

Second Edition

Encyclopedia of SOILS IN THE ENVIRONMENT

Volume One: Soil Biology



BIODIVERSITY,
ANTIBIOTICS, FOOD SECURITY
HUMAN AND ANIMAL HEALTH
PATHOGENS,
SOCIAL COHESION

Editors-in-Chief

Michael Goss and Margaret Oliver

Section Editors

Evgenia Blagodatskaya & Adrian Unc

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

SECOND EDITION

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

SECOND EDITION

EDITORS-IN-CHIEF

Professor Michael J Goss

Professor Emeritus, School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G 2W1, CANADA

Professor Margaret Oliver

Visiting Professor, Soil Research Centre, Department of Geography and Environmental Science, Russell Building, University of Reading, Whiteknights, Reading, RG6 6DR, United Kingdom

VOLUME 1

SOIL BIOLOGY

SECTION EDITORS

Evgenia Blagodatskaya

Dept. of Soil Ecology, Helmholtz-Zentrum für Umweltforschung GmbH – UFZ, Theodor-Lieser Str. 4, D-06120 Halle (Saale), GERMANY

Adrian Unc

Boreal Ecosystems and Agricultural Sciences, Grenfell Campus, Memorial University of Newfoundland, 20 University Drive, Corner Brook, Newfoundland A2H 5G4, CANADA



ELSEVIER

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge MA 02139, United States

Copyright © 2023 Elsevier Ltd. All rights reserved

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

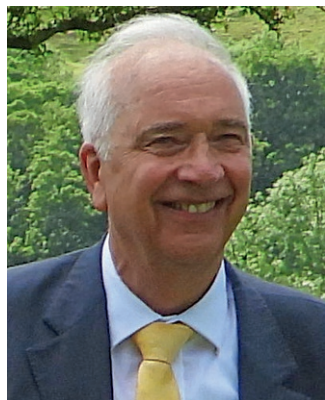
ISBN 978-0-12-822974-3

For information on all publications
visit our website at <http://store.elsevier.com>



Publisher: Oliver Walter
Acquisition Editor: Priscilla Braglia
Content Project Manager: Sona Chahal
Associate Content Project Manager: Nandhini Mahendran
Designer: Christian Bilbow

EDITORS BIOGRAPHY



Michael J Goss, PhD, MSc, BSc, is a professor emeritus in the School of Environmental Sciences at the University of Guelph. Before taking on the joint editor-in-chief responsibilities with Margaret Oliver, he held a similar position for *Soil Use and Management*, a journal of the British Society of Soil Science of which he is an honorary fellow. He was an inaugural member of the European Society for Agronomy. Michael started his career at the Letcombe Laboratory before moving to Rothamsted Experimental Station and the Macaulay Land Use Research Institute (now the James Hutton Research Institute). In 1990 he was appointed to the Research Chair in Land Stewardship at the University of Guelph and in 2005 became the director of its Regional Campus near Ottawa.

Michael's research has covered soil plant relations, soil management, and agriculture's footprint on the environment. His focus has been on the physical properties of soil and their impacts on root growth, which he showed to be influenced by the role of the root cap. He broadened his approach to include the effects of tillage and land drainage on root exploration. This work stimulated his interest in the transport of materials, such as plant nutrients, pathogenic microbes, antibiotics, and biologically active compounds through the soil and their contamination of water resources. Michael's knowledge in this field resulted in his involvement in the public inquiry into the death of seven people, including children, from a microbially contaminated drinking water supply. He contributed, both to the investigation into how the contamination occurred and if any fault could be attributed to local agricultural practices, and to how to prevent future events.

Michael also developed an interest in the interrelationships between soil microbes and root systems, particularly the tripartite symbiosis between roots of legumes, mycorrhizal fungi, and rhizobia. He has been an energetic member of collaborative research programs in France and Germany, and in Portugal, where he continues his interest in the role of mycorrhiza in sustainable land management.

Michael has authored more than 180 publications including two books, one in its 2nd edition, and edited two others.



Margaret Ann Oliver (BSc, PhD)

Margaret's interest in soil began in her penultimate year at school when learning about different types of soil and their formation. She chose the University of Bristol, United Kingdom for her first degree because the geography course offered soil science and she was also able to do a degree in geology alongside. Her love for both subjects remains undiminished. She did her PhD in soil spatial analysis at the University of Birmingham. Her early research focused on the multivariate and geostatistical analysis of soil data, which she continued to use in various research projects. Her research interests expanded to include the application of a wide range of numerical and geostatistical methods to soil and other data including pollen counts, forestry, radon emission, remotely sensed imagery, precision agriculture, and the incidence of childhood cancer. Her specific interests have been in sample design for spatial analysis and eventual mapping, risk analysis associated with prescribed thresholds in relation to soil pollutants or deficiencies of crop nutrients, and the relations between different scales of spatial variation. Her most

recent research was in the field of precision agriculture. At both the Universities of Birmingham and Reading, she taught topics in soil science, statistical methods, and spatial analysis. At Reading she had a research group working on various projects such as imagery, sampling and mapping, precision agriculture, and human health.

She has published more than 100 academic papers and has made several contributions to books. She is the co-author of three books: *Statistical Methods in Soil and Land Resource Survey*, Oxford University Press (Webster, R and Oliver, M. A. 1990); *Geostatistics for Environmental Scientists*, Second Edition, John Wiley & Sons (Webster, R and Oliver, M. A. 2007), and *How to do the Basic Steps in Geostatistics: The Variogram and Kriging*, Springer, Dordrecht (Oliver, M.A. & Webster, R.2015). She has edited two books: *Geostatistical Applications for Precision Agriculture* (Springer, Dordrecht, 2010) and *Precision Agriculture for Sustainability and Environmental Protection* (Earthscan from Routledge, 2013).

Margaret retired officially from Reading in 2004, but as yet has not managed to escape from academia. She has retained a link and a base at the University as a visiting professor since January 2005. After retirement, she developed and taught geostatistics courses to outside groups for several years. She then started on the long path of editorial work, as an associate editor of the *European Journal of Soil Science* and eventually the editor-in-chief, assistant editor of *Mathematical Geosciences*, co-editor of *Precision Agriculture* with John Stafford, and finally the editor-in-chief of the *Encyclopedia of Soils in the Environment* with Michael Goss.

Margaret was made an honorary member of the British Society of Soil Science in 2021 to reflect her commitment to soil science through fellow research, teaching, and editorial work. She was honored by the Pedometrics Commission of the International Union of Soil Science by having an award named after her in 2015. It is the Margaret Oliver award for young pedometricians.

EDITORIAL BOARD

Editors in Chief

Michael J Goss

Professor Emeritus, School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G 2W1, CANADA

Margaret Oliver

Visiting Professor, Soil Research Centre, Department of Geography and Environmental Science, Russell Building, University of Reading, Whiteknights, Reading, RG6 6DR, United Kingdom

Section Editors

Dr Evgenia Blagodatskaya

Dept. of Soil Ecology, Helmholtz-Zentrum für Umweltforschung GmbH – UFZ, Theodor-Lieser Str. 4, D-06120 Halle (Saale), GERMANY

Professor Adrian Unc

Boreal Ecosystems and Agricultural Sciences, Grenfell Campus, Memorial University of Newfoundland, 20 University Drive, Corner Brook, Newfoundland A2H 5G4, CANADA

Dr. Siobhán Staunton

INRA, UMR Eco&Sols, Montpellier SupAgro, bâtiment 12, 2 Place Viala, 34060 Montpellier Cedex 2, Montpellier, FRANCE

Professor Baoshan Xing

University of Massachusetts – Amherst, College of Natural Sciences, Stockbridge School of Agriculture, Environmental & Soil Chemistry, Program, 410 Paige Laboratory, Amherst, MA 01003, USA

Dr Karin Mueller

New Zealand Institute for Plant & Food Research Ltd., Ruakura Research Centre, Private Bag 3123, Hamilton, New Zealand

Professor Christina Siebe

Instituto de Geología, Circuito de la Investigación Científica s/n, Ciudad Universitaria, Ciudad de México, CP 04510, México

Professor Rupert Baeumler

Institute of Geography, Friedrich-Alexander-University of Erlangen-Nuremberg, Wetterkreuz 15, 91058 Erlangen, Germany

Dr Uta Stockmann

Soil Processes and Function Organisational Unit, CSIRO Agriculture and Food, Black Mountain, GPO BOX 1700, Canberra ACT 2601, AUSTRALIA

Professor Paul Hallett

*Room 2.23, Cruickshank Building, School of Biological Sciences, University of Aberdeen, Aberdeen, AB24 3UU,
UNITED KINGDOM*

Professor Dani Or

*Department of Environmental Systems Science (D-USYS), Institute of Biogeochemistry and Pollutant Dynamics (IBP),
Soil and Terrestrial Environmental Physics (STEP), CHN F 29.1 Universitaetstrasse 16, 8092 Zurich, SWITZERLAND*

FOREWORD

Soil is the porous “skin” on the surface of the Earth that has formed over thousands of years of weathering, comprising minerals, organisms (both dead and alive), water, and air. Soil is more than just the sum of its parts. It serves several crucial ecological functions, from food and fiber production to water storage and purification to waste decomposition.

Increasingly we understand that everything is connected, and that soil is a “Critical Zone” that intersects with the hydrosphere (water), lithosphere (minerals), atmosphere (air), and biosphere (all living organisms). To quote Daniel Hillel, the 2012 World Food Prize laureate and editor of the First Edition of *The Encyclopedia of Soils in the Environment*, “[Soil] is the crucible of terrestrial life, within which biological productivity is generated and sustained... Our civilization depends on the soil...” He also promoted the importance of understanding soil to manage it effectively. This revised and reorganized second edition of the Encyclopedia aims to reflect the changing status of soil in the world. Within these five volumes, over 724 experts from many countries have provided updated, easy-to-understand chapters on soil, its properties, and its role in the environment to affect climate change, plant productivity, human health, and biodiversity.

Many technological advances have been made since the first edition of this encyclopedia that has allowed us to investigate soil physical, chemical, and especially biological properties more precisely, enabling us to identify and understand their roles in ecosystem services. The study of soils, especially in the environmental context, is critical to meeting the challenges of feeding, clothing, and cleaning up the world. The information presented in these volumes will be useful to many, including soil and environmental scientists, policymakers, land managers, educators, students, and anyone interested in learning more about soil.

Soil is under considerable threat from many sources—building and infrastructure, climate change, population pressure, and pollution, yet it is vital to agronomic productivity and continues to provide critical ecosystem services. Society needs to be engaged in how soil is managed and conserved. Most natural scientists are very dedicated and generous, including all of the authors in this Encyclopedia who have freely given their time to make these volumes so comprehensive. The study of soil, while often challenging, is fascinating—soil is a beautiful earthy material that can delight our senses with its variegated colors, gritty to smooth textures, and especially its characteristic aroma when freshly turned. Perhaps that is why gardeners and even some farmers tend to overwork their soil to its detriment—it *just feels good to get your hands in the dirt!*

Soil science is one of the most exciting, interesting, and rewarding disciplines, largely because it is multi-dimensional and interdisciplinary, basic and applied. It affects all of our lives. Thus, I encourage you to peruse these volumes and discover all kinds of interesting aspects of soil in the environment.

April Ulery

Former President, Soil Science Society of America
Professor, Soil & Environmental Sciences
New Mexico State University
Las Cruces, NM
The United States

PREFACE

This second edition of the *Encyclopedia of Soils in the Environment* follows in the footsteps of a highly rated first edition overseen by Professor Daniel Hillel as editor in chief, ably assisted by six editors. That edition provided a very valuable resource on the “state-of-the-art” in soil science in the context of the environment in the closing years of the twentieth century and the first three years of the current century. It set a standard that the editorial team for the second edition has aimed to match. The soil and environmental sciences have made huge strides in the almost 20 years since the first edition was published, and the new team has had to embrace these developments while retaining the basic components of these sciences. The project has taken 4 years to complete. It began with the appointment of the two editors-in-chief, professors Michael J. Goss and Margaret A. Oliver, followed by that of the nine subject editors and a pedology adviser.

The project was in its infancy (a few months only) when COVID-19 struck. The editors-in-chief discussed the format and structure of the second edition around the time that the first “lockdowns” around the world were being introduced. The first edition was alphabetically organized, whereas we decided to have a section-oriented approach in this edition, which we hope gives a more structured feel for the various topics embraced. The principal sections are soil biology, soil chemistry, the soil environment and its management, pedology (the origins and development of soil), pedometrics (the application of mathematics and statistics to soil), and soil physics. We were able to give our full attention to this development and to the selection of the scientists to join us on the road ahead. Furthermore, we were fortunate to choose a team that has maintained its enthusiasm despite the many tribulations we have faced along this path. The team comprises Jane Blagodatskaya and Adrian Unc (biology and microbiology), Siobhan Staunton and Baoshan Xing (chemistry), Karin Müller and Christina Siebe (environment and management), Rupert Bäumler (pedology), Uta Stockmann and Margaret Oliver (pedometrics), and Paul Hallett and Dani Or (physics). The editorial board has been mindful in their choice of subjects of the need for basic soil science and for state-of-the-art developments, with the aim that the audience for these volumes can be as large as possible.

During the gestation period of the Encyclopedia, methods of working for those in academia were turned on their heads. The change from “face-to-face” teaching to formulating and establishing online courses, initially for lockdown periods but then for much of the next 2 years, had huge consequences for academic staff worldwide. This meant that they had less time to devote to outside activities, such as writing chapters for an encyclopedia. As a result, some authors were unable to complete their chapters in time, leaving a few unfortunate gaps. The whole team has done as much as they can to limit any adverse effects, but we ask readers to appreciate the struggle that authors have had with illness (personal, within their families and work colleagues) and a great increase in their workload. Despite this, we are presenting a comprehensive document that should stand the test of the next 20 years of soils in the environment.

This century has seen an almost unprecedented interest in the soil and environmental sciences. It has been triggered not only by issues that the Earth is facing from global climate change, increasing population pressure, soil and land degradation, and food insecurity but also by the many new and sophisticated methods available, such as the DNA determination of microbial populations and their functional capabilities, proximal sensing, machine learning methods, and an ever-increasing array of technological tools. These developments have greatly increased our knowledge and appreciation of the complexity of soils, which has changed our understanding of some well-established aspects of the science. The soil and its environment are at the critical interface of the atmosphere, biosphere, hydrosphere, and lithosphere. They deserve to be taken seriously by the international community in the same way that air and water have been. Governments are now paying attention

to the soil and environment, and a legal and governance framework is emerging to look after these parts of our life support resources.

We have aimed, through our editing of chapters, to make the language of the chapters accessible to all, regardless of educational background. This does not mean that we have diminished the scientific content, but that readers should be able to gain the foothold they desire in these subjects to progress to greater knowledge and understanding of them. We have retained the overview of soil written by Daniel Hillel for the First Edition, first to honor his great contribution to soil science and second to demonstrate the significance of soil. It follows this Preface.

*Michael J. Goss
Margaret A. Oliver*

SOIL

The term “soil” refers to the weathered and fragmented outer layer of our planet’s land surfaces. Formed initially through the physical disintegration and chemical alteration of rocks and minerals by physical and biogeochemical processes, soil is influenced by the activity and accumulated residues of a myriad of diverse forms of life. As it occurs in different geologic and climatic domains, soil is an exceedingly varied body with a wide range of attributes.

Considering the height of the atmosphere, the thickness of the earth’s rock mantle, and the depth of the ocean, one observes that soil is an amazingly thin body—typically not much more than one meter thick and often less than that. Yet it is the fount of terrestrial life, within which biological productivity is generated and sustained. It acts like a composite living entity, a home to a community of innumerable microscopic and macroscopic plants and animals. A mere fistful of soil typically contains billions of microorganisms, which perform vital interactive biochemical functions. Another intrinsic attribute of the soil is its sponge-like porosity and its enormous internal surface area. That same fistful of soil may actually consist of several hectares of active surface, upon which physicochemical processes take place continuously.

Realizing humanity’s utter dependence on the soil, ancient peoples, who lived in greater closeness to nature than many of us today, actually revered the soil. It was not only their source of livelihood but also the material from which they built their homes and that they learned to shape, heat, and fuse into household vessels and writing tablets (ceramics, made of clayey soil, being the first synthetic material in the history of technology). In the Bible, the name assigned to the first human was Adam, derived from “adama,” meaning soil. The name given to the first earthling’s mate was Hava (Eve, in transliteration), meaning “living” or “life-giving.” Together, therefore, Adam and Eve signified quite literally “Soil and Life.”

The same powerful metaphor is echoed in the Latin name for human species—Homo, derived from humus, the material of the soil. Hence, the adjective “human” also implies “of the soil.” Other ancient cultures evoked equally powerful associations. To the Greeks, the earth was a manifestation of Gaea, the maternal goddess who, impregnated by Uranus (god of the sky), gave birth to all the gods of the Greek pantheon.

Our civilization depends on the soil more crucially than ever because our numbers have grown while available soil resources have diminished and deteriorated. Paradoxically, however, even as our dependence on the soil has increased, most of us have become physically and emotionally detached from it. Many of the people in the so-called developed countries spend their lives in the artificial environment of a city, insulated from direct exposure to nature, and some children may now assume as a matter of course that food originates in supermarkets.

Detachment has bred ignorance, and out of ignorance has come the delusion that our civilization has risen above nature and has set itself free of its constraints. Agriculture and food security, erosion and salination, degradation of natural ecosystems, depletion and pollution of surface waters and aquifers, and decimation of biodiversity—all these processes, which involve the soil directly or indirectly, have become abstractions to many people. The very language we use betrays disdain for that common material underfoot, often referred to as “dirt.” Some fastidious parents prohibit their children from playing in the mud and hurry to wash their “soiled” hands when the children nonetheless obey an innate instinct to do so. Thus, soil is devalued and treated as unclean though it is the terrestrial realm’s principal medium of purification, wherein wastes are decomposed and nature’s productivity is continually rejuvenated.

Scientists who observe soil closely see it in effect as a seething foundry in which matter and energy are in constant flux. Radiant energy from the sun streams onto the field and cascades through the soil and the plants growing in it. Heat is exchanged, water percolates through the soil’s intricate passages, and plant roots extract

water and transmit it to their leaves, which transpire it back to the atmosphere. Leaves absorb carbon dioxide from the air and synthesize it with soil-derived water to form the primary compounds of life. Oxygen emitted by the leaves makes the air breathable for animals, which consume and in turn fertilize plants.

Soil is thus a self-regulating bio-physio-chemical factory, processing its own materials, water, and solar energy. It also determines the fate of rainfall and snowfall reaching the ground surface—whether the water thus received will flow over the land as runoff, or seep downward to the subterranean reservoir called groundwater, which in turn maintains the steady flow of springs and streams. With its finite capacity to absorb and store moisture, and to release it gradually, the soil regulates all these phenomena. Without the soil as a buffer, rain falling over the continents would run off entirely, producing violent floods rather than sustained river flow.

Soil naturally acts as a living filter, in which pathogens and toxins that might otherwise accumulate to foul the terrestrial environment are rendered harmless. Since time immemorial, humans and other animals have been dying of all manner of diseases and have then been buried in the soil, yet no major human disease is transmitted by it. The term antibiotic was coined by soil microbiologists who, as a consequence of their studies of soil bacteria, specifically actinomycetes, discovered streptomycin (an important cure for tuberculosis and other infections). Ion exchange, a useful process of water purification, also was discovered by soil scientists studying the passage of solutes through beds of clay.

However unique in form and function, soil is not an isolated body. It is, rather, a central link in the larger chain of interconnected domains and processes comprising the terrestrial environment. The soil interacts both with the overlying atmosphere and the underlying strata, as well as with surface and underground bodies of water. Especially important is the interrelation between the soil and climate. In addition to its function of regulating the cycle of water, it also regulates energy exchange and surface temperature.

When land is cleared of vegetation and turned into a cultivated field, the native biomass above the ground is often burned and the organic matter within the soil tends to decompose. These processes release carbon dioxide into the atmosphere, thus contributing to the Earth's greenhouse effect and to global warming. On the other hand, the opposite act of afforestation and soil enrichment with organic matter, such as can be achieved through conservation management, may serve to absorb carbon dioxide from the atmosphere. To an extent, the soil's capacity to store carbon thus helps to mitigate the greenhouse effect.

Thousands of years are required for nature to create life-giving soil out of sterile bedrock. In only a few decades, however, unknowing or uncaring humans can destroy that wondrous work of nature. In various circumstances, mismanaged soils may be subject to erosion (the sediments of which tend to clog streambeds, estuaries, lakes, and coastal waters), to leaching of nutrients with attendant loss of fertility and eutrophication of water bodies, to waterlogging and impaired aeration, or to an excessive accumulation of salts that may cause a once-productive soil to become entirely sterile. Such processes of soil degradation, sometimes called "desertification," already affect large areas of land.

We cannot manage effectively and sustainably that which we do not know and thoroughly understand. That is why the tasks of developing and disseminating sound knowledge of the soil and its complex processes have assumed growing urgency and importance. The global environmental crisis has created a compelling need for a concentrated, concise, and definitive source of information—accessible to students, scientists, practitioners, and the general public—about the soil in all its manifestations in nature and in relation to the life of humans.

Daniel Hillel
May 2004

LIST OF CONTRIBUTORS FOR VOLUME 1

Nathalie Amacker

Ecology and Biodiversity group, Institute of Environmental Biology, University of Utrecht, Utrecht, The Netherlands

Jasmine Anenberg

School of Forestry, Northern Arizona University, Flagstaff, AZ, United States

Anita Antoninka

School of Forestry, Northern Arizona University, Flagstaff, AZ, United States

Lydia Bailey

School of Forestry, Northern Arizona University, Flagstaff, AZ, United States; Center for Ecosystem Science and Society, Northern Arizona University, Flagstaff, AZ, United States

Matthew Bakker

Department of Microbiology, University of Manitoba, Winnipeg, Canada

Isabelle Basile-Doelsch

Aix Marseille Univ, CNRS, IRD, INRAE, CEREGE, Aix-en-Provence, France

Karen Baumann

Faculty of Agricultural and Environmental Sciences, Soil Science, University of Rostock, Rostock, Germany

Ernest C Bernard

Department of Entomology & Plant Pathology, University of Tennessee, Knoxville, TN, United States

Li Bi

School of Agriculture and Food, Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, VIC, Australia

Nataliya Bilyera

Geo-Biosphere Interactions, Department of Geoscience, University of Tuebingen, Tuebingen, Germany

Sergey Blagodatsky

Institute of Zoology, University of Cologne, Cologne, Germany

Matthew A Bowker

School of Forestry, Northern Arizona University, Flagstaff, AZ, United States; Center for Ecosystem Science and Society, Northern Arizona University, Flagstaff, AZ, United States

Maria JI Briones

Departamento de Ecología y Biología Animal, Universidad de Vigo, Vigo, Spain

Isabel Brito

MED – Mediterranean Institute for Agriculture, Environment and Development & CHANGE – Global Change and Sustainability Institute, University of Évora, School of Science and Technology, Évora, Portugal

Cynthia Brown

Agricultural Biology, Colorado State University, Fort Collins, CO, United States

F Buscot

Helmholtz Centre for Environmental Research (UFZ), Halle (Saale), Germany

Peng Cai

State Key Laboratory of Agricultural Microbiology, College of Resources and Environment, Huazhong Agricultural University, Wuhan, China

Alberto Canarini

Centre for Microbiology and Environmental Systems Science, University of Vienna, Vienna, Austria; Center for Ecological Research, Kyoto University, Otsu, Shiga, Japan

Mário Carvalho

MED – Mediterranean Institute for Agriculture, Environment and Development & CHANGE – Global Change and Sustainability Institute, University of Évora, School of Science and Technology, Évora, Portugal

Pippa Chapman

School of Geography, University of Leeds, Leeds, United Kingdom

Tiphaine Chevallier
*IRD, INRAE, CIRAD, Institut Agro, Université
Montpellier, Eco&Sols, Montpellier, France*

Bee Khim Chim
*University of Maine at Presque Isle, Presque Isle, ME,
United States*

Louise Comas
*United States Department of Agriculture—Agricultural
Research Service, Fort Collins, CO, United States*

KA Congreves
*Department of Plant Sciences, University of Saskatchewan,
Saskatoon, SK, Canada*

EJ Daly
*Department of Renewable Resources, University of Alberta,
Edmonton, AB, Canada*

Aidan De Sena
*Department of Bioresource Engineering, Macdonald
Campus of McGill University, Ste-Anne-de-Bellevue, QC,
Canada*

Marie-France Dignac
*INRAE, CNRS, Sorbonne Université, IRD, Institute of
Ecology and Environmental Sciences of Paris (iEES-Paris),
Paris, France*

Fang Ding
*Chinese Academy of Sciences, Institute of Applied Ecology,
Shenyang, China*

Patrice Dion
*Institut de biologie Intégrative et des systèmes (IBIS),
Pavillon Charles-Eugène Marchand, Laval University,
Québec, QC, Canada*

Maxim Dorodnikov
*Department of Soil Science of Temperate Ecosystems,
University of Goettingen, Goettingen, Germany; Institute of
Landscape Ecology, University of Muenster, Muenster,
Germany*

Kari Dunfield
*School of Environmental Sciences, University of Guelph,
Guelph, ON, Canada*

Amandine Erktan
*IRD, INRAE, CIRAD, Institut Agro, Université
Montpellier, Eco&Sols, Montpellier, France*

Monica A Farfan
*School of Global Environmental Sustainability,
Colorado State University, Fort Collins, CO, United
States*

Grégoire Freschet
*Centre National de la Recherche Scientifique, Station
d'Écologie Théorique et Expérimentale, Moulis, France*

Lucia Fuchslueger
*Centre for Microbiology and Environmental
Systems Science, University of Vienna, Vienna,
Austria*

Marcelo Valadares Galdos
*Rothamsted Research, Sustainable Soils and Crops,
Harpenden, United Kingdom*

Elisa Gamalero
*Dipartimento di Scienze e Innovazione Tecnologica,
Università del Piemonte Orientale, Alessandria, Italy*

Chunhui Gao
*State Key Laboratory of Agricultural Microbiology, College
of Resources and Environment, Huazhong Agricultural
University, Wuhan, China*

Magda Garbowski
*Department of Botany, University of Wyoming,
Laramie, WY, United States*

Stefan Geisen
*Laboratory of Nematology, Department of Plant Sciences,
Wageningen University and Research, Wageningen, The
Netherlands*

Charles P Gerba
University of Arizona, Tucson, AZ, United States

Bernard R Glick
*Department of Biology, University of Waterloo, Waterloo,
ON, Canada*

María Gómez-Brandón
*Grupo de Ecología Animal (GEA), Centro de
Investigación Mariña (CIM), Universidad de Vigo, Vigo,
Galicia, Spain*

Michael J Goss
*School of Environmental Sciences, University of Guelph,
Guelph, ON, Canada*

RM Goss
*Plant & Environmental Sciences Department, New Mexico
State University, Las Cruces, NM, United States*

Andrey Guber
*Department of Plant, Soil, and Microbial Sciences,
Michigan State University, East Lansing, MI,
United States*

Shamim Gul
*Department of Natural Resource Sciences, Macdonald
Campus of McGill University, Sainte-Anne-de-Bellevue,
QC, Canada; Department of Botany, University of
Balochistan, Quetta, Pakistan*

Vadakattu VSR Gupta
CSIRO Ag & Food, Waite Road, Urrbrae, SA, Australia

Li-Li Han

*State Key Laboratory of Urban and Regional Ecology,
Research Center for Eco-Environmental Sciences, Chinese
Academy of Sciences, Beijing, China*

PB Hatzinger

*Biotechnology Development and Applications, APTIM,
Lawrenceville, NJ, United States*

Ji-Zheng He

*School of Agriculture and Food, Faculty of Veterinary and
Agricultural Sciences, The University of Melbourne,
Parkville, VIC, Australia*

Angela Hodge

*Department of Biology, University of York, York, United
Kingdom*

Joseph Holden

*School of Geography, University of Leeds, Leeds, United
Kingdom*

Lijun Hou

*Department of Natural Resource Sciences, Macdonald
Campus of McGill University, Sainte-Anne-de-Bellevue,
QC, Canada*

Hang-Wei Hu

*School of Agriculture and Food, Faculty of Veterinary and
Agricultural Sciences, The University of Melbourne,
Parkville, VIC, Australia*

Qiaoyun Huang

*State Key Laboratory of Agricultural Microbiology, College
of Resources and Environment, Huazhong Agricultural
University, Wuhan, China*

PW Inglett

*University of Florida, Soil and Water Sciences Department,
Wetland Biogeochemistry Laboratory, Gainesville, FL,
United States*

Heribert Insam

*Department of Microbiology, University of Innsbruck,
Innsbruck, Austria*

Kristina Ivashchenko

*Agrarian-Technological Institute, RUDN University,
Moscow, Russia; Institute of Physicochemical and Biological
Problems in Soil Science, Russian Academy of Sciences,
Pushchino, Russia*

Louise Jackson

*Department of Land, Air, and Water Resources, University
of California, Davis, CA, United States*

Yutong Jiang

*Department of Natural Resource Sciences, Macdonald
Campus of McGill University, Sainte-Anne-de-Bellevue,
QC, Canada*

François-Xavier Joly

*Centre for Microbiology and Environmental Systems
Science, University of Vienna, Vienna, Austria; Biological
and Environmental Sciences, University of Stirling,
Scotland, United Kingdom*

Alexandre Jousset

*Ecology and Biodiversity group, Institute of Environmental
Biology, University of Utrecht, Utrecht, The Netherlands*

Josiah B Judson

*School of Geography, University of Leeds, Leeds, United
Kingdom*

Patrick Jung

*Applied Logistics and Polymer Sciences, University of
Applied Sciences Kaiserslautern, Pirmasens, Germany*

Hilina Yohannes Kebede

*School of Forestry, Northern Arizona University, Flagstaff,
AZ, United States*

JW Kelsey

*Environmental Science and Chemistry, Muhlenberg
College, Allentown, PA, United States*

Thomas Klammersteiner

*Department of Microbiology, University of Innsbruck,
Innsbruck, Austria; Department of Ecology, University of
Innsbruck, Innsbruck, Austria*

Maria Korneykova

*Agrarian-Technological Institute, RUDN University,
Moscow, Russia; Institute of North Industrial Ecology
Problems, Kola Science Centre of Russian Academy of
Sciences, Apatity, Russia*

Ekaterina Kozlova

*Agrarian-Technological Institute, RUDN University,
Moscow, Russia*

Alexandra Kravchenko

*Department of Plant, Soil, and Microbial Sciences,
Michigan State University, East Lansing, MI, United States*

Iryna Loginova

*Geo-Biosphere Interactions, Department of Geoscience,
University of Tuebingen, Tuebingen, Germany*

Naga Raju Maddela

*Departamento de Ciencias Biológicas, Facultad de Ciencias
de la Salud, Universidad Técnica de Manabí, Portoviejo,
Ecuador*

Chandra A Madramootoo

*Department of Bioresource Engineering, Macdonald
Campus of McGill University, Ste-Anne-de-Bellevue, QC,
Canada*

Tolulope Mafa-Attoye

*School of Environmental Sciences, University of Guelph,
Guelph, ON, Canada*

Ali Akbar Mahboubi

Department of Soil Science, College of Agriculture, Bu-Ali Sina University, Hamadan, Iran; School of Engineering, University of Guelph, Guelph, Ontario, Canada

Evan Mayer

School of Environmental Sciences, University of Guelph, Guelph, ON, Canada

Mallavarapu Megharaj

Global Centre for Environmental Remediation (GCER), and Cooperative Research Centre for Contamination Assessment and Remediation of the Environment (CRC CARE), College of Engineering, Science and Environment, The University of Newcastle, Callaghan, NSW, Australia

Eduardo K Mitter

School of Environmental Sciences, University of Guelph, Guelph, ON, Canada

Mehrdad Mohammadiani

Department of Microbiology, University of Manitoba, Winnipeg, Canada

Mohammad Reza Mosaddeghi

Department of Soil Science, College of Agriculture, Isfahan University of Technology, Isfahan, Iran

Krisztina Mosdossy

Department of Natural Resource Sciences, Macdonald Campus of McGill University, Sainte-Anne-de-Bellevue, QC, Canada

Florence Mus

Institute of Biological Chemistry, Washington State University, Pullman, WA, United States

Deborah A Neher

Department of Plant and Soil Science, University of Vermont, Burlington, VT, United States

Jeanette M Norton

Department of Plants, Soils and Climate and Ecology Center, Utah State University, Logan, UT, United States

Dasiel Obregon

School of Environmental Sciences, University of Guelph, Guelph, ON, Canada

Maureen O'Callaghan

Resilient Agriculture, AgResearch, Christchurch, New Zealand

Yang Ouyang

Department of Microbiology and Plant Biology, Institute of Environmental Genomics, University of Oklahoma, Norman, OK, United States

NS Panikov

Chemistry and Chemical Biology Department, Northeastern University, Boston, MA, United States

Carlos B Pires

Kansas State University, Manhattan, KS, United States

Thomas O Powers

Department of Plant Pathology, University of Nebraska, Lincoln, NE, United States

Balasubramanian Ramakrishnan

Division of Microbiology, ICAR-Indian Agricultural Research Institute, New Delhi, India

K Ramesh Reddy

University of Florida, Soil and Water Sciences Department, Wetland Biogeochemistry Laboratory, Gainesville, FL, United States

Natasha Rayne

University of Wisconsin River Falls, River Falls, WI, United States

Marc Redmile-Gordon

Department of Environmental Horticulture, Royal Horticultural Society, Wisley, United Kingdom

Charles W Rice

Kansas State University, Manhattan, KS, United States

Andreas Richter

Centre for Microbiology and Environmental Systems Science, University of Vienna, Vienna, Austria; International Institute for Applied Systems Analysis, Advancing Systems Analysis Program, Laxenburg, Austria

Sarah S Roley

School of the Environment, Washington State University, Richland, WA, United States

Azadeh Safadoust

Department of Soil Science, College of Agriculture, Bu-Ali Sina University, Hamadan, Iran

D Said-Pullicino

University of Torino, Department of Agricultural, Forest and Food Sciences, Torino, Italy

Gustavo Santoyo

Instituto de Investigaciones Químico Biológicas, Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Mexico

Marcos VM Sarto

Kansas State University, Manhattan, KS, United States

Mikhail V Semenov

Dokuchaev Soil Science Institute, Laboratory of Soil Carbon and Microbial Ecology, Moscow, Russian Federation

Balwant Singh

Sydney Institute of Agriculture, School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW, Australia

Heather Slinn
*School of Environmental Sciences, University of Guelph,
Guelph, ON, Canada*

Donald L Sparks
*Department of Plant and Soil Sciences, University of
Delaware, Newark, DE, United States*

Micaela Tosi
*School of Environmental Sciences, University of Guelph,
Guelph, ON, Canada*

AL Ulery
*Plant & Environmental Sciences Department, New Mexico
State University, Las Cruces, NM, United States*

Adrian Unc
*School of Science and the Environment, Memorial
University of Newfoundland, Corner Brook, NL, Canada*

Jeremiah D Vallotton
*Environmental Science Program, Memorial University of
Newfoundland, St. John's, NL, Canada*

Viacheslav Vasenev
*Soil Geography and Landscape Group, Wageningen
University, Wageningen, The Netherlands*

Maria Vasilieva
*Agrarian-Technological Institute, RUDN University,
Moscow, Russia*

Kadiyala Venkateswarlu
*Formerly Department of Microbiology, Sri Krishnadevaraya
University, Anantapuramu, India*

Diana H Wall
*School of Global Environmental Sustainability, Colorado
State University, Fort Collins, CO, United States*

Chaoqun Wang
*Biogeochemistry of Agroecosystems, University of
Goettingen, Goettingen, Germany*

Joann K Whalen
*Department of Natural Resource Sciences, Macdonald
Campus of McGill University, Sainte-Anne-
de-Bellevue, QC, Canada; Chair of Soil Science,
Mohammed VI Polytechnic University, Ben Guerir,
Morocco*

David Wright
*Resilient Agriculture, AgResearch, Christchurch, New
Zealand*

Hsin-Hua Wu
*Institute of Biological Chemistry, Washington State
University, Pullman, WA, United States; Department of
Chemistry, Washington State University, Pullman, WA,
United States*

Yichao Wu
*State Key Laboratory of Agricultural Microbiology, College
of Resources and Environment, Huazhong Agricultural
University, Wuhan, China*

Sabina Yeasmin
*Department of Agronomy, Bangladesh Agricultural
University, Mymensingh, Bangladesh*

Dan-Ting Yu
*State Key Laboratory for Subtropical Mountain
Ecology of the Ministry of Science and Technology
and Fujian Province, Fujian Normal University, Fujian,
China*

CONTENTS OF ALL VOLUMES

VOLUME 1: SOIL BIOLOGY

Overview

Edaphology	1
<i>RM Goss and AL Ulery</i>	

Biodiversity	10
<i>Monica A Farfan and Diana H Wall</i>	

Soil organisms

Bacteriophage (Viruses)	16
<i>Li Bi, Ji-Zheng He, Dan-Ting Yu, Li-Li Han, and Hang-Wei Hu</i>	

Archaea	25
<i>Mikhail V Semenov</i>	

Bacteria	31
<i>Mikhail V Semenov</i>	

Soil protists	39
<i>Stefan Geisen, Nathalie Amacker, and Alexandre Jousset</i>	

Fungi including mycorrhizal fungi	49
<i>F Buscot</i>	

Soil algae	61
<i>Karen Baumann and Patrick Jung</i>	

Soil arthropods: Underfoot and all around	70
<i>Ernest C Bernard</i>	

Nematodes	105
<i>Deborah A Neher and Thomas O Powers</i>	

Earthworms: Essential ecosystem engineers providing vital ecosystem services	112
<i>Maria JI Briones</i>	

Complex structures

Soil biofilms: Evolving concepts and ecological functions	120
<i>Yichao Wu, Marc Redmile-Gordon, Peng Cai, Chunhui Gao, and Qiaoyun Huang</i>	

Biocrusts	127
<i>Matthew A Bowker, Jasmine Anenberg, Hilina Yohannes Kebede, Anita Antoninka, and Lydia Bailey</i>	

Functioning

Studying soil microbial communities: Culture-dependent approaches	139
<i>Eduardo K Mitter, Dasiel Obregon, Micaela Tosi, Heather Slinn, and Kari Dunfield</i>	

Studying soil microbial communities: Culture-independent approaches	146
<i>Tolulope Mafa-Attoye, Heather Slinn, Micaela Tosi, Evan Mayer, Dasiel Obregon, and Kari Dunfield</i>	

Studying microbial function in soil: Culture-independent techniques <i>Micaela Tosi, Dasiel Obregon, Heather Slinn, and Kari Dunfield</i>	155
Kinetics of microbial processes: General principles <i>NS Panikov</i>	168
Kinetics of microbial processes: Applications to soil biology <i>NS Panikov</i>	186
Biogeochemical cycles	
Carbon cycle in soils: Dynamics and management <i>Charles W Rice, Carlos B Pires, and Marcos VM Sarto</i>	219
Carbon storage in soils <i>Isabelle Basile-Doelsch, Tiphaine Chevallier, Marie-France Dignac, and Amandine Erktan</i>	228
Soil organic nitrogen <i>EJ Daly and KA Congreves</i>	243
Nitrification: Process, organisms and management <i>Jeanette M Norton and Yang Ouyang</i>	253
Microbial denitrification <i>Michael J Goss, Sergey Blagodatsky, and Adrian Unc</i>	261
Symbiotic nitrogen fixation <i>Florence Mus and Hsin-Hua Wu</i>	270
Nitrogen: Non-Symbiotic Nitrogen fixation in soils <i>Vadakattu VSR Gupta and Sarah S Roley</i>	283
Phosphorus in soils—Biological interactions <i>Nataliya Bilyera and Iryna Loginova</i>	293
Biological transformations of sulfur in soil <i>Chaoqun Wang and Maxim Dorodnikov</i>	303
Plant uptake of mineral nutrients <i>Angela Hodge</i>	310
Higher plants and their interactions with soil organisms	
Soil Biology: Root form and function <i>Magda Garbowski, Grégoire Freschet, Louise Jackson, Cynthia Brown, and Louise Comas</i>	321
Root mucilage: Chemistry and functions in soil <i>Yutong Jiang, Lijun Hou, Fang Ding, and Joann K Whalen</i>	332
Root exudates and microorganisms <i>Aidan De Sena, Krisztina Mosdossy, Joann K Whalen, and Chandra A Madramootoo</i>	343
Root interactions with the microbiome from the rhizoplane to the bulk soil: An overview <i>Joann K Whalen and Shamim Gul</i>	357
Microorganisms & biological processes	
Soil respiration <i>Jeremiah D Vallotton, Sergey Blagodatsky, and Adrian Unc</i>	369
Anaerobic soil processes <i>PW Inglett, D Said-Pullicino, and K Ramesh Reddy</i>	379
Mineral-organic-microbial interactions <i>Balwant Singh, Sabina Yeasmin, and Donald L Sparks</i>	387

Microbial ecology & management

Habitat for soil organisms and their functions <i>Alexandra Kravchenko and Andrey Guber</i>	407
Plant growth promoters <i>Elisa Gamalero, Gustavo Santoyo, and Bernard R Glick</i>	418
Inoculation with microorganisms <i>Maureen O'Callaghan and David Wright</i>	428
Biocontrol of soil-borne plant diseases <i>Mehrdad Mohammadiani and Matthew Bakker</i>	439
Management of soil biology for crop protection from biotic and abiotic stresses <i>Michael J Goss, Isabel Brito, and Mário Carvalho</i>	447
Trees, hedges, agroforestry and microbial diversity <i>Josiah B Judson, Joseph Holden, Pippa Chapman, and Marcelo Valadares Galdos</i>	469
Urban soil microbiome: Activity, diversity and functioning <i>Maria Korneykova, Ekaterina Kozlova, Kristina Ivashchenko, Maria Vasilieva, and Viacheslav Vasenev</i>	480
Microbiology of extreme soil environments <i>Patrice Dion</i>	494
Microbial mobility and transport in soils <i>Mohammad Reza Mosaddeghi, Adrian Unc, Ali Akbar Mahboubi, and Azadeh Safadoust</i>	512

Biology of soil amendments & contaminant degradation

Biology of compost <i>Heribert Insam, Thomas Klammersteiner, and María Gómez-Brandón</i>	522
Landfarming and its microbiology <i>Naga Raju Maddela, Balasubramanian Ramakrishnan, Kadiyala Venkateswarlu, and Mallavarapu Megharaj</i>	533
Biodegradation of organic contaminants <i>PB Hatzinger and JW Kelsey</i>	547
Biological aspects of manure management <i>Natasha Rayne and Bee Khim Chim</i>	558
Septic systems and municipal waste disposal on land <i>Charles P Gerba, Michael J Goss, and Adrian Unc</i>	567

Climate change

Climate change impacts on soil biology <i>Alberto Canarini, Lucia Fuchslueger, François-Xavier Joly, and Andreas Richter</i>	578
---	-----

VOLUME 2: CHEMISTRY**Basic Soil Chemistry**

Acidity and acidification <i>Siobhan Staunton</i>	1
pH <i>Siobhan Staunton</i>	8
Buffering capacity <i>Siobhan Staunton and Ravi Naidu</i>	15
Soil specific surface area <i>Emmanuel Arthur, Markus Tuller, and Kurt D Pennell</i>	22

Soil Analytical Chemistry

Cation exchange capacity and reactions <i>Donald S Ross and Ulf Skyllberg</i>	32
Testing for nutrients in the soil and the environment <i>Antonio P Mallarino</i>	43
Metal extraction and analysis <i>Guo Wang and Honghong Li</i>	50
Soil organic matter: Composition <i>Markus Kleber</i>	58
Physical fractionation techniques <i>Sam J Leuthold, Michelle L Haddix, Jocelyn Lavallee, and M Francesca Cotrufo</i>	68
The use of stable isotopes in soil science: Low atomic number elements <i>Chiara Pistocchi, Isabelle Bertrand, Kristýna Kantnerová, Astrid Oberson, Federica Tamburini, and Roland A Werner</i>	81
The use of radioactive isotopes to study soil pools and processes <i>Timothy I McLaren</i>	96
The use of stable isotopes in soil science: Metals <i>Cécile Quantin and Damien Guinoiseau</i>	107

Periodic Table

Alkali metals in soils <i>Guy J Levy and Asher Bar-Tal</i>	116
Alkaline earth metals in soil <i>Justin B Richardson</i>	126
Transition metals in soil <i>Siobhan Staunton</i>	135
Nitrogen and ammonia in soils <i>Nicholas T Girkin and Hannah V Cooper</i>	142
Phosphorus in soil <i>Luisella Celi, Maria Martin, and Elisabetta Barberis</i>	152
Aluminum in soils <i>Syuntaro Hiradate</i>	167
Iron in soils <i>Trolard Fabienne and Guilhem Bourrié</i>	176
Sulfur and selenium in soil <i>Dongli Liang, Fei Zhou, and Mingxing Qi</i>	192
Arsenic in soil <i>Md Imran Ullah Sarkar, Afsana Jahan, Md Tofail Hossain, Mohammad Mahmudur Rahman, and Ravi Naidu</i>	201
Non-transition elements: Metals and metalloids in soils <i>Natasha, Sana Khalid, Muhammad Shahid, Faiz Rabbani, and Camille Dumat</i>	214
Lanthanides, actinides and radon <i>Franco Ajmone Marsan and Mery Malandrino</i>	225
Halogens in soils <i>Svensson Teresia, Thiry Yves, Bueno Maïté, and Oelmann Yvonne</i>	234
Transformation of metals and metalloids by microorganisms <i>Susan M Glasauer, Euan P Burford, and Geoffrey Michael Gadd</i>	243
Dynamics of radionuclides <i>Siobhan Staunton</i>	256
Silicon and other non-metal elements from Group 14 <i>Bruno Delvaux and Zimin Li</i>	263

Soil Physical Chemistry

Chemical equilibria	274
<i>Arthur Paul Schwab</i>	
Dissolution and precipitation processes	284
<i>Zhenqing Shi, Shiwen Hu, Rong Li, and Yang Lu</i>	
Reactions, redox potential and kinetics	297
<i>Trolard Fabienne and Guilhem Bourrié</i>	
Sorption of organic chemicals in soil	315
<i>Joseph J Pignatello</i>	
Sorption—Oxyanions	327
<i>CP Schulthess, H Wijnga, W Yang, Anahita Khosravi, and Baoshan Xing</i>	
Sorption—Metals	336
<i>Daniel G Strawn and Donald L Sparks</i>	
Sorption-desorption kinetics	343
<i>Gautier Landrot and Donald L Sparks</i>	
Sorption-desorption hysteresis	355
<i>Wenhao Wu, Hongxia Zhu, and Kun Yang</i>	
Surface charge and zero-charge points	368
<i>J Chorover</i>	
Flocculation and dispersion	376
<i>Guy J Levy and Isaac Shainberg</i>	

Soil Carbon Chemistry

Soil organic matter dynamics	387
<i>Yu Luo and Jianming Xu</i>	
Soil proteins	396
<i>Siobhan Staunton</i>	
Biochar: Properties, mechanisms, and interactions in the soil	406
<i>Balwant Singh, Shamim Mia, and Marta Camps-Arbestain</i>	
Dissolved organic carbon	421
<i>Jason James and Rob Harrison</i>	
Black carbon, pyrogenic carbon	432
<i>Heike Knicker</i>	
Fire effects on soil	448
<i>Stefan H Doerr, Cristina Santín, and Jorge Mataix-Solera</i>	
Interfacial reactions of microorganisms with minerals and organic matter	458
<i>Rota Wagai, Puu-Tai Yang, and Klaus Kaiser</i>	
Pesticide fate in soils	470
<i>Pierre Benoit, Laure Mamy, Carole Bedos, and Enrique Barriuso</i>	
Soil organic matter in molecular simulations	483
<i>Martin H Gerzabek, Daniel Tunega, Edgar Galicia-Andrés, and Chris Oostenbrink</i>	

Soil Physical Analysis

Raman spectroscopy	497
<i>Anupam Das and Huiyuan Guo</i>	
Fourier transform infrared spectroscopic methods of soil analysis	510
<i>Derek Peak</i>	
Electron paramagnetic resonance spectroscopy: Part I Historical Perspectives	519
<i>Giorgio S Senesi and Nicola Senesi</i>	

Electron paramagnetic resonance spectroscopy: Part II the view forward <i>Robert L Cook, Emmanuel Ampiah, Phillip T Sprunger, and Randall W Hall</i>	533
Fluorescence Spectroscopy: Part I Principles <i>Yong Zhang, Shuai Guo, and Chaoxian Wei</i>	544
Fluorescence spectroscopy: Part II practical applications <i>Yong Zhang, Chaoxian Wei, and Shuai Guo</i>	552
Near and mid infrared soil spectroscopy <i>Wartini Ng, Brendan Malone, Budiman Minasny, and Sangho Jeon</i>	567
Electron microscopic methods (TEM, SEM and energy dispersal spectroscopy) <i>Jian Zhao and Xia Liu</i>	575
Application of atomic force microscopy (AFM) on environmental interfaces <i>Saikat Ghosh, Zhang Di, Bo Pan, and Baoshan Xing</i>	589
Applications of nuclear magnetic resonance for the study of soils <i>Daniel H Lysak, Myrna J Simpson, and Andre J Simpson</i>	605
Neutron-gamma analysis of soil elements <i>Aleksandr Kavetskiy, Galina Yakubova, Nikolay Sargsyan, Stephen A Prior, and H Allen Torbert</i>	625
X-Ray Diffraction <i>Chao Wang and Cheng Gu</i>	642
Soil analysis using laser-induced breakdown spectroscopy <i>Débora Marcondes Bastos Pereira Milori, Ladislau Martin-Neto, Hero T Gollany, Diego Victor de Babos, Luis Carlos Leva Borduchi, and Paulino Ribeiro Villas-Boas</i>	654
Calorimetric analysis <i>Xingmin Rong and Qiaoyun Huang</i>	662
Understanding of soil processes at the microscale—Use of NanoSIMS in soil science <i>Carsten W Mueller, Carmen Hoeschen, and Ingrid Koegel-Knabner</i>	670
Nanoparticles in soil <i>Di Lang, Quan Chen, Yang Liu, and Bo Pan</i>	681
Overview	
Soil chemistry and climate change <i>Jianming Xu, Yong Li, and Zhongmin Dai</i>	692
VOLUME 3: ENVIRONMENT AND MANAGEMENT	
Soil functions, potential uses of soil and ecosystem services	
Soil functions, soil potential uses and ecosystem services <i>Karl Stahr</i>	1
Farming systems to conserve land for nature <i>Rattan Lal</i>	12
Geoarchaeology and soils <i>Carlos E Cordova</i>	20
Civilizations and the role of soil, agriculture, and climate change <i>Jeffrey A Homburg, Eric C Brevik, and Jonathan A Sandor</i>	29
Ecosystem services provided by soils in highly anthropized areas (SUITMAs) <i>Jean Louis Morel, Geoffroy Séré, Viatcheslav Vasenev, and Thomas Nehls</i>	37
Structure and function of forested soils <i>Peter Clinton and Jen Owens</i>	56
Agroforestry <i>NV Thevathasan, AM Gordon, and PKR Nair</i>	68

Rangeland management <i>Alexandra Sandhage-Hofmann</i>	88
Effects on soil of grassland management for pasture, hay and silage <i>Carmen Segura, Laura Cardenas, Richard McDowell, Sarah Morgan, and Martin SA Blackwell</i>	102
Dryland farming: Technological and management options for sustainable agriculture and food systems <i>Ch Srinivasarao, G Ranjith Kumar, Ravula Manasa, Kiran Pilli, Samaresh Sahoo, S Rakesh, Sumanta Kundu, K C Nataraj, KV Rao, JVNS Prasad, and S Malleswari</i>	113
The effect of crop rotations on soil <i>Caio Fernandes Zani, Arlete Simões Barneze, Rogério Peres Soratto, and Charles Andrew Francis</i>	125
The contribution of organic farming systems to soil fertility—A systems perspective <i>Bernhard Freyer, Pierre Ellssel, Jürgen K Friedel, and Kurt Möller</i>	135
Soil functions to support natural plant communities and crops: Soil multifunctionality <i>Leidivan Almeida Frazão, Luiz Arnaldo Fernandes, Adriana Marcela Silva Olaya, and Igor Costa de Freitas</i>	146
Methane and nitrous oxide emissions from land <i>Joanna M Cloy and Keith A Smith</i>	153
Soil and land degradation	
Overview chapter on soil degradation <i>K Müller, Margaret A Oliver, and C Siebe</i>	165
Soil monitoring <i>Ben Marchant and Nicolas Saby</i>	172
Soil quality and health key indicators <i>Edoardo AC Costantini and Simone Priori</i>	181
Water induced soil erosion <i>R Jane Rickson</i>	193
Erosion by wind <i>John Tatarko</i>	208
Human exposure to aeroparticles emitted from unpaved roads due to traffic resuspension and wind erosion <i>Arón Jazcilevich and Ann G Wellens</i>	223
Soil salination processes and management <i>Uri Nachshon and Guy J Levy</i>	236
Degradation of soil structure and soil organic matter <i>Andrew S Gregory</i>	246
Drainage, soil water storage, buffering and filtering <i>Jeffrey Strock and Norm Fausey</i>	257
Environmental effects of irrigation <i>Sevilay Topcu</i>	266
Trade-offs of wastewater irrigation <i>Manzoor Qadir and Christopher A Scott</i>	277
Pollution and risks of trace elements in the soil environment <i>Peng Wang, Fang-Jie Zhao, and Steve P McGrath</i>	288
Pollutants—Persistent organic <i>Rai S Kookana and Divina A Navarro</i>	296
Pollution by microplastic in agricultural areas <i>Nicolas Beriot and Esperanza Huerta-Lwanga</i>	309

Municipal solid waste management: Production, management, and environmental effects <i>Bryant M Serre and Lynda H McCarthy</i>	321
Remediation of metals and organic contaminants in soil <i>Peter Sanderson, Md Mezbaul Bahar, Bhabananda Biswas, and Ravi Naidu</i>	333
Soil change under different scenarios	
Soil change under different scenarios <i>Joann K Whalen</i>	344
The effect of global change on the soil body <i>Sophie Cornu, David Montagne, and Antra Boca</i>	360
The effects of adequate and excessive application of mineral fertilizers on the soil <i>Bijay-Singh and Tek B Sapkota</i>	369
Foliar Application of Nutrients <i>Thomas Eichert and Victoria Fernández</i>	382
Mulches and cover crops part I: Types <i>KK Bandyopadhyay, CL Acharya, and KM Hati</i>	392
Mulches and cover crops part II: Role in soil health and climate resilient agriculture <i>KK Bandyopadhyay, CL Acharya, and KM Hati</i>	401
The potential agronomic and environmental applications of biochar: Prospects and challenges <i>Yu-Jun Wang, Ping Wu, Nanthi S Bolan, and Hai-Long Wang</i>	414
Soil management for carbon sequestration <i>Christopher Brock, Uwe Franko, and Martin Wiesmeier</i>	424
Soil tillage <i>Mathieu Lamandé, Lars Juhl Munkholm, and Trond Børresen</i>	433
Weed and soil management: A balancing act <i>Trevor Kenneth James and Charles Norman Merfield</i>	439
Terraces as traditional agricultural landscapes. Stability amidst change <i>Gerardo Bocco</i>	450
Soil quality protection policies and plans to ensure sustainability <i>Ram Swaroop Meena and Sandeep Kumar</i>	457
Construction of Technosols for green roofs and vertical green system <i>Eleonora Flores-Ramírez, Stefan Abel, Elizabeth Chávez-García, and Geoffroy Séré</i>	473
Management of park areas, sport fields, and school yards, including golf courses and public right of ways <i>Maxine J Levin and Anna Paltseva</i>	488
Soil-society relations	
Overview chapter on soil-society relations <i>C Siebe, Margaret A Oliver, and K Müller</i>	505
Soil art: Sensory and symbolic engagement with soils <i>Alexandra R Toland and Daniel Wolter</i>	509
Soil science in the criminal justice system <i>Lorna Anne Dawson</i>	521
Soil governance at the international, regional and national level <i>Harald Ginzky</i>	532
The economics of soils' contribution to human well-being <i>Bartosz Bartkowski and Julian R Massenber</i>	547

Soil and human health	555
<i>Margaret A Oliver and Eric C Brevik</i>	
Soil Science Education, Awareness and Outreach in the Modern World	572
<i>Laura Bertha Reyes-Sánchez and Eric C Brevik</i>	
Soils in landscape planning	580
<i>Seth Denizen</i>	
VOLUME 4: PEDOLOGY	
Pedology	
Pedology—Basic principles	1
<i>Rupert Bäumler, Uta Stockmann, and Margaret A Oliver</i>	
Factors of soil formation, pedogenic processes and mineralogy	
Overview chapter on factors of soil formation	7
<i>Rupert Bäumler, Margaret A Oliver, and Uta Stockmann</i>	
Factors of soil formation: Climate	14
<i>Stefaan Dondeyne, Stephan Mantel, and Seppe Deckers</i>	
Factors of soil formation: Soil biota and organic matter as a factor for soil formation and weathering processes	25
<i>Scott Werts</i>	
Soil forming factors: Parent material	34
<i>Stephen Nortcliff</i>	
Parent materials: Loess	42
<i>Tobias Sprafke</i>	
Time	48
<i>Susan E Melzer, Caroline M Yonker, and Eugene F Kelly</i>	
Factors of soil formation—Human impacts	54
<i>Jonathan A Sandor, C Lee Burras, Michael Thompson, and Skye A Wills</i>	
Dynamic soils	69
<i>Markus Egli and Dennis Dahms</i>	
Interactions between soils and climate change	82
<i>Cornelia Rumpel</i>	
Primary minerals	91
<i>Balwant Singh</i>	
Mineral components of the soil clay fraction	109
<i>Darrell G Schulze</i>	
Short-range ordered aluminosilicates	121
<i>Tetsuhiro Watanabe, James B Harsh, and Rota Wagai</i>	
Metal oxides	135
<i>Andreas C Scheinost and Balwant Singh</i>	
Soil morphology	149
<i>Phillip R Owens and Zamir Libohova</i>	
Soil micromorphology	171
<i>Rosa M Poch and Georges Stoops</i>	
Soil Classification and distribution	
History and principles of soil classification	185
<i>Maria Gerasimova and Dmitry Konyushkov</i>	

Soil classification systems: World reference base and soil taxonomy	197
<i>David Badía-Villas</i>	
World reference base for soil resources (WRB)	206
<i>Stephan Mantel, Stefaan Dondeyne, and Seppe Deckers</i>	
World soil map based on soil taxonomy	218
<i>Eswaran Padmanabhan and Paul F Reich</i>	
Soils as archives of city history	232
<i>Hermine Huot and Jean Louis Morel</i>	
Classification of soils by ecozones	
Ecozones and soils – An introduction	246
<i>Rupert Bäumler</i>	
Cold Region Soils: Part I—Distribution, Properties and Processes	250
<i>Nicolas A Jelinski, Chien-Lu Ping, and John CF Tedrow</i>	
Cold region soils: Part II – Genesis and classification	263
<i>Nicolas A Jelinski, Chien-Lu Ping, and John CF Tedrow</i>	
Boreal forest soils	280
<i>Gabriele Broll</i>	
Soils of humid cool temperate regions	289
<i>Nicolas A Jelinski, Justin B Richardson, and Edward A Nater</i>	
Grassland soils in the cool–arid–temperate ecozone (Steppe, grassland)	299
<i>Mengyang You, Xiaobing Liu, Rupert Bäumler, and Margaret A Oliver</i>	
Soils from a warm-temperate semi-tropical ecozone (Mediterranean) with humid winter months	319
<i>J Torrent, Marco Pfeiffer, and Juan José Ibañez</i>	
Arid and semi-arid subtropical and tropical ecozone (II)	332
<i>Bernhard Lucke</i>	
Humid subtropical ecozone with emphasis on Acrisols and Alisols	343
<i>Nilton Curi, Sérgio Henrique Godinho Silva, and Alberto Vasconcellos Inda</i>	
Soils of the humid and sub-humid tropics	353
<i>Karl Stahr and Ludger Herrmann</i>	
Classification of soils by properties and environmental settings	
Sodic soils	368
<i>Guy J Levy and Uri Nachshon</i>	
Organic soils	383
<i>Heike Schimmel and Wulf Amelung</i>	
The pedology of anaerobic soils	398
<i>John Hollis and Allan Lilly</i>	
Alluvium and alluvial soils	417
<i>Janis L Boettinger</i>	
Hydric Soils	424
<i>MJ Vepraskas and JF Berkowitz</i>	
Wetland soils	436
<i>Cornelius Wilhelm van Huyssteen</i>	
Paddy soils: Characteristics, productivity and management	447
<i>Stephan M Haefele</i>	
Volcanic soils and non-volcanic andic soils	458
<i>Rupert Bäumler and G Uehara</i>	

PEDOMETRICS**Pedometrics**

Overview of Pedometrics	471
<i>Alexandre MJ-C Wadoux, Inakwu OA Odeh, and Alex B McBratney</i>	

Statistics and soil data analysis

Statistics in soil science	486
<i>Richard Webster</i>	

Multivariate analysis for soil science	499
<i>Si Yang Han, Patrick Filippi, Mercedes Román Dobarco, Januar Harianto, Mathew S Crowther, and Thomas FA Bishop</i>	

Machine learning and processing of large data	509
<i>Madlene Nussbaum</i>	

The issue of scale and change of support in the spatial analysis of environmental data	521
<i>A Castrignanò, A Belmonte, and N Romano</i>	

Soil measurements in space and time

Sampling for soil survey	535
<i>DJ Brus</i>	

Handling of soil samples in the field, in transport to the laboratory and in the laboratory	551
<i>Claudy Jolivet, Céline Ratié, Déborah Chavrit, Nicolas Soler-Dominguez, and Antonio Bispo</i>	

Digital Soil Morphometrics	568
<i>Yakun Zhang, Alfred E Hartemink, Jingyi Huang, and Budiman Minasny</i>	

Proximal soil sensing in the field	579
<i>Mikaela Tilse, Uta Stockmann, and Patrick Filippi</i>	

Proximal sensing of soil moisture	591
<i>Markus Tuller, Ebrahim Babaeian, and Scott B Jones</i>	

Proximal sensing of land surface temperature	600
<i>Ebrahim Babaeian and Markus Tuller</i>	

Proximal sensing of evapotranspiration	610
<i>Ebrahim Babaeian and Markus Tuller</i>	

Remote sensing of soil moisture	618
<i>Alexander Gruber and Jian Peng</i>	

Remote Sensing of Soil Organic Carbon	631
<i>Bas van Wesemael and Sabine Chabrilat</i>	

Pedotransfer functions and their application to soil water dynamics	642
<i>Marcel G Schaap, Yonggen Zhang, and Attila Nemes</i>	

Spatial and temporal soil variation

Geostatistics: Modelling spatial variation	655
<i>Richard Webster</i>	

Uncertainty assessment of spatial soil information	671
<i>Gerard BM Heuvelink and Richard Webster</i>	

Digital soil mapping: Evolution, current state and future directions of the science	684
<i>Brendan Malone, Dominique Arrouays, Laura Poggio, Budiman Minasny, and Alex B McBratney</i>	

Geographical information systems (GIS) and soils	696
<i>S Grunwald and J Böhner</i>	

Precision agriculture	710
<i>James A Taylor</i>	

Foundations, measurements and trends in pedodiversity <i>Juan José Ibañez and Marco Pfeiffer</i>	726
Soil landscape modelling	
Modeling soil development in a landscape context <i>Peter A Finke, Budiman Minasny, and Arnaud JAM Temme</i>	739
Climate and land surface models: Role of soil <i>Toby Richard Marthews, Holger Lange, Alberto Martínez-de la Torre, Richard J Ellis, Sarah E Chadburn, and Martin G De Kauwe</i>	751
VOLUME 5: SOIL PHYSICS	
Soil structure & mechanics	
Soil structure <i>Steffen Schlüter and John Koestel</i>	1
Aggregation <i>Teamrat A Ghezzehei</i>	8
Porosity and pore-size distribution <i>John R Nimmo, Sheela Katuwal, and Maik Lucas</i>	16
Pore-scale modelling of flow and transport phenomena in soils <i>Kirill M Gerke and Marina V Karsanina</i>	25
Soil texture: Its relevance, measurement and representation internationally <i>Attila Nemes</i>	35
Soil tilth <i>Thomas Keller, Douglas L Karlen, and Paul D Hallett</i>	48
Crusts and seals: Structural <i>Shmuel Assouline and Sally Thompson</i>	57
Soil mechanics <i>Markus Berli and Paul D Hallett</i>	69
Compaction <i>JJH van den Akker, L ten Damme, M Lamandé, and Thomas Keller</i>	85
Bioturbation—Physical processes <i>Siul Ruiz, Paul D Hallett, and Dani Or</i>	100
The reinforcement of soil by plant roots <i>Alexia Stokes and Anil Yildiz</i>	115
Physical effects of soil salinity <i>David Russo</i>	123
Shrink-Swell processes in soils <i>Pascal Boivin</i>	139
Water retention	
Water properties <i>N Lu, C Zhang, and D Hillel</i>	146
Water potential <i>Dani Or, Markus Tuller, and Jon M Wraith</i>	159
Water content and water potential measurement <i>Wolfgang Durner</i>	168
Soil water retention and characteristic curve <i>Markus Tuller and Dani Or</i>	187
Soil water repellency <i>Jörg Bachmann and Marc-Oliver Goebel</i>	203

Flow & transport processes

Aeration	216
<i>Rodney T Venterea and Dennis E Rolston</i>	
Capillarity	225
<i>Dani Or and Markus Tuller</i>	
Darcy's law	240
<i>Conrad Jackisch and Eva Kröner</i>	
Diffusion	247
<i>TKK Chamindu Deepagoda and Shoichiro Hamamoto</i>	
Hydrodynamics in soil	256
<i>Ty Ferre and Art Warrick</i>	
Thermodynamics of soil water	262
<i>Pieter H Groenevelt</i>	
Immiscible Fluids	271
<i>Yukie Tanino and Dubravka Pokrajac</i>	
Overland flow	287
<i>Tammo S Steenhuis, Rajith Mukundan, and Seifu A Tilahun</i>	
Direction-dependency of soil hydraulic and mechanical properties	294
<i>Steffen Beck-Broichsitter, Heiner Fleige, Jose Dörner, and Rainer Horn</i>	
Macropores and macropore flow	306
<i>Nicholas Jarvis and Mats Larsbo</i>	
Unstable preferential flow in porous media	314
<i>Naaran Brindt, J-Yves Parlange, and Tammo S Steenhuis</i>	
Leaching processes	321
<i>Brent Clothier and Steve Green</i>	
Colloid transport	336
<i>Yan Jin, Chongyang Shen, and Volha Lazouskaya</i>	
Scaling of transport properties	353
<i>Robert Ewing and Behzad Ghanbarian</i>	
Soil influence on groundwater	364
<i>David Rudolph and Edwin Cey</i>	

Thermal properties and behaviour

Temperature effects of soil hydraulic properties	379
<i>Jan W Hopmans and Robert Horton</i>	
Heat and vapor transfer	385
<i>Sunny Eng Giap Goh, Joshua L Heitman, and Robert Horton</i>	
Radiation Balance	393
<i>Jerry L Hatfield, Kaitlin Togliatti, Thomas J Sauer, and John H Prueger</i>	
Freezing and thawing cycles	400
<i>Alan W Rempel</i>	
Arctic permafrost	410
<i>Julia Boike, Andrey Abramov, Katrina E Bennett, and Lars Kutzbach</i>	
Permafrost and cold region hydrology	419
<i>Masaki Hayashi and Ylva Sjöberg</i>	

Sensing Techniques

Soil water sensing by neutron scattering	427
<i>Steven R Evett</i>	

Time-domain reflectometry	436
<i>GC Topp and Ty Ferre</i>	
Geophysical methods for soil applications	444
<i>Sarah Garré, Guillaume Blanchy, David Caterina, Philippe De Smedt, Alejandro Romero-Ruiz, and Nataline Simon</i>	
Non-invasive imaging of soil processes	459
<i>Craig J Sturrock</i>	
Hydrology & water balance	
Soil processes in the hydrologic cycle	469
<i>Cosimo Brogi, Harry Vereecken, Heye Reemt Bogena, and Luca Brocca</i>	
Global water cycle from a soil perspective	482
<i>Michael H Young and Dani Or</i>	
Water evaporation from bare soil	490
<i>Peter Lehmann</i>	
Evapotranspiration	501
<i>Yuting Yang</i>	
Plant available water	509
<i>Quirijn de Jong van Lier, Sally D Logsdon, Everton Alves Rodrigues Pinheiro, and Paulo Ivonir Gubiani</i>	
Plant-water relations	516
<i>Ian C Dodd, Andrew D Hirons, and Jaime Puértolas</i>	
Water use efficiency: A review of spatial and temporal variability	527
<i>Simone Fatichi, Athanasios Paschalis, Sara Bonetti, Gabriele Manoli, and Christoforos Pappas</i>	
Root water uptake and myco-rhizosphere hydraulic properties	543
<i>Andrea Carminati, Maria Marin, Valentin Couvreur, Mohsen Zarebanadkouki, Stéphane Declerck, and Mathieu Javaux</i>	
Plant-soil-water relations	552
<i>Jan Vanderborght, Quirijn de Jong van Lier, Andrea Schnepf, and Mathieu Javaux</i>	
Crop and landscape water requirements	571
<i>Luis S Pereira, Isabel Alves, and Paula Paredes</i>	
Water harvesting	593
<i>Wolde Mekuria and Desalegn Tegegne</i>	
Irrigation methods	608
<i>Shmulik P Friedman</i>	
Soil–plant–atmosphere continuum	624
<i>John M Norman, Martha C Anderson, and William P Kustas</i>	
The Penman-Monteith equation	634
<i>Richard G Allen, Ayse Kilic, Guido D'Urso, and Simona Consoli</i>	
Infiltration	646
<i>Ty Ferre and Art Warrick</i>	
Solute transport	652
<i>MC Sukop and M Dentz</i>	
Lysimetry	667
<i>Thomas Pütz and Jannis Groh</i>	
Recognition of notable past soil scientists	680
<i>Michael J Goss, Evgenia Blagodatskaya, Ariena HC van Bruggen, Monika Joschko, Stefan Schrader, and Christoph Tebbe</i>	
<i>Index</i>	691

SOIL BIOLOGY

Edaphology

RM Goss and AL Ulery, Plant & Environmental Sciences Department, New Mexico State University, Las Cruces, NM, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	2
Edaphology research	3
Sustainable agriculture and organic production	3
Edaphic properties	4
Edaphic factors affecting water and air	5
Edaphic factors affecting plant nutrients	7
Edaphic factors affecting root and microbial growth	8
Conclusion	9
References	9

Abstract

Edaphology is the study of soil and soil organism relationships, and how these relationships affect plant growth. The study of edaphic factors is important in understanding the Earth's Critical Zone, the outer skin of our planet where rock, soil, water, air, and organisms interact to regulate the environment and support life. Edaphic factors are soil conditions and properties that affect the fauna, flora and microbial organisms living in the root zone, rhizosphere, or landscape. Such factors are numerous but the most important include water content, aeration, nutrient levels, pH, and soil temperature.

Glossary

Agroecology The science of applying ecological concepts and principles to the design and management of sustainable agroecosystems. See also: sustainable agriculture.

Base-cation saturation ratio (BCSR) A method of interpreting soil tests based on the relative ratio balance of the base cations (calcium, magnesium, potassium and sodium) that would be found in an ideal soil. This somewhat controversial method is used in some sustainable agriculture and some sectors of urban horticulture.

Ca²⁺ Calcium ion.

CEC Cation exchange capacity.

Climax The most advanced successional community of plants capable of development under, and in dynamic equilibrium with, the prevailing environment.

Earth's Critical Zone The near-surface, life-sustaining environment that includes living organisms, rock, soil, water, and air.

EC Electrical conductivity, related to the concentration of salts in solution or soil salinity.

Ecology The science of the relationship between organisms and their environment.

Edaphology The science that deals with the influence of soils on living things; particularly plants, including human use of the land for plant growth. The term may also refer to man's use of soil for plant growth or the overall human use of the land.

Gleyed/gleying A condition resulting from prolonged soil saturation and reducing conditions, indicated by the presence of bluish or greenish colors from ferrous iron in the soil mass or as mottles.

Humus Organic soil compounds exclusive of undecayed plant and animal tissues, their 'partial decomposition' products, and the soil biomass. The term is often used synonymously with soil organic matter.

K⁺ Potassium ion.

Macronutrient A plant nutrient found at relatively high concentrations (greater than 500 mg kg⁻¹) in plants; usually refers to nitrogen, phosphorus, and potassium, but may also include calcium, magnesium, and sulfur.

Mg²⁺ Magnesium ion.

mg kg⁻¹ Milligrams per kilogram of dry soil or plant material. Similar to parts per million (ppm).

Micronutrient A plant nutrient found in relatively small amounts (less than 100 mg kg⁻¹) in plants. These include boron, chloride, copper, iron, manganese, molybdenum, nickel, cobalt, and zinc.

mm Millimeters or 1 × 10⁻³ (0.001) meters. Used to define the length or diameter of soil particles, flora and fauna including plant roots, soil organisms, etc.

Mottles Spots or blotches of different color or shades of color interspersed with the dominant soil color.

NH₄⁺ Ammonium ion.

Pedology The science that deals with study of soils in their natural environment. The science generally includes soil formation (pedogenesis), soil morphology, and soil classification.

Physiography The study of the earth's natural features including land formation, climate, currents and distribution of life.

Rhizosphere The region of the soil that surrounds plant roots. This zone of soil serves as the interface of plant roots and the surrounding soil matrix. Microbial populations, plant roots, root secretions, soil water, soil air, and soil nutrients are dynamically linked in this region of the soil.

Root zone The zone of the soil profile penetrated by plant roots. In natural environments, the root zone is comprised of the upper portions of the native soil profile (mostly A and B soil horizons). In urban environments and other human-affected locations, the root zones are created from non-native soils or loose materials such as sand, clay and organic matter to meet specific construct criterion.

Sodium adsorption ratio (SAR) A relation between soluble sodium and soluble divalent cations that can be used to predict the exchangeable sodium fraction of soil equilibrated with a given solution. It is defined as the concentration of sodium divided by the square root of the sum of calcium and magnesium concentrations.

Soil salinity The salt content of the soil. Soil salinity is generally increased by the accumulation of salts from mineral weathering, water evaporation and/or application of saline irrigation water.

Soil quality/soil health The ability of a soil to support plant growth and development for human purposes based on the physical, chemical, and microbial characteristics of a soil. Improving and sustaining soil quality and soil health is a major principle of sustainable agriculture and organic production. In addition, soil organic management, crop rotations, mixed plant communities and cover crops are used to maintain a diverse soil microbial community, thus maintaining soil quality and soil health.

Sufficiency nutrient levels Traditional method for analyzing soil tests by determining the available soil nutrients that are found in sufficient levels for plant growth. Nutrients found in excess, or detrimental levels, are also addressed in this analysis technique. Also named Liebig's Law of the Minimum.

Sustainable agriculture The use of ecological principles to manage soil quality and soil health while cultivating plants for food, fiber and forage to sustain human life.

Worms m⁻² The number of earthworms per square meter of soil.

Key points

- Edaphology is the study of soil in relation to plant growth.
- Edaphic factors include the physical, chemical, and biological conditions of soil.
- Supporting flora and fauna is one of the major functions of soil.

Introduction

Edaphology is the study of soil and soil organism relationships, and how these relationships affect plant growth. The [Soil Science Society of America \(2008\)](#) defines edaphology as “the science that deals with the influence of soils on living things; particularly plants, including man's use of land for plant growth.” The root of the word is ‘édaphos,’ Greek for ‘foundation,’ ‘soil,’ ‘ground,’ or ‘land.’ ‘Soil science’ is the term more commonly used today and includes the study of soil as a natural body in the landscape (pedology), as well as a medium to be managed for optimum crop and rangeland productivity, environmental waste disposal, and construction. Cooperatively, pedology focuses on the description, classification, and formation of soil in place ([Richter, 2019](#)). Edaphology concentrates specifically on soil as a medium for fauna (animal), flora (plant) and microbial growth, or generally the human use and management of soil to grow plants.

The study of edaphic factors is important in understanding the Earth's Critical Zone, the outer skin of our planet where rock, soil, water, air, and organisms interact to regulate the environment and support life ([Committee on Basic Research Opportunities in the Earth Sciences, 2001](#)). Edaphic factors have been described as those soil conditions and properties that affect the fauna, flora and micro-organisms living in a particular root zone, rhizosphere, or even landscape. Such factors are numerous but the most important include water content, aeration status, nutrient levels, pH, and soil temperature ([Lindbo et al., 2012](#); [Weil and Brady, 2017](#)). Climatic or physiographic factors also influence living organisms, but these factors have a larger scale and tend to affect aboveground activities more than the below ground soil environment.

Ecosystems are thrown out of balance when major disturbances occur such as overgrazing of rangeland, or the conversion of native forests or grassland to agriculture (Brantley et al., 2006). Newly disturbed ecosystems cannot be adequately described by traditional theories of plant succession or climax models. The maintenance of agricultural land by humans prevents natural plant succession. An edaphic climax is a localized vegetative community that differs from the surrounding vegetation because of different soil types and properties, including water-holding capacity, drainage class, soil depth, and soil fertility. Nonequilibrium ecology and vegetation dynamics describe the interactions between the soil resource and the associated vegetative community. In these models, the condition of the soil is directly connected to aboveground vegetation. The organisms that live in a soil influence soil formation and are in turn affected by the edaphic properties that govern water, air, and nutrient supply. Thus, a continual feedback relationship is established between the soil flora and fauna, and the physical and chemical characteristics of the soil.

Soil is dynamic, teeming with organisms, and is an integral, interactive part of the environment. Soil is the foundation upon which our civilization stands, and it must be used wisely and conserved if we are to continue benefiting from its numerous functions, including food and fiber production, environmental protection, water and nutrient storage, and engineering roles.

Edaphology research

Studies conducted by ecologists, plant scientists, agronomists, and soil scientists (or edaphologists) are designed to measure, characterize, monitor, and assess the soil status as it serves as a medium for plant growth. Edaphology research is largely divided into ecologic and agronomic studies.

In the science of ecology, the spatial variation or distribution of vegetation as a function of soil type or edaphic property is important to characterize an ecosystem (e.g., Luo et al., 2021). The information gleaned from these studies provides a baseline of native vegetation in the landscape and helps identify changes due to anthropogenic and climatic effects. The long-term goal of this kind of study is to manage or protect native or desired vegetation by understanding the relationships among plant communities, the underlying soil, disturbance regimes, and physiographic features. Ecological studies that correlate plants with specific soil types, properties, or microbial activity in addition to landscape position and climate information are other examples of edaphology. Evaluating or predicting the effect of management on soil–plant interactions and edaphic factors is often the objective of studies on ecosystem health, soil status, and land degradation.

The second field of edaphology research focuses on agronomic management to maximize crop yields. Irrigation and soil-fertility management are examples of this approach. Agronomic edaphology is the study of agricultural foundational principles where long-term cropping practices are used to sustain soil fertility and productivity. Intensive conventional agriculture can deplete soil organic matter and beneficial microorganisms, thereby deteriorating soil status, reducing soil ecosystem functions and degrading the soil. Employing conservation practices may slow, halt or reverse negative impacts of intensive conventional agricultural management practices. By monitoring indicators of soil status and various edaphic factors, scientists can assess the usefulness and sustainability of adopted management practices on agronomic ecosystems (or agroecosystems) (USDA, 2020). The two broad areas of ecological and agricultural edaphology have also been combined to allow comparison of intensively managed agroecosystems with virgin soils and native vegetation.

Sustainable agriculture and organic production

The growing focus on sustainable agriculture and organic farming has heightened the awareness of edaphology and edaphic factors. The FAO definition of sustainable agricultural development is “the management and conservation of the natural resource base, and the orientation of technological and institutional change in such a manner as to ensure the attainment and continued satisfaction of human needs for present and future generations. Such development... conserves land, water, plant and animal genetic resources, is environmentally non-degrading, technically appropriate, economically viable and socially acceptable.” (<https://www.fao.org/3/u8480e/U8480E01.htm>). In the United States, sustainable agriculture is defined by U.S. Code Title 7, Section 3103, as amended in 2011, as “an integrated system of plant and animal production practices having a site-specific application that will, over the long-term – (a) satisfy human food and fiber needs; (b) enhance environmental quality and the natural resource base upon which the agriculture economy depends; (c) make the most efficient use of nonrenewable resources and on-farm resources and integrate, where appropriate, natural biological cycles and controls; (d) sustain the economic viability of farm operations; and (e) enhance the quality of life for farmers and society as a whole” ECFR (Electronic Code of Federal Regulations), 2011. This legal definition of sustainable agriculture differs from organic farming and various other soil management philosophies including ecological, regenerative, biodynamic, community-based, environmentally sensitive, extensive, low-input, permaculture, wise-use agriculture. Each of these terms has a slightly different perspective, but they all have similar goals, which is to ensure that plant yields are maximized without harming the soil. Successful management practices are intended to further improve soil status with time.

Sustainable agriculture focuses not only on the sustainability of the means of production, including the land and other resources, but also the economic sustainability of agricultural system. On the other hand, organic production identifies specific, restricted agricultural practices that cannot be used for the agricultural products to be classified as organic. To be considered an organically

produced agricultural product in the United States, they must not have been grown with synthetic- and petroleum-based products or other prohibited practices during crop production and handling. U.S. Code Title 7.B.I.M.205.105, as amended in 2021 (<https://www.ecfr.gov/current/title-7/subtitle-B/chapter-I/subchapter-M/part-205/subpart-B/section-205.105>), requires the product to “have been produced and handled without the a) use of synthetic chemicals and ingredients, . . .; b) nonsynthetic substances prohibited . . .; c) nonagricultural substances used in or on processed products . . .; d) nonorganic agriculture substances used in or on processed products . . .; e) excluded methods, except for vaccines: *Provided*, That the vaccines are approved. . .; f) ionizing radiation, as described in Food and Drug administration regulation . . .; and g) sewage sludge.” Although the rules for organic production in the European Union are lengthier and more specific, they follow the spirit of the US guidelines (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2018.150.01.0001.01.ENG&toc=OJ:L:2018:150:TOC). Regardless of the depth and scope of the regulation policies, sustainable agriculture, organic production and other soil management philosophies use edaphological practices to ensure the soil will provide long-term benefit to the farmers and society. Balancing the soil physical, chemical, and biological properties for the benefit of the crop increases the plant’s ability to produce and the land’s ability to support humanity.

The production and consumption of organic produce has increased since the 1990’s, resulting in the organic market becoming one of the fastest growing sectors of agriculture. The overall U.S. agricultural market share for organic fruit and vegetables was 15% in 2019 (Willer et al., 2021). Increased acceptance and consumption of organic produce will likely continue to grow. In addition to the elimination of petrochemicals and synthetic fertilizers from their management practices, organic farmers place an emphasis of “soil health” or “soil quality” to increase yields of produce. However due to the restrictions on synthetic pesticides, increased pest management efforts and additional human labor are often required to produce desirable yields. Soil health and soil quality in this context refer to the status or ability of a soil to support plant growth and development for human purposes based on the beneficial integration of soil physical, chemical, and microbial characteristics (Doran et al., 1994).

Edaphic properties

Edaphologists study the soil by examining a variety of factors or soil properties that govern life in the soil. Edaphic factors include those physical, chemical, and biological properties of a soil that influence plant growth. The various properties that are important for plant growth are interrelated but can be distilled to a few simple requirements. Essentially, what plants need from the soil are water, air, nutrients, and physical support. Sunlight is required for photosynthesis but is provided aboveground and may be affected by aspect, latitude, and other physiographic and climatic factors.

Edaphic factors affect the ability of soil to sustain biological production and diversity, regulate and partition water, filter and buffer contaminants, store and cycle nutrients, and provide plant support. Generally, soils of better status will provide a superior growing environment for plants, resulting in plants of greater quality with increased productivity. Also, beneficial microbial soil communities will flourish in such soils. When soils function well in supporting plant growth, they are considered to be fertile and productive but also resist erosion, and help maintain good air and water quality. Soils that are resilient to natural and anthropogenic stresses are therefore resistant to degradation. Farmers, homeowners, turfgrass managers and other managers of the land can use a variety of management practices to produce and sustain soils. These best management practices result from edaphic research and include supplementation of root zones with composts, manures, and other organic amendments; utilization of cover crops and crop rotations; application of integrated pest management practices, including biological pest control methods. Soil quality or health indicators are those physical, chemical, and biological properties that measure a soil attribute or function and are sensitive to changes in environment and management (Tables 1–3) (Doran et al., 1994).

Table 1 Some physical edaphic properties and soil status indicators that affect plant growth by influencing water and air movement, plant rooting and soil microbial activity.

<i>Soil property</i>	<i>Water and air movement</i>	<i>Plant rooting</i>	<i>Microbial activity</i>
Texture (sand, silt, and clay)	X	X	X
Structure (especially porosity and macropore arrangement)	X	X	X
Bulk density (dry mass/bulk volume)	X	X	X
Soil depth or distance to impermeable layer	X	X	
Drainage class	X	X	X
Water-holding capacity (related to texture and structure)	X		X
Infiltration (water movement into the soil surface)	X		X
Hydraulic conductivity (water movement through the soil)	X		X
Soil-water potential (or availability of water)	X		X

Table 2 Some chemical edaphic properties and soil status indicators that affect plant growth by influencing water and air movement, nutrients, or rooting depth in soil.

<i>Chemical property</i>	<i>Water and air movement</i>	<i>Nutrient availability</i>	<i>Rooting</i>	<i>Microbial activity</i>
pH		X		X
Cation exchange capacity		X		X
Organic matter content	X	X	X	X
Nitrate nitrogen		X		X
Organic nitrogen		X		X
Total or soluble phosphorus		X		X
Total or soluble potassium		X		X
Micronutrients		X		X
Clay mineralogy	X	X	X	
Salinity or EC	X	X		X
Toxic ions or heavy metals		X		X

Table 3 Some biological edaphic properties or soil status indicators that affect plant growth by influencing water and air movement, nutrients, or rooting depth in soil.

<i>Biological property</i>	<i>Water and air movement</i>	<i>Nutrient availability</i>	<i>Root support or impedance</i>
Earthworm number	X	X	X
Respiration rate (measure of microbial population and activity)	X	X	
Microbial diversity	X	X	
Composition of organic matter	X	X	X
Plant root depth	X	X	X

Edaphic factors affecting water and air

Soil physical properties have the greatest effect on water movement into and through the soil. The role of water is so important in plant growth that some definitions refer to edaphology as the science of physics applied to soil and water. Water and air are considered together because they use the same pathways to enter and move through the soil (Hillel, 1998). The water-holding capacity of a soil refers to the amount of water contained in a soil and is determined largely by texture, structure, porosity, organic matter content, and depth (Weil and Brady, 2017).

Soil texture refers to the particle-size distribution of the inorganic soil material which ranges over at least three orders of magnitude from sand (2–0.05 mm) to clay (<2 μm) sizes. Soil texture and structure (the arrangement of particles) both have a direct impact on soil porosity, a property that reflects the number and size of pores in the soil (Weil and Brady, 2017; Lindbo et al., 2012). In general, finer-textured clayey soils and soils with larger organic matter content will have greater total porosity than coarser-textured sandy soils or soils with less organic matter. Even though clayey soils have greater porosity and hold more total water than sandy soils, the water is held so tightly in the micropores that more of it is not available to the plants. The micropores prevalent in clayey soils store water in the root zone for a longer time but greatly resist its movement through the soil because of their small cross sectional area and interconnectivity. In addition the water is subject to a higher matric potential acting (the attractive force of a solid for water). Sandy soils, on the other hand, have large, interconnected macropores that fill and drain easily. Macropores provide for water and air movement through the soil, but do not hold water very long in the root zone (Fig. 1). While the texture refers specifically to the distribution of particles less than 2 mm in diameter, also known as the fine earth fraction; another valuable piece of information is the percentage of gravel or amount of particles greater than 2 mm. The larger particles affect soil packing, porosity, bulk density, nutrient status, and root growth.

Porosity results from the arrangement of soil particles into aggregates with inter- and intra-aggregate pores, as well as from root growth and faunal activity. Macropores can form after roots decay or as a result of burrowing by soil fauna. One of the most dominant groups of soil animals are earthworms, which may number from approximately 50 to 300 worms m^{-2} in agricultural soils up to 500 worms m^{-2} in forest and grassland soils. Earthworms shred and consume organic matter and excrete wastes as casts, which help create stable soil aggregates and enhance soil porosity (USDA, 2020). Their burrows are persistent over time and provide a major conduit for drainage and preferential water movement through the soil, but also minimize surface water erosion by increasing infiltration. In agricultural soils, the presence and abundance of earthworms often indicate, productive, organic-rich soils (Stroud, 2019).

Bulk density is the dry soil weight per volume of soil and is a good measure of porosity and compaction. Generally, uncompacted soils containing 50% pore space with a bulk density of 1.1–1.6 g cm^{-3} are ideal for agricultural soils. Soil minerals have an average particle density of $\sim 2.65 \text{ g cm}^{-3}$. Pore space is included in the volume of a soil but the air content does not weigh

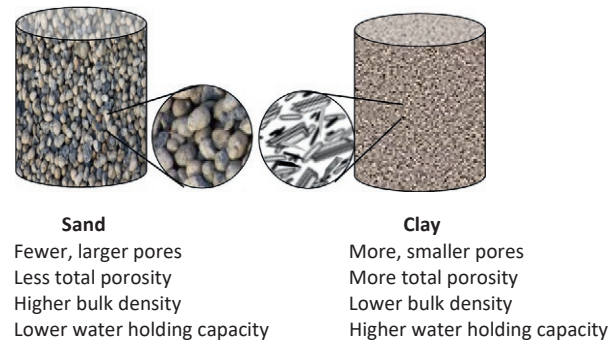


Fig. 1 The size and amount of pores in sand and clay results in different bulk densities and water holding capacities.

much, thus as porosity increases, the dry weight decreases for a given volume, resulting in a lower bulk density. Compacted soils have large bulk densities ($1.5\text{--}1.8\text{ g cm}^{-3}$), few macropores, and tend to restrict root growth, water and air flow. Texture and organic matter both affect bulk density as well. Sandy soils have a larger bulk density than clayey soils because of smaller total porosity. Organic matter weighs very little, takes up a large volume, and so creates lower bulk density. Some minerals have a large particle density (mass per particle volume) and can contribute to a greater soil bulk density if they are dominant in the soil. Soils with a smaller bulk density contain more plant available water than soils with a large bulk density, but they tend to drain quickly and become droughty (Fig. 1).

Organic matter contributes to water-holding capacity in several ways. It aggregates soil and produces a crumb-like structure at the surface that facilitates infiltration, percolation, storage, and diffusion of water and air. Sticky substances, such as waxes and tannins left behind when soil organisms decompose organic residues, help bind soil particles together into porous clusters that hold and release water, nutrients, and humus. Soil organic matter or humus holds as much as 20 times its weight in water and has many beneficial physical, chemical, and biological properties (Magdoff and Van Es, 2021). The O and A horizons in soil profiles have the largest amounts of organic matter and are often measured in field studies to evaluate their influence on nutrient supply or relationship to vegetation. Organic carbon or total organic matter is also commonly measured in edaphic studies. Too much organic matter can cause other problems due to the waxy coating left on coarse textured soils, such as those 80–100% sand golf course putting greens, that results in hydrophobic soils with decreased infiltration rates despite the dry soil conditions. Naturally occurring water-repellent soils are found in other parts of the world where sandy soil particles become coated in plant waxes or other plant degradation products (e.g., see <https://www.csiro.au/en/research/plants/crops/farming-systems/water-repellent-soils>). Once dried, organic matter is often hydrophobic and may repel water.

The depth of the soil and the water-holding capacity also directly determine the total amount of water present. Many trees and desert shrubs have root systems that exploit soil water and nutrients to depths below 2 m, but most agronomic plant roots are concentrated in the upper meter of the soil. Urban root zones are often shallower. Water movement through a root zone is hampered by any changes to soil physical properties including changes in texture, compaction, porosity or organic matter content. The bottom of a soil profile may be an impermeable layer or hardpan, bedrock, or a shallow groundwater table. Shallow, eroded soils hold less water than deep, uniformly textured soils. Gleying and mottles are evidence of water-saturated soil and poor aeration or anaerobic conditions prevailing for extended periods of time.

Mineralogy, specifically the kind and amount of certain clay minerals, may also affect water and air movement through the soil. Some well weathered soil minerals such as montmorillonite, a member of the smectite family, absorb water and swell to fill up cracks and macropores, resulting in restricted hydraulic conductivity. This is particularly problematic in sodic soils or alkali-affected areas. Sodium (Na) is a cation with low charge density and therefore has a greater hydrated radius than higher valency ions, that tends to neutralize and disperse negatively charged clay particles and the aggregates formed from clay and organic matter. When aggregates disperse, they form an impermeable seal at the soil surface and along cracks that limits the movement of water and air into the soil (Fig. 2). Other problems with salinity are discussed in a later section.

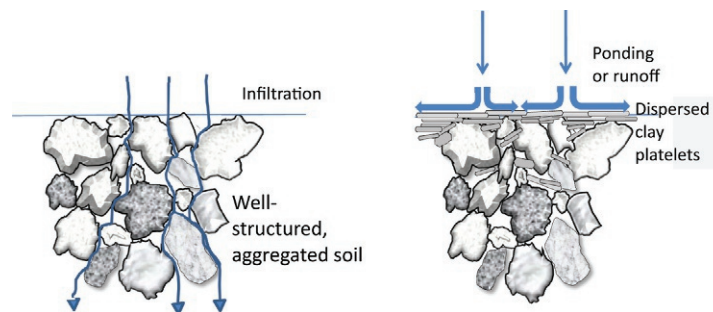


Fig. 2 Infiltration and percolation is enhanced by well-structured soil aggregates while dispersed clay particles in sodic soils may cause ponding.

Edaphic factors affecting plant nutrients

For optimum growth, plants require many inorganic nutrients in varying amounts and forms. Macronutrients have concentrations of at least 500 mg kg⁻¹ in plants, while micronutrients are required in much smaller amounts, usually less than 100 mg kg⁻¹. Nitrogen (N), phosphorus (P), and potassium (K) are the most commonly measured and applied macronutrients by farmers, gardeners, and other land managers. Macronutrients are not necessarily the most important nutrient in determining the ability of that plant to grow; macro- refers to the large quantity used by the plant. Liebig's Law of the Minimum (or sufficiency levels) describes the fertilizer needs of plants by stating that plant growth (yield) is controlled by the most limited (needed) nutrient. In other words, an application of the most limited nutrient in the form needed by the plant will result in additional plant growth. The next most limited nutrient will then control the rate of growth of the plant. The exact requirements of each nutrient vary for each plant and environment. Recommended ranges are determined through edaphology research and experience.

Fertilizers are labeled with their N-P-K amounts and should be applied to the soil only when soil or tissue test results indicate that the nutrients are below optimum levels. Fertilizer applications for nutrients that are not limiting will not result in additional growth and may leave the plant root zone by leaching or runoff impacting offsite environments. Due to the rapid fluxes of soil nitrogen content, on-site tissue tests for nitrogen are often more accurate than soil tests. Specific edaphic factors that may be measured to monitor soil fertility include nitrate-N, available P, and soluble K (Table 2). Other commonly measured nutrients include calcium (Ca), magnesium (Mg), sulfate (SO₄²⁻), and iron (Fe). Recent debates in the scientific community have considered not just the total amount of each nutrient affecting plant growth and plant landscape distribution, but also the relative proportions of nutrients. Base-cation saturation ratio (BCSR) has been considered an important tool for determining soil health in sustainable agriculture and some sectors of urban horticulture. For example, the calcium-to-magnesium ratio may be used to explain the occurrence of some endemic plants growing on serpentine soils in California, USA. However, other research has found higher plant health results when the minimal nutrient levels for sustainable growth are used in fertilizer programming (Woods et al., 2016).

The soil reaction or pH is related to nutrient availability and optimum plant growth. Soil pH is one of the most common and important soil chemical properties measured. Some plants prefer acidic soils, while others are best suited for alkaline soils. There are few plants that will survive in extremely acid or alkaline soils however, and most flora and fauna prefer conditions between pH 5.5 and 8.5. For most plants, having soil pH close to 6.5 is ideal as all essential plant nutrients are very available. Some pH-based nutrient problems can be avoided with the use of foliar-absorbed nutrients. Soil pH also affects the soil microbial population necessary to mineralize soil organic matter to inorganic nutrient forms.

The combined physical, chemical, and biological benefits of organic matter make it one of the most important edaphic factors affecting plant growth (Magdoff and Van Es, 2021). Most microbes prefer warm, moist, near-neutral pH conditions to mineralize organic matter. These conditions produce an active microbial community that will decompose manure, compost, and plant residues into their final inorganic products of carbon dioxide, water, and plant-available inorganic nutrients. The nutrients found in these sources are generally found in smaller quantities and are more slowly available to the plant than those in manufactured inorganic fertilizers. Soils rich in organic matter are darker and tend to warm up faster in the spring. For these reasons, management of soil organic matter is essential for those practicing sustainable agriculture and organic production.

The ability of a soil to adsorb or hold cations is called the cation exchange capacity (CEC). Soil organic matter and clay minerals are usually negatively charged, thus allowing them to attract and retain positively charged cations from the soil solution. This is an important feature of soil materials because many plant nutrients are cations (e.g., potassium (K⁺), calcium (Ca²⁺), magnesium (Mg²⁺), ammonium (NH₄⁺) and many micronutrients). Cations are held loosely and temporarily by negatively charged exchange sites in and around soil particles that prevent leaching with every rainfall or irrigation event. However, their attraction is not permanent and the cations are easily available to plants and microbes as conditions allow. Soils that have abundant humus and smectite or vermiculite clays have larger CEC than sandy soils with little organic matter (Weil and Brady, 2017). Organic matter is also valuable in complexing or chelating some metals to make them more soluble and plant-available. In addition, soil organic matter has been shown to negate the toxicity of aluminum (Al) in some soils.

Several elements are toxic to plants and may require characterization to determine whether they are affecting plant growth and distribution on the landscape (Table 4). Aluminum, nickel (Ni), manganese (Mn), lead (Pb), copper (Cu), chromium (Cr), and cobalt (Co) are among the metals that have been analyzed and correlated with vegetation growth around mines or smelters. Some elements are plant micronutrients at low concentrations but become toxic at high concentrations in the soil or solution. Examples of these elements are boron (B), Cu, Mn and zinc (Zn). In addition, excessive quantities of some nutrients can result in deficiencies in other nutrients due to preferential uptake (Ackova, 2018).

Soil mineralogy also influences the availability of some important nutrients such as phosphate and sulfate (Weil and Brady, 2017). Allophane or clay materials derived from volcanic glass can fix or permanently adsorb phosphate and sulfate anions. Iron and aluminum oxides are prevalent in leached, acidic soils of humid zones and can combine with phosphate to make insoluble phosphate minerals that are unavailable to plants. In arid regions, where calcareous soils are common, the calcium combines with and binds the phosphate. Phosphorus is generally most plant-available at near-neutral pH and in soils not dominated by allophane, iron or aluminum oxides, or calcium carbonate. Available or extractable phosphorus is analyzed to distinguish it from organic or mineral forms that are not accessible to plants.

Potassium and ammonium may also become "fixed" or unavailable in soils high in vermiculite. The size of these ions and their charge density allow them to become permanently imbedded in the interlayer region of high-charge vermiculite clay minerals that

Table 4 Toxic elements that may affect plant growth and soil status.

<i>Trace element</i>	<i>Maximum concentration allowed in irrigation water^a (mg L⁻¹)</i>	<i>Essential to plants at low concentrations?^b</i>
Aluminum (Al)	5.0	Probably not
Arsenic (As)	0.10	Probably not
Beryllium (Be)	0.10	No
Bismuth (Bi)	NA	No
Boron (B)	NA	Yes
Cadmium (Cd)	0.01	No
Chromium (Cr)	0.10	Probably not
Cobalt (Co)	0.05	Probably not
Copper (Cu)	0.20	Yes
Fluorine (F)	1.0	Probably not
Iron (Fe)	5.0	Yes
Lead (Pb)	5.0	Probably not
Lithium (Li)	2.5	Probably not
Manganese (Mn)	0.20	Yes
Molybdenum (Mo)	0.01	Yes
Nickel (Ni)	0.20	Yes
Selenium (Se)	0.02	Probably not
Silver (Ag)	NA	No
Tin (Sn)	NA	No
Vanadium (V)	0.10	Probably not
Zinc (Zn)	2.0	Yes

NA, no maximum concentration has been designated.

^aCriteria established by the Food and Agricultural Organization (FAO, 1985) of the United Nations.

^bSome elements are required by plants at low concentrations but are toxic at higher levels (therefore 'Yes'), and some elements are still under investigation or only partially accepted as being essential, thus the term 'Probably not', but as of 2021, these are more likely to be 'No', but there are no updated universally accepted lists published.

collapse around them in response to opposite electrical charges. Thus, soluble potassium or clay mineralogy is measured to determine the amount of K available for plant uptake.

Salinity, often measured by electrical conductivity (EC), can be a problem in closed basins with shallow water tables, arid regions where evapotranspiration rates exceed precipitation rates, and in areas with poor quality irrigation water. Inorganic fertilizers are also salts and may contribute to soil EC and saline stress when applied in excess. Salinity refers to all ionic compounds in a soil solution while sodicity refers specifically to the Na concentration relative to calcium and magnesium concentrations. Not all salts are equally destructive to plants and soil. For example, while Na may disperse clay aggregates, forming an impermeable seal at the soil surface and along cracks that limits water and air movement into the soil, high Na levels can also be toxic to plants and impede water movement into plant roots. The sodium adsorption ratio (SAR) or exchangeable sodium percentage (ESP) are two measures of the sodium content affecting plant growth, especially when rainfall is insufficient to leach salts out of the soil profile. Closed basins with shallow groundwater are often sodic, saline, and detrimental to plant growth (Hopmans et al., 2021).

Salts can affect plant growth in two ways, either directly as a toxic specific ion or indirectly by lowering the osmotic potential of the soil water, which hinders the plants' ability to uptake water. Increased salts lower the osmotic potential, which in turn lowers the total soil-water potential and requires additional energy to extract water from the soil solution (Hopmans et al., 2021). Specific ion toxicity can be a problem when B, chloride (Cl), selenium (Se), or Na concentrations are excessive. Some plants are especially adapted to salinity. Adaptations include ion selectivity to prevent absorption of toxic ions through roots, sequester toxic ions into plant tissues or organelles away from critical plant tissues and plant functions, osmotic adjustment in plant tissues to ensure normal enzymatic activity, and osmotic adjustment in the rhizosphere to improve water uptake from soil matrix.

Edaphic factors affecting root and microbial growth

Many of the edaphic factors previously discussed affect plant root growth and development. Water or air movement and nutrient availability are linked to the growth of plant roots. In addition, beneficial microbial communities are found in larger quantities when the associated plants are healthier. Plant root development, microbial activity, and edaphic factors are strongly correlated (e.g., Yao et al., 2018). Additional edaphic factors that can also directly impact rooting and microbial growth are discussed below.

Soils having extremely large or small bulk density are not beneficial for microbial activity or plant growth. Compaction and high bulk density limit the root system due to the limited pore space and reduced pore space continuity for unimpeded development. Therefore, plants must use additional energy to push their roots through the soil to access more soil water and nutrients. Limited

porosity also reduces water and air movement which limits both root and microbial growth. Consequently, the symbiotic relationship of soil microorganisms and plant roots is weaker (Gregory, 2006; Weil & Brady, 2017). Conversely, in soils with extremely low bulk density, the application of moderate compaction may enhance plant growth by increasing contact between soil and developing roots. Microbial activity may also be enhanced by closer proximity to the soil water and nutrients.

Soil temperature is an important edaphic factor, influencing plant growth and root development but has limited impact on air movement in soil and moderate impact on plant nutrient availability. Soil temperature is affected by several physical properties including porosity, water content, soil color, organic matter content, bulk density, and soil depth. Temperature is also a function of location in the landscape and physiographic features such as aspect, elevation, and latitude. Agriculture management practices that moderate soil temperature vary by region, time of year, and crop.

Shallow soils resulting from wind and water erosion are problematic for proper root growth of most desirable plants. Agricultural management practices should be employed to minimize soil loss, especially between crops when the soil is most exposed to the weather. The maintenance of good plant cover year-round is one of the best ways to minimize soil loss, and thus is an important component in soil conservation and management. Physiographic factors such as slope percentage, aspect, and location on the landscape are often recorded along with edaphic factors to evaluate plant–soil–landscape relations.

Most urban soils are shallow and contain a smaller root zone volume than needed by the plant. In addition, these soils may serve a utilitarian purpose, such as a drivable surface for emergency traffic or temperature moderation to reduce urban heat island effects. The soil conditions surrounding homes and other edifices are usually not ideal because of the loss of topsoil and displacement of deeper soil horizons during construction. These newly exposed soils will contain smaller concentrations of organic matter and less diverse microbial communities. Due to soil limitations and other urban hazards, urban plants have shorter life spans than those same plants found outside urban centers. However, with continued plant growth, organic matter inputs and careful management, these soils can become more favorable for future plants and microorganisms.

Conclusion

Edaphology is the study of the soil as a plant growth medium. To support healthy, high functioning ecosystems, one needs good soil conditions in which to grow the necessary flora and fauna. Soils serve a variety of functions in the environment and are a reusable, but not a quickly renewable resource. Edaphology research provides guidance on the management of this precious resource for both agronomic and ecologic functions.

References

- Ackova DG (2018) Heavy metals and their general toxicity on plants. *Plant Science Today* 5: 15–19. <https://doi.org/10.14719/pst.2018.5.1.355>.
- Brantley S, White T, White A, Sparks D, Richter D, Pregitzer K, et al. (2006) Frontiers in exploration of the Critical Zone. In: *Frontiers in exploration of the Critical Zone: Report of a workshop sponsored by the National Science Foundation (NSF) (1–30)*. Newark, DE.
- Committee on Basic Research Opportunities in the Earth Sciences (2001) The critical zone: Earth's near-surface environment(link is external). In: *Basic Research Opportunities in Earth Science*, pp. 35–45. Washington, D.C.: National Academy Press.
- Doran JW, Coleman DC, Bezdicek DF, and Stewart BA (eds.) (1994) *Defining Soil Quality for a Sustainable Environment*. In: *SSSA Special Publication No. 35*, Madison, WI: Soil Science Society of America.
- ECFR (Electronic Code of Federal Regulations) (2011) Title 7. Agriculture. <https://ecfr.federalregister.gov>. Accessed 8 Sep 2021.
- FAO (1985) *Water Quality Guidelines for Maximum Crop Production*. UN: Food and Agricultural Organization. <http://www.fao.org/3/T0234e/T0234E06.htm>.
- Gregory PJ (2006) Roots, rhizosphere and soil: The route to a better understanding of soil science? *European Journal of Soil Science* 57: 2–12. <https://doi.org/10.1111/j.1365-2389.2005.00778.x>.
- Hillel D (1998) *Environmental Soil Physics*. San Diego, CA: Academic Press.
- Hopmans JW, Qureshi AS, Kisekka I, Munns R, Grattan SR, Rengasamy P, et al. (2021) Critical knowledge gaps and research priorities in global soil salinity. *Advances in Agronomy* 191.
- Lindbo DL, Koslowski DA, and Robinson C (eds.) (2012) *Know Soil Know Life*. Madison, WI: Soil Science Society of America.
- Luo W, Griffin-Nolan RJ, Ma W, Liu B, Zuo X, Xu C, Yu Q, Luo Y, Mariotte P, Smith MD, Collins SL, Knapp AK, Wang Z, and Han X (2021) Plant traits and soil fertility mediate productivity losses under extreme drought in C3 grasslands. *Ecology*. <https://doi.org/10.1002/ecs.3465>. e03465.
- Magdoff F and Van Es H (2021) Building soils for better crops: Ecological management for healthy soils. *Sustainable Agriculture Research and Education (SARE), supported by the National Institute of Food and Agriculture (NIFA)*. 4th edn, *Handbook Series Book 104th edn*?, College Park, MD: University of Maryland.
- Richter DD (2019) Game changer in soil science: The Anthropocene in soil science and pedology. *Journal of Plant Nutrition and Soil Science* 2019: 1–7. <https://doi.org/10.1002/jpln.201900320>.
- Soil Science Society of America, Inc. (2008) *Glossary of Soil Science Terms*. <https://doi.org/10.2136/2008.glossarysoilscienceterms>.
- Stroud JL (2019) Soil health pilot study in England: Outcomes from an on-farm earthworm survey. *PLoS One* 14(2): e0203909. <https://doi.org/10.1371/journal.pone.0203909>.
- USDA (2020) *Cropland In-Field Soil Health Assessment Guide. Soil Health Technical Note No. 450-06*. Washington, DC: U.S. Department of Agriculture, Natural Resources Conservation Service.
- Weil RR and Brady NC (2017) *The Nature and Properties of Soils*, 15th edn, Pearson Education. ISBN: 978-0133254488.
- Willer H, Travnicek J, Meier C, and Schlatter B (eds.) (2021) *The World of Organic Agriculture - Statistics & Emerging Trends 2021*. Bonn: Research Institute of Organic Agriculture FIBL, Frick, and IFOAM – Organics International. (v20210301).
- Woods MS, Stowell LJ, and Gelernter WD (2016) *Minimum soil nutrient guidelines for turfgrass developed from Mehlich 3 soil test results*. No. e2144v1. PeerJ Preprints.
- Yao X, Zhang N, Zeng H, and Wang W (2018) Effects of soil depth and plant-soil interaction on microbial community in temperate grasslands of northern China. *Science of the Total Environment* 630: 96–102. <https://doi.org/10.1016/j.scitotenv.2018.02.155>. Epub 2018 Feb 21. PMID: 29475117.

Biodiversity

Monica A Farfan and Diana H Wall, School of Global Environmental Sustainability, Colorado State University, Fort Collins, CO, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of D.H. Wall, BIODIVERSITY, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 136–141, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00140-5>.

Introduction	10
Diversity and abundance	10
Distribution	12
Soil biodiversity and ecosystem functioning	13
Global change and impacts on soil biodiversity	14
Sustaining soil biodiversity	15
References	15

Abstract

Soil biodiversity contributes to ecosystem services such as nutrient cycling, water filtration, pest control, carbon storage, and soil stabilization, all important for human well-being. Diversity in the soil ranges from the microbes to vertebrates that sometimes use soils for only part of their lives. Scientists use many tools to study soil biodiversity depending on the organisms being studied. Recently molecular techniques have been employed to understand global distributions and changes in soil biodiversity. Many factors affect biodiversity, including organic material and climate. Understanding the threats to biodiversity is important for conservation of this vital resource for life on Earth.

Key points

- Soil biodiversity provides many ecosystem services that support other life on Earth including human well-being including nutrient cycling, water filtration, and soil fertility.
- The diversity in soil is very high, ranging from soil microbes to invertebrates and vertebrates.
- Understanding the distribution of soil biodiversity globally is challenging and many different techniques are used for this, including molecular techniques.
- Many factors influence soil biodiversity, including climate and organic material.
- Understanding emerging threats is key to conserving soils for human health and well-being.

Introduction

Soils teem with life. Representatives of almost every phylum of organism known aboveground also occur in soil. In less than a handful of soil, there can be millions of types of microbes and hundreds of species of microscopic invertebrates. The identities and natural histories of these microscopic flora and fauna and many of the larger, visible soil fauna are the least-known biota in terrestrial ecosystems. However, new evidence is rapidly revealing the major contributions of soil organisms to the maintenance of life on Earth, opening a new frontier for exploration and a rising concern about the loss of soil biodiversity with increasing degradation of the soil habitat locally and globally. Soil biota regulate numerous ecosystem processes. Their activities provide a wealth of essential ecosystem services for humans, such as breaking down organic matter, filtering water, stabilizing soil, generating and renewing soil fertility, providing nutrients for plant growth, modifying the hydrologic cycle (including mitigating floods and controlling erosion), and controlling pest and pathogens of plants and animals. Plants, soil biota, and their consumers are tightly coupled and increasingly soil organisms are recognized for their influence on broader ecological functions such as productivity, decomposition rates, patterns of biodiversity and global change (Table 1). Soil biodiversity influences ecosystem processes in all ecosystems, terrestrial and aquatic including the atmosphere, through regulation of CO₂ flux and carbon sequestration in soils. Understanding how soil biodiversity will change and as a consequence, alter ecosystem functioning as fertile soils continue to decline globally is important knowledge for policymakers and members of the public who need to develop and implement strategies for the future.

Diversity and abundance

Soil biodiversity includes a plethora of life that is hidden from our everyday view: viruses; bacteria; actinomycetes; fungi; algae; protists (single celled eukaryotes); microscopic invertebrates such as rotifers, tardigrades (water bears), soil planaria (flatworms),

Table 1 Estimated diversity and abundance of soil taxa based on published literature and supported by expert judgment.

Taxon	Diversity per amount soil or area (taxonomic units indicated below)	Abundance (approximate)
Bacteria and Archaea	100–9000 cm ⁻³	4–20 × 10 ⁹ cm ⁻³
Fungi (including mycorrhizal fungi)	200–235 g ^{-1a}	100 mg ⁻¹
AMF (species)	10–20 m ⁻²	81–111 m cm ⁻³
Protists (sequences)	253–961 g ^{-5b}	10 ⁴ –10 ⁷ m ⁻²
Nematodes (genera)	10–100 m ⁻²	2–90 × 10 ⁵ m ⁻²
Enchytraeids	1–15 ha ⁻¹	12,000–311,000 m ⁻²
Tardigrades	NA	490 mg ^{-1c}
Collembola	20 m ⁻²	1–5 × 10 ⁴ m ⁻²
Mites (Oribatida)	100–150 m ⁻²	1–10 × 10 ⁴ m ⁻²
Isopoda	10–100 m ⁻²	10 m ⁻²
Diplopoda	10–2500 m ⁻²	110 m ⁻²
Earthworms (Oligochaeta)	10–15 ha ⁻¹	300 m ⁻²

^aOperational taxonomic units.^bGeisen et al. (2015).^cCzernerková et al. (2015).

and nematodes (roundworms); the microarthropods, Acari (mites) and Collembola (springtails); and larger invertebrates, easily seen by the naked eye such as terrestrial gastropods (snails and slugs), isopods (pill bugs, sow bugs), Oligochaeta (enchytraeids and earthworms), spiders, scorpions, beetles, centipedes, millipedes, crustaceans, ants, and termites. Because of this diversity, the invertebrates are frequently grouped by size (body width) as microfauna, mesofauna, and macrofauna (Orgiazzi et al., 2016; Fig. 1). Vertebrates also depend on the soil as a habitat, for example, moles, prairie dogs, meercats, wombats, small rodents, and some species of lizards, snakes, frogs, and even birds. Many aboveground invertebrates may be temporary inhabitants (such as nematode parasites of insects; immature Diptera such as tipulid crane flies and tephritid fruit flies; cicadas; and Coleoptera such as

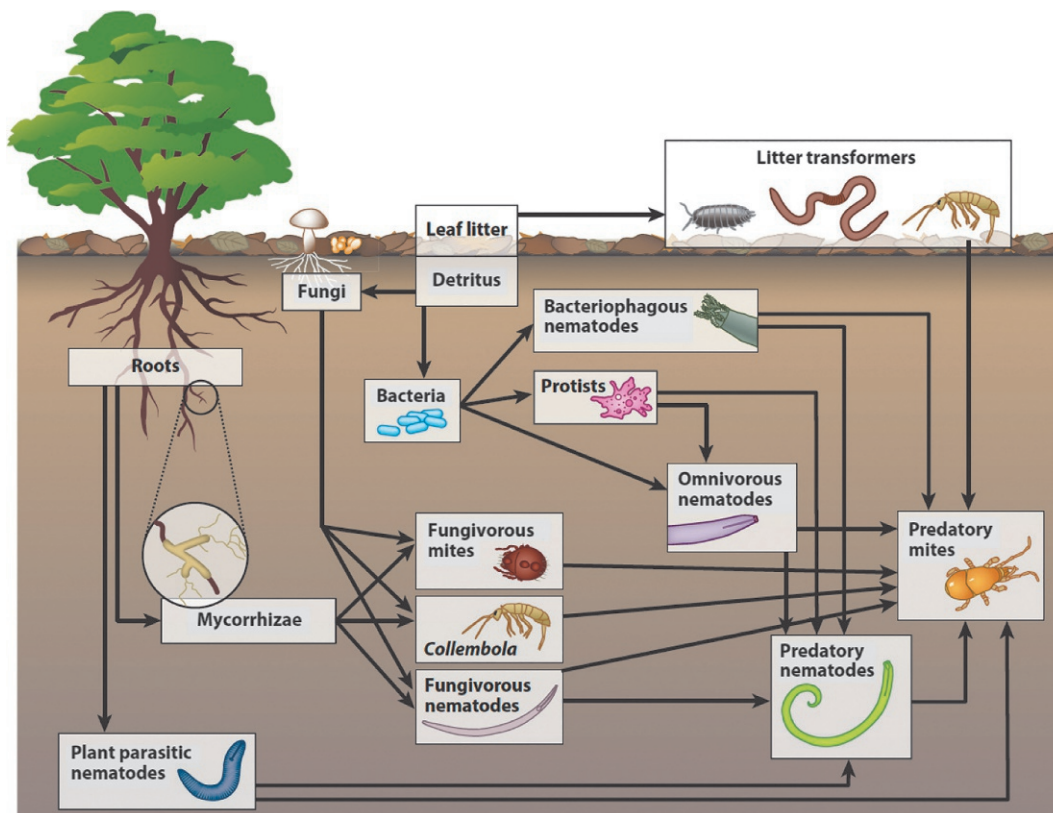


Fig. 1 A simplified soil food web representing the substantially more complex real-life food soil food web that includes many other groups of soil fauna. From Nielsen UN, Wall DH, and Six J (2015) Soil Biodiversity and the Environment. *Annual Review of Environment and Resources* 40: 63–90.

scarab beetle) living in the soil for only one stage of their life cycle or occurring seasonally. Thus, soil diversity is dynamic and tightly connected to the aboveground biotic and climatic systems.

It appears that billions of bacteria, meters of fungal hyphae, and millions of protozoa, nematodes, oligochaetes, and arthropods can inhabit a square meter of soil (Table 1). The diversity of these species appears so immense that knowledge of the numbers of species (species richness) in just two groups, the microarthropods and nematodes, could substantially increase overall global biodiversity estimates. Yet our understanding of the total numbers of soil species even within any one group is surprisingly poor. Data suggest 90–95% of soil biota are unidentified and unknown to science. There is no single location where the total number of all soil species or molecular types of microbes, invertebrates, or vertebrates has been assessed, although up to 1000 invertebrate species have been enumerated by traditional methods in a square meter of forest soil. Global estimates have been based on many studies of varying numbers and sizes of soil samples from a variety of ecosystems and generally only to shallow (0–20 cm) depths. Unfortunately, this leads to an underestimate of soil biodiversity given that mites and nematodes occur to 15 m depths in deserts and bacteria and nematodes to greater depths elsewhere. Nonetheless, global estimates of the total numbers of species identified morphologically from soils (Fig. 1) include 600 species of enchytraeids, 1500 species of soil protozoa, 5000 species of nematodes, and 20,000–30,000 species of mites (Freckman and Virginia, 1989).

Even with the enormity of soil biodiversity, there is excitement about the species yet to be identified and their role in ecosystems. Advancements in technology such as molecular tools, biochemical methods, and microscopic advancements (scanning electron microscope, image analysis) are enabling faster identification or characterization of organisms. For example, many studies now use new molecular technologies to measure both microbial (Fierer and Jackson, 2006) and eukaryotic biodiversity and compare assemblages quantitatively among sites. Examples of the eukaryotes include soil arthropods (Oliverio et al., 2018) and earthworms (Decaëns et al., 2013). By examining multiple gene regions, researchers can simultaneously assess and compare functional (e.g., diversity of protein-coding gene categories), phylogenetic and taxonomic diversity of soil communities. Advances in bioinformatics allow comparisons among huge molecular datasets as well as identification of soil organisms to standardized taxonomic levels (van den Hoogen et al., 2019). Stable isotope technology has improved analysis of the contribution of different groups of taxa to carbon and nitrogen flow in the soil food web. Scientists, distributed globally with expertise in specific soil biotic groups, have increased evidence of the number of species and what they do in an ecosystem, by sharing data, images and results via the internet, and creating web pages for the public.

These new methods, when combined with traditional methods, are accelerating knowledge of belowground biodiversity and ecosystem functioning. However, there are few taxonomists and systematists remaining who specialize in the identification, natural history, and phylogenies of the soil biota. Specialists that are versed in the techniques for collecting, extracting, and identifying the various organisms provide essential information for analysis of different ecosystems, but there are no standard techniques for the collection or extraction of even a single group of fauna. For example, a handful of soil might suffice for extracting and identifying many species of microscopic nematodes, mites, protists, and microbes, but a different technique would be required for identification of earthworms. Technology, new educational and data analysis tools, better training and the excitement of conquering these challenges are attracting students to this new frontier of science (FAO et al., 2020; Orgiazzi et al., 2016).

Distribution

Soils, like air and water, are resources that have developed over geologic time and are non-renewable in our lifetime. Both the type of soil and the evolutionary adaptation and dispersal mechanism of an organism to a particular location play a role in establishing the structure of a soil biotic community. Soil chemical and physical factors combine to make each soil a unique habitat. Chemical (e.g., pH, salinity, type and amount of organic matter and nutrients) and physical properties (e.g., texture -amount of sand, silt, and clay, pore size, moisture, temperature) along with the composition of vegetation and geologic history, all contribute to the evolution of the diversity of species in soil. These components are all intertwined, and changes in any one factor can result in a new habitat for the biota. For example, an increase in organic matter can also increase the soil water-holding capacity, which tends to result in a darker soil that will have higher soil temperatures. Soil texture is also linked to microclimate. As clay content increases, soil fertility, soil moisture, and soil temperature can increase. Clay content is in turn largely dependent on the parent material (granite related to sandy soils versus limestone related to clayey soils) or landscape position (a floodplain is more fine-grained, a ridge top more coarse-grained). Thus, beneath our feet, there are thousands of soil habitats due to many combinations of factors such as pH, fertility, organic matter, soil moisture, climate, and geologic history.

Soil biota occupy remarkably different habitats within the soil with some species living only in decaying leaves and dead wood, or in the rich organic matter horizon, while others occur several centimeters to meters deeper in the mineral soil (Coleman et al., 2017). For example, some species of earthworms occur in litter-organic horizons and others exist in deeper mineral soil. Species of organisms may be very different even at millimeter scales near a growing root than further away in bare ground. Increasing to larger scales, within a backyard or agricultural field, the diversity (species richness) might be similar. However, comparison of an agricultural field to a nearby natural area will generally show a very different community. At a global scale, primary factors affecting the biogeography of soil organisms are climate, the type (trees, grasses, forbs) and amount of vegetation (organic matter), and soil texture (composition of sand, silt, and clay). But these can vary with the type of organism, for example a primary factor for the distribution of nematode functional groups is soil carbon (van den Hoogen et al., 2019). In general, microarthropod-rich soil

communities are found in forest ecosystems which have lower soil pH, root biomass, mean annual temperature, low soil inorganic N and higher C:N, litter and moisture than grasslands. Hot and cold desert ecosystems have lower levels of diversity than forest, grassland or tundra ecosystems. For example, earthworms are generally absent from dry ecosystems, illustrating an evolutionary adaptation to moist soils (Phillips et al., 2019). Bacterial diversity appears unrelated to site temperature, latitude, and other variables used to predict plant and animal diversity and instead correlates strongly with soil pH (Fierer and Jackson, 2006). A growing focus on using computer learning and in modeling diversity, abundance and activity of soil microorganisms across the globe is providing some perspective to help scientists characterize community-level and regional understanding of how soil organisms affect carbon storage and where to concentrate efforts to mitigate effects of climate change and biodiversity loss (Crowther et al., 2019). As data are compiled on organisms and their natural history and then combined with geographic information systems (GIS) maps of soil and vegetation, reliable estimates of the global biogeography of soil microflora and fauna will become feasible.

Dispersal mechanisms of biota are a key to their distribution at local, regional, and global scales. Soil organisms are transported with soil: by humans transporting plants or by cars, trucks, or other machinery; in rainwater, floods, seawater; carried externally and internally by animals such as birds, ungulates, and insects; in plant debris; and by wind. Protists, rotifers, tardigrades, and nematodes are aquatic biota, living in water films around soil particles, but as soil dries, they enter a dormant state, anhydrobiosis (meaning life without water), reduce their body surface area, water content, and metabolism, and can live for years until they revive in water. While in anhydrobiosis, they can be wind-transported over hundreds of miles. Collembola live in air pores of soil and in organic matter and were long thought to be confined to small geographic areas. However, they are now known to survive and move to other continents in seawater, because of a physiological advantage, their hydrophobic (unwettable) cuticles. This finding on a new means of dispersal will accelerate hypotheses on the global distribution patterns of Collembola. Other dispersal mechanisms are based in geologic time. For example, earthworm distribution in North America has been patterned largely by glaciation.

Soil biodiversity and ecosystem functioning

Soil organisms moderate numerous ecosystem processes (e.g., decomposition, C and N transformation, hydrologic cycles) that are a major component of global cycles. Land-use change, climate change, pollution, desertification and an increase in invasive species are all components of 'global change': whether global change results in the resilience or alteration of species composition within the soil community, and how this may impact ecosystem function, are important questions in the research of soil biology today. Little knowledge exists about the functional attributes of the majority of individual species, but there is considerable information on the consequence of larger species on soil processes. Macrofauna (earthworms and termites) species have a significant effect on soil structure and hydrology, and transfer and reconfigure organic matter, affecting nutrient mobility. The influence of these animal species on surface-soil plant litter can be measured at local and larger regional scales, but for smaller species, scientists have used other approaches to determine their influence on ecosystem processes. Scientists lump species with similar morphologies and feeding sources into functional groups and measure the transformations of carbon and nitrogen through each functional group within the soil food web. A couple of examples illustrate this in the following paragraphs.

Perhaps some of the greatest diversity is involved in the decomposition (decay) component of the soil food web. In natural terrestrial ecosystems, the bulk of the world's dead plant material (leaves, stems, dead roots, also known as detritus) falls to the ground to be decomposed. The rate of decomposition in an ecosystem is affected by climate, the quantity and quality (type or chemical composition, generally C, N, and lignin) of the substrate or detritus (leaf versus wood versus animal material, for example), and a 'detritivore' functional group in the soil food web that includes many species of generalist feeders from numerous phyla. If the substrate is more recalcitrant (has more lignin or a larger C:N ratio, e.g., wood of trees) decomposition will occur primarily by fungi and fungal consumers. If the organic matter is more labile (less lignin or smaller C:N, ratio e.g., some grasses) bacteria and their consumers generally are the major decomposers.

The breakdown of detritus is a product of a succession of the species of invertebrates and fungi, actinomycetes and other bacteria, all with specific roles in decaying complex organic substrates into inorganic nutrients that are then available for plant growth. For example, some faunal species (e.g., arthropods) shred leaves into smaller particles; others channel through the soil, increasing the distribution of microbes and the organic matter while affecting soil porosity in the soil profile; other fauna ingest the pieces of organic matter, further exposing it to enzymes that aid the decay, and then pass it from their bodies in a different chemical form. Smaller microscopic animals, nematodes, mites, and Collembola prey on the decomposers, bacteria and fungi, maintaining them in an accelerated growth phase, accelerating the transfer of nutrients to soil. As with other biotic food webs, predators are at the top of the soil food web. Interestingly, predators in soil food webs can be microflora (e.g., fungi that prey on nematodes) and microfauna (predaceous nematodes, or mites or tardigrades feeding on nematodes), as well as macrofauna. Understanding predator-prey relationships of species can contribute to biocontrol management of soil pathogens of humans and plants.

Aboveground and belowground taxa are tightly coupled ecologically and interactions between them are known as plant-soil feedbacks (Nielsen et al., 2015). For example, species with mutualistic symbionts such as mycorrhizal fungi can enhance access to limiting nutrients, with a positive feedback on productivity. Alternatively, the accumulation of pathogens or root herbivores in early successional soils may weaken pioneer plant species and pave the way for later successional stages through negative feedbacks.

Invasive plant species can produce chemicals that reduce local mycorrhizal fungi or promote soil pathogens more harmful to surrounding native species than the introduced species. The theoretical framework around these feedbacks will help explain the dynamics and succession of vegetation, the invasiveness of introduced species in new habitats and how terrestrial ecosystems may respond to land use and climate change.

Global change and impacts on soil biodiversity

Land-use change is the major driver affecting soil biodiversity and future soil sustainability. Land-use changes (tillage, pollution, paving with concrete, erosion, dams, change in plant species) affects soil physical and chemical properties, soil structure, and the base of the soil food web (chemical composition of plants, amount of organic matter, manure) and thus have direct impacts on species composition. The conversion of a natural grassland or forest to a managed system for agriculture, pasture, urban, or industrial use changes the determinants of soil biodiversity, the vegetation, soil structure, and microclimate. The disruption to the natural vegetation and the soil habitat with land-use change decouples the nutrients provided by the decomposition food web from plant uptake. The result is a loss in soil fertility provided by the original soil and its inhabitants. Additional fertilizer is required, and pesticides may be necessary especially with more intensive agriculture because the new soil food web generally has fewer predators, resulting in a change in or loss of biocontrol of plant pathogens (Tsiafouli et al., 2015). While conservation tillage methods are beneficial in conserving carbon in soils and creating a food web that is more detritus-based, often a lower level of herbicides and pesticides is used. There are examples from tropical wet and dry forests, grasslands, deserts, and other ecosystems showing that these land-use changes affect the total soil biota (macrofauna, microfauna, and microflora), generally reducing species diversity. Desertification resulting from land-use change has a considerable impact on soil processes, including reduction of soil carbon, soil structure, soil biota, and soil fertility.

Invasive species of plants and animals also influence changes in soils and biotic communities. The movement of plants, people, and machinery has not only increased the numbers of species invasions into soils globally, but has significantly increased impacts on soil biodiversity and some characteristic of the ecosystem (e.g., nutrient cycling, species richness and or abundance, plant factors, soil physical or chemical factors) and increased costs of eradication. In some cases, freedom from native pathogens, root parasites and herbivores allow introduced plant species to proliferate in a new habitat. Furthermore, invasive species often influence surrounding biodiversity through negative plant soil feedbacks. An earthworm species introduced to New York has changed forest-floor litter quality and composition, soil chemistry, and water infiltration rates. European nations are examining ways to eradicate the New Zealand planarian flatworm, *Artioposthia triangulata*, a predator of earthworms. This invasive species has impacts aboveground (removing the food source for birds) and belowground (removing an animal species that influences organic matter transformation, soil hydrology, and structure). Invasive plant species with differing rooting depths and plant chemical composition can have repercussions for soil communities. Woody plant invasions into grasslands of the Great Plains of the USA have greater rooting depths (affecting soil carbon storage) and, at these depths, a more depauperate nematode community.

Increasing levels of CO₂ and climate change (temperature and rainfall patterns, and extreme, infrequent events such as droughts and flooding) indirectly affect soil biodiversity through impacts on plant community composition or soil chemistry. Soil biota play a crucial role in terrestrial carbon cycling, and can be sensitive to climate change. Although data vary with ecosystem, modifications in the soil community and changes in belowground herbivory and decomposition pathways for soil food webs have been documented. Even in the Antarctic Dry Valleys, where there are no visible plants, climate change has decreased soil nematode populations through changes in temperature and available soil water. At a global scale, warmer, wetter conditions associated with climate change will result in altered distributions of aboveground taxa with unpredictable consequences for soil biota. Due to the differing abiotic controls acting on aboveground and belowground diversity, their response to climate change will likely differ temporally and in magnitude. Since these systems are intricately linked together, decoupling of aboveground/belowground diversity may result in overall lower productivity and reduced diversity. Direct effects of elevated CO₂ are considered to be less important, since soils have high ambient levels of CO₂ owing to root and microbial respiration.

Changes in rainfall are now known to disrupt the balance between the abundance of root-feeding nematodes in grasslands, which are major constraints to ecosystem primary production, and their predators in favor of the root feeders (Franco et al., 2019). These nematode responses increase in magnitude from arid to moist grasslands, and further aggravate the effects of drought on grassland primary productivity and above-belowground plant biomass partitioning (Franco et al., 2020).

Soil food webs can be modeled to simulate the loss of species due to global change and to examine the effect, if any, on nutrient cycling and plant production. For example, in the short grass steppe of Colorado, USA, a model was assembled using field data on the microbes and micro- and mesofauna (Hunt and Wall, 2002). A single factor global change scenario (elevated CO₂) and resulting change of plant species composition, which might indirectly result in the loss of soil biodiversity, was simulated for each of the 15 functional groups in the soil food web. For example, loss of bacteria and fungi led to extinctions of other groups in the soil food web. Removal of six of 15 groups impacted the abundance of other groups, and three of these groups (bacteria, saprophytic fungi, and root-feeding nematodes) caused up to a 10% change in two ecosystem processes, nitrogen mineralization, and plant primary production. More recently, studies show that multiple global change factors acting simultaneously affect soil communities and the processes they control that is unpredictable from single- or two-factor studies (Rillig et al., 2020; Thakur et al., 2019).

Soil food web models may not reflect the complexity of individual species interactions and the underlying habitat but are critical for scientific analysis and forecasting for future scenarios of C and N cycling. The C and N cycling of the soil food web ties the soil to

the aboveground ecosystem by relating the soil biota to soil carbon storage, nutrients available for plant growth and to atmospheric CO₂ and trace gas fluxes. Models, along with studies at species and functional group resolution, are currently being examined in experiments that manipulate biodiversity in field, laboratory (microcosms with generally less than 10 species), and in large soil experiments in plant growth facilities. Together these provide a theoretical and quantitative basis for integrating the soil biota into global ecosystem studies.

Sustaining soil biodiversity

Soil biodiversity is an important resource that regulates ecosystem processes essential to the functioning of earth's ecosystems. Our understanding of the species, their interactions, and effect on processes occurring in the soil food web in natural systems are an important contribution to management of land, particularly agriculture. The link between aboveground and belowground diversity is strong, although occurring at different temporal scales for organisms, and changes affecting aboveground diversity and function are reflected in belowground ecosystems. An immediate effect is a decrease in the biological capacity of soils and a change in the regulation of interactions and processes. Knowledge on whether all or a few key taxa are important in this regulation of ecosystem processes is a high priority for planning for future sustainability.

Current priorities in the study of soil biodiversity include: (1) understanding the functions of rarely sampled fauna so that species estimates and global distributions are based on an accurate and current database; (2) determining which habitats are most vulnerable for soil biodiversity loss, e.g., where the 'hot spots' of biodiversity are, and which habitats and what time frames are most amenable to restoration; (3) synthesizing data and determining which invertebrate and microbial 'species' are key to ecosystem processes; (4) identifying gaps in knowledge of multispecies interactions, including plant soil feedbacks and their influence on ecosystem functioning; (5) collaborating on long-term (more than 3 years) experiments to examine effects of global changes on aboveground-belowground biodiversity linkages and ecosystem functioning; (6) determining, based on natural history information, which species are more likely to be invasive if introduced, and using this information as a management tool to reduce spread and identify threats to other species and (7) predicting the effect of a warmer wetter planet, brought about through climate change, on resilience of soil biodiversity and function. Given advancements in methods for assessing soil biodiversity, and by building upon current knowledge, mapping and assessing soil biodiversity from regional to global scales is now possible. Continued research into soil biodiversity is essential if we are to measure the effect of the many threats and determine options to conserve this vital resource for human health and well-being.

References

- Coleman DC, Callahan MA, and Crossley DA Jr. (2017) *Fundamentals of Soil Ecology*. Academic Press.
- Crowther TW, van den Hoogen J, Wan J, et al. (2019) The global soil community and its influence on biogeochemistry. *Science* 365: eaav0550.
- Czernerková M, Jönsson KI, Hajer J, and Devetter M (2015) Evaluation of extraction methods for quantitative analysis of tardigrade populations in soil and leaf litter. *Pedobiologia* 70: 1–5.
- Decaëns T, Porco D, Rougerie R, Brown GG, and James SW (2013) Potential of DNA barcoding for earthworm research in taxonomy and ecology. *Applied Soil Ecology* 65: 35–42.
- FAO, ITPS, GSBI, SCBD, and EC (2020) *State of Knowledge of Soil Biodiversity - Status, Challenges and Potentialities, Report 2020*. Rome: FAO.
- Fierer N and Jackson RB (2006) The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences* 103: 626–631.
- Franco ALC, Gherardi LA, de Tomasel CM, et al. (2019) Drought suppresses soil predators and promotes root herbivores in mesic, but not in xeric grasslands. *Proceedings of the National Academy of Sciences* 116: 12883–12888.
- Franco ALC, Gherardi LA, de Tomasel CM, et al. (2020) Root herbivory controls the effects of water availability on the partitioning between above and belowground grass biomass. *Functional Ecology* 34: 2403–2410. <https://doi.org/10.1111/1365-2435.13661>.
- Freckman DW and Virginia RA (1989) Plant-feeding nematodes in deep-rooting desert ecosystems. *Ecology* 70: 1665–1678.
- Geisen S, Tveit A, Clark I, et al. (2015) Metatranscriptomic census of active protists in soils. *ISME Journal* 9: 2178–2190.
- Hunt HW and Wall DH (2002) Modelling the effects of loss of soil biodiversity on ecosystem function. *Global Change Biology* 8: 33–50.
- Nielsen UN, Wall DH, and Six J (2015) Soil biodiversity and the environment. *Annual Review of Environment and Resources* 40: 63–90.
- Oliverio AM, Gan H, Wickings K, and Fierer N (2018) A DNA metabarcoding approach to characterize soil arthropod communities. *Soil Biology and Biochemistry* 125: 37–43.
- Orgiazzi A, Bardgett RD, and Barrios E (2016) *Global Soil Biodiversity Atlas*. European Commission.
- Phillips HR, Guerra CA, Bartz ML, et al. (2019) Global distribution of earthworm diversity. *Science* 366: 480–485.
- Rillig MC, Ryo M, Lehmann A, et al. (2020) The role of multiple global change factors in driving soil functions and microbial biodiversity. *Science* 366: 886–890.
- Thakur MP, Del Real IM, Cesarz S, et al. (2019) Soil microbial, nematode, and enzymatic responses to elevated CO₂, N fertilization, warming, and reduced precipitation. *Soil Biology & Biochemistry* 135: 184–193.
- Tsiafouli MA, Thébaud E, Sgardelis SP, et al. (2015) Intensive agriculture reduces soil biodiversity across Europe. *Global Change Biology* 21: 973–985.
- van den Hoogen J, Geisen S, Routh D, et al. (2019) Soil nematode abundance and functional group composition at a global scale. *Nature* 572: 194–198.

Relevant Website

<https://globalsoilbiodiversity.org>—Global Soil Biodiversity Initiative.

Bacteriophage (Viruses)

Li Bi^a, Ji-Zheng He^a, Dan-Ting Yu^b, Li-Li Han^c, and Hang-Wei Hu^a, ^aSchool of Agriculture and Food, Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Parkville, VIC, Australia; ^bState Key Laboratory for Subtropical Mountain Ecology of the Ministry of Science and Technology and Fujian Province, Fujian Normal University, Fujian, China; ^cState Key Laboratory of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing, China

© 2023 Elsevier Ltd. All rights reserved.

This is an update of M Radosevich, KE Williamson, KE Wommack, Bacteriophage, Editor(s): D Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 122–129, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00463-X>.

Introduction	17
Diversity and life strategies of phages	17
Advances in methodologies to study soil viruses	18
Potential roles of soil phages in the terrestrial ecosystem	20
Contribution of soil phages to biogeochemical element cycles	20
Potential roles of soil phages in plant nutrient uptake and adaption	20
Potential roles of phages in regulating crop health	21
Feedback responses of soil phages to global changes	22
Conclusions and outlook	23
References	23

Abstract

Viruses are an essential component of soil biodiversity and contribute to important ecosystem functions. In the past, soil viruses have been largely understudied due to methodological challenges of extracting viruses from the complex soil environment. This has gradually changed with the availability of recent metagenomic analysis. In this chapter, we revisit phage diversity and life strategies, introduce recent advances in utilizing metagenomic methods to characterize soil phages, and summarize the potential roles of soil phages in mediating biogeochemical element cycles and plant health by influencing their microbial hosts. Finally, we discuss the feedback responses of soil phages to global changes and highlight the importance of incorporating the roles of soil phages into global biogeochemical models.

Glossary

Auxiliary metabolic genes (AMGs) Phage-contained genes that can modulate host metabolism and improve host fitness. Phages often obtain those genes from their hosts via horizontal gene transfer.

Bacteriophages (phages) Viral particles that infect bacteria, accounting for the largest proportion of virosphere in the natural environment.

Microdiversity Intra-population genetic variation. The viral microdiversity in natural environments is largely underexplored.

Phyllosphere The aerial part of a plant, harboring diverse microbes on the surface and within the plant organs. Phyllosphere microbiome is critical to plant growth.

Prophage A temperate phage that can lead to a lysogenic infection, integrate its genome into host chromosome and replicate in conjunction with host genomes.

Superinfection exclusion A phenomenon that a primary phage infection can protect hosts via preventing the replication of a second infection by a closely related phage.

Total soil metagenomes A culture-independent and sequencing-dependent approach to study soil microbiome, which extracts and analyzes DNA of all entities from soil sample.

Viral ecology The studies focusing on viral abundance, community composition, biogeography, host prediction and their potential ecological functions.

Viral shunt A process during which hosts are lysed by phages and host biomass is transferred into available nutrients (i.e., dissolved organic matter) for neighboring microorganisms, with consequences for global biogeochemical cycles.

Viromes Viral metagenome that is retrieved using a series of procedures including size-fractionation to remove most cellular contamination, concentration of virus-like particles, and extraction of viral DNA.

Key points

- Phages have enormous morphological and genomic diversity and diverse life strategies.
- Metagenomic approaches have radically improved our understanding of soil viral ecology.

- Soil phages play an important role in carbon and other biogeochemical cycles.
- Soil phages are critical to plant nutrient uptake and health through modulating plant associated microbial communities.
- Global changes have important consequences for soil phage diversity and potential ecological functions.

Introduction

Viruses are highly diverse and ubiquitous in the environment. They can infect every cellular organism and are recognized as a critical component in food web ecology. *Viral ecology* has been widely studied in marine systems where viruses are known to be instrumental in a range of ecological functions through infecting and regulating marine microbial communities. For example, viruses lyse approximately 20–40% of marine prokaryotes, acting as a control of host abundance and community compositions. During the viral lysis process, nutrients are released from host cells, with important consequences for global element cycles (Breitbart et al., 2018). Viruses can also reprogram host metabolism and modulate biogeochemical cycles through expressing auxiliary metabolic genes (AMGs) that are involved in a wide variety of metabolic functions and can extend the ecological niches of hosts during infection (Hurwitz and U'Ren, 2016). Marine viruses are also known to drive microbial evolutionary trajectories, through their co-evolution with hosts or functioning as major horizontal gene transfer agents.

It is highly plausible that cell infection by soil phages and subsequent cell lysis also regulate soil microbial community diversity and composition, thereby influencing soil processes, in a manner similar to the one described for marine ecosystems. The heterogeneous soil environment provides diverse microhabitats for organisms (e.g., bacteria, archaea, fungi, protists, and fauna) and supports an enormous diversity of viruses. Bacteria are the predominant component of soil organisms and viruses associated with them (*Bacteriophages/phages*) are generally the most abundant group of the virosphere. As metagenomic analyzes are being refined (i.e., *total soil metagenomes* and soil viromes), they are beginning to provide deeper insights into soil viral community composition and diversity, and their roles in regulating host populations and evolution dynamics and nutrient cycles (Sommers et al., 2021). Recent viromic studies demonstrate the potential ecological importance of soil phages in impacting soil carbon cycling process. Soil phages may also influence plant nutrient acquisition and act as bio-control agents of bacterial pathogens in agroecosystems. However, the underlying mechanisms of those processes require further exploration.

In this chapter, we revisit phage morphological and genomic diversity and their potential life strategies, summarize the current major state-of-the-art methods for studying the diversity and functions of soil phages, and outline current knowledge of viral contribution to soil element cycling and their potential roles in plant growth and health. In the context of global change, a large body of studies investigating climate warming, extreme weather, acidification, flooding or other changes, are revealing the effects on soil microbial communities and their feedback responses (Eisenhauer et al., 2012; Dubey et al., 2019). Here we further discuss the responses of soil viral diversity and function to global changes and highlight the theoretical and practical importance to incorporate soil phages into global biogeochemical models.

Diversity and life strategies of phages

Morphological observations with transmission electron micrograph classify phages into several types according to their capsids (protein shell), including polyhedral (e.g., *Microviridae*, *Corticoviridae*), pleomorphic (e.g., *Plasmaviridae*), filamentous (e.g., *Inoviridae*), and tailed particles (e.g., *Caudovirales*). Among those, the *Caudovirales* order of tailed phages includes *Siphoviridae*, *Myoviridae*, *Podoviridae*, *Ackermannviridae*, and *Herelleviridae*, and represents the most isolated and characterized phage group. Phages have various genome types, including double-stranded DNA (dsDNA), single-stranded DNA (ssDNA), single-stranded RNA (ssRNA), and double-stranded RNA (dsRNA) (Sommers et al., 2021). The dsDNA genomes of tailed phages account for the largest proportion of phage genomes in the public databases. Traditionally, the International Committee on Taxonomy of Viruses (ICTV) classified phages according to their genome structure, morphology diversity, and host range. This method, however, can only characterize a limited number of phage taxa as the vast majority of phages cannot be cultivated under laboratory conditions. Recent application of culture-independent metagenomic approaches revealed a more comprehensive picture of the virosphere and identified an extremely high viral diversity from natural environments. For example, an analysis of ~5 Tb metagenomic sequence data from more than 3000 global samples revealed over 125,000 partial DNA viral genomes, which led to a 16-fold expansion in the number of known viral genes (Paez-Espino et al., 2016). Viral genomes include a great proportion of unclassified and uncovered viruses, comprising >90% in some samples. It remains a great challenge for the ICTV to establish guidelines for robust classification of rapidly emerging unknown phages.

To identify uncultivated phages, which can provide unprecedented insights into the virosphere, the ICTV has included genomic features as a fundamental component of taxonomic classification. A 15-rank classification hierarchy has been developed to explore viral macroevolution and establish evolutionary relationships among distantly related viruses (International Committee on Taxonomy of Viruses Executive Committee, 2020). Virus taxonomy keeps dynamic to accommodate new discoveries of viral lineages (Dutilh et al., 2021). Some studies have clustered viral genome sequences into unique

viral operational taxonomic units (vOTU), based on genome similarity, to capture evolutionarily and ecologically cohesive units and assess viral genomic diversity. This classification, however, may sometimes conflict with ICTV-ratified taxa. Studies based on the vOTU analysis have tried to link viral diversity parameters with different biomes, locations, niches, or abiotic and biotic factors to reveal how viruses interact with their surrounding environments, enabling us to identify the key environmental drivers of phage community and distribution patterns. For example, soil viral communities showed temporal dynamics and were significantly correlated to soil properties (e.g., pH, soil organic matter, soil nutrition) in agricultural soils (Santos-Medellin et al., 2021).

Soil phages have diverse replication strategies, typically including lytic, lysogenic, and chronic. During lytic infection, hosts are lysed by virulent phages with virus progeny and cellular lysates being released into the environment. Phages that only initiate lytic infection are termed “virulent phages”. In lysogenic infections, phages integrate their genomes into host chromosome or maintain them extrachromosomally. They can replicate without lysing hosts and no viral progeny is released outside of host cells. Temperate phages are able to enter into either a lysogenic or lytic lifestyle, and the switch is influenced by many factors, including hosts and phages (e.g., gene, host density, and superinfection), the environment (e.g., temperature, nutrient levels, salinity, and other environmental stresses) and other external signals (e.g., quorum sensing signaling molecules) (Correa et al., 2021). Specifically, according to the “kill the winner model” (Thingstad, 2000), lytic life cycles are generally favored when host abundance is high, whereas lysogenic infections are more common when host densities are low, enabling survival in an unfavorable environment. This view, however, is challenged by the finding that weak or even negative correlations are observed between the frequency of phages and host density in some habitats, such as coral reefs and mammalian gut, which supported the proposal of the “Piggyback-the-Winner” model (Knowles et al., 2016).

Phages can also initiate a chronic life cycle, during which progeny virus particles are released into the environment from host cells without killing their hosts. This life cycle has been mainly identified for the *Inoviridae* family which is detected widely in aquatic, terrestrial and host ecosystems, indicating the widespread environmental prevalence of chronic life cycle. In addition, other lifestyles also exist, such as “pseudolysogeny”, during which phages remain in the host cells as an extrachromosomal element (Chevallereau et al., 2022). The pseudolysogeny lifestyle may be associated with host starvation and phages are unable to replicate due to the low energy state of hosts. When environmental conditions favor host cells, productive cycles of phages may develop. However, which infection type of phages prevails in soil remains unclear, and it is important to understand phage infection strategies for prediction of their interactions with hosts and ecological roles. As soil is a heterogeneous environment and soil microorganisms and phages may experience environmental fluctuations over a short period, phage-host interaction in situ may be extremely complicated and dynamic. Phage lifestyle is traditionally determined by laboratory experiments, but this can be labor-intensive and seems impossible for the continuous discovery of newly uncultivated viruses. Some computational methods have been developed to predict phage lifestyles, e.g., PHACTS, PropagATe and BACPHLIP. The accuracy and universality of those methods awaits to be verified with more environmental sequence data and experiments.

Advances in methodologies to study soil viruses

Metagenomic methods (e.g., total soil metagenomes and viromes) have been increasingly used as an instrumental approach to study soil viral ecology. The total soil metagenomic method directly obtains metagenomic sequences from free phage particles, soil organisms (bacteria, archaea, fungi, and other eukaryotes), and lysogenic-state prophage. It can link the potential performance of phages with their host community composition and metabolic functions. In particular, DNA stable isotope probing (SIP) technique can be combined with metagenomic approach to identify active linkages of phages and hosts from the individual to population level. For example, a recent DNA-SIP-metagenomic study linked phages to a critical biogeochemical process of methane oxidation, and revealed active interactions of phages with hosts at different trophic levels including both methylotrophs and non-methanotrophic methylotrophs (Lee et al., 2021). H_2^{18}O SIP-metagenomics were used to track all viral activities and virus-host dynamics, and identify their ecological roles in frozen northern peat soils (Trubl et al., 2021). This method requires expensive isotopically labeled substrate to ensure sufficient labeled nucleic acids as the failure rate of labeling viral metagenomes may be relatively high. As a promising tracing method, it could be widely applied in the future to decipher the mechanisms of the roles of soil phages in targeted processes (e.g., nitrogen fixation and organic matter degradation).

The virome method requires a number of steps (e.g., size fractionation and centrifugation) before DNA extraction to eliminate cellular contaminants and ideally extract sequences from virus like particles (VLPs). Compared with the soil total metagenomes method, viromes recover a higher proportion of viral data (especially the rare virosphere) and reduce the difficulty of assembling viral sequences (Santos-Medellin et al., 2021). It is increasingly becoming an important method to investigate phage communities and their involvement in ecosystem functions. Soil virome procedures include the following major steps (Fig. 1) (Bi et al., 2022). Soil samples (~50–500 g) are mechanically or artificially mixed with elution media to release free phages from the soil. Elution media containing VLPs are concentrated using tangential flow filtration (TFF), polyethylene glycol concentration or centrifugal ultrafiltration. Some studies add a CsCl gradient ultracentrifugation purification step after concentration. Samples are then treated with DNase to remove free cellular DNA before extracting nucleic acids of phages. Amplification of viral nucleic acids is usually

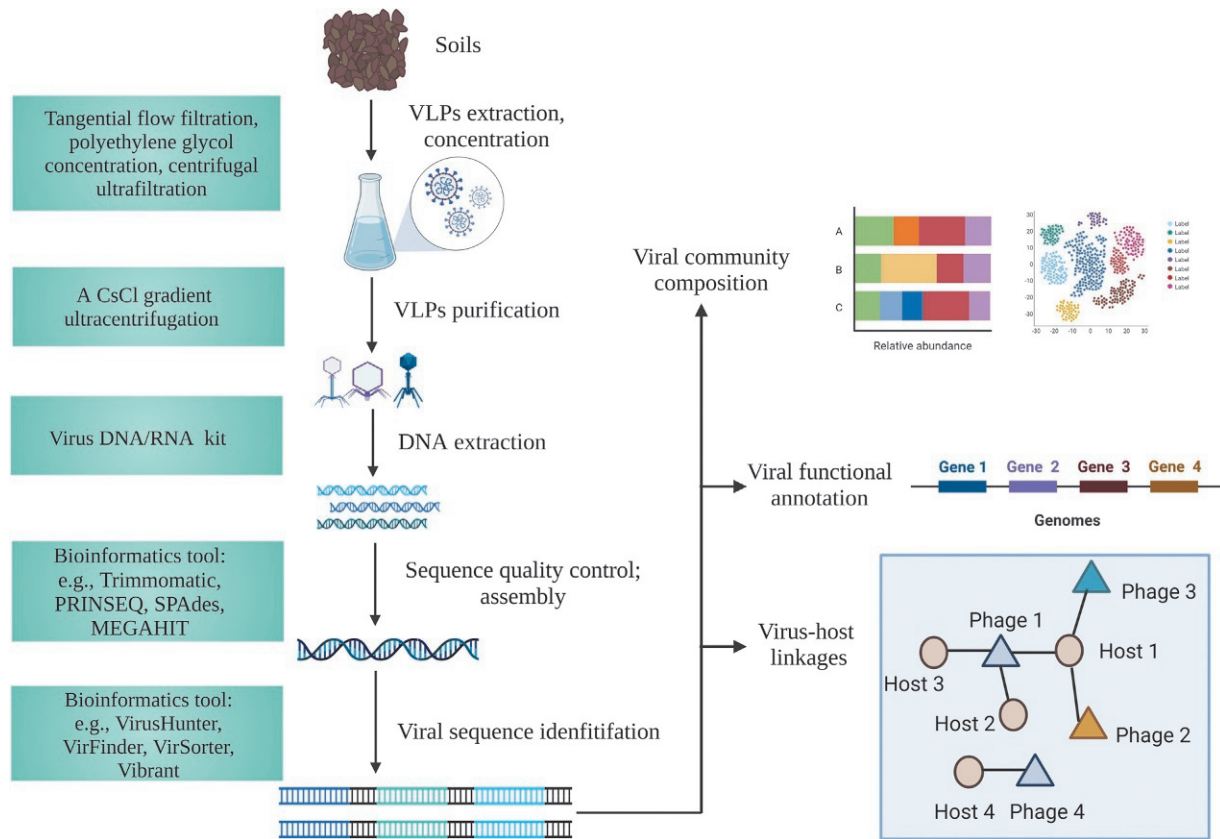


Fig. 1 Soil virome experiments and analysis workflow. Soil virome experimental procedures generally include extracting and concentrating VLPs from soil suspensions, VLPs purification, and DNA extraction; viral metagenomic analysis mainly includes sequence quality control and assembly, viral consequence prediction, viral community composition and function annotation, and virus-host prediction. Specific methods and examples of computational tools are shown in green box.

required for metagenomic sequencing as viral genomes are small and DNA yield sometimes is extremely low. Optimized extraction protocols have been recently developed to extract dsDNA phages from agricultural soils (Goller et al., 2020). However, some issues of soil virome method remain: (i) the normal use of 0.2 μm threshold to remove large soil particles or prokaryotic and eukaryotic cells could deplete giant viruses, which are important components of virosphere; (ii) potential strong absorption of viruses to soil matrix hinders us from describing the full content of soil virosphere; and (iii) there is no standard procedure for bioinformatic analysis for phage identification, phage abundance, and gene function prediction.

The new single-virus genomics (SVG) method can characterize individual phage genetic information and genomic *microdiversity* within phage populations (Martínez et al., 2020). This method requires small amounts of samples. The procedures generally include collecting phages from samples, fluorescently staining dsDNA phages, separating or sorting individual phage particles based on fluorescent signals, lysing virion capsids, extracting and amplifying phage DNA, as well as sequencing and analyzing metagenome data. This approach can reconstruct and uncover genetic heterogeneity in phage population and potentially assembles genomes from variant-rich sample, which may be discarded during the assembly process of the metagenomic methods. However, SVG has poor detection of ssDNA and RNA viruses and has a relatively high cost. This method is now commonly used for marine samples but has been used in only a few soil studies, as the complex soil matrix leads to the difficulties with the fluorescent stain and VLPs separation. A representative example is the use of fluorescence activated cell sorting to separate potential nucleocytoplasmic large DNA viruses (NCLDVs) from forest soil. Combined with mini-metagenomics, it greatly expanded the phylogenetic diversity of NCLDVs (Schulz et al., 2018). Other sequencing technology—long-read sequencing e.g., Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), can yield large assembly-free genomes and recover more complete viral genomes. The long-read sequencing combined with Illumina short sequencing has been used to reveal agricultural slurry viromes (Cook et al., 2021). That study reported a temporal-stable and lytic-dominated slurry phage community, which encodes abundant AMGs including virulence-associated genes and antimicrobial resistance genes. In addition, some other tools are capable of describing phage-host interactions and viral dynamics and overcoming some limitations in metagenomic methods. For example,

bioorthogonal noncanonical amino acid tagging (BONCAT) combined with metagenomics, nanoscale secondary ion mass spectroscopy (NanoSIMS), single-cell colony, PhageFISH, and Virocell-FISH can be used to better characterize soil virosphere and viral ecology (Sommers et al., 2021).

Potential roles of soil phages in the terrestrial ecosystem

Contribution of soil phages to biogeochemical element cycles

Phages can impact biogeochemical element cycling via top-down control through lysing microbial hosts. This process can transfer microbial biomass into dead particulate organic matter and dissolved organic matter (*viral shunt*) and impact microbes that drive biogeochemical cycles. Phages can also impact biogeochemical element cycling via bottom-up control through carrying and expressing AMGs and modulating host metabolisms (Trubl et al., 2018). Some pioneer studies investigated the roles of soil phages in carbon cycling processes. Viral sequences were retrieved from soil metagenomes, size-fractionated metagenomes, and viromes of permafrost peatlands (Emerson et al., 2018). Virus hosts were predicted in silico and a certain proportion of predicted hosts were linked to key carbon-cycling microorganisms related to methane production and oxidation and organic matter degradation. Divergent AMGs associated with central carbon metabolism, polysaccharide binding and soil organic matter degradation were also identified in the viral genomes. Among those AMGs, active function of one viral glycoside hydrolase was further expressed and verified with enzymatic assays (Emerson et al., 2018). Virus-contained AMGs related to carbon cycling are also observed in other soil habitats (e.g., agricultural soils and mangrove soils) (Jin et al., 2019; Bi et al., 2021), suggesting that viral involvement in carbon cycling process may be prevalent in the terrestrial environment.

Soil phages can also influence other nutrient cycling processes through lysing functional microbes and encoding AMGs related to sulfur, phosphorus or nitrogen cycling. The underlying mechanisms of why and how phages encode host-derived genes remain largely unknown. The well-studied AMGs include *psbA* and *psbD*, encoding core photosynthetic reaction centers, which are carried by cyanophages in the marine environment. Expression of *psbA* and *psbD* genes during infection process can promote the photosynthetic activity of hosts to generate energy for viral production and increase viral abundance. Some researchers further proposed that AMGs may maximize soil phage reproductivity by improving host fitness under temperate conditions or increase the chance of phage reproductivity by aiding their host survival under harsh conditions (Hwang et al., 2021). Despite the ecological importance of AMGs, there are great knowledge gaps in the diversity and proportion of AMGs carried by soil phages and under what conditions they are activated and expressed in microbial hosts.

Potential roles of soil phages in plant nutrient uptake and adaption

Plants provide various niches (e.g., rhizosphere, endosphere, and phyllosphere) for a wide spectrum of microorganisms. Some eukaryotic viruses (emerging as plant pathogens) are harmful to plant health and may cause crop diseases and yield decline. Phages mainly infect bacteria and modulate soil- and plant-associated bacterial community assemblies and turnover, thereby they can be beneficial, neutral or pathogenic to plants. Potentially, beneficial phages may increase plant nutrient uptake, mitigate environmental stress, and protect plants from pathogens via direct and indirect impacts on microorganisms within the soil food webs (Fig. 2). For example, phage-induced release of available carbon and other nutrients from bacterial cells is prevalent in the rhizosphere, which benefits plant nutrient acquisition (Fig. 2). However, the phage-induced mortality of beneficial bacteria that have plant growth promotion functions would impede plant nutrient acquisition. For example, phage lysis of *Rhizobium* bacteria and phosphate solubilizing microorganisms can negatively impact beneficial microbial functions and decrease plant nitrogen and phosphate acquisition. The *phyllosphere*, as the aboveground plant part, is subjected to environmental stresses (e.g., temperature change, drought, UV radiation, and nutrient deficiency). In such a harsh environment, phyllosphere bacteria may tend to be infected by temperate phages with lysogenic cycle and establish mutualistic host-phage relationships. Prophages may encode tolerance featured AMGs (such as DNA-repair proteins against radiation) or other AMGs supporting host metabolisms. The presence of those phages potentially improves the resistance of phyllosphere bacteria to the changing environment and promotes host survival under harsh conditions. Contradictory circumstances also remain that phage-encoded virulence factors (e.g., toxin, invasion and motility) can increase pathogenic bacteria infection of plants and have the potential to spread the virulence genes via horizontal gene transfer.

In addition to direct effects on plants, phage-host interactions indirectly influence plant growth. For example, the lysis of competitors of plant growth-promoting bacteria (e.g., inefficient soil rhizobia) may promote niche-competitiveness and early colonization by beneficial bacteria (e.g., efficient soil rhizobia), thereby facilitating later plant nitrogen acquisition (Fig. 2). Some phages can infect bacteria with lysogenic lifestyle and protect hosts against subsequent phage infection, which to some extent improves bacterial competitive fitness compared with other bacteria. As soil bacteria can evolve to escape viral infection and phages counter-adapt to this resistance, their co-evolutionary dynamics drive bacterial phenotypic and genotypic variation. The change in microbial community structure might have cascading effects on other microbiomes within the soil food web. In addition, root exudates may help the formation of mutualistic plant-virus association to increase plant fitness for nutrition acquisition or disease suppression. To date, however, plant-bacteria-phage interactions have been largely overlooked. Future studies are greatly needed to decipher the linkages among plant, bacteria, and phages and to reveal the roles of soil phages in plant-soil systems, which can promote the application of viruses for improving plant health and productivity (Pratama et al., 2020).

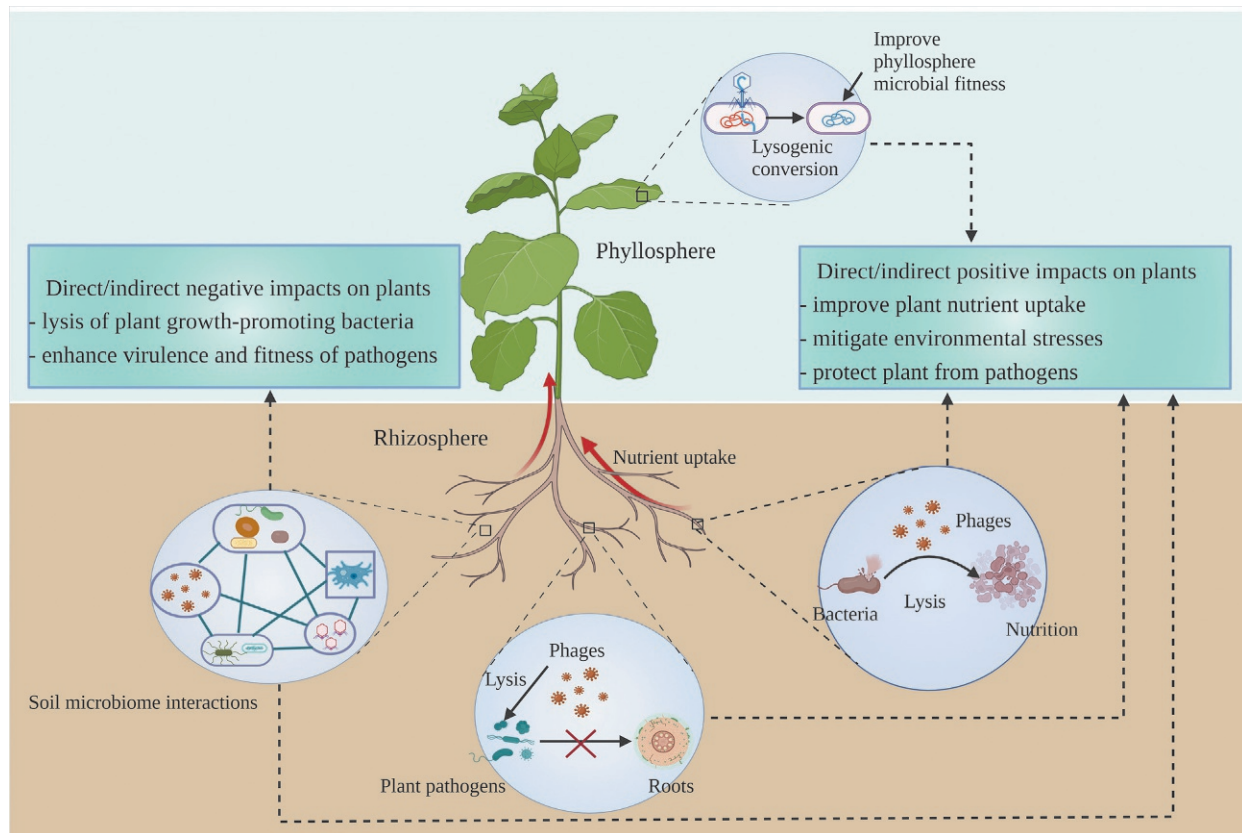


Fig. 2 Phages directly and indirectly influence plant growth and health. Phages in the rhizosphere soil potentially promote plant nutrient uptake via “viral shunt”; beneficial phyllosphere bacteria may be infected by prophages which can improve host fitness, thereby improving plant adaption to environmental stresses. Phages as an important biocontrol agent can protect plant by infecting and lysing bacterial pathogens. The phage-host interactions may have negative or positive impacts on plant growth by modulating bacterial competition and their functional performance within the food web.

Potential roles of phages in regulating crop health

Bacterial pathogens are known to cause great crop yield losses globally, representing a major threat to crop health and productivity. To reduce economic loss and ensure food security, multiple agricultural management practices (e.g., application of copper-based pesticides, antibiotics) have been widely applied to protect crops against bacterial pathogens. However, some bacterial pathogens have evolved to resist to the widely used agrochemicals and the agrochemical-based methods have unintended environmental and human health risks. Environmentally friendly control strategies are urgently required to ensure sustainable agricultural production. Phages can selectively infect pathogenic strains and propagate along with their hosts, and thus they are regarded as a potential biosafety and biosecurity biological control agent, compared with chemical options. Phage therapy (generally termed when used in human and animal protection) or phage biocontrol (generally applied term, when used in plants protection) was firstly proposed and applied in the early 20th century, but was hindered by the long-term and wide use of antibacterial agents. Most studies showed promising effects of phages on controlling pathogens, reducing disease incidence and promoting plant survival in the field or greenhouse conditions (Holtappels et al., 2021). One well-studied example is the application of phage biocontrol to combat the pathogenic bacterium *Ralstonia solanacearum* in the tomato rhizosphere (Wang et al., 2019). Four model lytic phages, belonging to the order *Caudovirales*, were isolated from four geographically distant tomato fields across China. The combination of those phages reduced pathogen densities in both greenhouse and field experiments, changed microbial community composition and diversity, and enriched bacteria antagonistic toward *R. solanacearum* in the rhizosphere.

It remains a great challenge to develop efficient biocontrol strategies in agricultural production through harnessing the potential of soil phages. Phages with an exclusive lytic (virulent) life strategy are more suitable than temperate phages in biocontrol, because temperate phages may replicate as part of host genomes without lysing bacteria, improve pathogenic host fitness and virulence, and even spread virulence genes to benign bacteria. In addition, it should be noted that *superinfection exclusion* induced by phages during their infection can prevent secondary infection of hosts by other phages. An ideal phage for biocontrol would have a relatively wide host range, which allows them to infect most targeted pathogen genera or species. Alternatively, a complementary set of isolated phages (i.e., phage mixture, cocktails) can be used to improve the efficiency of eliminating bacterial pathogens. Despite great potential of phage biocontrol in sustainable agriculture, development of efficient phage-based biocontrol products under fluctuating environmental conditions is still in its infancy.

Feedback responses of soil phages to global changes

Global changes, including warming, soil acidification, extreme drought, heavy precipitation and flooding, and land-use change, present grand challenges to the sustainable provisions of ecosystem functions and services. In particular, global temperature is predicted to rise continuously in coming decades and will lead to a significant increase in the frequency and intensity of extreme weather events. Many studies have reported that global changes influence the diversity and composition of soil microbial communities, with important consequences for their mediated functions. For example, microbial metabolic rates normally increase with temperature, and climate warming can alter soil microbial community compositions and functional genes in various terrestrial habitats (e.g., glacier forefield, forest and grassland) (Jansson and Hofmockel, 2020; Wang et al., 2020). A global meta-analysis based on 1235 observations found that global change factors have inconsistent impacts on microbial diversity, with rare microbial species being more vulnerable to global changes than common species (Zhou et al., 2020). As phages have intimate relationships with bacterial hosts, they are inevitably subjected to the changes in these abiotic factors. Variations in temperature, soil moisture, local geochemical properties, and host community can affect the infection, lifestyle, abundance, community composition, and functions of soil phages as well as their interactions with hosts. Among those global change factors, effects of temperature on phage infection have been well documented in laboratory and field experiments. Before reaching a phage-specific thermal optimum (T_{opt}), temperature increase is predicted to shorten the latency period and enhance the release of phage progeny, but a reversal trend is observed beyond the T_{opt} (Wieczynski et al., 2021). Temperature change (e.g., a slight increase) might induce changes in viral life strategies, switching from a lysis to chronic lifestyle which supports the coexistence of phages and hosts in unfavorable situations. Seasonal dynamics of phage abundance have been observed in agricultural and forest soils (Roy et al., 2020). However, in general, patterns of temperature-driven shifts in phage infections, lifestyles and abundance are complex and remain to be elucidated in natural environments, especially under the scenarios of combined biotic and abiotic factors.

To determine the impacts of warming on soil viral community composition, studies have sampled soils from simulated warming experiments established in a boreal peatland and found that viral community composition was significantly affected by soil depth, water content, and concentrations of methane and carbon dioxide, but not by temperature (Ter Horst et al., 2021). This insignificant effect of temperature on soil viruses might be due to the short time of only two-years for the warming treatment. Global warming results in the thawing of permafrost soils, leading to a shift of “soil-like” viral populations in drier soils to ‘aquatic-like’ viral populations in wetter soils. Viral communities are significantly related to host community composition, pH, soil moisture content, and soil depth, indicating that global change factors such as warming may alter viral populations through modulating soil properties and host communities (Schulz et al., 2018). Soil pH can modify the adsorption of phages to soil organic materials and minerals and influence the survival and infectivity of some phages. As soil pH is an important factor driving soil phage community structure (Bi et al., 2021; Lee et al., 2022), global soil acidification due to anthropogenic activities may profoundly alter viral communities. Nevertheless, we still lack knowledge of how global change factors influence the ecological functions of soil phages, with unknown consequences for soil biogeochemistry and microbiome evolution.

Soil phages can also directly and indirectly respond to global changes. Viral shunt induced by viruses drives carbon flow, increases carbon and nutrient bioavailability, and enhances surrounding microbial respiration, which may facilitate the release of CO_2 from soil to atmosphere. In particular, phage-induced mortality of autotrophs stimulates the growth of a myriad of heterotrophic bacteria that utilize soil organic matter and release CO_2 , thereby contributing to greenhouse gas emissions and exacerbating climate warming (Fig. 3). However, the net flux of CO_2 induced by soil phages has not been measured. Soil phages also

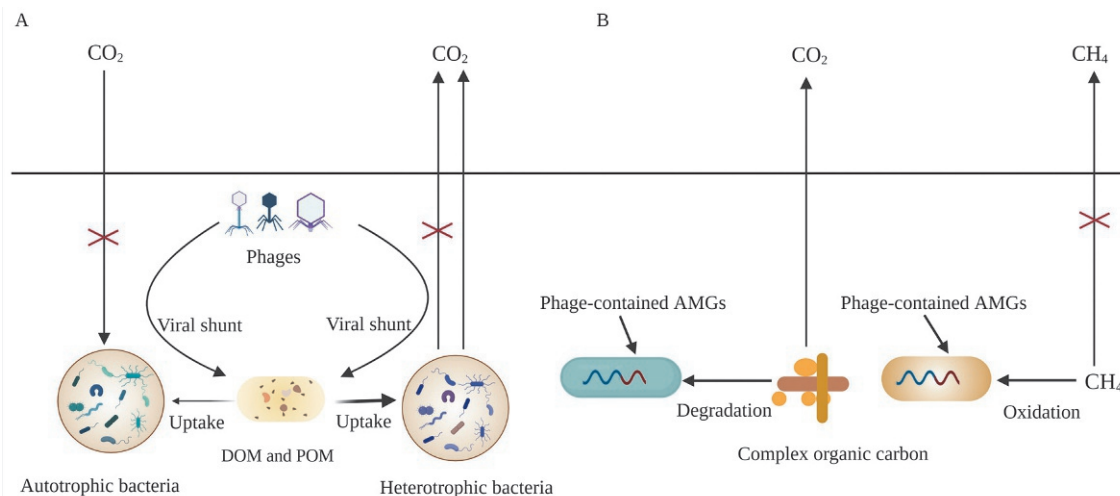


Fig. 3 Examples of the responses of soil phages to climate change and potential consequences. (A) Phage-induced mortality of soil autotrophs may stimulate the growth of heterotrophic bacteria that utilize soil organic matter and release CO_2 , thereby exacerbating greenhouse gas emissions. (B) Soil phages can also modulate soil organic matter degradation and CH_4 oxidation via expressing related AMGs, which influence the CO_2 and CH_4 emissions from soil to atmosphere.

indirectly mediate microbial metabolic response to global warming and enhance host adaption to extreme environmental conditions. For example, phage-contained AMGs related to soil organic cycling can help hosts to mineralize soil organic matter, while AMGs involved in CH₄ oxidation can decrease CH₄ flux from soil to atmosphere, indicating the dichotomous role of phages in regulating greenhouse gas emissions and influencing global warming. Moreover, viral abundances might be a significant predictor of CH₄ dynamics in some soil habitats. The $\delta^{13}\text{C}$ of porewater CH₄ in thawing permafrost soil was significantly related to the abundance of phages of methanogens and methanotrophs, rather than the single index of abiotic factors or methanogen and methanotroph abundances. At present, how phages mediate soil microbial responses to warming at different trophic levels and whether soil phages mitigate or exacerbate climate change are poorly understood. Future studies are required to incorporate phage feedbacks and contribution to global biogeochemical models, which will enable a comprehensive understanding and prediction of soil microbial responses to global changes.

Conclusions and outlook

Our knowledge of the community diversity and ecological roles of soil phages and their interactions with hosts has been advanced in recent decades, due to the rapid development in sequencing technologies, molecular-omics, SIP techniques, and new bioinformatic tools. In this chapter, we describe several important aspects of viral potential functions in soil ecosystems, including the contribution of soil phages to global biogeochemical element cycling and their potential application in regulating plant growth and health. Some critical and specific questions remain to be addressed in the study of soil viruses, including (i) how phage lifestyles (lytic, lysogeny or chronic) impact soil microbial communities and how the shift in life strategies of phages changes their ecological roles in soil ecosystems; (ii) how to reduce biases in laboratory experiments and bioinformatic analysis to comprehensively illustrate the virosphere; (iii) how to measure quantitatively the proportion and biomass of hosts undergoing phage lysis in the soil environment; and (iv) to what extent phages improve plant nutrition and protect plants from bacterial pathogen infection and how they can be used in future agricultural production and application. In addition, global change is increasingly impacting terrestrial ecosystems and changing soil microbial communities and functions, but their influences on phage diversity and functions are largely unknown yet deserved exploration. Collectively, we envision those proposed studies will unlock viral roles in various soil ecosystems and incorporate them into global biogeochemical configurations.

References

- Bi L, Yu DT, Du S, et al. (2021) Diversity and potential biogeochemical impacts of viruses in bulk and rhizosphere soils. *Environmental Microbiology* 23: 588–599.
- Bi L, Yu D, Han L, et al. (2022) Unravelling the ecological complexity of soil viromes: Challenges and opportunities. *Science of the Total Environment* 812: 152217.
- Breitbart M, Bonnain C, Malki K, and Sawaya NA (2018) Phage puppet masters of the marine microbial realm. *Nature Microbiology* 3: 754–766.
- Chevallereau A, Pons BJ, van Houte S, and Westra ER (2022) Interactions between bacterial and phage communities in natural environments. *Nature Reviews Microbiology* 20: 49–62.
- Cook R, Hooton S, Trivedi U, et al. (2021) Hybrid assembly of an agricultural slurry virome reveals a diverse and stable community with the potential to alter the metabolism and virulence of veterinary pathogens. *Microbiome* 9: 65.
- Correa AM, Howard-Varona C, Coy SR, et al. (2021) Revisiting the rules of life for viruses of microorganisms. *Nature Reviews Microbiology* 19: 501–513.
- Dubey A, Malla MA, Khan F, et al. (2019) Soil microbiome: A key player for conservation of soil health under changing climate. *Biodiversity and Conservation* 28: 2405–2429.
- Dutilh BE, Varsani A, Tong Y, et al. (2021) Perspective on taxonomic classification of uncultivated viruses. *Current Opinion in Virology* 51: 207–215.
- Eisenhauer N, Cesarz S, Koller R, et al. (2012) Global change belowground: Impacts of elevated CO₂, nitrogen, and summer drought on soil food webs and biodiversity. *Global Change Biology* 18: 435–447.
- Emerson JB, Roux S, Brum JR, et al. (2018) Host-linked soil viral ecology along a permafrost thaw gradient. *Nature Microbiology* 3: 870–880.
- Goller PC, Haro-Moreno JM, Rodríguez-Valera F, et al. (2020) Uncovering a hidden diversity: Optimized protocols for the extraction of dsDNA bacteriophages from soil. *Microbiome* 8: 17.
- Holtappels D, Fortuna K, Lavigne R, and Wagemans J (2021) The future of phage biocontrol in integrated plant protection for sustainable crop production. *Current Opinion in Biotechnology* 68: 60–71.
- Hurwitz BL and U'Ren JM (2016) Viral metabolic reprogramming in marine ecosystems. *Current Opinion in Microbiology* 31: 161–168.
- Hwang Y, Rahlff J, Schulze-Makuch D, et al. (2021) Diverse viruses carrying genes for microbial extremotolerance in the Atacama Desert hyperarid soil. *mSystems* 6: e00385-21.
- International Committee on Taxonomy of Viruses Executive Committee (2020) The new scope of virus taxonomy: Partitioning the virosphere into 15 hierarchical ranks. *Nature Microbiology* 5: 668–674.
- Jansson JK and Hofmockel KS (2020) Soil microbiomes and climate change. *Nature Reviews Microbiology* 18: 35–46.
- Jin M, Guo X, Zhang R, et al. (2019) Diversities and potential biogeochemical impacts of mangrove soil viruses. *Microbiome* 7: 58.
- Knowles B, Silveira C, Bailey B, et al. (2016) Lytic to temperate switching of viral communities. *Nature* 531: 466–470.
- Lee S, Sieradzki ET, Nicolas AM, et al. (2021) Methane-derived carbon flows into host–virus networks at different trophic levels in soil. *Proceedings of the National Academy of Sciences of the United States of America* 118: e2105124118.
- Lee S, Sorensen JW, Walker RL, et al. (2022) Soil pH influences the structure of virus communities at local and global scales. *Soil Biology and Biochemistry* 166: 108569.
- Martínez JM, Martínez-Hernández F, and Martínez-García M (2020) Single-virus genomics and beyond. *Nature Reviews Microbiology* 18: 705–716.
- Paez-Espino D, Eloe-Fadrosch EA, Pavlopoulos GA, et al. (2016) Uncovering Earth's virome. *Nature* 536: 425–430.
- Pratama AA, Terpstra J, de Oliveria ALM, and Salles JF (2020) The role of rhizosphere bacteriophages in plant health. *Trends in Microbiology* 28: 709–718.
- Roy K, Ghosh D, DeBruyn JM, et al. (2020) Temporal dynamics of soil virus and bacterial populations in agricultural and early plant successional soils. *Frontiers in Microbiology* 11: 1494.
- Santos-Medellín C, Zinke LA, ter Horst AM, et al. (2021) Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *ISME Journal* 15: 1956–1970.
- Schulz F, Alteio L, Goudeau D, et al. (2018) Hidden diversity of soil giant viruses. *Nature Communications* 9: 4881.
- Sommers P, Chatterjee A, Varsani A, and Trubl G (2021) Integrating Viral Metagenomics into an Ecological Framework. *Annual Review of Virology* 8: 133–158.

- Ter Horst AM, Santos-Medellin C, Sorensen JW, et al. (2021) Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome* 9: 233.
- Thingstad TF (2000) Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnology and Oceanography* 45: 1320–1328.
- Trubl G, Jang HB, Roux S, et al. (2018) Soil Viruses are underexplored players in ecosystem carbon processing. *mSystems* 3: e00076-18.
- Trubl G, Kimbrel JA, Lique-Gonzalez J, et al. (2021) Active virus-host interactions at sub-freezing temperatures in Arctic peat soil. *Microbiome* 9: 208.
- Wang X, Wei Z, Yang K, et al. (2019) Phage combination therapies for bacterial wilt disease in tomato. *Nature Biotechnology* 37: 1513–1520.
- Wang Y, Ma A, Liu G, et al. (2020) Potential feedback mediated by soil microbiome response to warming in a glacier forefield. *Global Change Biology* 26: 697–708.
- Wieczynski DJ, Yoshimura KM, Denison ER et al. (2021) Viral infections mediate microbial food web controls on the global carbon cycle under warming. *Authorea Preprints*.
- Zhou Z, Wang C, and Luo Y (2020) Meta-analysis of the impacts of global change factors on soil microbial diversity and functionality. *Nature Communications* 11: 3072.

Archaea—Soil biology

Mikhail V Semenov, Dokuchaev Soil Science Institute, Laboratory of Soil Carbon and Microbial Ecology, Moscow, Russian Federation

© 2023 Elsevier Ltd. All rights reserved.

This is an update of J.E.T. McLain, ARCHAEA, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 88–94, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00525-7>.

Introduction	25
The archaeal cell	26
Taxonomy and ecology	27
Methanogens	27
Ammonia oxidizers	28
Biotechnological potential of archaea	28
Summary	29
Acknowledgments	29
References	29
Further reading	30

Abstract

Since the discovery of the Archaea as the third domain of life along with prokaryotes and eukaryotes, much has changed in terms of understanding their ecology and distribution. New molecular methodologies have revealed the wide distribution of Archaea in soils and their crucial roles in global carbon and nitrogen cycles, including influencing greenhouse gas emissions. This chapter examines the taxonomy and ecology of soil archaea, as well as some aspects of their biotechnological potential.

Glossary

Ammonia oxidation A fundamental core process in the global biogeochemical nitrogen cycle. Oxidation of ammonia (NH_3) to nitrite (NO_2^-) is the first and rate-limiting step in nitrification.

***amoA* gene** a gene that encodes subunit A of the enzyme ammonia monooxygenase (AMO), responsible for autotrophic ammonia oxidation. *amoA* has become a widely used functional and phylogenetic marker gene for bacterial and archaeal ammonia oxidizers.

Lithoautotrophs organisms which derive energy from reactions of reduced compounds of mineral (inorganic) origin.

***mcrA* gene** a gene that encodes methyl-coenzyme M reductase (MCR), that catalyzes the final methane-forming step of methanogenesis in methanogens, and the initial methane activation step in the anaerobic oxidation of methane (AOM) in anaerobic methanotrophic archaea (ANME).

Methanogenesis an anaerobic respiration exclusive to archaea that generates methane as the final product of metabolism.

Key points

- Archaea play crucial roles in global carbon and nitrogen cycles, including influencing greenhouse gas emissions
- Archaea are ubiquitous in soils and are often one of the dominant phyla in the soil microbial community
- A single gram of soil can harbor tens or hundreds of millions of archaeal cells
- Methanogenesis is a unique terminal energy-yielding process exclusive to methanogenic archaea
- Ammonia-oxidizing archaea are the major drivers of nitrification in many soils

Introduction

Since the discovery of the Archaea as the third domain of life along with prokaryotes and eukaryotes (Woese et al., 1990), much has changed in terms of understanding their ecology and distribution. Archaea were initially considered to be restricted to environments with extreme conditions such as high levels of salinity and high temperature, and environments in which they evolved to metabolize organics, such as methane, and nitrogen. However, new molecular methodologies have revealed the presence of

Archaea in a wide variety of habitats including oceans, sediments, and soils. Based on cell counts and molecular studies, Archaea typically account for 10^7 – 10^8 cells per gram of soil and 1–15% of the soil prokaryotes (Bates et al., 2011; Semenov et al., 2016, 2018).

Archaea have evolved a variety of energy metabolisms using organic or inorganic electron donors and acceptors (including fixing carbon from inorganic sources) and thus play crucial roles in global carbon and nitrogen cycles, including influencing greenhouse gas emissions (Offre et al., 2013). For example, methanogenesis and anaerobic methane oxidation are important steps in the carbon cycle that are performed by anaerobic Archaea (Liu and Whitman, 2008; Shen et al., 2019). Although both Archaea and Bacteria contribute to the globally important process of aerobic ammonia oxidation, the wide distribution of ammonia-oxidizing Archaea in virtually all investigated aerobic soils indicates a prominent role for these organisms (Leininger et al., 2006; Alves et al., 2018).

Archaeal communities and their potential roles in soil are affected by many edaphic and environmental factors. pH, carbon and nitrogen contents, vegetation, and depth may all result in changes in the abundance and composition of archaeal communities in a soil environment (Tripathi et al., 2013; Stempfhuber et al., 2015). The adaptations of archaea to energy stress provide a competitive advantage compared to bacteria. Archaea can survive under extreme salinity and pH, in which bacteria and eukaryotes cannot grow. They also have advantages when faced with energy limitation, including limited substrate supply, such as for nitrification, and minimal free-energy yields from catabolism, such as for methanogenesis and anaerobic methane oxidation (Valentine, 2007).

The archaeal cell

Archaeal cells display a wide variety of morphological types. Some, similar to species of the bacterial domain, are strictly rod-shaped or spherical. Others are disks, spirals, or filaments, or exhibit amebal irregularity, with variable protuberances. Still others have a mineral-like geometry, with shapes similar to cubes or triangles. Other irregular cells are bumpy spheres (e.g., *Thermococcus*) or are vaguely rod-shaped, but of highly variable diameter (e.g., *Pyrodictum*).

Archaeal cells possess many characteristics similar to those of the Bacteria and Eukarya, and others that are unique to the archaeal domain (Table 1). With rare exceptions, Archaea contain a cell wall that, like the cell wall of bacteria, functions to prevent osmotic lysis and to define cell shape. The ability of Archaea to adapt to extreme environments is assisted by the possession of unique cell wall types, which vary from those containing molecules composed of pseudopeptidoglycan, closely resembling bacterial peptidoglycan, to cell walls completely lacking a polysaccharide component. A common wall type is a paracrystalline surface layer (S-layer), consisting of protein or glycoprotein, generally of hexagonal symmetry. Although S-layers are common to all groups of Archaea, the biochemical makeup of S-layers among species is very diverse and, in some cases, this layer is too supple to contribute to the stability of the cell. In these cases, the function of the S-layer is unknown, but it has been proposed that the space between the cytoplasmic membrane and the outer surface of the S-layer may fulfill the role of a periplasmic space reminiscent of Gram-negative bacteria.

Archaeal RNA polymerases are complex, consisting of up to 14 subunits (compared with 4 in the bacterium *Escherichia coli*). Like Bacteria, Archaea possess a single RNA polymerase, but the archaeal RNA polymerase resembles those of eukaryotes in multisubunit complexity and sequence homology. In addition, archaeal RNA polymerases are unable to initiate transcription *in vitro*, a feature also seen in eukaryotes where general transcription factors are required for initiation.

Table 1 Some cellular features of the Archaea.

Feature of cell	Most similar to domain
Morphology	Some unique
Cell wall with pseudopeptidoglycan	Resembles bacteria
Genome consisting of single circular piece of DNA	Bacteria
DNA polymerase, DNA helicase, DNA ligase	Bacteria
Protein trafficking for sugars and inorganic ions	Bacteria
Multisubunit RNA polymerase	Eukarya
Ether-linked lipids	Unique
Promoter site for transcription	Eukarya
Lack of nuclear membrane	Eukarya
Flagellar composition and assembly	Unique
Chaperonin heat-shock proteins	Bacteria

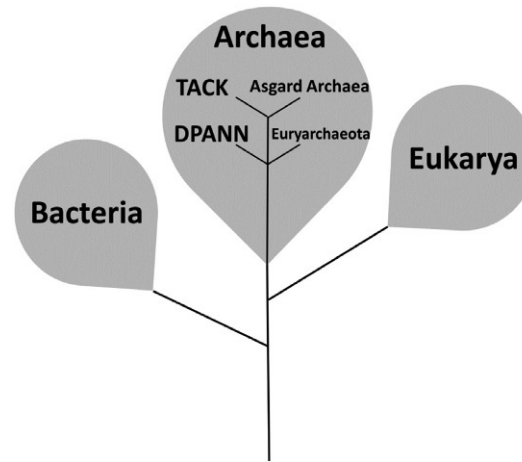


Fig. 1 Phylogenetic tree of life.

Taxonomy and ecology

Archaea are subdivided into four superphyla: (1) Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, and Nanoarchaeota (DPANN); (2) Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota (TACK); (3) Euryarchaeota; and (4) the newly discovered Asgardarchaeota (Zaremba-Niedzwiedzka et al., 2017) (Fig. 1). Thaumarchaeota and Euryarchaeota are the two most common and abundant phyla in soils. Cultivated archaea can be divided into several broad physiological (halophiles, thermophiles, and acidophiles) or metabolic (nitrifiers, methanogens, and anaerobic methane oxidizers) groups. In further subsections, the two most important ecological groups of soil archaea will be considered—methanogens (Euryarchaeota) and nitrifiers (Thaumarchaeota).

Methanogens

Methanogens are a group of phylogenetically cohesive microbes from the domain Archaea that are responsible for the production of methane (CH_4), the second most important greenhouse gas with a global warming potential 25 times greater than that of CO_2 (Bridgham et al., 2013). Methanogenesis is a unique terminal energy-yielding process exclusive to methanogenic archaea. Methanogens can grow by reducing one-carbon compounds (CO_2 , CO, methanol, methylamines, and methyl sulfides) or acetate to CH_4 through one of several methanogenesis pathways (Table 2). Regardless of the substrate used, a methyl-coenzyme M molecule is ultimately reduced to methane by the methyl-coenzyme M reductase enzyme (MCR). Methanogens are distinguished from bacteria and other archaea that may produce methane as a byproduct of metabolism by their obligate need to synthesize methane to conserve energy using the Wolfe Cycle (Buan, 2018).

Table 2 Reactions for methanogenesis by methanogenic archaea.

Substrates		Products
1. CO_2-type		
$4\text{H}_2 + \text{CO}_2$	→	$\text{CH}_4 + 2\text{H}_2\text{O}$
$4\text{H}_2 + \text{HCO}_3^- + \text{H}^+$	→	$\text{CH}_4 + 3\text{H}_2\text{O}$
$4 \text{Formate} + 4\text{H}^+$	→	$\text{CH}_4 + 3\text{CO}_2 + 2\text{H}_2\text{O}$
$4 (2\text{-Propanol}) + \text{CO}_2$	→	$\text{CH}_4 + 4 \text{acetone} + 2\text{H}_2\text{O}$
$4 \text{CO} + 2\text{H}_2\text{O}$	→	$\text{CH}_4 + 3\text{CO}_2$
2. Methylated C1 compounds and ethanol		
$2 \text{Ethanol} + \text{CO}_2$	→	$\text{CH}_4 + 2 \text{acetate} + 2\text{H}^+$
4Methanol	→	$3\text{CH}_4 + \text{CO}_2 + 2\text{H}_2\text{O}$
4Methanol	→	$3\text{CH}_4 + \text{HCO}_3^- + \text{H}^+ + \text{H}_2\text{O}$
$4 \text{Methylamine} + 2\text{H}_2\text{O}$	→	$3\text{CH}_4 + \text{CO}_2 + 4\text{NH}_4^+$
$\text{Methanol} + \text{H}_2$	→	$\text{CH}_4 + \text{H}_2\text{O}$
3. Acetate		
$\text{Acetate} + \text{H}^+$	→	$\text{CH}_4 + \text{CO}_2$
$\text{Acetate} + \text{H}_2\text{O}$	→	$\text{CH}_4 + \text{HCO}_3^-$

All methanogens characterized so far belong to the Euryarchaeota. Currently, there are seven orders of methanogens recognized based on *mcrA* and ribosomal gene phylogeny (Adam et al., 2017). Five of the methanogen orders (*Methanopyrales*, *Methanococcales*, *Methanobacteriales*, *Methanomicrobiales*, and *Methanocellales*) so far only contain hydrogenotrophic methanogens, while the *Methanomassiliicoccales* order is defined as having obligate methyl-respiring methylotrophic methanogens. The *Methanosarcinales* have many members capable of more than one methanogenesis pathway, as well as obligate hydrogenotrophic or obligate acetoclastic members.

Methanogenesis is strongly dependent on the soil water regime and organic carbon and total nitrogen contents. Low levels of soil moisture reduce the abundance of methanogens and the rate of methanogenesis (Timonen and Bomberg, 2009; Semenov et al., 2019). In contrast, waterlogging favors the development of methanogens, which makes them an important part of microbial communities of peat, gleyic, and rice paddy soils (Kharitonov et al., 2021). Temperature also affects the diversity and abundance of soil methanogenic archaea. CH₄ emissions markedly increase with rising temperatures and are linked to the transcriptional activities of methanogens (Yvon-Durocher et al., 2014).

Ammonia oxidizers

Until recently, it was assumed that only lithotrophic bacteria and partly heterotrophic microorganisms were involved in nitrification. The discovery that lithoautotrophic archaea, thriving in oxic and moderate habitats, could oxidize NH₃ to NO₂ was unexpected. Further molecular investigations have confirmed their widespread occurrence in terrestrial ecosystems and their potential roles in nitrification, one of the most important processes in the nitrogen cycle (Leininger et al., 2006). Phylogenetic analyses based on ribosomal-proteins, other core genes, and comparative genomics strongly suggested that these mesophilic archaea should be assigned to a novel archaea phylum, Thaumarchaeota (Brochier-Armanet et al., 2008; Pester et al., 2011). After this, many more non-thermophilic Group 1.1a and 1.1b Thaumarchaeota have been isolated in pure cultures and sequenced, such as the Group 1.1a *Nitrosopumilus maritimus*, *Candidatus Nitrosopumilus salaria*, *Candidatus Nitrosopumilus sediminis* and *Candidatus Nitrosopumilus koreensis*, the Group 1.1a-associated *Nitrosotalea devanattera*, and the Group 1.1b *Nitrososphaera viennensis*, *Nitrososphaera gargensis*, and *Nitrososphaera evergladensis* (Bomberg, 2016). No representative of the Group 1.1c has yet been obtained in pure culture. The 1.1c group has been found in acidic boreal forest soils, deep peats, boreal fens, shallow peat from elevated oligotrophic subarctic bogs, and tropical peat swamp forest soils (Bomberg, 2016).

Thaumarchaeota are ubiquitous in soil and are often one of the dominant phyla in the soil microbial community (Alves et al., 2018). For example, the relative abundance of Thaumarchaeota can reach up to 20–28% in chernozem (Semenov et al., 2018) or 15–20% in phaeozem (Semenov et al., 2021). Most soil Thaumarchaeota are represented by the *Nitrososphaeria* class, which are ammonium-oxidizing archaea (AOA). The diversity and ecology of AOA have been intensively studied, often in relation to that of ammonia-oxidizing bacteria (AOB), and has relied mainly on analysis of *amoA* genes, which encode subunit A of the key metabolic enzyme ammonia monooxygenase (AMO) (Alves et al., 2018). Consequently, *amoA* became the second most frequently sequenced marker gene in microbial ecology after the 16S ribosomal RNA (rRNA) gene. Analysis of the archaeal *amoA* gene provided strong evidence that pH is a key factor driving ecology and niche specialization of AOA (Nicol et al., 2008; Gubry-Rangin et al., 2011; Pester et al., 2011; Stempfhuber et al., 2015). The C:N ratio, total organic carbon and total nitrogen contents, water content, temperature and oxygen concentration are all environmental factors affecting AOA communities (Bates et al., 2011; Alves et al., 2018).

Since the abundance of archaeal *amoA* genes often exceeds that of bacterial *amoA* genes (Leininger et al., 2006), AOA are thought to be major drivers of nitrification in many soils. This is particularly evident in acidic soils where the low pH values could drive the conversion from ammonia to ammonium, decreasing substrate availability that might selectively favor AOA over AOB as the dominating ammonia-oxidizers (Zhang et al., 2012). There is also evidence that differences in affinity for ammonia may lead to differential growth of AOA and AOB, contributing to niche separation. The growth of AOB has been linked to nitrification activity following amendment with high levels of ammonium, while the growth and nitrification activity of AOA is associated with supply of ammonia at small concentration through mineralization of organic matter (Verhamme et al., 2011).

Biotechnological potential of archaea

Biotechnologically useful enzymes represent the main focus of industrial interest in the Archaea, as a result of the abilities of these microbes to function at the temperature, salinity, and pH limits of life. Heat-tolerant enzymes are currently the most investigated of all extremoenzymes because, for many reasons, performing biotechnologically related processes at higher temperatures is often advantageous. In chemical reactions involving organic solvents, decreased viscosity and increased diffusion at elevated temperatures result in higher reaction rates. In addition, performing reactions at higher temperatures reduces the possibility of complications resulting from contamination. One thermophilic compound of particular interest is DNA polymerase, an enzyme that is responsible for the elongation of the primer strand of a growing DNA molecule and is thus central to the polymerase chain reaction for DNA amplification. DNA polymerases from various hyperthermoarchaea (including *Pwo* from *Pyrococcus woesei* and *Pfu* from *P. furiosus*) are showing biotechnological promise, based on their stringent proofreading abilities and suitability for the amplification of longer DNA fragments. These hyperthermophilic DNA polymerases possess error rates that are 5- to 10-fold less than that of the widely used thermobacterial *Taq* polymerase from *Thermus aquaticus*.

Applied uses exist or have been proposed for a variety of other archaeally derived materials. The extremely stable lipids of archaeal membranes may represent a novel drug delivery system because of their enhanced stability under temperature extremes. Archaeal components such as the S-layer glycoprotein have drawn interest for their use as possible vaccine carriers and other nanotechnological potentials, and it has been shown that much stronger immune responses in mice are shown to protein antigens encapsulated in archaeosomes than in conventional liposomes. A thermostable ligase for the ligase chain reaction (an amplification method that involves the ligation of two sets of adjacent oligonucleotides) would be of obvious benefit because the ligation must be carried out near the melting temperature of the DNA, and the ligase enzymes must be stable during the dissociation step that follows. Currently, a ligase from *T. aquaticus* is used, but a more stable equivalent may be available from hyperthermophilic Archaea. Haloarchaeal polymers have been considered as raw materials for biodegradable plastics. Hydrolases from hyperthermophiles could be used in the food-processing industry to hydrolyze fats at high temperatures, reducing bacterial contamination problems. Addition of polymer-hydrolyzing extremoenzymes such as beta-glycanases from psychrophiles to detergents would allow for efficient washing in cold water. The food industry could exploit pectinases that act at lower temperatures in the processing of fruit juices or cheeses.

Harnessing methanogens is useful to produce methane from renewable carbon feedstock. Anaerobic digestion of waste to produce biogas is a highly efficient process. Biogas comprises 30–90% methane and can be combusted to generate electricity or refined and compressed to power transportation. Methanogens have also been used to reduce the biological oxygen demand of agricultural or municipal wastewater while simultaneously producing renewable methane that can be captured from anaerobic digesters and used as fuel (Buan, 2018).

Summary

Over the past 15 years, information on the isolation, characterization, description, and applications of soil Archaea has expanded dramatically. Every few years there is news about the discovery of another archaeal phylum. It is now recognized that archaea are abundant in soil and they play important roles in global biogeochemical cycles. Methanogens and sulfate-dependent anaerobic methane-oxidizers belonging to the Euryarchaeota phylum are essential for global methane cycling under anaerobic conditions, while Thaumarchaeota are the major drivers of the aerobic oxidation of ammonia in soil. Nevertheless, the taxonomy and ecology of soil archaea remain poorly understood. Soil metagenomic surveys, coupled with efforts to cultivate these microbes, should greatly expand our knowledge of soil archaea in the future.

Acknowledgments

The contribution of the author was supported by the Russian Science Foundation, project no. 21-76-10025.

References

- Adam PS, Borrel G, Brochier-Armanet C, and Gribaldo S (2017) The growing tree of archaea: New perspectives on their diversity, evolution and ecology. *The ISME Journal* 11: 2407–2425.
- Alves RJ, Minh BQ, Ulrich T, von Haeseler A, and Schleper C (2018) Unifying the global phylogeny and environmental distribution of ammonia-oxidising archaea based on amoA genes. *Nature Communications* 9: 1–17.
- Bates ST, Berg-Lyons D, Caporaso JG, Walters WA, Knight R, and Fierer N (2011) Examining the global distribution of dominant archaeal populations in soil. *The ISME journal* 5(5): 908–917.
- Bomberg M (2016) The elusive boreal forest Thaumarchaeota. *Agronomy* 6(2): 36.
- Bridgham S, Cadillo-Quiroz H, Keller J, and Zhuang Q (2013) Methane emissions from wetlands: Biogeochemical, microbial, and modeling perspectives from local to global scales. *Global Change Biology* 19: 1325–1346.
- Brochier-Armanet C, Boussau B, Gribaldo S, and Forterre P (2008) Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Reviews Microbiology* 6(3): 245–252.
- Buan NR (2018) Methanogens: Pushing the boundaries of biology. *Emerging Topics in Life Sciences* 2(4): 629–646.
- Gubry-Rangin C, Hai B, Quince C, Engel M, Thomson BC, James P, Schlöter M, Griffiths RI, Prosser JL, and Nicol GW (2011) Niche specialization of terrestrial archaeal ammonia oxidizers. *Proceedings. National Academy of Sciences. United States of America* 108(52): 21206–21211.
- Kharitonov S, Semenov M, Sabrekov A, Kotsyurbenko O, Zhelezova A, and Schegolkova N (2021) Microbial communities in methane cycle: Modern molecular methods gain insights into their global ecology. *Environments* 8(2): 16.
- Leininger S, Ulrich T, Schlöter M, Schwark L, Qi J, Nicol GW, Prosser JL, Schuster SC, and Schleper C (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* 442(7104): 806–809.
- Liu Y and Whitman WB (2008) Metabolic, phylogenetic, and ecological diversity of the methanogenic archaea. *Annals of the New York Academy of Sciences* 1125(1): 171–189.
- Nicol GW, Leininger S, Schleper C, and Prosser JL (2008) The influence of soil pH on the diversity, abundance and transcriptional activity of ammonia oxidizing archaea and bacteria. *Environmental Microbiology* 10(11): 2966–2978.
- Offre P, Spang A, and Schleper C (2013) Archaea in biogeochemical cycles. *Annual Review of Microbiology* 67: 437–457.
- Pester M, Schleper C, and Wagner M (2011) The Thaumarchaeota: An emerging view of their phylogeny and ecophysiology. *Current Opinion in Microbiology* 14(3): 300–306.
- Semenov MV, Manucharova NA, and Stepanov AL (2016) Distribution of metabolically active prokaryotes (Archaea and Bacteria) throughout the profiles of chernozem and brown semidesert soil. *Eurasian Soil Science* 49(2): 217.

- Semenov MV, Chernov TI, Tkhakakhova AK, Zhelezova AD, Ivanova EA, Kolganova TV, and Kutovaya OV (2018) Distribution of prokaryotic communities throughout the Chernozem profiles under different land uses for over a century. *Applied Soil Ecology* 127: 8–18.
- Semenov MV, Manucharova NA, Krasnov GS, Nikitin DA, and Stepanov AL (2019) Biomass and taxonomic structure of microbial communities in soils of the right-bank basin of the Oka River. *Eurasian Soil Science* 52: 971–981.
- Semenov MV, Krasnov GS, Semenov VM, Ksenofontova N, Zinyakova NB, and van Bruggen AH (2021) Does fresh farmyard manure introduce surviving microbes into soil or activate soil-borne microbiota? *Journal of Environmental Management* 294: 113018. <https://doi.org/10.1016/j.jenvman.2021.113018>.
- Shen LD, Ouyang L, Zhu Y, and Trimmer M (2019) Active pathways of anaerobic methane oxidation across contrasting riverbeds. *The ISME Journal* 13(3): 752–766.
- Stempfhuber B, Engel M, Fischer D, Neskovic-Prit G, Wubet T, Schöning I, et al. (2015) pH as a driver for ammonia-oxidizing archaea in forest soils. *Microbial Ecology* 69: 879–883.
- Timonen S and Bomberg M (2009) Archaea in dry soil environments. *Phytochemistry Reviews* 8: 505–518.
- Tripathi BM, Kim M, Lai-Hoe A, Shukor NA, Rahim RA, Go R, and Adams JM (2013) pH dominates variation in tropical soil archaeal diversity and community structure. *FEMS Microbiology Ecology* 86(2): 303–311.
- Valentine DL (2007) Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nature Reviews Microbiology* 5(4): 316–323.
- Verhamme DT, Prosser JL, and Nicol GW (2011) Ammonia concentration determines differential growth of ammonia-oxidising archaea and bacteria in soil microcosms. *The ISME Journal* 5(6): 1067–1071.
- Woese CR, Kandler O, and Wheelis ML (1990) Towards a natural system of organisms: Proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences* 87(12): 4576–4579.
- Yvon-Durocher G, Allen AP, Bastviken D, Conrad R, Gudas C, St-Pierre A, et al. (2014) Methane fluxes show consistent temperature dependence across microbial to ecosystem scales. *Nature* 507(7493): 488–491.
- Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, et al. (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 541(7637): 353–358.
- Zhang LM, Hu HW, Shen JP, and He JZ (2012) Ammonia-oxidizing archaea have more important role than ammonia-oxidizing bacteria in ammonia oxidation of strongly acidic soils. *The ISME Journal* 6(5): 1032–1045.

Further reading

- Oton EV, Quince C, Nicol GW, Prosser JL, and Gubry-Rangin C (2016) Phylogenetic congruence and ecological coherence in terrestrial Thaumarchaeota. *The ISME Journal* 10(1): 85–96.
- Semenov MV, Krasnov GS, Semenov VM, and van Bruggen AH (2020) Long-term fertilization rather than plant species shapes rhizosphere and bulk soil prokaryotic communities in agroecosystems. *Applied Soil Ecology* 154: 103641.

Bacteria—Soil biology

Mikhail V Semenov, Dokuchaev Soil Science Institute, Laboratory of Soil Carbon and Microbial Ecology, Moscow, Russian Federation

© 2023 Elsevier Ltd. All rights reserved.

This is an update of L.J. Halverson, BACTERIA | Soil, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 115–122, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00135-1>.

Introduction	32
Bacterial phylogenetic and physiological diversity	32
The microbe as the scale of study	33
Bacterial life strategies and survival	33
Avoidance	35
Stress tolerance	35
Water availability	35
Extracellular polysaccharide production	36
Competition for resources	36
Cooperation among bacteria	36
Siderophores	36
Quorum-sensing	37
Summary	37
Acknowledgments	37
References	37

Abstract

Bacteria are the most abundant organisms in soil. Despite small size and relative simplicity of prokaryotic cells, bacteria are the key players in soil processes, participating in biogeochemical cycles and maintaining soil health. The ability of bacteria to thrive in soil is due to their unequalled metabolic versatility and phenotypic plasticity, which allow them to colonize different soil habitats. This chapter examines the phylogenetic and physiological diversity of soil bacteria, as well as some aspects of their survival and functioning in soils.

Glossary

16S rRNA gene A gene that encodes the RNA component of the smaller subunit of the bacterial ribosome, responsible for converting genetic messages to functional cell components via the translation of mRNA to proteins. 16S rRNA is highly conserved in the evolutionary process of bacteria, permitting the detection of relatedness among very distant bacterial species.

Horizontal gene transfer The movement of genetic material between unicellular and/or multicellular organisms other than by reproduction.

Metabarcoding The large-scale taxonomic identification of complex environmental samples via analysis of DNA sequences for short regions of one or a few genes called DNA barcodes.

Metagenomics The direct genetic analysis of collective genomes contained within an environmental sample via whole-genome DNA sequencing.

Rhizosphere A narrow zone of soil adjacent to the roots of living plants that is directly influenced by root exudates.

Key points

- Soil bacteria carry out key ecosystem services, including cycling of carbon and other nutrients and sustaining plant growth
- A single gram of soil can harbor billions bacterial cells and tens of thousands bacterial species
- Soil bacteria can be categorized on the basis of their ecological strategies and used to predict soil functions
- Microscale spatial patterns regulate bacterial activity in soil
- Cooperative and coordinated behaviors are required for the development of bacterial community

Introduction

Bacteria are the most abundant organisms in soil, second only to fungi in their contribution to global soil microbial biomass (7 Gt C) (Bar-On et al., 2018). The majority of soil bacteria have a diameter of 0.5–1.0 μm . There are also many ultramicrobacteria (also known as “nanobacteria”) in soil, which are less than 0.2 μm in diameter (Lysak et al., 2010). Therefore, the average mass of a soil bacterial cell is only 2×10^{-14} g dry weight (or 10^{-14} g C) (Polyanskaya et al., 2017). Despite small size and relative simplicity of prokaryotic cells, bacteria are the key players in soil processes, participating in biogeochemical cycles and maintaining soil health (Crowther et al., 2019; van Bruggen et al., 2019). Soil bacteria can regulate nutrient availability, aggregate stability, C sequestration, pollutant degradation, plant health and plant growth promotion (Fierer, 2017). They are also involved in the cycling of trace gases, such as nitrous oxide (N_2O), nitric oxide (NO), carbon monoxide (CO), hydrogen (H_2) and reduced sulfur species, which affect the chemistry of the atmosphere and, thus, indirectly also affect the radiation balance of Earth.

The species concept for bacteria is still obscure. It has been proposed that the species in prokaryotes correspond roughly to the genera of eukaryotic organisms, and that prokaryotic ecotypes correspond to eukaryotic species. The most accepted species definition, the phylo-phenetic species, is defined as a monophyletically and genomically coherent cluster of individual organisms that show a high degree of overall similarity in many independent characteristics and is diagnosable by a discriminative phenotypic property (Bergey, 2001; Torsvik and Øvreas, 2006). The 97% 16S rRNA similarity level has been chosen as a threshold criterion for defining bacterial species. Its use for the description of bacterial diversity is, however, limited since strains of the same bacterial species, namely organisms whose indicator 16S rRNA sequences are >97% similar, typically only share some genes, while other genes are strain-specific and unique (Fernández et al., 2019). The traditional species concept as isolated clusters of genetically compatible isolates may be shifted to a cohesive speciation network in which such clusters are interconnected by genetic processes.

The most important issues related to soil bacteria are their phylogenetic and physiological diversity, patterns of spatial distribution and metabolic activities at the scale of the microbe, as well as life strategies and survival of bacteria in the soil environment. The ability of bacteria to thrive in soil is due, in part, to their unequaled metabolic versatility and phenotypic plasticity, which allow them to colonize the vastly different habitats that make up soils. Bacteria, probably more than any other type of organism, can quickly adapt to fluctuations in environmental conditions, both in physiological and genetic terms. Life in soil has clearly driven bacteria to fine-tune their genetic systems and genomes over evolutionary time allowing them to cope directly with the local conditions offered by the soil environment (Standing and Killham, 2006).

Bacterial phylogenetic and physiological diversity

Soil is the most complex environment with greatest microbial diversity (Walters and Martiny, 2020). A single gram of soil can harbor up to 10^{10} – 10^{11} bacterial cells and an estimated diversity of tens of thousands of species (Wagg et al., 2014). Bacteria account for up to 99% of the genes in the soil metagenome (Semenov, 2019). Until recently, it was impossible to detect the vast majority of bacterial diversity, since most soil microorganisms (95–99%) do not grow readily under standard laboratory conditions. However, direct extraction and characterization of microbial community DNA through PCR amplicon surveys (metabarcoding) and metagenomics has revolutionized our ability to characterize microbial diversity by enabling the investigation of community composition at a much greater phylogenetic resolution than ever before (Nesme et al., 2016; Semenov, 2019). Metagenomic analysis of nucleic acids provides direct access to the genomes of the “uncultivated majority” (Nesme et al., 2016). According to 16S rRNA gene amplicon sequencing, all soil bacterial communities are dominated by Acidobacteria, Actinobacteria, Bacteroidetes, Proteobacteria, and Verrucomicrobia (Fierer et al., 2012). Additional phyla including Chloroflexi, Cyanobacteria, Firmicutes, and Gemmatimonadetes are also found in nearly all soils (Fierer et al., 2012). In contrast to the other dominant bacterial taxa, most Verrucomicrobia have not been yet isolated in pure culture. However, this phylum is one of the main dominants in prokaryotic communities; their proportion typically varies from 5 to 15–25% of prokaryotic sequences (Semenov et al., 2018). Despite such high abundance, we lack a predictive understanding of the ecological attributes of most soil individual verrucomicrobial taxa, with their environmental preferences, traits, and metabolic capabilities remaining largely unknown.

From an ecological point of view, information regarding the metabolic functions of soil bacterial communities is certainly the most relevant (Baldrian, 2019). Currently, one of the most relevant issues in soil microbial ecology is the conversion of the soil bacterial taxonomic profile obtained by metabarcoding or metagenomics into a functional profile (Alteio et al., 2021; Blagodatskaya et al., 2021). The functional diversity of a bacterial community has been defined as the occurrence and distribution of physiological and metabolic traits among member of that community (Torsvik and Øvreas, 2006; Malik et al., 2020). Functional diversity can be expressed as the number of different functional groups (guilds) present. For instance, soil bacteria can be described by their physiological and ecological properties (Table 1).

Adaptive loss of function or genome streamlining, convergent evolution and horizontal gene transfer all erode the phylogenetic signal of many traits (Louca et al., 2018). Horizontal gene transfer also leads to low genetic linkage of traits within genomes and hence to re-assortment of traits between genomes. Some *Escherichia coli* strains, for example, overlap by less than 40% in their protein-coding genes. The phylogenetic scale on which functions are conserved varies strongly between functions, and even for single functions phylogenetic conservatism can vary between clades (Fernández et al., 2019). As a result, the contrast between stable functional composition and variable taxonomic composition seen in the aforementioned studies reflects a weak association

Table 1 Examples of physiological and ecologic classifications of soil bacteria.

Category	Classification	Description
Nutritional	Heterotroph	Able to use organic compounds as energy and carbon source
	Phototroph	Able to use light as its energy source
	Autotroph	Able to use carbon dioxide as a sole carbon source
	Chemolithotroph	Able to use an inorganic substrate as an electron donor
Functional	Saprophyte	Degrades nonliving organic material
	Nitrifiers	Oxidation of ammonia to nitrite and nitrate
	Denitrifiers	Reduction of nitrate to dinitrogen gas
	Nitrogen-fixers	Reduction of dinitrogen gas to ammonia
	Sulfate-reducers	Reduction of sulfate to hydrogen sulfide
	Sulfate-oxidizers	Oxidation of elemental sulfur to sulfate
	Decomposers	Breakdown of complex compounds into simpler compounds
Interactions	Commensalism	One organism benefits while the other is neither harmed nor benefits
	Parasitism	One organism benefits while the other is harmed
	Mutualism	Both organisms benefit

between many functions and bacterial phylogeny (Baldrian, 2019; Alteio et al., 2021). Quantification of microbial diversity involved in various metabolic functions also revealed that communities typically exhibit high ‘functional redundancy’ with respect to a multitude of functions, in the sense that each metabolic function can be performed by multiple coexisting, taxonomically distinct organisms (Louca et al., 2018).

A multitude of biotic and abiotic factors affects the structural and functional diversity of soil bacteria. Soil pH is considered as the best predictor of bacterial community composition (Delgado-Baquerizo et al., 2018). Also, important factors that have notable influences on the structure of soil bacterial communities are nitrogen availability, soil organic carbon content, temperature, moisture, and redox status (Fierer, 2017).

The microbe as the scale of study

It is important not to neglect the element of scale as we try to understand bacterial behavior and processes (Blagodatskaya et al., 2021; Hugoni et al., 2021). While macro- and mesoscale distributions of microbes indicate the average abundance of microorganisms through volumes of soil, it is the patterns of distribution and metabolic activities at the scale of the microbe (the microscale) that will control many microbial behaviors (Blagodatskaya et al., 2021; Raynaud and Nunan, 2014). Studies on microbial systems (whether they focus on microbial activity or diversity) are generally carried out at scales many orders of magnitude larger than those at which microorganisms interact with other organisms or with their surrounding environment (Baldrian, 2019). From the perspective of a bacterium, a volume of soil or a root several centimeters away can represent the distance of tens kilometers. The 1–10 µm scale will be of greater significance for growth and activity of bacteria than bulk soil properties.

Solid aggregates separate the soil into many fully or partially isolated microniches with contrasting conditions (Standing and Killham, 2006). Bacteria inhabit these niches in accordance with their needs for nutrition and survival, resulting in an uneven distribution of the microbiome inside soil aggregates (Fig. 1). The high heterogeneity of bacterial communities is also due to the complexity of the soil pore system. For example, bacteria are able to enter and colonize small soil pores, potentially protecting them from predation. Moreover, micropores and biofilms on macropore walls often contain higher concentrations of organic substances that can be used by microorganisms as substrates.

Microscale spatial patterns may have a regulatory effect on bacterial activity, as a result of the diffusive delivery of substrate to, and the dispersion of product from bacterial cells, which may have rate-limiting or stimulatory effects on microbially mediated processes (Raynaud and Nunan, 2014). Furthermore, bacteria can release chemical signals that modulate bacterial functions once a critical concentration of chemical is reached (known as quorum-sensing). Spatial separation controls horizontal gene transfer in bacterial populations.

Bacterial life strategies and survival

Heterotrophic bacterial taxa can be divided into broad categories on the basis of shared life strategies and survival (Fierer, 2017; Malik et al., 2020; Blagodatskaya et al., 2021). For instance, soil bacterial taxa could be grouped by growth and efficiency into r- and K-strategists. The r-selected strategists have short life expectancy and large reproductive effort, whereas K-selected strategists have long life expectancy and invest a smaller proportion of energy and resources into reproduction. The copiotroph-oligotroph continuum is proposed as analogous to the r- and K-selection theory and is based mainly on microbial substrate preferences and trophic strategy. Another option is to modify a trait-based competitor–stress tolerator–ruderal (C-S-R) concept and adopt it to

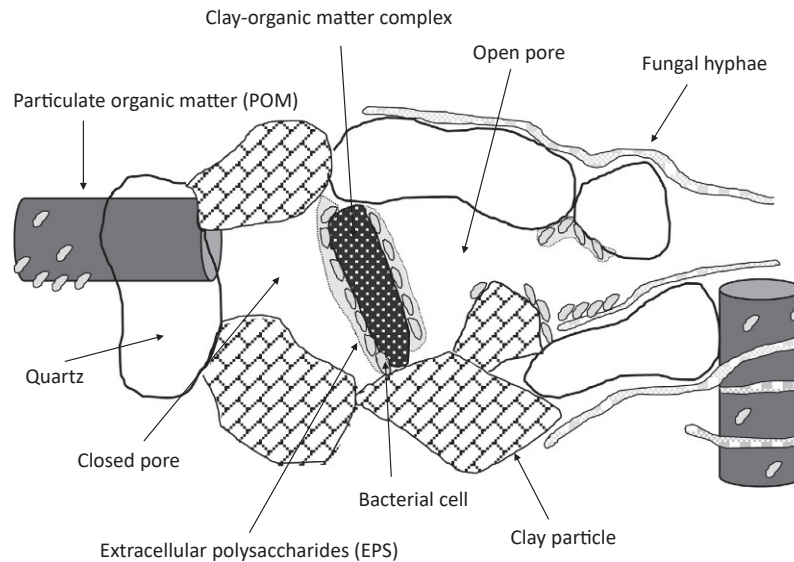


Fig. 1 Microscale distribution of soil bacteria.

categorize the broad spectrum of soil bacterial lifestyles (Fierer, 2017) (Table 2). ‘Stress-tolerant’ taxa can persist under low-resource or suboptimal abiotic conditions. ‘Competitor’ taxa outcompete other soil microbial taxa for space or resources. ‘Ruderal’ taxa grow rapidly and exploit unoccupied niches generated as a result of biotic or abiotic disturbances.

The C-S-R strategies do not fully reflect microbial systems due to the reliance of heterotrophic microorganisms on an external source of carbon and energy that drive differential cellular allocation and lead to tradeoffs in traits. This makes the quantity and quality of resources in the surrounding environment a key factor in influencing species distribution. In addition, the way microbes respond to disturbance and stress may not be entirely distinguishable. To overcome these limitations, a new Y-A-S (yield-acquisition-stress tolerance) life strategy theory for soil bacteria was recently proposed (Malik et al., 2020) (Table 3). High yield (Y) strategists maximize the fraction of resource uptake that is allocated to biosynthetic processes by investing in central metabolism and associated assimilatory pathways such as amino acid, nucleotide, and fatty acid synthesis to build cellular components using these precursor compounds. The absence of resource limitation and stress are expected to favor the high yield strategy. The resource acquisition (A) strategy prevails in low-resource conditions where bacteria would be under selection to increase resource capture at the expense of growth yield. Microbes exposed to suboptimal abiotic conditions would possess traits linked to stress tolerance (S) at the expense of other traits (Malik et al., 2020). Recent developments in molecular approaches have provided potential for microbial trait differentiation based on information regarding genome size, number of ribosomal gene copies per genome, and quantification of functional marker genes or their transcripts by -omics approaches (Li et al., 2021).

Table 2 Examples of physiological and ecologic classifications of soil bacteria (based on Fierer, 2017).

Classification	Description	Examples
Stress-tolerators	Auxotrophic	Acidobacteria
	Slow growth	Chloroflexi
	Few rRNA operon copies	Verrucomicrobia
	High substrate affinity	
	Production of extracellular polymeric substances (EPS)	
Competitors	Larger genomes	Many Actinobacteria (e.g. <i>Streptomyces</i>)
	Antibiotic production	Many α - and γ -Proteobacteria (e.g. <i>Pseudomonas</i>)
	Filamentous growth	
	High catabolic diversity	
	Siderophore production	
Ruderal	Rapid growth	Many Firmicutes (e.g. <i>Bacilli</i>)
	Multiple rRNA operon copies	Many β -Proteobacteria (e.g. <i>Burkholderia</i>)
	Low catabolic diversity	Many Bacteroidetes (e.g. <i>Flavobacterium</i>)
	Spore formation	
	Low C/N ratio	

Table 3 Y-A-S life history strategies and underlying traits of soil bacteria (based on Malik et al., 2020).

<i>Classification</i>	<i>Conditions</i>	<i>Description</i>
Growth yield (Y)	Resource abundance Ambient	Growth per unit resource
Resource acquisition (A)	Resource limitation Ambient	Degradation of complex substrates Motility: resource discovery Uptake of simple substrates
Stress tolerance (S)	Resource limitation Stressed	Biomolecular damage repair Osmolyte production Extracellular polymeric substances (EPS) synthesis Maintenance of cellular integrity

Avoidance

To avoid a stress, bacteria need to actively reposition themselves toward a more hospitable site. Avoidance can also happen passively when bacteria are physically repositioned in the soil profile by root movement or through some other physical disturbance, but this is not an effective strategy for surviving stressful times. If bacteria grow or survive at particular sites in soils, then the ability to move to such sites is an important arsenal in their survival strategies. Bacterial motility is likely to play a role in positioning them in a niche that is abundant in desirable nutrients, low in toxic or undesirable compounds, or provides protective sites. Although flagellar motility requires a sufficiently thick water film, which is more common in wetter soils, it is possible that twitching or surface motility may play a role in positioning the cell in a favorable site, particularly when water films are too thin for flagellar motility, which more commonly occurs in unsaturated soils. It is uncertain over what distances could bacteria move using this form of motility, but it would be shorter than that covered by flagellar motility.

Once a bacterium reaches a favorable site, the ability to resist removal may be advantageous. The strength of the selective pressure depends on the strength of the removal pressure, such as water percolation following rain, tillage, or movement of a root through soil. Once associated with a surface, bacteria can adhere to it by specific attachment structures such as cellulose fibrils, membrane proteins, pili, or extracellular polymeric substances (EPS), such as polysaccharides. It has been hypothesized that bacteria employ specific adhesins in environments where there are strong physical forces. Due to the improbability of continuous strong shear forces in soil, bacteria probably do not use adhesion extensively, except when they occupy sites that offer some protection, or are nutrient-rich, or on roots, because root movement could provide sufficiently strong physical forces to cause their removal.

Stress tolerance

Bacteria in soil are exposed to a variety of environmental conditions that are going to affect them such as, for example, fluctuations in temperature, availability of water and nutrients, the presence of toxic environmental pollutants, or toxic metabolic wastes produced by themselves or by their neighbors. Survival in a continually changing environment requires a wide range of fast, adaptive responses, some of which do not require transcriptional activation of genes whose products facilitate coping with a given stress, since these responses take too long to complete. For example, modification of fatty acid composition in pseudomonads provides an extremely rapid response to maintain membrane fluidity when they are exposed to stresses such as desiccation or organic solvents that perturb membrane properties. Transcriptional activation of genes whose products facilitate coping with stress provides more long-term solutions. It is not feasible to review bacterial responses to every stress they will potentially encounter in soil. Here, reduced water availability is highlighted, and this is likely to be a dominant factor influencing bacteria in many soils; the mechanisms of adaptation to this stress are likely to be similar to the mechanism for adaptation cold or heat stress, two common stresses bacteria frequently experience in surface soils, since they all alter physical properties of bacterial membranes.

Water availability

If solutes are abundant in free soil-water, they could become sufficiently concentrated upon drying to stress the resident bacteria. Osmotic stress is only one component of the total water stress that a bacterium may encounter in soil as it dries. The total soil-water potential is a quantitative term reflecting water availability. It is comprised primarily of the sum of the osmotic potential, which is due to the interaction of water with solutes, and the matric potential, which is due to the interaction of water with soil constituents that increases as the soil dries. From a bacterial perspective, the primary difference between these two stresses is that with an osmotic stress bacteria are bathed in water (albeit water with diminished activity), whereas with a matric stress bacteria become desiccated by removal of water from its environment, and the availability of the water that is remaining is reduced due to its sorptive interactions with soil constituents. The stress imposed by a low matric potential has a stronger influence on a bacterial cell than an equivalent osmotic potential. This is due in part to the greater cellular dehydration that occurs under a matric than an osmotic stress. Dehydration produces DNA damage, denatures proteins, and causes membranes to become less fluid and potentially damaged. Resistance to dehydration is crucial for microbial life in soil habitats. These include various strategies to counter the life-threatening,

lipid-solidifying effects of desiccation, synthesizing solutes compatible with cellular physiology to create an intracellular water potential that is in equilibrium with the external environment, protecting and repairing DNA, and producing EPS.

Extracellular polysaccharide production

There is considerable evidence that soil bacteria are capable of producing EPS, especially polysaccharides, and increased EPS production by soil bacteria has been shown to improve soil aggregation, aggregate stability, and water retention properties. These polysaccharides are likely to have a dominant role in the ecology of these organisms, since their presence defines the environment immediately surrounding the bacteria. As such, they mediate the interactions of the bacteria with the surface, other microorganisms, and the physical and chemical environment. The EPS matrix that surrounds them can have ion-exchange capabilities, which can concentrate nutrients from dilute sources in the vicinity of the cell or provide protection from predators and shield cells from the action of lytic enzymes, antibiotics, and other inhibitory compounds. EPS production may be particularly important during desiccating conditions, because many EPS are hygroscopic and their production results in a more hydrated microenvironment in the immediate vicinity of the cell relative to the bulk environment. The greater water content in the immediate vicinity of cell membranes may reduce the physical damage to membrane properties that might occur otherwise. This is an example of how habitat modification can benefit the resident bacteria.

Many studies have revealed aggregates of bacteria in bulk soil or on root surfaces. These bacteria are closely packed and covered with an amorphous material that is presumably EPS. This association of aggregates of bacteria in a matrix is analogous to that of biofilms, which are defined as assemblages of bacteria on surfaces encased in EPS of their own making.

Competition for resources

Soil bacterial communities are characterized by high temporal dynamics which mirrors permanent fluctuations and recurrent variations in the community structure, composition or function (Zelenev et al., 2005). Temporal fluctuations of soil bacteria are governed not only by environmental conditions, but also by intra-species and inter-species interactions relationships (Zelenev et al., 2006). Bacteria of different species co-occur in soil and, if these species occupy the same microsites, the survival of any one of them depends on its ability to compete successfully for shared resources or to coexist by utilizing different resources (Zelenev et al., 2005; Zelenev et al., 2006). These resources can include the same physical site or be nutritionally based, and competition for nutrients is a driving factor in microbial ecology of bacteria in soil. Thus, a broad nutrient utilization profile would increase the potential for an organism to survive in the presence of others and the ability to utilize an abundant resource should be highly advantageous as long as other resources are not limiting. Many soil bacteria, and in particular, rhizosphere (the root surface and the region immediately surrounding a root) colonists have broad nutrient utilization profiles. Likewise, the ability of soil bacteria to tolerate prolonged nutrient deprivation would also be highly advantageous. If microbes must compete for limited resources, production of antimicrobial metabolites, such as an antibiotic, could improve their competitive ability. Many soil bacteria are capable of producing antibiotics that target dissimilar organisms or bacteriocins that target more closely related strains of the same species, and their production would conceivably provide a competitive advantage.

The rhizosphere constitutes an ecologic niche where nutrients are more readily available compared with the bulk soil. The ability to compete effectively for these nutrients is likely to be an important factor contributing to the ability of a bacterium to become a rhizosphere colonist. The deposition of organic material from the roots (rhizodeposition) enhances microbial growth, drives the structuring of microbial communities, and controls their metabolic activities in the rhizosphere (Kuzyakov and Blagodatskaya, 2015). The interactions between the plant and the surrounding bacteria in the bulk soil select for the establishment of certain populations (rhizobacteria). While rhizodeposition strongly influences the size and activity of microbial populations at the soil–plant interface, the activity of these populations in turn affects plant health and thus influences both the quantity and quality of rhizodeposition. The potential for either an exudative response to bacteria or a response by bacteria to exudation suggests a certain degree of coevolution between plants and rhizobacteria.

Cooperation among bacteria

The ability to modify a habitat may be augmented by cooperative interactions among bacteria and these interactions may occur among both homogeneous and heterogeneous populations. The occurrence of aggregated populations of bacteria in soil or in association with plant roots provides an opportunity to maximize opportunities for interactions among community members. There have been several recent descriptions illustrating sophisticated cooperation among community members, even though at times they may appear not to be cooperative attributes.

Siderophores

Although iron is abundant, the extreme insolubility of ferric hydroxide limits the amount of free iron available in aerobic soil environments since the iron exists in the form of insoluble oxides at neutral or alkaline pH soils. Many bacteria use siderophores and corresponding membrane receptors to acquire this essential nutrient (Souza et al., 2015). Siderophores, which are low-

molecular-weight molecules produced under iron-limiting conditions, bind the ferric iron (Fe^{3+}), the membrane receptor recognizes a particular siderophore–iron complex, and the iron is then transported into the cell. Any particular bacterial species or strain generally produces one or more siderophores and the corresponding membrane receptors for transporting each siderophore–iron complex into the cell. The bacterium releases the siderophore into the environment and then the siderophore needs to chelate ferric iron and diffuse back to the cell for the cell to benefit from the energy spent to synthesize the siderophore. Yet many soil bacterial species have the capacity to utilize siderophores produced by other bacterial strains or species and possibly fungi (Souza et al., 2015). This cross-feeding clearly provides advantages to community members, since there is a greater potential return on the metabolic investment of siderophore production by taking up ferric–siderophore complexes that it did not synthesize rather than depending solely on diffusion of the siderophore it produced back into the cell. This illustrates that bacterial survival may not depend solely on the efforts of the individual but also, in part, on the efforts of the community.

Quorum-sensing

Bacteria can communicate with one another using chemical signaling molecules. Specifically, they release, detect, and respond to the accumulation of molecules called autoinducers. Detection of autoinducers allows bacteria to distinguish between low and high cell population density, and to control gene expression in response to changes in cell number. In reality, this is not necessarily a function of high cell density, but is likely to be related more directly to high signal concentrations, which are likely to be achieved at lower population densities in soil due to their low water volumes that result in relatively high autoinducer concentrations. This process, termed ‘quorum-sensing,’ allows a population of bacteria to coordinately control gene expression of the entire community and allows the bacteria to behave as multicellular organisms and to reap the benefits that would be unattainable to them as individuals (Hawver et al., 2016). Many microbial behaviors are regulated by quorum-sensing, including, for example, expression of nodulation genes in the *Rhizobium*–legume symbioses, virulence traits of bacterial pathogens of plants, conjugal transfer of plasmids between bacteria, antibiotic production, and biofilm formation (Bogino et al., 2015).

Many bacteria have also developed mechanisms to interfere with bacterial quorum sensing, such as secreting enzymes that destroy autoinducers or producing autoinducer antagonists. Furthermore, some plants secrete substances that mimic or interfere with *N*-acyl-homoserine lactone autoinducer signal activities and affect population density-dependent behaviors in associated bacteria. These autoinducer signal-mimic compounds could prove to be important in determining the outcome of interactions between higher plants and a diversity of pathogenic, symbiotic, and saprophytic bacteria. Furthermore, some bacteria are capable of recognizing autoinducers produced by different bacterial species, indicating that interspecies quorum-sensing may play an important synergistic or competitive role in the dynamic behaviors of soil bacterial communities (Hawver et al., 2016).

Summary

Soil bacterial communities are surprisingly diverse and abundant; a single gram of soil can harbor billions bacterial cells and tens of thousands bacterial species. Soil bacteria carry out key ecosystem services, including cycling of carbon and other nutrients and sustaining plant growth. Recent advances in marker gene and metagenomic analyses have greatly expanded our ability to characterize the soil bacterial communities, taxonomically and functionally, and identify the factors that shape them across space and time. One of the most relevant issues in soil microbial ecology is the conversion of the data on bacterial taxonomic profile into a functional status of the soil. Although most soil microorganisms remain undescribed, we are able to categorize soil bacteria on the basis of their ecological strategies and thereby predict the soil functions.

The ability of bacteria to survive in soil with its inherent complexity and heterogeneity under conditions that are often severe is a testament to their metabolic versatility and phenotypic plasticity (adaptability). Many bacteria are able to modify their local environment, such as by producing EPS, to facilitate growth, and their increased numbers facilitate further habitat modification. Cooperative and coordinated behaviors are required for the development of bacterial community. Each individual member of the community, through its own actions, based in part on the signals it receives from other community members, initiates a spectrum of adaptive strategies that are appropriate for the particular habitat they reside in and the specific environmental challenges they are experiencing to optimize chances for survival.

Acknowledgments

The contribution of the author was supported by the Russian Science Foundation, project no. 21-76-10025.

References

- Alteio LV, S  neca J, Canarini A, Angel R, Guseva, et al. (2021) A critical perspective on interpreting amplicon sequencing data in soil ecological research. *Soil Biology and Biochemistry* 160: 108357.
- Baldrian P (2019) The known and the unknown in soil microbial ecology. *FEMS Microbiology Ecology* 95: fiz005.

- Bar-On YM, Phillips R, and Milo R (2018) The biomass distribution on Earth. *Proceedings of the National Academy of Sciences* 115: 6506–6511.
- Bergey DH (2001) *Bergey's Manual® of Systematic Bacteriology*. vol. 2. New York: Springer Science & Business Media.
- Blagodatskaya E, Tarkka M, Knief C, Koller R, Peth S, et al. (2021) Bridging microbial functional traits with localized process rates at soil interfaces. *Frontiers in Microbiology* 12: 625697.
- Bogino PC, Nieves FL, and Giordano W (2015) A review: Quorum sensing in *Bradyrhizobium*. *Applied Soil Ecology* 94: 49–58.
- Crowther TW, Van den Hoogen J, Wan J, Mayes MA, Keiser AD, et al. (2019) The global soil community and its influence on biogeochemistry. *Science* 365: eaav0550.
- Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, et al. (2018) A global atlas of the dominant bacteria found in soil. *Science* 359: 320–325.
- Fernández SL, Větrovský T, and Baldrian P (2019) The concept of operational taxonomic units revisited: Genomes of bacteria that are regarded as closely related are often highly dissimilar. *Folia Microbiologica* 64: 19–23.
- Fierer N (2017) Embracing the unknown: Disentangling the complexities of the soil microbiome. *Nature Reviews Microbiology* 15: 579–590.
- Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, et al. (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences* 109: 21390–21395.
- Hawver LA, Jung SA, and Ng WL (2016) Specificity and complexity in bacterial quorum-sensing systems. *FEMS Microbiology Reviews* 40: 738–752.
- Hugoni M, Nunan N, Thioulouse J, Dubost A, Abrouk D, et al. (2021) Small-scale variability in bacterial community structure in different soil types. *Microbial Ecology*. <https://doi.org/10.1007/s00248-020-01660-0>.
- Kuz'yakov Y and Blagodatskaya E (2015) Microbial hotspots and hot moments in soil: Concept & review. *Soil Biology and Biochemistry* 83: 184–199.
- Li H, Yang S, Semenov MV, Yao F, Ye J, et al. (2021) Temperature sensitivity of SOM decomposition is linked with a K-selected microbial community. *Global Change Biology* 27: 2763–2779.
- Louca S, Polz MF, Mazel F, Albright MB, Huber JA, et al. (2018) Function and functional redundancy in microbial systems. *Nature Ecology and Evolution* 2: 936–943.
- Lysak LV, Lopygina EV, Konova IA, and Zvyagintsev DG (2010) Population density and taxonomic composition of bacterial nanoforms in soils of Russia. *Eurasian Soil Science* 43: 765–770.
- Malik AA, Martiny JB, Brodie EL, Martiny AC, Treseder KK, and Allison SD (2020) Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. *The ISME Journal* 14: 1–9.
- Nesme J, Achouak W, Agathos SN, Bailey M, Baldrian P, et al. (2016) Back to the future of soil metagenomics. *Frontiers in Microbiology* 7: 73.
- Polyanskaya LM, Pinchuk IP, and Stepanov AL (2017) Comparative analysis of the luminescence microscopy and cascade filtration methods for estimating bacterial abundance and biomass in the soil: Role of soil suspension dilution. *Eurasian Soil Science* 50: 1173–1176.
- Raynaud X and Nunan N (2014) Spatial ecology of bacteria at the microscale in soil. *PLoS One* 9: e87217.
- Semenov MV (2019) Metabarcoding and metagenomics in soil ecology research: Achievements, challenges, and prospects. *Zhurnal Obshchei Biologii* 80: 403–417.
- Semenov MV, Chernov TI, Tkachkova AK, Zhelezova AD, Ivanova EA, et al. (2018) Distribution of prokaryotic communities throughout the Chernozem profiles under different land uses for over a century. *Applied Soil Ecology* 127: 8–18.
- Souza RD, Ambrosini A, and Passaglia LM (2015) Plant growth-promoting bacteria as inoculants in agricultural soils. *Genetics and Molecular Biology* 38: 401–419.
- Standing D and Killham K (2006) The soil environment. In: Van Elsas JD, Jansson JK, and Trevors JT (eds.) *Modern soil Microbiology*, 2nd edn, pp. 1–22. New York: CRC Press.
- Torsvik V and Øvreås L (2006) The soil environment. In: Van Elsas JD, Jansson JK, and Trevors JT (eds.) *Modern Soil Microbiology*, 2nd edn, pp. 23–54. New York: CRC Press.
- van Bruggen AH, Goss EM, Havelaar A, van Diepeningen AD, Finckh MR, and Morris JG Jr. (2019) One Health-Cycling of diverse microbial communities as a connecting force for soil, plant, animal, human and ecosystem health. *Science of the Total Environment* 664: 927–937.
- Wagg C, Bender SF, Widmer F, and van der Heijden MGA (2014) Soil biodiversity and soil community composition determine ecosystem multifunctionality. *Proceedings of the National Academy of Sciences* 111: 5266–5270.
- Walters KE and Martiny JB (2020) Alpha-, beta-, and gamma-diversity of bacteria varies across habitats. *PLoS One* 15: e0233872.
- Zelenev VV, Van Bruggen AHC, and Semenov AM (2005) Short-term wavelike dynamics of bacterial populations in response to nutrient input from fresh plant residues. *Microbial Ecology* 49: 83–93.
- Zelenev VV, Van Bruggen AHC, Leffelaar PA, Bloem J, and Semenov AM (2006) Oscillating dynamics of bacterial populations and their predators in response to fresh organic matter added to soil: The simulation model 'BACWAVE-WEB'. *Soil Biology and Biochemistry* 38: 1690–1711.

Soil protists

Stefan Geisen^a, Nathalie Amacker^b, and Alexandre Jousset^b, ^aLaboratory of Nematology, Department of Plant Sciences, Wageningen University and Research, Wageningen, The Netherlands; ^bEcology and Biodiversity group, Institute of Environmental Biology, University of Utrecht, Utrecht, The Netherlands

© 2023 Elsevier Ltd. All rights reserved.

This is an update of W. Foissner, PROTOZOA, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 336–347, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00162-4>.

Introduction	39
Diversity and ecology	40
Adaptations to the soil ecosystem	40
Abundance and diversity in soils	42
Geographical distribution and diversity patterns	42
Ecological relevance	44
Predation	44
Parasites	45
Nutrient cycling: C, N, P and Si	45
Soil protists in agricultural systems	46
Impact of agricultural management on soil protists	46
Soil protist to support plant performance	46
Soil protist as bioindicators	47
Conclusion	47
References	47

Abstract

Protists are the most abundant and diverse eukaryotes that inhabit virtually all soils. They are active players in soil food webs as phototrophic algae, plant and animal parasites, and microbiome predators. As predators, protists lead to modification of their prey community composition typically promoting functions linked to (plant-)pathogen suppression. The diverse functional roles of protists are further connected to cycling of carbon, nitrogen and other nutrients such as phosphorous and silicon. Thanks to their link to plant growth and their dynamic response to environmental changes, soil protists are further suggested as bioindicators of soil and plant health.

Key points

- Protists are the most abundant and diverse eukaryotes in soils
- Soil protist can have local or global distribution
- Protists drive microbiome dynamics and soil fertility
- Protists can favor pathogen suppressive microbial communities
- Protists are valuable bioindicators to monitor soil health

Introduction

Protists are among the most abundant and diverse microorganisms. Protists are a paraphyletic group referring to all Eukaryotes except plants, animals and Fungi. Soil protists are mostly unicellular and display a size range of few micrometers to millimeters. As to be expected from their broad phylogenetic coverage, protists exhibit a huge variety of morphotypes with the classical division between amoebae, flagellates, ciliates and shell-forming (testate) types; we provide some examples of protist species from different sizes and morphotypes in Fig. 1.

In the soil context, protistologists have primarily focused their efforts on the heterotrophic, free-living protists, traditionally referred to as “protozoa.” Many heterotrophic protists are predators and feed on relatively small prey such as bacteria and yeasts by ingesting them via phagocytosis, but others graze or attack bigger organisms such as hyphal-forming fungi and other micro-eukaryotes (Geisen et al., 2018). While protist predators have recently been reported to be the most abundant functional group in many natural soil ecosystems (Oliverio et al., 2020; Xiong et al., 2021), soil protists can also be autotrophic, mixotrophic or parasitic. Soil protists thus exhibit a range of ecological functions from primary producers to decomposers, and include important predators and pathogens (Geisen et al., 2018).

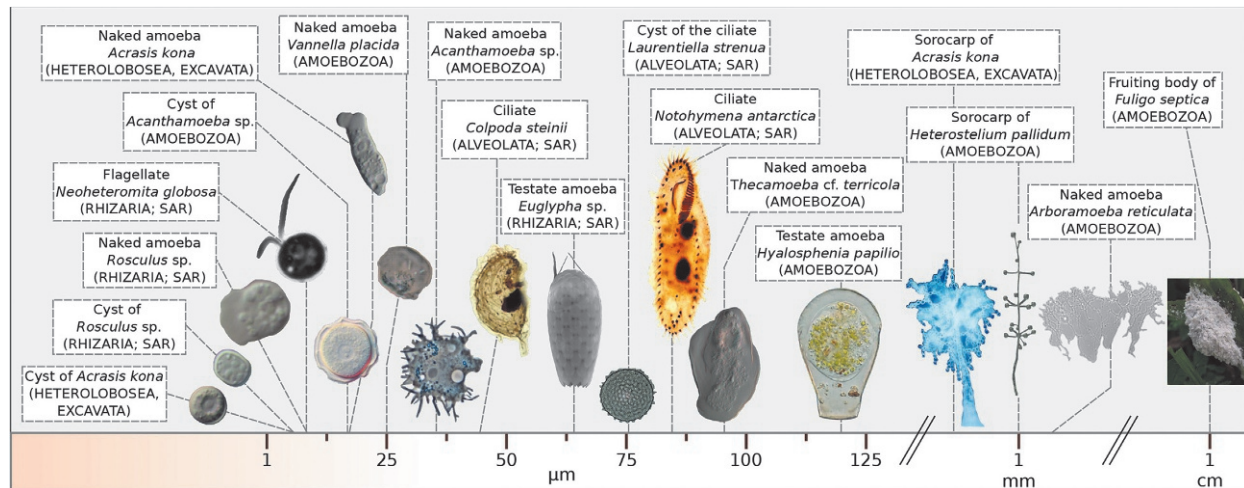


Fig. 1 Common free-living soil protists as visualized by size (lengths), morphology and phylogenetic affiliation. Reproduced with permission from Geisen S, Mitchell EAD, Wilkinson DM, Adl S, Bonkowski M, Brown MW, Fiore-Donno AM, Heger TJ, Jassey VEJ, Krashevska V, Lahr DJG, Marcisz K, Mulot M, Payne R, Singer D, Anderson OR, Charman DJ, Ekelund F, Griffiths BS, Rønn R, Smirnov A, Bass D, Belbahri L, Berney C, Blanderier Q, Chatzinotas A, Clarholm M, Dunthorn M, Feest A, Fernández LD, Foissner W, Fournier B, Gentekaki E, Hájek, M, Helder J, Jousset A, Koller R, Kumar S, La Terza A, Lamentowicz M, Mazei Y, Santos SS, Seppely CVW, Spiegel FW, Walochnik J, Winding A, and Lara E (2017) Soil protistology rebooted: 30 fundamental questions to start with. *Soil Biology and Biochemistry* 111: 94–103. <https://doi.org/10.1016/j.soilbio.2017.04.001>.

As a consequence of such ecological breadth, the study of protists has been spread over different research fields that have historically been disjunct. Autotrophic protists were, for instance, typically studied by botanists, while heterotrophic protists were considered as part of zoology. In the early 2000s, a framework called the eukaryote Tree of Life (abbreviated eToL) was established to reconnect this knowledge. Five to eight so-called ‘supergroups’ were created as broad assemblages to help classify the protists. These supergroups were initially based on morphological or cell biological features (e.g., the posterior flagellum in motile cells for the Opisthokonta, the red alga-derived plastids for the Chromalveolata) and have, since then, been strongly remodeled, increasingly based on (multigene) molecular phylogenetic information and less so on distinctive biological features. The contemporary discovery of novel taxa also continues to contribute to the remodeling of these supergroups (Burki et al., 2020). We present in Fig. 2 a schematic representation of the phylogenetic tree of the eukaryotes, highlighting with brown dots common taxonomic groups found in soils.

Soil protistology has received a constant interest over the last decades with an average of 93 ± 11 publications (per 5 years) for the period between 1980 and 2010. The related research effort is further very likely to follow a steady growing curve as witnessed for the aquatic environments with an observed 20% relative increase in the number of publications (per 5 years) for marine systems from 1995 to 2020 and 30% relative increase for freshwater and aquatic systems for the period 2005–2020 (see Fig. 3 below). Indeed, the number of publications related to soil protists has more than doubled in the last 10 years (83 publications reported for the 2005–2010 period, versus 183 for 2015–2020) and include important research and reviews that especially highlight soil protist taxonomic and functional diversity. Due to their link to plant growth, the essential role of soil protists in agroecosystems is also increasingly recognized. In addition to their relation to plant growth, the dynamic response of soil protists to environmental changes has further suggested protist community composition to be indicative of soil condition (Foissner, 1999).

In this article, we aim at providing an overview of the current knowledge on soil protists. We will first present the main adaptations allowing protists to thrive in the soil ecosystem. We then present the biogeography and the ecological roles of soil protists. We further highlight some particularities of the impact of agroecosystems on soil protists and on the relation between protists and plant performance. We also present the potential to apply protists as valuable bioindicators, before concluding with a synthesis and providing perspectives of the immense potential of soil protistology.

Diversity and ecology

We here first briefly review some of the main adaptations of protists required to inhabit the soil ecosystem and provide estimations on protist abundances and diversity in soils. We present the geographical distribution and diversity patterns of terrestrial protists. We finish this section on the ecological role of soil protists as predators and parasites and how they relate to nutrient cycling.

Adaptations to the soil ecosystem

Most soil protists produce viable resistant structures usually refer to as cysts to withstand the unique set of challenges associated with life in soil. The soil is a complex matrix characterized by pores of different dimensions and connectivity. In addition, a distinct set of physicochemical variables, such as pH and moisture content, typically governs soil microbial biodiversity. Water content, especially,

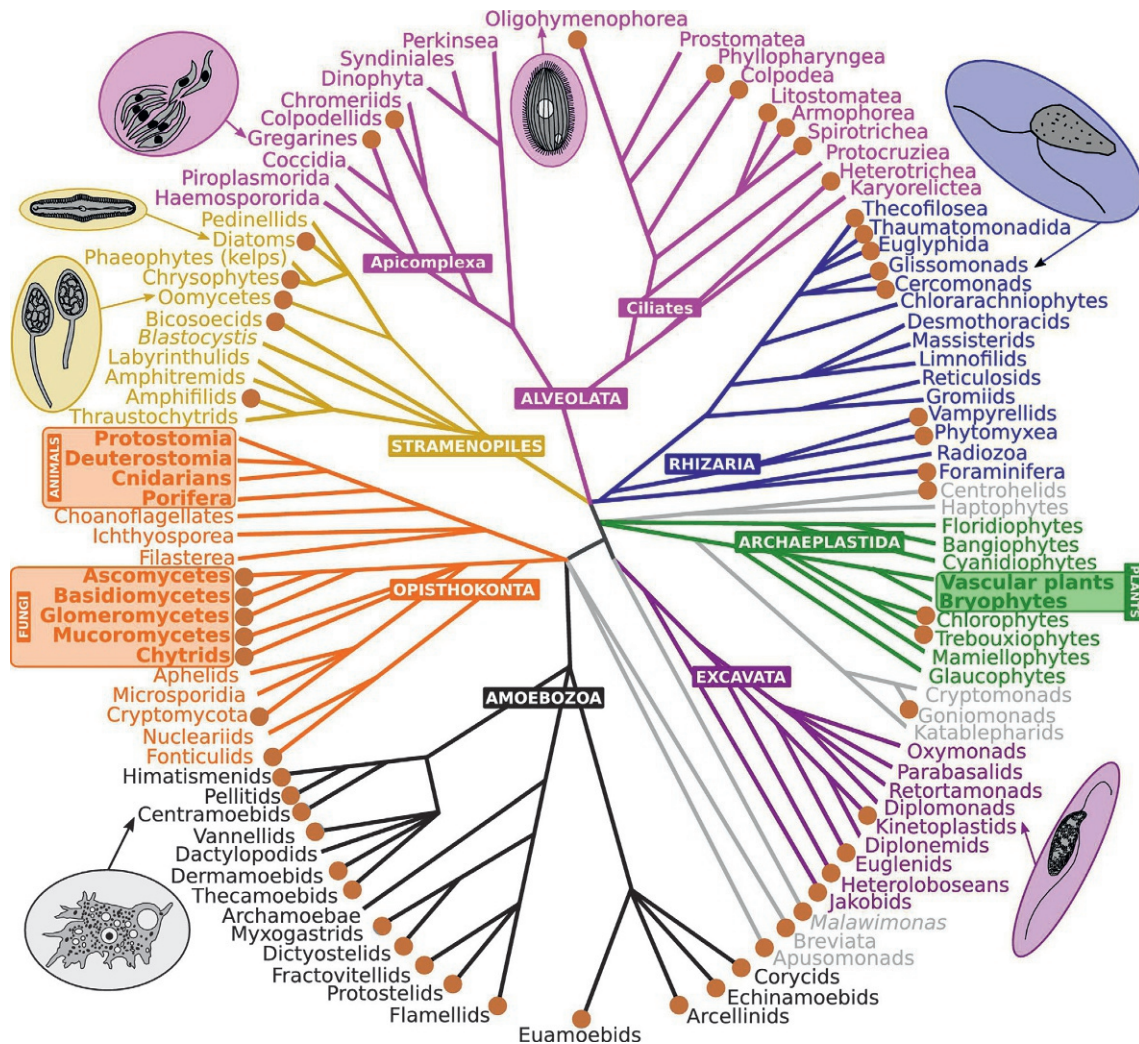


Fig. 2 Schematic representation of the phylogenetic tree of eukaryotes. The brown dots indicate taxonomic groups that are particularly diverse and abundant in soils. Non-protist taxa (land plants, animals, and fungi) are highlighted in boxes. Reproduced with permission from Geisen S, Mitchell EAD, Adl S, Bonkowski M, Dunthorn M, Ekelund F, Fernández LD, Jousset A, Krashevskaya V, Singer D, Spiegel FW, Walochnik J, and Lara E (2018) Soil protists: A fertile frontier in soil biology research. *FEMS Microbiology Review* 42: 293–323. <https://doi.org/10.1093/femsre/fuy006>.

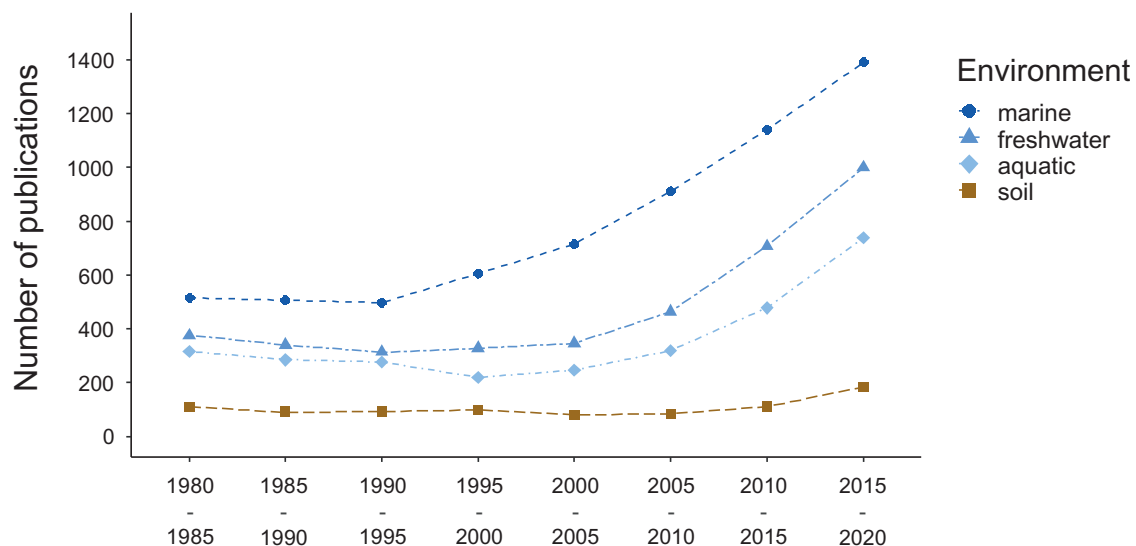


Fig. 3 Overview of studies specifically mentioning protists and the studied habitat (marine, freshwater, aquatic or soil) in the title in 5-year intervals from 1980 to 2020. The literature research has been done on web of science on the 11.04.2021 based on the keywords previously used in the study of Geisen et al. (2017).

can vary dramatically and impose an important pressure on soil protists, which are particularly sensitive to dehydration and rely on water for movement (Geisen et al., 2018). When encysting, the cells typically round up, lose water and secrete a thick wall (the exact composition of which is specific to distinct protist groups) to form a coccoid structure. The longevity of the cysts typically depends on the environmental conditions in which the cells live: cysts formed in arid environments can survive desiccation for decades, while those produced in rainforests only survive for a few weeks or months. In addition to water availability, environmental variables such as pH and salinity as well as biotic variables, such as prey abundance and activity (e.g., production of cyst-inducing compounds), can also induce encystment (Geisen et al., 2018). Cyst formation is thus valuable in a heterogeneous and dynamic environment such as the soil, where moisture content fluctuates as well as the abundance and identity of prey varies.

The physical structure of the soil, which can be described in terms of pore geometry, connectivity and hydration status, further governs the ability of soil organisms to sense and access their food or prey. This is especially true for micro-organisms such as bacteria, protists and nematodes that are bound to the given soil structure and are not able to modify it by, for example, physically creating pores (Erktan et al., 2020). Small pores typically provide valuable refuges for small bacteria against predators too big to access them, but amoeboid protists are particularly well-adapted to reach these small cavities. Amoeboid protists have a flexible body that they can extend in arm-like projections, called pseudopods, to reach such cavities and access their prey (Erktan et al., 2020). A representation of the soil structural complexity and its impact on predator-prey interactions is given in Fig. 4 below.

Abundance and diversity in soils

In terms of abundance, soil protists are generally estimated to range from a few hundred to a million individuals per gram of dry soil (Geisen et al., 2018). With an average of 1.6 gigaton of carbon (Gt C), the soil protist biomass contributes to ca 7% of the terrestrial biomass (the percentage is given after excluding plant biomass and based on the supplementary Table S23 of Bar-On et al., 2018). The protist biomass follows that of soil bacteria with 7 Gt C (32%) and fungi with 12 Gt C (55%) and exceeds that of archaea and animals. Therefore, soil protists contribute far more to the terrestrial biomass than other higher trophic organisms; about $8 \times$ more than arthropods (including all insects on the planet) and $250 \times$ more than nematodes (the most abundant animals).

Protists are estimated to be one of the most diverse eukaryotic groups in soils. In addition, soils have been recently highlighted as hotspots of protist richness compared to freshwater and marine ecosystems (Singer et al., 2021). About 21,000 protist species have been identified in soils, but this has been estimated to represent only about 2% of the true total species richness ($>1,000,000$ species; Geisen et al., 2018). Noteworthy, these numbers are more than 10 times larger than in 2005, when Foissner reported ca 1600 known terrestrial species and an estimated total richness of 3700 species in the first edition of this book. Such an increase in less than two decades is mostly explained by the development of molecular methods allowing for higher taxonomic identification possibilities.

Not all protist groups, however, contribute equally to the soil protist community with some species being relatively rare and others rather dominant. The dominant taxonomic groups in soils are Cercozoa and the Ciliophora, as they together constituted 73% of all reads obtained from 180 soils across six continents (Oliverio et al., 2020). In terms of ecological functions, the most abundant group in soils is further suggested to be formed by predators which can contribute to up to 75% of the total reads (Oliverio et al., 2020; Xiong et al., 2021). The proportion of other functional groups (i.e., phototroph and parasites) is further determined by the ecosystem type. For example, arid ecosystems tend to have relatively more autotrophs, which can even represent up to 40% of the total protist community, while tropical forests are typically characterized by higher abundance of parasites that contribute up to 38% of the total protist community (Mahé et al., 2017; Oliverio et al., 2020).

Geographical distribution and diversity patterns

Species richness is usually not randomly distributed but follows distinct diversity patterns. Such patterns are typically investigated in biogeography, i.e., the study of the distribution of biodiversity over space and time (Martiny et al., 2006). Biogeography of micro-organisms has been the focus of many studies supporting various models going from a random distribution of species to the dominant contribution of either historical events or contemporary environmental conditions (Martiny et al., 2006). The soil protists are no exception as illustrated by the debate between the 'ubiquity model' of Fenchel and Finlay (2004) and the 'moderate endemicity model' of Foissner (2008). While both models assume that most protists are cosmopolitan, Foissner (2008) argued that some species show local endemism. Local endemism is predicted for organisms with relatively high speciation rate and low dispersal potential. High speciation is usually expected for most soil protists as they have relatively short generation times that can thus foster mutations leading to speciation. The dispersal potential, however, may vary between species, as some protists are relatively large and some even lack the ability to produce long-term viable cysts (Foissner, 2008). The complete absence of an organism in a given environment cannot be proven with certainty as this could be due to under-sampling or sampling biases. Yet, there are some well-known examples of flagship species that are easy to observe and study, and appear to have a restricted biogeographic distribution. The relatively large (130–210 μm) testate amoeba *Apodera vas* has been recognized as such a flagship species: it has only been found in South and Central America and Sub-Saharan Africa, but is lacking in North America and Eurasia. Other clear examples of local endemism are found within parasitic protists, such as some *Phytophthora* species, which distribution is directly dependent on the presence of their host. *Phytophthora agathidicida* is, for instance, only found in New Zealand, where its host, the kauri tree (*Agathis australis*), is endemic. Hence, with the enormous taxonomic breadth and corresponding life strategies

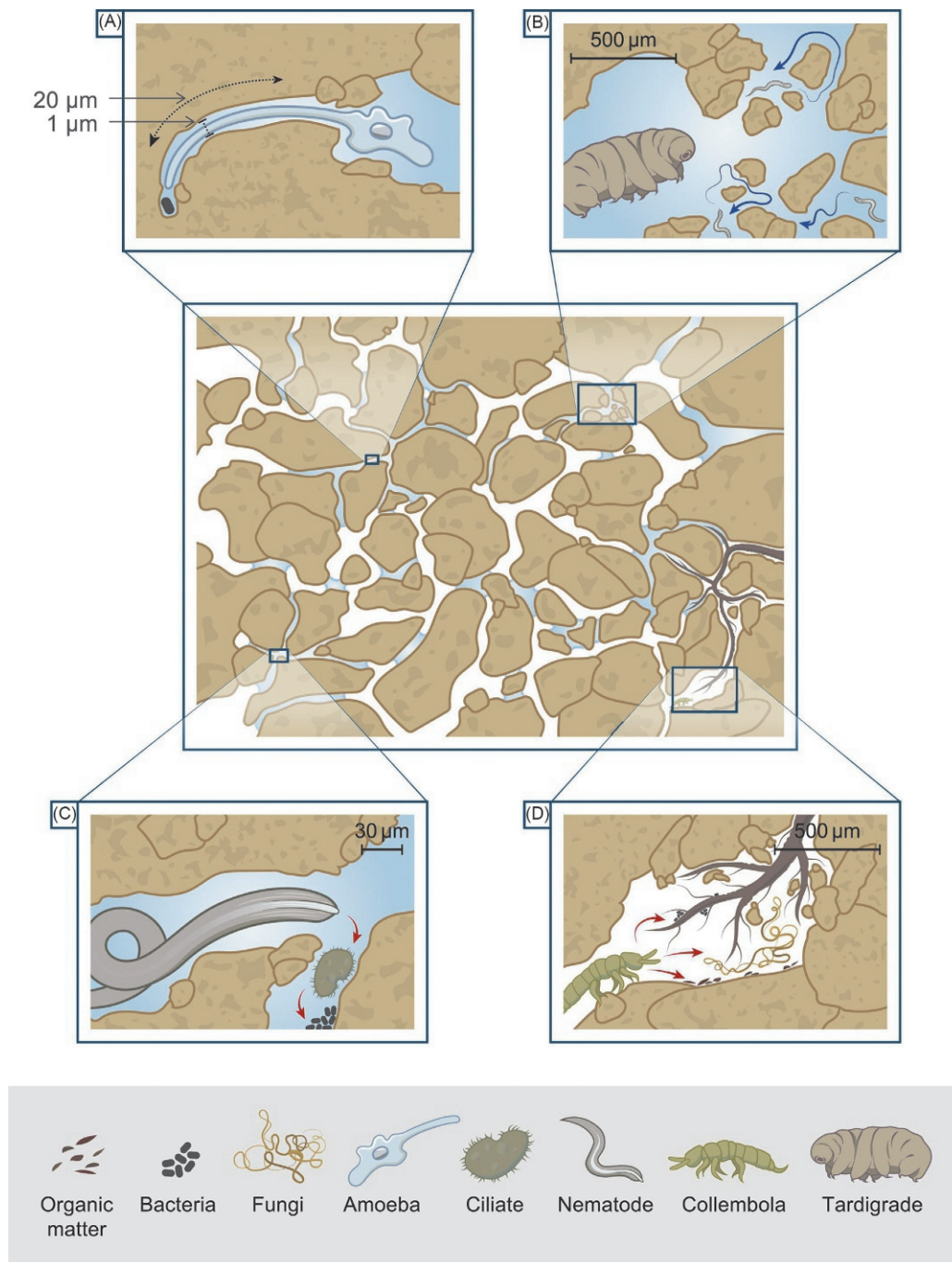


Fig. 4 Representation of the soil structural complexity and its impact on predator-prey interactions for organisms of different sizes. The ability of amoebae to reach prey in small pores is highlighted in panel A, while panel C represents a tri-partite interaction providing access to bacterial prey for nematodes through smaller sized protists (in this case a ciliate) functioning as intermediate predators. Reproduced with permission from Erktan A, Or D, and Scheu S (2020) The physical structure of soil: Determinant and consequence of trophic interactions. *Soil Biology and Biochemistry* 148: 107876. <https://doi.org/10.1016/j.soilbio.2020.107876>.

understood behind the term ‘protist’, there is a clear support for a ‘moderate endemicity model’ where some taxa are globally distributed (i.e., cosmopolitan) and others are locally distributed (i.e., endemic).

Interestingly, community composition of soil protists tends to follow patterns typically reported for macro-organisms such as the latitudinal (tendency for species richness to increase from poles to the equator) or the elevational gradient (tendency for species richness to decrease with elevation) (Geisen et al., 2018). The similarity of these patterns may further be explained by using the conceptual deterministic-stochastic framework and by considering body sizes instead of the dichotomous view of multicellularity vs

unicellularity. Deterministic processes are associated with fitness differences across organism (e.g., survival growth and reproduction) and ecological selection: abiotic and biotic factors determine the presence or absence and relative abundance of organisms. On the contrary, stochastic processes are not the consequences of environmental determined fitness, but are mostly characterized by probabilistic dispersal and random changes in community structure. Bacterial communities were reported to be relatively more influenced than larger organisms by dispersal-based stochastic processes. Larger organisms, including protists, fungi and nematodes were more influenced by selection-based deterministic processes. Within the group of unicellular protists, both stochastic and deterministic processes were reported depending on their size. Smaller protists, like bacteria, were mostly influenced by dispersal-based stochastic processes, while larger protists were mostly influenced by selection-based deterministic processes similarly to fungi and nematodes (Luan et al., 2020).

Ecological relevance

Protist trophic interactions are at the basis of most of their functional roles in the soil ecosystem. Soil protists contribute up to 70% of the total soil respiration originated from microbial eukaryotes and further participate to carbon, nitrogen, phosphorous and silicon cycles, being responsible for 14–66% and 20–40% of the C and N mineralization, respectively (Foissner, 1999).

Predation

Soil protists are mostly known for their role as main predators of bacteria. As such, protists typically reduce the bacterial biomass. Counterintuitively, however, the bacterivorous activity of protists is usually linked with an increase in microbial activity. This discrepancy is generally explained by the increased nutrient turnover that results from protist consumption as well as the removal of less active, senescent bacterial cells, which are more likely to be preyed upon first. The removal of senescent cells lowers competition for space and makes nutrients available for the more active bacterial cells.

Another crucial characteristic of protist consumption patterns is prey selectivity. In the context of bacterivory (the best described feeding preference), protists can sense and discriminate between different prey based on size, shape and chemical properties including specific membrane-bound proteins but also volatiles compounds (VOCs). Such prey discrimination and selectivity cause protists to modify bacterial community composition, favoring a specific subset of species. Noteworthy, distinct food preferences have been observed for phylogenetically closely related protists with similar morphotypes suggesting that each species has its own food preferences. Such preferential feeding is further unlikely to be very specific to only few bacterial species as the heterogeneity of the soil environment offers rather unpredictable prey community composition. It is more reasonable to expect predator protists to feed on most of the bacteria present in soils with, however, some bacterial species better fitting their nutritional needs and thus being preferred when given the choice. Thanks to such distinct preferential feeding with low overlap between protist species, the protist community is essential in maintaining a diverse taxonomical and functional bacterial community (Gao et al., 2019).

These predator-prey interactions are further likely to constantly co-evolve. As such, bacteria have evolved different strategies to escape or combat predators, including changes in bacterial cell morphology, colony formation, escaping movement and production of toxic compounds (Matz and Kjelleberg, 2005). Many toxic, anti-microbial compounds have received a lot of attention as they were primarily recognized and studied for their potential to fight various pathogens. Indeed, an important source of antibiotics comes from such research and naturally occurring anti-microbial compounds are further increasingly considered as alternatives to synthetic pesticides. These anti-microbial compounds have typically a rather untargeted mode of action and can inhibit various eukaryotes. As such, numerous compounds produced by soil pseudomonads that have long been known for their antifungal activity, such as 2,4-diacetylphloroglucinol (DAPG), hydrogen cyanide, pyrrolnitrin or phenazines, also help protect bacteria against protist predation. This observed overlap between pathogen inhibition and inhibition of predator protists highlights the role of predation pressure on the very appearance and maintenance of such biocontrol traits in the microbial community. Soil protists may thus play a valuable role in keeping soils healthy with low (plant-) pathogen densities by promoting biocontrol traits (Amacker et al., 2020).

While most studies focus on bacterivorous protists, free-living heterotrophic protists also feed on other organisms such as fungi, nematodes, and other protists (Geisen et al., 2018). Some protists became even highly specialized on a specific prey type. For example, the Grossglockneriidae, a distinct family of ciliates, feed solely on fungi: they have an adapted structure to penetrate fungal hyphae, while their cytostome does not allow the ingestion of a prokaryotic cell. In some soils, these ciliates and other obligate fungivorous protists represent a substantial proportion of the protist population likely to play important functional roles similarly to the one typically connected to bacterivorous protists (i.e., nutrient cycling and modification of the prey community composition). In addition to these highly specialized feeders, many presumably bacterivorous protists have been shown to equally ingest and grow on yeasts, as well as on spores of hyphal-forming fungi. Many protists can thus better be described as plurivorous as they can feed on different prey such as bacteria, yeasts and spores. Some protists may even be described as omnivores as they can prey on bacteria, yeasts, spores but also on higher trophic organisms such as nematodes. For example, the small testate amoebae (shell length: 17 µm, shell width: 15 µm) of the genus *Cryptodifflugia*, presumably bacterivorous, were not only able to feed and grow on yeasts, but they were even shown to collaboratively and effectively attack and prey on bacterivorous nematodes following a pack hunting strategy (Geisen et al., 2016).

Protists are eventually themselves preyed upon, channeling nutrients from small prey in unreachable pores to higher trophic predators too big to access the micro-niches (see also the Fig. 3, panel C displays such tripartite interaction). Predators of protists include a range of organisms such as nematodes, collembola, earthworms and insect larvae.

Parasites

Another important heterotrophic category of protists is 'parasite.' In soils, the most notorious group of parasites are the apicomplexans, which also include the infamous human pathogens *Plasmodium* spp., causative agents of malaria. In soils another taxonomic group of the apicomplexans, the Gregarines, is particularly diverse and abundant and they predominantly infect arthropods and other invertebrates. Parasitic protists are present in most soils, and they can contribute to more than half of the total protist diversity in tropical rainforests. In such environment, the obligate parasitic apicomplexans are suggested to play a role in controlling animal populations and local animal diversity (Mahé et al., 2017). Thus, similar to free-living, heterotrophic protists, parasitic protists might be an essential part of the ecosystem, modulating the presence and abundance of host organisms.

Some important plant pathogens are also found among soil protists. Most of the known plant pathogens are oomycetes (fungi-like protists belonging to the supergroup of the Stramenopiles). Plant-pathogenic protists have been and are still the cause of severe economic damages linked to crop losses. The oomycete *Phytophthora infestans* is a well-known infamous example as it destroyed potato fields in the mid-1840 leading to a general food crisis in Northern Europe, particularly affecting countries whose diet was primarily potato-based, such as Ireland. Especially in moist conditions, *P. infestans* is still the cause of major yield destruction if left uncontrolled. Other well-known oomycete plant pathogens include the downy mildews, *Pythium* spp. (e.g., damping off disease), *Albugo* spp., and *Plasmopara viticola*. In addition to oomycetes, the supergroup of the Rhizaria also comprises a distinct group of obligate biotrophic parasites, the phytomyxids, which include *Plasmodiophora brassicae* (clubroot disease in Brassicaceae) and *Spongospora* spp. (powdery scab in potatoes) (Geisen et al., 2018).

Several protists found in soils can also be opportunistic human pathogens. These pathogens may infect humans via ingestion of contaminated food or water, or via contaminated hands. Body openings such as the nose and eyes as well as skin lesions further offer possibilities for soil-borne protists to infect humans. Examples of such opportunistic pathogens include *Acanthamoeba* spp. causing keratitis and granulomatous amoebic encephalitis, *Balamuthia mandrillaris* and *Sappinia* spp. both also causing granulomatous amoebic encephalitis and *Vermamoeba vermiformis* causing eye infections. In addition to the direct pathogenicity of some protists, a number of amoebae have been reported to act as "Trojan horses" by hosting pathogenic bacteria, fungi and viruses. By doing so, the amoebae indirectly help the pathogens to evade measures to eradicate them or prime them to attack their host even more aggressively. A well-documented example is the bacteria *Legionella pneumophila* that can cause serious lung infection and that has been shown to be more invasive for human cells after passage through amoebae. Thus, soil is also an important reservoir and vector of human pathogens (Geisen et al., 2018).

Parasitic protists are an important component of soils with the potential to impact the presence and abundance of their host with potential consequences for ecosystem functioning. In addition, parasitic protists directly participate in nutrient cycling: by parasitizing and eventually killing larger, macroscopic hosts, they release bioavailable nutrients into the surrounding soil.

Nutrient cycling: C, N, P and Si

As highlighted above, soil protists are mostly known for their contribution to nutrient cycling as predators, but they participate to the carbon cycle at every level, from fixation to decomposition. Indeed, photosynthetic protists do occur as free-living organisms in the upper layers of soils (e.g., *Bacillariophyta*, *Chrysophyceae*, *Xanthophyceae*) and in symbiotic associations, such as with fungi in lichens (e.g., *Trebouxiphyceae*). Some soil protists even contribute to the decomposition of organic matter. Free-living, saprotrophic oomycetes, for example, decompose organic matter via the production of exoenzymes and many protists are able to take up nutrients directly from their environment via osmotrophy. Thanks to osmotrophy, *Acanthamoeba* spp. can, for example, be grown solely on nutrient-rich media without any addition of bacterial prey, allowing the generation of axenic cultures (Geisen et al., 2018).

The predator activity of soil protists is intimately linked to nitrogen. As soil protists respire, carbon is released; Because the C:N ratio of prey bacteria and fungi is either similar or lower than the one of protists, the excess nitrogen needs to be excreted by the predator protists. The important role of predators in the release of nitrogen and nutrients in general from the microbial biomass was first described under the concept of "microbial loop" in aquatic systems. A similar process was reported in soils and in particularly in the rhizosphere soil (i.e., soil under influence of plant roots). While the rhizosphere constitutes a carbon-rich environment due to the release of root exudates, nitrogen is locked up in the microbial biomass, unavailable for plants and thus limiting their growth. The presence and activity of protists is necessary to release nitrogen from the microbial biomass. About one third of the consumed nitrogen is liberated as such and becomes available to other microorganisms and plants. Indeed, an increase of up to 50% of dry weight and nitrogen content was observed for plants grown in soils containing both bacteria and protists, compared to soils lacking protists (Clarholm, 1985).

Protist activity is reported to support plant symbiotic fungi. Arbuscular mycorrhizal fungi are such plant symbionts and improve the phosphorous uptake of their host. These arbuscular mycorrhizal fungi are, however, unable to mineralize nitrogen and depend on nitrogen released in the surrounding via, for example, the microbial loop. This way, the predator activity of protists is further related to the phosphorous cycle by supporting the activity of mycorrhizal fungi (Geisen et al., 2018).

Some shell-bearing protists participate in the silicon cycle: terrestrial diatoms and testate amoebae take up silicon to build their shell, being both high silica consumers and suppliers. In some ecosystems the contribution of testate amoebae to the silica cycle is even comparable to that of higher plants, usually considered as the main contributors (Aoki et al., 2007).

In summary, soil protists contribute to every step of the carbon cycle, release nitrogen for other organisms via the soil microbial loop, thus indirectly supporting mycorrhizal transfer of phosphorous to plant, and also participate in the silicon cycle.

Soil protists in agricultural systems

Impact of agricultural management on soil protists

Protists inhabit every soil type and basically every habitat on Earth. Therefore, they are also an inherent part of agricultural soils. Agricultural soils are modified by humans with constant management to produce food, feed, fiber and fuel. As such, soil biodiversity including protists are affected in anthropogenic agricultural soils.

The first and most evident aspect managed by humans is the vegetation type, which itself strongly influences protist community composition. Each plant releases root exudates that attract a distinct microbial community comprising of bacteria and fungi, and these micro-organisms further attract predatory protists (Geisen et al., 2018). A diverse vegetation is expected to correlate with a large below-ground diversity as each plant would attract a distinct microbial community. In contrast, in human-managed areas, monocultures are expected to reduce protist diversity compared with plant rich habitats, which are expected to support and maintain greater biodiversity.

Various methods are used in the control of plant-pathogens. Different types of pesticides are thus commonly applied to protect plants against pathogens. Such pesticides are known to affect many non-target organisms, including non-pathogenic soil protists. The effects of pesticides are, however, not straightforward: some compounds directly inhibit protist viability (e.g., the fungicide azoxystrobin can reduce the density of ciliates and flagellates), while others such as glyphosate can stimulate bacterial growth and potentially further support protist growth. In addition, the effects of pesticides typically vary between protist species and are dose dependent, i.e., depend on the concentrations of the tested compounds (Lo, 2010). While it is clear that chemical compounds can affect soil protists, the extent of the impacts and the ecological consequences remain largely unknown.

To increase yield, the addition of fertilizer is also common in agricultural soil and protists are reported to be particularly affected by such addition. Such effects are mostly indirect: protist diversity was, for example, reduced following modifications of abiotic properties (Zhao et al., 2019). Which specific taxa was favored or reduced was related to the soil type (red soil, fluvo-aquic soil or black soil) and the sampling season (summer or autumn); bacterial and fungal communities only varied between soil types but not with season nor due to fertilizer application. These findings highlight a particularly dynamic response of the protist population to environmental change.

In general, any type of soil management is expected to affect the protist community composition through modifying soil structure and physicochemical properties as well as the above- and below-ground biodiversity. It is also increasingly recognized that soil protists may be more strongly affected compared with lower trophic levels, such as bacteria and fungi, thus deserving a particular focus. Indeed, due to their dynamic response to environmental changes and their pivotal role in the food web, soil protists possess a unique potential to monitor the effect of agricultural practices on the soil microbial biodiversity and functionality.

Soil protist to support plant performance

A better understanding of the protist ecology in natural and managed soils could lead to improved soil health. Indeed, plant performance is intimately linked to its microbiome. While the focus is usually centered on bacteria and fungi, protists are also essential partners of the plants.

In general, the presence of an abundant and diverse protist community should support a greater nutrient turnover (due to the microbial loop in soil) and a diverse prey community (due to preferential feeding). While predation by protists typically induces changes in the bacterial community composition due to preferential feeding, it is also expected to generally prevent the dominance of few fast-growing bacteria, thus allowing for a more diverse bacterial community composition including rare taxa. Supporting diverse and abundant protist communities thus enables a more efficient use of organic fertilizers and should favor a greater functional diversity. Better ecological knowledge and technological development may even allow the enhancement of specific, targeted microbiome functions by exploiting the feeding preferences of protists. The application of protists has been already shown to increase the survival and plant-beneficial activity of different plant-growth promoting bacteria (Gao et al., 2019).

The increased survival of plant-growth beneficial bacteria in presence of predators is likely related to their ability to either escape their predators (e.g., high speed motility, biofilm formation) or to fight back via the production of anti-microbial compounds. Anti-microbial compounds produced by bacteria to inhibit their predators also affect many other eukaryotes (incl. Plant pathogens), therefore acting as biocontrol agents. In addition to benefiting plant growth, soil protists may thus be used to support plant health. The presence of predator protists can affect the ability of the soil microbiome to reduce plant pathogens in multiple ways. First, presence of protists induces the production of anti-microbial compounds. Second, protists may further support the successful establishment of biocontrol bacteria in soils because the predator protist is expected to preferentially feed on undefended, non-toxic bacteria, thus lowering the competition in favor of the biocontrol agents. Third, protists may indirectly prime plant immune response as a response to over-produced antimicrobial compounds as a bacterial defense strategy. Finally, protists can directly contribute to disease suppression by consuming or parasitizing pathogens, reducing their survival in soil and potentially protecting plants (Gao et al., 2019).

Soil protist as bioindicators

Because of their link to plant growth, their unique nature (e.g., rapid growth rate, small size, delicate external membranes) and position in the food web, soil protists have been suggested as indicative of soil status and functioning. Such potential as bioindicators was first highlighted in protist groups displaying specific morphological traits enabling easier identification. Thus, ciliates which are in general more easily counted and identified than other protists, could be used as bioindicators in polluted and degraded soils (Foissner, 1999). Another group of protists was recognized as valuable indicator in soils: the shell-bearing testate amoebae. They are used to distinguish between different humus types (i.e., mor, moder and mull) and further to infer other environmental changes, such as pollution, land use, fire history, nutrient status, and even as forensic indicators for dating the time of death. In addition, thanks to the relative high resistance of their shells to decay and the presence of distinct features enabling identification, testate amoebae are often used to reconstruct the water table in peatlands. Such studies enable the reconstruction of the paleo-environment to better understand effects of climate change (Geisen et al., 2018). With the technological advancement of the last decades, it has become possible to study the response of not only ciliates and testate amoebae, but of the general soil protist population. It was demonstrated that not only do protists respond dynamically to environmental changes at the community-level, but they are even more sensitive than lower trophic organisms such as bacteria and fungi (Zhao et al., 2019).

In addition to their indicative potential for soil status, soil protists are suggested as model organisms for ecotoxicological bioassays to predict the effect of various biocides, including pesticides used in agriculture and other pollutants that unintentionally reach the soils. Some bioassays have been used to test, for example, the effect of pesticides on *Dictyostelium discoideum* or the influence of heavy metals on ciliates (Geisen et al., 2018). Such bioassays using protists have already been developed and are used in aquatic systems to evaluate the toxicity of various chemicals (e.g., with the freshwater ciliate *Tetrahymena pyriformis* or various green algae), but to our knowledge none are routinely used to assess impacts of biocides in soils.

While the validity of methods using soil protists is usually accepted, practical applications are still limited by the difficulties related, for example, to time-consuming morphological identifications. In addition to the time limitation, only few skilled specialists can accurately identify the protists, thus strongly restricting its usage. Simplified direct observations, approaches based on functional traits and development of new molecular screening tools are likely to be necessary to enable the routine application of soil protists as indicators of soil health (Geisen et al., 2018).

Conclusion

Soil protists are now increasingly recognized as an abundant, diverse and essential component of the soil ecosystem. Evidence for both global and endemic distributions have been reported, as well as various diversity patterns related to environmental variables or biotic interactions. Predator and parasitic protists determine their prey or host community composition and activity. At the same time, protists drive the cycling of nutrients, including carbon, nitrogen, phosphorous and silicon. Finally, soil protists are an inherent component of agricultural soils, dynamically responding to soil managements and having the potential to be both detrimental and beneficial to plants, animals and humans. Understanding the ecology of soil protists will help us supporting a healthy and diverse soil ecosystem.

The true diversity of soil protists, however, remains unknown. Identifying and describing this huge taxonomical diversity is thus still an important agenda. In addition to the taxonomy, the ecological role of protists has been revealed as essential for soil status, but many aspects are still ambiguous. For example, it is clear that protists induce distinct modification of their bacterial prey community composition, but the mechanisms behind this apparent preferential feeding have not yet been fully identified. This prevents us from fully harvesting the potential of protists in supporting a beneficial soil microbiome. Finally, it is also essential to recognize and investigate the unique contribution of each protist species and start linking them to functional traits that enable approaches focused on their ecological roles. Altogether, we conclude that protists can be considered as a rising star in soil biodiversity research with future studies undoubtedly deciphering their full complexity in the ecosystem.

References

- Amacker N, Zhilei G, Betina CA, Latz E, Kowalchuk G, Valverde CF, Jousset A, and Weidner S (2020) Biocontrol traits correlate with resistance to predation by protists in soil *Pseudomonads*. *Frontiers in Microbiology* 11: 1–13.
- Aoki Y, Hoshino M, and Matsubara T (2007) Silica and testate amoebae in a soil under pine–oak forest. *Geoderma* 142: 29–35. <https://doi.org/10.1016/j.geoderma.2007.07.009>.
- Bar-On Y, Phillips R, and Milo R (2018) The biomass distribution on Earth. *PNAS* 115(25): 6506–6511. www.pnas.org/cgi/doi/10.1073/pnas.1711842115.
- Burki F, Roger AJ, Brown MW, and Simpson AGB (2020) The new tree of eukaryotes. *Trends in Ecology & Evolution* 35: 43–55. <https://doi.org/10.1016/j.tree.2019.08.008>.
- Ciarholm M (1985) Interactions of bacteria, protozoa and plants leading to mineralization of soil nitrogen. *Soil Biology and Biochemistry* 17: 181–187.
- Erktan A, Or D, and Scheu S (2020) The physical structure of soil: Determinant and consequence of trophic interactions. *Soil Biology and Biochemistry* 148: 107876. <https://doi.org/10.1016/j.soilbio.2020.107876>.
- Fenchel T and Finlay BJ (2004) The ubiquity of small species: Patterns of local and global diversity. *BioScience* 54: 777–784. [https://doi.org/10.1641/0006-3568\(2004\)054\[0777:tuossp\]2.0.co;2](https://doi.org/10.1641/0006-3568(2004)054[0777:tuossp]2.0.co;2).
- Foissner W (1999) Soil protozoa as bioindicators: pros and cons, methods, diversity, representative examples. *Agriculture, Ecosystems and Environment* 74: 95–112.
- Foissner W (2008) Protist diversity and distribution: some basic considerations. *Biodiversity and Conservation* 17: 235–242. <https://doi.org/10.1007/s10531-007-9248-5>.

- Gao Z, Karlsson I, Geisen S, Kowalchuk G, and Jousset A (2019) Protists: puppet masters of the rhizosphere microbiome. *Trends in Plant Science* 24: 165–176. <https://doi.org/10.1016/j.tplants.2018.10.011>.
- Geisen S, Koller R, Hünninghaus M, Dumack K, Ulrich T, and Bonkowski M (2016) The soil food web revisited: Diverse and widespread mycophagous soil protists. *Soil Biology and Biochemistry* 94: 10–18. <https://doi.org/10.1016/j.soilbio.2015.11.010>.
- Geisen S, Mitchell EAD, Wilkinson DM, Adl S, Bonkowski M, Brown MW, Fiore-Donno AM, Heger TJ, Jassey VEJ, Krashevska V, Lahr DJG, Marcisz K, Mulot M, Payne R, Singer D, Anderson OR, Charman DJ, Ekelund F, Griffiths BS, Rønn R, Smirnov A, Bass D, Belbahri L, Berney C, Blandenier Q, Chatzinotas A, Clarholm M, Dunthorn M, Feest A, Fernández LD, Foissner W, Fournier B, Gentekaki E, Hájek M, Helder J, Jousset A, Koller R, Kumar S, La Terza A, Lamentowicz M, Mazei Y, Santos SS, Seppey CVW, Spiegel FW, Walochnik J, Winding A, and Lara E (2017) Soil protistology rebooted: 30 fundamental questions to start with. *Soil Biology and Biochemistry* 111: 94–103. <https://doi.org/10.1016/j.soilbio.2017.04.001>.
- Geisen S, Mitchell EAD, Adl S, Bonkowski M, Dunthorn M, Ekelund F, Fernández LD, Jousset A, Krashevska V, Singer D, Spiegel FW, Walochnik J, and Lara E (2018) Soil protists: a fertile frontier in soil biology research. *FEMS Microbiology Review* 42: 293–323. <https://doi.org/10.1093/femsre/fuy006>.
- Lo C-C (2010) Effect of pesticides on soil microbial community. *Journal of Environmental Science and Health. Part. B, Pesticides, Food Contaminants, and Agricultural Wastes* 45(5): 348–359. <https://doi.org/10.1080/10934520903467873>.
- Luan L, Jiang Y, Cheng M, Dini-Andreote F, Sui Y, Xu Q, Geisen S, and Sun B (2020) Organism body size structures the soil microbial and nematode community assembly at a continental and global scale. *Nature Communications* 11: 6406. <https://doi.org/10.1038/s41467-020-20271-4>.
- Mahé F, de Vargas C, Bass D, Czech L, Stamatakis A, Lara E, Singer D, Mayor J, Bunge J, Sernaker S, Siemensmeyer T, Trautmann I, Romac S, Berney C, Kozlov A, Mitchell EAD, Seppey CVW, Egge E, Lentendu G, Wirth R, Trueba G, and Dunthorn M (2017) Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nature Ecology and Evolution* 1: 0091. <https://doi.org/10.1038/s41559-017-0091>.
- Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR, Morin PJ, Naeem S, Øvreås L, Reysenbach A-L, Smith VH, and Staley JT (2006) Microbial biogeography: Putting microorganisms on the map. *Nature Reviews Microbiology* 4: 102–112. <https://doi.org/10.1038/nrmicro1341>.
- Matz C and Kjelleberg S (2005) Off the hook – how bacteria survive protozoan grazing. *Trends in Microbiology* 13: 302–307. <https://doi.org/10.1016/j.tim.2005.05.009>.
- Oliverio AM, Geisen S, Delgado-Baquerizo M, Maestre FT, Turner BL, and Fierer N (2020) The global-scale distributions of soil protists and their contributions to belowground systems. *Science Advances* 6: eaax8787. <https://doi.org/10.1126/sciadv.aax8787>.
- Singer D, Seppey CVW, Lentendu G, Dunthorn M, Bass D, Belbahri L, Blandenier Q, Debroas D, de Groot GA, de Vargas C, Domaizon I, Duckert C, Izaguirre I, Koenig I, Mataloni G, Schiaffino MR, Mitchell EAD, Geisen S, and Lara E (2021) Protist taxonomic and functional diversity in soil, freshwater and marine ecosystems. *Environment International* 146: 106262. <https://doi.org/10.1016/j.envint.2020.106262>.
- Xiong W, Jousset A, Li R, Delgado-Baquerizo M, Bahram M, Logares R, Wilden B, de Groot GA, Amacker N, Kowalchuk GA, Shen Q, and Geisen S (2021) A global overview of the trophic structure within microbiomes across ecosystems. *Environment International* 151: 106438. <https://doi.org/10.1016/j.envint.2021.106438>.
- Zhao Z-B, He J-Z, Geisen S, Han L-L, Wang J-T, Shen J-P, Wei W-X, Fang Y-T, Li P-P, and Zhang L-M (2019) Protist communities are more sensitive to nitrogen fertilization than other microorganisms in diverse agricultural soils. *Microbiome* 7: 33. <https://doi.org/10.1186/s40168-019-0647-0>.

Fungi including mycorrhizal fungi

F Buscot, Helmholtz Centre for Environmental Research (UFZ), Halle (Saale), Germany

© 2023 Elsevier Ltd. All rights reserved.

This is an update of L.M. Egerton-Warburton, J.I. Querejeta, M.F. Allen, S.L. Finkelman, Mycorrhizal Fungi, Reference Module in Earth Systems and Environmental Sciences, Elsevier, 2013, ISBN 9780124095489, <https://doi.org/10.1016/B978-0-12-409548-9.05226-X>.

Updated from the chapter written by K. Ritz Emeritus Professor of Soil Ecology, University of Nottingham, Sutton Bonington Campus, Sutton Bonington, Leicestershire, LE12 5RD, UK.

Introduction	49
Classification	49
Biology	50
Morphology and form	50
Physiology and reproduction	50
Biotic interactions	52
Fungal associations with other microbes	52
Fungal associations with plants	52
Mutualism	52
Parasitism	52
Fungal associations with fauna	54
Microbiome and holobiome	54
Fungal contributions to soil-system formation and function	55
Soil formation and pedogenesis	55
Nutrient cycling	55
Soil architecture	55
Methods of study	56
Isolation methods	56
Direct observation	57
Physiological and molecular methods	58
Activity	58
Exploitation of soil fungi	58
References	59

Abstract

The morphological, reproductive and physiological traits of fungi are presented, followed by their contrasting biotic interactions with organisms living in soils, particularly mycorrhiza (the symbiotic associations with plant roots) that are crucial to the functioning of terrestrial ecosystems. Their broad range of metabolic capacities indicate that fungi are a main group of the soil microbiome, importantly participating in pedogenesis and element cycling, and of the plant holobiont. Recent integration of microbiological, physiological, new molecular ecological techniques and bioinformatics has dramatically increased knowledge of soil fungi, their role in soil functioning and potential development of more sustainable soil and land-use practices.

Introduction

Fungi are ubiquitous in soils, but comprise a varying proportion of the microbial biomass in different ecosystems and soil horizons. They tend to dominate over bacteria in soils rich in organic matter but constitute a smaller proportion in intensively managed and mineral soils. They are involved in a plethora of functional roles, encompassing biological, chemical, and physical interactions, and are of great ecological and economic significance. The study of fungi is termed mycology (after *mukēs*, Greek for ‘fungus’).

Classification

The ‘true fungi’ belong to the kingdom of Fungi and are eukaryotic, exclusively heterotrophic organisms with cell walls that contain chitin or chitosan as a major constituent for almost all taxa, and typically have a thread-like hyphal growth-form, although unicellular forms are common in the yeasts (Richards et al., 2017). Recent multigene analyses separate

19 phyla: Aphelidiomycota, Ascomycota, Basidiobolomycota, Basidiomycota, Blastocladiomycota, Calcarisporiellomycota, Caulochytriomycota, Chytridiomycota, Entomophthoromycota, Entorrhizomycota, Glomeromycota, Kickxellomycota, Monoblepharomycota, Mortierellomycota, Mucoromycota, Neocallimastigomycota, Olpidiomycota, Rozellomycota and Zoopagomycota (Wijayawardene et al., 2020). The former Fungi imperfecti, lacking a sexual phase, are now ascribed to different phyla. Representatives from most phyla are commonly found or can be expected in soils, where fungal biodiversity and density are greater than in aboveground ecosystem compartments and vary differently among geographic gradients than do other taxa (e.g. plants) (Teder et al., 2014).

Biology

Morphology and form

A fundamental characteristic of the filamentous (eucarpic) fungi is their growth-form, which consists of thread-like hypha, growing by apical extension and periodic branching to form a mycelium that permeates the environment explored by the fungus (Fricker et al., 2017) and Fig. 1A–E. The diameters of individual hyphae vary according to age, species, and nutritional conditions but are typically 3–10 µm. In extreme cases, up to 2 km of hyphae per gram of soil have been measured in some soils; a typical cultivated arable soil contains several meters per gram. Mycelia are indeterminate structures that grow in three dimensions and therefore conceptually are difficult to size. In soil, they may occupy almost any volume, and the largest organisms on the planet are probably mycelia of *Armillaria bulbosa*, individual clones of which are reported to extend continuously over tens of hectares in North American forests (Fig. 1G) (Smith et al., 1992). In Asco- and Basidiomycota, hyphae are generally compartmentalized by cross-walls (septa), which occur at varying intervals depending on species (Fig. 1B). Septa may be complete or perforated to varying degrees and play a role in regulating the passage of materials, including organelles, between adjacent compartments. In particular, tubular vacuoles, extend through pores and by their pulsating activity, they control directed transfers of solutes between hyphal compartments. If mycelia are fractured, septal pores become occluded, which prevents leakage of cytoplasm from the open ends. In certain phyla such as Glomeromycota and Mucoromycota, mycelia have almost no septa and are largely coenocytic (multinucleated).

Hyphae may aggregate to form higher-order structures such as strands (cords; Fig. 1E) or rhizomorphs, which enable long-distance foraging (decameters in the case of woodland basidiomycetes) (Agerer, 2001). Hyphal aggregation and differentiation can also produce morphologically complex macrostructures such as the intricate fruiting bodies (basidiocarps) of some Basidiomycota or ascocarps of some Ascomycota.

Those fungi that do not form hyphae (predominantly yeasts) have a holocarpic growth form, comprising individual cells that replicate by budding or binary fission. Some fungi are dimorphic and can switch between holocarpic and eucarpic growth forms depending on environmental conditions (e.g., *Mucor hiemalis*).

Physiology and reproduction

All fungi are obligate heterotrophs, i.e., they utilize fixed (organic) C sources as substrate (Carlile and Watkinson, 2001). Across the kingdom an extremely wide range of compounds can be utilized, ranging from C1 compounds (methylotrophic yeasts) through to lignin and industrially produced polyphenols – almost any natural or synthetic ring macromolecule occurring in soils can be degraded by some fungi. A concomitantly wide range of enzymes is secreted locally by hyphae, and thus targeted metabolic capabilities are expressed at the vicinity of or within specific substrates, which is regulated by nutrient availability in each micro-niche (Schinner et al., 2012). This makes mycelia highly efficient and flexible nutrient acquisition systems, well adapted to the heterogeneity of soils in time and space (Dighton et al., 2019). Respiration can be aerobic or anaerobic, with obligate or facultative types of both forms occurring. Some fungi obtain their nutrition from other living organisms (plants, animals, other fungi) via parasitic or mutualistic associations, whilst others are saprotrophs (Carlile and Watkinson, 2001). Many soil fungi are very efficient scavengers of nutrients and grow sparsely but effectively under the oligotrophic conditions that prevail in many soils. A wide variety of secondary metabolites are produced by soil fungi, including many volatile organic compounds (VOCs), which may act as signals or substrates as they diffuse through pore networks (Macheleidt et al., 2016). If fungal spores or mycelial fragments are added to nonsterile soils, they rarely germinate. This phenomenon, known as fungi stasis (or mycostasis) is probably due to an inadequate supply of available nutrients (the extant microflora normally having depleted the microbially available nutrient pool), the presence of inhibitory (antibiotic) compounds produced by the existing microflora or a combination of both (Garbeva et al., 2011).

Reproduction can be sexual or asexual, and a variety of mating systems occur (Hawker, 2016). Many fungi have genetic systems that prevent mating between genetically identical cells or mycelia. If hyphae encounter each other, they may fuse (anastomose). If they are somatically compatible, the fusion results in the formation of larger mycelial networks, with attendant functional consequences in terms of spatial exploration, activity, translocation, and combativeness. Somatic incompatibility, however, results in a rejection response, and the cytoplasm in fused compartments is destroyed. Fungal propagules generally take the form of asexual or sexual spores, sclerotia (dense, heavily pigmented hyphal aggregations), or hyphal fragments. Some species sporulate very profusely (e.g., *Penicillium*), others have never been known to produce any form of spore (e.g., *Rhizoctonia solani*). In soil, propagules are dispersed by mass flow in soil water, carried by animals internally or externally to their bodies, or via physical (often faunally mediated) soil movement.

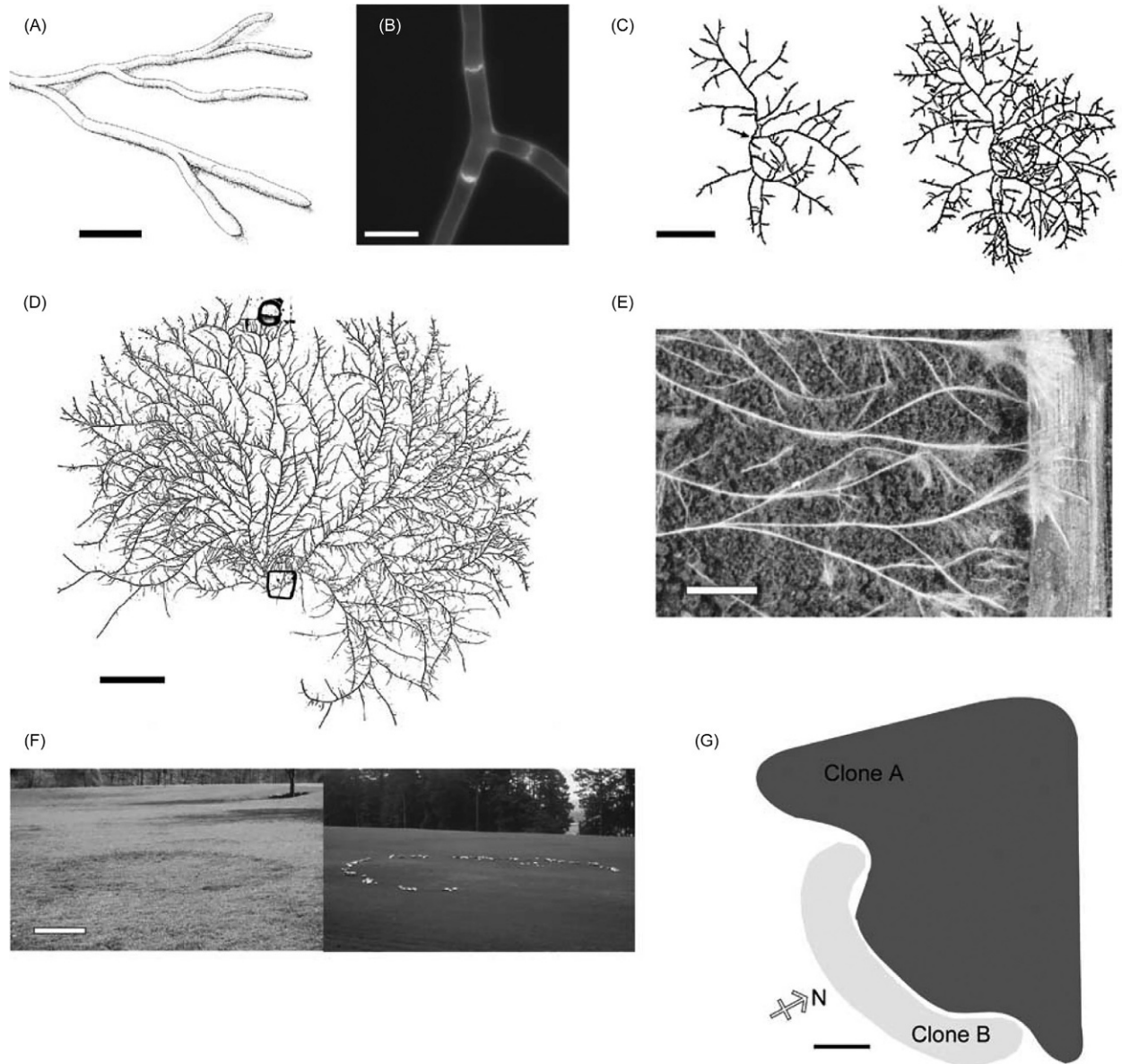


Fig. 1 Aspects of mycelia of soil fungi at different scales. (A) stylization of fungal hyphae, showing branching habit and morphology of tips. Scale bar 50 µm; (B) micrograph of hyphae of *Rhizoctonia solani* showing branch and septa. Scale bar 50 µm; (C) mycelium of *Trichoderma viride* 44 h (left) and 48 h (right) after germination, grown on spatially uniform nutrients. Scale bar 100 µm; (D) mycelium of *T. viride* grown on spatially nonuniform nutrients (point of inoculation is denoted by a square, localized nutrient source denoted by a circle). Scale bar 5 mm; (E) colonizing fans at tips of foraging cords of *Agrocybe gibberosa* when a piece of straw (to right of image) is encountered. Scale bar 5 mm; (F) fairy ring in grassland, resulting from enhanced mineralization of soil nutrients by subterranean mycelium, expressed as enhanced plant growth (left image) and following basidiocarp emergence (right image). Scale bar 1 m; (G) mapped extent of two clones of *Armillaria bulbosa* in a Michigan forest. Scale bar 100 m. (A) Adapted from Jennings DH and Lysek G (1999) *Fungal Biology*. Oxford, UK: Bios Scientific, with permission. (B) courtesy of Karl Ritz. (C) Reproduced with permission from Ritz K and Crawford JW (1990) Quantification of the fractal nature of colonies of *Trichoderma viride*. *Mycological Research* 94: 1138–1142. (D) Adapted from Crawford JW et al. (1993) Quantification of fungal morphology, gaseous transport and microbial dynamics in soil: An integrated framework utilising fractal geometry. *Geoderma* 56: 157–172, with permission. (E) Reproduced with permission from Robinson CH et al. (1993) Resource capture by interacting fungal colonies on straw. *Mycological Research* 97: 547–558. (F) Reproduced with permission from Eric Nelson, Cornell University. (G) is adapted from Smith ML et al. (1992) The fungus *Armillaria bulbosa* is among the largest and oldest living organisms. *Nature* 356: 428–431 © 1992 Macmillan Magazines, with permission.

Biotic interactions

In soils, as in other environment compartments, no organism grows in isolation, but always in the context of a wider microbial community, and soil organisms have evolved under such circumstances over millions of years. It is not surprising, therefore, that there are extensive interactions between soil fungi and all other classes of organisms, including other fungi, bacteria, protozoa and viruses. At present, we only have a limited understanding of the degree and nature of such interactions, since the majority of microbiological research has been on the growth and physiology of microbes often cultivated in isolation, or in simple combination, and often in vitro. But the ecological importance of interaction networks among fungi and with members of other kingdoms is actually considered in the emerging concepts of soil microbiomes and holobionts (Fierer, 2017).

Fungal associations with other microbes

Fungal associations with algae or cyanobacteria, in the form of lichens with combination specific morphology, play a role in soil genesis via rock weathering (Jackson, 2015). Bacteria can also glide at the surface of fungal hyphae colonizing and connecting soil compartments. In soils proper, when fungal mycelia meet, there can be a variety of outcomes from such interactions. If they are somatically compatible, they will fuse to form a larger mycelium (see above); other outcomes include an intermingling without fusion, chemically mediated repellence, physical exclusion by the development of insulated barriers, or death and replacement of one mycelium by the other (Fricker et al., 2017). There are complex hierarchies of combativity between different species and strains, and environmental factors such as temperature, water, pH, and nutritional status impact the balance between niche conditions and stochasticity in shaping community composition of soil fungi (Beck et al., 2015). Fungal communities in soil are therefore in a state of perpetually responding to stress and interaction. There are also many examples of specifically hyperparasitic fungi (e.g., *Arthrobotrys oligospora*, *Coniothyrium minitans*) and bacterial mycoparasites (e.g., *Pseudomonas* spp.). Mycoviruses do not have an extracellular phase, and since they do not necessarily affect the fungal phenotype they may be more prevalent than was thought hitherto.

Fungal associations with plants

Plants are the primary autotrophs in terrestrial systems and therefore responsible for most deposition of fixed C into soils. Litter input and root exudation provide the majority of substrate for the heterotrophic fungal soil biomass. Biotic associations between vascular plants and fungi are extensive.

Mutualism

Mutualistic associations between plant roots and fungi are termed ‘mycorrhizas’ (Fig. 2A–E) and first evolved, almost simultaneously, with the emergence of land plants (Smith and Read, 2010). Such associations are extremely common, and the natural state for the majority of root systems is to be infected to some extent by mutualistic fungi. Seven major types of mycorrhizal association exist that differ anatomically and physiologically and have variable host ranges (Table 1). In the majority of cases, the fungal partner derives carbon from the host, and the plant benefits from a more secure uptake of nutrients such as P or N and water, mediated by extensive external (extraradical or extramatrical) mycelia. In some cases, however, non-chlorophyllic plants receive organic carbon from their mycorrhizal partner (mycotrophism). Mycorrhizal infection may also confer resistance against pathogens and abiotic stresses such as drought or toxic metals. Given that most soil microbes are C-limited, it is likely that mycorrhizal hyphae constitute a significant proportion of the fungal biomass in vegetated soils, since they are intimately associated with a readily available source of C via their host plants. Mycorrhizas strongly affect the competitive ability of plants (often mediated via nutrient acquisition) and thus play a role in governing the community structure of vegetation. Mycorrhizal networks can infect a number of plants of the same or different species simultaneously and thus form an interconnected common mycelium network, providing the possibility for nutrient and signal molecule transfer between plants via interconnecting hyphae (Johnson and Gilbert, 2015). For example, it has been suggested that arbuscular mycorrhiza influence the acquisition of mineral phosphorus either positively or negatively among plant communities via common mycelium networks (Smith et al., 2011). Research on the ecological significance of such phenomena is still in its infancy. The network of extraradical growing hyphae of mycorrhizal fungi, also affects soil formation and functioning. Owing to its particular role, especially in forests, in relating plants and soil compartments and acting as nutrient, water and information transfer channels, this fungal network is a central element of what has been termed the “wood wide web” (Simard and Durall, 2004). Mycorrhizal fungi that also have saprotrophic potential (e.g., ectomycorrhizal) have a competitive advantage over exclusive saprotrophs. In forest biomes dominated by ectomycorrhiza, this competitive advantage lowers the litter turnover and triggers accumulation of organic matter (Fernandez and Kennedy, 2016). Vice versa, arbuscular mycorrhiza are obligate biotrophs and have no direct saprotrophic capabilities. Biomes where they dominate are characterized by high levels of litter decomposition and accumulate less soil organic matter.

Parasitism

Plant pathogenic fungi cause very significant economic losses to crops in both temperate and tropical agricultural and forestry systems (Doehlemann et al., 2017). They also occur in natural systems and play a role in mediating plant community structure,

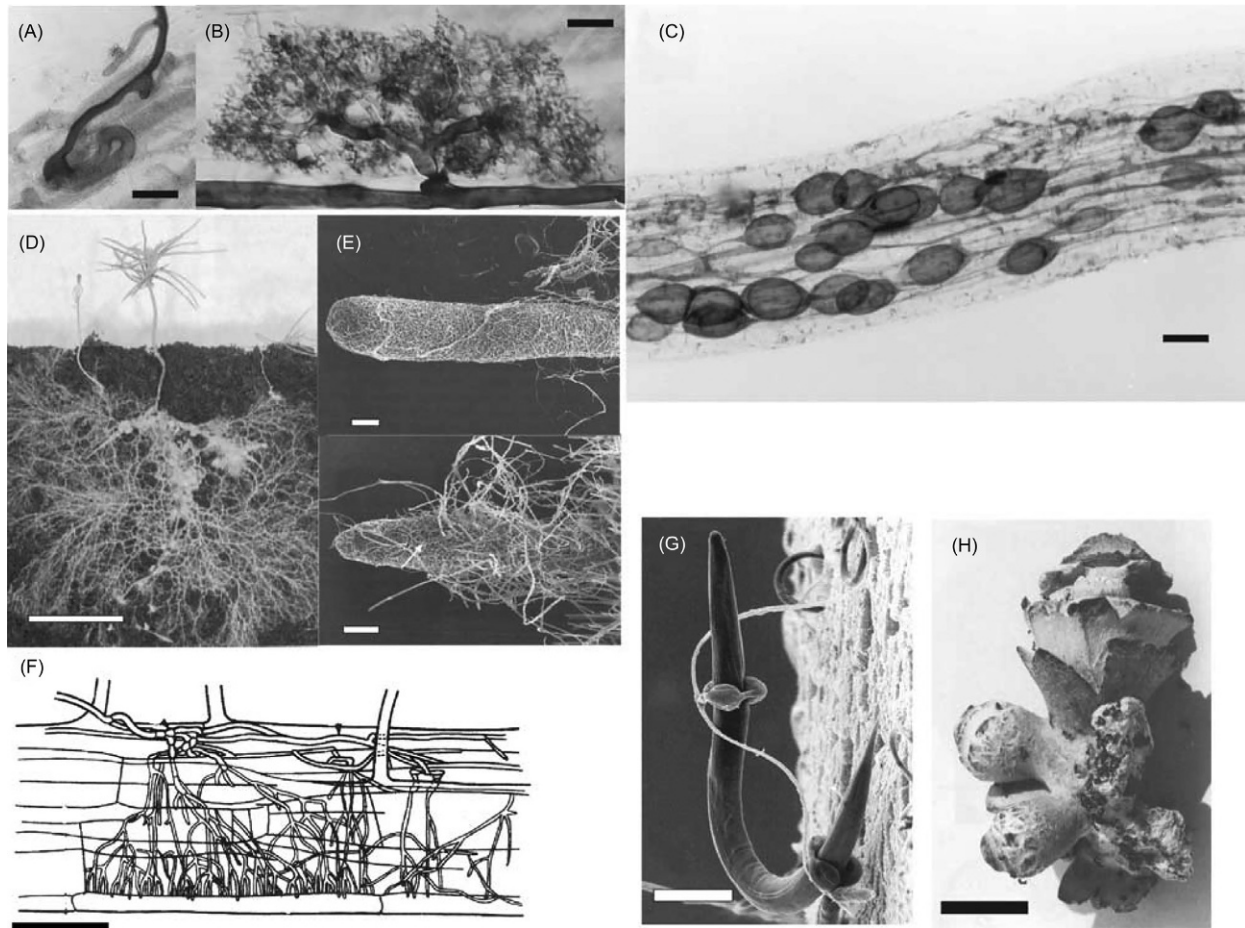


Fig. 2 Mutualistic associations involving soil fungi. (A) Infection point of arbuscular mycorrhizal (AM) fungus on root of *Plantago lanceolata* (plantain). Scale bar 50 μ m; (B) arbuscule in root cell of *Allium porrum* (leek). Scale bar 3 μ m; (C) intraradical mycelium and vesicles of AM fungus proliferating in root cortex of *P. lanceolata*. Scale bar 200 μ m; (D) external mycelium of *Suillus bovinus* on *Pinus sylvestris* (pine) in microcosm system. Scale bar 5 cm; (E) ectomycorrhizal colonization of roots of *Betula alleghaniensis* (yellow birch) by *Pisolithus tinctorius*. Top: thin mantle of mycelium formed on root surface; bottom: compact mantle entirely enveloping roots, and development of external mycelium. Scale bar 100 μ m; (F) infection of *Triticum aestivum* (wheat) roots by *Gaeumannomyces graminis* (take-all). Scale bar 250 μ m; (G) nematode trapped by constricting rings of *Arthrobotrys anthonia*. Scale bar 100 μ m; (H) basidiomata of *Leucoagaricus gongylophorus* removed from nest of *Atta cephalotes* (ant). Scale bar 2 cm. (A, C) Reproduced with permission from Karl Ritz. (B) Reproduced with permission from Brundrett MC et al. (1983) A new method for observing the morphology of vesicular- arbuscular mycorrhizae. *Canadian Journal of Botany* 62: 2128–2134. (D) Reproduced with permission from Harley JL, Smith S, and Reid DJ (1996) *Mycorrhizal Symbiosis*. 2nd edn. with permission from Academic Press. (E) Reproduced with permission from Massicotte HB et al. (1990). Structure and ontogeny of *Betula alleghaniensis*-*Pisolithus tinctorius* ectomycorrhizae. *Canadian Journal of Botany* 68: 579–593. (F) Adapted from Skou JP (1981) Morphology and cytology of the infection process. In: Asher MJC and Skipton PJ (eds) *The Biology and Control of Take-all*. London, Academic Press, with permission. (G) Reproduced with permission from George Barron. (H) Reproduced with permission from Fisher PJ et al. (1994) *Leucoagaricus basidiomata* from a live nest of leaf-cutting ant *Atta cephalotes*. *Mycological Research* 98: 884–888.

although relatively little is known about the extent of such regulation. Such fungi can have biotrophic, necrotrophic or hemibiotrophic life styles (Horbach et al., 2011). Importantly, whereas plant fungal pathogens are less virulent in species-rich plant communities, they cause serious diseases in intensive crop monocultures.

Soil-borne plant pathogens occur across almost all fungal taxonomic groups and vary in their host specificity (Crous et al., 2015); important examples include *Gaeumannomyces graminis* (take-all of wheat; Fig. 2F), *Rhizoctonia solani* (root and stem rots, including potato tubers, with a wide host range), *Armillaria mellea* (honey fungus, with a wide host range of trees), *Fusarium oxysporum* (a wide variety of forms, which affect specific plants – e.g., *F. cubense* causes wilt in banana).

The primary site of infection for soil pathogens is normally roots, although some pathogens attack the hypocotyls or stem bases of plants on or near the soil surface, and there are a range of tuber diseases. Many pathogens show strong taxis toward roots, mediated via soluble or volatile exudates, or electrical charges (electro-taxis). The pathogenic mode of action takes many forms between different pathogens, from necrotrophic attack and wholesale degradation of tissue, through cortical invasion, root hypertrophy, to the blocking of vascular tissue and hence interference with water and solute transport (wilts) (Horbach et al., 2011). There is often a proliferation of mycelia on root surfaces prior to penetration via localized aggregations of hyphae (infection cushions).

Table 1 Basic characteristics of the seven mycorrhizal types (see Smith and Read (2010) for details).

Mycorrhizal type	Habit	Host range	Examples	Notes
Arbuscular (AM)	Endotrophic; form arbuscules and sometimes vesicles in root cortical cells	ca. 90% of vascular plants; exceptions include members of Brassicaceae	Exclusively Glomeromycota (e.g. <i>Glomus</i> , <i>Gigaspora</i> , <i>Acaulospora</i>)	Obligate biotrophs; can only be cultured in association with plants (including root-organ tissue culture systems)
Ecto- (ECM)	Ectotrophic; form extensive mycelial mantles around root, penetrate between cells of root cortex to form a Hartig net)	Mainly woody plant of boreal and temperate regions but also some major tropical trees	Basidiomycetes (e.g., <i>Amanita</i> , <i>Suillus</i>); Ascomycetes (e.g. Tuber); some Zygomycetes (Endogone)	Also grow saprotrophically in absence of host plant, and in pure culture
Ectendo-	Ectendotrophic; same anatomy as ECM but with penetration into root cells	Young trees such as spruce, frequent in nurseries	Ascomycetes (e.g. <i>Cenococcum</i>) Basidiomycetes	Same as for ECM
Arbutoid	Ectendotrophic; same anatomy as ECM but intense penetration into root cells	Essentially mediterranean Ericaceae of the genus <i>Arbutus</i>	Basidiomycetes and Ascomycetes that form ECM on neighbor trees	As <i>Arbutus</i> resists forest fire, it helps ECM formation on trees regenerating burnt forests
Monotropoid	Ectendotrophic; same anatomy as ECM but with penetration into root cells limited to hyphal pegs	Exclusively achlorophyllous plants that parasitize tree roots (e.g. <i>Monotropa</i>)	Basidio and ascomycetes forming ECM on the parasite host tree	Mycotrophy: transfer of carbohydrates from the parasitized ECM host tree via the common mycelium
Ericoid	Endotrophic; intracellular coils formed inside host cortex	Ericaceae	Hymenoscyphus, Oidiodendrum	Occupy up to 80% of root cell volume, almost no extra-radical mycelium. Secrete acid proteinases that attack recalcitrant litter in acidic soils
Orchidoid	Endotrophic; short-lived intracellular coils in host root cells	Seeds and achlorophyllous orchids	<i>Rhizoctonia</i> , <i>Marasmius</i>	Mycotrophy: provide carbohydrates to orchid seed endosperm for stimulating growth of the protocorm, on which plant embryos differentiate

Control of root pathogens is often based upon use of resistant cultivars, crop rotation to mitigate build-up of pathogen populations, and appropriate soil management such as crop rotation and tillage practices, organic matter incorporation, or liming. Persistent growth of the same cereal crop can result in so-called take-all decline, where disease severity is reduced after several seasons, due to the build-up of microbial communities that are antagonistic to *Gaeumannomyces*. Chemical control is rarely feasible due to the buffering capacity of soil against bioactive compounds and the rapid degradation of such compounds by the soil microflora.

Fungal associations with fauna

Fungal associations with soil fauna are extremely common, including a wide range of mutualistic and parasitic relationships. For example, in predatory, nematode-trapping species (e.g., *Arthrobotrys* and *Dactylella* spp.) an extensive variety of mechanisms to procure their prey has evolved (Su et al., 2017). This includes sticky hyphae, branches, nets, and knobs, which are sufficiently adhesive to prevent escape of nematodes prior to digestion by the fungus (analogous to insectivorous sundew plants), and hyphal rings (lassoes), some types of which contain inflating trigger cells, which grip the bodies of prey to prevent escape (Fig. 2G). Other fungi (e.g., *Verticillium* spp.) infect nematodes via spore attachment and hyphal penetration of cysts, juveniles, or adults. Virtually all insects studied to date appear to have a range of parasitic and mutualistic (principally in their guts) fungi associated with them (Wang and Wang, 2017), including during larval and pupal stages by vertical transmission. Fungal hyphae and spores are also attacked or grazed by representatives from most faunal groups, including protozoa, nematodes, worms, insects, and mites (Crowther et al., 2012). Subtle interactions occur. For example, there is a wide variation in the degree of palatability of different fungi to different fauna, which results in preferential grazing of mycelia by soil fauna, analogous to aboveground herbivory (Bonkowski et al., 2000). Soil fauna play a strong role in dispersing fungi through soil, either as spores attached to their bodies or via egestion in feces at sites remote from where the fungus was ingested. Some Basidiomycetes (e.g., *Termitomyces*, *Acromyrmex*, *Leucoagaricus* spp.; Fig. 2H) are 'farmed' by termites and ants, maintained within the insect colonies in essentially pure culture on imported plant material.

Microbiome and holobiome

Microorganisms are intermingled in their functioning within the plant-soil ecosystem (Vandenkoornhuysen et al., 2015). Mycorrhiza helper bacteria support fungi in forming mycorrhiza (Frey-Klett et al., 2007), and nitrogen-fixing bacteria support the activity of fungi in decaying dead wood (Hoppe et al., 2014), a substrate with very sparse available nitrogen resources. More generally, every biogeochemical transformation or cycling activity in soils involves dynamic communities of interacting microorganisms, which

comprise, in addition to fungi, protozoa, archaea, bacteria and viruses. These inter-kingdom microbial communities constitute the “soil microbiome,” and the community structure and functions of these complex biocoenoses can now be explored with soil molecular ecological approaches (see below). Similarly, for larger multicellular organisms such as animals and plants, the concept of holobiont has emerged, which considers the sum of these organisms plus their associated microbiomes as functional and evolutionary entities (hologenomes) (Zilber-Rosenberg and Rosenberg, 2008).

Fungal contributions to soil-system formation and function

Soil formation and pedogenesis

Soil formation (pedogenesis) occurs via concomitant weathering of parental rock and decay of plant residues, accompanied by selective vertical transfers and edifications of complexes between the minerals and organic fractions that result from the two processes. Physical, chemical and biological factors drive these intermingled transformation and complexation steps, and fungi contribute to the three levels. So called rock-eating fungi, free-living or in lichen or mycorrhizal associations, contribute to the weathering of parental rock material and bore micro-tunnels at rock surface, which opens the way for their physical fractionation into smaller particles with enhanced chemical reactivity. Fungi are major agents of plant residue transformation, and the biomolecules produced during the decay of their own necromass enter into the composition of the stable soil organic matter (humus). As described below, fungi are also crucial contributors to formation and stabilization of mineral-organic complexes and soil aggregates (Rillig, 2004).

Nutrient cycling

Soil fungi play a crucial role in nutrient cycling in terrestrial systems, due to the primary role they play as decomposers, mediated by a large repertoire of catabolic enzyme activities (Waring et al., 2013). They are particularly involved in cycling C, N, and P, but have roles in most of the other soil elemental cycles (Read and Perez-Moreno, 2003). Fungi are primarily involved in the decomposition of plant material, the lignocellulosic components of which are relatively recalcitrant to bacteria. There is typically a succession of fungal (and other microbial) species on new substrate when it enters soil, based upon the initial decomposition of simpler compounds and a subsequent degradation of more complex polymers (Purahong et al., 2016). Cellulases are produced by a wide variety of fungi across the taxonomic classes; these include the brown rots, so-called because their activity results in a darkening of wood as the cellulose is decomposed and lignin (which they are not able to decompose) remains. Within fungi, polyphenol oxidases (mainly laccases), and mangan- or lignin-peroxidases involved in the decomposition of lignin, are largely confined to Basidiomycetes and some Ascomycetes. These include the white rots species, which normally also have strong cellulase activity but render wood a light color during selective lignin decay.

As well as invoking the enzymatically mediated degradation of polymers, many fungi also produce a variety of compounds such as organic acids or siderophores, which solubilize or immobilize essential or toxic metals, with consequences for both their own growth and the growth of proximal organisms (Priyadarshini et al., 2021). Fungi also act to immobilize nutrients in soil through sequestration in mycelia, which are further released following death and lysis or attack by pathogens or grazers. Eucarpic fungi play a significant role in physically transporting nutrients through the soil fabric via translocation within mycelia, thus short-circuiting other diffusion and mass-flow transport mechanisms, both in saprotrophic and mutualistic (mycorrhizal) contexts. Soil humus was long considered to consist in partly degraded and transformed plant residues, but is now recognized as residual biomolecules of microorganisms and hence, in large part, originate from fungi, that are the main contributors to humus formation (Liang et al., 2020).

Soil architecture

Eucarpic fungi influence soil structural dynamics via a number of mechanisms and on different spatial scales. Hyphal networks enmesh soil particles together, rather like the bars in reinforced concrete. Sparse mycelia in nutrient-poor mineral soils are unlikely to be particularly influential in this respect, but the dense mycelia that can form where substrate is abundant, for example in forest litter layers, can be quite resistant to physical disruption. These mechanisms operate on scales of millimeters to centimeters. The variety of compounds exuded by hyphae as they grow bind soil particles together via adhesive or charge-based mechanisms, on scales from nanometers (charge) to millimeters (adhesive). Such binding results in increased aggregation of soil and an associated structural generation. Degradative enzymes prevalent near hyphal tips may also decompose organic matter binding soil particles together, resulting in structural de-generation. Fungi produce large quantities of a diverse array of hydrophobic proteins (hydrophobins), which serve to insulate hyphal walls. This is important due to the large surface area-to-volume ratio that arises from the mycelial growth-form. However, these compounds also coat soil particles and can strongly influence the water sorptivity and repellency characteristics of the soil. This in turn affects the hydraulic properties on larger scales. It has recently been suggested that glomalin, a glycoprotein produced in abundance by arbuscular mycorrhizal (AM) fungi and which is resistant to decomposition, may play a key role in effecting mechanisms of adhesion and soil stabilization (Rillig, 2004). As hyphae extend through soil, they may also cause physical restructuring of soil, by rearranging particles, and orientate clay lamellae parallel to their walls (Fig. 3).

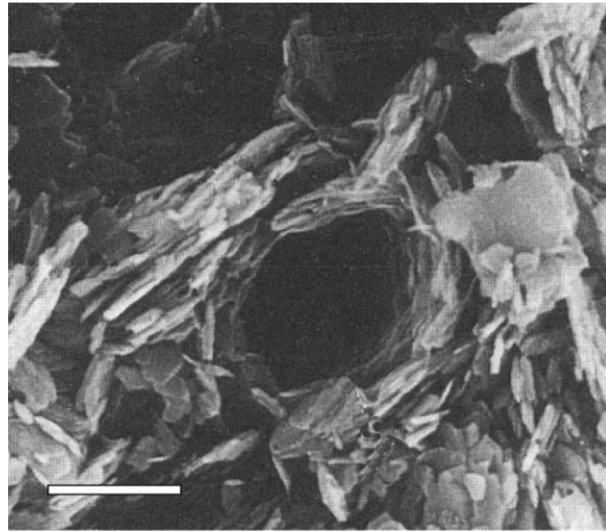


Fig. 3 Orientation of kaolinite clay platelets by hyphae of *Chetomium* sp. Scale bar 2 μm . Reproduced with permission from Dorioz JM et al. (1993) The role of roots, fungi and bacteria on clay particle organization. An experimental approach. *Geoderma* 56: 179–194.

Although fungi certainly affect soil structural dynamics, the physical architecture of the soil also affects fungal growth in that it provides the physical framework through which fungi must grow to locate resources. The eucarpic growth-form is particularly effective for growth in such spatially structured environments, where complex pore networks prevail and water and substrate are distributed in a heterogeneous manner within the three-dimensional matrix. Hyphae are able to grow across the surfaces of pore walls but also bridge wider pores; since mycelia form an interconnected network, materials can be transported within them from zone to zone, potentially supplementing growth factors where they are in demand in particular regions. Translocation within most soil fungi is yet poorly understood. There is evidence that different materials can be simultaneously transported in opposite directions within mycelia. Most substrate in soils is distributed patchily, and path lengths between resource depots can be extended greatly by the tortuosity of pore networks. Many species show foraging strategies, which balance explorative and exploitative growth strategies relative to the distribution of resources encountered in their normal habitats. Such strategies are highly developed in the cord-forming Basidiomycetes. Physical exclusion mechanisms, where pore diameters are too small to permit penetration by hyphae, probably operate; the extent to which hyphae can physically separate soil particles to allow penetration, akin to plant roots, is largely unknown. Physical protection of organic matter and organisms (including hyphae) from degradation or predation via such spatial exclusion also occurs. Hyphae are generally poor at growing through water-filled pores and hence are less prevalent in waterlogged soils. There are thus strong interactions between soil structure, water, and mycelial growth.

Methods of study

A comprehensive range of methods has been developed and applied to study the presence, abundance, activity, and distribution of fungi in soils. ‘Quantification’ of soil fungi presents a particular challenge, since there is a diversity of functionally important components, ranging from total hyphal length through mass of cytoplasm and number of propagules (Bridge and Spooner, 2001). Methods must be carefully matched to the nature of the study being undertaken. A limitation is the time investment required for each analytical approach. Time-consuming approaches only permit analyses of limited subsets of samples, which only enable us generating small data sets, which do not reflect the high heterogeneity and multi-compartmentation of soil habitats, only provide indicative insights and preclude large scale monitoring or studies.

Isolation methods

A wide diversity of fungi can be isolated from soils by traditional methods of dilution plating and enrichment growth on broad-scale or specific gel-based media supplemented with a variety of nutrient sources. These range from oligotrophic (based on silica gels, which are devoid of organic compounds) through to C-rich media such as Czapek-Dox medium (an agar gel based on 3% sucrose). Since relatively small dilutions of soil are often necessary compared with those appropriate for bacteria (typically 10^{-3} in contrast to 10^{-5}), media may be supplemented with antibacterial compounds to inhibit contaminating overgrowth. Many soil prokaryotes are apparently not cultivable in vitro, but the extent to which the ‘noncultivability syndrome’ may apply to soil fungi is unclear. Although such plating methods may be appropriate to determine fungal richness (i.e., a basic range of cultivable species

present), meaningful interpretation of 'colony-forming units' (CFUs) by such means is often dubious. This is because colonies can form from hyphal fragments as well as from spores, and the number of CFUs will therefore be strongly affected by the propensity for, and degree of fragmentation of, mycelia during sample preparation – it is difficult to relate this to the probable status of any mycelia in the soil. In addition, the expression of colonies on media can be dependent on the presence or absence of other species within the same plate. Soil fungal colonies with a high sporulation rate, such as *Penicillium*, tend to be overestimated in quantification analyses based on isolation. In addition, the numerous dormant spores of non-active fungal colonies, which germinate on isolation media bias the characterization of active soil fungal communities. Techniques to wash out spores prior to isolation help to circumvent such biases. Hyphal fragments or (larger) spores can be directly extracted from soils by micromanipulation, on density gradients or by elutriation and sieving. 'Baiting' methods are based on the insertion of sterile substrates such as gel-filled perforated tubes, cotton strips, or hair into undisturbed soil, followed by removal and isolation of fungi colonizing these substrata. If an air gap surrounds the baits, it is assumed that any fungi entering them were in an active hyphal state in the soil.

Direct observation

Measurement of total hyphal length is possible by direct observation of thin films of dispersed soil using a microscope. It is normally necessary to stain hyphae using colored (e.g., aniline blue, acid fuchsin) or fluorescent (e.g., Mg ANS, an 8-anilino-1-naphthalene sulfonic acid salt; Fluorescent Brighteners – FB24, also known as Calcofluor W M2R) stains to render them visible against the complex background of soil particles. Metabolically active hyphae can be distinguished from inactive forms by use of vital stains such as fluorescein diacetate (FDA) or tetrazolium dyes. By measuring hyphal diameters as well, estimates of bio-volume and thence biomass can be made using conversion factors. Soil films can be produced by smearing small volumes of suspension onto glass slides or pulling suspensions down onto membrane filters. Soil thin-sections, produced by impregnation of undisturbed soil with hard-curing resin, and cutting and polishing the resultant blocks with lapidary (jeweller's) pastes, enable the in situ observation of fungal features (Fig. 4). Computerized image processing and analysis can be used to automate identification and measurement procedures, but such methods rely on a high degree of contrast being obtained between hyphae and background. It is rarely possible to identify fungi by these methods, since most hyphae do not have a sufficiently distinct morphology.

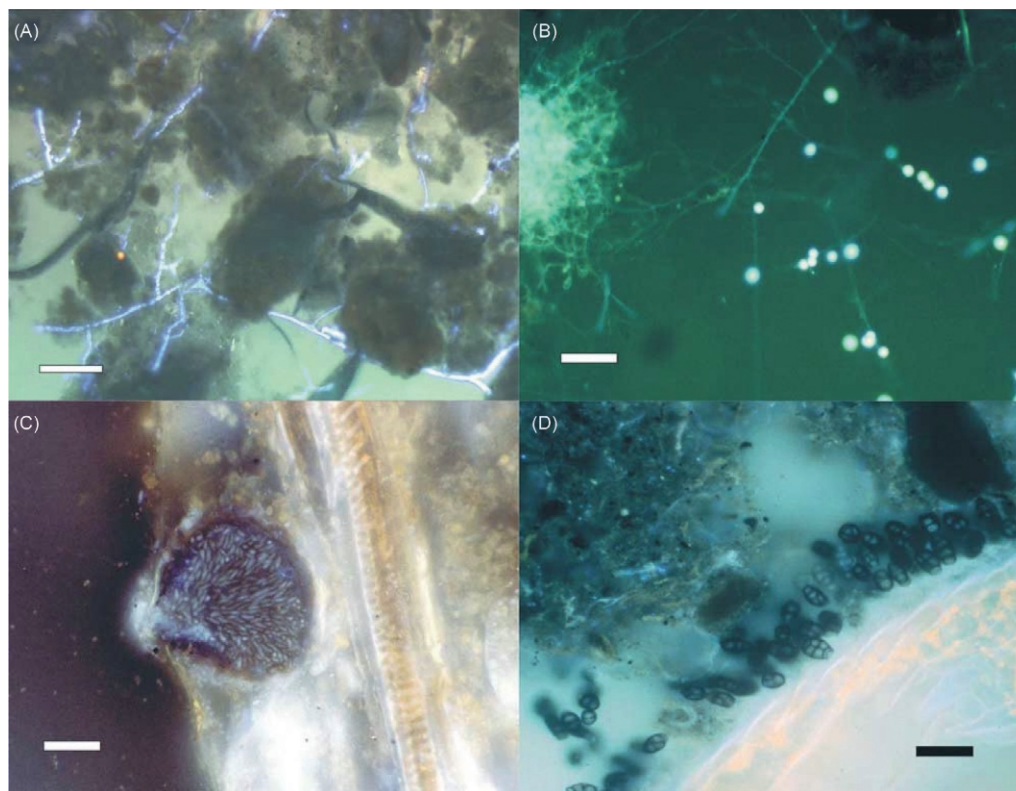


Fig. 4 Fungal features visualized in thin sections of undisturbed soil. (A) Hyphae of *Rhizoctonia solani* predominating in soil pores. Scale bar 100 μm ; (B) unidentified mycelium in grassland soil showing extensive branching in soil pore, and associated sporophores. Scale bar 50 μm ; (C) perithecial of unidentified fungus embedded in cortex of root in thin-section of arable soil (note orientation of orifice through which spores are released into soil pore). Scale bar 50 μm ; (D) accumulation of fungal spores in vicinity of decomposing root (located in lower-right quadrant of image). Scale bar 50 μm . Reproduced with permission from Karl Ritz, David Crabb, Kirsty Harris, ScottishCrop ResearchInstitute.

Infection of roots by AM fungi is normally assessed by direct microscopic observation of roots cleared of cytoplasm by treatment with alkali, and stained using dyes such as trypan blue or chlorazol black E. Morphological and anatomical morphotyping of ectomycorrhizal roots permits characterization of the diversity and frequency of ectomycorrhizal fungi and, when connections to fruiting bodies exist, even their identification. Analyzing the network, ramification and length of rhizomorphs attached to mycorrhiza also provides an insight into the proximate to distant soil compartment from which ectomycorrhiza mobilize plant nutrients.

Physiological and molecular methods

Complementary to cultivation and direct-observation techniques are so-called molecular analyses, which focus on biochemicals that are more or less specific to fungi. These methods circumvent many of the problems associated with *in vitro* cultivation of fungi and offer a wide range of scales of resolution in quantitative analysis. Total fungal biomass in soils can be estimated by direct extraction and measurement of biomarker compounds such as chitin, ergosterol, phospholipid fatty acids or DNA (Baldrian et al., 2013). The effectiveness of this approach relies on such compounds being specific to fungi and present in a constant proportion relative to the amount of biomass. Few compounds meet these criteria strictly. For example, chitin is also present in arthropods and ergosterol concentrations can vary significantly depending on the age, nutritional status, and species of fungus. More potential is shown by phospholipid fatty acids (PLFAs); these membrane-bound compounds vary in composition with reasonable consistency between different organisms, can be readily extracted from soils and have wide-ranging PLFA profiles determined by gas chromatography and mass spectrometry. PLFA 18:2 ω 6,9 and 18:1 ω 9 are as fungal indicators for Mucoromycotina, Ascomycota, and Basidiomycota, but presence of plant material creates biases. PLFA 16:1 ω 5 is apparently specifically associated with AM fungi. Immunological techniques based on polyclonal or monoclonal antibodies, particularly via the enzyme-linked immunosorbent assay (ELISA) can be used to detect the presence of specific fungi in soil samples.

Analysis of nucleic acids directly extracted from soils has revolutionized soil ecology, including soil mycology (Sandrini et al., 2022). With burgeoning knowledge of the specific nucleic acid sequences associated with particular fungi, especially in relation to ribosomal DNA (rDNA; for fungi typically 28S, 18S, and 5.8S, and associated internally transcribed spacer, ITS, regions), it is possible to prescribe probes that detect fungal DNA with virtually any degree of specificity, from the broadest of scales to strain-specific. Such probes can be labeled with fluorescent, radioactive, magnetic, or enzyme tags, which can enable their direct detection in soil-based systems. Application of the polymerase chain reaction (PCR) can enable the detection of extremely small concentrations of specific DNA sequences in complex environmental samples and can also be used in a broad-scale or specific sense. Many techniques have emerged to analyze PCR products, with cloning and sequencing for the Sanger sequencing to the next generation sequencing that no-longer requires cloning. Sequencing technologies including pyrosequencing (454), Illumina, Ion Torrent, Pac Bio, MinION differ in their principle, the length of the sequenced fragment and the degree of sequencing errors. Apart sequencing of PCR products gained from nucleic acids extracted from soils (barcoding), so called shotgun sequencing analyze directly DNA extracts from soil and enable assembly of large portions or even whole genomes of soil fungi and ideally to capture the metagenome of soil fungi, i.e., the assembly of all fungal genomes present in a sample. All these technologies require bioinformatics tools to assemble the produced fragments, filter out chimera. By use of reference data banks, such as UNITE (<https://unite.ut.ee/>) and FUNGuild (<http://www.funguild.org/>) sequences can be assigned to taxa and functional ecological guilds. Owing to the power of these techniques, monitoring diversity of soil fungi at global scale has become possible.

Activity

Specific assessment of fungal activity in soils is difficult, since it will invariably occur in combination with the activity of the rest of the biomass (Nannipieri et al., 2017). One approach is to measure bulk respiration in the presence and absence of metabolic inhibitors specific to prokaryotes such as streptomycin; assuming the inhibitors are actually specific and fully effective, the difference in respiration is attributable to fungi. More effective procedures are based on measuring the incorporation of stable or radioactive isotopes into fungal marker compounds, such as the incorporation of ^{14}C -labeled acetate into ergosterol. Since RNA is produced to a greater extent in metabolically active cells and hyphae, analysis of directly extracted RNA, in a similar manner to that used for DNA (see “Physiological and molecular methods” section) offers great potential. Although, RNA is difficult to extract from soils and mycological applications to date have been relatively limited, the era of soil fungal transcriptome and meta-transcriptome analyses is now opened. Fractionation of substrate labeled with stable isotopes (e.g. ^{15}N , ^{13}C , ^{18}O deuterium) enables their breakdown trajectory along soil food webs to be followed, and hence improve understanding of soil microbiome functioning.

Exploitation of soil fungi

Soil fungi are unquestionably vital to the functioning of terrestrial ecosystems but they also play an important and increasing role in biotechnology and environmental management. Culinary exploitation and cultivation has occurred for millennia, with mushroom-growing industries present on most continents. However, in comparison with their species richness, only a minute part of edible and otherwise useful fungi are produced under controlled culture conditions, owing to the complexity of factors that trigger sexual reproduction of mushrooms. As an example, the Perigord truffle (*Tuber melanosporum*) remains resistant to routine, large-scale cultivation and is consequently one of the most expensive epicurean foods on a weight basis, retailing at potentially

thousands of dollars per kilogram. On a contrasting scale, the production of single-cell protein in large-scale fermentors from *Fusarium graminearum* is a successful and established industry. Many fungi isolated from soil provide a range of important primary and secondary metabolites, which are exploited industrially. These include various organic acids (e.g., *Aspergillus*, *Rhizopus* spp.), polysaccharides (e.g., *Aureobasidium* spp.), and lipids (e.g., *Fusarium* spp.), as well as antibiotics (e.g., *Penicillium*, *Aspergillus*, *Sordaria* spp.), other pharmaceuticals (e.g., *Penicillium*, *Tolypocladium* spp.), and agro-chemicals (e.g., *Strobilurus* spp.). Several thousand antibiotics derived from fungi have been identified, but relatively few have found clinical use, since most are too toxic to humans. Fungi that are antagonistic to pathogens (see above) have actual and potential use as biocontrol agents against other fungi, nematodes, and weeds. Establishment of plants in horticultural or restoration scenarios can be compromised by the absence of suitable mycorrhizal hosts; inoculation with appropriate fungal partners can be effective and economically viable. Use of fungi in bioremediation of soils contaminated by industrial pollutants such as phenolics, hydrocarbons, and heavy metals has potential – claims are often made for Basidiomycete species that show strong phenol oxidase activity, such as *Phanerochaete chrysosporium* – and, in combination with plants (phytoremediation), metal-solubilizing fungi may offer effective soil cleanup mechanisms. Increasing use of soil fungi is likely given further research and understanding of the biology of these highly diverse and versatile organisms. However, as mentioned above, survival of fungal inoculates in soils is limited by the mycostasis phenomenon mentioned in the Section “Physiology and reproduction,” which challenge attempts to inoculate favorable fungal single strains or consortia.

Therefore, the greatest actual challenge in research on soil fungi and microbial ecology is to understand the processes of microbiome community assembly and functioning, how they answer to soil use, and how they can be steered by management practices to improve soil sustainability and ecosystem services such as high productivity, carbon storage and resistance to biotic and abiotic stresses. This is the objective of agroecology, which promotes the concept of soil health (Wilhelm et al., 2022).

References

- Agerer R (2001) Exploration types of ectomycorrhizae: A proposal to classify ectomycorrhizal mycelial systems according to their patterns of differentiation and putative ecological importance. *Mycorrhiza* 11: 107–114.
- Baldrian P, Větrovský T, Cajthaml T, Dobášová P, Petránková M, Šnajdr J, and Eichlerová I (2013) Estimation of fungal biomass in forest litter and soil. *Fungal Ecology* 6(1): 1–11.
- Beck S, Powell JR, Drigo B, Cairney JW, and Anderson IC (2015) The role of stochasticity differs in the assembly of soil- and root-associated fungal communities. *Soil Biology and Biochemistry* 80: 18–25.
- Bonkowski M, Griffiths BS, and Ritz K (2000) Food preferences of earthworms for soil fungi. *Pedobiologia* 44(6): 666–676.
- Bridge P and Spooner B (2001) Soil fungi: Diversity and detection. *Plant and Soil* 232: 147–154.
- Carille MJ and Watkinson SC (2001) *The Fungi*. Gulf Professional Publishing.
- Crous PW, Hawksworth DL, and Wingfield MJ (2015) Identifying and naming plant-pathogenic fungi: Past, present, and future. *Annual Review of Phytopathology* 53: 247–267.
- Crowther TW, Boddy L, Jones T, and H. (2012) Functional and ecological consequences of saprotrophic fungus–grazer interactions. *The ISME Journal* 6(11): 1992–2001.
- Dighton J, Sridhar K, and Deshmukh S (2019) *The roles of macro fungi (Basidiomycota) in terrestrial ecosystems. Advances in Macrofungi: Diversity, Ecology and Biotechnology*. Boca Raton, FL: CRC Press 70–104.
- Doehlemann G, Ökmen B, Zhu W, and Sharon A (2017) Plant pathogenic fungi. *Microbiology Spectrum* 5(1): 5.1. 14.
- Fernandez CW and Kennedy PG (2016) Revisiting the ‘Gadgil effect’: Do interguild fungal interactions control carbon cycling in forest soils? *New Phytologist* 209(4): 1382–1394.
- Fierer N (2017) Embracing the unknown: disentangling the complexities of the soil microbiome. *Nature Reviews Microbiology* 15(10): 579–590.
- Frey-Klett P, Garbaye J, and Tarkka M (2007) The mycorrhiza helper bacteria revisited. *New Phytologist* 176(1): 22–36.
- Fricker MD, Heaton LL, Jones NS, and Boddy L (2017) The mycelium as a network. *The Fungal Kingdom* 335–367.
- Garbeva P, Hol WG, Termorshuizen AJ, Kowalchuk GA, and De Boer W (2011) Fungistasis and general soil biostasis—a new synthesis. *Soil Biology and Biochemistry* 43(3): 469–477.
- Hawker LE (2016) *The Physiology of Reproduction in Fungi*. Cambridge University Press.
- Hoppe B, Kahl T, Karasch P, Wubet T, Bauhus J, Buscot F, and Krüger D (2014) Network analysis reveals ecological links between N-fixing bacteria and wood-decaying fungi. *PLoS One* 9(2): e88141.
- Horbach R, Navarro-Quesada AR, Knogge W, and Deising HB (2011) When and how to kill a plant cell: Infection strategies of plant pathogenic fungi. *Journal of Plant Physiology* 168(1): 51–62.
- Jackson TA (2015) Weathering, secondary mineral genesis, and soil formation caused by lichens and mosses growing on granitic gneiss in a boreal forest environment. *Geoderma* 251: 78–91.
- Johnson D and Gilbert L (2015) Interplant signalling through hyphal networks. *New Phytologist* 205(4): 1448–1453.
- Liang C, Kästner M, and Joergensen RG (2020) *Microbial Necromass on the Rise: The Growing Focus on Its Role in Soil Organic Matter Development*. Elsevier.
- Macheleidt J, Mattern DJ, Fischer J, Netzer T, Weber J, Schroeckh V, Valiente V, and Brakhage AA (2016) Regulation and role of fungal secondary metabolites. *Annual Review of Genetics* 50: 371–392.
- Nannipieri P, GRECO S, and Ceccanti B (2017) Ecological significance of the biological activity in soil. *Soil Biochemistry* 293–356.
- Priyadarshini E, Priyadarshini SS, Cousins BG, and Pradhan N (2021) Metal-Fungus interaction: Review on cellular processes underlying heavy metal detoxification and synthesis of metal nanoparticles. *Chemosphere* 274: 129976.
- Purahong W, Wubet T, Lentendu G, Schloter M, Pecyna MJ, Kapturska D, Hofrichter M, Krüger D, and Buscot F (2016) Life in leaf litter: Novel insights into community dynamics of bacteria and fungi during litter decomposition. *Molecular Ecology* 25(16): 4059–4074.
- Read D and Perez-Moreno J (2003) Mycorrhizas and nutrient cycling in ecosystems—A journey towards relevance? *New Phytologist* 157(3): 475–492.
- Richards TA, Leonard G, and Wideman JG (2017) What defines the “kingdom” fungi? *Microbiology Spectrum* 5(3): 5.3. 23.
- Rillig MC (2004) Arbuscular mycorrhizae, glomalin, and soil aggregation. *Canadian Journal of Soil Science* 84(4): 355–363.
- Sandrini M, Nerva L, Sillo F, Balestrini R, Chitarra W, and Zampieri E (2022) Abiotic stress and belowground microbiome: The potential of omics approaches. *International Journal of Molecular Sciences* 23(3): 1091.
- Schinner F, Öhlinger R, Kandler E, and Margesin R (2012) *Methods in Soil Biology*. Springer Science & Business Media.
- Simard SW and Durall DM (2004) Mycorrhizal networks: A review of their extent, function, and importance. *Canadian Journal of Botany* 82: 1140–1165. <https://doi.org/10.1139/b04-116>.
- Smith ML, Bruhn JN, and Anderson JB (1992) The fungus *Armillaria bulbosa* is among the largest and oldest living organisms. *Nature* 356(6368): 428–431.

- Smith SE, Jakobsen I, Grønlund M, and Smith FA (2011) Roles of arbuscular mycorrhizas in plant phosphorus nutrition: Interactions between pathways of phosphorus uptake in arbuscular mycorrhizal roots have important implications for understanding and manipulating plant phosphorus acquisition. *Plant Physiology* 156(3): 1050–1057.
- Smith SE and Read DJ (2010) *Mycorrhizal symbiosis*. Academic press.
- Su H, Zhao Y, Zhou J, Feng H, Jiang D, Zhang KQ, and Yang J (2017) Trapping devices of nematode-trapping fungi: Formation, evolution, and genomic perspectives. *Biological Reviews* 92(1): 357–368.
- Tedersoo L, Bahram M, Põlme S, Kõljalg U, Yorou NS, Wijesundera R, Ruiz LV, Vasco-Palacios AM, Thu PQ, and Suija A (2014) Global diversity and geography of soil fungi. *Science* 346(6213), 1256688.
- Vandenkoornhuyse P, Quaiser A, Duhamel M, Le Van A, and Dufresne A (2015) The importance of the microbiome of the plant holobiont. *New Phytologist* 206(4): 1196–1206.
- Wang C and Wang S (2017) Insect pathogenic fungi: Genomics, molecular interactions, and genetic improvements. *Annual Review of Entomology* 62: 73–90.
- Waring BG, Averill C, and Hawkes CV (2013) Differences in fungal and bacterial physiology alter soil carbon and nitrogen cycling: Insights from meta-analysis and theoretical models. *Ecology Letters* 16(7): 887–894.
- Wijayawardene NN, Hyde KD, Al-Ani LKT, Tedersoo L, Haelewaters D, Rajeshkumar KC, Zhao R, Aptroot A, Leontyev D, and Saxena R (2020) Outline of Fungi and fungus-like taxa. *Mycosphere Online: Journal of Fungal Biology* 11(1): 1060–1456.
- Wilhelm RC, van Es HM, and Buckley DH (2022) Predicting measures of soil health using the microbiome and supervised machine learning. *Soil Biology and Biochemistry* 164: , 108472.
- Zilber-Rosenberg I and Rosenberg E (2008) Role of microorganisms in the evolution of animals and plants: The hologenome theory of evolution. *FEMS Microbiology Reviews* 32(5): 723–735.

Soil algae

Karen Baumann^a and Patrick Jung^b, ^aFaculty of Agricultural and Environmental Sciences, Soil Science, University of Rostock, Rostock, Germany; ^bApplied Logistics and Polymer Sciences, University of Applied Sciences Kaiserslautern, Pirmasens, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	61
Organisms	62
Prokaryotic algae	62
Eukaryotic micro-algae	64
Identification of soil algae	65
Habitat effects on soil algae	65
Algal distribution in the soil profile	65
Habitat niches	65
Effect of solar radiation on soil algae	66
Effect of soil moisture on soil algae	66
Effect of soil pH and nutrient availability on soil algae	66
Algae as soil conditioners	67
Soil algae in nutrient cycles	67
Macronutrient cycling	67
The role in the C cycle	67
The role in the N cycle	67
The role in the P cycle	67
Micronutrient cycling	68
Biotic interactions of soil algae	68
Outlook	69
Further reading	69

Abstract

Photosynthetically active prokaryotic cyanobacteria as well as eukaryotic microalgae, including green algae and diatoms frequently colonizing the top soil environment, group as soil algae. This taxonomically diverse group fulfils important ecological functions including carbon and nitrogen fixation, bio-weathering activities and various microbial interactions. In contrast to their aquatic relatives, the ecophysiology of soil algae is highly adapted towards the terrestrial environment that also offers new possibilities for biotechnology.

Key points

- The most important phyla of soil algae include members of Cyanobacteria, green algae and diatoms.
- Soil algae play an important role in nutrient cycles and facilitate microbial interactions within the soil environment.
- As pioneer organisms and ecosystem engineers, they are often the first microbes to colonize new terrain and prime soils.
- In contrast to their aquatic relatives, soil algae possess unique adaptations for life within the soil environment which can also be used in biotechnology.

Introduction

Phototrophic bacteria (Cyanobacteria) and eukaryotes (e.g., green algae), which inhabit terrestrial locations and reproduce sexually, via gametes, or asexually, via viable spores, are grouped as soil algae. They are often also referred to as ‘terrestrial microalgae’. Soil algae occur as unicells or filaments, solitary or aggregated in colonies, as free-living organisms forming biofilms or in symbiotic associations, such as in lichens (Figs. 1 and 2). In addition, they are a frequent constituent of the so-called biological soil crusts (biocrusts) where they—together with lichens, mosses and heterotrophic microbes—colonize the top millimeters of soils. Biocrusts are distributed world-wide. They are particularly well-studied in arid regions such as deserts where soil algae are a main component of biocrusts. The distribution of soil algae takes place via wind, water, soil relocation, human disturbance or via animal transport including cattle or migratory birds.

The cells of soil algae are between 0.5 µm and 40 µm but most of them are about 10 µm in diameter. The most common soil algae are prokaryotic Cyanobacteria as well as eukaryotic green algae and diatoms (Fig. 1, Table 1). Cyanobacteria appear

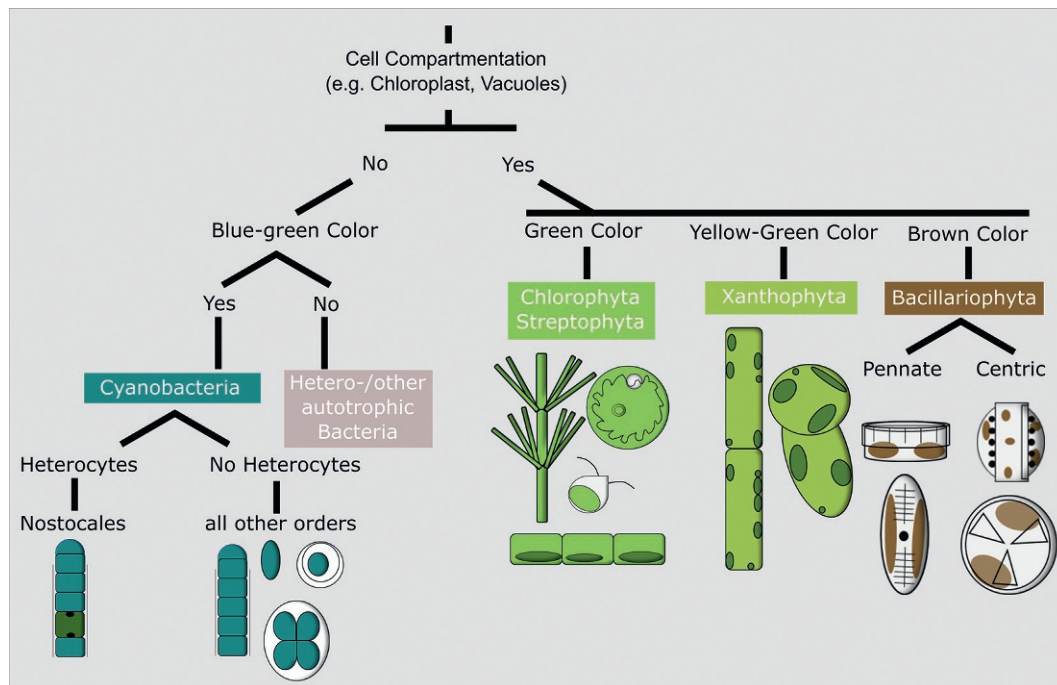


Fig. 1 Simplified taxonomic key to the most common types of soil algae. Depicted are prokaryotic filamentous Nostoclean cyanobacteria with heterocytes (e.g., *Scytonema*), cyanobacteria of non-heterocytous orders (e.g., filamentous *Oscillatoria*, unicellular *Synechococcus*, unicellular *Gloeocapsa* with layered mucilaginous sheath, aggregated *Chroococcus* colony with mucilaginous sheath). Eukaryotic Chlorophytes and Streptophytes (e.g., multi-branched macroscopic *Chara*, unicellular *Trebouxia* with a pyrenoid in the lobed chloroplast, unicellular *Chlamydomonas* with flagella, filamentous *Klebsormidium* with parietal chloroplasts). Xanthophytes (e.g., filamentous *Vaucheria* with multiple chloroplasts). Diatoms of the bilaterally symmetric pennate and radially symmetric centric type, both with their glass-like cell wall (frustule) consisting of an upper (epitheca) and bottom box (hypotheca).

blue-greenish because of their special pigments phycoerythrin (red) and phycocyanin (blue) that mask the chlorophylls. Their cells can be surrounded by a thick mucilaginous sheath made of extracellular polymeric substances (EPS) that contains sugars but also proteins. Eukaryotic green algae on the other hand are surrounded by a cellulose-rich cell wall and contain at least one chloroplast with a light greenish color reflecting chlorophylls as their main photosynthetic pigments. Diatoms are remarkably diverse due to a high variability of their characteristic structured silica cell wall and appear brownish due to fucoxanthin as their dominant pigment.

Most soil algae have a photo-litho-autotrophic metabolism, are able to perform oxygenic photosynthesis and are poikilohydric. This means that they are highly dependent on the water availability of their environment as they possess only limited mechanisms to maintain cell turgidity. Besides liquid water from, e.g., rain, they can use relative air humidity, fog and dew for photosynthesis. As such they are organisms, predestined to inhabit soils and unweathered rocks, and are regarded as pioneers. They can outcompete higher plants and mosses under certain circumstances such as when the water availability drops below a certain limit.

Due to their ability to fix atmospheric carbon dioxide (CO_2) they play an important role in the carbon (C) cycle. Further, some Cyanobacteria can fix atmospheric nitrogen (N_2) or are reported to mobilize soil phosphorus (P), thus also representing an essential part in the N and P cycles. This means that soil algae are often the first organisms that colonize barren ground and prime this material by an enrichment in C, N and other substances essential for life throughout trophic levels. Soil algae often intimately interact with other microbes such as (lichen-forming) fungi and heterotrophic bacteria and also interact with each other in multipurpose ways. This does not only include symbiotic or parasitic relationships but is most often also important for whole ecosystem functions from the micro- to the macro scale.

Starting in the early 19th century, soil algae were included in plant taxonomic studies before lists of soil algae were compiled in the 20th century for many parts of the world. Nowadays much attention is also paid to soil algae in soil ecological studies and to their role in soil fertility, soil stabilization, pest control, and biotechnology.

Organisms

Prokaryotic algae

As one of the oldest forms of life, Cyanobacteria are prokaryotic cells which do not contain any nucleus but free DNA and 70S ribosomes in the cytoplasm (Table 1). They are Gram-negative bacteria without any flagellum. Oxygenic photosynthesis is performed in variously shaped thylakoids which always contain chlorophyll *a* and additionally may contain chlorophyll *d* and *f* which absorb light in the far red. The accessory pigments, the so called phycobilins, phycocyanin (blue) and phycoerythrin (red)

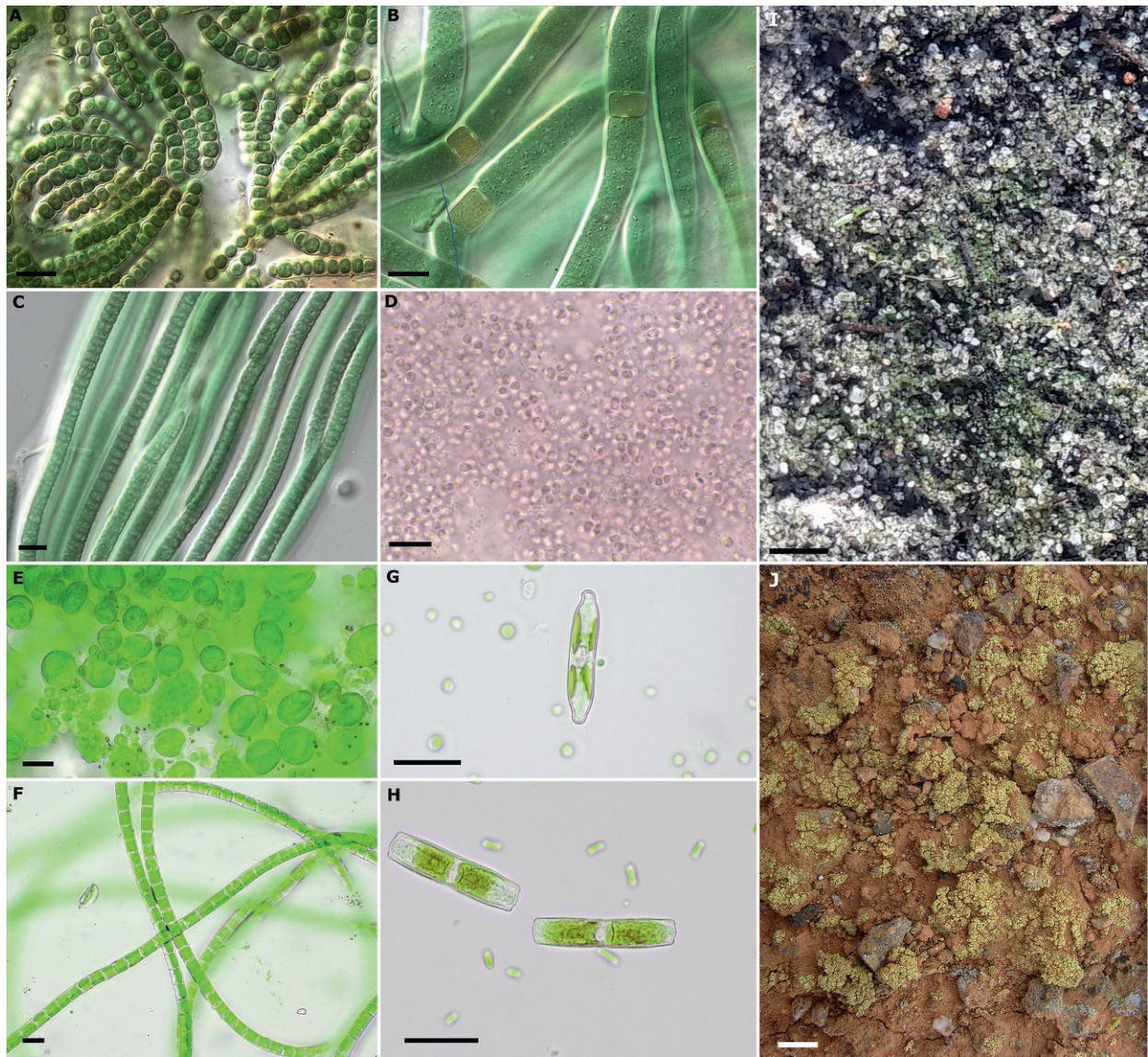


Fig. 2 Common soil algae. (A) motile hormogonia of *Nostoc* sp. from Spitsbergen, Arctic. (B) *Scytonema hyalinum* with heterocytes from biocrusts of the Atacama Desert, Chile. (C) *Microcoleus vaginatus* from Antarctica. (D) *Gloeobacter violaceus* from soil in a cave, Spain. (E) *Bracteacoccus* sp. releasing spores from soil, Germany. (F) *Klebsormidium* sp. from dunes of the Baltic Sea. (G and H) Diatoms from arable soil, Germany. (I) Algae on sandy soil surface, Germany. (J) Lichen dominated biocrust of the Atacama Desert, Chile. Scale bar (A–H) 10 μm , (I) 1 cm, (J) 5 cm. Photos: A–H, J P. Jung, I K. Baumann.

are organized in phycobilisomes. They allow photosynthesis also to take place under low light conditions by exploiting the so called 'green gap' in the light spectrum. Due to the proportions of all pigments the color of Cyanobacteria can range from violet (e.g., low-light conditions) to black (e.g., high-light conditions), but most of them appear blue-greenish. Cyanobacteria fix atmospheric CO_2 within polyhedral structures the so called carboxysomes. At these sites the CO_2 concentrating mechanism (CCM) is located so that the enzyme ribulose-1,5-bisphosphate-carboxylase-oxygenase (RuBisCO) can operate in an elevated CO_2 atmosphere that makes the C-fixation more efficient. Nonetheless, most of them are also facultative photoheterotrophs taking up compounds including glucose, fructose, pyruvate, acetate and some amino acids. Prominent storage compounds of Cyanobacteria encompass cyanophycin starch which is similar to glycogen, cyanophycin (an arginine- and asparagine polymer), and volutin (a polyphosphate). For example, above an internal P concentration of about 0.7% dry weight, granules of polyphosphate start to form in *Nostoc*.

The cell size of cyanobacteria varies between 0.5 μm and 40 μm . They occur as single cells, as cell colonies or filaments (Figs. 1 and 2). The filaments can be either unbranched or truly or untruly branched including various special types. Some species show cell differentiation. For example, in one of the soil-ecologically most important orders, the Nostocales, vegetative cells, akinetes and heterocytes (formally known as heterocysts) exist. Heterocytes are specialized in N_2 fixation. They contain the enzyme complex nitrogenase, an Fe—S—Mo containing enzyme which is only active if no other N-source such as nitrate (NO_3^-) is available and no oxygen (O_2) is present. For these reasons the heterocytes are enclosed by a thick cell wall. The nitrogenase reduces N_2 to ammonia (NH_3). On the other hand, akinetes are resting cells which are rich in storage compounds.

Table 1 Selected soil algae with a common world-wide distribution.

Domain	Phylum	Class	Traditional name	Common terrestrial species	Habitats
Bacteria	Cyanobacteria	Cyanophyceae	Blue-green algae	<i>Nostoc commune</i>	Soils, lichen photobiont
				<i>Nostoc flagelliforme</i>	Arid soils
				<i>Microcoleus vaginatus</i>	Soils
				<i>Microcoleus steenstrupii</i>	Desert soils
				<i>Chroococcidiopsis</i> spp.	Endolithic, lichen photobiont
				<i>Gloeocapsa</i> spp.	Epilithic, soils
				<i>Scytonema hyalinum</i>	Soils
				<i>Stigonema</i> spp.	Epilithic, lichen photobiont
				<i>Oculatella</i> spp.	Soils, epilithic
				<i>Gloeobacter violaceus</i>	Soil
Eukaryotes	Chlorophyta	Chlorophyceae	Green algae	<i>Bracteacoccus minor</i>	Soil
		Trebouxiophyceae	Green algae	<i>Chlamydomonas reinhardtii</i>	Soil
				<i>Chlorella vulgaris</i>	Soil, tree bark
				<i>Myrmecia bisecta</i>	Soil
	Streptophyta	Klebsormidiophyceae	Green algae	<i>Stichococcus bacillaris</i>	Soil, tree bark
				<i>Trebouxia</i> spp.	Lichen photobiont
				<i>Klebsormidium crenulatum</i>	Soil
				<i>Interfilum terricola</i>	Soil
	Xanthophyta	Xanthophyceae	Yellow-green algae	<i>Streptosarcina</i> spp.	Soil
				<i>Streptofilum</i> spp.	Soil
				<i>Botrydium</i> spp.	Soil
				<i>Vaucheria</i> spp.	Soil, flooded areas
	Bacillariophyta	Bacillariophyceae	Diatoms	<i>Navicula seminulum</i>	Soil
				<i>Eunotia diodon</i>	Soil
				<i>Stauroneis anceps</i>	Soil
				<i>Hantzschia amphioxys</i>	Soil
				<i>Pinnularia borealis</i>	Soil

Cyanobacteria only reproduce asexually by, e.g., fragmentation of filaments (development of hormogonia) (Fig. 2). Cyanobacterial cells may produce EPS such as polysaccharides and polypeptides (Fig. 1). They can form jelly sheaths which protect cells from rapid dehydration and enhance quick re-hydration. Further, substances such as gloeocapsin or scytonemin can be placed into the EPS which protects the cells from high UV radiance. The EPS layer can also be used as a short-termed storage for proteins and other substances that can later be re-absorbed. This is often the case when the C:N ratio in the cells is outbalanced e.g., during dehydration events.

The taxonomy of microalgae, including terrestrial Cyanobacteria, is under constant revision. The class Cyanophyceae, formally also known as blue-green algae, includes eight orders (Table 1). One of the most important orders of the terrestrial Cyanophyceae is the Nostocales. Here, one of the most common representatives is the taxonomically diverse genus *Nostoc* that can be found in soils world-wide including hot and cold deserts. *Nostoc* is a filamentous genus with rounded cells and mostly coiled arrangement encapsulated by a wide EPS sheath (Fig. 2). The colonies mainly form biofilms or rounded thalli that mostly reach macroscopic size due to the massive amounts of EPS material. They form heterocytes for N₂ fixation, akinetes as resting stages and have a complex life cycle that often mimics the forms and shapes of other cyanobacterial genera including motile hormogonia. The genus *Nostoc* is a major player in biocrusts where it fulfils a variety of ecosystem services including N₂-fixation and prevention of soil erosion. Further, it can be the photosynthetic symbiont in lichens or plant roots (Cycads). In biotechnology it is used for EPS production in the cosmetic industry, for pigment extraction (phycocyanin) in the food industry (food dye), and for waste-water treatment due to its toxins absorbing abilities.

Eukaryotic micro-algae

Eukaryotic soil algae are characterized by a cell structure that includes a DNA-containing nucleus, 80S ribosomes and mitochondria in the cytoplasm as well as a cell wall made up from pectin in fibrillary base structure of cellulose, xylans, and mannans (Fig. 1). The cell wall can be encrusted by silica (SiO₂), as in the algal class of Bacillariophyceae, also called diatoms (Table 1). Chloroplasts of the micro-eukaryotic algae are made up of at least two membranes and contain free chloroplast DNA and 70S ribosomes (endosymbiotic theory NOTEREF _Ref111579902 ¶ 1). The chloroplasts contain thylakoids with photosystems II and I located in the thylakoid membranes. The photosystems are made up of accessory pigments such as carotenoids or xanthophylls as well as of chlorophyll *a + b* (phylum Chlorophyta) or *a + c* (phyla Xanthophyta, Bacillariophyta), a reaction center (protein complex P680, P700) and a water-splitting complex. Eukaryotic soil algae build up starch via the Calvin cycle. At species-specific pyrenoids, a structurally

NOTEREF _Ref111579902 ¶ 1 The endosymbiotic theory proposes that organelles, such as mitochondria and chloroplasts found in eukaryotic cells were once independent prokaryotic microbes.

identifiable round complex often in the center of the chloroplast(s) of some green algae, the CO₂ concentrating mechanism takes place so that RuBisCO can operate more efficiently. The number of the chloroplasts per cell are species specific. Many Chlorophyta and Xanthophyta are facultative chemoheterotrophs which shows the potential for growth of these algae beneath the soil surface.

In contrast to Cyanobacteria, eukaryotic soil algae reproduce asexually by disintegration of filaments, or binary fission, as well as sexually by the development of zoospores with or without two flagella. Due to these flagella, eukaryotic soil algae are able to move actively and redistribute in water films that form within the porous space of soil particles.

Within the Plant kingdom terrestrial eukaryotic algae mainly group into the phyla Chlorophyta (classes Chlorophyceae (green algae), Trebouxiophyceae), Streptophyta (class Klebsormidiophyceae), Xanthophyta (class Xanthophyceae) and Bacillariophyta (class Bacillariophyceae) (Table 1). Chlorophyta are mainly free living or occur as symbiotic partner in biocrusts and lichens. They contain chlorophyll *a + b*. Klebsormidiophyceae are known to integrate soil particles due to their filamentous habitus. They can also stick particles together due to EPS production. Xanthophyceae, traditionally known as yellow-green algae, contain chlorophyll *a + c*, β -carotene and xanthophyll but no fucoxanthin, whereas Bacillariophyceae, also known as diatoms, also contain fucoxanthin which covers the chlorophyll *a + c* so that plastids appear brownish. Bacillariophyta can be up to several hundred μm in size. Their cell wall (frustule) is box-like, consisting of an upper (epitheca) and bottom box (hypotheca) part. They are either bilaterally (pennate type) or radially symmetric (centric type).

One of the most abundant green algae are members of the genus *Chlorella* which belong to the class Trebouxiophyceae of the phylum Chlorophyta (Table 1). Their cells are small, usually between 2 μm and 10 μm . They have a light green color and contain one chloroplast that often fills half the cell. They grow quickly in nature and especially in the lab where they form slimy, light green biofilms in which their cells mostly do not aggregate. The cell content is rich in lipids and also proteins which makes *Chlorella* the most used alga in commercial biotechnology. Dried and pressed into pills it is often sold as a food supplement. *Chlorella* is mostly grown in open-pond systems or closed bioreactors comprising hundreds of liters. In those, it gains biomass very quickly and also has been a promising candidate to generate bio-fuel.

Identification of soil algae

Soil algae growth can be directly examined in soil by microscopic techniques. For an exact determination of species, however, algae are extracted from soil by, e.g., soil-water suspensions. Subsequently, soil algae are cultured in liquid media or are grown on agar plates. Only organisms of axenic or at least unialgal cultures are then identified at a species level using microscopy and literature keys.

Cultured soil algae can also be examined by genetic approaches. A polyphasic approach using microscopic as well as genetic techniques produces the most reliable results and reflects the scientific state of the art. In this approach, group-specific primers are used to amplify and subsequently sequence specific gene regions, on which the scientific community has agreed as valid reference points. Bioinformatic analyses such as phylogenetic tree reconstructions of these single or multi locus gene regions are then subjected to statistical methods that generate the relatedness of strains compared to DNA information from other species and strains. Describing new species follows detailed international rules and keys. Cultured species that were examined by this approach are finally transferred to public culture collections that maintain the curation of these strains and make them publicly available for the international scientific community. In some cases, phylogenetic constructions of one or even several sequenced gene regions do not create sufficient information to properly determine the taxonomic position of a strain. In such a case, full genome sequencing is an approach that accesses the full DNA information of a single organism as the *ultima ratio*. This technique becomes more and more affordable and offers new possibilities, such as genome mining approaches, in which the genetic information is systematically screened for the potential abilities to produce natural products for human use, such as antibiotics.

Habitat effects on soil algae

Algal distribution in the soil profile

On average, $(5.5 \pm 3.4) \times 10^6$ soil algae inhabit each gram of Earth's surface soil with particularly high abundances in acidic, moist and vegetated soils. They mainly dwell in the upper few millimeters of the terrestrial soil where there is sufficient light to perform photosynthesis. Some soil algae have developed strategies to deal with high light conditions, such as in vegetation-free soils. Here, they slide, for example, on their EPS matrix a few millimeters below the soil surface. Depending on the soil texture as well as on the moisture content of the soil and thus penetration of light, their occurrence is usually limited to the upper millimeters (in fine soils) or centimeters (in coarse soils) of the soil. However, cracks in dry soil may be considerably deeper and due to the downward-movement of water, agricultural activity, root growth or soil faunal activity, soil algae have also been found at depths to two meters. Thereby, their abundance decreases with increasing soil depth. At lower soil depths, cells are mainly inactive and in resting stages.

Habitat niches

Generally, five different lifestyles can be distinguished for soil algae: they can live either edaphically (in or on the soil), *endo*- or *hypolithically* (in or beneath stones), *chasmoendolithically* (in cracks of stones and rocks), *epilithically* (on top of rocks), and symbiotically as, e.g., photobiont partner in lichens. Depending on their habitat niche they encounter different environmental

conditions due to variation in factors such as solar radiation, temperature, water and nutrient availability or pH value. These factors shape their preferential niche and affect soil algal species richness and diversity and subsequently the ecosystem services that they can provide.

Effect of solar radiation on soil algae

The solar radiation reaching Earth's surface consists of 8% ultraviolet (UV), 42.3% visible (VIS) and 49.4% infrared (IR) radiation. As such it is an important environmental factor for algal distribution within the soil profile, for differentiation of soil algal cells as well as for regulation of sex and zoosporegenesis in eukaryotic soil algae. Due to a variety of tolerance mechanisms including the activation of several photo/dark repair mechanisms, antioxidant systems and the biosynthesis of UV-photoprotectants (e.g., mycosporine-like amino acids, scytonemin, carotenoids, polyamines), soil algae are very resilient to high UV radiation stress.

Depending on their chlorophyll type and accessory pigment setting, soil algae can use visible and infrared light with absorption maxima ranging between 430 nm and 710 nm wavelength for photosynthesis. Generally, visible light penetration of soil is stronger at longer wavelengths and deeper in wet soil compared with dry soil, thus creating distinct habitats for soil algae. Cyanobacteria, especially those containing chlorophyll *d* or *f*, can make use of the very low light conditions that occur further down the soil profile, within gypsum blocks or even in almost dark caves. In addition, and similar to other phototrophs, Cyanobacteria can regulate functions of light-harvesting antennas in response to changes in the environment, such as the spectral composition and intensity of light or temperature amplitude.

Infrared radiation affects soil surface temperature and depends on soil moisture thus, directly affecting soil algae in the upper few mm of the soil. Here, algal activity is governed by the range and fluctuation of temperature. Soil algae are extremely resistant to low temperatures and are even found in the Arctic and Antarctic regions where they can show metabolic activity below the freezing point. Also, soil algae such as *Microcoleus vaginatus* can tolerate high temperature peaks of around 110 °C.

Soil algae have developed several strategies to cope with high radiation. Some invest in the ability to move, so that they can glide into deeper soil layers. Others, rather than invest in motility, produce UV protection pigments. In addition, the symbiosis with lichen-forming fungi also can be found for some cyanobacteria and green algae. In this association, the fungal partner (mycobiont) produces pigments to protect the algal photobiont from harmful radiation. In arid deserts with high radiation, some types of soil algae prefer a hypolithic lifestyle in which they grow on the ventral site of quartz stones. These stones are whitish and remain cooler compared to their surrounding which supports water condensation on their surfaces. In addition, the translucent stones filter harsh UV radiation but at the same time allow visible radiation to penetrate which is sufficient for photosynthesis (also referred to as 'microbial cabana hypothesis').

Effect of soil moisture on soil algae

The moisture status of the soil is of great importance for soil algal growth. It ensures physiological functionality of the organisms including photosynthesis, is essential for nutrient availability and affects O₂ level as well as light and heat travel in soil. Most soil algae are poikilohydric organisms showing equilibration to ambient humidity and rapid responses upon water gain. However, both Bacillariophyceae and Xanthophyceae are very susceptible to desiccation. Some of them even depend on inhabiting very wet soil (*Botrydium* sp.) or salt marshes (*Vaucheria* sp.). A low redox potential of waterlogged soils can also lead to increased growth and enhanced N₂ fixation by cyanobacteria. For other soil algae, too much water can be harmful as the water film surrounding the cells acts as a diffusion barrier that slows down O₂ and CO₂ gas exchange thus, leading to an inhibited or reduced photosynthesis. Bacillariophyceae, for example, exhibit vertical movement within biocrusts in response to light and water availability. Many cyanobacteria are even well adapted to limited water availability. They are abundant in arid regions where they are able to maintain metabolic functions using water from dew. Green algae can also use high relative air humidity above 90% for their photosynthetic activity but also runoff water from trees (stem flow) or stones can be a sufficient water source. As an integral part of biocrusts, some soil algal communities have also been reported to decrease in their abundance and to alter their community composition under frequent wet-dry cycles.

Effect of soil pH and nutrient availability on soil algae

Soil pH is a controlling factor of the abundance and diversity of algal communities. While Cyanophyceae as well as Bacillariophyceae predominate in alkaline soils, Chlorophyceae can mainly be found in acid soil. However, soil algae can also actively change the pH of their micro-environment by mediating bio-weathering activities. Both effects, alkalinolysis and acidolysis have been reported by different soil algae that etch their substrate during nutrient acquisition.

The addition of inorganic nutrients such as available N or P can suppress algal growth but has also been found to increase algal diversity. Most algae react sensitively to increasing N to the levels at which their species richness often declines. For these reasons agricultural areas are mostly free of cyanobacteria but can be strongly colonized at least temporarily by green algae.

Although micro-habitat preferences have been observed for xeric algae either on, under, or within rock, gravel, and diaphanous substrates, little is known about the influence of soil texture on algal distribution. However, the texture of mineral soils has been stressed as an important factor for determining distribution of diatoms and cyanobacteria as it influences light, heat, water retention and infiltration rate.

Algae as soil conditioners

Free-living or symbiotically associated soil algae occur on and within rocks, gravel and stones where their biogenic weathering activities contribute to the disintegration of the solid substrate. These processes include physical impacts as a result of, e.g., swelling and shrinking of cells due to cell turgor changes. Further, chemical processes mediated by soil algae such as alkalization and acidification impact on the substrate. Over geological timescales, these processes contribute to the formation of a young soil, e.g., a terrestrial protopedon in the Southern Atacama Desert.

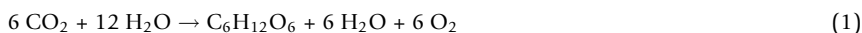
As primary producers, soil algae play an important role in the colonization of barren soils and their soil conditioning effects on soils, including agriculturally used soils, has been recognized. They improve the soil's C content by photosynthetic production of organic matter. Especially the production of EPS can aggregate soil particles thus, improving soil structure and preventing soil erosion by wind and water. Particularly the formation of biocrusts by diverse microbial communities is very successful in this respect. For example, desertification can be actively stopped by raising specific soil algae in the lab and re-introducing them into desert areas where they can stabilize the top soil layer and prevent further erosion or desertification. An increase in soil organic matter also has positive effects on soil water retention and water holding capacity. Further, soil algae can improve edaphic biomass and activity as they function as food for many soil organisms. Turnover and leakage by soil algae release nutrients, permanently but slowly, and makes them a welcome reservoir also for plant nutrients. Further, phytohormones produced by soil algae can stimulate plant (root) growth. Their metabolites are also known to contribute to phytopathogen and pest control. In addition to the above-mentioned positive effects by both, Cyanobacteria as well as eukaryotic soil algae, only the former can further condition soil N content by fixing atmospheric N₂. Cyanobacteria are also known to increase plant available P by solubilizing immobile P and can develop close association with roots. All of the effects mediated by soil algae have been shown to improve plant performance, including germination, N- and P-uptake, plant height and weight as well as crop yield.

Soil algae in nutrient cycles

Macronutrient cycling

The role in the C cycle

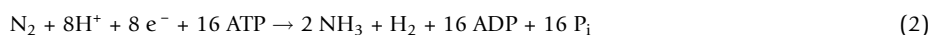
Soil algae play a pivotal role as primary producers in the carbon cycle. In the process of photosynthesis, they convert atmospheric carbon dioxide and water into oxygen and glucose (Eq. 1). For this process light energy is required.



Globally, soil algae take up around 3.6 Pg C per year. This relates to about 6% of the net primary production of the terrestrial vegetation. Within the soil, they thus contribute to the global soil organic matter pool of 2400 Gt C. Within the algal cells, glucose is converted to many different organic compounds which are vital for any functioning cell. Once in soil due to algal release of certain compounds (e.g., EPS, enzymes), leakage from algal cells or after algal cell death, organic compounds are available to all soil organisms. Eventually, any C-containing compound is mineralized again to CO₂ and water (H₂O) by soil microorganisms and the C-cycle can start again.

The role in the N cycle

In addition to carbon, some Cyanobacteria can also fix atmospheric N₂. This N fixation is vital for the replenishment of the soil N pool which provides the most important N supply for all other terrestrial auto- and heterotrophs. Globally, atmospheric N-fixation by cryptogams, of which soil algae are part of, amounts to about 49 Tg N per year, which corresponds to 46% of the total biological N fixation. Due to their nitrogenase complex, about 15% of all prokaryotes can convert N₂ to NH₃, a process which is very energy consuming (adenosine triphosphate (ATP)-expensive) (Eq. 2). Finally, ammonium (NH₄⁺; equilibration result of NH₃ in water) and α-ketoglutarate react with glutamate, an amino acid which is used in protein biosynthesis.

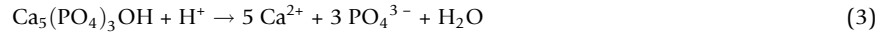


The NH₄⁺ or organically bound N is then released from the cyanobacterial cell into the soil from where it subsequently can be taken up by plants and heterotrophic soil microorganisms. In symbioses of cyanobacteria with plants or fungi (e.g., *Azolla*, *Gunnera*, Cycads, lichens) the ways of N-loss are minimized by intimate associations of the two partners. Only anaerobic denitrification products and NH₃ losses from soil replenish the atmospheric N pool.

The role in the P cycle

Soil algae can solubilize immobilized phosphorus (P) thus, essentially contributing to the P-availability in soils. Given that most organisms can only take up soluble P (PO₄³⁻, HPO₄²⁻ or H₂PO₄⁻), solubilization of P from insoluble P forms, such as primary minerals (apatites), secondary minerals (Ca, Fe, Al, Mn phosphates), sorbed P (onto clays and Al, Fe oxides) or organic compounds (e.g., phytate), is an important step in the P-cycle. P mobilization by soil algae can take place by acidification, enzymatic cleavage or chelation.

Acidification results from respiration (respired CO_2 reacts with H_2O of the soil solution to H_2CO_3), production of NH_4^+ or excretion of organic acids. For example, under low pH, P from e.g., hydroxyapatite can be set free from the crystal lattice according to Eq. (3). Also, carbonate-bound P is freed by a decrease in soil pH.



Further, to solubilize P, soil algae can produce and release phosphatases. Most cyanobacteria, green algae and lichens have phosphatases in their cell walls and mucilaginous sheaths. These extracellularly working enzymes hydrolyze organic phosphates thus, liberating P from organic compounds such as inositol phosphates or sugar phosphates. The most intensively investigated phosphatases belong to the group of either phosphomonoesterases or phosphodiesterases. Phosphomonoesterases are further classified according to their pH optima as alkaline or acid phosphatases. Phosphatase expression is mainly substrate induced. In the case of phosphomonoesterase, activity is inhibited by the hydrolytic product phosphate. Cyanobacteria, however, may respond to internal P concentrations rather than external phosphate concentrations.

Cyanobacteria can also produce chelators to solubilize P. Metal chelators such as siderochromes thereby complex Ca, Al, and Fe to release P originally bound by these metals. Also, other water-soluble substances such as peptide-N or riboflavin can be secreted.

Micronutrient cycling

Soil algae are not just vital for the macronutrient cycles of C, N and P but also play a role in micronutrient cycling. Commonly, micronutrients are metal constituents of enzymes mediating biochemical reactions. This includes reactions such as those in the photosynthetic process, the respiratory chain or the nitrogenase complex for N fixation. For soil algal growth, elements such as Fe, Mn, Cu, Zn, Mo, V, B, Cl, Co and Si are essential at concentrations below 10^{-5} M and are added as micronutrient solutions to culture media in the lab. The most commonly used culture media (including various adaptations) for green algae is Bold's Basal Medium (BBM) and that for Cyanobacteria is BG 11. Soil algae are able to mobilize those micronutrients from soil, take them up and accumulate them within their biomass. Excreted chelators may help in the process of mobilization. Once in a soluble state these micronutrients can also be taken up by other organisms. As such, soil algae growth can contribute to an increased supply of micronutrients for other organisms including crop plants and can even influence plant properties such as taste and aroma or the ability of plants to withstand drought periods. The accumulation of micronutrients within soil algal biomass further provides a high amount of these nutrients to the soil after soil algal death.

Biotic interactions of soil algae

As primary producers, soil algae convert atmospheric CO_2 into glucose and from there also into other organic carbon compounds. Hence, they are an important carbon source for many heterotrophic soil organisms including protozoa, tardigrade, earthworms or isopods which feed on living algae. If not eaten, they represent a reservoir for carbon but also inorganic nutrients before they function as a food for those organisms which live on compounds released from dead organic matter.

Soil algae are also known to be partners in a facultative mutualistic symbiosis. Cyanobacteria *Nostoc* sp. and *Anabaena azollae*, both of which can fix atmospheric N_2 in heterocytes, live in close association with plants. While *Nostoc* sp. can be found in the roots of Cycads or the rhizome of *Gunnera* sp., *Anabaena* sp. resides in leaf cavities of the freshwater fern *Azolla* sp. Here, they provide N as amino acids and extracellular NH_3 to the plant. In the fast-growing *Azolla* fern, *Anabaena azollae* may fix up to about $2 \text{ kg N ha}^{-1} \text{ day}^{-1}$ thus making *Azolla* application a good practice for sustaining soil fertility and crop productivity in paddy rice cultivation worldwide.

In the obligate mutualistic symbiosis of lichens, soil algae represent the photobiont which is closely associated with the fungal mycobiont. The mycobiont forms a thallus in which the photobiont is concentrated within an algal layer or algal stacks (heteromorous thallus) or is irregularly distributed within the whole thallus (homoimorous thallus). Photobiont species belong either to the phyla Cyanophyta or Chlorophyta. Of all known lichens, about 10% are cyanobacterial lichens (cyanolichens) with *Nostoc* sp. as the predominate photobiont. In green algae lichens (chlorolichens), 24 different genera may act as photobiont but *Trebouxia* sp. occurs in more than 75% of all green algal lichens. Both types of photobionts provide sugars for the mycobiont. Cyanobacteria, in addition, provide N in some cases. The mycobiont is one of about 13,500 species belonging to either the ascomycetes (predominating) or basidiomycetes. It protects the photobiont from desiccation and strong UV radiation, improves its P supply and offers a joint reproduction strategy. Generally, in lichen symbiosis, photobionts reproduce only asexually while mycobionts reproduce asexually as well as sexually. The meiospores are produced within apothecia which often are characteristic structures of conspicuous character on the lichen thallus. For simultaneous reproduction of photo- and mycobiont, special reproduction organs are formed by the lichen. Isidia are outgrowths which get distributed by wind and water after easy separation at a breaking point. Soralia on the other hand are breakup regions. Both structures are made of fungal hyphae wrapped around photobiont cells which act as propagules and regrow full lichens under adequate conditions.

As free-living organisms or as part of lichens, soil algae can be observed in biocrusts. In these associations they play the role of primary producers which are tightly associated with heterotrophic organisms as well as their byproducts in the upper few mm of soil. A characteristic biocrust type has been detected in the Atacama Desert. Here, up to 80% of ~ 0.6 cm grit stones are colonized by a diverse lichen community which even concatenate these grit stones. Biocrusts can also be found on ice free areas of the Arctic and Antarctica.

Outlook

Using microalgae in biotechnology, offers innovative solutions to meet the growing demand for bioactive ingredients, food and feedstock with low resource consumption. To date, however, limitations in physiology and bioreactor design have prevented an economically feasible transfer of numerous microalgae-based production processes into an industrial scale. Soil algae are especially interesting for their biotechnological potential due to their low maintenance and useful properties. The previous focus on aquatic microalgae for biotechnological approaches might be due to the fact that they are highly abundant in freshwater and marine habitats displaying the main driver of aquatic primary production. Moreover, they are—at the first glance—easy to cultivate in liquid cultures. The outstanding role of terrestrial microalgae, e.g., in nutrient-limited ecosystems as soil stabilizer and N supplier, was only recognized three decades ago. Compared to aquatic strains, soil algae can be grown under low light conditions, ambient air temperature and low nutrient concentrations. In addition, high value products such as pigments can easily be concentrated in soil algae once stressed by short-term high-light exposure. The high content of proteins and lipids of soil algae will be a commercial factor influencing the future agronomical systems.

Due to the adaptability of cyanobacteria, astrobiologists are drawn to the idea that cyanobacteria could be used to terraform Mars. Terraforming is defined as a process of planetary engineering, specifically directed at enhancing the capacity of an extra-terrestrial planetary environment to support life including that of human beings. Based on data collected so far, Mars is fairly different from Earth but remains the same in a few key characteristics. While some climatic parameters are comparable and atmospheric composition, and surface pressure are relatively similar to Earth, Mars has lower gravity, lower average sunlight and longer years/seasons compared to Earth. Hence, Earth's soil algae may contribute to terraforming Mars given that they are able to inhabit also low light environments and can cope with long dry periods.

In summary, the microbial world of soil algae at our feet is often overlooked but significantly contributes to major global processes from the micro- to the macro scale and holds a great potential for biotechnology and human use.

Further reading

- Adhikari K, Bhandari S, and Acharya S (2021) An overview of *Azolla* in rice production: A review. *Reviews in Food and Agriculture (RFNA)* 2. <https://doi.org/10.26480/rfna.01.2021.04.08>.
- Alvarez AL, Weyers SL, Goemann HM, Peyton BM, and Gardner RD (2021) Microalgae, soil and plants: A critical review of microalgae as renewable resources for agriculture. *Algal Research* 54: 102200. <https://doi.org/10.1016/j.algal.2021.102200>.
- Baumann K, Jung P, Samolov E, Lehnert LW, Büdel B, Karsten U, Bendix J, Achilles S, Schermer M, Matus F, Osés R, Osés P, Morshedizad M, Oehlschläger C, Hu Y, Klysubun W, and Leinweber P (2018) Biological soil crusts along a climatic gradient in Chile: Richness and imprints of phototrophic microorganisms in phosphorus biogeochemical cycling. *Soil Biology & Biochemistry* 127: 286–300.
- Belnap J. (2011) Biological phosphorus cycling in dryland regions. In: Bünemann, E. K., Oberson, A. & Frossard, E. (eds.) *Phosphorus in Action. Biological Processes in Soil Phosphorus Cycling*. 1st edn., pp 371–406, Berlin, Heidelberg: Springer-Verlag.
- Belnap J, Weber B, and Büdel B (2016) Biological soil crusts as an organizing principle in drylands. In: *Biological Soil Crusts: An Organizing Principle in Drylands*, pp. 3–13. Chambridge: Springer-Verlag.
- Elbert W, Weber B, Burrows S, Steinkamp J, Büdel B, Andreae MO, and Pöschl U (2012) Contribution of cryptogamic covers to the global cycles of carbon and nitrogen. *Nature Geoscience* 5: 459–462.
- Jassey VEJ, Walcker R, Kardoly P, Geisen S, Heger T, Lamentowicz M, Hamard S, and Lara E (2022) Contribution of soil algae to the global carbon cycle. *New Phytologist*. <https://doi.org/10.1111/nph.17950>.
- Jung P, Schermer M, Briegel-Williams L, Baumann K, Leinweber P, Karsten U, Lehnert L, Achilles S, Bendix J, and Büdel B (2019) Water availability shapes edaphic and lithic cyanobacterial communities in the Atacama Desert. *Journal of Phycology* 55: 1306–1318.
- Kula-Maximenko M, Zieliński KJ, and Ślesak I (2021) The role of selected wavelengths of light in the activity of photosystem II in *Gloeobacter violaceus*. *International Journal of Molecular Sciences* 22: 4021. <https://doi.org/10.3390/ijms22084021>.
- Metting B (1981) The systematics and ecology of soil algae. *The Botanical Review* 47: 195–312.
- Rastogi RP, Madamwar D, Nakamoto H, and Incharoensakdi A (2020) Resilience and self-regulation processes of microalgae under UV radiation stress. *Journal of Photochemistry & Photobiology C: Photochemistry Reviews* 43: 100322.

Soil arthropods: Underfoot and all around

Ernest C Bernard, Department of Entomology & Plant Pathology, University of Tennessee, Knoxville, TN, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	71
Detrital food webs and functional groups	73
Collecting soil arthropods	75
Arachnida	77
Acari (mites)	77
Food	77
Habitats	77
Interactions	77
Phoresy or not	77
Araneae (spiders)	79
Food	82
Habitats	82
Interactions	83
Opiliones	84
Pseudoscorpiones	86
Basal hexapods	87
Collembola (springtails)	87
Food	87
Habitats	89
Interactions	89
Diplura	89
Protura	92
Food	92
Habitats	92
Interactions	92
Interesting fact	92
Archaeognatha	92
Food and habitat	92
Interactions	93
Zygentoma	94
Food, habitats, interactions	94
Crustacea	94
Food	94
Habitats and interactions	94
Myriapoda	95
Chilopoda (centipedes)	96
Food	97
Habitats	97
Interactions	97
Diplopoda	97
Food	98
Habitats	98
Interactions	98
Paupoda	98
Food	99
Habitats	99
Interactions	99
Symphyla	99
Food	99
Habitats	99
Interactions	99
Camouflage, mimicry and defense	99

Conclusion	102
Acknowledgments	102
References	102
Further reading	103

Abstract

This chapter is a synopsis of the soil arthropods that which exist in soil and its associated organic litter layers and form part of the detrital food web. The arachnids include mites, spiders, harvestmen and pseudoscorpions; the hexapod classes are Protura, Collembola, Diplura and Insecta; crustaceans are represented by the pillbugs; and the myriapods consist of centipedes, millipedes, pauropods and symphylans. Each taxon is a separate and distinct entity, not like any of the others and deserving of study. Phoresy by mites is covered, as are the myriad approaches some arthropods take to avoid being noticed by a predator.

Glossary

Basal hexapod An informal grouping of three separate classes of six-legged non-insect arthropods (Protura, Collembola, Diplura) plus two orders of true insects (Archaeognatha, Zygentoma). The five taxa are distinctive in never having had winged ancestors, unlike all the other insects. "Apterygota" is an increasingly obsolete term for these arthropods.

Batesian mimicry Protective mimicry in which a harmless species has evolved to resemble the appearance of a harmful or more dangerous species. Counterpart to Müllerian mimicry.

Clade A grouping of related taxa based on phylogenetic analysis. The taxon level can be at any level from species to phylum.

Class The first major taxonomic division below phylum. In this paper, several classes of the phylum Arthropoda are discussed.

Müllerian mimicry Protective mimicry in which two or more unrelated species with similar noxious or unpleasant defenses develop similar appearances. Often contrasted with Batesian mimicry.

Phoresy A dispersal behavior of many invertebrates, especially larval mites and pseudoscorpions, in which they cling or attach usually to a larger arthropod for transport to new localities or food sources.

Sympatry The condition of organisms living together in the same place at the same time. Usually applied to related species living and presumably interacting in the same restricted area but using available resources somewhat differently. The opposite term is allopatry.

Key points

- Soil arthropods form an important part of the detrital and grazing food webs, deconstructing organic matter so that it can eventually be recycled into nutrients usable by photoautotrophs, especially plants.
- The soil arthropod "universe" comprises a full range of detritivores, predators and parasites with innumerable interactions that keep undisturbed systems in balance.
- Mites (Acari) are by far the most abundant and diverse soil arthropods, followed by Collembola (springtails). Spiders and beetles are their most significant predators, but many other smaller taxa, such as centipedes, are effective predators as well.
- Evolution has produced a dizzying array of morphological, behavioral and reproductive specializations to enhance the survival of each species.
- Camouflaged appearance, mimicry of stinging insects and production of noxious exudates are important means by which many soil arthropods protect themselves.

Introduction

The forces of gravity and weather sooner or later bring terrestrial life back to earth as dead organic matter. This dead matter provides the energy to drive the detrital food web (DWB), the components of which include nearly every class and order of invertebrates. At the same time, the grazing food web, which we usually think of as above the soil surface, is active on plant roots and other subterranean plant parts (e.g., plant-parasitic nematodes, cicada nymphs, beetle larvae). These grazing organisms eventually die and thus become incorporated themselves into the DWB. The ultimate function of the detrital food web is to break down complex organic materials into nutrients that can be used by photosynthetic organisms (photoautotrophs), which thus drives the maintenance of terrestrial life on the planet.

The DWB operates in the zone between bedrock and the surface, where organic material is present for degradation. Organisms of the DWB typically are absent from deep soil layers that contain little organic matter. In general, most decomposition occurs in the top meter of soil where most of the organic matter resides. Much of this decomposition is fostered, especially in its early stages, by arthropods living in this soil depth.

Table 1 Subhorizons of the O-horizon and adjacent A-horizon.

Designation	Name	Description
<i>E&W</i>	<i>Chapin et al.</i>	
O _i	O _e	Litter layer
O _r	O _i	Plant debris not yet decomposed
O _h	O _i	Partially decomposed plant debris, recognizable fragments of vegetation
O _h	O _a	Fermentation layer
A _h	A	Humus layer
		Plant residues so decomposed that definite structure is not observed with microscopy
		Enriched A-horizon
		Humus-enriched, dark upper mineral layer

From Eisenbeis G and Wichard W (1987) *Atlas on the Biology of Soil Arthropods*. Springer-Verlag: Berlin. 437 pp; Chapin FS III, Matson PA and Vitousek PM (2011) *Principles of Terrestrial Ecosystem Ecology*, 2nd edn. Springer: New York. 529 pp.

The differentiation of organisms strictly associated with soil from those in detrital layers covering the soil surface may be difficult since movement between the layers is common and sometimes both layers are necessary for life cycles to be completed. The soil profile conventionally is divided into four horizons: O (organic), A (surface), B (subsoil) and C (substratum). The O-horizon, in turn, is expansively refined (Table 1) by sublayers, which are important because the size and connectivity of the pores within, help determine the distribution and movement of soil fauna in the organic horizon. Roots, the channels left after roots die, earthworm and arthropod tunnels and vertebrate nests also are significant means allowing non-burrowers to move about more efficiently.

Arthropoda is the largest phylum of animals, differentiated by possession of an exoskeleton and jointed appendages. Its major classes are Arachnida (primarily mites, spiders, scorpions), Crustacea (copepods, crabs, lobsters, woodlice), Hexapoda (basal hexapods, true insects) and Myriapoda (millipedes, centipedes). Soil arthropods comprise those members of the phylum that spend a significant amount of their time living in or actively modifying the micro-environment of their soil or detrital habitat. Every extant terrestrial arthropod class has soil-dwelling members (Table 2) and nearly every order of insects has members adapted to life in the soil in at least one stage; even in Lepidoptera (moths and butterflies) some larvae live in the organic layers and many species pupate in soil. Ants and termites are major players in soil turnover and nutrient cycling (along with earthworms), but they will be mentioned only incidentally in relation to other groups as there is a large literature devoted to them. This chapter is focused mainly on the many non-insect arthropods that occupy every nook and cranny of the strata under our feet (Figs. 1 and 3). Short synopses are provided for each order within the groups of soil arthropods that are covered, together with numerous original images to interest and inform the reader. Further Reading may be consulted for additional sources and web sites, respectively.

Table 2 Higher classification and sources of general information on major taxa of soil-dwelling arthropods.

<i>Core sources</i>	<i>Coleman and Hendrix (2000), Coleman et al. (2018), Crossley Jr. et al. (1991), Dindal (1990), Eisenbeis and Wichard (1987) and Minelli (2011, 2015)</i>	
Class, Order	Common name	References
Arachnida		
Acari	Mites	Houck (1994), Walter and Proctor (1999), and Krantz and Walter (2009)
Araneae	Spiders	Foelix (2011) and Ubick et al. (2017)
Opiliones	Harvestmen	Pinto-da-Rocha et al. (2007)
Pseudoscorpiones	Pseudoscorpions	Buddle (2010) and Harvey (2011)
Collembola ^a	Springtails	Hopkin (1997)
Diplura ^a	Campodeids, japygids	Allen (2002)
Protura ^a	Proturans	Tuxen (1964), Imadate (1974) and Szeptycki (2007)
Insecta		
Archaeognatha ^a	Jumping bristletails	Sturm and Machida (2001), Wygodzinsky and Schmidt (1980)
Zygentoma ^a	Silverfish	Chick (2018), Smith (2017) and Wygodzinsky (1972)
Crustacea, Isopoda	Woodlice, Pill bugs, sow bugs	Warburg (1987)
Myriapoda		
Chilopoda	Centipedes	Lewis (1981) and Minelli (2011)
Diplopoda	Millipedes	Minelli (2015)
Pauropoda	Pauropods	Scheller (2011)
Symphyla	Symphylans; garden centipedes	Szucsich and Scheller (2011)

^aBasal hexapods.



Fig. 1 Maintenance of healthy soils is dependent upon our knowledge of and respect for decomposition processes: a sign along an interpretive trail in Frozen Head State Park, Tennessee, USA. Photo courtesy of Adrian Smolis.

Detrital food webs and functional groups

The DWB comprises all of the organisms, from prokaryotes to the annelids and arthropods, that contribute to the reduction of organic matter to a state where its basic constituents can be reabsorbed by plants. This process mineralizes carbon and releases nutrients that are utilizable by autotrophs (Potapov et al. 2022). The species richness of organisms in soil is enormous, albeit imperfectly known, but richness is a critical factor in ecosystem functioning, as . . . "environmental heterogeneity strengthens the relationship between biodiversity and ecosystem functioning. . . ." (Albrecht et al. 2021). This result highlights the importance of species assembling into groups of organisms doing similar activities with similar results. Organisms in soil environments have been grouped into a small number of functional groups whose ecological niche dimensions vary from group to group. The application of molecular techniques has revealed numerous species complexes that enhance the understanding of functional group effects.

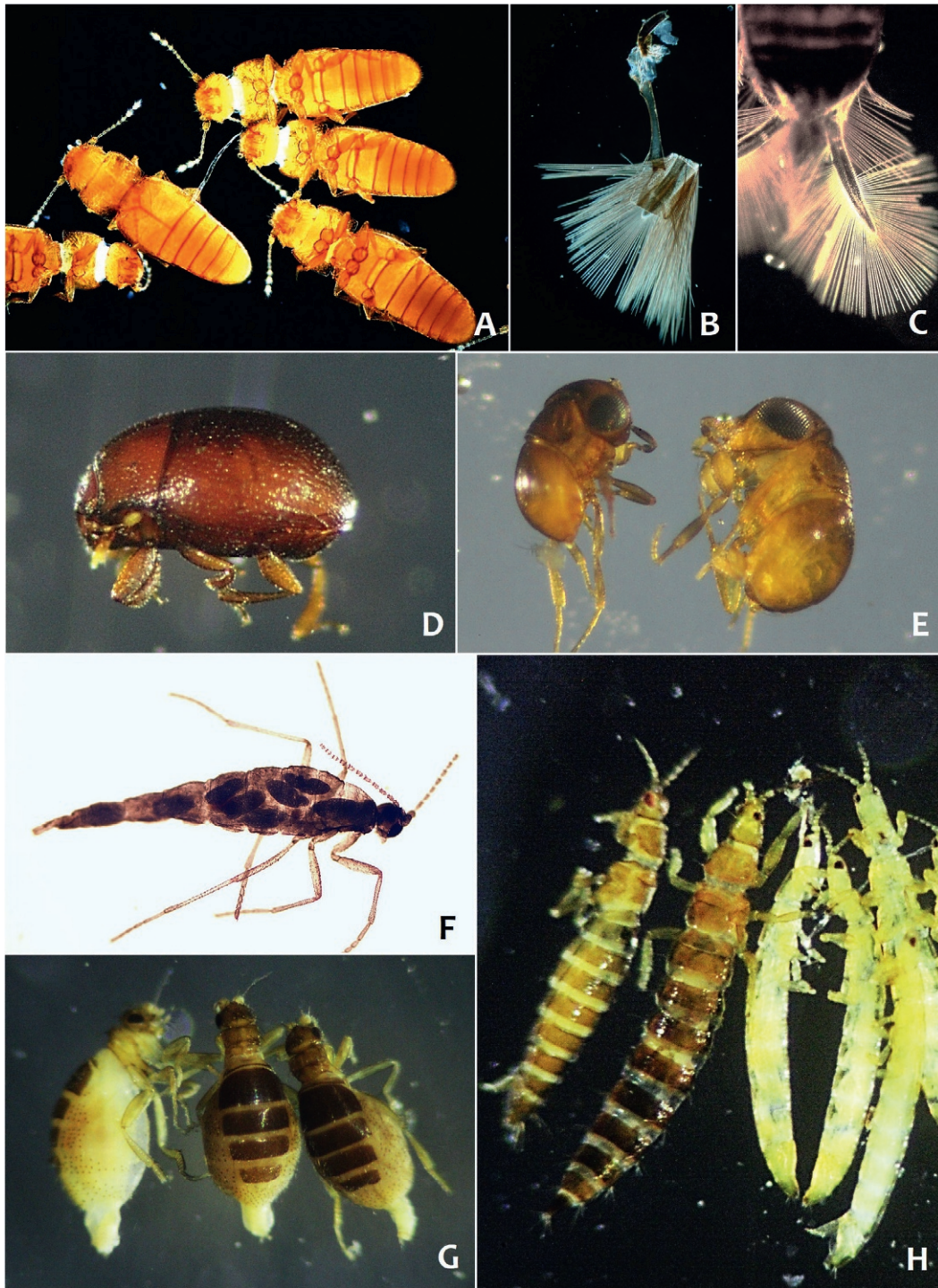


Fig. 2 Think small! The litter/soil layers contain the some of the smallest known insects (length in mm). A–C, featherwing beetles (Ptiliidae). (A) *Ptinella* sp., a wingless species (0.9). (B and C) Folded and extended wings of winged species. (D) A hister beetle (0.8) (Histeridae). (E) Two wingless scelionid wasp females, *Baeus* spp. (Scelionidae). (F and G), Wingless female flies. (F) Sciaridae (1.6). (G) Phoridae (1.2). (H) Thrips (1.8) (Thysanoptera).



Fig. 3 (A) Typical Tullgren funnel extraction from a mixed deciduous forest. The beetle at lower left is a ptilodactylid; two centipedes are present; the large springtail in the lower center is a *Pogonognathellus* sp. (Tomoceridae); most of the white individuals are onychiurid springtails, the light violet ones are mostly Isotomidae; and the many shiny brown specimens are oribatid mites. (B) An extraction of mites, with several orders represented.

Collecting soil arthropods

"If you don't look, you'll never find anything." — Gary J. Phillips (2014).

Soil arthropods can be collected using several methods, the choice of which depends on the type of investigation and the expected sizes of the organisms (Fig. 2). Tullgren funnels and pitfall traps are two standard approaches but by no means the only ones. Many entomology texts provide descriptions of methodologies and should be consulted for more extensive descriptions (Fig. 3).

The Tullgren funnel (Fig. 4A and B, a modification of the Berlese funnel) uses light and heat from a low-wattage incandescent bulb to drive organisms downward through the substrate placed on the screen. Fleeing arthropods fall through the screen into a container of preservative (usually ethanol) or a culture chamber. The volume or mass of substrate can be measured in advance to provide quantifiable estimates of arthropod density. The unit in Fig. 4 uses a small 15-watt bulb over a 20-cm-diam #4 sieve (4.75-mm apertures). Specimens fall directly into 70–95% ethanol. Isopropanol can be used, as well, at a pinch. Both compounds are suitable for DNA preservation. Typically, there is no need for a surfactant; after the container is capped it can be inverted a few times to wet any floating specimens so that they sink to the bottom. This kind of Tullgren funnel is most suitable for the organic horizons. Mineral soils can be more efficiently extracted by using a finer-meshed sieve or laying a piece of window screen in the sieve. A more elegant and efficient modification of the Tullgren funnel for soil is the high-gradient extractor.

Pitfall traps (Fig. 4C and D) are widely used for soil arthropod surveys. A cylinder of soil is removed with a bulb planter or dug with a trowel to allow the insertion of a cup flush with the substrate surface. A small bulb planter will work in loose soils, but the planter style in Fig. 4C is much more efficient in heavier soils and when many traps are being deployed. The cup is partially filled with a preservative, typically ethylene glycol mixed with other liquids; however, propylene glycol is better because it is less toxic to curious wildlife that occasionally disturb traps. Traps should be covered with a lid as protection from rain and debris and weighted with a rock or brick to prevent smaller mammals from lifting the lid. Pitfall traps are best used for repeated field bioassays on a monthly, bimonthly or seasonal basis to measure phenologies of the target taxa. Pitfall traps provide a relative measure of arthropod



Fig. 4 Collection methods. (A and B) Tullgren funnels. (C–E) Pitfall traps. (F) Shaker box. (G) An aspirator.

field density that fall into the traps, but cannot measure true density because we don't know the relative propensity for each taxon to fall into the trap, and we do not know the extent of the surrounding area from which individuals are drawn.

A shaker or sieve made of food containers and a piece of coarse screen is a good way to separate small arthropods from organic litter (Fig. 4D and E). A handful of surface debris is placed in the apparatus and shaken over a white pan, after which individuals can be gathered with an aspirator or dumped directly into preservative. The method can be made somewhat quantitative by shaking equal masses for a predetermined period of time before aspiration.

It must be noted that molecular analysis of specimens for identification, intestinal contents and chemical composition is becoming a near-essential approach in ecological and behavioral studies for objectives, such as identification, identification of food and microbiomes, and chemical composition. Tullgren funnel extraction, collection via shakers or individual collection by sight into 95–100% ethanol will provide much better specimens for molecular work than most pitfall trap collection liquids.

Arachnida

Spiders, mites, harvestmen, scorpions, pseudoscorpions and about a dozen other small taxa are included in the class Arachnida. They are united in the possession of a body with two tagmata (major sections): the prosoma, equivalent to the head + thorax (thus also termed the cephalothorax), and the opisthosoma (or abdomen). Even when the body appears to be a single unit, as in mites and harvestmen, the distinction is still usually obvious, at least on the ventral side. Arachnids have eight legs as adults and two pairs of distinctive mouthparts, the chelicerae and the pedipalps. These organisms have a limited number of juvenile stages and do not molt after reaching adulthood.

Acari (mites)

The most biodiverse taxon of non-insect soil and litter-dwelling arthropods. Their higher classification has been altered in recent years due to new molecular-based studies, but no matter the arrangement, mites are extraordinarily diverse, with nearly 350 recognized families and over 48,000 described species. Nevertheless, many authors believe that may be just a fraction of the actual number. The discussion below is very general as mites are so morphologically, ecologically and behaviorally diverse. Ticks also are members of the Acari and occasionally fall into pitfall traps but are not considered further in this discussion.

In addition to being mega-diverse, mites occupy every conceivable terrestrial niche, with an astonishing variety of morphologies (Fig. 3B and Figs. 5–9) and life strategies. Mites have two tagmata: an anterior prosoma (or cephalothorax) that combines the feeding and sensory functions of the head with the more posterior part with the eight legs, and an opisthosoma, the body region posterior to the legs and bearing the genital orifice and the anus. Like other arachnids, mites have a pair of pedipalps and a pair of chelicerae for sensory and feeding functions. Some families of mites have a single pair of eyes, but otherwise mites are largely blind. Mites navigate their environment with sensory input from pores, setae and other organs on the body and appendages. Mites lay eggs, which hatch as six-legged larvae (Fig. 6A), followed generally by three instars before coming adult. Mites do not molt after reaching maturity.

Food

The anterior mouthparts function to grip, macerate and usually liquefy food items before ingestion. Pedipalps detect suitable food and may be modified to hold or tear prey. Reddish trombidiform mites (Fig. 5A–D) are common in spring and are easily seen, sometimes in large numbers, running over the ground or on logs and low plant material, seeking small arthropods for food. The pedipalps are modified for piercing and are much larger than the inconspicuous chelicerae. In most mites, however, the chelicerae are better developed as pincer-like claws (chela). The shapes of these chelicerae indicate the probable food of mites. Predacious mesostigmatid mites (Fig. 5E and F) have strongly toothed chelicerae for gripping and tearing prey, whereas detritus-feeding oribatid mites (Fig. 5G–J) have heavy pliers-like chelicerae for cutting tough bits of organic matter. Other predators include bdellid mites (Fig. 6B and C), with long, slender chelicerae that can reach deep inside prey and churn it up. The distinctive yellow astigmatid genus *Labidostoma* (Fig. 6D and E) has the movable digit on the claw interlocking with the fixed digit, reducing the chance for prey to get away. These variations are just a fraction of the variability in chelicera shape and function. Body form in soil mites tends to be correlated with herbivory or predation. Predators have long legs, long pedipalps and have slightly to moderately thickened or leathery cuticles. These mites can flee when confronted with a larger enemy. Slower-moving detrital feeders, exemplified by the enormous suborder Oribatida, usually have an armored body, resistant (although not impervious) to penetration by claws and teeth. Some oribatids (“trash mites”) have modified setae that allow accumulation of shed exoskeletons and other debris on their backs to discourage or fool predacious mites, spiders and insects (Fig. 21A).

Habitats

Mites are abundant nearly everywhere other life exists. Most ground-dwelling mites live in the leaf litter zone but the more active predators may roam freely up and down plant stems searching for their next meal. In this behavior they overlap with the many plant-feeding mites.

Interactions

Given their abundance, mites constitute a significant food resource for other soil invertebrates. Several families of small beetles are specialists on armored oribatid mites, including the innocuous-looking Ptiliidae (Fig. 2A), which use a variety of specialized methods to breach the cuticles of these mites.

Phoresy or not

“We gotta get out of this place, if it’s the last thing we ever do...” — The Animals (1965).

Phoresy is the behavior by which invertebrates, usually with limited dispersal capabilities, are transported by another organism to new habitats and resources. This mechanism is particularly common for organisms that exploit ephemeral food resources in patchy environments, for example, cattle droppings in a large pasture. These rich sources of food for nematodes, insects and mites are rapidly exploited and depleted, requiring transport, usually via beetles and flies, to a fresher source. Even in seemingly stable

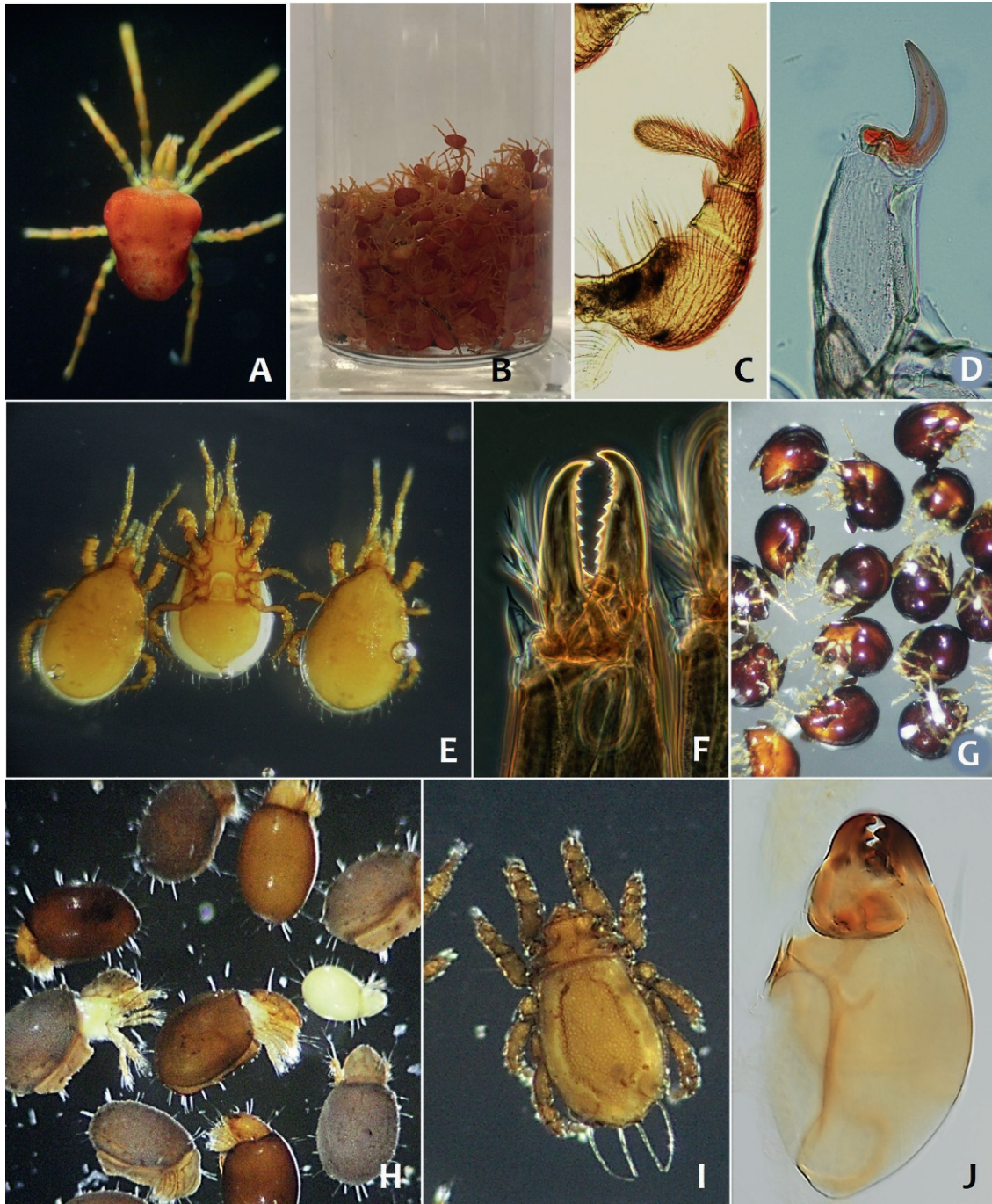


Fig. 5 Mites I. (A–D) Trombididae (Order Trombidiformes). (A) Habitus of a red velvet mite. (B) Approximately 300 collected in a single pitfall trap. (C) Pedipalp. (D) Chelicera. (E and F) Order Mesostigmata. (E) Two of many forms. (F) Chelicera. (G–J) Oribatida. (G) Galumnidae, note lateral “wings”. (H) Phthiracaridae. (I) Nothridae. (J) Oribatid chelicera.

environments, a distance of a few centimeters may be too much for some small arthropods, and they may need to hitch rides on forest-floor insects or millipedes. Failure to reach a new favorable micro-environment can mean failure to reproduce.

Several groups of mites have evolved supremely effective means of catching rides on other arthropods, and some avail themselves of the opportunity to feed on their transport (Figs. 7–9). Phoresy is especially common on insects and millipedes. The third instar (deutonymph) is the usual passenger; it may be passively along for the ride or also actively feed on its transport. Bess beetles (Passalidae) are almost always infested with phoretic mites of more than one order. Three mesostigmatid species on the same beetle are imaged in Fig. 7A–D: a large species, a very small species (Fig. 7B and C), both favoring the thoracic venter, and a related species found only on the abdominal sternites (Fig. 7D, note differing setae on anterior end). Astigmatid deutonymphs

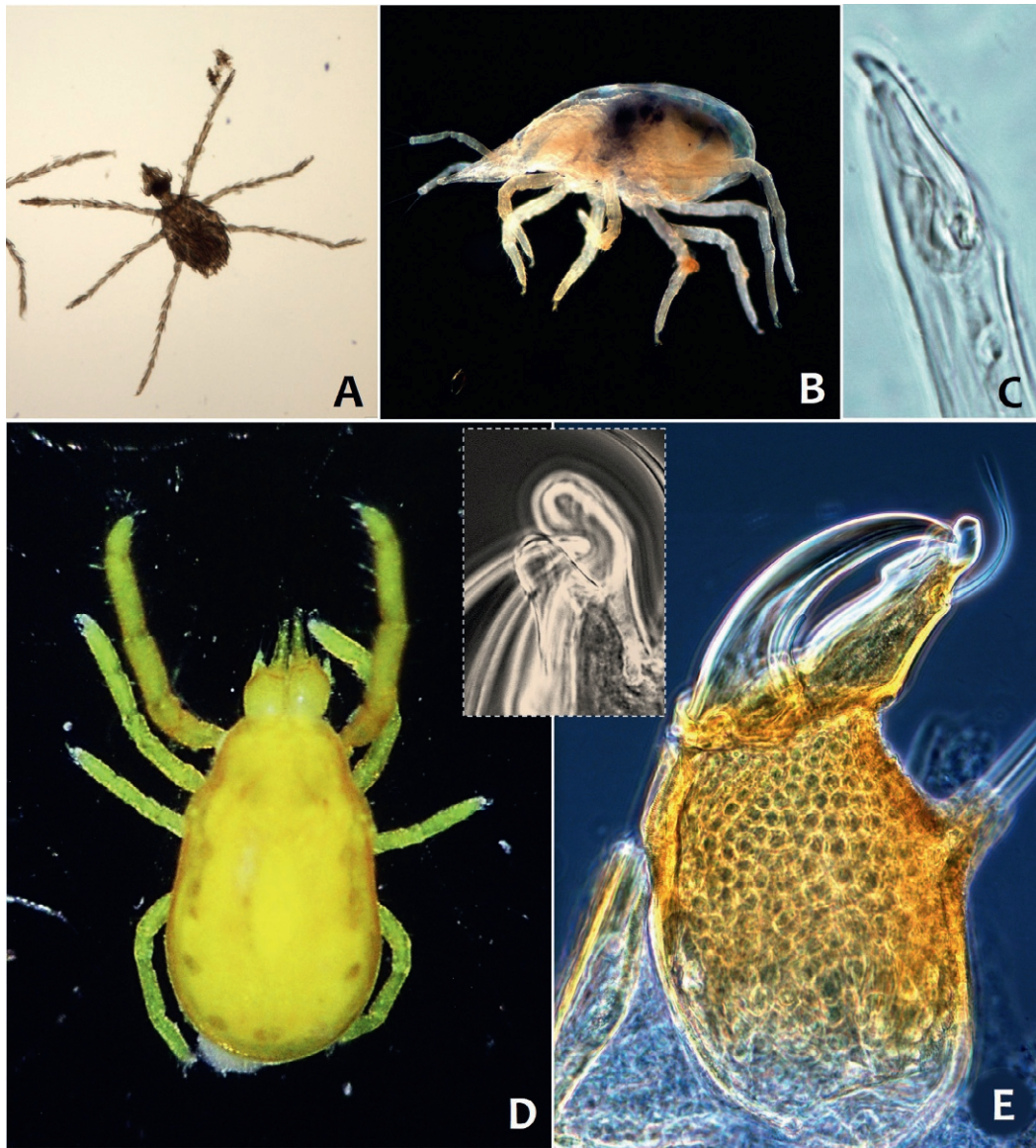


Fig. 6 Mites II. (A) Six-legged larva of Cunaxidae, body length 0.15 mm. (B and C) Bdellidae, entire specimen and chelicera. (D and E) *Labidostoma* sp. (Labidostomatidae), dorsal view and chelicera. Close-up of chelicera tip shows interlocking of digit and claw.

occur on millipedes (Fig. 7E–G), sometimes in large numbers. These mites have a specialized cluster of adhesive pads on the ventral side of the body, giving them a firm hold on the millipede. These phonetics can be gently removed and placed on agar streaked with bacteria. In a few days they will mature to adults and begin to lay eggs.

Another interesting approach to phoresy is that of uropodid deutonymphs that produce a stalk firmly cemented onto a beetle's body or appendages (Fig. 8A–C) (Bajerlein et al. 2016). The imaged scolytid beetle would seem to be aerodynamically disadvantaged, but it was caught in a flight intercept trap and must not have been significantly impaired.

Many apparent arthropod-mite phoretic relationships are really the larval mite stage feeding on a host and just incidentally getting moved somewhere else (Figs. 8D and 9). In Figs. 8A and 9E mite mouthparts appear to be firmly embedded in the host. These larvae may eventually detach and continue their development, but they lack the specializations of phoretic deutonymphs and therefore are not true phoretic mites.

Araneae (spiders)

Eight-legged predacious arthropods that make silk webs. Spiders have always been unjustifiably feared because of their venomous bites; but only a few of the 50,000 known world species are capable of doing any harm to humans whatsoever, and none actively chase humans to bite them. Rather, spiders occupy a key role in regulation of invertebrate communities by their feeding activity. All

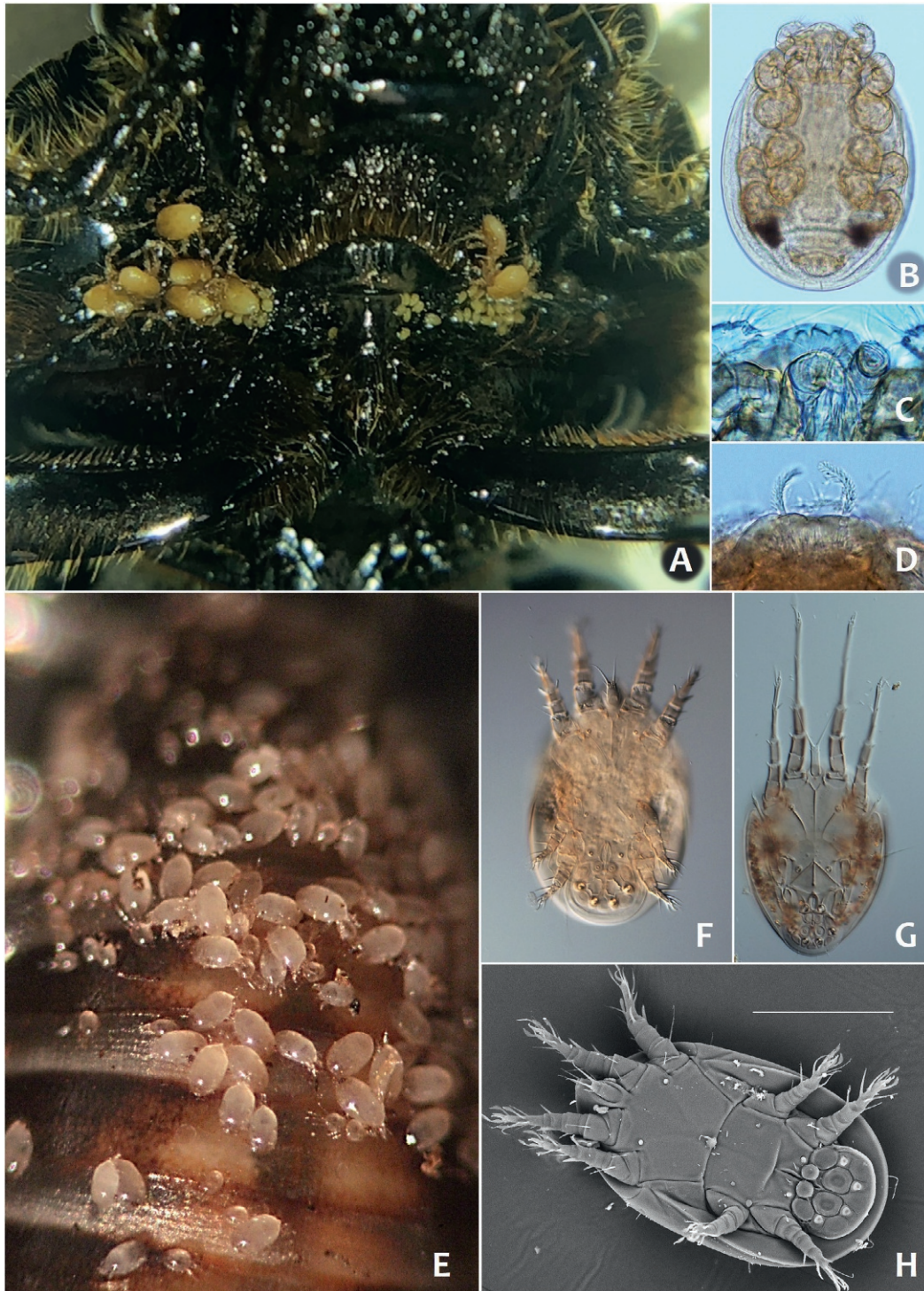


Fig. 7 Mite phoresy I. (A–D), mites on a bess beetle (*Odontotaenius disjunctus*). (A) Mesostigmatid (larger) and presumed hypopi (smaller) at the ventral cephalic-prothoracic junction. (B and C) Closeup view of smaller mites and head end. (D) Related species collected from venter of abdomen. Note feathery anterior setae. (E–H), mites phoretic on the millipede *Choctella hubrichti*. (E) Two phoretic species intermingled on cuticle. (F) *Sancassania anomala* (Acaridae), ventral view. (G) *Histiotoma feroniarum* (Histiotomatidae), ventral view. (H) SEM of *S. anomala* clearly showing ventral suckers. Identifications of mites in (F–H) by Barry O'Connor, SEM by Gary Phillips.

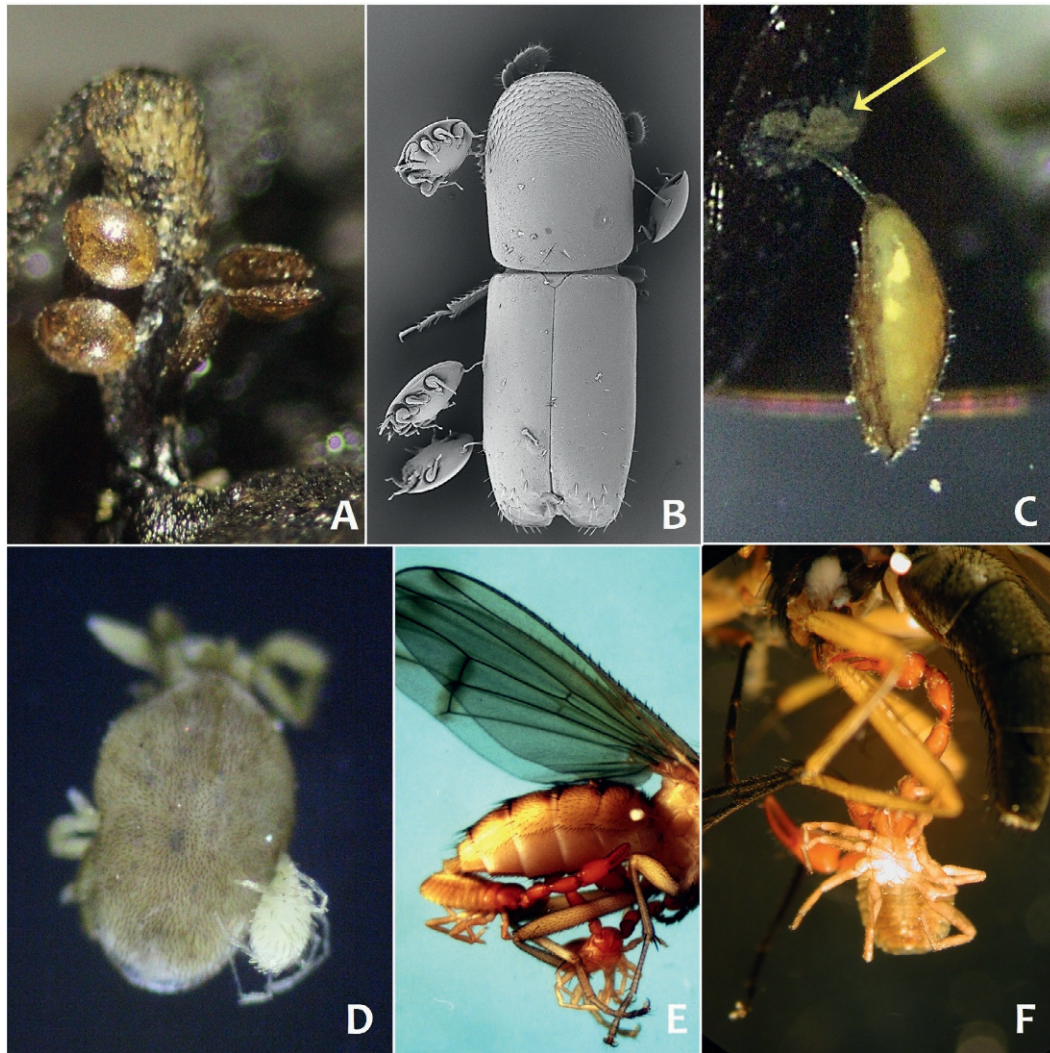


Fig. 8 Phoresy II. (A–C), Uropodid deutonymphs attached to insects. (A) On leg of a pinned ironclad beetle (Zopheridae). (B) Attached to a bark beetle (Curculionidae, Scolytinae). (C) On leg of a nitidulid beetle; arrow indicates stalk base. (D) Trombidiform mite fed upon by larval hitchhiker. (E and F) Pseudoscorpions gripping legs of flies, collected from Malaise trap.

spiders are predators on other animals, with the largest species capable of subduing and feeding on small mammals. Some species of semi-aquatic Pisauridae can catch small fish. All spiders produce silk from spinning glands in the abdomen for myriad uses but many active ground-dwelling species do not make prey-catching webs.

The order Araneae contains two infraorders. Mygalomorphae has chelicerae that parallel each other (paraxial), pointing posteriorly (Fig. 10A), while chelicerae of Araneomorphae point toward each other or cross (diaxial) (Fig. 10B). Mygalomorphs include the large tarantulas of movie horror as well as many smaller, even minute spiders. They are secretive, more common and more widespread than usually recognized. However, one tiny, 3-mm species, the spruce-fir moss spider (*Microhexura montivaga*) is a U.S. federally endangered species. Its habitat on the tallest mountains in Great Smoky Mountains National Park and other nearby mountains is disappearing due to the loss of the conifers that grace those peaks.

Araneomorphae contains the vast majority of spider species. Male spiders have enlarged pedipalps (Fig. 10F) with an intricate sclerotized apparatus for transferring sperm to a female. The intricacies are of vital importance in identifying araneomorphs to genus. Nearly all spiders have eight eyes, the arrangement of which is often useful in identifying them to family. The eyes provide depth perception for these predators to better insure a successful strike. Salticidae (jumping spiders) (Figs. 10D and 11C–E) have an enormous pair of medial eyes that detect slight movements of potential prey and allow the spider to calculate the length of jump needed to catch its prey. Foelix (2011) likened this behavior to that of a cat stalking a mouse. Wolf spiders (Lycosidae) also have large medial eyes for active hunting (Fig. 10E). In contrast, ambush hunters like crab spiders (Thomisidae) have smaller eyes (Fig. 10F); they tend to wait for prey to come to them. Grass or funnel-web spiders (Agelenidae) have smaller eyes (Fig. 10G) and construct webs for snaring prey, rather than actively hunting. The tiny spiders in Linyphiidae (Fig. 10C) also build webs but many species have their eyes on an enlarged, elevated cephalothorax.

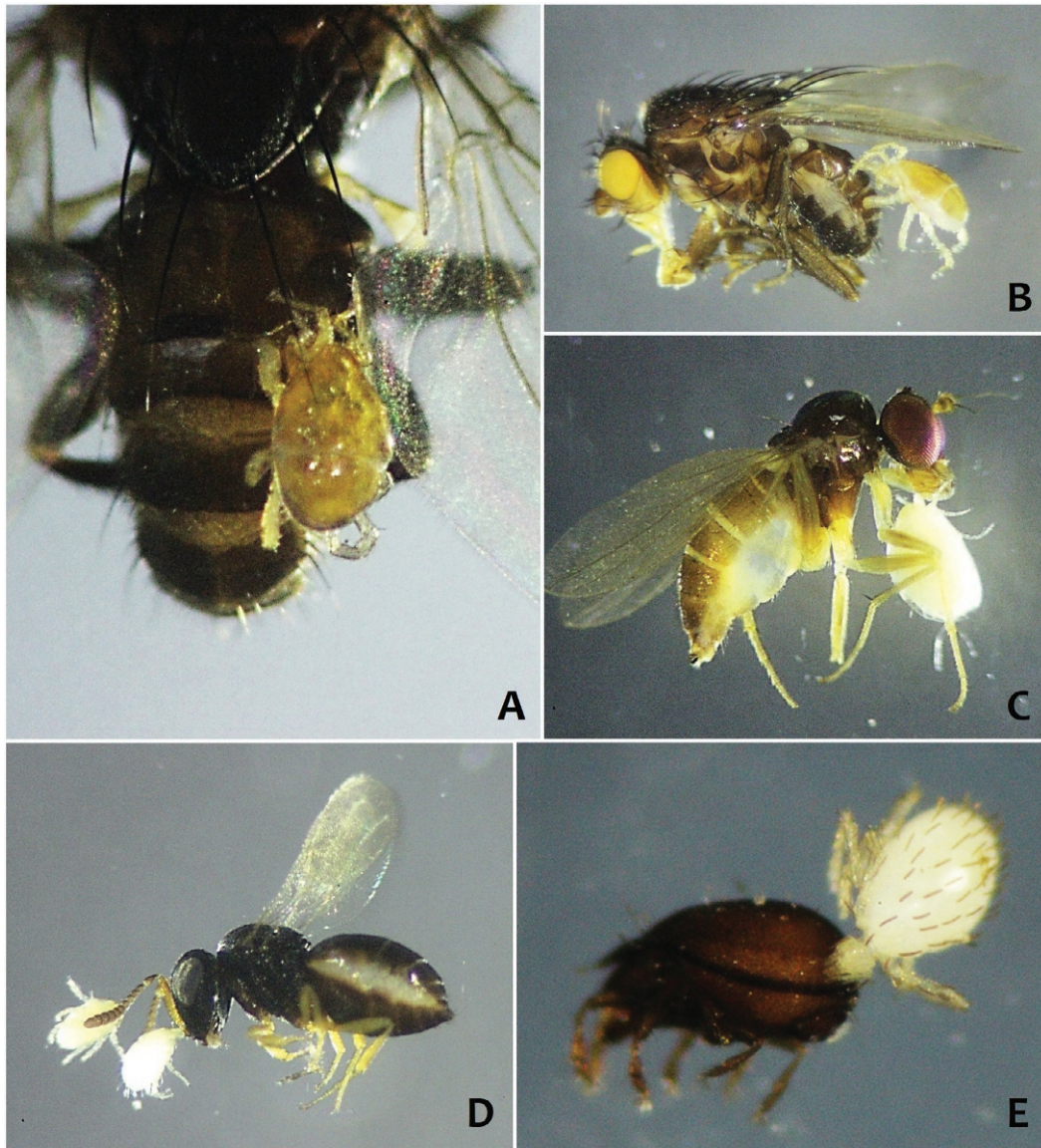


Fig. 9 Doubtful mite phoresy: six-legged larval stages on small arthropods. (A–C) On flies; mites in (A) and (B) likely were feeding on the insect while traveling for free. (D) On Scelionidae (Hymenoptera). (E) Larval stage feeding on an oribatid mite. Transportation on these hosts likely is an accidental result of feeding. Although the fly in (C) seems badly unbalanced with the mite attached to its mouthparts, it was nevertheless able to be caught in a flight intercept trap. The scelionid was collected in a pitfall trap.

Food

All spiders are predators and dine almost exclusively on insects, although some augment their diet with nectar and other non-animal foods. Myriapods and other spiders can also be on the menu, but noxious insects such as stinkbugs and those that fight back, like bees and wasps, are generally avoided. Spiders are generalist predators that attack prey small enough for them to overcome, self-selecting their prey by the qualities of the insect-catching webs they build.

Habitats

Soil and litter-dwelling spiders are closely tied to their environments, so that the spider fauna of a forest will differ markedly from that of a nearby meadow. Colonization of new habitats is difficult if walking is the only way to get there and phoresy is very rare in the Araneae. Instead, young spiders of many families disperse by ballooning. When atmospheric conditions are favorable, spiders will climb to a high point, raise their abdomens and release a long silk strand that can catch a slight breeze and lift the spider into the air. Needless to say, undirected dispersal has a high mortality associated with it, but theoretically only a male and female need to



Fig. 10 Predators with eyes upon you: spiders and a beetle. (A–G), Araneae (spiders). (A) Parallel cheliceral fangs of Mygalomorphae. (B) Opposed cheliceral fangs of Araneomorphae. (C) *Origanates rostrata* (Linyphiidae), length 2 mm: above, entire male, subdorsal view; below, cephalic lobe. (D) Anterior view of a jumping spider (Salticidae). (E) Anterior view of a wolf spider (Lycosidae). (F) Anterior view of a crab spider (Thomisidae). (G) Anterior view of a grass spider (Agelenidae). (H and I) Anterior view and dorsal view of *Stenus colon* (Staphylinidae), a specialist predator of Collembola.

make the flight successfully, land in the same place and find each other to establish the species in a new place. Thousands of linyphiid spiders may launch at the same time, enhancing the chances for successful colonization.

Interactions

Spiders generally are solitary creatures but do not defend large areas around them, although exceptions occur. Spiders tend to avoid other spiders unless it is mating time. Predators of spiders include most famously the spider wasps (Pompilidae) and some members of the thread-waisted wasp family Sphecidae, especially mud daubers. Spiders also are prey for small insectivorous mammals and birds. Numerous parasitoid insects develop in spiders or their egg cases. Mermithid nematodes also develop inside spiders (Fig. 17C).

Several families of spiders, especially Salticidae, contain species highly modified to resemble ants (Formicidae), bethylid wasps (Bethyidae) or velvet ants (Mutillidae) (Fig. 11). Salticids are most advanced in this respect. The species shown in the figure were all

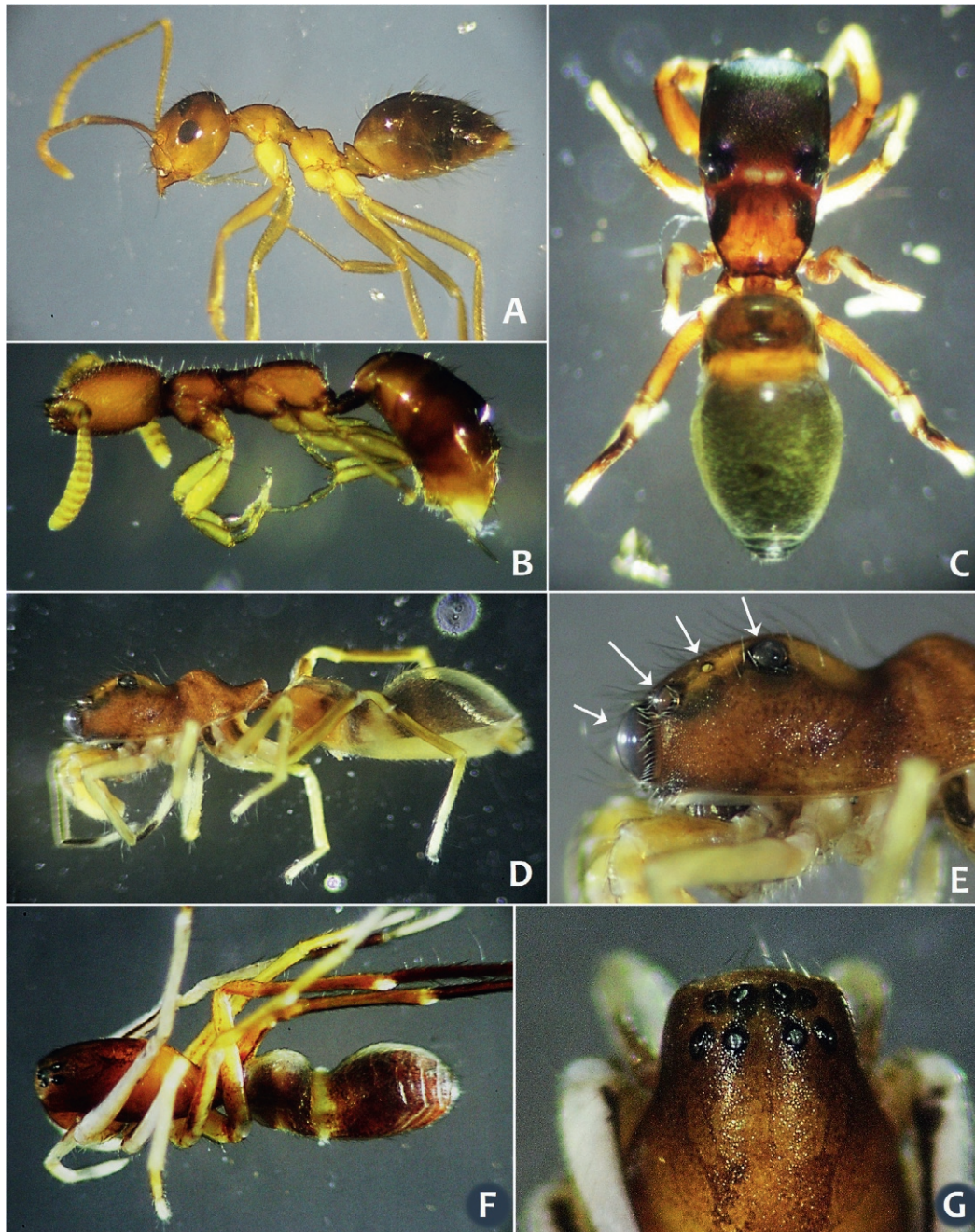


Fig. 11 Ant mimics: who's mimicking whom? (A) Ant (Formicidae). (B) Wingless bethylid wasp. (C) Jumping spider (Salticidae) slightly modified by constriction of abdomen. (D and E) Lateral view and left-side eyes (arrows) of highly modified salticid. (F and G) Lateral view and anterior cephalothorax of a similar spider (Corinnidae).

collected in the same general area in central Tennessee. Tropical salticids may be much more highly modified. A corinnid spider nearly identical to the salticids was collected in the same area but was easily distinguished by eye size and arrangement. It seems likely that these outwardly similar forms provide "group protection" and additionally avoid each other.

Opiliones

Long-legged arachnids known as harvestmen or daddy longlegs. These common arthropods are among the most unusual of all, with a little oval body suspended from eight impossibly long legs and the prosoma surmounted with a two-eyed observation knob, the ocularium (Fig. 12B and C). Harvestmen usually have cryptic patterns and colors that blend them in with their background

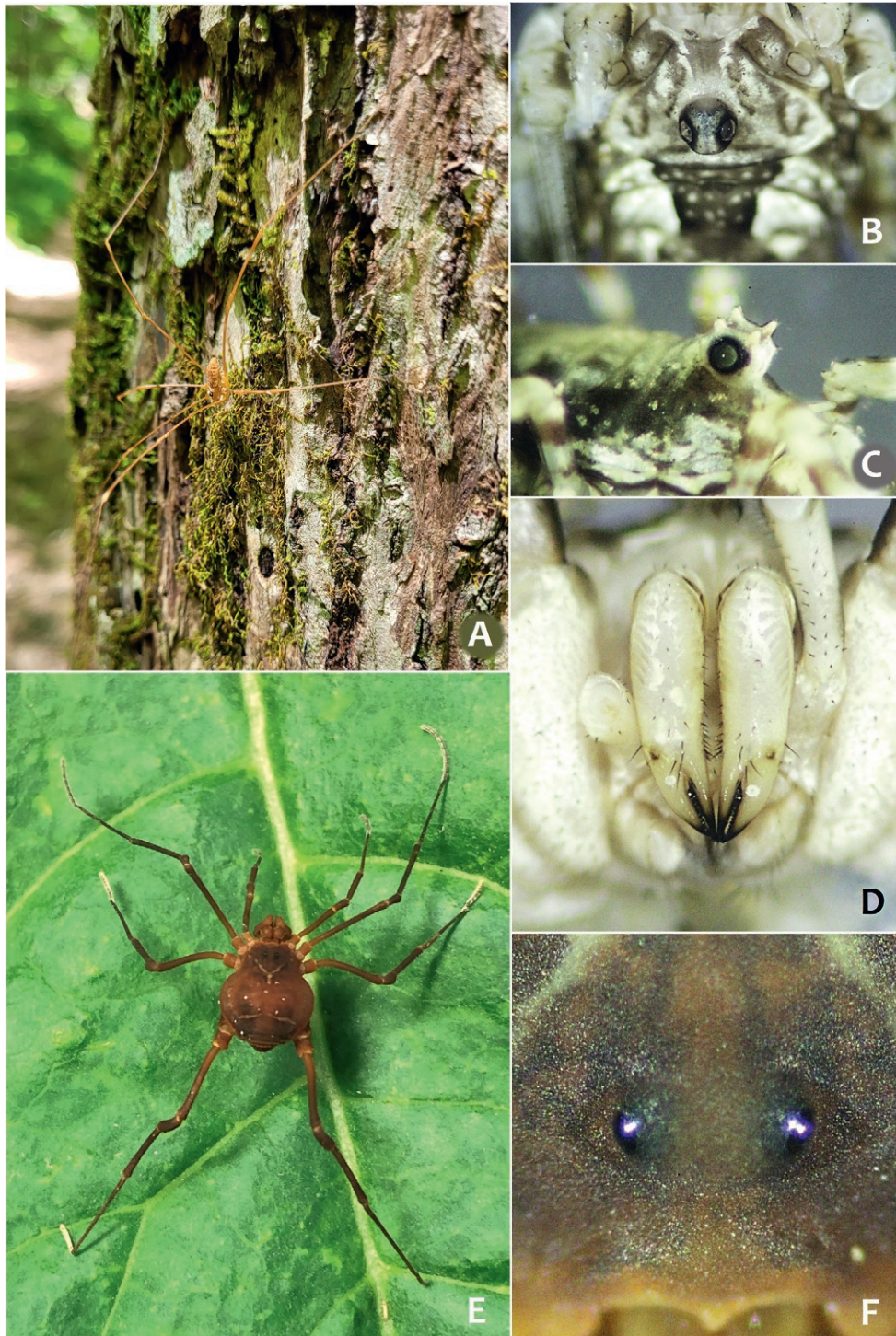


Fig. 12 Harvestmen. (A–D) Sclerosomatid harvestmen: habitus, eye tubercle of two species, chelicerae. (E and F) An armored harvestman, *Vonones sayi* (Cosmetidae): habitus (posed), paired eyes not on distinct tubercle. Fig. 5A courtesy of Adrian Smolis.

(Fig. 12A). Most people are used to seeing the very long-legged species but there is surprising variation in appearance, for instance the family Cosmetidae (Fig. 12E and F). Unlike the “typical” harvestmen, the ocularium is rudimentary and the eyes are distantly spaced. The following paragraphs are derived primarily from information in a splendid volume by Pinto-da-Rocha et al. (2007).

Harvestmen are typically included with the soil fauna, but many of the most common species are more likely to be on leaves or tree trunks. Pitfall traps, however, can collect many immatures and smaller species, especially in spring. Most harvestmen are generalist predators, using their pedipalps to hold the prey in place and the long, mobile chelicerae to tear it apart (Fig. 12D).



Fig. 13 The basal hexapods. (A) Protura. (B) Collembola, Isotomidae. (C) Collembola, Dicyrtomidae. (D) Diplura, Campodeidae. (E) Diplura, Japygidae. (F) Archaeognatha. (G) Zygentoma, Nicoletiidae. (E) by Mark Muegge, (F) by Matthew L. Bowser.

Harvestmen host an impressive array of parasites and parasitoids, as well as being potential food for members of every vertebrate class. They are, however, not defenseless. Harvestmen secrete noxious compounds from glands opening on the lateral margins near the coxae of the anterior legs. Consequently, harvestmen are distasteful and repellent to many would-be predators (although not all). The gland aperture is an ozopore, analogous to the defensive system of millipedes.

Pseudoscorpiones

Arthropods that resemble tiny scorpions that have lost their tails (Fig. 8E and F). They all are predators. Several species are anthropophilic but most of the nearly 3500 species are inhabitants of the soil, leaf litter, caves, bird nests and mammal burrows.

Pseudoscorpions have stout pincers-like chelicerae and prominent pedipalps with large pinching chelae reminiscent of scorpions. In most species the chelae have venom glands for subduing the prey. One or two small eyes are usually present on each side of the head. The life cycle consists of the egg, three juvenile stages and the adult, and is quite long for such little creatures, from 6 months to more than a year. Pseudoscorpions produce silk from glands on their chelicerae, which many species use it to spin a chamber for protection while they brood their eggs.

Pseudoscorpions are secretive and adept at hiding in small spaces. The pedipalps have long setae sensitive to the slightest air movement. If one moves an object close to a live pseudoscorpion it will energetically back up and raise its claws. They prey on small Collembola and other small arthropods. Phoresy is a common dispersal feature of pseudoscorpions. Unlike mite deutonymphs, it is the adult stage of pseudoscorpions that latch on to insects with their pedipalps (Fig. 8E and F). For them it is strictly a ride; they have never been reported feeding on their winged hosts, at least in flight.

Basal hexapods

Basal hexapods (Fig. 13) are three arthropod classes (Collembola, Diplura, Protura) and two insect orders (Archaeognatha, Zygentoma) whose ancestors never developed flight capabilities. These five taxa often are referred to as apterygotes. All are clearly recognized as separate but for convenience Archaeognatha and Zygentoma are sometimes discussed together. The term ‘apterygote’ refers to being wingless, which is true for the classes and orders indicated, but also can apply to any wingless insects such as fleas and lice, which are descended from winged forms. The five should be called basal hexapods; they all have six legs as in insects but lack the lateral pleural regions that are part of the later-evolving thoracic “box” for powered flight. Collembola, Diplura and Protura differ from true insects in having mouthparts enclosed by the head capsule, so that the mandibles and maxillae are hinged to internal tentoria. True insects have these parts hinged externally on the head capsule. Several molecular studies have confirmed this phylogenetic separation. Unlike the winged insects all basal hexapods practice external fertilization; males deposit sperm packets or droplets that nearby females pick up, sometimes after a mating dance.

Individuals of Collembola, Diplura, Archaeognatha and Zygentoma molt continuously throughout their lives, during which they can regenerate appendages lost to predation or accident. Protura, on the other hand, have five or six post-embryonic stages depending on subclass. Molting proturan adults have not been reported in the literature.

Collembola (springtails)

An exceedingly ancient class of hexapods dating at least to the lower Devonian. About 9000 species have been described worldwide, but as with most of the higher taxa mentioned in this chapter, that number probably is just a fraction of the true number. Many of the families are abundant and widespread (Fig. 14). The common name derives from an appendage on the venter of the fourth abdominal segment, the furcula, formed by the fusion of the former legs of that segment. However, many soil springtails have a reduced furcula or lack one entirely. Genial habitats may have 15–60 species that can reasonably be considered sympatric. Collembola living in the litter layers often exhibit vivid colors and patterns or are covered with scales that form light-reflective patterns.

Springtails lay eggs. Developing juveniles grow by molting through several stages but also continue to molt as adults. *Folsomia candida* (Isotomidae), the standard laboratory research springtail, may have up to 40 molts, with egg-laying commencing after the sixth molt. Females alternate egg-laying and non-egg-laying stages. Many species of Collembola are parthenogenetic, apparently facilitated by endosymbiotic *Wolbachia* bacteria.

Food

Springtails often are often thought to be mostly fungal feeders, but the evidence speaks to a far more varied diet for many species (Chahartaghi et al. 2005). Larger species skeletonize leaves and eat lichens that have fallen to the ground (Fig. 15A and B). Many species will readily consume a wide range of plant, fungal and animal tissues presented to them in the laboratory, although that is not always an accurate assessment in natural conditions; litter quality is important in determining the rate at which Collembola and other detritivores process organic matter (Ma et al. 2019). Collembola have two separate approaches to ingestion. Most families have chewing mouthparts, while a few families (e.g., Neanuridae, Odontellidae) have modified mouthparts that suggest a piercing-sucking mode of feeding. However, within sympatric species of a single genus these modified mouthparts may differ enough to suggest resource partitioning. For example, three sympatric species of *Morulina* have markedly different mouthparts (Fig. 16A–C) and the intestinal contents of each differs from the others (Bernard 2006). *Morulina crassa* (body length 1.4–2.2 mm) contained large pieces of hyphae, whereas *Morulina callowayia* (3.0–3.4 mm) ingested much finer hyphae, and the intestine of *M. delicata* (2.4–3.8 mm) was devoid of solid material. Likewise, species with chewing mouthparts co-exist in the same microhabitats by having food preferences. For instance, three ground-dwelling species in a fescue field had distinctly different arrays of food in their intestines (Fig. 16D–H). Members of the genus *Metisotoma* (Isotomidae) have mandibles and maxillae highly modified for prey capture and nestled in an atypically large head with enhanced muscles. These springtails wait in ambush for small prey (especially other springtails) to walk by; they then rush forth to bite their prey and consume it. Predation or carnivory can frequently be observed in culture. *Pogonognathellus* (Tomoceridae) are almost instantly attracted to an injured springtail and will eat it (Fig. 16F).

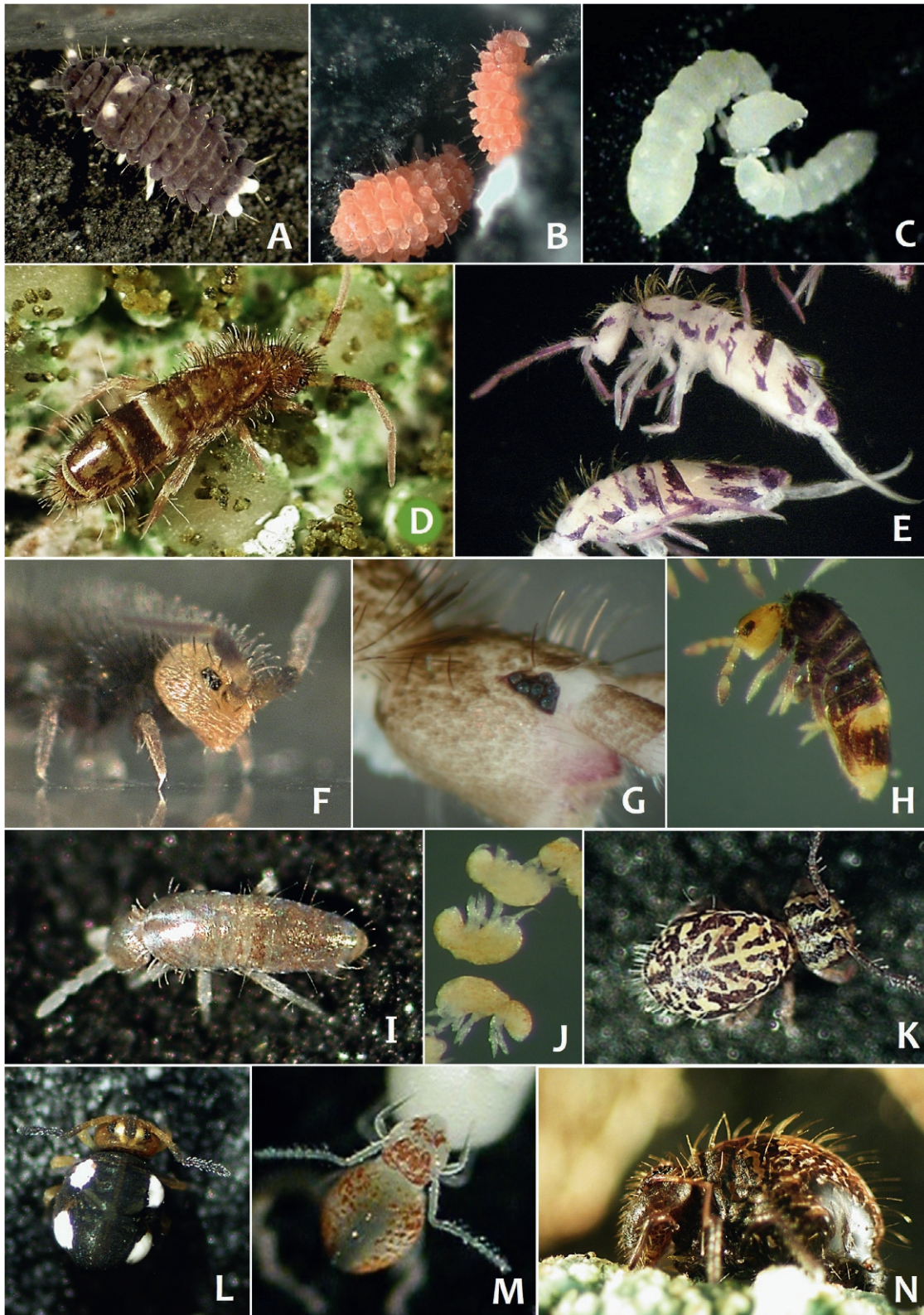


Fig. 14 Collembola diversity. (A) *Lobella palmeri* (Neanuridae). (B) *Vitronura* sp. (Neanuridae). (C) *Paronychiurus ramosus* (Onychiuridae). (D) *Orchesella cincta* (Orchesellidae). (E) *Entomobrya zona* (Entomobryidae). (F) *Orchesella carneiceps* (Orchesellidae); note that Collembola stand on the tips of their claws. (G) Head of *Pogonognathellus* sp. (Tomoceridae), note 6 eyes and body scales. (H) *Lepidocyrtus beaucatcheri* (Entomobryidae). (I) *Pseudosinella* sp. (Entomobryidae). (J) *Megalothorax maculosus* (Neelidae). (K) *Ptenothrix marmorata* (Dicyrtomidae). (L) *Sminthurinus quadrimaculatus* (Katiannidae). (M) *Pararrhopalites* sp. (Arrhopalitidae). (N) *Allacma fusca* (Sminthuridae).

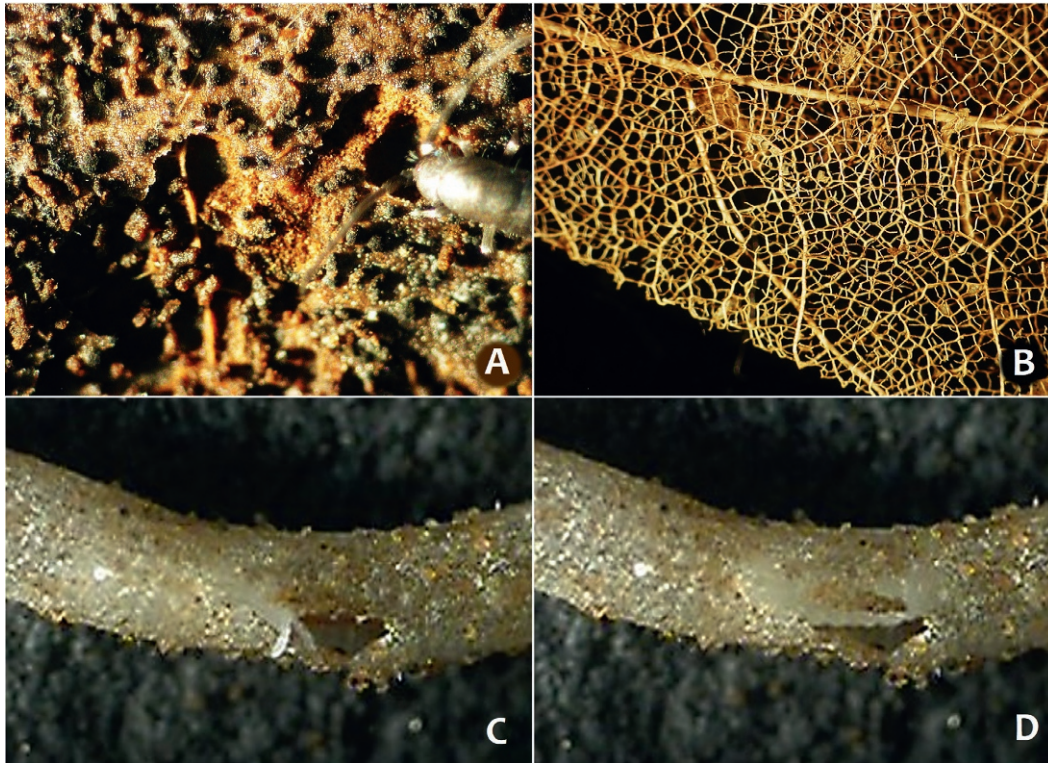


Fig. 15 Springtail detritivores and herbivores at work. (A) Reduction of dead leaf tissue by a large springtail, *Pogonognathellus* sp. (Tomoceridae). Note holes chewed in leaf and scattered blackish fecal pellets. (B) A dead leaf thoroughly skeletonized by detritivores, likely including small millipedes, isopods and Collembola. (C and D) Video captures of *Protaphorura fimata* (Onychiuridae) tunneling in roots of rooted maple cuttings. This species is one of the very few economically important pest springtails on living plants. (C) Antennae protruding from hole in excavated root. (D) Moving through tunneled root.

Habitats

Springtails are the most abundant and widespread of all hexapods, occurring wherever some suitable food is available. Several species are endemic to the Antarctic mainland, and a diverse fauna exists above the Arctic Circle to the northern rims of the continents. Habitat variety is endless; every vegetation site will produce a range of species adapted to it. Several families are partly or wholly arboreal, often being collected as stray specimens that have fallen to the ground. Other species in the Tomoceridae and Sminthuridae are not truly arboreal but climb tree trunks to feed and hide, and many species of Entomobryidae and Katiannidae climb herbaceous plants early in humid mornings to feed on pollen and fungal spores. *Hydroisotoma schaefferi*, associated with streams, hides between moss scales to escape predation by rove beetles (members of the family Staphylinidae). A few species are plant pests; *Sminthurus viridis* is an economic pest of alfalfa, scraping the epidermis of leaves and dehydrating them, while *Protaphorura fimata* tunnels through roots of tree seedlings, resulting in their decline and death (Fig. 15C and D).

Interactions

According to USDA entomologist Ian Stocks, Collembola can be thought of as “cows of the forest floor” since by their great abundance they are a major food source for the many small predators that roam the same environments, especially mites, spiders, ants and beetles. Ants (Formicidae) and crab spiders (Thomisidae) can often be seen scuttling around with springtails clutched in their chelicerae and may hang onto them even after falling into pitfall traps (Fig. 17D and E). *Stenus* species (Staphylinidae) (Fig. 10H and I) are specialist springtail predators. These beetles have enormous eyes for their size, allowing them excellent vision for spotting springtails and an extensible sticky labium for catching their prey, which is immediately hauled back to the mandibles and maxillae for consumption. Springtails can be colonized by *Entomophthora*-like fungi (Fig. 17A) and occasionally are infected with mermithid nematodes (Fig. 17B), although these are probably accidental infections of aquatic nematodes coming into contact with the springtails due to stream flooding.

Diplura

A class of basal hexapods consisting of about 700 species worldwide (Fig. 13D and E). Two of the three superfamilies (Campodeoidea, Japygoidea) are found worldwide in soil, litter and under rocks or moss mats on rocks in moist environments. These arthropods all lack eyes and are at least partially carnivorous on small, soft-bodied soil fauna, but campodeids also are detritivores and fungivores. Diplurans lay their eggs suspended on one or more stalks in a soil cavity or space. Females guard the eggs

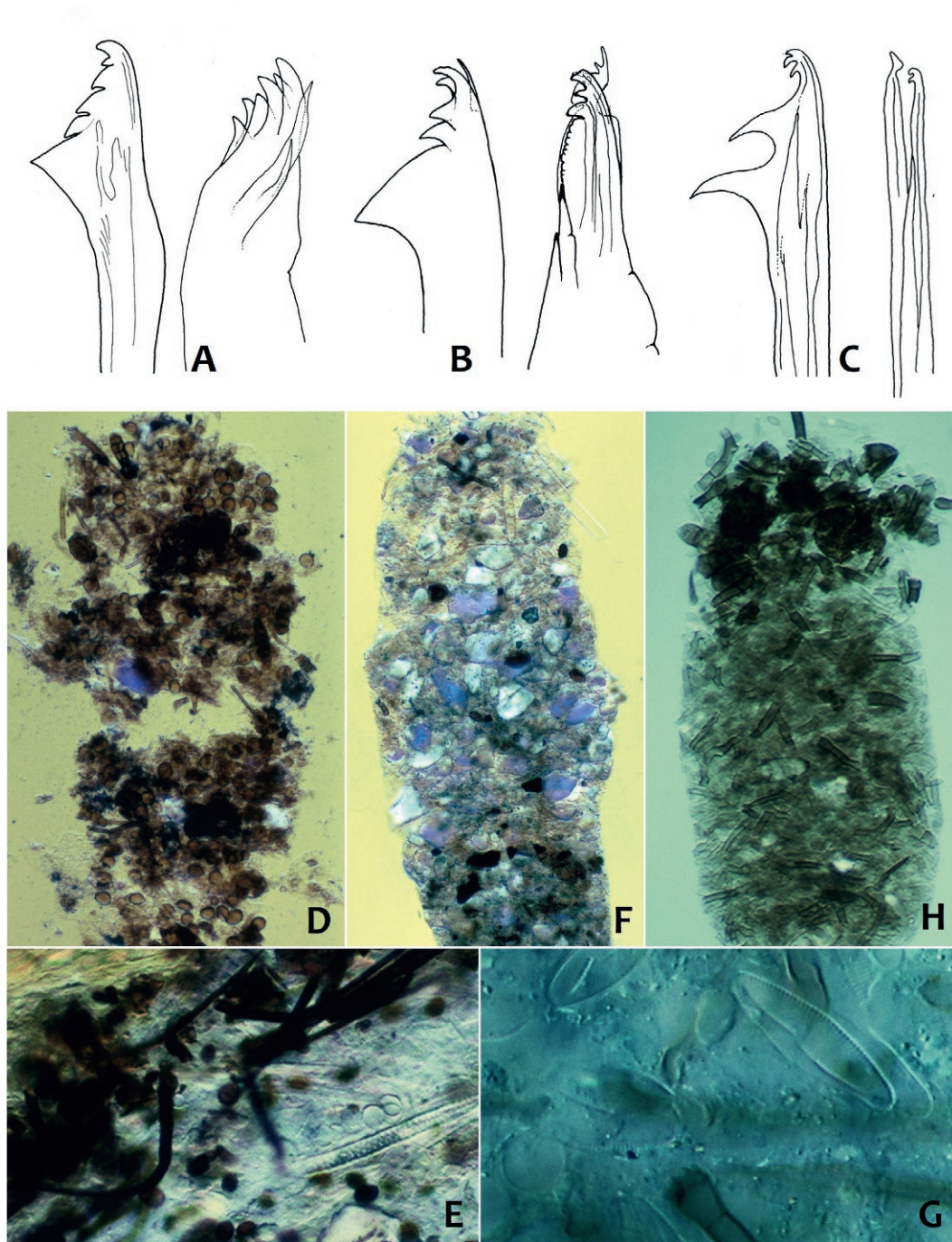


Fig. 16 Resource partitioning by Collembola. (A–C), mandible (left), maxilla (right) of three species of Appalachian *Morulina* (Neauridae) with different diets. (A) *M. crassa*. (B) *Morulina callowayia*. (C) *M. delicata*. (from Bernard 2006). (D–H), intestinal contents of three Collembola species collected from the same field. (D and E) *Isotoma viridis*, with fungal spores, plant tissue and amorphous material. (F and G) *Isotomurus bimus*, with fungal spores and spines, mineral particles and diatoms. (H) *Lepidocyrtus cinereus*, with dematiaceous hyphae and a few spores. Photos by Kelly Silas Parman.

and clean them of contaminants and potential pathogens until they hatch. Diplurans molt throughout life and so vulnerable appendages are regenerated over time.

Campodeoidea have long, slender cerci (Fig. 13E). They are very fast runners, moving quickly to find shelter when exposed. As they run, they wag their cerci, presumably as a distraction for potential predators. Japygoidea are instantly recognizable by the presence of pincer-like cerci (Fig. 13D), thus bearing a fleeting resemblance to young earwigs. The body is weakly sclerotized except for the final segment and the cerci. Members of Japygidae feed by capturing prey with their mandibles, then reaching over the body with the cerci to hold it while mandibles and maxillae chew it up. One family of very small Japygoidea, Parajapygidae, is a true soil dweller and feeds on plant tissue.

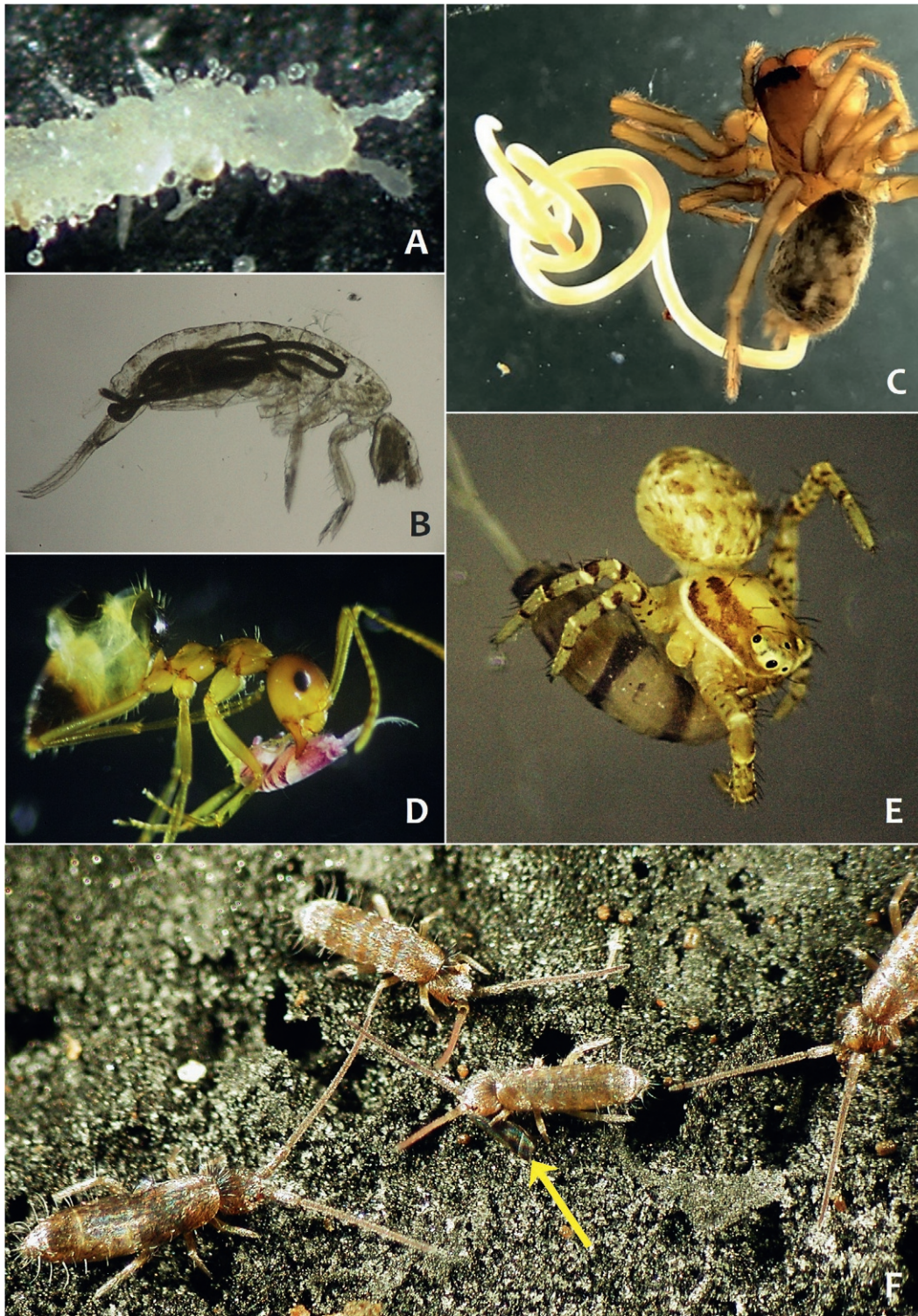


Fig. 17 Dangerous times. (A) Soil springtail (*Mesaphorura yosiii*, Tullbergiidae) colonized and killed by *Conidiobolus* sp. (Entomophthoromycetes). (B) *Pogonognathellus* sp. (Tomoceridae) infected with mermithid nematode. (C) Mermithid nematode exiting its spider host. (D and E) Entomobryid Collembola captured by an ant and a crab spider, respectively. (F) *Pogonognathellus* individuals attracted to and devouring an injured springtail (*Lepidocyrtus paradoxus*, arrow).

Protura

A class of six-legged arthropods separate and distinct from insects (Fig. 13A). The first proturans were discovered in 1907; since then, more than 800 species have been described from every continent except Antarctica. This number undoubtedly is far short of the actual number of species, as regional surveys typically find numbers of undescribed species. Adult proturans are less than 2 mm long and amber-colored. Protura lack eyes and antennae are absent, but the front legs are enlarged and held forward to take their place, with the middle and hind legs doing most of the walking. The mouthparts, which are held within the head capsule, except when feeding, are slender and more-or-less pointed.

All proturans have five postembryonic stages: prelarva (nonfeeding), larva I, larva II, matus junior, adult. Additionally, in Eosentomidae, developing males have a separate pre-imago stage before molting to functional males. The stages are easily separated from one another, thus large collections can be used to develop phenologies.

Tullgren funnels are best for collection of most Protura. For the small species that dwell in mineral soils high-gradient extractors are useful and sucrose flotation methods used for nematodes can lift them free from mineral components, though often with the loss of one or both forelegs. Protura rarely fall into pitfall traps and direct examination of the habitat for these secretive arthropods usually is an exercise in frustration.

Food

Proturans feed on fungi and may favor ectomycorrhizal fungi, but so little observation has been done that proturan food limits are not yet well defined.

Habitats

Protura are most often collected from upper soil layers, especially the O-layers (Table 2) of forest leaf litter and can be quite abundant, with nearly 40,000 individuals per surface square meter (Krauß & Funke 1995), and they are often abundant in secondary oak forest, suggesting a close correlation to actively developing ectomycorrhizal fungi. Proturans are also found in grassy areas, but as grasses are not symbiotic with ectomycorrhizal fungi, proturans must have a broader range of fungal food sources such as individual hyphae or mycelial mats.

Interactions

Proturans are likely to be food for small beetles, spiders, pseudoscorpions and predacious mites, but their importance as prey has not been documented. Proturans seem to have at least two active defense mechanisms. When disturbed from behind, a proturan may suddenly extend its abdomen to poke the offender. Proturans also exude a sticky substance from glands on the eighth abdominal tergite that may enmesh the mouthparts of predators (Hansen et al. 2010).

Interesting fact

Protura hatch with 9 abdominal segments and add new ones as they develop and molt, to a total of 12 (anamorphosis). However, they don't add one at each molt. The five instars have 9 (prelarva), 9, 10, 12, 12 (adult); there is no 11-segmented stage. In the subclass Eosentomata there is a "preimago" male, in which all characters are those of a mature male except the genital apparatus is not fully developed. If the female stage has a similar molt it is identical to the mature female.

Archaeognatha

The jumping bristletails were once included with the silverfish in the now obsolete order Thysanura. The order is also called Microcoryphia. Archaeognatha comprise about 500 species distributed worldwide. This insect order is exceedingly ancient; their fossils represent some of the first insects to appear on earth, and still bear many archaic traits. The extinct order Monura is now recognized as early Archaeognatha. The members of this order are the only monocondylic insects, referring to the mandibles being hinged to the head capsule by only one distinct condyle. All extant archaeognaths are robust insects with a cylindrical, tapering, arched body that can be forcibly snapped, propelling the insect upward some distance when it is disturbed. The ventral abdominal plates each bear a pair of coxites, remnants of the abdominal legs seen in the earliest fossils. All species have a body covered in frequently iridescent scales often forming a cryptic pattern, enormous compound eyes, large ocelli below the compound eyes, a pair of long cerci and an even longer terminal filament. This terminal filament plays an important role in stabilizing an airborne individual; damage to the filament adversely affects the jump (Yanoviak et al. 2009). Sturm & Machida (2001) describe collection and rearing of these ancient insects, as well as presenting keys to the world genera.

Food and habitat

Most archaeognaths tend to occur on rocks, piles of stone or rock crevices where their scale pattern blends with the rock surface (Fig. 13F), but in forest they can sometimes be found on tree trunks. Food consists of algae, lichens, detritus scraped from leaves or wood surfaces and sometimes leaf surfaces themselves. Investigation of the ability of these insects to use a wide range of plant foods would help complete gaps in knowledge of this resilient insect order.

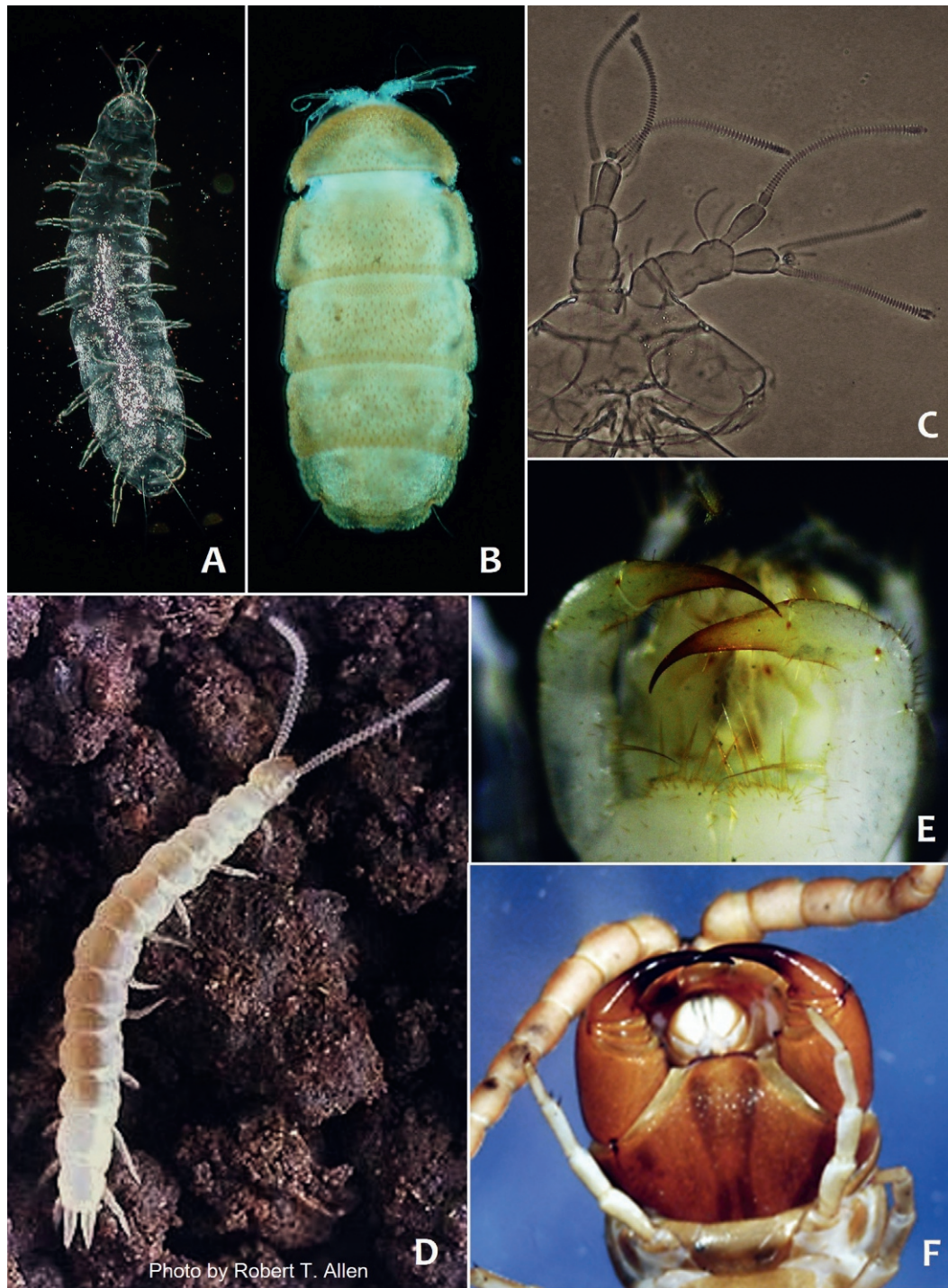


Fig. 18 Pauropoda, Symphyla, Chilopoda. (A–C), Pauropoda. (A) *Allopauropus carolinensis* (Pauropodidae). (B) *Eurypauropus spinosus* (Eurypauropodidae). (C) Pauropod antennae. (D) Symphyla. (E and F) Chilopoda. (E) Forcipules of *Scutigera coleoptrata*. (F) Forcipules of a lithobiomorph centipede. Symphyla image by Robert T. Allen.

Interactions

Gregarine protists (see Diplopoda section) are present in the midgut of many archaeognaths but their effects, if any, are unknown. Mermithid nematodes have been reported but could well be accidental infections. The few reports of predators are mostly for spiders and the occasional rove beetle. Interactions of these relatively large charismatic insects with parasites and predators could be studied to good effect without expensive equipment.

Zygentoma

The silverfish and related taxa consist of about 600 species. The best-known, of course, are the anthropophilic silverfish in the family Lepismatidae, but five other families exist including Nicoletiidae and Tricholepidiidae. This latter is known from one species in a small area of California. Zygentoma are rather flattened, quick-moving insects with three “tails” (two cerci, terminal filament) of about the same length. Lepismatidae are scaled and have a small cluster of ocelli on each side of the head, while in Nicoletiidae eyes are lacking and scales are present or absent. Zygentoma mouthparts are dicondylic, with two hinge points for each mandible.

The lepismatidae genera *Lepisma*, *Ctenolepisma* and *Thermobia* contain most of the species considered to be domestic pests (Wygodzinsky 1972), but they thrive outside if the climate is suitable, especially *Ctenolepisma*. They can easily slip back inside due to their flattened bodies and nocturnal movements. The damage they cause by eating starchy material, book sizing, etc., can be significant but is not often seen today, perhaps due to a widespread adoption of air conditioning. Rust & Millard (2009) provide management guidelines. My home has a light infestation of the four-lined silverfish, *Ctenolepisma lineatum*, and several buildings on the University of Tennessee campus have the gray silverfish, *C. longicaudatum*, but I like the philosophy expressed by the entomologist F.E. Lutz (1948, p. 40): “If such a creature is eating your wallpaper...or other belongings, your sorrow may be mitigated by your interest in seeing the most primitive insect you are likely to observe without special effort.”

Although much is known about these domestic companions, much less attention has been paid to the other members of the family. Pitfall traps in temperate areas remote from human habitation sometimes contain lepismatidae, but identification is usually not possible because those biotopes have never been surveyed. When collections are made in a region and modern molecular techniques are deployed as well as morphological description for species analysis, numerous new taxa may be identified (see Smith 2017).

Nicoletiidae is a widespread subtropical-tropical family with a few representatives reaching to the mid-temperate latitudes. These secretive members of Zygentoma do not inhabit homes although one or two species are widespread in greenhouses. Nicoletiidae has two subfamilies: Nicoletiinae are slender, have long tail filaments and tend to live in soil, litter and caves, while Atelurinae (Fig. 13G) are wedge-shaped with short tail filaments and are often associated with ants. Both subfamilies contain many undiscovered species.

Food, habitats, interactions

Zygentoma are generalist feeders, eating a wide range of materials. The diets of Nicoletiinae have not been elucidated, but corn root traps in caves have successfully attracted them. Atelurinae are largely associated with ants but not in a tight symbiotic relationship or commensal relationship. Tolerance of the ant host for the presence of inquilines and social parasites is dependent on the ability of the “guest” to exhibit behavioral and chemical mimics that mark it as a fellow ant (von Beeren et al. 2021). Atelurines were more subject to attack by their hosts than were some other guests due to their quick movements that signaled a colony intruder.

Crustacea

Crustacea are overwhelmingly aquatic organisms, with enormous diversity in both marine and freshwater habitats. However, several classes and orders have managed to find a semblance of life ashore. Most of these groups are dependent upon a very moist environment and periodic retreat to water, among them amphipods, crabs and crayfish. Others such as fairy and tadpole shrimps (Anostraca) complete their life cycles in ephemeral pools, then lay eggs resistant to desiccation until the return of free water. Small copepods are occasionally found in soil extracts. Ostracods occasionally show up in pitfall traps, an indication that they have been disturbed by flooding from a nearby stream; ostracods are strictly aquatic. The only fully terrestrial crustaceans are Isopoda of the superfamily Oniscoidea (woodlice, pill bugs, sow bugs, roly-polies) (Fig. 19A). Oniscoidean isopods, of which there are about 5000 species, have a cosmopolitan distribution.

Most crustaceans respire through gills, but terrestrial isopods have a highly modified lung-like system under the abdominal pleopods (pleopodal lungs, Unwin 1931; <https://www.bmig.org.uk/image/woodlice-pleopodal-lungs>); this modification allows them to live in a wide range of climates, although sufficient moisture is a necessity. Oniscoidea carry their eggs in a brood pouch between the leg bases. Hatched juveniles leave the brood pouch soon after hatch. Juveniles resemble the adults and go through four or five molts to reach maturity.

Food

Woodlice are omnivores, although some species may have decided preferences for certain kinds of plant litter. In laboratory culture they readily eat fresh vegetables and thrive alongside the millipedes therein. Woodlice also are coprophages.

Habitats and interactions

Terrestrial isopods are active and abundant in most mesic biotopes in leaf litter, loose soil, under rocks or beneath loose bark or pieces of wood. Regardless of habitat, terrestrial isopods are dependent upon high humidity and some free water to maintain internal water balance and operate their pleopodal lungs (Warburg 1987). Woodlice are vulnerable to beetles, spiders and other predators capable of penetrating their strong cuticle.



Fig. 19 Isopoda (pillbugs) and Diplopoda (millipedes). (A) A common isopod, *Armadillidium vulgare*. (B) *Narceus americanus* (Spirobolida); lower left, a tiny *Oxidus gracilis* plus assorted snails. (C) *O. gracilis* (Polydesmida). (D) Two mating individuals in tribe Apheloriini (Polydesmida). (E) *Brachycybe petasata* (Platydesmida). (F) Polyxenidae (Polyxenida). (G) *Abacion* sp. (Callipodida) chewing into *Russula* sp. (H) *Floridobolus* sp. (Spirobolida) eating bell pepper. (A–D, H) imaged in culture; (G) courtesy of Adrian Smolis.

Myriapoda

“*Pauropus huxleyi* is a bustling, active, neat, and cleanly little creature. It has, too, a look of cheerful intelligence, which forms a great contrast to the dull stupidity of the Diplopods, or the melancholy ferocity of most Chilopods.” — Sir John Lubbock (1868).

"One who once has seen a living pauropod, like a miniature mouse scurrying around on the dark underside of a stone with quick elegance, will certainly concur with him." — Ulf Scheller (2011).

(Referring to the house centipede, which eats insects and spiders): "Dr. Felt says 'its presence in a house should be welcomed, since it is capable of inflicting no injury aside from a somewhat poisonous bite, the latter being extremely rare.' I confess that any found in our house get stepped on." — Frank E. Lutz (1948).

Myriapoda is a clade of four arthropod classes (Chilopoda, Diplopoda, Pauropoda, Symphyla), with the distinction of containing the world's largest known land arthropod, an extinct millipede relative that may have been more than 2.5 m long with a mass of 50 kg (Davies et al. 2021). The extant members are all much smaller, with the African millipede *Archispirostreptus gigas* reaching 30 cm long and weighing 100 g (Fig. 20A), and several scolopendrid centipedes nearly as long. In contrast, most pauropods are less than 2 mm long. Myriapods are mandibulate arthropods with two main body divisions (head, trunk), one pair of antennae and many legs. They are all terrestrial and most occur in soil and litter.

Chilopoda (centipedes)

Myriapod class of predators with one pair of legs per segment. Anthropomorphism is an outdated approach to understanding the nature of invertebrates, but most people recoil upon the sight of a centipede...fast-moving, fangs, too many legs and overwrought stories of their venomous bite. The 3000 species of centipedes are easily recognized by their slender, amber-colored bodies and rather fearsome "fangs" (forcipules) that are really the highly modified first pair of legs modified for injecting toxin into prey (Fig. 18E and F).

Adults of the four common centipede orders are readily identified. Scutigermorpha (of which the house centipede is best known) have 15 pairs of legs and large compound eyes. The other orders have small eyes or are blind. Lithobiomorpha have 15 body segments bearing a pair of legs; Geophilomorpha are slender and have 31 to more than 180 segments with legs; and Scolopendromorpha have 21 or 23 leg-bearing segments. A fifth order, Craterostigmomorpha, occurs in Tasmania and New Zealand. Interestingly, the venom composition among the centipede orders varies substantially among the orders (Jenner et al. 2019).

Females of geophilomorphs and scolopendromorphs brood their eggs and hatchlings, which have the same complement of legs as the adults. Lithobiomorphs and scutigermorphs lay eggs singly, and hatchlings have only a few pairs of legs. The number of



Fig. 20 (A) The African Giant Millipede, *Archispirostreptus gigas*; largest individual 28 cm (Order Spirostreptida). (B) Gregarine protists (Apicomplexa) attached to inner wall of everted intestine of the American Giant Millipede, *Narceus americanus*.

post-embryonic molts is variable among the orders, generally 7–11 (Lewis 1981). Chilopoda are typically long-lived with only one generation per year. Centipede densities are smaller than those of their prey, with an upper limit of about 1000/m² for geophilomorphs, but other orders typically have much smaller densities. However, the bulk of this work has been done in northwestern Europe, with few tabulations from other regions.

Food

All centipedes are predators inhabiting soil or litter. Specific enzymes are produced in centipede venom to digest prey tissues. Choice of prey is not very selective; various observations have shown that any invertebrate encountered could be a target. In laboratory studies, some potential prey items may be rejected, but whether rejection is uniform among an isolate, population or species is an open question. The larger scolopendrids can be predators of small vertebrates, including bats. Herbivory has been reported but is probably due to incidental organic matter ingested along with prey or consumed in stressful conditions.

Habitats

Scutigermorph centipedes are long-legged, swift surface runners native to tropical and subtropical regions. The house centipede, *Scutigera coleoptrata*, is a nearly cosmopolitan inhabitant of houses, but can also occur on the soil surface, in crevices and among log piles in the vicinity of human habitation. Other species may opportunistically inhabit rodent burrows (Voigtländer, 1981). Most scolopendromorph and lithobiomorph centipedes inhabit leaf litter, soil cracks and spaces and under the bark of cut and fallen trees and logs. A good place to locate them is on soil under rocks, especially flat ones with worm and insect tunnels that provide space for them to dwell. Some centipedes are capable of burrowing, especially as a defense against desiccation. Most centipedes are nocturnal organisms, although scutigermorphs may also be active in lighter conditions.

Interactions

Snakes, lizards, birds and small to medium sized mammals, from shrews to foxes, have all been reported to prey on centipedes, although there are few or no records of predation on scutigermorphs. Carabid beetles may be significant predators of smaller individuals. Centipede digestive tracts contain a diverse array of gregarine, coccidian and other protists that probably are harmless commensals. Nematode infection is very rare; oxyuridomorph nematodes, which are common in millipedes and soil insects, are unknown in Chilopoda except perhaps for one report of a *Cephalobellus* sp. However, centipedes apparently can serve as an intermediate host for the human parasite *Angiostrongylus cantonensis* (rat lungworm), a causal agent of eosinophilic meningitis (<https://www.cdc.gov/parasites/angiostrongylus/index.html>). Wang et al. (2018) described cases in China of persons who acquired this parasite from eating raw centipedes.

Diplopoda

The millipedes, a myriapod class of more than 7000 described species (Figs. 19B–H and 20A). Millipedes are instantly recognizable by the two pairs of legs on each body segment, formed in the distant past by fusion of two body segments each with a single pair of legs. This fused segment is better termed a diplosegment or body ring. Fusion must have occurred nearly a half-billion years ago, as all known millipede-like fossils have this feature. In living species, this fusion strengthens body architecture and allows a stronger push when millipedes are burrowing through soil or leaf litter. The body architecture further allows most millipedes to roll up in a tight spiral protecting the more vulnerable underside. In cross-section millipedes are largely cylindrical, hemispherical or flat, but Glomeridae and Sphaeriodesmidae resemble large isopods and similarly can roll up into tight spheres. Millipedes typically are solitary animals but platydesmids (Fig. 19E) usually are found in small groups on covered damp wood or bark.

Another signal feature of millipedes is the presence in most species of a chemical defense system, consisting of a pair of glands in each segment with an outlet (ozopore) through the body wall. Hydrogen cyanide (HCN) is common in the order Polydesmida but many different combinations of noxious chemicals exist among most of the other orders. The common polydesmid family Xystodesmidae (Fig. 19D) seems particularly pungent. Paradoxically, flies adapted to feeding or laying eggs on dead arthropods (Phoridae, Sarcophagidae) are drawn within a minute to a crushed millipede, presumably due to the release of the pungent defense chemicals. In our laboratory experience, millipedes in culture for months or gently handled repeatedly become acclimated to these situations and cease to release defensive chemicals.

Millipedes range in size from tiny soft-bodied polyxenids (Fig. 19F) to *Archispirostreptus gigas* (Fig. 20A), an African species that can grow to a length of more than 30 cm. The record for most legs is 1306 for the recently described Australian female of *Eumillipes persephone* (Marek et al. 2021), collected from a 60-m-deep exploratory drill hole. As discussed by Marek et al. (2021) and others, subterranean systems have numerous, cracks, fissures, faults and chambers where an almost unknown fauna may be hiding. Sinkholes likely include an underground chamber or cave that usually gets filled in with earth or concrete, destroying any ecosystem that exists in it.

Millipede development is anamorphic, with new segments added at each molt, at least until the adult stage is reached. The first-stage juvenile usually has five body segments and three pairs of legs; at first glance one might think they are springtails or strange beetle larvae. Some taxa continue to molt and add segments even after reaching adulthood, while others reach their segmental limit but need to molt a few more times to become sexually mature. Millipedes possess anteroventral gonopores, usually on the seventh body ring. At maturity males have a pair of highly modified, sometimes fantastically shaped legs, the gonopods, on the ventral side of the seventh body ring. These gonopods manipulate a sperm globule for insertion into the female's gonopore.

Mating is preceded by the male accompanying the female as she walks around or even riding on top of her. Eventually the two entwine venter to venter and the fertilization process occurs (Fig. 20D).

Food

Millipedes are major components of the detrital food web, consuming dead leaves, wood, fecal material from herbivores and their own droppings. Bacterial fermentation within the intestine is an important source of supplemental nutrition. Some species will eat live basidiomycete fungi (Fig. 20G). Fungal interactions with millipedes may be extraordinarily complex. In one extensive study *Brachycybe lecontii* (Fig. 20 is *Brachycybe petasata*) was associated with 176 fungal genera in 39 orders (Macias et al. 2019). The large species *Narceus americanus* (Fig. 20B) is often found on tree trunks two or more meters from the surface eating organic matter on bark. In our laboratory cultures spirobolida and spirostreptida are ready consumers of many fresh vegetables and denser fruits such as apple, with particular attraction to sweet corn and bell pepper (Fig. 19B, C, H and 20A), but in nature are found mostly in forest ecosystems, not in gardens. Millipedes can also be opportunistic carrion feeders on other invertebrates and even on dead millipedes. The reduction of large pieces of dead plant material to smaller pieces was long thought to enhance fungal and bacterial colonization, but Rawlins et al. (2007) demonstrated that digestion removes biochemical components useful to soil microorganisms, leaving behind biologically recalcitrant materials. However, weathering of fecal pellets probably disperses individual fragments so that even recalcitrant materials such as lignin are more exposed to their microbial specialists. Damage to living plants is rare but the introduced species, *Ommatoiulus moreleti*, has damaged crops in Australia (Douglas et al. 2019).

Habitats

Diplopods are most abundant in moist deciduous forests, where several species can be found moving about openly during the day, but more often at night. Millipedes also occur in meadow and grassland but are not as well known; hand-collecting and use of pitfall traps in these habitats will yield species that do not live in forest. Millipedes associated with the forest floor often move under bark on standing dead trees.

Interactions

The chemical defenses of millipedes repel many potential predators, but others are undeterred. David (2015) has summarized these successful predators, both mammalian and arthropodan. The large European staphylinid beetle *Ocyopus olens* (up to 32 mm length) is a generalist predator that feeds readily on *Ommatoiulus moreleti*, a medium-length millipede (up to 45 mm). The lack of suitable predators in Australia likely facilitated the dispersal and abundance of *O. moreleti*.

Internal parasites and commensals of millipedes exist in several kingdoms and phyla, including Protista (ciliates, gregarines), “trichomycete” Zygomycota fungi, Acanthocephala, Nematomorpha, Nematoda, and Arthropoda. Detailed discussion of the numbers and complexities of these groups is beyond the scope of this Chapter. However, as most of these taxa occur in the gut of the millipede, one should think of a “universe” within a millipede intestine. Most millipedes more than 2 mm wide are likely to contain nematodes, trichomycetes and ciliate protists in the long hindgut. Gregarines, a group of apicomplexan protists, will be easily noticed in the trophozoite stage protruding from the inner lining of the intestine (Fig. 20B). Nematodes are the most prominent members of millipede intestinal fauna. Hundreds of species in the nematode orders Rhigonematomorpha and Oxyuridomorpha have been described. Both of these orders are considered bacterivores and do not feed on millipede tissues. Adults of most of these nematodes exceed lengths of 2 mm and several can reach 10 mm. Although a large millipede may harbor hundreds of nematodes it remains active and appears healthy, suggesting no harm. The ease of millipede dissection and identification of intestine-inhabiting nematodes provides an excellent means for students to explore natural relationships of hosts and commensals directly from the field (Phillips et al. 2019). An unrelated bacteria-feeding nematode, *Rhabditis necromena*, is useful in millipede biocontrol due to carrying a pathogenic bacterium in its intestine that is released into the millipede lumen after the nematode has entered the host (Schulte 1989). It must be noted that many other species of Rhabditidae may invade millipedes when they are sick or recently dead, colonizing the hemocoel, collapsing tissues and reproducing rapidly. One must distinguish these opportunistic scavengers from nematodes transporting pathogenic bacteria.

Paupropoda

Tiny, inconspicuous arthropods with a head, 12 trunk segments and 8–11 pairs of legs (Fig. 18A–C). Development is anamorphic, with five active stages following egg hatch; total time from egg to adult is thought to be 3–4 months. The antennae are unique among the myriapods; they have two distinct branches (biramous), one of which has a basket-like sensory process, the globulus. Most species are whitish-translucent, but the family Eurypaupodidae (Fig. 18B) has enlarged, sclerotized tergites. Pauropods move in short runs, stopping frequently then continuing. Their distinctive darting movements make them easy to spot, even though most species are only about 1 mm long. About 830 species have been described but many await discovery; every in-depth survey reveals numerous undescribed taxa.

Starling (1944) studied the Pauropoda of Duke Forest (North Carolina) in field observations and laboratory cultures of *Paupopus carolinensis*. In both pine and oak forest soils pauropod densities were largest at soil depths of 2.5–5.0 cm, soil temperatures of 18 °C or above, soil moisture averaging 20% and high relative humidity. Pauropod densities may reach several thousand per m² in favorable environments.

Food

Pauropods are generalist herbivores or fungivores (Potapov et al. 2022). One species has been reported damaging root hairs of greenhouse-rooted African Violet (Scheller et al. 2004).

Habitats

Pauropods are creatures of moist soil, small enough to live in soil spaces. They also seem to favor the underside of rocks; they can be found by overturning a stone and looking for tiny white creatures darting about on the surface, from which they can be collected with a moistened paintbrush.

Interactions

There seems to be little or nothing known of predation on pauropods. Two families, including Eurypauropodidae (Fig. 18B), are “armored”, with thickened, more sclerotized terga, suggesting defense against predators. The tiny size of pauropods allows them to move through connected soil pores or between aggregates to avoid larger, less mobile predators.

Symphyla

Slender white myriapods with 11 or 12 pairs of legs, a pair of filiform antennae and lacking eyes (Fig. 20D). About 200 species of symphylans have been described. Most species are less than 5 mm long, but several may reach 15 mm. The posterior end of the body has a pair of cerci, each of which has a spinneret that produces silk. Symphylan development is anamorphic, with newly hatched juveniles having six pairs of legs and short antennae with only a few segments. They reach adulthood after five juvenile stages but continue to molt after maturity. In culture, adults may molt more than 50 times. Symphylans are surprisingly long-lived, up to 7 years in one study; this life span is comparable to that of the largest millipedes.

Although a common name for this class is “garden centipede” they are harmless to people. Note, however, that the centipede, *Lithobius forficatus*, also has the common name of garden centipede (e.g., see <https://homeguides.sfgate.com/garden-centipedes-76177.html>), but this arthropod is much longer and brown. Possible confusion of the two organisms is a good reason to just call Symphyla symphylans. Symphylan mouthparts resemble those of millipedes and even the largest species cannot bite. Symphylans are classified into two easily separated families. In Scolopendrellidae the posterior margins of the tergites have pointed posterior projections, whereas the Scutigereidae have rounded posterior margins.

Food

Symphylans are considered omnivores, capable of feeding on the softer parts of plants and animals. *Scutigereella immaculata*, a worldwide pest of live plants, was studied thoroughly by Michelbacher (1938), who fed his cultures with lettuce. Seedlings suffer the most damage due to root feedings and germinating plants may die before emergence above the soil surface. Plants that are several weeks old generally suffer little damage because the root systems are more robust, and roots and root cell walls are thicker.

Habitats

Symphylans live in leaf litter and soil. They can be found in most biotopes. Their slender, very flexible bodies move easily through soil channels formed by roots or soil structure and aggregation, enabling them to migrate deeply into soil.

Interactions

Because symphylans typically reside in soil, they are unlikely to encounter the usual suspects (spiders, harvestmen, small beetles, centipedes) unless they come to the surface to feed. Soil-dwelling beetle larvae (Carabidae, Staphylinidae, etc.) may be more likely enemies. Symphylans produce silk from glands on the cerci. Other arthropods that make silk tend to have multiple uses for it; for symphylans, these include anchorage during the molting process or as a defense against predators. Quite possibly, silk smeared on the mouthparts of a predator will deter the attacker.

Camouflage, mimicry and defense

Staying alive and well long enough to reproduce is a constant challenge for all organisms. Arthropods have evolved a wide array of adaptations to lessen the chances of being eaten, including camouflage, mimicry, chemical defenses and impersonation. Only a few examples can be given here but readers are invited to find others.

Some arthropods disguise themselves as debris, which not only blurs the arthropod’s dimensions but also reduces the chance of a successful attack through the debris. The “trash” oribatid (Fig. 21A) has already been mentioned. Larvae of Chrysopidae (green lacewings) are similarly modified with cuticular lobes and long setae to hold debris on their backs as a disguise. Camouflage also can be an intrinsic part of the arthropod’s body. Some springtails (Fig. 21B) have scale patterns that break up their outlines and resemble the substrate on which they are resting. Species of Dicyrtomidae (Fig. 14K) mimic linyphiid spiders and even have a similar walk.

Chemical defense is an important deterrent mechanism for several families of poduromorph Collembola, especially Hypogasturidae, Neanuridae, Onychiuridae and Tullbergiidae. When a threat is perceived or they are physically attacked, these springtails

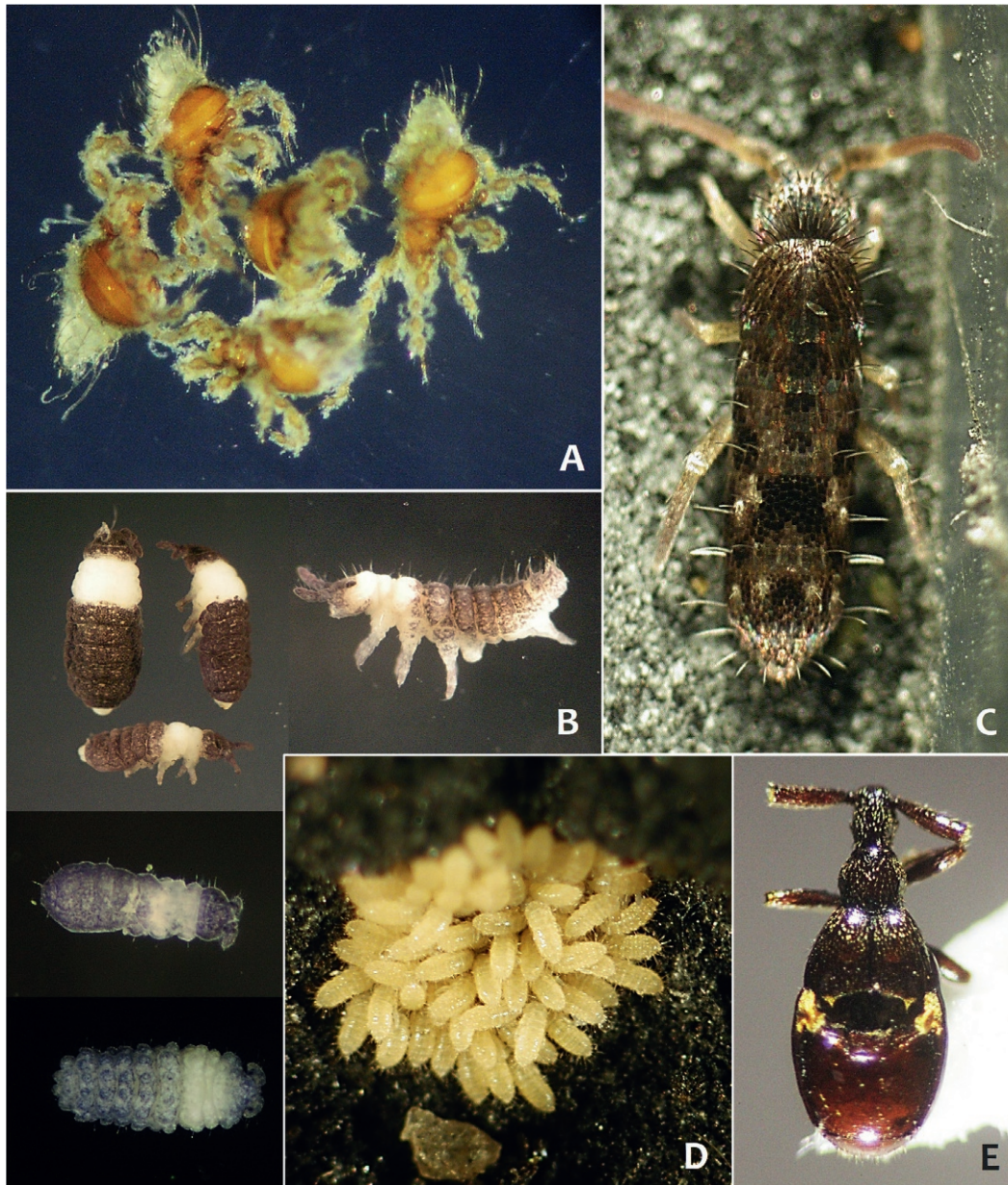


Fig. 21 Camouflage, mimicry, defense and “I’m with you”. (A) Oribatid “trash” mites with old exuviae and debris clinging to long body setae and feathery leg setae. (B) *Pseudachorutes aureofasciatus* (Neanuridae) (left), putative neanurid mimics *Anurida* sp. (below) and *Vitronura* sp. (bottom), and *Ceratophysella* sp. (Hypogastruridae) (right). (C) *Pogonognathellus* sp. (Tomoceridae) with disruptive scale pattern. (D) Juveniles of *Isotoma* sp. crowding together after hatching. (E) *Adranes coecus*, a clavigerite staphylinid beetle and closely associated inquiline of ants. This 2-mm beetle is perceived as a member of the colony due to secretions from specialized glands.

exude hemolymph through specialized body pores (pseudocelli) or have pre-existing deterrent chemicals externally or internally. Bitzer et al. (2004) described the attack by the rove beetle, *Stenus comma*, which has an extensible, sticky-tipped labium for catching prey, on a hypogastrurid. Following the attack (no harm to springtail), the beetle rubbed its labium on the substrate and regurgitated gut contents. The active chemicals were two isomers of benzoic acid on the springtail integument. Predacious mites attacking *Neanura muscorum* (Neanuridae) fed only momentarily or not at all, backing off and exhibiting enhanced mouthpart cleaning (Messer et al. 1999). Four different compounds presumably in the hemolymph were implicated in this deterrence. The hemolymph of another neanurid, *Brasilimeria assu*, was collected following reflex bleeding and the hemolymph was analyzed; it contained 32 compounds of several different classes, including benzyl benzoate, an ingredient in insect repellents (Zeppelini et al. 2019). Beetles rubbed with secretions and hemolymph exuded from the pseudocelli of the onychiurid *Tetradontophora bielanensis* exhibited disorientation and exaggerated mouthpart cleaning (Dettner et al. 1996). Fluids from the pseudocelli of *Onychiurus* spp. can paralyze predacious mites.

Chemical defense and mimicry may be closely linked. Fig. 21B shows several Collembola of two different families, Neanuridae and Hypogastruridae, in four separate genera. All four species are roughly sympatric geographically. Both families have species that have been shown to produce repellent chemicals. Strikingly, all four have the same basic color scheme of a white thorax and dark body. This contrasting pattern could be warning coloration for four unpalatable hosts, a potential example of Müllerian mimicry among Collembola.

Isotoma species are large (3 mm) springtails. In culture they lay clusters of eggs that hatch nearly simultaneously. These first-stage juveniles do not feed; instead, they cluster tightly together with their heads down and abdomens pointing outward (Fig. 21D). After about 2 weeks, individuals begin to leave the group, feed and develop. This grouping appears to be a defensive approach as the mouthparts and other organs of these juveniles mature. Interestingly, each individual has four small horns near the posterior end, which may serve as a deterrent to small predators. In the field, females might be more likely to lay egg clusters in a hole where the hatchlings would not be exposed so much.

Some soil-dwelling arthropods want to be friends with their potential enemies. The clavigerate beetle (Staphylinidae) in Fig. 21E is an obligate commensal of ants. Its survival in an ant colony is dependent on chemicals produced in glands covered by the orange tufts of setae. These chemicals tell the ants that this beetle is one of them.

Apparent mimicry is widespread among insects as well as the spiders in Fig. 11. All of the small hymenopterans in Fig. 22 are functionally wingless unlike most of their relatives and they more or less resemble ants. However, ants are Hymenoptera as well, so one would expect a certain similarity of design. Bethyids, dryinids and rhopalosomatids (Fig. 22A, B, D, H) are capable of delivering a sting, as are many ant species, which would provide a real deterrent and reinforce a shape to be avoided; this would seem to be a case of Mullerian mimicry for the stinging Hymenoptera and Batesian mimicry for the rest.



Fig. 22 Possible mimicry. Many hymenopterous families include members that have given up flight for a life on the surface and in litter. Their general resemblance to ants is not surprising, given that ants are also Hymenoptera. However, the specimens in (A–E) do appear to show significant convergence to a more ant-like design. (A and B) Bethyidae. (C) Braconidae. (D) Dryinidae (*Gonatopus* sp.). (E) Pteromalidae. (F and G) Scelionidae. (H) Rhopalosomatidae (*Olixon* sp.). Arrows indicate vestigial wings.

Conclusion

Plant health, and indeed the health of the terrestrial biosphere, is dependent upon the efficient functioning of the soil detrital food web, which returns nutrients to the soil for reuse. Soil arthropods are among the critical biotic components that drive this vast engine of decomposition and renewal. Nearly all of the arthropods described in this work can be found easily in backyards, fields and forests. Flipping over rocks, looking on the underside of discarded boards or tapping pine cones in a pan to dislodge the inhabitants can reveal all kinds of tiny yet interesting creatures. A simple hand lens is all that is necessary to begin examining the biodiversity and behaviors of the many thousands of small arthropods that scurry around under our feet.

Acknowledgments

All uncredited images were made by the author, except for those that are acknowledged. For those images I am grateful to Gary Phillips, Adrian Smolis, Kelly Silas Parman, Joseph Berger and Matthew Bowser for the use of their images and especially Suzanne Allen, who gifted me the archives of the late Dr. Robert T. “Tommy” Allen. Fig. 13E is an image made by the late Dr. Mark Muegge while a student at the University of Tennessee. David E. Walter provided valuable advice on mites. I also owe a large debt to former graduate students Cindy Williver Bilbrey, Kelly Felderhoff Oten, Matthew Petersen, Gary Phillips, Nicholas Reynolds and Haley Shae Smith for their enthusiastic interest in all things small.

References

- Albrecht J, Peters MK, Becker JN, Behler C, et al. (2021) Species richness is more important for ecosystem functioning than species turnover along an elevational gradient. *Nature Ecology & Evolution*. <https://doi.org/10.1038/s41559-021-01550-9>. 25 pp.
- Allen RT (2002) A synopsis of the Diplura of North America: Keys to higher taxa, systematics, distributions and descriptions of new taxa (Arthropoda: Insecta). *Transactions of the American Entomological Society* 128(4): 403–466.
- Bajerlein D, Adamski Z, Kacalak W, Tandecka K, Wiesner M, and Jurga S (2016) To attach or not to attach? The effect of carrier surface morphology and topography on attachment of phoretic deutonymphs of *Uropoda orbicularis* (Acari). *Science of Nature* 103: 61. 17 pp <https://doi.org/10.1007/s00114-016-1385-9>.
- Bernard EC (2006) *Morulina delicata*, new species from Great Smoky Mountains National Park and redescription of *M. callowayia* Wray and *M. crassa* Christiansen & Bellinger (Collembola: Neanuridae). *Proceedings of the Biological Society of Washington* 119(4): 540–556.
- Bitzer C, Brasse G, Dettner K, and Schulz S (2004) Benzoic acid derivatives in a hypogastrurid collembolan: Temperature-dependent formation and biological significance as deterrents. *Journal of Chemical Ecology* 30: 1591–1602.
- Chahartaghi M, Langel R, Scheu S, and Russ A (2005) Feeding guilds in Collembola based on nitrogen stable isotope ratios. *Soil Biology & Biochemistry* 37: 1718–1725.
- David J-F (2015) Diplopoda—Ecology. In: Minelli A (ed.) *The Myriapoda. Diplopoda*, vol. 2, pp. 303–324. Leiden: Brill.
- Davies NS, Garwood RJ, McMahon WJ, Schneider JW, and Shillito AP (2021) The largest arthropod in Earth history: Insights from newly discovered *Arthropleura* remains (Serpukhovian Stainmore Formation, Northumberland, England). *Journal of the Geological Society* 179(3): 1–18. <https://doi.org/10.1144/jgs2021-115>.
- Dettner K, Scheuerlein A, Fabian TP, Schulz S, and Francke W (1996) Chemical defense of giant springtail *Tetradontophora bielensis* (Waga) (Insecta: Collembola). *Journal of Chemical Ecology* 22(5): 1051–1074.
- Douglas J, Hoffmann A, Umina P, and Macfadyen S (2019) Factors influencing damage by the Portuguese Millipede, *Ommatoiulus moreleti* (Julida: Julidae), to crop seedlings. *Journal of Economic Entomology* 112(6): 2695–2702. <https://doi.org/10.1093/jeet/toz180>.
- Eisenbeis G and Wicher W (1987) *Atlas on the Biology of Soil Arthropods*. Springer-Verlag: Berlin. 437 pp.
- Foelix RF (2011) *Biology of Spiders*, 3rd edn Oxford: Oxford University Press. 419 pp.
- Hansen JA, Bernard EC, and Moulton JK (2010) A defensive behavior of *Acerentulus confinis* (Berlese) (Protura: Acerentomidae). *Proceedings of the Entomological Society of Washington* 112(1): 43–46.
- Jenner RA, von Reumont BM, Campbell LI, and Undheim EAB (2019) Parallel evolution of complex centipede venoms revealed by comparative proteotranscriptomic analyses. *Molecular Biology and Evolution* 36(12): 2748–2763. <https://doi.org/10.1093/molbev/msz181>.
- Krauß J and Funke W (1995) Collembolen- und Proturen gesellschaften in Fichtenforsten Auswirkungen von Winwurf, Kalk- und Kompostgaben. *Mitteilungen der Deutschen Gesellschaft für allgemeine und angewandte Entomologie* 27: 99–102.
- Lubbock J (1868) Ill. On *Paupopus*, a new type of centipede. *Transactions of the Linnean Society of London* 26(1): 181–190.
- Lutz FE (1948) *Field Book of Insects*, 3rd edn New York: G.P. Putnam's Sons, 510.
- Ma C, Yin X, Kou X, Wang Z, Li X, Jiang Y, Wang H, and Bernard EC (2019) Effects of soil fauna on cellulose and lignin decomposition of plant litter in the Changbai Mountain, China. *Environmental Entomology* 48(3): 592–602. <https://doi.org/10.1093/ee/nvz035>.
- Macias AM, Marek PE, Morrissey EM, Brewer MS, Short DPG, Stauder CM, Wickert KL, Berger MC, Metheny AM, Stajich JE, Boyce G, Rio RVM, Panaccione DG, Wong V, Jones TH, and Kasson MT (2019) Diversity and function of fungi associated with the fungivorous millipede, *Brachycybe lecontei*. *Fungal Ecology* 41: 187–197. <https://doi.org/10.1016/j.funeco.2019.06.006>.
- Marek PE, Buzatto BA, Shear WA, Means JC, Black DG, Harvey MS, and Rodriguez J (2021) The first true millipede—1306 legs. *Scientific Reports* 11: 23126. <https://doi.org/10.1038/s41598-021-02447-0>.
- Messer C, Dettner K, Schulz S, and Francke W (1999) Phenolic compounds in *Neanura muscorum* (Collembola, Neanuridae) and the role of 1,3-dimethoxybenzene as an alarm substance. *Pedobiologia* 43: 174–182.
- Michelbacher A (1938) The biology of the garden centipede, *Scutigera immaculata*. *Hilgardia* 11(3): 55–148. <https://doi.org/10.3733/hilg.v11n03p055>.
- Phillips GJ (2014) *Unpublished Plea Made to Insect Taxonomy Class (EPP548) to Use Microscopes*. University of Tennessee. September 16, 2014.
- Phillips G, Yates DI, Shelley RM, Orstadt PR, and Bernard EC (2019) Investigating commensal relationships of nematodes in millipedes: Life in unexpected places. *American Biology Teacher* 81(4): 278–283. <https://doi.org/10.1525/abt.2019.81.4.278>.
- Pinto-da-Rocha R, Machado G, and Giribet G (eds.) (2007) *Harvestmen*, p. 597. Cambridge, MA: Harvard University Press.
- Potapov AM, Beaulieu F, Birkhofer K, Bluhm SL, Degtyarev MI, Devetter M, Goncharov AA, Gongalsky KB, Klamer B, Korobushkin DI, Lieke DF, Maraun M, McDonnell RJ, Pollierer MM, Schaefer I, Shrubovych J, Semenyuk II, Sendra A, Tuma J, Tümová M, Vassilieva AB, Chen TW, Geisen S, Schmidt O, Tiunov AV, and Scheu S (2022) Feeding habits and multifunctional classification of soil-associated consumers from protists to vertebrates. *Biological Reviews* 2022. <https://doi.org/10.1111/brv.12832>.

- Rawlins AJ, Bulla ID, Ineson P, and Evershed IP (2007) Stabilisation of soil organic matter in invertebrate faecal pellets through leaf litter grazing. *Soil Biology & Biochemistry* 39: 1202–1205. <https://doi.org/10.1016/j.soilbio.2006.10.010>.
- Rust, M.K. & Millard, M.R. 2009. *Silverfish and firebrat management guidelines—UC IPM*. <http://ipm.ucanr.edu> Pest Notes Publication 7475.
- Scheller U (2011) Pauropoda. In: Minelli A (ed.) *The Myriapoda*. vol. 1, pp. 467–508. Leiden: Brill.
- Scheller U, Berg MP, and Jansen MGM (2004) Pauropoda (Myriapoda), a class new to the Dutch fauna, with description of a new species. *Entomologische Berichten* 64: 3–9.
- Schulte F (1989) The association between *Rhabditis necromena* Sudhaus & Schulte, 1989 (Nematoda: Rhabditidae) and native and introduced millipedes in South Australia. *Nematologica* 35: 82–89.
- Smith GB (2017) The Australian silverfish fauna (order Zygentoma)—Abundant, diverse, ancient and largely ignored. *General & Applied Entomology* 45: 9–58.
- Starling JH (1944) Ecological studies of the Pauropoda of Duke Forest. *Ecological Monographs* 14(3): 293–310.
- Szucsich N and Scheller U (2011) Symphyla. In: Minelli A (ed.) *The Myriapoda*. vol. 1, pp. 445–466. Brill: Leiden.
- The Animals (1965) “We Gotta Get Out Of This Place.” (Written by Barry Mann, Cynthia Weil.) *Animal Tracks*. MGM Records.
- Unwin EE (1931) On the structure of the respiratory organs of the terrestrial Isopoda. In: *Papers and Proceedings of the Royal Society of Tasmania* 37–104.
- Voigtländer K (1981) Chilopoda—Ecology. In: Minelli A (ed.) *The Myriapoda. Chilopoda, Symphyla, Pauropoda*, vol. 1, pp. 309–325. Leiden: Brill.
- von Beeren C, Brückner A, Hoelne PO, Ospina-Jara B, Kronauer DJC, and Blüthgen N (2021) Multiple phenotypic traits as triggers of host attacks towards ant symbionts: Body size, morphological gestalt, and chemical mimicry accuracy. *Frontiers in Zoology* 18: , 46. <https://doi.org/10.1186/s12983-021-00427-8>.
- Wang H, Lu L, She D, Wen Z, Mo Z, Li J, and Li H (2018) Eating centipedes can result in *Angiostrongylus cantonensis* infection: Two case reports and pathogen investigation. *The American Journal of Tropical Medicine and Hygiene* 99(3): 743–748. <https://doi.org/10.4269/ajtmh.18-0151>.
- Wygodzinsky, P. and Schmidt K. 1980. Survey of the Microcoryphia (Insecta) of the northeastern United States and adjacent provinces of Canada. *American Museum Novitates* Number 2710, 1980, 7 pp.
- Yanoviak SP, Kaspari M, and Dudley R (2009) Gliding hexapods and the origins of insect aerial behaviour. *Biology Letters* 5: 510–512. <https://doi.org/10.1098/rsbl.2009.0029>.
- Zeppelini D, Queiroz GC, Lopes NP, and Mendonca-Junior FJB (2019) Chemical analysis of *Brasilimeria* Stach, 1949 (Hexapoda, Collembola, Neanuridae) hemolymphatic secretion, and description of a new species. *PLoS One* 14(2), e0212451. <https://doi.org/10.1371/journal.pone.0212451>.

Further reading

Some comprehensive web sites for learning about soil arthropods

- Silverfish (n.d.) *1726 Silverfish Images, Stock Photos & Vectors*. <https://www.shutterstock.com/search/silverfish>. Commercial platform showing silverfish from every conceivable angle.
- Bellinger PF, Christiansen KA and Janssens F (1996–2023) *Checklist of the Collembola of the World*. <http://www.collembola.org>. Comprehensive site on everything relating to springtails, including thousands of images from around the world; indispensable resource for anyone who likes or studies Collembola.
- Bowser, ML (2012) *Bristletails of North America*. <https://archaeognatha.myspecies.info/node/1>. Lots of good images of different jumping bristletails and a comprehensive list of species and literature.
- BugGuide: Identification, Images & Information for Insects, Spiders & Their Kin for the United States & Canada (2003–2022) *Iowa State University*. <https://bugguide.net/node/view/15740>. A huge mass of knowledge and images growing larger by the day, with identifications curated by experts.
- Canadian Journal of Arthropod Identification (n.d.) *Canadian Journal of Arthropod Identification*. <https://cjai.biologicalsurvey.ca/>. This web-based journal features comprehensive reviews of a wide range of insect taxa, some of which are associated with soil. Most articles are Canada-centric but still of value far outside Canadian borders. The color images alone are worth hours of browsing.
- Edaphobase (n.d.) *Edaphobase is a pan-European data warehouse for soil biodiversity, with more than 400,000 observations from more than 30,000 localities for thousands of different soil taxa*. <https://www.senckenberg.de/en/science/research-infrastructure/databases-and-digital-resources/edaphobase-database-on-soil-zoology/>. This site includes a searchable database and associated maps, mostly for Europe but also with many worldwide datapoints.
- Galerie (n.d.) *Galerie du Monde des insectes*. 2002–2022. <https://www.galerie-insecte.org/galerie/>. A valuable resource with thousands of high-quality images.
- Harvey MS (2011) *Pseudoscorpions of the world, version 2.0*. Western Australian Museum, Perth. <http://www.museum.wa.gov.au/catalogues/pseudoscorpions>. One-of-a-kind site for pseudoscorpions.
- Murray A (2022) *A Chaos of Delight. Exploring Life in the Soil*. <https://www.chaosofdelight.org/>. The finest photomicrography, not just showing images but also telling stories via the imaging.
- Orgiazzi A, Bardgett RD, Barrios E, Behan-Pelletier V, Briones MJL, Chotte JL, De Deyn GB, Eggleton P, Fierer N, Fraser T, Hedlund K, Jeffrey D, Johnson NC, Jones A, Kandeler E, Kaneko N, Lavelle P, Lemanceau P, Miko L, Montanarella L, de Souza Moreira FM, Ramirez KS, Scheu S, Singh BK, Six J, van der Putten WH and Wall DH (2016) *Global Soil Biodiversity Atlas*. European Commission, Publications Office of the European Union, Luxembourg. <https://www.globsoilbiodiversity.org/atlas-introduction>. A big international effort with substantial image galleries and effective summaries of biodiversity and ecology of soil organisms.
- Shelley RM (2022) *The myriapods, the world's leggiest animals*. <https://myriapodology.org/nadiplochilo/index.html>. The late Dr. Rowland Shelley, a world authority on millipedes and centipedes and adjunct professor at the University of Tennessee, developed *Nadiplochilo* as a comprehensive source of images of millipedes, centipedes and the many scientists who have studied them for the past 250 years. Through the kindness of Lourdes Shelley, the site contents were archived at the University of Tennessee and later transferred to the Centre International de Myriapodologie - International Society for Myriapodology (<https://myriapodology.org/>) where they can be updated and made publicly available.
- Walter DE (2009–2017) *Macromite's blog*. <https://macromite.wordpress.com/>. A delightful site covering myriad topics on mites, making them far more fascinating than anyone could have imagined.

General References

- Chapin FS III, Matson PA, and Vitousek PM (2011) *Principles of Terrestrial Ecosystem Ecology*, 2nd edn New York: Springer, 529.
- Coleman DC, Callaham MA Jr., and Crossley DA Jr. (2018) *Fundamentals of Soil Ecology*, 3rd edn London: Academic Press, 369.
- Coleman DC and Hendrix PF (eds.) (2000) *Invertebrates as Webmasters in Ecosystems*, p. 336. Wallingford: CABI Publishing.
- Crossley DA Jr., Coleman DC, Hendrix PF, Cheng W, Wright DH, and Edwards CA (1991) Modern techniques in soil ecology. *Agriculture, Ecosystems & Environment* 34: 1–510.
- Dindal DL (ed.) (1990) *Soil Biology Guide*. New York: John Wiley & Sons. 1349 pp.
- Gibb TJ and Oseto C (2020) *Insect Collection and Identification: Techniques for the Field and Laboratory*, 2nd edn London: Elsevier, 339.
- Holldobler B and Wilson EO (1990) *The Ants*. Cambridge, MA: Belknap Press, 746.
- Imadate G (1974) *Protura (Insecta)*. Tokyo: Keigaku Publishing, 351.
- IUSS Working Group (2015) *World Reference Base for Soil Resources 2014, Update 2015*. World Soil Resources Reports 106 Rome: FAO. ISBN: 978-92-5-108369-7.

Arachnida

- Beaulieu F, Knee W, Nowell V, Schwarzfeld M, Lindo Z, Behan-Pelletier VM, Lumley L, Young MR, Smith I, Proctor HC, Mironov SV, Galloway TD, Walter DE, and Lindquist EE (2019) Acari of Canada. In: Langor, D.W. & Sheffield, C.S. (eds.). *The Biota of Canada—A Biodiversity Assessment*. Part 1: The Terrestrial Arthropods. *ZooKeys* 819: 77–168. <https://doi.org/10.3897/zookeys.819.28307>.

- Buddle CM (2010) Photographic key to the pseudoscorpions of Canada and the adjacent USA. *Canadian Journal of Arthropod Identification*. <https://doi.org/10.3752/cjai.2010.10.No.10>.
- Harvey MS (1988) The systematics and biology of pseudoscorpions. In: Austin AD and Heather NW (eds.) *Australian Arachnology*, pp. 75–85. Brisbane: Australian Entomological Society.
- Harvey MS (2002) The neglected cousins: What do we know about the smaller arachnid orders? January 2002. *Journal of Arachnology* 30(2): 357–372.
- Harvey MS (2011) *Pseudoscorpions of the World, Version 2.0*. Perth: Western Australian Museum. <http://www.museum.wa.gov.au/catalogues/pseudoscorpions>.
- Houck MA (ed.) (1994) *Mites: Ecological and Evolutionary Analyses of Life-History Patterns*, p. 357. New York: Chapman and Hall.
- Krantz GW and Walter DE (eds.) (2009) *A manual of Acarology*, 3rd edn, p. 807. Lubbock, Texas: Texas Tech University Press.
- Stockmann R and Ythier E (2010) *Scorpions of the World*. Paris: NAP Editions, 572.
- Ubick D, Paquin P, Cushing PE, and Roth V (eds.) (2017) *Spiders of North America, an Identification Manual*, p. 425. Keene, NH: American Arachnological Society.
- Walter DE and Proctor HC (1999) *Mites: Ecology, Evolution and Behavior*. Wallingford: CABI Publishing, 322.

Phoresy

- Bartlow AW and Agosta SJ (2021) Phoresy in animals: Review and synthesis of a common but understudied mode of dispersal. *Biological Reviews* 96: 223–246. <https://doi.org/10.1111/brv.12654>.

Basal hexapods

- Chick AIR (2018) A revised checklist of the UK silverfish (Zygentoma: Lepismatidae). *Zootaxa* 4504: 447–450. <https://doi.org/10.11646/zootaxa.4504.3.10>.
- Hopkin SP (1997) *Biology of the Springtails (Insecta: Collembola)*. Oxford: Oxford University Press, 330.
- Sturm H and Machida R (2001) Archaeognatha. In: *Handbuch der Zoologie IV*, p. 213. Berlin: Walter de Gruyter.
- Szeptycki A (2007) Catalogue of the world Protura. *Acta Zoologica Cracoviensis* 50B(1): 1–210.
- Tuxen SL (1964) *The Protura*. Paris: Hermann, 360.
- Wygodzinsky P (1972) A review of the silverfish (Lepismatidae, Thysanura) of the United States and the Caribbean area. *American Museum Novitates* 2481: 26.

Crustacea

- Warburg MR (1987) Isopods and their terrestrial environment. In: Cragg JB (ed.) *Advances in Ecological Research*. vol. 17, pp. 187–242. London: Academic Press.

Myriapoda

- Lewis JGE (1981) *The Biology of Centipedes*. Cambridge: Cambridge University Press, 476.
- Means JC, Francis EA, Lane AA, and Marek PE (2015) A general methodology for collecting and preserving xystodesmid and other large millipedes for biodiversity research. *Biodiversity Data Journal* 3: e5665. <https://doi.org/10.3897/BDJ.3.e5665>.
- Minelli A (ed.) (2011) *The Myriapoda*. In: *Chilopoda, Symphyla, Pauropoda*, vol. 1. Leiden: Brill. 530 pp.
- Minelli A (ed.) (2015) *The Myriapoda*. In: *Diplopoda*, vol. 2. Leiden: Brill. 482 pp.
- Shear WA (2015) The chemical defenses of millipedes (Diplopoda): Biochemistry, physiology and ecology. *Biochemical Systematics and Ecology* 61: 78–117. <https://doi.org/10.1016/j.bse.2015.04.033>.

Nematodes

Deborah A Neher^a and Thomas O Powers^b, ^aDepartment of Plant and Soil Science, University of Vermont, Burlington, VT, United States;
^bDepartment of Plant Pathology, University of Nebraska, Lincoln, NE, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction – Life in a soil pore	105
Nematode distribution and abundance	109
Physiology	109
Ecology	110
Monitoring	110
References	111

Abstract

A hallmark of the Phylum Nematoda is their exceptional abundance and ubiquitous presence in soil as microscopic fauna. Nematodes require free water for normal functioning, but many are capable of entering a cryptobiotic state when conditions turn unfavorable. Communities span several trophic positions within the soil food web, collectively contributing directly to the biogeochemistry of soils by regulating decomposition and nutrient cycling. Nematode communities serve as biological indicators of soil health because they respond predictably to land-management practices. Valuing nematode communities as biological indicators of soil condition represents a tremendous shift in emphasis in the science of nematology that historically focused on plant-parasitic nematodes.

Key points

- Nematodes are the most abundant multicellular animals in the planet.
- Nematodes have colonized nearly every habitat on earth.
- Nematodes can exist in extreme environmental conditions.
- Parasitic nematodes can be lethal to plants and animals.
- Numerous ecosystem services are provided by nematodes but the extent and significance of these ecosystem services is yet to be determined.

Introduction – Life in a soil pore

Few people are aware of the most numerous of all soil-dwelling animals, the nematodes. These minute, unsegmented roundworms are actually aquatic organisms, inhabiting water films (1–5 μm thick) that coat and surround soil particles. This chapter focuses on nematodes that typically complete their entire life cycle within terrestrial soils or within plant roots. Numerous parasites of vertebrates have soil dwelling stages, but their development is dependent on entering an appropriate host species. Vertebrate nematode parasites include hookworms, intestinal roundworms, pinworms, whipworms, and other serious pests of humans and domesticated animals (Maggenti, 1981). Information about vertebrate parasites is readily found in textbooks of parasitology (Roberts et al., 2013). Their size, typically 0.40–1.0 mm in length, is ideal for navigating the porous matrix of soil. Although fully functional within a water film, they are ultimately bounded by soil structure, restricted in movement within soil pores less than 30 μm diameter. Nematodes are simple organisms (Fig. 1). Comprised of approximately 1000 somatic cells in the adult stage, these worm-like organisms are an example of functional and anatomical economy. Lacking eyes, appendages, and true segmentation, nematodes use mechanosensory and chemosensory neurons embedded in the cuticle to orient and respond to a wide range of environmental stimuli. At the nematode's anterior end is a cirlet of sensilla arranged around an oral opening (Fig. 2). These sensilla, including two larger, laterally placed chemosensory organs called 'amphids,' detect subtle chemical gradients, providing directional information to the nematode (Fig. 3). Plant-feeding nematodes respond to slight variations in CO_2 and root exudates. A positive chemosensory signal initiates a snake-like sinuous movement characteristic of most nematodes, created by longitudinal body muscles working in opposition to a hydrostatic skeleton. Directional movement more than a few centimeters per day is unlikely in the soil environment. Once they contact the root surface, plant-feeding nematodes will probe with their stylet, a hollow, protrusible, hypodermic needle-like feeding tube (Fig. 4A). Some plant-feeding nematodes will feed externally on the root surface, using this stylet to puncture cells, withdrawing the cytoplasmic contents. Others will penetrate the root and establish permanent feeding sites within the root cortex, or migrate cell-to-cell, leaving a trail of damaged necrotic tissue.

Free-living nematodes also occupy the interstitial spaces in soil. Unlike the plant-feeders, free-living nematodes are seldom sedentary, continually moving to feed on a diverse array of food, including bacteria, algae, fungi, protozoa, small invertebrates, and

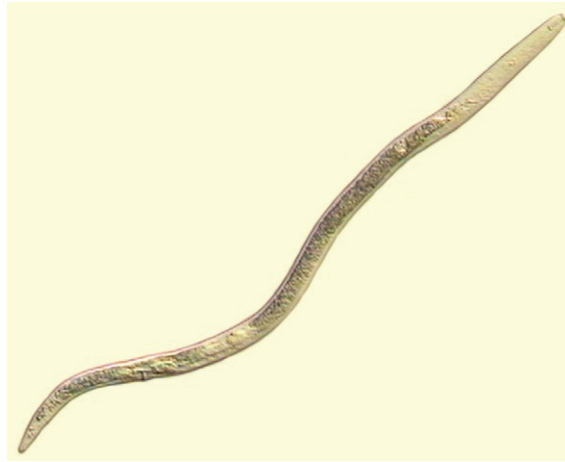


Fig. 1 Entire body view of a female *Pratylenchus agilis* ($\times 100$ magnification), collected on the Konza Prairie (96° W $35'$ 39° N $05'$) beneath Scribner's panicum (*Panicum scribnerianum*) and bluegrass (*Poa pratensis*) near Manhattan, Kansas. Photograph is provided courtesy of Peter Mullin.



Fig. 2 Anterior view of a member of the order Dorylaimida with an intricate pattern of striations in the cuticle, collected beneath tallgrass prairie near the Audubon Spring Creek Prairie outside of Lincoln, Nebraska. Photograph is provided courtesy of Abigail Borgmeier.

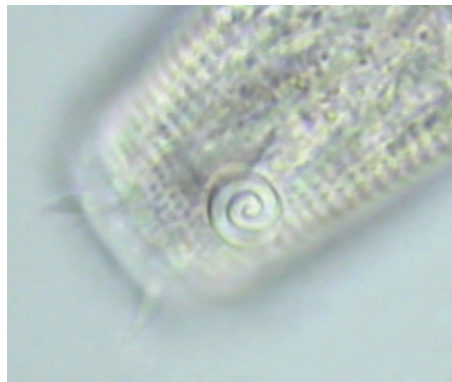


Fig. 3 Spiral-shaped chemosensory organs called 'amphids' in an anterior position of *Achromadora* sp. collected from soil of Jumbo Valley fen in Cherry County, Nebraska.

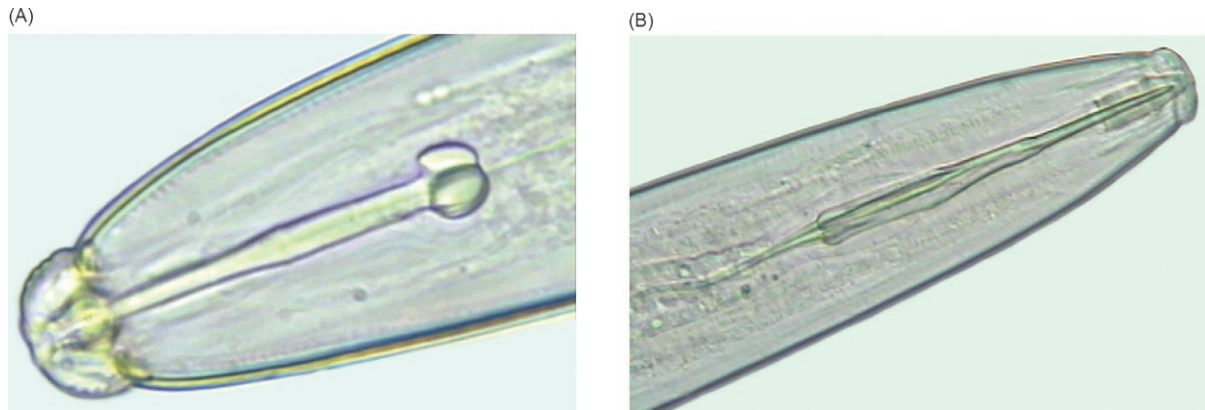


Fig. 4 Variation in morphology of spear-like structure in oral opening: (A) male plant-parasite *Hoplolaimus galeatus* ($\times 1000$ magnification), collected from soil with big bluestem (*Andropogon gerardii* Vitman) in the Konza Prairie (96° W 35° 39° N $05'$), near Manhattan, Kansas; and (B) female fungivore *Enchodelus hopedorus* ($\times 400$ magnification), collected from the summit of Long's Peak, Colorado (105° W 35° 40° N $16'$). Photographs are provided courtesy of Peter Mullin.

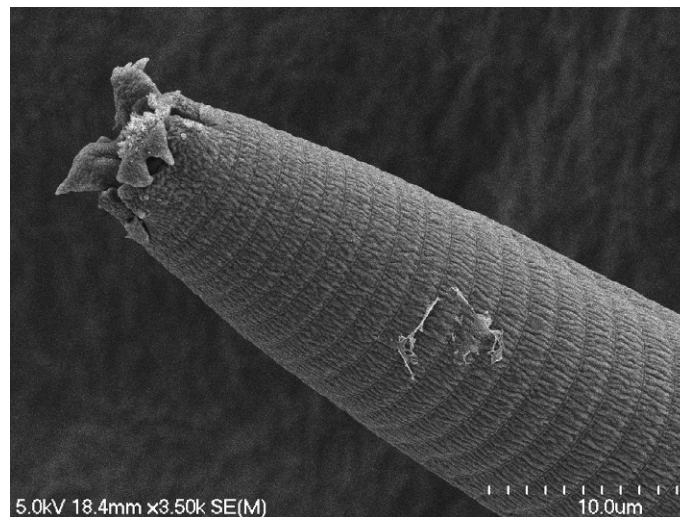


Fig. 5 Cuticle elaborations of the lips around the oral opening *Chiloplacus* species (3500 magnification, collected in soil from Spring Creek Prairie near Lincoln Nebraska. Photograph is provided courtesy of Abigail Borgmeier.

other nematodes. They use the same set of sensilla to track their food source; however, their feeding structures are modified to suit their meal (Freckman and Baldwin, 1990). Bacterial-feeders (bacterivores) graze using a relatively simple tubular mouthpart, although the cuticle surrounding the oral opening may be modified elaborately to direct food toward the stoma (Fig. 5). Predators may be adorned with 'teeth,' used to puncture or shred the integument of various invertebrates (Fig. 6). Fungal-feeders (fungivores) and omnivores have a stylet similar in appearance to that of plant-feeders, distinguished by the lack of stylet 'knobs,' points of muscle attachment for stylet protrusion (Fig. 4B).

Early observations of nematodes by light microscopy led scientists to refer to nematodes as a 'tube within a tube.' Food entering the nematode is channeled quickly into a tubular pharynx and then passed into an intestine that comprises the bulk of the body cavity. Waste products are eliminated through a posterior ventral anus. There is no circulatory or respiratory system within the nematode. In addition to the alimentary system, the reproductive system is a conspicuous feature in the adult nematode (Fig. 7). Females and hermaphrodites are characterized by a branched genital system, which produces eggs that exit the body via a ventral genital opening in the body wall. Males possess a spicule, a cuticularized, protrusible modification of the genital system that guides sperm into the female for internal fertilization (Fig. 8). Many nematodes reproduce by amphimixis, but parthenogenesis and self-fertilizing hermaphroditism are also common among species. Regardless of mode of reproduction, nematodes are prolific animals. Some species complete a lifecycle of 3 or 4 days. Others require several months to go from egg to egg-laying female. *Meloidogyne* spp., an economically important group of plant parasites, will typically begin to produce eggs within 3 weeks of hatching; each female is capable of producing hundreds of eggs during an adult life.

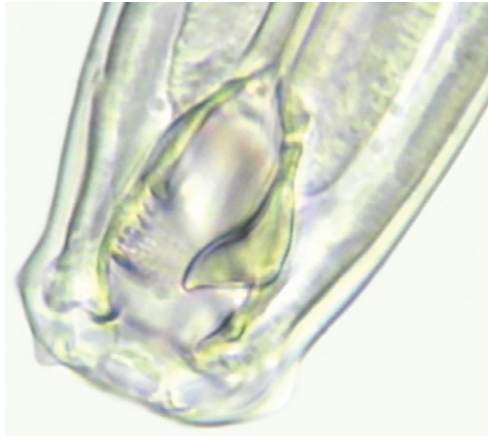


Fig. 6 Teeth of oral opening of predator *Mylonchulus montanus* ($\times 1000$ magnification), collected in soil with big bluestem in the Konza Prairie (96° W 35° 39' N $05'$) near Manhattan, Kansas. Photograph is provided courtesy of Peter Mullin.



Fig. 7 Ovary in reproductive tract of female *Axonchium micans* ($\times 400$ magnification), collected in soil with big bluestem/Scribner's panicum in the Konza Prairie, near Manhattan, Kansas. Photograph is provided courtesy of Peter Mullin.

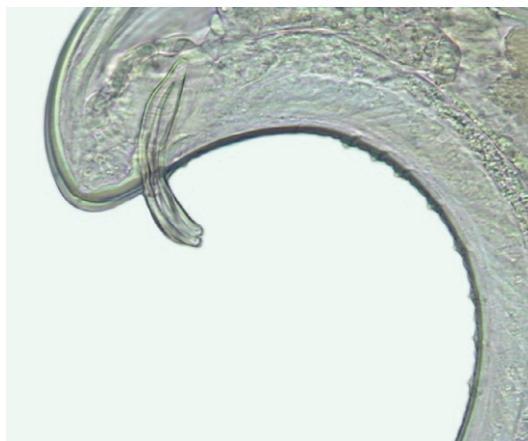


Fig. 8 Protrusible modification of male genitalia of *Aporcelaimellus obscuroides* ($\times 400$ magnification), collected in soil with big bluestem/Scribner's panicum in the Konza Prairie, near Manhattan, Kansas. Photograph is provided courtesy of Peter Mullin.

Table 1 Nematode communities in selected studies.

Ecosystem type	Abundance (10^6 m^{-2})	Genera (n)	Species (n)	Bacterivore (%)	Fungivore (%)	Herbivore (%)	Predator/omnivore (%)
Grassland	2.3–20	30–124	71–384	24–38	5–20	16–41	26–40
Deciduous forest	2.3–3.7	22–81	34–175	27–46	15–25	13–31	5–32
Tropical rainforest	NA	107	204	35	5	29	30
Tundra	NA	58	162	55	3	22	19
Heath	1.2	NA	NA	57	NA	2	NA
Agroecosystems	3.5–5.0	19–44	33–100	36–46	13–19	33–41	5–7
Temperate desert	NA	18	23	33	22	17	28
Antarctic coastal	NA	15	27	47	27	0	27
Antarctic dry valley	NA	3	3	67	0	0	33

NA, not available.

Nematode distribution and abundance

The hallmark of Phylum Nematoda is the exceptional abundance and ubiquitous presence of nematodes (Table 1). It is a challenging exercise to identify a habitat not occupied by a diverse community of nematodes. Consider the dry valleys of Antarctica, i.e., so extreme environment that it is thought to provide the bare minimum for sustaining life on earth. There, in McMurdo Dry Valleys, four nematode species occupy the top levels in the soil food chain, three bacterivores and a fourth species, *Eudorylaimus antarcticus*, an omnivore that may feed on the two bacterivorous species (Freckman and Virginia, 1997). Soils of the high Arctic are populated so heavily with nematodes that biologists have searched for explanations for an Arctic nematode biodiversity that exceeds that of the humid tropics. Temperate grasslands appear to sustain the greatest levels of diversity and abundance in the phylum (Todd et al., 2006). On the Konza Prairie, one of the last remaining large stands of true tallgrass prairie in North America, each handful of soil is likely to contain 50–100 different nematode species. An estimated 1.5 million nematodes can be extracted from a square meter of soil, just within the surface 10 cm. Some researchers have considered the plant-feeding or herbivorous species as ‘belowground buffalo,’ referring to their impact as grazers of plant roots. Nearly all researchers agree that the approximately 25,000 described species of nematodes represent a small percentage of the species that actually exist in nature (van den Hoogen et al., 2019).

Finding agreement on those factors responsible for extraordinary nematode abundance and diversity is not so simple. Clearly, the evolutionary age of nematodes has provided considerable time for nematode diversification. Recent molecular studies indicate that the roots of the nematode evolutionary tree are hundreds of millions of years older than the corresponding tree for insects. Marine nematodes, which inhabit even the deepest ocean floors, have a nucleotide diversity that suggests nematodes flourished before life colonized land. The same nematode survival mechanisms that allowed them to be among the first colonizers of land most likely contribute to a competitive advantage in current-day harsh environments (Wharton, 2003).

Physiology

Although nematodes depend on water films for normal functioning, under harsh conditions such as freezing or drying many nematodes are capable of entering a cryptobiotic state, essentially a reversible state of suspended animation until favorable conditions return (Perry and Wright, 1998). Anhydrobiotic success, the ability to withstand the lack of moisture, depends on a slow rate of dehydration, rapid decline in cuticle permeability, and accumulation of compounds such as trehalose and glycerol to stabilize phospholipids and proteins. Species including *Acroboloides* spp., *Aphelenchus avenae*, and *Scutellonema brachyurum* start to enter anhydrobiosis by coiling at gravimetric water contents of 3.7% (–300 kPa), 9% (–30 kPa), and 15% (–10 kPa), respectively. It is unknown how long nematodes may persist in this state and still survive hydration. Specimens of the wheat gall nematode *Anguina tritici* have been revived after more than 30 years in anhydrobiosis.

Nematodes are not alone among invertebrates in their ability to enter a cryptobiotic state. Therefore, additional clues to their successful colonization of soil are found in their physiological versatility. Physiological adaptations have allowed nematodes to invade habitats in which few other animals can survive and avoid interspecific competition and many environmental selection pressures. Nematodes are capable of regulating their uptake of oxygen over a wide range of partial pressures from 100 to 5%. When oxygen falls below 5%, nematodes convert their use of energy reserves. For example, *Aphelenchus avenae* uses reserves of neutral lipids under aerobic conditions and glycogen under anaerobic conditions. Not only does a hydrostatic skeleton facilitate movement, but also an ability to expand or contract body size in response to gradual changes in concentration of solutions containing sodium, calcium, magnesium, or potassium ions. The adjustment process, osmoregulation, is possible by modifying the permeability of their cuticle. Furthermore, nematodes can tolerate a wide pH range, with some species capable of withstanding strong acids or bases (pH 1.6–11.0). Nematode tolerance for temperature is equally remarkable. In hot springs, nematodes live at temperatures as high as or higher than any other Metazoa. *Aphelenchoides parietinus*, a cosmopolitan species, has been reported from 58 °C springs in Chile and 61 °C springs in New Zealand. Moreover, nematode species recently discovered in Mono Lake, California can withstand 500× the human lethal dose of arsenic (Shih et al., 2019).

Ecology

Although nematodes have adapted mechanisms to survive extremities of climate, their activity is stimulated by the return of more moderate conditions. For example, communities of nematodes are revived after rain in desert soils or after a relatively warm period in soils of polar regions. Species are found interacting with other organisms in a variety of occupations including competitor, opportunist, parasite, host, predator, or prey. As a competitor, nematodes rival protozoa for access to bacterial food sources. However, differences in size and generation time can alleviate competition by specializing on microhabitats fine-tuned to minute spatial and temporal scales. As opportunists, nematode species are phoretically transported by insects, adhering to their bodies, providing an opportunity to reach food sources at a longer distance than they would be capable of reaching alone. For example, Psychodidae flies carry nematodes to fermenting organic matter; dung beetles carry nematodes to fresh dung; and Scolytidae bark beetles transport Tylenchida and Rhabditida to their tunnels, where they feed on bacteria and fungi. Nematodes that are parasites of plants include both specialists and generalists. Host ranges may extend to hundreds of plant species or be restricted to a single plant variety. Some nematodes specialize in precise feeding sites along the root. Plant-parasitic nematodes affect primary productivity of plants by altering uptake of water and nutrients. These abnormalities may result from changes in root morphology and/or physiology resulting in reduced productivity. Other nematodes infect insects, killing them within 48 h by releasing insect-pathogenic bacteria into the insect. The insect host dies from the bacterial infection, and nematodes feed on the bacteria and develop and reproduce inside the insect cadaver. In this case, the nematode and bacteria have a mutualistic association. Sometimes, a nematode finds itself in a role reversal, becoming the host or target of fungi specialized for trapping nematodes, using constricting rings, sticky knobs, or hyphal nets. Being in the middle of a food chain, nematodes eventually become prey to higher-order consumers. Nematodes provide a portion of the diet for many kinds of small soil arthropods. For example, a symphylan can hold seven large nematodes in its gut at one time. In 1 day, a mite may consume two large nematodes or several smaller nematodes.

Genetic diversity, nematode abundance, and the variety of niches occupied by nematodes has led to the assessment of nematode communities as a means of making predictions about past or present soil conditions. The species composition and relative nematode abundance of species comprising a nematode community allow nematodes to serve as biological indicators for soil disturbance as well as 'soil health.'

Monitoring

Nematodes contribute directly to biogeochemistry of soils by regulating processes such as organic matter decomposition and nutrient cycling (Freckman, 1988; Ingham et al., 1985). Nematodes do not feed on decaying organic matter directly, but on bacteria, fungi, algae, and actinobacteria that colonize and decompose decaying plant and animal debris. Indirectly, nematodes control availability of nutrients by regulating the abundance or activity of these organisms, releasing nitrogen and phosphorus from microbes they digest, immobilizing nutrients in their live tissues, and excreting excess nitrogen as ammonium. Under field conditions, bacterivorous and predatory nematodes contribute 8–19% of nitrogen mineralization in conventional and integrated farming systems, respectively (Neher et al., 2012).

Nematode community structure and function are known to change in response to land-management practices such as nutrient enrichment through fertilization by organic or inorganic nitrogen, cultivation, liming, drainage, plant community composition and age, and toxic substances such as heavy metals, pesticides, and polycyclic aromatic hydrocarbons (Neher, 2010). Disturbance can affect the survival of individuals directly, or indirectly by changing resource levels. For example, a larger ratio of fungivorous to bacterivorous nematodes suggests less effect of cultivation than does a smaller ratio. Any disturbance that results in compaction reduces soil porosity, and the numbers of relatively large-sized nematodes. General biocides such as methyl bromide can nearly eliminate nematodes, returning the ecological succession of soil communities to that assembling a depauperate soil matrix. Although recovery occurs eventually, abundance and diversity of nematode communities may take years to achieve prefumigation levels. Alternatively, some herbicides affect nematodes indirectly by reducing vegetation and smaller additions of organic matter to soil.

The composition of nematode communities may reflect the frequency of disturbance to soil, whether the disturbance is primarily physical and/or chemical in origin (Neher, 2010). This concept is based on the principle that different species have contrasting levels of sensitivity or tolerance to stress because of unique survival and/or reproductive traits. Smaller values of an index that integrates relative abundance of species within each sensitivity/tolerance class are indicative of a more disturbed environment, and larger values may indicate a less-disturbed environment. Ideally, indicators of soil health would correspond with ecological processes occurring in soil. Because initial experiments have been correlative in nature, it is too soon to claim such an association with great confidence.

The concept of nematodes as indicators of soil condition and their predictive use in soil management represents a tremendous shift in emphasis in the science of nematology (Beaumelle et al., 2021). Formerly a science focused on the control of parasitic and harmful species, now nematodes are an integral and potentially useful component of soil systems. Considering this new perspective, it is likely that the study of nematodes will contribute significantly to our understanding of biological processes in the soil.

References

- Beaumelle L, Thouvenot L, Hines J, et al. (2021) Soil fauna diversity and chemical stressors: A review of knowledge gaps and roadmap for future research. *Ecography* 44: 1–15.
- Freckman DW and Baldwin JG (1990) Nematoda. In: Dindal DL (ed.) *Soil Biology Guide*, pp. 155–200. New York: John Wiley.
- Freckman DW (1988) Bacterivorous nematodes and organic-matter decomposition. *Agriculture, Ecosystems and Environment* 24: 195–217.
- Freckman DW and Virginia RA (1997) Low-diversity Antarctica soil nematode communities: Distribution and response to disturbance. *Ecology* 78: 363–369.
- Ingham RE, Trofymow JA, Ingham ER, and Coleman DC (1985) Interaction of bacteria, fungi, and their nematode grazers: effect on nutrient cycling and plant growth. *Ecological Monographs* 55: 119–140.
- Maggenti A (1981) *General Nematology*. New York: Springer-Verlag.
- Neher DA (2010) Ecology of plant and free-living nematodes in natural and agricultural soil. *Annual Review of Phytopathology* 48: 371–394.
- Neher DA, Weicht TR, and Barbercheck ME (2012) Linking invertebrate communities to decomposition rate and nitrogen availability in pine forest soils. *Applied Soil Ecology* 54: 14–23.
- Perry RN and Wright DJ (1998) *The Physiology and Biochemistry of Free-living and Plant-parasitic Nematodes*. New York: CABI Publishing.
- Roberts L, Janovy J, and Nadler S (2013) Nematodes: Trichinellida and Dioctophymatida, Enoplean Parasites. In: Schmidt GD and Roberts LS (eds.) *Foundations of Parasitology*, 6th ed., pp. 377–390. New York: McGraw-Hill.
- Shih P-Y, Lee JS, Shinya R, et al. (2019) Newly identified nematodes from Mono Lake exhibit extreme arsenic resistance. *Current Biology* 29: 3339–3344. <https://doi.org/10.1016/j.cub.2019.08.024>.
- Todd TC, Powers TO, and Mullin PG (2006) Sentinel nematodes of land-use change and restoration in tallgrass prairie. *Journal of Nematology* 38: 20–27.
- van den Hoogen J, Geisen S, Routh D, et al. (2019) Soil nematode abundance and functional group composition at a global scale. *Nature* 572: 194–198. <https://doi.org/10.1038/s41586-019-1418-6>.
- Wharton DA (2003) The environmental physiology of Antarctic terrestrial nematodes: A review. *Journal of Comparative Physiology B* 173(8): 621–628.

Relevant websites

- <https://nemys.ugent.be/>—World Database of Nematodes.
- <http://www.wormbook.org/>—WormBook: The Online Review of *C. elegans* Biology.
- <http://www.marinespecies.org/>—World Register of Marine Species.

Earthworms: Essential ecosystem engineers providing vital ecosystem services

Maria JI Briones, Departamento de Ecología y Biología Animal, Universidad de Vigo, Vigo, Spain

© 2023 Elsevier Ltd. All rights reserved.

Introduction	112
Environmental filters shaping their geographical range	113
Functional diversity: Habitat and dietary requirements	114
Interactions with other soil organisms and aboveground vegetation	115
Role of earthworms in the delivery of ecosystem services	115
Conclusion	117
References	117

Abstract

Earthworms are considered to be ‘ecosystem engineers’ because their impact on the soil environment significantly influence the activities of other soil organisms (plants and microorganisms). Their pivotal role in ecosystem functioning translates into the delivery of essential ecosystem services including plant growth, soil structure, nutrient cycling, water and climate regulation and cultural services. Despite the extensive research into their taxonomy and ecology, there is substantial lack of detailed knowledge on the potential effects of current and future environmental threats on their distribution and performance of vital soil functions. Actions to protect these organisms are needed to effectively conserve our soils.

Glossary

Ecological categories Species groupings based on morphological, ecological and feeding characteristics defined by Bouché (1977) and that include epigeic, anecic, and endogeics together with some intermediate categories.

Cast Fecal material egested by earthworms that is deposited in the burrows or at the soil surface; they can adopt different shapes, including pellet cast, turret cast and mass cast.

Drilodefensins Unique metabolites that counteract the inhibitory effects of polyphenols on earthworm gut enzymes.

Drilosphere The area of soil directly affected by earthworm burrowing activities; it typically refers to the thin (few millimeters thick) layer around the burrows.

Midden The mound of cast, litter and organic materials accumulated by an anecic earthworm at the burrow entrance.

Rhizosphere The area immediately surrounding the plant roots of intense biological activity attracted by root exudates.

Key points

- Earthworms are considered to be “ecosystem engineers” because they create, maintain and modify the soil environment and thus affect the activities of other soil organisms (either directly or indirectly).
- Their high taxonomical and functional diversity enables them to exploit different trophic niches and withstand adverse environmental conditions; however, perturbations in soil abiotic and biotic properties can have profound alterations on their community structure and activities.
- They are bioturbating agents as they create burrows, incorporate litter along the soil profile, mix soil and produce casts; this makes them one of the most important decomposers in soils and responsible for the delivery of provisioning, supporting, regulating and cultural ecosystem services.
- Despite their pivotal role in ecosystem functioning, research is biased toward their ecology and the most economically beneficial services (primary productivity, vermicomposting, fish bait).
- Importantly, they are excluded from soil protection regulations and the threatened status of the majority of the earthworm species is currently unknown.

Introduction

Earthworms are terrestrial segmented animals (phylum Annelida) which at maturity develop a “clitellum” (the glandular swelling on several segments in the anterior half of the body that forms the cocoons or egg capsules); this is why they are included in the class Clitellata. Because each body segment bears a low number of short setae, they are placed in subclass Oligochaeta (usually eight on

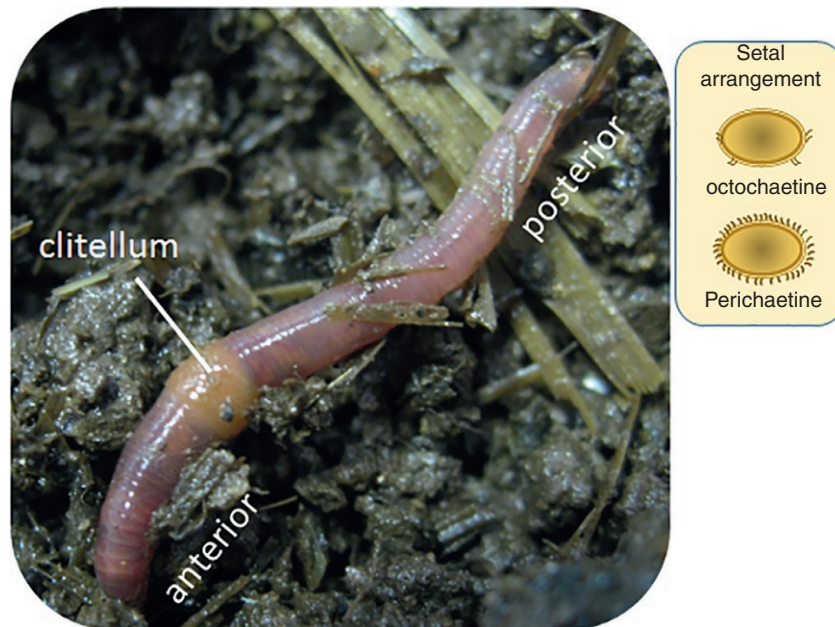


Fig. 1 *Dendrodrilus rubidus*, a small size earthworm living in the upper layers of the soil. The two main types of seate arrangement: (i) octochaetine, eight seate on each segment (four pairs closely or distantly paired) and (ii) perichaetine, numerous setae forming a ring around each segment.

each segment (four pairs) = octochaetine arrangement, but those included in the Megascolecidae family can have more than 30 seate per segment (40–200 around the equator of each segment = perichaetine arrangement) (Fig. 1). Further subdivisions include order Haplotaxida and suborder Lumbricina that contains 38 families, of which Acanthodrilidae, Megascolecidae and Lumbricidae contains the highest number of species. Approximately 7000 of valid species have been described but it is estimated that there might be around 30,000 globally (Orgiazzi et al., 2016).

Their body size ranges from a few cm to 2–3 m in length, with most species fall into the interval of 5 to 15 cm. Therefore, according to the body size classification proposed by Swift et al. (1979), they are included in the macrofauna gross group. This includes those soil organisms whose body width is greater than 2 mm. Although the majority of the species are not very colorful (including some unpigmented worms), some tropical species exhibit bright colorations, stripped patterns and even glow in the dark (bioluminescent). This large variability in pigmentation can be attributed to their ecological behavior and hence, whether they spend most of the time in the deeper layers of the soil (pale in color) or make frequent incursions to the upper soil and litter layers and thereby, become more exposed to UV-light (darker coloration).

Environmental filters shaping their geographical range

Earthworms are distributed worldwide, except for the poles, with the greatest abundance and diversity found in the temperate areas (Phillips et al., 2019). This is due to the strong influence that climatic variables (namely temperature and precipitation regimes) have on their life cycle and activities, with extreme conditions (prolonged heat and drought, but also freezing temperatures) strongly affecting their abundances and the species composition of their communities. Despite these strong environmental filters, some species are known to exhibit high tolerance to adverse environmental conditions through specific physiological and behavioral adaptations. For example, by secreting cryoprotectant substances against freezing temperatures (Jørgensen et al., 2008) or by finding refuge in the deeper soil layers, such as in the case of *Megascolides australis*, a giant earthworm that is able to extend its burrow system deeper than 1.5 m (Kretzschmar and Aries, 1992). That is why changing climatic conditions may facilitate the invasion of earthworms at higher latitudes (Singh et al., 2020), with important implications for nutrient cycling. A study comparing earthworm-invaded and earthworm-free zones, in soil types that are representative of the boreal forest, showed that the invasion of non-native earthworms caused the thinning of the surface organic layer in most cases, and the development of a vermimull humus form (Lejoly et al., 2021).

In addition to climatic factors, soil properties also play an important role in determining community composition and species abundance. In particular, soil pH is a strong determinant of their habitat preferences, with most species benefiting from nearly neutral pH values (5.5–7.0). However, a few acido-tolerant species such as *Dendrobaena octaedra*, *Dendrodrilus rubidus* and *Lumbricus rubellus* can tolerate pH values between 4.0 and 5.5, and are often the main species colonizing forest soils and organic rich environments (e.g. acid grasslands, compost piles, etc.). Furthermore, clay content also seems to affect earthworm densities and

individual growth (Klok et al., 2007), possibly through changes in other soil properties, such as organic matter content and water retention.

Intensive agricultural practices also pose serious risks to earthworm survival. Soil compaction due to the use of heavy machinery, removal of the potential food sources (e.g. crop residues), excessive use of mineral fertilizers as well as agrochemicals, and plowing severely reduced their populations, causing the disappearance of the surface dwellers and larger burrower species and the majority of adult individuals (Briones and Schmidt, 2017). Land degradation through soil pollution (heavy metals, organic pollutants, microcystins, antibiotics) can also be harmful to certain earthworm species, whereas others show a high degree of tolerance by accumulating high amounts of these toxic substances in their tissues (bioaccumulation) or by exhibiting cleansing mechanisms via hydrolysis and oxidation followed by conjugation (Katagi and Ose, 2015).

In view of this, it is clear that current drivers of ecosystem degradation worldwide, particularly climate and land use changes, are expected to have severe effects on earthworm communities across the globe, which could have cascading effects on the other organisms (both above- and below-ground) with which earthworms interact.

Functional diversity: Habitat and dietary requirements

Earthworms can be differentiated according to their external morphology, burrowing behavior and feeding habits. This led Bouché (1977) to classify them into three ecological categories: (i) the small sized and heavily pigmented epigeic worms colonize the litter layer and top soil horizons, where they can consume fresh organic substrates; (ii) the larger anecic are darker in color and build more or less vertical burrows to feed on the soil surface; (iii) the medium sized endogeic worms are usually unpigmented or with a light coloration and construct branching sub- and horizontal burrows in the mineral soil horizons where they feed on more recalcitrant sources (Fig. 2). Despite not being universally accepted (e.g. Lavelle, 1979), this functional classification is nevertheless widely used among earthworm ecologists because it provides more valuable information about their role in ecosystem functioning than species assemblages (Briones, 2014).

Indeed, these three groups show different rates of food consumption, assimilation and egestion, although previous observations show a high variability depending on the substrate quality, earthworm activity and experimental conditions (Curry and Schmidt, 2007). Egestion rates are also variable and a reflection of their feeding preference. For example, the endogeic species *Aporrectodea caliginosa* can egest 349 mg (dry matter) of feces per gram of worm (fresh weight, fwt) daily, whereas *Lumbricus terrestris* (anecic) has a daily egestion rate of 165 mg dry weight (dwt) feces per gram of worm (fwt) (Taylor and Taylor, 2014). The degree to which the ingested material is digested is influenced by their enzymatic capabilities (including a wide array of enzymes such as



Fig. 2 Earthworms can be ecologically differentiated according to their external morphology, burrowing behavior and feeding habits into three groups (Bouché, 1977): (i) epigeic earthworms are small sized and their burrowing and casting activities concentrate in the top soil horizons where they can consume fresh organic substrates; (ii) anecic earthworms are larger in size and build more or less vertical burrows to feed and defecate on the soil surface; (iii) endogeic earthworms are medium sized and construct branching sub- and horizontal burrows in the mineral soil horizons where they feed on more recalcitrant sources.

phenol-oxidase, catalase, protease, cellulase, glycosidase, phosphatase, among others), which ultimately rely on the microorganisms living in their gut or present in the ingested food. In the particular case of those species feeding on polyphenol-rich plant litter, such as in woodland soils, they make use of the unique surface-active metabolites in their gut called 'drilodefensins,' which act as biological surfactants to impede the precipitation of proteins by polyphenols (Liebeke et al., 2015).

These different strategies could represent mechanisms for avoiding competition for the same spatial and trophic resources, but also be a reflection of different adaptations to changes in the soil environment, such as droughts or winter freezing. Because the first protective barrier of earthworms is their tegument, differences in their body wall thickness across these functional groups have been linked to adaptations to specific soil habitats, allowing the use of the earthworm tegument for assigning species to certain ecological groups (Briones and Álvarez-Otero, 2018). According to these authors, not only do anecics have the thickest cuticle and epidermis compared with epigeics and endogeics, but also two types of anecic worms can be identified: (i) 'shelters' that live in permanent burrows (e.g. *Lumbricus terrestris*), and (ii) 'explorers' that build temporary burrows to feed at deeper layers and to find a shelter during those periods when they remain inactive awaiting for better environmental conditions (e.g. *Aporrectodea longa*). In support of this, morphological and functional differences between earthworm burrows produced by species belonging to the same ecological grouping have been observed with the help of X-ray tomography, a technique that has greatly increased our understanding earthworm bioturbation activities (e.g. Capowiez et al., 2015; Le Bayon et al., 2021). Thus, in contrast with Bouché's definitions, these categories are by no means exclusive and more research is needed for a better integration of the existing variability (Bottinelli et al., 2020).

Interactions with other soil organisms and aboveground vegetation

Different earthworm activities influence the diversity and population density of other organisms (both above- and below-ground) either directly or indirectly. For example, soil microorganisms (both fungi and bacteria) are food sources for earthworms; hence, their numbers in soils can be reduced when earthworms are present, but they can also be promoted through a priming effect after gut transit. Consequently, earthworms are able to change the microbial community composition in their area of influence, the so-called "drilosphere." It has been suggested that fungal spores are destroyed or inhibited by the conditions in their digestive system, but bacteria cells are stimulated because they can adapt to the anaerobic environment in the intestine (Don et al., 2008). Accordingly, more Proteobacteria, Actinobacteria, Firmicutes, Acidobacteria, Planctomycetes, Bacteroidetes, Nitrospirae, and Chloroflexi have been observed in the presence of earthworms (Medina-Sauza et al., 2019).

In contrast, although plant roots are also ingested while working through the soil, there is little evidence that live plant material greatly contributes to earthworm nutrition. Instead they seem to be more attracted toward root exudates (Curry and Schmidt, 2007). In relation to this, $^{13}\text{CO}_2$ pulse labelling experiments have shown that although earthworms assimilate recently photosynthesized carbon, it is not their primary food source (Ostle et al., 2007). However, their role as granivores and seedling herbivores has been demonstrated in laboratory experiments (Eisenhauer et al., 2010) indicating that these organisms can indeed digest seeds. This together with the fact they can also passively disperse seeds, spores and viable propagules, it is clear that they also have an important role in shaping plant communities. In addition, earthworms appear to modify plant growth, development and defense through the emission of signal molecules (Puga-Freitas et al., 2012). For example, presence of earthworms can affect plant photosynthetic activity (measured as increased chlorophyll leaf concentration) and lead to higher aboveground biomass and tolerance to parasites (Blouin et al., 2005).

Tunnelling the soil is another way by which earthworms modify the soil habitat. They create porous space that facilitates soil colonization by aerobic microorganisms (specifically Gram-negative bacteria and fungi; Banfield et al., 2017) and provide channels for roots, which can extend their network without investing mechanical energy to penetrate the soil (Canellas et al., 2002).

Finally, undigested materials deposited in their casts and burrow linings provide nutrients for microorganisms and plants. They are rich in nitrogen, either as ammonium that attracts microbes (Lavelle and Martin, 1992) or as a nitrate. Some plant species (e.g., *Brachypodium distachyon*) can take up as much as 20–30% of the nitrogen for their requirements by foraging N from 46 days old casts (Agapit et al., 2018). In particular, anecic species, which form permanent or semi-permanent burrows, tend accumulate organic and inorganic materials together with their casts around the burrow entrance (the so-called "middens") that are also a hotspot for microorganisms and can also function as an "external rumen" that aid them in the breakdown of the plant material.

Role of earthworms in the delivery of ecosystem services

Earthworms are responsible for the delivery of provisioning, supporting, regulating and cultural ecosystem services (Fig. 3). Although this research is biased toward the most economically beneficial services (such as crop productions and soil structure), substantial gaps of knowledge remain (Blouin et al., 2013).

The effects of earthworms on the provisioning of food for human and animal consumption has been intensively investigated both in the field and under laboratory conditions. After reviewing the data derived from staple crops (maize, rice and wheat), pastures, as well as many other food crops across the globe between 1910 and 2013, van Groenningen et al. (2014) found that, on average, earthworms led to a 25% increase in crop yield and a 23% increase in aboveground biomass. Five mechanisms have been identified as potential responsible for these positive effects of earthworms on plant growth (reviewed by Blouin et al., 2013): (i) increased soil

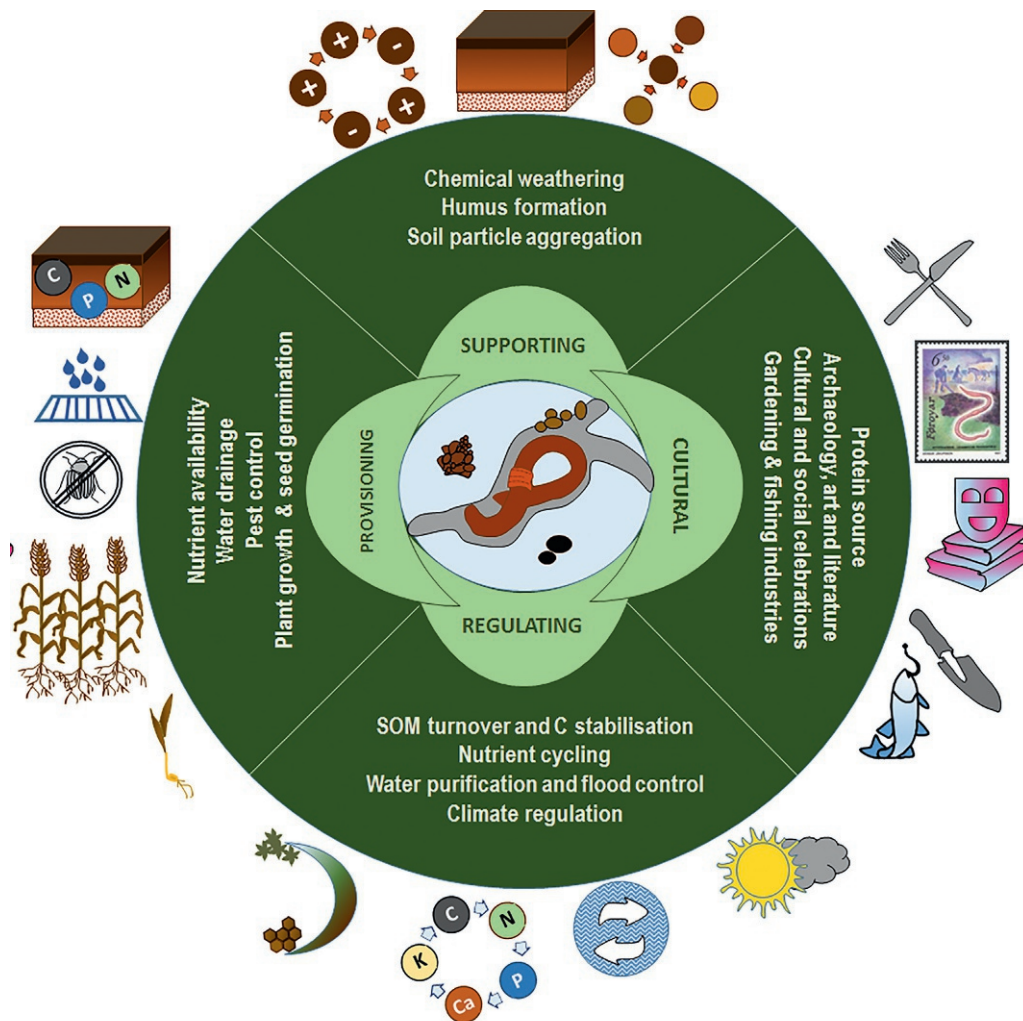


Fig. 3 Ecosystem services provided by earthworms can be both direct (by providing benefits that can be directly used) and indirect (through the interactions with plants and other organisms): (i) provision services (food, fresh water and raw materials); (ii) supporting services (soil formation, habitat sustenance); (iii) regulating services (hydrological services, climate regulation, disease and pest control); (iv) cultural services (natural capital).

nutrient availability due to acceleration of soil organic matter decomposition and mineralization; (ii) improved soil structure that increases water mobility and oxygen concentrations; (iii) biological control of potential pests and parasites; (iv) stimulation of the secretion of plant growth regulators by microorganisms; (v) enhancement of symbiotic relationships. An important component that drives many of these responses are earthworm casts (or vermicasts) that are deposited at the soil surface or on the walls of their burrows (Fig. 2). This cast material produced by earthworms is far more fertile than bulk soil in terms of cation exchange capacity, base saturation, total organic carbon, total phosphorus and total nitrogen concentrations, as well as for plant available pools of N and P (van Groenigen et al., 2019). As a result, earthworm casts not only enhance plant growth but also the germination and establishment of certain plant species (Clause et al., 2016), providing the basis for the profitable business of supplying vermicompost for agriculture and gardening activities worldwide (Fig. 3).

Supporting services by earthworms include soil formation and maintenance. Earthworms are important agents for the breakdown of primary minerals (chemical weathering) and organic materials (humus formation) as well as for pore and aggregate formation (Fig. 3). Their burrows and biopores are highly stable structures thanks to greater compaction at their walls, the fine texture of subsurface casts and the frequent secretion of biopolymers and mucus lining the burrows (Jegou et al., 2001; Ruiz et al., 2015). In addition, earthworm casts cement soil particles together in water-stable aggregates, which can represent 40–60% of the soil weight in the upper 15 cm of soil (Lavelle et al., 2020) and help to rebuild the soil. It has been estimated that, under favorable conditions, in temperate climates, they can deposit 50 t/ha annually of these biogenic aggregates, enough to form a layer from 5 to 25 mm thick layer (Bertrand et al., 2015). Indeed, the term “vermiform” was introduced in the US Soil Taxonomy to define a new type of soils with at least 50% or more of the A horizon and > 25% of the B horizon volume consisting of earthworm cast material and their biostructures (Feller et al., 2003). However, earthworm species that either compact or loosen the soil have also been

identified, with crucial implications for soil bulk density, porosity, water infiltration and soil erosion. Overall, more positive effects on soil formation and improvement of soil structure than negative ones have been described in the literature (Blouin et al., 2013).

Bioturbating activities by earthworms, including burrowing, litter incorporation and decomposition, soil mixing and aggregate formation, directly affect *regulating services* such as SOM turnover and C storage, nutrient cycling, water purification and flood control, and climate regulation (Fig. 3). While the accelerating effects of earthworms on organic matter decomposition, N mineralization and hydraulic properties appear to be undisputed (Blouin et al., 2013; Briones, 2021), their effects on C sequestration and greenhouse gas mitigation remain unresolved. Part of the problem is the difficulties in obtaining reliable measurements of the earthworm contribution to the dynamic transformations (mobilization vs. stabilization) of the basic nutrient elements (C and N) and their separation from other interactive factors (plant activities, interactions with microorganisms). Many studies have shown that earthworms induce C stabilization, for example by favoring the decomposition of leaf- over root-derived SOM, which contributes to the gradual accumulation of the latter fraction in soil (Angst et al., 2020). They also contribute to the formation of microaggregates and the integration of these into macroaggregates through stimulation of the microbial glue between the mineral particles and hence, not only they increase C stability but also its resilience against environmental perturbations (Angst et al., 2019). In contrast, several studies have suggested that earthworm activities increase greenhouse gas (GHG) emissions. Thus, in controlled experiments without plants, earthworms increased the emissions of N₂O by 42% and CO₂ by 33% (Lubbers et al., 2013), a finding that has been hotly disputed (e.g. Zhang et al., 2013). On the other hand, endogeic worms consistently reduced the impact of periodic flooding on CH₄ emissions from riparian soils (Kernecker et al., 2015). A more recent review concluded that physicochemical properties of soil, residue incorporation, species feeding strategies and the duration of the experiments modulate earthworm impacts on GHG emissions (Singh and Singh, 2019). Clearly, more field and long-term evidence is required to establish the direction and magnitude of these effects and to enable realistic extrapolations of the effects of earthworms on GHG balance.

Another important, although far less studied, regulating service provided by earthworms is the control on the populations of other organisms. One of the key aspects that have led to consider earthworms as “ecosystem engineers” is the fact that their biogenic structures (aggregates, burrows) serve as habitats for different species other than themselves (Lavelle et al., 1997). For example, the large amounts of bacterial and fungal decomposers in earthworm biopores are a clear indication that they represent hotspots of microbial activities (Banfield et al., 2017). In an agronomic context, the earthworm effects on pest suppression were confirmed by substantial reductions in the number of infested plants by parasitic nematodes (Blouin et al., 2005) and oomycetes (Lagerlöf et al., 2020), as well as enhanced resistance against herbivores (Xiao et al., 2018). This reduction in pest susceptibility and disease symptoms were attributed to both direct (consumption of the pathogenic organisms) and indirect effects (such as soil disturbance and improved plant fitness). For example, it has been shown that the presence of earthworms can increase the production of chemical defenses by 31% when plants are attacked by cell-feeders (thrips), resulting in an 81% increase in resistance against these plant-sucking insects (Xiao et al., 2018).

Although far less studied, *cultural services* provided by earthworms have been long acknowledged and include a variety of benefits to humans (Fig. 3), from a source of protein for some tribal communities in the Amazon (Moreno and Paoletti, 2004) to medicinal products (Cooper et al., 2012). Their burial activities can be seen either as positive (i.e. preserving archeological remains; Canti, 2003) as well as negative (i.e. destruction of heritage sites; Motiejūnaitė et al., 2019), but exploitation for stratigraphic interpretation is only minor (Canti, 2003). In addition, earthworms are commonly represented in art, literature, stamps and crafts (Motiejūnaitė et al., 2019) and a theme in cultural and social celebrations such as the “Worm Charming” events in some localities in England. Much more appreciated is their value as gardeners’ friends and as fish bait, and the commercial value of growing earthworms to be sold for vermicomposting and as bait for fish stores, animal growers (frogs, poultry, game birds) has become a very profitable business in several parts of the world. For example, in North America one pound of *Lumbricus rubellus* sells for around \$15–30 on the commercial market depending on the supplier, and in Ontario (Canada) the supply of *L. terrestris* has become the most lucrative business in the region with farmers leasing land to worm-pickers for over CA\$1000 per year (Steckley, 2020).

Conclusion

The crucial role of earthworms as “nature’s plow,” increasing soil fertility and helping plants to grow, has been recognized from the time of Charles Darwin. However, despite their quantifiable benefits for soil functioning, no specific regulation has been formulated to protect these essential ecosystem engineers (either nationally or internationally) and the potential threatened status of most of the earthworm species is currently unknown. More research is needed to fill the current knowledge gaps on their physiological responses to environmental threats, habitat requirements and restrictions, variability in feeding strategies and their complex interactions with their soil physical and chemical environment and with other organisms to ensure their protection and the continuously delivery of the ecosystem services they provide.

References

- Agapit C, Gigon A, Puga-Freitas R, Zeller B, and Blouin M (2018) Plant-earthworm interactions: Influence of age and proportion of casts in the soil on plant growth, morphology and nitrogen uptake. *Plant and Soil* 424: 49–61. <https://doi.org/10.1007/s11104-017-3544-y>.
- Angst G, Mueller CW, Prater I, Angst S, Peterse F, and Nierop KGJ (2019) Earthworms act as biochemical reactors to convert labile plant compounds into stabilized soil microbial necromass. *Communications Biology* 2: 1–7. <https://doi.org/10.1038/s42003-019-0684-z>.

- Angst G, Angst Š, Frouz J, Peterse F, and Nierop KGJ (2020) Preferential degradation of leaf- vs. root-derived organic carbon in earthworm-affected soil. *Geoderma* 372: 114391. <https://doi.org/10.1016/j.geoderma.2020.114391>.
- Banfield CC, Dippold MA, Pausch J, Hoang DTT, and Kuzakov Y (2017) Biopore history determines the microbial community composition in subsoil hotspots. *Biology and Fertility of Soils* 53: 573–588. <https://doi.org/10.1007/s00374-017-1201-5>.
- Bertrand M, Barot S, Blouin M, Whalen J, de Oliveira T, and Roger-Estrade J (2015) Earthworm services for cropping systems. A review. *Agronomy for Sustainable Development* 35: 553–567. <https://doi.org/10.1007/s13593-014-0269-7>.
- Blouin M, Zuily-Fodil Y, Pham-Thi A-T, Laffray D, Reversat G, Pando A, Tondoh T, and Lavelle P (2005) Belowground organism activities affect plant aboveground phenotype, inducing plant tolerance to parasites. *Ecology Letters* 8: 202–208. <https://doi.org/10.1111/j.1461-0248.2004.00711.x>.
- Blouin M, Hodson ME, Delgado EA, Baker G, Brussaard L, Butt KR, Dai J, Dendooven L, Peres G, Tondoh JE, Cluzeau D, and Brun J-J (2013) A review of earthworm impact on soil function and ecosystem services. *European Journal of Soil Science* 64: 161–182. <https://doi.org/10.1111/ejss.12025>.
- Bottinelli N, Hedde M, Jouquet P, and Capowiez Y (2020) An explicit definition of earthworm ecological categories – Marcel Bouché's triangle revisited. *Geoderma* 372: 114361. <https://doi.org/10.1016/j.geoderma.2020.114361>.
- Bouché MB (1977) Stratégies lombriciennes. In: Lohm U and Persson T (eds.) *Soil Organisms as Components of Ecosystems. Ecology Bulletin*, vol. 25, pp. 122–132. Stockholm: Oikos Editorial Office.
- Briones MJJ (2014) Soil fauna and soil functions: A jigsaw puzzle. *Frontiers in Environmental Science* 2: art7. <https://doi.org/10.3389/fenvs.2014.00007>.
- Briones MJJ (2021) Towards agroecological intensification: Relying on soil biodiversity for delivery of ecosystem services in agricultural soils. In: Soto Gómez D, Fernández Calviño D. & Shanskiy M. (Eds.). *Interactions Between Agricultural Management and Soil Biodiversity: An Overview of Current Knowledge*. 2nd edn., pp. 8–25. Zenodo. <https://doi.org/10.5281/zenodo.4742198>.
- Briones MJJ and Álvarez-Otero R (2018) Body wall thickness as a potential functional trait for assigning earthworm species to ecological categories. *Pedobiologia* 67: 26–34. <https://doi.org/10.1016/j.pedobi.2018.02.001>.
- Briones MJJ and Schmidt O (2017) Conventional tillage decreases the abundance and biomass of earthworms and alters their community structure in a global meta-analysis. *Global Change Biology* 23(10): 4396–4419. <https://doi.org/10.1111/gcb.13744>.
- Canellas LP, Olivares FL, Okorokova-Facanha AL, and Facanha AR (2002) Humic acids isolated from earthworm compost enhance root elongation, lateral root emergence, and plasma membrane H⁺-ATPase activity in maize roots. *Plant Physiology* 130: 1951–1957. <https://doi.org/10.1104/pp.007088>.
- Canti MG (2003) Earthworm activity and archaeological stratigraphy: A review of products and processes. *Journal of Archaeological Science* 30: 135–148. <https://doi.org/10.1006/jasc.2001.0770>.
- Capowiez Y, Bottinelli N, Sammarino S, Michel E, and Jouquet P (2015) Morphological and functional characterisation of the burrow systems of six earthworm species (Lumbricidae). *Biology and Fertility of Soils* 51: 869–877. <https://doi.org/10.1007/s00374-015-1036-x>.
- Clause J, Barot S, and Forey E (2016) Earthworms promote greater richness and abundance in the emergence of plant species across a grassland-forest ecotone. *Journal of Plant Ecology* 9: 703–711. <https://doi.org/10.1093/jpe/rtw008>.
- Cooper EL, Balamurugan M, Huang C-H, Tsao CR, Heredia J, Tommaseo-Ponzetta M, and Paoletti MG (2012) Earthworms dilong: Ancient, inexpensive, noncontroversial models may help clarify approaches to integrated medicine emphasizing neuroimmune systems. *Evidence-based Complementary and Alternative Medicine* 2012: 164152. <https://doi.org/10.1155/2012/164152>.
- Curry JP and Schmidt O (2007) The feeding ecology of earthworms – A review. *Pedobiologia* 50: 463–477. <https://doi.org/10.1016/j.pedobi.2006.09.001>.
- Don A, Steinberg B, Schöningh I, Pritsch K, Joschko M, Gleixner G, and Schulze E-D (2008) Organic carbon sequestration in earthworm burrows. *Soil Biology and Biochemistry* 40: 1803–1812. <https://doi.org/10.1016/j.soilbio.2008.03.003>.
- Eisenhauer N, Butenschoen O, Radsick S, and Scheu S (2010) Earthworms as seedling predators: Importance of seeds and seedlings for earthworm nutrition. *Soil Biology and Biochemistry* 42: 1245–1252. <https://doi.org/10.1016/j.soilbio.2010.04.012>.
- Feller C, Brown GG, Blanchart E, Deleporte P, and Chernyanski SS (2003) Charles Darwin, earthworms and the natural sciences: Various lessons from past to future. *Agriculture, Ecosystems and Environment* 99: 29–49. [https://doi.org/10.1016/S0167-8809\(03\)00143-9](https://doi.org/10.1016/S0167-8809(03)00143-9).
- Jegou D, Schrader S, Diestel H, and Cluzeau D (2001) Morphological, physical and biochemical characteristics of burrow walls formed by earthworms. *Applied Soil Ecology* 17: 165–174. [https://doi.org/10.1016/S0929-1393\(00\)00136-0](https://doi.org/10.1016/S0929-1393(00)00136-0).
- Jørgensen S, Overgaard J, Holmstrup M, and Westh P (2008) Cryoprotectants are metabolic fuels during long term frost exposure in the earthworm *Dendrobaena octaedra*. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology* 150: S159. <https://doi.org/10.1016/j.cbpa.2008.04.414>.
- Katagi T and Ose K (2015) Toxicity, bioaccumulation and metabolism of pesticides in the earthworm. *Journal of Pesticide Science* 40: 69–81. <https://doi.org/10.1584/jpestics.D15-003>.
- Kernecker M, Whalen JK, and Bradley RL (2015) Endogeic earthworms lower net methane production in saturated riparian soils. *Biology and Fertility of Soils* 51: 271–275. <https://doi.org/10.1007/s00374-014-0965-0>.
- Klok C, Faber J, Heijmans G, Bodt J, and van der Hout A (2007) Influence of clay content and acidity of soil on development of the earthworm *Lumbricus rubellus* and its population level consequences. *Biology and Fertility of Soils* 43: 549–556. <https://doi.org/10.1007/s00374-006-0135-0>.
- Kretzschmar A and Aries F (1992) An analysis of the structure of the burrow system of the giant gippsland earthworm *Megascolides australis* McCoy, 1878 using 3D-images. *Soil Biology and Biochemistry* 24: 1583–1586. [https://doi.org/10.1016/0038-0717\(92\)90154-P](https://doi.org/10.1016/0038-0717(92)90154-P).
- Lagerlöf J, Ayuke F, Heyman F, and Meijer J (2020) Effects of biocontrol bacteria and earthworms on *Aphanomyces euteiches* root-rot and growth of peas (*Pisum sativum*) studied in a pot experiment. *Acta Agriculturae Scandinavica Section B Soil and Plant Science* 70: 427–436. <https://doi.org/10.1080/09064710.2020.1761995>.
- Lavelle P (1979) Relations entre types écologiques et profils démographiques chez les vers de terre de la savanne de Lamto (Coted'Ivoire). *Revue d'Écologie et de Biologie du Sol* 16: 85–101.
- Lavelle P and Martin A (1992) Small-scale and large-scale effects of endogeic earthworms on soil organic matter dynamics in soils of the humid tropics. *Soil Biology and Biochemistry* 24: 1491–1498. [https://doi.org/10.1016/0038-0717\(92\)90138-N](https://doi.org/10.1016/0038-0717(92)90138-N).
- Lavelle P, Bignell D, Lepage M, Wolters V, Roger P, Ineson P, Heal OW, and Dhillon SP (1997) Soil function in a changing world: The role of invertebrate ecosystem engineers. *European Journal of Soil Biology* 33: 159–193.
- Lavelle P, Spain A, Fonte S, Bedano JC, and Blanchart E (2020) Soil aggregation, ecosystem engineers and the C cycle. *Acta Oecologica* 105: 103561. <https://doi.org/10.1016/j.actao.2020.103561>.
- Le Bayon R, Guenat C, Schlaepfer R, Fischer F, Luiset A, Schomburg A, and Turberg P (2021) Use of X-ray microcomputed tomography for characterizing earthworm-derived belowground soil aggregates. *European Journal of Soil Science* 72: 1113–1127. <https://doi.org/10.1111/ejss.12950>.
- Lejoly J, Quideau S, and Laganier J (2021) Invasive earthworms affect soil morphological features and carbon stocks in boreal forests. *Geoderma* 404: 115262. <https://doi.org/10.1016/j.geoderma.2021.115262>.
- Liebekke M, Strittmatter N, Fearn S, Morgan AJ, Kille P, Fuchser J, Wallis D, Palchykov RJ, Lahive E, Spurgeon DJ, McPhail D, Takáts Z, and Bundy JG (2015) Unique metabolites protect earthworms against plant polyphenols. *Nature Communications* 6: 7869. <https://doi.org/10.1038/ncomms8869>.
- Lubbers JM, van Groenigen KJ, Steven JF, Six J, Brussaard L, and van Groenigen JW (2013) Greenhouse-gas emissions from soils increased by earthworms. *Nature Climate Change* 3: 187–194. <https://doi.org/10.1038/nclimate1692>.
- Medina-Sauza RM, Álvarez-Jiménez M, Delhal A, Reverchon F, Blouin M, Guerrero-Analco JA, Cerdán CR, Guevara R, Villain L, and Barois I (2019) Earthworms building up soil microbiota, a review. *Frontiers in Environmental Science* 7: 81. <https://doi.org/10.3389/fenvs.2019.00081>.

- Moreno AG and Paoletti MG (2004) *Andiorrhinus (Andiorrhinus) kuru* sp. nov. (Oligochaeta: Glossoscolecidae), a giant earthworm as food resource for Makiritare indians of the alto rio Padamo, Amazonas, Venezuela. *Canadian Journal of Zoology* 82: 1000–1004. <https://doi.org/10.1139/z04-056>.
- Motiejūnaitė J, Børja I, Ostonen I, Bakker MR, Bjarnadóttir B, Brunner I, Iršėnaitė R, Mrak T, Oddsdóttir ES, and Lehto T (2019) Cultural ecosystem services provided by the biodiversity of forest soils: A European review. *Geoderma* 343: 19–30. <https://doi.org/10.1016/j.geoderma.2019.02.025>.
- Orgiazzi A, Bardgett RD, Barrios E, Behan-Pelletier V, Briones MJ, Chotte J-L, De Deyn GB, Eggleton P, Fierer N, Fraser T, Hedlund K, Jeffery S, Johnson NC, Jones A, Kandeler E, Kaneko N, Lavelle P, Lemanceau P, Miko L, Montanarella L, Moreira FMS, Ramirez KS, Scheu S, Singh BK, Six J, van der Putten WH, and Wall DH (eds.) (2016) *Global Soil Biodiversity Atlas*. Luxembourg: European Commission, Publications Office of the European Union.
- Ostle N, Briones MJ, Ineson P, Cole L, Staddon P, and Sleep D (2007) Isotopic detection of recent photosynthate carbon flow into grassland rhizosphere fauna. *Soil Biology and Biochemistry* 39: 768–777. <https://doi.org/10.1016/j.soilbio.2006.09.025>.
- Phillips HRP, Guerra CA, Bartz MLC, Briones MJ, Brown GG, Ferlian O, Gongalsky KB, Krebs J, Orgiazzi A, Schwarz B, Bach EM, Bennett J, Brose U, Decaëns T, De Vries FT, König-Ries B, Loreau M, Mathieu J, Mulder C, van der Putten WH, Ramirez KS, Rillig MC, Russell D, Rutgers M, Thakur MP, Wall DH, Wardle D, Cameron EK, and Eisenhauer N (2019) Global distribution of earthworm diversity. *Science* 366: 480–485. <https://doi.org/10.1126/science.aax4851>.
- Puga-Freitas R, Barot S, Taconnat L, Renou J-P, and Blouin M (2012) Signal molecules mediate the impact of the earthworm *Aporrectodea caliginosa* on growth, development and defence of the plant *Arabidopsis thaliana*. *PLoS One* 7: e49504. <https://doi.org/10.1371/journal.pone.0049504>.
- Ruiz S, Or D, and Schymanski SJ (2015) Soil penetration by earthworms and plant roots—Mechanical energetics of bioturbation of compacted soils. *PLoS One* 10: e0128914. <https://doi.org/10.1371/journal.pone.0128914>.
- Singh A and Singh GS (2019) Is earthworm a protagonist or an antagonist in greenhouse gas (GHG) emissions from the soil? *International Journal of Environmental Science and Technology* 16: 1145–1158. <https://doi.org/10.1007/s13762-018-1922-5>.
- Singh J, Schädler M, Demetrio W, Brown GG, and Eisenhauer N (2020) Climate change effects on earthworms - A review. *Soil Organisms* 91: 113–137. <https://doi.org/10.25674/so91iss3pp114>.
- Steckley J (2020) Cash cropping worms: How the *Lumbricus terrestris* bait worm market operates in Ontario, Canada. *Geoderma* 363: 114128. <https://doi.org/10.1016/j.geoderma.2019.114128>.
- Swift MJ, Heal OW, and Anderson JM (1979) *Decomposition in Terrestrial Ecosystems*. Oxford: Blackwell Science.
- Taylor A and Taylor AFS (2014) Assessing daily egestion rates in earthworms: using fungal spores as a natural soil marker to estimate gut transit time. *Biology and Fertility of Soils* 50: 179–183. <https://doi.org/10.1007/s00374-013-0823-5>.
- van Groenigen J, Lubbers I, Vos H, Brown GG, De Deyn GB, and van Groenigen KJ (2014) Earthworms increase plant production: A meta-analysis. *Scientific Reports* 4: 6365. <https://doi.org/10.1038/srep06365>.
- Van Groenigen JW, van Groenigen KJ, Koopmans GF, Stokkermans L, Vos HMJ, and Lubbers IM (2019) How fertile are earthworm casts? A meta-analysis. *Geoderma* 338: 525–535. <https://doi.org/10.1016/j.geoderma.2018.11.001>.
- Xiao Z, Wang X, Koricheva J, Kergunteuil A, Le Bayon R-C, Liu M, Hu F, and Rasmann S (2018) Earthworms affect plant growth and resistance against herbivores: A meta-analysis. *Functional Ecology* 32: 150–160. <https://doi.org/10.1111/1365-2435.12969>.
- Zhang W, Hendrix PF, Dame LE, Burke RA, Wu J, Neher DA, Li J, Shao Y, and Fu S (2013) Earthworms facilitate carbon sequestration through unequal amplification of carbon stabilization compared with mineralization. *Nature Communications* 4: 2576. <https://doi.org/10.1038/ncomms3576>.

Soil biofilms: Evolving concepts and ecological functions

Yichao Wu^a, Marc Redmile-Gordon^b, Peng Cai^a, Chunhui Gao^a, and Qiaoyun Huang^a, ^aState Key Laboratory of Agricultural Microbiology, College of Resources and Environment, Huazhong Agricultural University, Wuhan, China; ^bDepartment of Environmental Horticulture, Royal Horticultural Society, Wisley, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction	120
The evolving concepts of soil biofilm	120
The EPS production in soil	123
Ecological functions of soil biofilms	123
Future prospects	124
Conclusion	125
Reference	125

Abstract

Biofilms are recognized as essential habitats for the majority of bacterial and archaeal life on Earth. These microbial assemblages are hotspots for trophic, genetic, and inter- and intra-species communication. The sessile lifestyle confers survival advantages to biofilm inhabitants under harsh environmental conditions. As one of the major microbial biomes, soil provides a huge internal surface-area for biofilm formation. However, due to the opaque nature and spatio-temporal heterogeneity of soil system, the distribution of biofilms in this medium remains obscure. In this chapter, we present fundamental criteria and context for understanding biofilms and their ecological importance in soil.

Key points

- Review of the current state of knowledge regarding soil biofilm concepts.
- Summary of the ecological functions of biofilms.
- Discussion of the challenges and future opportunities for soil biofilm research.

Introduction

It is widely accepted that most naturally occurring microbes exist in a biofilm at some stages (if not all) of their lifecycle. The ubiquity of biofilms in natural, industrial and clinical environments was initially described by Costerton et al. (1987). It has been estimated that 20–80% of cells on Earth currently inhabit biofilms. More than 95% of bacteria and archaea on Earth live in the ‘big five’ habitats: deep oceanic subsurface (33.3%), upper oceanic sediment (4.2%), deep continental subsurface (25.0%), soil (25.0%) and oceans (8.3%). Microbial life tends to attach to solid surfaces, where it subsequently begins the processes of biofilm formation (Flemming and Wurtz, 2019). As one of the ‘big five’ habitats, soil offers a huge internal surface area and hosts approximately one quarter of all microorganisms on Earth. Microscopic observations show that most soil microbes are attached on soil mineral surfaces and are immersed in thin films (Fig. 1). Microbial cells are known to attach to surfaces in soil pores and develop multicellular aggregates and many believe that biofilms are the dominant growth form in soil (Flemming and Wurtz, 2019). Meanwhile, recent experimental evidence revealed that soil biofilm formation is a dynamic strategy for soil microbes which responds to labile carbon input (Redmile-Gordon et al., 2020; Redmile-Gordon et al., 2015; Wu et al., 2019). This concept will be discussed in detail in the following sections.

The evolving concepts of soil biofilm

Given the high density of attached microbial cells in soil, the term ‘soil biofilm’ was generally accepted for describing the lifestyle of soil microorganisms within their associated polymeric matrices. In comparison to other natural or industrial habitats, the volume of space available for biofilm development in soil is restricted by mineral fragments and other constituents. Uncertainties in the spatial distributions of biofilm-dwelling microorganisms have ignited a debate over their existence (Bar-On and Milo, 2019; Baveye, 2020). However, microscopic observations of soil show that multicellular microbial aggregates exist in soil, but as patchy biofilms (Fig. 1) rather than the thick and uniform biofilms that are imagined in wastewater treatment and biofouling processes (Eickhorst and Tippkötter, 2008; Flemming et al., 2021; Nunan et al., 2003; Watteau and Villemin, 2018).

The debates over biofilms require a clarification on related definitions. In the terminology of biomaterials, biofilms are aggregates of microorganisms where cells are embedded in a matrix of extracellular polymeric substances (EPS)

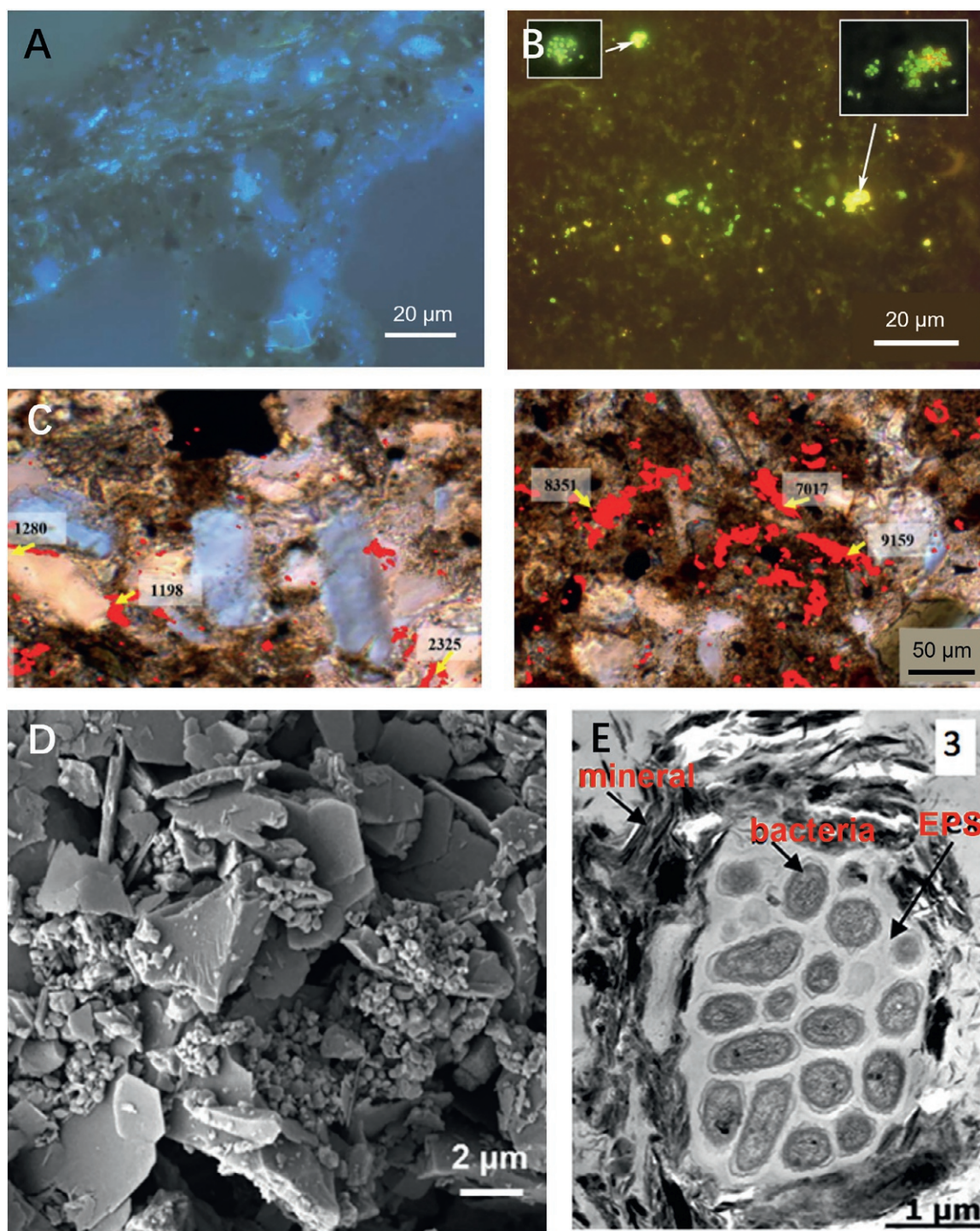


Fig. 1 Microscopy images of soil biofilms. Calcofluor White-stained biofilms in a surface sandy silt loam (A) (Nunan et al., 2003). Biofilm colonies were heterogeneously distributed and separated by uncolonized regions. FISH-stained biofilms in an undisturbed sandy soil (B) (Eickhorst and Tippkötter, 2008). The biofilm cells were visualized with the eubacterial probe Eub338-FITC. The distribution of biofilms in mineral grains (left panel) and soil aggregates (right panel) (C) (Gutiérrez Castorena et al., 2016). The numbers represent the bacterial population size of individual biofilms. Biofilms in an artificial soil (D) (Wu et al., 2019). Soil biofilm formation induced the development of micro-meter sized aggregates. Biofilm cells embedded in self-produced EPS matrix in a calcic soil (E) (Watteau and Villemin, 2018). (A) Nunan N, Wu K, Young IM, Crawford JW, Ritz K (2003) Spatial distribution of bacterial communities and their relationships with the micro-architecture of soil. *FEMS Microbiology Ecology* 44(2): 203–215.; (B) Eickhorst T, Tippkötter R (2008) Detection of microorganisms in undisturbed soil by combining fluorescence in situ hybridization (FISH) and micropedological methods. *Soil Biology and Biochemistry* 40(6): 1284–1293.; (C) Gutiérrez Castorena EV, Gutiérrez-Castorena MdC, González Vargas T, Cajuste Bontemps L, Delgadillo Martínez J, Suástegui Méndez E, et al. (2016) Micromapping of microbial hotspots and biofilms from different crops using digital image mosaics of soil thin sections. *Geoderma* 279: 11–21.; (D) Wu Y, Cai P, Jing X, Niu X, Ji D, Ashry NM, et al. (2019) Soil biofilm formation enhances microbial community diversity and metabolic activity. *Environment International* 132: 105116.; (E) Watteau F, Villemin G (2018) Soil microstructures examined through transmission electron microscopy reveal soil-microorganisms interactions. *Frontiers in Environmental Science*.

(Vert et al., 2012). Based on this, and also the definitions from other disciplines, soil biofilms were defined as *aggregates of microbes connected by EPS* (Redmile-Gordon et al., 2014). However, there are no existing criteria to define a soil biofilm either for the minimum number of cells, or the size of the EPS matrix.

Interestingly, a diagnostic criterion of clinical biofilms was established in the context of human disease. Apart from the multicellular structure, it is based mainly on enhanced resistance to antibiotics, and persistence against host defenses (Hall-Stoodley and Stoodley, 2009). This raises the question as to whether small biofilms in soil retain the innate resistance seen in mature biofilms. The reduced susceptibility of biofilm cells to antimicrobial agents and environmental stresses has been attributed to physically restricted diffusion rates and chemical interactions between solutes and EPS. Although mature, thick biofilms were initially believed to be a prerequisite for biofilm resistance, thin biofilms (5–10 μm) also exhibit altered physiology, including increased microbial resistance to heavy metals (Teitzel and Parsek, 2003). Therefore, soil biofilms of similar size impart resistance to antimicrobial stresses and exhibit biofilm emergent properties, further differentiating them functionally from their free-living counterparts. Therefore, it may be useful to think of biofilms as *aggregated microbial cells embedded in a matrix of EPS which exhibit emergent properties absent in free-living cells* (Flemming et al., 2021). This view should help facilitate much needed progress in this area.

Nonetheless, biofilms remain a fundamental form of microbial life on Earth with phenotypic expression of prokaryotes being highly dependent on the stage of life-cycle, e.g., biofilm forming vs. free-living (Bar-On and Milo, 2019). The triggers or 'cues' for changes in life-cycle are largely environmental, but include bacterial density, quality of carbon in the soil or growing medium and environmental stresses (Redmile-Gordon et al., 2015). Knowing the proportion of soil microbial biomass that exists in a biofilm phase of growth at any given time, will help to improve predictions related to soil biogeochemistry and the soil physical properties (Cai et al., 2019).

The concept of 'microbial biomass carbon' specifically relates to cell carbon content, and accordingly is measured as 'biomass carbon'. These cells can be either (i) free-living (planktonic or sessile), or (ii) members of a biofilm. If one adopts a working definition of *aggregated microbial cells in a matrix of EPS* during visual investigations of bulk soil, a large proportion of the soil microbial biomass may exist independently of biofilms. Although most soil bacterial isolates form biofilms in rich media, only about 0.1–2% of the microbial biomass in non-rhizosphere soils has sufficient carbon to grow (Blagodatskaya and Kuzyakov, 2013). Micrographs reveal that the spatio-temporal distribution of soil microorganisms is highly heterogeneous. Only 0.000001% of internal soil surfaces are colonized by microbes (Young and Crawford, 2004), while about 1–0.0001% cells have a neighbor within an 'interaction distance' of approximately 20 μm (Gutiérrez Castorena et al., 2016; Nunan et al., 2003). Therefore, in most bulk soils, microorganisms occupy a relatively vast expanse, with the fraction residing in biofilms potentially being very small and restricted to specific carbon-rich interfaces (Fig. 2).

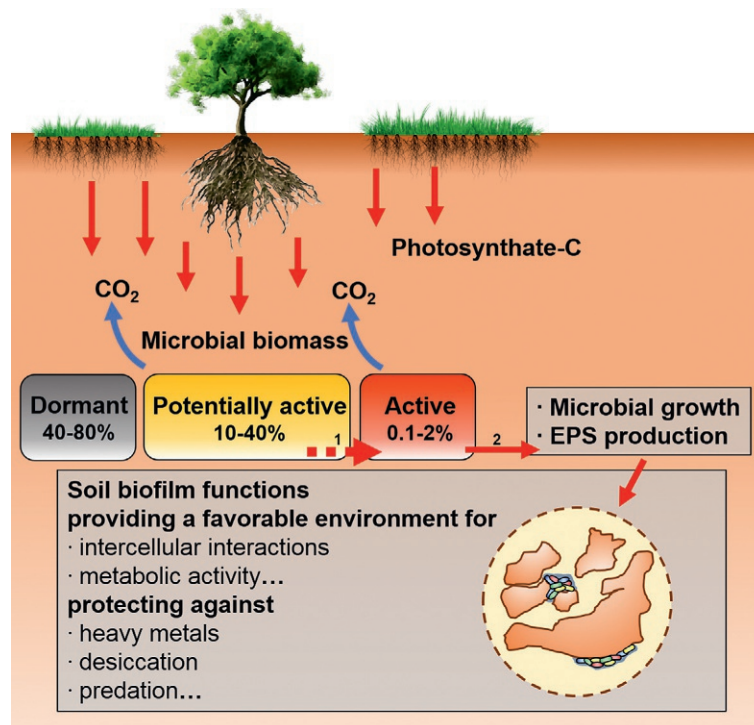


Fig. 2 Schematic diagram of increases in active microbial biomass with the capacity to form biofilms in response to inputs of photosynthate-C. Dormant microorganisms can be activated by an influx of labile substrates (process 1). Subsequently active microbes assimilate labile C into microbial biomass and extracellular polymeric substances (EPS) (process 2). Continued microbial growth and EPS production forms a biofilm. The micrometer sized films exhibit an extended phenotype, providing a favorable environment for intercellular interactions and metabolic activity, while protecting against heavy metals, desiccation, and predation.

The EPS production in soil

As EPS is a crucial component of biofilms, its production is a prerequisite for their formation in soil. Soil microorganisms process organic matter and synthesize metabolites, which can be stored extracellularly as EPS. When nutrients are depleted, microorganisms can selectively use certain components in EPS as nutrient source to survive. For example, in the absence of carbon sources, cyanobacteria secrete polypeptidase and cellulase to hydrolyze and assimilate the self-produced organic carbon (Stuart et al., 2016). Soil EPS can also serve as a communal growth substrate for microorganisms in soil. *Actinobacteria* and *Proteobacteria* widely distributed in soil can use homogeneous polysaccharides in EPS, such as cellulose, as carbon sources, while microorganisms such as *Planctomycetes* and *Armatimonadetes* selectively decompose and utilize the heteropolysaccharides (Wang et al., 2015). Moreover, microorganisms can also obtain inorganic nutrients such as nitrogen and phosphorus by decomposing the proteins and DNA in EPS (Bohórquez et al., 2017).

The secreted EPS is mainly distributed on the surface of cells and minerals, providing a feasible living environment for soil microorganisms (Buckeridge et al., 2022; Zhang et al., 2021). The association of EPS with soil minerals can significantly reduce its bioavailability via various physical and chemical protection mechanisms (Mikutta et al., 2011). The protective effect makes it possible for EPS to persist in the soil and eventually form stable soil organic matter to expand the soil carbon pool (Buckeridge et al., 2022). Thus, the synthesis and decomposition of EPS essentially contribute to the turnover of soil organic carbon. However, most investigations on soil EPS are based on the cultivation of pure bacteria in the laboratory, and little is known about the conversion efficiency, EPS products and corresponding degradation process in the soil environment. More efforts need to be devoted to elucidate the fate of these exopolymers in soil.

The production of the EPS matrix and maintenance of biofilm structural integrity requires considerable investments of energy (Flemming et al., 2016). EPS typically accounts only for 2–15% of soil organic matter (Redmile-Gordon et al., 2020), but under copiotrophic conditions, soil microorganisms can store additional carbon as EPS (Redmile-Gordon et al., 2015; Wu et al., 2019). Supporting this, EPS are transient bonding agents and are produced rapidly in response to land-use change and correlated with soil structural improvements (Redmile-Gordon et al., 2020). Microorganisms assimilate nutrients and incorporate part of soil organic matter as extracellular polymers (Liang et al., 2017). The EPS synthetic efficiency is influenced by microbial species, growth status, nutrient types and other biotic and abiotic factors. Analyses of the EPS synthetic efficiency of 38 soil isolates showed that most had a higher synthetic efficiency for EPS than for the synthesis of cellular components (Fernandes Jr. et al., 2011). The synthesis efficiency of EPS in natural soil ($\mu\text{g EPS nmol}^{-1} \text{ATP}$) was found to be substantially improved when inorganic N was not available in excess (Redmile-Gordon et al., 2020).

Different organic carbon sources affect EPS synthetic pathways according to limits of the metabolic production of exopolymer precursors (Flemming et al., 2016). In addition, the production of EPS also depends on other aspects of the soil environment, including temperature, pH and the availability of nutrients such as carbon, nitrogen and phosphorus. For example, compared with low molecular weight organic acids (formic acid, acetic acid), microorganisms can secrete more exopolysaccharides when using glucose as a carbon source (Li and Yang, 2007). The substrate composition was also found to influence soil EPS production. When the substrate C:N ratio decreased from 20 to 4, the carbohydrate content in loosely bound EPS decreased by 200%, while the extracellular protein content increased by 300% (Ye et al., 2011). It suggested that a low C:N ratio enhanced the synthesis and secretion of extracellular protein (Ye et al., 2011). Conversely, in soil containing a substantial organic matter pool, an increased C:N ratio of inputs, caused both the production of extracellular polysaccharide and protein to be significantly enhanced (Redmile-Gordon et al., 2015). Therefore, EPS production would seem to interact with organic matter availability but is a microbial survival strategy that responds to the stoichiometric ratio of labile inputs (Redmile-Gordon et al., 2015). With the same C:N ratio, sufficient nutrient supply was the prerequisite for the formation of soil biofilm (Wu et al., 2019). A 10-times increase in carbon and nitrogen contents resulted in a decreased microbial biomass growth and a 14.2-fold increase in EPS production, which triggered the formation of soil microaggregates (Wu et al., 2019).

Intrinsically, EPS is a key element in soil biofilms. High organic carbon input stimulates soil EPS production. The increased substrate availability activates dormant cells to initiate biofilm formation, and habitats such as the rhizosphere and detritosphere are hotspots for the formation of dense microbial assemblages (Gutiérrez Castorena et al., 2016; Nunan et al., 2003). Accordingly, land-uses with continual inputs of labile carbon, such as grasslands were found to yield more EPS than their arable or fallow counterparts (Redmile-Gordon et al., 2020).

Ecological functions of soil biofilms

Biofilms consist mainly of microbial cells and self-produced EPS. The EPS matrix determines adhesive and cohesive properties of biofilms and enhances the adaptability of soil microorganisms (Or et al., 2007b). Accordingly, the soil biofilm confers unique features and functions. For example, EPS matrix protects microbes against dehydration and antimicrobial agents, anchors extracellular enzymes, captures nutrient and cellular products, and facilitates genetic exchange among populations (Flemming et al., 2016; Tecon and Or, 2017).

Protection against desiccation is one of the best-known properties of a soil biofilm. It is the hydrophilic components in EPS that play a critical role in resisting dehydration (Flemming et al., 2016). EPS secreted by microbes establishes a stable interface between

biological and physical components in soil via interfacial interactions including ligand exchange, hydrogen bonding and electrostatic interactions (Cai et al., 2018). During the reduction of hydraulic conductivity, increased water retention in the EPS matrix results in hydraulic decoupling of the biofilm from drier soil pores. EPS thereby maintains local hydrated areas and protects the biofilm cells. The hydrated matrix also enhances soil moisture content at a macro level and promotes nutrient diffusion under dry conditions (Chenu and Roberson, 1996).

Besides protection from desiccation, biofilm also provides physicochemical barriers against toxic substances. Due to the absorption and diffusion barrier, biofilm exhibits significantly enhanced tolerance to antimicrobial agents. For example, heavy metals can be accumulated in biofilm matrix via biosorption, bioprecipitation, intracellular accumulation and redox immobilization. The anionic charged functional groups, including hydroxyls, carboxyls, amides, amines and phosphorylase, interact with the positively charged metal ions. The microaerobic microenvironment in biofilm also facilitates *in situ* metal bioreduction. Metals such as Pd, Se, Cr, Tc and U were found to be reduced to low valent species and were immobilized in biofilm (Ahmed et al., 2012).

The binding of an EPS matrix with microbial cells in close proximity may establish stable spatial interactions. Soil microorganisms share their habitat with genetically diverse neighbors, leading to competitive or cooperative microbial interactions and determining community diversity and functional traits. The proximity of cells in biofilms facilitates the conjugative transfer of antibiotic resistance genes (ARGs). The EPS matrix provides a feasible environment to increase the stability of plasmids and the host range of mobile genetic elements. The close interaction between biofilm cells also promotes intercellular interaction like quorum sensing to enhance the conjugation of ARGs. Compared with free-living cells in the surrounding environment, the antibiotic-resistant plasmids exhibit a significantly higher conjugation transfer frequency in biofilms (Wu et al., 2021). Moreover, the spatial or physical orientation of biofilms can affect profoundly the way populations share genetic information. Biofilm environments also confer selective effects on plasmid-encoded traits. By evaluating the persistence of plasmids in their receptors and the local competitiveness of the host, biofilms can affect the transconjugants during horizontal gene transfer and further promote the spread of genetic elements via vertical gene transfer (Wu et al., 2021).

Soil biofilms have developed very effective strategies to capture water soluble nutrients, for example, from the substratum. This strategy depends mainly on the passive sorption characteristics of the sponge-like EPS matrix, which influence the exchange of nutrients, gases and other molecules between biofilms and the environment (Or et al., 2007a). Biofilms contain a variety of anionic and cationic binding sites able to bind nutrients at very low concentrations. The trapped and accumulated nutrients can sustain biofilm growth in severely undersupplied environments. In addition to soluble substances, suspended solids like particulate organic matters can be captured by biofilms and incorporated into the matrix (Flemming et al., 2016).

Future prospects

While soil biofilms may be expected to contain a small fraction of the total microbial biomass, they may contribute substantially to soil physical and biogeochemical processes. However, most of our existing knowledge about soil biofilms was obtained through parallel understanding of biofilms in clinical and aquatic systems. There are still fundamental questions and basic technical issues needed to be resolved. These include distinguishing biofilm-cells from free-living sessile or planktonic cells, elucidating the soil biofilm development processes, understanding triggers for biofilm phenotypes, visualizing spatial distributions of different populations within assemblages, and clarifying their functional roles. Simplified model system like artificial soil and synthetic biofilm communities can be initiated to examine the fundamental properties of soil biofilms. Improvements, refinements and combinations of existing techniques and novel approaches are needed to gain a comprehensive understanding of soil biofilm processes (Table 1). For example, non-invasive planar optode imaging can be employed in microcosm systems to resolve the temporal-spatial distribution of microbial activities in soil biofilms. The advent of next-generation physiological analyses presents opportunities to reveal the phenotypes of individual cells in biofilms via non-destructive approaches (Hatzenpichler et al., 2020). These advances will provide a clearer picture of spatial heterogeneity and microbially-mediated biogeochemical processes, and thus help to develop more sustainable soil management practices.

Soil ecosystems exhibit extreme heterogeneity in biological and physicochemical properties. Biogeographical assessments across landscape scales are important to understand the ecological significance of biofilms in soils. The development of omics techniques enables the survey of soil microbial features at multiple scales, which outlines the microbial diversity, spatial organization and functionality. However, due to the intrinsic complexity and the limited understanding of soil biofilm processes, there is a lack of specific molecular, taxonomic and functional signatures for biofilms. Therefore, it remains challenging to implement currently available omics techniques to quantify and characterize soil biofilm community.

Microfluidics, a widely applied approach in environmental microbiology, is a promising technique for biofilm research. Soil microfluidic chips provide optical access to natural opaque soil systems and enable direct characterization of microbial growth and communal interactions in microstructure or chemically controlled environments. However, *in situ* analysis of complex interactions with the high spatial and temporal resolution is still lacking. Microfluidic devices, combined with appropriate fluorescent tagging and imaging techniques, can detect chemical changes in the microbial biofilm matrix and metabolic activities like cell growth, movement and nutrient usage (Pucetaite et al., 2021). The application of advanced micro-spectral techniques in microfluidic soil chips, namely infrared absorption, Raman scattering and X-ray micro-spectral techniques based on synchrotron radiation are also

Table 1 Overview of techniques for soil biofilm research.

Techniques	Measured Parameters	Advantages	Disadvantages
Synthetic biofilm community	–	Precise control of community composition	Limited resemblance to natural ecosystem
Planar optode	Temporal-spatial distribution of microbial activities in soil biofilms	Non-invasive, real-time, and high resolution	Multi-analyte sensing unavailable
Meta-omics	Microbial diversity, spatial organization and functionality	Characterize biofilm complexity and variability at multiple scales; high-throughput; culture-independent	Each platform has different technological limitations due to sample preparation, sequencing platforms and depth, limits in instrumentation, and dynamic range.
Next-generation physiology approaches	Examine the physiological responses of biofilm communities	Characterization of uncultured microorganisms with unprecedented speed; high spatial resolution; non-destructive analysis	Mechanisms not well understood
Soil microfluidic chips	Enable direct characterization of microbial growth and communal interactions in well-defined microstructure	Precise control of physicochemical properties at micrometer scale and simulate specific growth conditions	Simulation still highly unnatural
Advanced microspectral techniques coupled with microfluidic chips	Physicochemical and biological reactions within biofilms with high spatial and temporal resolution	High reproducibility; in situ measurement at micro-scale resolution; simplified sample preparation	

making further breakthroughs at the frontier of soil science (Pucetaite et al., 2021). These chemical analyses can reveal physicochemical and biological reactions within soil system with the high spatial and temporal resolution. In future, microfluidic devices should be designed to represent and simulate unique changes in the soil environment, for example the drying and wetting with the natural occurrence of water infiltration and evaporation events. It is expected that microfluidic devices operating in controllable unsaturated soil conditions will deepen our understanding of soil biofilms and provide new perspectives for understanding soil microbial ecological processes (Cai et al., 2019).

Evidence-based soil management holds the key to improvements in the sustainability of agriculture and horticulture. Researchers seek to establish a quantitative framework to track individual cells and communities in a virtual soil environment, and to provide guidance for using experimental and monitoring activities. This includes incorporating soil dynamic processes into mechanical models and refining the models with new insights from more controlled observations. In future, the basic characteristics of soil biofilms should be assessed by studying increasingly complex synthetic model communities in the laboratory (Tecon and Or, 2017).

Conclusion

Soil biofilms are aggregates of microbial cells embedded in a self-produced EPS matrix. The EPS matrix binds cells in a biofilm and creates a unique environment for biofilm-dwelling cells that exhibit emergent properties absent in free-living cells. Soil biofilm formation has been shown to enhance microbial survival, community diversities and metabolic activities. The combined future study of natural and artificial biofilm communities can help reveal the mechanistic geochemical interactions driving key soil and biofilm processes. These advances in research will help us to obtain a more comprehensive, accurate and in-depth understanding of both soil and the microbiome.

Reference

- Ahmed B, Cao B, McLean JS, Ica T, Dohnalkova A, Istanbul O, Paksoy A, Fredrickson JK, and Beyenal H (2012) Fe (III) reduction and U (VI) immobilization by *Paenibacillus* sp. Strain 300A, isolated from Hanford 300A subsurface sediments. *Applied and Environmental Microbiology* 78(22): 8001–8009.
- Bar-On YM and Milo R (2019) Towards a quantitative view of the global ubiquity of biofilms. *Nature Reviews. Microbiology* 17(4): 199–200.
- Baveye PC (2020) “Soil biofilms”: Misleading description of the spatial distribution of microbial biomass in soils. *Soil Ecology Letters* 2(1): 2–5.
- Blagodatskaya E and Kuzyakov Y (2013) Active microorganisms in soil: Critical review of estimation criteria and approaches. *Soil Biology and Biochemistry* 67: 192–211.
- Bohórquez J, McGenity TJ, Papaspyrou S, García-Robledo E, Corzo A, and Underwood GJ (2017) Different types of diatom-derived extracellular polymeric substances drive changes in heterotrophic bacterial communities from intertidal sediments. *Frontiers in Microbiology* 8: 245.
- Buckeridge KM, Creamer C, and Whitaker J (2022) Deconstructing the microbial necromass continuum to inform soil carbon sequestration. *Functional Ecology* 00: 1–5.

- Cai P, Lin D, Peacock CL, Peng W, and Huang Q (2018) EPS adsorption to goethite: Molecular level adsorption mechanisms using 2D correlation spectroscopy. *Chemical Geology* 494: 127–135.
- Cai P, Sun X, Wu Y, Gao C, Mortimer M, Holden PA, Redmile-Gordon M, and Huang Q (2019) Soil biofilms: Microbial interactions, challenges, and advanced techniques for ex-situ characterization. *Soil Ecology Letters* 1–9.
- Chenu C and Roberson EB (1996) Diffusion of glucose in microbial extracellular polysaccharide as affected by water potential. *Soil Biology and Biochemistry* 28(7): 877–884.
- Costerton JW, Geesey GG, Ladd TI, Nickel JC, Dasgupta M, and Marrie TJ (1987) Bacterial biofilms in nature and disease. *Annual Review of Microbiology* 41(1): 435–464.
- Eickhorst T and Tippkötter R (2008) Detection of microorganisms in undisturbed soil by combining fluorescence in situ hybridization (FISH) and micropedological methods. *Soil Biology and Biochemistry* 40(6): 1284–1293.
- Fernandes PI Jr., de Oliveira PJ, Rumjanek NG, and Xavier GR (2011) Poly-beta-hydroxybutyrate and exopolysaccharide biosynthesis by bacterial isolates from pigeonpea [*Cajanus cajan* (L.) Millsp.] root nodules. *Applied Biochemistry and Biotechnology* 163(4): 473–484.
- Flemming HC and Wuert S (2019) Bacteria and archaea on Earth and their abundance in biofilms. *Nature Reviews. Microbiology* 17(4): 247–260.
- Flemming HC, Wingender J, Szewzyk U, Steinberg P, Rice SA, and Kjelleberg S (2016) Biofilms: An emergent form of bacterial life. *Nature Reviews. Microbiology* 14(9): 563–575.
- Flemming HC, Baveye P, Neu TR, Stoodley P, Szewzyk U, Wingender J, and Wuert S (2021) Who put the film in biofilm? The migration of a term from wastewater engineering to medicine and beyond. *NPJ Biofilms Microbiomes* 7(1): 10.
- Gutiérrez Castorena EV, Gutiérrez-Castorena MDC, González Vargas T, Cajuste Bontemps L, Delgadillo Martínez J, Suástegui Méndez E, and Ortiz Solorio CA (2016) Micromapping of microbial hotspots and biofilms from different crops using digital image mosaics of soil thin sections. *Geoderma* 279: 11–21.
- Hall-Stoodley L and Stoodley P (2009) Evolving concepts in biofilm infections. *Cellular Microbiology* 11(7): 1034–1043.
- Hatzenpichler R, Krukenberg V, Spietz RL, and Jay ZJ (2020) Next-generation physiology approaches to study microbiome function at single cell level. *Nature Reviews. Microbiology* 18(4): 241–256.
- Li XY and Yang SF (2007) Influence of loosely bound extracellular polymeric substances (EPS) on the flocculation, sedimentation and dewaterability of activated sludge. *Water Research* 41(5): 1022–1030.
- Liang C, Schimel JP, and Jastrow JD (2017) The importance of anabolism in microbial control over soil carbon storage. *Nature Microbiology* 2: 17105.
- Mikutta R, Zang U, Chorover J, Haumaier L, and Kalbitz K (2011) Stabilization of extracellular polymeric substances (*Bacillus subtilis*) by adsorption to and coprecipitation with Al forms. *Geochimica et Cosmochimica Acta* 75(11): 3135–3154.
- Nunan N, Wu K, Young IM, Crawford JW, and Ritz K (2003) Spatial distribution of bacterial communities and their relationships with the micro-architecture of soil. *FEMS Microbiology Ecology* 44(2): 203–215.
- Or D, Phutane S, and Dechesne A (2007a) Extracellular polymeric substances affecting pore-scale hydrologic conditions for bacterial activity in unsaturated soils. *Vadose Zone Journal* 6(2): 298–305.
- Or D, Smets BF, Wraith J, Dechesne A, and Friedman S (2007b) Physical constraints affecting bacterial habitats and activity in unsaturated porous media—a review. *Advances in Water Resources* 30(6–7): 1505–1527.
- Pucetaite M, Ohlsson P, Persson P, and Hammer E (2021) Shining new light into soil systems: Spectroscopy in microfluidic soil chips reveals microbial biogeochemistry. *Soil Biology and Biochemistry* 153: 108078.
- Redmile-Gordon M, Brookes P, Evershed R, Goulding K, and Hirsch P (2014) Measuring the soil-microbial interface: Extraction of extracellular polymeric substances (EPS) from soil biofilms. *Soil Biology and Biochemistry* 72: 163–171.
- Redmile-Gordon MA, Evershed RP, Hirsch PR, White RP, and Goulding KW (2015) Soil organic matter and the extracellular microbial matrix show contrasting responses to C and N availability. *Soil Biology and Biochemistry* 88: 257–267.
- Redmile-Gordon M, Gregory AS, White RP, and Watts CW (2020) Soil organic carbon, extracellular polymeric substances (EPS), and soil structural stability as affected by previous and current land-use. *Geoderma* 363: 114–143.
- Stuart RK, Mayali X, Lee JZ, Craig Everrood R, Hwang M, Bebout BM, Weber PK, Pett-Ridge J, and Thelen MP (2016) Cyanobacterial reuse of extracellular organic carbon in microbial mats. *The ISME Journal* 10(5): 1240–1251.
- Tecon R and Or D (2017) Biophysical processes supporting the diversity of microbial life in soil. *FEMS Microbiology Reviews* 41(5): 599–623.
- Teitzel GM and Parsek MR (2003) Heavy metal resistance of biofilm and planktonic *Pseudomonas aeruginosa*. *Applied and Environmental Microbiology* 69(4): 2313–2320.
- Vert M, Doi Y, Hellwich K-H, Hess M, Hodge P, Kubisa P, Rinaudo M, and Schué F (2012) Terminology for biorelated polymers and applications (IUPAC Recommendations 2012). *Pure and Applied Chemistry* 84(2): 377–410.
- Wang X, Sharp CE, Jones GM, Grasby SE, Brady AL, and Dunfield PF (2015) Stable-isotope probing identifies uncultured *Planctomycetes* as Primary degraders of a complex heteropolysaccharide in soil. *Applied and Environmental Microbiology* 81(14): 4607–4615.
- Watteau F and Villemain G (2018) Soil microstructures examined through transmission electron microscopy reveal soil-microorganisms interactions. *Frontiers of Environmental Science & Engineering* 6: 106.
- Wu Y, Cai P, Jing X, Niu X, Ji D, Ashry NM, Gao C, and Huang Q (2019) Soil biofilm formation enhances microbial community diversity and metabolic activity. *Environment International* 132: 105116.
- Wu S, Wu Y, Cao B, Huang Q, and Cai P (2021) An invisible workforce in soil: The neglected role of soil biofilms in conjugative transfer of antibiotic resistance genes. *Critical Reviews in Environmental Science and Technology* 52: 1–29.
- Ye F, Ye Y, and Li Y (2011) Effect of C/N ratio on extracellular polymeric substances (EPS) and physicochemical properties of activated sludge flocs. *Journal of Hazardous Materials* 188(1–3): 37–43.
- Young IM and Crawford JW (2004) Interactions and self-organization in the soil-microbe complex. *Science* 304(5677): 1634–1637.
- Zhang M, Peacock CL, Cai P, Xiao K-Q, Qu C, Wu Y, and Huang Q (2021) Selective retention of extracellular polymeric substances induced by adsorption to and coprecipitation with ferrihydrite. *Geochimica et Cosmochimica Acta* 299: 15–34.

Biocrusts

Matthew A Bowker^{a,b}, Jasmine Anenberg^a, Hilina Yohannes Kebede^a, Anita Antoninka^a, and Lydia Bailey^{a,b}, ^aSchool of Forestry, Northern Arizona University, Flagstaff, AZ, United States; ^bCenter for Ecosystem Science and Society, Northern Arizona University, Flagstaff, AZ, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	128
Global distribution	128
Composition and global biodiversity	131
Functional properties and roles in ecosystems	132
Desiccation tolerance	132
Tolerance to other stresses	132
Ultraviolet radiation	132
Temperature	133
Ecosystem functional roles	133
Soil surface stability	133
Carbon fixation	133
Nitrogen fixation and nutrient cycling	133
Hydrological functions	134
Soil surface albedo and heat storage	134
Interactions with other biota	134
Multifunctionality	134
Threats and Recovery	134
Physical disturbance	135
Climate change	135
Natural recovery from disturbance	136
Assisted recovery	136
Management considerations	136
Conclusion	137
References	137

Abstract

Biocrusts are a thin, distinct, cohesive layer at the surface of terrestrial soils, engineered by microbes and cryptogams. They are found throughout the world, especially in drylands, but are also known from cold-limited environments and some disturbed mesic temperate environments. Biocrusts are multifunctional. They are photosynthetic, reduce erosion, accumulate nutrients, influence water redistribution and storage, and interact with other biotic communities. Biocrusts are sensitive to both climate change and physical disturbances, and can be lost much faster than they can subsequently recover. Assisted biocrust recovery techniques are being developed as biocrusts are increasingly incorporated into land management decisions.

Glossary

Bryophyte True land plants that lack vascular tissue and do not reproduce from seeds. There are three groups of bryophytes: mosses, liverworts and hornworts.

Cryptogam A “plant” (sensu Linnaeus) that does not flower or reproduce from seeds. This is a historical collective term for related and unrelated organisms including algae, lichens, mosses, ferns, fungi, and cyanobacteria.

Cyanobacteria A phylum of photosynthetic bacteria. They are also commonly referred to as blue-green algae, though they are no longer classified as algae because they are prokaryotes.

Desiccation tolerance The ability of an organism to lose a large proportion of its body water, but remain alive in an inactive, dry state. Desiccation tolerance also involves the ability to rehydrate and reactivate metabolic processes after a dry, inactive period.

Lichen A symbiotic partnership between at least two organisms, always including a fungus that houses either a cyanobacterial or algal partner within it. Some lichens have fungal, cyanobacterial and algal components, while others may have more than one fungal species involved.

Poikilohydry Lack of ability of an organism to regulate water uptake or loss, leading to equilibration of cell or tissue water content with the environment. Poikilohydric organisms that occupy dry environments are often desiccation tolerant.

Key points

- Biocrusts are a distinct, thin cohesive and photosynthetic layer at the soil surface held together by biota
- Biocrusts are found throughout the world in drylands, cold-limited environments and disturbed mesic ecosystems
- Biocrust biota include photosynthesizing organisms, such as cyanobacteria, lichens, algae, mosses and liverworts, and heterotrophs, such as fungi, other bacteria, and microfauna
- Biocrusts can tolerate drying without dying, reduce erosion, accumulate nutrients, influence water redistribution and storage, and interact with other biotic communities
- Biocrusts may be lost or degraded by physical disturbances and climate change, but increasingly are being considered in ecological restoration and land management

Introduction

Biological soil crusts, increasingly referred to in the shorthand as “biocrusts”, may have originated on land long before the colonization and radiation of vascular plants. Multiple lines of evidence suggest the existence of terrestrial communities extending billions of years into Earth’s past. Cyanobacteria are one group of organisms that possessed many of the traits that would have enabled life on land - oxygenic photosynthesis, ultraviolet tolerance mechanisms, desiccation tolerance, nitrogen fixation (in some taxa) - and possibly occupied land ~2400 million years ago. Because cyanobacteria are photo-autotrophic, they may have formed biocrusts or biocrust-like communities on soil surfaces (Weber et al., 2016).

Eukaryotic algae are much younger, but still older than and ancestral to all land plants. They may have existed on terrestrial surfaces ~500 million years ago or earlier. Bryophytes, including mosses and liverworts, and lichens are younger still, and originated on land between ~400–500 million years ago, either evolving from green algal ancestors in the case of bryophytes or arising from fungal symbioses with cyanobacteria or algae in the case of lichens. Therefore, it is highly plausible that complex biocrust communities consisting of any combination of cyanobacteria, green algae, mosses, liverworts or lichens, and associated microbes, predated the fossil record of all tracheophytes (vascular plants), and very likely preceded the phanerogams (seed plants) that eventually came to dominate the land surface (Weber et al., 2016). Biocrusts are still extant, widespread and sometimes abundant today in many different ecosystems, serving as modern analogs of early terrestrial communities.

Recognition of biocrusts as a distinctive community type, and interest in their ecological roles came about relatively recently, mostly after the mid- twentieth century. Ecological studies of biocrusts became more common in the 1980s and 1990s, and have proliferated at an increasingly fast rate, subsequently becoming a major subject of ecological research (Weber et al., 2016). Multiple synonyms have been used to refer to these communities with replacements over time. Among others, these included cryptogamic, microbiotic, microfloral, microphytic, and cryptobiotic crusts; these were either inconsistently used, or fell out of favor, for example because they were not considered to be inclusive of all the biota that generate biocrusts. Most authors are currently using the terms biological soil crusts or biocrusts (Belnap and Lange, 2003, Weber et al., 2016). Increasingly, biocrusts are being compared to a “living skin” on the soil surface (Bowker et al., 2018).

A recent synthesis proposed a universal definition for biocrusts, encompassing the wide variety of communities referred to as such in the ever-expanding literature.

Biological soil crusts (biocrusts) result from an intimate association between soil particles and differing proportions of photoautotrophic (e.g., cyanobacteria, algae, lichens, bryophytes) and heterotrophic (e.g., bacteria, fungi, archaea) organisms, which live within, or immediately on top of, the uppermost millimeters of soil. Soil particles are aggregated through the presence and activity of these often extremotolerant biota that desiccate regularly, and the resultant living crust covers the surface of the ground as a coherent layer.

(Weber et al., 2022)

This definition brings together 4 major intersecting elements: habitat, physical structure, functional characteristics, and taxonomic composition (Weber et al., 2022). Biocrusts inhabit terrestrial soils in habitats in which the soil surface dries regularly. Biocrusts form a thin - usually millimeters thick, or a few centimeters at most - distinctly structured layer at the soil surface that differs from underlying soil. This structure arises specifically because of the functional attributes of biocrusts: (a) biocrusts always contain photosynthetic organisms that must occur at the soil surface where light is available, (b) biocrust organisms aggregate soil, generating the characteristic physical structure, and (c) because of their soil surface habitat, they are tolerant of extreme biological conditions such as desiccation. Biocrusts are dominated by certain lineages of photoautotrophs belonging to the historical grouping of “cryptogams” and exclude tracheophytes. Though this grouping of taxa is not phylogenetically supported, it does express commonality in several functional traits (small size, poikilohydric common habitat preferences, etc.). Several examples of communities that fall within this definition are illustrated in Fig. 1.

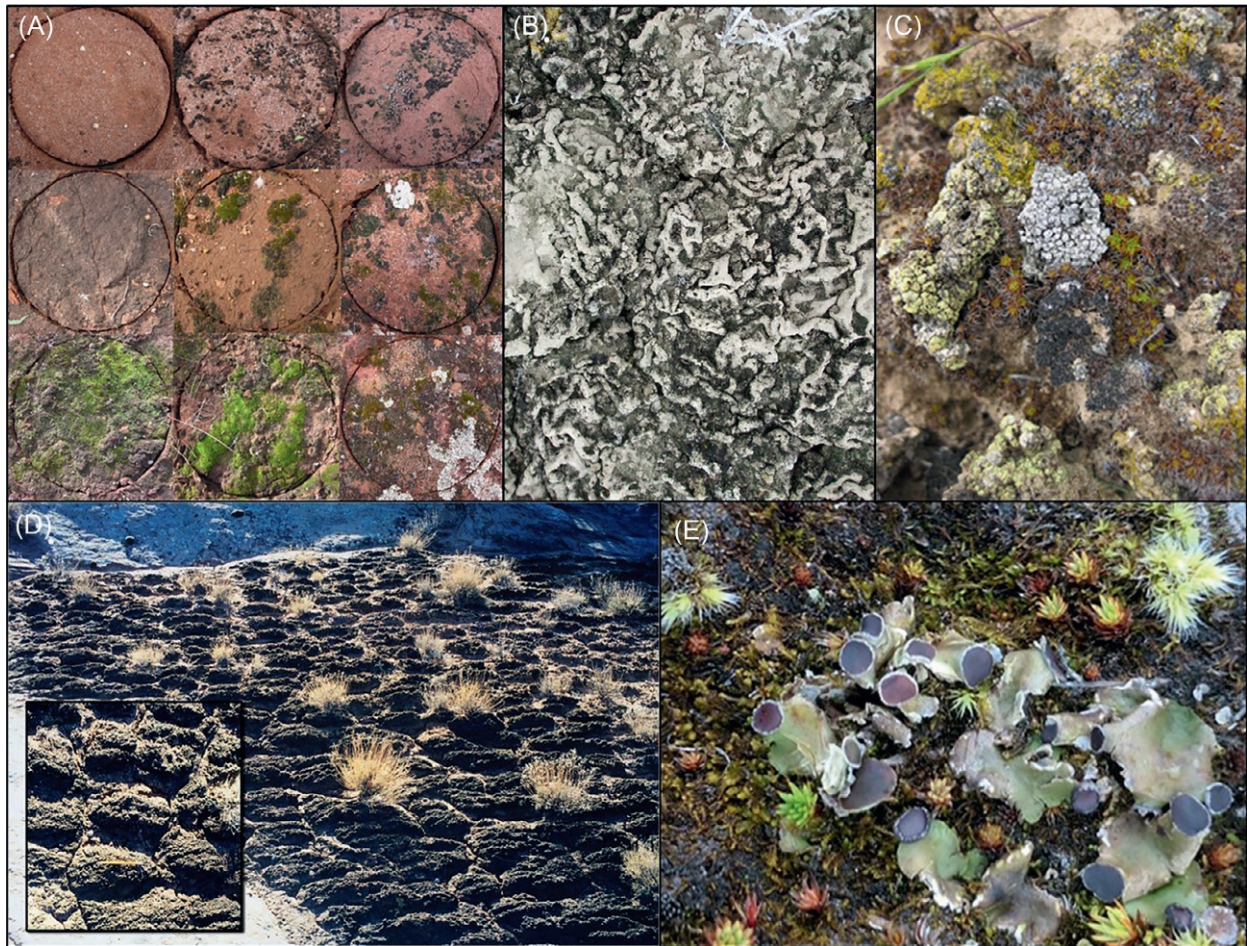


Fig. 1 Varied forms, compositions, colors and textures of biocrusts. (A) A series of distinct biocrust communities in close proximity differing in relative abundance of cyanobacteria, mosses and lichens, diversity, and spatial heterogeneity; only the upper right panel depicts a surface lacking biocrust (New South Wales, Australia). (B) A cyanobacteria-dominated biocrust with dark patches indicative of microbial sunscreens (Utah, USA). (C) A species-rich biocrust featuring multiple variously colored lichens, interspersed with mosses and patches of cyanobacterial biocrust (Idaho, USA). (D) A microtopographically-complex biocrust, dominated by cyanobacteria; inset highlights elongated micro-ridges and valleys (Utah, USA). (E) A well-developed biocrust composed of lichens, liverworts and mosses growing in a disturbed mesic area, approximately 20 years after ecological restoration activities (Iceland).

Global distribution

Biocrusts are a global community type, found on all continents in many different habitats (Weber et al., 2016). One of the most common shared attributes among biocrust habitats is an open vegetation canopy that allows substantial light to reach the soil surface. This is because biocrusts are photosynthetic, but extremely limited in height, and therefore cannot compete for light with most vascular plants. Biocrusts inhabit only terrestrial habitats, not aquatic or marine ones, and are not typically found in habitats flooded with water for extended periods of time.

Biocrusts are perhaps best documented in the drylands of the world, regions with a high ratio of potential evapotranspiration to precipitation, where water limits vascular plant distribution (Fig. 2). This includes a broad range from subhumid grasslands and savannas and some open forests, semi-arid shrublands, steppes, grasslands, savannas, desert shrublands and grasslands, and some hyper-arid regions. Notable ecoregions supporting well-documented biocrusts include the Negev Desert of Israel, the Kalahari Desert and Sahel region of Africa, the Colorado Plateau and other deserts of North America, semi-arid alpha grass steppes of Spain, Eucalypt woodlands and drier deserts of Australia, northern cool and cold deserts of China (Gurbantunggut, Tennger, Mu Us deserts and the Loess Plateau), among many others (Weber et al., 2016, 2022). In these environments, biocrusts may occupy and sometimes fill the interspaces in between vascular plants, and may also be found underneath the canopy of vascular plants.

Given the wide variety of potential dryland habitat types, dryland biocrusts have been documented and studied on all continents (Rodríguez-Caballero et al., 2018; Fig. 2). In a recent study focusing on biocrusts documented in warm and cool deserts (excluding cold deserts), it was estimated that about 24% of the terrestrial land surface was suitable for dryland biocrusts and that about 12% of the terrestrial surface is currently occupied by biocrusts (Rodríguez-Caballero et al., 2018; Fig. 2).

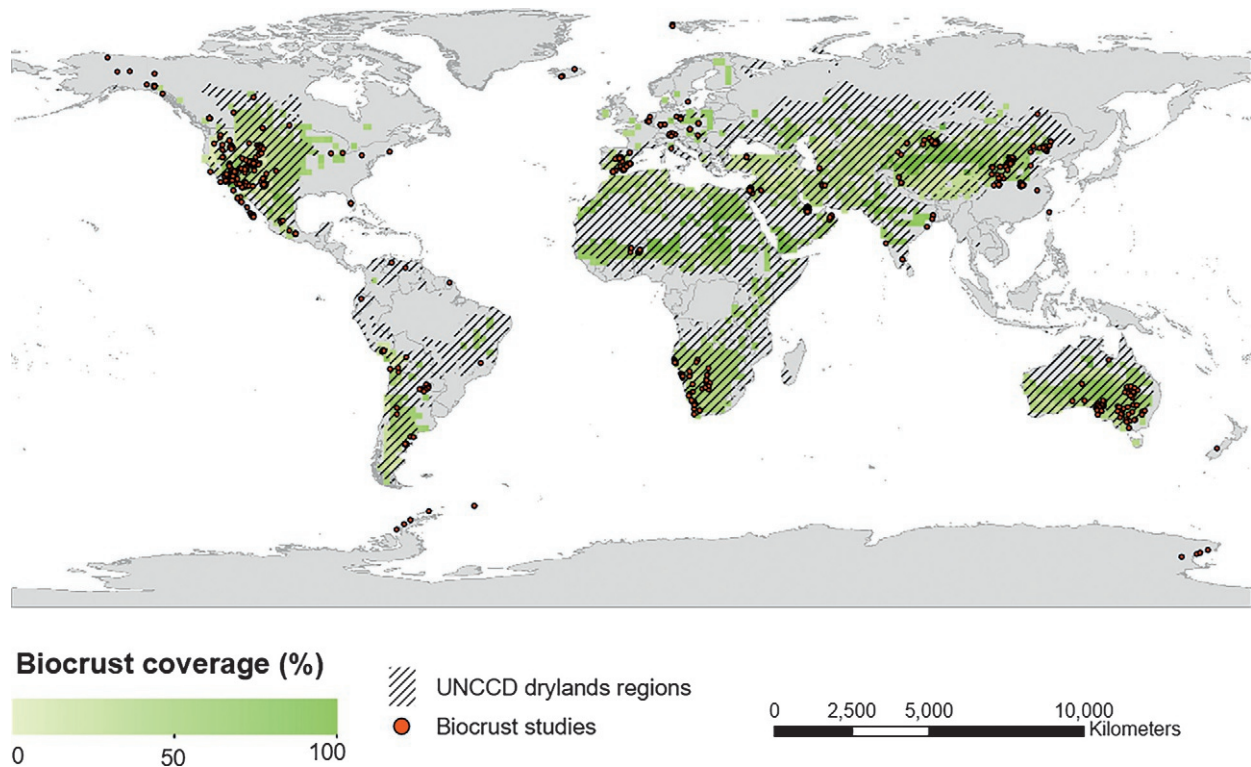


Fig. 2 The global distribution of biocrust studies in both dryland and non-dryland environments (map: Emilio Rodríguez-Caballero). The niche and coverage of biocrusts (modeled in Rodríguez-Caballero et al., 2018; training data included biocrust occurrences in hot and cool drylands but excluded polar regions and anthropogenic landscapes) is shown alongside dryland regions according to UNCCD (UNEP-WCMC, 2007).

Biocrusts are also documented to occur in cold-limited environments such as alpine and high latitude regions, including Antarctica (Jung et al., 2018). Some notable examples include Devon Island and the Canadian Arctic, Svalbard, Greenland, highlands of Iceland, the Austrian Alps, and the Antarctic dry valleys and maritime regions (Weber et al., 2022). Biocrusts of these high altitude or latitude cold regions are not as widely documented or studied as in the drylands, thus their full distribution is not known. In these regions, water and cold temperatures may be co-limiting, either because of little precipitation or because water is often in a frozen state (Weber et al., 2022).

The final climatic zone where biocrusts may occur is in disturbed, mesic regions or in unique soil environments within mesic regions (Corbin and Thiet, 2020). Because climatically-induced stress is less in this type of environment than other typical biocrust environments, the disturbance provides a window of opportunity for biocrust development before a complete coverage of vascular plants can reestablish. Biocrusts are found in primary successional environments, for example colonizing land exposed by glacial retreat (Weber et al., 2022) and in secondary successional settings, such as following stand-destroying wildfires in forested regions throughout the world, where fast-growing moss biocrusts may form (Weber et al., 2022). Biocrusts have occasionally been documented in agricultural soils, either after abandonment, or between tillage events. In some cases, stressful or unusual soil conditions might limit vascular plant canopy closure, and combined with occasional disturbance, create a long-term niche for biocrusts, for example in pine barrens of North America or mesic inland dunes of Germany (Weber et al., 2022). The occurrence of biocrusts in the wide array of situations, described above, expands the probable map of biocrust occurrence, even if only temporarily, to most terrestrial ecoregions.

The factors controlling biocrust occurrence, degree of development and composition are best understood in the drylands and differ strongly with the spatial scale (Weber et al., 2016). Biogeographical patterns emerge, for example, geographical distance or time of land mass isolation, mostly at global and continental scales, that influence biocrust species composition and structure. Climatic (and micro-climatic) factors are a ubiquitous influence at all scales from global to micro-scale. Chief among the climatic factors are the interactive effects of temperature and precipitation regimes, which co-determine the amount, mode and timing of soil moisture availability at the soil surface. Soil surface moisture availability for biocrusts is best understood not as an amount of water, but rather the residence time of a sufficient amount of water to activate biocrust metabolism. Extremely low moisture availability, such as in hyper-arid regions, tends to limit both the total amount of biocrust present and restrict biocrust composition to only the most stress-tolerant species. Additional moisture availability, e.g., arid to semi-arid regions, promotes greater biocrust cover and a wider variety of possible compositions (including lichen- or moss-dominated types).

Composition and global biodiversity

Biocrusts are dominated by photoautotrophic organisms (those with the ability to photosynthesize and fix their own carbon) including cyanobacteria, eukaryotic algae (multiple lineages), mosses, liverworts and lichens (Weber et al., 2022). Cyanobacteria are among the most ancient and widespread members of the biocrust community, and are often dominant or co-dominant in drylands and elsewhere. There are three different functional types of cyanobacteria: 1) filamentous biocrust builders, which create long sheathed colonies which bind soil particles together and create stability, 2) biocrust residents which add additional functionality such as N-fixation, and 3) cyanobacteria that arrive via stochastic processes and do not permanently reside in biocrusts (Weber et al., 2016). Eukaryotic Algae, such as chlorophytes and diatoms, are another ancient component of biocrusts, widely dispersed, co-dominant or dominant in some environments but often found in low abundance in drylands (Weber et al., 2016). Lichens are prominent symbiotic components of biocrusts, both functionally and visually, and are mostly macroscopic. They harbor an impressive amount of diversity in morphologies and functional traits that determine ecosystem process rates (Mallen-Cooper and Eldridge, 2016). Bryophytes, a collective term for mosses, liverworts, and hornworts, are conspicuous, macroscopic, widespread and sometimes dominant in biocrusts. Similar to lichens, the functional trait diversity of bryophytes allows for a large diversity of influential traits within the biocrust community (Mallen-Cooper and Eldridge, 2016). Both lichens and bryophytes tend to be less common where soil moisture is extremely limiting (e.g., hyperarid regions), but attain a higher degree of dominance as effective precipitation increases (e.g. semiarid regions; Weber et al., 2016).

Biocrusts are also inhabited by heterotrophs, some of which may also play roles in biocrust formation alongside the photoautotrophic components (Weber et al., 2022). These include fungi, bacteria (other than cyanobacteria), and microfauna, including protozoa, nematodes, tardigrades, rotifers, and microarthropods. Fungi (excluding lichen-forming types) are important biocrust components, but are poorly studied in comparison to lichenized fungi because of their microscopic nature and because the majority are not readily culturable. Ascomycetes are particularly well-represented in biocrusts. Because of their filamentous hyphae, many fungi likely contribute to biocrust building, though that role is understudied. Fungi may also mediate nutrient exchanges between biocrusts and vascular plants. Some fungi live on or within other biocrust biota as commensals, parasites or pathogens (Weber et al., 2016). Bacteria (phyla other than Cyanobacteria) comprise a large proportion of the microbial community members in biocrusts. Bacterial diversity estimates and compositions in biocrusts (and in general) are highly dependent on the methodology employed. Only in the past three decades have researchers been able to advance from culture-dependent methods to early culture-independent molecular methods to next-generation high-throughput sequencing approaches, capturing an increasingly complete picture of bacterial communities (Weber et al., 2016). Bacterial communities can be highly diverse; it is not unusual for several phyla to be detected in biocrusts containing a large number of operational taxonomic units (OTU) or amplicon sequencing variants (ASV) using modern methodologies. Actinobacteria, Proteobacteria, Bacteroidetes and Acidobacteria are often among the most abundant phyla (Weber et al., 2016). Protozoa, nematodes, tardigrades, rotifers and microarthropods are the major groups of animals and animal-like protists that complete a complex food web within biocrusts (Weber et al., 2016).

Worldwide, at least 68 liverworts, 255 mosses, 553 lichens, 320 species of cyanobacteria and 353 species of eukaryotic algae have been reported to inhabit biocrusts (Weber et al., 2016; updated in Bowker et al., 2017). These are pooled estimates from papers which include species lists and are not expected to be a complete accounting of the global species pool of inhabitants for the following reasons:

1. The available literature has multiple geographic gaps.
2. In some of these groups, cyanobacteria in particular, taxonomy is in flux and increasingly reliant on molecular approaches to resolve taxa that are morphologically very similar (e.g. *Microcoleus vaginatus* and the *Microcoleus steenstrupii* complex); similar appearance hides multiple species and even families (Fernandes et al., 2021).
3. The estimates above exclude important groups of heterotrophic biocrust biota. No attempts have been made to estimate the global diversity of bacterial or fungal diversity at OTU or ASV levels within biocrusts, and most studies of biocrust microfauna are not conducted at the species-level.

Therefore, the global species pool is likely much larger than the 1549 species identified to date.

There are several multi-continental or even global taxa that are commonly reported from biocrusts and which may be among the dominant species (Bowker et al., 2017). In fact, some geographically distant biocrust floras have substantial overlap. Nonetheless, there is considerable variation among biocrust communities from place to place, regulated by differing forces at different scales of observation. At the largest scales, dispersal limits may create differentiation between biocrust floras of different regions. Climatic and edaphic factors also influence compositions at large to small scales because these are two major elements of species' ecological niches.

Because biocrusts vary in their composition from place to place, a few classification schemes have been used which facilitate comparisons among regions. Most commonly, biocrust communities are classified by their dominant photoautotrophic component, e.g., moss, lichen or cyanobacterial biocrusts, among other possibilities (e.g., Havrilla et al., 2019). Cyanobacterial biocrusts are sometimes further divided into "light" and "dark" types, based on their pigmentation, which is reflective of the relative abundance of certain genera. Ordinal scales of development have been proposed, which simultaneously capture darkness, succession, community compositional shifts, and differences in function in the whole biocrust community (Belnap et al., 2008). The physical structure and morphology of biocrusts has also been used to differentiate different biocrust types. For example, flat, rugose, pinnated, and rolling morphologies are linked to different habitat types and are expected to have different functional

properties in ecosystems (Weber et al., 2022). Functional-morphological groups have been proposed for some taxonomic groups of biocrust organisms to reduce the need for taxonomic expertise in biocrust studies; these groupings effectively encapsulate multiple functional traits (Mallen-Cooper and Eldridge, 2016).

Functional properties and roles in ecosystems

Desiccation tolerance

Biocrusts lack roots, do not have vascular tissue for water transport (with the exception of a handful of moss taxa), and do not have stomata to regulate water loss. Rather, water is absorbed directly from the environment (e.g., rainfall, fog droplets, wet surfaces) and is subsequently lost by evaporation, leaving biocrust organisms in a desiccated state. Biocrust organisms dry without dying. They can suspend metabolic processes, then resume them upon rehydration, often within seconds (Fig. 3A). Therefore, they are considered desiccation tolerant, and many biocrust organisms can survive months or longer in the dry state.

There are limits to this ability, however. Desiccating biocrust organisms endure cellular damage, e.g., to their membranes, or oxidative stress, thus repair is often an early priority upon reactivating metabolic processes. Because of the cost of repair, respiration rates are often higher than photosynthetic rates early in a rehydration event (Weber et al., 2016). If the hydration event is too short to attain a positive carbon balance, the organism will experience a net loss of carbon during the event. A series of such events can be lethal (Weber et al., 2016).

Tolerance to other stresses

Ultraviolet radiation

Because of their position on the surface of the soil, and the open-canopied nature of their habitats, biocrusts are exposed to the full light spectrum, including ultraviolet radiation (UV-A and UV-B). Many biocrust organisms have at least one strategy or mechanism to cope with this stressor, some apparently ancient. Bundled, motile, filamentous cyanobacteria have an avoidance strategy. These



Fig. 3 Biocrust functional properties and roles in ecosystem functions. (A) Biocrust organisms, such as the moss *Syntrichia ruralis*, are poikilohydric and desiccation tolerant. The significance of this is that they can dry without dying, entering an inactive state, and resume activity upon rehydration (photo: Theresa Clark; Arizona, USA). (B) A greenhouse-grown biocrust generates aggregation of the soil surface through a proliferation of cyanobacterial filaments that weave a cohesive surface layer (photo: Kristina Young; Arizona, USA). (C) Stark differences in surface pigmentation due to experimental repeated trampling of biocrusts, increasing albedo (photo: Sasha Reed; Utah, USA). (D) Low-growing vascular plants cluster within and directly adjacent to gray patches of liverwort-dominated biocrust (highlands of Iceland).

species may glide within a polysaccharide sheath onto the surface of the soil, or downward under the protection of soil particles and other organisms (Belnap and Lange, 2003). Non-motile organisms, e.g., surface-bound cyanobacteria, lichens, and bryophytes cannot pursue this avoidance strategy, but have a variety of pigments that assist with either reducing damage caused by UV exposure or mitigating it (Belnap and Lange, 2003). For example, cyanobacteria and cyanolichens may have sunscreen pigments, like scytonemin or mycosporine-like amino acids which absorb UV radiation, intercepting it before it can damage living cells and tissues (Belnap and Lange, 2003). Some families of pigments, carotenoids and xanthophylls found in all major groups of photosynthetic biocrust organisms, are anti-oxidants capable of quenching free radicals and reducing oxidative stress caused by UV exposure (Belnap and Lange, 2003; Weber et al., 2016).

Temperature

Many biocrust organisms appear tolerant of temperature extremes because they inhabit some of the hottest and coldest habitats on Earth. However, this assessment is deceptive because often the most extreme temperatures occur when biocrusts are desiccated and inactive (Weber et al., 2016). The periods of time in which biocrusts are actually hydrated and active often occur in the portions of the year with milder climates (for example in the coolest season of warm climates). When hydrated, many biocrust organisms (e.g., mosses, most lichens) may have relatively cool temperature optima. Temperatures above 30 °C are often suboptimal if not stressful for many species, and some taxa have the ability to remain active at temperatures close to 0 °C (Belnap and Lange, 2003).

Ecosystem functional roles

Soil surface stability

Biocrusts are important in soil surface stabilization and dust trapping. Filamentous structures of biocrust biota weave through the top few millimeters of soil, binding loose soil particles together and generating a distinguishable, thin aggregated layer on the original soil surface (Belnap and Lange, 2003; Fig. 3B). These structures include filaments (trichomes) of cyanobacteria, rhizoids of mosses, rhizines of lichens, and hyphae of fungi. Biocrusts also stabilize soil surfaces through the exudation of *exo*-polymeric substances, notably polysaccharides, which can act as glues or provide charged surfaces for adsorption. Some of the major cyanobacteria create a matrix of polysaccharide sheaths in which their colonies also live. Biocrusts may also improve soil surface stabilization through the accumulation of organic matter and fine soil particles. Soil surfaces aggregated by biocrusts are much more resistant to both water and wind erosion than soils that lack them. Notably, researchers estimated that global dust flux - an outcome of wind erosion - would be 60% greater in the absence of soil stabilization by biocrusts (Rodríguez-Caballero et al., 2022). Biocrusts not only reduce dust emissions at the source but also trap and accumulate dust particles originating from other areas.

Carbon fixation

Biocrusts are net photosynthetic communities. Although the photosynthetic machinery is the same as in vascular plants, biocrust photosynthesizers are active only when hydrated, which may be only a small proportion of the year in some biocrust habitats. Because of this, the overall amount of carbon fixed on a per area basis will be much smaller than that achieved by vascular plants (even after taking size into account; Weber et al., 2016). Both chlorophyll content and maximal photosynthetic rate varies widely, even within functional groups (Weber et al., 2016). Maximal photosynthetic rates of biocrust organisms differ as a function of the dominant phototroph. Under optimal conditions it can be comparable to or exceed that of conifer needles, such as in lichens, but may be much smaller in cyanobacterial biocrusts (Belnap and Lange, 2003). In general, biocrust organisms with less dry mass per area have a greater maximum net photosynthetic rate (Weber et al., 2016). Several biocrust species have been characterized as “sun plants” in the laboratory (Belnap and Lange, 2003), but in the field the same species may be more successful in less exposed habitats because period of hydration there may be longer. Measuring net carbon uptake of biocrusts *in situ* is challenging because biocrust C uptake may be smaller than respiration from the underlying soil column which might contain vascular plant roots. Long-term studies of biocrust-covered soils tend to show net respiration for much of the year, punctuated by periods of net photosynthesis when the climate is optimal for biocrust photosynthesis and growth (Weber et al., 2016). Partitioning the gas exchange activity of natural biocrusts from activity below the biocrusts in the field over the long-term may require technological advances (Weber et al., 2016).

Nitrogen fixation and nutrient cycling

Biocrusts harbor species with the capability of fixing atmospheric di-nitrogen gas and converting it to ammonium, a key form of mineral nitrogen, which is an essential nutrient for plants and microbes. This is a crucial process because most terrestrial ecosystems are limited by the availability of mineral nitrogen. Some biocrust cyanobacteria, and some biocrust lichens with cyanobacterial photobionts have this ability, along with some heterotrophic bacteria that reside in biocrusts. Annual nitrogen fixation by biocrusts varies by orders of magnitude from study to study as the biocrust communities and methodologies change (Weber et al., 2016). Elbert et al. (2012) estimated a global average rate of 6 kg N ha⁻¹ y⁻¹, based on published values in the literature. In that same study, a total of 25.7 Tg y⁻¹ was estimated to be fixed by cryptogamic ground covers (mostly various types of biocrusts) in desert and steppe environments, which corresponds to about a quarter of the nitrogen fixation on land. Rodríguez-Caballero et al. (2018) arrived at a very similar estimate of global nitrogen fixation attributable to dryland biocrusts. In addition to fixing nitrogen, biocrusts harbor organisms that contribute to mineralization and other transformations of carbon, nitrogen, phosphorus and other nutrients (Delgado-Baquerizo et al., 2018; Weber et al., 2016).

Hydrological functions

Biocrusts have a pervasive influence on most soil processes pertaining to water (Eldridge et al., 2020). In some regions biocrusts slow infiltration and promote runoff, while in others they are reported to do the opposite (Weber et al., 2016). Biocrusts often retain more moisture than soils lacking biocrusts, but may concentrate it near-surface (Eldridge et al., 2020). Biocrusts may increase or impede evaporation rates. Biocrusts increase the capture of fog, dew and vapor, but the magnitude of this effect varies widely by ecosystem (Li et al., 2021). This wide array of influences comes about because biocrusts change many soil properties that affect water retention and transport such as temperature, texture, surface roughness, aggregation, pore size distribution and connectivity to the surface, and organic matter content (Belnap and Lange, 2003). In some patterned ecosystems, biocrusts are part of a system in which water is transported from runoff zones to run-on zones, boosting overall productivity of the ecosystem (Belnap and Lange, 2003).

Soil surface albedo and heat storage

Albedo refers to the proportion of solar radiation reflected from a surface, such as a soil. Albedo would generally be expected to be negatively related to the amount of heat absorbed and stored by surface soils. Biocrusts modify many soil surface properties such as darkness, roughness and moisture content, all expected to reduce albedo. Many, but not all, biocrusts darken as they develop and mature due to increasing prevalence of highly pigmented species, and disturbances and other stressors may reverse development and decrease darkness (Fig. 3C). Shifts in community composition induced by experimental climate change treatments, increased soil surface albedo by 30% through loss of darker taxa and replacement with lighter taxa (Rutherford et al., 2017). Other authors have focused on the effect of biocrust darkness and development on temperature. Mature cyanobacterial biocrusts can warm soil surfaces by as much as 10 °C, and the degree of warming can be very well predicted by the concentration of a single sunscreen pigment (Couradeau et al., 2016). Albedo and soil surface temperature do not always change in lockstep, however. Xiao and Bowker (2020) found that moss-dominated biocrusts consistently lower albedo year-round, especially when dry, but were only warmer than bare soils under dry conditions and were cooler under wet conditions.

Interactions with other biota

Because biocrusts modify many properties of the soil habitat, they also shape the structure and composition of other assemblages of biota. Biocrust influences on vascular plants are the most widely documented examples but are also extremely variable and contingent on the life history stage and traits of the vascular plants, in addition to the type of biocrust (Havrilla et al., 2019). Biocrusts may promote (Fig. 3D), inhibit or have neutral effects on a given plant species, and these effects may shift across plant ontogeny (e.g., germination vs. subsequent growth). These effects may be attributable to a wide variety of mechanisms including, but not limited to, prevention of radical penetration into soil, or concentration or redistribution of key soil resources such as water and nutrients. Rare studies of biocrust effects on entire plant communities have found that biocrust properties such as total cover, richness or relative moss dominance can influence key properties of plant communities such as native plant dominance over exotic plants (Bowker et al., 2022).

Though less widely studied, biocrusts also modify the habitat properties for soil bacteria and fungi, in addition to supplying potential food resources for these groups. As an example, in a study on three continents, the presence of biocrust mosses constrained the variability of microbial communities attributable to aridity, suggesting that the mosses buffer the microclimate experienced by soil microbes (Delgado-Baquerizo et al., 2018). Biocrust influences may be detectable more broadly in soil food webs, for example more successional developed biocrusts tend to support greater abundance and diversity of soil microfauna (Weber et al., 2016).

Multifunctionality

Ecologists are increasingly focusing on multifunctionality of ecosystems, or the ability of the ecosystem to carry out multiple functions simultaneously, as opposed to focusing on single ecosystem functions like productivity. Biocrusts are multifunctional, because they simultaneously contribute to many of the functions listed above and others. For example, working with samples from three continents, Delgado-Baquerizo et al. (2016) compared indicators of carbon, nitrogen and phosphorus cycling in biocrust patches compared to adjacent patches lacking biocrusts. All indicators were integrated into two types of multifunctionality index and the authors found that biocrusts enhanced multifunctionality in semi-arid and, especially arid environments, but not dry-subhumid or humid environments. The degree of multifunctionality of biocrusts has been linked to properties of the biocrust community, for example species richness, evenness or community composition (Bowker et al., 2017). Recent work suggests that global change factors, such as nitrogen deposition or drought, may strengthen the influence of biocrusts on multifunctionality (Dias et al., 2020; Su et al., 2021).

Threats and Recovery

Biocrusts are sensitive to both climate change and physical disturbances. Many of the ecosystems where biocrusts thrive are among the most threatened and disturbed as a result of human activities (Právělie, 2016). In addition, it is predicted that biocrust coverage will decrease substantially with land use and climate change activities within the 21st century (Rodríguez-Caballero et al., 2018).

Understanding the ways in which biocrusts are damaged and their subsequent recovery trajectories, natural or assisted, is necessary to predict ecosystem health for drylands and other ecosystems where biocrusts are common. Below we describe the effects of physical disturbance and climate change on biocrusts, and then discuss natural and assisted recovery for biocrusts in response to physical or climate disturbance.

Physical disturbance

In ecosystems in which abiotic stress is high, such as drylands, disturbances that break or compress the soil surface are detrimental to biocrusts and tend to have similar outcomes across ecosystems (Weber et al., 2016). These disturbances include land clearing and tillage, grazing (hoof action), military training (heavy vehicle traffic) and mining (surface soil removal). Of these, grazing is the most extensive land use in the world and is especially prevalent in drylands (FAO, 2020). Damage to biocrusts due to grazing is primarily an outcome of trampling, rather than consumption. The degree of damage sustained may vary among taxa with differing morphologies. The degree of damage sustained is also influenced by elements of the grazing regime, such as intensity or season; higher stocking rates tends to result in greater impacts on biocrusts, as does grazing during dry, warm seasons. Fire can also exert a detrimental effect on biocrusts, through combustion or exposure to heat. The degree of fire damage to biocrusts depends strongly on the fire intensity and patchiness, which is influenced by the amount, connectivity and type of fuel. Response of biocrusts to these varied disturbances are surprisingly similar in many cases (Weber et al., 2016). The most commonly reported changes include decreases in biocrust total abundance, and a simplified community structure consisting mostly of early successional community members such as cyanobacteria and a loss of later successional mosses and lichens (Ferrenberg et al., 2015).

The effect of the scale of disturbances is an understudied aspect of their impacts upon biocrusts. Small disturbances are likely to have limited effects because local propagule sources remain, allowing biocrusts from surrounding areas to colonize the disturbed area, assuming conditions are adequate for recovery. Large-scale disturbances could lead to a much slower recovery time because soil erosion could be initiated, leading to loss of soil fertility and other resources, alongside degradation of the vascular plant community. Recovery in these scenarios, compounded by a lack of local propagule sources, would hypothetically be much slower and may require active rehabilitation (Bowker, 2007).

In less stressful ecosystems (e.g., mesic regions), biocrusts may have a very different relationship to disturbance. Normally, these ecosystems would not provide much niche space for biocrusts owing to high vascular plant productivity and canopy closure. Disturbances can create an opportunity, at least temporarily, for biocrust establishment by creating open space in which sunlight penetrates to the exposed soil surface (Bowker, 2007). For example, biocrusts composed primarily of mosses or liverworts may proliferate after a forest fire. In most such cases, the biocrusts are a type of pioneer vegetation that will eventually be replaced by vascular plant cover, but can help to prevent soil, nutrient and water loss until vascular plants can recolonize (Bowker, 2007). In some cases, unusual soil conditions, such as excessive drainage or extreme shallowness, can interact with disturbance to allow for a more permanent presence of biocrusts in mesic ecosystems (Corbin and Thiet, 2020).

Climate change

Although biocrusts are adapted to life in harsh terrestrial environments, they can be stressed by environmental conditions experienced when active. Thus, if precipitation and temperature regimes shift, biocrusts may be vulnerable to climate change. Most regions on Earth are currently experiencing warming trends. Though precipitation trends are more varied, warming alone is often hypothesized to make dryland environments drier through enhanced evapotranspiration. Thus, a general expectation for dryland biocrusts might be that warmer temperatures or reduced rainfall would reduce the time when the surface soil is hydrated, thereby decreasing opportunities for biocrusts to grow. Experiments have employed various methods to induce warming such as the field installation of passive open-topped warming chambers with infrared heat lamps (Weber et al., 2016). Likewise, researchers have experimentally installed rain-reduction shelters to intercept a portion of the rainfall and induce drier conditions or have altered the pattern of precipitation (Weber et al., 2016). Elevational transplantation has been used to induce warmer and drier conditions simultaneously. To date, experimental studies investigating biocrust susceptibility to climate change suggest that: (1) Warming reduces the prevalence of at least some biocrust components, often lichens or mosses, (2) Effects of rainfall amount are more subtle, but altered rainfall frequency can induce the rapid mortality of mosses (Weber et al., 2016). For example, an experiment increasing temperature and modulating precipitation to come in small pulses led to mature biocrusts losing moss and lichen components, and resetting to an early successional filamentous cyanobacterial community (Ferrenberg et al., 2015). These experimental findings were supported in a long-term observational field study in which nitrogen-fixing biocrust lichens declined sharply in response to high temperatures during a severe drought and did not recover again over the next two decades; this study appears to capture detrimental climate change effects on biocrusts occurring in real time without experimental manipulation (Finger-Higgins et al., 2022). How extensive and impactful might effects of climate change be on biocrusts? A global modeling study projected that due to the synergistic effects of climate change and expected land use change, the global land area suitable for biocrusts may decline by 16–33% by 2070, while actual biocrust coverage might decline by 27–39% (Rodríguez-Caballero et al., 2018). These biocrust losses are estimated to impact global net primary production only a small amount, but could reduce global N-fixation on land by 6–9% (Rodríguez-Caballero et al., 2018), and increase global dust emission by 5–15% (Rodríguez-Caballero et al., 2022). Potential impacts to other biocrust-mediated ecosystem functions have not yet been estimated at the global scale.

Natural recovery from disturbance

Biocrusts can recover from disturbance, but time required for recovery has been estimated at anywhere between a few years to a few millennia (Weber et al., 2016). Biocrust recovery is fastest in climates with greater effective precipitation and greater inherent stability of soils and when disturbances are less severe. Most of the very slow recovery rate estimates of centuries to millennia are associated with arid or hyper-arid conditions, extreme disturbance types, such as tank training or complete removal, or both. Excluding these extreme cases, most estimates seem to suggest that biocrusts typically recover from disturbance in several years to a few decades, depending on the factors listed above. The operational definition of recovery can also be a complicating factor. Recovering disturbed biocrusts may attain an appearance similar to undisturbed biocrusts relatively fast, but the full recovery of biomass, diversity, composition and multifunctionality may take much more time. Another consideration when estimating recovery rate is that recovery is likely not to be linear, but instead works in pulses when climate conditions are good for growth, interspersed with long periods of slow or static growth (Weber et al., 2016).

The barriers to natural recovery are hypothesized to be hierarchical, and include active soil erosion, resource limitations, and propagule limitations; as these barriers are lifted, recovery may occur (Bowker, 2007). Biocrusts are often described as following a successional sequence as they recover. Light-pigmented cyanobacterial communities are often among the early colonizers, followed by dark-pigmented cyanobacterial communities, followed by either mosses, lichens or both (Weber et al., 2016). For the most part, this sequence is a series of addition of organisms rather than complete replacement, and the sequence may only attain the later stages where the environment permits later successional forms. Though this successional sequence is commonly reported from disparate localities, it is not universal. In some locations mosses are early successional taxa, whereas cyanobacteria can be late successional. Further, biocrusts of some ecosystems appear to have alternative states and successional pathways depending on the specific nature of the disturbance regime. It must be noted that most of what we know about natural recovery and successional pathways comes from studies that are not able to observe the complete pathway, so estimates of time to recovery and true trajectories are often incompletely known.

Assisted recovery

Assisted recovery of biocrusts is an active area of interest and research in which considerable advances have been made over the last 20 years (Weber et al., 2016; Antoninka et al., 2020). The goal of assisted recovery is to increase the rate of biocrust recovery and associated ecosystem functions through human intervention. Interventions can include methods to stabilize soils and retain biocrust propagules, enhance nutrient availability, increase water availability, propagule cultivation and subsequent addition, and combinations of the above. Interventions should be guided by specific recovery targets as well as the economic and technical constraints of the activities. Further, assessing the site conditions and addressing the barriers to recovery are critical first steps to ensure that the best interventions are chosen. Considerable progress has been made in methods to stabilize soil and retain propagules on site, as well as methods to cultivate biocrusts in the laboratory, greenhouse and field. Cultivation allows for biocrust inoculum to be produced on demand with lesser reliance on sourcing material from the field; this step is critical to the sustainability of inoculation interventions. Major barriers still to be fully addressed include success of cultured organisms in the field, storage of propagules, scaling up the areas that can be treated, and finally, guidance for land managers.

Management considerations

As our understanding of biocrusts as integral ecosystem components grows, biocrusts can increasingly be considered in land management plans. Key steps to integrating biocrusts into management activities include:

1. Monitoring the abundance, community, condition and function of biocrusts as indicators for the health and trend of landscapes (Belnap and Lange, 2003).
2. Identifying trends in the monitoring data to guide decisions such as site selection for conservation, restoration or construction projects or to trigger shifts in management strategies.
3. Educating the public on the importance of biocrusts and managing their land use behavior, for example limiting off-highway vehicle recreation to designated areas.

The requisite data for monitoring biocrusts across landscapes is becoming more widely available to support management activities. While remote sensing of biocrusts still faces many challenges due to the presence of vegetative overstories with similar reflectance, advances are being made in both coarse, large-scale monitoring with satellite imagery and finer-scale monitoring with unmanned aerial systems (Havrilla et al., 2020). Ground-based monitoring of biocrusts is also increasing, as with the US Bureau of Land Management's Assessment, Inventory and Monitoring (AIM) Strategy which, alongside other indicators, includes multi-year data on the abundance of biocrusts on soils (Herrick et al., 2021) and has been implemented across the western US. As biocrust monitoring at multiple scales becomes more widespread, land managers will have increasing access to important data that can guide their decision-making process.

Because they are a sensitive element of landscapes and ecosystems, loss or gain of biocrusts can signal the success or failure of a management or land use strategy to meet its goals. For example, large-scale soil conservation efforts, such as the Grain for Green

program in China, have found that altering land management regimes can dramatically reduce soil erosion (Gao et al., 2017). A major reason for and an excellent indicator of this landscape transformation was the widespread proliferation of biocrusts which strongly reduced erosion rates (Gao et al., 2017). Public land managers in the US are also starting to include the return of biocrust cover as a metric of successful recovery after disturbance. As more biocrust research focuses on land management decision support, the breadth of applications for biocrust monitoring in management plans may increase.

Educational outreach about biocrusts has taken many forms including interpretive signage in biocrust-rich recreation areas, workshops targeted toward land managers and recreation industry professionals, and programming in elementary schools. While these relatively targeted methods are effective at reaching their audience, more work is needed to ensure public support for incorporating biocrusts in land management. López-Rodríguez et al. (2020) call for the biocrust research community to adopt a “social-ecological research perspective” which prioritizes inter-disciplinary collaboration between researchers and non-academic stakeholders to advance the use of biocrust in conservation practices.

Conclusion

Biocrusts are a globally-distributed ecological community type composed of cryptogams and microbes found directly on the soil surface creating a distinctive thin, aggregated layer that is also photosynthetic. They are currently estimated to cover 12% of the land surface. Biocrusts are multifunctional, contributing to a wide variety of ecosystem functions such as soil stabilization, accumulation and maintenance of fertility, interception, retention and redistribution of water, among others. They interact with and influence the development and composition of other co-occurring assemblages such as those of vascular plants and soil microfauna. Thus, biocrusts are both widespread and highly influential to the functional state of the ecosystems that they occupy.

Despite the breadth of their niche and their impressive abilities to tolerate key abiotic stresses such as desiccation and UV exposure, they are vulnerable to physical disturbance, either due to natural disturbance regimes or human land use. Biocrusts are also vulnerable to climate change. Simultaneously, most ecosystems are warming and experiencing greater water evaporation from soils. Such changes are likely to impact biocrusts negatively by reducing their hydration time and therefore periods of activity. To reduce some of these negative effects on biocrusts and their functional contributions to ecosystems, it is necessary to consider the impacts on biocrusts in land management decision making. Currently, development of techniques to assist recovery of biocrusts after a disturbance is a major research theme. Researchers are learning to grow biocrusts for re-entry into degraded ecosystems to improve ecosystem function.

Since the early 2000s biocrust research is growing at a fast rate and becoming increasingly international as more scientists recognize the distinctive attributes and importance of these communities and their vulnerabilities. Knowledge gaps remain. For example, there are geographic gaps in our knowledge of biocrusts to fill and we do not currently have a complete understanding of the global roles of biocrusts in regulating the Earth system nor how to incorporate them into models. Nevertheless, new knowledge is accumulating quickly as new tools and perspectives are being applied (Rodríguez-Caballero et al., 2022), and each year we know substantially more than the last.

References

- Antoninka A, Faist A, Rodríguez-Caballero E, et al. (2020) Biological soil crusts in ecological restoration: Emerging research and perspectives. *Restoration Ecology* 28: S3–S8.
- Belnap J and Lange OL (eds.) (2003) *Biological Soil Crusts: Structure, Function, and Management*. Berlin, Germany: Springer-Verlag.
- Belnap J, Wilcox B, Van Scoyoc MW, and Phillips S (2008) Successional stage of biological soil crusts: An accurate indicator of ecohydrological condition. *Journal of Arid Environments* 72: 1257–1264.
- Bowker MA (2007) Rehabilitation of biological soil crusts in theory and practice: An underexploited opportunity. *Restoration Ecology* 15: 13–23.
- Bowker MA, Büdel B, Maestre FT, Antoninka AJ, and Eldridge DJ (2017) Bryophyte and lichen diversity on arid soils: Causes and consequences. In: Steven B (ed.) *The Biology of Arid Soils. Life in Extreme Environments Series*, pp. 73–96. Berlin, Germany: DeGruyter.
- Bowker MA, Reed SC, Maestre FT, and Eldridge DJ (2018) Biocrusts: The living skin of the Earth. *Plant and Soil* 419: 1–7.
- Bowker MA, Doherty KD, Antoninka AJ, et al. (2022) Biocrusts influence community plant development, promoting native plant dominance. *Frontiers in Ecology and Evolution* 10: 840324.
- Corbin JD and Thiet RK (2020) Temperate biocrusts: Mesic counterparts to their better-known dryland cousins. *Frontiers in Ecology and the Environment* 18: 456–464.
- Couradeau E, Karaoz U, Lim HC, et al. (2016) Bacteria increase arid-land soil surface temperature through the production of sunscreens. *Nature Communications* 7: 10373.
- Delgado-Baquerizo M, Maestre FT, Eldridge DJ, et al. (2016) Biocrust-forming mosses mitigate the negative impacts of increasing aridity on ecosystem multifunctionality in drylands. *New Phytologist* 209: 1540–1552.
- Delgado-Baquerizo M, Maestre FT, Eldridge DJ, et al. (2018) Biocrust-forming mosses mitigate the impact of aridity on soil microbial communities in drylands: Observational evidence from three continents. *New Phytologist* 220: 824–835.
- Dias T, Crous CJ, Ochoa-Hueso R, Manrique E, Martins-Loução MA, and Cruz C (2020) Nitrogen inputs may improve soil biocrusts multifunctionality in dryland ecosystems. *Soil Biology and Biochemistry* 149: 107947.
- Elbert W, Weber B, Burrows S, et al. (2012) Contribution of cryptogamic covers to the global cycles of carbon and nitrogen. *Nature Geoscience* 5: 459–462.
- Eldridge DJ, Reed SC, Travers S, et al. (2020) The pervasive and multifaceted influence of biocrusts on water in the world's drylands. *Global Change Biology* 26: 6003–6014.
- FAO (2020) *FAO Statistical Database (FAOstat)*. <http://www.fao.org/faostat/en/#home>.
- Fernandes VMC, Giraldo-Silva A, Roush D, and Garcia-Pichel F (2021) Coleofasciculaceae, a monophyletic home for the *Microcoleus steenstrupii* complex and other desiccation-tolerant filamentous cyanobacteria. *Journal of Phycology* 57: 1563–1579.
- Ferrenberg S, Reed SC, and Belnap J (2015) Climate change and physical disturbance cause similar community shifts in biological soil crusts. *PNAS* 112: 12116–12121.
- Finger-Higgins R, Duniway MC, Fick S, et al. (2022) Decline in biological soil crust N-fixing lichens linked to increasing summertime temperatures. *PNAS* 119: e2120975119.

- Gao L, Bowker MA, Xu MX, et al. (2017) Biological soil crusts decrease erodibility by modifying inherent soil properties on the Loess Plateau, China. *Soil Biology & Biochemistry* 105: 49–58.
- Havrilla C, Chaudhary V, Ferrenberg S, et al. (2019) Towards a predictive framework for biocrust mediation of plant performance: A meta-analysis. *Journal of Ecology* 107: 2789–2807.
- Havrilla CA, Villarreal ML, DiBiase JL, Duniway MC, and Barger NN (2020) Ultra-high-resolution mapping of biocrusts with unmanned aerial systems. *Remote Sensing in Ecology and Conservation* 6: 441–456.
- Herrick JE, Van Zee JW, McCord SE, et al. (2021) Monitoring manual for grassland, shrubland and savanna ecosystems. In: *Volume 1: Core Methods*, 2nd edn. Las Cruces, New Mexico: USDA - ARS Jornada Experimental Range.
- Jung P, Briegel-Williams L, Simon A, Thyssen A, and Büdel B (2018) Uncovering biological soil crusts: Carbon content and structure of intact Arctic, Antarctic and alpine biological soil crusts. *Biogeosciences* 15: 1149–1160.
- Li S, Bowker MA, and Xiao B (2021) Biocrusts enhance non-rainfall water deposition and alter its distribution in dryland soils. *Journal of Hydrology* 595: 126050.
- López-Rodríguez MD, Chamizo S, Cantón Y, and Rodríguez-Caballero E (2020) Identifying social-ecological gaps to promote biocrust conversion actions. *Web Ecology* 20: 117–132.
- Mallen-Cooper M and Eldridge DJ (2016) Laboratory-based techniques for assessing the functional traits of biocrusts. *Plant and Soil* 406: 131–143.
- Právělie R (2016) Drylands extent and environmental issues. A global approach. *Earth-Science Reviews* 161: 259–278.
- Rodríguez-Caballero E, Belnap J, Büdel B, et al. (2018) Dryland photoautotrophic soil surface communities endangered by global change. *Nature Geoscience* 11: 185–189.
- Rodríguez-Caballero E, Stanelle T, Egerer S, et al. (2022) Global cycling and climate effects of aeolian dust controlled by biological soil crusts. *Nature Geoscience* 15: 458–463.
- Rutherford WA, Painter TH, Ferrenberg S, et al. (2017) Albedo feedbacks to future climate via climate change impacts on dryland biocrusts. *Scientific Reports* 7: 44188.
- Su Y, Liu J, Zhang Y, and Huang G (2021) More drought leads to greater significance of biocrusts to soil multifunctionality. *Functional Ecology* 35: 989–1000.
- UNEP-WCMC (2007) A spatial analysis approach to the global delineation of dryland areas of relevance to the CBD Programme of Work on Dry and Subhumid Lands. <https://resources.unep-wcmc.org/products/60440a42a27643c9ad50d024617f451f>.
- Weber B, Büdel B, and Belnap J (eds.) (2016) *Biological Soil Crusts: An Organizing Principle in Drylands*. Switzerland: Springer.
- Weber B, Belnap J, Büdel B, et al. (2022) What is a biocrust? A refined, contemporary definition for a broadening research community. *Biological Reviews* 97: 1768–1785.
- Xiao B and Bowker MA (2020) Biocrusts strongly decrease soil surface albedo, altering land-surface energy balance in a dryland ecosystem. *Science of the Total Environment* 741: 140425.

Studying soil microbial communities: Culture-dependent approaches

Eduardo K Mitter, Dasiel Obregon, Micaela Tosi, Heather Slinn, and Kari Dunfield, School of Environmental Sciences, University of Guelph, Guelph, ON, Canada

© 2023 Elsevier Ltd. All rights reserved.

Introduction	140
Microbial culturing and culture media	140
Enumeration of viable microbial populations: Plate count and most probable number	141
Community-level physiological profiling	142
Single-cell manipulation and isolation	142
Flow cytometry	142
Optical tweezers	142
Diffusion chambers	143
Culturomics	143
Concluding remarks	145
References	145

Abstract

One of the major challenges in microbial ecology is the inability to culture most microorganisms under laboratory conditions. Based on comparisons with culture-independent techniques, it is thought that less than 1% of microbial species has been isolated through standard cultivation techniques. Nonetheless, cultivation remains the only way to amplify the microbial material, fully characterize their functions and metabolism, predict their impact on an environment and validate phenotype predictions based on genomic analysis. This chapter describes traditional and novel culture-dependent techniques to isolate, grow and study soil microorganisms.

Glossary

Community-level physiological profiling (CLPP) A rapid method to assess the metabolic potential of a microbial community by simultaneously testing their growth with different carbon sources (i.e., substrate utilization patterns).

Culture medium Solution that contains all the necessary nutrients required to grow specific microorganisms or groups of microorganisms.

Culture-dependent methods All microbiological methods based on growing microorganisms by incubating them in a culture medium under controlled environmental conditions.

Culturomics High-throughput culturing of microorganisms based on testing hundreds of culture conditions to comprehensively identify strains or species in complex samples.

Diffusion chambers Devices to grow microorganisms in their natural environment by restricting their movement but allowing the diffusion of chemicals to and from the exterior.

Dilution to extinction Process where a sample is diluted until, on average, there is only a single viable cell left in a given dilution.

Flow cytometry A fluorescence-based method used to identify and quantify different cell types in a solution of heterogeneous cells.

Matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) Technique used to identify microbes based on the comparison of their protein mass spectra with a database of known microbes and their protein mass spectral signatures.

Microbial culturing Process of growing microorganisms in a lab setting by inoculating them onto a previously sterilized culture medium.

Microfluidics Technology involving the manipulation of small volumes of fluids (e.g., in the range of microliters to picoliters) using channels on the scale of tens to hundreds of micrometers.

Most probable number (MPN) An indirect method to quantify viable microbial cells based on the serial dilution of a sample and the presence of growth at each dilution level.

Optical tweezers Systems consisting of a neodymium laser tightly focussed to a tiny region in space using a microscope objective as a lens.

Plate count A method based on the cultivation of microorganisms on a solidified culture medium in order to immobilize microbial cells and count the number of colony forming units (CFUs).

Pure culture A microbial culture containing a single kind of microorganism, which can be isolated from an environmental sample.

Single-cell manipulation Techniques used to selectively isolate individual cells, useful for isolating slow-growing or low-abundance microorganisms that would not be detected in enrichment cultures.

Key points

- Culture-dependent techniques applied to soil microbiology
- Basics of culture media (types and uses)
- Quantification of viable microorganisms
- Functional or physiological profiling
- Manipulation of single cells
- Improved culturing conditions and isolation via diffusion chambers
- High-throughput culturing or culturomics

Introduction

Microbial culturing techniques are among the oldest tools to study soil microbial communities, and they have revolutionized biology and biotechnology, leading to significant advances in fields such as medicine, waste management, and food science. Culture-dependent techniques have played an important role in identifying and characterizing new microbial species but are still low-throughput and time-consuming processes. One of the major challenges in microbial ecology is the inability to culture most microorganisms under laboratory conditions, as only a small fraction (<1% of species) has been isolated through standard cultivation techniques. This limitation leads to a phenomenon known as the “great plate count anomaly,” which refers to the relatively reduced number of microorganisms from a natural sample that can be grown in any culture medium compared to the cell counts observed using microscopy (Madigan et al., 2021). One of the reasons behind this limitation is the failure to mimic their natural growth environment and the lack of information regarding specific requirements in terms of nutrients, temperature, pH, and other environmental factors. Other reasons why microorganisms cannot be cultured in isolation are related to their need, in some cases, for the presence of another organism to survive (e.g., obligate symbionts). In environmental samples, microorganisms can also be present in very low abundance or in a dormant-like state, known as viable but nonculturable (VBNC), where visible growth is not achieved even if conditions are suitable for that particular organism.

In recent decades, genomics studies revealed that unculturable microorganisms represent a large pool of novel species and biotechnological potential. Omics tools have an enormous potential to explore the structure and function of complex microbial communities, but they still present challenges due to the complexity of the obtained datasets and the considerable cost involved in carrying out multiple omics analyses. Culturing still remains the only way to fully characterize microbial functions and metabolism, validate phenotype predictions based on genomic analysis and generate microbial biomass. Realizing the crucial importance of bringing more microbial species into cultivation, different techniques have marked the rebirth of culture-dependent approaches by trying to better understand the specific requirements of each microorganism and mimic their natural growth conditions.

Microbial culturing and culture media

Microorganisms can be grown in the laboratory by inoculating them on a previously sterilized culture medium while maintaining aseptic conditions during the whole process. Culturing allows to obtain pure cultures by isolating and growing cells from a single microorganism. Culture media are solutions that include all the nutrients necessary to grow a particular microorganism or group of microorganisms, and they can be liquid or solid. Solid media are obtained by adding agar to the solution and they are typically formed in a Petri dish. These media have the advantage of immobilizing microbial cells, allowing them to form visible growth colonies which can be used for characterization and counting, as well as isolation. In contrast, liquid media do not permit separate colonies to be distinguished but they allow for rapid growth and hence can be used for enrichment or resuscitation. A broad classification of media separates them into those that are defined, which means they contain known amounts of specific organic or inorganic chemicals, and those that are complex or made from digests of microbial, animal or plant products (e.g., yeast extract, beef extract and tryptic soy broth, respectively). Minimal media are those that contain the minimum amount of nutrients possible for growth and are useful for growing oligotrophic organisms that would not be able to grow in a nutrient-rich medium. Previous information from targeted organisms must be known to determine their minimal media requirements. In terms of the organisms able to grow in them, culture media can be general (i.e., allowing the growth of several different species) or selective, tending toward the requirements for a specific organism, promoting its growth but inhibiting or preventing the growth of undesired organisms (e.g., using antibiotics or other selective agents, or via changes in pH or osmotic conditions). Enrichment media, on the other hand, are liquid media that promote a particular type of organism from a complex mixture (e.g., soil sample) with or without the use of inhibitors (selective or elective, respectively). A special type of solid culture medium, known as differential, allows the distinguishing of different microorganisms growing together by producing specific colony morphology (e.g., color). This is usually achieved by including an indicator (e.g., pH indicator) that responds to changes in the medium due to different substrate utilization or production of specific compounds. Besides the culture medium, other important factors that are taken into consideration during microbial culturing are growth conditions, such as incubation temperature, oxygen levels, pH and osmotic conditions.

Enumeration of viable microbial populations: Plate count and most probable number

Enumeration methods to determine viable soil microorganisms may be accomplished by direct (e.g., standard plate count) or indirect (e.g., most probable number or MPN) techniques. Yet, it is important to emphasize that bacteria classified as viable in reality only include those capable of growing on the media used and with the incubation conditions employed. The fundamental principles of both these methods are (i) creating a series of dilutions of the original sample, (ii) dispensing an aliquot to a growth medium, (iii) incubating under appropriate conditions, and (iv) counting the developed colonies or detecting the presence or absence of growth (Germida and de Freitas, 2008). Both methods are inexpensive and can give a qualitative or quantitative assessment of the number of viable microorganisms in a sample. However, these techniques are often time consuming and rely on morphological (e.g., colony shape and size) and biochemical (e.g., Gram stain) methods, which may have a very severe bias. Therefore, expanding the repertoire for microbial identification also relies on other methods based on carbon source utilization, synthesis of lipids, and oxygen requirements.

The standard plate count method uses Petri dishes with solid culture media, which allow to count the number of colony forming units (CFUs) per milliliter of diluted sample plated (Pelczar et al., 2002). The term CFU is used because two or more cells can be clumped together forming a single colony, even though the assumption is that each colony is formed by a single organism. Plate counts can be done using the spread-plate method (Fig. 1), where the diluted sample is spread on solidified agar, or the pour-plate method, where molten agar is poured on top of the sample. To avoid errors, the number of colonies cannot be too large or too small, so a standard practice is to prepare and plate several 10-fold serial dilutions of the original samples and count only plates that have 30–300 colonies (Fig. 1).

The MPN method is used to estimate microbial population sizes when counting single viable cells or colonies is not possible. The liquid culture media used in MPN allow a greater nutrient exchange and is more suitable for specific groups of microorganisms that will not readily grow on solid agar (e.g., a variety of protozoa, autotrophic-nitrifying and sulfur-oxidizing bacteria) (Germida

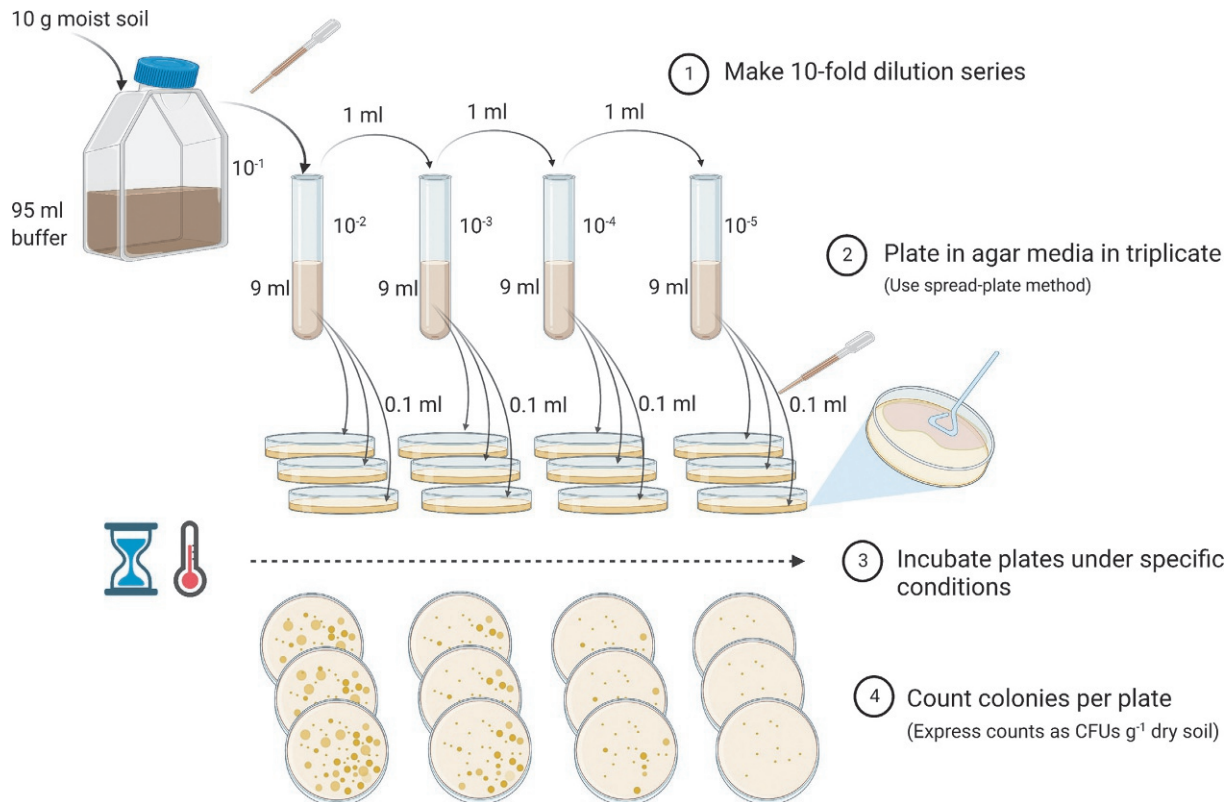


Fig. 1 Representative diagram of the spread-plate method. In this technique, the soil sample is diluted (1:100) in a specific buffer [often phosphate buffered saline (PBS) is used], and then subject to 10-fold serial dilutions ①. Each dilution is then plated in solid culture medium (in duplicate or triplicate) ② using a specific agar medium based on the type of microorganism to be quantified. An inoculum volume of 0.1 mL should be used to ensure spreading, absorption, and proper counting. The inoculum should be spread shortly, using flame-sterilized glass or metal spreader. After incubating under controlled conditions ③, the plates should reflect the predictable drop in colonies number based on the serial dilutions. Plates from dilutions yielding 30–300 colonies should be counted, and the results expressed as CFUs g^{-1} of dry soil (colony forming units per gram of dry soil) ④. Adapted from Pepper IL and Gerba CP (2015) 'Cultural methods', in Pepper IL, Gerba CP, and Gentry TJ (eds) *Environmental Microbiology*. 3rd edn. Academic Press, pp. 195–212, created with BioRender (<https://biorender.com/>).

and de Freitas, 2008). This method also starts with serial dilutions of the initial sample, and each dilution is inoculated in liquid medium with 5–10 replicates. Each tube is classified as a positive or negative based on the presence of growth (e.g., turbidity, gas, substrate utilization) and the number of positive/negative tubes per dilution is used to calculate using publicly available MPN tables (Woomer, 1994). Some disadvantages are that it is time-demanding and less precise than plate counting.

Community-level physiological profiling

The main principle of community-level physiological profiling (CLPP) is to characterize specific organisms or groups of organisms based on their ability to metabolize different substrates. Microorganisms are cultivated on microplates containing different carbon sources (e.g., carbohydrates, amino acids, amines, carboxylic acids, and polymers) and wells without a growth substrate are used as a control (Garland, 1997). The technique uses a redox system, where tetrazolium violet is frequently used as an indicator of oxidation. During the period of incubation, soil microorganisms oxidize substrates in the plate wells, while simultaneously reducing the colorless tetrazolium dye to a violet formazan. These color changes are measured spectrophotometrically, based on absorbance (i.e., optical density), using an automated plate reader that interprets the results and compares them with a known database. The inoculation of soil solutions directly into plates allows for a rapid and convenient tool to obtain physiological profiles of whole microbial communities (Weber and Legge, 2010). Yet, it is important to note a few limitations. For example, due to cross-species interactions and different nutrient requirements, it is impossible to know precisely, which groups of microorganisms contribute to substrate use on a plate. Moreover, there are biases among the different abilities of bacteria to use the substrates, where non-culturable or slow-growing microorganisms, incapable of growing in such conditions, are not included in the analysis. Other variants of this methodology, including culture-independent ones, are described in Chapter 00115 Studying soil microbial function: Culture-independent approaches.

Single-cell manipulation and isolation

In the techniques mentioned above, isolation of targeted organisms is most often achieved by dilution to extinction. This is where a sample is diluted until, on average, there is only a single viable cell left in a given dilution. In single cell manipulation systems, individual cells may be selectively isolated, rather than leaving the choice of cells being investigated to chance. This is particularly useful for isolating slow-growing microorganisms or organisms present at such low numbers that would not be detected in enrichment cultures.

Flow cytometry

Flow cytometry (FCM) is a well-established fluorescence-based method in medical science that is used to identify different cell types in a solution of heterogeneous cells. In microbial ecology, FCM has been gaining significant attention as a direct and rapid method to assess soil bacterial abundance and analyzing microbial cells at the single-cell level (Frossard et al., 2016). The method first separates microbial cells from soil particles by: (i) detaching the cells by mechanical (i.e., shaking or sonication) or enzymatic treatments, and (ii) separating these cells from soil organic and inorganic particles (i.e., by filtration or centrifugation) before quantification. Then, microbial cells are stained with fluorescent markers (e.g., SYBR Green and acridine orange) and analyzed in a flow cytometer so that the light is absorbed and then emitted in a band of wavelengths (Khalili et al., 2019). Hence, an efficient separation method maximizing cell recovery while reducing interference from background debris is critical to the procedure. A flow cytometry analyzer can distinguish background noise signal caused by non-biological particles, and it uses relative size, fluorescence intensity and wavelength to allow the quantification of many thousands of cells per second. Although the fluorescent signal can be masked by nonspecific dye binding and auto-fluorescence, FCM is more accurate than epifluorescence microscopy-based (EPM) enumeration in terms of human biases, detection of small cells (<0.5 μm) or cells hidden under soil particles. Moreover, it is less time- and labor-intensive, which facilitates the analysis of larger sample sizes (Deng et al., 2019).

Optical tweezers

The main mechanisms of optical tweezer (OT) systems consist of a neodymium laser tightly focussed to a tiny region in space using a microscope objective as a lens. This region becomes an optical trap for small objects such as a biological cells or organelles. Trapping a single cell is possible because the laser beam creates a force that pushes down on a cell and holds it in place (Eriksson et al., 2007). The system may also contain a glass capillary tube with a predetermined breaking point (i.e., notches in the tube wall to determine the breaking location), which is used as a separation chamber between a suspension of mixed microbial population and a fresh sterile medium. The capillary creates a linear microchannel into which single cells could

be optically manipulated. When the laser beam is moved by a computer-controlled microscope stage, the trapped cell moves along with it. With this movement, a single cell from the mixed culture is then separated into the fresh medium section of the tube. The capillary is then broken at the predetermined breaking point, where the single cell is transferred to a new tube of sterile medium (Madigan et al., 2021). This technique therefore allows microbes to be cultivated by eliminating the competition of faster growing organisms. Additionally, the forces applied in the optical trap can be measured as well, which opens up a broad range of possibilities in detecting the environmental effects on microbial cells, such as cell elasticity and flagellar motion. For example, a few studies have combined OT with microfluidics to move microbial cells from one microenvironment to another. The cells were exposed to these different environments, where the adaptation to changes in nutrient concentrations or responses to osmolarity can then be studied (Kelothe et al., 2018).

Diffusion chambers

One effective strategy to improve microbial cultivation methods is to mimic their natural environment in the laboratory. For this purpose, diffusion chambers allow microorganisms to grow into pure culture by using the required chemical components from their natural environment (Bollmann et al., 2007). First, soils are serially diluted, mixed with sterile agar and inoculated between two semipermeable membranes. The chambers are then placed on the same soils from which dilutions were carried out and moistened if necessary. The membranes are permeable to molecules of various sizes, thus allowing for the diffusion of necessary elements from the soil into the inner chamber, but also acts as a barrier restricting the passage of microbial cells from the chamber to the external environment (Pham and Kim, 2012). This permits microbial cells trapped in the inner chamber to have access to their naturally occurring growth components (e.g., oxygen, nutrients, and metabolites) and provides a natural environment for microbe-microbe interactions to occur. After incubation, diffusion chambers are opened, the biomass is collected and, if necessary, transferred to new sterile chambers to enhance recovery of environmentally relevant isolates.

Diffusion chambers improve microbial recovery and *in vitro* cultivation of otherwise poorly cultivable species when compared to conventional plating. But one limitation is its relatively small throughput, producing hundreds of microcolonies from multiple species which needs laborious separation under a microscope. Therefore, the isolation chip or “ichip” has been designed based on hundreds of miniature diffusion chambers (Nichols et al., 2010). Each diffusion “minichamber” is loaded with a single cell that results in a culture that is monospecific. Yet, for growth, some microbial species require a close proximity to their synergistic partner(s), such as hydrogen-producing bacteria and methanogenic archaea. Thus, having these two partners on different sides of the membrane will compromise this synergy and the recovery of these species. On the other hand, the simplicity of the ichip construction, the ease of its application and the high recovery of microorganisms are great advantages of this method. After one round of *in situ* cultivation, the colonies that remain “uncultivable” can then be passed through more rounds, with each cycle leading to isolation of additional species (Berdy et al., 2017).

Culturomics

Culturomics identifies microbial species using high-throughput culturing with hundreds of culture conditions and matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) (Fig. 2) (Lagier et al., 2012). This technique emerged from studies on the human gut microbiota, with the aim of optimizing culture conditions of “yet to be cultured” organisms and to address the shortcomings of metagenomics, where some rare taxa do not have adequate sequencing coverage to be detected (Lagier et al., 2012). In addition, culturomics yields information about live microbial communities, whereas metagenomics is unable to distinguish between DNA from live or dead organisms. Culturomics led to the recovery of microbes from environments that were thought to be uninhabitable, and it is now expanding to other fields such as microbial ecology (Sood et al., 2021).

The first step of culturomics is to divide an environmental sample into different culture conditions (e.g., culture media, antibiotics and oxygen levels) and then generate pure cultures of microbes (Fig. 2). These conditions are designed to suppress fast growers and improve the isolation of fastidious microorganisms. Once isolates are obtained, the next step is identification, where MALDI-TOF mass spectrometry is used as a simple and fast (<1 h) method that relies on the comparison of the isolate’s protein mass spectra with a database of microbes and their protein mass spectral signatures. If this identification fails, colonies may be analyzed through ribosomal RNA sequencing. Finally, if there is less than 98.65% similarity to the closest official strain, the isolate could be identified as a putative novel species. However, one important limitation of culturomics is the major laboratory workload and the inability to test as many samples at the time as other methods (e.g., metagenomics), although, in some cases, the number of culture conditions can be optimized.

Since culturomics is an approach based on the recovery of the greatest number of microbial species present in the sample, the results obtained can only be qualitative but not quantitative. However, this limitation can be easily overcome by combining it with quantitative PCR. Moreover, culturomics is highly complementary to metagenomics, where one can enhance the output of the other (Table 1). Together they can be used to describe the composition of different microbiomes more exhaustively (Lagier et al., 2018).

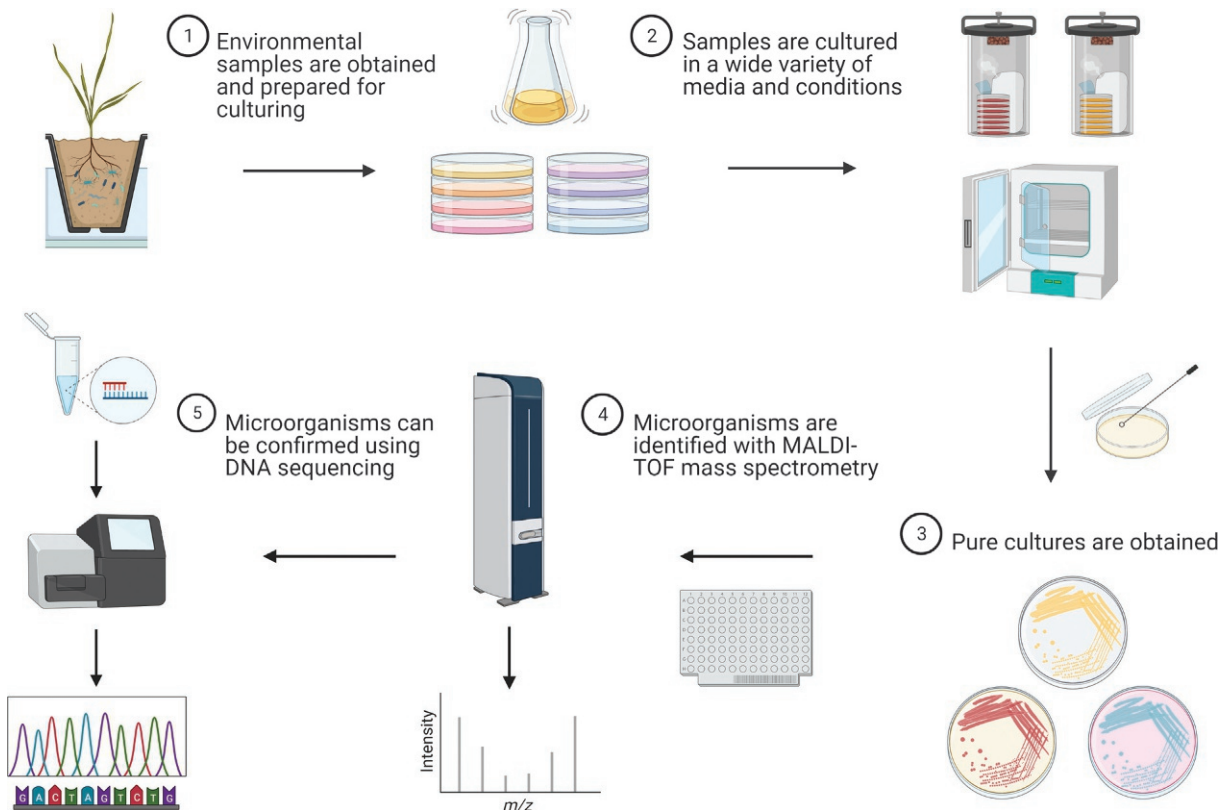


Fig. 2 Overview of culturomics pipeline from microbial sampling to identification. Culturomics uses high-throughput culturing of microorganisms obtained from environmental samples ① on many different types of media and with many different culture conditions ②, to maximize their recovery. In the original methods, microorganisms are isolated to obtain pure cultures ③ so that they can be identified using protein spectra generated by MALDI-TOF mass spectrometry ④. Protein spectra generated are compared to a database of spectra correlated with known microorganisms. Mass spectrometry is a much faster and much cheaper way to identify microorganisms when there are hundreds to thousands of pure cultures to identify. However, if spectra are unidentified in the database and a new species is suspected, sequencing can be used to confirm the identification ⑤. Created with BioRender (<https://biorender.com/>).

Table 1 Limitations and advantages of metagenomics approaches vs. culturomics.

	Metagenomics	Culturomics
Limitations	DNA extraction: extraction protocols differ in cell lysis efficiency with a major impact on the recovery of genomic content from difficult-to-lyse organisms. These limitations could lead to significant differences in diversity and taxonomic distribution Amplification (amplicon sequencing): highly dependent on the choice of primers and on the efficacy of DNA polymerases (e.g., samples with high GC content are not amplified as efficiently) Bioinformatics: variations in the methodology used (pipelines, clustering, taxonomic assignment) and difficulty in discriminating between species due to short amplicon length (amplicon sequencing) Viability: failure to discriminate if the sequences belong to active, quiescent or dead microorganisms Depth: difficulty in detecting microorganisms at very low concentrations (depth bias) Low taxonomic resolution: cannot identify novel species and often relies on comparisons with databases of known genomic sequences of microbial cultures	Sample material: steps to divide the sample and establish different culture conditions may require more sampling material to work with Recovery: the discovery of a novel or “not yet cultivable” species may still not be achieved despite the different conditions established (e.g., novel species that could only grow in the presence of other community members) Growth conditions: can still be difficult to source or formulate. Dealing with multiple organisms at a time with different growth rates and nutrient requirements can be challenging Time: major laboratory workload is often required, which limits the ability to test many samples at a time Quantification: results are only qualitative and not quantitative since it targets the maximum recovery of microbial species Function and metabolism: does not provide information on enzymatic capabilities, only taxonomic identification

Table 1 (Continued)

	Metagenomics	Culturomics
Advantages	<i>Processing a large number of samples:</i> allows the analysis of thousands to tens of thousands of genes or gene regions simultaneously <i>Relatively easy to assign taxonomic classification:</i> uses several validated tools and curated databases (variety of software available for analysis) <i>Detection of uncultured microbes:</i> metagenomics considers both culturable and unculturable organisms where the latter can be grouped without taxonomic assignment <i>Time:</i> sample preparation, analysis and comparison of large cohorts in record time <i>Functional Analysis:</i> the function of microbial communities can be inferred by shotgun sequencing	<i>Detects only viable bacteria:</i> avoids false discovery rates and over-representation of species by the identification of DNA from dead organisms <i>Recovery of rare species:</i> can detect minority populations by suppressing majority populations and selectively promoting growth of fastidious microorganisms at lower concentrations <i>Discovery of new species:</i> sequences isolated by culturomics can be compared, at the species level, to those generated by previous metagenomic studies, thus facilitating process for new species announcement <i>Costs:</i> relatively inexpensive since taxonomic identification by MALDI-TOF MS requires a low level of training and lower costs compared to 16S rRNA sequencing <i>Experimental testing:</i> increases the availability of strains for in vitro/ in vivo studies and facilitates the discovery of new targets and functions within the microbiome

Concluding remarks

Culture-dependent approaches provide valuable information about soil microbial communities, as they allow for the isolation and characterization of specific microorganisms, providing means to study their growth requirements, physiological and biochemical properties, as well as interactions with other microorganisms. Even though culture-dependent approaches have lost popularity with the advent of culture-independent approaches (e.g., genomics, transcriptomics and proteomics), they are now seeing a re-birth as we acknowledge their importance for characterizing microorganisms and their functions. Part of this re-birth was in fact greatly influenced by culture-independent techniques, which provided valuable information on how to grow so-called “unculturable” microorganisms in the laboratory. As technology advances, new methods for the isolation and characterization of soil microorganisms continue to emerge, offering new opportunities to understand the role of these communities in ecosystem functioning. While these methods still have limitations (e.g., bias toward easily culturable microorganisms, potential for isolating non-representative microorganisms, specific growth requirements), they remain a valuable tool in the study of soil microbiota. Most likely, it will be the combination of both culture-dependent and -independent approaches that will significantly advance our understanding of soil microbial communities, their functions and their relationship with ecosystem processes. A more complete understanding of these complex communities will be key to develop potential applications in sustainable agriculture, ecological restoration and bioremediation.

References

- Berdy B, et al. (2017) In situ cultivation of previously uncultivable microorganisms using the ichip. *Nature Protocols* 12(10): 2232–2242. <https://doi.org/10.1038/nprot.2017.074>.
- Bollmann A, Lewis K, and Epstein SS (2007) Incubation of environmental samples in a diffusion chamber increases the diversity of recovered isolates. *Applied and Environmental Microbiology* 73(20): 6386–6390. <https://doi.org/10.1128/AEM.01309-07>.
- Deng L, et al. (2019) Improving the accuracy of flow cytometric quantification of microbial populations in sediments: Importance of cell staining procedures. *Frontiers in Microbiology* 10(APR): 1–13. <https://doi.org/10.3389/fmicb.2019.00720>.
- Eriksson E, et al. (2007) A microfluidic system in combination with optical tweezers for analyzing rapid and reversible cytological alterations in single cells upon environmental changes. *Lab on a Chip* 7(1): 71–76. <https://doi.org/10.1039/B613650H>.
- Frossard A, Hammes F, and Gessner MO (2016) Flow cytometric assessment of bacterial abundance in soils, sediments and sludge. *Frontiers in Microbiology* 7(JUN): 1–8. <https://doi.org/10.3389/fmicb.2016.00903>.
- Garland JL (1997) Analysis and interpretation of community-level physiological profiles in microbial ecology. *FEMS Microbiology Ecology* 24: 289–300.
- Germida JJ and de Freitas JR (2008) Cultural methods for soil and root-associated microorganisms. In: Carter MR and Gregorich EG (eds.) *Soil Sampling and Methods of Analysis*, 2nd edn, pp. 341–353. Boca Raton, FL: CRC Press Taylor & Francis Group.
- Keloth A, et al. (2018) Single cell isolation using optical tweezers. *Micromachines* 9(9): 434. <https://doi.org/10.3390/mi9090434>.
- Khalili B, et al. (2019) Optimization of a method to quantify soil bacterial abundance by flow cytometry. *mSphere* 4(5). <https://doi.org/10.1128/mSphere.00435-19>. G. Suen.
- Lagier J-C, et al. (2012) Microbial culturomics: Paradigm shift in the human gut microbiome study. *Clinical Microbiology and Infection* 18(12): 1185–1193. <https://doi.org/10.1111/1469-0691.12023>.
- Lagier J-C, et al. (2018) Culturing the human microbiota and culturomics. *Nature Reviews Microbiology* 16(9): 540–550. <https://doi.org/10.1038/s41579-018-0041-0>.
- Madigan MT, et al. (2021) *Brock Biology of Microorganisms*, 16th edn Harlow: Pearson.
- Nichols D, et al. (2010) Use of ichip for high-throughput in situ cultivation of “uncultivable” microbial species. *Applied and Environmental Microbiology* 76(8): 2445–2450. <https://doi.org/10.1128/AEM.01754-09>.
- Pelczar MJ, Chan ECS, and Kreig NR (2002) *Microbiology*, 5th edn New Delhi: Tata McGraw-Hill.
- Pham VHT and Kim J (2012) Cultivation of unculturable soil bacteria. *Trends in Biotechnology* 30(9): 475–484. <https://doi.org/10.1016/j.tibtech.2012.05.007>.
- Sood U, Kumar R, and Hira P (2021) Expanding culturomics from gut to extreme environmental settings. *mSystems* 6(4). <https://doi.org/10.1128/mSystems.00848-21>.
- Weber KP and Legge RL (2010) Community-level physiological profiling. In: Walker JM (ed.) *Methods in Molecular Biology*, pp. 263–281. Humana Press. https://doi.org/10.1007/978-1-60761-439-5_16.
- Woomer PL (1994) Most probable number counts. In: Weaver RW, et al. (ed.) *Methods of Soil Analysis: Part 2 Microbiological and Biochemical Properties*, pp. 59–79. Soil Science Society of America, Inc.

Studying soil microbial communities: Culture-independent approaches

Tolulope Mafa-Attoye, Heather Slinn, Micaela Tosi, Evan Mayer, Dasiel Obregon, and Kari Dunfield, School of Environmental Sciences, University of Guelph, Guelph, ON, Canada

© 2023 Elsevier Ltd. All rights reserved.

Introduction	147
Traditional techniques to study soil microbial abundance and biomass	147
Enumeration techniques using microscopy	147
Biomass: Substrate-induced respiration	148
Biomass: Fumigation-based techniques	148
Biomass: Biomarker-based techniques	148
Biomass of specific microbial groups	148
PCR-based techniques to detect and quantify specific microbial groups	149
Fluorescence in situ hybridization (FISH) to identify specific microbial groups	149
Microbial community profiling using fatty acids-based approaches	149
Microbial community profiling using traditional molecular techniques	150
Hybridization-based methods: Microarrays	150
Clone library construction methods	150
PCR-based fingerprinting methods	150
High-throughput techniques for community composition and diversity analyses	151
Amplicon-based sequencing	151
Metagenomics (shotgun sequencing)	153
Single-cell sequencing	153
Concluding remarks	153
References	153

Abstract

Soils are known to host remarkably diverse microbial communities. The first tools to investigate these communities relied on culture-based approaches and could not detect the vast majority of the microbial community. Over the last few decades, numerous culture-independent methods have been developed to improve the characterization of microbial communities. Most notable among these methods is the advent of PCR, clone library construction, high-throughput sequencing, and metagenomics, with high-throughput sequencing and metagenomics being the approach currently most used by microbial ecologists.

Glossary

Bioinformatics Interdisciplinary field based on computer science, statistics and mathematics to analyze and interpret biological data such as DNA sequencing data.

Biomarker A measurable and characteristic compound or molecule used to identify or quantify the presence or abundance of a specific organism or type of organism.

Community profiling Techniques used to identify and quantify microbial structure in terms of the organisms present and their relative abundance.

Culture-independent methods Techniques to study microorganisms without growing them in culture media.

Fingerprinting Techniques used to distinguish between different organisms based on their DNA.

Fluorescent labels Chemicals that emit light when excited by a specific wavelength of light which are added to nucleotides for detection in different methodologies.

Gel electrophoresis Technique used to separate DNA fragments based on their size and charge.

Metagenome The collective genetic material present in an environmental sample, consisting of the genomes of many different organisms.

Microbial abundance Number or count of microorganisms in a given environment or sample.

Microbial biomass Total mass or weight of living microorganisms in a given environment or sample.

Plasmid Small circular piece of DNA that can replicate independently of the main chromosome in bacteria.

Pyrosequencing DNA sequencing method that uses detection of pyrophosphate release from nucleotide incorporation to determine the sequence of a DNA molecule.

Read length next-generation sequencing (NGS) term referring to the number of base pairs (bp) sequenced from a DNA fragment.

Sequencing adapters Oligonucleotides or short DNA fragments (~80 bp in length) added to the end of DNA fragments for sequencing, which allows the sequencing of multiple samples in a single high-throughput sequencing run.

Sequencing coverage Proportion of a genome that is represented by sequence reads in a sequencing dataset.

Sequencing-by-ligation DNA sequencing method that uses ligation of short oligonucleotides to a DNA template.

Sequencing-by-synthesis DNA sequencing method where nucleotides are added one at a time to a growing DNA chain, with each addition detected and recorded.

Ultrastructural studies Studies based on the examination of biological specimens at a high magnification level using electron microscopy with the goal of visualizing cellular and subcellular structures at a much higher resolution than using traditional light microscopy.

Key points

- Microbial abundance and biomass
- Microbial community profiling using fatty acids-based approaches
- Microbial community profiling using traditional molecular techniques
- Amplicon-based sequencing
- Metagenomics (shotgun sequencing)

Introduction

The extremely diverse and abundant soil microorganisms include members from the three life domains: Bacteria, Archaea, and Eukarya. These microbes do not exist in isolation but thrive in communities. Understanding and describing the huge diversity of soil microbial communities requires methodologies that may allow for a comprehensive, broad view of all organisms present in a soil sample. The first methods to enumerate and identify soil microbes were culture-dependent, which were eventually shown to account for less than 0.1–1% of the community. The progression of soil microbial ecology as a scientific discipline is undeniably linked to the emergence of comprehensive, faster, specific, and more sensitive culture-independent molecular techniques, which have allowed new insights into the diversity, abundance, and activity of soil microbial communities. These techniques mostly focus on analyzing genetic-coded information from soil-extracted nucleic acids (DNA or RNA) to assess the diversity of soil microbial communities. In the last two decades, the development of high-throughput sequencing methods for nucleic acid sequencing has made possible the taxonomic profiling of soil microbiota at high resolution. In this chapter, we describe the utility and methodological basis of traditional and current culture-independent approaches used to study soil microbial community structure.

Traditional techniques to study soil microbial abundance and biomass

There are several ways to estimate the abundance and biomass of microbial communities in soil. Techniques can be either direct or indirect, may target a whole community or just a subgroup, and can capture microorganisms under different physiological conditions (e.g., active, dormant, and sometimes dead). Before the advent of culture-independent methods, enumeration techniques like microscopy and culturing were commonly used to quantify the abundance of microorganisms in a soil sample. These methods have been replaced by more accurate and less labor-intensive techniques, especially in the field of soil microbial ecology where the subject of study is the whole microbial community.

Microbial biomass represents the mass of living microorganisms present in the soil, and, as such, it is a fraction of soil organic matter. Given the simplicity and robustness of the soil microbial biomass assessment methods, it has widely been used as an indicator of soil quality and health, and thus an indirect indicator of the soil's functional status. There are several methods to quantify both soil microbial biomass and abundance, most of which rely on the extraction and quantification of a compound or biomarker, or physiological parameters such as the respiratory response of the community.

Enumeration techniques using microscopy

Microscopy allows for direct counts of microbial cells. It usually involves preparing a 1:5 soil: water suspension, selective staining (e.g., acridine orange for fluorescence microscopy), and counting the bacterial cells or fungal mycelia. Compared to culture-dependent techniques (e.g., plate counts), microscopy is direct and not restricted to the very small fraction (~1%) of the soil microbial community that is represented by culturable microorganisms. Hence, it is not surprising that enumeration via microscopy yields larger numbers of bacteria than culturing methods. Still, microscopy also presents its challenges, such as being more suitable for bacteria than fungi, because of their simpler unicellular morphology. Depending on the principle of acquiring images, microscopy may be differently catalogued. Light microscopy is ideal for naturally pigmented cells. However, most cells are

not pigmented, so they need to be thermally killed (i.e., fixed) and stained with dyes to facilitate contrast and visualization. Alternatively, phase contrast microscopy allows imaging of live cells in black and white. With electron microscopy, beams of electrons, instead of diffusing light, focus on the specimen. A more advanced type of microscopy, fluorescence microscopy, uses laser generators to supply incident light of a defined wavelength to the specimen, which then generates defined wavelength fluorescence; this can be employed to repeatedly capture live cell images and thus visualize time lapses of changes in living cells. These methods in microscopy provide high-resolution images and are effective in ultrastructural studies of soil particles and microorganisms. Since there are several complexities involved in soil preparation, no universal microscopy technique applies to all microbes in the soil.

Biomass: Substrate-induced respiration

Soil microbial biomass can be determined using physiological methods, such as substrate-induced respiration (SIR), fumigation-incubation, or oxygen-uptake assessment. SIR quantifies the respiratory response of the active fraction of the aerobic and heterotrophic microbial community upon being supplied with defined carbon substrates. These respiratory response measurements can be converted to units of total microbial biomass by calibration against fumigation-based techniques. Respired CO₂ can be measured via titration or by chromatographic analysis of headspace gas. An empirically determined conversion factor of 40.04, estimated from the correlation between SIR and fumigation techniques, is commonly used to convert the rate of substrate-induced CO₂ to total microbial biomass. A standardized protocol to quantify soil microbial biomass via SIR is described in ISO 14240-1:1997.

Biomass: Fumigation-based techniques

Fumigation-incubation, also based on respiratory responses, is a direct measurement of soil microbial biomass. Here, soil samples are fumigated with chloroform to kill microbial cells, inoculated with a non-fumigated soil suspension aliquot, and then incubated for 24 h. It is assumed that the newly inoculated cells will consume the dead biomass from fumigated samples and respire CO₂. The difference in CO₂ production between fumigated and unfumigated samples is thus assumed proportional to the dead biomass in the fumigated sample and can be expressed as microbial biomass carbon using a conversion factor of 0.45. Microbial biomass nitrogen can be calculated by extracting and quantifying ammonium instead of measuring respired CO₂. A similar approach is used in the fumigation-extraction method, where organic C from both fumigated and non-fumigated samples is extracted using 0.5 M K₂SO₄ and quantified, and microbial biomass carbon is estimated as the difference between both. The fumigation-extraction approach also allows the extraction and quantification of N and P, to estimate the size of microbial pools for these two elements. The standardized protocol for the fumigation-extraction method is described in ISO 14240-2:1997.

Biomass: Biomarker-based techniques

Biomarkers are simply biochemical compounds found in cells, but sometimes also present extracellularly. For biomarkers to be reliable estimators of biomass, ideally, they should only be present at a relatively constant concentration in living cells, quickly degrading after release from dead cells. A good example of bioindicators of microbial biomass is total, extractable fatty acid methyl esters (FAMES) or, more particularly, methyl esters of phospholipid fatty acids (PLFAs) (Li et al., 2020). These lipid biomarkers are associated with vital functions, such as storing energy and signaling, and thus their abundance and profiles are indicators of living soil microbial communities. FAMES comprise PLFAs, as well as neutral lipids and glycolipids. The PLFA fraction is constituted of polar lipids from the cell membrane lipid bilayer that have a very short life in soil once the cell dies. Thus, PLFAs represent only viable microorganisms.

An alternative approach is to extract total DNA and quantify it as an estimator of microbial biomass, using a method that maximizes DNA yield (ISO 11063:2012) as opposed to commercial DNA extraction kits (Petric et al., 2011).

A key limitation of these techniques is that extracted biomarkers might not be derived only from living organisms. Moreover, variable extractability can make the techniques less accurate.

Biomass of specific microbial groups

Most of the methods described above target microbial communities without distinction. Nevertheless, some biomarkers are found only in certain microbial groups and are thus suitable for targeted quantification. An example widely used is ergosterol, a membrane sterol found in most fungi but rarely in other microorganisms. Similarly, muramic acid and diaminopimelic acid were used to quantify prokaryotic biomass. Both FAMES and PLFAs have also been used to quantify specific groups such as saprophytic fungi, arbuscular mycorrhizal fungi, and bacteria (and within them Gram-positive, Gram-negative, anaerobic bacteria, cyanobacteria, etc.) (Frostegård et al., 1996). Alternatively, selective inhibition using antibiotics can be coupled with biomass measurements like SIR to estimate the biomass of bacteria (e.g., cycloheximide inhibits eukaryotes) or fungi (e.g., streptomycin inhibits prokaryotes). However, the level of suppression of the non-target group is variable with these inhibition methods.

PCR-based techniques to detect and quantify specific microbial groups

Polymerase chain reaction (PCR) methods like real-time quantitative PCR (qPCR) and digital PCR (dPCR) allow for an estimation of the abundance of specific microbial groups in a soil sample by amplifying a targeted indicator gene sequence. These methodologies are rooted in the assumption that each cell has an approximately equal count of the targeted indicator gene. The basic steps of PCR involve cycles of DNA denaturation, primer annealing, and extension (i.e., DNA polymerase synthesizes new target DNA strands). Nucleotides, Taq polymerase, DNA templates, and primers specific to the target genes are involved in the amplification of a small, specific segment of DNA. The cycles are repeated several times producing more than one billion exact copies of the initial DNA template.

The qPCR method uses fluorescence to recurrently quantify the amplified product after each PCR cycle, instead of a single end-point measurement (Heid et al., 1996). A variant of qPCR is reverse transcription qPCR (RT-qPCR), which allows work with RNA templates and thus quantifies RNA transcripts instead of gene copies. This is done by adding an initial step where RNA is reverse transcribed into complementary DNA (cDNA). Both qPCR and RT-qPCR rely on fluorescence intensity responses proportional to the gene copies or transcripts per unit reaction volume. A large number of copies makes quantification easier. If the amplification rate is known, and amplification inefficiencies are also known, then the final counts of the copies of the templates are proportional to the starting counts. The two detection chemistries used in qPCR are double-strand (ds) DNA-binding molecules and fluorophore-labeled short strands of DNA (oligonucleotides) that can serve as the starting point (i.e., primers) for amplification of a target sequence. For qPCR, SYBR Green is the most commonly used dsDNA-binding fluorescent dye. However, SYBR Green can bind to nonspecific PCR products resulting in overestimation. Oligonucleotides, can be modified by attaching a fluorophore at one end (i.e., fluorescent molecule at the 5' position?) and a fluorescence quencher at the other end (the 3' position). These are called TaqMan probes. If the probe is intact, then the proximity of the quencher to the fluorophore inhibits fluorescence. As the DNA strands are synthesized and the double strand DNA is produced the TaqMan probes are cleaved during hybridization and thus the quencher is separated from the fluorophore allowing it to fluoresce. As more copies of the target gene sequence are produced, the intensity of the fluorescence increases proportionally, allowing a direct quantification. Compared with SYBR Green, TaqMan-based detection is highly specific and involved in more applications, such as multiplexing, single nucleotide polymorphism (SNP) genotyping, copy number variation, pathway analysis, microRNAs, and protein analysis.

Digital PCR (dPCR) is a variation that relies on the same chemistry as the qPCR. However, for dPCR, the sample is diluted and partitioned into multiple subsamples on plates. Dilution purifies the target from inhibiting compounds, and the number of partitions improves precision. The partitioning ensures that the detection limit is enhanced as the small reaction volume increases the concentration of the target molecule. This makes dPCR useful for targets with only a small abundance, and applications in complex environments. Digital PCR platforms for partitioning can be classified as droplet-based (droplet digital PCR or ddPCR) and chamber-based (chamber digital PCR or cdPCR). The cdPCR utilizes physical structures (i.e., microfluidic chambers) as partitions, whereas ddPCR utilizes a water-oil emulsion droplet system that can generate millions of suspended droplets acting as partitions. After end-point amplification, any fluorescence is detected in droplets, with results digitized as 0 and 1 for negative and positive droplets, respectively. Using the Poisson distribution law, the fraction of positive droplets are converted to the number of molecules in the initial sample, using absolute quantification (standard curves not needed).

Fluorescence in situ hybridization (FISH) to identify specific microbial groups

Fluorescence in situ hybridization (FISH) is a technique that uses fluorescent probes (i.e., RNA or DNA oligonucleotides) that are complementary to a target sequence in a cell, to detect and identify specific groups or organisms in the sample. Common target sequences are the 16S or 23S rRNA genes, or mRNA for gene expression studies. A fluorescent dye is attached to the probe, so it can be visualized in situ using epifluorescence, a form of fluorescence microscopy. The basic steps of the FISH procedure involve denaturation, hybridization, probe detection, and analysis. Over the years, improvements in the specificity, sensitivity, and resolution of the FISH technique together with advances in fluorescence microscopy have resulted in the diversification of the protocol into various procedures that have increased our knowledge of the chemical properties of nucleic acids. For example, a variant suitable for gene expression studies or phylogenetic studies of microorganisms that are slow-growing or of small abundance is the catalyzed reporter deposition-FISH (CARD-FISH). In CARD-FISH, the probe does not have a fluorophore attached but the peroxidase enzyme, so that after hybridization it can be treated with a fluorescent labeled substrate of this enzyme. Other related techniques are fiber-FISH, useful for mapping genes, and rainbow-FISH for the detection and quantification of at least seven different microbial groups in a microscopic field. An in silico version of this technique is e-FISH, a BLAST-based simulation program useful as a bioinformatics tool.

Microbial community profiling using fatty acids-based approaches

The FAME and PLFA techniques described above can also be used for characterizing soil microbial community structure. For example, these techniques have been used to evaluate contrasting ecosystems or agricultural management practices, or the effect of biotic or abiotic stresses or disturbances. Both FAME and PLFA are rapid, sensitive, inexpensive techniques which use alkaline reagents to extract fatty acids from soil converting them into FAMEs. Based on the presence of unique fatty acid markers in different

microbial groups, these techniques allow the detection of specific microbial groups, such as Gram-positive and Gram-negative bacteria, actinobacteria, arbuscular mycorrhizal fungi (AMF), or saprophytic fungi. Specific fatty acids are usually detected and quantified using gas chromatography–mass spectrometry (GC–MS), based on structural criteria, such as the number of carbon atoms, or the number and location of double bonds. Although both PLFA and FAME share the same set of microbial FAME biomarkers, both methods differ during lipid extraction, separation, and transesterification. Whereas the FAME method extracts PLFA, glycolipids, and neutral lipids together, the PLFA methodology elutes and discards the last two fractions to purify only polar lipids. As mentioned earlier, because phospholipids have a short turnover time once the cell dies, they provide information on living organisms. A recent study also found that, despite having a higher cost and complexity than FAME, PLFA was more sensitive to changes in soil physicochemical properties (Li et al., 2020). Some fatty acids have been considered as indicators of physiological stress. A major limitation of these methods is that many fatty acids are common to several microorganisms, so they can be variably produced by diverse organisms under the same environmental conditions. This limits the resolution of FAME or PLFA profiles and makes their interpretation challenging (Frostegård et al., 2011).

Microbial community profiling using traditional molecular techniques

As mentioned in the previous sections, microbial ecology has been revolutionized in the last few decades, as emphasis has shifted from identifying individual microbes by culturing techniques to advanced methods employed for the assessment of natural communities. The analysis of molecular, DNA-based methods that involve cloning and sequencing of DNA fragments or amplification of target genes using PCR are some of the various approaches to community profiling (Nocker et al., 2007; Hirsch et al., 2010). Genetic profiling methods, such as denaturing gradient gel electrophoresis (DGGE), temperature gradient gel electrophoresis (TGGE), and automated ribosomal intergenic spacer analysis (ARISA), were among the first methods to be applied in the study of microbial composition and diversity. However, from the start of the 21st century, the rapid development of sequencing technologies has revolutionized the study of soil microbial communities and most of these technologies have been superseded.

Hybridization-based methods: Microarrays

In environmental microbiology, microarrays (or microchips) are used to characterize the microbiome (e.g. PhyloChip via phylogenetic oligonucleotide arrays) or to identify gene expression (e.g. GeoChip via functional gene arrays) of microbes in samples (Letowski et al., 2003; He et al., 2007). Microarrays use hybridization of either oligonucleotide probes or cDNA with target DNA or RNA sequences that have been tagged with fluorescent labels. The oligonucleotide probes are selected based on what is expected to be found or targeting organisms or functional genes of interest. When hybridization occurs, the fluorescent label is turned on indicating the presence of the feature of interest. Microarrays can be incredibly useful as they offer a high-throughput alternative (thousands of features can be assessed at a time) that is faster and costs less than sequencing. They are also good for detecting rare taxa in environmental samples and, depending on the probes used, can give a species-level identification of taxa. However, their hybridization technology can lead to false positives from unspecific complementary strand binding. Another limitation is that relatively large quantities of DNA or RNA are needed for hybridization to work.

Clone library construction methods

In the field of recombinant DNA technology, gene cloning has emerged as one of the most important developments of all time (Rougeon et al., 1975). The first phase in cloning a specific gene involves the construction of cloned DNA fragments, DNA library or gene libraries containing the gene of interest. Then, the DNA fragment that contains the target gene is inserted into purified DNA genome or a plasmid (cloning vector). Once cloned into the vector, sequencing of the insert produces the highest phylogenetic resolution and species identification can be achieved by using sequence databases. Sequencing of clone libraries from environmental samples can provide information about microbial diversity.

As part of the advancement in molecular biology techniques, ribosomal RNA gene sequencing emerged as a versatile approach to study microbes present in environmental samples. The 16S rRNA gene, which codes for the RNA that forms part of the small ribosomal subunit and is important for functional protein formation, is the most used sequencing target for identification of prokaryotic DNA. This gene comprises both highly conserved regions, which allows the development of primers to target a wide array of taxa, but also hypervariable regions that can provide high taxonomic resolution. The extraction of nucleic acids and subsequent amplification is followed by cloning the 16S rRNA genes and sequencing. Cloning of the PCR-amplified sequences allows for accurate species identification. Another common application is the cloning of cDNA for the assessment of gene expression.

PCR-based fingerprinting methods

DNA fingerprinting examines part of the genetic information contained in nucleic acids extracted directly from environmental samples. For decades, various techniques such as denaturing gradient gel electrophoresis (DGGE), temperature gradient gel electrophoresis (TGGE), terminal restriction fragment length polymorphism analysis (T-RFLP), and automated ribosomal

intergenic spacer analysis (ARISA) were developed for DNA fingerprinting microbial communities. DNA extracted from environmental samples comprises fragments from all microbes in the community. Primers targeting the variable regions of the 16S rRNA gene are used in amplification and then gel electrophoresis is used to visualize and differentiate amplified fragments based on their size and sequence variability. These techniques provide rapid, inexpensive information on the community composition profiles and diversity between soil samples without the need to grow microbes on culture media.

Both DGGE and TGGE are based on the differential bond strength between double DNA strands as a function of the proportion and order of DNA bases. Strands richer in guanine-cytosine nucleotide base pairs (GC), or with more collocated GC, are less easily denatured. Differential denaturation can be achieved along a polyacrylamide gel containing an increasing gradient of chemical denaturants (usually urea and formamide; DGGE) or exposed to a temperature gradient (TGGE). As the double-stranded DNA fragments partially denature, their *electrophoretic mobility* is significantly reduced. Unequal denaturation among strands with different genetic codes (different order and abundance of DNA bases) will lead to differences in electrophoretic mobility. This produces different band counts, at different locations in the gel and of variable band intensity corresponding to the relative number of microbial taxa and the relative abundance respectively. Gels can be analyzed with imaging software to determine diversity indices and bands can be excised from the gel and sequenced to identify microbial taxa.

T-RFLP technique involves DNA amplification using a fluorescent-labeled primer, followed by restriction enzyme digestion, and fragment size separation through gel electrophoresis. This is possible using an automated capillary gel electrophoresis in a DNA sequencer platform. The fluorescent labeled fragments travel through the gel at variable, size-dependent speeds, and the fluorescence intensity (or peak size) is read by a detector. An electropherogram is thus produced, which can be analyzed using automated fragment analysis programs such as T-REX which calculate peak sizes (TRF length in base pairs—bp). TRF patterns, such as presence or absence and relative abundance of peaks, can be then analyzed using diversity indices and multivariate statistics.

Unlike DGGE or TGGE and T-RFLP, ARISA targets the internal transcribed spacer (ITS) region between the 16S rRNA gene and the gene coding for the large ribosomal subunit (LSU) or 23S rRNA, using specific primers for PCR amplification. In ARISA, electrophoretic separation of fluorescently labeled PCR products generates a community fingerprint that is detected by a sensitive laser on an automated sequencing system. A variation of this method is the quantitative fingerprinting (“qfingerprinting”), which involves a combination of two methods: community fingerprinting analysis and a dilution endpoint PCR. This is a quantitative version of ARISA, qARISA, which can be used for the quantitative estimation of targeted OTUs from complex communities (Ramette, 2009).

High-throughput techniques for community composition and diversity analyses

High-throughput sequencing (HTS) or next-generation sequencing (NGS) is the broad term used to refer to DNA sequencing methods developed over the last couple decades following the capillary sequencing methods (Sanger sequencing). These technologies enable many processes to take place in parallel, which allows for the sequencing of thousands to millions of molecules simultaneously, thus enabling the sequencing of large genomes in a fast and cost-effective manner, but also sequencing of all the genetic material (metagenome) in a sample (Fig. 1). Because NGS increases sequencing coverage, it offers more detailed estimations of community composition and diversity of the soil microbiome than earlier technologies. Consequently, in a very short time NGS techniques have replaced the construction of traditional clone libraries and fingerprinting techniques, revolutionizing the fields of soil microbiology and microbial ecology. The first NGS technology was pyrosequencing (e.g., platform Roche 454), a sequencing-by-synthesis method based on the detection of DNA polymerase activity (i.e., pyrophosphate release) through the emission of visible light (Ronaghi et al., 1998). Pyrosequencing, now rarely used, was followed by other sequencing-by-synthesis technologies developed by Illumina (e.g., MiSeq and HiSeq) that use dNTPs with fluorescent labels, and which are still highly popular. At around the same time as Illumina, a different type of technology known as Sequencing by Oligonucleotide Ligation and Detection (SOLiD), came to market and is still available today. Ion semiconductor sequencing, or Ion Torrent sequencing, is another available sequencing platform on the market which does not use optics, like pyrosequencing and Illumina, but rather the production of hydrogen ions that are generated during DNA polymerization. In 2022, we are already into the next wave of sequencing technologies called ‘third generation’, which is characterized by increased read length (e.g., Pacific Biosciences, Oxford Nanopore) and portable technologies which can be brought to the field (e.g., Oxford Nanopore MinION sequencer). The performance of these platforms varies in terms of sequence capacity, genome depth, and quality, which are also associated to differences in terms of cost. Because of the large amount of data (i.e., sequences) produced by NGS, several bioinformatics pipelines and tools have been developed that properly process and interpret results. In this section, we address the main NGS approaches used to explore the composition and diversity of soil microbial communities.

Amplicon-based sequencing

Amplicon sequencing involves HTS and the detailed analysis of one particular gene of interest (marker gene) within an environmental sample, in the way that the genetic diversity reflects the phylogenetic and taxonomic diversity of the microbial communities (Fig. 1). Any gene can be selected for amplicon sequencing, but their choice must be based on the taxonomic resolution that is expected. Target genes are used as taxonomic marker genes for different domains or kingdoms. For example, the 16S rRNA gene is used for bacteria and archaea, the 18S rRNA gene and ITS region are used for fungi, and 18S rRNA and the cytochrome *c* oxidase

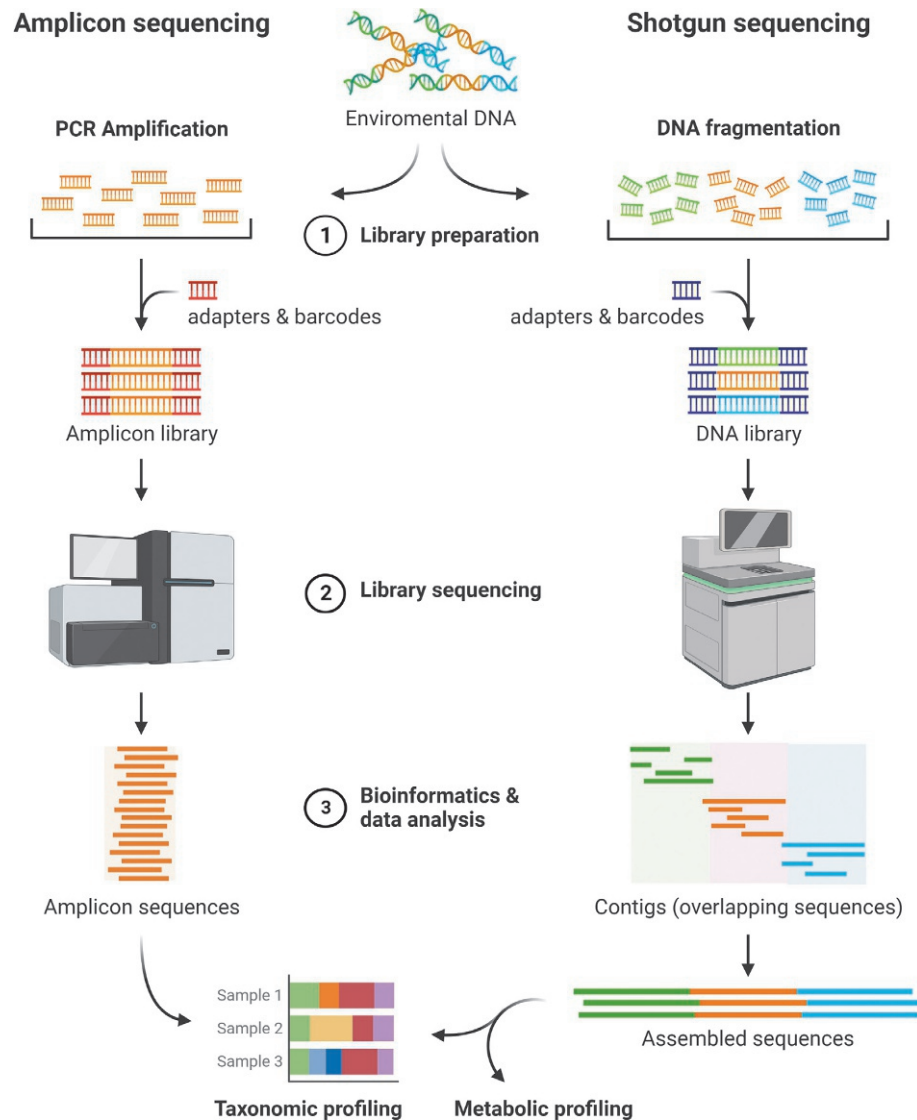


Fig. 1 Scope and methodology comparison between amplicon and shotgun sequencing approaches. Amplicon sequencing only targets specific regions of the DNA (e.g., a region of the 16S rRNA gene), whereas shotgun sequencing targets all DNA molecules randomly. The main differences are in the steps of library preparation (1), library sequencing (2) and bioinformatics and data analysis (3). For library preparation (1), amplicon sequencing involves PCR amplification to amplify the target gene region and then incorporate the adapters and barcode sequences needed for parallel DNA sequencing from different samples, while random sequencing does not require PCR amplification and DNA samples are directly barcoded together for sequencing. Additionally, the output of these sequencing technologies also differs during and after bioinformatic processing (3). Whereas amplicon sequencing data allows for taxonomic profiling of the microbiota, with precise taxonomic resolution down to the genus level, shotgun sequencing allows a more accurate taxonomic resolution to the species level and, more importantly, it can be used to produce a metabolic profile of the microbiome based on the content of function-coding genes.

subunit I (COI or CO1) gene are often used for other eukaryotes. Nevertheless, in some circumstances, a functional gene involved may be a better option. This is particularly useful to study microbial transformation of interest (e.g., to study nutrient cycling or biodegradation) but also for low abundance microbial groups. For example, the genes *amoA* and *pmoA* are suitable for studying autotrophic ammonia-oxidizing and methane-oxidizing bacteria, respectively, two functions that are highly relevant for nitrogen and methane cycling and carried out by organisms that represent less than 1% of the soil microbial community.

There are two steps to the amplicon sequencing process: library preparation and parallel sequencing of the target amplicon. The library construction involves two steps: (1) a PCR amplification to yield a pool of target sequences (amplicons), and (2) the addition of sequencing adapters, which are small pieces of DNA designed with 3 functional segments: binding site, index (barcode) and primer binding site that interacts with the NGS platform. The barcode allows multiple samples to be sequenced in a single sequencing run (multiplexing), in a way that the sequences can be identified and tracked by samples (demultiplexing). Finally, amplicons are sequenced using various NGS platforms from the different technologies (i.e., Illumina, PacBio, Oxford Nanopore,

etc.). In this sequencing method, similar to the other NGS technologies, it is possible to sequence a single DNA strand (single-end sequencing) or both forward and reverse strands (paired-end sequencing). This advantage is particularly important for Illumina platforms (the most commonly used), which are limited to producing short reads (150–300 bp). Using paired-end sequencing it is possible to target a long fragment (~500 bp) and then merging forward and reverse reads into a single sequence.

Once amplicons are sequenced, they are analyzed using standardized bioinformatic pipelines. These pipelines are computer software that optimizes sequence denoising, screening, and taxonomic annotation based on reference databases (e.g., SILVA, RDP or Greengenes for prokaryotic 16S rRNA sequences, Unite for fungal ITS data, and BOLD for eukaryotic CO1 amplicons).

Metagenomics (shotgun sequencing)

Shotgun metagenome sequencing involves random and parallel sequencing of all the DNA from all organisms within the microbial community in an environmental sample, and simultaneously adding barcode sequences to each fragment enabling sequencing of several samples at once (multiplexing). In this approach, DNA is randomly broken up into numerous small segments, and then sequenced. Because of the overlapping of the short reads amplified, it is possible to assemble it in long-length sequences named contigs (Fig. 1). Due to this characteristic, shotgun sequencing also makes it possible to sequence and assemble the complete genome of a microorganism of interest, either cultured or directly from an environmental sample.

In terms of taxonomic profiling of the microbial communities, shotgun sequencing enables high sequencing coverage of all the DNA sequences in the sample, including rare species that are often undetected by amplicon sequencing. In addition, since multiple reads from several genes are targeted, it allows species-level detection of the community members. The process of sequence cleaning, assembly, and annotation of shotgun metagenomic data is done via computer software. Many programs or pipelines exist for processing shotgun reads, with two of the most used programs in soil science being MG-Rast and the combination DIAMOND-MEGAN. The RefSeq database hosted by the National Center for Biotechnology Information (NCBI) is one of the main reference databases for taxonomic annotation of shotgun sequence data.

Overall, the main difference between amplicon and shotgun sequencing is their scope: amplicon sequencing targets one specific gene, while shotgun sequencing targets all DNA in a sample. Some advantages to using amplicon sequencing over shotgun sequencing are the smaller cost and that it provides greater coverage for the gene of interest avoiding contamination from non-target DNA. On the other hand, shotgun sequencing has greater scope (DNA from all the domains present in the sample), enabling both taxonomic and functional annotation and profiling of soil microbiome, and provides greater coverage and resolution of the composition of microbial communities.

Single-cell sequencing

Single-cell sequencing (SCS) is HTS sequencing of whole genomes of individual cells in a sample. The process of single-cell sequencing begins with the isolation of an individual cell from its environment, which can be accomplished using a variety of methods such as fluorescence-activated sorting, patch pipettes, and microfluidics techniques. Once cells are isolated, their total DNA is extracted, followed by library preparation and DNA sequencing using NGS techniques. The general functions of SCS include observing genomic heterogeneity between cells within a population and evolutionary relationships between organisms. Applications of SCS specific to soil science include studying rare or unculturable microorganisms, elucidating microbial niches, and studying horizontal gene transfer within a microbial population (Kvist et al., 2007; Walker and Parkhill, 2008).

Concluding remarks

Culture-independent approaches have revolutionized our understanding of soil microbial communities, providing a more comprehensive view of their composition, diversity, and functional potential. These methods have enabled the detection and identification of microorganisms that cannot be cultivated in the lab, offering a more accurate picture of the microbial communities present in soil ecosystems. Despite the advantages of culture-independent approaches, it is important to acknowledge their limitations (e.g., potential biases and limitations in detecting low-abundance microorganisms) when designing studies and interpreting results. Ultimately, the choice of a particular culture-independent approach will depend on the research question, experimental setup, and resource availability. Overall, culture-independent methods represent a powerful tool for studying soil microbial communities and will continue to shape our understanding of microbial ecology and ecosystem functioning in the future.

References

- Frostegård A, Tunlid A, and Bååth E (1996) Changes in microbial community structure during long-term incubation in two soils experimentally contaminated with metals. *Soil Biology and Biochemistry* 28(1): 55–63. [https://doi.org/10.1016/0038-0717\(95\)00100-X](https://doi.org/10.1016/0038-0717(95)00100-X).
- Frostegård Å, Tunlid A, and Bååth E (2011) Use and misuse of PLFA measurements in soils. *Soil Biology and Biochemistry* 43(8): 1621–1625. <https://doi.org/10.1016/j.soilbio.2010.11.021>.
- He Z, et al. (2007) GeoChip: A comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *The ISME Journal* 1(1): 67–77. <https://doi.org/10.1038/ismej.2007.2>.

- Heid CA, et al. (1996) Real time quantitative PCR. *Genome Research* 6(10): 986–994. <https://doi.org/10.1101/gr.6.10.986>.
- Hirsch PR, Mauchline TH, and Clark IM (2010) Culture-independent molecular techniques for soil microbial ecology. *Soil Biology and Biochemistry* 42(6): 878–887. <https://doi.org/10.1016/j.soilbio.2010.02.019>.
- Kvist T, et al. (2007) Specific single-cell isolation and genomic amplification of uncultured microorganisms. *Applied Microbiology and Biotechnology* 74(4): 926–935. <https://doi.org/10.1007/s00253-006-0725-7>.
- Letowski J, Brousseau R, and Masson L (2003) DNA microarray applications in environmental microbiology. *Analytical Letters* 36(15): 3165–3184. <https://doi.org/10.1081/AL-120026566>.
- Li C, et al. (2020) A comparison between fatty acid methyl ester profiling methods (PLFA and EL-FAME) as soil health indicators. *Soil Science Society of America Journal* 84(4): 1153–1169. <https://doi.org/10.1002/saj2.20118>.
- Nocker A, Burr M, and Camper AK (2007) Genotypic microbial community profiling: A critical technical review. *Microbial Ecology* 54(2): 276–289. <https://doi.org/10.1007/s00248-006-9199-5>.
- Petric I, et al. (2011) Inter-laboratory evaluation of the ISO standard 11063 "Soil quality—Method to directly extract DNA from soil samples" *Journal of Microbiological Methods* 84(3): 454–460. <https://doi.org/10.1016/j.mimet.2011.01.016>.
- Ramette A (2009) Quantitative community fingerprinting methods for estimating the abundance of operational taxonomic units in natural microbial communities. *Applied and Environmental Microbiology* 75(8): 2495–2505. <https://doi.org/10.1128/AEM.02409-08>.
- Ronaghi M, Uhlén M, and Nyrén P (1998) A sequencing method based on real-time pyrophosphate. *Science* 281(5375): 363–365. <https://doi.org/10.1126/science.281.5375.363>.
- Rougeon F, Kourilsky P, and Mach B (1975) Insertion of a rabbit β -globin gene sequence into an *E. coli* plasmid. *Nucleic Acids Research* 2(12): 2365–2378. <https://doi.org/10.1093/nar/2.12.2365>.
- Walker A and Parkhill J (2008) Single-cell genomics. *Nature Reviews Microbiology* 6(3): 176–177. <https://doi.org/10.1038/nrmicro1862>.

Studying microbial function in soil: Culture-independent techniques

Micaela Tosi, Dasiel Obregon, Heather Slinn, and Kari Dunfield, School of Environmental Sciences, University of Guelph, Guelph, ON, Canada

© 2023 Elsevier Ltd. All rights reserved.

Introduction	156
Traditional techniques to study microbial function in soil	156
Microbial respiration	157
Catabolic capacity and biodegradation	157
Simple substrate utilization assays	157
Complex substrate biodegradation assays	158
Microbial nitrogen transformations	158
Nitrogen mineralization and nitrification	158
Denitrification	159
Nitrogen fixation	159
Enzymatic activity	159
Dehydrogenase activity	159
Hydrolysis of fluorescein diacetate	159
Specific enzymatic activities	160
Isotopic techniques to study microbial function in soil	161
Stable isotopes techniques	161
Natural abundance and isotopic fractionation	161
Isotopic enrichment	161
Stable isotope probing	162
Radioisotope techniques	162
Radiotracer techniques	162
Thymidine and leucine incorporation	163
Omics techniques to study microbial function in soil	163
Sequence-based metagenomics	163
Functional metagenomics	163
Functional gene prediction from amplicon-sequence data	164
Metatranscriptomics	164
Metaproteomics	165
Metabolomics	165
Multimomics	165
Concluding remarks	166
References	166
Further reading	167

Abstract

Microorganisms have numerous functions in soils. They are key drivers of soil biogeochemical cycling, which determines the fate of nutrients and organic matter. This gives microbial activity a key role in soil ecosystem services such as plant growth, carbon sequestration and regulation of greenhouse gas emissions. Since the beginnings of microbial ecology, technological advancements have increased our understanding soil microbial communities and shed light on the “black box” of microbial activity. This chapter revisits the most influential techniques and methodologies that have helped us study microbial function in soil.

Glossary

Amplicon sequencing DNA or RNA sequencing technique that analyzes a specific region of the genome previously targeted and amplified using PCR.

Biochemical techniques Techniques developed in the mid to late 20th century to measure biochemical reactions carried out by active microbial cells or their extracellular enzymes.

Biodegradation Process of breaking down complex substrates into simpler components.

Biomagnification Process by which the concentration of a substance increases in the tissues of organisms at higher trophic levels.

Black box measurements Techniques that quantify substrates and products without exact knowledge of the reactions or organisms involved.

Catabolic capacity Capacity of a microorganism or microbial community to metabolize a given substrate.

Community-level physiological profiling (CLPP) A rapid method to assess the metabolic potential of a microbial community by simultaneously testing their growth with different carbon sources.

Isotope Variations of a chemical element differing in its number of neutrons and thus atomic mass.

Isotopic fractionation Preferential depletion or enrichment of one isotope in comparison to the others, leading to changes in isotopic proportions.

Isotopic labeling Process of marking specific substrates with a stable or radioactive isotope to trace their fate and transformations.

Machine learning algorithms Algorithms that can learn from and make predictions on data, using statistical models to recognize patterns and relationships.

Net activity rates Activity rates that represent the balance between production and consumption of reaction products.

Next-generation sequencing (NGS) Also known as second-generation sequencing, this term that comprises all high-throughput technologies for massively parallel sequencing of DNA or RNA.

Omics Any of several areas or disciplines in biology defined by the characterization or quantification of a specific type of biomolecule or the totality of a molecular process within an organism (e.g., genomics, transcriptomics, proteomics).

Radioisotope Radioactive isotopes, meaning they decay over time, emitting ionizing radiation until reaching a stable state.

Radiotracer Compound that has been labeled with a radioactive isotope of an element.

Stable isotope probing (SIP) Technique based on the labeling of substrates with a stable isotope followed by the quantification of a specific microbial biomarker (e.g., DNA).

Stable isotope Isotope that is non-radioactive, which means it does not decay over time or emit ionizing radiation.

Trophic level Position of an organism in a food chain, determined by the number of steps it is away from the primary producers.

Key points

- Microbial metabolic activity
- Microbial nitrogen transformations
- Enzymatic activities
- Isotopic techniques to study microbial functions
- Omics techniques for functions analysis

Introduction

The soil microbiota plays a major role in ecosystem functioning. In general terms, microbial communities are not only abundant and genetically diverse, but also have ample metabolic capacity and versatility. Hence, microbes are the main drivers of soil biogeochemical cycling, with consequent effects on the fate of nutrients and organic matter, as well as the degradation of xenobiotics or contaminants. This makes microbial activity key to plant nutrition and fitness, carbon sequestration and regulation of greenhouse gas emissions, among other ecosystem services provided by soils. Some soil microorganisms can also interact directly, and sometimes symbiotically, with plants and animals, either promoting their growth or acting as pathogens. The success of each interaction depends on many factors, including the influence of other microbes in the community (e.g., competition which leads to disease suppression). In addition, soil microorganisms are intimately associated with the soil mineral and organic phases, and can influence the soil environment (e.g., through particle aggregation, and contributing to the dynamics of water, and air).

Due to the key roles of the soil microbiota in soil functioning, as well as its sensitivity to environmental changes (e.g., resulting from anthropogenic activities), the study of microbial functions in soil has become of great interest in soil science. The study of microbial ecology, including microbial activity, began during the 19th century. Since then, more and increasingly sophisticated tools have allowed us to shed light on what is known as the “black box” of microbial activity. This chapter revisits the most influential techniques and methodologies that have helped the understanding of microbial functions in soil. These techniques are organized in terms of their application to soil sciences (i.e., the activity or function being measured).

Traditional techniques to study microbial function in soil

This section includes techniques that were developed in the mid to late 20th century to measure biochemical reactions carried out by active microbial cells or their extracellular enzymes (hence the name “biochemical techniques”). These techniques have been key

to the study of biogeochemical cycling of elements in soil and are among the first ones to describe soil microbial communities without relying on isolation and culturing, which can only recover a small fraction of the microorganisms present in soil.

Many of the techniques described in this section are “black box” measurements, meaning that substrates and products are quantified without exact knowledge of the reactions or organisms involved. These techniques usually estimate the activity carried out by a wide array of organisms and enzymes and are therefore representative of the whole microbial community. Measurements may target general parameters (e.g., respiration, N mineralization), that characterize the overall microbial activity and are directly related to microbial biomass. Other, “specific” techniques, measure the activity of hydrolytic enzymes, which can be extracellular (exoenzymes) and are not necessarily associated with overall microbial biomass and activity. Often, measurements represent net activity rates, as products of the measured reaction may be consumed or transformed in subsequent reactions. Although all methods described here can be carried out in laboratory incubations (*in vitro*), some can be adapted to measure microbial activity under field conditions (*in situ*). *In vitro* techniques have the obvious advantage of allowing control of the incubation conditions. *In situ* methods reflect real conditions more accurately and can facilitate the incorporation of spatial and temporal variability of these functional measurements.

Despite the great technological advancements of recent years, including the development of functional omics techniques, these traditional approaches are still widely used. This is partly due to their small cost, simplicity, and accessibility, but also because they still constitute the most straightforward way to measure microbial and enzymatic activity in soil. Additionally, many of them already have internationally standardized protocols that facilitate comparisons across research laboratories.

Microbial respiration

Methods for determining soil microbial respiration are among the oldest and most commonly used to measure the global metabolic activity of microbial communities. These assays measure the oxidation of organic matter by aerobic heterotrophic microorganisms without supplementation of organic carbon substrates (i.e., basal respiration). Aerobic respiration uses oxygen (O_2) as an electron acceptor and generates CO_2 and water as end products. Both O_2 consumption and CO_2 production can be measured to estimate soil microbial respiration, but the latter is more sensitive due to the smaller concentration of CO_2 in the atmosphere. *In vitro* methods can be either static (i.e., a closed incubation system) or dynamic (i.e., continuous aeration with CO_2 -free air). In addition, CO_2 measurements can be taken once, after a set period of time (~24–72 h), or several times during a single incubation. For longer incubations (e.g., weeks or months), measurements are carried out at regular intervals or continuously (e.g., with respirometers or gas flux chambers). Emitted CO_2 can be captured in an alkali trap (NaOH) followed by chemical titration or electrical conductivity measurements (electro titration), or by measuring the headspace CO_2 concentration using gas chromatography or infrared gas analyzers. A commonly used static approach combines the incubation of soil samples with NaOH traps in air-tight containers, followed by precipitation with $BaCl_2$ or $SrCl_2$ and titration with acid. For *in vitro* approaches, incubations are conducted under optimized temperature and moisture conditions. Alternatively, the quick CO_2 -Burst or 24 h evolved CO_2 assay uses thin-layer gel chromatography and a pH-sensitive gel to provide a colorimetric measure of the respiratory response. *In vitro* basal respiration can be used to compare the general activity of different soils, although it is also highly correlated with the availability of mineralizable C. For that reason, some protocols recommend the pre-incubation of samples for 7–10 days before carrying out the assays. In addition to general microbial activity, information about ecosystem maturity or disturbance level can be obtained from the metabolic quotient (qCO_2), calculated as the ratio of microbial respiration to biomass. Standardized protocols for soil respiration measurements are available in ISO 16072:2002 and ISO 17155:2012. Field methods are also available to measure soil respiration at the ecosystem level, but these will measure net ecosystem exchange of CO_2 , including the respiration of all present organisms, minus the CO_2 used in photosynthesis, when applicable. Some field methodologies include static or dynamic chambers, solid-state infrared gas analyzer (IRGA) probes buried in soil, and eddy covariance.

Catabolic capacity and biodegradation

Different methodologies are available to assess the metabolic response of soil microbial communities to the incorporation of an organic carbon substrate. These may employ simple substrates (e.g., sugars, amino acids, organic acids), or evaluate the degradation of complex substrates (e.g., plant litter). While the simple-substrate methods have been primarily used to assess catabolic capacity or diversity, the latter focus on the capacity of a soil community to degrade a complex substrate of interest.

Simple substrate utilization assays

Substrate-induced respiration (SIR) is known for its utility as a microbial biomass measurement. However, the respiratory response of a microbial community to the addition of a simple substrate can provide information on the capacity of the community to metabolize a given substrate. Additionally, evaluating the respiratory response of microbes to the addition of a range of carbon substrates (i.e., carbon source utilization patterns) can be useful to estimate the catabolic capacity or diversity of a community. All these methods measure potential activity levels under optimized incubation conditions and non-limiting substrate concentrations. A simple way to carry out this analysis is to run multiple SIR assays in parallel (*multi-SIR*), by amending soil with different C substrates, incubating the resulting soil slurries for a pre-determined amount of time and measuring CO_2 in the headspace (Degens and Harris, 1997). Less labor-intensive methods are the high-throughput techniques that use different types of microtiter plates (microplates). These methods are known widely as community-level physiological profiling (CLPP), although

the term is sometimes used to refer exclusively to one specific methodology using Biolog™ plates or similar approaches (Garland and Mills, 1991). *Biolog-CLPP* is carried out in soil suspensions amended with different carbon sources and with a redox indicator dye to detect the production of CO₂. These plates can then be analyzed in a microplate reader or spectrophotometer. For that reason, the method is considered culture-dependent, which is prone to selective enrichment, meaning that it is more likely to detect fast-growing organisms (e.g., bacteria). A fluorescence-based method carried out in a BD Oxygen Biosensor system (*BDOxy-CLPP*) solved some of these limitations, although its use was never as widespread (Garland et al., 2003). *BDOxy-CLPP* improves sensitivity and enables testing of lower substrate concentrations compared to other methods (multi-SIR, *Biolog-CLPP*, *MicroResp*™), by measuring the depletion of oxygen, which is less soluble in water. Furthermore, in comparison to *Biolog*, *BDOxy-CLPP* allows testing of soil suspensions that are more turbid, and even of intact soil “microbags” by using a fluorophore located at the bottom of the wells. In contrast, a different method called microrespirometry or *MicroResp* works directly on soil samples (i.e., culture-independent) (Campbell et al., 2003). With *MicroResp*, the amended samples are in a deep-well microplate, while a smaller microplate on top of it has the detection gel with the redox indicator dye, where it is sealed in a way that each sample well is only connected to one detector well. This top microplate can be read with the same microplate readers used for *Biolog-CLPP*. Even though *MicroResp* also requires relatively large substrate concentrations, it can be coupled with ¹⁴C-labeled substrates to work at smaller concentrations.

Complex substrate biodegradation assays

Biodegradation assays can be used to measure the degradation or decay rate of complex substrates (e.g., plant biomass, cellulose, lignin, pollutants, xenobiotics) to simpler intermediates or end products. Degradation of these complex substrates can be carried out in one or several steps by living cells or extracellular enzymes. Thus, measurements can reflect the activity of a single microbial species, a group of microbes, or the presence of extracellular enzymes capable of degrading the added substrate or its intermediates. Commonly, methods involve the addition of a specific substrate to a soil sample to then measure the rate of decomposition as weight loss or change in substrate or product concentration. A common procedure to evaluate plant biomass decomposition is the *litterbag* method (Schinner, 1996). Plant litter is weighed into nylon mesh bags that can be positioned on or below the soil surface (sometimes stratified, at different depths). The mesh allows air, water and nutrients to be exchanged freely with the surrounding soil, and different mesh sizes can be used to include or exclude different organisms (e.g., fauna). After days or months the residual weight of dried litter is measured and a linear decomposition rate is calculated from the difference relative to the initial dry weight. Changes in chemical composition can be also assessed using the same litter. In biodegradation assays, including *litterbag*, substrates can be labeled with radioactive or stable isotopes. These assays can be carried out under optimized conditions (*in vitro*, microcosms or mesocosms) or in the field.

Microbial nitrogen transformations

Biochemical methods to study microbial N transformations are available for four major steps of N cycling: ammonification, nitrification, denitrification and N₂ fixation.

Nitrogen mineralization and nitrification

Generally speaking, N mineralization is understood as the conversion of organic to inorganic N and it starts with the depolymerization, or aminization, of large N-containing compounds (e.g., proteins) to N-containing monomers (e.g., amino acids, nucleic acids). Then, heterotrophic microorganisms hydrolyze those simpler organic N compounds to ammonia (NH₃) in a process known as ammonification. As ammonia can be rapidly oxidized to nitrate in a two-step process known as nitrification these processes are commonly analyzed together. Autotrophic nitrification occurs in two steps: the oxidation of ammonia to nitrite (NO₂⁻) by particular bacterial and archaeal species, followed by the oxidation of nitrite to nitrate (NO₃⁻) by some bacteria. Under certain conditions, heterotrophic nitrification can be carried out by bacteria and fungi. Soil N mineralization is generally quantified as the production of mineral N (ammonia and nitrate). Nitrite is not usually measured because its concentration in soils is very small. One type of *in vitro* testing involves incubations under aerobic conditions with optimized temperature and moisture levels, followed by mineral N extraction and quantification. In incubation-leaching methods, soil samples, with or without organic amendments, are incubated in a column and leached mineral N is regularly determined. Some techniques exclusively quantify *ammonification*, which is usually the rate limiting step of N mineralization. This is conducted by incubating samples under waterlogged conditions and quantifying the production of ammonia (Kandeler, 1996). Anaerobic conditions inhibit nitrification and may also reduce immobilization and protease activity. Methods that quantify ammonification can also be used to estimate mineralizable N, as the substrate is limiting, or it can be carried out with amendments such as arginine (i.e., arginine ammonification method). Alternatively, potential *nitrification* can be measured by incubating samples with ammonium amendments. Field measurements of soil N mineralization usually require working with buried samples in columns (undisturbed) or buried bags (disturbed), with ammonium and nitrate content being determined by laboratory analysis. Ion-exchange resins can be applied to the bottom of columns to account for leaching losses, and at the top to account for deposition (atmospheric, run-off). Depending on the protocol, mineral N can be quantified via spectrophotometry, acid titration or using an automated analyzer. All these assays measure net N mineralization, as some ammonia and nitrate will be immobilized or transformed via other reactions. Gross mineralization can be estimated using labeled substrates. Standardized protocols for N mineralization are available in ISO 14238:2012.

Denitrification

Denitrification is the reduction of nitrate to nitrogen gases, primarily nitrous oxide (N_2O) and dinitrogen (N_2). In this stepwise process, nitrate (NO_3^-), nitrite (NO_2^-), nitric oxide (NO) and N_2O can act as electron acceptors in anaerobic respiration. Although, this process is not restricted to exclusively anaerobic soils, and can occur in hotspots or microsites where O_2 is limiting. Depending on the organism and the environmental conditions, denitrification can be complete (N_2 as product) or incomplete (N_2O or NO as products). Denitrification can be carried out only by active microbial cells, mostly heterotrophic prokaryotes but also, as found more recently, by some fungi and autotrophic bacteria (i.e., nitrifier denitrification). The most common method to measure denitrification is through the *acetylene-inhibition technique* (AIT) (Alef, 1995b). In AIT, the N_2O -reductase enzyme is inhibited by acetylene and the released N_2O , equivalent to losses as both N_2O and N_2 , is measured using gas chromatography. Incubations are carried out in air-tight containers under optimized temperature and moisture conditions for between 48 h (actual rates) and 200 h (potential rates), in sieved soil samples or undisturbed soil cores. Potential denitrification activity can also be measured when adding nitrate as a substrate. A variant of the AIT, known as the soil cover method, allows the making of this estimation in situ, under field conditions. There are also several strategies to distinguish autotrophic from heterotrophic denitrification, for example, using inhibitors, ^{15}N labeling, or different acetylene concentrations. ^{15}N -labeled substrates can also allow gross denitrification rates to be measured, as opposed to net denitrification rates. Standardized protocols are available in ISO 20131-1:2018 and 20131-2:2018.

Nitrogen fixation

The reduction of atmospheric dinitrogen (N_2) to ammonia (NH_3) by nitrogenases is commonly known as N_2 fixation or biological N fixation. This process is exclusively carried out by prokaryotes both symbiotically and non-symbiotically (i.e., free-living). The most popular method to quantify this process is via the *acetylene reduction method*. Acetylene can be reduced by the nitrogenase enzyme complex, yielding ethylene as a product. The assay consists of the incubation of soil samples in airtight containers with an atmosphere enriched in acetylene, under optimized temperature and moisture conditions. The ethylene produced can be measured in gas samples using gas chromatography. Fixed-N can be obtained using a conversion factor obtained with $^{15}\text{N}_2$ assays. For accurate results, it is recommended that the proportion of ethylene produced by microbes via other processes is extracted and quantified (e.g., inhibiting acetylene reduction by nitrogenases with carbon monoxide). The acetylene reduction method can be complemented with glucose amendments. This assay has been carried out on intact soil cores, sieved soil samples, or in situ. Actual N fixation rates can be obtained by using dinitrogen labeled with the stable isotope ^{15}N (see below). No standardized protocols are available.

Enzymatic activity

Hydrolytic enzymes break macromolecules down to smaller molecules (e.g., cellulose to sugars, proteins to amino acids). This can take place intracellularly or attached to active cells, intracellularly in dead or dormant cells, or extracellularly (e.g., adsorbed to soil colloids). While most of the extracellular enzymes are of microbial origin, the measured activities are not necessarily associated with active microorganisms because active enzymes can be stabilized on soil colloids. Most common methods are based on the use of defined substrates (mostly synthetic) with the measurement of the reaction product carried out under optimized conditions (e.g., temperature, pH, incubation time). This means the data generated by these methods describe potential enzymatic activities and not actual activity rates. These accessible tools to measure soil enzymatic activity became available in the 1960s–70s and have been widely used ever since. Enzymatic methods commonly used *in vitro* are listed in Table 1.

Dehydrogenase activity

Dehydrogenases are intracellular enzymes that catalyze oxidation-reduction reactions in the electron transport chain during the aerobic oxidation of organic substances. Consequently, measurements of dehydrogenase activity represent total oxidative activity of aerobic heterotrophic microorganisms and can be used as a measurement for general microbial activity. The assay is based on the use of synthetic electron acceptors instead of O_2 : 2,3,5-triphenyltetrazolium chloride (TTC) or 2-(4-iodophenyl)-3-(4-nitrophenyl)-5-phenyltetrazolium chloride (INT) (see Table 1). The products of these reactions, triphenyl formazan (TPF) and idonitrotetrazolium chloride (INF), respectively, can be measured by spectrophotometry.

Hydrolysis of fluorescein diacetate

Fluorescein derivatives such as fluorescein diacetate or 3,6-diacetyl-fluorescein (FDA) can be hydrolyzed by extracellular and membrane-bound lipases, esterases and proteases (Table 1). These non-specific hydrolases are produced by a wide range of microorganisms, which is why the hydrolysis of FDA is considered an estimator for global hydrolytic activity. This method consists of incubating soil samples with FDA under optimized temperature and pH, extracting the produced fluorescein and quantifying it via spectrophotometry. Standardized methods are available in ISO 23753-1:2019 and 23753-2:2019.

Table 1 Most commonly used assays of soil enzymatic activity

Enzyme or process	Enzyme Commission Number (EC) ^a	Reaction/s targeted	Original publication/s	Available protocols
General activity				
Dehydrogenase	1.1.1.	Oxidation of OM by several oxidoreductases. Used as measurement of overall microbial activity.	Lenhard (1956), Thalmann (1968), von Mersi and Schinner (1991)	Schinner et al. (1996), Bloem et al. (2006), Alef and Nannipieri (1995) ISO 23753:2019
Hydrolysis of fluorescein diacetate		Global hydrolytic activity (lipases, esterases and proteases).	Adam and Duncan (2001)	Bloem et al. (2006), Alef and Nannipieri (1995)
C-cycling enzymes				
Cellulases	3.2.1.4,	Cleavage of cellulooligosaccharides from non-reducing ends (exo), cleavage of glucosidic linkages (endo)	Hope and Burns (1987), Schinner and von Mersi (1990)	Schinner et al. (1996), Alef and Nannipieri (1995)
(endoglucanase + exoglucanase + β -2 β -glucosidase (aka cellobiase)	3.2.1.91, 3.2.1.21			
Saccharase (aka invertase)	3.2.1.26	Hydrolysis of sucrose to fructose and glucose	Hoffman and Dedeken (1965), Eivazi and Tabatabai (1988) Schinner and von Mersi (1990), Hoffman and Pallauf (1965)	Schinner et al. (1996), Alef and Nannipieri (1995)
Xylanase	3.2.1.8	Hydrolysis of xylans to oligosaccharides (<i>endo</i> -) and oligosaccharides to monomers (<i>exo</i> -)	Schinner and von Mersi (1990)	Schinner et al. (1996), Alef and Nannipieri (1995)
P-cycling enzymes				
Phosphomonoesterases (phosphoric monoester hydrolases)	3.1.3.-	Hydrolysis of organic phosphomonoesters to phosphate	Tabatabai and Bremner (1969), Eivazi and Tabatabai (1977)	Schinner et al. (1996), Alef and Nannipieri (1995) ISO 20130:2018
Phosphodiesterase (phosphoric diester hydrolases)	3.1.4.-	Hydrolysis of organic phosphodiester to phosphomonoesters	Eivazi and Tabatabai (1977), Browman and Tabatabai (1978)	Schinner et al. (1996), Alef and Nannipieri (1995)
N-cycling enzymes				
Proteases	3.4.21.-	Proteins to polypeptides, oligopeptides to amino acids	Hoffmann and Teicher (1957), Ladd and Butler (1972)	Schinner et al. (1996), Alef and Nannipieri (1995)
Amidase	3.5.1.4	Hydrolysis of amides to ammonia and carboxylic acid	Frankenberger and Tabatabai (1980)	Alef and Nannipieri (1995)
Urease	3.5.1.5	Hydrolysis of urea to ammonia and carbon dioxide	Tabatabai and Bremner (1972), Kandeler and Gerber (1988)	Schinner et al. (1996), Alef and Nannipieri (1995)
Nitrate reductase	1.7.99.4	Dissimilatory reduction of nitrate to nitrite (anaerobic conditions)	Abdelmagid and Tabatabai (1987)	Schinner et al. (1996)
Arylamidase	3.4.11.2	Hydrolysis of N-terminal amino acid from peptides, amides or arylamides	Acosta-Martínez and Tabatabai (2000)	
S-cycling enzymes				
Arylsulfatase	3.1.6.1	Hydrolysis of aryl sulfates (organic aromatic sulfate esters) to sulfate.	Tabatabai and Bremner (1970)	Schinner et al. (1996), Alef and Nannipieri (1995)

^a<https://www.brenda-enzymes.org/ecexplorer.php?browser=1>.

Specific enzymatic activities

Other intra- but mostly extracellular hydrolases are involved in phosphorus, nitrogen, sulfur and carbon cycling. Related assays commonly employ substrates that change color or gain fluorescence as products of the enzymatic activities accumulate to then be assessed spectrophotometrically or fluorometrically (see Table 1). Conversely, quantification of substrate depletion, rather than production can be measured but these assays are less common. Many colorimetric methods use synthetic substrate analogs such as conjugates of *p*-nitrophenol. Fluorometric methods are more sensitive as they use an artificial fluorescent molecule bound to one or more natural molecules. Fluorogenic substrates are conjugates of 4-methylumbelliferone (MUB) and 7-amino-4-methyl coumarin (AMC). Bench-scale and high-throughput microplate methods are available for both colorimetric and fluorometric approaches. Standardized protocols for high-throughput colorimetric analysis of several hydrolases (e.g., arylamidase, arylsulfatase, β -galactosidase, α -glucosidase, β -glucosidase, *N*-acetyl-glucosaminidase, phosphatases and urease) can be found in ISO 20130:2018.

Isotopic techniques to study microbial function in soil

Isotopes are variations of the same chemical element differing in their number of neutrons and thus mass. Stable isotopes are non-radioactive (i.e., they do not decay over time or emit ionizing radiation), which makes them safer to manipulate than radioactive isotopes also known as radioisotopes. In contrast, radioisotopes decay over time, emitting ionizing radiation until reaching a stable state. Labeling specific substrates with a stable or radioactive isotope makes it possible to trace their fate and utilization by soil microorganisms and through the soil food web. Isotopic labeling addresses some of the limitations imposed by the traditional techniques described earlier. For example, it allows calculation of gross transformation rates and substrate residence times. Some of these techniques (e.g., stable isotope probing) assist with the elucidation of “black box” microbial activity mentioned earlier, by detecting which organisms are using the compound of interest. An additional application of stable isotopes in microbial ecology is to study the natural isotopic signature of compounds and their isotopic fractionation.

Stable isotopes techniques

There are two main approaches to working with stable isotopes. The first is based on their natural abundance and fractionation, while the other consists of enriching a sample with a particular stable isotope to study its transformation or use by specific microbial groups. The most studied stable isotopes in soil microbial ecology are C and N, and to a lesser extent S and O. All these elements exist in nature mostly in one form (^{12}C , ^{14}N , ^{32}S and ^{16}O), but they also have less abundant stable isotopes that are useful to study their dynamics in nature (^{13}C , ^{15}N , ^{34}S and ^{18}O). Because phosphorus only has one stable isotope (^{31}P), no stable isotope techniques are available for it. These elements also have the advantage of being relatively light, which enhances differences in isotopic fractionation, in addition to being relevant to many biological processes. The isotope ratio (δ , in in parts per thousand (‰) relative to a standard) is generally quantified by converting the compounds to simple gases (e.g., CO_2 , SO_2 , N_2) followed by isotope ratio mass spectrometry (IRMS).

Natural abundance and isotopic fractionation

In nature, preferential depletion or enrichment of one isotope in comparison to the others leads to changes in isotopic proportions. This process is known as isotopic fractionation and can be used to measure changes in the isotope ratio (δ), which represents the abundance of an isotope relative to an international standard (‰) (e.g., $\delta^{13}\text{C}$, $\delta^{34}\text{S}$, $\delta^{15}\text{N}$). In biological reactions, it is thermodynamically more favorable to use lighter isotopes over heavier isotopes and, therefore, lighter isotopes are usually enriched in the product of a reaction compared to the substrate. Isotopic fractionation can thus be a useful method to identify if a reaction was carried out by microorganisms or not. Isotopic composition, in particular ^{13}C and ^{15}N enrichment, can allow the elucidation of trophic levels of soil organisms, as higher trophic levels are more enriched in heavier isotopes than lower levels. This occurs because heavier isotopes are more selectively incorporated into the tissues of these organisms and then concentrated due to the process of biomagnification. Lighter isotopes, on the other hand, are predominantly lost as they preferentially participate in biochemical processes that transform them into lighter molecules more easily lost or respired. Another widespread approach using stable isotopes makes use of the different natural ^{13}C abundance or ^{13}C to ^{12}C ratio of C_3 and C_4 plants (–40‰ to –20‰ vs. –17‰ to –9‰, respectively) to evaluate decomposition and fate of plant inputs. The isotopic signature of each vegetation remains in soil organic matter even if microbial transformations cause some fractionation.

Isotopic enrichment

Artificially enriching substrates with stable isotope tracers to track their fate makes them useful to estimate gross, as opposed to net, microbial transformation rates. Since these methods require the labeling of a specific compound, they are less suitable to analyze general microbial activities. In C cycling studies, labeling can be done by enriching simple substrates (e.g., glucose, herbicides) or more complex substrates (e.g., plant residues) with ^{13}C (Gerzabek et al., 1996). In plant-soil systems, pulse labeling of plants with $^{13}\text{CO}_2$ has been applied to study the final fate of plant-derived C in soils. Tracing experiments with the stable isotope ^{15}N has improved our understanding of several N-cycling processes in soil (Brumme and Aden, 1995). For example, ^{15}N -labeled fertilizers, or litter of plants grown with such fertilizer, can be used to track their final fate in soils or plants. Likewise, denitrification losses can be estimated after applying ^{15}N -fertilizer to soils and measuring labeled denitrification products ($^{15}\text{N}_2$ and $^{15}\text{N}_2\text{O}$). Gross N nitrification or N mineralization rates can be assessed using an isotopic dilution technique. Here, the soil nitrate or ammonium pools are labeled and the abundance and isotopic ratio of the pool is measured as it is diluted by unlabeled N from mineralization and nitrification. Isotopic enrichment experiments have also been used to measure rates of heterotrophic and autotrophic nitrification, separately. Another common application of the isotope ^{15}N is to estimate biological N_2 fixation. While the use of ^{15}N has been carried out more frequently in plants, it has also been used to estimate the amount of N derived from symbiotic associations with diazotrophic bacteria. To estimate N_2 -fixation in soil samples, samples are incubated in an atmosphere containing $^{15}\text{N}_2$, and then the ratio of ^{14}N to ^{15}N in the soil is measured. The isotope ^{18}O has also been applied in soil microbiology and biochemistry, for example, to study P cycling using $\text{P}^{18}\text{O}_2^{4-}$ or microbial growth and degradation using H_2^{18}O (Tamburini et al., 2012; Zhran et al., 2021).

Stable isotope probing

Stable isotope probing (SIP) is based on the labeling of substrates with a stable isotope followed by the quantification of microbial biomarkers, such as phospholipid fatty acids (PLFA), DNA, RNA or proteins (Radajewski et al., 2000). Labeling commonly involves ^{13}C , although ^{15}N and ^{18}O have also been used. Substrates are usually chosen based on the microbial group of interest (e.g., to target a specific functional group). Once the labeled substrate is applied to the soil sample and incubated, microbial communities can be studied by extracting the biomarker or cellular component (e.g., DNA), separating lighter and heavier fractions using density-gradient centrifugation, and analyzing those fractions as needed (e.g., DNA sequencing) (Fig. 1). By integrating the use of specific substrates with the identification of microbes capable of utilizing them, SIP constitutes a useful tool to link structural and functional aspects of soil microbial communities. A common application of SIP is to evaluate the use and incorporation of a labeled substrate by different microbial groups within a complex community.

Radioisotope techniques

Radiotracer techniques

Radiotracers are compounds that have been labeled with a radioactive isotope of an element. The presence of the radiotracer in different pools or fractions can be detected by measuring radioactivity, usually by liquid scintillation counting. Like stable isotope techniques, these methods track the fate of a specific compound and calculate gross turnover rates. Additionally, the greater sensitivity of radiotracer methods allows detection of smaller substrate concentrations that are more representative of real conditions in soil. However, radiotracer techniques are more useful for focusing on the transformation of a specific compound, than to measure general activity. The radioisotopes most used in soil microbial ecology are ^{14}C , ^{35}S , ^{32}P and ^{33}P , with half-lives long enough to allow for accurate quantification. The most restrictive half-life is from ^{32}P (~14 days), followed by ^{33}P (~24 days) and ^{35}S (~88 days), while ^{14}C has a half-life of around 5000 years. There are no long-lived radioisotopes of nitrogen and oxygen, which is why radiotracer techniques for these two atoms are not widespread. The radioisotope ^{14}C can be used to label substrates (e.g., glucose, cellulose, plant residues, pesticides) and track their use by soil microorganisms (Alef, 1995a). For example, an organic substrate labeled with ^{14}C is added to the soil and the evolved $^{14}\text{CO}_2$ from respiration is measured. The stable isotope ^{13}C is also available for pulse labeling of plants with $^{13}\text{CO}_2$, which facilitates tracking the use of plant-derived C compounds by the soil biota. Moreover, isotopic dilution techniques are available to measure gross nutrient transformation rates in soil. For example, ^{35}S has been applied to study the use and transformation of sulfur compounds in soil, including: immobilization and mineralization and plant uptake of sulfate ($^{35}\text{SO}_4^{2-}$), reduction to hydrogen sulfide (H_2^{35}S) by sulfate-reducing bacteria during anaerobic respiration (e.g., in flooded soils), and oxidation of elemental sulfur ($^{35}\text{S}^0$) (McLaren et al., 1985). Radiotracer techniques for phosphorus

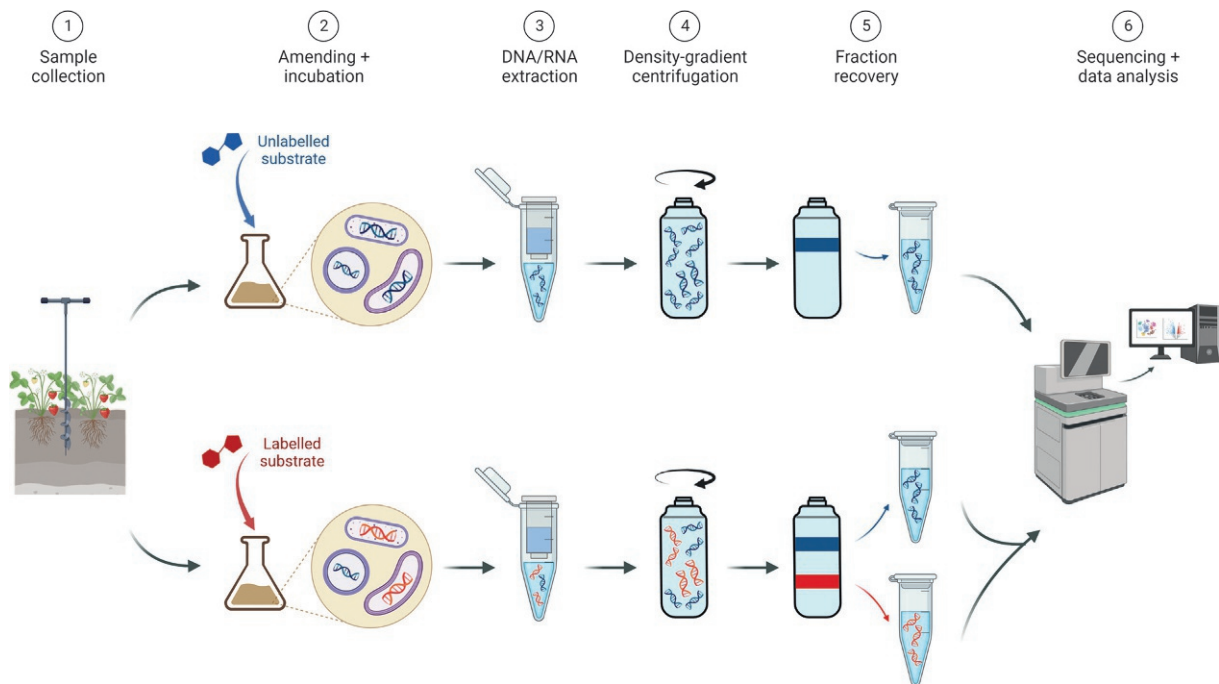


Fig. 1 Schematic overview of stable isotope probing for nucleic acid analyses. This analysis is frequently carried out on environmental samples such as soil ①, which are incubated with a ^{13}C - or ^{15}N -labeled substrate ②. Following incubation, DNA or RNA is extracted from the samples ③ and the heavier labeled nucleic acids from the organisms who assimilated the labeled substrate are separated from the lighter unlabeled fraction using density gradient centrifugation ④. Both fractions are then recovered and used for downstream analysis (e.g., sequencing) ⑤ and ⑥. The flow at the top represents a control with unlabeled substrate which is carried out in parallel. Created with Bio Render (<https://biorender.com/>).

usually involve labeling phosphate fertilizers ($^{32}\text{PO}_4^{2-}$ or $^{33}\text{PO}_4^{2-}$) to study P transformations in plant-soil systems such as immobilization, mineralization, movement and plant or mycorrhizal uptake (Harrison, 1982; Zapata and Axmann, 1995). Radiotracer techniques can be carried out in laboratory incubations, microcosms, and field studies.

Thymidine and leucine incorporation

Incorporation of thymidine, labeled with the radioisotope ^3H ($[^3\text{H}]$ thymidine), or ^{14}C -labeled leucine ($[^{14}\text{C}]$ leucine), has been used to estimate gross bacterial growth rate in soils (Bloem and Bolhuis, 2006). Since thymidine is a precursor of the DNA nucleotide base thymine, its incorporation gives information about DNA synthesis. On the other hand, leucine is an amino acid and its incorporation is indicative of protein synthesis. These methods are more suitable indicators of bacterial growth, rather than simply measuring biomass, because they assess gains and losses (e.g., death or predation). Although these isotope methods were developed independently, when carried out simultaneously they lead to greater accuracy, as they both have different advantages and limitations. For example, thymidine is only incorporated by bacteria, although it excludes some groups, whereas leucine can be incorporated by bacteria but also other microbial groups.

Omics techniques to study microbial function in soil

Despite the major role that soil microorganisms play in soil and ecosystem functioning, understanding the functioning of complex soil microbial communities is still a major challenge owing to the high proportion of non-culturable taxa. In recent decades, the advancement of next-generation sequencing (NGS) technologies led to the emergence of new scientific disciplines known as omics such as metagenomics, metatranscriptomics, metaproteomics and metabolomics. Omics techniques enable the analysis of the functional traits of microbial communities in soil. Omics are defined as high-throughput biochemical assays that comprehensively and simultaneously measure molecules of the same type from a biological sample. These omics approaches delve into the different processes of DNA-to-protein translation, as well as protein and metabolite production (Fig. 2), which is caused by the functioning of the microbial community in different contexts (e.g., bulk soil, rhizosphere, wetlands). As such, omics techniques and their integration are promising tools to illuminate the relationship between structure and function in complex microbial communities.

Sequence-based metagenomics

Whereas metagenomics is the study of all genetic material from a mixed community of organisms that are recovered directly from the environment, sequence-based functional metagenomics refers to the analysis of metagenomic data focusing on the abundance of gene families, pathways and systems, instead of the taxonomy. These analyses facilitate studying the functional capabilities of soil communities in different contexts (e.g., environmental conditions, treatments). For metagenomic functional profile analyses, one of the most used approaches is the DIAMOND + MEGAN pipeline. DIAMOND is a high-throughput DNA-to-protein alignment tool that rapidly aligns translated DNA sequences against a reference database of protein sequences such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. The output of DIAMOND is then processed using MEGAN, which is a highly efficient software for interactive analysis of the functional gene content of the microbiomes from hundreds of samples. MEGAN supports the use of metadata and provides several databases for functional profiling such as SEED, eggNOG, KEGG and InterPro2GO (Huson et al., 2016). The selection of a database for annotation of functional genes depends on the type of samples under analysis (i.e., according to similar studies on the subject). All these databases are up-to-date and mostly share similar annotations, but their individual characteristics may be critical for certain data under study. For instance, Clusters of Orthologous Genes (COG) only includes information about bacteria and archaea, while EuKaryotic Orthologous Groups (KOG) is a eukaryote-specific version of the COG, and eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) database generates functional annotations for eukaryotic, prokaryotic and viral sequences. There are also other bioinformatics pipelines and platforms for metagenome analysis that are frequently upgraded with a focus on optimizing metagenome data elucidation (Meyer et al., 2022).

Functional metagenomics

In contrast to sequence-based metagenomics, functional metagenomics uses a very specific function-based “wet-lab” method that enables reliable assessment of a specific function of the microbiome. This assessment is based on the phenotypic selection or screening of clones harboring the DNA sequence encoding that function (Charles et al., 2017). In this approach, the DNA is extracted directly from environmental samples and subsequently cloned into a suitable vector to create DNA libraries. The libraries are inserted into alternative host cells that, once replicated under controlled conditions, can express the target DNA (e.g., screening soil metagenomic libraries for antimicrobial activities) (Biver et al., 2013). This approach is particularly important for the discovery of novel genes encoding microbial functions in soil and can contribute to the annotation of genome and metagenome sequences. Therefore, functional metagenomics has the potential to make major contributions to the study of soil microbiota functioning and the development of biotechnology solutions. The most common expression platform used in functional metagenomics is *Escherichia coli*, however, the quest for robust and flexible screening platforms is an active process in functional metagenomics research (Lewin et al., 2017).

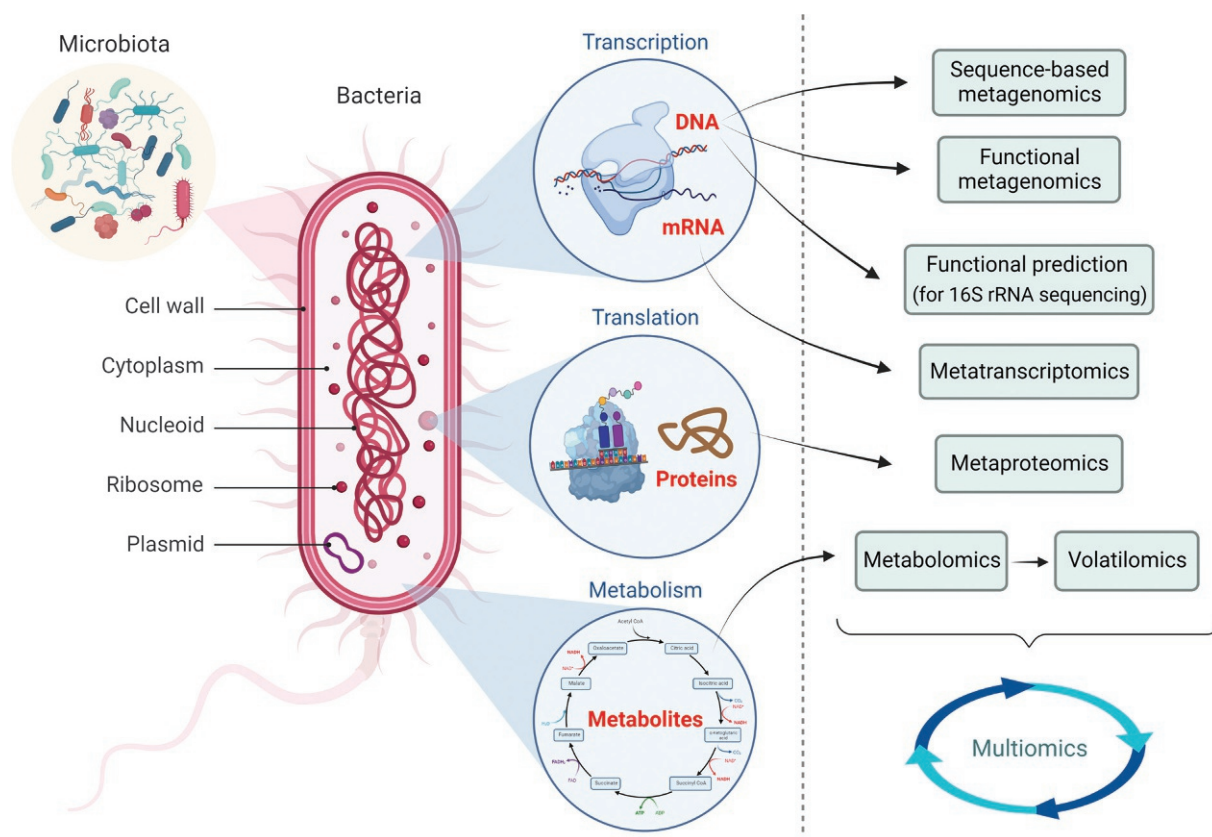


Fig. 2 Schematic diagram of biological molecules targeted from prokaryotic communities by different functional omics technologies. DNA is contained in the nucleoid of prokaryotes, and by sequencing it, it is possible to reveal the taxonomic identity and functional capabilities of the microbiome (e.g., through *metagenomic functional analysis*). Genetic information is transferred in the form of mRNA through the process of DNA transcription. Thus, using high-throughput RNA sequencing (RNA-Seq) it is possible to estimate potential microbial function as changes in the microbial transcriptome (*metatranscriptomics*). Ribosomal RNA (rRNA), which does not harbor relevant functional data, is highly abundant in RNA samples and hence it is usually removed or depleted before sequencing. Cellular messenger RNA (mRNA) then passes to the ribosomes or rRNA, where genetic information is used to synthesize proteins in a process known as translation. The analysis of the complete set of proteins produced by all microbial cells is known as *metaproteomics*. The synthesized proteins, as well as compounds assimilated by the prokaryotic cells, are combined in complex metabolic pathways that take place in the cytoplasm and cell wall. By analyzing the complete metabolites profile, *metabolomics* brings an instantaneous snapshot of the microbial community functioning. Within metabolomics, *volatilomics* constitutes a non-invasive approach to study microbial volatile organic compounds. The multi-level integration and analysis of different omics datasets (*multiomics*) emerges as a promising approach to explore the association between the genotypical and phenotypical features of microbial communities, and their response to environmental factors. Created with BioRender (<https://biorender.com/>).

Functional gene prediction from amplicon-sequence data

Functional gene prediction arose as an alternative to metagenomics sequencing. It consists of predicting microbial functional gene families from their taxonomic compositions inferred from more cost-effective 16S rRNA amplicon sequencing. This approach takes advantage of the evolution of bioinformatics and machine learning algorithms and the abundant characterization of genomes, functional genes and metabolic functions accumulated in databases over the last few decades. Several different pipelines have been proposed including PICRUSt1, PICRUSt2, Piphillin, PanFP, Tax4Fun2, and FAPROTAX (Sun et al., 2020). For example, in PICRUSt2, the amplicon sequence variants (ASVs) are placed into a reference phylogenetic tree containing more than 20,000 full 16S rRNA genes from bacterial and archaeal genomes. From this, the copy numbers of individual gene-families are predicted for each ASV (Douglas et al., 2020). The main limitation of amplicon-based functional prediction is their dependence on existing reference genomes, which can affect the identification of rare environment-specific functions. The accuracy of functional prediction tools will improve as the availability of high-quality genomes from non-clinical (or environmental isolates) increases.

Metatranscriptomics

Metatranscriptomics is the analysis of the whole microbiome gene expression in a particular environment. It is based on shotgun sequencing of RNA from environmental samples and the subsequent analysis of transcriptome profiles. Since mRNA carries out the

genetic information from the DNA to ribosomal sites of protein synthesis, gene expression profiling (i.e., assessing thousands of genes at once) can provide a snapshot of the microbially mediated processes occurring at a particular moment in a specific environment. Therefore, metatranscriptomics can provide information into the underlying functioning of microbial communities and their responses to environmental conditions. The development and popularity of transcriptomics has been attributed to the evolution of RNA sequencing technologies, passing from hybridization-based microarrays to current NGS sequencing technology (e.g., RNA-Seq), which consists of sequencing of complementary DNA (cDNA). Furthermore, RNA-Seq is also evolving from short to long-read RNA-Seq technologies. Long-read RNA-Seq is more cost-effective and more feasible for de novo transcriptome assembly, isoform expression quantification, and in-depth RNA species analysis. The functional annotation of RNA-Seq data can be made directly using reads or using assembled contigs as the annotation of metagenomic data. In addition, several bioinformatics workflows exist for the RNA-seq data analysis. Some of these workflows start with raw sequencing reads, and provide the functional gene profiling and the differentially expressed transcripts (Shakya et al., 2019).

Metaproteomics

Metaproteomics is the high-throughput identification and quantification of all the proteins recovered in an environmental sample. Since proteins are the results of gene expression and because they constitute the largest cellular material, the metaproteomics screening represents a powerful molecular tool to assess microbial community functioning. Metaproteomics also provides information about community composition (in terms of per-species biomass contribution), carbon sources, and the consumption of substrates by community members. Compared to metatranscriptomics, metaproteomics confirms the presence of proteins. This is particularly relevant to study microbial function because mRNA is not always translated into proteins, the amount of protein produced for a given amount of mRNA can vary according to the physiological status of the cell, and post-translational modifications of proteins (e.g., phosphorylation and glycosylation) can be crucial for their function. Metaproteomics uses shotgun liquid chromatography-tandem mass spectrometry (LC-MS/MS) to identify thousands of peptides and proteins per sample. Metaproteomics identifies peptide data by matching it against protein sequence databases and assigning it to proteins or protein groups, which are further assigned to functional groups using various annotation databases. Several bioinformatics pipelines have been developed to characterize and assess the proteomic profile from microbial community samples (e.g., eggNOG-mapper, MEGAN, MetaGOmics, MPA, Unipept, and ProPHAn). These programs differ in several aspects, and each one has its advantages, but there is currently no evidence on which one is more effective than the other (Sajulga et al., 2020).

Metabolomics

Metabolomics is the identification and quantification of the complete set of metabolites present in biological systems. Metabolomics offers unique advantages relative to the other functional profiling strategies and can be highly complementary with other omics interventions. Changes in metabolite concentrations often result from modified enzyme activity or differences in the rates of nutrient consumption and excretion. Thus, metabolite profiling provides a real-time measurement of biochemical activity, taking place in the subject, at the time of sampling. Metabolomics analyses typically involve extracting metabolites, with the addition of internal standards for calibration purposes, and quantification of metabolites. Quantification is carried out by a separation method, such as high-performance liquid chromatography (HPLC) or gas chromatography (GC), coupled with a detection method such as mass spectrometry (MS) or nuclear magnetic resonance (NMR) spectroscopy. Nevertheless, the use of metabolomics in soil science is still in its early stages, despite its extensive use in biomedical science and plant biology, its application in soils remains uncommon. A particular branch that is gaining the interest of the soil-science community, due to its non-invasiveness, is the volatilomics approach. Soil volatilomics deals with the wide array of volatile organic compounds (VOCs) produced or degraded by living organisms in soils. Because soil VOC emissions are largely microbial, dynamic and sensitive to changes in soil conditions, VOC emission profiles are a promising tool to study the metabolic activity of soil microbial communities (Meredith and Tfaily, 2022).

Multimomics

Multimomics is an integrative approach in which multiple omics datasets (e.g., genome, proteome, transcriptome, epigenome, or metabolome) from the same samples are analyzed in parallel and their findings are explored in a complementary way. The multilevel data mining of diverse omics sources enables insight into the linkages between genotype, phenotype, and environmental factors. Multimomics integration approaches could be conceptual (i.e., insight from several omics data sources for a more comprehensive understanding of the biological system) or statistical (i.e., based on mathematical and computational models). Several statistical tools have been proposed for multimomics data integration and analyze, including correlation-based, network-based, Bayesian network-based, machine learning-based, kernel-based, or matrix factorization-based methods (Santiago-Rodriguez and Hollister, 2021); nevertheless, despite multiple advances, multimomics integrative analyses remains challenging. Deep learning is emerging as the most promising class of machine learning algorithms in multimomics data analysis, particularly in large-scale datasets with high dimensionality. It uses multilevel networks algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to automatically extract meaningful features from the data, and learn intricate relationships (i.e., nonlinear and hierarchical patterns) between different omics datasets that may be difficult to detect using traditional statistical methods

(Picard et al., 2021). An emergent branch of multiomics is single-cell multiomics which encompasses the integrative analysis of diverse single-cell datasets. This approach offers unprecedented resolution into the multiple molecular modalities at the single-cell level. Several recent single-cell data integration approaches have been developed, including Seurat v3, Harmony and iNMF (Gao et al., 2021).

Concluding remarks

We have described a range of methodologies to study soil microbial functions, ranging from traditional techniques that have been used and validated for decades, to novel techniques such as omics tools. While the latter tools harbor great potential, we firmly believe that traditional techniques still have a valuable place in studying soil microbiology. These tools provide affordable and reliable means to evaluate microbial activity in soils, particularly in areas where access to sophisticated omics tools may be limited. Meanwhile, high-throughput tools that collect large volumes of data are becoming not only more accessible but also indispensable to understand the structural and functional complexity of the soil microbiota. As we transition to a future likely driven by more sophisticated technologies and big data, combining the strengths of traditional and novel techniques can provide a more comprehensive understanding of assembly and functions of soil microbial communities. Ultimately, this understanding will allow us to develop effective strategies (e.g., sustainable agricultural practices) to protect and improve soil health.

References

- Alef K (1995a) Aerobic biodegradation of ^{14}C -labelled organic matter in soils. In: Alef K and Nannipieri P (eds.) *Methods in Applied Soil Microbiology and Biochemistry*, pp. 220–222. London: Academic Press.
- Alef K (1995b) Assay of denitrification. In: Alef K and Nannipieri P (eds.) *Methods in Applied Soil Microbiology and Biochemistry*, pp. 285–286. London: Academic Press.
- Biver S, Steels S, Portetelle D, et al. (2013) *Bacillus subtilis* as a tool for screening soil metagenomic libraries for antimicrobial activities. *Journal of Microbiology and Biotechnology* 23(6): 850–855. Available at: <https://doi.org/10.4014/jmb.1212.12008>.
- Bloem J and Bolhuis P (2006) Thymidine and leucine incorporation to assess bacterial growth rate. In: Bloem J, Hopkins D, and Benedetti A (eds.) *Microbiological Methods for Assessing Soil Quality*, pp. 142–149. CABI Publishing.
- Brumme R and Aden G (1995) The use of ^{15}N to study the nitrogen turnover in soils. In: Alef K and Nannipieri P (eds.) *Methods in Applied Soil Microbiology and Biochemistry*, pp. 246–257. London: Academic Press.
- Campbell CD, Chapman SJ, Cameron CM, et al. (2003) A rapid microtiter plate method to measure carbon dioxide evolved from carbon substrate amendments so as to determine the physiological profiles of soil microbial communities by using whole soil. *Applied and Environmental Microbiology* 69(6): 3593–3599. Available at: <https://doi.org/10.1128/AEM.69.6.3593-3599.2003>.
- Charles TC, Liles MR, and Sessitsch A (2017) *Functional Metagenomics: Tools and Applications*. Cham: Springer International Publishing. Available at: <https://doi.org/10.1007/978-3-319-61510-3>.
- Degens BP and Harris JA (1997) Development of a physiological approach to measuring the catabolic diversity of soil microbial communities. *Soil Biology and Biochemistry* 29(9–10): 1309–1320.
- Douglas GM, Maffei VJ, Zaneveld JR, et al. (2020) PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology* 38(6): 685–688. Available at: <https://doi.org/10.1038/s41587-020-0548-6>.
- Gao C, Liu J, Kriebel AR, et al. (2021) Iterative single-cell multi-omic integration using online learning. *Nature Biotechnology* 39(8): 1000–1007. Available at: <https://doi.org/10.1038/s41587-021-00867-x>.
- Garland JL and Mills AL (1991) Classification and characterization of heterotrophic microbial communities based on patterns of community-level sole carbon-source utilization. *Applied and Environmental Microbiology* 57(8): 2351–2359.
- Garland JL, Roberts MS, Levine LH, et al. (2003) Community-level physiological profiling performed with an oxygen-sensitive fluorophore in a microtiter plate. *Applied and Environmental Microbiology* 69(5): 2994–2998. Available at: <https://doi.org/10.1128/AEM.69.5.2994-2998.2003>.
- Gerzabek M, Danneberg O, and Kandeler E (1996) Soil organic matter turnover by carbon-stable isotopes. In: Schinner F, et al. (ed.) *Methods in Soil Biology*, p. 119. Berlin Heidelberg: Springer.
- Harrison AF (1982) 32P-method to compare rates of mineralization of labile organic phosphorus in woodland soils. *Soil Biology and Biochemistry* 14(4): 337–341. Available at: [https://doi.org/10.1016/0038-0717\(82\)90003-7](https://doi.org/10.1016/0038-0717(82)90003-7).
- Huson DH, Beier S, Flade I, et al. (2016) MEGAN community edition—Interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Computational Biology* 12(2): 1–12. Available at: <https://doi.org/10.1371/journal.pcbi.1004957>.
- Kandeler E (1996) N-mineralization under waterlogged conditions. In: Schinner F, et al. (ed.) *Methods in Soil Biology*, pp. 141–143. Berlin Heidelberg: Springer.
- Lewin A, Lale R, and Wentzel A (2017) Expression platforms for functional metagenomics: Emerging technology options beyond *Escherichia coli*. In: Trevor C, Liles MR, and Sessitsch A (eds.) *Functional Metagenomics: Tools and Applications*, pp. 13–44. Springer. Available at: https://doi.org/10.1007/978-3-319-61510-3_2.
- McLaren R, Keer J, and Swift R (1985) Sulphur transformations in soils using sulphur-35 labelling. *Soil Biology & Biochemistry* 17(1): 73–79.
- Meredith LK and Tfaily MM (2022) Capturing the microbial volatilome: An oft overlooked ‘ome’. *Trends in Microbiology* 30(7): 622–631. <https://doi.org/10.1016/j.tim.2021.12.004>.
- Meyer F, Fritz A, Deng Z, et al. (2022) Critical assessment of metagenome interpretation—The second round of challenges. *Nature Methods* 19: 429–440. Available at: <https://doi.org/10.1038/s41592-022-01431-4>.
- Picard M, Scott-Boyer MP, Bodein A, et al. (2021) Integration strategies of multi-omics data for machine learning analysis. *Seminars in Perinatology* 45(6), 151456 Available at: <https://doi.org/10.1016/j.csbj.2021.06.030>.
- Radajewski S, Ineson P, Parekh NR, et al. (2000) Stable-isotope probing as a tool in microbial ecology. *Nature* 403(6770): 646–649.
- Sajulga R, Easterly C, Riffle M, et al. (2020) Survey of metaproteomics software tools for functional microbiome analysis. *PLoS One* 15(11): 1–20. Available at: <https://doi.org/10.1371/journal.pone.0241503>.
- Santiago-Rodriguez TM and Hollister EB (2021) Multi ‘omic data integration: A review of concepts, considerations, and approaches. *Seminars in Perinatology* 45(6), 151456 Available at: <https://doi.org/10.1016/j.semperi.2021.151456>.
- Schinner F (1996) In situ studies of litter decomposition. In: Schinner F, et al. (ed.) *Methods in Soil Biology*, p. 114. Berlin Heidelberg: Springer.
- Shakya M, Lo C-C, and Chain PSG (2019) Advances and challenges in metatranscriptomic analysis. *Frontiers in Genetics* 10: Available at: <https://doi.org/10.3389/fgene.2019.00904>.

- Sun S, Jones RB, and Fodor AA (2020) Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. *Microbiome* 8: 46. <https://doi.org/10.1186/s40168-020-00815-y>.
- Tamburini F, Pfahler V, Bünemann EK, et al. (2012) Oxygen isotopes unravel the role of microorganisms in phosphate cycling in soils. *Environmental Science & Technology* 46(11): 5956–5962. Available at: <https://doi.org/10.1021/es300311h>.
- Zapata F and Axmann H (1995) ^{32}P isotopic techniques for evaluating the agronomic effectiveness of rock phosphate materials. *Fertilizer research* 1995 41(3): 189–195. Available at: <https://doi.org/10.1007/BF00748308>.
- Zhran M, Ge T, Tong Y, et al. (2021) Assessment of depth-dependent microbial carbon-use efficiency in long-term fertilized paddy soil using an $^{18}\text{O}\text{-H}_2\text{O}$ approach. *Land Degradation & Development* 32(1): 199–207. Available at: <https://doi.org/10.1002/ldr.3708>.

Further reading

- Abdelmagid HM and Tabatabai MA (1987) Nitrate reductase activity of soils. *Soil Biology and Biochemistry* 19(4): 421–427. [https://doi.org/10.1016/0038-0717\(87\)90033-2](https://doi.org/10.1016/0038-0717(87)90033-2).
- Acosta-Martínez V and Tabatabai M (2000) Arylamidase activity of soils. *Soil Science Society of America Journal* 64(1): 215–221.
- Adam G and Duncan H (2001) Development of a sensitive and rapid method for the measurement of total microbial activity using fluorescein diacetate (FDA) in a range of soils. *Soil Biology and Biochemistry* 33(7–8): 943–951. [https://doi.org/10.1016/S0038-0717\(00\)00244-3](https://doi.org/10.1016/S0038-0717(00)00244-3).
- Alef K and Nannipieri P (1995) *Methods in Applied Soil Microbiology and Biochemistry*. Academic Press.
- Bloem J, Schouten AJ, Sørensen SJ, Rutgers M, van der Welf A, Breure AM, and Benedetti A (eds.) (2006) Monitoring and evaluating soil quality. *Microbiological Methods for Assessing Soil Quality*, pp. 73–77. CABI.
- Browman MG and Tabatabai MA (1978) Phosphodiesterase activity of soils. *Soil Science Society of America Journal* 42(2): 284–290. <https://doi.org/10.2136/SSSAJ1978.03615995004200020016X>.
- Eivazi F and Tabatabai MA (1977) Phosphatases in soils. *Soil Biology and Biochemistry* 9(3): 167–172. [https://doi.org/10.1016/0038-0717\(77\)90070-0](https://doi.org/10.1016/0038-0717(77)90070-0).
- Eivazi F and Tabatabai MA (1988) Glucosidases and galactosidases in soils. *Soil Biology and Biochemistry* 20(5): 601–606. [https://doi.org/10.1016/0038-0717\(88\)90141-1](https://doi.org/10.1016/0038-0717(88)90141-1).
- Frankenberger WT and Tabatabai MA (1980) Amidase activity in soils: I. Method of Assay. *Soil Science Society of America Journal* 44(2): 282–287. <https://doi.org/10.2136/SSSAJ1980.03615995004400020016X>.
- Hoffman G and Dedeken M (1965) Eine methode zur kolorimetrischen bestimmung der β -glucosidaseaktivität in böden. *Z Pflanzenernaehr Bodenkd* 108: 195–201.
- Hoffman G and Pallauf J (1965) Eine kolorimetrische methode zur bestimmung der saccharaseaktivität in böden. *Z Pflanzenernaehr Düng Bodenkd* 110: 193–201.
- Hoffmann G and Teicher K (1957) Das Enzymsystem unserer Kulturböden VII. Proteasen II. *Zeitschrift für Pflanzenernährung, Düngung, Bodenkd.* 77(3): 243–251. <https://doi.org/10.1002/JPLN.19570770308>.
- Hope CFA and Burns RG (1987) Activity, origins and location of cellulases in a silt loam soil. *Biology and Fertility of Soils* 5(2): 164–170. <https://doi.org/10.1007/BF00257653>.
- Kandeler E and Gerber H (1988) Short-term assay of soil urease activity using colorimetric determination of ammonium. *Biology and Fertility of Soils* 6(1): 68–72. <https://doi.org/10.1007/BF00257924>.
- Ladd JN and Butler JHA (1972) Short-term assays of soil proteolytic enzyme activities using proteins and dipeptide derivatives as substrates. *Soil Biology and Biochemistry* 4(1): 19–30. [https://doi.org/10.1016/0038-0717\(72\)90038-7](https://doi.org/10.1016/0038-0717(72)90038-7).
- Lenhard G (1956) Die Dehydrogenaseaktivität des Bodens als Mass für die Mikroorganismenaktivität im Boden. *Z. Pflanzenernähr. Düng. Bodenkd.* 73: 1–11.
- Schinner F and von Mersi W (1990) Xylanase-, CM-cellulase- and invertase activity in soil: An improved method. *Soil Biology and Biochemistry* 22(4): 511–515. [https://doi.org/10.1016/0038-0717\(90\)90187-5](https://doi.org/10.1016/0038-0717(90)90187-5).
- Schinner F, Öhlinger R, Kandeler E, and Margesin R (1996) *Methods in Soil Biology*. Springer Science & Business Media.
- Tabatabai MA and Bremner JM (1969) Use of p-nitrophenyl phosphate for assay of soil phosphatase activity. *Soil Biology and Biochemistry* 1(4): 301–307. [https://doi.org/10.1016/0038-0717\(69\)90012-1](https://doi.org/10.1016/0038-0717(69)90012-1).
- Tabatabai MA and Bremner JM (1970) Arylsulfatase activity of soils. *Soil Science Society of America Journal* 34(2): 225–229. <https://doi.org/10.2136/SSSAJ1970.03615995003400020016X>.
- Tabatabai MA and Bremner JM (1972) Assay of urease activity in soils. *Soil Biology and Biochemistry* 4(4): 479–487. [https://doi.org/10.1016/0038-0717\(72\)90064-8](https://doi.org/10.1016/0038-0717(72)90064-8).
- Thalmann A (1968) Zur methodik der bestimmung der dehydrogenaseaktivität in boden mittels triphenyltetrazoliumchlorid (TTC). *Landwirtschaftliche Forschung* 21: 249–258.
- von Mersi W and Schinner F (1991) An improved and accurate method for determining the dehydrogenase activity of soils with idonitrotetrazolium chloride. *Biology and Fertility of Soils* 11(3): 216–220. <https://doi.org/10.1007/BF00335770>.

Kinetics of microbial processes: General principles

NS Panikov, Chemistry and Chemical Biology Department, Northeastern University, Boston, MA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	169
State variables and growth parameters	170
Amount of microorganisms	170
The concentration of limiting substrate	171
Growth yield	171
Maintenance energy	172
Energy overflow	172
Cell quota. Variation of yield on anabolic substrates	173
Products	173
Kinetic models of microbial bioprocesses	174
Chemical and enzyme kinetics	174
Exponential growth	175
Logistic growth	175
Monod model	175
Cell quota model	176
Structured models	176
Genome-scale models (GEM)	176
Models of intermediate complexity	178
Growth dependence on environmental factors	180
Hierarchy of factors, the primary significance of the nutrient resources	180
Effects of pH	180
Effects of temperature	182
The special case of near-zero growth kinetics	183
Concluding comments	184
References	184
Further reading	185

Abstract

Kinetics describes rates and mechanisms of processes. Microbiological kinetics studies all dynamic behaviors, such as growth, survival, death, mutations, adaptations, product formation, cell cycles, using a combination of experiments and mechanistic mathematical simulations. This chapter introduces the basic terms, variables, and concepts moving from chemical and enzyme kinetics, to unstructured growth models (logistic, Monod, cell quota), structured models simulating self-adaptive microbial behavior (cybernetic, SCM), and to modern genome-scale reconstructions. Effects of environmental factors are illustrated as an interaction of variable pH and temperature with nutrient concentration. The final topic is the near-zero growth of microbial populations prevailing in the biosphere.

Key points

- Microbiological kinetics aims to elucidate the underlying mechanisms of all dynamic manifestations of microbial life by using mechanistic mathematical models. This covers growth, survival and death, mutations, differentiation of cells into forms with contrasting function or activity, product formation, adaptations of metabolism to changing environments, cell cycling and interactions with other organisms.
- The relationship between microbial biomass and cell numbers or length of mycelium is not straightforward, because the mass of individual cells or width of hyphae vary among different organisms and even those of the same species depending on growth conditions.
- As all potentially available nutrients cannot normally be assessed, the focus is on one or few individual compounds or classes of molecules representing the limiting nutrient substrate, i.e., the nutrients that control the growth and activity of the microorganisms investigated.
- Kinetic models are either deterministic (clear, determined and regular processes) or stochastic (dealing with random variables participating in biological processes). The main variables used in deterministic models (concentrations of cell mass, substrates, and products) are the same as in chemical kinetics, whereas stochastic models consider variables such as probabilities, frequency distributions, variance.

- Under natural conditions in the soil, microbial growth is mostly continuous and very slow because of severe nutrient limitations. Even in eutrophic habitats (e.g., the rhizosphere) microorganisms divide no more than 1–2 times per day, whereas in mesotrophic and oligotrophic habitats the doubling time can be months and years.

Nomenclature

$D=F/V$	Chemostat dilution rate
F	Medium feeding rate
K_s	Saturation constant, at $s=K_s$, $\mu=0.5\mu_m$
m	Maintenance term, the catabolic rate $q_s \rightarrow m$, as $\mu \rightarrow 0$
NZG	Near-zero growth under chronic starvation
ODE	Ordinary differential equation
p	Product concentration
PPA	Plant photosynthetic activity
$q_p=(1/x) \cdot dp/dt$	Specific product formation rate (SPR)
$q_s=(1/x) \cdot ds/dt$	Specific substrate uptake rate (SUR)
ROS	Reactive oxygen species
s	Limiting nutrient substrate concentration
SCM	Synthetic Chemostat Model
SGR	Specific growth rate
t	Time
V	Volume of cell culture kept constant in the chemostat
VBNC	Viable but nonculturable cells
x	Cell mass concentration
$Y=-dx/ds$	Cell yield per unit of consumed substrate
$\mu=(1/x) \cdot dx/dt$	SGR calculation
μ_m	Maximum SGR, $\mu \rightarrow \mu_m$ as $s \rightarrow \infty$

Introduction

Kinetics (from the Greek κινετικός, forcing to move) is a branch of natural science that deals with the rates and mechanisms of any processes, physical, chemical, or biological. Microbiological kinetics studies all dynamic manifestations of microbial life: including growth itself, survival and death, mutations, adaptive reconfigurations of metabolism to the changeable environment, products formation, cell cycles, interactions with other microorganisms or microbial hosts (humans, animals, plants), the functional differentiation of cells into active-or-dormant, motile-or-sessile, resistant-or-susceptible. Contrary to descriptive monitoring of process dynamics, kinetic studies attempt to elucidate the underlying mechanisms by using mechanistic mathematical models. The epithet ‘mechanistic’ implies that the respective model translates the postulated essential point of the studied processes into the mathematical equations with variables and coefficients having clear physical, chemical, or biological meaning (contrary to, say, the polynomial regression equations). Then the comparison of observation and the model’s prediction allows researchers to discard incorrect hypotheses or modify them to resolve the interpretive contradictions.

A hypotheses-driven approach to acquiring knowledge is a common part of any [scientific method](#) established in science since at least the 17th century. Testing of hypotheses can be done against a wide range of data and be based on any revealed inconsistencies. The distinctive feature of kinetics is the added strength of the failed dynamic simulation as inconsistency, at the same time typical kinetic study requires an experimental component of high standard, that should be based on reliable *quantitative dynamic* data, such as precise rates measurements under fully controlled experimental conditions or observations of process dynamics over time, preferentially with continuous logging of a set of data interconnected by cause-and-effect relationships. Descriptive information that cannot be expressed numerically is generally not acceptable.

Microbial kinetics is tightly linked to the growth stoichiometry (from the Ancient Greek στοιχείον ‘element’ and μέτρον ‘measure’), which is the quantitative relationship between substrates and the products of the bioprocess. In practical terms, stoichiometry addresses mainly problems of a static nature, like “how much?” and “in what proportion?” while kinetics considers the dynamics and deals with questions like “how fast?” and “by which mechanism?”

There could be reluctance in applying such a biokinetic approach to environmental and biospheric studies, including soil biology, on the basis that these natural systems are too complex for precise rate measurements and fully controlled experimental conditions. This is not true. The complexity of soil as a part of nature does not prevent its deeper understanding by using precise rate measurements and other quantitative methods. Moreover, the combination of mechanistic mathematical simulations with field

observations, laboratory soil incubations, and planted soil microcosms should be highly beneficial for a better understanding of biospheric processes including on-going global climatic changes. It could even be highly beneficial in the practical developments of the mitigation strategies and novel environmental biotechnologies aimed at establishing sustainable land use.

The first chapter in this two-parts series has the objective of introducing the reader to the basic terms, variables and principles of microbiological kinetics that apply equally to industrial bioreactors, laboratory cultures and soil communities. The second chapter focuses on specific applications of the general principles to soil biology. We recommend that an interested reader read both parts, although the sequence order is not critical!

State variables and growth parameters

There are two types of kinetic models, *deterministic* and *stochastic*. The former describes clear, determined, and regular processes. The latter models deal with random variables participating in bioprocesses. The main variables used in deterministic models are the same as in chemical kinetics - concentrations of cell mass, substrates, and products - while stochastic models consider instead probabilities, frequency distributions, variance, etc. For example, a deterministic model can describe the process of N₂-fixation as a function of energy and nutrient supply to the rhizosphere, while a stochastic model considers the probability of rhizobia infecting legume roots for a given frequency distribution of bacteria across the environmental space. Any real biospheric process has both deterministic and stochastic components, so ideally the kinetic model should be a 'hybrid' and contain a deterministic core with a stochastic interface, generating probability distributions and confidence corridors rather than single values or curves.

The major state variables of kinetic models describing microbial processes in nature are summarized in Table 1 and described in detail below.

Amount of microorganisms

Quantitatively, the abundance of microbes can be expressed as biomass (x) or cell numbers (N) of unicellular organisms (bacteria, yeasts, protozoa, fungal spores) per unit of soil or water mass, volume, or surface area. Filamentous organisms (fungi, actinobacteria) are characterized by the length (L) of mycelium and the number of hyphal tips, n , initiated by hyphal branching. Note that n is equivalent to N when branching is equivalent to cell division.

The relationship between x and N or L is not straightforward because the mass of individual cells and width of hyphae vary among different organisms and even for the same species depending on growth conditions. Generally, a plentiful supply of nutrients fosters the formation of bigger cells or wider filaments, whereas starvation causes cells to shrink. To estimate microbial biomass from direct microscopic counting, one should generate several arrays of data: the number $N = [N_1, N_2, \dots, N_n]$ of cells with specifying biovolume $V = [V_1, V_2, \dots, V_n]$, where N_i is the number of cells in the i th volume class V_i and n is the number of such classes, the volumes being calculated from the geometry of all detected cell shapes (spherical, ellipsoid, rod-shaped, irregular). For filamentous microorganisms, one should have arrays of hyphal length $L = [L_1, L_2, \dots, L_m]$ and corresponding hyphal diameters $D = [D_1, D_2, \dots, D_m]$. The total biomass, x , is calculated as a sum of unicellular and mycelial soil microorganisms:

$$x = V \times N \times (1 - \theta) + \pi D^2 \times L \times (1 - \theta) \quad (1)$$

where θ is the cellular water content.

Advanced microscopy (epifluorescence, electron- and atomic force microscopy) and computer-aided image-analysis techniques make cell counting and sizing more accurate and less tedious. Additionally, Fluorescence In Situ Hybridization (FISH) allows targeting of specific taxonomic or functional groups based on the use of RNA-oligonucleotide probes. The total microbial biomass, x , can be estimated by extracting from environmental samples one of the following unique cell constituents: (a) DNA, (b) ATP, (c) membrane phospholipids, (d) microbial cell wall components (chitin for fungi and muramic acid for bacteria), or (e) the sum of labile or extractable C released after natural sample fumigation with chloroform or drying-rewetting (Blagodatskiy et al., 1987).

Selection of either biomass x or cell number N or mycelium length L depends on the specific objectives of the study. Biomass, x , has the obvious advantage in studies of carbon and nutrient cycling, while cell number, N , is preferred in population studies when, say, mutation or plasmid transfer is crucial for the understanding of the studied natural processes.

A large microbial biomass does not necessarily imply a fast growth rate or vice versa, because low abundance could be caused by intensive *elimination* (e.g., washout, chemotactic movement, grazing, and lysis by parasites, see discussion below) of growing cells while high abundance could be due to *migration* of cells from other habitats or accumulation of dormant cells. In the absence of

Table 1 The major dynamic variables used in deterministic kinetic models.

Variable	Symbol	Dimension (examples)
Concentration of cell biomass	x	$\mu\text{g cell mass (g soil)}^{-1}$, $\text{g cell mass m}^{-2}$
Cell number	N	$\text{count (g soil)}^{-1}$
Mycelium length	L	m (g soil)^{-1}
Hyphal tips number	n	$\text{count (g soil)}^{-1}$
Concentration of limiting substrate	s	mg (g soil)^{-1} , $\text{mmol (g soil)}^{-1}$
Concentration of products (CO ₂ , CH ₄ ,...)	p	mg (g soil)^{-1} , $\text{mmol (g soil)}^{-1}$

elimination, the microbial growth rate can be found by recording one of the used quantities (x , N , L , ATP content, etc.) over time t . The *absolute growth rate* is calculated as a simple slope of cell mass x versus time:

$$\frac{dx}{dt} \approx \frac{\Delta x}{\Delta t} = \frac{x_2 - x_1}{t_2 - t_1} \quad (2)$$

This slope normalized to the biomass represents the *specific growth rate* (SGR) of microorganisms, μ :

$$\mu = \frac{1}{x} \frac{dx}{dt} \approx \frac{1}{\bar{x}} \frac{\Delta x}{\Delta t} = \frac{1}{\bar{x}} \times \frac{x_2 - x_1}{t_2 - t_1}, \text{ where } \bar{x} = \frac{x_2 + x_1}{2} \quad (3)$$

In the case of explicit or hidden elimination of microorganisms accompanied microbial growth, the formulae above are no longer valid and the calculated from Eq. (3) SGR should be referred to as the *apparent growth rate*.

The concentration of limiting substrate

Normally we cannot assess all potentially available nutrients and consequently we focus on one or few individual compounds or classes of molecules representing the limiting nutrient substrate, i.e., such nutrient(s) that control the growth and activity of microorganisms in question. Typically, the growth of chemoorganotrophic microorganisms in soil and subsoil is limited by available organic compounds, while photosynthetic microbiota is limited either by light or sources of P, N, or Fe.

The absolute rate of substrate consumption by microorganisms is measured as a time derivative ($-ds/dt$) of the substrate depletion curve, $s(t)$, while the *specific rate of substrate consumption*, q is calculated by normalizing the absolute rate to the instant cell mass concentration, x :

$$q = -\frac{1}{x} \frac{ds}{dt} = Q \frac{s}{K_s + s} \quad (4)$$

Eq. (4) is similar to the Michaelis–Menten equation established in enzymology with parameters Q standing for maximum uptake rate and saturation constant K_s numerically equal to such substrate concentration at which uptake rate $q = 0.5Q$. SGR μ is related to q by a mass balance Eq. (6) (see below) and therefore can be derived from (4) as follows:

$$\mu = Yq = \mu_m \frac{s}{K_s + s} \quad (5)$$

where $\mu_m = YQ$ is the maximal SGR. This expression is known as the Monod equation. Note that both Eqs. (4) and (5) are empirical and should be applied with caution. The substrate uptake rate not only depends on its concentration but also on other environmental factors (temperature, tonicity, pH, etc.) and the physiological state of cells (macromolecular cell composition, proteome expression profile, and other cell qualities). Even more complex is the intracellular machinery responsible for growth (see Section “Models of intermediate complexity”). Therefore Eqs. (4) and (5) should be applied only to relatively simple experiments with s as the only variable and under conditions in which studied cells preserve the same physiological state at all tested values of s .

There are two groups of microbial nutrient substrates: (1) *catabolic substrates* which are sources of energy, and (2) *anabolic or conserved substrates* which are sources of biogenic elements forming cellular material. Examples of catabolic substrates are H_2 for hydrogen-oxidizing microorganisms, NH_4^+ and NO_2^- for nitrifying bacteria, S^0 for sulfur-oxidizing bacteria, and oxidizable or fermentable organic substances for diverse heterotrophic species. Their consumption is accompanied by oxidation and dissipation of chemical substances into waste products, which are no longer reusable as an energy source (H_2O , NO_3^- , SO_4^{2-} , CO_2 , etc.). The anabolic substrates are consumed and incorporated into de novo synthesized cell components, being conserved in biomass; that is why they are referred to as the *conserved substrates*. Contrary to catabolic substrates, the conserved substrates can be reabsorbed after excretion or cell lysis. They include nearly all non-carbon sources of biogenic elements (N, P, K, Mg, Fe, and trace elements), CO_2 for autotrophs, as well as amino acids and growth factors for respective auxotrophic species.

Growth yield

Two variables, cell mass x and substrate concentration s , are linked via stoichiometric parameter *growth yield*:

$$Y = -\frac{dx}{ds} \approx -\frac{\Delta x}{\Delta s} \quad (6)$$

where Δx is the increase in microbial biomass consequent on utilization of the amount Δs of the substrate. Dividing both parts of (6) by xdt gives the relationship between SGR and substrate production:

$$Y = -\frac{dx}{ds} = \frac{1}{x} \frac{dx}{dt} : \frac{1}{x} \frac{ds}{dt} = -\frac{\mu}{q} \quad (7)$$

The reason for Y variation is different for catabolic and anabolic substrates.

Maintenance energy

In the case of energy sources, some fraction of the total substrate flux is diverted from growth per se to meet so-called *maintenance requirements*. Specific maintenance functions are resynthesis of self-degrading macromolecules (proteins, nucleic acids, cell wall), cell motility, osmoregulation, leakage, proofreading, ROS detoxification, etc. The energy balance is described by the following equation:

$$\frac{\text{total energy source uptake}}{q} = \frac{\text{consumption for growth}}{\mu/Y^{\max}} + \frac{\text{consumption for maintenance}}{m} \quad (8)$$

where m is the maintenance coefficient and $Y^{\max}$ is the hypothetical maximum yield that would be observed with zero maintenance, $m = 0$.

To account for the maintenance requirements, Eq. (4) should be modified as follows:

$$q = \frac{\mu}{Y^{\max}} + m = Q \frac{s}{K_s + s} \cdot \mu = \mu_m \frac{s - s^*}{K_s + s} \quad (9)$$

where $s^* = Y^{\max}K_s m$ is the threshold substrate concentration below which cells are not expected to sustain growth, if $s < s^*$, then $\mu < 0$.

If m and $Y^{\max}$ are constants, then the plot of q versus μ should be a straight line with y -intercept equal to m . Such a plot is the only possible experimental way of finding the maintenance term m . In practice, a chemostat culture of the tested microorganism is run sequentially at several dilution rates until steady-state ($D = \mu$) is established; at each D steady-state biomass and some catabolic rate (respiration, heat production, uptake of the energy source, etc.) are monitored and q plotted versus $D = \mu$ as illustrated in Fig. 1. Note that the described technique is indirect and time-consuming but unfortunately there are no *direct ways* to estimate the maintenance as a part of cellular energy balance. Numerous maintenance processes (osmoregulation, resynthesis, and repair of damaged polymers, stress-response) are mixed up with the growth-associated processes and could not be accurately translated into the energetic cost. The only manifestation of the maintenance is a decrease of apparent cell yield at low SGR. Under very deep nutrient limitation typical for the soil environment, the maintenance term m can no longer be considered as a constant, so the Eq. (9) must be replaced by more advanced kinetic models, see the Section “Near-zero growth kinetics.”

Energy overflow

Maintenance requirements should be distinguished from other non-productive use of energy called *wasteful respiration* or *energy-spilling reactions*. They are observed under several special circumstances: (i) when growth is energy-sufficient but limited by anabolic substrates (sources of N, P, K, Fe, vitamins, etc.); (ii) over extended lag-phase of batch culture; (iii) in transient experiments when starving cells are brought to rich medium and (iv) when cells are exposed to uncoupling inhibitors. A common feature of all the listed processes is so-called *energy overflow*: the energy flux significantly exceeds the level required for biosynthetic processes constrained by inhibitors or a depleted pool of ribosomes or by other than energy source nutrients. As a result, the diversion of energy consumption from growth is much greater than expected from the basic maintenance functions observed in the energy-limited chemostat culture. Note that the energy overflow has been observed in laboratory experiments; this phenomenon has not been demonstrated in situ.

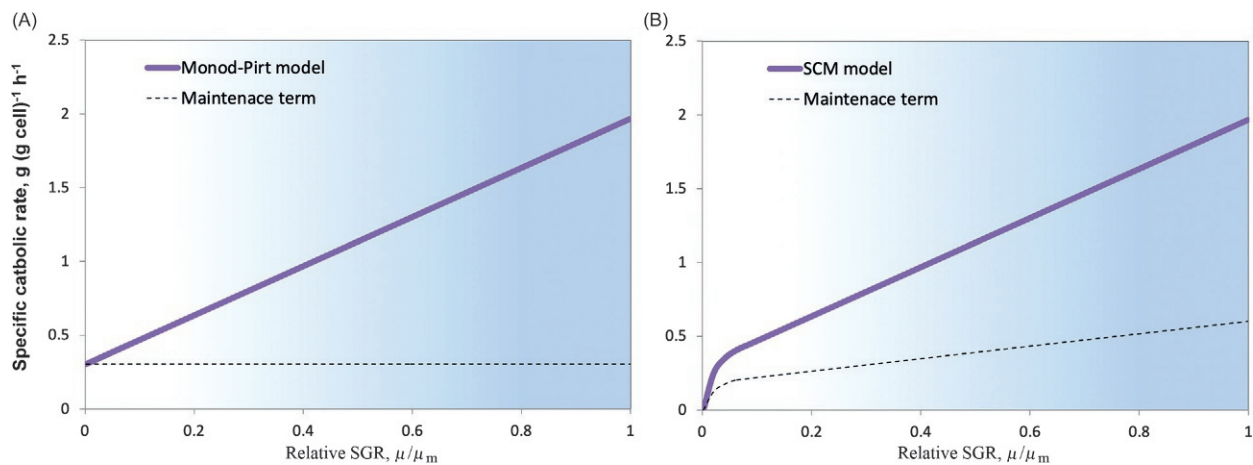


Fig. 1 Two ways of representing the maintenance term. (A) The Monod-Pirt model assumes the constancy of the maintenance term. The tested microorganisms are grown at several dilution rates D in a chemostat limited by the energy source. The specific catabolic rates q_s (respiration, catabolic substrate uptake) are plotted vs $D = \mu$, and fitted to Eq. (9) giving $1/Y^{\max}$ (slope) and the maintenance coefficient, m (intercept and the dotted line). (B) The same plot is simulated by a more realistic SCM model, which allows adaptive self-regulation of the maintenance term. Note a progressive decrease in maintenance with slowing growth. Modified from Panikov NS (1995) *Microbial Growth Kinetics*. The Netherlands: Springer. 378 p.

Cell quota. Variation of yield on anabolic substrates

Cell yield on anabolic substrates varies mainly as a result of alterations in biomass chemical composition as expressed by the parameter σ_s , the intracellular content of a deficient element, or *cell quota*. For example, the variation in N content in bacteria from 5% to 15% gives the σ_s range 0.05–0.15 g N (g cell mass)⁻¹. For most known cases, the quota σ_s increases parallel to growth acceleration, because the higher growth rate requires higher *intracellular content* of proteins and RNA (contain N, P, S) as well as K⁺, Mg²⁺ and vitamins participating in primary metabolic reactions. The yield and cell quota are inversely related to each other, e.g., small N-content ($\sigma_N = 0.05$ g N/g) corresponds to large cell yield ($Y_N = 1/\sigma_N = 20$ g cell per g of N utilized), large N-content in rapidly growing cells can be attained only with a small cell yield ($Y_N = 1/0.15 = 6.67$ g cell/g N) (Fig. 2). The principal difference between Y and inverse σ ($1/\sigma$) is that the first stoichiometric parameter is subjected to instantaneous changes as dependent on s level in various transient shift-up or shift-down processes, while the second parameter is essentially slower required deep reconfiguration in cellular composition.

Products

There could be *intermediate* and *end* metabolic products accompanying microbial growth. The typical intermediates of the glycolytic and the Entner–Doudoroff pathways are:

- glucose $\rightarrow$ 2 pyruvate⁻ + 2H⁺;
- glucose $\rightarrow$ gluconic acid $\rightarrow$ pyruvate;
- pyruvate $\rightarrow$ acetate + CO₂

biological activity due to the wide availability of gas analyzers (GC, IR- and MS). For the case of aerobic decomposition, the measured CO₂ evolution is a significant fraction of the C budget as described by the following mass-balance equation:

The only major end product of aerobic microbial decomposition is CO₂. Under anaerobic conditions, the spectrum of products is much wider and includes several organic acids, alcohols, and ketones, H₂, CO₂, CH₄, and various reduced inorganic compounds (S²⁻, Fe²⁺...). CO₂ evolution rate is probably the most popular parameter to characterize soil

$$\frac{\Delta C-s}{\Delta t} = \frac{\Delta C-x}{\Delta t} + \frac{\Delta C-CO_2}{\Delta t} + \frac{\Delta C-p}{\Delta t} \quad (10)$$

Division by time increment dt gives:

$$\frac{ds}{dt} = \frac{dx}{dt} + \frac{d[CO_2]}{dt} + \frac{dp}{dt} \quad (11)$$

Thus, microbial consumption of C-substrate is the sum of growth, respiration, and product formation, if all four variables are expressed in the same mass-C units, e.g., mg C/g soil. Based on such a C-balance, we can estimate with confidence some variables that are hard to measure directly. For example, cell mass, x in a laboratory soil incubation study can be calculated as the difference between the amount of initially added substrate s_0 and the time-dependent qualities of residual substrate s and products CO₂ + p . However, it would be incorrect to identify the CO₂ evolution rate alone with total soil decomposition, as some degraded C is incorporated into soil organisms or their extracellular products.

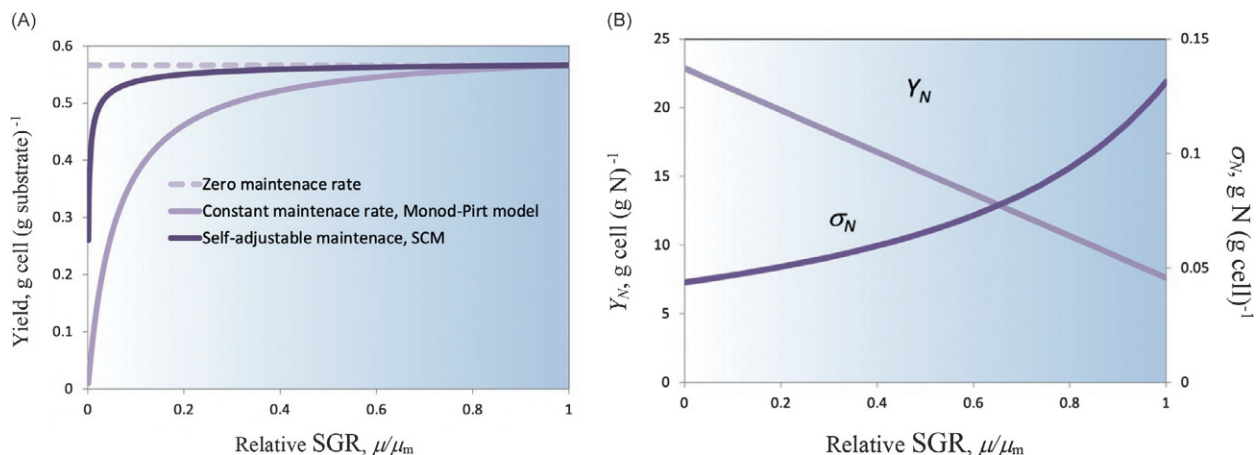


Fig. 2 Variation of the cell mass yield on catabolic (A) and anabolic (B) substrates related to SGR. All curves were calculated using SCM or its simplified versions. Modified from Panikov NS (1995) *Microbial Growth Kinetics*. The Netherlands: Springer, 378 p.

Kinetic models of microbial bioprocesses

Chemical and enzyme kinetics

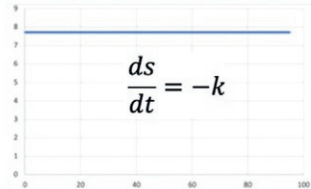
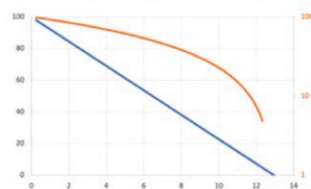
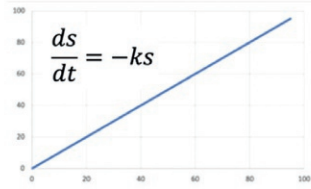
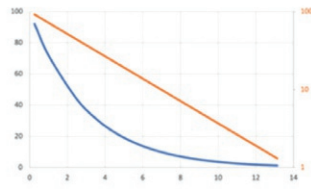
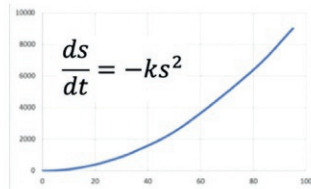
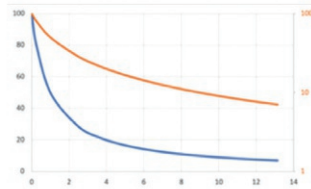
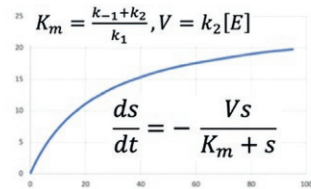
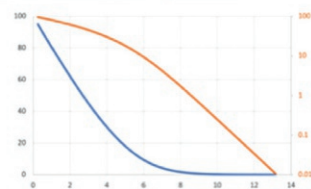
Microbial growth is a complex bioprocess that includes numerous enzymatic and chemical reactions. The basic equations of chemical and enzymatic kinetics are presented in Table 2. There are notions of *molecularity* (the number of molecules that are altered in the reaction) and *kinetic order*, which should not be confused. According to the *law of mass action*, the rate of the chemical reaction is directly proportional to the product of the activities or concentrations of the reactants:

$$\frac{ds}{dt} = -k_n s^n \quad (12)$$

where n is the kinetic order, and k_n is a rate constant. For a simple single-step reaction, the order is generally the same as the molecularity. For a complex reaction involving a sequence of unimolecular and bimolecular steps, the molecularity is not the same as its order. Reactions of molecularity greater than 2 are common, but reactions of order greater than 2 are very rare. For instance, a trimolecular reaction, such as $A + B + C \rightarrow P$, as a rule proceeds through two elementary steps, $A + B \rightarrow X$ and $X + C \rightarrow P$, each of which are of the second or first order. Table 2 gives expressions for three typical cases (zero, first and second order), the left column shows reaction rate as a function of reactant concentration (s), the right column presents results of integration under the initial conditions: $s = s_0$ at time $t = 0$.

A unique feature of all enzymatic reactions is reversible formation of the intermediate enzyme-substrate complex $ES: E + S \xrightleftharpoons[k_{-1}]{k_1} ES \xrightarrow{k_2} E + P$, where k_{-1} and k_2 are the first order and k_1 is the second order rate constants. Under steady-state assumption

Table 2 Basic equations of chemical and enzyme kinetics.

Rate vs. substrate concentration		Reaction progress, $s(t)$		
Zero-order				
		$s = s_0 - kt$	Residual substrate, linear scale, log-scale	
$\frac{ds}{dt} = -k$				
First-order				
		$s = s_0 e^{-kt}$		
$\frac{ds}{dt} = -ks$				
Second order				
		$s = \frac{s_0}{1 + ks_0 t}$		
$\frac{ds}{dt} = -ks^2$				
Mixed order, Michaelis-Menten equation				
		$K_m \ln\left(\frac{s_0}{s}\right) + s_0 - s = Vt$		
$K_m = \frac{k_{-1} + k_2}{k_1}, V = k_2[E]$ $\frac{ds}{dt} = -\frac{Vs}{K_m + s}$				
Substrate concentration, mM		Time, min		

(constancy of ES concentration over time), this mechanism is translated into the Michaelis-Menten equation (Table 2, the bottom line), predicting a mixed reaction order (the first order at small s and zero order at large s), a unique kinetic feature of reactions driven by enzymes. Based on experimental data including the reactant depletion curve $s(t)$, we can distinguish abiotic vs enzymatic reactions and identify the reaction kinetic order.

Exponential growth

It is the simplest growth pattern displayed by all microorganisms (binary dividing bacteria, budding yeasts, fungal mycelium). Exponential growth is observed temporarily under favorable physicochemical conditions before depletion of nutrient substrates and accumulation of self-inhibitory products. The basic assumption is that the current growth rate is proportional to the instant cell mass, x , the quotient μ remaining constant:

$$\frac{dx}{dt} = \mu x \quad (13)$$

The integration of (13) at initial condition, $x = x_0$ at time $t = 0$, gives the exponential equation:

$$x = x_0 e^{\mu t} \text{ or } \ln x = \ln x_0 + \mu t \quad (14)$$

Any metabolic rate, such as intensity of methane production, the evolution of CO_2 , rate of sulfate reduction recorded in microbial culture or nutrient-amended soil samples, is proportional to the growth rate of microorganisms:

$$\begin{array}{ccccc} \text{Metabolic rate} & & \text{Stoichiometric factor} & & \text{Cell growth} \\ \frac{dp}{dt} & = & \frac{1}{Y'} & \times & \frac{dx}{dt} \\ & & & & = \mu \frac{x}{Y'} = \mu \frac{x_0}{Y'} e^{\mu t} = A e^{\mu t} \end{array} \quad (15)$$

where p is the concentration of metabolic product expressed for simplicity in the same carbon units as biomass, mg C per g sample, and Y' is the cell mass yield per mass unit of the metabolic product. Integration of (15) under initial condition $p = 0$ at $t = 0$ gives:

$$p = \frac{x_0}{Y'} (e^{\mu t} - 1) = B(e^{\mu t} - 1) \quad (16)$$

Comparing (14) and (16) we see that x and p time courses are not identical, although both imply exponential cell growth, the $p(t)$ curve deviates from the exponential law, especially at small t . In particular, the “ $\ln x - t$ ” plot is linear, but the plot of $\ln p$ vs t is not. As t increases, the difference progressively decreases. However, the rate of product formation, e.g., soil respiration measured as $\text{mg CO}_2\text{-C/h/g soil}$, follows the exponential pattern (Eq. 14).

Logistic growth

In contrast to the exponential growth equation described above, the logistic equation established by Verhulst in 1838 sets an upper limit for microbial growth by assuming that SGR is dependent on the microbial density, x :

$$\frac{dx}{dt} = \mu(x)x = (\mu - a)x \left(1 - \frac{x}{x_m}\right) = rx \left(1 - \frac{x}{K}\right) \quad (17)$$

where x_m is the upper limits of cell mass. In ecological literature, x_m is called the *carrying capacity* of the ecosystem, K , and the difference between true growth rate, μ , and mortality rate, a , is presented as apparent growth rate, r . This equation mimics the S-shaped growth curve frequently observed in nature. However, it represents nothing more than an empirical relationship ignoring real ecological mechanisms and interactions; note that parameters of logistic equation r and K cannot be predicted *before* actual observation.

Monod model

This model was developed by a famous French scientist in 1941 but it is still very popular due to its elegant simplicity. There are two basic premises: (a) the SGR is dependent on the concentration of limiting substrate (Eq. 4); and (b) biomass formation is linked to substrate uptake by mass-conservation condition (Eq. 6):

$$\begin{array}{l} \frac{dx}{dt} = \mu(s)x = \mu_m \frac{s}{K_s + s} x \\ \frac{ds}{dt} = -\frac{1}{Y} \frac{dx}{dt} \end{array} \quad (18)$$

Equation set (18) contains three parameters: yield Y , maximal SGR μ_m , and saturation constant K_s , which can be thought of as ‘ID’ for individual organisms and used to predict their growth dynamics. Remarkably, this model was used to develop a chemostat theory before actual experiments with continuous culture had been undertaken, one of the very rare events in the history of mathematical biology! For this purpose, the equation set (18) was modified as follows:

$$\begin{aligned}\frac{dx}{dt} &= (\mu - D)x, \mu = \mu_m \frac{s}{K_s + s} \\ \frac{ds}{dt} &= D(s_r - s) - \frac{1}{Y}\mu x, D = \frac{F}{V},\end{aligned}\quad (19)$$

where s_r is the concentration of limiting substrate in the fed fresh medium delivered with pump from a reservoir, F is the pumping rate (cm^3/h), V is culture volume (cm^3), and D is the dilution rate defined as the ratio $D=F/V$, h^{-1} . The model predicted several counter-intuitive features of the chemostat, e.g., that SGR can be set up by an experimentalist by changing the medium flow at any values between 0 and μ_m (before specific growth was believed to be a constant μ_m) and that growth rates depend on s (the *residual* nutrient concentration in bioreactor) rather than on s_r (the nutrient concentration in the feed medium).

However, the Monod model fails to explain many dynamic phenomena observed experimentally, namely: lag-phase, death of starving cells, product formation, and any kind of adaptive changes in the microbial population, such as induction-repression of enzymes, yield variation, changes in the cell RNA content, etc. These gaps were intended to be filled by the structured models.

Cell quota model

The model is based on the empirical observation that SGR is proportional to the intracellular content of a deficient element σ , such as N, P, Fe, and other conserved nutrient substrates (Droop, 1973). As applied to the chemostat culture, it contains the same two variables (x and s) as Eq. (19), but a different expression for specific growth rate:

$$\begin{aligned}\frac{dx}{dt} &= \mu x - Dx \\ \frac{ds}{dt} &= D(s_r - s) - \mu x \sigma \\ \mu &= \frac{\mu_m \sigma_m (\sigma - \sigma_0)}{\sigma (\sigma_m - \sigma_0)}\end{aligned}\quad (20)$$

It was the first attempt to express the relationship between SGR and cell quota, the variable characterizing the chemical composition of the cell biomass. However, it is wrong to consider cell quota as an independent variable. As a matter of fact, σ itself is a function of the environmental factors, such as the concentration of respective sources of biogenic elements in situ.

Structured models

These models explicitly describe variations in cell composition (Fredrickson, 1976), they include mass balance equations not only for *external* substrates and products concentrations but also several *intracellular* components, $C_1, C_2, \dots, C_n$, such as DNA, RNA, proteins, cell wall constituents, reserved polysaccharides, ATP-pool, etc. For each variable C_i a differential equation is written which accounts for all sources, r_+ , and sinks, r_- , as well as its dilution due to cell mass expansion (growth),

$$\frac{dC_i}{dt} = r_+(s, C_1, \dots, C_n) - r_-(s, C_1, \dots, C_n) - \mu C_i \quad (21)$$

The earliest structured models contained just 2–3 highly aggregated internal variables, e.g., the total content of cellular proteins, RNA, and DNA. More advanced models covered up to 50 real cellular constituents and the lumped variables such as pools of amino acids and nucleotides, rNTP and dNTP, cell wall precursors, stable (rRNA), and short-lived RNAs (mRNA and tRNA), regulatory metabolite ppGpp, glycogen, peptidoglycan. These structured models produced modest results and never gained a wide recognition blamed by biochemists for oversimplification of real microbial metabolism and by modelers for extreme complexity. Two of their main disadvantages were (i) a subjective selection of internal variables, and (ii) a problem of experimental verification of the aggregated data.

Genome-scale models (GEM)

The ambitious goal of GEM is the accurate prediction of the phenotype from the respective genotype under specified environmental conditions. High throughput genome sequencing makes it possible to screen nearly all cultivable microorganisms of whatsoever medical, industrial, or environmental significance, including soil isolates. Presently, NCBI lists more than 300,000 completed whole-genome projects including ~7500 sequenced fungal species and more than 31,000 prokaryotic (archaeal and bacterial) species. Soil microbial communities are still underrepresented in culture collections because most soil inhabitants resist isolation. However, the metagenomic approach (shotgun sequencing of the DNA directly extracted from the soil) potentially allows in silico reconstruction of phenotypes for any microorganisms, well-known or mysterious uncultivable species. We are still far away from this ultimate goal, but in the case of success, GEMs would be able to reconstruct their growth dynamics under a specified state of the environment.

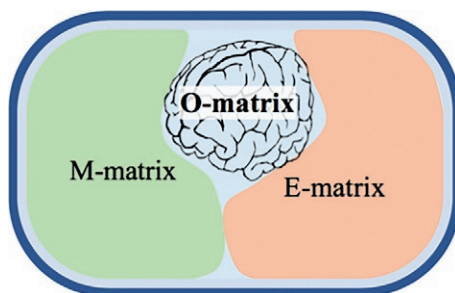


Fig. 3 Simplified view on the bacterial genome structure through the prism of GEM-modeling.

The progress of GEM development is impressive. The first GEMs were constructed for the simplest parasitic bacteria and *E. coli* in 1995–1997, and then the number of GEM publications has [risen exponentially](#). GEM is the metabolic network reconstruction. Contrary to the structured models above, GEM uses real individual intracellular metabolites and metabolic reactions extracted from genomic data.

GEMs cover only the part of the genome, mostly up to 35%. From a modeling perspective, there are four distinct parts of the genome (Fig. 3): (i) junk material and sequencing errors needed to be removed (blue background), (ii) the **M-matrix** (M for metabolism) standing for genes encoding enzymes catalyzing the entire intermediary metabolism, which is about one-third in prokaryotes and smaller fraction in eukaryotes; (iii) the **E-matrix** (E for expression) accommodating genes encoding enzymes and RNAs responsible for proteins synthesis, transcription, and translation; (iv) finally, the **O-matrix** (O for operon) containing the transcriptional regulatory network, a part of global regulatory machinery orchestrating cellular functions via transcription factors, TFs. Respective GEMs are called **M**-, **E**- and **O**-models. Reconstruction of metabolic networks (M-models) is now a well-established process. The E-models have recently been started by using a similar approach and are produced as combined **ME**-models. Both models convert genomic data into stoichiometric matrices of intermediate (M-matrix) and biosynthetic reactions (E-matrix) that are made interactive. Unfortunately, there is no way to predict in theory or based on sequence similarity the functionality of the O-matrix, ‘the brain’ of the metabolic network. Instead, GEMs apply the optimality principle, assuming that natural regulation of metabolic network is performed in such a way that maximizes the growth rate of microorganisms under given growth conditions. There could be other objective functions, e.g., the minimizing cost of protein biosynthesis, minimizing the ATP-pool size or the energy consumption.

The most popular is the M-models called the Flux Balance Analysis (FBA). It seeks only the steady-state solutions for metabolic fluxes and uses constraints to define a closed solution space for the flux vectors by setting up the low and high boundaries for each flux. Additional constraints are coming from the reaction thermodynamics as well as from the metabolomic, transcriptomic, or proteomic data.

The starting point of FBA is the annotated genome sequence. The product of each gene is annotated by homology searches to identify all the metabolic enzymes. It happens frequently that some enzymes in a known pathway are missing, so other reactions are added to close the mass balance. In addition to enzymatic reactions, all transport mechanisms must be accounted for, such as diffusion, active or facilitated transport through membrane, and secretion of exometabolites. In the next step, a dynamic mass balance is derived for all the metabolites using a matrix notation:

$$\frac{d\mathbf{X}}{dt} = \mathbf{S} \times \mathbf{v} \quad (22)$$

where $\mathbf{X}$ is the vector of cellular metabolites, $\mathbf{S}$ is the stoichiometric matrix and $\mathbf{v}$ is the matrix of the fluxes.

The main FBA advantage is its computational efficiency: the COBRA toolbox in Matlab or COBRApy in Python installed on a regular laptop executes full-size M-models (thousands of metabolic reactions) in a few seconds. The FBA is not designed for dynamic simulations and cannot present growth as a time series of x , s , and p to demonstrate the agreement. Instead, it produces the metabolic map with fluxes for each metabolic reaction. It correctly reproduces specific rates of unlimited growth (observed in the batch culture exponential phase), metabolic reconfigurations after switching from one C-source to another, from aerobic to anoxic growth conditions; it also accurately predicts genes essentiality and the impact of the gene knockdown on metabolic flows including the exometabolic products that is especially valuable for bioengineering.

The dynamic FBA (dFBA) applies the *quasi-steady-state* approximation. The metabolic fluxes can be differentiated into slow and fast categories as dependent on reaction volume. The exchange reactions between cell and environment (nutrient uptake and product release) take place in the volume range 10^0 – 10^4 cm³. The *intracellular* metabolic reactions occur in a much smaller volume of a single cell, 10^{-8} – 10^{-9} cm³. The volumetric difference 10^8 – 10^{13} times implies that intracellular reactions can be safely assumed to be at a steady-state, and therefore the flux distribution can be resolved by the FBA. Few remaining slow exchange variables are to be numerically solved as a set of ordinary differential equations (ODEs). Numerous dFBA models were developed to simulate growth dynamics of the model (*Saccharomyces cerevisiae* and *E. coli*) and industrially important organisms, some of which (e.g., *Pseudomonas putida*) are used for soil remediation.

The ME-models and the whole-cell models aimed to simulate a self-regulatory behavior of microorganisms under fluctuating environment have been constructed for smallest parasitic prokaryotes as well as for the two best studied organisms, *E. coli* and baker yeasts. Unfortunately, they remain too cumbersome for execution and too modest in discovery of the emergent systems qualities of the cell (Panikov, 2021). Indisputably, GEMs are the most promising tools in biokinetic studies, but their practical implementation to environmental sciences is still not possible.

Models of intermediate complexity

Several growth models occupy an intermediate position between oversimplified Monod-type equations and gigantic genome-scale metabolic reconstructions operating with thousands of variables. In real life, we face a necessity to have a ‘fundamentally correct’ and compact representation of microbial dynamics that can be instantly solved with regular computers and standard software. A perfect equivalent would be the [ideal gas law](#) in physical chemistry, the simplified but dependable model reproduces correctly the behavior of many gases under many conditions. It would be wrong to simplify GEM by selecting a limited subset of genes (see above a failure of the structured models), a more promising option would be to rely on compact models able to reproduce the behavior of *an ideal cell* in a real variable environment. A loose analogy with ideal gas (having zero intermolecular attractions) allows us to picture such a cell as free from numerous specific details and having only two essential functional qualities – *to reproduce itself and adapt to the fluctuating environment by conditional expression of subcellular structures*. From the kinetic perspective, both qualities are unique to cells and absent in enzymes, which do not proliferate or change their composition. There is no problem with reproduction, it is manifested kinetically as the exponential x increases with the constant SGR under permanently optimal conditions (Eqs. 13 and 14). Under fluctuating conditions, we typically observe increases or decreases of SGR parallel to changes in the macromolecular composition of cells (MMCC), including proteomic and transcriptomic profiles, the composition of lipids, metabolic intermediates, and other results of the *omics* molecular inventories. We apply here the term ‘models of intermediate complexity’ to those non-GEM growth models that make a very important step forwards to the genome-scale level by attempting to relate the macroscopic growth kinetics (SGR, yield, respiration, maintenance, etc.) with MMCC and variable environmental conditions. There are two types of such models: (i) the cybernetic model and (ii) the synthetic chemostat model (SCM). Both reached some success in the simulation of the self-regulated behavior of cells in response to environmental changes.

The cybernetic models (Kompala et al., 1986) were introduced initially as simple structured models, the word ‘cybernetic’ implying a parallel between a living cell and electronic gadgets, both dealing with circular causalities and feedback. Two cybernetic variables were incorporated as Boolean ON/OFF switches into the conventional Monod-style growth models applied to situations when the cell must take a decision, e.g., select the preferential substrate in the mixture of glucose and lactose:

$$\begin{aligned}\frac{dx}{dt} &= x \sum_{i=1}^2 \mu_i v_i - Dx; v_i = \frac{\mu_i}{\max(\mu_i)} \\ \frac{ds_i}{dt} &= D(s_{ri} - s_i) - x \left(\frac{\mu_i v_i}{Y_i} + m_i v_{mi} \right) \\ \frac{de_i}{dt} &= q_i u_i - \mu e_i; u_i = \frac{\mu_i}{\mu}; \mu = \sum_{i=1}^2 \mu_{mi} e_i \frac{s_i}{K_{si} + s_i}\end{aligned}\quad (23)$$

Here the elements of MMCC are two transporters e_1 and e_2 specializing on uptake of glucose and lactose respectively. The cybernetic variables regulate transcription (u) and uptake rates (v) as dependent on the difference between the current values of μ_1 and μ_2 , the SGR on each of the competing substrates: $v_1 = \begin{cases} 1, & \text{if } \mu_2 < \mu_1 \\ 0, & \text{if } \mu_2 \geq \mu_1 \end{cases}$; $v_2 = \begin{cases} 0, & \text{if } \mu_2 < \mu_1 \\ 1, & \text{if } \mu_2 \geq \mu_1 \end{cases}$.

Not surprisingly, the model successfully predicted a correct choice of glucose as a preferential substrate with the following diauxic pattern, unfortunately only in the simplest setup of the canonic exponential phase with constant SGR and MMCC. The modern cybernetic models use similar v and u cybernetic variables but are scaled up to the genomic level (Kim et al., 2008; Song and Ramkrishna, 2011). Their value as a self-dependent subtype of GEM is questionable because of the artificial nature of these operational variables. So far, the cybernetic models have not been applied to complex microbial biodynamics, such as non-steady state and unbalanced growth in bioreactors or in situ.

The SCM (Panikov, 1995) has been built upon all the chemostat models listed above, including the Monod model and its modifications, cell quota and structured models (that explains the epithet ‘synthetic’). The basic SCM contains three differential equations for the limiting substrate, s , cell biomass, x , and the r -variable compactly characterizing the MMCC (the ‘quality’ of cells):

$$\frac{ds}{dt} = D(s_r - s) - q_s x, q_s = rQ \frac{s}{K_s + s} \quad (24)$$

$$\frac{dx}{dt} = \mu x - Dx, \mu = Y(q_s - m) \quad (25)$$

$$\frac{dr}{dt} = \mu \left(\frac{s}{K_r + s} - r \right) \quad (26)$$

The r -variable (Eq. 26) has been introduced as follows. There are at least two functional groups of cellular constituents: the content of P-components correlates with SGR because they participate in the Primary metabolic reactions (ribosomes, key biosynthetic

enzymes). Other conditionally-expressed cell constituents are associated with the secondary metabolism responsible for cell survival (U-components) under stresses, and their content negatively correlates with SGR. Examples of U-components are enzymes involved in drug resistance and synthesis of antibiotics, protective pigments, reserved polymers, high-affinity transporters, etc. The first and second groups are combined respectively into vectors $\mathbf{P}$ and $\mathbf{U}$ satisfying conservation condition $\mathbf{P} + \mathbf{U} = 1.0$ if each P_i and U_i are normalized to cell dry mass (the contribution of small molecules is neglected for clarity). The constitutive (unchangeable) parts are denoted as $P_i^{\min}$ and $U_i^{\min}$ while $P_i^{\max}$ and $U_i^{\max}$ are the upper limits respectively. The SCM makes three assumptions: (1) all P-components change in response to environmental signal synchronously and in the same direction (otherwise sustainable growth would be impossible); (2) the main environmental signal controlling the vectors $\mathbf{P}$ and $\mathbf{U}$ expression is the concentration of limiting substrate, s that is sensed by cells; (3) the expression degree is approximated by a hyperbolic function of s :

$$\mathbf{P} + \mathbf{U} = \begin{bmatrix} P_1^{\min} \\ \dots \\ P_n^{\min} \end{bmatrix} + \begin{bmatrix} U_1^{\min} \\ \dots \\ U_m^{\min} \end{bmatrix} + r \begin{bmatrix} P_1^{\max} - P_1^{\min} \\ \dots \\ P_n^{\max} - P_n^{\min} \end{bmatrix} + (1-r) \begin{bmatrix} U_1^{\max} - U_1^{\min} \\ \dots \\ U_m^{\max} - U_m^{\min} \end{bmatrix}; \quad r = \frac{\bar{s}}{K_r + \bar{s}} \quad (27)$$

Constitutive part Changeable P part Changeable U part Steady state r

where r has been introduced as a common quotient to all elements of the vector $\mathbf{P}$. It is a scalar function of s , which varies between 0 (long-term starvation) and 1.0 (established unlimited growth) and expresses the quality of cells: the higher is r , the higher is the content of P-components. These assumptions and expression 27 agree with available chemostat data on the growth rate driven changes in the composition of cells (Fig. 4A) and the observed patterns in gene expression (Panikov, 2021) as well as with the well-established knowledge on the master role of transcription factors orchestrating simultaneous transcription of the long list (100–1000) genes responsible for stress resistance (Hengge-Aronis, 2002; Gasch et al., 2000). Thus, gene expression is likely to vary in a well-coordinated manner and SCM accurately mimics this trend although unable to specify individual conditionally expressed macromolecules. Nonetheless, the combining kinetic terms with MMCC improves significantly the predictive power of the SCM. The model can reproduce and predict before the actual experiment, the time course of batch culture (Fig. 4B) starting from the initial lag phase to the following growth phases leading eventually to starvation death of cells in depleted culture. To run the simulation,

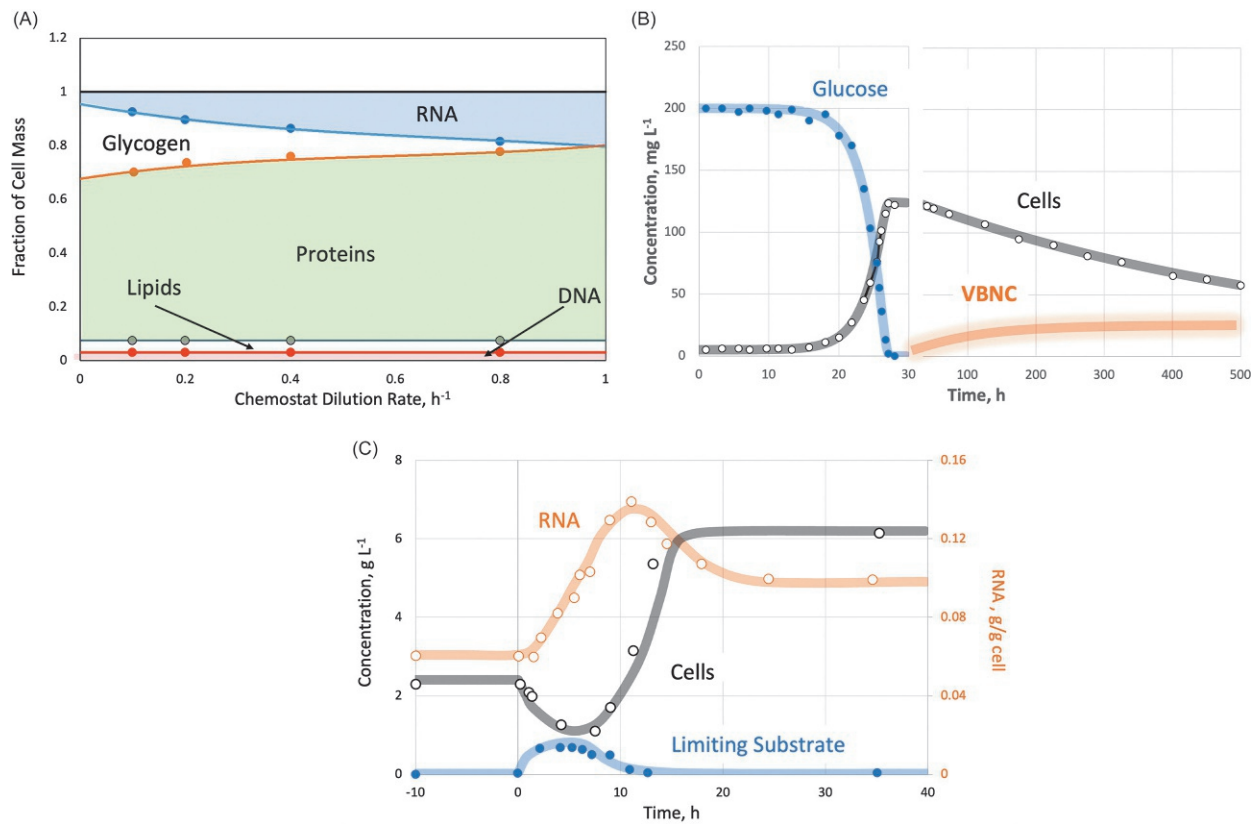


Fig. 4 Simulation success of the SCM model. (A) Cell composition of *A. aerogenes* as dependent on dilution rate in a NH_4^+ -limited chemostat culture (Tempest et al., 1967). (B) Batch growth of *Schwanniomyces vanrijiae* on the minimal glucose–mineral medium. VBNC stands for the viable-but-not-cultivable cells (Panikov, 2019). (C) The shift-up experiment in a chemostat culture of *Aerobacter aerogenes* limited by glycerol; at time zero, the dilution rate was changed from 0.004 to 0.24 h^{-1} (Tempest et al., 1967). With permission from Panikov NS (2019) 1.18-Microbial growth dynamics, In: *Comprehensive Biotechnology* (3rd edn.) Moo-Young M. Oxford: Pergamon. p. 231–273. Copyright (2019) Elsevier.

one needs to know the initial conditions (x_o , s_o , r_o) and the basic growth parameters as a set of constant coefficients specific to every cultivable microorganism. Further, the SCM successfully simulates the steady-state and transient growth in various continuous bioprocesses (chemostat, turbidostat, pH-stat, retentostat and dialysis culture), including the most challenging transient shift-up and shift-down regimes (Fig. 4C). This model will be mostly used below for analysis of the effect of environmental factors on microbial growth, as well as in the linked chapter, where it is devoted to soil applications.

Growth dependence on environmental factors

The factors affecting growing cells in culture and in situ can be physical (temperature, tonicity, moisture content, texture, conductivity, gas exchange rates); chemical (pH, Eh, ionic strength, surface chemistry of particulate matter, medium or soil solution composition); and biological, e.g., the positive, or negative interactions with other microorganisms or a host.

Hierarchy of factors, the primary significance of the nutrient resources

In the ecological literature, all factors are viewed as important, their priority is dependent on local circumstances. For example, the radiation regime and temperature are supposed to be the key environmental factors in polar ecosystems, water availability in deserts, metals toxicity in areas affected by industrial pollution, salinity in salt marsh soil, and so on. However, from the biokinetic perspective, the *nutritional* factor stands alone as a *primary* factor under any circumstances, while the rest of the *modifying* factors (temperature, pH, Eh, pressure, precipitation, moisture content, and many others) can accelerate, slow or stop a microbial process but can never produce new cell mass. The state variables representing the nutritional factor, such as s , the limiting substrate concentration, must be present in any kinetic model of microbial growth. The variables characterizing the modifying factors are incorporated into specialized kinetic models focusing on a particular problem. The following three arguments justify the stated difference between the primary and modifying factors. *First*, the nutrients participate in growth stoichiometry, they are reactants initiating the entire cellular metabolic network eventually producing new cells and products. The modifying factors do not participate in growth stoichiometry or mass-balance of microbial growth. *Second*, contrary to most modifying factors that characterize the *state* of the environment, nutrients represent the (actually or potentially) limited *consumable resources* shared between various competitors: coexisting species in communities, subpopulations of the same species (e.g., the spontaneous mutant vs wild parental cells), alternative pathways of the cellular metabolic network. It makes the nutritional factor especially influential in the simulations of complex systems aimed at the selection of the most probable scenario from numerous alternative options. *Third*, the temperature and other modifying environmental factors are important; however, simulation of their action can be meaningful only in combination with the nutritional factors. The opposite is not true: the effect of nutrients can be fruitfully explored separately from the modifying factors, provided they remain constant. The reason is that microbial response to any environmental factor heavily depends on the metabolic status of cells, gene expression profile, and proteome – all internal characteristics being shaped by the nutritional status. Specifically, the stress response to heat, cold, or drugs, is mediated by chaperones and other reparation biomolecules that require energy and biosynthetic precursors. Interestingly, the maximal stress resistance is observed not under plentiful nutrient supply but rather under deep substrate limitations (Panikov et al., 2015), a phenomenon simulated and explained by SCM.

Below we demonstrate how two archetype modifying factors (pH and temperature) interact with the nutritional factor in the characterization of growth kinetics. The final example is a special case of the exceptionally deep nutrient limitation neglected in laboratory studies and widely occurring in nature responsible for so-called, near-zero microbial growth (NZG).

Effects of pH

In enzymology, the effect of pH is explained by the ionization of several functional groups within the active center. The first assumption is that an enzyme is the dibasic acid H_2E , which undergoes dissociation:



The second assumption is that only the HE^- form is active, it cuts the reaction rate by the pH-dependent factor $\theta = [HE^-]/([H_2E] + [HE^-] + [E^{2-}])$, where all three concentrations can be expressed via $[H^+]$, K_1 and K_2 . The final assumption is that all cellular enzymes have similar pH profiles, then the proposed mechanism can be applied to the SGR:

$$\mu_{pH} = \mu_m \theta = \mu_m \frac{K_1 [H^+]}{[H^+]^2 + K_1 ([H^+] + K_2)} \quad (29)$$

Eq. (29) does not account other than pH restriction factors. So, in the case of nutrient limitation and non-optimal temperature, we should add respectively Eqs. (18), (19), (24)–(26) and (36) (see below) in a multiplicative way. Eq. (29) reproduces the experimentally observed symmetrical bell-shaped curve with a maximum at $(pK_1 + pK_2)/2$ (Fig. 5). There could be other

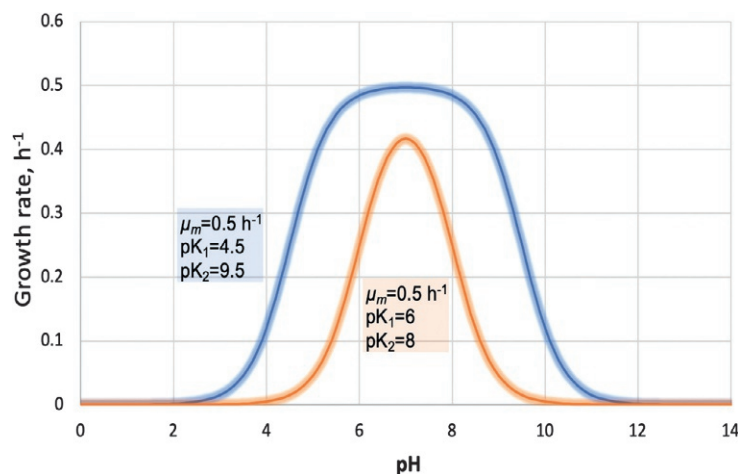


Fig. 5 Effect of pH on microbial growth, Eq. (28).

mechanisms involved at the cellular level. However, for practical purposes, Eq. (29) works reasonably well in finding the optimal pH, especially useful when data points are rare and scattered.

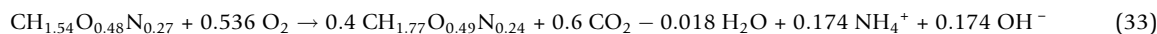
Contrary to the most enzymatic reactions, microbial growth is accompanied by dramatic pH changes. A widely spread misconception is that growth is predetermined to release acidic metabolites (acetic, pyruvic, citric acids) produced in glycolysis and TCA cycles. This idea is not accurate, microbial growth can be associated with either metabolic acidification or alkalization and is dependent on media composition; acidic metabolites are not the only reason for pH drop, it is better to state that growth is accompanied by uptake and release of ionized species which can cause the acid-base disbalance. Very often, the main reason for disbalance is the intensive uptake of the N-source rather than release of organic acids. Let us identify where the problem lies.

The starting point in the understanding and prediction of the acid-base disbalance, is to draft the stoichiometric equation of bacterial growth. Cell biomass is represented as a pseudo-compound with the empirical formula $\text{CH}_h\text{O}_o\text{N}_n$ derived from the average elemental composition (below the *E. coli* data are used) and normalized to the C-atoms. The lower-case indexes h, o, and n indicate the amount (very rarely the integer as in individual chemical compounds) of respective elements H, O, and N per one C atom. Other macro-elements (P, K, S, Ca, Mg, ...) and trace elements are ignored because of their minor effect on the ionic status of a typical microbial culture. The following three stoichiometric equations are just examples. They are complete analogs of well-known chemical equations, each side must represent the same number of atoms, and the same holds true for the total electric charge (condition of neutrality) through conversion of three 'reactants' (O_2 , C- and N-sources) into de novo cell mass, CO_2 , and H_2O with assumed cell yield 0.4 g cell-C per g glucose-C:

Conditions	C-source	Oxygen	N-source	Cells	Carbon dioxide	Water	Acid-base disbalance
Glucose + NH_4Cl :	CH_2O	+ 0.593 O_2	+ 0.096 NH_4^+	$\rightarrow 0.4 \text{ CH}_{1.77}\text{O}_{0.49}\text{N}_{0.24} + 0.600$	CO_2	+ 0.790 H_2O	+ 0.096 H^+ + 0 OH^- (30)
Glucose + NaNO_3 :	CH_2O	+ 0.401 O_2	+ 0.096 NO_3^-	$\rightarrow 0.4 \text{ CH}_{1.77}\text{O}_{0.49}\text{N}_{0.24} + 0.600$	CO_2	+ 0.598 H_2O	+ 0 H^+ + 0.096 OH^- (31)
Glucose + urea:	CH_2O	+ 0.593 O_2	+ 0.048 CON_2H_4	$\rightarrow 0.4 \text{ CH}_{1.77}\text{O}_{0.49}\text{N}_{0.24} + 0.648$	CO_2	+ 0.742 H_2O	+ 0 H^+ + 0 OH^- (32)

To comply with the conservation conditions, these Eqs. (29)–(31) produce either H^+ (NH_4Cl as N-source) or OH^- (KNO_3 as N-source) that disturb the acid-base balance. The uptake of urea, an uncharged N-source, does not change the ionic status of cultural liquid, its only effect is slight stimulation of the CO_2 release.

The next interesting case is the complex protein-rich media of uncertain composition, which is the most popular type of media in laboratory practice. We selected the casein as a standardized commercial product with known elemental composition serving as a combined source of C- and N:



It turns out that casein is the strongest among the tested nutrient sources alkalization agent, nearly two times stronger than nitrate. Other widely used alkalization nutrients are the sodium and potassium salts of organic and amino acids.

The next biokinetic objective is the reconstruction of the pH time course in real batch culture or it could be the reconstruction of the auto-titration dynamics aimed at neutralizing metabolic acidification or alkalization. To fulfill this task, we need to specify not only the growth stoichiometry (Eqs. 30–32) but also the buffering capacity of cultural or environmental liquid, and to reconstruct the growth dynamics, expressed as a set of ODEs for cell biomass, nutrient substrates consumption and products formation. The

parameters required for coupling the growth of microorganisms with their degree of acid-base disbalance can be found again in Eqs. (30)–(32). For example, we can easily convert the cell ‘moles’ of these equations into the conventional dry weight of cell biomass and compare it with the moles of produced acidity or alkalinity. For example, the production of one gram of cells on glucose + NH_4Cl requires the cumulative use of 7.8 mmol of NaOH to neutralize the metabolic acidity. We can also state a stoichiometric yield-like parameter $Y_n = 7.8 \text{ mmol (g cell)}^{-1}$, stating that production of one gram of cell biomass causes metabolic acidification equivalent to release of 7.8 mmol of strong acid. Such data can be fruitfully used for the planning purposes, e.g., to calculate in advance the composition of nutrient medium (amount of energy, C- and N-sources, buffer molarity and respective buffering capacity) required for cultivation without instrumental pH control. Another more creative biotech solution would be to revise the medium composition in such a way that minimizes acid-base disbalance.

Effects of temperature

In chemical kinetics, the dependence of the reaction rate on temperature is explained by the Eyring transition-state theory. It is based on the use of thermodynamics and principles of quantum mechanics. The reaction proceeds through a continuum of energy states and must surpass the state of maximum energy when a transient activated complex is formed. Then, the dependence of the reaction rate constant k on the absolute temperature, T , is expressed as follows:

$$\frac{d \ln k}{dT} = \frac{\Delta H^* + RT}{RT^2} \quad (34)$$

where R is the gas constant and ΔH^* is the enthalpy of activated complex formation. The more popular Arrhenius equation is obtained by integration of Eq. (34) and a simplified substitution $\Delta H^* + RT \approx \Delta H = E_a$:

$$\ln k = \ln A - \frac{E_a}{RT} \quad \text{or} \quad k = Ae^{-E_a/RT} \quad (35)$$

where E_a is the activation energy and A is the integration constant, interpreted as the frequency of molecular collisions.

The Arrhenius equation fits well the biokinetic data under a narrow temperature range from zero to 20–60 °C as dependent on tolerance to the thermal denaturation (minimal in psychrophiles, maximal in thermophiles). Let us assume that denaturation is reversible with the equilibrium constant $K_T = [E']/[E]$, where E and E' stand for the active and denatured molecules respectively. Then, the combination of Eq. (35) with the van Hoff expression for K_T ($-RT \ln K_T = \Delta G^\circ = \Delta H^\circ - T\Delta S^\circ$) results in the following updated Arrhenius equation and its exponential approximation in Celsius temperature:

$$\mu_m = \frac{Ae^{-\frac{E_a}{RT}}}{1 + e^{\frac{\Delta S^\circ}{R} - \frac{\Delta H^\circ}{RT}}} = \frac{Ae^{-\frac{E_a}{RT}}}{1 + e^{B(1 - \frac{C}{T})}} \approx \frac{Ae^{\alpha T_c}}{1 + e^{B(1 - \frac{C}{273+T_c})}} \quad (36)$$

where ΔG° , ΔS° , and ΔH° are the standard Gibbs free energy, enthalpy, and entropy of the denaturation reaction, respectively, aggregated into constants B and C . In this equation, we made substitution of the reaction rate k for the maximum SGR.

The best experimental setup for verification and estimation of kinetic parameters is the exponential phase of the batch culture or turbidostat and pH-auxostat. The chemostat is not appropriate because of SGR dependence on two strong factors, temperature and limiting substrate concentration. In the case of soil incubation studies, soil community should have either continuous supply of nutrients, or experiments must be a short-term to avoid the effect of a hidden change of growth rate. In the example (Fig. 6), we measured the intensity of methane production by stabilized methanogenic community characterized by low SGR, that was continuously supplied by gaseous substrate (H_2 : CO_2).

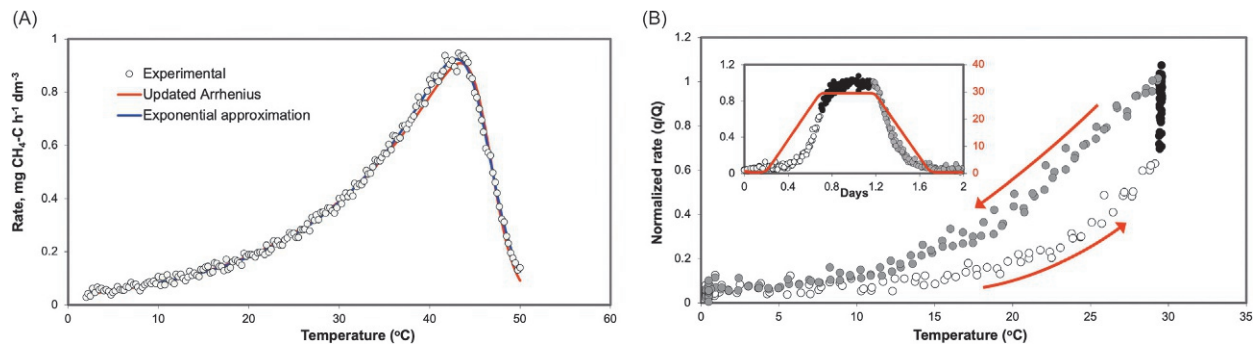


Fig. 6 The effect of temperature on microbial activity in a methanogenic bioreactor (Panikov, unpublished). Experimental conditions: the *Sphagnum* peat core was continuously flushed with gas mixture N_2 : CO_2 : $\text{H}_2 = 90$:5:5 and off-gas mass-spectrometric recording of methane. After stabilizing at 3 °C for 2 days, the incubation temperature was programmed to rise linearly with the rate of 1.0 deg. C h^{-1} from 3 to 50 °C (panel A) or to cycle between 1 and 30 °C as shown by the red line (panel B insert). (A): The 10-min averaged CH_4 production rates plotted versus temperature fitted to Eq. (36). (B) The hysteresis in microbial response to temperature variation, red arrows indicate the temporal progression, and the insert shows the time series of temperature (red curve) and methane formation rate. Note different symbols (unfilled, grey, and black-filled circles) for three parts of the temperature gradient.

Within the biologically meaningful range of temperatures (0–60 °C), the exponential approximation does not deviate from the more accurate Arrhenius equation (Fig. 6A) but is more stable in solving the inverse problem. Contrary to chemical reactions, microorganisms demonstrate significant *hysteresis* in their response to cyclic temperature changes (Fig. 6B). In other words, the effect of temperature depends on how it varied in the past. It is explained by the fact that adaptation to temperature extremes (too cold, too warm) involves conditionally expressed macromolecules, e.g., the heat-shock proteins, desaturating enzymes, and chaperones (Strocchi et al., 2006). The SCM is capable of simulating the process after the update of Eq. (26): the r -variable should be multiple functions of s and T .

The special case of near-zero growth kinetics

Laboratory cultivation mainly occurs under an excess of nutrients with SGR close to the maximum, μ_m . In nature, microbial growth follows diverse patterns, remaining mostly continuous and very slow ($\mu \ll \mu_m$) because of severe nutrient limitations. Even in relatively rare *eutrophic* habitats, such as a plant rhizosphere or soil invertebrate gut, microorganisms divide no more than 1–2 times per day, while in *mesotrophic* and *oligotrophic* habitats (plant-free soil, subsoils, sediments) the doubling time can be months and years. Here we call such a low-intensity process a *near-zero growth* (NZG) and summarize below its specific features. Formally, any growth model, simple or structured gives a prediction of how microorganisms should react to extremely low s . According to the unstructured models, the maintenance term should play a crucial role: if we ignore the maintenance ($m = 0$ as in the original Monod model), the NZG is nothing more than a smooth growth deceleration ($\mu \rightarrow 0$ as $s \rightarrow 0$), however, if we accept the non-zero maintenance ($m > 0$, Eq. 9), then NZG is predicted to stop at certain threshold substrate concentration, $s^* = Y^m K_s m$, below which μ becomes negative implying net death of starving cells (Fig. 7B).

The laboratory reproduction of continuous NZG is a challenging problem. Batch culture at low starting substrate concentration (s_0) is not adequate because, at any s_0 producing a measurable amount of cells, the growth initially is inevitably too intensive and then abruptly stops and becomes negative after substrate depletion. In theory, a chemostat is a perfect option as the dilution rate, D , can be set at any level. However, the NZG range requires an unrealistically long stabilization period. For example, a chemostat culture would need to be operated uninterrupted for at least 4 to 5 months to establish a steady division frequency of once per month. Over this time period, there is a high probability of accidental equipment malfunction, wall growth, mutations, or contamination invalidating experimental results.

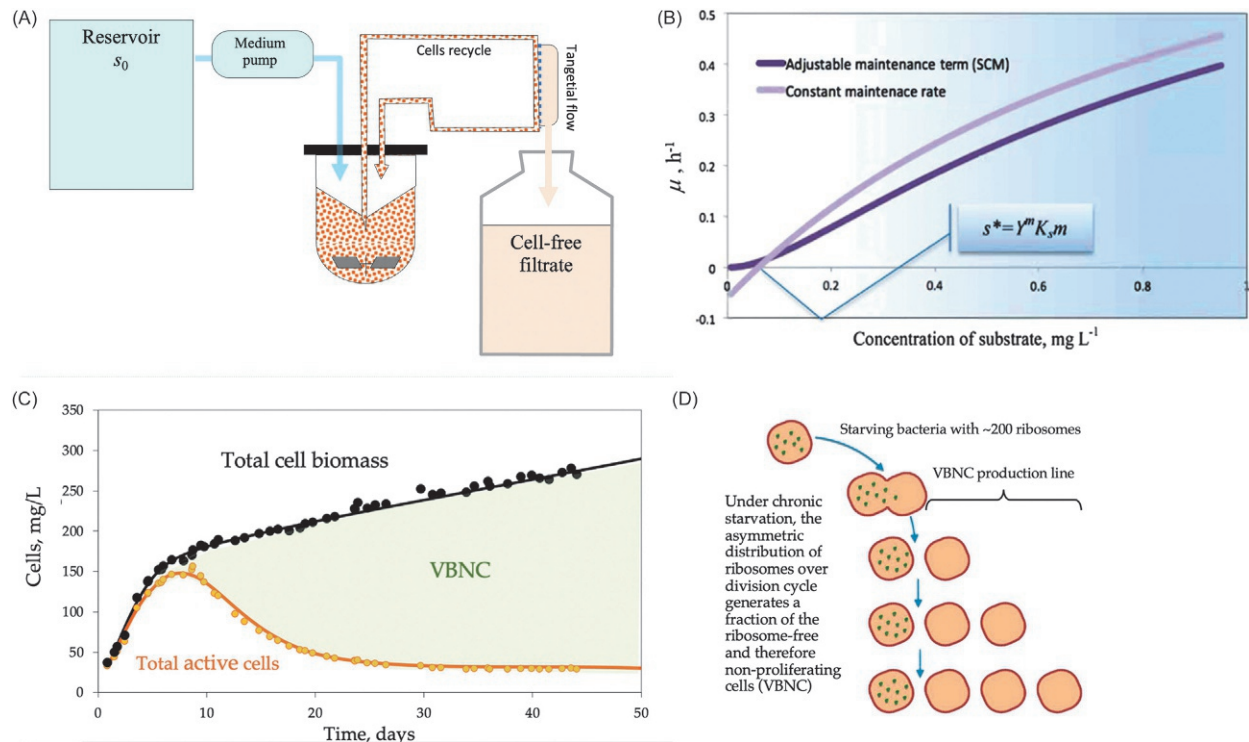


Fig. 7 Laboratory reproduction of the near-zero microbial growth (NZG) typical for soil environment (Panikov et al., 2015). (A) Schematic of continuous cultivation with cell retention (CCR). Medium is continuously delivered from the reservoir, while the outflow is passed through the tangential-flow filter, returning cells to the bioreactor. (B) The effect of the energy source concentration on microbial SGR. The non-structured Monod-Pirt model assuming a constant maintenance rate predicts cell death (negative μ) at $s < s^*$. The more advanced SCM model predicts an adjustable maintenance term allowing the positive NZG. (C) Experimental data on the NZG growth dynamics of *Pseudomonas putida* fitted to the SCM. (D) The hypothetical mechanism explaining the loss of reproductive capacity in the VBNC (viable-but-not-cultivable) cells caused by an uneven distribution of ribosomes during the cell division.

Continuous culture with cell retention (CCR) by using a dialysis membrane or tangential-flow filtration (Fig. 7A), provides a better practical alternative, bringing microbial cells to a monthly division frequency in only 1–2 weeks. The Monod chemostat model with constant positive m as applied to CCR (Eq. 9 combined with 19) predicts attaining of the so-called *state of maintenance*, the transition of microbial cells to the non-proliferating status ($\mu = 0$) preserving relatively high the maintenance catabolic activity. Such a hypothetical state has never been confirmed experimentally (Fig. 7). Instead, it was found (Panikov et al., 2015) that under chronic starvation in CCR, microorganisms undergo deep metabolic reconfiguration (about 40% of proteome replaced as compared with the cells grown at conventional SGR range). An immediate result of such adaptive changes was the progressive decrease of the maintenance term allowing a near linear increase of cell biomass concentration over time (Fig. 7C). The second conclusion was that a stable long-term NZG requires the differentiation of starving cells into growing and non-replicating VBNC (viable-but-not-nonculturable cells) forms, the latter preserving measurable catabolic activity. Over cultivation time, the proliferating cells eventually attain a steady state, their slow growth being balanced by VBNC production. Since growth deceleration is always accompanied by a dramatic decrease in the rRNA content, it was hypothesized that VBNC cells are produced because of asymmetry in the distribution of ribosomal particles between chronically starving mother and daughter cells (Fig. 7D). It is interesting that extremely slowly growing cells acquiring broad resistance to stresses other than starvation, including suboptimal temperature, toxic compounds, desiccation, and osmotic stress. Apart from the described *epigenetic changes*, there is a strong possibility of spontaneous mutations and selection of the GASP (Growth Advantage of Stationary Phase) phenotypes.

Concluding comments

The chapter introduces the basic terms, variables, and concepts established in microbiological growth kinetics. There is a wide range of models that present different degrees of perception of bioprocesses. It is a normal phenomenon, there is no perfectly correct model, and all of them including the simplest formulation could be helpful in scientific exploration provided their selection has been appropriately justified. The choice of biokinetic models should be based on available data and depends on the objectives of a particular study. Specifically, a soil scientist should estimate the degrees of involvement in a process of soil microorganisms and other soil biotas. We can envisage several situations:

- 1) The short-term (the time scale minutes-hours) response of the soil community to some specific factor. The process duration is too short to induce a measurable growth of soil microorganisms. In this case, the process kinetics should follow the equations listed in Table 1. It gives basic information about the kinetic order of reactions that could be very useful. Any growth models are not applicable.
- 2) The time scale of biomonitoring is extended to a few days, and there is a good chance of observing the initial growth of soil microorganisms but without deep changes in the community structure and accompanying deep shifts in environmental conditions. A typical example is soil incubation with easily degradable organic compounds, inducing intensive exponential growth. Such growth can be categorized as *simple biodynamics* with constant SGR and MMCC. Such growth can obey the simplest growth models (Monod, cell quota, logistic equation) or requires more advanced models (SCM, GEM) in the case of clearly expressed lag-phase indicating metabolic reconfiguration during the *complex biodynamics*.
- 3) The next step is the issue of *microbial heterogeneity*; how diverse the acting community is. If the studied bioprocess is carried out by a mixture of species with similar biokinetic characteristics, then the SCM could be still well appropriate to explore the dynamic patterns and help its better mechanistic understanding. Otherwise, researchers must address the multispecies interacting models. That topic will be touched on in the 2nd chapter of this series.

References

- Blagodatskiy SA, et al. (1987) A rehydration method of determining the biomass of microorganisms in soil. *Eurasian Soil Science* 19(3): 119–126.
- Droop M (1973) Some thoughts on nutrient limitation in algae. *Journal of Phycology* 9(3): 264–272.
- Fredrickson AG (1976) Formulation of structured growth models. *Biotechnology and Bioengineering* 18(10): 1481–1486.
- Gasch AP, et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell* 11(12): 4241–4257.
- Hengge-Aronis R (2002) Signal transduction and regulatory mechanisms involved in control of the sigma^S (RpoS) subunit of RNA polymerase. *Microbiology and Molecular Biology Reviews* 66(3): 373–395.
- Kim JI, Varner JD, and Ramkrishna D (2008) A hybrid model of anaerobic *E. coli* GJT001: Combination of elementary flux modes and cybernetic variables. *Biotechnology Progress* 24(5): 993–1006.
- Kompala DS, et al. (1986) Investigation of bacterial growth on mixed substrates: Experimental evaluation of cybernetic models. *Biotechnology and Bioengineering* 28(7): 1044–1055.
- Panikov NS (1995) *Microbial Growth Kinetics*, p. 378. The Netherlands: Springer.
- Panikov NS (2019) 1.18-Microbial growth dynamics. In: Moo-Young M (ed.) *Comprehensive Biotechnology*, 3rd edn, pp. 231–273. Oxford: Pergamon.
- Panikov NS (2021) Genome-scale reconstruction of microbial dynamic phenotype: Successes and challenges. *Microorganisms* 9(11). <https://www.mdpi.com/2076-2607/9/11/2352>.
- Panikov NS, et al. (2015) Near-zero growth kinetics of *Pseudomonas putida* deduced from proteomic analysis. *Environmental Microbiology* 17(1): 215–228.
- Song HS and Ramkrishna D (2011) Cybernetic models based on lumped elementary modes accurately predict strain-specific metabolic function. *Biotechnology and Bioengineering* 108(1): 127–140.
- Strocchi M, et al. (2006) Low temperature-induced systems failure in *Escherichia coli*: Insights from rescue by cold-adapted chaperones. *Proteomics* 6(1): 193–206.
- Tempest DW, Herbert D, and Phipps PJ (1967) Studies on the Growth of *Aerobacter aerogenes* at Low Dilution Rates in a Chemostat, in *Continuous Cultivation of Microorganisms*. Salisbury: H.M.Stat.Office240–253.

Further reading

- Abner K, et al. (2014) Single-cell model of prokaryotic cell cycle. *Journal of Theoretical Biology* 341: 78–87.
- Adamberg K, Valgepea K, and Vilu R (2015) Advanced continuous cultivation methods for systems microbiology. *Microbiology* 161(9): 1707–17019.
- Bergkessel M, Basta DW, and Newman DK (2016) The physiology of growth arrest: Uniting molecular and environmental microbiology. *Nature Reviews. Microbiology* 14(9): 549–562.
- Bruggeman FJ, et al. (2020) Searching for principles of microbial physiology. *FEMS Microbiology Reviews* 821–844.
- Bull AT (2010) The renaissance of continuous culture in the post-genomics age. *Journal of Industrial Microbiology and Biotechnology* 37(10): 993–1021.
- De Jong H, et al. (2017) Mathematical modelling of microbes: Metabolism, gene expression and growth. *Journal of the Royal Society Interface* 14(136): 1–14.
- Esser DS, Leveau JHJ, and Meyer KM (2015) Modeling microbial growth and dynamics. *Applied Microbiology and Biotechnology* 99(21): 8831–8846.
- Gagliardi A, et al. (2016) Proteomics analysis of a long-term survival strain of *Escherichia coli* K-12 exhibiting a growth advantage in stationary-phase (GASP) phenotype. *Proteomics* 16(6): 963–972.
- Garza DR, et al. (2018) Towards predicting the environmental metabolome from metagenomics with a mechanistic model. *Nature Microbiology* 3: 456–460.
- Maier RM (2009) Bacterial growth. In: Maier RM, Pepper IL, and Gerba CP (eds.) *Environmental Microbiology*, 2nd edn, pp. 37–54. San Diego: Academic Press.
- Minkevich IG, Krynetskaya AY, and Eroshin VK (1989) Bistat—A novel method of continuous cultivation. *Biotechnology and Bioengineering* 33(9): 1157–1161.
- Orth JD, Thiele I, and Palsson BO (2010) What is flux balance analysis? *Nature Biotechnology* 28(3): 245–248.
- Palsson BØ (2006) *Systems biology: Properties of Reconstructed Networks*. Cambridge University Press.
- Palsson BØ (2015) *Systems Biology: Constraint-based Reconstruction and Analysis*. Cambridge: Cambridge University Press.
- Panikov NS (1995) *Microbial Growth Kinetics*, p. 378. The Netherlands: Springer.
- Panikov NS (2019) Microbial growth dynamics. In: Moo-Young M (ed.) *Comprehensive Biotechnology*. 3rd edn, 1: pp. 231–273. Oxford: Pergamon.
- Panikov NS (2021) Genome-scale reconstruction of microbial dynamic phenotype: Successes and challenges. *Microorganisms* 9(11): 2352.
- Pirt SJ (1975) *Principles of Microbe and Cell Cultivation*. Oxford, London, Edinburgh, Melbourne: Blackwell Science.
- Skandamis PN and Jeanson S (2015) Colonial vs. planktonic type of growth: mathematical modeling of microbial dynamics on surfaces and in liquid, semi-liquid and solid foods. *Frontiers in Microbiology* 6: 1178.
- Wang G, et al. (2014) Representation of dormant and active microbial dynamics for ecosystem modeling. *PLoS One* 9(2): e89252.
- Zinn M, Witholt B, and Egli T (2004) Dual nutrient limited growth: Models, experimental observations, and applications. *Journal of Biotechnology* 113(1-3): 263–279.

Kinetics of microbial processes: Applications to soil biology

NS Panikov, Chemistry and Chemical Biology Department, Northeastern University, Boston, MA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	188
Estimation of microbial growth rates in nature	189
Understanding true vs apparent growth rates	189
On-site direct microscopy	189
Methods based on the analysis of the cell-division cycle	189
Nonreplicating genetic markers	190
Approaches developed for steady-state populations	190
Water streams	190
Digestive track of invertebrates	190
Stopping microbial grazing	191
Isotope incorporation techniques	191
Heterotrophic CO ₂ fixation	191
DNA synthesis rate	191
Protein synthesis rate	191
Respiration rate	191
Microbial productivity inferred from periodic fluctuations of soil biodynamics	192
First observation and the IBP renaissance	192
The original interpretation of microbial periodicity, calculation of SGR, and deficiencies of the method	192
Contemporary view on periodic fluctuations	193
What is wrong with the original calculation of SGR and how to correct it?	193
Seasonal decomposition of organic matter	195
Principle of the method and its deficiencies	195
Correction for the overstated maintenance requirements	196
Correction for cryptic microbial growth	196
Prospects for correction of the underestimated DOC input to soil	197
Resolving interpretive and computational controversies of the ¹⁴ CO ₂ pulse-chase method	197
Diversity of soil microhabitats	200
Principle of habitats classification	200
Representing examples	200
Growth dynamics in various habitats	202
Diversity of growth characteristics. Life strategy concept	204
Explanation of the concept	204
How the survival triangle is constructed?	205
r-Selected organisms	206
K-Selected organisms	206
L-Selected organisms	207
Molecular basics of biokinetic diversity associated with life strategies	207
Life strategies and the problem of growth recalcitrance	208
Decomposition kinetics	208
Decomposition as self-decay	208
The mean residence times and half-life: the interpretive nuances	210
Decomposition of simple individual compounds	211
Decomposition of pure cellulose	211
Decomposition of plant litter	215
Conclusion and prospects for future developments	215
How to handle soil complexity?	215
Still, what is the rate of microbial growth in the soil?	216
What did we learn from kinetic studies?	216
What we missed	216
References	216
Further reading	217

Abstract

The chapter summarizes the advance in microbial biokinetics aimed at the mechanistic understanding of soil biodynamics. The covered topics include methods to estimate microbial growth rates in nature, the origin of sustained soil fluctuations, navigation to the multitude of microbial habitats, the life strategy concept explaining the biokinetic diversity of soil microorganisms and the uncultivability phenomenon, decomposition kinetics of simple organic compounds, cellulose, and plant litter. The review specifies typical misconceptions and computational handicaps and gives recommendations on the correct account for the maintenance, seasonal microbial production, root exudation rates, and the organism-centered models of soil decomposition.

Key points

- The main challenge in the assessment of microbial growth rates in nature is that the apparent growth (the observed increase of cell number) could be only a small fraction of the true growth hidden by parallel elimination such as grazing by soil animals, lysis, active migration, or wash-out.
- Growth rate of microorganisms in soil and other natural habitats is essentially slower than in laboratory cultures because of non-optimal conditions, primarily nutrient deficiency
- There are few types of microbial habitats in the soil as dependent on nutrient supply, elimination of growing cells and the spatial assembly, these habitats support distinct dynamic pattern of microbial growth. The fastest growth rate is found in habitats with continuous nutrient supply and elimination (rhizosphere, digestive tract of soil animals), the slowest microbial growth is characteristic to the expanded dispersion zone with continuous and slow diffusional input of nutrients without biological elimination.
- The intensity of rhizodeposition as the main source of labile decomposable organic compounds (DOC) can be quantitatively defined using experimental ^{14}C pulse-chase data and the mechanistic kinetic model of the full chain of events: the photosynthetic uptake of CO_2 , Calvin cycle, phloem translocation, release of root exudates and their oxidative metabolism by the rhizomicrobial community. The earlier interpretation of the two-phase $^{14}\text{CO}_2$ evolution as a measure of microbial contribution is incorrect.
- Weekly (with period of 6–8 days) fluctuations of microbial biomass in the planted soils (in situ and in the plant-soil microcosms in climatic chambers) are induced by PPA and superimposed on the prey-predator cycles
- Diversity of biokinetic characteristic of soil microorganisms is explained by the Darwinian evolution under different circumstances of natural selection. We distinguish the products of r-selection (pioneer stages of succession), K-selection (climax communities), and L-selection (stressful environment), each of which has unique biokinetic and genomic characteristics. Modern growth kinetics reproduces the biodynamic patterns empirically observed in the opportunistic species (syn.: zymogenic and copiotrophic) as manifestation of the r-selection. The oligotrophic and autochthonous species are identified as the products of K-selection, and the extremophiles and stress-resistant species as the products of L-selection
- Substrate decomposition in soil does not follow the first-order degradation kinetics, the paradigm widely accepted since introduced by Jenny in 1941, because different plant litter C-pools are NOT degraded *independently* and *simultaneously* with the same rate over time, but *strongly interact* with each other (catabolic repression, inhibition, co-metabolism through ROS formation). Besides the degradation (depolymerization, oxidation) reactions are combined with the biosynthetic processes (growth of microbial decomposers and formation of their extracellular products). Instead, we suggest the organism-oriented decomposition model applied as a specific illustration to the process of cellulose decomposition.

Nomenclature

$D = F/V$	Chemostat dilution rate
DOC	Decomposable organic compounds
F	Medium feeding rate
K_s	Saturation constant defined as $K_s = s$ at $\mu = 0.5\mu_m$
m	Maintenance term, the fraction of catabolic rate that is not associated with growth
NZG	Near-zero growth under chronic starvation
ODE	Ordinary differential equation
p	Product concentration
PDE	Partial differential equation
PPA	Plant photosynthetic activity
$q_p = (1/x) \cdot dp/dt$	Specific product formation rate (SPR)
$q_s = (1/x) \cdot ds/dt$	Specific substrate uptake rate (SUR)
RE	Root exudation
ROS	Reactive oxygen species

s	Limiting nutrient substrate concentration, mostly DOC
SCM	Synthetic Chemostat Model
SGR	Specific growth rate
t	Time
V	Volume of cell culture kept constant in the chemostat
VBNC	Viable but nonculturable cells
x	Cell mass concentration
$Y = -dx/ds$	Cell yield per unit of consumed substrate
$Y_{x/p} = dx/dp$	Cell yield per unit of produced CO_2
$\mu = (1/x) \cdot dx/dt$	SGR calculation
μ_m	Maximum SGR, $\mu \rightarrow \mu_m$ as $s \rightarrow \infty$

Introduction

The first chapter of this series on microbial kinetics outlined the general biokinetic terms, principles, and models. Here our objective is to demonstrate some applications of the described biokinetic approach to soil microbiology. Soil is commonly recognized as an especially tough research object for microbial ecologists. Because of its solid state and opaqueness, soils do not allow direct non-destructive microscopy, flow cytometry, or even a uniform distribution of soluble reactants (e.g., ^3H -thymidine to measure the DNA replication rates) without artefactual distortion of the native structure. Further, any soil is a mosaic of multiple microbial microhabitats, a single soil crumb may be the home of organisms with dramatically different tolerance limits and metabolic types (aerobic-anaerobic, autotroph-heterotroph, prototroph-auxotroph, etc.), each in its own tiny niche that differs chemically or physically from other niches. It makes the soil microbial community especially diverse, accounting for up to one-quarter of Earth's diversity. Most (70–90%) soil prokaryotes are elusive not-yet-culturable species known only from their genomic signatures; this intriguing biokinetic puzzle presents both challenges and opportunities to learn previously unknown molecular principles, and their discovery could impact the entire modern biology.

There are many other not-yet-explained biokinetic features observed in soil, e.g., the periodic fluctuations of microbial activity, soil microbiostasis, poorly understood declines in microbial decomposition rates resulting in paludification and litter accumulation in some soils, and the catastrophic decline of soil organic matter in other soils. For centuries, soil biologists fruitlessly try to estimate the intensity of root exudation as a main source of easily decomposable organic matter and explain the coexistence of the fast- and slow-growing species violating the law of competitive exclusion. These and many other intriguing questions belong to the area of biokinetics, a scientific discipline dealing with the rates and mechanisms of biological processes, an approach combining precise measurement of dynamic variables with mechanistic mathematical modeling.

Soil biokinetic studies were started in 1940–50 focusing on the decomposition (Jenny, 1941), nitrification (Saunders and Bazin, 1973), denitrification (Smith, 1980), and the rhizosphere effect (Newman and Watson, 1977). The initially slow development of soil biokinetics was dramatically accelerated by the International Biological Program (IBP, 1964–1974) aimed to explore global biological productivity as a basis of human welfare. The ecological term *productivity* is synonymous with the biokinetic term *growth rates* addressing the previously ignored dynamic status of natural populations. The IBP demonstrated that microbial ecology lagged far behind the other ecological subdisciplines dealing with plants and animals, and the gap has been quickly filled due to the energetic efforts of several research teams discussed below. The second even more significant impact was produced by the Global Change agenda that made soil biology part of modern biospheric interdisciplinary science. It became imperative to provide quantitative information on the soil resources, C-pools, and fluxes of the greenhouse gases from the soil to the atmosphere. In addition to fluxes and pools, soil biologists were charged with the task of deciphering the native regulatory mechanisms involved in the ongoing climatic changes to be included in the predictive biospheric models and development of the mitigation strategy. The third accelerating factor was the proteogenomic revolution in biology affecting environmental research. It provided powerful molecular tools, generating reliable phylogenetic data on the soil community composition and offered the opportunity to follow the population dynamics of individual taxonomic units and selected taxonomic groups.

This chapter starts with the basic biokinetic question: how to measure the rates of microbial growth in soils and how fast soil microorganisms grow? It is not a trivial task, and to evaluate the true growth rate of microbial populations, we often need to reveal hidden sources and sinks for growing microorganisms. The next issue addressed is the previously mentioned multitude of soil microhabitats. We show how to find a way through this labyrinth and systematize the observations into the space of permissible combinations. Each element of this system has respective analogs in laboratory cultivation opening the way for purposeful modeling and selection of correct growth models. The third part of this article addresses the issues of the biokinetic diversity of soil microorganisms that evolved under different circumstances of natural selection. It answers one of the puzzling ecological paradoxes: why the fast- and slow-growing populations competing for the same nutrient resources do co-exist in nature, apparently violating the law of competitive exclusion. The introduced concept of multiple life strategies explains the paradox and sheds the light on the nature of uncultivability phenomena. Finally, the wealth of accumulated knowledge is applied to the core problem of soil biology, the kinetics of organic matter decomposition.

Estimation of microbial growth rates in nature

Any kinetic study critically depends on the quality of the primary experimental data including the measured rates of a process. Microbiological kinetics is no exception. Unfortunately, the accurate estimation of microbial growth rate in nature is not a trivial task even in the postgenomic era. Thank to culture-independent molecular techniques (such as metagenomic and metatranscriptomic analyses, FISH, SIP, oligonucleotide probes) soil microbiologists now can resolve the taxonomic structure of the whole soil community, its active part, or a subset of species responsible for a given soil process. It is also possible to monitor the changes in the microbial biomass or cell numbers over time $x(t)$. However, it is not enough to answer the question about the growth rates. Let us see where the problem lies by looking at one specific case.

Understanding true vs apparent growth rates

Suppose, our task is to assess the growth performance of bacterial strains intended to use as soil ameliorants. We start by recording the growth curves $x(t)$ of all strains using a plate reader with shaking and temperature controls. Growth is automatically recorded as an increase of optical density (OD) at 600 nm every 5 min. At the end of the experiment, the recorded data are exported to Excel, and SGR for each strain is found by numeric differentiation applying the Excel function SLOPE and AVERAGE for three consecutive OD points vs time:

$$\text{SGR} = \frac{1}{x} \frac{dx}{dt} \approx \frac{1}{\text{AVERAGE}(\text{OD}_1 : \text{OD}_3)} \times \text{SLOPE}(\text{OD}_1 : \text{OD}_3, t_1 : t_3) \quad (1)$$

Let us assume that we found the best candidate strain (short lag phase and high SGR), and want to test its long-term stability using continuous culture with medium feeding rate F , ml h^{-1} and culture volume V , ml. Again, we measure OD in periodically withdrawn samples of bacterial suspension and calculate the SGR, but now the mathematical expression is different:

$$\text{SGR} = \frac{1}{x} \frac{dx}{dt} + D, D = \frac{F}{V} \quad (2)$$

In 1–2 days of cultivation, the derivative dx/dt of Eq. (2) becomes very small and eventually disappeared when the culture attained a steady state when $dx/dt = 0$ and $\text{SGR} = D$. Now we can formulate the principal challenge faced by microbial ecologists exploring growth kinetics in nature: how to distinguish the *apparent* growth rates dx/dt from the *true* rates such as the sum of two terms in Eq. (2), one of which is hidden from the observer. In laboratory experiments, it is not a problem because we design our own experiments. However, the natural growth patterns could remain enigmatic. Over the past several decades, microbial ecologists made some progress in resolving the outlined problem. There are no one-size-fits-all solutions, rather we have a collection of successful procedures that we can group into three clusters: **group B** (from the Batch) applicable to the simple growth regimes without hidden sinks when the apparent growth rate is a true estimate; **group C** (from the Continuous) applicable to a steady-state continuous growth when growth is balanced by the elimination; and finally, **group U** (from Uniform) of the uniformly applicable methods for any growth regime but having other restrictions.

On-site direct microscopy

Regular glass slides (Staley, 1971) or specially designed *artificial substrates*, such as rectangular capillaries (Perfil'ev and Gabe, 1969) were installed at the field sites (ponds, wetlands, submerged soils) to track the colonization and following growth by native microbial populations. The compound phase-contrast microscopes were available nearby or permanently immersed in water (Staley, 1971). Interestingly, it was found that *Chlorella* cells grew without division during the day, and divided into four daughter cells at night. Occasionally this cycle was interrupted, with no cell division until the second night. This technique requires exceptional microscopic skills and patience (James Staley was one of such high-rank professionals) and cannot be recommended for wide use. It is also limited to clean aquatic habitats and is selective to the relatively large cells adherable to the glass surface. Today underwater microscopy is supported by a specialized video-recording hardware (Mullen et al., 2016). Potentially, the Staley approach can be upgraded today to the automated surveillance system based on the multiple microfluidic devices permanently installed in a pond, fen, or saturated soil. Research diving will be not required, and video recording of the growth dynamics can be transmitted via optical fiber. Thus, on-site microscopy has a good chance to provide exceptionally valuable data on division frequency, responses to fluctuating environments, and other autecological features of native populations. The described approach is qualified as a B group, and any immigration and emigration or other elimination events must be accounted for.

Methods based on the analysis of the cell-division cycle

In eukaryotes, the mitotic cells can be recognized morphologically, and the time of mitosis (TM) is constant, although the total cell cycle duration (equivalent to the generation time $g = \ln 2/\mu$) varies. In prokaryotes, the equivalent of TM , the nearly constant part of the cell cycle, starts with the formation of division septa and ends with the separation of daughter cells. TM must be established for a specified eukaryotic microorganism (fungi, algae, protozoa, slime mold) in preliminary experiments with pure cultures. To find the

generation time g , a researcher needs to perform a direct microscopic examination of the environmental sample and count the total number of the targeted organism (N) and the number of mitotic cells (N^*). Then the generation time g is found as

$$g = \frac{TM}{R} \text{ or } g = \frac{TM}{1.44R}, R = \frac{N^*}{N} \quad (3)$$

where 1.44 is empirical correction factor (Brock, 1971). As applied to the prokaryotes, the FDC (Frequency of Dividing Cells) method is not restricted to a selected taxon, instead, it was suggested that the method be calibrated using the enriched chemostat culture of bacteria from the studied habitat (Hagström et al., 1979). The method belongs to the U-group, as in theory it does not depend on a hidden microbial sink. There are several drawbacks, the most essential computational inconsistency being Eq. (3), which implicitly assumes that the native population is not synchronized and, as a result, the fraction R remains constant over time. However, this assumption contradicts the observations on the nocturnal division event in *Chlorella* (Staley, 1971). Another drawback is that the critical parameter TM must be known and based on the ex situ culturing. Therefore, this technique is applicable to only known culturable species. First, many native microorganisms cannot be cultivated. Second, even well-known organisms are not easily distinguishable without FISH, immunofluorescent staining, or other advanced microscopic methods. It does make sense, to apply this approach to relatively simple synthetic or native communities, e.g., the rumen, hydroponic plant culture, mycorrhiza of well-studied plants, and soil-plant inoculation experiments including the application of mycorrhiza-infected roots.

Nonreplicating genetic markers

Two such markers have been tested: (i) superinfecting mutants of phages that enter the cells of the pathogenic strain of *Salmonella typhimurium* but do not replicate (Meynell, 1959) and (ii) ^3H -thymidine labeling of bacteria after label removal (Brock, 1971). In both cases, the labeled cells are mixed with the unlabeled indigenous populations. At each cell division, the fraction of the population which contains the label is halved, therefore, the doubling time can be calculated from the rate of marker dilution. Computationally, the isotopic dilution is equivalent to the elimination term D in Eq. (2) if we use as a dynamic variable the concentration of labeled cells $x^*(t)$, therefore it belongs to the C-category. In practice, this approach can be applied only to a limited number of genetically explored species and to very simple communities or single-species studies, such as infected organs of animals. Specific drawbacks: a, the accuracy of SGR estimation drops dramatically over time parallel to the label dilution, and b, the method is based on assumption that the biokinetic qualities of the labeled and the native cells are identical, but the labeling procedure involves laboratory cultivation that significantly affects the phenotype. In any case, the approach played a historically important role (Meynell, 1959): for the first time, it demonstrated a dramatic decline in bacterial growth rates in vivo as compared with laboratory cultivation. *S. typhimurium* dividing in nutrient broth every 30 min after intravenous inoculation into a mouse grew 20 times slower!

Approaches developed for steady-state populations

Here we outline three techniques that belong to the C-type and are designed to measure the normally hidden elimination factors playing the same role as cells washout in continuous culture.

Water streams

Some natural habitats have a visible similarity with the continuous cultivation of cells like chemostats. Moreover, the volume replacement rate equivalent to dilution rate D is often easier to measure than SGR. The classic example is the hot-spring drain ways as a habitat of thermophilic algae (Brock, 1971). The technique involved the measurement of the algal washout rate after growth was stopped by darkening the system with a wide non-transparent screen. The following exponential decrease of $x(t)$ in the effluent stream was monitored to calculate the dilution rate equal numerically to SGR under normal illumination. It was found to be 0.4 d^{-1} ($g = 40 \text{ h}$).

Digestive track of invertebrates

The soil-related example is the digestive systems of the earthworms, isopods, millipedes, and other widely spread in nature invertebrates that continuously consume detritus (e.g., plant litter) converting the ingested matter into excrements highly enriched with microorganisms. Assuming the chemostat-type bioprocess for animals with a short gut (isopods), we can estimate the growth rate of microbial commensals as:

$$\text{SGR} = D \frac{(x_{\text{out}} - x_{\text{in}})}{x_{\text{out}}} \quad (4)$$

where x_{in} is microbial biomass content in the animal feed, x_{out} is the enlarged biomass in fresh excrements. Here D is an equivalent of the dilution rate in a chemostat found experimentally as $D = 1/TT$, the reciprocal of the food transient time TT , how long it takes for ingested food to travel from an animal's mouth to anus. Parameter TT is measured by feeding the animal with stained plant litter or soil labeled with inert visible particles, e.g., silicone beads, see below the difference between TT and half-decay time.

Stopping microbial grazing

In many natural environments including soils, the main elimination factor for growing microbial cells is their grazing by protozoa, microarthropods, nematodes, and other small animals. This elimination rate cannot be measured directly as in method described under *Water streams*. Instead, the microbivores are removed by filtration, based on the solid fact that even the smallest predators are bigger than bacteria. In marine and freshwaters ecosystems, bacterial cells have a size of 1–3 μm and do not form multicellular colonies or flocks, hence filtrations through 5- μm filters allow efficient removal of the protozoa and metazoan predators without major disturbance of bacterial populations. Then the temporal turbidity increase versus unfiltered control is a good measure of their true growth rate in situ. This approach must be used with care since filtration can reduce the nutrient flux produced by large photosynthetic organisms including algae.

Isotope incorporation techniques

Heterotrophic CO_2 fixation

Dark $^{14}\text{CO}_2$ fixation rate was suggested as a measure of the total heterotrophic bacterial production (Romanenko, 1964). The carboxylation of pyruvate produces oxalacetate, it is one of several *anaplerotic* metabolic reactions, serving to refill the intermediates of the TCA cycle used for the biosynthetic processes. In theory, the stoichiometry of anaplerotic reactions should be tightly coupled with the overall cell growth. However, the experimental data display a wide range of 0.01–0.12 between the fixed CO_2 and the total assimilated C, most probably because of the multitude of alternative metabolic pathways. Thus, it is advisable to establish the conversion factor for a given soil and for specific conditions in preliminary incubation experiments. The method is very sensitive and robust for diverse environments including soils. It can be used as a non-destructive way to introduce the label (^{13}C or ^{14}C) through the soil aeration pores to practically any metabolically active soil chemotrophic organisms for following molecular studies, such as SIP. Phototrophic organisms can also consume the label under light exposure. The last case no longer belongs to the heterotrophic category, it is the autotrophic CO_2 fixation rate with a simpler 1:1 stoichiometry, i.e., $\text{SUR} = \text{SGR}$, the uptake rate is equivalent to the growth rate if the biomass $x(t)$ is consumed substrate is expressed in the same C-units, e.g., mg C cm^{-3} .

DNA synthesis rate

Another popular method is the measurement of the uptake rate of labeled precursors of nucleic acids, thymidine (DNA), uridine (RNA), and adenine (both macromolecules) (Karl, 1980). The use of isotopes with high specific activity guarantees minimal alterations of the in-situ growth conditions. At the same time, the amount of added nucleoside should be large enough to suppress their synthesis *de novo* from endogenous cell compounds. The main deficiency of this technique, as applied to soil, is an uneven distribution of precursors within the soil mass. Typically, nucleosides are added as a buffered solution making a soil slurry, which can dramatically affect the growth rates of soil microorganisms. Another problem is the ambiguity of conversion factors from a nucleoside uptake rate to a microbial growth rate. Also, the resulting SGR could be overestimated because of nucleoside utilization as a C source.

Protein synthesis rate

Similar to the previous method, but using labeled leucine instead of thymidine to estimate the SGR from the rate of label incorporation into cellular proteins.

All these isotope methods are sensitive enough to reduce the incubation time down to a few minutes (which must be extended to a few hours in extreme and poor environments, e.g., the polar, and hypersaline waters, permafrost, etc.), they measure the instant biosynthetic activity of environmental cells, and results are not affected by the parallel slower elimination processes. Therefore, the isotope methods belong to the U-category and can be applied to any taxonomic or metabolic type of microorganisms. There are no restrictions in applying all three subcategories to homogeneous habitats (fresh and marine waters, wastewater treatment, rumen). However, soils and other heterogeneous habitats cannot be uniformly mixed with labeled biosynthetic precursors (see above comment about DNA synthesis rate) without at least partial disturbance, but the heterotrophic CO_2 fixation method is the most universally applicable.

Respiration rate

The respiration intensity can be estimated from the rates of dissolved oxygen consumption (the most appropriate in aquatic habitats) or CO_2 formation in terrestrial environments, including soils. CO_2 can be recorded with an infra-red gas analyzer (IRGA), portable mass spectrometers, and other analyzers, either in situ with soil chambers or in laboratory incubation experiments with undisturbed soil cores or homogenized sieved and substrate-amended soil samples. In the field and plant-soil microcosm studies, the crucial problem is to distinguish microbial versus plant root respiration, the topic was exhaustively reviewed by Kuzyakov and Larionova (2005). If root respiration is successfully resolved, then the total respiration rate (RR), expressed as $\text{mg CO}_2\text{-C (h kg soil)}^{-1}$ can be used to calculate the average SGR of heterotrophic soil microorganisms:

$$SGR = \mu = Y_{x/p} \frac{RR}{x_{soil}} \quad (5)$$

where x_{soil} is an independently measured total biomass of respiring soil microorganisms, mg cell-C (g soil)⁻¹, and $Y_{x/p}$ is the cell yield per mass unit of produced CO₂. Microbial growth limited by the energy and C-sources is not accompanied by a noticeable release of products other than CO₂, therefore growth C-balance is very simple:

$$\begin{array}{ccccc} \text{Consumed substrate} & & \text{Produced cells} & & \text{Produced CO}_2 \\ \Delta s - C & = & \Delta x - C & + & \Delta p - C \end{array} \quad (6)$$

Dividing both parts of Eq. (6) by $\Delta x - C$, we have

$$\frac{1}{Y} = 1 + \frac{1}{Y_{x/p}} \text{ or } Y_{x/p} = \frac{Y}{1 - Y} \quad (7)$$

where Y is the conventional yield factor expressed as g cell-C/g substrate-C. For most heterotrophic microorganisms Y varies between 0.5 and 0.6 g C (g C)⁻¹ (Payne, 1970) implying an $Y_{x/p}$ range 1.0–1.5. It can be more accurately determined in preliminary soil incubation experiments under specified conditions (see below). The main advantages of this method are (i), well-established instrumental techniques for monitoring CO₂ emission from soils to the atmosphere at different footprints from a closed chamber covering a fraction of m² to the Eddy covariance flux towers with footprints 10³–10⁴ m²; (ii) a central role of respiration in the metabolism of any organism and its well-predicted stoichiometry (Payne, 1970); (iii) availability of the CO₂ fluxes data for the entire biosphere like FLUXNET that makes possible to generate microbial productivity patterns in seasonal and year-to-year dynamics without dedicated experimental efforts. The limitations: (a) significant deviations from a simple stoichiometry (Eq. 6) under anoxic conditions (should be added specified fermentation products to Eq. (6)) and under deficiency of mineral nutrients (additional products, mostly the polycarbonic acids); (b) complexity of the measured CO₂ fluxes that are not equivalent to the RR, being affected by the transient processes at the soil-atmosphere boundary, e.g., the temperature inversions; (c) possible contribution of the abiotic decarboxylation reactions and release of the trapped CO₂ especially significant in soils with low biological activity.

There are other microbial processes associated with gas exchange, e.g., methanogenesis (uptake of H₂ and CO₂, formation of CH₄), methane consumption (uptake of CH₄, release of CO₂), denitrification (formation of N₂, NO, and N₂O), CO and H₂ consumption by the carboxydobacteria and the hydrogen bacteria, etc. Growth rates of microbial populations responsible for the listed processes can be also estimated using equations similar but not identical to Eq. (5): the total microbial biomass x_{soil} and the yield factor $Y_{x/p}$ must be updated for a respective physiological group of soil microorganisms; fortunately, it can be done relatively easy in a preliminary calibration (Panikov and Sizova, 1996).

Microbial productivity inferred from periodic fluctuations of soil biodynamics

First observation and the IBP renaissance

More than 100 years ago, a team of protozoologists led by Ward Cutler at the Rothamsted Experimental Station performed one of the most scrupulous soil studies (Cutler et al., 1923): every day at 10 am for 365 (!) consecutive days, they undertook a full taxonomic survey of soil protozoa with a differential count of the active and dormant (tolerant to 2% HCl forms as well as the total bacterial count by plating one soil dilution 4×10^{-6} on mannitol-asparagine-mineral agar. They made two important observations for us: (i) a statistically significant correlation with time delays between the protozoan and bacterial counts characteristic for the prey-predator interactive system, two amoebae were recognized as the most active soil microbivores; and (ii) a remarkably sustained periodic fluctuation in the numbers of bacteria and protozoa with periods 4–8 days that were not related to environmental factors. Such fluctuations would not be detected with a less regular sampling schedule.

After 40 years, the same high-frequency sampling was continued as a systematic soil biodynamic study across the East European Plain and Siberia from polar deserts to subtropics. It was a part of the IBP with the goal of estimating microbial growth rates in diverse terrestrial ecosystems. The plating was replaced by the direct count of bacteria stained by erythrosine. The periodicity of bacterial fluctuations was confirmed but not for all soils, the highest amplitudes were found in the tundra, where the count could rise in a few days by more than three orders of magnitude, up to 1500 times (Parinkina, 1973). Moving to lower latitudes, from tundra to taiga, deciduous forest, and steppe, fluctuation amplitude progressively declined, and no fluctuations were detected in semi-desert soils (Aristovskaya, 1974).

The original interpretation of microbial periodicity, calculation of SGR, and deficiencies of the method

It was hypothesized that, apart from the predator-prey interactions, the fluctuations were caused by the periodic accumulation in the soil of some unknown autoinhibiting metabolites susceptible to spontaneous oxidation or volatilization, e.g., H₂, ethylene oxide, a hypothetical compound *periodine* (Aristovskaya, 1974). The overall production of the fluctuating bacterial growth was calculated as follows: the bacterial abundance curve over time, $x(t)$ was split into the fragments separated by the local extremes $x_i^{\max}$ and $x_i^{\min}$, then the bacterial production was calculated as a sum of all positive increments, $\Delta x_i = x_i^{\max} - x_i^{\min}$. Each negative increment was interpreted as cells consumed by predators and was either added to the total (Aristovskaya, 1974) or ignored, resulting in a two-times smaller bacterial production (Parinkina, 1973). The generation time, g , was found to vary from 7 to 100 h, surprisingly intensive growth as compared with other published results. We will show below that this calculation was not correct.

The described approach was not widely accepted as a technique to estimate SGR. It misses fungi and actinobacteria, the mycelial organisms whose abundancies cannot be expressed in cell numbers. The procedure is too laborious, cumbersome, and destructive to the field site because of extensive soil coring (humorously called 'Swiss cheese syndrome'). Another discouragement stems from the stochastic nature of soil fluctuations, they rarely look like regular *oscillations*, rather the distinct rises and drops sooner or later become a random scatter. There were also doubts about the authenticity of the phenomenon since the observed variations of bacterial count over time $x(t)$ could be interpreted as a result of the x variation across the space, the different bacterial numbers in the replicate soil samples. The most radical criticism was the logical argument denying the mere possibility of the discussed phenomenon based on the following reasoning: an observable periodicity requires a global growth synchronization of soil microorganisms, but it is impossible in a mosaic soil environment because bacterial cells occupying separated soil microsites *a priori* should not communicate with each other. Such extreme nihilism is not fair as we show below. Periodicity patterns may be not the best primary data to estimate SGR, but they do exist as a biodynamic phenomenon that deserved to be thoroughly explored for a better understanding of soil ecology.

Contemporary view on periodic fluctuations

All doubts about the authenticity were adequately addressed by complementing the manual and destructive procedure of bacterial count with other techniques including the instrumental non-destructive recording of gas emission. Fluctuations of the expected duration of 4–8 days were established for CO₂ emission rates and soil gas profiles using the permanently installed membrane gas probes and the collars of the gas flux chambers with removable tops (Panikov and Gorbenko, 1992; Panikov et al., 2007). Even more accurate recording of the diurnal and day-to-day variation of gas exchange dynamics was achieved in the laboratory plant-soil microcosms kept in climatic chambers with artificial illumination (Panikov, 1995). On the other hand, periodic fluctuations were never found in the unplanted soil, indicating the essential role of plants as inducers of soil fluctuations. It was also confirmed by the time-series statistical (harmonic and cluster) analyses of the field experiment (Panikov and Gorbenko, 1992). It revealed that every rise in the fluctuating soil biodynamics started with increasing of the plant photosynthetic activity (PPA) followed by the delayed response of soil microorganisms (soil respiration, increased bacterial count), and then even more delayed protozoan activation.

Contrary to the earlier hypotheses explaining microbial fluctuations by the internal soil factors, the plant-centered concept does allow for *synchronization* of microbial activity throughout the soil mosaic because the plant roots and associated mycorrhiza penetrate the entire soil layer reaching most of the microbial microhabitats and combine them into an integrated cross-communication system. Diurnal cycles of PPA and root exudation force rhizosphere bacteria to divide synchronously roughly once a day in the rhizosphere. Day-to-day variations in rhizodeposition can affect microbial activity in two ways, by changing the concentration of decomposable organic compounds (DOC) and the number of soil microsites within the zone of plants influence.

Last question, what causes the PPA of plants to fluctuate? A fully competent answer should be given in the future based on interdisciplinary research crossing the borders of plant and soil biology. The known facts primarily based on the CO₂ fluxes can be summarized as follows: (i) fluctuations occur either outdoors or indoors; (ii) the fluxes are synchronized across the geographical scale well above the ecosystem or landscape levels (Fig. 1); (iii) variable environmental conditions (moisture, temperature, illumination) affect the flux intensity but not the *periodicity* of fluctuations. Thus, a *master* oscillator is likely to be of an extraterrestrial radiation nature, such as *solar wind*. Curiously, the atmospheric chemistry data (von Savigny et al., 2019) has a clear signature of the 27-day Sun rotation cycle, and it produces subharmonics remarkably similar to soil fluctuations.

What is wrong with the original calculation of SGR and how to correct it?

The original method of calculation (Aristovskaya, 1974; Parinkina, 1973) is based on intuition and is not accurate. To find the true SGR, we simulated the fluctuating growth using the following mathematical model:

$$\text{Plant – derived DOC – pool, } s \quad \frac{ds}{dt} = F - q_s x; F = A \left(B + \pi \frac{t}{T} \right); q_s = \frac{Q_s s}{K_s + s} \quad (8)$$

$$\text{Microbial biomass, } x \quad \frac{dx}{dt} = \mu x - q_x z; \mu = Y q_s; q_x = \frac{Q_x (x - x_0)}{K_x + x - x_0} \quad (9)$$

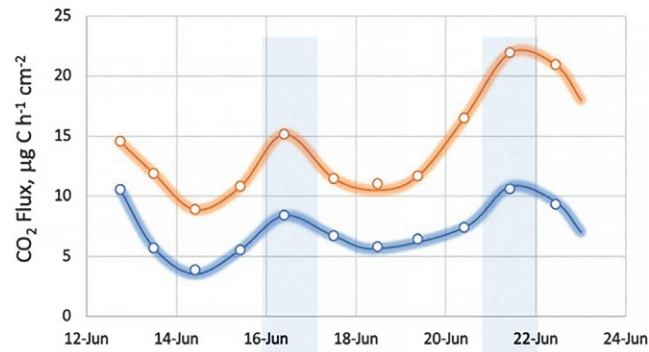
$$\text{Protozoa, } z \quad \frac{dz}{dt} = \eta z - a_z z; \eta = Y' q_x \quad (10)$$

$$\text{CO}_2, p \quad \frac{dp}{dt} = (1 - Y) q_s x + (1 - Y') q_x z + a_z z - D \left(\frac{p}{\Theta} - c_0 \right) \quad (11)$$

$$\text{The sum of all variables} \quad \frac{ds}{dt} + \frac{dx}{dt} + \frac{dz}{dt} + \frac{dp}{dt} = F - D \left(\frac{p}{\Theta} - c_0 \right) \quad (12)$$

The DOC are generated by the plants through the process of rhizodeposition, F , proportional to PPA (green curve in Fig. 2). DOC are consumed by microorganisms producing new cell mass and CO₂, and microbial cells are eliminated by protozoa. The sum of four variables, s , x , z , and p close the C-balance, so Eq. (12) is used for the verification based on the conservation law. The diurnal variations are neglected for clarity but the day-to-day variations are simulated using the sine function (Eq. (8), F depends on time,

Moscow, 55.75° N 37.62° E



Pushchino, 54.94° N 37.35° E

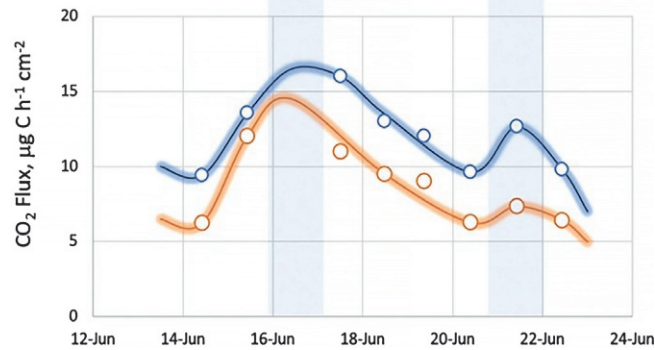


Fig. 1 Example of the soil biodynamics synchronization at the geographical scale. CO₂ emission from soil was simultaneously recorded at two field sites 100 km apart. The shaded areas mark two synchronous emission peaks (Panikov and Gorbenko, 1992).

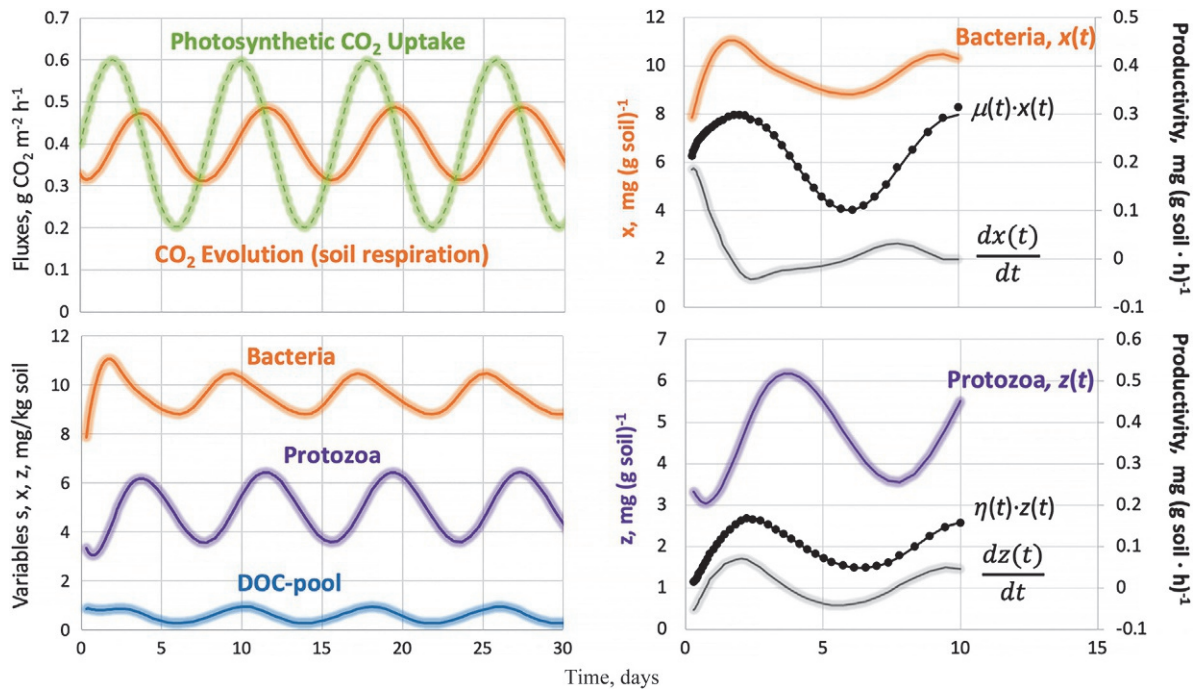


Fig. 2 Simulation of the periodic soil fluctuations and demonstration of the way to reconstruct the SGR dynamics based on the observed amplitude and frequency of fluctuations.

A and B are the constants, and T is the period). θ stands for the soil air-filled porosity, required to define the concentration gradient between soil CO_2 , p , and the atmospheric CO_2 concentration, c_0 .

Simulation results (Fig. 2, left column) show the idealized fluctuation patterns that are not always observed in nature but other essential features (growth intensities, time delays, CO_2 emission, soil biomass) do agree with the published field observation (Panikov and Gorbenko, 1992). The right column compares the former intuitive and the correct ways of calculating microbial productivity: the top panel shows results for bacteria and the bottom panel shows similar curves for protozoa. The primary observation data collected in the field are the bacterial $x(t)$ and protozoan count $z(t)$. Normally it is obtained daily but we use for clarity the smaller time step of one min to produce a continuous curve. The former intuitive calculation is shown as gray curves; for clarity, the positive increments $\Delta x_i/\Delta t_i$ are also presented with the 1 min temporal resolution as of the positive derivatives $dx(t)/dt$ and $dz(t)/dt$. The perfectly correct productivity of bacteria $\mu(t) \cdot x(t)$ and protozoa $\eta(t) \cdot z(t)$ was obtained from numeric integration and plotted as continuous black curves. We see that intuition fails badly. Two variants (with and without adding the negative terms Δx_i) gave respectively 15- and 30-times underestimated bacterial and 5-times underestimated protozoan productivity vs correct results. The reason is a wrong instinctive assumption that bacteria are not grazed over periods when Δx_i are positive and do not grow when Δx_i are negative. The truth is that fluctuations never stop the growth of either bacteria or their grazers, rather their growth rates vary in response to the fluctuating environmental conditions. In fact, bacterial SGR depends on soil temperature and moisture and on available DOC and does not depend on the density of grazers or the rate of elimination. Another curious correction result is that the true SGR of soil bacteria fluctuates between 0.7 and 0.27 day^{-1} equivalent to the division frequency once per day and per 2.5 days respectively. The same SGR is found for microbivores. Another wrong intuitive expectation is that the microbial generation time should match the fluctuation period. The soil weekly periodicity is completely independent of bacterial division frequency contrary to the diurnal cycles (see observation on *Chlorella* (Staley, 1971)): bacteria divide once per day when DOC reaches maximum, the next division happens in 2 or 3 days. Note that the discussed relatively high growth intensity is limited to the rhizosphere and the top soil enriched with DOC, rather than to the entire soil community.

How to fix the problem of the SGR calculation based on weekly fluctuations? Is it possible to recalculate the published data? Yes, the problem can be fully resolved but only if there are both sets of primary data, $x(t)$ and $z(t)$. The first practical step is the smoothing of the time series using splines followed by numeric differentiation to obtain the apparent rates dx/dt and dz/dt , both steps can be completed with regular Excel. Then we add to both derivatives the hidden parts of the growth mass balance (marked below in red):

$$\text{Protozoan productivity} \quad \eta(t) \cdot z(t) = Y' q_z z = \frac{dz}{dt} + a_z z \quad (13)$$

$$\text{Bacterial productivity} \quad \mu(t) \cdot x(t) = Y q_s x = \frac{dx}{dt} + \frac{1}{Y} \left(\frac{dz}{dt} + a_z z \right) \quad (14)$$

Eqs. (13) and (14) are obtained by rearranging Eqs. (9) and (10) respectively. The correction is simpler for protozoa, it is just a mortality term. For bacteria, the hidden production part is the consumed microbial mass required to produce the total mass of microbivores including the dyed part. Fig. 2, the right column presents the corrected results as filled circles, and as we see, the circles perfectly match the continuous curve indicating that two independent computing ways produce the same results. Correction success critically depends on the selection of two parameters, Y' and a_z , which can be borrowed from the published sources (Curds and Cockburn, 1968) or adjusted using computing procedures minimizing simulation error (Excel users are encouraged to implement the Solver add-in). The count of bacteria alone without grazing data (or other elimination processes) is not sufficient for adequate correction.

Seasonal decomposition of organic matter

Principle of the method and its deficiencies

As in measurements based on respiration rate, this approach is based on growth stoichiometry (Eq. 6), the only difference is that the term $\Delta p\text{-C}$ (produced CO_2) is replaced by $\Delta s\text{-C}$ (consumed substrate). Rearrangement of Eqs. (6) and (7) gives the following equivalent of Eq. (5):

$$\text{SGR} = \mu = Y \frac{\text{SUR}}{x_{\text{soil}}} \quad (15)$$

where SUR stands for substrate uptake rate. As most soil microorganisms are decomposers, SUR is equivalent to the round-year average *soil decomposition rate*. Eqs. (5) and (15) look similar but as a source of primary observational data SUR is not defined as clearly as RR. In addition, the CO_2 determination is done instantly with high precision and accuracy using well-developed instrumental procedures. However, soil decomposition is a process studied for centuries but still poorly quantified because of its natural complexity (see below). It is a gigantic matrix of biotic and abiotic, catabolic, and biosynthetic reactions, thousands of reactants, products, and regulatory factors. Some recalcitrant DOC components have turnover times of up to decades and centuries, some are degraded in hours, and some remain unknown. To summarize, seasonal decomposition cannot be recommended as an experimental protocol but it can be useful as a computational experiment for establishing the upper limit for microbial productivity allowed by the C-balance of an ecosystem. For this purpose, it has been implemented in several studies of diverse ecosystems (see references in Panikov, 1995) revealing rather slow growth of soil microorganisms with generation time from 10 to 100 days and sometimes producing absurd results of negative SGR. Below we discuss three sources of computational inconsistencies and suggest relevant corrections.

Correction for the overstated maintenance requirements

The meaning of the maintenance term, m - the non-productive uptake of catabolic substrates - is clear but its direct measurement is impossible. Instead, it is found indirectly as a deviation from the proportionality between the catabolic substrates uptake rate, q_s , and the SGR, μ :

$$\mu = Yq_s = Y^m(q_s - m) \quad (16)$$

$$Y = Y^m \frac{q_s - m}{q_s} \quad (17)$$

where Y^m is the constant standing for the substrate-to-biomass conversion efficiency under imaginary condition $m = 0$. Classic growth kinetics assumes the m -term to be a constant, and it is indeed acceptable for a narrow SGR range from μ_m down to $\sim 0.1\mu_m$. But under deep nutrient limitation ($\mu \ll \mu_m$), cells undergo adaptive metabolic reconfigurations, remarkably reducing their maintenance requirements. Soil modelers in their growth equations occasionally predicting the negative SGR make two mistakes: (i) apply inadequate growth models with a constant m -term ignoring the adaptive self-adjustment of cellular catabolic machinery, and (ii) borrow from published sources the maintenance coefficient m in combination with the apparent yield Y rather than Y^m , making a 'double charge' because the Y -values already account for the maintenance. Both mistakes can be fixed by using more advanced and fortunately relatively simple kinetic models like SCM (see Section "Decomposition kinetics").

Correction for cryptic microbial growth

Cryptic growth refers to hidden microbial growth as a cannibalistic reutilization of the microbial necromass. For example, the depleted batch culture displays a net decline in total cell biomass, but some cells remain viable and continue an invisible slow growth by utilizing the necromass and the autolytic products of other cells. Most soil models ignore cryptic growth (Table 1, Eq. 18), assuming that microbial cells grown on the plant-derived DOC mysteriously disappear with the rate constant k_1 rather than stay in the soil and further participate in the soil C-cycle.

More realistic options for the grown cells are: (i) to stay in the soil and continue the DOC consumption; (ii) enter the dormancy state, (iii) be killed by parasites (phages, Bdellovibrio, etc.) or consumed by microbivores to produce the exploiter's biomass, CO_2 and the excretory or lysis products refilling the DOC-pool, (iv) be transformed into stable humus-like compounds with a long turnover time. For the climax soil community at a quasi-steady state, option (i) is accounted for simply by removing a meaningless rate constant k_1 leading to nowhere; we also neglect (ii) and (iv) based on the reasoning that transition to dormancy is compensated by germination, and humification is balanced by humus degradation. So, the only pathway to be considered is iii) as shown in the second graph and translated to the second more realistic set of ODEs. Eqs. (19), (20), and (21) give three improved estimations of microbial productivity respectively for microbial decomposers, for microbivores, and for their sum. Note, that we combine predation and parasitism into one elimination factor. Bacteriophages already attracted the attention of microbial ecologists as a strong factor impacting soil microorganisms. However, adequate biokinetic data on their lytic and lysogenic cycles are sparse. To economize on words, we refer to protozoa as the most active bacterial predator. The common opinion is that soil fungi are consumed mostly by nematodes, and both groups (bacteria and fungi) are eliminated to different degrees by microarthropods, viruses, and bacterial Bdellovibrio-like parasites.

As expected, the highest production is predicted by Eq. (21) simply because it sums up two trophic groups. Interestingly, with $Y=Y' \geq 0.5$, the cumulative microbial production exceeds the primary DOC input, and this conclusion does not contradict the mass conservation law. As a matter of fact, growth with recycling occurs as a multistep process likely involving a succession of species. Each step obeys the first and second laws of thermodynamics, with some consumed energy being dissipated. Eq. (19) for soil microorganisms alone also predicts their production exceeding the input F under conditions that $Y \geq 0.66$ at a fixed $Y' = 0.5$. Both

Table 1 Mathematical expressions designed for estimation of microbial productivity in soils.

Graph	Mass-balance ODEs	Productivity	
	$\dot{S} = F - q_s X$ $\dot{X} = Y(q_s - m)X - k_1 X = \mu X - k_1 X$ $\dot{P} = (1 - Y)q_s X - k_D P$	$\mu X = Y(F - mX)$	18
	$\dot{S} = F - q_s X + k_2 N$ $\dot{X} = Yq_s X - q_X Z; \mu = Yq_s$ $\dot{Z} = Y'q_X Z - k_1 Z; \eta = Y'q_X$ $\dot{N} = Y'q_X Z - k_1 Z$ $\dot{P} = (1 - Y)q_s X + (1 - Y')q_X Z - k_D P$	$\mu X = F \frac{Y}{1 - Y Y'}$ $\eta Z = F \frac{Y Y'}{1 - Y Y'}$ $\mu X + \eta Z = F Y \frac{1 + Y'}{1 - Y Y'}$	19 20 21

Red arrows indicate the external inputs or outputs fluxes to the microbial compartment. The variables: S, the decomposable substrates; X, microorganisms; Z, microbivore protozoa; N, necromass and excretory products. The symbols over arrows stand for: PP, the fraction of the primary production filling the DOC-pool, μ and η are the SGR of respectively soil microorganisms and predators, k_1 and k_2 are the rate constants for respectively death of protozoa and conversion of necromass to the DOC-pool.

yield factors are still within the published range for the bacteria (Payne, 1970) and protozoa (Curds and Cockburn, 1968). The reutilization has a modest impact on predator production (Eq. 20) because these organisms are assumed to be unable to grow cryptically on the lytic products, and a minor positive impact of reutilization is due to the consumption of the cryptically grown prey.

Prospects for correction of the underestimated DOC input to soil

The improved expressions (19)–(21) can still substantially underestimate microbial productivity if we use the wrong input parameter F . The conservative estimate of F is often limited to the seasonal formation of plant litter, which varies from 50 to 150 g m⁻² in the tundra to 2500 g m⁻² in the virgin steppe (Panikov, 1995). These estimates are not accurate and could be misleading if we take the recalcitrant plant necromass as the only DOC source. A more abundant and accessible nutrient source is the **rhizodeposition**, which includes 3–23% of the particulate matter (the sloughing-off border cells, mucilage, and dead root hair), while the bulk 77–97% belongs to the soluble pool of *root exudates* (RE), a mixture of sugars, amino acids, polycarbonic acids and soluble proteins and polysaccharides with MW 10–50 kD (Warembourg and Billes, 1979; Nguyen, 2009). Rhizodeposition has been studied for centuries, and still is actively discussed in contemporary periodicals. Unfortunately, the data on rates of RE remains controversial. The review by Cheng and Gershenson (2007) lists five approaches to estimate the C-fluxes in the rhizosphere: (i) direct chemical analyses of the hydroponic solutions, (ii) ¹⁴CO₂ pulse-chase, (iii) continuous labeling, (iv) natural isotope abundance, and (v) reporter genes. The conclusion is that each approach has drawbacks and none is fully satisfactory. The impediments arise from the vulnerability and biocomplexity of roots, their irregular architecture, tight binding to the soil, and close association with mycorrhiza and prokaryotic cells. There is also the kinetic handicap of an unknown sink (e.g., the produced RE are immediately consumed by microbial commensals) that we discussed from the very start (see above) and that can be resolved by a combination of experimental studies with mechanistic modeling. Below we demonstrate this strategy as applied to the most popular and the most controversial pulse-chase method.

Resolving interpretive and computational controversies of the ¹⁴CO₂ pulse-chase method

Pulse-chase method (Fig. 3) involves a short-term (down to 20 min) exposure of the aboveground plant to the ¹⁴CO₂, the following 2–4 days the relocation of ¹⁴C is evaluated within the plant and in the rhizosphere. The technique was warmly accepted as a simpler and faster option than continuous labeling. Especially encouraging was the finding that the ¹⁴C that passed through the plant was recovered as a consequence of two phases in ¹⁴CO₂ evolution from the root (Fig. 3F), the first peak was attributed to root respiration (RR), and the second to rhizomicrobial respiration (RMR). The conclusion was supported by the fact that plant sterilization removed the 2nd phase (Warembourg, 1975; Warembourg and Billes, 1979). But soon other researchers (Todorovic et al., 2001) presented a high-quality continuous record of ¹⁴CO₂ demonstrating that plant sterilization does not remove the second peak (Fig. 3I). The case is still not closed, some supporters of the succession concept split the ¹⁴CO₂ output into RR and MR parts irrespective of the peak number (Fig. 3H), it is called *the C-flow model* but it looks more like an empirical curve fitting satisfying the condition that RMR is close to RR ignoring any mechanistic details (Domanski et al., 2001; Kuzyakov and Larionova, 2005).

To resolve the conflict of opinions, we suggest here a minimal kinetic model of the soil-plant system (Fig. 3A–C) that mimics the known mechanisms: fixation of ¹⁴CO₂ in the chloroplasts (Calvin cycle), movement of the ¹⁴C-sap (sucrose and other photoassimilates) downward from leaves to roots through the phloem according to the commonly recognized **pressure flow hypothesis** (Knoblauch et al., 2016), utilization of ¹⁴C sap for the plant cells growth and dark respiration, and, finally, exudation of the ¹⁴C sap near root tips and its consumption by the rhizobacteria controlled by the protozoan grazers. The crucial issue is the spatial distribution of ¹⁴C: should we picture the pulse movement in a plant like a **plug flow** (no mixing of the moving liquid elements across the phloem pipe) or a mixing flow such as a series of stirred vessels after adding a color tracer (Fig. 3C). Truth is somewhere between these two extremes, but we select the second option based on the following reasoning. The arrows (Fig. 3A) indicate possible points of the ¹⁴C entries to the plant's leaves during the labeling step (green arrows), followed by the entries of ¹⁴C-sap into the phloem (blue arrows) and finally the exudation steps (black arrows). All these entries are multiple points evenly distributed across the vertical ordinate. There could be a **random distribution** of ¹⁴C at the microscopic scale but there are no chances for the **macroscopic spatial gradients** of ¹⁴C associated with plug flow which requires PDEs to explore the combined effect of **time and space**. Thus, a pulse labeling protocol produces a cloud of ¹⁴C intermediates in the aboveground (AG) compartment, resulting in the extended entry of ¹⁴C to the belowground (BG) compartment. The BG traffic further increases the randomness of the ¹⁴C distribution because of the irregular geometry of the root's phloem: the long and wide segments provide a fast downward ¹⁴C movement while the short and narrow pipes support mostly a slow lateral movement. The following set of ODEs represents the after-pulse ¹⁴C flows:

Aboveground

$$^{14}\text{CO}_2 \quad \frac{dp_1}{dt} = -q_p \gamma_1, \quad (22)$$

$$^{14}\text{C} - \text{sap, photoassimilates} \quad \frac{ds_1}{dt} = q_p \gamma_1 - q_s \gamma_1 - TR s_1 \quad (23)$$

$$^{14}\text{C} - \text{phytomass (shoots)} \quad \frac{dy_1}{dt} = Y_1 q_s \gamma_1 - k_1 \gamma_1 \quad (24)$$

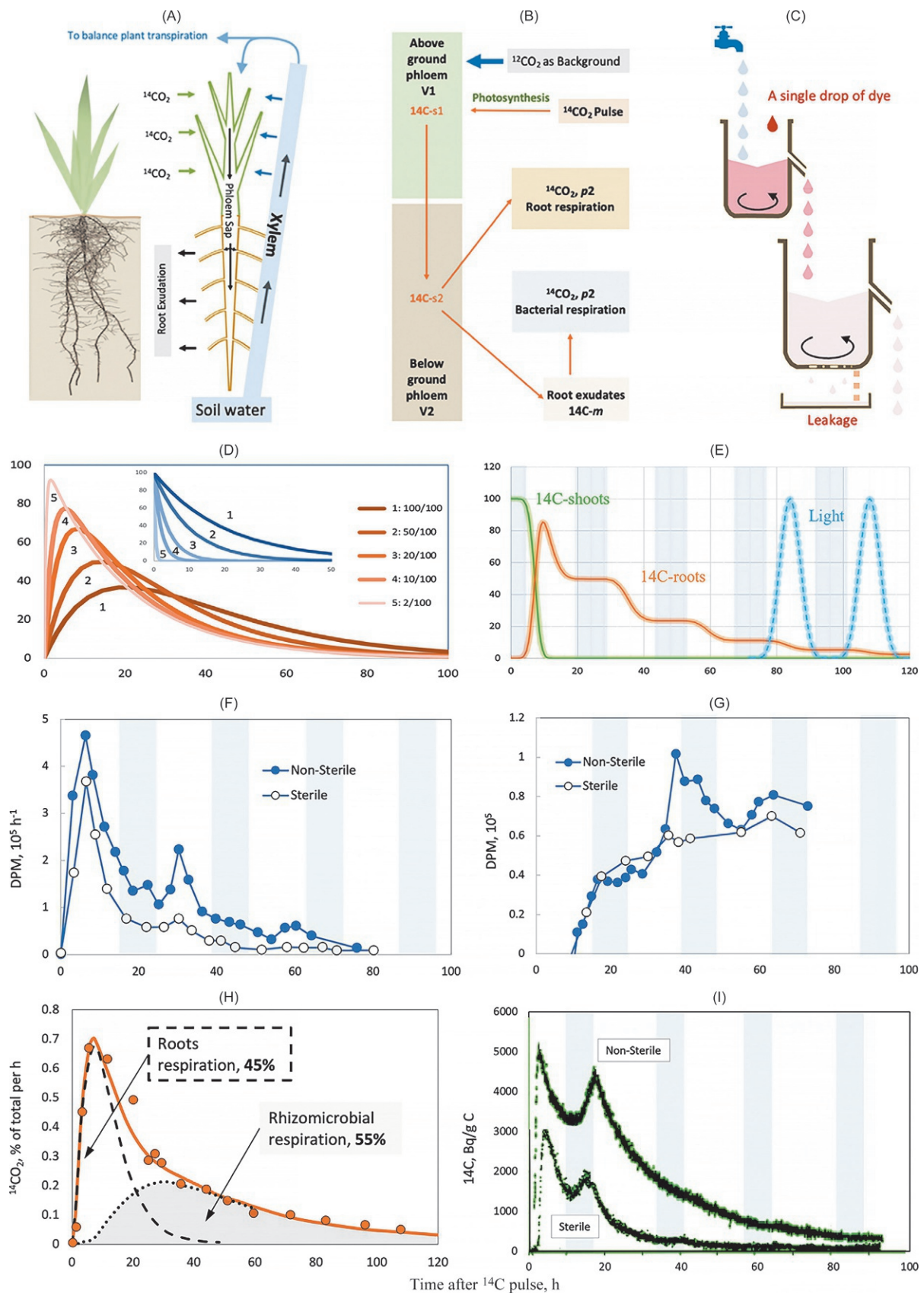


Fig. 3 Kinetic exploration of the pulse-chase technique to quantify root exudation (RE): (A) Cartoon illustrating relationship between RE and water flow based on the Münch hypothesis. (B) Model's flowchart. (C) A simplified analog of the pulse chase: a constant water flow through two vessel one with leaky bottom. A red dye is a tracer analogous to the $^{14}\text{CO}_2$ pulse. (D) Example of the sensitivity analysis: the tracer washout dynamics in the AG (blue curves) and BG compartment (orange) as dependent on their volumes. The V1:V2 ratio is a crucial factor affecting the tracer profile. (E) The effect of diurnal cycles on the washout profile. (F) and (G). Experimental data (Warembourg and Billes, 1979) on $^{14}\text{CO}_2$ release rate (F) and ^{14}C concentration in solution (G). (H) A non-recommended way to differentiate the roots and microbial respiration, data redrawn from (Kuzakov and Larionova, 2005). (I) A continuous high resolution recording of $^{14}\text{CO}_2$ (Todorovic et al., 2001). The shaded rectangles in (E) through (I) panels indicate the dark periods.

Belowground

$$^{14}\text{C} - \text{sap, photoassimilates} \quad \frac{ds_2}{dt} = TR s_1 - q_s y_2 - \psi s_2 \quad (25)$$

$$^{14}\text{C} - \text{phytomass (roots)} \quad \frac{dy_2}{dt} = Y_2 q_s y_2 - k_2 y_2 \quad (26)$$

$$^{14}\text{CO}_2 \quad \frac{dp_2}{dt} = (1 - Y_2) q_s y_2 + (1 - Y') q_x z + (1 - Y) q_m x - D p_2 \quad (27)$$

$$^{14}\text{C} - \text{root exudates} \quad \frac{dm}{dt} = \psi s_2 - q_m x + k_z z \quad (28)$$

$$^{14}\text{C} - \text{Microbial biomass} \quad \frac{dx}{dt} = Y q_m x - q_x z \quad (29)$$

$$^{14}\text{C} - \text{Protozoan biomass} \quad \frac{dz}{dt} = Y' q_x z - k_z z \quad (30)$$

where, q stands for specific uptake rates, subscripts p , s , m , and x indicating uptake of CO_2 , sap, RE, and bacterial cells respectively; Y , Y_1 , Y_2 and Y' are the growth yield of microorganisms, plant shoots, roots, and protozoa respectively; k_1 , k_2 , and k_z are the mortality rates for plant shoots, roots, and protozoa respectively; ψ is the exudation rate.

Fig. 3C pictures a simplified hydrological analog of the plant system for easier understanding of the pulse-chase technique. There are two stirred vessels analogous to AG and BG compartments with continuous water flow mimicking the phloem translocation. At time zero, a red tracer is added to the first vessel as the equivalent of a $^{14}\text{CO}_2$ pulse. Finally, to mimic the root exudation, the second vessel is assumed of having a leaky bottom, the lost liquid being collected in a tub and pumped back to imitate the rhizobacterial uptake (the dotted line). What are the conclusions from this allegory? *First*, it is a reminder that the system is driven by the unlabeled feeding of CO_2 , the decline of the ^{14}C is primarily caused by its washout (= dilution with ^{12}C) in the same way as the color tracer is washed out by the continuously running clear water. It is the ^{12}C flow that causes exponential decay of radioactivity observed in all experiments (Fig. 3F, H, and I) rather than sequential processing through RR and RMR discussed in publications. *Second*, the leakage (an analog of RE) cannot be detected from the washout profile, to assess the leakage, we must record the volume collected in the tub below, therefore recording of the extraradical ^{14}C is absolutely required as suggested by Kuzyakov and Siniakina (2001). *Third*, the only useful information to be extracted from the washout profile, is the turnover rate, TR , an equivalent of the dilution rate in chemostat $TR=D=F/V$. Not too much, but as applied to plants, it allows estimation of the hard-to-measure phloem translocation flux as a product $TR \times \text{sap pool}$. The sap pool is a measurable variable, but not a phloem volumetric flow, so the washout profile provides important information. Further discussion of TR is given below.

Some processes are slow (^{14}C -phytomass) and can be assumed to be constant during the pulse-chase experiment, while others are so fast (respiratory oxidation of ^{14}C -sap) that can be simplified to the steady-state variables to be found from the zeroed ODEs. Between these two extremes is the phloem translocation having the intermediate characteristic time (Eqs. 23 and 25) that defines the pace of the entire process. Fig. 3D shows solutions of these equations simulating the ^{14}C -sap profile as dependent on the AG-to-BG sizes ratio, some of these curves fit the experimentally recorded $^{14}\text{CO}_2$ evolution rates (Fig. 3F, H, and I) indicating high probabilities that the full model would reproduce the pulse-chase process. The next simulation (Fig. 3E) sheds the light on the nature of the observed bimodal shape of the $^{14}\text{CO}_2$ emission. Two or more interruptions are associated with the diurnal cycles, specifically with slowing the phloem traffic in dark periods with the possible additional impact of the plant dark respiration.

What about the RE? Is it possible to estimate the exudation rates based on the pulse-chase principle? Yes, it is possible, and there is a clear signature of RE in the first pulse-chase experiments performed 44 years ago (Warembourg and Billes, 1979). It is a mostly ignored extraradical pool of ^{14}C compounds in the nutrient solution (Fig. 3G). *First*, the release of ^{14}C from roots started after the latent period of 10–12 h required for ^{14}C -sap to reach the root tips, exuded, and metabolized by microorganisms. The published data on the speed of translocation varies (Knoblauch et al., 2016) and the observed delay is within the range. *Second*, the authors overlooked a very important difference in the behavior of native vs sterilized plants, the native system displayed ^{14}C fluctuations as a characteristic pray-predator pattern (Fig. 2). This pattern was undeniably destroyed by sterilization. These two observations confirm that RE does exist and plays an important role in the ^{14}C budget. To convert the extraradical ^{14}C -pool size into the exudation flux, one needs to fit the whole set of Eq. (22)–(30) to the experimental data set. Unfortunately, the paper of Warembourg and Billes (1979) does not contain all the required information.

To summarize, the pulse-chase technique can be fruitfully used for the estimation of RE. The minimum set of required data includes (i) the CO_2 exchange fluxes separately for the AG and BG compartments and (ii) the BG extraradical pool of the soluble organic compounds. The measurements (i) and (ii) must be duplicated for ^{14}C and the total $^{12}\text{C} + ^{14}\text{C}$ isotopes before and after the pulse to estimate the specific radioactivity and express the obtained results using conventional mass units, e.g., the exudation fluxes expressed as $\text{mg C (kg soil h)}^{-1}$ or $\text{g C m}^{-2} \text{ h}^{-1}$. It is more reasonable to express rhizodeposition and RE as a part of the gross PPA because it is the only origin of RE. Units such as %, mmol, or mg C per plant or per g of the root are meaningless. It is also highly desirable: a, to monitor continuously the fluxes using data logging to a computer rather than manually collecting CO_2 with the alkaline trap; b, to have a record of microbial (bacteria, fungi) populations; and c, to have a record of microbial predator (protozoa, nematodes). The protozoa and nematodes are always present in rhizospheres of laboratory-grown plants or in situ. The hydroponic

and aeroponic plant cultures are no exception. It sounds paradoxical, but the account of microbial elimination is an absolutely required element of the realistic mass-balance and kinetic models dealing with the chemistry of the soil-plant systems. Finally, the two-phase succession concept and the associated deconvolution procedure illustrated by Fig. 3 H must be recognized as a hypothesis that failed to pass the experimental verification. It is a normal event in science, nothing personal!

Diversity of soil microhabitats

Principle of habitats classification

Usually, the soil is viewed as a mosaic of diverse microbial habitats representing independent microsites that are unique in many characteristics. They can be drastically dissimilar in the amount and qualities of growth substrates, aeration level, texture, moisture, etc. However, there are just a few general features, which determine the *population dynamics* of microorganisms in any microsite. We list below three such features:

1. **Substrate input** that can be continuous (1) or discontinuous (2). In the first case, the input flux of nutrient compounds can vary over time but remains non-zero because its microbial uptake is compensated by substrate delivery from some external source. The discontinuous pattern implies rare occasional pulses of nutrients without refill that result in short-term explosive growth followed by a period of complete starvation.
2. Grown cells can be **retained** within a habitat (α) or **eliminated** (β). Examples of microbial elimination include cell washout, grazing by protozoa, nematodes, and other soil animals, lysis by parasites (phages, *Bdellovibrio*), and the active migration of microorganisms between different soil habitats. Migration occurs through (i) flagellar motion across the soil water-filled channels (motile bacteria and zoospores), (ii) sliding, gliding, and twitching of bacteria devoid of flagella (*Mycobacterium*, *Myxococcus*, *Flavobacterium*, *Cytophaga*), and (iii) through the vectorized apical growth of soil fungi, actinobacteria, and other mycelial organisms. The last migration style is the most efficient in a heterogeneous soil environment, it can proceed across any soil pore space including those filled with air and organo-mineral gel-like matter rather than only the water-filled voids accessible to the motile cells.
3. Habitats vary in **spatial assembly**, being either (a) homogeneous or (b) heterogeneous. In the first type of habitat, all variables (cells, nutrient substrates, and products) are distributed evenly or randomly across the space (like we found in the pulse-chase labeling of rhizodeposition, see above). In the second type of habitat, we observe the spatial gradients of the biokinetic variables (x , s , p) essential from the biokinetic perspective for the understanding of how microbial growth is controlled. The homogenization factors in arable soils include irrigation, plowing, harrowing and other forms of tillage, the explorative activity of the fossorial soil animals. Soil homogeneity is never as perfect as in the stirred bioreactor or even natural water. Close to true homogeneity can be established through the air-filled soil pores for the volatile and gaseous nutrient substrates.

Three independent ordination axes produce eight possible combinations (Table 2), and two of them are logically forbidden because any kind of spatial gradients of growth substrates implies the *continuous* rather than *discontinuous* delivery of nutrients.

Representing examples

Fig. 4 depicts several characteristic microbial habitats occurring in the soil. The most extensive and extremely oligotrophic soil habitats belong to the category of *dispersion zone* (1a β); this zone embraces the root-free parts of the soil profile and the underlying subsoil that do not receive seasonal input of fresh plant litter or any other plant-derived DOC. Numerous microbial populations inhabiting these habitats grow very slowly on volatile or soluble mobile compounds, which continuously diffuse from richer soil parts. The diffusion flux is extremely small as most of the nutrients are intercepted by other microorganisms residing closer to the source. Examples of such remote sources of mobile compounds are (i) the poorly aerated anoxic soil layers harboring anaerobic microbial communities that produce diverse fermentation products including gases (H_2 , CH_4) and volatile aliphatic compounds (formic, acetic, and propionic acids, alcohols, ketones, and aldehydes); (ii) rhizosphere soil and (iii) the top organic soil layer with fresh plant litter discussed below. Some microorganisms in the dispersion zone are chemolithotrophs using inorganic compounds as energy sources: H_2 , H_2S , S, Fe^{2+} , Mn^{2+} , etc.

Table 2 Matrix of growth patterns in situ and ex situ (Panikov, 1995).

Spatial assembly	Substrate input			
	Continuous (1)		Discontinuous (2)	
	Cell removed (α)	Cell retained (β)	Cell removed (α)	Cell retained (β)
Homogeneous (a)	1a α	1a β	2a α	2a β
Heterogeneous (b)	1b α	1b β	Forbidden	Forbidden

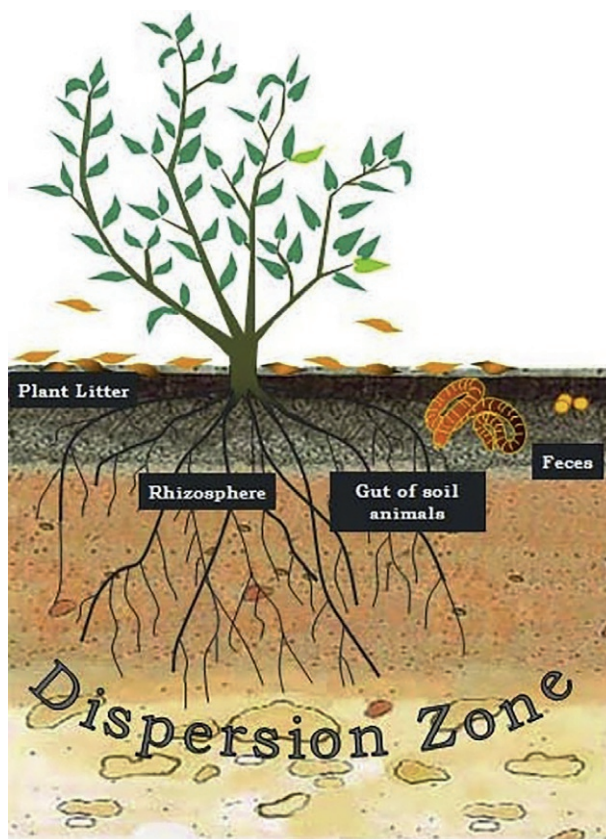


Fig. 4 Some microbial habitats in nature.

The greatest intensity and mostly uninterrupted microbial activity are localized in the topsoil under the *plant litter* layer, *rhizosphere*, and *digestive tract* of soil animals (habitats of the types 1aβ - 1bβ). The C-substrates for microbial populations are the monomeric labile compounds (sugars, aliphatic, and aromatic organic acids) derived from plants either as root exudation in the rhizosphere or released by extracellular hydrolytic enzymes from the lignocellulose, hemicellulose, pectin, and other polymers of dead plant matter. The belowground plant necromass (root litter and root sloughing) provides microbial C-substrates in the rhizosphere. Root exudation is proportional to the PPA and, therefore, follows a diurnal pattern with mid-day maximum and nocturnal decline. The process of plant litter decomposition is more monotonous, being under the control of the activity of the hydrolytic microorganisms and their extracellular enzymes. Spatially, all these habitats are rather heterogeneous; however, the simpler non-distributed models of homogeneous growth kinetic are often justified because of random variation of cell mass and other variables (rather than clearly expressed spatial gradients) combined with the macroscopic scale of experimental data.

Clear vertical gradients are formed in the litter layer from uncolonized fresh plant debris on the top to the highly decomposed sublayer at the interface with mineral soil. At the microscale, the spatial heterogeneity of microbial colonies (mainly fungi and some actinobacteria) is manifested as their differentiation into growing extension zone and non-growing reproductive compartments (Fig. 5). Within the rhizosphere, there are also several spatial gradients of different scales: (a) the vertical gradient of the photosynthate from green aboveground part of plants to the roots, (b) the horizontal gradient between distant trees or tussocks and the bare soil, and (c) the micro-scale gradient around root hair with the maximal concentration of microbial cells and substrates on the root surface (rhizoplane) and exponential decline outwards.

Soil millipedes, isopods, some earthworms, and other soil animals representing the *primary decomposers* inhabit the litter layer and feed on plant debris. The ingested lignocellulose material is mechanically disrupted by the mandible, moistened with saliva, and then passed to the midgut and then to the hindgut. The digestive tract of various soil invertebrates harbors not only specific symbionts but also the normal free-living microorganisms occurring in soil or plant litter, e.g., *Pseudomonas*, *Flavobacterium*, *Vibrio*, *Enterobacter*, *Streptomyces*. The acceleration of their growth in the hindgut is due to favorable conditions: neutral pH, optimal moisture, elevated concentrations of nutrient and growth factors (amino acids, peptides, vitamins) as well as the continuous input of fresh substrate and concomitant removal of digestion products (glucose) to prevent the catabolic repression slowing down biodegradation of plant polymers. Due to peristaltic motion, the content of the gut is mixed and homogeneous. *Secondary decomposers* (i.e., the earthworm *Allolobophora chlorotica*) pass soil mass containing bacterial cells and fungal mycelia through their digestive track; they digest some microorganisms while other resistant species thrive in the gut accelerating their growth.

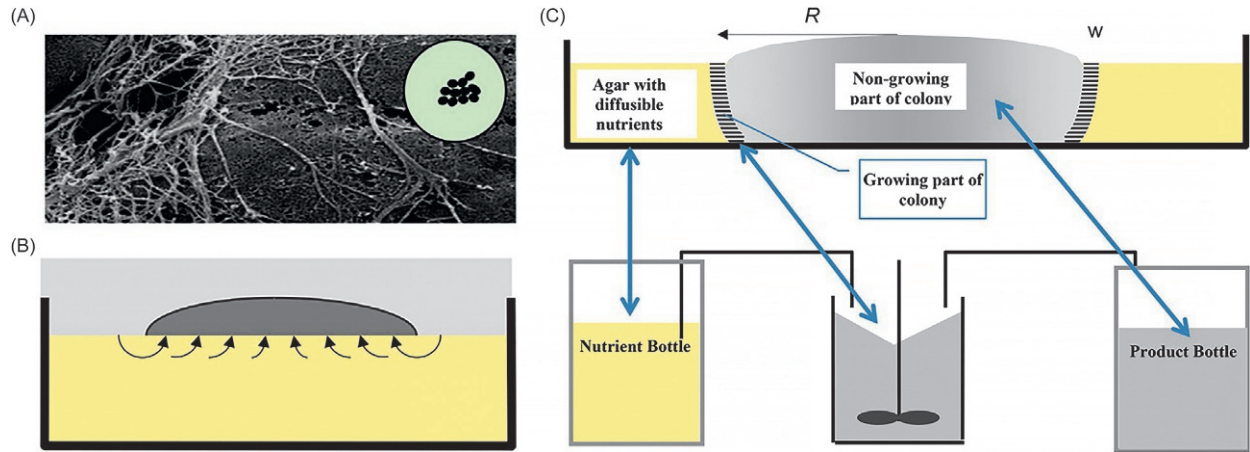


Fig. 5 Mechanisms of colonial growth. (A) microcolonies of bacteria, and fungal mycelium observed in the soil. (B) bacterial growth on the agar surface. (C) fungal growth on agar. Note that bacterial cells colonized the surface of gel consuming nutrients diffusing from the underlying agar layer. The mycelial organisms can spread inside the agar layer and colonize the space faster independent of nutrient diffusion. Under closer examination, we can see a similarity between growing mycelial colony and chemostat. On both systems, the growth pattern is (a) continuous, (b) steady-state, and (c) substrate-limited. The role of the bioreactor in the fungal colony plays the leading front called the peripheral zone of a colony; the non-growing core of the colony is equivalent to the product bottle, while the pump is substituted by the outward growth of the colony invading the non-yet occupied gel volume.

Discontinuous, explosive microbial growth occurs within the decomposition *hot spots* initiated by relatively rare events of input to the soil of the large amounts of easily degradable organic compounds ($2a\alpha - 2a\beta$), such as feces and carcasses of animals, rain washing of mobile organic compounds from plant foliage, manure particles, dry-rewet or freeze-thaw cycles releasing soluble C, etc. Intensive microbial growth tends to attract predators and parasites and therefore is accompanied by the elimination of produced microbial biomass but only after the latent period required for a build-up of microbial biomass.

Growth dynamics in various habitats

Microbial growth expressed as a time series of microbial biomass, x , and the limiting substrate concentration, s , in any type of habitats is described by the following mass balance equations:

$$\begin{array}{cccccc} \text{Net change} & \text{Input} & \text{Loss} & \text{Growth} & \text{Maintenance} & \\ \frac{ds}{dt} & = & F(t) & - & G(s) & - & \frac{\mu x}{Y} & - & mx \end{array} \quad (31)$$

$$\begin{array}{cccccc} \text{Net change} & \text{Immigration} & \text{Elimination} & \text{Growth} & \\ \frac{dx}{dt} & = & V(t) & - & H(x) & + & \mu(s)x \end{array} \quad (32)$$

To specify kinetic expressions, we can rely on kinetic models developed for the laboratory cultivation system by selecting the closest analog to the respective natural habitat (Table 3). For example, the growth of soil bacteria in the intestine of earthworms is

Table 3 Examples of typical natural habitats as their counterparts in laboratory cultivation systems.

Type	Lab cultivation system	Soil habitats		
		Location	Substrate input	Elimination factor
1a α	Chemostat, auxostat	Rhizosphere	Rhizodeposition: exudation and sloughing	Predation by protozoa and nematodes, phagolysis, environmental stresses
1a β	Fed batch culture, dialysis culture	Dispersion zone	Diffusion of gases, volatile and soluble OM	Death from stress factors. Grazing is not operational at low prey density
1b α	Plug-flow culture	Digestion tract of invertebrates	Predigested plant litter	Digestion and washout with feces
1b α	Colonies on agar, solid-phase fermentation	Plant litter	Litter fall, root sloughing	Predation by microarthropods, apical hyphal extension
1b β	Column packed with microbial biofilms	Soil water channels	Trickling solutions	Cell motility
2a β	Simple batch	Temporal decomposition hot-spots without grazers	Occasional release of DOC	Starvation death after DOC depletion

remarkably similar to the tubular (plug-flow) reactor. The rhizosphere community can be viewed as the continuous growth of microorganisms in the chemostat or the phased culture with the fluctuating input of the limiting C-substrate synchronized with diurnal photosynthetic cycle and rhizodeposition (see above review on soil periodic fluctuations). Slow microbial growth in the dispersion zone is adequately simulated by dialysis culture, and the growth of fungi degrading plant litter has many common kinetic features with colonies developing on agar.

Growth models based on the SCM for various homogeneous cultivation systems are shown in Table 4. The listed cultivation protocols reproduce a wide range of microbial physiological states that potentially occur in situ. Fig. 5 illustrates additionally the

Table 4 The major cultivation systems relevant to the purposes of microbial ecology.

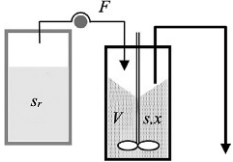
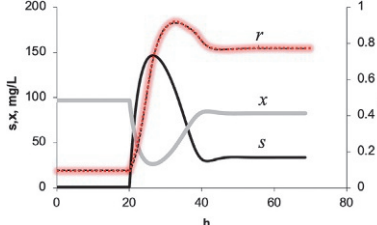
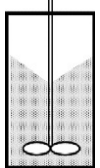
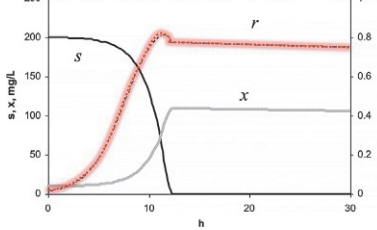
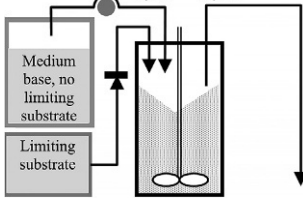
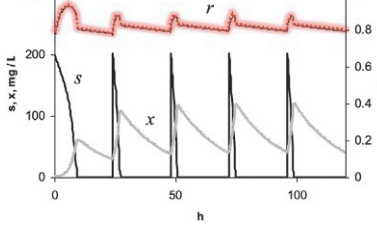
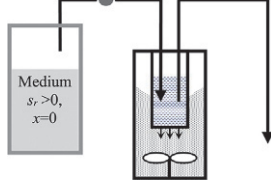
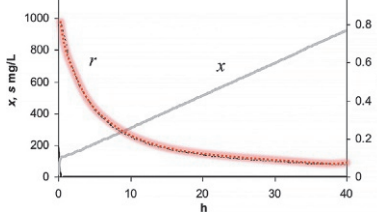
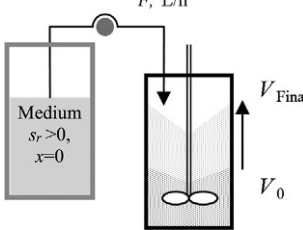
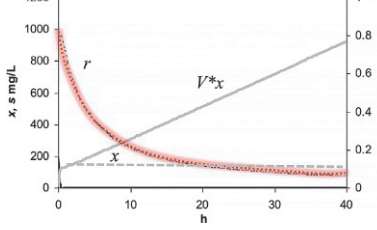
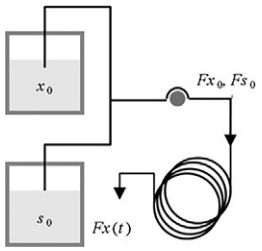
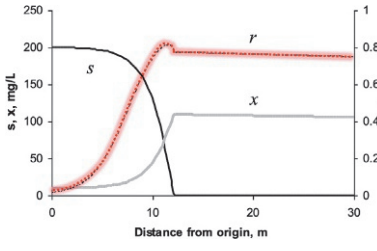
Cultivation	Design	Equations of mass balance	Growth Dynamics
Chemostat		$\frac{ds}{dt} = D(s_r - s) - q_s x$ $\frac{dx}{dt} = \mu x - D x, \mu = Y(q_s - m)$ $\frac{dr}{dt} = \mu \left(\frac{s}{K_r + s} - r \right)$ $F, V, D=F/V, \text{ and } s_r \text{ are constant}$	33 
Batch		As the Eq. (33) at $D = 0$ The initial conditions: at time zero, $x = x_0$ (inoculation dose of microbial cells), and $s = s_0$, initial substrate concentration	
Phased culture (repetitive batch cycles with washout)		As the Eq. (33) with s_r following a square wave pattern (input pulses at regular time intervals): $s_r = \begin{cases} >0, & \text{if } t = \tau n \\ 0, & \text{if } t \neq \tau n, n = 1, 2, \dots, n \end{cases}$ where n is the number of cycles of duration τ h.	
Continuous dialysis, the shortest way to reproduce NZG		There is no cell washout, $-Dx = 0$. The absolute growth rate A is constant, while SGR progressively declines: $\frac{dx}{dt} = A = YF(s_r - s_m) = \mu x;$ $\mu = \frac{A}{x_0 + At}$	34 
Fed-batch culture, another way to reproduce NZG		Parameter D progressively declines: $D = \frac{F}{V_0 + Ft}$ Note the similarity with the dialysis culture: the total cell mass in the bioreactor, Vx , increases linearly over time.	35 

Table 4 (Continued)

Cultivation	Design	Equations of mass balance	Growth Dynamics
A tubular or plug-flow continuous culture that produces steady state batch growth at any phase		$\frac{\partial s}{\partial t} + f \frac{\partial s}{\partial z} = D_s \frac{\partial^2 s}{\partial z^2} - q_s x, f = \frac{F}{A}$ $\frac{\partial x}{\partial t} + f \frac{\partial x}{\partial z} = D_x \frac{\partial^2 x}{\partial z^2} - \mu(s)x \quad [14a]$ where D_s and D_x are diffusion coefficients, A is the cross-section area. Note that the right side of Eq. (27) is the <i>along-the flow-growth rate</i>. At low D_x and D_s, the tubular culture reproduces the batch process unrolled over space, z, instead of time, t	

Based on Panikov NS (1995) *Microbial Growth Kinetics*. Springer Netherlands.

colonial growth, which is predominantly used as a preparatory step (isolation, purity test, etc.) but can be also performed in a precise quantitative à la bioreactor fashion. One of the best criteria of adequacy of the selected cultivation system to mimic the natural state of microbial population is the growth intensity expressed as SGR. If the ex situ and in situ values coincide, then the cultivation technique may be regarded as ecologically adequate.

Diversity of growth characteristics. Life strategy concept

Explanation of the concept

Jacque Monod once wrote that “anything that is true of *E. coli* must be true of elephants, except more so”. He was probably right as far as the basics of molecular biology are concerned, but NOT about behavioral stereotypes. Growth kinetics was mostly studied using *E. coli* and several other model bacteria (*Aerobacter aerogenes*, *Bacillus subtilis*, *Pseudomonas putida*, *Mycobacterium smegmatis*) as well as yeasts (e.g., *Saccharomyces cerevisiae*, *Pichia pastoris*). It is a tiny fraction of microbial diversity but large enough to convince us that the dynamic behavior of different microorganisms is remarkably different even if they use the same or similar spectrum of nutrients. Some grow fast others slowly; some are highly resistant to an adverse environment while others are vulnerable. Surprisingly, the natural biokinetic diversity remains sustainable, slow, and vulnerable populations coexist with fast and resistant individuals rather than being eliminated as dictated by the famous Gause’s law of competitive exclusion. One explanation of the paradox is just *spatial-temporal inhomogeneity* of natural habitats: potential competitors grow as separate colonies in distant loci of soils without strong direct interactions, and temporal (seasonal, weekly, or diurnal) fluctuations further neutralize individual advantages and disadvantages within a diverse community. However, there is another more important and often neglected factor of sustained species diversity: the pressure of natural selection varies under specific circumstances in such a way that competitive advantage not always is given to the fastest or the most resistant species. Fig. 5 illustrates a survival triad, three axes of ordination (could be more) showing the competitive advantage of organisms relative to each other: to invade free ecological space or promptly recover after devastating perturbation (*r*-axis); the ability to compete for limited resources in a stable but overcrowded community (*K*-axis), and survive under unfavorable environmental conditions (*L*-axis). Respectively one can distinguish the following three extreme types of natural selection which correspond to three types of life strategy.

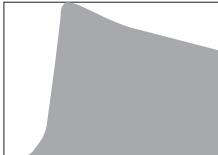
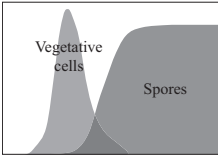
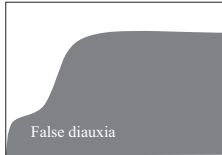
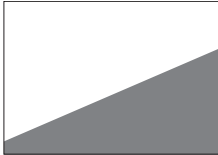
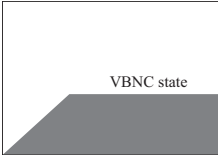
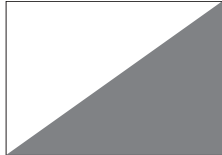
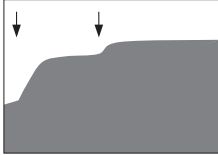
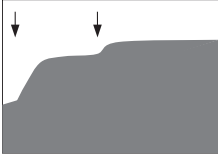
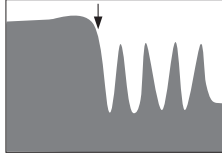
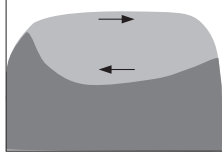
- (1) **r-Selection.** The symbol *r* stands for apparent growth rate, the difference $\mu - a$ between the true growth rate and the mortality rates in the logistic equation (Panikov, 2023, Eq. 17, Part 1) it defines the start of the logistic growth curve. Respectively, *r*-selection operates on the pioneer stages of successions typically caused by perturbations, such as a forest fire that kills most of the organisms in a mature ecosystem. Nutrients are temporary in excess, therefore *r*-selection acts in favor of the opportunistic organisms, the *jackal* type of strategy, a kind of ‘first come, first served.’
- (2) **K-Selection.** The parameter *K* of the logistic equation stands for the carrying capacity of the ecosystem, the maximum biomass that could be attained at available resources. Fluctuations of biomass *x* around *K*, determine the state of mature ecosystems saturated by organisms under stable environmental conditions. Respectively, *K*-selection acts in favor of highly competitive species, the ‘lions’ type of life strategy.
- (3) **L-selection.** The symbol *L* is not shown explicitly in the logistic equation, but it plays the role of a hidden parameter indicating the minimum abundance that provides sustainable existence in a natural community. The condition $x < L$ stands that the population is very close to extinction. The *L*-selection operates under adverse environmental conditions caused by various stresses: starvation, too cold or too hot, extreme pH, desiccation, intoxication, etc. The products of *L*-selection mostly fluctuate around *L*-level unable to withstand competition with stronger *r*- and *K*-selected species under favorable conditions and survive mostly due to raised resistance to stresses. It is the ‘camel’ type of life strategy.

How the survival triangle is constructed?

We start from field observations or lab competition experiments with mixed cultures, in which we follow time series of two and more competitors (up to 10–20, if we use appropriate barcode markers to distinguish individual strains). The second option is much easier and less ambiguous. Different circumstances of natural selection are provided by using respective cultivation methods shown in Table 5. The r-selection is simulated using the exponential phase of simple batch culture (none of the nutrients yet depleted), while K-selection is reproduced in one of the substrate-limited continuous cultivations, e.g., chemostat at reasonable low D , fed-batch, or dialysis cultures. Finally, L-selection can be simulated by exposing growing mixed cultures to continuous or pulsed adverse growth conditions, e.g., suboptimal pH, temperature, oxygen level, etc. The pulsed exposure is performed as a temporal departure from optimum with the following periods of recovery that gives a more realistic pattern of the combined L/r-selection operating in most soils from tropics to tundra. Continuous exposure is a more appropriate choice for lab reconstruction of the permanently pessimal soil environments, e.g., the arid and polar deserts, hot springs, salt marshes, etc.

As a simplified illustration, let us consider competition between three species shown in Fig. 6: *Pseudomonas fluorescens* (P), *Arthrobacter globiformis* (A), and *Caulobacter crescentus* (C). The exponential growth in the batch culture (r-selection), the outcome of the competition was as follows: P (the winner) > A (midway) > C (the weakest competitor). The principal kinetic predictor of competition is the maximum SGR, μ_m : 0.45 (P), 0.25 (A), and 0.15 h^{-1} (C). The same three bacteria have completely different competitive strengths under K-selection reproduced in the mixed glucose-limited dialysis culture: $C > A > P$ with C as an evident winner. The most informative predictors were the substrate saturation constants (K_s 5.0, 1.0, and $0.3 \mu\text{M}$ for P, A, and C respectively) as well as more parsimonious growth stoichiometry of C and A with higher yields and lower maintenance term, m as compared with P. Results of described competition in the batch and dialysis cultures are plotted on axes 1 and 2 respectively, with the distances between P, A, and C proportional to their growth rates, $\mu_i(s)$ observed in experiments. Precautionary note: it is tempting to call C and P ‘bacterial lion’ and ‘jackal’ respectively as permanent labels to be used for the interpretation of numerous descriptive ecological data in the past or expected from future observations. But it would not be perfectly correct because the ordination procedure does make sense only within the analyzed subset of chosen isolates. In other circumstances, P could display itself as a ‘lion’ in the fight with other than A and C competitors. We should also realize the multidimensional nature of the real ecological space with a multitude of factors affecting the outcome of a real-life competition. Having said that, we anyway continue the description of general trends and attempt to summarize the specific biokinetic features of microorganisms associated with their predominant life strategy.

Table 5 Typical growth patterns of soil bacteria associated with their life strategies.

Cultivation conditions	Products of r-selection (<i>Pseudomonas</i> , <i>enterobacteria</i>)	Products of L-selection (<i>Bacillus</i> , <i>Clostridium</i>)	Products of K-selection (<i>Arthrobacter</i> , <i>Caulobacter</i>)
Batch culture from lag-phase to starvation decline			
Dialysis culture: cell mass dynamics			
Chemostat transients: dynamics of cell mass after increase of dilution rate, D ($\downarrow$)			
C-limited chemostat culture: steady-state biomass versus D . Arrow indicates direction of D changes			

Based on Panikov NS (1995) *Microbial Growth Kinetics*. Springer Netherlands.

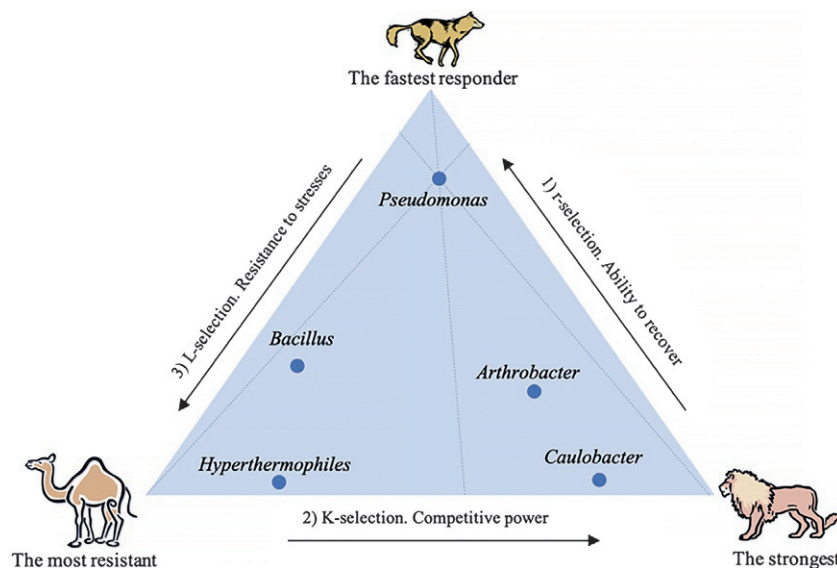


Fig. 6 The concept of life strategies is graphically illustrated using animalistic symbols. Five familiar microorganisms are placed within a triangle space as a crossing point (shown for *Pseudomonas*) of three ordination positions. The position is determined either in real or computer-simulated competition experiments with mixed cultures of microorganisms competing for limited resources under specified environmental conditions, see the text. Based on Panikov NS (1995) *Microbial Growth Kinetics*. Springer Netherlands.

r-Selected organisms

Table 5 summarizes the results of kinetic studies of microorganisms with contrasting degrees of impact of either r, K, or L-selection. In the most general terms, all prokaryotes and single-cell eukaryotes are much more impacted by the r-selection as compared with higher forms of multicellular life. They are simpler organized, reproduce with higher rates, and recover after any perturbation with the speed unthinkable for multicellular Animalia and Plantae. Anyway, some microorganisms carry more features beneficial for the opportunistic lifestyle. Examples are those microorganisms, which have been isolated already in the 19th century and are most thoroughly studied today: enterobacteria including the model #1 *E. coli*, *pseudomonas*, baker, and fodder yeasts, 'saccharolytic fungi,' etc. Their dynamic behavior is erroneously considered to be typical for *all* microbes: rapid and balanced growth, short lag periods, and smooth transients. They dominate in those natural habitats that are frequently 'rejuvenated' to the pioneer succession stage: hot spots of substrate amendment, animal gut and feces, rhizosphere with diurnal fluctuations in exudation rate, and perpetual changes in 'address' of exudation loci due to apical extension of root hair. Exponential batch culture and plating are good simulations of such hot spots; that is why r-selected species appear first on isolation plates or in enrichment culture when we undertake a traditional survey of soils by plating and serial dilutions.

K-Selected organisms

K-selected bacterial species are much less amenable to isolation and cultivation. Probably most of the unculturable microbial species belong to this type of life strategy. The bacterial 'lions' are diverse microorganisms, many of them are called *oligotrophic* because they are found in great numbers in poor habitats, such as oligotrophic lakes or soil dispersion zone. The word *poor* stands for low levels of essential nutrients caused by either (i) aggressive consumption by microbiota or (ii) small supplies from external sources. The first option implies highly competitive pressure for use of nutrient resources rather than poorness (respective ecosystems can be quite productive and flourishing vs. deserts belonging to the 2nd option). The microbial community attains a mature successional stage when all available niches have been already occupied by the winners, i.e., those populations that successfully outcompeted others in a struggle for shared deficient nutrients. Thus, 'bacterial lions' instead of strong muscle, stamina, and deadly jaws, have extremely powerful high-affinity transporters that allow efficient nutrient uptake even from extremely diluted solutions, 10^{-7} – 10^{-10} M. In addition, they are characterized by low maintenance requirements and deliberately unbalanced growth (false diauxia, see below) without strong coupling between uptake and biosynthesis. That is a very valuable quality to keep the leading positions in the oligotrophic overcrowded community. On the other hand, when growing in the laboratory under traditional cultivation conditions, they look rather sluggish, fastidious, and unpredictable. We can picture such behavior as an artefactual response of an otherwise strong competitor to a wrong domestication protocol, a kind of *lion in captivity*. The best option for their cultivation is any continuous culture with cell retention: fed-batch, dialysis culture, or batch bioreactor with the C-supply via gas phase (volatile C-substrates, such as ethanol or VFA). Under these cultivation conditions, we avoid substrate and product

inhibition, and most oligotrophic microorganisms display a high yield and almost 100% viability even at extremely slow cell division (generation time up to months). In other cultivation protocols aimed at intensification of bioprocess, the growth patterns are characterized by: (1) the biphasic growth on the single C-source in a batch culture called the *false diauxia* (the first phase is just a filling of the C-storage compartment), (2) long survival under starvation, (3) oscillations and bistability (two or more steady-states at one dilution rate) in a chemostat at high D, indicating dynamic instability. To mathematically simulate the described abnormal behavior of *Arthrobacter* and other oligotrophic species, the SCM was elaborated to include the ROS type of intermediates with autoinhibition functions and the shortcut pathway of the C-storage (Table 5, the right column).

L-Selected organisms

Many L-selected microorganisms (bacilli, actinomycetes, some fungi) share the following common features: transition to dormancy as endospore under stressful conditions, use of antibiotics to suppress competitors, and a wider range of utilizable nutrients including recalcitrant polymeric compounds (pectin, starch, cellulose, lignocellulose), all of which require an extracellular breakdown by secreted hydrolytic enzymes. All the listed adaptations are combined to compensate for the competitive weakness of the L-selected organisms, their tendency to 'retreat from the battlefield' by switching to lower quality resources and repelling the enemies with poisons. If it does not help, then the last resort is to fully stop the metabolic activity and form endospores. Sometimes such qualities as the use of recalcitrant molecules and antibiotics are erroneously interpreted as a strength rather than a weakness of microorganisms, while sporulation is interpreted as a perfect survival mechanism. We have a convincing counter-argument based on the long-term experiments in dialysis culture (Table 5, the middle column): bacilli and clostridia were shown to be unable to sustain the near-zero growth and when the residual limiting substrate (glucose) concentration drops down to $\sim 10^{-7}$ M their cells entered a non-proliferating state without formation of the normal stress-resistant endospores. Note that the feast-to-famine transition observed in the regular batch culture is fully compatible with the sporulation process. In this case, the explosive growth of bacilli is stopped after depletion of the limiting nutrients (C- or N-source), and the following starvation serves as a trigger to start the multistep sporulation process producing eventually normal stress-resistant endospores able to germinate under adequate growth conditions. However, the deep substrate limitation in dialysis culture is a kind of 'trap' for bacilli and clostridia; they are provoked to start sporulation but are prevented from its completion by a slow non-zero feeding flux of nutrients through the dialysis membrane. The halfway dormancy is a pathological non-replicative state of cells with a low chance of recovery on the full-strength medium. As we discussed above, dialysis culture mimics an oligotrophic soil environment like a dispersion zone. The described phenomenon explains why sporulating bacteria never dominate the oligotrophic microbial communities.

Molecular basics of biokinetic diversity associated with life strategies

Are there any unique genes or specialized molecular mechanisms associated with different variants (r-K-L) of natural selection? Over the last two decades, several research laboratories tried to answer this question (see references under 'Further reading').

We start from the sporulating bacteria, impacted by the L-selection. It seems to be a straightforward task for the modern bioinformatic to distinguish them from the rest of the asporogenic prokaryotes. Classic microbiology textbooks usually refer to the *Bacillus* and *Clostridium* genera as typical spore-forming representatives of the microbial world. Now the list is approaching 100 genera, many of them found in soils. In the best-studied model organism *B. subtilis*, the process of sporulation involves over 500 genes. This list has been generated based on (i) proteomic profiling of endospores, (ii) RNA-seq profiling of gene expression during sporulation, and (iii) genome sequencing of asporogenic mutants. Many genes from the list are not uniquely attributed to the process because sporulation is tightly linked with cell division. Screening of other than *B. subtilis* aerobic spore-formers produced a shorter list of 60 orthologous genes essential for sporulation. Among them are Spo0A (the sporulation master regulator), sspA (acid-soluble spore protein A), dpaA, and dpaB (dipicolinate synthase subunit A and B). Further, it was concluded that firmicutes with genomes smaller than 2300 kb (equivalent to about 2100 protein-coding genes) are not able to produce endospores, therefore genome size can be used as an additional search filter. However, the most extensive recent survey of up to 400 completely sequenced firmicute revealed serious complications (Galperin et al., 2012): some firmicutes lack several essential sporulation genes, while other key sporulation genes including Spo0A were found in microorganisms that do not form endospores, e.g., cyanobacteria, proteobacteria, and spirochaetes. In summary, the bioinformatic approach so far failed to provide an easy way to estimate the abundance of the sporulating bacteria from the metagenomic soil surveys. Specific molecular mechanisms responsible for bacteriostasis of bacilli and clostridia in dialysis culture also remain unclear and are awaiting further investigations.

More genomic data are now available to draw the line between oligotrophic and copiotrophic microorganisms that evolve respectively under impacts of the K- and r-selection. First, a tight correlation was found between bacterial growth rate and the copy number of the rRNA gene involved in the ribosomal synthesis of proteins. Typical r-selected copiotrophic species (*E. coli*, *Vibrio*, *P. putida*, *Burkholderia cepacia*) contain 6–9 copies, while the oligotrophic *Caulobacter*, *M. smegmatis*, and *Sphingomonas* sp. contain 1–2 copies. The largest copy number was found in *Bacillus* and *Clostridium* (10–13 copies) which agrees with a common opinion that these bacteria have evolved as products of a mixed r/L-selection (Klappenbach et al., 2000).

Oligotrophic and copiotrophic microorganisms have a clear distinction in their permeases and other proteins with transporting functions. The balanced growth of copiotrophic microorganisms is at least partly explained by a tight coupling between transport and the following intracellular primary metabolic pathways (the PTS mechanism of glucose uptake). The more efficient growth of oligotrophic bacteria (higher yield, lower maintenance) is explained by the fact that they minimize the number of energy-intensive

ATP-binding cassette (ABC) transporters each having broad substrate specificity and high affinity to the wide range of nutrients (Lauro et al., 2009). The copiotrophs tend to express specialized transporters that involve greater energetic costs but provide more flexibility in transient kinetics minimizing the lag phase.

Life strategies and the problem of growth recalcitrance

Dependent on the method of calculation, from 70 to 99% of soil prokaryotes do not grow in the laboratory media (Stewart, 2012), the phenomenon is called *uncultivability* or *growth recalcitrance*. Several useful tactics have been developed, such as diluted media with soil extracts, antioxidants, growth factors, dilution to extinction, and many more (Stewart, 2012). Yet, the problem is not resolved. A common opinion is that recalcitrant cells grow in situ but microbiologists fail to reproduce their growth requirements in the laboratory. An alternative explanation is that recalcitrant cells belong to the category of VBNC (Viable but nonculturable) cells, a kind of dormancy state of low metabolic activity and zero growth rate caused by adverse environmental conditions. We described (Panikov et al., 2015) the irreversible formation of VBNC as ribosome-deficient cells accumulating in the continuous recycle culture of *P. putida* (Panikov, 2023, Part 1) and *B. subtilis* (Table 5, middle column). However, a more plausible explanation is that most not-yet-culturable microorganisms are the products of the K-selection (Watve et al., 2000). Indeed, the conventional isolation techniques are based on cell cloning, a process deriving a population of cells from a single cell exposed to excess nutrients. In other words, the isolation protocol is an ideal reproduction of the r-selection, or a mixed r/L selection if a single cell is allowed to divide at extreme pH, temperature, osmolarity, etc. A completely different isolation protocol should be implemented for the K-selected species, which evolved to grow in the overcrowded environment at low nutrient concentrations established in the climax communities. These bacteria grow slowly in situ but during the standard isolation, their growth is stopped by a high concentration of catabolic substrates (actual inhibitors are likely to be ROS as by-products). Instead of conventional cell cloning (subcritical dilution and exposure to nutrient excess), the K-strategists must be isolated by using a continuous enriched culture with heavy inoculum and long incubation time sufficient for self-maturation.

Now we move to the last section of this overview to see how the biokinetic diversity linked to life strategy can help to understand decomposition in soil.

Decomposition kinetics

Decomposition is one of the most essential soil processes potentially affecting the Earth's atmosphere, a key point of the Global Change agenda. Not surprisingly, decomposition is presently the subject of diverse interdisciplinary studies involving mathematical simulations aimed at a better mechanistic understanding of the process and the prediction of its dynamics. Below we outline several alternative ways of presenting the biokinetics of substrate decomposition in soil.

Decomposition as self-decay

The simplest assumption is that the rate of decomposition is proportional to the instant content of DOC (s) coming to the soil as plant litter. Such assumption implies the first-order decomposition kinetics introduced to soil science about 80 years ago (Jenny, 1941):

$$\frac{ds}{dt} = -ks \quad (37)$$

The rate constant k is allowed to vary as dependent on climatic conditions (temperature, soil moisture) and soil properties (texture, trophic status, biological activity). The next logical step was to account for the notoriously known heterogeneity of organic matter subjected to decomposition, it was done by presenting the variable s in Eq. (37) as a sum of n components, each one having its own self-decay rate constant:

$$\frac{ds}{dt} = -k_1s_1 - k_2s_2 - \dots - k_ns_n = -\sum_{i=1}^n k_is_i \quad (38)$$

Typically, n varies from 3 to 10 to keep mathematical simulation reasonably compact although some authors suggest a non-reductionistic approach of accounting for the whole array of DOC detectable by modern analytical techniques such as solid-state ^{13}C NMR spectrometry (Mazzoleni et al., 2012). Fig. 7 shows as an example the popular CENTURY ecosystem model (Parton et al., 2001), which stratifies organic carbon into nine pools associated with three subgroups: (i) the aboveground plant litter, (ii) root litter, and (iii) soil organic matter (SOM). Inside each subgroup, there are labile and resistant C-pools as dependent on the lignin-to-N content. According to Eq. (38), the chemical structure of the selected C-pools plays the leading role in the decomposition kinetics rather than the soil organisms, which perform the process. Most probably it stems from the last-century paradigm established for the rare model microorganisms (*E. coli*, *S. cerevisiae*) that respond specifically to single individual C-sources, some compounds are utilized instantly, some slowly, and some left intact. In a similar way, the ecosystem modelers assumed that every soil compound has its own unique degradability potential, and an adequate simulation model should solve two major tasks: (i) to combine compounds with a similar degree of degradability into a few pools or clusters (too many pools make model cumbersome),

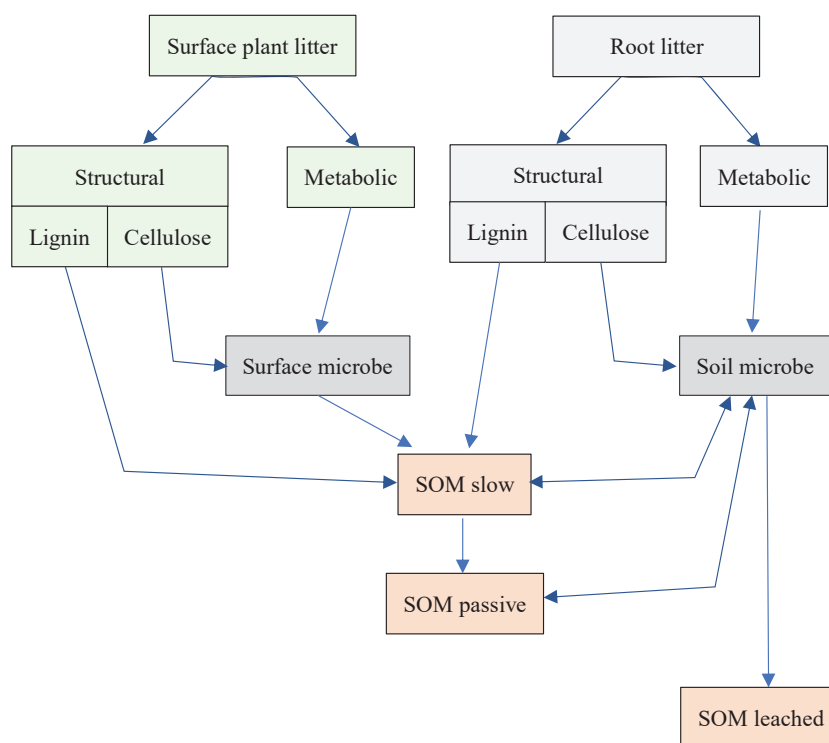


Fig. 7 Soil decomposition simulated by the CENTURY model. The boxes stand for various C-pools that belong to the aboveground plant litter (light green), root litter (light gray), and soil organic matter (light orange). There are two microbial pools, the first one is localized within a surface plant litter (green frame), and the second one spreads between the mineral soil layers. The chart based on Parton B, Ojima D, Del Grosso S, and Keough C (2001) *CENTURY Tutorial. Supplement to CENTURY User's Manual*. NREC Pub. Natural Resource Ecology Laboratory, Colorado State University.

and (ii) to account for the modifying effect of environmental factors, e.g., fluctuating temperature, moisture, as well as specific soil qualities potentially affecting process kinetics, e.g., the clay minerals that bind organic molecules and slow down their degradation.

This paradigm has been gently criticized. For example, Davidson and Janssens (2006) stated that one and the same organic compound can display dramatically different rates of decomposition depending on the soil micro-environment, e.g., some molecules may become physically protected in the interior of soil aggregates. Another more serious deficiency is that Eq. (38) misinterprets the basic mechanisms of soil decomposition ignoring its essential systems qualities:

1. It assumes that SOM degradation proceeds *spontaneously* while in real life the process is driven by soil organisms (microbes and invertebrates), therefore decomposition dynamics must be simulated by equations containing the terms characterizing the community of the degraders, their abundance, and biokinetic parameters that establish feedback to the changeable soil environment.
2. Eq. (38) assumes that decomposable substrate is a mixture of compounds each of which is degraded *independently* and *simultaneously* as if plant litter has been split by magical chromatographic procedure into hundreds of purified individual compounds, each placed into a separate microtube and offered to soil degraders. However, real dead plant material remains consolidated and decomposition proceeds *sequentially* starting from the most accessible part (glucose, glycolytic and TCA intermediates, and other mobile compounds) that usually acts as catabolic repressors preventing depolymerization of even slightly more resistant plants polymers (starch, hemicellulose, cellulose). Lignocellulose is decomposed as a whole; the carbohydrate part is aerobically consumed with the formation of ROS (reactive oxygen species) that prime the oxidative degradation of the lignin. Thus, different C-pools *strongly interact* with each other, and these interactions have a striking effect on the process kinetics.
3. According to Eq. (38), decomposition in soil is a set of one-way degradation reactions converting SOM into waste products (CO_2 and H_2O), however, the real soil process combines the catabolic decay reactions with the biosynthesis of new compounds, such as degrader cells, their extracellular metabolites including macromolecules. Thus, the simulation of decomposition dynamics should include equations of microbial growth under very specific soil conditions of near-zero microbial growth encouraging secondary metabolism, formation of stress-resistant pigments (some resemble humic compounds), biofilms cementing soil particles into stable aggregates, and other multicellular structures essential for a better understanding of mature soil as a habitat and part of nature.

The biosynthetic reactions discussed above are usually referred to as the *transformation* reactions associated with decomposition. Recently, several attempts were made to improve Eq. (38) by making it mathematically more comprehensive and accounting for the combined effect of decay with transformations. It presents the SOM as a *continuum* of degradability using PDE, a differential equation in the partial derivatives (Carpenter, 1981):

$$\frac{\partial f(z, t)}{\partial t} + \frac{\partial V(z, t)}{\partial z} = -U(z, t)f(z, t) \quad (39)$$

Here $f(z, t)$ is the distribution of detritus components with respect to their susceptibility to microbial degradation, z , where z ranges between 0 and the maximum value z_m ; $V(z, t)$ is the amount of detritus converted into more or less available compounds; and $U(z, t)$ is the degradation rate. However, this elegant equation has never been filled with specific biological information and applied to soil decomposition studies. Two subsections below dealing with the decomposition of simple individual compounds and cellulose give an alternative to the self-decay paradigm placing the emphasis on soil organisms responsible for decomposition, and their specific biokinetic features imprinted in their life strategies.

The mean residence times and half-life: the interpretive nuances

Decomposition kinetics is often characterized using half-life ($T_{0.5}$) and the mean residence time, MRT (τ). The half-life is the time required for a given individual compound or C-pool to be decomposed by 50%. The relationship between $T_{0.5}$ and k is found after integration of Eq. (37) at initial conditions, $s = s_0$ at $t = 0$, and $s = 0.5 s_0$ at $t = T_{0.5}$:

$$s = s_0 e^{-kt}, \quad \text{if } t = T_{0.5}, \quad \text{then } 0.5s_0 = s_0 e^{-kT_{0.5}} \quad \text{and} \quad T_{0.5} = \frac{\ln 2}{k} \quad (40)$$

As applied to Eq. (38), the half-life is calculated for each C-pool as $T_{0.5i} = \ln 2/k_i$.

The *mean residence time*, MRT is the average length of time that an entity, in this case, a decomposable i th soil compound will remain in a studied system. The best relevant MRT illustration comes from soil zoology (Fig. 8): the earthworms swallow pieces of plant litter, and the moving food lump undergoes digestion, being eventually released as excrement. Here the MRT is the time taken for a single litter bite to move from the mouth to the anus; MRT can be estimated by feeding the worms with the stained plant material. The rate of decomposition is calculated as (note similarity with Eq. 4):

$$\frac{ds_i}{dt} = \frac{s_i^{\text{in}} - s_i^{\text{out}}}{\tau_i} \quad (41)$$

The soil decomposition does not allow such a straightforward interpretation of MRT, rather it is closer to the variant of MRT called *turnover time* (τ). It is defined as the ratio of the pool size to the flux in or out of the pool; then as applied to Eq. (38), $\tau_i = s_i/k_i$; $\tau_i = 1/k_i$. The expressions for τ_i and $T_{0.5}$ look very similar; they differ just by the constant quotient $\ln 2 \approx 0.693$ although their original premises are completely different. Let us see where the problem lies by looking at the process kinetics obeying Eq. (38). We need to update this equation, first, by adding the source of DOC, F_i and, second, by assuming the steady-state status of the soil system achieved when the input exactly matches the output:

$$\frac{ds_i}{dt} = F_i - k_i s_i = 0 \quad \therefore \quad \bar{s}_i = \frac{F_i}{k_i} = F_i \times \frac{1}{k_i} = F_i \times \tau_i \quad (42)$$

Here $\bar{s}_i$ is the steady-state pool size, which is the product of two terms, the input F_i multiplied by τ_i . It clarifies the meaning of the parameter τ distinct from the half-life $T_{0.5}$: the MRT stands for the time required to rebuild the soil pool from a continuously running source, F , while $T_{0.5}$ is associated only with decay, the negative component of the C-balance. A simple numeric example: let us assume that the content of free amino acids (AA) extracted from field soil varies over the summer season around 10 ppm (mg/kg soil), let it be a steady-state or a near steady-state pool $\bar{s}$. In an independent manipulation experiment, added to the same soil AA was degraded with an average rate of $k = 0.5 \text{ d}^{-1}$. From Eq. (33) we can easily estimate the source strength, $F = 10 \times 0.5 = 5 \text{ ppm/d}$, which otherwise is a very challenging task. The MRT, $\tau = 1/0.5 = 2 \text{ d}$ implies a very high turnover rate, once per 2 days, but does not say anything about the decomposition. In a thinking experiment, if we stop the input ($F=0$), then after $\tau = 2$ days the pool will be depleted down to $\sim 37\%$ and will proceed with the same pace; the 50% declines every $T_{0.5} = 1.39$ days. Thus, we see that τ and $T_{0.5}$ values are not identical although close to each other.

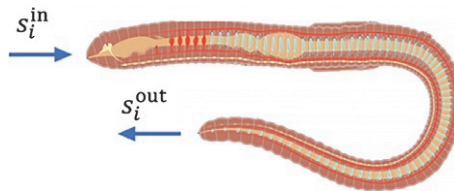


Fig. 8 Digestive tract of earthworm.

In conclusion, the MRT does make sense when applied to the open soil systems having both positive (PP, rhizodeposition, etc.) and negative (decomposition, leaching, crop harvesting) parts of the C balance. In this case, MRT or turnover parameters allow estimation of the true decomposition hidden by soil pool restoration from some non-accounted sources. On the other hand, for soil decomposition under conditions excluding refill (e.g., the litter bags technique), the $T_{0.5}$ is a better option.

Decomposition of simple individual compounds

The realistic kinetic models represent the decomposition process as a result of the growth and activity of decomposer organisms. DOC entering soil are their C-source. Therefore, the changes in biomass must be accounted for in the kinetic expressions. Unfortunately, monitoring the status of the microbial community remains a very challenging procedure. Either classic (direct microscopy, plating, enrichments) or modern molecular techniques (metagenomic surveys, RNA seq, FISH - Fluorescence In Situ Hybridization, SIP - Stable Isotope Probe, etc.) are tedious or expensive and provide mostly single-point data rather than the required dynamic picture with high temporal resolution. The good news is that microbial growth can be *inferred* from the dynamic record of decomposition. In the case of laboratory incubations of soil, test organic compounds are added and the decomposition time course is followed by respirometry (CO_2 evolution rate) or by analyses of the residual organic material. The first technique is especially convenient because it is non-destructive and can be performed automatically in a series of multiplexed soil samples. The principle of reconstruction is as follows. Experimental data on substrate (s) uptake and CO_2 produced (p) are fitted to the SCM or any other advanced growth models able to simulate a self-regulatory behavior of microorganisms. The model describes sources and sinks for three variables, e.g., s , p , and x (cell biomass) using the non-linear regression as a curve-fitting procedure minimizing the simulation error. The missing experimental data on cell mass are constrained by the top and bottom limits and found as a part of inverse problems solution together with the coefficients of the model. Note that this approach is meaningful only in the case of precise lab experiments with soil amended with readily available organic compounds that demonstrate high rates of degradation. These compounds must be added to soil at an amount sufficient to double (at least) the background respiration level.

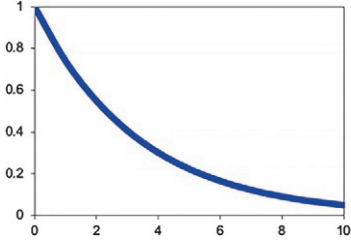
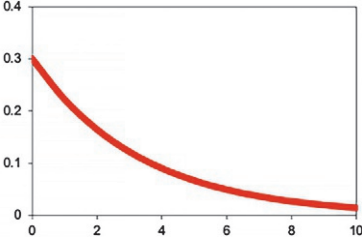
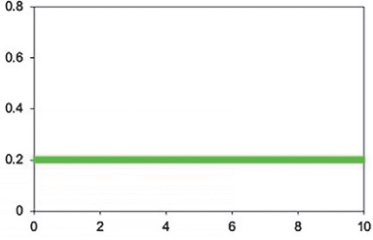
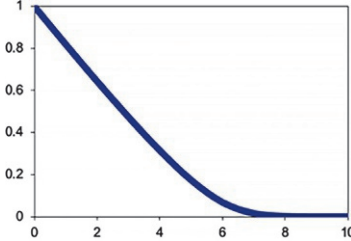
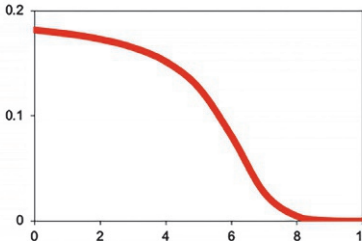
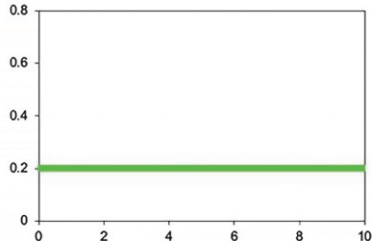
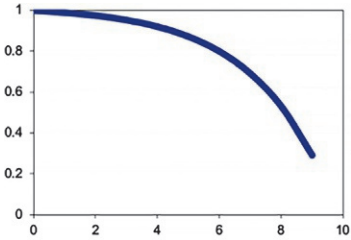
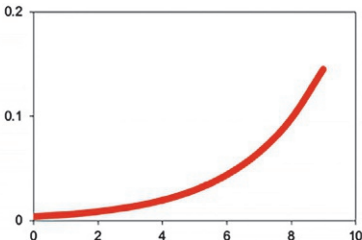
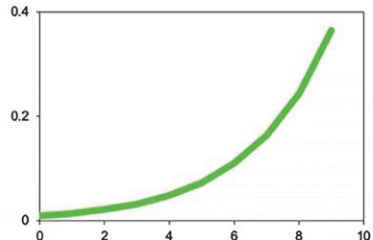
As shown in Table 6, we can confidently identify those rare cases when the decomposition dynamics are not associated with microbial growth, when degradation starts instantly and then progressively declines over time. It could be a spontaneous abiotic self-oxidation of unstable organic compounds that is expected to follow the first-order kinetics (Table 6, row 1), or the degradation process catalyzed by either the extracellular or soil enzymes, or by the non-growing but catabolically active cells (growth arrest is discussed below) that follows the enzymatic Michaelis-Menten kinetics (Table 6, row 2). Microbial growth associated with decomposition is manifested as an exponential increase in the degradation rate (Table 6, rows 3 and 4). The short-term soil incubation experiments limited by the exponential phase of the growth (row 3) do not give a full picture of the decomposition cycle but can be used in practice as a quick indirect method to estimate the biomass of certain physiological groups of soil microorganisms utilizing given organic compounds (Panikov and Sizova, 1996). It is equivalent to the classic SIR (Substrate Induced Respiration) with one essential step forward: the conversion factor from SIR to biomass is found individually for each sample as parameter x_0 of the verified mechanistic SCM model (Panikov, 1995). Besides, there is a bonus of qualitative characterization of the microbial community (parameter r_0), which is a reporter for the macromolecular composition of cells in situ.

As expected, the soil is much more complex than a trivial batch culture. Soil bacteria are still mostly unknown, and some recently discovered species can deviate significantly from the canonic exponential pattern. Notably, the dominating K-selected oligotrophic bacteria can display a *false diauxia* (Table 5, first row). In soil samples taken from the dispersion zone, glucose decomposition shows a characteristic wavy dynamic pattern (Table 6, row 5) with two peaks of glucose uptake, one taking place at the start of incubation and the second in the middle of the process. A remarkably similar pattern was found in a batch culture of *Arthrobacter globiformis* and other K-selected oligotrophic species. The first phase is attributed to filling the storage compartment, while the second one is due to normal exponential growth. This pattern can be one of the major reasons for the *priming effect* phenomenon, usually interpreted as the co-metabolism of recalcitrant indigenous SOM by an added labile compound. Our interpretation states that labile amendment 'wakes up' the half-dormant microbial residents that remain activated for a long time period after depletion of the primer.

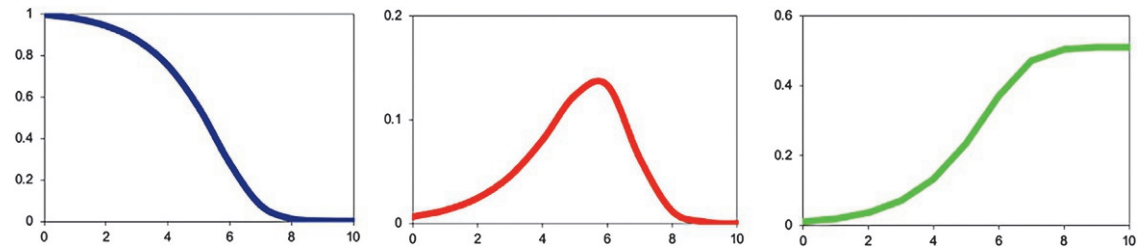
Decomposition of pure cellulose

Significantly more challenging is the decomposition kinetics of the dead structural polymeric plant material, containing lignocellulose, pectin, tannin, waxes, and other recalcitrant OM. Contrary to monomeric substrates, plant polymers are degraded by extracellular enzymes, and the depolymerization products are shared by the rest of the community (the *cheaters* in the modern molecular vocabulary). In addition, structural plant tissues are poor in the biogenic elements (N, P, Mg, Fe...), while the production of extracellular enzymes require extra amounts of these elements for ribosomal protein synthesis. The deficiency of biogenic elements coupled with C-sufficiency provides a high risk of metabolic imbalance and arrest of microbial activity. The non-intentional metabolic arrest called *bacteriostasis* and *fungistasis* is still a mystery in microbial physiology; it can occur without the deliberate application of toxic materials. One example is the 'substrate-accelerated death', severe catabolic repression of cells exposed to a high concentration of catabolite (e.g., glucose) after non-growing conditions, especially caused by deficiency of anabolic substrates (N, P, K...). The natural manifestations of microbial metabolic arrest include accumulated dead organic matter in soils of various degrees, from shallow plant litter layers on the forest floor to multimeter peat deposits in ombrotrophic and

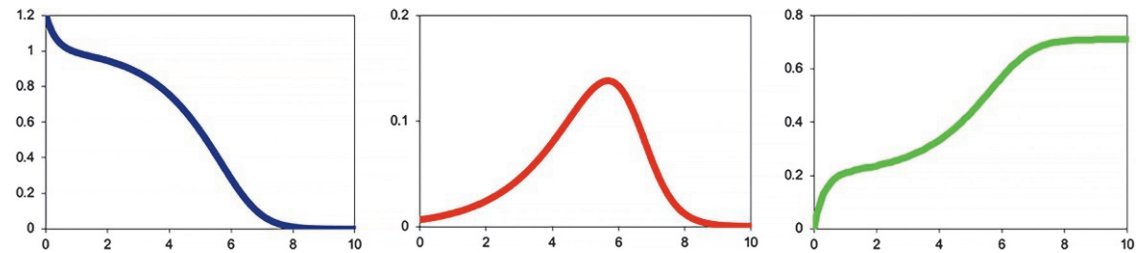
Table 6 Decomposition kinetics of simple individual compounds added to soil, based on (Panikov, 1995).

Condition, model	Residual substrate	CO ₂ production rate	Microbial biomass
1. Negligible effect of microbial growth, the process rate does not depend on microbial biomass. Eq. (37), the first-order kinetics typical for abiotic reactions			
2. Negligible effect of microbial growth, the process does not depend on microbial biomass. The Michaelis-Menten kinetics typical for enzymatic reactions			
3. Degradation process is accompanied by unlimited microbial growth. A simple exponential growth, or Eq. (33) under simplified conditions, $s \gg K_r$			

4. Degradation process is accompanied by substrate-limited microbial growth.
The Monod model, or Eq. (33), the full SCM.



5. Degradation process is accompanied by the growth of K-selected microorganisms.
The customized version of SCM to simulate the false diauxia



Time, h

mesotrophic bogs. A sustainable microbial decomposition under described challenging conditions is made possible only by the cooperative efforts of the entire community.

For the sake of clarity, we illustrate the process complexity using pure ash-free cellulose incubated in forest soil under controlled environmental conditions (Fig. 9). The minimal structure of the decomposer community turned out to contain three microbial groups:

- 1) L-selected hydrolytic populations that produce extracellular cellulases. As we argued in the previous section, switching to low-grade nutrients is a sign of L-selection typically combined with two other qualities, the production of antibiotics and easy transition to metabolic arrest. Since all L-selected species fluctuate around the point of extinction, then the production de novo of extracellular depolymerization enzymes is a vulnerable decomposition step that can be very sensitive to the availability of these organisms and their physiological state.
- 2) r-selected copiotrophic organisms under adequate energy supply capable of explosive growth and mitigating the deficiency of mineral nutrients. Specifically, they can fix atmospheric N_2 under a shortage of N, and mobilize K and P by dissolving insoluble salts through the secretion of chelating exometabolites.
- 3) K-selected oligotrophic populations capable of scavenging monomeric products diffusing from decomposition hot spots. Their remote location makes them less dependent on the deficiency of biogenic elements.
- 4) In addition to the live microbial cellulose degraders, there are free cellulases immobilized on soil particles providing some background level of depolymerization activity.

The chain of events after cellulose is added to soil is as follows: (i) added polymeric substrate induces microbial synthesis of extracellular cellulases releasing mono- and oligosaccharides from polymeric substrate within the decomposition hot-spots; (ii) sugar uptake induces a short-term (ephemeral) microbial growth of the group 1 until one of the nutrients, for example, N

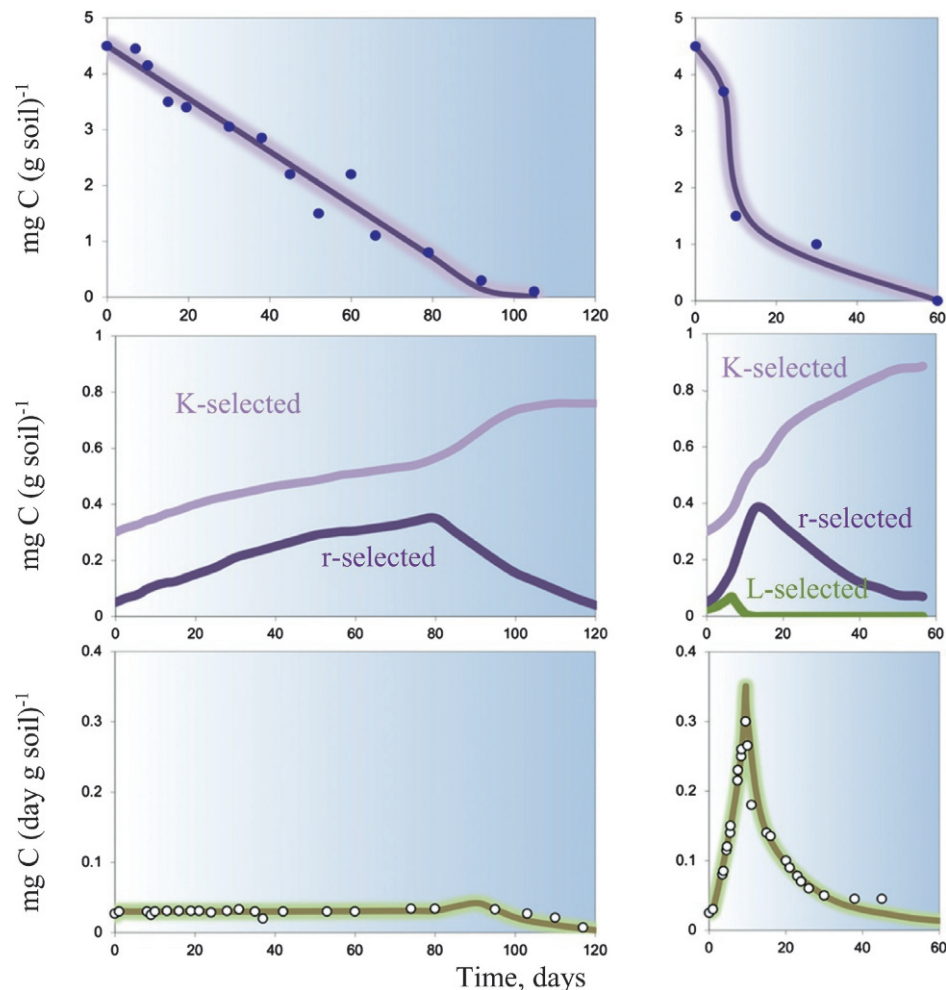


Fig. 9 Cellulose decomposition dynamics. Cellulose (as ash-free filter paper) was added either to the freshly collected forest soil (**left** column) or the same soil after drying and rewetting for activation of the cellulolytic microorganisms (**right** column). The incubation took place at constant moisture and 25 °C with recording cellulose weight loss (**top** panel) and CO₂ evolution rate (**bottom** panel). Microbial community dynamics (**middle** panel) and other curves were simulated by the SCM-based model (Panikov, 1995) assuming the involvement of extracellular hydrolytic enzymes and three indicated microbial groups, see explanations in the text.

becomes depleted; (iii) N-starvation prevents further production of new hydrolytic enzymes, growth of the L-selected organisms is stopped and they enter a dormant stage; (iv) cellulose is still degraded by the accumulated extracellular enzymes, the released oligo- and monosaccharides are accumulated dramatically, increasing the C:N ratio and provoking explosive copiotrophic growth combined with N_2 fixation; (v) microbial group 2 depletes accumulated C-monomers, starvation causes their lysis accompanied by release of NH_4^+ , the N-deficiency progressively evolves into the double N:C and then a pure C-limitation; (vi) group 3 of the K-selected oligotrophs continues to function as a sink for depolymerization products, preventing product inhibition and eventually resulted in the complete degradation of cellulose. Verbal description is translated into a mathematical model which gives fair agreement with observation (Fig. 9).

The inadequacy of the self-decay model (Eq. 38) is evident from the described experiments. In the same soil and under identical hydrothermal conditions, cellulose is decomposed at a dramatically different rate, depending on the state of the decomposer community. Two perturbation factors, soil NPK fertilization, and drying-rewetting 'wake up' the L-selected hydrolytic microorganisms and accelerated the decomposition. The most pronounced effect was achieved by using one dry-rewet cycle before the incubation (compare the left and right parts of Fig. 9). This perturbation is a stressful procedure causing partial damage to microbial cells accompanied by the leakage of cell constituents into the soil solution. Probably, some inhibitory or repressing metabolites responsible for metabolic arrest are also released from their cells during the rehydration cycle giving a chance for the hydrolytic populations to proliferate. It is interesting that healthy native soils have, by default, a repressed state of microorganisms with high hydrolytic activity, although soil still has a background depolymerization activity level due to the presence of soil enzymes (slow zero-order decomposition, Fig. 9, left column). Thus, a restricted soil decomposition looks like a normal physiological state contrary to the pathological state of the NPK-overdosed arable soils where high hydrolytic activity is combined with toxicity caused by the L-selected species, mostly fungi.

Decomposition of plant litter

The last step from cellulose to lignocellulose and the whole plant litter brings the following complications: (a) in lignocellulose the polysaccharide component is made less accessible to microbial enzymes by a combined effect of low ash content, tough texture, and toxicity of the lignin, (b) the higher heterogeneity of plant litter due to the presence of other than cellulose polysaccharides, as well as waxes, and aromatic compounds often considered as inhibitors and (c) a more complex soil community. However, the degradation of cellulose and lignin is carried out jointly; the carbohydrate component is used as an available C-source generating during catabolic reactions the peroxides and other ROS instrumental in the spontaneous oxidative destruction of lignin. Although lignin slows the decay rate, lignocellulose is decomposed as a whole, and pure lignin is not accumulated at late decomposition stages although the lignin-to-N ratio often increases. Therefore, the kinetic mechanisms of decomposition revealed for cellulose alone are also applicable to lignocellulose.

Conclusion and prospects for future developments

How to handle soil complexity?

Complexity increases with organizational levels, and so do process kinetics. In *chemistry*, reaction rates obey the *law of mass action* translated into simple zero-, first-, or second-order kinetics. The *enzymatic* reactions proceed through the dynamic enzyme-substrate complexes, and their rates follow a mixed-order Michaelis-Menten kinetic that opens a multitude of subtle regulations, e.g., reversible inhibition and activation, allosteric effects, cooperativity, etc. At the *cellular level*, the differential gene expression adds a dramatic increase in complexity since the variable environmental conditions (temperature, substrate concentrations, toxicity, etc.) affect the growth rates directly and through the self-adjustable changes of the macromolecular composition of cells. Finally, the *soil community* produces further complexities: (i) multiple species of the same metabolic type interacting through competition or metabiosis, (ii) other species including microbial predators and parasites, (iii) a multifaceted mosaic environment subjected to occasional perturbations and stresses. The conventional microbial kinetics focuses on the first three organizational levels (chemical, enzyme, and microbial growth kinetics). Here we demonstrate several new approaches that allow us to address the soil complexity. *First*, the soil biokinetic diversity was shown to be categorized into a few types (we suggested six), each one corresponding to the unique laboratory cultivation. It reduces soil complexity and allows experimental verification of each kinetic module based on reliable laboratory data. *Second*, the nutrient supply (flux, pattern, spectrum) has been selected as the *primary environmental factor*, and it is obligatory to include the respective variable, the concentration of limiting nutrient. Other factors are optional, and it helps to keep the model compact and manageable. *Third*, introducing the concept of life strategies allowed a justified aggregation of the multitude of potential soil competitors into a limited number of behavioral groups (e.g., L-selected hydrolytic, r-selected copiotrophic, and K-selected oligotrophic microbial clusters in the model of cellulose decomposition). According to Occam's razor principle of parsimony, the number of groups is not formally limited but should be kept at a pragmatic minimum sufficient to explain the observed dynamic pattern or its self-regulation. *Fourth*, similar aggregation is made for microbivore soil organisms (or parasites) accounting for the biological elimination. It is included in most cases; the only exceptions are the dispersion zone (too sparse prey density provides immunity) and the short-term soil incubations (within a predator's lag phase). *Fifth*, biokinetic studies benefit from combining field observations with laboratory experiments. Specifically, soil

incubations with added substrates and soil-plant microcosms reduce complexity by excluding stochastic events and allow a thorough recording of transient regimes, e.g., perturbations, and ecological successions with required precision and accuracy.

Still, what is the rate of microbial growth in the soil?

Eight different approaches for estimating microbial growth rates in nature have been reviewed. Some were included just for the sake of historical accuracy. The most practical for soil studies are the non-destructive methods using gaseous substrates and products, e.g., heterotrophic $^{14}\text{CO}_2$ fixation under isotope incorporation techniques and CO_2 formation under soil respiration. However, it would be wrong to have the SGR assessment as the only research goal. The first four approaches discussed dealt with simple communities in a defined environment. But most of the soil-based ecosystems display remarkably wide SGR variation over time and across the community. Therefore, a single average SGR is a worthless characteristic apart from compact microbial communities at near steady state and having a narrow SGR distribution. Examples are long-term fallow soil (vegetation-free plots dominated by K-selected slow-growers) and the invertebrate gut microbiome (dominated by r-selected fast-growers). Otherwise, a more reasonable goal would be to identify all essential sources and sinks for microbial growth, record the process, and fit the data to the growth model, the SGR dynamics will be found as a part of the model's outputs. We used this way when discussed soil fluctuations (Fig. 2) and found that growth rate cannot be constant, instead, the SGR fluctuates between the high and low limits, the upper limit is attained once per week, that day a rhizomicrobial population completes a full division cycle, the following 5–7 days bacteria grow slower making another 2 or 3 divisions. It is very fast growth; similar growth intensity can be detected in the gut of animals and in hotspots of decomposition. Soil microorganisms located in the dispersion zone grow much slower, displaying long generation times extended to few weeks or months. Soil fungi responsible for decomposition of plant litter grow with SGR between these two extremes.

What did we learn from kinetic studies?

The most beneficial feature of biokinetics is that it helps to organize empirical data (field observations, experimental results) into the non-contradictory concept and respective mathematical model. Any inadequate assumptions are discarded during the model's construction based on the mismatch between observation and simulation. This chapter gives several examples of incremental progress in the understanding of soil biodynamics: periodic soil fluctuations, the priming effect, ecological successions accompanying cellulose decomposition, the mechanisms of microbiostasis, and the uncultivability phenomenon. Application of the SCM model to short-term soil incubations with added substrates allowed the responder biomass estimation without any empirical conversion factors. In the attempt to correct the absurd simulation results (too low and negative average SGR, described under Seasonal decomposition of organic matter), we detected the published misinterpretations of the physiological state of soil biota (maintenance, cryptic growth) and revised a theoretical basis of the chase-pulse method designed for assessment of root exudation (RE), the most significant C-source for microbial decomposers in soil. We treated RE as an intermediate of the plant production process appended with a microbial loop (Eqs. 22–30). It is a standard modeling approach in biokinetics to describe the reactant A conversion to the product P via the intermediate I. The main obstacle in the past was that A (PPA and phloem translocation) is beyond the traditional scope of soil biology. The 'moral of the story' is that a full understanding of soil processes cannot be achieved without touching other parts of the ecosystem, we need to look inside the green aboveground parts of plants as well.

What we missed

Sure, it is impossible to embrace all the areas potentially accessible using the biokinetic approach. We missed the fungi and actinomycetes as mycelial forms of microbial life growing as a network of branched hyphae rather than unicellular bacteria. Another important missed point is the symbiosis of prokaryotes (Rhizobium and Frankia) and mycorrhiza with plants, which has tremendous theoretical significance and great biotechnological potential. Other missed points include (although not limited to) the phages, Bdellovibrio, and microarthropods as selective elimination factors of the growing microorganisms, plate count as a kinetic tool to characterize the microbial physiological state, the quantitative microbial morphology helping extraction of useful information from direct soil microscopy, beneficial spontaneous mutations as one of the mechanisms affecting microbial dynamics and other topics. On the practical side, biokinetic knowledge is especially important and helpful for the optimization of various bioengineering projects aimed at soil bioremediation from pollutants, and for the development of sustainable 'green' agriculture.

References

- Aristovskaya T (1974) The regularities of bacteria biomass reproduction in the soils of different geographical zones. In: *First International Congress of Ecology 'Structure, Functioning and Management of Ecosystems*. The Netherlands: The Hague.
- Brock TD (1971) Microbial growth rates in nature. *Bacteriological Reviews* 35(1): 39–58.
- Carpenter SR (1981) Decay of heterogeneous detritus: A general model. *Journal of Theoretical Biology* 89(4): 539–547.
- Cheng W and Gershenson A (2007) Carbon fluxes in the rhizosphere. In: *The Rhizosphere*, pp. 31–56. Elsevier.
- Curds CR and Cockburn A (1968) Studies on the growth and feeding of *Tetrahymena pyriformis* in axenic and monoxenic culture. *Microbiology* 54(3): 343–358.

- Cutler DW, Crump LM, and Sandon HA (1923) A quantitative investigation of the bacterial and protozoan population of the soil with an account of the protozoan fauna. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character* 211: 317–350.
- Davidson EA and Janssens IA (2006) Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* 440(7081): 165–173.
- Domanski G, Kuzakov Y, Siniakina SV, and Stahr K (2001) Carbon flows in the rhizosphere of ryegrass (*Lolium perenne*). *Journal of Plant Nutrition and Soil Science* 164(4): 381–387.
- Galperin MY, Mekhedov SL, Puigbo P, Smirnov S, Wolf YI, and Rigden DJ (2012) Genomic determinants of sporulation in Bacilli and Clostridia: Towards the minimal set of sporulation-specific genes. *Environmental Microbiology* 14(11): 2870–2890.
- Hagström U, Larsson PH, and Normark S (1979) Frequency of dividing cells, a new approach to the determination of bacterial growth rates in aquatic environments. *Applied and Environmental Microbiology* 37(5): 805–812.
- Jenny H (1941) *Factors of Soil Formation: A System of Quantitative Pedology*. New York: McGraw-Hill.
- Karl DM (1980) Cellular nucleotide measurements and applications in microbial ecology. *Microbiological Reviews* 44(4): 739–796.
- Klappenbach JA, Dunbar JM, and Schmidt TM (2000) rRNA operon copy number reflects ecological strategies of bacteria. *Applied and Environmental Microbiology* 66(4): 1328–1333.
- Knoblauch M, Knoblauch J, Mullendore DL, Savage JA, Babst BA, Beecher SD, Dodgen AC, Jensen KH, and Holbrook NM (2016) Testing the Munch hypothesis of long distance phloem transport in plants. *eLife* 5.
- Kuzakov Y and Larionova AA (2005) Root and rhizomicrobial respiration: A review of approaches to estimate respiration by autotrophic and heterotrophic organisms in soil. *Journal of Plant Nutrition and Soil Science* 168(4): 503–520.
- Kuzakov Y and Siniakina SV (2001) A novel method for separating root-derived organic compounds from root respiration in non-sterilized soils. *Journal of Plant Nutrition and Soil Science* 164(5): 511–517.
- Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S, DeMaere MZ, Ting L, Ertan H, Johnson J, Ferriera S, Lapidus A, Anderson I, Kyripides N, Munk AC, Detter C, Han CS, Brown MV, Robb FT, Kjelleberg S, and Cavicchioli R (2009) The genomic basis of trophic strategy in marine bacteria. *Proceedings of the National Academy of Sciences* 106(37): 15527–15533.
- Mazzoleni S, Bonanomi G, Giannino F, Incerti G, Piermatteo D, Spaccini R, and Piccolo A (2012) New modeling approach to describe and predict carbon sequestration dynamics in agricultural soils. In: Piccolo A (ed.) *Carbon Sequestration in Agricultural Soils: A Multidisciplinary Approach to Innovative Methods*, pp. 291–307. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Meynell G (1959) Use of superinfecting phage for estimating the division rate of lysogenic bacteria in infected animals. *Microbiology* 21(2): 421–437.
- Mullen AD, Treibitz T, Roberts PLD, Kelly ELA, Horvitz R, Smith JE, and Jaffe JS (2016) Underwater microscopy for in situ studies of benthic ecosystems. *Nature Communications* 7(1): 12093.
- Newman E and Watson A (1977) Microbial abundance in the rhizosphere: A computer model. *Plant and Soil* 48: 17–56.
- Nguyen C (2009) Rhizodeposition of organic C by plant: Mechanisms and controls. *Sustainable Agriculture* 97–123.
- Panikov NS (1995) *Microbial Growth Kinetics*. Netherlands: Springer.
- Panikov NS (2015) Near-zero growth kinetics of *Pseudomonas putida* deduced from the proteomic analysis. *Environmental Microbiology* 17(1): 215–228.
- Panikov NS and Gorbenko AY (1992) The dynamics of gas exchange between soil and atmosphere in relation to plant-microbe interactions: fluxes measuring and modelling. *Ecological Bulletins* 42: 53–61.
- Panikov NS and Sizova MV (1996) A kinetic method for estimating the biomass of microbial functional groups in soil. *Journal of Microbiological Methods* 24(3): 219–230.
- Panikov NS, Mastepanov MA, and Christensen TR (2007) Membrane probe array: Technique development and observation of CO₂ and CH₄ diurnal oscillations in peat profile. *Soil Biology and Biochemistry* 39(7): 1712–1723.
- Parinkina OM (1973) Determination of Bacterial Growth Rates in Tundra Soils. *Bulletins from the Ecological Research Committee* 17: 303–309.
- Parton B, Ojima D, Del Grosso S, and Keough C (2001) *CENTURY Tutorial. Supplement to CENTURY User's Manual*. NREC Pub. Natural Resource Ecology Laboratory, Colorado State University.
- Payne WJ (1970) Energy yields and growth of heterotrophs. *Annual Reviews in Microbiology* 24(1): 17–52.
- Perfil'ev BV and Gabe DR (1969) *Capillary methods of investigating micro-organisms*. Oliver and Boyd: Edinburgh.
- Romanenko VI (1964) Heterotrophic CO₂ assimilation by aquatic bacterial flora. *Mikrobiologiya* 33: 679–683.
- Saunders PT and Bazin MJ (1973) Nonsteady state studies of nitrification in soil: Theoretical considerations. *Soil Biology and Biochemistry* 5(5): 545–557.
- Smith K (1980) A model of the extent of anaerobic zones in aggregated soils, and its potential application to estimates of denitrification. *Journal of Soil Science* 31(2): 263–277.
- Staley JT (1971) Growth rates of algae determined in situ using an immersed microscope. *Journal of Phycology* 7(1): 13–17.
- Stewart EJ (2012) Growing unculturable bacteria. *Journal of Bacteriology* 194(16): 4151–4160.
- Todorovic C, Nguyen C, Robin C, and Guckert A (2001) Root and microbial involvement in the kinetics of 14C-partitioning to rhizosphere respiration after a pulse labelling of maize assimilates. *Plant and Soil* 228(2): 179–189.
- von Savigny C, Peters DHW, and Entzian G (2019) Solar 27-day signatures in standard phase height measurements above central Europe. *Atmospheric Chemistry and Physics* 19(3): 2079–2093.
- Warembourg F (1975) Application de techniques radioisotopiques a l'etude de l'activite biologique dans la rhizosphere des plantes. *Revue d'Ecologie et de Biologie du Sol* 12(1): 261–272.
- Warembourg F and Billes G (1979) Estimating carbon transfers in the plant rhizosphere. In: *The Soil–Root Interface*, pp. 183–196. Elsevier.
- Watte M, Shejval V, Sonawane C, Rahalkar M, Matapurkar A, Shouche Y, Patole M, Phadnis N, Champhenkar A, Damle K, Karandikar S, Kshirsagar V, and Jog M (2000) The 'K' selected oligophilic bacteria: A key to uncultured diversity? *Current Science* 78(12): 1535–1542.

Further reading

Growth in-situ versus laboratory cultivation

- Brock TD (1971) Microbial growth rates in nature. *Bacteriological Reviews* 35(1): 39–58.
- Gause GF (1934) *The Struggle for Existence*. Baltimore: Williams and Wilkins. (Reprint, 1964), 163 p.
- Hoehler TM and Jorgensen BB (2013) Microbial life under extreme energy limitation. *Nature Reviews. Microbiology* 11(2): 83–94.
- Lee ZM and Schmidt TM (2014) Bacterial growth efficiency varies in soils under different land management practices. *Soil Biology and Biochemistry* 69: 282–290.
- Lynch JM (1990) *The Rhizosphere*. Chichester: Wiley.
- Panikov NS, et al. (2015) Near-zero growth kinetics of *Pseudomonas putida* deduced from the proteomic analysis. *Environmental Microbiology* 17(1): 215–228.
- Panikov NS (1995) *Microbial Growth Kinetics*. p. 378. Netherlands: Springer.
- Paul EA and Clark FE (eds.) (1996) *Soil Microbiology and Biochemistry*. San Diego: Academic Press.
- Resat H, Bailey V, McCue L, and Konopka A (2012) Modeling microbial dynamics in heterogeneous environments: Growth on soil carbon sources. *Microbial Ecology* 63(4): 883–887.

Life strategies, general principles, and molecular-genetic approaches

- Andrews JH and Harris RF (1986) r- and K-selection and microbial ecology. *Advances in Microbial Ecology* 99–147.

- Frioux C, Singh D, Korcsmaros T, and Hildebrand F (2020) From bag-of-genes to bag-of-genomes: Metabolic modelling of communities in the era of metagenome-assembled genomes. *Computational and Structural Biotechnology Journal* 18: 1722–1734.
- Gerson U and Chet I (1981) Are allochthonous and autochthonous soil microorganisms *r*- and *K*-selected? *Revue d'Ecologie et de Biologie du Sol*. 18(3): 285–289.
- Jansson JK and Hofmockel KS (2018) The soil microbiome — from metagenomics to metaphenomics. *Current Opinion in Microbiology* 43: 162–168.
- Klappenbach JA, Dunbar JM, and Schmidt TM (2000) rRNA operon copy number reflects ecological strategies of bacteria. *Applied and Environmental Microbiology* 66(4): 1328–1333.
- Lauro FM, et al. (2009) The genomic basis of trophic strategy in marine bacteria. *Proceedings of the National Academy of Sciences* 106(37): 15527–15533.
- Malik AA, et al. (2020) Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. *The ISME Journal* 14(1): 1–9.
- Niewerth H, et al. (2012) Complete genome sequence and metabolic potential of the quinaldine-degrading bacterium *Arthrobacter* sp. Rue61a. *BMC Genomics* 13: 534.
- Ortiz-Álvarez R, et al. (2018) Consistent changes in the taxonomic structure and functional attributes of bacterial communities during primary succession. *ISME Journal* 12: 1658–1667.

Soil decomposition

- Carpenter SR (1981) Decay of heterogeneous detritus: A general model. *Journal of Theoretical Biology* 89(4): 539–547.
- Davidson EA and Janssens IA (2006) Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* 440(7081): 165–173.
- Hunt HW (1977) A simulation model for decomposition in grasslands. *Ecology* 58(3): 469–484.
- Jenkinson DS (1990) The turnover of organic carbon and nitrogen in soil. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 329(1255): 361–368.
- Komarov AS (2008) Use of mathematical models for assessing the pool and dynamics of carbon in forest soils. *Eurasian Soil Science* 41(13): 1387–1397.
- Parton B, Ojima D, Del Grosso S, and Keough C (2001) *CENTURY Tutorial. Supplement to CENTURY User's Manual*. NREC Pub. Natural Resource Ecology Laboratory, Colorado State University.
- Parton WJ, et al. (1993) Observations and modeling of biomass and soil organic matter dynamics for the grassland biome worldwide. *Global Biogeochemical Cycles* 7(4): 785–809.

Carbon cycle in soils: Dynamics and management

Charles W Rice, Carlos B Pires, and Marcos VM Sarto, Kansas State University, Manhattan, KS, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of C.W. Rice, CARBON CYCLE IN SOILS | Dynamics and Management, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 164–170, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00183-1>.

Introduction	219
Fluxes	220
Soil-forming factors	220
Functions	221
Regulation of soil organic carbon dynamics	223
Composition of soil organic carbon	223
Management	224
Conclusion	226
References	226
Further reading	227

Abstract

Globally, the amount of soil organic carbon (SOC) is more than twice the carbon in the atmosphere or living vegetation. SOC is an important indicator of the condition of a soil because it influences almost all soil biological, chemical, and physical properties and processes. Loss of SOC, accelerated by soil erosion and disturbance, are affected by C management. This chapter explores the carbon cycle, soil-forming factors, composition and functions of soil carbon and the management practices affecting the levels of soil carbon.

Key points

- Soils contain several important C pools and play an essential role in the global C cycle.
- Development of conservation agriculture practices have increased SOC.

Introduction

Carbon is the fundamental building block of all life on Earth. It is present in the atmosphere, plant and animal life, nonliving organic matter, fossil fuels, rocks, and is dissolved in oceans. Carbon is the sixth-most abundant element in the universe, after hydrogen, helium, oxygen, neon, and nitrogen. Concerns about increasing carbon dioxide levels in the atmosphere have resulted in greater public and scientific interest in the global carbon cycle. Since the mid to late 1800s, fossil fuel use, expansion of cultivated agriculture, and forest clearing have led to an increase in atmospheric CO₂ from 260 ppm to current levels of approximately 410 ppm (Smith et al., 2014). Most of the recent increase in atmospheric CO₂ is attributed to the burning of fossil fuels. Current levels of CO₂ are higher now than in the past 1000 years. The increase in CO₂ is of concern because CO₂ is one of the three primary greenhouse gases, the other two being methane (CH₄) and nitrous oxide (N₂O) (Paustian et al., 2019). These three gases retain heat that normally radiates from the Earth's surface. The concern is that elevated levels of CO₂ could increase the heat retained in the atmosphere, thus leading to global warming. Observations have recorded a 0.6 °C increase in the global temperature since 1900 (NOAA, 2014, Smith et al., 2014).

The basic carbon cycle of life is: (1) the conversion of atmospheric carbon dioxide to carbohydrates by photosynthesis in plants; (2) the consumption and oxidation of these carbohydrates by animals and microorganisms to produce carbon dioxide and other products; and (3) the return of carbon dioxide to the atmosphere (Fig. 1). On a global level, the carbon cycle is more complex and involves carbon stored in fossil fuels, soils, oceans, and rocks.

Soils contain several important C pools and play an essential role in the global C cycle. We can organize all the carbon on Earth into five main pools, listed in order of the size of the pool:

1. Lithosphere (Earth's crust): This consists of fossil fuels and sedimentary rock deposits such as limestone, dolomite, and chalk. This is the largest carbon pool on Earth. The amount of carbon in the lithosphere is 66–100 million Pg (1 Pg = 10¹⁵ g). Of this amount, 4000 Pg consists of fossil fuels.
2. Oceans: Ocean waters contain dissolved carbon dioxide and calcium carbonate in marine organisms. The amount of carbon in oceans is 38,000–40,000 Pg.
3. Soil organic matter: The amount of carbon stored in soil organic matter is 2500 Pg (Lal, 2004);

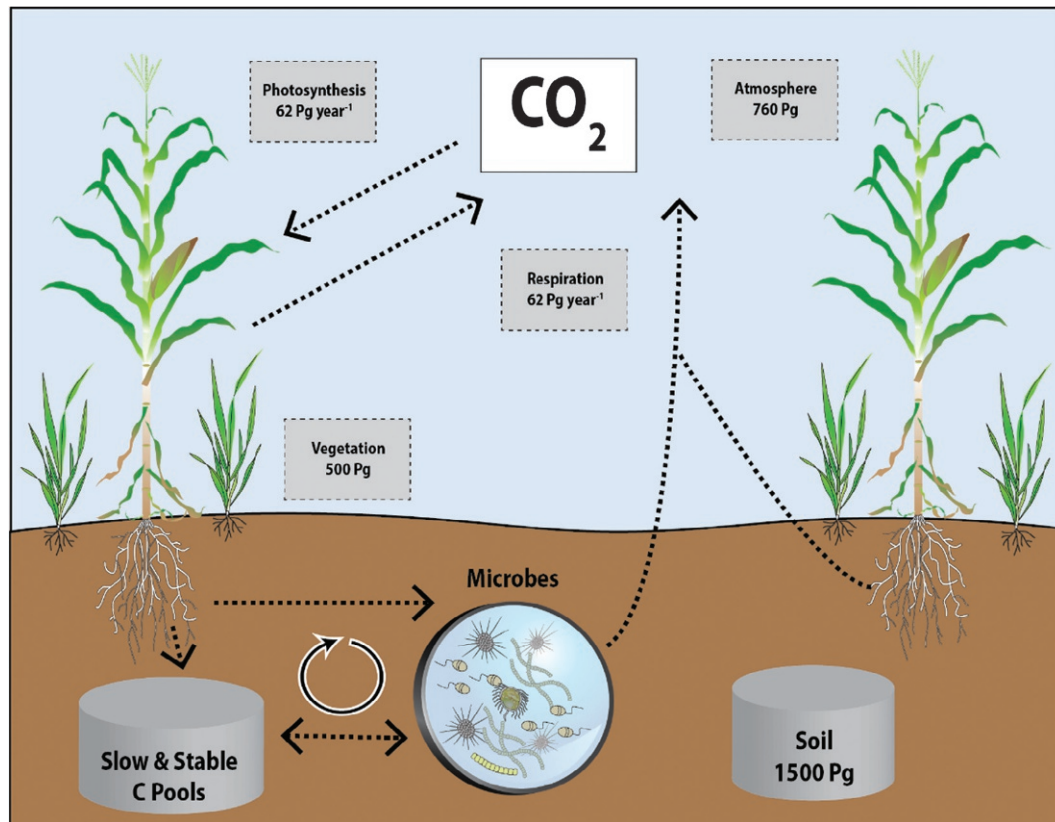


Fig. 1 Simplified representation of the terrestrial carbon cycle between the atmosphere, plants, and soil.

4. Atmosphere: The atmosphere consists primarily of carbon dioxide, carbon monoxide, and methane. The amount of carbon in the atmosphere has increased from 578 Pg in 1700 to about 766 Pg in 1999, and continues to increase at the rate of about 6.1 Pg year⁻¹.
5. Biosphere: The biosphere consists of all living and dead organisms not yet converted into soil organic matter. The amount of carbon that resides in the biosphere is 540–610 Pg.

In the terrestrial system, soils play a key role both in reservoirs and fluxes in the plant–soil–atmosphere continuum. Carbon in soils is present both in inorganic and organic forms. Carbon comprises up to 10% of the soil mass, except for waterlogged soils such as Histosols that contain up to 30% carbon. The surface horizons of most soils contain less than 3% carbon. Most of the carbon that resides in the soil is in the organic form, except for arid soils and some soils formed from carbonate parent material. While carbon is a relatively minor component in terms of mass, it serves an important function in soil and the environment.

Fluxes

The largest fluxes of the carbon cycle are those that occur between the atmosphere and the vegetation and oceans. Carbon enters the soil primarily through plants (Fig. 1). During photosynthesis, plants convert CO_2 into organic carbon for growth. The estimated fixation by photosynthesis is approximately 62 Pg year⁻¹. This plant carbon enters the soil through root exudates, root turnover, and the addition of aboveground plant material. Upon their death, plant tissues decompose, primarily by soil microorganisms, with much of the carbon in plant material eventually being released back into the atmosphere as CO_2 . Respiration is estimated to release 62 Pg C year⁻¹, which balances photosynthesis. This compares with the 5 Pg year⁻¹ released by burning of fossil fuels; however, the release of C by burning of fossil fuels is not offset by a sink, thus resulting in the dramatic increase in atmospheric CO_2 . In some cases, under anaerobic conditions, methane (CH_4) is produced by microorganisms. However, soil microorganisms also consume CH_4 under aerobic conditions. Approximately 7% of the atmospheric C is recycled between plant uptake and microbial respiration. Thus, plant productivity and microbial respiration are the two key mediators between atmospheric CO_2 uptake and input.

Soil-forming factors

The level of organic carbon in soils is a function of climate, topography, biology (plant, fauna, and microorganisms), parent material, and time. These are the same factors that affect soil development. The first few years after a disturbance, such as soil tillage,

soil carbon increases relatively dramatically but slows until equilibrium is attained under the conditions of the existing environment. This is easily seen after reclamation of mined lands or conversion of cultivated cropland to permanent grasslands. A change in the environmental conditions with climate or management can alter the equilibrium levels. The equilibrium is reached as plant inputs balance the outputs of microbial respiration. The soil organic C levels are a result of biological recalcitrance and physical and chemical stabilization of the carbon, to be discussed later.

Climate is the most important factor governing soil organic carbon levels. Climate through water and temperature affects plant productivity and microbial activity. Hot climates, where water is restricted, have the lowest soil organic carbon levels because plant production is limited. Deserts are characterized by this environment. In the USA, precipitation increases from the Rocky Mountains eastward, thus resulting in greater soil carbon levels. This general increase in soil carbon is due to precipitation enhancing plant productivity.

Vegetation type can also influence the soil organic carbon levels through composition and production of the plant biomass. Plant materials vary in their decomposability due to chemical differences. Lignin and other polyphenolic substances decrease the decomposability of the plant material. Grasses tend to promote higher soil organic carbon than agricultural systems. Grasses have not only high productivity but also allocate more photosynthate belowground. The large densities and intense turnover of grass roots tend to favor formation of soil organic carbon. Although forests also sequester carbon, a greater proportion of that carbon is in woody biomass and litter (O horizon) (Sarto et al., 2020c). There are differences among crop species in the ability to sustain soil organic carbon levels. In general, high residue-producing annual crops such as corn, wheat, and grain sorghum replenish the supply of carbon to the soil. Low residue-producing crops, such as cotton and soybeans, tend to deplete soil organic carbon levels.

Parent material affects levels of soil carbon through the effect on texture, mineralogy, and pH. The amount and type of clay in soils also influences the ability of the soil to adsorb carbon compounds to clay surfaces and form stable soil aggregates, which contribute to the protection and stabilization of soil organic carbon.

Topography affects soil organic carbon levels through effects on microclimate, runoff and erosion, and evaporation. Lower slope positions often have higher organic carbon levels because of deposition and greater plant production. Slope positions with excess water also restrict microbial decomposition, resulting in the buildup of soil organic C.

Functions

Soil carbon is the most important addition to soils as it affects their properties. Carbon as soil organic matter alters the physical, chemical, and biological properties of the soils. Organic matter is a primary indicator of soil health defined as “the capacity of soil to function as a vital living system, within ecosystem and land-use boundaries, to sustain plant and animal productivity, maintain or enhance water and air quality, and promote plant and animal health” (Doran and Zeiss, 2000) (Fig. 2). Improvements in soil organic matter create a beneficial environment, leading to increases in plant growth. Increasing soil carbon stocks would increase agricultural resilience and fertility and reduce GHG emissions from soils (Paustian et al., 2019).

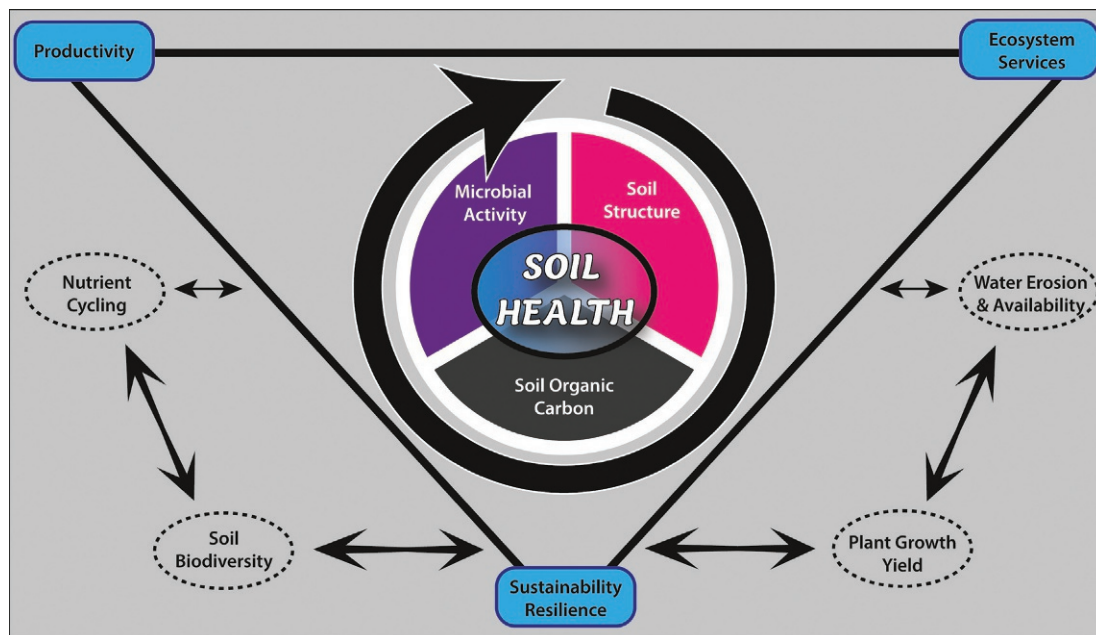


Fig. 2 Relationship between soil organic carbon, soil structure, and microbial activity.

Physically, soil organic matter improves aggregation of soil particles. Improved aggregation results in better soil structure, allowing for movement of air and water through the soil as well as better root growth. More stable soil structure results in less soil erosion, which retains nutrients on the land and protects water quality. Higher levels of soil organic carbon reduce bulk density, thus providing an improved rooting environment. In addition, organic matter holds water, an important attribute for plant growth in arid and mesic environments. Higher soil organic matter decreases soil crusting and increases water infiltration rates which enhance plant productivity and thus the return of plant material to the soil.

Chemically, soil organic carbon increases the cation exchange capacity of the soil. Soil organic matter can contribute 20 to 80% of the soil's cation exchange capacity. Cation exchange is important for the retention of many nutrients. Associated with the organic carbon are organic-bound nitrogen, phosphorus, and sulfur which, upon decomposition, provide slow release of these nutrients for plant production. In some cases, soil organic compounds enhance the chelation of metals, thus increasing the bioavailability of trace elements required for plant growth. Soil organic carbon often provides binding sites for many anthropogenic chemicals, thus minimizing leaching of hazardous chemicals through the soil profile or making them less available, which reduces toxicity.

Biologically, soil organic carbon is the source of carbon and energy for most soil microorganisms and fauna. Increased soil organic carbon enhances the biomass and diversity of the soil biota (McGowan et al., 2019; Sarto et al., 2020a). Since the soil microbiome drives energy and related biogeochemical transformations in soil, plant nutrient availability is often enhanced with the increase in soil microbial biomass and activity. Some organic compounds in the soil also exhibit plant growth-promoting properties, further enhancing plant productivity. Increasing crop frequency and species diversity in agricultural systems promote nutrient cycling, microbial biomass and activity, aggregate stability, and soil protection. Crop diversification increases the soil microbial biomass (Bonini Pires et al., 2020). The conversion of native biomes to agricultural systems shifts the soil microbial community composition, enzyme activity, C and N, and soil structure (Sarto et al., 2020b) (Fig. 3).

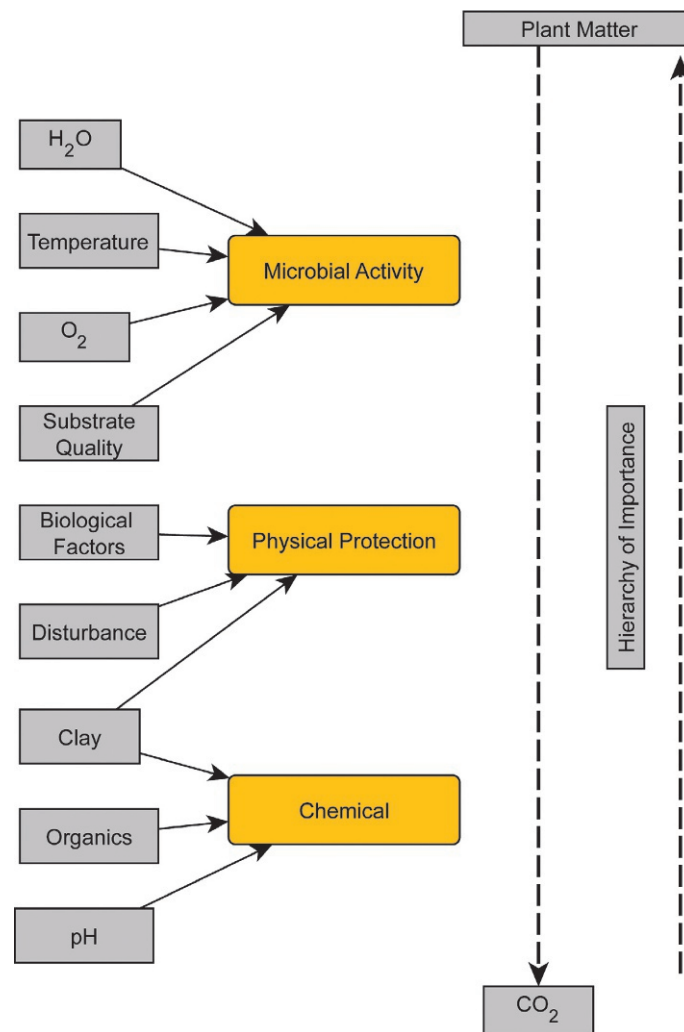


Fig. 3 Regulation of carbon stabilization in soil.

Regulation of soil organic carbon dynamics

As much as 20% of plant carbon remains in the soil as organic matter, sometimes referred to as 'humus.' Some of this carbon can remain in soils for hundreds and even thousands of years. The quantity of organic carbon in soil is a function of the amount of plant material entering the soil, the microbial decomposition rates of those residues, and the soil chemistry and mineralogy.

Soil C may be stabilized due to biochemical recalcitrance, chemical stabilization, and physical protection (Fig. 2). Any environmental factor that affects microbial activity influences the decomposition of organic material. These factors include soil water, temperature, pH, and O₂. Plant production and soil microbial activity are recognized as the biological processes governing soil carbon dynamics (Fig. 2). The amount of plant material produced is a function of plant species, as well as climate (temperature and available water) and nutrients. The composition of the plant material entering the soil is also a key regulator of microbial decomposition. Typical plant components include soluble sugars, hemicellulose, cellulose, and lignin, which vary in proportions between plant species, within plant species, and within plant organs. Of these components, lignin is the most difficult to degrade.

The quality of the carbon compounds (e.g., N content, C:N ratio, and lignin: N ratio) are important factors regulating decomposition processes (Fig. 2). The C:N ratio often serves as a guide to potential for decomposition; a ratio of more than 30 slows the rate of decomposition because of the lack of sufficient N for microbial growth; a ratio less than 20 provides sufficient N for microbial growth and allows decomposition to proceed (Robertson and Groffman, 2007). A ratio of 20–30 results in an equilibrium state between N mineralization and immobilization (Brust, 2019). Lignin content or lignin: N ratios are also used as a guide of organic matter degradability.

The organic carbon is transformed to greater biological recalcitrance as soil microorganisms process the C inputs from plant material in the course of microbial succession. Soil carbon may be stabilized because of its biochemical recalcitrance, e.g., lignin derivatives or melanin produced by fungi and other soil organisms. The rate of transformation is a function of temperature and soil moisture. As temperature increases, microbial activity increases; doubling for every 10 °C increase, ($Q_{10} = 2$). Soil-water content is also important; optimum microbial activity occurs at near 'field capacity,' which is equivalent to 60% water-filled pore space (Linn and Doran, 1984). As soils become waterlogged, decomposition slows and becomes less complete due to the lack of O₂. Peat soils are a result of these waterlogged conditions. As soils dry below 60% water-filled pore space, decomposition is also slowed due to water limiting conditions.

Soil structure plays a dominant role in controlling microbial access to substrates, and thus turnover (Fig. 2). Relatively labile material may become physically protected from decomposition by incorporation into soil aggregates. Disruption of aggregates, either by natural forces (freeze–thaw, wet–dry) or human activity (tillage) stimulates decomposition of the organic C protected inside the aggregates. One of the primary causes of soil organic C loss is tillage. The role of soil structure in the protection of soil organic carbon from decomposition is demonstrated by increased C mineralization in disrupted aggregates compared to intact aggregates. Cultivation of soils strongly affects the structural stability and reduces the amount of soil organic C, and this loss reduces the proportion of soil macroaggregates. No-till systems with crop rotation increase stabilization of easily degradable organic carbon via either physical protection or association with soil minerals (Arachchige et al., 2018).

Chemical stabilization of C refers to binding of organic C to clay surfaces or precipitation. Chemical stabilization by clay is regulated by clay type, pH, and other organics (Fig. 2). Clays that have a high adsorption capacity, such as 2:1 clays (e.g., montmorillonite), protect organic carbon to a greater extent than 1:1 clays (e.g., kaolinite), which have a lower adsorption capacity. Certain cations such as Ca, Al, and Fe are known to enhance the stabilization of organic carbon through cation bridges with clay surfaces. Clay type is an inherent property of soil and therefore cannot be directly manipulated as easily as the biological and physical mechanisms of stabilization. Inorganic forms of carbon in soil are the result of the equilibrium of CO₂ with water to form carbonic acid (H₂CO₃), bicarbonate (HCO₃⁻¹), and carbonate (CO₃⁻²) in the soil solution. Secondary minerals form in arid soils from the precipitation of Ca and Mg with carbonate. Dissolution of minerals also occurs depending on pH.

Composition of soil organic carbon

The most widely accepted theory on the formation of soil humus carbon is that, as plant material is decomposed by microorganisms, the altered compounds and new compounds synthesized by soil microbes polymerize through chemical or enzymatic reactions. Soil organic carbon is undergoing constant transformation. Typically, most models separate soil organic carbon into pools of organic matter that differ in composition and decomposability (Fig. 4). For example, the 'active' pool, comprised of microbial biomass and labile organic compounds, makes up less than 5% of the soil organic C. The slow pool usually makes up 20–40% of the total organic C, and the recalcitrant pool makes up 60–70% of the soil C. The microbial biomass, while a smaller proportion of the organic C in the soil, is the processing agent. The recalcitrant pool is material that is difficult to degrade and contains what is referred to in the historical literature as humic and fulvic acids – materials obtained by chemical fractionation procedures. However, a more recent view of recalcitrance is much more complex and is related to accessibility and ability of microbes to degrade complex carbon structures (Lehmann and Kleber, 2015). The active pool has turnover times in the order of months to years; the slow pool takes decades to turn over, while C in the recalcitrant pool takes from hundreds to thousands of years to turn over completely. However, 2–5% of the recalcitrant pool is degraded annually (Rice, 2006). Since the recalcitrant pool is generally in equilibrium in natural systems, the formation rate may equal the degradation rate (Rice et al., 2021).

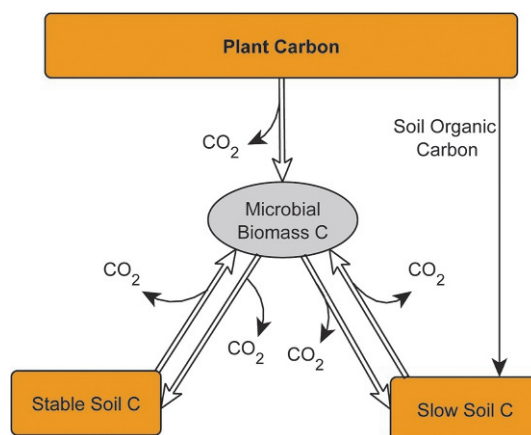


Fig. 4 Conceptualization of soil carbon pools. Adapted from Paul EA and Clark FE (1996) *Soil Biology and Biochemistry*. San Diego, CA: Academic Press, with permission.

Management

Most of the world's productive soils are now under cultivation. Historically, agriculture relied upon plowing the soil. In some cases, poor crop yields and removal of crop residues reduced the amount of organic material (carbon) to the soil. As a result, soil C content has decreased in many agricultural soils by as much as 50% over the last 50- to 100-years. A typical soil carbon-loss curve is shown in Fig. 4. Losses are rapid initially and then decline with time after the conversion to agriculture from native ecosystems. In recent decades, higher yields, return of crop residues, and development of conservation tillage practices have started to reverse the loss of soil carbon. These and other advancements in crop and soil management practices have the potential to increase soil C. Table 1 lists several practices affecting the soil's ability to sequester C. These practices can be grouped into three guiding principles: (1) minimization of soil disturbance; (2) maximizing plant production through nutrients and water availability; and (3) maximizing return of plant biomass (Fig. 5).

Intensive tillage, poor crop yields, lack of crop diversification, and excessive removal of crop residues have reduced C replenishment to the soil and thus negatively affected soil productivity. Tillage promotes disruption of soil aggregates, which exposes 'protected' carbon inside aggregates to microbial attack. In addition, tillage increases aeration, over-stimulating microbial activity, which results in the conversion of organic carbon to CO_2 . A reduction in tillage intensity, with no-tillage being the least disruptive in cropped soils, allows macroaggregate formation and reduces the degradation of soil organic C. *Minimum* or conservation tillage also conserves carbon in soil. Losses of soil carbon by erosion are reduced by conservation tillage. No-till can also enhance crop productivity (Daigh et al., 2018) and improve several soil properties, such as soil organic carbon (Nicoloso et al., 2018), bulk density (Blanco-Canqui et al., 2009), soil aggregation (Fabrizzi et al., 2009), and soil microbial community biomass (Bonini Pires et al., 2020). Organic waste and organic fertilizer, such as cattle manure, may replenish and maintain soil nutrient equilibrium, thus increasing nutrient status and crop yield. Organic amendments and crop rotation can increase soil organic carbon (Nicoloso et al., 2020).

Table 1 Management strategies for soil carbon sequestration.

Practice	Land use			
	Cropland	Rangeland	Forestry	Agroforestry
Soil Management	Reduced tillage intensity and residue retention, residue management, fertility, composting, improved water management, biochar application	Biodiversity/Perennial deep root, nutrient management, animal management, fire management	Reduce Deforestation, Afforestation and/or Reforestation, improved and sustainable forest Management, Fire Management, reduce conversion of savannas and Grasslands, reduce conversion and degradation of peatlands	woody biomass (including trees and woody shrubs) with crops, C allocation in different depths, synergism between species.
Crop Management	Crop varieties, crop rotation, cover crops, perennial cropping systems, integrated production system, crop diversification,			

Adapted from Lal R, Kimble JR, Follett RF, and Cole CV (1998) *The Potential of US Cropland to Sequester Carbon and Mitigate the Greenhouse Effect*. Ann Arbor, MI: Ann Arbor Press.

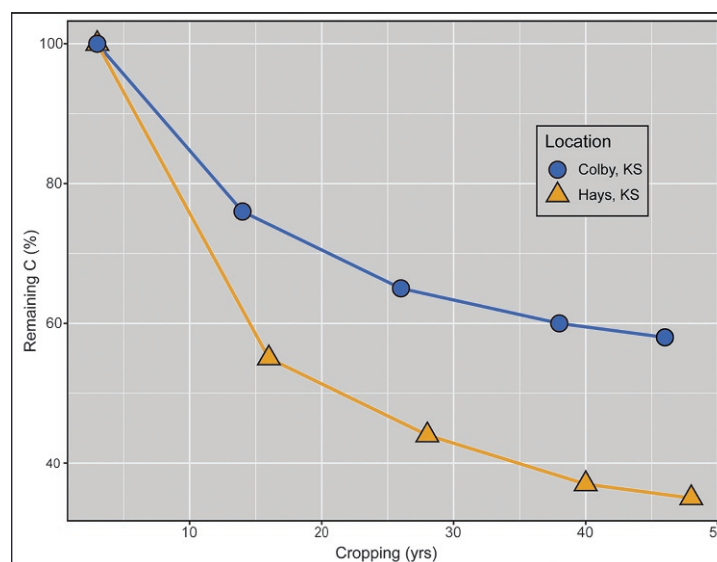


Fig. 5 Loss of soil carbon at the onset of cultivation for two locations in Kansas, Hays and Colby. Redrawn from Haas, H.J., Evans, C.E., and Miles, E.F.: 1957, Nitrogen and Carbon Changes in Great Plains Soils as Influenced by Cropping and Soil Treatment, USDA Technical Bulletin 1164, Washington, DC, US Department of Agriculture.

Cropping intensity affects the equilibrium level of soil organic carbon. Crop rotations and diversification can impact soil carbon levels by varying the amount and composition of the plant material added to the soil (Nicoloso et al., 2020; Bonini Pires et al., 2020). Crop rotation with forage crops can also, potentially increase P and K contents in the topsoil layer (Bassegio et al., 2015; Sarto et al., 2021). Cover crops capture nutrients such as phosphorus (P) and potassium (K) from deeper depths and store them in their aboveground biomass. The cover crop biomass will then be laid upon the soil surface, decompose and release nutrients through mineralization, making them available at the soil surface (Bassegio et al., 2015). In agriculture, high residue-producing crops such as corn, wheat, and sorghum tend to sustain or increase soil C. Elimination of summer fallow, a practice used in the Great Plains can increase soil C by providing more plant material on an annual basis. In those areas where sufficient soil water is available, double cropping is a viable practice, where two crops are produced per year. In more humid regions, winter cover crops are also a viable option for returning plant material to the soil. In grassland systems, perennial crops also tend to increase soil C because of the lack of tillage and the addition of organic C through root turnover. Integrated crop–livestock systems (ICLS) can increase agricultural sustainability and C sequestration potential. Sarto et al. (2020a) verified that the inclusion of *Eucalyptus* spp. to an *Urochloa* pasture added more C below 0.6 m depth, while *U. brizantha* increased C in the first 0.2 m, showing complementarity between the two species. However, the integration of trees into grass-dominated vegetation includes changes in above-and below ground productivity (Sarto et al., 2020c).

Those practices that directly affect plant production are also important in increasing soil carbon. Improved water management alleviates short-term water stress during the growing season, which improves the production of plant biomass and the return of plant material to the soil. In arid regions, irrigation can increase plant production and microbial activity. Water management also affects microbial activity, as discussed earlier. Fertility management eliminates nutrient limitations of plant growth. Proper management of nutrients supplies sufficient nutrients for the plant while minimizing the excess that could lead to degradation of the environment.

In agriculture, practices that increase soil organic matter include increase crop intensification, reduced tillage intensity and residue retention, crop management (e.g., improved crop varieties, crop rotation, use of cover crops, perennial cropping systems, integrated production systems (agroforest, intercropping systems), crop diversification, agricultural biotechnology), improved water management, manure or biochar application, maintained soil cover. Conservation agriculture combines the use of minimum soil disturbance, crop rotation (including cover crop), and residue retention to ensure minimal soil tillage intensity, maximum soil cover and increasing soil microbial community composition.

For grazing land, improved species, the addition of legumes, and fertilizer management can increase soil C in tame pastures. In native lands, proper grazing rotation and management of burning are strategies to improve soil carbon, especially for previously poorly managed grazing lands. Fire management includes the improved use of fire for sustainable ecosystem management of forested and savanna ecosystems, including wildfire prevention and prescribed burning. Prescribed burning is used to reduce the risk of large, uncontrollable fires in forest areas. Controlled burning is an effective economic method of reducing fire danger and stimulating natural reforestation under the forest canopy, and after clear felling, prevention of soil erosion and land degradation and is used in rangelands to conserve biodiversity and to enhance forage quality.

Table 2 Estimates of C sequestration potential of agricultural practices of US cropland.

Agricultural practice	(MTC ha ⁻¹ year ⁻¹)
Conservation Reserve Program	0.3–0.7
Conservation tillage	0.24–0.40
Fertilizer management	0.05–0.15
Rotation with winter cover crops	0.1–0.3
Summer fallow elimination	0.1–0.3

Adapted from Lal R, Kimble JR, Follett RF, and Cole CV (1998) The Potential of US Cropland to Sequester Carbon and Mitigate the Greenhouse Effect. Ann Arbor, MI: Ann Arbor Press.

Reducing deforestation and forest degradation supports the conservation of existing carbon pools in forest vegetation and soil by avoiding tree cover loss and disturbance. Reforestations are activities that convert land to forest, where reforestation applies to land that has previously grown forests and afforestation to land that historically has not supported forest. Reducing the conversion of grasslands and savannas to croplands prevents soil carbon losses by oxidation, and to a smaller extent, biomass carbon loss due to vegetation clearing.

Restoration of degraded soils and ecosystems offers a high potential for storing carbon in soils. Marginal croplands could be reverted to forests and grasslands. The Conservation Reserve Program of the US Department of Agriculture is designed to take marginal lands out of production (Table 2). Restoration of soils of mined land offers many opportunities for storing carbon in soil.

Conclusion

Soil carbon is a key pool of the global carbon cycle. The amount of carbon stored in soils is a function of the soil-forming factors climate, topography, soil type, time, vegetation. Human activity principally through agriculture and forest management influences soil carbon to the extent that we have lost about 50% of the carbon stored in soils globally. However, with improved agricultural management through increased plant productivity and forest management, we can restore the soil carbon. Soil carbon is the key component of soil status as it affects the soil's chemical, physical, and biological properties. With carbon as a critical component of soil health and offsets for climate change, interest in managing soils for carbon will increase interest and the management of soil carbon.

References

- Arachchige PSP, Hettiarachchi GM, Rice CW, et al. (2018) Sub-micron level investigation reveals the inaccessibility of stabilized carbon in soil microaggregates. *Scientific Reports* 8: 16810. <https://doi.org/10.1038/s41598-018-34981-9>.
- Bassegio D, Santos RF, Secco D, et al. (2015) Short-term effects of crop rotations on soil chemical properties under no-tillage condition. *Australian Journal of Crop Science* 9: 49–54.
- Blanco-Canqui H, Stone LR, Schlegel AJ, Lyon DJ, Vigil MF, Mikha MM, Stahlman PW, and Rice CW (2009) No-till Induced Increase in Organic Carbon Reduces Maximum Bulk Density of Soils. *Soil Science Society of America Journal* 73: 1871. <https://doi.org/10.2136/sssaj2008.0353>.
- Bonini Pires CA, Amado T, Reimche G, Schwalbert R, Sarto M, Nicoloso R, Fiorin JE, and Rice CW (2020) Diversified crop rotation with no-till changes microbial distribution with depth and enhances activity in a subtropical Oxisol. *European Journal of Soil Science* 71: 1173–1187. <https://doi.org/10.1111/ejss.12981>.
- Brust GE (2019) Management strategies for organic vegetable fertility. In: Biswas D and Micallef SA (eds.) *Safety and Practice for Organic Food*, pp. 193–212. London, U.K: Academic Press.
- Daigh ALM, Dick WA, Helmers MJ, Lal R, Lauer JG, Nafziger E, Pederson CH, Strock J, Villamil M, Mukherjee A, and Cruse R (2018) Yields and yield stability of no-till and chisel-plow fields in the Midwestern US Corn Belt. *Field Crops Research* 218: 243–253. <https://doi.org/10.1016/j.fcr.2017.04.002>.
- Doran JW and Zeiss MR (2000) Soil health and sustainability: managing the biotic component of soil quality. *Applied Soil Ecology* 15(1): 3–11.
- Fabrizzi KP, Rice CW, Amado TJC, Fiorin J, Barbagelata P, and Melchiori R (2009) Protection of soil organic C and N in temperate and tropical soils: effect of native and agroecosystems. *Biogeochemistry* 92: 129–143. <https://doi.org/10.1007/s10533-008-9261-0>.
- Lal R (2004) Soil carbon sequestration impacts on global climate change and food security. *Science* 304: 1623–1627. <https://doi.org/10.1126/science.1097396>.
- Lehmann J and Kleber M (2015) The contentious nature of soil organic matter. *Nature* 528: 60–68. <https://doi.org/10.1038/nature16069>.
- Linn DM and Doran JW (1984) Effects of water-filled pore space on carbon dioxide and nitrous oxide production in tilled and non-tilled soils. *Soil Science Society of America Journal* 48: 1267–1272. <https://doi.org/10.2136/sssaj1984.03615995004800060013x>.
- McGowan AR, Nicoloso RS, Diop HE, Roozeboom KL, and Rice CW (2019) Soil organic carbon, aggregation, and microbial community structure in annual and perennial biofuel crops. *Agronomy Journal* 111: 128–142. <https://doi.org/10.2134/agronj2018.04.0284>.
- Nicoloso RS, Rice CW, Amado TJC, Costa CN, and Akley EK (2018) Carbon saturation and translocation in a no-till soil under organic amendments. *Agriculture, Ecosystems and Environment* 264: 73–84. <https://doi.org/10.1016/j.agee.2018.05.016>.
- Nicoloso RS, Amado TJC, and Rice CW (2020) Assessing strategies to enhance soil carbon sequestration with the DSSAT-CENTURY model. *European Journal of Soil Science* 71: 1034–1049. <https://doi.org/10.1111/ejss.12938>.
- NOAA National Centers for Environmental Information (2014) State of the Climate: Global Climate Report for Annual 2014. published online January 2015, retrieved on April 19, 2021 from <https://www.ncdc.noaa.gov/sotc/global/201413>.

- Paustian K, Collier S, Baldock J, Burgess R, Creque J, DeLonge M, Dungait J, Ellert B, Frank S, Goddard T, Govaerts B, Grundy M, Henning M, Izaurrealde RC, Madaras M, McConkey B, Porzig E, Rice C, Searle R, et al. (2019) Quantifying carbon for agricultural soil management: from the current status toward a global soil information system. *Carbon Management* 10(6): 567–587. <https://doi.org/10.1080/17583004.2019.1633231>.
- Rice CW (2006) Organic matter and nutrient dynamics. In: *Encyclopedia of Soil Science*, 2nd Edn, Boca Raton, FL: Taylor and Francis Group.
- Rice CW, Bonini Pires CA, Lin J, and Sarto MVM (2021) Soil organic carbon assessment methods. In: Karlen DL, Stott DE, and Mikha MM (eds.) *Soil Health: Vol. 2: Laboratory Methods for Soil Health Assessment*. Madison, WI: Soil Science Society of America (SSSA) & Wiley International, SSSA. Chapter 3.
- Robertson GP and Groffman PM (2007) Nitrogen transformation. In: Paul EA (ed.) *Soil Microbiology, Biochemistry, and Ecology*, pp. 341–364. New York, New York, USA: Springer.
- Sarto MVM, Borges WLB, Sarto JRW, Rice CW, and Rosolem CA (2020a) Deep soil carbon stock, origin, and root interaction in a tropical integrated crop–livestock system. *Agroforestry Systems* 94: 1865–1877. <https://doi.org/10.1007/s10457-020-00505-6>.
- Sarto MVM, Borges WLB, Bassegio D, et al. (2020b) Soil microbial community, enzyme activity, C and N stocks and soil aggregation as affected by land use and soil depth in a tropical climate region of Brazil. *Archives of Microbiology* 202: 2809–2824. <https://doi.org/10.1007/s00203-020-01996-8>.
- Sarto MVM, Borges WLB, Sarto JRW, Rice CW, and Rosolem CA (2020c) Root and shoot interactions in a tropical integrated crop–livestock–forest system. *Agricultural Systems* 181: 102796. <https://doi.org/10.1016/j.agsy.2020.102796>.
- Sarto MVM, Borges WLB, Bassegio D, Rice C, and Rosolem CA (2021) Maize and sorghum root growth and yield when intercropped with forage grasses. *Agronomy Journal* 113: 1–16. <https://doi.org/10.1002/agj2.20920>.
- Smith P, Bustamante M, Ahammad H, Clark H, Dong H, Elsiddig EA, Haberl H, Harper R, House J, Jafari M, Masera O, Mbow C, Ravindranath NH, Rice CW, Robledo Abad C, Romanovskaya A, Sperling F, and Tubiello F (2014) Agriculture, forestry and other land use (AFOLU). In: Edenhofer O, Pichs-Madruga R, Sokona Y, Farahani E, Kadher S, Seyboth K, ... Minx JC (eds.) *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Further reading

- Karlen DL and Rice CW (2015) Soil degradation: Will human kind ever learn? *Sustainability* 7: 12490–12501.
- Lal R and Follett RF (2009) Priorities in soil carbon research in response to climate change. *Soil Carbon Sequestration and the Greenhouse Effect* 57: 401–410. <https://doi.org/10.2136/sssaspecpub57.2ed.c23>.
- Lal R (2015) Sequestering carbon and increasing productivity by conservation agriculture. *Journal of Soil and Water Conservation* 70: 55A–62A. <https://doi.org/10.2489/jswc.70.3.55a>.
- Rice CW (2005) Carbon cycling in soils: Dynamics and management. In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*, pp. 164–169. Oxford, U.K: Elsevier Ltd.

Carbon storage in soils

Isabelle Basile-Doelsch^a, Tiphaine Chevallier^b, Marie-France Dignac^c, and Amandine Erktan^b, ^aAix Marseille Univ, CNRS, IRD, INRAE, CEREGE, Aix-en-Provence, France; ^bIRD, INRAE, CIRAD, Institut Agro, Université Montpellier, Eco&Sols, Montpellier, France; ^cINRAE, CNRS, Sorbonne Université, IRD, Institute of Ecology and Environmental Sciences of Paris (iEES-Paris), Paris, France

© 2023 Elsevier Ltd. All rights reserved.

Introduction	229
Quantification of soil C stocks and storage	229
C content and density: Measurement parameters to calculate the soil C stock	229
C stock calculation	230
Soil C storage	232
The soil C stock is dynamic	233
Carbon inputs to soil	233
Organic inputs kickstart the soil food chain	234
Organic matter biotransformation in soil: A carbon recycling business	234
Organic matter in close interaction with mineral constituents	235
Agricultural practices drive the C storage dynamics	236
Practices and land use changes with a negative impact on the soil C stock	236
Practices and land use changes with a positive effect on soil carbon storage	237
Practices with uncertain effects on soil C stocks	238
The benefits of soil C storage from the societal viewpoint	238
Conclusion	240
Acknowledgments	241
References	241

Abstract

Soils are a pivotal component in the global carbon cycle, while carbon storage in soils is a natural phenomenon involving organic carbon. Maintaining or increasing soil carbon levels is beneficial for many ecosystem services. Soil carbon is also a soil condition indicator and a key focus of several Sustainable Development Goals. This chapter describes the forms of carbon in soils, the quantification of carbon stocks and storage, the processes underlying the heterogeneous distribution of carbon stocks across the planet and their dynamics, land-use changes and practices that affect soil carbon stocks, as well as the socioeconomic benefits of soil carbon storage.

Glossary

Soil C content The C content in a soil is the quantity of C (organic or inorganic) in a soil mass. Total organic carbon (TOC) is expressed in gC.kg^{-1} soil.

Soil C stock The C stock is the quantity of C in a soil volume. In practice, it is expressed as a quantity of C related to a given surface area (tC.ha^{-1} or kgC.m^{-2}), for a given soil thickness.

Soil C storage Carbon storage represents the increase in a carbon stock over time, while generally being determined for a given soil, depth, land management change, duration and climate setting. It is expressed in $\text{tC.ha}^{-1}.\text{yr}^{-1}$. This storage can take a negative value when there is a loss of C. A distinction is made between potential storage, additional storage and effective storage (Fig. 3).

Land use Land use is the term applied to describe the human use of land. In the soil C stock framework, soils that bear forests, grasslands, cropland and soils of urban environments are differentiated under this term.

Agricultural practices For a given type of land use, agricultural practices refer to the techniques used for farming. They include, for example, vineyard grass cover cropping, establishment of temporary grasslands, intercropping, plowing or no-till farming, agroforestry, irrigation. . .

Key points

- Carbon occurs in different forms in soils: in living organisms, dead organic matter with particulate organic matter, organic molecules, and as inorganic carbon
- Soil carbon stock and soil carbon storage are different concepts: a carbon stock is quantified at a given date while its storage is quantified over a period of time

- The carbon stock distribution across the planet is heterogeneous, and complex biotic and abiotic processes underlie its dynamics
- For a given soil and climate, the largest stocks are found in forests and grasslands. In comparison, cultivated soils are carbon depleted
- Land-use changes and agricultural practices have a major impact on soil carbon stocks: they may have negative, positive or diverse effects on carbon storage, depending on climate and soil conditions
- Soil carbon storage provides socioeconomic benefits as targeted by five Sustainable Development Goals

Introduction

Plants fix atmospheric CO_2 through photosynthesis and incorporate it into the soil as shoot and root litter residues, and root exudates. This fixed carbon (C) remains soilborne as organic matter over variable periods of time before being largely converted back into CO_2 via the respiration of decomposer organisms. Soil organic C, estimated at more than 1500 billion tonnes, represents a major reservoir of the C cycle on continents. About half of this amount is present in the atmosphere (830 Gt), and about a third in aboveground plant biomass (550 Gt). Soils are at the heart of global carbon flows. Carbon sequestration in soils, i.e., the capture and storage of atmospheric carbon, contributes to climate regulation. In the context of increasing soil degradation and climate disruption, interest in soil organic carbon (SOC) has increased to the point where the topic emerged on the political agenda with the launch of the 4 per 1000 initiative at COP21 in Paris.

C occurs in soils as organic matter and inorganic mineral phases such as carbonates (e.g., CaCO_3). Inorganic C, especially present in soils developed on carbonate sedimentary bedrock such as limestone, generally follows reactive processes on long time scales. Although about 30% of soils worldwide contain inorganic C, its short-term dynamics and influence on organic C dynamics has barely been studied. Hence, only processes related to soil organic C will be discussed in detail hereafter in this chapter. In soils, the different types of organic matter that contain C include simple and complex biopolymers (~75%), particulate organic debris (including carbonized debris) (~20%), microorganisms (~2%), fauna (<1%) and very fine roots (<2%) (Fig. 1). These relative proportions are approximate and vary according to the climate, soil type, horizon, depth, land use and agricultural practices. C stocks are generally much higher in forests and grasslands than in croplands. As organic matter is essential for soil biological activity and health, increasing soil organic carbon stocks in croplands should contribute to food security and in turn limit global warming. In this chapter, we first define and explain how to measure soil C stocks and calculate C storage. The main mechanisms that control C storage are briefly presented. We then review agricultural practices and land use changes that favor or impair C storage in soil and finally discuss the importance of societal issues that arise from the SOC management strategies.

Quantification of soil C stocks and storage

The soil C stock is the quantity of C in a volume of soil. In practice, it is expressed as a quantity of C (measured in a soil sample) relative to a land surface area (tC.ha^{-1} or kgC.m^{-2}) in a given soil layer. To be representative of the real C stock in a studied plot, the soil samples being analyzed should be representative of the plot, with the sampling strategy adapted to the environmental heterogeneity. In highly heterogeneous soils, it is important to sample at fine plot scales across the surface (horizontal heterogeneity) and down the vertical profile. Storage (or destocking) represents the gain (or loss) of stock over a given time period. It is usually expressed in $\text{tC.ha}^{-1}.\text{yr}^{-1}$. Controversies regarding the quantification of C stocks and storage often arise due to ambiguities or lack of accuracy in the calculations. Therefore, we recall here: (i) the measurements needed (C content and soil density) to quantify soil C stocks, (ii) the calculation methods and possible corrections to be made, as well as (iii) the ranges of C stocks encountered in different types of soil at the global scale. We then detail the methods available to estimate soil C storage.

C content and density: Measurement parameters to calculate the soil C stock

To quantify a soil C stock, we have to measure: (1) the total soil C content (TOC), and (2) the soil bulk density, layer by layer of the soil profile, down to the depth considered in the stock calculation.

The TOC of a soil sample is the proportion of C (in mass) contained as organic matter in dry soil (dried at 105°C). It is expressed in gC.kg^{-1} of soil (gC.kg^{-1} is the International System unit, but % is also used, with $1\% = 10 \text{ gC.kg}^{-1}$). Unless otherwise specified, the soil mass is the dry mass at 105°C of 'fine soil,' i.e., soil sieved through a 2 mm mesh, without plant debris or coarse elements (gravel and pebbles, whose mass must be weighed independently). The organic C content of fine soil is usually measured directly, in a subsample of a few mg, by dry combustion with an elemental analyzer (after decarbonation of the sample if needed). The measurement uncertainty is about 0.1 gC.kg^{-1} . Other thermal and chemical methods exist that are often less precise or generate polluting products, e.g., the wet oxidation method ('Anne' or 'Walkley-Black'). Other indirect methods (visible and infrared spectroscopy) are under development (Allory et al., 2019). The latter are time-effective and inexpensive methods, but are not accurate enough (estimated uncertainty of 2.6 gC.kg^{-1} soil to 4.8 gC.kg^{-1}), while requiring complex calibration that is dependent

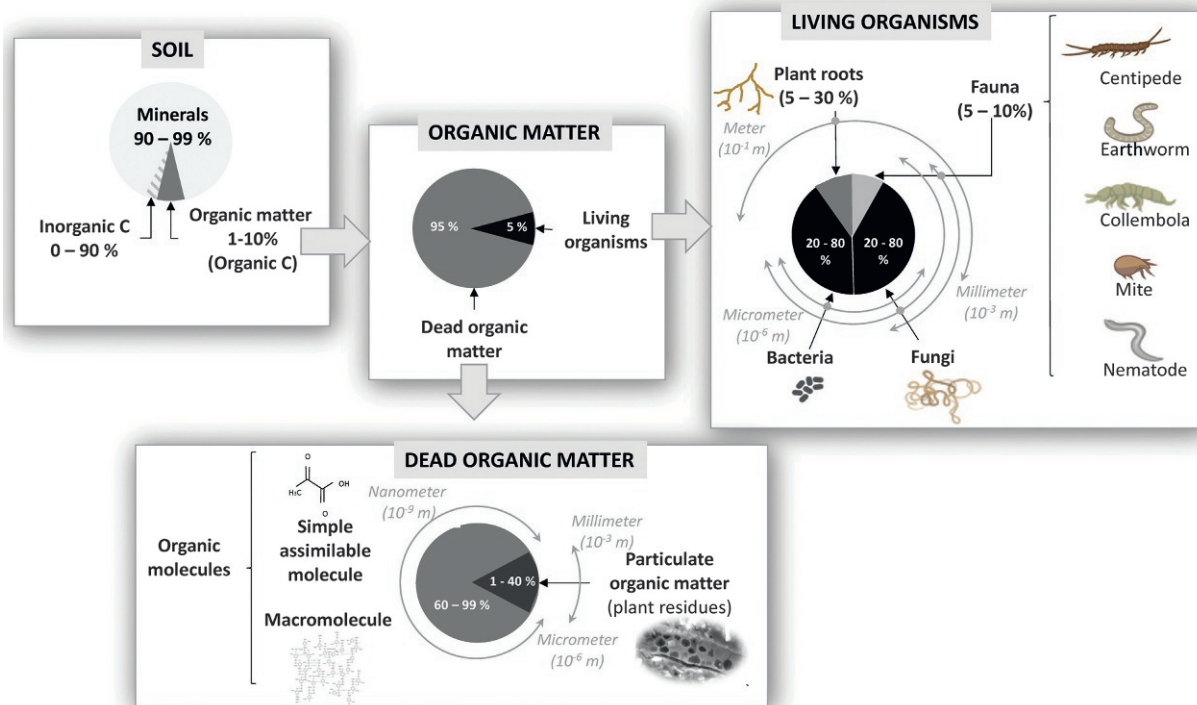


Fig. 1 Nature, size range and indicative proportions of organic matter in the <2 mm soil fraction. Organic matter accounts for less than 10% of soil constituents. The soil C stock primarily consists of dead organic matter molecules that mainly result from microbial activity. Adapted from Basile-Doelsch I, Balesdent J and Pellerin S (2020) Reviews and syntheses: The mechanisms underlying carbon storage in soil. *Biogeosciences* 17: 5223–5242.

on the pedoclimatic context. Fig. 2 shows TOC profiles measured in different soil types. Soil organic carbon (SOC) is vertically distributed, generally with a marked decreasing concentration gradient from the soil surface to depth. Soil C contents can range from 400 gC.kg⁻¹ soil in organic horizons at the soil surface to nearly 100 gC.kg⁻¹ soil in the first cm of the organo-mineral horizon, and down to less than 5 gC.kg⁻¹ at 1 m depth.

The dry bulk density (or apparent density) of fine soil (expressed in g.cm⁻³) is the second measurement parameter required to calculate C stocks. It is defined by the mass of dry soil (105 °C) divided by its volume. The soil volume can be deduced by the immersion of an undisturbed soil sample in water or directly obtained by harvesting samples of undisturbed soil in cylinders of given volume. Contrary to TOC, the soil density increases from surface to depth. The soil density is difficult to measure, especially in soils with a high proportion of coarse elements. This parameter is thus often not available in databases and may be estimated but this adds uncertainty to the C stock calculation.

C stock calculation

The organic carbon stock is the total amount of carbon contained in a given volume of soil. It is usually expressed, for a given depth, in kg per m² or in Mg per ha (t/ha is also used, with 1 kgC.m⁻² = 10 MgC.ha⁻¹ = 10 tC.ha⁻¹). It is calculated by multiplying the TOC by the mass of fine soil contained in the layer.

For a soil layer:

$$\text{SOC stock}_i = \text{BD}_i(1 - \text{CE}_i) \times \text{TOC}_i \times e_i \div 10 \quad (1)$$

For the soil, over the considered thickness:

$$\text{SOC stock}_n = \sum_{i=1}^n \text{SOC stock}_i \quad (2)$$

where i is the considered soil layer, n is the number of layer increments, TOC is the carbon concentration in fine soil (gC.kg⁻¹), BD is the bulk density (g.cm⁻³), CE is the proportion of coarse elements ($0 < \text{CE} < 1$), and e is the layer thickness (cm).

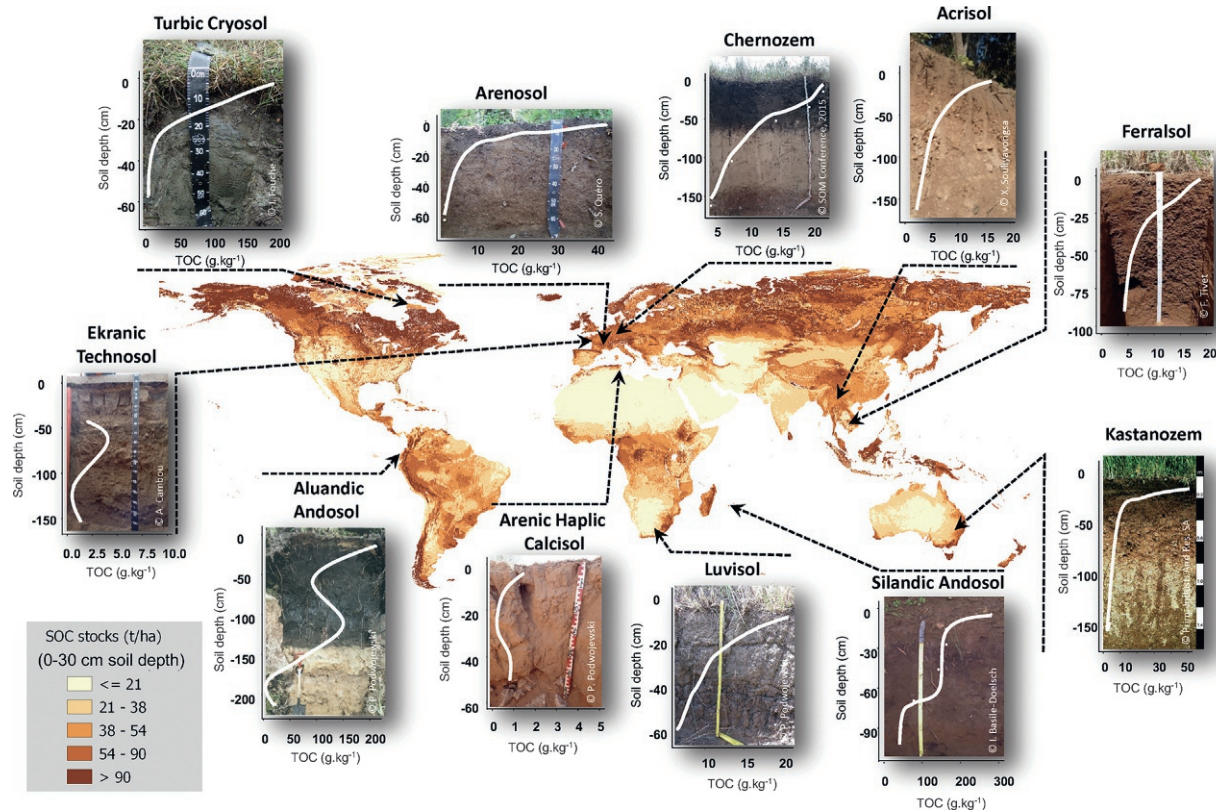


Fig. 2 TOC soil depth profiles for various pedogenesis and C stocks at the global scale. Map source: FAO. 2022. Global Soil Organic Carbon Map—GSOCmap v.1.6. Technical Report. Rome. DOI: <https://doi.org/10.4060/cb9015en>. Data sources (TOC profiles at different soil depth): (i) Turbic Cryosol: Fouche et al., (2014) <https://doi.org/10.1016/j.soilbio.2013.10.007>; (ii) Arenosol: Quérou et al. (2022) <https://doi.org/10.5194/soil-2021-115>; (iii) Chernozem: SOM Conference, 2015; (iv) Aluandic Andosol: J. Poulénard and P. Podwojewski; (v) Arenic Haplic Calcisol: P. Podwojewski; (vi) Luvisol: Podwojewski et al. (2020) <https://doi.org/10.1002/esp.4960>; (vii) Silandic Andosol: I. Basile-Doelsch; (viii) Supracalcic Calcarosol: Hobley, and Wilson (2016) <https://doi.org/10.1002/ecs2.1214>; (ix) Acrisol: Souliyavongsa et al. (2022) <https://doi.org/10.1016/j.geodrs.2021.e00457>; (x) Oxisol: Florent Tivet; (xi) Technosol: Aurelie Cambou.

Soil C stocks can range from a few tonnes of C per hectare to several hundred tonnes per hectare. Fig. 2 shows the distribution of C stocks in soils (0–30 cm) worldwide. C stocks are highly variable and depend on the climate, topography, parent rock mineralogy, vegetation, soil use history and cultivation practices in agricultural soils.

A few noteworthy points when expressing or comparing C stocks in soils:

- A soil C stock is expressed in C, not in tonnes of CO₂ or CO₂ equivalent (tCO₂eq) as is the case for greenhouse gases.
- A soil C stock is a quantity of C for a given soil thickness. The soil thickness must be specified with the C stock measurement. In general, the soil C stock is quantified for a 0–30 cm soil layer (Fig. 2, IPCC reference thickness). Most historical research on C dynamics has been devoted to this soil layer and C stocks in the 0–30 cm soil layer are also usually taken into account in C dynamics models. Soil C stocks also account for deep soils (below 30 cm) but it is seldom measured.
- C stocks in soils under two land uses (or practices) must be rigorously compared for equivalent soil mass (Ellert and Bettany, 1995). The soil mass for a given soil layer thickness may differ between different land uses if the soil tends to be compacted, for example. Therefore, we recommend avoiding comparisons of C stocks at equivalent depth without correction for equivalent soil mass. It cannot be concluded that there is a greater soil C stock for a given land use if the difference noted is the result of the fact that a greater soil mass was taken into account. In practice, at a given depth, if the reference soil mass is the layer with the heaviest mass per unit volume, the correction applied for all cumulative increments is:

$$\text{SOC stock}_{n-\text{corr}} = \text{SOC stock}_n + \left(\text{TOC}_{n+1} \times (1 - \text{CE}_{n+1}) \times \frac{M_{n-\text{heaviest}} - M_n}{10} \right) \quad (3)$$

Here SOC stock_{n-corr} is the corrected cumulative SOC stock (tC.ha⁻¹) down to the depth of layer n; SOC stock_n is the uncorrected cumulative SOC stock (tC.ha⁻¹) down to the depth of layer n (Eq. 2); M_{n-heaviest} is the heaviest cumulative fine soil mass (g.cm⁻²); M_n is the cumulative fine soil mass (g.cm⁻²) and TOC_{n+1} is the fine soil carbon concentration of the underlying layer.

- Mulch or forest litter is a portion of the organic surface horizon that is often excluded from carbon budget calculations, although it is generally several decades old and can represent a significant portion of carbon stocks. H horizons in peatlands or wetlands and grasslands are also often inadequately considered.
- The soil carbon stock is at equilibrium when the inflow—through carbon inputs—is equal to the outflow—through mineralization, loss as dissolved organic carbon or erosion. The soil C stock will remain constant over time when the carbon stock is at equilibrium. Soil carbon is said to be in a steady state if the stock is at equilibrium and if the input flux does not vary over time.

Soil C storage

Carbon storage is the increase in carbon stock over time. It is defined for a given soil, climate, management change, soil depth and time period. It is expressed in $\text{tC}\cdot\text{ha}^{-1}\cdot\text{yr}^{-1}$ or $\text{kgC}\cdot\text{m}^{-2}\cdot\text{yr}^{-1}$. When the storage is negative, there is a C loss. The 4 per 1000 target represents a yearly increase of 0.4% of the C stock (Minasny et al., 2017). For instance, for a soil of $30 \text{ tC}\cdot\text{ha}^{-1}$ (0–30 cm), the target would be $0.12 \text{ tC}\cdot\text{ha}^{-1}\cdot\text{yr}^{-1}$ or $0.012 \text{ kgC}\cdot\text{m}^{-2}\cdot\text{yr}^{-1}$. Storage calculations are often performed over a 20 years period (Chenu et al., 2019).

When comparing two agricultural practices (e.g., Fig. 3, practices B and D), the additional storage related to the agricultural practice implemented in B is the difference between: (i) the carbon stock measured in soil after implementation of the practice over a given period (T_1 - T_2) (plot B), and (ii) the stock of the same soil under a reference practice (plot D) in the same common initial state at T_1 , without any change in practice over the same period (T_1 - T_2). The additional storage is thus defined for two practices, at a given site, and over a time period. A minimum of 5–10 years is generally required to measure the additional storage, but isotopic approaches are being developed to reduce these time lags (Jamoteau et al., 2021).

As for soil C stocks, we add a note of caution regarding several points pertaining to soil C storage estimation:

- The average annual C storage reflects the extent of storage over the observation time and should not be extrapolated to longer durations because the maximum possible storage potential is limited. Extrapolating the observed storage by a linear relationship entails a high risk of overestimation (Chenu et al., 2019).
- When carbon stocks (current or future) are not at equilibrium, the actual storage over time after a practice has been adopted and the additional storage resulting from the practice are separate data items (Fig. 3). Thus, for a situation where the SOC stock is decreasing due to past and current practices (plot C), it is possible that the additional storage following the adoption of a new practice will not fully offset the soil carbon stock downtrend. In this case, the additional storage is positive, yet the effective storage remains negative, which means that the soil continues destocking C, but at a slower rate. The carbon stock evolution trend expected under the current management method is called the 'baseline' (Fig. 3, scenario in plot D). Additional storage is always estimated as the deviation from baseline.
- For a given soil unit, the carbon storage potential is defined as the maximum gain in soil C stock that can be reached under a given climate and time period (Chenu et al., 2019). In practice, it is often hard to identify a reference soil so as to be able to accurately set the target C stock level.
- Storage and sequestration are two concepts slightly distinct although the underlying mechanisms are the same. Soil C storage considers only the increase or decrease of the C stock in the soil (the organic C accumulated from the soil is counted) while the notion of sequestration integrates the atmosphere together with the soil (we refer to the C of CO_2 subtracted from the atmosphere). Soil carbon sequestration is the net removal of CO_2 from the atmosphere that results from the transfer of

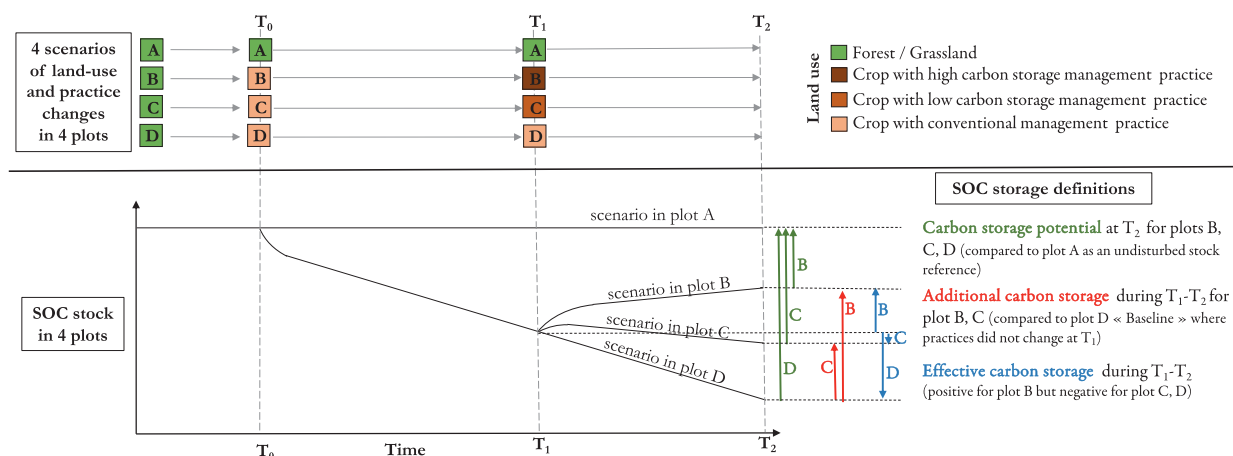


Fig. 3 Theoretical C stock trends as a function of changes in agricultural management practices: differences between carbon storage, additional carbon storage and potential carbon storage. Adapted from Pellerin S, Bamière L, Launay C, et al. (2019) *Stocker du carbone dans les sols français, Quel potentiel au regard de l'objectif 4 pour 1000 et à quel coût? Synthèse du rapport d'étude*, INRAE, p. 114.

atmospheric carbon into soil organic carbon compartments with slow turnover compared to a situation where the CO_2 would remain in the atmosphere. Olson et al. (2014) define soil C sequestration as ‘the process of transferring CO_2 from the atmosphere into the soil of a land unit, through plants, plant residues and other organic solids, which are stored or retained in the unit as part of the soil organic matter. Retention time of sequestered carbon in the soil (terrestrial pool) can range from short-term (not immediately released back to atmosphere) to long-term (millennia) storage.’

The soil C stock is dynamic

Carbon pools in soils are not inert carbon stocks. Soil carbon has a wide range of ages, from a few days to several thousand years old (Fig. 4) (Balesdent et al., 2018). The soil carbon stock is the sum of what remains from each past annual input. It depends on carbon influxes, biotransformation and the stabilization duration before the release of carbon from the soil, mainly as CO_2 produced by the respiration of decomposers. The processes that regulate carbon dynamics in soils are further described below (see Basile-Doelsch et al. (2020) for a detailed description).

Carbon inputs to soil

Organic matter entering the soil system is mainly synthesized by higher plants. It reaches the soil through the roots (dead roots or root exudates), shoot litter, and unharvested aboveground plant parts. The belowground input flux is considered to be the prime contributor to soil organic matter (Rasse et al., 2005). Rhizodeposition represents the contribution of carbon to the soil by living plants via roots. This may include root renewal, release of cells or tissues, macromolecules such as mucilage and extracellular enzymes, or small molecules, i.e., exudates. The rhizodeposition flux is estimated to represent 20–50% of the net root production (Jones et al., 2009). Belowground inputs are still largely unknown, highly variable, and constitute a definite but still relatively unexplored lever that drives soil carbon storage.

Diverse organic molecules are contained in and exuded by plants: cellulose and hemicellulose, lignin, pectin, protein from plant constituents, lipid from waxes and cuticles, tannin, and a multitude of other small molecules (sugar, organic acid) constituting root rhizodeposits (Kogel-Knabner, 2017). In comparison with a structural OM source, small molecules probably have a greater impact

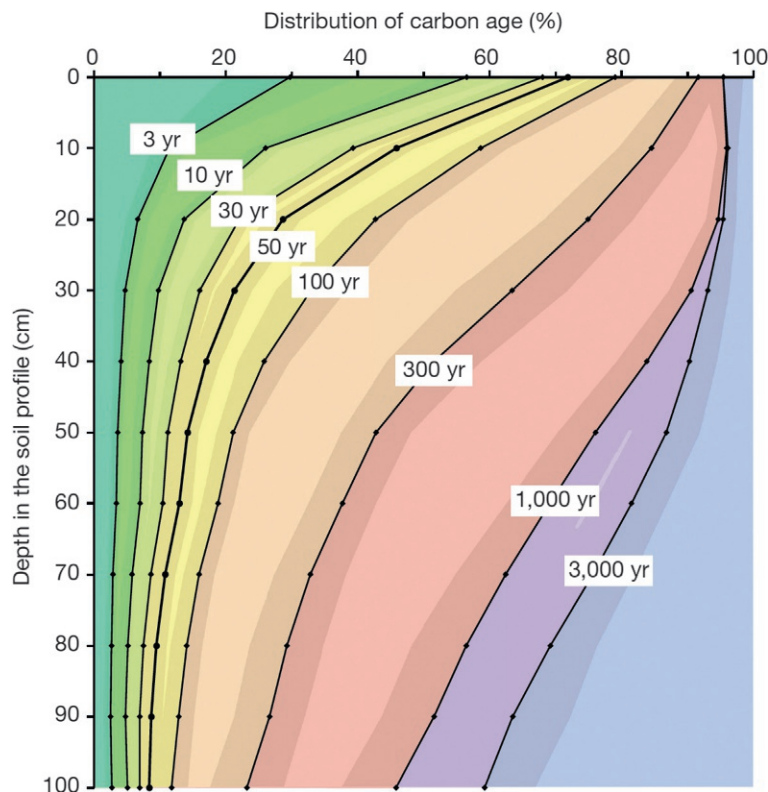


Fig. 4 Meta-analysis of the carbon age distribution in 55 tropical grassland and forest soil profiles. At each depth, black curves indicate the proportion of carbon with an age of less than the specified time (i.e., 3 years, 10 years, etc.). Gray bands represent ± 1 standard error of the estimated mean. From Balesdent J, Basile-Doelsch I, Chadoeuf J et al. (2018) Atmosphere-soil carbon transfer as a function of soil depth. *Nature* 559: 599–602.

on carbon dynamics through their effects on microorganisms and organic matter (OM) mobilization via their priming effect (Keiluweit et al., 2015). Compared to plant OM, microbial products are enriched to different extents in other polysaccharides, lipids, proteins, amino-sugars, nucleic acids, chitin and a very diverse range of metabolites (Kallenbach et al., 2016). Soil organic matter is thus generated from all of these plant or microbial biopolymers or monomers.

Additional inputs may come from organic waste products (e.g., manure from poultry, cattle or pig, composed of mixtures of plant and microbial molecules or their monomers, as well as animal-derived organic compounds), compost and sewage sludge, or byproducts of incomplete combustion or pyrolysis (plant coal, biochar). Plastics may also be present in organic inputs. Moreover, soils may contain geogenic organic carbon, particularly when the parent rocks are organically rich, such as black shale which blackens the soil.

Organic inputs kickstart the soil food chain

Once in the soil, much of the organic matter feeds soil organisms that gradually degrade it. Biochemical reactions that occur during OM decomposition are mainly induced by microorganisms (fungi, bacteria, and archaea) that are soilborne or associated with soil fauna and herbivores. Recent studies have highlighted the close complementarity of all living organisms in the soil with regards to OM transformation:

- Macrofauna (earthworms, termites, ants, etc.) fragment the litter, incorporate it into the soil profile and mix the soil within the profile by bioturbation. Soil transit through the digestive tract of macrofauna (mainly earthworms) promotes contact between microbes and OM. Digestion alters the chemical structure of OM: many groups of soil fauna are thus recognized as stimulating microbial activity and soil organic matter biodegradation. Since some macrofauna organisms (like earthworms) feed on both soil and OM, they can also promote close contact between organic and mineral constituents and enhance the formation of organo-mineral complexes that stabilizes C in soil. Moreover, during bioturbation processes, soil fauna (e.g., earthworms, ants, termites, voles and moles) bury plant residue, gradually mix the soil, or move material between the surface and deep horizons.
- Micro- and meso-fauna (mites, springtails, collembola, tardigrades, protozoa, etc.) form a food web that regulates decomposing microorganisms, e.g., protozoa and bacteria-feeding nematodes tend to decrease the density of microbial populations (Trap et al., 2016).
- Microorganisms are the main drivers of OM chemical biotransformation. They represent the most taxonomically and functionally diverse component of living soil.

Organic matter biotransformation in soil: A carbon recycling business

Biotransformation processes in soil are chemical reactions catalyzed by enzymes produced by living soil organisms, particularly microorganisms.

Organic compound degradation reactions (so-called catabolic reactions) mainly involve hydrolytic or oxidative depolymerization processes (Lehmann and Kleber, 2015) that follow each other according to a continuum model (Fig. 5).

- Incoming plant compounds are mainly large molecules. Due to their size, their depolymerization first takes place outside microbial cells. Co-location between substrates and microorganisms at the microbial habitat scale is essential for the reactions to occur. Substrate-enzyme contact can take place by diffusion and advection of substrates and enzymes, or by microorganism growth (mainly for fungi) and mobility (mainly for bacteria). In addition, local environmental conditions (e.g., oxygenation, pH, water content, temperature) at the micrometer spatial scale must be favorable for microbial activity.
- The action of extracellular enzymes continues until smaller reaction products (sugars, phenolic compounds, amino acids, lipids smaller than around 600 Da) can be transported through the microbial cell membranes. The extracellular nature of reactions has several consequences. First, biodegradation has a high energy cost for organisms (e.g., enzyme transport through cell membranes) and cells have to invest C, N, P and S, whereas some compounds escape from the cells and are diluted in the soil solution or adsorb to other organic or mineral compounds. Small molecules resulting from biodegradation can thus aggregate via weak bonds (hydrogen bonds or hydrophobic interactions) with each other to form supramolecular assemblies (Sutton and Sposito, 2005) or with minerals to form organo-mineral associations (Kleber et al., 2015).
- Small molecular weight organic compounds (organic acids, sugars, amino acids) can be transported into the intracellular environment of microorganisms for further biotransformation. Oxidative degradation can continue until its ultimate stage when the elements are mineralized (CO_2 , NH_4^+ , H_2O , HPO_4^- , SO_4^{2-}). The entire biodegradation chain is shown in Fig. 5.

Unlike oxidative degradation, synthesis of new organic molecules from small organic molecular weight compounds occurs in microbial cells in so-called anabolic reactions. These new molecules, cellular components or excreted metabolites (e.g., organic acids, polysaccharides, extracellular enzymes) contribute to C stocks in soils. C incorporated by microorganisms and re-incorporated into soil OM is repeatedly recycled. A molecule quickly consumed by microorganisms does not necessarily mean that its C will be rapidly mineralized into CO_2 . The most biodegradable compounds have large long-term soil organic matter formation yields (Cotrufo et al., 2013). They display high reactivity that is easily stabilized by association with minerals. Relative to plants, microorganisms are the main producers of organic compounds that are sustainably stabilized through microbial growth, turnover, necromass formation and re-utilization (Kallenbach et al., 2016).

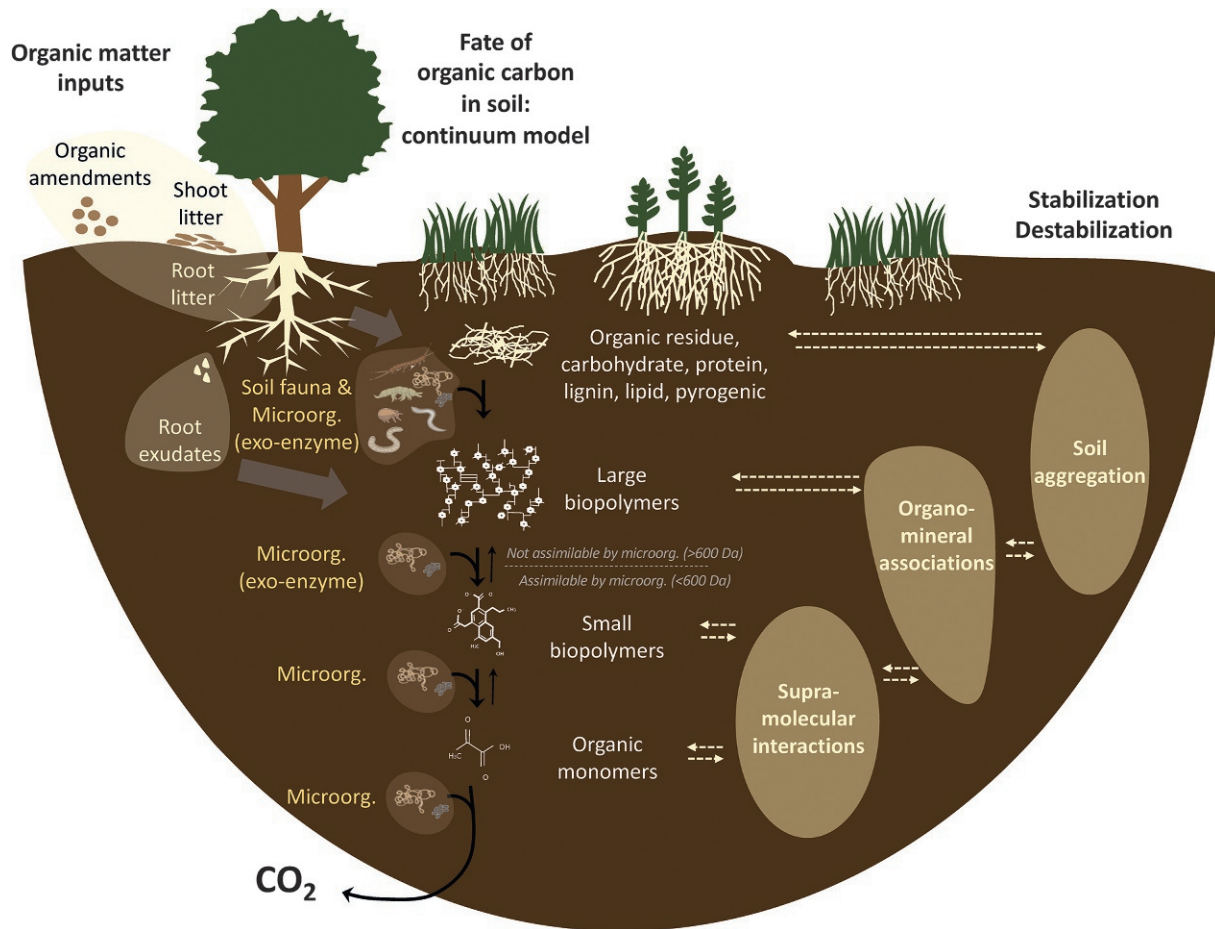


Fig. 5 Illustration of the soil OM biotransformation continuum model (Lehmann and Kleber, 2015). Organic matter enters the soil surface and deeper horizons in different forms. It is continuously degraded from plant and animal debris into tiny molecules by the decomposer community. 600 Da (approximately 1 nm) is roughly the maximum molecule size that can be absorbed by microorganisms. Simultaneously, the increase in the oxidation state of OM carbon increases the water solubility of the compounds. This also enhances their potential protection against further decomposition through increased reactivity with OM (supramolecular associations) and mineral surfaces (organo-mineral interactions) and incorporation into aggregates (aggregation). Solid arrows represent biotic processes and dashed arrows indicate abiotic processes. Adapted from Basile-Doelsch I, Balesdent J and Pellerin S (2020) Reviews and syntheses: The mechanisms underlying carbon storage in soil. *Biogeosciences* 17: 5223–5242.

The so-called ‘priming effect’ is another major process that may affect soil C dynamics and stocks. Here, the supply of complex decomposable substrates (from roots or exogenous inputs) provides competent microorganisms with the energy resources required to biodegrade stabilized OM through various mechanisms (stoichiometric decomposition, co-metabolism or nitrogen mining). Small organic acids of root exudates may also induce biodegradation of stabilized OM through abiotically-mediated priming effects (Bernard et al., 2022). Whatever the mechanisms, the rhizosphere is the main soil compartment concerned with priming.

Organic matter in close interaction with mineral constituents

Soil minerals account for over 90% of soil solid constituents (Fig. 1) and over the last decade organo-mineral associations have become recognized as a key factor in stabilizing soil organic matter. Mineral-associated organic matter (MAOM) is a critical soil nutrient reservoir that accounts for the majority of soil carbon (C) and nitrogen (N) in mineral soils (Kleber et al., 2015). MAOM has historically been considered a relatively passive reservoir of stable OM, but a growing body of research now points to the dynamic nature of mineral-organic associations.

MAOM is represented by aggregates at the micrometer scale (Fig. 5). MAOM aggregation affects the OM dynamics. Aggregates, especially microaggregates (<50 µm), are often used as fractions indicating the ‘degree’ of physical protection of carbon. Conceptual

models describe the C dynamics in the different aggregates by considering the formation-destruction cycles of the aggregates, but parameterizing these models is a complex endeavor.

Studies at the nanometric scale are necessary to gain insight into the mechanisms involved at the interface between mineral phases and organic molecules that drive C storage in soils. Mineral phases involved in organo-mineral associations can be classified into three main categories: well-crystallized minerals, short-range order phases and metallic monomers. Each category is associated with a distinct conceptual organo-mineral association model:

- Adsorption: organic molecules are adsorbed on the mineral surface of well-crystallized minerals (e.g., smectite, vermiculite or goethite). The adsorption model is conceptually the most intuitive.
- Coprecipitation: organic molecules are coprecipitated with short-range order phases or amorphous small polymers of Al, Fe and Si (e.g., allophane, ferrihydrites, amorphous Al hydroxides). They form a loose irregular 3D Al, Si, Fe network skeleton in which organic compounds are linked by various bonds (Tamrat et al., 2019). Co-precipitation would be the most common model in terms of the amount of stabilized C (Rasmussen et al., 2018).
- Complexation and bridging: organic molecules are bound to monomers (single ions). Al and Fe are the main metals involved in complexation in soils with a quite acidic pH. At higher pH, Ca dominates organo-mineral associations by cation bridging (Rasmussen et al., 2018). In both models, cations are surrounded by one or more organic molecules acting as stabilizers of the consortium of small organic molecules forming a so-called 'supramolecular aggregate' tertiary structure.

The existence of these different types of organo-mineral models should thus be considered when assessing MAOM dynamics. A further degree of complexity related to the binding energies should also be considered. Regardless of the model, the five following types of organic-inorganic bonds can be present: covalent bonding, electrostatic cation bridges, hydrophobic interactions, Van der Waals forces and H bonding. The binding strength varies greatly depending on the type of bond. These bonds between organic molecules and inorganic phases tend to slow down the OM degradation process, thereby disturbing the linearity of the biotransformation chain described in Fig. 5.

Part of the MAOM remains accessible to microorganisms but its mineralization is slowed down. The extent of this decrease in the mineralization rate depends both on the nature of the organic compounds and of the minerals. However, some MOAM compounds do not seem to be mineralizable by microorganisms. Thus, rather than using a binary terminology, suggesting that OM is, or is not, stabilized in organo-mineral associations, it would be more rigorous to use a stabilization gradient concept. This gradient is likely a key parameter that controls carbon dynamics in soils and its sensitivity to agricultural management.

Agricultural practices drive the C storage dynamics

Soil management practices affect soil C stocks by modulating the proportion of carbon-containing biomass returned to the soil and by altering the physical, chemical and biological properties of soils through direct or indirect actions (Fig. 6):

- physical modification of the soil surface horizon (tillage, compaction by the passage of machinery, etc.)
- export of plant C produced in the soil (harvesting, etc.)
- selection of crop species (generally these have low biodiversity)
- inputs of mineral products (fertilizers, pH regulators, etc.)
- inputs of exogenous organic matter (organic amendments)
- changes in the soil biological and microbiological activity (indirect effect of land management or use of biofertilizers).

Practices and land use changes with a negative impact on the soil C stock

Some land use changes are known to destock soil C: conversion of natural ecosystems to agricultural or pastoral systems (e.g., peatland drainage), or some agricultural practices (e.g., bare fallow) (Guo and Gifford, 2002). Land use change has resulted in the degradation of over a third of the world's soils (FAO and ITPS, 2015), with C contents greatly decreased compared to their historical values. All practices that maximize plant biomass export, thereby decreasing soil C inputs, prompt a decrease in soil C stocks.

The effects of practices must be assessed at the landscape level, since those implemented locally on a soil can impact the surrounding natural ecosystems. For example, in addition to the effects of climate change, treatments carried out to lower the pH of cultivated soils in Canada have favored the development of invasive earthworm species. These invasive organisms are colonizing boreal forests and causing the mineralization of C stored in the forest floor (Lejoly et al., 2021).

There is expected to be an increase in areas affected by salinity due to global change, with a projected concomitant effect on soil C stocks. Indeed, soil salinity reduces plant growth and thus C input while reducing microbial activity, thereby affecting the C dynamics. The SOC content of saline soils is generally low. Management practices such as phytoremediation, organic matter and biochar amendment have been proposed to rehabilitate saline soils by improving their structure, reducing the osmotic stress for soil organisms, increasing their C content and restoring their fertility.

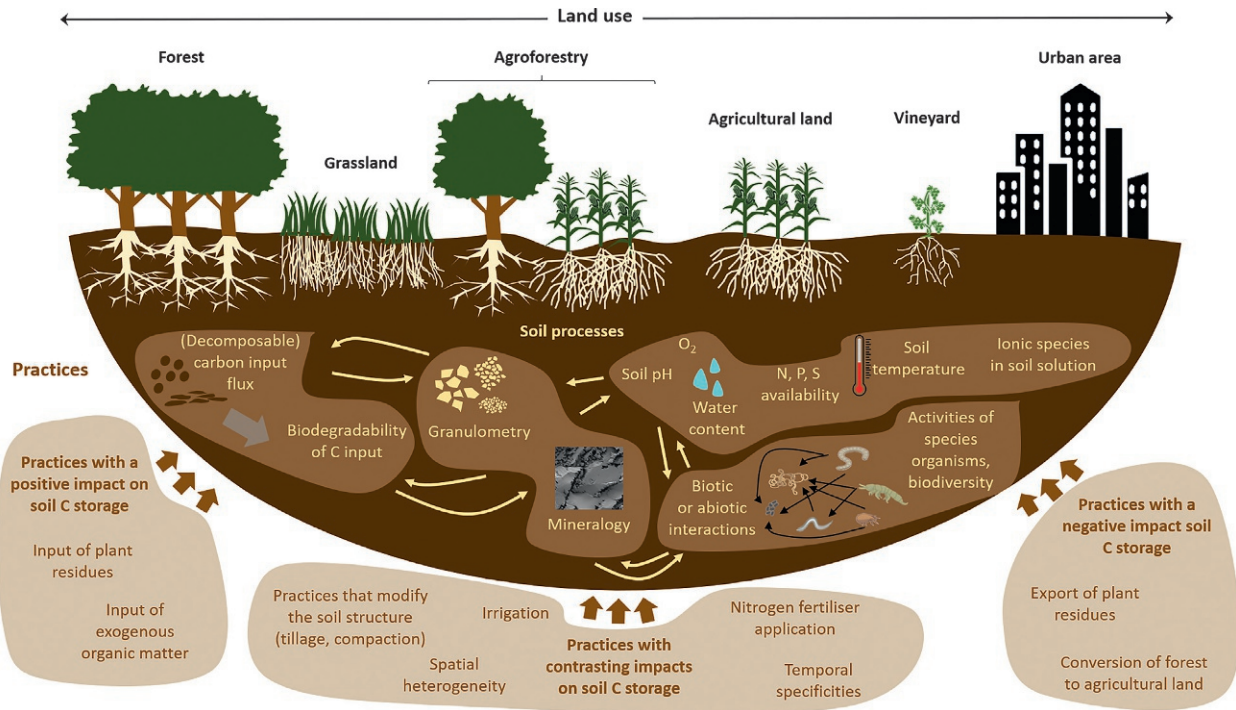


Fig. 6 Different parameters that soil control C stocks: top—land uses; center—biotic and abiotic factors intrinsic to the soil; bottom—management practices that could be favorable or unfavorable to C storage, or have contrasting effects. Adapted from Basile-Doelsch I, Balesdent J and Pellerin S (2020) Reviews and syntheses: The mechanisms underlying carbon storage in soil. *Biogeosciences* 17: 5223–5242.

Practices and land use changes with a positive effect on soil carbon storage

Some land uses and management practices can increase soil C stocks by increasing C inputs or decreasing losses through the protection of soil C (Chenu et al., 2019). For example, land use changes, such as vegetated fallowing, natural state restoration, conversion of arable land to grassland, and practices that avoid bare fallows or increase crop diversification, can promote an increase in soil C stocks (Poeplau and Don, 2015).

In agricultural systems, C inputs can be increased by returning crop residues to the soil, adding exogenous organic matter (urban or agricultural waste compost, digestate from the methanization of agrifood residues used in the bioenergy sector, biochar), or by establishing a permanent soil cover. In grasslands, grazing increases C inputs compared to mowing. In forest, leaving slash scattered on forest floors increases C inputs as compared to clear-felling systems in which this slash is exported. In vineyard, grass-covered inter-rows favor C inputs. Combining several different land-uses on a soil increases the diversity and quantities of C inputs. This can be done, for example, by introducing temporary grass cover in crop rotations or by planting trees in croplands (agroforestry) (Cardinael et al., 2017).

Practices that increase primary production promote C inputs to the soil, either through belowground inputs, or through the aboveground portion that is not exported. Fertilizer application practices and crop selection can affect the soil C dynamics. For instance, the introduction of nitrogen-fixing legumes generates a soil nitrogen supply and can benefit primary production and C stocks. The organic form of nutrients improves the coupling of C, N and P biogeochemical cycles, and the plant nutrient uptake efficiency (e.g., Recous et al., 2019). Crop and intercrop plant selection could be a key way to maximize C inputs (Poeplau and Don, 2015). In particular, plants with large root systems or hyper-exudative roots increase belowground C inputs (see section [Carbon inputs to soil](#)), even if the aboveground parts are harvested and exported. Selecting crop varieties or grassland or forest species with deep root development could be an especially interesting opportunity to increase OM in deep horizons and store C throughout the soil profile.

Practices that limit soil C mineralization improve C stocks. These practices include the addition of mineral amendments, such as lime, that increase the soil pH, or of mining or quarrying waste to soils so as to promote organo-mineral associations, or OM protection within aggregates (Chenu et al., 2019).

Recent research has highlighted the importance of soil organisms in C storage as the activity of soil microorganisms and fauna may promote the formation of organo-mineral associations while safeguarding OM (see section [Organic matter in close interaction with mineral constituents](#)). Adopting and implementing practices that benefit soil organisms in the long term, such as reduced tillage, can thus promote C storage. Yet, paradoxically, soil fauna and microbiological activity contribute to both soil C storage and

mineralization (Lejoly et al., 2021). Research is still necessary to identify other practices that promote C protection through biological activity without increasing its mineralization.

Management practices that increase soil biomass inputs are not only beneficial for soil C stocks. There are many co-benefits, e.g., permanent plant soil cover reduces soil erosion, promoting the coupling of biogeochemical cycles (particularly of N and C) contributes to soil fertility (see section Agricultural practices drive the C storage dynamics). Meanwhile, the extent of the positive effect of the latter practices is very dependent on soil type and properties. Agricultural practices that enhance soil C stocks should therefore be implemented on the basis of local scientific studies focused on site-specific soil and ecosystem properties such as the pH, clay and initial soil C content. These are essential parameters to take into account regarding additional C storage in the form of organo-mineral assemblages (Cotrufo et al., 2019).

Practices with uncertain effects on soil C stocks

Reviews on soil additional C storage after a change of agricultural practices often show great variability in the values. It is likely that C input is the main factor in this variability, but a positive impact is not necessarily sustainable in the long term if the C losses induced become greater than the C input. Studies on the soil priming effect suggest that C addition to soil through readily degradable organic matter (plant residue, compost, digestate, etc.) can stimulate microbial growth and activity, although not increasing the C stock (Bernard et al., 2022). The antagonistic stocking and destocking effects are not yet fully understood and are dependent on the quality of the applied OM, soil depth and pedoclimatic conditions.

Tillage is a practice that has uncertain effects on carbon stocks since it is highly dependent on environmental parameters and soil type. Theoretically, no-till farming can help protect soil C by preserving aggregates and promoting the mixing of plant residues with soil mineral particles. However, the no-till impacts on carbon stocks can be transient or unquantified (Whitehead et al., 2018). Otherwise, no-till agriculture is recommended to control erosion—another source of soil C loss—so it has a positive effect on soil C stocks from this perspective.

Research on the effects of irrigation on C stocks have also produced conflicting results (Whitehead et al., 2018), because irrigation can jointly favor C input (decreased by drought) and C mineralization (slowed down by drought). Irrigation also modifies the plant root:shoot ratio (decreased by drought). Irrigation is assumed to have a positive impact in dry ecosystems and negative or no effect in humid climatic conditions (Whitehead et al., 2018).

Mineral N fertilizer is often applied to croplands and grasslands to increase the production of aboveground biomass, which is usually exported. Yet N fertilizer application can decrease carbon allocation to roots, and thus potentially decrease root carbon inputs to the soil. Since these belowground inputs are crucial for soil C stock accrual, mineral fertilizer application combined with plant export can have a negative effect. Whereas the direct effect of mineral N application on soil C stocks is uncertain, excessive use of mineral N may, in some cases, lead to pollution (nitrate and ammonia leaching) and N₂O emissions (Guenet et al., 2021), which has a global warming potential 261-fold greater than that of CO₂. Moreover, this practice consumes substantial fossil energy (fertilizer production and transport). It would therefore be essential to consider the global greenhouse gas balance of a given practice (e.g., through a life-cycle assessment), not only its impact on soil C stocks. The environmental, climatic and socio-economic repercussions of agricultural practices should be assessed.

Reforestation is often promoted to increase C stocks, but its ecological and socioeconomic impacts are more complex than they appear at first glance (Di Sacco et al., 2021). The tree species composition, stand density and pedoclimatic conditions (Cotrufo et al., 2019), as well as the economic contexts, vary markedly. Reforestation can also have negative impacts, including natural non-forest ecosystem destruction, loss of arable land which may in turn threaten food production, water cycle disruption, increase in invasive species and pollinator threats, decrease in C stored in vegetation and loss of soil organic carbon (Di Sacco et al., 2021). Reforestation must be planned at the landscape level, in a mosaic of land uses, but taking into account the overall needs of society and with the involvement of all stakeholders (business, government, NGOs, scientists, operational agents, landowners, farmers, citizens, etc.). When the economic conditions and the expected positive effects on the soil are satisfied, the reforestation program should favor a mixture of local species capable of mutualistic interactions while excluding invasive species. Safeguarding existing forests remains the priority.

Different urban soil uses and practices, such as unsealed or sealed soils in the vicinity of roads, parks and playgrounds, are poorly studied regarding their carbon stocks. Yet these soils can represent pools of organic carbon but they are highly variable and poorly quantified.

The benefits of soil C storage from the societal viewpoint

Soil organic matter (SOM) is pivotal factor with regard to soil functioning and properties, including soil fertility and biodiversity. SOM decomposition and mineralization contributes to nutrient recycling (N, P, K, etc.), which is essential for plant growth. SOM constitutes energy sources for heterotrophic soil organisms, helps maintain soil structure, contributes—along with clay minerals—to ion exchange capacity, thereby retaining pollutants and contributing to the underground water quality. On a global scale, soils capture more carbon than they release and are considered to be carbon sinks and contribute to climate regulation. Indeed, most of the terrestrial C (2000–3000 Gt C, Gt = gigatons or billions of tonnes of C) is in soils, 1500 Gt C of which occurs in the first meter of soil, compared to 830 Gt C in the atmosphere and 450–650 Gt C in plant biomass. As seen above, SOM and C

stocks are not homogeneous. Different organic pools, with different properties and dynamics, are involved in the soil C cycle. The most stable compartments associated with fine soil fractions (clay and fine silt) have a small role in nutrient cycling and agricultural production in the short term, but are more involved in long-term carbon stabilization (climate objective), in contrast to the particulate organic matter compartments (food security objective).

Maintaining organic matter levels in soils is beneficial for a range of ecosystem services that soil provides to our societies. Soil carbon and the land uses and management that affect its stocks are thus pivotal to many of the United Nations Sustainable Development Goals (SDG), including SDG #2 'Zero Hunger,' #13 'Climate Action,' #15 'Life on Land,' and #12 'Responsible Consumption and Production' (Fig. 7).

SOM losses, especially in dry regions or sandy soils, where initial levels are small, invariably results in the degradation of soil and associated functions, e.g., agricultural production. This often triggers a vicious degradation loop (Lal, 2004). Conversely, increasing SOM improves soil condition and fertility, thereby contributing to ecosystem and agroecosystem resilience and sustainability. The three international environmental conventions on Climate Change (UNCCC), Biological Diversity (UNCBD) and Desertification (UNCCD) have thus adopted the soil organic C stock as a land and soil quality monitoring indicator.

Knowledge of soil carbon stocks and dynamics is often considered to be a soil and agronomic research matter. This is no longer strictly the case and interdisciplinary research has been developed on all C fluxes and their implications for agriculture, energy and fiber production, socio-economics, and the climate. Increasing or maintaining soil carbon stocks is a global objective, but carbon storage can only be achieved locally through a multitude of soil uses and practices, and involve various stakeholders. These

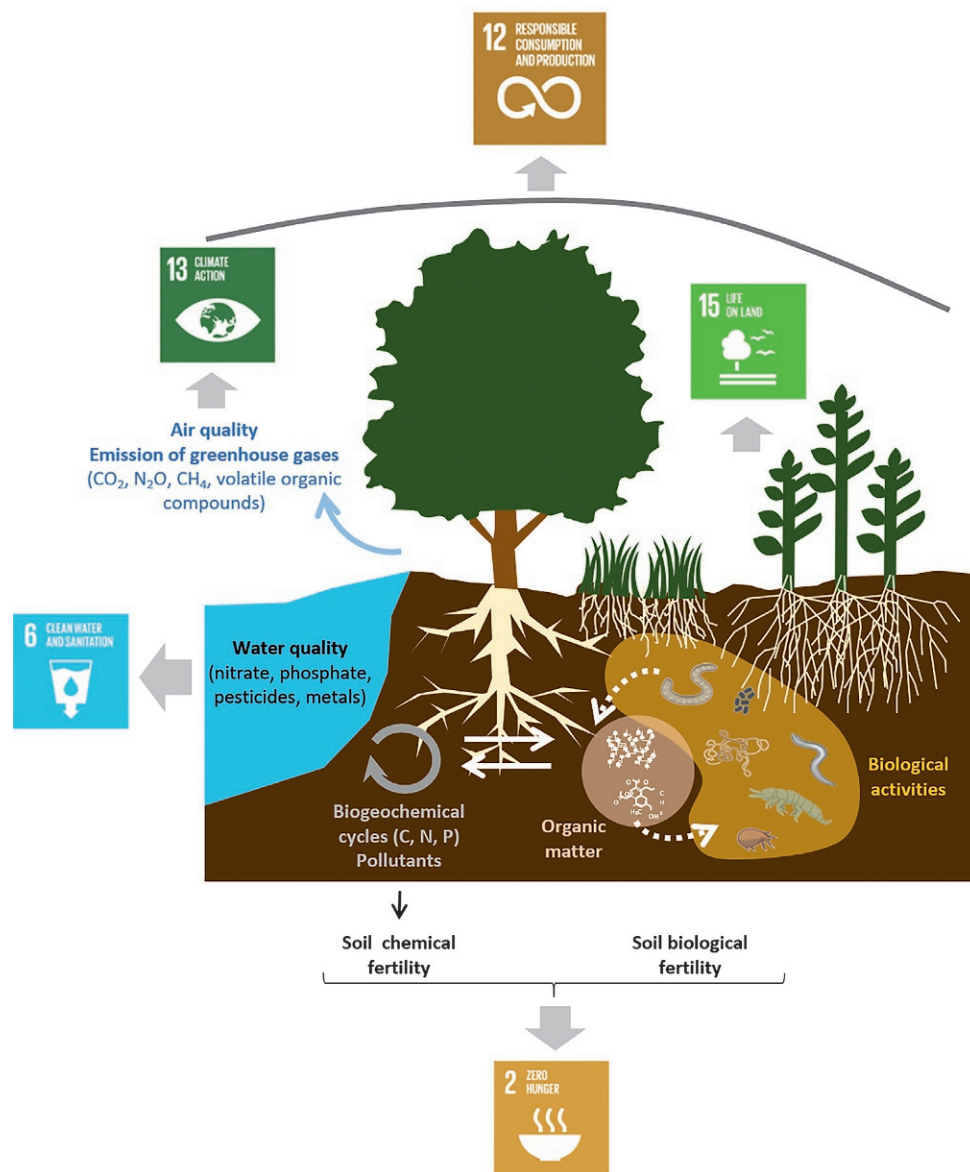


Fig. 7 The benefits of soil C storage for soil functioning and ecosystem services and for the UN Sustainable Development Goals.

initiatives are sometimes associated with the transfer of organic resources (e.g., organic waste, biomass from crop residues, forage, mulch, forest products and manure) between different parts of a territory: such as from urban areas to periurban market gardening areas, from grazing areas to cultivated areas. Agricultural practices and land-use planning, such as sharing land for forestry, crop or livestock farming with their different extents of integration (e.g., agroforestry), have an impact on C fluxes and stocks. Programs aimed at increasing soil carbon stocks must therefore not only consider practices at the agricultural field level, but also the organization of the farm, territory, supply chains and markets to be effective, suitable and sustainable. Their multiple impacts, trade-offs between uses and technical, socio-economic and cultural limits, as well as the motives of the different stakeholders, must be taken into account.

Carbon storage in soils affects all activity sectors and can be considered at levels ranging from mineral particles that stabilize C in soils to large areas governed by public policies that regulate the use of agricultural, natural and urban spaces, and even international governance that can foster a public policy or funding in favor of soil C (Chevallier et al., 2020). Various disciplines (e.g., geography, agronomy, rural economics, sociology) are therefore necessary to understand and manage these C stocks and fluxes.

To reconcile the interests and constraints of each stakeholder with the C storage biophysical and technical parameters, socio-economic scenarios have to be developed to predict C stock variation patterns and their implications for people. These modeled biophysical, economic and sociological assessments are the basis for designing public policies adapted to the context. However, these assessments are still seldom performed.

Biophysical assessments concern C storage as well as all greenhouse gas emissions (especially CH₄ and N₂O) related to land use and management. There are still uncertainties regarding these complete greenhouse gas budgets in all contexts that research is striving to overcome through modeling. These uncertainties arise from the lack of data in some contexts. The diversity of soil type, climatic conditions and agricultural practices complicates the assessment of all impacts on C stocks and GHG emissions. In-depth characterizations of the context through both laboratory measurements and field surveys are therefore necessary to specify the local C storage potential. Factors responsible for the variability in these potentials must be identified to develop spatiotemporal extrapolation methods for monitoring and predicting C stock variation trends. These factors are, for example, climate, crop production, soil type, soil texture, and the amount of crop residues returned to the soil. C stock variation patterns must be evaluated both with and without changes in land use or agricultural practices to accurately assess the effect of the practice on C stocks and budgets (Fig. 3).

Therefore, it is essential to develop databases and methods that quantify the effects of changes in agricultural use or practice on soil carbon storage. Such a quantification and monitoring system would help to recognize and finance the service provided by agriculture to mitigate climate change. Indeed, one of the main challenges of the implementation of C storage or sequestration in contrasting environments is its social and economic acceptability and feasibility. This C sequestration objective cannot be disconnected from other issues and requires effort and investment at the same time as involving risk. Different parties could take on these investments and risks: the producer, the landowner or the society or community through subventions or guarantee mechanisms. Indeed, efforts to enhance C sequestration may not be attractive for a producer, who is not the landowner and who only benefits from the land in the short term. The benefits from a farmer's efforts are sometimes only perceptible in the medium or long term (e.g., degraded land restoration, agroforestry). Several studies have revealed that the technical potential for soil C sequestration may be high but its implementation limited by the absence of economic incentives to support the adoption of techniques conducive to C storage. In France, Pellerin et al. (2019) showed that additional SOC storage in French agricultural soils is limited to 0.66 MtC/yr in the absence of economic incentives, while its technical (or biophysical) potential is 8.43 MtC/yr. The cost of these economic incentives varies greatly between studies and countries, from around 5 euros per tCO₂ (Rakotovo et al., 2021) to around 100 euros (Soussana et al., 2019).

Assigning an economic value to SOC and bolstering the C market could be an effective incentive to improve SOC sequestration (Amelung et al., 2020). In agriculture, CO₂ certificates for sequestered soil carbon are being developed in recognition of farmer effort (Demenois et al., 2021). There are two main types of certification: those based on results achieved, i.e., quantification of GHG emission reductions (in CO₂ equivalents) and those based on the means implemented to reduce GHG emissions, i.e., recognition of the implementation of sequestering practices. The first type places the responsibility for implementing the practices on the farmers or foresters who initiate them, while the second type places the responsibility for implementing the practices on the certification authority. Implementation of the first type is more challenging since it requires actual C storage measurement and verification. It may be perceived as 'safer' and fairer but has little incentive. The second option seems simpler and may be more attractive if they are associated with a clear description of the specifications of the practices to be implemented. The second option also recognizes all ecosystem services that are enhanced by practice changes and C storage, not just their effects on GHG emissions and the climate.

Conclusion

Soil carbon storage is a natural phenomenon related mainly to organic forms of soil carbon. Atmospheric CO₂ captured by vegetation is integrated into the soil as organic matter, which represents 50–60% of all soil organic carbon. Although soils and vegetation emit CO₂, they serve as a C sink and host major stocks of carbon worldwide. Soils are the largest terrestrial reservoir of organic carbon in contact with the atmosphere. Soils and soil organic matter are key compartments in the global C cycle. Their

management may contribute to limiting greenhouse gas emissions, while also being crucial in combating desertification and enhancing soil biological activity and biodiversity.

Organic C present in soils in the form of organic matter has many roles (e.g., soil fertility, soil structure, retention of mineral elements and pollutants). Soil organic C is consumed by soil organisms and thus is a dynamic compartment that promotes soil biological activity. The organic matter mineralization that results from this biological activity releases nutrients that are essential for plant growth. The balance between organic matter stabilization and mineralization depends on processes that impact soil biological activity in the short or long term. These processes depend on the climate, soil type, land use and management practices. At the global scale, soil C stocks vary from 5 to 350 tC ha⁻¹ depending on pedoclimatic conditions and soil management practices. Soil use and management can lead to non-negligible variations in soil C storage. The eventual negative variations under cultivated soils are to avoid soil erosion and soil nutrient depletion. The positive variations, i.e., C storage, varies from 0 to about 2 tC ha⁻¹ years⁻¹, contribute to preservation of soil functioning, favor soil sustainable crop productivity and may help attenuate negative impacts of the global climate change.

There are major differences between current, additional and potential soil C storage. We have outlined the methods and highlighted potential pitfalls in measuring and calculating soil C stock and storage. Carbon stocks are regularly measured directly in soil samples and must be tailored to the spatiotemporal heterogeneity of the target soil, but such point measurements cannot be used to assess global patterns or be frequently repeated due to cost and time limitations. Spatial and temporal extrapolation methods are used to monitor and predict changes in soil C stocks in response to changes in land use and agricultural practices. Considering the interests and constraints of each stakeholder in the application of carbon storage techniques by developing socio-economic scenarios is essential. Prediction models and contextually appropriate public policies could be further designed based on biophysical, economic and sociological assessments.

Soil C data and knowledge on are sometimes partial or lacking, due to the heterogeneity of soils, varying climatic conditions, soil types and the diverse potential range of soil cover, uses and management. Long-term studies and data compilation are essential to gain insight into the drivers of the soil-C storage. Ongoing long-term, multidisciplinary studies are needed to better understand this complex issue. Nevertheless, initiatives geared toward boosting soils and soil C stocks must not be hampered or paralyzed by this complexity. Solutions that can be implemented are already available. In this chapter, we have reviewed land-use changes and practices that promote soil C storage and those that could lead to soil C loss. There are also practices whose impacts are uncertain because they depend on the pedoclimatic context. There is no single universal solution and choices to store or maintain C in the long term must be adapted to the local context.

Acknowledgments

We thank Sylvain Pellerin and Jerome Balesdent for discussions and exchanges during the INRAE 4 per 1000 study (Pellerin et al., 2019) which served as a guide in writing the first parts of the chapter. We are also thankful to the French network CarboSMS (Carbon Stabilization Mechanisms in Soil), which provided a framework for discussion and exchange on soil C storage. We thank Christian Hartmann, Florent Tivet and Aurélie Cambou, Pascal Podwojewski and Julien Fouché for providing images of soil profiles and associated data on the TOC concentration with soil depth. We thank Y. Yigini and M. Mainka for guiding us on how to use the latest version of the FAO Global Soil Organic Carbon Map.

References

- Allory V, Cambou A, Moulin P, et al. (2019) Quantification of soil organic carbon stock in urban soils using visible and near infrared reflectance spectroscopy (VNIRS) in situ or in laboratory conditions. *Science of the Total Environment* 686: 764–773.
- Amelung W, Bossio D, de Vries W, et al. (2020) Towards a global-scale soil climate mitigation strategy. *Nature Communications* 11: 5427.
- Balesdent J, Basile-Doelsch I, Chadoeuf J, et al. (2018) Atmosphere-soil carbon transfer as a function of soil depth. *Nature* 559: 599–602.
- Basile-Doelsch I, Balesdent J, and Pellerin S (2020) Reviews and syntheses: The mechanisms underlying carbon storage in soil. *Biogeosciences* 17: 5223–5242.
- Bernard L, Basile-Doelsch I, Derrien D, et al. (2022) Advancing the mechanistic understanding of the priming effect on soil organic matter mineralisation. *Functional Ecology* 36.
- Cardinael R, Chevallier T, Cambou A, et al. (2017) Increased soil organic carbon stocks under agroforestry: A survey of six different sites in France. *Agriculture, Ecosystems & Environment* 236: 243–255.
- Chenu C, Angers DA, Barré P, et al. (2019) Increasing organic stocks in agricultural soils: Knowledge gaps and potential innovations. *Soil and Tillage Research* 188: 41–52.
- Chevallier T, Loireau M, Courault R, et al. (2020) Paris climate agreement: Promoting interdisciplinary science and stakeholders' approaches for multi-scale implementation of continental carbon sequestration. *Sustainability* 12.
- Cotrufo MF, Wallenstein MD, Boot CM, et al. (2013) The microbial efficiency-matrix stabilization (MEMS) framework integrates plant litter decomposition with soil organic matter stabilization: Do labile plant inputs form stable soil organic matter? *Global Change Biology* 19: 988–995.
- Cotrufo MF, Ranalli MG, Haddix ML, et al. (2019) Soil carbon storage informed by particulate and mineral-associated organic matter. *Nature Geoscience* 12: 989–994.
- Demenois J, Dayet A, and Karsenty A (2021) Surviving the jungle of soil organic carbon certification standards: An analytic and critical review. *Mitigation and Adaptation Strategies for Global Change* 27: 1.
- Di Sacco A, Hardwick KA, Blakesley D, et al. (2021) Ten golden rules for reforestation to optimize carbon sequestration, biodiversity recovery and livelihood benefits. *Global Change Biology* 27: 1328–1348.
- Ellert BH and Bettany JR (1995) Calculation of organic matter and nutrients stored in soils under contrasting management regimes. *Canadian Journal of Soil Science* 75: 529–538.
- FAO and ITPS (2015) Status of the World's Soil Resources (SWSR) – Main Report. In: *Food and Agriculture Organization of the United Nations and Intergovernmental Technical Panel on Soils*, Rome, Italy. <http://www.fao.org/3/a-i5199e.pdf>.
- Guenet B, Gabrielle B, Chenu C, et al. (2021) Can N2O emissions offset the benefits from soil organic carbon storage? *Global Change Biology* 27: 237–256.

- Guo LB and Gifford RM (2002) Soil carbon stocks and land use change: A meta analysis. *Global Change Biology* 8: 345–360.
- Jamoteau F, Balesdent J, Basile-Doesch I, et al. (2021) Can stable isotopes quantify soil carbon build-up from organic fertilizers? *Isotopes in Environmental and Health Studies* 57: 470–491.
- Jones DL, Nguyen C, and Finlay RD (2009) Carbon flow in the rhizosphere: Carbon trading at the soil-root interface. *Plant and Soil* 321: 5–33.
- Kallenbach CM, Frey SD, and Grandy AS (2016) Direct evidence for microbial-derived soil organic matter formation and its ecophysiological controls. *Nature Communications* 7.
- Keiluweit M, Bougoure JJ, Nico PS, et al. (2015) Mineral protection of soil carbon counteracted by root exudates. *Nature Climate Change* 5: 588–595.
- Kleber M, Eusterhues K, Keiluweit M, et al. (2015) Chapter One—Mineral–organic associations: Formation, properties, and relevance in soil environments. In: Donald LS (ed.) *Advances in Agronomy*, pp. 1–140. Academic Press.
- Kogel-Knabner I (2017) The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter: Fourteen years on. *Soil Biology & Biochemistry* 105: A3–A8.
- Lal R (2004) Soil carbon sequestration impacts on global climate change and food security. *Science* 304: 1623–1627.
- Lehmann J and Kleber M (2015) The contentious nature of soil organic matter. *Nature* 528: 60–68.
- Lejoly J, Quideau S, and Laganère J (2021) Invasive earthworms affect soil morphological features and carbon stocks in boreal forests. *Geoderma* 404: 115262.
- Minasny B, Malone BP, McBratney AB, et al. (2017) Soil carbon 4 per mille. *Geoderma* 292: 59–86.
- Olson KR, Al-Kaisi MM, Lal R, and Lowery B (2014) Experimental consideration, treatments, and methods in determining soil organic carbon sequestration rates. *Soil Science Society of America Journal* 78: 348–360.
- Pellerin S, Bamière L, Launay C, et al. (2019) *Stockage du carbone dans les sols français, Quel potentiel au regard de l'objectif 4 pour 1000 et à quel coût ? Synthèse du rapport d'étude*. INRAE114.
- Poepplau C and Don A (2015) Carbon sequestration in agricultural soils via cultivation of cover crops—A meta-analysis. *Agriculture, Ecosystems & Environment* 200: 33–41.
- Rakotovoah NH, Chevallier T, Chapuis-Lardy L, et al. (2021) Impacts on greenhouse gas balance and rural economy after agroecology development in Itasy Madagascar. *Journal of Cleaner Production* 291: 125220.
- Rasmussen C, Heckman K, Wieder WR, et al. (2018) Beyond clay: Towards an improved set of variables for predicting soil organic matter content. *Biogeochemistry* 137: 297–306.
- Rasse D, Rumpel C, and Dignac M-F (2005) Is soil carbon mostly root carbon? Mechanisms for a specific stabilisation. *Plant and Soil* 269: 341–356.
- Recous S, Lashermes G, Bertrand I, et al. (2019) C–N–P decoupling processes linked to arable cropping management systems in relation with intensification of production. In: Lemaire G, Carvalho P, Kronberg S, and Recous S (eds.) *Agroecosystem Diversity*, pp. 35–53. Academic Press.
- Soussana J-F, Lutfalla S, Ehrhardt F, et al. (2019) Matching policy and science: Rationale for the '4 per 1000—Soils for food security and climate' initiative. *Soil and Tillage Research* 188: 3–15.
- Sutton R and Sposito G (2005) molecular structure in soil humic substances: The new view. *Environmental Science and Technology* 39: 9009–9015.
- Tamrat WZ, Rose J, Grauby O, et al. (2019) Soil organo-mineral associations formed by co-precipitation of Fe, Si and Al in presence of organic ligands. *Geochimica et Cosmochimica Acta* 260: 15–28.
- Trap J, Bonkowski M, Plassard C, et al. (2016) Ecological importance of soil bacterivores for ecosystem functions. *Plant and Soil* 398: 1–24.
- Whitehead D, Schipper LA, Pronger J, et al. (2018) Management practices to reduce losses or increase soil carbon stocks in temperate grazed grasslands: New Zealand as a case study. *Agriculture, Ecosystems & Environment* 265: 432–443.

Relevant websites

The international "4 per 1000" initiative: <https://4p1000.org> <https://regenerationinternational.org/4p1000/>
 Soil C storage potential in France <https://www.inrae.fr/en/news/storing-4-1000-carbon-soils-potential-france>
 Soils for food security and the climate: <https://www.inrae.fr/en/news/soils-food-security-and-climate>
 Advocating for Soil health and Climate at COP26 <https://www.inrae.fr/en/news/advocating-soil-health-and-climate-inrae-contributions-cop26>
 FAO soils portal <https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/global-soil-organic-carbon-map-gsocmap/en/>

Soil organic nitrogen

EJ Daly^a and KA Congreves^b, ^aDepartment of Renewable Resources, University of Alberta, Edmonton, AB, Canada; ^bDepartment of Plant Sciences, University of Saskatchewan, Saskatoon, SK, Canada

© 2023 Elsevier Ltd. All rights reserved.

Introduction	243
Soil organic nitrogen forms and properties	244
Soil organic nitrogen cycling	246
Depolymerization and microbial cycling of dissolved organic nitrogen	247
Soil organic matter persistence	247
Plant uptake of soil organic nitrogen	248
Dissolved organic nitrogen leaching	249
Sampling, identification, and quantification of organic nitrogen	249
Quantification of total organic nitrogen	249
Sampling dissolved organic nitrogen	250
Fractionation, identification, and quantification of organic nitrogen compounds	250
Soil organic nitrogen: An emerging component of soil condition in managed systems	251
Conclusion	251
References	251

Abstract

Soil organic nitrogen (N) is a paramount component of the terrestrial N cycle. Knowledge of soil organic N can inform a greater understanding of soil biogeochemical relationships, mitigate environmental degradation, and improve soil conditions in managed systems. This chapter provides an overview of soil organic N forms and classification, the cycling of organic N within soil, plant uptake, leaching, approaches to identify and quantify soil organic N, together with a discussion of how future research on soil organic N can be applied to the enhancement of agricultural productivity and sustainability.

Glossary

Amino acid Organic compounds that contain both amino (NH₂) and carboxylic acid (COOH) functional groups.

Biologically active N Nitrogen that is actively cycled by biogenic processes.

Monomer A simple molecule capable of covalent bonding with other monomers to form larger molecules.

Oligomer A molecule that consists of a few repeating monomers.

Oligopeptide A peptide comprised of a few amino acids.

Peptide A compound consisting of at least 2 amino acids linked by covalent bonds formed from the carboxyl group of one amino acid to the amino group of another, known as a peptide bond.

Polymer A macromolecule consisting of many repeating subunits.

Polypeptide An organic polymer consisting of many amino acids.

Reactive N Biologically active N. Includes all N forms except inert N₂.

Key points

- Soil organic N is a highly heterogeneous mixture containing many unique compounds.
- Depolymerization is the predominant pathway by which microbes and plants utilize organic N.
- The dynamics of soil organic matter are strongly linked to soil organic N transformations.
- Improved understanding of soil organic N cycling in natural and managed ecosystems may mitigate environmental issues.

Introduction

The terrestrial nitrogen (N) cycle includes soil, plant, and animal pools that contain relatively small amounts of biologically active N, in comparison with the substantial pools of inert N in the lithosphere and atmosphere. Nevertheless, the terrestrial N cycle and the transformations of N forms therein are integral to the global biogeochemical N cycle, as N is essential for life as a building block of DNA, RNA, chlorophyll, proteins, and enzymes. Nitrogen is commonly the nutrient most limiting primary productivity.

Despite the importance of N to all life, too much N in reactive forms triggers negative climatic and environmental effects. For example, over the past half century, reactive N inputs have accrued to unprecedented levels, associated directly with the addition of synthetic inorganic N fertilizer to agricultural soils and indirectly through increased atmospheric N deposition (Galloway et al., 2008). Consequently, the processing of augmented levels of reactive N has increased greenhouse gas production, the release of N in the form of N-oxides, and leaching losses of N to ground and surface water. Much of this N processing takes place in soils; as such, cycling of soil N, when functioning tightly, is a key ecosystem service that supports the terrestrial and global N cycles; but when producing excessive N losses, provides an ecosystem disservice.

Over 80% of N in soil is in the organic form, and the largest terrestrial reservoir of organic N is soil organic matter (SOM), which is a highly heterogeneous mixture of N-containing compounds that include proteins, amino acids, amino sugars, nucleosides, peptides, and some phospholipids (Warren, 2014a). Traditionally, research of the soil N cycle was largely focused on inorganic N forms (ammonium $[\text{NH}_4^+]$ and nitrate $[\text{NO}_3^-]$). This was motivated by the prominent role of the inorganic N in agricultural management and thus the perceived dependency of agricultural production on this N pool (Näsholm et al., 2009). However, a paradigm shift in our understanding of the terrestrial N cycle has questioned the common assumptions underlying N management that plants *solely* utilize inorganic N sources, and instead are capable of utilizing a wide variety of N sources, including organic N forms, in addition to inorganic N (Schimel and Bennett, 2004). Thus, the soil N cycle is now viewed as being driven by the depolymerization of high molecular weight organic N to bioavailable organic N monomers that can be directly utilized by microbes and plants, in addition to utilizing soil inorganic N. For agronomically important plants, this may present an opportunity to transition away from the currently exclusive reliance on synthetic N fertilizers (Farzadfar et al., 2021). Further, significant improvements in our understanding of the nature of SOM, which is a pillar of the soil condition and an important component of climate change mitigation, have highlighted the importance of organic N for primary productivity and as a component of persistent, mineral-associated organic matter (Cotrufo and Lavelle, 2022; Lehmann and Kleber, 2015). As such, a greater understanding of organic N pools and cycling is key to better understanding the functioning of natural ecosystems, improved management of agricultural systems, and stabilization of SOM stocks.

Soil organic nitrogen forms and properties

Soil organic compounds are described on a continuum from intact organic material to large biopolymers, small biopolymers, and monomers (Lehmann and Kleber, 2015). As a component of SOM, most soil organic N is in amide bonds, the bonds present in protein molecules, with a lesser proportion in amine and heterocyclic N bonds. The vast majority (>90%) of soil organic N is sorbed onto soil minerals or protected within aggregates, whereas a small portion of organic N (1–4%) exists as a component of particulate organic matter (POM), which is a transitory pool in the decay continuum of SOM (Warren, 2014a). Dissolved organic nitrogen (DON; 0.1–2%) represents the soluble N in the soil solution, the exact chemical composition of which is controlled by soil organic N sources and varies with soil type, land use, and environmental conditions.

Functional and ecological differences in DON molecules are directly or indirectly related to molecular weight. Based on molecular weight, DON can be broadly divided into three major classes: high molecular weight (HMW), medium molecular weight (MMW) and low molecular weight (LMW) (Fig. 1). The LMW fraction of DON is cycled within hours while the MMW and HMW fractions are cycled much more slowly (~days to months) (Warren, 2014a).

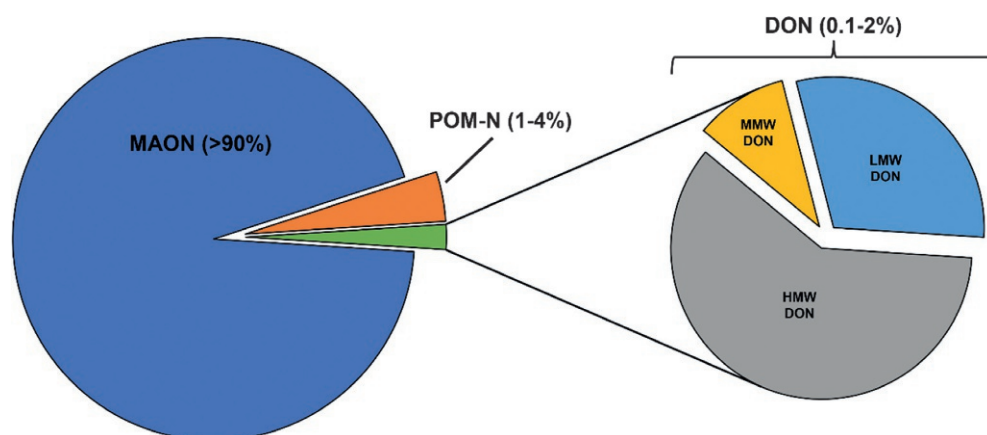


Fig. 1 The approximate distribution of soil organic nitrogen (N). MAON: mineral associated organic nitrogen; POM-N: particulate organic matter nitrogen; DON: dissolved organic nitrogen; HMW DON: high molecular weight dissolved organic nitrogen >100 kDa; MMW DON: medium molecular weight dissolved organic nitrogen from 1 to 100 kDa; LMW DON: low molecular weight dissolved organic nitrogen <1 kDa. Adapted from Warren CR, (2014a) Organic N molecules in the soil solution: What is known, what is unknown and the path forwards. *Plant and Soil*, 375(1): 1–19.

The HMW class is composed of oligomers and polymers of amino acids with molecular weights >100 kDa. Most soil DON falls into the HMW class, of which proteins comprise the majority; up to 99.5% of amino acids in soil are contained in HMW polymeric N compounds, such as proteins and peptides. The HMW class also includes smaller amounts of bacterial cell wall-derived peptidoglycan and fungal-derived chitin, the composition of which varies between species. Typically, peptidoglycan is composed of *N*-acetylglucosamine and *N*-acetylmuramic acid cross-linked by peptide bonds of 3–5 amino acids attached to *N*-acetylmuramic acid (Fig. 2). Polymeric linkages of *N*-acetylglucosamine are the basis of chitin.

The MMW class is the smallest pool of DON, consisting of peptides and proteins of molecular weights ranging from 1 to 100 kDa. The LMW class is comprised predominantly of protein amino acids, with smaller quantities of non-protein amino acids, quaternary ammonium compounds (QACs), amino sugars, purine and pyrimidine bases, nucleic acids, and short oligopeptides weighing <1 kDa (Fig. 3).

As the major component of the LMW class, all 20 protein amino acids have been identified in the DON fraction of soils, and common non-protein amino acids include gamma aminobutyric acid (GABA), citrulline, and ornithine. Amino acids can be present as *L*- or *D*-enantiomers, i.e., as stereoisomers that are mirror images of each other, based on their atomic arrangement. The *D*-amino acids are less abundant compared to their *L*- counterparts, however *D*-amino acids can occur naturally in bacterial cell walls, microbial metabolites, antibiotics, eukaryotic biomass, and from the biotic or abiotic racemization of *L*-amino acids. Compared to their *L*- counterparts, *D*-amino acids may undergo slower mineralization rates in soil.

Also abundant in the LMW pool are QACs, which include proline, proline betaine, ectoine, hydroxyectoine, and hercynine. Most QACs are product of microbial synthesis and may account for 1–28% of non-peptide compounds in the LMW organic N class.

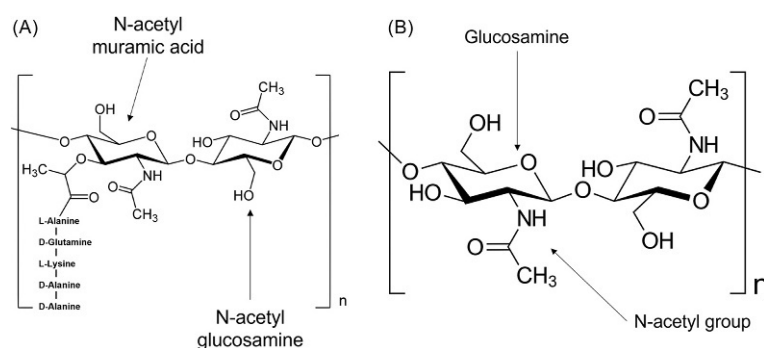


Fig. 2 (A) Bacterial cell wall-derived peptidoglycan comprised of *N*-acetylglucosamine and *N*-acetylmuramic acid cross-linked by chains of *L*- and *D*- enantiomers of amino acids bound by peptide bonds. (B) Fungal derived chitin comprised of polymeric linkages of *N*-acetylglucosamine. Adapted from Myrold DD (2021) Transformations of nitrogen. In: Principles and Applications of Soil Microbiology, pp. 385–421. Elsevier.

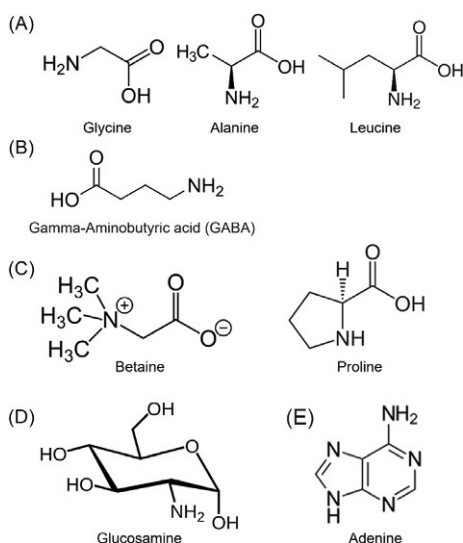


Fig. 3 Common low molecular weight compounds found in soil. (A) Protein amino acids, (B) a non-protein amino acid, (C) quaternary ammonium compounds, (D) an amino sugar, (E) a nucleic acid. Adapted from Myrold DD (2021) Transformations of nitrogen. In: Principles and Applications of Soil Microbiology, pp. 385–421. Elsevier.

Soil organic nitrogen cycling

As a component of SOM, soil organic N is the largest reservoir of N in the terrestrial biosphere; therefore, its cycling is crucial for the global N cycle. Soil organic N inputs are determined by primary productivity, mainly entering the soil in the form of particulate organic matter (POM), via inputs of plant litter, rhizodeposits, animal excreta, and microbial turnover. Leaching from surface litter and root exudation, as well as small amounts of pyrogenic organic N in fire-affected systems can also contribute as DON (Knicker, 2011). The largest input of organic N to soil is proteinaceous material, via the addition of above and belowground plant material and microbial residues (Greenfield et al., 2020). Plants and microbes contain thousands of diverse proteins, which differ from one another in terms of solubility, charge, size, and structure. A smaller fraction of the organic N input to soil includes other polymeric compounds that are of microbial origin, such as peptidoglycan and chitin (Liang et al., 2019). In soil, organic N is subject to the following processes: (1) depolymerization from HMW to LMW DON compounds, (2) microbial cell uptake, (3) mineralization to inorganic N, (4) incorporation into persistent SOM, (5) direct root uptake, and (6) DON leaching (Fig. 4).

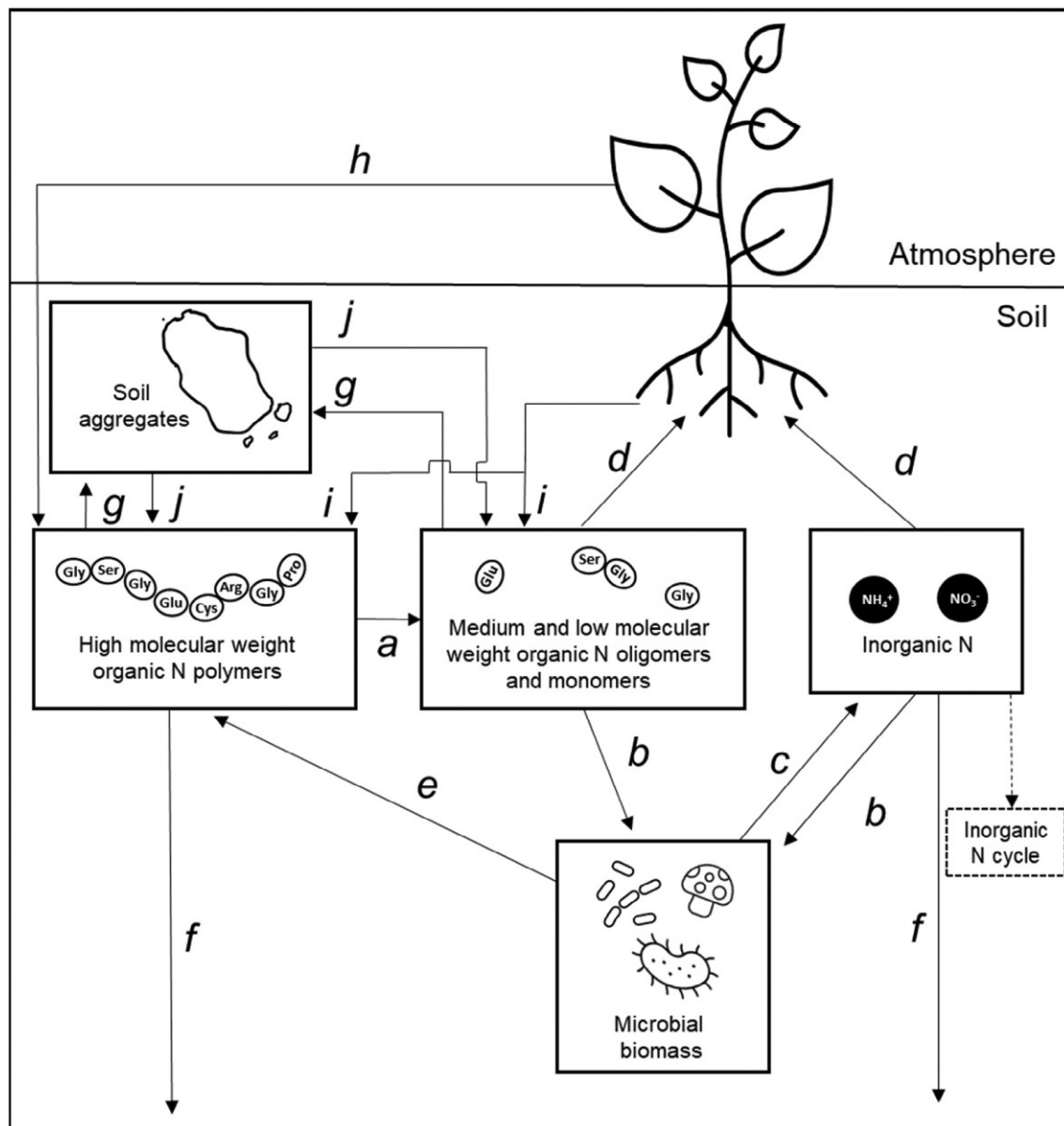


Fig. 4 A simplified soil organic nitrogen cycle. *a*: depolymerization; *b*: microbial uptake, i.e., immobilization; *c*: mineralization; *d*: plant uptake; *e*: microbial turnover and necromass formation; *f*: leaching; *g*: protection of organic nitrogen by various mineral associations; *h*: inputs of organic N, i.e., leaching from aboveground plant litter; *i*: root contributions to soil organic N via cell sloughing, cell lysis, and exudation; *j*: desorption of organic N from mineral surfaces.

Depolymerization and microbial cycling of dissolved organic nitrogen

In general, HMW polymeric material is not readily available for plant and microbial uptake; although some plants may have the capacity to take up HMW organic N compounds, it is not a main pathway of N acquisition (Paungfoo-Lonhienne et al., 2008). Instead, depolymerization of HMW organic N forms to LMW polymers and monomers is the predominant pathway by which plants and microbes can assimilate organic N. The conversion from polymeric organic compounds into bioavailable N regulates N availability in the ecosystem, therefore, depolymerization of HMW organic N is often referred to as the 'rate-limiting step' of the soil N cycle, as depolymerization is central to the provision of bioavailable N (Schimel and Bennett, 2004). However, recent experimental results challenge this paradigm, suggesting that it is not protein depolymerization but instead the rate of protein supply from plant and microbial turnover in soils that controls N availability in soil (Greenfield et al., 2020).

Depolymerization of protein is facilitated by extracellular protease enzymes, which are produced and released by bacteria, fungi, and plant roots. Based on their mode of action, proteases can be classified into exopeptidases and endopeptidases. Both groups facilitate the breakdown of proteins via hydrolyzation, but exopeptidases cleave peptide bonds adjacent to the amino or carboxy termini of peptide chains to produce individual amino acids, whereas endopeptidases cleave peptide bonds present on the inside of the polypeptide molecules, forming oligopeptides (Fig. 5). Importantly, it is sometimes assumed that complete depolymerization of proteins to amino acids occurs, however, evidence suggests that enzymatic degradation of proteins can produce similar quantities of peptides and amino acids, and both can act as an N source for microbes (Warren, 2021). Other HMW organic N, including nucleic acids, cell-wall derived peptidoglycan, and fungal chitin must also undergo depolymerization prior to being utilized as an N source by microbes, albeit at much lower rates than protein-derived peptides and amino acids. Extracellular enzymes catalyze the cleavage of peptidoglycan to muropeptides and muramic acid, chitin to *N*-acetyl-glucosamine and glucosamine, and nucleic acids to individual nucleotides.

After depolymerization of HMW organic N to LMW organic N, compounds can be directly assimilated by microbes, a process that is referred to as the direct route. Conversely, the low C:N ratio of peptides and amino acids can lead to mineralization and release of NH_4^+ into the soil solution, thus becoming available to microbes and plants, or removed by exchange sites on soil minerals and SOM. In processing the available NH_4^+ , microbes may carry out mineralization-immobilization-turnover (MIT), or nitrifying bacteria may oxidize NH_4^+ to NO_3^- , which can then be used as a terminal electron acceptor by denitrifying bacteria and reduced to gaseous nitrogen oxides and dinitrogen and lost to the atmosphere. The quantitative and relative importance of the direct vs. MIT routes of NH_4^+ use during decomposition is determined by factors such as the form of N available, the source of C, and the availability of N relative to C (Geisseler et al., 2010). Overall, cycling of LMW organic N occurs rapidly in soil, leading to small pool sizes of biologically available DON but significant fluxes that can turnover multiple times per day in some ecosystems.

Notably, the rate of soil organic N cycling can be altered by the priming effect, which is the short-term acceleration or retardation of SOM mineralization in response to fresh inputs of C or N to the soil, such as litter, rhizodeposits, or N fertilizer (Kuzyakov et al., 2000). The precise mechanisms underlying this phenomenon are still debated, however, evidence suggests that multiple microbially-driven mechanisms are simultaneously acting in soils leading to observations of the priming effect (Daly and Hernandez-Ramirez, 2020). The N mining hypothesis posits that the priming effect is a response of the soil microbial community to N-limitation after fresh C inputs to soil, thus the acceleration of SOM mineralization is a strategy to liberate assimilable organic N forms for microbial utilization. Conversely, the stoichiometric decomposition hypothesis suggests that the priming effect is due to stoichiometric C and N needs of microbes being met by exogenous inputs, which results in the temporary enhancement of microbial activity and amplified depolymerization of organic N forms (Chen et al., 2014). Therefore, the priming effect is a critical consideration for depolymerization and the stability of SOM, as this enhancement in microbially-mediated turnover can enhance N loss from the soil system.

Soil organic matter persistence

The dynamics of SOM are strongly linked to soil N transformations, as organic N is a major component of SOM and subject to microbially-mediated transformations. When considering the role of microbes in organic N depolymerization, it may be considered paradoxical that microbial cycling of organic N is simultaneously critical for the formation and persistence of SOM. However,

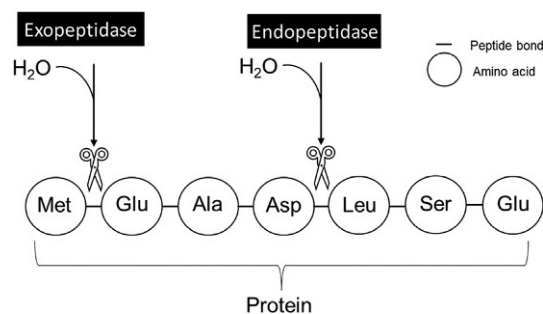


Fig. 5 Depolymerization of protein via hydrolyzation of peptide bonds by *exo*- and *endo*-peptidases.

Table 1 Factors controlling the incorporation of microbially-derived N-containing organic matter into the persistent SOM pool.

Factor	Importance
C: N ratio	Substrate Affects the efficiency (i.e., anabolism vs. catabolism [or assimilation vs. respiration]) and quantity of microbial processing. High throughput of better quality substrate (i.e., materials with a low C:N ratio) can result in a greater mineral-stabilization of SOM due to enhanced microbial residue production
Microbial community structure	Microbial processing Affects the efficiency and quantity of microbial processing and thus the formation of persistent SOM. Microbial communities with high substrate use efficiency may produce greater amounts of N-rich necromass, a primary component of stable OM
Edaphic and environmental conditions	Conditions that favor greater microbial activity will allow greater production of microbially processed OM and therefore may enhance the production of N-rich necromass
Mineralogy	Adsorption Clay minerals play a role in OM persistence. The large surface area of clay particles leads to a greater number of charged sites and thus binding sites for microbial by-products. The net negative charge, small pore spaces, and swelling nature of some phyllosilicate clay minerals can further increase this capacity
Protein size and structure	Affects adsorption to mineral surfaces. Larger organic N compounds (ex. proteins) may have several sites for binding to minerals which can enhance bonding strength and limit active sites for enzyme attachment. Conversely, larger proteins may show increased folding and result in fewer bonding sites
pH	Affects the surface charge of the minerals and the degree of ionization of proteins. The prevalence of ligand exchange, a strong bonding mechanism, increases as pH decreases between proteins and Fe-oxides and Al-silicates
Antecedent OM	The presence of previously sorbed OM can decrease further adsorption capacity if attachment sites are saturated.
Microaggregate formation	Spatial separation Occlusion within aggregates or within small pores may protect SOM from microbial decomposers
Soil depth	Reduced diffusion of water and nutrients to microbial decomposers at depth limits microbial processing of SOM

Adapted from Bingham AH and Cotrufo MF (2016). Organic nitrogen storage in mineral soil: implications for policy and management. *Science of the Total Environment*, 551, pp.116–126.

microbially-derived residues are the main contributor to SOM via biomass turnover and necromass accumulation. Overall, more than 80% of soil organic N is estimated to originate from microbial sources (Hu et al., 2020). Soil microbes contribute a source of chemically diverse, recalcitrant organic matter resulting from anabolic processes that incorporate N into microbial cell walls. Incorporation of microbially-derived organic matter into the persistent SOM pool—the pool that is long-lived and resistant to degradation—is dependent on the substrate composition, microbial community characteristics, the stabilization of organic matter by soil minerals via sorption to cation exchange complexes, and the formation of organo-mineral associations (Hu et al., 2020; Lehmann and Kleber, 2015) (Table 1).

Adsorption of organic N containing molecules onto mineral surfaces reduces their susceptibility to degradation by blocking the attachment of extracellular enzymes, as well as retarding their transport in soil solution and thus reducing their availability (Bingham and Cotrufo, 2016). Absorptive capacity differs between soil environments, dependent on mineralogy, pH, and DON composition. Soils high in phyllosilicate clay minerals and Fe and Al oxides increase the number of binding sites for organic N, whereas pH can alter the charge present on mineral surfaces. In low pH environments, Fe and Al oxides are dominant and predominantly positively charged, which results in bonding via cation exchange. At neutral pH, phyllosilicate clays are negatively charged, and cation bridging becomes more prevalent (Bingham and Cotrufo, 2016). Organic N may also be protected by incorporation within soil aggregates or physical separation in occluded soil pores that are too small for microbes or their enzymes to enter.

Plant uptake of soil organic nitrogen

Previous research into plant nutrition mainly focused on plant reliance of inorganic N forms, i.e., NH_4^+ and NO_3^- , despite organic N comprising up to 90% of the soil N pool. However, an updated understanding suggests that plants, including those of agronomic importance, can take up and utilize a wide variety of organic N forms, including amino acids, peptides, and even proteins, albeit in small but significant amounts (Farzadfar et al., 2021). Notably, organic N compounds play a particularly important role in plant nutrition in N-limited ecosystems such as the tundra or boreal forest, where inorganic N levels are too low to support plant productivity as observed, suggesting that amino acid uptake is a feature many different species possess (Nordin et al., 2004; Persson and Näsholm, 2001).

Like inorganic N, plant acquisition of organic N requires direct contact with roots, with or without the aid of mycorrhizal associations (Näsholm et al., 2009). Direct contact between roots and organic N forms can occur via mass flow, diffusion, and root interception, with diffusion being most significant in the rhizosphere. Plant uptake of organic N compounds is an energy-driven process facilitated by both high and low affinity transport systems located in the plasmalemma, which require the plasma membrane H^+ -ATPase to generate a proton motive gradient. Numerous amino acid and peptide transporters in various species have been identified to date, the former of which differ in their specificity for neutral, basic, or acidic amino acids (Tegeeder, 2012).

Once inside the root cells, organic N compounds are used for biomass and energy production. Specifically, organic N compounds are catabolized in the cytosol of plant root cells to release NH_4^+ and intermediaries such as glutamate. Cytosolic NH_4^+ is used to synthesize glutamine which is then transported into the plastid and assimilated by the GS/GOGAT cycle, producing glutamate. Likewise, the glutamate generated in the cytosol (via catabolism of organic N) enters the GOGAT cycle within the plastid. Subsequently, glutamate is subjected to further aminotransferase reactions to produce other amino acids or is cycled back to glutamine by cytosolic glutamine synthetase and plastidial glutamine synthetase. Similarly, any NO_3^- that is transported into the cytosol of plant root cells is reduced to NO_2^- , transported into the plastid for further reduction to NH_4^+ , and thereafter enters the GOGAT cycle.

Because the rhizosphere is a zone of intense competition for resources, the relevance of organic N as an N source for plants has been questioned, namely because microbes tend to be more efficient at amino acid assimilation due to their high uptake capacity (Kuzuyakov and Xu, 2013). The ability to take up organic N has therefore been suggested as a mechanism to recapture amino acids passively “leaked” from root cells, as opposed to organic N acting as a significant nutrition source for plants (Jones et al., 2005). However, plants have several strategies which may allow them to compete for organic N resources in soil. These strategies include a longer life span than microbes, which release previously bound organic N compounds thus maintaining a temporal niche differentiation, mycorrhizal associations, and root-derived proteases (Kuzuyakov and Xu, 2013).

Certain plants from a variety of ecosystems form associations with particular mycorrhizal fungi, which facilitates the uptake and utilization of organic N forms. The uptake, transformation, and transfer of organic N to plants by mycorrhizal fungi is multi-step process, beginning with the depolymerization of HMW organic N to LMW forms. Ectomycorrhizal and ericoid mycorrhizal fungi can produce a suite of exoenzymes that depolymerize HMW organic N to LMW compounds (Moreau et al., 2019). However, arbuscular mycorrhizal fungi (AMF), which form mutualistic relationships with approximately two-thirds of all land plants, have poor exoenzyme production capability but may increase plant organic N acquisition by hyphal proliferation and cooperation with other soil microorganisms from organic N enriched hotspots in soil (Jansa et al., 2019). Assimilation of organic N by mycorrhizal fungi is facilitated by membrane transport proteins that are general enough to take up several amino acids, or specific enough to target one type of amino acid. Once acquired, the compound may require transformation prior to transfer to the host plant, which requires energy and resource investment by the fungi. The degree to which transfer of organic N from fungi to host plant occurs depends on complex interactions, cost-benefit relationships, and plant and fungal species.

Plants may also increase organic N acquisition by secretion of root-derived proteases, namely in the form of root-bound proteases (Adamczyk, 2021; Greenfield et al., 2020). Enhanced proteolysis on the root surface may enhance plant organic N uptake by reducing competition with microbes, however, the importance of this process for N nutrition is not fully known (Greenfield et al., 2020).

Dissolved organic nitrogen leaching

Leaching of DON is a central pathway of N loss and therefore is an essential component in the soil organic N cycle. Excess leaching of DON into surface waters and drinking water reservoirs can lead to eutrophication, acidification, and potentially pose risks to human health (van Kessel et al., 2009). Organic N leaching occurs as water moves through the soil, but, as opposed to inorganic N losses which are controlled by biological N demand, the losses of organic N are controlled by the depolymerization rates of DON and solubility (Ros et al., 2009). Most DON that is lost to leaching is HMW DON. Because of the HMW form, this loss may not be considered a major loss of bioavailable N for plant and microbial utilization, however, leaks can be large even under extremely N-limited conditions (van Kessel et al., 2009).

In forest ecosystems, our understanding of the N cycle has long included losses due to DON leaching, which can exceed losses of NO_3^- , especially in undisturbed forested ecosystems where DON leaching can account for up to 95% of total dissolved N lost (Perakis and Hedin, 2002). Notably, large leaching losses of DON in forests might explain persistent N limitation in these ecosystems. Conversely, due to the large solubility and mobility of NO_3^- in soil, studies of leaching losses in agroecosystems are often focused on inorganic N and neglect organic N leaching. However, an updated understanding suggests that DON leaching can represent a significant leak of N from the soil system in agroecosystems. Namely, DON leaching can account for up to one-third of total N lost to leaching, (van Kessel et al., 2009).

Sampling, identification, and quantification of organic nitrogen

Quantification of total organic nitrogen

Total organic N in solid soil samples may be determined by the Kjeldahl method or the Dumas method. In brief, the Kjeldahl method digests samples with the aid of a catalyst at high temperature in sulfuric acid (H_2SO_4) then analyzes the distillate NH_4^+ by titration. The total Kjeldahl-N pool usually includes soil soluble NH_4^+ -N in addition to the organic bounded N groups. If the strongly fixed NH_4^+ in minerals is also desired, hydrofluoric acid (HF) may be used in place of H_2SO_4 (Leinweber et al., 2013). For the Dumas method, samples are combusted at very high temperatures in pure oxygen, which releases nitrogen oxide gases that are reduced to N_2 and quantified using a thermal conductivity detector. This provides total N in the sample; organic N must then be determined by subtracting total inorganic N from total N. Inorganic N (i.e., NO_3^- , NO_2^- and NH_4^+) is determined by reacting a subsample with specific reagents for target analytes and followed by ion chromatography or colorimetric analysis. If removal of

fixed NH_4^+ is required, pre-treatment with a strong oxidizing agent such as potassium hypobromite (KBrO) or potassium hydroxide (KOH) is needed (Leinweber et al., 2013).

Sampling dissolved organic nitrogen

Obtaining a sample of organic N from the soil solution can be done ex-situ, via extraction followed by centrifugation or filtration, or in-situ, via microfiltration or microdialysis. Significantly, some literature distinguishes between samples obtained via extraction and those obtained from in-situ methods, as extraction methods also may include soluble organic N compounds that are sorbed to soil particles. Therefore, ex-situ procedures obtain samples of total soluble organic N (SON), whereas in-situ techniques capture DON in the soil solution alone.

Bulk soil samples can be extracted using water or salt solutions including KCl, K_2SO_4 , BaCl_2 , NH_4Cl , NaHCO_3 . The analyte can then be separated from soil using either filtration or centrifugation. This method is ideal for dry soils, but recovery of SON can vary dependent on soil type and extraction procedure. Alternatively, samples can be directly centrifuged to separate the solution from the solid phase. This method requires field moist soils, and like extraction, involves disruption of the soil structure, which may result in deviations between results and actual SON concentrations available in soil.

In-situ methods, such as microfiltration, use suction lysimeters with a semi-permeable membrane that can be installed in soil to collect samples of pore water via mass flow after the application of partial negative pressure. These samples require less soil disturbance compared to their ex-situ counterparts; however, samples can be biased towards larger pores and adsorption of organic compounds to the lysimeter membrane can reduce concentrations in the solution (Jones and Willett, 2006; Warren, 2014b). Microdialysis is a technique that permits sampling of solutes via diffusion, which is driven by the concentration gradient between a perfusate and the soil solution. In this method, a microdialysis probe with a semi-permeable membrane is inserted into soil. Soil solutes then diffuse through the probe membrane into the perfusate which flows at a constant known rate, and which is then continuously sampled (Buckley et al., 2020).

Fractionation, identification, and quantification of organic nitrogen compounds

Several chromatographic methods can be used to separate and quantify organic N compounds in the soil solution, including gas chromatography–mass spectrometry (GC–MS), liquid chromatography–mass spectrometry (LC–MS), capillary electrophoresis–mass spectrometry (CE–MS), pyrolysis–gas chromatography mass spectrometry (Py–GC–MS). Each of these methods combines different compound separation techniques with mass spectrometry (MS) for the profiling of charged molecules, and have detection limits capable of identifying various organic N classes in soil solution when a pre-concentration process is applied to samples prior to analysis. Namely, combining a separation technique prior to MS allows for the separation and identification of ions with the same nominal mass, but different exact mass, as well as the identification of structural isomers. The drawback of these methods is that no single analytical platform provides comprehensive identification and quantification of all organic N compounds.

Compounds must be volatile for analysis by GC–MS, which is not applicable for several organic N molecules (i.e., QACs, oligomers, and polymers). As such, this method is largely limited to the identification of amino acids. If using LC–MS, neutral and anionic compounds can be analyzed, but not polyamines. In contrast, CE–MS can analyze cationic compounds and polyamines, but lacks the ability to identify *N*-acetyl-glucosamine and ureides (Warren, 2014b). Finally, pyrolysis in Py–GC–MS can result in secondary reactions, producing artefacts such as aromatic heterocyclic N compounds that may not be present in the original sample (Leinweber et al., 2013). Therefore, the use of multiple coupled separation–MS methods may be required for a comprehensive analysis of organic N compounds (Warren, 2021).

Solid-state ^{15}N nuclear magnetic resonance (NMR) analysis has been used to identify organic N compounds, by identifying and providing the relative abundance for functional groups. This method is non-destructive and independent of the solubility of different organic N molecules. Therefore, it does not require any pre-treatment, minimizing sample disturbance. However, NMR results can be strongly affected by instrumental parameters, as NMR has limited detection capabilities and therefore, it tends to be less sensitive than MS methods (Knicker, 2011). Improvement in this technique may be possible by combining NMR with various separation techniques, including liquid chromatography (LC–NMR). Analysis via a coupled NMR–MS technique can provide identification of organic N compounds with a specificity that may not be possible with either NMR or MS alone.

Other analytical techniques using MS for the characterization of organic N include pyrolysis field ionization MS (Py–FIMS), Fourier-transform ion cyclotron resonance MS (FT–ICR MS), and nanoscale secondary ion MS (NanoSIMS). Py–FIMS combines temperature resolved pyrolysis within an ion source with ionization. Py–FIMS may reduce the secondary synthesis of heterocyclic aromatic N compounds formed from Py–GC–MS by heating the sample stepwise and transferring less energy to the sample, and by application of correction factors that account for pyrolytic formation of artefacts (Leinweber et al., 2013). In FT–ICR–MS, ions are produced by pressure ionization techniques, which are then detected using an oscillating electric field impulse and converted into single frequencies by applying a Fourier transformation. This method is highly sensitive, and capable of high resolution and accuracy. NanoSIMS is a surface analysis technique that provides a wealth of spatially resolved data on a sample, including the molecular, elemental, and isotopic composition (Leinweber et al., 2013). In nanoSIMS, the sample surface is bombarded with a high energy beam comprised of primary ions, which produces secondary ions and ion complexes that can be detected by MS. A drawback of this method is the combination of secondary ion complexes that may interfere with the signal of target species (Leinweber et al., 2013).

Spectroscopic techniques include X-ray photoelectron spectroscopy (XPS) and XANES spectroscopy. In XPS, electrons are excited by focusing energy from an X-ray source onto a sample. Excited electrons can then be used to understand the configuration of N-atoms and individual N-species can be identified and quantified. This methodology requires ultraclean vacuum conditions, as adsorbates from the atmosphere can lead to contamination of the sample. The use of XANES spectroscopy yields electronic and structural information about the compounds of interest, however, XANES regions for identification of compounds overlap significantly, and as such this method is only able to distinguish between general groups of N moieties and thus lacks specificity (Leinweber et al., 2013).

Soil organic nitrogen: An emerging component of soil condition in managed systems

Soil organic N has historically been overlooked as a source of crop nutrition in agricultural systems, which are reliant on inputs of synthetic N fertilizers for their productivity. Consequently, estimated N availability in agroecosystems is often poorly synchronized with crop demand, resulting in economic and environmental consequences (Grandy et al., 2022). As C input in these systems is generally too small to promote microbial immobilization of the excess N fertilizer added, there can be significant leaks via leaching and gaseous N losses (Finney et al., 2015).

Recent developments in agricultural management are centered on the importance of soil condition to improve sustainability and productivity. The soil health concept is an integrative concept linking a diverse set of indicators to the capacity of soil to function as a living ecosystem (Lehmann et al., 2020). However, there is currently a disconnect between C and N in soil health frameworks, and greater integration of organic N cycling within these frameworks will be a step towards improving agroecosystem health (Grandy et al., 2022). A foundational aspect of management decisions guided by the concept of soil health is the recognition that solely managing inorganic N nutrient availability is not sufficient for optimizing plant productivity. As such, recoupling C and N cycles by building up or maintaining organic C and N reservoirs in soil by promoting plant and microbial processes may be a powerful tool to optimize soil N cycling, align soil N supply with plant demand, and minimize environmental losses.

Although the role of organic N in plant nutrition is not fully understood, evidence suggests that agronomically important crops can access a wide array of N sources, which may provide an opportunity to reduce N fertilizer reliance while maintaining or even improving productivity. Improving N use efficiency (NUE), an established metric used to benchmark N management in agricultural systems that has historically ignored the contributions of organic N, may be one means to accomplish this (Congreves et al., 2021). There are several proposed mechanisms by which a greater emphasis on organic N nutrition may improve crop NUE, including: reduced energy costs for the crop, increased root growth, support of healthy soil microbial communities, re-capture of N compounds lost from roots, and matching root uptake capacity to soil N supply (Farzadfar et al., 2021). However, our current understanding of crop organic N uptake and soil bioavailable N uptake is incomplete and closing of knowledge gaps is needed to permit movement toward management strategies that prioritize soil condition, while maintaining crop productivity.

Conclusion

Soil organic N is a term that describes a highly heterogeneous mixture of N-containing organic compounds of various sizes and compositions. Its transformations and cycling are key components of the terrestrial N cycle, including: depolymerization of HMW organic N to LMW organic N, microbial and plant uptake, incorporation into persistent SOM, and leaching of DON to ground and surface waters. Significant advances in analytical techniques and instrumentation in recent years have supported an improved understanding of soil organic N forms and cycling, and future research can build upon this to increase efficiency and mitigate environmental degradation in agroecosystems and other managed systems.

References

- Adamczyk B (2021) Root-derived proteases as a plant tool to access soil organic nitrogen; Current stage of knowledge and controversies. *Plants* 10(4): 731–740. <https://doi.org/10.3390/plants10040731>.
- Bingham AH and Cotrufo MF (2016) Organic nitrogen storage in mineral soil: Implications for policy and management. *Science of the Total Environment* 551: 116–126.
- Buckley S, Brackin R, Jämtgård S, Näsholm T, and Schmidt S (2020) Microdialysis in soil environments: Current practice and future perspectives. *Soil Biology and Biochemistry* 143: 107743.
- Chen R, Senbayram M, Blagodatsky S, Myachina O, Dittert K, Lin X, Blagodatskaya E, and Kuzyakov Y (2014) Soil C and N availability determine the priming effect: Microbial N mining and stoichiometric decomposition theories. *Global Change Biology* 20(7): 2356–2367.
- Congreves KA, Otchere O, Ferland D, Farzadfar S, Williams S, and Arcand MM (2021) Nitrogen use efficiency definitions of today and tomorrow. *Frontiers in Plant Science* 12: 637108.
- Cotrufo MF and Lavalley JM (2022) Soil organic matter formation, persistence, and functioning: A synthesis of current understanding to inform its conservation and regeneration. *Advances in Agronomy* 172: 1–66.
- Daly EJ and Hernandez-Ramirez G (2020) Sources and priming of soil N₂O and CO₂ production: Nitrogen and simulated exudate additions. *Soil Biology and Biochemistry* 149: 107942.
- Farzadfar S, Knight JD, and Congreves KA (2021) Soil organic nitrogen: An overlooked but potentially significant contribution to crop nutrition. *Plant and Soil* 462(1): 7–23.
- Finney DM, Eckert SE, and Kaye JP (2015) Drivers of nitrogen dynamics in ecologically based agriculture revealed by long-term, high-frequency field measurements. *Ecological Applications* 25: 2210–2227.

- Galloway JN, Townsend AR, Erisman JW, Bekunda M, Cai Z, Freney JR, Martinelli LA, Seitzinger SP, and Sutton MA (2008) Transformation of the nitrogen cycle: Recent trends, questions, and potential solutions. *Science* 320(5878): 889–892.
- Geisseler D, Horwath WR, Joergensen RG, and Ludwig B (2010) Pathways of nitrogen utilization by soil microorganisms—A review. *Soil Biology and Biochemistry* 42(12): 2058–2067.
- Grandy AS, Daly AB, Bowles TM, Gaudin AC, Jilling A, Leptin A, McDaniel MD, Wade J, and Waterhouse H (2022) The nitrogen gap in soil health concepts and fertility measurements. *Soil Biology and Biochemistry* 175: 108856.
- Greenfield LM, Hill PW, Seaton FM, Paterson E, Baggs EM, and Jones DL (2020) Is soluble protein mineralisation and protease activity in soil regulated by supply or demand? *Soil Biology and Biochemistry* 150: 108007.
- Hu Y, Zheng Q, Noll L, Zhang S, and Wanek W (2020) Direct measurement of the in situ decomposition of microbial-derived soil organic matter. *Soil Biology and Biochemistry* 141: 107660.
- Jansa J, Forczek ST, Rozmoš M, Püschel D, Bukovská P, and Hřelová H (2019) Arbuscular mycorrhiza and soil organic nitrogen: Network of players and interactions. *Chemical and Biological Technologies in Agriculture* 6(1): 1–10.
- Jones DL and Willett VB (2006) Experimental evaluation of methods to quantify dissolved organic nitrogen (DON) and dissolved organic carbon (DOC) in soil. *Soil Biology and Biochemistry* 38(5): 991–999.
- Jones DL, Healey JR, Willett VB, Farrar JF, and Hodge A (2005) Dissolved organic nitrogen uptake by plants—An important N uptake pathway? *Soil Biology and Biochemistry* 37(3): 413–423.
- Knicker H (2011) Soil organic N—An under-rated player for C sequestration in soils? *Soil Biology and Biochemistry* 43(6): 1118–1129.
- Kuzyakov Y and Xu X (2013) Competition between roots and microorganisms for nitrogen: Mechanisms and ecological relevance. *New Phytologist* 198(3): 656–669.
- Kuzyakov Y, Friedel JK, and Stahr K (2000) Review of mechanisms and quantification of priming effects. *Soil Biology and Biochemistry* 32(11–12): 1485–1498.
- Lehmann J and Kleber M (2015) The contentious nature of soil organic matter. *Nature* 528(7580): 60–68.
- Lehmann J, Bossio DA, Kögel-Knabner I, and Rillig MC (2020) The concept and future prospects of soil health. *Nature Reviews Earth & Environment* 1(10): 544–553.
- Leinweber P, Kruse J, Baum C, Arcand M, Knight JD, Farrell R, Eckhardt KU, Kiersch K, and Jandl G (2013) Advances in understanding organic nitrogen chemistry in soils using state-of-the-art analytical techniques. *Advances in Agronomy* 119: 83–151.
- Liang C, Amelung W, Lehmann J, and Kästner M (2019) Quantitative assessment of microbial necromass contribution to soil organic matter. *Global Change Biology* 25(11): 3578–3590.
- Moreau D, Bardgett RD, Finlay RD, Jones DL, and Philippot L (2019) A plant perspective on nitrogen cycling in the rhizosphere. *Functional Ecology* 33(4): 540–552.
- Näsholm T, Kielland K, and Ganeteg U (2009) Uptake of organic nitrogen by plants. *New Phytologist* 182(1): 31–48.
- Nordin A, Schmidt IK, and Shaver GR (2004) Nitrogen uptake by arctic soil microbes and plants in relation to soil nitrogen supply. *Ecology* 85(4): 955–962.
- Paungfoo-Lonhienne C, Lonhienne TG, Rentsch D, Robinson N, Christie M, Webb RI, Gamage HK, Carroll BJ, Schenk PM, and Schmidt S (2008) Plants can use protein as a nitrogen source without assistance from other organisms. *Proceedings of the National Academy of Sciences of the United States of America* 105(11): 4524–4529.
- Perakis SS and Hedin LO (2002) Nitrogen loss from unpolluted South American forests mainly via dissolved organic compounds. *Nature* 415(6870): 416–419.
- Persson J and Näsholm T (2001) Amino acid uptake: A widespread ability among boreal forest plants. *Ecology Letters* 4(5): 434–438.
- Ros GH, Hoffland E, Van Kessel C, and Temminghoff EJ (2009) Extractable and dissolved soil organic nitrogen—A quantitative assessment. *Soil Biology and Biochemistry* 41(6): 1029–1039.
- Schimel JP and Bennett J (2004) Nitrogen mineralization: Challenges of a changing paradigm. *Ecology* 85(3): 591–602.
- Tegeder M (2012) Transporters for amino acids in plant cells: Some functions and many unknowns. *Current Opinion in Plant Biology* 15(3): 315–321.
- van Kessel C, Clough T, and van Groenigen JW (2009) Dissolved organic nitrogen: An overlooked pathway of nitrogen loss from agricultural systems? *Journal of Environmental Quality* 38(2): 393–401.
- Warren CR (2014a) Organic N molecules in the soil solution: What is known, what is unknown and the path forwards. *Plant and Soil* 375(1): 1–19.
- Warren CR (2014b) Development of liquid chromatography mass spectrometry method for analysis of organic N monomers in soil. *Soil Biology and Biochemistry* 78: 233–242.
- Warren C (2021) What are the products of enzymatic cleavage of organic N? *Soil Biology and Biochemistry* 154: 108152.

Nitrification: Process, organisms and management

Jeanette M Norton^a and Yang Ouyang^b, ^aDepartment of Plants, Soils and Climate and Ecology Center, Utah State University, Logan, UT, United States; ^bDepartment of Microbiology and Plant Biology, Institute of Environmental Genomics, University of Oklahoma, Norman, OK, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of J.I. Prosser, NITROGEN IN SOILS I Nitrification, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 31–39, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00512-9>.

Introduction	253
Nitrification processes	254
Trace gas production by nitrification	255
The organisms of nitrification in soils	255
Nitrification in the soil environment and agricultural management	256
Determination of nitrification rates in soil systems	257
Management of nitrification in agricultural systems	258
Nitrification inhibitors	259
Summary and outlook	259
References	259

Abstract

Nitrification is the biological conversion of ammonia or ammonium to nitrite or nitrate. Nitrate is mobile in soil and subject to loss by leaching and denitrification. In agricultural soils, loss of nitrate leads to decreased plant available nitrogen and reduces nitrogen use efficiency. Nitrification contributes both directly and indirectly to the production of greenhouse gases, nitrous oxide and nitric oxide. The rate of nitrification in soil is controlled by substrate availability (such as ammonia or urea fertilizers) and environmental parameters of oxygen, water, and temperature. Nitrification can be controlled by limiting substrate availability or with biological or synthetic nitrification inhibitors.

Glossary

Ammonia oxidation First step of nitrification where ammonia is converted to hydroxylamine and then to nitrite.

Anammox Anaerobic ammonium oxidation in which **nitrite** and **ammonium ions** are converted directly into **nitrogen** gas (N₂) and water.

Chemolithoautotrophy A type of microbial metabolism which derives energy from chemical reactions of **reduced** compounds of **mineral** (inorganic or litho) origin, these are used as their electron source. Auto refers to their carbon source which is carbon dioxide or carbonates.

Comammox Complete ammonia oxidation to nitrite and then to nitrate in one organism.

Nitrapyrin Is an organic compound is an organic compound (with the formula ClC₅H₃NCCL₃) used as a **nitrification** inhibitor and bactericide.

Nitrification Is the biological **oxidation** of **ammonia** to **nitrite** followed by the oxidation of the **nitrite** to **nitrate**.

Nitrite oxidation Second step in nitrification where nitrite is converted to nitrate.

Key points

- Nitrification changes the form of nitrogen and increases nitrogen loss from agroecosystems
- Nitrifying microorganisms have unique metabolisms that oxidize nitrogen for energy and use carbon dioxide to grow
- Key enzymes of nitrifiers include ammonia monooxygenase, hydroxylamine oxidoreductase (dehydrogenase) and nitrite oxidoreductase
- Nitrification rates may be controlled by limiting the availability of ammonium and ammonia or by natural or synthetic nitrification inhibitors

Introduction

Nitrification is the biological conversion of reduced nitrogen (N) generally in the form of ammonia or ammonium to oxidized N in the form of nitrite or nitrate. Nitrification converts positively charged ammonium (NH_4^+) into negatively charged nitrite (NO_2^-) or nitrate (NO_3^-), thus changing their interactions with the soil matrix and their availability for plant uptake. The products of nitrification are mobile in the soil and therefore subject to losses from the plant-soil system via leaching and denitrification (see Fig. 1). For these reasons, it is often desirable to manage agricultural soils to reduce nitrification and subsequent N loss, and thereby increase the efficiency of N fertilizers. In animal or municipal waste treatment systems, nitrification is often a desirable process as nitrate is less toxic to aquatic systems than ammonia (NH_3) and can be removed from the waste stream through the denitrification process. The nitrification process is known to produce both directly and indirectly, nitrogenous trace gases, nitric and nitrous oxides important in greenhouse gas and ozone atmospheric chemistry. Since natural and agricultural soils account for a large proportion (>60%) of the total global nitrous oxide production (Tian et al., 2020), managing nitrification and associated downstream processes is important for mitigating the greenhouse effect resulting from terrestrial emissions.

Nitrification processes

The broad definition of nitrification includes a variety of processes by distinct groups of microorganisms that oxidize reduced N to oxidized N, generally in the form of nitrite or nitrate. Nitrification is classically thought of as a process mediated by bacteria that gain energy from the oxidation of inorganic N and fix inorganic carbon (CO_2) and therefore have a chemolithoautotrophic metabolism. The biological nature of this process and the organisms involved were first studied in the late 19th century by Winogradsky (1890-1891) and by Frankland and Frankland (1890). The conversion of N takes place in two phases: in the ammonia oxidizing bacteria such as *Nitrosomonas* or *Nitrospira*, NH_3 or NH_4^+ is converted to NO_2^- . Ammonia oxidizing archaea, such as *Nitrososphaera*, have also been shown to oxidize NH_3 to NO_2^- (Schleper and Nicol, 2010) although by a significantly distinct metabolism from the bacteria. Ammonia oxidizing bacteria and archaea both have ammonia monooxygenase (AMO) enzymes that catalyze the conversion of ammonia to hydroxylamine (Fig. 2). The gene encoding the ammonia monooxygenase subunit-A is often

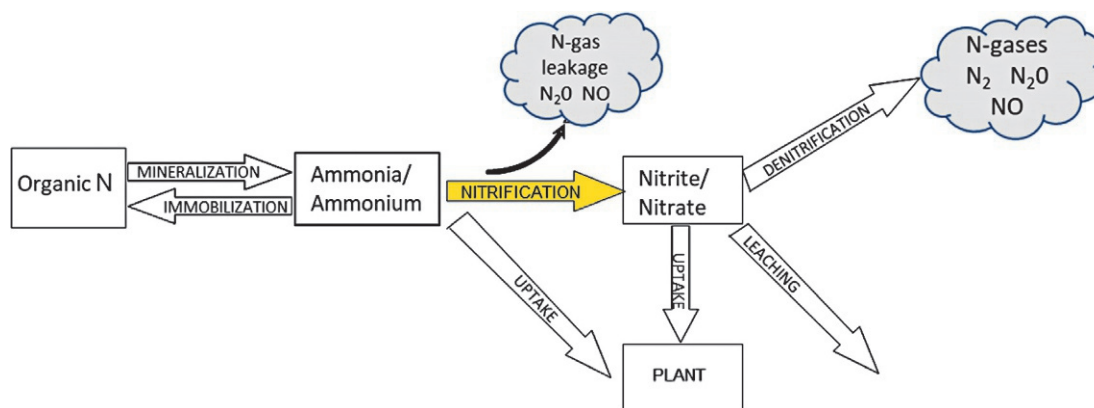


Fig. 1 Overview of the important role of nitrification in affecting the availability and fate of ammonia/ammonium and nitrite/nitrate for plant uptake or leading to loss of N from the root-zone via leaching or denitrification.

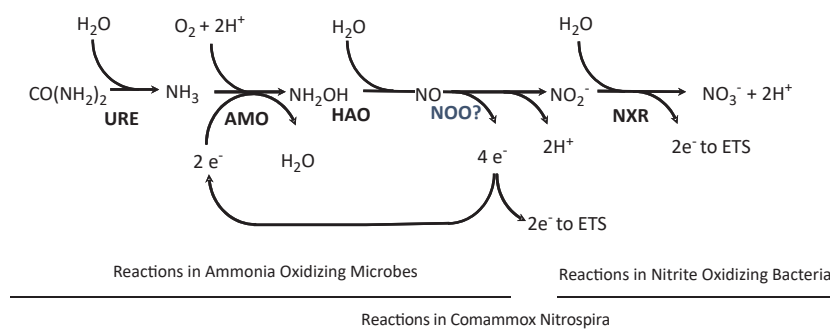


Fig. 2 Outline of nitrification pathway showing some of the important enzymes including: (1) URE: urease is present in some but not all ammonia oxidizing microbes, (2) AMO: ammonia monooxygenase (3) HAO: hydroxylamine oxidoreductase now also known as hydroxylamine dehydrogenase (may not be present in archaeal ammonia oxidizers) (4) NOO: a postulated nitric oxide oxidase and (5) NXR: nitrite oxidoreductase. Only the comammox *Nitrospira* would contain all of these enzymes in a single organism.

used as a molecular marker for ammonia oxidizers. The product of the enzyme hydroxylamine oxidoreductase (HAO) is thought to be nitric oxide (NO), which is then converted to nitrite by an unidentified NO oxidase (NOO) (Lancaster et al., 2018). In archaeal ammonia oxidizers, the similar HAO enzyme is lacking but it is thought that a Cu-enzyme converts hydroxylamine and nitric oxide to nitrite (Lancaster et al., 2018) with NO as a key intermediate. Nevertheless, the net outcome of ammonia oxidation is nitrite as shown in Eq. (1).



In phase 2, nitrite oxidizing bacteria such as *Nitrobacter* or *Nitrospira* convert nitrite to nitrate. This involves the key enzyme nitrite oxidoreductase (NXR) (Fig. 2).



Certain *Nitrospira* bacteria have been found that mediate the entire reaction from ammonium to nitrate within one organism in the Complete Ammonia Oxidation to nitrate known as comammox (Daims et al., 2015; van Kessel et al., 2015).



Other nitrification processes have also been identified as important in specific environments. Heterotrophic nitrification may be broadly defined as the oxidation of reduced N compounds (including organic N) producing nitrite or nitrate. Heterotrophic nitrification is catalyzed by a variety of microorganisms including fungi, and bacteria, including actinomycetes, presumably using a wide variety of metabolic pathways and has been found to be significant in some acidic forest soils. However, the process of heterotrophic nitrification has not been shown to be associated with energy production or to support autotrophic growth of the organisms involved (Stein, 2011).

Anaerobic oxidation of ammonium to dinitrogen gas via hydrazine (Eq. 4) termed the *anammox* process is catalyzed by nitrite dependent anaerobic bacteria belonging to the phylum *Planctomycetes* (Jetten et al., 2009). This process was first found in and is desirable for wastewater treatment systems but also occurs in paddy soils and in oxygen minimum zones of marine systems.



Although the diversity of processes included under a broad definition of nitrification has recently expanded, the aerobic oxidation of ammonia and nitrite by chemolithotrophic bacteria and archaea remains the primary mechanism of nitrification in aerobic agricultural soils.

Trace gas production by nitrification

The main products of nitrification are NO_2^- and NO_3^- which are subject to denitrification reactions that lead to the production of the greenhouse gases nitrous oxide (N_2O) and nitric oxide (NO). In addition to this important but indirect pathway, nitrification processes themselves are known to produce these trace gases as side products and in internal reactions known as nitrifier denitrification. Nitrifiers may produce N_2O by hybrid formation under oxic conditions, namely the abiotic reaction of hydroxylamine, NO, and NO_2^- together or with ferrous iron (Fe^{2+}). Nitrifier denitrification occurs under hypoxic conditions and may release both NO and N_2O . In most aerobic agricultural soils, the amount of direct emission of these trace gases during nitrification is a small proportion of the total emission (Liang and Robertson, 2021; Stein, 2011). However, because NO_2^- and NO_3^- are the substrates for subsequent denitrification, inhibiting nitrification is an effective method for preventing trace gas emissions from agroecosystems.

The organisms of nitrification in soils

The majority of bacteria and archaea that mediate nitrification in soil environments gain metabolic energy from the oxidation of N and use inorganic carbon (i.e. carbon dioxide and bicarbonate) as the primary carbon source for growth of their biomass. This mode of metabolism means that they are highly specialized organisms filling an important role in soil N cycling and N loss. An outline of the process of aerobic chemolithotrophic nitrification showing some of the key enzymes and the groups of organisms involved is shown in Fig. 2. The genes encoding subunits of the key enzymes AMO and NXR are often used as molecular markers for the different groups. The major functional groups of nitrifiers and a representative organism for each from soil if possible is given in Table 1.

In soil environments, the ammonia oxidizing bacteria generally belong to the *Beta-proteobacteria* and include the well-known *Nitrosomonas* and *Nitrospira* genera. Pure culture isolates from soil environments are well-represented in culture collections and by genome sequences (Norton, 2011; Prosser et al., 2014). In most soils, *Nitrospira* outnumber *Nitrosomonas*. The community structure of ammonia oxidizing bacteria is affected by soil agricultural management especially fertilizer application and changing

Table 1 Some well-known strains of soil or environmental nitrifying organisms by functional group with reference to their most complete genome sequence.

Genus and species	Strain	Lineage or cluster	Genome Accession	References
Ammonia oxidizing bacteria (<i>Beta-Proteobacteria</i>)				
<i>Nitrospira multiformis</i>	ATCC 25196	cluster 3	GCA_000196355.1	Norton et al. (2008)
<i>Nitrosomonas oligotropha</i>	Nm45	cluster 6A	NZ_QRBZ00000000.1	Sedlacek et al. (2019)
<i>Nitrosomonas europaea</i>	ATCC 19718	cluster 7	GCF_000009145.1	Chain et al. (2003)
Ammonia oxidizing archaea (<i>Thaumarchaeota</i>)				
<i>Nitrososphaera viennensis</i>	EN-76 ^T	<i>Nitrososphaera</i>	GCF_000698785.1	Stieglmeier et al. (2014)
Ca. <i>Nitrosocosmicus franklandus</i>	C13 NFRAN1	<i>Nitrososphaera</i> sister	GCF_900696045.1	Lehtovirta-Morley et al. (2016)
Nitrite oxidizing bacteria (<i>Proteobacteria</i> , <i>Nitrospirae</i>)				
<i>Nitrobacter winogradskyi</i>	NB-255	Alpha-proteobacteria	GCA_000012725.1	Starkenburger et al. (2006)
<i>Nitrospira moscoviensis</i>	NSP-S1	lineage II <i>Nitrospira</i>	NZ_CP011801.1	Koch et al. (2015)
Comammox bacteria (<i>Nitrospirae</i>)				
<i>Nitrospira inopinata</i>	ENR4	lineage II (clade A.1)	GCF_001458695.1	Daims et al. (2015)
Anammox bacteria (<i>Planctomycetes</i>)				
Ca. <i>Jettenia caeni</i>	KSU-1		GCA_000296795.1	Ali et al. (2015)
Ca. <i>Brocadia sinica</i>	JPN1		NZ_BAFN01000001.1	Oshiki et al. (2016)
Heterotrophic Nitrifier (<i>Proteobacteria</i> and others)				
<i>Paracoccus pantotrophus</i>	DSM 2944	Wastewater source	GCA_003633525.1	Stein (2011)

pH (Prosser, 2011). A meta-analysis of studies quantifying the effect of N fertilizers on the N cycling functional genes found that ammonia oxidizing bacterial abundance increased in response to larger applications and longer duration of N fertilizer use (Ouyang et al., 2018).

The importance of the ammonia oxidizing archaea in soils is indicated by their observed numerical dominance in many systems and their ability to process ammonium in acidic soils. The phylum *Thaumarchaeota* covers the recently discovered diversity of this group of important soil ammonia oxidizing organisms. The groups found in soils are primarily in the order *Nitrososphaerales* and in the *Nitrosotales* order for acid soils (Alves et al., 2018). The diversity and community structure of these organisms are generally determined using the functional gene for the archaeal AMO. A meta-analysis of studies quantifying the effect of N fertilizers on the N cycling functional genes found that ammonia oxidizing archaeal abundance increased in response to larger applications of organic nitrogen amendments at pH >6–8, (Ouyang et al., 2018). This meta-analysis and recent characterization of a soil isolate (Lehtovirta-Morley et al., 2016) suggest that both bacterial and archaeal ammonia oxidizers co-exist and are important in agricultural soils receiving applied nutrients.

Bacteria that oxidize NO_2^- are quite diverse and include members from several phyla including the *Proteobacteria*, *Nitrospirae* and *Nitrospiniae* suggesting that NO_2^- -oxidizing metabolism may have evolved multiple times in evolutionary history. *Nitrobacter* are often isolated from agricultural soils receiving fertilizers, but *Nitrospira* may be numerically dominant. *Nitrospira* also includes the comammox organisms, which as already mentioned, are capable of complete oxidation of ammonium to nitrate. There is convincing metagenomic and isotopic evidence of comammox presence and activity in the soil environment although pure cultures from soil are not currently available. Linking the ecophysiology of the different groups of nitrite oxidizers to their activity and molecular markers remains to be fully realized.

Anammox bacteria are responsible for a significant amount of global N loss from marine, aquatic and terrestrial systems. Their presence and activity are found in various soils mostly in anaerobic or wet environments including permafrost soils, agricultural soils, peat soils and paddy rice soils. The ecological importance of anammox in soil N cycling has not yet been fully explored.

Heterotrophic nitrification may be mediated by a wide range of both bacteria and fungi. Some known heterotrophic nitrifiers including *Paracoccus pantotrophus* are also known aerobic denitrifiers and therefore are of biotechnological and wastewater engineering interest (Duan et al., 2021). While similar bacteria to known heterotrophic nitrifiers are present in most soils their importance to soil nitrification activity remains poorly characterized.

Nitrification in the soil environment and agricultural management

In the majority of aerobic circum-neutral soils, ammonium is rapidly converted to nitrate by nitrification, which increases the mobility of N through the soil matrix, strongly influencing N retention in the system. The factors which control the rate and extent of nitrification and therefore also the populations and community structure of the nitrifying organisms are sometimes typical of general soil microbes but key factors are unique to the nitrifiers. The controlling factors on the rate and extent of nitrification in soils include: substrate supply, environmental conditions, abundance and diversity of nitrifiers, plant and microbial interactions and nitrification inhibitors (Norton and Ouyang, 2019) (see Table 2).

Table 2 Main factors controlling the rates of nitrification in aerobic soils.**Substrate supply**

- 1) Fertilization with ammonia and urea
- 2) Inorganic nitrogen: pools and turnover of ammonium/ammonia, nitrite
- 3) Decomposition - rates of ammonium production via mineralization of organic nitrogen
- 4) Organic nitrogen pools - amounts and characteristics (C/N and complexity)
- 5) Nitrogen deposition to inorganic pools
- 6) Oxygen availability

Environmental soil conditions

- 1) Temperature
- 2) Moisture and Aeration (Oxygen)
- 3) Soil pH

Microbial actors

- 1) Populations of ammonia oxidizers, nitrite oxidizers
- 2) Community structure of ammonia oxidizers, nitrite oxidizers

Nitrification inhibition

- 1) Biological inhibitors produced by plants or other microbes
- 2) Synthetic nitrification or urease inhibitors applied with fertilizers

Adapted from Norton JM and Ouyang Y (2019) Controls and adaptive management of nitrification in agricultural soils. *Frontiers in Microbiology* 10: 1931.

The ammonia oxidizers require NH_3 or NH_4^+ , oxygen and carbon dioxide to proliferate and grow. Since carbon dioxide is always available in soil environments and under well-drained soil conditions oxygen is sufficient for nitrification to proceed, the availability of NH_3 or NH_4^+ limits both the rate of nitrification and the size of the resultant nitrifier population in most agricultural soils. The amounts of substrates available for nitrification include those added as fertilizers, mostly as urea and NH_3 or NH_4^+ in arable soils, urea added by grazing animals, N in wet and dry deposition, and NH_3 or NH_4^+ produced by decomposition of organic N (Table 2). As indicated by Eqs. (1) and (2), both ammonia oxidation and nitrite oxidation require oxygen as a substrate so the aeration status of the soil is a key factor as well.

Both ammonia oxidizing bacteria (including comammox *Nitrospira*) and archaea have AMO enzymes. The substrate for AMO is believed to be NH_3 and is therefore its availability is partially controlled by the pH of the environment. The ammonia: ammonium equilibrium ($\text{pK}_a = 9.25$) means that at pH 6 and lower the availability of free ammonia is very small. Urea is also used directly by some but not all soil ammonia oxidizers as they have a cytoplasmic urease enzyme and have urea transporters. The ability to use urea is an advantage to these organisms especially in acidic soils where ammonia availability is small.

Nitrite does not usually accumulate in soils except under transient conditions that have decreased the population or activity of the NO_2^- -oxidizers. One example is in fertilizer bands of anhydrous ammonia where ammonia toxicity may inhibit nitrite oxidation. When nitrite does accumulate, this often leads to nitric acid that may decompose or be transformed to NO and N_2O gases. The extent of NO_2^- accumulation and subsequent trace gas production is controlled by soil pH, soil buffering capacity, fertilizer formulation and depth of incorporation, and soil moisture (Venterea and Rolston, 2000). A lag effect where ammonia-oxidizers grow and convert ammonia faster than the NO_2^- -oxidizers can respond has been attributed to direct ammonia toxicity or nitrification induced acidity. This effect of the slower growth of nitrite-oxidizers has been counteracted by adding *Nitrobacter* cells to soil slurries (Giguere et al., 2017).

Determination of nitrification rates in soil systems

The term nitrification rate has at least three distinct experimentally defined meanings: *net*, *gross* and *potential* nitrification rates. The net nitrification rate is the rate of accumulation of NO_2^- or NO_3^- and is equal to the rate of NH_3 or NH_4^+ oxidation minus any consumption. This is the most commonly measured rate in both laboratory and field conditions. The gross nitrification rate is the actual rate of conversion of NH_3 or NH_4^+ to NO_2^- and NO_3^- regardless of their consumption by plants, other microbes, and abiotic mechanisms. Determination of gross nitrification rates generally requires the use of isotopes of N specifically the stable isotope ^{15}N . Using modeling and pool-dilution equations investigators determine gross rates of the simultaneous production and consumption of NO_2^- and NO_3^- . The potential nitrification rate is the maximum rate of nitrification in a mixed system with non-limiting substrate (ammonium or ammonia) supply with the elimination of NO_2^- and NO_3^- consumption. Nitrification potentials measure short-term NO_2^- and NO_3^- production in shaken soil slurries with non-limiting substrate supply. These assays are useful indicators of the enzymatic potential for nitrification but are not necessarily predictive of in-situ rates, which are often constrained by substrate availability (Table 2). The details of the methods used for these rate determinations are given in Norton and Stark (2011). Measurements of these different operational rates are often combined to give a clearer view of the nitrification process.

Management of nitrification in agricultural systems

Agriculture is responsible for over half of the input of reactive N to terrestrial systems and improving N availability remains the primary management technique to increase crop yields in most regions. Management actions that decrease nitrification are desirable because they reduce N losses and increase N fertilizer use efficiency. Several approaches are available to control ammonium substrate availability while maintaining crop production. These include nitrogenous fertilizer management based on the '4Rs' approach of applying the right source, at the right rate, at the right time, in the right place (Clarke and Beegle, 2014). Crop rotations and cover and catch crops may be used to decrease ammonium and nitrate pools that are subject to loss. Nitrification inhibitors may be applied with fertilizers to directly control and delay nitrification. Although many appropriate technologies are currently available to reduce nitrification and subsequent greenhouse gas emissions and N losses; these may require appropriate incentives for farmers to adopt. China has some of the most intensive application programs for N fertilizers and the associated high levels of nitrification and nitrate contamination. In a meta-analysis of Chinese agriculture, management practices designed to minimize N loss were assessed including: the application of controlled-release N fertilizers, nitrification and urease inhibitors, higher splitting frequency of fertilizer N application, lower basal N fertilizer proportions, deep placement of N fertilizer, and optimizing N rate based on soil N test (Xia et al., 2017). These knowledge-based N fertilizer application practices were generally effective at reducing N loss by leaching, runoff and greenhouse gas emission and showed some increases in economic return. Split applications of fertilizer-N and the use of enhanced efficiency fertilizers, including those with polymer coatings and urease and nitrification inhibitors, will make increased economic sense if they are used selectively under those environmental conditions where the potential for N loss is high. The goal of nitrification management is to promote a low nitrifying system and reduce N loss and greenhouse gas emissions (Fig. 3).

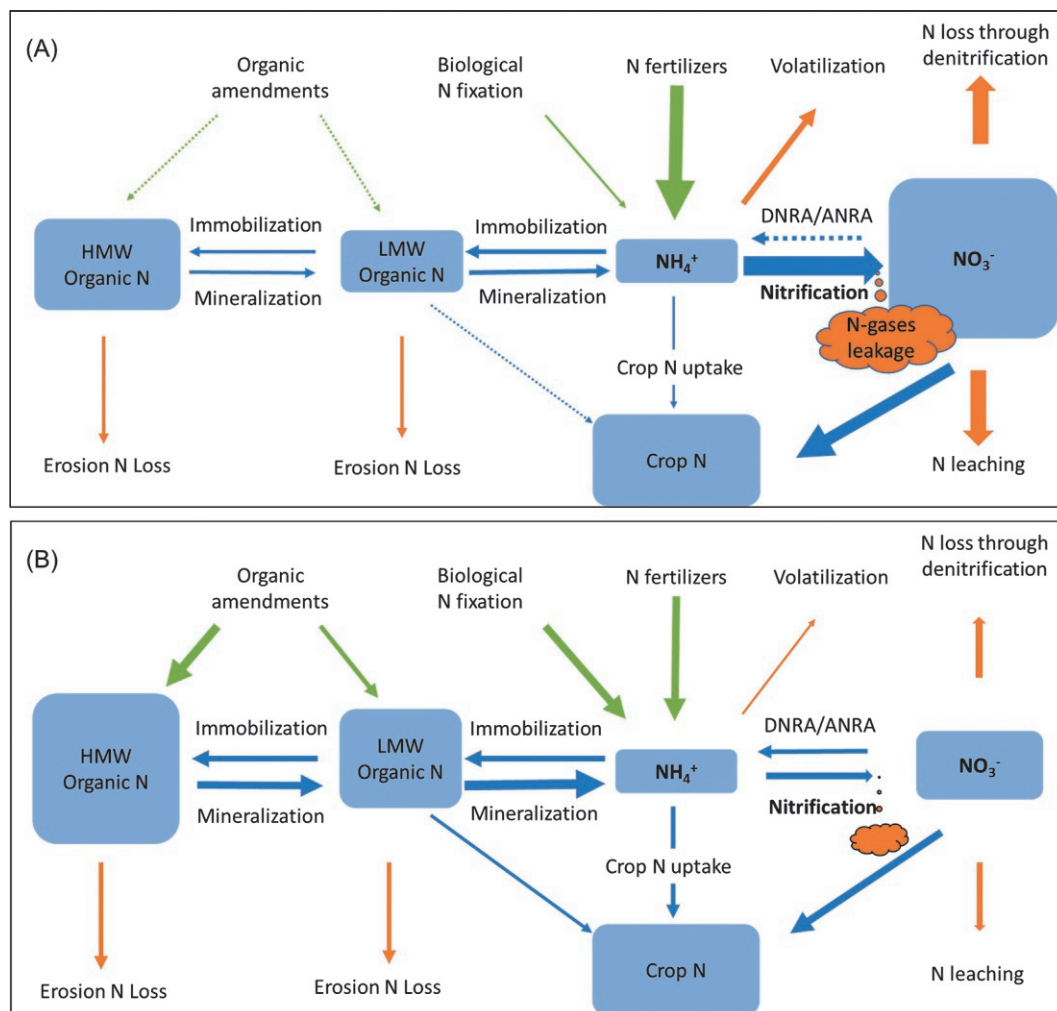


Fig. 3 Comparison of agricultural management systems resulting in high-nitrifying and low nitrifying agricultural systems. Hypothetic nitrogen pools and flows are represented. Arrows represent nitrogen inputs (green), losses (orange) and transformations (blue). Abbreviations: HMW, high molecular weight; LMW, low molecular weight; DNRA, dissimilatory nitrate reduction to ammonium; ANRA, assimilatory nitrate reduction to ammonium. Adapted from Norton JM and Ouyang Y (2019) Controls and adaptive management of nitrification in agricultural soils. *Frontiers in Microbiology* 10: 1931.

Nitrification inhibitors

Inhibition of nitrification is an important consideration in three main areas: (1) agricultural N management to increase fertilizer use efficiency and minimize N loss, (2) non-target effects of inhibitory substances in waste treatment systems and in soil ecology, and (3) research investigations of nitrification or nitrifying organisms. The use of synthetic nitrification inhibitors in agriculture is most likely to be economical when nitrate losses are significant, such as in sandy soils, poorly drained soils, or when ammoniacal N is applied in the fall.

Nitrification inhibitors (NIs) slow the microbial conversion of NH_4^+ to NO_3^- , thereby reducing the risk of loss through leaching or denitrification and consequently increasing the N use efficiency of fertilizers. Generally, the first step of nitrification is targeted because accumulation of nitrite is undesirable. Many synthetic nitrification inhibitors act on the AMO enzyme (for example acetylene). Several nitrification inhibitors that are widely used in agriculture include: (1) 2-chloro-6-(trichloromethyl) pyridine (nitrapyrin), (2) dicyandiamide (DCD), and (3) 3,4-dimethylepyrazole phosphate (DMPP). Urease inhibitors, such as N-(n-butyl) thiophosphoric triamide (NBPT), are used to decrease urea hydrolysis and volatilization. Urease and nitrification inhibitors may be used together to delay N movement and slow losses of N. Meta-analyses report that the application of urease and nitrification inhibitors significantly reduced inorganic N leaching (−48%), N_2O emission (−44%), and NO emission (−24%) (Thapa et al., 2016) while increasing crop yield (+7.5%) and N use efficiency (+12.9%) (Abalos et al., 2014). Achieving a beneficial outcome from nitrification inhibitors may depend on the soil environment (e.g. soil pH, texture, temperature) and other management factors (e.g. irrigation and N fertilizer rate). The longevity of the inhibitors under soil conditions as affected by temperature is of key importance for their effectiveness. The effects of several inhibitors on known nitrifiers in pure culture were assessed revealing that nitrapyrin remains one of the best inhibitors of both ammonia oxidizing bacteria and archaea (Papadopolou et al., 2020). The use of synthetic nitrification inhibitors has limitations as they are not allowed in organic production and inhibitors have recently been found off-site in streams (Woodward et al., 2016). In New Zealand, DCD residues were detected in milk resulting in the suspension of DCD use in pastures (Thapa et al., 2016). These observations suggest that organic formulations and biological nitrification inhibitors are desirable alternatives.

Certain plants have been observed to slow nitrification by producing chemicals that act as biological nitrification inhibitors. Inhibitory substances have been isolated from the root systems of tropical pasture grasses that are adapted to low-N environments, in particular the roots of *Brachiaria* spp. have high inhibitory activity. Among field crops, sorghum (*Sorghum bicolor*) root extracts have been observed to inhibit nitrification (Subbarao et al., 2015). Incorporation of these plants into rotations or pasture systems may help to retain N in these soils systems by preventing nitrate accumulations. The search continues for biological nitrification inhibitors for use in the major grain crops especially maize.

Summary and outlook

Nitrification and nitrifying organisms have been studied for well over 100 years yet scientists have recently discovered novel organisms and processes that are involved in soil nitrification around the globe. The diverse nitrifying microbes remain an important example of how management of a microbial process and community may impact agricultural sustainability and the production of greenhouse gases. As we learn more about the ecophysiology of the different groups of microbes responsible for the oxidation of ammonium and ammonia to nitrite and nitrate, we may have the ability to fine-tune the nitrification process in our agroecosystems, improving the use of a vital but limiting soil resource, plant available nitrogen.

References

- Abalos D, Jeffery S, Sanz-Cobena A, Guardia G, and Vallejo A (2014) Meta-analysis of the effect of urease and nitrification inhibitors on crop productivity and nitrogen use efficiency. *Agriculture, Ecosystems & Environment* 189: 136–144.
- Ali M, Oshiki M, Awata T, Isobe K, Kimura Z, Yoshikawa H, Hira D, Kindaichi T, Satoh H, and Fujii T (2015) Physiological characterization of anaerobic ammonium oxidizing bacterium, 'Candidatus Jettenia caeni'. *Environmental Microbiology* 17: 2172–2189.
- Alves RJE, Minh BQ, Ulrich T, von Haeseler A, and Schleper C (2018) Unifying the global phylogeny and environmental distribution of ammonia-oxidising archaea based on amoA genes. *Nature Communications* 9: 1517.
- Chain P, Lamerdin J, Larimer F, Regala W, Lao V, Land M, Hauser L, Hooper A, Klotz M, Norton J, Sayavedra-Soto L, Arciero D, Hommes N, Whittaker M, and Arp D (2003) Complete genome sequence of the ammonia-oxidizing bacterium and obligate chemolithoautotroph *Nitrosomonas europaea*. *Journal of Bacteriology* 185: 2759–2773.
- Clarke K and Beegle DB (2014) *Nutrient Management to Improve Nitrogen Use Efficiency and Reduce Environmental Losses*. Pennsylvania State University Extension. Retrieved 14-August from https://extension.psu.edu/programs/nutrient-management/educational/soil-fertility/nutrient-management-to-improve-nitrogen-use-efficiency-and-reduce-environmental-losses/extension_publication_file.
- Daims H, Lebedeva EV, Pjevac P, Han P, Herbold C, Albertsen M, Jehmlich N, Palatinszky M, Vierheilig J, and Bulaev A (2015) Complete nitrification by *Nitrospira* bacteria. *Nature* 528: 504–509.
- Duan S, Zhang Y, and Zheng S (2021) Heterotrophic nitrifying bacteria in wastewater biological nitrogen removal systems: A review. *Critical Reviews in Environmental Science and Technology* 52(13): 1–37.
- Frankland PF and Frankland GC (1890) V. The nitrifying process and its specific ferment.—Part I. *Philosophical Transactions of the Royal Society of London B*: 107–128.
- Giguere AT, Taylor AE, Suwa Y, Myrold DD, and Bottomley PJ (2017) Uncoupling of ammonia oxidation from nitrite oxidation: Impact upon nitrous oxide production in non-cropped Oregon soils. *Soil Biology and Biochemistry* 104: 30–38.
- Jetten MSM, Niftrik LV, Strous M, Kartal B, Keltjens JT, and Op den Camp HJM (2009) Biochemistry and molecular biology of anammox bacteria. *Critical Reviews in Biochemistry and Molecular Biology* 44: 65–84.

- Koch H, Lucker S, Albertsen M, Kitzinger K, Herbold C, Spieck E, Nielsen PH, Wagner M, and Daims H (2015) Expanded metabolic versatility of ubiquitous nitrite-oxidizing bacteria from the genus *Nitrospira*. *Proceedings of the National Academy of Sciences of the United States of America* 112: 11371–11376.
- Lancaster KM, Caranto JD, Majer SH, and Smith MA (2018) Alternative bioenergy: Updates to and challenges in nitrification metalloenzymology. *Joule* 2: 421–441.
- Lehtovirta-Morley LE, Ross J, Hink L, Weber EB, Gubry-Rangin C, Thion C, Prosser JI, and Nicol GW (2016) Isolation of '*Candidatus Nitrosocosmicus franklandus*', a novel ureolytic soil archaeal ammonia oxidiser with tolerance to high ammonia concentration. *FEMS Microbiology Ecology* 92: fiw057.
- Liang D and Robertson GP (2021) Nitrification is a minor source of nitrous oxide (N_2O) in agricultural landscapes and declines with increasing management intensity. *Global Change Biology* 27(21): 5599–5613.
- Norton JM (2011) Diversity and environmental distribution of ammonia-oxidizing bacteria. In: Ward BB, Klotz MG, and Arp DJ (eds.) *Nitrification*, pp. 39–55. Washington, D.C.: ASM Press.
- Norton JM, Klotz MG, Stein LY, Arp DJ, Bottomley PJ, Chain PSG, Hauser LJ, Land ML, Larimer FW, Shin MW, and Starkenburg SR (2008) Complete genome sequence of *Nitrosospora multififormis*, an ammonia-oxidizing bacterium from the soil environment. *Applied and Environmental Microbiology* 74: 3559–3572.
- Norton JM and Ouyang Y (2019) Controls and adaptive management of nitrification in agricultural soils. *Frontiers in Microbiology* 10: 1931.
- Norton JM and Stark JM (2011) Regulation and measurement of nitrification in terrestrial systems. In: Klotz MG (ed.) *Research on Nitrification and Related Processes*. Vol. 486, pp. 343–368. San Diego, CA USA: Academic Press.
- Oshiki M, Satoh H, and Okabe S (2016) Ecology and physiology of anaerobic ammonium oxidizing bacteria. *Environmental Microbiology* 18: 2784–2796.
- Ouyang Y, Evans SE, Friesen ML, and Tiemann LK (2018) Effect of nitrogen fertilization on the abundance of nitrogen cycling genes in agricultural soils: A meta-analysis of field studies. *Soil Biology and Biochemistry* 127: 71–78.
- Papadopoulou ES, Bachtsevani E, Lampronikou E, Adamou E, Katsaouni A, Vasileiadis S, Thion C, Menkissoglu-Spiroudi U, Nicol GW, and Karpouzias DG (2020) Comparison of novel and established nitrification inhibitors relevant to agriculture on soil ammonia- and nitrite-oxidizing isolates. *Frontiers in Microbiology* 11: 2795.
- Prosser JI (2011) Soil nitrifiers and nitrification. In: Ward BB, Arp DJ, and Klotz MG (eds.) *Nitrification*, pp. 347–383. Washington, DC: American Society for Microbiology.
- Prosser JI, Head IM, and Stein L (2014) The family nitrosomonadaceae. In: Rosenberg E, DeLong E, Lory S, Stackebrandt E, and Thompson F (eds.) *The Prokaryotes: Alphaproteobacteria and Betaproteobacteria*, pp. 901–918. Berlin Heidelberg: Springer-Verlag.
- Schleper C and Nicol GW (2010) Ammonia-oxidizing archaea - physiology, ecology and evolution. In: Poole RK (ed.) *Advances in Microbial Physiology*. vol. 57, pp. 1–41. London: Academic Press Ltd-Elsevier Science Ltd.
- Sedlacek CJ, McGowan B, Suwa Y, Sayavedra-Soto L, Laanbroek HJ, Stein LY, Norton JM, Klotz MG, and Bollmann A (2019) A physiological and genomic comparison of *Nitrosomonas* cluster 6a and 7 ammonia-oxidizing bacteria. *Microbial Ecology* 78: 985–994.
- Starkenburg SR, Chain PSG, Sayavedra-Soto LA, Hauser L, Land ML, Larimer FW, Malfatti SA, Klotz MG, Bottomley PJ, Arp DJ, and Hickey WJ (2006) Genome sequence of the chemolithoautotrophic nitrite-oxidizing bacterium *Nitrobacter winogradskyi* Nb-255. *Applied and Environmental Microbiology* 72: 2050–2063.
- Stein LY (2011) Heterotrophic nitrification and nitrifier denitrification. In: Ward BB, Arp DJ, and Klotz MG (eds.) *Nitrification*. Washington, D.C: ASM Press.
- Stieglmeier M, Klingl A, Alves RJE, Rittmann S, Melcher M, Leisch N, and Schleper C (2014) *Nitrososphaera viennensis* gen. nov., sp. nov., an aerobic and mesophilic, ammonia-oxidizing archaeon from soil and a member of the archaeal phylum Thaumarchaeota. *International Journal of Systematic and Evolutionary Microbiology* 64: 2738–2752.
- Subbarao GV, Yoshihashi T, Worthington M, Nakahara K, Ando Y, Sahrawat KL, Rao IM, Lata J-C, Kishii M, and Braun HJ (2015) Suppression of soil nitrification by plants. *Plant Science* 233: 155–164.
- Thapa R, Chatterjee A, Awale R, McGranahan DA, and Daigh A (2016) Effect of enhanced efficiency fertilizers on nitrous oxide emissions and crop yields: A meta-analysis. *Soil Science Society of America Journal* 80: 1121–1134.
- Tian H, Xu R, Canadell JG, Thompson RL, Winiwarter W, Suntharalingam P, Davidson EA, Ciais P, Jackson RB, and Janssens-Maenhout G (2020) A comprehensive quantification of global nitrous oxide sources and sinks. *Nature* 586: 248–256.
- van Kessel MAHJ, Speth DR, Albertsen M, Nielsen PH, Op den Camp HJM, Kartal B, Jetten MSM, and Lucker S (2015) Complete nitrification by a single microorganism. *Nature* 528: 555.
- Venterea RT and Rolston DE (2000) Mechanistic modeling of nitrite accumulation and nitrogen oxide gas emissions during nitrification. *Journal of Environmental Quality* 29: 1741–1751.
- Winogradsky S (1890-1891) Recherches sur les Organismes de la Nitrification. *Annales de l'Institut Pasteur* 4: 215–231.. 257–275, 760–811; 5: 92–100, 577–616.
- Woodward EE, Hladik ML, and Kolpin DW (2016) Nitrapyrin in streams: The first study documenting off-field transport of a nitrogen stabilizer compound. *Environmental Science & Technology Letters* 3: 387–392.
- Xia L, Lam SK, Chen D, Wang J, Tang Q, and Yan X (2017) Can knowledge-based N management produce more staple grain with lower greenhouse gas emission and reactive nitrogen pollution? A meta-analysis. *Global Change Biology* 23: 1917–1925.

Microbial denitrification

Michael J Goss^a, Sergey Blagodatsky^b, and Adrian Unc^c, ^aSchool of Environmental Sciences, University of Guelph, Guelph, ON, Canada; ^bInstitute of Zoology, University of Cologne, Cologne, Germany; ^cSchool of Science and the Environment, Memorial University of Newfoundland, Corner Brook, NL, Canada

© 2023 Elsevier Ltd. All rights reserved.

This is an update of the chapter written by the late D.A. Martens, Denitrification, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 378–382, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00138-7>.

Introduction	261
Definitions and pathways	262
Substrates, mechanisms and organisms	262
Aerobic oxidation of ammonia – Nitrification	262
Anaerobic oxidation of ammonia	263
Anaerobic oxidation of methane	264
Canonical denitrification	264
Genetic insights into microbial diversity, and diversity of genes encoding denitrification enzymes	265
Denitrification under field conditions	266
Heterotrophic denitrification to N₂O	267
Nitrifier denitrification	267
Dissimilatory nitrate reduction to ammonia or nitrate ammonification	267
Environmental factors	267
Organic C availability	268
Controls on O ₂ availability	268
Nitrate supply	268
Temperature	269
pH	269
Conclusion	269
References	269

Abstract

Denitrification is the transformation of plant available nitrogen (N) to gaseous forms that can return N to the atmospheric pool. Nitrous oxide, N₂O, an important greenhouse gas is formed during both aerobic and anaerobic processes but the concentration of oxygen in the soil determines which processes predominate. The varied nature and uncertainties within the pathways, together with the range of gaseous products makes quantification of losses from soil particularly difficult. Fertilizer N accounts for about 43% of global N₂O releases to the atmosphere from agriculture. The chapter summarizes the fundamental processes of N-transformation that generate gaseous forms of N.

Key points

- Denitrification is the process returning fixed N to the atmospheric pool
- Oxygen depletion triggers denitrification, thereby creating conditions for NO₂[−] and NO₃[−] to be employed as electron acceptors
- Denitrification is carried out mostly by facultative heterotrophic Bacteria, Archaea and, more limited, by certain fungi, often by same organisms that may also carry genes encoding for nitrogenases (N fixers)
- Despite advances in molecular biology, the relationship between gene abundance or enzyme production and denitrification is yet to be consistently quantified
- Abiotic factors still offer the most easily measured inputs for limited modelling of denitrification rates and fluxes

Introduction

Denitrification is the transformation of nitrogen (N) compounds, which are plant available in soil, into gaseous forms, including the climate-forcing N oxides, nitric oxide (NO), nitrogen dioxide (NO₂), nitrous acid (HNO₂) and nitrous oxide (N₂O), and the inert dinitrogen (N₂), that are returned to the atmospheric N-pool. Early research on denitrification was stimulated by the inability

to obtain a mass balance of total inputs and outputs of N in agricultural systems. The large portion of the N left unaccounted was determined to be lost as gaseous N, and which resulted, agronomically, in decreased N fertilizer efficiency.

Research into global change processes identified that one of the gases lost, N_2O , is a potent greenhouse gas with important impacts on our environment. It is currently considered to have a global warming potential over a 100-year that is about 273 times as strong as carbon dioxide, partly due its atmospheric lifetime of about 120 years and because of its significant depletion of the earth's stratospheric ozone layer. As the amounts of industrial or biologically fixed N_2 used for crop production increase, the production of NO and N_2O due to nitrification and denitrification processes will also increase and potentially cause and contribute to a warming of the earth's surface by influencing the radiative budget of the troposphere.

Increasing concern over N_2O as a climate-forcing gas has focused the spotlight on soil use and management for agriculture and forestry, which in combination form the largest source of N_2O emissions in the USA and accounted for 75% of total N_2O emissions in USA during 2021. Rates of N loss from fertilizer applied to agricultural soils through denitrification can vary tremendously, ranging from 0 to 70% of applied N, with mean estimates being about 25%.

The abiotic transformation of N, chemodenitrification, is one mechanism that contributes to total denitrification, especially in some paddy soils and in marine sediments, with nitrate (NO_3^-) and nitrite (NO_2^-) potentially being chemically reduced to N_2O in the presence of Fe^{2+} in an anoxic, neutral to alkaline environment. The soil organic matter (SOM) content is also important, with NO_2^- being preferentially reduced if there is a smaller content but the NO_3^- is preferred at high levels. However, as it is the soil population of microorganisms that plays the dominant role the regulating the N cycle for plant-available N, including the return of N to the atmosphere, the chapter will focus on microbial denitrification.

Definitions and pathways

A number of microbial processes form N_2O and other reduced gaseous forms of N in soil, including nitrification, the oxidation of ammonia, which in water at small concentration is dissociated as the ammonium cation (NH_4^+), to NO_2^- and NO_3^- . The formation of NH_4^+ can be seen as the first step in making N, present in organic form in soil, available as a nutrient to higher plants. However, being a cation, it readily adsorbs to negatively charged surfaces and is therefore of very limited mobility compared with the negatively charged oxides.

Soil organic matter mineralization and inorganic N-fertilizer inputs, now predominantly urea-based ($\text{CO}(\text{NH}_2)_2$), are the main sources of NH_4^+ in the environment. Ammonia oxidation is carried out by several groups of microorganisms through four pathways, including the autotrophic aerobic ammonia oxidization pathway by ammonia-oxidizing bacteria (AOB), archaea (AOA), or complete ammonia oxidization (comammox) bacteria (limited to a few species in the genus *Nitrospira*); heterotrophic aerobic ammonia oxidization (or heterotrophic nitrification) pathway by heterotrophic nitrifying bacteria (HNB, or heterotrophic AOB) and fungi as well as the anaerobic autotrophic ammonia oxidization (anammox) pathway by anaerobic AOB (Table 1).

Substrates, mechanisms and organisms

Aerobic oxidation of ammonia – Nitrification

Nitrification is the oxidation of NH_4^+ or NH_3 via NO_2^- to NO_3^- . Heterotrophic oxidation begins with the oxidation of organic forms of nitrogen to NH_4^+ . Most aerobic oxidation of ammonia in soil is carried out by chemolithoautotrophic members of bacteria (AOB) and archaea (AOA). These organisms first use the enzyme 'ammonia monooxygenase' (amoA) (the polypeptide at the active site of is encoded by the amoA functional gene) to form hydroxylamine (NH_2OH). The archaeal form of amoA has greater affinity for NH_3 than bacterial amoA, which may allow Archaea to be more competitive when substrate concentrations are small, or soil pH is more acid.

In both AOA and AOB the formation of NO_2^- depends on the oxidation of NH_2OH . However, enzyme responsible, 'hydroxylamine oxidoreductase,' only forms NO and to date no enzyme has been identified for oxidizing the NO to NO_2^- . The exact pathway operating in AOA is also still uncertain, possibly following that for AOB or it is possible that NH_2OH reacts with NO to form two molecules of NO_2^- , one of which is reduced to NO and cycles back to react with another NH_2OH molecule. Completion of the oxidation process to form NO_3^- depends on another group of chemolithoautotrophs, the nitrite-oxidizing bacteria (NOB).

Both AOA and AOB release N_2O and HNO_2 following the synthesis of NH_2OH , which is a potentially unstable intermediate in nitrification. Both gases may form without enzyme activity (abiotic), although it is possible that some HNO_2 is produced via an unidentified enzyme. The formation of NO from NH_2OH is the second point where N_2O can be released by AOB together with NO and NO_2 (Fig. 1). In AOA, it appears that it is only NO_2 that is produced at this point. The formation of NO and HNO_2 may also occur following the formation of NO_2^- .

Table 1 Main microbial processes in which reduced nitrogen gasses are formed from organic or mineral forms available to plants.

Microbial process	Main Groups of organisms	Key pathway steps	Reduced gaseous N forms
Nitrification	A. Autotrophic ammonium oxidisers: ammonia-oxidizing archaea (AOA)	AOA: $\text{NH}_4^+ + \text{O}_2 + \text{H}^+ \rightarrow \text{NH}_2\text{OH} + \text{H}_2\text{O}$ $\text{NH}_2\text{OH} + 2\text{H}_2\text{O} + \text{NO} \rightarrow 2\text{NO}_2^- + 7\text{H}^+ +$ $\text{NH}_2\text{OH} + \text{NO} \rightarrow \text{N}_2\text{O} \uparrow + \text{H}_2\text{O} + \text{H}^+$	
	and bacteria (AOB)	AOB: $\text{NH}_4^+ + \text{O}_2 + \text{H}^+ \rightarrow \text{NH}_2\text{OH} + \text{H}_2\text{O} \rightarrow$ $\text{NO}_2^- + 5\text{H}^+$ $2\text{NO} + 2\text{H}^+ \rightarrow \text{N}_2\text{O} + \text{N}_2\text{O} \uparrow + \text{H}_2\text{O}$	N_2O , NO, NO_2 , HNO_2
	B. Heterotrophic ammonium oxidizers (HAOx)	HAOx $\text{Organic-N} \rightarrow \text{NH}_4^+ \rightarrow \text{NO}_2^- \Rightarrow \text{NO}_3^-$	
	C. Nitrite oxidizing bacteria (NOB) D. Comammox	$\text{NO}_2^- + \text{H}_2\text{O} \rightarrow \text{NO}_3^- + 2\text{H}^+ + 2\text{e}^-$ $\text{NH}_4^+ + 2\text{O}_2 \rightarrow \text{NO}_3^- + \text{H}_2\text{O} + 2\text{H}^+$	None N_2O , NO, NO_2 , HNO_2
Nitrifier-denitrification	AOA, AOB, HAOx	$\text{NH}_4^+ \rightarrow \text{NO}_2^- \rightarrow \text{NO} \rightarrow \text{NO} \rightarrow$ $\text{N}_2\text{O} \uparrow \Rightarrow \text{N}_2 \uparrow$	NO, N_2O , N_2
Denitrification	HAOx	$\text{NO}_3^- \rightarrow \text{NO} \rightarrow \text{N}_2\text{O} \text{ or } \text{N}_2$	$\text{NO} \rightarrow \text{N}_2\text{O} \text{ or } \text{N}_2$
Dissimilatory nitrate reduction to ammonia	DNR(A) bacteria and fungi	$\text{NO}_3^- \rightarrow \text{NO}_2^- \rightarrow \text{NH}_4^+$ ↓ $\text{N}_2\text{O} \text{ or } \text{N}_2$	N_2O , N_2
Anaerobic ammonium oxidation	Anammox bacteria	$\text{NH}_4^+ + \text{NO}_2^- \rightarrow \text{N}_2 \uparrow + 2\text{H}_2\text{O}$	N_2O , N_2
N-dependent methane oxidation	Anaerobic methane-oxidizing bacteria	a) Nitrite $3\text{CH}_4 + 8\text{NO}_2^- + 8\text{H}^+ \rightarrow 3\text{CO}_2 + 4\text{N}_2 + 10\text{H}_2\text{O}$ b) Nitrate $5\text{CH}_4 + 8\text{NO}_2^- + 8\text{H}^+ \rightarrow 5\text{CO}_2 + 4\text{N}_2 + 14\text{H}_2\text{O}$	N_2 , N_2O

Organisms that are capable of the complete oxidation of to NH_4^+ to NO_3^- - comammox bacteria – also release N_2O , NO, NO_2 and HNO_2 during aerobic nitrification at the same points as indicted above (Han et al., 2021). Aerobic production of N_2O is considered to be the dominant source of this gas in soils with less than 60% water-filled pore space.

Although the number of archaeal cells carrying out ammonia oxidation is often greater than bacterial cells, the range in yield of N_2O per NO_2^- molecule produced per cell for AOB is from 0.1 to 8%, for AOA it is only 0.04 to 0.3%. AOA are more active than AOB in acidic soils or when concentrations of NH_4^+ are at levels typically associated with organic matter undergoing mineralization. In well-fertilized soils, AOB are more active, being responsible for nearly four times more N_2O production than AOA. The release of NO and HNO_2 is also greater from AOB than AOA or comammox bacteria.

Some nitrifying bacteria that oxidize NH_3 to NO_2^- have some of the denitrifying enzymes that lead to N_2O production. For example, *Nitrosomonas europaea* has a nitrite reductase enzyme and a nitric oxide reductase but no enzyme capable of reducing N_2O . This pathway is known as nitrifying denitrification.

Nitrification has also been found to occur in fungi and heterotrophic bacteria that utilize the same substrate, intermediates and products as autotrophic nitrifiers, and are often able to form N_2O under aerobic conditions. Release of N_2O by heterotrophic nitrification is considered a minor contribution to total N_2O production in soil.

Anaerobic oxidation of ammonia

Formation of N_2O not only occurs during aerobic nitrification but can also be released by autotrophic anaerobic ammonia oxidation (anammox) bacteria. In this case NH_3 is activated by reaction with NO to form hydrazine (N_2H_4), a very powerful reductant. The NO is probably produced by the reduction of NO_2^- by a hydroxylamine oxidoreductase (HAO) reductive homolog HAO_r. This step is the likely source of N_2O , which is one of the gases released in the anammox process. The final step in anammox is the dehydrogenation of N_2H_4 to N_2 gas. The main catabolic enzymes for the process are found in an anammoxosome, a large vacuolar cell organelle with a very dense membrane that prevents the outwards diffusion of the N_2H_4 .

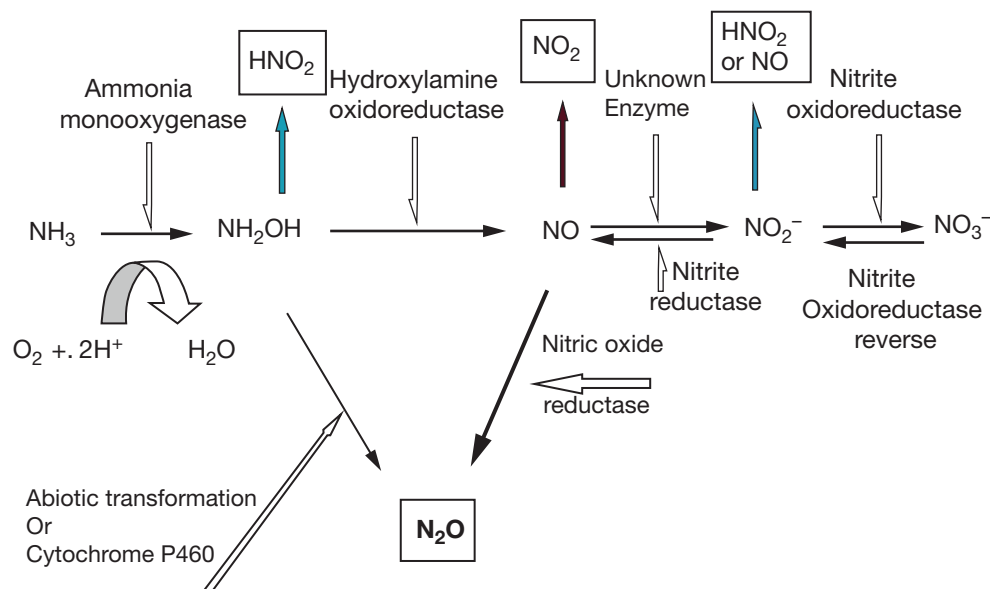
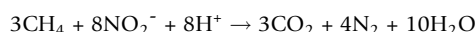


Fig. 1 Pathway and enzymes involved in nitrification.

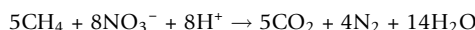
Anaerobic oxidation of methane

Methanotrophic members of the Archaea and Bacteria use methane as a carbon and energy source. Methanotrophy is an important process in global methane cycling in that it reduces the net methane emission to the atmosphere. Under anoxic conditions in soils and rice paddies, as well as in marine, estuarine and lake conditions, specific methanotrophs have specialized to use different electron acceptors or transferring electrons to syntrophic (sulfate-reducing) bacterial partners. In NO₂⁻ dependent anaerobic methane oxidation (DAMO) metal and metalloid ions, such as Fe³⁺ and Mn⁴⁺, as well as SO₄²⁻, NO₂⁻ and NO₃⁻, are used as electron acceptors and reduced directly. A two-member consortium of a bacteria and an archaeon distantly related to marine methanotrophic Archaea was first identified to carry out DAMO (Raghoebarsing et al., 2006). Eventually, members of the genus *Methyloirabilis*, belonging to the uncultured NC10 phylum, were identified to have DAMO capacity (Guerrero-Cruz et al., 2019). However, while some evidence of *Methyloirabilis* spp. presence in soil has been provided (Bannert et al., 2012), to date its presence is most clearly confirmed for consistently saturated paddy soils, as it is suggested that is a slow grower, sensitive to variability in the redox state and establishes more consistently in more stable anaerobic conditions (Su et al., 2023).

For Nitrite-DAMO, the reaction for methane oxidation is:



And for Nitrate-DAMO:



In both cases, the oxide is reduced to NO and then to N₂ gas and oxygen. The latter is used to oxidize methane mediated by the enzyme methane mono-oxygenase. In a peatland investigation, it appeared that NO₃⁻ acted to inhibit the use of NO₂⁻ as the electron acceptor and was much less effective itself in that role. The known microbes able to carry out this n-Damo process are *Candidatus Methyloirabilis oxyfera* and *Candidatus Methanoperedens nitroreducens*.

Canonical denitrification

Canonical denitrification is defined as the “microbial reduction of nitrate or nitrite coupled to electron transport phosphorylation resulting in gaseous N either as molecular N₂ or as an oxide of N.” The important point being the capture of energy in the high energy phosphate bond in adenosine triphosphate (ATP).

This denitrification process involves a stepwise reduction of N oxides by bacteria or archaea to gaseous products of N₂O or N₂ under limited O₂ conditions (Fig. 2). As indicated above, the process involves the use of NO₂⁻ and NO₃⁻ as terminal electron acceptors instead of molecular O₂ and is irreversible once NO is formed. Thus, a source of organic C is required for microbial metabolism and sufficient NO₃⁻ must be available to act an electron acceptor in an environment with limited O₂. All three conditions, a C source, a low level of O₂ and sufficient NO₃⁻ must be met for denitrification to take place.



Fig. 2 Pathway and enzymes involved in denitrification. Although this Figure was of acceptable resolution, it was also redrafted and resubmitted.

Table 2 Reductases of oxides of nitrogen used in denitrification processes.

Gene markers	Enzyme and function	Organisms
amoA	ammonia monooxygenase subunit A Oxidation of NH_3 to NH_2OH	Autotrophic and heterotrophic Archaea, Bacteria, Fungi.
hao	Hydroxylamine oxidoreductase Oxidation of NH_2OH to NO	Autotrophic and heterotrophic Archaea, Bacteria, Fungi.
hdh	Hydrazine dehydrogenase Oxidation of N_2H_4 to N_2	Anammox bacteria
hzo	Hydrazine oxidoreductase Oxidation of N_2H_4 to N_2	Anammox bacteria
hzs	Hydrazine synthase Reduction of NO to NH_2OH and subsequent condensation with NH_3 to form N_2H_4	Anammox bacteria
nap	Dissimilatory periplasmic nitrate reductase Reduces nitrate to nitrite	Heterotrophic denitrifiers DNRA
nar	Membrane-bound nitrate reductase (narG) Reduction of NO_3^- to NO	Heterotrophic denitrifiers DNRA
nif	Nitrogenase Reduction of N_2 to NH_3	Diazotrophic bacteria and archaea
nir, nirB, nirS, nirK	Nitrite reductases Reduction of NO_2^- to NO by cytoplasmic (nirB), haeme-containing (cd1-S type,) cytochrome or copper-containing (Cu-type, K) enzymes	AOB, AOA, Heterotrophic denitrifiers DNRA organisms Anammox bacteria
nod	Nitric oxide oxidase. Direct oxidation of NO to NO_3^-	Anammox bacteria
nor, cnorB, qnorB	Nitric oxide reductase Reduction of NO to N_2O Cytochrome-containing (cnorB) or quinol-dependent (qnorB)	AOB, Heterotrophic denitrifiers
nos	Nitrous oxide reductase (nosZ, multi-copper) Reduction of N_2O to N_2	Denitrifiers
nrf	dissimilatory periplasmic cytochrome <i>c</i> nitrite reductase Reduction of NO_2^- to NO or NH_3	DNRA
nxr	Nitrite oxidoreductase Oxidation of NO_2^- to NO_3^-	Anammox
$\text{nxr}_{(\text{rev})}$	NO_2^- reduction to HNO_2	

N_2O is produced by the sequential action of NO_3^- , NO_2^- and NO reductases (Table 2), which, however, are generally tolerant of O_2 .

Both the iron-containing cytochrome (S type) nitrite reductase and nitric oxide reductase can catalyze the four-electron reduction of O_2 to water, essential for energy production as ATP in aerobic organisms.

The integrated relationship between the various processes forming the nitrogen cycle is illustrated in Fig. 3.

Genetic insights into microbial diversity, and diversity of genes encoding denitrification enzymes

The advent of Next Generation of Sequencing tools has shown that denitrification genes are widespread in the microbial communities. The stepwise respiratory reduction of NO_3^- to gaseous end products has been shown to involve diverse bacterial taxa of Proteobacteria, Firmicutes, Actinomycetes, Bacteroidetes and Aquificae, and as well as fungi. The polyphyletic origin of denitrification makes it difficult to apply a simple taxonomical approach to describe the primary taxa involved in denitrification. Thus, to understand the diversity of the microbial community involved in respiratory denitrification the diversity of specific denitrification-related bacterial genes is commonly employed. Fig. 2 shows the generic gene markers for the different steps. Reduction of nitrate to nitrite (NO_3^- to NO_2^-), mediated by both denitrifiers and nitrate respirers, involves nitrate reductase encoded by the nar (e.g., respiratory nitrate reductase subunit alpha [narG], respiratory nitrate reductase subunit beta [narH]) and nap (e.g., periplasmic nitrate reductase [napA]) genes. Nitrite reductases encoded by the nir (e.g., Cu-type nitrite reductase [nirK], cytochrome cd1-type nitrite reductase [nirS]) genes are responsible for the reduction of nitrite to nitric oxide (NO_2^- to NO). Reduction to nitrous oxide (N_2O) is catalyzed by nitric oxide reductase, which is encoded by nor (e.g., quinol-oxidizing

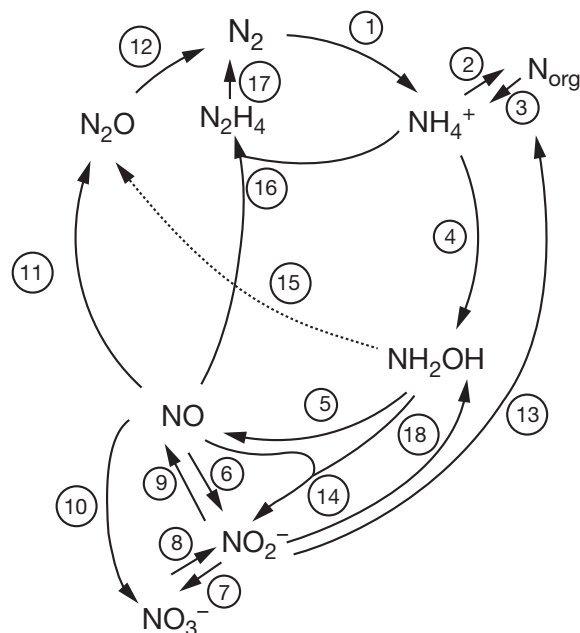


Fig. 3 Key steps in the nitrogen cycle and the formation of nitrous oxide, nitric oxide and dinitrogen. Step 1, microbial conversion of dinitrogen gas to ammonia (NH_3) – biological nitrogen fixation by diazotrophic bacteria and archaea– mediated by nitrogenase (*nif*) Step 2, assimilation of NH_3 into amino acids through the glutamine synthetase (*gln*)-glutamate synthase (glutamine 2-oxoglutarate amidotransferase (*glt*) [*GOGAT*]) pathway, which occur in most organisms. Step 3, ammonification of organic matter to release ammonia can proceed with reductive deamination using glutamate dehydrogenase (*gdh*). Step 4, the oxidation of NH_3 to hydroxylamine (NH_2OH) mediated by ammonia monooxygenase (*amo*) in ammonia oxidizing bacteria (AOB), archaea (AOA) and complete ammonia oxidation by Comammox bacteria. Step 5, the oxidation of NH_2OH to nitric oxide mediated by hydroxylamine oxidoreductase (*hao*) in AOB, organisms capable of complete oxidation of ammonia (Comammox) and by hydroxylamine oxidase by some of those capable of anaerobic ammonia oxidation (Anammox). Step 6, the oxidation of NO to nitrite (NO_2^-) by an unknown enzyme. Step 7, the oxidation of NO_2^- to nitrate (NO_3^-) by nitrite oxidoreductase (*nxr*) in nitrite oxidizing bacteria, including Comammox. Step 8, the reduction of NO_3^- to NO_2^- by the same enzyme acting in the reverse. Step 9, the reduction of NO_2^- to NO by nitrite reductase (*nir*) in AOA, AOB, Comammox organisms and heterotrophic denitrifiers. It can also be induced abiotically. Step 10, direct oxidation of NO to NO_3^- mediated by nitric oxide oxidase (*nod*). Step 11, the reduction of NO to N_2O by nitric oxide reductase (*nor*) in AOB, which is a particularly significant source of N_2O . Step 12, the reduction of N_2O to N_2 mediated by nitrous oxide reductase (*nos*) in denitrifier organisms and with Step 13 forms the dissimilatory nitrate reduction to ammonia (DNRA) pathway. Step 13, the reduction of NO_2^- to NH_3 by nitrite reductase (*nrfa*) in DNRA. Step 14, In AOA, there is no known *hao* homolog, so it is hypothesized that NO and NH_2OH react together via a copper enzyme complex to form two molecules of NO_2^- . Step 15, the oxidation of NH_2OH can occur abiotically as well as via cytochrome P460. Step 16, the formation of hydrazine (N_2H_4) from NO and NH_2OH mediated by hydrazine synthetase in Anammox bacteria. Step 17, the formation of N_2 from N_2H_4 mediated by hydrazine dehydrogenase in Anammox bacteria. Step 18, the reduction of NO_2^- to NH_2OH rather than to NO has been found in some Anammox bacteria.

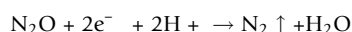
single-subunit class [qnorB], cytochrome *bc*-type complex [cnorB]) genes. The final step, leading to the production of dinitrogen (N_2) is catalyzed by the oxide reductase encoded by the *nosZ* group of genes. The types and amount of the organic carbon can affect microbial activity and thus the proportions of these genes. Some denitrifiers maintain the whole four-gene pathway (*nar*, *nir*, *nor* and *nos*), but others possess partial pathways. It is unclear when ones or the other are favored in a particular environment. Arguably, changes in the ratios of these different genes might inform on the status of the denitrification processes in a soil. Genetic variation in these genes is, however, not easily translated into an extended phenotype for a population of denitrifiers. Furthermore, the sensitivity to environmental conditions of genes encoding equivalent functions may vary; for e.g., *nirS*-type denitrifiers was reported to be more sensitive to the soil type, soil moisture, pH, and rhizosphere effect than *nirK*. Such variability suggests both that denitrification can be affected by changes in the environment, and also that redundancies in encoding genes make it resilient to same changes. Generally, the rates of the transient production of NO_2^- , NO and N_2O that occurs during transition from oxic to anoxic respiration are better related to the abiotic conditions and seem not to be dependent on the taxonomy of the communities of denitrifiers.

Denitrification under field conditions

Measurements of loss of gaseous forms of N from soil under field conditions have large coefficient of variation, which has been attributed to the formation of soil “hotspots” that exhibit increased soil respiration. Spatial variability in N_2O fluxes typically causes coefficients of variation from 30 to 100% to be measured at a spatial scale of several meters with field plots. Hotspots of organic debris can stimulate denitrification due to rapid microbial mineralization and release of NH_4^+ leading to NO_3^- with concomitant depletion of O_2 concentrations. The best field indicator for denitrification has been found to be drainage class, which determines the potential for the soil to become waterlogged at times of intense precipitation. Seasonal differences in frequently noted for denitrification events (less in spring vs. more in fall) have been attributed to smaller concentrations of NO_3^- in spring in temperate soils. Larger inputs of organic material close to the soil surface supports greater microbial activity and can result in increased

denitrification compared with that in deeper layers with less C available. Furthermore, 30–80% of the N_2O produced at deeper soil layers can be reduced to N_2 before exiting the soil (Clough et al., 2005).

The genera of denitrifying bacteria are diverse. The capacity for denitrification has been reported to be present in more than 50 genera of bacteria. Although the distribution of denitrifiers is ubiquitous, the activity or measurement of denitrifier biomass and synthesis of denitrifying enzyme activity from soil field cores is poorly correlated with measured N gas loss. Yeast and filamentous fungi are also capable of gaseous N production. Fungal N_2O production may be of great significance because fungi often dominate the microbial biomass of temperate soils, yet fungi are not tolerant of waterlogged conditions. Denitrifiers are not fermentative (facultative anaerobes), hence their growth under oxygen-limited conditions is solely dependent on the presence of N oxides. In this process, N_2O is an intermediate and so the ratio of $\text{N}_2\text{O}:\text{N}_2$ fluxes can be affected by various environmental factors. The proportion of N_2O compared to N_2 is increased if the soil pH is decreased, if the amount of NO_3^- is increased, and if smaller O_2 concentrations are present. In each case, the environmental factors have an influence on the enzyme nitrous oxide reductase.



A clear distinction between this canonical denitrification pathway and other denitrification pathways is necessary as the relative proportion of N_2O production from the various mechanisms is affected by the environmental conditions that prevail. The denitrification pathway is prevalent once NO_3^- is formed and the correct environmental conditions of O_2 concentrations, and soluble C are present with microbial reduction of the N oxides resulting.

Heterotrophic denitrification to N_2O

Production of N_2O and NO following the reduction of NO_3^- or NO_2^- to NH_3 increases as the O_2 concentration decreases from fully oxygenated (21%). Under hypoxic conditions, between 3 and 0.5%, the contribution to total N_2O production by heterotrophic denitrification can be as much as 50% and can account for all N_2O production under anoxia. The fungus *Aspergillus flavus*, a model organism for heterotrophic nitrification, especially in relation to acidic forest soils, produces NO_3^- for denitrification by oxidation of amino acids, such as aspartate, when its preferred carbon and energy source has been depleted.

Nitrifier denitrification

When O_2 concentrations become insufficient to support NO_2^- oxidation to NO_3^- by NOB, ammonium oxidizers can produce N_2O from NO_2^- in a process known as nitrifier denitrification. This reaction, first identified in AOB, differs from that of canonical denitrifiers because it is not coupled with the oxidation of organic C. The pathway operates under aerobic oxidation of NH_3 but at a slower rate than in micro-aerobic conditions. First, the enzyme nitrite reductase (nir) reduces NO_2^- to NO, which is reduced in turn to N_2O via nitric oxide reductase. It is important to recognize that this production of N_2O does not need anoxic conditions. A portion may be further reduced to N_2 gas.

The relative proportion of nitrifier denitrification to the N_2O budget ranges from 0 to 30% of the total N_2O released. Use of molecular probes specific for genes encoding for heme-type dissimilatory nitrite and nitrous oxide reductase gene homologies derived from *N. europaea* gene sequences has found that prominent nitrifiers of the genera *Nitrosospira* and *Nitrosolobus* may differ from *N. europaea* in the mechanism and capacity for denitrification.

Dissimilatory nitrate reduction to ammonia or nitrate ammonification

Chemoorganoheterotrophic bacteria, methane-oxidizing archaea, and fungi can carry out dissimilatory nitrate denitrification to ammonia (DNRA) in which NO_3^- is first reduced to NO_2^- , replacing O_2 as the electron acceptor in a respiratory electron transport chain. The associated oxidative phosphorylation provides energy in the form of ATP. The further reduction of NO_2^- to NO_3^- can take place, either in the same respiratory process (the respiratory pathway) or using energy from phosphorylation of a carbon substrate (fermentation pathway). Reduction of NO_3^- to NO_2^- is mediated by a periplasmic nitrate reductase encoded by narG. Reduction of NO_2^- in the fermentation pathway involves nitrite reductase enzymes coded by nirB, whereas the enzyme involved in the respiratory pathway is encoded by nrfA. The latter is found in bacteria capable of DNRA from different phyla (e.g., Bacteroides, Firmicutes, Planctomycetes and Proteobacteria) and can be used as a molecular marker for this pathway, with positive associations between gene abundance and DNRA activity (Pandey et al., 2020).

Although the reduction to NH_4^+ can account for over 90% of the NO_3^- entering the DNRA pathway, ^{13}N -labelling has shown that there is some N_2O produced in the process. Under conditions with greater carbon input elevated C:N ratio, production of NH_4^+ by DNRA is enhanced relative to the products of denitrification. In contrast, high levels of nitrate, small C:N ratio, favor denitrification over DNRA.

Environmental factors

Denitrification supports microbial life through respiration as mineralization of a reduced substrate (mainly organic C, but reduced iron or sulfur may also donate electrons) donates electrons via carriers to an oxidized N oxide. The decline in free energy is coupled

with the generation of ATP. In the presence of O_2 , the ATP yield per number of electrons transported is about 1.7 times greater than if NO_3^- is the terminal electron acceptor, with the final efficiency apparently being species dependent.

Organic C availability

The availability of a suitable reduced C source is important because of the strong relationship between denitrification and soil organic matter. More important, the water-soluble portion of the soil organic matter has been found to be a good indicator for potential denitrification. Recent work has also found that organic litter with low C:N ratios has a greater potential to promote higher denitrification activity than litter with higher C:N ratios. Animal manures with a readily decomposable C source and a high N content can greatly enhance soil denitrification rates.

The presence or absence of plants has also been found to affect denitrification rates. Higher denitrification rates have been noted in the plant rhizosphere as compared with bulk soil several millimeters away from the roots if NO_3^- was sufficient, but if NO_3^- availability was limited, the presence of plants decreased denitrification activity. Plant roots have the potential to impact denitrification activity by providing C to support microbial populations and supply electrons for NO_3^- reduction. Roots can also create anaerobic zones through respiratory O_2 consumption when O_2 supply is limited by water conditions. The rooting environment can create a drier soil through evapotranspiration processes and result in increasing O_2 diffusion in the drier soil. Roots can also assimilate NH_4^+ and NO_3^- from the soil solution reducing the availability of substrate to the denitrification pathways.

Due to a typical decrease in available soil C with depth, denitrification activity has also been found to decrease with soil depth and is potentially limited by a lack of available C below 60 to 75 cm. Other soil characteristics can also change with depth concomitant with organic C content such as porosity/permeability, pH, temperature and depth to water table that may also have an impact in denitrification rates.

Controls on O_2 availability

Reduced availability or absence of O_2 is required for both the synthesis and activity of enzymes involved in denitrification. Soil parameters can control overall O_2 availability and influence denitrification rates. The universally recognized inverse relationship between water content and denitrification is due to the slower the diffusion of O_2 to metabolically active sites with greater the amounts of water present in the soil matrix. Bulk soil parameters that have been used to describe O_2 control of denitrification include soil moisture content, soil water potentials, partial pressure of O_2 , O_2 concentration in solution, percentage air-filled porosities and redox potentials. Bulk measurements have shown promise for predicting denitrification activity in laboratory studies, but in field measurements, spatial variability due to hot spots can mask the predictive nature O_2 diffusion or C mineralization rates.

Soil texture is an important factor for determining critical air-filled porosity, as a finer textured soil (less sand content) will have a higher critical air-filled porosity value than found for coarser textured soils. Oxygen concentrations in soil atmospheres at which denitrification has been observed for different soil types ranged from 4% to 17%. Although the bulk soil O_2 concentrations are important, it has been noted that the O_2 concentration at the microsite level are more important for determination of denitrification activity.

The occurrence of anaerobic microsites can explain how an O_2 sensitive process is possible in a well-structured soil having bulk O_2 measurements indicating aerobic conditions. The microsite concept proposes that anaerobic sites occur within centers of water saturated soil aggregates because of limited diffusion of O_2 and consumption of O_2 by microbial metabolism. Research has shown that the occurrence of anaerobic microsites is largely a function of localized C availability and is possible within aggregates exceeding 9 mm. The microsites are more likely to occur in the rhizosphere and with particles decomposing plant litter in zones where water can limit O_2 diffusion. Again, denitrification activity is a function of available C, limited O_2 potentials and a source of N oxides.

Nitrate supply

Research has indicated that the rate of denitrification activity in pure culture studies was independent of NO_3^- concentration. Recent research has reported first order kinetics with NO_3^- concentrations below 40 mg l^{-1} and the kinetics changed from first order to zero order for NO_3^- concentrations above 100 mg. Microbial NO_3^- reduction is an enzyme catalyzed reaction and use of Michaelis-Menten kinetics has found K_m values for NO_3^- reduction to be <15 μM NO_3^- or about 93 μg NO_3^- l^{-1} for several bacterial isolates. The use of K_m values determined that concentrations required for NO_3^- reduction are higher for soils than those found for pure culture or enzyme studies. The rate kinetics also change if the soil has recently received chemical fertilizers. However, Michaelis-Menten kinetics has not been found to accurately describe NO_3^- reduction dependence on available C and NO_3^- concentrations in soil. Several studies have found that kinetic analysis may not reflect denitrification rates, but the rate of NO_3^- diffusion from solution into the actively denitrifying soil aggregate. The evidence indicates that even very low NO_3^- concentrations present in soil solutions are in excess of enzyme and microbial needs for terminal electron acceptors during O_2 stress, but denitrifying activity may actually be limited by the diffusion of NO_3^- into the soil microsites.

Temperature

Denitrification, because of enzyme dependence on temperature for the activity rate, would be expected to increase exponentially with increased temperature within the range of enzyme activity. The temperature effect is impacted by the interaction of temperature on O_2 solubility and O_2 consumption as well as O_2 diffusion. Decreasing temperature will increase O_2 solubility and decrease O_2 consumption. Research has found a minimum soil temperature range of 2.7–10 °C for denitrifying activity and a maximum temperature of about 50 °C. Above 50 °C, chemical decomposition reactions of NO_2^- can complicate the denitrification-temperature relationship. Pure culture experiments with soil isolates found 30 °C to be the optimum temperature for growth, but evidence suggests that denitrifiers are adapted to soils and growth conditions of the specific soil. Bacterial isolates from tropical soils (33 total) showed no growth when incubated at 10 °C whereas 68% of temperate soil isolates (95 total) grew at 10 °C.

pH

The pH requirement for denitrifying organisms appears to be similar to heterotrophic organisms. In the neutral pH range from 6 to 8, there is a limited effect of pH on the activity of denitrification, but denitrification rates can be limited in acid pH ranges. Adjusting a naturally acid soil (pH 3.5) with reduced microbial activity to pH 6.5 rapidly stimulated a strongly increase in denitrification rates and suggests that a general population with neutral growth optimum was present in a low pH environment. Research has suggested that the decrease in denitrification activity in a moderately acid environment favors NO_3^- reduction to NH_4^+ and that increased soil acidity may favor NO_3^- reduction rather than denitrification. Isotopic evidence has suggested that although moderate change in pH (pH 6.7 to 5.2) had little effect on the total rate of N_2 and N_2O production, the pH adjustment resulted in a rapid change from predominately N_2 production at pH 6.7 to predominately N_2O at pH 5.2. Due to the ratio change noted with change in soil pH value, the enzyme N_2O reductase may be extremely sensitive pH.

Conclusion

Several microbiological mechanisms occurring in soils can give rise to gaseous loss of N. Denitrification and nitrifier denitrification are biological pathways that are very closely linked by the production of NO_3^- in soils and sediments. The requirements for denitrification activity include a reduced C source, a NO_3^- pool and an O_2 limited environment. Under anoxic conditions, many facultative heterotrophic microbes carry out denitrification and reduce NO_3^- to NO_2^- , which is then reduced to N_2O and, in microaerobic or anaerobic conditions, to N_2 (coupled nitrification–denitrification). The $N_2O:N_2$ ratio produced by the denitrification pathway can be increased by increased NO_2^- and NO_3^- concentration. The ratio can also be increased by decreased O_2 concentration, pH values and increasing the C availability. The spatial variation of soil C and N and low O_2 tensions required for denitrification during field quantification and the importance of nitrifier denitrification processes are crucial areas of research that need to be understood before management of gaseous N loss can be achieved.

References

- Bannert A, Bogen C, Esperschütz J, Koubová A, Buegger F, Fischer D, Radl V, Fuß R, Chroňáková A, Elhottová D, Šimek M, and Schlöter M (2012) Anaerobic oxidation of methane in grassland soils used for cattle husbandry. *Biogeosciences* 9: 3891–3899.
- Clough T, Sherlock R, and Rolston D (2005) A review of the movement and fate of N_2O in the subsoil. *Nutrient Cycling in Agroecosystems* 72: 3–11. <https://doi.org/10.1007/s10705-004-7349-z>.
- Guerrero-Cruz S, Stultiens K, Kessel MAHV, Versantvoort W, Jetten MSM, Camp HJMOD, and Kartal B (2019) Key physiology of a nitrite-dependent methane-oxidizing enrichment culture. *Applied and Environmental Microbiology* 85: e00124–19.
- Han P, Wu D, Sun D, Zhao M, Wang M, Wen T, Zhang J, Hou L, Liu M, Klümper U, Zheng Y, Dong HP, Liang X, and Yin G (2021) N_2O and NO production by the comammox bacterium *Nitrospira inopinata* in comparison with canonical ammonia oxidizers. *Water Research* 190: 116728.
- Pandey CB, Kumar U, Kaviraj M, Minick KJ, Mishra AK, and Singh JS (2020) DNRA: A short-circuit in biological N-cycling to conserve nitrogen in terrestrial ecosystems. *Science of the Total Environment* 738. <https://doi.org/10.1016/j.scitotenv.2020.139710>. 139710.
- Raghoebarsing AA, Pol A, van de Pas-Schoonen KT, Smolders AJP, Ettwig KF, Rijpstra WIC, Schouten S, Damsté JSS, Op den Camp HJM, Jetten MSM, and Strous M (2006) A microbial consortium couples anaerobic methane oxidation to denitrification. *Nature* 440: 918–921.
- Su G, Lehmann MF, Tischer J, Weber Y, Lepori F, Walser J-C, Niemann H, and Zopfi J (2023) Water column dynamics control nitrite-dependent anaerobic methane oxidation by Candidatus "Methyloirabilis" in stratified lake basins. *The ISME Journal* 17: 693–702.

Symbiotic nitrogen fixation

Florence Mus^a and Hsin-Hua Wu^{a,b}, ^aInstitute of Biological Chemistry, Washington State University, Pullman, WA, United States;

^bDepartment of Chemistry, Washington State University, Pullman, WA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	271
Biological nitrogen fixation and nitrogenase enzymes	271
Diversity of nitrogen-fixing plant-microbe associations	272
Associative nitrogen-fixing interaction	273
Endophytic nitrogen-fixing associations	273
Endosymbiotic nitrogen-fixing associations	273
Regulation of nitrogenase activity in diazotrophs	274
Transcriptional regulation	274
Post-transcriptional regulation	276
Nutrient exchange and symbiotic biological nitrogen fixation	276
Carbon metabolism	276
Nitrogen metabolism	277
Environmental factors influencing symbiotic biological nitrogen fixation	277
Soil pH, moisture, temperature and light	277
Soil nutrient availability	278
Oxygen status	278
Mycorrhizal interaction	279
The importance of nitrogen-fixing symbioses to soil fertility	279
Strategies for engineering symbiotic nitrogen fixation	279
Engineering the legume symbiosis into cereals	280
Engineering expression of nitrogenase in organelles of plants	280
Engineering nitrogen fixation in plant colonizing bacteria	281
Conclusions	281
References	281

Abstract

Nitrogen is an essential nutrient in plant growth. The ability of a plant to supply its requirements from biological nitrogen fixation through interactions with associative, endophytic and endosymbiotic symbionts, is of tremendous importance to the environment and world agriculture. Symbiotic nitrogen fixation is part of a mutualistic relationship, in which plants provide a niche and fixed carbon to bacteria in exchange for fixed nitrogen, and relies on complex regulatory circuits between both partners. This process is restricted mainly to legumes in agricultural systems, and there is considerable interest in exploring whether similar symbioses can be developed in non-legumes.

Key points

- Availability of nitrogen (N) is one of the principal elements limiting growth and development of crops.
- Biological nitrogen fixation is the ability of diazotrophic bacteria to reduce atmospheric N₂ to NH₃, which is readily assimilated into biological molecules.
- Most legumes and a few non-legumes form intimate, nitrogen-fixing symbioses with diazotrophs that provide the plants with ammonia.
- Many important food species cannot establish effective nitrogen-fixing symbioses with diazotrophs.
- Large-scale use of industrially-produced N-fertilizer has doubled the influx of N into the terrestrial biogeochemical N-cycle; therefore, the long-term sustainability of massive N-fertilizer inputs in agriculture has come into question.
- The threat to sustainability has created a strong driver to provide an alternative to synthetic nitrogen by engineering biological nitrogen fixation in plants.

Introduction

In agroecosystems, nitrogen is one of the major nutrients limiting plant growth. To meet the increased nitrogen demand in agriculture, synthetic fertilizers have been used extensively which have led to environmental challenges such as nitrate pollution. Biological nitrogen fixation is the conversion of atmospheric nitrogen (N_2) to ammonia (NH_3), a form that can be used by plants. However, the process is restricted to bacteria and archaea and does not occur in eukaryotes. The bacteria responsible for nitrogen fixation are called diazotrophs; they encode nitrogenase, the oxygen-sensitive enzyme complex that catalyzes the conversion of N_2 to NH_3 . Diazotrophs are found in several phyla and representatives from most (if not all) of these phyla are known to engage in nitrogen-fixing symbiosis with plants. Symbiotic nitrogen fixation is part of a mutualistic relationship in which plants provide a niche and fixed carbon to bacteria in exchange for fixed nitrogen. Depending on environmental conditions, symbiotic nitrogen fixation may be tuned to the plant's nitrogen demand or specifically inhibited, or even prevented. Therefore, diazotrophs have evolved elegant strategies to tightly regulate the expression and activity of components required for biological nitrogen fixation. Biological nitrogen fixation by legumes, and by associative, endophytic, and endosymbiotic nitrogen fixation in non-legumes play major roles in reducing the use of synthetic nitrogen fertilizer in agriculture, increased plant nutrient content, and improving soil fertility. However, symbiotic nitrogen fixation, largely restricted to legumes, is not directly available to the most agriculturally important crops. It has thus been a focus to develop different strategies to transfer the ability to fix nitrogen to cereals.

Biological nitrogen fixation and nitrogenase enzymes

Biological nitrogen fixation refers to the enzyme-catalyzed reduction of N_2 to NH_3 , and occurs exclusively in microorganisms called diazotrophs. Diazotrophic activity accounts for approximately two-thirds of the fixed N available to the biosphere (Rubio and Ludden, 2008) and diazotrophs are distributed across multiple phyla of Bacteria and Archaea (Boyd et al., 2011). Diazotrophs represent ~7.8% of all bacterial and archaeal cultivars with sequenced genomes and are characterized by organisms operating aerobic, anaerobic, phototrophic, chemotrophic, autotrophic, and heterotrophic metabolisms. Known bacterial diazotrophs were found in 13 taxonomic phyla: Actinobacteria, Aquificae, Chlorobi, Chloroflexi, Chrysiogenetes, Cyanobacteria, Deferribacteres, Firmicutes, Fusobacteria, Nitrospirae, Proteobacteria, Spirochaetes, and Verrucomicrobia. Unlike bacterial species, nitrogen fixation in Archaea is contained only within the phylum Euryarchaeota. Diazotrophs are found in a wide variety of states and habitats including free-living in soils and water, associative symbioses with grasses, symbiotic actinorhizal associations with woody plants, cyanobacterial symbioses with various plants and root-nodule symbioses with legumes (Dos Santos et al., 2012).

Biological nitrogen fixation is catalyzed by the enzyme complex nitrogenase. The nitrogenase enzyme occurs in three different isoforms: Mo-nitrogenase, V-nitrogenase, and Fe-only nitrogenase. The Mo-nitrogenase (Nif) is found in all diazotrophs and is the only nitrogenase known in diazotrophs forming symbioses with higher plants. The two alternative nitrogenases, the V-nitrogenase (Vnf) and the Fe-only nitrogenase (Anf), are found in free-living soil bacteria, archaea, and cyanobacteria. Biological nitrogen fixation is usually considered to rely primarily on the more efficient canonical Mo-nitrogenase, whereas the V-nitrogenase and the Fe-only nitrogenase are often considered "backup enzymes", used when Mo is limiting. Indeed, nitrogen-fixing symbionts with higher plants such as *Rhizobium* and *Frankia* do not possess alternative nitrogenase genes and are believed to be the major contributors to biological nitrogen fixation. If the environmental distribution and diversity of alternative nitrogenases remains largely unknown, numerous lines of evidence suggest that alternative nitrogenases are significant contributors to biological nitrogen fixation in terrestrial ecosystems and can contribute up to 70% of the total biological nitrogen fixation in some diazotrophs (e.g., cyanolichens).

Despite differences in their metal content, the three nitrogenase enzymes are structurally, mechanistically, and phylogenetically related. The most intensively characterized system is the canonical Mo-nitrogenase, a complex two component metalloenzyme comprised of two proteins. The smaller dimeric component, known as the iron protein or dinitrogenase reductase (NifH protein), functions as an ATP-dependent electron donor to the larger heterotetrameric component, known as the molybdenum-iron protein or dinitrogenase (comprising the NifD and NifK component proteins), which contains the enzyme catalytic site (Fig. 1) (Seefeldt et al., 2009). Both nitrogenase-component proteins are extremely oxygen sensitive. The two other nitrogenases, V- and Fe-only nitrogenases, are enzyme homologs, except for an additional subunit (VnfG or AnfG) in the dinitrogenase component, and the absence of the heteroatom Mo. The three nitrogenases differ greatly in their catalytic activities: the Mo-nitrogenase requires 16 mol of ATP per mole of N_2 (Fig. 1), versus 24 and 40 mol of ATP per mole of N_2 for the V- and Fe-only nitrogenases, respectively, as assessed with in vitro assays (Mus et al., 2018).

The Mo-, V-, and Fe-only nitrogenases are genetically distinct as they are encoded by separate gene clusters designated, *nif*, *vnf*, and *anf*, respectively. The structural components of the Mo-nitrogenase are encoded by *nifH*, *nifD*, and *nifK* genes. The V-nitrogenase is encoded by *vnfH*, *vnfD*, *vnfG*, and *vnfK*, and the Fe-only nitrogenase components are products of *anfH*, *anfD*, *anfG*, and *anfK* genes. Apart from the catalytic components, additional gene products are required to produce fully functional nitrogenase enzymes. Although the number of proteins involved in the activation of nitrogenases seems to be species-specific and varies according to the physiology of the organism and environmental niche. Extensive biochemical and genetic characterization has implicated at least 82 genes products in the formation and regulation of the three forms of nitrogenases. Global transcriptional profiling studies of nitrogen-fixation reported in different diazotrophic bacteria were valuable for defining the full extent of the *nif*, *vnf*, and *anf* gene clusters and for identifying genes that are coordinately expressed with particular nitrogenase but not directly involved in nitrogen

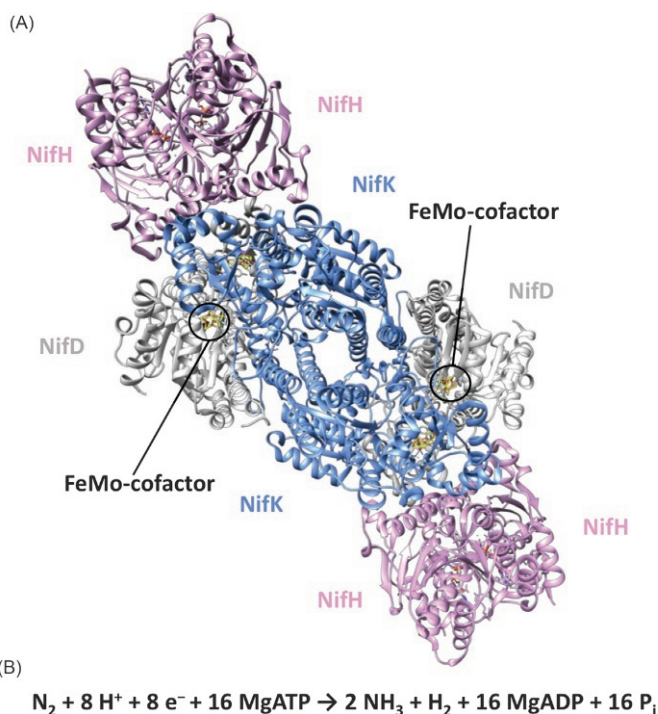


Fig. 1 Structure and stoichiometry of the Mo-nitrogenase enzyme. (a) Structure of the *Azotobacter vinelandii* Mo-nitrogenase enzyme complex (NifHDK; PDB 1M34; Schmid et al., 2002) showing the MoFe (NifD, in grey; NifK in blue) and Fe protein (NifH, in pink) components. The active-site FeMo-cofactor of Mo-nitrogenase is circled. (b) Stoichiometry for N_2 reduction catalyzed by the Mo-nitrogenase enzyme.

fixation (non-specific nitrogen-fixation genes involved in different functions such as transport and secretion, iron-sulfur cluster assembly system, respiratory and energy metabolism, electron transport, nitrogen and carbon metabolism, secondary metabolism, translation and transcription) (Mus et al., 2018).

Diversity of nitrogen-fixing plant-microbe associations

Various types of associations/interactions occur between diazotrophs and their host plants including associative interactions, endophytic associations, and endosymbiotic associations (legume and non-legume symbioses) (Fig. 2). In all these associations and symbioses, for the host plants the expected benefit of the interaction is the fixed nitrogen provided by the symbiotic partner,

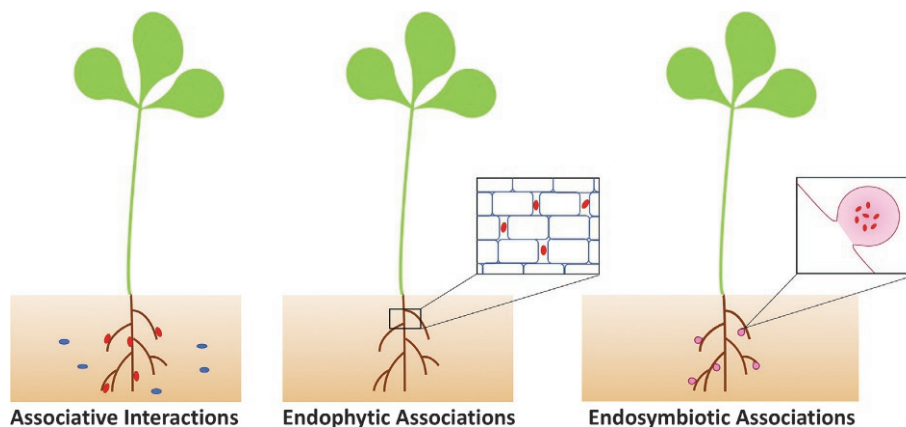


Fig. 2 Nitrogen-fixing plant-microbe interactions or associations exhibit different degrees of intimacy: associative interactions, endophytic associations, and endosymbiotic associations. Free-living and associative diazotrophs (e.g., *Azospirillum*, *Azotobacter*, *Nostoc*) may be enriched within the rhizosphere and benefit directly from plant root exudates. The plant may benefit from released nitrogen, but has little control over the exchange. Endophytic diazotrophs spread and multiply within plant tissues without causing damage and eliciting significant defense reactions as they do not reside intracellularly in living-plant cells. Endosymbiotic diazotrophs (e.g., *Azoarcus*, *Herbaspirillum*, *Gluconacetobacter*, *Nostoc*) colonize different plant organs intracellularly. Endosymbiotic diazotrophs (e.g., *Rhizobia*, *Frankia*) accommodated inside plant cells within plant-derived membranes known as nodules. The plant has a high degree of control and symbiosis is very specific.

which, in return, receives reduced carbon and possibly all the other nutrients it requires. In addition, the symbiotic or endophytic plant structure colonized by the nitrogen-fixing microorganisms may provide the appropriate conditions for protecting the nitrogenase complex from oxygen exposure.

Associative nitrogen-fixing interaction

Bacteria that colonize the rhizosphere are called rhizobacteria, and rhizobacteria with beneficial effects on plant development are referred to as plant-growth-promoting rhizobacteria (PGPR). Some of these PGPR are diazotrophic bacteria and have the ability to develop root associations with different plants including grasses. Interactions between plants and associative nitrogen-fixing bacteria are the simplest form of nitrogen-fixing symbiosis (Fig. 2) (Mus et al., 2016). Diazotrophic rhizobacteria have been identified in several genera of alpha- and β -Proteobacteria including *Acetobacter*, *Azoarcus*, *Azospirillum*, *Azotobacter*, *Burkholderia*, *Enterobacter*, *Herbaspirillum*, *Gluconobacter* and *Pseudomonas*. These associative diazotrophic nitrogen-fixing bacteria respond to root exudates via chemotaxis to, and colonization of, the rhizosphere of many plants but do not invade plant tissues. Many of these bacteria can live in association with maize, rice and wheat, and to contribute to improved plant growth (Santi et al., 2013).

Endophytic nitrogen-fixing associations

Many species of diazotrophic bacteria have evolved beyond surface colonization to spread and multiply within plant tissues without causing damage and eliciting significant defense reactions, they are usually designated endophytic nitrogen-fixing bacteria (Fig. 2). Endophytic bacteria invade plant tissues, but they differ from endosymbionts, as they do not reside intracellularly in living plant cells and their colonization does not induce the formation of any differentiated plant structure. Endophytes are ubiquitous, likely form colonies in all plants, and have been isolated from almost all plants examined to date. Their association can be obligate or facultative and causes no harm to the host plants. They exhibit complex interactions with their hosts which involves mutualism and antagonism. *Azoarcus* spp., *Herbaspirillum seropedicae* and *Gluconobacter* are recognized as endophytes. They differ from other rhizobacteria such as *Azospirillum* and *Azotobacter*, in that 'they are tightly associated with plants and do not survive well in soil' (Santi et al., 2013). Cyanobacteria are also frequently found within plant tissues. *Nostoc* is endophytic with two genera of liverworts (*Blasia* and *Cavicularia*) and all hornworts. Colonization can take place in dome-shaped auricles on the thallus of liverworts or in slime cavities of the thallus or mucilage-filled canals that run parallel to the thallus of hornworts. *Nostoc* is also able to endophytically colonize coralloid roots of cycads. The mechanism of recruitment is unknown, but the cyanobacteria are found embedded in mucilage in a specific cortical layer of the coralloid root between elongated specialized cells.

Endosymbiotic nitrogen-fixing associations

The highly specific and most efficient processes for nitrogen fixation involve the formation of root nodules on legumes and non-legumes (Fig. 2). The diazotrophic bacteria involved in these endosymbiotic interactions include *Rhizobia* (Gram negative) members of the alpha-subgroup of the phylum Proteobacteria that associate with legumes (family Fabaceae) and the non-legume *Parasponia* species (family Cannabaceae); *Frankia* sp. (Gram positive) members of the actinomycete family that associate with a broad spectrum of plants belonging to eight families collectively called actinorhizal plants; and nitrogen-fixing cyanobacteria (mainly *Nostoc* sp.) that have also been found to colonize different plant organs intracellularly in the family Gunneraceae (Santi et al., 2013).

Rhizobium-legume symbiosis

The nitrogen-fixing symbiosis is predominant in leguminous plants belonging to the family Fabaceae. Legume-*Rhizobium* symbiosis starts with a molecular dialogue between the two partners. The legume secretes a cocktail of phenolic molecules (flavonoids and isoflavonoids) into the rhizosphere. These signals are taken up by rhizobia, bind the transcriptional regulator NodD, and activate a suite of bacterial nodulation genes. These nodulation genes are responsible for the production of lipochitooligosaccharides (LCOs) called Nod factors which are indispensable in the specific host-*Rhizobium* interaction and at later stages in the infection process and nodule organogenesis. In the initial stages of infection, rhizobia attach to the root hair surface in a process that may involve a number of different mechanisms, mediated by means of adhesion proteins on the bacterial cell surface that recognize specific determinants of the plant cell wall. Invasion of the plant starts with an infection-thread, through which bacteria enter the root hair and move, at a high rate of proliferation, towards the root cortex, where they differentiate into nitrogen-fixing bacteroids within a unique plant organelle called the symbiosome. The symbiosome is delimited by a plant-derived membrane that controls nutrient exchange between the symbionts. Two main types of nodules are formed on the various legume species, indeterminate (apical meristem and developmental gradient) or determinate (no apical meristem and no developmental gradient), depending on whether or not the meristem remains active for the life of the nodule, respectively. The symbiosomes of determinate nodules usually contain a small number of bacteroids that are of a similar shape and size as free-living bacteria, while symbiosomes from indeterminate nodules usually contain a single swollen and pleomorphic bacteroid. Both of types of legume nodules have a peripheral vasculature, in contrast to roots (Rashid et al., 2015).

Actinorhizal symbiosis

Actinorhizal plants have the ability to develop an endosymbiosis with the nitrogen-fixing soil actinomycete *Frankia*. *Frankia* is a genus of soil actinomycetes in the family Frankiaceae that fix nitrogen, both under symbiotic and free-living aerobic conditions, while most rhizobia do not. Over 200 strains of *Frankia* have been isolated from many actinorhizal plant species in the plant

families *Betulaceae*, *Casuarinaceae*, *Coriariaceae*, *Datisceae*, *Eleagnaceae*, *Myricaceae*, *Rhamnaceae*, and *Rosaceae*. In both the free-living and the symbiotic state, nitrogen fixation takes place within the swollen tips of hyphae termed vesicles that are encased in multilayered hopanoid-rich membrane envelopes that provide the necessary O₂ protection to prevent nitrogenase inactivation. The strategies used by *Frankia* spp. to infect actinorhizal plants are quite similar to those used by rhizobia. The establishment of the symbiotic process results in the formation of root nodules in which *Frankia* provides fixed nitrogen to the host plant in exchange for reduced carbon. Depending on both the host species and *Frankia* clade, infection can either take place via infection threads or by crack entry. Unlike legume root nodules, the root nodules of these non-legumes are modified lateral roots with a central vascular system and a peripheral infected expanded cortex (Santi et al., 2013; Rashid et al., 2015).

Parasponia-rhizobium symbiosis

In the order Rosales, in addition to actinorhizal nitrogen-fixing plants, *Parasponia* species (family Cannabaceae) also display an original nitrogen-fixing root symbiosis. *Parasponia* is the only non-legume host plant known to be nodulated by rhizobia. Like in most legume-*Rhizobium* symbioses, the *Parasponia*-rhizobium symbiosis depends on Nod factors. The infection process that leads to nodule development occurs via the so-called crack entry mode of infection. The rhizobia remain in threads throughout the symbiotic process and are not released from the threads, unlike the process of bacteroid formation in legume-*Rhizobium* symbioses. The final structure of the *Parasponia* nodule lobe is similar to that observed in actinorhizal symbiosis and resembles lateral roots (Santi et al., 2013; Rashid et al., 2015).

Cyanobacterial endophytic and endosymbiotic associations

The intimacy of the symbiosis *Gunnera-Nostoc* is therefore greater than in other cyanobacterial associations since the cyanobacteria are hosted intracellularly, whereas in bryophytes, *Azolla* and cycads, the cyanobionts occur as endophytes in mucus-filled cavities. The cyanobiont *Nostoc* enters *Gunnera* plants through specialized glands located on the stem. The glands secrete polysaccharide-rich mucilage that attracts specific symbiotic cyanobacteria, supports their growth on the gland surface, and contributes to their differentiation. After entering the gland through existing channels, cyanobacteria induce divisions in the host cells lining the channel and are subsequently taken up into the host cells. The conversion of vegetative filaments into motile and short-lived hormogonia is an essential step for the establishment of the symbiotic process. Different environmental stimuli and plant factors released during nitrogen starvation can stimulate the induction of gliding filaments known as hormogonia. Hormogonia play a major role in short-distance dispersal of filamentous cyanobacteria and in many strains display positive phototaxis, which is important in the colonization of illuminated portions of the habitat by these photoautotrophic organisms. After entering the host, hormogonia revert to non-motile vegetative filaments, the differentiation of nitrogen-fixing heterocysts is observed, and vegetative cells show altered morphology compared with their free-living counterparts. The development of heterocysts after plant infection is essential for a functional nitrogen-fixing symbiosis (Santi et al., 2013; Rashid et al., 2015).

Regulation of nitrogenase activity in diazotrophs

The reaction catalyzed by the nitrogenase enzyme is an extremely energy-demanding process since reduction of 1 mol of N₂ requires 16 mol of ATP in vitro. Moreover, the relatively slow turnover time of the enzyme requires diazotrophs to synthesize large quantities of nitrogenase to use dinitrogen as a sole nitrogen source. In addition, the oxygen sensitivity of nitrogenase imposes considerable physiological constraints on diazotrophy as there is an obligation to protect the enzyme from oxygen damage. Therefore, the synthesis of both molybdenum and the molybdenum-independent nitrogenases is stringently regulated at the transcriptional level in response to the availability of fixed nitrogen and to environmental oxygen levels. The degree to which each stimulus affects transcription is characteristic of the particular diazotroph. In addition to regulation at the transcriptional level, nitrogenase activity is also subject to post-translational regulation in some diazotrophs (*Rhodospirillum rubrum*, *Rhodobacter capsulatus*, and *Azospirillum brasilense*). The necessity to respond to the concentrations of fixed nitrogen and external oxygen, and to provide sufficient energy for nitrogen fixation, imposes common regulatory principles among diazotrophs. However, there is considerable plasticity in the regulatory networks, which differ among microorganisms and are dependent on host physiology.

Transcriptional regulation

Regulation of nitrogen fixation is governed by several genes in response to the environmental signals of molecular oxygen and fixed nitrogen. The σ^{54} -dependent transcriptional activator NifA is the central key to this regulation (Fig. 3).

Nitrogen regulation

Diazotrophs can also access nitrogen through uptake of externally available nitrogen sources, which can include both low- and high-molecular-weight organic N sources. Given the energy costs of symbiotic nitrogen fixation, it is generally downregulated by increasing nitrogen availability, as diazotrophs use external nitrogen in favor of fixed nitrogen. Ammonium, the direct product of nitrogen fixation, is well known to inhibit nitrogen fixation and has been shown to inhibit nitrogenase synthesis at the transcriptional level through the regulation of NifA expression or modulation of its activity. NifA (a member of the bacterial enhancer-binding family) is the master regulator of nitrogen fixation in diazotrophic Proteobacteria which acts as a transcriptional activator responsible for control of all major *nif* gene cluster transcription to ensure that nitrogenase expression is stringently regulated by the demand for fixed nitrogen. In many diazotrophic Proteobacteria, expression of NifA is controlled by the

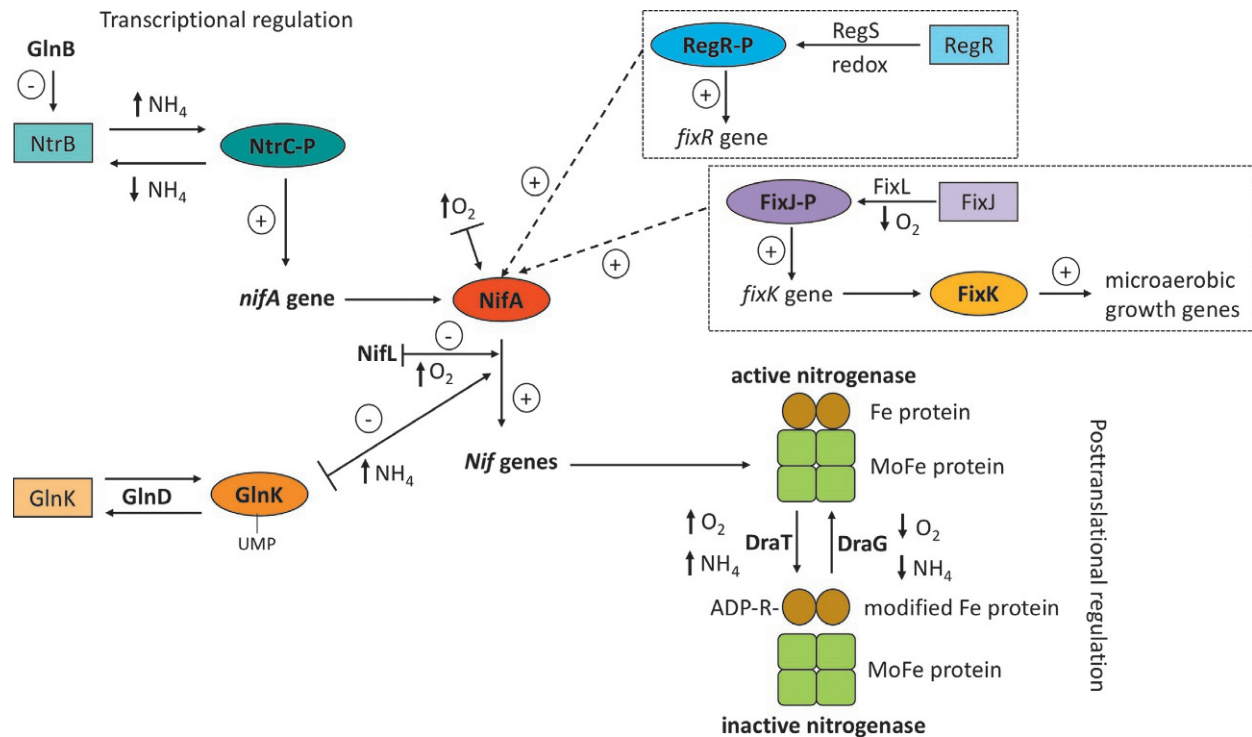


Fig. 3 Regulation of nitrogen fixation. Transcriptional control operates by a regulatory cascade with three levels: activation of NtrC by NtrB in response to low nitrogen, induction of *nifA* gene expression by phosphorylated NtrC, and induction of *nif* genes by NifA in *Klebsiella pneumoniae*. Transcription of *nifL-nifA* operon is not subject to nitrogen regulation in *Azotobacter vinelandii*. In both organisms NifA is active under conditions of nitrogen limitation. In *A. vinelandii* this is promoted by binding of 2-oxoglutarate to NifA, which prevents inhibition by NifL, and also by uridylylation of GlnK controlled by GlnD, preventing its interaction with NifL. By contrast, in *K. pneumoniae* GlnK is required to prevent interaction of NifL and NifA under nitrogen-limiting conditions. Oxygen regulates *nif* gene expression by affecting NifA activator either directly (*K. pneumoniae*) or through the NifL protein (*A. vinelandii*), and in some bacteria additional bicomponent systems like FixL/FixJ (*Sinorhizobium meliloti*, *Bradyrhizobium japonicum*) and RegS/RegR (*B. japonicum*) also control expression of *nifA* or microaerobic growth genes. In *Rhodobacter* and *Rhodospirillum*, nitrogenase activity is regulated by a switch-on/switch-off mechanism (posttranslational control) in response to oxygen and ammonium, which involves the DraT/DraG two component system.

two-component NtrB-NtrC nitrogen regulatory system in response to covalent modification and the ligand-binding state of the PII protein (GlnB) which provides global control in response to the nitrogen source (Fig. 3) (Dixon and Khan, 2004). This first level of the nitrogen regulatory cascade is absent in the aerobic nitrogen fixing soil bacteria *Azotobacter vinelandii* since NtrB-NtrC system is not required for activation of *nifLA* transcription.

In addition, the activity of NifA itself is stringently regulated either directly or indirectly by signal transduction protein (GlnK) (Fig. 3). In many cases, the nitrogen signal transduction mechanism involves direct interaction with PII protein. However, in some Proteobacteria, NifA activity is controlled by a partner protein NifL, which senses environmental cues and interacts with PII protein, to modulate NifA activity in response to the nitrogen status, providing a second level of the regulatory hierarchy. For example, in *A. vinelandii*, the non-modified form of GlnK is required for the formation of the inhibitory complex between NifL and NifA under nitrogen excess conditions, as opposed to *Klebsiella oxytoca*, where GlnK is required to relieve inhibition of NifA activity by NifL under nitrogen-limiting conditions (Dixon and Khan, 2004).

Oxygen regulation

Four proteins that regulate nitrogen fixation are oxygen/redox sensors: the histidine kinase FixL, the histidine kinase RegB, the anti-activator NifL and the activator NifA (which is only oxygen-sensitive in those diazotrophs that lack NifL) (Fig. 3) (Bueno Batista and Dixon, 2019).

In some rhizobia, the two-component FixL-FixJ system (Fig. 3) is involved in the low oxygen-dependent transcriptional control of nitrogen fixation genes. FixL protein contain N-terminal PAS domain which is commonly found in prokaryotic and eukaryotic sensors of oxygen. Low oxygen concentrations activate the oxygen sensor protein FixL, which in turn phosphorylates and thereby activates the transcriptional activator FixJ. The activated FixJ protein induces transcription of *nifA* and *fixK*, and the protein products of these genes induce the transcription of different genes encoding proteins involved in the process of nitrogen fixation.

In the free-living diazotrophs *Klebsiella pneumoniae*, *A. vinelandii*, and *Enterobacter cloacae*, which belong to the γ -Proteobacteria, NifA activity is not directly sensitive to oxygen. In these organisms NifA activity is regulated in response to molecular oxygen and combined nitrogen through a second regulator NifL; the *nifL* and *nifA* genes together form an operon (Fig. 3). In *Klebsiella*

pneumoniae, the expression of the *nifLA* operon itself is regulated by the nitrogen status of the cell, via the NtrB/NtrC two component regulatory system; in *A. vinelandii*, however, *nifLA* is expressed constitutively. The domain structures of NifL and FixL are analogous. The sensing method is different in each case, however, as the NifL PAS domain contains flavin-adenine dinucleotide whereas FixL contains a haem prosthetic group. The FAD-containing PAS domain of NifL is required for redox sensing and regulates the ability of NifL to inhibit NifA activity in response to the oxygen status. Inhibition of NifA activity by NifL occurs possibly via direct protein-protein interaction. In diazotrophic Proteobacteria that lack NifL, NifA proteins have a slightly different domain structure to those that are subject to NifL inhibition, containing an invariant Cys-X₄-Cys motif. The presence of these cysteine residues seems to correlate with the oxygen sensitivity of these proteins. Indeed, all of the conserved cysteines are indispensable for NifA activity, and metal ions are required for *in vivo* activity in *Bradyrhizobium japonicum*. This might suggest a model in which metal ion coordination to the cysteine residues controls the activity of these proteins in response to the redox status.

The global redox-responsive RegB-RegA two-component system (Fig. 3), directly regulates *nif* transcription in some diazotrophic members of the α -Proteobacteria. RegB is a histidine sensor kinase and contains a conserved cysteine residue. This conserved cysteine is essential for redox sensing and on oxidation. Phosphorylation of the response regulator RegA increases its DNA-binding activity, transcriptional modulation and interactions with other regulators. In *R. capsulatus*, RegA function in conjunctions with NtrC to indirectly activate nitrogenase synthesis by binding to and activating the expression *nifA2* promoter. In *B. japonicum*, homologs of RegB and RegA (known as RegS and RegR, respectively) have been shown to control *nifA* expression during the symbiosis.

Post-transcriptional regulation

In addition to regulation at the transcriptional level, nitrogenase activity is also subject to posttranslational regulation in a few free-living diazotrophs (*R. rubrum*, *R. capsulatus*, and *A. brasilense*). At least three different mechanisms, involving PII proteins (GlnB or GlnK), are known to switch-off nitrogenase activity under excess nitrogen conditions in Proteobacteria (Bueno Batista and Dixon, 2019). GlnB and GlnK are signal transduction proteins known to integrate various metabolic signals including the levels of glutamine, ADP, ATP, and 2-oxoglutarate to control both nitrogen fixation and nitrogen assimilation. The first mechanism involves rapid inactivation of nitrogenase Fe protein by ADP ribosylation, catalyzed by DraT (ADP-ribosyltransferase). This post-translational covalent modification can be reversed by another enzyme with opposing activity named DraG (ADP-ribosyl glycohydrolase) (Fig. 3). The activity of both enzymes is regulated by interaction with GlnB and GlnK proteins depending on the nitrogen status. Under excess nitrogen conditions, non-modified GlnB interacts with DraT-stimulating ADP ribosylation. In addition, GlnK mediates formation of a ternary complex between AmtB, GlnZ and DraG, that sequesters DraG to the membrane, inhibiting its ability to remove the covalent modification from nitrogenase. Under conditions of nitrogen limitation, the uridylylated forms of GlnB cannot interact with either DraT or DraG. Under these circumstances, DraT is inactive and DraG is active, leading to the removal of the ribosyl modification from the nitrogenase Fe protein, thus reactivating nitrogenase. A second mechanism for nitrogenase switch-off apparently involves inhibition of the Rnf1 electron transport system (encoded by *rnfABCDE*), which potentially provides a dedicated electron transfer pathway to nitrogenase in *A. vinelandii*. Genes encoding the Rnf1 complex are co-located with the *nifLA* operon in diverse diazotrophs. In *A. olearius* BH72, the interaction GlnK and the Rnf1 complex has been proposed to control electron transfer to nitrogenase. A third unknown mechanism involving GlnK and the ammonium transporter AmtB may be responsible for nitrogenase switch-off during ammonium upshift in organisms lacking both DraT-DraG and the Rnf1 electron transport system.

Nutrient exchange and symbiotic biological nitrogen fixation

The exchange of nutrients between a plant and a nitrogen-fixing microorganism is critical for symbiotic nitrogen fixation. In return for fixed nitrogen, the plant typically provides its bacterial symbiont with a carbon source and other crucial nutrients, depending on the intimacy of the symbiosis. Both organisms change their metabolic routines to accommodate to each other's needs, a process that is monitored and regulated by both partners (Fig. 4) (Mus et al., 2016).

Carbon metabolism

Nitrogen fixation is an energetically and, therefore, carbon expensive process and diazotrophs can only fix nitrogen when adequate supplies of carbon are available. Plant root exudates, carbon-rich secretions consisting of low-molecular-weight compounds such as sugars, organic acids, and mucilage are potential source of carbon capable of meeting energy demands of associative diazotrophs. During legume-*Rhizobium* symbiosis, the carbon supply, required to fuel nitrogenase activity in the bacteroid, is derived from plant photosynthate that is transported to the nodules via the phloem as sucrose. In nodule cells, sucrose metabolism releases hexoses that subsequently enter glycolysis to produce phosphoenolpyruvate, which in turn is converted to dicarboxylic acids. Dicarboxylic acids such as malate are assimilated by gluconeogenesis or catabolized via enzymes of the tricarboxylic acid (TCA) cycle to provide the reductant and ATP required for nitrogen fixation (Alemayehu et al., 2018). Several other compounds such as phosphorus, iron, and sulfur are made available to symbiotic microbes, especially in the case of endosymbionts, which rely on the host for all their essential nutrients. The exchange of metabolites between the plant and bacteroids does not happen freely but is facilitated by specialized transporters. Among these are genes encoding putative sugar transporters, amino acid transporters, and sulfate transporters.

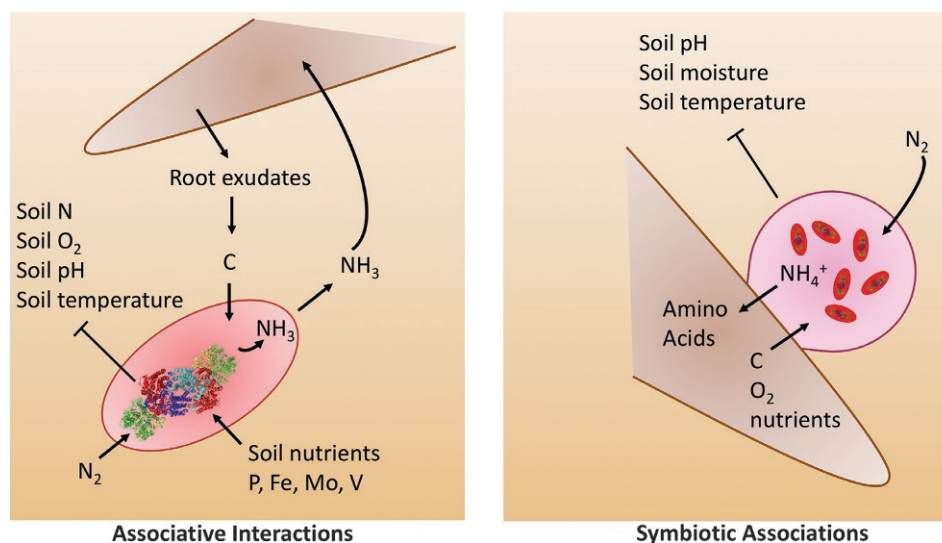


Fig. 4 Associative nitrogen-fixing interactions carried out by a diverse array of free-living nitrogen fixers and symbiotic nitrogen-fixing associations performed by endophytic or endosymbiotic nitrogen fixers (e.g., *Rhizobia*). Associative nitrogen-fixing interactions is supported by root exudates compounds while symbiotic nitrogen fixing associations receive a constant supply of simple C compounds directly from the host plant. Oxygen concentration in the rhizosphere is highly variable and driven by soil structure and texture and respiration by microbes and roots. Conversely, symbiotic nitrogen fixers are supplied with oxygen at low concentrations by their host plant. Nutrients necessary to support associative nitrogen-fixing interactions must be acquired by the diazotroph from the soil. However, these nutrients are delivered to symbiotic nitrogen fixers by the host plant. Associative nitrogen fixers can access nitrogen via nitrogen fixation or through uptake of externally available nitrogen sources. The availability of external nitrogen sources inhibits the nitrogenase enzyme. Ammonium, the direct product of nitrogen fixation, released by associative nitrogen fixers is available to plants. All symbiotically fixed nitrogen is delivered to the host plant.

Nitrogen metabolism

The exchange of fixed nitrogen is another nutrient important for the symbiosis to be mutually beneficial. Specifically, bacterial nitrogen metabolism must be altered so that nitrogen is excreted rather than incorporated into the microbial biomass. The plant host appears to directly interfere with bacterial amino acid biosynthesis and thereby forces the release of nitrogen.

In nitrogen-fixing *Rhizobium* bacteroids, nitrogen metabolism is significantly altered during bacteroid differentiation, and ammonia assimilation is effectively shut via an unknown and presumably plant-regulated post-translational modification of the enzyme glutamine synthetase (GS). Ammonia produced by nitrogenase is delivered to the plant cell as NH_4^+ or NH_3 via passive diffusion or active transport involving transporter AmtB. Once inside the plant cell, ammonia is assimilated into amino acids mainly by the action of GS, glutamate synthase (GOGAT), and aspartate aminotransferase. The expression of genes encoding these enzymes is induced during nodule development (White et al., 2007).

In actinorhizal plant-*Frankia* symbiosis, the bacterial GS remains fully functional, but downstream components of amino acid biosynthesis are downregulated. Fixed nitrogen is released to the plant in the form of amino acids or amides. These are then broken back down to NH_4^+ , which is then assimilated by the actinorhizal host by the action of GOGAT (White et al., 2007).

In symbiosis with both *Azolla* and *Gunnera*, the cyanobacterial GS is downregulated (unlike in legume-*Rhizobium* symbiosis), resulting in ammonium release. This ammonium is subsequently assimilated by the GS-GOGAT system of the plant host. Unlike other cyanobacterial symbioses, in *Nostoc*-cycad associations, the GS-GOGAT system of *Nostoc* is not downregulated, and nitrogen is transferred to the host in the form of citrulline, glutamine, or both, depending on the cycad host (White et al., 2007).

Environmental factors influencing symbiotic biological nitrogen fixation

Biological N fixation relies on a delicate partnership between plant and bacterial species and is maximized when niche requirements are met for both species. Factors that limit plant health and photosynthetic capacity will likewise limit nitrogen fixation potential. However, even when optimal conditions are met for plant growth and development, the establishment of symbiotic association can be independently inhibited by factors including nutrient availability, moisture, soil pH, temperature, and oxygen status (Fig. 4) (Alemayehu et al., 2018).

Soil pH, moisture, temperature and light

Soil pH is likely to be highly variable in the rhizosphere and may be an important environmental control on symbiotic biological nitrogen fixation. Indeed, rhizobia growth, survival and abundance, in addition to their competitiveness in nodulation, are highly influenced by soil pH. Increased H^+ concentration and increased solubility of the toxic metal ions Al^{3+} , Cu^{2+} and Mn^{2+} are the primary causes of intercellular pH instability leading to growth inhibition in low pH soils. In low pH soil, nodule formation has been reported to be reduced by >90% and nodule dry weight by >50% in species such as soybean, pea, cowpea, *Medicago* and

lucerne. In common bean, low soil pH is reported to reduce the number, ultrastructure and weight of nodules, in addition to the nitrogenase activity. Nitrogen fixation is typically reported to be reduced in acidic soil, with molybdenum deficiency in low pH soil further hinders nitrogen fixation, as it is a key component of the nitrogenase enzyme complex. Low pH conditions also disrupt the signal exchange between the host plant and microsymbiont by reducing plant flavonoid secretion, decreasing *Rhizobia* *Nod* gene induction and restricting Nod factors and Nod metabolite excretion. Low pH conditions have also been shown to affect *Rhizobia* attachment to root hairs and root colonization, leading to reduced nodule formation.

The activity of diazotrophs is also controlled by other abiotic factors such as soil moisture content and ambient temperatures, which can promote or inhibit nitrogen fixation. For instance, sufficient moisture content is required for nitrogen fixation activity and has been put forward as a main driver of nitrogen fixation. Studies, which have been conducted in the high Arctic on free-living- as well as on moss and lichen-associated diazotrophs, have shown a strong moisture dependence of nitrogen fixation. Whereas mosses, lichens and free-living nitrogen fixers easily lose water under warm and dry conditions, diazotrophs associated with legumes should be able to remain active for longer exposed to the same conditions. As for all enzymatic processes, temperature is a key factor, affecting nitrogenase activity positively, with reported temperature optimums between 25 °C and 30 °C. Soil temperature also disturbs the fixation processes, as it can exceed 40 °C at 5 cm depth, imposing limitations on the biological nitrogen fixation and nodule viability in the soil (Hungria and Vargas, 2000). The soybean crop thrives under favorable temperatures from 25 °C to 30 °C during its developmental stages in the soil. In legumes, the optimal temperature for induction of biological nitrogen fixation ranges from 27 °C to 40 °C, although temperatures exceeding 37 °C cause the death of the rhizobia and some genetic changes, rendering them less efficient, when compared with the selected strains.

Another important factor controlling nitrogen fixation in cyanobacteria is light. Nitrogen fixation is energy “costly,” and energy in the form of photosynthates is required to sustain nitrogenase activity in autotrophs. Heterotrophic diazotrophs on the other hand have to cover their energy demand through the breakdown of organic matter. As for all enzymatic processes, temperature is a key factor, affecting nitrogenase activity positively, with reported temperature optimums between 25 °C and 30 °C.

Soil nutrient availability

Soil nutrient status has a tremendous influence on the symbiosis as well as independent growth and survival of both partners.

Several field studies also demonstrate that nodulation and nitrogen fixation can be inhibited by high field nitrogen levels. High levels of nitrate in or near the nodules inhibits nitrogenase activity through a feedback mechanism, thereby reducing nitrogen fixation. Above a certain concentration, excess nitrogen can inhibit nodule initiation entirely. Estimates for nitrogen levels that will eliminate nodulation vary widely from lower value of 50 kg/ha to high value of 120 kg/ha. Tolerance of nitrogen fixation to high soil nitrate levels varies across species and even among genotypes of the same species (Zahran, 1999). Ammonium, the direct product of nitrogen fixation, is well known to inhibit nitrogen fixation (see sections above “Transcriptional regulation” and “Nitrogen regulation”) and has been shown to inhibit nitrogenase synthesis at the genetic level through the regulation of *nifA* gene transcription. Expression of nitrogenase in symbiotic diazotrophs is fairly insensitive to ammonium because export of ammonium to their symbiont suppresses ammonium levels. The expression of *nif* genes in free-living diazotrophs is more sensitive to cellular ammonium levels.

The availability of phosphorus and micronutrients, including Fe, Mo, and V, is known to influence nitrogen fixation (Zahran, 1999). In low phosphorus-content soil, phosphorus has been shown to control the rates of heterotrophic biological nitrogen fixation and to limit biological nitrogen fixation in rhizobial symbioses. Phosphorus appears essential for both nodulation and nitrogen fixation, as phosphorus deprivation is often associated with decreased nodule tissue formation and low rates of nitrogen fixation. Legume-*Rhizobia* symbiosis requires micronutrients including boron, cobalt, copper, iron, manganese, molybdenum, nickel, selenium, and zinc, sometimes at higher rates than the plant or free-living bacteria require alone. Iron, sulfur, molybdenum and vanadium are essential components of nitrogenases (Nif, Vnf, and Anf) (see section above, “Biological nitrogen fixation and nitrogenase enzymes”), and the availability of these trace metals may be critical for the nitrogen cycle of terrestrial ecosystems.

Oxygen status

Oxygen irreversibly inhibits nitrogenase and so can greatly influence the efficiency of symbiotic nitrogen fixation. Therefore, diazotrophs must employ protection mechanisms to maintain nitrogen fixation when oxygen is present. This includes avoidance of oxygen via growth strategy, spatial or temporal isolation of nitrogenase from oxygen, and production of biofilms as oxygen diffusion barriers. Diazotrophs can also remove oxygen by increasing substrate utilization, which increases respiration rates, thereby decreasing oxygen concentrations.

Microaerobic conditions favorable to nitrogenase are established in legume, *Parasponia*, and actinorhizal nodules by various mechanisms that include: O₂ diffusion resistance in outer cell layers of nodules; binding and transport of O₂ by leghemoglobins in infected cells of the nodule interior; restriction of oxygen diffusion into bacteria by external mucilage; rapid consumption of O₂ by bacteria and plant mitochondria. The crucial role of leghemoglobins was demonstrated for nitrogen fixation in legume root nodules (Ott et al., 2005). Leghemoglobins accumulate to millimolar concentrations in the cytoplasm of infected plant cells prior to nitrogen fixation and are thought to buffer free oxygen in the nanomolar range as low as 20–50 nM, avoiding inactivation of oxygen-labile nitrogenase while maintaining high oxygen flux for respiration. Rhizobia respond to the microaerobic conditions in the nodule through a complex signaling cascade involving multiple O₂ sensors (FixL/FixJ, hFixL/FxkR, FnrN), and rhizobial regulatory mechanisms insensitive to O₂ tension are essential for successful differentiation into bacteroids and the establishment of a productive symbiosis (Inomura et al., 2017).

Oxygen concentrations in the rhizosphere are dynamic and can range from near 0% (anaerobic) to 21% (ambient). However, because of the diversity of diazotrophs present and potentially active in the rhizosphere, it is difficult to pinpoint one optimal oxygen concentration for biological nitrogen fixation. For example, Inomura et al. (2017) demonstrated that oxygen concentrations around 3% resulted in significantly greater nitrogenase activity by the plant-associated aerobic diazotroph *A. vinelandii*. Various mechanisms for oxygen protection of nitrogenase in *A. vinelandii* have been proposed (Maier and Moshiri, 2000). These include a high respiratory rate involving a specialized cytochrome that keeps oxygen levels low inside the cell, a protein that binds and protects nitrogenase (but renders it temporarily inactive) under conditions of oxygen stress, and the production of an alginate capsule that presumably slows oxygen diffusion into the cell. Similar mechanisms have been described for *Gluconacetobacter*.

Cyanobacteria have evolved several mechanisms to protect nitrogenase from oxygen toxicity (Stal and Krumbien, 1985). These strategies involve spatial or temporal separation of photosynthesis and nitrogen fixation. Some filamentous cyanobacteria, such as *Nostoc* and *Anabaena*, develop specialized non-photosynthetic cells, called heterocysts, where nitrogen fixation occurs. Heterocysts lack the oxygenic photosystem II and are able to maintain microaerophilic conditions by their thick cell walls acting as an oxygen barrier and through active respiration. Other cyanobacteria have the ability to carry out photosynthesis and nitrogen fixation in the same cell by fixing nitrogen at a time when photosynthesis is depressed, typically at night.

Mycorrhizal interaction

In natural environments, legume roots form tripartite interactions, and are simultaneously colonized by both arbuscular mycorrhizal fungi and *Rhizobia*. Various studies have demonstrated that tripartite interactions improve plant productivity, seed yield, phosphorus and nitrogen acquisition, and photosynthetic rates (Bournaud et al., 2018). Arbuscular mycorrhizal fungi are known to have significant positive effect on biological nitrogen fixation by direct or indirect interaction with nitrogen-fixing microorganisms. Indeed, arbuscular mycorrhizal fungi can facilitate the colonization of legume roots by symbiotic nitrogen-fixing bacteria and have been shown to increase the biological nitrogen fixation by root nodules by at least 50%, providing their hosts with elements that are essential for nitrogen fixation, including phosphorus, zinc, iron, manganese and molybdenum.

The importance of nitrogen-fixing symbioses to soil fertility

Nitrogen input through biological nitrogen fixation is approximately 122 million tons of nitrogen per year of which 55–60 million tons is fixed by agricultural crops. The potential of biological nitrogen fixation providing nitrogen in ecosystems is increasingly being exploited in agricultural practices, mostly through legume cultivation (e.g., soybean, lupin, alfalfa, chickpea, cowpea). The nitrogen fixed by legumes can also represent an important renewable source of nitrogen for agricultural soils. However, the net inputs of fixed nitrogen into soils depend upon the amounts of nitrogen fixed relative to the amounts of nitrogen removed in the protein-rich legume products which can represent between 45% and 75% of the nitrogen in the above-ground biomass of crop legumes (Peoples et al., 2009). The residual contribution of fixed nitrogen into agricultural systems following legume cropping might be then relatively small. Despite this, the yields of crops are often greater when legumes are included in cropping sequences than when they are not. Several non-leguminous plants, mainly cereals (maize, barley, sugarcane), have also developed multiple strategies in association with diazotrophs (*Azospirillum*, *Azotobacter*, *Pseudomonas*, *Bacillus*) to cope with nitrogen deficiency. The most sophisticated associations are root nodule endosymbioses between *Frankia* and actinorhizal plants, *Rhizobium* and *Parasponia* sp., and cyanobacteria that associate with *Gunnera* sp.

In addition to their nitrogen-fixing abilities, diazotrophic bacteria can also solubilize phosphorus so that it is available to the plant. They can also provide specific iron chelators (siderophores) and synthesize phytohormones (such as cytokinins, gibberellins, indole-3-acetic acid), that promote plant growth and yield and cause positive changes in soil structure and microbial community. Many diazotrophic strains belonging to *Rhizobia*, *Azotobacter*, *Azospirillum*, *Pseudomonas*, *Klebsiella*, and *Bacillus* genera were reported to enhance the plant growth and grain yield of chickpea, common bean, pea, wheat and rice through phytohormones, siderophores and secondary metabolites production. *Rhizobia* has also been reported to act as biocontrol agents against pathogenic fungi through hydrocyanic acid, antibiotics and/or mycolytic enzymes production.

Strategies for engineering symbiotic nitrogen fixation

In agriculture nitrogen limitation is circumvented by application of fertilizers derived from the Haber-Bosch process of inorganic nitrogen-fixation. Despite the significant contribution of synthetic fertilizers, nitrogen requirements for food production increase from year to year, while the overuse of agrochemicals compromise soil health and agricultural sustainability. With the increasing global population and problems caused by unintended nitrogen wastage and pollution, more sustainable ways of managing the nitrogen cycle in soil and utilizing biological nitrogen fixation have become imperative. Symbiotic nitrogen fixation is largely limited to legumes in agricultural systems, but there are a number of microorganisms, including some diazotrophs, that inhabit the rhizosphere of other crop plants, some of which have been shown to enhance plant growth. One of the greatest challenges in agriculture is supplying sufficient plant-available nitrogen to cereal crops, which provide 45% of the world's dietary energy consumed directly by humans in addition to providing feed for animal agriculture and a feedstock for bio-based fuel production and manufacturing. Therefore, there is growing interest in increasing the contribution of biological nitrogen fixation to the growth of crop plants in agriculture.

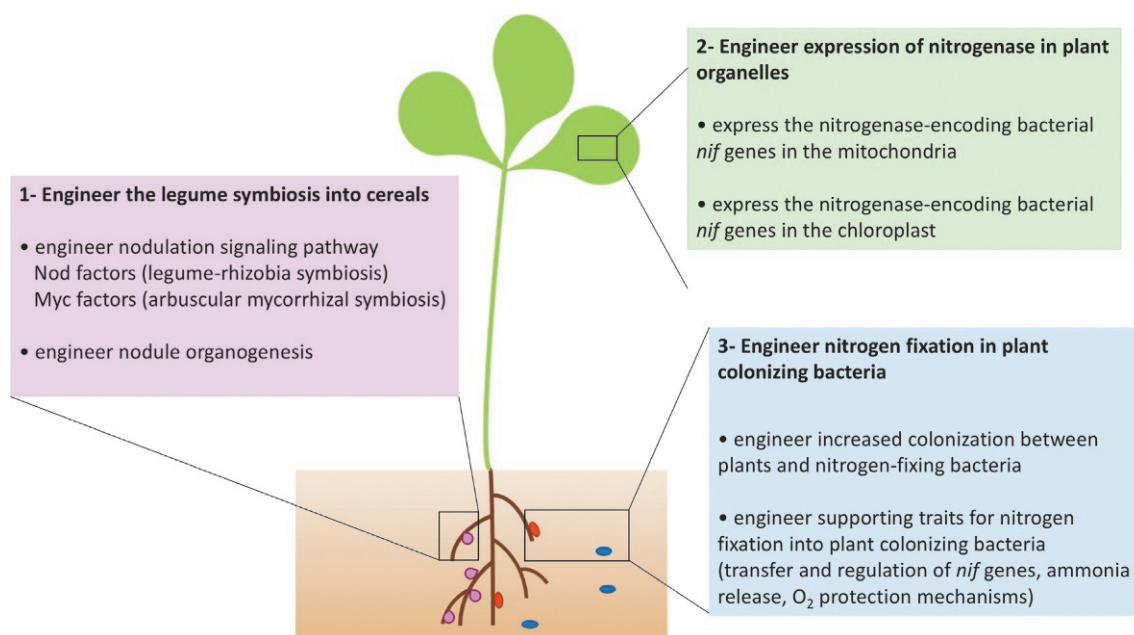


Fig. 5 Main approaches to engineer or improve biological nitrogen fixation in cereals: 1-Engineer the legume symbiosis into cereals; 2-Engineer expression of nitrogenase in plant organelles; 3-Engineer nitrogen fixation in plant colonizing bacteria (Oldroyd and Dixon, 2014; Rogers and Oldroyd, 2014; Mus et al., 2016; Pankiewicz et al., 2019).

Three approaches have arisen with the common goal of engineering biological nitrogen fixation in plants, either by (1) engineering the legume symbiosis into cereals or by (2) transferring the components required for nitrogenase activity into plant organelles such as chloroplast and mitochondria, or by (3) involving the enhancement of naturally occurring plant-associated diazotrophs by generating strains that release fixed nitrogen to benefit the crop (Fig. 5). Although the two first approaches are reliant on complex synthetic biology strategies, which will probably take considerable time to develop as viable products for farmers, the third approach could potentially provide a means to reduce the use of synthetic nitrogen fertilizers thus providing a short-term solution to the nitrogen crisis.

However, the ability to stringently control this association and eliminate escape mutants and DNA release to the environment will be an aspect to take into consideration when developing strategies and tools for engineering symbiotic nitrogen fixation.

Engineering the legume symbiosis into cereals

Most land plants, including cereals, can form arbuscular mycorrhizal associations but are unable to form nitrogen-fixing root nodule symbioses. Several components of the legume symbiotic signaling (SYM) pathway also play a role in the arbuscular mycorrhizal symbiosis (Oldroyd and Dixon, 2014). Since cereals contain the SYM pathway for arbuscular mycorrhizal associations, nodulation could be established in them by engineering the perception of rhizobial signaling molecules to activate this pathway, as well as by engineering its outputs of activation into an oxygen-limited nodule-like root organ for nitrogen fixation. Engineering synthetic symbiosis in cereal crops involves the manipulation of both nitrogen-fixing microorganisms and nitrogen fixation associations to exchange appropriate signals to establish successful colonization and nitrogen fixation. In this approach, plants can be engineered to secrete a specialized carbon source that specifically enhances the competitiveness of newly introduced nitrogen-fixing microbes in the rhizosphere. However, the availability of a wide range of well characterized promoter elements in cereals poses a challenge in engineering symbiotic nitrogen fixation in cereals. Indeed, the use of the same promoter to express multiple genes in transgenic plants can induce gene silencing. The construction of large multigene synthetic cassettes and the persistent need to improve cereal transformation technology cause other immediate barriers to engineer nitrogen-fixing capability in cereal crops.

Engineering expression of nitrogenase in organelles of plants

The introducing nitrogenase-encoding bacterial *nif* genes into plants is a direct approach to engineer nitrogen fixation in non-legumes. However, the complexity of nitrogenase biosynthesis and the sensitivity of nitrogenase to oxygen present a significant challenge. Extensive genetic and biochemical studies have identified the common core set of genes/gene products required for functional nitrogenase biosynthesis (Mus et al., 2016). In addition, plastids and mitochondria offer potential subcellular low-oxygen environments to express active nitrogenase in plants, making this engineering strategy feasible.

Engineering nitrogen fixation in plant colonizing bacteria

Two distinct approaches have been proposed to enhance pre-existing interactions between plants and bacteria, either engineer increased colonization between plants and highly efficient nitrogen-fixing microbes or engineer transfer of efficient nitrogen fixation into bacteria that already associate closely with cereals (Fig. 4) (Geddes et al., 2015). During the last 40 years, there has been a rapid expansion in genetic engineering tools, and advances made in the field of synthetic biology (precision expression control, multi-gene and assembly, synthetic regulation) may aid the engineering and transfer the supporting traits for nitrogen fixation. However, the factors that are required for nitrogen fixation are more defined than those that govern colonization. Therefore, the approach of engineering robust colonizers that fix nitrogen has arguably more practical merit.

The large nitrogen fixation island from *Pseudomonas stutzeri* was successfully transferred to the aerobic associative bacterium *Pseudomonas protegens* Pf-5, conferring the ability to grow micro-aerobically using N_2 as a sole nitrogen source. Moreover, inoculation of Arabidopsis, alfalfa, tall fescue, and wheat with transgenic *P. protegens* resulted in significant growth promotion effects compared to the near-isogenic wild-type under nitrogen-limited conditions (Setten et al., 2013). Although engineering new more-robust microbe-mediated nitrogen fixation associations requires many genes that are tightly regulated in their native host and are sensitive to environmental conditions that are not desirable in agriculture (e.g., repressed by high ammonia). To overcome this issue, it requires either transferring the pathway from one organism to another or un-silencing a cluster in a native host. However, the pathway is very sensitive to small changes in gene expression, and the regulatory control in many organisms is not well characterized. In addition, the microbe needs to establish a symbiotic relationship with the crop, which either requires engineering the plant or using endophytes, for which there may be few genetic tools (e.g., transposon mutagenesis).

Ammonia release by bacteria is another critical attribute for translating engineered nitrogen fixation by bacteria into nitrogen assimilation by plants (Geddes et al., 2015). The misregulation of the *P. stutzeri* *nif* cluster in *P. protegens* Pf-5 resulted in significant ammonia release by the bacteria. Mutations of the key regulators *nifA* or *nifL* have also resulted in ammonium excretion in *A. vinelandii*. The deletion of two ammonium transporters, encoded by *amtB1* and *amtB*, enhanced ammonia secretion in *P. stutzeri* that was further increased when combined with expression of *nif* genes from a constitutive promoter. Ammonia release by an *A. brasilense* glutamate synthetase (*glnA*) mutant may have increased growth of wheat under nitrogen-limited conditions. *A. vinelandii* with combined *nifL* and glutamine synthetase active site mutations also showed high ammonia release.

A significant challenge of transferring nitrogen fixation to aerobic microorganisms is that the nitrogenase is irreversibly inactivated by oxygen. Some aerobes such as *A. vinelandii* and *Azorhizobium caulinodans* are capable of free-living nitrogen fixation and have sophisticated mechanisms of oxygen protection. This is thought to be through maintaining high rates of oxygen consumption at the cell membrane via respiration, but it may also involve an alginate oxygen diffusion barrier and conformational protection of nitrogenase by interaction with a specific iron-sulfur protein termed the Shethna protein. Efforts are in progress to transfer oxygen protection systems, like the Shethna protein of *A. vinelandii*, to other diazotrophs.

Conclusions

The association of plants with nitrogen-fixing bacteria range from casual coexistence in the same environment to complex intracellular symbioses and allow plants to flourish in situations in which plant-assimilable nitrogen is limiting or even absent. The establishment and functioning of an effective symbiotic nitrogen fixation are dependent on genetic determinants in both plant and bacteria which are controlled by intricate regulatory mechanisms to ensure that fixed nitrogen is readily assimilated into biomass and not released to the environment. Symbiotic nitrogen fixation is restricted mainly to legumes in agricultural systems, and there is considerable interest in exploring whether similar symbioses can be developed in non-legumes, which produce the bulk of human food. Therefore, the creation of artificial symbioses or associations between diazotrophs and crops has been a primary goal in agriculture to reduce the demand for chemical nitrogen fertilizers, to improve crop yields, promote soil fertility, and reduce the environmental impacts of chemical fertilizer application.

References

- Alemayehu D, Zerihun A, and Solomon B (2018) Limitations and strategies to enhance biological nitrogen fixation in sub-humid tropics of Western Ethiopia. *Journal of Agricultural Biotechnology and Sustainable Development* 10: 122–131.
- Bournaud C, James EK, de Faria SM, Lebrun M, Melkonian R, Duponnois R, Tisseyre P, Moulin L, and Prin Y (2018) Interdependency of efficient nodulation and arbuscular mycorrhization in *Piptadenia gonoacantha*, a Brazilian legume tree. *Plant, Cell & Environment* 41: 2008–2020.
- Boyd ES, Hamilton TL, and Peters JW (2011) An alternative path for the evolution of biological nitrogen fixation. *Frontiers in Microbiology* 2: 205.
- Bueno Batista M and Dixon R (2019) Bueno Batista M, Dixon R. Manipulating nitrogen regulation in diazotrophic bacteria for agronomic benefit. *Biochemistry Society Transactions* 47: 603–614.
- Dixon R and Khan D (2004) Genetic regulation of biological nitrogen fixation. *Nature Reviews. Microbiology* 2: 621–631.
- Dos Santos PC, Fang Z, Mason SW, Setubal JC, and Dixon R (2012) Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC Genomics* 13: 162.
- Geddes BA, Ryu MH, Mus F, Garcia CA, Peters JW, Voigt CA, and Poole P (2015) Use of plant colonizing bacteria as chassis for transfer of N_2 -fixation to cereals. *Current Opinion in Biotechnology* 32: 216–222.
- Hungria M and Vargas MA (2000) Environmental factors affecting N_2 fixation in grain legumes in the tropics, with an emphasis on Brazil. *Field Crops Research* 65: 151–164.

- Inomura K, Bragg J, and Follows MJ (2017) A quantitative analysis of the direct and indirect costs of nitrogen fixation: A model based on *Azotobacter vinelandii*. *The ISME Journal* 11: 166–175.
- Maier RJ and Moshiri F (2000) Role of the *Azotobacter vinelandii* nitrogenase-protective shethna protein in preventing oxygen-mediated cell death. *Journal of Bacteriology* 182: 3854–3857.
- Mus F, Crook MB, Garcia K, Garcia CA, Geddes BA, Kouri ED, Paramasivan P, Ryu MH, Oldroyd GS, Udvardi MK, Voigt CA, Ané JM, and Peters JW (2016) Nitrogen fixation and the challenges to its extension to nonlegumes. *Applied and Environmental Microbiology* 82: 3698–3710.
- Mus F, Alleman AB, Pence N, Seefeldt LC, and Peters JW (2018) Exploring the alternatives of biological nitrogen fixation. *Metallomics* 10: 523–538.
- Oldroyd GED and Dixon R (2014) Biotechnological solutions to the nitrogen problem. *Current Opinion in Biotechnology* 26: 19–24.
- Ott T, van Dongen JT, Gunther C, Krusell L, Desbrosses G, Vigeolas H, Bock V, Czechowski T, Geigenberger P, and Udvardi MK (2005) Symbiotic leghemoglobins are crucial for nitrogen fixation in legume root nodules but not for general plant growth and development. *Current Biology* 15: 531–535.
- Pankievicz VCS, Irving TB, Maia LGS, and Ané JM (2019) Are we there yet? The long walk towards the development of efficient symbiotic associations between nitrogen-fixing bacteria and non-leguminous crops. *BMC Biology* 17: 1–17.
- Peoples MB, Brockwell J, Herridge DF, Rodchester IJ, Alves BJR, Urquiaga S, Boddey RM, Dakora FD, Bhattarai S, Maskey SL, Sampet C, Rerkasem B, Khan DF, Hauggaard-Nielsen H, and Jensen ES (2009) The contributions of nitrogen-fixing crop legumes to the productivity of agricultural systems. *Symbiosis* 48: 1–17.
- Rashid H, Krehenbrink M, and Akhtar MS (2015) In: Lichtfouse E (ed.) *Nitrogen-Fixing Plant-Microbe Symbioses*. vol. 15, pp. 193–234. Springer International Publishing.
- Rogers C and Oldroyd GE (2014) Synthetic biology approaches to engineering the nitrogen symbiosis in cereals. *Journal of Experimental Botany* 65(8): 1939–1946. <https://doi.org/10.1093/jxb/eru098>.
- Rubio LM and Ludden PW (2008) Biosynthesis of the iron-molybdenum cofactor of nitrogenase. *Annual Review of Microbiology* 62: 93–111.
- Santi C, Bogusz D, and Franche C (2013) Biological nitrogen fixation in non-legume plants. *Annals of Botany* 111: 743–767.
- Schmid B, Einsle O, Chiu HJ, Willing A, Yoshida M, Howard JB, and Rees DC (2002) Biochemical and structural characterization of the cross-linked complex of nitrogenase: Comparison to the ADP-AIF4(–)-stabilized structure. *Biochemistry* 41: 15557–15565.
- Seefeldt LC, Hoffman BM, and Dean DR (2009) Mechanism of Mo-dependent nitrogenase. *Annual Review of Biochemistry* 78: 701–722.
- Setten L, Soto G, Mozzicafreddo M, Fox AR, Lisi C, Cuccioloni M, Angeletti M, Pagano E, Díaz-Paleo A, and Ayub ND (2013) Engineering *Pseudomonas protegens* Pf-5 for nitrogen fixation and its application to improve plant growth under nitrogen-deficient conditions. *PLoS One* 8: 1–14.
- Stal LJ and Krumbien WE (1985) Nitrogenase activity in the non-heterocystous cyanobacterium *Oscillatoria* sp. grown under alternating light-dark cycles. *Archives of Microbiology* 143: 67–71.
- White J, Prell J, James EK, and Poole P (2007) Nutrient sharing between symbionts. *Plant Physiology* 144: 604–614.
- Zahran HH (1999) Rhizobium-legume symbiosis and nitrogen fixation under severe conditions and in an arid climate. *Microbiology and Molecular Biology Reviews* 63(4): 968–989.

Nitrogen: Non-Symbiotic Nitrogen fixation in soils

Vadakattu VSR Gupta^a and Sarah S Roley^b, ^aCSIRO Ag & Food, Waite Road, Urrbrae, SA, Australia; ^bSchool of the Environment, Washington State University, Richland, WA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	283
Ecology of NS-N ₂ fixation	284
Significance of NS-N ₂ fixation in soils and terrestrial environments	285
Nitrogenase and diazotroph communities	287
Factors affecting functionality	289
Future research directions	290
Outlook	290
References	291

Abstract

Non-symbiotic nitrogen fixation (NS-N₂ fixation) is the microbially-catalyzed conversion of dinitrogen gas to ammonia in the absence of a plant symbiont. A wide range of Bacteria and Archaea are capable of NS-N₂ fixation, although Phyla Proteobacteria and Cyanobacteria are often the most abundant NS-N₂ fixing microorganisms present in soils. NS-N₂ fixation rates vary both seasonally and geographically, depending on soil physicochemical conditions, identity of overlying plants, and microbial community composition. This process can be an important source of N to ecosystems. In agriculture, NS-N₂ fixation rates can potentially be enhanced by optimizing physicochemical and biological conditions for NS-N₂ fixation, which may include customized plant- NS-N₂ fixing microorganism combinations.

Key Points

- Non-symbiotic nitrogen fixers contribute the nitrogen fixed to many ecosystems.
- A wide and diverse array of microbial taxa are capable of Non-Symbiotic (NS) Nitrogen fixation, although many soils are dominated by organisms in Phylum Proteobacteria.
- Ideal conditions for NS-N₂ fixation include low soil nitrogen, high soil carbon, low oxygen, and high soil moisture.
- Conditions suitable for NS-N₂ fixation are found in many ecosystems.
- Future research priorities should include better annual or seasonal estimates of NS-N₂ fixation, improved characterization of the plant-N₂ fixing microorganism relationship, and identification of functionally active taxa.
- Stimulating NS-N₂ fixation in agricultural fields may improve sustainability by reducing reliance on synthetic fertilizers.

Introduction

Nitrogen (N) is an essential element for all organisms to synthesize proteins and DNA; it is one of the most abundant elements in organisms, after carbon (C), oxygen (O₂), and hydrogen. Because of the high demand for N, it is often in short supply in soils and is the nutrient that most often limits growth of terrestrial vegetation. Prior to the development of synthetic N fertilizer, the primary source of N to terrestrial ecosystems was biological N₂ fixation, the microbial catalysis of dinitrogen gas (N₂) into biologically available ammonia (NH₃).

N is abundant on Earth, with 78% of the atmosphere containing N₂ gas. This N is mostly unavailable, though, because its triple bond is difficult to break. N₂-fixing microorganisms break the triple bond with the nitrogenase enzyme and high energy inputs, in the form of ATP ($\text{N}_2 + 8\text{H} + 8\text{e}^- + 16\text{ATP} + \text{Nitrogenase} \rightarrow 2\text{NH}_3 + \text{H}_2 + 16\text{ADP} + 16\text{Pi}$). Some N₂-fixing microorganisms form a symbiosis with plants where the plants provide energy, as C, in exchange for N from the N₂ fixed. Well-known N₂-fixing symbioses include legumes (partnered with members of the Rhizobiales), and actinorhizal plants, such as alder, (partnered with *Frankia*). However, a wide array of Bacteria and some Archaea can fix N₂ without a symbiont, a process known as non-symbiotic N₂ fixation (NS-N₂ fixation). NS-N₂ fixation occurs across a wide variety of habitats, with a range of non-leguminous plants, cereals and grasses and NS-N₂ fixing organisms greatly outnumber symbiotic N₂ fixing bacteria (Gaby and Buckley, 2011).

The term NS-N₂ fixation includes three groups of microorganisms: (i) free-living N₂ fixing organisms (habitats include the bulk soil, decomposing plant residues, soil aggregates, and termite habitats), (ii) associative N₂-fixing organisms (habitats include the roots and rhizosphere of non-legume plants), and (iii) endophytic N₂ fixing microorganisms, that live inside growing plants.

Ecology of NS- N_2 fixation

The nitrogenase enzyme catalyzes N_2 fixation and all known N_2 -fixing organisms contain the nitrogenase enzyme complex (Fig. 1). This complex has three isozymes, characterized by the transition metal present at the active site. The molybdenum (Mo)-nitrogenase is most common, but vanadium (V) and iron (Fe)-only nitrogenases also occur and are often referred to as alternative nitrogenases. N_2 -fixing organisms with alternative nitrogenases are mostly found in Mo-limited soils. In addition to the transition metal, all nitrogenases contain phosphorus (P) and Fe.

Nitrogenase activity demands large amounts of ATP (energy) but the enzyme is inactivated by O_2 . Because of these requirements, N_2 fixation generally occurs where there is an abundant supply of energy (C for heterotrophs, sunlight for photoautotrophs) and sufficient metals and P for nitrogenase synthesis. In addition, N_2 fixation tends to occur where N is limiting because when N is abundant, it takes less energy to scavenge for N than to fix it.

N_2 -fixing microorganisms, also known as diazotrophs, have been found in a wide array of soils and other environments. Within the soil environment, they have been found in both bulk and rhizosphere soil, on and within roots, and in decomposing residues.

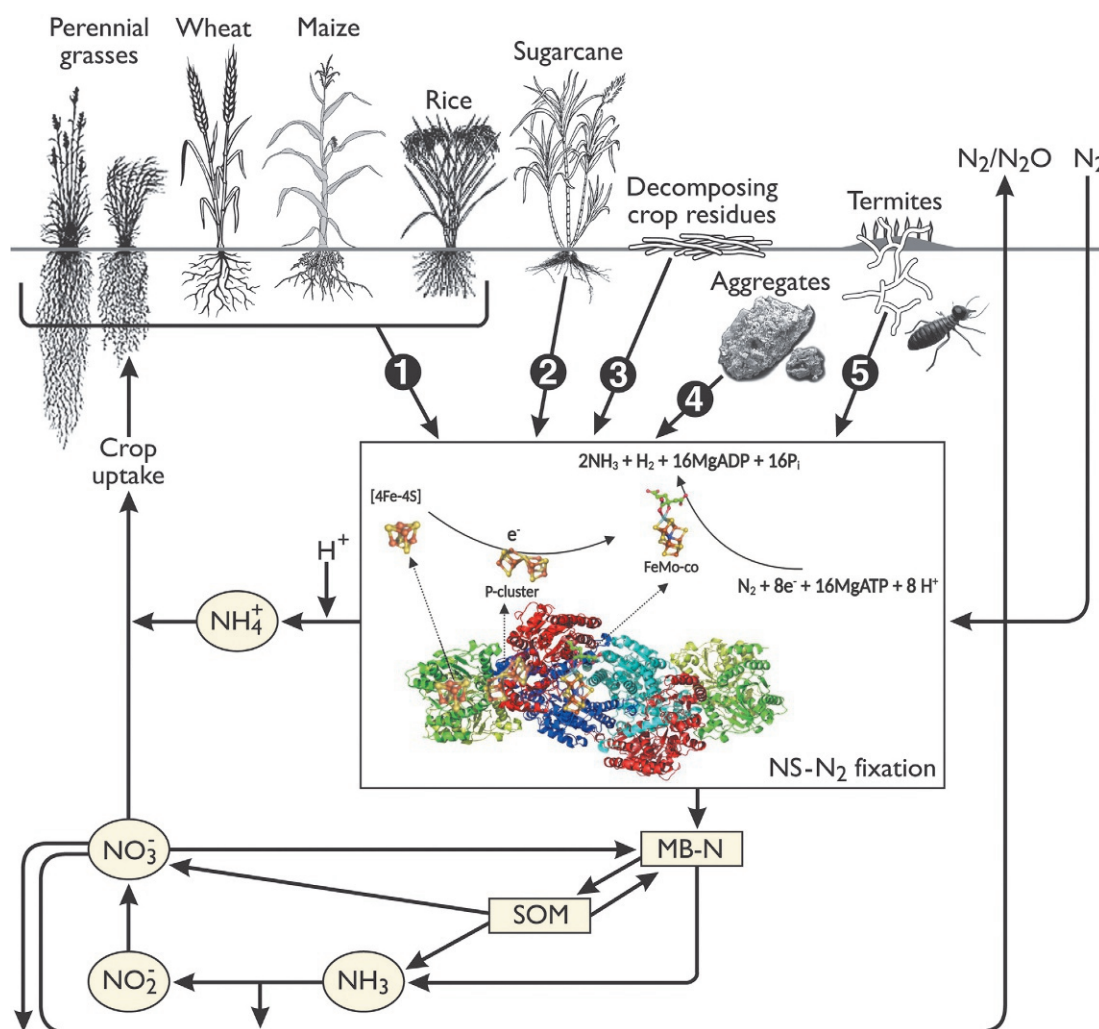


Fig. 1 A conceptual diagram highlighting the diverse set of microsites that can support Non-Symbiotic N_2 fixation and its role in the soil N cycle. (1) A diverse array of bacterial genera have been found in the rhizosphere and endophytic environment of a variety of cereals and other plants; the amount of N_2 fixation in the field environment is yet to be properly quantified, (2) A significant amount of N_2 fixation has been shown to occur in the below- and above-ground plant environments with sugarcane, (3) fresh decomposing residues, especially with wide C:N ratios, provide optimal conditions for N_2 fixation, (4) stable aggregates not only provide protective sites for free-living N_2 -fixing bacteria but also provide the required low oxygen conditions for nitrogenase activity, and (5) a diverse group of microflora in termite guts and nests possess *nifH* genes that show potential for N_2 fixation in natural environments mainly in semi-arid and arid ecosystems (Modified from Roper MM and Gupta VSR (2016) Enhancing Non-symbiotic N_2 fixation in Agriculture. *The Open Agriculture Journal* 10: 7–27. <https://doi.org/10.2174/1874331501610010007>). Details of the mechanism for N_2 fixation and nitrogenase involvement in diazotrophs is adapted from and Aasfar A, Bargaz A, Yaakoubi K, Hilali A, Bennis I, Zeroual Y, Kadmiri IM (2021) Nitrogen fixing *Azotobacter* species as potential soil biological enhancers for crop nutrition and yield stability. *Frontiers in Microbiology* 12: 628379. <https://doi.org/10.3389/fmicb.2021.628379>.

Diazotrophs also live on leaf surfaces (the phyllosphere), in termite mounds and guts, and within other animals, including marine clams (Table 1; Fig. 1).

In soils, N_2 fixing organisms are a source of N for the ecosystem, but their interactions with other organisms are somewhat uncertain. Newly-fixed N has been found in a wide variety of non-leguminous and non-actinorhizal terrestrial plant species (Table 1). The relationship between plants and NS- N_2 -fixing microorganisms is not yet defined, but three potential relationships exist (Roley, 2021). First, plant benefits may be incidental, with diazotrophs fixing N solely for their own needs and the fixed-N cycling rapidly through the soil to be taken up by plants. Second, plants and diazotrophs may have a mutualistic relationship, in which plants exude carbohydrates or other materials to attract diazotrophs, and the plant and the diazotroph exchange materials. Third, plants may have a predatory relationship with microbes, in which they excrete oxidative species that remove the Fixed-N from diazotrophs. The exact nature of the relationship likely varies with host plant, diazotroph species and possibly with physicochemical conditions.

Likewise, the relationship of N_2 -fixing endophytes within the plant remains poorly defined. A diverse and abundant N_2 -fixing community has been found inside plants (Moyes et al., 2016). In some cases, these organisms clearly provide Fixed-N to the plant. In other cases, they provide plant growth-promoting benefits, in addition to or instead of, N_2 fixation, such as phytohormone and vitamin synthesis and improved nutrient uptake (Dobbelaere et al., 2003). But in other cases, the functional role of the N_2 -fixing organisms is unclear. Some diazotrophs in plant tissues could be there incidentally, moving into plant tissues via the transpiration stream without providing benefits, while others could be recruited by the plant. Several species of termites acquire N by hosting NS- N_2 fixing bacteria in their gut. Fixation rates and diazotroph abundance are highest in the workers and species with an N-poor diet (i.e., wood-feeding (N-poor) compared to dung-feeding (N-rich) species).

The transfer of Fixed-N to plants depends upon whether fixation occurs in the soil or the endosphere. Nitrogen fixed in the soil needs to be released from microbes prior to plant uptake. N release can occur via excretion (Bageshwar et al., 2017) or upon cell death and subsequent decomposition. The released N is subjected to N cycling processes (Fig. 1) (Aasfar et al., 2021), whereas the N fixed by endophytic diazotrophs can be directly available to the host plant. Direct provision of N from diazotrophs to plants may occur in the rhizosphere, likely in exchange for C, but this has yet to be confirmed by observation.

Significance of NS- N_2 fixation in soils and terrestrial environments

A growing body of evidence suggests that NS- N_2 fixation is a significant contributor to the N cycle in many ecosystems. In general, NS- N_2 fixation inputs are greatest where N availability is low and the C:N ratio is high. These conditions, along with NS- N_2 fixation, are present in all ecosystem types (agricultural fields, grasslands, forests, deserts), although inputs from fixation vary tremendously within and among ecosystems (Fig. 1, Table 1).

N_2 fixation is a tantalizing alternative to fertilizer addition in sugarcane and cereal cropping ecosystems; hence, it has been studied extensively in these systems. Supplemental N (as synthetic fertilizer or organic N) is typically used on these crops, but some of the added N will be removed via leaching, denitrification, and immobilization before it can be used by the crop. As a result, during peak N demand, N requirements can exceed supply from N mineralization and N fertilizers. Sugarcane has particularly high NS- N_2 fixation rates, receiving more than half of its N from endophytic diazotrophs (roots, stems, and leaves), with rates as high as $150 \text{ kg N ha}^{-1} \text{ year}^{-1}$ (Boddey et al., 2003). Furthermore, in agricultural soils with a lower level of fertility and organic matter, e.g., in southern and western Australia, India, Africa and Mediterranean regions, non-leguminous crops would benefit from increased N inputs from NS- N_2 fixation.

NS- N_2 fixation rates are generally smaller in cereal crops, but are still of agronomic significance. Globally, N-fertilized rice, wheat, and maize obtain up to 24% of their N from NS- N_2 fixation (Ladha et al., 2016). Rice paddies offer ideal conditions for fixation; the water-logged soils are low in O_2 and high in C. Rice also harbors a diverse community of diazotrophic endophytes (Minamisawa et al., 2005); however, inoculation with known active NS- N_2 -fixing organisms increases growth rate, grain yield, and protein content (Fig. 2) (Gaby and Buckley, 2014). Cyanobacteria and other photosynthetic NS- N_2 -fixing microorganisms inhabit rice paddies, fixing N that cycles through the soil and into rice. Wheat is grown in drier upland soils, but still receives agronomically significant levels of N from NS- N_2 fixation ($10\text{--}40 \text{ kg N ha}^{-1} \text{ year}^{-1}$, Ladha et al., 2016). NS- N_2 fixation likely contributes 29–92% of N nutrition in ancient Mexican maize land races (Van Deynze et al., 2018).

In croplands, NS- N_2 fixation has generally been measured in the rhizosphere, but it also occurs in association with crop residues, which contain large amounts of C. Conservation agricultural practices, which leave crop residues on the field, may increase N inputs from NS- N_2 fixation, particularly during non-crop periods. For example, residue NS- N_2 fixation contributed $<10 \text{ kg}$ to 38 kg N ha^{-1} during non-crop periods in southern and eastern Australia (Gupta et al., 2006).

Grasslands vary widely in their NS- N_2 fixation inputs (Table 1). In tropical grasslands, the overlying species appears to be important, with rapidly-growing species such as bahiagrass accumulating up to $37 \text{ kg N ha}^{-1} \text{ year}^{-1}$ from fixation. Slow-growing species, such as *Aristida laevis*, accumulate much less N from fixation ($<10 \text{ kg N ha}^{-1} \text{ year}^{-1}$), but still receive up to a third of their annual N needs from NS- N_2 fixation (Marques et al., 2017). In temperate grasslands, mass balance estimates are as high as $54 \text{ kg N ha}^{-1} \text{ year}^{-1}$ (Chapman et al., 1949), but estimates from laboratory incubations are often much lower, $<10 \text{ kg N ha}^{-1} \text{ year}^{-1}$. The discrepancy in rate measurements could be due to strong spatial or temporal variability: N_2 fixation may occur episodically, with occasional bursts of N_2 fixation activity responsible for most fixation on an annual scale (Roley et al., 2019). Although NS- N_2 fixation has been reported in the phyllosphere of perennial grasses, their contribution to overall plant N nutrition is not known (Gupta et al., 2019).

Table 1 Examples of non-symbiotic N₂-fixation and diazotroph taxa in different habitats.

System	Location	Common members/groups	Function (N ₂ -fixed)	References
Temperate grasslands	Rhizosphere	α -, β - and δ -Proteobacteria	4–54 kg N ha ⁻¹ year ⁻¹	Roley et al. (2019) and Jesus et al. (2016)
Tropical grasslands	Rhizosphere	<i>Azotobacter</i> , <i>Azospirillum</i> , <i>Herbaspirillum</i>	0.6–37.6 kg N ha ⁻¹ year ⁻¹	Marques et al. (2017)
Perennial grasses, Australia	Roots and Rhizosphere	α -, β -, and λ -Proteobacteria, <i>Opitutae</i>	0.9–2.35 mg N kg soil ⁻¹ d ⁻¹ and 0.75–1.25 mg N kg root ⁻¹ d ⁻¹	Gupta et al. (2014) and Gupta et al. (2019)
Rice	Rhizosphere, soil, roots	<i>Azotobacter</i> , <i>Azomonas</i> , <i>Beijerinckia</i> , <i>Azospirillum</i> , <i>Gluconacetobacter diazotrophicus</i> , <i>Herbaspirillum</i> , <i>Burkholderia</i> , <i>Azoarcus</i> sp., <i>Clostridia</i>	18–51 kg N ha ⁻¹ crop ⁻¹	Ladha et al. (2016), (Ladha and Reddy, 2019), Boddey et al. (2000) and Bei et al. (2013)
Rice	Soil	<i>Azolla</i> , Cyanobacteria	10–80 kg N ha ⁻¹ crop ⁻¹	Roger and Ladha (1992)
Sugarcane	Rhizosphere, roots, stems, leaves	<i>Azospirillum</i> , <i>Beijerinckia</i> , <i>Pseudomonas</i> , <i>G. diazotrophicus</i> , <i>Herbaspirillum seropedicae</i> , <i>Herbaspirillum rubrisubalbicans</i> , and <i>Burkholderia silvatlantica</i>	24–150 kg N ha ⁻¹ crop ⁻¹	Boddey et al. (2003), Li et al. (2017) and Urquiaga et al. (2012)
Wheat	Rhizosphere	<i>Azotobacter chroococcum</i>	40 kg N ha ⁻¹ crop ⁻¹	Bageshwar et al. (2017)
Wheat	Rhizosphere, roots	α - and β -Proteobacteria, <i>Verrucomicrobia</i> , <i>Opitutae</i> , Bacillales	0.2–2.3 mg N kg ⁻¹ d ⁻¹	Gupta et al. (2014)
Maize	Rhizosphere	<i>Bradyrhizobium</i> , <i>Rhodopseudomonas</i> , <i>Methylocella</i>	0.33 kg N ha ⁻¹ d ⁻¹	Hsu and Buckley (2009) and Wang et al. (2018)
Tropical rainforest	Rhizosphere	α -, β -, λ - and δ -Proteobacteria, Cyanobacteria	0.1–60 kg N ha ⁻¹ year ⁻¹	Mirza et al. (2014) and Reed et al. (2011)
Temperate forest (deciduous and coniferous)	Rhizosphere	α -, δ -, λ - and β -Proteobacteria,	0.3–42 kg N ha ⁻¹ year ⁻¹	Bormann et al. (2002), Reed et al. (2011) and Berthrong et al. (2014)
Rainforests	Phyllosphere	λ -proteobacteria, Cyanobacteria	1–6 μ mol N ₂ sq. M ⁻¹ d ⁻¹	Goosem and Lamb (1986), Furnkranz et al. (2008) and Bao et al. (2019)
Alpine forest, United States	Interior of <i>Pinus flexilis</i> needles	α -Proteobacteria, Bacteroidetes; Acetobacteraceae, <i>Caulobacterales</i> , <i>Sphingomonadales</i> , <i>Rickettsiales</i> , <i>Rhodospirillales</i> , <i>Rhizobiales</i>	¹⁵ N uptake based confirmation	Moyes et al. (2016)
Fruit trees	Phyllosphere	<i>Skermanella</i> , <i>Bradyrhizobium</i> , <i>Erwinia</i> , <i>Pseudomonas</i> , <i>Chroococcidiopsis</i>	N/A	Liang et al. (2019)
Termites	Gut, mounds	<i>Clostridia</i> , <i>Proteobacteria</i> , <i>Citrobacter freundii</i> , <i>Enterobacter agglomerans</i> , <i>Desulfovibrio</i> sp., <i>Fibrobacter succinogenes</i> , Spirochaetes	<1–5 μ mol C ₂ H ₄ formed d ⁻¹ g ⁻¹ (wet wt)	Ohkuma et al. (1999) and Sapountzis et al. (2016)
Desert	Soil crust	α -, β - and λ -Proteobacteria (early successional), Cyanobacteria (late successional)	0.01–13 kg N ha ⁻¹ year ⁻¹	Miao et al. (2020), Che et al. (2018), Pepe-Ranney et al. (2016) and Reed et al. (2011)
Salt marshes, Atlantic Coast, United States	Sediment	α -, β - and λ -Proteobacteria	N budget estimates	Davis et al. (2018)

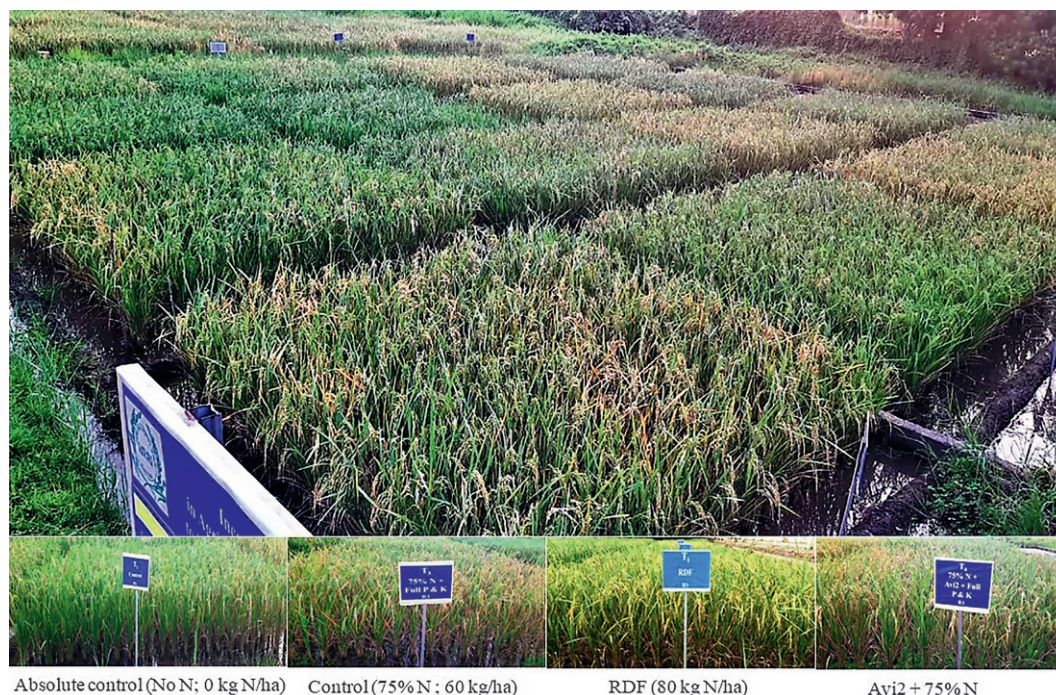


Fig. 2 Field experiments evaluating the benefits from endophytic N_2 -fixing inoculant *Azotobacter chroococcum* (Avi2) in rice crop var. Naveen at ICAR-NRRI, Cuttack in India (Kumar et al., 2021).

Forests appear to be particularly adept at N_2 fixation. NS- N_2 fixation has been measured in temperate deciduous and coniferous forests across a wide geographic range including most continents. Rates of up to $62 \text{ kg N ha}^{-1} \text{ year}^{-1}$ have been reported, with the highest rate in red pine in North America (Bormann et al., 2002). Similarly, NS- N_2 fixation in tropical rainforests ranges from 0.1 to $60 \text{ kg N ha}^{-1} \text{ year}^{-1}$ (Reed et al., 2011). Tropical forest soils are often highly weathered, and species that can fix their own N have a competitive advantage because they can harness more resources to extract P and micronutrients from the depleted soils. In forests, NS- N_2 fixation has primarily been measured in soils, but there is also evidence for fixation by needle endophytes (Moyes et al., 2016).

In deserts, biological soil crusts are an important source of N. The crusts occupy inter-plant spaces over the desert soil and consist of an array of microbial taxa, mostly in the Phyla Cyanobacteria and Proteobacteria (Pepe-Ranney et al., 2016). N_2 fixed in these crusts percolates downward to the soil and is a key N input, with contributions of up to $13 \text{ kg N ha}^{-1} \text{ year}^{-2}$ (Reed et al., 2011). NS- N_2 fixation also occurs in desert soils, in the absence of crusts, and can account for nearly 20% of the total soil N pool (Miao et al., 2020).

Nitrogenase and diazotroph communities

Diazotrophs are microbial taxa that contain the *nifH* gene, which codes for nitrogenase. The *nifH* gene is widespread across the Bacteria (heterotrophic and autotrophic) and Archaea, which demonstrates the adaptability of diazotrophs to a wide range of conditions and locations.

More than 50 species of diazotrophs have been cultured from soil and plant environments. Some of the extensively studied free-living bacterial species include *Azospirillum brasilense* (aerobe), *Azotobacter vinelandii* (an obligate aerobe), *Gluconacetobacter diazotrophicus*, *Herbaspirillum* sp., and *Klebsiella pneumoniae* (facultative anaerobes), *Clostridium pasteurianum* (an obligate anaerobe), *Rhodobacter capsulatus* (a photosynthetic bacterium) and various *Anabaena* and *Nostoc* species (heterocyst-forming cyanobacteria). Culturable taxa have been used in laboratory experiments that addressed the effects of environmental conditions (e.g., O_2 concentration, C sources, nutrients, temperature) on NS- N_2 fixation. A number of diazotroph cultures are available for use in sugarcane and rice fields, including *Azospirillum brasilense*, *Azotobacter chroococcum*, *A. vinelandii* and Cyanobacteria (Kumar et al., 2021, Fig. 2). However, most NS- N_2 fixing organisms are difficult to culture on artificial media. Our knowledge of their existence mainly comes from studies that use molecular techniques.

The nitrogenase enzyme complex is encoded by *nifH*, *nifD* and *nifK* genes, although the *nifH* gene is used as the signature gene for molecular diversity studies. The phylogenies of *nifD*, *nifK*, *nifE* and *nifN* appear to agree with *nifH*-based nitrogenase clusters, suggesting that the use of *nifH* is justified. Five major clusters with homology to *nifH* have been described, based on the *nifH* sequences available in public databases (Gaby and Buckley, 2011). Aligned sequences in such databases are required for

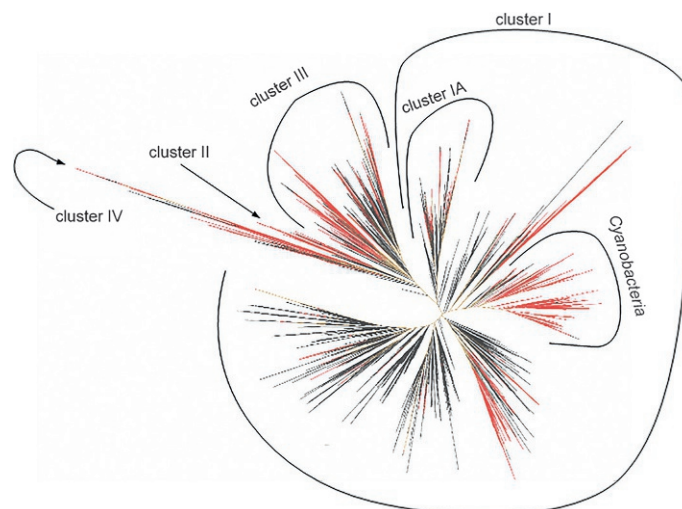


Fig. 3 The phylogenetic tree of all *nifH* homologs in the database. Red branches are low G + C sequences. Major phylogenetic groups are labelled. There are 7030 sequences (21.35% of the total database) in *nifH* cluster III and subcluster IA, the clusters that contain primarily obligate anaerobes. There are 24,799 sequences (75.31%) in *nifH* Cluster I, which mainly contains members of Proteobacteria, Firmicutes, and Actinobacteria (aerobic and facultative anaerobes). The database also contains 578 sequences (1.76%) in *nifH* Cluster II, which comprises the iron-dependent alternative nitrogenase cluster and obligate anaerobes including methanogenic Archaea, and 524 sequences in cluster IV which contains genes not functionally associated with nitrogenase. From Gaby JC, Buckley DH (2011 and 2014) A global census of nitrogenase diversity. *Environmental Microbiology* 13(7): 1790–1799. <https://doi.org/10.1111/j.1462-2920.2011.02488.x>.

phylogenetic and diversity analyses and for designing PCR probes for taxa and functional gene abundance. The diversity of diazotrophs, based on *nifH* sequence homology, is not distributed evenly across phylogenetic groups or environments and a majority of the soil *nifH* diversity is still undiscovered, particularly the anaerobic taxa (Gaby and Buckley, 2011).

The use of *nifH* gene-based analysis has provided a more complete picture of diazotrophic community composition and diversity and has expanded the list of known environments for NS- N_2 fixing microorganisms. Sequencing analysis places the *nifH* gene in phylogenetically diverse groups of Bacteria and Archaea. Based on the *nifH* gene homologs, diazotrophs are divided into 5 phylogenetic clusters (Fig. 3). In soils, most taxa are often in Phylum Proteobacteria. Alpha (α), beta (β), gamma (γ), and delta (δ) Proteobacteria are all common, with α -Proteobacteria often dominating. Within the α -Proteobacteria, Order Rhizobiales are most widely found. Other common phyla include Cyanobacteria, particularly in soil crusts. At the class, order, and genus-level, there is tremendous among-site variation, with that variation driven by physicochemical and biological factors. Nonetheless, there are some genera that show up in a variety of environments, including *Bradyrhizobium*, *Geobacter*, *Gluconacetobacter*, *Azospirillum*, *Azoarcus* and *Burkholderia*.

The composition of NS- N_2 fixing communities is shaped by physicochemical factors, including pH, soil texture, and nutrient availability. Soil management, which includes tillage type, plant type, and chemical application, also influences the NS- N_2 fixer community, likely through physicochemical factors. Bacterial phylotype diversity and richness are maximized at circum-neutral pH. In general, fertilizer addition reduces the abundance and diversity of diazotrophs (Ouyang et al., 2018), while organic matter inputs increase diversity and abundance, most likely by increasing the C:N ratio. Conversion to row crops reduces microbial biomass and changes the diazotrophic community, presumably as a result of the numerous plant and physicochemical changes associated with row crop conversion. The diversity of diazotroph communities has a significant relationship with the particulate organic carbon and water-soluble organic C to N ratio in soil. Agricultural pollution reduces diazotroph diversity and alters community composition (Jing et al., 2015). At the field and regional scale, management and plant type drive diazotroph community composition, whereas on a continental scale, climate and soil properties are the principal drivers.

Diverse and abundant communities of diazotrophs have been found in the above-ground plant parts (phyllosphere) of a variety of annual and perennial crops. Phyllosphere diazotroph community composition can vary with plant genotype and can be distinct from the root diazotrophic community (Gupta et al., 2019; Bao et al., 2019).

Sequencing is a powerful tool for understanding the diversity of the soil microbiome, but it has some limitations. Extracted DNA comes from both living and dead organisms and so does not indicate which individuals are active in the community. For N_2 fixation, researchers can target the *nifH* gene to determine which NS- N_2 -fixing microorganisms are present, but it does not determine which taxa are actively fixing N. More advanced molecular techniques can help identify the active taxa. Extraction of *nifH* RNA allows identification of organisms that are actively transcribing *nifH* and therefore fixing N (Bahulikar et al., 2014). Stable isotope probing determines which organisms have Fixed-N and subsequently incorporated it into their tissues (Pepe-Ranney et al., 2016). These approaches can be limited though, because they represent a single point in time and are expensive and technically challenging.

Molecular approaches can identify the plant genetic factors that determine rhizosphere NS- N_2 -fixer community composition. For example, multiple genes control rhizosphere N_2 fixation rates in rice (Wu et al., 1995). Within a plant species, cultivars can vary in rhizosphere NS- N_2 -fixer community composition, *nifH* gene abundance, and NS- N_2 fixation rates. Additionally, there is some evidence that older crop varieties and ancestral parents support more NS- N_2 fixation than modern, high-yielding varieties. This cultivar-based variation has been observed in wheat, barley, rice, grasses, maize, millet, and sugarcane, suggesting a plant-based genetic basis for these factors (Roper and Gupta, 2016). With a few exceptions, the specific plant genes responsible for these differences are unknown. Further knowledge of gene-based mechanisms would help in the development of new germplasm (cultivars) with enhanced NS- N_2 fixation potential.

Factors affecting functionality

NS- N_2 fixation is mainly influenced by energy availability along with soil physicochemical properties, environmental and biotic factors, including O_2 availability, soil moisture content, soil texture, nutrient availability, and overlying plant communities (Fig. 4).

The rhizosphere is a C-rich environment conducive to NS- N_2 fixation. It contains root exudates, decomposing soil organic matter, and plant residues, all of which provide C to diazotrophs. Plant roots exude labile C (e.g., sugars), which can be readily used by NS- N_2 -fixing microorganisms. The quality and quantity of that C varies by plant species and life stage, and this may be responsible for observed seasonal variation in NS- N_2 fixation rates. In cereal crops and grasses, rhizosheaths are enriched with organic materials and have a high density of diazotrophs, which support substantial NS- N_2 fixation. Deep and extensive root systems, exemplified in perennial grasses and trees, provide root exudates to a large volume of soil and also provide C via decomposing roots. Surface residue inputs also stimulate NS- N_2 fixation rates in cereal crops (Roper and Gupta, 2016).

NS- N_2 fixation rates are controlled by availability of key nutrients. In general, soil mineral N inhibits NS- N_2 fixation because diazotrophs save energy by taking up soil N. However, when C is abundant, excess mineral N can be immobilized by microbes, lowering mineral N availability and making NS- N_2 fixation energetically favourable. As a result, in some locations, the soil C:N ratio can be a good predictor of NS- N_2 fixation rates, because it indicates the ratio of available energy to the N required for microbial growth. For example, crop residues from cereal crops and grasses can promote NS- N_2 fixation more so than residues with a low C:N ratio, such as those from legumes. In addition, other nutrients such as Mo, Fe, and P are needed for nitrogenase synthesis. As a result, application of P can increase NS- N_2 fixation especially with crops and grasslands in nutrient-poor soils.

The enzyme nitrogenase is denatured by O_2 . In symbiotic N_2 fixation, a special O_2 -transporting protein, leghemoglobin, facilitates a carefully controlled O_2 supply to the N_2 -fixing bacteria. In contrast, NS- N_2 fixation often occurs in microaerophilic or anaerobic microsites in the soil. Some NS- N_2 -fixing microorganisms have evolved strategies for separating O_2 from nitrogenase, and this allows them to fix N_2 in aerobic soils (Puppo and Rigaud, 1986). Nonetheless, NS- N_2 fixation rates generally decrease with O_2 concentration. Similarly, NS- N_2 fixation rates generally increase with soil moisture (Roper and Gupta, 2016), because water-filled pore spaces slow O_2 diffusion and maintain low- O_2 conditions.

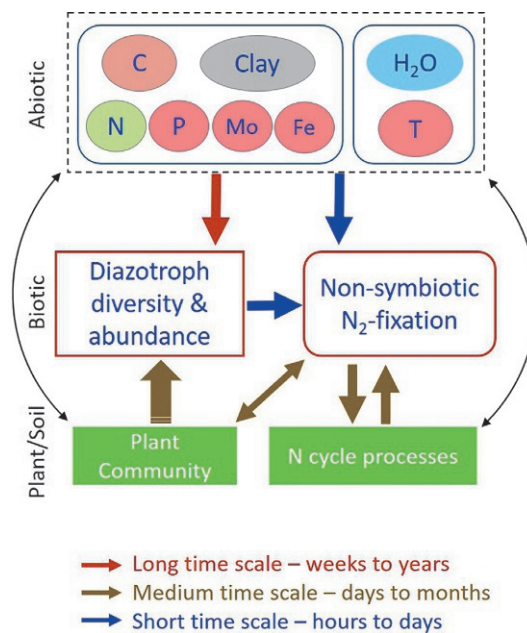


Fig. 4 Edaphic and environmental factors influencing the diversity, abundance, and functional capacity of non-symbiotic N_2 fixation.

Other important physicochemical and environmental factors include pH, soil texture, heavy metal concentration, and soil temperature. Nitrogenase activity peaks at pH 7–7.5, and NS- N_2 fixation rates decline at higher or lower pH. The amount and type of clay in soils can alter soil structure and habitat conditions affecting NS- N_2 fixation; for example, the presence of montmorillonite clay increases NS- N_2 fixation rates. Soil contamination by heavy metals reduces the abundance of diazotrophs and rate of NS- N_2 fixation. The most favourable temperatures for NS- N_2 fixation are between 30 °C and 35 °C, but the best temperature range depends upon the diazotroph taxa and habitat conditions; N_2 fixation can occur from near 0 °C to 45 °C.

Soil aggregates are important mediators of biogeochemical processes, such as denitrification and organic C accrual, and they may also influence NS- N_2 fixation rates. Survival of NS- N_2 -fixing bacteria is related to the ability of soil aggregates to protect them against biocides; biocides can reduce the amount and duration of N_2 -fixing activities in soils exposed to adverse environmental conditions. Aggregates formed in the rhizosphere tend to be water-stable and water-stable aggregates can maintain greater soil moisture and lower O_2 concentrations than the surrounding soil (Kuz'yakov and Blagodatskaya, 2015), potentially leading to higher rates of NS- N_2 fixation. Furthermore, aggregates formed in the rhizosphere environment can harbor larger populations of NS- N_2 fixing bacteria. Hence, physicochemical and biological conditions in aggregates likely favour high rates of NS- N_2 fixation.

Plant communities differ in terms of the quantity and quality of plant exudates (i.e. C sources) and signal molecules, which in turn influence both diazotroph community composition and NS- N_2 fixation rates (Fig. 4). For example, perennial grasses with their extensive root systems and large amounts of root derived C generally support more N_2 fixation compared to modern cereal crop varieties. Endophytic diazotrophs gain benefits from living inside the plant, an environment conducive to NS- N_2 fixation because the diazotrophs: (i) have direct access to plant-sourced energy, (ii) can avoid competition from other rhizosphere microbes, and (iii) are in a low- O_2 environment in the plant interior. Because of their location, N fixed by endophytic diazotrophs can be transferred more efficiently to where it is needed in the host.

Future research directions

NS- N_2 fixation remains an area of active research interest. There are several key areas where new research will help us more fully understand NS- N_2 fixation: more reliable annual and seasonal rate estimates, identification of key diazotrophs, and characterization of the underlying ecological relationships. Application of this acquired knowledge to agricultural ecosystems may lead to increased NS- N_2 fixation and improved economic and environmental sustainability.

Accurate crop season and annual estimates of NS- N_2 fixation at a biome and global scale will allow a better understanding of the importance of N_2 fixation in the global N cycle. In addition, the underlying data required to achieve accurate annual estimates will also help inform agricultural management. Most NS- N_2 fixation estimates are based on short-term measurements, conducted under laboratory conditions with samples from a small geographic area. Annual and seasonal scaling is often an extrapolation from these limited samples, which makes biome-scale estimates tenuous. Future experiments may improve scaling by measuring a comprehensive suite of physicochemical predictors and by designing experiments to adequately capture spatial and temporal variability of a given habitat (Soper et al., 2021), an approach that was successfully deployed by Gupta et al. (2006) to derive estimates for different southern Australian agricultural zones. Molecular tools have revealed a wide diversity of N_2 fixing microorganisms across different habitats, but it is usually unclear which taxa are actively fixing N. New molecular techniques, such as mRNA analysis and stable isotope probing, will allow researchers to simultaneously identify the active taxa and their functional capacity at a given time. These tools will help us address key questions, such as: 'do the active taxa vary with physicochemical conditions, geography, or overlying vegetation?' and 'are most *nifH* taxa fixing N or just a few?'

Similarly, the ecological relationship between diazotrophs and plants is unclear. Plants can influence the rhizosphere and root microbiome, including the diazotrophic community and their function. This would require a better understanding of the host-genetic factors regulating diazotroph assembly in roots for identification of key taxa specific to plant type, which could ultimately lead to management changes that stimulate NS- N_2 fixation in cereal crops. As we begin to identify microbial taxa that are actively fixing N, we can tailor investigations to those relevant taxa. In addition, we can apply existing ecological theory on plant-microbe interactions to guide studies of plant-diazotroph interactions. The combination of new molecular tools and existing ecological theory may lead to new insights on NS- N_2 fixation in the near future.

Outlook

As the relationship between active NS- N_2 fixing microorganisms and plants is better understood, we may be able to enhance NS- N_2 fixation in agricultural settings especially in low-fertility soils. Economically affordable and environmentally friendly alternatives to synthetic N fertilizers can improve global food production, particularly in low-income regions. Ideally, we will be able to identify candidate microbes that fix N at higher rates and develop cultivars that select for effective diazotroph populations. Eventually, it could lead to designer plant-diazotroph combinations. One path forward may be the reintroduction of traits from older crop varieties that promote diazotroph colonization into modern varieties to enhance biological N_2 fixation in agricultural systems.

Ideally, development of plant-diazotroph microbe systems in agriculture that behave like the legume—*Rhizobium* symbiosis will lead to improved economic and ecological sustainability for agroecosystems.

References

- Asfar A, Bargaz A, Yaakoubi K, Hilali A, Bennis I, Zeroual Y, and Kadmiri IM (2021) Nitrogen fixing *Azotobacter* species as potential soil biological enhancers for crop nutrition and yield stability. *Frontiers in Microbiology* 12: 628379. <https://doi.org/10.3389/fmicb.2021.628379>.
- Bageshwar UK, Srivastava M, Pardha-Saradhi P, Paul S, Gothandapani S, Jaat RS, Shankar P, Yadav R, Biswas DR, Kumar PA, Padarla JC, Mandal PK, Annappurna K, and Das HK (2017) An environmentally friendly engineered *Azotobacter* strain that replaces a substantial amount of urea fertilizer while sustaining the same wheat yield. *Applied and Environmental Microbiology* 83. <https://doi.org/10.1128/AEM.00590-17>. e00590-17.
- Bahulika RA, Torres-Jerez I, Worley E, Craven K, and Udvardi MK (2014) Diversity of nitrogen-fixing bacteria associated with switchgrass in the native tallgrass prairie of Northern Oklahoma. *Applied and Environmental Microbiology* 80(18): 5636–5643. <https://doi.org/10.1128/aem.02091-14>.
- Bao L, Cai W, Zhan X, Liu J, Chen H, Wei Y, Jia X, and Bai Z (2019) Distinct microbial community of phyllosphere associated with five tropical plants on Yongxing Island, South China Sea. *Microorganisms* 7: 525. <https://doi.org/10.3390/microorganisms7110525>.
- Bei QC, Liu G, Tang HY, Cadisch G, Rasche F, and Xie ZB (2013) Heterotrophic and phototrophic N-15(2) fixation and distribution of fixed N-15 in a flooded rice-soil system. *Soil Biology and Biochemistry* 59: 25–31. <https://doi.org/10.1016/j.soilbio.2013.01.008>.
- Berthrong ST, Yeager CM, Gallegos-Graves L, Steven B, Elchors SA, Jackson RB, and Kuske CR (2014) Nitrogen fertilization has a stronger effect on soil nitrogen-fixing bacterial communities than elevated atmospheric CO₂. *Applied and Environmental Microbiology* 80: 3103–3112.
- Boddey RM, da Silva LG, Reis V, Alves BJR, and Urquiaga S (2000) Assessment of bacterial nitrogen fixation in grass species. In: Triplett EW (ed.) *Prokaryotic Nitrogen Fixation: A Model System for Analysis of a Biological Process*, pp. 705–726. Horizon Scientific Press.
- Boddey RM, Urquiaga S, Alves BJR, and Reis VM (2003) Endophytic nitrogen fixation in sugar cane: Present knowledge and future applications. *Plant and Soil* 252: 139–149.
- Bormann BT, Keller CK, Wang D, and Bormann FH (2002) Lessons from the sandbox: Is unexplained nitrogen real? *Ecosystems* 5(8): 727–733. <https://doi.org/10.1007/s10021-002-0189-2>.
- Chapman HD, Liebig GF, and Rayner DS (1949) A lysimeter investigation of nitrogen gains and losses under various systems of covercropping and fertilization, and a discussion of error sources. *Hilgardia* 19(3): 57–128.
- Che RX, Deng YC, Wang F, Wang WJ, Xu ZH, Hao YB, Xue K, Zhang B, Tang L, Zhou HK, and Cui XY (2018) Autotrophic and symbiotic diazotrophs dominate nitrogen-fixing communities in Tibetan grassland soils. *Science of the Total Environment* 639: 997–1006. <https://doi.org/10.1016/j.scitotenv.2018.05.238>.
- Davis DA, Malone SL, and Lovell CR (2018) Responses of salt marsh plant rhizosphere diazotroph assemblages to drought. *Microorganisms* 6(1). <https://doi.org/10.3390/microorganisms6010027>.
- Dobbelaere S, Vanderleyden J, and Okon Y (2003) Plant growth-promoting effects of diazotrophs in the rhizosphere. *Critical Reviews in Plant Sciences* 22(2): 107–149. <https://doi.org/10.1080/713610853>.
- Furnkranz M, Wanek W, Richter A, Abell G, Rasche F, and Sessitsch A (2008) Nitrogen fixation by phyllosphere bacteria associated with higher plants and their colonizing epiphytes of a tropical lowland rainforest of Costa Rica. *The ISME Journal* 2(5): 561–570. <https://doi.org/10.1038/ismej.2008.14>.
- Gaby JC and Buckley DH (2011) A global census of nitrogenase diversity. *Environmental Microbiology* 13(7): 1790–1799. <https://doi.org/10.1111/j.1462-2920.2011.02488.x>.
- Gaby JC and Buckley DH (2014) A comprehensive aligned nifH gene database: A multipurpose tool for studies of nitrogen-fixing bacteria. *Database* Vol. 2014, Article ID bau001, <https://doi.org/10.1093/database/bau001>.
- Goosem S and Lamb D (1986) Measurements of phyllosphere nitrogen fixation in a tropical and two sub-tropical rain forests. *Journal of Tropical Ecology* 2(4): 373–376. <https://doi.org/10.1017/S0266467400001000>.
- Gupta VSR, Roper MM, and Roget DK (2006) Potential for non-symbiotic N₂-fixation in different agroecological zones of southern Australia. *Australian Journal of Soil Research* 44(4): 343. <https://doi.org/10.1071/sr05122>.
- Gupta VSR, Kroker SJ, Hicks M, Davoren CW, Descheemaeker K, and Llewellyn R (2014) Nitrogen cycling in summer active perennial grasses in South Australia: Non-symbiotic nitrogen fixation. *Crop & Pasture Science* 65: 1044–1056. <https://doi.org/10.1071/CP14109>.
- Gupta V, Zhang BZ, Penton CR, Yu JL, and Tiedje JM (2019) Diazotroph diversity and nitrogen fixation in summer active perennial grasses in a mediterranean region agricultural soil. *Frontiers in Molecular Biosciences* 6. <https://doi.org/10.3389/fmolb.2019.00115>.
- Hsu SF and Buckley DH (2009) Evidence for the functional significance of diazotroph community structure in soil. *The ISME Journal* 3(1): 124–136. <https://doi.org/10.1038/ismej.2008.82>.
- Jesus ED, Liang C, Quensen JF, Susilawati E, Jackson RD, Balser TC, and Tiedje JM (2016) Influence of corn, switchgrass, and prairie cropping systems on soil microbial communities in the upper Midwest of the United States. *Global Change Biology. Bioenergy* 8: 481–494.
- Kumar U, Kaviraj M, Dangar TK, Pannierselam P, Nayak AK (2021) Liquid bioinoculant of endophytic (*Azotobacter chroococcum*) and rhizospheric (*Azotobacter vinelandii*) nitrogen fixing bacteria for rice crop. NRRI Technical Bulletin 156, pp. 9, Cuttack, India.
- Jing HM, Xia XM, Liu HB, Zhou Z, Wu C, and Nagarajan S (2015) Anthropogenic impact on diazotrophic diversity in the mangrove rhizosphere revealed by nifH pyrosequencing. *Frontiers in Microbiology* 6. <https://doi.org/10.3389/fmicb.2015.01172>.
- Kuznyakov Y and Blagodatskaya E (2015) Microbial hotspots and hot moments in soil: Concept & review. *Soil Biology and Biochemistry* 83: 184–199. <https://doi.org/10.1016/j.soilbio.2015.01.025>.
- Ladha JK and Reddy PM (2019) The soil microbiome and crop nitrogen nutrition in anaerobic systems: A case study in rice. In: Zeigler RS (ed.) *Sustaining global Food Security—The nexus of science and policy*, pp. 278–302. Australia: CSIRO Publishing.
- Ladha JK, Tirol-Padre A, Reddy CK, Cassman KG, Verma S, Powlson DS, van Kessel C, Richter DD, Chakraborty D, and Pathak H (2016) Global nitrogen budgets in cereals: A 50-year assessment for maize, rice, and wheat production systems. *Scientific Reports* 6. <https://doi.org/10.1038/srep19355>.
- Li HB, Singh RK, Singh P, Song QQ, Xing YX, Yang LT, and Li YR (2017) Genetic diversity of nitrogen-fixing and plant growth promoting pseudomonas species isolated from sugarcane rhizosphere. *Frontiers in Microbiology* 8. <https://doi.org/10.3389/fmicb.2017.01268>.
- Liang S, Liu H, Wu S, Xu S, Jin D, Faiola F, Zhuang X, Zhang G, Fan H, and Bai Z (2019) Genetic diversity of diazotrophs and total bacteria in the phyllosphere of *Pyrus serotina*, *Prunus armeniaca*, *Prunus avium*, and *Vitis vinifera*. *Canadian Journal of Microbiology* 65: 1–11.
- Marques ACR, de Oliveira LB, Nicoloso FT, Jacques RJS, Giacomini SJ, and de Quadros FLF (2017) Biological nitrogen fixation in C-4 grasses of different growth strategies of South America natural grasslands. *Applied Soil Ecology* 113: 54–62. <https://doi.org/10.1016/j.apsoil.2017.01.011>.
- Miao L, Feng W, Zhang YQ, Bai YX, Sun YF, She WW, Mao HN, Lai ZR, and Qin SG (2020) Chemoheterotrophic diazotrophs contribute to nitrogen incorporation in a semi-arid desert. *Biology and Fertility of Soils* 56(8): 1165–1176. <https://doi.org/10.1007/s00374-020-01492-7>.
- Minamisawa K, et al. (2005) Biological nitrogen fixation, sustainable agriculture and the environment. In: Wang YP, Lin M, Tian ZX, Elmerich C, and Newton WE (eds.) *Current Plant Science and Biotechnology in Agriculture*. 41: Dordrecht: Springer. https://doi.org/10.1007/1-4020-3570-5_83.
- Mirza BS, Potisap C, Nusslein K, Bohannan BJM, and Rodrigues JLM (2014) Response of free-living nitrogen-fixing microorganisms to land use change in the Amazon rainforest. *Applied and Environmental Microbiology* 80(1): 281–288. <https://doi.org/10.1128/aem.02362-13>.
- Moyes AB, Kueppers LM, Pett-Ridge J, Carper DL, Vandehey N, O'Neil J, and Frank AC (2016) Evidence for foliar endophytic nitrogen fixation in a widely distributed subalpine conifer. *New Phytologist* 210(2): 657–668. <https://doi.org/10.1111/nph.13850>.
- Ohkuma M, Noda S, and Kudo T (1999) Phylogenetic diversity of nitrogen fixation genes in the symbiotic microbial community in the gut of diverse termites. *Applied and Environmental Microbiology* 65(11): 4926–4934.

- Ouyang Y, Evans SE, Friesen ML, and Tiemann LK (2018) Effect of nitrogen fertilization on the abundance of nitrogen cycling genes in agricultural soils: A meta-analysis of field studies. *Soil Biology and Biochemistry* 127: 71–78. <https://doi.org/10.1016/j.soilbio.2018.08.024>.
- Pepe-Ranney C, Koechli C, Potrafka R, Andam C, Eggleston E, Garcia-Pichel F, and Buckley DH (2016) Non-cyanobacterial diazotrophs mediate dinitrogen fixation in biological soil crusts during early crust formation. *The ISME Journal* 10(2): 287–298. <https://doi.org/10.1038/ismej.2015.106>.
- Puppo A and Rigaud J (1986) Superoxide-dismutase - an essential role in the protection of the nitrogen-fixation process. *FEBS Letters* 2: 187–189.
- Reed SC, Cleveland CC, and Townsend AR (2011) Functional ecology of free-living nitrogen fixation: A contemporary perspective. In: Futuyma DJ, Shaffer HB, and Simberloff D (eds.) *Annual Review of Ecology, Evolution, and Systematics, Annual Review of Ecology Evolution and Systematics*, vol. 42, 489–512. <https://doi.org/10.1146/annurev-ecolsys-102710-145034>.
- Roger PA and Ladha JK (1992) Biological N₂ fixation in wetland rice fields—Estimation and contribution to nitrogen balance. *Plant and Soil* 141(1–2): 41–55. <https://doi.org/10.1007/bf00011309>.
- Roley SS (2021) *Diazotrophic Nitrogen Fixation in the Rhizosphere and Endosphere. Rhizosphere Biology: Interactions Between Microbes and Plants*. Singapore, Singapore: Springer Nature.
- Roley SS, Xue C, Hamilton SK, Tiedje JM, and Robertson GP (2019) Isotopic evidence for episodic nitrogen fixation in switchgrass (*Panicum virgatum* L.). *Soil Biology and Biochemistry* 129: 90–98. <https://doi.org/10.1016/j.soilbio.2018.11.006>.
- Roper MM and Gupta VSR (2016) Enhancing Non-symbiotic N₂ fixation in Agriculture. *The Open Agriculture Journal* 10: 7–27. <https://doi.org/10.2174/1874331501610010007>.
- Sapountzis P, de Verges J, Rousk K, Cilliers M, Vorster BJ, and Poulsen M (2016) Potential for nitrogen fixation in the fungus-growing termite symbiosis. *Frontiers in Microbiology* 7. <https://doi.org/10.3389/fmicb.2016.01993>.
- Soper FM, Taylor BN, Winbourne JB, Wong MCY, Dynarski KA, Reis CRG, Peoples MB, Cleveland CC, Reed SC, Menge DNL, and Perakis SS (2021) A roadmap for sampling and scaling biological nitrogen fixation in terrestrial ecosystems. *Methods in Ecology and Evolution* 12(6): 1122–1137. <https://doi.org/10.1111/2041-210x.13586>.
- Urquiaga S, Xavier RP, Morais RF, Batista RB, Schultz N, Leite JM, Maia Sá J, Barbosa KP, Resende AS, Alves BJR, and Boddey RM (2012) Evidence from field nitrogen balance and ¹⁵N natural abundance data for the contribution of biological N₂ fixation to Brazilian sugarcane varieties. *Plant and Soil* 356(1–2): 5–21. <https://doi.org/10.1007/s11104-011-1016-3>.
- Van Deynze A, Zamora P, Delaux P-M, Heitmann C, Jayaraman D, Rajasekar S, Graham D, Maeda J, Gibson D, Schwartz KD, Berry AM, Bhatnagar S, Jospin G, Darling A, Jeannotte R, Lopez J, Weimer BC, Eisen JA, Shapiro H-Y, AneÅ J-M, and Bennett AB (2018) Nitrogen fixation in a landrace of maize is supported by a mucilage-associated diazotrophic microbiota. *PLoS Biology* 16(8), e2006352. <https://doi.org/10.1371/journal.pbio.2006352>.
- Wang C, Zhen MM, Hu AY, Zhu CQ, and Shen RF (2018) Diazotroph abundance and community composition in an acidic soil in response to aluminum-tolerant and aluminum-sensitive maize (*Zea mays* L.) cultivars under two nitrogen fertilizer forms. *Plant and Soil* 424: 463–478.
- Wu P, Zhang G, Ladha JK, McCouch SR, and Huang N (1995) Molecular marker facilitated investigation on the ability to stimulate N₂ fixation in the rhizosphere by irrigated rice plants. *Theoretical and Applied Genetics* 91(8): 1177–1183. <https://doi.org/10.1007/bf00220926>.

Phosphorus in soils—Biological interactions

Nataliya Bilyera and Iryna Loginova, Geo-Biosphere Interactions, Department of Geoscience, University of Tuebingen, Tuebingen, Germany

© 2023 Elsevier Ltd. All rights reserved.

This is an update of MD Mullen, Phosphorus in Soils I Biological Interactions, Editor(s): D Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 210–216, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00161-2>

Introduction	293
Forms and sources of soil phosphorus	294
Microbial necromass as a source of labile organic P	295
The phosphorus cycle	297
Mineralization and immobilization of P	298
Microbial solubilization of inorganic P	299
P-solubilizing microorganisms and plant growth	299
Phosphorus and the environment	300
Summary	301
References	301
Further readings	288

Abstract

This chapter describes the biological interactions of the phosphorus cycle. It introduces the mechanism of mineralization and solubilization as well as immobilization of both inorganic and organic soil phosphorus by microorganisms and their dependence on C:P stoichiometry. Along with the importance of the microbial biomass to making P available for plants, the contribution of microbial necromass to soil organic P is crucial. Viruses may be the one of the major drivers of bacterial mortality and necromass-P supply in soil. Further, this chapter discusses the effect of phosphorus solubilizing microorganisms on plant nutrient uptake and describes environmental aspects (specifically, P over-application) of the biological transformations of phosphorus.

Key points

- Soil microorganisms are responsible for organic P mineralization and immobilization of inorganic P into organically-bound forms.
- C:P ratio of the microbial biomass defines the extent to which P is either mineralized or immobilized.
- Both microbial biomass and necromass represent an important part of the organic P pool in soil.
- Phosphorus-solubilizing microorganisms potentially contribute to plant P nutrition and solubilize relatively insoluble inorganic P compounds (e.g. Al-, Fe-, and Ca-phosphates).
- Over-application of both inorganic and organic P-containing materials may result in potential offsite degradation of water quality.

Introduction

Phosphorus (P) is essential to life; it is a primary component of many important biomolecules such as adenosine triphosphate (ATP), nucleic acids (DNA and RNA), phospholipids, phosphopyridine nucleotides (NADP⁺), and phytin (P storage molecule—derivative of inositol hexaphosphate). Plant requirements for P generally account for between 0.2% and 0.4% of plant dry weight. Despite the relatively modest P requirements compared with nitrogen (N), the main challenge with P lies in maintaining enough soluble phosphate to prevent agricultural crops becoming deficient. Taking into account the quick fixation of applied P in the soil and the rapidly depleting stock of phosphate rocks, management of P in agriculture is of major interest (Cordell et al., 2009).

Plant-available phosphate ions, weathered from the primary mineral, apatite, are present in soil in varying concentrations. Phosphorus enters the soil environment through a number of other pathways, for example as anthropogenic P from incinerators that may be transported through the atmosphere and reach soils with precipitation (around 0.02–0.3 kg ha⁻¹ y⁻¹). Phosphorus is applied as inorganic fertilizers, as rock phosphate, or in organic materials such as animal manures, bone meal, seaweed, municipal and other biosolids, as recycled P fertilizers (e.g. struvite-P extracted from wastewaters), or, in undisturbed soil ecosystems, it is simply recycled from plant and animal matter.

Soil native microbial necromass is considered as an important source of labile organic P, becoming plant available after microbial death due to external factors including viral infection. Additionally, significant amounts of P further contribute to the formation of SOM.

The P cycle in soils includes both abiotic and biotic components. The abiotic component is primarily related to the dissolution of P-containing minerals, the formation of phosphates of Al, Fe, Ca, and Mg, and their sorption to clays, Al- and Fe-oxides, sesquioxides and organic compounds in soil. Abiotic and biotic components present the total P pool in the soil, which requires various transformations to become plant-available.

Biological transformations of P in the soil are crucial for recovering both residual inorganic and organic P (as up to 54% of residual P in agricultural soils may be stored in organic form) (Stutter et al., 2012). Microorganisms, nematodes and earthworms effectively increase availability of P in the soil. The biotic components of P cycle primarily involve the mineralization of organic P sources and immobilization of soil solution orthophosphates during microbial metabolism.

Soil microorganisms, including bacteria and fungi living in the bulk soil and the rhizosphere, as well as mycorrhizal fungi, produce a variety of organic acids and chelating agents which facilitate the solubilization of phosphate from the inorganic P and supply P to the plant. The differences among crop varieties in the intensity of residual inorganic and organic P utilization, through the release of phosphatases and organic acids as root exudates, can also be considered as a way to reduce the level of mineral fertilizer application. The capacity of soil to adsorb P is important for understanding the potential stock of residual phosphorus. Mismanagement or over application of P-containing materials can result in excessive amounts of inorganic P, so-called 'legacy P' (Haygarth et al., 2014), which can have both positive agronomic and negative environmental impacts. The 'legacy P' stock, accumulated over past decades, has the great potential to obviate the necessity for further agronomic P inputs. However, although the amount of P stored in soil might be sufficient to satisfy crop requirements to create high yields for many years, without any or with only minimal P input, much of it cannot be taken up by plants because its presence, in forms that are readily available, is limited. For the environment, the entry into surface waters of 'legacy P' and P applied to soil in manure or waste water can result in eutrophication, and further degradation of the quality of the habitat for aquatic organisms.

Forms and sources of soil phosphorus

The original source of most soil P is from the primary mineral apatite, or rock phosphate (RP). Apatite is a highly insoluble Ca-phosphate compound with the general formula of $\text{Ca}_{10}(\text{PO}_4)_6 \times 2$, where X is F^- (fluoroapatite), OH^- (hydroxyapatite), or Cl^- (chloroapatite). In acidic soils, little of these apatite minerals remain in the soil. The P in these minerals is released, either chemically or by biological interactions with microorganisms and plant roots. Over time, available P is taken up by plants or microorganisms and incorporated into organic matter, some is retained in relatively insoluble inorganic P minerals with Al and Fe in acidic soils, and with Ca in neutral to alkaline soils. In highly weathered soils, P is tied up by Al- and Fe-oxides. These compounds are only sparingly soluble (Fig. 1). Generally, solubility of phosphorus from Al and Fe sources increases with pH levels in the range of about pH 5.5–6.5. As the pH rises, however, Ca-phosphate minerals form, which control the concentration of soluble P. Hence, soluble P concentrations decrease significantly as pH increases above neutrality. The low solubility of inorganic soil P makes the use of P fertilizer materials necessary in many high-yielding agricultural systems. The practice of liming acid soils to increase pH to about 6.0–7.0 has the practical effect of enhancing P availability for plant growth.

The total amounts of P in natural soils vary widely, depending on parent material and the extent of weathering. In sandy soils with low organic matter content, there may be $\sim 100 \text{ mg kg}^{-1}$ of total P, while soils high in primary P minerals (e.g., high-phosphate limestone parent materials), clay content, or organic matter may contain well in excess of 2000 mg kg^{-1} soil. In most soils, inorganic P is the dominant form, although organic P increases with increasing soil organic matter. Cultivated soils may accumulate high levels of total P due to regularly P input, especially when large rates are applied.

The plant-available forms of P in soils are orthophosphate ions, H_2PO_4^- (dominant form at pH 2.2–7.2) or HPO_4^{2-} (dominant form at pH 7.2–12.4). Orthophosphate from soluble phosphate fertilizers or mineralized from organic sources rapidly precipitates as Al-, Fe-, or Ca-phosphates or is sorbed by Al- and Fe-oxides and soil organic matter. Due to these reactions, soluble P is present at very low concentrations in soil solution, typically in the range of less than $0.1\text{--}1 \text{ mg L}^{-1}$. In unfertilized soils the solution orthophosphates must be continuously replenished by chemical or biological dissolution of inorganic P compounds or by microbial mineralization of organic P.

Organic P commonly accounts for 30–50% of the total soil P in most soils, but it can vary from 3% to 90%. The P content of soil organic matter (SOM) ranges from about 1 to over 3%. Total organic P in a soil is a function of organic matter content, so that as organic matter increases, so does organic P. Organic P also decreases with soil depth, as does organic C.

The major types of organic P compounds, which are added to soil by recycling plant, animal, and microbial remains in the soil, are shown in Fig. 2. Under aerobic conditions, phytins or inositol phosphates, are usually found in the greatest quantities, making up to 40% of the total organic P content. The inositol phosphates are typically in a polymeric state, making these compounds relatively resistant to decomposition. Other organic P compounds include more labile phosphomonoesters (up to 43% of all organic P) and phosphodiester, such as DNA and RNA (making up 17%) (Darch et al., 2014; Spohn, 2020). Remaining forms of organic P, including phosphonates and organic polyphosphates, are present in trace amounts only and differ markedly in their bioavailability and transformation pathways in the soil.

Since inorganic forms of P are quite rapidly and strongly bound to or within secondary minerals, their availability to plants and leaching from soil are essentially restricted. Dissolved organic P forms can be more mobile within the soil, compared to inorganic P, due to weaker sorptive retention.

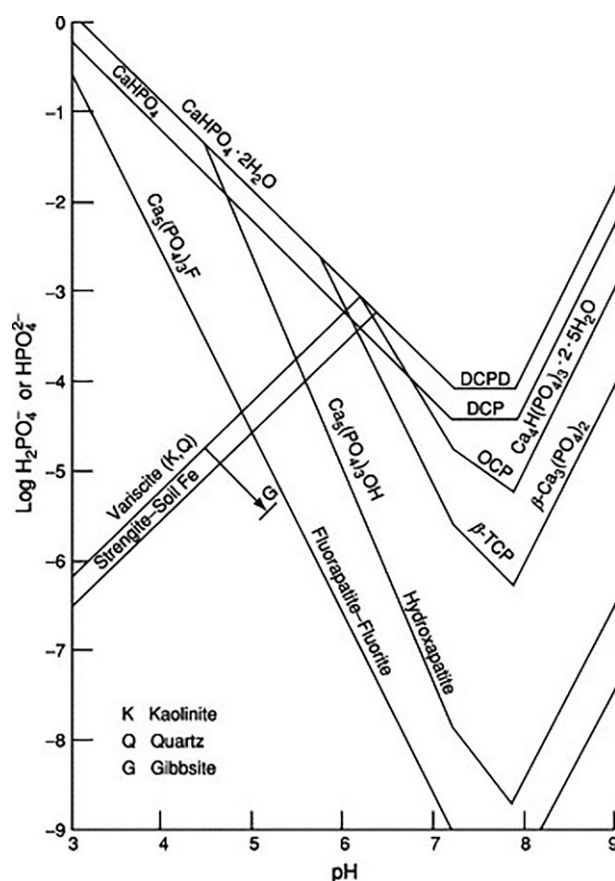


Fig. 1 Solubility of orthophosphate from variscite ($\text{AlPO}_4 \cdot 2\text{H}_2\text{O}$), strengite ($\text{FePO}_4 \cdot 2\text{H}_2\text{O}$), and various Ca-phosphates as a function of solution pH. These equilibria assume a Ca^{2+} concentration of $10^{-2.5} \text{ mol L}^{-1}$ and CO_2 at 0.03 kpa. DCPD, calcium phosphate dehydrate, brushite; DCP, calcium phosphate, monetite; OCP, octacalcium phosphate; β -TCP, β -tricalcium phosphate. Reproduced with permission from Lindsay WL (1979) *Chemical Equilibria in Soils*. New York: John Wiley.

Organic P compounds sorb strongly to positively charged mineral surfaces (e.g. amorphous Al and Fe oxides and the facets of clay minerals). This protects organic compounds against decomposition because sorption physically hinders enzymatic catalysis. As a result, organic P compounds may contribute to the sorptive stabilization of SOM against microbial decomposition in soil (Kleber et al., 2007).

The microbial biomass represents an important part of the organic P pool in soil. This pool represents actively cycling P in the soil, and is a part of the labile or readily available organic P. Microbial activity is responsible for mineralization and immobilization reactions that convert organic P into inorganic P and inorganic P into organically bound P. The microorganisms are also able to solubilize the relatively insoluble Al-, Fe-, and Ca-phosphates. In arable soils the biomass-P fraction accounts for up to 5% of the total organic P pool, while in undisturbed grassland and forest soils, the fraction may be as high as 40% (Turner et al., 2013). Soil disturbance, e.g., tillage, reduces organic matter and total organic P content, and in general the microbial biomass P fraction decreases more rapidly than the total organic P pool. Therefore, the labile organic P fraction is more susceptible to disturbance than the more stable, complex organic P forms. This indicates the importance of the microbial biomass to making P available for plant use in stable terrestrial ecosystems.

Microbial necromass as a source of labile organic P

In recent years there has been increasing evidence of a large contribution from microbial necromass to both SOM and soil organic P. Microbial necromass accumulation is an important way of sequestering C, making up as much as 30–60% of the topsoil SOM (Liang et al., 2019). By degrading plant-derived organic compounds to grow and reproduce, microorganisms immobilize P and it is stabilized as necromass in soil. Further transformations of this necromass include its mineralization and involvement in its turnover or stabilization (by aggregation and bonding to the mineral matrix of the soil).

As a source of available C and N, microbial necromass influences microbial P utilization. The global ratio of C:N:P for microbes is 60:7:1 (Cleveland and Liptzin, 2007). Such a narrow ratio compared to other sources of soil P makes microbial necromass an ideal source of nutrients, especially under P deficiency.

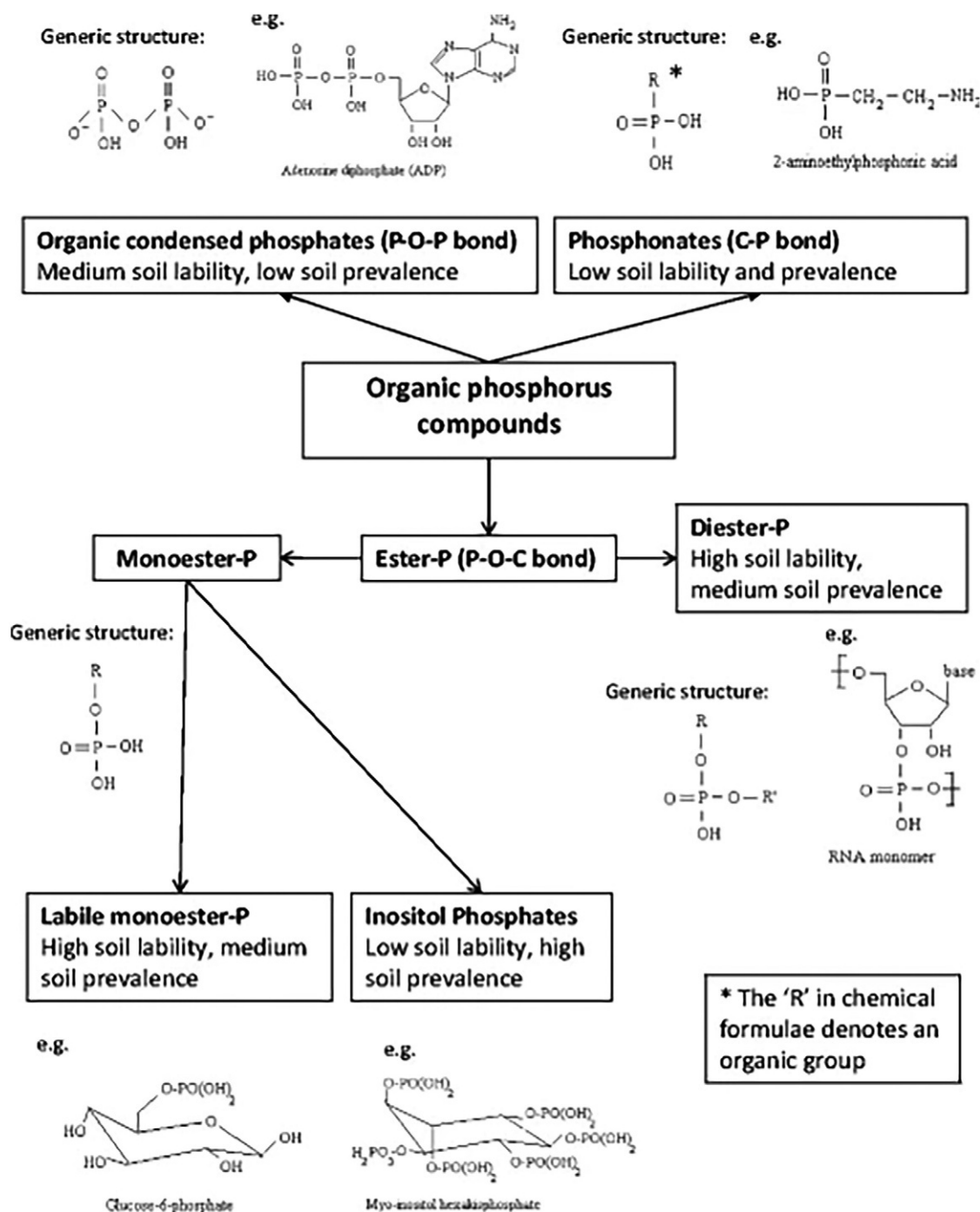


Fig. 2 Organic phosphorus forms with generic and example structures, information on the lability and prevalence (George et al., 2018).

Microbial processes in the soil that promote soil C turnover depend on the intensity of land use. It is suggested that the turnover of living microbial biomass is faster in croplands and grasslands compared with forest soils, and the contribution of fungal derived necromass C to SOM is consistently greater than bacterial necromass C in all soils (approx. 65–35%, respectively, with a higher ratio of fungal:bacterial necromass C for forest soils) as fungal necromass is preferentially protected in the soil and involved in the formation and stabilization of soil aggregates. At the same time, microbial biomass is preferentially utilized by bacteria, whereas fungi, especially ectomycorrhizal, are better adapted to utilize plant litter (López-Mondéjar et al., 2020).

Soil fauna and viruses are also involved in microbially-mediated C transformation processes and consequently influence microbial necromass production, recycling and SOC dynamics. Bacterial death is generally caused by external factors. Viruses

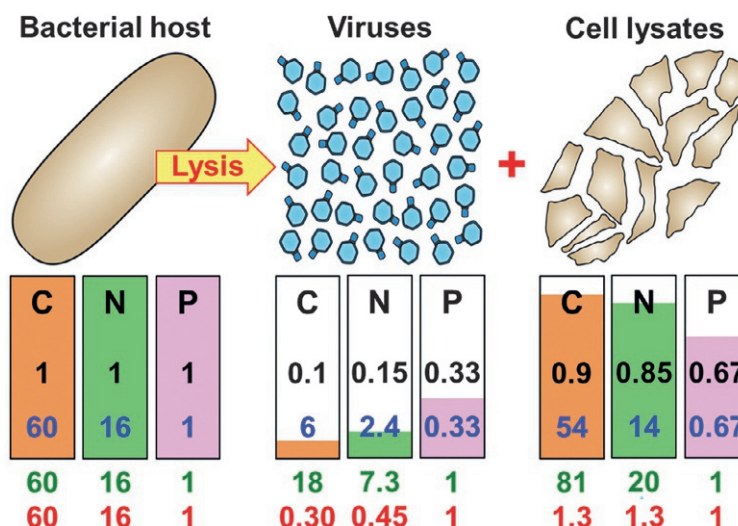


Fig. 3 Changes in carbon (C), nitrogen (N) and phosphorus (P) stoichiometry in initial bacterial host, viruses, and remaining bacterial lysates after viral lysis (relative to the initial content in bacterial host cells). The upper numbers (in black) show the C:N:P ratio of each component, the lower numbers (in blue) represent normalized data to the total amount of each element in the bacterial host (C:N:P = 1:1:1) (Kuzakov and Mason-Jones, 2018).

(i.e., bacteriophages) may be one of the major drivers of bacterial mortality in soil. Considering very high infection rates for bacteria and the short time to lysis, the average lifetime of bacteria in soil may be weeks, at most. It is suggested that viruses induce bacteria turnover in soil. Viral infection turns over the microbial population by cell lysis, thus releasing immobilized nutrients including P for root uptake (Kuzakov and Mason-Jones, 2018).

Unlike C and N, where the ratio in soluble cell lysates is very close to that of whole cells, host P is incorporated into nucleic acids of new virions, thus increasing the C:P ratio of the lysate and leading to a microscale divergence of C:N:P ratio in bacterial lysates after viral lysis (Fig. 3).

Higher P enrichment of viruses, compared with C and N, may lead to a strong depletion of P in the remaining lysates and it becoming temporarily inaccessible. However, although these concepts still need experimental verification, they are reasonable to consider in assessing P transformations in the soil.

The phosphorus cycle

The dynamics of P in the soil is governed by an interaction of chemical, physical and biological factors, which in turn are affected by environmental conditions and land use (Fig. 4). In some ways, the phosphorus cycle is not as complex as the nitrogen and sulfur cycles, because P typically does not undergo oxidation–reduction reactions or exist in gaseous forms. However, as for the N and S cycles, the biological processes of mineralization and immobilization as well as plant uptake are present. Additionally, the microbial fraction is active in the solubilization of relatively insoluble inorganic P materials, making P available for plant or microbial uptake or other transformations.

In the biological subcycle, soluble orthophosphates are either taken up by plants or immobilized into the microbial biomass. When organic sources of P, such as crop residues, manures and municipal biosolids, are land-applied, P may suffer three distinct fates. Organic forms, especially inositol phosphates, may be incorporated directly into stable SOM, becoming relatively unavailable for plant and microbial use. If the material has a relatively large P content relative to C and N, a portion of it will be mineralized to orthophosphate during microbial decomposition. Finally, part of the P will be incorporated into the microbial biomass (immobilized) during decomposition. This immobilized P may become available as soluble P when the microbial biomass fraction dies (e.g. forming a necromass) and is itself mineralized by other microorganisms, as described above.

Although this model of the P cycle segregates biological activity from geochemical activity, it should be noted that the dissolution of apatite and other secondary P minerals can be enhanced by the activities of microorganisms and plants. Dissolution or solubilization of these minerals releases P as orthophosphate into solution, where it can then be utilized by plants or microorganisms. Essential portions of this P will again precipitate as secondary P minerals or be adsorbed onto Al- and Fe-oxides or clay surfaces.

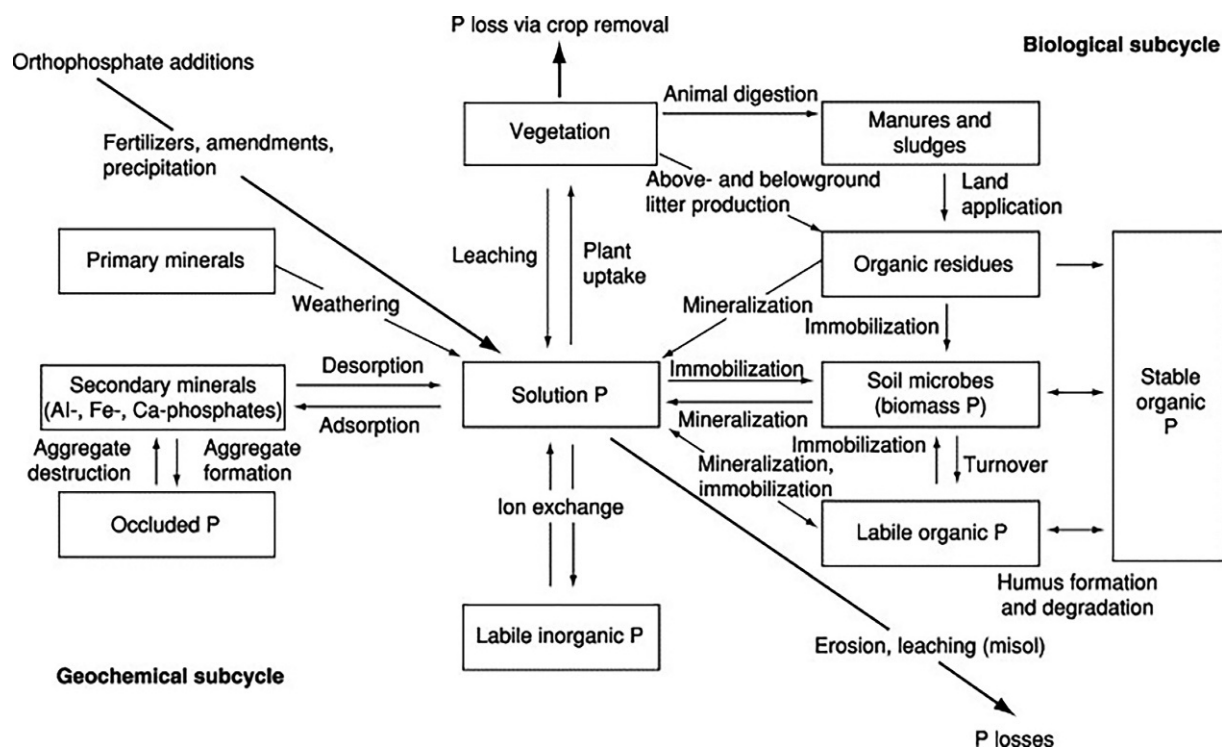


Fig. 4 The phosphorus cycle showing inputs, losses, and transformations. This model illustrates the relationship of the chemical and biological P subcycles through the solution orthophosphate pool. Adapted with permission from Walbridge MR (1991) Phosphorus availability in acid organic soils of the lower North Carolina coastal plain. *Ecology* 72: 2083–2100.

Mineralization and immobilization of P

As briefly discussed above, key components of the biological cycling of P in soil are mineralization (the transformation of organic P into inorganic orthophosphate P), and immobilization (the transformation of inorganic P back into organic forms). These reactions are catalyzed by those microorganisms that are active in the soil, and are largely the consequence of heterotrophic microbial catabolic and anabolic processes. To acquire energy, electron-rich organic compounds are oxidized and degraded to produce energy. As these compounds are decomposed, inorganic nutrients are released. To build new cells, the biomass must acquire inorganic P and other nutrients. The extent to which P is either mineralized or immobilized during these reactions depends on the C:P ratio of the substrate being degraded. If the C:P ratio is 200:1 or less, then the residue is rich in P and the excess will be mineralized. Conversely, if the C:P of the residue is greater than about 300:1 there will not be enough P available in the residue to facilitate complete degradation. Therefore, immobilization of orthophosphate from the soil into microbial biomass will occur if decomposition of the residue and biomass growth is to continue (Bilyera et al., 2021).

Although immobilization decreases the amount of orthophosphate immediately available to plants, it is typically a very short-lived phenomenon and may protect orthophosphate from being tied up in relatively insoluble inorganic compounds. After organic amendments are mineralized, the microbial biomass will begin to die back to previous levels, resulting in the subsequent mineralization of the microbial P pool.

Mineralization is simply the enzymatic release of orthophosphate or another organic P intermediates (i.e., diesters → monoesters → orthophosphate), during the decomposition of organic materials that contain P. The phosphatase enzyme group is extracellular and responsible for the hydrolysis of esters and anhydrides of orthophosphate-P. A number of different phosphatase enzymes are observed in the soil environment. The most commonly measured soil phosphatases are the phosphomonoesterases, which hydrolyze phosphomonoester bonds. Two phosphomonoesterases have been widely studied: acid phosphatase and alkaline phosphatase. Acid phosphatase is produced by microorganisms and plants and predominates in acid soils, while alkaline phosphatase is only produced by soil microorganisms and is dominant in alkaline soils (although it is present in acid soils as well). Phytase is also an important enzyme as phytate (i.e., inositol phosphates, Fig. 3) can make up as much as 40% of the organic P in a soil. Therefore, phytate hydrolysis can be an important source of orthophosphate, supporting plant nutrition. These extracellular enzymes, actively excreted into soils by organisms, can also accumulate in soil, protected by sorption to organo-clay complexes in the soil. These enzymes are important, not only for the degradation of freshly added organic residues, but also for the release of inorganic P from soil organic matter. Activity of phosphatases (phosphomonoesterase) is generally inhibited in plant roots and rhizosphere soil after application of inorganic P and stimulated under P deficient conditions or at organic P application (Janes-Bassett et al., 2022).

Microbial solubilization of inorganic P

Microorganisms are also important in the solubilization of P from mineral sources. Orthophosphate is typically present in soil solution at very low concentrations. This solution P is replenished either from mineralization of organic P sources, as discussed above, or from the dissolution of P from phosphate minerals, which in soils is primarily apatite. In acidic soils, secondary inorganic P minerals in the form of aluminum and iron phosphates, such as variscite ($\text{AlPO}_4 \cdot 2\text{H}_2\text{O}$) and strengite ($\text{FePO}_4 \cdot 2\text{H}_2\text{O}$), dominate. In alkaline soils, it is calcium phosphates that dominate (e.g., $\text{Ca}_3(\text{PO}_4)_2$).

Most forms of inorganic P in soil have limited solubility. It is generally agreed that microorganisms can facilitate the enhanced dissolution of these compounds by at least two mechanisms: production of organic acids and excretion of protons during NH_4^+ assimilation, resulting in chelation of cations or reduction in pH to release P. These mechanisms are performed by soil fungi and bacteria and appear to be potentially important, especially in soils receiving little or no phosphatic fertilizer materials.

Phosphate solubilization has been observed in many different species of bacteria, including actinomycetes, fungi, and algae, collectively labeled as Phosphorus Solubilizing Microorganisms (PSM). Bacterial species reported include those in the genera *Bacillus*, *Pseudomonas*, *Agrobacterium* and *Micrococcus*, *Nitrosomonas*, and others. The most commonly reported P-solubilizing fungi are *Aspergillus*, *Azospirillum* and *Penicillium* species, although a number of other species have been reported to solubilize P. In soil, P-solubilization activities are greatest in the rhizosphere. The plant root system provides ample quantities of readily utilized carbon, which results in high microbial activity. The activity of P-solubilizing organisms in the rhizosphere can be potentially important for the P nutrition of plants. In general, it appears that fungi may be more important to the solubilization of inorganic phosphate in soil than bacteria. Fungal hyphae can grow across distance in soil microenvironments, allowing for exploitation of P resources not available to relatively immobile bacteria. However, according to so-called 'biological markets' the amount of P provided by fungus depends on the amount of C received from a partner plant (van't Padje et al., 2021).

Organic acids produced by bacteria and fungi have been shown to solubilize various forms of inorganic phosphates (Alori et al., 2017). Citric, and other organic acids, can be quite effective in solubilizing largely insoluble forms of P, such as AlPO_4 . The efficacy of organic acids as P-solubilizing agents in liquid media is due to their effect in lowering the pH, which facilitates acid hydrolysis of AlPO_4 , and the released Al^{3+} . The latter is then complexed with the organic anion. This chelation or complexation of the released Al helps to minimize re-precipitation of Al-phosphates and the potential for Al-toxicity on acid soils. While organic acids appear to be effective in solubilizing Al-phosphates, solubilization is typically more effective in the presence of microbial cells than that observed with organic acids alone. In addition, some organisms solubilize Al-phosphates with no appreciable production of organic acids. Proton accumulation during the assimilation of NH_4^+ or due to respiratory activities appears to be a primary mechanism in the absence of organic acids (Alori et al., 2017).

Dissolution of Ca-phosphates, such as rock phosphate or $\text{Ca}_3(\text{PO}_4)_2$, also occurs due to organic acid production or the accumulation of protons during ammonium assimilation or respiration. Typically, a decrease in the pH of the growth substrate occurs during solubilization of P from these minerals. The pH decrease is most pronounced, and solubilization is usually highest, in the presence of ammonium N sources. While organic acids can be effective in solubilization of Ca-phosphates, they are not always produced by all organisms capable of the process.

P-solubilizing microorganisms and plant growth

Inoculation of crops with PSM has produced controversial results. In the 1950s in the former Soviet Union, researchers reported excellent results when spores of *Bacillus megatherium* var. *phosphaticum* (also referred to as phosphobacterin) were applied to soils that were neutral to alkaline in pH and with a high organic matter content. The mode of action was thought to be the mineralization of organic P. Subsequent research, mainly in the United States, was not able to verify the claims of significant crop yields and the practice has never been adopted. Recently, the pH decrease has been reported as the most probable mechanism. Phosphorus solubilizing bacteria are acidifying microorganisms, which mobilize Ca bounded phosphate via proton secretion. Other mechanisms of P solubilization include the production of organic anions, which can either complex P by ligand-promoted mechanisms, or they can directly desorb orthophosphate P from soil minerals. The production of a wide range of phosphatase enzymes by microorganisms can also mobilize orthophosphate from organic P forms in soil.

The effects of PSM are variable under controlled and field conditions. Under controlled conditions, inoculations with PSM improved plant growth and P shoot content, especially on low P soils. In some studies under field conditions, there was little or no effect, whereas other research has shown positive effects of inoculation on plant yields and P uptake. The effectiveness of PSM decrease under regions of increased temperatures and smaller precipitation. Another reason for the lack of field PSM performance on crop productivity and P uptake may be related to poor soil colonization due to competition with other native microbial communities. Expectedly, soil pH is also important for the efficacy of PSM to solubilize soil P. Generally, inoculation is more effective at higher soil pH (i.e. calcareous soils), where acidification is considered to be the main cause of increased P availability.

Availability of P in soil is also regarded as a necessary requirement for PSM performance. To achieve positive plant response to PSM, some threshold P availability should be met. A meta-analysis revealed that inoculation resulted in a greater yield increase under conditions with moderate P availability in soil, but was less effective in more P-deficient environments (Schütz et al., 2018).

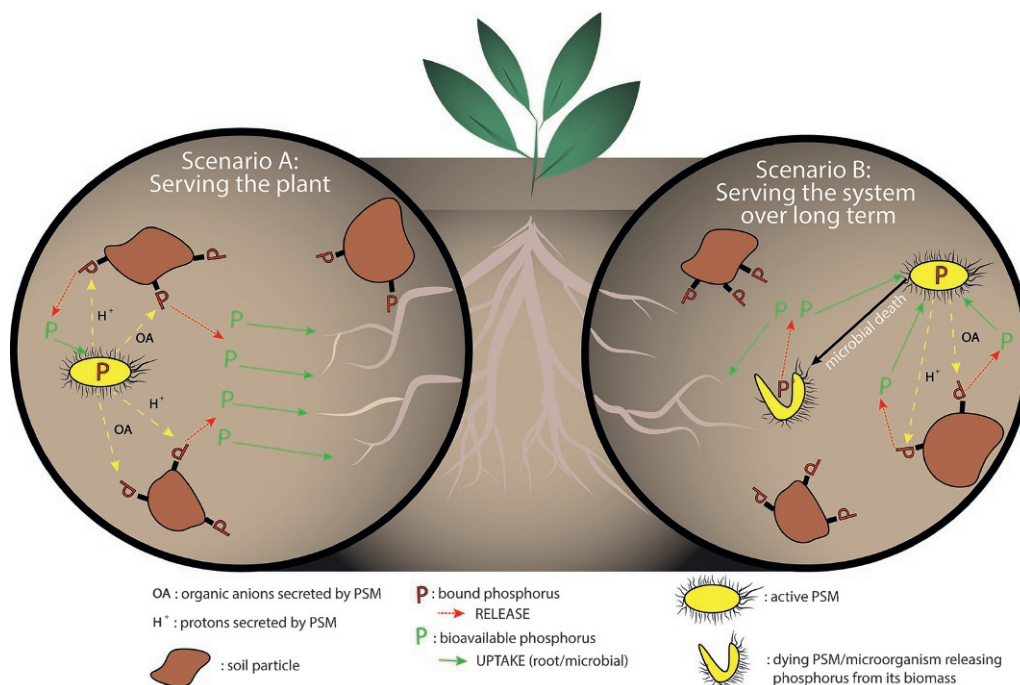


Fig. 5 Overview of the two conceptual pathways for phosphate-solubilizing microorganism(s) (PSM). Scenario A represents the ideal PSM contribution to plant P nutrition in which excess P is solubilized in soil to meet plant needs considering adequate nutrient supply (i.e. carbon and nitrogen). Scenario B illustrates PSM solubilizing P to sustain their own need and serve the system over the long term by releasing P from their biomass turnover (Raymond et al., 2020).

Ideally, microorganisms contribute directly to plant P nutrition solubilizing P from soil to meet plant needs, as shown in Scenario A (Fig. 5). Another possible pathway (Scenario B, Fig. 5) is to facilitate microbial biomass turnover to serve over a long term (Raymond et al., 2020). Microbial P mobilization may be enhanced by a general decrease in microbial biomass, which depends on C and N availability and their ratio in soils. In a P-rich environment, soil microorganisms will use available P, whereas in a P-limited environment it will drive P-solubilization activity within the microbial biomass.

Although, other factors than the putative P availability alone may contribute to field performance of PSM. Among them are stimulating effects on plant growth and biological nitrogen fixation, and availability of other essential nutrients (e.g., Zn and Fe).

Phosphorus and the environment

Concern over the impact of P on the environment has re-emerged over the past decade. Over application of P-containing materials, whether inorganic fertilizers or organic amendments, can result in high levels of extractable soil inorganic phosphate. If these soils are close to surface waters, then this P can be exported to those waters either as soluble P in runoff water or as P adsorbed to eroded soil particles. In both cases, this can result in the increase of bioavailable P in these waters. This, in turn, can potentially lead to eutrophication, which occurs when excess nutrients, typically N or P, accumulate in the water, resulting in accelerated growth of algae. These algae can cloud the water, decreasing light penetration, resulting in decreased growth of aquatic plants. Increased algal biomass production stimulates higher rates of degradation in the water column by bacteria. This accelerated decomposition then decreases available oxygen, often making the habitat unfit for many forms of aquatic life.

Many soils that have received regular applications of manure over time have accumulated large values of total and available P. If manure is applied to a crop to satisfy N requirements, too much P is typically applied, resulting in a P buildup. Although much of the P in these solid manures is added in the organic form, the result is typically a significant increase in the total amount of inorganic phosphate due to hydrolysis by phosphatase enzymes in both soil and water. The last fact highlights the importance of organic P for eutrophication and the necessity to focus on organic P in addition to inorganic when considering P transfer from agricultural lands.

Microbial mineralization of manure releases the phosphate ion, with a significant proportion remaining relatively labile. This lability may be due to the concomitant increase in soluble soil C that may then competitively bind to P-sorption sites or with soil Al and Ca, reducing P complexation potential. Microbial mineralization of these organic P sources over long periods can therefore result in potential degradation of the quality of nearby water bodies.

Modern environmental initiatives to decrease the CO₂ concentration in the atmosphere, by promoting land management practices that sequester organic C in soils, have received both approval and criticism. One of the concerns is simultaneous sequestration of organic P along with carbon, especially in mineral soils. More P input will be needed to sequester C and such co-sequestered P will be stored in soil, but not available to plants, if SOM management aims to keep increased SOM levels (Spohn, 2020).

Summary

Despite of relatively modest P requirements, phosphorus deficiency remains the second largest concern for the world agriculture after nitrogen. Considering the depletion of mined phosphate rocks and the quick fixation of applied P in soil, understanding P mineralization and immobilization by soil microorganisms is essential. Mineralization occurs if the C:P ratio is 200:1 or less, when the residue is rich in P. Mineralization is simply the enzymatic release of orthophosphate or another organic P intermediates. At C: P > 300:1 there will not be enough P available in the residue to facilitate complete degradation and immobilization occurs. Furthermore, viruses are one of the major drivers of mortality among bacteria in soil. As viral infection drives cell lysis this accelerates the turnover of the microbial biomass, in the course of which immobilized nutrients including P may be released and made available for uptake by plants. Microbial necromass consisting of the residual mixtures of intracellular and extracellular biomolecules may be an important source of labile organic phosphorus. The global ratio of C:N:P for microbes (60:7:1) is narrow compared to other sources of soil P making microbial necromass an ideal source of nutrients, especially under P deficiency. Phosphorus solubilizing microorganisms might, under the right conditions, facilitate P nutrition and thus enhance plant growth. Normally, P solubilization occurs mainly in the rhizosphere, which is both a microbial hotspot and a zone for P uptake. The release of organic acids produced by bacteria and fungi, as well as decrease in pH are the most probable mechanisms to solubilize various forms of inorganic phosphates. Over application of both inorganic and organic P-containing materials may result in potential offsite degradation of water quality. Environmental initiatives like C sequestration may lead to co-sequestered P in soil, which causes increased requirements for P input.

References

- Alori ET, Glick BR, and Babalola OO (2017) Microbial phosphorus solubilization and its potential for use in sustainable agriculture. *Frontiers in Microbiology* 8. <https://doi.org/10.3389/fmicb.2017.00971>.
- Bilyera N, et al. (2021) Maize genotype-specific exudation strategies: An adaptive mechanism to increase microbial activity in the rhizosphere. *Soil Biology and Biochemistry* 162. <https://doi.org/10.1016/j.soilbio.2021.108426>.
- Cleveland CC and Liptzin D (2007) C:N:P stoichiometry in soil: Is there a “Redfield ratio” for the microbial biomass? *Biogeochemistry* 85(3): 235–252. <https://doi.org/10.1007/s10533-007-9132-0>.
- Cordell D, Drangert JO, and White S (2009) The story of phosphorus: Global food security and food for thought. *Global Environmental Change* 19(2): 292–305. <https://doi.org/10.1016/j.gloenvcha.2008.10.009>.
- Darch T, et al. (2014) A meta-analysis of organic and inorganic phosphorus in organic fertilizers, soils, and water: Implications for water quality. *Critical Reviews in Environmental Science and Technology* 44(19): 2172–2202. <https://doi.org/10.1080/10643389.2013.790752>.
- George TS, et al. (2018) Organic phosphorus in the terrestrial environment: A perspective on the state of the art and future priorities. *Plant and Soil* 427(1–2): 191–208. <https://doi.org/10.1007/s11104-017-3391-x>.
- Haygarth PM, et al. (2014) Sustainable phosphorus management and the need for a long-term perspective: The legacy hypothesis. *Environmental Science & Technology* 48(15): 8417–8419. <https://doi.org/10.1021/es502852s>.
- Janes-Bassett V, et al. (2022) A meta-analysis of phosphatase activity in agricultural settings in response to phosphorus deficiency. *Soil Biology and Biochemistry* 165: 108537. <https://doi.org/10.1016/j.soilbio.2021.108537>.
- Kleber M, Sollins P, and Sutton R (2007) A conceptual model of organo-mineral interactions in soils: self-assembly of organic molecular fragments into zonal structures on mineral surfaces. *Biogeochemistry* 85(1): 9–24. <https://doi.org/10.1007/s10533-007-9103-5>.
- Kuzuyakov Y and Mason-Jones K (2018) Viruses in soil: Nano-scale undead drivers of microbial life, biogeochemical turnover and ecosystem functions. *Soil Biology and Biochemistry* 127: 305–317. <https://doi.org/10.1016/j.soilbio.2018.09.032>.
- Liang C, et al. (2019) Quantitative assessment of microbial necromass contribution to soil organic matter. *Global Change Biology* 25(11): 3578–3590. <https://doi.org/10.1111/gcb.14781>.
- López-Mondéjar R, et al. (2020) Metagenomics and stable isotope probing reveal the complementary contribution of fungal and bacterial communities in the recycling of dead biomass in forest soil. *Soil Biology and Biochemistry* 148: 107875. <https://doi.org/10.1016/j.soilbio.2020.107875>.
- Raymond NS, et al. (2020) Phosphate-solubilising microorganisms for improved crop productivity: A critical assessment. *New Phytologist*. nph.16924. <https://doi.org/10.1111/nph.16924>.
- Schütz L, et al. (2018) Improving crop yield and nutrient use efficiency via biofertilization—A global meta-analysis. *Frontiers in Plant Science* 8. <https://doi.org/10.3389/fpls.2017.02204>.
- Spohn M (2020) Increasing the organic carbon stocks in mineral soils sequesters large amounts of phosphorus. *Global Change Biology* 26(8): 4169–4177. <https://doi.org/10.1111/gcb.15154>.
- Stutter MI, et al. (2012) Recovering phosphorus from soil: A root solution? *Environmental Science & Technology* 46(4): 1977–1978. <https://doi.org/10.1021/es2044745>.
- Turner BL, et al. (2013) Soil microbial biomass and the fate of phosphorus during long-term ecosystem development. *Plant and Soil* 367(1–2). <https://doi.org/10.1007/s11104-012-1493-z>.
- van't Padje A, Werner GDA, and Kiers ET (2021) Mycorrhizal fungi control phosphorus value in trade symbiosis with host roots when exposed to abrupt “crashes” and “booms” of resource availability. *New Phytologist* 229(5): 2933–2944. <https://doi.org/10.1111/nph.17055>.

Further readings

- Bai X, Dippold MA, An S, Wang B, Zhang H, and Loeppmann S (2021) Extracellular enzyme activity and stoichiometry: The effect of soil microbial element limitation during leaf litter decomposition. *Ecological Indicators* 121: 107200. <https://doi.org/10.1016/j.ecolind.2020.107200>.
- Bilyera N, Dippold MA, Bleicher J, Maranguit D, Kuzuyakov Y, and Blagodatskaya E (2021) Microbial tradeoffs in internal and external use of resources regulated by phosphorus and carbon availability. *European Journal of Soil Biology* 106: 103353. <https://doi.org/10.1016/j.ejsobi.2021.103353>.
- Bünemann EK (2015) Assessment of gross and net mineralization rates of soil organic phosphorus—A review. *Soil Biology and Biochemistry* 89: 82–98. <https://doi.org/10.1016/j.soilbio.2015.06.026>.

- Jones DL and Oburger E (2011) Solubilization of phosphorus by soil microorganisms. In: Bünemann E, Oberson A, and Frossard E (eds.) *Phosphorus in Action, Soil Biology*, pp. 215–243. Berlin, Heidelberg: Springer. <https://doi.org/10.1007/978-3-642-15271-9>.
- Koester M, Stock SC, Nájera F, Abdallah K, Gorbushina A, Prietzel J, Matus F, Klysubun W, Boy J, Kuzyakov Y, Dippold MA, and Spielvogel S (2021) From rock eating to vegetarian ecosystems—Disentangling processes of phosphorus acquisition across biomes. *Geoderma* 388: 114827. <https://doi.org/10.1016/j.geoderma.2020.114827>.
- Le Bayon RC and Milleret R (2009) Effects of earthworms on phosphorus dynamics—a review invited mini-review dynamic soil, dynamic plant effects of earthworms on phosphorus. *Dynamic Soil, Dynamic Plant* 3 (special issue 2): 21–27. Global Science Books. URL: <https://www.researchgate.net/publication/40783737>
- Oburger E, Jones DL, and Wenzel WW (2011) Phosphorus saturation and pH differentially regulate the efficiency of organic acid anion-mediated P solubilization mechanisms in soil. *Plant and Soil* 341: 363–382. <https://doi.org/10.1007/s11104-010-0650-5>.
- Richardson AE and Simpson RJ (2011) Soil microorganisms mediating phosphorus availability. *Plant Physiology* 156(3): 989–996. <https://doi.org/10.1104/pp.111.175448>.
- Turner BL, Richardson AE, and Mullaney EJ (eds.) (2007) *Inositol Phosphates: Linking Agriculture and the Environment*. Wallingford, UK: CABI.

Biological transformations of sulfur in soil

Chaoqun Wang^a and Maxim Dorodnikov^{b,c}, ^aBiogeochemistry of Agroecosystems, University of Goettingen, Goettingen, Germany; ^bDepartment of Soil Science of Temperate Ecosystems, University of Goettingen, Goettingen, Germany; ^cInstitute of Landscape Ecology, University of Muenster, Muenster, Germany

© 2023 Elsevier Ltd. All rights reserved.

This is an update of S.D. Siciliano, J.J. Germida, SULFUR IN SOILS | Biological Transformations, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 85–90, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00462-8>.

Introduction	303
The global sulfur cycle	304
The biological availability of sulfur	304
Mineralization	304
Assimilation	305
Sulfur oxidation	307
Disproportionation of sulfur	307
Reduction of sulfur	308
Sulfur transformations and environmental quality	308
Conclusion	309
References	309

Abstract

The element sulfur (S) plays an essential role in the development and proper cellular functions of all organisms. Biological S transformations are key processes to provide available S for organisms. These transformations consist of multiple pathways such as oxidation, reduction, assimilation, and mineralization of various sulfur compounds by microorganisms. This chapter summarizes the current knowledge about the main pathways of the biological S transformations in the soil. The significance of oxidation, reduction, assimilation, and mineralization of S is discussed in relation to soil characteristics and environmental factors in terrestrial and aquatic ecosystems.

Key points

- Sulfur cycling is driven by the processes of oxidation, reduction, assimilation, and mineralization.
- Microbial cell- and extracellular enzymatic reactions catalyze sulfur mineralization from organic matter.
- Assimilation converts inorganic sulfur into organic sulfur.
- Some microorganisms gain energy from sulfur oxidation.
- Organisms perform sulfur disproportion by using sulfur as electron donor and acceptor.
- Assimilation and dissimilation trigger sulfur reduction coupled with the oxidation of hydrogen or organic compounds.

Introduction

In soil, sulfur (S) exists in a variety of forms, some of which are beneficial to the environment, and other forms that are typically considered pollutants. The wide array of oxidation states and forms of S results in a rich diversity of biological transformations involving this important element. These transformations are critical for many basic ecological processes such as plant growth, but also influence the movement of environmental contaminants.

The biological transformations of S can be divided into four major classes: (1) oxidation; (2) reduction; (3) assimilation; and (4) mineralization (Fig. 1). In addition, microbial metabolites can chemically fix or solubilize inorganic S compounds. The first two classes are related to the acquisition of energy for growth and the last two are typically related to nutrient acquisition. For energy acquisition, S can act as an electron acceptor, an electron donor, and sometimes both at the same time. The sulfate anion (SO_4^{2-}) acts as an electron acceptor and is reduced to S^{2-} by a group of bacteria known as sulfate-reducing bacteria. These bacteria grow in the absence of oxygen. In contrast, S^{2-} can act as an electron donor and is oxidized by a group of bacteria known as S oxidizers. These bacteria grow in the presence of oxygen. Interestingly, some bacteria can disproportionate S, i.e. intermediate-valence S species are simultaneously reduced and oxidized to form different products. In this reaction, one of the two S atoms, present in thiosulfate as S^{1-} , acts as an electron donor and the other S atom, present as S^{5+} , acts as an electron acceptor.

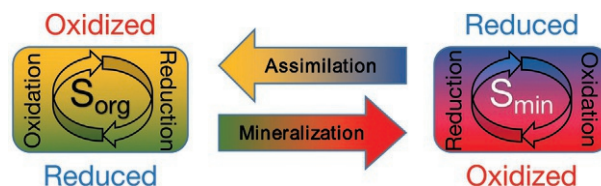


Fig. 1 The biological transformations of sulfur (S). Four main processes (oxidation, reduction, assimilation, mineralization) are shown. S_{org} : organic S forms, S_{min} : mineral S forms. Oxidized and reduced states of S correspond to both forms of S.

The last two classes of S transformation are directly linked to the nutritional status of the cell. Mineralization is a transformation in which organic S contained in a growth substrate is released by the cell in a mineral form such as sulfate. In contrast, assimilation is a process in which S present in the environment is incorporated into biomass. In the former case, the mineralized S is available for other organisms to use, whereas in the latter the assimilated S is no longer available to other organisms. Each of these transformations of S has important side-effects on the environment. The mineralization of S provides sulfate, which can potentially stimulate sulfate-reducing bacteria that play important roles in the arsenic and mercury biogeochemical cycles. The oxidation of reduced S by S oxidizers results in acid mine drainage, one of the primary detrimental impacts of mining.

The global sulfur cycle

The global cycle of S begins with S volatilized largely as dimethylsulfide from marine algae, marshlands, mud flats, plants, and soil entering the atmosphere. This dimethylsulfide is converted photo-chemically to methanesulfonic acid, which then deposits on to the Earth and is rapidly converted by bacteria to carbon dioxide and sulfate. This sulfate is assimilated by plants and immobilized into compounds like sulfoquinovosyl diacylglycerol, or alternatively the S is released back to the atmosphere as dimethylsulfide. As plants die, bacteria metabolize the S present in plant biomass and either immobilize this S into their biomass or release it as sulfate for uptake by plants. In addition to this natural S cycle, human activities have at least doubled the input of S to terrestrial systems. Approximately 1.5×10^{11} kg S per year is deposited due to atmospheric pollution, largely in the form of sulfuric acid (sulfate), with small amounts of sulfite and bisulfite also being deposited. This deposited sulfate can either be used by bacteria as a terminal electron acceptor or assimilated by organisms requiring S as a nutrient. Despite this recent disruption of the global S cycle, the S levels present in soil are largely due to pedogenic factors like climate, vegetation, and parent material. As a result, S levels in soil range from 0.002 to 10%: most of this S, 90%, is found in organic form.

The biological availability of sulfur

Although 90% of S in soil is present as organic S, much of this organic S is not available for assimilation or dissimilation by microorganisms and plants. Microorganisms need to convert the organic S in soil into sulfate before plants can take up the sulfate as a nutrient. This conversion of sulfate from organic S is the end result of a complicated biogeochemical cycle. Sulfur enters the soil typically with the residues of plants and animals and is converted and assimilated by microorganisms into their biomass. Predators then consume the prey bacteria and excess S is released into the soil solution. The S released into the soil solution is available for uptake by other organisms or plants. However, the predators do not mineralize all the S present in prey organisms and, as a result, a small fraction of S enters what is termed the 'resistant' pool of organic S present in soil (Fig. 2). This resistant pool of organic S is found as large polymers that are in close association with clays. These large polymers resist microbial attack, either because they are physically protected or because it is not energetically favorable for microbes to attack these polymers. There are three distinct pools of S in soil: (1) inorganic sulfate that is readily taken up by plants and microbes alike; (2) short-lived organic S compounds like taurine, cysteine, cystine, and methionine which are mineralized by microorganisms and either released as sulfate or incorporated into biomass; and (3) long-lived, resistant organic-S polymers found in soil humus can be stabilized through associations with soil clays. Microbial transformations are the primary process by which S is transferred between the pools of readily available and recalcitrant organic S. These transformations are discussed below.

Mineralization

Almost exclusively, microorganisms mediate S mineralization. By definition, mineralization is the metabolic conversion of an organic form of an element to an inorganic form. Mineralization processes can be divided into cell-mediated or extracellular enzymatic processes. Cell-mediated mineralization occurs by oxidative decomposition under aerobic conditions or by desulfurization under anaerobic conditions. In cell-mediated mineralization, S present in carbon-containing compounds is mineralized as organisms consume the carbon. Thus, if there is more S in the substrate than organism requires, the S will be mineralized and released to the environment. If there is not enough S, then any released S is consumed and immobilized by the organism. The

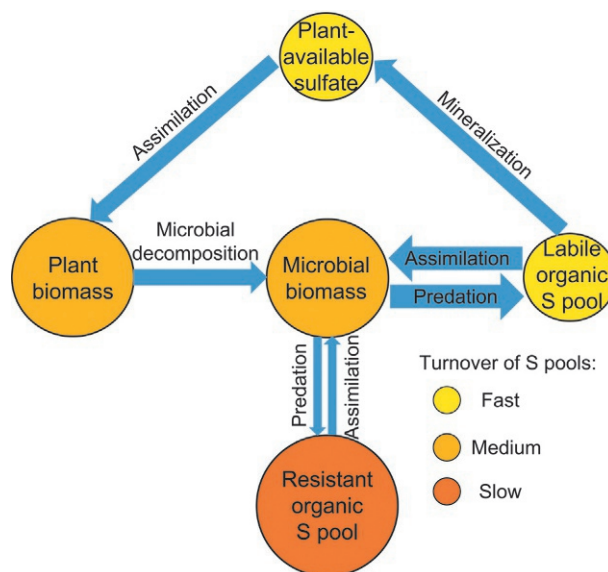


Fig. 2 Different sulfur (S) pools present in soil and their relationship to plant-available S. Size of circles reflects the relative size of respective S pools. Color of circles shows a relative turnover of S pools from fast (yellow) to slow (brown). Arrows and their sizes demonstrate the direction and intensity of S fluxes between pools, respectively.

break-even point for mineralization can be calculated based on the C:S ratio of the substrate, the C:S ratio of the organism, and the yield coefficient, i.e., how much of the consumed C is incorporated into biomass. As a general rule of thumb, if the C:S ratio of a substrate is 200 or less, then S is released. If the C:S ratio is greater than 400, then S is not going to be released because the microorganisms require the S for their own metabolism. In this latter case, organisms rely on a second, enzymatic method of mineralization to obtain the S essential for growth.

Enzymatic mineralization involves extracellular enzymes, such as arylsulfatases, released by the microorganism. These enzymes, once released from cells, will hydrolyze sulfate esters present in the soil. This process releases sulfate into the soil solution for use by the cells. The activity of these enzymes depends on a wide range of soil factors with reaction rates varying by a factor of 4 between different soils. These enzymes attack the vast amount of S present in the soil as organically bound S. Organic S in soil constitutes more than 90% of the total S present in soils. This organic S can be grouped into carbon-bonded S and organic sulfates. Organic sulfates, which comprise between 30 and 75% of S in soil, typically include compounds such as sulfate esters (C—O—S), sulfamates (C—N—S), and sulfated thioglycosides (N—O—S)- the forms of sulfate that enzymatic mineralization processes attack.

Mineralization processes are biological transformations of S that are intimately linked to cellular growth. Thus, factors that influence microbial activity, such as moisture, temperature, or plant growth, will influence mineralization. Increasing the growth of microorganisms by either adding a carbon source or plant growth typically results in reduced mineralization.

Assimilation

Assimilation is the process of converting inorganic S into organic S present in an organism. This assimilatory process is intimately linked to mineralization. In fact, the two processes, assimilation and mineralization, occur simultaneously. If the net effect is to release inorganic sulfate then this is termed mineralization; if instead the net effect is for the S to be incorporated into the biomass, then this is termed assimilation. Thus, the first step of many assimilation processes is the conversion of S into a form that can be incorporated into biomass. Taken alone, such steps might be considered as mineralization processes but we term them assimilatory processes if: (1) the sum of the steps is to incorporate S into biomass and (2) they are induced by S limitation, suggesting that the organism is using the process to obtain S. There are different biological assimilatory transformations that can occur depending on what form the available S is in. The assimilation of sulfate can be summarized in four steps (Droux, 2004): (1) uptake of sulfate; (2) activation of sulfate; (3) reduction of sulfate; (4) synthesis of L-cysteine (Cys). For inorganic SO_4^{2-} , an enzyme called adenosine triphosphate (ATP) sulfurylase mediates the reaction between sulfate and ATP and sulfate to form adenosine-5'-phosphosulfate (APS) (Fig. 3, top side). In turn, this APS is transformed by APS kinase with another ATP molecule to form adenosine-3'-phosphate-5'-phosphosulfate (PAPS). This high-energy molecule reacts with PAPS reductase and two sulphydryl groups to form sulfite, and sulfite is reduced by NADPH_2 to form sulfide which is incorporated into L-cysteine by O-acetylserine sulphydrylase. This pathway will efficiently scavenge excess sulfate present in the soil solution that organisms require. So much so, that in the presence of an energy source, S added as SO_4^{2-} is quickly incorporated into the organic fraction of the soil biomass and later found in the fulvic acid fractions. However, sulfate is not commonly found in soil solution and organisms have developed other pathways to assimilate S for their metabolic needs.

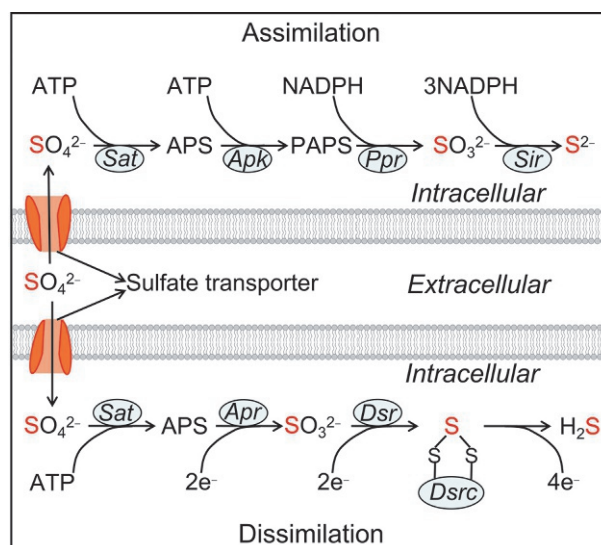


Fig. 3 Metabolic pathways of assimilation (top side) and dissimilation (bottom side) of inorganic sulfur (S). Abbreviations: ATP, adenosine triphosphate; APS, adenosine-5'-phosphosulfate; PAPS, 3'-phosphoadenosine-5'-phosphosulfate; Sat, ATP sulfurylase; Apk, APS kinase; Ppr, PAPS reductase; Sir, sulfite reductase; Apr, adenylyl-sulfate reductase; Dsr, dissimilatory (bi)sulfite reductase.

Aromatic sulfonates, SO_3^{1-} groups bound to aromatic rings, can also serve as a S source for assimilation. This assimilatory pathway is controlled by a genetic cluster called the 'sulfate-stimulation-induced stimulon' that encodes three distinct gene clusters, *asf*, *ssu*, and *ats*. Together these three genes control the assimilation of S by bacteria in soil. The *ats* gene cluster is responsible for the binding of aromatic sulfonates and transportation into the cell. The *ssu* genes encode for the cleavage for the C—S bond by a monooxygenase. The *asf* gene cluster provides the reducing equivalents from the oxidation of NADH necessary to assimilate aromatic sulfonates. The overall result of this pathway is the production of sulfite, which reacts with PAPS reductase to form sulfide, as described above. Alkane sulfonates are also a source of S for assimilation. For most alkane sulfonates the pathway is very similar to that described above, with the exception that the transport mechanisms differ. There is a distinctly different pathway for taurine, involving an α -ketoglutarate-dependent dioxygenase which oxidizes both taurine and α -ketoglutarate to release sulfite. This pathway is regulated by *TauD* and is specific for taurine, with little activity seen for other alkane sulfonates.

Sulfate esters are assimilated by the action of three separate enzyme families (Linder, 2018), including arylsulfatases, metallo- β -lactamase (MBL)-like sulfatases, and α -ketoglutarate/Fe(II)-dependent dioxygenases. Arylsulfatases are broadly grouped by their pH optima, with alkaline sulfatases having a pH optimum of 8.3–9.0 and acid sulfatases having a pH optimum of 6.5–7.1. In essence, arylsulfatases hydrate the S—O bond and release sulfate (Fig. 4; top reaction), which can then be assimilated by the pathway described above for sulfate. These assimilatory processes should be clearly differentiated from processes that use organo-sulfur compounds as a C and energy source. Organo-sulfur assimilation genes are regulated by the sulfate-stimulation-induced stimulon, which in turn is regulated by levels of cysteine, thiosulfate, and sulfite present in the cell. The presence of these

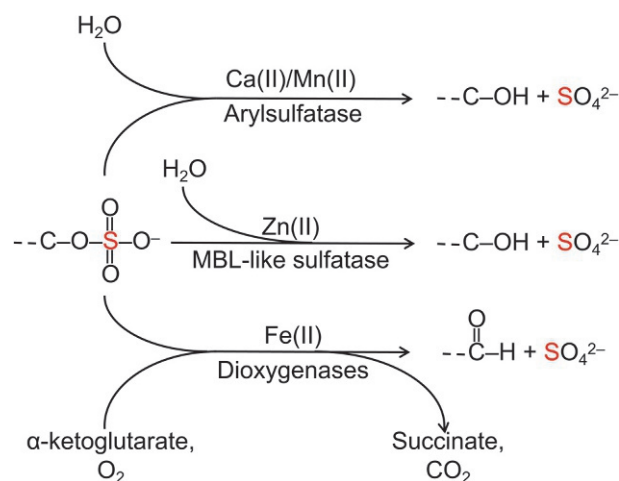


Fig. 4 Reaction mechanisms for desulfurization of sulfate esters. Metal ion requirements for each enzyme family are indicated.

compounds inside the cell will downregulate the sulfate-stimulated-induced genes and thereby repress expression of assimilatory enzymes. This is a logical regulatory control, since if there is an excess of cysteine and sulfate present in the cell, the cell likely does not need to assimilate S. The MBL-like sulfatases are involved in aliphatic sulfate ester desulfurization and produce sulfate and the corresponding aliphatic alcohol (Waddell et al., 2017) (Fig. 4; middle reaction). Unlike arylsulfatases, the MBL-like sulfatases hydrate the C—O bond. The α -ketoglutarate- and Fe(II)-dependent dioxygenases utilize sulfate esters and α -ketodicarboxylic acids as co-substrates. The α -ketoglutarate- and Fe(II)-dependent dioxygenases hydrate the C—O bond with the help of one oxygen atom derived from molecular oxygen and produce sulfate and aliphatic aldehyde (Fig. 3; bottom reaction). Meanwhile, the α -ketodicarboxylic acids are oxidatively decarboxylated into succinate.

Sulfur oxidation

Many organisms, such as chemoautotrophs and photoautotrophs, use S as a source of energy. This occurs by oxidizing S^0 to SO_3^{2-} and finally to SO_4^{2-} and in the process stripping six electrons from S and using oxygen or nitrate as a terminal electron acceptor. A great variety of thiobacilli can perform these reactions. Some thiobacilli are chemoautotrophs (only able to use S as an energy source), others are facultative chemoautotrophs (able to use C as an energy source if necessary) and some are mixotrophs (able to use C and S as an energy source at the same time). Most thiobacilli require the presence of oxygen but some, such as *Thiobacillus denitrificans*, can use nitrate as a terminal electron acceptor. Similarly, some thiobacilli can use Fe^0 in addition to S as an electron donor.

The research on the importance of thiobacilli on S oxidation comes largely from pure culture studies. The actual situation in soil involves a wide diversity of organisms. There are gliding S oxidizers, including bacteria, the cells of which are arranged in trichomes, which show a gliding motion on the substrate. The most important members of this group in relation to S-oxidation in soils are species of *Beggiatoa*, bacteria that participate in sulfide oxidation in the root zone of rice. All strains of *Beggiatoa* deposit S in the presence of H_2S :

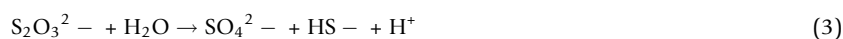


Phototrophic bacteria, such as *Chromatium* and *Chlorobium*, also play an important role in sulfide oxidation in rice paddy soil, but not in aerobic agricultural soils. A number of nonfilamentous, chemolithotrophic sulfur-oxidizing bacteria, such as *Sulfolobus*, *Thiospira*, or *Thiomicrospira*, have also been isolated from special habitats, but the importance of these bacteria in S oxidation in soils has yet to be determined. Some thiobacilli, such as *Thiobacillus denitrificans*, can oxidize S or thiosulfate and use nitrate as an electron acceptor. Other aerobic bacteria, such as *Arthobacter*, *Bacillus*, and *Pseudomonas*, oxidize S in soil during their normal metabolism of oxidizing organic compounds in soil and using oxygen as a terminal electron acceptor. These reactions have not been characterized but it is thought that this is occurring as a side reaction to the primary transformations being carried out by these organisms. In the soil ecosystem, it is widely assumed that thiobacilli are the dominant S oxidizers but this view is based on the observation that if you add elemental S to soil, the numbers of thiobacilli increase. However, no consistent correlation between S-oxidation rates and the prevalence of thiobacilli has been demonstrated, except in very broad terms. In general, it is assumed that initially heterotrophs oxidize elemental S until the pH is low enough that oxidation of S by chemoautotrophs is energetically favorable and then thiobacilli dominate. Thus, under energy-limiting conditions in an environment dominated by dissolved, reduced S compounds with little organic matter, thiobacilli will dominate. In contrast, in a soil in which there are significant amounts of organic matter and less reduced S, heterotrophs will dominate.

Microorganisms responsible for S oxidation are generally confined to the surface layer of organic-rich sediments, where they can take advantage of steep chemical gradients of S and oxidants (e.g., nitrate and oxygen). To bridge the spatial or temporal gap between S and oxidants, bacteria have developed interesting adaptations such as the motility and storage of both nitrate and elemental S in the cells and the formation of several-centimeter long chains, i.e., cable bacteria, conducting electron currents (Schulz and Jørgensen, 2001).

Disproportionation of sulfur

In the latter part of the twentieth century, investigators discovered that not only can bacteria oxidize and reduce S, but they can also disproportionate it. Disproportion reactions are best likened to a fermentation in which one portion of a molecule acts as an electron donor and the other, an electron acceptor. Bacteria capable of S disproportionation can be found in *Desulfovibrio*, *Desulfobacter*, and *Desulfocapsa* spp., among others. Bacteria are able to disproportionate thiosulfate, in which the S atom that carries a +5 charge is oxidized to one with a charge of +6. This S^{6+} is released as SO_4^{2-} . Simultaneously, bacteria oxidize the other S atom in thiosulfate that carries a -1 charge to one with a charge of -2. This S^{2-} is released as hydrogen sulfide:



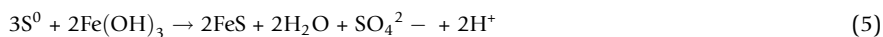
This metabolic process has immense implications for the global S cycle since typically in sediments sulfate predominates in the upper layers and sulfide in the lower layers. Hence, this metabolic reaction reveals an entirely new and, previously unsuspected

ecological niche, which is responsible for up to 60% of the thiosulfate transformations in sediment. Since then, other bacteria have been found that can disproportionate sulfonates such as taurine to sulfate and sulfide. The bacteria capable of this organic S disproportionation reaction form a new species called *Desulfonisporea*.

In contrast to the disproportionation of thiosulfate, the disproportionation of S^0 is not thermodynamically favorable. Low concentrations (<1 mM) of sulfide products, which usually require scavenging with iron or manganese, can make S^0 disproportionation thermodynamically favorable (Finster, 2008). The net stoichiometry of S^0 disproportionation is:



however, this net stoichiometry will change if sulfide is scavenged by iron:



Disproportionation is a mechanism for depleting ^{34}S in sedimentary sulfides. The disproportionation of S^0 produces ^{34}S -depleted sulfides and ^{34}S -enriched sulfates relative to S^0 . The sulfide produced by this process is reduced by approximately 4–9‰ relative to S^0 . Thus, the production of ^{34}S -depleted sulfides would increase with repeated cycles of sulfide oxidation to S^0 and S^0 disproportionation.

Reduction of sulfur

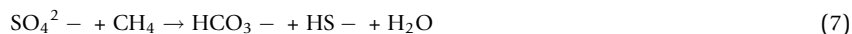
The reduction of S can occur for two reasons: assimilation and dissimilation. The assimilation pathway has been discussed above – bacteria convert sulfate to sulfide for inclusion in amino acids such as cysteine. Dissimilatory sulfate reductase refers to a process whereby sulfate is used as a terminal electron acceptor. Sulfate-reducing bacteria use such an enzyme, which allows them to oxidize H_2 or organic substrates such as lactate, malate, acetate, succinate, fumarate, fructose, glucose, fatty acids, and ethanol, and use sulfate as a terminal electron acceptor. Hence, these bacteria are primarily found in environments where there is little or no oxygen and live off the fermentation end products produced by other bacteria. Sulfate-reducing bacteria catalyze the following process (Fig. 3; bottom side):



In this process, sulfate is transported into cells by sulfate transporters and becomes activated with ATP in the cytoplasm by ATP sulfurylase to form APS. The APS is reduced by adenylyl-sulfate reductase (Apr) to form sulfite, and the (bi)sulfite is further reduced by (bi)sulfite reductase (Dsr) to finally form H_2S . Then, H_2S diffuses extracellularly through the cell membrane.

Since hydrogen sulfide is extremely toxic, these organisms require some sort of metal to react with the hydrogen sulfide produced and precipitate as an insoluble metal sulfide. In soil, this metal is typically iron in the form of FeS, which is often found in areas occupied by sulfate-reducing bacteria. The reaction catalyzed by sulfate-reducing bacteria reaches an optimum at an E_h of -300 mV at a pH of 7. These conditions are typically reached in the strictly anaerobic zones. Typically, terminal electron acceptors are depleted in the following order: oxygen, nitrate, nitrite, manganite, iron, sulfate, and finally, carbon dioxide. Thus, there are many alternative electron acceptors that will be used by a microbial community before sulfate is reduced. Despite this limitation, sulfate-reducing bacteria and their activity are commonly found in soils, sediments, polluted water, oil-bearing strata, and shales.

Sulfate-reducing bacteria can implement an alternative survival strategy if sulfate levels in the water are too low to be used. The H^+ produced by organic matter oxidation is transferred directly to a methanogen which then consumes that H^+ to reduce CO_2 and produce methane. Because the methanogens keep the partial pressure of hydrogen so low, the sulfate-reducing bacteria can use this mechanism to obtain energy from organic matter degradation even when there is not enough sulfate around them to transform to sulfide. It is estimated that 3–4% of the global flux of organic carbon flowing to the seafloor is converted to methane. Methane then diffuses upward and is quantitatively oxidized by anaerobic methanotrophic archaea upon encountering sulfate serving as the electron acceptor (Reeburgh, 2007):

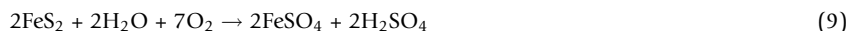


Sulfur transformations and environmental quality

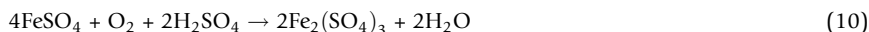
The biological transformations of S can lead to significant environmental problems in the soil. The enzyme sulfate-reducing bacteria use to transfer methyl groups can also accidentally methylate mercury to form methylmercury. Sulfate-reducing bacteria have been identified as one of the primary causes of the increased mercury accumulation in many ecosystems because methylmercury is the form of mercury that most readily accumulates in food chains. Typically, this reaction occurs in anaerobic, lowland soils where the activity of sulfate-reducing bacteria is closely linked to methylmercury production. However, sulfate-reducing bacteria can also mitigate environmental pollution through the production of metal sulfides by the following reaction:



Sulfate-reducing bacteria like *Desulfovibrio* spp. can form sulfides of Sb, Co, Cd, Fe, Pb, Ni, and Zn during the reduction of sulfate. The extent of metal sulfide genesis depends on many factors, such as the amount of sulfate present and the relative toxicity of the metal ion. In nature, this toxicity is reduced when the metal ions are adsorbed on clays or complexed with organic matter. The oxidation of metal sulfides in soil involves both chemical and microbial processes and, as a result, is a more complex process than is the oxidation of S^0 . Chalcocite (Cu_2S), chalcopyrite ($CuFeS_2$), galena (PbS), pyrite (FeS_2), and nickel sulfide (NiS) are just a few examples of metal sulfides that are subject to microbial transformations. For example, the biological oxidation of pyrite (FeS_2) can follow one of two pathways. The first pathway follows a series of oxidation steps, described in Eqs. (9) to (12). The second mechanism of pyrite oxidation involves thiosulfate as an intermediate but in the end requires the oxidation of elemental S to sulfate by *Thiobacillus*. These biotic oxidations are responsible for the formation of acid mine drainage and acid soil formation in surface mine spoils. First, ferrous sulfate is formed as the result of an abiotic oxidation step:



This reaction is then followed by the bacterial oxidation of ferrous sulfate, generally by *T. ferrooxidans*:



This reaction occurs chemically but can be accelerated 10^6 – 10^8 times by thiobacilli. This bacterial oxidation of ferrous iron plays an important role in the bioleaching of metal sulfides in the environment because it cycles the iron between +2 and +3 and the iron is then free to oxidize sulfide minerals abiotically.

Subsequently, ferric sulfate is reduced and pyrite oxidized by a strictly chemical reaction:



The elemental S is finally oxidized by *T. thiooxidans*, and the acidity produced helps the whole process to continue.



Note the net production of 10 molecules of H_2SO_4 during the process.

Although several S-oxidizing thiobacilli and heterotrophs can be isolated from acid sulfate soils in which pyrite is being oxidized, they appear not to play an important role in the process, with the exception of *T. ferrooxidans*. The biological oxidation of sulfides and other reduced S compounds can have severe consequences for the environment. For example, acid mine drainage contaminates several thousand kilometers of streams in the Appalachian coal-mining region of the USA.

Conclusion

There is a large amount of S present in soil and most of it is found in a form susceptible to biological transformations. These transformations include mineralization, assimilation, oxidation, and reduction. Since the industrial revolution, anthropogenic activity has impacted the S cycle by increasing the deposition of sulfate in the form of acid rain and by exposing more S to oxidation during mining operations. The four basic transformations discussed in this article influence the fate of the movement of S in terrestrial and aquatic ecosystems and the environmental impact of anthropogenic activities altering the S cycle.

References

- Droux M (2004) Sulfur assimilation and the role of sulfur in plant metabolism: A survey. *Photosynthesis Research* 79(3): 331–348.
- Finster K (2008) Microbiological disproportionation of inorganic sulfur compounds. *Journal of Sulfur Chemistry* 29: 281–292.
- Linder T (2018) Assimilation of alternative sulfur sources in fungi. *World Journal of Microbiology and Biotechnology* 34(4): 1–7.
- Reeburgh WS (2007) Oceanic methane biogeochemistry. *Chemical Reviews* 107: 486–513.
- Schulz HN and Jørgensen BB (2001) Big bacteria. *Annual Review of Microbiology* 55: 105–137.
- Waddell GL, Gilmer CR, Taylor NG, Reveral JRS, Forconi M, and Fox JL (2017) The eukaryotic enzyme Bds1 is an alkyl but not an aryl sulfohydrolase. *Biochemical and Biophysical Research Communications* 491: 382–387.

Plant uptake of mineral nutrients

Angela Hodge, Department of Biology, University of York, York, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

This is an update of A. Hodge, NITROGEN IN SOILS | Plant Uptake, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 39–46, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00159-4>.

Introduction	310
Soil formation factors impact upon nutrient availability	312
Nutrients in the soil-plant system	313
Movement of ions to the root surface	315
Root responses	316
Conclusion	319
References	320

Abstract

Plants require essentially the same set of elements for normal growth, acquired mainly as ions from the soil environment. Yet soils vary widely in their capacity to supply these ions due to large variations in soil development and inherent heterogeneity within the soil system. Roots predominantly access nutrients from the soil solution phase but concentrations are low, thus replenishment from the solid and, depending on the nutrient, the gaseous phase, is important. Roots and ions must also come into contact so that uptake can occur, and the responses of roots and the movement of ions will be discussed.

Key points

- Nutrients (macro- and micro-nutrients) essential for plant growth and development.
- The five factors of soil formation, and how these impact nutrient availability.
- Nutrients in the soil-plant system as influenced by the particular ion and the different phases (solid, solution and gaseous), the organic and mineral components of soil and the microbial population.
- Root-ion contact including movement of ions to the root.
- Root responses to nutrients including ideotype root traits for acquisition of N and P and mycorrhiza colonization.

Introduction

Despite the vast diversity observed among green plants, they all require the same set of essential elements to grow, develop and successfully complete their lifecycle. These elements can be further grouped into the so called ‘macronutrients,’ because plants require these elements at relatively large concentrations, and the ‘micronutrients,’ which, as the name suggests, are required at much smaller concentrations, but, nevertheless, are still essential to ensure the plants lifecycle can be completed. In reality, demonstrating that elements are actually ‘essential’ can be problematic especially in the case of those that are required in such small amounts analytical detection may be difficult or may be subject to contamination issues. Indeed, it is only relatively recently that nickel (Ni) was added to the ‘essential’ element list (Brown et al., 1987) and others may still yet be identified and subsequently included.

While plants can obtain some of these essential elements, such as carbon, oxygen and hydrogen, through photosynthesis, the majority are mainly acquired as ions from the soil environment (Fitter and Hay, 2002; Table 1). The elements shown in Table 1 are required by plants for the same essential functions and when plants are unable to take up these elements in sufficient quantities then nutrient deficiencies occur. Many symptoms of different nutrient deficiencies however, can look remarkably similar with, for example, interveinal chlorosis (i.e., where yellowing of leaf tissue between the veins occurs while the veins remain green) being indicative of magnesium, iron, or manganese deficiency. Symptoms of nitrogen and sulfur deficiency, which, in both cases involves yellowing of the leaves, may also appear very similar and, as such, hard to distinguish, especially during the early stages of deficiency. While the location (i.e., whether in young or old leaves) of the deficiency symptoms may also provide an indication of the particular nutrient in question (and its associated mobility), it often can be difficult to decipher the exact nutritional problem by visual observation alone. While, in other cases, such as is frequently the case for phosphorus deficiency, visible symptoms may not be obvious even though phosphate supply may be limiting for the plant. If deficiency symptoms are not identified early enough and remedial action taken, then large crop losses may occur. In some cases, such as in the case of boron deficiency, such crop losses cannot be recovered by later applications. Nutrient deficiency symptoms may also look slightly different depending on the plant species in question and should always be compared with a healthy specimen of the same species. Under field conditions further

Table 1 The elements, with the exclusion of carbon (C), hydrogen (H) and oxygen (O) which are incorporated during photosynthesis, required for healthy plant growth, their typical concentrations, main function and deficiency symptoms.

Element	Typical concentration in plant tissue ($\mu\text{g g}^{-1}$)	Main function	Deficiency symptoms
Macronutrients			
Nitrogen (N)	20,000	Constituent of amino acids, hence proteins, and of chlorophyll.	Stunted growth and chlorosis; Yellowing of older leaves; Also pink-purple tints to stems, petioles, leaves.
Phosphorus (P)	2,000	Nucleic acids (e.g. DNA, RNA); Adenosine triphosphate (ATP); Membranes (i.e. fatty phospholipids)	Leaf tips have a burnt appearance; Leaves have an intense green or reddish-purple coloration.
Potassium (K)	10,000	Electric charge balance in cells; Osmoregulation.	Leaf tips with brown or burnt appearance; Area between leaf veins becomes yellow or white in appearance while veins are green (called interveinal chlorosis)
Calcium (Ca)	500–5,000	Cell wall and membrane integrity and function; Second messenger.	New leaves have twisted or stunted appearance; Causes blossom-end rot (fruit with a dark spot at bottom).
Magnesium (Mg)	200–2,000	Constituent of the chlorophyll molecule; Required for numerous enzyme reactions; Control of cation - anion balance.	Yellowing of older leaves between leaf veins (i.e. interveinal chlorosis)
Sulfur (S)	1,000	Component of many primary and secondary plant metabolites including some amino acids.	Chlorosis and yellowing of young leaves sometimes followed by older leaves.
Micronutrients			
Iron (Fe)	100	Operation of redox systems; Chlorophyll synthesis.	Interveinal chlorosis in young leaves.
Manganese (Mn)	50	Various including: Redox processes; Chlorophyll formation; Photosynthesis; Respiration; Cell division.	Interveinal chlorosis in young leaves; Sometimes with tan or dark spots on chlorotic areas; Plant growth stunted.
Boron (B)	20	Diverse roles including: Cell wall formation and stability; Integrity of membranes.	Depression of growing points (root tips, buds etc.) and stunted/deformed growth.
Zinc (Zn)	20	Enzyme functioning; Membrane structure; Plant hormones (e.g. auxin).	Interveinal chlorosis particularly in younger leaves. May have necrosis spots. Terminal leaves may be more clustered in rosette form.
Copper (Cu)	5	Redox reactions.	Plant is stunted; Leaf tips may turn blueish-green color or brown depending on plant species.
Chlorine (Cl)	1,000–10,000	Various including roles in osmotic and stomatal regulation, Evolution of oxygen in photosynthesis etc.	Wilting and curling of leaves at margins; Chlorosis then necrosis in severe cases.
Molybdenum (Mo)	0.2	Enzymes of N metabolism.	Older leaves more affected and have yellow appearance; Rest of the plant may become light green.
Nickel (Ni)	0.5	Enzymes of N metabolism.	Necrosis of leaf tips; Chlorosis of younger leaves.

Adapted from Fitter AH and Hay RKM (2002) *Environmental Physiology of Plants*. (3rd edn.). London: Academic Press.

complications arise including more than one nutrient may be deficient at the same time; one element may be limiting and another at excessive (toxic) levels; plants may also be subject to pathogen attack; all of which may further complicate the correct identification of any nutrient deficiency to which the plant may be subject. Consequently, analytical analysis of leaf material to confirm the limiting nutrient is normally advised. This is particularly important in an agricultural context to provide the appropriate addition to alleviate the nutrient deficiency without exacerbating other issues given that the difference among deficient-adequate-toxic levels can be quite narrow.

In addition to the 'essential' elements there is another class often referred to as 'beneficial' elements. Examples of beneficial elements are silicon (Si), sodium (Na), cobalt (Co), aluminum (Al) and selenium (Se). The members of this class can stimulate plant growth or promote survival but are considered non-essential for plants to complete their life cycle. This is despite the fact that these 'beneficial' elements may be essential for some plant species (e.g., such as Si in horsetail (*Equisetum spp.*) and some C4 species), or under certain specific conditions (e.g., Co to ensure nitrogen-fixation rates in plant-bacterial nitrogen-fixation associations given Co is a component of the coenzyme cobalamin (vitamin B₁₂) required by nitrogen-fixing bacteria).

Unsurprisingly, therefore, the definition between what constitutes an essential versus a beneficial element has long been a controversial topic, and perhaps the arguments on this subject have been most passionately debated in the case of Si (Epstein, 2009; Coskun et al., 2019). Plants can be grown to maturity and complete their life cycle when grown in solution culture containing no Si source, hence, Si does not meet one of the key classical criteria for being classed an 'essential' element. However, Si uptake and its subsequent accumulation in leaves and stems has often been found to enhance growth through, for example, an increased erectness of leaves, improved light interception and enhanced photosynthetic rates. Furthermore, Si uptake enables plants to tolerate abiotic

and biotic stress in the field (Coskun et al., 2019) which may, in turn, be fundamental to the very survival of the plant in that environment. Thus, while Si may not be classified as essential in a nutritional context, it may be considered 'quasi-essential' for plants when under stressed conditions (Epstein, 2009).

When ions are at the root surface they can then be taken up by the plant. Selective uptake of ions occurs so that the plant can both accumulate ions it requires at higher levels than that in the soil solution and reduce or exclude those ions that it requires in lower amounts or that may be harmful. Uptake into the root may occur via diffusion (passive diffusion or facilitated or restricted by transport proteins) or by active transport. However, before ion uptake can occur, first, roots and ions have to encounter each other and it is the factors that influence ion availability and these encounters that will be discussed in this chapter.

Soil formation factors impact upon nutrient availability

Soils vary widely in their capability to supply the elements shown on Table 1 due to the interactions among the five factors of soil formation namely parent material, climate, topography, organisms and time, with the latter being the time the other four factors have had to work on the system. As soils age, nutrient levels tend to decline due to losses through leaching or erosion processes, consequently, some of the most ancient soils on earth, such as those found in parts of South Africa and Australia, are also among the most nutrient impoverished. However, some soils can also become rejuvenated and nutrient levels can increase through, for example, deposition of material from elsewhere, glaciation events and so forth.

The parent material is the initial state of the system from which the subsequent soil develops. The climate that the parent material is under influences how rapidly the parent material weather and thus, how quickly soil formation occurs with warm, moist climates increasing soil development compared to cold, dry climates. The parent material is composed of mineral or organic material or both and can be considered, with the notable exception of nitrogen, as setting the upper limit for the nutrient content of the soil. Nitrogen levels are the exception because N can enter the soil through fixation from the atmosphere. Although weathering of rocks was previously considered to release only negligible amounts of nitrogen, more recently this viewpoint has been questioned with the evidence that rocks, particularly sedimentary rocks, may release more N to soils than previously realized (Houlton and Morford, 2015; Houlton et al., 2018). This in turn, has led to suggestions that this may have implications for global climate change predictions if the capacity of plants to respond to elevated CO₂ levels are not so constrained by N limitations to growth and development as previously thought (Houlton et al., 2018). However, whether these effects actually result in global, rather than local, impacts still has to be fully determined, but may help to explain why some younger soils contain more N than perhaps would be predicted based upon estimates of nitrogen fixation alone.

Soil that is derived from different parent materials can have distinctive properties. For example, soils formed from limestone are, at least initially, very base-rich while those formed on granite or sandstone are very base-poor. Parent material which is base-poor has a tendency to be predisposed to acidification and podzolization processes especially in climatic areas that are cool and with considerable precipitation. In areas where the vegetation is coniferous forest, an acidic humus is produced. The acidic humus reduces microbial decomposition as well as earthworm activity so that the layers (or horizons) of the soil profile are not merged and instead tend to be sharply defined. Podzols have a grey or light colored 'E' or eluvial horizon due to severe leaching (or eluviation) washing out everything except the silica (SiO₂) sand. Instead, the aluminum and iron oxides accumulate in the lower, subsoil horizons. The accumulation of the iron oxides may result in a hardpan layer forming which, in turn, can impede water drainage. While podzols can be found at all elevations, topography is another important factor which influences the rate and type of soil formation, e.g., soils on slopes tend to suffer from erosion so have a tendency to be shallower than soils at either the bottom or top of hills. Soils at the bottom of hills also tend to be more nutrient-rich due to downward movement of nutrients as a result of erosion or leaching processes. The topography of an area can also have an important impact upon the biodiversity of the plant species present. For example, in fen peats, where the plant species present are controlled by the soil water chemistry, if the topography allows the accumulation of base-rich ground water above limestone rock the dissolved calcium contained within the water acts as a pH buffer to maintain a more basic pH which, in turn, enables more plant species to survive in these particular ecosystems and so promotes biodiversity.

The impact of organisms (including humans) on soil formation is complex. Humans have a large impact upon soil formation through large scale changes to land-use including conversion of natural areas for agricultural production and the subsequent intensive management practices these normally entail. Other organisms including other animals, microorganisms and plants also impact on soil properties through bioturbation activity or altering the chemical environment through the release or selective uptake of substances from the soil. Studies on the Glacier Bay region in south eastern Alaska have proved to be useful to study soil development over relatively short timescales given records of the glacier retreat exist, and thus, the ages of the glacial moraines are known. This helps facilitate the study of soil properties that relate directly to the vegetation. The plant species that are present in soil systems may produce litter inputs that are acidic or basic which, in turn, will influence the organisms that will be able to persist under these altered soil conditions. At Glacier Bay it was found that under alder the uppermost soil horizons had a dramatic reduction in pH from 8.0 or more, to pH 5.0, within a 35 to 50-year period. In contrast, other plant genera present (such as *Salix* and *Populus*) did not demonstrate such a dramatic impact upon the soil pH as alder (Crocker and Major, 1955). With time, spruce replaces the alder at Glacier Bay thus the impact of organisms also involves ecological succession where the mix of species and the habitat changes over time. Plant species present are also impacted by the climate thus emphasizing that the factors that influence soil formation should not be considered in isolation from each other.

Nutrients in the soil-plant system

The five factors of soil formation result in variations in soils within the landscape but variations at a more local level, and at scales relevant to plant roots, can also occur. Thus, while soil development helps explain the variation among different soils to supply nutrients to plants, generally much shorter time scales are relevant over the lifetime of a plant. Plants have to acquire the majority of their nutrients from the soil and soils are complex environments, comprising solid, liquid, and gas phases. Plant roots predominantly obtain their nutrients from the soil solution phase but concentrations of ions in the soil solution phase are generally small, thus the rate at which this nutrient supply is subsequently replenished becomes important. Mineral nutrient cycling in soil systems involves losses, inputs and transfers among components of the nutrient cycle but it is the equilibrium among these three different phases that largely impacts upon actual nutrient availability for the plant which, in turn, depends on the nutrient in question. For example, in the case of nitrogen, inputs may be via biological fixation of N_2 from the atmosphere or wet (in the form of nitrate (NO_3^-) and ammonium (NH_4^+)) and dry (nitrous oxides (NO_x), and ammonia (NH_3)) deposition whereas losses may occur due to NH_3 volatilization, denitrification and leaching of highly mobile NO_3^- ions. Both NH_3 and NO_x are emitted from the burning of fossil fuels. Agricultural systems also contribute to NH_3 emissions predominantly due to livestock manures and both organic and inorganic fertilizer additions. Moreover, these N forms are not necessarily returned to the same system from which they were originally emitted and can cause problems if deposited on naturally N-sensitive soils resulting in a dramatic reduction in the biodiversity of the plant community as well the diversity of other species associated with that community. Furthermore, NO_x , together with sulfur dioxide (SO_2), which is also emitted predominantly from the burning of fossil fuels, can react with water and oxygen in the atmosphere to form nitric and sulfuric acid respectively which could then fall as precipitation: a phenomenon commonly referred to as acid rain. Acid rain can fall on plant-soil systems at some considerable distance from the original source of the emission and in addition to increasing the acidity of the soil can reduce the fertility of affected soils through leaching of calcium and other base cations. Other impacts upon soil chemistry include an increase in accumulation of N and S, increased Al mobilization and an increase in Al levels in the soil solution. The latter can subsequently negatively impact upon plant growth as well as potentially leaching into water courses and causing further environmental problems. The impact of acid rain upon soil chemistry will, however, depend upon the buffering capability of the particular soil experiencing the acid deposition. Well buffered soils will be less prone to these changes than soils that have low buffer power. The gaseous phase is thus clearly important in N cycling, in common with other element cycles such as carbon and sulfur. In contrast, there is no gaseous phase in the phosphorus cycle and so no significant quantities of phosphorus in the atmosphere.

Most ions in soil occur not in the solution phase that plants mainly access but, rather, in the solid phase as a component of soil organic matter or inorganic soil minerals. For example, while the total soil P content may be relatively large, the concentrations in the soil solution are typically only c. 0.1–10 μM (Schachtman et al., 1998; Hinsinger, 2001), and it is these small concentrations to which the plant has access. Mineralization of organic material releases nutrients including N, P and S into the soil-plant system some of which will enter the solution phase. Mineralization is carried out by a wide range of soil animals and microorganisms. However, while microorganisms play a central role in transformation and accessibility of N, the supply of P tends to be dominated instead by inorganic equilibria because of the propensity of phosphate ions to react with charged surfaces including organic matter and the dominant cations in most soils (i.e., Al^{3+} , Fe^{3+} or Ca^{2+}). The resulting adsorption of phosphate ions to these surfaces leads to sequestration of P in the solid phase into both labile and non-labile pools. Of these two, the labile P pool is less strongly adsorbed so is more likely to come into the solution phase and is in equilibrium with it. Over time some non-labile P will react to become labile P although much of the non-labile P will remain in this form and thus, inaccessible to the plant.

In the case of N, microorganisms play a central role in the transformation and accessibility of N forms in soils through the decomposition of complex organic compounds into more plant accessible forms. However, in most natural systems, mineralization rates tend to be slow and the subsequent conversion of NH_4^+ through to NO_3^- were generally considered to be the rate-limiting steps in the terrestrial N-cycle. Due to their rapid turnover times, rapid adaptation rate and proximity to the N-reaction sites, microorganisms are well positioned to utilize released inorganic N sources first, with the plant only being able to access what is left over after microbial uptake. Studies using ^{15}N labelled organic materials of varying complexity added as discrete zones or patches to the soil-plant system have shown that microbial sequestration of this N however, was only short term and that plants did capture most of the mineralized N from the organic material. This has been attributed to the longer lifespan of roots in comparison to most soil microbes (Hodge et al., 2000a).

Most transformations of N in soil also depend on the supply of carbon (C) and thus these two nutrient cycles are considered closely linked. The C:N ratio of the substrate determines the rates at which N (and C) are released and mineral N made available for plant uptake. Also influential are the composition of the decomposer microbial community and its requirements for N. Fungi tend to have a greater substrate assimilation efficiency and a smaller N requirement per unit of C assimilated than bacteria. Thus, if the decomposer population is predominantly fungal, more N may be released than if it was predominantly bacterial. However, fungi also tend to turn over less rapidly than bacterial populations, so the N contained in their structural components will be locked up, and hence unavailable for use by other organisms including plants, for longer than bacterial N. Generally net mineralization, which results in the release of N, occurs if the C:N ratio of the substrate being decomposed is less than approx. 25:1. At higher C:N ratios, net immobilization occurs and the microbial decomposer community requires additional N from the soil system to decompose the substrate. Net immobilization of N reduces the N available for the plant and can reduce overall plant growth. Although other factors influence the release of N, for example lignin content of the substrate, there is a strong inverse relationship between the amount of

Table 2 The percentage N captured by plants from organic material which in all cases was *Lolium perenne* shoot material but of varying C:N ratio.

Plant species	C:N ratio	Duration (days)	N from organic material captured by plant (%)	References
<i>L. perenne</i>	12:1	30	22	Hodge et al. (2000b)
<i>L. perenne</i>	12:1	70	26	Hodge et al. (2000c)
<i>Poa pratensis</i>	12:1	70	26	Hodge et al. (2000c)
<i>L. perenne</i>	21:1	49	11	Hodge et al. (2000d)
<i>L. perenne</i>	31:1	39	5	Hodge et al. (1998)
<i>P. pratensis</i>	31:1	39	3	Hodge et al. (1998)
<i>L. perenne</i>	31:1	56	9	Hodge et al. (1999)
<i>P. pratensis</i>	31:1	56	5	Hodge et al. (1999)

plant N captured from organic substrates and the C:N ratio of the substrate, as shown in Table 2 from the same organic material (*Lolium perenne* shoots) but of varying C:N ratio.

Whereas the relationship between the C:N ratio of the substrate determines if net N mineralization or immobilization will occur is well established for N, this is not the case for P. Moreover, in forest surface soils net C, N and P mineralization are not closely linked. Differences in the reactivity of P versus N chemical forms with other ions in the mineral soil system together with increased microbial immobilization of N compared to P in both decomposing litter and organic horizons were thought to be responsible for this uncoupling between N and P mineralization (Brödlén et al., 2019).

In some ecosystems such as the arctic, alpine and boreal forest, annual plant N uptake was found to be higher than should be available through inorganic N capture alone. This led to the conclusion that many plant species in these ecosystems are accessing organic N sources directly without the prior need for microbial mineralization. Forms of organic N in these soils range from the complex (e.g., humus) to the simple (e.g., amino acids and urea), although most of the evidence for direct uptake of intact organic N compounds via plant roots comes from studies on amino acid N or small peptides and it is now widely accepted that roots from a wide range of plant species can take up these simple organic N forms directly. The findings of some other studies suggest that plant roots may release proteases to promote the breakdown of soil proteins and are also able to take up intact proteins into root cells, all without the involvement of any microbial activity (Paungfoo-Lonhienne et al., 2008). In most soils proteins are a considerable organic N reservoir, thus, such an N acquisition pathway could potentially be very important for the plant. However, microorganisms can also access these N sources so plant-microbial competition will occur for these soil N forms also. Moreover, the results from studies on root proteases are contradictory with some concluding that proteases activity was mainly root-bound rather than secretory and that the use of exogenous protein was an indirect by-product linked to other root functions rather than having direct involvement in plant N acquisition *per se* (Greenfield et al., 2020). Thus, the ecological importance of root produced proteases as an important N uptake pathway in plants has yet to be fully resolved.

Rather than competing with the plant for nutrient resources some microorganisms may aid nutrient capture by plants. Such associations can be rather loose, such as the plant growth-promoting rhizobacteria (PGPRs) which covers bacteria from a number of different species but which have been found to be beneficial to promoting plant growth. This can occur via a wide range of mechanisms including via an impact upon nutrient availability through, for example, enhanced phosphate solubilization or biological nitrogen fixation. PGPRs although present in the rhizosphere are only loosely associated with the plant. Much closer symbiotic associations are formed between plants such as legumes and alder and bacteria such as rhizobium and *Frankia* respectively. These bacterial symbionts form nodules on the roots which are capable of fixing dinitrogen and thus supply the plant with N directly. In return the bacteria obtain a supply of carbon from the plant principally in the form of organic acids. The majority of plants, however, do not form N₂-fixing symbiotic associations and must rely on N uptake via their roots or mycorrhizas which are associations between plant roots and soil fungi (see “Root responses” section).

In addition to microorganisms, soil animals also play an important role in the mineralization process through physical, chemical and biological processes. Soil animals can physically break up and redistribute organic materials in the soil. They also ingest organic materials which are subsequently broken down by a range of enzymes or by microflora present in their guts. Much of the C they consume is used for metabolic processes and the excreted waste products, in the form of NH₄⁺, urea, or amino acids, are a much more readily available form of N for plant uptake than the organic material originally ingested. Soil animals feed on microorganisms as well as other soil animals. The plant benefits from the regulation of the size and activity of the microbial population and from the release of more nutrients (including N and P) back into the soil-plant system: for example, an increase in plant N capture of up to 75% has been reported as a result of protozoan grazing on soil microorganisms (Clarholm, 1985). Thus, although microorganisms may initially sequester inorganic N, grazing by protozoa ensures that this N is released back into the plant-soil environment and becomes available for root uptake.

In terms of soil fertility, clays are the most important of the inorganic soil minerals solid phase because of their reactivity. Clays consist of aluminosilicate sheets composed of aluminum, silicon and oxygen held together by various cations (e.g., Al³⁺, Mg²⁺, Fe³⁺, Na⁺, K⁺). Where clays fracture, surfaces with a net negative charge are exposed. This negative charge is balanced by cations that are ‘held’ exchangeably on these exposed surfaces which gives rise to the cation exchange capacity (CEC) of the soil. Soil organic matter

also contributes to the CEC of the soil but in this case is due to the dissociation of organic acids resulting in a net negative charge which is then balanced by the cations present in the soil. This dissociation of organic acids is, in turn, influenced by soil pH such that a soil with a low pH will have a low CEC due to the organic matter fraction. The base saturation, i.e., the percentage of CEC sites occupied by base cations such as Ca^{2+} , Mg^{2+} , Na^+ and K^+ , can also be calculated and as base saturation increases, so does pH. Soils that have a high base saturation are therefore more strongly buffered against acid cations released by plant roots and through other soil processes, such as nitrification, organic matter decay and acid rain. In the more local vicinity of the soil directly surrounding the plant root the form of inorganic N taken up by the plant can strongly influence the pH of the rhizosphere. To maintain the electrochemical gradient between root cells and soil (cells typically have electrochemical potentials of less than -100 mV) which drives cation uptake, plants taking up N as nitrate will release bicarbonate (HCO_3^-) or hydroxyl (OH^-) ions, causing the rhizosphere pH to increase. In contrast, plants taking up ammonium as the main form of N will tend to release protons, causing the rhizosphere pH to decrease. However, even when grown on nitrate as the sole source of N, the rhizosphere pH of some plants (such as chickpea) can decrease. This decrease is almost certainly due to a higher ratio of cation to anion uptake. Leguminous plants with active N_2 -fixing nodules also tend to have a large cation to anion uptake ratio and so a tendency to excrete protons and so decrease the pH of their rhizosphere soil. Thus, the consequences for the rhizosphere pH will depend not only on the form of N (i.e., NH_4^+ , NO_3^- or N_2 fixation) the plant is taking up but also on the plant species present and the buffering capacity of the soil.

Overall, soil pH can have a large impact upon nutrient availability for the plant and soils with a low pH tend to have little calcium and magnesium availability. Other nutrients, such as molybdenum (Mo), are also affected and become unavailable and consequently Mo deficiencies may occur, while phosphate can react with iron and aluminum hydroxides also reducing its availability for the plant. Microbial, particularly bacterial, activity is also reduced at low pH thus slowing organic matter decomposition rates and reducing nutrient release particularly of available N. In addition, there are other problems associated with low pH including Al toxicity which can severely impact upon normal plant, particularly root, growth. Liming of acidic soils to increase pH can improve plant growth, yield and plant species richness and diversity as demonstrated by the Park Grass Experiment at Rothamsted, UK, the oldest ecological experiment in the World which has been running since 1856 (although regular applications of lime have only been applied since 1903; [Silvertown et al., 2006](#)). However, at high soil pH the availability of some nutrients also becomes problematic with, for example, phosphate forming insoluble compounds this time with calcium and magnesium. A slightly acidic to near neutral pH is generally considered optimal for the availability of most plant nutrients.

Although a small CEC means that more nutrients will potentially be in the soil solution rather than held on exchange sites this also means these nutrients can be lost from the soil system through leaching processes and so forth, thereby reducing the overall fertility of the soil. In contrast, soils with a large CEC will have a large nutrient reservoir but these nutrients will be strongly held on the charge surfaces and therefore unavailable to the plant. Thus, CEC gives an indication of the potential nutrient supply for the plant rather than the actual availability. The addition of cations to the soil (via liming, fertilizer etc.) results in some of these held cations being exchanged and hence coming into the soil solution so that they are potentially available for acquisition by plant roots.

Movement of ions to the root surface

For nutrient ions to be acquired by the plant they have to come into contact, and there are three main ways in which this may occur. The first is root interception, whereby roots growing through the soil come into direct contact with nutrients held on soil aggregates. Root interception has been found to be more important for some nutrient ions (such as magnesium) than others, however, given that roots typically only contact c. 1% or so of the soil volume the potential for root interception is rather limited. Moreover, any factor that has a negative impact on root growth (such as soil compaction) will also reduce the capacity for root interception. Thus, root interception is not considered a major mechanism for nutrient acquisition by the plant.

The second way in which contact between roots and ions may occur is via mass flow. Mass flow, or convective flow, is driven by the transpiration pull of the plant resulting in a water potential gradient. Water that moves to the roots to satisfy this transpiration demand will also contain soluble ions and these will be available for root acquisition. The mass flow of water is equal to the water that the plant transpires and that is subsequently released to the atmosphere. Transpiration rates can be substantial, e.g., in warm, forested areas between the tropics 32×10^{15} kg yr⁻¹ of water vapor passes through the stomata, a value that is equivalent to double the water vapor content of the atmosphere (see [Hetherington and Woodward, 2003](#)). The amounts of nutrient ions that can move by mass flow can be determined by the particular nutrient concentration in the soil solution and the water flux. If uptake at the root surface is greater than the rate of supply to the root surface by mass flow, then a concentration gradient will develop from the soil toward the root. Ions will then move down this concentration gradient toward the root surface via diffusion, which is the third way in which nutrient ions can move toward the root surface. Similarly, if movement by mass flow results in an oversupply of a particular ion at the root surface compared to plant demand, then diffusion of that ion will be away from the root surface. Although movement of ions via mass flow is generally a rapid process providing water flux and soil solution concentrations of ions is large enough, movement of ions via diffusion, by contrast, is a very slow process (see [Tinker and Nye, 2000](#) for a full discussion on this topic).

While both mass flow and diffusion represent the two main ways in which ions move to the root surface for subsequent root uptake, which is the more predominant movement pathway depends largely on the particular nutrient ion in question as well as the prevailing soil conditions. [Barber \(1974\)](#) estimated the relative contribution of mass flow and, by difference, diffusion, for Mg, Ca, P and K for a maize crop. Mass flow was more than adequate to meet the crops demand for Ca^{2+} and Mg^{2+} . This subsequently resulted in an over-supply of these ions which accumulated at the root surface as supply exceeded crop uptake. Diffusion occurs down a

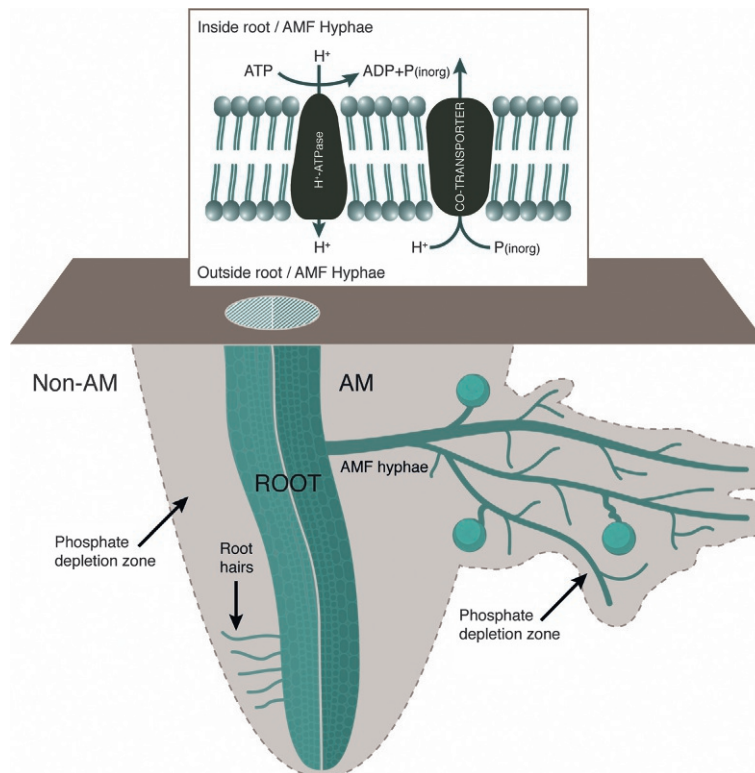


Fig. 1 Schematic representation of the phosphate depletion zones that develop around a non-mycorrhizal root with root hairs (left) and a root colonized by arbuscular mycorrhizal fungi (AMF) (right) where the depletion zone is extended due to the AMF hyphae exploring a larger soil volume. The box above shows the uptake of phosphate across the plasma membrane and is common to both AMF hyphae and roots. Protons (H^+) are pumped out the cell by H^+ -ATPases on the plasma membrane to generate the proton-motive force to enable H^+ -phosphate co-transport by high affinity phosphate transporters at the root surface or extraradical mycelium (ERM) of the AMF hyphae. Reproduced from Hodge A (2017) Accessibility of inorganic and organic nutrients for mycorrhizas. In: Johnson NC, Gehring C, and Jansa J. (eds) *Mycorrhizal Mediation of Soil: Fertility, Structure, and Carbon Storage*. pp. 129–148, with kind permission from Elsevier Science.

concentration gradient, hence, in this case, diffusion of Ca^{2+} and Mg^{2+} was away from the root and therefore not an important supply mechanism for these ions. In contrast, mass flow supplied only 11% of the crop uptake of K^+ and less than 1% of P (as $H_2PO_4^-$) where supply by diffusion was more important. However, as noted previously diffusion tends to be a slower ion movement pathway hence ions cannot arrive quickly enough to satisfy the demand of the plant resulting in a zone of depletion developing from the root surface out into the surrounding bulk soil. The size and intensity of these depletion zones (i.e., the volume of soil the root effectively can acquire the ion from) depends on the particular ion in question and is also influenced by other factors including if the root is in a mycorrhizal symbiosis (Fig. 1).

The depletion zones that develop due to the diffusion of phosphate for example, are narrow and sharply defined. Hence, are extremely 'depleted' in phosphate. If the root is in mycorrhizal symbiosis however this depletion zone becomes extended due to the mycorrhizal fungal hyphae exploring a larger soil volume than the root alone (Fig. 1). If, prevailing conditions in the soil are such that mass flow fails to meet plant demand, NO_3^- will also move by diffusion. In contrast to the depletion zone for phosphate however the resulting depletion zone for NO_3^- tends to be both large and shallow. This means that even if movement is via diffusion, NO_3^- can still be effectively acquired by the plant. Hence, in the case of NO_3^- the problem is not movement to the root *per se* but if there are sufficient quantities of NO_3^- present in the soil.

Root responses

All plants face the same basic economic problem of where best to place their resources but are aided in this by being able to show a high degree of morphological and physiological plasticity at several organization levels: from cellular to between organs. Root systems are modular which allows them to be very plastic in their response to prevailing environmental factors, including the distribution of nutrient resources. When nutrient availability is low, plants generally increase their root weight ratio (RWR; ratio of root weight to total plant weight) because more biomass is allocated to the root system. However, the RWR is not necessarily the best root trait to measure for assessing the ability of plants to subsequently capture nutrient resources. The RWR will include old,

lignified roots not actively involved in nutrient uptake *per se*, and also changes in the architecture of the root system (i.e., the spatial configuration of the root system in the soil) in response to nutrient availability can occur independent of changes in root biomass. Uptake of nutrients is more closely related to root length than root mass, and a greater length of finer roots aids in the exploration of the soil environment (Hodge, 2004). Root mineral nutrient uptake is largely reliant on the absorption of ions from the soil that are transported to them by diffusion and mass flow. Beneficial physiological changes in ion uptake rates by roots can enhance uptake of mobile ions such as nitrate. In contrast, more immobile ions (e.g., phosphate) are limited more by diffusion to the root rather than uptake at the root surface so would be less affected by changes in root uptake rates. Although NO_3^- will move toward the root rapidly via mass flow if water flux and ion concentrations are large enough, if this is not the case then NO_3^- ions will move to the root via diffusion (i.e., down a concentration gradient). The diffusion coefficient for NO_3^- in the soil environment is approx. $10^{-10} \text{ m}^2 \text{ s}^{-1}$. In contrast, the diffusion coefficient for phosphate is approx. 10^{-13} – $10^{-15} \text{ m}^2 \text{ s}^{-1}$. The smaller diffusion coefficient of phosphate is due to the reaction of phosphate ions with charged surfaces and insoluble complexes formed with cations such as Al^{3+} , Fe^{3+} , and Ca^{2+} . Diffusion coefficients for NH_4^+ range between approx. 10^{-10} and $10^{-12} \text{ m}^2 \text{ s}^{-1}$, but the availability of NH_4^+ is further reduced by sorption in soil. Thus, the distribution of NO_3^- in soils should be more uniform than that of phosphate and, in principle, even a relatively low root length density (RLD) should be enough to capture sufficient N from the soil when present in the form of NO_3^- . In actuality, plants have to capture N, and indeed other nutrients, in competition with other plant roots as well as soil microorganisms, also a large number of inputs in natural systems occur in organic form via leaf litter, dead roots, and animals, leading to a heterogeneous distribution of resource supply both spatially and temporally. Thus, the distribution of resources in soil, even of NO_3^- , is not uniform but heterogeneous or patchy.

Root systems can respond to nutrient-rich zones or patches through proliferation of roots as demonstrated in the classic experiments conducted by Drew and colleagues in the 1970s. Barley roots exposed to a localized supply of nitrate, ammonium, and phosphate increased both the initiation (i.e., number) and extension (i.e., length) of first- and second-order laterals within that localized zone. Moreover, shoot weight was only slightly depressed when these nutrients were supplied locally compared with controls receiving a uniformly high supply of these nutrients throughout their entire root system. However, Drew's experimental system involved plants grown in sand units supplied with nutrient solutions and did not reflect the complexity of the soil environment. Moreover, responses observed under ideal, and often controlled, conditions may not necessarily relate to those that actually occur in the natural environment. Nevertheless, the study demonstrated that roots respond, often spectacularly, to inorganic N as well as phosphate-rich zones (Drew, 1975; Fig. 2).

The benefit of root proliferation in terms of increasing acquisition of immobile ions from the soil is relatively easy to interpret given the increased volume of soil explored. For mobile ions placing resources into increased root growth may appear counter-intuitive despite such root responses being widely reported. However, when two different plant species were grown in competition for resources from an organic N-rich patch (grass shoots dual-labelled with the stable isotopes ^{15}N and ^{13}C) added to soil, plant ^{15}N capture was related to RLD produced in the patch. No ^{13}C enrichments were detected in the plant tissue, therefore N must have been taken up as mineralized N (Hodge et al., 1999). Thus, when in competition for N, even when the N is in mineralized form and potentially highly mobile, RLD is important because the supply of N from the decomposing organic patch is finite. In contrast, in the absence of plant–plant competition, the supply of N can be thought of as invariant and RLD is less important.

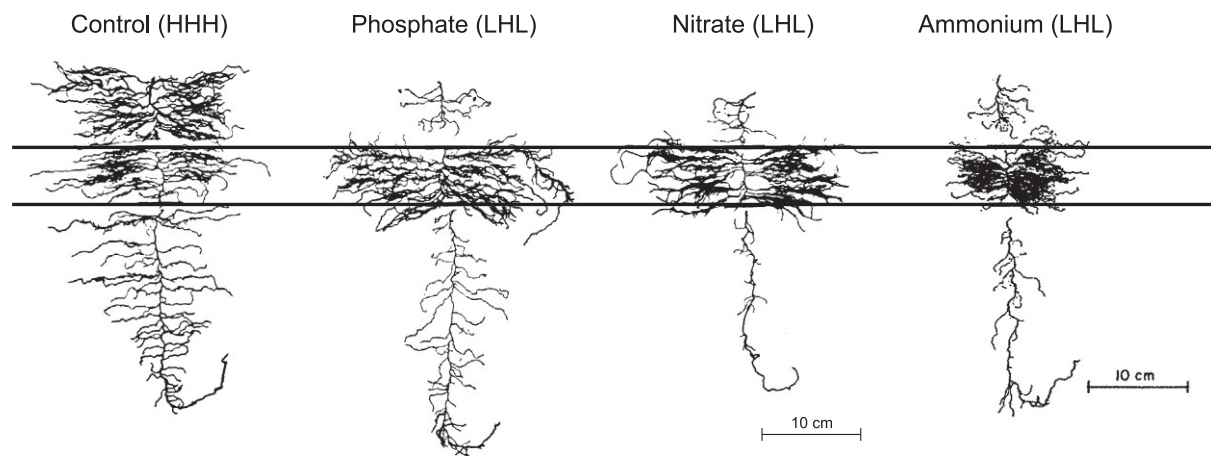


Fig. 2 The proliferation of primary and secondary laterals of barley (*Hordeum vulgare*) root systems when supplied with a 100-fold greater concentration of phosphate, nitrate or ammonium to the middle part of their root system (indicated by the lines). H, high; L, low, which refer to the nutrient concentrations of phosphate, nitrate or ammonium received by the top, middle and bottom of the root system, respectively. The control which received uniformly high levels of nutrients to all parts of its root system is shown as a comparison. Reproduced from Drew MC (1975) Comparison of the effects of a localized supply of phosphate, nitrate, ammonium and potassium on the growth of the seminal root system, and the shoot, in barley. *New Phytologist* 75: 479–490, with kind permission of the trustees of the *New Phytologist* and Wiley-Blackwell.

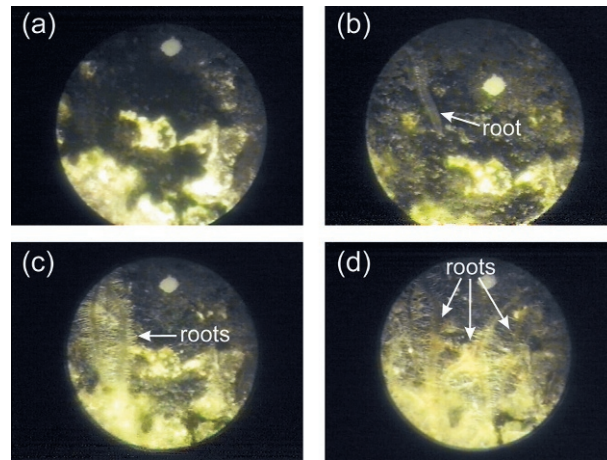


Fig. 3 Root growth and proliferation in an N-rich organic material source added as a spatially discrete patch and comprising lyophilized algal cell organic material (C:N ratio of 3.2:1). (a) One day after the patch was added no roots in the patch zone were observed (b) 21 days after the patch was added, a single root was observed growing into the N-rich zone. (c) The same image 25 days and (d) 48 days after the patch was added: increased root production and root hair development in the N-rich zone was clearly observed. After 49 days, the plants were harvested and found to contain 54% of the N originally added in the algal cell patch in their tissues. Adapted from Hodge A, Robinson D, and Fitter A (2000) Are microorganisms more effective than plants at competing for nitrogen? *Trends in Plant Science* 5: 304–308, with kind permission from Elsevier Science.

The timescales to construct new roots in these nutrient-rich zones are generally weeks rather than days (Fig. 3). In contrast, microbial populations can respond much more rapidly in timescales of hours to days. Thus, root proliferation responses are considered to be more important in terms of plant–plant competition rather than plant–microbial competitive interactions. High-order laterals produced upon root proliferation also tend to have higher N concentrations and thus are more expensive to construct but presumably the benefits of exploiting a nutrient-rich patch and the capture of N from it outweigh the disadvantages of construction costs for these roots. Moreover, increased root turnover, i.e., a decrease in root longevity, is often observed in both fertile soils and fertile patches, but again presumably the advantages of having a young population of roots actively involved in nutrient-uptake processes outweigh the disadvantages of high root turnover rates. Plants can extract nutrients, including N, from root structures shortly after construction thus also considerably lowering initial construction costs.

A number of other root traits have also been demonstrated to be important for exploiting the soil environment for particular nutrient resources, although most research has focused upon the acquisition of the macro-nutrients P and N. The ideotype root traits for enhanced phosphate capture may include shallow basal roots with small diameter, increased specific root length (SRL), long and dense root hairs and arbuscular mycorrhizal (AM) associations. Certainly, inhibition of primary root growth, but increased second-order lateral root growth resulting in a more branched root system with roots concentrated in the upper soil layers where most P sources are located, is frequently reported under low P conditions. Plants may also induce high affinity P transporters in roots and alter their internal biochemical pathways to conserve P when under P-stress conditions. Some plants may also release H^+ ions under P-stressed conditions which is suggested to increase the solubility of calcium phosphates through acidification in the locality of the rhizosphere (Hinsinger, 2001) although if soils become too acidic this also presents problems for the plant as previously discussed.

Some plants, such as *Brassica napus*, show a large increase in the release of organic acids such as the carboxylic acids, malic and citric acid from their roots under low P availability. This could aid in P acquisition through the carboxylic acids chelating with cations such as Al^{3+} , Fe^{3+} or Ca^{2+} and so releasing bound phosphates associated with these cations into the soil environment for subsequent plant capture. While in other plant species, such as those that can form cluster roots (i.e. short lateral rootlets that are closely spaced) or the functionally similar dauciform roots (i.e. short lateral roots with dense clusters of root hairs), there is a marked, tightly regulated release of carboxylates including malic and citric acid that displace the held P so that it can be taken up by the dauciform or the cluster roots aided by the large surface of the clusters of root hairs (dauciform roots) or rootlets (cluster roots). Cluster roots may also release enzymes that enable organic P forms in the soil to be broken down while the carboxylic acids may also play a role in the hydrolysis of organic P forms. Thus, this would give these plant species access to P sources not normally available. This P acquisition strategy including the directed release of carboxylic acids is costly to the plant but it does allow these plants to extract P from forms that would not normally be available and hence allows these plant species to grow in extremely P-impoverted soils.

In contrast to the root traits that help increase P acquisition, the root trait ideotype for N acquisition involve a steep growth angle of crown roots, architectural and anatomical traits that reduce the cost of exploration and deep root placement in the soil to capture leached N, or ‘steep, cheap and deep’ root traits (Lynch, 2013). As these are ideotypes not all plant species show all these traits and there may be further trade-offs for example, in the case of P acquisition between root hair density and the formation of mycorrhizal associations. Moreover, plants need both N and P as well as the other macro- and micro-nutrients of varying mobility for normal

plant growth and function. Thus, further research is required to ensure breeding for the most appropriate traits for the particular environmental conditions in question but, nevertheless, root trait selection for crops that reduce the cost of soil exploration while promoting climate-resilience are urgently required to help ensure long term food security (Lynch et al., 2022).

Many plants rely on associations with mycorrhizal fungi to help acquire nutrients. Mycorrhizal fungi are important ecosystem engineers and can provide a range of benefits for their host plant including enhanced water relations, improved soil aggregation and reduced disease occurrence in addition to nutritional benefits. It is not surprising therefore that promotion of mycorrhizal associations is often suggested as a means to promote agriculture sustainability. Mycorrhizal associations are widespread with greater than 80% of plant species forming some type of mycorrhizal association. The arbuscular mycorrhizal (AM) association between fungi of the Mucoromycota subphylum Glomeromycotina and about two-thirds of all land plants is the most ubiquitous, and, together with mycorrhiza-like associations with fungi of the Mucoromycota subphylum Mucoromycotina, are contemporaneous with the early land plants (Feijen et al., 2018). It is well established that AM fungi aid their host plant acquire poorly mobile nutrients such as phosphate through the small diameter fungal hyphae exploring a larger volume of soil, including accessing soil pores that would otherwise be inaccessible to plant roots, at a reduced C cost compared to new root construction (Hodge, 2017). The inorganic P then has to be taken up by the AM hyphae (or indeed the roots) against a steep electrochemical gradient due to the difference in concentrations between that in the soil versus those inside the AM hyphae or roots. This is achieved by H⁺-phosphate co-transport via high affinity phosphate transporters (as shown in Fig. 1). Under P-deficient conditions AM colonized plants generally show increased growth compared to their non-mycorrhizal counterparts but when P levels improve then AM plants may do no better and may even show repressed growth. However, even when total plant P is not enhanced the contribution of the AM fungi to plant P acquisition may still be substantial. For example, while some plants demonstrate no growth response, or improved total P acquisition, as a result of colonization by different AMF fungi, application of the radioisotope, ³³P, revealed that the AM fungi may still be responsible for most, if not nearly all, of the P acquired by the plant (Smith et al., 2004). AM fungi can also help plants acquire N, predominantly in inorganic forms. The values of the contribution of this N to total shoot N however vary considerably: from a few percent to c. 75% depending on the environmental conditions including that of water availability (see Hodge and Storer, 2015 for a review on this topic). AM associations can also help with plant acquisition of other nutrients such as sulfur and zinc, but again values vary considerably among studies suggesting diversity among AM fungi in their nutritional benefit which may also differ depending upon the prevailing environmental conditions (Giovannetti et al., 2017; Cavagnaro, 2008).

Over evolutionary time, various transitions have resulted in the occurrence of other mycorrhizal associations, such as the ectomycorrhizal (EcM), ericoid (ErM) and orchid types, adapted to the prevailing environmental conditions although the AM symbiosis remains the most prevalent among plant species. In contrast to the AM association, the EcM association involves a much smaller number of plant species (c. 6000), but a much larger number of fungal species (c. 20,000). Most of these plant species are woody species, hence, although the EcM association only involves c. 2% of plant species (van der Heijden et al., 2015), it is of economic as well as ecological importance. The fungal species involved in the EcM symbiosis are from the Basidiomycota and Ascomycota fungal phyla and have evolved multiple times from saprotrophic ancestors. Consequently, the capability of these fungi to acquire nutrients from complex organic sources is equally diverse with some (but by no means all) having the ability to operate oxidative decomposition pathways to release nutrients including N and P for their host plant. The ericoid (ErM) association has an even more limited host plant range and involves only c. 1% of plant species (comprising mainly members of the Ericaceae and some liverworts). The fungi are mainly from the Ascomycota but also include some Basidiomycota. These fungi have been demonstrated to access organic forms of nutrients through their enzymatic capabilities reflecting the nutrient sources prevalent in the environments in which this association is found. Thus, in both the ErM and EcM associations the mycorrhizal fungi may allow the plant access to nutrient sources that would otherwise be inaccessible.

Conclusion

Variability in nutrient supply among and within soils has a large impact upon plant growth. Nutrient ions which are essential for plant growth and development, have widely contrasting mobilities depending on the degree of interaction with the prevailing soil environment. For some ions that barely interact with charged surfaces in the soil the impact is slight, while for others movement is severely impeded. Ions with high mobility can move readily to the root surface via mass flow providing water flux and ion concentrations are high enough, otherwise movement is via the much slower pathway of diffusion. Diffusion down a concentration gradient is the main way in which immobile ions move to the root. However, as diffusion is a slow ion movement pathway, ions do not arrive fast enough to satisfy plant demand so a depletion zone develops around the root. The size of this depletion zone is important as it determines the volume of soil the root effectively has access to for that particular ion. It is therefore unsurprising that plants have evolved a number of strategies to help them acquire limiting nutrients from their environment. These include changes in the root system architecture (RSA), chemical extraction and associations with microorganisms from the relatively loose association with PGPRs in the rhizosphere to the tight symbiotic associations with nitrogen-fixing bacteria and mycorrhizal fungi. These strategies also represent targets for enhancing sustainable crop food security in the future through, for example, the application of microbial inoculants or genetic modification of RSA. If these strategies are successful or not will depend upon if the environmental context (i.e., organic versus inorganic additions, mono- versus mixed cropping, multi-ion environment and so forth) is also considered as not all will work in all types of soils and under all circumstances.

References

- Barber SA (1974) Influence of the plant root on ion movement in the soil. In: Carson EW (ed.) *The Plant Root and Its Environment*, pp. 525–563. Virginia: University Press.
- Brödlin D, Kaiser K, and Hagedorn F (2019) Divergent patterns of carbon, nitrogen and phosphorus mobilisation in forest soils. *Frontiers in Forests and Global Change* 2: e66.
- Brown PH, Welch RM, and Cary EE (1987) Nickel: A micronutrient essential for higher plants. *Plant Physiology* 85: 801–803.
- Cavagnaro TR (2008) The role of arbuscular mycorrhizas in improving plant zinc nutrition under low soil zinc concentrations: A review. *Plant and Soil* 304: 315–325.
- Clarholm M (1985) Interactions of bacteria, protozoa and plants leading to mineralization of soil nitrogen. *Soil Biology and Biochemistry* 17: 181–187.
- Coskun D, Deshmukh R, Sonah H, et al. (2019) The controversies of silicon's role in plant biology. *New Phytologist* 221: 67–85.
- Crocker RL and Major J (1955) Soil development in relation to vegetation and surface age at Glacier Bay, Alaska. *Journal of Ecology* 43: 427–448.
- Drew MC (1975) Comparison of the effects of a localised supply of phosphate, nitrate, ammonium and potassium on the growth of the seminal root system, and the shoot, in barley. *New Phytologist* 75: 479–490.
- Epstein E (2009) Silicon: Its manifold roles in plants. *Annals of Applied Biology* 155: 155–160.
- Feijen FAA, Vos RA, Nuytinck J, and Merckx VSFT (2018) Evolutionary dynamics of mycorrhizal symbiosis in land plant diversification. *Scientific Reports* 8: 10698.
- Fitter AH and Hay RKM (2002) *Environmental Physiology of Plants*, 3rd edn London: Academic Press.
- Giovannetti M, Volpe V, Salvioli A, and Bonfante P (2017) Fungal and plant tools for the uptake of nutrients in arbuscular mycorrhizas: A molecular view. In: Johnson NC, Gehring C, and Jansa J (eds.) *Mycorrhizal Mediation of Soil: Fertility, Structure, and Carbon Storage*, pp. 107–128. Elsevier: Amsterdam.
- Greenfield LM, Hill PW, Paterson E, Baggs EM, and Jones DL (2020) Do plants use root-derived proteases to promote the uptake of soil organic nitrogen? *Plant and Soil* 456: 355–367.
- Hetherington AM and Woodward FI (2003) The role of stomata in sensing and driving environmental change. *Nature* 424: 901–908.
- Hinsinger P (2001) Bioavailability of soil inorganic P in the rhizosphere as affected by root-induced chemical changes: A review. *Plant and Soil* 237: 173–195.
- Hodge A (2004) The plastic plant: Root responses to heterogeneous supplies of nutrients. *New Phytologist* 162: 9–24.
- Hodge A (2017) Accessibility of inorganic and organic nutrients for mycorrhizas. In: Johnson NC, Gehring C, and Jansa J (eds.) *Mycorrhizal Mediation of Soil: Fertility, Structure, and Carbon Storage*, pp. 129–148. Elsevier: Amsterdam.
- Hodge A and Storer K (2015) Arbuscular mycorrhiza and nitrogen: Implications for individual plants through to ecosystems. *Plant and Soil* 386: 1–19.
- Hodge A, Stewart J, Robinson D, Griffiths BS, and Fitter AH (1998) Root proliferation, soil fauna and plant nitrogen capture from nutrient-rich patches in soil. *New Phytologist* 139: 479–494.
- Hodge A, Robinson D, Griffiths BS, and Fitter AH (1999) Why plants bother: Root proliferation results in increased nitrogen capture from an organic patch when two grasses compete. *Plant, Cell and Environment* 22: 811–820.
- Hodge A, Robinson D, and Fitter AH (2000a) Are microorganisms more effective than plants at competing for nitrogen? *Trends in Plant Science* 5: 304–308.
- Hodge A, Stewart J, Robinson D, Griffiths BS, and Fitter AH (2000b) Plant N capture and microfaunal dynamics from decomposing grass and earthworm residues in soil. *Soil Biology and Biochemistry* 32: 1763–1772.
- Hodge A, Stewart J, Robinson D, Griffiths BS, and Fitter AH (2000c) Spatial and physical heterogeneity of N supply from soil does not influence N capture by two grass species. *Functional Ecology* 14: 645–653.
- Hodge A, Stewart J, Robinson D, Griffiths BS, and Fitter AH (2000d) Competition between roots and soil microorganisms for nutrients from nitrogen-rich patches of varying physical and chemical complexity. *Journal of Ecology* 88: 150–164.
- Houlton BZ and Morford SL (2015) A new synthesis for terrestrial nitrogen inputs. *The Soil* 1: 381–397.
- Houlton BZ, Morford SL, and Dahlgren RA (2018) Convergent evidence for widespread rock nitrogen sources in Earth's surface environment. *Science* 360: 58–62.
- Lynch JP (2013) Steep, cheap and deep: An ideotype to optimize water and N acquisition by maize root systems. *Annals of Botany* 112: 347–357.
- Lynch JP, Mooney SJ, Strock CF, and Schneider HM (2022) Future roots for future soils. *Plant, Cell and Environment* 45: 620–636.
- Paungfoo-Lonhienne C, Lonhienne TGA, Rentsch D, et al. (2008) Plants can use protein as a nitrogen source without assistance from other organisms. *Proceedings of the National Academy of Sciences U. S. A.* 105: 4524–4529.
- Schachtman DP, Reid RJ, and Ayling SM (1998) Phosphorus uptake by plants: From soil to cell. *Plant Physiology* 116: 447–453.
- Silvertown J, Poulton P, Johnston E, et al. (2006) The Park Grass experiment 1856–2006: Its contribution to ecology. *Journal of Ecology* 94: 801–814.
- Smith SE, Smith FA, and Jakobsen I (2004) Functional diversity in arbuscular mycorrhizal (AM) symbiosis: The contribution of the mycorrhizal P uptake pathway is not correlated with mycorrhizal responses in growth or total P uptake. *New Phytologist* 162: 511–524.
- Tinker PB and Nye PH (2000) *Solute Movement in the Rhizosphere*. New York: Oxford University Press.
- van der Heijden MGA, Martin FM, Selosse M-A, and Sanders IR (2015) Mycorrhizal ecology and evolution: The past, the present, and the future. *New Phytologist* 205: 1406–1423.

Relevant websites

<https://www.rhs.org.uk/prevention-protection/nutrient-deficiencies>—More information regarding nutrient deficiencies in plants.

<http://www.rothamsted.ac.uk/long-term-experiments>—Information regarding the long term experiments in place at Rothamsted Research.

Soil Biology: Root form and function

Magda Garbowski^a, Grégoire Freschet^b, Louise Jackson^c, Cynthia Brown^d, and Louise Comas^e, ^aDepartment of Botany, University of Wyoming, Laramie, WY, United States; ^bCentre National de la Recherche Scientifique, Station d'Écologie Théorique et Expérimentale, Moulis, France; ^cDepartment of Land, Air, and Water Resources, University of California, Davis, CA, United States; ^dAgricultural Biology, Colorado State University, Fort Collins, CO, United States; ^eUnited States Department of Agriculture—Agricultural Research Service, Fort Collins, CO, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of LE Jackson, Root architecture and growth, Editor(s): D Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 411–421, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00164-8>

Introduction	321
Root anatomy, development, and classification	322
Trade-offs in root traits	324
Root traits for stress tolerance	325
Genetic analysis of root traits	327
Root trait effects on plant communities and ecosystem processes	328
Conclusion	330
References	330

Abstract

Research focused on root traits and functions has expanded in recent decades and findings have identified how specific traits influence plant performance under environmental stressors, including drought, compaction, nutrient limitation, and salinity. Integrating global information has advanced understanding of how roots drive ecosystem functions, such as productivity, water and nutrient cycling, and erosion. Studies on anatomy and control of root growth are elucidating root response to changing environments. Modern methods in quantitative genetics and genomics enhance understanding of genetic underpinnings of root traits and functions. Advanced understanding of root form and function has important implications for ecology and land management.

Key points

- Roots are essential components of plant strategies.
- Growth and development of roots is governed by genetics and hormones and root responses to their environments.
- Trade-offs in root form and function influence how plants interact with soils and cope with environmental stressors such as drought, salinity, and nutrient limitation.
- Root traits influence various ecosystem functions such as nutrient cycling, productivity, and erosion.

Introduction

Many aspects of root form and function, including root architecture, morphology, physiology, and growth, are important for exploration of soil for water and nutrients. Root systems are comprised of segments with unique properties and a complexity that is often under appreciated. Research continues to uncover distinctive functions associated with root segments that vary in anatomy, morphology, and physiology. Several important concepts govern the development of roots and their direct and indirect effects on plant growth:

- 1) Plants construct roots of various forms and characteristics to optimize resource acquisition in particular conditions.
- 2) Trade-offs exist in root form and function with the costs of producing roots yielding benefits in terms of the acquisition of soil water and nutrients and ultimately, plant growth, and fitness.
- 3) Enormous variation exists among plant species in their root responses to abiotic (e.g., drought, salinity) and biotic (e.g., competition, pathogens) stress.
- 4) Genetic analyses have revealed that root architectural traits are often multi-gene or quantitative traits.
- 5) Root traits and functioning influence attributes of soil environments, plant communities, and ecosystems.

Root anatomy, development, and classification

Given the unique viewpoints from which researchers study roots, numerous schemes have been adopted to classify components of root systems. Here we first briefly discuss root anatomy, developmental-based schemes of root classification, and hormonal control of root growth. We then highlight functional traits and functional classifications based on root orders.

A cross-section of the mature root has several distinctive zones (Fig. 1a; Ranathunge, 2005), with characteristics that vary by root age, among different groups of plants, and among species. The epidermis is the outer layer of the root, which often has a protective exodermis immediately underneath it. The exodermis is composed of very compact cells with a Casparian strip composed of waxy substances, and suberin lamellae composed of glycerolipid polymers, both of which curtail apoplastic water flow, confining the flow of water to the symplast. This increases the selective uptake of nutrients, limits water flux out of the root, protects from microorganism intrusion, and further increases root mechanical resistance. These layers surround the cortex, which stores nutrients and supports colonization by mycorrhizal fungi (Comas et al., 2012). The endodermis is the innermost layer of the cortex, which also harbors a Casparian strip. Within the vascular cylinder, xylem tissues, including hollow tracheids, are important for transport of water and dissolved minerals to upper parts of the plant, and phloem elements, which conduct sugars and organic compounds from leaves to roots and other parts of the plant. In monocotyledon plants vascular bundles are organized in a circular pattern around a central pith. In dicotyledon plants xylem elements are present in the center of the root and are separated from phloem bundles by vascular cambium. The physical arrangement of xylem and phloem tissues can further vary among species. Within xylem tissues, protoxylem is the first formed elements of the primary xylem. Metaxylem is the part of the primary xylem that differentiates after the protoxylem but before the secondary growth if plants undergo secondary development. The pericycle is the outermost boundary of the vascular cylinder, located inside the endodermis. Lateral roots are initiated from the pericycle, and then grow outwards through the cortex and epidermis. The vascular cambium, which is responsible for secondary growth in roots, also originates within the pericycle in woody plants.

Root development begins at embryogenesis when embryonic plant cells give rise to seminal or primary roots. Many dicotyledonous plants develop one primary root, which thickens in some species due to secondary cambial activity, whereas most monocots develop several primary roots that do not undergo secondary growth. Embryonic plant cells also give rise to the quiescent centers and central root caps of these first roots. Several key zones become present in developing roots and remain present in root tips or apical zones as plants mature (Fig. 1b; Crang et al., 2019): the root cap that protects the meristem, the meristem itself, a region of elongation, and a region of maturation. Several processes are concentrated in this apical zone, including root elongation, the maturation of cells, and the development of vascular tissues, organs (e.g., root hairs, xylem elements, phloem elements), and lateral primordia, which give rise to lateral roots.

Roots branch in response to endogenous and exogenous factors. Numerous phytohormones, as well as biomolecules, regulate root growth and development (Jung and McCouch, 2013; Mukherjee and Baluška, 2021). The plant hormone, auxin, has a key role in the initiation of lateral root primordia. It can accumulate at the root apex, which usually exerts some degree of apical dominance,

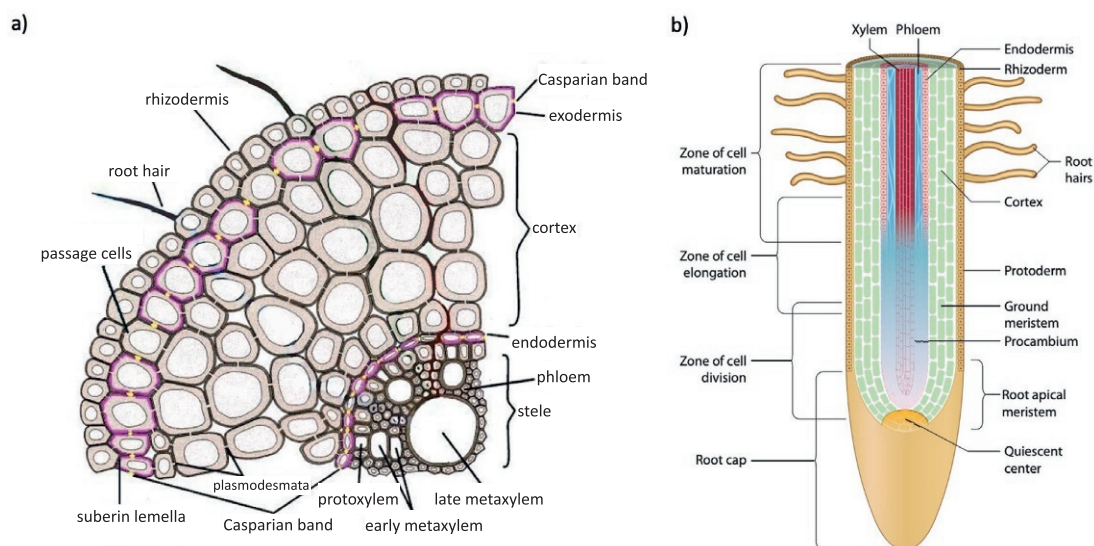


Fig. 1 (a) Horizontal and (b) longitudinal cross sections of a root showing the different tissues, root sections, root features, and xylem and phloem elements. The outermost layer is the epidermis and root hairs, and the protective exoderms. In the cortex, there are intercellular air spaces between the cells. The endodermis surrounds the vascular cylinder, which contains the xylem and phloem. (A) Modified from Ranathunge KM (2005) *The Role of the Apoplastic Transport Barriers for Radial Water and Ion Uptake in Rice (*Oryza sativa* L.) and corn (*Zea mays* L.) Roots*. Doctoral thesis, University of Bayreuth, Faculty of Biology, Chemistry and Earth Sciences. (B) From Crang R, Lyons-Sobaski S, Wise R (2019) Roots. In: Crang R, Lyons-Sobaski S, Wise R (eds.). *Plant Anatomy: A Concept Based Approach to the Structure of Seed Plants*. Springer. "Roots" in Plant Anatomy.

affects developing lateral root primordial, or both. Cytokinins, another class of plant hormones, inhibit lateral root formation. Low nutrient levels, especially low nitrogen levels, can act as stress signals that decreases cytokinin content, thus, increase root growth by releasing the inhibition on lateral root formation. The ratio of auxin and cytokinin concentrations is thought to be important for the regulation of developmental steps for the complete formation of lateral roots. Ethylene, a plant hormone that is involved in root response to stress, can either enhance or repress root growth, depending on its concentration. Lateral root formation and the relative growth rate of roots are stimulated by nitrate, ammonium, and phosphate, such that highly branched roots and greater root length density are found in enriched soil patches (Hodge, 2004). For immobile ions like phosphate, or poorly mobile ions such as ammonium, local root proliferation is important to increase nutrient exploitation from such patches. In contrast, a mobile ion such as nitrate may be more likely captured by a root system with a more extensive growth (i.e., explorative roots with small root branching density). Plant hormones and other biomolecules clearly have important effects on root branching and growth and the complexity of their interactions continue to be discovered.

According to a development-based classification scheme, root systems are composed of seminal roots that originate as part of the developing embryo in the seed (from only one seminal root in most eudicots to several in monocots), postembryonic shoot-borne roots (or adventitious, crown, or nodal roots), and lateral roots that emerge from all root types. In monocotyledonous plants, such as grasses, seminal roots grow out of the developing embryo, while shoot-borne roots subsequently arise from the shoot, and gradually replace the seminal axes. In contrast, in dicotyledonous plants, the entire root system originates from a single root in the embryo, from which laterals arise, and later can become a taproot, or thicken and become woody due to cambial activity. Numerous structures and entities comprise whole-plant root systems. Key entities of roots and their main functions are discussed here, but we refer the reader to Box 1 of Freschet et al. (2021a) for a more comprehensive list.

Root hairs are extensions of epidermal cells (i.e., trichoblasts) originating from the differentiation zone of root tips and vary in length from 0.2 to 2 mm depending on the species. By increasing the effective radius of the root, root hairs improve the acquisition of relatively mobile (e.g., nitrate) and immobile nutrients (e.g., phosphate) as well as water. Given an inverse relationship between root hair and mycorrhizal abundance, root hairs are thought to serve similar roles in resource acquisition as mycorrhizal hyphae. In addition, root hairs influence the formation of rhizospheres (i.e., soil being influenced by roots) and rhizosheaths (i.e., soil particles adhered to roots) to further regulate interactions among roots and their environments (York et al., 2016). Associated processes such as nutrient absorption kinetics, root exudates, and microbial activity in the rhizosphere affect nutrient availability and therefore plant uptake of nutrients from the soil.

Numerous other structures and unique root types are associated with specific functions. Mycorrhizal roots and root nodules are essential to plant-fungi and plant-bacteria associations, which are key drivers of resource acquisition (Bergmann et al., 2020; Tedersoo, 2017). Rhizomes, stolons, and tubers, which are not truly root structures but rather underground stems, along with true root clonal structures (i.e., root spacers) allow plants to spread clonally and regrow after disturbances (Klimešová et al., 2019). Sinker roots, buttress, and stilt roots are adapted for increased anchorage and support.

Root classification methods use the branching of root systems to describe and compare roots among and within species (Fig. 2). To classify roots using these methods roots or root segments can be counted starting from the outermost (i.e., distal) roots increasing counts inward (e.g., functional, morphometric, stream-order, centripetal or topological link-based approaches) or starting from basal (i.e., proximal) roots increasing counts outward (e.g., the developmental or centrifugal approach). Morphometric approaches are useful in situations where extracting whole root systems is challenging (e.g., woody or perennial species) and allows for functional classification of root orders, whereas the developmental approach is used to characterize whole root systems.

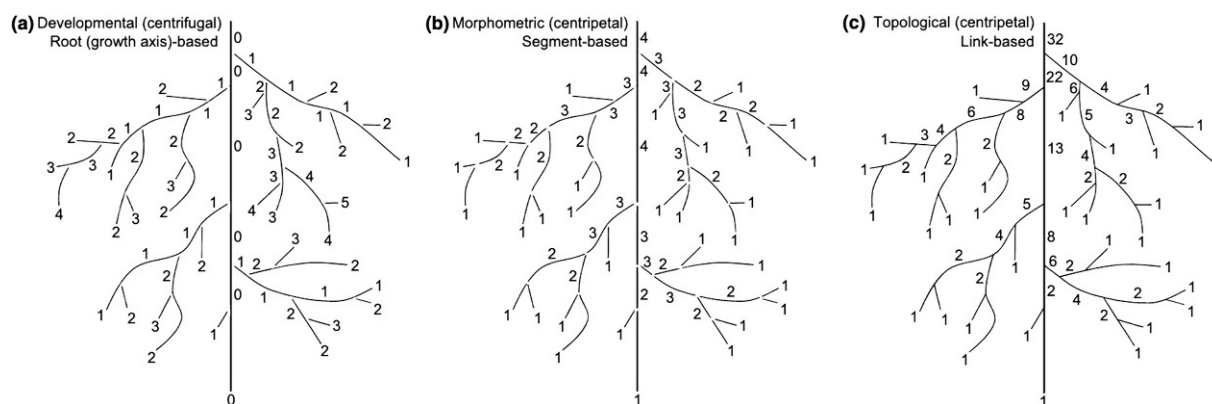


Fig. 2 Schematic presentations of systematic root classification approaches. (a) Developmental approach; (b) morphometric approach; and (c) topological approach. Root growth axis (a) and root order (b, c) levels are indicated by adjacent numbers; note that highest root orders are 5, 4 and 32, respectively (a–c), corresponding to 6, 4 and 12 classification levels (a–c). The ‘0’ root order in (a) can be replaced by root class (e.g., basal root, tap root); lower root order in (b) can be refined further, for example by adding information on the mycorrhizal status; the highest order number in (c) represents the numbers of root tips. Figure and caption from Freschet GT, Pagés L, Iversen CM et al. (2021) A starting guide to root ecology: Strengthening ecological concepts and standardizing root classification, sampling, processing and trait measurements. *New Phytologist* **232**: 973–1122. doi: 10.1111/nph.17572.

Roots can also be classified based on diameter classes, but diameter-based classification is now recognized as creating difficulties for distinguishing functional categories needed for many studies. Under diameter-based classification, roots of shared diameter classes are assumed to share specific functions, but roots of similar diameters can have entirely different functions among different species. It is often useful or desirable to separate portions of root systems associated with soil resource acquisition from the portion associated with providing structural support or transport. Although the entire root system in herbaceous plants can function for soil resource acquisition, typically the most distal two orders and a portion of the third order roots typically remain non-woody in woody plants and function as absorptive roots, while higher order roots undergo secondary development and function as structural or transport roots.

Trade-offs in root traits

Trade-offs in root traits can result from individual plant responses to environmental conditions or from selective pressures across generations. The resulting root traits are often optimized to balance resource limitation belowground with light limitation aboveground and to ultimately enhance growth, survival, or fitness. Most simply, the balanced growth hypothesis suggests that total plant growth will be largest when the root system maximizes nutrient and water acquisition per unit resource supplied by the shoot. However, allocation of large amounts of carbohydrates and nutrients to roots may detract from shoot growth, leaf area, and photosynthetic carbon gain, thereby reducing total plant growth, especially if the increased root allocation does not achieve increased rates of nutrient uptake.

Research over the last several decades has identified key trade-offs in aboveground plant form and function associated with fast growth and rapid resource acquisition versus slow growth and longevity. Key trade-offs associated with belowground resource acquisition, root tissue construction, extensiveness of root foraging (amount of root length in soil) and root collaboration with mycorrhizal fungal partners have been identified (Comas et al., 2012; Bergmann et al., 2020; Weigelt et al., 2021). A major gradient termed the “collaboration gradient” opposes “do-it-yourself” plants that have long specific root length and independently acquire resources with “outsourcing” plants that have large root diameters and cortical fractions, which support mycorrhizal associations (Fig. 3). A second major axis mirrors the aboveground “fast-slow” gradient such that one end of the gradient is characterized by root traits that promote high metabolic activity and rapid resource acquisition (e.g., large root nitrogen concentration) and the other end of the gradient is characterized by traits that promote longevity and survival of tissues (e.g., large root tissue density). Correlations between specific root length and root length density have also been identified (Comas et al., 2012), as has an independent axis of rooting depth (Weigelt et al., 2021). However, because data for traits associated with these axes are more limited, the strength and ubiquity of these patterns remain inconclusive. These recent findings have clarified our understanding of belowground trade-offs but further elucidating how plants allocate resources and adjust growth and morphology to optimize the capture of both above and belowground resources is necessary to further advance understanding of whole-plant trait integration.

Plants need to balance above versus belowground investments because acquisition of carbon via light capture and photosynthesis depends on acquisition of nutrients and water belowground, and vice versa. Aboveground, plants often adjust leaf

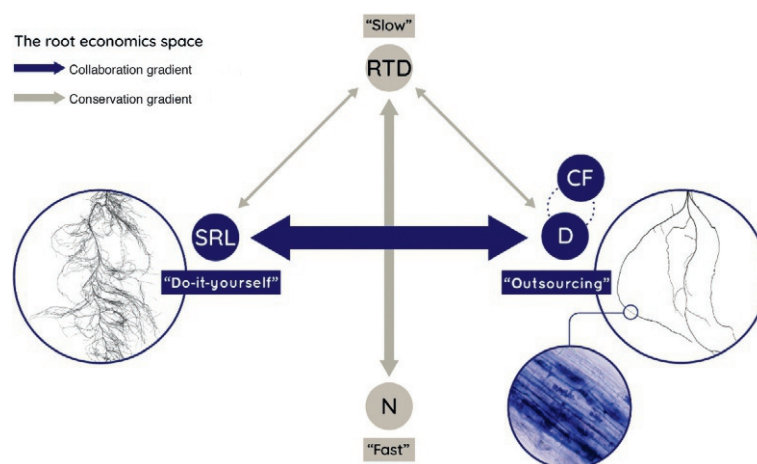


Fig. 3 Conceptual framework of the root economics space depicting (i) a collaboration gradient ranging from do-it-yourself soil exploration using large specific root length (SRL) to outsourcing by investing carbon into a mycorrhizal partner and hence into extraradical hyphae, which requires a large cortex fraction (CF) and root diameter (D) and (ii) a conservation gradient ranging from roots with high root tissue density (RTD) that show a slow resource return on investment but are long-lived and well-protected, to fast roots with a large nitrogen content (N) and metabolic rate for fast resource return on investment but a short life span. Arrows indicate negative correlations between the single traits. Figure and caption from Bergmann J, Weigelt A, van Der Plas F et al. (2020) The fungal collaboration gradient dominates economics space in plants. *Science Advances* 6: eaba3756. doi: 10.1126/sciadv.aba3756.

morphology (e.g., increase specific leaf area), allocation, physiology, or architecture in response to light limitation. Similarly, belowground, root traits and spatial distribution shifts in response to soil resource limitation, with plasticity in allocation and physiology being most critical for improving resource acquisition (Freschet et al., 2018).

Root traits for stress tolerance

Plants cope with abiotic stresses such as drought, nutrient limitation and imbalances, soil compaction, salinity, acid soil, heavy metal toxicity, and waterlogging via various mechanisms and adaptations (Koevoets et al., 2016). Plants must also maintain growth and survive in the face of biotic stresses such as competition, herbivory, and pathogens. Some plant species show highly specific responses and adaptations to stressors, many of which are moderated by root growth, architecture, traits, and physiological mechanisms.

Root systems and root traits are key components of plant strategies for coping with drought. Many drought-resistant plants avoid the negative consequences of moisture stress by optimizing water capture. For example, plants with deep root systems can access water from deep soil layers, and capture water that would otherwise percolate down and be lost from the system. Roots can exploit deep moisture via hydraulic lift, whereby deep moisture is transferred at night through passive physical processes to upper soil layers. Greater root length density and nutrient concentrations in upper soil layers make water and nutrient uptake possible during the following day.

Traits that effectively increase root contact with soil to improve hydraulic continuity have been shown to improve crop performance under drought (Comas et al., 2014). These include thin fine root diameters, long specific root length, high root length densities (especially in layers where soil moisture is greatest), large numbers of root hairs, root tips, and deep roots. Research suggests that many of these traits are regulated by specific genes and breeding for improved production of crops under drought stress is a promising avenue for agricultural research and practice (Comas et al., 2014; Li et al., 2021). Similar root characteristics are likely responsible for promoting plant growth under moderate moisture deficit, particularly in species with short life cycles, short growing seasons, or those species that use pulses of moisture for growth and survival, such as in monsoonal climates.

Numerous root traits are involved with drought tolerant plants strategies, i.e., the ability of plants to maintain low hydration and survive drought. Plants with large root tissue density have been shown to improve survival of perennial grass species in some Mediterranean systems and deep roots have been shown to improve seedling survival in annual systems. Root anatomical traits are also emerging as important drivers of drought-coping strategies in plants. For example, ratios of root tissues (i.e., aerenchyma, cortex, stele or xylem) explain adaptations of grass species to moisture gradients such that small cortex to stele and xylem to stele ratios, or large aerenchyma to stele ratios are advantageous in limited moisture conditions (Fig. 4; Yamauchi et al., 2020). Aerenchyma in extremely dry soil may provide a gap to limit water loss from roots back to the soil. The presence of storage organs and bud banks on belowground roots and stems is also of importance for regrowth of aboveground parts after senescence during a drought period. Despite these recent clarifications about how root characteristics influence drought tolerance, few studies, particularly from non-crop species, focus on root anatomical and physiological mechanisms of drought tolerance. Given that root anatomy governs hydraulic function of roots, further research in this area is needed to understand fully how roots and their functions influence plant performance in response to limited soil moisture.

As in aboveground responses to drought, root growth in response to moisture stress is regulated in large part by hormonal cues. Roots in contact with dry soil can be the source of inhibitors of shoot growth and stomatal opening, which in turn, limits water loss from the plant. Hydraulic signals that elicit physiological responses to drought move from the roots to the shoots in the transpiration stream, and the hormone abscisic acid (ABA), root-derived cytokinins, brassinosteroids, and ethylene phytohormones play a role in this signaling process (Gupta et al., 2020). The accumulation of ABA in roots and aboveground tissues activates signaling of various hormonal pathways and modulates the expression of many genes and physiological processes that protect cells

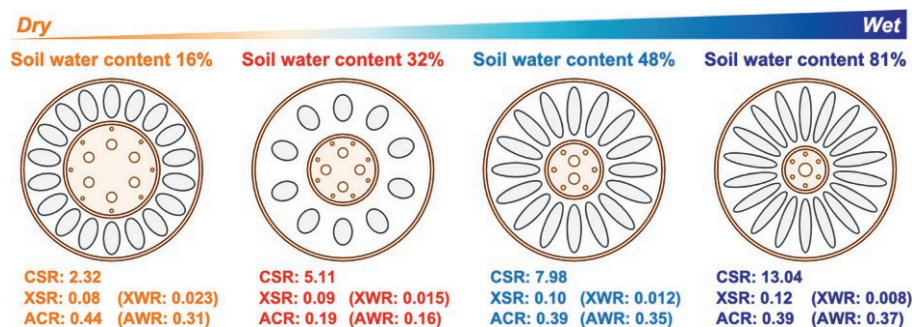


Fig. 4 Adaptive anatomical root traits for the soil water gradients, as predicted by the models of cortex to stele ratio (CSR) (a), xylem to stele ratio (XSR) (b) and aerenchyma to cortex ratio (ACR) (c). The xylem to whole root ratio (XWR) and aerenchyma to whole root ratio (AWR) were calculated by using the values of the CSR and XSR, and CSR and ACR, respectively. Figure and caption from Yamauchi T, Pedersen O, Nakazono M, Tsutsumi N (2020) Key root traits of Poaceae for adaptation to soil water gradients. *New Phytologist* 6(229): 3133–3140. doi: 10.1111/nph.17093.

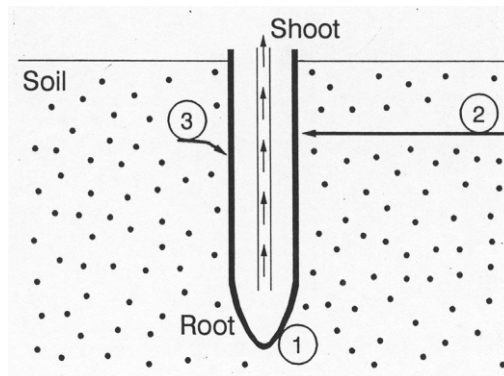


Fig. 5 Schematic presentation of the movement of elements to the root surface of soil-grown plants. (1) Root interception: soil volume displaced by roots. (2) Mass flow: transport of soil solution along the water potential gradient (driven by transpiration). (3) Diffusion: element transport along a concentration gradient. ● = available nutrients (as determined, e.g., by soil testing). Figure and caption from Marschner P and Rengel Z (2012) Nutrient availability in soils. In: Marschner P (ed.). *Marschner's Mineral Nutrition of Higher Plants*, 3rd edn, Elsevier. doi: 10.1016/B978-0-12-384905-2.00012-1.

from excessive water loss. ABA is also involved in hydrotropism, or a root's ability to sense and respond to water potential gradients in soil. Because of its key role in regulating plant responses to drought, the ABA pathway is a key target of many schemes aimed at improving water use efficiency and drought resistance in plants.

Nutrient limitations are well known to alter root growth in ways that increase the capacity of the root system to take up nutrients. Nutrient bioavailability is mediated by root interception, mass flow, and diffusion (Fig. 5; Marschner and Rengel, 2012). As roots grow, they displace soil and intercept available nutrients, for example, adsorbed onto soil particles. This process of root interception, however, is thought to provide only a small proportion of the plant's total nutrient uptake with mass flow, the transfer of water and dissolved nutrients to the root surface driven by water through the plant shoot via transpiration, typically accounting for a larger portion. Rates of diffusion, i.e., transport in soil along the concentration gradient, depend on the mobility and effective diffusion coefficients of specific ions, soil replenishment, and plant uptake. When transpiration rates are high and soil moisture adequate, large quantities of water and dissolved solutes travel to the root via mass flow, but as soils dry and mass flow of water is limited through large soil pores and channels, diffusion of nutrients from proximate soil may facilitate some plant uptake until soil moisture becomes further limited. When moisture is depleted from the soil matrix plant nutrient uptake also ceases.

Both mature and newer regions of the root can play a role in nutrient absorption. Even though older parts of the root system are more suberized and may be surrounded by depletion zones, they have been shown to have high potentials for nutrient uptake in laboratory experiments (Hawkins et al., 2014). Since they constitute a large proportion of the root surface area of the plant, they may be responsible for a considerable proportion of total nutrient uptake. However, the largest rates of nutrient uptake per unit root length are found in the region just behind the root apex, where root hairs are most abundant; more mature roots in older plants generally assume other primary functions (e.g., structural support).

Nutrient-poor soils elicit several changes in root morphology, as well as physiology, but the degree of these responses differs among plant taxa (Freschet et al., 2018). In addition, because individual nutrients have different movement and accumulation patterns in soils, limitations of specific nutrients are mediated via root systems in different ways. Some nutrients, such as phosphate and potassium, have limited mobility and thus accumulate in topsoil after litter decomposition. In contrast, more mobile nutrients such as nitrate and chloride are readily leached and thus accumulate in deeper soil layers. In phosphate limited soils, root growth tends to be concentrated in shallow layers, roots become thinner, and the elongation of laterals increases, especially those at the most distal orders, i.e., the finest, most absorptive roots. More numerous and longer root hairs are produced with faster elongation rates and the greatest benefit for absorption of ions that diffuse slowly in soil. In nitrate limited soils, larger concentrations of the nutrient may be found in deeper soil layers. Roots foraging for nitrate from deeper soil layers often results in greater root:shoot ratios. In addition, roots selectively proliferate in high-nitrate soil patches when nitrate distributions are heterogenous.

Compaction, like soil drying, increases the mechanical impedance to root penetration of soil, affects numerous root traits, and ultimately plant growth and fitness. Compaction increases bulk density, decreases soil porosity, and reduces water and nutrient availability. Because of these changes, compaction influences how and where roots grow. Plants with root systems that are susceptible to compaction will respond by producing lateral roots with shallow angles, and reducing total root length, number, and elongation rate. Plants with root systems more tolerant to compaction may increase root diameters, grow roots with higher tortuosity, and may grow many roots in available larger pore spaces (Correa et al., 2019). While growing through soil, roots must overcome radial, axial, and frictional forces, and there is typically a negative correlation between root elongation rate and penetrometer resistance (Bengough et al., 2011). The osmotic potential of root cells increases in impeded root tips, possibly as a mechanism to increase cell turgor and soil penetrability. Sloughing of the root cap as well as mucilage secretion may protect root meristems and reduce friction between roots and soil particles as they penetrate dense soils. Root impedance by mechanical forces in soil also triggers root signals that reduce leaf expansion and transpiration rates, also in part due to ABA signaling, and ethylene and auxin responses (Jacobsen et al., 2021). Reduced leaf expansive growth and shoot water loss induced by mechanical stress, even

under high nutrient and water supply, may prime plants growing in compacted soil for scarce availability of soil resources resulting from reduced root growth.

Salinity inhibits root elongation growth and development in non-halophytes, and this in turn, indirectly limits the ability of roots to explore soil for nutrients and water, which decreases the growth rate of the plant. Direct effects of salinity on roots include the osmotic stress and low soil water potential that can induce both the biosynthesis of costly organic osmolytes to regulate osmotic adjustment in the root, and the accumulation of specific ions in the root tissue, which can lead to induced toxicities and deficiencies. Salt tolerance is achieved by several potential mechanisms taking place in either shoots or roots. In roots, a ratio of low levels of Na^+ and high levels of K^+ in the cytoplasm can be achieved via selective ion uptake mechanisms or salt excretions, or both. Elevated Ca^{++} concentrations in the soil medium can increase the root elongation rate in some plant taxa in saline soils, possibly because there is less displacement of membrane-associated Ca^{++} by Na^+ than in less salt-tolerant genotypes.

Acid soils limit the availability of some nutrients, such as P, increase the availability and toxicity of heavy metals such as aluminium, and reduce colonization of roots by mycorrhizal fungi, all of which can reduce root growth. Aluminium toxicity decreases both root elongation and cell division, and root cells become shorter and wider giving roots a stubby appearance. The root tips are the primary sites of aluminium toxicity. In many tropical soils that experience low pH and aluminium toxicity, P is either deficient or is readily fixed to soil particles. Thus, the small root length density of plants in response to aluminium toxicity is an important limitation to P uptake and total plant growth. Other heavy metals, such as cadmium, copper, and zinc, also decrease root elongation, and root dry mass to a lesser extent, but the mechanisms for these responses are not clearly understood.

In waterlogged or flooded soils, air is displaced from soil pore spaces, and various degrees of oxygen depletion typically occur in heterogeneous microsites in the soil. Flooding of a previously well-aerated soil can cause immediate cessation of root growth, root respiration, and eventually root death if the lack of oxygenation persists. This is partly due to the production of toxic substances, such as nitrite, or acidification of the root cells due to lactic acid produced during anaerobic fermentation. In grass species adapted to restricted oxygen supply due to flooding, similar to those adapted to low soil moisture, plants tend to have high aerenchyma to stele area ratios and high cortical to stele area ratios for supporting aerenchyma but for different reasons (Fig. 4; Yamauchi et al., 2020). In waterlogged or flooded soils, these characteristics promote oxygen diffusion from the shoot base to root tips.

In some plant species after waterlogging, adventitious roots with well-developed aerenchyma, large air spaces in the root cortex, can emerge from the base of the stem, or existing roots can develop aerenchyma. The formation of aerenchyma under hypoxic conditions is associated with the production and accumulation of the plant hormone ethylene. Aerenchyma increases root porosity, and the long-distance transport of oxygen from the shoot base to apical zones of the root system. Pressurized gas transport, driven by temperature gradients between aboveground aerated parts of the plant and submerged root systems, also exists in some flooding tolerant species.

In addition to abiotic stressors, root systems are key components of plant responses to biotic stressors such as competition, herbivory, and pathogens. Research has linked root traits associated with a high level of root investment (e.g., greater root length density in soil, large root:shoot ratio and total root length) to competitive outcomes among plants, i.e., likely through increasing the ability of a plant to take up soil resources and negatively affect neighbor performance. Root traits also play an important role in defense against herbivory. Root derived toxins, traits associated with root toughness (e.g., structural macromolecules), and rhizosheaths are important components of belowground herbivore defense (Moore and Johnson, 2017).

Mycorrhizal symbioses are key to many plant root strategies and help plants overcome various types of stress (Latef et al., 2016; Tedersoo, 2017). Mycorrhizal colonization increases the surface area of the root system by extending hyphae beyond the root surface, and consequently increases nutrient uptake, especially of immobile ions in soil, e.g., phosphorus, zinc, and copper. Mycorrhizae increase the distance of nutrient exploration away from the root and are important in sites with patchy nutrient distributions. Water uptake is also increased for some plant species via mycorrhizae. There are two broad types of mycorrhizal associations. Arbuscular mycorrhizae (AM) are endophytes with intracellular colonization by Zygomycete fungi, whereas ectomycorrhizae (EcM) are endophytes with extracellular colonization by either Ascomycete or Basidiomycete fungi that form a mantle or sheath enclosing the tip of the rootlet. Dependence on AM tends to be greatest in plants with thicker diameter roots and large cortical areas that can host fungal partners. Many herbaceous species with extensive root length density in soil, long specific root length, and thin roots are facultative mycotrophs that show less colonization with increased phosphorus availability. For EcM roots, the bulky tissue of the mycorrhizal mantle surrounding the root tips serves a role in nutrient storage and in root-fungus contact but does not greatly increase the surface area for nutrient uptake. Instead, the extraradical mycelium that extends from the mantle and their rhizomorphs, conglomerations of hyphae with distinct structural shapes, perform this function. Ectomycorrhizal fungi exude hormones such as ethylene and auxin that stimulate root proliferation and branching for increased fungal colonization (Tedersoo, 2017). Both AM and EcM colonization frequently result in allocation of a smaller proportion of total plant weight to root dry weight. Additionally, most EcM fungi can produce exoenzymes capable of degrading soil organic matter and increasing organic nutrients, whereas this ability is generally absent from AM fungi that primarily acquire inorganic nutrients mobilized by soil saprotrophs. Consequently, EcM associations are more prevalent in nutrient-poor organic soils, whereas AM associations are generally found in relatively more fertile soils.

Genetic analysis of root traits

As with many aspects of plant growth and development, root architectural and morphological traits are usually under the control of a complex gene system in which several genes, rather than a single gene, operate together to create a given phenotype. Multi-gene

control of root traits and root responses make the study of root genetics difficult. Also, roots are hard to study due to the large numbers of plants needed for genetic studies, especially plants grown in soil under field conditions where phenotypic evaluation is most valid.

Methods in quantitative genetics and genomics such as quantitative trait locus (QTL) analysis, next-generation sequencing, and high throughput genotyping have allowed identification of the genetic underpinnings of root architecture and many root traits. New sensing and quantification approaches are being used in rapidly screening multi-dimensional phenotypes at cellular, organ, plant, and population levels to further advance understanding of how phenotypes with specific genes interact with environments (i.e., phenomics). In addition, molecular phenotyping approaches (e.g., transcriptomics, proteomics, metabolomics, and ionomics) provide information about trait inheritance in relation to plant physiology.

QTL associated with enhanced drought performance have been identified in numerous crop species (Siddiqui et al., 2020; Ye et al., 2018). These include QTL linked to rooting depth, root length density, root number, root hairs, root angles, and root xylem. Some of these traits are likely associated with stable QTL and have also been shown to be heritable, indicating that breeding for them is a promising avenue for improved plant performance in diverse agroecosystems. Single-gene traits for root morphological traits have been identified in some plants but are most often studied in a tiny crucifer, *Arabidopsis thaliana*, a model species used by plant geneticists. Mutants for traits such as failure to initiate lateral roots, auxin-related lateral root growth, and variable lateral root initiation in response to the soil environment exist in this species. Transgenic *Arabidopsis* plants have also been used to demonstrate the function of a specific gene controlling the proliferation of lateral roots in nitrate-rich patches. In normal, non-transgenic plants, the expression of this gene is induced rapidly when nitrate is supplied, but in transgenic plants in which the gene is repressed, roots no longer proliferate in response to nitrate. The gene is therefore a key factor in determining the developmental plasticity of roots in response to an environmental cue. Specific genes associated with high investment in roots, rooting depth, and root crown angle have also been identified in wheat, maize, and rice varieties (e.g., Jung and McCouch, 2013).

The strong effect of the abiotic and biotic environment in the soil on root development and branching patterns is important for understanding the inheritance of root traits and their adaptation to specific habitats. For many crops, extensive genotypic variation exists for root system traits among cultivars, and research continues to unravel the genetic basis of these traits for improved breeding. Integration of genes from wild varieties related to plant responses to various stresses such as drought can further improve plant performance for more sustainable use of water and nutrients. The editing of genomes using the clustered regularly interspaced short palindromic repeat (CRISPR) technology is being developed to enhance drought tolerance, yield, and domestication, and root traits influencing these characteristics.

Root trait effects on plant communities and ecosystem processes

Plants shape and alter ecosystems and associated properties and processes through their interactions with one another, with other organisms, and with the environments in which they grow. Root systems and their traits are critical components of these interactions and key drivers of ecosystem functions including productivity, carbon cycling and sequestration, nutrient cycling, water budgets and hydraulic redistribution, and soil aggregation and stability (Fig. 6) (Bardgett et al., 2014; Freschet et al., 2021b).

Recent estimates suggest that approximately one quarter of standing plant biomass exists belowground with the largest ratios between below and aboveground fractions found in grasslands (c. 65%), followed by shrublands (c. 50%) and forests (c. 20%). This belowground biomass contributes substantially to overall plant productivity and carbon sequestration. However, these estimates also vary within each of these ecosystems globally and are differentially impacted by root turnover rates in different ecosystems. For example, ecosystems with high mean annual temperatures tend to have greater turnover rates, such that standing biomass underrepresents total biomass produced more so in these regions. Root biomass, surface area, and distribution through the soil profile are particularly important for measures of fine root productivity and carbon cycling due to their relatively short lifespan, small C:N ratio, which enhances turnover rates, and their role in nutrient absorption, and, thus, net primary production. Various root traits affect carbon in ecosystems by modifying carbon inputs, losses, transformations, and retention in soil. Root traits associated with decelerated cycling of carbon include symbiosis with ectomycorrhizal fungi, slow root turnover rates, and long root lifespans. Roots with low specific root length, small N:C ratio and high lignin concentrations tend to have longer lifespans and slower rates of decomposition. Root respiration and exudation alter carbon storage in soil via their effects on CO₂ flux. Roots also promote specific microbial communities that further alter carbon dynamics with fungal communities tending to sequester more carbon than communities dominated by bacteria.

Spatial and temporal patterns of changes in root growth, architecture, and distribution, among other traits, affect nutrient availability in soil, as well as retention and loss of these resources to the atmosphere, runoff, and leaching. As described earlier, the distribution of roots in response to nutrient availability in different soil layers affects nutrient acquisition throughout the soil profile. In addition, numerous root traits are associated with increased or decreased nitrogen and phosphorus inputs into soils. Roots with large nitrogen concentrations show positive correlations with nitrogen mineralization, nitrification, and negative relationships to nitrogen immobilization. Root associations with mycorrhizas also influence nutrient cycling such that communities dominated by AM fungi primarily utilize inorganic nitrogen forms and have overall faster nitrogen cycling rates, whereas those dominated by EcM fungi are capable of using more complex forms of nitrogen and have slower rates of nutrient cycling. Root phosphorus concentrations can also be key drivers of phosphorus inputs in soils via litter shedding and root turnover, especially in phosphorus limited systems. The solubilization of phosphorus from bedrock by deep roots and associated mycorrhizal fungi is also

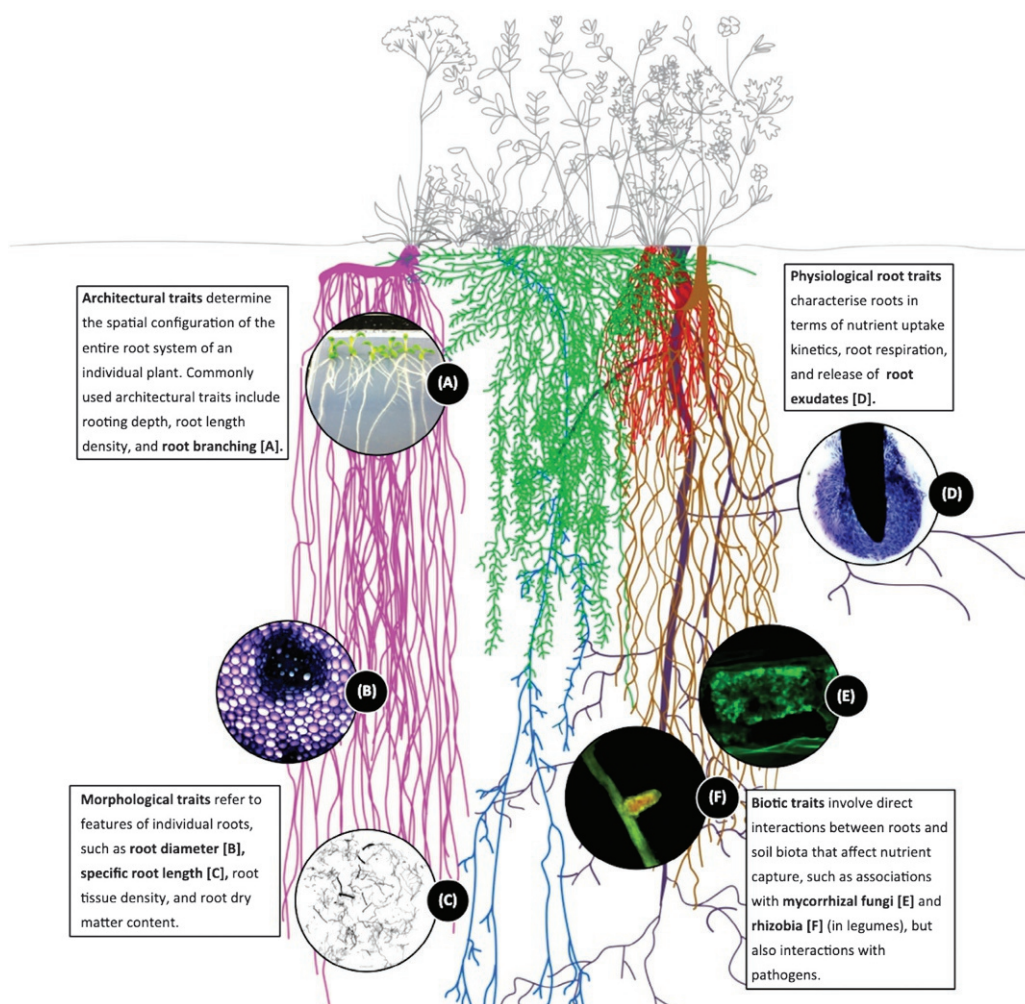


Fig. 6 Categories of root traits that have potential to impact ecosystem processes, including architectural, morphological, physiological, and biotic traits. The general idea behind a trait-based framework is that variation in root traits across individual plant species, communities, and ecosystems captures variation in a range of ecosystem processes, including carbon and nutrient cycling, and the structural stability of soil. Images represent the different root trait categories with potential to impact ecosystem processes: (A) root systems of *Leucanthemum vulgare* growing on agar plates by Liesje Mommer, Wageningen University (WU), representing architectural traits; (B) cross section of root of *L. vulgare* stained toluidine blue by Eric Visser, Radboud University, Nijmegen; (C) root length scan of washed roots from diverse grassland community by Natalie Oram (WU), both representing morphological traits; (D) root nodule in *Medicago truncatula* by Erik Limpens (WU); (E) arbuscules of AM fungi in *M. truncatula* root by Rik Huiskes (WU), both representing biotic traits; and, finally (F) rhizodeposition around *Zea mays* root by Davey Jones, Bangor University, representing physiological traits. Figure and caption from Bardgett RD, Mommer L, De Vries FT et al. (2014) Going underground: Root traits as drivers of ecosystem processes. *Trends in Ecology and Evolution* **29** (12): 692–699. doi: 10.1016/j.tree.2014.10.

a major source of phosphorus uplift to the topsoil horizon, which explains the retrogression of very old ecosystems progressively disconnected from bedrock as soils grow thicker and bedrock becomes inaccessible to roots (Peltzer et al., 2010).

Root architecture and traits influence relationships among plants and water budgets and cycling at both small and large scales. Competition between plants in the same community affects root distribution and water uptake. Studies that use isotopic techniques along with measurements of root biomass and leaf gas exchange provide an assessment of root distribution, short- and long-term physiological activity, and source water use to inform competitive interactions. In tree-grass mixtures, for example, grasses exert a negative effect on tree root growth, but usually not vice versa. Drought, however, typically has a more negative effect on grass than on trees, which have more conservative leaf gas exchange characteristics, and deeper roots that utilize deeply stored water. As previously mentioned, plants with deeper roots, such as trees, can further redistribute water from the deep soil to the shallow soil horizons overnight, a process called hydraulic lift or redistribution which also benefits more shallow rooted species. Isotopic techniques as well as experimental and observational studies continue to sort out these complex, root-mediated, competitive relationships. Root traits may also influence competitive dynamics between native and invasive species. For example, rapid root growth of the invasive annual grass, *Bromus tectorum*, in cool conditions allows it to pre-empt resources and outcompete co-occurring native grass species, whereas root traits such as large numbers of root tips and large root to shoot ratios have been shown to improve native plant performance in response to competition from invasives *B. tectorum* and *Centaurea stoebe*.

At a regional scale, root distribution can affect hydrology. An example is the replacement of perennial, deep-rooted (~20 m) trees and shrubs, by shallow-rooted pasture and crop plants in the seasonally-dry Murray-Darling Basin in Australia. Historically, the native vegetation would have relied on deep soil recharge and groundwater, while the shallow-rooted agroecosystems today occupy only the top meter of soil and utilize far less water than the forest and woodland communities, whose evergreen canopies had higher evapotranspiration rates.

Many aspects of soil structural stability (Bardgett et al., 2014; Rillig et al., 2015) and slope stability (Stokes et al., 2009) are altered by root systems and specific root traits. Roots promote the formation and stabilization of soil aggregates by compressing or entangling soil particles, by providing food sources (e.g., organic residues from roots sloughing or dieback) for microbes that secrete organic polymers or directly via their root exudates, both of which bind together soil particles. Roots mitigate soil detachment and erosion by enhancing soil cohesion (a component of soil shear strength) and aggregate stability (i.e., the ability to resist disintegration under force). As roots extend through soil vertically and laterally, they reinforce soil to increase its shear strength and overall soil cohesion. The distribution of roots, their length, orientation, and tensile strength (i.e., the force required to break a root divided by the cross-sectional area of a root) also influence soil and slope stability.

Conclusion

Research focused on root form and function has improved understanding of the mechanisms by which plants cope with environmental stress, the dynamics of plant-plant and plant-microbe interactions, and how roots influence various ecosystem processes, properties, and services. Despite significant advances, additional research is needed to understand how root growth and traits influence plant performance, vegetation dynamics, and associated ecosystem processes in changing conditions. Increased focus belowground on dynamic plant traits (e.g., turnover rates, root lifespan), regenerative traits (e.g., belowground bud bank size and storage), anatomical traits (e.g., cortex and stele area ratio), and physiological traits (e.g., hydraulic conductance) are areas where knowledge remains particularly scarce. In addition, a more comprehensive understanding of between- and within-species variation in root growth and traits may improve our ability to select plants with specific traits for increased crop productivity and improved restoration in the face of various global change pressures. Luckily, methods for root trait sampling, processing, and measuring are continuously being improved and refined to ensure consistency across studies and ecosystems (Freschet et al., 2021a, 2021b). In addition, as data accumulate from diverse systems and become available via large databases (e.g., Global root database, Guerrero-Ramirez et al., 2021) researchers will be able to further clarify linkages between root traits and vegetation dynamics across diverse ecosystems and scales.

References

- Bardgett RD, Mommer L, De Vries FT, et al. (2014) Going underground: Root traits as drivers of ecosystem processes. *Trends in Ecology & Evolution* 29(12): 692–699. <https://doi.org/10.1016/j.tree.2014.10>.
- Bengough AG, McKenzie BM, Hallet PD, and Valentine TA (2011) Root elongation, water stress, and mechanical impedance: A review of limiting stresses and beneficial root tip traits. *Journal of Experimental Botany* 62: 59–68. <https://doi.org/10.1093/jxb/erq350>.
- Bergmann J, Weigelt A, van Der Plas F, et al. (2020) The fungal collaboration gradient dominates economics space in plants. *Science Advances* 6: eaba3756. <https://doi.org/10.1126/sciadv.aba3756>.
- Comas LH, Mueller KE, Taylor LL, et al. (2012) Evolutionary patterns and biogeochemical significance of angiosperm root traits. *International Journal of Plant Sciences* 173: 584–595. <https://doi.org/10.1086/665823>.
- Comas LH, Becker SR, Cruz MV, Byrn PF, and Dierig DA (2014) Root traits contributing to plant productivity under drought. *Frontiers in Plant Science* 4: 1–16. <https://doi.org/10.3389/fpls.2013.00442>.
- Correa J, Postma JA, Watt M, and Wojciechowski T (2019) Soil compaction and the architectural plasticity of root systems. *Journal of Experimental Botany* 70(21): 6019–6034. <https://doi.org/10.1093/jxb/erz383>.
- Crang R, Lyons-Sobaski S, and Wise R (2019) Roots. In: Crang R, Lyons-Sobaski S, and Wise R (eds.) *Plant Anatomy: A Concept Based Approach to the Structure of Seed Plants*. Springer.
- Freschet GT, Violle C, Bourget MY, et al. (2018) Allocation, morphology, physiology, architecture: The multiple facets of plant above- and below-ground responses to resource stress. *New Phytologist* 219: 1338–1352. <https://doi.org/10.1111/nph.15225>.
- Freschet GT, Pagès L, Iversen CM, et al. (2021a) A starting guide to root ecology: Strengthening ecological concepts and standardizing root classification, sampling, processing and trait measurements. *New Phytologist* 232: 973–1122. <https://doi.org/10.1111/nph.17572>.
- Freschet GT, Roumet C, Comas LH, et al. (2021b) Root traits as drivers of plant and ecosystem functioning: Current understanding, pitfalls and future research needs. *New Phytologist* 232: 1123–1158. <https://doi.org/10.1111/nph.17072>.
- Guerrero-Ramirez N, Mommer L, Freschet GT, et al. (2021) Global Root Traits (GRoT) Database. *Global Ecology and Biogeography* 30(1): 25–37. <https://doi.org/10.1111/geb.13179>.
- Gupta A, Rico-Medina A, and Caño-Delgado AI (2020) The physiology of plant responses to drought. *Science* 648(368): 266–269. <https://doi.org/10.1126/science.aaz7614>.
- Hawkins BJ, Robbins S, and Porter RB (2014) Nitrogen uptake over entire root systems of tree seedlings. *Tree Physiology* 34: 334–342. <https://doi.org/10.1093/treephys/tpu005>.
- Hodge A (2004) The plastic plant: Root responses to heterogeneous supplies of nutrients. *New Phytologist* 162(1): 9–24. <https://doi.org/10.1111/j.1469-8137.2004.01015.x>.
- Jacobsen AG, Jervis G, Xu J, et al. (2021) Root growth responses to mechanical impedance are regulated by a network of ROS, ethylene and auxin signalling in Arabidopsis. *New Phytologist* 231: 225–242.
- Jung KH and McCouch S (2013) Getting to the roots of it: Genetic and hormonal control of root architecture. *Frontiers in Plant Science* 4(186): 1–32. <https://doi.org/10.3389/fpls.2013.00186>.
- Klimešová J, Martinková J, Pausas JG, et al. (2019) Handbook of standardized protocols for collecting plant modularity traits. *Perspectives in Plant Ecology, Evolution, and Systematics* 40: 125485. <https://doi.org/10.1016/j.ppees.2019.125485>.

- Koevoets IT, Venema JH, Elzenga JTM, and Testerink C (2016) Roots withstanding their environment: Exploiting root system architecture responses to abiotic stress to improve crop tolerance. *Frontiers in Plant Science* 7: 1335. <https://doi.org/10.3389/fpls.2016.01335>.
- Latef AA, Hashem A, Saiema R, et al. (2016) Arbuscular mycorrhizal symbiosis and abiotic stress in plants: A review. *Journal of Plant Biology* 59: 407–426. <https://doi.org/10.1007/s12374-016-0237-7>.
- Li C, Li L, Reynolds MP, Wang J, et al. (2021) Recognizing the hidden half in wheat: Root system attributes associated with drought tolerance. *Journal of Experimental Botany* 72(14): 5117–5133. <https://doi.org/10.1093/jxb/erab124>.
- Marschner P and Rengel Z (2012) Nutrient availability in soils. In: Marschner P (ed.) *Marschner's Mineral Nutrition of Higher Plants*, 3rd edn Elsevier <https://doi.org/10.1016/B978-0-12-384905-2.00012-1>.
- Moore BD and Johnson SN (2017) Get tough, get toxic, or get a bodyguard: Identifying candidate traits conferring belowground resistance to herbivores in grasses. *Frontiers in Plant Science* 7: 1925. <https://doi.org/10.3389/fpls.2016.01925>.
- Mukherjee S and Baluška F (2021) Rhizobiology: Molecular physiology of plant roots. *Springer Nature*. <https://doi.org/10.1007/978-3-030-84985-6>.
- Peltzer DA, Wardle DA, Allison VJ, et al. (2010) Understanding ecosystem retrogression. *Ecological Monographs* 80: 509–529. <https://doi.org/10.1890/09-1552.1>.
- Ranathunge KM (2005) *The Role of the Apoplastic Transport Barriers for Radial Water and Ion Uptake in Rice (Oryza sativa L.) and corn (Zea mays L.) Roots*. Doctoral thesis University of Bayreuth, Faculty of Biology, Chemistry and Earth Sciences.
- Rillig MC, Aguilar-Trigueros CA, Bergmann J, et al. (2015) Plant root and mycorrhizal fungal traits for understanding soil aggregation. *New Phytologist* 205(4): 1385–1388. <https://doi.org/10.1111/nph.13045>.
- Siddiqui MN, Léon J, Naz AA, and Ballvora A (2020) Genetics and genomics of root system variation in adaptation to drought stress in cereal crops. *Journal of Experimental Botany* 72: 4. <https://doi.org/10.1093/jxb/eraa487>.
- Stokes A, Atger C, Bengough AG, Fourcaud T, and Sidle RC (2009) Desirable plant root traits for protecting natural and engineered slopes against landslides. *Plant and Soil* 324: 1–30. <https://doi.org/10.1007/s11104-009-0159-y>.
- Tedersoo L (2017) *Biogeography of Mycorrhizal Symbiosis*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-56363-3>.
- Weigelt A, Mommer L, Andrzejek K, et al. (2021) An integrated framework of plant form and function: The belowground perspective. *New Phytologist* 232(1): 42–59. <https://doi.org/10.1111/nph.17590>.
- Yamauchi T, Pedersen O, Nakazono M, and Tsutsumi N (2020) Key root traits of Poaceae for adaptation to soil water gradients. *New Phytologist* 229(6): 3133–3140. <https://doi.org/10.1111/nph.17093>.
- Ye H, Roorkiwal M, Valliyodan B, et al. (2018) Genetic diversity of root system architecture in response to drought stress in grain legumes. *Journal of Experimental Botany* 69: 3267–3277. <https://doi.org/10.1093/jxb/ery082>.
- York LM, Carminati A, Mooney SJ, Ritz K, and Bennett MJ (2016) The holistic rhizosphere: Integrating zones, processes, and semantics in the soil influenced by roots. *Journal of Experimental Botany* 67(12): 3629–3643. <https://doi.org/10.1093/jxb/erw108>.

Root mucilage: Chemistry and functions in soil

Yutong Jiang^a, Lijun Hou^a, Fang Ding^b, and Joann K Whalen^a, ^aDepartment of Natural Resource Sciences, Macdonald Campus of McGill University, Sainte-Anne-de-Bellevue, QC, Canada; ^bChinese Academy of Sciences, Institute of Applied Ecology, Shenyang, China

© 2023 Elsevier Ltd. All rights reserved.

Introduction	332
Production and chemistry of root mucilage	333
Factors affecting root mucilage production	336
Abiotic stress and root mucilage	336
Biotic stress and root mucilage	337
Functions of root mucilage	337
Water retention	337
Protective barrier	337
Nutrient absorption	337
Binding of trace metals	338
Rhizobiome: Mucilage-microorganism interaction	338
Mucilage and root-associated bacteria	338
Research methods for mucilage collection	339
Summary and outlook	341
References	341

Abstract

Root mucilage is a sticky polymer that is secreted through root border cells. Primarily composed of polysaccharides, with some protein and lipid components, mucilage causes soil particles to adhere to the roots, creating the characteristic rhizosheath that is visible around many plant root systems. The mucilage layer retains water, which is important for plants growing in water-limited environments. Plant signal molecules are transmitted through the mucilage to symbiotic and free-living soil microorganisms that promote plant growth. Furthermore, mucilage contains defensive compounds that can prevent infection by root pathogens. This chapter discusses multiple functions of root mucilage for plant health and resilience.

Key points

- Root mucilage is a complex mixture of polysaccharides, proteins and lipids, secreted by plants in response to genetic and environmental factors.
- Mucilage creates an adhesion between roots and soil particles, bringing roots into closer proximity to the soil water and dissolved nutritive ions.
- Mucilage retains water when soils become drier, creating a humid microenvironment that attracts soil microorganisms, assembling the rhizobiome.
- Root mucilage protects plants from pathogens. Mucilage is a physical barrier between soil-borne pathogens and the root surface, and it contains extracellular DNA, histone and other antimicrobial compounds that halt disease-causing microorganisms before they can infect the plant.

Introduction

The rhizosphere is the unique soil microenvironment that exists around growing plant roots (Fig. 1). The rhizoplane is the most dynamic zone of the rhizosphere. Located at the interface between the external root surface and the soil particles surrounding the root, the rhizoplane is enriched with plant root secretions and debris called rhizodeposits (Fig. 1). Rhizodeposits include the abraded root cap and border cells, dead root cells, water-soluble root exudates, respired gasses and mucilage. The mucilage released from roots, combined with root hairs, is responsible for the formation of a rhizosheath, the mass of soil adhering to plant roots that does not fall away when roots are gently shaken.

The rhizosheath forms because mucilage is an adhesive substance that causes soil particles to stick together, creating porous soil aggregates near the root surface. The rhizosheath protects plants from desiccation in dry and semi-arid environments, for two reasons. First, mucilage has a polymeric gel-like structure that attracts and retains water molecules. Second, porous soil aggregates hold water under tension, which prevents the water from draining out of the rhizosphere. Another feature of soil aggregates is that they contain clay, which has a large surface area occupied by exchangeable ions such as Ca^{2+} , Mg^{2+} , NH_4^+ and K^+ , essential plant

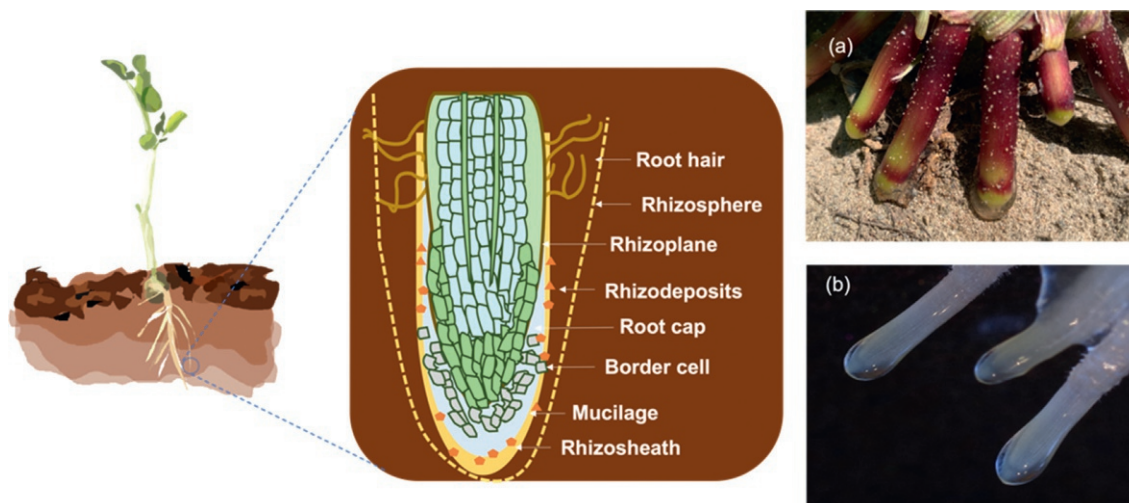


Fig. 1 Plant rhizosphere, showing the various components of the root system, secretions and surrounding soil. Mucilage is a polysaccharide-rich compound that is produced in border cells in the plant root tip. Photos show (a) mucilage being released from the brace root tip of corn (*Zea mays*) at the reproductive stage, and (b) mucilage secreted from the fine root tips of 3-day old wheat (*Triticum aestivum*) seedlings. Photos from Holz M, Leue M, Ahmed MA, Benard P, Gerke HH and Carminati A (2018) Spatial distribution of mucilage in the rhizosphere measured with infrared spectroscopy. *Frontiers in Environmental Science* 6: 87. doi: 10.3389/fenvs.2018.00087.

nutrients. The presence of nutrient-rich clays in the rhizosheath reduces the distance that ions must diffuse to reach the root surface, where they are absorbed by the plant. Furthermore, porous soil aggregates are a microhabitat for root symbionts, endophytes and free-living microorganisms that help plants to acquire water and nutrients.

The rhizosheath was first described in desert grasses, as a sheath of agglutinated sand particles around the roots (Price, 1911). Plants with a visible rhizosheath secrete mucilage and have additional characteristics that favor the root-soil adhesion, such as their root hair length, density, morphology and a diverse root microbiome. Mucilage secretions can account for nearly half of the rhizodeposits from grasses. Root mucilage is a viscous, gel-like material. It is composed of polysaccharides (~78%), glycoproteins (~7%), minerals (~6%) and lipids (~3%) (Nazari, 2021). Most plant mucilage is produced in the Golgi apparatus and it is released predominantly from fine root tips, which extend from the distal tips of the lateral roots into the surrounding soil. The rhizosheath forms because the gelatinous mucilage layer around border cells, outside of the fine root tip, adheres strongly to the surrounding soil particles by hydrogen bonds between the hydroxyl groups of the polysaccharides and soil particles (Watt et al., 1993). Mucilage lubricates growing roots as they extend into soil pores and coats the surrounding soil particles to protect the delicate root tip from abrasion.

Mucilage varies in its chemical composition, and this affects its functions in different plant species (Table 1). For example, some organic molecules like pectins and arabinogalactan protein have bioadhesive properties, meaning that they bind microbial cells to the root surface. Mucilage molecules like xyloglucan are responsible for soil adhesion because they act as a glue between the root surface and soil particles (Table 1). Other functional molecules in mucilage are responsible for stress response, pathogen control, cation exchange and hydraulic properties at the root-soil interface (Table 1). Enzymatic degradation of polysaccharides present in mucilage could provide carbon substrates that are metabolized by growing soil microorganisms. Therefore, root mucilage contains diverse compounds that support many physical, chemical and biological processes around roots.

Production and chemistry of root mucilage

Most plant roots are coated with mucilage, but the root tip has the thickest mucilage layer. This is because mucilage at the root tip adheres to cellular debris (sloughed border cells) that are scraped off the root as it grows (Fig. 2a). The thick mucilage layer protects the root apical meristem as the growing root tip penetrates pores, cracks and voids in the soil matrix. Biosynthesis and production of mucilage occurs mainly in root epidermal cells. As illustrated in Fig. 2b, mucilage is produced in the Golgi apparatus inside the cells of the root apical meristem, then transferred into the cytoplasm via the cisternae, the flattened membrane vesicles within the Golgi apparatus. Mucilage is stored temporarily in the cisternae, mixed with aqueous solution and moved through the outgoing transport vesicles. These vesicles traverse the cytoplasm and fuse to the meristematic cell membrane, where their contents are discharged into the periplasm as mucilage droplets. Finally, mucilage droplets are secreted through pores in the meristematic cell wall, onto the surface of the root epidermal cells and form a gelatinous layer that adheres to the surrounding soil (Fig. 2).

Early studies of root mucilage relied upon visual images of stained root cross-sections, viewed with light microscopy. Soaking root extracts in acid-Schiff reagent with methylene blue dye proved that mucilage contains carbohydrates with cis hydroxyl groups. Ruthenium red dye binds to pectins in mucilage. Lipids in mucilage are visible under a light microscope when triglycerides react

Table 1 Mucilage components found on the root surface of representative plant species, and their functions.

Mucilage components	Compounds	Molecule	Functions	Plant species							
				Corn (Zea mays)	Barley (Hordeum vulgare)	Lupin (Lupinus angustifolius)	Rapeseed (Brassica napus)	Tomato (Solanum lycopersicum)	Wheat (Triticum aestivum)	Thale cress (Arabidopsis thaliana)	Pea (Pisum sativum)
Polysaccharides (78–94%) ^a	Pectin	Homogalacturonan	Lubrication of the root tip	✓	✓		✓	✓	✓	✓	✓
		Xylogalacturonan	Soil adhesion			✓	✓	✓		✓	✓
	Hemicelluloses	Rhamnogalacturonan	Soil adhesion ^b	✓	✓	✓	✓	✓	✓	✓	✓
		Xyloglucan	Root defense ^c	✓	✓	✓	✓	✓	✓	✓	✓
		Xylan	Cell adhesion ^d	✓			✓	✓	✓		✓
		Heteromannan	Root defense ^c	✓	✓			✓	✓		✓
	Cellulose		Cell adhesion ^e	✓	✓	✓	✓	✓	✓	✓	✓
Glycoproteins (6–7%) ^a	Hydroxyproline-rich glycoproteins	Arabinogalactan protein	Soil adhesion	✓	✓	✓	✓	✓	✓	✓	
			Lubrication of the root tip								
			Binding microbial cell to root cell								
Minerals (~6%) ^a		Extensins	Root defense ^c		✓			✓		✓	✓
		Mg ²⁺ , Ca ²⁺ , Cu ²⁺ , K ⁺ , Zn ²⁺	Cation exchange	✓	✓	✓	✓	✓	✓	✓	
Lipids (~1–3%) ^a	Phosphatidylcholines	Fatty acids	Viscosity								
			Phosphate desorption	✓	✓	✓	✓	✓	✓	✓	
			Lubrication of the root tip								
			Modifies soil hydraulic properties ^e								

^aNazari (2021).^bKaczmarska et al. (2021).^cCastilleux et al. (2018).^dRalet et al. (2016).^eRopitiaux et al. (2019).

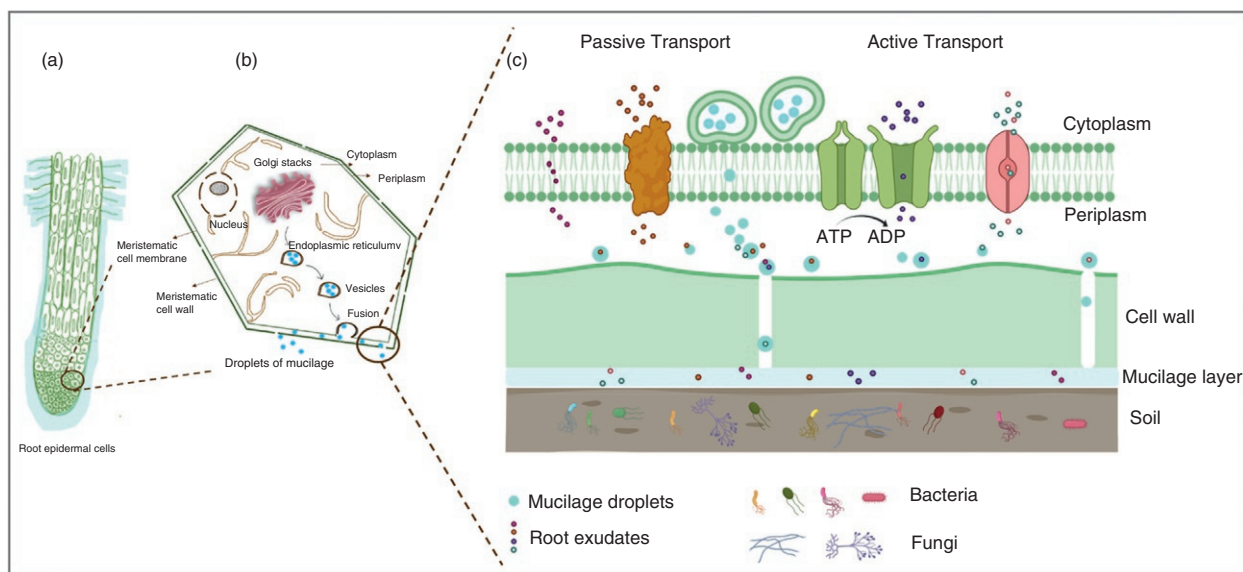


Fig. 2 Production and secretion of mucilage from root tip cells. At the root tip, (a) the epidermal cells are coated with a mucilaginous layer. Mucilage biosynthesis occurs in (b) the Golgi apparatus of a meristematic cell, producing mucilage droplets that travel through the cytoplasm. In the periplasmic space, (c) root exudates are encapsulated within the mucilage droplets, and these compounds move jointly through the cell wall before they are deposited in the mucilage layer. The root exudates diffuse gradually through the mucilage layer into the adhering soil, which is inhabited by free-living soil bacteria and fungi. Adapted from Driouch A, Gaudry A, Pawlak B and Moore JP (2021) Root cap-derived cells and mucilage: A protective network at the root tip. *Protoplasma* **258**: 1179–1185. doi: 10.1007/s00709-021-01660-y; Silva J, Ferraz R, Dupree P, Showalter AM and Coimbra S (2020) Three decades of advances in arabinogalactan-protein biosynthesis. *Frontiers in Plant Science* **11**: 610377. doi: 10.3389/fpls.2020.610377.

with fat-soluble Sudan III red dye. These visual observations led to a mucilage classification scheme, which distinguished the cellulose-based mucilage (containing cellulose microfibrils) from the amorphous pectin-based mucilage.

Today, the complex chemistry of root mucilage and its physical functions are studied with advanced chemical techniques like gas chromatography-mass spectrometry, ^1H and ^{13}C nuclear magnetic resonance spectroscopy, and Fourier-transform infrared spectroscopy. Analysis of the cellulosic polysaccharides in mucilage, such as xyloglucan and heteromannan, showed that these compounds create stability and cohesion in mucilage polymers. Mucilage hydration is maintained by pectic polysaccharides, namely homogalacturonan, rhamnogalacturonan-I, and xylogalacturonan. In some cases, the cellulose and pectin components of mucilage bind together to create an extensive polymeric network. Mucilage also contains cell wall glycoproteins like arabinogalactan proteins and extensins that protect the integrity of root cells beneath the mucilage layer.

Mucilage has multiple functions that are beneficial to plants. Mucilage forms the rhizosheath, which brings nutrient-rich clay particles close to the root surface, allowing for ion exchange and interception by growing roots. Mucilage protects elongating root tips from abrasion. The polymeric, gel-like structure of mucilage retains water in the rhizosphere. Soil ions and water pass through the mucilage layer to reach the root surface, but substances also flow in the other direction, from the roots into the mucilage layer and then into the surrounding soil. Many low molecular weight root exudates move through the mucilage layer into soil, including tannins, alkaloids, steroids, polyphenols, terpenes, flavonoids, phenolic acids and anthocyanins. As illustrated in Fig. 2c, root exudates are transported out of root cells through active and passive transport, then discharged into the periplasm along with mucilage droplets. Root exudates that become encapsulated in mucilage droplets are secreted together through cell wall pores, into the mucilage layer on the external surface of the root epidermal cell. This is how root exudates reach soil microorganisms in the rhizosheath and surrounding soil (Fig. 2c).

Root exudates are often described as a source of energy and carbon for soil microbial populations, but they have other functions. Some root exudates like flavonoids and strigolactones are signal molecules that stimulate microbial movement or growth towards the root. Attracting nitrogen-fixing bacteria, mycorrhizal fungi and other plant growth-promoting microorganisms (free-living and endophytic) to roots is expected to improve plant growth. However, signal molecules may also induce the growth of plant pathogenic fungi and parasitic nematodes that infect root tissues, an undesirable outcome. Some root exudates elicit microbial biosynthesis of secondary metabolites like phytohormones and siderophores in the rhizosheath. For example, the root exudate L-tryptophan is the catabolic precursor of the plant growth-stimulating hormone indole-3-acetic acid produced by rhizosphere bacteria (Kravchenko et al., 2004). Microbial byproducts may stimulate or inhibit plant growth, depending on the nature of the compound, the environmental conditions and the plant species involved. Mucilage seems to distribute root exudates where they are needed to enhance the connection between plant roots and soil microorganisms, thereby controlling plant-microbial interactions in the rhizosphere.

Factors affecting root mucilage production

Most angiosperms produce mucilage, but rhizosheath formation is more common in grasses and cereal crops in the Gramineae family. Well-developed rhizosheath is visible around the roots of cultivated cereals like corn (*Zea mays*), wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), oat (*Avena sativa*) and sorghum (*Sorghum bicolor*) (Duell and Peacock, 1985). Since as much as 50% of the rhizodeposits from these cereal crops is allocated to mucilage production, which is equivalent to about 10% of the total carbon fixed through photosynthesis, this suggests a specialized growth strategy. We expect plant species to allocate more energy and carbon to mucilage production when exposed to abiotic stress (e.g., drought, heat, trace metals) or biotic stress from pathogens and parasites.

Abiotic stress and root mucilage

Drought occurs when there is insufficient rainfall or irrigation to replenish the soil water supply for plants. As the soil water content declines, plants must exert more hydraulic force to pull water under tension out of soil pores, to the root surface. Plants with an inadequate water supply must close their stomates to avoid desiccation from evapotranspiration, which limits photosynthesis. Since water is retained in the polymeric structure of mucilage and in porous soil aggregates adhering to the mucilage layer, plants may deliberately secrete mucilage to prevent root dehydration. Mucilage-producing plants that can attract mycorrhizal fungi to colonize their roots will have extra capacity to acquire water because fungal hyphae can extract water from soil micropores that are too small for plant roots to penetrate. There is evidence that plants release more mucilage in response to soil water shortage. For example, corn genotypes growing in semi-arid regions produce more mucilage than corn genotypes in humid regions (Table 2; Nazari et al., 2020). The amount of galactose, fructose, mannose, glucose, arabinose, xylose, and glucuronic acid contained in nodal root mucilage also vary among corn genotypes in different agroecological regions (Table 2). It is not fully understood why corn plants modify the proportions of polysaccharides in mucilage when they grow in a water-limited environment. It could be to optimize the water-holding capacity of mucilage, which depends upon the number of hydroxyl and carboxyl groups present in polysaccharides. Therefore, changes in the quantity of mucilage produced and its chemical composition could be a plant strategy to survive water shortages, although this remains to be confirmed.

Heat stress occurs when plants cannot regulate their internal temperature in hot environments. Heat stress limits plant root development, which could affect mucilage production and chemical composition. In corn plants, the root cap grows optimally at 25 °C, producing about 3600 cells d⁻¹, which is an order of magnitude greater than 360 cells d⁻¹ produced at 15 °C and four times more than 850 cells d⁻¹ produced at 35 °C (Clowes and Wadekar, 1988). A decline in root cell production at hotter temperatures will probably reduce the amount of mucilage produced around the root tip. Furthermore, extreme heat stress causes plants to close their stomates, which lowers their transpiration and photosynthesis rates. Less carbon fixation constrains many biosynthetic processes of plants, including mucilage production and is expected to alter the relative proportions of polysaccharides, proteins and lipids contained in mucilage.

Contaminants like trace metals affect root mucilage production, although the response depends on the metal ion and the exposure dose (i.e., from the smallest observed effect concentration to the lethal concentration). Root epidermal cells exposed to trace metals often experience reactive oxygen stress, which disrupts the cell membrane due to lipid peroxidation and impaired thiolation of proteins, decreases cellular respiration, impairs biochemical reactions and damages cellular DNA. Mucilage contains uronic acids and pectin molecules that bind to metal ions, thereby protecting root epidermal cells from reactive oxygen stress. When saltmarsh mallow (*Kosteletzkya pentacarpos*) was exposed to Zn²⁺ from 0.1 mM ZnCl₂ solution, root mucilage production increased from 12 to 18 mg g⁻¹ plant (dry mass), and the uronic acid concentration in mucilage rose from 50 to 180 µg mg⁻¹ of mucilage on a dry weight basis (Zhou et al., 2020). Plants vary the pectin content in mucilage when exposed to trace metals, possibly because

Table 2 Nodal root mucilage dry weight and proportion of monosaccharides in the nodal root mucilage of eight corn genotypes from four climate regions.

Climate region	Genotype of maize	Country	Mucilage production (µg mm ⁻²)	Galactose	Fucose	Mannose	Glucose	Arabinose	Xylose	Glucuronic acid
Temperate	Kentos	Germany	4.3 ± 0	~42%	~24%	~12%	~4%	~9%	~1%	~5%
	KXB8383	France	5.2 ± 1	~41%	~26%	~11%	~2%	~11%	~4%	~3%
Mediterranean	Keravnos	Turkey	12 ± 1	~40%	~28%	~12%	~3%	~10%	~2%	~3%
	Kerubino	Italy	7.5 ± 2	~41%	~22%	~14%	~3%	~9%	~3%	~4%
Tropical	DH 04	Kenya	6.6 ± 2	~40%	~26%	~12%	~2%	~10%	~3%	~5%
Semi-arid	DH 02	Kenya	24 ± 4	~40%	~22%	~12%	~4%	~11%	~4%	~4%
Semi-arid	900MGold	India	25 ± 10	~39%	~30%	~11%	~1%	~10%	~2%	~4%
Semi-arid	30V92	India	20 ± 4	~42%	~24%	~13%	~2%	~8%	~2%	~5%

Data are the mean ± standard error of the mean.

Adapted from Nazari M, Riebeling S, Banfield CC, Akale A, Crosta M, Mason-Jones K, Dippold MA and Ahmed MA (2020) Mucilage polysaccharide composition and exudation in maize from contrasting climatic regions. *Frontiers in Plant Science* 11: 587610. doi: 10.3389/fpls.2020.587610.

pectin is a selective binder of trace metals. For example, Cai et al. (2011) reported that soybean root tips exposed to Al^{3+} had a thicker mucilage layer and more pectin methylesterase activity, which was associated with proton release and Al^{3+} binding to pectin in the mucilage layer.

Biotic stress and root mucilage

Plant roots are exposed to biotic stress from viruses, pathogenic fungi and bacteria, soil nematodes and other biota, which activate plant defense mechanisms. Once a pathogen is detected, plants upregulate the phytohormone signaling pathways of salicylic acid, jasmonic acid, and ethylene that initiate pathogenesis-related peptides and genes, triggering the systemic acquired resistance defenses. However, plants can also rely on root mucilage for physical and chemical defense. Mucilage combined with the root cap-derived cells creates the mucilaginous root extracellular trap, a sticky adhesive that prevents pathogens from moving to and infecting the growing root tip. Once the pathogen is physically immobilized, it is deactivated by defensive molecules, including extracellular DNA, defensive peptides and proteins, contained in the mucilage. Mucilage fills xylem vessels in injured plant roots, which prevents pathogen invasion. For example, a wounded corn root produced mucilage with abundant pectin that filled 3 cm of xylem vessels in the injured area after 1 day, and this restricted pathogen infection of the damaged root tissues (Crews et al., 2003).

Plants in the Fabaceae family respond to signals from invading pathogens by releasing 'extracellular trapping' substances such as extracellular DNA and granular proteins (e.g., histone) from lysed border cells in their root mucilage. For example, pea roots released extracellular DNA from the border cells when exposed to the flagellin signal molecule of pathogenic *Ralstonia solanacearum* but did not respond to the presence of the *R. solanacearum* mutants that lack flagellin or the non-pathogenic bacteria *Pseudomonas aureofaciens*, *Ensifer meliloti* and *Escherichia coli* (Gunawardena and Hawes, 2002; Tran et al., 2016). This implies that root mucilage production is part of the defensive response to root pathogens. Mucilage may function as a physical barrier to root invasion, as discussed above. Another possibility is that extracellular DNA and other compounds in the mucilage bind to chemical receptors in the pathogen cell membrane to interfere with pathogen growth and root infection. The ability of pea roots to alter their mucilage production after detecting flagellin from a pathogen suggests that mucilage is responsible for specific defense mechanisms that prevent plant diseases.

Functions of root mucilage

Water retention

Mucilage is a gel-like, hygroscopic substance that can absorb water at a rate of 25–600 times its dry weight, due to its high hydroxyl content (Huang and Gutterman, 1999). For example, mucilage-treated soil maintained a volumetric water content of $0.24\text{--}0.25\text{ cm}^3\text{ cm}^{-3}$ when dried, but the soil water content without mucilage declined from $0.25\text{ cm}^3\text{ cm}^{-3}$ to $0.17\text{ cm}^3\text{ cm}^{-3}$ in 200 min (Kroener et al., 2014). When soil is moist, mucilage becomes hydrated and retains water around the rhizosheath. The presence of mucilage is expected to increase the volumetric water content of the rhizosheath by 3–5% at water potentials of -10 kPa , with about 2% more volumetric water in the rhizosheath at -10 MPa (Ahmed et al., 2014; Ghezzehei and Albalasmeh, 2015). Transpiration pulls water from deeper soil layers to the roots, and water is likely to be adsorbed onto or absorbed into the mucilage layer of the rhizosheath, due to the hygroscopic character of mucilage. When soils become drier, the long-chain lipids in mucilage create a temporary water repellent layer, which prevents water loss from the rhizosheath. Furthermore, mucilage is viscous, which enhances the physical contact among soil particles around the rhizosheath. Formation of soil aggregates allows for the development of a soil microstructure, with small pores that hold water more tightly than large pores. Mucilage acts as an adhesive between the root surface and soil aggregates, which improves the connectivity of water-filled soil pores and the rhizoplane. Cohesive water tends to remain longer in soil micropores, whereas it drains readily from soil meso- and macro-pores. Together, these mucilage-induced effects in the rhizosheath mean that the root surface remains moist, even as the surrounding soil gets drier.

Protective barrier

Root mucilage acts as a protective barrier for the root tips because hydrated mucilage is a lubricant. Growing roots exert tremendous pressure ($>700\text{ kPa}$) on root tips to push them through the soil (McNear Jr, 2013). The lubricating effect is due to the polysaccharides and lipids in mucilage. Effectively, these compounds reduce friction between the growing root tip and soil particles, so that the growing root tip can easily penetrate into the soil without being damaged.

Nutrient absorption

Since the mucilage layer around the rhizosheath is relatively well-hydrated, compared to the surrounding soil, it enhances nutrient uptake around the root. Most plant nutrients are acquired from dissolved ions in the soil pore water, which move from the bulk soil to the root surface through diffusion (e.g., H_2PO_4^- , K^+ , NH_4^+ , Zn^{2+}) and mass flow (e.g., Ca^{2+} , NO_3^-) processes. Thus, plant nutrient uptake is controlled by the amount of water that reaches the root, as well as the length of time the ion-rich water is in contact with the rhizosphere. The rhizosheath facilitates the transfer of dissolved ions to the root surface. For example, the rhizosheath improved the Zn absorption into oat roots by 40% in dry soil with $<1.5\text{ MPa}$ water potential (Nambiar, 1976). Acquisition of phosphate ions is enhanced by root mucilage, for two reasons. First, lipids (mainly phosphatidylcholines and saturated lipids) in root mucilage reduce the capillary force needed to mobilize H_2PO_4^- that is adsorbed onto soil particles through intermolecular cohesion and adhesion forces. This phenomenon can increase the phosphate ion concentration in soil solution by 10% (Read et al., 2003).

Second, the mucilage provides a moist layer around the roots that favors ion diffusion, increasing the likelihood that phosphate ions will enter the rhizosphere rather than remain in the bulk soil. Therefore, mucilage production is expected to increase plant nutrient uptake from soil.

Binding of trace metals

Mucilage is a negatively charged organic molecule with an abundance of carboxylic group (e.g., in uronic acid monomers) that can bind cations, including trace metals. Uronic acids in the mucilage of plants like Arabidopsis, cress (*Lepidium sativum*) and saltmarsh mallow (*Kosteletzkya virginica*) generally bind large amounts of Ca^{2+} but may exchange Ca^{2+} for other monovalent and divalent cations in the soil pore water, such as Na^+ , Pb^{2+} and Cd^{2+} (Ghanem et al., 2010). The cation exchange capacity of mucilage can enhance the transfer of such ions from soil water to the plant. For example, saltmarsh mallow grown in a Na^+ soil environment had a greater Na^+ concentration in the mucilage and shoot tissue, relative to the Ca^{2+} concentration, probably because Na^+ was bound preferentially to the uronic acid and rhamnose in mucilage (Ghanem et al., 2010). However, the plant family Melastomataceae produces mucilage with unmethylated uronic acid that has high affinity for trivalent cations, such as Al^{3+} . In a hydroponic experiment with melastoma (*Melastoma candidum*), plants that produced root mucilage accumulated 83% more Al^{3+} in the shoots than plants without mucilage (Watanabe et al., 2008). It is equally likely that Al^{3+} will remain bound in the mucilage, which is beneficial for plants that are sensitive to Al toxicity. For example, cowpea (*Vigna unguiculata*) produced mucilage that retained about 50% of the added Al^{3+} within 5 mm of the primary root tips, without transferring the Al into the root tip meristems (Horst et al., 1982). Given the capacity of root mucilage to retain cations, some excluder plants may produce uronic acid-rich mucilage to limit their exposure to non-essential cations like Na^+ , Al^{3+} and trace metals. This could be a metal tolerance mechanism to limit the contaminant uptake in plant tissues.

Rhizobiome: Mucilage-microorganism interaction

Mucilage and root-associated bacteria

Mucilage is expected to affect the growth and activity of microorganisms living in and around plant roots. There are several ways that mucilage attracts motile microorganisms to the root surface, and assembles a diverse microbial community known as the rhizobiome: (1) mucilage retains water, providing a moist microenvironment, (2) signal molecules diffuse through the mucilage, allowing for communication between plant roots and microorganisms in and around the rhizosphere, (3) some of the polysaccharides, lipids and proteins in mucilage are labile compounds that microorganisms metabolize for energy and use to build cellular components, as they grow and reproduce, (4) within the mucilage, polysaccharides with highly branched structure create a sticky, adhesive barrier that restrain microbial cells, including pathogens, from making direct contact with root cells, and (5) mucilage contains extracellular DNA that can control pathogen infections of the root.

Mucilage provides a moist microenvironment for the rhizobiome

All organisms need water to survive. Mucilage released from the root tips attracts microorganisms because it changes soil hydraulic properties in the rhizosphere (soil water content, water potential and saturated hydraulic conductivity). The moist conditions encourage hyphal growth and motile bacteria towards the root. Also, the passive movement of spores and non-motile bacteria into the rhizosphere is enhanced by the water-attracting properties of mucilage. In general, the moist mucilage-rich rhizosheath supports more microorganisms than bulk soil. There were 1×10^8 CFU g^{-1} of culturable bacteria in a grass (*Calamovilfa longifolia*) root rhizosheath, which was two orders of magnitude greater than the 3×10^6 CFU g^{-1} in bulk soil from a sand dune (Bergmann et al., 2009). Water that is held by mucilage in the rhizosheath may contain less oxygen and have fewer dissolved ions than the bulk soil, as these essential substances are either absorbed into the root or used by the actively-growing rhizobiome. This is expected to increase competition in the rhizobiome community, which will strive to acquire more nutrients to meet their requirements and continue growing, reproducing and colonizing the moist mucilage-rich coated roots and rhizosheath.

Mucilage transfers signals from plant roots to the rhizobiome

The rhizobiome contains soil microorganisms that are specially adapted to grow in the rhizosphere, as mutualists, commensals and parasites. Roots communicate with their rhizobiome by releasing chemical signals in root exudates. Plant signal molecules are generally low-molecular weight compounds like organic acids, sugar, aliphatic acids, fatty acids, amino acids, flavonoids, plant hormones and secondary metabolites. Plant signal molecules elicit a response from the rhizobiome. This could be a specific growth response that entices the microorganism towards the root, which occurs with root symbionts such as nitrogen-fixing bacteria and arbuscular mycorrhizal fungi. However, some plant signaling molecules such as coumarins, triterpenes, flavonoids, and benzoxazinoids are known to repress the movement and growth of certain species in the rhizobiome (Jalmi and Sinha, 2022). Plant signal molecules also cause some root-associated bacteria to synthesize auxin-like substances and soil fungi to produce volatile organic compounds that alter root length, branching and other morphological features (Schenkel et al., 2018). Plants with a rhizosheath must transfer chemical signals through the mucilage layer, as illustrated in Fig. 2c, to make contact with their rhizobiome. Soluble compounds released as signal molecules are more likely to be received in a moist environment, probably because the water streaming over the root surface brings dissolved compounds into contact with microorganisms living in water films. Roots coated with mucilage may transfer more signal molecules to a responsive rhizobiome than roots without mucilage or in dry environments, but the idea still needs to be tested.

Mucilage is a carbon source for the rhizobiome

Mucilage is composed primarily of polysaccharides, some of which are a labile carbon source for the rhizobiome. Many soil bacteria and fungi produce glycan-degrading enzymes such as galactanases, arabinofuranosidases and glucuronidases that cleave bonds in the polysaccharides, releasing glucose, fructose and other simple sugars that are absorbed and metabolized in microbial cells. Gram-negative bacteria are the most active decomposers of mucilage polysaccharides. Based on ^{13}C isotopic tracers and phospholipid fatty acid analysis, Ahmed et al. (2018a) deduced that 40% of the carbon in cell membranes of gram-negative bacteria was derived from corn nodal root mucilage after 15 d incubation. Therefore, mucilage is considered a source of carbon and energy that fuels the metabolic processes, including growth and reproduction, of the rhizobiome.

The rhizobiome produces extracellular enzymes to degrade mucilage, but these enzymes are also capable of hydrolyzing other organic compounds, such as those found in soil organic matter. It is common to observe accelerated soil organic matter mineralization after microorganisms are activated by labile carbon sources, which is known as the “priming effect” (Kuzyakov, 2002). As it contains labile carbon for microorganisms, mucilage can induce a priming effect, depending on the soil water content. In moist soil at ~80% water holding capacity, 98% of the mucilage from corn nodal roots was decomposed after 15 d incubation and this was accompanied by a significant carbon dioxide flux, indicating a positive priming effect (Ahmed et al. (2018b)). However, little mucilage was decomposed and no carbon dioxide flux was observed when soil water content was at ~30% water holding capacity in the same experiment, meaning no priming effect in drier soil (Ahmed et al. (2018b)). This may be due in part to the fact that mucilage becomes water repellent in dry soils, which could prevent polysaccharide degradation by hydrolytic enzymes (i.e., that require water to catalyze the biochemical reaction). Furthermore, if the soil organic matter was coated with water-repellent mucilage, it may be physically protected from decomposition. Therefore, we must know the mucilage chemistry and environmental conditions to predict whether root mucilage will stimulate priming by the rhizobiome.

Mucilage is a barrier that prevents pathogens from contacting the root

Root mucilage forms a physical barrier around root tips of plants in the Fabaceae and Brassicaceae families, and this barrier immobilizes fungal hyphae and spores (Fig. 3). A mucilage barrier is constructed from several compounds (e.g., arabinogalactan proteins, pectin and hemicellulose) that form a rigid biofilm-like structure. Fungal hyphae cannot penetrate the thick gelatinous mucilage covering the root tips, which prevents contact between the pathogen and the root, stopping the pathogen from infecting the root tissues. Mobile spores are also halted by the mucilage barrier. For example, a mucilage-free soybean root tip was colonized rapidly by zoospores from the Oomycete *Phytophthora*, which began to germinate within 15 min. However, there was no contact between a mucilage-coated soybean root tip and *Phytophthora* zoospores or hyphae after 30 min (Ropitiaux et al., 2020). The protective effect of soybean root mucilage was attributed to abundant xyloglucan and heteromannan molecules that cross-linked with cellulose fibers. In pea root mucilage, arabinogalactan proteins attracted 57% of the zoospores of the pathogenic Oomycete *Aphanomyces euteiches* after 4 h, thereby preventing infection of the pea root (Cannesan et al., 2012). The key role of arabinogalactan proteins was confirmed by adding β -glucosyl Yariv reagent, which deactivated the arabinogalactan proteins and lowered the attraction efficiency for zoospores from 57% to 32% (Cannesan et al., 2012). Pathogen infection by *A. euteiches* is the result of chemotactic attraction of zoospores to the root, encystment within root tissues followed by cyst germination. This suggests that mucilage acts like a viscous shield that can physically protect roots from such infections.

Extracellular DNA inactivates infectious root pathogens

Mucilage inactivates infectious root pathogens because it contains chemical substances that inhibit spore germination and limit mycelial growth. These chemical substances include extracellular DNA and histones, which bind to pathogenic bacteria and kill them. Extracellular DNA can inactivate root pathogens in two ways. First, extracellular DNA can ‘trap’ pathogens, as described in the section “Biotic stress and root mucilage”. Second, extracellular DNA is directly toxic to pathogens because it disrupts cell membranes and causes cell lysis by chelating cations that stabilize the microbial cell structure (Mulcahy et al., 2008). Pea root tips that secreted extracellular DNA had only 5% root infection due to the fungal pathogen *Nectria haematococca*, but the infection rate reached 100% after adding exogenous DNase I, which degrades the extracellular DNA (Wen et al., 2007). Histone also appears to be directly toxic to pathogens. Live pathogenic bacteria populations were reduced by 55% following exposure to $10\ \mu\text{g mL}^{-1}$ of histone H4 extracted from pea root, and the effect vanished when the anti-histone H4 antibody was added to the exposure media (Tran et al., 2016). The protective effect of extracellular DNA and histones may be enhanced by the presence of other antimicrobial substances that are exuded by roots (e.g., flavonoids, strigolactones), but this still needs to be established. These examples are compelling evidence that mucilage helps plants to inactivate dangerous root pathogens.

Research methods for mucilage collection

Mucilage is released from roots, and most roots grow below the soil surface. It is difficult to measure mucilage from these roots without destructively sampling the plant. However, it is possible to get mucilage from undisturbed, living plants that have aerial roots, such as the brace roots of corn. Aerial roots grow above-ground, so they can be cut at the soil surface and immersed in a plastic cylinder filled with water for 24 h to stimulate mucilage production. When the cylinder is removed, the brace root is covered with mucilage, which is then collected with a syringe to quantify the daily mucilage production and analyze its chemical composition. This is a suitable method to evaluate mucilage production in aerial roots of plants growing in field or controlled environments, since

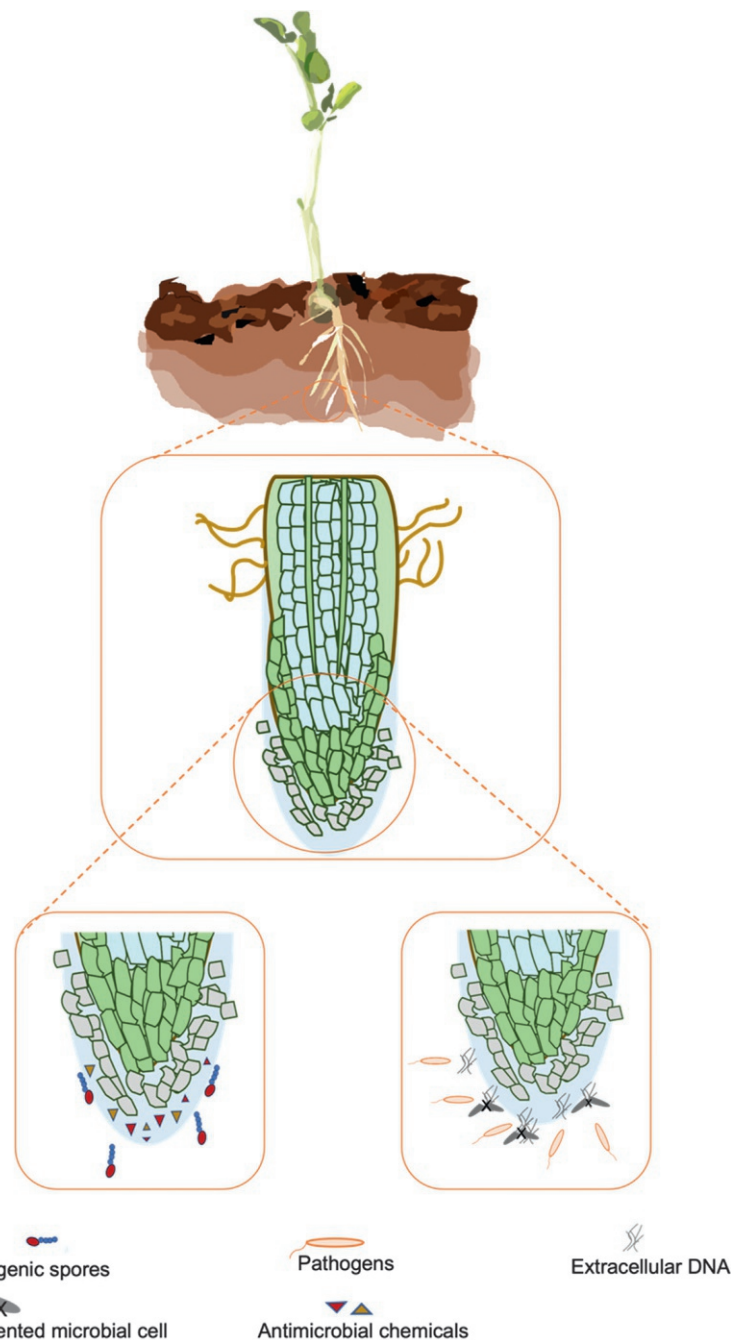


Fig. 3 Schematic representation of plant root defenses associated with mucilage secretion. Mucilage creates a sticky physical barrier that adhere to microbial spores, preventing the spore from coming into contact with root cells. Mucilage also contains lytic substances like extracellular DNA, histones and other antimicrobial substances that are directly or indirectly toxic to pathogenic microorganisms.

it allows for repeated sampling of the same plant to understand its response to abiotic and biotic stresses, and to monitor the mucilage production at key growth stages.

In contrast, roots growing in the soil profile must be sampled destructively by excavating the root mass, then manually removing most of the adhering soil, followed by brushing and washing to clean the roots. Intense root manipulation is likely to create artifacts that affect mucilage production and its chemical composition. Consequently, hydroponic and aeroponic methods are generally considered to be more reliable ways to collect mucilage from plants with fibrous and taproot systems.

Mucilage is commonly sampled from plant roots in axenic hydroponic systems, which are free of other organisms. After the seedling is established and its roots are partially or completely immersed in the nutrient solution, which is replenished regularly with drip irrigation, the hydroponic system is flushed to remove salts and other residuals. During the measurement period, mucilage secreted into fresh nutrient solution is recovered by freeze-drying, followed by centrifugation, filtration and dialysis steps to isolate

mucilage fractions. The method provides a snapshot of the mucilage production and its characteristics at a defined point in time. However, it is not the best way to collect mucilage for follow-up studies because mucilage loses its structural properties after it is exposed to the nutrient solution, and these cannot be restored after freeze-drying. Mucilage collected in hydroponic systems is similar to, but not identical to the mucilage released under field conditions. Clearly, plants grown in hydroponic environments are not exposed to the same physical, chemical and biological stresses (e.g., inadequate water supply, nutrient deficiency, pathogens) that are experienced routinely by field-grown plants. Therefore, mucilage production and mucilage chemistry in hydroponic systems can only approximate how plant roots act in natural field environments.

Aeroponic systems, which expose plant roots to nutrient-containing aerosol droplets, are an interesting alternative for sampling root mucilage. A 'mucilage separator' for aeroponic systems allowed researchers to collect hydrated, intact mucilage gel (Zickenrott et al., 2016). Briefly, this involves seed germination under aeroponic condition, followed by seedling placement inside a polyethylene bottle that was installed on a custom-designed, rotating tray. As the tray rotated, the centrifugal force separated individual droplets of mucilage from the tip of the developing radicle, and the mucilage droplets were collected within 15 min of secretion from the root. This is a reliable method to recover mucilage from young, recently-germinated seedlings in the first 10 d of their growth. However, more work is needed to adapt the method to older plants (>10 d), which need more space to accommodate their root growth. The aeroponic method for mucilage collection has advantages because it avoids excessive root manipulation associated with excavating plant roots from soil, and it prevents root exposure to salts in hydroponic nutrient solutions. Further development of the aeroponic method will provide us with deeper insight into the multiple functions of root mucilage in diverse plant species.

Summary and outlook

Mucilage is responsible for rhizosheath formation and protects root meristems as they extend vertically and laterally in the soil profile. However, it is emerging that mucilage has several other vital functions, including water retention, rhizobiome assembly and plant defense responses in the rhizosphere. The water-attracting functions of mucilage are critically important in regions where soil water supply is less than crop water requirements, either due to natural aridity or because of an emerging water shortage associated with global climate change. Since mucilage sustains the water reserves in the rhizosphere, it is interesting to consider how plant species adjust their mucilage production in water-deficient soils. This could lead to the identification of drought-tolerant root phenotypes for agricultural use. Cultivars with superior mucilage properties could be included in crop rotations or retained in crop breeding programs. Since mucilage mediates the release of plant signals to potential symbiotic and beneficial soil microorganisms, it needs to be considered in understanding crop associations with plant growth-promoting rhizobacteria that occur naturally in field soils or that are added as inoculants. Finally, the defensive attributes of mucilage need further investigation, as certain mucilage traits could boost the inherent resistance of plants to infective pathogens. Enhancing the ability of plants to protect themselves from diseases is expected to reduce our reliance on synthetic pesticides. Therefore, ongoing research into root mucilage and its functions will improve our understanding of plant resilience in an increasingly stressful world.

References

- Ahmed MA, Kroener E, Holz M, Zarebanadkouki M, and Carminati A (2014) Mucilage exudation facilitates root water uptake in dry soils. *Functional Plant Biology* 41(11): 1129–1137.
- Ahmed MA, Banfield CC, Sanaullah M, Gunina A, and Dippold MA (2018a) Utilisation of mucilage C by microbial communities under drought. *Biology and Fertility of Soils* 54(1): 83–94.
- Ahmed MA, Sanaullah M, Blagodatskaya E, Mason-Jones K, Jawad H, Kuzyakov Y, and Dippold MA (2018b) Soil microorganisms exhibit enzymatic and priming response to root mucilage under drought. *Soil Biology and Biochemistry* 116: 410–418.
- Bergmann D, Zehfus M, Zierer L, Smith B, and Gable M (2009) Grass rhizosheaths: Associated bacterial communities and potential for nitrogen fixation. *Western North American Naturalist* 69(1): 12. <https://scholarsarchive.byu.edu/wnan/vol69/iss1/12>.
- Cai M-Z, Wang F-M, Li R-F, Zhang S-N, Wang N, and Xu G-D (2011) Response and tolerance of root border cells to aluminum toxicity in soybean seedlings. *Journal of Inorganic Biochemistry* 105(7): 966–971.
- Cannesan MA, Durand C, Burel C, Gangneux C, Lerouge P, Ishii T, Laal K, Follet-Gueye M-L, Driouch A, and Vire-Gibouin M (2012) Effect of arabinogalactan proteins from the root caps of pea and *Brassica napus* on *Aphanomyces euteiches* zoospore chemotaxis and germination. *Plant Physiology* 159(4): 1658–1670.
- Castilleux R, Plancot B, Ropitiaux M, Carreras A, Leprince J, Boulogne I, Follet-Gueye ML, Popper ZA, Driouch A, and Vire M (2018) Cell wall extensins in root-microbe interactions and root secretions. *Journal of Experimental Botany* 69(18): 4235–4247.
- Clowes FAL and Wadekar R (1988) Modelling of the root cap of *Zea mays* L. in relation to temperature. *New Phytologist* 108(3): 259–262.
- Crews LJ, McCully ME, and Canny MJ (2003) Mucilage production by wounded xylem tissue of maize roots—time course and stimulus. *Functional Plant Biology* 30(7): 755–766.
- Duell RW and Peacock GR (1985) Rhizosheaths on mesophytic grasses 1. *Crop Science* 25(5): 880–883.
- Ghanem ME, Han RM, Classen B, Quetin-Leclercq J, Mahy G, Ruan C-J, Qin P, Perez-Alfocea F, and Lutts S (2010) Mucilage and polysaccharides in the halophyte plant species *Kosteletzkya virginica*: Localization and composition in relation to salt stress. *Journal of Plant Physiology* 167(5): 382–392.
- Ghezzehei TA and Albalasmeh AA (2015) Spatial distribution of rhizodeposits provides built-in water potential gradient in the rhizosphere. *Ecological Modelling* 298: 53–63.
- Gunawardena U and Hawes MC (2002) Tissue specific localization of root infection by fungal pathogens: Role of root border cells. *Molecular Plant-Microbe Interactions* 15(11): 1128–1136.
- Horst WJ, Wagner A, and Marschner H (1982) Mucilage protects root meristems from aluminium injury. *Zeitschrift für Pflanzenphysiologie* 105(5): 435–444.
- Huang Z and Gutterman Y (1999) Water absorption by mucilaginous achenes of *Artemisia monosperma*: floating and germination as affected by salt concentrations. *Israel Journal of Plant Sciences* 47(1): 27–34.
- Jalmi SK and Sinha AK (2022) Ambiguities of PGPR-induced plant signaling and stress management. *Frontiers in Microbiology* 13: 899563. <https://doi.org/10.3389/fmicb.2022.899563>.

- Kaczmarek A, Pieczywek PM, Cybulska J, and Zdunek A (2021) Structure and functionality of rhamnogalacturonan I in the cell wall and in solution: A review. *Carbohydrate Polymers* 278: 118909. <https://doi.org/10.1016/j.carbpol.2021.118909>.
- Kravchenko LV, Azarova TS, Makarova NM, and Tikhonovich IA (2004) The effect of tryptophan present in plant root exudates on the phytostimulating activity of rhizobacteria. *Microbiology* 73(2): 156–158.
- Kroener E, Zarebanadkouki M, Kaestner A, and Carminati A (2014) Nonequilibrium water dynamics in the rhizosphere: How mucilage affects water flow in soils. *Water Resources Research* 50(8): 6479–6495.
- Kuzakov Y (2002) Review: Factors affecting rhizosphere priming effects. *Journal of Plant Nutrition and Soil Science* 165(4): 382–396.
- McNear DH Jr. (2013) The rhizosphere—roots, soil and everything in between. *Nature Education Knowledge* 4(3): 1. <https://www.nature.com/scitable/knowledge/library/the-rhizosphere-roots-soil-and-everything-in-between>.
- Mulcahy H, Charron-Mazeno L, and Levenza S (2008) Extracellular DNA chelates cations and induces antibiotic resistance in *Pseudomonas aeruginosa* biofilms. *PLoS Pathogens* 4(11): e1000213. <https://doi.org/10.1371/journal.ppat.1000213>.
- Nambiar EKS (1976) Uptake of Zn65 from dry soil by plants. *Plant and Soil* 44(1): 267–271.
- Nazari M (2021) Plant mucilage components and their functions in the rhizosphere. *Rhizosphere* 18: 100344. <https://doi.org/10.1016/j.rhisph.2021.100344>.
- Nazari M, Riebeling S, Banfield CC, Akale A, Crosta M, Mason-Jones K, Dippold MA, and Ahmed MA (2020) Mucilage polysaccharide composition and exudation in maize from contrasting climatic regions. *Frontiers in Plant Science* 11: 587610. <https://doi.org/10.3389/fpls.2020.587610>.
- Price SR (1911) The roots of some North African desert-grasses. *New Phytologist* 10(9): 328–340.
- Ralet M-C, Crepeau M-J, Vigouroux J, Tran J, Berger A, Salle C, Granier F, Botran L, and North HM (2016) Xylans provide the structural driving force for mucilage adhesion to the *Arabidopsis* seed coat. *Plant Physiology* 171(1): 165–178.
- Read DB, Bengough AG, Gregory PJ, Crawford JW, Robinson D, Scrimgeour CM, Young IM, Zhang K, and Zhang X (2003) Plant roots release phospholipid surfactants that modify the physical and chemical properties of soil. *New Phytologist* 157(2): 315–326.
- Ropitiaux M, Bernard S, Follet-Gueye M-L, Vicré M, Boulogne I, and Driouich A (2019) Xyloglucan and cellulose form molecular cross-bridges connecting root border cells in pea (*Pisum sativum*). *Plant Physiology and Biochemistry* 139: 191–196.
- Ropitiaux M, Bernard S, Schapman D, Follet-Gueye ML, Vicre M, Boulogne I, and Driouich A (2020) Root border cells and mucilage secretions of soybean, *Glycine Max* (Merr) L.: Characterization and role in interactions with the oomycete *Phytophthora parasitica*. *Cell* 9(10): 2215. <https://doi.org/10.3390/cells9102215>.
- Schenkel D, Maciá-Vicente JG, Bissell A, and Splivallo R (2018) Fungi indirectly affect plant root architecture by modulating soil volatile organic compounds. *Frontiers in Microbiology* 9: 1847. <https://doi.org/10.3389/fmicb.2018.01847>.
- Tran TM, MacIntyre A, Hawes M, and Allen C (2016) Escaping underground nets: extracellular DNases degrade plant extracellular traps and contribute to virulence of the plant pathogenic bacterium *Ralstonia solanacearum*. *PLoS Pathogens* 12(6): e1005686. <https://doi.org/10.1371/journal.ppat.1005686>.
- Watanabe T, Misawa S, Hiradate S, and Osaki M (2008) Characterization of root mucilage from *Melastoma malabathricum*, with emphasis on its roles in aluminum accumulation. *New Phytologist* 178(3): 581–589.
- Watt M, McCully ME, and Jeffree CE (1993) Plant and bacterial mucilages of the maize rhizosphere: comparison of their soil binding properties and histochemistry in a model system. *Plant and Soil* 151(2): 151–165.
- Wen FS, VanEtten HD, Tsapraillis G, and Hawes MC (2007) Extracellular proteins in pea root tip and border cell exudates. *Plant Physiology* 143(2): 773–783.
- Zhou MX, Classen B, Agneessens R, Godin B, and Lutts S (2020) Salinity improves zinc resistance in *Kosteletzkya pentacarpos* in relation to a modification in mucilage and polysaccharides composition. *International Journal of Environmental Research* 14(3): 323–333.
- Zickenrott IM, Woche SK, Bachmann J, Ahmed MA, and Vetterlein D (2016) An efficient method for the collection of root mucilage from different plant species—a case study on the effect of mucilage on soil water repellency. *Journal of Plant Nutrition and Soil Science* 179(2): 294–302.

Root exudates and microorganisms

Aidan De Sena^a , Krisztina Mosdossy^b , Joann K Whalen^b , and Chandra A Madramootoo^a , ^aDepartment of Bioresource Engineering, Macdonald Campus of McGill University, Ste-Anne-de-Bellevue, QC, Canada; ^bDepartment of Natural Resource Sciences, Macdonald Campus of McGill University, Sainte-Anne-de-Bellevue, QC, Canada

© 2023 Elsevier Ltd. All rights reserved.

Introduction	343
Composition of root exudates	344
Root exudate production	345
Origin and secretion of inorganic atoms and compounds as root exudates	345
Origin and transport of organic compounds to root epidermal cells	347
Secretion of organic root exudates	348
Factors affecting root exudation	349
Plant physiology	349
Abiotic controls	350
Biotic controls	351
Functions of root exudates	351
Plant root architecture and sensing	351
Nutrient mobilization	352
Microbial selection	352
Plant defense	353
Methods of root exudate characterization and analysis	353
Root exudate characterization	353
Analysis of root exudate effects	354
Summary and outlook	354
Acknowledgements	355
References	355

Abstract

Plant root exudates are atoms and compounds that are released from living root cells. Root exudate chemistry and concentrations depend on the plant species and environmental conditions. These exudates increase microbial growth and activity in the rhizosphere, compared to the bulk soil. Some root exudates are signal molecules that attract beneficial and mutualistic microorganisms to grow in and around the roots. Other root exudates provide chemical inhibition and biological control of disease-causing organisms. Methodological advances that improve our fundamental understanding of root exudation will be important to improve plant health and agricultural sustainability in the future.

Key points

- Root exudates are the medium by which plants communicate with the soil microbiome.
- Plant physiology determines the inherent composition of root exudates, which is then modulated by abiotic and biotic factors.
- Root exudates allow plants to sense their soil environment and adapt their root architecture accordingly.
- Root exudates can enhance nutrient availability by solubilizing immobile phosphorus while restricting export of mobile nitrogen.
- Root exudation exerts the selecting force that determines the rhizosphere microbiome.
- Root exudates released by constitutive and induced processes protect plants against root pathogens.

Introduction

The rhizosphere is at the interface of the plant root system and the soil environment. Here, plant roots sense the abiotic and biotic factors in the surrounding soil and respond directly to them by modifying root growth and interacting with the matrix of the root-associated soil. The ability of a plant to adjust root growth directly impacts the development and survival of plants and indirectly affects soil ecological functions. This chapter explains how plant roots interact with the abiotic and biotic components of soil, transforming soil that was originally part of the bulk environment into the dynamic rhizosphere, all through the secretion of root exudates.

Root exudates are chemical forms released from plant roots into the surrounding soil. Possessing diverse chemistry, root exudates include inorganic atoms and compounds, as well as organic compounds of both low- and high-molecular weight. The amount and chemical composition of root exudates released vary in space and time, creating a dynamic interface between roots and their associated soil. At the scale of an individual plant, root exudation depends on plant genetics, stage of development, root system architecture and nutritional status. Furthermore, the type of exudates and rate of release are influenced by environmental factors related to abiotic (e.g., light, atmosphere, temperature, soil texture, moisture, salinity) and biotic conditions (e.g., plant competition, herbivory, microbial relationships). Many of these environmental factors fluctuate with time, which means that root exudation adapts to maintain the rhizosphere functions.

Root exudates often change the soil microenvironment around the roots for the benefit of plants. For example, root cells secrete organic acids that lower the soil pH and dissolve phosphate minerals, which liberates phosphate ions that can be absorbed by the root. Some root exudates act as signal molecules, like the strigolactones. These terpenoid-derived molecules allow symbiotic fungi to detect a potential plant host. In addition, strigolactones function as plant hormones that regulate several developmental processes, allowing shoot and root architecture to adapt to environmental conditions. Many root exudates are simple organic substrates that are readily metabolized by free-living and root-associated microorganisms (e.g., archaea, bacteria, fungi). The free-living microorganisms derive energy and nutrients from these substrates to support their activities. Free-living microorganisms living near the root are known to transfer nutrients to plants, provide hormonal control of plant growth and support plant immunity. Hence, the connection between root exudates and microbial responses in the rhizosphere is a key determinant of plant tolerance to fluctuating growing conditions.

From this perspective, root exudation appears to be an adaptive strategy that ensures plant survival and allows for effective communication between plant roots and soil microorganisms. While immense progress has been made in the analysis of root exudates and their impact on processes in the plant-soil system, more insights are forthcoming. In this chapter, we discuss the emerging techniques that can deepen our understanding of root exudates and the outcomes of plant-microbial communication for rhizosphere functions. Such knowledge may allow us to use root exudates as an indicator of plant health, with regards to plant nutrition and disease suppression, as well as provide insight into the role of the rhizosphere as both a source and sink of greenhouse gases via the release of root exudates. Root exudation could reflect the degree to which inherent biological processes are optimized. There are indications that plants with a particular profile of root exudates might be sufficiently well-nourished that they require less fertilizer or have adequate immunity and growth potential to reduce pesticide use. Likewise, the control that root exudates exert on microbial metabolism suggests the potential for engineering the rhizosphere to manage greenhouse gas emissions from soil. Hence, there is considerable interest in studying root exudation as a means to achieve economical crop yield targets while lowering agrochemical inputs and greenhouse gas emissions. These possibilities are fully aligned with the principles for sustainable intensification of agriculture.

Composition of root exudates

There is a vast diversity of root exudates. We define root exudates as atoms and compounds released passively and actively from plant roots as solids, liquids and gases into the soil environment. Consequently, root exudates range from atoms to complex macromolecules. They can be inorganic atoms and compounds, like protons, bicarbonate, and carbon dioxide (Table 1), as well as low- and high-molecular weight organic compounds like amino acids, simple sugars, organic acids, fatty acids, sterols, plant growth factors, vitamins, enzymes and proteins, phenolics, terpenes and terpenoids (Table 2).

Plants release inorganic and organic substances in their root exudates. As is evident from the divergent chemical structures of inorganic and organic exudates, different mechanisms are responsible for their synthesis, transport and release. As such, there are unique metabolic processes, organelles, cells and tissues involved in the production pathways of various exudates.

Table 1 Examples of inorganic atoms and compounds exuded by plants and exporter protein families related to their release.^{a–c}

Chemical Class	Exporter Families ^d	Examples
Ionic Atoms and Compounds	ATPase, Anion Channels ^e	Bicarbonate (HCO ₃ ⁻), hydroxide (OH ⁻), hydrogen proton (H ⁺)
Gas Compounds		Carbon dioxide (CO ₂), hydrogen gas (H ₂)

^aAdapted from Haichar F, Santaella C, Heulin T and Achouak W (2014) Root exudates mediated interactions belowground. *Soil Biology and Biochemistry* 77: 69–80. <https://doi.org/10.1016/j.soilbio.2014.06.017>.

^bAdapted from Sasse J, Martinoia E and Northen T (2018) Feed your friends: Do plant exudates shape the root microbiome? *Trends in Plant Science* 23: 25–41. <https://doi.org/10.1016/j.tplants.2017.09.003>.

^cAdapted from Vives-Peris V, de Ollas C, Gómez-Cadenas A and Pérez-Clemente RM (2020) Root exudates: From plant to rhizosphere and beyond. *Plant Cell Reports* 39: 3–17. <https://doi.org/10.1007/s00299-019-02447-5>.

^dExporter protein families known or suspected to release these exudate classes are listed.

^eSimple diffusion.

Table 2 Examples of organic compounds exuded by plants and exporter protein families related to their release.^{a–d}

Chemical Class	Exporter Families	Examples
Amino acids	GDU, BAT1 ^e , SIAR2 ^e , CAT ^e	α -Alanine; β -alanine; γ -aminobutyric acid; α -aminoadipic acid; arginine; asparagine; aspartic acid; citrulline; cystathionine; cysteine; cystine; deoxymugineic acid; 3-epihydroxymugineic acid; glutamine; glutamic acid; glycine; histidine; homoserine; isoleucine; leucine; lysine; methionine; mugineic acid; ornithine; phenylalanine; proline; serine; threonine; tryptophan; tyrosine; valine
Sugars	MFS ^e	Arabinose; deoxyribose; fructose; galactose; glucose; maltose; mannose; oligosaccharides; raffinose; rhamnose; ribose; sucrose; xylose
Organic acids	ALMT, MATE	Acetic; aconitic; ascorbic; aldonic; benzoic; butanoic/butyric; caffeic; <i>t</i> -cinnamic; citric; <i>p</i> -coumaric; erythronic; ferulic; formic; fumaric; glutaric; glycolic; lactic; malic; malonic; oxalacetic; oxalic; <i>p</i> -hydroxybenzoic; piscidic; propionic; pyruvic; succinic; syringic; tartaric; tetric; valeric; vanillic acids
Fatty acids and derivatives	ABC	(<i>E</i>)-2-hexenal; L-ascorbyl 2,6-dipalmitate; hexadecanal; lauric acid; linoleic acid; linolenic acid; oleic acid; palmitic acid; stearic acid
Sterols	ABC	Campesterol; cholesterol; sitosterol; stigmasterol
Growth factors	ABC	<i>p</i> -Aminobenzoic acid; auxins; choline; inositol; jasmonic acid, strigolactone (strigol; sorgolactone)
Vitamins	– ^f	Biotin; niacin; pantothenate; riboflavin; thiamin; pyridoxine
Enzymes and other proteins	– ^f	Amylase; β -1,3-glucanase; invertase; lipase; pathogenesis-related proteins; peroxidase; phenolase; acid/alkaline phosphatase; polygalacturonase; protease
Nucleotides/purines	P-Type ATPase, ANT ^e , ABC ^e	Adenine; cytidine; guanine; uridine
Phenolics and derivatives	MATE ^e , ABC ^e	Benzoxazinoids: AMPO (2-amino-7-methoxy-3H-phenoxazin-3-one); APO (2-amino-3H-phenoxazin-3-one); DIBOA (4-dihydroxy-2H-1,4-benzoxazin-3(4H)-one); DIMBOA (2,4-dihydroxy-7-methoxy-2H-1,4-benzoxazin-3(4H)-one) Chalcones: 4,4'-dihydroxy-2'-methoxychalcone; isoliquiritigenin Coumarins: esculetin; esculin; fraxetin; herniarin; scopoletin; scopoline; umbelliferone Coumestans: coumestrol (Iso)Flavonoids: flavanones (4',7'-dihydroxyflavanone; eriodictyol; liquiritigenin; naringenin); flavones (4',7'-dihydroxyflavone; 7,3'-dihydroxy-4'-methoxyflavone; 3,5,7,3'-tetrahydroxy-4'-methoxyflavone; luteolin); flavonols (+ and – catechin); isoflavones (daidzein; genistein; prunetin); proanthocyanidins (selliguenin A) Others: 2,6-ditertbutyl- <i>p</i> -cresol; rosmarinic acid; salicylic acid Terpenes: diterpenes (brachialactone; rhizathalene A); homoterpenes (DMNT [(<i>E</i>)-4,8-dimethyl-1,3,7-nonatriene]); sesquiterpenes ((<i>E</i>)- β -caryophyllene) Terpenoids: diterpenoids (momilactone A; kauralexin A); monoterpenoids (1,8-cineole; camphor); sesquiterpenoids (geijerene; pregeijerene); triterpenoids (glycinoeclepin A, B, and C; solanoeclepin A; avenacin A)
Others	– ^f	Al-induced polypeptides, alcohols, benzoquinones (sorgoleones), camalexin, ethanol, glucosides (glucosinolates), glycinebetaine, phthalates (dibutyl and dimethyl phthalate), shikonins, sulfur-containing volatiles (diallyl disulfide, dimethyl sulfide, dimethyl disulfide, dimethyl sulfoxide, isothiocyanates, methanethiol, thiocyanates, thiocyanic acid), unidentified ninhydrin positive compounds, unidentifiable soluble proteins

^aAdapted from Haichar F, Santaella C, Heulin T and Achouak W (2014) Root exudates mediated interactions belowground. *Soil Biology and Biochemistry* 77: 69–80. <https://doi.org/10.1016/j.soilbio.2014.06.017>.

^bAdapted from Massalha H, Korenblum E, Tholl D and Aharoni A (2017) Small molecules below-ground: The role of specialized metabolites in the rhizosphere. *The Plant Journal* 90: 788–807. <https://doi.org/10.1111/tpj.13543>.

^cAdapted from Sasse J, Martinoia E and Northen T (2018) Feed your friends: Do plant exudates shape the root microbiome? *Trends in Plant Science* 23: 25–41. <https://doi.org/10.1016/j.tplants.2017.09.003>.

^dAdapted from Vives-Peris V, de Ollas C, Gómez-Cadenas A and Pérez-Clemente RM (2020) Root exudates: From plant to rhizosphere and beyond. *Plant Cell Reports* 39: 3–17. <https://doi.org/10.1007/s00299-019-02447-5>.

^eSuspected export protein families.

^fUnknown export mechanism.

Root exudate production

Origin and secretion of inorganic atoms and compounds as root exudates

The secretion of inorganic root exudates is crucial to basal processes, like gas diffusion and ion absorption by roots. For example, catabolic activities in roots continually release carbon dioxide (CO₂). In addition, the steady absorption of cations (e.g., ammonium) by plant roots requires the exudation of inorganic hydrogen protons to maintain the electrostatic charge balance of cellular membranes. Thus, the exudation of these inorganic atoms and compounds are vital to root physiology.

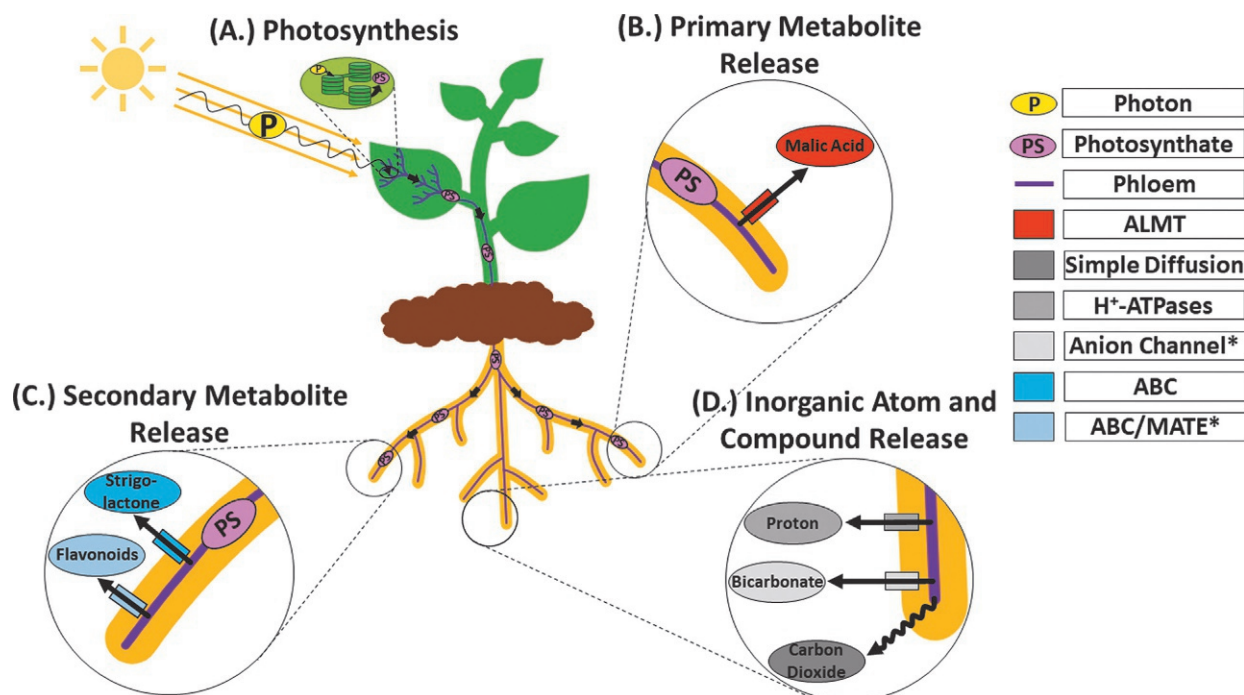


Fig. 1 Root exudate synthesis, transport and release processes. Most root exudates are organic compounds that originated as photosynthates, the carbon-based compounds produced from the conversion of photons into chemical energy via photosynthesis (A). Photosynthates are conveyed from chloroplasts in leaves to the root system through the phloem vascular tissue. In roots, photosynthates can be released directly from the epidermal cells as is or modified by other metabolic processes prior to release as root exudates. Root exudates must cross at least one plasma membrane passively or actively, and often via an enzyme transporter. Primary metabolites (B) include organic acids, amino acids and sugars. Some organic acids have a well-characterized export mechanism, like malic acid by ALuminum-activated Malate Transporters (ALMT), but the export pathways are unknown for most organic acids, amino acids and sugars. The known export pathways for secondary metabolites (C) include transporters of the ATP-Binding Cassette (ABC) family that are responsible for strigolactone release from roots. ABC transporters may be responsible for the export of flavonoids, along with Multidrug and Toxic compound Extrusion (MATE) transporters. Root exudates that are inorganic atoms (e.g., hydrogen protons), or compounds, like bicarbonate ions and carbon dioxide (D), do not originate directly from photosynthates but could be transferred from above-ground tissues to roots by the phloem. Hydrogen protons are exported from roots by H^+ -ATPases transporters, while anion channels are purported to facilitate the diffusion of bicarbonate ions into the soil environment. Carbon dioxide does not require an enzyme transporter for its exudation because it is released passively from roots via simple diffusion. Export pathways requiring experimental validation are indicated with an asterisk. Adapted from Sasse J, Martinoia E and Northen T (2018) Feed your friends: Do plant exudates shape the root microbiome? *Trends in Plant Science* 23: 25–41. <https://doi.org/10.1016/j.tplants.2017.09.003>.

Although some gases and inorganic solutes could originate from above-ground plant organs, the vast majority of inorganic root exudates are derived from the basal physiological processes occurring in roots. These processes are distributed throughout the root cells. In the mitochondria, the tricarboxylic acid cycle generates CO_2 that is released via root respiration, while the autoionization of water in the cytosol produces hydrogen protons and hydroxides. These processes occur continually, ensuring a steady emission of gases and ongoing production of inorganic root exudates.

Root exudates must cross the plasma membrane of the root cells before they enter the surrounding soil (Fig. 1). Some inorganic root exudates do so by passive transport. This mode allows the free movement of inorganic compounds across the membrane along steep electrochemical gradients and includes both simple and facilitated diffusion. Simple diffusion permits small, uncharged exudates, like gases (e.g., CO_2), to freely move through the lipid bilayer (Canarini et al., 2019). In contrast, charged inorganic exudates must pass the lipid bilayer by facilitated diffusion. This transport mode is likely the case for hydroxide and bicarbonate ions, which must move through channel or uniporter transmembrane proteins across the plasma membrane. These forms of passive diffusion permit the free flow of inorganic root exudates into the rhizosphere.

While passive transport accounts for the release of most inorganic root exudates, a direct expenditure of energy is required for hydrogen protons to cross the plasma membrane. Energy-driven movement of hydrogen protons and other substances is known as active transport. Direct export of hydrogen protons requires primary active transport, where uniporter transmembrane proteins, known as proton pumps (H^+ -ATPases), are coupled with adenosine triphosphate (ATP) hydrolysis to transfer hydrogen protons across the lipid bilayer. However, hydrogen protons can also pass through the plasma membrane indirectly if coupled to the secondary active transport of another compound via symporter and antiporter transmembrane proteins (discussed in section “Secretion of organic root exudates”). Active transport pathways guarantee hydrogen proton export into the rhizosphere, which is essential for electrochemical charge balance and controls the pH around the root.

Origin and transport of organic compounds to root epidermal cells

Most root exudates are organic compounds. The reason is that plants allocate a considerable amount of their total fixed carbon (5–21%) to root secretion, even during periods of stress (Haichar et al., 2014). Many of these organic compounds are enriched in nitrogen. In fact, the nitrogen content in root exudates can be up to 15% of the total nitrogen acquired by the plant during the growing season (Sasse et al., 2018). This substantial dedication of carbon and nitrogen to manufacture compounds that are then released from the plant reflects the necessity of root exudation for the plant's well-being.

Of the organic compounds, those categorized as primary metabolites (e.g., amino acids, organic acids, sugars) are exuded in greater quantities than secondary metabolites (Canarini et al., 2019). Primary metabolites are the dominant exudates because they participate in a number of fundamental physiological processes, such as nutrient acquisition. In contrast, secondary metabolites (e.g., flavonoids, aromatics) are a large group of more than 100,000 specialized organic compounds that are released under specific circumstances, unrelated to the basal physiology of plants (Haichar et al., 2014).

Whether primary or secondary metabolites, all organic root exudates and their precursors depend upon the carbon and energy obtained from photosynthesis (Fig. 1). The chlorophyll-containing shoot tissues are where photosynthesis converts photons into chemical energy via the *de novo* synthesis of glyceraldehyde-3-phosphate, the foundational molecule of photosynthates (i.e., sugars and other organic compounds). Normally, complete *de novo* synthesis is energetically expensive, but photosynthesis harnesses the free energy of photons. Consequently, photosynthesis permits the energetically efficient production of organic root exudates.

Other metabolic pathways (e.g., tricarboxylic acid cycle, phenylpropanoid pathway, glutamine synthetase-glutamate synthase cycle) can further modify the glyceraldehyde-3-phosphate into a variety of compounds that are eventually secreted by roots. For primary metabolites, such as sugars and amino acids, only a few biochemical steps are required to transform glyceraldehyde-3-phosphate into these compounds. In contrast, engineering macromolecules for exudation is a complex process. For example, the construction of extracellular enzymes and other proteins involves amino acid synthesis, transcription, translation and folding into their functional structures. Thus, the rate of organic root exudate synthesis is related to the number of metabolic pathways required for their creation.

The metabolic pathways involved in the synthesis of organic root exudates can occur either in a specific organelle or at multiple sites within plant cells. For example, photosynthesis occurs in chloroplasts whereas the tricarboxylic acid cycle occurs in mitochondria. As the tricarboxylic acid cycle supports basal metabolism in all cells, it occurs in shoots and roots and supports a number of functions, in addition to the production of root exudates. In contrast, the phenylpropanoid pathway involves metabolons in the cytosol and endoplasmic reticulum (Biała and Jasiński, 2018). Phenylpropanoids produced within a particular cell can then be transferred to other cells for further biosynthesis into secondary metabolites, like flavonoids and coumarins. For example, phenylpropanoid production may originate in leaves using photosynthates obtained from chloroplasts, which are then transported to the root system so isoflavonoid synthesis can occur in the root cortex. This network means that a significant amount of cellular machinery throughout the plant is required for the production of the diverse root exudates detected in the rhizosphere.

In summary, organic root exudates are formed from photosynthates, byproducts of the tricarboxylic acid cycle, phenylpropanoids and many other organic compounds that originate in shoot cells and are gradually conveyed to the roots. These organic substances may be exuded through the root epidermis without further modification or altered within root cells before their secretion. The degree to which a particular organic root exudate is synthesized in the shoot versus modified in the root is not known, nor do we fully comprehend the time between the inception of a given exudate—or its precursor—in the shoot and subsequent transfer to the root. Transport of organic substances from the shoots will be controlled by the rate at which these materials are loaded into the vascular tissue of the phloem. Phloem loading occurs continuously, but plants can up- and down-regulate phloem transport rates in response to environmental triggers (Gargallo-Garriga et al., 2018). For example, water stress can result in increased loading of certain photosynthates (e.g., proline, pinitol) into the phloem (Canarini et al., 2016). Hence, phloem loading can be both a constitutive and induced response for the transfer of photosynthates and other organic compounds to roots, which eventually results in their exudation into the soil.

Transport of photosynthates from leaves to roots via the phloem requires that photosynthates first travel symplastically from the mesophyll cells to bundle sheath cells and finally reach the phloem parenchyma cells (Thompson and Wang, 2017). To enter the phloem itself, these organic compounds may continue to travel symplastically via plasmodesmata, which serve as a conduit between the phloem parenchyma and sieve element-companion cell complex of the phloem. If plasmodesmata are absent, organic compounds will be released into the apoplast but require active transport to enter the sieve element-companion cell complex. The availability of symplastic or apoplastic pathways for metabolites into the phloem depends on the plant species. Transfer of organic compounds through the phloem is controlled by concentration gradients that create pressurized bulk flow from the source to the sink. Phloem loading with organic metabolites causes water diffusion from the xylem into the phloem. The resulting change in turgor creates sufficient pressure for mass transport through the phloem to the sink, in this case, roots. Rapid phloem unloading will empty the sieve tube elements and permit water diffusion back into the xylem, creating a positive feedback loop that assures continued movement of the photosynthates into roots. Photosynthates are unloaded from the sieve element-companion cell complex of the phloem through funnel plasmodesmata to the phloem-pole pericycle. Phloem unloading for simple metabolites occurs in a convective manner with mass flow and diffusion, whereas larger macromolecules are unloaded in batches to the phloem-pole pericycle. If the compounds are <27 kDa, the organic metabolites may diffuse throughout the root cortex and reach the root epidermis cells (Ross-Elliott et al., 2017). Our current understanding is that larger metabolites remain in the phloem-pole pericycle, where they are likely degraded. Thus, while root exudates include large macromolecules like proteins, these compounds do not appear to originate from shoot biomass and must be constructed in the root cells.

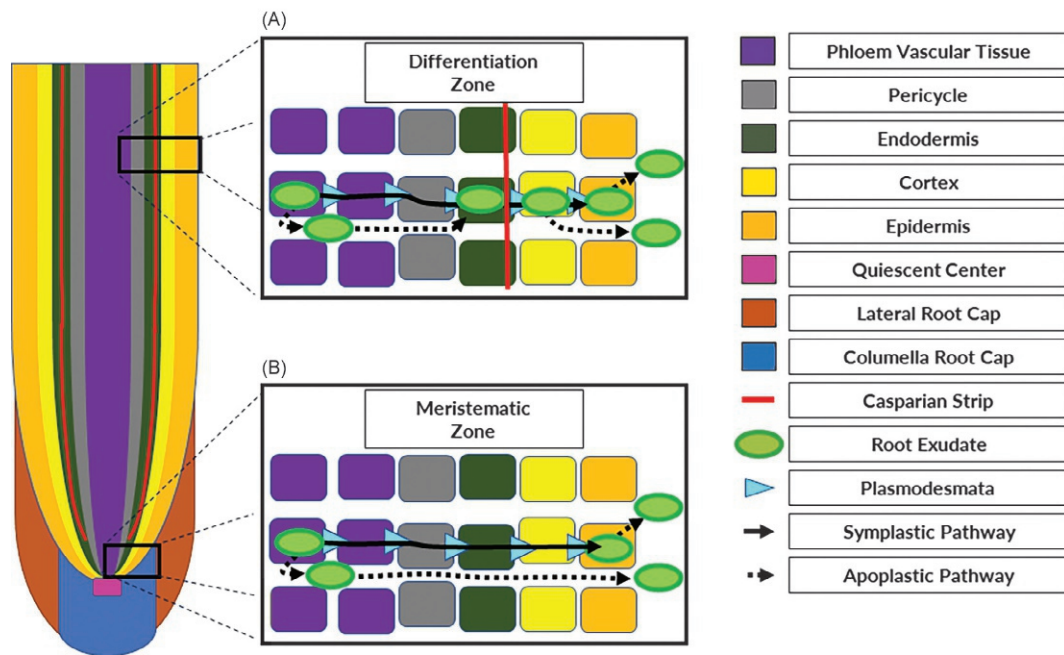


Fig. 2 Release of root exudates. Root exudates can travel from the phloem by the apoplastic (extracellular) or symplastic (intercellular via plasmodesmata) pathways. In the root differentiation zone (A), root exudates move through the pericycle by apoplastic or symplastic processes but must reenter the symplast at the endodermis to traverse the Casparian strip. Upon crossing the Casparian strip, root exudates travel by apoplastic or symplastic routes until they reach the root epidermal cells, where they pass the cell membrane via the apoplast to enter the soil environment. In the undifferentiated meristematic zone (B), root exudates can use apoplastic or symplastic pathways to move through all root tissues (i.e., pericycle, endodermis, cortex) due to the absence of the Casparian strip. However, exudates in the symplast must move into the apoplast to exit the plasma membrane of root epidermal cells and reach the surrounding soil. Adapted from Canarini A, Kaiser C, Merchant A, Richter A and Waneke W (2019) Root exudation of primary metabolites: Mechanisms and their role in plant responses to environmental stimuli. *Frontiers in Plant Science* 10: 157. <https://doi.org/10.3389/fpls.2019.00157>.

Movement of simple organic compounds from the phloem-pole pericycle to the root epidermis depends on whether their exudation occurs in undifferentiated or differentiated root tissue (Fig. 2). In undifferentiated root tissue, organic compounds can travel by either apoplastic or symplastic pathways through the endodermis, cortex and epidermis (Canarini et al., 2019). In differentiated tissue, the simple organic compounds must travel symplastically—at least once—to cross the Casparian strip of the developed endodermis. Additionally, differentiated tissue has fewer plasmodesmata than undifferentiated tissue, which limits the quantities and movement of organic compounds. These constraints mean that far less exudation of simple organic compounds occurs from differentiated root tissues, such as lignified structural roots, than from undifferentiated tissue, like those of the root tip meristematic zone, where the majority of root exudation is expected to take place.

Secretion of organic root exudates

All organic root exudates must exit the plasma membrane of root cells before entering the rhizosphere (Fig. 1). Like the inorganic root exudates, organic root exudates can cross the lipid bilayer of root cells by passive or active means, depending on the size and charge of the compound. Although small, nonpolar organic compounds (e.g., urea, glycerol) can traverse the plasma membrane by simple diffusion, most organic root exudates are passively transported via facilitated diffusion (Sasse et al., 2018; Canarini et al., 2019). Transmembrane proteins like Aluminum-activated Malate Transporters (ALMT) are known to facilitate the exudation of malic acid. Other substrate-specific transporters are likely responsible for facilitating the diffusion of other organic acids as well as amino acids and sugars. Thus, both simple and facilitated diffusion are responsible for the passive release of organic exudates from root cells into the rhizosphere.

Simple and facilitated diffusion allows for the spontaneous exudation of organic compounds from roots, but plants do exert some control over the passive transport of such root exudates. These mechanisms include downregulating gene expression for the transmembrane proteins performing facilitated diffusion, the active reassimilation of root exudates or actively storing root exudates in the plant cell vacuole (Canarini et al., 2019). For example, wheat (*Triticum aestivum* L.) roots can take up some organic exudates at rates 1–175% greater than their efflux via secondary active transport (Warren, 2015). Uncertainty exists regarding whether the benefits from conserving carbon resources outweigh the energy costs required for these controls over the passive release of organic root exudates. Yet, both passive root exudation and the carbon-conserving mechanisms that control this process are physiological traits conserved by plants throughout their life history. As such, the persistence of these plant behaviors despite evolutionary pressures suggests their essential roles in the survival of plants.

Active transport of some organic root exudates occurs by primary active transport, similar to protons. Like H^+ -ATPases, the primary active transport of organic root exudates occurs through uniporters directly coupled with ATP hydrolysis that move organic substances across the plasma membrane against their electrochemical gradient. Nucleotides, peptides, fatty acids, siderophores and strigolactones are exuded in this manner via ATP-Binding Cassette (ABC) transporters, a family of uniporters that is ubiquitous in plants (Vives-Peris et al., 2020). ABC transporters are also thought to export a number of other secondary metabolites from roots into the soil environment, but as yet this function has not been confirmed.

Alternatively, organic root exudates can move across the root epidermis by secondary active transport and vesicle transport mechanisms. Secondary active transport of organic root exudates requires ATP hydrolysis, but the energy is transferred to a H^+ -ATPase rather than the transmembrane protein (Canarini et al., 2019). This pump generates an electrochemical gradient of protons across the plasma membrane, creating energetically favorable conditions that allow the export of organic root exudates. Once this gradient is established, the organic root exudates and protons may be transferred by the transmembrane protein in parallel (via symporters) or in opposite directions (by antiporters). Symporter systems for organic compounds, like inositols, are found in the plasma membranes of vacuoles but are not yet associated with the release of organic compounds into the rhizosphere (Sasse et al., 2018). However, antiporters are known to be responsible for the export of citrate and some phenolics and alkaloids from roots into the rhizosphere, such as through the Multidrug And Toxic compound Extrusion [MATE] transporters (Sasse et al., 2018). Vesicle transport (exocytosis) differs from primary and secondary active transport because it does not involve a transmembrane protein. Instead, this process relies on the formation of vesicles containing root exudates near the endoplasmic reticulum or the Golgi apparatus (Weston et al., 2012). After the vesicle membrane fuses with the plasma membrane and unfolds, the organic root exudates cross the lipid bilayer and are extruded through the cell wall into the soil environment. Vesicle transport is responsible for the exudation of high-molecular weight compounds, such as secondary metabolites.

Overall, most organic root exudates seem to be exported to the soil environment via passive transport rather than active transport. This observation is supported by the fact that individual plant metabolites are more concentrated in the cytosol of root cells ($0.1\text{--}10\text{ mmol L}^{-1}$) than in the rhizosphere ($<0.001\text{ mmol L}^{-1}$; Oburger and Jones, 2018). The 100- to 1000-fold concentration difference of organic substances within the root versus outside the root creates a steep electrochemical gradient that allows for simple or facilitated diffusion. Passive transport requires no direct energy expenditure, and thus should be a preferred pathway for most organic exudates. However, larger organic compounds must be transported via active processes. Research is needed to identify the transporter proteins, energetic requirements and physiological factors that control the active transport of the high-molecular weight organic compounds found in root exudates.

Factors affecting root exudation

The release of root exudates into the rhizosphere is determined by plant physiology and growing conditions. Characteristics of the plant (i.e., genetics, life stage, root morphology, nutrition) determine the intrinsic quantity and chemical composition of root exudates (Fig. 3). Still, plants exhibit considerable phenotypic plasticity that allows them to respond to the external environment by tailoring their profile, concentration and release rate of root exudates. The environmental triggers can be both abiotic and biotic in nature. Some abiotic factors, for example water stress (drought and flooding), can increase the total carbon exuded by plants, whereas exposure to biotic factors, such as herbivory and pathogens, lead to an upregulation of secondary metabolite exudation by the plant. Abiotic and biotic factors can have a profound effect on the chemistry, concentration and rate that root exudates are released into the rhizosphere.

Plant physiology

Root exudation is affected by the life history of plant species. Fast-growing, exploitive plant species (*r*-strategists) tend to have greater rates of carbon release through root exudation than slow-growing, conservative plant species (*K*-strategists; Guyonnet et al., 2018). There are also distinct patterns of root exudation at the phylogenetic level. For example, some root exudates are unique to specific taxonomic groups, like benzoxazinoids, which are secondary metabolites released only by cereal species of the *Poaceae* (Massalha et al., 2017; de Bruijn et al., 2018). However, root exudate profiles can differ among cultivars of the same species, as demonstrated for thale cress (*Arabidopsis thaliana*). Root exudate analysis of 19 cultivars of *A. thaliana* showed more variation in the glucosinolates, followed by flavonoids and then phenylpropanoids (Sasse et al., 2018). This finding suggests that root exudation might explain plant fitness, meaning the ability of plants to survive and reproduce in environments that are sometimes stressful with difficult growing conditions.

The developmental stage of the plant also affects quantity and composition of root exudation. In *A. thaliana*, there are more sugars released into the rhizosphere during the early growth stages, but more amino acids and flavonoids are exuded in the later growth stages (Vives-Peris et al., 2020). The amount of exudate produced by roots peaks during the vegetative growth and onset of reproduction and is least at seedling emergence and senescence stages (Haichar et al., 2014). In general, this temporal variation in the exudation rate is associated in part with the effects of a particular growth stage on photosynthesis, since the generated photosynthates are the main source of energy and carbon required to produce, transport and release root exudates into the rhizosphere.

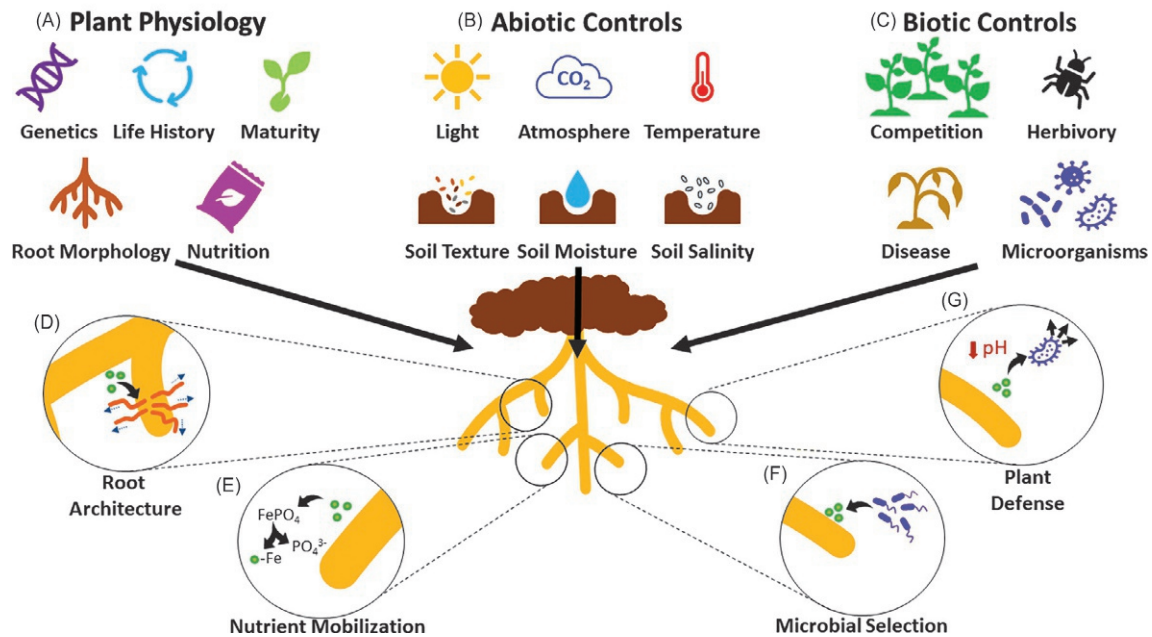


Fig. 3 Factors that affect root exudation (A–C) and its functions (D–G). Plant physiology (A) as well as abiotic conditions (B) and biotic stressors (C) will cause plants to modify the composition and rate of root exudation. In turn, root exudates serve a variety of functions for plant growth and development. Release of strigolactones can modify root architecture by elongating root hairs and increasing their density (D). Some root exudates can mobilize nutrients, like organic acids that solubilize phosphorus by chelating iron (E). Root exudates are signals that alert soil microorganisms to move towards the root surface. For example, flavonoids attract rhizobia to bind onto root hairs, which initiates their mutualistic symbiosis (F). Plants can also protect themselves with root exudates, such as with protons that acidify the rhizosphere, creating an inhospitable environment for some pathogens (G).

Spatial variation in root exudation is clearly related to the root system architecture. The exudation of primary metabolites occurs largely at the root tip, most likely from the meristematic zone where both the lack of a Casparian strip and abundance of plasmodesmata would permit the efficient export of these exudates from the phloem (Canarini et al., 2019). Sloughed border cells from the root are an important source of root exudates, releasing secondary metabolites, like flavonoids and proteins, into the rhizosphere (Massalha et al., 2017; Sasse et al., 2018). Variation in this pattern occurs in some plants with specialized root structures and unique exudation patterns. For example, white lupin (*Lupinus albus*) grows cluster roots that exude large quantities of organic acids, as well as protons, phenolics and enzymes (Sasse et al., 2018). To encourage nodulation, legume species (*Fabaceae* spp.) release specific flavonoids from the root hairs and cortical cells of branching roots, rather than from sloughed root border cells (Haichar et al., 2014). Thus, root morphology will affect the spatial pattern of root exudation.

Plants change the chemistry and rates of root exudation according to their nutritional status. Nitrogen availability is well-known to influence root exudation. During periods of nitrogen stress, plants tend to reduce the quantities of amino acids transported to the roots (Canarini et al., 2019). For example, maize (*Zea mays*) growing under nitrogen deficient conditions reduced the amino acid concentration of root exudates by 75%, on average, compared to the control (Carvalho et al., 2011). Increasing the nitrogen availability also affects root exudates, but the magnitude and direction of the response vary among plant species and growing conditions. Studies with perennial ryegrass (*Lolium perenne*) demonstrated a 153% decrease in below-ground carbon allocation with higher levels of ammonium nitrate fertilizer application (Pausch and Kuzyakov, 2018), whereas an experiment with maize found that increasing the urea-nitrogen application by 130 kg N ha⁻¹ boosted total root exudation by 80% (Zhu et al., 2016). This disparity in effects may result from the greater absolute rate of nitrogen (560 kg N ha⁻¹) in the perennial ryegrass study, whereas maize received up to 160 kg N ha⁻¹. The form of mineral nitrogen fertilizer may have affected the root exudation response as well. Other reports indicate that ammonium fertilizers stimulate more sugar exudation than nitrate fertilizers (Canarini et al., 2019). Similar effects were shown for phosphorus-stressed plants. In soil environments with low phosphorus availability, some plants change their root exudate profile, increasing the production of strigolactones (See section “Plant root architecture and sensing”), organic acids and phosphatase enzymes (See section “Nutrient mobilization”). It is likely that the fertilizer application rate and fertilizer source will induce changes in the composition of root exudates, but this remains to be confirmed for other nutrients and plant species.

Abiotic controls

Root exudation is also affected by abiotic factors, like light, atmosphere, temperature, soil texture, moisture and salinity. Similar to the effects of plant growth and development on the release of organic root exudates, light, atmospheric composition and

temperature strongly influence the photosynthesis rate of plants, and thus the intensity of root exudation. Extended photoperiods stimulate root exudation, likely due to greater photosynthesis. Higher CO₂ levels provide the plant with more carbon for photosynthesis and substantially increase the amount of carbon exuded by roots. For example, a study using ¹⁴C-isotope tracing found that an elevated CO₂ level of 720 ppm increased carbon in the rhizosphere by 18% for perennial ryegrass and by 47% for wheat, on average (Pausch and Kuzyakov, 2018). Temperature can also increase the exudation of organic acids and other compounds (Vives-Peris et al., 2020). The relevance of some of these effects on root exudation (i.e., atmosphere and temperature) is expected to grow in the future, as anthropogenic activities continue to elevate CO₂ concentrations in the atmosphere, subsequently raising the global temperature via the greenhouse effect. Consequently, the release of root exudates may increase with climate change, but this assumption requires robust analysis to account for all factors at play (e.g., flooding) when considering the future climate's impact.

Soil characteristics, like texture, moisture and salinity, can also alter root exudation rate and composition. Soil texture is thought to impact root exudates since clay particles are more reactive and have greater surface area than silt or sand particles (Nguyen, 2009), which could impede the diffusion of root exudates in soils with higher clay content. Two factors that create osmotic stress for roots, moisture and salinity, will likely have a profound impact on root exudation. Water stress increases the amount of carbon-based compounds released by roots. For example, drought conditions increased the proportion of carbon exuded by crested wheatgrass (*Agropyron cristatum*) by 71%, while flooding conditions increased the proportion of exuded carbon by 45% (Vives-Peris et al., 2020). Roots experiencing osmotic stress also modulate the chemistry and concentration of their root exudates. Drought and saline conditions stimulate the exudation of organic acids, flavonoids and terpenoids, and it is postulated that these exudates were signal compounds meant to recruit microbes and initiate symbioses that could alleviate the stress in species like maize and soybean (*Glycine max*; Vives-Peris et al., 2020; Williams and de Vries, 2020). However, other studies found no change in root exudate composition under water stress or saline conditions, as found for sunflower (*Helianthus annuus*), suggesting a species-dependent effect (Canarini et al., 2016). Nevertheless, these stress-induced changes in root exudation may be an adaptive strategy that improves plant survival and warrant further investigation to determine which root exudate compounds are the key to improve plant tolerance to stress.

Biotic controls

In natural environments, plants are confronted with numerous biotic challenges. Competition with other plants, herbivory from predators, disease-causing pathogens and other antagonistic interactions will impact root exudation. To defend against competitors, some plants release allelochemicals from their roots to repel the advances of the neighboring plants into their rhizosphere via inhibiting seed germination and preventing root cell elongation (Vives-Peris et al., 2020). While plant species have adapted to allelochemicals, this mechanism still protects the exuding plant to a degree. Likewise, herbivory triggers the emission of volatile organic compounds from roots that can signal the danger to all plants in the vicinity and enlist microorganisms to protect the plant. This response is seen in maize cultivars that emit sesquiterpenes to recruit nematodes that feed on the herbivorous larvae of Western corn root worm (*Diabrotica virgifera*; Preece and Peñuelas, 2020). Concurrently, the plant will often exude readily-metabolizable organic compounds and signals molecules that stimulate the action of nutrient-cycling microorganisms to aid in compensatory plant growth (Williams and de Vries, 2020). Plants exude metabolites that attract immune response-inducing microbes, to provide natural biocontrol (Sasse et al., 2018). Such mechanisms can partially protect the plant from biotic stresses.

There are other biotic interactions that affect root exudation. Rhizosphere microorganisms can rapidly assimilate the simple organic compounds in root exudates, based on the elevated respiration in the rhizosphere compared to the bulk soil. Microbial depletion of these organic substrates will preserve the steep electrochemical gradients between the root cytoplasm and soil solution, which is necessary for passive diffusion (Haichar et al., 2014). In addition, microbes release metabolites that bind with receptors on the root surface cells and modulate root exudation (Zipfel and Oldroyd, 2017). There is also evidence that certain microbial metabolites increase the solubility of the plasma membrane in root cells, which enhances the diffusion of root exudates (Oburger and Jones, 2018). Through these mechanisms, rhizosphere microorganisms exert some control over root exudation.

Functions of root exudates

There is a growing appreciation for the multifunctional nature of root exudates. Root exudation allows plants to adjust to and tolerate fluctuations in their growing environment, maintain their nutrient status, interact with their microbial neighbors, and ensure their healthy growth and development (Fig. 3).

Plant root architecture and sensing

Once released, root exudates interact with the soil, whether by adsorbing to soil particles, dissolving in soil pore water, or diffusing into the air-filled compartment of the pore. Although the residence time of a given root exudate is short due to their rapid assimilation by soil microbes, their brief interaction with the soil matrix can signal the quality of the soil microsite (i.e., nutrient availability) to the neighboring roots, as well as the origin root itself. This will allow the plant to optimize their root architecture in variable environments and explains the plasticity in root development that is characteristic of successful, high-yielding plants.

Changes in root morphology are associated with some of the primary metabolites. For example, high concentrations of L-glutamate (10^{-5} M) trigger root tip branching in *A. thaliana*, likely to enhance nutrient foraging (Canarini et al., 2019). Organic acid anions are another indicator of the soil nutrient status that induce changes in root growth. Malate indicates low phosphorus availability by transporting iron to the root tip (Canarini et al., 2019). This signal triggers a modification in the root morphology, arresting growth at the root apical meristem and stimulating lateral root branching. This response coincides with sugar accumulation in roots, suggesting sugars may also function as a signal molecule in phosphorus-deficient soil conditions (Canarini et al., 2019). This proposed mechanism is supported by the exogenous application of sugars to roots and their subsequent suspension of growth at the root tip. Thus, primary metabolites initiate nutrient starvation pathways that modify root development.

Secondary metabolites also influence root architecture. For example, the exudation of strigolactones increases the development of dense root hairs and stimulates root hair elongation under poor soil phosphorus conditions (Haichar et al., 2014). Strigolactones are also released by plant roots that are associated with symbiotic arbuscular mycorrhizal fungi (AMF) to trigger hyphal branching of potential fungal symbionts prior to colonization (Haichar et al., 2014). These strigolactone-induced modifications to root hairs and hyphae generate an extensive network that can efficiently acquire the limited supply of phosphate ions and other essential substances in low nutrient environments.

Nutrient mobilization

Root exudates also help mobilize immobile nutrients, such as phosphorus, through chemical and biochemical reactions. The release of organic acids (e.g., malic acid, citric acid) dissolves metal-phosphate complexes by chelating metals (e.g., iron, aluminum) and solubilizing phosphate ions, which are then available to plants (Sasse et al., 2018). Protons exuded by roots alter the pH of the rhizosphere, which accelerates the chemical weathering of clay minerals and releases phosphate and other ions adsorbed to soil surfaces. Phosphatase enzymes are another mobilizing agent exuded from the roots of some plant species (e.g., *L. albus*). These enzymes can release available phosphate ions via the hydrolysis of certain organic phosphorus compounds—such as simple phosphomonoesters, phosphodiester and inositol hexakisphosphates—if in the proximity of the active phosphatase (Preece and Peñuelas, 2020). These foraging strategies, mediated by root exudates, assist plants in meeting their phosphorus nutrition requirements, in conjunction with the phosphorus obtained through any mycorrhizal associations.

Some root exudates can restrain the transformation of nitrogen forms into mobile species, such as nitrate. High nitrate levels in soil are undesirable because nitrate is susceptible to leaching and denitrification reactions that result in its loss from the soil-plant system. One strategy exhibited by some plants, like sorghum (*Sorghum* sp.), is to release root exudates that prevent the microbial transformation of ammonium into nitrate via ammonia and nitrite oxidation (Haichar et al., 2014). These biological nitrification inhibitors released in root exudates, such as methyl 3-(3-hydroxyphenyl) propionate, sorgoleone, and brachialactone, hold promise for restraining the nitrate-forming reactions in soil, as well as their production of nitrous oxide, a greenhouse gas. Currently, genes capable of producing nitrification inhibitors are being incorporated into commercial cultivars of wheat to improve nitrogen use efficiency and prevent nitrate loss through leaching or denitrification (Subbarao et al., 2021). Future use of this technology should result in improved crop performance, reduced nitrate pollution in waterways, and lower nitrous oxide emissions from cultivated fields.

Microbial selection

Root exudates can attract or repel soil microorganisms, which selects for microbial colonizers in the rhizosphere. This effect is driven by chemical gradients, created by the release of specific root exudates as signal molecules in the soil pore water. The proximity of microbial taxa to the release point of the root exudate (e.g., lateral root, root tip meristem) will determine their success in colonizing the growing root surface (Sasse et al., 2018). Microorganisms sense chemical gradients, and those that are mobile will move towards or against the strength of the signal via chemotaxis (Haichar et al., 2014). Once free-living microorganisms move into the rhizosphere or attach to the root surface, they can use the inorganic and organic root exudates for a variety of metabolic processes. Electron transport, respiration and other biological processes in microbial cells depend upon the energy, hydrogen protons, organic acids and carbon substrates in the root exudates. Free-living microorganisms may metabolize substrates exuded from the root into the soil pore water or acquire them indirectly, by cross feeding on the byproducts of microbial metabolism. Plants may deliberately target a particular free-living microorganism. For example, drought-stressed maize plants exude malic acid as a chemoattractant for *Bacillus subtilis*, a bacterium that can secrete osmolytes which maintain plant stomatal integrity and limit their water loss (Williams and de Vries, 2020).

Plants also use root exudates to select for symbiotic microorganisms, which are more intimately associated with roots since they can establish a mutualistic relationship that allows microbial colonization inside the root. To establish the mutualistic relationship, microorganisms and plants must communicate via chemical signals. The chemical signals from plants are transmitted through root exudation. Root-derived compounds, like strigolactones and flavonoids, are responsible for initiating the microbial symbiosis. These signals initiate a cascade of events that allow microorganisms like *Frankia*, rhizobia and mycorrhizal fungi to enter the root, proliferate and function in a way that is mutually beneficial to the plant host and the symbiotic microorganism (Haichar et al., 2014). The simplest mutualism is one in which the plant host provides carbon substrates, and the microbes supply the plant with growth-limiting nutrients. There is evidence that symbiotic plant-microbial relationships improve plant defenses, as discussed in the next section.

Plant defense

Mycorrhiza, the symbiotic relationship between plant roots and fungi, improves plant defenses through physical and biochemical means. The extensive hyphae growing on and around the root surface can function as a physical barrier that prevents pathogens from infecting roots. Second, mycorrhizal fungi can modify root metabolism and root exudate composition to protect the plant host from pathogens. These effects were shown for tomato plants (*Solanum lycopersicum*) that had a 108% lower rate of root penetration by root-knot nematodes (*Meloidogyne incognita*) when 79% of the root surface was colonized by AMF than when no AMF were present (Vos et al., 2012). Furthermore, root exudates collected from the AMF-colonized tomato plants were able to briefly paralyze the nematode pathogens. Thus, these symbiotic relationships initiated by root exudates are able to defend the plant from pathogenic attack via mycorrhiza-induced resistance.

Nevertheless, plants have an innate capacity for defense that does not rely upon the mycorrhizal symbiosis. For instance, root exudates like indols, flavonoids, and other secondary metabolites exert an antimicrobial effect that can inhibit the growth of pathogenic microorganisms in the rhizosphere (Haichar et al., 2014). Extracellular enzymes are another way that plants defend against pathogens because enzymes can inactivate the cellular machinery that root pathogens use to enter the root. Enzymes and antimicrobial compounds can be released preemptively or upon attack. Additionally, hydrogen protons in root exudates can acidify the rhizosphere and make it an inhospitable environment for some pathogenic microorganisms. Consequently, root exudation itself can protect plants from infections that limit their growth and development.

Methods of root exudate characterization and analysis

Root exudates are an ephemeral and dynamic pool of substances that are continually produced and metabolized by rhizosphere microorganisms. This makes it challenging to collect sufficient exudates to monitor and track the atoms and compounds contained in root secretions. Several creative methods have been developed to improve our ability to collect root exudates, quantify their chemistry and evaluate their effects on plants and the root-associated microbes, as described in the next sections.

Root exudate characterization

Characterizing root exudation by plants is crucial to understanding their intrinsic composition during the plant's lifespan, as influenced by plant physiology. Furthermore, we lack information about how abiotic and biotic factors affect the root exudation rate and concentration of the various compounds of interest within this complex mixture. Root exudates are generally collected from hydroponic systems, where plants are grown in a nutrient solution with minimal physical support (Williams et al., 2021). At the time of sampling, the entire root system is removed from the hydroponic system and soaked in fresh nutrient solution for a period of time to solubilize root exudates for analysis. However, many factors influence root exudation in hydroponic solutions, including the hydroponic design (reservoir vs. ebb and flow system, free root vs. inert substrate) and the aeration, pH and sterility of the nutrient solution (Oburger and Jones, 2018). In addition, the spatial scale of analysis needs to be considered, to understand whether root exudates will be collected from the entire root system or specific root regions (e.g., with split root systems or exudation traps; Oburger and Jones, 2018). Another challenge with hydroponic methods is that a solution-based system is unlike a natural soil environment and lacks the physicochemical and biological stimuli that prompt plant roots to release exudates. However, collecting root exudates from soil is complicated by the fact that most root exudates are transient and rapidly metabolized by rhizosphere microorganisms.

Researchers have the option of growing plants in natural soil environments before transferring them into a hydroponic system. There are several ways to obtain root exudates in this process. One of the most common is to carefully remove intact plants from a soil container, rinse the root system and transfer the roots into a nutrient solution, assuming that any root exudates that are collected in the solution represent the plant response in soil (Williams et al., 2021). This technique requires substantial disturbance of the root system during root washing. An alternative is to use a modified soil container known as a rhizobox, which can physically segregate part of the root system for precise sampling of root exudates (Pantigoso et al., 2021). We still need innovative strategies that allow plants to grow in realistic soil environments that give us access to the root system for the simple collection of root exudates.

Once root exudates are collected, the solution is prepared for analysis. Normally, filtration or centrifugation are required to remove root materials, microbial cells and other biological debris that would interfere with root exudate analysis (Oburger and Jones, 2018). This treatment purifies the extract, but the resulting solution generally contains a small concentration of exudates. The next step involves freeze-drying or rotary evaporation to concentrate the inorganic and organic solutes. These preparatory steps are needed to preserve the sample quality and ensure sufficient analyte for chemical assessment.

Characterization of the root exudates relies on advanced instrumentation, including high pressure liquid chromatography, ion chromatography, gas and liquid chromatography combined with tandem mass spectrometry, and nuclear magnetic resonance (Oburger and Jones, 2018; Sun et al., 2021). It seems impossible to get a complete analysis of all root exudates with a single instrument. Ion chromatography is suitable to characterize the polar root exudates (e.g., inorganic ions, amino acids, proteins), while gas chromatography is appropriate for the analysis of volatile organic exudates, gaseous molecules and primary metabolites. The chemical composition of dissolved secondary metabolites can be quantified with liquid chromatography or nuclear magnetic resonance spectroscopy. Thus, a sequential approach using several analytical methods is necessary to fully characterize root exudates.

Analysis of root exudate effects

Innovative techniques have been developed to assess the response of plants and the soil microbiome to root exudates. Applying synthetic exudates to bare soil and measuring the soil microbial response is one way to accomplish this objective (Sun et al., 2021). Although this does not reflect the actual exudates released from roots and it ignores the influence of the physical root structure on soil microorganisms, this approach can isolate the effects of specific exudates or cocktail of exudates on soil microbial processes in a realistic environment. An incisive technique to evaluate the effects of root exudates involves knocking out genes that are known to be important for root exudate production. Gene knockout procedures are often used to assess the direct effect of root exudates on soil pathogens, as well as the ability of these root exudates to illicit biocontrol responses by antagonistic organisms (Vives-Peris et al., 2020; Sun et al., 2021). However, these methods are often tested *in vitro* without soil, which obscures our understanding of root exudate processes occurring in natural soil environments. Most studies that use the gene knockout approach involve *A. thaliana*, a model organism that is simple to grow and has a fully sequenced genome. Relying on one plant species will naturally limit our understanding of root exudation, and this should be considered when interpreting root exudate effects on the rhizosphere.

During the past decades, isotope technology has revolutionized our knowledge of root exudation and its effects, especially regarding soil carbon and nitrogen dynamics. Typically, plants grown in soil are exposed to ^{13}C — CO_2 , which is fixed through photosynthesis and used by the plant to produce ^{13}C -root exudates. Root exudates can also be labeled with ^{15}N by exposing the plant to ^{15}N -nitrogen compounds via plant stem infiltration, leaf tip feeding or foliar application. Isotope tracing makes it possible to follow the movement of labeled compounds from shoots into the root tissues, and then track the quantities of labeled root exudates that are released, absorbed and metabolized by soil microorganisms before any tracer remaining in the lysed cellular debris enters the soil organo-mineral pool (Pett-Ridge and Firestone, 2017). We can also quantify the proportion of root exudates that are leached from the soil-plant system as ^{13}C -dissolved organic carbon and ^{15}N -dissolved organic nitrogen or lost as gaseous products like $^{13}\text{CO}_2$, $^{13}\text{CH}_4$ and $^{15}\text{N}_2\text{O}$. Such work improves our understanding of the transport and fate of root exudates in realistic soil environments, as well as the direct contribution of root exudation to greenhouse gas emissions. If the soil microbial biomass is sufficiently enriched with ^{13}C or ^{15}N , stable isotope probing is possible (Pett-Ridge and Firestone, 2017). This method would involve sequencing labeled nucleic acids or proteins to determine which microorganisms were preferentially metabolizing the root exudates and their subsequent activity. Inevitably, the carbon and nitrogen isotopes in soil pools may also come from other sources besides root exudates, such as from mucilage and dead root tissue. Nevertheless, isotope technology is an effective tool to resolve the fate of root-derived carbon and nitrogen, including root exudates.

Isotope tracing methods are limited by the destructive soil sampling at some point during the plant growth. This means that each sample is a ‘snap-shot’ of the root exudation and microbial responses at a particular point in time. To avoid the issues associated with repeated discrete sampling, researchers can label plants with radioisotopes (i.e., ^{14}C) instead, and use imaging technology like autoradiography to analyze the fate of root exudates (Oburger and Jones, 2018). This method enables continuous monitoring and assessment of the spatiotemporal distribution of root exudation in the short-term. However, as ^{14}C will be present in both root exudates and the root itself, it can sometimes be difficult to delineate the fate of root exudates in the soil environment from the physical root system of the plant. As such, the resolution of such images can pose a challenge in their interpretation. Radioisotope imaging also causes some soil disturbance when the intact root-soil system is placed in the imaging device. This experimental bias can obscure the findings, which is why radiography images of soil-root systems must be interpreted with caution.

Zymography is another imaging method used to measure extracellular enzyme activities in the rhizosphere. This is done by placing a substrate-loaded agar(ose) gel in the root-soil system. When roots come into contact with the gel, the extracellular enzymes released from the roots will hydrolyze the substrate and this appears as a color change on the gel (Spohn et al., 2013; Sanaullah et al., 2016). Similarly, specifically treated membranes or agar(ose) gels loaded with dyes can be placed in the rhizosphere to detect changes in pH and ion concentrations caused by root exudation. Fluorescence imaging is another way to visualize root exudation and the impact of specific root exudates on microorganisms that emit fluorescent light in the rhizosphere (Sun et al., 2021). However, these imaging methods measure rhizosphere effects rather than that of specific root exudates. Additionally, such methods are often performed at discrete points in time, though recent developments have substantially improved their temporal resolution to near-continuous measurements, like time-lapse zymography (Guber et al., 2021). Thus, imaging techniques are best used in conjunction with other analytical approaches to properly interpret the effects of root exudates on rhizosphere processes.

Summary and outlook

Root exudation is a fundamental controller of plant-microbial functions in the rhizosphere. From their origin as photosynthates in leaves that are transported and transformed to bioactive substances in the roots, exudates are integral to plant growth and development. These exudates transform the oligotrophic soil around the root into a carbon- and nutrient-rich microenvironment with intense biological activity. If roots engineer the rhizosphere, then root exudates are the building blocks supporting a diverse community of free-living and symbiotic microorganisms that are directly responsible for maintaining plant productivity and health.

Root exudates are a broad class of substances that serve multiple functions for plants once they are secreted into the soil environment. However, future discoveries about root exudates will require us to improve the methodologies for their collection from realistic soil environments, along with continuous sampling protocols and integrated analytical platforms that can quantify the range of inorganic and organic substances that roots exude into the rhizosphere. Information on the activity of root transporter systems could be a helpful indicator of secondary metabolite and extracellular enzyme movement from the root to the surrounding environment.

Improved methodologies are essential for characterizing the breadth of root exudates that exist, but innovative strategies are also required to grasp the scale of root exudate effects on the rhizosphere. This includes microRNA (miRNA), which has garnered immense interest during the past years with recent discoveries on their potential ability to silence microbial genes, but whose full impact on microbe-plant interactions remains to be assessed. Understanding the response of microbes to exuded miRNA with incisive studies is key to unlocking the extent of their function in regulating microbial activity in the rhizosphere. We also should be able to document the response of rhizosphere microorganisms to root exudation as a whole with advanced techniques like RNA and protein stable isotope probing. These promising methods are still in their infancy for understanding nutrient transfers and other processes in the plant-microbial system, and merit further use.

Better understanding of root exudation in agricultural crops would not only reveal much about plant tolerance to abiotic stresses, their ability to acquire essential nutrients and capacity for self-defense against growth-limiting biota, but could also lead to microdosing of pesticides and fertilizer rather than uniform field applications. Additionally, further research into the root exudate profiles that regulate microbial metabolism could identify root exudates that either trigger (e.g., labile carbon substrates) or impede (e.g., biological nitrification inhibitors) greenhouse gas production. In the face of climate change, these findings could help us to one day engineer crops that sequester fixed carbon in and mitigate nitrous oxide emissions from cultivated soils. Such developments are consistent with the goals of precision agriculture and align well with global initiatives aiming to achieve food security, protect the agroenvironment and stabilize rural income in sustainable agroecosystems, now and in the future.

Acknowledgements

This work was supported by the Agricultural Greenhouse Gases Program (AGGP) of Agriculture and Agri-Food Canada (AAFC) [Project number: AGGP—020] and the Fonds de recherche du Québec—Nature et technologies (FRQNT) [Award number: B2X-321305]. We thank Fang Ding, Lijun Hou and Yutong Jiang for their support in the preparation of this chapter.

References

- Biala W and Jasirski M (2018) The phenylpropanoid case—It is transport that matters. *Frontiers in Plant Science* 9: 1610. <https://doi.org/10.3389/fpls.2018.01610>.
- Canarini A, Merchant A, and Dijkstra FA (2016) Drought effects on *Helianthus annuus* and *Glycine max* metabolites: From phloem to root exudates. *Rhizosphere* 2: 85–97. <https://doi.org/10.1016/j.rhisph.2016.06.003>.
- Canarini A, Kaiser C, Merchant A, Richter A, and Wanek W (2019) Root exudation of primary metabolites: Mechanisms and their role in plant responses to environmental stimuli. *Frontiers in Plant Science* 10: 157. <https://doi.org/10.3389/fpls.2019.00157>.
- Carvalhais LC, Dennis PG, Fedoseyenko D, et al. (2011) Root exudation of sugars, amino acids, and organic acids by maize as affected by nitrogen, phosphorus, potassium, and iron deficiency. *Journal of Plant Nutrition and Soil Science* 174: 3–11. <https://doi.org/10.1002/jpln.201000085>.
- de Bruijn WJC, Gruppen H, and Vincken JP (2018) Structure and biosynthesis of benzoxazinoids: Plant defence metabolites with potential as antimicrobial scaffolds. *Phytochemistry* 155: 233–243. <https://doi.org/10.1016/j.phytochem.2018.07.005>.
- Gargallo-Garriga A, Preece C, Sardans J, et al. (2018) Root exudate metabolomes change under drought and show limited capacity for recovery. *Scientific Reports* 8: 12696. <https://doi.org/10.1038/s41598-018-30150-0>.
- Guber A, Blagodatskaya E, Juyal A, Razavi BS, Kuzyakov Y, and Kravchenko A (2021) Time-lapse approach to correct deficiencies of 2D soil zymography. *Soil Biology and Biochemistry* 157: 108225. <https://doi.org/10.1016/j.soilbio.2021.108225>.
- Guyonnet JP, Cantarel AAM, Simon L, Haichar F, and e. Z. (2018) Root exudation rate as functional trait involved in plant nutrient-use strategy classification. *Ecology and Evolution* 8: 8573–8581. <https://doi.org/10.1002/ece3.4383>.
- Haichar F, Santaella C, Heulin T, and Achouak W (2014) Root exudates mediated interactions belowground. *Soil Biology and Biochemistry* 77: 69–80. <https://doi.org/10.1016/j.soilbio.2014.06.017>.
- Massalha H, Korenblum E, Tholl D, and Aharoni A (2017) Small molecules below-ground: The role of specialized metabolites in the rhizosphere. *The Plant Journal* 90: 788–807. <https://doi.org/10.1111/tpj.13543>.
- Nguyen C (2009) Rhizodeposition of organic C by Plant: Mechanisms and Controls. In: Lichtfouse E, Navarette M, Debaeke P, Véronique S, and Alberola C (eds.) *Sustainable Agriculture*, 1st ed, pp. 97–123. Springer. https://doi.org/10.1007/978-90-481-2666-8_9.
- Oburger E and Jones DL (2018) Sampling root exudates—Mission impossible? *Rhizosphere* 6: 116–133. <https://doi.org/10.1016/j.rhisph.2018.06.004>.
- Pantigoso HA, He Y, DiLegge MJ, and Vivanco JM (2021) Methods for root exudate collection and analysis. In: Carvalhais LC and Dennis PG (eds.) *The Plant Microbiome*. 1st edn, *Methods in Molecular Biology*, 1st edn, 2232nd volume, pp. 291–303. Springer. https://doi.org/10.1007/978-1-0716-1040-4_22.
- Pausch J and Kuzyakov Y (2018) Carbon input by roots into the soil: Quantification of rhizodeposition from root to ecosystem scale. *Global Change Biology* 24: 1–12. <https://doi.org/10.1111/gcb.13850>.
- Pett-Ridge J and Firestone MK (2017) Using stable isotopes to explore root-microbe-mineral interactions in soil. *Rhizosphere* 3: 244–253. <https://doi.org/10.1016/j.rhisph.2017.04.016>.
- Preece C and Peñuelas J (2020) A return to the wild: Root exudates and food security. *Trends in Plant Science* 25: 14–21. <https://doi.org/10.1016/j.tplants.2019.09.010>.
- Ross-Elliott TJ, Jensen KH, Hanning KH, et al. (2017) Phloem unloading in *Arabidopsis* roots is convective and regulated by the phloem-pole pericycle. *eLife* 6: e24125. <https://doi.org/10.7554/eLife.24125>.
- Sanaullah M, Razavi BS, Blagodatskaya E, and Kuzyakov Y (2016) Spatial distribution and catalytic mechanisms of β -glucosidase activity at the root-soil interface. *Biology and Fertility of Soils* 52: 505–514. <https://doi.org/10.1007/s00374-016-1094-8>.
- Sasse J, Martinola E, and Northen T (2018) Feed your friends: Do plant exudates shape the root microbiome? *Trends in Plant Science* 23: 25–41. <https://doi.org/10.1016/j.tplants.2017.09.003>.
- Spohn M, Carminati A, and Kuzyakov Y (2013) Soil zymography—A novel *in situ* method for mapping distribution of enzyme activity in soil. *Soil Biology and Biochemistry* 58: 275–280. <https://doi.org/10.1016/j.soilbio.2012.12.004>.
- Subbarao GV, Kishii M, Bozal-Leorri A, et al. (2021) Enlisting wild grass genes to combat nitrification in wheat farming: A nature-based solution. *Proceedings of the National Academy of Sciences* 118: e2106595118. <https://doi.org/10.1073/pnas.2106595118>.

- Sun H, Jiang S, Jiang C, Wu C, Gao M, and Wang Q (2021) A review of root exudates and rhizosphere microbiome for crop production. *Environmental Science and Pollution Research* 28: 54497–54510. <https://doi.org/10.1007/s11356-021-15838-7>.
- Thompson GA and Wang HL (2017) Phloem. In: Thomas B, Murphy DJ, and Murray BG (eds.) *Encyclopedia of Applied Plant Sciences*. 2nd edn, 1st volume, pp. 110–118. Elsevier. <https://doi.org/10.1016/B978-0-12-394807-6.00071-X>.
- Vives-Peris V, de Ollas C, Gómez-Cadenas A, and Pérez-Clemente RM (2020) Root exudates: From plant to rhizosphere and beyond. *Plant Cell Reports* 39: 3–17. <https://doi.org/10.1007/s00299-019-02447-5>.
- Vos C, Claerhout S, Mkandawire R, et al. (2012) Arbuscular mycorrhizal fungi reduce root-knot nematode penetration through altered root exudation of their host. *Plant and Soil* 354: 335–345. <https://doi.org/10.1007/s11104-011-1070-x>.
- Warren CR (2015) Wheat roots efflux a diverse array of organic N compounds and are highly proficient at their recapture. *Plant and Soil* 397: 147–162. <https://doi.org/10.1007/s11104-015-2612-4>.
- Weston LA, Ryan PR, and Watt M (2012) Mechanisms for cellular transport and release of allelochemicals from plant roots into the rhizosphere. *Journal of Experimental Botany* 63: 3445–3454. <https://doi.org/10.1093/jxb/ers054>.
- Williams A and de Vries FT (2020) Plant root exudation under drought: Implications for ecosystem functioning. *New Phytologist* 225: 1899–1905. <https://doi.org/10.1111/nph.16223>.
- Williams A, Langridge H, Straathof AL, et al. (2021) Comparing root exudate collection techniques: An improved hybrid method. *Soil Biology and Biochemistry* 161: 108391. <https://doi.org/10.1016/j.soilbio.2021.108391>.
- Zhu S, Vivanco JM, and Manter DK (2016) Nitrogen fertilizer rate affects root exudation, the rhizosphere microbiome and nitrogen-use-efficiency of maize. *Applied Soil Ecology* 107: 324–333. <https://doi.org/10.1016/j.apsoil.2016.07.009>.
- Zipfel C and Oldroyd GED (2017) Plant signalling in symbiosis and immunity. *Nature* 543: 328–336. <https://doi.org/10.1038/nature22009>.

Root interactions with the microbiome from the rhizoplane to the bulk soil: An overview

Joann K Whalen^{a,b} and Shamim Gul^{a,c}, ^aDepartment of Natural Resource Sciences, Macdonald Campus of McGill University, Sainte-Anne-de-Bellevue, QC, Canada; ^bChair of Soil Science, Mohammed VI Polytechnic University, Ben Guerir, Morocco; ^cDepartment of Botany, University of Balochistan, Quetta, Pakistan

© 2023 Elsevier Ltd. All rights reserved.

Introduction	357
Chemical signaling between roots and their rhizobiome	358
Root signals that attract compatible microorganisms to the rhizobiome	358
Root endophytes	361
Ecological interactions of roots and their rhizobiome	361
Physical and chemical gradients from the rhizoplane to bulk soil	362
Physical gradients	362
Water acquisition by plant roots and their rhizobiome	363
Chemical gradients	363
Rhizosphere soil pH	363
Redox conditions in the rhizosphere	365
Nutrient and salt concentrations in the rhizosphere	365
Conclusions	367
References	367

Abstract

Roots assemble a diverse community of soil microorganisms in their rhizosphere, the biologically active microenvironment that exists between the root surface and the bulk soil. Roots and their rhizobiome, which include root-associated and endophytic microorganisms, produce and respond to chemical signals. This cross-talk involves the exchange of hundreds of metabolites that alter plant growth, development and defense responses at genetic, cellular, physiological and morphological levels. This allows the root-rhizobiome partners to engineer their shared rhizosphere and optimize its physical, chemical and biological properties for their joint benefit and survival.

Key points

- Plant roots associate with a diverse community of soil bacteria, fungi and archaea to create a rich, biologically-active rhizosphere.
- The rhizobiome consists of microorganisms that live on or around the root surface, as well as endophytes that colonize the root epidermis, cortex and vascular tissues.
- Interactions in the root-rhizobiome result from the exchange of chemical signal molecules, which are produced and responded to by both biological partners.
- The rhizobiome acts as an early warning system of environmental stresses, leading to phenotypic plasticity in plant root and shoot development, or the upregulation of protective functions like induced systemic resistance and antioxidant production. Plants benefit when their rhizobiome senses changes in the soil environment.
- Root-rhizobiome interactions have considerable potential for sustaining food and fiber production systems.

Introduction

Plant roots occupy, and modify, the soil profile of vegetated land. Root-associated soil between the rhizoplane and the bulk soil is referred to as the rhizosphere. Constant interactions between plant roots and soil biota make the rhizosphere a zone of intense biological activity. However, the rhizosphere is also spatially variable. Plant roots extend vertically to a depth of a few to hundreds of centimeters, depending on the depth of the soil profile and the plant species. Thale cress (*Arabidopsis thaliana*), a model organism in plant biology research, has a primary root that reaches a depth of 6.5 cm with about 10 lateral roots extending from the primary root, the longest being nearly 2 cm, after 9 days (Aceves-García et al., 2016). Perennial tree and vine crops like apple (*Malus domestica*), peach (*Prunus persica*), Asian pear (*Pyrus pyrifolia*), grape (*Vitis* spp.) and kiwifruit (*Actinidia deliciosa*) produce woody and fine roots that explore 55–99% of the soil volume to a depth of 1 m, based on cores taken within 2 m of the main woody stem in

established orchards (from 3 to 8 year old; Hughes et al., 1995). It is unlikely that areas in the soil profile are completely unexplored by roots, since the continual turnover of fine roots suggests the absence of roots is only temporary.

Microorganisms that inhabit plant roots and the rhizosphere are known as the rhizobiome. Plant root-microbial interactions have existed for millennia and coincide with the appearance of terrestrial plants on Earth, about 450 million years ago. The common symbiosis pathway, a conserved molecular signaling pathway, is found in all terrestrial plants that interact with bacteria and fungi through rhizobial, actinorhizal and mycorrhizal fungal symbioses. As details of symbiotic plant-microbial interactions are covered in other chapters, they will not be repeated here. Rather, we focus upon free-living soil microorganisms that colonize plant roots and the rhizosphere to share carbon and other resources. The active recruitment of such microorganisms is most evident in young plant seedlings. For example, about 27% of the photosynthates allocated to the root system of spring rye (*Secale cereale*) were respired through root and microbial metabolism during early vegetative growth, and this declined to about 5–8% of photosynthates respired during the stem elongation and ear emergence stages (Remus et al., 2016). The microbial community benefiting from these photosynthates includes around 2500–3000 bacteria, 300–350 fungi and 50 ammonia-oxidizing archaea, based on the number of unique operational taxonomic units in the rhizosphere of *S. cereale* (Li and Wu, 2018).

Plant roots use photosynthates to build a deep, extensive system of structural roots that anchor the plant and help it find water, nutrients and interact with soil microorganisms. Upon reaching a favorable soil microenvironment, epidermal root cells behind the root tip differentiate to produce root hairs that send phytochemicals into soil pores. Dissolved and volatile chemicals attract soil biota towards the root. Communicating through chemical signals, plant root cells and free-living soil microorganisms interact in ways that are beneficial to one, if not both, of the organisms. Cooperative interactions in the rhizosphere give the rhizobiome more capacity than microorganisms in the bulk soil to solubilize inorganic phosphates, dissolve metal oxides, chelate trace metals and mineralize organic matter. The rhizobiome also produces plant growth promoting hormones (phytohormones) that control the extent of root growth. Furthermore, the rhizobiome often acts in a protective way, by directly inhibiting the growth of infectious plant pathogens and parasites, or by up-regulating the induced systemic resistance that defends plants from these and other harmful organisms.

This chapter explains how plant roots attract soil bacteria, fungi and archaea to the rhizosphere. We concentrate on the well-studied rhizobiomes of plants used as food and medicine, as forage for grazing animals, and that provide fibrous materials for textiles, buildings and energy. In this chapter, the rhizobiome is also considered to include microorganisms that penetrate the root epidermis and become resident as root endophytes. This is justified because the biological influence of root endophytes extends into the rhizosphere. Finally, we examine how plant roots and their rhizobiome engineer the soil matrix to create a favorable environment for their biological activities. We discuss the soil physical and chemical gradients that develop from the rhizoplane to the bulk soil through root-microbial interactions, with relevant examples.

Chemical signaling between roots and their rhizobiome

Like all living cells, plant root cells are aware of their surrounding environment. Roots respond in real time to cues from soil microorganisms and to conditions in the soil matrix. When an external signal molecule binds to a compatible receptor in the target root cell, it elicits a cellular response. The chemical signal received by a single root cell will be transmitted throughout the entire plant by cell-to-cell communication. This occurs because every plant cell is connected directly to neighboring cells via the plasmodesmata, a water-filled channel ≤ 50 nm in diameter that transfers intracellular mediator substances (ions and molecules up to ~ 500 Da in size) from one cell to the next. This means that the plant can use its entire cellular network to respond to a chemical stimulus. Consequently, plant growth depends upon the actual environmental conditions, and no two plants have an identical response. This gives rise to the phenotypic variation in plant communities, where genetically similar plants have unique shapes, sizes and productivity because they have adapted to live in a local microenvironment.

Root signals that attract compatible microorganisms to the rhizobiome

One of the ways that plants contend with environmental variation is by interacting with microorganisms. Signaling between plant roots and soil microorganisms was traditionally studied in the context of plant symbioses with rhizobial bacteria and mycorrhizal fungal but is now understood to be a widespread phenomenon involving diverse, non-symbiotic microorganisms that become part of the rhizobiome. Signaling processes are complex and multi-directional (Venturi and Keel, 2016). Three levels of cellular communication occur in the rhizosphere: (1) signaling from plants to microorganisms via small plant-secreted molecules, (2) signaling from microorganisms to plants, based on microbially-produced compounds that affect plant gene expression, root architecture, and plant defense responses, and (3) microbial intraspecies and interspecies signaling, which occurs mainly via quorum-sensing signal molecules that synchronize the growth and activity of the rhizobiome (Fig. 1).

Roots release soluble and volatile chemical molecules that are recognized by soil microorganisms. In the simplest cases, motile microorganisms in the soil pore water receive the chemical signal and respond by swimming towards the root, colonizing the root surface and multiplying in the rhizosphere. Maize (*Zea mays*) releases benzoxazinoids from their roots that causes the soil bacteria *Pseudomonas putida* to transcribe genes for benzoate catabolism and chemotaxis, increasing *P. putida* colonization of roots in wild-type maize compared to a modified maize cultivar that does not produce these compounds (DIMBOA-deficient bx1 mutant; Neal et al., 2012). Sessile microorganisms depend upon root-based signals to locate the closest plant root. When roots exude

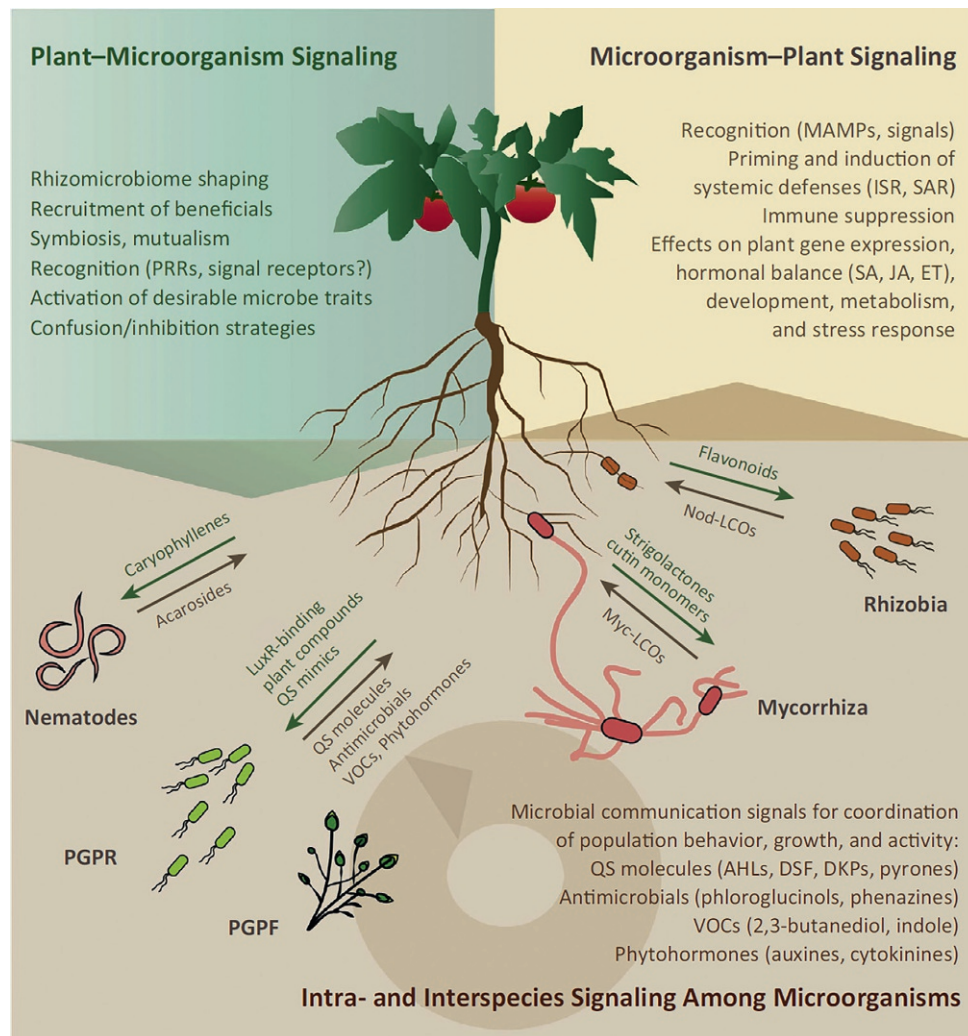


Fig. 1 Known molecules and events involved in intra- and interspecies signaling among microorganisms and interkingdom signaling between microorganism and plants in the rhizosphere. The rhizosphere is populated by a diverse community of microorganisms, including rhizobia (bacteria), mycorrhiza (fungi), plant growth promoting bacteria and fungi (PGPR and PGPF), and nematodes. The interactions of these organisms with the plant are often beneficial. To coordinate their behavior and to control their growth and activity, these microorganisms produce a diversity of signals, among them quorum-sensing (QS) molecules such as N-acyl homoserine lactones (AHLs), diffusible signal factor (DSF), and diketopiperazines (DKPs), antibiotics at subinhibitory concentrations, phytohormone-like molecules, volatile organic compounds (VOCs), and, in the case of nematodes, specific pheromones (acaricides). These communication molecules can also function as interkingdom signals that elicit various effects on plant developmental processes and on local and systemic immune responses (i.e., ISR, induced systemic resistance; and SAR, systemic acquired resistance) and involve interactions with plant hormonal signaling via salicylic acid (SA), jasmonic acid (JA), and ethylene (ET). The plant recruits, shapes, activates and sustains its rhizobiome via the release of root exudates and signaling molecules. To recognize its associated rhizomicrobiome, the plant uses dedicated pattern recognition receptors (PRRs) and presumably other as yet unknown receptors to detect particular microbe-associated molecular patterns (MAMPs) and likely the signals presented by the microorganisms. Recruitment of symbiotic rhizobia and mycorrhiza is initiated by the exchange of specific plant signals (flavonoids and strigolactones, respectively) and microbial signals (Nod factors and Myc, respectively). Signal exchange is also involved in the attraction and colonization of roots with PGPR and PGPF. In addition to hosting beneficial microorganisms as guards, the plant also uses confusion and inhibition strategies (e.g., QS mimicry and quenching) to ward off harmful microorganisms. From Venturi V and Keel C (2016) Signaling in the rhizosphere. *Trends in Plant Science* 21: 187–198. doi: 10.1016/j.tplants.2016.01.005.

strigolactones, a group of phytohormones, the nearby arbuscular mycorrhizal fungi grow longer, branching hyphae before they form the appressorium or hyphopodium, specialized cells used by the fungus to penetrate the plant root. Successful fungal colonization begins with physical contact between the appressorium and the root surface, followed by detection of plant chemicals like hexadecanoic acid in the cutin of root cells, cyclic adenosine monophosphate (cAMP) and ethylene.

The rhizobiome recognizes plant signals, but it also synthesizes compounds that elicit a plant response. Volatile organic compounds released by the rhizobiome include natural products that act as antimicrobials, while promoting plant growth, inducing plant disease resistance and supporting signal transduction pathways that allow plants to quickly respond to natural gradients in light, water, nutrients and other vital resources. The main benefit of volatile organic compounds, relative to water-soluble chemicals, is their enhanced ability to permeate cell membranes and the ease with which they diffuse through the air-filled pores in soil. In raspberry

Table 1 Volatile organic compounds produced by the root-associated bacteria *Bacillus siamensis*, and the antifungal activity of selected compounds against raspberry postharvest diseases caused by the fungal pathogens *Botrytis cinerea* and *Rhizopus stolonifer*.

Compound name	Chemical formula	Inhibition rate in vitro (%)	
		<i>B. cinerea</i>	<i>Ranunculus stolonifer</i>
3-Hydroxy-2-butanone	C ₄ H ₈ O ₂	0	0
5-Aminovaleric acid	C ₅ H ₁₁ NO ₂	0	0
2(Benzyloxy) ethanamine	C ₉ H ₁₃ NO	0	0
Hexanoic acid, 2-methyl-	C ₇ H ₁₄ O ₂	0	0
p-Xylene	C ₈ H ₁₀	ND	ND
2-Heptanone	C ₇ H ₁₄ O	2	<1
Oxime-, Methoxy-phenyl-	C ₈ H ₉ NO ₂	ND	ND
2-Heptanone, 6-methyl-	C ₈ H ₁₆ O	2	<1
2-Heptanone, 5-methyl-	C ₈ H ₁₆ O	2	1
Phenol	C ₆ H ₆ O	ND	ND
Benzene acetaldehyde	C ₈ H ₈ O	25	10
Acetophenone	C ₈ H ₈ O	ND	ND
2-Nonanone	C ₉ H ₁₈ O	2	2
2-Nonanol	C ₉ H ₂₀ O	6	2
2-Decanone	C ₁₀ H ₂₀ O	ND	ND
2-Decanol	C ₁₀ H ₂₂ O	ND	ND
Acetic acid, 1-(S)-phenylethyl ester	C ₁₀ H ₁₂ O ₂	ND	ND
1-Heptanol, 2-propyl-	C ₁₀ H ₂₂ O	11	3
2-Undecanone	C ₁₁ H ₂₂ O	ND	ND
2-Undecanol	C ₁₁ H ₂₄ O	6	5
Phenol, 2,4,6-trichloro-	C ₆ H ₃ Cl ₃ O	ND	ND
2-Dodecanone	C ₁₂ H ₂₄ O	ND	ND
Dodecane, 2,6,10-trimethyl-	C ₁₅ H ₃₂	ND	ND
2-Tridecanone	C ₁₃ H ₂₆ O	2	6
BHT, Butylated Hydroxy Toluene	C ₁₅ H ₂₄ O	100	100
Phenol, 2,4bis(1,1-dimethylethyl)	C ₁₄ H ₂₂ O	100	100
2-Tetradecanone	C ₁₄ H ₂₈ O	10	4
2-Pentadecanone	C ₁₅ H ₃₀ O	10	3
2-Nonadecanone	C ₁₉ H ₃₈ O	10	5
2-Pentadecanone, 6,10,14-trimethyl-	C ₁₈ H ₃₆ O	ND	ND
Hexadecanoic acid, methyl ester	C ₁₇ H ₃₄ O ₂	0	<1
9-Octadecenoic acid (Z), methyl ester	C ₁₉ H ₃₆ O ₂	0	0

The inhibition rate in vitro was between 0% and 100% inhibition, or not determined (ND).

Adapted from Zhang XY, Gao ZF, Zhang XX, Bai WB, Zhang L, Pei HB, and Zhang YJ (2020) Control effects of *Bacillus siamensis* G-3 volatile compounds on raspberry postharvest diseases caused by *Botrytis cinerea* and *Rhizopus stolonifera*. *Biological Control* 141: 104135. doi: 10.1016/j.biocontrol.2019.104135.

(*Rubus idaeus*), there was up to 100% growth inhibition of the fungal pathogens *Botrytis cinerea* and *Rhizopus stolonifer* by two volatile organic compounds produced by the root-associated bacteria *Bacillus siamensis*, namely 2, 6-di-tert-butyl-4-methylphenol (BHT) and 2,4,2,4-di-tert-butylphenol (2,4-DTBP; Zhang et al., 2020; Table 1). Besides inducing plant defenses for better resistance to resist diseases, pests and abiotic stresses, the rhizobiome also produces growth-promoting substances, described in the following sections.

Finally, bacterial members of the rhizobiome exchange quorum-sensing molecules to control their community composition and regulate growth. In general, these molecules help bacteria to synchronize their gene expression in response to cell density, a process known as quorum sensing. One class of quorum-sensing molecules are the N-acyl homoserine lactones, produced by proteobacteria in the rhizosphere. The N-acyl homoserine lactones initiate biofilm formation, causing bacterial cells to aggregate on the root surface. Next, the colony produces extracellular polymeric substances (exopolymers) that protect it from desiccation and other environmental stresses. Production of N-acyl homoserine lactones and other quorum-sensing molecules (e.g., diffusible signaling factors that are cis-2-unsaturated fatty acids, auto-inducing peptides) also influence plant gene expression, induce systemic resistance, and affect plant growth and development. However, plant roots do not passively accept bacterial colonizers. Plants can biosynthesize inhibitory substances that prevent bacteria and fungi from living on their roots. Some of these compounds are phenolics, including benzoates, phenyl propanoids, stilbenes, flavonoids, gallotannins, proanthocyanidins and coumarins, as well as terpenes (monoterpenes, sesquiterpenes, diterpenes and triterpenes), quinones and organosulfur compounds. Secreting these inhibitory substances allows plants to control quorum sensing and biofilm formation on their roots, stems and leaves (Ta and Arnason, 2016). This is an example of cross-talk (exchange of signal molecules) that is characteristic of root-microbial interactions in the rhizosphere.

Root endophytes

Often originating from the soil, root endophytes are microorganisms that become resident in plant root tissues and live there for at least part of their life cycle, without causing disease. Bacteria root endophytes are common in the intercellular space of roots, presumably because there is enough space for their growth plus access to abundant water, carbohydrates, amino acids and inorganic nutrients. In the early stages of colonization, bacteria root endophytes are first observed in root hairs, and subsequently in the root cortex. For example, grape roots inoculated with *Burkholderia* sp. supported the bacterial endophytes in their cortical cells, endodermis and xylem vessels (Compant et al., 2005); colonization was greatest in the primary and secondary roots, and at the base of lateral roots and root tips. Fungal endophytes are routinely isolated from the roots of apparently healthy and vigorous plants. They have regularly septate hyphae with variable texture (translucent, glassy) and color (from lightly to darkly pigmented). The hyphae grow on, in, and in between cells of the epidermis, cortex, and sometimes in the vascular tissue of roots. Virtually every terrestrial plant in the world has fungal root endophytes. In contrast, the archaea root endophytes are an under-investigated group that were detected in the rhizosphere of crops by culture-independent metagenomics technologies. To date, no root-associated archaea endophyte has been isolated and cultured, which is a gap in our scientific knowledge.

Endophytes gain access to plant roots through active and passive mechanisms (Fig. 2). After they attach to the rhizoplane, bacterial endophytes move along the root surface and seek potential entry into the root. Wounds and cracks, as well as openings where root hairs or lateral roots emerge, are the main points where bacterial endophytes enter the root. Furthermore, some bacterial endophytes secrete cell wall cellulolytic enzymes such as cellulases, xylanases, pectinases, and endoglucanases to weaken the root cell wall and enter the root. During the early stages of rice root colonization, the endophytic diazotrophic bacterium *Gluconacetobacter diazotrophicus* expressed genes that deactivated oxidative stress and upregulated the antioxidant response (Alquéres et al., 2013). This suggests a well-developed stealth strategy that allows the bacterial endophyte to penetrate the root without activating a plant defensive response. Once established, bacterial endophytes support plant growth by producing phytohormones, increasing nutrient acquisition and improving plant resilience to abiotic and biotic stresses. On balance, this makes root endophytes beneficial members of the rhizobiome.

Ecological interactions of roots and their rhizobiome

The co-evolution of plants and microorganisms allows for multiple interactions between these biological partners. From an ecological perspective, the interactions can be negative for both organisms (competitive, antagonistic), positive for both organisms (mutualistic, symbiotic), or positive for one and negative for the other (commensal, pathogenic). Interactions change the physiology and morphology of the plant host and its microbial guests, altering the function of one or both partners.

The rhizobiome is populated by microbial cells that reached the root by chance or in response to signal molecules, also referred to as infochemicals, that are released from the plant root and guide microorganisms towards the root surface. Upon colonizing the plant root, the antagonistic microbe may produce specialized molecules that repel other microorganisms from the root. For instance, the root nodule tissue of field-grown alfalfa (*Medicago sativa*) harbors an accessory bacterial community of non-Rhizobiales members from the phyla Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria, which tend to be more abundant on older nodules (Hansen et al., 2020). Some dominant bacteria inhabiting the nodule are cooperative partners (e.g., *Paenibacillus* sp. and *Pseudomonas* sp.). However, the Gram-positive aerobic bacteria *Brevibacillus brevis* produces antibiotics (novel

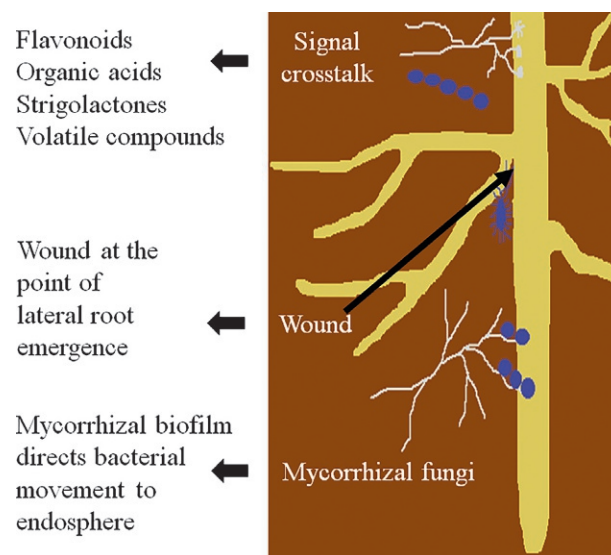


Fig. 2 Passive and active pathways leading to microbial colonization of the endosphere of plant roots.

gramicidin-family metabolites called brilacidin) that inhibit the growth of other bacteria (Hansen et al., 2020). Colonization of root tissues by antibiotic-producing bacteria is significant because these antagonists will control the composition and functions of the rhizobiome. Hansen et al. (2020) interpreted the bacteria-bacteria antagonism to be a strategy that would protect nodules from an infectious bacterial pathogen; thus, antagonism in the rhizobiome could be important to preserve biological N₂ fixation, a critical function in alfalfa.

An example of a mutualistic interaction is when a root-associated microorganism alters the root physiology and morphology for the benefit of the root and the rhizobiome. Mutualistic microorganisms can use hormonal, genetic and regulatory mechanisms to change the root architecture and anatomical phenotype of their plant host. Hormonal effects occur when the microorganism secretes phytohormones, or when it produces secondary metabolites that control hormonal pathways within the root. For instance, rice roots are longer when indole-3-acetic acid production is induced by *Azospirillum brasilense*, *Bacillus altitudinis* and *Klebsiella pneumoniae*, while more lateral roots grow when the cytokinins level is modulated by *A. brasilense*, *Phomopsis liquidambari* and *Rhizophagus irregularis* (Verma et al., 2022). Greater root mass, root thickness, length, and more lateral roots are produced when rice is colonized with *Rhizophagus irregularis* because the fungal partner modulates the expression of OsCYCLOPS1 and OsLRT1 genes, which regulate several phytohormones (auxin, cytokinins, abscisic acid and ethylene). These observations should encourage further study into the mutualistic interactions of roots and their rhizobiome.

Finally, plant roots are negatively impacted by pathogens, disease-causing organisms that are fungi, bacteria, protists, nematodes or viruses. Proliferation of root pathogens causes a range of symptoms, from mild impairment of root functions to plant death, depending on the severity of the infection. Plant constitutive and induced defense systems can partially control root pathogens, but the rhizobiome can also provide protection. Some antibiotics produced by the rhizobiome are effective against pathogens because they disrupt the cell wall and cell membrane structures, while inhibiting ribosomal biogenesis during pathogen cell division. Rhizobacteria like *Bacillus* spp. secrete extracellular enzymes such as chitinases, β -1,4-glucanases, proteases, cellulases and xylanases that inhibit the mycelial growth of fungal pathogens like *Botrytis cinerea*, *Fusarium oxysporum*, *F. solani* and *Rhizoctonia solani* (Dutta et al., 2022). By occupying the root surface and competing for water, Fe³⁺ and other nutrients, the rhizobiome may be able to indirectly displace root pathogens. Rhizobacteria also induce systemic resistance in plants through the jasmonate and ethylene signaling pathways. Furthermore, various rhizobacteria are known to increase the mRNAs encoding for phenylalanine ammonia lyase (PAL), the first key enzyme in the phenyl propanoid pathway that biosynthesizes phenolics and phytoalexins, as well as upregulating salicylic acid production, another pathway for induced systemic resistance in plants (Dutta et al., 2022). These examples illustrate the potential of root-rhizobiome interactions to reduce the severity of plant diseases caused by pathogens.

Physical and chemical gradients from the rhizoplane to bulk soil

The rhizosphere, starting at the rhizoplane and ending at the bulk soil, has unique physical and chemical properties. These characteristics are the result of the growth and biological activity of the roots and their rhizobiome. However, the rhizosphere is spatially variable, depending upon the plant species and its root architecture. Furthermore, the magnitude of root-rhizobiome activity is determined by the plant growth stage, soil edaphic factors and environmental stresses. The next sections explain how root-rhizobiome interactions create a distinctive rhizosphere that has unique physical and chemical properties from the rhizoplane and the bulk soil.

Physical gradients

Roots extend vertically and expand horizontally through pores in the soil matrix, but the force of a growing root is sufficient to alter the physical environment, making it better suited for anchoring the plant and acquiring resources. Roots seek a path of least resistance, which encourages growth in the existing cracks and fissures, although bypassing rocks and compacted layers. However, roots can also create space by deforming the friable soil mass. This occurs because roots exert pressure that displaces the soil particles when they enter soil pores that are smaller than the root diameter. The maximum axial root growth pressure ($\sigma_{\max}$) is 200–1300 kPa for various species and genotypes of annual crops (Hosseini et al., 2019). We expect larger primary roots and lateral roots with a smaller diameter to exert a similar amount of axial root growth pressure because the force needed to push aside soil particles is proportional to the root diameter. Root hairs are unlikely to modify soil physical properties due to their small size (about 15 μ m diameter and $\leq$ 1 mm long). Root elongation rates depend on the cell turgor, cell wall extensibility and cell wall pressure of the roots, each of which responds to environmental conditions.

Soil conditions are not always favorable to root growth. Compacted soils and degraded soils often have high mechanical impedance to root penetration. Plants respond by growing shorter, thicker roots. This occurs because ethylene, a plant hormone and gas released from root tissue, accumulates around roots in a compacted soil. Rice (*Oryza sativa* L.) roots exposed to a high ethylene concentration will produce other plant hormones, principally auxin, which reduces root elongation, and abscisic acid that promotes radial cell expansion (Huang et al., 2022). Drought is another barrier to root growth since it creates osmotic stress, which affects root cell turgor, and increases the soil mechanical strength because water is held more tightly to the soil matrix. One of the ways that roots adapt to penetrate dry, hard soils is through their rhizobiome. For example, when maize was colonized by the fungal root endophyte *Piriformospora indica* (basidiomycete of the Sebacinaceae family), it could produce at least double the maximum axial root growth pressure (from 505 to 780 kPa) than without the fungal partner (186–344 kPa; Hosseini et al., 2019).

Water acquisition by plant roots and their rhizobiome

Plants get most of their water from soil, through their roots. Due to their diverse morphology, roots vary in their ability to acquire water from soil. In general, younger, finer roots have greater water uptake rates per surface area than mature roots. Still, it is challenging to acquire plant-available water from the soil matrix. Sometimes the soil becomes so dry that it reaches the permanent wilting point for plants. Once a plant wilts and cannot recover its turgidity after 12 h in a fully watered soil, it is considered to have permanently wilted. Normally, the soil still contains water at the permanent wilting point, but the water is adhering too tightly to soil particles to be extracted by the plant, or the water is within micropores ($<10\ \mu\text{m}$ in diameter) that are not accessible to plant roots (i.e., the smallest root hairs are about $15\ \mu\text{m}$ in diameter). Thus, roots cannot always obtain enough soil water to fulfill plant metabolic needs.

Plants rely on the rhizobiome to overcome soil water deficits. Fungal root endophytes are particularly interesting in this regard. After the fungus colonizes the plant root, it grows long branching, filamentous hyphae (from 3 to $30\ \mu\text{m}$ in diameter) that extend into the soil pores. The smallest hyphae can enter micropores of about $2\ \mu\text{m}$ diameter that are inaccessible to plant roots and root hairs (Kakouridis et al., 2022). Water and solutes in the micropore then move along the surface of the fungal hyphae through capillary action, towards the plant root. Fungal hyphae often connect multiple plants. For example, hydraulic redistribution of water from pine trees (*Pinus ponderosa*) to their seedlings was facilitated by the hyphae of ectomycorrhizas (mostly *Rhizopogon* sp.), which transferred about 15% of the water required for seedling growth (Warren et al., 2008). Another way that fungal hyphae change soil hydraulic properties in the rhizosphere is by promoting macroaggregate formation. Structurally, a macroaggregate is an agglomeration of microaggregates ($<0.25\ \text{mm}$ in diameter containing clay, metal oxides and bacterial debris) that are held together with sand, silt and organic matter by fungal hyphae. A rhizosphere with more macroaggregates also has more macropores for root growth and that conduct water towards the roots. When switchgrass (*Panicum virgatum*) was inoculated with mycorrhizal fungal (*Piriformospora indica*), the greater porosity allows roots and fungal hyphae to partially fill macropores, decreasing the saturated hydraulic conductivity and allowing them to access more water in saturated and unsaturated soils (Marcacci et al., 2022). These examples of fungal-mediated physical changes in the rhizosphere soil have a direct impact on water acquisition by plants.

There are two additional ways that fungal partners increase the water supply to plants. Fungal hyphae exude hundreds of metabolites, including carbohydrates and polyols (9% of the total number of metabolites identified), amino acids and amines (43%), nucleic acids (10%), organic acids (21%), fatty acids (1%), homologs (5%), phosphate esters (9%), phytohormones (1%), and others (1%) (Luthfiana et al., 2021). Some of these substances will attract water, similar to plant mucilage, and may help to retain water in the rhizosphere during dry periods. Finally, fungal hyphae can directly transfer water through cell-to-cell interactions, from soil to plants. Slender wild oat (*Avena barbata*) colonized by *Rhizophagus intraradices* relied on the fungal hyphae to bridge air gaps, penetrate small pores and access water that was inaccessible to roots. Nearly 35% of transpired water was directly transferred from the fungal partner to the plant through cellular transport and via an extracytoplasmic pathway, along the external surface of the fungal hyphae (Kakouridis et al., 2022). This further confirms the importance of root-fungal interactions to maintain plant productivity in dry environments.

Chemical gradients

Rhizosphere soil pH

The rhizosphere is generally more acidic than the bulk soil (Fig. 3). When roots absorb nutrient ions from the soil water, they release a counter-ions into the environment. Uptake of positively charged ions such as NH_4^+ , K^+ , Ca^{2+} , Mg^{2+} and micronutrients like Fe^{2+} , Mn^{2+} , Cu^{2+} and Zn^{2+} is accompanied by the release of a proportionate amount of H^+ into the soil solution. Root hairs have epidermal cell membranes containing proton pumps (H^+ ATPases) that use ATP as an energy source to pump H^+ out of the cell and into the soil, against their electrochemical gradient. These proton pumps create a strong electrochemical gradient with a high concentration of protons and a strong positive charge outside of the cell, and a low concentration of protons and relatively negative charge inside of the cell. The electrochemical gradient causes cations to diffuse to the root hairs, and they may enter the root through passive diffusion (e.g., through porins) or active transport via carrier proteins (i.e., against a concentration gradient). Electrochemical and concentration gradients are also important for plant uptake of negatively charged ions like NO_3^- , H_2PO_4^- , SO_4^{2-} , Cl^- and MoO_4^{2-} . Plants also release anions like OH^- and HCO_3^- to maintain the electrical neutrality of the anion uptake process. Consequently, the area closest to the root surface has a distinct pH from the rest of the rhizosphere, and the bulk soil (Taylor and Bloom, 1998).

Roots and their rhizobiome also secrete organic acids that acidify the rhizosphere. Organic acids contain one or more carboxylic groups. Root exudation of organic acids can acidify the rhizosphere from near neutrality (around pH 7) to pH 4–5, depending upon the type and quantity of organic acids secreted into the rhizosphere. Roots produce acetic, citrate, malate, oxalate, shikimate and malonate acids, among others, with concentrations of 5–50 mM in the root cell vacuoles being as much as 10-fold greater than organic acids concentration in the cytosol, depending on the plant species and growing conditions (Sandnes et al., 2005). Monocarboxylic acids are the most common organic acid produced by roots, followed by the di- and tri-carboxylic acids. Furthermore, organic acids are released by the rhizobiome and from the decomposition of organic matter and plant residues in soil. However, roots and the rhizobiome often secrete $>1\ \text{mM}$ of organic acids in the rhizosphere, which is about two orders of magnitude greater than the bulk soil with $<0.01\ \text{mM}$ of soluble organic acids (Sandnes et al., 2005). These levels of organic acids in the rhizosphere are of biological importance to roots and their rhizobiome.

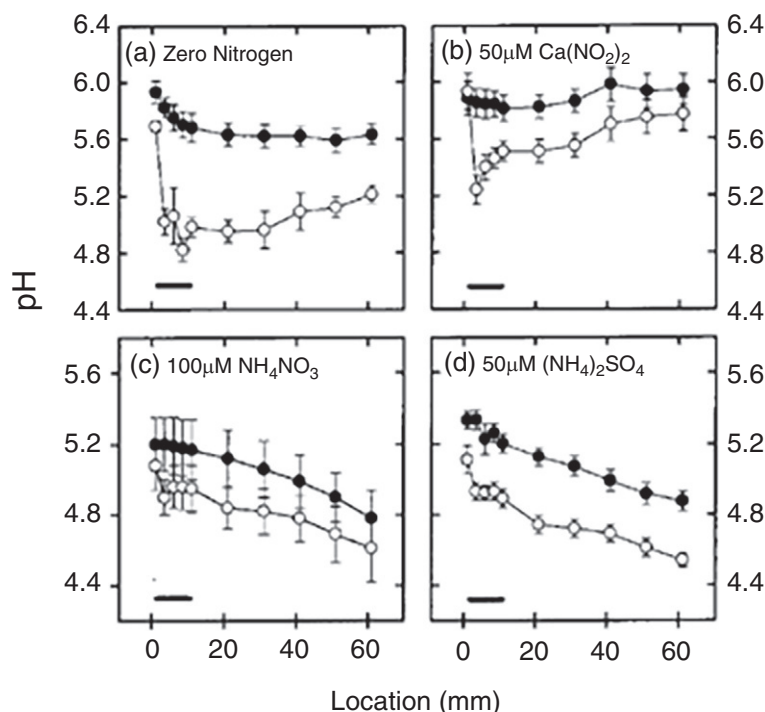
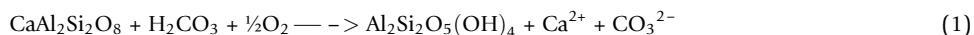


Fig. 3 Root surface pH profiles along the primary root axis of 4-d-old maize seedlings. The length of the primary root is indicated by the black bar, above the x-axis. pH readings were taken at multiple distances (in mm) at the root surface (open symbols) and in the bulk solution (closed symbols). Maize seedlings were grown in hydroponic solutions with zero nitrogen, 50 μM $\text{Ca}(\text{NO}_3)_2$, 50 μM $(\text{NH}_4)_2\text{SO}_4$ and 100 μM NH_4NO_3 . The pH is always lower at the root surface than in the bulk solution, confirming that the rhizosphere tends to be acidic, relative to the surrounding soil. From Taylor AR and Bloom AJ (1998) Ammonium, nitrate, and proton fluxes along the maize root. *Plant, Cell & Environment* 21: 1255–1263. doi: 10.1046/j.1365-3040.1998.00357.x.

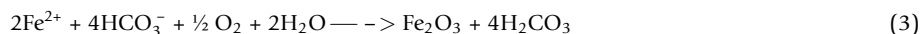
Organic acids in the rhizosphere are involved in soil weathering, which chemically transforms rocks into mineral forms. For example, plagioclase feldspar ($\text{CaAl}_2\text{Si}_2\text{O}_8$) is hydrolyzed to kaolinite clay ($\text{Al}_2\text{Si}_2\text{O}_5(\text{OH})_4$) in the presence of carbonic acid (H_2CO_3) in an acidic environment.



Acidic conditions also favor oxidation reactions. The dissolution of olivine (a ferromagnesian silicate, Fe_2SiO_4) in the presence of carbonic acid produces dissolved iron, carbonate, and silicic acid (H_4SiO_4).



When oxygen is present, the dissolved iron is quickly converted to hematite, Fe_2O_3 :



These reactions release soluble Ca^{2+} and Fe^{2+} , which are essential nutrient ions for biological activity. Such reactions, facilitated by organic acids, assure the root-rhizobiome of a steady supply of ions that are released from minerals, including K^+ , Mg^{2+} , Mn^{2+} , Cu^{2+} and Zn^{2+} .

Organic acids are also involved in the dissolution of phosphates that precipitated with calcium, iron and aluminum in soil. Calcium phosphates such as dicalcium phosphate ($\text{CaHPO}_4 \bullet 2\text{H}_2\text{O}$), octacalcium phosphate ($\text{Ca}_8\text{H}_2(\text{PO}_4)_6 \bullet 5\text{H}_2\text{O}$) and hydroxyapatite ($\text{Ca}_{10}(\text{PO}_4)_6(\text{OH})_2$) are the most common of the stable, insoluble crystalline phosphate forms in alkaline soil. Although calcium phosphates are sparingly soluble in water, all dissolve in acids, including organic acids. In acidic soils, phosphate ions precipitates with iron, often strengite ($\text{FePO}_4 \bullet 2\text{H}_2\text{O}$) and aluminum as wavellite [$\text{Al}_3(\text{PO}_4)(\text{OH})_3 \bullet 5\text{H}_2\text{O}$] and variscite ($\text{AlPO}_4 \bullet 2\text{H}_2\text{O}$). Furthermore, phosphate becomes chemisorbed to the hydroxides and oxides of iron and aluminum in soil. Organic acids chelate with iron and aluminum (hydr)oxides, which either replace phosphates in sorption sites or prevents the ions from binding with the iron and aluminum.

Roots control the concentration and composition of organic acids secreted into the rhizosphere based on their genetics and environmental factors, like soil acidity, the presence of insoluble minerals, phosphorus deficiency and metal toxicity. More phosphate is solubilized with the root-produced tricarboxylic acids (citric and tartaric acids) than dicarboxylic acids (maleic and fumaric acids), followed by monocarboxylic acids (oxalic and acetic acids; Sindhu et al., 2022). Rhizobiome strains of *Bacillus* secrete a mixture of lactic, isovaleric, isobutyric and acetic acids that contribute to mineral weathering. Phosphate-solubilizing bacteria in the genera *Burkholderia*, *Cronobacter*, *Curtobacterium*, *Enterobacter*, *Massilia*, *Pantoea*, *Pseudomonas* and *Ralstonia* release organic acids like glycolic, oxalic, malonic and succinic acids. Rhizosphere-compatible fungi such as *Alternaria*, *Aspergillus*, *Fusarium*,

Table 2 Chemical forms of organic acids in the rhizosphere that are biosynthesized by plant roots or by soil microorganisms.

Aliphatic carboxylic acids			Aromatic carboxylic acids	Miscellaneous carboxylic acids	Substituted acids
Monocarboxylic	Dicarboxylic	Tricarboxylic	p-Coumaric acid	Phthalic acid	1-aminocyclopropane-1-carboxylic acid
Acetic acid	Fumarate	Aconitic acid	Ferulic acid	Shikimic acid	2,4-dichlorophenoxyacetic acid
Butyric acid	Maleic acid	Citric acid	Gallic acid		2,4,5-trichlorophenoxy acetic acid
Formic acid	Malic acid		p-hydroxybenzoic acid		Absciscic acid
Lactic acid	Malonate		Protocatechuic acid		Gibberellic acid
Propionic acid	Oxalate		Salicylic acid		Indole-3-acetic acid
Pyruvic acid	Succinate		Sinapic acid		
Valeric acid	Tartarate		Syringic acid		
			Vanillic acid		

Adapted from Sindhu SS, Sehrawat A, and Glick BR (2022) The involvement of organic acids in soil fertility, plant health and environment sustainability. *Archives of Microbiology* 204: 720. doi: 10.1007/s00203-022-03321-x.

Helminthosporium and *Penicillium* also produce oxalic, maleic, succinic, citric, and fumaric acids that release the precipitated and sorbed phosphates from the mineral phase into soil pore water (Table 2). The cooperative phosphorus-solubilizing activity of roots and their rhizobiome improves their uptake of phosphates, which are essential for the growth and survival of all living organisms.

Redox conditions in the rhizosphere

The activity of roots and their associated rhizobiome is affected by the availability of molecular oxygen (O_2). The O_2 concentration from the root surface to the bulk soil depends upon the balance between O_2 consumption from biological respiration, and the amount of O_2 that diffuses into the rhizosphere from the atmosphere and O_2 -rich soil pores. Biological respiration in the rhizosphere usually relies on aerobic metabolism of carbon substrates with O_2 as the dominant electron acceptor in the roots and rhizobiome. If the rhizosphere has a low O_2 concentration because the pores are filled with water (i.e., O_2 diffuses more slowly through water-filled pores than through air-filled pores), the microorganisms might switch from aerobic to anaerobic respiration. In anaerobic respiration, microorganisms will use alternative terminal electron acceptors in the respiratory electron transport chain, in this order: nitrate, NO_3^- ; manganese oxide, MnO_2 ; iron hydroxide, $Fe(OH)_3$; sulfate, SO_4^{2-} ; and carbon dioxide, CO_2 . The likelihood that microorganisms will switch from aerobic to anaerobic reactions is indicated by the redox potential of soil. When the redox potential is >400 mV, O_2 is the main electron acceptor. From 200 to 400 mV, the weakly reducing conditions favor O_2 , NO_3^- and MnO_2 as electron acceptors. However, a redox potential of 200 to -100 mV is moderately reducing ($Fe(OH)_3$ as the electron acceptor) and <-100 mV is strongly reducing (SO_4^{2-} and CO_2 are the electron acceptors). Therefore, the O_2 concentration and the redox potential determines the oxidation-reduction status of the rhizosphere.

Nutrient and salt concentrations in the rhizosphere

Water moves from the bulk soil to the rhizosphere, mainly through mass flow in response to differences in matric potential, carrying with it soluble nutrient ions and low molecular weight compounds. These substances are needed for the proper growth of roots and their rhizobiome. However, the rhizobiome also contributes directly to the nutrient supply for plants. This can be by acquiring nutrients from the atmosphere (e.g., biological nitrogen fixation), solubilizing soil minerals to release mineral-associated nutrients (e.g., phosphate-solubilizing bacteria) or mineralizing soil organic matter to recycle organically bound nutrients.

One way that roots create gradients to increase nutrient availability is through the positive priming effect, which transfers carbon-rich root exudates into the rhizosphere (Fig. 4). In sunflower (*Helianthus annuus*) and soybean (*Glycine max*) crops, the root system had 27–245% more soil C mineralization, and 36–62% greater N mineralization than the bulk soil (Zhu et al., 2014). Living roots increased the microbial biomass carbon by up to 28% and stimulated the activity of cellulose-degrading and lignin-oxidizing enzymes. The extent to which the enzyme activity is stimulated by roots depends upon the microorganisms involved, as well as crop and soil properties. For instance, the enzymes that degrade organic carbon and organic nitrogen substrates were elevated within 1 mm of the root, whereas phosphate-hydrolyzing enzymes increased up to 3.5 mm from the root of lentil (*Lens culinaris*) and maize (Razavi et al., 2016), relative to the bulk soil. Phosphorus cycle. This suggests greater activity of carbon-degrading and nitrogen-mineralizing microorganisms near the root surface, whereas phosphorus mineralization may be associated with endophytic fungi, whose hyphae extend a greater distance from the root surface. Phosphorus mineralization could be the result of phosphatase enzymes secreted by the fungi but may also be from phosphatase-producing bacteria that swim along the fungal hyphae to microsites with abundant organic phosphorus. This underscores that physical, chemical, and biochemical factors influence root-rhizobiome interactions, illustrating the need to fully describe the environmental context and to look beyond biological explanations for observed phenomenon.

Soil salinity is a growing global problem because of the naturally high salt concentration in many arid and semi-arid soils, and in irrigation water. As the salt concentration increases in the rhizosphere, this lowers the osmotic potential and prevents root cells from absorbing water. Roots that absorb Na^+ and Cl^- from the soil water become imbalanced in their osmotic and redox potentials, leading to a reduced photosynthesis rate and to oxidative stress in cells. Furthermore, a high concentration of Na^+ in the rhizosphere can induce Ca^{2+} and K^+ deficiencies in the plant, due to competitive inhibition in the uptake of these essential nutrients resulting from the ionic imbalance of the soil water. However, plants have a natural tolerance for salts that is enhanced by the presence of naturally occurring bacteria and fungi in the rhizosphere (Table 3; (Abbas et al., 2019)). The rhizobiome can directly stimulate

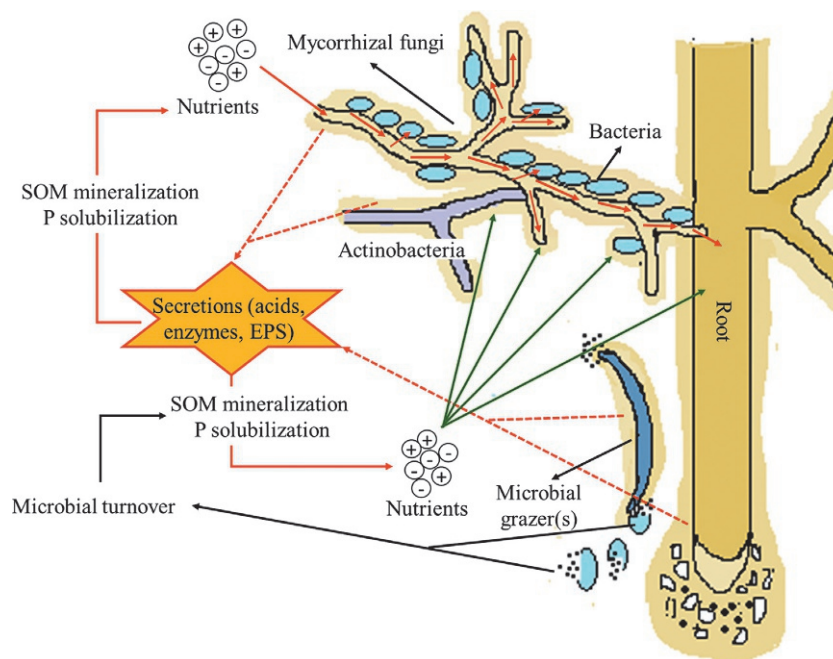


Fig. 4 Biological interactions that increase the rate of soil organic matter (SOM) mineralization and nutrient biogeochemical cycling in the rhizosphere and the mycorrhizosphere (i.e., through bacteria and fungal iterations). EPS: extracellular polymeric substances.

Table 3 Salt tolerance in crops is enhanced by root-associated bacteria that support plant growth through a variety of mechanisms.

Crop	Root-associated bacteria	Mechanism for salt tolerance
Tomato (<i>Lycopersicon esculentum</i>)	<i>Achromobacter</i> sp. <i>Streptomyces</i> sp.	Reduce ethylene level Produce proline
Cucumber (<i>Cucumis sativus</i>)	<i>Acinetobacter</i> sp. <i>Burkholderia</i> sp.	Reduce ethylene content More water uptake and chlorophyll
Wheat (<i>Triticum aestivum</i>)	<i>Aeromonas</i> sp. <i>Serratia</i> sp. <i>Enterobacter</i> sp. <i>Klebsiella</i> sp. <i>Microbacterium</i> sp.	Produce extrapolymeric substances Reduce ethylene production Increase K ⁺ content, produce proline
Barley (<i>Hordeum vulgare</i>)	<i>Curtobacterium</i> sp. <i>Hartmannbacter</i> sp.	Produce proline Produce extrapolymeric substances
Rice (<i>Oryza sativa</i>)	<i>Micrococcus</i> sp.	Production of IAA and siderophores
Maize (<i>Zea mays</i>)	<i>Azotobacter</i> sp. <i>Geobacillus</i> sp.	Improved nutrition Produce proline, more photosynthesis
Wild mint (<i>Mentha arvensis</i>)	<i>Exiguobacterium</i> sp.	Produce extrapolymeric substances
Strawberry (<i>Fragaria ananassa</i>)	<i>Kocuria</i> sp.	Improve phosphate nutrition
Sunflower (<i>Helianthus annuus</i>)	<i>Azospirillum</i> sp.	Greater chlorophyll content
Pea (<i>Pisum sativum</i>)	<i>Arthrobacter</i> sp. <i>Variovorax</i> sp.	Increase nutrient uptake Produce ACC deaminase
Soybean (<i>Glycine max</i>)	<i>Bacillus</i> sp. <i>Pseudomonas</i> sp.	More ACC deaminase activity More proline and IAA
Peanut (<i>Arachis hypogaea</i>)	<i>Brachybacterium</i> sp. <i>Brevibacterium</i> sp. <i>Haererohalobacter</i> sp. <i>Ochrobactrum</i> sp.	Increase K ⁺ content Produce IAA and ACC deaminase
Mung bean (<i>Vigna radiata</i>)	<i>Enterococcus</i> sp. <i>Pantoea</i> sp.	Lower Na ⁺ uptake Greater ACC deaminase
Lentil (<i>Lens esculenta</i>)	<i>Oceanobacillus</i> sp.	Produce extrapolymeric substances

ACC: 1-Aminocyclopropane-1-carboxylic acid; IAA: Indole-3-acetic acid.

Adapted from Abbas R, Rasul S, Aslam K, Baber M, Shahid M, Mubeen F, and Naqqash T (2019) Halotolerant PGPR: A hope for cultivation of saline soils. *Science* 31: 1195–1201. doi: 10.1016/j.jksus.2019.02.019.

plants to extend their root architecture. Microorganisms secrete phytohormones like gibberellins and indole acetic acid that make plant roots grow longer, with greater root surface area and more root tips. Consequently, the roots will gain access to water and nutrients from the soil profile, beyond an area of soil salinity.

Conclusions

Roots attract and associate with a diverse community of bacteria, fungi and archaea that form their rhizobiome. Biological interactions among the root-rhizobiome are controlled by a cell-to-cell communication through the exchange of chemical signal molecules. Advances in phytochemistry and plant metabolomics have made it possible to quantify the hundreds of compounds that are produced and received by roots, as well as the root-associated and endophytic microorganisms in the rhizobiome. We anticipate more discoveries about how these infochemicals control plant responses at the cellular, physiological and anatomical levels. Future investigations will be enhanced by using metagenomic, transcriptomic and proteomic analyzes to understand the genetic basis for the two-way transmission of information between the root and its rhizobiome. This research should provide clues about characteristics of the compatible plant and microbial genotypes. It will also provide insight about the role of non-culturable microorganisms, including archaea, that are generally under-represented in the literature. One outcome from this research could be to identify suitable microbial consortia with potential to enhance plant growth. However, we must remain alert to the environmental factors that induce phenotypic variation in plants and microorganisms. As much of the evidence for beneficial root-rhizobiome interactions comes from well-controlled growth bench and laboratory studies, we recommend confirmatory field studies. This will corroborate the pathways for assembly of diverse rhizobiomes that support resilient crops in sustainable food and fiber production systems.

References

- Abbas R, Rasul S, Aslam K, Baber M, Shahid M, Mubeen F, and Naqqash T (2019) Halotolerant PGPR: A hope for cultivation of saline soils. *Science* 31: 1195–1201. <https://doi.org/10.1016/j.jksus.2019.02.019>.
- Aceves-García P, Álvarez-Buylla ER, Garay-Arroyo A, García-Ponce B, Muñoz R, and Sánchez MP (2016) Root architecture diversity and meristem dynamics in different populations of *Arabidopsis thaliana*. *Frontiers in Plant Science* 7: 858. <https://doi.org/10.3389/fpls.2016.00858>.
- Alquéres S, Meneses C, Rouws L, Rothballer M, Baldani I, Schmid M, and Hartmann A (2013) The bacterial superoxide dismutase and glutathione reductase are crucial for endophytic colonization of rice roots by *Gluconacetobacter diazotrophicus* PAL5. *Molecular Plant-Microbe Interactions* 26: 937–945. <https://doi.org/10.1094/MPMI-12-12-0286-R>.
- Compant S, Reiter B, Nowak J, Sessitsch A, Clément C, and Barka EA (2005) Endophytic colonization of *Vitis vinifera* L. by plant growth-promoting bacterium *Burkholderia* sp. strain PsJN. *Applied and Environmental Microbiology* 71: 1685–1693. <https://doi.org/10.1128/AEM.71.4.1685-1693.2005>.
- Dutta P, Muthukrishnan G, Gopalasubramaiam SK, Dharmaraj R, Karuppaiah A, Loganathan K, Periyasmy K, Pillai MA, Upamanyu GK, Boruah S, Deb L, Kumari A, Mahanta M, Heisnam P, and Mishra AK (2022) Plant growth-promoting rhizobacteria (PGPR) and its mechanisms against plant diseases for sustainable agriculture and better productivity. *Biocell* 46: 1843–1859. <https://doi.org/10.32604/biocell.2022.019291>.
- Hansen BL, Pessotti RD, Fischer MS, Collins A, El-Hifnawi L, Liu MD, and Traxler MF (2020) Cooperation, competition, and specialized metabolism in a simplified root nodule microbiome. *MBio* 11. <https://doi.org/10.1128/mBio.01917-20>. e01917–e01920.
- Hosseini F, Mosaddeghi MR, Dexter AR, and Sepehri M (2019) Effect of endophytic fungus *Piriformospora indica* and PEG-induced water stress on maximum root growth pressure and elongation rate of maize. *Plant and Soil* 435: 423–436. <https://doi.org/10.1007/s11104-018-03909-7>.
- Huang GQ, Kilic A, Karady M, Zhang J, Mehra P, Song XY, Sturrock CJ, Zhu WW, Qin H, Hartman S, Schneider HM, Bhosale R, Dodd IC, Sharp RE, Huang RF, Mooney SJ, Liang WQ, Bennett MJ, Zhang DB, and Pandey BK (2022) Ethylene inhibits rice root elongation in compacted soil via ABA- and auxin-mediated mechanisms. *Proceedings of the National Academy of Sciences* 119: e2201072119. <https://doi.org/10.1073/pnas.2201072119>.
- Hughes KA, Gandar PW, and de Silva HN (1995) Exploration and exploitation of soil by apple, kiwifruit, peach, Asian pear and grape roots. *Plant and Soil* 175: 301–309. <https://doi.org/10.1007/BF00011366>.
- Kakouridis A, Hagen JA, Kan MP, Mambelli S, Feldman LJ, Herman DJ, Weber PK, Pett-Ridge J, and Firestone MK (2022) Routes to roots: Direct evidence of water transport by arbuscular mycorrhizal fungi to host plants. *New Phytologist* 236: 210–221. <https://doi.org/10.1111/nph.18281>.
- Li S and Wu FZ (2018) Diversity and co-occurrence patterns of soil bacterial and fungal communities in seven intercropping systems. *Frontiers in Microbiology* 9: 01521. <https://doi.org/10.3389/fmicb.2018.01521>.
- Luthfiana N, Inamura N, Tantriani ST, Saito K, Oikawa A, Chen WG, and Tawarayama K (2021) Metabolite profiling of the hyphal exudates of *Rhizophagus clarus* and *Rhizophagus irregularis* under phosphorus deficiency. *Mycorrhiza* 31: 403–412. <https://doi.org/10.1007/s00572-020-01016-z>.
- Marcacci KM, Warren JM, Perfect E, and Labbé JL (2022) Influence of living grass roots and endophytic fungal hyphae on soil hydraulic properties. *Rhizosphere* 22: 100510. <https://doi.org/10.1016/j.rhisph.2022.100510>.
- Neal AL, Ahmad S, Gordon-Weeks R, and Ton J (2012) Benzoxazinoids in root exudates of maize attract *Pseudomonas putida* to the rhizosphere. *PLoS One* 7: e35498. <https://doi.org/10.1371/journal.pone.0035498>.
- Razavi BS, Zarebanadkouki M, Blagodatskaya E, and Kuzyakov Y (2016) Rhizosphere shape of lentil and maize: Spatial distribution of enzyme activities. *Soil Biology and Biochemistry* 96: 229–237. <https://doi.org/10.1016/j.soilbio.2016.02.020>.
- Remus R, Hüve K, Pörschmann J, and Augustin J (2016) Determining the timepoint when ¹⁴C tracer accurately reflect photosynthate use in the plant-soil system. *Plant and Soil* 408: 457–474. <https://doi.org/10.1007/s11104-016-3002-2>.
- Sandnes A, Eldhuset TD, and Wollebæk G (2005) Organic acids in root exudates and soil solution of Norway spruce and silver birch. *Soil Biology and Biochemistry* 37: 259–269. <https://doi.org/10.1016/j.soilbio.2004.07.036>.
- Sindhu SS, Sehrawat A, and Glick BR (2022) The involvement of organic acids in soil fertility, plant health and environment sustainability. *Archives of Microbiology* 204: 720. <https://doi.org/10.1007/s00203-022-03321-x>.
- Ta CAK and Amason JT (2016) Mini review of phytochemicals and plant taxa with activity as microbial biofilm and quorum sensing inhibitors. *Molecules* 21: 29. <https://doi.org/10.3390/molecules21010029>.
- Taylor AR and Bloom AJ (1998) Ammonium, nitrate, and proton fluxes along the maize root. *Plant, Cell & Environment* 21: 1255–1263. <https://doi.org/10.1046/j.1365-3040.1998.00357.x>.
- Venturi V and Keel C (2016) Signaling in the rhizosphere. *Trends in Plant Science* 21: 187–198. <https://doi.org/10.1016/j.tplants.2016.01.005>.

- Verma PK, Verma S, and Pandey N (2022) Root system architecture in rice: Impacts of genes, phytohormones and root microbiota. *3 Biotech* 12: 239. <https://doi.org/10.1007/s13205-022-03299-9>.
- Warren JM, Brooks JR, Meinzer FC, and Eberhart JL (2008) Hydraulic redistribution of water from *Pinus ponderosa* trees to seedlings: Evidence for an ectomycorrhizal pathway. *New Phytologist* 178: 382–394. <https://doi.org/10.1111/j.1469-8137.2008.02377.x>.
- Zhang XY, Gao ZF, Zhang XX, Bai WB, Zhang L, Pei HB, and Zhang YJ (2020) Control effects of *Bacillus siamensis* G-3 volatile compounds on raspberry postharvest diseases caused by *Botrytis cinerea* and *Rhizopus stolonifera*. *Biological Control* 141: 104135. <https://doi.org/10.1016/j.biocontrol.2019.104135>.
- Zhu B, Gutknecht JLM, Herman DJ, Keck DC, Firestone MK, and Cheng WX (2014) Rhizosphere priming effects on soil carbon and nitrogen mineralization. *Soil Biology and Biochemistry* 76: 183–192. <https://doi.org/10.1016/j.soilbio.2014.04.033>.

Soil respiration

Jeremiah D Vallotton^a, Sergey Blagodatsky^b, and Adrian Unc^c, ^aEnvironmental Science Program, Memorial University of Newfoundland, St. John's, NL, Canada; ^bInstitute of Zoology, University of Cologne, Cologne, Germany; ^cSchool of Science and the Environment, Memorial University of Newfoundland, Corner Brook, NL, Canada

© 2023 Elsevier Ltd. All rights reserved.

Introduction: Soil respiration as indicator of biological activity	370
Respiration and global GHG balance	371
Aerobic vs. anaerobic respiration in soils	371
Drivers and controls of soil respiration	371
Soil temperature	371
Soil moisture	372
Available soil carbon	372
Soil carbon pools	373
Nutrient availability	373
pH	373
Soil organic matter supply and quality	374
Interplay of temperature and hydrology	374
Impact of management on soil respiration	374
Drainage	374
Tillage	375
Nutrient management (mineral vs organic)	375
Cover crops	375
Application of soil respiration measurements in agronomy and ecology	375
Estimation of soil organic matter decomposition rates and soil C balance	375
Index of soil quality and soil health	375
Measuring respiration	376
Soil respiration measurements as an integrated proxy for soil biological activity	376
Measuring respiration on site	376
Use of isotopes for partitioning of soil respiration	376
Ex-situ measurements of respiration	376
Respiration and climate change	377
Summary	378
References	378

Abstract

Soil respiration is a useful indicator of global carbon balance and terrestrial carbon exchange with the atmosphere. Different soil carbon pools metabolize at variable rates based on environmental constraints on microbiological activity. Respiration is thus primarily controlled by climatic effects (temperature and hydrology) and soil management. Soil respiration can be measured both on site and in laboratory, with each method possessing distinct advantages and disadvantages relating to actual or potential respiration rates. These tests help us understand the importance of soil respiration as an indicator of soil conditions and C cycling and enable some predictions of changes in soil carbon under climate change.

Key points

- Soil respiration is controlled by complex interactions between climate (temperature and hydrology), land management, and soil microbiology.
- Soil carbon can be divided into pools that respire at different rates depending on their accessibility to microbes.
- Soil respiration can be measured via in-field and in laboratory methods, each of which have advantages in describing the complete picture of soil respiration.
- Soil respiration is a useful soil health and global carbon cycling indicator that is of particular interest under conditions of expanding human society and climate change.

Introduction: Soil respiration as indicator of biological activity

Soil respiration integrates the metabolic processes of all types of organisms living in soil, (including the roots of plants) as they break down organic molecules (i.e., carbonaceous compounds), to harvest the energy necessary for their metabolism. The obtained energy is stored in form of ATP or other compounds capable of transferring electrons to release energy that is employed in downstream metabolic processes. The two main products of respiration are water and gaseous carbon dioxide (CO₂), which can be measured as it leaves the soil. As energy is required for any cellular biochemical processes, respiration is the gatekeeping process regulating metabolic energy fluxes in the cell and can be employed to understand the functional state of a soil. This is why measurements of soil respiration are utilized as best direct indicators of soil activity and health.

As with any biochemical process, temperature and water availability (Fig. 1) are key controlling parameters, and respiration in most soil types is largely an aerobic process governed by the availability of oxygen. Other than in bogs and marshes, in-situ respiration measurements are commonly assumed to reflect aerobic respiration; ex-situ measurement methods actively create aerobic conditions. Furthermore, respiration depends on the availability of carbonaceous molecules and on the availability of all macro- (N, P, K) and microelements (Ca, Mg, Mn...) that are necessary for high-functioning of biochemical processes involved in metabolism of an organism. These elements may be provided during the decomposition of plant or microbially derived organic material or can be made available in ionic or mineral forms through natural biogeochemical processes or as a result of their addition to soils in form of various fertilizers.

Therefore, soil respiration is a proximate indicator tightly linked to many individual processes, primarily respiration of autotrophic organisms, such as plants and algae, the breakdown of carbonaceous compounds by heterotrophic organisms, including bacteria or fungi, and predation. Respiration reflects the stability of the soil food web, which governs the pathways and speed of the transformation of carbonaceous matter along its trophic nodes, and also reflects the impact of management on altering the proportions and types of organic matter, minerals and altering soil heat and water fluxes.

Though total soil respiration is not specific to particular groups or types of organisms (e.g., bacteria or fungi), it nonetheless is a very useful measure of soil condition, as it indicates overall biological activity driven by the presence of carbonaceous substrates in soil and roots, as affected by any external constraints. The term “soil respiration” addressed in this chapter differs from the term “CO₂ emission from the soil surface,” which presumes additional consideration of physico-chemical processes of gas transport in soil profile, while our topic (soil respiration) is more intricately linked with biological nature of CO₂ production in soil.

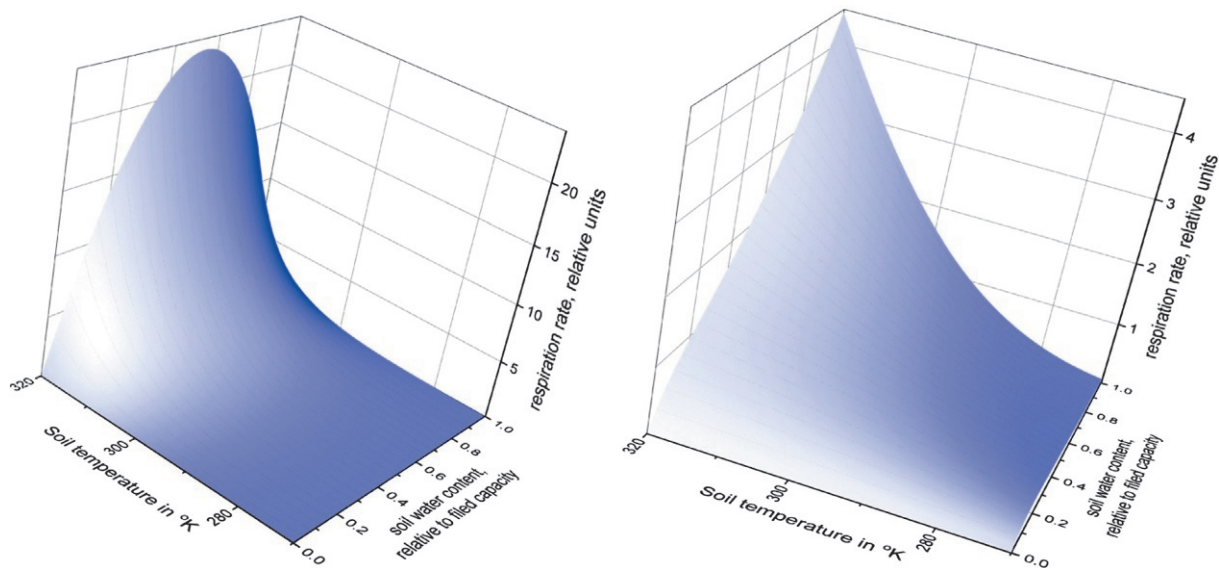


Fig. 1 Soil respiration response to changes in soil moisture and temperature. (A) empirical function for mineral soil combining an Arrhenius type temperature response - $r = r_{ref} \cdot \exp\left(-\frac{E_a}{R} \left(\frac{1}{T} - \frac{1}{T_{ref}}\right)\right)$ and moisture scaling function based on relative water saturation $f(\theta_s) = a\theta_s - b\theta_s^2$ (Eqs. 20 and 21, respectively, in Moyano et al., 2018); (B) empirical function reflecting soil surface CO₂ efflux (root and soil respiration) optimized for the set of forest and shrubland sites in Europe and USA - $f(T_{soil}, RSWC) = \exp\left(E_0(RSWC) \cdot \left(\frac{1}{T_{ref} - T_0} - \frac{1}{T_{soil} - T_0}\right)\right)$, where $E_0(RSWC) = a_{E_0} + b_{E_0} \cdot RSWC$ and RSWC is relative soil water content (Eqs. 2–5 in Reichstein et al., 2003).

Respiration and global GHG balance

Soil respiration, together with CO₂ assimilation by plants, is routinely estimated as global primary productivity (GPP) and is the driving force behind global carbon cycling between terrestrial and atmospheric pools of carbon. Just as GPP draws down 123.4 PgC of carbon annually and fixes it in various organic carbon forms within the terrestrial ecosystem, so respiration allows for the conversion of C from organic forms back into inorganic CO₂ and its return to the atmosphere (121.6 PgC year⁻¹). Without human usage of fossil fuels, these two forces are in near equilibrium, with a minor increase in terrestrial carbon of +1.8 PgC year⁻¹. The increase in land sink was 3.4 ± 0.9 PgC year⁻¹ during 2010–2019, while CO₂ emission increased by 1.6 PgC year⁻¹ during the same period (Friedlingstein et al., 2020). Thus, while soil respiration is a major contributor to carbon cycling, it is a natural process that counterbalances GPP and ensures global C cycling. However, human activity, such as land clearance, disturbance, and above all, the utilization of fossil fuels contributes additional carbon emissions to this cycle and causes an imbalance, creating a net gain to the atmosphere of 4 PgC year⁻¹, driving climate change.

Aerobic vs. anaerobic respiration in soils

Aerobic respiration is the primary source of CO₂ within soil, as oxygen is critical for determining the rate of soil respiration as it is required for basic metabolism. Water content controls the availability of oxygen via diffusion, and thus determines whether conditions are aerobic or anaerobic. In the absence of oxygen (anaerobiosis), anaerobic respiration employs molecules other than oxygen as electron acceptors, and thus can occur when soils are waterlogged. Anaerobic respiration is energetically less efficient and thus not preferred, and releases a variety of compounds, particularly CO₂, CH₄, and NO_x. Thus, the contribution of anaerobic respiration to the C efflux from soil is much smaller than the main source of CO₂ in soil, which is related to the oxidation of organic substances with O₂. Therefore, CO₂ released from a soil reflects both aerobic and anaerobic respiration. Even if a soil is overwhelmingly aerobic, there will always be anaerobic microsites present within the soil microaggregates, where many small pores are filled with water, causing slower diffusion of oxygen and often high microbial activity, resulting in very small oxygen concentrations. However, these sites account for very small percentages of carbon processing (background levels) and are often ignored in respiration measurements (Smith et al., 2003). In a well-drained soil, anaerobiosis will only continue in soil micropores, but water remains the primary controller on carbon cycling in bogs, which have proportionally more carbon due to functional stasis of plant material in various stages of decomposition due to anaerobiosis.

Plant roots often create specific conditions in the ‘rhizosphere’ by pumping carbonaceous and other compounds into the soil matrix so as to preferentially select for particular microbial consortia, raise the pH, and otherwise create ideal conditions for plant growth. Roots also, however, respire, increasing the concentration of CO₂ and reducing that of oxygen.

Drivers and controls of soil respiration

Soil temperature

Generally, rising soil temperature increases soil respiration (see e.g., Fig. 2), while temperatures near or below zero Celsius suppress respiration. The sensitivity of soil respiration to increases in temperature can be expressed as the Q₁₀ value; that is, the factor by which soil respiration increases for each 10 °C increase in temperature. However, temperature sensitivity of soil respiration (Q₁₀) is also affected by changes in the availability of soil organic carbon, nutrients, and in the soil water potential. Hence, when hydrology is factored in, a substantial amount of Q₁₀ values, especially in areas that are considered very susceptible to climate change (e.g., boreal forests, wetlands) are perhaps less at risk than previously thought, while tropical systems appear to be mostly unaffected by hydrology (Fig. 2).

As enzymes carrying out metabolic activities become inactive at lower temperatures, their activation rates, that is the amount of energy (expressed in kJ mol⁻¹) required to trigger the biochemical reaction carried out by an enzyme, decreases with an increase in temperature. However, the enzymatic apparatus of microbial communities that have acclimated to different temperatures are also particularly adapted to the respective temperature regimes. Cold adapted organisms and cold soil microbiomes are more sensitive to changes in temperature, having lower activation temperatures. At the upper end (greater than 30 °C) microbes slow down or stop, and it is agreed that respiration declines vertiginously for temperatures over 40 °C.

The relationship between temperature and soil respiration might be fitted using an exponential equation:

$$R_T = a \times \exp^{b \times T}$$

Q₁₀, namely when the values are calculated relative to a 10 °C change, may be calculated as:

$$Q_{10} = \exp^{10 \times b}$$

The terms a and b are empirically fitted and vary as a function of water and carbon availability and differ according to the level of adaptation of the soil community to the local climate (i.e., cold vs. hot biomes).

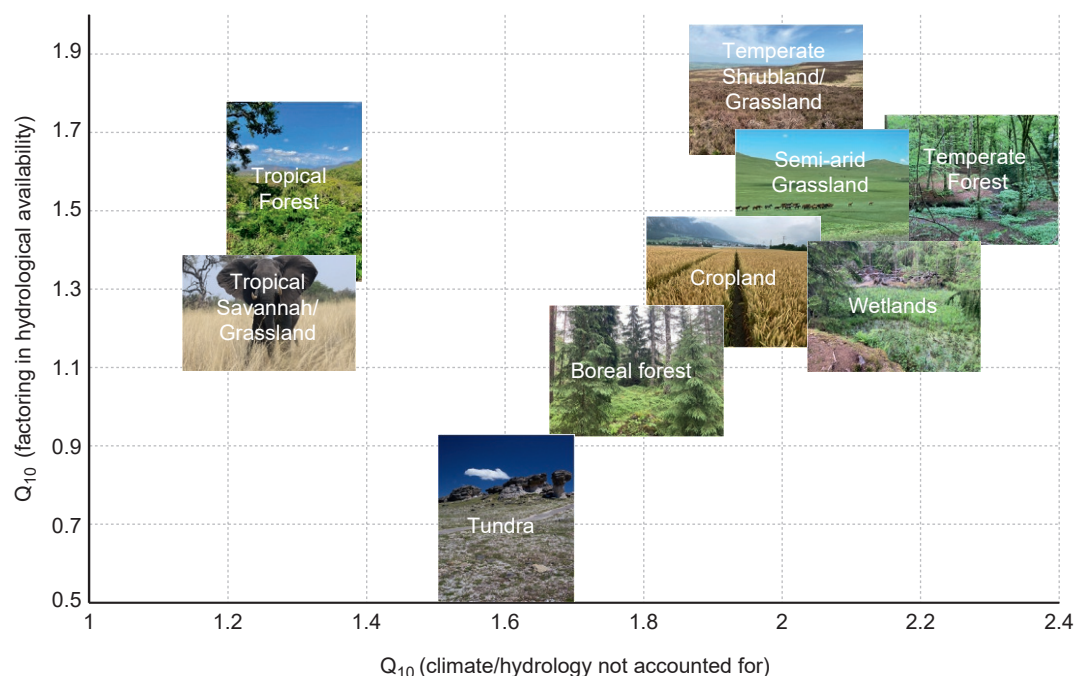


Fig. 2 While Q_{10} is linked to temperature it might be more dependent on hydrology than is usually thought. When hydrology is factored in, high Q_{10} may be lowered such as for boreal forest (compared to tropical savannah or grassland where both are the same); x-axis is temperature-only Q_{10} , while y-axis is hydrologically-sensitive Q_{10} (i.e. reflecting hydrological effects on temp dependency of C degradation). Graph is created using data from Fan et al. (2022). Images copyrighted to the chapter's authors.

Soil moisture

As biochemical processes require water, its availability is a colimiting factor. It is common to describe soil water content in terms of volumetric units, but this does not truly describe its availability. The upper limit of soil available water content, a parameter dependent on its capacity to retain water against gravitational pull, and which varies with the soil texture, is achieved at the field capacity for water, commonly taken to be -33 kPa (-6 to -33 kPa, soil texture and aggregation dependent) matric potential. The lower limit of water availability to plants is the permanent wilting point (PWP) at a matric potential of -1500 kPa; this indicates that only water in pores >0.2 μm diameter is available to plants. At PWP fine textured soils may have as much as 33% water content volumetrically, while sandy soils might have $<10\%$. In direct contact, microbes might be able to access some of this water. However, the water content also affects the diffusion rates of soluble nutrients required by soil microbiota. Drier soils will interfere with the consistent solubilization and diffusion of necessary nutrients (Moyano et al., 2018). In mineral soils diffusion of nutrients ceases at a matric potential of $-14,000$ kPa. In plant litter the nutrients might be available for matric potentials as low as $-36,000$ kPa. as microbes are in immediate proximity of rich organic materials, but below this, desiccation stops microbial activity; these thresholds are uniform across biomes (Manzoni et al., 2012).

It is generally considered that water contents at matric potentials equivalent to field capacity or above (i.e., closer to saturation, matric potential 0 kPa) is available with little restriction. Excessive water, however, will lead to anaerobiosis, that is lack of oxygen, which will halt aerobic respiration. As a rule of the thumb, soil moisture levels of $\sim 50\%$ water-filled pore space are considered ideal for allowing maximal respiration rate without oxygen starvation that results in anaerobiosis (Franzluebbers, 1999). Thus, soil moisture and temperature are closely and intimately linked (Fig. 1): a hot soil without water may have minimal respiration, while a cool soil with ideal moisture may have higher rates.

Another mechanism affecting availability of water to respiring microbes is excess salts in the soil solution. This reduces availability of water to microbial cells through osmotic stress. High ionic strength of the solution can inhibit microbial growth independent of the type of ions. However, ion specific toxicities can affect soil microbial respiration, which is less inhibited by salts containing SO_4^{2-} than Cl^- (Rath et al., 2016). The drying of soil also increases the ionic strength of the soil solution. Fungi are more resilient to increases in the ionic strength of soil solution.

Available soil carbon

As most soil microbes are heterotrophic, soil microbial respiration depends on the availability and accessibility of carbon-based substrate. Soluble and exposed carbon compounds are the immediate source of carbon for soil microbes. Exposed, complex carbon compounds (e.g., cellulose, hemicellulose, chitin, lignin, proteins) need to be degraded before uptake and thus microbes produce

the necessary enzymes. The less available a carbon compound the more the energy expenditure of microbes and thus the less efficient the catabolism. Less efficient organic carbon catabolism results in more respiration per biomass unit. A simple means to describe the efficiency of carbon utilization by a soil microbial community is the CO_2 quotient ($q\text{CO}_2$), which describes the amount of CO_2 respired by a unit of microbial biomass carbon ($\text{unit CO}_2\text{-C unit}^{-1} \text{C}_{\text{mic}} \text{h}^{-1}$). Sugars are an efficient source of carbon that can be immediately employed as energy sources, and thus lead to immediate increases in respiration rates. Bacteria can also take up amino acids that can be directly employed in cell anabolic and catabolic processes, and thus are particularly energy-efficient sources of carbon. Complex organic substances (celluloses, chitin, lignin), which are less available to microorganisms can be degraded more efficiently by fungi, some species of which possessing of specific enzyme systems; this longer degradation and utilization pathway leads to a slower utilization, and thus an apparent lower respiration rate. If carbon sources are easily available, soil microbes deal with environmental stresses through an acceleration of carbon utilization and thus accelerated metabolism, leading to more C to be respired as CO_2 .

Soil carbon pools

Soil organic carbon (SOC) can be conceptually assigned to labile, intermediate and stable carbon pools. The proportion of labile carbon is the main driver of soil respiration, modulated by temperature and moisture conditions. In practice this partitioning is less clearly defined as the definition of each pool is dependent on variably imprecise methods of measurement.

As organic materials enter the soil it is physically shredded and chemically degraded by digestion in the alimentary canal of mesofauna, and through exoenzymes produced by microfauna. Depending on the propensity of a certain type of organic matter to be degraded soil organic matter pools of variable stability form. This leads to a series of particulate organic matter (POM) pools from coarse (4000–250 μm equivalent diameter) to fine (250–50 μm equivalent diameter). The coarser POM is less accessible to microbes. Exposed fine fraction POM (i.e., not physically protected by strong attachment onto charged mineral particles or enclosed within micropores of soil aggregates and not readily accessible to microbes) is more easily degradable by microbes and has been shown to be correlated to soil microbial respiration (Koutika et al., 2008).

Another approach to describe active carbon is to quantify the permanganate oxidizable carbon (POXC). Although POXC is generally positively related to POM and to the soil microbial biomass carbon (MBC), it is more closely related to POM <250 μm and also to heavier (>1.7 g cm^{-3}) rather than lighter POM (Culman et al., 2012). Thus, POXC may be an easy to measure, proximate indicator of the labile fraction of SOM and thus a good indicator of the respiratory potential of a soil.

Dissolved organic carbon (DOC), a complex mixture of mainly hydrophilic fractions of variable aromaticity, is assumed to be easily available to microorganisms and thus a potential driver of respiration. DOC is produced under both anaerobic and aerobic conditions; however, under aerobic conditions the DOC is also rapidly consumed and thus measuring DOC production is less accurate.

SOC that is physically separated from microbial biomass such as physically protected SOC in aggregates, or SOC in the subsurface layers of acid soils, does not participate in microbial metabolism, but it may be made available when soil aggregates are broken apart through soil management such as tillage.

Nutrient availability

As cell metabolism requires nutrients other than C, their availability will affect the interaction between soil microorganisms and soil C pools. Nitrogen (N) and phosphorus (P) are the nutrients required in the largest amounts as they are critical to the production of organic molecules essential to the functioning of a cell. N is an essential component of proteins and thus critical to enzymatic activities, whereas P is essential for management of cellular energy (i.e., ATP). Each organism assimilates C and nutrients to proportions necessary for their structural and functional components. Imbalance in C and nutrient provision will affect cell physiology and thus the proportional likelihood for C to be retained in microbial biomass or lost in respiration. The C:N ratio is the most common indicator used in soil management to explain the likelihood for C to be retained by soil heterotrophs (i.e., fixed) or be lost to respiration. The C:N ratio controls the balance between immobilization and mineralization of N. Insufficient N will limit the amount of soil organic C that can be retained by heterotrophs. As fresh litter with a C:N:P ratio of 3000:46:1 is decomposed, the nutrients are concentrated in soil microbial biomass with a ratio of 60:70:1 (Sinsabaugh et al., 2009). The excess C is thus not included in the biomass and can be respired as CO_2 . Nevertheless, given that some complex organic polymers, such as lignin, are not easily decomposed, the average organic matter of soil has a ratio of 186:13:1.

P affects soil heterotrophic respiration in a manner partially similar to N. If available P is added to nutrient rich and stable soil systems it may increase SOC, suggesting it is the limiting factor. Co-addition of N and P may have logically consistent but variable effects depending on the original state of the receiving soil: a minor increase in respiration is likely.

pH

For constant moisture and temperature conditions, microbially driven respiration will decline as soils acidify. Neutral to slightly alkaline pH conditions favor microbial activity and thus the greatest respiration rates permitted by conditions of temperature, moisture and carbon availability. Survival and growth at greater acidity (lower pH) needs additional energy expenditure, meaning that larger fraction of C will be diverted to catabolism and relatively more CO_2 will be respired by soil microorganisms. So, the net effect of soil acidification on cumulative CO_2 production depends on the two opposite trends: decrease of general decomposition activity and increase of respired C fraction due to lower carbon use efficiency at low pH.

Soil organic matter supply and quality

Soil organic matter, as the supplier of carbon to subsurface heterotrophic organisms, is the primary driver of soil respiration, which means that land managers must consider its supply and quality. Carbon supply comes from living organisms, most of which is generated by continuous plant cover (as found in nature), but actions that reduce or eliminate this supply (e.g., residue burning or removal, clearcutting, land clearance) will, over the long-term, starve the soil of its carbon and decrease its respiration, often changing the community of organisms that reside and respire. As described above (see "Available Soil Carbon"), carbon quality matters as well; more complex plant matter will add in a greater variety of carbon, which in turn can support more diverse and larger populations of organisms. Indeed, one of the hallmarks of natural landscapes is that they often accumulate carbon over time, while human land management (agriculture) has historically depleted soil carbon through reducing the quality of plant inputs and the quantity of matter returned to the soil. Adding compost or manure can help alleviate these problems, and soils can be regenerated through the addition of different and complex forms of carbon (thus favoring stable, long-term respiration).

Interplay of temperature and hydrology

For much of the biosphere, respiration increases with increasing temperature, unless it is hydrologically limited. Arid regions, as previously mentioned, have excess heat but limited water, hence respiration is controlled by water access and plant presence (which decreases albedo and increases soil moisture), not by temperature. At the opposite end of the spectrum, boreal regions would seem to be primarily influenced by lack or presence of heat, but there is more complexity than this. In particular, the interplay of temperature and water is what controls respiration. For example, winters with less snowfall cause deeper frost penetration of longer duration and more mechanical degradation of C through cryoturbation, which in turn means there is more C available for metabolism. On the other hand, warmer winters (with shallower freeze-thaw and snowmelt) can cause massive flushes of DOC into local waterways (Haei et al., 2013). Increased rainfall can protect C through saturation of the soil, whereas moderate drying events can result in runaway respiration as C is exposed. In areas such as the southern Siberian forest (e.g., Mongolia), which have a transect with arid and boreal conditions, a slight decrease in moisture or increase in temperature can be enough to cause natural degradation of forests and their replacement by grasslands (Munkhjargal et al., 2020). Finally, podzolic forests in particular have interplays with pH, which, although caused by coniferous vegetation and high rainfall, is not a direct control of respiration but rather a useful predictor of its viability (more acidity often is linked to more recalcitrant litter, and thus lower respiration).

Impact of management on soil respiration

Practically, soil respiration is a near-absolute measure of carbon leaving the soil in a gaseous form, and in this way is closely related to global carbon dynamics: increased respiration means more carbon lost from the terrestrial system to the atmosphere. It also closes the carbon cycle that starts with its photosynthetic fixation by autotrophs. However, the value for the rate of soil respiration is also a highly contextual number that varies depending on the climatic or soil conditions that drive it, even under conditions of 'equal carbon.' One example is arid soils; many of these soils (e.g., Mediterranean, North American desert) are characterized by often high temperatures but rainfall that is sporadic and often confined to a single season or month of the year. Soil microbial populations adapt to these conditions by having a very low steady-state respiration followed by rapid metabolism when rare rain events occur, often hundreds of times larger than when dry. Wetlands are another example: high water table and constant anaerobic conditions predispose them to have reduced respiration, but when drained or during dry spells, the extremely high carbon content of these soils is very susceptible to destructive metabolism. Respiration in agricultural soils often does not match natural soils of the same climate or topography, as the transition from natural carbon-rich systems to highly-managed agricultural systems can render the soils more homogenous in nutrients, carbon, and water content compared to natural systems. Although there are ways to improve carbon retention and input in modern agriculture (e.g., diverse cropping rotations, organic inputs, no-till), the relatively carbon-starved nature of modern farming means that these systems have lower stable reserves and inputs of carbon, less diverse microbial populations, and therefore far smaller respiration rates than natural systems. Additionally, the proportion of C respired versus the total is often higher (more lost) than in natural systems, where higher respiration is outstripped by higher inputs. Natural systems such as forests or grasslands will still have differences between types of natural systems, e.g., grasslands may have lower respiration than forests due to water deficit, despite the low C:N nature of grassland soils that otherwise promotes respiration. Finally, different soil types with similar climatic conditions may have very different respiration: a Podzol adjacent to a Luvisol will often have lower respiration, as the Podzol is more acidic and nutrient starved. Hence, interpretation of respiration information requires a whole-system approach that considers how a given location's unique topographical/climatic/managed state affects its respiration.

Drainage

In natural systems, poor drainage often results in large accumulations of C. Since anaerobic soil conditions reduce growth and yield or kill plants, farmers maximize drainage via tillage, rock removal, leveling land, or introducing artificial drainage schemes. This

often increases respiration, as the soil is better aerated and more frequently tilled. However, at the same time, better plant growth and less anaerobiosis means more C inputs to soil and decreased emission of CH_4 and NO_x .

Tillage

Inversion and intensive tillage act to break up compaction, mix and homogenize the soil, and aerate it, all of which increase respiration in the short term by providing maximum oxygenation to the soil. However, in the long term, in addition to accelerating respiration, tillage can also retard it by causing compaction, anaerobiosis, poor structure, and impaired drainage, but also by burying C and rendering it less accessible to microbial activity (especially via 'deep plowing,' which inverts the uppermost 0.5–1 m of the soil with the subsoil beneath it). No-till is often recommended as a way to retain or increase soil C and seems to be particularly suited to water management in dry regions, but it is by no means a complete solution, for it primarily redistributes C in the profile (concentrating it in the uppermost 5 cm) rather than evenly throughout the plow-layer as under tillage. Hence, while no-till improves soil moisture and biological activity compared to the drier and highly aerobic conditions of tillage, it still does not consistently increase carbon sequestration.

Nutrient management (mineral vs organic)

Choice of fertilizer can affect respiration. Since most soils are N-limited rather than C-limited, mineral fertilizers tend to promote maximum respiration at the expense of soil organic matter and C retention (since plenty of N allows for even recalcitrant C to be metabolized), and further, select for r-strategist bacteria suited to rapid metabolization. Organic fertilizers (e.g., compost, manure) can have similar effects, but since most also contain a significant source of C alongside N rather than pure N, they are more likely to either have only mild effects on soil C, or in the case of some soils, increase it long term, a strategy that has been used for millennia to control soil fertility. Either way, fertilizer application will increase overall soil respiration, and very recent applications can cause artifacts when measuring it.

Cover crops

The use of cover crops can have contrasting effects on respiration that occur simultaneously. Increased or continuous plant cover can reduce respiration by improving aggregation, infiltration, increased diversity of soil biota (earthworm, insect, mite, etc.) resulting in better plant litter incorporation into soil, and functionally stable microbial community (dominating K-strategists). There can also be an increase in soil respiration since superior aggregation and infiltration allow for more ideal conditions for respiration, while the increased C processing, litter diversity, and penetration into the soil matrix can likewise increase available labile C to fuel respiration.

Application of soil respiration measurements in agronomy and ecology

Estimation of soil organic matter decomposition rates and soil C balance

Soil respiration measurements can be used to estimate soil organic matter decomposition rates as well as the overall soil carbon balance. This can be conceptually done through the use of three numbers: maximum soil respiration ($\text{mg CO}_2\text{-C kg}^{-1}$ soil) calculated on a basis of time (e.g., hours or days) and calibrated by temperature and moisture (e.g., water filled porosity), total yearly soil inputs of carbon (via plant biomass, e.g., g C kg^{-1} soil year^{-1}), and the soil organic matter content of the soil (e.g., percentage of SOM which is assumed to have a mixture of C, N, P, cations, etc.). Soil respiration numbers can be used to extrapolate the carbon leaving the soil yearly and this can be compared to plant inputs to determine the state of the soil carbon balance (e.g., $\text{mg CO}_2\text{-C kg}^{-1}$ soil lost per year compared to g C kg^{-1} soil year^{-1} input for a net gain of g C kg^{-1} soil year^{-1}), or to organic matter to determine how much native soil carbon is lost (e.g., 1000 g C kg^{-1} soil year^{-1} respired is about 1.3% of total SOM lost per year). If inputs balance output or exceed respiration, then the soil will have a net gain of carbon.

In practice, potential carbon mineralization rates can be estimated in laboratory incubation experiments, where relative size of theoretical carbon pools and their rate of decomposition can be estimated and results then used for comparative analysis of, for example, soils under different managements.

Index of soil quality and soil health

Soil respiration is an important tool currently utilized as an integral part of soil quality and health indices (e.g., Cornell Comprehensive Assessment of Soil Health, EU Soil Health Law). Since soil respiration measures biological activity constrained by abiotic factors, it is a very important response variable that connects abiotic features and soil management to the reality of soil life. Generally speaking, greater soil respiration values indicate better soil conditions (respiration as proxy for biological activity and C availability), and by referencing the soil to others in different climates and locations, specific land management practices can be assessed for their effect on soil respiration dynamics and thus soil health or quality. However, in some cases a large respiration value may be indicative of unsustainable biological activity (ultimate degradation of soil C). Therefore, interpretation and contextualization of respiration data is vital to make indices a useful tool, both by comparing to other locations of a similar type (e.g., agricultural to agricultural, not forest, within same climatic zone) and by measuring other important soil features of the soil quality or health index (e.g., soil aggregation or labile C).

Measuring respiration

Soil respiration measurements as an integrated proxy for soil biological activity

Soil respiration represents the sum of biological processes in the soil that produce CO₂ and is therefore a gross measure that does not specify the activity of any one organism. Given that plants, as autotrophs, contribute to both respiration and fixation of carbon (GPP), it is helpful to try and partition out different sources of CO₂ flux. Further, the response of autotrophic and heterotrophic respiration to changes in environmental factors (such as solar light intensity, soil moisture, and air and soil temperature) can differ. Therefore, separating CO₂ flux from decomposing soil organic matter in bulk soil from that found in the rhizosphere of living plant roots can help in understanding the mechanisms driving soil respiration. This separation is methodologically difficult however, as direct mechanical methods are required for extraction of root fragments from soil, and the destruction of fine root hairs may cause spikes in respiration that differ from undisturbed roots, and this is also true for respiration in disturbed soil samples without roots (Hanson et al., 2000). Currently no universal method applicable for separation of autotrophic and heterotrophic respiration is available, though a number of different indirect methods are available for researchers. They can be applied in situ with no disturbance, or ex-situ with evaluation of possible bias caused by measurements (Gavrichkova et al., 2019; Kuzyakov, 2006). These methods are detailed in the following sections.

Measuring respiration on site

Soil respiratory flux can be measured in-situ utilizing semi-permanent chambers that are connected to an instrument. A common practice is to place a base frame, usually a ring, into the ground with a chamber temporarily topping the frame to form a seal. This has the advantage of minimal disturbance but can result in root cutting and artificially increase the levels of CO₂ measured. The challenge with in-situ respiration is the extremely large variability; environmental factors, such as sunlight, temperature, moisture, and time of day are not controlled, roots of many different plant species may be present and respiring, and smaller organisms may also contribute actively. It must be stressed, that direct in-situ measurements are labor and time consuming and need to be organized correspondingly to account for temporal and spatial variability. Hence, contextualizing laboratory measurement (e.g., basal respiration) to partition the proportion of respiration caused by microbes versus these other sources, can help understand the full implication of in-situ measurements collected under high variability (Ananyeva et al., 2020). Nevertheless, more accurate, high-end equipment (e.g., laser spectroscopy) is now available for field measurements; with proper data management it can help overcome the weaknesses of traditional methods.

Use of isotopes for partitioning of soil respiration

Carbon dioxide emitted by soil originates from several sources (e.g., root and microbial respiration) and usually these sources have different isotopic-C signatures. The most common method to distinguish the origin of CO₂ is to estimate the ¹³C isotopic abundance of the carbon atom in CO₂. The ¹³C enrichment of organic-C originating from the plants with different types of photosynthesis (C3 and C4) can be easily differentiated, and this is used in so-called natural enrichment difference method, when, for example, soil organic matter is formed in long term from the C3-plant residues, whereas currently growing C4-plants respire CO₂ with isotopic signature remarkably different from the soil-derived CO₂. ¹³C isotopic enrichment can differ for CO₂ produced from different soil horizons, or from different organic compounds in soil organic matter. The historical timeline of soil C accumulation can be followed also using ¹⁴C tracer. ¹⁴C produced by past nuclear tests leaves a distinct trace (radioactivity) in the organic matter formed at that time. Finally, growing plants under an atmosphere of ¹³C-enriched CO₂ in the field or lab conditions can help in disclosing the temporal pattern of root and rhizosphere (rhizomicrobial) respiration and partitioning of soil and root derived carbon. This kind of research necessitates additional costs and efforts (so-called free-air carbon enrichment experiments, see e.g., Hungate et al. (2009)).

Ex-situ measurements of respiration

Laboratory measurements are preferred in respiration work due to their reduction of the inherent variability of nature and the possibility of partitioning out the different sources of respiration in a systemic manner. That said, laboratory measurements do not necessarily reflect in-situ conditions. Ex-situ respiration is often at a constant, controlled temperature and moisture, may involve substantial processing (and thus disturbance) of soil, including sieving, rock and plant detritus removal, and often uses volumes of soil that are miniscule (e.g., sometimes an aliquot of small as 1 g soil) compared to the complete, sometimes multi-meter deep soil profile. All of this creates particular challenges for ex-situ measurements, centered around two questions: how do we avoid introducing sources of bias (e.g., by incubating too small a soil volume that introduces respiration artifacts that are not reflective of field conditions), and how do we factor soil sensitivity to altered conditions from in-situ, especially the disturbance inherent to ex-situ measurement? The following sections summarize some of the ex-situ measurements available.

Basal respiration is a measurement of heterotrophic microbial respiration in coarse-sieved, root-free soil under controlled temperature and moisture conditions (but no drying or other processing of soil). If roots are present, they are separated from soil before analysis, and soil is standardized by coarse sieving (e.g., 4 mm sieve). The soil at optimal moisture for microbial activity (40–60% of water holding capacity) is then investigated under controlled temperature conditions. Basal respiration is a powerful tool, but basic techniques have not changed much in the past 100 years: it focuses on optimizing incubation conditions for all soil microorganisms to maximize response. Nevertheless, the temperature might be larger than the annual means in many soils (i.e., 25 °C). Sieving also has the benefit of eliminating rock fragments and other large particles that cause erratic respiration due

to surface-to-area considerations. However, elimination of variability may also result in respiration readings that are less relevant to field conditions (even if they more carefully partition respiration from other sources), and so basal respiration can overestimate in-situ respiration since soil is disturbed moving to the laboratory, and it gives little indication of the maximum potential respiration of the soil, nor can it partition different microbial community respiration contribution. An alternative approach for basal respiration is to incubate for some time (e.g., 20–30 days) while keeping the soil at a steady moisture through gentle rewetting before measuring the samples (e.g., Comeau et al., 2022). This has the advantage of eliminating the effect of the disturbance but takes a vast amount of time and resources committed to allow for such long-term incubations and is therefore difficult to justify commercially. Although used frequently by research scientists, basal respiration is difficult for commercial operations due to the need of the customer to keep the sampled soil moist (and therefore in climate-controlled conditions) to avoid microbial die-off or anaerobiosis. One option that allows farmers to perform basal in-situ (i.e., sieving a volumetric quantity in-field, placing in an airtight container, and measuring it under controlled temperature conditions) is the Solvita paddle (Haney et al., 2008).

Substrate-induced respiration (SIR) aims to determine the microbial response to the addition of a particular substrate, whereas basal respiration gives a gross estimate of overall microbial activity in soil without its stimulation. For SIR, the response of a consortium of organisms serving a particular biological function (e.g., hemicellulose degraders) is tested by the addition of an appropriate easily-available substrate, necessary for the growth. This substrate is expected to support rapid growth by this consortium if present in the soil, and the resultant respiration indicates potential activity of the microbial population responding to this substrate. After soil sieving and optimization of water content (similar to preparation for basal respiration measurement), the samples are incubated for 24–96 h, during which CO₂ concentration (or production rate) is recorded by corresponding equipment (e.g., GC or IRGA). The greatest benefit of SIR is providing quantification of microbial biomass of microorganisms responding to the specific substrate. After calibration, the determination of initial (4–6 h) SIR after glucose amendment was proposed as standard method for the total soil microbial biomass (Anderson and Domsch, 1978) and since then widely used by soil scientists. Measurements of CO₂ production rate over longer periods (up to 48 h) after soil amendment with carbon and nutrient sources can be used for the estimation of microbial growth kinetics and partitioning of dormant and active fractions of microbial populations (Blagodatsky et al., 2000; Panikov and Sizova, 1996).

Burst Respiration relies on the Birch phenomenon noted with some soils, wherein soil drying followed by wetting induces a ‘burst’ of respiration, presumably as inert microbial populations in hibernation reactivate with the reintroduction of water and rapidly consume available carbon to increase their populations (possibly including microbial biomass alongside other labile C). Burst is performed by drying soil between 35 °C and 75 °C (hotter or cooler introduces too much variability), followed by 2 mm sieving or rolling, and then remoistening an aliquot of soil (e.g., 30 cc) to 50% water filled pore-space conditions (Franzluebbers, 1999). The burst of respiration is measured at uniform temperature, usually 20–25 °C, for maximum microbial response. Burst can be read over 1–72 h, though the respiratory fluxes tend to diminish after 24 h. Burst respiration is a measure of the maximum potential of a soil to respire carbon, and therefore is a good way to measure soil sensitivity to unusual (or regular) climate conditions, such as drought-wet cycles, and can serve as a measure of how much C can be easily metabolized and lost under ideal moisture and temperature conditions (Laffely et al., 2020). Its greatest advantage is that it creates far less variable and highly uniform conditions for soil testing, and it can be performed at any time since the soil is dried immediately. Burst is a useful predictor of the susceptibility of soil C stocks to drying-wetting cycles, and it is well suited to rapid laboratory analysis. Its primary flaw is the yet unclear relationship to actual (in situ) respiration, and whether burst can properly characterize soils that never experience dry-wet cycles.

Respiration and climate change

Climate change has the potential to drastically accelerate soil respiration, but with actual absolute changes being ecosystem and management dependent. Annual net plant primary productivity exceeds respiration and continues to do so along with removing 25% of annual fossil fuel emissions. The remaining 75% of emissions, however, accumulates in the atmosphere and drives global climate change. Furthermore, the ability of natural and human-managed ecosystems to function at peak efficiency has decreased in the past century, with global soil degradation resulting in the loss of ~50–75 Gt of soil C to the atmosphere, and the mechanism by which this occurs is accelerated C respiration, resulting from formerly-protected C being exposed to microorganisms through erosion, overzealous tillage, inadequate plant cover, and other poor land-management techniques. Further, climate change itself can act to accelerate soil respiration, especially when temperature-limited (e.g., boreal regions). However, shifting precipitation patterns may create differing effects; too much or too little water could slow or stop respiration in some regions. Increased CO₂ concentrations and warmer temperatures could also increase plant primary productivity, which could cause a negative feedback loop stabilizing the C cycle. However, the greatest risk is that erratic cycles of temperature and moisture could cause much of the ecosystem to be increasingly exposed to wet-dry cycles, which could increase naturally occurring ‘burst’ respiration events, and thus frequent flushes of maximum respiration from soils.

Summary

Soil respiration is intimately related to the C cycle. It is driven by heterotrophic and autotrophic processes and is responsible for cycling C between the terrestrial and aerial environments. Soil respiration is controlled by complex interactions between temperature, hydrology, land management, and biology. Soil C can be divided into different pools based on their origin and availability to microbes; their lability, or lack thereof, drives soil respiration. Soil respiration is strongly affected by human management, primarily via agriculture, but peripherally through human land clearance from natural states, which often causes inevitable losses of respired C to the atmosphere. It is this relationship with humanity that is of particular interest, and our ability to manage global climate change is very closely tied to the global balance of soil respiration and autotrophic productivity. Soil respiration is measured by a variety of in-situ and ex-situ tests, most of which are complementary, not exclusionary, since no single test accurately measures the entirety of soil respiration pools and paths, and each answers different questions.

References

- Ananyeva ND, Sushko SV, Ivashchenko KV, and Vasenev VI (2020) Soil microbial respiration in subtaiga and forest-steppe ecosystems of European Russia: Field and laboratory approaches. *Eurasian Soil Science* 53: 1492–1501.
- Anderson JPE and Domsch KH (1978) A physiological method for the quantitative measurement of microbial biomass in soils. *Soil Biology and Biochemistry* 10: 215–221.
- Blagodatsky SA, Heinemeyer O, and Richter J (2000) Estimating the active and total soil microbial biomass by kinetic respiration analysis. *Biology and Fertility of Soils* 32: 73–81.
- Comeau L-P, MacKinley K, Unc A, and Vallotton J (2022) Ex situ soil respiration assessment using minimally disturbed microcosms and dried–sieved soils; comparison of methods to assess soil health. *Canadian Journal of Soil Science* 103: 143–151.
- Culman SW, Snapp SS, Freeman MA, Schipanski ME, Beniston J, Lal R, Drinkwater LE, Franzluebbers AJ, Glover JD, Grandy AS, Lee J, Six J, Maul JE, Mirsky SB, Spargo JT, and Wander MM (2012) Permanganate oxidizable carbon reflects a processed soil fraction that is sensitive to management. *Soil Science Society of America Journal* 76: 494–504.
- Fan N, Reichstein M, Koirala S, Ahrens B, Mahecha MD, and Carvalhais N (2022) Global apparent temperature sensitivity of terrestrial carbon turnover modulated by hydrometeorological factors. *Nature Geoscience* 15: 989–994.
- Franzluebbers AJ (1999) Microbial activity in response to water-filled pore space of variably eroded southern Piedmont soils. *Applied Soil Ecology* 11: 91–101.
- Friedlingstein P, O'Sullivan M, Jones MW, Andrew RM, Hauck J, Olsen A, Peters GP, Peters W, Pongratz J, Sitch S, Le Quéré C, Canadell JG, Ciais P, Jackson RB, Alin S, Aragão LEOC, Arneeth A, Arora V, Bates NR, Becker M, Benoit-Cattin A, Bittig HC, Bopp L, Bultan S, Chandra N, Chevallier F, Chini LP, Evans W, Florentie L, Forster PM, Gasser T, Gehlen M, Gilfillan D, Gkritzalis T, Gregor L, Gruber N, Harris I, Hartung K, Havard V, Houghton RA, Ilyina T, Jain AK, Joetzier E, Kadono K, Kato E, Kitidis V, Korsbakken JI, Landschützer P, Lefèvre N, Lenton A, Lienert S, Liu Z, Lombardozi D, Marland G, Metzl N, Munro DR, Nabel JEMS, Nakaoka SI, Niwa Y, O'Brien K, Ono T, Palmer PI, Pierrot D, Poulter B, Resplandy L, Robertson E, Rödenbeck C, Schwinger J, Séférian R, Skjelvan I, Smith AJP, Sutton AJ, Tanhua T, Tans PP, Tian H, Tilbrook B, van der Werf G, Vuichard N, Walker AP, Wanninkhof R, Watson AJ, Willis D, Wiltshire AJ, Yuan W, Yue X, and Zaehle S (2020) Global carbon budget 2020. *Earth System Science Data* 12: 3269–3340.
- Gavrichkova O, Evdokimov I, and Valentini R (2019) Comparative study of soil respiration partitioning methods for herbaceous ecosystems. In: Vasenev V, Dovletyarova E, Cheng Z, Prokof'eva TV, Morel JL, and Ananyeva ND (eds.) *Urbanization: Challenge and Opportunity for Soil Functions and Ecosystem Services*, pp. 106–111. Cham: Springer International Publishing.
- Haei M, Öquist MG, Kreyling J, Ilstedt U, and Laudon H (2013) Winter climate controls soil carbon dynamics during summer in boreal forests. *Environmental Research Letters* 8: 024017.
- Haney RL, Brinton WF, and Evans E (2008) Soil CO₂ respiration: Comparison of chemical titration, CO₂ IRGA analysis and the Solvita gel system. *Renewable Agriculture and Food Systems* 23: 171–176.
- Hanson PJ, Edwards NT, Garten CT, and Andrews JA (2000) Separating root and soil microbial contributions to soil respiration: A review of methods and observations. *Biogeochemistry* 48: 115–146.
- Hungate BA, Van Groenigen K-J, Six J, Jastrow JD, Luo Y, De Graaff M-A, Van Kessel C, and Osenberg CW (2009) Assessing the effect of elevated carbon dioxide on soil carbon: A comparison of four meta-analyses. *Global Change Biology* 15: 2020–2034.
- Koutika LS, Dassonville N, Vanderhoeven S, Chapuis-Lardy L, and Meerts P (2008) Relationships between C respiration and fine particulate organic matter (250–50 µm) weight. *European Journal of Soil Biology* 44: 18–21.
- Kuzakov Y (2006) Sources of CO₂ efflux from soil and review of partitioning methods. *Soil Biology and Biochemistry* 38: 425–448.
- Laffely A, Erich MS, and Ohno T (2020) Dissolved organic carbon chemical composition controls the rate of CO₂ release from rewetted soil. *Soil Science Society of America Journal* 84: 483–493.
- Manzoni S, Schimel JP, and Porporato A (2012) Responses of soil microbial communities to water stress: results from a meta-analysis. *Ecology* 93: 930–938.
- Moyano FE, Vasilyeva N, and Menichetti L (2018) Diffusion limitations and Michaelis–Menten kinetics as drivers of combined temperature and moisture effects on carbon fluxes of mineral soils. *Biogeosciences* 15: 5031–5045.
- Munkhjargal M, Yadamsuren G, Yamkhin J, and Menzel L (2020) Ground surface temperature variability and permafrost distribution over mountainous terrain in northern Mongolia. *Arctic, Antarctic, and Alpine Research* 52: 13–26.
- Panikov NS and Sizova MV (1996) A kinetic method for estimating the biomass of microbial functional groups in soil. *Journal of Microbiological Methods* 24: 219–230.
- Rath KM, Maheshwari A, Bengtson P, and Rousk J (2016) Comparative toxicities of salts on microbial processes in soil. *Applied and Environmental Microbiology* 82: 2012–2020.
- Reichstein M, Rey A, Freibauer A, Tenhunen J, Valentini R, Banza J, Casals P, Cheng Y, Grünzweig JM, Irvine J, Joffre R, Law BE, Loustau D, Miglietta F, Oechel W, Ourcival J-M, Pereira JS, Peressotti A, Ponti F, Qi Y, Rambal S, Rayment M, Romanya J, Rossi F, Tedeschi V, Tirone G, Xu M, and Yakir D (2003) Modeling temporal and large-scale spatial variability of soil respiration from soil water availability, temperature and vegetation productivity indices. *Global Biogeochemical Cycles* 17.
- Sinsabaugh RL, Hill BH, and Follstad Shah JJ (2009) Ecoenzymatic stoichiometry of microbial organic nutrient acquisition in soil and sediment. *Nature* 462: 795–798.
- Smith KA, Ball T, Conen F, Dobbie KE, Massheder J, and Rey A (2003) Exchange of greenhouse gases between soil and atmosphere: Interactions of soil physical factors and biological processes. *European Journal of Soil Science* 54: 779–791.

Anaerobic soil processes

PW Inglett^a, D Said-Pullicino^b, and K Ramesh Reddy^a, ^aUniversity of Florida, Soil and Water Sciences Department, Wetland Biogeochemistry Laboratory, Gainesville, FL, United States; ^bUniversity of Torino, Department of Agricultural, Forest and Food Sciences, Torino, Italy

© 2023 Elsevier Ltd. All rights reserved.

Introduction	379
Physical characteristics	380
Biological characteristics	380
Chemical characteristics	381
Biogeochemical processes	382
Agronomic, ecological and environmental significance	385
References	386

Abstract

Anaerobic soils are characteristic of a number of different wetland environments including bogs, fens, swamps, marshes, floodplains, coastal marshes and mangroves, and paddy rice fields. They represent a critical feature of the global landscape because of their unique characteristics and role in regulating global biogeochemical cycles. The productivity and functioning of complex wetland ecosystems is greatly determined by the physical, chemical and biological processes of anaerobic soils. The oxygen-depleted environments that characterize these soils lead to a systematic modification of the microbial community structure and related processes, with important implications on the agronomical, ecological and environmental functions of these soils.

Key points

- Anaerobic soils occupy an important niche in the biosphere, and their importance in wetlands and paddy rice soils is widely recognized. They occur when or where oxygen consumption by soil biota exceeds oxygen supply via diffusion into the soil matrix.
- Saturated soil conditions support microbial populations adapted to anaerobic environments, largely facultative and obligate anaerobic bacteria and archaea.
- Fluctuations in available oxygen and other electron acceptors have significance for regulating biogeochemical cycles of nutrients in anaerobic soils. The redox potential (E_h) of a system may be seen as a characteristic or predictor of the prevailing biogeochemical processes.
- Microbial metabolism including both reduction and oxidation processes directly impacts the forms and fate of nitrogen, iron, manganese, sulfur, and carbon in anaerobic soil systems.
- Wetlands are productive ecosystems and store 20–30% of the global SOC pool (estimated 650 Pg carbon). They are valued for their function as sinks, sources, transformers of nutrients, but are also a significant source of greenhouse gas emissions including both nitrous oxide and methane.

Introduction

Anaerobic soils occur when or where oxygen consumption by soil biota exceeds oxygen supply via diffusion into the soil matrix. This condition is also termed ‘soil anaerobiosis’ and results in a predominantly oxygen-depleted environment in the soil profile. The creation of anaerobic soils is generally linked with water saturation or flooding, though anaerobic conditions may also occur in soils amended with heavy rates of organic materials, animal wastes, biosolids, and composts, as well as soils treated with ammoniacal fertilizers. Once a soil becomes water saturated, the supply of oxygen is immediately reduced due to the displacement of oxygen contained in the available pore space. Following consumption of the relatively small amount of available oxygen in the pore water, oxygen can only be supplied to respiring organisms through the process of diffusion from the nearest aerobic zone. This process is comparatively slow under saturated soil conditions as oxygen diffusion in water is approximately four orders of magnitude slower than through air. Under these conditions, even moderate rates of soil or root respiration can quickly deplete available oxygen and result in anaerobic soil conditions.

Anaerobic soils are characteristic to a number of different wetland environments including bogs, fens, swamps, marshes, floodplains, coastal marshes and mangroves, and paddy rice fields. Wetland soils are widely distributed throughout the world and can be found in all climates, ranging from the tropics to tundra, except for Antarctica. Approximately 6% of the Earth’s land surface, which equals approximately 800 million hectares, is covered by wetlands. They represent a critical feature of the global

landscape because of their unique role in regulating global biogeochemical cycles. In these landscapes, anaerobic conditions are strongly linked to their hydrology and may arise due to inundation (e.g., intertidal zones) or saturation by surface or ground water at a frequency and duration sufficient to strongly limit the diffusion of oxygen into the soil. In any given landscape, wetlands are generally located in areas with a low elevation and high water table.

Anaerobic soils are generally associated with poorly drained, heavy textured soils that retain water during rainy periods or after irrigation events, or with soils that have a superficial water table. However, even in well-drained, seemingly aerobic soils, limited oxygen diffusion into structural units such as peds or aggregates may lead to the establishment of anaerobic microsites that are more ubiquitous than previously assumed and have important implications on element cycling in soils. In upland environments, anaerobic conditions may however be temporary and may not last more than a few days, while in wetland environments, soil anaerobic conditions may last for several months. The functioning of these complex wetland ecosystems is greatly determined by the physical, chemical and biological characteristics of anaerobic soils. Since most of our knowledge on anaerobic soils has been gained through research in wetlands and paddy rice fields, this chapter will largely focus on the morphological and biogeochemical features of these types of soils (Kögel-Knabner et al., 2010).

Physical characteristics

Soil volume primarily comprises solid matter, water, and air, with upland mineral soils generally consisting of about 50% by volume of solids, 25% of water, and 25% of air. Variations in soil moisture contents influence the air-filled pore space of soils, which are extremely important for oxygen diffusion into the soil from the atmosphere. In water-saturated mineral soils, approximately 50% of the soil volume is occupied by solid constituents, while the remaining 50% is occupied by water. In wetland organic soils, a large proportion (up to 80%) of soil volume is occupied by water, with SOM and mineral matter occupying less than 20%. Since gaseous exchange with the atmosphere in upland soils is not restricted because of the connectivity of air spaces, the gaseous composition of soil pores in about 10–21% oxygen, 0.03–1% carbon dioxide and trace amounts of nitrous oxide and ammonia. However, an increase in water-filled pore space limits gaseous exchange with the atmosphere, which is greatly reduced (ca 10,000 times slower) and the gaseous composition of air spaces becomes depleted in oxygen and dominated by reduced chemical forms that are regulated by associated biogeochemical processes (Ponnamperuma, 1972) (Fig. 1). In recently flooded soils, nitrous oxide (N_2O) can be present as a result of denitrification of nitrate nitrogen (NO_3^- -N). In moderately reduced soils, H_2S can be observed, followed by methane (CH_4) under more reducing conditions. In highly reduced soils, ethylene (C_2H_4) and phosphine (PH_3) can even be observed.

Soil structure can also be influenced by water saturation. For example, rapid water saturation can physically break down aggregates due to slaking when air inside aggregates is compressed or swelling occurs in the presence of expandable 2:1 clay minerals. Similarly, anaerobic soil conditions may also enhance particle dispersion and aggregate breakdown via chemical mechanisms, such as reduction of structural Fe, the reductive dissolution of Fe (hydr)oxide coatings, or an increase in soil pH that may enhance surface charge and repulsive forces between soil particles. Moreover, reducing conditions may promote the dissolution of organic matter and Fe (hydr)oxides that serve as strong aggregate binding agents.

Biological characteristics

Saturated soil conditions support microbial populations adapted to anaerobic environments (Liesack et al., 2000). After the soil becomes saturated, most of the aerobic organisms become quiescent or die, and new groups, largely facultative (organisms which can function under both aerobic and anaerobic environments) and obligate anaerobic bacteria and archaea, take over. Aerobic microbial populations are then restricted to zones where O₂ is available. Fungi, which are active in upland environment, are also a significant component of the decomposer community in wetlands (e.g., surface soils), but are generally inhibited, primarily due to

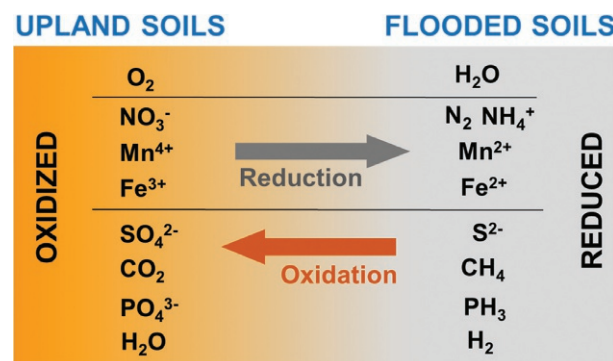


Fig. 1 Various inorganic oxidized and reduced compounds in upland and wetland soils.

absence of O_2 and alteration in soil pH (acid to neutral) under anaerobic conditions. Despite this inhibition, select fungi remain active under low oxygen levels and have an important role in anaerobic processes such as fermentation.

At the soil–floodwater interface, saturated soils can support the growth of microphytic communities, including a variety of planktonic, epiphytic, and benthic algae and fungi. Many of the species in these microbial assemblages have the capacity to carry out photosynthesis. Diel fluctuations in dissolved O_2 produced as a consequence of photosynthesis often increase the O_2 levels in the floodwater beyond saturation levels during daytime which then decline to low levels during night-time. These large fluctuations in available oxygen have special significance in wetlands for regulating biogeochemical cycles of nutrients.

Wetland or anaerobic soil conditions also support the presence of hydrophytic vegetation, or plants that are adapted to the reducing wetland environment. These plants include almost 2000 species that are only found under wetland or flooded soil conditions. Hydrophytic plants have unique characteristics to adapt to oxygen-deficient conditions, including physiological adaptations (such as capability to respire anaerobically), anatomical adaptations (such as development of intercellular air spaces), and morphological adaptations (such as water roots and adventitious roots). Some plants can facilitate (passively or actively) O_2 transport into their roots, some of which may leak out creating aerobic rhizosphere zones in the soil. With these adaptations, hydrophytic plants are able to survive under reducing conditions considered toxic to other macrophytes. In many cases, adapted plant communities become the dominant source of organic matter in wetland systems.

The presence of aerobic and anaerobic zones affects the composition and distribution of organisms in flooded soils, with different compartments (e.g. the oxic surface soil, the rhizosphere, the anoxic bulk soil) being characterized by different physiochemical conditions leading to microscale chemical gradients that control the activity and distribution of functional groups of organisms. These zones consist of metabolically related organisms, e.g. oxygen-respiring bacteria and archaea, nitrate reducers, iron reducers, sulfate reducers, as well as fermenting bacteria and fungi, and methanogenic archaea. The concerted action of all functional groups of microorganisms drive element cycling in anaerobic soils.

Besides the spatial distribution there is also a temporal development of organism groups driven by the transition of soil conditions from aerobic to anaerobic metabolism. The shift from aerobic to anaerobic respiration causes an overall decrease in microbial biomass and activity under saturated soils conditions, and the rates of many microbially-mediated reactions (e.g., respiration) decline, and some reactions may be eliminated and replaced by new ones. Dynamic environmental conditions are also known to yield microbial communities with high functional diversity and species richness because they host periodically variant conditions for various metabolisms.

Chemical characteristics

When O_2 availability becomes limited, microorganisms must utilize other compounds as electron acceptors to maintain their metabolism. These compounds, many of which are nutrient elements, can exist in both dissolved and solid phases, and include oxidized forms of elements such as N, Fe, Mn and S. As they are utilized during respiration, these elements gain electrons and thus become chemically reduced. The result of microbial metabolism therefore is the conversion of oxidized elements to the corresponding reduced form under anaerobic conditions. When wetland soils are drained, many of the reduced compounds are re-oxidized either by chemical or biochemical reactions. Therefore, in upland or drained soils, oxidized forms of chemical species dominate, while reduced forms dominate in wet soil systems (Fig. 1).

Reduction–oxidation, or redox potential (E_h), reflects the free energy available for reduction (intensity of reduction) or a measure of electron (e^-) activity (Reddy et al., 2022; Zhang and Furman, 2021). The E_h of a system may be seen as a characteristic or predictor of the prevailing biogeochemical processes. Depending on soil characteristics, E_h generally decreases with time and approaches a steady value after flooding. Redox potential is the most common parameter used to measure degree of soil water saturation or intensity of soil anaerobic conditions. The range of E_h values typically observed in soils is from +700 to –300 mV (Fig. 2). Negative values represent high electron activity and intense anaerobic conditions typical of permanently waterlogged soils, whereas positive values represent low electron activity and aerobic to moderately anaerobic conditions typical of wetlands in transition zone. Soils with $E_h > 300$ mV are considered aerobic or upland, where oxygen is used as the dominant electron acceptor. Soils with $E_h < 300$ mV are considered anaerobic or wetland.

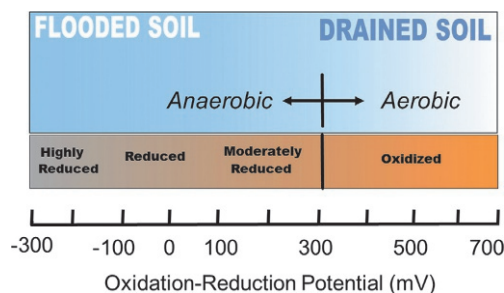


Fig. 2 Schematic showing relationship between oxidation–reduction potential and oxidized and reduced conditions in wetland and flooded soils.

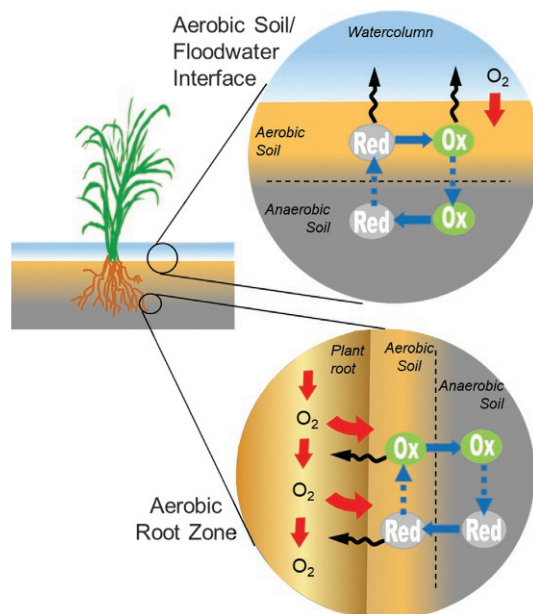


Fig. 3 Diagram of aerobic and anaerobic zones and processes of spatial coupling of reduction and oxidation affecting movement of elements in anaerobic soils.

The chemical nature of the reduction process also affects soil pH, electrical conductivity, cation exchange capacity, and sorption and desorption processes. In general, the decrease in soil redox potential under saturated soil conditions is accompanied by an increase in electrical conductivity and ionic strength. The pH of most soils tends to approach neutrality under flooded conditions, with acid soils increasing and alkaline soils decreasing in pH. The increase in pH of acidic soils depends on the activities of oxidants (such as NO_3^- , Fe^{3+} , Mn^{4+} , and SO_4^{2-}) and proton consumption released during reduction of these oxidants under flooded conditions. In alkaline soils, pH is controlled (and generally lowered) by the accumulation of dissolved CO_2 and organic acids.

Accumulation of reduced compounds in anaerobic soils results in the establishment of sharp concentration gradients across aerobic–anaerobic interfaces (e.g. in the rooting zone). The concentration of reduced compounds is usually higher in the anaerobic zones, which results in diffusion toward aerobic areas where they are oxidized (Fig. 3). Similarly, some of the dissolved, oxidized compounds diffuse toward anaerobic areas where they will be reduced. The exchange rates between aerobic and anaerobic zones will determine whether the wetlands soils or sediments are functioning as a sink or source for nutrients. The rate of exchange of dissolved species depends upon: (1) concentration gradient of dissolved species in soil pore water between zones; (2) soil type and other related physicochemical properties (porosity, pH, cation exchange capacity, SOM content); and (3) kinetics of related redox-driven biogeochemical processes that remove or add species to the soil solution.

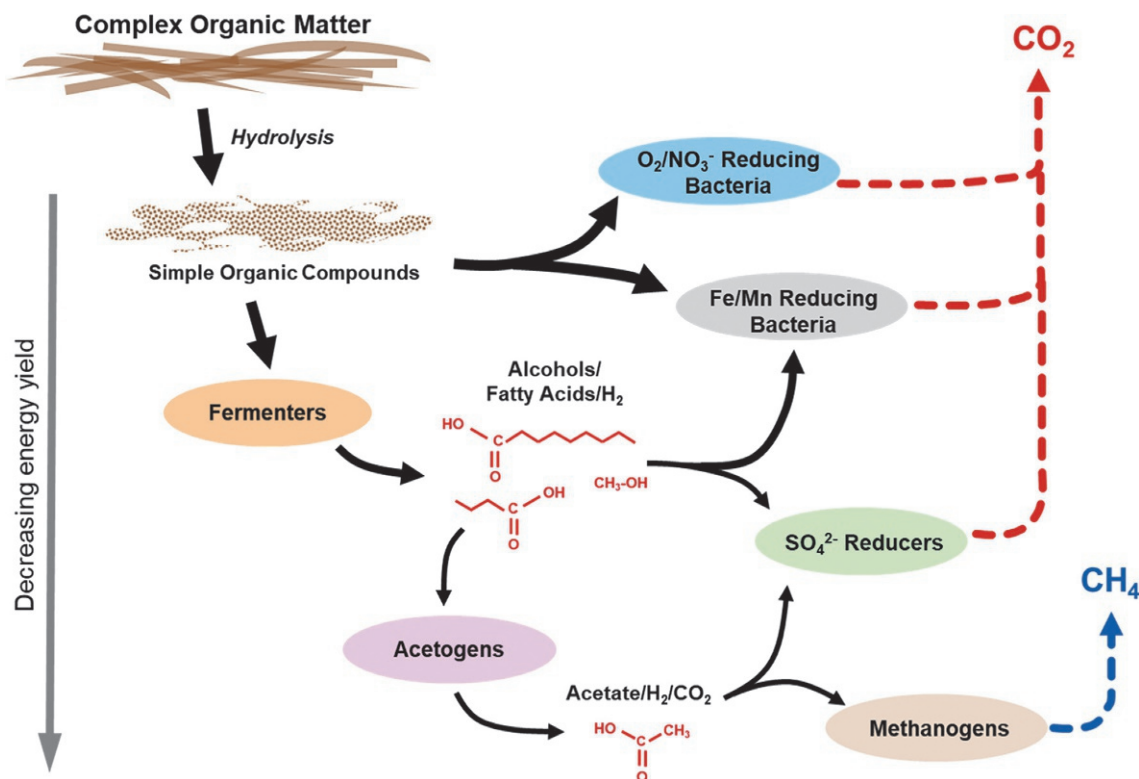
Biogeochemical processes

The biogeochemical cycling of elements in anaerobic soils are generally very complex and highly dynamic as they (1) depend on the establishment of concentration gradients of oxidized or reduced forms across aerobic–anaerobic interfaces, (2) are influenced by mobilization processes linked to diffusion or mass flow of dissolved species across these gradients or with the predominant water fluxes, as well as losses of gaseous species to the atmosphere, (3) often involve various redox-driven interactions through both abiotic, e.g. adsorption, (co)precipitation, and biotic (i.e. microbial respiration) processes (Fig. 3) (Kirk, 2004; Reddy et al., 2022).

Microbial communities in anaerobic soils play a key role in regulating a number of essential biogeochemical cycles such as carbon, nitrogen, phosphorus, sulfur and metals. Organic matter released by primary producers is degraded by microorganisms to supply carbon, energy and nutrients for growth. Enzymatic hydrolysis of SOM produces various monomers, including glucose, xylose, fatty acids, and amino acids. These simple reduced compounds then serve as substrates for microbial metabolism as chemical energy is released during their oxidation. Bioavailable organic matter therefore contributes to the reduction capacity of a soil as an electron reservoir, however this is also largely modulated by the chemical composition of the SOM available. If present, oxygen provides an electron acceptor (oxidant) for supporting microbial oxidation of reduced carbon compounds (respiration). Thermodynamically, oxygen is the most preferred electron acceptor by microorganisms (Table 1), and as a result, aerobic microorganisms maintain a competitive advantage while oxygen is present. Oxygen is used preferentially because it receives electrons from the reductant material (organic matter) more readily than do other oxidants. The greater energy yield during the aerobic process is due to: (1) complete oxidation of carbon atoms in the organic substrates to CO_2 , and (2) the oxygen redox couple having a relatively high, positive reduction potential. This leads to a large net difference in electrical potentials between electron donor (organic substrate) and terminal electron acceptor (oxygen).

Table 1 Microbial groups involved in various redox reactions in wetland soils.

Redox potential (mV)	Electron acceptor	Decomposition end products	Microbial groups
Aerobic (>300 mV)	O ₂	CO ₂ , H ₂ O	Aerobic fungi and bacteria
Fermenting (less than -100 to +300 mV)	Organics	Organic acids, CO ₂ , H ₂ , alcohols, amino acids	Fermenting bacteria and fungi
Facultative anaerobic (+100 to +300 mV)	NO ₃ ⁻	N ₂ O, N ₂ , CO ₂ , H ₂ O	Denitrifiers
	Mn ⁴⁺	Mn ²⁺ , CO ₂ , H ₂ O	Mn(IV) reducers
	Fe ³⁺	Fe ²⁺ , CO ₂ , H ₂ O	Fe(III) reducers
Obligate anaerobic (less than -100 mV)	SO ₄ ²⁻	HS ⁻ , CO ₂ , H ₂ O	Sulfate reducers
	CO ₂ and acetate	CH ₄ , CO ₂ , H ₂ O	Methanogens
	Organic acids	Acetate, CO ₂ , H ₂	H ₂ -producing bacteria

**Fig. 4** Pathways involved in organic matter decomposition in anaerobic wetland soils. Adapted from Reddy RK, DeLaune RD, and Inglett PW (2022) Biogeochemistry of Wetlands: Science and Applications. 2nd edn. Boca Raton, FL, USA: CRC Press.

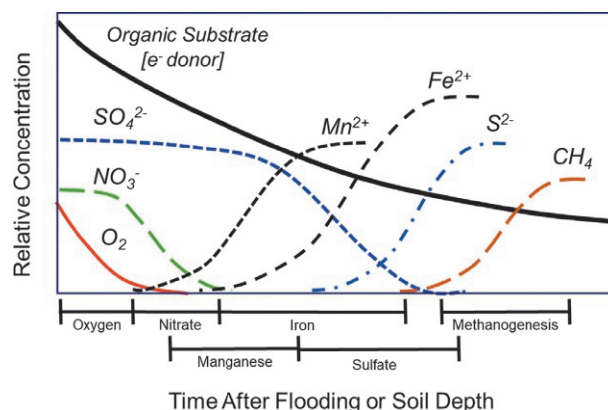
Once oxygen is depleted in the soil, bacteria must respire anaerobically. Anaerobic decomposition of organic matter is more complex and less energetically favorable than aerobic decomposition. It is mediated by a complex group of physiologically different microorganisms which participate in the decomposition pathways (Fig. 4). Often, the end product of one metabolic pathway is substrate for the next until the decomposition is complete. Initially, fermenting bacteria mediate the extracellular hydrolysis of high-molecular-weight polymers (i.e., proteins, polysaccharides, lipids, and nucleic acids) and transform their respective monomers (e.g., amino acids, sugars, fatty acids, and nucleotides) to CO₂, H₂, acetate, propionate, butyrate, and other fatty acids and alcohols.

In the absence of oxygen, the microbial use of alternative electron acceptors proceeds in a sequential manner dependent on their electron affinity, energy yield, and related enzyme systems in the bacteria. In this manner, the thermodynamic order of electron acceptor use is NO₃⁻, Mn⁴⁺, Fe³⁺, SO₄²⁻ and finally CO₂ (Table 2). Anaerobic respiration is typically reflected in temporal variations in the concentration of reduced/oxidized species of the different redox couples involved, as the most favorable electron acceptors are utilized first (Fig. 5). However, the exclusion of less-favorable respiration pathways is not complete, resulting in considerable overlap between the different forms of organic matter mineralization. The rate at which electron acceptors are consumed in soil systems depends on their concentration, the availability of organic compounds, and the activity of the microbial population involved in the reductive processes.

Nitrate reduction can occur under hypoxic to anoxic conditions through two major pathways: (1) Denitrification, the progressive microbial reduction of nitrate or nitrite to N oxides and molecular N₂ (NO₃⁻ → NO₂⁻ → NO → N₂O → N₂); (2) Dissimilatory

Table 2 Summary reactions for microbial respiration pathways in anaerobic soils.

Electron acceptor	Reaction coupled to glucose oxidation	ΔG° (kJ mol ⁻¹)
O ₂	C ₆ H ₁₂ O ₆ + 6O ₂ → 6CO ₂ + 6H ₂ O	-2879
NO ₃ ⁻	5C ₆ H ₁₂ O ₆ + 24NO ₃ ⁻ + 24H ⁺ → 30CO ₂ + 12 N ₂ + 42H ₂ O	-2713
MnO ₂	C ₆ H ₁₂ O ₆ + 12MnO ₂ + 24H ⁺ → 6CO ₂ + 12Mn ²⁺ + 18H ₂ O	-1916
Fe(OH) ₃	C ₆ H ₁₂ O ₆ + 24Fe(OH) ₃ + 48H ⁺ → 6CO ₂ + 24Fe ²⁺ + 66H ₂ O	-418
SO ₄ ²⁻	C ₆ H ₁₂ O ₆ + 4H ₂ O → 2CH ₃ COO ⁻ + 2HCO ₃ ⁻ + 4H ₂ + 4H ⁺	-207
	CH ₃ COO ⁻ + SO ₄ ²⁻ + 3H ⁺ → 2CO ₂ + H ₂ S + 2H ₂ O	-63
CO ₂	C ₆ H ₁₂ O ₆ + 4H ₂ O → 2CH ₃ COO ⁻ + 2HCO ₃ ⁻ + 4H ₂ + 4H ⁺	-207
	2CH ₃ COO ⁻ + 2H ₂ O → 2CH ₄ + 2HCO ₃ ⁻	-31

**Fig. 5** Sequential reduction of oxidants (oxidized compounds) and accumulation of reductants (reduced compounds) in wetland soils. Adapted from Reddy RK, DeLaune RD, and Inglett PW (2022) Biogeochemistry of Wetlands: Science and Applications. 2nd edn. Boca Raton, FL, USA: CRC Press.

nitrate reduction to ammonia (DNRA), the reduction of nitrate to ammonium via nitrite ($\text{NO}_3^- \rightarrow \text{NO}_2^- \rightarrow \text{NH}_4^+$) (Reddy et al., 2022). Denitrifiers are a diverse group of heterotrophic bacteria (most of them facultative anaerobic), archaea, and fungi that couple the oxidation of substrates to the reduction of NO_3^- to either N_2O or N_2 . The process can be either organotrophic (organic electron donors including methane) or lithotrophic (inorganic electron donors such as Fe^{2+} or sulfide). This reaction occurs in moderately reduced conditions in the absence of oxygen and is one of the dominant mechanisms responsible for the loss of N from wetland soils. The rate of denitrification is largely determined by the concentration of NO_3^- as the electron acceptor. Dissimilatory reduction of NO_3^- is carried out by a diverse group of heterotrophic and facultative anaerobic chemolithoautotrophic bacteria. In anaerobic conditions, NO_3^- is first reduced to NO_2^- and then directly to NH_4^+ without being assimilated into microbial tissues.

Denitrification and DNRA are catalyzed by different strains of bacteria and Archaea, and their relative contribution to the overall nitrate reduction is often attributed to the availability of organic matter and the presence of sulfide, which has been shown to inhibit denitrification and stimulate DNRA. More recently, a variety of other reduced compounds have been shown capable of catalyzing NO_3^- reduction including Mn^{2+} , Fe^{2+} , and CH_4 . Other processes similar to denitrification have been identified. Among these are anaerobic ammonium oxidation (Anammox) which, as its name implies, involves NH_4^+ oxidation and the simultaneous reduction of NO_2^- to final product N_2 . Bacteria of the order Brocadiales including the genera Kuenenia, Brocadia, Jettenia, Candidatus Anammoxoglobus and Scalindua have been shown to possess hydrazine oxidoreductase, the enzyme required for Anammox. Measured rates have been reported in both estuarine and freshwater wetland soils, but N_2 production by Anammox is generally much lower than competing denitrification, thus its importance relative to other N cycling reactions is not well known.

The reduction of Fe and Mn in many soils is possibly one of the main processes of soil formation, and the relatively ubiquitous presence of Fe and Mn oxides in soils and sediments makes respiration by this pathway a major contributor of organic matter mineralization (Kappler et al., 2021). A large variety of microorganisms are capable of reducing Fe, including fungi. Fe^{3+} -reducing bacteria, including Geobacter and Shewanella species, couple the reduction of ferric iron with the oxidation of organic (fatty acids, carbohydrates, amino acids, aromatic compounds) or inorganic (dihydrogen, elemental sulfur, or sulfide) electron donors. When Fe^{3+} is reduced, Fe^{2+} is the reduced end product. Like Fe, Mn^{2+} is generally accepted as the end product of Mn^{4+} reduction though Mn^{3+} may also be encountered as an intermediate species. Fe^{3+} reducers can use complexed dissolved Fe^{3+} as well as short-range-ordered (e.g. ferrihydrite) or crystalline (e.g. goethite, hematite and magnetite) mineral phases as an electron acceptor, though the energy gained by such electron transfer varies depending on the mineral.

The redox potentials of diverse Fe(II)-Fe^{3+} redox couples lie between those of oxidized and reduced carbon, nitrogen, oxygen and sulfur redox species. Changes in solubility caused by Fe redox transformations also indirectly influence the mobility of other

elements with implications for the fate of carbon, nutrients (e.g., phosphorus) and contaminants (e.g., arsenic and cadmium). For example, Fe^{3+} readily complexes with available phosphate to lessen its availability. Upon reduction of Fe^{3+} to Fe^{2+} , however, phosphorus becomes more soluble leading to increased flux into the water column. More recently, oxidized forms of Fe^{3+} and Mn^{4+} were also shown to oxidize ammonium anaerobically in the processes Feammox and Mnammo, respectively (Reddy et al., 2022).

Sulfate reduction and oxidation processes are very important in wetlands affected by marine inputs of sulfur. Sulfate reducers are obligate anaerobes (bacteria and archaea) that couple the oxidation of a variety of organic substrates to CO_2 with the chemical reduction of the terminal electron acceptor SO_4^{2-} to sulfides ($-\text{S}^{2-}$) (Barton and Fauque, 2009). The strains isolated from rice paddy fields are mostly affiliated with the genera *Desulfovibrio* and *Desulfotomaculum*. Sulfate-reducing organisms do not synthesize enzymes to hydrolyze polymers such as polysaccharides. Also, many groups of sulfate-reducers cannot use monomers such as monosaccharides (e.g., glucose) as substrates for energy. Thus, sulfate reducers are mainly dependent on fermenters to produce simple organic compounds (e.g., acetate, propionate).

Sulfate reduction and oxidation frequently involves compounds and processes which are toxic to most organisms. For example, the product of sulfate reduction, hydrogen sulfide, is extremely toxic to aerobic organisms because it reacts with the heavy metal groups of the cytochrome system. Sulfate reducers are also responsible for converting mercury into the bioavailable methylated-Hg form. Hydrogen sulfide is very reactive with metals and usually results in the precipitation of metallic sulfides (e.g., FeS , ZnS , HgS), so in high sulfide areas toxic metals may be completely bound (and Hg methylation prevented). However, when these soils or sediments are exposed to oxygen (e.g., when dredged or coastal soils are drained) the sulfides are oxidized chemically and biologically (e.g., by *Thiobacillus ferrooxidans*) releasing the bound metals and lowering the pH through the formation of sulfuric acid $\text{H}_2\text{S} + 2\text{O}_2 \rightarrow \text{H}_2\text{SO}_4$.

Methanogens and Methanotrophs: The terminal step in the anaerobic degradation of organic macromolecules, in the absence of all other electron acceptors, is the conversion of acetate and H_2/CO_2 to methane (Conrad, 2007). Methanogens, typically found in the archaeal families of the *Methanobacteriales*, *Methanomicrobiales* and *Methanosarcinales*, are obligate anaerobes that can produce methane through the conversion of acetate to CH_4 and CO_2 (acetoclastic pathway) or through the conversion of H_2 and CO_2 to CH_4 (hydrogenotrophic pathway). Like sulfate reducers, methanogens cannot directly utilize large-molecular-weight polymers; so methanogens must depend on at least three groups of microbes, including hydrolytic, fermentative, and H_2 -producing acetogenic bacteria.

Sulfate reducers also consume acetate and H_2 and are direct competitors for substrates with methane. Thus, in areas of sufficient sulfate availability, methanogenesis can be completely inhibited. Once produced, methane can find its way to the atmosphere, but not all the CH_4 produced by methanogens is emitted into the atmosphere. Approximately half of the CH_4 is oxidized by methanotrophic microorganisms that live at the interface between anoxic and oxic zones. These organisms can be aerobic or anaerobic coupling the oxidation of CH_4 with the reduction of nitrate, Fe^{3+} , or sulfate compounds.

Agronomic, ecological and environmental significance

Anaerobic soils occupy an important niche in the biosphere, and their importance in wetlands and paddy rice soils is widely recognized by scientists, environmental managers, and policy-makers. Agronomically, anaerobic soils are widely used throughout the world for paddy rice production. Anaerobic wetland soils have unique biogeochemical properties and are primary drivers of the natural ecosystem function, as many of the processes have important feedback to ecosystem productivity and function. These soils are primary nutrient sources to plants grown in the paddy soils or wetlands. The decomposition process described here results in production of bioavailable nitrogen and phosphorus, which supports the productivity of plants. Furthermore, the extent of Fe^{3+} or Mn^{4+} reduction, or the presence of sulfide can strongly influence the distribution of toxic trace metals and availability of P.

Environmentally, anaerobic soils may have both positive and negative attributes. One negative aspect is that wetlands are one of the primary sources of greenhouse gas emissions including both nitrous oxide and methane. Nitrous oxide is produced during anaerobic nitrogen transformations such as nitrate reduction. Nitrogen loaded wetlands are therefore an important source of N_2O . Approximately 25% of methane emitted to the atmosphere is derived from wetlands (about $115\text{--}227 \text{ Tg CH}_4 \text{ year}^{-1}$). On a molecular basis and a time frame of 100 years, the global warming potential of CH_4 is about 25–40 times stronger than that of CO_2 . Disturbance and nutrient inputs can shift the balance of gaseous wetland C emissions toward CH_4 , therefore even though they accumulate organic matter, younger restored systems are more prone to have positive global warming potentials than older less impacted wetlands which emit a majority of CO_2 .

On the other hand, wetlands are very productive ecosystems and despite being only about 5–8% of the terrestrial landscape, serve as important terrestrial sinks of atmospheric CO_2 storing around 20–30% of the global SOC pool (around 650 Pg carbon). Natural wetlands, especially peatlands, generally accumulate C as a result of a greater rate of inputs (organic matter produced in situ and ex situ) than outputs (mainly slow decomposition of recalcitrant compounds under waterlogged conditions). There is growing evidence suggesting that organic matter turnover and stabilization in redoximorphic soils is also strongly linked to the dynamic interactions with redox-active minerals, in particular Fe oxyhydroxides. In northern peatlands wetland soil C is stabilized by cold temperatures and saturated conditions. Thus, climate change, and in particular warming, has a significant effect on boreal C storage where it both increases peat decomposition as well as enhancing releases of CH_4 from decomposition and thawing of frozen soils.

Anaerobic soils in wetlands are also highly valued for their function as sinks, sources, transformers of nutrients and contaminants, and their role in improving water quality. Wetlands and their significant storages of organic matter contribute significantly to

the export of necessary of secondary production for productivity of associated systems (e.g., fisheries). Wetland abundance and hydrological connectivity in watersheds is also a major contributor to pollutant and nutrient load reductions. This function of anaerobic soils has resulted in anaerobic soils or zones being incorporated in sustainable landscape designs (e.g., stormwater management) as well as low-cost constructed wetland technologies for water treatment. At present thousands of such wetlands are in operation throughout the world including a treatment wetland (360 ha) in arid climate at the Nimr Water Treatment Plant in Oman and the world's largest treatment wetland known as the Everglades Stormwater Treatment Areas system (Florida, USA) with more than 23,000 ha of total wetland area.

References

- Barton LL and Fauque GD (2009) Biochemistry, physiology and biotechnology of sulfate-reducing bacteria. *Advances in Applied Microbiology* 68: 41–98.
- Conrad R (2007) Microbial ecology of methanogens and methanotrophs. *Advances in Agronomy* 96: 1–63.
- Kappler A, Bryce C, Mansor M, Lueder U, Byrne JM, and Swanner ED (2021) An evolving view on biogeochemical cycling of iron. *Nature Reviews Microbiology* 19: 360–374.
- Kirk G (2004) *The biogeochemistry of submerged soils*. Chichester, UK: John Wiley and Sons Ltd.
- Kögel-Knabner I, Amelung W, Cao Z, Fiedler S, Frenzel P, Jahn R, Kalbitz K, Kölbl A, and Schloter M (2010) Biogeochemistry of paddy soils. *Geoderma* 157: 1–14.
- Liesack W, Schnell S, and Revsbech NP (2000) Microbiology of flooded rice paddies. *FEMS Microbiology Reviews* 24: 625–645.
- Ponnamperuma FN (1972) The chemistry of submerged soils. *Advances in Agronomy* 24: 29–96.
- Reddy RK, DeLaune RD, and Inglett PW (2022) *Biogeochemistry of Wetlands: Science and Applications*, 2nd edn Boca Raton, FL, USA: CRC Press.
- Zhang Z and Furman A (2021) Soil redox dynamics under dynamic hydrologic regimes – A review. *Science of the Total Environment* 763: 143026.

Mineral-organic-microbial interactions

Balwant Singh^a, Sabina Yeasmin^b, and Donald L. Sparks^c, ^aSydney Institute of Agriculture, School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW, Australia; ^bDepartment of Agronomy, Bangladesh Agricultural University, Mymensingh, Bangladesh; ^cDepartment of Plant and Soil Sciences, University of Delaware, Newark, DE, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of P.M. Huang, M.C. Wang, M.K. Wang, MINERAL–ORGANIC–MICROBIAL INTERACTIONS, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 486–499, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00220-4>.

Introduction	388
Processes of organo-mineral interactions	388
Ligand exchange or inner-sphere complexation	388
Ion exchange	388
Polyvalent cation or water bridges	388
Van der Waals forces	390
Hydrogen bonding	390
Hydrophobic interactions	390
Intercalation	390
Co-precipitation	390
Physical protection of SOM	390
Mineral-organic matter-microbiota interactions	391
Abiotic and biotic controls on soil organo-mineral interactions	393
Organic matter dynamics	393
Mineral phase properties	394
Soil solution chemistry	395
Insights into organo-mineral associations	396
Isolation of mineral associated organic matter	397
Physical procedures	397
Chemical procedures	397
Combination of physical and chemical procedures	398
Characterization of mineral associated organic matter	399
NMR spectroscopy	399
FTIR spectroscopy	400
X-ray photoelectron spectroscopy	400
Natural carbon isotope	400
Radiocarbon dating	400
Micro-imaging	401
Synchrotron-based techniques	401
Organo-mineral interactions and terrestrial environment	403
Soil ecosystem services	403
Soil structural stability	403
Biochemical cycling of nutrients in soil	403
Mobility and bioavailability of heavy metals	403
Climate change	404
Conclusions and the future	405
References	405

Abstract

Mineral-organic-microbial interactions drive many important biogeochemical reactions in soil environments. Microbial mineralization of soil organic matter (SOM) is an important process in which carbon (C) and nutrients are transformed to CO₂ and plant-available forms of nutrients that sustain life on the Earth. Microbes can solubilize, precipitate, and transform minerals and thus can have a significant influence on physical and chemical properties of minerals. There is an emerging view that microbial residues and dead microbial cells (or necromass) are involved in forming mineral-associated organic matter (MAOM), which is considered to be the main mechanism of OM stabilization in the soil. In this chapter we explore the complex interactions between OM and minerals and the role of the microbes in such interactions. Different procedures to isolate MAOM and various characterization techniques are discussed.

Key points

- Mineral-associated organic matter (MAOM) confers stability to organic carbon in the soil.
- Organic carbon in MAOM is mostly of microbially origin and the stabilization of organic matter in MAOM depends on the soil mineralogical properties, composition of the OM involved and environmental conditions in the soil.
- Iron and Al oxides, in particular their poorly crystalline forms, play a significant role in OM stabilization in soils due to their large specific surface area and highly reactive surfaces.
- Spectroscopic and micro-imaging techniques provide useful information about MAOM including the nature of interactions between mineral and organic compounds involved in MAOM formation.
- Mineral-organic-microbial interactions play a vital role in influencing agricultural sustainability and environment health.

Introduction

Minerals, organic matter (OM) and microorganisms are the constituents of the solid component of the soil. Soil organic matter (SOM) is a heterogeneous mixture originating from plants, animals and microbial residues. Soil OM is arguably the most dynamic, complex and least understood component of the soil. It represents the largest terrestrial carbon (C) pool, with variable size, composition and turnover time. Soil organic C (SOC) adsorbed to the reactive surfaces of soil minerals can be protected against biodegradation. This is an important natural process that preserves SOM, reduces CO₂ emission from soils, and provides benefits to crop productivity and many other soil ecosystem services. Several chemical processes are involved in the organo-mineral interactions and their stability are highly dependent on the composition of organic inputs, soil mineralogy, and environmental conditions. Soil minerals, including phyllosilicates, Fe and Al oxides, and short-range and poorly crystalline minerals, provide the reactive surface area for the adsorption of SOC. Organic C (OC) associated with minerals accounts for a major portion of the total SOC, and mineral-bound C has longer turnover times than the other fractions, such as free OM or physically protected OC in soil aggregates. Changes in the amount and stability of the OC associated with soil minerals, therefore, can have a significant influence on the bulk soil C storage in the long term and ultimately impact global warming and climate change.

In view of the significance of soil mineral-OM associations, this chapter presents an overview of the current knowledge of the proposed mechanisms, controlling factors and impact of organo-mineral interactions, as well as analytical techniques to investigate the interaction mechanisms. Important aspects of mineral-OM-microbiota interactions are also discussed.

Processes of organo-mineral interactions

Organic matter interacts with minerals through adsorption reactions on mineral surfaces and physical incorporation (occlusion) within soil aggregates. These interactions are defined by the intrinsic nature of the organic materials (e.g., their functional groups and molecular size), the type of minerals (e.g., 1:1 or 2:1 phyllosilicates or Fe oxides) and their surface properties (surface functional groups and specific surface area) as well as the soil properties (e.g., pH, type of cations, moisture status). The main mechanisms involved in the adsorption of OM to mineral surfaces are briefly described below and a schematic representation of these is given in Fig. 1.

Ligand exchange or inner-sphere complexation

A ligand exchange or substitution reaction is exactly what it says—a reaction in which one ligand in a complex ion is replaced by a different one; this mechanism is also called inner-sphere complexation or specific adsorption. Ligand exchange involves the exchange of acidic hydroxyl groups of OM (e.g., carboxyl and phenolic) with the hydroxyl groups on mineral surfaces. Ligand exchange occurs mostly in acidic soils containing Fe and Al oxides. It is one of the important mechanisms for the formation of organo-mineral associations, involving Fe (or Al)-O-C bonds.

Ion exchange

Organic cations such as protonated amines and quaternary ammonium cations can be adsorbed on to the surface of permanent charge clay minerals by cation exchange reactions or by forming outer-sphere complexes in soils (encircled-A in Fig. 1). The organic cations can exchange metal cations on the surfaces of clay minerals. Organic cations are often preferred over metal cations on soil exchange sites because of size considerations.

Polyvalent cation or water bridges

Negatively charged mineral surfaces repel organic anions, but binding occurs when a polyvalent cation is present on the exchange complex. Polyvalent cations neutralize the negative charge on the solid surface as well as the acidic group of the organic molecules

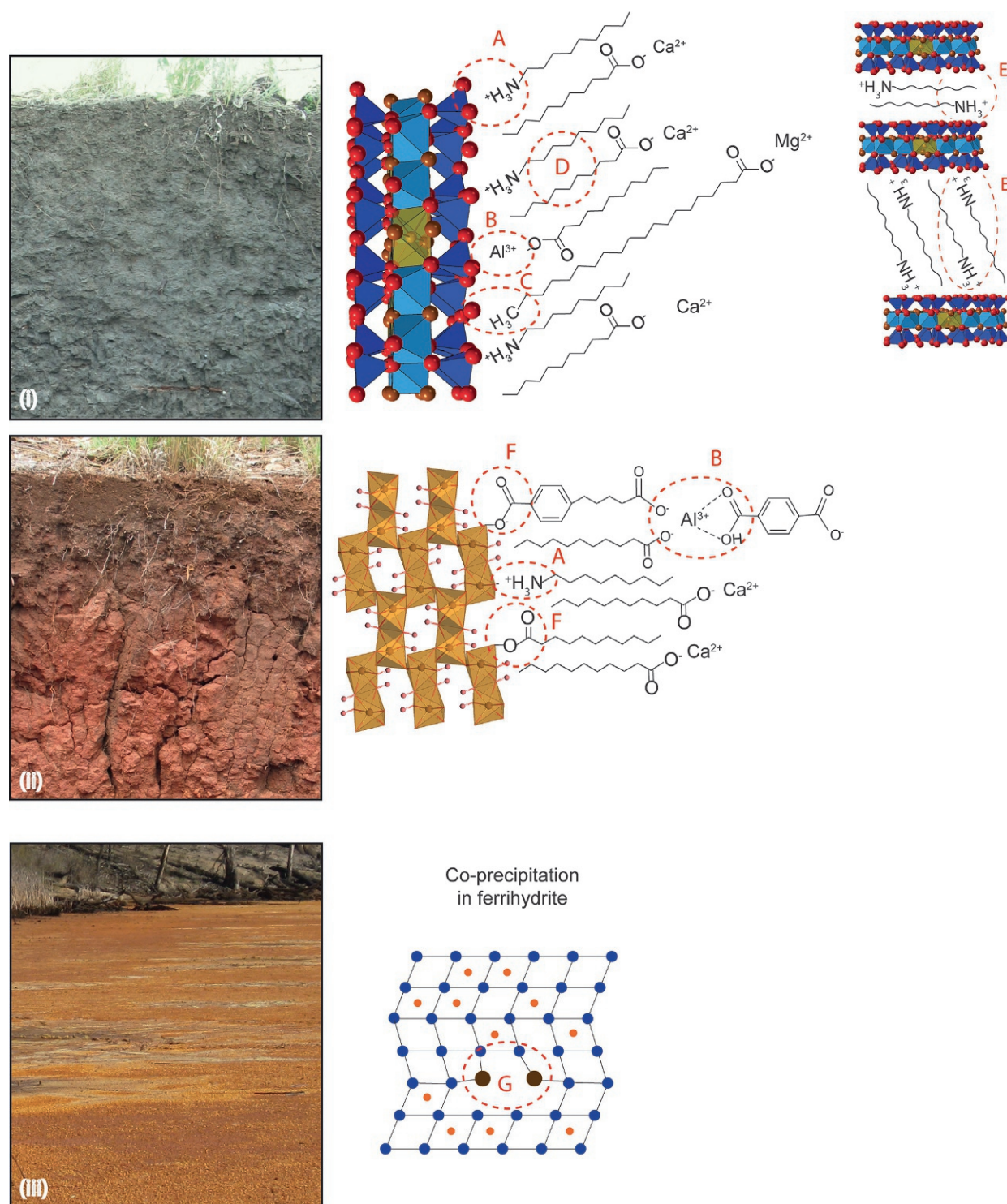


Fig. 1 Adsorption mechanisms of organic substances in three contrasting soils, (I) a Vertisol dominated by permanent charge mineral—smectite with small amounts of kaolinite and illite; (II) a Luvisol, containing hematite, goethite and phyllosilicates (smectite, illite and kaolinite); and (III) Thionic Fluvisol, consisting mostly of Fe oxides—ferrihydrate and goethite. Key mechanisms for organo-mineral interactions speculated in these soils are presented. The highlighted dotted-circled areas show—(A) electrostatic interactions (outer-sphere complexation), (B) polyvalent cation bridge, (C) Van der Waals forces, (D) hydrophobic interactions, (E) intercalation in the interlayers of expandable clay minerals, (F) inner-sphere complexation, and (G) co-precipitation in a poorly crystalline Fe oxide.

by forming a salt bridge, as shown in the encircled-B areas in Fig. 1. The predominant polyvalent cations in neutral and alkaline soils are Ca^{2+} and Mg^{2+} , whereas in acidic soils Fe^{3+} and Al^{3+} may exist in significant amounts. It has been suggested that higher valence cations, i.e., Fe^{3+} and Al^{3+} , form stronger coordination complexes than the lower valence cations, i.e., Ca^{2+} and Mg^{2+} .

Van der Waals forces

Van der Waals forces are weak electrostatic forces ($4\text{--}8 \text{ kJ mol}^{-1}$) that result from changes in the electric charge density of individual atoms. An electrically positive fluctuation in one atom causes an electrically negative fluctuation in a neighboring atom, resulting in a net attractive force. These forces are important in the adsorption of neutral polar and nonpolar molecules, particularly those with high molecular weight (encircled-C in Fig. 1).

Hydrogen bonding

When a hydrogen atom bonded to a strongly electronegative atom (e.g., an O—H bond in a water molecule) exists in the vicinity of another electronegative atom with a lone pair of electrons, hydrogen bonding is formed. For example, when a polar organic molecule or substance is unable to displace a water molecule solvating an exchangeable metal cation, it can attach directly coordinated water by forming a hydrogen bond. This type of arrangement is also referred to as 'water bridge,' where a water molecule acts as the link between an exchangeable metal cation and a polar organic molecule. Like Van der Waal's forces, hydrogen bonds are relatively weak ($2\text{--}40 \text{ kJ mol}^{-1}$) and not important for long-term protection of OM in soils or sediments.

Hydrophobic interactions

Nonpolar residues are excluded from water through entropy-related interactions, which force the nonpolar groups together. These interactions become abundant at low pH because of the protonation of carboxyl and hydroxyl groups of OM. These types of binding are less important for highly soluble low molecular weight organic acids and may be a dominant mechanism for larger macromolecules adsorption to mineral surfaces, although the bonding is very weak (4 kJ mol^{-1}).

Intercalation

Intercalation is a special type of adsorption of organic ligands within the interlayer spaces of expandable 2:1 phyllosilicates (as shown by encircled region E in Fig. 1), particularly at $\text{pH} < 5$. Organic cations can intercalate beyond the cation exchange capacity if Van der Waals forces play a major role in binding molecules to surfaces. Contrasting observations have been made about the intercalation of OM into interlayers of expandable phyllosilicates in acid soils ($\text{pH} < 4$). Because of the lack of specific methods to characterize and quantify intercalated OM, this issue has remained unresolved.

Co-precipitation

Co-precipitation of OM can occur when adequate metal (Al or Fe) and dissolved OM co-exist in certain soil environments, such as acid sulfate soils or soils that frequently experience periodic fluctuations in redox conditions, such as peat and wetlands. In this process a monomeric or polymeric aqueous metal species (Al or Fe) forms a mixed metal–organic solid from the solution after reacting with organic compounds (encircled-G in Fig. 1). This pathway may increase the potential quantity of mineral-associated OM per unit mineral surface compared to surface adsorption. For example, in a waterlogged paddy soil condition, coprecipitation resulted in greater OC contents ($49\text{--}213 \text{ mg g}^{-1}$) compared to adsorbed systems ($18\text{--}47 \text{ mg g}^{-1}$), and also enhanced the inclusion of OM within highly aggregated associations (Sodano et al., 2017).

Physical protection of SOM

Physical protection of SOM can occur through its incorporation into stable aggregates by the agglomeration or aggregation of the colloidal particles. Soil OM may act as a glue that holds inorganic soil particles together into aggregates. The aggregates can be grouped as macro ($>250 \mu\text{m}$)- or microaggregates ($<250 \mu\text{m}$) depending on their sizes. The initial unit of aggregation is called a microaggregate; it forms through the interaction of clay colloids with organic compounds and inorganic cementing materials. The microaggregates are bound together by organic compounds of different origins (e.g., plants or microbes) or structural binding agents (e.g., root hair and fungal hyphae). Aggregates of various sizes may also interact with one another in complex ways, conceptualized as the 'aggregate hierarchy' where aggregates form around microaggregates, and facilitate formation of additional macroaggregates, and mineral-associated OM (MAOM) within themselves. As a result of the occlusion of organics into large aggregates, their decomposition rate decreases, but to a lesser degree than the MAOM. Organic C in macroaggregates has a lifetime of only a few years whereas it is a century of lifetime for C in microaggregates. Soil aggregation and the protection of SOM are affected by several factors, including moisture content, clay content, mineralogy, and quantity and quality of OM in soils. The stability of aggregates is also affected by management practices; microaggregates are often formed as a result of disruption or decomposition of macroaggregates due to different management activities including tillage.

Mineral-organic matter-microbiota interactions

Soil microbial biomass constitutes about 5% of OM in the soil. Microorganisms live in the soil pore space, and therefore water content, pore-size distribution and pore inter-connectivity determine their abundance and activity in the soil. Soil microbial community structure and metabolic potential are greatly influenced by soil management ecosystems (e.g., arable, grassland and forest), environmental factors that influence C-input, and soil properties, such as soil pH, soil texture, mineralogy. Simultaneously, microbial-derived products and residues, such as, polysaccharides, proteins, cell wall polymers, and other organic molecules also become integral components of OM during formation, mobilization, transformation and storage of SOM. Thus, microbial community structure and composition, range and magnitude of the microbial activities, ability of microbes to metabolize or transform OM, and bioavailability of C substrates in OM are intimately connected, but these complex and highly dynamic interactions are still not well recognized.

Microorganisms interact with soil minerals in several ways, such as, microbially facilitated formation and dissolution of minerals, and formation of organo-mineral associations via the accumulation of microbial products on mineral surfaces, and ultimate acquisition and utilization of OM in organo-mineral associations for microbial metabolism (Kleber et al., 2015). Microorganisms have been shown to significantly facilitate weathering of minerals. Rhizosphere bacteria can directly solubilize minerals from soil. The mineral dissolution resulting from the microbial activity can be either an active or a collateral process. Free-living microorganisms and symbiotic fungi weather minerals through the formation of biofilms on their surface. The key mechanisms usually linked to bacterial weathering include pH changes adjacent to the mineral particle and proton promoted dissolution, chelation of elements present in a mineral matrix or the soil, and redox reactions (Samuels et al., 2020). Microbial weathering of mafic silicates, basalts, and glasses produce a variety of clay minerals, such as smectite, zeolite, and serpentine. Generally, microbes do not precipitate clay minerals directly, but the products of their weathering reactions reprecipitate to form clay minerals when surroundings become favorable.

A wide range of bacteria and archaea can reduce iron by typically attaching themselves to clay mineral surfaces (Dong 2012). An example of this is presented in Fig. 2, where Fe(III) in the structure of nontronite (an iron-rich smectite) is reduced by microbes, which is coupled with the oxidation of lactate to acetate. The reduction of structural Fe(III) to Fe(II) causes a decrease in the surface area, interlayer spacing, water swellability, and hydraulic conductivity of clay minerals. The reduction of structural Fe in smectite increases the negative layer charge and cation exchange capacity of the mineral, while the interlayer cations tend to be less exchangeable.

Microbial dissolution of minerals is demonstrated in Fig. 3, where *Shewanella putrefaciens* promotes the dissolution of nontronite (Perdrial et al., 2009). The authors hypothesized that the unstable octahedral Fe(III) at the edge of nontronite was complexed by the organic catecholate ligands and promoted the formation of Fe chelating agents, which enhanced the dissolution of nontronite (Fig. 3A). Subsequently, Ca^{2+} ions released (Fig. 3A) from the interlayers of the nontronite formed a bridge between bacteria and the nontronite surface, and later Si^{4+} was released (Fig. 3B) from the tetrahedral sheet of the mineral and precipitated as amorphous silica.

Microorganisms interact with minerals and metals for their own benefits. For example, prokaryotes use redox active metals as electron donors or acceptors to gain energy for growth. This leads to dissolution or precipitation of metals and this process ultimately influences the cycling of metals and fate of MAOM. Microbes also require metals for protein synthesis and have element specific sensory systems and uptake processes. For example, microorganisms can release siderophores to locate and chelate the

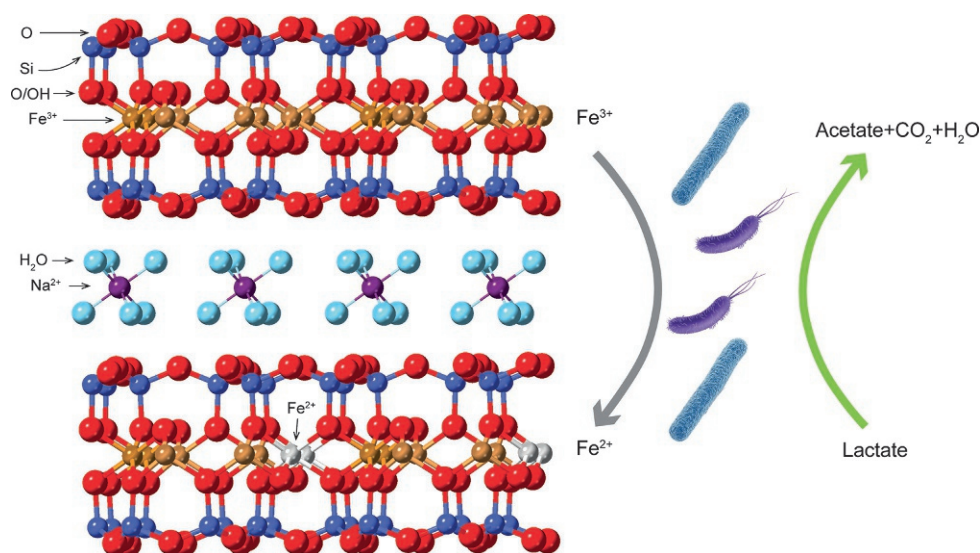


Fig. 2 Schematic diagram showing the structure of a 2:1 clay mineral nontronite (an iron-rich smectite) and microbial reduction of structural Fe(III) that is present in the octahedral sheet. The bioreduction of octahedral Fe(III) is coupled with the oxidation of lactate to acetate. Electron transfer is shown to take place parallel to the clay layers, but this does not exclude other possibilities. Based on Dong H (2012) Clay-microbe interactions and implications for environmental mitigation. *Elements* 8(2): 113–118.

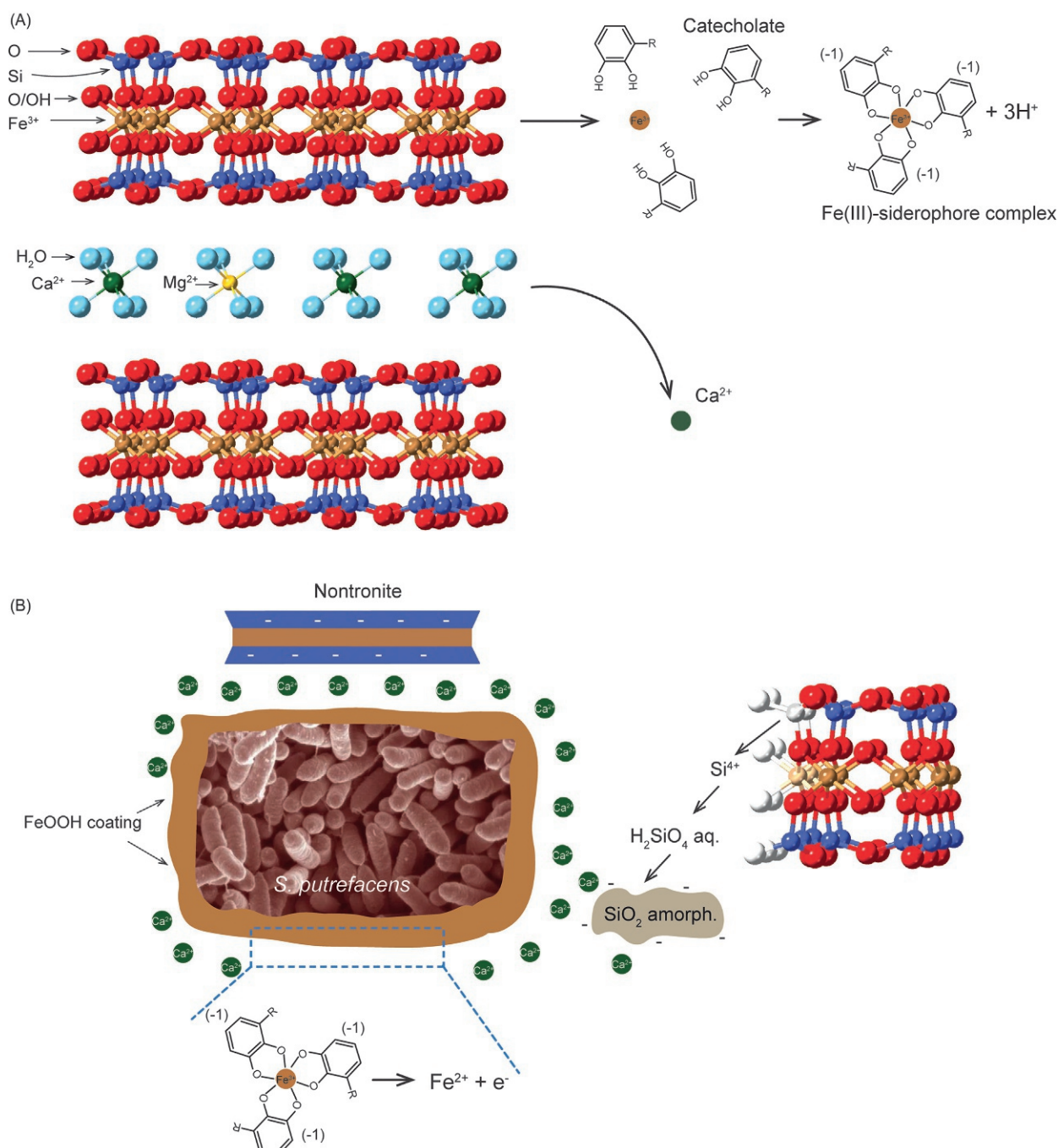


Fig. 3 Schematic representation of the interaction of nontronite (an iron-rich smectite) and *Shewanella putrefaciens* at the lattice layer scale. (A) An octahedrally coordinated Fe(III) is mobilized at an unstable edge site and Ca²⁺ ion from is released from interlayer sites into solution. Bacterially produced ligands, catecholate, form siderophore complexes with Fe(III) as the central cation, and (B) released Ca²⁺ ions acts as a bridge between smectite and bacteria. The Fe(III) complex is transported to the cell, dismantled and reduced or involved in the formation of Fe(III)-coatings at the cell surface. As the complexation leads to continuous nontronite dissolution, Si is released and precipitated as amorphous silica, which is presumably associated with EPS. Based on Perdril JN, Warr LN, Perdril N, Lett MC and Elsass F (2009) Interaction between smectite and bacteria: Implications for bentonite as backfill material in the disposal of nuclear waste. *Chemical Geology* **264**(1–4): 281–294.

target metals from solid phases and transport them to the cell. To avoid adverse intracellular metal concentrations, microbes sometimes bind metals at the cell surface or in cytoplasm or on extracellular polymeric substances (EPS).

Generally, both the bacterial cells and mineral surfaces are negatively charged and bind to each other by forming a polyvalent cation bridge. Microbial cells can also bind to negatively charged surfaces of minerals through electrostatic forces when mineral colloids are coated with positively charged hydroxy polymers. Microbes produce EPS or cell peripherals like flagella or pili to adhere or attach to the mineral surfaces. EPS are composed of polysaccharides, proteins and lipids and have a very important role in the

interaction of minerals and OM in the soil. Preferential sorption of proteins on goethite, kaolinite and montmorillonite, and phosphate rich compounds and polysaccharides on Al hydroxides have been found. Microbe cell-connected siderophores can form inner sphere complexes between catechol groups and hydroxylated mineral surfaces. Also, Fe reducing bacteria can recognize Fe oxide surfaces and control surface attachment via outer membrane proteins.

Microbially produced enzymes can be easily adsorbed at mineral surfaces. Enzymes, like most proteins, are made up of amino acids with properties ranging from hydrophilic to hydrophobic, and carry negative, positive or no charge. Enzymes exhibit a strong affinity for soil mineral surfaces. In earlier studies, ionic exchange, electrostatic attraction, and entropic effects such as hydrophobic interactions were thought to be the main mode of bonding between enzymatic proteins and mineral surfaces. Interlayer adsorption of proteins on to the expandable clay minerals has also been considered in some instances. In some soils, short-range ordered or poorly crystalline metal oxides, such as Al and Fe oxides, may play a dominant role in regulating enzyme stability and activity.

Abiotic and biotic controls on soil organo-mineral interactions

Organic matter dynamics

Soil OM is a heterogeneous mixture of organic compounds of various origin, form and stability. Organic compounds in SOM can vary in molecular size, solubility, molecular complexity, aromaticity, aliphaticity and other properties. Generally, simple organic compounds, such as carbohydrates, amino acids, are easily mineralized with a short turnover time (labile) compared to complex compounds, such as, lignin and aromatics, with a longer turnover time (stable). The molecular variations in SOM originate from the source of origin and the decomposition stage of OM. Plant litter is the primary source for OM formation in soils. Soil microorganisms alter plant residues, resulting in changes in the chemical structure and synthesis of organic products of microbial origin. Cotrufo et al. (2013) proposed the 'microbial efficiency-matrix stabilization' framework for explaining the relation of plant litter input to the decomposition and formation of MAOM. The framework hypothesizes that soil microbial community mineralizes plant litter and the microbial products of decomposition then become the main precursors of MAOM, as shown in the left panel, microbial filter, in Fig. 4. The proficiency of MAOM formation is considered to be greatly affected by the litter quality, or in other words, substrate C use efficiency. The substrate use efficiency is greater for labile organic compounds, such as, carbohydrates and proteins, compared to the stable litter components like lignin. However, Mikutta et al. (2019) observed little effect of litter quality on the accumulation of MAOM. According to these authors, plant litter C can bypass microbial assimilation and be sorbed directly onto minerals to form MAOM, which is depicted in the right panel (mineral filter) in Fig. 4. The 'mineral filter' hypothesis assumes that POM and plant litter may generate dissolved OM which can be sorbed onto minerals in surface soils. From the dissolved OM,

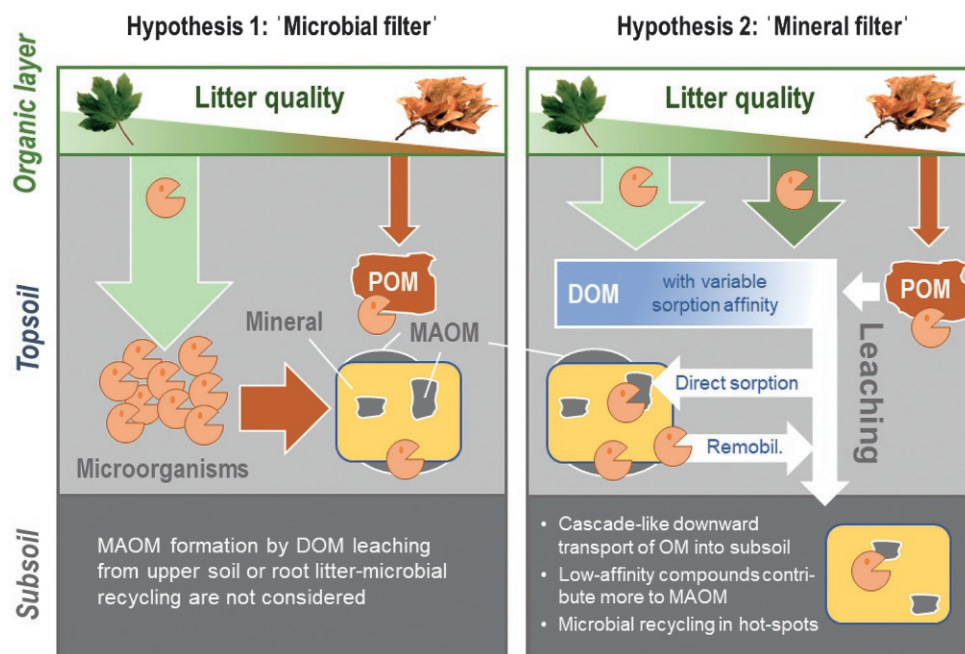


Fig. 4 Two opposing hypotheses on pathways for the formation of mineral-associated organic matter (MAOM) in soils. According to the 'microbial filter' hypothesis (left) litter quality is the major driver for MAOM formation, with easily decomposable high-quality litter being preferentially assimilated by microorganisms due to higher substrate use efficiency and microbial assimilates then become sorbed to mineral surfaces. Low-quality litter with greater content of structural lignin components can accumulate in mineral soils as residual particulate organic matter (POM). The 'mineral filter' hypothesis (right) presumes that microbial degradation of plant litter and POM produces a wide spectrum of dissolved organic matter (DOM) components that interact with mineral surfaces. According to this hypothesis, both microbial and plant-derived OM components are involved in the MAOM formation. From Mikutta R, Turner S, Schippers A, Gentsch N, Meyer-Stüve S, Condron LM, Peltzer DA, Richardson SJ, Eger A, Hempel G and Kaiser K (2019) Microbial and abiotic controls on mineral-associated organic matter in soil profiles along an ecosystem gradient. *Scientific Reports* 9(1): 1–9.

phenolic and proteinaceous compounds are mentioned as preferentially sorbed to soil minerals, while the less strongly sorbed compounds, such as polysaccharides, are vertically transported to the sub-soil, where they may experience repeated sorption and remobilization cycles. Recently, Islam et al. (2022) postulated that the stabilization of OC into MAOM is a complex phenomenon that not only depends on the microbial C use efficiency of organic substrates but the abundance of active clay surfaces in soils. They postulated that microbial use and recycling of plant- and microbially-derived substrates decline with increased abundance of active clay surfaces. They further suggested that consideration of the interactions between biotic and abiotic controls on SOM stabilization is necessary to understand how efficiently C substrates of different chemistries are incorporated into MAOM in different soils.

The “humification” concept of forming a highly condensed macromolecular structure and the “selective preservation” concept of preferentially decomposing labile OM for forming stable OM in soils have been challenged in recent years. Lehmann and Kleber (2015) argued that “the available evidence does not support the formation of large-molecular-size and persistent ‘humic substances (HS)’ in soils” and proposed a new concept “soil continuum model (SCM)”. This concept indicates the importance of mineral-OM association for forming stable OM. According to this model, OM exists as a continuum of organic fragment sizes. The decomposition of large molecules leads to a decrease in the size of primary plant material with concurrent increases in polar and ionizable groups and thus to an increased solubility in water. At the same time, the chance for protection against further decomposition increases through greater reactivity towards mineral surfaces. In contrast, De Nobili et al. (2020) presented evidence to demonstrate that the SCM is not valid and that humification does occur along with synthesis of molecules that are not simply products of lysis of plant components. De Nobili and other researchers have argued that the SCM does not account for the complexity of SOM and does not consider the enormous amount of scientific literature that proves the high reactivity of plant and microbial metabolites.

Distinct affinity of minerals towards organic compounds has been shown based on the chemical composition of different mineral-organic associations (Kögel-Knabner et al., 2008; Yeasmin et al., 2014, 2017a). However, the information regarding the nature of MAOM is often inconsistent, with preferential association of aromatic C and protonated amide N with smectite and illite, polysaccharides with kaolinite, and aromatic and carboxylate C with Fe and Al oxides being observed.

Mineral phase properties

The soil clay fraction encompassing phyllosilicates, Fe and Al oxides, short-range-order Fe oxides, and poorly ordered aluminosilicates provides the majority of the surface area for the adsorption of OM in soils. Phyllosilicates are layered aluminosilicates with 1:1 and 2:1 unit layers which have different surfaces. The basal surfaces of 2:1 clay minerals (e.g., illite, smectite, vermiculite) consist of oxygen ions of the tetrahedral sheet, which carry permanent negative charge caused by isomorphic substitution (e.g., Al^{3+} for Si^{4+} in the tetrahedral sheet or Mg^{2+} for Al^{3+} in the octahedral sheet) in the mineral structure. The 1:1 layer phyllosilicates (e.g., kaolinite) are comprised of a tetrahedral and an octahedral sheet and have two types of basal surfaces, one consisting of basal oxygen ions of the tetrahedral sheet and the other made up of the hydroxyls attached to Al in the octahedral sheet. There is a little isomorphic substitution in the structure of 1:1 clay minerals. Both 2:1 and 1:1 phyllosilicates have variable charge developed on the surfaces at the broken edges of the minerals. Owing to the permanent negatively charged surfaces of 2:1 layer silicates, organic anions (e.g., carboxylates) are repelled from the surfaces and binding occurs through cation bridging only if polyvalent cations are present in soils. Cationic organic compounds (e.g., protonated amide) can interact directly with negatively charged phyllosilicate surfaces. Apolar organic compounds (e.g., alkyl C) can associate with hydrophobic basal surfaces of layer silicates through hydrophobic interactions. Organic compounds may also bind to phyllosilicate basal surfaces via weak interactions, such as Van der Waals forces and hydrogen bonding. The surface hydroxyls at the broken edges of layer silicates are positively charged under acidic conditions and associate with anionic organic molecules via ligand exchange. Non-expandable layer silicates, such as kaolinite, are believed to have only weak bonding affinities to organics due to the lack of layer charge and interlayer spaces on their surface compared with expandable 2:1 clay minerals such as smectites.

Iron and Al oxides carry variable surface charge. The hydroxyl groups on the oxide surfaces are positively charged in acidic soils and associate with anionic carboxyl groups of OM through ligand exchange. Organic molecules can also interact with Fe or Al oxide surfaces by forming hydrogen bonds and hydrophobic interactions. Soils, richer in oxide minerals (e.g., Ferralsols or Oxisols) and poorly crystalline aluminosilicates (e.g., Andosols), show greater C in MAOM than do soils richer in phyllosilicates (Yeasmin et al., 2017a). The significance of Fe and Al oxides in the OM stabilization in soils is demonstrated by a strong correlation between Fe oxide and SOM contents, reduction in the OM sorption after removal of Fe and Al oxides from soils and decreased mineralization of dissolved OM in the presence of Fe oxides.

Generally, short-range-order and poorly crystalline minerals of Fe, Al and aluminosilicates, (e.g. ferrihydrite, allophane) have a large ($>200 \text{ m}^2 \text{ g}^{-1}$) specific surface area (SSA); goethite, montmorillonite or illite have an intermediate SSA; but kaolinite and gibbsite have a much smaller SSA. The sorption of OC has been shown to increase with increasing SSA of soil minerals. The sorption of OC in the soil occurs preferentially at reactive sites, such as edges of phyllosilicates, particularly on Al—OH groups, rough surfaces (e.g., crystal surfaces of Fe oxides) or soil micropores. Due to preferential binding at the reactive sorption sites, the accumulation of OM at mineral surfaces occurs in patches instead of as a continuous coating. The patchy surface coverage by OM, however, is inconsistent with the use of SSA or micropores in predicting the amount of OM stabilized in MAOM in soils (Kögel-Knabner et al., 2008). The capacity of soils to stabilize OC can be restricted by OC saturation, which corresponds to the saturation of reactive sites on mineral surfaces. Using data from 84 sites in 27 countries, Fujisaki et al. (2018) determined the C saturation value in the fine fraction (MAOM) of tropical soils as 53 g C kg^{-1} , which is smaller than the global SOC storage value (78 g C kg^{-1}), which is based mostly on temperate soils. Fujisaki et al. (2018) also found the MAOM fractions very close to their SOC saturation capacity in forest soils as opposed to croplands.

It has been suggested that the organic molecules adsorbed first might be strongly stabilized via multiple ligand interactions. With increasing surface loading, sorption of newly added OC occurs with fewer ligand attachments, thereby leaving parts of unattached organic molecules that are more susceptible to degradation. Consequently, there is a greater potential for OM preservation in mineral rich sub-soils, as evidenced by a larger proportion of mineral bound OC held with stronger bonding than surface soils (Kögel-Knabner et al., 2008). However, less decomposition of SOM in sub-surface soils has also been explained by decreased microbial activity, resulting from reduced oxygen diffusion and lack of fresh organic residue input rather than the formation of MAOM. The stabilization of SOM in subsoil needs further research, particularly underlying mechanisms for the stabilization of SOM in MAOM. Keiluweit et al. (2017) observed that a shift in microbial metabolism to less efficient anaerobic respiration at anaerobic microsites, resulted in the selective accumulation in soil of reduced organic compounds, such as lipids and waxes. Quantitative data are needed to confirm the relationship between the abundance of anaerobic microsites and the amount of SOM stabilized or protected.

Soil solution chemistry

Differences in soil hydrological conditions, pH and ionic strength influence the formation and stock of MAOM in soils. For example, total adsorption of D-gluconic acid on gibbsite in 5 mM NaClO₄ solution was approximately tenfold greater at pH 9 and 12 than at pH 4 and 7 (Fig. 5, Schneckenburger et al., 2018). Changes in hydrological conditions can shift soil pH and redox conditions that can alter the reactivity of some soil minerals, such as, Fe and Al oxides. The potential of oxide minerals to stabilize OM is mostly relevant in acidic soils. The complexation of OM involving ligand exchange reactions on Fe oxides surfaces increases with decreasing solution pH (Yeasmin et al., 2014), which can be explained by the dissociation constant of the organic anions and increasing positively charged surfaces on oxides. Even under alkaline conditions Fe and Al oxides can possess a small positive surface charge and thus can bind anionic organic compounds. Additionally, an increase in ionic strength can enhance the adsorption of OM on minerals; this so-called 'salting-out effect' occurs due to a decrease in solubility of the nonelectrolyte in water in the presence of a salt.

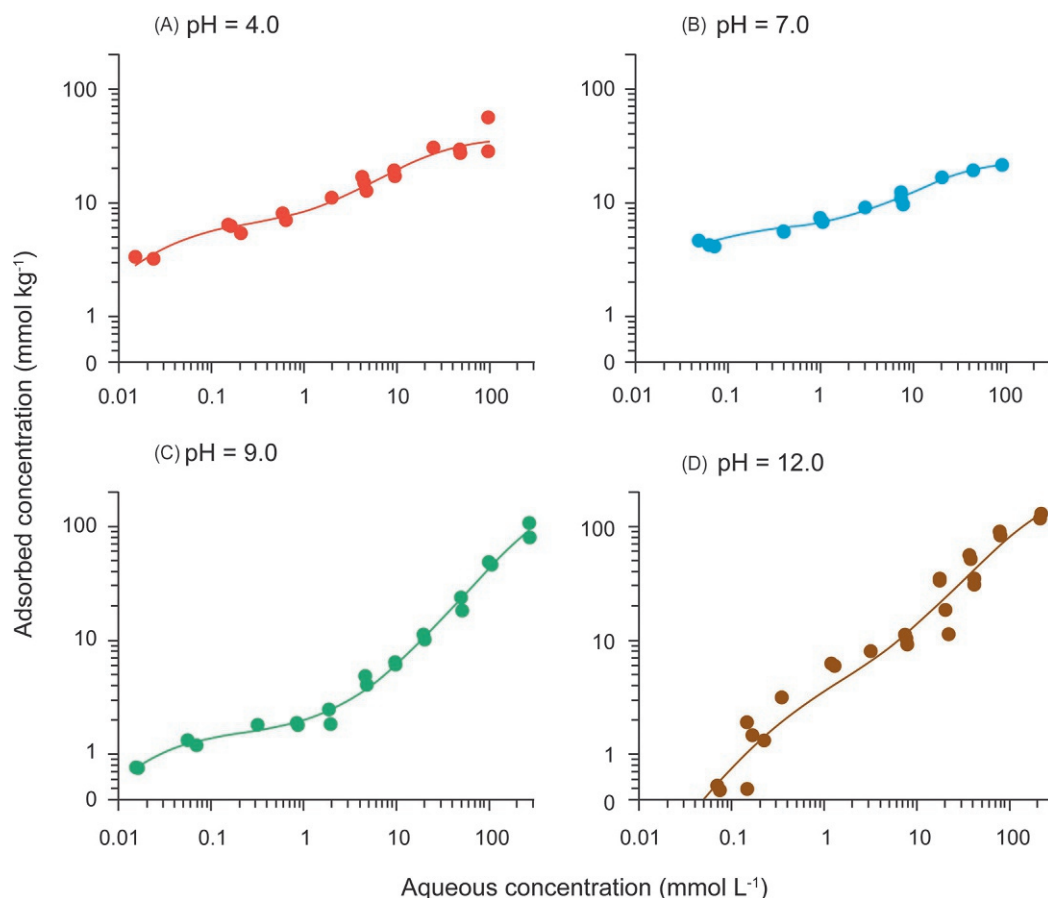


Fig. 5 Sorption isotherms of aliphatic (poly)hydroxy carboxylic acid, D-gluconic acid, on gibbsite at four pH values, i.e., pH 4, 7, 9 and 12. Adsorption experiments were conducted using 5 mM NaClO₄ as the background solution. From Schneckenburger T, Riefstahl J and Fischer K (2018) Adsorption of aliphatic polyhydroxy carboxylic acids on gibbsite: pH dependency and importance of adsorbate structure. *Environmental Sciences Europe* 30(1): 1–15.

Insights into organo-mineral associations

Because of the complex nature of MAOM, the processes and mechanisms for their formation in different soil types are not well understood. The involvement of multiple mineral and organic phases of varying compositions and structures, and the existence several simultaneous abiotic and biotic reactions make the characterization of MAOM very difficult if not impossible. Different procedures have been used by researchers for the isolation and characterization of both organic and minerals phases present in MAOM, and some of the main procedures are given in Fig. 6.

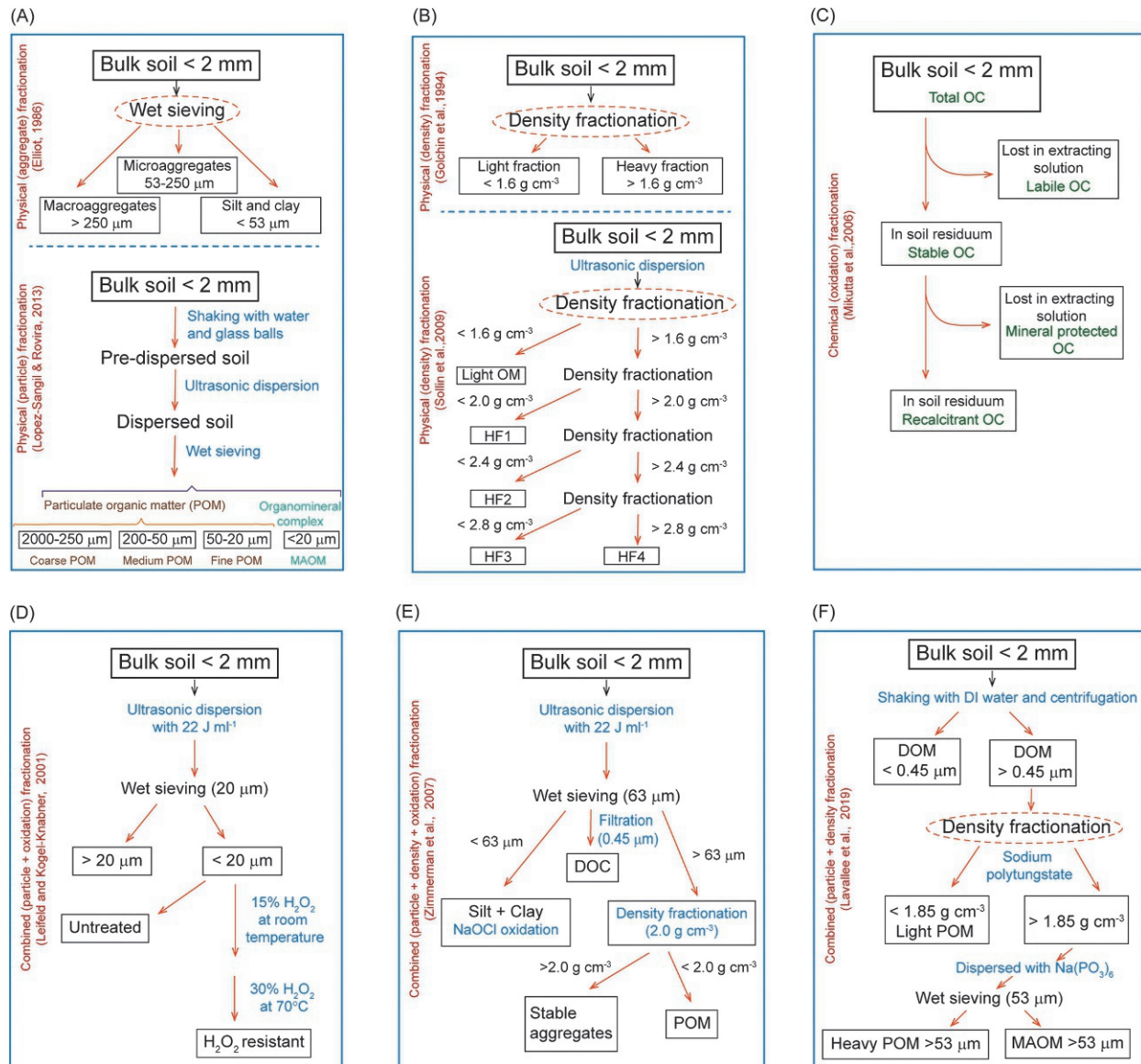


Fig. 6 Schematic graph of physical (A and B), chemical (C) and combined chemical and physical (D and E) fractionation procedures of soil organic matter (SOM) (adapted from Poeplau C, Don A, Six J, Kaiser M, Benbi D, Chenu C, Cotrufo MF, Derrien D, Gioacchini P, Grand S and Gregorich E (2018) Isolating organic carbon fractions with varying turnover rates in temperate agricultural soils—A comprehensive method comparison. *Soil Biology and Biochemistry* **125**: 10–26). (A) Aggregates and particle fractionation: separated (i) SOM into aggregate size classes (macroaggregates >250 μm; microaggregates 250–53 μm; silt and clay <53 μm) by wet sieving yielded three particle fractions matching three classical fractions in texture analysis; and (ii) the second procedure isolated SOM into four particle sizes: particulate OM-POM-2000–200 μm (coarse sand), 200–50 μm (fine sand), 50–20 μm (coarse silt) and organo-mineral complexes (<20 μm). (B) Density fractionation: (i) procedure separated SOM into light fraction (LF; <1.6 g cm⁻³) and heavy fraction (HF; >1.6 g cm⁻³) by density fractionation using sodium polytungstate (SPT); and (ii) SOM separated into LF, and four HF consisting of organo-mineral complexes of different mineralogies. (C) Chemical fractionation: Oxidation separated fractions according to protection mechanisms i.e., mineral-protected and recalcitrant organic carbon. (D) Particle + oxidation procedure isolated SOM into >20 μm and <20 μm, and <20 μm fraction was treated with H₂O₂ to obtain H₂O₂ oxidation resistant SOM. (E) Particle + density + oxidation procedure: Isolated SOM into DOC = dissolved organic carbon, SC = silt + clay, rSOC = oxidation resistant organic carbon, POM, SA = stable aggregate that are related to the model pools of the Rothamsted carbon model. (F) Size + density fractions (based on Lavalley JM, Soong JL and Cotrufo MF (2019) Conceptualizing soil organic matter into particulate and mineral-associated forms to address global change in the 21st century. *Global Change Biology* **26**: 261–273): separates SOM into DOM = dissolved OM, water extractable, <0.45 μm, POM = particulate OM, not water extractable, >0.45 μm; light POM ≤1.60–1.85 g cm⁻³, heavy POM ≥1.85 g cm⁻³ and >53 μm, MAOM = mineral associated OM, >1.85 g cm⁻³ and <53 μm. The details of the references are given in Poeplau et al. (2018).

Isolation of mineral associated organic matter

Physical procedures

Physical fractionation procedures involve various degree of soil dispersion, followed by density or size separation (Fig. 6A and B) to isolate pools of SOM associated with minerals of different sizes and density. Organic substances isolated using these procedures have been useful to study the interactions between organic and mineral components of soils and the fractions have been related to the turnover of SOM.

Particle size fractionation

Particle size fractionation (Fig. 6A) is based on the concept that SOM associated with particles of different sizes and different mineralogical composition varies in its structure and function. Generally, quartz is the dominant mineral in the sand fraction whereas phyllosilicates and Fe or Al oxides are abundant in the clay fraction of soils. A meta-analysis of the typical allocation of OM in particle size fractions of temperate arable top-soils show about 50–70% of the total SOC is associated with the clay fraction ($<2\ \mu\text{m}$), 20–40% with the silt fraction ($2\text{--}63\ \mu\text{m}$) and $<10\%$ with the sand fraction ($>63\ \mu\text{m}$) (von Lützow et al., 2007). The greater association of SOC with smaller particles does not always correspond with a longer turnover time. von Lützow et al. (2007) reported a shorter turnover time range (76–190 years) for SOC associated with the clay fraction compared to the OC associated with the silt (115–676 years) and sand (0.5–374 years) fractions. Therefore, particle size fractions may provide a crude measure of the composition and turnover time of OM associated with different size fractions.

Density fractionation

Density fractionation (Fig. 6B) is used to isolate free or unassociated OM (light fraction) from MAOM (heavy fraction) in soils. Soil OM typically has density varying between 1.24 and $1.64\ \text{g cm}^{-3}$ and soil minerals have a particle density greater than $1.6\ \text{g cm}^{-3}$. Aqueous solutions of various inorganic salts, such as, magnesium sulfate (MgSO_4), zinc bromide (ZnBr_2), sodium iodide (NaI), and sodium polytungstate ($\text{Na}_6(\text{H}_2\text{W}_{12}\text{O}_{40})$), have been commonly used for the density fractionation of SOM. Sodium polytungstate is the most widely used, since a wide range of densities can be obtained using the salt (from 1.0 to $3.1\ \text{g cm}^{-3}$) and it has a low viscosity, little toxicity and can be easily recycled. The light fractions of SOM contain predominantly easily decomposable plant residues, whereas the heavy fractions comprise MAOM, with highly decomposed stable OM. However, the presence of charred material, in addition to the plant litter, in the light fraction has also been reported, which can lead to the overestimation of the OM in the light fraction and underestimation of MAOM in the heavy fractions.

Separating SOM into multiple sequential density fractions has also been used to isolate OM associated with individual clay minerals, with the aim to investigate the interaction mechanism and composition of OM associated with individual mineral matrices in the density fractions (Sollins et al., 2009; Jones and Singh, 2014; Yeasmin et al., 2017a). Generally, OM associated with phyllosilicates is present in the density fraction between 1.8 and $2.6\ \text{g cm}^{-3}$ and with Fe oxides in density fractions $>1.8\ \text{g cm}^{-3}$ (Yeasmin et al., 2017a). Nevertheless, sequential density fractionation is a useful procedure to determine OM-mineral associations and OC turnover dynamics in MAOM.

Chemical procedures

Chemical isolation procedures involve either a selective oxidation of SOM or selective dissolution of mineral phases to characterize SOM.

Oxidation of SOM

Soil OM can be oxidized using reagents, such as hydrogen peroxide (H_2O_2), sodium hypochlorite (NaOCl), potassium permanganate (KMnO_4), and sodium persulfate ($\text{Na}_2\text{S}_2\text{O}_8$), all of which remove easily decomposable or labile OM and leave behind a relatively stable OM fraction in the residues. The chemical oxidation is expected to mimic the biodegradation of SOM under natural conditions. Among the SOM oxidizing reagents, NaOCl is preferred for isolating the labile and stable MAOM because of the oxidation efficiency and minimal effects on soil minerals (Mikutta et al., 2005, Fig. 6C). The removal of about 26–96% of OC from soils with different mineralogies, and 77–95% of the initial OC from 196 mineral soil fractions by oxidation with NaOCl support the potential of this reagent to separate OM associated with minerals (Kaiser and Guggenberger, 2003). The residues of NaOCl treated OM were characterized by the presence of aromatic and alkyl C, which had been sequestered for a longer time than that released by other oxidants (Mikutta et al., 2005; Yeasmin et al., 2017a). Whether the isolated residues represent the stable SOM pool, due to the selective oxidation of labile organic compounds or due to the association of certain SOM substances with the mineral matrix, still needs further investigation.

Dissolution of minerals

Hydrofluoric acid (HF) has been used to isolate MAOM by the dissolution of silicates, carbonates and metal oxides. Dissolution of total OC up to 57% in the mineral rich sub-surface soil horizons led to the conclusion that HF dissolves the MAOM fractions (Rumpel et al., 2008). However, the mineral removal efficiency from soils and the composition of dissolved OM with the HF treatment can vary with soil types and depth. In general, the HF soluble OM fraction was found to be older OC than the HF resistant OM that was left in the residues. Yeasmin et al. (2017b) found the HF treatment to be better than NaOCl and dry ashing treatments for characterizing SOC using IR spectroscopy, the HF treatment rendered mineral interference-free spectra with relatively least

affected and accentuated organic bands. The authors also pointed that HF treated spectra may reflect the total OC qualitatively, but one should be careful in using spectra for quantitative analysis after the HF treatment, particularly in mineral rich soils.

Combination of physical and chemical procedures

Combined physical and chemical fractionation procedures (Fig. 6D–F) have been used to overcome the limitation of the individual physical or chemical procedures. In this approach, physical fractions are expected to capture the spatial arrangement of different MAOM particles, whereas the chemical isolation can be suitable for elucidating the molecular-level interactions between SOM and minerals. Various schemes have been employed combining physical and chemical procedures, e.g., size or density fractions followed by oxidation or acid hydrolysis. Yeasmin et al. (2017a) used a combined procedure of physical (sequential density fractionation) and chemical (NaOCl, HF) to examine the role of minerals in protecting OC in four different soil types (Ferralsol, Luvisol, Vertisol and Solonetz) (Fig. 7). They effectively isolated MAOM into three distinct groups in the soils: (i) OM associated with Fe and Al oxides (1.8 to >2.6 g cm⁻³ in the Ferralsol); (ii) OM associated with phyllosilicates (1.8 – 2.6 g cm⁻³) and (iii) OM associated with quartz and feldspar (>2.6 g cm⁻³) in the Luvisol, Vertisol and Solonetz by a sequential density fractionation procedure. Later, after chemical fractionations, it was observed that the Fe and Al oxides dominated fractions were relatively more resistant to NaOCl oxidation and more OC was dissolved with the HF treatment than other fractions, which suggests that OC is protected better against oxidative degradation in MAOM formed in association with Fe or Al oxides than MAOM formed in association with phyllosilicates, and quartz and feldspar matrices.

Poeplau et al. (2018) evaluated 20 different SOC fractionation methods for their suitability to isolate fractions with varying turnover rates, using agricultural soils with vegetation change from C₃ to C₄. Based on different criteria, including differentiation, redundancy, recovery, reproducibility, workload, and overall performance indicators, they identified five best methods and among them three methods used combined physical and chemical approaches, i.e., particulate + density, aggregate + oxidation, and particulate + density + oxidation.

Thus, combinations of physical and chemical procedures have been found to be more effective in isolating the MAOM fractions than a single method. Physical isolates represent the effects of the spatial arrangement of primary and secondary organo-mineral associations on SOM dynamics, whereas the chemical fractions are more appropriate to explicate molecular-level interactions between OM and mineral particles in the soil (von Lütow et al., 2007). Preceding spatial separation of the discrete mineral-OC assemblages by physical fractionation might facilitate the chemical treatments to produce more homogenous OC fractions.

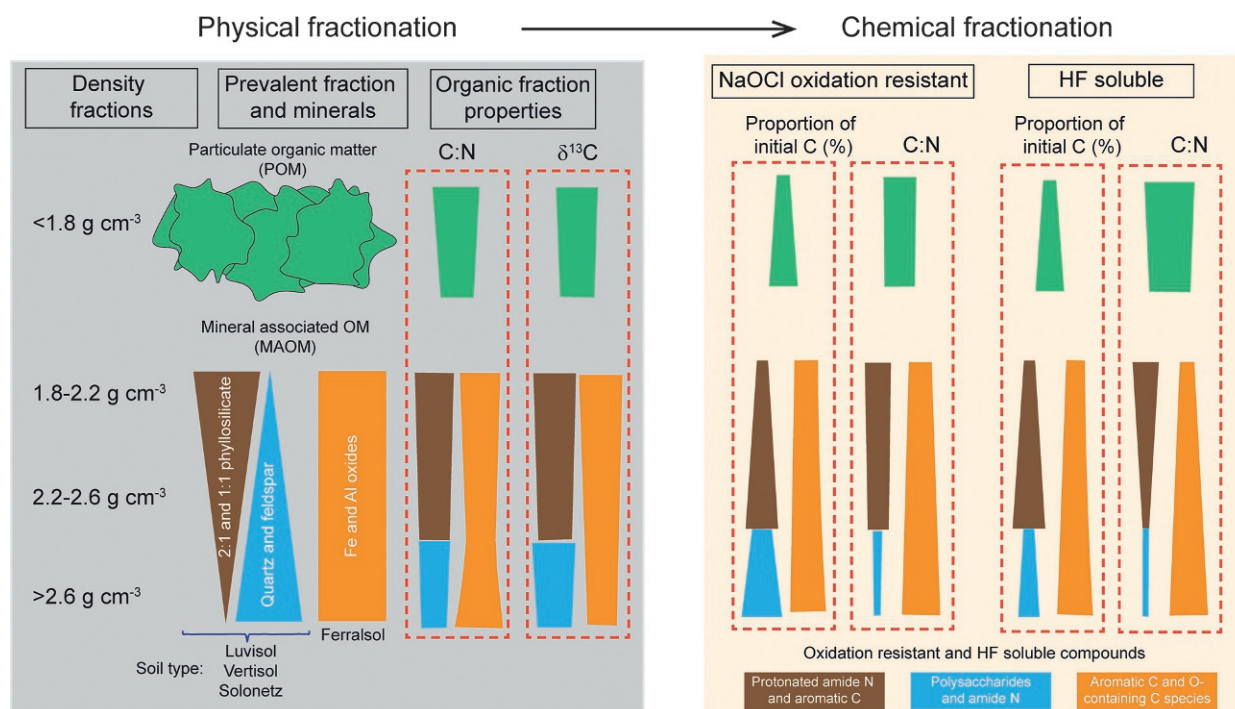


Fig. 7 A conceptual model showing the qualitative and quantitative composition of organic carbon (OC) across different density fractions with varying proportion of phyllosilicates (kaolinite, illite and smectite), Fe or Al oxides (hematite, goethite and gibbsite) and primary minerals (quartz and feldspars) in four soil types, and changes after organic matter oxidation using NaOCl and HF demineralization. By comparing the residuals, it was observed that the size and composition of labile and preserved OC pools were greatly influenced by the abundance and surface properties of soil minerals. The points of cones and arrows indicate the direction of change between density fractions and the thickness of the cones indicate the magnitude of change (increase or decrease). Specific colors assigned to different fractions and the same color represents the same fraction in the properties' panels. From Yeasmin S, Singh B, Johnston CT and Sparks DL (2017a) Organic carbon characteristics in density fractions of soils with contrasting mineralogies. *Geochimica et Cosmochimica Acta* **218**: 215–236.

Characterization of mineral associated organic matter

The chemical composition of OM associated with minerals in the isolated OM fractions and laboratory prepared complexes of organo-mineral complexes can be examined directly using advanced techniques, such as spectroscopic techniques, use of stable isotopes and radiocarbon, and microbial markers. Among the spectroscopic methods, solid state ^{13}C nuclear magnetic resonance (NMR) is considered to be the gold-standard for SOM characterization. In addition to NMR, Fourier transform infrared (FTIR), X-ray photoelectron spectroscopy (XPS) and isotopic ratio mass spectroscopy (IRMS) are widely acknowledged for SOM characterization. Using these techniques, samples can be analyzed without major pre-treatment and extraction.

NMR spectroscopy

In the solid-state ^{13}C NMR spectroscopy, ^{13}C nuclei found in various functional groups of SOM are distinguished on the basis of 'chemical shift' regions of the magnetic field of the spectrometer. Each chemically distinct C functional group has a characteristic chemical shift region, e.g., alkyl C = 0–45 ppm, O/N-alkyl C = 45–110 ppm, aromatic C = 110–160 ppm. By integrating the signal intensity contained within each chemical shift region, the proportion of different C groups can be determined. NMR analysis is very useful in identifying the state of OM biodegradation in the samples using the ratio of alkyl to O-alkyl C and aromaticity. This technique has the potential to distinguish between charred and lignin derived aromatic C, which is not possible with DRIFT spectroscopy. However, this technique has some drawbacks—it is not very sensitive for samples with low OC, it causes low resolution and low signal to noise ratios of spectra, and there can be interference of paramagnetic species present in the soil. The mineral interference in soils can be overcome via dissolution of minerals with HF.

Solid-state ^{13}C NMR spectroscopy is generally used to determine the proportions of the main functional groups of SOMs in whole soil samples and in isolated physical fractions of soils. Poirier et al. (2005) compared the chemical composition of the density separated SOM fractions (free POM, aggregate occluded POM and MAOM) by using a range of spectroscopic methods including NMR (Fig. 8). NMR spectroscopy data showed that free POM was dominated by carbohydrates, while aggregate-occluded POM was dominated by aromatic signals (Fig. 8). Both OM fractions had weaker alkyl and carboxyl signals than MAOM, but exhibited almost similar abundance of N-alkyl, acetal and phenolics. Large contributions of carbohydrates, O-alkyls and aromatics

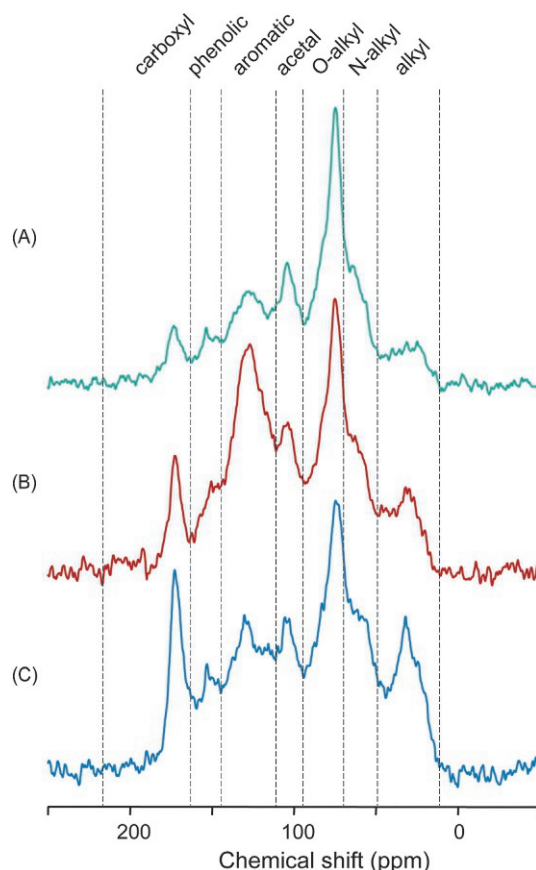


Fig. 8 Solid-state ^{13}C NMR spectra of (A) free particulate organic matter (POM), (B) aggregate-occluded POM and (C) mineral-associated OM fractions of a Chromic Luvisol from the Broadbalk experiment at Rothamsted in the UK. From Poirier N, Sohi SP, Gaunt JL, Mahieu N, Randall EW, Powlson DS and Evershed RP (2005) The chemical composition of measurable soil organic matter pools. *Organic Geochemistry* **36**(8): 1174–1189.

were prominent in the MAOM fractions. The ratio of peak area in the O-alkyl relative to alkyl C regions was 4.06, 3.01 and 0.86 for free POM, aggregate occluded POM and MAOM, respectively, which suggests that MAOM with a smaller ratio has experienced greater decomposition than the other two fractions.

FTIR spectroscopy

FTIR spectroscopy has proven to be an extremely valuable tool in determining SOM chemistry and elucidating binding mechanisms involved in MAOM formations. FTIR spectroscopy uses polychromatic radiation to measure the excitation of molecular bonds and the relative absorbance of different bands provide an index of the abundance of various functional groups. A variety of FTIR techniques have been used to investigate organic-mineral interactions, e.g., transmission FTIR; diffuse reflectance infrared Fourier transform (DRIFT); attenuated total reflectance (ATR)-FTIR. ATR-FTIR spectroscopy is the most powerful technique to study the molecular level interfacial organic chemistry on mineral surfaces due to its high sensitivity to a small molecular change in the medium. Indications of different coordination complexes of organic compounds on mineral surfaces and the types of surface complex structures formed can be obtained from FTIR spectra. Formation of inner-sphere complexes between organics with a carboxylic group and mineral surfaces is often designated by substantial shifts in the frequency of asymmetric and symmetric COO^- stretching relative to a spectrum of the aqueous solution.

For example, an ATR-FTIR study of glutamate sorption on rutile by Parikh et al. (2011) indicated three potential surface configurations: bridging bidentate through inner-sphere coordination of both carboxyl groups, chelating monodentate via inner-sphere coordination of the γ -carboxyl and bridging bidentate with inner-sphere coordination of the α -carboxyl and outer-sphere coordination of the γ -carboxyl. Yeasmin et al. (2014) also made an ATR-FTIR study of carboxylic (citric and oxalic) and amino (glutamic, alanine, phenylalanine and lysine) acid sorption on kaolinite, illite, montmorillonite, ferrihydrite and goethite. Their results indicated that citric and oxalic acids possibly formed monodentate or bidentate inner-sphere, and outer-sphere complexes with Fe oxide and kaolinite surfaces. Iron oxides interacted with amino acids via the carboxylate group, whereas amino groups or both carboxylate and amino groups were involved in the sorption on phyllosilicates. Amino acids sorbed to the minerals mostly via electrostatic interaction or H-bond or Van der Waals interaction.

DRIFT spectroscopy is the most commonly used method for characterization of SOM because it is a quick and nondestructive technique. DRIFT spectral data can also be used to identify and quantify the presence of important organic functional groups such as aliphatic-C, aromatic-C, carboxyl-C in SOM fractions. Recent work has also shown the usefulness of the band ratio approach as an indicator for characterizing the decomposability of the SOM, such as the ratio of carboxylate to polysaccharides and the ratio of aromatic to aliphatic bands. Fractionated samples have been investigated to delineate the chemistry of OC associated with different mineral surfaces. Yeasmin et al. (2017a) found that oxidation resistant and HF soluble OC of the phyllosilicate-dominated fractions were enriched with aromatic-C, polysaccharides and amide-N; OM associated with Fe and Al oxides was predominantly aromatic- and carboxylate-C; and quartz and feldspars associated OM was comprised of N-containing organic compounds and polysaccharides. However, the application of DRIFT spectroscopy to characterize the composition of OM fractions can be compromised by the interference of mineral absorbance in the sample.

X-ray photoelectron spectroscopy

X-ray photoelectron spectroscopy (XPS) provides a direct chemical analysis of mineral surfaces without any extraction procedure. It consists of exposing a surface to a soft X-ray beam (typically ~ 1.5 keV) to induce the ejection of photoelectrons from the atoms present in this surface. The kinetic energy of the ejected electrons is measured, and the electron binding energy of the emitting atom is calculated as the difference between the known energy of the X-ray and the kinetic energy. It records a spectrum of electron binding energy and each peak of the spectrum is characteristic of a given energy level of an element. Due to inelastic scattering of the electron in the sample, XPS provides information about the outermost 2–10 nm layer only. This technique can distinguish between different bonding environments of C, N and O by measuring the shift in the electron binding energy with electron density (oxidation state) of the emitting atom. Jones and Singh (2014) determined carboxylate functional groups and Fe concentration in six density fractions of four soil types and found a strong positive correlation between the abundance of carboxylate functional groups and the Fe concentration in soil density fractions (Fig. 9). The authors speculated that OM is bonded strongly via carboxylate groups to Fe oxides surfaces. These results show the important role of Fe oxides in the preservation of SOM in a variety of soils.

Natural carbon isotope

Since plant residues are the main source of SOC, its isotopic signature resembles the vegetation from which it has been derived. During the microbial decomposition of plant residues, the $\delta^{13}\text{C}$ value increases. Thus, stable isotope analysis ($\delta^{13}\text{C}$ and also $\delta^{15}\text{N}$) acts as a powerful tool to assess the source and also the degree of microbial transformation of SOM or OM fractions in the soil. Generally, larger $\delta^{13}\text{C}$ and $\delta^{15}\text{N}$ values are reported to be associated with longer turnover times for MAOM than the plant residues, which suggest the stabilization of microbially transformed OM on mineral surfaces.

Radiocarbon dating

Radiocarbon dating of SOM can be done by in situ labelling with the 'bomb' ^{14}C approach, based on the enrichment of ^{14}C in the atmosphere due to the atmospheric testing of nuclear weapons in the 1950s and 1960s. The information on the degree to which this 'bomb' ^{14}C is found in SOC helps to differentiate OC pools with different turnover rates, ranging from seasonal to millennial time scales. The ^{14}C contents are calculated from the measured $^{14}\text{C}/^{12}\text{C}$ ratio of the sample compared to that of an international standard

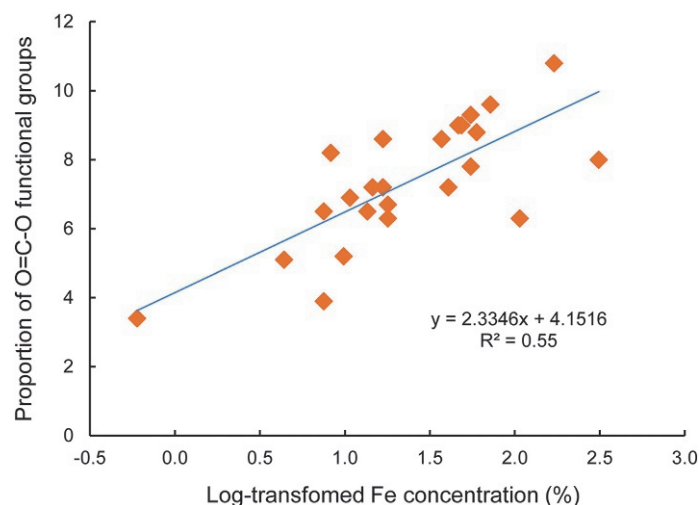


Fig. 9 The relationship between log transformed Fe concentration (% w/w) and proportion of O=C—O functional groups derived from the XPS analysis of six density fractions i.e., <1.6, 1.6–1.8, 1.8–2.0, 2.0–2.2, 2.2–2.6 and >2.6 g cm⁻³ of four contrasting soils from New South Wales, Australia. From Jones E and Singh B (2014) Organo-mineral interactions in contrasting soils under natural vegetation. *Frontiers in Environmental Science* 2: 2.

from 1950. The primary modern radiocarbon standard is oxalic acid I, 95% of the activity of the oxalic acid I is equal to 100 pMC (percent Modern C) or to a hypothetically ¹⁴C-activity from the year 1950. ¹⁴C values greater than 100 pMC indicate the presence of bomb-produced ¹⁴C, and lower values indicate the predominance of C fixed from the atmosphere long enough ago for significant radioactive decay of ¹⁴C to have occurred. The ¹⁴C age of different MAOM fractions may be indicative of the mechanisms that cause OC to persist in soils over millennial time scales.

Generally, MAOM fractions (HF dissolved or oxidation resistant OM) have been found to contain older OC than the free OM (HF resistant or oxidized OM) (Rumpel et al., 2008). However, MAOM fractions do not necessarily always contain old C. For example, a radiocarbon measurement of extracted mineral-associated OM pools found recently deposited plant inputs (modern $\Delta^{14}\text{C}$) in the mineral protected OM (Mikutta et al., 2006). The age of OM may vary depending on the mineral properties and association of OM in organo-metal complexes or mineral-associated OM. Identical age for OM in MAOM and free OM or particulate OM has also been found in soils.

Micro-imaging

Scanning electron microscopy (SEM) and transmission electron microscopy (TEM) are useful techniques for obtaining information about the shape and size of OM and minerals, however, because of the high vacuum and beam damage to samples, these techniques are not useful for characterization of SOM and MAOM. Atomic force microscopy (AFM) can determine the topography of surfaces as well as the bonding type and mechanical properties of the surface. Dynamic force spectroscopy (DFS) uses a functionalized AFM cantilever, which is allowed to interact with a surface so that a bond between the functionalized tip and the surface is formed. This in situ technique has also been used to make quantitative molecular scale measurements of OM bound to a mineral surface (Newcomb et al., 2017). These authors showed that DFS can perform on pristine, flat as well as on heterogeneous mineral surfaces. This technique is suited to measure the binding strength of biomolecules at OM-mineral interfaces even during biological activity, such as in the presence of enzymes or live microbes (Newcomb et al., 2017). Additionally, these authors showed monolayer adsorption of organics on mineral surfaces as the mechanism of organo-mineral associations, which is contrary to the theory of heterogeneously multilayered distribution of OM on mineral surfaces.

Synchrotron-based techniques

Synchrotron-based approaches are useful to study the mineral-OM associations. Scanning transmission X-ray microscope (STXM) combined with near-edge X-ray absorption fine structure (NEXAFS) spectroscopy has been used to explore microscale heterogeneity of OM within a soil aggregate and on mineral surfaces. STXM and NEXAFS spectroscopy were further combined with isotopic sensitive high resolution secondary ion mass spectrometry (SIMS) imaging to investigate the formation, heterogeneity and potential importance of newly formed mineral-OM associations over time (Vogel et al., 2014). Quantitative NanoSIMS results (Vogel et al., 2014) revealed that almost all the OM existed in the clay-sized fraction and was bound in mineral clusters. A preferential association of new OM was found with organo-mineral clusters with rough surfaces containing pre-existing OM. This is confirmed by an increase in area enriched in ¹³C and ¹⁵N, while the total area was covered with a constant amount of OM. Additionally, only a partial percentage of the total clay-sized mineral particles were involved in OM sequestration, which was unexpected for clay-sized mineral particles (illite, mixed-layer clay minerals and pedogenic iron oxides, such as goethite) in this soil with reactive surfaces.

Chen et al. (2014) investigated C associations with Ca, Fe, Al, and Si species in the clay fraction of soils from an upland pasture hillslope using STXM-NEXAFS spectroscopy. Chemical imaging and NEXAFS spectra were obtained using image sequence scans over

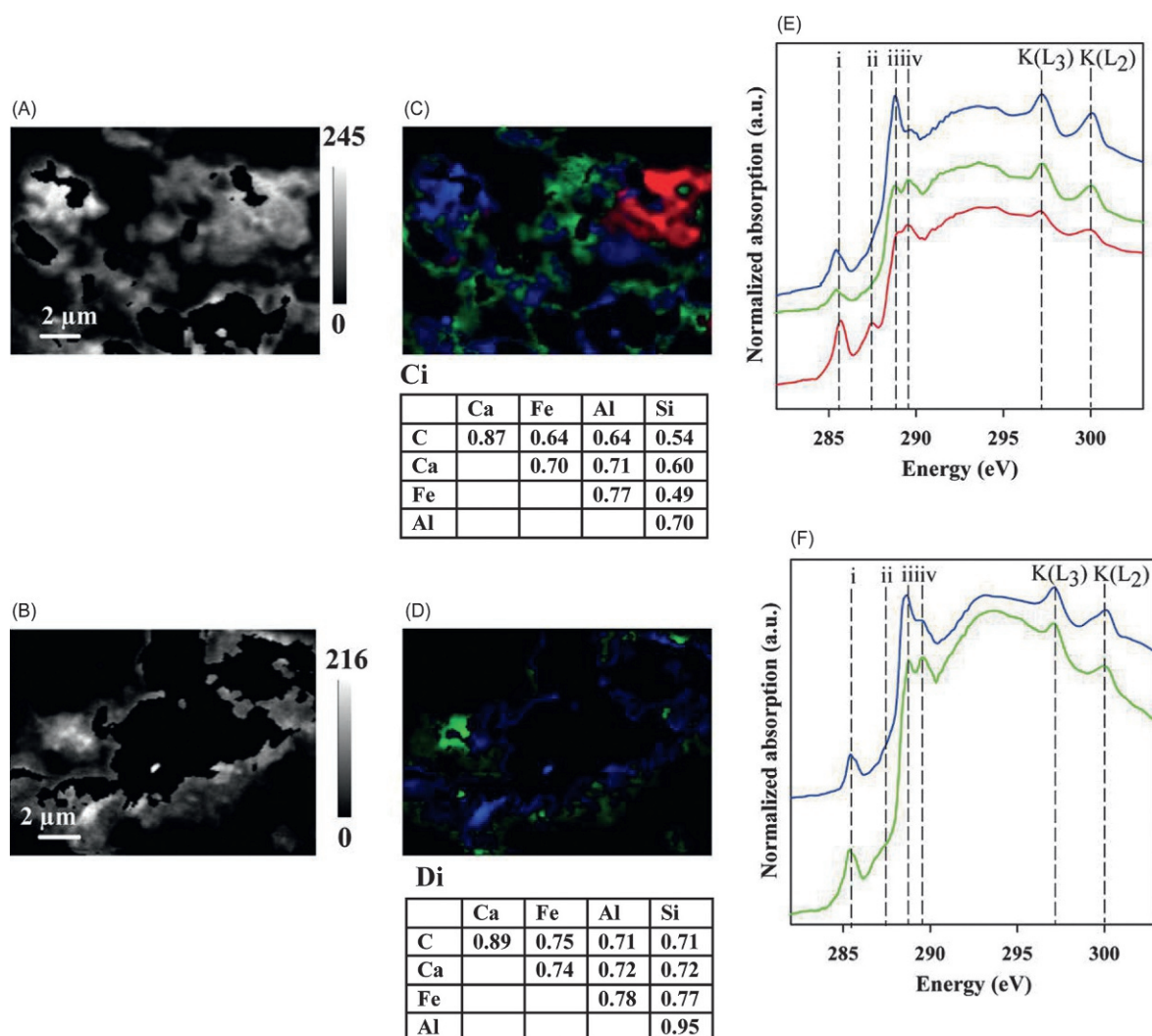


Fig. 10 Carbon image difference maps ($OD_{289}-OD_{282}$) of the thin regions from the (A) summit and (B) footslope soil clay particles. The gray scale indicates thickness in nanometers. C 1s cluster indices map showing the distribution of C functional groups in the thin regions from the (C) summit soil clay particles, with three distinct regions (red, green, and blue) and (D) footslope soil clay particles, with two distinct regions (green and blue). The C 1s NEXAFS spectra were extracted from the regions in the cluster indices maps for the (E) summit and (F) footslope soil clay particles, respectively. Note the color of the spectra and the region's color from which they were extracted are the same. Spectra features identified by the vertical dashed lines correspond to (i) aromatic (C=C) (285.5 eV), (ii) aliphatic C—H (287.4 eV), (iii) carboxylic (COOH) (288.6 eV), and (iv) polysaccharides C—OH (289.5 eV) functional groups. The peaks at higher energies are from K and correspond to its L_3 and L_2 edges. Correlation coefficients values of C with different elements within distribution maps of the summit and footslope soil clay particles are given in Ci and Di, respectively. From Chen C, Dynes JJ, Wang J, Karunakaran C and Sparks DL (2014) Soft X-ray spectroscopy study of mineral-organic matter associations in pasture soil clay fractions. *Environmental Science and Technology* 48(12): 6678–6686.

a range of photon energies covering C 1s, Ca 2p, Fe 2p, Al 1s and Si 1s edges (Fig. 10). The C distribution map was obtained by subtracting the below-edge optical density ($OD = \ln(I_0/I)$, I = transmitted flux through the sample and I_0 = incident flux measured in an empty region adjacent to the analyzed particles) map from its above-edge OD map and showed fine-scale heterogeneity in the compositional chemistry of organic C of summit and footslope soil clays (Fig. 10C and D). Three (red, green, and blue) and two (green, blue) distinct regions were identified in the summit and footslope soils, respectively using principal component analysis (PCA) and cluster analysis (CA). The C NEXAFS analyses show that in both soil clays aromatic-C, carboxyl-C, and polysaccharides were prevalent, with the relative abundance of carboxyl-C and polysaccharides varying spatially at the sub-micrometer scale. Only the clay fraction of the summit soil showed the presence of aliphatic-C, which is indicated by the red region in Fig. 10C and the corresponding spectrum in Fig. 10E shows a peak at 287.4 eV for aliphatic-C. The results further showed that C was associated with Ca, Fe, Al, and Si and there was no discrete phase of C. Carbon was associated with both Fe oxides and aluminosilicate minerals (Chen et al., 2014). The study demonstrates progress being made in the compositional chemistry and associative interactions between OM and natural minerals at the sub-micrometer scale.

Organo-mineral interactions and terrestrial environment

Soil ecosystem services

Soil structural stability

Soil OM turnover is intimately associated with soil structure. The aggregation of fundamental soil particles not only distribute OM heterogeneously, but it also produces various reaction sites and pathways, and environmental conditions in the soil. Soil structure results from the bonding of elementary soil particles by cementing substances such as root exudates, microbial products and other organic substances, Al or Fe-oxides, colloidal silica and calcium carbonates. Both biotic and abiotic processes contribute to soil aggregation and structural stability. Biotic processes, such as plant exudates and plant roots, macrofauna when they ingest organic and mineral soil components together, and microbes by acting on their close organo-mineral environment, all have significant role in soil aggregation. Organo-mineral associations can be considered as structural units of soil aggregates and the nanoparticulate fraction of soil microaggregates. The important role of short-range ordered phases of Fe, Mn, and Al in MAOM is well recognized and thus their role in the formation and stabilization of soil aggregates is central as well. Abiotic C stabilization mechanisms in soil are strongly related to biotic mechanisms. The mechanisms in the formation of micro- and macro-aggregates might be different, which may lead to contrasting OM mineralization rates in soils. Other factors such as existence of different microbial communities in micro- and macro-aggregates, and the differences in the composition of OM and stability of micro- and macro-aggregates can also regulate the mineralization of OM.

Several novel techniques allow insight into mechanisms and processes involved soil aggregation and the formation soil structure, which influences OM mineralization. Imaging techniques like X-ray microtomography (μ CT) can provide details of the physical structure, including pore size distribution, pore connectivity and distances between pores and occluded particulate OM in undisturbed soil, at a spatial resolution of a few microns. Nanoscale secondary ion mass spectrometry (nanoSIMS) and synchrotron based STXM-NEXAFS spectroscopy not only can locate OM at the sub-micrometer scale, but also provide chemical and structural information. These novel techniques, in combination with traditional experimental methods are expected to improve our understanding on the role of soil structure on OM dynamics and incorporating the drivers into emerging C models.

Biochemical cycling of nutrients in soil

Organic matter is a reservoir of nutrients, and its turnover is central to nutrient cycling in the soil. The turnover of OC is often closely associated with the dynamics of N, P and S in the soil. The relationship between nutrients and SOM mineralization is not straightforward, as microbes need nutrients to degrade SOM. Therefore, the addition of nutrients in the soil can result in a decrease or an increase in OM, depending on the initial SOM stoichiometry, the formation MAOM incorporating the microbial products of decomposition, and the simultaneous effects on plant productivity and OM inputs to soils. Overall, net gain or loss in the SOM (and the preservation and availability of nutrients) depends on interactions between the soil mineral matrix, plants, and microbes (Cotrufo et al., 2013). Thus, C:N:P:S stoichiometry is important in SOM cycling and there are suggestions to include this in future C models.

Mobility and bioavailability of heavy metals

Heavy metal pollution has been a global problem which presents a severe threat to the environment and human health. The mobility and biotoxicity of heavy metals depend on their speciation in terrestrial and aquatic environments, particularly their associations with solid fractions of soils including OM. These biotic and abiotic components naturally interact with each other to form MAOM composites, complexes or aggregates involving various processes, such as adsorption, coating and co-precipitation. These interactions have profound control on the speciation, mobility, bioavailability and ultimately the fate of heavy metals in soils and sediments.

Humic acid (HA) and fulvic acid (FA) form complexes with soil minerals via their surface carboxylic and hydroxyl groups, hydrogen bonds and hydrophobic forces. These complex interfacial reactions significantly influence the retention and release of heavy metals in soil environments. Generally, there is increase in the adsorption of heavy metals in the presence of HA and FA. The increased metal adsorption has been attributed to the decreased surface potential of Fe (hydr)oxides induced by HS, changes in affinity of HS for heavy metals or the formation of ternary composites. The interfacial interactions between Fe (hydr)oxides and HA associations strongly influence the immobilization and transport of heavy metals, particularly in acid soils in tropical and subtropical regions, acid sulfate soils and acid mine sediments.

Organic compounds accelerate the adsorption rate of heavy metals and increase the adsorption capacity to the solid assemblages in soils in comparison to pure minerals. For example, Engel et al. (2021) observed up to three times greater adsorption of Cd, Zn and Ni on organic-ferrihydrite assemblages compared to metal adsorption on pristine ferrihydrite (Fig. 11).

Du et al. (2018a, 2018b) investigated the adsorption behavior of Pb and Cd on iron-humic acid composites formed by adsorption (adsorption of OM on pre-synthesized ferrihydrite) and coprecipitation (formation of ferrihydrite in the presence of HA). The ferrihydrite-HA coprecipitated composites have a greater C content and smaller surface area compared to the equivalent adsorbed composites. They found a faster adsorption rate of Pb on co-precipitated ferrihydrite-HA composites than the Fe-OM composites made by OM adsorption. In a study of As(V) adsorption onto Fe-humic composites formed by adsorption and co-precipitation at different Fe:C ratios, the fastest reaction was observed on the co-precipitates formed at high C content, which

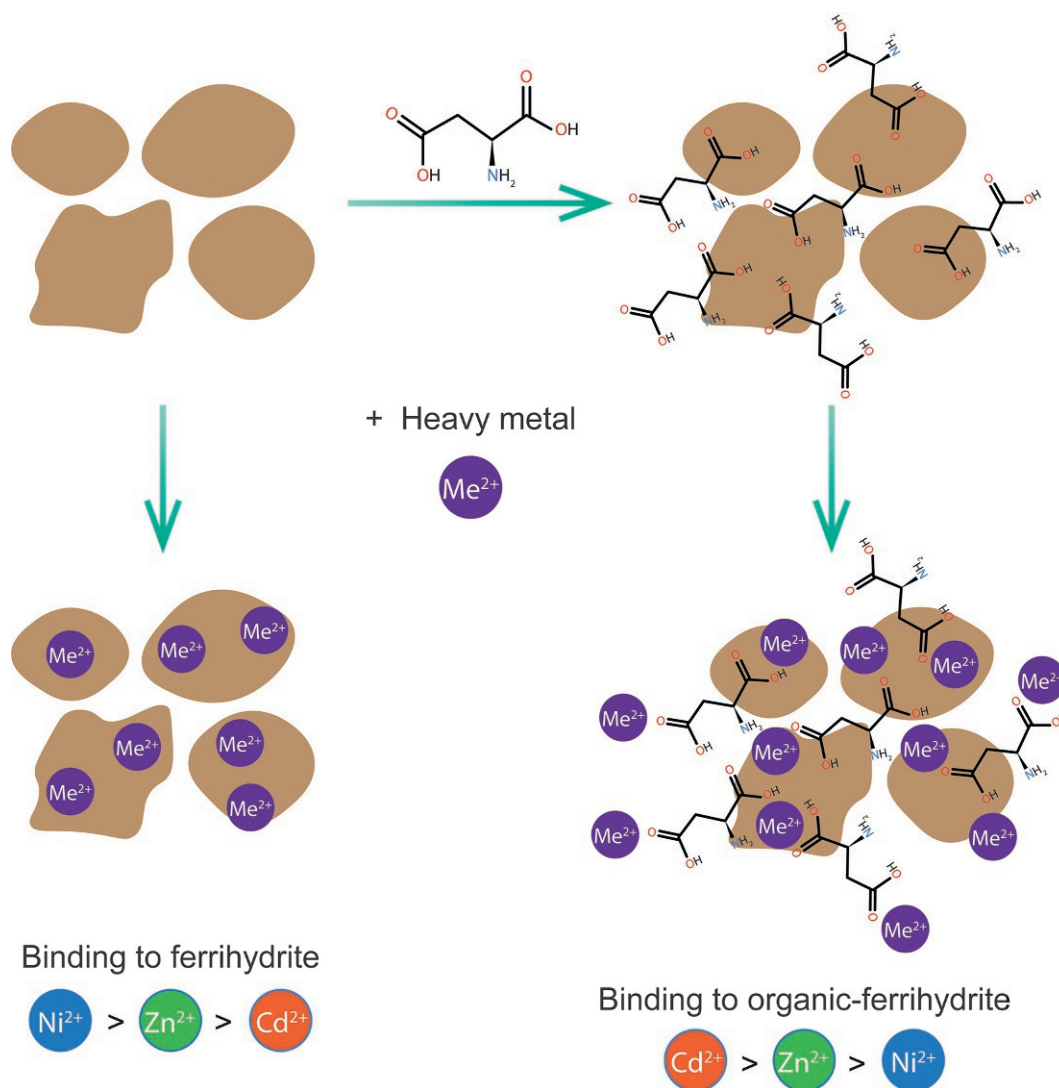


Fig. 11 A schematic representation of the effects of organic compounds on the sorption of heavy metals (Cd, Zn and Ni) on ferrihydrite. The sorption rate and the binding capacity of metals on ferrihydrite increased in the presence of organic compounds. The metal loading order on ferrihydrite was also changed, in pristine ferrihydrite (i.e., without organic compounds) the metal loading followed the order: $\text{Ni} > \text{Zn} > \text{Cd}$, which changed to $\text{Cd} > \text{Zn} > \text{Ni}$ in the presence of organic compound. From Engel M, Pacheco JSL, Noël V, Boye K and Fendorf S (2021) Organic compounds alter the preference and rates of heavy metal adsorption on ferrihydrite. *Science of the Total Environment* **750**: 141485.

was facilitated by desorption of weakly bonded OM and disaggregation of mineral-organo composites (Mikutta et al., 2014). Spectroscopic and surface complexation analyses suggest that As is largely adsorbed by forming inner-sphere complexes on Fe (hydr)oxide-HS composites. There is ample evidence of enhanced adsorption of heavy metals on HS adsorbed phyllosilicates.

Climate change

Climate change and soil warming can accelerate microbial SOM decomposition rates causing increasing CO_2 concentration in the atmosphere. A meager reduction in soil C stocks is likely to have a drastic impact on the global C cycle and climate change. Approximately one third of the world's soil C stock exist in permafrost soils of the Arctic, which are perhaps most vulnerable to the projected global warming (Schuur et al., 2015). There are contrasting observations about the effects of increasing temperature and CO_2 concentration in the atmosphere on Arctic soil C, with modelling showing increased C sequestration (McGuire et al., 2018) and accelerated decomposition of SOM (Commans et al., 2017). These contradictory modelling results highlight knowledge gaps on the effects of climate change on a highly significant pool of the global soil C stock. In these models many important aspects are

not considered, for example, the transition from unprotected to mineral-protected soil C pools under warming, the effects of root exudates on MAOM, and the role of Fe and Al oxides in stabilizing SOM in deeper soils. Also, changed soil environmental conditions, particularly soil thermal and hydraulic changes, will disrupt the primary mechanism (freezing) for SOM persistence in the Arctic region. It remains uncertain how under changed conditions MAOM may attenuate the extent and rate of permafrost SOM decomposition following thaw.

Some earlier studies in soils from regions other than Arctic have suggested that the rise in CO₂ will lead to increases in plant biomass, and hence terrestrial C storage. However, these are an oversimplification, since belowground C storage dominates in some terrestrial ecosystems. Elevated CO₂ concentration in the atmosphere can cause complex effects on OM, including interacting belowground responses. The indirect effects of elevated CO₂ on the plant community, such as changes in quantity and quality of plant litter inputs and root exudation determine SOM responses, and these processes could affect POM and MAOM differently. Drake et al. (2011) showed that under elevated CO₂ enhanced rates of net primary production (NPP) in a forest ecosystem were sustained by a C-cascade through the root-microbe-soil system. The flux of C belowground increased under elevated CO₂ stimulated microbial activity, which accelerated the rate of SOM decomposition and stimulated tree uptake of N bound to this SOM. This process propelled and maintained greater C gain under elevated CO₂ because of increases in canopy N content and greater photosynthetic N-use efficiency. Increased root exudation under elevated CO₂ may increase the mineralization of native SOM or cause destabilization of MAOM (Phillips et al., 2010), thus may prevent soil C accumulation. More research is needed to get a clear understanding of the effects of elevated CO₂ on SOM accumulation under different land use systems.

Since a significant portion of the SOC is present in the MAOM fraction, a clear understating of the expected temperature increase on the stability of C in this fraction is required. Also, SOC sequestration via MAOM is a potential pathway to mitigate the effects of increased soil temperature and to reduce CO₂ emissions from soils. Lavallee et al. (2019) argued that targeted investigations on the direct effects of warming on the persistence and formation of POM and MAOM are needed to elucidate specific mechanisms of change. Such studies should be conducted in intact ecosystems so as to quantify the additive direct and indirect effects of warming on POM and MAOM and ultimately predict future SOM stocks with climate change.

Conclusions and the future

Organic C associated with soil minerals is relatively stable, and controls the long-term sequestration of SOM. The emerging view is that OM in MAOM largely consists of microbial residues and dead microbial cells or necromass, which are adsorbed on to minerals. In addition to chemical interactions, OM can be trapped into soil aggregates, which hinders access to soil microbes and thus remain protected. The type of adsorption mechanism involved and associated strength to protect the OM from microbial decomposition are governed by several factors including the type and properties of minerals present in the soil, the quantity and quality of OM, microbial decomposition products and soil environmental conditions.

Organo-mineral association in the soil is an important biogeochemical process that exerts influence on soil ecosystem services, such as, the cycling and preservation of soil nutrients, soil structure, and mobility and bioavailability of heavy metals. The preservation of OC in MAOM is acknowledged as a potential pathway to reduce CO₂ emissions and control global C cycle. However, the stability of C in MAOM in different soil types, land use and climatic conditions is not clear and current models provide conflicting results for projected global warming and climate change. Therefore, insights into mineral-OM interactions at molecular scale and stability of MAOM to microbial mineralization are crucial to manage the SOM and regulate the CO₂ emission from the soil. Modern techniques, such as nanoSIMS, STXM and NEXAFS, along with traditional techniques are being used investigate the key questions in relation to the stability and formation of MAOM. Often, physically fractionated pools are used to determine the spatial arrangement of primary and secondary MAOM; and chemically fractionated pools to elucidate molecular scale interactions of OM-mineral. The combined fractionation (physical + chemical) procedures are more effective than a single method and facilitate both spatial and molecular information. However, there is a lack of consistent approach in the fractionation procedures and subsequent characterization which is an hinderance to obtain consistent data for a range of soil systems. Most studies have been focused on mineral-OM-microbial interactions in surface soils, however, a significant portion of OM occurs in deeper soil horizons that needs more attention. Many investigations of the stability of OM associated with minerals have been laboratory based. It is important to evaluate mineral-OM-microbial interactions and the stability of C in MAOM in the rhizosphere of living plants. Such research may provide further insights into microbial contribution to organo-mineral association.

References

- Chen C, Dynes JJ, Wang J, Karunakaran C, and Sparks DL (2014) Soft X-ray spectromicroscopy study of mineral-organic matter associations in pasture soil clay fractions. *Environmental Science and Technology* 48(12): 6678–6686.
- Commane R, Lindaas J, Benmergui J, Luus KA, Chang RYW, Daube BC, Euskirchen ES, Henderson JM, Karion A, Miller JB, and Miller SM (2017) Carbon dioxide sources from Alaska driven by increasing early winter respiration from Arctic tundra. *Proceedings of the National Academy of Sciences* 114(21): 5361–5366.
- Cotrufo MF, Wallenstein MD, Boot CM, Deneff K, and Paul E (2013) The microbial efficiency-matrix stabilisation (MEMS) framework integrates plant litter decomposition with soil organic matter stabilization: Do labile plant inputs form stable organic matter? *Global Change Biology* 19(4): 988–995.
- De Nobili M, Bravo C, and Chen Y (2020) The spontaneous secondary synthesis of soil organic matter components: A critical examination of the soil continuum model theory. *Applied Soil Ecology* 154: 103655.

- Dong H (2012) Clay–microbe interactions and implications for environmental mitigation. *Elements* 8(2): 113–118.
- Drake JE, Gallet-Budynek A, Hofmøckel KS, Bernhardt ES, Billings SA, Jackson RB, Johnsen KS, Lichter J, McCarthy HR, McCormack ML, and Moore DJ (2011) Increases in the flux of carbon belowground stimulate nitrogen uptake and sustain the long-term enhancement of forest productivity under elevated CO₂. *Ecology Letters* 14(4): 349–357.
- Du H, Huang Q, Lei M, and Tie B (2018a) Sorption of Pb (II) by nanosized ferrihydrite organo-mineral composites formed by adsorption versus coprecipitation. *ACS Earth and Space Chemistry* 2(6): 556–564.
- Du H, Peacock CL, Chen W, and Huang Q (2018b) Binding of Cd by ferrihydrite organo-mineral composites: Implications for Cd mobility and fate in natural and contaminated environments. *Chemosphere* 207: 404–412.
- Engel M, Pacheco JSL, Noël V, Boye K, and Fendorf S (2021) Organic compounds alter the preference and rates of heavy metal adsorption on ferrihydrite. *Science of the Total Environment* 750: 141485.
- Fujisaki K, Chapuis-Lardy L, Albrecht A, Razafimbelo T, Chotte JL, and Chevallier T (2018) Data synthesis of carbon distribution in particle size fractions of tropical soils: Implications for soil carbon storage potential in croplands. *Geoderma* 313: 41–51.
- Islam MR, Singh B, and Dijkstra FA (2022) Stabilisation of soil organic matter: Interactions between clay and microbes. *Biogeochemistry*. <https://doi.org/10.1007/s10533-022-00956-2>.
- Jones E and Singh B (2014) Organo-mineral interactions in contrasting soils under natural vegetation. *Frontiers in Environmental Science* 2: 2.
- Kaiser K and Guggenberger G (2003) Mineral surfaces and soil organic matter. *European Journal of Soil Science* 54: 219–236.
- Keiluweit M, Wanzek T, Kleber M, Nico P, and Fendorf S (2017) Anaerobic microsites have an unaccounted role in soil carbon stabilization. *Nature Communications* 8(1): 1–10.
- Kleber M, Eusterhues K, Keiluweit M, Mikutta C, Mikutta R, and Nico PS (2015) Mineral-organic associations: Formation, properties, and relevance in soil environments. *Advances in Agronomy* 130: 1–140.
- Kögel-Knabner I, Guggenberger G, Kleber M, Kandeler E, Kalbitz K, Scheu S, Eusterhues K, and Leinweber P (2008) Organo-mineral associations in temperate soils: Integrating biology, mineralogy, and organic matter chemistry. *Journal of Plant Nutrition and Soil Science* 171: 61–82.
- Lavallee JM, Soong JL, and Cotrufo MF (2019) Conceptualizing soil organic matter into particulate and mineral-associated forms to address global change in the 21st century. *Global Change Biology* 26: 261–273.
- Lehmann J and Kleber M (2015) The contentious nature of soil organic matter. *Nature* 528: 60–68.
- McGuire AD, Lawrence DM, Koven C, Clein JS, Burke E, Chen G, Jafarov E, MacDougall AH, Marchenko S, Nicolisky D, and Peng S (2018) Dependence of the evolution of carbon dynamics in the northern permafrost region on the trajectory of climate change. *Proceedings of the National Academy of Sciences* 115(15): 3882–3887.
- Mikutta R, Kleber M, Kaiser K, and Jahn R (2005) Review: Organic matter removal from soils using hydrogen peroxide, sodium hypochlorite, and disodium peroxodisulfate. *Soil Science Society of America Journal* 69: 120–135.
- Mikutta R, Kleber M, Torn MS, and Jahn R (2006) Stabilization of soil organic matter: Association with minerals or chemical recalcitrance? *Biogeochemistry* 77(1): 25–56.
- Mikutta R, Lorenz D, Guggenberger G, Haumaier L, and Freund A (2014) Properties and reactivity of Fe-organic matter associations formed by coprecipitation versus adsorption: Clues from arsenate batch adsorption. *Geochimica et Cosmochimica Acta* 144: 258–276.
- Mikutta R, Turner S, Schippers A, Gentsch N, Meyer-Stüve S, Condron LM, Peltzer DA, Richardson SJ, Eger A, Hempel G, and Kaiser K (2019) Microbial and abiotic controls on mineral-associated organic matter in soil profiles along an ecosystem gradient. *Scientific Reports* 9(1): 1–9.
- Newcomb CJ, Qafoku NP, Grate JW, Bailey VL, and De Yoreo JJ (2017) Developing a molecular picture of soil organic matter–mineral interactions by quantifying organo-mineral binding. *Nature Communications* 8(1): 1–8.
- Parikh SJ, Kubicki JD, Jonsson CM, Jonsson CL, Hazen RM, Sverjensky DA, and Sparks DL (2011) Evaluating glutamate and aspartate binding mechanisms to rutile (α -TiO₂) via ATR-FTIR spectroscopy and quantum chemical calculations. *Langmuir* 27: 1778–1787.
- Perdrial JN, Warr LN, Perdrial N, Lett MC, and Elsass F (2009) Interaction between smectite and bacteria: Implications for bentonite as backfill material in the disposal of nuclear waste. *Chemical Geology* 264(1–4): 281–294.
- Phillips RP, Finzi AC, and Bernhardt ES (2010) Enhanced root exudation induces microbial feedbacks to N cycling in a pine forest under long-term CO₂ fumigation. *Ecology Letters* 14: 187–194.
- Poepelau C, Don A, Six J, Kaiser M, Benbi D, Chenu C, Cotrufo MF, Derrien D, Gioacchini P, Grand S, Gregorich E, et al. (2018) Isolating organic carbon fractions with varying turnover rates in temperate agricultural soils—A comprehensive method comparison. *Soil Biology and Biochemistry* 125: 10–26.
- Poirier N, Sohi SP, Gaunt JL, Mahieu N, Randall EW, Powlson DS, and Evershed RP (2005) The chemical composition of measurable soil organic matter pools. *Organic Geochemistry* 36(8): 1174–1189.
- Rumpel C, Chaplot V, Chabbi A, Largeau C, and Valentin C (2008) Stabilisation of HF soluble and HCl resistant organic matter in sloping tropical soils under slash and burn agriculture. *Geoderma* 145: 347–354.
- Samuels T, Bryce C, Landenmark H, Marie-Loudon C, Nicholson N, Stevens AH, and Cockell C (2020) Microbial weathering of minerals and rocks in natural environments. In: Katerina Dontsova K, Balogh-Brunstad Z, and Le Roux G (eds.) *Biogeochemical Cycles: Ecological Drivers and Environmental Impact*, *Geophysical Monograph* 251, 1st edn, 59–79. American Geophysical Union. Published 2020 by John Wiley & Sons, Inc: New York.
- Schneckenburger T, Riefstahl J, and Fischer K (2018) Adsorption of aliphatic polyhydroxy carboxylic acids on gibbsite: pH dependency and importance of adsorbate structure. *Environmental Sciences Europe* 30(1): 1–15.
- Schuur EAG, McGuire AD, Schädel C, Grosse G, Harden JW, Hayes DJ, Hugelius G, Koven CD, Kuhry P, Lawrence DM, Natali SM, Olefeldt D, Romanovsky VE, Schaefer K, Turetsky MR, Treat CC, and Vonket JE (2015) Climate change and the permafrost carbon feedback. *Nature* 520(7546): 171–179.
- Sodano M, Lerda C, Nisticò R, Martin M, Magnacca G, Celi L, and Said-Pullicino D (2017) Dissolved organic carbon retention by coprecipitation during the oxidation of ferrous iron. *Geoderma* 307: 19–29.
- Sollins P, Kramer MG, Swanston C, Lajtha K, Filley T, Aufdenkampe AK, Wagai R, and Bowden RD (2009) Sequential density fractionation across soils of contrasting mineralogy: Evidence for both microbial- and mineral-controlled soil organic matter stabilization. *Biogeochemistry* 96: 209–231.
- Vogel C, Mueller CW, Höschen C, Buegger F, Heister K, Schulz S, Schlöter M, and Kögel-Knabner I (2014) Submicron structures provide preferential spots for carbon and nitrogen sequestration in soils. *Nature Communications* 5(1): 1–7.
- von Lützow M, Kögel-Knabner I, Ekschmitt K, Flessa H, Guggenberger G, Matzner E, and Marschner B (2007) SOM fractionation methods: Relevance to functional pools and to stabilization mechanisms. *Soil Biology and Biochemistry* 39: 2183–2207.
- Yasmin S, Singh B, Kookana RS, Farrell M, Sparks DL, and Johnston CT (2014) Influence of mineral characteristics on the retention of low molecular weight organic compounds: A batch sorption-desorption and ATR-FTIR study. *Journal of Colloid and Interface Science* 432: 246–257.
- Yasmin S, Singh B, Johnston CT, and Sparks DL (2017a) Organic carbon characteristics in density fractions of soils with contrasting mineralogies. *Geochimica et Cosmochimica Acta* 218: 215–236.
- Yasmin S, Singh B, Johnston CT, and Sparks DL (2017b) Evaluation of pre-treatment procedures for improved interpretation of mid infrared spectra of soil organic matter. *Geoderma* 304: 83–92.

Habitat for soil organisms and their functions

Alexandra Kravchenko and Andrey Guber, Department of Plant, Soil, and Microbial Sciences, Michigan State University, East Lansing, MI, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	407
Abiotic conditions	408
Solid phase	408
Light and temperature	409
Water	410
pH	411
Biotic conditions	412
Soil porosphere	413
Special habitats: Rhizosphere and detritusphere	415
Rhizosphere	415
Detritusphere	415
References	416
Further reading	417

Abstract

Soil is a major global terrestrial ecosystem that offers the greatest variability of environments for the organisms populating it, ensuring the Earth's greatest biodiversity. Focusing on micro-environmental conditions, this chapter examines the importance of soil solid phase and pore structure on water and gas availability to soil organisms, along with importance of soil pH and other chemical characteristics. The vicinity and surfaces of pores is where most of soil microorganisms and micro- meso-fauna live, thus, properties and habitat value of different pore classes, as well as rhizosphere and detritusphere, the two soil hotspots of biological activity, are separately discussed.

Key points

- Soil offers the most diverse range of habitats among terrestrial ecosystems.
- Solid phase provides building blocks of the soil matrix, largely defining all other physical habitat characteristics.
- Distributions of water, O₂, pH, and organic matter within the soil matrix are the highly spatially and temporally variable drivers of soil biodiversity.
- Porosphere is where most soil organisms live and its special water and gas holding and transporting characteristics drive the life strategies of soil biota.
- Rhizosphere and detritusphere are two special "hotspot" habitats substantially differing in their environmental conditions from the bulk soil matrix.

Introduction

Habitat can be loosely defined as a space with a certain assemblage of biotic and abiotic characteristics enabling organisms to sustain themselves. Soil is one of the major global terrestrial habitats, hosting the greatest diversity of living organisms. As a layer of interaction between the Earth's crust, biosphere, and atmosphere, soil is uniquely different from any other habitat on Earth (Voroney, 2007).

Specific properties of soil habitats develop during soil formation, a process driven by five key factors: parent materials, climate, living organisms, topography, and time. Some of the characteristics of present-day soils that define their habitat features took hundreds to thousands of years to form, e.g., soil texture, mineralogy, and their vertical distribution down the soil profile. However, characteristics, such as soil organic matter content and soil structure can change much more rapidly, and thus, are greatly susceptible to the impact of what can be viewed as the sixth factor of soil formation – the anthropogenic influence.

When considering the soil as a habitat, it is important to keep in mind the vast diversity of soil inhabitants in terms of their size, life strategy, place in the food web, and ecosystem role. Soil organisms can range in body width from 1 to 2 µm for most bacteria,

10–100 μm for microfauna, 100–2000 μm mesofauna, and $>2000 \mu\text{m}$ for macrofauna. Soil megafauna is represented by vertebrates and large invertebrate organisms $>1 \text{ cm}$ in body length; however, some, such as the giant earthworm *Megascolides australis*, can grow to 1–3 m in length.

Soil as a habitat is perceived differently by the various groups of soil biota: for large borrowing animals, for small borrowing animals (earthworms), and for non-architectural soil meso- and micro-biota. The former two groups build their living space within the soil. The latter group are the organisms that rely on the existing physical framework of the soil space, which they populate but are not able to modify through direct physical actions. Of course, all organisms, especially the two latter groups, have a very substantial indirect effect on the soil formation and are continuously modifying many characteristics of soil habitats.

Soil microorganisms are among the most diverse soil inhabitants. They include all possible variants of food procurement: photoautotrophs, chemoautotrophs, and heterotrophs of all kinds. Remarkably diverse in environmental preferences, soil microorganisms can live in aerobic or anaerobic, wet or dry, cold or warm conditions; and include an equally remarkable abundance of facultative organisms with high versatility in their environmental preferences. This vast diversity has formed in response to a very wide range of conditions offered by the soil's living space to its inhabitants, and persistent soil heterogeneity continues to support this diversity. The biotic and abiotic conditions and resources that a soil environment can provide to one group of organisms might be completely different from what is available to another, even when the two are separated by just a few centimeters of physical space.

The following sections provide only a very general outline of the key characteristics to be considered when looking at what soil can potentially offer a particular organism. The primary focus is on the aspects of soil habitats that might be of key importance to microorganisms and micro- and mesofauna, only briefly touching on selected items of relevance to macrofauna and excluding the vertebrates all together. Habitat requirements of the vertebrates differ too much from those of the small soil dwellers to be discussed in sufficient detail in this article; they deserve a distinct article of their own!

Abiotic conditions

Solid phase

The soil solid phase plays a major role in determining properties and functions of soil habitats. This is what provides the physical scaffolding and building blocks of the soil matrix, largely defining all other physical characteristics. Yet, it is also the driver of soil chemical characteristics and the source of nutrients for plants and for many of the soil organisms. The solid phase consists of mineral grains of varying sizes interlaced with non-particulate soil organic matter (SOM) and mixed with particulate organic matter (POM), i.e., sizeable ($>0.053 \text{ mm}$) fragments of organic residues at various stages of decomposition (Fig. 1A).

The basic physical characteristic of the mineral portion of the solid phase of a soil is its particle size distribution, also known as its texture. Soil texture is commonly viewed as a composition of three particle-size categories: clay, silt, and sand. Proportions of these three key constituents, and especially, their spatial arrangement, define the structure of the soil's pore space, including pore size and shape distributions, pore connectivity, and tortuosity. They also largely determine the size and properties of soil surfaces. Pore volume and pore surface area form the dwelling spaces for most of the soil organisms and constitute a fundamental basis for the soil habitat.

The solid phase is an outcome of weathering of primary rocks, and the clay fraction is the common intermediate outcome of such weathering. Among the three particle-size categories, the surface area of clay particles is the largest due to its plate structure and spans from 20 to 40 $\text{m}^2 \text{ g}^{-1}$ in kaolinite to 500–800 $\text{m}^2 \text{ g}^{-1}$ in smectite. For comparison, the surface area in coarse sands is less than 0.1 $\text{m}^2 \text{ g}^{-1}$. The importance of clays is not just in their large surface areas: clay surfaces are also typically negatively charged, attracting and holding organic compounds, metal oxides and hydroxides, cations, extracellular enzymes of plants and microorganisms, and serving as the places of attachment for microorganisms themselves. Sizable surface areas of the soil solid phase, being largely the outcome of clay content, stimulate growth of microbial populations. Clays are often found to be associated with extracellular polymeric substances together forming particularly favorable environments for archaea and bacteria (Fig. 1B).

Soil solids are not just a physical environment but a source of nutrients for many organisms. Some organisms rely on procurement of specific nutrients from surfaces of soil minerals while others utilize SOM and POM. The attractiveness of the surfaces of soil particles as a habitat may vary for different organisms depending on their nutrient needs. For example, since NH_4^+ is often strongly absorbed by soil surfaces, ammonium oxidizing organisms tend to live in stronger contact with the surfaces compared to most of the heterotrophic bacteria.

As a contributor to soil habitat, SOM plays a unique and multifaceted role. It is the key gluing agent causing formation of soil structure, driving properties and connectivity of soil pores. At the same time, SOM decomposition is the main source of energy for soil microbes, as well as food for micro- and mesofauna organisms. Variability of SOM content at large spatial scales, such as across agricultural fields, is typically relatively modest; thus, it is often not sufficiently appreciated that SOM variability can be enormous at scales of few microns to few millimeters within the soil matrix. Indeed, SOM concentrations might vary from 2 to 10% in soils across a temperate forest watershed, yet, across a 1 mm distance within the soil matrix of any single spot within that watershed, SOM can span from a virtual zero to 100%. Thus, of two 1–2 μm sized bacteria separated by a 1 mm distance within the soil matrix, one might be enjoying an unlimited carbon and nutrient supply, while the other is living in an extremely inhospitable and resource poor

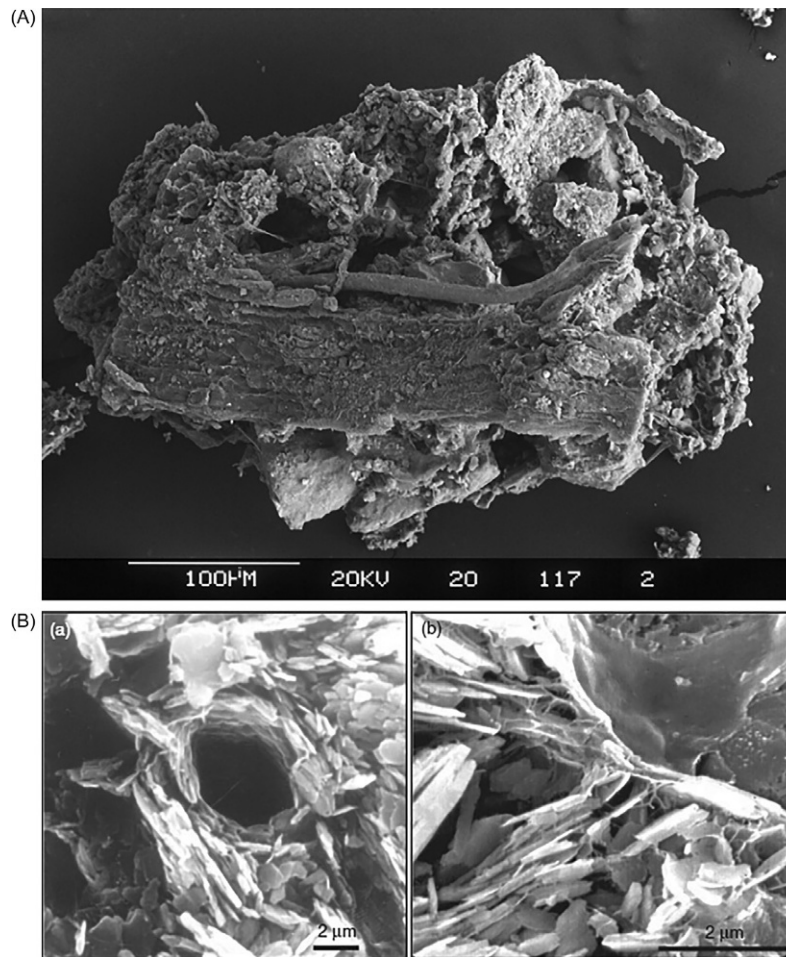


Fig. 1 (A) An intact soil aggregate formed by mineral grains tied around a fragment of POM, used as a living space by a fungal hypha. (B) Interactions between bacteria and platelets of soil clay. (A) From Fig. 3.4 in Coleman DC, Callahan MAJ, and Crossley DAJ (2018) *Fundamentals of Soil Ecology*. 3rd edn, Elsevier Inc. (B) From Fig. 3.1. in Chenu C and Cosentino D (2011) Microbial regulation of soil structural dynamics. *The Architecture and Biology of Soils: Life in Inner Space*. Wallingford, UK: CABI, 37–70.

environment without any means to move to a better place. It is that microscale variability, together with a number of other sources of soil micro-heterogeneity, what results in an enormous diversity of soil microorganisms.

Light and temperature

Availability of light affects only the organisms living on the soil surface, in the top couple of millimeters of the profile, or in occasional wide cracks. It is of only minor relevance for soils of most ecosystems, with soil crusts of arid regions being one notable exception.

Temperature is as crucial a component in defining soil habitats as it is for any other terrestrial ecosystem, yet its importance is modulated by soil characteristics. The buffering of temperature fluctuations within the soil is greatly enhanced under wet conditions due to the large specific heat of water. Thus, spatial patterns in variability of soil temperature across landscapes typically follow the spatial distributions of soil water, with higher temperatures being associated with the soils of dry upslope locations and colder soils in wet depressions.

Only in the top few centimeters of the soil profile are organisms directly exposed to the atmospheric temperature fluctuations and, at times, the temperatures here can be more extreme than those in the air immediately above. Owing to the small heat conductivity of soil, at greater depths the temperature is less temporally variable and is more moderate than daily and seasonal aboveground extremes. The deeper within the soil profile the less variable are the temperatures, and the more insulated are the inhabitants from rapid temperature changes, including rapid freezing events. Reduction in temperature variability with depth is counted on in the life strategies of many soil mesofauna and microorganisms. Yet for many organisms inhabiting the top 5–10 cm of the soil profile, e.g., nematodes and microarthropods, aboveground factors mediating temperature fluctuations in the topsoil can

be of major importance. For example, the type of vegetation cover, such as a shady forest vs. an open pasture, can substantially influence warm season soil temperatures, while snow cover can restrict cold season temperature drops.

Water

The availability of rainfall water to soil biota is greatly affected by a multitude of topographical, vegetative, and edaphic factors. Only a portion of precipitated water infiltrates into the soil. The exact quantities are a function of a number of relatively stable environmental characteristics, including topography, vegetation, and soil properties, as well as some highly spatially and temporally variable features, e.g., antecedent soil water content and characteristics of individual precipitation events. Depending on rainfall characteristics, soils of runoff-prone areas in topographical depressions can receive an order of magnitude or more of the fallen precipitation as compared with hillslopes. The outcome is a substantial spatial variability in soil moisture distribution across topographically diverse landscapes with far-reaching effects on soil inhabitants. It drives spatial patterns in surface soil temperatures, redistribution of nutrients carried with surface and subsurface flow (such as NO_3^-), and of soil physical characteristics due to transport of soil particles, SOM, and POM with surface runoff.

The water that does infiltrate into the soil is subject to a multitude of influences that further enhance variability in its spatial distribution and role in soil habitat formation (Fig. 2). First, the incoming water rarely travels through the soil profile as a uniform front, instead, a large portion of it quickly moves through via large pores or established flow paths in a phenomenon called preferential flow (primary paths, Fig. 2). In some instances, as much as 90% of the incoming water can be directly transported into the subsoil by preferential flow paths, leaving the bulk of the topsoil relatively dry. The distribution of the water that does enter the soil matrix is also highly heterogeneous. In relatively dry finer-textured soils the water is quickly absorbed into the fine pores of the matrix (tertiary paths, Fig. 2), leaving large pores empty. In wet conditions water flow takes place through non-preferential paths of both large and small pores of the soil matrix. Besides the influence of antecedent water content, the rates and volumes of water fluxes are driven by soil texture and structure, while the magnitude of the preferential flow into deeper soil horizons is also affected by the quantity and intensity of surface water supply via precipitation and runoff.

Water flow pathways play an important role in dispersion of microfauna and microorganisms through the soil and affects their spread to new regions within the soil profile vertically, as well as laterally. The rate of water flow through the soil matrix under unsaturated conditions is significantly (orders of magnitude) slower than the saturated one. The unsaturated flow is driven mostly by capillary forces when water moves along the gradients of matric potential. While slow and short-distanced as compared to the saturated flow, it provides yet another important mode of transport and means for dispersal for bacteria, viruses, and similarly sized organisms. They can be absorbed and tightly held on the air-water interfaces, that are ubiquitous in unsaturated soil, and be carried along as the capillary flow moves such interfaces across the soil matrix (Mills, 2003). Their own motility can take such organisms only over very short distances and even then only when the right water conditions, such as large, deep, connected water menisci, are present.

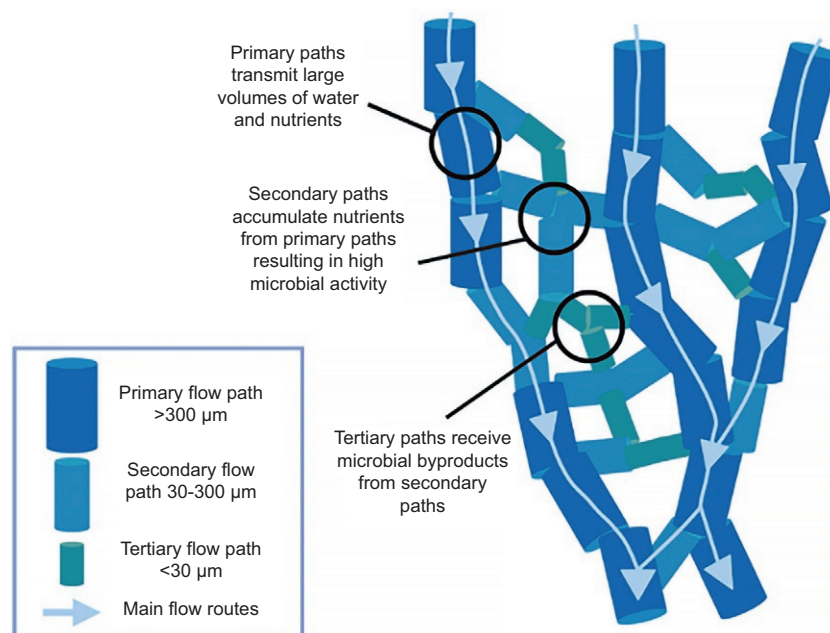


Fig. 2 A conceptual model of the three types of water flow paths within the soil matrix highlighting their contributions to flow and transport processes (Franklin et al., 2021). From Franklin SM, Kravchenko AN, Vargas R, Vasilas B, Fuhrmann JJ, and Jin Y (2021) The unexplored role of preferential flow in soil carbon dynamics. *Soil Biology & Biochemistry* 161.

In soils of humid regions, wet conditions prevail and the soils are often at either fully saturation or at field capacity, i.e., a state when gravimetric forces drained the largest of pores. There the large pores are filled as well, yet in a very uneven manner, with menisci and partially filled spaces. Recently, with advances in X-ray and neutron computed micro-tomography imaging, the magnitude of micro-scale variability in water distribution within the soil matrix started to be fully appreciated. Air-pockets within apparently fully-saturated soil, as well as water-pockets within large pores of an unsaturated soil, might be a rule rather than an exception, as opposed to an expectation of relatively evenly distributed water menisci within the soil matrix in a manner consistent with capillary theory. High variability in water distribution suggests that even when soil becomes unsaturated the water bodies sufficiently large to support movement in water might persist for some periods of time. On the other hand, soils of dry and temperate regions are below field capacity either year-round or most of the time during the warm season. In such settings there will be hardly any water menisci or other connected water bodies big enough to sustain motility of microorganisms. Thus, only the pore surface provides the lasting stable environment in the form of water films or menisci of varying thickness, where microorganisms can live on for lengthy periods of time, allowing species persistence and reproduction. Sedentary existence of microorganisms attached to the surfaces of pores is likely the rule. Together with high variability at small-scale for resources and environmental conditions, it leads to the exceptionally high diversity of soil microorganisms, with a seemingly contradictory observations of diversity as high within a few-cm of soil as within a few-km watershed.

For many organisms the challenge might not be how to survive the short-term full immersion, but how to quickly switch from sedentary living on the air-solid or water-solid interfaces to benefitting from planktonic lifestyle and or slow movement on thin water films. It is also a question of how to survive onsets of dryness. Impressive diversity of motility mechanisms and flexibility of their use in soil bacteria as well as production of extracellular polymeric substances to stay hydrated are examples of adaptations to high temporal and spatial heterogeneity in water distribution.

Water flows and water distribution are closely coupled with gas fluxes and gas exchange with the atmosphere. Soil pores are the conduits of gases, and their characteristics, specifically size distribution and shape, are important. Large empty pores with low tortuosity that directly open into the atmosphere are oxygen rich and supply oxygen to the surrounding soil matrix. Gas flow through tortuous pore space is slower. Gas diffusion in water-filled pore space is by orders of magnitude slower than in the air. Joint influences of pore distribution through the soil matrix and water distribution within them greatly amplify the heterogeneity of the soil micro-environments in terms of oxygen levels. Some of the fine pore regions, removed from large atmosphere-connected pores, might remain low on oxygen during most of the year, others, like preferential flow pathways, might stay mostly aerobic, except for brief periods of saturation, yet still others might be changing from oxic to anoxic conditions often.

It should be noted that even in air-filled pores the humidity of the soil atmosphere is typically very high (~98%), enabling survival and functioning of many invertebrate and microbial organisms. Availability of fully- to partially-air filled pores is an important condition for fungi growth and for dispersion of fungi and its colonization of the soil habitat. Many soil invertebrates are sensitive to temperature and drought, thus abiotic habitat characteristics can be more important than trophic resources in defining suitability of particular soil habitat.

pH

Soil pH, the concentration of H^+ ions, can directly and indirectly affect the functioning of many soil organisms. The indirect influence can be manifested through the role of pH in governing availability of nutrients to soil organisms, notably P and Fe. Among the indirect influences are modulations of the activity of extracellular enzymes (EE). While EE may differ in terms of specifics of their optimal pH levels, any deviations from the optimal pH will reduce the enzyme activity and thus the returns on investment for those who produced it.

Among the many mechanisms driving soil pH are the following:

- 1) Dissolution of atmospheric CO_2 in the water and subsequently generated carbonic acid.
- 2) Dissolution of SO_2 and N oxides in the water generating sulfuric and nitric acids.
- 3) Microbial activities such as nitrification, redox reactions, and organic matter decomposition, where the latter is enhanced by agricultural activities, particularly plowing.
- 4) Plant activities; plants continuously release H^+ or OH^- to balance uptake of cations and anions at the root surface, along with organic acids and other exudates. Root respiration, also increases the concentration of CO_2 in soil with subsequent carbonic acid dissolution, as well as alters pH via redox-coupled reactions.

Anthropogenic influences play a substantial role in modulating soil pH. In agricultural soils NH_4^+ and NO_3^- additions as chemical fertilizers decrease pH, whereas lime and organic amendments, e.g., manure, increase it. Sulfur-rich acid rains enhance soil acidification worldwide. Yet, natural sources, such as sulfide enriched parent materials (e.g., pyrite), emissions of acid-generating gases during volcanic eruptions, or their production during thunderstorms, also play a role.

Most of the world's soils are acid with a pH range of 3.0–5.0, including soils of tropical and of boreal regions. It is the soils of grasslands and forests in temperate regions that tend to have neutral, 6.0–7.0, pH, whereas those of arid regions have neutral-to-alkaline pH of 7.0–8.0. However, in many soils, pH varies with depth, e.g., pH can be 3.5 in the A-horizon while measuring 5.0 in a C-horizon of the same soil profile. Moreover, the microscale patterns in spatial distribution of pH within the soil matrix can be extremely spatially and temporally variable. For example, pH can easily fluctuate within a 1–2 unit range across just a few millimeters

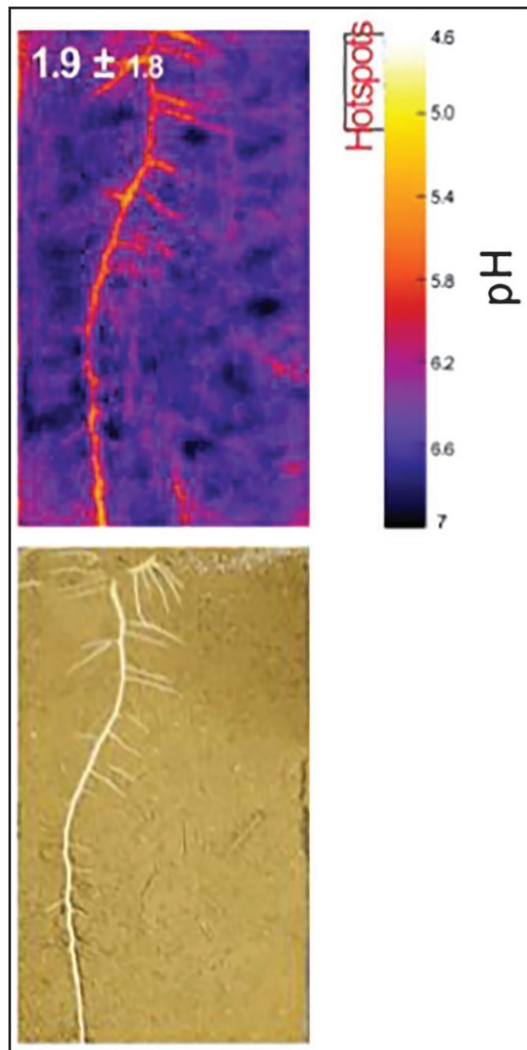


Fig. 3 An example of variability in distribution of pH in the rhizosphere and bulk soil (Ma et al., 2021). From Ma X, Liu Y, Shen W, and Kuzyakov Y (2021) Phosphatase activity and acidification in lupine and maize rhizosphere depend on phosphorus availability and root properties: Coupling zymography with planar optodes. *Applied Soil Ecology* 167: 104029.

of the soil matrix, driven by microscale variations in the factors listed above. It is notable that rhizosphere soil often has pH that is 1–2 units less than that of the bulk soil, with the contrast especially large when plants are experiencing P or Fe deficiencies (Fig. 3). Local nitrification spots in well-aerated pores can increase acidity locally, whereas the absence of oxygen and subsequent denitrification can reduce it. High spatial variability at the microscale is often coupled with high temporal variability, and the local pH changes in the rhizosphere can occur within the timeframe of a few hours. The microbial and micro-faunal organisms that lack the mobility for them to escape the unfavorable change must thus be able to adjust rapidly to pH fluctuations. Switching from acid to alkaline phosphatase production as a function of soil pH is an example of such acquired adjustment capabilities.

Biotic conditions

Most of the carbon in soil originates from atmospheric CO₂, which is photo-assimilated and placed belowground by plants. The rest of the nutrients, N, P, K, etc., originate primarily from soil minerals, where weathering is instrumental in their availability to plants and soil organisms. For example, the manganese (Mn), needed by plants for their photosynthetic activity, is obtained in a soluble form from weathering of primary aluminosilicate minerals and secondary minerals, often from deep within the soil profile. The Mn-rich leaves return to the soil surface with litter fall; this enriches the topsoil with soluble Mn providing it for subsequent uptake by plants, but also for use by a variety of soil organisms.

Soil heterotrophs are arguably the most abundant group of soil organisms. They get their nutrients from organic residues, with those of plant origin, either from live plants or their detritus, being the main source. Decomposition of inputs might be regarded as the key process underlying most of the activities in the soil.

The main types of plant inputs are from aboveground plant biomass (leaves and shoots), root biomass, and rhizodeposition. Chemical composition of the organic detritus is of major importance: it defines how much C is turned into the soil, quality of the added organic, thus its suitability for soil mesofauna and microorganisms, the form, quantity, and variety of nutrients added, soil pH, and other factors. Characteristics of soil organic matter influence its suitability as a trophic resource for soil inhabitants, both microorganisms and invertebrates that consume it directly. Yet, the importance of the indirect positive effects of soil organic matter accumulation on optimizing soil pH and increasing soil water holding capacity, while improving structure formation and thus oxygen supply, cannot be overstated.

Aboveground biomass contributes to soil C supply and nutrient turnover, when decomposed on the soil surface, with dissolved organic C and other decomposition products seeping into the topsoil. It also can be incorporated into the soil by animal activity, such as when earthworms drag leaves into their burrows, and further processes by fungi. In agricultural soils, tillage is a key route through which aboveground plant residues end up within the topsoil.

Roots play by far the largest role in building the soil habitat, and fine ($< 2 \text{ mm } \varnothing$) roots with their large spatial spread and fast turnover are especially important. Biomass of senescent roots is a very important source of organic inputs for soil organisms. The rhizodeposits, including root exfoliations, exudates, and other secretions (e.g., mucilage), as well as inputs of root organics into the soil via mycorrhizae, are other major contributors of organic inputs. Physical characteristics, such as root biomass, architecture, and morphology, directly define the formation of soil pore structure. The chemical composition of roots also defines their decomposability, similar to that of the aboveground biomass. Roots actively construct soil structure and soil pores, where pores are built by the direct physical impact of the roots. But the organic inputs from live and decomposing roots glue soil particles together, with subsequent wetting, drying, freezing, and thawing indirectly contributing to soil pore formation and stability.

Soil porosphere

It should be noted that all soil dwellers live within the soil pore space, either inside of the pores of their own creation, e.g., earthworms, or on pore-solid interfaces of existing soil pores, e.g., mesofauna and microorganisms. Earthworms and many other members of soil meso- and macrofauna are actively building their own living quarters within the soil creating their own habitat and acting as architects of the soil as a whole.

The special physical features of porous materials described above are relevant, both to fully developed soils and to all possible porous materials, justifying the concept of “the porosphere” (Vannier, 1987). Porosphere is defined as all solid materials with internal, i.e., porous, space, and soils is one of the most important members of that group. The characteristics of the porosphere that separate the conditions within it from any other soil environment are:

- 1) limited light access,
- 2) high CO_2 levels,
- 3) special features of its water, including water absorption on the surfaces of soil pores, water being held at negative potential, and frequent and widespread high relative humidity of the gaseous phase, and
- 4) exceptionally high micro-scale heterogeneity in spatial distribution of solid, liquid, and gaseous components exacerbated by a nearly infinite amount of completely separated or poorly connected pore spaces and particularly high volume of the small pores.

These characteristics played a major role in the evolution of not only soil microorganisms, but all members of the soil fauna. For example, high integumental permeability and resistance to high levels of CO_2 , characteristic of soil mesofauna, are responses to porosphere environment. It is argued that high humidity and water availability in the porosphere of the ancient Earth made the soil an intermediate step for movement of the organisms from the oceans and lakes. Likewise, great volumes of pore spaces inaccessible to plant roots but accessible to fungi in the soil porosphere drove the development of mycorrhizae as a more efficient way for plants to fully explore the resources of the soil matrix.

Although variations due to differences in mineralogy and organic inputs can be expected, the overall patterns in the microenvironmental conditions provided by pores of different sizes can be regarded as universal. They are based on organism and root size requirements, nutrient input patterns, and gas and water flow drivers (Fig. 4).

The large ($>500 \text{ } \mu\text{m}$ diameter) pores are of suitable size to host soil micro- to mesofauna, while very large ($>2000 \text{ } \mu\text{m}$ diameter) pores are needed for larger members of the mesofauna as well as for the macrofauna (earthworms, millipedes, etc.). Conditions within all large pores are similar in terms of almost invariably high levels of oxygen, due to ample connections with the atmosphere, but little water availability. The origin of the large pores in combination with other soil physical characteristics defines nutrient availability within them. Abiotic cracks receive nutrients only in a form of mobile organic C flowing through them during the saturation events and deposited as SOM- and clay-rich cutans formed on their walls, yet such events are infrequent, and the flows are fast. Large pores of biological origin, biopores, are highly enriched in nutrients originating from earthworm slimes and feces, root exudates, and root residues and are easily available to small fauna and microorganisms. But despite their potentially high nutrient levels, all large pores present challenges to their inhabitants in terms of water availability, since they are the first to empty of water and dry completely, and then to remain arguably the driest portion of the soil body. Moreover, the biopores of earthworm or plant

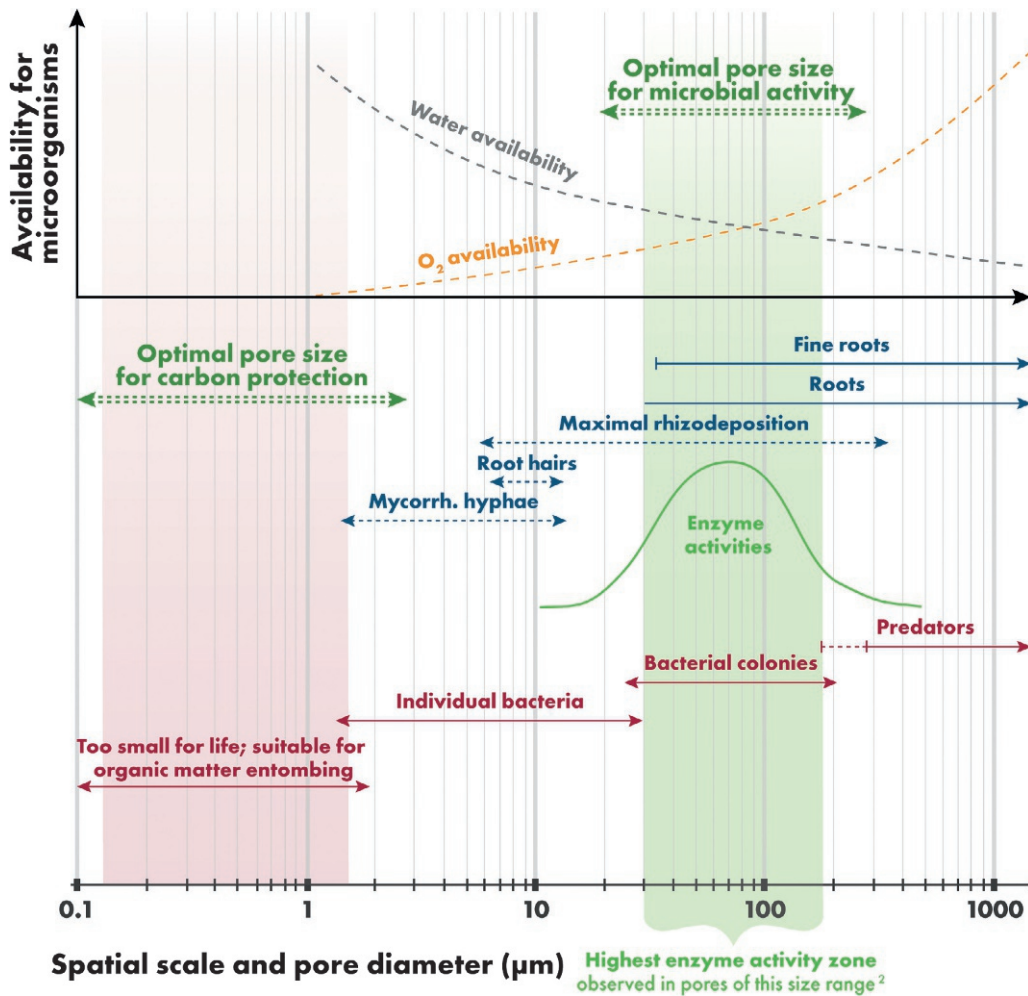


Fig. 4 Pore habitats and their key features (Kravchenko et al., 2020) specifying water and oxygen availability trends, root input contributions to pores of different size ranges, and assessments of size-wise suitability of pores to different types of inhabitants. Also highlighted are pore-size ranges contributing the most to soil C protection and biological activity. From Kravchenko AN, Guber AK, Razavi BS, Koestel J, Quigley MY, Robertson GP, and Kuzyakov Y (2020) Reply to: “Variables in the effect of land use on soil extrapore enzymatic activity and carbon stabilization” by Glenn (2020). *Nature Communications* 11.

origin as well as abiotic cracks often serve as preferential flow pathways, subject to what could be seen as micro-scale flash-floods, with a powerful impact on their inhabitants.

The medium-sized (tens to hundreds μm diameter) pores are appropriate to host the microfauna (e.g., protozoa and nematodes) as well as the smallest members of the mesofauna groups. Many microorganisms form colonies embedded in hydrated extracellular polymeric substances – medium-sized pores are big enough to accommodate such structures. Overall, the space of medium-sized pores has several advantages over that of the large pores. It provides a greater balance between oxygen availability - frequent and sizeable, and water availability - at optimal levels for longer periods of time than in the large pores. Although medium-sized pores might well suit soil organisms in terms of their physical characteristics, it is the availability of nutrients which is the necessary condition for turning such pores into optimal habitats. Among the sources of nutrients in these pores are:

- 1) fine plant roots and root hairs – the most active components of plant root systems in providing C inputs via exudates and rhizodeposits;
- 2) decomposition products leaching out of POM fragments and senescent fine roots; and
- 3) mobile organic C brought by the inflowing soil solution.

It should be noted that, compared with large pores, the flow velocities within the medium pores are slow, leaving more time and opportunities for the entrained organic C and nutrients to interact with minerals of pore wall surfaces and to become accessible to their inhabitants. Medium pores do experience sporadic anoxic conditions. This happens during soil saturation and during periods of high microbial activity when the rate of oxygen consumption exceeds that of its diffusion, the latter especially relevant for the

nutrient-rich pore spaces. The connectivity of large and medium-sized pore spaces is a major contributor to dispersion and interactions among micro- to mesofauna species, and the lack of it stimulates diversity not only of microbes but also of soil mesofauna.

Small-sized pores can be especially variable in what they offer as habitats to soil microorganisms and microfauna. Their suitability greatly depends on their relative position with respect to other elements of the soil matrix, such as large and medium pores and nutrient sources, (e.g., plant roots and POM). Whereas, overall the small pores often have low levels of oxygen and tend to stay water-saturated for the longest time, the oxygen deficiencies can be frequently alleviated if they are located nearby air-filled large and medium pores. However, the small pores deep within the soil matrix, far from any air-filled bigger pores, likely remain persistently anaerobic. The small pores in the rhizosphere are more likely to have constant inputs of C and nutrients of root origin, while the ones within the bulk (i.e., not rhizosphere or detritosphere) soil have a very limited nutrient supply. The only mechanisms through which C and nutrients can reach small pores within the bulk soil are capillary flow of soil solution and diffusion of the chemicals; both processes are slow and inefficient in providing the nutrients to the inhabitants.

The relative proportions and volumes occupied by different pore-size groups can vary greatly among soils of different types and geographies, reflecting a multitude of influences on soil formation and inherent soil characteristics, including texture, soil history, vegetation, and management. Pore space constitutes 40–60% of the total soil volume, but a sizeable (>30%) proportion of it are pores with diameters <1 µm, thus not suitable as a habitat even for the smallest of soil microorganisms. In mineral soils the volumes of small- medium- and large-sized pores can constitute 20–80, 10–40, and 5–10% of the pore space, respectively.

Special habitats: Rhizosphere and detritosphere

Due to their substantial C and nutrient availability, rhizosphere and detritus present two very special environments within the soil matrix and can be regarded as separate habitats in their own right. They also differ from the other portions of the soil matrix in terms of physical characteristics. They can be viewed as “hot-spot” habitats, existence stemming from concentrated sources of labile C that are highly variable, both spatially and temporally (Kuzakov and Blagodatskaya, 2015):

Rhizosphere

The rhizosphere possesses a number of characteristics that make it arguably an optimal soil habitat for soil microorganisms, micro-fauna and, possibly, meso-fauna. Plants provide the soil around them with organic compounds of exceptionally wide chemical diversity creating a drastic contrast between rhizosphere and bulk soil nutrient supply. Many of the compounds exuded by the roots come out in quantities that are below what would be of nutritional value to the organisms, but instead are signal molecules that recruit microbial symbionts, and indirectly influence many other microorganisms and fauna. Plants also generate fluxes of H⁺, OH⁻, cations, anions in the rhizosphere, majorly affecting its biochemistry. The rates and quantities of plant outputs and intakes are strongly affected by plant species, root age, soil texture, soil moisture content, and the pore structure encountered by the roots.

Roots tend to grow into pores somewhat bigger than themselves, preferring the pores that are relatively well interconnected and connected to the atmosphere. Often a root does not fully occupy the entirety of the pore, leaving what is known as a “root gap.” Formation of a root gap, together with high oxygen consumption by the roots themselves and by the active microbial and faunal life in the rhizosphere, leads to highly dynamic oxygen availability. When soil becomes wet, and the pores formed or modified by roots become the pathways for preferential water flow, subsequent oxygen consumption and slow gas diffusion will lead to temporal onset of anoxic conditions. Yet, upon soil drying, speeded up by the root water uptake, oxygen availability may be restored within a few hours. Moreover, roots greatly influence the water content within the rhizosphere, they either absorb it, generating water flow toward the root surface and often resulting in a smaller water content in the root vicinity, or they maintain the content somewhat above that of the bulk soil when the latter becomes dry. The magnitude and dynamic of such fluctuations greatly depend on the plant species and soil physical characteristics.

Though the area of the most active rhizosphere expands to 1–3 mm from the root, the overall sphere of root-influence can extend as far as 20 mm for fast diffusing small-molecule compounds consumed by the roots, e.g., depletion of NO₃⁻ in the root vicinity. Furthermore, the sphere of influence can be <1 mm for some of the heavy organic molecules of root exudates or extracellular enzymes. Thus, the outcome is a gradient of micro-environmental conditions expanding from the root surface into the bulk soil, rather than an abrupt rhizosphere boundary.

While we generally refer to all soil in close proximity to roots as the rhizosphere, it should be noted that specific inputs and environmental conditions vary drastically and very rapidly along the length of a growing root. What was once a micro-region receiving inputs from a root tip and promoting fast-growing bacteria, becomes a location of root hairs supporting mycorrhizal fungi, and then a place adjacent to an older root benefiting root-grazing fauna. These changes may take place within a few hours to a few days.

Detritosphere

Plant and animal residues, i.e., their detritus, and soil in their immediate vicinity, i.e., the detritosphere, form yet another unique soil habitat (Fig. 5). It resembles rhizosphere in terms of a plentiful supply of C and nutrients available to soil microorganisms and micro- and mesofauna, but has a number of differences.

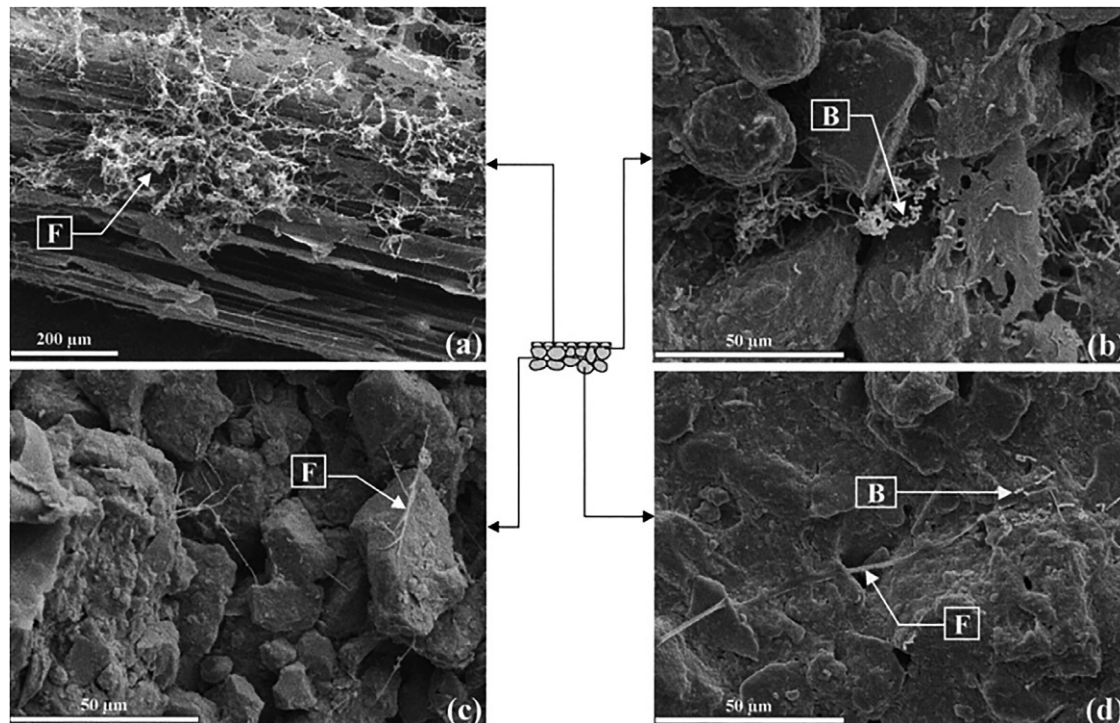


Fig. 5 Detritus from straw (A) incubated within the soil for 30 days and colonized by fungi (F) and bacteria (B) in proximity ((B) and (D)) and at greater distance (C) from the residue (Gaillard et al., 1999). From Gaillard V, Chenu C, Recous S, and Richard G (1999) Carbon, nitrogen and microbial gradients induced by plant residues decomposing in soil. *European Journal of Soil Science* 50: 567–578.

The properties of the detritosphere are a function of (i) the nature of detritus itself, i.e., its origin, chemical composition and physical characteristics, (ii) the properties of the soil surrounding the detritus, i.e., soil texture and pore characteristics, and (iii) the soil moisture regime. How available is the detritus to decomposition and the quantity and quality of nutrients it can provide vary substantially, and also change with time as decomposition progresses. Whether a fragment of detritus is surrounded by the tightly adherent soil with limited gas diffusion or is in a relatively large air-filled pore will lead to contrasting conditions for the activity of heterotrophic organisms and the decomposition rates. The former case might be typical for fragments of aboveground plant biomass incorporated into the soil, while the latter for the detritus of the root origin. The decomposition products diffuse out of the decomposing residue into the surrounding soil supporting the organism within. The distance for such diffusion varies from a few tens of microns for some large organic molecules to a few millimeters for NO_3^- .

The overall soil moisture content will affect the longevity of a particular detritosphere spot through its influence on decomposition. For example, speeding it up in wet soil with good oxygen supply, or slowing it down when the conditions are too dry or too wet and anoxic. The soil moisture also affects the spatial extent of the detritosphere influence - it can be expected to be larger under wetter conditions compared to drier conditions. Yet due to their fine pore structure, tissues of many plant residues, especially at the early stages of decomposition, can act like a sponge, absorbing water from the surrounding soil. Thus, they may provide water for the decomposers even when the overall soil moisture is relatively low.

References

- Franklin SM, Kravchenko AN, Vargas R, Vasilas B, Fuhrmann JJ, and Jin Y (2021) The unexplored role of preferential flow in soil carbon dynamics. *Soil Biology & Biochemistry* 161.
- Gaillard V, Chenu C, Recous S, and Richard G (1999) Carbon, nitrogen and microbial gradients induced by plant residues decomposing in soil. *European Journal of Soil Science* 50: 567–578.
- Kravchenko AN, Guber AK, Razavi BS, Koestel J, Quigley MY, Robertson GP, and Kuzyakov Y (2020) Reply to: "Variables in the effect of land use on soil extrapore enzymatic activity and carbon stabilization" by Glenn (2020). *Nature Communications* 11.
- Kuzyakov Y and Blagodatskaya E (2015) Microbial hotspots and hot moments in soil: Concept & review. *Soil Biology & Biochemistry* 83: 184–199.
- Ma X, Liu Y, Shen W, and Kuzyakov Y (2021) Phosphatase activity and acidification in lupine and maize rhizosphere depend on phosphorus availability and root properties: Coupling zymography with planar optodes. *Applied Soil Ecology* 167: 104029.

Mills AL (2003) Keeping in touch: Microbial life on soil particle surfaces. *Advances in Agronomy* 78: 1–43.
Vannier G (1987) The porosphere as an ecological medium emphasized in Professor Ghilarov's work on soil animal adaptations. *Biology and Fertility of Soils* 3: 39–44.
Voroney RP (2007) *The Soil Habitat, Soil Microbiology, Ecology and Biochemistry*. Elsevier 25–49.

Further reading

Coleman DC, Callahan MAJ, and Crossley DAJ (2018) *Fundamentals of Soil Ecology*, 3rd edn Elsevier Inc.
Paul EA (2015) *Soil Microbiology, Ecology and Biochemistry*, 4th edn Elsevier Inc.
Ritz K and Young I (2011) *The Architecture and Biology of Soils: Life in Inner Space*. UK: CABI.

Plant growth promoters

Elisa Gamalero^a, Gustavo Santoyo^b, and Bernard R Glick^c, ^aDipartimento di Scienze e Innovazione Tecnologica, Università del Piemonte Orientale, Alessandria, Italy; ^bInstituto de Investigaciones Químico Biológicas, Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Mexico; ^cDepartment of Biology, University of Waterloo, Waterloo, ON, Canada

© 2023 Elsevier Ltd. All rights reserved.

Introduction	419
Need for non-chemical (natural) means of promoting plant growth because of increasing population and global climate change	419
Prevalence of different organisms in soil	419
Long-term relationship (mutual dependency) between soil microbes and plants	420
Plant growth-promoting bacteria: Rhizospheric, endophytic, mutualistic symbiotic, phyllospheric	420
Mechanisms	421
Direct mechanisms	421
N fixation	421
Fe sequestration	422
P and K solubilization	422
IAA production	422
Gibberellin synthesis	423
Cytokinin synthesis	423
ACC deaminase	423
Indirect mechanisms	423
Antibiotics	423
HCN	424
Cell wall degrading enzymes	424
Siderophores	424
Induced systemic resistance	424
Plant growth-promoting fungi	424
Mycorrhizal fungi	424
Trichoderma	424
Microbial consortia	425
Bacteria-bacteria	425
Bacteria-mycorrhizal fungi	425
Bacteria-Trichoderma	425
Commercialization of plant growth-promoting soil microorganisms	426
Bacteria	426
Fungi	426
Future prospects	426
References	426

Abstract

There is insufficient food to adequately sustain the population of the world, either now or in the future. However, the food supply might be increased sustainably by employing plant-beneficial soil microorganisms in agriculture. These microorganisms have naturally facilitated plant growth and development for millions of years and can support plant growth directly or indirectly by decreasing the damage caused by phytopathogens. The beneficial microorganisms include plant growth-promoting bacteria (PGPB), mycorrhizal and endophytic fungi. PGPB may be rhizospheric, endophytic, symbiotic, or phyllospheric. Microorganisms may act as consortia to benefit plant growth including bacteria-bacteria, bacteria-mycorrhizae, and Bacteria-Trichoderma consortia.

Key points

- There is not enough food to adequately feed the population of the world.
- One way to sustainably improve food production is using plant-beneficial soil microorganisms.
- Many different soil microorganisms facilitate plant growth.
- These microorganisms use a range of mechanisms to promote plant growth, both functioning individually and working together in consortia.
- Beneficial microorganisms include plant growth-promoting bacteria, mycorrhizal fungi, and *Trichoderma* spp.
- Plant growth-promoting bacteria may be rhizospheric, endophytic, symbiotic, or phyllospheric.

- Mixtures of microorganisms that act as consortia to benefit plant growth include bacteria-bacteria, bacteria-mycorrhizal fungi, and Bacteria-Trichoderma consortia.
- Some plant growth promoting bacteria, mycorrhizal fungi and Trichoderma have already been commercialized.

Abbreviations

AMF	Arbuscular Mycorrhizal Fungi
BCA	BioControl Agents
PGPB	Plant Growth-Promoting Bacteria

Introduction

Need for non-chemical (natural) means of promoting plant growth because of increasing population and global climate change

The production of food on a global scale is dependent on agriculture, the main objective of which is to provide enough food for the world population. In 2020, a United Nations report on the global State of *Food Security and Nutrition* concluded that almost 690 million people suffered from hunger in 2019, which was 10 million more than in 2018 and the number had increased by slightly less than 60 million over the previous 5 years. In addition, the fraction of undernourished people was approximately 9% of the world's population, and over the last 5 years the trend in this parameter has increased, suggesting that food demand relative to food availability has increased proportionally with the world population (<http://www.fao.org/3/ca9692en/online/ca9692en.html#>).

At the same time, due to the increase in the amount of greenhouse gasses in the atmosphere the world is witnessing dramatic climate change especially with respect to temperature and rainfall patterns, creating stressful conditions for plant growth and productivity. Abiotic factors such as salinity (10%), low and high temperatures (7% and 20%, respectively), drought (9%), and other stress conditions (4%) can lead to a ~50% global loss of crop productivity. It is also true that climate change favors the propagation of pests and soil borne diseases, potentially bringing additional significant losses in global agricultural yield.

Intensive agriculture has been viewed as one solution to this problem, however, this approach is based on the use of large amounts of chemical pesticides and fertilizers to maximize plant and crop yields to keep up with the growing global food demand. However, these chemicals accumulate in soils and water, polluting ecosystems, resulting in a plethora of environmental and human health issues. At the same time, consumers have become more aware of the impact of these chemicals on their diet and health. Thus, the demand for healthy, nutritious foods is rapidly increasing in both developed and developing countries. Therefore, a significant reduction in the use of chemical inputs in agriculture is necessary, and the application of environmental sustainability principles has become an essential target shared by many farmers and consumers.

Genetic engineering and plant breeding are two of the available tools that may be used to enhance crop resistance to stressful conditions. However, these strategies include time-consuming and costly processes, and they are not always positively perceived by the consumer market.

In this context, an alternative option for increasing agricultural productivity and overcoming various environmental stresses includes the deliberate use of microbial biofertilizers. These microbes offer several advantages to other approaches, they are typically non-toxic and are rarely associated with resistance development by the target organism.

Prevalence of different organisms in soil

Soil contains many micro- (viruses, bacteria, algae, fungi, protozoa, and nematodes) and macro- (earthworms, myriapods, and insects) organisms, whose population density varies according to the taxonomic group considered. Among microorganisms, fungi are dominant in soil as their amount is estimated to reach 10^5 – 10^6 cells/g of soil with a biomass of a 100–1500 g/m². On the other hand, the bacterial density approximates 10^8 – 10^9 cells/g of soil, but due to their relatively small size, their total biomass is 2–3 times less than that of fungi (Glick, 2020). The number of microorganisms in soil changes with land management, as well as soil depth. Because of their high tolerance level to disturbance, bacteria and protozoa are prevalent in tilled soil, while fungi and nematodes dominate in untilled soils. According to Torsvik et al. (1990), 1 g of soil contains up to 1.5×10^{10} bacterial cells with ~40,000 different species, of which fewer than 1% are able to grow and develop colonies on artificial media. The availability of organic substances and oxygen decrease through different soil horizons: consequently, the population density of microorganisms decreases with depth in soil. Soil biomass, together with soil biodiversity are the main drivers of ecosystem functions. Thus, greater soil biodiversity is expected to lead to more stable ecosystem functioning. Soil microbial diversity has often been considered a mysterious black box and it has been only with the development of molecular tools, and more recently, -omics techniques that this box has been opened.

Long-term relationship (mutual dependency) between soil microbes and plants

Plants evolved during the Mid-Ordovician period, ~460 to 470 million years ago. Since then, a huge number of diverse microorganisms established interactions with land plants, leading to colonization with different degree of intimacy.

Mycorrhizal fungi are believed to have co-evolved with their host plants since the Early Devonian (~393 to 419 million years ago). At that time, plants were very simple, lacking leaves and roots, being characterized by a rhizoid devoted to nutrient uptake and living in a hostile environment. Interestingly, fossil plants of this type recovered in Scotland have been found to be associated with fungi able to produce arbuscule-like structures like those of the Glomeromycotina. It is reasonable to think that one factor that motivated the development of mycorrhizal symbiosis was the need of both partners to survive in such a hostile environment.

Within bacteria, several taxonomic groups can fix atmospheric nitrogen, and this widespread ability suggests that occurrence of multiple independent regulatory molecular mechanisms evolved through convergent evolution. According to a second hypothesis, Angiosperms belonging to the nitrogen-fixing clade developed a predisposition to nodulate, thus leading to the full establishment of the legume-rhizobia symbiosis (Martin et al., 2017).

Recent data indicate that plants actively recruit their associated microorganisms from the soil, rhizosphere, phyllosphere, anthosphere, spermosphere and carposphere (Gamalero and Glick, 2020). This intricate net of interactions connecting the plant host with its microbiome and all the members of the microbiota to each other is called a “holobiont”. Beneficial microorganisms and plants have interacted and facilitated one another's growth for around 50 million years. While most soil microorganisms and plants can grow and develop on their own, their survival and persistence are greater when they cooperate, especially when facing harsh environmental conditions.

Plant growth-promoting bacteria: Rhizospheric, endophytic, mutualistic symbiotic, phyllospheric

Plants host microorganisms in and around all their tissues. The rhizosphere (the area around the plant roots) is characterized by an intense exudate release by plant roots (mainly composed of organic acids, sugars, amino acids, fatty acids, and vitamins) that attracts a large amount of microbial biodiversity. The phyllosphere (the surface of leaves and stems) is a relatively unfavorable environment for bacteria, being subjected to extremely variable temperature, radiation, and fluctuating water activity. Microorganisms living close to or on the surface of plant tissues are defined as epiphytes, whereas microorganisms able to colonize the internal plant tissues (stems, leaves, roots, flowers, fruits, and seeds) are called endophytes. Symbiotic mutualistic relationships can be established between plants and specific group of bacteria. For example, the symbiosis between various strains of nitrogen fixing rhizobia and about 15,000 different legume species is one of the oldest and most studied. Once an association between the plant and the microbiota is established, the microorganisms may cause beneficial, neutral, or detrimental interactions with the host plant.

Bacterial community structure in the rhizosphere can vary according to the plant species, its phenological stage and health, as well as to environmental and soil parameters. However, the core microbiome, a group of microbial taxa or genes that are shared by all or most local plant species, is generally composed of alpha and beta Proteobacteria followed by Actinobacteria, Firmicutes, Bacteroidetes, Planctomycetes, Verrucomicrobia, and Acidobacteria. Rhizodeposition selects the associated microbiota but in turn this affects the composition of root exudates.

Those bacteria colonizing the phyllosphere are very well adapted to the low nutrient availability and mainly colonize the underside of the leaf on veins, hairs, stomata and glandular trichomes. As for the rhizosphere, there is a core microbiome but composed mainly of alpha and gamma Proteobacteria, followed by Bacteroidetes and Actinobacteria. The physiological group of methylotrophs are quite abundant and typically present on the leaves where they use C1 compounds (i.e., one carbon atom molecules, thus excluding a C—C bond) such as methanol and methylamine as carbon sources. Moreover, lactic acid bacteria dominate the phyllosphere of several plants in the Mediterranean area especially during the summer season (Remus-Emsermann and Schlechter, 2018).

Finally, the endosphere, represented by the internal plant tissues, are typically colonized by non-pathogenic bacterial strains originating mainly from the bulk soil or the rhizosphere and entering through cracks or lateral root branching junctions. This type of horizontal transmission of beneficial bacteria may represent an advantage for the plant that recruits a diverse set of symbionts, thus leading to a prompt response to stressful conditions. Frequently, the penetration into the plant tissues is supported by the release of lytic enzymes degrading the plant cell wall. Once inside the apoplastic space, bacteria replicate reaching densities up to 10^7 – 10^8 colony forming units (cfu) per gram of tissue fresh weight. At the root level, these bacteria can spread to the xylem vessels and then are translocated towards the epigeous portion of the plant. It should also be considered that the possible inheritance of bacterial endophytes is through vertical transmission. If a seed endophyte is a plant beneficial bacterial strain, it is reasonable to suggest that a form of mutualism has been established between the bacterial strains and the plant host (Frank et al., 2017).

The relationship established between the host plant and the members of the bacteriome can be very diverse according to the final effects on the plant growth (neutral, positive, or deleterious) (Fig. 1). Those bacteria involved in positive plant-microbe interactions, leading to beneficial effects on plant development and health are defined as Plant Growth-Promoting Bacteria (PGPB) (Glick, 2020). The mechanisms by which PGPB improve plant growth, health and increase yield are described in the following sections.

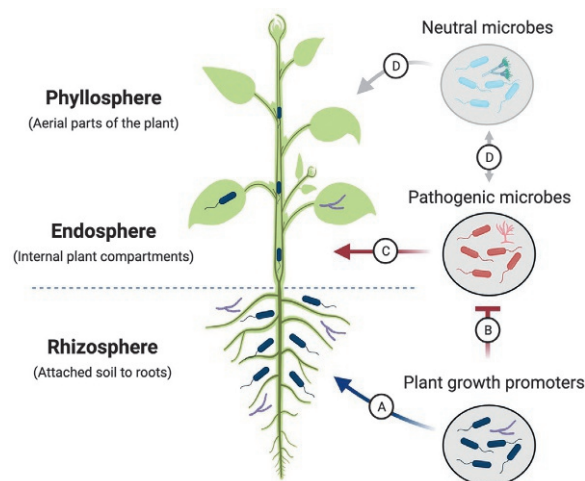


Fig. 1 A general overview of plant growth-microbial promoters residing in the phyllosphere, endosphere and rhizosphere. (A) Beneficial interactions that improve plant growth and fitness (i.e., modulation of phytohormones, nutrient uptake improvement, etc.); (B) Biocontrol of pathogens is also carried out by plant growth promoters, and indirectly, improve the plant health; (C) Harmful interaction exerted by pathogens towards the plant; (D) Neutral or commensal microbes interacting with the plant or other microbes.

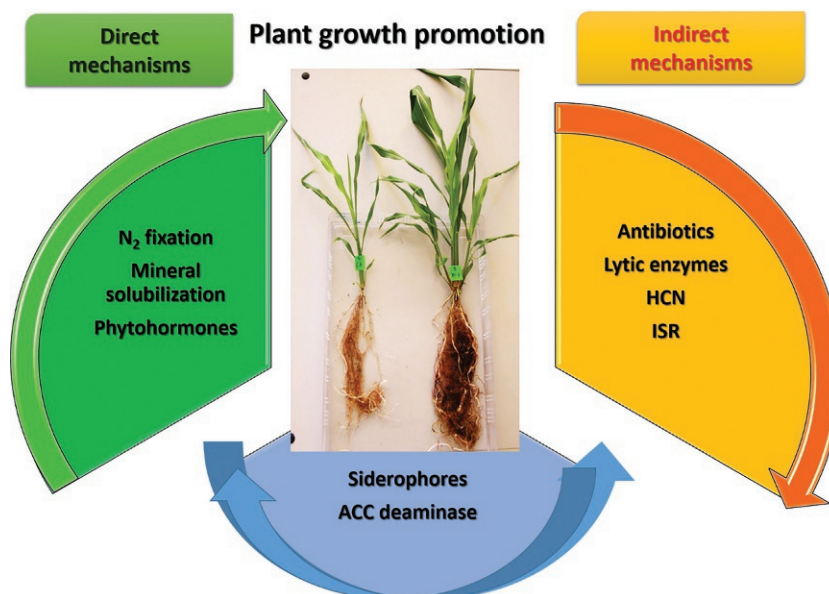


Fig. 2 Schematic overview of the main mechanisms used by PGPB to promote plant growth. Direct mechanisms are based on plant nutrition improvement through nitrogen fixation, mineral solubilization (mainly phosphate and potassium) and phytohormones synthesis such as auxins, gibberellins and cytokinines. Indirect mechanisms involved production of antibiotics (2,4-diacetylphloroglucinol, pyoluteorin, etc.), lytic enzymes such as chitinases and cellulase, cyanuric acid and induced systemic resistance (ISR). The synthesis of iron chelators (siderophores) and of the enzyme 1-aminocyclopropane-1-carboxylic acid are involved in both direct and indirect mechanisms of plant growth promotion.

Mechanisms

While direct mechanisms carried out by PGPB enhance plant growth in the absence of pathogens (mainly through nutrient uptake and phytohormone concentration modulation), the indirect mechanisms of plant growth promotion consist of biocontrol of deleterious microorganisms (by antibiosis, lytic enzyme release, induced systemic resistance, siderophore and cyanide production) (Fig. 2).

Direct mechanisms

N fixation

The amount of nitrogen gas in the atmosphere is very high (~78%). However, plants are unable to use nitrogen in this form, therefore N₂ needs to be reduced to ammonia before becoming part of macromolecules such as proteins and nucleic acids.

Biological nitrogen fixation (BNF) is mediated by some free living and symbiotic bacteria that can convert atmospheric nitrogen into plant available ammonia through an energy expensive reaction mediated by the enzyme complex nitrogenase. While free-living nitrogen-fixing bacteria include several bacterial species that are associated with plant roots, symbiotic nitrogen fixers include a limited range of rhizobia, each specific to one or a few legume species. Although quantification of the BNF contribution can be problematic, it has been estimated that ~80% of the biologically fixed nitrogen comes from the legume-rhizobia symbiosis. However, under appropriate conditions, the contribution of free-living nitrogen fixers to plant nutrition can be quite significant, reaching 60 kg N ha⁻¹ year⁻¹ (Aasfar et al., 2017).

Nitrogen is one of the macronutrients limiting crop yield. Consequently, agricultural crops are highly dependent on the provision of fixed nitrogen in chemical fertilizers. BNF provides a significant fraction of the available nitrogen to the total plant requirements, therefore decreasing the need for synthetic nitrogen fertilizers during plant growth (Masson-Boivin and Sachs, 2018). Included in the wide range of bacteria that are able to fix nitrogen are several genera of cyanobacteria; the symbiotic genera Rhizobia, Sinorhizobium, Bradyrhizobia, Mesorhizobium and Frankia; rhizospheric and endophytic genera Azospirillum, Azotobacter, Klebsiella, Bacillus and Pseudomonas (Glick, 2020).

Fe sequestration

Despite its great abundance within the earth's crust, in aerobic soils, iron occurs mostly in precipitated forms making the amount of this element available to organisms very small (10^{-7} to 10^{-23} M at pH 3.5 and 8.5, respectively). However, iron is required by both plants and soil microorganisms (10^{-4} to 10^{-9} and 10^{-5} to 10^{-7} M, respectively), leading to an intense competition for this nutrient, especially in the rhizosphere where plant, bacteria, and fungi coexist. Many bacteria respond to iron starvation by synthesizing (i) low-molecular weight iron chelators, called siderophores, with exceptional affinity for ferric iron and (ii) membrane receptors that bind and internalize the ferri-siderophore complex. According to their chemical structure, siderophores are classified into catecholates, hydroxamates, and carboxylates. Although the capability of synthesizing siderophores is widely distributed among bacteria and fungi, mixed ligands produced by fluorescent pseudomonads (i.e., pyoverdines, pseudobactins and pyochelin) are the most studied. Siderophores have a role in improvement of plant nutrition since they supply iron to plants that are able to uptake the bacterial ferri-siderophores complex. Moreover, the increased amount of iron available to plants and beneficial soil microorganisms may lead to the proliferation of beneficial microbes in soils, thus affecting the composition of soil microbiota (Saha et al., 2016).

P and K solubilization

While the amount of phosphorus (P) is typically between 400 and 1200 mg kg⁻¹ of soil the availability of phosphorus in a soluble form (i.e., phosphate ions, H₂PO₄⁻ and HPO₄⁻²) is very small at ~1 mg kg⁻¹ or less. Since P is one of the macronutrients that can limit plant growth it is usually introduced into agricultural soils as a chemical fertilizer. However, only 10–25% of this added phosphate is generally used by plants since it forms metal cation complexes that precipitate in soils. In this context, phosphate solubilizing bacteria (PSB) including species of Pseudomonas, Agrobacterium, Bacillus, Azotobacter, Burkholderia, Enterobacter, Erwinia, Kushneria, Paenibacillus, Ralstonia, Rhizobium, Rhodococcus, Bradyrhizobium, Sinomonas, and Thiobacillus (Alori et al., 2017) able to solubilize inorganic phosphates through the synthesis of organic acids that release soluble phosphate, or able to solubilize organic phosphates by synthesizing enzymes that break down these compounds may be useful (Alori et al., 2017).

After N and P, potassium (K) is one of the main essential macronutrients for plant growth. Since the amount of potassium in a soluble form in soils is usually very low, plants can show potassium deficiency symptoms such as poorly developed roots, leaf chlorosis localized on the margin of the dorsal side, reduced plant biomass and yield. Solubilization of potassium rock by PSB occurs through the release of organic acids lowering of the pH. Strains belonging to the species Acidithiobacillus ferrooxidans, Paenibacillus spp., Bacillus mucilaginosus, Bacillus edaphicus, and Bacillus circulans showed this interesting ability which, together with other plant beneficial traits, can lead to a reduction of the use of chemical fertilizers in agricultural soil management (Etesami et al., 2017).

IAA production

Among the auxins, indole-3-acetic acid (IAA) is the most common and the most studied. As often happens for hormones, different amounts of auxins show different impacts on plant growth. For plants, the main consequences related to auxin are the modulation between cell division and extension, differentiation of plant cells and tissues, seed and tuber germination, increases in the rate of xylem and root development, increases in the emergence of lateral and adventitious roots, mediation of responses to light and gravity, florescence, and fructification of plants. IAA also affects photosynthesis, pigment formation, biosynthesis of various metabolites, and resistance to stressful conditions.

The effect of auxin in planta on root elongation is dose dependent. Moreover, the effective concentration of auxin in planta depends on the plant species and on the sensitivity of the target tissue. Bacterial auxin levels below an optimal amount may have little to no effect on plant development, while levels above this concentration of hormone can inhibit growth.

The ability to synthesize IAA is a common trait in many soil bacteria, both PGPB and pathogens, belonging to a wide range of different taxonomic groups. There are seven known metabolic pathways leading to IAA biosynthesis in bacteria, with five of them based on tryptophan as a precursor (Duca et al., 2014). Tryptophan synthesis is energy requiring for bacterial cells. Consequently, its concentration inside the cell is small, and the bacterial cells need to find a balance between the use of this amino acid for protein synthesis and the production of IAA. To obtain large amounts of IAA, bacterial cells use the tryptophan exuded by the plant root systems. IAA of bacterial origin supplements the endogenous plant IAA pool potentially leading to different and opposite

consequences (stimulation or suppression) for plant growth and root morphogenesis. One of the most relevant consequences of bacterial IAA synthesis is an increase in lateral and adventitious roots, leading to improved nutrient uptake and, in turn, to the enhancement of root exudation which attracts microorganisms that colonize the roots. Therefore, IAA, together with a plethora of other molecules released by plants, is involved in the build-up of the plant microbiota (Keswani et al., 2020).

Gibberellin synthesis

Gibberellins are phytohormones produced by higher plants, fungi, and bacteria, with a chemical structure based on diterpenoid acids consisting of an isoprene residue. Affecting both cell division and elongation, gibberellins play an important role in several stages of plant development such as seed germination, stem elongation, flowering, fruit setting, and delay of senescence in many organs. Moreover, they regulate root hair abundance thus positively impacting nutrient uptake efficiency. The ability to synthesize gibberellin is quite common in both Gram positive and negative bacterial strains belonging to the genera *Azospirillum*, *Azotobacter*, *Arthrobacter*, *Azospirillum*, *Pseudomonas*, *Bacillus*, *Acinetobacter*, *Flavobacterium*, *Micrococcus*, *Agrobacterium*, *Clostridium*, *Rhizobium*, *Burkholderia*, and *Xanthomonas*. Plant growth stimulation by PGPB able to produce gibberellins has been widely demonstrated and is often associated with other plant beneficial physiologic traits as well as the interplay with other hormones and the modulation of the overall hormonal balance in plants (Salazar-Cerezo et al., 2018).

Cytokinin synthesis

The chemical structure of cytokinins is based on N6-substituted aminopurine. These molecules are involved in diverse plant physiological processes, such as cell division, interruption of the quiescence of dormant buds, activation of seed germination, promotion of branching, root growth, accumulation of chlorophyll, leaf expansion, and delay of senescence. One plant response to cytokinins is the modification of cell size and shape due to regulation of the expression of protein expansion, which determines the loosening of plant cell walls and the subsequent facilitation of plant cell expansion.

Although several PGPB strains including *Azotobacter*, *Azospirillum*, *Rhizobium*, *Bacillus*, and *Pseudomonas* spp., can synthesize this hormone, cytokinins synthesized by bacteria have received much less attention than the bacterial synthesis of IAA. In fact, the synthetic pathway of bacterially encoded cytokinins and their effect on plant development are not yet completely understood. Finally, because of loosening plant cell walls, bacterial cytokinins can increase the release of root exudates that function as attractants for many bacteria, including plant beneficial ones therefore increasing the possibility of being efficiently colonized by PGPB.

ACC deaminase

The enzyme ACC deaminase can degrade the ethylene precursor, ACC, to ammonia and alpha-ketobutyrate and is widespread in soil bacteria and some soil fungi. By breaking down some of the ACC in plants, this enzyme can significantly lower plant ethylene levels and thus decrease the extent of ethylene inhibition of plant growth, especially following various biotic and abiotic stresses.

The gaseous phytohormone ethylene, produced in all higher plants, mediates a wide range of plant responses and developmental steps. It is involved in seed germination, tissue differentiation, formation of root and shoot primordia, root branching and elongation, lateral bud development, flowering initiation, anthocyanin synthesis, flower opening and senescence, fruit ripening and degreening, production of volatile organic compounds responsible for aroma formation in fruits, leaf senescence, leaf and fruit abscission, Rhizobial stimulation of nodule formation, mycorrhizal fungi-plant interaction, and the response of plants to a wide range of biotic and abiotic stresses. In some cases, ethylene is stimulatory while in others it is inhibitory, depending upon the amount of ethylene that is produced.

Exposure of plants to stressful conditions (either biotic or abiotic) results in the overproduction of plant ethylene levels. High ethylene levels can cause a plant to incur increased damage as a result. For example, when plants are infected by phytopathogens, a large fraction of the damage to the plant is due to high ethylene levels. According to the model proposed by Glick et al. (2007), following the onset of an environmental stress the plant produces a small amount of ethylene leading to the transcription of defensive genes. Subsequently, if the stress becomes more intense or persists over time, the ethylene level in the plant increases and induces the initiation of processes inhibitory to plant growth and survival such as senescence, chlorosis, and abscission. Bacteria able to synthesize ACC deaminase, cleave the immediate ethylene precursor, ACC, into ammonia and -alpha-ketobutyrate, lowering the second ethylene peak, thus improving plant tolerance to a wide range of stresses including stress caused by pathogen infection of plants.

Indirect mechanisms

Antibiotics

The rhizosphere is a hot spot where the members of the microbiota are constantly competing for nutrients. Therefore, the synthesis and release of antibiotic molecules is a survival strategy for some rhizosphere microorganisms to help them to successfully colonize this ecological niche. As observed during take-all decline, wheat plants infested with the phytopathogenic fungus *Gaeumannomyces graminis* var. *tritici* (Ggt) through the exudates released by necrotized roots, recruit fluorescent *Pseudomonas* able to produce the antibiotic 2,4 diacetylphloroglucinol which inhibits the growth of the phytopathogenic fungus. Fluorescent *Pseudomonas*, releasing 2,4 diacetylphloroglucinol, phenazine-1-carboxylic acid, phenazine-1-carboxamide, pyoluteorin, pyrrolnitrin, oomycinA, and pyocyanin and the *Bacillus* group, synthesizing surfactin, iturins, fengycins, kanosamine and zwittermycin A are two of the main producers of antibiotics in the rhizosphere. All these molecules have been tested both in *in vitro* and *in vivo* conditions and showed their ability to suppress diseases induced by fungal pathogens (Santoyo et al., 2021a, 2021b, 2021c).

HCN

According to the temperature, the oxygen availability and the bacterial strain involved, several bacterial species, mainly fluorescent pseudomonads, Burkholderia and Chromobacterium, can synthesize different amount of cyanide in the rhizosphere. The chemical reaction is based on glycine decarboxylation, involving the enzyme cyanuric acid synthase releasing CO₂ and HCN, the latter being toxic to phytopathogens through respiration inhibition at the cytochrome *c* oxidase level.

Bacteria able to carry out this reaction are usually resistant to cyanide. For pseudomonads, a **thiosulfate: cyanide sulfurtransferase** (rhodanese) converts cyanide to **thiocyanate** reducing the toxicity for these cells. In addition to its use for biocontrol of phytopathogens, HCN is involved in metal chelation, thereby increasing the amount of phosphate available for plant growth (Anand et al., 2020).

Cell wall degrading enzymes

Lytic enzymes released by PGPB such as β -1,3-glucanases, lipases, cellulases, and chitinases are involved in endosphere colonization as well as in nutrient cycling in soil but, due to the cell wall composition of these PGPB, they are among the main actors of antagonism towards fungal pathogens. Moreover, lytic enzymes produced by PGPB can damage both egg shells and cuticles of plant parasitic nematodes which are composed of a protein matrix and a chitin and a proteinaceous membrane, respectively. These two structures are the target of chitinases, proteases, peptidyl-peptide hydrolases, and gelatinolytic proteins (Gamalero and Glick, 2020).

Siderophores

In addition to being involved in plant growth promotion, siderophores can inhibit soil-borne diseases. Bacteria able to synthesize these chelating molecules can outcompete pathogenic microorganisms that are unable to synthesize iron chelators, or produce only a low level of siderophores, or release siderophores with a low affinity for ferric iron. Moreover, bacterial siderophores can sometimes induce systemic resistance in plants infected by pathogens. Due to their ability to bind a wide array of toxic metals, such as Cr³⁺, Al³⁺, Cu²⁺, Eu³⁺, and Pb²⁺, siderophores are often involved in detoxification of heavy-metal-contaminated substrates, becoming an effective and eco-friendly tool for heavy metal remediation (Swapan et al., 2020).

Induced systemic resistance

Plant beneficial bacteria can also boost intrinsic plant defenses against pathogens through induced systemic resistance (ISR) which also affects the non-colonized parts of the plant. This allows the plant to become less susceptible to further infections by a wide variety of pathogens including microbes, nematodes, and insects. The mechanism is based on microbial specific elicitors called microbe-associated molecular patterns (MAMPs), such as bacterial flagella, siderophores, antibiotics, biosurfactants and volatile organic compounds, which are detected by immunosensors located on the extracellular side of the plant cell plasma membrane known as pattern recognition receptors (PRRs). When signal perception occurs, the response of the plant is quite rapid, and typically within a few minutes the plant produces reactive oxygen species, phytoalexins, antimicrobial enzymes, phenolic compounds, or pathogenesis-related proteins, together with cell wall morphology changes. Although the triggering of ISR is strain and plant species dependent, this bacterial ability represents an innovative alternative to chemical control of plant diseases as well as biotic stresses (Meena et al., 2020).

Plant growth-promoting fungi

Mycorrhizal fungi

Soil-dwelling mycorrhizal fungi form a complex symbiosis with plant roots, forming a mycorrhiza. As this relationship may be up to 450 million years old, the communication between the two organisms is very intimate (Strullu-Derrien et al., 2018). The first studies of this symbiosis carried out by Frank (1885) distinguished two types of mycorrhizal fungi-plant interactions, those that could penetrate the cell wall (endomycorrhizae) and those that merely surrounded the roots (ectomycorrhizae). Later studies were able to determine that ectomycorrhizal fungi (ECM) mainly formed associations with trees and shrubs, whereas endomycorrhizae or arbuscular mycorrhizae (AM) are formed through a much wider range of associations with plants that are not only part of natural ecosystems, but also with crops of agronomic interest. The main contributions of arbuscular mycorrhizal fungi (AMF) to plants are direct by improving their nutrition (mainly P and poorly mobile micronutrients) as well as to increase their natural defenses against pathogens and herbivory, and indirect by enhancing soil structure and thereby improving water retention. Moreover, mycorrhizal fungi can form complex and abundant interactions with the roots of the plant, extending their area by a large percentage to obtain nutrients and water from the soil.

Trichoderma

Fungi belonging to the genus *Trichoderma* (teleomorph *Hypocrea*), which includes about 100 species, occur in all climatic zones, both in soil and the rhizosphere, and are usually classified as opportunistic, avirulent plant symbionts, and mycoparasites of plant pathogenic fungi. Thus, *Trichoderma* spp. are considered as promising fungal biocontrol agents. Plant inoculation with *Trichoderma* can provide several advantages, such as the suppression of soil borne disease carried out by phytopathogenic fungi, and the consequent improvement of plant health and growth. The main mechanisms used by *Trichoderma* spp. are based on efficient root colonization leading to niche exclusion against other possible deleterious root colonizers, antibiotic molecules production, increased resistance of plants through induced localized acquired resistance (LAR) and induced systemic resistance (ISR) and,

typically, on the synthesis of constitutively expressed fungal cell wall lytic enzymes. Several lytic enzymes (cellulase, chitinase, β -1,3 glucanases and proteases) synthesized by *Trichoderma* strains have been isolated, and their activity has been ascertained against both fungal pathogens and fungus-like organisms whose cell walls lack chitin e.g., *Pythium* and *Phytophthora*. The microparasitic action of *Trichoderma* begins with host recognition, through sensing oligochitin that are components of the pathogen cell wall. Subsequently, *Trichoderma* grows chemotropically in a targeted manner and its hyphae contact the pathogen mycelium. Then, the hyphae of *Trichoderma* synthesize lytic enzymes which induces the formation of pores leading to hyphae collapse and to the release of cellular contents that become nutrients for the biocontrol agent. Besides parasitizing fungi and oomycetes, several *Trichoderma* species can behave as parasites against eggs and larvae of nematodes and insects. Finally, many strains of *Trichoderma harzianum* and *T. atroviride* are efficient plant growth promoters, significantly increasing crop productivity (Sood et al., 2020).

Microbial consortia

Plants during their development and growth can be stimulated by associated microbiota. The microbiomes associated with plants can interact with each other and have consequences for their hosts. Some interactions that occur between the microbiomes are commensal, amensal, neutral, competition or predation, and cooperation. The interactions that are important for plant growth promotion are those where two or more organisms cooperate with each other to exhibit additive or synergistic actions. Additive interactions add up to twice the activity, while synergistic interactions can have actions that are more than twice the individual activities. Thus, a microbial consortium can be defined as a mixture of two or more microorganisms that work additively or synergistically to promote plant growth (Santoyo et al., 2021a, 2021b, 2021c).

In the literature, different mixtures of microorganisms have been reported that can function as consortia that benefit the growth of different plant crops under different growing conditions, whether in laboratory, greenhouse, or open field studies. Three types of consortia are described below that involve the joint application of bacteria-bacteria, bacteria-mycorrhizal fungi, and bacteria-*Trichoderma*.

Bacteria-bacteria

One of the widely studied consortia includes those where two or more bacterial groups participate in promoting plant growth. As mentioned above, PGPB exhibit different mechanisms of action, including nitrogen fixation, production of iron chelating compounds (siderophores), production of phytohormones, among others. Sometimes a bacterium can be very efficient in only one of these mechanisms. For example, rhizobia are excellent nitrogen fixers in legume plants, such as beans or alfalfa, but their antifungal or antibacterial activities are generally not very efficient. However, this deficiency can be supplied by another bacterial genus, such as *Pseudomonas* or *Bacillus*, bacterial genera with strong antagonistic activities towards plant pathogens, in addition to having ISR activities. Such activities are due to a diverse production of antimicrobial compounds such as antibiotics and lytic enzymes. On the other hand, these non-rhizobial bacterial groups are incapable of forming nodules on the roots of legumes. In this example, each bacterial group contains different mechanisms for promoting plant growth and their activities are complementary.

Bacteria-mycorrhizal fungi

Another example of microbial consortia where there is a complementarity of activities is between bacteria and mycorrhizal fungi. Bacteria have excellent activities to promote plant growth directly and indirectly, while mycorrhizal fungi are characterized by their effective improvement of plant nutrition. AMF are also widely known for improving the physical properties of the soil, since they form aggregates that allow them to retain more water, avoiding water stress in plants. The glomalins produced by AMF may be relevant for this function, which is unknown in bacteria. AMF can also function as soil bioremediators, thus improving soil health. Therefore, consortia composed of mixtures of bacteria and mycorrhizal fungi, can have beneficial effects on plants that are subjected to abiotic or biotic stress. For example, a consortium including the PGPB *Variovorax paradoxus* strain 5C-2, plus a nodule-forming and nitrogen-fixing strain, *Rhizobium leguminosarum* RCAM1066, and an arbuscular mycorrhizal fungus of the genus *Glomus* were able to confer resistance and improve the growth of pea plants (*Pisum sativum* L.) in soils contaminated with cadmium (Belimov et al., 2020).

Bacteria-Trichoderma

There is considerable evidence of the role played by microbial consortia composed of PGPB and different species of *Trichoderma*. Some of the synergistic functions that have been proposed include: stimulation of plant growth under normal and stress conditions, increase in plant nutrition, improvement of the root system, induction of ISR, and biocontrol of potential phytopathogens. Both groups of organisms contain species that, in addition to exhibiting the aforementioned beneficial mechanisms, are also excellent competitors in environments, such as the rhizosphere. The capabilities of this type of consortium have been proven not only in laboratory experiments, also in the field. Such is the case of the co-inoculation of *Brevibacterium halotolerans* and *Trichoderma harzianum* that synergistically enhanced the growth and oil yield in *Mentha arvensis* under both greenhouse and field conditions (Singh et al., 2019). There is evidence that PGPB and *Trichoderma* consortia carry out additive and synergistic interactions that boost the growth and health of crop plants (Kong et al., 2018; Santoyo et al., 2021a, 2021b, 2021c).

Commercialization of plant growth-promoting soil microorganisms

Bacteria

The selection of new PGPB as biofertilizers or biocontrol agents is based on a complex procedure consisting of: (i) isolation of microorganisms from soil, or different plant compartments, (ii) characterization of the bacterial physiological and metabolic activities involved in plant growth promotion, (iii) assessment of the effect of the PGPB on the growth of plants cultivated both under controlled and greenhouse conditions, and under both optimal and stressed environmental conditions, (iv) determination of the ecological safety of the PGPB, (v) inoculum formulation to satisfy the farmer's needs, and ensuring bacterial survival and viability on storage and shipping, and (vi) marketing and registration of the bacterial strain(s). However, it is rare to find a precise match between the traits determined *in vitro* and the behavior observed on plant growth in the field. Moreover, the use of microorganisms in agriculture is regulated in different countries by measures that ensure the efficacy of the microbiological product, as well as the quality control check and the protection of the intellectual property (for a recent review on regulatory policies for microorganism use, see Sundh et al., 2021).

The global biofertilizer market is expected to reach \$3.5 billion US by 2025 with free living and symbiont nitrogen fixing bacteria as the main commercial products (80%) (Basu et al., 2021). Recently, several manufacturers have begun to commercialize consortia based on a mixtures of rhizobia strains and rhizospheric or endophytic PGPB, bacteria and AMF, bacteria and plant growth-promoting fungi. Given that plant microbiomes contain hundreds to thousands of different microbes, complex natural consortia are beyond the ability of our current technology. On the other hand, it is also true that thanks to the recent development of -omics techniques, within the next 10–15 years, researchers will be able to formulate stable and efficacious complex microbial consortia based on a limited number of key microbial species.

Fungi

The method generally used to produce AMF inocula on a large-scale is based on systems of trap plant cultivation on substrate or in aeroponic boxes, as well as under *in vitro* conditions. Overall, the procedure is labor-intensive, and the costs related to the maintenance and reproduction of single AMF isolates, the formulation of the carrier and the management of the optimal conditions for the establishment of the mycorrhizal symbiosis are quite substantial. However, the benefits from mycorrhizal inocula are remarkable and can be categorized as both marketable (so that they can be evaluated as economic gain) and non-marketable. The first group includes improvement of mineral nutrition, productivity enhancement, reduction of watering and pesticide cost, and facilitation of seedling establishment. On the other hand, the improvement of food nutritional value, reduction of soil erosion and nutrient leaching, soil phytoremediation of organic and inorganic pollutants as well as the social benefits of education are all included in the second category (Gupta and Abbott, 2021). Similar to bacterial inoculants, a great increase in the number of companies working on mycorrhizal fungal technologies and associated products and services occurred in the past few decades. In 2019, the AMF-based biofertilizer market was evaluated to be close to \$268.8 million US, and is expected to reach \$621.6 million US by 2025. However, the increase of the AMF inocula market is accompanied by inconsistent performance under field conditions.

A thorough knowledge of the factors that affect mycorrhizal inocula efficacy in agroecosystems including climate conditions, plant host species, season, geographic location, soil characteristics, or even the good quality in the preparation of the inoculum should allow the achievement of the expectations from AM symbiosis.

Future prospects

The way to facilitate plant growth and feed the growing world population by judiciously employing the use of plant growth promoting microorganisms is challenging but promising. Crop plants typically recruit beneficial microorganisms from the soil and (as best as they can) exclude phytopathogens. These beneficial microorganisms (mostly bacteria and fungi), using a wide range of mechanisms, facilitate plant growth. Beneficial microorganisms may act individually or in concert with one another. As our knowledge of the fundamental mechanisms employed by these microorganisms continues to increase so does our ability to utilize these microorganisms in sustainable agricultural practice. In the future, it is hoped that we will eventually be able to limit or replace the bulk of the potentially deleterious chemicals that are using in agriculture and replace them with well-characterized and well-understood soil microorganisms.

References

- Aasfar A, Bargaz A, Yaakoubi K, Hilali A, Bennis I, Zeroual Y, and Meftah Kadmiri I (2017) Nitrogen fixing *Azotobacter* species as potential soil biological enhancers for crop nutrition and yield stability. *Frontiers in Microbiology* 25(12): 628379.
- Alori ET, Glick BR, and Babalola OO (2017) Microbial phosphorus solubilization and its potential for use in sustainable agriculture. *Frontiers in Microbiology* 8: 971.
- Anand A, Chinchilla D, Tan C, Mène-Saffrané L, L'Haridon F, and Weisskopf L (2020) Contribution of hydrogen cyanide to the antagonistic activity of *Pseudomonas* strains against *Phytophthora infestans*. *Microorganisms* 8: 1144.
- Basu A, Prasad P, Das SN, Kalam S, Sayyed RZ, Reddy MS, and El Enshasy H (2021) Plant Growth Promoting Rhizobacteria (PGPR) as green bioinoculants: Recent developments, constraints, and prospects. *Sustainability* 13: 1140.
- Belimov AA, Shaposhnikov AI, Azarova TS, Makarova NM, Safronova VI, Litvinskiy VA, Nosikov VV, Zavalin AA, and Tikhonovich IA (2020) Microbial consortium of PGPR, Rhizobia and Arbuscular Mycorrhizal Fungus makes pea mutant SGEcdt comparable with indian mustard in Cadmium tolerance and accumulation. *Plants (Basel, Switzerland)* 9: 975.

- Duca D, Lorv J, Patten CL, Rose D, and Glick BR (2014) Indole-3-acetic acid in plant-microbe interactions. *Antonie Van Leeuwenhoek* 106: 85–125.
- Etesami H, Ernami S, and Alikhani HA (2017) Potassium solubilizing bacteria (KSB): Mechanisms, promotion of plant growth, and future prospects: A review. *Journal of Soil Science and Plant Nutrition* 17: 4.
- Frank AB (1885) Ueber die auf Wurzelsymbiose beruhende Ernährung gewisser Bäume durch unterirdische Pilze. *Berichte der Deutschen Botanischen Gesellschaft* 3: 128–145. [English translation (2005), *Mycorrhiza* 15: 267–275].
- Frank AC, Saldierna Guzmán JP, and Shay JE (2017) Transmission of bacterial endophytes. *Microorganisms* 10: 70.
- Gamalero E and Glick BR (2020) The use of Plant Growth-Promoting Bacteria to prevent nematode damage to plants. *Biology* 9: 381.
- Glick BR (2020) *Beneficial Plant-Bacterial Interactions*, 2nd edn Heidelberg: Springer. 383 pages.
- Glick BR, Cheng Z, Czarny J, and Duan J (2007) Promotion of plant growth by ACC deaminase-containing soil bacteria. *European Journal of Plant Pathology* 119: 329–339.
- Gupta MM and Abbott LK (2021) Exploring economic assessment of the arbuscular mycorrhizal symbiosis. *Symbiosis* 83: 143–152. <https://doi.org/10.1007/s13199-020-00738-0>.
- Keswani C, Singh SP, Cueto L, García-Estrada C, Mezaache-Aichour S, Glare TR, Borriess R, Singh SP, Blázquez MA, and Sansinenea E (2020) Auxins of microbial origin and their use in agriculture. *Applied Microbiology and Biotechnology* 104: 8549–8565.
- Kong Z, Hart M, and Liu H (2018) Paving the way from the lab to the field: Using synthetic microbial consortia to produce high-quality crops. *Frontiers in Plant Science* 9: 1467.
- Martin FM, Uroz S, and Barker DG (2017) Ancestral alliances: Plant mutualistic symbioses with fungi and bacteria. *Science* 26(356): eaad4501.
- Masson-Boivin C and Sachs JL (2018) Symbiotic nitrogen fixation by rhizobia—The roots of a success story. *Current Opinion in Plant Biology* 44: 7–15.
- Meena M, Swapnil P, Divyanshu K, Kumar S, Harish Tripathi YN, Zehra A, Marwal A, and Upadhyay RS (2020) PGPR-mediated induction of systemic resistance and physiochemical alterations in plants against the pathogens: Current perspectives. *Journal of Basic Microbiology* 60: 828–861.
- Remus-Emsermann MNP and Schlechter RO (2018) Phyllosphere microbiology: At the interface between microbial individuals and the plant host. *New Phytologist* 218: 1327–1333.
- Saha M, Sarkar S, Sarkar B, Sharma BK, Bhattacharjee S, and Tribedi P (2016) Microbial siderophores and their potential applications: A review. *Environmental Science and Pollution Research International* 23: 3984–3999.
- Salazar-Cerezo S, Martínez-Montiel N, García-Sánchez J, Pérez-y-Terrón R, and Martínez-Contreras RD (2018) Gibberellin biosynthesis and metabolism: A convergent route for plants, fungi and bacteria. *Microbiological Research* 208: 85–98.
- Santoyo G, Gamalero E, and Glick BR (2021a) Mycorrhizal-bacterial amelioration of plant abiotic and biotic stress. *Frontiers in Sustainable Food Systems* 5: 672881.
- Santoyo G, Guzmán-Guzmán P, Parra-Cota FI, Santos-Villalobos SDL, Orozco-Mosqueda MDC, and Glick BR (2021b) Plant Growth stimulation by microbial consortia. *Agronomy* 11: 219.
- Santoyo G, Urtis-Flores CA, Loeza-Lara PD, Orozco-Mosqueda MDC, and Glick BR (2021c) Rhizosphere colonization determinants by Plant Growth-Promoting Rhizobacteria (PGPR). *Biology* 10: 475.
- Singh S, Tripathi A, Maji D, Awasthi A, Vajpayee P, and Kaira A (2019) Evaluating the potential of combined inoculation of *Trichoderma harzianum* and *Brevibacterium halotolerans* on oil yield on *Mentha arvensis* under greenhouse and field conditions. *Industrial Crop Production* 131: 173–181.
- Sood M, Kapoor D, Kumar V, Sheteiwy MS, Ramakrishnan M, Landi M, Araniti F, and Sharma A (2020) Trichoderma: The “secrets” of a multitasking biocontrol agent. *Plants (Basel)* 18(9): 762.
- Strullu-Derrien C, Selosse M-A, Kenrick P, and Martin FM (2018) The origin and evolution of mycorrhizal symbioses: From paleomycology to phylogenomics. *New Phytologist* 220: 1012–1030.
- Sundh I, Del Giudice T, and Cembalo L (2021) Reaping the benefits of microorganisms in cropping systems: Is the regulatory policy adequate? *Microorganisms* 9: 1437.
- Swapna KG, Tanmay B, and Ananda MC (2020) Microbial siderophore—A boon to agricultural sciences. *Biological Control* 144: 104214.
- Torsvik V, Salte K, Sørheim R, and Goksoyr J (1990) Comparison of phenotypic diversity and DNA heterogeneity in a population of soil bacteria. *Applied and Environmental Microbiology* 56: 776–781.

Inoculation with microorganisms

Maureen O'Callaghan and David Wright, Resilient Agriculture, AgResearch, Christchurch, New Zealand

© 2023 Elsevier Ltd. All rights reserved.

Introduction	428
Functions of soil microbial inoculants	429
Biofertilizers	429
General plant growth promotion and protection against soil-borne pests and diseases	431
Enhanced plant stress tolerance	431
Bioremediation	431
Climate change mitigation	432
Enhancing soil characteristics	432
Microbial consortia	432
Soil and environmental factors impact on inoculation	433
Competition between inoculants and resident soil microbial communities	433
Quality and delivery of soil inoculants	434
Field efficacy	435
Safety	436
Emerging approaches	436
Targeted design of microbial consortia	436
Novel formulation and delivery systems	436
Genetically modified microorganisms as inoculants	437
Outlook	437
References	437
Further reading	438

Abstract

Soil microbial inoculants are microorganisms that are used as sustainable alternatives to conventional inorganic fertilizers (biofertilizers), carry out specific functions including biocontrol of pests and diseases (biopesticides), or are used for bioremediation and enhancement of soil characteristics. Some soil inoculants such as nitrogen-fixing rhizobia have a long and successful history of use, but many new products perform inconsistently in the field. Better understanding of how soil inoculants interact with plants and the soil environment is helping to improve their effectiveness, as is the development of improved formulations and delivery systems.

Key points

- Significant gains in crop productivity have been achieved in part through increased use of agrichemicals including synthetic fertilizers and pesticides.
- Over-use of synthetic fertilizers is linked to deterioration of water quality and there is growing concern over pesticide residues accumulating in food chains and causing population decline in beneficial species such as pollinators
- Beneficial soil microorganisms capable of improving the supply of nutrients in soil or reducing the impact of soil pests and diseases can be used to inoculate soil.
- Microorganisms applied to soil face many challenges to become established in soil so they must be formulated and applied carefully.
- Selection of multi-functional microbial consortia is a promising approach to development of future inoculant products.
- Soil microbial inoculation provides a sustainable approach toward reduction of some synthetic agricultural inputs.

Introduction

Demand for food to support the rapidly increasing global population has driven widespread and ongoing intensification of agricultural production with greater crop productivity achieved largely through improved plant varieties and heavy use of agrichemicals. The environmental impacts from over-use of agricultural inputs such as fertilizers and pesticides are increasingly evident, for example nitrogen and phosphorus from agricultural fertilizers has led to significant deterioration of water quality across the globe. Detection of pesticide residues in food chains and within previously pristine environments is raising concern and has led

to the de-registration of many effective but environmentally persistent and highly toxic synthetic pesticides, leaving growers with reduced options for management of crop pests and diseases. At the same time, intensification of agriculture has led to significant global soil loss and degradation. Given the importance of soil in sustaining the world's human population, there is a clear need for more sustainable approaches that maintain agricultural production while better protecting the environment.

Decades of (predominantly laboratory) research on a wide range of soil and plant-associated microorganisms has demonstrated their potential benefits, and current pressure to develop more sustainable agriculture has led to a resurgence in research interest and a growing global industry around the use of microbial inoculants in agriculture. Soil microbial inoculants are defined as bacteria, fungi (and more rarely other microorganisms) that are introduced into soil to perform a specific function such as biocontrol or plant growth promotion (O'Callaghan et al., 2022).

Many soil and plant-associated microorganisms have beneficial properties that can be utilized to provide much needed solutions for agriculture (Fig. 1). For example, some strains can act as biopesticides or biocontrol agents to protect emerging seedlings from soil-borne pests and diseases, providing an alternative to synthetic pesticides. Others have the capacity to enhance plant uptake of soil nutrients and are referred to as biofertilizers and are a viable option as alternatives to conventional inorganic fertilizers. The mechanisms that underpin the beneficial impacts of microbial inoculants are varied and complex, so successful use of inoculants depends heavily on a sound understanding of the ecology of the microorganisms and their interactions with the soil environment.

Functions of soil microbial inoculants

Biofertilizers

As nitrogen (N) is the essential plant nutrient needed in the greatest amount, it is not surprising that nitrogen-fixing microorganisms are the most widely used soil microbial inoculants. Plant-associated symbiotic rhizobia have been used extensively in leguminous crops where their ability to meet plant N requirements has been shown to significantly reduce, and sometimes, eliminate the need for synthetic N fertilizers. For many legume varieties, inoculation with the correct rhizobial partner is essential for crop establishment where the required strain is not already present in soil and legume inoculation has been standard agricultural practice since the middle of the last century. In Brazil, inoculation of soybean with N-fixing *Bradyrhizobium* spp. can meet the N demand of the crop, eliminating the need for N fertilizers (a significant cost saving for farmers) with concomitant reductions in greenhouse gas emissions associated with fertilizer use (Zilli et al., 2021).

In addition to many plant-associated rhizobial inoculants, free-living N-fixing microorganisms such as *Azospirillum* and *Azotobacter* spp. are frequently used as soil inoculants. *Azospirillum* is one of the most studied genera of plant growth-promoting bacteria (PGPB) in the world and there are currently over 100 products based on this genus in South America alone (Cassan et al., 2020). Although some strains of *Azospirillum* have an affinity for certain crops, the major advantage of this genus is that useful strains are not always as plant specific as many rhizobia inoculants and can enhance the growth of numerous plant species. While *Azospirillum* is usually unable to supply all the required N for cereal and maize crops, reductions in fertilizer inputs of 25–50% have been reported. Beneficial effects of inoculation with *Azospirillum* result from multiple mechanisms including phytohormone production and stimulation of root growth, as well as N fixation (Cassan et al., 2020). Co-inoculation of soil with both symbiotic rhizobia and *Azospirillum* is also being used to boost crop production, as are other microbial consortia (see below).

With issues around security of supply, the ethics and cost of extraction, and the environmental impacts of non-renewable phosphorus (P) fertilizers on water quality, there is increasing interest in better management of the pool of P existing in soil using P-solubilizing microorganisms. Arbuscular mycorrhizal fungi (AMF) are widely used as soil inoculants, particularly for inoculation of trees and horticultural crops or for use in degraded or disturbed soils to assist plant colonization. AMF increase plant P uptake by physically extending the root systems through extension of the hyphae through soil and through their ability to acquire orthophosphate from soil solution at lower concentrations than roots. Mycorrhizal products have been shown to aid plant survival under stressful conditions, but scientific opinion is divided as to whether AMF inoculants deliver consistent benefits in cropping situations (Ryan and Graham, 2018; Zhang et al., 2019).

A wide range of P-solubilizing bacteria (PSB) have been identified from soil based on their capacity to solubilize sparingly soluble phosphate compounds in artificial media through secretion of low molecular weight organic ions, which acidify the medium. PSB (commonly *Pseudomonas*, *Bacillus*, and *Burkholderia* spp.) are reported to aid in de-sorption of inorganic P and complex organic P compounds from clay particles in soil by acidifying the rhizosphere, thereby increasing the solubility of precipitated inorganic P salts. Some soil microorganisms also produce various enzymes including phytases that enable microbes to access soil P but the relative contribution of microbial enzymes, in comparison with plant-derived soil enzymes remains unknown. While numerous strains have been shown to solubilize P in laboratory assays, it has not been possible to consistently demonstrate a clear link between improved soil P availability by PSB and crop productivity and it is possible that P-solubilizing microbes (PSM) are solubilizing soil P to meet their own requirements, rather than those of the plant (Raymond et al., 2021).

Better results in the field have been achieved with free-living P-solubilizing fungi. For example, Canadian growers reported that inoculation of wheat and canola with *Penicillium bilaii* resulted in average wheat and canola yield benefits of ~6%, but up to 66% (Harvey et al., 2009). While reported yield benefits vary, products based on *P. bilaii* are currently being marketed successfully, for example JumpStart® (Bayer Crop Science, 2022).

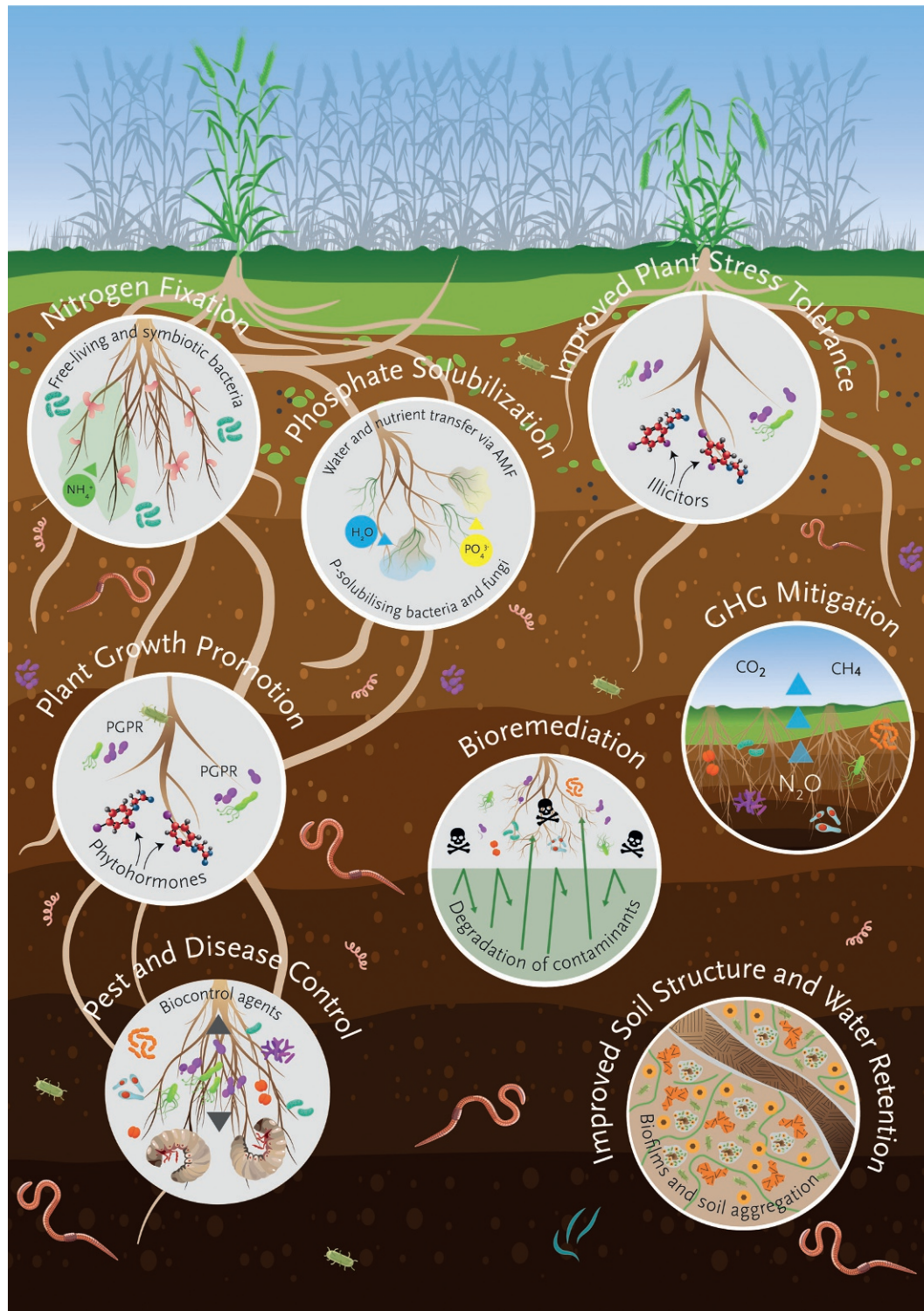


Fig. 1 Current and potential uses of soil microbial inoculants: Increased plant nutrient availability through activity of symbiotic and free-living nitrogen-fixing bacteria and solubilization of soil P by bacteria and fungi; enhanced plant tolerance to environmental extremes including induction of stress response pathways in plants; plant growth promotion via many mechanisms, e.g., production of phytohormones; bioremediation by microbial degradation of contaminants and co-inoculation to aid phytoremediation; reduced greenhouse gas (GHG) emissions through microbial activities; protection against soil-borne insect pests and plant pathogens; improved soil structure and water retention. Modified from: O'Callaghan, M., Ballard, R., Wright D. (2022). Soil microbial inoculants for sustainable agriculture: Limitations and opportunities. *Soil Use and Management* doi: 10.1111/sum.12811.

General plant growth promotion and protection against soil-borne pests and diseases

A wide range of plant-growth promoting bacteria and fungi are used as soil inoculants for many agriculturally significant plant species, including legumes, cereals, and vegetable crops. Many studies have reported that application of plant growth-promoting bacteria (PGPB) resulted in significant increases in root and shoot mass and reproductive yields and effects were even greater for above ground traits under drought conditions.

Beneficial microorganisms utilize many different mechanisms to enhance plant growth, including production of phytohormones, siderophores, antibiotics and other secondary metabolites (Lopes et al., 2021). Some also have capacity to protect plants from soil-borne diseases and pests and it is difficult to separate the mechanisms underpinning improved plant growth as in some cases inoculants are likely to be providing multiple benefits to the plant. For example, fungal inoculant products based on *Trichoderma* spp. enhance plant growth by multiple mechanisms including biocontrol of soil-borne pathogens, enhancing plant systemic resistance to pathogens and insects, and increased root proliferation (Schuster and Schmoll, 2010). Because of their multiple modes of action, *Trichoderma* strains are the active ingredient in hundreds of agricultural products commercialized worldwide.

Plant growth-promoting microorganisms (PGPM) used as soil inoculants are most often applied at time of sowing, when they provide protection of seeds and emerging seedlings against soil-borne pests and seed or root rot pathogens. While most often applied early in plant growth, their benefits can sometimes extend well beyond plant establishment, for example treatment of maize seed with two microbial biocontrol agents reduced populations of the pathogen *Fusarium verticillioides* and its associated mammalian-active mycotoxins in harvested grain (Pereira et al., 2011).

Soil inoculants are a good plant protection option when no chemical treatments are available, which is often the case for soil-borne plant pathogens and pests which are notoriously difficult to control. Many soil dwelling pests and pathogens are attracted toward germinating seeds and young roots, resulting in rapid destruction of newly planted crops. Introduction of an effective biocontrol agent directly into the root zone is therefore a useful strategy. This approach also has potential for management of soil dwelling insect pests, which in the past have been managed through application of persistent agrichemicals. Proof-of-concept of soil microbial inoculants for insect pest control has been achieved in the field through application of a soil bacterium (*Serratia entomophila*) for management of a scarab beetle larvae attacking pasture and cereal crops (O'Callaghan, 2016).

Enhanced plant stress tolerance

Abiotic stresses harm plant growth and physiology and reduce yield by obstructing cell metabolism or gene regulation within the cells. Abiotic stresses at play in soils include extreme temperatures (both high and low), moisture stress (drought or floods, rewetting cycles), pH, salinity and presence of heavy metals and other contaminants. Most crop plants don't possess advanced morphological or physiological adaptations necessary to withstand drought or high salinity situations. However, inoculation with selected soil and plant-associated microorganisms can increase plant tolerance to abiotic stress (Goswami and Deka, 2020). As discussed above, plant-growth-promoting bacteria can produce various plant growth regulators and hormones that improve the plant's antioxidant potential and accumulation of osmolytes. For example, a novel root-colonizing salt tolerant strain of *Bacillus* sp. was selected and tested for its capacity to maintain growth of wheat in a gradient of soil salinity (Ansari et al., 2019). Large salt concentrations adversely affected wheat growth and biochemical parameters relevant to photosynthesis, transpiration, and stress response in uninoculated plants, whereas seed yield and protein content of grain was maintained in inoculated plants. Another well-known inoculant, *Azospirillum brasilense*, has been shown to alleviate drought stress by multiple mechanisms including induction of systemic resistance, restricting Na⁺ absorption and increasing K⁺ and Ca⁺⁺ uptake by plants, and induction of root branching and density of root hairs (Fukami et al., 2018).

Efforts have been made to select for inoculants that are themselves tolerant to target abiotic stress. Drought tolerant bacteria produce exopolysaccharides (EPS) on their surfaces which improve water retention in the rhizosphere, protecting the root surface from mechanical damage. Microorganisms that have already evolved tolerance to soil abiotic stresses can also be selected directly from the environments where they are needed as soil inoculants. Acidic soils comprise more than 30% of the global land area and are characterized by a pH value lesser than 5.5 throughout the year. Acidic soils are typically deficient in available beneficial elements including P, molybdenum and magnesium and can contain inhibitory levels of aluminum and iron. Microorganisms survive within acidic soils by maintaining the intracellular levels of elements at near neutral levels or by dispensing with peripheral cellular structures that are in direct contact with the soil acidity. However, microbial inoculants suited to acidic soil conditions are rare, and greatly needed to support plant growth under these conditions.

Bioremediation

Anthropogenic activities, such as mining, intensive agriculture, e-wastes, and use of fossil fuels, can result in soil contaminants reaching hazardous levels. Increasing loss of soil functionality due to contamination by heavy metals, excessive trace element loadings and organic pollutants is a global environmental issue that poses significant risks to human health and both productive and natural ecosystems. Use of soil inoculants for bioremediation of contaminated soils is not new but interest in utilizing microorganisms to effect in situ remediation of contaminated soils has intensified due to the expense of conventional remediation. In particular, the combined use of carefully selected inoculants with site-tailored cropping systems (phytomanagement) is proving a useful tool to help plants cope with soil contamination. Inoculation with plant growth-promoting bacteria and AMF have been

most used, but responses can vary according to specific combinations and with different plant species. For example, some strains of *Burkholderia* sp. have been shown to reduce cadmium accumulation in plants while other research has reported increased mobilization and increased plant accumulation (Guo et al., 2020).

Much better understanding of the efficiency and the mechanisms by which soil inoculants function in complex contaminated soil environments is needed. Recent work showed that a bacterial inoculant *Comamonas testosteroni*, capable of degrading polycyclic aromatic hydrocarbons (PAHs), also increased synergistic effort to degrade PAHs among indigenous soil bacteria (Lu et al., 2022).

Many microorganisms with capacity to degrade various widely used pesticides have been identified but recent research has begun to focus on the use of microbial consortia rather than single strains. Microbial consortia are discussed further below, but in the bioremediation context, it is thought that consortia may be able to degrade pesticides more quickly and efficiently by sharing the metabolic labor to attack and breakdown the molecular structure.

Climate change mitigation

Microbial inoculants potentially offer sustainable solutions to climate change indirectly through their influence on plant growth and delivery of various agroecosystem functions and services, allowing reductions in traditional agrichemicals. Examples of direct mitigation are scarce to date but it is an area of increasing research interest. Reduction of methane emissions from agriculture is a key goal of greenhouse gas mitigation strategies and recent work has demonstrated that co-inoculation of efficient methane-utilizing bacterium *Methylobacterium oryzae* and plant growth-promoting bacterium *Paenibacillus polymyxa* reduced cumulative methane emissions in rice paddy fields (Rani et al., 2021). N₂O-reducing bacteria have also been reported to decrease soil N₂O emissions in soil microcosms and genetically engineered CO₂-fixing microorganisms reduced soil emission of CO₂ by promoting carbon sequestration (in culture only) (Liu et al., 2022).

Enhancing soil characteristics

Soil bioengineering—using microbial inoculants to modify soil chemistry and structure—is a novel approach that has not yet been widely explored. Microorganisms play a critical role in soil formation and soil aggregate stabilization and some EPS-producing bacteria can improve soil aggregation and adherence of soil to plant roots, demonstrating their potential as inoculants. For example, a strain of *Pseudomonas putida* (selected for its ability to survive at low soil moisture potential) colonized the soil adhering to sunflower roots and increased the development of stable soil aggregates. Increased plant biomass and stress tolerance was attributed to plant growth-promoting compounds produced by the bacteria within the biofilm produced on the roots of the seedlings, and through production of EPS which increased soil aggregation, thereby maintaining higher water potential around the roots (Sandhya et al., 2009). Improvements in the macroporosity of soil have been reported with a range of other plant-growth promoting bacteria, suggesting that EPS-producing bacteria could be used as inoculants to regulate water stress (Gouzou et al., 1993). Restoration of soil quality and function is a key goal of rehabilitation at mine sites and the use of microbial inoculants to improve soil aggregation in the post-mining context is also being considered (da Silva et al., 2022).

Soil bacteria can also produce a range of other surface-active agents or biosurfactants which aid in colonization of soil. For example, *Pseudomonas fluorescens* strain SBW25 produces the surfactant viscosin capable of altering local soil water distribution in soil microcosm experiments (Fechtner et al., 2011). It remains to be seen whether the beneficial properties of EPS- and surfactant-producing bacteria could be harnessed at large scale but given the increasing areas of cropping land impacted by drying conditions and salt stress, this approach warrants further attention.

The enhancement of the strength of soil through the activity of microorganisms or their by-products is a relatively new area of research known as biocementation. Urease-producing microorganisms such as *Sporosarcina pasteurii* were found to induce carbonate precipitation which improved the geotechnical properties of sand (Badiie et al., 2019). The studies referenced here hint at the potential for soil inoculants to be used to improve soil characteristics, including increased resistance to erosion, but much more proof is needed in this area.

Microbial consortia

Because of the highly complex nature of soil microbial communities and the diversity of niches available for colonization in soil, it is reasonable to expect that a mixture of compatible microorganisms with beneficial traits will enhance the likelihood of inoculated microorganisms establishing and persisting in different locations in soil. In addition, there is now ample evidence that microbial communities often work cooperatively to carry out specific functions. To date, most soil inoculants currently on the market contain a single inoculant strain but there is increasing evidence that better outcomes are achieved when multiple microorganisms are inoculated simultaneously. Co-inoculation of wheat seeds with the well-studied *Penicillium bilaiae* and a strain of *Bacillus simplex*, isolated from hyphae of *P. bilaiae* (and previously shown to stimulate *P. bilaiae* growth and P solubilization in the laboratory), increased plant P uptake from soil containing little available P above levels achieved with either inoculant alone (Hansen et al., 2020). Another field study examined the impact on growth of faba bean and wheat of inoculation with plant growth-promoting rhizobacteria (PGPR) alone; rhizobia alone; a mixture of PGPR and rhizobia; AMF alone; and mixture of PGPR, rhizobia and AMF. Inoculation with all three groups of microorganisms together resulted in significant improvement in crop biomass and improved the nutrient content of the crops, over that achieved without inoculation or the use of only single inoculants (Raklami et al., 2019).

Similar outcomes were found when soils were inoculated with a PGPR consortium of three strains effective against soil-borne disease of sweet pepper (Yang et al., 2014). Disease was suppressed, plant yield was increased, and improved soil nutrient levels were highly correlated with abundances of the inoculated species. These detailed studies are beginning to build the knowledge base needed to guide development of efficacious soil inoculant products based on consortia.

The studies described above used combinations of well-characterized microorganisms but many polymicrobial soil inoculants on the market today contain as many as 10 or more different microorganisms, sometimes with relatively low levels of each. Product claims are often very general such as “probiotic,” “living biostimulant,” “soil conditioner” and the mode of action of the consortium members are not well understood. It is unclear how the combinations of microorganisms brought together within such products have been selected and whether they do in fact work in unison or are competitive.

Developing a microbial consortium-based inoculum will involve selection of the desired strains, their ability to survive within a consortium and in the soil environment to confer the desired benefits in the field. Systematic approaches to development of multi-species inoculants that account for the specific properties and biocompatibility of the co-inoculant species are needed. While very laborious, deliberate design of a multi-species and AMF-biocompatible inoculant of phosphate rock (PR)-solubilizing bacteria yielded two effective combinations of co-inoculants that significantly enhanced maize biomass and uptake of N, P and K (Magallon-Servin et al., 2020), likely through the inclusion of P-solubilizing strains that utilized a range of different P-plant acquisition strategies. Similar outcomes with co-inoculation have been reported in other recent studies (Ribeiro et al., 2022).

Soil and environmental factors impact on inoculation

The heterogeneity of the soil as the receiving environment provides both opportunities and challenges for establishment and persistence of soil inoculants. Neutral pH, adequate moisture, adequate organic carbon, and low salinity favor microbial colonization, function and persistence and unfavorable soil conditions (such as pH extremes) are best managed prior to application unless the inoculant has been selected to operate under specific environmental conditions (Lopes et al., 2021). Impacts of pH have been reported for several inoculants including the P-solubilizing fungus *Penicillium* spp. where greater inoculation response was achieved in neutral and alkaline soils, while rhizobia inoculants vary widely in their ability to function in acidic soils (O’Callaghan et al., 2022). Soil salinity can also restrict microbial activity in soil, sometimes requiring the selection of salt tolerant inoculant strains. An example is a salt tolerant strain of *Sinorhizobium medicae* selected for successful nodulation of the pasture legume messina (*Melilotus siculus*) in saline soil, because of its improved survival over summer (Bonython et al., 2011).

For broad-acre crops grown in rain-fed agricultural systems, adequate soil moisture and temperature support use of soil inoculants, but in Mediterranean climates, conditions for microbial function may only be optimal around the time of crop establishment. However, some studies report better inoculation responses when plants are under environmental stress, such as in dryland agriculture (Schütz et al., 2018). Drying and wetting cycles in soil can be detrimental to microbial survival so inoculants are generally best introduced into the soil when moisture is adequate for seed germination. Microbes able to survive on seed in high numbers may have an advantage in these environments.

Knowledge of the soil nutrient status should guide use of inoculants as N-fixing and P solubilizing inoculants will likely decrease as levels of soil N and P increase. Mathematical modeling suggests that inoculation with PGPM is more efficient in nutrient poor soils, or stressed soils, because the development of the resident microflora is inhibited (Lopes et al., 2021). Other nutrient deficiencies may limit success of inoculants, for example molybdenum is important for N fixation by rhizobia and other elements may be important in microbial degradation of pesticides. It is important that deficiencies are addressed at time of application and interactions between elements must also be considered; inoculants intended to build soil organic C and improve soil structure will only be effective in the presence of sufficient soil N.

Competition between inoculants and resident soil microbial communities

After inoculation, cell numbers typically undergo a rapid decline, in part because of intense competition with the resident soil microbiome for niches and resources in soil, and predation by protozoa and nematodes. Soil microbial communities are highly complex and diverse and are typically resistant to invasion by applied microorganisms. In addition, the inoculant will almost always be numerically disadvantaged. For example, high quality peat inoculants will typically be applied at a field rate equivalent to 2.5×10^{10} rhizobia per ha, while naturalized soil communities of rhizobia often exceed 2.5×10^{12} rhizobia per ha. Placement of high numbers of inoculants very close to the roots to facilitate colonization of roots is therefore very important.

The presence of competitive microbes in the soil is less of a problem where host specificity is strongly expressed, where the function targeted by the inoculant microbe is specific rather than generic, or where the number of competitors has declined over time because the plant host is not consistently present, or soil conditions are unfavorable to the survival of the competing microflora.

Soil disturbances that disrupt microbial communities, such as tactical cultivation, may provide the opportunity to tip the balance in favor of inoculant microbes or reduce the impact of antagonists. Carefully designed formulations which provide the inoculant with a competitive edge, through provision of resources or safe niches within soil are also under investigation (see below). Such

tailored resource pulses may be able to disrupt the equilibrium within the resident soil microbial community and allow the inoculant to become established or proliferate.

Quality and delivery of soil inoculants

Standard production methods for inoculants utilize nutrient rich cultivation media and often produce cells that are sensitive to the abiotic and biotic conditions encountered in soil. Pre-conditioning cells to stresses encountered in the soil environment during production and formulation processes can enhance the ability of applied cells to survive and establish. Microbial inoculants require formulation to ensure acceptable shelf life prior to application, to enhance survival in the field, and to enable delivery of inoculants using common agricultural machinery. Peat-based formulations have been traditionally used for the delivery of soil inoculants. Peat provides a good niche for microbial growth when outside the soil and protection once applied. However, as peat is a non-renewable resource, new formulations are required utilizing cheap, abundant carriers; nevertheless, most alternatives tested to date have been considered inferior to peat (Bashan et al., 2014) and while synthetic formulations based on polymers and more recently nanomaterials have been proposed, raw materials for these formulations are expensive compared to peat and other organic matrices and have only been used experimentally to date. The shelf life of inoculants can be limiting compared to traditional agrichemicals and some inoculants require refrigeration until use. Upskilling of merchants and growers, accustomed to handling chemicals, to store and use biological products correctly, will be critical to ensure the success of inoculants in the field. Although inoculant formulations based on solid carriers, such as peat, generally have a shorter shelf life than those formulated as liquid products, which may contain additional nutrient and cell protectant compounds, solid formulations provide superior protection for the cells in the soil post application.

Most soil inoculants are applied early in plant development, usually at time of sowing in furrow as liquid or granules, on seed as slurry or as pre-coated seed treatments (Fig. 2). In-furrow application of inoculants is often successful but can be more expensive as large quantities of inoculant are required. Granules are increasingly used as they are convenient to apply, and correct placement and inoculant application rate are easily managed. Granules can also protect the inoculant from seed-applied fungicides and pesticides. Liquid formulations for in-furrow treatment are also popular with growers as they are easily handled but may not provide the inoculant with sufficient protection against unfavorable soil conditions since they lack the protection of a carrier such as peat and are more likely to be washed through the soil profile under irrigation or heavy rainfall.

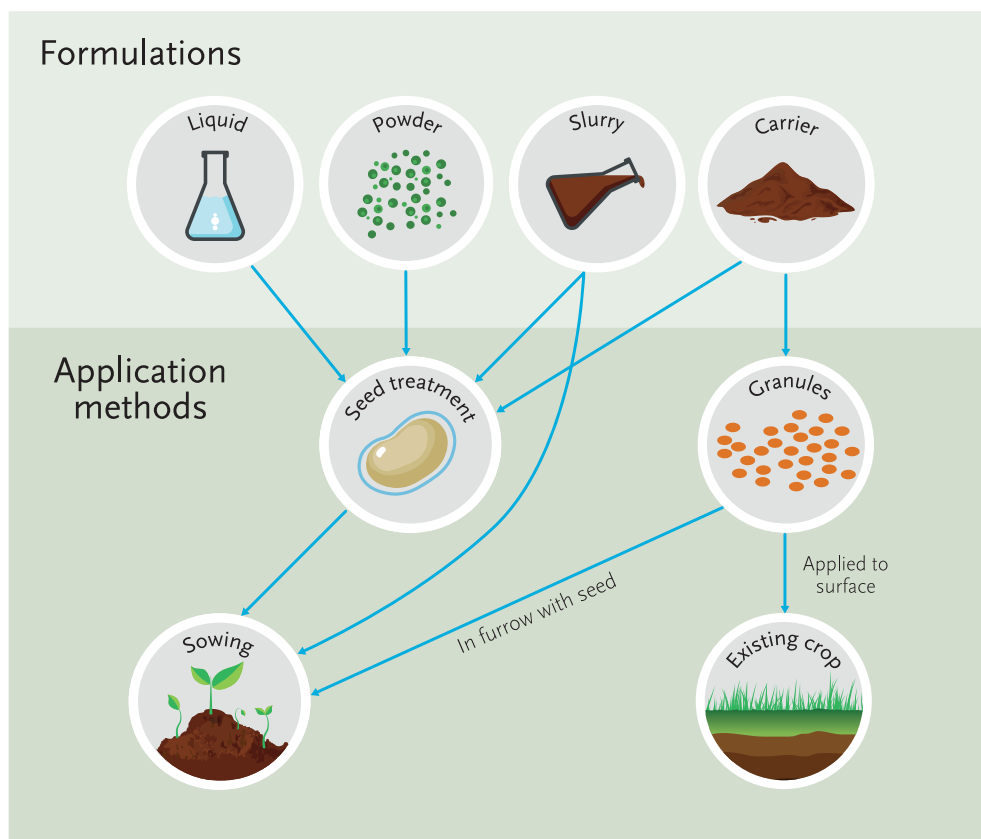


Fig. 2 Commonly used formulations and application methods for soil inoculation.

Seed treatments have been used to deploy beneficial microbes for decades and they tend to deliver consistent results and are very cost effective (O'Callaghan, 2016). Seed inoculation typically requires less inoculant to be applied than other application methods, although (small) size of seed can limit the amount of inoculant applied. However, seed inoculation is far more challenging than coating of seed with agri-chemicals as the microorganisms must survive the environmental stresses of the seed treatment process, in particular high temperature and desiccation. In addition, incompatibility with fungicides and insecticides applied to seed can be detrimental to inoculant survival, and careful formulation may be required. These issues may be compounded where seed is pre-inoculated and stored, sometimes for months, before distribution to growers. Seed inoculation close to the time of sowing reduces this risk. There remains a need to find biomaterials and seed treatment processes that provide a microenvironment to protect microbes from desiccation while also having the mechanical properties needed to form a robust but functional layer around the seed. The inclusion of various secondary metabolites (flavonoids, phytohormones and nodulation factors) to formulations and seed treatments also shows promise in improving efficacy of inoculants. For example, the addition of small amounts of flavonoids to rhizobial formulations alleviates the effects of adverse conditions and enhances nitrogen fixation (Morel et al., 2016). Improved seed treatment processes are needed to ensure consistent loadings of functional mixtures of physiologically active and primed inoculants on seed that maximize seedling protection and vigor (Fig. 3).

In situations where application is required outside of the traditional in-furrow or seed treatment methods, such as inoculation of established pasture or into areas inaccessible to farm machinery, methods of application are limited. There are a few liquid formulations of soil inoculants on the market that are designed to be sprayed onto the soil or foliage. While there are several biological mechanisms that may assist the movement of inoculants through soil, the success of such applications is likely to be limited as microbes applied by spray will be subjected to desiccation and UV light and will to a lesser or greater extent rely on precipitation to move into the soil profile. Formulations designed to provide a protective environment enabling the survival of inoculants under adverse conditions that release when environmental conditions are optimal for inoculant movement have been described (Wright et al., 2017) and have potential for aerial application to inaccessible areas.

Field efficacy

Despite the potential demonstrated by many beneficial microorganisms, relatively few have been developed into commercially available products, and even fewer of these have been subjected to robust validation under a range of soil and environmental conditions. Field efficacy of microbial inoculants can vary widely across sites and climatic conditions, and inconsistent performance of some inoculant products is an ongoing science and industry challenge. However, it must be remembered that an inoculant will be beneficial only if it is targeting a significant limitation in the soil where it is applied. Products claiming improved nutrient availability will only be effective where a nutrient is lacking, in the same way that fertilizers provide benefit where specific nutrients are limiting. Similarly, use of an inoculant to control a soil-borne plant disease will only provide benefit where the targeted pathogen is present, and conditions are conducive to expression of disease. Agronomic decisions around use of inoculants should



Fig. 3 Future biological seed treatments may deliver multiple beneficial microorganisms and metabolites that work together to increase seed germination and plant establishment and yield.

mainly be based on prior knowledge and testing for soil constraints, as garnered from pre-testing soil fertility, or farmers observations of production-limiting populations of pests. Under some circumstances inoculant application can be justified on the basis that it provides insurance against infection and crop loss, so long as the cost is not prohibitive. Rhizobia (biofertilizer) inoculants are often used as an insurance policy when legumes are sown, because the cost of a crop failure can be large, relative to the small cost of inoculation.

Where scientifically validated products are used according to the manufacturer's instructions, significant increases in crop production have been reported, especially for inoculants used with legumes. Because legume inoculants have been used for many decades in association with a range of crops and under various soil and environmental conditions, robust meta-analyses of their efficacy can be undertaken, showing the important contribution of inoculants to global agriculture (e.g., Schütz et al., 2018).

Safety

Knowledge on the safety of soil inoculant products varies widely and has not been well studied for many products. Registration of inoculants as biopesticides usually requires submission of field trial data to support efficacy claims and evidence of mammalian and environmental safety, including lack of phytotoxicity and impact on key non-target species, such as earthworms, pollinators, and the indigenous soil microbial community. Users of biopesticidal inoculants can therefore have some confidence in the product claims as the data is often published in peer reviewed journals or independently evaluated as required by regulatory authorities. In contrast, inoculants described as biofertilizers, biostimulants and plant growth promoters have generally not been subjected to the same level of scrutiny in terms of their potential non-target effects.

Other factors to be considered in assessing the likelihood of non-target effects of inoculants include the establishment and extent of spread of microorganisms beyond the site of application. There are challenges in tracking applied strains within the complexity of the resident soil microbial community (Manfredini et al., 2021) and establishment and spread will be impacted by a multitude of environmental and biological factors, including the innate invasive capacity of the inoculant and its persistence under wide ranging field conditions. Research shows that many inoculants modify the soil microbial communities in various ways in the short or long term (Cornell et al., 2021). In some studies, the soil microbial community did not revert to its initial state after inoculation, even when the inoculant strain(s) had declined quickly following application. In contrast, other studies report only short-lived changes in the soil microbial community, that are minor in comparison with seasonal fluctuations. Mechanisms underpinning changes in the structure of soil microbial communities range from resource competition to direct antagonism and it is often not clear whether soil functions have also been impacted.

The deliberate introduction of microbial inoculants into new ecosystems is attracting increasing scrutiny, reflecting growing awareness of the need to protect indigenous soil microbial diversity, and increasing biosecurity concerns around movement of microorganisms across borders and the potential for widespread application of inoculants to trigger invasions by formerly non-invasive crops or co-occurring plants. Inoculant products containing large numbers of diverse, unnamed, and uncharacterized microorganisms is cause for concern, especially for soil inoculants intended for global markets, as the release of alien species poses a potential risk of disrupting integrity of indigenous communities. This is particularly true for inoculants composed of microbial consortia, because of the technical challenges and complexities of monitoring multiple microorganisms simultaneously in the environment.

Emerging approaches

Targeted design of microbial consortia

Microbial consortia are ubiquitous in soil and can outperform single microorganisms at multiple tasks. When used as inoculants, they have the potential to endure more changeable environments, as they may occupy a broader resource niche in soil together versus alone, allowing them to compete more effectively with resident soil microorganisms. In addition, the complexity of soil microbiome–plant interactions suggests that use of consortia may be a more useful strategy than using single microorganisms as inoculants. While inoculant products containing multiple species are available on the market, the rationale behind selection and collective performance of the consortia is often questionable. Several recent studies have demonstrated robust scientific approaches to selection and design microbial consortia capable of delivering greater benefits than single-species inoculant products. For example, Paredes et al. (2018) described a new method to design and predict the impact of novel bacterial combinations that modified phosphate accumulation in the plant shoot by constructing synthetic communities and testing their effects on plant phenotypes. Another approach—state-of-the-art metabolic modeling—was used to predict the efficiency of degradation of the herbicide atrazine by constructed microbial consortia, as well as the mechanisms underpinning herbicide inactivation. While not yet validated in field trials, this approach is another promising tool for consortia design (Faust, 2019).

Novel formulation and delivery systems

The spatial distributions of microbial populations across landscapes are highly variable and broadacre application of inoculants is expensive. Remote sensing and targeted application technologies currently in development for the precision application of pest and

disease controls, and nutrient management programs have potential as tools to enable the targeted application of soil inoculants, reducing the quantity of inoculant needed to achieve a response. Novel formulations that provide physical refuge for inoculants enable sustained, multiple invasions of inoculant into the local environment over a longer period than traditional formulations and improve the chances for establishment of inoculum in the environment.

Incorporation of growth substrates within a protective matrix allows multiplication of the inoculum in the environment after application and provides sustained delivery of inoculum. The introduction of inoculants via the seed microbiome using the seed as a protective carrier is a novel concept. This is achieved by spraying bacteria onto the flowers of crop plants to colonize the flower and subsequently become incorporated into the seed where they are protected from competition in the soil and rhizosphere and can colonize the developing plant at an early stage of development (Sessitch et al., 2019.) The feasibility of this approach has yet to be validated at scale.

Genetically modified microorganisms as inoculants

The potential use of genetically modified microorganisms (GMMs) as soil or plant inoculants has been proposed for decades but they have not yet seen practical use in the field. Advances in synthetic biology-based approaches is leading to generation of precisely engineered bacteria with additional beneficial functions. For example, inoculation of model experimental plant species *Arabidopsis thaliana* with transgenic bacterial strains engineered to over-express several phytase enzymes supported improved plant growth under phosphorus-limited conditions (Shulse et al., 2019). While these experiments were not conducted in soil, the approach can be considered a first proof-of-concept in development of effective phosphate-mining bacteria for future use in agriculture. However, while genetic modification of crops is widely accepted in agriculture, the use of GMMs has not yet become common practice. Further work on the stability and biosafety of engineered microbial inoculant species is needed.

Outlook

Progress is being made toward development of more reliable and well-defined soil inoculants, with use of some recent scientifically validated products being guided by detailed information on individual strains, their mode of action, and when and how to best store and apply the product. Deeper knowledge of the plant microbiomes of key agricultural crops and their interactions with the soil microbiome, in combination with detailed modeling is facilitating improved selection and development of microbial consortia with multiple functionalities capable of establishing under a range of geographical and environmental conditions.

References

- Ansari FA, Ahmada I, and Pichtel J (2019) Growth stimulation and alleviation of salinity stress to wheat by the biofilm forming *Bacillus pumilus* strain FAB10. *Applied Soil Ecology* 143: 45–54.
- Badiee H, Sabermahani M, Tabandeh F, and Saeedi Javadi A (2019) Application of an indigenous bacterium in comparison with *Sporosarcina pasteurii* for improvement of fine granular soil. *International Journal of Environmental Science and Technology* 16: 8389–8400.
- Bashan Y, de Bashan LE, Prabhu SR, and Hernandez JP (2014) Advances in plant growth-promoting bacterial inoculant technology: formulations and practical perspectives (1998–2013). *Plant and Soil* 378: 1–33.
- Bayer Crop Science (2022) Product Information for JumpStart. Accessed 30/11/2022. <https://www.cropscience.bayer.us/seedgrowth/acceleron/other-solutions/jumpstart-wettable-powder-inoculant>.
- Bonython AL, Ballard RA, Charman N, Nichols PG, and Craig AD (2011) New strains of rhizobia that nodulate regenerating messina (*Melilotus siculus*) plants in saline soils. *Crop & Pasture Science* 62: 427–436.
- Cassan F, Coniglio A, Lopez G, et al. (2020) Everything you must know about *Azospirillum* and its impact on agriculture and beyond. *Biology and Fertility of Soils* 56: 461–479.
- Cornell C, Kokkoris V, Richards A, et al. (2021) Do bioinoculants affect resident microbial communities? A meta-analysis. *Frontiers in Agronomy* 3: 753474. <https://doi.org/10.3389/fagro.2021.753474>.
- da Silva GOA, Southam G, and Gagen EJ (2022) Accelerating soil aggregate formation: A review on microbial processes as the critical step in a post-mining rehabilitation context. *Soil Research*. <https://doi.org/10.1071/SR22092>.
- Faust K (2019) Microbial consortium design benefits from metabolic modelling. *Trends in Biotechnology* 37: 123–125.
- Fechtner J, Koza A, Dello Sterpaio P, Hapca SM, and Spiers AJ (2011) Surfactants expressed by soil pseudomonads alter local soil-water distribution, suggesting a hydrological role for these compounds. *FEMS Microbiology Ecology* 78: 50–58.
- Fukami J, Cerezini P, and Hungria M (2018) *Azospirillum*: Benefits that go far beyond biological nitrogen fixation. *AMB Express* 8: 73.
- Goswami M and Deka S (2020) Plant growth-promoting rhizobacteria—Alleviators of abiotic stresses in soil: A review. *Pedosphere* 30: 40–61.
- Gouzou L, Burtin G, Philipp R, Bartoli F, and Heulin T (1993) Effect of inoculation with *Bacillus polymixa* on soil aggregation in the wheat rhizosphere: preliminary examination. *Geoderma* 56: 479–490.
- Guo J, Muhammad H, Lv X, et al. (2020) Prospects and applications of plant growth promoting rhizobacteria to mitigate soil metal contamination: A review. *Chemosphere* 246: 125823.
- Hansen V, Bonnichsen L, Nunes, et al. (2020) Seed inoculation with *Penicillium bilaiae* and *Bacillus simplex* affects the nutrient status of winter wheat. *Biology and Fertility of Soils* 56: 97–109.
- Harvey PR, Warren RA, and Wakelin S (2009) Potential to improve root access to phosphorus: The role of non-symbiotic microbial inoculants in the rhizosphere. *Crop & Pasture Science* 60: 144–151.
- Liu X, Le Roux X, and Salles JF (2022) The legacy of microbial inoculants in agroecosystems and potential for tackling climate change challenges. *iScience* 25(3). <https://doi.org/10.1016/j.isci.2022.103821>.

- Lopes MJS, Dias-Filho MB, and Gurgel ESC (2021) Successful plant growth-promoting microbes: Inoculation methods and abiotic factors. *Frontiers in Sustainable Food Systems*. <https://doi.org/10.3389/fsufs.2021.606454>.
- Lu X, Sun X, Jiang X, Cui Y, Li X, and Cui J (2022) Effects of *Comamonas testosteroni* on dissipation of polycyclic aromatic hydrocarbons and the response of endogenous bacteria for soil bioremediation. *Environmental Science and Pollution Research* 29. <https://doi.org/10.1007/s11356-022-21497-z>.
- Magallon-Servin P, Antoun H, Taktek S, and de Bashan LE (2020) Designing a multi-species inoculant of phosphate rock-solubilizing bacteria compatible with arbuscular mycorrhizae for plant growth promotion in low-P soil amended with PR. *Biology and Fertility of Soils* 56: 521–536.
- Manfredini A, Malusà E, Costa C, et al. (2021) Current methods, common practices, and perspectives in tracking and monitoring bioinoculants in soil. *Frontiers in Microbiology* 12: 698491. <https://doi.org/10.3389/fmicb.2021.698491>.
- Morel MA, Cagide C, and Castro-Sowinski S (2016) The contribution of secondary metabolites in the success of bioformulations. In: Arora N, Mehnaz S, and Balestrini R (eds.) *Bioformulations for Sustainable Agriculture*, pp. 235–250. New Delhi: Springer.
- O'Callaghan M (2016) Microbial inoculation of seed for improved crop performance: Issues and opportunities. *Applied Microbiology and Biotechnology* 100: 5729–5746.
- O'Callaghan M, Ballard R, and Wright D (2022) Soil microbial inoculants for sustainable agriculture: Limitations and opportunities. *Soil Use and Management*. <https://doi.org/10.1111/sum.12811>.
- Paredes S, Gao T, Law T, et al. (2018) Design of synthetic bacterial communities for predictable plant phenotypes. *PLoS Biology* 16: e2003962.
- Pereira P, Ibáñez SG, Agostini E, and Etcheverry E (2011) Effects of maize inoculation with *Fusarium verticillioides* and with two bacterial biocontrol agents on seedlings growth and antioxidative enzymatic activities. *Applied Soil Ecology* 51: 52–59.
- Raklami A, Bechtaoui N, Tahiri A, et al. (2019) Use of rhizobacteria and mycorrhizae consortium in the open field as a strategy for improving crop nutrition, productivity and soil fertility. *Frontiers in Microbiology* 10: 1–11.
- Rani V, Bhatia A, and Kaushik R (2021) Inoculation of plant growth promoting-methane utilizing bacteria in different N-fertilizer regime influences methane emission and crop growth of flooded paddy. *Science of the Total Environment* 775: 145826. <https://doi.org/10.1016/j.scitotenv.2021.145826>.
- Raymond NS, Gómez-Muñoz B, van der Bom FJT, et al. (2021) Phosphate-solubilising-microorganisms (PSM) for improved crop productivity: A critical assessment. *New Phytologist* 229: 1268–1277.
- Ribeiro VP, Gomes EA, de Sousa SM, et al. (2022) Co-inoculation with tropical strains of *Azospirillum* and *Bacillus* is more efficient than single inoculation for improving plant growth and nutrient uptake in maize. *Archives of Microbiology* 204: 143. <https://doi.org/10.1007/s00203-022-02759-3>.
- Ryan MH and Graham JH (2018) Little evidence that farmers should consider abundance or diversity of arbuscular mycorrhizal fungi when managing crops. *New Phytologist* 220: 1092–1107.
- Sandhya V, Ali S, Grover M, Reddy G, and Venkateswarlu B (2009) Alleviation of drought stress effects in sunflower seedlings by exopolysaccharides producing *Pseudomonas putida* strain GAP-P45. *Biology and Fertility of Soils* 46: 17–26.
- Schuster A and Schmöll M (2010) Biology and biotechnology of *Trichoderma*. *Applied Microbiology and Biotechnology* 87: 787–799.
- Schütz L, Gättinger A, Meier M, et al. (2018) Improving crop yield and nutrient use efficiency via biofertilization—A global meta-analysis. *Frontiers in Plant Science* 8. <https://doi.org/10.3389/fpls.2017.02204>.
- Sessith A, Pfaffenbichler N, and Mitter B (2019) Microbiome applications from lab to field: Facing complexity. *Trends in Plant Science* 24: 194–198.
- Shulze CN, Chovatia M, Agosto C, et al. (2019) Engineered root bacteria release plant-available phosphate from phytate. *Applied and Environmental Microbiology* 85: e01210–e01219. <https://doi.org/10.1128/AEM.01210-19>.
- Wright D, Zydenbos SM, Wessman P, et al. (2017) Surface coating aids survival of *Serratia entomophila* (Enterobacteriaceae) in granules for surface application. *Biocontrol Science and Technology* 27: 1383–1399.
- Yang W, Zheng L, Liu HX, Wang KB, and Yu Y (2014) Evaluation of the effectiveness of a consortium of three plant-growth promoting rhizobacteria for biocontrol of cotton Verticillium wilt. *Biocontrol Science and Technology* 24: 489–502.
- Zhang S, Lehmann A, Zheng W, You Z, and Rillig MC (2019) Arbuscular mycorrhizal fungi increase grain yields: A meta-analysis. *New Phytologist* 222: 543–555.
- Zilli JE, Pacheco RS, and Gianluppi V (2021) Biological N₂ fixation and yield performance of soybean inoculated with *Bradyrhizobium*. *Nutrient Cycling in Agroecosystems* 119: 323–336.

Further reading

- Ahmad M, Pataczek L, Hilger TH, et al. (2018) Perspectives of microbial inoculation for sustainable development and environmental management. *Frontiers in Microbiology* 9: 2992. <https://doi.org/10.3389/fmicb.2018.0992>.
- Barrow N and Lambers H (2022) Phosphate-solubilising microorganisms mainly increase plant phosphate uptake by effects of pH on root physiology. *Plant and Soil*. <https://doi.org/10.1007/s11104-021-05240-0>.
- Batista BD and Singh BK (2021) Realities and hopes in the application of microbial tools in agriculture. *Microbial Biotechnology* 14: 1258–1268.
- Dejonghe W, Boon N, Seghers D, Top EM, and Verstraete W (2001) Bioaugmentation of soils by increasing microbial richness: Missing links. *Environmental Microbiology* 3: 649–657.
- Kaminsky LM, Trexler RV, Malik RJ, Hockett KL, and Bell TH (2019) The inherent conflicts in developing soil microbial inoculants. *Trends in Biotechnology* 37: 140–151.
- Khan ST (2022) Consortia-based microbial inoculants for sustaining agricultural activities. *Applied Soil Ecology* 176: 104503. <https://doi.org/10.1016/j.apsoil.2022.104503>.
- King WL and Bell TH (2022) Can dispersal be leveraged to improve microbial inoculant success? *Trends in Biotechnology* 40. <https://doi.org/10.1016/j.tibtech.2021.04.008>.
- Mawarda PC, Le Roux X, van Elsas JD, and Salles JF (2020) Deliberate introduction of invisible invaders: A critical appraisal of the impact of microbial inoculants on soil microbial communities. *Soil Biology & Biochemistry* 148: 107874.
- Richardson AE, Barea JM, McNeill AM, and Pringet-Combaret C (2009) Acquisition of phosphorus and nitrogen in the rhizosphere and plant growth promotion by microorganisms. *Plant and Soil* 321: 305–339.
- Rubin RL, van Groenigen KJ, and Hungate BA (2017) Plant growth promoting rhizobacteria are more cost effective under drought: A meta-analysis. *Plant and Soil* 416: 309–323.
- Zhang H, Yuan X, Xiong T, Wang H, and Jiang L (2020) Bioremediation of co-contaminated soil with heavy metals and pesticides: Influence factors, mechanisms and evaluation methods. *Chemical Engineering Journal* 398: 2567.
- Zvinavashe AT, Lim E, Sun H, and Marelli B (2019) A bioinspired approach to engineer seed microenvironment to boost germination and mitigate soil salinity. *PNAS* 116(51): 25555–25561. <https://doi.org/10.1073/pnas.1915902116>.

Biocontrol of soil-borne plant diseases

Mehrdad Mohammadiani and Matthew Bakker, Department of Microbiology, University of Manitoba, Winnipeg, Canada

© 2023 Elsevier Ltd. All rights reserved.

Introduction	439
Soil-borne disease management	440
Chemical control	440
Solarization	440
Sanitation	441
Crop rotation	441
Resistant varieties	441
Biological control	442
Biological control agents currently in use	443
Mechanisms contributing to biological control	444
Microbiome management for biological control	445
Outlook	445
References	446

Abstract

Plant diseases caused by soil-borne pathogens are a major limitation to crop production every year, and new tools are needed for more effective management. In this chapter, we argue for an expanded role for biological control, which involves the use of other living organisms to antagonize plant pathogens. Biological control approaches are complementary to our existing tools for managing soil-borne diseases, which include chemical pesticides, soil solarization, sanitation, crop rotation, and the use of genetically resistant crop varieties.

Key points

- Soil-borne diseases cause major economic losses in crop production.
- Management of soil-borne diseases is challenging because of pathogen survival and the difficulty of treating the whole soil volume.
- Soil harbors both detrimental and beneficial microbes; some of these microbes cause plant disease, while others constrain pathogen success.
- Disease control should not depend on a single strategy, because of off-target effects and the potential for pathogen evolution to circumvent that control.
- Biological control uses living organisms to reduce the impact of pests including soil-borne pathogens.
- Biological approaches are an attractive tool to complement traditional control strategies of soil-borne pathogens.

Introduction

Until the nineteenth century, many scientists believed that various plants are normally sterile and free from microbes within their tissues, as hypothesized by well-known scientists such as Pasteur and Chamberland. In the late 1800's, when Galippe reported that microbes from soil colonized interior parts of vegetable plants (Galippe, 1887), he was highly criticized by other scientists. By the early twentieth century, however, it had been established that many soil microorganisms can be harbored within internal plant tissues, and today it is well known that in both managed and natural ecosystems, a wide range of microbes colonize plant tissues. In fact, the mutualism of terrestrial plants with fungi has been found in fossilized tissues from 400 million years ago. Thus, the presence of microbes along with plants is now considered to be the rule, rather than an exception.

Among these microbes that associate with plants, soil-borne pathogens are a major threat to crop production. Common pathogenic genera include *Fusarium*, *Rhizoctonia*, *Sclerotinia*, *Verticillium*, *Pythium*, and *Phytophthora*. Yield failure resulting from soil-borne diseases such as seedling damping-off, root rot, stunting and wilting have destroyed entire agricultural industries. There is no country or geographic area without soil-borne disease problems, and some soil-borne diseases have been a serious problem for many decades and are responsible for restrictions on agricultural yield.

Management of soil-borne diseases in cropping systems is often highly challenging. In many cases, soil-borne diseases share similarities in symptoms, which makes the accurate diagnosis of a particular disease more difficult and subsequently harder to manage. Once established, soil-borne pathogens often survive for long periods in the soil, even in the absence of host plants. They

can survive either as saprotrophs or as living resting spores, which in the case of certain pathogens have the capacity to survive for at least 15 years waiting for a suitable host. Furthermore, the depth and complex structure of soil precludes the use of many of the tools that can be used to effectively control foliar diseases.

Managing soil-borne diseases presents a tremendous practical challenge, but the persistent nature of this problem also suggests the limitations of our understanding of mycology, bacteriology, nematology, genetics, molecular biology, ecology, and soil microbiology; with a sufficient understanding of these aspects of life in soil, surely effective management of soil-borne diseases would follow!

The objective of this chapter is to summarize the current practices for control of soil-borne diseases, with an emphasis on biological control approaches.

Soil-borne disease management

Integrated disease management draws tools from a diverse toolbox to effectively and economically respond to the particularities of each soil-borne disease scenario. Some of the most frequently used tools for preventing and managing soil-borne disease include chemical control technologies, soil solarization, sanitation, employment of crop rotations that promote the decline of pathogen populations, and the use of genetically resistant crop varieties. To this set of practices, biological control approaches offer unique opportunities and complementary mechanisms of control, and we advocate for expanded research and use of biological control for soil-borne diseases.

Chemical control

At an instinctive level, perhaps the most straightforward approach to managing soil-borne pathogens is to attempt to kill those pathogens prior to introducing a crop. However, in practice this impulse is exceptionally difficult to enact effectively; it is very challenging to expose the full soil volume to substances or conditions that would be lethal to pathogens, due to factors such as the depth and volume of soil, heterogeneity within soil, including discontinuities among patches of soil air and water, as well as chemical sorption phenomena and the potential for rapid biologically-mediated modification of bioactive compounds.

Nevertheless, chemical pesticides have been widely used to control soil-borne pathogens. These can be categorized according to their mode of application and intended zone of protection: chemical fumigants represent an attempt to eradicate a pathogen from the entire soil volume via exposure to toxic gases, while chemicals applied in the furrow or as seed coatings aim for narrower control of pathogens within the germination zone of the seed. Some foliar-applied pesticides may be taken up by the plant and translocated systemically, such that soil-borne pathogens in the rooting zone may be controlled.

One of the issues that accompany repeated use of chemical control technologies is the potential for the development within pathogen populations of resistance toward the action of a given chemical pesticide. In particular, broad use of pesticides that target a single biochemical step within the pathogen's cells can quickly select for resistance. There are thousands of polymorphisms within the genomes of soil-borne pathogens, and in large populations, rare genetic variants will likely exist that impact the efficacy of chemical pesticides. As a result of widespread pesticide use, lineages that are less sensitive to the effects of the pesticide will increase in frequency in response to selection, at the expense of more sensitive members of the population (Hollomon, 2015).

Environmental concerns represent a further downside of chemical pesticides; to the extent that these compounds contribute to ecological disturbance, human health hazards, and damage to aquatic ecosystems, their use in agriculture is problematic. For instance, chemical fumigation has been a major tool for soil disinfestation, yet fumigants are associated with harmful side effects, such as environmental pollution and ozone depletion (Mihajlović et al., 2017). The safe and effective use of soil fumigants requires consideration of site conditions (soil, air, weather), buffer zone establishment, emergency response planning, and most importantly, sealing the soil surface immediately after injection of the fumigant, usually with impermeable film, which adds to the capital and labor expense. In many parts of world, the fumigant methyl bromide was extensively used to control soil-borne diseases before the implementation of the Montreal Protocol in 1992 (Mihajlović et al., 2017). In recent years, progressive restrictions on the use of methyl bromide, because of its contribution to thinning of the ozone layer, have brought into sharp focus the need for alternative strategies for the management of soil-borne diseases.

Solarization

An innovative approach to removing pathogens without the application of toxic chemicals is to use soil solarization. Soil solarization, introduced by Katan (1981), involves management practices such as the deployment of plastic tarpaulins under very sunny conditions to elevate soil temperature by 2 to 15 °C above the ambient temperature of untreated soil. This heating has a profound impact on soil microbial communities, and can cause the death of plant pathogens, especially those that cannot endure temperature above 37–40 °C. Similarly, the seedbank of weed seeds, as well as various insect and nematode pests, can be rendered non-viable through soil solarization.

However, the efficacy of soil solarization declines with soil depth, and many agricultural regions do not have suitable environmental conditions for effective solarization. Furthermore, solarization approaches share with chemical fumigants the issue of leaving the post-treatment soil 'wide open' for rapid spread of new colonists; if pathogens are among the first microbes

to re-establish after treatment, they may experience an environment of low competition, enabling their rapid proliferation and spread. One promising approach may be to couple soil solarization with the immediate application of beneficial microorganisms, such that the establishment of beneficial organisms is promoted in the less competitive environment immediately following solarization, erecting a barrier toward the re-establishment of pathogens.

Sanitation

An approach that aims to prevent the introduction of pathogen propagules to soil, and to remove resources and refugia that promote pathogen survival or reproduction, is complementary to strategies for actively destroying pathogen populations. A key aspect of this approach is to maintain sanitation. Most soil-borne diseases are caused by eukaryotic pathogens that are able to produce resting structures like chlamydospores, microsclerotia, oospores or sclerotia. Furthermore, many of these pathogens also subsist on crop debris as saprotrophs. Therefore, preventative sanitation practices can be seen as attempting to avoid a disease problem before it becomes severe. Examples of sanitation include removing infected plant material, as well as cleaning the tools and equipment that are used in the field, and limiting transfer of soil among fields. Tillage and residue management, such as harvesting, burning, or chopping residue, have been used to accelerate the displacement of pathogens from infected crop debris. These practices may expose the debris to conditions that are less favorable for the pathogen or may accelerate processes of decomposition. Somewhat related are practices that aim to reduce susceptibility to disease by adjusting environmental conditions to be close to the optimum for a given crop species, such as through pH adjustment, provision of adequate nutrition, or moisture management. Here, the goal is preventative; by enhancing plant vigor and limiting pathogen success in the present, fewer epidemics of soil-borne disease are expected in the future.

Despite its benefits, sanitation alone is frequently insufficient to prevent disease outbreaks. In the contemporary context, there are also increasingly competing pressures from additional expectations being brought to agriculture. For instance, society is increasingly coming to expect that our agricultural soils will hold large amounts of carbon (i.e., as one of our strategies for mitigating the rise in atmospheric carbon dioxide concentration). Managing for high soil carbon is frequently at odds with managing for sanitation and removal of pathogen refugia.

Crop rotation

One effective strategy for steering pathogen populations into decline is to rotate the plant cover over an entire field to a very different plant species from 1 year to the next. Indeed, rotating among crops has been used for hundreds of years to promote productivity, even if the specific mechanisms were not understood in detail. Crop rotations have a marked effect on the incidence and severity of soil-borne disease, by avoiding the buildup of inoculum in soil. Under a scenario of repeated monoculture, cultivating the same crop species in a field year after year, pathogen inoculum can become progressively more abundant. There are also mechanisms by which diversifying crop rotations can have beneficial impacts on soil structure and fertility, as well as control of weeds (Reddy, 2017). For effective control of soil-borne pathogens, crop rotation typically aims for a sequence of very dissimilar crops, such as switching between plant families. For example, rotation with a legume is an effective approach for management of fusarium wilt in tomato. The concept is that the more widely divergent the plant species are, the less likely it is that they will share pathogens.

Despite its apparent simplicity and power, crop rotation has limited efficacy in the case of some soil-borne diseases, notably those caused by pathogens with a wide range of host plants, and those caused by pathogens with exceptionally long-lived survival structures like sclerotia, oospores or chlamydospores. For pathogens with a very wide host range, it may be difficult to find non-host plants that can be economically produced. For pathogens with exceptional longevity, the rotation system that would be needed to effectively drive down pathogen population density may be prohibitively long; for instance, in some cases it may be necessary to rotate away from the host crop for 5 years or longer. With such an infrequent return interval, it may be challenging to maintain the equipment necessary for a particular crop, and investment in regional-scale infrastructure, such as processing facilities, may also be disincentivized because of the limitation imposed on the acreage available to a particular crop. Some pathogens, such as *Sclerotinia sclerotiorum*, combine both challenges in that they have a broad host range as well as produce resilient survival structures (Liang and Rollins, 2018), making these pathogens particularly difficult to manage via rotation alone.

In addition to rotating between main crop species from 1 year to the next, there are also practices that increase plant diversity over time by adding plant species from which no harvest is intended. Such plantings are referred to as cover crops, and may be inserted into rotations during periods of the year that are not sufficient for growth of a harvestable crop. Cover crops are intended to provide services and benefits other than harvested grain. Displacement or control of pathogens may be one of the services that are available from cover crops. However, care should be taken, as it is also possible that cover crops could serve as hosts or refugia for pathogens, limiting the efficacy of rotations. For example, the common cover crop winter cereal rye hosts multiple pathogens capable of also infecting a subsequent corn crop (Bakker et al., 2016); thus, farmers have to select crops for rotation very deliberately and empirical testing of performance is critical.

Resistant varieties

Developing crop varieties that have genetic resistance toward pathogens is complementary with all other control strategies, and is unparalleled in terms of its efficacy and cost effectiveness. For instance, crown and root rot of sugar beet (*Beta vulgaris*) caused by

Rhizoctonia solani can be controlled to a large extent by selecting highly resistant cultivars. The beauty of resistant varieties is that they can obviate the need for other inputs or management practices that may add to the cost of production or produce undesirable side effects. Where particular genes have a large impact on resistance or susceptibility, and where genetic markers are available, plant breeders can very rapidly improve the genetic resistance of crops toward certain pathogens. In some cases, genetic modification has been used to engineer resistance toward plant disease. Most famously, a transgenic approach saved the papaya industry in Hawaii from being rendered untenable by a ringspot virus. With respect to soil-borne diseases, traits related to root exudation may be particularly relevant to interactions with pathogens. In one case, root exudates of a resistant tobacco cultivar showed inhibitory activity against the pathogen *Phytophthora nicotinae*, whereas the exudates of a susceptible cultivar stimulated growth of the pathogen (Zhang et al., 2020). Much more research is needed to understand in detail how root exudation could be modified during cultivar development and the extent to which related traits may be selected upon to improve resistance toward soil-borne pathogens.

There are several clever ways to multiply the benefits of genetic resistance via management decisions. One of these approaches involves growing cultivar mixtures, in which a resistant cultivar is paired with one or more other varieties, which may lack resistance or may possess a different resistance profile. For example, compared to growth in a pure stand, a wheat cultivar that is susceptible to soil-borne wheat mosaic virus (which is vectored by nematodes in soil) demonstrated lower incidence of infection when grown in a mixture with a resistant cultivar (Hariri et al., 2001). In these cultivar mixtures, the development and spread of disease can be slowed through epidemiological mechanisms (Mundt, 2002). However, the efficacy of this approach is dependent on the life cycle of the pathogen, host specificity, and spatial distributions of both the pathogen in the soil and of the crop cultivars.

Grafting is a second technique that maximizes the utility of genetic resistance. This technique is popular in fruit trees as well as in high-value vegetable crops such as peppers and tomatoes, where it is used increasingly to manage the Verticillium wilt (Mackey et al., 2018). Some bacterial wilt and root knot nematodes of vegetables are also managed by grafting techniques. The advantage of this technique is that it permits separate consideration of desirable root traits vs. yield or quality traits of the portion of the plant that is actually harvested. In particular, the rootstock can be selected for effective resistance to soil-borne disease, without the need to consider potential impacts on traits of the aboveground portion of the plant. Although grafting can be done in a moderately high throughput fashion, this technique is limited by high costs and the need to transplant grafted seedlings; it is thus not applicable to most commodity crops.

Despite genetic resistance being a fantastic tool for management of soil-borne diseases, there are also some limitations and drawbacks to this tool. In similar fashion to the use of chemical pesticides, the largescale deployment of genetic resistance can exert substantial directional selection on pathogen populations, driving evolution of pathogen strains that are able to overcome the mechanism of resistance. Thus, vigilance is required to recognize new pathotypes and to improve genetic resistance. If resistance genes are deployed in a careless or uncontrolled manner, they may quickly fail; this represents a tremendous waste of what may be an irreplaceable resource. For some soil-borne pathogens, genetic resistance depends not on a few genes of large effect, but on many genes that each have only small effects; in these cases, progress in improving resistance is very slow and genetic resistance may not be available or adequate. Furthermore, there are some resistant cultivars that may be unacceptable due to other horticultural characteristics or insufficient quality traits.

Biological control

Complementary to all previously discussed methods for controlling soil-borne disease, is biological control: the use of another living organism for the control of a pathogen. Soil sanitation and solarization, crop rotation, breeding for new resistant varieties and grafting, and chemical pesticides are insufficient to control soil-borne disease in many agricultural systems. An increasing number of recent studies demonstrate that many fungi and bacteria can serve as antagonists of pathogens. These microbes associate with roots and can protect their host plants against soil-borne diseases. There is now substantial evidence that some beneficial microbes, particularly endophytic fungi and bacteria, are recruited by plants and these microbes can act as a second, microbiological layer for plant defense against biotic stresses (Dini-Andreote, 2020).

The biocontrol of soil-borne diseases via the application of living microorganisms has been researched for more than 80 years. In 1963, butt rot of pines, a disease caused by the fungus *Heterobasidion annosum*, was controlled by inoculation with a non-pathogenic fungus, *Phlebiopsis gigantea*. Successful biocontrol agents have been found in diverse microbial taxa, including fungi, bacteria, viruses, and protozoa. Some of these microbes have suppressive effects toward selected soil-borne diseases and also contribute to the success of plant defense strategies under greenhouse and field conditions.

Frequently, the goal with biological control, as for several of the control strategies discussed earlier, is not to eradicate a pathogen from the soil, but more modestly, to provide a barrier between susceptible plant tissues and the attacking pathogen. As a result, biocontrol agents that are able to colonize plant tissues either epiphytically or endophytically have been viewed as particularly attractive. Endophytic colonization refers to microbial growth within internal plant tissues that causes no apparent adverse symptoms. For instance, endophytic colonization of wheat by *Sarocladium zeae* successfully slowed the spread of the pathogen *Fusarium graminearum* through wheat tissue (Kemp et al., 2020). In addition, microbial associates of plants have many other functional roles, such as being plant growth promoters, protectors against plant herbivores, and aiding in nutrient acquisition. Thus, using biological control agents, especially those that colonize internal tissues of their host plant, is increasingly capturing the imagination of many microbiologists and plant pathologists.

Biological control agents currently in use

Many of the effective biological control technologies that have been developed are formulations of live fungi. Fungi play a functional role in habitat adaptation and enhanced plant resistance to both biotic and abiotic stresses because they are an important element of the microbiota of agricultural and forest ecosystems. In terms of taxonomic classification, a wide range of fungi inhibit the growth of soil-borne plant pathogens; more than 30 genera and 80 species of soil fungi (such as *Botrytis cinerea*, *Penicillium chrysogenum*, *P. citrinum*, *Gliocladium virens*, *Trichoderma glaucum*, *T. harzianum*, *T. koningii* and *T. viride*) have demonstrated antagonistic activity against soil-borne diseases in addition to their well-known plant growth promotion and enhancement (Tariq et al., 2020). A lengthier summary of fungal taxa that have been used in biocontrol is provided in Table 1. Non-pathogenic strains of *Fusarium* can live in soil, in the rhizosphere and in planta and show antagonistic activities against fungal plant pathogens. Other isolates such as some belonging to *Aspergillus*, *Saccharomyces*, and *Gliocladium* have shown great promise to inhibit and suppress the activity of pathogens such as *Alternaria*, *Aspergillus*, *Fusarium*, *Rhizoctonia*, *Pythium*, and *Phytophthora*. In addition to these genera, *Trichoderma* spp. have received a lot of research attention and have been developed into effective microbial biocontrol technologies that significantly reduce the incidence and severity of soil-borne diseases.

Under certain circumstances within an integrated pest management approach, some fungi that feed on nematodes can be used for biocontrol. Nematophagous fungi, like *Arthrobotrys*, *Dactylella*, *Dactylaria*, are used to trap root-knot nematodes, and a talc-based formulation of *Paecilomyces lilacinus* has been used for the control of plant-parasitic nematodes, as this fungus can parasitize nematode eggs (Pendse et al., 2013). Arbuscular mycorrhizal fungi have also been reported to reduce severity of many soil-borne diseases, including root-knot nematodes, although in some of these interactions, the antagonistic mechanisms and relevant functions are still unknown.

Microbes that colonize internal plant tissues (endophytes) are promising biocontrol agents because of the potential of this growth habit for close interactions with the plant, the increased likelihood that the biocontrol agent is present at the site of pathogen attack, and because a less competitive environment inside the plant (relative to the open soil) may facilitate longer persistence of the biocontrol agent. Fungal endophytes such as *Heteroconium chaetospora* and *Acremonium alternatum* have potential to be effective biocontrol agents against clubroot, caused by *Plasmodiophora brassicae* (Khalid et al., 2022). Frequently, close relatives of pathogens are found to be effective at colonizing plants endophytically. This adds some difficulty to the search for beneficial endophytes, as the close taxonomic affiliation to pathogens can raise regulatory hurdles or otherwise give pause. For instance, nonpathogenic *Fusarium* species can colonize internal plant tissues endophytically (Al-Ani, 2019), but *Fusarium* is best known as a genus containing many damaging plant pathogens.

It is particularly attractive to consider that some biocontrol agents may be able to bring multiple services to the crop production environment. Notably, various genera of entomopathogenic fungi (that is, fungi which parasitize or feed on insects) have been also shown to act as endophytes in plant hosts (Quesada Moraga, 2020) and demonstrate antagonistic activity against soil-borne pathogens. This dual action suggests that one biocontrol technology may be useful in providing protection against soil-borne diseases, and also against insect pests. Several strains with relatively broad host ranges have been developed into crop protection technologies, the majority based on filamentous Ascomycota within the genera *Beauveria* and *Metarhizium* (Clavicipitaceae) (Rasool et al., 2020). Many of these products have demonstrated antagonistic activities against soil-borne pathogens. For instance, *B. bassiana* strain ATCC 74040, applied on squash, has been reported to significantly suppress the incidence and severity of the zucchini yellow mosaic virus. Both *Beauveria* spp. and *Metarhizium* spp. play an important role in habitat adaptation of many species of host plants, resulting in improved plant performance.

Table 1 Fungal taxa reported as biocontrol agents against fungi and oomycetes causing soil-borne plant diseases.

Biocontrol agent	Test plant	Target pathogen
<i>Aspergillus fumigatus</i>	cocoa	<i>Phytophthora palmivora</i>
<i>Beauveria bassiana</i>	broad bean	<i>Botrytis cinerea</i>
<i>Beauveria bassiana</i>	grapevine	<i>Plasmopara viticola</i>
<i>Beauveria bassiana</i>	tomato	<i>Pythium myriotylum</i>
<i>Cladosporium</i> spp.	notoginseng	<i>Fusarium solani</i>
<i>Fusarium oxysporum</i>	soybean	<i>Sclerotinia sclerotiorum</i>
<i>Metarhizium anisopliae</i>	maize	<i>Fusarium graminearum</i>
<i>Trichoderma afroharzianum</i>	cucumber	<i>Sclerotinia sclerotiorum</i>
<i>Trichoderma asperellum</i>	thale cress	<i>Verticillium dahliae</i>
<i>Trichoderma atroviride</i>	beans	<i>Botrytis cinerea</i>
<i>Trichoderma gamsii</i>	rice	<i>Phoma herbarum</i>
<i>Trichoderma hamatum</i>	thale cress	<i>Phytophthora nicotianae</i>
<i>Trichoderma hamatum</i>	cabbage	<i>Sclerotinia sclerotiorum</i>
<i>Trichoderma harzianum</i>	rice	<i>Bipolaris oryzae</i>
<i>Trichoderma harzianum</i>	soybean	<i>Sclerotinia sclerotiorum</i>
<i>Trichoderma harzianum</i>	melon	<i>Fusarium oxysporum</i>
<i>Trichoderma virens</i>	tobacco	<i>Rhizoctonia solani</i>
<i>Trichoderma</i> spp.	avocado	<i>Rosellinia necatrix</i>
<i>Trichoderma</i> spp.	peanut	<i>Fusarium solani</i>

A major constraint to the development of biocontrol technologies lies in the requirements for stable formulation and long shelf-life under minimally controlled environmental conditions. These factors are often not given consideration in early research, which may explain why there is a much longer list of taxa that have shown early signs of being effective biocontrol agents, but a much narrower list of taxa that have been successfully developed into commercial products. For instance, bacterial biocontrol products are dominated by species of *Bacillus*, which are able to produce exceptionally robust spores.

Mechanisms contributing to biological control

Although a wide range of taxa have been shown to be capable of providing biological control over soil-borne plant pathogens, there are only a handful of distinct mechanisms underlying the provision of this service. These mechanisms can be fit into categories of resource competition, interference competition, and impacts on the susceptibility of the host plant.

Resource competition in biological control includes scenarios in which the beneficial strain occupies a niche that overlaps with that of the pathogen, such that the two organisms attempt to exploit the same resources, or to colonize the same physical spaces. Presumably, this competition will occur primarily prior to the pathogen's attack on the plant, although in the case of endophytic biocontrol strains, resource competition may take place in planta. The rhizosphere, the zone of soil immediately surrounding plant roots and under their heavy influence, is a site of intense competition. Because of the importance of this zone within soil, biocontrol strains are commonly evaluated for their rhizosphere competence, or the ability to colonize the rhizosphere at high densities and to persist over time in this environment. As an example of this category of biological control, the presence of endophytic fungi in the root of potatoes reduces incidence of *Rhizoctonia solani*, likely as a result of competition for ecological niche.

Interference competition refers to direct interactions between a biocontrol agent and the pathogen. For example, fungi such as *Acremonium alternatum*, *Acrodonitum crateriforme* and *Cladosporium oxysporum* can physically interact with and destroy the mycelia of plant pathogens. Antibiosis, or the release of compounds which are toxic to other microorganisms, is a common trait among biocontrol strains (Tariq et al., 2020). These toxins are typically secondary metabolites, a term which indicates that they are not essential to the basic cellular functioning of the organism. As 'accessory' functions, the ability to produce such metabolites can vary widely among strains and is not typically a conserved or reliable trait in a taxonomic sense. This functional variability motivates efforts to screen very large numbers of potential biocontrol strains to identify those rare strains that may either produce novel metabolites or produce prodigious quantities of some known metabolite of interest. Numerous studies have indicated that several established fungi in plants produce antibiotic and toxic compounds such as monoterpenes and sesquiterpenes to compete with soil-borne pathogenic organisms. Thus, specific compounds produced by *Trichoderma harzianum*, like harzianolide, 8-dihydroxy-3-methyl-anthraquinone, and azaphilones, inhibit the growth of *Phytophthora cinnamomi* as well as suppress the incidence of diseases caused by *Botrytis cinerea*, *Rhizoctonia solani* (Vinale et al., 2009), and *Pythium ultimum* (Tariq et al., 2020). The production of the antibiotic beauvericin by the fungus *B. bassiana* has a significant bioactivity and can inhibit the development of damping-off caused by *R. solani* and *P. myriotylum* in tomato and cotton (Jaber and Ownley, 2018). A cautionary note seems warranted: interference competition is a concept borrowed from traditional ecology, but whether microbial interactions are direct or 'intentionally' oriented toward a particular competitor is much more ambiguous. For instance, control of plant pathogens via antibiotic compounds produced by *Streptomyces* bacteria may be largely a side effect of, or damage collateral to interference competition among saprotrophic *Streptomyces* strains (Kinkel et al., 2012).

Finally, suppression of soil-borne disease development may arise through functions of a biocontrol strain that serve to enhance plant health or vigor. Microbial processes that are involved in plant growth enhancement can include production of phytohormones, altering the growth pattern of the plant, or increased provision of scarce nutrients like iron (e.g., via production of microbial siderophores, which are iron-chelating compounds that can be taken up from the soil solution together with their cargo of iron) or nitrogen (via fixation from atmospheric N₂) (Vessey, 2003). In each case, the effect relies on more vigorous plants being able to mount a more effective defense against pathogen attack. Plant growth promoting interactions are very complex to untangle as there are typically numerous confounded effects and feedbacks are likely. For instance, if interactions with a microbial partner alter root traits such as root hair density, root hair length, branching pattern, and so on, then there will be knock-on effects in terms of water and nutrient availability that are only indirectly associated with the action of the microbial partner.

Root colonization of soybean plants with *M. anisopliae* has been shown to produce greater amounts of jasmonic acid, increasing shoot length in comparison with fungal-free control. Some endophytic fungi, like *Aspergillus fumigatus* and *Fusarium proliferatum*, can produce phytohormones such as gibberellins, and thus, may alleviate some harmful effects of biotic stresses. The gibberellins and tetracyclic diterpenoid acid compounds, the most valuable group of hormones, are well known to contribute to plant development and regulate cell division. They benefit plants in different ways such as growth promotion, improvement in crop productivity and vegetative fitness, stem and leaf growth, floral initiation, and flower and fruit growth (Bilal et al., 2018). Furthermore, plant growth promotion by beneficial fungi also has been reported due to production of indole-3-acetic acid (IAA), which promotes plant root growth, root hair development, and is indispensable for plant growth and development.

Exposure to other microbes prior to pathogen attack, including beneficial microbes that may be deliberately applied, also has an impact on the speed and magnitude with which plants are able to mount their innate defense responses. This effect has been termed 'defense priming' (Conrath et al., 2015). There is, however, some uncertainty about the extent to which this effect may be an artifact of simplified experimental conditions during research, as most plants grown in a field setting are continually exposed to a wide range of microorganisms. Related concepts, which reflect the capacity of plants to respond with enhanced resistance within tissues that are spatially distant from the stimulus, have been termed systemic acquired resistance and induced systemic resistance (Vlot et al., 2021). In a regulatory context, products developed around microorganisms that interact primarily with the plant may be administered according to different rules (i.e., for 'plant growth promotion' products, rather than for 'disease control' products).

Microbiome management for biological control

Efforts to isolate and identify potential biocontrol strains have sometimes been successful by prospecting in areas with pronounced disease development, but an alternate strategy seeks to identify biocontrol strains in areas where disease fails to develop as expected (that is, in disease-suppressive soils). In the disease-suppressive soils that have been most carefully studied, pathogen control has sometimes been attributed to particular members of the soil microbial community, or even to particular secondary metabolites produced by certain members of the soil microbial community (Mendes et al., 2011).

The identification and study of disease suppressive soils that have arisen in various places and cropping systems has also given rise to a new aim for biocontrol: management of indigenous microbial communities to foster traits or community members that suppress pathogens, rather than inoculation with an exogenous strain to introduce this function. Management practices that could conceivably be used in service of microbiome management aims include crop rotation, crop genotype selection, amendments of various types, and perhaps even targeted disruption events.

Plant root-associated microbiomes, such as rhizosphere and the endospheric microbial community of roots, are considered as a critical factor for understanding the interactions between plants and soil. Root-associated microbes are recruited from the soil, and many provide benefits for the host plant, such as facilitating nutrient availability. Endophytic microbes have an interspecific interaction with their host plants for a portion of or their entire lifecycle and cause no apparent adverse symptoms. Numerous studies have shown microorganisms such as bacteria and fungi have been successfully adopting a lifestyle as endophytes by colonizing plant tissues in many crop plants such as potato, broad bean, maize, cotton, wheat, tomato and oil seed rape. There have been demonstrations of colonization by endophytes leading to significant alleviation of abiotic stress, as for example due to salinity (Gupta et al., 2021).

Outlook

There is an on-going need for improved technologies to control soil-borne plant diseases, as made clear by persistent losses to soil borne-disease, together with off-target effects of conventional management practices and the rise of goals for agriculture that go beyond food production. Furthermore, relying on a single strategy to combat the threats posed by soil-borne diseases is not an effective solution over the long term. On the other hand, non-chemical options such as soil sanitization and solarization, crop rotation, using resistant varieties, and microbial biological control agents may not satisfy growers' expectations. However, applying all these tools together as an integrated system can hold the population of soil-borne pathogens below the economic threshold level and in an environmentally friendly way without imposing adverse impacts on the fauna and flora of the agricultural system (Fig. 1).

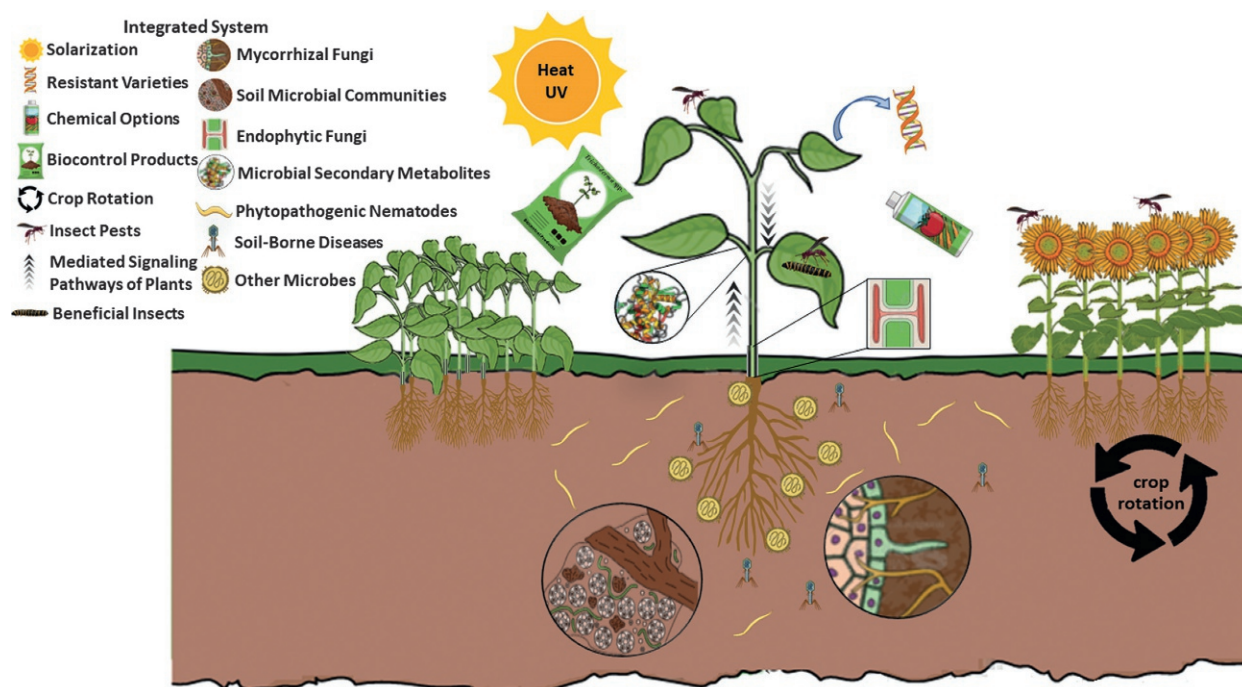


Fig. 1 Overview of tools that can be used in the integrated management of soilborne diseases, including biological control approaches. Pathogens should be understood as members of the soil microbiome, able to be influenced through management of diverse aspects of the abiotic environment and of the community of interaction partners.

References

- Al-Ani LKT (2019) Secondary metabolites of non-pathogenic *Fusarium*: Scope in agriculture. In: Singh H, Keswani C, Reddy M, Sansinenea E, and García-Estrada C (eds.) *Secondary Metabolites of Plant Growth Promoting Rhizomicroorganisms*, pp. 59–76. Singapore: Springer.
- Bakker MG, Acharya J, Moorman TB, Robertson AE, and Kaspar TC (2016) The potential for cereal rye cover crops to host corn seedling pathogens. *Phytopathology* 106: 591–601.
- Bilal L, Asaf S, Hamayun M, et al. (2018) Plant growth promoting endophytic fungi *Aspergillus fumigatus* TS1 and *Fusarium proliferatum* BRL1 produce gibberellins and regulate plant endogenous hormones. *Symbiosis* 76: 117–127.
- Conrath U, Beckers GJM, Langenbach CJG, and Jaskiewicz MR (2015) Priming for enhanced defense. *Annual Review of Phytopathology* 53: 97–119.
- Dini-Andreote F (2020) Endophytes: the second layer of plant defense. *Trends in Plant Science* 25: 319–322.
- Galippe V (1887) *Note sur la présence de micro-organismes dans les tissus végétaux. Comptes Rendus Hebdomadaires de la Société de Biologie*. 410–416. Paris.
- Gupta S, Schillaci M, Walker R, et al. (2021) Alleviation of salinity stress in plants by endophytic plant-fungal symbiosis: current knowledge, perspectives and future directions. *Plant and Soil* 461: 219–244.
- Hariri D, Fouchard M, and Prud'homme, H. (2001) Incidence of soil-borne wheat mosaic virus in mixtures of susceptible and resistant wheat cultivars. *European Journal of Plant Pathology* 107: 625–631.
- Hollomon DW (2015) Fungicide resistance: facing the challenge—a review. *Plant Protection Science* 51: 170–176.
- Jaber LR and Ownley BH (2018) Can we use entomopathogenic fungi as endophytes for dual biological control of insect pests and plant pathogens? *Biological Control* 116: 36–45.
- Katan J (1981) Solar heating (solarization) control of soilborne pests. *Annual Review of Phytopathology* 19: 211–236.
- Kemp ND, Vaughan MM, McCormick SP, Brown JA, and Bakker MG (2020) *Sarocladium zeae* is a systemic endophyte of wheat and an effective biocontrol agent against fusarium head blight. *Biological Control* 149: 104329.
- Khalid M, Kayani SI, Khan AA, Gul H, and Hui N (2022) *Plasmodiophora brassicae*—the causal agent of clubroot and its biological control/suppression with fungi—a review. *South African Journal of Botany* 147: 325–331.
- Kinkel LL, Schlatter DC, Bakker MG, and Arenz BE (2012) *Streptomyces* competition and co-evolution in relation to plant disease suppression. *Research in Microbiology* 163: 490–499.
- Liang X and Rollins JA (2018) Mechanisms of broad host range necrotrophic pathogenesis in *Sclerotinia sclerotiorum*. *Phytopathology* 108: 1128–1140.
- Mackey M, Kurosky A, Robb EJ, and Nazar RN (2018) A graft mimic strategy for *Verticillium* resistance in tomato. *Molecular Biotechnology* 60: 665–669.
- Mendes R, Krijt J, de Bruijn I, et al. (2011) Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* 332: 1097–1100.
- Mihajlović M, Rekanović E, Hrustić J, Grahovac M, and Tanović B (2017) Methods for management of soil-borne plant pathogens. *Pesticidi i fitomedicina* 32: 9–24.
- Mundt CC (2002) Use of multiline cultivars and cultivar mixtures for disease management. *Annual Review of Phytopathology* 40: 381–410.
- Pendse MA, Karwande PP, and Limaye MN (2013) Past, present and future of nematophagous fungi as bioagents to control plant parasitic nematodes. *Journal of Plant Protection Sciences* 5: 1–9.
- Quesada Moraga E (2020) Entomopathogenic fungi as endophytes: their broader contribution to IPM and crop production. *Biocontrol Science and Technology* 30: 864–877.
- Rasool S, Vidkjær NH, Hooshmand K, et al. (2020) Seed inoculations with entomopathogenic fungi affect aphid populations coinciding with modulation of plant secondary metabolite profiles across plant families. *New Phytologist* 229: 1715–1727.
- Reddy PP (2017) Crop rotation. In: Reddy PP (ed.) *Agro-Ecological Approaches to Pest Management for Sustainable Agriculture*, pp. 229–242. Singapore: Springer.
- Tariq M, Khan A, Asif M, et al. (2020) Biological control: a sustainable and practical approach for plant disease management. *Acta Agriculturae Scandinavica, Section B—Soil & Plant Science* 70: 507–524.
- Vessey JK (2003) Plant growth promoting rhizobacteria as biofertilizers. *Plant and Soil* 255: 571–586.
- Vinale F, Flematti G, Sivasithamparan K, et al. (2009) Harzianic acid, an antifungal and plant growth promoting metabolite from *Trichoderma harzianum*. *Journal of Natural Products* 72: 2032–2035.
- Vlot AC, Sales JH, Lenk M, et al. (2021) Systemic propagation of immunity in plants. *New Phytologist* 229: 1234–1250.
- Zhang CS, Zheng Y, Peng L, and Cao J (2020) Rootstock-scion interaction affects the composition and pathogen inhibitory activity of tobacco (*Nicotiana tabacum* L.) root exudates. *Plants* 9: 1652.

Management of soil biology for crop protection from biotic and abiotic stresses

Michael J Goss^a, Isabel Brito^b, and Mário Carvalho^b, ^aSchool of Environmental Sciences, University of Guelph, Guelph, ON, Canada; ^bMED – Mediterranean Institute for Agriculture, Environment and Development & CHANGE – Global Change and Sustainability Institute, University of Évora, School of Science and Technology, Évora, Portugal

© 2023 Elsevier Ltd. All rights reserved.

Introduction	448
Major biological components of the soil that positively influence plant growth	448
A definition of management in the context of soil microbes	449
Potential approaches to soil microbe management	449
Inoculation	449
Inoculation of rhizobia	449
Inoculation of plant growth promoting rhizosphere bacteria and archaea	452
Inoculation with mycorrhizal fungi	453
Reliability of inoculation	454
Exploitation of indigenous microbes	454
Pre-requisites for ensuring appropriate levels of biological and functional diversity in soil microbiomes	454
Farming systems and agronomic management practices	454
Tillage	454
Nutrient management	456
Cropping system	463
Chemical crop protection	463
Disease suppressive soils	464
Approaches to manage soil microbes in farming systems	464
Examples of managing AMF	464
Prerequisites for managing AMF	464
Abiotic stress	465
Biotic stress	466
Complex interactions – the case of Mn toxicity in legumes	466
Conclusions	467
References	467

Abstract

A vast number and types of microorganisms occur in soil. They perform a wide array of processes and fulfil numerous functions, ranging from biogeochemical cycling to plant protection. The complexity of habitats and ecosystems in which microorganisms operate and the level of current understanding greatly limit our ability to manage them. One possible approach is to apply inocula of 'laboratory-grown' organisms that can complete missing or inadequate processes; however, its universal effectiveness poses considerable challenges. Alternatively, agronomic practices are selected, based on tillage and cropping systems, that are conducive to enhanced functioning of indigenous microbes against biotic and abiotic stresses.

Glossary

1-aminocyclopropane-1-carboxylic acid (ACC) 1-Aminocyclopropane-1-carboxylic acid (ACC) deaminase is an enzyme that degrades ACC, the precursor of ethylene, a gaseous plant hormone present in all higher plants. Numerous bacteria, which can enhance plant growth, produce ACC-deaminase.

Operational taxonomic unit (OTU) Operational taxonomic unit (OTU), is the basic taxonomic unit identified by numerical genome analysis following high-throughput sequencing, and are so called because the units prescribed are not always comparable to those established by formal taxonomy. The information used is restricted to the nucleotide sequence of specific regions of DNA or RNA prepared by polymerase chain reaction (PCR), which contain copying errors at varying frequencies that depend on the techniques employed. A commonly accepted value for the variability within an OTU is 3%, resulting in its members sharing 97% similarity in base sequences. Numerical analysis tools are used to distinguish biological variations in sequences from those resulting from experimental copying errors. The OTUs that have gone through this 'denoising' process are referred to as zero radius OTUs or zOTU in short.

Key points

- Soil contains a large number of microbes, with bacteria ($\sim 10^8 \text{ g}^{-1}$), algae ($\sim 10^6 \text{ g}^{-1}$) archaea (10^5 g^{-1}), fungi (10^5 g^{-1}), protozoa ($\sim 10^4 \text{ g}^{-1}$) and small-size animals such as nematodes ($\sim 10^4 \text{ g}^{-1}$) contributing the majority of individuals.
- These organisms contribute to a wide array of processes in soil that affect the growth and development of the plants and animals that inhabit each locality, as well as contributing to the major geochemical cycles.
- Given the numbers, types and activities of these microbes, and the ecosystems involved, coupled with the confines of current knowledge, our ability to manage them is very restricted.
- One approach is to apply an inoculum of organisms grown under controlled conditions to the soil, but such microbes may not be able to compete with the indigenous population or edaphic conditions might be sub-optimal. The alternative is to use agronomic practices that are conducive to the growth of indigenous microbes.
- A strategy for the deliberate manipulation of the indigenous arbuscular mycorrhizal fungi (AMF) population and its functional diversity. It requires the intentional use of selected host plants to form an extensive extraradical mycelium (ERM) involving a consortium of indigenous AMF.

Introduction

Over the last 30 years, the consequences for global climate change of the rapidly increasing greenhouse gas composition of the atmosphere coupled with the demand for food, fuel and fiber from an inexorably growing human population have highlighted key issues in agricultural production. These issues can be summarized as availability of land, water and essential inputs to ensure sustainable utilization of the other two resources. The area of land available for arable crops is declining because of urban expansion and soil degradation. Changes in the distribution of precipitation have exposed more crops to drought. Inadequate through drainage can result in the build-up of fertilizer residues and allow saline groundwater to enter the root zone. Constraints have been made to inputs of fertilizer and other agrochemicals to protect the wider environment and reduce greenhouse gas emissions, but this can also provide opportunities for altered ranges for plant pests and diseases.

The biological component of soil is critical for the breakdown of plant residues and their incorporation into soil organic matter (SOM). Soil fauna, including earthworms, collembola and mites, shred animal and plant remains to make them more accessible to chemical degradation by microbes, that operate the ultimate steps of mineralization. Microorganisms also have great potential to protect agro-ecosystems from stressors because of their ability (i) to improve soil structure by releasing mucilage that can stabilize soil particles into aggregates; (ii) to increase the availability of plant nutrients; (iii) to form different kinds of symbioses that enhance the growth of plants and protect them against different stresses. The current challenge is therefore, to establish and implement the criteria by which the beneficial functions of soil biology can be supported or even enhanced through agronomic practices that ensure sustainable food security. It also means maintaining or increasing microbial biodiversity to optimize the potential to counter biotic stressors, including pathogens and herbivores, and abiotic stresses, such as deficiencies or toxicity of mineral nutrients. In part that requires maintaining functional activity of microbes in the rhizosphere of plants, including endosymbionts. Biofilms appear to be an important mechanism by which microbes maintain their activity in soil, so limiting actions that can disrupt them is another prerequisite.

Major biological components of the soil that positively influence plant growth

The development of new generation molecular methods has greatly changed our knowledge of the organisms present in the soil. The genetic and functional diversity, positions in the 'tree of life,' inter-relationships, interactions and contributions to ecosystem functioning are, in some aspects, becoming clearer, but it also has extended the perception of what is yet unknown. Numbers of organisms are vast (Table 1), for example the total number of bacteria and archaea in the soil worldwide is estimated to be $\sim 3 \times 10^{29}$ cells; many can be detrimental to plants. The root systems of plants create a localized zone of influence by their extension, radial expansion, uptake of water and selective uptake of nutrients, release of exudates, and loss of cells. Organic material entering the soil as plant litter, harvest residues and rhizodeposition is decomposed by soil organisms, which use it as an energy source but can also add additional material, particularly their necromass. The rhizobiome is also important to the biogeochemical cycling of carbon, particularly in the turnover of the component present in the soil. Roots and the microbes of their rhizobiome determine the structure and stability of the rhizosphere at the same time as influencing the carbon cycle and the interactions among them. Within the species of the archaea, bacteria and fungi found in the rhizobiome, strains have been isolated that can greatly enhance the availability of mineral nutrients, or stimulate the extension of roots, act as biocontrol agents against pathogens and support plant dry matter production (Table 2). The host plant genotype is important in the formation of this rhizobiome (Fig. 1).

Table 1 The concentration of main groups of microbes, including nematodes in soil.

Organisms		Number	Units
Archaea		$0.2\text{--}0.4 \times 10^5$	amoA gene copies ^a g ⁻¹ × 10 ³
Bacteria^b		$4.4\text{--}29.1 \times 10^5$	cfu g ⁻¹ × 10 ³
Fungi		$5\text{--}8 \times 10^2$	cfu g ⁻¹ × 10 ³
Algae^c		$0.1\text{--}1.3 \times 10^4$	cfu g ⁻¹ × 10 ³
Total protozoa	Pasture	29.20	g ⁻¹ × 10 ³
	Grass	1.13	g ⁻¹ × 10 ³
	Arable soil	5.55	g ⁻¹ × 10 ³
Flagellates	Annual grass	0.50	g ⁻¹ × 10 ³
	Arable soil	0.57–13.95	g ⁻¹ × 10 ³
Ciliates	Annual grass	0.01	g ⁻¹ × 10 ³
	Arable soil	0.03–0.28	g ⁻¹ × 10 ³
Amoebae	Cropped soil	2.33–51.98	g ⁻¹ × 10 ³
Testacae	Pasture	0.6	g ⁻¹ × 10 ³
Nematodes			
Grassland			
Total		15.18	g ⁻¹ × 10 ³
Plant feeder		6.32	g ⁻¹ × 10 ³
Fungal feeder		2.56	g ⁻¹ × 10 ³
Bacterial feeders		3.88	g ⁻¹ × 10 ³
Omnivores		1.10	g ⁻¹ × 10 ³
Barley			
Total		9.92	g ⁻¹ × 10 ³
Plant feeder		2.34	g ⁻¹ × 10 ³
Fungal feeder		2.56	g ⁻¹ × 10 ³
Bacterial feeders		4.80	g ⁻¹ × 10 ³
Omnivores		0.22	g ⁻¹ × 10 ³

^aAssumes there are between 1 and 3 amoA (ammonia monooxygenase) gene copies per archaeal cell.

^bAssumes that less than 0.1 to 10% of total bacteria are culturable.

^cCounts of autotrophic microbes in the soil's crust (1 cm depth samples).

Given the wide range of soil organisms and the significance of the functions they carry out, there are obvious challenges to any attempt at directing their collective activities.

A definition of management in the context of soil microbes

Overall, soil and plant management strategies must maintain genetic and functional diversity by preventing the main causes of land degradation (Table 3). With our current knowledge, management of the microbial component of a soil has to depend on modifying the factors that determine the particular microorganisms present and those that influence their activities. Importantly, there is clearly a need to ensure that changes in the population of microbes do not adversely impact the range of functions that they perform. Equally, requirement for functional diversity must allow for sufficient genetic diversity. Activities of microbes are very dependent on temperature, the supply and quality of organic matter, mineral nutrients and the availability of water and oxygen.

Potential approaches to soil microbe management

Approaches to changing the composition of the microbe population involve addition of a specific inoculum or modifying conditions to stimulate components of the soil microbiome.

Inoculation

Techniques and opportunities for inoculation are discussed by Lopes et al. (2021) and O'Callaghan et al. (2022).

Inoculation of rhizobia

Rhizobia are commonly applied in agriculture to give improved nitrogen fixation in root nodules of pulse crops as well as other grain, fodder and pasture legumes (Fabaceae) relative to indigenous microbes. The competitiveness of an inoculant has to exceed

Table 2 Rhizosphere microorganism groups that have strains reported to promote growth and some processes by which it is achieved.

Domain	Phylum (Class)	Order (Family)	Example species	Nutrient availability				Effect on Plant or other microbes		
				N	P	K	Micro-nutrients	Plant Growth	Biocontrol	Biocide
Archaea	Halobacteriota (Halobacteria)	Halobacteriales (Halococcaceae)	<i>Halococcus hamelinensis</i>		√					
Bacteria	Actinobacteriota (Actinomycetia)	Actinomycetales (Frankiaceae)	Frankia casuarinae	√						
		Actinomycetales (Microbacteriaceae)	<i>Microbacterium arborescens</i>	√	√			√		√
		Mycobacteriales (Micromonosporaceae)	<i>Micromonospora lupini</i>				√	√		
		Streptomycetales (Streptomycetaceae)	<i>Streptomyces</i> spp.		√		√	√		
	Bacteroidota; (Bacteroidia)	Flavobacteriales (Flavobacteriaceae)	<i>Flavobacterium anhuiense</i>					√		
	Cyanobacteria	Cyanobacteriales (Nostocaceae)	Nostoc sp.	√				√		
	Firmicutes (Bacilli)	Bacillales (Bacillaceae)	<i>Bacillus amyloliquefaciens</i> ^a	√	√		√	√	√	√
			<i>Bacillus thuringiensis</i>	√	√		√	√	√	√
		Staphylococcales (Staphylococcaceae)	<i>Staphylococcus succinus</i>		√			√		
	Firmicutes A (Clostridia)	Clostridiales (Clostridiaceae)	<i>Clostridium pasteurianum</i>	√	√			√		
	Proteobacteria (Alphaproteobacteria)	Azospirillales (Azospirillaceae)	<i>Azospirillum amazonense</i>	√	√			√		
		Rhizobiales (Beijerinckiaceae)	Beijerinckiafluminensis NR_116306.1	√	√	√	√	√		
		Rhizobiales (Rhizobiaceae)	<i>Rhizobium leguminosarum</i>	√			√	√		√
		Rhizobiales (Xanthobacteraceae)	<i>Bradyrhizobium japonicum</i>	√			√	√		
		Sphingomonadales (Sphingomonadaceae)	<i>Novosphingobium aromaticivorans</i>	√	√		√	√		
		Burkholderiales (Burkholderiaceae)	<i>Burkholderia vietnamiensis</i>	√	√			√		
		Enterobacterales (Enterobacteriaceae)	<i>Klebsiella pneumoniae</i>	√	√			√		
		Enterobacterales (Yersiniaceae)	<i>Rahnella aquatilis</i>		√			√		
		Pseudomonadales_ (Moraxellaceae)	<i>Acinetobacter baumannii</i>		√		√	√	√	
		Pseudomonadales (Pseudomonadaceae)	<i>Pseudomonas fluorescens</i>		√		√	√	√	√
Fungi	Ascomycota (Dothideomycetes)	Cladosporiales (Cladosporiaceae)	Cladosporium cladosporioides							√
	Ascomycota (Eurotiomycetes)	Eurotiales (Aspergillaceae)	Penicillium chrysogenum	√				√	√	
	Ascomycota (Saccharomycetes)	Saccharomycetales (Pichiaceae)	Pichia membranifaciens						√	√
	Ascomycota (Sordariomycetes)	Hypocreales (Hypocreaceae)	Hypomyces aurantius						√	√

	Pleosporales (Pleosporaceae)	Curvularia spicifera						√
Basidiomycota (Agaricomycetes)	Agaricales (Amanitaceae)	Amanita	√					
	Cantharellales (Ceratobasidiaceae)	Rhizoctonia solani						√
Basidiomycota (Exobasidiomycetes)	Robbaurales (Robbauraceae)	Robbauera albescens					√	
Basidiomycota (Microbotryomycetes)	Sporidiobolales (Sporidiobolaceae)	<i>Rhodotorula mucilaginosa</i> YR07				√		
Mucoromycota (Glomeromycetes)	Diversisporales (Acaulosporaceae)	Acaulospora colombiana	√	√	√	√	√	
	Diversisporales (Gigasporaceae)	Gigaspora rosea	√	√	√	√	√	
	Glomerales (Claroideoglomeraceae)	Claroideoglomus walkeri	√	√	√	√	√	
	Glomerales (Glomeraceae)	Funneliformis mosseae	√	√	√	√	√	
		Rhizophagus irregularis	√	√	√	√	√	

^aOperational group: combined results for *B. amyloliquifaciens*, *B. vazezensis* and *B. siamensis*

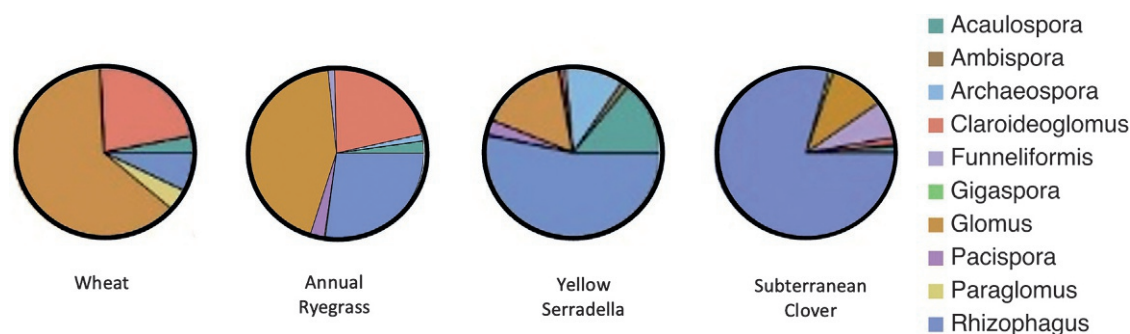


Fig. 1 Diagrammatic representation of the major genera of indigenous arbuscular mycorrhizal fungi recruited into the rhizobiome of four plant species (wheat - L.; annual ryegrass – *Lolium rigidum* L.; yellow serradella – *Ornithopus compressus* L. and subterranean clover – *Trifolium subterraneum* L.) grown in pots of the same soil. There are differences in the genera present and in their relative numbers.

Table 3 Processes contributing to land degradation, possible remedies and consequences for the soil and environment.

Processes in land degradation	Possible remedial actions	Consequences of degradation
Protect the soil surface from erosion by wind and water.	Adopt reduced and conservation tillage. Enhance residue or living-crop cover. Establish hedges and windbreaks	Eutrophication of surface waters, flooding. Poor plant development, structural decline, reduced biodiversity.
Protect the soil from compaction and hard-setting	Build up soil organic matter, adopt machinery with appropriate axle weight limits	Landslides, poor plant establishment and restricted root development.
Protect the soil from acidification	Apply lime and animal manure, avoiding over application of N-fertilizers and manure.	Poor plant development, structural decline, reduced biodiversity.
Protect the soil from organic matter decline.	Adopt reduced and conservation tillage. Optimize crop growth to support rhizodeposition and apply organic amendments.	Greater release of greenhouse gases. Structural decline, reduced biodiversity. Loss of cation exchange capacity leading to greater loss of nutrient ions by leaching.
Protect the soil from contamination and pollution	Avoid amendments containing potentially toxic metals, use pesticides with care. Apply appropriate conservation measures to prevent wildfires.	Contamination of surface waters. Accumulation of toxic metals in plants and their entry into the human food chain.
Protect the soil from salinization or sodification	Avoid over application of fertilizers, and water and waste water for irrigation.	Greater competition for good quality water for agriculture.
Maintain the biodiversity of soil micro-, meso- and macrobiota	Adopt reduced and conservation tillage. Enhance diversity and optimize crop growth to support rhizodeposition and apply organic amendments. Apply appropriate conservation measures to prevent wildfires.	Risks from pests and invasive species.

that of any indigenous rhizobia. A sufficient concentration of microbes applied to soil and seed is necessary as rhizobia require a minimum number to act together to effect root colonization. Typically, inoculation is beneficial if the host-crop species has not been grown on the land previously so that no compatible rhizobia are present, and the supply of mineral nitrogen from the soil is inadequate for plant growth. For example, Goss et al. (2002) reported that nitrogen fixation in soybean (*Glycine max* (L.) Merrill) is reduced by 0.67 kg N ha⁻¹ for every kg ha⁻¹ of mineralized or fertilizer N present in the soil at planting. A reduction in the number of nodules is one of the factors involved, but both transcription and functioning of nitrogenase are negatively affected by the presence of ammonium (NH₄⁺). Free-draining soils are also preferred to ensure appropriate aeration. Excess water can impair oxygen supply that is essential for optimum fixation by the nodules. However, the rate of fixation in leguminous plants is also reduced when water is restricted. Co-inoculation of non-symbiotic plant growth promoting rhizobacteria with rhizobia has been used to increase nodule formation or enhance the benefits from nitrogen fixation in rhizobial symbiosis for growth. However, Zeffa et al. (2020) found no evidence for increased grain yield, following co-inoculation, from a meta-analysis of papers published over 31 years from 1987.

Inoculation of plant growth promoting rhizosphere bacteria and archaea

Concerns for impacts on the environment of nitrogen and phosphatic fertilizers together with the use of synthetic insecticides and herbicides have increased interest in the application of rhizosphere bacteria (other than rhizobia) and archaea to nursery and field crops. Bacteria that secrete antibiotics or induce systemic resistance in target plants have been applied to protect crops from disease

and pathogens. To increase nutrient availability, a broad spectrum of bacteria has been applied, such as diazotrophic (N), phosphatase-producing, phosphorus-solubilizing or hydrogen cyanide-releasing (P), ionophore or organic acid releasing (iron (Fe) and micronutrients). Inocula that contains these groups of microbes, often together with others that release plant growth promoters, including plant hormones, salicylic acid and the ethylene formation inhibitor ACC-deaminase, have been marketed as biofertilizer. For Plant Growth Promoting Rhizobacteria (PGPR) to be applied successfully in the field, they must demonstrate 'rhizosphere competence' and colonize the soil, particularly in the rhizosphere or the root (endophytes). To an extent, colonization depends on their competitiveness with other microorganisms but plants recruit specific microbial communities from the bulk soil into their rhizosphere using signal molecules, which are secondary metabolite components of their root exudates. The composition of exudates varies across plant species and with their age and the season, thereby allowing plants to influence the dynamics of their rhizobiomes. However, this is not entirely a plant-driven process. Some PGPR can change the secondary metabolites released by their rhizosphere host, thereby enhancing their coordination. Persistence of an organism within the rhizosphere will depend on its ability to maintain its metabolic functions. The soil characteristics within the rhizosphere, such as pH, nutrient availability, water supply and redox conditions, must also be appropriate. For the majority of PGPR to function fully will mean participation in biofilms that can locate in niches that allow access to sources of macro- and micro-nutrients. The requirements for colonization and persistence in the rhizosphere will influence the specifications of the inoculum delivery system and may also contribute to the challenges for successful application of biofertilizers to field crops.

The importance of the microbe selection by plants and their subsequent interactions can also have significant consequences for the efficacy of microbial inoculation for ecosystem restoration. For example, the direction of ecosystem change and the development of native plant communities from arable to grassland or dry heathland, on the same sandy glacial deposit, were greater after inoculation with unsterilized soil if the topsoil was removed rather than being left in place (Wubs et al., 2016). This highlights the significance of the population size, diversity and ability of indigenous microbes to compete for niche habitat and persist.

A final point about the use of inoculum derived from rhizosphere microbes is that there is a considerable similarity between the organisms present in rhizobiomes and those in the human gut, including some that can be pathogenic. Appropriate biosecurity protocols are necessary when soil microbiomes are used as inoculum, both to protect operatives and the wider environment. Inoculant organisms may alter the balance of the existing consortium of microbes in the rhizosphere, possibly modifying the microbiome functioning, and if persistent could pose the risk of later invasive spread in the presence of other crops. Significantly, the use of fungal inoculants has changed bacterial populations in a majority of examples assessed, generally leading to greater diversity, whereas bacterial inoculants increased diversity in only 66% of examples (Cornell et al., 2021).

Inoculation with mycorrhizal fungi

Although mycorrhizal fungi are commonly associated with many species of host plants, similar to the situation with rhizobacteria, there is also strong evidence of host-driven selectivity for the AMF present in the roots of individual plants and species. The signal exchange dialogue between plant and fungus, which takes place before root colonization starts, is an important component of the selectivity. Land use and management intensity affect the abundance and species richness of AMF from which host selection can take place. Very degraded soils probably have an inadequate supply or quality of AMF propagules and hence can benefit from inoculation. Other soils, such as those supporting temperate grassland, may have an overwhelming abundance of propagules that originate from an AMF community of considerable richness and can out-compete any introduced inoculum.

Colonization results from three types of propagules: spores, extraradical mycelium (ERM), and colonized root fragments. Spores are the easiest propagules to produce as commercial inoculum, but there are issues related to germination and viability after production and limitations to the diversity of species that are readily amenable to the processes involved. Colonized root fragments, together with spores, are likely to be present when freshly collected soil samples are used as inoculum. Although ERM can remain viable in soil after freezing or drought, it rapidly loses infectivity following disturbance (Addy et al., 1997; Goss et al., 2017). Thus, for ERM to be important as a propagule, it needs to have developed from colonized roots, and for roots of target plants to explore the same undisturbed soil volume as that explored by the mycelium. This has resulted in experimental systems that grow 'nurse' plants, colonized by one or more selected AMF to produce ERM that infects batches of serially germinated target plants. The members of each batch all become colonized at the same time and more rapidly than typically achieved using spores. This concept can be exploited in agriculture and horticulture (see Section "Examples of managing AMF") but does not lend itself to being the basis for a commercial inoculum.

The establishment of a mutualistic symbiosis between a mycorrhizal fungus and the roots of its plant host takes place in the presence of all other microorganisms in the immediate rhizobiome. There is evidence from using fresh soil as an inoculum in the field and direct laboratory studies, that some soil microbes, especially bacteria, contribute to the success of the process and promote mycorrhizal development. Both AMF and ectomycorrhizal fungi (EMF) are associated with rhizobacterial strains, 'mycorrhiza helper bacteria' (MHB), that support the development and establishment of their symbioses. There are also bacterial endophytes found within the hyphae of AMF that also support the formation of the mycorrhiza. The specificity of MHB for the fungal partner is still uncertain. Within the MHB, there are strains that stimulate the germination of mycorrhizal fungus spores, modify the radial growth of fungal mycelium or increase the branching of hyphae, all important for the initial colonization process. At the same time, some MHB can influence the architecture of the host root system leading to enhanced branching and to longer branch roots. The combination of these effects on fungus and host roots could contribute to increased root colonization. However, the formation of more root tips and hyphal branches could be equally beneficial in enhancing the signaling between potential symbionts. Greater exchange of signals together with release of plant hormones or any deficient nutrients could also result in greater colonization. The

identification of appropriate MHB of specific mycorrhizal fungi can increase the efficacy of mycorrhiza formation, requiring less propagule material. Other bacteria associated externally on fungal hyphae appear to operate symbiotically, with the physiology of both partners showing changes.

Mycorrhizal fungi and soil bacteria typically form biofilms in soil, and fungal hyphae can act as the substrate for their formation by bacteria or be part of a consortium of microbes in a common biofilm. There is significant interest in the inoculation of organisms, which on being applied to soil readily form a common biofilm and can recruit additional members. Such a technique or the application of organisms, already part of a biofilm, to soil could help overcome the problems of ensuring the persistence of beneficial microbes after inoculation and keeping synergistic partners together for long enough to influence target plants.

Reliability of inoculation

What is clear from assessments of empirical studies is that the success of inoculation to manipulate the soil biota depends on effecting positive changes to the biodiversity of the indigenous community. Changes to the biodiversity of bacterial communities can be achieved through both fungal and bacterial inoculants. Although positive changes are reported, negative change and no-change are frequently reported, indicating a lack of reliability in implementing this management practice.

Exploitation of indigenous microbes

In a critical review of the performance of P-solubilizing bacteria in the field, [Raymond et al. \(2021\)](#) argued that they do not mobilize enough P to alter the nutrition of a crop under field conditions. These authors recommend that a more promising approach would be to support the indigenous soil microbiome in the mobilization and cycling of P. Numerous studies have reported that the application of local inoculum from a similar soil, which contains vital whole microbial communities, is more effective in improving plant performance than the addition of individual exotic species or strains. In agricultural ecosystems, the opportunity for in-situ exploitation of indigenous microbes is even greater than for restoration of natural ecosystems.

Pre-requisites for ensuring appropriate levels of biological and functional diversity in soil microbiomes

Establishment of a resilient soil biology depends on maintaining suitable habitats for the organisms that comprise the soil food web. Restricting disturbance of soil mechanically, chemically or biologically helps stabilize soil structure, maintain niche habitats, biofilms and mycorrhizal ERM networks, prevents the accumulation of excess mineral nutrients, toxic metals and organic molecules, and encourages cooperative activity or syntrophy. Effectiveness of the food web depends on ensuring a supply of easily accessible organic matter that can support the numbers and diversity of soil microbes. In turn, this supply of energy drives the soil carbon and nitrogen cycles and releases other macro- and micro-mineral nutrients. Rhizodeposition, root turnover and plant residues all contribute to the provision of organic matter but additional external sources, such as compost or animal manure, might be required. The actions of roots, soil fauna and microbes link soil particles to organic matter and form microaggregates, which combine into macroaggregates and then into structured soil. The pores within and between aggregates provide trophic niches and store or transport water and air. Soil organisms and roots further mold the soil to create larger pores for their own passage. The organic molecules released by roots of different plant genotypes attract consortia of microbes into the rhizosphere from the bulk soil that differ in the taxa present and relative numbers of individuals present. Precise membership of the consortia changes with the age or growth stage of the plants and is further influenced by soil and environmental factors.

Many of the actions associated with agronomic activity can impact, positively or negatively, on this complex balance within the soil. In the context of microbe management, agronomic practices need to be selected by taking into account their potential influence on soil microbial diversity and activity.

Farming systems and agronomic management practices

Three major farming systems dominate crop production: conventional systems that rely on mineral fertilizers to provide the main nutrients for the crops and annually till the soil to prepare a seedbed; organic systems that use animal manure, vegetable compost and cover crops to supply most of the nutrients but the soil is nearly always tilled annually; conservation systems that aim to protect the soil from erosion and conserve its ability to provide key ecosystem services, particularly by reducing excessive cultivation in favor of limited or strategic tillage practices. In addition to nutrient management and tillage practices, weed and disease control tend to differ between farming systems, with conventional and conservation systems relying on chemical control. Individual producers may adopt a package of measures selected from some or all the major systems.

Practices associated with each of the main management areas – tillage, cropping system, nutrient and disease management – may have positive or negative effects on soil biota, therefore, they provide opportunities to manage the direction and extent of any resulting change.

Tillage

The main tillage systems and their consequences for soil properties and microorganisms are summarized in [Table 4](#).

Table 4 Characteristics, aims and some essential effects of various tillage systems and their potential impacts on soil microbes.

<i>Tillage technique</i>	<i>Main machines involved</i>	<i>Principal aims and effects achieved</i>	<i>Possible effects on microorganisms</i>
Inversion tillage	Mouldboard plough, offset disks	Effects inversion of the soil. Disturbs and fragments the soil; buries weeds and residues from the previous crop, incorporates manure and both organic and inorganic amendments. Conventional tillage is usually associated with the use of the mouldboard plough. Increased risk of soil erosion.	At the plough layer, roots, AMF mycelium and biological channels are disrupted. Soil organic matter (SOM), less mobile nutrients and spores are distributed more uniformly. Biofilms are likely to be disrupted. AMF community is changed in favor of species that produce more spores.
Ridge till	Ridge tillage cultivator and modified planter	Soil is pushed by pairs of winged tines to form ridges up to 20 cm deep. The harvest residues on the surface are removed and the top of the ridge flattened by a horizontal blade so that the seed is planted in the ridges without residue cover in the row. Contrasting soil conditions are created in the ridge and between the ridges. Residues and organic amendments are mixed with soil in the sides of the ridges each year. The ridges warm up quicker than flat land in spring, in part because the ridges drain more readily.	At the inter-rows, there is little root development. In the ridges, SOM builds up and conditions for root development and microbial growth are improved. Once ridges are formed, there is little disruption of microbes. Biofilms near the center of ridges are largely protected as are AMF that preferentially start new infections from intact ERM.
Reduced tillage	Rigid or sprung tines	No soil inversion. Typically harvest residues are left on the soil surface.	The disruption of roots, AMF mycelium and biopores is partial. Stratification of SOM, less mobile nutrients and spores is favored.
Deep reduced tillage	Rigid tines	Non-inversion tillage to a depth greater than 100 to 150 mm.	The disruption of roots, AMF mycelium and vertical biopores is partial.
Shallow (reduced) tillage	Rigid or sprung tines	Non-inversion tillage to a depth of less than 100 mm.	The disruption of roots, AMF mycelium and biopores occurs in the disturbed soil. Residues remain covering the soil surface
Stubble-mulch tillage	Tines or disks	Residues retained on the soil surface, with soil disturbance without inversion	Disruption extensive to the depth of the tillage tool.
Zero-tillage/no-till/direct drilling	Drill with combination of tines or disks	Seed is drilled into the stubble of the previous crop with minor soil disturbance confined to the row and does not extend below seeding depth. Soil conditions are closer to the ones found in the natural environment if compaction can be avoided. At the micro-scale there is the opportunity for contrasting ecological niches to develop in the soil.	Roots, AMF mycelium and root channels are left undisturbed. Stratification of SOM, less mobile nutrients and spores is maximized. Biofilms are not disrupted, except in the seed row. AMF species that produce less spores are protected.
Zone or Strip-tillage	Rigid or sprung tines	Zone-tillage is a development of no-tillage crop production. It consists of tillage with coulters or rigid tines (down to 15 cm below the soil surface) in a narrow strip of soil. Seeding takes place in the tilled zones, with the inter-row area remaining undisturbed and retaining crop residues.	The disruption of roots, AMF mycelium and roots channels are partial. Stratification of SOM, less mobile nutrients and spores is favored. In a large scale, contrasting soil conditions are induced between tilled and no-tilled zones. In the latter, soil conditions will be similar to no-till systems.
Strategic tillage	Rigid or sprung tines, mouldboard plough	A system predominantly using reduced tillage equipment but includes the use of mouldboard ploughing as a flexible responsive to seedbed concerns (typically weeds) or the need to incorporate organic amendments, manure or crop residues. Avoids long-term development of soil conditions typical of natural soils.	Roots and AMF mycelium periodically disrupted
Rotational ploughing	Rigid or sprung tines, mouldboard plough	Reduced tillage or no-till is used routinely for soil preparation but a plough is used at key points in the rotation	Roots and AMF mycelium periodically disrupted
Secondary tillage	Powered or unpowered tines, disks, rotary implements, rollers	Tillage practices used after the principal tool to create a seedbed and provide additional weed control	Completes disruption of roots, AMF mycelium and biopores throughout the depth of operation. Homogenises the distribution of soil organic matter, immobile nutrients and spores.
Remedial tillage	Subsoiler tines, rotary spades or mole plough	Tillage used when required to remove compacted layers or improve water movement	Improves air-filled porosity and the continuity of vertical structural pores and increases the depth of rooting

The effect of conventional tillage on soil biology is both physical and chemical, with the direct impact of implements, disruption of structure and abrupt changes in aeration and water content. The latter is mainly the result of moist soil from depth being brought to the surface. Action of ploughs and tines disrupts fungal hyphal networks and breaks up root systems and stems infected by soil-borne pathogens, reducing the size of individual sources and consequently their inoculum potential. The uniformity of soil organic carbon (SOC) content and pH in the profile and the imposed changes in soil structure, together with the direct impact on earthworm populations and burrows, also reduce the diversity of trophic microsites within the tilled layer and the populations of microorganism within them. The effects of reducing tillage depth and adoption of conservation measures are similar but confined to shallower depths. The lack of inversion likely reduces the scale of any change in aeration and water, except in zones of previously compacted soil. Under no-till, drying of the soil surface may be reduced because of the presence of crop and harvest residues, which also aid infiltration and protect from erosion and raindrop impact. The reduced physical disruption and absence of inversion improves the stability of soil aggregates, which protects against erosion and consolidation, and helps to maintain porosity and pore continuity. The inherent heterogeneity in structure can lead to differences in microbial biomass and microbial diversity. Similarly, the contrast in distribution of SOC and variability in food resources between tillage systems can influence the location and diversity of soil organisms within the profile. In terms of inducing the effects of both structural and chemical heterogeneity, chisel ploughs can be as effective as no-till (Zuber and Villamil, 2016).

Within the soil macrofauna, gastropods, arthropods, earthworms and insects have been considered sensitive to soil disturbance, although experimental results are more variable.

Archaea and bacteria show responses to tillage but the scale of study is important. At the scale of soil aggregates, no effects on abundance are found but where tillage reduces aggregation or number of aggregates, then abundance is less. Effects of conventional tillage tend to reduce the abundance or biomass of microbes near the soil surface, but when considered over the depth of the soil profile differences between ridge-till and no-till are not significant. In shallow-tilled soil, impacts in the disturbed layer are similar to those of conventionally tilled soil. Certain groups of archaea, bacteria and fungi are more associated with conventional cultivation, whereas other groups are mostly found in untilled soil (Table 5). For example, anaerobic methanogenic archaea are more commonly reported in no-till. However, a reduction in soil disruption is associated with increased diversity of organisms (especially minor representatives) and their level of activity (Table 6). The added input of crop residues under conservation tillage can also support a larger abundance of microbes (Table 6).

Given the changes that can result from full inversion tillage, it can be used as a means of resetting the microbial population to a point that a new crop can establish its own rhizobiomes, free from well-established sources of root disease organisms and other imbalances. It is a common practice to use tillage to overcome replant issues for ginseng (*Panax ginseng* C. A. Meyer).

Nutrient management

At maturity, plants have accumulated different contents of nutrients according to their needs for growth and development (Table 7, Tubeileh and Goss, 2022). Even for the macronutrients N, P and potassium (K) the variation is considerable between plants in the same family (e.g., maize – *Zea mays* L. – and rice – *Oryza sativa* L. or chickpea – *Cicer arietinum* L. – and alfalfa – *Medicago sativa* L.). To meet plant requirements for nutrients in agricultural and forestry ecosystems, mineral fertilizers and organic amendments, such as animal manure, are applied to maintain the amounts available in soil. In addition to mineral forms of N and P, manure also contains organic forms, micronutrients and other carbon compounds. The latter contribute to the soil stock of SOC that can greatly enhance the number of bacteria in soil, but the effect on fungal populations is small in the absence of organic-based nutrient additions (Fig. 2).

The contributions of soil organisms to the availability of macro- and micro-nutrients in soil (described above) are also influenced by the form of nutrients used in replenishing them. Application of mineral fertilizers, particularly ammoniacal-N fertilizer, can result in acidification of soil, whereas animal manure tends to increase soil pH. Variation in soil pH greatly influences bacterial abundance, with decreases in total microbial biomass below pH 5 but increases above and in community composition (Fig. 3). Mineral N fertilizer alone, particularly at annual applications of 100 kg N ha⁻¹ (e.g., Zhou et al., 2017), tends to decrease microbial biomass, whereas smaller applications and meeting crop needs for P and K as well as N may result in increased microbial biomass. The effects on Gram positive (GM⁺) bacteria are different from those of Gram negative (GM⁻), resulting in an increase in the ratio of GM⁺:GM⁻. Effects on bacterial diversity tend to reflect the differences in biomass. However, diversity may not be reduced if N is applied together with either P or K rather than combined with P and K. The exception to this pattern of response to N-fertilizer is in anaerobic soils, such as those growing paddy rice, where the application of N alone increases bacterial diversity.

Nutrient management practices, including the application of lime to decrease acidity, greatly influence the composition of soil bacterial communities. For example, Chinnadurai et al. (2014) reported that the greatest number of phyla was associated with soil receiving mineral fertilizers, which also contained one phylum not found in other treatments (Fig. 4B). Furthermore, the relative abundance of bacteria in the major phyla varied between soil receiving mineral and organic forms of nutrients, and both differed from that in the control where no nutrients were applied (Fig. 4A and C).

The abundance of the amoA genes from ammonia oxidizing archaea (AOA) and ammonia oxidizing bacteria (AOB) reflects the numbers of organisms present. Application of N produces greater increases in amoA gene abundance in AOB than in AOA. The response is unrelated to the rate of N application but to the total load of N applied. However, AOB are more responsive to mineral-N than to organic-N but AOA respond similarly to both forms. The smaller AOB populations in soils with no history of receiving additions of N show a Greater reaction than those under agricultural or grazed grassland management. In contrast, AOA show a greater response to N in agricultural arable soils than in pasture soils.

Table 5 Examples of Archaea, Bacteria and Arbuscular Mycorrhizal Fungi that are more commonly associated with particular tillage regimes at various taxonomic levels.

<i>Domain</i>	<i>Phylum</i>	<i>Order</i>	<i>Family</i>	<i>Genus (Species)</i>
a) Conventional tillage: disturbed habitat				
Archaea	Nanoarchaeota	Woesearchaeales		
Archaea	Thermoproteota			
Archaea	Thermoproteota	Nitrososphaerales	Nitrososphaeraceae	Nitrosocosmicus
Archaea	Thermoplasmatota	Nitrososphaerales	Nitrosopumilaceae	Nitrosotalea
Bacteria	Cyanobacteria			
Bacteria	Actinobacteriota	Mycobacteriales	Geodermatophilaceae	
Bacteria	Actinobacteriota	Streptomycetales	Streptomycetaceae	
Bacteria	Actinobacteriota	Actinomycetales	Dermatophilaceae	
Bacteria	Actinobacteriota	Actinomycetales	Kineosporiaceae	Kineosporia
Bacteria	Actinobacteriota	Actinomycetales	Micrococcaceae	
Bacteria	Bacteroidota	Chitinophagales	Chitinophagaceae	Flavisolibacter
Bacteria	Bacteroidota	Sphingobacteriales	Sphingobacteriaceae	Pedobacter
Bacteria	Deinococcota	Deinococcales	Thermaceae	Thermus
Bacteria	Planctomycetota			
Bacteria	Proteobacteria	Burkholderiales	Burkholderiaceae	Herbaspirillum, Ramlibacter
Bacteria	Proteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingomonad
Fungi	Mucoromycota	Glomerales	Glomeraceae	Glomus (microaggregatum)
Fungi	Mucoromycota	Glomerales	Glomeraceae	Funneliformis mosseae
Fungi	Mucoromycota	Diversisporales	Diversisporaceae	Diversispora (versiformis)
b) No-till: sensitive to change				
Archaea	Methanobacteriota	Methanobacteriales	Methanobacteriaceae	Methanobacterium
Archaea	Halobacteriota	Methanomicrobiales	Methanomicrobiaceae	Methanomicrobium
Archaea	Halobacteriota	Methanomicrobiales	Methanoregulaceae	Methanolinea
Archaea	Halobacteriota	Methanomicrobiales	Methanoregulaceae	Methanoregula
Archaea	Halobacteriota	Methanosarcinales	Methanosarcinaceae	Methanomethylovorans
Archaea	Methanomicrobia	Methanosarcinales	Methanosarcinaceae	
Bacteria	Acidobacteriota	Acidobacteriales		
Bacteria	Actinobacteriota	Propionibacteriales	Nocardoidaceae	
Bacteria	Bacteroidota	Chitinophagales	Chitinophagaceae	Flavisolibacter
Bacteria	Bacteroidota	Chitinophagales	Chitinophagaceae	Terrimonas
Bacteria	Chlamydiae	Chlamydiales	Chlamydiaceae	Chlamydia
Bacteria	Chlorobi	Chlorobiales		
Bacteria	Firmicutes			
Bacteria	Nitrospirota			
Bacteria	Proteobacteria	Xanthomonadales	Xanthomonadaceae	Xanthomonas
Bacteria	Proteobacteria	Rhizobiales	Hyphomicrobiaceae	Hyphomicrobium
Bacteria	Verrucomicrobiata	Verrucomicrobiales		
Fungi	Mucoromycota	Diversisporales	Acaulosporaceae	Acaulospora (colombiana)
Fungi	Mucoromycota	Archaeosporales	Archaeosporaceae	Archaeospora (undulata)
Fungi	Mucoromycota	Diversisporales	Gigasporaceae	Gigaspora (gigantea)
Fungi	Mucoromycota	Glomerales	Glomeraceae	Rhizophagus (diaphanum)

Table 6 The effects of different tillage systems on key parameters of microbial functioning.

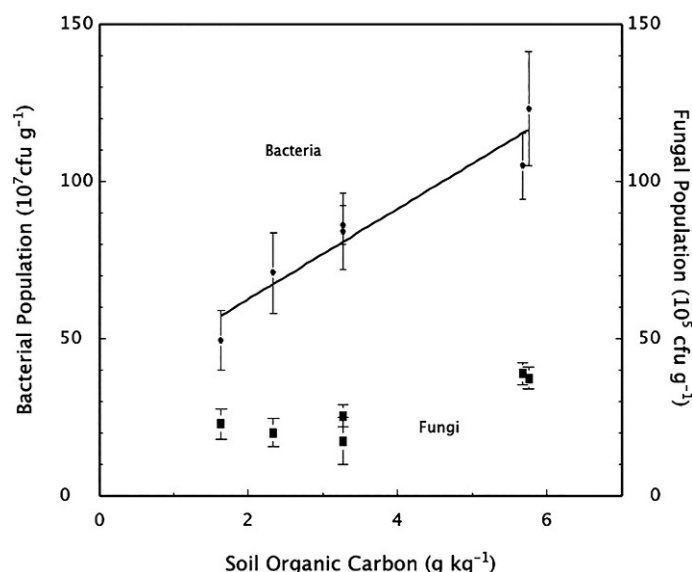
<i>Microbial parameter</i>	<i>Tillage System</i>			
	<i>Conventional tillage</i>	<i>Ridge-till</i>	<i>No-till</i>	<i>Conservation tillage</i>
Microbial biomass/abundance	++	+++ ^a	+++ ^a	+++
Microbial diversity	++	+++	+++	+++
Biological activity	++	+++	+++	+++

^aIndicates that increases occur in some layers but summed over the soil profile, effects are not statistically significant from conventional tillage.

Table 7 Approximate content of macronutrients in some agricultural crops at harvest and in typical animal manures from farms, expressed in g kg⁻¹ dry matter, together with their ratio, based on P.

	Nutrient content (g kg ⁻¹)			Nutrient ratio
	N	P	K	N:P:K
Crop				
Maize (<i>Zea mays</i> L.)	13.5	1.8	14.1	7.5:1: 7.8
Rice (<i>Oryza sativa</i> L.)	9.3	1.6	10.3	5.5:1: 6.4
Wheat (<i>Triticum aestivum</i> L.)	12.4	1.3	13.8	9.5:1:10.6
Oat (<i>Avena sativa</i> L.)	9.0	1.8	16.9	5.0:1: 9.4
Soybean (<i>Glycine max</i> L. Merr.)	31.7	3.8	15.4	8.4:1: 4.1
Chickpea (<i>Cicer arietinum</i> L.)	19.9	4.0	10.7	5.0:1: 2.7
Lentil (<i>Lens culinaris</i> Medik.)	15.9	2.9	13.7	5.5:1: 4.7
Alfalfa (<i>Medicago sativa</i> L.)	39.9	3.3	29.3	12.1:1: 8.9
Manure				
Liquid manure, dairy cattle	3.5	0.8	2.4	4.3:1: 3.0
Solid manure, beef cattle	7.4	2.4	5.7	3.1:1: 2.4
Liquid manure, pig	4.0	1.3	1.8	3.1:1: 1.4
Solid manure, chicken (layer)	19.3	8.9	8.0	2.2:1: 0.9

Modified from Tubeileh AM and Goss MJ (2022) Assessing the effects of using animal manure on soil health. In Horwath R (ed.) *Improving Soil Health*. pp. 281–308. Cambridge, UK: Burleigh Dodds Science Publishing.

**Fig. 2** Effects of changes in soil organic carbon on the populations of bacteria and fungi in a sandy loam (Typic Haplustalf). The largest populations of both bacteria and fungi were present in soil receiving organic amendments (composted cattle manure). Data from Chinnadurai et al. (2014).

The form of nitrogen application, organic amendment or mineral fertilizer, also changes community composition of both bacteria and fungi (Toljander et al., 2008). The application of ammonium sulphate, (NH₄)₂SO₄, fertilizer or sewage sludge affects community structure very differently. Some changes result from arrival in the soil of dominant phyla present in the sewage sludge, whereas others stem from stimulation of the indigenous population by some of its components. Organic amendments can favor Ascomycota fungi, although saprotrophic fungi may not be affected by nutrient addition, whether as NPK mineral fertilizer or as cattle manure. However, application of mineral nutrients can result in reductions of AMF colonization and growth of ERM. In contrast, increases of both these variables, together with the number of AMF species present, can result if soil only receives nutrients with the application of manure. Toljander et al. (2008) found that the number of AMF species and composition of the community were influenced by the form of N applied. These changes can also be associated with modifications to soil pH resulting from the mineral fertilizers (increased acidity, lower pH) and manure (decreased acidity and increased as pH) as an important factor modifying the microbial community structure, particularly of bacteria.

Changing the availability of N in the soil, such as by making N applications that are greater than 100 kg N ha⁻¹, can be associated with a decline in root colonization by AMF and a reduction in their abundance. However, as such effects are not uniform across all AMF taxa, the relative abundance of taxa and hence the AMF community composition can be modified. In turn, that can influence the community composition of host plants. Nevertheless, addition of mineral N may have no significant effects on species richness

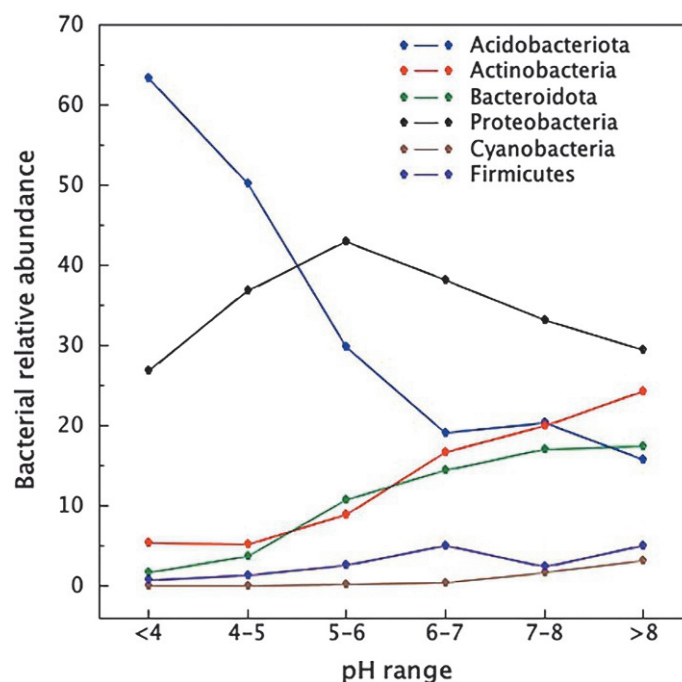


Fig. 3 Relative abundance of some major bacterial phyla in communities across 88 soils from North and South America grouped according to pH. Data from Lauber et al. (2009).

or diversity of AMF, a response that is strongly linked to the relative availability of soil P (i.e., the soil available N:P ratio). Application of N to soil with an adequate content of available P appears to increase the likelihood of a negative effect on AMF. In contrast, increased N supply, when P availability to host plants is limiting in the absence of mycorrhiza, may stimulate plants to provide more carbon and thereby enhance AMF presence, consequently obtaining more P.

The carbon cost to the plant per unit of N gained can be three times less when it is mycorrhizal than when it is not. Furthermore, AMF are more efficient in obtaining N from organic sources than from mineral fertilizer (Chowdhury et al., 2022). Although AMF hyphae lack the key enzymes required to mineralize N from organic matter, they can take up many amino acids. The AMF take up both NH_4^+ (via transporters) and NO_3^- (following conversion to NH_4^+ by nitrate (NO_3^-) and nitrite (NO_2^-) reductases) and convert the mineral-N into amino acids. Importantly, NH_4^+ is the form of N transported across the arbuscule-plant membrane interface and into the plant. The release of exudate from ERM hyphae helps recruit a selection of soil bacteria to the hyphosphere (the equivalent of the rhizosphere for ERM), including the hyphal surfaces, which allows the ERM from mycorrhizal plants to exert control over mineralization of organic N through the functioning bacterial population that accompanies it.

No unique relationships have been identified between individual decomposers, which mineralize N but only make it available to the ERM of specific AMF. This generalized interaction between AMF and microbial decomposers in the provision of N contrasts with the release and uptake of P from phytate. The AMF have been shown to provide a route, via the surface of extraradical hyphae, by which particular mobile microbes can reach and release mineral-P from the organic source in soil, as exemplified by *Rahnella aquatilis*, a member of the Gammaproteobacteria, and the AMF, *Rhizophagus irregularis* (Jiang et al., 2021).

A key aspect of the response of AMF to P is the effect on the symbiosis and how the interdependency might change, especially in agroecosystems. The AMF colonization usually declines with increasing concentrations of available P (Covacevich et al., 2007). However, depending on the soil and crop, there is a critical value of P (between 25 and 30 mg P kg⁻¹) beyond which the level of colonization stabilizes (Fig. 5). The number of AMF species and community structure present in the soil and the density of ERM hyphae (m g⁻¹ soil) decline with soil available P (Lang et al., 2022). Furthermore, the reduction in species abundance can limit the range of mutualism that is available to host plants to meet their requirements for nutrients. The question then arises, "How important to AM symbiosis is the provision of a greater supply of P to the host plant?"

Plant roots typically take up P as dihydrogen phosphate (H_2PO_4^- , P_i) from soil via high-affinity transporters in their epidermal, root hair, and outer cortical cells – the direct uptake pathway (DUP). But when colonized by AMF, a second, symbiotic (SUP) pathway is provided by the fungal symbionts. At larger concentrations of soil P, growth and P_i uptake by mycorrhizal plants may not differ greatly from non-mycorrhizal controls, suggesting that there is no benefit from the mycorrhiza formation. However, AMF vary in the ways they influence the SUP and DUP pathways. It is clear that in mycorrhizal plants, the DUP can be modulated independently of the P_i status through the interaction with SUP (Fig. 6). Importantly, ERM networks can also mediate a reduced supply of P_i by DUP locally, but without affecting that in distant roots, which may not yet be colonized. It also means that solely basing assessment of the benefits of mycorrhiza formation on enhanced P_i provision is inadequate (Facelli et al., 2014).

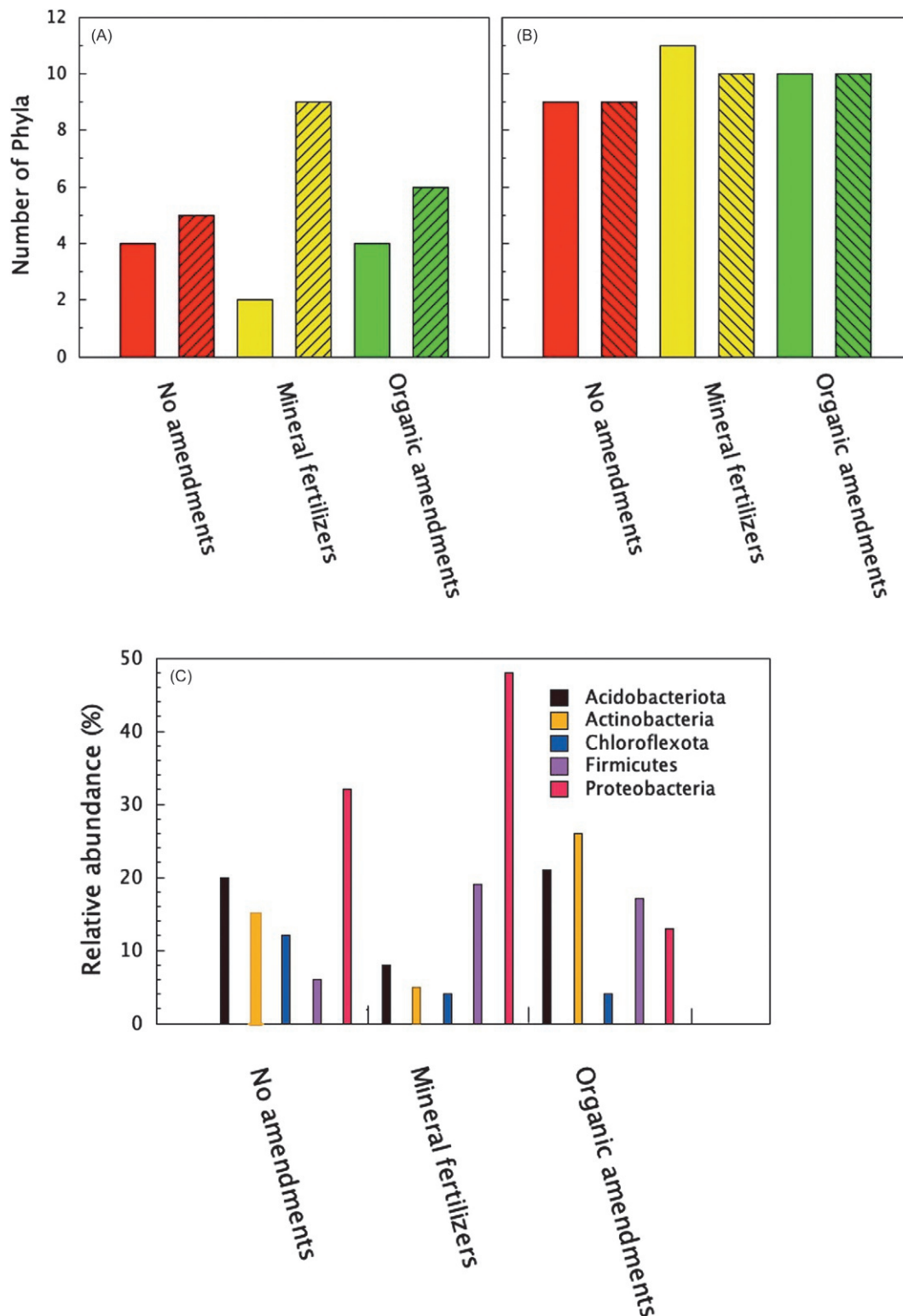


Fig. 4 Effects of nutrient management practices on bacterial community structure. (A) Number of major (plain bars) and minor phyla (hatched bars) present under different nutrient management regimes (B) Variation in the total number of phyla present in soil (plain bars) and the number of phyla also present in soils under the other nutrient practices (hatched bars). (C) The relative abundance of bacteria in the major phyla varies between soils with the distribution showing more evenness where organic forms of nutrients were applied. The largest populations of both bacteria and fungi were in soil receiving organic amendments (composted cattle manure). In the 'Mineral fertilizers' treatment, nutrients in the form of urea (N), super phosphate (P) and muriate of potash (K) was applied in the elemental ratio of 7:1:1. No nutrients were supplied to the 'No amendments' treatment. Data from Chinnadurai et al. (2014).

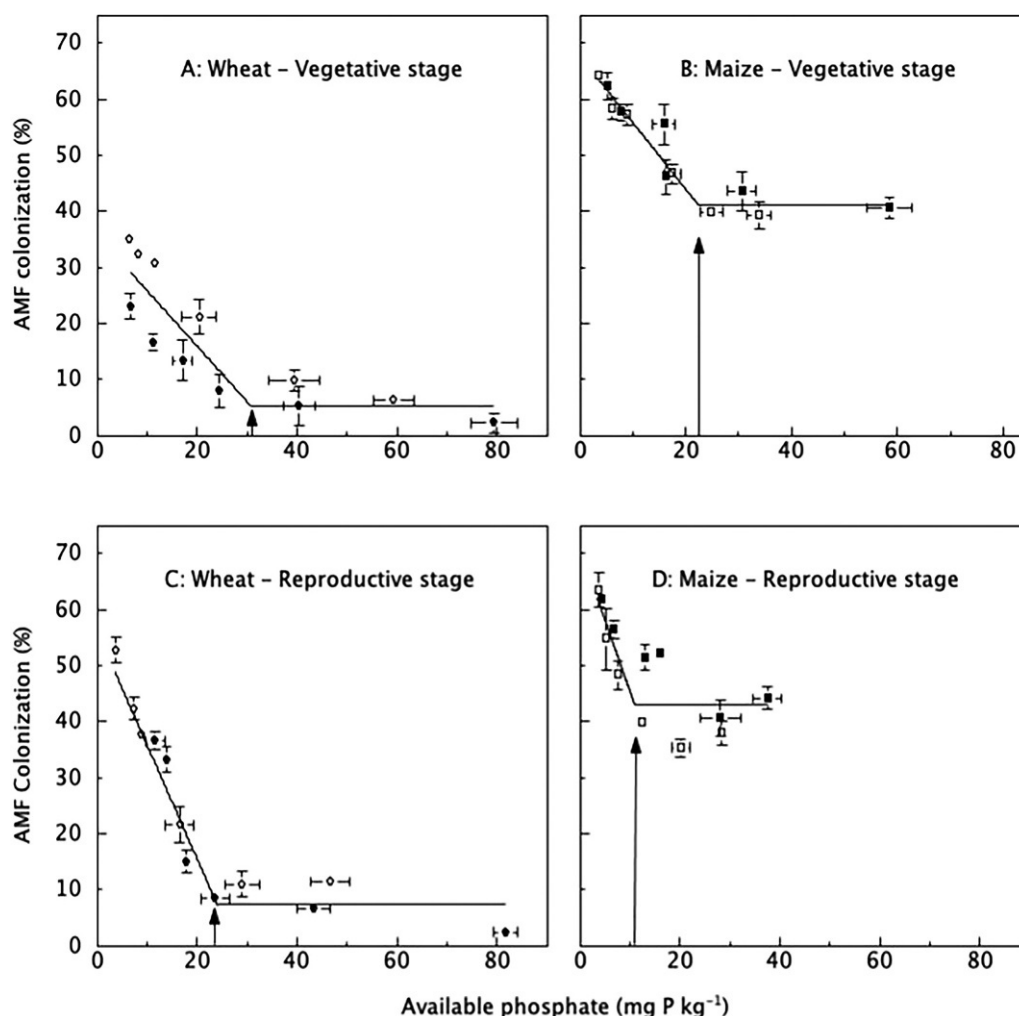


Fig. 5 Effect of phosphate availability on AMF colonization in wheat and maize at vegetative and reproductive stages of growth. Closed and open markers indicate results for 2010 and 2011, respectively. Arrows indicate the break from a general decline in colonization with P concentration to a plateau level of colonization, which is greater for maize than wheat. Critical values of available P range between 10 and 31 mg P kg⁻¹ (Data from [Deng et al., 2017](#)).

The impact of P on AMF colonization suggests that consideration of the distribution of fertilizer P in the soil is important. Under no-till systems, a strong stratification in P concentration results in greater concentrations of P closer to the soil surface than with inversion tillage. Broadcasting of fertilizer enhances this effect. Such information supports the view that fertilizer banding can offer a strategy to reduce the proportion of roots that are exposed to negative impacts of large amounts of soluble phosphate. In similar vein, for small grain cereal crops grown under no-till, broadcasting P fertilizer might reduce maximum soil P concentrations compared with deep banding, and have less negative effect on AMF colonization. Under intensive agricultural systems, deeper soil layers with smaller P concentrations could encourage colonization by a more diverse AMF community than in enriched superficial soil layers.

The abundance and diversity of AMF appear to be enhanced under organic agriculture relative to conventional farming, probably owing to the positive impact of using organic forms of P. However, the differences in abundance and diversity are related to a loss of taxa under intensive conventional agriculture rather than changes in species. [Såle et al. \(2015\)](#) found a greater number of AMF species under organic farming only when reduced tillage was used, but under inversion tillage there were no differences in the number of AMF species between organic and conventional systems. However, benefits from organic forms of P also depend on the type and amount of organic amendment. Consequently, as there are generally several differences in cropping systems between farming types, the impact of the latter is not always easy to interpret.

The cropping system is also a factor in the AMF response to P. For example, abundance of AMF species present in soil was significantly greater under a 7-year crop rotation than under a maize monoculture receiving the same application of mineral P fertilizer ([Oehl et al., 2003](#)), and benefits of rotation occurred even where large amounts of nutrients were applied.

The contribution of AMF to plant P uptake and growth seems to be greater when the nutrients are released from organic sources rather than provided by mineral fertilizers, even where plant P content is the same. Therefore, the large-scale adoption of strategies

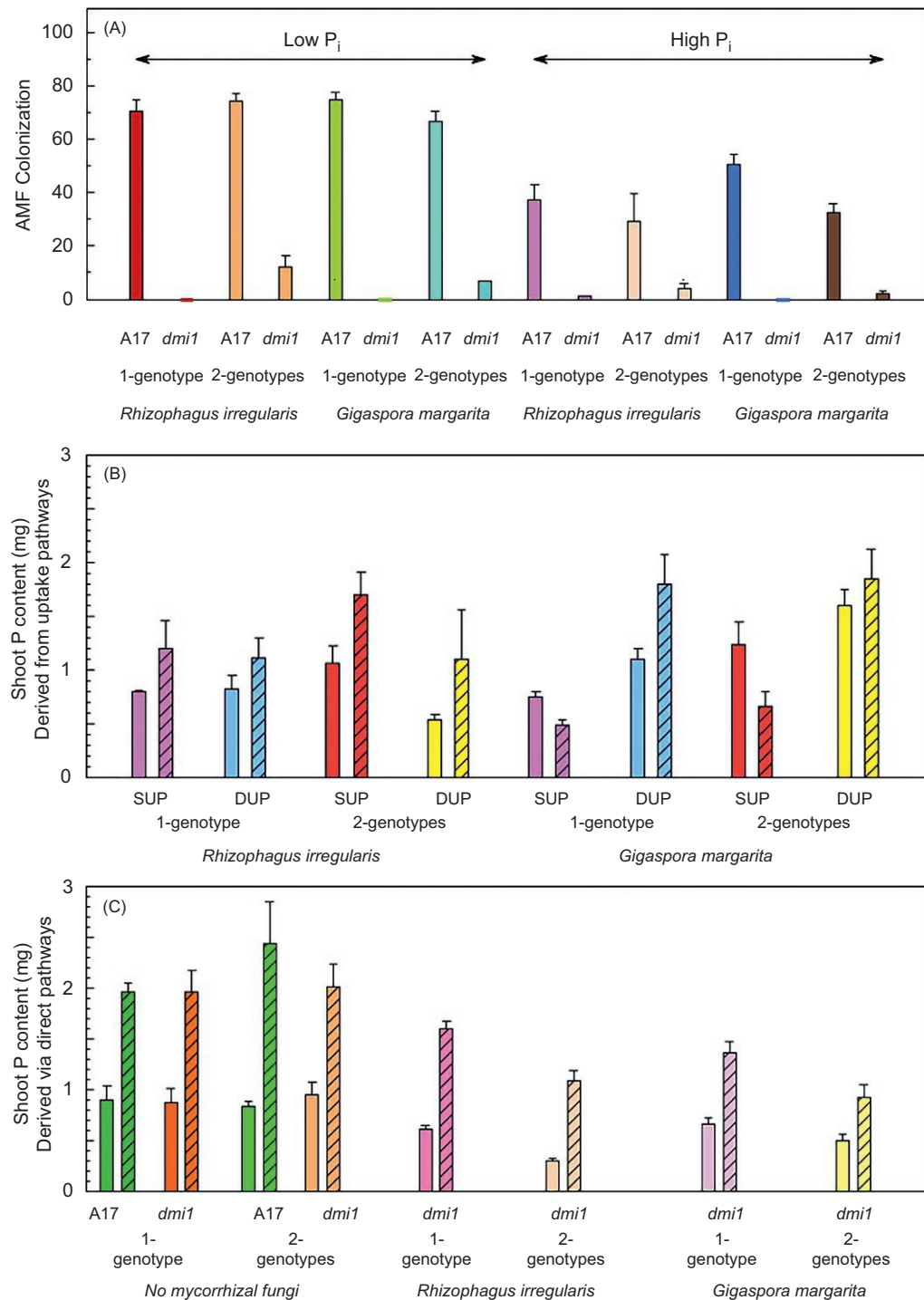


Fig. 6 Effect of soil phosphate (P_i) supply on colonization and uptake of P of two genotypes of the annual medic *Medicago truncatula* (wildtype A17 and the mutant, *dmi1*, that is unable to establish symbioses with rhizobia or AMF) in association with 2 AMF (*Rhizophagus irregularis* or *Gigaspora margarita*). (A) Increased P_i supply reduced the percent root length colonized by either fungal symbiont. Colonization of mutant *dmi1* by AMF was zero or negligible when the plants were grown in single culture (1-genotype) but was somewhat greater when grown in mixed culture with wildtype plants, although very few arbuscules were present. (B) Shoot P content of wildtype A17 derived from the direct uptake pathway (DUP) through roots and that derived from uptake via extraradical hyphae of the AMF (SUP). With *R. irregularis* the SUP pathway consistently provided less P to the shoots from low P_i (plain bars) than high P_i available soils (hatched bars) but the separate supplies via the two pathways were similar, although when the plants were grown in mixed culture with mutant *dmi1*, SUP was greater than DUP. With *G. margarita*, the SUP pathway supplied less P from soil of high P_i compared with low P_i , whereas the supply via DUP was greater from the soil of high P_i . Furthermore, the supply from DUP was always greater than that from SUP. When the two genotypes of *M. truncatula* were grown together, the DUP was similar to the SUP attained with *R. irregularis*. (C) Shoot P content of both genotypes derived from the direct uptake pathway (DUP) through roots. Even in the absence of AMF, DUP for P by the wildtype, A17, was greater when grown in mixed culture with mutant *dmi1* in high P_i soil. Uptake by mutant *dmi1* was similar to that of the wildtype in the absence of AMF in both high and low P_i soils, but in their presence, uptake was reduced when the plants were grown in mixed culture with wildtype and when grown in single culture if the AMF was *G. margarita*. Data from Facelli et al. (2014).

to improve SOM, such as conversion to reduced or no-till systems, maintenance of crop residues on the soil surface, the inclusion of cover crops and use of organic fertilizers, would certainly make a major contribution to improvement in the sustainability of agricultural systems. Such strategies would also enhance the potential role that AMF can play, either in agricultural systems where excessive levels of fertilizer application are a concern for water quality and AMF biodiversity, or in those regions of the world where inputs of nutrients are only moderate, or even too small, and consequently cause concern for soil productivity.

Cropping system

Crop rotation is consistent with the aims of sustainable agriculture in maintaining soil fertility and increasing productivity, restricting buildup of pathogens and weeds, improving crop nutrition and soil structure through fuller exploration by contrasting root systems.

A typical crop rotation involves a repeated sequence of monocot and dicot crops grown in a pure stand, although mixed stands and strip cropping may be used. Ideally, the mixture of plants includes some that are shallower and others that are deeper rooting. Rotations may include crops that are not grown to maturity, such as forage, cover or catch crops. Some systems include a perennial crop with annual crops sown between the individual perennial plants. Variation in the components of the residues from different crops is a significant factor in determining soil properties, with leguminous plants providing a source of nitrogen that can help alleviate yield loss associated with plough-down of grassy pastures. However, some crops, such as brassicas, may be particularly beneficial in inhibiting pathogens but may also have adverse effects on beneficial microbes: being non-mycotrophic they inhibit both the diversity and numbers of individual AMF species, in part because of their inhibitory effects on spore germination.

Soybean and maize in a crop rotation improve the nutrient content of soil primarily by influencing composition of the bacterial community, especially the Gram-positive bacteria. [Dorr de Quadros et al. \(2012\)](#) compared rotations that were a combination of cereals (oat- *Avena strigosa* Schreb, grown in winter and spring-sown maize.) or a simple combination of a legume (*Vicia sativa* L.) grown in winter with spring-sown maize, or a combination of cereal and legume grown in each season (oat and *V. sativa* grown in winter and maize and cowpea (*Vigna unguiculata* L.) in summer. Based on the analysis of 16S rRNA gene sequences, cropping systems had a large effect on microbial community structure. The oat-maize rotation had the greatest diversity of bacteria, based on the Shannon index. However, the rotations that included legumes resulted in soils that were significantly more acidic, which would tend to reduce abundance and diversity ([Fig. 3.](#)).

Improving the potential contribution of AMF to agricultural systems requires crop rotations (including cover crops) to be designed that take account of the key role of plant hosts to maintain species richness and diversity. Legumes are associated with greater diversity than grasses or herbs, with C_4 grasses being more responsive to AMF than C_3 . Alternating cereals and legumes in cropping systems seems to improve the benefits of AMF, which might be because grasses and broadleaved herbaceous plants act as distinct groups of functional host plant, or offset negative feedback mechanisms if grown in monoculture. Such results highlight the potential of particular cropping systems to have contrasting effects on different groups of microorganisms.

Cover crops, particularly combined with no-till, should also be considered as a key element in the management of microorganisms, and can be an important means to improve the diversity of soil microbes in perennial systems, such as orchards. A diversified crop rotation, provided the component species are mycotrophic, can be especially important for the abundance and diversity of AMF. Cover crops of mixed species can be advantageous for these same reasons. Restoration of AMF inoculum potential in soil following non-mycotrophic canola (*Brassica napus* L.), fallows, or lengthy periods made unfavorable for plant growth by drought or cool temperatures, can be achieved using cover crops ([Kabir and Koide, 2002](#)). Importantly, intact hyphae can survive and remain infective after a dry hot summer (Mediterranean climate) or a cold temperate winter. Such use of cover crops might have less scope under rainfed agriculture, where there is insufficient water to grow two crops in a year. Any mycotrophic weeds that germinate at the beginning of the planting season can increase colonization of the crop, its growth and P uptake if they are controlled by herbicides and the crop is seeded without soil disturbance (see [Goss et al., 2017](#)).

Chemical crop protection

The main crop protection chemicals aim to control weeds and pests, the latter focused on fungi, insects and nematodes. Management of weeds, fungal diseases together with the impact of nematodes and insects also involves the risks associated with increased resistance of target species because the chemicals employed lack diversity, including their modes of action. Furthermore, although reduction in target groups, such as fungi or nematodes, would be anticipated, direct adverse effects on other organisms could be of concern if environmental services are threatened. Indirect effects may occur if the removal of target organisms changes the availability of resources to competitors. The direct effects on non-target soil organisms of the range of pesticides, commonly applied in world agriculture, was tabulated by [Bünemann et al. \(2006\)](#) who concluded that these increased from herbicides to insecticides and fungicides. There was considerable variation between chemicals in responses to single or repeated applications, and in the affected non-target species. The herbicide, glyphosate, which is widely used world-wide, modified microbial community structure leading to increases in fungi but reduced the presence of phosphatase enzymes in soil. For bacteria, [Newman et al. \(2016\)](#) found that members of the phyla Proteobacteria, Acidobacteriota, and Actinobacteriota dominated maize and soybean rhizospheres, but application of glyphosate increased the relative abundance of Proteobacteria (particularly in the class, Gammaproteobacteria) but decreased that for Acidobacteriota. The effect of herbicides on AMF colonization is also inconsistent, for example glyphosate has been reported to have no effect (see [Goss et al., 2017](#)). Negative effects of the herbicides bromoxynil and paraquat on AMF colonization have been reported when they were incorporated directly into the soil ([Abd-Alla et al., 2000](#)), but diclofop-methyl, diflufenican and bromoxynil applied to soil had no effects, neither did spraying weeds with paraquat. Regardless

of the impact, soil tillage is the alternative to the use of herbicides for weed control, which can have negative effects on the diversity and abundance of soil microbes, especially AMF.

Neonicotinoids have been the most widely used insecticides in the last decade. However, their impact on non-target insects has been severe, partly because of their persistence within crop plants, which can last for up to 200 days in wheat. Commonly applied as a seed coating, 80% may remain in the soil for up to 3 years with 20% taken into the plant. In a 3-year soybean–maize–soybean rotation on a clay-loam soil, 294 bacterial zOTU were significantly different in abundance between neonicotinoid-treated samples and controls; 68 were more abundant, whereas 226 (>60 genera) had reduced abundance (Parizadeh et al., 2021).

Fungicides are potentially the most deleterious pesticide for AMF. However, when applied at field rates, effects on mycorrhizal development depend on the active compound. Systemic fungicides do not appear to have greater effect on the formation of mycorrhiza than contact types. Fungicides are probably more detrimental to AMF development when applied as a seed coating than if plants, already mycorrhizal, are treated. However, development also seems to depend on the active compound and some have no-effect or even stimulate AMF (Plenchette et al., 2005). When methyl bromide was an acceptable soil fumigant, it was evident that species richness and diversity of AMF could recover from a single fumigation (67% methyl bromide, 33% chloropicrin) by the end of the growing season, even though repeated use could result in the complete absence of root colonization.

Disease suppressive soils

A particular issue for achieving appropriate levels of biological and functional diversity in soil occurs in disease suppressive soils, where genera of particular Phyla can dominate the rhizosphere of crops and suppress root pathogens. One, well-researched system is ‘Take-all suppression’ in soils on which a cereal rotation of wheat and barley (*Hordeum vulgare* L.) or oat (*Avena sativa* L.) is grown continuously for several years. Take-all in wheat is caused by the fungus *Gaeumannomyces tritici*, which also infects barley and some other grasses but not oat. The fungus enters the root system and invades the phloem and xylem of the plant.

Detailed analysis of Take-all shows that within the rhizosphere of wheat are numerous genotypes belonging to the *Pseudomonas fluorescens* complex, which produce the antibiotic phenazine-1-carboxylic acid (PCA) that helps to prevent root invasion by fungi, including *G. tritici* (Kwak and Weller, 2013). The level of root colonization and PCA accumulated after production by these bacteria is inversely related to the amount of rainfall or irrigation water received, whereas the severity of Take-all is less under dry conditions. Biofilm formation is more robust and antibiotic formation is enhanced by the PCA-producing microbes in drier soil. Another antibiotic, also produced by pseudomonads and some other bacteria in the rhizosphere, is 2,4-diacetylphloroglucinol (DAPG), which inhibits viruses, bacteria, fungi, and oomycetes. In addition to its direct action against microbes, DAPG primes the induced systemic resistance mechanism of plants, which is an enhanced defensive capability that acts against a broad spectrum of pathogens. The DAPG-producers are particularly significant in the suppression of soilborne pathogens in monoculture, for example, pea (*Pisum sativum* L.) and small-grained cereals. Under cereal monoculture, DAPG-producers accumulate within the rhizosphere until they reach 10^5 CFU g⁻¹ root, sufficient to suppress the invasive actions of *G. tritici* and ‘Take-all decline’ begins. The complex actions of DAPG appear to be sufficient to prevent *G. tritici* from developing resistance to the antibiotic (Kwak and Weller, 2013). It is clear that in these soils, during monocropping, the microbiome is not as biologically diverse as in mature natural ecosystems.

Approaches to manage soil microbes in farming systems

There is evidence that key aspects of agricultural field management affect the soil microbial population non-randomly and modify the potential for plant growth. The addition of different plants has a marked effect on the rhizosphere population because of the selection and recruitment by the introduced plant roots. A rational use of organic and inorganic nutrients needed to maintain soil reserves is compatible with maintaining an abundant and diverse microbial population as long as other elements of the cropping systems, including reduced or no-till, are also included. Incorporating a diversified cropping system can reduce the load of individual plant pathogens and adds new root systems that modify the microbial population of the rhizosphere. Differences in quality and quantity of organic residues returned to the soil from the various crops in a rotation and inclusion of organic nutrient sources help to enhance the soil organic content that reduced tillage can help maintain. This in turn can improve soil structure and structural resilience that will increase the capacity for exchange of nutrients and improve their efficient utilization.

Examples of managing AMF

Local indigenous communities of AMF are usually quite diversified, certainly well adapted, definitely cheaper and more efficient than commercial inoculum and form mutualist symbiosis with most cultivated plants. Furthermore, functional complementarity occurs within the AMF community colonizing a single root system (Jansa et al., 2008), and a more diverse population expands the opportunity of beneficial outputs. Practical examples of their management in agronomic systems have been described (Goss et al., 2017).

Prerequisites for managing AMF

Despite the complex relationships between crop production techniques, environmental factors and microbial diversity, there are clearly identified traits in the agronomic system that affect microbial performance. These include crop sequence, tillage regime or

fertilizers and manure amendments (Table 4, Fig. 4) that can be managed to define a strategy to promote beneficial microbial functioning. Identification of major concerns in a particular field, such as a prevailing stress for plant growth and nutrient supply, together with soil characteristics and cropping history, should be considered in defining appropriate strategies. Although current understanding is limited and the underpinning research insufficient, the management of microbial functioning in agricultural conditions should be considered in the context of sustainable agronomic practices. It must be consistent with natural ecological processes that include resilience of the system, more efficient use of inputs, and supporting long-term productivity.

Considering the issues associated with inoculation, a key aspect is to ensure colonization of the target crop by members of the indigenous AMF community. Potential sources are spores, colonized root fragments and ERM. Of these inoculum sources, ERM, when intact, commonly constitutes the most infective propagule by providing an earlier and faster AMF colonization, as documented both in pot experiments and agricultural ecosystems (Goss et al., 2017). Colonization started by an intact ERM optimizes the potential benefits from AMF to the host plant because it can capitalize on them sooner after exposure. This property of intact ERM is particularly important when crops are challenged by biotic or abiotic stresses existing in the soil at sowing, as the extent of AMF colonization at the time the host plant meets the stressor agent is directly related to the level of bioprotection achieved.

Given the preferential association between certain AMF and host plants, together with the characteristics of intact ERM as preferential inoculum, identifying a plant sequence and adopting reduced or no-till techniques are key components of a strategy that can ensure a consistent and predictable manipulation of the indigenous inoculum. It allows capitalization on the manifold benefits potentially provided by AMF and their functional diversity.

A possible strategy for the deliberate manipulation of the indigenous AMF population and its functional diversity consists of.

- 1 the intentional use of selected host plants to establish an extensive extraradical mycelium (Developer-plants) involving a consortium of the AMF,
- 2 keeping this ERM intact by the adoption of appropriate non-inversion tillage techniques,
- 3 support this preferential source of AMF inoculum to establish earlier and faster colonization of the following plant.

With this approach, effective protection has been achieved against abiotic (Mn soil toxicity) and biotic (*Fusarium oxysporum* or *Magnaportheopsis maydis* root pathogens) stresses in different host plants. Knowledge of the preferential association between AMF and a particular host plant (as Developer-plant) means that this strategy can also be used to manage AMF functional diversity (Brito et al., 2021). It can easily be applied at field scale in both low- and high-input cropping systems. The strategy may require adjustments to the cropping system, such as employing no-till or reduced tillage, selecting crop rotations or cover crops, that are easy to adopt and realistically can be implemented at a field level. Irrespective of other aspects needed to be considered (e.g., soil type and pH), choosing the Developer-plant involves specific criteria: for abiotic stress plants need to be tolerant of the stress, and representatives of the natural vegetation at the site provide the best opportunities for recruitment; for biotic stress the Developer-plant should not host the pathogen.

Abiotic stress

The ability of intact ERM to promote earlier and faster colonization and the efficacy of these strategies to protect wheat and subterranean clover plants against Mn toxicity was demonstrated in a two-phase pot experiment (Goss et al., 2017). Manganese toxicity resulted from the soil parent material and low pH. Unsterilized field soil was used, relying on a diverse indigenous population of AMF adapted to the conditions. The selected Developer-plants were annual ryegrass (*Lolium rigidum* L.) and yellow serradella (*Ornithopus compressus* L.), a legume, both highly mycotrophic plants. These plants were grown separately to develop the ERM. After 7 weeks, growth was stopped chemically with Glyphosate. The soil was undisturbed in half the pots of each developer treatment, the ERM was kept intact. In the remaining pots, the soil was disturbed and ERM disrupted by passing through a 4-mm sieve before being replaced in the same pots to act as a negative control. Two weeks later, an Mn susceptible crop, wheat (*Triticum aestivum* L.), was planted in half the disturbed and undisturbed pots of each Developer-plant treatment, and in the other pots subterranean clover was planted. Mycorrhizal colonization started earlier and developed faster, both in wheat and subterranean clover grown in undisturbed soil. Consequently, there was significant improvement in plant growth. There was no ability of AMF to protect plants against Mn toxicity in disturbed soil where intact ERM was absent.

When the AMF consortium forming mycorrhizal roots of a host plant has been formed from spores indigenous to the soil, each plant influences the choice of AMF in the assemblage (see Fig. 1. Note the similarity between consortia in wheat and annual ryegrass, and in yellow serradella and subterranean clover), and is usually consistent at the family level of plant taxa. However, when the consortium is formed from intact ERM associated with the Developer-plant, the AMF community present in the target plant shows a close relationship with the consortium in that plant (i.e., with that of the first plant of the succession, irrespective of its botanical group (Fig. 7, van Tuinen et al., 2017)). There is some evidence that the biochemical recognition dialogue between AMF and the host plant, which leads to a functional mycorrhiza, does not follow the same pattern but depends on whether the inoculum source is based on spores or intact ERM, as the plant symbiotic program depends to some extent on the colonizing propagule type. Gene expression by wheat was differentially activated according to the preceding Developer. The assemblage developed by yellow serradella activated processes in wheat related to cell division and growth, with only a few related to stress. However, when annual ryegrass was the Developer-plant, the wheat genes induced were those related to oxidative stress, disease protection and metal binding. It suggests that wheat genes activated by the consortium from annual ryegrass were less effective in protecting the plant from Mn toxicity than those activated by AMF from yellow serradella, which had a more positive effect on the shoot dry weight of wheat.

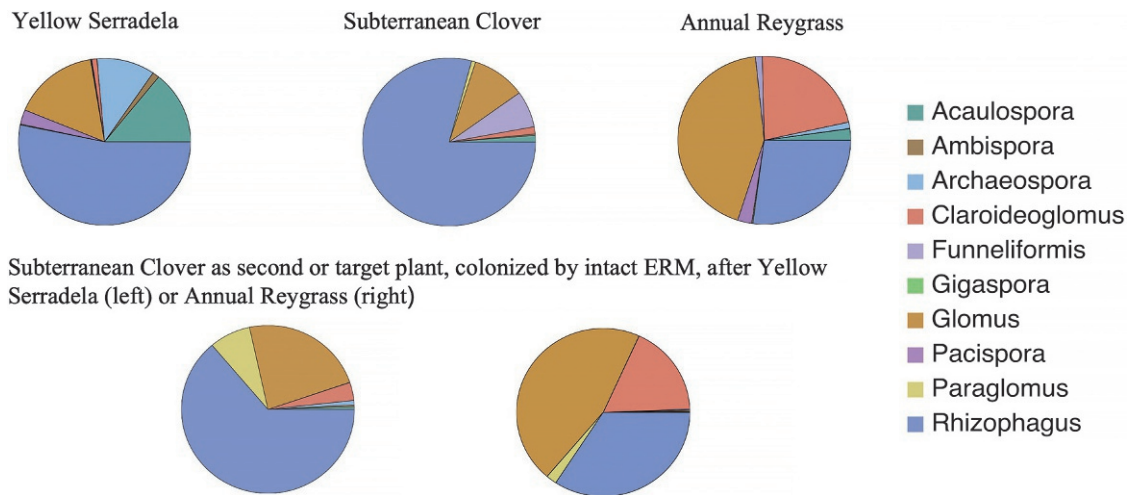


Fig. 7 Changes in taxonomic affiliation of the operational taxonomic units (OTUs) of AMF in a target or second plant in a cropping sequence, following colonization initiated by intact extraradical mycelium (ERM), are greatly influenced by the ERM-Developer involved. There are clear differences between the OTUs colonizing yellow serradella, subterranean clover and annual ryegrass grown following soil disturbance, with spores being the major form of propagule (top row). However, when yellow serradella (Fabaceae) or annual ryegrass (Poaceae) act as Developer-plants, and subterranean clover (Fabaceae) is the target or second plant, with intact ERM as the main propagule the AMF community present in the second plant shows a close similarity to that of the preceding Developer-plant, irrespective of its botanical group. Based on van Tuinen D, Brígido C, Brito I, Goss MJ, Alho L, and Carvalho M (2017) Cover plants, a tool to manage arbuscular mycorrhizal fungi diversity? 9th International Conference on Mycorrhiza. 30th July– 4th August 2017. Prague, Czech Republic. P(ID 384), pp 88.

Biotic stress

The effectiveness of AMF colonization in bio-protection has also been reported in relation to biotic stresses. Its success is frequently associated with AMF inoculation before the presence of a pathogen infection. Such observations indicate the need for well-established colonization prior to contact with the stressor agent. This apparent limitation can be overcome if AMF colonization begins earlier, which occurs when it is initiated by an intact ERM. Warm temperatures together with irrigation make spring or summer crops particularly vulnerable to biotic stresses, especially when monoculture is practiced over long periods. Under these circumstances the bio-protective role of AMF can also be brought to the cropping system by adjusting management practices. A judicious choice of over-winter cover crop, which can also function as a developer for ERM and maintain ERM integrity using appropriate land preparation, can be combined in a realistic bio-protection approach. Such an approach has already been tested in tomato (*Solanum lycopersicum* L.) and maize.

For field-grown tomato at risk from *Fusarium oxysporum* wilt disease, the modified production system adopted barley as the cover crop and an Developer-plant. Intensive deep tillage in spring was moved to the immediate post-harvest period and planting beds prepared, after which barley was sown. In spring, the Developer-plants were killed with systemic herbicide, planting beds had minor adjustments, and the tomato crop planted. The combination of cover crop and shallow soil disturbance resulted in significantly greater arbuscular colonization of tomato and less plant mortality from fusarium wilt, compared with a control area where standard practices were retained. Tomato production increased by 20 t ha⁻¹ - a 20% increase in total fruit production. In addition, an increase in general indicators of soil microbial activity (phosphatase, β -glucosidase, arylsulfatase, and urease activity) were also observed, underlining the beneficial effect of these practices on the general soil microbial population.

The same strategy was tested under field conditions for the bioprotection of maize against the soilborne fungus *Magnaportheopsis maydis* (was *Cephalosporium maydis*). The Developer-plant selected was Italian ryegrass (*Lolium multiflorum* L.). Tillage by moldboard ploughing and disking was compared with no-till. The presence of the soilborne fungus in maize was assessed by quantitative polymerase chain reaction. Only when a cover crop was grown and no-till adopted was there a significant reduction in *M. maydis* in the maize crop, which also resulted in greater dry matter production and a 19% increase in grain production (Patanita et al., 2020).

These examples demonstrate the ability to manage the indigenous microbiology and to use it to achieve a biological solution to soilborne diseases, for which no chemical option was available.

Complex interactions – the case of Mn toxicity in legumes

Subterranean clover shows little evidence of susceptibility to Mn toxicity when its N is obtained from mineral-N sources. However, the effects of Mn toxicity develop when plants become dependent on N₂-fixation in root nodules by the symbiotic diazotroph *Rhizobium leguminosarum* bv. *trifolii*. Hence, managing AMF to protect subterranean clover against the abiotic stress posed by Mn toxicity probably depends on the response of nodule bacteroids, or on both them and the plant. Using intact ERM as the preferential source of inoculum, Mn concentration in the roots was less, shoot and nodule dry weight, and N₂-fixation were greater than when other propagules were used. This is consistent with the hypothesis that better growth of the plant depended on protecting the rhizobia in the root nodules from the adverse effects of Mn. In this case, the interactions among participants in the tripartite

symbiosis have significantly increased the proportion of Mn uptake that was distributed to the shoots, maintaining the tissue concentration there, but reducing that in root tissue to overcome the inhibition of N₂-fixation. The efficacy of protection by AMF depended on the timing of early colonization events by both symbiotic microbes due to their use of common transduction pathways. Against this abiotic stress, there was little difference in effectiveness between the natural soil spore content and a combination of the spores and infected root fragments (Goss et al., 2017).

Conclusions

Soil contains a large number of microbes, with bacteria, archaea, algae, fungi, protozoa and small-size animals such as nematodes contributing the majority of individuals. These organisms participate in a wide array of processes in soil that affect the growth and development of the plants and animals that inhabit each locality, as well as contributing to the major geochemical cycles.

Given the numbers and wide range of soil organisms there are clear challenges to any attempt at directing their collective activities. Furthermore, their importance for soil and ecosystem functioning makes manipulation a potentially risky endeavor. Approaches to changing the composition of the microbe population are limited and either involve the addition of an inoculum of specific microbes taken from a different habitat, or modifying conditions such that particular components of the soil microbiome are stimulated. There is uncertainty in the effects, safety and reliability of inoculation as a general approach for extensive application, except where the local indigenous population has been lost. The evidence points to adopting agronomic practices as the main tools of manipulation. Tillage can be detrimental to balanced ecosystems but can be used to reset the system. Appropriate cropping systems can be used to attract and enhance specific groups of organisms into the rhizosphere where an effective rhizobiome can develop. In addition, the ERM of mycorrhizal fungi also harbor a distinct mycorrhizosphere community that are important in the functioning of the mycorrhiza. Thus, the development of a strategy for managing AMF can also influence other microbial components associated with the plants linked in a common ERM network. The strategy described for integrating bio-protection and other beneficial roles of native AMF into cropping systems, uses a Developer-plant for establishing ERM, which can be derived from numerous species. That ERM acts as the immediately available propagule for the colonization of crop plants that may be susceptible to biotic or abiotic stresses in the soil. Implementing such a strategy is versatile, requiring selection of the Developer-plant within the cropping system, potentially including common weed species, and ensuring, by adopting appropriate tillage systems, that the ERM developed is kept intact prior to the planting of the target crop. Increased knowledge of the functioning of rhizo- and mycorrhizosphere organisms is essential to future developments in soil microbe management.

References

- Abd-Alla MH, Omar SA, and Karanxha S (2000) The impact of pesticides on arbuscular mycorrhizal and nitrogen-fixing symbioses in legumes. *Applied Soil Ecology* 14: 191–200. [https://doi.org/10.1016/S0929-1393\(00\)00056-1](https://doi.org/10.1016/S0929-1393(00)00056-1).
- Addy HD, Miller MH, and Peterson RL (1997) Infectivity of the propagules associated with extraradical mycelia of two AM fungi following winter freezing. *New Phytologist* 135: 745–753. <https://doi.org/10.1046/j.1469-8137.1997.00707.x>.
- Brito I, Carvalho M, and Goss MJ (2021) Managing the functional diversity of arbuscular mycorrhizal fungi for the sustainable intensification of crop production. *Plants, People, Planet* 1–15. <https://doi.org/10.1002/ppp3.10212>.
- Bünemann EK, Schwenke GD, and Van Zwieten L (2006) Impact of agricultural inputs on soil organisms—a review. *Australian Journal of Soil Research* 44: 379–406. <https://doi.org/10.1071/SR05125>.
- Chinnadurai C, Gopalaswamy G, and Balachandrar D (2014) Impact of long-term organic and inorganic nutrient managements on the biological properties and eubacterial community diversity of the Indian semi-arid Alfisol. *Archives of Agronomy and Soil Science* 60: 531–548. <https://doi.org/10.1080/03650340.2013.803072>.
- Chowdhury S, Lange M, Malik AA, Goodall T, Huang J, Griffiths RI, Gleixner G, and G. (2022) Plants with arbuscular mycorrhizal fungi efficiently acquire nitrogen from substrate additions by shaping the decomposer community composition and their net plant carbon demand. *Plant and Soil* 475: 473–490. <https://doi.org/10.1007/s11104-022-05380-x>.
- Cornell C, Kokkoris V, Richards A, Horst C, Rosa D, Bennett JA, and Hart MM (2021) Do bioinoculants affect resident microbial communities? A meta-analysis. *Frontiers in Agronomy* 3: 753474. <https://doi.org/10.3389/fagro.2021.753474>.
- Covacevich F, Echeverría HE, and Aguirrezabal LAN (2007) Soil available phosphorus status determines indigenous mycorrhizal colonization of field and glasshouse-grown spring wheat from Argentina. *Applied Soil Ecology* 35: 1–9. <https://doi.org/10.1016/j.apsoil.2006.06.001>.
- Deng Y, Feng G, Chen XP, and Zou CQ (2017) Arbuscular mycorrhizal fungal colonization is considerable at optimal Olsen-P levels for maximized yields in an intensive wheat-maize cropping system. *Field Crops Research* 209: 1–9. <https://doi.org/10.1016/j.fcr.2017.04.004>.
- Dorr de Quadros P, Zhalnina K, Davis-Richardson A, Fagen JR, Drew J, Bayer C, Camargo FAO, and Triplett EW (2012) The effect of tillage system and crop rotation on soil microbial diversity and composition in a subtropical acrisol. *Diversity* 4: 375–395. <https://doi.org/10.3390/d4040375>.
- Facelli E, Duan T, Smith SE, Christophersen HM, Facelli JM, and Smith FA (2014) Opening the black box: Outcomes of interactions between arbuscular mycorrhizal (AM) and non-host genotypes of Medicago depend on fungal identity, interplay between P uptake pathways and external P supply. *Plant, Cell & Environment* 37: 1382–1392. <https://doi.org/10.1111/pce.12237>.
- Goss MJ, de Varennes A, Smith PS, and Ferguson JA (2002) N₂ fixation by soybeans grown with different levels of mineral nitrogen, and the fertilizer replacement value for a following crop. *Canadian Journal of Soil Science* 82: 139–145. <https://doi.org/10.4141/S01-003>.
- Goss MJ, Carvalho M, and Brito I (2017) *Functional Diversity of Mycorrhiza and Sustainable Agriculture: Management to Overcome Biotic and Abiotic Stresses*, 1st edn Academic Press/Elsevier. eBook ISBN: 9780128042861 -Paperback ISBN: 9780128042441.
- Jansa J, Smith FA, and Smith SE (2008) Are there benefits of simultaneous root colonization by different arbuscular mycorrhizal fungi? *New Phytologist* 177: 779–789. <https://doi.org/10.1111/j.1469-8137.2007.02294.x>.
- Jiang F, Zhang L, Zhou J, George TS, and Feng G (2021) Arbuscular mycorrhizal fungi enhance mineralisation of organic phosphorus by carrying bacteria along their extraradical hyphae. *New Phytologist* 230: 304–315. <https://doi.org/10.1111/nph.17081>.
- Kabir Z and Koide RT (2002) Effect of autumn and winter mycorrhizal cover crops on soil properties, nutrient uptake and yield of sweet corn in Pennsylvania, USA. *Plant and Soil* 238: 205–215. <https://doi.org/10.1023/A:1014408723664>.

- Kwak Y-S and Weller DM (2013) Take-all of wheat and natural disease suppression: A review. *Plant Pathology Journal* 29: 125–135. <https://doi.org/10.5423/PPJ.SI.07.2012.0112>.
- Lang M, Zhang C, Su W, Chen X, Zou C, and Chen X (2022) Long-term P fertilization significantly altered the diversity, composition and mycorrhizal traits of arbuscular mycorrhizal fungal communities in a wheat-maize rotation. *Applied Soil Ecology* 170(2022), 104261. <https://doi.org/10.1016/j.apsoil.2021.104261>.
- Lauber CL, Hamady M, Knight R, and Fierer N (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Applied and Environmental Microbiology* 75: 5111–5120. <https://doi.org/10.1128/AEM.00335-09>.
- Lopes MJS, Dias-Filho MB, and Gurgel ESC (2021) Successful plant growth-promoting microbes: Inoculation methods and abiotic factors. *Frontiers in Sustainable Food Systems* 5: 606454. <https://doi.org/10.3389/fsufs.2021.606454>.
- Newman MM, Hollett N, Lorenz N, Dick RP, Liles MR, Ramsier C, and Kloepper JW (2016) Glyphosate effects on soil rhizosphere-associated bacterial communities. *Science of the Total Environment* 543: 155–160. <https://doi.org/10.1016/j.scitotenv.2015.11.008>.
- O'Callaghan M, Ballard RA, and Wright D (2022) Soil microbial inoculants for sustainable agriculture: Limitations and opportunities. *Soil Use and Management* 38: 1340–1369. <https://doi.org/10.1111/sum.12811>.
- Oehl F, Ewald Sieverding E, Ineichen K, Mäder P, Boller T, and Wiemken A (2003) Impact of land use intensity on the species diversity of arbuscular mycorrhizal fungi in agroecosystems of Central Europe. *Applied and Environmental Microbiology* 69: 2816–2824. <https://doi.org/10.1128/AEM.69.5.2816-2824.2003>.
- Parizadeh M, Mimee B, and Kembel SW (2021) Neonicotinoid seed treatments have significant non-target effects on phyllosphere and soil bacterial communities. *Frontiers in Microbiology* 11: 619827. <https://doi.org/10.3389/fmicb.2020.619827>.
- Patanita M, Campos MD, Félix MR, Carvalho M, and Brito I (2020) Effect of tillage system and cover crop on maize mycorrhization and presence of *Magnaportheopsis maydis*. *Biology* 9: 46. <https://doi.org/10.3390/biology9030046>.
- Plenchette C, Clermont-Dauphin C, Meynard JM, and Fortin JA (2005) Managing arbuscular mycorrhizal fungi in cropping systems. *Canadian Journal of Plant Science* 85: 31–40. <https://doi.org/10.4141/P03-159>.
- Raymond NS, Gómez-Muñoz B, van der Born FJT, Nybroe O, Jensen LS, Müller-Stöver DS, Oberson A, and Richardson AE (2021) Phosphate-solubilising microorganisms for improved crop productivity: A critical assessment. *New Phytologist* 229: 1268–1277. <https://doi.org/10.1111/nph.16924>.
- Säle V, Aguilera P, Laczko E, Mäder P, Berner A, Zihlmann U, van der Heijden MGA, and Fritz Oehl F (2015) Impact of conservation tillage and organic farming on the diversity of arbuscular mycorrhizal fungi. *Soil Biology & Biochemistry* 84(2015): 38–52. <https://doi.org/10.1016/j.soilbio.2015.02.005>.
- Toljander JF, Santos-González JC, Tehler A, and Finlay RD (2008) Community analysis of arbuscular mycorrhizal fungi and bacteria in the maize mycorrhizosphere in a long-term fertilization trial. *FEMS Microbiology Ecology* 65: 323–338. <https://doi.org/10.1111/j.1574-6941.2008.00512.x>.
- Tubeileh AM and Goss MJ (2022) Assessing the effects of using animal manure on soil health. In: Horwath R (ed.) *Improving Soil Health*, pp. 281–308. Cambridge, UK: Burleigh Dodds Science Publishing.
- van Tuinen D, Brígido C, Brito I, Goss MJ, Alho L, and Carvalho M (2017) Cover plants, a tool to manage arbuscular mycorrhizal fungi diversity? In: *9th International Conference on Mycorrhiza. 30th July– 4th August 2017. Prague, Czech Republic*. P(ID 384), p. 88.
- Wubs ERJ, van der Putten WH, Bosch M, and Bezemer TM (2016) Soil inoculation steers restoration of terrestrial ecosystems. *Nature Plants* 2: 16107. <https://doi.org/10.1038/NPLANTS.2016.107>.
- Zeffa DM, Fantin LH, Koltun A, de Oliveira ALM, Nunes MPBA, Canteri MG, and Gonçalves LSA (2020) Effects of plant growth-promoting rhizobacteria on co-inoculation with *Bradyrhizobium* in soybean crop: A meta-analysis of studies from 1987 to 2018. *PeerJ* 8: e7905. <https://doi.org/10.7717/peerj.7905>.
- Zhou Z, Wang C, Zheng M, Jiang L, and Luo Y (2017) Patterns and mechanisms of responses by soil microbial communities to nitrogen addition. *Soil Biology and Biochemistry* 115: 433–441. <https://doi.org/10.1016/j.soilbio.2017.09.015>.
- Zuber SM and Villamil MB (2016) Meta-analysis approach to assess effect of tillage on microbial biomass and enzyme activities. *Soil Biology and Biochemistry* 97: 176–187. <https://doi.org/10.1016/j.soilbio.2016.03.011>.

Trees, hedges, agroforestry and microbial diversity

Josiah B Judson^a, Joseph Holden^a, Pippa Chapman^a, and Marcelo Valadares Galdos^b, ^aSchool of Geography, University of Leeds, Leeds, United Kingdom; ^bRothamsted Research, Sustainable Soils and Crops, Harpenden, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

This is an update of P.K.R. Nair, AGROFORESTRY, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 35–44, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00244-7>.

Introduction	469
Soil microbes and system complexity	470
Trees, shrubs, and soil biodiversity	472
Hedges	473
Agroforestry	473
Summary	478
References	478

Abstract

Soil microbial communities are critical for provision of above-ground ecosystem services, supporting biodiversity and regulating material exchange between earth and atmosphere. What is less evident is the importance of trees and shrubs in sustaining soil microorganisms. We consider system interdependencies between agricultural soils, plants and microbes, demonstrating negative effects of historic removals of woody vegetation from the landscape for the sustainability of soil microbial diversity. In contrast, reintroduction of shrubs and trees to agricultural systems, including agroforestry and hedgerow planting, has potential to restore soil functioning and create space for nature while offering many other benefits such as economic diversification and increased climate resilience.

Key points

- Trees, hedges and agroforestry play a critical role in sustaining soil biodiversity and terrestrial ecosystem services
- Removal of trees and hedges, under pressure from intensive agriculture and industry, has negative consequences for soil functioning
- As multifunctional landscape features, hedges and agroforestry create space for nature above and below ground while allowing space for economic use of land
- Although research has considered ecosystem service changes in large and complex landscapes, there is much less focus on small-scale aggregated benefits, synergies and trade-offs impacting soil functioning and practitioner decision-making

Introduction

Soil microbial communities contain a substantial majority of Earth's biodiversity. This abundant and largely hidden group of organisms is critical for both soil system functioning and the delivery of a wide range of above-ground ecosystem benefits. Soil microorganisms support life both above and below ground, providing energy, timely nutrient supply and a stable growth medium for plants, among many other ecological benefits.

The importance of well-functioning soil systems for the growth of plants is intuitive. What is less apparent, however, is the reverse: the critical role that trees, hedges and their roots play in supporting microbial diversity. In doing so, vegetation in the landscape indirectly sustains many of the benefits attributable to microorganisms. Understanding this interactive relationship requires a system-level view of the interdependencies between soil, microbial communities and vegetation. This chapter considers how the presence (and removal) of trees, hedges, and agroforestry influences microbial diversity, and the potential for the reintroduction of perennial woody vegetation to revitalize soil function, and simultaneously bring systems of economic production on land back in step with biosphere interdependencies. We first consider interdependencies which exist between vegetation and soil microbial communities, focusing on trees and hedges and the consequences of their removal for whole system functioning. Then we consider how hedges and agroforestry, as intentional and multifunctional landscape features, may support and restore the health of below-ground systems. This may be direct, via rhizosphere effects, and indirect, through effects on hydrology and microclimate. We conclude by summarizing the potential for planting strategies to generate change for soil systems and priorities for further research.

Soil microbes and system complexity

Microbial communities are integral to soil function, responsible for mineralization and cycling of nutrients within soil, as well as controlling both formation and decomposition of soil organic carbon (SOC). Abiotically, microbes contribute to soil structure and formation of aggregates, and the formation and transformation of organic matter (OM) (Prescott et al., 2019). Soil supporting services such as nutrient availability and water holding capacity are thus substantially mediated by microbial activity (Dollinger and Jose, 2018). Plants receive material benefits from microbial communities such as timely nutrient supply and reliable soil water access. Associations with mycorrhizal fungal networks improve nutrient supply and resilience against drought by increasing root-hyphal surface area. These networks protect against pathogens both physically and by production of signaling compounds that can induce a coordinated defense response in the plant community (Prescott et al., 2019). The centrality of these interactions means that soils are widely considered to be degraded once microbial activity is suppressed, unable to offer supporting ecosystem services vital for life.

Organic matter drives the soil food web, and the majority of soil microbes rely on carbon as a source of energy. Approximately 1500 Gt of the organic matter stored in the top meter of global soil is carbon, and this chapter will focus on this critically significant component of the terrestrial ecosystem. SOC is statistically correlated with other indicators of soil function, particularly aggregation, soil formation, and water holding capacity (Shukla et al., 2006). SOC is a significant reservoir in the context of sustainability more widely, both for soil fertility and maintenance of microbial diversity, but also wider public goods such as resilience of food and climate systems, resistance to desertification and protection of global biodiversity both above and below ground. Yet since the onset of ancient agriculture and widespread land-use change for cultivation, large quantities of C have been lost from the world's soils. SOC losses have received particular attention due to implications for climate and impacts on soil fertility for agriculture. It is estimated that 133 Gt C has been released to the atmosphere as carbon dioxide (CO₂) through land-use change with a particular acceleration in the last 200 years (Sanderman et al., 2017). Land-use change has produced wider soil impacts, including greater propensity toward drought, salinization, acidification, compaction, nutrient depletion, and soil losses through erosion. Impacts on soil structure have arisen both through SOC losses and physical disturbance: practices such as frequent tillage or intensive grazing degrade structural elements of the soil, including physical alteration of soil aggregates and collapse of macropores, resulting in compaction. Disturbance alters soil habitat for both smaller and larger soil organisms, eventually diminishing below-ground functional diversity and the ecosystem services which are sustained by soil. Plant-mycorrhizal symbioses are consequently affected, and soil microbial respiration rates increase once SOC stabilized within soil aggregates or as mineral-associated organic matter is exposed to decomposition by organisms.

Negative outcomes of land-use change collectively impact soil microbial communities. However, individual negative outcomes are not produced in isolation, and significant interdependencies exist within the terrestrial ecosystem (Fig. 1). For example, reduction of SOC stocks diminishes available energy for organisms; as soil microbial biomass is consequently diminished, the ability to provide supporting services to vegetation is weakened (Fig. 1, R1). Yet removal of vegetation, or reliance on external sources of nutrition, in turn impacts microbial communities. SOC stocks are commonly in a dynamic equilibrium between removal via metabolic respiration by soil organisms (as CO₂) and inputs such as plant litter, root material and C-rich exudates produced by trees and other plants to regulate the soil environment in the vicinity of their roots. Forests, trees, shrubs, and grasslands are all examples of perennial vegetation which return litter and exudates to soil in this way. Once they are removed, the rate at which C is photosynthetically fixed and transferred to the subsurface is likely to be diminished, reducing available energy for soil microorganisms. Annual crops planted for agriculture in place of perennial vegetation fix and transfer some C to the soil, yet the process is significantly weakened in comparison. Wheat roots, which cease growth by early June in the northern hemisphere, transfer less than half the C to soil annually compared with perennial grass-clover ley. Trees, shrubs, and forests return considerably more even than perennial grasses due to their deeper roots and more abundant litter. Thus, removal of crop biomass at harvest, frequent tillage and lack of biological diversity in agricultural landscapes all reduce net C fluxes into soil systems. Reduction in C inputs to soil is a problem because negative plant-soil feedbacks are generated, whereby disrupted C supply to soil microorganisms diminishes their ability to make nutrients available to plants, which in turn inhibits growth and further C addition. Similarly, exudation benefitting soil microorganisms is energetically expensive for plants, such that they are potentially less likely to produce exudates in exchange for nutrition if nutrition instead comes from synthetic inputs. Studies have demonstrated the negative feedbacks that management choices such as tillage and agrochemical use have on plant-soil feedbacks (Menalled et al., 2020).

Further complex feedbacks arise elsewhere in the soil system. Both reduction in SOC stocks and compaction resulting from soil disturbance decrease soil water holding capacity. More compact soil has less available pore space, such that water infiltration can be reduced, and surface runoff increased, resulting in greater erosion risk and removal of C-rich habitat from the top soil. Meanwhile, removal of trees and shrubs, or other vegetation, leads to physical exposure of soil, making it vulnerable to disturbance. Disturbance, in turn, exposes SOC to decomposition by microorganisms, depleting C stocks and releasing CO₂ to the atmosphere.

What is evident from these interactions is that no component of the system can be considered in isolation. Disruption to one component produces unintended consequences elsewhere. Consequences may be reinforcing (R) or balancing (B), as shown in Fig. 1. For reinforcing feedbacks, gain or loss of system function will be exponential, whereas changes to balancing feedbacks set new equilibrium points for the system (Meadows, 1999). Within undisturbed forest soils, interactions in the soil system are remarkably circular—leaching beneath forests is very small in comparison to large internal fluxes within the plant-soil system itself (Prescott et al., 2019).

Trees, shrubs, and soil biodiversity

Interactions between plants and soil microorganisms in natural ecosystems can be mutually beneficial, harmful or simply neutral (Lynch, 1990). Plant species produce exudates from fine roots, with which they regulate the soil environment surrounding roots and microbial dynamics within it. The rhizosphere (Fig. 2), in its earliest definition, refers to the volume of soil surrounding living roots which is influenced by root activity (Hiltner, 1904; Haichar et al., 2014), into which between 5% and 21% of photosynthetically fixed C is transferred through root exudates (Whipps, 1990). Rhizodeposition is a broader term incorporating not just exudation but also deposits of root hairs, fine branches, border cells, and other cells from the root cap and epidermis into soil surrounding plant roots. It enables the plant to cope with herbivores and parasites, changes the physical and chemical properties of soil and inhibits the growth of competing plant species (Haichar et al., 2014) (Fig. 2). Exudates include amino acids, organic acids, sugars, and other secondary metabolites which are released into the soil through passive diffusion from roots.

A steep concentration gradient of these exudates exists as they are swiftly consumed by microbes in the vicinity of roots. Secreting C-dense exudates is costly to the plant in terms of energy, yet in return microbes make soil nutrients accessible to plants in a timely manner for growth. Exudate composition is plant species-specific, allowing plants to attract and recruit a range of beneficial microbes from bulk soil in addition to modification of soil structure and resisting pathogens. Exudates change according to the growth stage of plants, and plant-soil interactions are active, in that plants recognize and actively respond to different microorganisms within the rhizosphere. Beneficial interactions occur with rhizobia, mycorrhiza and rhizobacteria, including protection for plants against both biotic and abiotic stresses. Negative interactions occur with parasitic plants, pathogenic microbes, and invertebrates (Fig. 2) (Haichar et al., 2014). The consequence of these interactions is that the rhizosphere has distinctively different chemical, physical and biological characteristics to soil outside the rhizosphere.

The complexity and diversity of interactions within the rhizosphere illustrates the importance of woody vegetation and roots for the functioning of soil systems. Although globally the C stock of terrestrial vegetation is smaller than that of soils, vegetation is nonetheless critical in regulating exchange of material between soil and atmosphere. Trees and woody vegetation regulate soil microbial communities through their control on exchange of C, water and other nutrients between pedosphere and atmosphere, and abiotic management of soil structure. Above-ground biodiversity interacts with below-ground habitat to promote soil functioning. Although the precise mechanisms of material exchange remain are yet to be fully characterized, it may be presumed that when ecological systems are diverse, the rhizosphere is permitted to function as normal, sustaining feedbacks which lead to a thriving and resilient soil system capable of delivering multiple benefits (Aislabie and Deslippe, 2013). Indeed, in arable systems, taxonomic diversities of key species of arbuscular mycorrhizal fungi have become reduced compared with woodland systems, implying impaired soil system functioning (Helgason et al., 1998). The removal of woodland, forest, hedges, and native grasslands across the world to make way for agriculture has substantially impaired interactions between above and below ground systems, with negative implications for sustainability and soil functioning. The reintroduction of trees and shrubs and other forms of biodiversity in the landscape is thus critical in sustaining soil microbial communities.

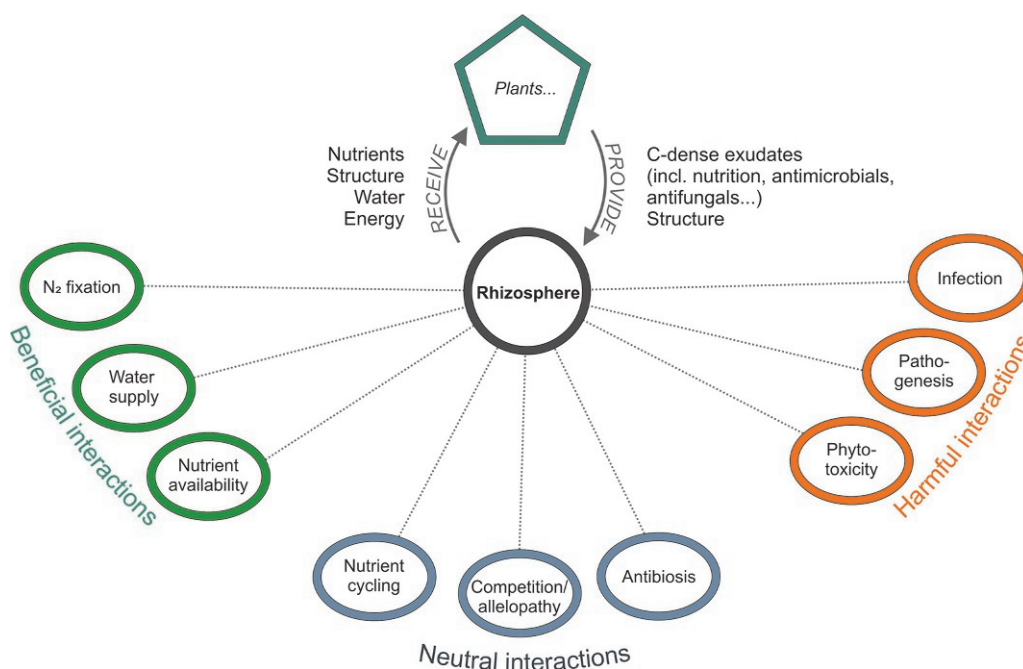


Fig. 2 Examples of rhizosphere interaction between plants, soil and soil microorganisms.

Hedges

Species-rich semi-natural and natural habitat remnants play a key role in maintaining biodiversity as well as delivering a range of other ecosystem services in agricultural landscapes (Table 1). Hedgerows are an example of such habitat—linear features composed of shrubs and trees that were traditionally planted to delineate fields, simultaneously providing stock-proof boundaries, as well as being sources of food, fuel, and timber. Increases in population density, intensifying competition for productive land and mechanized farming produce a need for delineation of holdings and systems of management such that useful space is maximized and tools or machinery may be used as efficiently as possible. Hedges have been effective in this regard, not only providing linear separation and efficient use of space, but also other ecosystem benefits such as space for nature within cultivated landscapes. However, after the Second World War, farmers in many countries were encouraged by policy and financial incentives to remove hedgerows to make farming even more efficient, incorporating bigger farm machinery and goals of greater food self-sufficiency. At its peak, hedgerow removal in the United Kingdom reached 6400 km per year. Hedge removal produces negative impacts for biodiversity and wider ecosystem service delivery. Given that hedges can trap large quantities of sediment moving downslope during storm events, removals can lead to sediment loss from agricultural landscapes, with associated deterioration in downstream water quality. In the EU hedge systems have made way for growing agricultural operations since 1990, resulting in less space for nature and increasing ecosystem damage, such that European agricultural areas are among the most ecologically degraded of all habitat types on the continent (Pe'er et al., 2014). Restoration of hedges is thus a policy priority in many temperate zones, particularly where hedges are simultaneously capable of providing space for nature without substantial encroachment on agriculturally productive land.

Extensive hedge networks are commonly found in temperate landscapes. A well-known example is the bocage of northwest France (Fig. 3), although hedge networks are also common in the United Kingdom and across much of Western Europe. Hedgerows enhance ecological corridors for biodiversity, protect soil and limit wind and water erosion. Soil carbon storage is enhanced, flood risk may be reduced in some circumstances and hedge networks offer considerable landscape amenity. Soils under hedge systems are more enriched in organic matter than the fields they surround, resulting in lower bulk density, higher infiltration rates and increased vertical drainage. N uptake in the vicinity of hedges may also be enhanced, improving water quality. Large canopies facilitate higher interception while demand for water and deeper roots under hedges increase evapotranspiration, resulting in drier soil. This, in-turn, reduces the development of saturation-excess overland flow, maintaining subsurface storage capacity for storm rainfall and incoming flows from upslope. Water tables below hedgerows are commonly depressed compared with adjacent fields, and the uppermost soil layer is likely to contain significantly less moisture than adjacent grass margins, pasture or arable fields. Hedges may influence soil water content in winter up to 10 m on either side due to the hydraulic gradient generated by drier soil in the vicinity of roots (Caubel et al., 2003; Holden et al., 2019). Scaling up hydrological effects such as these from field to whole catchments is challenging—once separate drainage systems are combined, the complexity of interactions between flow processes increases to such an extent that it becomes very challenging to model the impact of hedgerows on landscape flood mitigation. Nonetheless, evidence illustrates the importance of hedge soils as a valuable sink for excess water to mitigate downstream flooding particularly when hedges are aligned across-slope, reducing sediment loss from agricultural fields during storm events and loss of soil habitat from within catchments.

Whereas the above-ground landscape benefits of hedges are more visible, effects on below-ground soil functioning are much less well known. However, hedge soils show very different characteristics to either conventional or even organic fields. Fungal biomass and diversity are elevated beneath hedges due to the lack of fungicide applications and extra plant diversity. Fungal populations are capable of decomposing recalcitrant organic compounds which are prevalent beneath hedge soils. This means that hedge soils are commonly fungal-dominated, unlike soils which receive high inputs of synthetic fertilizer or are beneath monocultures which are typically bacterial-dominated (Bardgett et al., 1999). Hedge soils and grassy field margins typically support more diverse communities of macrofauna such as earthworms, which can be 1–2 orders of magnitude more abundant than in adjacent fields. Hedge soils store a third more C in the top half-meter than beneath intensively managed grassland, and can accumulate C at a rate of 1.5 Mg C ha⁻¹ yr⁻¹. Soil temperatures are additionally better regulated, although curiously the drying effect of soil beneath hedges can give rise to higher solute concentrations of NO₃⁻ and PO₄³⁻—a counterintuitive result given the lack of fertilizer applications beneath hedges (Holden et al., 2019; Biffi et al., 2022).

The relative complexity of soil systems beneath hedges, in contrast to conventionally cultivated areas, generates habitat for subsurface diversity to drive enhanced soil functioning and ecosystem service delivery. Enhanced agroecosystem complexity in landscapes with hedges is correlated with many beneficial factors in the vicinity of hedges, such as increased stocks of SOC, which in turn supports a more dynamic microbial population. The reintroduction of hedge ecosystems is becoming a priority for governments including both the UK and EU countries, as reports increasingly recognize the role that hedgerows can play in achieving biodiversity gain and net zero targets on land. What remains to be seen is whether hedge systems alone can reinvigorate soil system functioning at sufficient scale for benefits to have a landscape effect and deliver on both economic and ecological goals.

Agroforestry

Agroforestry is the cultivation of two or more species of plants (or plants and animals) within the same space, at least one of which is a woody perennial (Nair, 2004). Agroforestry thus offers many of the same benefits as hedges, such as shelter and protection of soil,

Table 1 Contributions of trees, hedges, agroforestry, and soil microbial communities to ecosystem services.

<i>Ecosystem service</i>	<i>Description</i>	<i>Soil microbes</i>	<i>Trees, hedges, and agroforestry</i>
Raw material	Soils have historically been used as raw materials, e.g., as peat for fuel or as building material (rammed earth, bricks)	Microbes contribute to soil formation through nutrient cycling and organic matter production	Timber trees offer wood as a building material, and short rotation woody crops may be grown for bioenergy
Promotion of vegetation	Vegetation and associated products are used in many contexts including as food, energy and building materials. Plants consequently provide services of their own, such as carbon cycling, filtration and public amenity	Soil microbes make otherwise sparingly soluble nutrients, such as Fe and P, available to plants to support growth	Hedges and agroforestry create shelter, allowing plants to be established in their vicinity
Natural flood management	Storage of excess water in the landscape may alleviate flood risk following extreme precipitation events; degree of storage is mediated by soil structural variables such as bulk density, infiltration rate and permeability, and surface understory variables influencing surface roughness interactions with overland flow	Soil microbes directly influence soil structure, and are additionally a food source for ecosystem engineers such as earthworms, responsible for formation of macropores. Microbes support plant roots and their contribution to soil structure	Soils under hedges and agroforestry demonstrate higher levels of infiltration, allowing water to be stored in the landscape during storm events, mitigating downstream flood peaks at local scales. Additionally, tree and hedge canopies intercept water, creating above ground storage reducing flood peaks
Nutrient cycling	Soil cycles nutrients by facilitating decomposition of organic matter and mobilizing bedrock and aggregate material	Soil bacteria, archaea and fungi drive nutrient cycling in soils and are involved in weathering minerals	Leguminous trees and shrubs are capable of fixing N from the atmosphere in soils
Waste recycling and detoxification	Soils absorb, detoxify and recycle waste materials, agrochemicals and fuel or oil spills, reducing environmental hazard	Microbial mineralization and immobilization processes mitigate the effects of pollutants. Microbes are the main way that organic pollutants are degraded in soil	High transpiration rates, root biomass and longevity allow perennial vegetation to process waste through phytoremediation
Filtration of contaminants	Leached materials such as synthetic fertilizer may contaminate streams and lakes and their ecosystems. Soils act as a buffer, absorbing contaminants within clay minerals and organic matter	By contributing to soil physical structuration and the wider soil food web, microbes promote water holding capacity of soils and their potential to absorb contaminants. Mycorrhiza reduce the leaching of mineral nutrients	Trees and plants absorb atmospheric pollutants including those harmful to human health, removing them from the environment
Natural habitat	Although comparatively less well studied, soil is home to an enormous proportion of global biodiversity which itself depends on natural soil function for survival	Soil bacteria, archaea and fungi comprise the vast majority of global biological diversity, while also being the foundation of soil food webs and underpinning higher trophic levels both above and below ground	Trees and hedges provide essential habitat for above ground biodiversity and ecological corridors in the landscape. Additionally, roots create habitat for below ground microorganisms and fungi
Control of pests, weeds and pathogens	By providing habitat for beneficial species capable of managing pest populations and pathogens, soil provides a buffer against invasive species	Many beneficial species are soil microbes which support plant growth and manage populations of pathogens	Trees manipulate the soil environment in the vicinity of roots (rhizosphere) using exudates. This includes control of pathogens and competing plants and herbivores
Regulation of greenhouse gas emissions	SOC represents the second-largest deposit of C after the oceans, making soils a significant influence over global greenhouse gas concentrations including CO ₂ and CH ₄	Soil microbes both mineralize and decompose soil C and other nutrients, thus influencing both the C storage capacity of soils and rates of exchange of greenhouse gases with the atmosphere	Woody plants absorb CO ₂ for photosynthesis. C is sequestered in the form of above ground biomass, and also transferred to soil through leaf and root litter, and as C-dense exudates from roots

Adapted from Aislabie J. and Deslippe J.R. (2013) Soil microbes and their contribution to soil services, *Ecosystem Services in New Zealand—Conditions and Trends*, Vol. 1(12), pp. 143–161. Lincoln, New Zealand: Manaaki Whenua Press



Fig. 3 Bocage landscape, northern France. Bocage landscape of Avesnois by pyje0 is licensed under CC BY-SA 2.0. To view a copy of this license, visit <https://creativecommons.org/licenses/by-sa/2.0/?ref=openverse>.

while substantially increasing the density of space for nature in agricultural land. Consequently, agroforestry systems are more complex, both ecologically and economically. Within agroforestry systems, trees may be planted both on the edges of fields and in among productive areas normally used solely for livestock or annual crops.

The first written use of the term *agroforestry* was in the late 1970s, at which time the focus was explicitly on provisioning benefits of incorporating trees into farming systems, only alluding to sustainability as a secondary benefit in service of provisioning. However, such a focus is not unusual; practices resembling agroforestry are ancient, and the productivity benefits of farming with trees have been known for millennia. Throughout history crops and trees have been deliberately cultivated together, and like other regenerative or agroecological practices, agroforestry can be found in ancient indigenous practice (Nair, 1993). Practitioners of agroforestry often do not use a word to describe it, instead considering simply normal or traditional practices. The Yoruba of Nigeria cite soil fertility and avoidance of nutrient leaching and erosion as principal benefits of combining herbaceous, shrub, and tree crops (Nair, 1993). Shade-grown coffee and cocoa comprise a considerable area of global agroforestry, and for many years communities have identified better food security and economic returns as reasons for combining the two crops. Similarly, some of the most successful examples of agroecological farming today revive ancient technologies, incorporating trees to promote productivity and simultaneously conserve resources, with potential to reclaim function within partially or wholly degraded soils otherwise unsuited to sustainable production.

More recently agroforestry has been defined simply as ‘farming with trees’ (Soil Association, 2019). Particularly in the industrialized world, the stated environmental benefits of agroforestry have grown in prominence as an alternative to conventional, intensive land management systems. Overreliance on agrochemicals such as fertilizers and pesticides and the systematic conversion of natural habitat to cultivation have detached agricultural practice from ecological cycles which sustain ecosystem service provision. Yield and efficiency gains through the 20th century, although economically significant, are not sustainable so long as natural systems simultaneously sustained by land are neglected. In contrast, the presence (or reintroduction) of trees on farms offers potential to restore and take advantage of ecological complexity for the benefit of provisioning and ecosystem regulation. In this sense, agroforestry is typical of agroecological methods, applying ecological principles to the design of food systems (Gliessman, 2014) and utilizing circularity to generate multiple benefits and simultaneously address challenges to soil functioning, food security and environmental sustainability.

For most of history, food production has been the justification for incorporating trees on farms, with ecological gains serving that purpose. The potential for agroforestry to deliver global goods beyond provisioning, in particular carbon sequestration or resilience against climate change, is comparatively modern and part of the growing trend in its popularity. Of particular relevance is the extent to which agroforestry can deliver benefits in temperate regions to mirror its success elsewhere (e.g., tropics) for both people and environment. Historic agroforestry practices are found in temperate areas—the concept of ‘dehesa’ (Fig. 4), a Spanish system (‘montado’ in Portugal) in which dispersed pastureland is interspersed with holm and cork oaks, has origins in the early Middle Ages, however the practice is almost certainly much older than the word (Álvarez, 2016). Similarly, trees planted within slash and



Fig. 4 Examples of agroforestry. Clockwise from top left: (a) Cattle graze in a silvopastoral dehesa system; (b) silvoarable alleys within poplar in Bedfordshire, UK; (c) shelterbelt planted for protection of bulb crops, Isles of Scilly, UK. Panel (a) *Toros en la dehesa* by Fernando Cuenca Romero is licensed under CC BY-NC-SA 2.0. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc-sa/2.0/?ref=openverse>, Panel (b) *Silvoarable agroforestry Silsoe 2002* Paul Burgess by agforward is licensed under CC BY-NC-SA 2.0. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc-sa/2.0/?ref=openverse>, Panel (c) *Bulb fields and shelter belts from the air*. St Mary's by Mary Gillham Archive Project is licensed under CC BY 2.0. To view a copy of this license, visit <https://creativecommons.org/licenses/by/2.0/?ref=openverse>.

burn systems in Germany date back to the Middle Ages. Lines of trees have often historically been introduced as shelterbelts, and some modern agricultural landscapes in Europe resemble agroforestry simply as a result of remnant trees surviving land-use change. What distinguishes modern agroforestry from the simple presence of trees on farms is its intentionality and focus on people and management as integral components of the system.

Agroforestry includes several subdivisions depending on system design (Fig. 4). Silvopasture describes integration of trees with livestock such as ruminants, pigs, or poultry, an example being the dehesa or montado system mentioned above. Silvoarable agroforestry is the integration of trees and crops in the same field. Within silvoarable systems, trees are normally planted in rows to allow access for machinery, with crops in the 'alleys' in-between; as such silvoarable agroforestry is often referred to as 'alley cropping' (Lawson et al., 2016; Soil Association, 2019). Additionally, all three elements (trees, crops, animals) may be combined into an agrosilvopastoral system. Additional systems are commonly classed as agroforestry: shelterbelts, windbreaks and riparian buffer strips may also be examples of trees on farms for the promotion of ecosystem and economic benefit, designed to protect livestock, crops or soil and water quality. Smaller-scale applications such as kitchen gardening make use of trees to contribute to vegetable production.

The reintroduction of trees to farmed landscapes in more recent times is thus with the explicit focus of strengthening soil function and wider ecosystem goods lost through land-use conversion to conventional agriculture. Trees provide ecological

corridors within the farmed landscape, while offering many of the benefits for soil function that come with increased diversity of vegetation and the ability to host higher levels of above and below ground biodiversity.

Introduction of conventional broad-leaved woodland (>400 stems ha^{-1}) on previously arable land generally delivers higher long-term SOC stocks in soils. Other soil function benefits can emerge in tandem with C improvements, such as increases in plant-available N, earthworm abundance and functional diversity (Ashwood et al., 2019). Planting densities for agroforestry are typically lower ($75\text{--}200$ stems ha^{-1}) than for woodland, yet enhanced SOC stocks are nonetheless commonly observed to a depth of 100 cm. The effect is dependent on the type of land-use change; combining global studies illustrates that conversion from arable to agroforestry produces the largest increase in SOC stocks ($+40\%$), with a much smaller increase for converting pasture or grassland to agroforestry ($+9\%$) (De Stefano and Jacobson, 2018). Agroforestry systems have larger SOC concentrations than treeless control sites (both diachronic, in which controls are baseline-sampled on the same site before agroforestry establishment; and synchronic, in which controls are a treeless neighboring plot with similar characteristics). Broadleaf systems lead to greater SOC addition than coniferous systems, and hedges lead to the greatest addition of SOC to soils in comparison with treeless or hedgeless sites, followed by silvoarable and silvopasture sites, respectively (Mayer et al., 2022).

When we consider agroforestry and SOC effects more closely in context, the picture becomes more complex. Evidence from a number of sources suggests the importance of understory vegetation in controlling SOC beneath trees and grassland. Fine root density is generally higher where understory vegetation is dense and thriving, which may in turn lead to higher rates of rhizodeposition and incorporation of root matter into the soil, elevating SOC concentration. Particularly for silvoarable systems, it is possible that the majority of above-ground organic C inputs come from herbaceous vegetation, and not from trees, and similarly when understory is absent or artificially eliminated, SOC stocks beneath agroforestry are commonly reduced. At a simple level, inputs of C to the soil come via three principal pathways: addition of leaf litter (or pruning residue) to the soil surface, addition of fine root litter within the soil and the production of organic carbon exudates within the rhizosphere (see Fig. 1). SOC stocks in soil are in a dynamic equilibrium between gains from exudates and organic matter, and two types of loss: firstly to the atmosphere as CO_2 as a result of decomposition, and secondly via the aquatic ecosystem through leaching of dissolved organic carbon (DOC) or erosion of particulate organic carbon (POC). Root matter helps to stabilize soil aggregates, which in turn protects organic molecules from exposure to microbial breakdown and loss from the system.

Interactions between trees and soil with respect to macronutrients are well known in tropical zones. For example, the addition of leguminous (N-fixing) trees in agricultural systems has been shown to improve circularity in N usage. Trees and crops compete for N in the surface soil in tropical soil systems; however, tree roots access leached N from deeper soil horizons otherwise inaccessible to crops and their more superficial rooting systems in a process known as nutrient 'pumping.' Alfalfa (*Medicago sativa* L.) used as an N-fixing intercrop within alleys of 17-year-old walnut trees (*Juglans nigra* x *Juglans regia* L.) in a Mediterranean experiment show that yield losses in shaded areas close to tree rows are buffered by facilitative mechanisms between plants and trees which increase light use efficiency in the intercrop (Querné et al., 2017). Moreover, agroforestry systems can add nutrients to surface soil directly through litterfall. In Zambia *Faidherbia albida* A.Chev., an N-fixing tree, is used to improve soil fertility in the short term. Soils amended with *F. albida* leaf litter demonstrate greater rates of N mineralization than control soil samples with fertility boosted beneath the canopy through longer-term modification of soil biological and chemical properties (Yengwe et al., 2018).

Physical canopies of trees create a microclimate for below-ground organisms, with some mature species hosting 35%–60% more microbial biomass C in soil beneath their canopies. Litterfall leads to extra soil fertility, but canopies also create habitat for larger animals such as birds whose fecal matter is added to soil beneath trees. Trees modify the soil environment for microorganisms through the effect of roots and canopies on catchment hydrology. By increasing SOC stocks, afforestation of mineral soils produces greater soil water holding capacity and infiltration capacity during storm events. By creating subsurface habitat for soil microorganisms and other ecosystem engineers, trees promote physical alteration of the soil structure and the formation of macropores and channels. These in turn promote lower bulk densities for soil and alter hydrological functioning.

The litterfall effect may be moderated by tree age and, in turn, litter quantity. Younger cacao trees in an agroforestry site in Mexico moved a greater supply of phosphorous (P), calcium (Ca), magnesium (Mg), iron (Fe) and copper (Cu) into soil because they produce a higher supply of litter than older trees (Pérez-Flores et al., 2018). But irrespective of age, the trees are able to compensate for the nutrient demand of the main crop. Litterfall supply of each of N, P, Ca, Mg and also potassium (K) is sufficient to recover nutrient concentration lost to the main cacao crop, demonstrating the greater circularity in nutrient use which is typical of complex agroforestry systems.

However, species choice in agroforestry systems is critical for maximizing nutrient benefits. Facilitation of increases in stocks of different nutrients in soil can be species dependent, with several examples from experimental studies. Isolated trees in a Nicaraguan silvopastoral system increased C, N, and P concentrations in the topsoil, with *Crescentia alata* Kunth improving N stocks and *Guazuma ulmifolia* Lam. favoring P stocks (Hoosbeek et al., 2018). N mineralization patterns varied among 10 plant species considered by Partey et al. (2018), with greatest N-supply (up to 93 mg N kg^{-1}) coming from two of them—*Tithonia diversifolia* (Hemsl.) A.Gray and *Gliricidia sepium* (Jacq.) Steud.

Knowledge of nutrient effects in temperate areas is more limited, although species-specific effects nonetheless emerge. For example, roots with larger N and P concentrations, such as those from legume crops, promote a smaller C:N ratio which facilitates faster decomposition of organic matter; furthermore, additive effects between litter and crop residues enhance efficiency of uptake and grow long term fertility (Zeng et al., 2010). To an extent, trees compete with alley crops for nutrients, and seasonal phase effects between each are critical for overall efficiency. As if to demonstrate this, yields and N use efficiency for agroforestry systems may be influenced by the presence or absence of artificial barriers between tree and crop. In a US Midwestern maize system combined with

black walnut trees (*Juglans nigra* L.), tree roots take up soil water and N before the maize is planted, reducing supply for the crop. With a 1.2 m soil barrier in place the effect is minimized (Jose et al., 2000). Yet there are caveats: fine tree roots due to their more rapid release of N and P, are able to bolster soil nutrient pools. Additionally, the pre-existing nutrient-depleted state of the soils may create a dependency for artificial N input. With stronger positive feedbacks associated with fertility and microbial activity, such inputs could potentially be reduced.

Introduction of trees to agroecosystems creates habitat for wildlife and opportunities to improve ecological connectivity. By reconciling higher levels of production with greater species diversity, the potential exists for agroforestry to contribute to restoration of biodiversity in farmed landscapes. Among the most damaging impacts of intensive agricultural practice in the last century has been declines in insect pollinators. Yet silvoarable and silvopastoral systems are capable of increasing airborne arthropod diversity compared with monocultures. Bees and hoverflies may be found to be twice as abundant, seed set several times greater and species richness of solitary bees an order of magnitude greater within agroforestry systems, compared with monocultures (Varah et al., 2020). By restoring space for nature within cultivated landscapes, agroforestry systems are capable of supporting pollinators and beneficial pest predators.

Agroforestry, if deployed at scale, carries the potential to reconcile need for land for cultivation with a similarly critical need to sustain soil system functioning. Greater SOC and nutrient stocks benefit soil microbial dynamics, ensuring the food web is supported by a steady supply of inputs. As discussed, increased plant diversity is capable of promoting greater shoot and root biomass, quantities of exudates, bacterial biomass and fungal biomass—demonstrating once again the close linkage between above ground system complexity and below ground functioning. Explicit links between above ground biodiversity benefits and subsurface dynamics are limited, requiring a greater focus on the centrality of soil biodiversity in the sustainability of agricultural landscapes. But as with hedges, evidence strongly points toward a close relationship between agroforestry, landscape resilience, and holistic functioning of soil systems.

Summary

Trees and hedges are critical in supporting microbial diversity within soil systems. Systems with greater density of woody vegetation promote ecosystem benefits including greater SOC stocks, efficient nutrient cycling, reduced erosion and protection of soil habitat, while enhancing hydrological functioning. Ecosystem services sustained by soil microbial diversity are thus directly supported by perennial woody vegetation. Although we can infer characteristics of these linkages however, quantitative study of the pathways themselves is challenging. Crop studies are one means of understanding linkages and have shown that greater plant diversity leads to greater shoot biomass, root biomass, quantity of exudates, bacterial biomass, and fungal biomass. Yet it is difficult with this method to isolate growth directly attributable to rhizodeposition. Some have tried to overcome the issue in the field by using isotope tracers such as $^{14}\text{CO}_2$, yet the difficulty of accurate in-field measurement persists making quantitative identification of fluxes elusive. Growing plants hydroponically offers better visibility of interactions, albeit in an unrealistic environment. The exact pathways by which soil microorganisms are supported by vegetation thus still need to be uncovered. Nonetheless, interdependency between above-ground vegetation and microorganisms is demonstrated in the marked loss of functioning once land systems become dominated by monoculture or systems of land use which are more linear than circular. Effective stewardship of soil system resources requires an appreciation of their complexity and circularity, and the design of systems which operate in step with these interactions.

Although the benefits of agroforestry and hedge restoration may be characterized at large scale, practical considerations remain in terms of addressing constraints on implementation within local geographies. Firstly, better evidence is needed for ecosystem effects in temperate and boreal zones. For example, dynamics of nutrient cycling are better mapped elsewhere in the world, especially in the tropics, and these must be verified or adapted for temperate settings to design strategies for reintroduction of trees or hedges to agricultural landscapes. Secondly, an understanding of causal dynamics, system complexity, and feedback mechanisms within agroforestry systems would facilitate understanding of wider ecosystem goods and the consequences of particular interventions, particularly at catchment scale and above. This is particularly the case for flood management and hydrological effects, for which nonlinear dynamics become important when small-scale interventions are aggregated. Thirdly, and building on a complex systems perspective, better understanding is needed of the aggregated benefits, synergies and trade-offs associated with agroforestry interventions in the plot-scale spatial domain. High level effects of practices such as agroforestry on whole landscapes have received considerable attention, yet much less is known about effects on simultaneous multiple benefits at smaller scale, including soil functioning. The number of large-scale studies far exceeds the extent of local, reusable, and diversified scientific knowledge which can inform practitioners how to make instrumental changes at scales over which they have control. With this knowledge, decentralized action in the landscape capable of regenerating soil systems becomes a realistic goal.

References

- Aislabie J and Deslippe JR (2013) Soil microbes and their contribution to soil services. *Ecosystem Services in New Zealand—Conditions and Trends*. vol. 1, pp. 143–161. Lincoln, New Zealand: Manaaki Whenua Press. (12).
- Álvarez JRG (2016) The image of a tamed landscape: Dehesa through history in Spain. *Culture and History Digital Journal* 5(1): 1–17. <https://doi.org/10.3989/chdj.2016.003>.

- Ashwood F, et al. (2019) Woodland restoration on agricultural land: Long-term impacts on soil quality. *Restoration Ecology* 27(6): 1381–1392. <https://doi.org/10.1111/rec.13003>.
- Bardgett RD, et al. (1999) The measurement of soil fungal:bacterial biomass ratios as an indicator of ecosystem self-regulation in temperate meadow grasslands. *Biology and Fertility of Soils* 29: 282–290. Springer-Verlag.
- Biffi S, et al. (2022) Soil carbon sequestration potential of planting hedgerows in agricultural landscapes. *Journal of Environmental Management* 307(June 2021), 114484. Elsevier Ltd <https://doi.org/10.1016/j.jenvman.2022.114484>.
- Caubel V, et al. (2003) Influence of a hedge surrounding bottomland on seasonal soil-water movement. *Hydrological Processes* 17(9): 1811–1821. <https://doi.org/10.1002/HYP.1214>. John Wiley & Sons, Ltd.
- De Stefano A and Jacobson MG (2018) Soil carbon sequestration in agroforestry systems: a meta-analysis. *Agroforestry Systems* 92(2): 285–299. <https://doi.org/10.1007/s10457-017-0147-9>. Springer Netherlands.
- Dollinger J and Jose S (2018) Agroforestry for soil health. *Agroforestry Systems* 92(2): 213–219. <https://doi.org/10.1007/s10457-018-0223-9>. Springer Netherlands.
- Gliessman SR (2014) *Agroecology: The Ecology of Sustainable Food Systems*, 3rd edn. Boca Raton: Taylor & Francis. Available at: <https://doi.org/10.1201/b17881>.
- Haichar FZ, et al. (2014) Root exudates mediated interactions belowground. *Soil Biology and Biochemistry* 77: 69–80. <https://doi.org/10.1016/j.soilbio.2014.06.017>. Elsevier Ltd.
- Helgason T, et al. (1998) Ploughing up the wood-wide web? *Nature* 394(6692): 431. <https://doi.org/10.1038/28764>. Nature Publishing Group.
- Hiltner L (1904) Über neuere Erfahrungen und Probleme auf dem Gebiete der Bodenbakteriologie unter besonderer Berücksichtigung der Gründüngung und Brache. *Arbeiten der Deutschen Landwirtschaftlichen Gesellschaft Berlin* 98: 59–78.
- Holden J, et al. (2019) The role of hedgerows in soil functioning within agricultural landscapes. *Agriculture, Ecosystems and Environment* 273(December 2018): 1–12. <https://doi.org/10.1016/j.agee.2018.11.027>. Elsevier.
- Hoosbeek MR, Remme RP, and Rusch GM (2018) Trees enhance soil carbon sequestration and nutrient cycling in a silvopastoral system in south-western Nicaragua. *Agroforestry Systems* 92(2): 263–273. <https://doi.org/10.1007/s10457-016-0049-2>. Springer Netherlands.
- Jose S, et al. (2000) Defining competition vectors in a temperate alley cropping system in the midwestern USA; 3. Competition for nitrogen and litter decomposition dynamics. *Agroforestry Systems* 48(1): 61–77. <https://doi.org/10.1023/A:1006241406462>.
- Lawson GJ, et al. (2016) Sustainable management criteria for agroforestry in the European Union. In: *3rd European Agroforestry Conference 2016 Book of Abstracts*, 375–378. (4).
- Lynch JM (1990) Introduction: Some consequences of microbial rhizosphere competence for plant and soil. In: Lynch JM (ed.) *The Rhizosphere*, pp. 1–10. New York; Chichester: John Wiley & Sons, Ltd.
- Mayer S, et al. (2022) Soil organic carbon sequestration in temperate agroforestry systems—A meta-analysis. *Agriculture, Ecosystems, and Environment* 323. <https://doi.org/10.1016/j.agee.2021.107689>.
- Meadows D (1999) *Leverage Points: Places to Intervene in a System*. The Sustainability Institute.
- Menalled UD, Seipel T, and Menalled FD (2020) Farming system effects on biologically mediated plant-soil feedbacks. *Renewable Agriculture and Food Systems* 36: 1–7. <https://doi.org/10.1017/S1742170519000528>.
- Nair PKR (1993) *An Introduction to Agroforestry, Outlook on Agriculture*. Dordrecht; Boston; London: Kluwer Academic Publishers. <https://doi.org/10.1177/003072709402300413>.
- Nair PKR (2004) Agroforestry. In: Hillel D and Hatfield JL (eds.) *Encyclopedia of soils in the environment*, 1st edn, pp. 35–44. New York, NY: Academic Press Imprint.
- Partey ST, et al. (2018) Improving maize production through nitrogen supply from ten rarely-used organic resources in Ghana. *Agroforestry Systems* 92(2): 375–387. <https://doi.org/10.1007/s10457-016-0035-8>.
- Pe'er G, et al. (2014) EU agricultural reform fails on biodiversity. *Science* 344(6188): 1090–1092. <https://doi.org/10.1126/science.1253425>.
- Pérez-Flores J, et al. (2018) Leaf litter and its nutrient contribution in the cacao agroforestry system. *Agroforestry Systems* 92(2): 365–374. <https://doi.org/10.1007/s10457-017-0096-3>.
- Prescott CE, et al. (2019) Rehabilitating forest soils after disturbance. In: *Global Change and Forest Soils*, pp. 309–343. Elsevier B.V. <https://doi.org/10.1016/b978-0-444-63998-1.00013-6>.
- Querné A, et al. (2017) Effects of walnut trees on biological nitrogen fixation and yield of intercropped alfalfa in a Mediterranean agroforestry system. *European Journal of Agronomy* 84: 35–46. <https://doi.org/10.1016/j.eja.2016.12.001>. Elsevier B.V.
- Sanderman J, Hengl T, and Fiske GJ (2017) Soil carbon debt of 12,000 years of human land use. *Proceedings. National Academy of Sciences. United States of America* 114(36): 9575–9580. <https://doi.org/10.1073/pnas.1706103114>.
- Shukla MK, Lal R, and Ebinger M (2006) Determining soil quality indicators by factor analysis. *Soil and Tillage Research* 87(2): 194–204. <https://doi.org/10.1016/j.still.2005.03.011>.
- Soil Association (2019) In: Raskin B and Osborn S (eds.) *The Agroforestry Handbook: Agroforestry for the UK*, 1st edn Bristol: Soil Association Limited.
- Varah A, et al. (2020) Temperate agroforestry systems provide greater pollination service than monoculture. *Agriculture, Ecosystems and Environment* 301(June): 107031. <https://doi.org/10.1016/j.agee.2020.107031>. Elsevier.
- Whipps JM (1990) Carbon economy. In: Lynch JM (ed.) *The Rhizosphere*, pp. 59–97. New York; Chichester: John Wiley & Sons, Ltd.
- Yengwe J, et al. (2018) Effects of Faidherbia albida canopy and leaf litter on soil microbial communities and nitrogen mineralization in selected Zambian soils. *Agroforestry Systems* 92(2): 349–363. <https://doi.org/10.1007/s10457-016-0063-4>. Springer Netherlands.
- Zeng DH, et al. (2010) Carbon mineralization of tree leaf litter and crop residues from poplar-based agroforestry systems in Northeast China: A laboratory study. *Applied Soil Ecology* 44(2): 133–137. <https://doi.org/10.1016/j.apsoil.2009.11.002>.

Urban soil microbiome: Activity, diversity and functioning

Maria Korneykova^{a,b}, Ekaterina Kozlova^a, Kristina Ivashchenko^{a,c}, Maria Vasilieva^a, and Viacheslav Vasenev^d,

^aAgrarian-Technological Institute, RUDN University, Moscow, Russia; ^bInstitute of North Industrial Ecology Problems, Kola Science Centre of Russian Academy of Sciences, Apatity, Russia; ^cInstitute of Physicochemical and Biological Problems in Soil Science, Russian Academy of Sciences, Pushchino, Russia; ^dSoil Geography and Landscape Group, Wageningen University, Wageningen, The Netherlands

© 2023 Elsevier Ltd. All rights reserved.

Introduction	480
Case studies of microbial activity, diversity, and functioning in urban soils	482
The effect of typology and composition of urban green infrastructures on soil microbiome for the case of university campus (RUDN University, Moscow)	482
From landscape variability to soil microbial diversity: A case study of a tremendous green roof in Zaryadye Park (Moscow, Russia)	485
Arctic urban soil microbiome (Apatity, Russia)	487
Conclusion	490
Funding	490
References	491
Further reading	492

Abstract

Urbanization coincides with direct and indirect anthropogenic impacts on soil health resulting in remarkable changes in diversity and functionality of soil microbiome. Soil sealing, pollution or salinization deplete soil microbial biomass and hamper microbial activity. Development of urban green infrastructures including constructing Technosols, introduction of new vegetation species, management and maintenance practices create favorable conditions for soil microbiome. Microbial indicators are very informative for urban soil monitoring and assessment since they reflect complex and diverse processes occurring in soil life phase in urban environment. Soil microbiome is sensitive to anthropogenic stressors, and dynamics in microbial properties is widely used to monitor anthropogenic and technogenic alteration of urban ecosystems. At the same time, ecosystem adaptation to these stresses is driven by mutual interactions and co-adaptation between soil microbiome and urban vegetation. This chapter reviews diversity of urban soils' microbial properties and explores climatic conditions, typology and composition of green infrastructures, management and maintenance as the driving factors behind microbial variability and functionality in urban soils.

Introduction

Soil microorganisms play a crucial role in all terrestrial ecosystems as they form a vast reservoir of genetic diversity and make a substantial impact on soil forming processes and functions. Soil is home for a diverse array of microorganisms including bacteria, archaea, microscopic fungi, and viruses. The whole complex of these microorganisms is called the soil microbiome and supports important soil functions and ecosystem services, such as organic matter mineralization (Li et al., 2018; Liu et al., 2022), regulation of nutrient cycling (Basu et al., 2021), biodiversity (Berg et al., 2017), biodegradation of pollutants (Diakhaté et al., 2016) and climate regulation (Joergensen and Emmerling, 2006).

Urban ecosystems are influenced not only by biotic and abiotic environmental factors, but also by human activities, which change soil composition and lead to high heterogeneity (Yang and Zhang, 2015). The impact of these activities on the soil microbiome is complex and can be influenced by a variety of factors, including urbanization intensity, soil properties, land use and management practices, as well as the level and type of pollution, and the presence of green infrastructure (Wuyts et al., 2020; Gómez-Brandón et al., 2022). The spatial structure of soil microbial communities can be affected by numerous factors, both direct and indirect. Soil pH, nitrogen availability, soil organic carbon content, temperature, and redox status are likely the most significant factors that can have a noticeable impact on the structure of soil communities (Pett-Ridge and Firestone, 2005; Cederlund et al., 2014).

Urban soils have been found to harbor a diverse range of bacterial and fungal microbial communities that play a crucial role in supporting green infrastructure such as parks, green roofs, and bioswales. These microbial communities contribute to enhanced functional potential in nutrient assimilation and cycling, bioremediation, and decomposition, which collectively enhance the overall health and resilience of urban ecosystems (Gill et al., 2020).

Soil microbiome of each city represents a complex, diverse, and dynamic system influenced by numerous factors, including climate, local environment conditions, land use patterns, industrial development, and the level and type of pollution. Anthropogenic activity resulting in contamination with oil, polychlorinated biphenyls, cement dust, or household waste can alter the bacterial communities in urban soils, leading to simplification of their structure and dominance of forms that are resistant to various pollutants. This shift can result in the accumulation of potentially pathogenic bacteria such as *Escherichia*, *Enterobacter*, *Klebsiella*, and *Alcaligenes*, as well as allergenic bacteria from genera such as *Rhodococcus* and *Micrococcus*. This indicates a violation of

the ecological function of the soil as a “bacterial filter” and poses a threat to human health. Additionally, urban soils can exhibit characteristics similar to those found in southern regions, including a greater prevalence of the *Aspergillus* genus, which could pose a risk to individuals with weakened immune systems (Zvyagintsev, 1987).

The soils of urban areas differ from their natural counterparts in characteristic features, which generally include high bulk density and pollution, alkaline pH, as well as low nutrient content, with some with some exceptions, such as phosphorus and nitrogen content (Chen et al., 2014), grass cover density and plant litter thickness. However, the diversity, heterogeneity, and spatial variability of urban soils, as well as the level of anthropogenic impact, can make it difficult to clearly distinguish them from their natural counterparts in some parameters, particularly from a microbiological perspective (Table 1).

Pollution, land use change, soil compaction, and transformation of vegetation cover affect soil microbial properties, shaping microbial community of urban soils different from natural reference. They have direct and indirect impacts on the soil microbiome, such as altering the microbial community composition, reducing microbial diversity, and affecting microbial activity (Fig. 1). Given that microorganisms play a crucial role in providing a wide range of soil ecological functions and are highly sensitive to environmental conditions (Dobrovol'skaya et al., 2015; Korneykova and Nikitin, 2021), they can serve as effective and sensitive indicators of these functions (Schloter et al., 2018). For instance, studies of the soil microbial properties of forest parks have demonstrated changes in the microbial indicators of the soil carbon and nitrogen cycles under the influence of urbanization, particularly vegetation management practices (Ananyeva et al., 2023; Wang et al., 2023). These changes are expressed in a decrease in the microbial biomass carbon content, respiration activity, and the availability of soil carbon and nitrogen pools, which considered as the main indicators of soil quality (Schloter et al., 2003).

The sustainable development of urban green areas is dependent on the interaction between microbial and plant communities, their self-regulatory mechanisms, and their ability to adapt to climate change and anthropogenic stress (Barthel et al., 2019). Soil-plant systems in urban environments are highly sensitive to natural and anthropogenic stresses such as climate change, pollution, and soil salinization, which can lead to a reduction in the provision of ecosystem services and ultimately the appearance of disservices and degradation of the ecosystem (Liu et al., 2022).

Table 1 Differences in the microbial community characteristics of urban soils compared to the background soils.

Parameter	City, Country	Climate	Method	Urban vs natural soil	References
Microbial biomass	Beijing, China	Temperate monsoon climate	FE ^a	Lower	Zhao et al. (2013)
	Kiel, Germany	Temperate	SIR ^b	Comparable	Beyer et al. (1995)
	Stuttgart, Germany	Temperate	FE, SIR	Comparable and lower	Lorenz and Kandeler (2006)
Microbial biomass	Moscow, Russia	Temperate continental	SIR	Lower	Ivashchenko et al. (2019)
Basal respiration			Gas chromatography		
Functional diversity	Beijing, China	Temperate monsoon climate	Biolog ^{ECO}	Higher	Zhao et al. (2013)
Taxonomic diversity	Berlin, Germany	Moderately continental	16S rRNA sequencing	Higher	Scholier et al. (2023)
	Jyväskylä; Kuopio; Mikkeli, Finland	Subarctic continental; Subarctic; Temperate continental	16S rRNA sequencing		
	New York City, United States	Humid subtropical	16S rRNA sequencing	Similar to the communities from other biomes worldwide	
Fungal diversity	Helsinki; Joensuu; Jyväskylä; Lahti; Tampere, Finland	Continental subarctic or boreal	eDNA metabarcoding	Lower	Abrego et al. (2020), Tedersoo et al. (2020) and Donald et al. (2021)
Bacterial diversity	Beijing; Dongguan, China	Monsoon-influenced humid continental; humid subtropical	16S rRNA sequencing	Higher	Yan et al. (2016) and Tan et al. (2019)
	Cayenne, French Guiana	Tropical	eDNA metabarcoding		Donald et al. (2021)
	Lancaster, Lancashire, England	Temperate	Phospholipid fatty acid analysis		Pereira et al. (2021)
	16 provincial capital cities of China	Tropics, subtropics, temperate, frigid temperate and plateau regions	16S rRNA sequencing	Little or no difference with natural counterparts	Xu et al. (2014)
Potential hazard	Los Angeles, United States	Subtropical	eDNA metabarcoding	Higher, associated with heavy-metal concentrations	Mejia et al. (2023)

^aFE, fumigation extraction method

^bSIR, substrate induced respiration method.

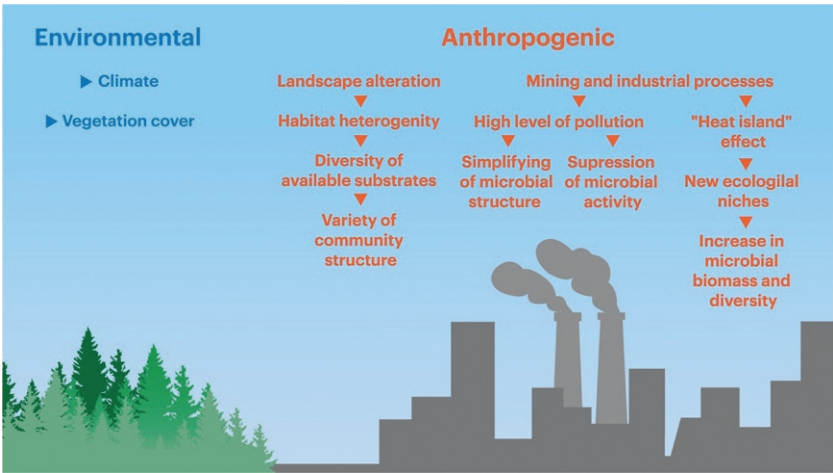


Fig. 1 Main factors influencing the urban soil microbiome and potential consequences of this influence.

This chapter aims to assess the impact of various factors of urban environment on the state of the soil microbiome using the microbial biomass, microbial respiration, community level physiological profile, taxonomic structure as informative indicators which widely used in urban soil science studies (Fig. 2).

Specifically, we examine the effect of different management practices on the soil microbiome properties, including (i) typical management practices of lawns with different trees for the case of a university campus, (ii) artificial landscaping of a tremendous green roof for the case of Zaryadye Park; and (iii) soils of urban green infrastructures under severe climatic conditions for the case of Arctic cities. We assume that climate. Aboveground vegetation and management practices are key drivers affecting soil microbial community in urban environment.

Case studies of microbial activity, diversity, and functioning in urban soils

The effect of typology and composition of urban green infrastructures on soil microbiome for the case of university campus (RUDN University, Moscow)

The relationship between the aboveground plant and the soil microbial communities remain poorly understood in urban environment. The composition and diversity of soil microbial communities appear to be strongly influenced by the structure of the above-ground vegetation, including the ratio of coniferous and deciduous species, with a greater impact on the community structure (Urbanová et al., 2015; Chen et al., 2019; Rodríguez-Rodríguez et al., 2023). Trees could influence by directly and indirectly on physiological status of soil microbial community through litter quality and root exudation (Attiwill and Adams, 1993; Kiikkilä et al., 2006; Augusto et al., 2015). However, there is lack of data on the effects of vegetation type on microbial activity and functional diversity, particularly in urban ecosystems.

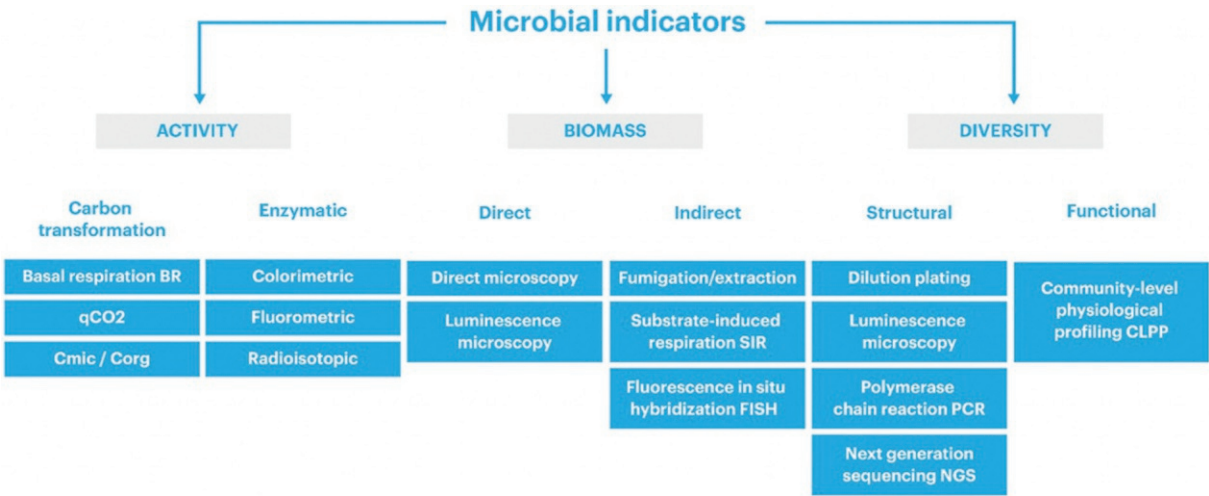


Fig. 2 The main microbial indicators and methods of their study.

The topsoil (0–10 cm) sampling campaign at RUDN University campus was organized in October 2019. Six tree species were selected: *Pinus sylvestris*, *Populus tremula*, *Acer platanoides*, *Tilia cordata*, *Picea abies*, *Betula pendula*. A composite soil sample was taken at a distance of 0.5 m from four opposite sides of the tree trunk from the top 0–10 cm layer at the beginning of October 2019 ($n = 5$ for each species, 30 in total) (Fig. 3). The following analysis were done in soil samples: pH_{water} (HANNA HI98129 pH meter, Romania), total carbon (C) and nitrogen (N) content (CHNS analyzer, Leco Corp., United States). Microbial biomass carbon (MBC) was measured by SIR method (Ananyeva et al., 2023). Basal respiration (BR) was measured as the rate of soil CO_2 release using gas chromatography (KrystaLLyuks-4000M; Meta-Chrom, Russia). The microbial metabolic quotient was calculated ($\text{BR}/\text{MBC} = q\text{CO}_2$). The capacity of the soil microbial community to metabolize diverse organic monomers (community-level physiological profiling, CLPP) was measured by the MicroResp™ technique for non-calcareous soils (Campbell et al., 2003). Microbial functional diversity was assessed through the Shannon-Wiener index: $H_{\text{CLPP}} = -\sum \text{Pi} \times \ln \text{Pi}$, where Pi is the ratio of respiration response to i substrate addition to total respiration response for all studied substrates. Samples for microbial properties measurements were taken from soils pre-incubated at 25 °C for 72 h (<https://estudogeral.uc.pt/bitstream/10316/24942/1/1-s2.0-S1470160X13003142-main.pdf>). Descriptive statistics were used to determine the mean, standard error, and variation coefficient. Significant differences in the variables between the studied tree species were examined by one-factor analysis of variance (ANOVA) with Tukey's multiple comparison test. Prior to the analysis, variance homogeneity was checked by Levene's test. The CLPP results were processed hierarchically using Ward's minimum variance method.

Among the studied chemical properties, only soil pH differed between lawns with different tree species (Table 2). Distribution of soil MBC and BR between studied groups didn't match (Fig. 4A and B). The lowest soil MBC content was under *Acer* tree (on average $538 \mu\text{g C g}^{-1}$), medium—for *Picea*, *Betula*, *Pinus* and (798, 805, 1124 $\mu\text{g C g}^{-1}$, respectively), and its highest content observed under *Populus* and *Tilia* (1378 and 1445 $\mu\text{g C g}^{-1}$). Soil BR didn't significantly differ between trees ($P = 0.1$). Specific respiration was higher for *Betula* and *Picea* by 1.3–2.1 times compared other species (Fig. 4C).

Considering the CLPP the studied samples were grouped in three clusters (Fig. 5A) (1) soil under *Tilia* tree with extremely high microbial response on diverse set of organic substrates (glucose, fructose, galactose, glycine, ascorbic and oxalic acids), (2) soil under *Acer* characterized by narrow spectra of organic substrates (ascorbic and oxalic acids) rapidly utilizing by microorganisms, (3) soil under *Betula*, *Picea*, *Pinus*, *Populus* with domination of microbial groups actively consumed two types of substrates—carbohydrates (fructose, glucose) and carboxylic acids (ascorbic and oxalic acids). The highest functional diversity of microbial community was found for *Betula*, but the lowest one for *Acer* (Fig. 5B). Hence, the metabolic status of soil microbial community and its capacity to decompose different organic compounds could be driven by the urban greening strategy. Besides, soil microbial community under *Betula* have high activity in utilizing phenolic acids by other words derivatives of aromatic hydrocarbons (organic pollutants). Perhaps, the high functional diversity of microbial community and capacity to utilize phenolic acids in soil under *Betula* could be explained by the relative lower its pH value (Table 2, Fig. 6). Interesting that soil MBC content and BR more related with N content than C ($r_s = 0.7$ vs. 0.6 and $r_s = 0.6$ vs. 0.5, respectively).

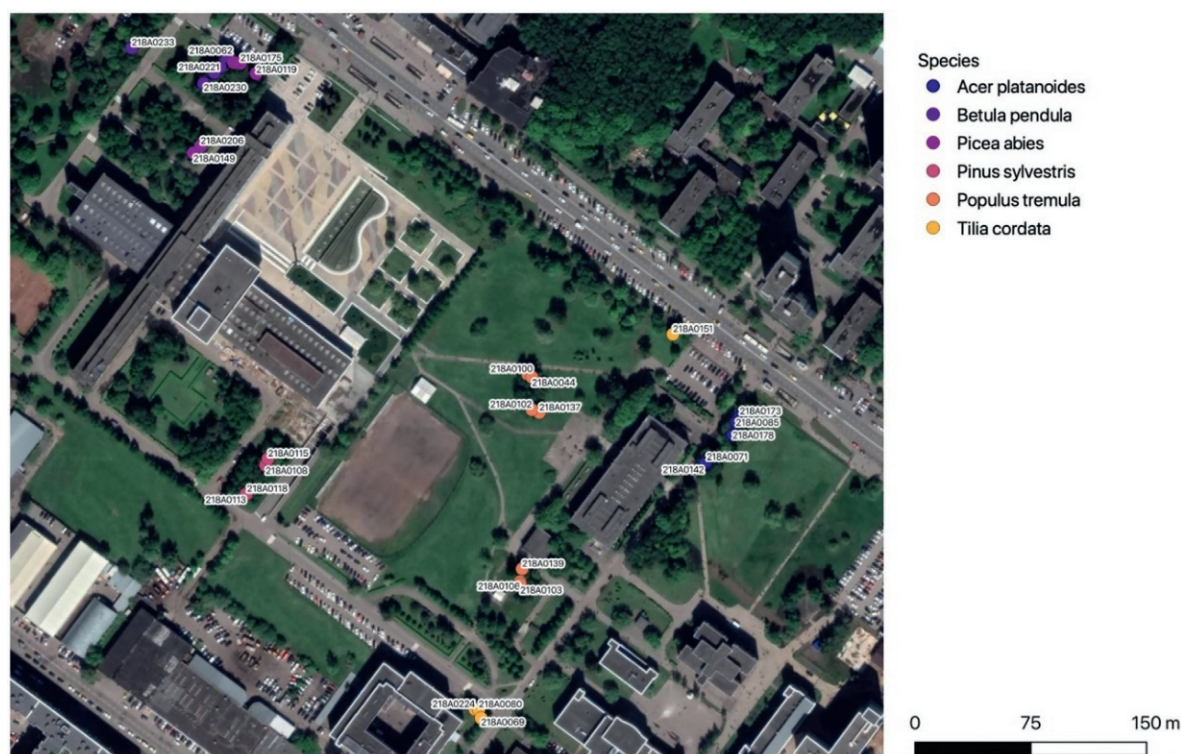
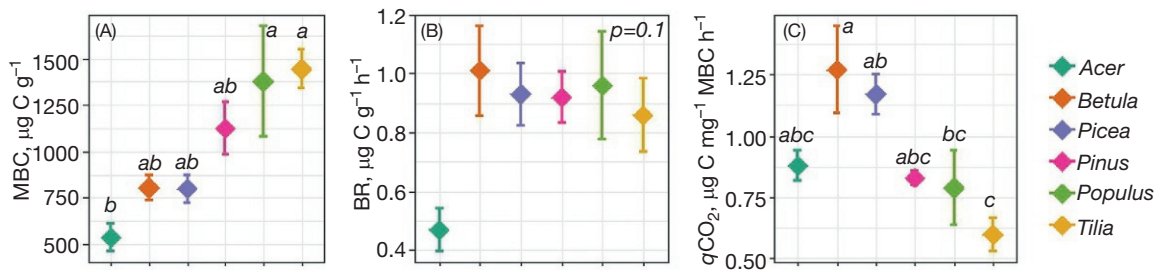
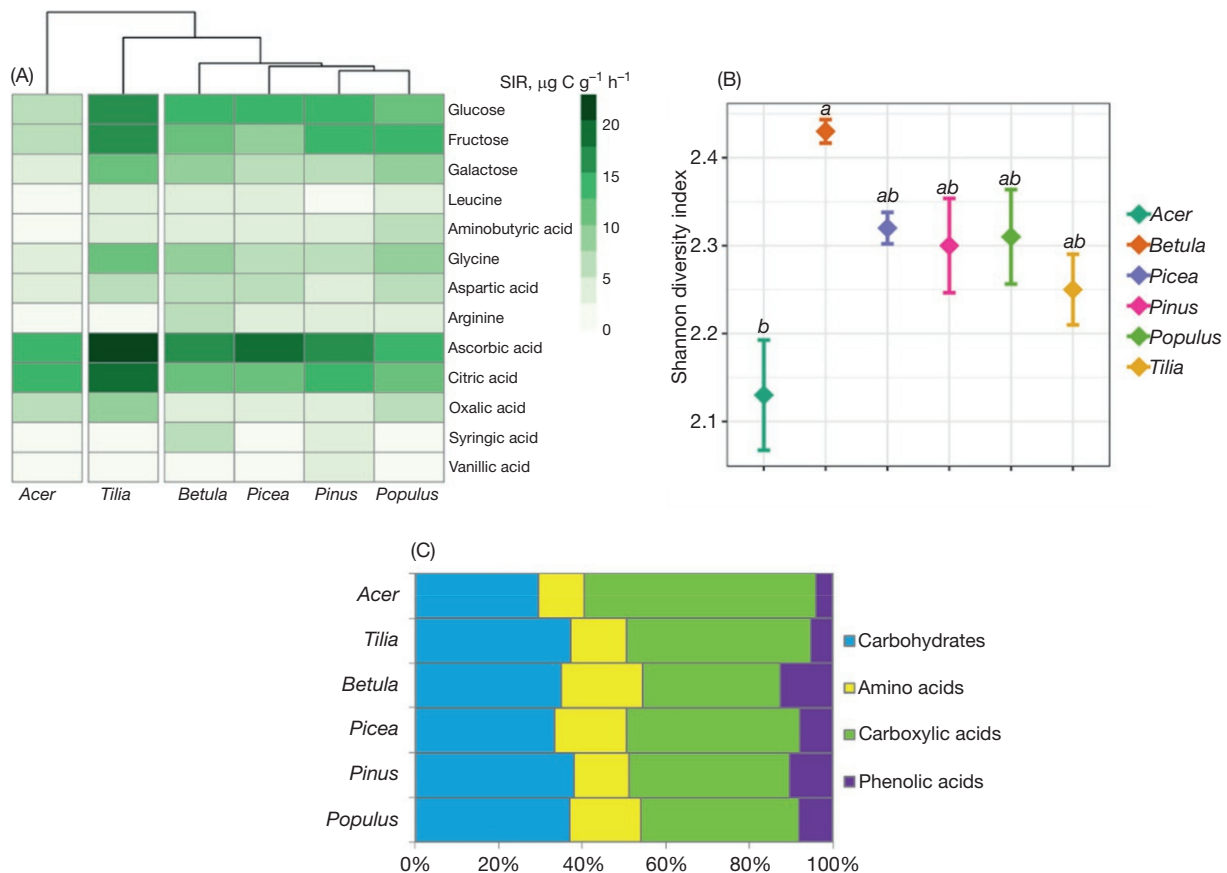


Fig. 3 The map of the soil sampling sites, RUDN University (55.652 N, 37.496 E, Moscow Russia).

Table 2 Topsoil properties (0–10 cm layer) for lawns with different tree species and diameter at breast height of trees (DBH).

Properties	Acer	Betula	Picea	Pinus	Populus	Tilia	P
pH	7.93 a(4)	6.48 c(7)	7.00 bc(12)	6.95 bc(10)	7.48 abc(1)	7.88 ab(4)	***
C, %	3.08/(20)	3.00/(23)	3.35/(34)	3.50/(10)	4.09/(29)	4.97/(54)	0.37
N, %	0.20/(33)	0.23/(19)	0.23/(22)	0.27/(11)	0.26/(22)	0.32/(43)	0.21
C:N	16.11/(21)	12.71/(8)	14.11/(19)	12.83/(9)	15.29/(13)	14.89/(15)	0.15
DBH	82.6/(8)	81.4/(10)	69.0/(13)	89.8/(20)	82.8/(7)	85.5/(8)	0.05

Values represent means with the variation coefficient (%) in brackets. Different letters indicate significant differences between groups (*** $P < .001$, ** $P < .01$, * $P < .05$, ANOVA with Tukey's multiple comparison test).

**Fig. 4** Mean plots with standard error of soil microbial biomass carbon (A), basal respiration (B), and microbial metabolic quotient (C) for lawns with different tree species. Different letters indicate significant differences between groups (ANOVA with Tukey's multiple comparison test, $P < .05$).**Fig. 5** Heatmap of CLPP represented by respiration response of microbial community on different organic substrates (A), mean plots with standard error of microbial functional diversity (B), contribution microbial groups utilized carbohydrates, amino-, carboxylic- and phenolic acids to CLPP (C) for lawns with different tree species.

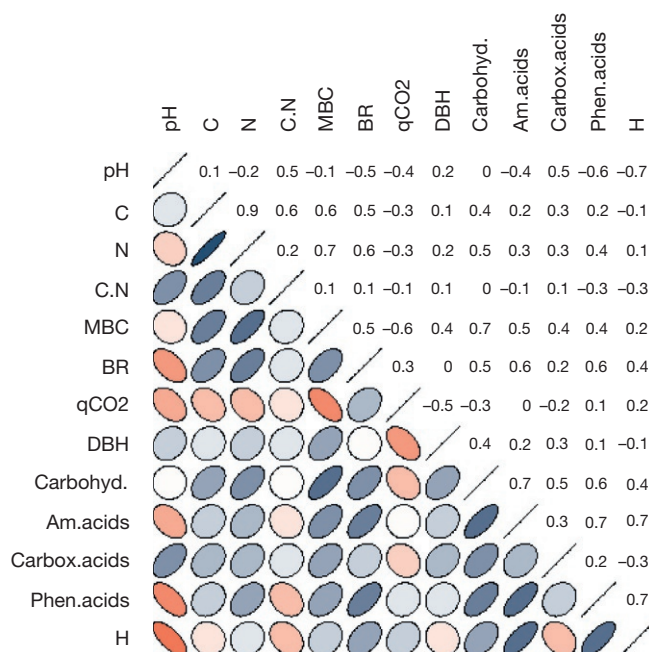


Fig. 6 Relationships between microbial (MBC, BR, qCO_2 , respiration response on carbohydrates, amino acids, phenolic acids, carboxylic acids), chemical (pH, C, N, C:N) properties and tree diameter at breast height, DBH ($n = 30$). Correlation matrix shows Spearman's correlation coefficients in upper triangle and ellipse (red and blue—negative and positive relationship) in lower triangle.

Thus, belonging tree to certain group (coniferous or deciduous) didn't contribute to microbial properties of urban soils. Instead, individual tree species and soil conditions (pH, C, N contents) determined the microbial biomass, microbial functional diversity and CLPP. Consequently, the most favorable conditions for microbial community (high biomass and diversity) were found for urban soil under *Betula*, whereas *Acer* determined the least favorable conditions for soil microbiome.

From landscape variability to soil microbial diversity: A case study of a tremendous green roof in Zaryadye Park (Moscow, Russia)

In recent years, there has been a growing trend in urban areas to incorporate green spaces using soil construction based on soils and soil substrates characteristic of various zonal ecosystems relocated to the centers of megacities (Brenneisen, 2006; Francis and Lorimer, 2011).

An exceptional example is Zaryadye Park, which was established in Moscow in 2017. This space can be described as a tremendous green roof built over underground infrastructures (exhibition halls, parking and technical facilities) with extensive utility systems. Despite its relatively small area of 10.2 ha, Zaryadye Park successfully replicates vegetation zones from the European part of Russia, including tundra, coniferous and broad-leaved forests, birch forests, meadows, and subtropics. Maintaining the growth and development of vegetation in these diverse landscapes, within a uniform climatic environment and on artificially created topography and substrates in the heart of a bustling metropolis, presents a unique and challenging task. It requires optimal soil and ecological conditions, including a diverse and specific microbiota that supports the requirements of each simulated landscape.

Studies on the soil microbiota of green roofs are limited, with vegetation diversity mainly associated with a thin fertile layer, emphasizing the fundamental role of soil microorganisms in its formation and maintenance (Brenneisen, 2006). The resilient and diverse soil microbiome plays a crucial role in the functioning of green roofs, contributing to improved vegetation cover, drought tolerance, pathogen protection, increased nutrient availability, salt tolerance, productivity, and substrate stabilization (John et al., 2014; Molineux et al., 2014; Ondoño et al., 2014; Fulthorpe et al., 2018) (Fig. 7). However, it is known that anthropogenic factors can significantly influence the evolution of the soil microbiome, which may be particularly evident in artificially created plant communities, leading to the formation of unique microbial communities.

Based on the available data, the dynamics of microbial biodiversity in urban soil ecosystems exhibit a multidirectional pattern. There is a noticeable decrease in the diversity and abundance of microscopic fungi compared to natural ecosystems, while bacterial communities tend to become more diverse (Abrego et al., 2020; Tedersoo et al., 2020; Donald et al., 2021). This shift in microbial composition is accompanied by a redistribution of bacterial community taxa, with an increase in the representation of fast-growing r-strategists, including human and plant pathogens (Donald et al., 2021). At the same time, data on the functional diversity of the urban soil microbial community, which is an important indicator of its quality, are limited, and its assessment is of particular interest.

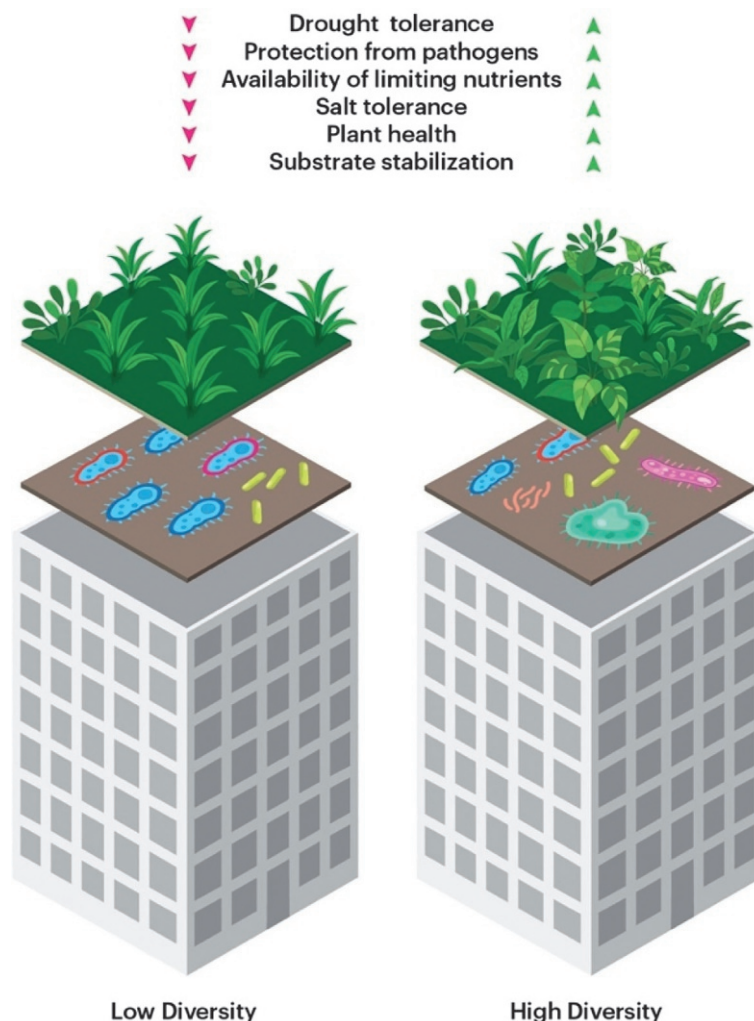


Fig. 7 The role of the soil microbiome in the formation and maintenance of green roofs (Fulthorpe et al., 2018). Zaryadye Park is an example of intensive green roof technology.

Soil survey in Zaryadye park, was carried out in summer 2022. Soil samples were collected from 0 to 10 cm in artificial green roofs with different vegetation types presented by zones: mountain tundra (MT), humid tundra (HT), pine forest (PF), spruce forest (SF), mixed forest (MF), birch forest (BF), steppe (S), floodplain (F), meadow (M), subtropical forest (SPF) and urban greening (UG). Disturbed and undisturbed areas were visually identified in each landscape type. A total of 22 zones were examined. CLPP and microbial functional diversity assessed using the methods described earlier in section “The effect of typology and composition of urban green infrastructures on soil microbiome for the case of university campus (RUDN University, Moscow)”.

The microbial communities of dissimilar natural zones demonstrated the highest physiological activity, manifested on a wider range of substrates: tundra, steppe, urban green areas, and the glass bark zone, in which plants of the subtropical zone are represented (Fig. 8A). The highest activity was observed in the tundra, which represents a landscape zone that is particularly challenging to replicate in Moscow due to its markedly different natural conditions. The revealed diversity of physiological groups of soil microorganisms indicates the existing potential of the microbial community for destructive activity not only in relation to readily available carbohydrates, but also in relation to substrates of amino acid groups, carboxylic and phenolic acids (Fig. 8B). Maximum functional diversity was also observed in undisturbed areas within the steppe landscape zone, as indicated by the Shannon index ($H = 2.47$) (Fig. 8C).

An analysis of the state of the park soils and their comparison with the parameters laid down at the opening of the park made it possible to conclude that the functional diversity depends on the content of organic matter. The tundra area, characterized by significantly higher levels of organic matter compared to the minimum acceptable values, displayed the most pronounced response to the introduction of various substrates (Fig. 9). The combined zones forming a single cluster represented less disturbed areas, further highlighting the impact of anthropogenic pressure intensity on the physiological activity of the microbial community.

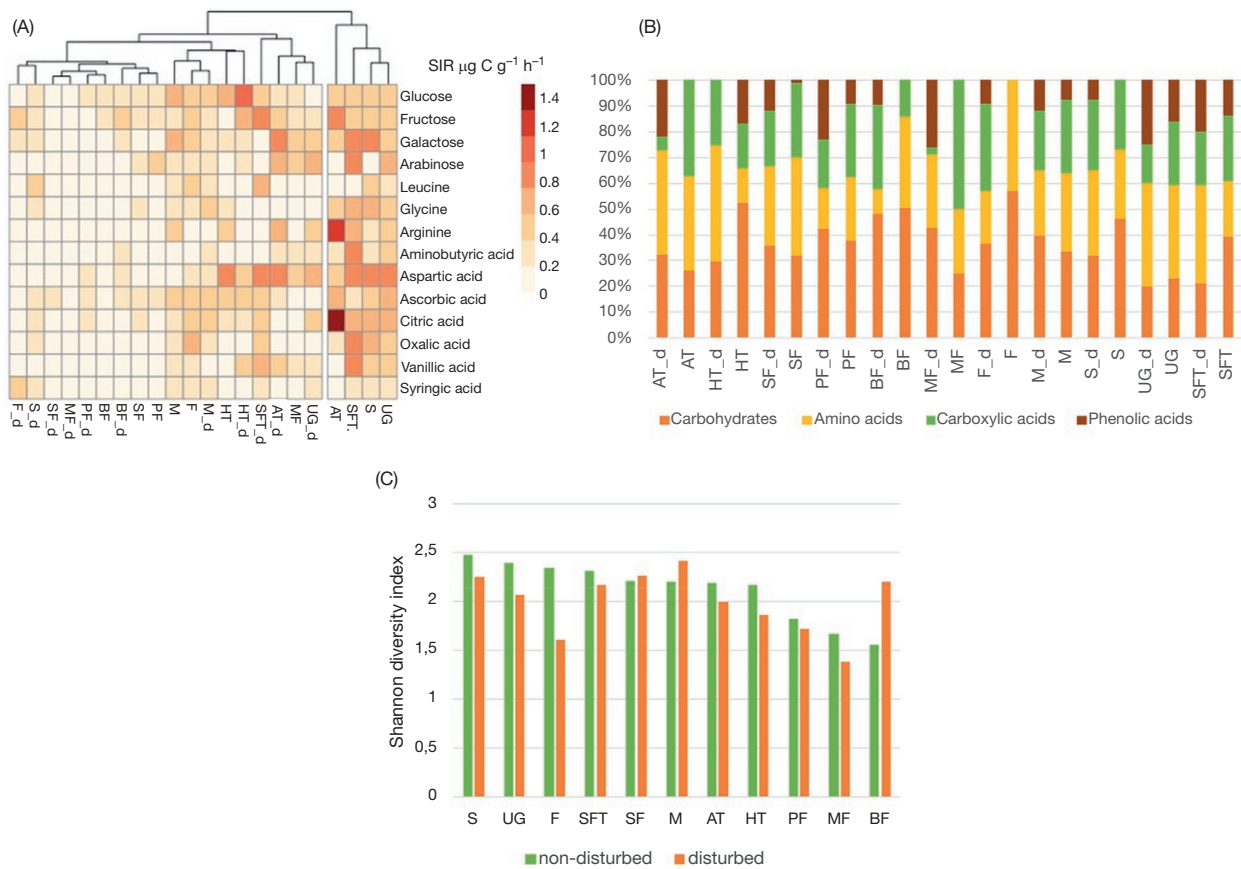


Fig. 8 Heatmap of CLPP represented by respiration response of microbial community on different organic substrates (A), microbial functional diversity (B), contribution microbial groups utilized carbohydrates, amino-, carboxylic- and phenolic acids to CLPP (C) for disturbed and non-disturbed areas of the landscape zones of Zaryadye Park. MT - alpine tundra; HT - humid tundra; PF - pine forest; SF - spruce forest; MF - mixed forest; BF - birch forest; S - steppe; F - floodplain; M - meadow; SPF - subtropical forest; UG - urban greening; d - disturbed area.

The microbial community of the park has patterns inherent in sustainable urban communities: reduction in species diversity and simplification of the structure of microscopic fungi, reduction in the number of cellulolytic fungi species, and redistribution of taxa towards potentially pathogenic species (Marfenina et al., 2011; Abrego et al., 2020). At the same time, the low share of actinomycetes, identified in the soil, on the contrary, indicates the initial stage of succession (Zviagintsev and Zenova, 2001).

These contradictions testify to the formation of unique communities of soil microorganisms in the park in which microbial successions can deviate classical assumptions. Perhaps, the territory of the park can be considered in the context of combining many ecotones-transitional environments between different plant communities (landscapes of the park), which enhances its role in the formation and conservation of biodiversity.

Arctic urban soil microbiome (Apatity, Russia)

The environmental consequences of urbanization are especially noticeable and challenging to evaluate in Arctic cities, where the restoration and development of ecosystems are constrained by climatic factors such as low temperatures and soil conditions such as nutrient deficiency (Goryachkin et al., 1999; Grosse et al., 2011).

The presence of organic substrates used in soil constructions in urban areas, combined with the warmer climate characteristic of “heat island” (urban area with significantly higher temperatures than surrounding rural areas), creates new ecological niches that facilitate the growth and development of soil microbial communities. For example, in the topsoil horizon of urban soils of Murmansk and Apatity, the microbial biomass and the number of copies of the rRNA genes of archaea, bacteria, and fungi were lower than in natural soils and amounted to (10^6 – 10^{10} , 10^9 – 10^{10} , and 10^7 – 10^9 , respectively). However, the opposite was shown for the subsoil horizons, where the neutral reaction, high carbon and nutrient content, and low pollution created favorable conditions for microbial development (Fig. 10). Similarly, total microbial biomass in urban soils (0.55–0.75 mg/g) was lower compared to 1.02 mg/g in background soil, while microbial biomass in urban subsoils was 2–2.5 times higher than in natural conditions. In addition, significant morphological changes were noted in microscopic fungi, with small spores predominating in urban soils, and mycelium predominating in forest soils. An increase in the diversity of microscopic fungi was also noted.

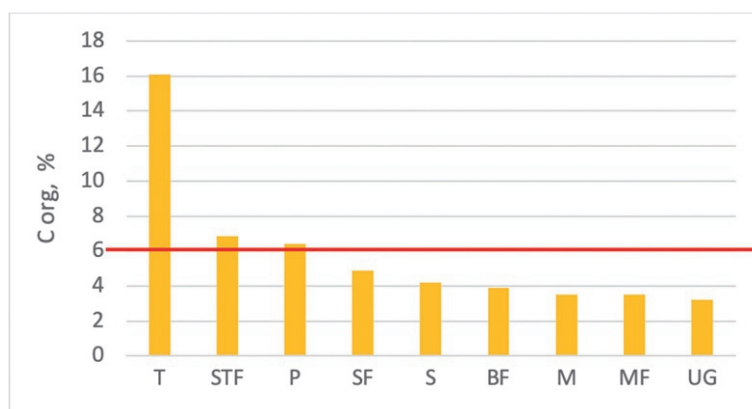


Fig. 9 Comparison of the content of organic matter in different landscapes. The red line represents the minimum permissible level of organic matter content in the soils during the creation of most landscapes. T – tundra; STF – subtropical forest; F – floodplain; SF – spruce forest; S – steppe; BF – birch forest; M – meadow; MF – mixed forest; SPF – subtropical forest; UG – urban greening.

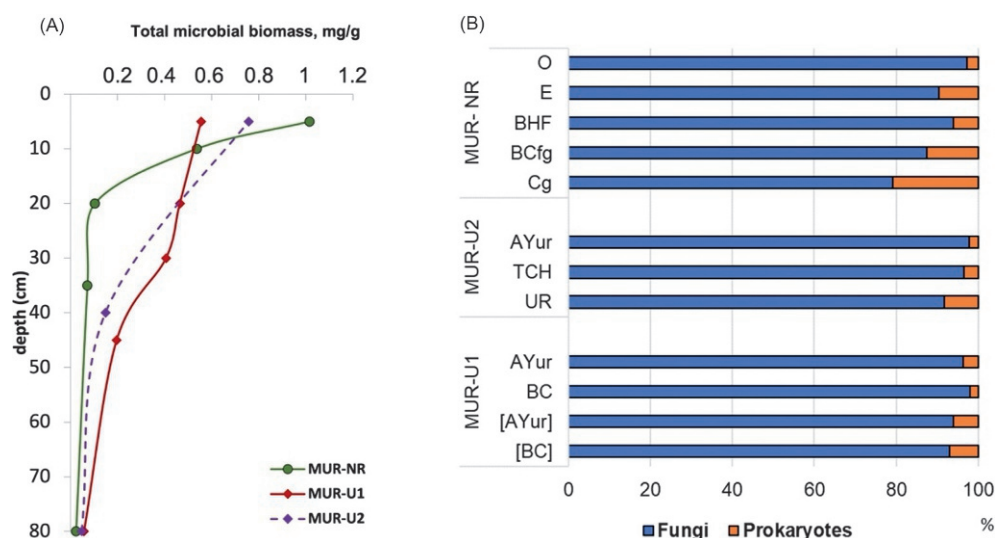


Fig. 10 Profile distribution of the total microbial biomass (A) and fungi: prokaryotes portion (B). MUR-U1, MUR-U2 – urban soil; MUR-NR – forest soil.

Arctic cities experience a strong anthropogenic load from industrial enterprises, the emissions of which have a strong negative impact on the state of terrestrial ecosystems. One of the brightest examples of such a city is Norilsk (Taimyr Peninsula), Monchegorsk (Kola Peninsula), etc. In the soils of such cities, microbiological activity is very suppressed. Among the cultivated bacteria, oligotrophs, adapted to existence in nutrient-poor soils, predominate. In general, soils are characterized by low numbers of microorganisms, biomass, and enzymatic activity compared to other northern regions that do not experience a strong anthropogenic load from industrial enterprises. However, compared with the background soils of the study area, in urban areas there was a slight increase in the biomass and diversity of soil bacteria and microscopic fungi, but at the same time, the proportion of opportunistic microfungi and bacteria of the *Escherichia coli* group, which can pose a danger to human health, increases.

In the soils of such cities, as a rule, there is a restructuring of the functional and species structure of the microbial community in comparison with the natural soils. For example, in different regions of Norilsk, soils were dominated by microorganisms capable of decomposing a group of amino acids and carbohydrates, while carboxylic acids dominated in natural soils. Among microfungi, melanin-containing species, which are known for their increased resistance to extreme conditions, were predominant.

When establishing Technosols, various types of organic materials in different ratios are commonly considered, primarily due to their availability (Ivashchenko et al., 2021). In the case of Apatity, peat, which is widely available, and mining wastes, which are industrial by-products, are frequently used as components in soil construction (Kremenetskaya et al., 2019).

During the monitoring of microbiological properties of Technosols of urban laws in Apatity for 26 months, the following trends were observed (Fig. 11).

For the microbial biomass carbon, there was an increase in all studies values (Table 3). Microbial activity increased during the first year of functioning, then there was a decrease and, accordingly, a return to the initial state. Specific respiration, which

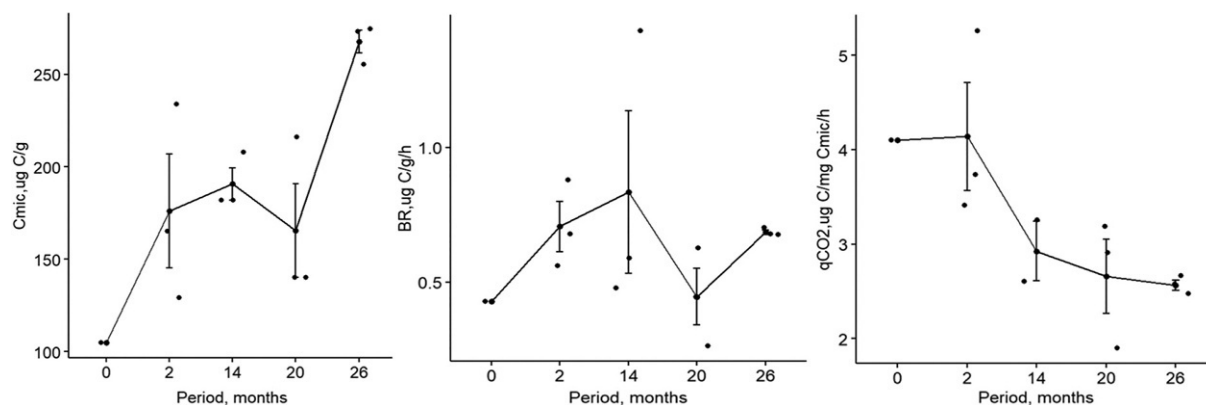


Fig. 11 Trend of microbial carbon indicators of Technosols of urban lawns in Apatity.

Table 3 The rate of increase in microbial parameters (comparison to zero moment).

Months	C_{mic}	BR	qCO_2
2	1.7	1.6	1.0
14	1.8	1.9	0.7
20	1.6	1.0	0.6
26	2.6	1.6	0.6
Result	Increased	Return to initial state	Decreased

characterizes the ecophysiological state of soil constructions, decreased over time, especially actively during the first year of functioning (2–14 months).

At the beginning of the functioning of the soil constructions, there was a significant variation in the observed values (Fig. 12). One year later, it decreased for microbial biomass carbon and specific respiration, and became strong for basal respiration. By the end of the observations for all parameters under examination, the variation was minor, which can probably be characterized as a more stable condition.

To characterize the state of the soil constructions, the indicators were compared with the values obtained from the analysis of the background soils, the state of which was the reference of sustainable functioning (Table 4).

For all parameters, a significant approximation to the background values was noted, and for the microbial metabolic quotient, the background values were reached, indicating the attainment of an equilibrium state.

In the functional structure of the microbial community in the soil structures of Apatity lawn, an increase in the consumption of substrates belonging to the carbohydrate group and a decrease in the consumption of carboxylic acids, which are the easiest and most accessible substrates for utilization, were observed (Fig. 13). Furthermore, there was an increase in the uptake of the more difficult-to-degrade phenolic acids.

Months	C_{mic}	BR	qCO_2
2	30	23	24
14	8	63	15
20	26	41	26
26	4	2	4

minor (<10)
middle (11-20)
significant (21-32)
strong (>32)

Fig. 12 Heatmap of the significance of variation coefficients at a different monitoring period.

Table 4 Comparison of microbial carbon indexes with background values.

Period	Apatity		
	C_{mic}	BR	qCO_2
0 moment	13.4 times lower	4.2 times lower	1.6 times higher
26 month	5.2 times lower	2.8 times lower	equal

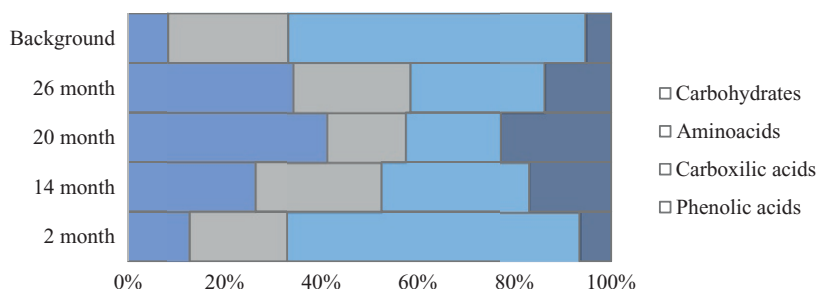


Fig. 13 Community level physiological profile in Apatity urban lawns soil constructions.

Arctic cities experience a strong anthropogenic load from industrial enterprises, which emit pollutants that strongly impact terrestrial ecosystems. Norilsk (Taimyr Peninsula) and Monchegorsk (Kola Peninsula) serve as prominent examples. In the soils of these cities, microbiological activity is heavily suppressed. Cultivated bacteria primarily consist of oligotrophs, specially adapted to survive in nutrient-depleted soils. In general, these soils exhibit lower microorganism counts, biomass, and enzymatic activity compared to other northern regions unaffected by intense industrial activities. However, when compared to the background soils in the study area, urban regions display a slight increase in both soil bacteria and microscopic fungi biomass and diversity. Nevertheless, this increase is accompanied by a higher proportion of opportunistic microfungi and bacteria from the *E. coli* group, which can pose risks to human health.

In the soils of such cities, as a rule, there is a restructuring of the functional and species structure of the microbial community in comparison with the natural soils. For example, in different regions of Norilsk, soils were dominated by microorganisms capable of decomposing a group of amino acids and carbohydrates, while carboxylic acids dominated in natural soils. Among microfungi, melanin-containing species, which are known for their increased resistance to extreme conditions, were predominant.

Urban green infrastructure can become a hotbed of microbial growth and soil diversity in a city from a polar region. This result should not be directly interpreted as a positive impact of urbanization on the soil microbiome, since the taxonomic and functional composition of microbial communities in urban soils can be very diverse. At the same time, urban soil is as a potential ecological niche for the development of the soil microbiome and a source of important ecosystem services, such as nutrient cycling and biodiversity control. The importance of ecosystem services and the harmful services provided by the urban microbiome will further increase after global warming scenarios and should be considered in sustainable soil management in Arctic cities.

Conclusion

In the urban context, the soil microbiome experiences a range of anthropogenic pressures, including contamination, land-use changes, and alterations in nutrient inputs. These factors impact the composition, diversity, and functioning of the soil microbiome. However, the extent and nature of these effects can vary greatly across different urban regions due to variations in urbanization patterns, local pollution sources, and management practices. For the development of a sustainable ecosystem in urban green areas, it is crucial for the microbiome to create conditions in terms of quantity, quality, and functionality.

The recognition of the importance of studying the soil microbiome has spurred significant advancements in methods for analyzing the microbiological properties of soil. Despite the active development of methods for analyzing the soil microbiome in the city, there are still no standards for soil microbiological properties, except those developed for human pathogens. The development of such standards must take into account the specifics of the region under study, including both natural and anthropogenic factors and is complicated by the variability of the results obtained and their non-obvious interpretation.

Currently, the assessment of urban soil is given by comparing its characteristics with nearby natural counterparts, using the main microbial indicators. These indicators provide valuable information on taxonomic diversity, community structure, microbial activity, and functional potential, enabling the characterization of ecosystem provisioning functions. However, the question remains open regarding the selection of the most informative indicators for assessing urban soil microbiomes.

Moreover, it is essential to advance our understanding of the relationships between microbial indicators and ecosystem functioning in urban soils. This includes investigating the contributions of urban soil microorganisms to vital processes such as nutrient cycling, organic matter decomposition, pollutant degradation, and plant-microbe interactions. By examining these relationships, we will be able to effectively evaluate the influence of the soil microbiome on the sustainability, resilience, and overall functionality of urban ecosystems.

Funding

The research was supported by RSF project #19-77-300-12.

References

- Abrego N, Crosier B, Somervuo P, Ivanova N, Abrahamyan A, Abdi A, et al. (2020) Fungal communities decline with urbanization-more in air than in soil. *The ISME Journal* 14: 2806–2815. <https://doi.org/10.1038/s41396-020-0732-1>.
- Ananyeva ND, Khatit RY, Ivashchenko KV, Sushko SV, Gorbacheva AY, Dolgikh AV, Kadulin MS, Sotnikova YL, Vasenev VI, Komarova AE, Yudina AV, and Dovletyarova EA (2023) Soil biophilic elements (C, N, P) and microbial activity in forest parks of Moscow and suburban forests. *Eurasian Soil Science* 56(1): 87–100.
- Attiwill PM and Adams MA (1993) Nutrient cycling in forests. *The New Phytologist* 124: 561–582.
- Augusto L, De Schrijver A, Vesterdal L, Smolander A, Prescott C, and Ranger J (2015) Influences of evergreen gymnosperm and deciduous angiosperm tree species on the functioning of temperate and boreal forests. *Biological Reviews* 90: 444–466.
- Barthel S, Isendahl C, and Visvanathan S (2019) Urban green commons: Insights on urban common property systems for ecosystem governance. *Ecosystem Services* 38: 100966.
- Basu S, Kumar G, Chhabra S, and Prasad R (2021) Role of soil microbes in biogeochemical cycle for enhancing soil fertility. In: *New and Future Developments in Microbial Biotechnology and Bioengineering*, pp. 149–157. Elsevier.
- Berg G, Köberl M, Rybakova D, Müller H, Grosch R, and Smalla K (2017) Plant microbial diversity is suggested as the key to future biocontrol and health trends. *FEMS Microbiology Ecology* 93(5), fix050.
- Beyer L, Blum H-P, Elsner D-C, and Willnow A (1995) Soil organic matter composition and microbial activity in urban soils. *The Science of the Total Environment* 168: 267–278. [https://doi.org/10.1016/0048-9697\(95\)04704-5](https://doi.org/10.1016/0048-9697(95)04704-5).
- Brenneisen S (2006) Space for urban wildlife: designing green roofs as habitats in Switzerland. *Urban Habitats* 4.
- Campbell CD, Chapman SJ, Cameron CM, Davidson MS, and Potts JM (2003) A rapid microtiter plate method to measure carbon dioxide evolved from carbon substrate amendments so as to determine the physiological profiles of soil microbial communities by using whole soil. *Applied and Environmental Microbiology* 69(6): 3593–3599. <https://doi.org/10.1128/AEM.69.6.3593-3599.2003>. PMID: 12788767. PMC161481.
- Cederlund H, et al. (2014) Soil carbon quality and nitrogen fertilization structure bacterial communities with predictable responses of major bacterial phyla. *Applied Soil Ecology* 84: 62–68.
- Chen F-S, Yavitt J, and Hu X-F (2014) Phosphorus enrichment helps increase soil carbon mineralization in vegetation along an urban-to-rural gradient, Nanchang, China. *Applied Soil Ecology* 75: 181–188. <https://doi.org/10.1016/j.apsoil.2013.11.011>.
- Chen L, Xiang W, Wu H, Ouyang S, Zhou B, Zeng Y, Chen Y, and Kuzyakov Y (2019) Tree species identity surpasses richness in affecting soil microbial richness and community composition in subtropical forests. *Soil Biology and Biochemistry* 130: 113–121.
- Diakhate S, Gueye M, Chevallier T, Diallo NH, Assigbetsé K, Abadie J, et al. (2016) Soil microbial functional capacity and diversity in a millet-shrub intercropping system of semi-arid Senegal. *Journal of Arid Environments* 129: 71–79.
- Dobrovolskaya TG, Zvyagintsev DG, Chernov IV, Golovchenko AV, Zenova GM, Lysak LV, et al. (2015) The role of microorganisms in the ecological functions of soils. *Eurasian Soil Science* 48: 959–967.
- Donald J, Muriene J, Chave J, Iribar A, Louisanna E, Manzi S, et al. (2021) Multi-taxa environmental DNA inventories reveal distinct taxonomic and functional diversity in urban tropical forest fragments. *Global Ecology and Conservation* 29: e01724. <https://doi.org/10.1016/j.gecco.2021.e01724>.
- Francis R and Lorimer J (2011) Urban reconciliation Ecology: The potential of living roofs and walls. *Journal of Environmental Management* 92: 1429–1437.
- Fulthorpe R, MacIvor JS, Jia P, and Yasui S-LE (2018) The green roof microbiome: improving plant survival for ecosystem service delivery. *Frontiers in Ecology and Evolution* 6: 5. <https://doi.org/10.3389/fevo.2018.00005>.
- Gill AS, Purnell K, Palmer MI, Stein J, and McGuire KL (2020) Microbial composition and functional diversity differ across urban green infrastructure types. *Frontiers in Microbiology* 11: 912. <https://doi.org/10.3389/fmicb.2020.00912>. PMID: 32582043. PMC7291602.
- Gómez-Brandón M, Herbón C, Probst M, Fornasier F, Barral MT, and Paradelo R (2022) Influence of land use on the microbiological properties of urban soils. *Applied Soil Ecology* 175: 104452.
- Goryachkin SV, Karavaeva NA, Targulian VO, and Glazov MV (1999) Arctic soils: Spatial distribution, zonality and transformation due to global change. *Permafrost and Periglacial Processes* 10(3): 235–250. [https://doi.org/10.1002/\(SICI\)1099-1530\(199907/09\)10:3<235::AID-PPP320>3.0.CO;2-4](https://doi.org/10.1002/(SICI)1099-1530(199907/09)10:3<235::AID-PPP320>3.0.CO;2-4).
- Grosse G, Harden J, Turetsky M, McGuire AD, Camill P, Tarnocai C, Frolking S, Schuur EAG, Jorgenson T, Marchenko S, Romanovsky V, Wickland KP, French N, Waldrop M, Bourgeau-Chavez L, and Striegl RG (2011) Vulnerability of high-latitude soil organic carbon in North America to disturbance. *Journal of Geophysical Research – Biogeosciences* 116(3), G00K06. <https://doi.org/10.1029/2010JG001507>.
- Ivashchenko K, Ananyeva N, Vasenev V, Sushko S, Seleznyova A, and Kudryarov V (2019) Microbial C-availability and organic matter decomposition in urban soils of megapolis depend on functional zoning. *Soil and Environment* 38(1): 31–41. <https://doi.org/10.25252/SE/19/6152>.
- Ivashchenko K, Lepore E, Vasenev V, Ananyeva N, Demina S, Khabibullina F, Vaseneva I, Selezneva A, Dolgikh A, Sushko S, Marinari S, and Dovletyarova E (2021) Assessing soil-like materials for ecosystem services provided by Constructed Technosols. *Landscape* 10: 1185. <https://doi.org/10.3390/land10111185>.
- Joergensen RG and Emmerling C (2006) Methods for evaluating human impact on soil microorganisms based on their activity, biomass, and diversity in agricultural soils. *Journal of Plant Nutrition and Soil Science* 169(3): 295–309.
- John J, Lundholm J, and Kernaghan G (2014) Colonization of green roof plants by mycorrhizal and root endophytic fungi. *Ecological Engineering* 71: 651–659. <https://doi.org/10.1016/j.ecoleng.2014.08.012>.
- Kiikkilä O, Kitunen V, and Smolander A (2006) Dissolved soil organic matter from surface organic horizons under birch and conifers: Degradation in relation to chemical characteristics. *Soil Biology and Biochemistry* 38: 737–746.
- Korneykova MV and Nikitin DA (2021) Qualitative and quantitative characteristics of the soil microbiome in the impact zone of the Kandalaksha aluminum smelter. *Eurasian Soil Science* 54(6): 897–906.
- Kremenetskaya I, Tereshchenko S, Alekseeva S, Mosendz I, Slukovskaya M, Ivanova L, and Mikhailova I (2019) Vermiculite-lizardite ameliorants from mining waste. *IOP Conference Series: Earth and Environmental Science* 368: 012027. <https://doi.org/10.1088/1755-1315/368/1/012027>.
- Li G, Sun G-X, Ren Y, Luo X-S, and Zhu Y-G (2018) Urban soil and human health: A review. *European Journal of Soil Science* 69: 196–215.
- Liu L, et al. (2022) Impact of urbanization on soil microbial diversity and composition in the megacity of Shanghai. *Land Degradation & Development* 33(2): 282–293.
- Lorenz K and Kandeler E (2006) Microbial biomass and activities in urban soils in two consecutive years. *Journal of Plant Nutrition and Soil Science* 169: 799–808. <https://doi.org/10.1002/jpln.200622001>.
- Marfenina OE, Makarova NV, and Ivanova AE (2011) Opportunistic fungi in soils and surface air of a megalopolis (for the Tushino Region, Moscow). *Microbiology* 80: 870–876. <https://doi.org/10.1134/S0026261711060142>.
- Mejia MP, Rojas CA, Curd E, et al. (2023) Soil microbial community composition and tolerance to contaminants in an urban brownfield site. *Microbial Ecology* 85: 998–1012. <https://doi.org/10.1007/s00248-022-02061-1>.
- Molineux CJ, Connop SP, and Gange AC (2014) Manipulating soil microbial communities in extensive green roof substrates. *Science of the Total Environment* 493: 632–638. <https://doi.org/10.1016/j.scitotenv.2014.06.045>.
- Ondoño S, Bastida F, and Moreno JL (2014) Microbiological and biochemical properties of artificial substrates: A preliminary study of its application as Technosols or as a basis in Green Roof Systems. *Ecological Engineering* 70: 189–199. <https://doi.org/10.1016/j.ecoleng.2014.05.003>.
- Pereira MC, O'Riordan R, and Stevens C (2021) Urban soil microbial community and microbial-related carbon storage are severely limited by sealing. *Journal of Soils and Sediments* 21: 1455–1465.

- Pett-Ridge J and Firestone MK (2005) Redox fluctuation structures microbial communities in a wet tropical soil. *Applied and Environmental Microbiology* 71: 6998–7007.
- Ramirez KS, et al. (2014) Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. *Proceedings of the Royal Society B: Biological Sciences* 281: 20141988.
- Rodríguez-Rodríguez JC, Fenton NJ, Bergeron Y, et al. (2023) Soil and tree phyllosphere microbial communities differ between coniferous and broadleaf deciduous boreal forests. *Plant and Soil*. <https://doi.org/10.1007/s11104-023-05959-y>.
- Schlöter M, Dilly O, and Munch JC (2003) Indicators for evaluating soil quality. *Agriculture, Ecosystems and Environment* 98(1–3): 255–262. [https://doi.org/10.1016/S0167-8809\(03\)00085-9](https://doi.org/10.1016/S0167-8809(03)00085-9).
- Schlöter M, Nannipieri P, Sørensen SJ, and van Elsas JD (2018) Microbial indicators for soil quality. *Biology and Fertility of Soils* 54: 1–10. <https://doi.org/10.1007/s00374-017-1248-3>.
- Scholier T, Lavrinienko A, Brila I, Tukalenko E, Hindström R, Vasylenko A, Cayol C, Ecke F, Singh NJ, Forsman JT, Tolvanen A, Matala J, Huitu O, Kallio ER, Koskela E, Mappes T, and Watts PC (2023) Urban forest soils harbour distinct and more diverse communities of bacteria and fungi compared to less disturbed forest soils. *Molecular Ecology* 32(2): 504–517. <https://doi.org/10.1111/mec.16754>.
- Tan X, Kan L, Su Z, Liu X, and Zhang L (2019) The composition and diversity of soil bacterial and fungal communities along an urban-to-rural gradient in South China. *Forests* 10: 797.
- Tedersoo L, Anslan S, Bahram M, Drenkhan R, Pritsch K, Buegger F, et al. (2020) Regional-scale in-depth analysis of soil fungal diversity reveals strong pH and plant species effects in Northern Europe. *Frontiers in Microbiology* 11: 1953. <https://doi.org/10.3389/fmicb.2020.01953>.
- Urbanová M, Šnajdr J, and Baldrian P (2015) Composition of fungal and bacterial communities in forest litter and soil is largely determined by dominant trees. *Soil Biology and Biochemistry* 84: 53–64.
- Wang Z, Tao T, Wang Y, Small GE, Chen J, and Sun X (2023) Soil quality in urban forests under different understory management practices. *Land Degradation & Development* 34(3): 899–910.
- Wuyts K, Smets W, Lebeer S, and Samson R (2020) Green infrastructure and atmospheric pollution shape diversity and composition of phyllosphere bacterial communities in an urban landscape. *FEMS Microbiology Ecology* 96(1), fiz173.
- Xu H-J, Li S, Su J-Q, Nie SA, Gibson V, Li H, and Zhu Y-G (2014) Does urbanization shape bacterial community composition in urban park soils? A case study in 16 representative Chinese cities based on the pyrosequencing method. *FEMS Microbiology Ecology* 87: 182–192.
- Yan B, Li J, Xiao N, Qi Y, Fu G, Liu G, and Qiao M (2016) Urban-development-induced changes in the diversity and composition of the soil bacterial community in Beijing. *Scientific Reports* 6: 38811.
- Yang J-L and Zhang G-L (2015) Formation, characteristics, and eco-environmental implications of urban soils—A review. *Soil Science and Plant Nutrition* 61: 30–46.
- Zhao D, Li F, Yang Q, Wang R, Song Y, and Tao Y (2013) The influence of different types of urban land use on soil microbial biomass and functional diversity in Beijing, China. *Soil Use and Management* 29: 230–239. <https://doi.org/10.1111/sum.12034>.
- Zviagintsev DG and Zenova GM (2001) *Ecology of Actinomycetes*. Moscow: GEOS. [in Russian].
- Zviagintsev DG (1987) In: Zviagintsev DG (ed.) *Soil and Microorganisms*. Moscow: Publishing House of Moscow State University.

Further reading

- Adhikari K and Hartemink AE (2016) Linking soils to ecosystem services—A global review. *Geoderma* 262: 101–111.
- Becker J and Kuzyakov Y (2018) Teatime on Mount Kilimanjaro: Assessing climate and land use effects on litter decomposition and stabilization using the Tea Bag Index. *Land Degradation & Development* 29(8): 2321–2329. <https://doi.org/10.1002/ldr.2982>.
- Blagodatskaya EV and Anderson T-H (1998) Interactive effects of pH and substrate quality on the fungal-to-bacterial ratio and qCO₂ of microbial communities in forest soils. *Soil Biology and Biochemistry* 30: 1269–1274.
- Blagodatskaya E and Kuzyakov Y (2013) Active microorganisms in soil: Critical review of estimation criteria and approaches. *Soil Biology and Biochemistry* 67: 192–211.
- Boeraeve M, Honnay O, and Jacquemyn H (2019) Local abiotic conditions are more important than landscape context for structuring arbuscular mycorrhizal fungal communities in the roots of a forest herb. *Oecologia* 190: 149–157.
- Delgado-Baquerizo M, Eldridge DJ, Liu YR, Sokoya B, Wang JT, Hu HW, et al. (2021) Global homogenization of the structure and function in the soil microbiome of urban greenspaces. *Science Advances* 7: eabg5809. <https://doi.org/10.1126/sciadv.abg5809>.
- Dilly O (2005) Microbial energetics in soils. In: Buscot F and Varma A (eds.) *Microorganisms in Soils: Roles in Genesis and Functions*. Berlin: Springer.
- Dilly O and Munch JC (1998) Ratios between estimates of microbial biomass content and microbial activity in soils. *Biology and Fertility of Soils* 27: 374–379.
- Docherty KM, Pearce DS, Lemmer KM, and Hale RL (2018) Distributing regionally, distinguishing locally: Examining the underlying effects of local land use on airborne bacterial biodiversity. *Environmental Microbiology* 20: 3529–3542.
- Faber JH, Creamer RE, Mulder C, Römbke J, Rutgers M, Sousa JP, Stone D, and Griffiths BS (2013) The practicalities and pitfalls of establishing a policy-relevant and cost-effective soil biological monitoring scheme. *Integrated Environmental Assessment and Management* 9: 276–284.
- Gattinger A, Ruser R, Schlöter M, and Munch JC (2002) Microbial community structure varies in different soil zones of a potato field. *Journal of Plant Nutrition and Soil Science* 165: 421–428.
- He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, Huang Z, Wu W, Gu B, Jardine P, et al. (2007) GeoChip: A comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *The ISME Journal* 1: 67–77.
- Hedlund K, Griffiths B, Christensen S, Scheu S, Setälä H, Tscharntke T, and Verhoef H (2004) Trophic interactions in changing landscapes: Responses of soil food webs. *Basic and Applied Ecology* 5: 495–503.
- Keuskamp JA, Dingemans BJJ, Lehtinen T, Searnel JM, and Hefting MM (2013) Tea Bag Index: A novel approach to collect uniform decomposition data across ecosystems. *Methods in Ecology and Evolution* 4: 1070–1075.
- Kim M, Morrison M, and Yu Z (2011) Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *Journal of Microbiological Methods* 84(1): 81–87. <https://doi.org/10.1016/j.mimet.2010.10.020>. PMID: 21047533.
- Liu YR, van der Heijden MGA, Riedo J, et al. (2023) Soil contamination in nearby natural areas mirrors that in urban greenspaces worldwide. *Nature Communications* 14: 1706. <https://doi.org/10.1038/s41467-023-37428-6>.
- Lu J and Scheu S (2020) Response of soil microbial communities to mixed forests of European beech and conifers: Variations with site conditions.
- Medriano CA, Chan A, De Sotro T, and Bae S (2023) Different types of land use influence soil physiochemical properties, the abundance of nitrifying bacteria, and microbial interactions in tropical urban soil. *Science of the Total Environment* 869: 161722. <https://doi.org/10.1016/j.scitotenv.2023.161722>.
- Mishustin EN, Pertsovskaya MI, and Gorbov VA (1979) *Sanitary Microbiology of Soil*. Moscow: Nauka. [in Russian].
- Mueller P, Schille-Beers LM, Mozdzer TJ, Chmura GL, Dinter T, Kuzyakov Y, de Groot AV, Esselink P, Smit C, D'Alpaos A, Ibáñez C, Lazarus M, Neumeier U, Johnson BJ, Baldwin AH, Yarwood SH, Montemayor DI, Yang Z, Wu J, Jensen K, and Nolte S (2018) Global-change effects on early-stage decomposition processes in tidal wetlands—implications from a global survey using standardized litter. *Biogeosciences* 15: 3189–3202. <https://doi.org/10.5194/bg-15-3189-2018>.
- Pulleman M, Creamer R, Hamer U, Helder J, Pelose C, Peres G, and Rutgers M (2012) Soil biodiversity, biological indicators and soil ecosystem services—An overview of European approaches. *Current Opinion in Environment Sustainability* 4: 529–538.

- Shi Z, Yin H, Van Nostrand JD, et al. (2019) Functional gene array-based ultrasensitive and quantitative detection of microbial populations in complex communities. *mSystems* 4: e00296-19. <https://doi.org/10.1128/mSystems.00296-19>.
- Thiele-Bruhn S, Schlöter M, Wilke B-M, Beaudette LA, Martin-Laurent F, Cheviron N, Mougin C, and Römbke J (2019) Identification of new microbial functional standards for soil quality assessment. *The Soil* 6(1): 17–34. <https://doi.org/10.5194/soil-2019-42>.
- Vasenev VI, Smagin AV, Ananyeva ND, et al. (2017) Urban soil's functions: Monitoring, assessment, and management. In: Rakshit A, Abhilash P, Singh H, and Ghosh S (eds.) *Adaptive Soil Management: From Theory to Practices*, pp. 359–409. Singapore: Springer.
- Wu S, Xiong J, and Yu Y (2015) Taxonomic resolutions based on 18S rRNA genes: A case study of subclass copepoda. *PLoS One* 10(6), e0131498.
- Xiang Q, Chen Q-L, Zhu D, An X-L, Yang X-R, Su J-Q, Qiao M, and Zhu Y-G (2018) Spatial and temporal distribution of antibiotic resistomes in a Peri-Urban area is associated significantly with anthropogenic activities. *Environmental Pollution* 235: 525–533.

Microbiology of extreme soil environments

Patrice Dion, Institut de biologie Intégrative et des systèmes (IBIS), Pavillon Charles-Eugène Marchand, Laval University, Québec, QC, Canada

© 2023 Elsevier Ltd. All rights reserved.

Introduction	495
Extreme soils and extreme microbes	496
Microbial adaptive potential to extreme soils	496
All soils are extreme, but some soils are more extreme than others	496
Microbial adaptation mechanisms in extreme soils	496
Carbon and nitrogen metabolism in extreme soils	500
Biological networks and dispersal: Examples of extreme soils	501
Cold soils	501
Presentation of cold soils and of their microbial communities	501
Microbial mechanisms for cold tolerance	503
Saline and hot desert soils	504
Sandy deserts	504
Rocky deserts and lithic microbial colonization	504
Hot soils	505
Volcanic and geothermal soils	505
Fire-affected soils	505
Saline and alkaline soils	506
Radioactive soils	506
Subsurface environments	506
Viruses in extreme soils	506
Biogeography of extreme soil microbial colonizers	507
Importance of extreme soil colonizing microorganisms	507
Contributions of extreme soil microorganisms to climate change	507
Contributions to astrobiology	507
Contributions to biodiversity conservation	508
Contributions to industry, agriculture, and medicine	508
Conclusion	508
References	509

Abstract

In soil, microbial community structure is greatly influenced by physical and chemical components, especially nutrients and toxic ions. When located at either ends of the physicochemical range of properties, soils are called “extreme” and tend to be oligotrophic environments or characterized by strong nutrient imbalances. Their colonization depends on the interplay between the spatial and temporal structure of environmental restrictiveness and the microbial adaptive potential, presence of biological networks and microbe movement to new habitats. Colonization is multiphasic and carried out by soil generalists, by endemic specialists, and by survivalists. Microorganisms inhabiting polar regions are critically involved in regulating Earth’s climate and carbon balance.

Key points

- The ability of microorganisms to thrive in inhospitable environments leads us to expand the boundaries of life and enter uncharted territory, where life itself may appear strange and foreign.
- Dispersal makes microbes ubiquitous, but selection by general and extreme soil characteristics decides their fate through strategies for survival that range from active growth to death.
- To overcome environmental stress, microorganisms can adopt strategies of compensation, conservation, protection, damage repair, and stress mitigation.
- Mechanisms for survival include development of DNA with high preponderance of specific base pairings, enhanced thermal stability and efficient repair mechanisms, a high level of crucial gene redundancy and genome plasticity with up-regulation of environmentally relevant genes, proteins with modified flexibility, changes in membrane composition, including amount and type of lipids, and regulation of fluidity.
- Knowledge of the microbiology of extreme soil environments has resulted in a vast field of practical applications, such as solvent and polymer production, agricultural enhancements, pollutant removal, and clean energy production.
- Microorganisms that inhabit polar regions are critically involved in regulating the Earth’s climate and carbon balance.

Introduction

Soils of the Earth represent a diverse set of heterogeneous and discontinuous ecosystems, fragmented into an incommensurate number of microhabitats, both discrete and variably interconnected. In these systems, self-organization (Young and Crawford, 2004) occurs at different scales, from the nanometer to the kilometer range, in a fractal mode sustained by biological, physical, and chemical interfacial processes. Depending of their location at any particular moment, 10^{30} bacterial and archaeal cells, which themselves can be grouped into more than 10^{12} species (Locey and Lennon, 2016), can be introduced into soil with a varying level of probability about their future. There, their fate, characterized as either growth, survival in the dormant state or death, will be determined by the interplay between their own physiology, morphology and genetics and the self-organizational processes occurring in that soil.

The impact of soil heterogeneity on diversity of colonizing microbial communities has been studied through a variety of approaches, one of these being to examine the microbiology of soils located at the extremes of the physicochemical range of soil properties. Such soils are called “extreme soils”. A microorganism introduced in an extreme soil will be subjected to the stressful conditions normally occurring in soil, and to an additional, intense, constant, and intrinsic selection pressure. This additional selection pressure arises because an extreme soil harbors one or several defining characters that superimpose a supplementary chemical or physical challenge to common obstacles faced by soil microbial colonizers. The microbial colonization patterns of extreme soils are defined by the interplay between the spatial and temporal structure of environmental restrictiveness with the: (i) microbial adaptive potential; (ii) biological networks; and (iii) dispersal, defined here as movement of microbes leading to their introduction in a new soil habitat (Fig. 1).

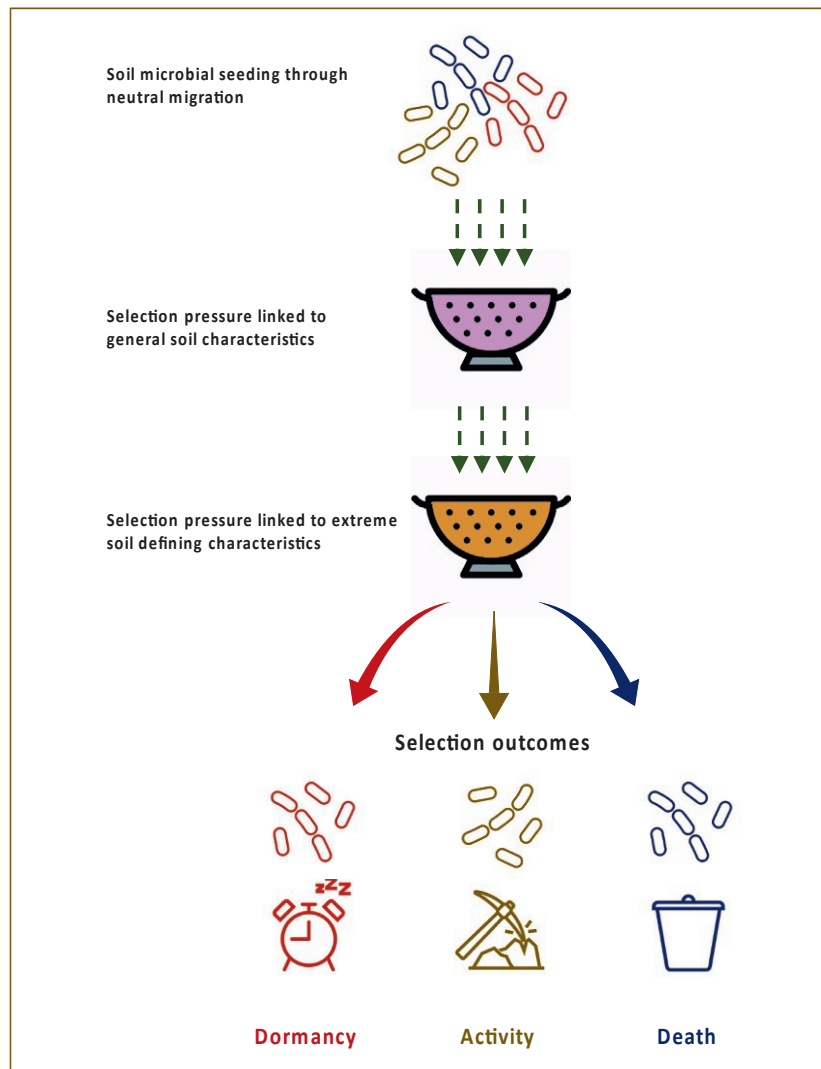


Fig. 1 Dispersal makes microbes ubiquitous, but selection by general and extreme soil characteristics decides their fate.

Extreme soils and extreme microbes

Soil microbial community structure is markedly influenced by soil organic carbon content and pH, together with other factors such as temperature, sand proportion, water content, electrical conductivity, as well as the phosphorus, sulfur, calcium, magnesium, iron, and aluminium content. These physicochemical properties vary widely, at scales ranging from the micrometer up to continental dimensions (Tecon and Or, 2017). In extreme soils, one or several soil environmental restrictions are pushed to an extreme value, defining the soil itself and the microbial community that inhabits it. All active members of an extreme soil community demonstrate genetic, physiological, and morphological adaptation to this defining character, in the form of resistance or tolerance. Other soil properties contribute to structure the microbial community, since adaptation to the defining character is not sufficient, by itself, to confer on a microorganism the capacity to inhabit a particular extreme soil, either as a dormant organism or else as active cells (Vera-Gargallo et al., 2019).

Extreme soils belong to a variety of Reference soil groups as defined by the World Reference Base, and it is not possible to establish an unequivocal and complete correlation between certain soil groups and the extreme character. Rather, those soils are typified as possessing one defining character or a set of defining characteristics, which are identified, not only by their intensity, but also by the fact that they are constantly expressed, or at least are maintained over extended time periods (Table 1).

Microbial adaptive potential to extreme soils

All soils are extreme, but some soils are more extreme than others

Harsh conditions, including low nutrient status, fluctuating water availability, occurrence of phenolics or other toxic compounds, as well as intra and interspecific competition, generally prevail in soils (Barberán et al., 2014). As decades of attempts at reliably introducing microbial inoculants in soil have shown, efficient microbial colonization of soil requires specific adaptive traits that are narrowly distributed. Colonization is a multiphasic phenomenon, carried out by soil generalists occurring in large numbers at a variety of sites and in different soil types, by specialists with a high degree of endemism, and by survivalists such as *Escherichia coli* (Moynihan et al., 2015; Van Elsas et al., 2011), capable of maintaining populations in soils for extended time periods following their transport from a different, preferential, habitat. Survival may be achieved through dormancy, characterized by a combination of reduced or abolished metabolism and maintenance of cellular integrity. The ultimate option open to soil microbes is death, whereby cell content is made available to other organisms. These descriptions of microbial strategies do not represent discrete categories, but rather compose a fluctuating continuum of adaptation levels. Microorganisms transition along this continuum because of spatial and nutritional heterogeneity (Joergensen and Wichern, 2018; Lennon and Jones, 2011), of soil community composition (Moynihan et al., 2015), and thanks to genomic evolution (Jang et al., 2017). Members of the microbial community locate themselves along this continuum, with the sum of microbial positions defining an overall community balance along a stress scale defining death, dormancy and growth strategies (Fierer, 2017). Extreme soil defining characters tip the overall community balance against growth and competition strategies, and in favor of stress tolerance (Fig. 2).

Successful soil microbes are found at both numerous and environmentally diverse sites. Efficient microbial soil colonizers share traits related to genome structure and functional potential. For example, bacteria with larger genomes and more metabolic and functional capabilities are also more successful at soil colonization (Barberán et al., 2014). On top of the regular traits common to all efficient soil colonizers, microbial activity in extreme soils is contingent on adaptation to a set of defining characteristics and on expression of specialized physiological responses (see Table 1 above). Thus, to the defining characteristics of extreme soils correspond equally defining properties of active microbial colonizers. Whereas extremophily is often associated with archaea, one lesson that can be extracted from the study of extreme soil microbiology is that extreme soils harbor microbial communities akin to those of normal soils, in terms of the general representation of the three domains of life. Archaea, bacteria, and eukaryotic microbes coexist in extreme soils as they do in normal soils. In particular, fungi, either unicellular or in filamentous form, and lichenized or non-lichenized, occur abundantly in all types of extreme soils and it has been suggested that they may be more resistant than bacteria under certain conditions prevalent in these environments (Dragone et al., 2020).

Being constantly present, extreme soil defining characteristics determine a corresponding stability in the colonizing microbial communities. Shaped by a constant historical environment, soil microbial assemblages become sensitive to change, leading to poorer functional acclimatization capacity as compared to communities inhabiting more fluctuating environments (Bardgett and Caruso, 2020). Microbial transformations in extreme soils result from activities of narrowly adapted specialists, existing in environments of low carrying capacity. When changes in the intensity of extreme soil defining characteristics occur, extremophile microbial communities express some level of resistance, resilience and functional redundancy, as do other communities from normal soils (Allison and Martiny, 2008). However, defining characteristics of extreme soils directly alter microbial survival and thus exert a strong negative impact on diversity or abundance of soil biology. Because functional redundancy may be low in extreme soil environments, and because of the rarity of changes (see above), microbial communities colonizing extreme soils may be more fragile and less resilient than communities found in other soils.

Microbial adaptation mechanisms in extreme soils

Agriculture, forestry and recreational activities showcase soils as life-sustaining devices (FAO and ITPS, 2015). Indeed, even the most barren soils generally harbor microbial life. Although no signs of life were detected in a particular region of Antarctica, facing a

Table 1 Extreme soils and their defining characteristics.

Extreme soil type	Defining characteristic	Other bio-limiting characteristics	Examples of microorganisms active in extreme soil			Characters of microorganisms	References ^a
			Bacteria	Archaea	Eukaryotes		
Defining character driven by climate							
Cold soils	Presence of permafrost below an active soil layer	Freeze-thaw cycles, desiccation, starvation, radiation, UV exposure	Psychrophilic and psychrotrophic bacteria; anaerobic bacteria: sulfate reducers, Fe(III) reducers and denitrifiers; methanotrophs	Acetoclastic and hydrogenotrophic methanogens	Fungi, including yeasts	Optimal growth temperature around 15 °C (growth down to –15 °C)	De Maayer et al. (2014) and Jansson and Tås (2014)
Hot arid soils	Ratio of precipitation to potential evapotranspiration of 0.05–0.2 for arid soils and of less than 0.05 for hyperarid soils	Extreme fluctuations in temperature, low nutrient status, high levels of ultraviolet radiation, strong winds	Communities dependent on microbial primary producers, such as cyanobacteria; Bacteroidetes and halophilic bacteria; aerobic heterotrophs from the Terrabacteria superphylum, including Actinobacteria and Chloroflexi; trace gas oxidizers	<i>Halobacteriaceae</i>	Green algae; melanin-containing fungi	Biological soil crust formation; hypolithic or endolithic colonization; presence of lichens	Bay et al. (2018), Jung et al. (2020) and Makhalanyane et al. (2015)
Defining character driven by soil chemistry and water content							
Hypersaline soils	Electrical conductivity of a saturated soil extract >15 dS m ^{–1}	Temperature, pH, nutrient availability, solar radiation	Moderate and extreme halophiles: Bacteroidetes, <i>Salinibacter</i> Balneolaeota, Rhodothermaeota; cyanobacteria of the <i>Halothece</i> cluster	Halophilic archaea: <i>Natronomonas</i> , <i>Halococcus</i> , <i>Halobacterium</i>	Black yeasts, e.g., <i>Hortaea werneckii</i> , filamentous fungi, e.g., <i>Cladosporium herbarum</i> ; <i>Chlorophyceae</i> , e.g., <i>Dunaliella</i>	Growth of halophiles at 2–5 M NaCl; water potential –1.5 to –40 Mpa; colonization of rock habitats in hypersaline arid soils	Crits-Christoph et al. (2016), Musa et al. (2018) and Vera-Gargallo et al. (2019)
Peatland soils	Flooded ecosystem where the production rate of biomass exceeds its decomposition rate	Anaerobiosis in deeper horizons; high acidity	Vertically stratified bacterial community; Bacteria, including methanotrophs of the Proteobacteria and Verrucomicrobia phyla; N-fixing <i>Burkholderia</i>	Archaea, including methanogens in the deep waterlogged horizon, or catotelm	Fungi, especially in the upper layer, or acrotelm: saprophytes, including polyphenol-degraders; mycorrhizal symbionts; green algae	Species-specific association between yeast and <i>Sphagnum</i> plants; unique acidophilic methanotrophs	Andersen et al. (2013)
Defining character driven by soil physics							
Deep subsurface	Environment characterized by darkness and anaerobiosis, where temperature and pressure increase with depth	Scarce water and organic matter, substrates limited to dissolved minerals; colonization dependent on fractures or porosity in rocks	Microbial hydrogen generation though fermentation, nitrogen fixation and other reactions; hydrogen consumption through acetogenesis, sulfur and iron reduction and other reactions; hydrogen-utilizing cyanobacteria as primary producers; Alphaproteo-bacteria, Bacteroidetes	Hydrogen consumption through methanogenesis	Ascomycota	Anaerobic oxidation of hydrogen using extracellular electron acceptors such as iron and manganese oxides, or phenolic compounds	Escudero et al. (2018) and Gregory et al. (2019)

(Continued)

Table 1 (Continued)

Extreme soil type	Defining characteristic	Other bio-limiting characteristics	Examples of microorganisms active in extreme soil			Characters of microorganisms	References
			Bacteria	Archaea	Eukaryotes		
Volcanic and geothermal soils	Ambient temperature 60–80 °C	Low pH, high temperatures, CO ₂ , CH ₄ , H ₂ S, H ₂ gas emissions	Thermophiles: <i>Geobacillus thermoleovorans</i> , <i>Rubrobacter spartanus</i>	<i>Thermoproteus thermophilus</i> , methanogenic archaea	Thermophilic fungi: <i>Thermomyces</i> , <i>Chaetomium</i> , <i>Mycothermus</i>	Use of H ₂ as electron donor and for hydrogenotrophic methanogenesis	de Oliveira and Rodrigues (2019) and Picone et al. (2020)
Defining character resulting from human activity Fire-affected soils	Temperature raised up to 700 °C transiently (for surface fires) or for long periods (underground fires)	Altered moisture, increase in carbon and mineral nutrient availability	Decreased prevalence of heterotrophic bacteria, increased prevalence of autotrophic bacteria; presence of thermophiles	Crenarchaeota; Marine Benthic Group A; Thermoprotei	Fungi are more sensitive to fire than bacteria	Enrichment in bacteria with small genomes	Lee et al. (2017), Pressler et al. (2019) and Sorensen et al. (2019)
Radioactive soils	Chronic ionizing radiation from γ , β , and α emitters at total dose rates of 5 μ Gy h ⁻¹ or more; these levels of radiation do not occur in natural soils	Other stressors common in soils and altering radiation resistance: non-optimal temperature and nutrient levels, toxic chemicals, UV exposure, desiccation, interspecific competition	A wide variety of bacteria, e.g., <i>Methylobacterium</i> , <i>Acinetobacter</i> , <i>Chroococcidiopsis</i> , <i>Deinococcus</i>	<i>Halobacterium</i> , <i>Thermococcus</i> , <i>Pyrococcus</i>	Fungi, e.g., <i>Penicillium</i> sp., <i>Cladosporium sphaerospermum</i> , <i>Aureobasidium pullulans</i>	Non-enzymatic antioxidants, physiological adjustments to combat oxidative stress, efficient DNA repair systems	Battista (2015), Gostinčar et al. (2015) and Shuryak (2019)

^aSee also general references: (Rothschild and Mancinelli, 2001; Sánchez-Otero et al., 2019; Singh et al., 2019).

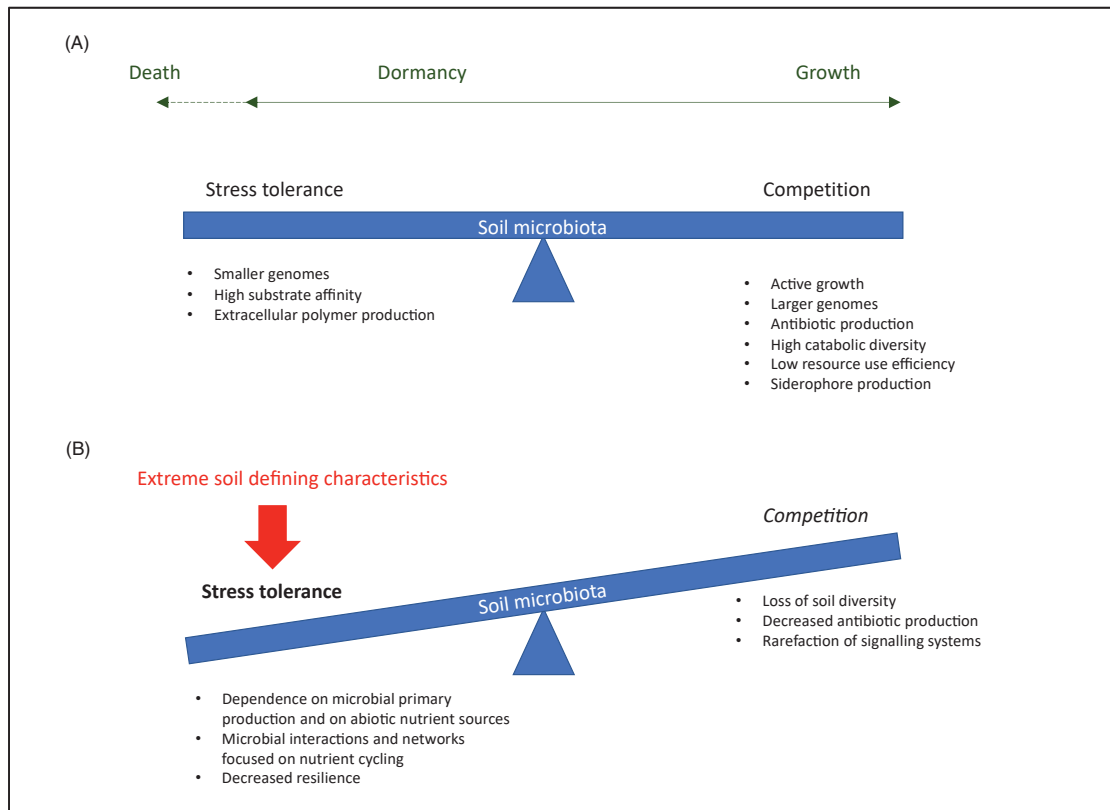


Fig. 2 Simplified view of soil microbiota, and of their growth and survival strategies. Strategies are arranged along a continuum (double-headed lines on top of the Figure going from near death dormancy to active growth). Soil microbiota, illustrated as a bar placed in equilibrium, are comprised of components performing activities or harboring properties related either to stress tolerance, or else to competition. (A) In normal soils, a fluctuating equilibrium exists between stress tolerance and competition. Properties attributed to each of these two microbial strategies are indicated below the bar representing the soil microbiota. (B) Extreme soil defining characters tip the balance in favor of stress tolerance. This results in loss of diversity, decrease in competition-related activities, increased dependence on microbial-driven nutrient cycling, and expression of various other community properties as indicated.

unique combination of cold temperatures, low water potentials, and high salt concentrations (Dragone et al., 2020), microorganisms have devised means to cope with most extreme soil defining conditions or sets of defining conditions.

Microorganisms resort to five possible strategies to overcome environmental stress: compensation, conservation, protection, damage repair, and stress mitigation.

(i) During the *compensatory response*, microorganisms seek to restore equilibrium and to maintain normal functions. For example, some halophiles adopt the “salt-in” strategy, where osmotic balance is achieved by accumulating high intracellular concentrations of inorganic salts, such as KCl (Gunde-Cimerman et al., 2018). (ii) When resorting to the *conservation response*, microorganisms enter dormancy. Whereas endospore formation represents a form of absolute dormancy, other forms are also possible, including differentiated resting structures such as arthrospores, conidia, cysts or akinetes. Many organisms can enter physiological dormancy without obvious morphological differentiation; this is a very common response, since 90% of soil microorganisms are dormant (Lennon and Jones, 2011). (iii) *Protective responses* are used to maintain the physical integrity of living organisms. Desiccation may cause water loss resulting in cell-volume collapse and in serious damage to cellular macromolecules such as proteins and nucleic acids. To protect themselves against desiccation, many bacterial cells accumulate compatible solutes, including carbohydrates, amino acids, quaternary amines, and tetrahydropyrimidines. (iv) Capacity for *damage repair* is exemplified by *Deinococcaceae*, which show an ability to withstand the lethal effects of DNA-damaging agents and to repair hundreds of DNA double and single-strand breaks as well as other types of DNA damages following desiccation or gamma irradiation (Heulin et al., 2017). (v) The *mitigation response* occurs when microorganisms avoid excessive stress exposure by choosing favorable niches within their heterogeneous environment, where they encounter conditions favorable to survival and growth.

Irrespective of their capacity to express one or several of these strategies, extremophiles and non-extremophiles co-exist in all types of soils. Thermophiles are ubiquitous in soils, since upper soil layers provide temporal niches for their activity (Santana and Gonzalez, 2015). Various examples also exist of moderate halophiles being isolated from non-saline soils (Wu et al., 2020). Whether they remain active or dormant in these temperate habitats, these resistant or tolerant organisms constitute a ‘microbial seed bank’ from which extreme soil communities will be structured upon development of an extreme character. Conversely, the observation of recovery of microbial communities to their original state following cooling of coal fire-heated soils (Lee et al.,

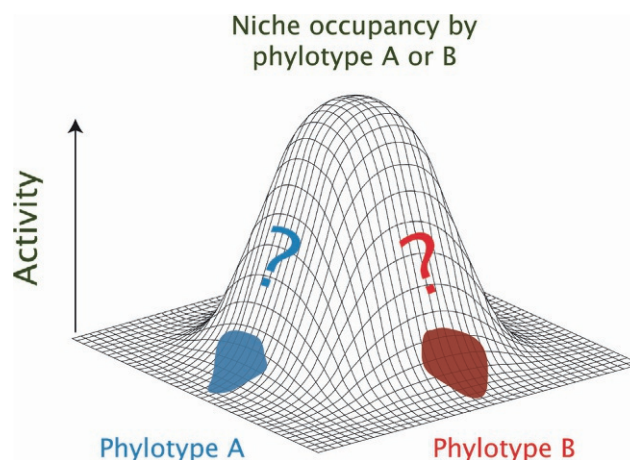


Fig. 3 Stochastic resuscitation model for extreme soil colonization. The diagram is constructed in a manner analogous to the adaptive landscape architecture, with activity levels being depicted in lieu of fitness levels as is usual in adaptive landscape representations. Through aeolian or other modes of dispersal as proposed by Bass Becking, phylotypes A and B are introduced as dormant populations in a newly formed extreme soil environment. The stochastic model states that either phylotype A or phylotype B will be resuscitated first in any given microenvironment, with this faster phylotype stably occupying the niche characterized by extreme physicochemical conditions and low carrying capacity.

2017) suggests that non extremophile microbes are constantly reintroduced in some extreme soils from neighboring locations. They may then encounter death, inactivation, or dormancy in their new extreme habitat, or else perform successful adaptation.

The combination of the lethal or inactivating potential of extreme soil defining characters and of the existence in soils of microbial seed banks comprised of dormant microbes awaiting resuscitation may be at the origin of a stochastic founding effect (Lee et al., 2017), responsible for the large beta- and gamma-diversity observed between different extreme soils sharing the same defining character. According to this hypothesis, extreme soil community structure would be largely determined by the kinetics of resuscitation among extremophiles comprising the microbial seed banks. Those microorganisms that are reactivated first in a newly formed extreme soil become firmly and stably established in this environment, where a lack of antagonisms and signaling activity leads to long-term or quasi-permanent colonization patterns. Fig. 3 provides a graphic depiction of the stochastic founding effect prevalent in newly formed extreme soil environments. This view is inspired by adaptive landscape representations (Cavrilets, 2008), with phylotype activity replacing genotype fitness as the varying property of microbial communities.

In addition to the wide variation of community structure of extreme soils sharing similar properties, this stochastic founding effect of newly formed extreme soils would allow microorganisms capable of deep dormancy, such as the spore-forming *Bacillaceae*, *Chloroflexi* and *Dormibacteraeota*, to occur frequently in various types of extreme soils, where they can perform physiological adaptation. Other microorganisms harboring a resting stage, such as *Actinobacteria* and *Cyanobacteria*, are also frequent extreme soil colonizers. Table 2 presents examples of adaptation mechanisms displayed by microbial colonizers of extreme soils.

The same survival strategy may be adopted as a microbial response that would be common to a range of environmental challenges. Table 3 provides examples of multipurpose stress resistance strategies, each of which is adopted by extremophiles in a variety of contexts. In a modular assembly process, multipurpose strategies are associated in environmentally-specified combinations, to constitute the integrated physiological, ecological, and genetic tolerance identity of individual organisms.

Carbon and nitrogen metabolism in extreme soils

Because of the essentially total absence of plants or large animals, most extreme soils are oligotrophic environments, or at least are characterized by strong nutrient imbalances. Carbon and nitrogen are supplied by microbial primary producers, or else come from scant and inanimate external sources (Table 4). One consequence of this, apart from low cell numbers and prevalence of dormancy, may well be the slow pace of microbial community succession (Knelman et al., 2014). Thus, initial community assembly, determined by random microbial dispersal (Herbold et al., 2014) and the stochastic founding effect presented above, would result in non-convergent community compositions over relatively short geographic ranges, with a high degree of phylotype endemism. This pattern would be particularly striking in the case of fragmented habitats, such as rocks supporting *endo*- or *hypolith*ic colonization in hot (Chan et al., 2012) or cold (Cowan et al., 2014) deserts, or else ice-wedge polygons that are geometric formations comprised of a combination of ice, bog and water body (van Dorst et al., 2017).

In normal soils, atmospheric H_2 and CO oxidizers are members of the rare biosphere, each comprising about 1% of the total bacterial community. However, in various extreme environments, such as Antarctic deserts (Bay et al., 2021) and deep subsurface environments (Gregory et al., 2019), trace gas oxidizers are the main primary producers in soils with otherwise low photosynthetic input (see also below).

Table 2 Examples of adaptation mechanisms expressed by microbial colonizers of extreme soils.

Type of extreme soil	Category of adaptation			References
	Environmental	Molecular	Physiological	
Cold soils	Colonization of protected niches or “refugia”, as occurs for hypolithic communities; dormancy	Genome: high G + C content of DNA; high level of crucial gene redundancy; high genome plasticity. Proteins: increased flexibility; increased enzymatic activity; decreased thermostability. Up-regulation of genes encoding cold-shock proteins	Increased membrane fluidity; optimized membrane function; accumulation of cryoprotectants, and of antifreeze and ice nucleation proteins; accumulation of trehalose	De Maayer et al. (2014)
Saline soils	Endolithic and hypolithic colonization by cyanobacteria and other microorganisms provides access to water through various mechanisms, including halite deliquescence and extraction of crystallization water from gypsum rock	Highly acidic proteome; changes in membrane composition and regulation of fluidity	Two alternative strategies: “salt-in strategy” with the intracellular accumulation of high concentrations of inorganic salts, mainly KCl; “compatible solute strategy”, or accumulation of organic compounds compatible with cell processes (glycerol often found in eukaryotes, glycine betaine in prokaryotes)	Crits-Christoph et al. (2016), Gunde-Cimerman et al. (2018) and Huang et al. (2020)
Hot soils	Fungal thermophily is an adaptation to transient seasonal and diurnal high temperatures; high soil surface and phyllosphere temperatures contribute to presence of thermophiles in cool soils	Proteins with increased ion-pair content and disulfide bonds, forming high-order oligomers and with decreased flexibility at room temperature. Adjustment in membrane composition, including amount and type of lipids (including ether-linked lipids). Enhanced thermal stability of DNA due to DNA-binding proteins and association with monovalent and divalent salts. High G + C content of ribosomal and transfer RNAs and presence of modified nucleosides	Presence of adaptive base excision DNA-repair system; high ploidy level; fast RNA turnover; ability to cope with thermolabile metabolites by metabolic channeling or other mechanisms; accumulation of compatible solutes	Gerday and Glansdorff (2007), Godon et al. (2020), Powell et al. (2012), Rothschild and Mancinelli (2001) and Santana and Gonzalez (2015)
Arid soils	For cyanobacteria, localization at the soil surface or below the surface depending on humidity levels; formation of biological soil crusts	Enhanced capacity for DNA repair in <i>Deinococcus</i>	Accumulation of compatible solutes as osmoprotectants; synthesis of extracellular polymers; cell differentiation into resting stages (endospores, akinetes)	Ferrenberg et al. (2017) and Heulin et al. (2017)
Radioactive soils	Protection against ionizing radiation is a consequence of adaptation to desiccation and other stresses	Efficient DNA repair mechanisms (repair of single-strand and double-strand breaks, and of base modification)	Protection against oxidative injury, notably by maintaining a high intracellular Mn/Fe ratio	Pavlopoulou et al. (2016)

Biological networks and dispersal: Examples of extreme soils

Cold soils

Presentation of cold soils and of their microbial communities

Permafrost is defined as ground, comprised of soil, sediment, or rock, with inclusion of ice and organic material, that remains at or below 0 °C for at least two consecutive years. During the short arctic summer, the surface zone of permafrost sediments thaws, giving rise to the active layer. Active layer depths range from a few cm in the high Arctic to more than 2 m in subarctic regions. Permafrost can be cemented by ice, as occurs in the Arctic, or it may be dry, as in the Antarctic polar deserts or rocky areas where interstitial water is lacking (Wagner, 2008).

Table 3 Examples of multipurpose physiological, ecological, and genetic strategies in extremophiles.

Type of multipurpose strategy	Strategy provides resistance to:	References
Endolithic colonization	Desiccation under hot or cold conditions	Cary et al. (2010) and Wierzbos et al. (2018)
Dormancy	General stress	Lennon and Jones (2011)
Accumulation of compatible solutes	Salinity, desiccation, heat ^a	Empadinhas and da Costa (2008), Santos et al. (2007) and Schmid et al. (2020)
DNA repair mechanisms	Desiccation, heat, ionizing radiation ^b	Pavlopoulou et al. (2016) and Ranawat and Rawat (2017)
Melanin and mycosporine-like amino acid production by fungi, cyanobacteria, and other microorganisms	Excessive heat or cold, extreme pH or osmotic conditions, UV and ionizing radiations	Gessler et al. (2014) and Gostinčar et al. (2015)

^aSome compatible solutes are restricted to hyperthermophiles; in some organisms, heat resistance is associated with radiation resistance.

^bIonizing radiation resistance is believed to arise as a by-product of desiccation and heat resistance involving DNA repair mechanisms and protection against reactive oxygen species.

Table 4 Metabolism of carbon and nitrogen in some examples of extreme soils.

Type of extreme soil	Metabolism of		References
	Carbon	Nitrogen	
Cold soils (permafrost layer)	Utilization of plant and animal residues accumulated over thousands of years	Nitrogen fixation with organic carbon as energy source	Mackelprang et al. (2011)
Cold soils (active layer)	Methanogenesis at sub-zero temperatures Degradation of organic carbon diffusing from permafrost, with production of carbon dioxide and methane Methanotrophy in the presence of oxygen	Assimilatory nitrate reduction Dissimilatory nitrate reduction, with production of molecular nitrogen and nitrous oxide intermediate	Mackelprang et al. (2011)
Cold desert soils (Antarctic soils)	Autotrophic carbon fixation; scavenging of H ₂ and CO from the atmosphere for use as energy sources	Functional capacity for nitrogen fixation in hypolithic communities	Cowan et al. (2014) and Ji et al. (2017)
Arid soils (halite endolith in the Atacama Desert)	Primary organic carbon production by photosynthetic cyanobacteria (<i>Chroococcidiopsis</i>) Possibility of photoheterotrophy by archaea with light-driven proton pumps Degradation of organic carbon by heterotrophic bacteria and archaea Genes for the 3-hydroxypropionate/4-hydroxybutyrate carbon fixation pathway in <i>Thaumarchaeota</i>	No nitrogen fixation; acquisition of bioavailable nitrogen from atmospherically-derived nitrate Genes for ammonia oxidation in <i>Thaumarchaeota</i>	Crits-Christoph et al. (2016) and Finstad et al. (2017)
Saline and alkaline soils	CO ₂ fixation through various pathways, including Calvin and reductive TCA cycles	Aerobic nitrate reduction; decreased nitrogen fixation	Hwang et al. (2020) Dendooven et al. (2010), Navarro-Noya et al. (2016) and Ren et al. (2018)
Geothermal soils	Chemolithoautotrophy, with carbon fixation using H ₂ and H ₂ S released by volcanic activity Methanotrophy from methane released by volcanic systems, occurring in soils with temperature <50 °C, pH >2.5 and low H ₂ S and H ₂ concentrations (due to competition with chemoautotrophic communities)	Nitrogenase genes detected; temperature upper limit for denitrification higher than for nitrogen fixation and nitrification	Burr et al. (2005), Gagliano et al. (2020) and Picone et al. (2020)
Deep continental biosphere (subsoil)	Primary organic carbon production by hydrogen-utilizing, autotrophic sulfate-reducing bacteria Utilization of small organic carbon molecules originating from serpentinization reactions of organic-rich graphite zones, by sulfate-reducing bacteria	Possibility of nitrogen fixation in deep subsoil, with hydrogen production	Lollar et al. (2019), Gregory et al. (2019) and Swanner and Templeton (2011)

Cold environments span more than 70% of the Earth and represent one of the largest biospheres. Microorganisms dominate these cold environments, and as such they represent the most abundant cold-adapted life forms on Earth. Among cold soils, the least hospitable are the cold desert soils of Antarctica and parts of northern Greenland. Of its 13,661,000 km², only 0.36% of the Antarctic continent is ice-free, and the ice-free areas consist primarily of permafrost. Mean air temperatures in January range from close to 0 °C in coastal areas to about -27 °C in inland zones. However, surface temperatures during summer may be much warmer. They are frequently above 0 °C for several days at a time and reach 20 °C, although wide fluctuations are common. Moisture supply is from snowfall or drifting snow, and yearly precipitation varies between 400 mm for wetter and warmer coastal areas to 20 mm or

even less in inland and colder zones. Cold desert soils have a slightly acid to alkaline pH reaction and high salinity, with salts that are principally chlorides, nitrates and sulfates of sodium, potassium, and magnesium. In comparison to Antarctica, the Arctic region climate is less extreme, supporting more plant and animal life. In both the Arctic and Antarctica, extended periods of darkness occur during winter months, while high UV radiation is experienced during summer.

Microorganisms thrive in cold climates, with phylum-level compositions of many microbial communities in Antarctic soils, occurring for example in the Dry Valleys and the Antarctic Peninsula, often equivalent to that of temperate soils. Bacterial phyla that are most abundant in temperate soils, such as Proteobacteria, Actinobacteria, Bacteroidetes and Firmicutes, are also dominant in polar soils. However, high specialization occurs at the species level, and diversity is reduced in the more extreme Antarctic environments. Microbial communities in polar soils are largely endemic and vary in distribution patterns over meter, kilometer, regional and continental scales. Various key biotic and abiotic factors, in particular pH, C/N ratio, assimilable N and phosphorus concentrations, as well as plant cover, influence polar microbial communities in these extreme environments (van Dorst et al., 2017). In Antarctic soils, many microorganisms are thought to principally exist in a dormant state, with metabolic energy directed towards cell maintenance rather than growth (Ji et al., 2017).

Many ice-free surface soils in Antarctica have been exposed for millions of years and have remained largely unchanged over that time. These soils have generally been unaffected by direct anthropogenic impacts and are often very isolated. Even if microbes can be dispersed to remote soil patches via aeolian transport, they often face a unique combination of cold temperatures, extremely low soil water potentials, and high salt concentrations that restrict the activity and survival of all but a few specifically adapted taxa. Some isolated Antarctic soils may have remained effectively uninhabited by microbes due to these constraints. Indeed, microbes could not be detected in many of the driest, higher elevation Antarctic soils, using cultivation-dependent, cultivation-independent, and metabolic assays (Dragone et al., 2020). Fungi are frequently found in Antarctic soils, including some soils from where no bacteria could be cultured.

In oligotrophic Antarctic soils, aeolian processes drive recruitment from less hostile microenvironments, and niche processes subsequently select for the most adapted microorganisms (Ji et al., 2017). The microbial community structure of Antarctic soils is shaped by selection for bacteria that can persist in these physically extreme, chemically deprived environments. Microbial phototrophs, including cyanobacteria and algae, are in low abundance, their development probably being restricted by climatic conditions such as low water and nutrient availability, summer radiation, and winter darkness. The barren soils of eastern Antarctica contain a particularly large proportion of chloroflexi, actinobacteria, and unknown, previously uncultured taxa (van Dorst et al., 2017). Most inhabitants are dormant aerobic bacteria that support energy generation and, in some cases, carbon fixation by oxidizing atmospheric trace gases. Antarctic surface soil microbial communities have the capacity to scavenge H_2 and CO from the atmosphere for use as energy sources, and CO_2 as a carbon source. Bacteria from the phyla Actinobacteria, *Dormibacteraeota*, and *Eremiobacteraeota* oxidize atmospheric H_2 and CO at rates that are theoretically sufficient to sustain the energy needs of their largely dormant microbial populations. Atmospheric H_2 and CO are dependable energy sources for dormant bacteria, as they are strong reductants, diffuse through cell membranes, and occur at low concentrations throughout the troposphere.

Primary production driven by metabolism of trace atmospheric gases in the absence of endogenous photosynthesizing organisms also occurs in other cold environments, besides Antarctic cold deserts. The Atacama Desert, in Northern Chile, comprises high-elevation volcano debris fields, with oligotrophic mineral soils. There, traces gases such as H_2 , CO and several organic C1 compounds are energy sources for chemoautotrophic microbes that then supply organic carbon to simple and low energy-flux communities (Lynch et al., 2014).

Microbial mechanisms for cold tolerance

Table 2 (see above) distinguishes three categories of mechanisms for physicochemical stress resistance. These will now be discussed in more detail for the specific case of cold tolerance.

- (i) *Environmental adaptation* occurs when psychrophilic communities selectively colonize protected niches or “refugia”, notably those provided by translucent rocks or by halites of Antarctic deserts. Rock-colonizing, also referred to as “endolithic”, microbial populations are protected from physical abrasion by wind-blown sand and buffered from rapid thermal fluctuations, despite being in equilibrium with the external temperature for much of the Antarctic year. Endolithic environments host extensive and varied microbial communities, some of which are dominated by cyanobacteria, such as *Chroococcidiopsis* and *Gloeocapsa*, Actinobacteria, Alphaproteobacteria, and *Deinococcus* spp. Others are composed primarily of eukaryotic algae, melanized fungi, lichens and heterotrophic bacteria (Cary et al., 2010; Wierzbosch et al., 2012). Composition of bacterial and fungal communities are influenced by rock type and by site-specific characteristics, such as local climate, and water abundance and chemistry (Choe et al., 2018; Walker and Pace, 2007).
- (ii) In *molecular adaptation*, genome and protein structure harbors various traits favorable to life at low temperatures. These include high G + C content, high level of redundancy in crucial genomic regions, and high genome plasticity conferred by an abundance of plasmids and other mobile genetic elements. Proteins of psychrophiles also show increased flexibility and enzymatic activity at low temperature. Differential gene expression in response to cold temperatures results in up-regulation of genes encoding cold-shock proteins. These act as regulators of various cellular processes, including transcription, translation, protein folding and membrane fluidity, thus leading to physiological adaptation (see below).
- (iii) Through *physiological adaptation*, psychrophiles increase membrane fluidity and optimize membrane function. This is achieved, for example, by increasing the ratio of polyunsaturated to saturated fatty acids in membrane phospholipids, and by inserting

carotenoid pigments as membrane fluidity regulators. To prevent protein denaturation, scavenge free radicals and stabilize proteins, cold-adapted bacteria accumulate cryoprotectants and antifreeze proteins in their cytoplasm, as well as ice-nucleating proteins and trehalose (De Maayer et al., 2014).

Fungi adapt to cold temperatures through increased production of unsaturated phospholipids in the cell membrane, maintaining membrane fluidity. Cryotolerant fungi may also accumulate cryoprotectants such as glycerol (Osborne et al., 2020). Mycosporines and mycosporine-like amino acids protect fungi against UV radiation and oxidative stress and possibly act as compatible solutes, together with trehalose, polyalcohols, betaine and proline (Gostinčar et al., 2015).

Bacteria and fungi colonizing cold soils produce a variety of pigments as a protective strategy against stresses such as low temperature, oxidative stress, and ultraviolet radiation. Carotenoids, prodigiosin, and indigoidine are examples of pigments made by bacteria and fungi inhabiting Antarctic and other cold soils. In the Antarctic deserts, rock-inhabiting fungi and certain bacteria produce melanin pigments that protect cells from extreme cold and heat, polychromatic UV radiations, extreme pH and osmotic conditions (Sajjad et al., 2020).

An abundant supply of organic carbon is required for microbial community response to temperature fluctuations in cold soils. In nutrient-poor soils, oligotrophic microorganisms remain strictly dependent on one another and thus establish temperature-independent nutritional networks. On the contrary, an abundance of soil organic carbon provides for an increased flexibility in community composition, which is then constantly and rapidly optimized as a function of current values of environmental variables.

Saline and hot desert soils

Deserts represent the single largest terrestrial ecosystem type on Earth, covering ~33.6% of the global land mass, excluding Antarctica, and are classified in terms of their aridity index, a ratio between precipitation (P) and potential of evapotranspiration (PET). This results in four desert categories: dry-semiarid ($0.5 < P/PET < 0.65$), semiarid ($0.2 < P/PET < 0.5$), arid ($0.05 < P/PET < 0.2$), and hyperarid ($P/PET < 0.05$). Hyperarid deserts generally receive annual precipitation of ≤ 70 mm and their soils display high (~ 7 to 9) pH, high salinity levels, high surface radiation fluxes, long periods of desiccation, and low water activity (Zablocki et al., 2016).

Extended periods of desiccation and oligotrophy are characteristics of hyperarid desert soils. Under these environmental constraints, microbial populations often form biofilms, where cells embedded in extracellular polymeric substance matrices are adsorbed to particle surfaces (Zablocki et al., 2016).

Within arid environments, relatively moist sites harbor richer and larger microbial communities, demonstrating the intense selective pressure of aridity (Osborne et al., 2020). However, desiccation may not be the sole or even the primary factor influencing microbial life in desert environments, where high salt and oxidizing agent concentrations are also crucial factors. An abrupt increase in water availability in hyperarid soils is extremely harmful to xerotolerant microorganisms because these cells are induced to transition from the dormant state to the metabolically active state while being exposed to attack from extreme temperature and UV radiation. In addition, excessive water causes high osmotic shock through the microbial semipermeable membrane and disturbs survival strategies adapted for limited moisture. After a heavy rainfall event in 2017 at the core region of the Atacama Desert, 75–87% of pre-rainfall species were undetectable, and no viable archaea or eukaryote were detected in undrained brines. Within the phylum *Deinococcus-Thermus*, dominance shifted from that of the order *Thermales*, who are mostly thermoresistant, to a dominance of the order *Deinococcales*, who are both thermoresistant and radioresistant (Shen et al., 2021).

Sandy deserts

Firmicutes (such as *Bacillus* and *Paenibacillus*) and Actinobacteria (such as *Arthrobacter* and *Geodermatophilus*) represent the dominant bacterial communities in the Sahara and Namib sandy deserts. Proteobacteria belonging to the four subgroups (alpha, beta, gamma, and delta) are diverse, whereas *Myxococcus* and *Azotobacter* are absent. Rhizobia occur in sandy deserts, especially strains able to nodulate the legume tree *Acacia*. Cyanobacteria are common inhabitants of sandy desert soils. For instance, the filamentous cyanobacterium *Oscillatoria* can migrate to the soil surface in response to wetting events or retreat below the surface in response to drying events (Heulin et al., 2017). Although diverse, microbial populations are limited in numbers, since microbial growth is restricted by the lack of water and oligotrophic conditions.

The absence of thermotolerant bacteria from Sahara sampling sites suggests that tolerance to desiccation is more important than tolerance to high temperature as an adaptive microbial trait to sandy desert conditions (Heulin et al., 2017).

Rocky deserts and lithic microbial colonization

The hyperarid region of the Atacama Desert receives less than 2 mm of precipitation annually (Huang et al., 2020). Together with desiccation, the intense solar irradiation limits microbial survival and growth in this environment. To overcome these extreme conditions, microbial consortia find refuge within or below rock substrates. Rock-inhabiting microbes benefit from the attenuation and filtration of the intense solar irradiation. Furthermore, lithic habitats capture liquid water from incoming fog formations and from water vapor condensation (Gómez-Silva, 2018).

Quartz, sandstone, granite, limestone, anhydrite, gypsum, ignimbrite and halites are lithic substrates that sustain microbial communities in the Atacama Desert. These niches are colonized by organisms designated as lithobionts, constituting epilithic,

endolithic and hypolithic microbial communities of unculturable members of the Archaea, Bacteria, and Eukarya domains. Abundance and diversity of the lithobionts depend on water availability, with a water gradient being established from the coastal areas to inland zones of the desert (Gómez-Silva, 2018).

Lithobiont microbial communities are usually dominated by photosynthetic cyanobacteria and contain a wide diversity of members of the phyla Actinobacteria, Acidobacteria, Bacteroidetes, and Proteobacteria. Although green algae are present and may contribute to the photosynthetic fixation of carbon, the main role in light harvesting is fulfilled by cyanobacteria (Zablocki et al., 2016).

Halites are evaporitic rocks that contain more than 95% NaCl, and as such they can be considered as poly-extreme habitats, coupling hyper-salinity with high temperatures, extremely low water activity, and high levels of UV radiation. As halites are typically colorless or white, they are partially translucent to solar radiation, and consequently photosynthesis can occur in their interiors. As a result of this photosynthetic activity, a carbohydrate-based nutrient environment exists within these rocks, which sustains heterotrophic microbial life (Gómez-Silva et al., 2019). To survive, microorganisms take advantage of the fact that halite deliquesces at an atmospheric relative humidity greater than 75%, as frequently happens due to marine fog intrusions, creating a highly saline solution in its mineral pores (Finstad et al., 2017). Microbial communities of halites are largely dominated by Halobacteriales and Bacteroidetes, together with algal and cyanobacteria species. Halite microbiomes are more similar to salt-adapted microbiomes than to those from other desert ecosystems (Gómez-Silva et al., 2019).

Gypsum rock-colonizing microbes also provide an example of an extreme survival strategy. Some cyanobacteria extract water of crystallization from this substrate, inducing a phase transformation from gypsum ($\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$) to anhydrite (CaSO_4) (Huang et al., 2020).

Endolithic communities harbor low diversity because of the extreme nature of their natural environment, with species richness and diversity being positively correlated to the availability of liquid water (Crits-Christoph et al., 2016). Nitrogen fixation cannot be detected in these arid, high-salt environments. Because productivity and growth in these environments are very low, abiotic sources of nitrogen are sufficient to sustain the communities. It has been hypothesized that atmospherically deposited nitrate and ammonium are the main sources of nitrogen for these communities (Finstad et al., 2017).

Halophilic eukaryotes, including fungi and protozoa (Oren, 2008), also occur in hot desert soils. Protozoa may control the abundance of prokaryotes and eukaryotes in hypersaline environments (Gómez-Silva et al., 2019). Endolithic microbial communities comprised mainly of eukaryotic microorganisms, such as melanized fungi and microalgae, have been found in soil gypsum crusts of the Atacama Desert. In these communities, melanized black fungi may play a role in hydration or protection of photobionts by dissipating excessive sunlight (Wierzechos et al., 2018).

Hot soils

Volcanic and geothermal soils

Volcanic and geothermal environments are characterized by low pH, high temperatures, and gas emissions consisting mainly of CO_2 with varied CH_4 , H_2S , and H_2 contents. In these environments, CO_2 and H_2 provide carbon and energy for the assembly of chemolithoautotrophic microbial communities, which perform the cycling of carbon, hydrogen, sulfur, and nitrogen.

H_2 is an abundant gas in volcanic and geothermal ecosystems that can be used by members of the microbial community as an electron donor. The low O_2 concentrations within these soils facilitate microaerobic H_2 oxidation using oxygen-sensitive hydrogenases. Zones devoid of O_2 are sites of anaerobic H_2 consumption, for example by autotrophic sulfate-reducing microorganisms or hydrogenotrophic methanogens. Methanogens are often detected in volcanic and geothermal environments, especially in the deeper anaerobic layers of the soil.

Proteobacteria are often found in volcanic and geothermal ecosystems, and this phylum includes methanotrophic and methylotrophic bacteria. The phylum Verrucomicrobia also contains methanotrophs, such as *Methylocandidiphilum*, which grows at temperatures of 50–60 °C and at pH of 2–5. Volcanic and geothermal ecosystems remain low in organic carbon (Picone et al., 2020).

Fire-affected soils

The town of Centralia, Pennsylvania, is the site of an underground coal mine fire that has been burning since 1962. Depending on the depth of the coal bed, the fire burns at an estimated 46–69 m below the surface. Heat, steam, and combustion products vent upward through the overlying soils. As steam and gases pass through the overlying rock and soil, soil temperatures increase while soil chemical composition is altered by both spontaneous and microbial-mediated chemical reactions. As the fire expands into new areas, it also retreats from some affected sites, which then recover to ambient temperatures (Lee et al., 2017).

A comparison of metagenomes from a 45 °C gradient of temperate-to-thermal soils that lie over the ongoing Centralia coal-seam fire showed that hot soils harbored distinct communities with small genomes and small cell sizes relative to those in ambient soils. Hot soils lacked genes that encode two-component regulatory systems, and genes for antimicrobial production and resistance. The small genomes in Centralia fire-affected soils are characteristic of thermophiles that were previously dormant in the temperate soil and were resuscitated by heat (Sorensen et al., 2019). Observations from the Centralia coal fire provide examples of the priority effect from microbial seed banks, alluded to above (Section “Microbial adaptation mechanisms in extreme soils”). In any given soil location submitted to a sudden rise in temperature, individual representatives of the thermophile seed banks react with

resuscitation from dormancy and rapid growth. Identity of transitioning individuals vary from site to site, thus accounting for the multiple equilibria among communities present in the various Centralia hot soil locations (Lee et al., 2017).

Saline and alkaline soils

The World Reference Base distinguishes two groups of salt-affected soils: (i) the Solonchaks, saline soils with an electrical conductivity of a saturated extract $>15 \text{ dS m}^{-1}$ in the top 125 cm, and with values that can reach above 150 dS m^{-1} ; (ii) the Solonetz, sodium-rich soils with an exchangeable sodium percentage $>15\%$ (Zaman et al., 2018).

The Solonchaks are widespread in all the arid and semi-arid areas of North Africa, Near East, Central Asia, India, Iran and Iraq, Australia, and the Americas. The Solonetz, mainly located in the steppe climatic regimes and flat landscapes with poor drainage, exist in Ukraine, the Russian Federation, Eastern Europe, China, India, United States, Canada, Southern and Eastern Africa and Australia (Otlewska et al., 2020).

In extremely saline soils with an electrical conductivity of more than 150 dS m^{-1} , mineralization of easily decomposable C-substrate is inhibited and large amounts of NH_4 are abiotically lost through NH_4 transformation or NH_3 volatilization. Additionally, microorganisms immobilize more $\text{NH}_4\text{-N}$ than the amount required for metabolic activity (Dendooven et al., 2010).

Whereas bacterial and archaeal diversity is lower in saline soils than in normal soils, this diversity does not decrease with an increase in salinity. Thus, other factors, in addition to salinity, contribute to limit biodiversity in saline soils. For example, strong sun exposure, low precipitation and high oxidative stress may be factors common to saline soils, irrespective of salt levels (Xie et al., 2017).

Radioactive soils

Radiation (including ionizing radiation and UV-B and UV-C) occurs at very low levels in natural soils and may be perceived as a constant environmental cue essential to normal bacterial metabolism (Castillo and Smith, 2017). Despite this constantly low radiation level, resistance to intense radiation is widespread among microorganisms. At least seven phyla of Bacteria and Archaea display this phenotype, as well as a wide variety of fungi, individual species within these phyla tolerating exposures in the laboratory that are as much as six orders of magnitude higher than the annual average exposure on Earth (Battista, 2015). This extreme radioresistance may have developed as a by-product of desiccation and heat resistance (Pavlopoulou et al., 2016). Organisms highly resistant to UV-C, including strains of *Deinococcus radiodurans*, were isolated from the Sonoran and Atacama Desert soils (Paulino-Lima et al., 2016). Combinations of oxidative stress and DNA damage are induced not only by ionizing radiation, but also by desiccation, UV radiation, heavy metals, and other agents (Shuryak, 2019).

When present in sufficiently high concentrations, as occurs in Chernobyl, Ukraine, and Fukushima, Japan, soil radionuclides apply a selective pressure that can shape prokaryotic community structure (Hoyos-Hernandez et al., 2019). For example, radiation encountered in and around the Chernobyl Nuclear Power Plant led to an increased abundance of pigmented fungi, including *Cladosporium sphaerospermum* and *Aureobasidium pullulans*, some of them showing unusual adaptations such as radiotropism (Gostinčar et al., 2015). The possibility of caesium accumulation by mushrooms growing on radionuclide-contaminated soils has been investigated, because of the perceived risk of food chain contamination. Edible mushrooms may accumulate radioactive caesium from the fallout of nuclear tests to an extent that varies over several orders of magnitude, depending on the species of fungi and on the mineral composition and organic matter content of the soil (Ohnuki et al., 2019).

Subsurface environments

The deep subsurface can be inhospitable to microorganisms, with combinations of extreme temperature, limited nutrient availability and limited energy sources. Furthermore, the lack of void space often constrains microbial activity as cells require pore networks for their development and metabolism.

Despite these limitations, microorganisms persist to great depths, and hotspots of activity may occur under appropriate physical conditions, and where energy can be obtained from electron donors such as methane, iron, or hydrogen. Hydrogen gas is a highly energetic electron donor when involved in sulfate, carbon dioxide, and ferric iron reduction, among other processes. In oligotrophic deep-subsurface environments, these processes may compete against each other for a limited supply of hydrogen (Gregory et al., 2019).

Viruses in extreme soils

Reported viral abundance in soils ranges from $2.2 \times 10^3 \text{ gdw}^{-1}$ (per gram dry weight of soil) in desert sands collected from Saudi Arabia to $5.8 \times 10^9 \text{ gdw}^{-1}$ in forest soil collected from eastern Virginia. In general, viral abundance appears to be lowest in hot deserts, intermediate in agricultural soils and in cold deserts, and highest in forested and wetland soils. Soil viral abundance is positively correlated with bacterial abundance, which indicates that most soil viruses infect bacterial hosts. Temperature influences viral persistence and abundance in soils, with lower temperatures fostering the persistence of viruses and higher temperatures

promoting thermal decay of viral particles (Williamson et al., 2017). Viral abundance is positively correlated with soil pH, total nitrogen and total carbon (Narr et al., 2017).

Lysogeny is widespread among soil bacteria, with up to 40% of terrestrial bacterial populations harboring inducible prophages. Temperate phages encode repressors and transcriptional regulatory proteins that silence, not only the expression of phage genes required for lytic replication, but also expression of homologous sequences in the host chromosome. This results in downregulation of metabolic functions in host bacteria and in reduction of energy expenditure, which in turn may contribute to host survival when nutrients are scarce. Hence, lysogenic phages may play important roles in extreme soils (Williamson et al., 2017).

Cold desert soils contain large extracellular virus titers. In contrast, hot hyperarid desert soils are characterized by high levels of bacterial lysogens and low extracellular virus counts (Zablocki et al., 2016). Host-virus interactions in the Atacama Desert hyperarid core may contribute to shape adapted microbiomes in this hostile environment. Diverse viruses infect a wide range of hosts, including Actinobacteria, Chloroflexi, Thaumarchaeota and Firmicutes. Viral genomes carry putative extremotolerance genes representing other versions of genes already used by microbes in their metabolism. For example, viral genes may code DNA repair proteins, enzymes against oxidative damage, spore formation proteins, and auxiliary metabolic functions. Thus, through horizontal transfer, desert viruses could mediate the spread of microbial resilience against environmental stress. These observations are indicative of a mutualistic model of host-virus interactions in desert environments, where viruses seek protection in microbial cells as lysogens, while viral extremotolerance genes aid survival of their hosts (Hwang et al., 2021).

Biogeography of extreme soil microbial colonizers

Extreme soils are the site of coexistence between endemic specialists and cosmopolitan airborne microorganisms, with those two types of soil colonizers adept at global dispersal. Aeolian transport of microbes (Després et al., 2012) is the primary assumption inherent to the hypothesis of Louren Baas Becking that ‘everything is everywhere, but the environment selects’. Aerosolized microbes may originate from agricultural fields, sewage treatment centers, geothermal springs, any surface exposed to sufficient wind force, and large-scale volcanic eruptions. Microorganisms associated with aerosolized desert dust can be transported over thousands of kilometers, across oceans and continents. Some of the microorganisms deposited on the Mont Blanc glacier following dust storms in the Sahara Desert belonged to phylogenetic groups frequently found in polar environments, such as the actinobacterium *Crossiella cryophilus* and *Deinococcus claudionis* (Chuvochina et al., 2011). Fungal spores can be transported in the presence or absence of dust (Kellogg and Griffin, 2006). High latitude regions are also a significant source of dust in the atmosphere (Bullard et al., 2016).

While survival of cells on aeolian particles is necessary to dispersal of microorganisms, successful colonization of new habitats is also required. Given the strong selection pressure applied by defining characters of extreme soils, the composition of bacterial communities of those soils is determined more by local environmental selection than by dispersal limitation or other historical contingencies (Chu et al., 2010).

Importance of extreme soil colonizing microorganisms

Contributions of extreme soil microorganisms to climate change

Permafrost is Earth’s largest terrestrial carbon sink, trapping plant and animal material in its frozen layers for centuries. It currently stores about 1600 billion tonnes of carbon—more than twice the current amount in the atmosphere. Microorganisms that inhabit polar regions are critically involved in regulating the Earth’s climate and carbon balance (van Dorst et al., 2017). As permafrost thaws because of climate change, enhanced microbial activity is responsible for increased emissions of carbon dioxide and methane (Hultman et al., 2015). In partially thawed muddy bogs, methane is produced through hydrogenotrophic methanogenesis, involving consumption of carbon dioxide and hydrogen. In fully thawed fens, the microbial community becomes more complex, and acetoclastic methanogenesis now occurs, in which acetate and carbon dioxide are used to produce methane. These two processes react differently to fluctuations in environmental conditions (Crill et al., 2014).

Climate change is expected to accelerate dryland expansion on a global scale, leading to reduced carbon sequestration and a positive feedback loop accelerating regional warming. The increasing aridity and enhanced warming will amplify the risk of desertification (Huang et al., 2016). Studies on arid soil microbiology, including examination of nutrient cycling and modes of microbial survival, appear useful to our understanding of global biological transitions in the context of climate change.

Contributions to astrobiology

Astrobiology deals with the search for extraterrestrial life and, in so doing, is drawn to investigate the nature of life and its expressions of resistance. Extreme soil studies may fulfill three functions in astrobiology: (i) to provide analogies for extraterrestrial geological, biological, and environmental conditions; (ii) to illustrate biological processes occurring on Earth that can be extrapolated to other planets; (iii) to provide testing grounds for astrobiological research instruments, field exploration procedures, and astronaut training (Gonçalves et al., 2019).

Extremophilic microorganisms represent model organisms, useful to elucidate the possibility of existence of extra-terrestrial life. For example, studies on carbon fixation by microbial communities colonizing hot desert rocks have demonstrated that more than half of the reductive tricarboxylic acid cycle, which functions in carbon fixation, can be photochemically activated in the presence of certain metals. Thus, this pathway may be involved in a primitive form of carbon anabolism in an acidic, metal-rich reducing environment (Gómez-Silva et al., 2019). Similarly, studies on terrestrial subsurface hydrogenotrophy and of hydrogen-based microbial chemolithotrophy as it occurs in the Earth's crust provide a factual basis for speculations on the evolution of hydrogen-driven ecosystems on Mars and other celestial bodies (Gregory et al., 2019). In some circumstances, eukaryotes may show a greater degree of resistance than that exhibited by prokaryotes. Fungi inhabiting Antarctic rocks display adaptations for colonization of hostile environments, making them potentially suitable eukaryotic models for astrobiology studies. *Cryomyces antarcticus*, a psychrophilic black fungus with optimal growth below 15 °C, was isolated from rocks and soil in the deserts of McMurdo Dry Valleys in Antarctica. *C. antarcticus* shows resistance to solar radiation, radioactivity, desiccation and oligotrophic conditions, which are conditions that may prevail on Mars or elsewhere (Gregory et al., 2019).

Contributions to biodiversity conservation

Soil organisms play an essential role in supporting life on Earth. Up to 90% of living organisms in terrestrial ecosystems are associated during their life cycle with below-ground habitats. As a complex and heterogeneous system, comprising organo-mineral aggregates of different sizes and organic components, the soil creates habitats for biodiversity at multiple spatial scales (FAO, ITPS, CBD, GSBI and EC, 2020). Preserving and understanding soil biodiversity is essential to the capacity of humans at conceiving and maintaining systems for food, health, and well-being provision. Extreme soils are illuminating soil biodiversity by revealing microbial adaptations to physicochemical conditions that differ vastly from those that humans commonly experience.

Less than 1% of total bacterial diversity and 5% of fungal taxa are amenable to usual culture techniques, while the remaining organisms are only observable through metagenomics. Composition of the cultivable and uncultivable components of active microbial communities differ between extreme and non-extreme soils. Given the selection pressures that are exerted in extreme soils, observation of these differences may lead to new knowledge and to applications in a wide range of scientific, technical, and industrial fields. Although microorganisms are not commonly considered in conservation efforts, extreme soils should be considered as microbial diversity hotspots warranting protection (van Dorst et al., 2017).

Contributions to industry, agriculture, and medicine

Extremophiles and their enzymes (called "extremozymes") have the potential to mediate catalysis in a wide range of operating conditions. Consequently, they can contribute to bioremediation, agriculture, polymer chemistry, bioenergy, and various other commercial applications. Examples pertaining to various branches of human activity are provided here.

(i) Extremophiles have been used for production of bioethanol, biobutanol, biodiesel, and gaseous fuels such as methane and hydrogen (Rathinam and Sani, 2018). (ii) The electrocatalytic activity of extremophiles has been harnessed for different bioelectrochemical applications (Dopson et al., 2016). Microbial fuel cells have been developed making use of electrocatalytic activity of extremophiles for bioelectricity generation from different substrates. Electrochemical activity of extremophiles has also made possible biohydrogen production in a microbial electrolysis cell. (iii) Archaea, bacteria and fungi isolated from the rhizosphere of plants growing under extremes of salinity, heat, cold, or pH value, can potentially improve crop plant development in a context of agricultural expansion and climate change (Saad et al., 2020; Verma et al., 2017). (iv) Extremophiles are promising candidates for bioremediation of contaminated soils because of their innate potential to resist toxic environments and to degrade a wide range of pollutants. (v) Biopolymers such as polyhydroxyalkanoates, extracellular polysaccharides, and alginate have been produced using extremophiles. (vi) Pigments from cold-adapted bacteria and fungi find applications, notably as food additives and in biomedicine (Sajjad et al., 2020). (vii) Fungi and actinobacteria producing novel antitumor and antimicrobial compounds were isolated from alkaline or acid soils (FAO, ITPS, CBD, GSBI and EC, 2020). (viii) Extremozymes offer promise in a wide range of sectors including diagnostics, therapeutics, detergents, bioenergy, food processing, tanning, and other industrial applications (Rathinam and Sani, 2018). (ix) Filamentous fungi and yeasts from hypersaline environments often show metal tolerance and can be exploited for the synthesis of a variety of metal-containing nanoparticles (Musa et al., 2018).

Conclusion

Microbial colonization of extreme soils resonates in multiple manners with curiosity, technological capacities, apprehension of the past and the future, and even spirituality of humans. The ability of microorganisms to thrive in inhospitable environments leads us to expand the boundaries of life and enter uncharted territory, where life itself seems to us strange and foreign. Microbes of extreme soils invite us to a travel through time and space, to other worlds populated by life forms that seem both incipient and highly evolved. Extreme soil microbiology also brings us home, to a vast field of practical applications, such as solvent and polymer production, agricultural enhancements, pollutant removal, and clean energy production. All this occurs in infinitely complex and highly structured soil environments, creating physical, biological, and conceptual spaces for stories that in many respects resemble our own. Given the contribution of biodiversity to ecosystem resilience (Myers, 1996) and more generally to sustainability of human civilizations, we should value and preserve extreme soil-colonizing microbes.

References

- Allison SD and Martiny JB (2008) Resistance, resilience, and redundancy in microbial communities. *Proceedings of the National Academy of Sciences* 105: 11512–11519.
- Andersen R, Chapman SJ, and Artz RRE (2013) Microbial communities in natural and disturbed peatlands: A review. *Soil Biology and Biochemistry* 57: 979–994.
- Barberán A, Ramírez KS, Leff JW, Bradford MA, Wall DH, and Fierer N (2014) Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. *Ecology Letters* 17: 794–802.
- Bardgett RD and Caruso T (2020) Soil microbial community responses to climate extremes: Resistance, resilience and transitions to alternative states. *Philosophical Transactions of the Royal Society B* 375: 20190112.
- Battista JR (2015) The origin of extreme ionizing radiation resistance. In: Bakermans C (ed.) *Microbial Evolution Under Extreme Conditions*, pp. 111–126. Berlin: De Gruyter.
- Bay S, Ferrari B, and Greening C (2018) Life without water: How do bacteria generate biomass in desert ecosystems. *Microbiology Australia* 39: 28–32.
- Bay SK, Dong X, Bradley JA, Leung PM, Grinter R, Jirapanjawan T, Arndt SK, Cook PL, LaRowe DE, and Nauer PA (2021) Trace gas oxidizers are widespread and active members of soil microbial communities. *Nature Microbiology* 6: 246–256.
- Bullard JE, Baddock M, Bradwell T, Crusius J, Darlington E, Gaiero D, Gasso S, Gisladdottir G, Hodgkins R, and McCulloch R (2016) High-latitude dust in the Earth system. *Reviews of Geophysics* 54: 447–485.
- Burr MD, Botero LM, Young MJ, Inskeep WP, and McDermott TR (2005) Observations concerning nitrogen cycling in a Yellowstone thermal soil environment. In: *Geothermal Biology and Geochemistry in Yellowstone National Park*, pp. 171–182. Bozeman, MT: Montana State University Publications.
- Cary SC, McDonald IR, Barrett JE, and Cowan DA (2010) On the rocks: The microbiology of Antarctic Dry Valley soils. *Nature Reviews Microbiology* 8: 129–138.
- Castillo H and Smith GB (2017) Below-background ionizing radiation as an environmental cue for bacteria. *Frontiers in Microbiology* 8: 177.
- Chan Y, Lacap DC, Lau MC, Ha KY, Warren-Rhodes KA, Cockell CS, Cowan DA, McKay CP, and Pointing SB (2012) Hypolithic microbial communities: Between a rock and a hard place. *Environmental Microbiology* 14: 2272–2282.
- Choe Y, Kim M, Woo J, Lee M, Lee E, and Lee Y (2018) Comparing rock-inhabiting microbial communities in different rock types from a high arctic polar desert. *FEMS Microbiology Ecology* 94: fty070.
- Chu H, Fierer N, Lauber CL, Caporaso JG, Knight R, and Grogan P (2010) Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. *Environmental Microbiology* 12: 2998–3006.
- Chuochina M, Alekhina I, Normand P, Petit J-R, and Bulat S (2011) Three events of Saharan dust deposition on the Mont Blanc glacier associated with different snow-colonizing bacterial phylotypes. *Microbiology* 80: 125–131.
- Cowan DA, Makhalanyane TP, Dennis PG, and Hopkins DW (2014) Microbial ecology and biogeochemistry of continental Antarctic soils. *Frontiers in Microbiology* 5: 154.
- Crill PM, Chanton JP, Rich VI, Tyson GW, and Saleska SR (2014) Methane dynamics regulated by microbial community response to permafrost thaw. *Nature* 514: 478–484.
- Crits-Christoph A, Gelsinger DR, Ma B, Wierzbos J, Ravel J, Davila A, Casero MC, and DiRuggiero J (2016) Functional interactions of archaea, bacteria and viruses in a hypersaline endolithic community. *Environmental Microbiology* 18: 2064–2077.
- De Maayer P, Anderson D, Cary C, and Cowan DA (2014) Some like it cold: Understanding the survival strategies of psychrophiles. *EMBO Reports* 15: 508–517.
- de Oliveira TB and Rodrigues A (2019) Ecology of thermophilic fungi. In: Tiquia-Arashiro SM and Grube M (eds.) *Fungi in Extreme Environments: Ecological Role and Biotechnological Significance*, pp. 39–57. Springer.
- Dendooven L, Alcántara-Hernández RJ, Valenzuela-Encinas C, Luna-Guido M, Perez-Guevara F, and Marsch R (2010) Dynamics of carbon and nitrogen in an extreme alkaline saline soil: A review. *Soil Biology and Biochemistry* 42: 865–877.
- Després VR, Huffman JA, Burrows SM, Hoose C, Safatov A, Buryak G, Fröhlich-Nowoisky J, Elbert W, Andreae M, and Pöschl U (2012) Primary biological aerosol particles in the atmosphere: A review. *Tellus Series B: Chemical and Physical Meteorology* 64: 15598.
- Dopson M, Ni G, and Sleutels TH (2016) Possibilities for extremophilic microorganisms in microbial electrochemical systems. *FEMS Microbiology Reviews* 40: 164–181.
- Dragone NB, Diaz MA, Hogg I, Lyons WB, Jackson WA, Wall DH, Adams BJ, and Fierer NB (2020) Exploring the boundaries of microbial habitability in soil. *bioRxiv*. <https://doi.org/10.1101/2020.08.03.234583>.
- Empadinhas N and da Costa MS (2008) Osmoadaptation mechanisms in prokaryotes: Distribution of compatible solutes. *International Microbiology* 11: 151–161.
- Escudero C, Vera M, Oggerin M, and Amils R (2018) Active microbial biofilms in deep poor porous continental subsurface rocks. *Scientific Reports* 8: 1–9.
- FAO and ITPS (2015) *Status of the World's Soil Resources (SWSR)—Main Report*. Rome, Italy: Food and Agriculture Organization of the United Nations and Intergovernmental Technical Panel on Soils.
- FAO, ITPS, CBD, GSBi and EC (2020) *State of Knowledge of Soil Biodiversity—Status, Challenges and Potentialities*. Full Report. Rome.
- Ferrenberg S, Tucker C, and Reed SC (2017) Biological soil crusts: Diminutive communities of potential global importance. *Frontiers in Ecology and the Environment* 15: 160–167.
- Fierer N (2017) Embracing the unknown: Disentangling the complexities of the soil microbiome. *Nature Reviews Microbiology* 15: 579–590.
- Finstad KM, Probst AJ, Thomas BC, Andersen GL, Demergasso C, Echeverría A, Amundson RG, and Banfield JF (2017) Microbial community structure and the persistence of cyanobacterial populations in salt crusts of the hyperarid atacama desert from genome-resolved metagenomics. *Frontiers in Microbiology* 8: 1435.
- Gagliano AL, Calabrese S, Daskalopoulou K, Kyriakopoulos K, Tagliavia M, and D'Alessandro W (2020) Methanotrophy in geothermal soils, an overlooked process: The example of Nisyros island (Greece). *Chemical Geology* 539: 119546.
- Gavrillets S (2008) Fitness landscapes. In: Jorgensen SE and Fath BD (eds.) *Encyclopedia of Ecology*, pp. 1612–1615. Oxford: Elsevier.
- Gerday C and Glansdorff N (2007) *Physiology and Biochemistry of Extremophiles*. ASM Press.
- Gessler N, Egorova A, and Belozerskaya T (2014) Melanin pigments of fungi under extreme environmental conditions. *Applied Biochemistry and Microbiology* 50: 105–113.
- Godon JJ, Galès A, Latrille E, Ouichanpagdee P, and Seyer J-P (2020) An “overlooked” habitat for thermophilic bacteria: The phyllosphere. *BioDiscovery* 23: e47033.
- Gómez-Silva B (2018) Lithobiotic life: “Atacama rocks are well and alive” *Antonie Van Leeuwenhoek* 111: 1333–1343.
- Gómez-Silva B, Vilo-Muñoz C, Galetović A, Dong Q, Castelán-Sánchez HG, Pérez-Llano Y, Sánchez-Carbente MDR, Dávila-Ramos S, Cortés-López NG, and Martínez-Ávila L (2019) Metagenomics of Atacama lithobiotic extremophile life unveils highlights on fungal communities, biogeochemical cycles and carbohydrate-active enzymes. *Microorganisms* 7: 619.
- Gonçalves VN, Alves IMS, de Oliveira FS, Schaefer CEGR, Turbay CVG, Rosa CA, and Rosa LH (2019) Rock-inhabiting fungi in Antarctica: New frontiers of the edge of life. In: Rosa LH (ed.) *Fungi of Antarctica*, pp. 99–126. Springer.
- Gostinčar C, Gunde-Cimerman N, and Grube M (2015) Polyextremotolerance as the fungal answer to changing environments. In: Bakermans C (ed.) *Microbial Evolution Under Extreme Conditions*. Berlin: de Gruyter.
- Gregory SP, Barnett MJ, Field LP, and Milodowski AE (2019) Subsurface microbial hydrogen cycling: Natural occurrence and implications for industry. *Microorganisms* 7: 53.
- Gunde-Cimerman N, Plemenitaš A, and Oren A (2018) Strategies of adaptation of microorganisms of the three domains of life to high salt concentrations. *FEMS Microbiology Reviews* 42: 353–375.
- Herbold CW, Lee CK, McDonald IR, and Cary SC (2014) Evidence of global-scale aeolian dispersal and endemism in isolated geothermal microbial communities of Antarctica. *Nature Communications* 5: 3875.
- Heulin T, De Luca G, Barakat M, Gommeaux M, de Groot A, Blanchard L, Ortel P, and Achouak W (2017) Bacterial adaptation to hot and dry deserts. In: *Adaptation of Microbial Life to Environmental Extremes*, pp. 75–98. Springer.
- Hoyos-Hernandez C, Courbert C, Simonucci C, David S, Vogel TM, and Larose C (2019) Community structure and functional genes in radionuclide contaminated soils in Chernobyl and Fukushima. *FEMS Microbiology Letters* 366: fnz180.
- Huang J, Yu H, Guan X, Wang G, and Guo R (2016) Accelerated dryland expansion under climate change. *Nature Climate Change* 6: 166–171.

- Huang W, Ertekin E, Wang T, Cruz L, Dailey M, DiRuggiero J, and Kisailus D (2020) Mechanism of water extraction from gypsum rock by desert colonizing microorganisms. *Proceedings of the National Academy of Sciences* 117: 10681–10687.
- Hultman J, Waldrop MP, Mackelprang R, David MM, McFarland J, Blazewicz SJ, Harden J, Turetsky MR, McGuire AD, and Shah MB (2015) Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature* 521: 208–212.
- Hwang Y, Schulze-Makuch D, Arens FL, Saenz JS, Adam PS, Bornemann TLV, Airo A, Schlöter M, and Probst AJ (2020) Leave no stone unturned: The hidden potential of carbon and nitrogen cycling by novel, highly adapted Thaumarchaeota in the Atacama Desert hyperarid core. *bioRxiv*. 2020.07.17.208546.
- Hwang Y, Rahlff J, Schulze-Makuch D, Schlöter M, and Probst AJ (2021) Diverse viruses carrying genes for microbial extremotolerance in the Atacama Desert hyperarid soil. *bioRxiv*. 2020.09.21.307520.
- Jang J, Hur HG, Sadowsky MJ, Byappanahalli M, Yan T, and Ishii S (2017) Environmental *Escherichia coli*: Ecology and public health implications—A review. *Journal of Applied Microbiology* 123: 570–581.
- Jansson JK and Taş N (2014) The microbial ecology of permafrost. *Nature Reviews Microbiology* 12: 414–425.
- Ji M, Greening C, Vanwonterghem I, Carere CR, Bay SK, Steen JA, Montgomery K, Lines T, Beardall J, and Van Dorst J (2017) Atmospheric trace gases support primary production in Antarctic desert surface soil. *Nature* 552: 400–403.
- Joergensen RG and Wichern F (2018) Alive and kicking: Why dormant soil microorganisms matter. *Soil Biology and Biochemistry* 116: 419–430.
- Jung P, Baumann K, Lehnert LW, Samolov E, Achilles S, Schermer M, Wraase LM, Eckhardt KU, Bader MY, and Leinweber P (2020) Desert breath—How fog promotes a novel type of soil biocenosis, forming the coastal Atacama Desert's living skin. *Geobiology* 18: 113–124.
- Kellogg CA and Griffin DW (2006) Aerobiology and the global transport of desert dust. *Trends in Ecology & Evolution* 21: 638–644.
- Knelman JE, Schmidt SK, Lynch RC, Darcy JL, Castle SC, Cleveland CC, and Nemergut DR (2014) Nutrient addition dramatically accelerates microbial community succession. *PLoS One* 9: e102609.
- Lee S-H, Sorensen JW, Grady KL, Tobin TC, and Shade A (2017) Divergent extremes but convergent recovery of bacterial and archaeal soil communities to an ongoing subterranean coal mine fire. *The ISME Journal* 11: 1447–1459.
- Lennon JT and Jones SE (2011) Microbial seed banks: The ecological and evolutionary implications of dormancy. *Nature Reviews Microbiology* 9: 119–130.
- Locey KJ and Lennon JT (2016) Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences of the United States of America* 113: 5970–5975.
- Lollar GS, Warr O, Telling J, Osburn MR, and Lollar BS (2019) 'Follow the water': Hydrogeochemical constraints on microbial investigations 2.4 km below surface at the Kidd creek deep fluid and deep life observatory. *Geomicrobiology Journal* 36: 859–872.
- Lynch RC, Darcy JL, Kane NC, Nemergut DR, and Schmidt SK (2014) Metagenomic evidence for metabolism of trace atmospheric gases by high-elevation desert Actinobacteria. *Frontiers in Microbiology* 5: 698.
- Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ, Rubin EM, and Jansson JK (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* 480: 368–371.
- Makhalanyane TP, Valverde A, Gunnigle E, Frossard A, Ramond J-B, and Cowan DA (2015) Microbial ecology of hot desert edaphic systems. *FEMS Microbiology Reviews* 39: 203–221.
- Moynihan EL, Richards KG, Brennan FP, Tyrrel SF, and Ritz K (2015) Enteropathogen survival in soil from different land-uses is predominantly regulated by microbial community composition. *Applied Soil Ecology* 89: 76–84.
- Musa H, Kasim FH, Gunny AAN, and Gopinath SC (2018) Salt-adapted moulds and yeasts: Potentials in industrial and environmental biotechnology. *Process Biochemistry* 69: 33–44.
- Myers N (1996) Environmental services of biodiversity. *Proceedings of the National Academy of Sciences* 93: 2764–2769.
- Narr A, Nawaz A, Wick LY, Harms H, and Chatzinotas A (2017) Soil viral communities vary temporally and along a land use transect as revealed by virus-like particle counting and a modified community fingerprinting approach (fRAPD). *Frontiers in Microbiology* 8: 1975.
- Navarro-Noya YE, Luna-Guido M, and Dendooven L (2016) Cultivable nitrogen fixing bacteria from extremely alkaline-saline soils. *Advances in Microbiology* 6: 412–423.
- Ohnuki T, Sakamoto F, Kozai N, Nanba K, Neda H, Sasaki Y, Niizato T, Watanabe N, and Kozaki T (2019) Role of filamentous fungi in migration of radioactive cesium in the Fukushima forest soil environment. *Environmental Science: Processes & Impacts* 21: 1164–1173.
- Oren A (2008) Microbial life at high salt concentrations: Phylogenetic and metabolic diversity. *Saline Systems* 4: 2.
- Osborne P, Hall LJ, Kronfeld-Schor N, Thybert D, and Haerty W (2020) A rather dry subject; investigating the study of arid-associated microbial communities. *Environmental Microbiome* 15: 1–14.
- Otlewska A, Migliore M, Dybka-Stepień K, Manfredini A, Struszczyk-Świta K, Napoli R, Białkowska A, Canfora L, and Pinzari F (2020) When salt meddles between plant, soil, and microorganisms. *Frontiers in Plant Science* 11: 1429.
- Paulino-Lima IG, Fujishima K, Navarrete JU, Galante D, Rodrigues F, Azua-Bustos A, and Rothschild LJ (2016) Extremely high UV-C radiation resistant microorganisms from desert environments with different manganese concentrations. *Journal of Photochemistry and Photobiology B: Biology* 163: 327–336.
- Pavlopoulou A, Savva GD, Louka M, Bagos PG, Vorgias CE, Michalopoulos I, and Georgakilas AG (2016) Unraveling the mechanisms of extreme radioresistance in prokaryotes: Lessons from nature. *Mutation Research, Reviews in Mutation Research* 767: 92–107.
- Picone N, Hogendoorn C, Cremers G, Poghosyan L, Pol A, van Aalen TA, Gagliano AL, D'Alessandro W, Quatrini P, and Jetten MS (2020) Geothermal gases shape the microbial community of the volcanic soil of Pantelleria, Italy. *mSystems* 5: e00517-20.
- Powell AJ, Parchert KJ, Bustamante JM, Ricken JB, Hutchinson MI, and Natvig DO (2012) Thermophilic fungi in an aridland ecosystem. *Mycologia* 104: 813–825.
- Pressler Y, Moore JC, and Cotrufo MF (2019) Belowground community responses to fire: Meta-analysis reveals contrasting responses of soil microorganisms and mesofauna. *Oikos* 128: 309–327.
- Ranawat P and Rawat S (2017) Radiation resistance in thermophiles: Mechanisms and applications. *World Journal of Microbiology and Biotechnology* 33: 112.
- Rathinam NK and Sani RK (2018) Bioprospecting of extremophiles for biotechnology applications. In: Sani RK and Rathinam NK (eds.) *Extremophilic Microbial Processing of Lignocellulosic Feedstocks to Biofuels, Value-Added Products, and Usable Power*, pp. 1–23. Springer.
- Ren M, Zhang Z, Wang X, Zhou Z, Chen D, Zeng H, Zhao S, Chen L, Hu Y, and Zhang C (2018) Diversity and contributions to nitrogen cycling and carbon fixation of soil salinity shaped microbial communities in Tarim Basin. *Frontiers in Microbiology* 9: 431.
- Rothschild LJ and Mancinelli RL (2001) Life in extreme environments. *Nature* 409: 1092–1101.
- Saad MM, Eida AA, and Hirt H (2020) Tailoring plant-associated microbial inoculants in agriculture: A roadmap for successful application. *Journal of Experimental Botany* 71: 3878–3901.
- Sajjad W, Din G, Rafiq M, Iqbal A, Khan S, Zada S, Ali B, and Kang S (2020) Pigment production by cold-adapted bacteria and fungi: Colorful tale of cryosphere with wide range applications. *Extremophiles* 24: 447–473.
- Sánchez-Otero M-G, Quintana-Castro R, Domínguez-Chávez JG, Peña-Montes C, and Oliart-Ros RM (2019) Unique microorganisms inhabit extreme soils. In: Kumar A and Sharma S (eds.) *Microbes and Enzymes in Soil Health and Bioremediation*, pp. 39–73. Singapore: Springer Singapore.
- Santana M and Gonzalez J (2015) High temperature microbial activity in upper soil layers. *FEMS Microbiology Letters* 362: fmv182.
- Santos H, Lamosa P, Faria TQ, Borges N, and Neves C (2007) The physiological role, biosynthesis, and mode of action of compatible solutes from (hyper) thermophiles. In: Gerday C and Glansdorff N (eds.) *Physiology and Biochemistry of Extremophiles*, pp. 86–103. Washington, DC: ASM Press.
- Schmid AK, Allers T, and DiRuggiero J (2020) SnapShot: Microbial extremophiles. *Cell* 180: 818–818.e1.
- Shen J, Wyness AJ, Claire MW, and Zerkle AL (2021) Spatial variability of microbial communities and salt distributions across a latitudinal aridity gradient in the Atacama desert. *Microbial Ecology* 82(2): 442–458.
- Shuryak I (2019) Review of microbial resistance to chronic ionizing radiation exposure under environmental conditions. *Journal of Environmental Radioactivity* 196: 50–63.

- Singh P, Jain K, Desai C, Tiwari O, and Madamwar D (2019) Microbial community dynamics of extremophiles/extreme environment. In: Das S and Dash HR (eds.) *Microbial Diversity in the Genomic Era*, pp. 323–332. London: Elsevier.
- Sorensen JW, Dunivin TK, Tobin TC, and Shade A (2019) Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. *Nature Microbiology* 4: 55–61.
- Swanner ED and Templeton AS (2011) Potential for nitrogen fixation and nitrification in the granite-hosted subsurface at henderson mine, CO. *Frontiers in Microbiology* 2: 254.
- Tecon R and Or D (2017) Biophysical processes supporting the diversity of microbial life in soil. *FEMS Microbiology Reviews* 41: 599–623.
- van Dorst J, Benaud N, and Ferrari B (2017) New insights into the microbial diversity of polar desert soils: A biotechnological perspective. In: Chénard C and Lauro FM (eds.) *Microbial Ecology of Extreme Environments*, pp. 169–183. Springer.
- Van Elsas JD, Semenov AV, Costa R, and Trevors JT (2011) Survival of *Escherichia coli* in the environment: Fundamental and public health aspects. *The ISME Journal* 5: 173–183.
- Vera-Gargallo B, Chowdhury TR, Brown J, Fansler SJ, Durán-Viseras A, Sánchez-Porro C, Bailey VL, Jansson JK, and Ventosa A (2019) Spatial distribution of prokaryotic communities in hypersaline soils. *Scientific Reports* 9: 1–12.
- Verma P, Yadav AN, Kumar V, Singh DP, and Saxena AK (2017) Beneficial plant-microbes interactions: Biodiversity of microbes from diverse extreme environments and its impact for crop improvement. In: Singh DP, Singh HB, and Prabha R (eds.) *Plant-Microbe Interactions in Agro-Ecological Perspectives*, pp. 543–580. Springer.
- Wagner D (2008) Microbial communities and processes in Arctic permafrost environments. In: Dion P and Nautiyal CS (eds.) *Microbiology of Extreme Soils*, pp. 133–154. Berlin, Heidelberg: Springer.
- Walker JJ and Pace NR (2007) Endolithic Microbial Ecosystems. *Annual Review of Microbiology* 61: 331–347.
- Wierzos J, de los Rios A, and Ascaso C (2012) Microorganisms in desert rocks: The edge of life on Earth. *International Microbiology* 15(4): 173–183.
- Wierzos J, Casero MC, Artieda O, and Ascaso C (2018) Endolithic microbial habitats as refuges for life in polyextreme environment of the Atacama Desert. *Current Opinion in Microbiology* 43: 124–131.
- Williamson KE, Fuhrmann JJ, Wommack KE, and Radosevich M (2017) Viruses in soil ecosystems: An unknown quantity within an unexplored territory. *Annual Review of Virology* 4: 201–219.
- Wu Q-L, Chen J-H, Deng L-Y, Liu Z-X, He J-W, Liu Z-X, Yang L-L, Tang S-K, and Chen Y-G (2020) *Sediminibacillus terrae* sp. nov., a moderate halophile isolated from non-saline farm soil. *International Journal of Systematic and Evolutionary Microbiology* 70: 1139–1144.
- Xie K, Deng Y, Zhang S, Zhang W, Liu J, Xie Y, Zhang X, and Huang H (2017) Prokaryotic community distribution along an ecological gradient of salinity in surface and subsurface saline soils. *Scientific Reports* 7: 1–10.
- Young IM and Crawford JW (2004) Interactions and self-organization in the soil-microbe complex. *Science* 304: 1634–1637.
- Zablocki O, Adriaenssens EM, and Cowan D (2016) Diversity and ecology of viruses in hyperarid desert soils. *Applied and Environmental Microbiology* 82: 770–777.
- Zaman M, Shahid SA, and Heng L (2018) *Guideline for Salinity Assessment, Mitigation and Adaptation Using Nuclear and Related Techniques*. Springer.

Microbial mobility and transport in soils

Mohammad Reza Mosaddeghi^a, Adrian Unc^b, Ali Akbar Mahboubi^{c,d}, and Azadeh Safadoust^c, ^aDepartment of Soil Science, College of Agriculture, Isfahan University of Technology, Isfahan, Iran; ^bSchool of Science and the Environment, Memorial University of Newfoundland, Corner Brook, NL, Canada; ^cDepartment of Soil Science, College of Agriculture, Bu-Ali Sina University, Hamadan, Iran; ^dSchool of Engineering, University of Guelph, Guelph, Ontario, Canada

© 2023 Elsevier Ltd. All rights reserved.

Introduction	513
Indicator organisms for pathogen contamination	513
Bacteria survival and persistence in soil systems	514
Soil water status	514
Soil type	514
Temperature	514
pH	515
Soil carbonate content	515
Properties of organic wastes and storage conditions	515
Type of organic waste and application method	515
Nutrient availability	515
Competition	515
Bacterial transport in soil systems	515
Physical processes	516
Physicochemical processes	516
Biological processes	516
Transport pathways	516
Soil type (texture and structure) and preferential flow	517
Impact of temperature	518
Soil water regime	518
Bacterial sorption and retention in soil systems	518
Surface charge properties of soil particles and bacterial cells	518
Organic waste properties and components affecting bacterial transport	519
Approaches to microbial transport studies	519
Field studies	519
Laboratory studies	520
Model studies	520
Role of new generation sequencing (NGS) for advancing microbial transport studies	520
Summary	520
References	520

Abstract

Organic wastes are C-rich, beneficial amendments for improving soil conditions and plant growth. However, their land application is a major source of pathogenic microorganisms entering surface and groundwater systems. Persistence and transport of these microorganisms depend on interrelationships between properties of the organic amendments and seasonally variable environmental properties of the soil. Ionic and particulate components of amendments appreciably alter the physical and chemical properties of soil and microbe charged surfaces. Microbial persistence in soils determines the load of cells available for transport. Microorganism transport depends on structure-related macropore flow and their adsorption and desorption to soil and suspended particles.

Key points

- Organic wastes are soil amendments but also major sources of microorganisms, including pathogens.
- Several interrelated factors affect survival and transport of pathogens in the soils.
- Organic waste properties can alter physicochemical properties of soils and pathogens.
- Soil type and water status play a dominant role in the fate of pathogens.
- Preferential flow and water repellency facilitate the microbial mobility.

Introduction

Organic manures and wastes are C-rich beneficial amendments used to improve soil physical and chemical conditions important to plant growth. They may improve soil aggregation and water infiltration, enhance soil water retention, decrease evaporation and the degree of compaction, increase resistance to water and wind erosion as well as soil cation exchange capacity, and serve as habitat or food source for soil life. Their dark color helps warm soils. As a source of several mineral nutrients, they can improve nutrient cycling. However, application of organic amendments to agricultural lands is a major source of pathogen microorganisms released into surface and groundwater systems. Pathogenic bacteria associated with agricultural wastes mainly include *Escherichia coli*, *Salmonella* sp., *Shigella* sp., *Campylobacter* sp., *Vibrio* sp. Pathogenic protozoans include *Cryptosporidium parvum*, and *Giardia* spp. The number of pathogens entering the environment is an important criterion for regulating the use of organic amendments in agriculture. Bacteria are single-celled organisms and constitute a large domain of prokaryotic microorganisms. Typically, a few micrometers in length, bacteria may have distinct shapes, ranging from spheres to rods (e.g., rod-shaped, curve-rods, spiral, ovoid, round or filamentous). In general, out of large totals of bacteria (e.g., 10^{10} bacterial colony forming units [CFU] per 1 g of dry matter) that can be found in manure, pathogenic bacteria commonly represent a very small percentage rarely exceeding 10^5 CFU per 1 g of dry manure. The most frequent manure carried bacteria are fecal coliforms and enterococci. Many bacteria, such as *E. coli*, are able to live under anaerobic conditions. Viruses observed in organic amendments as sub-microscopic infectious agents (30–300 nm) are able to survive outside the host for a long time. However, consistent evidence about survival and persistence of viruses in animal manures is infrequently available. Protozoa are single-celled eukaryotes, either free-living or parasitic, which feed on organic matter such as other microorganisms or organic tissues and debris.

Indicator organisms for pathogen contamination

Difficulties and expense involved in the testing for specific pathogens led to the use of indicator organisms for estimating the fate of enteric bacteria in the environment. Fecal coliforms (FC), a functional integrative category (Fig. 1), originally found in the intestinal gut of warm-blooded animals are the most commonly used indicators of pathogenic microorganisms. They are identified by their ability to ferment and produce gas from lactose at 44.5 °C.

Selected non-pathogenic strains of *Enterococcus* and *E. coli*, common and at high concentrations in manure and sewage, are also used as indicators of fecal contamination. *E. coli*, the most ubiquitous FC, is a Gram-negative, facultative anaerobic, rod-shaped (1–6 µm), coliform bacterium of the genus *Escherichia* and is commonly found in the lower intestine of warm-blooded organisms (endotherms). Most *E. coli* strains are harmless, but some strains, such as *E. coli* O157:H7 are pathogenic. Indicator organisms are selected for their easy culturing and counting but their persistence in the environment may not be the same as that of pathogens. However, the presence of a small number of indicator organisms in drinking water sources may signal a risk to human health.

The fate and survival of pathogens in soils can be affected in three ways: (1) retention in the soil matrix, (2) transport through soils to deeper layers and toward groundwater or tile drains, and (3) die-off due to unfavorable environmental conditions. Many factors may modify these including type of organic waste, pores size distribution and continuity in the soil. These soil characteristics are determined by the interplay between porosity and particle size distribution and influence permeability and

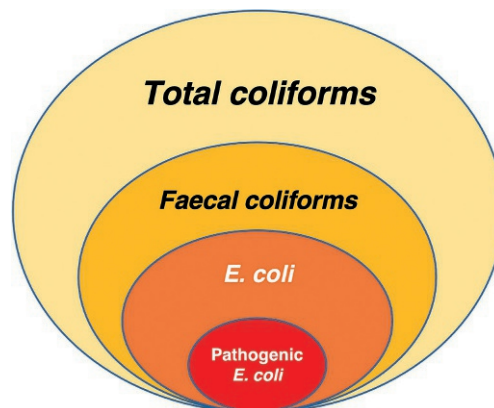


Fig. 1 The microbial pathogens in organic wastes are found in small proportions, harder to identify; thus, more common, encompassing groups such as fecal or total coliforms are used as indicators of potential risk.

water flow characteristics, as well as acidity and ionic strength of soil solution, all also affected by climatic conditions and management practices. Survival and transport of pathogens depend on the physicochemical properties of both soil and manure as well as the interactions between soil, water, microorganisms and the surrounding environment. Transport of bacteria from a diffuse field source (i.e., manure spreading) to surface and ground water bodies is mediated by precipitation, snow melt, and irrigation water.

Bacteria survival and persistence in soil systems

Enteric pathogens may survive for a long period of time outside their hosts. Numerous factors influence the survival and persistence of such pathogens in soil (Fig. 2). Survival of enteric bacteria in soil and groundwater can commonly range from 2 to 4 months. Several factors can affect bacteria survival in soil as follows: soil water status, soil type, temperature and pH, manure type and application method, nutrient availability, and competition.

Soil water status

The principal factor affecting the survival of enteric bacteria in soil systems is considered to be the soil water status or availability (i.e., soil water potential), overriding or obscuring the significance of other factors. Fecal microbes die more readily at low matric potentials in unsaturated soils. Survivability of *E. coli* or fecal coliform originating in animal manure increases as soil water content increases, with best survivability in waterlogged organic soils. While this phenomenon has been repeatedly reported, there is yet limited quantitative information to support prediction of the absolute relationship between soil water activity and microbial survival.

Soil type

Although soil type generally reflects pedogenesis, it can also be employed to infer physical and chemical properties relevant to soil biotic parameters. The single property that can be most easily related to bacterial survival is water retention, which is linked to particle size distribution and organic matter content. Reduced mortality could be primarily related to soil type, with lower mortality in soils with a higher matric potential.

Temperature

Readers are directed to the extensive review of the effects of temperature on *E. coli* die-off by Foppen and Schijven (2006). Most research reports an inverse relationship between temperature and FC bacterial mortality. Die-off rate of coliform bacteria is

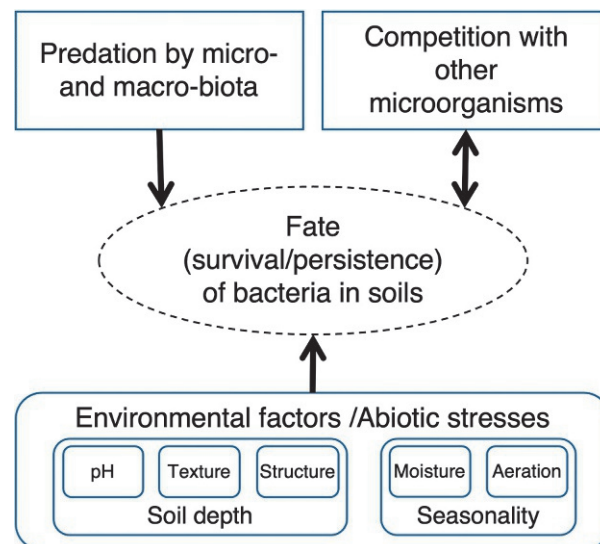


Fig. 2 Factors affecting the persistence of bacteria in soils.

approximately doubled for each 10 °C increase in soil temperature. As the soil cools, survival of fecal bacteria is enhanced: 90% of coliform bacteria die-off within 3.3 days of land application in the summer compared to 13.4 days in the winter; *E. coli* could survive at least 100 days in a water-soil matrix kept at 10 °C (Foppen and Schijven, 2006).

pH

Enteric bacteria are acclimated to neutral to slightly alkaline pH ranges, similar to the conditions in the lower digestive tract of warm-blooded animals. Thus, once discharged into the soil environments, the more acidic pH reduces their survival period. If the pH matches more closely the neutral pH in excreta, survival is improved. *E. coli* and fecal streptococci survive several weeks in limestone (pH 5.8–7.8) while dying in a few days in a peat soil (pH 2.9–4.5). If other properties, such as texture, organic matter and thus water activity, are comparable, the pH will still affect microbial survivability.

Soil carbonate content

Limited research has been carried out on the effect of carbonates on bacterial survival in soil. Carbonates do affect soil pH and, especially in arid soils (i.e., subsurface caliche layers), they can also control water retention. CaCO₃ supplementation to an arid loamy soil (at up to 25% of the total soil mass) was found to reduce the persistence of nalidixic acid-resistant *E. coli* (*E. coli* NAR), especially at high temperatures (i.e., 37 °C; Farhangi et al., 2013), even at a sufficient water content (i.e., at field capacity). On the other hand, survival of *Pseudomonas fluorescens* CHA0-Rif was enhanced with increased carbonate content, at the same time as the overall activity of the microbial community declined. This was evident for high temperatures as the temperature declined the presence of carbonate did not have an evident favorable effect on *P. fluorescens* survival (Farhangi et al., 2014).

Properties of organic wastes and storage conditions

Animal manures are the primary sources of non-native bacteria introduced to soils. Concentrations of indicator and pathogenic organisms in animal manures vary widely depending on animal species, age of animals, waste storage system (liquid or solid), storage period, and level of pre-treatment prior to land disposal. Usually, the number of pathogens may first increase as manure is stored, and then may decrease. The decay rate of bacterial counts is faster in solid manure than in liquid manure. Temperature, water content and pH can affect the decay rate of pathogens during manure storage.

Type of organic waste and application method

The total counts of pathogenic bacteria are greater in swine and poultry manure than in cattle manure. Due to the greater mobility of bacteria in the liquid phase, liquid organic waste tends to be more uniformly contaminated than their solid counterparts. Methods for the land application of depend on the type of organic waste (liquid or solid), soil, and crop. When liquid waste is injected into soil, microorganisms are unlikely to be destroyed by the UV in solar radiation. Injection also increases the possibility for microorganisms to be adsorbed onto soil particles. Normal tillage used in farming operations can influence the abiotic and biotic environments that waste microbes experience after application.

Nutrient availability

Nutrient availability is central to the survival of microbes in soil. Besides its role in mediating water retention, organic matter may also increase the retention of nutrients, provide an available carbon source for heterotrophic microbes. Hence, increased organic matter contents promote the survival of bacteria and also of enteric bacteria in soil.

Competition

Competing microorganisms will limit pathogen survival in soil. Bacteriophage and predatory microorganisms, such as *Bdellovibrio*, further limit the survival of fecal bacteria in soils. In the absence of soil microbes, i.e., in sterile soil experiments, pathogen regrowth might occur. On the other hand, addition of organic waste may favor recovery of *Escherichia coli* already in soil. Macrofauna, such as earthworms, may also limit persistence in soil of enteric microorganisms.

Bacterial transport in soil systems

Critical to the distribution of contaminant fecal bacteria in the soil and water environments is their likelihood to be transported through soils. Transport of bacteria through the soil depends on flow path geometry and stability and is mitigated by interaction

between soil, soil solution, air-water interfaces, and characteristics of microbial surfaces. After initial entry, the transport through soil depends on continued entrainment of bacteria and resuspension of those retained on the porous structure (Fig. 3).

Bacterial transport through heterogeneous porous media, such as soil, is affected by the interplay of a series of mechanisms which may be divided into physical, chemical and biological. Microorganisms can migrate through soil via advection and dispersion, while being subjected to the effects of filtration, adsorption, desorption, growth, decay, sedimentation and chemotaxis. Transport and retention of bacteria and solutes in soil are largely affected by soil texture, structure and their interactions, which affect the distribution of soil water content, water flux density, pore size distribution and continuity, filtration, adsorption and sedimentation. All these complicate the description and prediction of the movement of bacteria in soil. Furthermore, the electrochemical characteristics of bacterial cell surfaces and probably bacterial shape, together with the effect of manure on the ionic and electrostatic properties of the soil solution and surfaces of the soil matrix, affect bacterial transport and retention of manure-borne bacteria. Nevertheless, the transport of pathogens and indicator organisms in soils receiving manure is mainly facilitated by infiltrating water and surface runoff and associated with the movement of sediment and waste colloidal particles.

Physical processes

Principal physical processes for microbial movement through porous media soils include convection and advection or mass flow and hydrodynamic dispersion (mechanical mixing). Bacterial transport is mostly passive, being determined by the presence of rapid water fluxes (convection) and heterogeneity in water velocity at pore scale (dispersion), although cell motility can play a role in movement in suspension over short distances. Diffusion, as a short-distance process, is of little importance in bacterial transport compared to mechanical mixing. However, diffusion may be an important transport mechanism for small virus particles ($<1 \mu\text{m}$). Filtration and straining occur when a microorganism is prevented from flowing through a pore. Sedimentation is the gravitational deposition of particles (bacteria) on soil grain surfaces.

Physicochemical processes

Physicochemical adsorption and desorption processes are mainly attenuative in nature and hinder microbial transport through soil. Adsorption is defined as the process of accumulation of charged molecules or particles that are in aqueous suspension or solution on a suitable charged surface; the interfaces between air and liquid can also act as adsorption surfaces. Adsorption of microorganisms to soil surfaces is reversible and is best described by the Freundlich isotherm.

Biological processes

Microbial processes such as growth or death affect the concentration of microorganisms, thereby affecting microbial transport measurements. Transport of microorganisms may depend on the environmentally-driven phenotypic expression of the genetically controlled characteristics of microbial cells such as size, shape, and composition of the cell envelope. Motility of bacteria was sometimes found to contribute to transport velocities under static conditions in nutrient-saturated sand and sandstone with velocities of up to 0.4 cm h^{-1} . Motile microorganisms can migrate in response to chemical gradients. Such movement, known as chemotactic migration, is a directional motion toward or away from higher chemical concentrations of nutrients or toxicants.

Transport pathways

Bacteria and viruses move through soils with the distance travelled being dependent on the soil pore characteristics. Research results summarized as early as 1975 by Gerba et al. (1975) indicated that, under saturated conditions, whereas coliforms could travel 0.6 m in fine sandy loam soil they travelled up to 830 m in sand-gravel; bacteriophage T4 travelled up to 1.6 km in a carbonate rock terrain area. Transport of microorganisms through soil can be very fast with the average velocity of microorganisms moving through

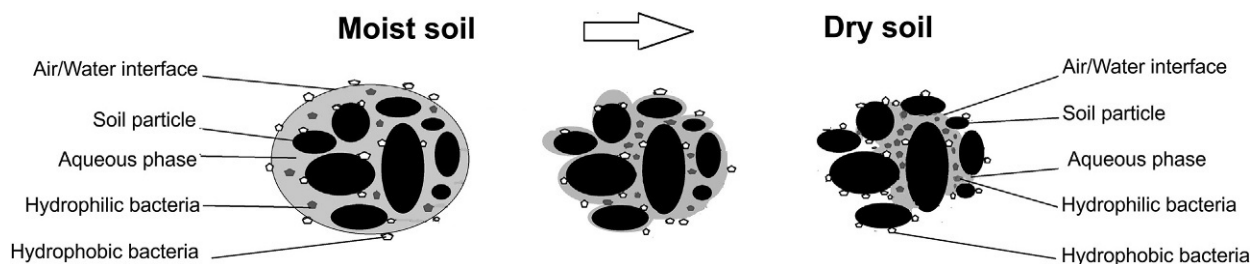


Fig. 3 Schematics of the impact of change in water content (i.e., passage from saturated to unsaturated soil conditions) on bacterial mobility and persistence.

soil greater than that of chemical tracers such as Cl or Br. This suggests mechanisms that accelerate microbial movement, and which differ from those involved in transport of soluble chemical tracers. This is attributed to preferential flow and pore size exclusion, which means that as bacterial cells are excluded from small pores, and thus micropore flow, due to their intrinsic size, they are constrained to remain in the larger pores, macropores, where water flows are fastest. Early breakthrough of bacteria, versus soluble tracers, has been confirmed for both saturated and unsaturated soils both in undisturbed and repacked columns.

Soil type (texture and structure) and preferential flow

In heterogeneous media, such as soils, bacterial transport is directly related to soil particle size distribution and texture. Bacterial cells tend to both flocculate and associate with charged surfaces, with cells rarely remaining independently suspended in the soil solution. Bacteria adsorbed by soil and organic waste colloids form colloid-bacterial complexes which create “bridges” between soil particles clogging soil pores and preventing the infiltration of other cells. Physical filtration is more likely for larger bacterial cells (0.2–5 μm) than for the smaller viruses (0.03–0.3 μm). Thus, the likelihood for soil pores to be clogged in this way depends on pore size, soil texture and aggregation, water velocity and bacterial cell dimensions. In disturbed soils, when bacterial cell size is at least 5% larger than the size of soil particles, the pores could filtrate or entrap bacteria. Straining in dead-end pores is an important process in fine-grained sediment. Water repellency of soil particle surfaces, such as can be induced by crude oil contamination, reduces attachment and enhances bacterial transport.

Thus, whereas soil texture, and associated charge properties can be used to predict bacterial movement through soils, this is more likely true for repacked or dispersed soils. In early research, well-controlled bacteria transport studies were conducted in columns filled with either glass beads or homogenized soils. Because homogenization and repacking processes destroy the natural soil structure, results obtained with repacked columns may not accurately represent the transport behavior of bacteria in the natural soils. In undisturbed natural soils, the soil structure and the aggregation dependent pore size distribution govern the effect of soil texture on contaminant transport. Reduced adsorption, a function of the reduced likelihood for bacterial cells to interact with surfaces of large pores, favors transport over retention. Thus, preferential flow due to continuous macropores and cracks is the main reason for rapid downward movement of bacteria in structured soils. In a classic example described by Smith et al. (1985), 44 to 79% of a pulse of streptomycin-resistant *E. coli* passed through a 28-cm column of undisturbed silt and silt loam soils. However, when the same soils were sieved and repacked to the same bulk density, less than 7% of the same pulse of *E. coli* passed through the columns with an equal elution volume.

Experimental observations indicate that preferential flow is the rule rather than the exception in most structured soils, and that continuous pores that are several times larger than a bacterial cell, allow bacterial transport over significant distances. Due to the preferential (bypass) flow paths that avoid the soil matrix, water can move through the soil profile faster than the mean water flux estimated by the Richardson-Richards equation. Suspended particles moved with this water flow through pores of a size that allows their passage (i.e., pore size exclusion). Therefore, suspended particles may be transported rapidly over a distance determined by the extent to which preferential pathways are continuous; namely, the extent of preferential flow has a direct influence on the distance to which particulates can be moved. In natural soils, macropores created by soil fauna (earthworms, insects, moles, and gophers) range in size from 1 to 50 mm and are usually tubular in shape. Pores formed by plant roots are also usually tubular. Cracks and fissures formed due to soil shrinkage and freeze-thaw cycles, are not usually tubular but might be significantly larger than the biopores described above. The erosive action of subsurface flows will also form natural soil pipes.

It is clear that structurally-stable fine-textured soils have greater potential for preferential flow and the rapid transport of microbial and solute contaminants than non-structured sandy soils. When fine-textured soils are disturbed, such as by tillage, the probability for interaction between bacteria and clay particles increases (i.e., increased collision rates), and the greater the clay contents, with greater surface charges, means that there is a greater the likelihood of these collisions resulting in adhesion of bacteria (increased collision efficiency). Preferential flow paths are even more important for the transport of larger cells or protozoans, such as *Cryptosporidium parvum*, that are less likely to enter small soil pores. Furthermore, in unsaturated fine-textured soil, water films and menisci are mostly connected and effectively transmit bacteria. Such observations confirm the dominant role for the physical filtration in the discontinuous pores of poorly structured soils (e.g., sands and loamy sands), whereas retention at adsorptive sites (i.e., chemical filtration) is more likely in structured, finer-textured soils. Nevertheless, when soil structure is altered, such as is common in disturbed soil columns, bacterial straining increases with clay content.

Due to variability in the transport of fecal bacteria, transport mechanisms need to be studied at both laboratory and field scales. Regularly tilling topsoil disrupts preferential flow pathways, favors pore discontinuity, and attenuates bacterial transport in soil. Thus, extrapolation of results of laboratory experiments to field conditions must be carefully considered to include multiple scales of heterogeneity due to three-dimensional flows, soil layering, interactions with native microbial communities, and temporal variations due to the transient nature of most environmental parameters. These factors usually confound the straightforward extension of laboratory results to field-scale predictions. Although controlled-system models are advantageous to understand the mechanistic details of microbial transport, field tests are required to determine whether large-scale complex environmental stochasticity may overwhelm the predictive capacity of mechanisms as described at laboratory scale.

Saturated flow transports more bacteria than unsaturated flow. More bacterial retention in the soil occurs for unsaturated flow conditions possibly due to increased area of air–water interfaces and reduced thickness of water films around the soil particles. Although most filtration occurs in the topsoil (e.g., ~20 cm), equivalent filtration was also measured for the deeper layers (e.g.,

40 cm) but this depended on variations of texture and structure down the soil profile. Under unsaturated flow, the breakthrough curve (BTC) for bacterial concentration is often bimodal with a peak immediately after application followed by a delayed, second peak. Although faster bacterial transport occurs under saturated flow, during unsaturated flow bacteria can appear earlier in the effluent than a nonreactive soluble tracer due to the pore size exclusion phenomenon.

The transport parameters can be modelled using mobile-immobile water (MIM), using attachment detachment approaches, allowing for the calculation of putative straining coefficients. Dual-permeability modelling, to account for both macropore and matrix flow, can be carried out in softwares such as in HYDRUS. Fitted attachment and straining coefficients increase with the soil clay content, and more evident if soils are repacked. Leaching from repacked columns typically produces symmetrical breakthrough curves, an indication of the dominant matrix flow. As macropores form and their proportion increases the breakthrough curves become skewed, with an early bacterial peak being followed by a long, low concentration leaching flow (i.e., tailing).

Impact of temperature

The risk of groundwater pollution due to application of liquid manure in the fall or spring has been the subject of debate. Livestock producers might prefer to apply the manure during the cold season as a means of reducing storage needs, providing more time for the application and reducing soil compaction when compared to the application on wet non-frozen soils in the spring. Due to bacterial die-off, bacterial densities in leachate after fall manure application might be significantly smaller than those after spring manure application. Nevertheless, winter application of manure is rarely recommended, as early spring excess runoff, from non-infiltrating frozen soils, may lead to manure nutrients and bacteria reaching nearby surface waters in surface runoff. Gharabaghi et al. (2015) studied the effect of ambient temperature (5 °C and 20 °C) on the saturated transport of *E. coli* NAR and Br through two soils which had been subjected to physical (freeze-thaw, and wet-dry cycles) and biological (by earthworms) weathering for 12 months. Leaching with cold water (5 °C) led to lower hydraulic conductivity and water flux and consequently enhanced bacterial filtration for both soils. Furthermore, very small values of the detachment coefficient for *E. coli* NAR at 5 °C meant an irreversible process of bacterial attachment in heterogeneous soils. Physical non-equilibrium MIM and attachment-detachment models were able to describe the BTCs of Br and *E. coli* NAR very well.

Decrease in surface soil temperature might reduce the bacterial release from manure and increase the bacterial die-off in the soil, while also affecting water viscosity.

Soil water regime

Transport fluxes in soil involve both vadose and saturated zones. Although the literature reports many field and laboratory experiments on the bacterial transport, most focus on the saturated zone transport presumably because saturated soils pose a greater risk to migration of bacteria than do unsaturated soils. Unsaturated transport is more complicated and has greater potential to remove bacteria and viruses from the flowing soil water. Water flow velocity, or more specifically pore water velocity, controls bacterial transport and filtration in soils; as the velocity of water decreases in unsaturated soils bacteria are more likely to be subjected to filtration processes in micropores and clogging in macropores necks. The opportunity for bacteria-soil interactions at lower water flow velocity also increases. Thus, retention probability, as described by collision efficiency, is largest over short distances, but eventually declines over long distances (i.e., straining rates coefficients decline with distance; Unc et al., 2012). This explains the propensity for the upper layers of soil to filter microbial cells most effectively.

Bacterial sorption and retention in soil systems

Adsorption is the main process which limits the movement of microorganisms. Bacteria are rarely free in the liquid phase of soil because most cells adhere to clay particles. Fine-grained soils, such as clays and silts, are more efficient in straining bacterial cells due to smaller pore sizes. The capacity of a soil to remove microorganisms is increased at low soil water contents and at greater soil clay contents and cation exchange capacity (CEC).

At first sight, this might suggest a repellent effect between bacteria and soil surfaces. At large electrolyte concentrations, and high thermal energy, the two surfaces can be brought sufficiently close for the short-range London-van der Waals forces to take effect. Although individually these attraction forces are weak, multiplicity of the connection points on the two charged surfaces can lead to an overall strong bonding. Presence of lipopolysaccharides (mucilage), produced by bacteria under specific conditions (e.g., starvation), may enhance their adhesion to mineral surfaces provided that the bacteria are brought in contact through other mechanisms and the contact period is long enough.

Surface charge properties of soil particles and bacterial cells

Most soil mineral and organic surfaces have a net negative surface charge. Changes in the environmental pH can change the ionization of surface groups and therefore change the electrical potential at the surface. Increase in the electrolyte concentration results in the neutralization of a greater number of charged surface loci by counter ions in the solution. This leads to smaller electrostatic repulsion forces as the electrical double layer thickness is reduced.

The charge properties of bacterial surfaces depend on the constituents of their cell wall and presence or absence of surface appendages. There are two basic organizational schemes for the bacterial cell wall: Phospholipids with negative charges, and proteins with positive charges. The relatively small amount of protein compared with phospholipids results in the bacterial wall mainly expressing hydrophobic and negative charges.

The effect of cell hydrophobicity on adhesion can be described as cell expulsion from the bulk of water and can result in large rates of bacterial adsorption at the solid–liquid interface. Hydrophobicity is positively correlated with bacteria adhesion to mineral particles, while hydrophilic strains are more likely to be found in bulk water. Hence, hydrophobic bacteria will attach to the liquid–solid interface at greater rates than would hydrophilic bacteria (Fig. 3). However, attachment is reversible, with a time scale of days or weeks. Consequently, hydrophobic strains move more slowly than hydrophilic ones, and thus any additive, such as surfactants, or changes in cell physiological state due to starvation (Borges et al., 2008), that modifies the hydrophobicity of cell surfaces will enhance transport (Liu et al., 2020). While most bacterial cells, both Gram-negative and Gram-positive, have a net negative charge, their hydrophobic character can vary within the same species or even strain.

Organic waste properties and components affecting bacterial transport

Organic wastes might affect soil physical and chemical conditions that are relevant to bacterial transport. Therefore, bacterial transport studies conducted in the absence of organic waste components may not correctly describe the transport potential of bacteria in organic waste-contaminated soil environments. Organic and inorganic components of organic wastes alter soil pH and increase salts together with soluble and insoluble organic compounds, all of which have an impact on the survival of bacteria, and hence their transport potential. Charged ions in solution can modify the electrostatic properties of both soil and microbial surfaces, altering the potential for the adsorption of microbes onto the soil. Increasing the ionic strength of the soil solution minimizes the thickness of the electrical double layer at the charged surfaces of microbial cells, of other suspended particles and soil surfaces. This phenomenon was also observed for gypsiferous soils: greater solubility of gypsum due to the presence of carbonate in the soil solution led to more *E. coli* transport in the high gypsum versus low gypsum soil (Sepehria et al., 2014).

Soluble and colloidal organic carbon components, and as well inorganic ionic compounds, can thus alter the electrochemical properties of microbial surfaces and thereby influence the likelihood that the microbes will remain in the pore water. Soluble organic compounds may compete with bacteria for retention at positively charged soil surface loci, reducing the surfaces available for bacteria sorption and enhancing bacteria repulsion. Also, soluble organic substances in the organic waste suspensions may attach to bacterial cell surfaces and diminish the water repellence of bacteria thus enhancing suspension of otherwise hydrophobic bacteria; similarly, adsorption of soluble organic compounds onto the surface of suspended organic particles will facilitate bacterial adsorption to such particles and favor their transport (Unc et al., 2012). The presence of organic particles can thus accelerate bacterial transport velocity, by as much as 10 times.

Larger, suspended organic particulates might offer surfaces for bacterial sorption which eventually facilitate co-transport of bacteria; organic electrolytes might increase the adsorption of bacteria on suspended charged organic particles. Attachment of waste-borne coliforms to soil can be inhibited in direct relation to increasing organic waste content. Greater detachment rates in the presence of waste have been reported for coliphages (Bradford et al., 2006a). Similarly, enhanced cell transport was observed for larger protozoans (i.e., *Giardia*) cells in the presence of waste material (Bradford et al., 2006b). On the other hand, large concentrations of colloidal manure compounds combined with increased ionic strength of the soil solution due to electrolytes in waste can lead to the production of larger flocs aggregating manure colloids and microbes. These flocs are then easier strained and retained within the soil pores, increasing the apparent filtration capacity of the soil. The concentration of charged waste particulates and electrolytes varies with type of waste: for example, greater ionic strength of soil solution following land application of poultry manure is more likely to increase such flocculation and thus apparent filtration. In contrast, the lower ionic strength of the soil solution after application of sewage or manure from cows can lead to more bacterial cells being transported through soil, even when the total bacterial count is smaller than in poultry manure (Mosaddeghi et al., 2009).

As with flocs, a large density of suspended organic particles might obstruct the bacteria-transmitting pores leading to more bacteria to be captured. Fibrous organic matter will encourage straining and filtration of microbes from infiltrating liquid. Moreover, bacteria survival may be extended in the presence of manure suspensions (Bradford et al., 2006a, b).

Approaches to microbial transport studies

Field studies

As indicated above, it is difficult to extrapolate results of laboratory experiments conducted on repacked soils to field conditions. Field studies are needed to explain general observations and to find solutions to specific problems, such as contamination of a well due to seepage from a nearby septic tank, or groundwater pollution because of land application of manure. Although field experiments are difficult to carry out, they are necessary to verify laboratory studies and to test conditions that cannot be simulated easily in the laboratory. An important issue is the compatibility of field results with those obtained in the laboratory. A quantitative comparison is difficult because microbial activity and mobility in the column is different from that at a field site. Thus, some researchers recommend the use of field experiments instead of laboratory soil columns to study microbial transport, as it is claimed that transport data determined from soil column experiments cannot duplicate actual soil structure in the field.

Laboratory studies

Laboratory studies increase our knowledge of microbial transport through disturbed (repacked) and undisturbed soil columns. Laboratory studies include the use of soil containers (batch) or soil columns (repacked and undisturbed). Batch experiments can examine adsorption phenomena without interference from other variables such as filtration. Repacked soil columns have been used in numerous bacterial and solute transport studies even though both the pore size distribution and pore continuity are altered. It is argued that repacked columns provide reproducibility and homogeneity, making it possible to study mechanistically different factors affecting bacterial movement in soils. The use of intact field soil columns is thus also common. While undisturbed soil cores have better predictive value for field situations than repacked ones, even relatively undisturbed columns may produce ambiguous results due to possible alteration in soil geometry during sampling. One proposed option to overcome such uncertainties is to use repacked columns that are allowed to weather before experimental use thus permitting for the natural re-formation of aggregates and pore networks which might then provide a more realistic, but controlled environment for transport simulations (Safadoust et al., 2011).

Model studies

Mathematical modelling of microbial transport through soils is an important tool in evaluating the long-term risk of undesirable microbes entering soil and groundwater and in identifying optimal sites for waste and wastewater disposal. Many mathematical descriptions of microbial transport through soils, such as attachment and detachment theory and filtration theory, were developed on the basis of the convection-dispersion equation (CDE). HYDRUS software is capable of simulating the transport of viruses, colloids, or bacteria by kinetic attachment and detachment together with filtration (chemical non-equilibrium) models. Two sorption-loci can be used to simulate sorption to solid phase and removal of particles due to attachment to air-water interface. In these models, first-order attachment coefficient (due to filling of favorable sorption sites) and first-order detachment coefficient are used to simulate the bacteria-solid interactions.

Role of new generation sequencing (NGS) for advancing microbial transport studies

While model development for microbial transport through soil has benefited from experiments with single, well-defined organisms, the advent of new generation sequencing (NGS) tools allows for population level and multispecies stochastic analysis of transport. Moreover, this can also provide insights into the transport of microbially associated functional genes (Bak et al., 2019) and thus allowing for the cell transport information to be augmented by information on the soil transfer of microbially associated functions. An observation that was possible due to NGS tools is that the microbes in the preferential flow paths, thus the ones likely to be transported, have functions distinct from microbes in smaller, soil matrix pores, which are less likely to be transported (Bak et al., 2019). Variability in probability for mobilization and transport of distinct bacterial species can also be rapidly parsed by assessing of the changes in the microbial community in soil flow via NGS tools (Dibbern et al., 2014). These examples highlight the utility of NGS tools to expand the scope of microbial transport modelling from the realm of contaminant microbiology to questions relevant to the vertical transfer of microbial functions in managed and natural soils. This creates the basis for unifying microbial transport knowledge, developed for assessing contaminant microbial transport, with the research modelling microbial processes relevant to geochemical processes.

Summary

Organic wastes, such as animal manures, are C-rich beneficial amendments used to improve soil physical and chemical conditions relevant for plant growth. However, application of such organic amendments to agricultural lands is recognized as a major source of pathogenic microorganisms released into surface and groundwater systems. We reviewed the interrelated factors affecting survival and transport of pathogens in the soil systems. Organic waste properties and components can significantly alter the physical and chemical properties of the soils and bacteria. Physiological state of bacteria cells can also affect their electrostatic surface properties. Soil type and water status play a dominant role in the fate of pathogenic microorganisms. Preferential flow due to soil structure and water repellency facilitate the bacterial movement through natural soils in the environment.

References

- Bak F, Nybroe O, Zheng B, Badawi N, Hao X, Haubjerg Nicolaisen M, and Aamand J (2019) Preferential flow paths shape the structure of bacterial communities in a clayey till depth profile. *FEMS Microbiology Ecology* 95(3): fiz008.
- Borges MY, Nascimento AG, Nunes Rocha U, and Tótola MR (2008) Nitrogen starvation affects bacterial adhesion to soil. *Brazilian Journal of Microbiology* 39(3): 457–463.
- Bradford SA, Tadassa YF, and Jin Y (2006a) Transport of coliphage in the presence and absence of manure suspension. *Journal of Environmental Quality* 35: 1692–1701.
- Bradford SA, Tadassa YF, and Pachepsky Y (2006b) Transport of *Giardia* and manure suspensions in saturated porous media. *Journal of Environmental Quality* 35: 749–757.

- Dibbern D, Schmalwasser A, Lueders T, and Totsche KU (2014) Selective transport of plant root-associated bacterial populations in agricultural soils upon snowmelt. *Soil Biology and Biochemistry* 69: 187–196.
- Farhangi MB, Safari Sinegani AA, Mosaddeghi MR, Unc A, and Khodakaramian G (2013) Impact of calcium carbonate and temperature on survival of *Escherichia coli* in soil. *Journal of Environmental Management* 119: 13–19.
- Farhangi MB, Safari Sinegani AA, Mosaddeghi MR, Unc A, and Khodakaramian G (2014) Survival of *Pseudomonas fluorescens* CHA0 in soil: Impact of calcium carbonate and temperature. *Arid Land Research and Management* 28: 36–48.
- Foppen JW and Schijven JF (2006) Evaluation of data from the literature on the transport and survival of *Escherichia coli* and thermotolerant coliforms in aquifers under saturated conditions. *Water Research* 40(3): 401–426.
- Gerba CP, Melnick JL, and Wallis C (1975) Fate of wastewater bacteria and viruses in soil. *Journal of the Irrigation and Drainage Division* 101: 157–174.
- Gharabaghi B, Safadoust A, Mahboubi AA, Mosaddeghi MR, Unc A, Ahrens B, and Sayyad G (2015) Temperature effect on the transport of bromide and *E. coli* NAR in saturated soils. *Journal of Hydrology* 522: 418–427.
- Liu G, Fang Z, Zhong H, Shi L, Yang X, and Liu Z (2020) Transport of *Pseudomonas aeruginosa* in porous media mediated by low-concentration surfactants: The critical role of surfactant to change cell surface hydrophobicity. *Water Resources Research* 56: , e2019WR026103.
- Mosaddeghi MR, Mahboubi AA, Zandsalimi S, and Unc A (2009) Influence of organic waste type and soil structure on the bacterial filtration rates in unsaturated intact soil columns. *Journal of Environmental Management* 90(2): 730–739.
- Safadoust A, Mahboubi AA, Gharabaghi B, Mosaddeghi MR, Voroney P, Unc A, and Sayyad G (2011) Bacterial filtration rates in repacked and weathered soil columns. *Geoderma* 167/168: 204–213.
- Sepehria N, Mahboubi AA, Mosaddeghi MR, Safari Sinegani AA, and Khodakaramian G (2014) *Escherichia coli* transport through intact gypsiferous and calcareous soils during saturated and unsaturated flows. *Geoderma* 217–218: 83–89.
- Smith MS, Thomas GW, White RE, and Ritonga D (1985) Transport of *Escherichia coli* through intact and disturbed soil columns. *Journal of Environmental Quality* 14: 87–91.
- Unc A, Goss MJ, Cook S, Li X, Atwill ER, and Harter T (2012) Analysis of matrix effects critical to microbial transport in organic waste-affected soils across laboratory and field scales. *Water Resources Research* 48: W00L12.

Biology of compost

Heribert Insam^a, Thomas Klammsteiner^{a,b}, and María Gómez-Brandón^c, ^aDepartment of Microbiology, University of Innsbruck, Innsbruck, Austria; ^bDepartment of Ecology, University of Innsbruck, Innsbruck, Austria; ^cGrupo de Ecología Animal (GEA), Centro de Investigación Mariña (CIM), Universidad de Vigo, Vigo, Galicia, Spain

© 2023 Elsevier Ltd. All rights reserved.

This is an update of T.L. Richard, COMPOST, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 294–301, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00142-9>.

Introduction	523
Definition of composting (process) and of compost (product)	523
Scope of this chapter	523
The substrates	523
Cellulose	524
Lignin	524
Hemicelluloses	524
Chitin	524
Murein	524
Other compounds like collagen or keratin	524
Supplements	525
The organisms and process phases	525
The commonly recognized composting at farm- and industrial scale	525
Initiation phase	525
High temperature phase (35–65 °C)	526
Cooling phase	526
Maturation phase	526
Curing phase	526
When is compost mature?	526
Vermicomposting	527
Black soldier fly composting	528
The products	529
Added value, carbon and nutrients	529
Method of quality assessment	529
Effects on soils	529
Microbial diversity, macro and micropores, aggregates and structural stability	530
Climate related issues	531
Conclusion	531
References	531

Abstract

The chapter addresses the source materials for composting processes, the microbes and phases involved in the process and quality criteria for composts. The focus is on composting in its common meaning, and two fauna-driven processes involving compost worms and the black soldier fly. Further, some inorganic amendments like biomass ashes and biochar are discussed. Ultimately, the overall effects of composts on soil, its microbiota and plants are discussed, and put in context with carbon mitigation and climate effects.

Glossary

Frass In natural environments, frass describes the excreta of insects. From the perspective of insect farming, however, frass is defined and marketed as an inhomogeneous mixture of feeding substrate residues, insect feces and body parts, exuviae, and dead insects that is collected after separating the harvest-ready insects at the end of the rearing process. Frass application introduces beneficial microorganisms to the soil and its chitin fraction is considered a resistance-promoting agent for plants.

Key points

- Substrates for composting processes
- Microbial vs. faunal-driven composting

- Effects of composts on soils, their microbiota and plants
- Composts for mitigation of climate warming

Introduction

In the context of global warming and a renewed debate on fossil energy, sustainable agricultural, horticultural and silvicultural approaches are receiving more attention. Indeed, the recycling of carbon (C) and nutrients to soil may alleviate the extensive use of fossil resources. Apart from rendering the soil as a sink for C, we should keep in mind that ~2% of the global anthropogenic energy use is attributed to the Haber-Bosch process to collect nitrogen from the atmosphere. With increasing prices for fertilizers, organic amendments such as manures or composts are becoming more attractive as sources of nutrients and organic matter.

Soil degradation as a result of extensive agricultural or horticultural use, combined with poor management, remains a major concern and often related to the loss of organic matter and a low level of nutrient availability. Composts may either be incorporated into the soil or used as mulch. Both approaches have advantages and disadvantages. Often, the effects are best detected in the vicinity of the applied material, which is of practical interest as roots grow toward such hotspots where they take immediate advantage of the nutrients released from the composts (Duong et al., 2013).

Definition of composting (process) and of compost (product)

Etymologically, composting (from Latin, *compositum*, mixture) refers to a biodegradation process, not of a single material but of a mixture of substrates. In addition, the process is carried out by a mixed community of organisms, be they only microorganisms, or microorganisms and fauna. The microbial conversion of pure substrates would be called fermentation or bio-oxidation, but not composting. Composting is commonly carried out on the basis of solid substrates (also rich in lignocelluloses), and is generally understood as an aerobic process. The exothermic process generates energy in the form of heat, resulting in an increase in temperature. A spontaneous process characteristically goes through five phases, an initial mesophilic phase, followed by a thermophilic and subsequent cooling (second mesophilic) phase, maturation and a terminal curing phase. During composting a temporary release of phytotoxins (intermediate metabolites, ammonia, etc.) is typical, while in the terminal phase this phytotoxicity is overcome and the end product benefits plant growth. The composting process results in CO₂, water, minerals and stabilized organic matter (often referred to as 'humic' material, a concept that has, however, been challenged by Lehman and Kleber, 2015). In brief, composting is based on a mixture of substrates and degrading organisms, particularly microorganisms, and is characterized by a constantly changing physico-chemical environment, including nutrient availability, pH, temperature and composition of intermediates (Insam and De Bertoldi, 2007; Insam et al., 2002). Although some authors term anaerobic treatment (in particular biomethanization) anaerobic composting, this process is explicitly not covered in the present chapter. The end product of anaerobic processes is called digestate, and does in many respects have different effects on soils than do composts (Gómez-Brandón et al., 2016). Post treatment of digestates by composting, however, may well be used to stabilize the material, again resulting in a product termed compost.

The main product of the process is called compost, which can be defined as stabilized and sanitized, compatible with and beneficial to plant growth. The conversion of waste organic material into compost is carried out for three main reasons, (i) to overcome the phytotoxicity of fresh, unstabilized organic material, (ii) to reduce the presence of pathogens (viruses, bacteria, fungi, parasites pathogenic to humans, animals and plants) to a level that does not pose a further health risk, and (iii) to produce an organic soil amendment that serves the purpose of nutrient addition and physico-chemical conditioning. This way biomass in the form of organic waste and crop residues is recycled.

Scope of this chapter

There is a plethora of different methods of composting, ranging in size from miniature backyard to huge industrial operations, and involving many different substrates from agricultural residues to kitchen waste and municipal sewage sludges. There are rapid processes, and others that take more time; there are pure microbiologically driven processes and ones that exploit the activities of mesofauna. In this chapter there will be a focus on medium- to large-scale, farm-scale and commercial composting, as well as on vermicomposting and organic waste treatment by the black soldier fly.

The substrates

Basic for any composting process is the availability of degradable organic matter which is usually of plant origin, but may also stem from faunal tissue or microbial cells. The basic materials are presented below, each with its own properties for the process. Owing to the biochemical composition of the substrate material, the composting process runs differently, and also the resulting compost quality may be different. Basic substrates range from manure or other agricultural or agro-industrial residues, to domestic organic

waste from source-separated collection and slurries from wastewater treatment plants. Also, mixtures of such wastes are often used, including lignocellulosic materials (bush- and tree cuttings), which serve as bulking agents, or digestates from biomethanization plants.

Cellulose

Cellulose as the most abundant component of plants is found in many types of organic waste. Waste rich in structural elements (such as wastes from the wood industry and straw from agriculture) has the highest cellulose content, while domestic organic (kitchen) waste contains less, and wastes derived from microbial or animal sources contain no cellulosic material. Cellulose molecules are chains of β -1,4-glycosidic bonded glucose with a polymerization degree of up to 40,000. Enzymatic degradation is brought about by the activity of three enzymes, (i) Endo- β -1,4-glucanase, (ii) *exo*- β -1,4-glucanase and (iii) β -glucosidase, the resulting glucose is taken up by the microorganisms and used for energy and as base molecules in anabolic processes.

Many fungi and bacteria are involved in the breakdown of cellulose. The catalytic effect (mechanical destruction) of micro- and mesofauna may be an important contributing factor. Fungi are generally more important than bacteria in breaking down cellulose, especially when cellulose is encrusted in lignin (e.g., in straw or wood). Because cellulose is rich in C but lacks N or other essential elements, the hyphal structure of fungi has a competitive advantage as the hydraulic force of fungal hyphae can break up lignocellulosic material. *Chaetomium*, *Fusarium* and *Aspergillus* are important fungi to mention in this context; cellulose-degrading bacteria mainly belong to the Myxomycetes or related classification groups (*Cytophaga*, *Polyangium*, *Sorangium*). *Pseudomonas* and related genera are also known to degrade cellulose, as are Actinobacteria.

Lignin

Lignin is a most recalcitrant structural component of plants, making up 2–30% of their biomass. It is composed of few monomeric units (phenylpropane derivatives, mainly coniferyl and sinapyl alcohol) but, due to the extraordinary variety of chemical bonds involved, its degradation is complex and yields little energy. Very often, the degradation of lignin is of a co-metabolic nature, primarily accomplished by fungi, in particular white-rot fungi such as *Trametes versicolor* or *Stereum hirsutum*. They break down the lignin, leaving behind cellulose parts that are light in color. Some fungi, such as *Pleurotus ostreatus*, degrade both cellulose and lignin.

Hemicelluloses

Among the hemicelluloses, xylan is most important, found in straw, bagasse and wood. It contains pentose sugars, such as arabinose and xylose, as well as hexoses, including galactose, glucose and mannose with a degree of polymerization of 30–100. Pectin, made up from unbranched chains of polygalacturonic acid, is degraded by pectinase, commonly produced by fungi and bacteria. Starch contains amylose (20%) and amylopectin. The former are unbranched helical (due to 1,4-position β -glycosidic bonding) chains of α -glucose, while the latter, containing phosphate residues and Ca and Mg ions, is also branched at the 1,6-position. Two types of enzymatic degradation are important, phosphorolysis (releasing glucose-1-phosphate) and hydrolysis (α -amylase cleaving the α -1,4-bonds within the molecule).

Chitin

Cellulose and chitin are chemically very similar. In chitin, glucose is replaced by N-acetylglucosamine, which makes a difference for degraders since chitin contains approximately 7% N (C:N-ratio = 7). Many fungi and bacteria use chitin both as an N and C source. Chitin is degraded to N-acetylglucosamine, which is resorbed, transformed to fructose-6-P and channeled into the carbohydrate metabolism. Chitin is an important structural compound in fungal cell walls, makes up the exoskeleton of crustaceans and insects, making it an important waste product from pharmaceutical and shellfish industries.

Murein

Murein, consisting of unbranched chains of N-acetylglucosamine and N-acetylmuramic acid bound to variable amino acids. Murein is the main cell wall component of most bacteria, and thus only a minor component of most waste streams. It may, however, be an important component of wastes derived from the pharmaceutical industry, and in wastewater slurries.

Other compounds like collagen or keratin

Proteinaceous compounds stemming from animal wastes serve as slow-release nitrogen sources and contribute to the value of the resulting compost. Often, they are contained in the digestates of biogas plants that are using slaughterhouse wastes.

Supplements

The supplementation of composts with by-products, such as biomass ash (Kuba et al., 2008; Bougnom et al., 2010) and biochar (Agegnehu et al., 2017) serves two purposes, the reduction of waste and the improvement of the product. Biomass ashes provide major nutrients like phosphorus (P), a multitude of micronutrients and also serve as a lime replacement, while biochar provides exchange sites, micro-niches for microorganisms and improves the physical structure of composts. There is consensus that the mixing of composting substrates and supplements prior to the composting process adds additional benefit compared with mixing prior to soil amendment. Other supplements, such as rock dust, are known for similarly positive effects, mainly through providing nutrients and facilitating the formation of mineral-organic complexes (aggregates). It has been reported that not only the soil microbiome is changed, but the effects extend as far as improving the vitamin C content in apples (Li et al., 2021).

The organisms and process phases

The commonly recognized composting at farm- and industrial scale

In contrast to small-scale backyard processes, larger scale operations do usually not allow for micro- or mesofauna driven processes, but rely on purely microbially driven turnover. This is because of the speed of the process (6–12 weeks) and the need for mechanical disturbance by either periodic or continuous turning. Such operations mostly rely on a well-designed process regulation based on measurements of carbon dioxide or oxygen content, temperature, pH and moisture. Also, pretreatment of the substrate material, including cutting, milling and sieving plays an important role, as does active or passive aeration, and moisture control. Five phases are typically involved (Fig. 1).

Initiation phase

In this first phase (also called starting phase or first mesophilic phase), easily degradable compounds like sugars and proteins are abundant and are initially degraded by mesophilic or thermotolerant (20–35 °C) fungi and bacteria, including actinobacteria. Ryckeboer et al. (2003) found few mesophilic fungi in source-separated household waste, but numerous thermophilic fungi and bacteria. Initially, food wastes containing vegetable residues often have a low pH (4.5–5.0) that stimulates the proliferation of fungi and yeasts. Ammonification causes an increase in pH, favorable for bacteria that subsequently colonize the substrate. Wastes that are rich in lignocellulosic matter show a different composition, still with a high microbial diversity but having a bias toward fungi.

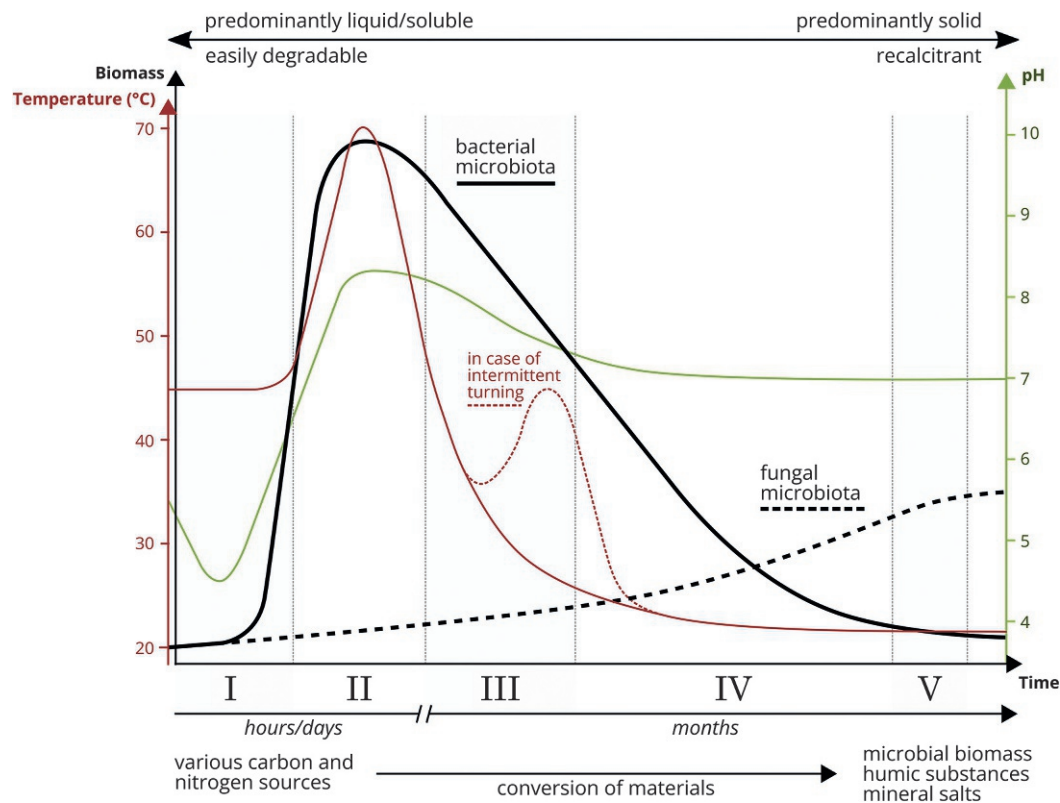


Fig. 1 Biomass, temperature and pH as they typically evolve in a composting process. The graph also shows dominant microbiota at two stages, as well as main material conversions.

However, due to the more rapid growth of bacteria, and a metabolism-driven temperature rise, fungi and yeasts are soon outcompeted. The high surface:volume ratio of bacteria allows a rapid transfer of soluble substrates into the cells. Bacteria are taxonomically and nutritionally the most diverse group, equipped with a broad range of enzymes to degrade a variety of organic materials. Also, the bacterial generation time is much shorter than those of fungi, which gives them a competitive advantage. Actinobacteria tend to grow slower than many other bacteria and are ineffective competitors when nutrient levels are high (Beffa et al., 1996); they thrive, however, as soon as the most easily degradable substrates are exploited. Due to their ability to produce various antibiotic substances, actinobacteria (e.g., streptomycetes) are most important for hygienization, inhibiting the growth of many human, animal and plant pathogens. If mechanical influences (like turning) are very limited, fauna, such as compost worms, mites, millipedes, and mesofauna develop, which mainly act as catalysts through their ability to physically disrupt substrates. Important regulators of microbial activities are protists, as well as viruses, including bacterial and fungal phages.

High temperature phase (35–65 °C)

When process temperature rises, thermophilic and thermotolerant organisms get a competitive advantage and gradually replace the mesophilic biota. Previously flourishing mesophiles die off and are eventually degraded by the succeeding thermophiles. Overall bacterial species diversity falls significantly during the thermophilic phase. The number and diversity of thermophilic actinobacteria increases, however, particularly at temperatures of 45–55 °C (Amner et al., 1988; Beffa et al., 1996). The decomposition runs fast and accelerates temperature increases up to about 62 °C. The rise in temperature is another reason why bacteria outcompete fungi. Temperatures exceeding 60 °C inhibit fungal growth, but they may survive as spores. Yeasts disappear during the thermophilic phase, but when the temperature cools down to 54 °C, they reappear (Ryckeboer et al., 2003). Endospore-forming bacteria, such as *Bacillus stearothermophilus*, are often considered typical for composting processes. They are very active at temperatures higher than 50 °C or even 60 °C (Ryckeboer et al., 2003). Non-spore forming bacteria such as *Hydrogenobacter* spp. and *Thermus* spp. are among the dominant ones in the hottest phase of the process, active at temperatures >70 °C and up to 82 °C.

Thermal inactivation of pathogens is required to obtain safe products, both in terms of phytohygiene and human diseases. The higher the temperature, the more efficient will be the destruction of pathogens (Ryckeboer et al., 2003) albeit temperatures exceeding 80 °C may trigger the Maillard reaction, leading to undesired odor and microbial inhibition. Millner et al. (1987) showed that pathogen suppression was more efficient in composts produced at 55 °C than at 70 °C. Additionally, regrowth of mesophilic populations may be delayed if too high temperatures are maintained for a long time and if this temperature is reached throughout the composting material, calling for appropriate reinoculation schemes.

Cooling phase

As soon as the easily available substrates are exhausted, the temperature starts to decrease. Mesophilic organisms recolonize the substrate, originating from surviving spores, through spread from protected microniches, or airborne inoculation, or as organisms attached to turning or sieving machinery. In this medium-temperature phase the available substrates are much different to those in the first low temperature phase. They are largely dominated by secondary degradation products, and recalcitrant polymers, such as starch, cellulose or lignin, and also thermally inactivated microbial cells.

Maturation phase

Lignin, cellulose and other recalcitrant compounds are decomposed by specialized microbiota equipped with ligninases, peroxidases and cellulases. The maturation phase is important for the build-up of humic acids that are either generated from aromatic compounds derived from the lignin, or are synthesized de novo. The humic matter is among the most valuable components of composts. During maturation, fungi are starting to outcompete bacteria, of which *Arthrobacter* form a numerically important fraction. Species diversity seems to be correlated with stability of the end product (Beffa et al., 1996). After incorporation into soils, immature composts may induce enhanced microbial activity and cause oxygen deficiency and a variety of phytotoxicity problems to plant roots.

Curing phase

Some researchers refer to a post-maturation, so-called, curing phase. In this phase, it is hardly possible to diagnose changes in physical or chemical compost quality. However, changes in the microbiota may still occur to the benefit to the product (Danon et al., 2008) but also to its disadvantage, for example, because of decreased plant disease suppressiveness (Franke-Whittle et al., 2019).

When is compost mature?

Classical maturity tests are discussed in detail by Itävaara et al. (2002) and Insam and de Bertoldi (2007). Benefits of compost application, such as suppression of diseases caused by soilborne plant pathogens, have been reported in agricultural and horticultural crops. Composts in their early stages often exert phytotoxic effects that may be associated with ion concentrations,

Table 1 Rottegrad, definition of process conditions.

Rottegrad	T_{max} in °C	Product
I	>60	Fresh compost
II	50–60	
III	40–50	
IV	30–40	Mature compost
V	<30	

excessive availability of heavy metals, toxic organic compounds or an adverse phytopathogenic microbiota. Phytotoxicity is defined as a delay of seed germination, inhibition of plant growth through, e.g., phytotoxins. During compost maturation these effects are commonly alleviated, rendering the composts compatible with plants and plant roots. Compost maturity tests based on phytotoxicity is a well-established methodology, worldwide (Bernal et al., 2017), as are tests based on ammonia or carbon dioxide release, or self-heating tests in Dewar flasks (Rottegrad test; Table 1) (Itävaara et al., 2002; Changa et al., 2003).

Vermicomposting

In general, small-scale composting favors faunal-driven composting, in contrast to large-scale operations that favor microbially dominated processes. Only in special cases, like vermicomposting or black-soldier-fly (BSF) driven operations, does large scale also enable a faunal-driven process.

Unlike composting, vermicomposting does not include a thermophilic phase and relies on the synergistic actions of earthworms and microorganisms to accomplish the breakdown of the organic matter. Earthworms have been classified, on the basis of their feeding and burrowing strategies, into three ecological categories: epigeic, anecic and endogeic. In particular, five epigeic earthworm species have been extensively used in vermicomposting facilities known as *Eisenia andrei* (Savigny), *Eisenia fetida* (Bouché), *Dendrobaena veneta* (Savigny) and, to a lesser extent the tropical species *Perionyx excavatus* (Perrier) and *Eudrilus eugeniae* (Kinberg). These earthworm species live in the soil organic horizon, in or near the surface litter, and mainly feed on fresh organic matter contained in forest litter, litter mounds and herbivore dung, as well as in man-made environments such as manure heaps. All display a natural ability to colonize organic wastes and are characterized by short life cycles, high reproductive rates and a high rate of consumption, digestion and assimilation of OM.

Vermicomposting systems sustain a complex food web that results in the recycling of organic matter (OM) and the release of nutrients (Domínguez et al., 2010). However, more than a century passed before vermicomposting was considered a viable technology, despite the important role of earthworms in the decomposition of dead plants and nutrient cycling. During vermicomposting, epigeic earthworms and microorganisms intensively interact accelerating the stabilization of OM and greatly modifying its physical and biochemical properties. Although microbes are responsible for the biochemical decomposition of the OM via the production of extracellular enzymes, earthworms are key players in the process through their effects on microbial communities. They may influence microbial decomposer activity by grazing directly on microorganisms, or by increasing the surface area available for microbial attack after the comminution of organic matter. Earthworm activity favors the aeration and fragmentation of OM, thereby enhancing its turnover and increasing the rate of decomposition. They can also affect other fauna directly through the ingestion of microfaunal groups (protozoa and nematodes) that are present in the detritus consumed, or indirectly by modifying the availability of resources for these groups.

In the short-term, the earthworm gut-associated processes, that is the ingestion, digestion and assimilation of OM and then casting, are major drivers of the vermicomposting process. Earthworms excrete large amounts of egested materials known as casts that result in an inoculum of nutrients and microorganisms into the environment. The contact between these materials and the raw substrate has been found to alter the nutrient dynamics and the rate of OM decomposition, similar to that when earthworms are present. The changes occurring during these initial stages of vermicomposting largely determine the properties of the final product and its potential usefulness as an organic amendment. A period of aging is also necessary to obtain a mature, phytotoxic-free vermicompost for its proper use to promote plant growth and suppress plant diseases. The duration of the maturation phase is not constant and may vary depending on how efficiently the initial phase of vermicomposting is performed.

Vermicompost is considered a nutrient-rich, peat-like material characterized by a large porosity and water holding capacity, but a small C:N ratio. Beneficial effects of vermicompost have been documented for a wide variety of agronomic and horticultural crops, when used as an amendment for soil or plant growth media. These positive effects have also been reported for extracts and teas made from vermicompost (Gómez-Brandón, 2015), even though they do not impart the same physical properties as solid vermicompost. Various factors such as an improved availability of air and water, the presence of plant-growth regulating substances, and the mitigation or suppression of plant diseases have been proposed as plausible, albeit not exclusive, mechanisms by which such improvement is achieved. There is also evidence of microbial-based mechanisms that may explain the positive influence of vermicompost on soil and plants (Gómez-Brandón et al., 2020). Compared with the initial substrate,

the resulting vermicomposts exhibited a greater functional diversity and higher abundance of putative genes related to specific metabolic processes, i.e., biosynthesis of antibiotics or plant hormone synthesis, potentially beneficial for plant growth and development.

Black soldier fly composting

Insects are increasingly gaining scientific and economic interest as agents for organic waste management. In terms of fertilizer production, converting organic waste into a stable soil amendment follows a similar principle as vermicomposting. In contrast to earthworms, however, industrially used insects are not primarily seen as producers of compost but rather as bioreactors capable of transforming low-value substrates into a source of value-added products that include protein, fat, chitin, and they also yield a by-product with fertilizing properties. Most of these products are directly derived from mechanically or chemically processing the raw insect biomass. The residue fraction, however, accumulates as the major by-product of the rearing process and has recently been defined by the European Commission as a mixture of digested and undigested substrate wastes, shed skins, and also to the smaller portion of dead insects. The beneficial effect of insect droppings on soil nutrient dynamics and their effect on soil respiration was described in detail in the past. With the emergence of insect farming, these benefits are now being exploited on an industrial scale. Thus, the term “frass,” originally used to describe these insect droppings in natural environments, has been adopted to include the rearing residues and is nowadays commonly used to describe this complex insect-derived compost.

Among the few species approved for commercial insect farming in the EU, the black soldier fly (*Hermetia illucens* Linnaeus, Diptera: Stratiomyidae) is considered the most promising contender to drive the worldwide expansion of this emerging industry. Also, in natural environments, the progressing globalization and increasing average temperature contribute to the cosmopolitanism of this subtropical fly, thereby driving its dispersion across all continents except Antarctica. In recent years, efforts were mainly directed toward determining organic waste streams from industries that are suitable for bioconversion by black soldier fly larvae (BSFL). Primarily agro-industrial by-products (e.g., wheat bran, spent grains) or fruit and vegetable wastes find current application in Western countries. Due to the novelty of large-scale insect farming, legal restrictions, and immature research on the safety of insect-derived products limit the use of a broader variety of substrates that would otherwise also include animal manure or food wastes (Liu et al., 2022).

In contrast to compost, which typically matures from an inconsistent and diverse mixture of organic wastes into a largely homogenous product with comparable properties, most of the frass physicochemical properties and nutrient content vary markedly according to the rearing substrates employed (Klammsteiner et al., 2020). Therefore, the blend of organic compounds making up the rearing substrate must be well described and available in consistent quality to ensure efficient production, as the ratio of its ingredients will consequently determine the quality and fertilizing properties of the frass. The conversion of substrate into insect biomass results in a reduced N and P content in the residues compared to the original material, though the bioavailability of the nutrients is increased through insect digestion. Commercial frass from BSFL shows on average an N:P₂O₅:K₂O ratio of 1:0.9:1.1 with a relatively high C:N ratio of 14.7 (Gärtling and Schulz, 2022). In comparison with other insect species, BSFL frass contains significantly larger concentrations of N (<130%) and K (<190%) (Beesigamukama et al., 2022). Although BSFL can thrive in a wide range of pH levels, the pH in frass is among the few parameters that show little variation across production systems and starting substrates (Gärtling and Schulz, 2022). Even frass from initially acidic substrates (pH 4) becomes alkaline (up to pH 9) during the process, approximating the high pH in the larvae's posterior gut (Bonelli et al., 2019).

Frass from BSFL also induces a higher germination rate and germination index than rearing residues from other industrialized insect species (Beesigamukama et al., 2022). This might be favored by the presence of plant growth-promoting rhizobacteria and other beneficial microbes (Barragán-Fonseca et al., 2022). The frass itself has been shown to have inhibitory effects on multiple plant pathogens and shows insecticidal properties (Vickerson et al., 2015). This effect is supported by the innate microbiota of the frass, in which the microbial community becomes more similar to the one inhabiting the BSFL guts as more and more of the substrate is digested (Gold et al., 2020). Especially bacteria (Gammaproteobacteria and Bacilli) and fungi (Ascomycota) present in the frass support the larvae in degrading the organic matter by driving aerobic fermentation (Barragán-Fonseca et al., 2022). This process could be further improved by inoculating the substrate with specific species such as *Bacillus subtilis*, as supplementing these bacteria is capable of enhancing frass maturation, seed germination, and enzymatic activity (Xiao et al., 2018). Unlike compost and vermicompost, frass also contains substantial amounts of chitin from insect skins and exuviae which not only serve as source of plant nutrients, but also enhance plant defense mechanisms by positively stimulating the soil microbiota (Barragán-Fonseca et al., 2022). While the inactivation of pathogenic microorganisms and phytotoxins in compost relies on one or more thermophilic phases, pathogens in BSFL frass are inhibited by antimicrobial peptides excreted by the larvae (Klammsteiner et al., 2020). To ensure the biological safety of frass, a first legal standard that requires a 70 °C heat treatment as post-processing measure has been recently introduced by the EU commission. To achieve this, frass could either be dried or composted prior to commercialization or soil application. However, frass that was treated via aerated composting was shown to cause 20% higher greenhouse gas emissions compared to its direct use, but could offer a feasible way to enhance nutrient availability and biosafety (Song et al., 2021). Any such post-treatment of frass, however, counteracts the very positive effects of untreated frass, as explained above, and a re-evaluation of the respective EU legislation.

Several studies ranging from pot to field scale indicate that frass application can achieve similar yields as common nutrient application regimes, however, with less negative consequences for the environment than, for example, manure (Liu et al., 2022).

Current environmental challenges associated with the use of manure or inorganic NPK fertilizers such as increased leaching or eutrophication could be alleviated by supplementing or partially substituting them with frass. At field-scale, a combined use has shown to enhance crop yields and reduce losses from plant diseases due to the promotion of several plant defense mechanisms (Barragán-Fonseca et al., 2022). With further expansion of the insect farming sector and additional efforts for standardizing rearing substrates and rearing systems, the high variation in macro- and micronutrients in frass is expected to decrease. N and P can be recaptured from the food chain by integrating these residues as biofertilizers in agriculture, thereby closing nutrient and material cycles according to the principles of the circular economy.

The products

Added value, carbon and nutrients

Beyond the delivery of nutrients, most of them in a slow-release-form, composts confer numerous additional benefits to soil that range from increased soil C stocks and increased water holding capacity to plant disease suppressiveness. All the effects of composts, however, are determined by the available input materials, and the prevailing process conditions. Thus, the above mentioned three levels of control, quality of input materials, process and end product, are required for reproducible end products with their distinct effects on soils and plants.

Method of quality assessment

Compost quality is defined both by the quality of the input material and the resulting end product, as well as by the process conditions. The composting process is designed to secure an 'end-of-waste status' for the substrates. Input material, ranging from lignocellulosic wastes to wastes with a higher proportion of proteins, sugars or other short-chained carbohydrates (see above), have inherent elementary properties. Apart from nutrients and C, they these also may contain heavy metals and xenobiotics and human, animal and plant pathogens. In particular, maximum allowed heavy metal contents are regulated, both in the substrate sources and in the end product in case of a destined use in agriculture, horticulture or forestry. The legal conditions, however, vary from country to country. For xenobiotics and pathogens, the composting process is critical for their elimination. Thus, process conditions are often strictly defined, such as the process temperatures that must be reached.

Effects on soils

An advantage of compost over mineral fertilizers is its ability to slowly release nutrients into its environment, often in line with plant needs, which increase as spring temperatures rise. As soon as the soil warms up and the plants start to grow, the activity of the microorganisms that release the nutrients also increases. Slow-release mineral fertilizers were thus developed to give mineral fertilizers this beneficial property, inherent to organic fertilizers. A disadvantage of compost is that the entire nutrient content is not released in the first year of use, but only fractions of it, which complicates the annual calculation of requirements (Insam and Merschak, 1997).

In addition to the nutrient content, compost has other properties that makes it valuable: compost improves the condition of the soil. It has a favorable influence on soil structure and soil texture (Fig. 2). As a result, nutrients, water and air are better retained, which ultimately increases yields in horticulture and agriculture.

In good soil, sand, clay, and loam mix together into aggregates that allow aeration and drainage. Ideally, the earth is granular and crumbly; a large sand content makes the soil too coarse-grained but with a high clay content it may be too compact. Too much clay

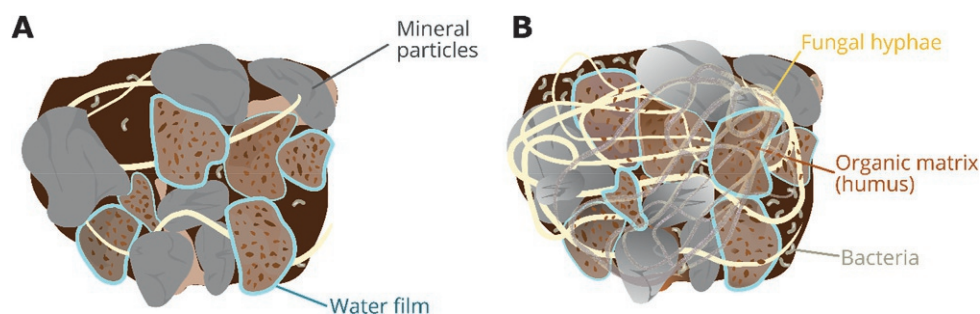


Fig. 2 Visualization of the effects of compost addition on the physical and microbiological properties of soil aggregates. Left: depleted soil with poorly developed pore structure and microbial diversity, right: soil that has received compost: formation of stable soil aggregates hosting pores of various sizes and an enhanced microbial diversity.

can promote waterlogging and the soil becomes impermeable to air. Compost creates pores of various sizes that fill with air and water, rendering them available for microbial metabolism and microbe proliferation, ultimately offering an ideal substrate for plants. On the other hand, if the soil is too sandy, the water drains away too quickly. In this case, the addition of compost contributes to the formation of larger granules that are surrounded by a film of water on their surface. The addition of compost compensates for a high proportion of sand or clay and promotes the formation of aggregates (clay-humus complexes). Between these aggregates, large and often air-filled pores are maintained that allow diffusion of air (oxygen) into the soil and the outward diffusion of CO₂. Compost amendment has also shown to change the pattern of emitted volatile organic compounds (VOC) (Seewald et al., 2010).

The greater the proportion of decomposed organic matter in its various stages protected from further decomposition, the larger is the water holding capacity of the soil, which counteracts the risk of desiccation, since 100 kg of humus are able to store 195 kg of water. The storage of water can counter the leaching of valuable minerals in drainage. If a soil is subject to chemical fertilizers over a long time, it loses its structural properties. Remedial tillage, fertilizers and irrigation must be used more. Maintaining a healthy soil structure is also a protection against erosion. Good soil structure also ensures optimal ventilation; oxygen promotes the uptake of potassium by plants and enables the synthesis of SO₂, CO₂ and the conversion of ammonia to nitrate.

Another positive influence of compost is its ability to neutralize toxins. Organic matter has a large capacity to bind heavy metals. Insoluble, stable complexes are formed that cannot be absorbed by the plant and are thus withdrawn from the food chain. Gómez-Brandón (2015), Kranz et al. (2020), and Ryals et al. (2014) provide the basis for an overview of the effects of compost (including vermicompost and compost teas) on plant growth and soil fertility, which may be summarized as follows:

- Enhanced nutrient retention
- Carbon sequestration
- Reduction of bulk density
- Enhanced water holding capacity, infiltration, hydraulic conductivity and plant available water
- Improved aeration (beneficial for both microorganisms and plant roots)
- Increased microbial diversity
- Increased plant disease suppression
- Enhanced resilience toward disturbances (high or low temperatures, drought)
- Decreased susceptibility to water and wind erosion
- Immobilization of cations, including heavy metals, rendering them unavailable to plants

The difference between inorganic and organic fertilizers on the one hand and compost on the other, is the dynamics of nutrient release. This may, depending on the composition of the fertilizers, be uncontrolled and fast, but it may also be retarded owing to specific formulations. The nutrient release dynamics of composts is mostly not as well defined, and depends a lot on the C:N:P-ratio and on the maturity of the compost. However, it accords quite well with plant demand because the nutrient release depends on the activities of microbes and microfauna, which are regulated by temperature and moisture availability, as is plant growth and nutrient uptake. The fertilizer effect of composts is often disputed, and indeed, it has been shown that a combination of slow-release compost with fast release fertilizer may be a way to combine the benefits of both (Ros et al., 2006). The high-nutrient component does not necessarily need to be mineral fertilizer, but could also be a fast-release organic fertilizer, such as manure. The choice of mineral or organic forms of nutrients could be based on whether the soil was dominated by bacterial or fungal biomass as the result of previous fertilizer strategies (Semenov et al., 2022). For nutrient balancing, agronomists assume that the nutrient release of compost is not greater than 20% of its content in the first, and even less in the subsequent years. Thus, the change of management practice from mineral to organic nutrient application, with the sole use of compost, may result in an initial nutrient deficit that is alleviated through repeated additions of compost in the following years.

Microbial diversity, macro and micropores, aggregates and structural stability

Biodiversity is an issue that is attracting increasing attention and so is diversity of soil microorganisms, a subject linked to resilience and resistance toward disturbances (Lynch and St. Clair, 2004). Abundant research has shown that the use of composts contributes to the microbial diversity in soils (Ros et al., 2006; Kurzemann et al., 2020). Microbial diversity benefits soils in two major ways: firstly, by safeguarding a steady nutrient supply through decomposition processes that cope with changing environmental conditions and, secondly, by supplying a microbiota that is disease suppressive. This may work through niche occupation and the generation of substances that are able to hamper the growth of pathogens (e.g., antibiotics produced by actinobacteria). An increased diversity of the microbiota is also the basis for a functioning food web, including protista and other eukaryotes, up to the earthworms. The activity of all these players contributes to the structural stability of soil, ranging from the formation of extracellular polymeric substances (EPS) by biofilm-producing bacteria, the structural network of actinobacterial and fungal mycelia to the secretion of mucilaginous substances by earthworms, strengthening the burrows. The co-occurrence of soil mineral and organic particles and EPS on one hand, and mycelia acting as a fiber-reinforcement on the other, makes resulting aggregates particularly strong.

In this context, the activity of soil fauna becomes important. Compost still contains digestible compounds, earthworms thrive and their burrows, strengthened by mucilaginous compounds, further contribute to the aeration and water conductivity of soils.

Climate related issues

The return of organic materials to soils is a natural way of closing nutrient and carbon cycles. The treatment process may, however, largely determine the climate-compatibility. The composting process releases CO₂, if managed properly. If aeration conditions are not maintained according to best practice, however, creation of considerable methane and nitrous oxide emissions can be a draw-back compared to anaerobic digestion (biomethanization) (Insam and Wett, 2008). Nevertheless, once applied to soil, several aspects contribute to the benefits of compost-based soil management regarding climate.

- Carbon sequestration. Under good process conditions, composting of organic wastes does generate CO₂ but not methane or nitrous oxide gas. The C remaining is in the form of some leftover plant material, (microbial) biomass and humic matter which is recalcitrant and may contribute to the buildup of long-lasting soil organic matter.
- The nutrients returned to soil substitute for chemical fertilizers that require a lot of energy for their production. The production of nitrogen fertilizer alone, based on the Haber-Bosch process consumes up to 2% of the global energy demand.
- Composts improve plant health, and in combination with organic agriculture, use of herbicides and pesticides may be averted.
- Soils amended with composts show an improved structural stability coupled with water retention, resulting in a reduced requirement for watering
- Improved water holding capacity may also help to reduce flooding that appears to become more frequent with climate change.

Conclusion

Nutrient and carbon cycling is a prerequisite for a sustainable economy and particularly concerning agriculture, horticulture and forestry. Bearing in mind the need for climate-friendly action and the need for maintaining healthy soils for the benefit of an increasing world population, composting is a way to get nutrient and carbon flows under control. Appropriately used, composts will render soils as sinks of greenhouse gases and make them less susceptible to erosion, increase their water holding capacity and microbial as well as faunal diversity and maintain their productivity on the long term. Moreover, composts may serve to convert waste lands to productive soils.

References

- Agegnehu G, Srivastava AK, and Bird MI (2017) The role of biochar and biochar-compost in improving soil quality and crop performance: A review. *Applied Soil Ecology* 119: 156–170. <https://doi.org/10.1016/j.apsoil.2017.06.008>.
- Amner W, McCarthy AJ, and Edwards C (1988) Quantitative assessment of factors affecting the recovery of indigenous and released thermophilic bacteria from compost. *Applied and Environmental Microbiology* 54(12): 3107–3112. <https://doi.org/10.1128/aem.54.12.3107-3112.1988>.
- Barragán-Fonseca KY, Nurfikari A, van de Zande EM, Wantulla M, van Loon JJA, de Boer W, and Dicke M (2022) Insect frass and exuviae to promote plant growth and health. *Trends in Plant Science* 27(7): 646–654. <https://doi.org/10.1016/j.tplants.2022.01.007>.
- Beesigamukama D, Subramanian S, and Tanga CM (2022) Nutrient quality and maturity status of frass fertilizer from nine edible insects. *Scientific Reports* 12(1): 7182. <https://doi.org/10.1038/s41598-022-11336-z>.
- Beffa T, Blanc M, and Aragno M (1996) Obligately and facultatively autotrophic, sulfur- and hydrogen-oxidizing thermophilic bacteria isolated from hot composts. *Archives of Microbiology* 165(1): 34–40. <https://doi.org/10.1007/s002030050293>.
- Bernal MP, Sommer SG, Chadwick D, Qing C, Guoxue L, and Michel FC (2017) Chapter three—Current approaches and future trends in compost quality criteria for agronomic, environmental, and human health benefits. In: Sparks DL (ed.) *Advances in Agronomy*. vol. 144, pp. 143–233. Academic Press. <https://doi.org/10.1016/bs.agron.2017.03.002>.
- Bonelli M, Bruno D, Caccia S, Sgambetterra G, Cappellozza S, Jucker C, Tettamanti G, and Casartelli M (2019) Structural and functional characterization of *Hermetia illucens* larval midgut. *Frontiers in Physiology* 10. <https://doi.org/10.3389/fphys.2019.00204>.
- Bougnom BP, Knapp BA, Elhottová D, Koubová A, Etoa FX, and Insam H (2010) Designer compost with biomass ashes for ameliorating acid tropical soils: Effects on the soil microbiota. *Applied Soil Ecology* 45(3): 319–324. <https://doi.org/10.1016/j.apsoil.2010.05.009>.
- Changa CM, Wang P, Watson ME, Hoitink HAJ, and Michel FC (2003) Assessment of the reliability of a commercial maturity test kit for composted manures. *Compost Science & Utilization* 11(2): 125–143. <https://doi.org/10.1080/1065657X.2003.10702119>.
- Danon M, Franke-Whittle IH, Insam H, Chen Y, and Hadar Y (2008) Molecular analysis of bacterial community succession during prolonged compost curing. *FEMS Microbiology Ecology* 65(1): 133–144. <https://doi.org/10.1111/j.1574-6941.2008.00506.x>.
- Domínguez J, Aira M, and Gómez-Brandón M (2010) Vermicomposting: Earthworms enhance the work of microbes. In: Insam H, Franke-Whittle I, and Goberna M (eds.) *Microbes at Work: From Wastes to Resources*, pp. 93–114. Springer. https://doi.org/10.1007/978-3-642-04043-6_5.
- Duong TTT, Verma SL, Penfold C, and Marschner P (2013) Nutrient release from composts into the surrounding soil. *Geoderma* 195–196: 42–47. <https://doi.org/10.1016/j.geoderma.2012.11.010>.
- Franke-Whittle IH, Juárez MF-D, Insam H, Schweizer S, Naef A, Topp A-R, Kelderer M, Rühmer T, Baab G, Henfrey J, and Manici LM (2019) Performance evaluation of locally available composts to reduce replant disease in apple orchards of central Europe. *Renewable Agriculture and Food Systems* 34(6): 543–557. <https://doi.org/10.1017/S1742170518000091>.
- Gärtling D and Schulz H (2022) Compilation of black soldier fly frass analyses. *Journal of Soil Science and Plant Nutrition* 22(1): 937–943. <https://doi.org/10.1007/s42729-021-00703-w>.
- Gold M, von Allmen F, Zurbrugg C, Zhang J, and Mathys A (2020) Identification of bacteria in two food waste black soldier fly larvae rearing residues. *Frontiers in Microbiology* 11. <https://doi.org/10.3389/fmicb.2020.582867>.
- Gómez-Brandón M (2015) Effects of compost and vermicompost teas as organic fertilizers. In: Sinha S and Pant K (eds.) *Fertilizer Technology. Synthesis*, 1st edn., pp. 301–318. Studium Press LLC.
- Gómez-Brandón M, Juárez MF-D, Zangerle M, and Insam H (2016) Effects of digestate on soil chemical and microbiological properties: A comparative study with compost and vermicompost. *Journal of Hazardous Materials* 302: 267–274. <https://doi.org/10.1016/j.jhazmat.2015.09.067>.

- Gómez-Brandón M, Aira M, and Domínguez J (2020) Vermicomposts are biologically different: Microbial and functional diversity of green vermicomposts. In: Bhat SA, Vig AP, Li F, and Ravindran B (eds.) *Earthworm Assisted Remediation of Effluents and Wastes*, pp. 150–170. Singapur: Springer Nature. https://doi.org/10.1007/978-981-15-4522-1_8.
- Insam H and De Bertoldi M (2007) Microbiology of the composting process. In: Golueke C, Bidlingmaier W, De Bertoldi M, and Diaz L (eds.) *Compost Science and Technology*, pp. 24–48. Elsevier.
- Insam H and Merschak P (1997) Nitrogen leaching from forest soil cores after amending organic recycling products and fertilizers. *Waste Management & Research* 15(3): 277–292. <https://doi.org/10.1006/wmre.1996.0084>.
- Insam H and Wett B (2008) Control of GHG emission at the microbial community level. *Waste Management* 28(4): 699–706. <https://doi.org/10.1016/j.wasman.2007.09.036>.
- Insam H, Riddech N, and Klammer S (2002) *Microbiology of Composting (2002)*. Springer.
- Itävaara M, Venelampi O, Vikman M, and Kapanen A (2002) Compost maturity—Problems associated with testing. In: Insam H, Riddech N, and Klammer S (eds.) *Microbiology of Composting*, pp. 373–382. Springer. https://doi.org/10.1007/978-3-662-08724-4_31.
- Klammsteiner T, Turan V, Juárez MF-D, Oberegger S, and Insam H (2020) Suitability of black soldier fly frass as soil amendment and implication for organic waste hygienization. *Agronomy* 10(10): 1578. <https://doi.org/10.3390/agronomy10101578>.
- Kranz CN, McLaughlin RA, Johnson A, Miller G, and Heitman JL (2020) The effects of compost incorporation on soil physical properties in urban soils—A concise review. *Journal of Environmental Management* 261: 110209. <https://doi.org/10.1016/j.jenvman.2020.110209>.
- Kuba T, Tschöll A, Partl C, Meyer K, and Insam H (2008) Wood ash admixture to organic wastes improves compost and its performance. *Agriculture, Ecosystems & Environment* 127(1): 43–49. <https://doi.org/10.1016/j.agee.2008.02.012>.
- Kurzemann FR, Plieger U, Probst M, Spiegel H, Sandén T, Ros M, and Insam H (2020) Long-term fertilization affects soil microbiota, improves yield and benefits soil. *Agronomy* 10(11): 1664. <https://doi.org/10.3390/agronomy10111664>.
- Lehmann J and Kleber M (2015) The contentious nature of soil organic matter. *Nature* 528: 60–68. <https://doi.org/10.1038/nature16069>.
- Li J, Mavrodi DV, and Dong Y (2021) Effect of rock dust-amended compost on the soil properties, soil microbial activity, and fruit production in an apple orchard from the Jiangsu province of China. *Archives of Agronomy and Soil Science* 67(10): 1313–1326. <https://doi.org/10.1080/03650340.2020.1795136>.
- Liu T, Klammsteiner T, Dregulo AM, Kumar V, Zhou Y, Zhang Z, and Awasthi MK (2022) Black soldier fly larvae for organic manure recycling and its potential for a circular bioeconomy: A review. *Science of the Total Environment* 833: 155122. <https://doi.org/10.1016/j.scitotenv.2022.155122>.
- Lynch JP and St. Clair SB (2004) Mineral stress: The missing link in understanding how global climate change will affect plants in real world soils. *Field Crops Research* 90(1): 101–115. <https://doi.org/10.1016/j.fcr.2004.07.008>.
- Millner PD, Powers KE, Enkiri NK, and Burge WD (1987) Microbially mediated growth suppression and death of *Salmonella* in composted sewage sludge. *Microbial Ecology* 14: 255–265.
- Ros M, Klammer S, Knapp B, Aichberger K, and Insam H (2006) Long-term effects of compost amendment of soil on functional and structural diversity and microbial activity. *Soil Use and Management* 22(2): 209–218. <https://doi.org/10.1111/j.1475-2743.2006.00027.x>.
- Ryals R, Kaiser M, Torn MS, Berhe AA, and Silver WL (2014) Impacts of organic matter amendments on carbon and nitrogen dynamics in grassland soils. *Soil Biology and Biochemistry* 68: 52–61. <https://doi.org/10.1016/j.soilbio.2013.09.011>.
- Ryckeboer J, Mergaert J, Vaes K, Klammer S, De Clercq D, Coosemans J, Insam H, and Swings J (2003) A survey of bacteria and fungi occurring during composting and self-heating processes. *Annals of Microbiology* 53: 349–410.
- Seewald MSA, Singer W, Knapp BA, Franke-Whittle IH, Hansel A, and Insam H (2010) Substrate-induced volatile organic compound emissions from compost-amended soils. *Biology and Fertility of Soils* 46(4): 371–382. <https://doi.org/10.1007/s00374-010-0445-0>.
- Semenov MV, Krasnov GS, Semenov VM, and van Bruggen A (2022) Mineral and Organic Fertilizers Distinctly Affect Fungal Communities in the Crop Rhizosphere. *Journal of Fungi* 8(3): 3. <https://doi.org/10.3390/jof8030251>.
- Song S, Ee AWL, Tan JKN, Cheong JC, Chiam Z, Arora S, Lam WN, and Tan HTW (2021) Upcycling food waste using black soldier fly larvae: Effects of further composting on frass quality, fertilising effect and its global warming potential. *Journal of Cleaner Production* 288: 125664. <https://doi.org/10.1016/j.jclepro.2020.125664>.
- Vickerson A, Radley R, Marchant B, Kaulfuss O, and Kabaluk T (2015) *Hermetia illucens* Frass Production and Use in Plant Nutrition and Pest Management (World Intellectual Property Organization Patent No. W02015013826A1). <https://patents.google.com/patent/W02015013826A1/en>.
- Xiao X, Mazza L, Yu Y, Cai M, Zheng L, Tomberlin JK, Yu J, van Huis A, Yu Z, Fasulo S, and Zhang J (2018) Efficient co-conversion process of chicken manure into protein feed and organic fertilizer by *Hermetia illucens* L. (Diptera: Stratiomyidae) larvae and functional bacteria. *Journal of Environmental Management* 217: 668–676. <https://doi.org/10.1016/j.jenvman.2018.03.122>.

Landfarming and its microbiology

Naga Raju Maddela^{a,b}, Balasubramanian Ramakrishnan^c, Kadiyala Venkateswarlu^d, and Mallavarapu Megharaj^e, ^aDepartamento de Ciencias Biológicas, Facultad de Ciencias de la Salud, Universidad Técnica de Manabí, Portoviejo, Ecuador; ^bInstituto de Investigación, Universidad Técnica de Manabí, Portoviejo, Ecuador; ^cDivision of Microbiology, ICAR-Indian Agricultural Research Institute, New Delhi, India; ^dFormerly Department of Microbiology, Sri Krishnadevaraya University, Anantapuramu, India; ^eGlobal Centre for Environmental Remediation (GCER), and Cooperative Research Centre for Contamination Assessment and Remediation of the Environment (CRC CARE), College of Engineering, Science and Environment, The University of Newcastle, Callaghan, NSW, Australia

© 2023 Elsevier Ltd. All rights reserved.

Introduction	533
Technological importance of landfarming	535
Properties of soil and chemical contaminants and environmental factors	535
Soil properties	535
Contaminant chemical properties	536
Concentrations of contaminants	536
Climatic factors	536
Biostimulation	537
Bioaugmentation	538
Microbiological features of landfarming	538
Bacteriological aspects of landfarming sites	538
Role of fungi in landfarming sites	539
Molecular microbial diversity and abundance	541
Evaluation of landfarming efficacy	542
Summary	542
References	545

Abstract

Landfarming, an aerobic, above-ground, engineered ex-situ bioremediation process, has been used globally since the 1970s for the remediation and restoration of sites impacted by diverse pollutants, principally petroleum hydrocarbons. One of the key actors in removing contaminants during landfarming is the microflora, mainly bacteria and fungi. However, despite the poor understanding of its microbial metabolic features, landfarming technology has become widespread. The present chapter provides an overview of the microbial aspects related to the diversity and composition, biostimulation and bioaugmentation for improving effectiveness, outcomes of recent field-scale studies, efficacy evaluation, and future research directions for the technological advances in landfarming.

Key points

- Bacteria and fungi are the most active degraders of organic contaminants in landfarming
- Bacteria exhibit greater diversity and abundance than fungi at landfarming sites
- Biostimulation and bioaugmentation are the essential landfarm management practices
- Insights into microbial metabolic networks can improve landfarming efficacy
- Continuing research effort is required to ensure the use of landfarming technology for diverse contaminants under extreme environmental conditions

Introduction

'Landfarming' involves natural processes for degrading organic contaminants in excavated soil or sediment *ex-situ*. The contaminated soil or sediment is collected and taken to the land farm, spread, and generally treated on top of the upper soil zone or in a specially designed land-treatment cell. This remediation technology is also referred to as land treatment or land application. The activities of landfarming technology include (i) tilling by agricultural equipment that is easily available, such as chisel plows, rotary tillers, harrows pulled by tractors and (ii) application of mineral or organic fertilizers, pH regulators, and bulking agents that improve structure, texture, aeration, nutrient status and microbial activities of the contaminated soils or sediments. This engineered ex-situ process includes spreading the excavated contaminated soils or sediments as a thin layer above-ground, usually <0.30 m deep, to facilitate tillage operations on a prepared soil surface. These treatment plots are also called 'landfarms.' Landfarms have had

significant success rates in treating the contamination of petroleum hydrocarbon wastes in soils and sediments, drill cuttings, low brine drilling fluids, oily sludges, tank sediments, and pit sludges. These complex mixtures of organic compounds may contain diverse alkanes, aromatics, phenols, paraffins, asphaltenes, and polynuclear aromatic hydrocarbons. Whereas the simple aliphatic and aromatic compounds are easily degradable, other compounds are recalcitrant to microbial degradation. This remediation technology has been used in industrial practice since the 1970s. About one-third of all US refineries in 1983 and nearly 60% of remediation in former military bases in South Korea have included landfarming (API, 2003; Kim et al., 2022). Wider adoption implies recognition of the importance and strengths of physical, chemical, and biological aspects of landfarming technology to degrade the contaminants and restore the contaminated sites. Effective landfarming management depends on the number and diversity of naturally occurring bacterial and fungal populations, which are the critical drivers of biodegradation of contaminants and that can be augmented or stimulated for optimal biotransformation and degradation. Microbial activities that are established during landfarming can degrade specific contaminants in two differentiated phases: the first phase involves the rapid degradation of labile fractions, and the second phase with slower degradation rates due to reduced bioavailability and recalcitrance. Those pollutant hydrocarbons of greater molecular weight, such as polycyclic aromatic hydrocarbons (PAHs), have slower biodegradation rates. But, highly volatile contaminants, such as light petroleum hydrocarbons, are unsuitable because of their high potential for volatilization.

Landfarming under aerobic conditions do not favor the remediation of soils or sediments contaminated with chlorinated and nitrate-containing compounds. The chemical structure, the concentration of the contaminants determines the effectiveness of the degradation processes. At landfarming sites, a biphasic degradation kinetics of contaminants occurs with gradual decreases in microbial respiration rates. The microbial biomass carbon (C), commonly written as C_{mic} , and community diversity indices, such as Shannon H, can decrease along with the microbial respiration rates. Even at small concentrations, contaminants (e.g., total petroleum hydrocarbons (TPHs) of 0.30%) can be toxic to the microbial community. Hence, repeated applications of contaminated soils can may become toxic, rather than acting as substrates for the microbial communities. Other factors that influence the effectiveness of landfarming include soil type and properties, climate and weather conditions at the treatment sites, and the cost involved and volume of the contaminated soil or sediment to be excavated and treated. In landfarms, aerobic microbial activity is stimulated by frequent turnover of the contaminated soil (or sediment) layers by tilling (or plowing) or adding nutrients or moisture. The commonly available tilling equipment limits the depth of contaminated soil to 0.3–0.6 m. Thus, optimization of the landfarming process depends on the soil moisture content, the frequency of aerating practices, and soil pH. Generally, no sophisticated equipment is required, and the standard equipment used for traditional agricultural operations is sufficient to treat the contaminated soil. At the landfarming treatment sites, a protective membrane layer above the natural soil surface of clay or composite liner is preferred when there is a risk of underground water contamination. Using an impermeable membrane such as high-density polyethylene (HDPE) membrane (of 1–10 mm thickness) can prevent groundwater contamination by controlling the leachate movement during landfarming.

Landfarming is thus a low-cost, energy-efficient, and eco-friendly remediation technology for biodegradable contaminants with minimal residue disposal problems. For these reasons, landfarming gained more public acceptance than incineration, landfilling, or deep-well injection. By recycling organic carbon and nutrients from the contaminants within the biosphere, landfarming plays a significant role in maintaining and sustaining soils or sediments critical to supporting healthy ecosystems, including through revegetation and reforestation (Sims and Sims, 2003). Rigorous studies on the bioavailability of pollutants, microbial community dynamics, and the formation of soil-bound residues informing risk assessment and management, are still required to provide valuable insights into understanding chemical-soil-microbe interactions for the risk assessment and management of landfarming technology. The basic scheme of landfarming technology is depicted in Fig. 1.

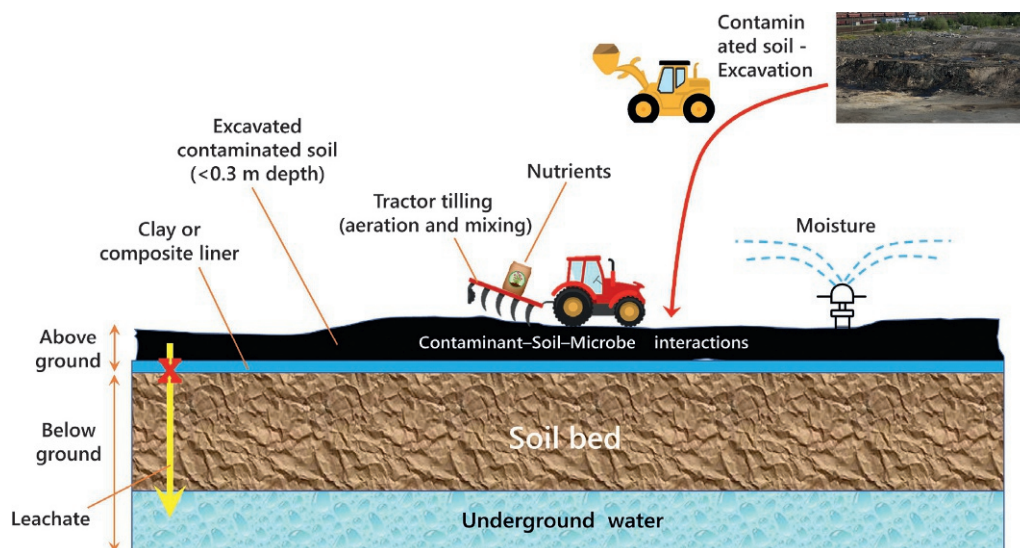


Fig. 1 A typical set-up of the landfarm technology for contaminated soils or sediments.

Technological importance of landfarming

Landfarming is a cost-effective biological treatment aimed at degrading contaminants by converting them to carbon dioxide and water. Variations in the physical, chemical, and biological parameters of this 'low tech' technology may contribute to or hamper the remediation process. Physicochemical processes such as volatilization, photooxidation, and humification can significantly decrease the contaminant concentration during the initial period. For example, gasoline, a volatile and lighter petroleum product, is removed by volatilization during the aeration process, particularly by tine cultivator or plowing (more intensive tilling) operations at landfarming sites. But the emissions of volatile organic compounds can accentuate air pollution at the landfarming sites. Enclosures such as a greenhouse or a plastic sheet cover will minimize the dissipation of constituent vapors emitted. Capturing and treating the vapors appropriately before venting is another control measure for contaminants with higher volatile constituents. Photooxidation can involve direct photolysis by sunlight and indirect photoreactions by the light-absorbing chromophores. This physicochemical process can oxidize and form new, less-stable reactive intermediate products. However, reactive intermediate products may be more toxic than the parent compounds. In addition, intermediate products are either water-soluble or insoluble polymers. The photo-produced insoluble polymers can form stable and persistent emulsions, making them challenging to degrade during clean-up operations. Humification is a process interconvertible with degradation. Humic substances with $-OH$, $-COOH$, $-CH_3$, $-CH_2$, and $C-H$ groups and aromatic substances can bind organic pollutants; these substances also shuttle electrons for the conversion of organic pollutants in the redox reactions. During the initial period, adsorption of contaminants on soil particles is essential for the dissipation of contaminants. Several mechanisms such as ion exchange, hydrogen bonding, charge transfer, London-van der Waals dispersion forces, and interaction with metallic cations, influence adsorption behavior, especially the bioavailability of contaminants. The significant biological aspect of landfarming is degradation of contaminants mediated by biota. Indigenous microorganisms are critical to biodegradation of contaminants. Both biodegradation and bioavailability determine the cost-effectiveness of landfarming. In some cases, the extent of biodegradation is inversely proportional to increasing contaminant concentrations.

For effectively treating contaminants by landfarming, soil may be used as an inoculum and as a supporting medium for biological growth under aerobic conditions. Factors that can be optimized to maximize remediation are moisture content, oxygen level, nutrients, pH, and soil bulking. Four significant steps in establishing landfarms for treating contaminated soils or sediments, are (i) application of mixing techniques for effective contact ($\sim 90\%$) between soil particles and additives; (ii) addition of inorganic nutrients in the proportion of 100:10:1 hydrocarbon: nitrogen: phosphorous; (iii) mixing the soils with sand, sawdust or wood chips to increase the porosity of the soils; and (iv) adjustment of soil pH by using lime, alum or phosphoric acid. With all these implementations, the typical land-farming cost can be US\$ 50–200 per cubic meter (Cookson, 1995). Nevertheless, the main advantages of landfarming technology are that it is a cost-effective technology, well suited to remote regions, does not require skilled persons, provides optimum conditions for microbial degradation, and has short treatment times (6 months to 2 years maximum) under optimal conditions. On the other hand, landfarming technology has its drawbacks, as it requires ample space, might lead to uncontrolled biodegradation conditions which increases the length of the operation time, is unsuitable for inorganic contaminants, can cause heavy metal toxicity to microorganisms, releases large amounts of particulate matter, may produce leachates, which at vulnerable sites can contaminate groundwater. Above all, the technology is inappropriate for the remediation of sites contaminated with halogenated organic contaminants, creosote, pentachlorophenol (PCP), and bunker C oil, all of which are not degraded by landfarming.

Properties of soil and chemical contaminants and environmental factors

The properties of contaminated soils, the nature of chemical contaminants, and the environmental factors prevalent at the treatment locations significantly affect the microbial activities with consequences for the degradation rates. Successful operation of the landfarming technology necessitates the consideration of those factors related to the soil, contaminants, and the environment (EPA, 2017):

Soil properties

A neutral pH range is optimal (i.e., 6–8) for successful landfarming. The optimal moisture content should be in the range of 40–85% of the water-holding capacity of the soil (field capacity) or 12–30% moisture by weight. Excess moisture should be avoided. Notably, tillage that causes evaporation of soil moisture and high precipitation or poor drainage that results in the excess accumulation of moisture have to be monitored regularly and managed to be in the optimal range for landfarming. Within the 10 °C to 40 °C temperature range, microbial activity doubles for every 10 °C rise; this implies that soil temperature can significantly impact landfarming efficiency. If the soil temperature is cooler than 10 °C, microbial activity substantially decreases; if the temperature goes below 5 °C, the activities of microbes related to contaminant degradation become slow and inefficient. Therefore, the period, when the soil temperature is within the optimal range, is called the 'landfarming season.' In colder regions, psychrophilic bacteria are introduced into the landfarms; these cold-adapted bacteria withstand temperatures cooler than 10 °C and help achieve the goals of landfarming.

The soil texture significantly influences aeration and moisture contents. For example, clay soils compact easily, leading to poor aeration, and have high water holding capacity, retaining water for extended periods following precipitation. Thus, poor aeration and excess moisture content in clay soils may adversely affect the microbial activity at the landfarming sites. Nitrogen and

phosphorus nutrients are added to support microbial cell growth and degradation as these nutrients are usually at insufficient concentrations at the contaminated sites. But, excess nutrients, such as phosphate and sulfate, can repress microbial activities. The concentration ratios of carbon:nitrogen:phosphorus (C:N:P) in contaminated soils or sediments that range from 100:10:1 to 100:1:0.5 are considered favorable for microbial growth. The ratios are decided based on the specific constituents and microbial species that are to be involved in the degradation process. The nutrients to be added are calculated as follows:

Example: The volume of contaminated soils is 90,000 m³, the average pollutant concentration in the contaminated soil is 1000 mg kg⁻¹, and the soil bulk density is 50 kg m⁻³.

Therefore,

$$\begin{aligned}\text{Soil mass} &= \text{Volume of contaminated soil} \times \text{Soil bulk density} \\ &= 90000 \text{ m}^3 \times (50 \text{ kg} \div \text{m}^3) \\ &= 4.5 \times 10^6 \text{ kg}\end{aligned}$$

$$\begin{aligned}\text{Contaminant mass} &= \text{Soil mass} \times \text{Pollutant concentration} \\ &= 4.5 \times 10^6 \times 1,000 \text{ mg kg}^{-1} \\ &= 4.5 \times 10^3 \text{ kg} = 4,500 \text{ kg} (\sim 10,000 \text{ lbs})\end{aligned}$$

For a C:N:P ratio of 100:10:1, 455 kg (or 1000 lbs) nitrogen and 45 kg (or 100 lbs) phosphorus would be required to treat contaminated sites using the landfarming technology.

Contaminant chemical properties

The volatility of chemical contaminants is an essential determinant for microbial degradation. Generally, petroleum hydrocarbon fractions have different proportions of volatile components, which evaporate into the atmosphere during tillage; these compounds are unavailable to microbial degradation. Biodegradation rates depend on the chemical structure of contaminants. The potential microbial degradability of different petroleum hydrocarbons during landfarming is provided in Fig. 2. n-Butane and n-pentane fractions of gasoline are easily degraded by microorganisms but naphthalene, fluoranthene, pyrenes, and acenaphthenes are less degradable.

Concentrations of contaminants

The concentration of contaminants at the sites is also critical to microbial degradation. For example, large concentrations of petroleum hydrocarbons (10,000–50,000 ppm) or heavy metals (>2500 ppm) are toxic and inhibitory to microbial growth. Soils from very contaminated sites may be diluted by adding fresh soil. On the other hand, very low levels of contaminants lead to diminished microbial activity.

Climatic factors

As temperature and moisture are critical to landfarming, the local climatic conditions are crucial. Rainfall determines the success rates of this remediation technology. During and after a significant precipitation episode, the soil moisture content will exceed the requirement for optimum bacterial activity. Drought results in soil moisture becoming below optimum. Both ample precipitation and drought can inhibit microbial activity, thereby extending the period required for successful landfarming.

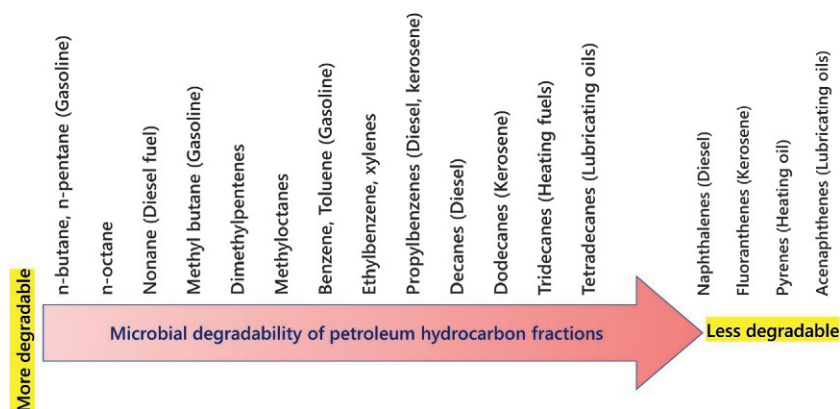


Fig. 2 Microbial degradability of different petroleum hydrocarbon fractions during landfarming.

Biostimulation

Though microbial activity can increase initially in the contaminated soils without any amendment, a rapid reduction in activity occurs due to a lack of nutrients. The stimulation of the activities of indigenous (autochthonous) contaminant-degrading microorganisms by adding nutrients or substrates to the contaminated soils or sediments is referred to as biostimulation. The adjustment of soil pH, the addition of nutrients (N, P, trace elements), and ensuring optimum moisture content and oxygen concentration are the main levers for the biostimulation of native microflora in contaminated soils or sediments. Biostimulation has been successfully used in landfarming technology to remove and restore PAHs-contaminated soils. Agricultural wastes, such as corn cobs, can be used to overcome nutrient limitations in the landfarming sites. In addition to providing nutrients, using agricultural or industrial wastes or composts can improve the porosity and aeration. Nearly 89% of PAHs were removed in the landfarms where ground rice hulls were added as a bulking agent and dried blood as a slow-release nitrogen source (Nduka et al., 2012). More importantly, the microbial population densities could triple under biostimulation, providing about 90% removal efficiency against <50% in the unamended soils.

Perhaps surprisingly, there is no significant difference between synthetic fertilizer (NPK and urea) and natural organic amendments (e.g., cattle or poultry manure, animal wastes) in stimulating the microbial activity during remediation of soils contaminated with petroleum hydrocarbons. Nevertheless, natural organic fertilizers are preferable as biostimulating agents due to their efficacy and low price. Landfarming of hydrocarbon-contaminated soil in the Niger-Delta (Nigeria) with a combination of nutrients (NPK), biochar and rhamnolipids resulted in the removal of 23% more TPHs than nutrients alone (Brown et al., 2017). Additionally, excess fertilizer application may have a negative impact on nutrient cycling in the landfarms. For example, urea applied at a rate of 893 g N m^{-2} inhibited nitrification at a landfarm site containing 4–6% (w/w) of mineral oil (Peltola et al., 2006). Also, addition of 889 g N m^{-2} methylene urea or 893 g N m^{-2} urea enhanced the removal of mineral oil from the contaminated soil. In another investigation of a 9-month plot-scale landfarm, the nutrient addition (N-P-K + surfactant) stimulated the maximum dehydrogenase activity at the end of 2nd, 4th, and 7th month in soil contaminated with diesel oil (10, 000 and 20,000 mg kg^{-1}) (Fig. 3, Silva-Castro et al., 2015). In a recent report, Johnsen et al. (2021) demonstrated that oil degradation was barely measurable in the absence of added mineral nutrients but greatly increased after their addition, confirming the importance of biostimulation process in activating the local microflora.

Oil spills from meteorological stations, military outposts and research stations occur in the Arctic. Landfarming is an appropriate technology for remediating contaminated soils as the hydrocarbon-degrading microorganisms are commonly prevalent in this region. In a full-scale Arctic landfarm, Johnsen et al. (2021) provided 1636 kg ha^{-1} fast-release fertilizer (total N 15.5%, CaO 26.3%), and 6523 kg ha^{-1} of the NPK fertilizer (total N 12%, total P 4%, K 13%, S 9%, Fe 0.06%, B 0.02%, Zn 0.01%, 3,4-dimethyl pyrazole phosphate as nitrification inhibitor 0.8%) to reach an initial N concentration $\sim 300 \text{ mg kg}^{-1}$ (15% water, wet wt) as amendments while remediating diesel-polluted soils. Finally, there was plowing (30 cm) for aeration and mixing of the soil, and sprinklers were used to release fertilizer from the pellets. This biostimulation method achieved the maximum population of diesel degraders ($1.2 \times 10^9 \text{ cells g}^{-1}$) in that arctic landfarm. Even when the metabolic diversity of the microbial degrader communities is adequate in the landfarms, biostimulation can ensure efficient degradation within a reasonable time frame.

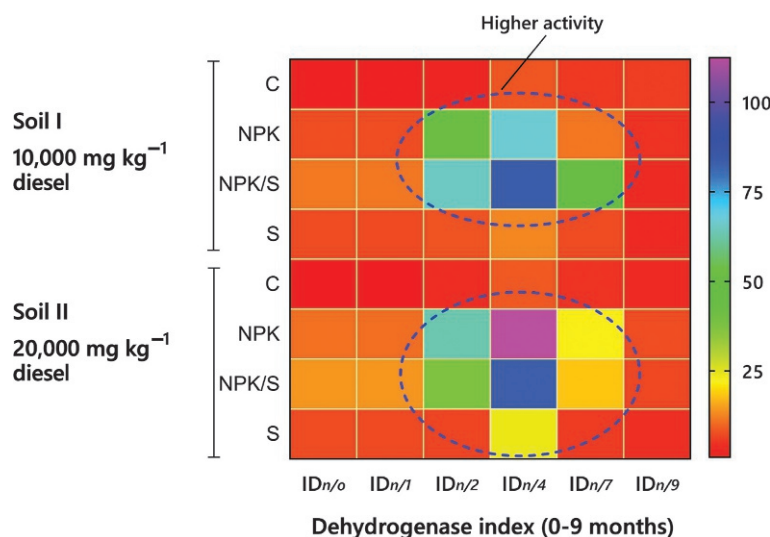


Fig. 3 Dehydrogenase indices in the diesel-polluted soil subjected to a 9-month biostimulation-assisted landfarm site. (C, Natural attenuation; S, Surfactant; NPK (Nitrogen:phosphorous:potassium) at 31:5:7; NPK/S, NPK + surfactant). Based on data from Silva-Castro GA, Uad I, Rodríguez-Calvo A, González-López J, Calvo C (2015) Response of autochthonous microbiota of diesel polluted soils to land-farming treatments. *Environmental Research* 137: 49–58.

Bioaugmentation

Bioaugmentation implies the addition of non-pathogenic contaminant degrading microorganisms to the soil. For successful bioaugmentation, selecting a consortium of autochthonous microflora is preferable for better adaptability to their new environment and increased degradation rates. The microorganisms are usually isolated from soil of the same type as that being remediated to remove any toxicity problem. The bioaugmentation-assisted landfarming shows about 16% and 29% greater efficiencies for removing TPHs than standard landfarming or natural attenuation, respectively (Guarino et al., 2017). In an ex-situ field-scale landfarm, the addition of a commercially available consortium at 1:50 inoculum-soil ratio resulted in the removal of 85% of TPHs after 155 days, with the TPH decay rate of 0.015 day^{-1} against the rate of 0.0069 day^{-1} under the intrinsic conditions (Chien et al., 2010). The survival of the added strains at the contaminated site is one of the significant challenges in bioaugmentation approaches. The competition from indigenous microorganisms and contaminant toxicity are two essential factors that affect the survival and activities of added strains.

Bioaugmentation is often used in combination with biostimulation to achieve better results in the restoration of contaminated sites by landfarming. The application of a consortium of selected bacterial species, combined with inorganic and oleophilic fertilizers can ensure even the efficient removal of PAHs from contaminated soils by landfarming. For example, whereas the biostimulation-based landfarm microcosms had about 34% and 57% removal rates for PAHs, and total BaP, respectively, these values increased to 87% and 67% when biostimulation and bioaugmentation were combined (Straube et al., 2003). The integration of landfarming with the composting of organic wastes can improve the bioremediation performance. In this combined strategy, composting organic waste promotes biostimulation and bioaugmentation, creating favorable conditions for biological activity. Such integrated approaches are economical and effective in removing organic pollutants from contaminated sites.

Microbiological features of landfarming

Large numbers of diverse microorganisms such as bacteria, algae, fungi, and many others, including protists, exist in the soils. However, the most numerous and biochemically active microorganisms in landfarms are bacteria. The activities of specific bacteria and fungi at the contaminated sites for degradation have been identified. For example, the bacteria *Pseudomonas aeruginosa*, *Acinetobacter* spp., *Myroides* spp., and *Bacillus* spp. and fungi belonging to the class *Eurotiomycetes*, *Saccharomycetes*, and *Sordariomycetes* are essential in the degradation of PAHs at spill sites. Many strains of *Alcaligenes* sp., *Enterobacter* sp., and *Ochrobactrum* sp. can degrade several classes of petroleum hydrocarbons.

Bacteriological aspects of landfarming sites

Bacteria are the key microbial degraders of contaminants at landfarming sites. Because of their genetic and metabolic diversity, their ubiquity in terrestrial and aquatic systems, they are naturally abundant at the contaminated sites, Bacteria are responsible for either degradation or mineralization of contaminants at landfarming sites. They act either as individual species or as communities. The preference for different classes of chemical compounds for degradation varies among diverse bacteria. The members of the genera such as *Sphingomonas*, *Sphingobium*, *Novosphingobium*, and *Sphingopyxis* belonging to *Sphingomonadaceae* are efficient degraders of aromatic compounds. For degrading aromatic or aliphatic hydrocarbons, the members of *Xanthobacteraceae* (*Alphaproteobacteria*) and *Pseudomonadaceae* (*Gammaproteobacteria*) have the necessary genes. A species of *Kaistobacter*, belonging to the family *Sphingomonadaceae*, was the most abundant (about 10–60% of the total 16S rRNA sequences reads) in control and the minimally managed oil sludge containing microcosms (Dörr de Quadros et al., 2016). Isolates of this genus are widely distributed in methane-enriched sites, diesel-contaminated soils, and isoprene-contaminated soils. Silva-Castro et al. (2012) identified the changes in the population in terms of the total number of cultural aerobic heterotrophic bacteria over 7 months of landfarming (six different treatments) used for restoring the hydrocarbon-contaminated soil (Fig. 4). Of the six treatments, the one containing a consortium of *Bacillus pumilus*, *Alcaligenes faecalis*, *Micrococcus luteus*, and *Enterobacter* sp. (i.e., treatment D) showed the largest number of culturable heterotrophic aerobic bacteria. In addition, it also removed 90% of PAHs, suggesting the importance of these bacterial species in restoring the contaminated sites by landfarming technology. The bacterial consortium A associated with treatment D was 2.6 times more effective in removing the PAHs than consortium B (associated with treatment E, which only achieved a PAH removal of 34%). However, the provision of NPK to consortium B led to a stimulation of PAH removal efficiency (treatment F, that removed 87% of PAHs). In another landfarming experiment, carried out over 90 days, addition of a bacterial consortium (22 isolates from the hydrocarbon-contaminated site, belonging to the genera *Acetobacter*, *Achromobacter*, *Comamonas*, *Delftia*, *Ochrobactrum*, *Pseudomonas*, *Pseudoxanthomonas*, *Sphingobium*, and *Stenotrophomonas*) improved the efficiency of TPHs removal by 12–80% in 7 out of 8 the landfarming treatments over the control which did not receive the bacterial consortium (Guarino et al., 2017).

The evolutionary relationship of isolated microbial species needs to be compared with the known microorganism degrading specific pollutants to explore the metabolic potentials of microorganisms that can be used in bioremediation. Very recently, two bacterial isolates (*Kocuria flava* DTU-1Y and *Rhodococcus pyridinivorans* DTU-7P) from contaminated soil and have been

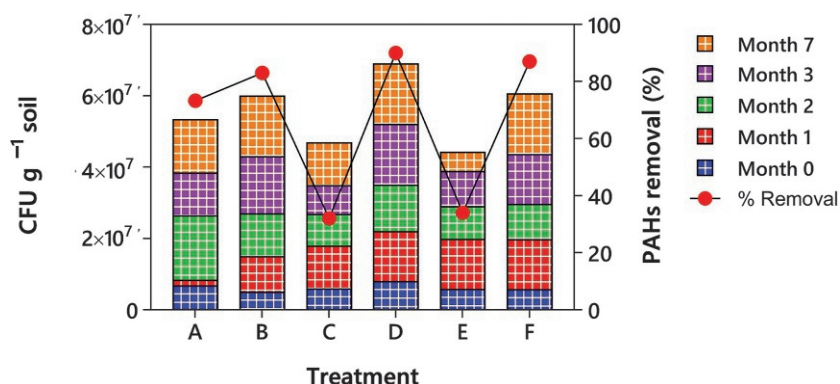


Fig. 4 Dynamic changes in the total number of culturable aerobic heterotrophic bacteria and PAHs removal (%) in hydrocarbon-contaminated soil subjected to different landfarming treatments (A, Control; B, amended with NPK fertilizer; C, amended with S200; D, amended with consortium A; E, amended with consortium B; F, amended with NPK fertilizer and consortium A. Consortium A was a mixture of *Bacillus pumilus* 28-11, *Alcaligenes faecalis* 212-2, *Micrococcus luteus* 212-4, and *Enterobacter* sp. 214-6; Consortium B was a mixture of *B. pumilus* 28-11, *Enterobacter* sp. 214-6, and *Ochrobactrum anthropi* AD2). Based on data from Silva-Castro GA, Uad I, González-López J, et al. (2012) Application of selected microbial consortia combined with inorganic and oleophilic fertilizers to recuperate oil-polluted soil using landfarming technology. *Clean Technologies and Environmental Policy* 14: 719–726.

phylogenetically compared with the known PAH-degrading bacteria having PAH-catabolic genes or enzymes and confirmed that strains DTU-1Y and DTU-7P had the potential to utilize PAHs as their sole carbon source for their growth (Sakshi et al., 2020). The microbial activities in the landfarms can be improved by 'surfactant flushing.' Hydrocarbon compounds can easily be removed from the soil through micelle formation and desorption mechanisms using surfactants as a flushing solution. Non-ionic surfactants are widely used in the remediation of soils impacted by hydrocarbon compounds due to their ability to solubilize hydrocarbons. Anaerobic biodegradation is one of the main tasks in cold anoxic regions. The search for novel microbial species that have the potential for oil degradation under anaerobic conditions is of paramount importance. The application of advanced molecular techniques [e.g., metagenomic-Stable Isotope Probing (SIP)] is important in the search for functional microbial species in pristine low-temperature conditions, especially in contaminated sites of Antarctic and Arctic environments. The use of extremozymes and extremophilic organisms is important in the degradation of various pollutants. There is also growing interest in 'nanozymes' to restore contaminated sites (Porto de Souza et al., 2022). Insights into bacterial degraders and their metabolic networks are critical to improving the efficiency of landfarming technology and restoring contaminated sites.

Role of fungi in landfarming sites

Fungi play a considerable role in the degradation of contaminants at the treatment sites with landfarming technology. Since fungi are metabolically diverse, they can degrade a wide range of contaminants such as textile dyes, leather tanning chemicals, petroleum hydrocarbons, PAHs, PPCPs (pharmaceuticals and personal care products), and agrochemicals (such as pesticides and herbicides). Fungal strains have been proven to be an economical, efficient, and environmentally benign agent to restore contaminant sites. The fungal isolate of *Penicillium* sp. 06 from a soil used for landfarming of oily waste degraded about 89% of phenanthrene and more than 75% of acenaphthene, fluorene, and fluoranthene after 30 days of incubation (Zheng and Obbard, 2003). The propensity of *Penicillium* sp. 06 isolate to oxidize high molecular weight PAHs (i.e., chrysene, benzo(a)pyrene, and dibenz(ah)anthracene) was inversely proportional to the molecular weight of the PAHs. *Fusarium oxysporum*, which was isolated from a petrochemical sludge landfarming site, was able to utilize a wide range of organics as its carbon sources, such as pyrene, anthracene, phenanthrene, naphthalene, catechol, gentisic acid, dihydroxybenzoic acid, benzene, ethylbenzene, toluene, xylene, 1-decene, 1-octene, hexane, ethanol, gasoline, and diesel oil (Jacques et al., 2007). The landfarming site for treating oil waste had significant population densities of fungi (total fungi, TF about 2.60×10^5 CFU g⁻¹); fungi are an essential group among microflora associated with the degradation of contaminants, other than heterotrophic aerobic bacteria (HAB) and heterotrophic anaerobic bacteria (HNB) (da Silva et al., 2013). These three microbial groups (TF, HAB, and HNB) were combinedly responsible for reducing TPH content by ~90% at a landfarming site. *F. oxysporum*, isolated from the landfarming site contaminated with PAHs, was one of the potential microorganisms in a mixed consortium of bacteria and fungi prepared for bioaugmentation (Jacques et al., 2008). The mixed consortium was found to degrade anthracene, phenanthrene, and pyrene in the soil over 70 days at removal rates of 99%, 99%, and 96%, respectively. Baron et al. (2021) isolated black fungi (or melanized fungi) from the landfarming site in the oil industry, capable of utilizing toluene as the sole carbon source. These strains were identified as *Cladophialophora minourae*, and *C. immunda*. Tables 1 and 2 provide details of two recent full-scale landfarming studies, one equatorial in Nigeria, the other in Greenland and illustrate the scope and understanding of the latest developments in the microbial features of landfarming technology. At the landfarming site in Niger Delta, 128 bacterial species and 52 fungal species were identified, and 92% of TPHs were removed in 56 days (Table 1). At the arctic landfarming site in Greenland (−10 °C), 64–88% of TPHs degradation was achieved, and the predominant microorganisms were the degraders of undecane, *m*-xylene, 2-methylnaphthalene and phenanthrene (Table 2).

Table 1 Field-scale landfarming of crude oil-impacted soil of the Niger Delta region, Nigeria.

Parameter	Details
Site and study	Landfarming study was carried out in a community in Emohua local government area in Rivers state, Nigeria (E 6.804116; N 4.969581) (Chikere et al., 2019). Duration of study was 56 days.
Microbiological analysis	- Soil samples of 0–2 m depth (from landfarm) were collected to analyze microbiological characteristics - Microbial cultures (bacteria and fungi) were isolated by using crude oil and paraffin wax as enrichment substrates. - Degradative potentials (diesel fuel) of these cultures were analyzed using culture-dependent methods (2,6-dichlorophenol indophenol (DCPIP) decolorization test). - Further hydrocarbon degradation potentials were tested using 96-well microtiter plate, and degradation potentials were monitored using Varioskan Flash Plate Reader for 0–144 h
Genomic analysis	- Soil genomic DNA (using Zymo Research fungal/bacterial DNA MiniPrep kit) was analyzed in this study. - Both bacterial and fungal sequences were subjected to phylogenetic analysis
Findings	- One hundred and twenty-eight bacterial and 52 fungal species were isolated in this study by enrichment techniques (crude oil and paraffin wax substrates). - In DCPIP decolorization test, 31 out of 128 isolates have shown decolorized diesel completely within 24 h, while 16 out of 31 strains have hydrocarbon degradation potential. - Crude oil and paraffin wax encouraged the growth of hydrocarbon utilizing bacteria (HUB) by 99.97% and 80.36%, respectively. - Both bacteria (of Proteobacteria and Firmicutes) and fungi (of Eurotiomycetes, Sordariomycetes, and Saccharomycetes) were recovered from the landfarming soil. - Maximum counts of culturable heterotrophic bacteria (CHB) and HUB were 6.80×10^7 (achieved on 18 and 36 day) and 6.80×10^7 (on day 36), respectively and their corresponding 0-day counterparts were 0.20×10^3 and 0.50×10^3, respectively. - TPHs concentration in landfarm was decreased from 8635.68 to 677.2 mg kg⁻¹ in 56 days

Table 2 Landfarming of diesel-polluted soil in Arctic region, Greenland.

Parameter	Details
Site and study	- This landfarming study was carried out at 9117 Mestersvig, sandy river delta (UTM zone 27: 400500 E; 8717500 N) in Greenland (Johnsen et al., 2021). It is an arctic climate with a yearly mean temperature and precipitation of -10°C and 300 mm year⁻¹, respectively. - Approximately 3000 m³ contaminated soil was distributed in an area of 8500 m² with 0.4 m thick layer in landfarm of 13,600 m² (160 × 85 m). - Fast release N-fertilizer (1636 kg ha⁻¹) and NPK (6523 kg ha⁻¹) were added to landfarms as a part of biostimulation
Microbiological analysis	Samples were most probable number (MPN)-enumerated for degraders that could grow on different hydrocarbons, i.e., arctic grade C diesel (10%, v/v), undecane (10%, v/v), <i>m</i>-xylene (15, v/v), 2-methylnaphthalene (2%, w/v), phenanthrene (saturated solution), biphenyl (25%, w/v), benzothiophene (0.50%, w/v), cyclodecane (10%, v/v), 1,3-dimethylcyclohexane (10%, v/v), <i>n</i>-butylbenzene (1%, v/v), sec-butylbenzene (1%, v/v), 1-hydroxynaphthalene (0.50%, w/v), or 1-naphthoic acid (0.50%, w/v) as single carbon sources.
Findings	- TPHs were removed by 64% in the first year (2012–13) and a total of 88% TPHs were removed by the end of landfarming (2012–17). - MPN of Arctic diesel degraders was changed from 1.30×10^6 to 1.20×10^9 cells g⁻¹ between 2012 and 2015 which was corresponded to 40% removal of TPHs during the above period. - There was a decline in MPN of diesel degraders from 1.20×10^9 to 1.10×10^8 cells g⁻¹ between 2015 and 2017, probably TPHs were already degraded by that period. - Higher log (MPN) (i.e., >6–10) in degraders of undecane, <i>m</i>-xylene, 2-methylnaphthalene and phenanthrene was found in three landfarms compared to log (MPN) of ~5 to 1 in background soil samples, in situ remediated soil, basin 1 water, basin 1 sediment, basin 2 water, and basin 2 sediment. - MPN counts for different hydrocarbons were as follows: diesel (<800 cells g⁻¹), cyclodecane (<800 cells g⁻¹), <i>n</i>-alkane (5.50×10^8 cells g⁻¹), <i>n</i>-butylbenzene (2.40×10^6 cells g⁻¹), sec-butylbenzene (1.50×10^8 cells g⁻¹), <i>m</i>-xylene (3.0×10^7 cells g⁻¹) and 1-hydroxynaphthalene (<800 cells g⁻¹). - MPNs were also detected for phenanthrenes, 1-methylnaphthalene, and 2-methylnaphthalene. Detection of MPNs for later two compounds indicate the microbial degradation of C1–naphthalene. - Cultivable benzothiophene MPNs were not detected, which implied that there was no biodegradation of benzothiophenes and dibenzothiophenes. - The ratio of 2MN/1MN (MN, methylnaphthalene) dropped to 84% in the landfarm in 2012, suggesting that biodegradation of naphthalenes had already started by the time of excavation and landfarm construction. In 2013, 2MN/1MN ratio in the landfarm was dropped further to 37%, indicating that there was high C1–naphthalene biodegradation. In the following period (i.e., 2014–17), 2MN/1MN ratio was almost constant which implied that C1–naphthalenes were almost completely removed by 2013. - Regarding alkane degradation, the ratios of nC17/Pr (Pr, Phenanthrene) and nC18/Pr were found to be 17% and 18%, respectively until 2016 landfarming, which suggested that there was continuous degradation of <i>n</i>-alkanes, and this was corroborated with the very high number of undecane degraders in the landfarm soils

Molecular microbial diversity and abundance

The microbial capabilities to degrade the contaminants have relied on the knowledge gained from the cultivated microorganisms before the introduction of molecular methods. The nucleic acid-based techniques such as electrophoresis-based fingerprinting, sequencing, metagenomics, and other omics-based analyses demonstrate the enormous taxonomic- and metabolic diversity of microorganisms involved in the degradation of contaminants at the landfarming sites. Metagenomics show the diversity and abundance of several microorganisms and their metabolic networks at the landfarming sites (Bergsveinson et al., 2019; Kim et al., 2021). In a study on landfarming of TPH-contaminated soil, the operational taxonomic units (OTUs), Chao1 richness index, and Shannon diversity index were found to decrease at the TPHs-contaminated site, relative to the control site, after the landfarming treatment (Kim et al., 2021). The respective values of OTUs, Chao 1 and Shannon indices for the landfarm site were 815 ± 111 , 1000 ± 101 , 7.06 ± 0.29 , while those for the control site were 1265 ± 308 , 1475 ± 340 , and 8.58 ± 0.62 . During the degradation of TPHs, there were increases in the relative abundance (%) of *Proteobacteria* and *Acidobacteria* at the landfarming site compared to the control site (Fig. 5). Nevertheless, other bacterial phyla (notably *Actinobacteria*, *Firmicutes*, *Chloroflexi*, *Gemmatimonadetes*, and *Bacteroidetes*) decreased at the landfarm site over the control site (Kim et al., 2021). Biochemically, there were enhanced microbial respiration rates, soil contents of microbial biomass-nitrogen, and increased urease activities (about 21–45% in 1–4 months) at the PAHs-contaminated landfarming sites over the control sites (Besalatpour et al., 2011).

The success of the landfarming used to restore the long-term contaminated soil can be assessed by novel approaches, such as the identification of prokaryotic communities using combined profiling of both rRNA genes and rRNA. This approach allows us to investigate pollution-dependent differences at a molecular level (i.e., 16S rRNA and rRNA gene pools). The investigations on how different biogeochemical cycling processes mediated by bacteria are affected at the landfarms show the significance of nitrogen cycling microbial communities on contaminant degradation. The abundance of ammonia-oxidizing bacteria (AOB) in the landfarming soil containing high concentrations of oil hydrocarbons ($10,000\text{--}15,000 \mu\text{g g}^{-1}$) and PAHs ($172\text{--}305 \mu\text{g g}^{-1}$) were in the range of 4×10^5 and 9×10^5 cells g^{-1} dwt soil, respectively (Kurola et al., 2005). Further phylogenetic analysis revealed that out of a total of 10 different AOB-like sequences recovered, six sequences were grouped with *Nitrospira* cluster 2 and four with *Nitrospira* cluster 3, suggesting that there were no major temporal changes in the AOB community composition during the studied period of landfarming (Kurola et al., 2005). Oil sludge containing microcosms treated by the landfarming technique was found to have the following predominant bacterial families: *Comamonadaceae*, *Ectothiorhodospiraceae*, *Pseudomonadaceae*, *Sphingomonadaceae*, *Xanthobacteraceae*, and *Xanthomonadaceae* (Dörr de Quadros et al., 2016). Soil amendments significantly impact the microbial community structure at the landfarming sites, which can be discerned by applying molecular methods. The addition of sludge, compost, nutrients, fern, or even the TPH-degrading bacteria to the diesel- and lubricant oil-contaminated soils treated by the enhanced landfarming technique resulted in an increased abundance of microbial species, with slight changes in the dominant species (Wang et al., 2016). However, molecular methods have made the identification of individual microbial species easier, which is otherwise difficult by the cultivation-based methods. In the same study, 34 specific TPH-degrading bacteria have been identified in the microcosm soils—these bacteria include *Acinetobacter* sp., *Acidovarax* sp., *Aquabacterium* sp., *Aquicola tertiarycarbonis*, *Azoarcus* sp., *Bordetella* sp., *Burkholderia* sp., *Caldimonas* sp., *Chlamydia* sp., *Chloroflexi* sp., *Eubacterium* sp., *Ideonella* sp., *Lautropia* sp., *Leptothrix* sp., *Leptothrix ginsengisoli*, *Limnohabitans* sp., *Methyloversatilis* sp., *Microbacterium* sp., *Pseudomonas aeruginosa*, *Pseudomonas* sp.,

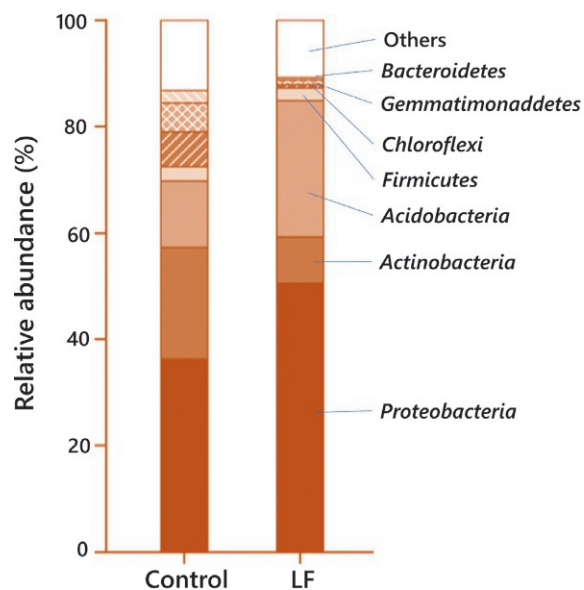


Fig. 5 Relative abundance (%) of bacteria at phylum level in a total petroleum hydrocarbon (TPH)-contaminated soil after landfarming (LF). Source: Kim J-W, Hong Y-K, Kim H-S, et al. (2021) Metagenomic analysis for evaluating change in bacterial diversity in TPH-contaminated soil after soil remediation. *Toxics* 9: 319.

Pseudoxanthomonas sp., *Ralstonia* sp., *Rhodoferrax* sp., *Thiothrix* sp., *Variovorax* sp., *Verrucomicrobium* sp., and *Zoogloea* sp. Recently, the impact of soil conditions on microbial communities in a landfarm for hydrocarbon waste (Regina, Canada) was investigated through the analysis of bacterial and fungal community structures using the next-generation sequencing (NGS) techniques (Bergsveinson et al., 2019). While the diversity and structure of bacterial communities were stable, the fungal community structure was less stable due to the influences of the variable soil oxygen conditions. In the landfarm metagenomics identified ~1200 and ~400 OTUs bacteria and fungi, respectively, belonging to 23 bacterial and seven fungal phyla. Bacterial diversity was significantly greater in the surface soil with different landfarming sections or depths than the fungal diversity.

Evaluation of landfarming efficacy

The effectiveness of landfarming technology is evaluated based on three significant components: soil, contaminants, and climatic conditions. The soil characteristics are assessed for many parameters, including the population densities of microbial degraders, pH, moisture content, temperature, nutrient concentrations, and soil texture (Fig. 6). The volatility, chemical structure, concentration, and toxicity are the essential chemical parameters considered for the contaminants (Fig. 7). The key elements of climate, such as ambient temperature, rainfall, and wind, are also appraised.

The microbial activities necessitate an evaluation of the type and degree of contaminants. The assessment helps to exclude the substances that might be toxic to certain microbial species at the site. Under the following conditions the contaminated site is usually considered hazardous and landfarming technology is not feasible due to toxicities to soil microorganisms (FCSAP, 2013).

- TPHs or total extractable hydrocarbons (TEHs) are >3%
- Total heavy metal concentration is >2500 ppm
- Electrical conductivity is >4 dS m⁻¹
- The sodium adsorption ratio (SAR) is >6

For effective landfarming, the population densities and dynamics of contaminant-degrading microbial communities are paramount. The preferred minimum heterotrophic plate count for effective landfarming is 10³ CFU g⁻¹ (Fig. 6). If the microbial population density is below the minimum value, landfarming can still be effective if one of the following two approaches is adopted—(i) stimulating existing bacteria by adding nutrients and (ii) soil amendments to increase the bacterial population. The addition of non-indigenous microflora to the contaminated sites has generally had limited success (FCSAP, 2013). It is always preferable to activate the indigenous microorganisms for achieving effective landfarming. Though commercially available microbial cultures are also popular, the cold-adapted native microorganisms for the arctic and sub-arctic regions are not readily available in landfarming technology.

Biotreatability studies are beneficial in determining the suitability of landfarming technology. Examining gas chromatography (GC) data is useful in pre-assessment of landfarming efficiency. The GC data improve understanding the influence of the chemical composition of a pollutant on landfarming efficacy. It is significant in deciding the site-specific suitability and estimating the necessity of other remedial approaches for the landfarming technology. The biotreatability studies determine how readily treatable are the contaminants or wastes present, and establish the disappearance, over time, of wastes due to the activities of either indigenous or exogenously added microorganisms. The chemical, physical and biological conditions are all considered during the biotreatability studies. Overall, treatability studies are set to determine the following parameters:

- Biodegradability of key chemical(s)
- Rate of O₂ consumption or CO₂ generation as an indicator of biological activity
- Rate of chemical substance(s) degradation, and
- The need for microbial species with degradation efficiencies and determine the effects of microorganisms on degradation rates.

The negative impacts (i.e., terrestrial ecotoxicity, aquatic ecotoxicity, and aquatic eutrophication) of landfarming can be evaluated by an appropriate life cycle assessment (LCA). The LCA of landfarming technology can help to treat the contaminated site in an effective eco-friendly manner. Evaluation of LCA at a landfarming site, using software such as SigmaPro 9.0™, makes it easy to identify environmental hotspots and to reduce the ecological impact.

Summary

Landfarming is an established ex-situ bioremediation technology, and has received wide public acceptance owing to its suitability for the remediation of different contaminant types, the biodegradation potential, the facility design and operation, and the clean-up levels that can be achieved. This bioremediation technology has successfully restored soils contaminated by kerosene, diesel, jet fuel, crude oil, oil-based drilling wastes, oily sludge, tank bottoms, and pit sludges. The susceptibility of contaminants to microbial attack during landfarming is greatly influenced by their molecular structure, chemical properties, and the mixture effects. The key factors for effective landfarming include, but are not limited to, the initial concentration of contaminants, soil pH, aeration or tillage, and

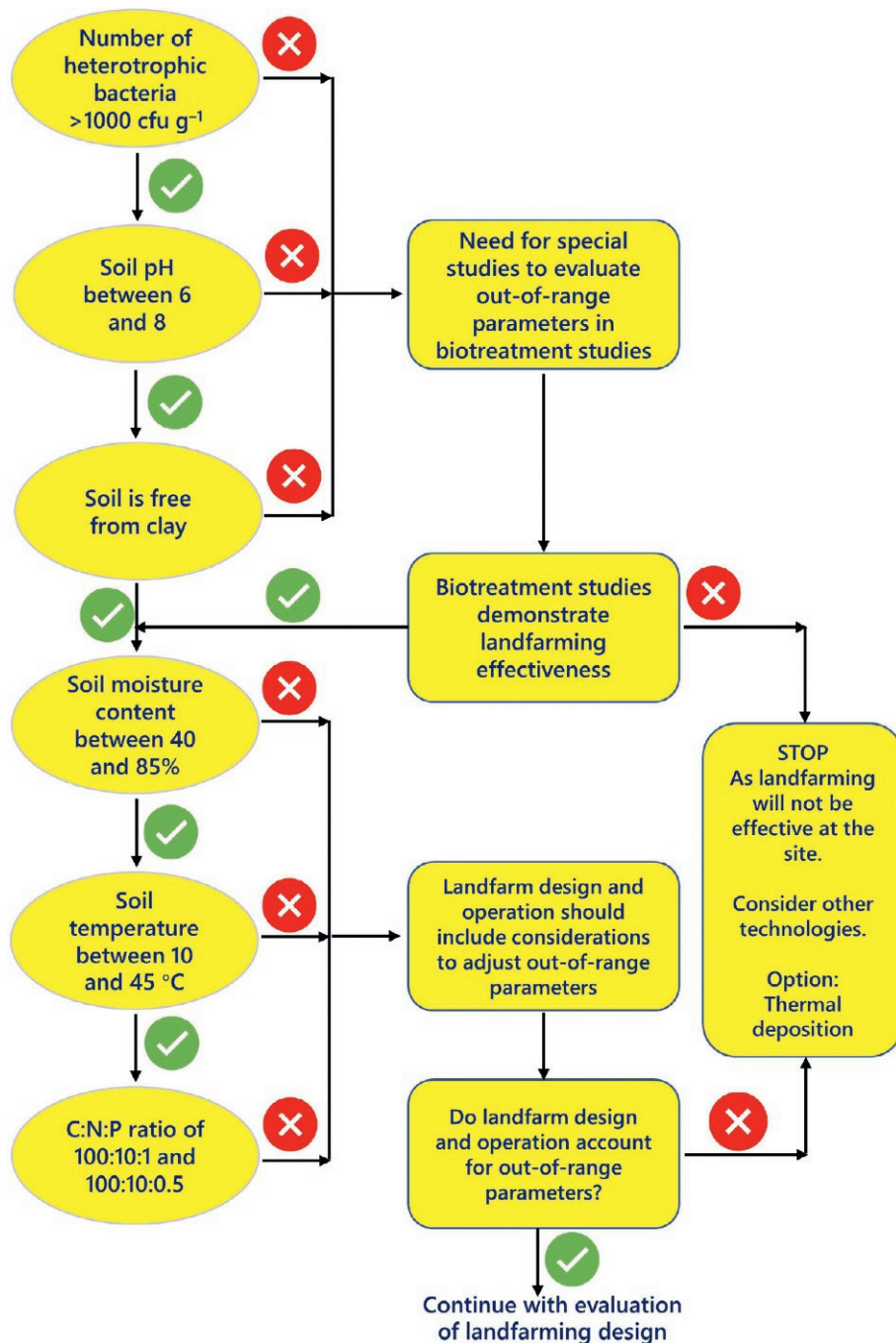


Fig. 6 Analysis of soil characteristics in the evaluation of landfarm efficacy. Source: EPA (2017) How to Evaluate Alternative Cleanup Technologies for Underground Storage Tank Sites, a Guide for Corrective Action Plan Reviewers Chapter 5 Landfarming. United States Environmental Protection Agency—October 2017.

the contents of salt and nutrients. The new insights related to biostimulation, bioaugmentation, and the diversity and metabolic activities of microorganisms can help further improve the landfarming technology's applicability to various contaminated sites.

It may thus be concluded that:

- the success of landfarming technology is considerably impacted by pH, moisture content, temperature, nutrient concentration, and texture of the soil, the chemical structure, concentration, volatility, and toxicity of contaminants, and the ambient temperature and rainfall at treatment sites.
- this ex-situ bioremediation technology in combination with biostimulation and bioaugmentation management strategies can effectively restore contaminated sites.

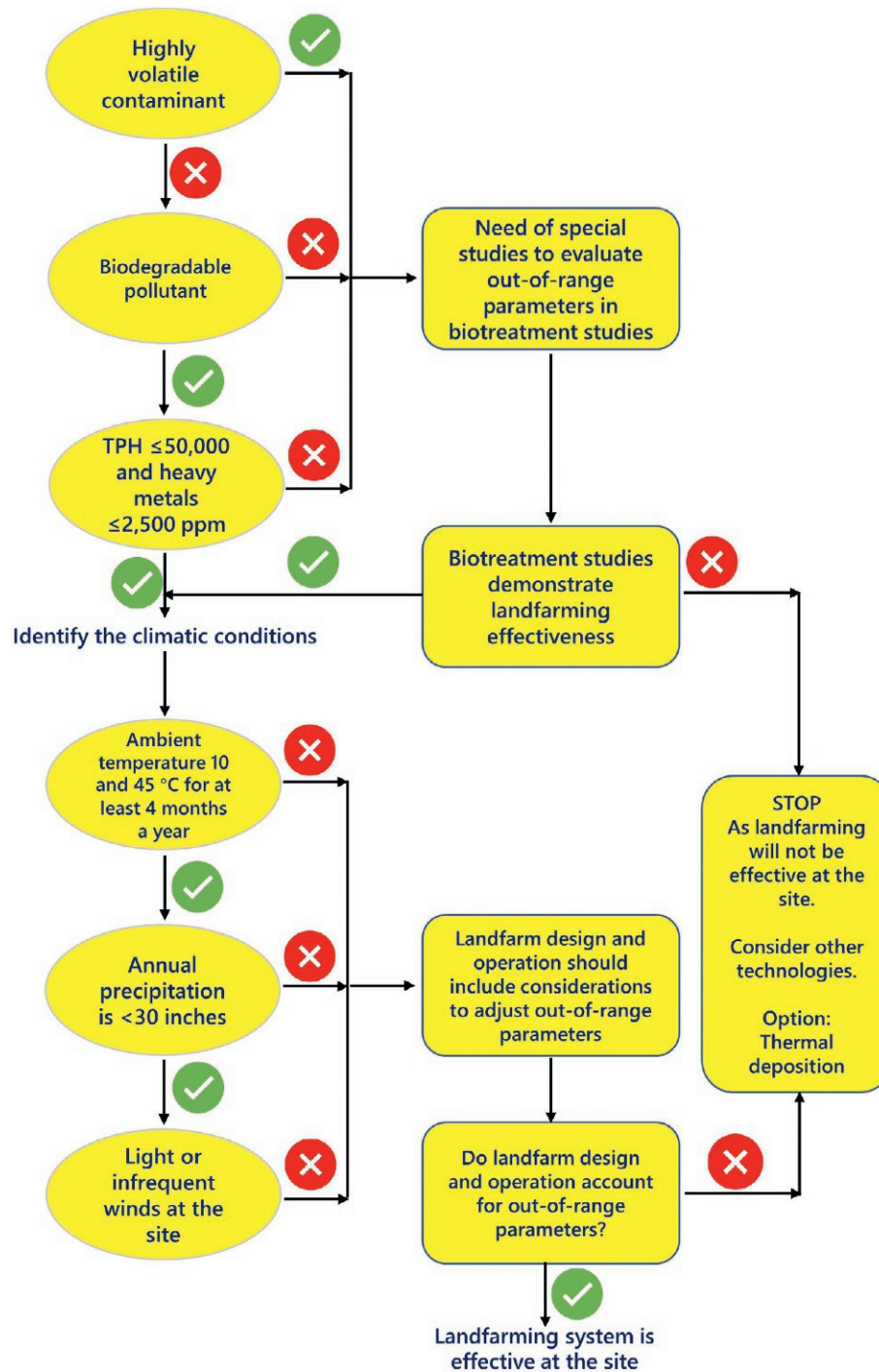


Fig. 7 Analysis of pollutant characteristics and ambient conditions in the evaluation of landfarm efficacy. Source: EPA (2017). How to Evaluate Alternative Cleanup Technologies for Underground Storage Tank Sites, a Guide for Corrective Action Plan Reviewers Chapter 5 Landfarming. United States Environmental Protection Agency—October 2017.

- bacteria and fungi are the primary active microbial groups at treatment sites; the richness and abundance of these microbial groups are greatly influenced by the management practices that are adapted at the treatment site and the nature of the contaminant.
- field-scale studies provide the scope and useful insights into understanding the substantial role and the microbial potential for contaminant degradation in landfarming technology.

Further insights into microbial degradation potential and their metabolic networks are required if we are to make landfarming more effective and more eco-friendly. Knowledge from future investigations may help to overcome the constraints and limitations of

landfarming technology related to moisture, temperature, pollutant bioavailability, nutrient availability, and aeration. Focusing on advanced molecular approaches, enzyme engineering, omics-based approaches, and other emerging techniques will provide even better tools for effectively applying diverse microbial groups at the landfarming sites. Such approaches can help identify the suitability of a microbial species or community for a specific pollutant. The new advances and a better understanding of microbial dynamics and networking will help optimize the environmental conditions for enhanced biodegradation activities during landfarming. Landfarming technology that employs the passive treatment with minimal costs and energy is very promising for the remediation of hydrocarbon-, PAH- and mineral oil-contaminated soils or sediments. The knowledge and information on when, and to what extent, the method can be optimized to improve microbial activities, the bioavailability of the contaminants, and the aerobic conditions are critical to popularize and apply this technology to other contaminants. Regular monitoring of microbial degradation efficiencies can provide the decision support on the efficacy of landfarming technology and tracking when regulatory targets are attained during the remediation of contaminants.

References

- API (2003) *American Petroleum Institute (1983) Land Treatment Practice in the Petroleum Industry*. report prepared by Washington, DC: Environmental Research and Technology Inc.
- Baron NC, Pagnocca FC, Otsuka AA, et al. (2021) Black fungi and hydrocarbons: An environmental survey for alkylbenzene assimilation. *Microorganisms* 9: 1008.
- Bergsveinson J, Perry BJ, Simpson GL, et al. (2019) Spatial analysis of a hydrocarbon waste-remediating landfarm demonstrates influence of management practices on bacterial and fungal community structure. *Microbial Biotechnology* 12: 1199–1209.
- Besalatpour A, Hajabbasi MA, Khoshgoftarmansh AH, and Dorostkar V (2011) Landfarming process effects on biochemical properties of petroleum-contaminated soils. *Soil and Sediment Contamination: An International Journal* 20: 234–248.
- Brown DM, Okoro S, van Gils J, et al. (2017) Comparison of landfarming amendments to improve bioremediation of petroleum hydrocarbons in Niger Delta soils. *Science of the Total Environment* 596: 284–292.
- Chien H, Kao C, Liu J, Takagi K, and Surampalli R (2010) Clean up of petroleum-hydrocarbon contaminated soils using enhanced bioremediation system: Laboratory feasibility study. *Journal of Environmental Engineering* 136: 597–606.
- Chikere CB, Tekere M, and Adeleke R (2019) Enhanced microbial hydrocarbon biodegradation as stimulated during field-scale landfarming of crude oil-impacted soil. *Sustainable Chemistry and Pharmacy* 14: 100177.
- Cookson JT Jr. (1995) *Bioremediation Engineering: Design and Application*. New York: McGraw-Hill, Inc.
- da Silva LJ, Oliveira FJS, and de França FP (2013) Oil waste management through large scale landfarming: A case study. *International Journal of Environment and Waste Management* 11: 233–243.
- Dórr de Quadros P, Cerqueira VS, Cazarolli JC, et al. (2016) Oily sludge stimulates microbial activity and changes microbial structure in a landfarming soil. *International Biodeterioration and Biodegradation* 115: 90–101.
- EPA (2017) How to evaluate alternative cleanup technologies for underground storage tank sites, a guide for corrective action plan reviewers. In: *Chapter 5 Landfarming*. United States Environmental Protection Agency. October 2017.
- FCSAP (2013) *Federal Guidelines for Landfarming Petroleum Hydrocarbon Contaminated Soils March 2006, Editorial Update 2013*. Federal Contaminated Sites Action Plan.. https://www.canada.ca/content/dam/eccc/migration/fcs-scf/B15E990A-C0A8-4780-9124-07650F3A68EA/Landfarming_en.pdf. (accessed 10 March 2022).
- Guarino C, Spada V, and Sciarillo R (2017) Assessment of three approaches of bioremediation (natural attenuation, landfarming and bioaugmentation—Assisted landfarming) for a petroleum hydrocarbons contaminated soil. *Chemosphere* 170: 10–16.
- Jacques RJS, Okeke BC, Bento FM, Peralba MCR, and Camargo FAO (2007) Characterization of a polycyclic aromatic hydrocarbon-degrading microbial consortium from a petrochemical sludge landfarming site. *Bioremediation Journal* 11: 1–11.
- Jacques RJS, Okeke BC, Bento FM, et al. (2008) Microbial consortium bioaugmentation of a polycyclic aromatic hydrocarbons contaminated soil. *Bioresource Technology* 99: 2637–2643.
- Johnsen AR, Boe US, Henriksen P, Malmquist LMV, and Christensen JH (2021) Full-scale bioremediation of diesel-polluted soil in an Arctic landfarm. *Environmental Pollution* 280: 116946.
- Kim J-W, Hong Y-K, Kim H-S, et al. (2021) Metagenomic analysis for evaluating change in bacterial diversity in TPH-contaminated soil after soil remediation. *Toxics* 9: 319.
- Kim SH, Woo H, An S, Chung J, Lee S, and Lee S (2022) What determines the efficacy of landfarming for petroleum-contaminated soils: Significance of contaminant characteristics. *Chemosphere* 290: 133392.
- Kurola J, Salkinoja-Salonen M, Aarnio T, Hultman J, and Romantschuk M (2005) Activity, diversity and population size of ammonia-oxidising bacteria in oil-contaminated landfarming soil. *FEMS Microbiology Letters* 250: 33–38.
- Nduka J, Umeh L, Okerulu I, Umedum L, and Okoye H (2012) Utilization of different microbes in bioremediation of hydrocarbon contaminated soils stimulated with inorganic and organic fertilizers. *Journal of Petroleum & Environmental Biotechnology* 3: 1–9.
- Peltola R, Salkinoja-Salonen M, Pulkkinen J, et al. (2006) Nitrification in polluted soil fertilized with fast- and slow-releasing nitrogen: A case study at a refinery landfarming site. *Environmental Pollution* 143: 247–253.
- Porto de Souza VL, Libardi Junior N, Valladares-Diestra KK, et al. (2022) Chapter 16—Enzymatic bioremediation: Current status, challenges, future prospects, and applications. In: Rodriguez-Couto S and Shah MP (eds.) *Development in Wastewater Treatment Research and Processes*, pp. 355–381. Elsevier.
- Sakshi K, Singh SK, and Haritash AK (2020) Evolutionary relationship of polycyclic aromatic hydrocarbons degrading bacteria with strains isolated from petroleum contaminated soil based on 16S rRNA diversity. *Polycyclic Aromatic Compounds*. <https://doi.org/10.1080/10406638.2020.1825003>.
- Silva-Castro GA, Uad I, González-López J, et al. (2012) Application of selected microbial consortia combined with inorganic and oleophilic fertilizers to recuperate oil-polluted soil using landfarming technology. *Clean Technologies and Environmental Policy* 14: 719–726.
- Silva-Castro GA, Uad I, Rodríguez-Calvo A, González-López J, and Calvo C (2015) Response of autochthonous microbiota of diesel polluted soils to land-farming treatments. *Environmental Research* 143: 49–58.
- Sims R and Sims J (2003) Landfarming framework for sustainable soil bioremediation. In: Šašek V, Glaser JA, and Baveye P (eds.) *The Utilization of Bioremediation to Reduce Soil Contamination: Problems and Solutions*, pp. 319–334. Dordrecht: Springer.
- Straube W, Nestler C, and Hansen L (2003) Remediation of polyaromatic hydrocarbons (PAHs) through landfarming with biostimulation and bioaugmentation. *Acta Biotechnologica* 23: 179–196.
- Wang S-Y, Kuo Y-C, Hong A, Chang Y-M, and Kao C-M (2016) Bioremediation of diesel and lubricant oil-contaminated soils using enhanced landfarming system. *Chemosphere* 164: 558–567.
- Zheng Z and Obbard JP (2003) Oxidation of polycyclic aromatic hydrocarbons by fungal isolates from an oil contaminated refinery soil. *Environmental Science and Pollution Research* 10: 173–176.

Relevant websites

CPEO—<http://www.cpeo.org/techtree/ttdescript/lanfarm.htm>

Enviro Wiki—<http://www.enviro.wiki/index.php?title=Landfarming>

En Wikipedia—<http://en.wikipedia.org/wiki/Landfarming>

FRTR—<http://frtr.gov/matrix/Landfarming/>

EPA—http://www.epa.gov/sites/default/files/2014-03/documents/tum_ch5.pdf

Biodegradation of organic contaminants

PB Hatzinger^a and JW Kelsey^b, ^aBiotechnology Development and Applications, APTIM, Lawrenceville, NJ, United States; ^bEnvironmental Science and Chemistry, Muhlenberg College, Allentown, PA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	547
Biodegradation	548
Aerobic transformation processes	548
Anaerobic transformation processes	549
Chemical properties affecting biodegradation	550
Environmental factors affecting biodegradation	552
Bioavailability	552
Nutrient availability	553
pH	554
Temperature	554
Water content	554
Presence of appropriate microorganisms	554
Predation	554
Presence of other compounds	555
Plants	555
Potential impacts of climate change on biodegradation	555
Changes in pH	555
Changes in Earth's temperature	556
Summary	556
References	557

Abstract

Perhaps the most important parameter determining the fate of an organic contaminant in soil is its susceptibility to biological degradation. This article summarizes current knowledge on chemical biodegradation, including microbial mechanisms, chemical and environmental factors that govern rates and extents of biodegradation in soils, and likely impacts of climate change on this process.

Key points

- Microorganisms biodegrade most organic chemicals.
- Aerobic and anaerobic transformation processes occur.
- Numerous chemical and environmental factors affect rates of contaminant biodegradation.
- Climate change is likely to impact biodegradation processes in many ways.

Introduction

Industrialized societies annually release large quantities of synthetic chemicals into Earth's natural environments, both intentionally, in the case of pesticides, and inadvertently, through the production, transport, refining, and combustion of fossil fuels, waste streams from manufacturing, chemical spills, disposal of products by consumers, and discharges from a wide variety of other sources. Currently, there are more than 86 000 chemicals registered in the US Environmental Protection Agency (US EPA) inventory of chemical substances manufactured or processed in, or imported into, the United States (USEPA, 2021). The number of potential chemical pollutants is substantially higher, as the EPA inventory does not include pesticides, radioactive materials, food additives, drugs, or cosmetics. Many of these compounds ultimately end up as contaminants in the atmosphere, water, sediment, or soil. To add to the complexity of the problem, the compounds of concern vary widely in their characteristics. A contaminant can belong to any of a number of disparate classes of compounds that differ in important properties such as charge, polarity, water solubility, acid dissociation constant (pK_a), molecular weight, and aromaticity. These chemical characteristics subsequently determine the fate of a contaminant in the environment and the hazard it poses to humans and other organisms.

Perhaps the most important parameter determining the fate of a given chemical is its susceptibility to biological degradation. There have been more than seven decades of research into chemical biodegradation, including microbial mechanisms and the chemical and environmental factors that govern its rate and extent in soils.

Biodegradation

Biodegradation is the partial or complete breakdown of a chemical pollutant resulting from the physiological reactions of microorganisms. In general, microorganisms carry out such processes to obtain necessary carbon, nutrients, and energy for their survival and growth. Through the millennia, bacteria have experienced extreme competition (primarily from each other) for limiting resources. As a result, they have evolved unrivaled metabolic capabilities, utilizing nearly every thermodynamically favorable reaction for energy and nearly every organic molecule for carbon. For example, whereas humans gain energy only through the oxidation of specific organic substrates using oxygen as an electron acceptor (i.e., through aerobic respiration), bacteria utilize virtually all natural and many man-made organics, numerous inorganic molecules (e.g., Fe(0); Fe(II), H₂S, S(0) H₂, NH₄⁺), and light for energy (Madigan et al., 2018). Moreover, besides oxygen, specific microorganisms can utilize a variety of chemicals, including an assortment of synthetic pollutants and pesticides, as terminal electron acceptors. Specific examples of microbial biodegradation processes under aerobic and anaerobic conditions are provided in subsequent sections.

The microbial degradation of contaminants can yield many different products. Generally, the biodegradation of a large organic molecule results in smaller, less complex organic molecules that may be more or less toxic than the parent compound. In some cases, an organic contaminant is completely converted to an inorganic product such as CO₂. A transformation of this type is referred to as “mineralization” and is associated with the elimination of toxicity of the parent compound. Although plants, animals, and microorganisms can all metabolize some pollutants, the mineralization of organic contaminants is generally carried out by bacteria, fungi and archaea only. These three groups of microorganisms are responsible for nearly all the relevant biotransformations of environmental contaminants.

Whereas biodegradation occurs naturally in soils and other environments without any intervention, bioremediation involves the application of natural processes to remove pollutants from contaminated sites (Alexander, 1999). In short, bioremediation is the use of microorganisms to clean up unwanted compounds, although the specific technology employed will depend on the nature of the site and the contaminant-of-concern. Contaminants may be treated *in situ* through the addition of nutrients to stimulate indigenous microorganisms (biostimulation) or by the inoculation of specific microorganisms capable of degrading the compound (bioaugmentation). Alternatively, contaminated soils can be excavated and treated *ex situ* using biological reactors, composting, or other approaches.

The biodegradation of soil pollutants is critical because of the impact the resultant transformations can have on human and environmental health. For example, in the absence of appreciable degradation, compounds released into soil may leach to groundwater. Contamination of groundwater poses a serious threat to drinking water supplies and public health. Moreover, compounds present in groundwater tend to persist much longer than those in soils due to the limited supplies of oxygen and inorganic nutrients and low microbial numbers generally found in subsurface aquifers. Soil pollutants and pesticides can also pose a threat to human and nonhuman health if they are bioconcentrated by soil organisms. Biological uptake directly from soil can adversely affect both individual organisms and food webs. A contaminant may be sufficiently concentrated in soil to harm the organism that contacts it initially. However, certain compounds, even if originally present at low concentration, can be accumulated and magnified through herbivory and predation. A biomagnified compound has the potential to damage organisms at higher trophic levels, despite their very limited contact with the contaminated soil. Alternatively, some compounds are so potent that they are a threat at very low concentrations. Microbial transformations can lower the risk associated with a particular contaminant because they reduce or eliminate the toxicity of the compound.

Aerobic transformation processes

Many different synthetic organic pollutants and pesticides have been shown to serve as growth substrates for specific bacteria and fungi under aerobic conditions, providing both carbon and energy to these organisms. Often, these compounds are completely mineralized by one organism or by a series of different organisms to CO₂ and cell constituents. During this process, the novel pollutant is generally modified by specific enzymes to one or more metabolites that are common in cellular metabolism (e.g., pyruvic acid, acetyl-CoA). Those metabolites are then converted to energy and biomass through the common metabolic pathways. Aerobic mineralization is typically a growth-linked process whereby microbial numbers increase as the chemical concentration decreases. Growth-linked biodegradation is desirable in that it can bring about a rapid loss of a pollutant from soil. Many different organic contaminants are subject to mineralization in soils, including most components of crude oil and gasoline, and numerous pesticides.

A variety of organic chemicals are biotransformed through a non-growth-linked process termed “cometabolism.” Cometabolism is the fortuitous transformation of a molecule by an enzyme synthesized within a cell for another purpose. By this process, the cell modifies the chemical structure, but gains no energy, carbon, or inorganic nutrients from the reaction. Several groups of aerobic bacteria present in soils, including methanotrophic bacteria (which oxidize methane for energy), nitrifying bacteria (which oxidize ammonia for energy), and ethane- propane-, isobutane-, benzene-, phenol-, and toluene-oxidizing bacteria, have been observed to transform a variety of different compounds through cometabolic reactions. Each of the above groups of bacteria produces nonselective monooxygenase or dioxygenase enzymes that are responsible for these reactions. For example, methanotrophic

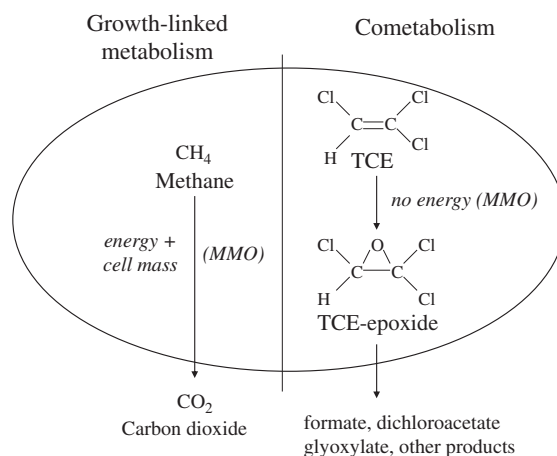


Fig. 1 Representation of an individual methanotrophic bacterium degrading methane to carbon dioxide through growth-linked metabolism and degrading trichloroethylene (TCE) to TCE-epoxide via cometabolism. MMO, methane monooxygenase.

bacteria produce methane monooxygenase (MMO), an enzyme that normally functions to oxidize methane for energy (Fig. 1). However, MMO has a broad substrate specificity and has been observed to catalyze the initial oxidation of numerous alkanes, alkenes, and aromatics as well as the degradation of an array of chlorinated aliphatics, such as *cis*- and *trans*-1,2-dichloroethylene (*cis*- and *trans*-1,2-DCE), trichloroethylene (TCE), and 1,2-dichloropropane (Frascari et al., 2015). As previously noted, bacteria do not gain carbon or energy from the cometabolic oxidation of most compounds; thus, an additional cosubstrate is often required for bacterial growth (e.g., toluene for toluene oxidizers, methane for methanotrophs, etc.). Although cometabolic reactions sometimes result in the accumulation of persistent intermediates, in many instances, the initial oxidative reaction produces readily degradable compounds that are then mineralized. For example, as shown in Fig. 1, the initial cometabolic attack of MMO on TCE creates an unstable TCE-epoxide, which further degrades to a variety of growth substrates, such as glyoxylate, formate, and dichloroacetate.

Most recently, cometabolic biodegradation research has focused on a wide variety of emerging contaminants, including methyl *tert*-butyl ether (fuel oxygenate), 1,4-dioxane (stabilizer in chlorinated solvents), *n*-nitrosodimethylamine (disinfection and liquid rocket fuel byproduct), 1,2-dibromoethane (soil fumigant and scavenger in leaded fuels), 1,2,3-trichloropropane (solvent and pesticide manufacturing byproduct), and a broad assortment of trace-level pharmaceuticals, antibiotics, estrogens, and personal care products detected in wastewater treatment plant sludges that may ultimately end up in soils (e.g., Fischer and Majewsky, 2014). Often, these compounds are present in the environment at concentrations too low to support microbial growth, but higher than relevant regulatory or toxicity thresholds. Because microbial growth is dependent on a secondary substrate (e.g., propane or methane) during cometabolism rather than the contaminant-of-concern, the process often can be used as a bioremediation strategy to effectively treat the aforementioned contaminants to very low concentrations (e.g., ng/L in water). This cometabolic bioremediation process typically entails adding a gaseous growth substrate, one of more inorganic nutrients (e.g., ammonia), and oxygen to the impacted environment to stimulate indigenous bacteria capable of carrying out the cometabolic transformations.

Anaerobic transformation processes

A wide variety of contaminants and pesticides are subject to biodegradation under anaerobic conditions. As described for aerobic degradation, mineralization and cometabolic reactions also occur in anaerobic environments. For some compounds, biodegradation will only occur under such conditions, since the organisms capable of carrying out the process are obligate anaerobes (i.e., they die in the presence of oxygen). Under anaerobic conditions, the degradation of specific pollutants and pesticides can be coupled to the reduction of terminal electron acceptors other than O_2 , such as NO_3^- , Mn(IV) , Fe(III) , SO_4^{2-} , or CO_2 . This process is termed "anaerobic respiration." The gasoline constituent toluene, for example, has been reported to be biologically degraded with each of these different electron acceptors. For a given substrate, the energy yield during anaerobic respiration is less than for the same reactions using oxygen ($\text{O}_2 > \text{NO}_3^- > \text{Mn(IV)} > \text{Fe(III)} > \text{SO}_4^{2-} > \text{CO}_2$) as an electron acceptor (Madigan et al., 2018). Therefore, the anaerobic reactions tend to occur more slowly, and microbial growth yields are lower than for aerobic reactions.

If a large quantity of a readily biodegradable organic pollutant is added to a soil, it is not uncommon for the dominant, terminal electron-accepting reaction to progress from aerobic respiration (with O_2 as the primary electron acceptor) all the way to methanogenesis (with CO_2 as the primary acceptor), passing through each dominant electron acceptor in the order shown in Fig. 2. Some microorganisms (e.g., facultative anaerobes) can utilize more than one electron acceptor, although rarely are these reactions performed simultaneously by a given organism. For example, many denitrifying bacteria grow aerobically in the presence of oxygen, then anaerobically on nitrate when oxygen levels are reduced. Other microorganisms, including many sulfate-reducing bacteria, are obligate anaerobes and are capable of using only one of the common electron acceptors (i.e., sulfate) for growth. As electron acceptors are depleted in sequence, the dominant metabolic reactions (and in many instances, the dominant microorganisms) accounting for the chemical losses shift. The reduction potential declines accordingly as the sequence of terminal electron

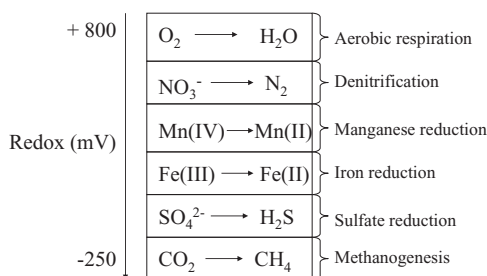


Fig. 2 Sequence of common electron-accepting reactions. Reactions are given in general order of occurrence from top to bottom, with the reactions at the top taking place at the highest redox potential and providing the highest energy yield to microorganisms during degradation of a given organic compound.

acceptors progresses from oxygen to carbon dioxide. In the absence of available external electron acceptors, some organic molecules are also subject to fermentation, whereby small amounts of energy are gained by an organism through internal oxidation-reduction reactions. However, when all common electron acceptors are consumed in a contaminated soil, the biodegradation reaction is likely to cease, even though biodegradable compounds are still present. It is for this reason that oxygen is frequently added via sparging or other means to enhance biodegradation in environments that have experienced heavy contamination with organic pollutants. It may also be necessary to add inorganic fertilizers in such instances, as the supply of macro- and micronutrients in the soil may be inadequate to sustain microbial degradative activity.

Some groups of chlorinated pollutants, including polychlorinated biphenyls (PCBs), chlorophenols, chlorinated solvents, and certain organochlorine pesticides can serve as alternate electron acceptors, by a process known as reductive dechlorination (also called dehalorespiration) (Fincker and Spormann, 2017). This process, which occurs only in the absence of oxygen, is a critical step in the degradation of these persistent pollutants. During reductive dechlorination, bacteria transfer electrons from a growth substrate (primarily hydrogen) to the chlorinated compounds, causing the release of chloride ion and the production of various organic daughter products. For example, during reductive dechlorination of perchloroethylene (PCE), the daughter products are TCE, *cis*-1,2-DCE, vinyl chloride (VC), and ethylene, respectively (Fig. 3). The occurrence of these degradation products (particularly *cis*-1,2-DCE and VC) suggests that reductive dechlorination of PCE or TCE is occurring in an anaerobic environment. A similar reaction has been noted for PCBs and several organochlorine pesticides (e.g., 1,1,1-trichloro-2,2-bis (4-chlorophenyl)ethane (DDT); pentachlorophenol (PCP); and 3,5,6-trichloro-4-amino-2-pyridinecarboxylic acid (picloram)) under strongly reducing conditions. In soils, reductive dechlorination of PCBs is observed as a shift in the representative PCB congeners found within an environment (there are 209 PCB congeners based on possible positions and numbers of chlorine atoms on the biphenyl molecule) from more chlorinated to less chlorinated congeners. Other chemical pollutants such as perchlorate (ClO_4^- ; an oxidant used in solid rocket fuel) and hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX; an explosive), have also been observed to serve as terminal electron acceptors in soils.

Chemical properties affecting biodegradation

All natural organic compounds are biodegradable. The occurrence and rates of biodegradation of these compounds vary widely depending on environmental conditions and chemical complexity (e.g., lignin biodegradation is extremely slow, even under ideal conditions), but it is generally believed that no natural organic materials are completely resistant to microbial destruction. Given hundreds of millions of years of evolution and the constant competition among microorganisms for limiting nutrients and energy, it is not surprising that microbial enzymes have naturally coevolved in step with new organic compounds. In contrast to nature, however, rapid advances in organic chemistry have allowed mankind to develop and produce an overwhelming array of new organic compounds in a short time period. A majority of these thousands of new compounds have been introduced to the Earth's

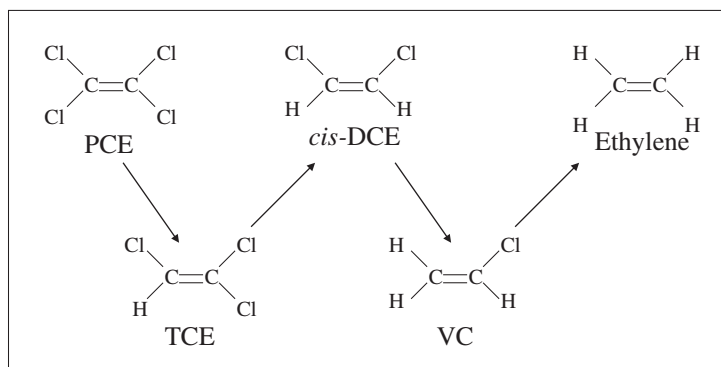


Fig. 3 Biodegradation pathway of perchloroethylene (PCE) during reductive dechlorination. *Cis*-DCE, *cis*-1,2-dichloroethylene; VC, vinyl chloride.

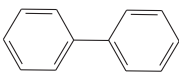
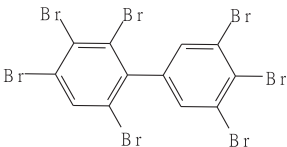
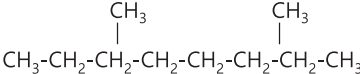
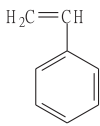
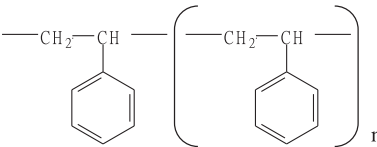
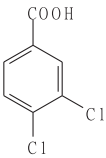
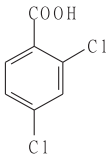
Readily Biodegradable	Poorly Biodegradable	Structural Difference
 Biphenyl	 Polybrominated biphenyl	Halogenation
$\text{CH}_3(\text{CH}_2)_6\text{CH}_3$ Octane	 3,6-Dimethyloctane	Branching
 Styrene	 Polystyrene (MW > 500)	Molecular size
 3,4-Dichlorobenzoic acid	 2,4-Dichlorobenzoic acid	Substituent position

Fig. 4 Influence of molecular structure on contaminant biodegradability.

environment for the first time since the mid-1940s. Many of these chemicals, particularly those that are similar in structure to naturally occurring organics, have proven to be readily biodegradable. Other, more novel synthetics, however, are highly resistant to microbial attack.

Some of the key factors that influence the degradability of a molecule include the type, number, and location of molecular substituents, the type of bonds, the degree of branching, and the molecular weight of the compound (Fig. 4) (Alexander, 1999). The addition of new organic or inorganic substituents to a common molecular structure is one means of creating novel synthetic compounds. Small differences in the type and location of these substituents can determine whether a molecule is readily biodegradable or highly persistent. Contaminants that are resistant to biodegradation owing to their chemical structure are termed “recalcitrant” molecules. A number of synthetic organic compounds and pesticides have been developed by adding halogens (particularly chlorine, fluorine, and bromine) to organic compounds. Because the halogen-carbon bond is very stable and resistant to microbial attack, this type of substitution tends to make a molecule recalcitrant, with increasing halogenation leading to increasing persistence. For example, organochlorine pesticides, such as DDT, toxaphene, 2,4,5-trichlorophenoxyacetic acid (2,4,5-T), pentachlorophenol, chlordane, aldrin, hexachlor, and methoxychlor, were once widely used. Many of these pesticides have now been banned from use in the United States and most European countries due to their long-term persistence and negative impacts in various ecosystems. These molecules proved to be highly resistant to aerobic biodegradation and persistent or very slowly degraded under anaerobic conditions. Halogenated pollutants such as PCBs, polybrominated biphenyls (PBBs), freon propellants and refrigerants (e.g., CCl_3F , CCl_2F_2), and chlorinated solvents (e.g., PCE, carbon tetrachloride) have also proven to be persistent contaminants.

In recent years, there has been growing concern over the long-term persistence and toxicity of per- and polyfluoroalkyl substances (PFAS), which represent a broad family of highly fluorinated molecules that have been used in many different products, including fire-fighting foams, stain-resistant fabrics, non-stick cookware, food packaging, electronics manufacture and electroplating, as well as numerous consumer products (Kwiatkowski et al., 2020). PFAS include more than 5000 different chemicals, the most commonly studied of which are perfluorooctanoic acid (PFOA) and perfluorooctane sulfonate (PFOS). The widespread use of aqueous film forming foam (AFFF) in fire training and firefighting has led to the most significant soil and groundwater contamination with PFAS, some of which are considered a risk at ng/L or ng/kg concentrations in environmental media. PFAS are extremely recalcitrant to mineralization and have been sometimes termed “forever chemicals”, because many have a fully or nearly fully fluorinated C_4 to C_{14} carbon backbone that is highly resistant to enzymatic attack. That being said, recent studies suggest that an

anaerobic iron-reducing bacterium (*Acidimicrobium* strain A6), which typically grows under acidic conditions using ammonium as an electron donor, is capable of substantially defluorinating some PFAS and cleaving the carbon backbone (Huang and Jaffe, 2019). In that the organism does not seem to gain carbon or energy from this reaction, it would be considered an anaerobic cometabolic process. Most importantly, however, this work suggests that biotransformation may play a role in PFAS fate and treatment in the future.

Other molecular traits that may decrease a chemical's biodegradability include high molecular weight, low water solubility, and extensive branching. As molecular weight increases to more than approximately 500–600 g per mole (g mol^{-1}), chemical biodegradability tends to decrease markedly. This trend reflects, in part, the inability of large molecules to enter the microbial cell. In some instances, extracellular enzymes produced by bacteria or fungi can initiate an attack on a large molecule, breaking it into smaller products that can then be metabolized further. For example, it is large size that prevents the biodegradation of many organic polymers that constitute plastics, such as polyethylene and polystyrene. If reduced to a molecular weight of less than 500 g mol^{-1} from their typical molecular weights of a few thousand to more than $100,000 \text{ g mol}^{-1}$, many of these polymer structures are biodegradable. Molecular branching also tends to reduce chemical biodegradability because molecular branches tend to sterically interfere with the common degradative enzymes used by microorganisms. For example, many alkanes are transformed by bacteria through the sequential oxidation and cleavage of two carbon molecules at a time, in a process known as "beta-oxidation". During this process, the alkane is sequentially oxidized (usually at the terminal carbon) to an alcohol, an aldehyde, and then an aliphatic acid. The acid is then cleaved to form acetyl-CoA, which the cell can utilize for energy. This beta-oxidation process continues until the alkane is completely degraded to acetyl-CoA. Branching of the alkane, particularly branching with one or more methyl groups, can block the two-carbon cleavage of beta-oxidation, causing an otherwise biodegradable molecule to become recalcitrant. For this reason, branched alkanes, such as pristane and phytane, tend to be among the slowest molecules to biodegrade during a crude oil spill. These molecules are frequently employed as internal standards to assess the biodegradation of other, more labile, components of crude oil after an environmental release.

Environmental factors affecting biodegradation

A compound that is readily metabolized in a laboratory may escape breakdown and persist unexpectedly in the field, even if an organism capable of causing its degradation is present. Unless pollutants are accessible to microorganisms and environmental conditions are conducive to microbial growth, contaminants will not undergo biotransformation and may remain in soil for extended periods. Thus, the extent of chemical biodegradation will depend in large part upon the physical, chemical, and biological properties of a contaminated site. A number of important factors influencing the growth of microorganisms and therefore biodegradation are described below.

Bioavailability

Contaminants released into soil will not be biodegraded if they are biologically unavailable. Even the most labile compound may persist if it does not come into contact with microorganisms. Bioavailability of a compound can be reduced if the compound is sorbed to soil particles, entrapped within the soil matrix, or dissolved in a nonaqueous solvent (Fig. 5).

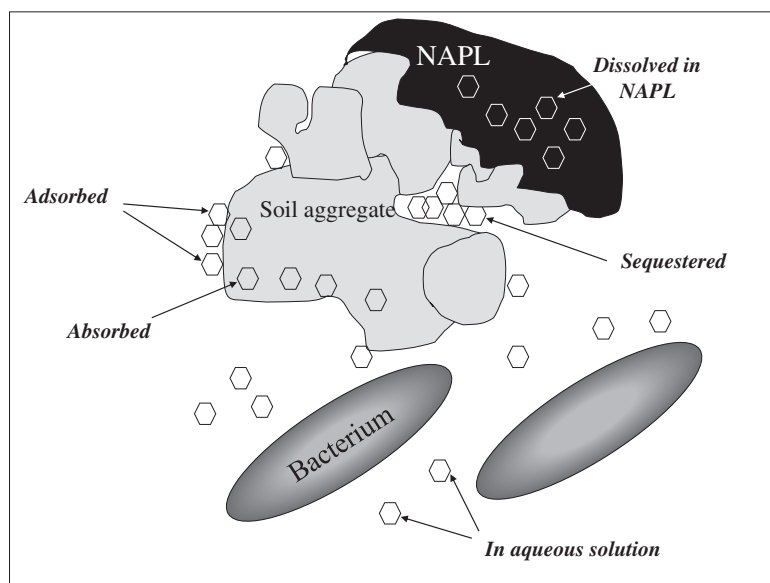


Fig. 5 Representation of different potential states of an organic pollutant (cyclohexane) in a soil environment. (Not drawn to scale.) NAPL, nonaqueous-phase liquid.

Sorption

Chemical compounds can bind to solid surfaces through the actions of several attractive forces (Schwarzenbach et al., 2016). In some instances, a compound will adhere to the outer surfaces of soil particles in a process known as “adsorption.” “Absorption” refers to retention within the interior matrix of a solid particle. In either case, a compound partitions from a liquid phase (usually from water) onto or into a solid phase. Typically, the nonspecific term “sorption” is used to describe this phenomenon. Since compounds must generally be in an aqueous phase and enter the microbial cell to be subject to biodegradation, sorption often increases the resistance of a contaminant to biological attack. After a given period in a soil, a pollutant establishes an equilibrium between the sorbed and the aqueous phase. The extent of sorption at equilibrium, as quantified by the pollutant’s sorption coefficient (K_d), reflects the characteristics of both the pollutant and the soil matrix. When comparing pollutants of a single structural class (e.g., polycyclic aromatic hydrocarbons, PAHs), those molecules with higher K_d values, which tend to be primarily present in soils in the sorbed phase, generally biodegrade more slowly than those with lower K_d values. To enhance contaminant desorption, some bacteria and fungi have been observed to produce extracellular biosurfactants. The release of these surfactant molecules can increase the rate and extent of pollutant desorption and, subsequently, accelerate contaminant biodegradation.

Clay minerals and soil organic matter contain the principal sites to which contaminants can sorb. Charged compounds such as cationic pesticides (e.g., diquat) are often retained by clay minerals. Nonpolar organic molecules, on the other hand, tend to associate with soil organic matter through both adsorption and absorption. For example, PAHs such as phenanthrene, anthracene, and pyrene are subject to sorption and diminished availability due to their interactions with humic materials. In some instances, strong adsorbents such as biochars, activated carbon compounds, and clays are specifically added to soils and other media to enhance the sorption and subsequently reduce the bioavailability of difficult-to-biodegrade organic compounds, such as PFAS.

Entrapment within the soil matrix

Extended residence time in soil, or aging, can cause pollutants to undergo a time-dependent decline in bioavailability as they slowly move from positions within the soil that are accessible to organisms into positions that are inaccessible (Ren et al., 2018). Nonpolar organic pollutants, for example, can diffuse deeply into hydrophobic regions of soil organic matter and, as a result, evade microbial cells that might otherwise degrade them. Other compounds may diffuse into soil nanopores, thus becoming physically protected from metabolism by microorganisms. Although the details of the mechanisms involved vary, a compound that has moved into remote sites is said to be “sequestered” in soil.

The degree to which sequestration occurs depends upon a number of variables, including the properties of the soil and the contaminant involved. Studies continue to define the soil properties most responsible for the entrapment of contaminants; however, soil organic matter content appears to be positively correlated with the extent of chemical sequestration. Likely, soil structure also plays a key role in the sequestration of some contaminants. Contaminants may be more protected from biodegradation in soils possessing highly developed aggregates than in less structured soils. The chemistry of a contaminant is clearly critical as well. Analyses suggest an inverse relationship between a compound’s water solubility and its potential for sequestration, but the relative importance of this and other properties remains largely unknown. Pollutants for which sequestration has been observed include the PAHs phenanthrene and anthracene, and various pesticides, such as 1,2-dibromoethane (EDB), DDT, and chlordane, among others.

Nonaqueous-phase liquids

Many relevant environmental contaminants are present in soils dissolved in non-aqueous solvents. For example, crude oil is a nonaqueous-phase liquid (NAPL) composed of multiple compounds, most of which have very low water solubility. Since microorganisms are typically active in the aqueous phase only, pollutants in NAPLs can escape biodegradation. The extent to which an individual compound dissolved in a NAPL will be degraded depends upon numerous factors, including the rate of partitioning between the nonaqueous and aqueous phases, the equilibrium concentration of the pollutants in the aqueous phase, the nature of the interface between the two phases, and the toxicity of the NAPL to the microorganisms responsible for degradation. Although individual compounds within a NAPL tend to be degraded at different rates, in general, pollutants in a nonaqueous phase can be extremely persistent.

Nutrient availability

Heterotrophic microorganisms in soil require a number of macro- and micronutrients, an energy source, and an electron acceptor to grow (Madigan et al., 2018). Often, a contaminant provides a carbon or energy source, but all other requirements must be fulfilled if biodegradation is to occur. A biodegradable compound released into a nutrient-poor soil (e.g., little available nitrogen or phosphorus) can persist unexpectedly, because microorganisms that possess the capacity to metabolize the contaminant cannot grow to sufficient cell numbers. In some instances, as during the biodegradation of the explosives 2,4,6-trinitrotoluene (TNT) and RDX under nitrogen-limiting conditions, specific organisms can degrade an organic compound to obtain a scarce noncarbon nutrient (in this case, nitrogen). Some organic phosphorus-containing compounds, such as triethylphosphate (TEP), can also be utilized by microorganisms to obtain phosphorus under conditions in which the compound is unavailable for microbial growth.

During bioremediation of soil contaminants, the addition of nitrogen, phosphorus, or other limiting nutrients is often conducted to stimulate the growth of indigenous or added microorganisms.

pH

Some bacteria, fungi, and particularly archaea can survive and grow under very acidic or alkaline conditions, but the optimum pH for most soil microorganisms is near neutrality. Unless organisms have evolved to live under extreme pH conditions (e.g., thermophilic or acidophilic bacteria or archaea), both high and low pH can lead to the denaturing of microbial enzymes and the hydrolysis of other crucial cell components. It is therefore not surprising that biodegradation tends to be fastest near pH 7 and slows considerably as the pH becomes appreciably more alkaline or acidic. In addition to its direct effect on cells, pH will also influence the chemistry of ionizable contaminants in soil. The sorption and subsequent bioavailability of cationic pesticides (e.g., paraquat, diquat), basic pesticides (e.g., atrazine), and acidic pesticides (e.g., 2,4-D, dicamba) will depend in part on the pH of the soil solution in which they are present. To enhance microbial growth and biodegradation, bioremediation protocols sometimes include chemical treatment of the contaminated soil (e.g., by lime addition) to adjust pH to the optimum for the biodegradative organisms.

Temperature

The growth of microorganisms is highly dependent on temperature, and the prevailing temperature in soil is an important factor controlling the rate and extent to which a contaminant is biodegraded. Most microorganisms in soil are mesophilic, with temperature optima ranging from approximately 25 °C to 35 °C, and growth possible from 15 °C to 45 °C. Therefore, optimal temperatures for biodegradation in soils tend to be within the best growth range for these organisms, with rates increasing as temperatures increase within this range. However, it should not be implied that biodegradation processes cease below 15 °C or above 45 °C. There are many bacteria and archaea that are psychrophilic, with growth optima less than 15 °C, and those that are thermophilic, with growth optima exceeding 40 °C and sometimes with extremes as high as 65 °C. These organisms can be responsible for contaminant biodegradation in extreme environments such as in polar climates (for psychrophiles) or in compost (for thermophiles). In fact, thermophiles have been applied in numerous *ex situ* bioremediation applications due to their ability to rapidly degrade some contaminants. Please see later section entitled *Climate Change and Biodegradation* for additional details.

Water content

The moisture content of soil influences the growth and metabolism of microorganisms. As with the other abiotic factors discussed above, tolerance to desiccation varies among organisms. Environments characterized by low water content tend to support only limited microbial activity. The optimal activity of most microorganisms occurs at soil moistures ranging from 60% to 80% of the soil water-holding capacity (WHC). Biodegradation can be very limited in dry soils, and pollutants may persist in arid or drought-affected regions longer than in wet areas. Conversely, excessive moisture can limit oxygen diffusion in soil and cause a shift from aerobic to anaerobic conditions. As previously noted, the absence of available oxygen can severely inhibit the disappearance of some pollutants and pesticides, although such conditions may be necessary to promote the degradation of other compounds (e.g., by reductive dechlorination).

Presence of appropriate microorganisms

Microorganisms capable of performing a given biodegradation reaction are frequently indigenous in soil. For example, bacteria that aerobically degrade the common gasoline components benzene, toluene, ethylbenzene, and xylenes (BTEX) are ubiquitous in most surface soils. If these contaminants persist for long times in a soil, it is likely that environmental factors such as nutrient availability, oxygen levels, contaminant availability, or the presence of toxic co-contaminants are impeding the biodegradation process. For some contaminants, however, biodegradative populations are less universally distributed or are present only at very low densities. For example, the gasoline oxygenate methyl *tert*-butyl ether (MTBE) has been observed to biodegrade readily in some aerobic soils and aquifers but to be highly persistent in others. In environments where the contaminant persists, the absence of a biodegrading population has been cited as a key factor limiting its destruction. Bioaugmentation has been used at some sites to promote MTBE biodegradation. Similarly, the absence of a specific dehalorespiring microorganism (*Dehalococcoides mccartyi*) has been cited as the key factor determining the extent of biodegradation of PCE and TCE in some anaerobic environments (see Fig. 3). If this organism is absent, dechlorination of PCE (or TCE) appears to stall at *cis*-DCE or VC, rather than progressing all the way to ethylene. Bioaugmentation has now been used to add this organism to thousands of sites to enhance chlorinated solvent biodegradation.

Predation

Bacteria are subject to predation by protozoa. Such grazing can influence the biodegradation of soil contaminants, with both inhibition and stimulation possible. Predation by soil protozoa can reduce the number of cells metabolizing a contaminant and cause a decline in the rate at which the compound disappears. Alternatively, due to the positive effect predation has on nutrient availability, grazing may stimulate the growth of microorganisms. Through their metabolism of microbial cells, protozoa release available nitrogen, phosphorus, and other nutrients back into soil and, consequently, stimulate the surviving microbial cells to grow

faster. The dynamics of predator-prey relationships in soil are complicated and are not completely understood. Whether predation decreases or increases biodegradation rates depends upon a number of factors, including the starting density of the bacterial prey and predator organisms, the size of the cells, and the properties of the soil in which the organisms live.

Presence of other compounds

Microorganisms metabolizing a pollutant can be affected by the presence of other soil contaminants. In particular, it is not uncommon for a co-contaminant (such as a heavy metal or low molecular weight hydrocarbon) to be present in toxic concentrations in a heavily polluted soil, thereby inhibiting the biodegradation of a target compound by killing the biodegradative population. An organism may also preferentially use a compound other than the contaminant as a carbon or energy source and, as a result, biodegradation of the selected contaminant can be slowed. Alternatively, the presence of a second compound can stimulate microbial activity and accelerate biodegradation. For example, bioremediation of a compound that is cometabolized can be enhanced by the addition of a substrate on which the cometabolizing microorganism can grow. The addition of biphenyl to soils has been observed to enhance the rate of degradation of some of the less-chlorinated PCBs. In some instances, however, the growth substrate will actually inhibit degradation of the cometabolized compound owing to competitive enzyme inhibition. The presence of toluene above a given concentration, for instance, can reduce the rates of TCE cometabolism by specific toluene-oxidizing strains. For bioremediation applications, intermittent addition of the growth substrate may be necessary to promote degradation of a cometabolized pollutant.

Plants

Plants play a critical role in the biodegradation of organic contaminants (Azaizeh et al., 2011). Growing plants cause many changes to soil structure, chemistry, and microbiology. Plants with extensive root networks disrupt soil aggregates, creating networks for penetration of oxygen and water, as well as new surfaces for microbial growth. Perhaps most importantly for biodegradation and bioremediation, plant roots excrete a wide variety of compounds that include enzymes, surfactants, growth factors, amino acids, carbohydrates, and carboxylic acids, among others, which are known collectively as “root exudates”. The plant root environment or “rhizosphere” typically harbors a larger and much different microbial community than the surrounding soil matrix, and one which can have very different capabilities in terms of contaminant biodegradation. A wide variety of contaminants including chlorinated solvents, chlorinated phenols, herbicides, petroleum hydrocarbons, polycyclic aromatic hydrocarbons, polychlorinated biphenyls, and explosives have all shown enhanced rates of biodegradation in rhizosphere compared to non-rhizosphere soils. These effects may be related to one or more factors including higher oxygen, carbon and inorganic nutrient levels, presence of primary organic substrates for cometabolism, greater density or diversity of degrading organisms, enhanced solubilization and bioavailability of the contaminant (e.g., due to surfactant effects), and changes in pH or soil water content. The observation of enhanced biodegradation in the rhizosphere combined with plant uptake of various contaminants (with or without enhanced degradation by plant enzymes) has led to the development of phytoremediation as a soil and water remediation technology over the past few decades. This has progressed to the development of transgenic plants that express microbial enzymes capable of biotransforming contaminants upon plant uptake.

Potential impacts of climate change on biodegradation

Decades of research have linked anthropogenic releases of greenhouse gases to several phenomena, notably, a steady increase in Earth’s average temperature since the late 19th Century, shifting climates, rising sea levels, and decreasing ocean pH. Some of the likely consequences of these changes for both natural and anthropogenic systems, including accelerated extinctions, redistribution of precipitation, and loss in soil fertility and food production, have been studied extensively; on the other hand, potential effects on biodegradation and bioremediation are not as well understood.

It is probable that global warming will influence rates and extents of contaminant biodegradation as well as bioremediation success at contaminated sites. Simply put, if the bacteria, archaea or fungi, possessing the ability to metabolize synthetic pollutants, grow successfully at a particular location, then biodegradation will proceed, perhaps even at accelerated rates; if the relevant microorganisms are inhibited by environmental changes, those contaminants will persist. The integration of natural systems presents challenges when predicting whether clean-up will be enhanced or impeded by altered conditions. However, many of the factors influencing biodegradation processes discussed in the previous section (e.g., temperature, pH, water content) may be influenced by climate change, and thus, may impact biodegradation and bioremediation in the future. Here we briefly summarize some important considerations.

Changes in pH

As noted above, soil microorganisms are most effective at their optimum pH. Higher levels of atmospheric carbon dioxide drive diffusion of this greenhouse gas into all of Earth’s water, including that in soil, and lowers its pH. In addition to the possible changes to microbial community structure that would follow, recalcitrant pollutants such as TCE could become even more persistent because of the inverse relationship between acidity and the rate of biological dichlorination reactions of such compounds.

Changes in Earth's temperature

Many models predict that the average temperature on Earth will increase by 1–2 °C before the year 2100, and in some areas, an increase in 1.5 °C has already occurred compared to pre-industrial times (Buis, 2019). The global climate system is quite complicated, however, so changes are not expected to be uniform. Indeed, although nearly all areas will experience appreciably warmer conditions, a few could become cooler. Our focus will be on the effects of higher temperature.

- *Increased reaction rates.* Like all chemical reactions, microbial transformations are positively correlated with temperature. If it did not interact with any other environmental variables, modest warming would, in principle, speed up the clean-up of contaminated sites (although it is unclear if a rise of 1–2 °C will have a substantive impact—those areas affected by more dramatic warming could see appreciable increases). Since temperature changes clearly do *not* occur in isolation, predictions of biodegradation rates must account for mutual influences among multiple variables (some of the most critical are described below).
- *Oxygen availability.* As described earlier, the biodegradation of many important soil pollutants depends on the actions of aerobic organisms, thus reduced O₂ concentration is of great concern. Elevated temperatures may diminish the amount of oxygen present in some systems for several reasons. First, because dissolved oxygen (DO) is inversely proportional to water temperature, soil water could become more hostile to oxygen-using bacteria and fungi. Second, warming will drive the melting of high-latitude permafrost and the release of sequestered soil organic matter (SOM); demand for oxygen will rise as more and more SOM is accessible for decomposition, potentially slowing degradation of synthetic pollutants that undergo aerobic biodegradation while accelerating rates for those that are subject to anaerobic degradation processes. Third, over time, aerobic microorganisms that both dominate their soil ecosystems and possess the capacity to metabolize pollutants could be replaced by species adapted to low-oxygen conditions, again shifting the balance of degradative processes from aerobic to anaerobic. We will return to this and other ecological-based considerations shortly.
- *Water availability.* Increased average temperatures have led to more net precipitation over terrestrial areas, although there is a great deal of variability with location. Some, like southern California (United States) have seen declining precipitation since 1901, whereas portions of upper mid-western and northeastern US states have received more rainfall during the same period (USEPA, 2016). Much evidence suggests that plant and microbial production would increase with modestly higher water availability; the problem is complex, because higher rates of evaporative losses with warming could effectively drive out any extra soil water. If water availability in soil decreases, desiccation-resistant organisms without the ability to metabolize pollutants could begin to dominate ecosystems. Alternatively, as described above, saturated soils that become anoxic will not support aerobic metabolism and very likely will slow the rate of aerobic degradation of pollutants, while potentially accelerating rates for those that undergo anaerobic biodegradation. Whether alterations in soil moisture content will be large enough to initiate relevant changes remains an open question.
- *Sea level rise.* Due to thermal expansion of sea water and melting of land-based glaciers, sea levels are expected to rise by between 0.5 and 0.8 m in the next 80 years (Oppenheimer et al., 2019). Clearly, high-elevation terrestrial areas will not be directly impacted by these changes. However, since human populations are disproportionately concentrated in coastal areas, the inundation of soil in low-lying areas with sea water could lead to profound environmental effects and disruption of biodegradation processes.
- *Changes to community structure.* Organisms dominating a soil ecosystem may lose their advantages as environmental conditions change. The problem is a nuanced one, though, because not all species are equally resilient. The effects of the resultant changes on microbial interactions (e.g., predation and competition) on biodegradation clearly will vary, and, like the concerns described above, are difficult to predict. Even without a change in the identity of species in a community, the activities of existing organisms will surely be affected.
- *Disruption of nutrient cycling.* Microbial activity is critical to the maintenance of soil ecosystems. For example, fixation, predation, and decomposition increase the biological availability of vital nutrients such as carbon and nitrogen. Changes in environmental conditions could put stress on the communities responsible for these processes and limit the growth of organisms carrying out biodegradation.
- *Changes in plant-microbe interactions.* The relationships between plants and microorganisms dictate success or failure of the participating organisms, affect soil fertility, and, as noted above, influence pollutant fate. A complex matrix of interacting variables limits our ability to generalize likely outcomes, but it is reasonable to expect that disruptions in mutualism, commensalism, competition, amensalism, parasitism, and predation could have appreciable effects on both microbial- and phytoremediation.

Summary

Biodegradation is an important process by which chemical contaminants may be removed from soil. Microorganisms have a remarkable ability to metabolize an extremely diverse range of synthetic compounds and can be invaluable in the remediation of contaminated environments. The rate and extent of biodegradation of a given contaminant in soil depends on a number of variables, including chemical structure, contaminant bioavailability, and an array of environmental factors, such as nutrient availability, pH, temperature, soil moisture, presence or absence of plants, predation, and the types and concentrations of

co-contaminants. Impacts of climate change on biodegradation processes are hard to predict, but increasing temperatures are likely to impact many different variables, including water, nutrient, and oxygen availability, which can subsequently influence rates, extents, and routes of pollutant biodegradation.

References

- Alexander M (1999) *Biodegradation and Bioremediation*, 2nd edn, New York: Academic Press.
- Azaizah H, Castro PML, and Kidd P (2011) Biodegradation of organic xenobiotic pollutants in the rhizosphere. In: Schroeder P and Collins CD (eds.) *Organic Xenobiotics and Plants*, pp. 191–215. New York: Springer Science and Business Media.
- Buis A (2019) A degree of concern: Why global temperatures matter. <https://climate.nasa.gov/news/2865/a-degree-of-concern-why-global-temperatures-matter/>.
- Fincker M and Spormann AM (2017) Biochemistry of catabolic reductive dehalogenation. *Annual Review of Microbiology* 86: 357–386.
- Fischer K and Majewsky M (2014) Cometabolic degradation of organic wastewater micropollutants by activated sludge and sludge inherent microorganisms. *Applied Microbiology and Biotechnology* 98: 6583–6597.
- Frascari D, Zanolli G, and Danko AS (2015) In situ aerobic cometabolism of chlorinated solvents: A review. *Journal of Hazardous Materials* 283: 382–399.
- Huang S and Jaffe PR (2019) Defluorination of perfluorooctanoic acid (PFOA) and perfluorooctane sulfonate (PFOS) by *Acidimicrobium* sp. strain A6. *Environmental Science & Technology* 53: 11410–11419.
- Kwiatkowski CF, Andrews DQ, Birnbaum LS, et al. (2020) Scientific basis for managing PFAS as a chemical class. *Environmental Science & Technology Letters* 7(8): 532–543.
- Madigan MT, Bender KS, Buckley DH, Sattley WM, and Stahl DA (2018) *Brock Biology of Microorganisms*, 15th edn, London: Pearson.
- Oppenheimer M., Glavovic, B., Hinkel, J. et al. (2019). Sea level rise and implications for low-lying islands, coasts and communities. In Pörtner, H.O., Roberts, D.C., Masson-Delmotte, V. et al. (eds) IPCC Special Report on the Ocean and Cryosphere in a Changing Climate. pp 321–445. https://www.ipcc.ch/site/assets/uploads/sites/3/2019/11/08_SROCC_Ch04_FINAL.pdf.
- Ren X, Zeng G, Tang L, et al. (2018) Sorption, transport and biodegradation—An insight into bioavailability of persistent organic pollutants in soil. *Science of the Total Environment* 610–611: 1154–1163.
- Schwarzenbach RP, Gschwend PM, and Imboden DM (2016) *Environmental Organic Chemistry*, 3rd edn, Hoboken, NJ: Wiley.
- USEPA (2016) *Climate Change Indicators in the United States, 2016*, 4th edn, United States Environmental Protection Agency. EPA 430-R-16-004 www.epa.gov/climate-indicators.
- USEPA (2021) *United States Environmental Protection Agency*. TSCA Chemical Substance Inventory. <https://www.epa.gov/tsca-inventory>.

Biological aspects of manure management

Natasha Rayne^a and Bee Khim Chim^b, ^aUniversity of Wisconsin River Falls, River Falls, WI, United States; ^bUniversity of Maine at Presque Isle, Presque Isle, ME, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of J.T. Sims, R.O. Maguire, MANURE MANAGEMENT, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 402–410, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00515-4>.

Introduction	559
Trends in manure production and application	559
Manure composition as affected by storage and treatment	559
Storage	559
Manure treatment	560
Anaerobic digestion	561
Composting	561
Effect of manure on soil conditions and biodiversity	561
Soil organic matter and soil organic carbon	562
Macro- and mesofauna	562
Microbes involved in nitrogen fixation, nitrification, and denitrification	563
Soil borne pathogens and antibiotic resistant genes	564
Summary	565
References	565

Abstract

Manure is associated with many benefits including increased soil microbial diversity and biomass - important indicators of soil conditions. Management of manure has, however, been associated with issues of air and water quality as well as threats to plant, animal, and human health. To mitigate some of these environmental concerns, various treatment strategies involving microorganisms have been developed. Microorganisms and various invertebrates break down organic compounds in manure, transforming them into energy and plant available nutrients. This chapter explores the biology associated with manure management, including its storage, composting and treatment, and emphasizes effect on certain species of soil organisms.

Glossary

Acetogenesis A stage in anaerobic digestion in which acetate is produced through the reduction of CO₂ or organic acids.

Acidogenesis A chemical process and stage in anaerobic digestion where sugars and amino acids are degraded to products such as alcohols, fatty acids, and volatile fatty acids.

Denitrification The process through which nitrate is reduced to gaseous forms of nitrogen under anaerobic conditions.

Methanogenesis A process that involves the reduction of carbon compounds under and is facilitated by archaea under anaerobic conditions.

Nitrification The sequential oxidation of ammonia to nitrite and of nitrite to nitrate.

Prion An abnormally folded protein that can transmit its abnormal folding pattern to normal proteins of the same kind.

Virulence The degree of harm a pathogen may cause to a host.

Key points

- Livestock manure when added to soil can increase microbial species richness and diversity
- Treatment and storage of manure affect the chemical and biological composition of manure
- Manure application may affect soil fauna positively or negatively depending on its effect on carbon, ammonia concentration, metal content, and salinity
- Manure can suppress disease causing pathogens in soil by enhancing pathogen antagonistic microorganisms

Introduction

Since early human civilization, animal manure has been used as a soil amendment. There is evidence of its use for growing crops as far back as 6000 BCE (Bogaard et al., 2013). Manure offers many physical, chemical, and biological benefits to soil and therefore to crop production. The application of manure to soil has been associated with improved hydrological characteristics, improved soil structure, increased organic matter, improved soil fertility, and increased microbial activity. Soil microbial activity and thus microbial processes in the soil have been identified as important indicators of soil condition. A healthy soil is the foundation of agriculture and the need for productive soils will only increase as the demand for food increases. However, intensification and expansion of agriculture around the world has led to the degradation of soils. Soil microbes govern nutrient cycling, break down pollutants, improve soil structure, and mediate organic matter decomposition. Livestock manure, when added to soil, can increase microbial species richness and diversity. These microbial species may play key roles in organic matter decomposition and plant growth promotion. However, there is increasing concern that manure could be a source of antibiotic resistance, which threatens human and animal health. When antibiotics are administered excessively to animals through feed or other means, antimicrobial resistant bacteria and genes can end up in the soil with animal feces. In addition, handling and application of manure has been associated with water and air quality issues. Thus, although manure provides many benefits to the soil, caution must be taken in handling, storage, and land application. Using manure to benefit the soil inevitably requires consideration of ways to mitigate any threats against soil and water quality. This chapter explores the biology associated with various steps in manure handling and treatment including storage, composting, and anaerobic digestion. This chapter also reviews the biological benefits, following manure application to soil, including organic matter, carbon content, and the dynamics of macro- and microorganisms already present in the soil.

Trends in manure production and application

Animal manure has long been used as a nutrient rich soil amendment. The discovery of manure as a nutrient source was likely more by happenstance with early farmers noticing enhanced crop growth in areas where dung had accumulated during the raising of animals. On a global scale, livestock production has increased greatly since the 1960s. In various developed countries, the average livestock density, expressed in the number of animals ha^{-1} has been steadily increasing, a development that has led to surplus plant nutrients on farms. In developing countries animal production is generally not as intensive but, as these countries develop, their production systems will likely become more intensive. The nature of the animal production system directly affects the amount of manure available for use as a soil amendment. The increase and concentrated geographical distribution of concentrated animal feeding operations (CAFOs) has resulted in excess manure nutrients and raised concerns about point and nonpoint source pollution of air, soil, and water. These concerns have increasingly brought manure handling practices to the attention of regulators and advisors. Globally, there has been a decline and stabilization of land applied manure N from dairy and non-dairy cattle since the early 2000s (Fig. 1). This trend is likely a function of the more concentrated production of manure in certain hot spots, with a decline in manure production being experienced in other regions. The type of animal manure applied to land depends on the manure most available in an area. It is a fact that animal production by species differs between regions; for example, sheep and goats are very common in Asia and Africa, whereas chickens are more prevalent in Asia and North America (Table 1).

Manure composition as affected by storage and treatment

The collection, storage, treatment, and land application of manure involves many biological processes, their by-products, and a diverse array of microbial species. Manure is commonly composed of urine, fecal matter and bedding materials thus containing proteins, fats, and carbohydrates. The biochemical composition of manure depends on animal species, animal feeding practices, the housing system, and the manure collection system (Table 2). It is well documented that, during handling, storage, and land application of manure, volatile organic compounds (VOC's), methane (CH_4), ammonia (NH_3), hydrogen sulfide (H_2S), and nitrous oxide (N_2O) are emitted into the atmosphere. The volatilization of ammonia and nitrous oxide emission result in a net loss of nitrogen (N), a nutrient needed in large amounts by crops. Ways to reduce or mitigate some of the threats associated with manure handling are to minimize uncontrolled losses of manure by building proper manure storage structures, treating manure through composting or anaerobic digestion (AD), controlling the timing, amount, method, and placement of manure when applied to land.

Storage

The purpose of manure storage is to contain the manure, process wastewater, and prevent contaminated runoff until it can be safely applied to land. There are various manure storage systems such as stockpiling of solid manure, tank or lagoon storage of liquid manure. Stockpiling is the practice of piling solid manure either inside in a stacking shed or outside, where it ages until it is land applied. Unlike composting, this system of storage is passive and anaerobic. Many of the characteristics associated with raw manure like nutrient concentrations and proportions, odor, and the presence of weed seeds and pathogens remain in stockpiled manure. Nevertheless, anaerobic biological processes can lead to the production of organic acids, acetic acid, methane, and carbon dioxide (CO_2). There are many types of tanks to store manure in liquid form including under-floor pits, fabricated steel or concrete storage

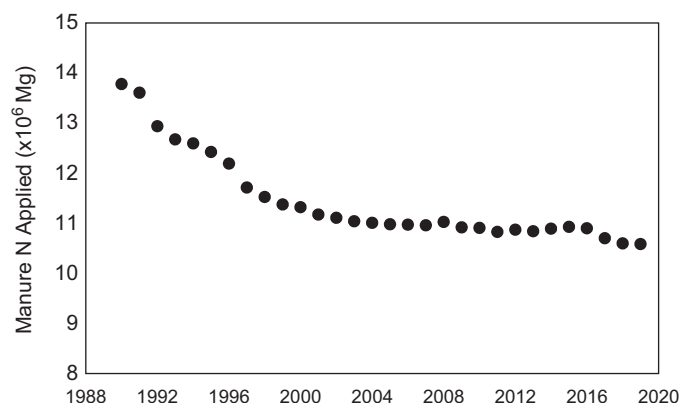


Fig. 1 Global trend in land applied manure nitrogen (N) from 1989 to 2019. Data from the Food and Agriculture Organization of the United Nations.

Table 1 Global animal production by animal type in several geographic regions in 2019.

Geographical region	Animal species					
	Cattle (Millions)	Swine (Millions)	Horses (Millions)	Sheep (Millions)	Goats (Millions)	Chicken (Millions)
Africa	360	44	7	408	476	2,064
Asia	471	429	14	530	574	15,611
Europe	117	186	5	127	16	2,336
North America	106	92	11	6	2	9,348
South America	361	67	12	65	23	2,670

FAOSTAT available at (<https://www.fao.org/faostat/en/#data/EMN>).

Table 2 Nutrients and management details in animal manure as reported in literature.

Manure	Total C (g kg ⁻¹)	Total N (g kg ⁻¹)	Total P (g kg ⁻¹)	Total K (g kg ⁻¹)	C:N ratio (g kg ⁻¹)	References
Cattle	296.0	13.4	2.5	7.7	22	Myint et al. (2009)
Cattle	—	18.3–20.2	4.9–5.8	1.92–2.75	—	Tadesse et al. (2013)
Cattle	181.6	9.3	2.1	1.7	19.53	Khater (2015)
Cattle	—	9.8	10.12	14.22	14.9	Iqbal et al. (2019)
Cattle	—	9.57	11.10	13.83	—	Iqbal et al. (2022)
Poultry	218.5	26.7	36.6	45.4	8	Myint et al. (2009)
Poultry	—	30.3	8.3	23	6.8	Agbede et al. (2017)
Poultry	—	12.65	7.32	9.76	12.98	Iqbal et al. (2019)
Poultry	—	12.98	9.35	10.06	—	Iqbal et al. (2022)
Sheep	—	1.68–1.76	1.32–1.63	1.18–1.33	—	Yang et al. (2021)
Sheep	—	20.3	4.04	31.75	—	Kanton et al. (2016)
Goat	250	26	4	46	—	Mnkeni and Austin (2009)

tanks, ponds, and lagoons. Lagoons are earthen structures that differ from storage tanks or ponds in that they provide treatment in addition to storage. The type of biological processes that occur in the lagoon depend on variations in light, temperature, access to oxygen, and the presence and type of organic matter. As a function of the presence or absence of oxygen, lagoons may contain aerobic, anaerobic, or facultative organisms (Pfoest et al., 2000). Both lagoons and stockpiles for manure storage and processing harbor facultative microbes in that both aerobic and anaerobic conditions exist.

Manure treatment

Treatment of manure involves a deliberate attempt to reduce the mass of the manure, control pathogens, concentrate or reduce nutrients, or produce by-products, such as compost or energy. Treatment of manure thus affects the chemical and biological composition of the manure and consequently how manure or its by-products will affect the soil to which it is applied.

Anaerobic digestion

Anaerobic digestion (AD) is a process through which organic feedstocks, including food wastes, biosolids or sewage sludge, and livestock manure are broken down in the absence of oxygen. The purpose of AD is two-fold; firstly, to produce renewable energy, biogas, thereby reducing the use of fossil fuels and secondly, to decontaminate biodegradable wastes allowing them to be safely used as a soil amendment. Biogas consists primarily of methane and carbon dioxide, but may also contain sulfur compounds, ammonia, volatile organic compounds (VOCs), and particulate matter. Microbial activity in AD systems is influenced by the composition of feedstock and operation conditions such as temperature, pH, and retention time. The efficiency of organic matter conversion to methane decreases with increasing total solids, while initial large concentrations of ammonium might inhibit methane production. Anaerobic digestion consists of hydrolysis, acidogenesis, acetogenesis, and methanogenesis and each one of these phases is carried out by a different group of microbes (Meegoda et al., 2018). Hydrolysis involves extracellular enzymes from hydrolytic bacteria degrading organic material into smaller, more soluble compounds such as lipids, proteins, and soluble carbohydrates. Unlike methanogenic species, which thrive at neutral pH, hydrolytic and acidogenic microorganisms are favored by a more acidic environment. During acidogenesis, products from the hydrolytic processes are fermented to short chain volatile fatty acids (VFAs), such as acetate and propionate in addition to gaseous products, such as, carbon dioxide, hydrogen, and ammonia. In acetogenesis, bacteria known as acid formers, convert products from the hydrolytic process into organic acids, such as acetic acid, butyric acid, propionic acid, ethanol and carbon dioxide and hydrogen. During the next step in the process, methanogenesis, is conducted by a group of archaea called methanogens. The production of methane in AD is thus a complex process driven by various microorganism and impacted by variables that when not adequately controlled can inhibit microbial activity. In agricultural production, anaerobic digestate can be used as soil amendment, fertilizer, and bedding for livestock.

Composting

Composting is different from stockpiling in that it is an actively managed process. Composting is an aerobic microbial process in which organic matter is transformed into a nutrient rich and stable product that can be applied to soil as a fertilizer and soil conditioner. During this process bacteria and fungi break down the organic matter and produce carbon dioxide, water, heat and stabilized organic matter. If composting is done correctly, the final product, compost, should be free of pathogens and viable seeds. Composting is not a continuous homogenous process but is rather variable with changing temperature, pH, oxygen flow, and moisture. Variable growth conditions favor different organisms growing and thriving at the various stages as organic matter breaks down. There are several types of composting, e.g., onsite composting, vermicomposting, aerated windrow composting, aerated static pile composting, and in-vessel composting. Different groups of organisms are involved in composting including macro- and mesofauna and microorganisms. The activity and abundance of microbes in compost is heavily impacted by various physicochemical properties of compost including the manure or bedding material used, water soluble organic carbon, temperature, and moisture. Manure type has been shown to significantly impact the diversity of microbial communities at different stages during composting. When it comes to manure type for example, Wan et al. (2021) found that sheep and cattle manure composts harbor more diverse microbial communities than pig and chicken manure composts, at both the beginning and the end of the composting process. Temperature is arguably the most important physicochemical variable in the composting process. Composting is categorized into three different stages as a function of the changing temperatures. The different phases, in chronological order, are the mesophilic phase, thermophilic phase, and the maturation phase. Various groups of microbes dominate at these different temperature regimes. The mesophilic phase is characterized by rapid growth of microorganisms facilitated by favorable conditions, with temperatures ranging between 25 and 40 °C. The thermophilic phase, with temperatures >40 °C, favors a high rate of decomposition with a rapid growth of thermophilic microorganisms, while the maturation phase includes cooling and stabilization of the process. Although both fungi and bacteria are involved in the decomposition and transformation of compost, bacteria dominate as they have higher thermal tolerance than fungi (Biyada et al., 2021). Both microbial biomass and basal respiration, a parameter for decomposition, decline with compost age. The relative abundance and diversity of bacteria and fungi has been shown to change during composting, with a decline in bacteria from the thermophilic to the maturation phase likely linked to a decline in available nutrients.

Effect of manure on soil conditions and biodiversity

That many soils around the world are degraded is common knowledge. Soil is considered to degrade when its physical, chemical, and biological characteristics deteriorate due to natural or human influences, leading to loss of important soil ecosystem functions. Various practices can be used to mitigate or reverse some impacts related to soil degradation some of which are reduced tillage, retention of crop residue, and the addition of organic soil amendments such as manure. Land application to meet crop nutrient requirements is by far the most widespread and longstanding use of animal manures. In addition to its impact on soil fertility and soil physical characteristics, manure also has a significant impact on biological properties including organic matter and carbon (C) content, and the abundance, diversity of organisms.

Soil organic matter and soil organic carbon

Soil organic matter (SOM) plays an important role in maintaining the ability of soil to provide a host of ecosystem services, including sustainable productivity. Soil organic matter has been associated with different benefits including a decrease in erosion and runoff, improved soil structure, and improved water infiltration. Moreover, there is an intimate relationship between organic matter and soil organisms, as they are the living component of SOM. Increasing SOM or SOC through manure application is a function of different factors such as the source, the percent solids, and bedding material used. Bedding materials in animal production facilities might include sand, sawdust, wood shavings, straw, or compost. Producers select bedding material for their stalls based on comfort and health of the animals, cost, availability, and labor efficiency. Bedding will ultimately affect the composition of the manure, its carbon to nitrogen ratio, its mineralization, and its effect on soil. Manure containing bedding material that has a large organic carbon content may result in a more significant increase in SOM and carbon (C). Additional factors, which will affect the degree of C increase, are weather related and include soil moisture and temperature, but the rate, frequency, and method of application also play a role. High rates of manure applied over several years will result in a greater increase in SOM compared to short term manure application. The carbon contained within SOM is receiving increased attention due to its role in mitigating the effects of climate change. It has been reported that the application of animal manure can potentially increase SOC stock by 32–50% (Gross and Glaser, 2021). However, given the many factors that this depends on it is reasonable to use caution in making estimations about the degree to which manure will increase C levels in soil. Soil organic carbon (SOC) is an important indicator of soil condition as it is also associated with microbial biomass. Soil respiration is an essential part of the carbon cycle and an important indicator of microbial activity in soil, although other processes such as root respiration are also involved in the production of CO₂ in soil. A high C: N ratio in manure and the slow release of nutrients from manure may cause a decrease in CO₂ flux when first applied. However, over time as microbes continue to decompose the manure, increased CO₂ flux might be detected. Although indicative of biological activity, soil is believed to be one of the most species rich habitats among terrestrial ecosystems harboring millions of species of organisms. Species may include archaea, algae, bacteria, fungi, nematodes, and a wide array of fauna.

Macro- and mesofauna

According to estimations, soil animals may represent roughly 23% of the total diversity of living organisms (Decaëns et al., 2006). There is generally a positive relationship between land applied manure and the abundance and diversity of fauna in soil as manure functions as a food and nutrient source for these organisms. However, manure may also have an adverse effect on the abundance and diversity of soil organisms depending on its biochemical composition. Macrofauna, the largest size class of invertebrates with body widths greater than 2 mm, are often thought of as the most familiar group of fauna as they can be seen with the naked eye. This group of fauna includes orders such as isopods, centipedes, millipedes, spiders, beetles, termites, ants, and earthworms with each playing a unique role in organic matter breakdown and nutrient cycling. Earthworms play a role in incorporation of organic material in soil, the formation of soil pores and stable macroaggregates, and greater nutrient availability. An increase of earthworms due to manure application is mainly attributed to the carbon in manure which functions as a food source for the earthworms. However, ammonia concentration, metal content, and salinity are among factors that may affect earthworms negatively. High levels of ammonia in manure are toxic to earthworms and may reduce their population size until conditions become more favorable (Edwards and Bohlen, 1996). However, the ammonia content of manure depends on different factors including, animal type, feed, manure storage and management, and whether the manure is liquid or solid. Poultry manure, for example, contains significantly more ammonia in comparison to beef or dairy manure and therefore would present a greater risk to soil fauna if not managed properly. In addition, long-term application of manure at high rates presents a salinization risk, a condition that generally affects soil fauna adversely. Electrical conductivity is a measure of the ability of soil to transmit an electrical current. Although primarily linked to potassium ions, this variable is used to estimate dissolved salts in soil. An increasing EC with the application of increasing rates of manure is not uncommon. An EC above $>4 \text{ mmho cm}^{-1}$ has been shown to inhibit the growth of most plants and affect many other soil biological factors. Mineral salts are an important component of livestock nutrition. In fact, several health issues including reduced productivity and a reduced appetite are associated with a lack of salt in the animal diet. Feeding large amounts of salt to livestock ultimately results in high amounts of salt in the animal excrements which can buildup in the soil. Similarly, when animal feed contains trace minerals such as cobalt (Co), copper (Cu), iron (Fe), and zinc (Zn) above dietary needs, the potential for heavy metal contamination of soil and water increases. These trace minerals are added as a supplement to animal feed to support proper growth and development of the animals. Heavy metal contamination and high salt concentrations have an inhibitory effect on the growth and abundance of certain species of earthworms and other fauna in soil. Mesofauna, of which acari (mites) and collembolans (springtails) are the predominant organisms, have also been found to be affected significantly by metal concentrations and other soil environmental factors. However, manure can affect variables of the soil environment differently and therefore may have different concurrent effects on organisms. For example, certain species of mites have been shown to be affected by metal pollution although not necessarily negatively when soil humidity and acidity are favorable (Manu et al., 2019).

Manure plays a role in various soil physical characteristics such as soil moisture, temperature, soil structure. These characteristics in turn may affect species richness and diversity of soil organisms. Although conclusions vary, manure application has been generally associated with greater soil water retention although this depends on the manure type and the bedding used. Decreasing soil moisture and droughty conditions have a negative effect on the population composition and numerical abundance of soil

fauna in which survival of the fittest becomes apparent. In general, it is highly likely that under drought conditions, the population of various species will show an abundance of drought resistance species and a reduction in the number of drought sensitive species. The maintenance or increase of soil moisture with the application of manure may thus promote a more balanced population of fauna species as such conditions facilitate the growth of a wider variety of species. However, although there is general agreement that the application of manure no matter the animal type, or the form of the manure whether liquid or solid, increase the abundance and density of fauna, the diversity of manure and manure management might make the extent of the effect somewhat challenging to predict. Importantly, during the treatment of manure specifically during composting, fauna might make their home in the compost pile which upon application can be transferred to the soil. Common fauna present in compost are mites, ants, snails, beetles, and earthworms.

Microbes involved in nitrogen fixation, nitrification, and denitrification

Nitrogen is an essential element for plant growth and is limiting in most ecosystems, although abundantly present in the atmosphere mostly as N gas (N₂). Nitrogen is a component of various compounds in plants and other organisms some of which are chlorophyll, proteins, and amino acids. The N is thus crucial for the productivity and proper functioning of ecosystems. Hence, while manure affects many other microbes and microbial processes (Table 3), the focus in this section is primarily on microbes involved in processes concerned with N.

Specialized prokaryotes can fix atmospheric N and reduce it to ammonia using the enzyme nitrogenase. Biological N fixation contributes approximately 200 million tons of N to the earth's ecosystem annually (Shamseldin, 2022) and could thus represent significant N-fertilizer costs savings. Several prokaryotes involved in N₂-fixation include aquatic organisms such as cyanobacteria, symbiotic diazotrophs such as Rhizobium and free-living non-symbiotic diazotrophs, such as Azotobacter. Long-term application of manure can significantly increase the abundance of soil autotrophic azotobacters relative to untreated soil. Manure containing little C as a food source and N has been shown to support a smaller diversity of soil nitrogenase and N₂-fixing bacteria. Rhizobia form a symbiotic relationship with legumes which results in the formation of nodules on the roots. It is well known that large applications of available N, of which manure is an important source, inhibit biological N fixation. However, application of manure in addition to synthetic fertilizer has been shown to increase the number and dry weight of nodules in soybean (Nagwanshi et al., 2018). This is likely due to the slow release of available N from manure, its large organic C content, and the favorable conditions the manure creates for the rhizobia.

Table 3 Relative abundance of soil bacterial phyla with manure application.

Manure type	Relative abundance		Source
Swine slurry	Phylogenetic Group	%	Rieke et al. (2018)
	Acidobacteria	30–5 ^a	
	Actinobacteria	9–12	
	Bacteroidetes	2–17	
	Proteobacteria	12–39	
Composted manure	Verrucomicrobia	12–6	Peng et al. (2021)
	Acidobacteria	16.5–15.1 ^b	
	Actinobacteria	11.7–14.7	
	Bacteroidetes	8.4–7.8	
	Proteobacteria	26.7–29.6	
Poultry litter	Firmicutes	6.0–5.4	Jangid et al. (2008)
	Acidobacteria	17.7, 18, 19 ^c	
	Bacteroidetes	7.6, 5.8, 7.6	
	Proteobacteria	40.5, 41.9, 31.7	
	β-Proteobacteria	14.4, 16.5, 11.4	
Organic manure	γ-Proteobacteria	6.3, 4.3, 2.7	Tang et al. (2020)
	Acidobacteria	11–14 ^c	
	Actinobacteria	8–17	
	Proteobacteria	8–13	
	α-proteobacteria	8–11	
	Firmicutes	9–14	

^aNumber on left represents the abundance at application (day 0), number on right represents abundance at 108 days after application.

^bNumber on left is abundance at pre-season and number on right represents abundance after manure application. \$numbers represent three agricultural management systems, respectively conventional tilled cropland, hayed pasture, and grazed pasture with poultry manure applied.

^cReported as a range at different of manure relative to chemical fertilizer.

Agricultural soils have been claimed as one of the largest sources of anthropogenic nitrous oxide (N_2O) emission. The main pathway to produce nitrous oxide is through microbial nitrification and denitrification. Soil nitrification, the oxidation of ammonia to nitrate, is driven by ammonia oxidizing bacteria (AOB) and archaea (AOA) and plays an important role in the N cycle. Nitrification is commonly described as a two-step process consisting of the oxidation of ammonia to nitrite and the oxidation of nitrite to nitrate. Manured soils have been found to have significantly greater nitrification rates compared with soil to which no manure is applied. When compared with synthetic fertilizer treatments, manure treated soil may yield a greater abundance of AOA and AOB which can be attributed to the ammonium ions released from manure during decomposition and the favorable soil environment created when manure is applied. However, the sensitivity to NH_4^+ concentrations, can be variable among ammonia oxidizers and large concentrations may even inhibit ammonia oxidation and cause a shift in community composition (Princic et al., 1998). Denitrification is the sequential process where nitrate or nitrite are reduced toward N_2 by a number of groups of denitrifying archaea, bacteria and fungi producing nitric oxide (NO) and N_2O as intermediate products. The efficiency of denitrification is not 100% and it does not always end with the formation of dinitrogen. The reduction of N_2O to N_2 however, is considered a critical step in denitrification as the completion of this process reduces the potential for N_2O emission. Various bacterial species, such as *Pseudomonas* and *Bacillus* among others are involved in this sequential reduction of nitrate and nitrite. This process is encouraged with high nitrate concentrations, high C availability, and under anaerobic conditions. Specific fungal phyla that possess the ability to denitrify are Ascomycota and Basidiomycota. Manure and the application of other organic soil amendments has been associated with an increase in the abundance of nitrate reduction and denitrifying species relative to synthetic fertilizer (Yang et al., 2022). However, when manure has been processed to AD and the viscosity has thus been reduced, fewer anaerobic microsites might contribute to a reduction in N_2O emission relative to farmyard manure (Möller, 2015). Thus, manure also affects microbes involved in nitrification and denitrification by altering soil properties that affect microbial activity, such as, soil aeration, pH, moisture, and temperature. Long-term excessive application of synthetic ammonium-based N fertilizers has been a major cause of soil acidification in agricultural land. The effect of livestock manure on soil acidity is quite variable depending on the biochemical composition and frequency of application. Manure has sometimes been shown to cause a decrease in soil pH due to nitrification of ammonium. However, in long term application of manure an increase in pH can also be achieved due to the presence of ions such calcium, magnesium, and compounds such as calcium carbonate in the manure. Changes in soil acidity in turn, result in changing in the soil microbial community structure. In general, fungi are less sensitive to acidic conditions than bacteria; a phenomenon that could cause a shift in the contribution of these communities to microbial processes. When the soil is acidified, N_2O -producing fungal species are more abundant. These conditions inhibit the activity of bacterial denitrifying reductases, which reduces the contribution of bacteria to N_2O emission (Arikan et al., 2009). Moreover, enriching the soil with organic matter through manure decreases the bulk density and improves soil aeration. Aeration is important in nitrification, and it is therefore no surprise that manure treated soil has been associated with greater rates of nitrification. Indeed, manure functions as a direct food source that increases microbial abundance by altering soil conditions so that they become favorable toward certain species of organisms.

Soil borne pathogens and antibiotic resistant genes

In addition to its effect on beneficial organisms, manure also plays a role in controlling soil borne pathogens. Various disease-causing pathogens which initially reside in intestinal tract of livestock include bacteria, fungi, viruses, nematodes, and protozoa are commonly present in manure. The abundance of these pathogens in manure depends on the animal species, the health of the animal, feed type, and antimicrobial additives in feed and prophylactics. During manure handling and application these pathogens can spread and become a pest to plants and cause diseases in animals and humans. Suppressing these pathogens is thus of utmost importance for animal, plant, soil, and human health. Bacteria, which have been found in manure and linked to outbreaks of human illness include *Escherichia coli*, *Salmonella* and *Campylobacter*. These organisms may transfer virulence factors through plasmids. Commonly associated with untreated manure these pathogens can then be transferred to crops consumed by humans. Pathogens may also end up in ground- and surface water through leaching and runoff after manure is land applied. In addition to pathogens, infected animals, might also excrete prions in their feces. Prions are proteins that cause the abnormal folding of proteins in the affected organism. During storage and treatment of manure pathogens and prions can be greatly reduced. Pre-application of manure treatments that include drying, heating, and the exposure to UV light have been shown to have a negative impact on pathogen density. Certain prions, however, might be able to survive extreme temperatures. Upon land application the duration of the survival of pathogens in manure depends on the environmental conditions pathogens are exposed to such as soil type, pH, moisture, temperature, aeration, UV radiation from sunlight, and competition with other organisms. Jiang et al. (2002) postulated that *E. coli* O157:H7 can survive in manure amended soil for extended periods of time even under dry conditions. Although there is thus a risk for the transfer of pathogens to the environment in manure amended soils, manure may, as previously noted, increase the abundance of native microorganisms in soil which will compete with the invasive pathogens for the same space, energy, and food resources. It is well documented that manure can suppress disease causing pathogens in soil by enhancing the activity and abundance of pathogen antagonistic microorganisms who eventually could outcompete the pathogens. The disease suppressing traits of manure can also be attributed to the effect on soil physicochemical properties such as SOM, pH, EC, nutrient availability, and soil water holding capacity which may favor the non-pathogens over pathogenic microbes. In a study evaluating 'Early Dying' in potato, composted beef manure significantly reduced disease incidence and increased potato yield relative to green manure and chemical treatment (Molina et al., 2014). Although manure can control soil borne pathogens and therefore benefit, plant, animal, and human health, it also contains a serious threat to human health in the form of antibiotic resistant genes (ARGs).

ARGs are genetic modifications that microbial organisms develop in resistance against antibiotics. These ARGs are transferrable between microorganisms in the environment and those in humans thus forming a serious threat to human health. Antibiotics used in livestock farming can produce ARGs in the animal which are then introduced to soil upon manure application. Indeed, antibiotic resistant bacteria (ARBs) and ARGs have become soil pollutants of significant concern. Manure amended soil has been associated with an increased abundance of ARGs, a phenomenon that is further enhanced by the ability of manure to increase the abundance of native soil microorganisms that may already carry ARGs. In fact, studies have shown that long term manure application significantly enhanced the abundance and diversity of ARGs in soil, in contrast with synthetic fertilizer which only shifted the microbial community structure without any significant impact on the abundance of ARGs (Wang et al., 2020).

Summary

Although manure has been used as a soil amendment since early civilization, the knowledge of its effect on soil microbes is its infancy. Manure application can improve and alter various soil biological characteristics including microbial abundance and diversity which in turn may restore some lost ecosystem functions. The degree of these effects, however, depends on many factors including manure composition, manure management, weather conditions, soil characteristics, and their complex interaction with each other. Moreover, the rich diversity of organisms in soil presents a wide range of metabolic and environmental needs, which means that manure with its soil altering properties will have different effects on the many species of organisms. Despite its many beneficial impacts, caution must be exercised when considering the use of manure as a fertilizer or soil amendment since improper handling, treatment and application of manure may release pathogens, excess nutrients, and greenhouse gasses into the environment. Various treatment methods including composting and anaerobic digestion among others have been shown to reduce some of these negative impacts. Despite the appearance of the insurmountable complexity of manure, its management, and interaction with the soil microbiome it is evident that manure has many properties, that benefit soil and the organisms in it. Continued development of handling, treatment, and application methods and technologies are therefore necessary if we are to maximize the benefits and minimize the drawbacks of manure to the soil environment.

References

- Agbede MT, Adekiya AO, and Efeidiyi EK (2017) Impact of poultry manure and NPK fertilizer on soil physical properties and growth and yield of carrot. *Journal of Horticultural Research* 25: 81–88.
- Arikan OA, Mulbry W, and Rice C (2009) Management of antibiotic residues from agricultural sources: Use of composting to reduce chlortetracycline residues in beef manure from treated animals. *Journal of Hazardous Materials* 164: 483–489.
- Bogaard A, Fraser R, Heaton TH, Wallace M, Vaiglova P, Charles M, Jones G, Evershed RP, Styring AK, and Andersen NH (2013) Crop manuring and intensive land management by Europe's first farmers. *Proceedings of the National Academy of Sciences* 110: 12589–12594.
- Byyada S, Merzouki M, Dmčenko T, et al. (2021) Microbial community dynamics in the mesophilic and thermophilic phases of textile waste composting identified through next-generation sequencing. *Scientific Reports* 11: 23624. <https://doi.org/10.1038/s41598-021-03191-1>.
- Decaëns T, Jiménez JJ, Gioia C, Measey GJ, and Lavelle P (2006) The values of soil animals for conservation biology. *European Journal of Soil Biology* 42: 23–38.
- Edwards CA and Bohlen PJ (1996) *Biology and Ecology of Earthworms*, 3rd edn London: Chapman & Hall.
- Gross A and Glaser B (2021) Meta-analysis on how manure application changes soil organic carbon storage. *Scientific Reports* 11: 5516. <https://doi.org/10.1038/s41598-021-82739-7>.
- Iqbal A, He L, Khan A, Wei S, Akhtar K, Ali I, Ullah S, Munsif F, Zhao Q, and Jiang L (2019) Organic manure coupled with inorganic fertilizer: An approach for the sustainable production of rice by improving soil properties and nitrogen use efficiency. *Agronomy* 9: 651. <https://doi.org/10.3390/agronomy9100651>.
- Iqbal A, He L, Ali I, Yuan P, Khan A, Hua Z, Wei S, and Jiang L (2022) Partial substitution of organic fertilizer with chemical fertilizer improves soil biochemical attributes, rice yields, and restores bacterial community diversity in a paddy field. *Frontiers in Plant Science* 13: 895230. <https://doi.org/10.3389/fpls.2022.895230>.
- Jangid M, Williams MA, Franzluebbers AJ, Sanderlin JS, Reeves JH, Jenkins MB, Endale DM, Coleman DC, and Whitman WB (2008) Relative impacts of land-use, management intensity and fertilization upon soil microbial community structure in agricultural systems. *Soil Biology and Biochemistry* 40: 2843–2853. <https://doi.org/10.1016/j.soilbio.2008.07.030>.
- Jiang X, Morgan J, and Doyle MP (2002) Fate of *Escherichia coli* O157:H7 in manure-amended soil. *Applied and Environmental Microbiology* 68(5): 2605–2609. <https://doi.org/10.1128/AEM.68.5.2605-2609.2002>. PMID: 11976144. PMC127522.
- Kanton RAL, Prasad PVW, Mohammed AM, Bidzakin JK, Ansoba EY, Asungre PA, Lamini S, Mahama G, Kusi F, and Sugri I (2016) Organic and inorganic fertilizer effects on the growth and yield of maize in a dry agro-ecology in Northern Ghana. *Journal of Crop Improvement* 30: 1–16. <https://doi.org/10.1080/15427528.2015.1085939>.
- Khater ESG (2015) Some physical and chemical properties of compost. *International Journal of Waste Resources* 5: 1. <https://doi.org/10.4172/2252-5211.1000172>.
- Manu M, Honciuc V, Neagoe A, et al. (2019) Soil mite communities (Acari: Mesostigmata, Oribatida) as bioindicators for environmental conditions from polluted soils. *Scientific Reports* 9: 20250. <https://doi.org/10.1038/s41598-019-56700-8>.
- Meegoda JN, Li B, Patel K, and Wang LB (2018) A Review of the processes, parameters, and optimization of anaerobic digestion. *International Journal of Environmental Research and Public Health* 15: 2224. <https://doi.org/10.3390/ijerph15102224>.
- Mkeni PNS and Austin LM (2009) Fertilizer value of human manure from pilot urine-diversion toilets. *Water* 35: 134–138.
- Molina OI, Tenuta M, El Hadrami A, et al. (2014) Potato early dying and yield responses to compost, green manures, seed meal and chemical treatments. *American Journal of Potato Research* 91: 414–428. <https://doi.org/10.1007/s12230-014-9365-0>.
- Möller K (2015) Effects of anaerobic digestion on soil carbon and nitrogen turnover, N emissions, and soil biological activity. A review. *Agronomy for Sustainable Development* 35: 1021–1041. <https://doi.org/10.1007/s13593-015-0284-3>.
- Myint AK, Yamakawa T, and Zenmyo T (2009) Plant growth, seed yield and apparent nutrient recovery of rice by the application of manure and fertilizer as different nitrogen sources in paddy soils. *Journal of the Faculty of Agriculture, Kyushu University* 54: 329–337.
- Nagwanshi A, Dwivedi AK, Dwivedi BS, and Dwivedi SK (2018) Effect of long term application of fertilizers and manure on leaf area index, nodulation and yield of soybean in a Vertisol. *Journal of Pharmacognosy and Phytochemistry* 7: 1962–1965.

- Peng M, Tabashsum Z, Millner P, Parveen S, and Biswas D (2021) Influence of manure application on the soil bacterial microbiome in integrated crop-livestock farms in Maryland. *Microorganisms* 9: 2586. <https://doi.org/10.3390/microorganisms9122586>.
- Pfost DL, Fulhage CD, and Rastorfer D (2000) *Anaerobic Lagoons for Storage/Treatment of Livestock Manure*. MU Extension: University of Missouri-Columbia.
- Princic A, Mahne II, Megusar F, Paul EA, and Tiedje JM (1998) Effects of pH and oxygen and ammonium concentrations on the community structure of nitrifying bacteria from wastewater. *Applied and Environmental Microbiology* 64: 3584–3590. <https://doi.org/10.1128/AEM.64.10.3584-3590>.
- Rieke EL, Soupir ML, Moorman TB, Yang F, and Howe AC (2018) Temporal dynamics of bacterial communities in soil and leachate water after swine manure application. *Frontiers in Microbiology* 21: 3197. <https://doi.org/10.3389/fmicb.2018.03197>.
- Shamseldin A (2022) Future outlook of transferring biological nitrogen fixation (BNF) to cereals and challenges to retard achieving this dream. *Current Microbiology* 79: 171. <https://doi.org/10.1007/s00284-022-02852-2>. PMID: 35476236.
- Tadesse T, Dechassa N, Bayu W, and Gebeyehu S (2013) Effects of farmyard manure and inorganic fertilizer application on soil physico-chemical properties and nutrient balance in rain-fed lowland rice ecosystem. *American Journal of Plant Sciences* 4: 309–316.
- Tang H, Li C, Xiao X, Shi L, Cheng K, Wen L, and Li W (2020) Effects of short-term manure nitrogen input on soil microbial community structure and diversity in a double-cropping paddy field of Southern China. *Scientific Reports* 10: 13540. <https://doi.org/10.1038/s41598-020-70612-y>.
- Wan J, Wang X, Yang T, Wei Z, Banerjee S, Friman VP, Mei X, Xu Y, and Shen Q (2021) Livestock manure type affects microbial community composition and assembly during composting. *Frontiers in Microbiology* 12: 621126.
- Wang F, Han W, Chen S, Dong W, Qiao M, Hu C, and Liu B (2020) Fifteen-year application of manure and chemical fertilizers differently impacts soil ARGs and microbial community structure. *Frontiers in Microbiology* 11: 62. <https://doi.org/10.3389/fmicb.2020.00062>. PMID: 32117108. PMC7015874.
- Yang C, Du W, Zhang L, and Dong Z (2021) Effects of sheep manure combined with chemical fertilizers on maize yield and quality and spatial and temporal distribution of soil inorganic nitrogen. *Complexity* 2021: 4330666. <https://doi.org/10.1155/2021/4330666>.
- Yang Y, Liu H, and Lv J (2022) Response of N_2O emission and denitrification genes to different inorganic and organic amendments. *Scientific Reports* 12: 3940.

Septic systems and municipal waste disposal on land

Charles P Gerba^a, Michael J Goss^b, and Adrian Unc^c, ^aUniversity of Arizona, Tucson, AZ, United States; ^bSchool of Environmental Sciences, University of Guelph, Guelph, ON, Canada; ^cSchool of Science and the Environment, Memorial University of Newfoundland, Corner Brook, NL, Canada

© 2023 Elsevier Ltd. All rights reserved.

This is an update of C.P. Gerba, WASTE DISPOSAL ON LAND I Liquid, Editor(s) Daniel Hillel, Encyclopedia of Soils in the Environment. Elsevier, Elsevier 2005, Pages 238-247, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00124-7>.

Introduction	568
Decentralized wastewater treatment systems	568
Wastewater treatment plants	568
Primary wastewater treatment	568
Secondary wastewater treatment	568
Trickling filter	569
Activated sludge	569
Tertiary wastewater treatment	569
Treatments for nitrogen and phosphorous	569
Land application of sewage biosolids	569
Land application of effluent from wastewater treatment plant	570
Types of land-treatment processes	571
Slow-rate process	571
Rapid-infiltration process	571
Overland-flow process	572
Treatment mechanisms	572
Filtration	572
Adsorption and precipitation	573
Chemical reactions	573
Volatilization	573
Biological mechanisms	573
Fate of specific contaminants	574
Metals and trace elements	574
Organic compounds	574
Disinfection by-products	574
Endocrine disruptors	575
Pharmaceuticals	575
Microorganisms	575
References	577

Abstract

On-site – mainly septic systems – and centralized wastewater treatment systems treat vast volumes daily. Most wastewater from septic systems reaches soils via covered leaching beds. Municipal biosolids are applied to cropland for their nutrient value after receiving treatments of various intensity but the soil is relied on to act as a treatment medium by adsorbing and degrading contaminants, inactivating viruses, and reducing viability of other pathogens. The efficacy of soil treatment depends on climate, hydraulic loading which is managed to induce fast or slow infiltration, the hydraulic properties of the soil, and the functional state of the soil biology.

Key points

- Wastewater must undergo treatment to reduce Biochemical Oxygen Demand as well as the chemical and microbial contaminant loads
- Sewage biosolids, the product of wastewater treatment, can nevertheless carry contaminants and harbor microbial pathogens
- After land application, soil can act as a waste treatment medium, and the treatment's efficacy depends on the soils' hydrological, physical and biological functional parameters and the waste loading rate
- There are three major land-treatment processes: slow rate, rapid infiltration, and overland flow.
- Chemical contaminants, inorganic and organic, can be filtered, adsorbed, precipitated, volatilized, and degraded

- Transport through soil of microbial contaminants can be delayed due to filtration and adsorption
- Unfavorable abiotic and biotic conditions in the soil can inactivate or kill microbial contaminants

Introduction

The application to land of human wastes has been practiced for centuries. The benefit of the nutrients provided in these wastes has been recognized as a major benefit in crop production. However, because of the presence of pathogenic organisms in such wastes, the use of untreated wastes is no longer practiced in developed countries. The main sources are decentralized or on-site wastewater treatment systems (e.g., septic tanks) and wastewater treatment effluents and treated sewage sludges (biosolids).

Decentralized wastewater treatment systems

Decentralized or on-site wastewater treatment systems serve an estimated 21.7 million families in the USA plus numerous business and recreational facilities (U.S. EPA, 2021). Worldwide, septic systems are the means by which 24% of waste water is treated (UN Habitat and WHO, 2021). The basic treatment system comprises a septic tank coupled to a distribution system that spreads the effluent below the soil surface via perforated pipes, the leaching bed. The pipes are surrounded by gravel in an absorption trench, below which is a sand bed on top of the underlying soil. As the wastewater moves from the gravel into the sand, some solids are filtered out and this process continues throughout the passage of the sand bed. A bacterial biofilm forms at the gravel-sand interface, which is responsible for removing organic material from the wastewater together with some minerals. This is an aerobic zone, allowing fast breakdown of organic material. Between 50% and 90% of the waste treatment takes place in the soil.

Newer systems employ a variety of treatment processes including aeration, filtration, clarification and disinfection. The aeration system adds a compartment where the material from the tank is subjected to air that is bubbled through. A filtration unit requires an additional compartment that is filled with natural or manufactured materials in which bacterial films, developed under aerobic conditions, breakdown the organic matter. A newer system involves the use of pumps that slowly draw the wastewater through fiber membranes. Each fiber has billions of microscopic pores that are small enough to prevent suspended solids, bacteria and viruses from passing through. Clarification, which increases the settling out of particulates, requires the presence of additional settling compartments or the addition of surfaces that will encourage adsorption. Disinfection relies mainly on the applying of compounds that release chlorine.

Septic tanks typically need to be pumped out, periodically. Many jurisdictions require the resulting septage to be handled through wastewater treatment plants but some still allow untreated material to be applied directly to land.

Wastewater treatment plants

There are nearly 24,000 wastewater treatment plants in Europe, over 16,000 in the USA and approximately 58,500 worldwide. Those in the USA treat about 150×10^9 L day⁻¹ of wastewater per day (U.S. EPA, 1997) but worldwide the total wastewater produced is approximately 3.6×10^{14} L year⁻¹, of which only 4.1×10^{12} L year⁻¹ go to a treatment plant (Ehalt Macedo et al., 2022). There are typically three treatment stages of piped sewage.

Primary wastewater treatment

Sewage and other waste streams entering a wastewater treatment plant passes through a screen to remove large floating objects, such as sticks, plastics (including wipes) and rags, before entering a grit chamber that allows small metal items, stones, sand-sized mineral particles, coffee grinds and eggshells to settle out. The remaining entrained solids are removed from the sewage in sedimentation tanks. These long narrow tanks hold the wastewater for approximately two hours to allow settling to take place under gravity. The material that settles (known as raw or primary sludge) is removed by mechanical scrapers, which can also remove floating material, such as fats.

The initial treatment involving further time in even slower movement of the wastewater to allow more solids to settle out, leaving only particles less than 10 µm still in suspension. The net effect is to reduce the level of oxygen required by the bacteria that breakdown the organic matter (biological oxygen demand) by a minimum of 20% before entering the next stage of treatment.

Secondary wastewater treatment

At this stage, the process of treating wastewater generally involves biological treatment together with further activities to remove at least 75% of the initial Chemical Oxygen Demand (COD). The trickling filter and the activated sludge processes are the most commonly used secondary treatments.

Trickling filter

A trickling filter is essentially a tank filled with porous media such as a deep bed of stones or slag or purposely designed porous plastic media. The stone beds are commonly up to 2.4 m deep but usually of large diameters, up to 60 m, while the plastic media is usually packed in towers that are tall (up to 12 m), but narrower (6–12 m in diameter). The plastic media can be designed to create crossflow, which has been shown to minimize clogging. The wastewater from the first treatment is continuously trickled over the top of the trickling filter from revolving arms and percolates to the bottom, where it is collected for further treatment. During this process, bacteria gather and multiply on surfaces of the filter material, creating biofilms. Because of this, the process may be also described as biofilm filtration. The steady unsaturated flow of dissolved and fine particulate organic matter entrained in the wastewater allows the bacteria to absorb the organics, thus lowering the biochemical oxygen demand (BOD) of the sewage. Air moving through the spaces between the filter's surfaces provides sufficient oxygen to meet the oxygen requirements for microbial metabolism. Two or more trickling filters may be connected in series so that wastewater can be recirculated for greater treatment efficiency.

Passage through settling tanks, secondary clarifiers, follow the trickling filters to collect the bacteria and their extracellular polymeric substances that are displaced by the flow of wastewater, including sloughed off biofilms.

Activated sludge

Activated sludge treatment consists of an aeration tank followed by a secondary clarifier tank. Wastewater that has passed through the secondary process is mixed with fresh sludge from the secondary clarifier in an aeration tank. Compressed air is then injected into the mixture from the bottom of the tank. As the bubbles rise to the surface of the wastewater the diffused air provides oxygen and a rapid mixing action. Alternatively, aeration can also be supplied by the stirring action of propeller mixing-blades at the surface. The duration of the initial part of the process, about six hours, allows for the production of microbes and the removal of the dissolved organic material. Secondary settling tanks deal with the removal of microbial waste from the process.

Tertiary wastewater treatment

Tertiary treatment allows for a reduction in the level of contaminants, such as key plant nutrients, nitrogen and phosphorus, where effluent is released into lakes or rivers.

Treatments for nitrogen and phosphorous

Oxidation ponds, also known as stabilization ponds, can be used to improve the quality of wastewater effluent through the interaction of sunlight, bacteria, and algae. They are large, shallow ponds or lagoons containing algae. Algal growth depends on photosynthesis using energy from the sun, to fix the carbon dioxide, nitrate and phosphorus released by bacteria in the different treatment stages. During photosynthesis, the algae release oxygen that is required by aerobic bacteria. Mechanical aerators are sometimes installed to supply yet more oxygen, thereby reducing the required size of the pond. Sludge deposits in the pond must eventually be removed by dredging. Algae remaining in the pond effluent can be removed by filtration or by a combination of chemical treatment and settling.

Rotating biological contactors consist of a series of large plastic blocks or disks mounted on a horizontal shaft. The units are partially submerged and thus as they rotate, the biofilms formed on their surfaces are alternatively exposed to anoxic conditions in the wastewater effluent, when they are partially submerged, or are well aerated when lifted above the liquid. During the latter stage, bacteria on the surface of the units oxidize ammonia to nitrite and nitrate but during the anoxic phase, these oxides of nitrogen can be reduced back to ammonia or to dinitrogen gas.

Land application of sewage biosolids

Biosolids from municipal wastewater treatment plants are commonly applied to agricultural land but this is usually subject to regulatory control. After general screening and grit removal, suspended organic material is separated out by sedimentation (Primary Treatment) and the microorganisms present in the waste digest the easily metabolized fraction (Secondary Treatment). In addition, inorganic nitrogen (N) and phosphorus (P) in solution may be removed (Tertiary Treatment). The material is stabilized in a final treatment stage. In this stage the material is subjected to heating and drying, alkaline stabilization to increase the pH, aerobic or anaerobic digestion together with heating, or composting. Stabilization is intended to reduce or eliminate pathogens and make the material less attractive to scavengers after land application. During these different treatment stages, the microbes in sewage are reduced in number, becoming concentrated in the sludge. However, some pathogens are still present in the effluent.

Biosolids are applied to land, either as a liquid or, if they undergo dewatering before being spread, as a cake. These materials are commonly referred to as municipal or sewage biosolids.

The projected production of sewage biosolids for 2010 in the USA and the original 15 countries of the European Union was about 16 Tg dry matter per year (Epstein, 2002). Although not all European countries use sewage sludge in agriculture, over the period 1996–1998 France applied 60% of its production to agricultural land and comparable values for Spain and the UK are 46%, Germany 40% and Italy 16%. Typically, between 50% and 70% of sewage biosolids produced are land applied (Epstein, 2002). Kelessidis and Stasinakis (2012) provide a comparative assessment of treatments and approaches to the disposal of sewage biosolids in Europe.

Land application of effluent from wastewater treatment plant

Land treatment is defined as the controlled application of wastewater onto the land surface to achieve a planned degree of treatment through natural physical, chemical, and biological processes within the plant–soil–water matrix. During land treatment of wastewater effluents, biological and chemical pollutants are removed by physical (settling, filtration), chemical (adsorption, precipitation), and biological (e.g., plant uptake, microbial transformation) processes. The key objectives of land application of wastewater are further effluent treatment, groundwater recharge, and provision of nutrients for agricultural crops. However, effluent is used in many water-short areas of the United States, with or without disinfection, for crop and landscape irrigation.

There are three major land-treatment processes: slow rate, rapid infiltration, and overland flow (Fig. 1, Table 1). A comparison of the design features of these land-treatment systems are seen as important in the recharge of drinking-water aquifers, and slow-rate systems as a means of crop production in the arid regions of the world. Because of the potential impact of contaminants on underground drinking-water supplies, the fate of trace organics and pathogens have received increasing attention where rapid infiltration is used to supplement groundwater supplies. Because of the transport of food crops around the world today, the fate of pathogens in slow-rate systems has also received more attention.

An expert review of the epidemiological evidence, sponsored by the World Health Organization, recommendation that treated wastewater contain less than one viable intestinal nematode egg L^{-1} (arithmetic mean) for restricted or nonrestricted irrigation, and less than 10,000 fecal coliform bacteria (FC) L^{-1} (geometric mean) for unrestricted irrigation (Blumenthal et al., 2000).

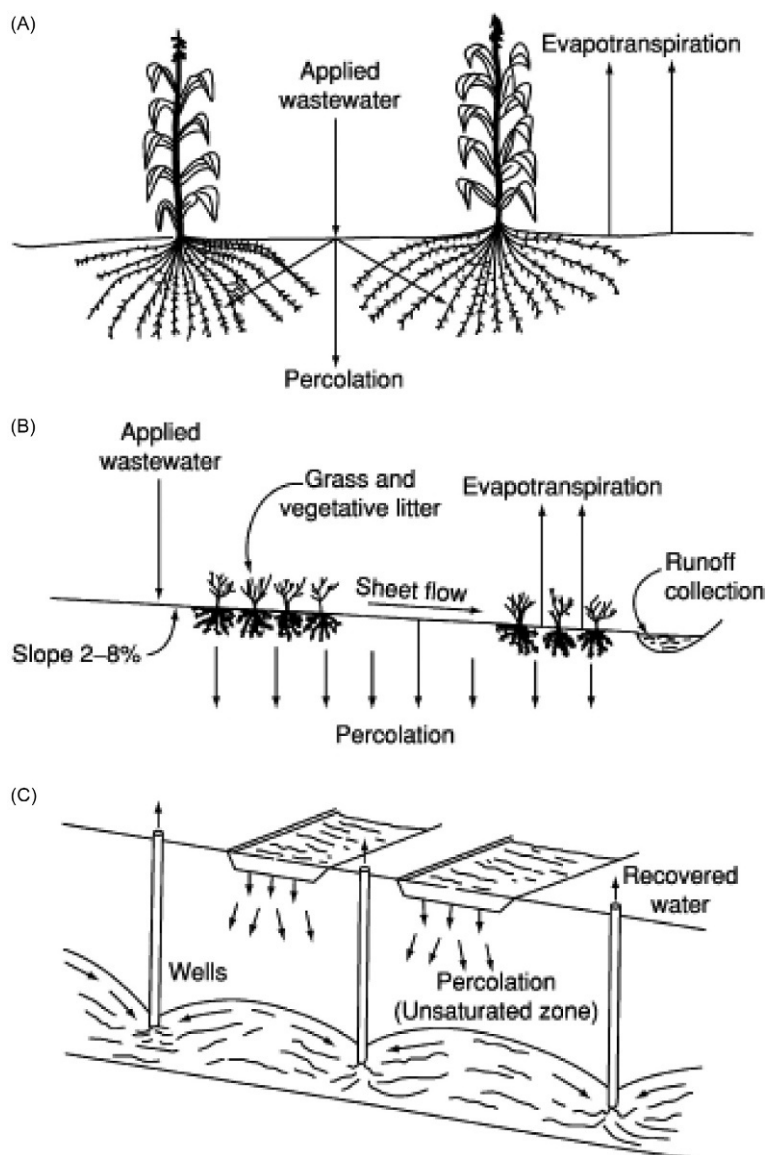


Fig. 1 (A) Slow-rate irrigation system; (B) overland-flow system; (C) rapid-infiltration system. Reproduced from the US Environmental Protection Agency (1981) *Technology Transfer Process Design Manual for Land Treatment of Municipal Wastewater*. EPA/1-81-013. Cincinnati, OH: USEPA.

Table 1 General characteristics of the three methods used for land application of sewage effluent.

Factor	Application method		
	Low-rate irrigation	Overland flow	High-rate infiltration
Main objectives	Reuse of nutrients and water Wastewater treatment	Wastewater treatment	Wastewater treatment Groundwater recharge
Soil permeability	Moderate (sandy to clay soils)	Slow (clay soils)	Rapid (sandy soils)
Need for vegetation	Required	Required	Optional
Loading rate	1.3–10 cm week ⁻¹	5–14 cm week ⁻¹	>50 cm
Application technique	Spray, surface	Usually spray	Surface flooding
Land required for flow of 10 ⁶ L day ⁻¹	8–66 ha	5–16 ha	0.25–7 ha
Required depth to groundwater	Approx. 2 m	Undetermined	5 m or more
BOD and suspended solid removal	90–99%	90–99%	90–99%
N removal	85–90%	70–90%	0–80%
P removal	80–90%	50–60%	75–90%

BOD, Biochemical oxygen demand.

Land-treatment systems are capable of reducing the levels of pathogenic microorganisms, biochemical oxygen demand (BOD), suspended solids, nutrients (N and P), toxic metals, and trace organics. Suspended solids are removed by filtration and sedimentation. Soluble organics are removed by microbial action in the soil. Nitrogen is removed by sedimentation-filtration (e.g., particle-associated organic nitrogen), adsorption to soil, volatilization, e.g., ammonia (NH₃), uptake by crops, and biological denitrification. Adsorption to soil particles, chemical precipitation, and uptake by vegetation combine to remove P.

Types of land-treatment processes

Slow-rate process

Slow-rate irrigation systems are the most frequently used land-treatment system. There are approximately 1200 systems in the USA. Slow-rate land treatment is the application of wastewater to a vegetated land surface, with the applied wastewater being treated as it flows through the plant–soil matrix. Some of the flow percolates to the groundwater and some is used by the vegetation. Off-site surface runoff of the applied wastewater is generally avoided in the design. Nitrogen is often the limiting factor in these systems because of strict standards for levels of nitrate in drinking water. In arid lands, salts at acceptable levels for crop production also may be limiting. Usually, sandy loams or clay loams are the preferred soil type.

In slow-rate systems, the soil can be as shallow as 0.3 m for grass crops, but 1.5 m is preferred for complete wastewater treatment. Retention of contaminants is a function of residence time of the wastewater in the soil and the degree of contact with the soil colloids.

BOD in most slow-rate treatment systems is applied at rates of less than 11 kg ha⁻¹ day⁻¹, which is an order of magnitude smaller than the capacity of the soil. High-strength food-processing wastewater systems are loaded as high as 9500 kg ha⁻¹ day⁻¹. The BOD reductions in most systems exceed 98%. Filtration through the soil results in removal of 99% or more of total suspended solids.

Nitrogen removal is achieved using a nitrogen balance that matches the expected removal plus a percolate nitrate-N of less than 10 mg L⁻¹, which is the drinking-water standard in the USA. Slow-rate systems use forage crops to remove much of the applied N. To achieve the nitrogen uptake rates expected for the crop, excess nitrogen must be applied so that the crop can compete effectively with soil microorganisms for the available nitrogen. Biological-N removal occurs by nitrification–denitrification. The loss by denitrification depends on the ratio of BOD-to-N and the soil temperature, pH, and moisture content. Intermittent application, which is characteristic of slow-rate systems, serves to enhance nitrification followed by denitrification. Ammonia volatilization losses of 10% can be expected if the soil pH is greater than 7.8 and the cation exchange capacity is low (little adsorption of ammonium by the soil). For soils with less than 2% organic matter, storage of N in the soil can be a significant loss for the first 3–4 years of operation of the system. Eventually, equilibrium is reached and the net storage of N stops.

Removal of P depends mainly on the formation of insoluble phosphates with metals at the soil surface. It is generally found that about 98–99% of P is removed in this way.

The concern with pathogens in slow-rate systems is related to their potential for reaching the groundwater, contamination of adjacent surface waters, or contamination of the crop. The small application rates used for wastewater generally limits the potential for groundwater contamination, and contamination of surface waters can be controlled by proper management and site selection. However, enteric viruses have been detected in groundwater beneath slow-rate systems when high rates of wastewater are applied to areas with shallow water tables. Food-crop contamination with pathogens easily occurs during furrow irrigation, but can be limited by drip irrigation and covering the soil surface with plastic sheeting.

Rapid-infiltration process

In rapid infiltration, most of the applied wastewater percolates through the soil and the treated effluent seeps through to the groundwater, where it may be collected in recovery wells for drinking water or irrigation. The wastewater is applied intermittently at

high loading rates ($6\text{--}125\text{ m year}^{-1}$) onto a permeable soil (e.g., sands or loamy sands). Application is usually in basins and vegetation is usually not planted. Recovery wells may be used to reclaim the treated water for irrigation or as a drinking-water source. The treatment potential of rapid-infiltration systems is less than in slow-rate systems. Removal of N is generally small but may be enhanced by encouraging denitrification. Denitrification requires adequate carbon (as found in primary wastewater effluents) and low oxygen levels, necessitating flooding periods as long as 9 days, followed by drying periods of approximately 2 weeks. Climate and season will affect the intervals of flooding and drying due to their influence on microbial activity.

Typical BOD loadings for rapid-infiltration systems using municipal wastewater range from 27 to $175\text{ kg ha}^{-1}\text{ day}^{-1}$ with removals ranging from 74% to 96%. BOD loadings of industrial systems range from 112 to $676\text{ kg ha}^{-1}\text{ day}^{-1}$. Suspended-solids loadings greater than $112\text{ kg ha}^{-1}\text{ day}^{-1}$ require more frequent diskings or scarifying of the infiltration basin surface to avoid plugging of the soil.

Nitrification–denitrification is the principal mechanism of ammonia (NH_3) and N removal from wastewater in rapid-infiltration systems. Ammonia adsorption of NH_3 also plays an important role in retaining ammonia in the soil long enough for biological conversion. Nitrification and denitrification are affected by temperature and the amount of organic carbon available. It has been found that nitrification rates of up to $67\text{ kg ha}^{-1}\text{ day}^{-1}$ can be achieved under favorable moisture and temperature conditions. Removal of N is a function of detention time, BOD-to-N ratio, and anoxic conditions. Detention time is related to hydraulic loading rates through the soil profile and the soil type dependent hydraulic conductivity. For effective nitrogen removal (80% or more), the loading rate should not exceed 3–4 cm. The BOD-to-N ratio needs to be 3:1 or more to ensure adequate carbon to drive the denitrification reaction. Secondary wastewater effluent will have a BOD-to-N ratio of approximately 1:1, whereas primary effluent usually has a BOD-to-N ratio of 3:1. To overcome the small ratio of BOD-to-N in secondary effluent, a longer period of application (7–9 days) is necessary. Typical removal of total nitrogen ranges from 38% to 93%.

Adsorption and chemical precipitation contribute to P removal. The adsorption occurs quickly and the chemical precipitation occurs more slowly; removal typically ranges from 40% to 97%.

Removal of bacteria, helminths, and protozoan parasites is largely accomplished by filtration. Under unsaturated conditions almost all are removed within a meter of the soil surface. Virus removal is dependent upon adsorption to the soil surface; viruses have the potential to travel long distances under the right conditions (50 m or more). The depth of the unsaturated or vadose zone is important in the degree of removal of viruses before the wastewater reaches the aquifer. Expected removals of virus should be 99% or more.

Overland-flow process

In overland-flow treatment, wastewater is applied at the upper reaches of grass-covered slopes of 2–10% grade and allowed to flow over the vegetated surface into runoff collection ditches. Typical slopes are 2–4% in grade and 36–45 m long. The overland-flow process is best suited to sites having relatively impermeable soils. The most suitable soils are clay or clay-loamy soils with a permeability equal to or less than 0.5 cm h^{-1} . The wastewater is treated by a thin biofilm and plants as it flows down the slope. Removal of nitrogen is due to nitrification followed by denitrification, and uptake by the grass or other plants. Phosphorus removal is via adsorption and precipitation.

In municipal systems, the BOD loading rate typically ranges from 12.3 to 4 kg ha^{-1} . Biological oxidation accounts for the 90–95% removal of BOD normally found in overland-flow systems. A typical BOD concentration in treated runoff water is approximately 10 mg L^{-1} . Overland flow is effective in removing most suspended solids, with effluent total suspended solids between 10 and 15 mg L^{-1} .

The removal of nitrogen depends on nitrification–denitrification and the crop uptake of nitrogen. Denitrification can account for 60–90% of the nitrogen removed, with denitrification rates of 160 kg ha^{-1} . Phosphorus removal in overland-flow systems is limited to approximately 40–50% because of the lack of soil–wastewater contact.

Overland flow is not very effective in removing microorganisms. Fecal coliforms can be reduced by approximately 90% when raw or primary effluent is applied; however, minimal removal occurs when secondary effluents are applied. Enteric virus removals of up to 85% have been observed with overland flow.

Treatment mechanisms

In the land application of wastewater, the soil is used as a treatment medium. Some substances pass through the soil and into the groundwater, some are utilized by growing plants, while others are retained almost indefinitely within the soil. Proper design of land-application facilities must relate the fate of pollutants to the properties of soil with which they may interact and minimize the fraction of contaminants passing through to groundwater. There are several separate-unit processes that can be adopted for removal of suspended solids, total dissolved solids, biochemical oxygen demand, nitrogen, phosphorus, and toxic substances. Waste-treatment mechanisms that occur in soils can conveniently be categorized as physical, chemical, and biological. Within each category, various processes act to remove or alter specific pollutants.

Filtration

As wastewater moves into and through soil pores, suspended solids are removed by mechanical filtration. The actual depth at which removal occurs varies with the size of suspended particles, soil texture, and the rate of application. The larger the hydraulic loading rate and the coarser the soil, the less the detention rate, and thus the greater the distance required. However, when loading rates are

such that much of the applied water is held in the soil, additional removal can occur as suspended materials settle or adhere to the surfaces of the soil particles. Some organic particles and some organisms such as worms and protozoa are large enough so that removal by simple blockage occurs. Both bacteria and viruses, however, are small enough that they can move through the soil pores. While filtration in the soil matrix limits the movement of most bacteria to a meter or less, viruses are too small to be affected, and filtration is not believed to play a significant role in their removal. In practice an organic mat or 'Schmutzdecke' forms at the soil surface, enhancing the removal of particulates. Suspended matter can evidently clog soil pores and thereby severely reduce the infiltration rate. In rapid-infiltration operations, flooding and drying cycles help restore infiltration rates. During flooding the infiltration rate is reduced over time. Drying of the Schmutzdecke results in its cracking and removal during the next infiltration event.

Adsorption and precipitation

Chemical reactions among dissolved ions or compounds and interactions with the soil solid-phase alter the mobility of waste pollutants. Some dissolved constituents are retained within the profile indefinitely, whereas the movement of others is only temporarily restricted or unaffected. Two processes, adsorption and precipitation, account for most of the retention. Adsorption refers to the net effect of interactions between dissolved and particulate (i.e., viruses) matter of the soil and humus particles. Precipitation is the formation of an insoluble form of a substance that was originally in solution. Ion exchange is one type of adsorption process. Anion exchange is of very minor importance in land-application systems. Phosphorus is the only anion appreciably retained in soils, though the primary mechanism is not anion exchange. The predominant ion exchange reactions in most soils involve cations and are related to the cation exchange capacity. In wastewater-treatment systems, the retention of dissolved cations depends largely on their concentration in solution entering the soil. Large concentrations overcome a higher degree of attraction of other cations toward the soil matrix. Many of the heavy metal cations such as zinc, copper, nickel, cadmium, mercury, lead, and chromium are present in wastewater at concentrations too low to be appreciably affected by cation exchange reactions. These cations seem to be incorporated into the soil solid-phase in a nonexchangeable form.

Chemical reactions

Phosphorus and nitrogen are at least partially controlled by chemical mechanisms in the soil. The most important chemical mechanism for nitrogen removal is the reaction of positively charged ammonium ions with the soil cation-exchange complex.

Volatilization

Volatilization refers to the evaporation of chemical vapor from soil or water and its subsequent loss to the atmosphere. Many organic compounds are volatile in water, as are some nitrogen compounds, such as ammonium (NH_4^+) and nitrous oxide (N_2O) generated from biological transformations. In addition, some inorganic chemicals (e.g., selenium compounds) may be rendered volatile through biological reactions. For a given chemical, the extent of volatilization is very dependent on the soil and atmospheric conditions. In general, volatilization is greatly reduced in soil compared with water, because the soil solid-phase retains the chemical mass, thereby reducing its vapor pressure. In addition, the soil can offer substantial resistance to the transport of the chemical from the soil profile to the surface, particularly if the soil is wet and little upward flow of water is occurring. In rapid-infiltration processes where the wastewater is ponded over the surface for prolonged periods of time, the primary route of volatilization loss is from the surface of the standing water. During the drainage cycle, when the soil becomes unsaturated, volatile constituents in solution near the surface can evaporate and escape to the atmosphere.

Ammonia present in wastewater is very volatile and vaporizes from the anhydrous form immediately upon exposure to air. When soil-water content is high and a source of organic carbon is present nitrite may be transformed anaerobically by biological denitrification to several gas species (primarily N_2O and N_2).

Biological mechanisms

Soil microorganisms alter the waste constituents through organic matter decomposition, inorganic transformations, and nutrient assimilation. These processes are largely restricted to the upper meter of soil. The ability of soil microorganisms to decompose organic matter is a function of their population complexity. The functional diversity of microbial communities enhances the capability of soil to degrade a wide variety of organic substances; some microorganisms prey on pathogenic bacteria and reduce the survival of viruses. The rate at which organic matter decomposition occurs and the exact nature of the intermediate and end products depend in part on the composition of the added organic matter in the wastewater. Soil factors, however, exert considerable control as well. The presence or absence of oxygen, more than any other single factor, determines the rate and end products. The oxygen status of the soil is a function of soil porosity, and properties that favor rapid transmission of water also favor oxygen movement unless the soil is completely saturated.

Decomposition proceeds most rapidly under aerobic conditions. Under aerobic conditions, up to 60% of the organic carbon may be respired by organisms as carbon dioxide in the initial stage of decomposition. Much of the rest is incorporated into microbial cells, and some of this is subsequently respired when the population declines. Through organic matter decomposition, N, P, sulfur (S), and a host of other trace elements are converted from organic to inorganic forms, i.e., mineralization. Many of the elements that are mineralized during organic matter decomposition are then subject to inorganic transformation in the soil and are assimilated by plants in overland-flow and irrigation-treatment systems. Vegetation thus functions as a sink where waste nutrients, most notably nitrogen and phosphorus, can be effectively immobilized. When harvested, vegetation plays an integral part in the renovation or 'treatment' of the applied wastewater.

Fate of specific contaminants

Metals and trace elements

Factors that affect the retention of trace elements by soils include soil texture, pH, soil organic matter, and contents of amorphous oxides of Fe, Al, and Mn. Most soils have a high capacity to attenuate the concentration of copper and lead. The retention of other trace elements is best correlated with the clay content of free iron oxides of the soil. Capacities for attenuating cationic elements (Cu, Pb, Be, Zn, Cd, Ni, and Hg) in soils tends to increase with their clay and iron oxide contents. For anionic trace elements (SeO_3 , VO_3 , AsO_4 , CrO_4), retention in the soil also increases with increasing clay and iron oxide content of the soils. In general, the solubility of the cationic trace-element species increases as the pH of the soil decreases. By contrast, the solubilities of anionic trace-element species in the soil tend to increase with the pH of the soil.

Metals and trace elements may be removed by a number of mechanisms during land application of wastewater. Those associated with suspended matter are retained at or near the soil surface. In rapid-infiltration systems, they accumulate in the colloid material in the surface-clogging layer, which eventually must be removed to restore infiltration. Smaller suspended particulates that can move through soil pores without becoming trapped are also attenuated by sorption to mineral surfaces in the soil matrix. For dissolved elements, ion exchange, precipitation, surface adsorption, and complexing with organic compounds are important. The principal mechanisms that immobilize dissolved trace elements in the wastewater within the soil appear to be absorption and precipitation. Most soils appear to have a high capacity to remove trace elements during land treatment of wastewater. The amounts of trace elements removed by crops are small compared with the amounts applied to the soils though application of wastewater (Table 2).

Because soil sorption may be finite for most trace elements, there may be long-term accumulation of metals in the soil. Groundwater concentrations of silver, barium, cadmium, cobalt, and chromium below the rapid-infiltration site at Hollister, California, have been unaffected by the additions of wastewater to the overlying soil. However, manganese, nickel, iron, zinc, lead, and copper are above background levels. Soil samples taken at the Whittier Narrows infiltration facility after more than 20 years of operation have shown elevated levels of cadmium, chromium, copper, nickel, lead, and zinc in the top 60 cm, but not below that depth, suggesting that the soil has the capacity to remove metals for many more years of operation before groundwater is affected.

Organic compounds

Organic compounds vary greatly in their mobility, volatility, and persistence in soil. In land-application systems, volatile compounds volatilize prior to application, and only soluble organic compounds enter the soil. The fate of the soluble compound depends on the degree to which it is chemically or biologically transformed during its passage through the system. Organic compounds degrade by hydrolysis, photo-decomposition, or redox reactions. Microbial conversion occurs chiefly at or near the soil surface, where bacterial populations and organic carbon levels are greatest. Low organic carbon levels limit microbial action in the deeper regions of the vadose zone during treatment in rapid-infiltration systems. The travel-time of an organic compound may be roughly estimated by its retardation factor in a given soil. The overall action of the chemical and microbiological processes transforming an organic compound moving through the soil may be crudely expressed as the half-life or degradation rate constant. The duration of the half-life compared with the travel-time can be used as an index of the potential for the compound to survive its contact and passage through the soil.

The dissolved organic carbon (DOC) in domestic, secondary-treated wastewater during rapid infiltration is reduced by 50% after passage of a few meters through the soil. DOC is primarily removed by microbial action during passage through the soil, with more than half being removed during passage through the top 8 cm of soil. With long residence times in the soil, aquifer DOC is reduced to 1 mg L^{-1} after 12–24 months.

Disinfection by-products

In most cases, wastewater is disinfected before land application with chlorine, chloramine, ozone, or ultraviolet light. In the case of chemical oxidizers, this results in the formation of disinfection by-products which may potentially be harmful to human health.

Table 2 Expected trace element removal by vegetation from wastewater-irrigated soils.

Element	Typical concentration in wastewater (g L^{-1})	Annual input (g ha^{-1})	Typical concentration in vegetation (g g^{-1})	Annual removal (g ha^{-1})	Removal (%)
As	<0.005	<60	1	5	8.3
Cd	0.005	60	0.5	2.5	4.2
Cr	0.025	300	0.5	2.5	0.8
Cu	0.10	1200	15	75	6.3
Hg	0.0009	11	0.02	0.1	0.9
Pb	0.05	600	2	10	1.7
Zn	0.15	1800	50	250	13.9

At an application rate of 1.2 m year^{-1} . Assuming annual dry matter yield of 5 t ha^{-1} e.g., potatoes.

Adapted from Page AL and Chang AC (1984). Fate of wastewater constituents in soil and groundwater: Trace elements. In: Petty grove GS and Asano T (eds) *Irrigation with Reclaimed Municipal Wastewater. A Guidance Manual*. Report Number 84–1. pp. 31–1 through 13–16. Sacramento, CA: California State Water Resources Control Board.

Table 3 Fate of disinfection by-products (DBP) during rapid infiltration.

DBP	Disinfectant	Estimated fate during recharge
Chloroform	Chlorine	Volatilization
	Chloramines	Sorption
Bromodichloromethane	Chlorine	Volatilization
	Ozone	Sorption
Trichloroacetic acid	Chlorine	Mineral sorption
MX	Chlorine	Sorption
		Degradation
Bromate	Ozone	Reaction with soil

MX, 3-chloro-4-(*d*-chloromethyl)-5-hydroxyl-2(*H*)-furanone.

Adapted from NRC (1994) *Groundwater Recharge Using Waters of Impaired Quality*. Washington, DC: US Government Printing Office.

Some of these compounds are listed in Table 3 along with their probable fate during land application of wastewater. In addition to known disinfection by-products, residual halogen is present when chlorine is used. This material may then appear as adsorbable organic halogen (AOX). Sorption processes do not play a significant role in their removal. Under anoxic conditions removal is probably based on co-metabolism and is related to the DOC that is needed as a substrate. Ozone disinfection of wastewater improves the biodegradability of refractory organic compounds and co-substrate concentration for AOX co-metabolism in soil columns. AOX removal in 1-m soil columns to which secondary effluent is applied averages 30%.

Endocrine disruptors

Many endocrine-disrupting compounds are present in wastewater at trace quantities (i.e., estrogens, pesticides, polychlorinated biphenyls, phthalates, pharmaceuticals). Estrogens are probably the largest contributors to the endocrine-disruption activity in domestic wastewater, since daily urination adds estrogen to household wastewater. Humans are known to excrete between 10,000 and 100,000 ng L⁻¹ of 17 β -estradiol per day. Synthetic estrogens such as 17 α -ethinyl estradiol have been detected in wastewater and in surface waters affected by effluent discharge. Measurements of endocrine-disrupting compounds displaced by a fluorescent estrogen assay or cell proliferation assay have shown the reduction of endocrine disruptors during rapid infiltration. By cell proliferation assay, the estradiol equivalent value of 71 ng L⁻¹ was found to be reduced to 13.7 ng L⁻¹ after passage of 63 m through the vadose zone. The equivalent 17 β -estradiol activity measured by a binding assay was reduced by 97% after travel through the vadose zone to reach the aquifer. Since the study site had been in operation for more than 10 years, it seems unlikely that endocrine disruptors would ever break through the vadose zone.

Pharmaceuticals

There has been increasing concern about the fate of micropollutants originating from pharmaceuticals and active ingredients in personal-care products (e.g., soaps, deodorants) that are introduced into domestic wastewater. There is evidence that substances of pharmaceutical origin are not completely eliminated during wastewater treatment or biodegraded in the environment. At two rapid-infiltration sites in Arizona, neither ibuprofen nor naproxen or any other acidic drug were detected in groundwater wells downgradient of the infiltration basins (Table 4). These results indicate a high potential for degradation of anti-inflammatory and lipid-regulator drugs. However, antiepileptic, carbamazepine, and primidone drugs have been detected in all of the wells, suggesting that they are able to persist in soil-treatment systems.

Microorganisms

From the standpoint of acute diseases, microbial contaminants are of greatest concern during the land application of wastewater. Microorganisms are responsible for more than 90% of all waterborne out- breaks reported each year in the USA. Hundreds of

Table 4 Fate of pharmaceuticals after rapid infiltration.

Use/origin	Compound	Secondary-treated wastewater (mg L ⁻¹)	Detection in groundwater (ng L ⁻¹)
Lipid regulator	Gemfibrozil	1235	Not detected
Antiepileptic	Carbamazepine	Not detected	455
	Primidone	110	115
Analgesic/anti-inflammatory	Ibuprofen	3380	Not detected
	Naproxen	6280	20
	Fenoprofen	35	Not detected
	Propyphenazone	20	15

Adapted from Drewes JE, Heberer T, and Reddersen K (2002) Fate of pharmaceuticals during indirect potable reuse. *Water Science and Technology* 46:73–80.

different types of pathogenic microorganisms (i.e., bacteria, viruses, protozoa, helminths) are excreted in fecal material of infected hosts and these can find their way into municipal wastewater. The number of types of pathogenic microorganisms present in wastewater varies by location and over time at a given location. A variety of factors influence pathogen content of wastewater, including the incidence of the disease in the population, the season of the year, the economic status of the population, and water use patterns. Diseases caused by waterborne organisms range from mild gastroenteritis to severe and life-threatening illness, such as hepatitis, cholera, typhoid, paralysis, meningitis, heart disease. New enteric pathogens are discovered almost yearly and the significance of new ones changes over time. Coronaviruses, which include SARS-CoV-2 which causes COVID-19, do consistently evolve and new variants are routinely identified and may be found in municipal wastewaters. Enteric bacteria are common in wastewaters, and they may pose health risk to wastewater treatment plants workers. Moreover, occasionally small proportions of enteric bacteria might survive treatment and reach farmland where the biosolids are used as nutrient sources. This is why biosolids application to crops that are intended to be used fresh for human consumption is generally regulated to avoid food-borne illnesses.

The resistance of enteric viruses to environmental stresses makes them good indicators of the efficacy of wastewater treatment. If conditions are effective for disinfection of enteric viruses, they are most likely effective for most other viruses and bacteria. Nevertheless, treated biosolids can be re-contaminated during handling and storage, commonly due to direct contact with rodents and wild birds. During land application of wastewater, pathogens are removed from the system by a combination of die-off (inactivation in the case of viruses) and physical filtration or adsorption by the soil or plant material. Pathogen removal by overland flow systems has been little-studied, but it is not expected to result in large reductions of enteric bacteria or viruses. Irrigation systems and rapid-infiltration systems have been studied more because of the potential for crop and aquifers.

Transmission of food-borne illness by enteric pathogens due to irrigation with untreated wastewater has been well-established for more than 100 years. For this reason, irrigation with untreated wastewater for food, crops is usually forbidden. Wastewater to be used for food crop production should meet the same standards as drinking water and be intensely monitored. Thus, almost all irrigation with wastewater is for the production of nonfood crops or fruit crops. However, outbreaks of disease associated with apples and raspberries caused by the parasites *Cryptosporidium* and *Cyclospora* suggest that only wastewater treated to potable standards should be used in any food crop production.

Factors controlling the survival of pathogens are shown in Table 5. Table 6 shows the range of reported survival times of selected enteric pathogens in the environment. Temperature is probably the most important factor controlling persistence of pathogens on plant material and soil. The die-off of pathogens increases as the temperature increases; however, under proper conditions (high humidity), some growth of bacterial pathogens may occur on the plants in the field. Rotavirus and enteroviruses appear to be inactivated very rapidly on irrigated grass under the conditions found in the arid southwestern USA.

Moisture and temperature govern survival of bacteria and inactivation viruses. *Salmonella* and other enteric bacteria can survive for several weeks on grass if sufficient organic matter and moisture is available. Helminth eggs such as *Ascaris* are believed to survive for 30–60 days, although they may survive many months in the soil itself.

Table 5 Factors influencing enteric pathogen survival in soils.

<i>Factor</i>	<i>Comments</i>
Temperature	Lower temperatures promote increased survival
Moisture content	Survival is decreased in drying soils. Increases may result in bacterial growth
pH	Survival is decreased pH extremes
Soil type	Survival influenced by chemical, textural, and mineralogical properties such as pH, exchangeable ion content, and water-retention capacity
Organic matter	Survival enhanced and growth of bacteria possible
Soil microflora	Survival is decreased in nonsterile soil
Salt species and concentration	Cationic type and ionic strength may influence survival
Soil surfaces	Survival may be enhanced by adsorption to soil particles
Saturation	Air–water interface under unsaturated conditions may enhance viral inactivation

Table 6 Survival of pathogens in soils.

<i>Organism</i>	<i>Survival time (days)</i>
Coliforms	38
Fecal streptococci	26–77
<i>Salmonellae</i>	15–90
<i>Salmonella typhi</i>	1–120
<i>Entamoeba histolytica</i> cysts	6–8
Enteroviruses	8–175
<i>Ascaris</i> ova	Up to 2 years

Table 7 Factors influencing microbial transport.

Factor	Comments
Size of microorganism	Transport of bacteria and parasites is limited because of size in most agricultural soils
Adsorption	Major factor in limiting virus transport through soils. Also plays a role in bacterial transport
pH	Lower pH enhances adsorption
Salt species and concentration	Increased cation valency and ion concentration increase retention; low ionic strength promotes desorption and transport
Organic matter	Some organics (e.g., humic and fulvic acids) interfere with adsorption and cause desorption
Metal oxides	Virus adsorption to iron oxides occurs readily in soils
Hydraulic conditions and moisture content	Increase flow rates and saturated flow decrease adsorption

Temperature significantly influences the survival of enteric pathogens in soil. Survival of bacteria in soils is also favored by cold temperatures. However, freeze–thawing increases mortality of enteric bacteria and protozoa. Fluctuating temperatures, even above freezing, also reduce survivability of fecal bacteria, when compared to constant temperature conditions. *Cryptosporidium* may persist in soils for several weeks, although its survival time is reduced at temperatures above 35 °C and if multiple freeze–thaw cycles in the soil take place. Extreme acidic or alkaline conditions (pH <6.0 or >8.0) tend to adversely affect the survival of both enteric bacteria and viruses in soil. Greater moisture also enhances survival of most enteric organisms in soil as does greater concentrations of organic matter in the wastewater.

Direct comparisons suggest that non-enveloped viruses (e.g., poliovirus, rotavirus, coxsackievirus) tend to resist inactivation better than enveloped viruses under environmental stress conditions. Temperature is critical. Poliovirus type one is inactivated at a rate of 0.06 log₁₀ h^{−1} and rotavirus SA-11 at 0.04 log₁₀ h^{−1} during the winter (4–10 °C) in Arizona (USA). The rates of die-off during the summer are 0.37 log₁₀ h^{−1} and 0.2 log₁₀ h^{−1}, respectively. It would appear that 8–10 h is needed during the summer and 16–24 h during the winter before inactivation of these viruses. SARS-CoV-2 survives best at low temperatures: on surfaces median estimated virus half-life was measured at >24 h at 10 °C and 40% RH, but ~1.5 h at 27 °C and 65% RH (Morris et al., 2021). However, in soil, enveloped viruses such as SARS-CoV-2 might survive up to 60 days at 10% moisture (Anand et al., 2021).

Transport of microorganisms through the soil is influenced by many of the same factors that affect survival of microorganisms (Table 7). Because of their relatively large size, bacteria, protozoan parasites, and helminths are usually removed within a meter or less of the soil surface. Because of their small size (20–200 nm), viruses have the potential to travel the greatest distances through soil. Removal largely occurs by adsorption to the soil surface and is influenced by the surface properties of the virus and soil, properties which are also affected by the ionic properties of the soil solution. Electrostatic and hydrophobic interactions control the degree of adsorption. Transport is favored in sandy soils with a low clay content. Adsorption largely occurs near the soil surface because of the presence of iron oxides to which the negatively charged viruses readily adsorb. Transport is more limited under unsaturated conditions and at pH levels greater than 8.0. Organic matter in wastewater reduces adsorption because of competition for adsorption sites on the soil surface.

References

- Anand U, Bianco F, Suresh S, Tripathi V, Núñez-Delgado A, and Race M (2021) SARS-CoV-2 and other viruses in soil: An environmental outlook. *Environmental Research* 198: 111297.
- Blumenthal UJ, Mara DD, Peasey A, Ruiz-Palacios G, and Stott R (2000) Guidelines for the microbiological quality of treated wastewater used in agriculture: Recommendations for revising WHO guidelines. *Bulletin of the World Health Organization* 78: 1104–1116.
- Ehalt Macedo H, Lehner B, Nicell J, Grill G, Li J, Limtong A, and Shakya R (2022) Distribution and characteristics of wastewater treatment plants within the global river network. *Earth System Science Data* 14: 559–577.
- Epstein E (2002) *Land Application of Sewage Sludge and Biosolids*, 1st edn CRC Press <https://doi.org/10.1201/9781420032116>.
- Kelessidis A and Stasinakis AS (2012) Comparative study of the methods used for treatment and final disposal of sewage sludge in European countries. *Waste Management* 32: 1186–1195.
- Morris DH, Yinda KC, Gamble A, Rossine FW, Huang Q, Bushmaker T, Fischer RJ, Matson MJ, Van Doremalen N, Vikesland PJ, Marr LC, Munster VJ, and Lloyd-Smith JO (2021) Mechanistic theory predicts the effects of temperature and humidity on inactivation of SARS-CoV-2 and other enveloped viruses. *eLife* 10: e65902.
- U.S. EPA (1997) *Environment Protection Agency, Office of Water. 1996 Clean Water Needs Survey Report to Congress. EPA 832-R-97-003*. Washington, DC: US EPA. <https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=200047ZL.txt>.
- U.S. EPA (2021) *Environment Protection Agency, Office of Water. Report to Congress on The Prevalence Throughout the U.S. of Low- and Moderate-Income Households Without Access to a Treatment Works and The Use by States of Assistance under Section 603(c)(12) of the Federal Water Pollution Control Act*. Available at https://www.epa.gov/system/files/documents/2022-01/low-mod-income-without-treatment_report-to-congress.pdf. Accessed 30.06.2023.
- UN Habitat and WHO (2021) *Progress on Wastewater Treatment – Global Status and Acceleration Needs for SDG Indicator 6.3.1*. Geneva: United Nations Human Settlements Programme (UN-Habitat) and World Health Organization (WHO).

Climate change impacts on soil biology

Alberto Canarini^{a,b}, Lucia Fuchslueger^a, François-Xavier Joly^{a,c}, and Andreas Richter^{a,d}, ^aCentre for Microbiology and Environmental Systems Science, University of Vienna, Vienna, Austria; ^bCenter for Ecological Research, Kyoto University, Otsu, Shiga, Japan; ^cBiological and Environmental Sciences, University of Stirling, Stirling, Scotland, United Kingdom; ^dInternational Institute for Applied Systems Analysis, Advancing Systems Analysis Program, Laxenburg, Austria

© 2023 Elsevier Ltd. All rights reserved.

Introduction	578
Direct versus indirect effects of climate and environmental change	579
Climate change experiments	580
Pulse and press changes and ecological stability	580
Effects of climate change on soil organisms	580
Bacteria, fungi and archaea	580
Soil fauna	581
Viruses	582
Effect of climate change on interactions	582
Global change effects on soil processes	583
Elevated CO ₂ effects	583
Warming effects	583
Changes in precipitation	584
Multiple climate change drivers	584
Summary and conclusion	584
References	585

Abstract

Human activities have caused a rapid climate change affecting all parts of the biosphere, including soils. Soil organisms from all three domains of life, their interactions, and all soil processes for which they are responsible are influenced by and in turn respond to climate change. The understanding of how soil organisms and their processes react to climate changes, is thus central to our ability to manage ecosystems and develop strategies to mitigate climate change. This chapter examines the current state of soil biology (from organisms and communities to the processes they control) in the context of climate change and identifies current gaps in knowledge and promising ways forward.

Key points

- Climate change experiments have many shortfalls in addressing realistic situations, including interactive effects and non-linear behaviors.
- Climate change affects soil biology both directly and indirectly.
- Climate change has large effects on soil communities, but with taxa-specific sensitivities.
- Studying the interactions of soil biota is challenging but represents an important new approach in climate change research in soils.
- Soil organisms are key players for almost any soil process and can potentially cause large feedback to climate change.

Introduction

It is unequivocal that human influence has caused rapid climate change that affects all parts of the biosphere, including soils. Soil temperature and moisture changes are evident in almost all regions of the world influencing the composition and functionality of soil communities. Globally, soils harbor the largest reservoir of organic carbon (around 1500–2400 Gt of carbon in the top meter), and any reduction in soil carbon stocks due to climate change can introduce strong positive feedbacks through increased CO₂ emission into the atmosphere. In contrast, targeted soil carbon management creates opportunities for carbon sequestration and for mitigating climate change.

The magnitude and direction of potential soil feedbacks to climate change depend on soil organisms. For example, both soil organic matter formation and decomposition are mediated by the growth, turnover and carbon use efficiency of soil microorganisms. The respiration by heterotrophic soil organisms is responsible for a CO₂ flux of around 60 Gt C per year from soils to the atmosphere, while autotrophic soil organisms are largely responsible for the production and uptake of the greenhouse gases nitrous oxide and methane. Soil organisms are thus at the very center of understanding the interaction between soil and climate change.

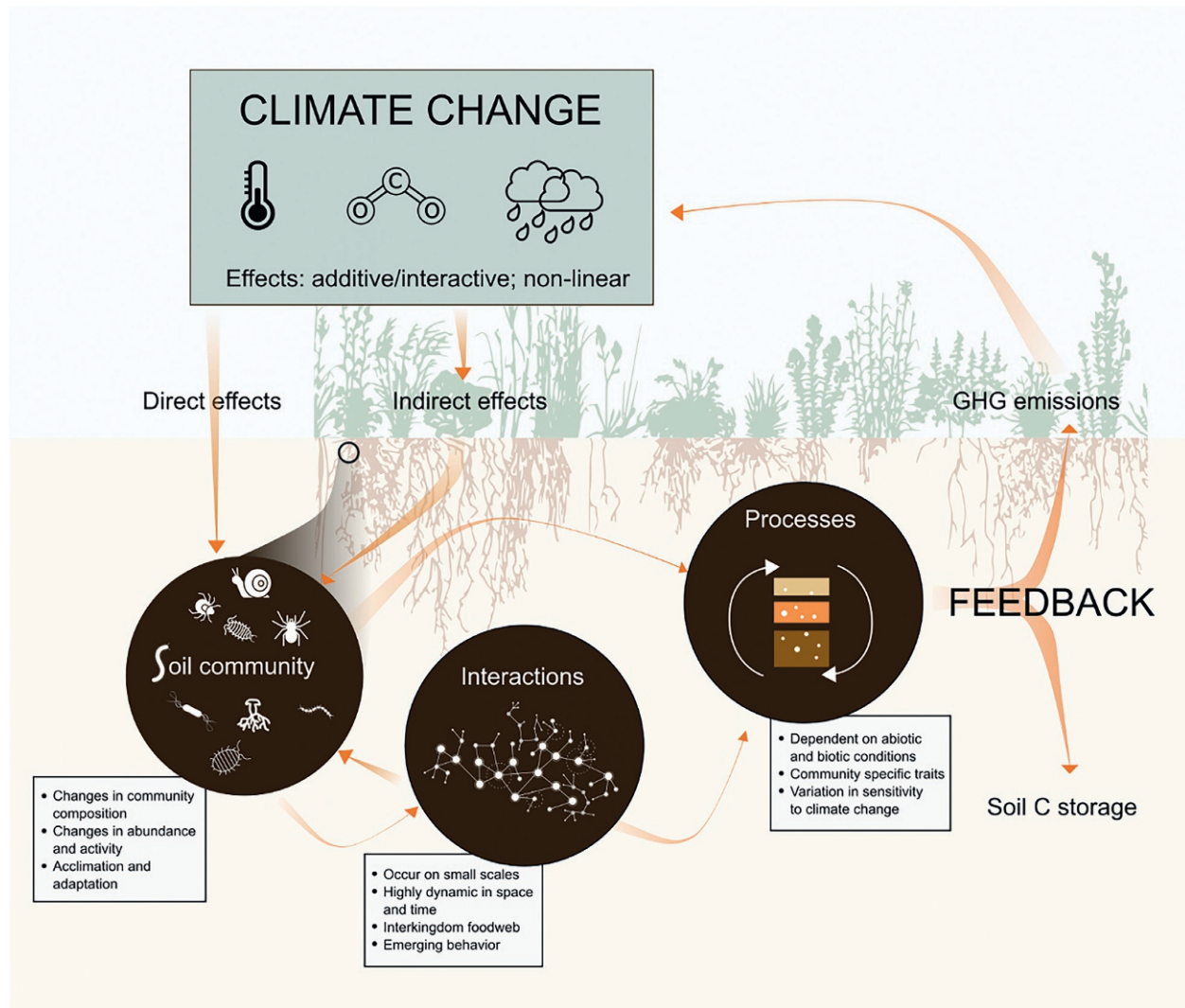


Fig. 1 Climate change affects soil communities, their interaction and soil processes at different levels. Experimental studies often focus on the impacts of single climate change factors, although interactive effects will emerge during co-occurring changes in elevated CO₂, temperature and precipitation. All these factors can either directly affect the physiology, activity and composition of the soil community, or indirectly, via changes of plant inputs and abiotic soil properties. Impacts of climate change on soil biology can be studied at different levels, from single organisms and their interactions up to the processes they control. Effects on soil organisms are challenging to predict given the high biodiversity present in soils, with species from all domains of life interacting with each other. Changes in abundance of soil biota (i.e., in the community composition) and their activity have been recorded, although there is still uncertainty on possible acclimation and adaptation to climate change. Climate change impacts on interactions are difficult to assess given the small scale at which the interactions occur and their high level of dynamicity in space and time, further complicating the prediction of emerging behavior under climate change. Changes at all the investigated levels will contribute to feedback to climate change, by regulating the amount of carbon stored in soil and released to the atmosphere via greenhouse gas emissions.

The inextricable link between climate and soil organisms represents a major scientific challenge. Soil is likely the most biologically diverse system on earth and encompasses taxa from all domains of life: archaea, bacteria, and eukaryotes (including taxa from the phyla fungi, animals, protists and plants), and from viruses (Fig. 1). All these organisms are interacting with each other and with their abiotic environment at various spatial and temporal scales, leading to an unparalleled functional complexity (Fig. 1). Thus, to understand soil, it needs to be considered as an integrated system. Yet, most studies focus on a single group of organisms, largely ignoring their interactions.

Direct versus indirect effects of climate and environmental change

Although most organisms respond directly to temperature or soil moisture, other environmental perturbations have little to no direct effects (Fig. 1). One of the many examples of an environmental perturbation that usually does not directly affect soil organisms, is elevated atmospheric carbon dioxide concentration. First, soil CO₂ levels are usually much higher than atmospheric CO₂ levels and second, most soil organisms are heterotrophic and only fix CO₂ anapleurotically, i.e., they replenish carbon to the metabolism when limited by organic carbon. Yet, CO₂ has a strong indirect effect on many groups of soil organisms: in response to

elevated CO₂ plants may increase biomass production, alters its biochemical composition and change their belowground carbon allocation and leaf litter input, affecting the amount and chemistry of organic carbon available as substrate for the soil food web. However, not only elevated CO₂, but all climate change drivers can cause indirect effects on soil organisms via changes in quality and quantity of plant carbon inputs (Fig. 1). Such indirect effects of climate change may be stronger than the direct ones. Keeping the two factors apart is one of the great challenges when investigating climate change effects on soil biology.

Climate change experiments

Although we are in an era of ongoing climate change, most knowledge on how climate change affects soil and soil organisms comes from climate change experiments. Such experiments are usually short- to medium-term. However, although many indirect effects are manifested only over decades, as for example, those from changes in plant species composition, 15-year studies are already regarded as long-term experiments in most of the scientific literature. In addition, many experiments manipulate only one or two climate change factors at a time. Climate change, however, does not alter only one factor alone. For example, both elevated CO₂ and temperature can cause changing soil moisture conditions, or induce changes in plant water use, thereby modifying the available soil water. Droughts often occur together with heat waves. Hence, predicting climate change effects from single factor experiments is only possible if their impacts on soil organisms were additive, something that, to date, the few multifactorial climate change studies have failed to show convincingly (Rillig et al., 2019).

Interactive effects of more than one co-occurring climate change factor cannot be predicted from studies varying single factors, as this would require a priori knowledge of the response and the interactions of all soil organisms to climate change factors—an impossible task in natural ecosystems. In addition, as climate change factors are often manipulated to a single pre-defined level, many experiments do not allow evaluation of non-linear effects commonly found in many biological systems. For example, soil moisture affects heterotrophic respiration non-linearly, with highest rates found at intermediate soil moisture content. Comparison of climate change experiments and generalization of results are complicated by such non-linear relationships or the interactive effects of different climate change factors. Seemingly contradicting results of climate change experiments have often been interpreted as being ‘context-dependent,’ i.e., being dependent on differences in the state of the ecosystems at the start of the experiments. It is not unlikely, however, that such contradicting results are in fact due to non-linear relations in climate change responses.

Pulse and press changes and ecological stability

Most climate change experiments address gradually developing, long-lasting climate-related changes (also called ‘press’ or ‘ramp’ changes), such as warming, but also investigate the impact of extreme events, such as of heat waves or torrential rain events (also considered ‘pulse’ changes). Such studies become more important as the frequency of these events is increasing. Both ‘press or ramp’ and ‘pulse’ changes can have strong and long-lasting effects on many ecosystem components, including soil. The ecological stability of an ecosystem is defined as the ability of a system to continue functioning under changing conditions or during disturbances. Thus, the stability can be measured by the degree of its resistance and resilience to change. Resistance is the degree to which a community or system withstands disturbances. The inverse of resistance is ecosystem sensitivity, which describes the degree to which an ecosystem or community responds to a disturbance. Resilience is the speed with which a system returns to its pre-disturbance level after a disturbance. Thus, resilience is a function of time. Based on the degree of resistance and resilience, ecosystems or communities can also shift to alternative (stable) states. Such shifts may easily go undetected as usually only a few groups of soil organisms and processes are measured in climate manipulation experiments. Finally, climate change experiments often fail to address how changes in the frequency of pulse disturbances or extreme events affects soil responses, which influence the dynamics of community living in soils.

Effects of climate change on soil organisms

In recent decades, our ability to characterize and describe the diversity and community composition of soil organisms has improved tremendously, largely due to the rapid advent of molecular techniques. Experimental evidence supports an important role of soil biodiversity for ecosystem processes such as plant growth and nutrient cycling. As such, understanding the effects of climate change on the diversity of soil micro- and macro-organisms represents a critical challenge (Zhou et al., 2020). Here, we summarize the general patterns for different groups of soil organisms in response to elevated atmospheric CO₂, warming and drought.

Bacteria, fungi and archaea

The effects of elevated atmospheric CO₂ on soil microbial communities are by no means consistent and occur rather indirectly in response to changed plant belowground carbon allocation and litter input. Overall, elevated atmospheric CO₂ seems to decrease bacterial diversity and evenness and to favor slow growing, resource-efficient oligotrophic microorganisms. For example, in a California grassland elevated atmospheric CO₂ decreased copiotroph Bacteroidetes, but increased bacteria with smaller rRNA copy numbers, a common trait of oligotrophs (Naylor et al., 2020). Similarly, a meta-analysis found that the response of fungal diversity to elevated atmospheric CO₂ was neutral, although fungal diversity increased when experiment duration was accounted for, indicating

that short-term experiments may underestimate the impact of elevated atmospheric CO₂ on soil microbial communities (Veresoglou et al., 2016).

Warming has been shown to stimulate microbial abundance, activity, and nutrient cycling rates, both in the short- and long-term, if water and nutrients are sufficiently available (Walker et al., 2018). Across several studies, warming alone did not change alpha-diversity (richness or Shannon-index) but clearly modified the community structure (Zhou et al., 2020). However, this is not always the case (Radujković et al., 2018), particularly for moderate increases in temperatures. Moreover, warming seems to have a stronger positive impact on fungal rather than on bacterial diversity, as well as to foster more specialized communities (Zhou et al., 2020). We would expect a particularly strong impact of warming in colder ecosystems as has been shown for other ecosystem components; however, across experiments, microbial diversity in tundra and temperate or boreal forests has shown responses in richness and Shannon index to warming similar to tropical and subtropical forests (Zhou et al., 2020). While scarce, information on the functional community diversity indicates no effects of warming or elevated atmospheric CO₂ (e.g., nitrifier communities, Séneca et al., 2020).

The response of microbial communities to changes in soil moisture, including drought periods and subsequent heavy rainfalls, is usually stronger than the response to warming. Common phylogenetically conserved features, such as osmolyte production, synthesis of capsules and exopolymeric substances to retain water, or dormancy and sporulation to avoid water deficit, can determine winning strategies between different microbial taxa. Yet, the effects of precipitation changes on soil microbial communities have shown contrasting effects across studies. Moderate increases in precipitation had stronger effects on microbial communities than moderate decreases, while the opposite was true for more extreme changes (Zhou et al., 2018). Thus, differences in the range of possible experimental manipulation (level of soil moisture changes compared to ambient conditions) and pre-existing environmental conditions (e.g., average precipitation experienced by different ecosystems) make comparisons across studies difficult.

The increase in frequency of drought events raises questions on the long-term impact of changing precipitation patterns. A recent study shows that long-term recurrent droughts can manifest changes in soil bacterial and fungal community and modify their response to a new drought event (Canarini et al., 2021). The authors show that recurrent droughts increase the dissimilarity of microbial communities, compared to control and single drought events, and enhance soil multifunctionality during drought (including key enzymes regulating soil organic matter decomposition). Thus, while drought events trigger specific changes in the community composition of both bacteria and fungi (Isobe et al., 2020), the dependency on drought severity and the long-term dynamics are still largely unknown. Finally, rewetting events that typically follow periods of drought, have also been shown to be very important for restructuring the soil community, and may have large effects on bacterial and fungal communities.

Soil fauna

Like microorganisms, soil animals are well-adapted to high CO₂ concentrations in soil, that regularly reach values of above 100,000 ppm in carbon rich soils. Elevated atmospheric CO₂ is thus unlikely to affect soil animals directly. However, its impact on plants could affect soil animals through modification of resource quality, quantity, soil moisture and other soil properties (e.g., changes in root exudation or modification of soil pH). In a meta-analysis of belowground responses to elevated atmospheric CO₂, A'Bear et al. (2014) reported limited changes in soil fauna community composition or abundance in response to elevated atmospheric CO₂. Observed changes were particularly limited to lower trophic levels, with reports of both reductions and increases in nematodes abundances. Collectively, elevated atmospheric CO₂ thus appears to have only limited and indirect effects on soil fauna, particularly with no changes at higher trophic levels which are less directly influenced by the plant community.

Moderate increases in temperature, i.e., a temperature increase that does not exceed the temperature optimum of a population, can reduce physiological constraints, which in turn favors the abundance and biomass of soil fauna, if other factors, such as water or nutrients, are not-limiting. This positive warming effect is the most likely driver of shifts in soil fauna species distribution toward the colder edge of species' range. Temperature increases can also negatively affect soil fauna if they exceed the optimal range of species. Warming can be indirectly detrimental to soil organisms by decreasing the soil water content, or by modifying the vegetation, which in turn alters the resource quality at the basis of the trophic network. The direction and magnitude of the warming effect may also depend on its timing as warming can reduce snow cover and thus expose soil animals to greater temperature fluctuations. Animal species requiring chilling period during their diapause may also be influenced by winter warming.

Drought is likely to reduce soil fauna density and biomass, albeit to varying degrees depending on soil fauna composition and ecosystem. Soil fauna encompass very diverse organisms, from single-celled protists to large multicellular organisms, such as earthworms or mammals. Consequently, their sensitivity to drought is not uniform, both among and within dominant groups. Microfauna, such as protists and nematodes are likely to be the most affected by drought. Many soil invertebrates can cope with seasonal or temporary drought through behavioral (avoiding drought by escaping in time or space) and physiological mechanisms (tolerance). Mainly, their mobility allows them to relocate to more favorable conditions, by burrowing deeper, sheltering in cavities, or surviving severe drought periods in developmental stages that are less prone to damage. Physiological mechanisms include the ability to obtain water from air vapor and to reduce their metabolism during dry periods. Larger soil fauna may also be more adapted to drought, by having lower surface-to-volume ratio and thus lower water loss rate. However, all these adaptations may not be sufficient to cope with more severe or prolonged drought. In addition to the group-specific nature of the soil fauna response to climate change, response may also depend on antecedent moisture conditions. Drought negatively affected predatory nematodes in mesic ecosystems, but not in xeric ecosystems (Franco et al., 2019). The nature of soil fauna response to drought may also depend on the timing of rainfall. In dryland ecosystems where soil moisture is limited and rainfall events are rare, soil biotic responses are controlled by a series of hierarchical thresholds in soil moisture with microorganisms responding to small rainfall events even at

low soil moisture, and soil invertebrates responding only to larger rainfall events inducing higher levels of soil moisture (Nielsen and Ball, 2015). Soil fauna may thus be more sensitive to changes in the timing and intensity of rainfall events than microorganisms. Yet, drought studies more often focus on changes in cumulative rainfall quantity using rainout shelters, leaving timing and frequency of rainfall events much less understood.

Viruses

Soil viruses are an important component of soil microbial communities, but the understanding of their role in the soil food-web is currently in its infancy, and their response to climate change is even less studied. Recent studies have shown that viruses may play an important role in mediating effects of climate change in certain ecosystems, by controlling the composition of the microbial community. In a permafrost thaw gradient, viral predators controlled key microorganisms for soil carbon cycling, including methanogens and methanotrophs (Emerson et al., 2018), and virus encoded enzymes were significant predictors of methane emissions. Thus, viruses may impact ecosystem functions in soils, but whether and how such phage-microbe interactions will be altered by climate change, remains to be elucidated.

Effect of climate change on interactions

Organisms do not live alone—they constantly interact with the individuals that live around them and with their abiotic environment. The sum of these interactions not only shapes community composition, but also greatly affects the ecosystem processes and services provided by soil (Fig. 1). The emerging behavior of micro-level interactions and soil food-web dynamics are thought to affect large-scale processes at the ecosystem level and can be prone to climate change effects. Climate change drivers thus not only affect each individual organism, but also affect the interaction of individuals and taxa at various levels. Soil fauna and microbial communities are very diverse, within a teaspoon of soil, for example, harboring tens of thousands of bacterial and fungal species. This diversity is mirrored by a large range of possible interactions (from chemical exchange to prey-predator relationship) spanning wide variation of the scale level. For example, microbial interactions mostly take place at a distance on the scale of tenth of μm , with an individual bacterial cell in soil interacting on average with about 120 other cells (Raynaud and Nunan, 2014). At the soil food-web level, inter-kingdom interactions can increase the scale to the centimeter or meter scale. The main challenge remains the ability to detect and measure these interactions within a community, especially at the lower end of possible spatial scales. A wide range of statistical (e.g., network analysis) and experimental approaches (e.g., microfluidics) are now being developed to increase our ability to characterize and manipulate microbial interaction at relevant scales. At a soil food-web level, inter-kingdom interactions can be approached by dividing soil food webs into bacterial, fungal, and a direct root energy channel (e.g., via root fungal symbionts and root-feeding nematodes). However, such different energy channels are not necessarily separated. Protists, for example, can have a wide range of feeding strategies, such as feeding on bacteria, fungi or even nematodes. While much work is underway to build classification systems of feeding relationships in soil, a detailed understanding of these traits is needed to improve our understanding of the soil food web. The 'channel concept' thus rather represents a framework that allows simplification of the testing of climate change effects on interactions at the soil community level.

Different climate change factors may of course cause different effects on the soil food web than on interactions among soil communities. Elevated atmospheric CO_2 induces rather indirect and modest alterations of soil biotic food webs by several mechanisms including changing soil water availability, plant productivity or rhizodeposition. Most plant species can form mutualistic associations with mycorrhizal fungi (arbuscular mycorrhizal fungi, AMF, or ectomycorrhizal fungi, EMF). Elevated CO_2 was shown to preferentially increase plant C allocation to AMF than to bacterial communities, followed by a slower release from AMF to the bacterial and fungal populations (Drigo et al., 2010). On the other hand, elevated atmospheric CO_2 also fosters fungal based food webs, such as of fungivorous relative to bacterivorous nematodes (Dam et al., 2017), or the density of ciliates, microarthropod detritivores or gamasid mites, which could exert a bottom-up effects on microarthropod taxa richness (Eisenhauer et al., 2012).

Warming can affect microbial interactions and their interaction with the soil fauna, but the direction of these effects is still largely unknown. Recent reports show that warming can either increase or decrease bacterial network properties (measured by network analysis) such as connectivity and stability (Zhou et al., 2021). When the analysis considers also other organisms (fungi and higher-level trophic groups) results vary across groups (Zhou et al., 2021). In two boreal-temperate ecotonal forests, warming was shown to alter energy flux to higher trophic organisms, with either increase or decrease depending on forest disturbance and presence of drought (Schwarz et al., 2017). This further reinforces the idea that climate change factors can have interactive effects also at the community interaction level.

Drought can strongly affect soil microbial interactions, with fungal-based food webs usually being more resistant to drought. The increase in resistance, however, has a trade-off, after drought ends with the lower resilience of fungal based energy channels relative to bacteria dominated energy channels. Indeed, drought promotes destabilizing network properties (analyzed by co-occurrence networks) in soil bacteria communities, but not in the fungal community (de Vries et al., 2018). Drought duration and the following rewetting events are also potentially important modulators of soil food web and microbial interactions. Reducing the frequency of precipitation without changing the overall rainfall amount, stabilized soil food webs (Andrés et al., 2017). Some studies highlight the possibility of stronger sensitivity of higher trophic levels during drought, and a stronger bottom-up control of soil processes (Siebert et al., 2019), with evidence that community and feeding activity of higher trophic groups (testate amoebae

and nematodes) were more strongly affected than soil microbial biomass and activity. Higher trophic groups were also reported to have unexpected interactive responses to changes in precipitation. For instance, increased growing-season precipitation reduced the overall nematode abundance (Franco et al., 2019). Although the litter-feeding activities of an isopod and millipede species were largely unresponsive to changes in precipitation, they consumed more litter in combination than separately, but this synergistic interaction disappeared under drought (Joly et al., 2021). Ultimately the type of ecosystem will determine the overall drought effects on soil food web. For example, primary inputs from plants can sustain the soil food web during drought, especially for species-rich plant communities (Cesarz et al., 2017), via strengthened mycorrhizal association or root-feeding nematodes.

Global change effects on soil processes

Soil faunal and microbial communities control and regulate soil fluxes of carbon and nutrients (e.g., soil C sequestration and soil organic matter stabilization, greenhouse gas emissions). Climate change can directly affect organisms, by causing changes to their growth, activity and turnover, and enzyme production rates, but also indirectly, by affecting water, carbon or nutrient availability. Soil biota are responsible for a wide range of processes occurring in the litter layer, bulk and rhizosphere soil. For example, soil meso- and macrofauna play a crucial role in fragmenting and decomposing plant organic matter. Similarly, soil fungal and bacterial communities are essential for plant and soil organic matter decomposition, as well as for depolymerizing and mineralizing nutrients bound in organic matter. Hence, their activity rates control soil carbon and nutrient turnover, as well as carbon outputs from heterotrophic soil respiration or nutrient cycling and availability for plants. Unsurprisingly, many of these processes will strongly respond to changing climatic conditions, which we discuss in more details below.

Elevated CO₂ effects

Increased atmospheric CO₂ can stimulate photosynthesis, which potentially increases the amount of carbon in plant biomass and litter. These observations have led to the 'CO₂-fertilization hypothesis,' which proposes that plant responses to increasing atmospheric CO₂ drive increases in ecosystem carbon storage, creating negative feedback to climate change by taking up atmospheric CO₂. Elevated atmospheric CO₂ has been shown to have negligible direct effects on soil processes (mainly at very high concentrations). However, in the short term, elevated CO₂ indirectly stimulates microbial activity and soil respiration rates, microbial biomass, and genetic signals for carbon cycling processes (Deng et al., 2015). Indeed, elevated atmospheric CO₂ can increase plant organic matter and root exudate inputs, which can promote fast-growing microorganisms in the rhizosphere and soil (Blagodatskaya et al., 2010). In the short term this can increase C inputs to the soil, the labile inputs can prime the break-down of labile and more complex soil carbon compounds (the 'priming effect') and level up heterotrophic soil respiration (Naylor et al., 2020). Over the long term, elevated atmospheric CO₂ may rather lead to increased 'new' C turnover in the soil or even induce C losses from ecosystems (van Groenigen et al., 2017). In contrast to C cycling processes, soil microbial N cycling seems to be less affected by elevated atmospheric CO₂, unless plants have a higher N (or P) demand, which could increase the competition of soil microbial communities and plants for N. A lack of either increased biological N fixation, or increased N mineralization rates has been shown to induce a progressive N limitation for plants, which on a decadal scale may reduce a potential 'CO₂ fertilization' effect on plants (Dieleman et al., 2010). Similarly, elevated atmospheric CO₂ may have little effects on soil microbial P cycling in the long-term (Ochoa-Hueso et al., 2019), although experimental data for effects on soil microbial P cycling are still scarce.

Warming effects

Warming can directly increase microbial decomposition by lowering the activation energy for enzyme-catalyzed reactions, that can induce higher specific activities of soil microbes (Walker et al., 2018). However, whereas warming might initially stimulate the metabolism of decomposers and thereby increase decomposition rates, evidence from long-term (> decade) experiments showed that the impact of warming rather decreases over time, so that in the long-term C losses induced by warming might be smaller than suggested by short-term warming studies. This pattern can be explained by a reduction in microbial biomass over time caused by a depletion of available resources (Walker et al., 2018), by adaptations or adjustments in microbial physiology of individual microbes or changes of microbial community composition and structure. Although warming may accelerate respiration and the growth of microorganisms, non-linear, interactive effects have been found when studying growth and respiration in response to elevated CO₂ and warming together (Simon et al., 2020).

The effect of warming on the contribution of soil fauna to soil processes remains unclear. Theory predicts that increased temperature should lead to increased animal activity to satisfy metabolic demand. Yet, in a 4-year climate change manipulation experiment in boreal forests, Thakur et al. (2018) reported negligible net effects of warming on detritivore feeding activity at ambient precipitation. The response of detritivore activity to warming may also be species-specific, with both positive and negative responses reported. Changes in soil fauna community as a result of warming however, could have important consequences on soil processes. For instance, potential detritivores expansion in subarctic ecosystems could stimulate litter decomposition (van Geffen et al., 2011). Earthworm expansion into the arctic was also found to accelerate organic matter turnover and N mineralization thereby facilitating plant productivity (Blume-Werry et al., 2020) and possibly leading to N losses.

Changes in precipitation

Moisture (or soil water availability) is essential for soil microbial activity. Changes in precipitation can thus have strong effects on soil processes. For example, drought reduces water in pore spaces, disconnects soil microhabitats, increases oxygen filled pore space and concentrates nutrients in the remaining soil water. All these changes have been shown to decrease soil microbial activity often measured as heterotrophic respiration rates. However, drier conditions can favor aerobic processes such as nitrification (although it may reduce anaerobic methane production). Overall soil carbon and nitrogen cycling accelerates with increased precipitation and slows down with decreased precipitation. Drought may also affect N and P cycling processes, although there is less information for increased precipitation. Drying and rewetting cycles mobilize available P to a greater extent than N (Gao et al., 2020). Although consequences for P losses or soil biology are not known, this could potentially induce an imbalance between N and P, especially in forest ecosystems. However, we currently lack a unifying theory explaining responses of microbial processes in soil to changes in precipitation. Currently-available data focuses mainly on trace gas fluxes as well as pools of carbon, nitrogen and phosphorous but the underlying processes remain largely unexplored. The main exception is soil respiration, of which, the response to varying soil moisture conditions is well studied. Furthermore, responses to changing precipitation regimes are often difficult to compare. Responses of soil respiration to drying and rewetting conditions largely depends on the intensity of drought and on the initial soil carbon content (Canarini et al., 2017). Especially the response of soil processes to rewetting can strongly change depending on the preceding drought conditions (Meisner et al., 2017). Many processes are difficult to measure during dry conditions, as measurements often require water addition. Beside changes in moisture and oxygen concentration, strong changes in substrate availability accompanied by changes in precipitation can affect soil processes. Therefore, responses of soil processes to drought might not be as straight forward to understand. For example, N₂O emissions during drought periods were dominated by anoxic denitrification processes, despite the drought leading to more oxygen filled pores in soil (Harris et al., 2021). The authors showed an enrichment in nitrogen-containing organic matter on soil microaggregates and suggested a role for chemo- or co-denitrification. The nitrification process, in turn, was quite resistant to drought, possibly due to enrichment in NH₄⁺ during drought and changes in community structure (Séneca et al. (2020).

Drought can negatively affect soil animals, yet faunal-driven litter decomposition appears to be relatively insensitive to drought, from arid to temperate or boreal ecosystems. In a Mediterranean forest, however, the increased rates of litter decomposition under a rainfall exclusion treatment, seemed to be driven by an increase in mesofauna abundance (Homet et al. (2021). This effect does not appear to be specific to dry environments, and litter consumption by an isopod and a millipede species was also not affected by halving rainfall quantity in a temperate forest (Joly et al., 2021). Collectively, these studies suggest that behavioral or physiological mechanisms to cope with drought allow soil animals to maintain their activity and contribution to soil processes.

Multiple climate change drivers

Different climate change drivers can have additive (similar or diametrical) or interactive effects on soil microbial C, N or P cycling. For example, warming could increase microbial activity and heterotrophic respiration, whereas simultaneous dry conditions may reduce microbial activity and reduce stimulating effects by higher temperatures. However, effects of different climate change drivers appear to be ecosystem-specific and depended on the history of the ecosystem. In a grassland ecosystem with experimentally simulated future climate (+3 °C warming in combination with +300 ppm CO₂ above ambient conditions), summer drought significantly increased microbial biomass-specific growth and respiration rates, while elevated atmospheric CO₂ alone did not have any impact on microbial C turnover (Simon et al., 2020). In a cold heathland ecosystem, climate manipulations (warming, drought, and elevated atmospheric CO₂) alone or in combination, did not change protein depolymerization rates, nitrogen mineralization, and amino acid and ammonium uptake showed no significant individual treatment effects (Wild et al., 2018). In a grassland ecosystem, warming and elevated atmospheric CO₂ had only minor effects on nitrifier communities and soil biogeochemical variables (Séneca et al., 2020).

Summary and conclusion

Soils and their biota are an integral part of all terrestrial ecosystems and are key components in the interaction of terrestrial ecosystems with the atmosphere. Climate change can directly or indirectly affect individual organisms and species, and their interactions, thereby affecting soil community composition and processes. Above all, soil biota plays the central role in organic matter decomposition in terrestrial ecosystems, regulating soil organic matter formation (via microbial growth or turnover and necromass production) as well as greenhouse gas release to the atmosphere, which in turn can lead to positive or negative feedbacks to climate change (Fig. 1). Moreover, microbial activity is closely dependent on soil faunal activity, via predation and resource provisioning. The large diversity of soil fauna with contrasting structures, life cycles, reproduction strategies, food sources, and adaptive strategies to environmental change makes climate change effects on soil organisms difficult to generalize. Given the highly interconnected nature of the soil food web and the small spatial scale at which many interactions occur, the effect of climate change on soil organisms cannot be simply understood based on individual responses of these contrasted soil organisms, but shows emergent behavior, making it difficult to predict alterations in soil functions and services in response to climate change.

We have shown above that global climate change affects all soil organisms, from bacteria to soil vertebrates and their interactions, in various ways, directly and indirectly. The concurrent drivers of climate change can act in additive as well as interactive ways and can exhibit nonlinear behaviors. Traditionally, large scale climate change experiments (i.e., field experiments) were set-up from a plant-centric view rather than an ecosystem perspective. For example, most experiments are only conducted during the growing seasons, ignoring the activity of soil organisms during winter or the dry season, when many plants are not active. Thus, we clearly need more integrated climate change experiments that recognize soils and soil biota as equally important components of ecosystems as vegetation. We also need more experiments which are explicitly designed to understand effects on soil biota and the processes they mediate, to disentangle underlying mechanisms and their feedback to vegetation in a future climate. For example, current studies often focus on the very top centimeters of soil, and we lack understanding on how deep soil layers (and the communities within) will be affected by climate change. We also need to initiate climate change experiments which specifically target the manipulation of soil organism and communities, to start addressing causal relationships between climate change effects on soil organisms and the processes they control.

Given all the challenges facing humanity, research into soil organisms, their interactions and processes in a changing climate is crucial if we are to understand the impacts of global climate change on terrestrial ecosystems and how soils can help mitigate climate change.

References

- A'Bear AD, Jones TH, and Boddy L (2014) Potential impacts of climate change on interactions among saprotrophic cord-forming fungal mycelia and grazing soil invertebrates. *Fungal Ecology* 10: 34–43.
- Andrés P, Moore JC, Cotrufo F, Deneff K, Haddix ML, Molowny-Horas R, Riba M, and Wall DH (2017) Grazing and edaphic properties mediate soil biotic response to altered precipitation patterns in a semiarid prairie. *Soil Biology and Biochemistry* 113: 263–274. <https://doi.org/10.1016/j.soilbio.2017.06.022>.
- Blagodatskaya E, Blagodatsky S, Dorodnikov M, and Kuzyakov Y (2010) Elevated atmospheric CO₂ increases microbial growth rates in soil: Results of three CO₂ enrichment experiments. *Global Change Biology* 16: 836–848. <https://doi.org/10.1111/j.1365-2486.2009.02006.x>.
- Blume-Werry G, Krab EJ, Olofsson J, Sundqvist MK, Väisänen M, and Klaminder J (2020) Invasive earthworms unlock arctic plant nitrogen limitation. *Nature Communications* 11: 1–10. <https://doi.org/10.1038/s41467-020-15568-3>.
- Canarini A, Kiazar LP, and Dijkstra FA (2017) Soil carbon loss regulated by drought intensity and available substrate: A meta-analysis. *Soil Biology and Biochemistry* 112: 90–99. <https://doi.org/10.1016/j.soilbio.2017.04.020>.
- Canarini A, Schmidt H, Fuchslueger L, Martin V, Herbold CW, Zezula D, Gündler P, Hasibeder R, Jecmenica M, Bahn M, and Richter A (2021) Ecological memory of recurrent drought modifies soil processes via changes in soil microbial community. *Nature Communications* 12: 5308. <https://doi.org/10.1038/s41467-021-25675-4>.
- Cesarz S, Ciobanu M, Wright AJ, Ebeling A, Vogel A, Weisser WW, and Eisenhauer N (2017) Plant species richness sustains higher trophic levels of soil nematode communities after consecutive environmental perturbations. *Oecologia* 184: 715–728. <https://doi.org/10.1007/s00442-017-3893-5>.
- Dam M, Bergmark L, and Vestergård M (2017) Elevated CO₂ increases fungal-based micro-foodwebs in soils of contrasting plant species. *Plant and Soil* 415: 549–561. <https://doi.org/10.1007/s11104-017-3191-3>.
- de Vries FT, Griffiths RI, Bailey M, Craig H, Girlanda M, Gweon HS, Hallin S, Kaisermann A, Keith AM, Kretschmar M, Lemanceau P, Lumini E, Mason KE, Oliver A, Ostle N, Prosser JL, Thion C, Thomson B, and Bardgett RD (2018) Soil bacterial networks are less stable under drought than fungal networks. *Nature Communications* 9: 3033. <https://doi.org/10.1038/s41467-018-05516-7>.
- Deng Y, He Z, Xiong J, Yu H, Xu M, Hobbie SE, Reich PB, Schadt CW, Kent A, Pendall E, Wallenstein M, and Zhou J (2015) Elevated carbon dioxide accelerates the spatial turnover of soil microbial communities. *Global Change Biology*. <https://doi.org/10.1111/gcb.13098>.
- Dieleman WJ, Luyssaert S, Rey A, De Angelis P, Barton CVM, Broadmeadow MSJ, Broadmeadow SB, Chigwerwe KS, Crookshanks M, Dufrêne E, Jarvis PG, Kasurinen A, Kellomäki S, Le Dantec V, Liberloo M, Marek M, Medlyn B, Pokorn R, Scarascia-Mugnozza G, Temperton VM, Tingey D, Urban O, Ceulemans R, and Janssens IA (2010) Soil [N] modulates soil C cycling in CO₂-fumigated tree stands: A meta-analysis. *Plant, Cell and Environment* 33: 2001–2011. <https://doi.org/10.1111/j.1365-3040.2010.02201.x>.
- Drigo B, Piji AS, Duyts H, Kielak AM, Gamper HA, Houtekamer MJ, Boschker HTS, Bodelier PLE, Whiteley AS, van Veen JA, and Kowalchuk GA (2010) Shifting carbon flow from roots into associated microbial communities in response to elevated atmospheric CO₂. *Proceedings of the National Academy of Sciences of the United States of America* 107: 10938–10942. <https://doi.org/10.1073/pnas.0912421107>.
- Eisenhauer N, Cesarz S, Koller R, Worm K, and Reich PB (2012) Global change belowground: Impacts of elevated CO₂, nitrogen, and summer drought on soil food webs and biodiversity. *Global Change Biology* 18: 435–447. <https://doi.org/10.1111/j.1365-2486.2011.02555.x>.
- Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, Singleton CM, Solden LM, Naas AE, Boyd JA, Hodgkins SB, Wilson RM, Trubel G, Li C, Frolking S, Pope PB, Wrighton KC, Crill PM, Chanton JP, Saleska SR, Tyson GW, Rich VI, and Sullivan MB (2018) Host-linked soil viral ecology along a permafrost thaw gradient. *Nature Microbiology* 3: 870–880. <https://doi.org/10.1038/s41564-018-0190-y>.
- Franco ALC, Gherardi LA, de Tomasel CM, Andriuzzi WS, Ankrom KE, Ashley Shaw E, Bach EM, Sala OE, and Wall DH (2019) Drought suppresses soil predators and promotes root herbivores in mesic, but not in xeric grasslands. *Proceedings of the National Academy of Sciences of the United States of America* 116: 12883–12888. <https://doi.org/10.1073/pnas.1900572116>.
- Gao D, Bai E, Li M, Zhao C, Yu K, and Hagedorn F (2020) Responses of soil nitrogen and phosphorus cycling to drying and rewetting cycles: A meta-analysis. *Soil Biology and Biochemistry* 148: 107896. <https://doi.org/10.1016/j.soilbio.2020.107896>.
- Harris E, Diaz-Pines E, Stoll E, Schlöter M, Schulz S, Duffner C, Li K, Moore KL, Ingrisch J, Reinthaler D, Zechmeister-Boltenstern S, Glatzel S, Brüggemann N, and Bahn M (2021) Denitrifying pathways dominate nitrous oxide emissions from managed grassland during drought and rewetting. *Science Advances* 7: eabb7118. <https://doi.org/10.1126/sciadv.abb7118>.
- Homet P, Gómez-Aparicio L, Matías L, and Godoy O (2021) Soil fauna modulates the effect of experimental drought on litter decomposition in forests invaded by an exotic pathogen. *Journal of Ecology* 109: 2963–2980. <https://doi.org/10.1111/1365-2745.13711>.
- Isobe K, Bouskill NJ, Brodie EL, Sudderth EA, and Martiny JHB (2020) Phylogenetic conservation of soil bacterial responses to simulated global changes. *Philosophical Transactions of the Royal Society, B: Biological Sciences* 375: 20190242. <https://doi.org/10.1098/rstb.2019.0242>.
- Joly FX, McAvoy E, and Subke JA (2021) Synergistic interactions between detritivores disappear under reduced rainfall. *Ecology* 102. <https://doi.org/10.1002/ecy.3299>.
- Meisner A, Leizeaga A, Rousk J, and Bååth E (2017) Partial drying accelerates bacterial growth recovery to rewetting. *Soil Biology and Biochemistry* 112: 269–276. <https://doi.org/10.1016/j.soilbio.2017.05.016>.
- Naylor D, Sadler N, Bhattacharjee A, Graham EB, Anderton CR, McClure R, Lipton M, Hofmockel KS, and Jansson JK (2020) Soil microbiomes under climate change and implications for carbon cycling. *Annual Review of Environment and Resources* 45: 29–59. <https://doi.org/10.1146/annurev-environ-012320-082720>.

- Nielsen UN and Ball BA (2015) Impacts of altered precipitation regimes on soil communities and biogeochemistry in arid and semi-arid ecosystems. *Global Change Biology* 21: 1407–1421. <https://doi.org/10.1111/gcb.12789>.
- Ochoa-Hueso R, Piñeiro J, and Power SA (2019) Decoupling of nutrient cycles in a Eucalyptus woodland under elevated CO₂. *Journal of Ecology* 107: 2532–2540. <https://doi.org/10.1111/1365-2745.13219>.
- Radujković D, Verbruggen E, Sigurdsson BD, Leblans NIW, Janssens IA, Vicca S, and Weedon JT (2018) Prolonged exposure does not increase soil microbial community compositional response to warming along geothermal gradients. *FEMS Microbiology Ecology* 94. <https://doi.org/10.1093/femsec/fix174>.
- Raynaud X and Nunan N (2014) Spatial ecology of bacteria at the microscale in soil. *PLoS One* 9: e87217.
- Rillig MC, Ryo M, Lehmann A, Aguilar-Trigueros CA, Buchert S, Wulf A, Iwasaki A, Roy J, and Yang G (2019) The role of multiple global change factors in driving soil functions and microbial biodiversity. *Science* 366: 886–890. <https://doi.org/10.1126/science.aay2832>.
- Schwarz B, Barnes AD, Thakur MP, Brose U, Ciobanu M, Reich PB, Rich RL, Rosenbaum B, Stefanski A, and Eisenhauer N (2017) Warming alters energetic structure and function but not resilience of soil food webs. *Nature Climate Change* 7: 895–900. <https://doi.org/10.1038/s41558-017-0002-z>.
- Séneca J, Pjevac P, Canarini A, Herbold CW, Zioutis C, Dietrich M, Simon E, Prommer J, Bahn M, Pötsch EM, Wagner M, Wanek W, and Richter A (2020) Composition and activity of nitrifier communities in soil are unresponsive to elevated temperature and CO₂, but strongly affected by drought. *ISME Journal* 14: 3038–3053. <https://doi.org/10.1038/s41396-020-00735-7>.
- Siebert J, Sünemann M, Auge H, Berger S, Cesarz S, Ciobanu M, Guerrero-Ramírez NR, and Eisenhauer N (2019) The effects of drought and nutrient addition on soil organisms vary across taxonomic groups, but are constant across seasons. *Scientific Reports* 9: 639. <https://doi.org/10.1038/s41598-018-36777-3>.
- Simon E, Canarini A, Martin V, Séneca J, Böckle T, Reinthaler D, Pötsch EM, Piepho HP, Bahn M, Wanek W, and Richter A (2020) Microbial growth and carbon use efficiency show seasonal responses in a multifactorial climate change experiment. *Communications Biology* 3. <https://doi.org/10.1038/s42003-020-01317-1>.
- Thakur MP, Reich PB, Hobbie SE, Stefanski A, Rich R, Rice KE, Eddy WC, and Eisenhauer N (2018) Reduced feeding activity of soil detritivores under warmer and drier conditions. *Nature Climate Change* 8: 75–78. <https://doi.org/10.1038/s41558-017-0032-6>.
- van Geffen KG, Berg MP, and Aerts R (2011) Potential macro-detritivore range expansion into the subarctic stimulates litter decomposition: A new positive feedback mechanism to climate change? *Oecologia* 167: 1163–1175. <https://doi.org/10.1007/s00442-011-2051-8>.
- van Groenigen KJ, Osenberg CW, Terrer C, Carrillo Y, Dijkstra FA, Heath J, Nie M, Pendall E, Phillips RP, and Hungate BA (2017) Faster turnover of new soil carbon inputs under increased atmospheric CO₂. *Global Change Biology* 1–10. <https://doi.org/10.1111/gcb.13752>.
- Veresoglou SD, Anderson IC, de Sousa NMF, Hempel S, and Rillig MC (2016) Resilience of fungal communities to elevated CO₂. *Microbial Ecology* 72: 493–495. <https://doi.org/10.1007/s00248-016-0795-8>.
- Walker TWN, Kaiser C, Strasser F, Herbold CW, Leblans NIW, Wobken D, Janssens IA, Sigurdsson BD, and Richter A (2018) Microbial temperature sensitivity and biomass change explain soil carbon loss with warming. *Nature Climate Change* 8: 885–889. <https://doi.org/10.1038/s41558-018-0259-x>.
- Wild B, Ambus P, Reinsch S, and Richter A (2018) Resistance of soil protein depolymerization rates to eight years of elevated CO₂, warming, and summer drought in a temperate heathland. *Biogeochemistry* 140: 255–267. <https://doi.org/10.1007/s10533-018-0487-1>.
- Zhou Z, Wang C, and Luo Y (2018) Response of soil microbial communities to altered precipitation: A global synthesis. *Global Ecology and Biogeography* 27: 1121–1136. <https://doi.org/10.1111/gcb.12761>.
- Zhou Z, Wang C, and Luo Y (2020) Meta-analysis of the impacts of global change factors on soil microbial diversity and functionality. *Nature Communications* 11: 3072. <https://doi.org/10.1038/s41467-020-16881-7>.
- Zhou Y, Sun B, Xie B, Feng K, Zhang Z, Zhang Z, Li S, Du X, Zhang Q, Gu S, Song W, Wang L, Xia J, Han G, and Deng Y (2021) Warming reshaped the microbial hierarchical interactions. *Global Change Biology*. <https://doi.org/10.1111/gcb.15891>.

Encyclopedia of SOILS IN THE ENVIRONMENT

Volume One: Soil Biology

EDITORS-IN-CHIEF

Michael J Goss

Professor Emeritus, School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G 2W1, CANADA

Margaret Oliver

Visiting professor, Soil Research Centre, Department of Geography and Environmental Science, Russell Building, University of Reading, Whiteknights, Reading, RG6 6DR, United Kingdom.

SECTION EDITORS

Dr Evgenia Blagodatskaya

Dept. of Soil Ecology, Helmholtz-Zentrum für Umweltforschung GmbH – UFZ, Theodor-Lieser Str. 4, D-06120 Halle (Saale), GERMANY

Professor Adrian Unc

Boreal Ecosystems and Agricultural Sciences, Grenfell Campus, Memorial University of Newfoundland, 20 University Drive, Corner Brook, Newfoundland A2H 5G4, CANADA

Dr. Siobhán Staunton

INRA, UMR Eco&Sols, Montpellier SupAgro, bâtiment 12, 2 Place Viala, 34060 Montpellier Cedex 2, Montpellier, FRANCE

Professor Baoshan Xing

University of Massachusetts – Amherst, College of Natural Sciences, Stockbridge School of Agriculture, Environmental & Soil Chemistry, Program, 410 Paige Laboratory, Amherst, MA 01003, USA

Dr Karin Mueller

New Zealand Institute for Plant & Food Research Ltd., Ruakura Research Centre, Private Bag 3123, Hamilton, New Zealand

Professor Christina Siebe

Instituto de Geología, Circuito de la Investigación Científica s/n, Ciudad Universitaria, Ciudad de México, CP 04510, México

Professor Rupert Baeumler

Institute of Geography, Friedrich-Alexander-University of Erlangen-Nuremberg, Wetterkreuz 15, 91058 Erlangen, Germany

Dr Uta Stockmann

Soil Processes and Function Organisational Unit, CSIRO Agriculture and Food, Black Mountain, GPO BOX 1700, Canberra ACT 2601, AUSTRALIA

Professor Paul Hallett

Room 2.23, Cruickshank Building, School of Biological Sciences, University of Aberdeen, Aberdeen, AB24 3UU, UNITED KINGDOM

Professor Dani Or

Department of Environmental Systems Science (D-USYS), Institute of Biogeochemistry and Pollutant Dynamics (IBP), Soil and Terrestrial Environmental Physics (STEP), CHN F 29.1 Universitaetstrasse 16, 8092 Zurich, SWITZERLAND



Set ISBN



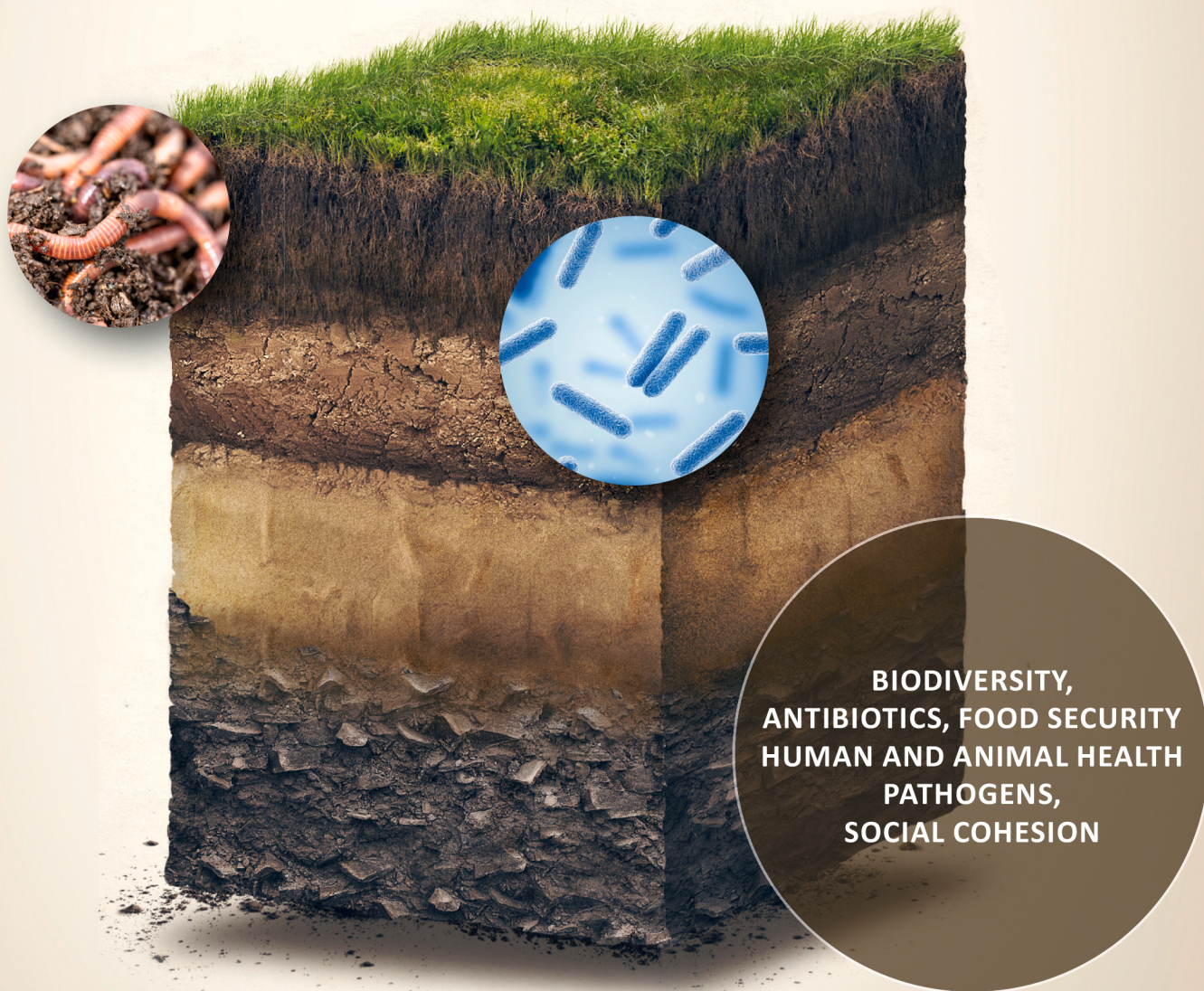
Volume ISBN



Second Edition

Encyclopedia of SOILS IN THE ENVIRONMENT

Volume Two: Chemistry



Editors-in-Chief

Michael Goss and Margaret Oliver

Section Editors

Siobhán Staunton & Baoshan Xing

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

SECOND EDITION

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

SECOND EDITION

EDITORS-IN-CHIEF

Professor Michael J Goss

Professor Emeritus, School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G 2W1, CANADA

Professor Margaret Oliver

Visiting Professor, Soil Research Centre, Department of Geography and Environmental Science, Russell Building, University of Reading, Whiteknights, Reading, RG6 6DR, United Kingdom

VOLUME 2

CHEMISTRY

SECTION EDITORS

Siobhán Staunton

INRA, UMR Eco&Sols, Montpellier SupAgro, bâtiment 12, 2 Place Viala, 34060 Montpellier Cedex 2, Montpellier, FRANCE

Baoshan Xing

University of Massachusetts – Amherst, College of Natural Sciences, Stockbridge School of Agriculture, Environmental & Soil Chemistry, Program, 410 Paige Laboratory, Amherst, MA 01003, USA



ELSEVIER

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge MA 02139, United States

Copyright © 2023 Elsevier Ltd. All rights reserved

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

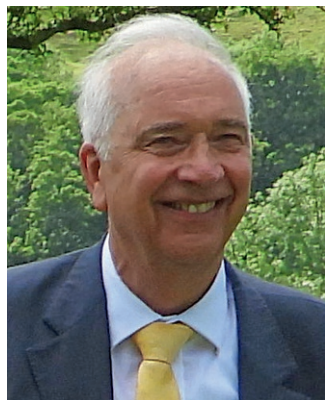
ISBN 978-0-12-822974-3

For information on all publications
visit our website at <http://store.elsevier.com>



Publisher: Oliver Walter
Acquisition Editor: Priscilla Braglia
Content Project Manager: Sona Chahal
Associate Content Project Manager: Nandhini Mahendran
Designer: Christian Bilbow

EDITORS BIOGRAPHY



Michael J Goss, PhD, MSc, BSc, is a professor emeritus in the School of Environmental Sciences at the University of Guelph. Before taking on the joint editor-in-chief responsibilities with Margaret Oliver, he held a similar position for *Soil Use and Management*, a journal of the British Society of Soil Science of which he is an honorary fellow. He was an inaugural member of the European Society for Agronomy. Michael started his career at the Letcombe Laboratory before moving to Rothamsted Experimental Station and the Macaulay Land Use Research Institute (now the James Hutton Research Institute). In 1990 he was appointed to the Research Chair in Land Stewardship at the University of Guelph and in 2005 became the director of its Regional Campus near Ottawa.

Michael's research has covered soil plant relations, soil management, and agriculture's footprint on the environment. His focus has been on the physical properties of soil and their impacts on root growth, which he showed to be influenced by the role of the root cap. He broadened his approach to include the effects of tillage and land drainage on root exploration. This work stimulated his interest in the transport of materials, such as plant nutrients, pathogenic microbes, antibiotics, and biologically active compounds through the soil and their contamination of water resources. Michael's knowledge in this field resulted in his involvement in the public inquiry into the death of seven people, including children, from a microbially contaminated drinking water supply. He contributed, both to the investigation into how the contamination occurred and if any fault could be attributed to local agricultural practices, and to how to prevent future events.

Michael also developed an interest in the interrelationships between soil microbes and root systems, particularly the tripartite symbiosis between roots of legumes, mycorrhizal fungi, and rhizobia. He has been an energetic member of collaborative research programs in France and Germany, and in Portugal, where he continues his interest in the role of mycorrhiza in sustainable land management.

Michael has authored more than 180 publications including two books, one in its 2nd edition, and edited two others.



Margaret Ann Oliver (BSc, PhD)

Margaret's interest in soil began in her penultimate year at school when learning about different types of soil and their formation. She chose the University of Bristol, United Kingdom for her first degree because the geography course offered soil science and she was also able to do a degree in geology alongside. Her love for both subjects remains undiminished. She did her PhD in soil spatial analysis at the University of Birmingham. Her early research focused on the multivariate and geostatistical analysis of soil data, which she continued to use in various research projects. Her research interests expanded to include the application of a wide range of numerical and geostatistical methods to soil and other data including pollen counts, forestry, radon emission, remotely sensed imagery, precision agriculture, and the incidence of childhood cancer. Her specific interests have been in sample design for spatial analysis and eventual mapping, risk analysis associated with prescribed thresholds in relation to soil pollutants or deficiencies of crop nutrients, and the relations between different scales of spatial variation. Her most

recent research was in the field of precision agriculture. At both the Universities of Birmingham and Reading, she taught topics in soil science, statistical methods, and spatial analysis. At Reading she had a research group working on various projects such as imagery, sampling and mapping, precision agriculture, and human health.

She has published more than 100 academic papers and has made several contributions to books. She is the co-author of three books: *Statistical Methods in Soil and Land Resource Survey*, Oxford University Press (Webster, R and Oliver, M. A. 1990); *Geostatistics for Environmental Scientists*, Second Edition, John Wiley & Sons (Webster, R and Oliver, M. A. 2007), and *How to do the Basic Steps in Geostatistics: The Variogram and Kriging*, Springer, Dordrecht (Oliver, M.A. & Webster, R.2015). She has edited two books: *Geostatistical Applications for Precision Agriculture* (Springer, Dordrecht, 2010) and *Precision Agriculture for Sustainability and Environmental Protection* (Earthscan from Routledge, 2013).

Margaret retired officially from Reading in 2004, but as yet has not managed to escape from academia. She has retained a link and a base at the University as a visiting professor since January 2005. After retirement, she developed and taught geostatistics courses to outside groups for several years. She then started on the long path of editorial work, as an associate editor of the *European Journal of Soil Science* and eventually the editor-in-chief, assistant editor of *Mathematical Geosciences*, co-editor of *Precision Agriculture* with John Stafford, and finally the editor-in-chief of the *Encyclopedia of Soils in the Environment* with Michael Goss.

Margaret was made an honorary member of the British Society of Soil Science in 2021 to reflect her commitment to soil science through fellow research, teaching, and editorial work. She was honored by the Pedometrics Commission of the International Union of Soil Science by having an award named after her in 2015. It is the Margaret Oliver award for young pedometricians.

EDITORIAL BOARD

Editors in Chief

Michael J Goss

Professor Emeritus, School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G 2W1, CANADA

Margaret Oliver

Visiting Professor, Soil Research Centre, Department of Geography and Environmental Science, Russell Building, University of Reading, Whiteknights, Reading, RG6 6DR, United Kingdom

Section Editors

Dr Evgenia Blagodatskaya

Dept. of Soil Ecology, Helmholtz-Zentrum für Umweltforschung GmbH – UFZ, Theodor-Lieser Str. 4, D-06120 Halle (Saale), GERMANY

Professor Adrian Unc

Boreal Ecosystems and Agricultural Sciences, Grenfell Campus, Memorial University of Newfoundland, 20 University Drive, Corner Brook, Newfoundland A2H 5G4, CANADA

Dr. Siobhán Staunton

INRA, UMR Eco&Sols, Montpellier SupAgro, bâtiment 12, 2 Place Viala, 34060 Montpellier Cedex 2, Montpellier, FRANCE

Professor Baoshan Xing

University of Massachusetts – Amherst, College of Natural Sciences, Stockbridge School of Agriculture, Environmental & Soil Chemistry, Program, 410 Paige Laboratory, Amherst, MA 01003, USA

Dr Karin Mueller

New Zealand Institute for Plant & Food Research Ltd., Ruakura Research Centre, Private Bag 3123, Hamilton, New Zealand

Professor Christina Siebe

Instituto de Geología, Circuito de la Investigación Científica s/n, Ciudad Universitaria, Ciudad de México, CP 04510, México

Professor Rupert Baeumler

Institute of Geography, Friedrich-Alexander-University of Erlangen-Nuremberg, Wetterkreuz 15, 91058 Erlangen, Germany

Dr Uta Stockmann

Soil Processes and Function Organisational Unit, CSIRO Agriculture and Food, Black Mountain, GPO BOX 1700 , Canberra ACT 2601, AUSTRALIA

Professor Paul Hallett

*Room 2.23, Cruickshank Building, School of Biological Sciences, University of Aberdeen, Aberdeen, AB24 3UU,
UNITED KINGDOM*

Professor Dani Or

*Department of Environmental Systems Science (D-USYS), Institute of Biogeochemistry and Pollutant Dynamics (IBP),
Soil and Terrestrial Environmental Physics (STEP), CHN F 29.1 Universitaetstrasse 16, 8092 Zurich, SWITZERLAND*

FOREWORD

Soil is the porous “skin” on the surface of the Earth that has formed over thousands of years of weathering, comprising minerals, organisms (both dead and alive), water, and air. Soil is more than just the sum of its parts. It serves several crucial ecological functions, from food and fiber production to water storage and purification to waste decomposition.

Increasingly we understand that everything is connected, and that soil is a “Critical Zone” that intersects with the hydrosphere (water), lithosphere (minerals), atmosphere (air), and biosphere (all living organisms). To quote Daniel Hillel, the 2012 World Food Prize laureate and editor of the First Edition of *The Encyclopedia of Soils in the Environment*, “[Soil] is the crucible of terrestrial life, within which biological productivity is generated and sustained... Our civilization depends on the soil...” He also promoted the importance of understanding soil to manage it effectively. This revised and reorganized second edition of the Encyclopedia aims to reflect the changing status of soil in the world. Within these five volumes, over 724 experts from many countries have provided updated, easy-to-understand chapters on soil, its properties, and its role in the environment to affect climate change, plant productivity, human health, and biodiversity.

Many technological advances have been made since the first edition of this encyclopedia that has allowed us to investigate soil physical, chemical, and especially biological properties more precisely, enabling us to identify and understand their roles in ecosystem services. The study of soils, especially in the environmental context, is critical to meeting the challenges of feeding, clothing, and cleaning up the world. The information presented in these volumes will be useful to many, including soil and environmental scientists, policymakers, land managers, educators, students, and anyone interested in learning more about soil.

Soil is under considerable threat from many sources—building and infrastructure, climate change, population pressure, and pollution, yet it is vital to agronomic productivity and continues to provide critical ecosystem services. Society needs to be engaged in how soil is managed and conserved. Most natural scientists are very dedicated and generous, including all of the authors in this Encyclopedia who have freely given their time to make these volumes so comprehensive. The study of soil, while often challenging, is fascinating—soil is a beautiful earthy material that can delight our senses with its variegated colors, gritty to smooth textures, and especially its characteristic aroma when freshly turned. Perhaps that is why gardeners and even some farmers tend to overwork their soil to its detriment—it *just feels good to get your hands in the dirt!*

Soil science is one of the most exciting, interesting, and rewarding disciplines, largely because it is multi-dimensional and interdisciplinary, basic and applied. It affects all of our lives. Thus, I encourage you to peruse these volumes and discover all kinds of interesting aspects of soil in the environment.

April Ulery

Former President, Soil Science Society of America
Professor, Soil & Environmental Sciences
New Mexico State University
Las Cruces, NM
The United States

PREFACE

This second edition of the *Encyclopedia of Soils in the Environment* follows in the footsteps of a highly rated first edition overseen by Professor Daniel Hillel as editor in chief, ably assisted by six editors. That edition provided a very valuable resource on the “state-of-the-art” in soil science in the context of the environment in the closing years of the twentieth century and the first three years of the current century. It set a standard that the editorial team for the second edition has aimed to match. The soil and environmental sciences have made huge strides in the almost 20 years since the first edition was published, and the new team has had to embrace these developments while retaining the basic components of these sciences. The project has taken 4 years to complete. It began with the appointment of the two editors-in-chief, professors Michael J. Goss and Margaret A. Oliver, followed by that of the nine subject editors and a pedology adviser.

The project was in its infancy (a few months only) when COVID-19 struck. The editors-in-chief discussed the format and structure of the second edition around the time that the first “lockdowns” around the world were being introduced. The first edition was alphabetically organized, whereas we decided to have a section-oriented approach in this edition, which we hope gives a more structured feel for the various topics embraced. The principal sections are soil biology, soil chemistry, the soil environment and its management, pedology (the origins and development of soil), pedometrics (the application of mathematics and statistics to soil), and soil physics. We were able to give our full attention to this development and to the selection of the scientists to join us on the road ahead. Furthermore, we were fortunate to choose a team that has maintained its enthusiasm despite the many tribulations we have faced along this path. The team comprises Jane Blagodatskaya and Adrian Unc (biology and microbiology), Siobhan Staunton and Baoshan Xing (chemistry), Karin Müller and Christina Siebe (environment and management), Rupert Bäumler (pedology), Uta Stockmann and Margaret Oliver (pedometrics), and Paul Hallett and Dani Or (physics). The editorial board has been mindful in their choice of subjects of the need for basic soil science and for state-of-the-art developments, with the aim that the audience for these volumes can be as large as possible.

During the gestation period of the Encyclopedia, methods of working for those in academia were turned on their heads. The change from “face-to-face” teaching to formulating and establishing online courses, initially for lockdown periods but then for much of the next 2 years, had huge consequences for academic staff worldwide. This meant that they had less time to devote to outside activities, such as writing chapters for an encyclopedia. As a result, some authors were unable to complete their chapters in time, leaving a few unfortunate gaps. The whole team has done as much as they can to limit any adverse effects, but we ask readers to appreciate the struggle that authors have had with illness (personal, within their families and work colleagues) and a great increase in their workload. Despite this, we are presenting a comprehensive document that should stand the test of the next 20 years of soils in the environment.

This century has seen an almost unprecedented interest in the soil and environmental sciences. It has been triggered not only by issues that the Earth is facing from global climate change, increasing population pressure, soil and land degradation, and food insecurity but also by the many new and sophisticated methods available, such as the DNA determination of microbial populations and their functional capabilities, proximal sensing, machine learning methods, and an ever-increasing array of technological tools. These developments have greatly increased our knowledge and appreciation of the complexity of soils, which has changed our understanding of some well-established aspects of the science. The soil and its environment are at the critical interface of the atmosphere, biosphere, hydrosphere, and lithosphere. They deserve to be taken seriously by the international community in the same way that air and water have been. Governments are now paying attention

to the soil and environment, and a legal and governance framework is emerging to look after these parts of our life support resources.

We have aimed, through our editing of chapters, to make the language of the chapters accessible to all, regardless of educational background. This does not mean that we have diminished the scientific content, but that readers should be able to gain the foothold they desire in these subjects to progress to greater knowledge and understanding of them. We have retained the overview of soil written by Daniel Hillel for the First Edition, first to honor his great contribution to soil science and second to demonstrate the significance of soil. It follows this Preface.

*Michael J. Goss
Margaret A. Oliver*

SOIL

The term “soil” refers to the weathered and fragmented outer layer of our planet’s land surfaces. Formed initially through the physical disintegration and chemical alteration of rocks and minerals by physical and biogeochemical processes, soil is influenced by the activity and accumulated residues of a myriad of diverse forms of life. As it occurs in different geologic and climatic domains, soil is an exceedingly varied body with a wide range of attributes.

Considering the height of the atmosphere, the thickness of the earth’s rock mantle, and the depth of the ocean, one observes that soil is an amazingly thin body—typically not much more than one meter thick and often less than that. Yet it is the fount of terrestrial life, within which biological productivity is generated and sustained. It acts like a composite living entity, a home to a community of innumerable microscopic and macroscopic plants and animals. A mere fistful of soil typically contains billions of microorganisms, which perform vital interactive biochemical functions. Another intrinsic attribute of the soil is its sponge-like porosity and its enormous internal surface area. That same fistful of soil may actually consist of several hectares of active surface, upon which physicochemical processes take place continuously.

Realizing humanity’s utter dependence on the soil, ancient peoples, who lived in greater closeness to nature than many of us today, actually revered the soil. It was not only their source of livelihood but also the material from which they built their homes and that they learned to shape, heat, and fuse into household vessels and writing tablets (ceramics, made of clayey soil, being the first synthetic material in the history of technology). In the Bible, the name assigned to the first human was Adam, derived from “adama,” meaning soil. The name given to the first earthling’s mate was Hava (Eve, in transliteration), meaning “living” or “life-giving.” Together, therefore, Adam and Eve signified quite literally “Soil and Life.”

The same powerful metaphor is echoed in the Latin name for human species—Homo, derived from humus, the material of the soil. Hence, the adjective “human” also implies “of the soil.” Other ancient cultures evoked equally powerful associations. To the Greeks, the earth was a manifestation of Gaea, the maternal goddess who, impregnated by Uranus (god of the sky), gave birth to all the gods of the Greek pantheon.

Our civilization depends on the soil more crucially than ever because our numbers have grown while available soil resources have diminished and deteriorated. Paradoxically, however, even as our dependence on the soil has increased, most of us have become physically and emotionally detached from it. Many of the people in the so-called developed countries spend their lives in the artificial environment of a city, insulated from direct exposure to nature, and some children may now assume as a matter of course that food originates in supermarkets.

Detachment has bred ignorance, and out of ignorance has come the delusion that our civilization has risen above nature and has set itself free of its constraints. Agriculture and food security, erosion and salination, degradation of natural ecosystems, depletion and pollution of surface waters and aquifers, and decimation of biodiversity—all these processes, which involve the soil directly or indirectly, have become abstractions to many people. The very language we use betrays disdain for that common material underfoot, often referred to as “dirt.” Some fastidious parents prohibit their children from playing in the mud and hurry to wash their “soiled” hands when the children nonetheless obey an innate instinct to do so. Thus, soil is devalued and treated as unclean though it is the terrestrial realm’s principal medium of purification, wherein wastes are decomposed and nature’s productivity is continually rejuvenated.

Scientists who observe soil closely see it in effect as a seething foundry in which matter and energy are in constant flux. Radiant energy from the sun streams onto the field and cascades through the soil and the plants growing in it. Heat is exchanged, water percolates through the soil’s intricate passages, and plant roots extract

water and transmit it to their leaves, which transpire it back to the atmosphere. Leaves absorb carbon dioxide from the air and synthesize it with soil-derived water to form the primary compounds of life. Oxygen emitted by the leaves makes the air breathable for animals, which consume and in turn fertilize plants.

Soil is thus a self-regulating bio-physio-chemical factory, processing its own materials, water, and solar energy. It also determines the fate of rainfall and snowfall reaching the ground surface—whether the water thus received will flow over the land as runoff, or seep downward to the subterranean reservoir called groundwater, which in turn maintains the steady flow of springs and streams. With its finite capacity to absorb and store moisture, and to release it gradually, the soil regulates all these phenomena. Without the soil as a buffer, rain falling over the continents would run off entirely, producing violent floods rather than sustained river flow.

Soil naturally acts as a living filter, in which pathogens and toxins that might otherwise accumulate to foul the terrestrial environment are rendered harmless. Since time immemorial, humans and other animals have been dying of all manner of diseases and have then been buried in the soil, yet no major human disease is transmitted by it. The term antibiotic was coined by soil microbiologists who, as a consequence of their studies of soil bacteria, specifically actinomycetes, discovered streptomycin (an important cure for tuberculosis and other infections). Ion exchange, a useful process of water purification, also was discovered by soil scientists studying the passage of solutes through beds of clay.

However unique in form and function, soil is not an isolated body. It is, rather, a central link in the larger chain of interconnected domains and processes comprising the terrestrial environment. The soil interacts both with the overlying atmosphere and the underlying strata, as well as with surface and underground bodies of water. Especially important is the interrelation between the soil and climate. In addition to its function of regulating the cycle of water, it also regulates energy exchange and surface temperature.

When land is cleared of vegetation and turned into a cultivated field, the native biomass above the ground is often burned and the organic matter within the soil tends to decompose. These processes release carbon dioxide into the atmosphere, thus contributing to the Earth's greenhouse effect and to global warming. On the other hand, the opposite act of afforestation and soil enrichment with organic matter, such as can be achieved through conservation management, may serve to absorb carbon dioxide from the atmosphere. To an extent, the soil's capacity to store carbon thus helps to mitigate the greenhouse effect.

Thousands of years are required for nature to create life-giving soil out of sterile bedrock. In only a few decades, however, unknowing or uncaring humans can destroy that wondrous work of nature. In various circumstances, mismanaged soils may be subject to erosion (the sediments of which tend to clog streambeds, estuaries, lakes, and coastal waters), to leaching of nutrients with attendant loss of fertility and eutrophication of water bodies, to waterlogging and impaired aeration, or to an excessive accumulation of salts that may cause a once-productive soil to become entirely sterile. Such processes of soil degradation, sometimes called "desertification," already affect large areas of land.

We cannot manage effectively and sustainably that which we do not know and thoroughly understand. That is why the tasks of developing and disseminating sound knowledge of the soil and its complex processes have assumed growing urgency and importance. The global environmental crisis has created a compelling need for a concentrated, concise, and definitive source of information—accessible to students, scientists, practitioners, and the general public—about the soil in all its manifestations in nature and in relation to the life of humans.

Daniel Hillel
May 2004

LIST OF CONTRIBUTORS FOR VOLUME 2

Franco Ajmone Marsan

Università di Torino, DISAFA—Chimica agraria e pedologia, Grugliasco, Torino, Italy

Emmanuel Ampiah

Department of Chemistry, Louisiana State University, Baton Rouge, LA, United States; Louisiana State University Superfund Research Center, Baton Rouge, LA, United States

Emmanuel Arthur

Department of Agroecology, Aarhus University, Tjele, Denmark

Elisabetta Barberis

Soil and Biogeochemistry, Department of Agricultural, Forest and Food Sciences, Grugliasco, Torino, Italy

Enrique Barriuso

Université Paris-Saclay, INRAE, AgroParisTech, UMR ECOSYS, Palaiseau, France

Asher Bar-Tal

Institute of Soil, Water and Environmental Sciences, ARO, Volcani Institute, Rishon LeZion, Israel

Carole Bedos

Université Paris-Saclay, INRAE, AgroParisTech, UMR ECOSYS, Palaiseau, France

Pierre Benoit

Université Paris-Saclay, INRAE, AgroParisTech, UMR ECOSYS, Palaiseau, France

Isabelle Bertrand

UMR Eco&Sols, INRAE, Montpellier, France

Luis Carlos Leva Borduchi

Embrapa Instrumentação, São Carlos, SP, Brazil

Guilhem Bourrié

Académie d'Agriculture de France, Paris, France

Euan P Burford

Geomicrobiology Group, School of Life Sciences, University of Dundee, Dundee, Scotland, United Kingdom

Marta Camps-Arbestain

School of Agriculture and Environment, Massey University, Palmerston, New Zealand

Luisella Celi

Soil and Biogeochemistry, Department of Agricultural, Forest and Food Sciences, Grugliasco, Torino, Italy

Quan Chen

Yunnan Provincial Key Laboratory of Soil Carbon Sequestration and Pollution Control, Faculty of Environmental Science & Engineering, Kunming University of Science & Technology, Kunming, P. R. China

J Chorover

Department of Environmental Science, The University of Arizona, Tucson, AZ, United States

Robert L Cook

Department of Chemistry, Louisiana State University, Baton Rouge, LA, United States; Louisiana State University Superfund Research Center, Baton Rouge, LA, United States

Hannah V Cooper

School of Water, Energy and Environment, Cranfield University, Cranfield, United Kingdom; Agricultural and Environmental Sciences, School of Biosciences, Sutton Bonington Campus, University of Nottingham, Nottingham, United Kingdom

M Francesca Cotrufo

Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, United States

Zhongmin Dai

Institute of Soil & Water Resources and Environmental Science, College of Environmental and Resource Sciences, Zhejiang Provincial Key Laboratory of Agricultural Resources and Environment, Zhejiang University, Hangzhou, China

Anupam Das

Department of Chemistry, The State University of New York at Binghamton, Binghamton, NY, United States

Diego Victor de Babos
Embrapa Instrumentação, São Carlos, SP, Brazil

Bruno Delvaux
*Université catholique de Louvain (UCLouvain),
Earth and Life Institute, Soil Science, Louvain-La-Neuve,
Belgium*

Zhang Di
*Yunnan Provincial Key Laboratory of Soil Carbon
Sequestration and Pollution Control, Faculty of
Environmental Science & Engineering, Kunming
University of Science & Technology, Kunming, P. R. China*

Stefan H Doerr
*Department of Geography, Swansea University, Swansea,
United Kingdom*

Camille Dumat
*DYNAFOR, Dynamiques et écologie des paysages
agriforestiers, UMR INRAE, INP, Toulouse, France*

Trolard Fabienne
*INRAE—UMR 1114 Environnement Méditerranéen et
Modélisation des Agro-hydrosystèmes—Bâtiment
Climat—AgroParis, Paris, France; Académie d'Agriculture
de France, Paris, France*

Geoffrey Michael Gadd
*Geomicrobiology Group, School of Life Sciences, University
of Dundee, Dundee, Scotland, United Kingdom*

Edgar Galicia-Andrés
*Department of Forest and Soil Sciences, Institute of Soil
Research, University of Natural Resources and Life Sciences
Vienna, Vienna, Austria; Department of Material Sciences
and Process Engineering, Institute of Molecular Modeling
and Simulation, University of Natural Resources and Life
Sciences Vienna, Vienna, Austria*

Martin H Gerzabek
*Department of Forest and Soil Sciences, Institute of Soil
Research, University of Natural Resources and Life Sciences
Vienna, Vienna, Austria*

Saikat Ghosh
*Yunnan Provincial Key Laboratory of Soil Carbon
Sequestration and Pollution Control, Faculty of
Environmental Science & Engineering, Kunming
University of Science & Technology, Kunming, P. R. China*

Nicholas T Girkin
*School of Water, Energy and Environment, Cranfield
University, Cranfield, United Kingdom; Agricultural and
Environmental Sciences, School of Biosciences, Sutton
Bonington Campus, University of Nottingham,
Nottingham, United Kingdom*

Susan M Glasauer
*School of Environmental Sciences, University of Guelph,
Guelph, ON, Canada*

Hero T Gollany
*USDA-Agricultural Research Service (USDA-ARS),
Columbia Plateau Conservation Research Center, Adams,
OR, United States*

Cheng Gu
*State Key Laboratory of Pollution Control and Resource
Reuse, School of the Environment, Nanjing University,
Nanjing, Jiangsu, P. R. China*

Damien Guinoiseau
Université Paris Saclay, UMR GEOPS, Orsay, France

Huiyuan Guo
*Department of Chemistry, The State University of
New York at Binghamton, Binghamton, NY, United States*

Shuai Guo
*State Key Laboratory of Marine Environmental Science of
China, Xiamen University, Xiamen, PR China; College of
Environment and Ecology, Xiamen University, Xiamen, PR
China*

Michelle L Haddix
*Department of Soil and Crop Sciences, Colorado State
University, Fort Collins, CO, United States*

Randall W Hall
*Department of Natural Sciences and Mathematics,
Dominican University of California, San Rafael,
CA, United States; Louisiana State University
Superfund Research Center, Baton Rouge, LA, United
States*

Rob Harrison
*School of Environmental and Forest Sciences, University of
Washington, Seattle, WA, United States*

Syuntaro Hiradate
*Laboratory of Soil Science, Division of Bioproduction
Environmental Sciences, Department of
Agro-environmental Sciences, Faculty of Agriculture,
Kyushu University, Fukuoka, Japan*

Carmen Hoeschen
*Technical University of Munich, TUM School of Life
Sciences, Soil Science, Freising-Weihenstephan, Germany*

Md Tofail Hossain
*Centre for Environmental Remediation (GCER), College of
Engineering, Science and Environment, The University of
Newcastle, Callaghan, NSW, Australia; CRC for
Contamination Assessment and Remediation of the
Environment (crcCARE), The University of Newcastle,
Callaghan, NSW, Australia; Department of Agricultural
Botany, Sher-e-Bangla Agricultural University, Dhaka,
Bangladesh*

Shiwen Hu
*State Environmental Protection Key Laboratory of
Integrated Surface Water-Groundwater Pollution Control,*

*School of Environmental Science and Engineering,
Southern University of Science and Technology, Shenzhen,
China*

*Qiaoyun Huang
College of Resources and Environment, State Key
Laboratory of Agricultural Microbiology, Huazhong
Agricultural University, Wuhan, China*

*Afsana Jahan
Soil Science Division, Bangladesh Rice Research Institute,
Gazipur, Bangladesh*

*Jason James
Exponent Inc., Environmental and Earth Science Practice,
Bellevue, WA, United States*

*Sangho Jeon
National Institute of Agricultural Sciences, Rural
Development Administration, Wanju, Republic
of Korea*

*Klaus Kaiser
Soil Science and Soil Protection, Martin Luther University
Halle-Wittenberg, Halle (Saale), Germany*

*Kristýna Kantnerová
University of Colorado Boulder, Geological Sciences &
INSTAAR, Boulder, CO, United States*

*Aleksandr Kavetskiy
USDA-Agricultural Research Service, National Soil
Dynamics Laboratory, Auburn, AL, United States*

*Sana Khalid
Department of Environmental Sciences, COMSATS
University Islamabad, Islamabad, Pakistan*

*Anahita Khosravi
University of Massachusetts, Amherst, Massachusetts,
United States*

*Markus Kleber
Department of Crop and Soil Science, Oregon State
University, Corvallis, OR, United States*

*Heike Knicker
Department of Biogeochemistry, Vegetal and Microbial
Ecology, Instituto de Recursos Naturales y Agrobiología de
Sevilla (IRNAS-CSIC), Sevilla, Spain*

*Ingrid Koegel-Knabner
Technical University of Munich, TUM School of Life
Sciences, Soil Science, Freising-Weihenstephan, Germany;
Technical University of Munich, Institute of Advanced
Study, Garching, Germany*

*Gautier Landrot
Synchrotron SOLEIL, L'Orme des Merisiers, Saint-Aubin,
France*

*Di Lang
Yunnan Provincial Key Laboratory of Soil Carbon
Sequestration and Pollution Control, Faculty of
Environmental Science & Engineering, Kunming
University of Science & Technology, Kunming, P. R. China*

*Jocelyn Lavallee
Department of Soil and Crop Sciences, Colorado State
University, Fort Collins, CO, United States; Environmental
Defense Fund, New York, NY, United States*

*Sam J Leuthold
Department of Soil and Crop Sciences, Colorado State
University, Fort Collins, CO, United States*

*Guy J Levy
Institute of Soil, Water and Environmental Sciences, ARO,
Volcani Institute, Rishon LeZion, Israel*

*Honghong Li
School of History and Geography, Minnan Normal
University, Zhangzhou, Fujian Province, PR China*

*Rong Li
The Key Lab of Pollution Control and Ecosystem Restoration
in Industry Clusters, Ministry of Education, School of
Environment and Energy, South China University of
Technology, Guangzhou, People's Republic of China*

*Yong Li
Institute of Soil & Water Resources and Environmental
Science, College of Environmental and Resource Sciences,
Zhejiang Provincial Key Laboratory of Agricultural
Resources and Environment, Zhejiang University,
Hangzhou, China*

*Zimin Li
Université catholique de Louvain (UCLouvain), Earth and
Life Institute, Soil Science, Louvain-La-Neuve, Belgium*

*Dongli Liang
College of Natural Resources and Environment, Northwest
A&F University, Yangling, Shaanxi, China; Key Laboratory
of Plant Nutrition and the Agri-environment in Northwest
China, Ministry of Agriculture and Rural Affairs, Yangling,
Shaanxi, China*

*Xia Liu
Institute of Coastal Environmental Pollution Control, Key
Laboratory of Marine Environment and Ecology, Ministry
of Education, Frontiers Science Center for Deep Ocean*

Multispheres and Earth System, Ocean University of China, Qingdao, P. R. China

Yang Liu

Yunnan Provincial Key Laboratory of Soil Carbon Sequestration and Pollution Control, Faculty of Environmental Science & Engineering, Kunming University of Science & Technology, Kunming, P. R. China

Yang Lu

South China Institute of Environmental Sciences, Ministry of Ecology and Environment (MEE), Guangzhou, People's Republic of China

Yu Luo

Institute of Soil & Water Resources and Environmental Science, College of Environmental and Resource Sciences, Zhejiang Provincial Key Laboratory of Agricultural Resources and Environment, Zhejiang University, Hangzhou, China

Daniel H Lysak

Environmental NMR Centre, University of Toronto Scarborough, Toronto, ON, Canada

Bueno Maïté

Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, Institut des Sciences Analytiques et de Physico-Chimie pour l'Environnement et les Matériaux, UMR 5254, Pau, France

Mery Malandrino

Department of Chemistry, Università di Torino, Torino, Italy

Antonio P Mallarino

Iowa State University, Ames, IA, United States

Brendan Malone

CSIRO Agriculture and Food, Black Mountain, ACT, Australia

Laure Mamy

Université Paris-Saclay, INRAE, AgroParisTech, UMR ECOSYS, Palaiseau, France

Maria Martin

Soil and Biogeochemistry, Department of Agricultural, Forest and Food Sciences, Grugliasco, Torino, Italy

Ladislau Martin-Neto

Embrapa Instrumentação, São Carlos, SP, Brazil

Jorge Mataix-Solera

Department of Agrochemistry and Environment, Environmental Soil Science Group, Miguel Hernández University, Alicante, Spain

Timothy I McLaren

School of Agriculture and Food Sciences, The University of Queensland, St Lucia, QLD, Australia

Shamim Mia

Department of Agronomy, Patuakhali Science and Technology University, Patuakhali, Bangladesh

Déborá Marcondes Bastos Pereira Milori

Embrapa Instrumentação, São Carlos, SP, Brazil

Budiman Minasny

Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Eveleigh, NSW, Australia

Carsten W Mueller

Department of Geosciences and Natural Resource Management, University of Copenhagen, Copenhagen, Denmark

Natasha

Department of Environmental Sciences, COMSATS University Islamabad, Islamabad, Pakistan

Wartini Ng

Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Eveleigh, NSW, Australia

Astrid Oberson

Institute of Agricultural Sciences, ETH Zurich, Lindau, Switzerland

Chris Oostenbrink

Department of Material Sciences and Process Engineering, Institute of Molecular Modeling and Simulation, University of Natural Resources and Life Sciences Vienna, Vienna, Austria

Bo Pan

Yunnan Provincial Key Laboratory of Soil Carbon Sequestration and Pollution Control, Faculty of Environmental Science & Engineering, Kunming University of Science & Technology, Kunming, P. R. China

Derek Peak

Department of Soil Science, College of Agriculture and Bioresources, University of Saskatchewan, Saskatoon, SK, Canada

Kurt D Pennell

School of Engineering, Brown University, Providence, RI, United States

Joseph J Pignatello

Department of Environmental Sciences, The Connecticut Agricultural Experiment Station, New Haven, CT, United States

Chiara Pistocchi

UMR Eco&Sols, INRAE, Montpellier, France

Stephen A Prior
 USDA-Agricultural Research Service, National Soil
 Dynamics Laboratory, Auburn, AL, United States

Mingxing Qi
 College of Natural Resources and Environment, Northwest
 A&F University, Yangling, Shaanxi, China

Cécile Quantin
 Université Paris Saclay, UMR GEOPS, Orsay, France

Faiz Rabbani
 Department of Environmental Sciences, COMSATS
 University Islamabad, Islamabad, Pakistan

Mohammad Mahmudur Rahman
 Centre for Environmental Remediation (GCER), College of
 Engineering, Science and Environment, The University of
 Newcastle, Callaghan, NSW, Australia; CRC for
 Contamination Assessment and Remediation of the
 Environment (crcCARE), The University of Newcastle,
 Callaghan, NSW, Australia

Ravi Naidu
 Centre for Environmental Remediation (GCER), College of
 Engineering, Science and Environment, The University of
 Newcastle, Callaghan, NSW, Australia; CRC for
 Contamination Assessment and Remediation of the
 Environment (crcCARE), The University of Newcastle,
 Callaghan, NSW, Australia

Justin B Richardson
 Department of Geosciences, University of Massachusetts
 Amherst, Amherst, MA, United States

Xingmin Rong
 College of Resources and Environment, State Key
 Laboratory of Agricultural Microbiology, Huazhong
 Agricultural University, Wuhan, China

Donald S Ross
 Department of Plant & Soil Science, Jeffords Hall,
 University of Vermont, Burlington, VT, United States

Cristina Santín
 Research Institute of Biodiversity (IMIB), Mieres, Spain;
 Department of Biosciences, Swansea University, Swansea,
 United Kingdom

Nikolay Sargsyan
 Department of Computer Science and Software
 Engineering, Auburn University, Auburn, AL, United
 States

Md Imran Ullah Sarkar
 Centre for Environmental Remediation (GCER), College of
 Engineering, Science and Environment, The University of
 Newcastle, Callaghan, NSW, Australia; CRC for

Contamination Assessment and Remediation of the
 Environment (crcCARE), The University of
 Newcastle, Callaghan, NSW, Australia; Soil Science
 Division, Bangladesh Rice Research Institute, Gazipur,
 Bangladesh

CP Schulthess
 University of Connecticut, Storrs, CT, United States

Arthur Paul Schwab
 Soil and Crop Sciences, Texas A&M University, College
 Station, TX, United States

Giorgio S Senesi
 Consiglio Nazionale delle Ricerche (CNR), Istituto per la
 Scienza e Tecnologia dei Plasmi (ISTP), Bari, Italy

Nicola Senesi
 Dipartimento di Scienza del Suolo, della Pianta e degli
 Alimenti, Università di Bari "Aldo Moro", Bari, Italy

Muhammad Shahid
 Department of Environmental Sciences, COMSATS
 University Islamabad, Islamabad, Pakistan

Isaac Shainberg
 Institute of Soil, Water and Environmental Sciences, ARO,
 Volcani Institute, Rishon LeZion, Israel

Zhenqing Shi
 The Key Lab of Pollution Control and Ecosystem
 Restoration in Industry Clusters, Ministry of Education,
 School of Environment and Energy, South China
 University of Technology, Guangzhou, People's Republic of
 China

Andre J Simpson
 Environmental NMR Centre, University of Toronto
 Scarborough, Toronto, ON, Canada

Myrna J Simpson
 Environmental NMR Centre, University of Toronto
 Scarborough, Toronto, ON, Canada

Balwant Singh
 Sydney Institute of Agriculture, School of Life and
 Environmental Sciences, The University of Sydney, NSW,
 Australia

Ulf Skyllberg
 Department of Forest Ecology and Management,
 Swedish University of Agricultural Sciences, Umeå,
 Sweden

Donald L Sparks
 Department of Plant and Soil Sciences, University of
 Delaware, Interdisciplinary Science and Engineering (ISE)
 Laboratory, Newark, DE, United States

Phillip T Sprunger

Department of Physics and Astronomy, Louisiana State University, Baton Rouge, LA, United States; Louisiana State University Superfund Research Center, Baton Rouge, LA, United States

Siobhan Staunton

UMR Eco&Sols, INRAE, INRAE-IRD-Cirad-Institut Agro-Université de Montpellier, Montpellier, France

Daniel G Strawn

University of Idaho, Moscow, ID, United States

Federica Tamburini

Institute of Agricultural Sciences, ETH Zurich, Lindau, Switzerland

Svensson Teresia

Department of Thematic Studies—Environmental Change, Linköping University, Linköping, Sweden

H Allen Torbert

USDA-Agricultural Research Service, National Soil Dynamics Laboratory, Auburn, AL, United States

Markus Tuller

Department of Environmental Science, The University of Arizona, Tucson, AZ, United States

Daniel Tunega

Department of Forest and Soil Sciences, Institute of Soil Research, University of Natural Resources and Life Sciences Vienna, Vienna, Austria; School of Pharmaceutical Science and Technology, Tianjin University, Tianjin, China

Paulino Ribeiro Villas-Boas

Embrapa Instrumentação, São Carlos, SP, Brazil; USDA-Agricultural Research Service (USDA-ARS), Columbia Plateau Conservation Research Center, Adams, OR, United States; Instituto de Física de São Carlos, São Carlos, SP, Brazil

Rota Wagai

Institute for Agro-Environmental Science, National Agriculture and Food Research Organization, Tsukuba, Ibaraki, Japan

Chao Wang

State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University, Nanjing, Jiangsu, P. R. China

Guo Wang

College of resources and environment, Fujian Agriculture and Forestry University, Fuzhou, Fujian Province, PR China

Chaoxian Wei

State Key Laboratory of Marine Environmental Science of China, Xiamen University, Xiamen, PR China; College of Environment and Ecology, Xiamen University, Xiamen, PR China

Roland A Werner

Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland

H Wijnja

Massachusetts Department of Agricultural Resources, Southborough, MA, United States

Wenhao Wu

Department of Environmental Science, Zhejiang University, Hangzhou, China

Baoshan Xing

University of Massachusetts, Amherst, Massachusetts, United States

Jianming Xu

Institute of Soil & Water Resources and Environmental Science, College of Environmental and Resource Sciences, Zhejiang Provincial Key Laboratory of Agricultural Resources and Environment, Zhejiang University, Hangzhou, China

Galina Yakubova

USDA-Agricultural Research Service, National Soil Dynamics Laboratory, Auburn, AL, United States

Kun Yang

Department of Environmental Science, Zhejiang University, Hangzhou, China

Puu-Tai Yang

Institute for Agro-Environmental Science, National Agriculture and Food Research Organization, Tsukuba, Ibaraki, Japan

W Yang

California Air Resources Board, Sacramento, CA, United State

Thiry Yves

Andra, Research and Development Division, Châtenay-Malabry, France

Oelmann Yvonne

Geocology, University of Tübingen, Tübingen, Germany

Yong Zhang

State Key Laboratory of Marine Environmental Science of China, Xiamen University, Xiamen, PR China; College of

Environment and Ecology, Xiamen University, Xiamen, PR China

Jian Zhao

Institute of Coastal Environmental Pollution Control, Key Laboratory of Marine Environment and Ecology, Ministry of Education, Frontiers Science Center for Deep Ocean Multispheres and Earth System, Ocean University of China, Qingdao, P. R. China; Laboratory for Marine Ecology and Environmental Science, Qingdao National

Laboratory for Marine Science and Technology, Qingdao, China

Fei Zhou

College of Natural Resources and Environment, Northwest A&F University, Yangling, Shaanxi, China

Hongxia Zhu

Department of Environmental Science, Zhejiang University, Hangzhou, China

CONTENTS OF ALL VOLUMES

VOLUME 1: SOIL BIOLOGY

Overview

Edaphology	1
<i>RM Goss and AL Ulery</i>	

Biodiversity	10
<i>Monica A Farfan and Diana H Wall</i>	

Soil organisms

Bacteriophage (Viruses)	16
<i>Li Bi, Ji-Zheng He, Dan-Ting Yu, Li-Li Han, and Hang-Wei Hu</i>	

Archaea	25
<i>Mikhail V Semenov</i>	

Bacteria	31
<i>Mikhail V Semenov</i>	

Soil protists	39
<i>Stefan Geisen, Nathalie Amacker, and Alexandre Jousset</i>	

Fungi including mycorrhizal fungi	49
<i>F Buscot</i>	

Soil algae	61
<i>Karen Baumann and Patrick Jung</i>	

Soil arthropods: Underfoot and all around	70
<i>Ernest C Bernard</i>	

Nematodes	105
<i>Deborah A Neher and Thomas O Powers</i>	

Earthworms: Essential ecosystem engineers providing vital ecosystem services	112
<i>Maria JI Briones</i>	

Complex structures

Soil biofilms: Evolving concepts and ecological functions	120
<i>Yichao Wu, Marc Redmile-Gordon, Peng Cai, Chunhui Gao, and Qiaoyun Huang</i>	

Biocrusts	127
<i>Matthew A Bowker, Jasmine Anenberg, Hilina Yohannes Kebede, Anita Antoninka, and Lydia Bailey</i>	

Functioning

Studying soil microbial communities: Culture-dependent approaches	139
<i>Eduardo K Mitter, Dasiel Obregon, Micaela Tosi, Heather Slinn, and Kari Dunfield</i>	

Studying soil microbial communities: Culture-independent approaches	146
<i>Tolulope Mafa-Attoye, Heather Slinn, Micaela Tosi, Evan Mayer, Dasiel Obregon, and Kari Dunfield</i>	

Studying microbial function in soil: Culture-independent techniques <i>Micaela Tosi, Dasiel Obregon, Heather Slinn, and Kari Dunfield</i>	155
Kinetics of microbial processes: General principles <i>NS Panikov</i>	168
Kinetics of microbial processes: Applications to soil biology <i>NS Panikov</i>	186
Biogeochemical cycles	
Carbon cycle in soils: Dynamics and management <i>Charles W Rice, Carlos B Pires, and Marcos VM Sarto</i>	219
Carbon storage in soils <i>Isabelle Basile-Doelsch, Tiphaine Chevallier, Marie-France Dignac, and Amandine Erktan</i>	228
Soil organic nitrogen <i>EJ Daly and KA Congreves</i>	243
Nitrification: Process, organisms and management <i>Jeanette M Norton and Yang Ouyang</i>	253
Microbial denitrification <i>Michael J Goss, Sergey Blagodatsky, and Adrian Unc</i>	261
Symbiotic nitrogen fixation <i>Florence Mus and Hsin-Hua Wu</i>	270
Nitrogen: Non-Symbiotic Nitrogen fixation in soils <i>Vadakattu VSR Gupta and Sarah S Roley</i>	283
Phosphorus in soils—Biological interactions <i>Nataliya Bilyera and Iryna Loginova</i>	293
Biological transformations of sulfur in soil <i>Chaoqun Wang and Maxim Dorodnikov</i>	303
Plant uptake of mineral nutrients <i>Angela Hodge</i>	310
Higher plants and their interactions with soil organisms	
Soil Biology: Root form and function <i>Magda Garbowski, Grégoire Freschet, Louise Jackson, Cynthia Brown, and Louise Comas</i>	321
Root mucilage: Chemistry and functions in soil <i>Yutong Jiang, Lijun Hou, Fang Ding, and Joann K Whalen</i>	332
Root exudates and microorganisms <i>Aidan De Sena, Krisztina Mosdossy, Joann K Whalen, and Chandra A Madramootoo</i>	343
Root interactions with the microbiome from the rhizoplane to the bulk soil: An overview <i>Joann K Whalen and Shamim Gul</i>	357
Microorganisms & biological processes	
Soil respiration <i>Jeremiah D Vallotton, Sergey Blagodatsky, and Adrian Unc</i>	369
Anaerobic soil processes <i>PW Inglett, D Said-Pullicino, and K Ramesh Reddy</i>	379
Mineral-organic-microbial interactions <i>Balwant Singh, Sabina Yeasmin, and Donald L Sparks</i>	387

Microbial ecology & management

Habitat for soil organisms and their functions <i>Alexandra Kravchenko and Andrey Guber</i>	407
Plant growth promoters <i>Elisa Gamalero, Gustavo Santoyo, and Bernard R Glick</i>	418
Inoculation with microorganisms <i>Maureen O'Callaghan and David Wright</i>	428
Biocontrol of soil-borne plant diseases <i>Mehrdad Mohammadiani and Matthew Bakker</i>	439
Management of soil biology for crop protection from biotic and abiotic stresses <i>Michael J Goss, Isabel Brito, and Mário Carvalho</i>	447
Trees, hedges, agroforestry and microbial diversity <i>Josiah B Judson, Joseph Holden, Pippa Chapman, and Marcelo Valadares Galdos</i>	469
Urban soil microbiome: Activity, diversity and functioning <i>Maria Korneykova, Ekaterina Kozlova, Kristina Ivashchenko, Maria Vasilieva, and Viacheslav Vasenev</i>	480
Microbiology of extreme soil environments <i>Patrice Dion</i>	494
Microbial mobility and transport in soils <i>Mohammad Reza Mosaddeghi, Adrian Unc, Ali Akbar Mahboubi, and Azadeh Safadoust</i>	512

Biology of soil amendments & contaminant degradation

Biology of compost <i>Heribert Insam, Thomas Klammersteiner, and María Gómez-Brandón</i>	522
Landfarming and its microbiology <i>Naga Raju Maddela, Balasubramanian Ramakrishnan, Kadiyala Venkateswarlu, and Mallavarapu Megharaj</i>	533
Biodegradation of organic contaminants <i>PB Hatzinger and JW Kelsey</i>	547
Biological aspects of manure management <i>Natasha Rayne and Bee Khim Chim</i>	558
Septic systems and municipal waste disposal on land <i>Charles P Gerba, Michael J Goss, and Adrian Unc</i>	567

Climate change

Climate change impacts on soil biology <i>Alberto Canarini, Lucia Fuchslueger, François-Xavier Joly, and Andreas Richter</i>	578
---	-----

VOLUME 2: CHEMISTRY**Basic Soil Chemistry**

Acidity and acidification <i>Siobhan Staunton</i>	1
pH <i>Siobhan Staunton</i>	8
Buffering capacity <i>Siobhan Staunton and Ravi Naidu</i>	15
Soil specific surface area <i>Emmanuel Arthur, Markus Tuller, and Kurt D Pennell</i>	22

Soil Analytical Chemistry

Cation exchange capacity and reactions <i>Donald S Ross and Ulf Skyllberg</i>	32
Testing for nutrients in the soil and the environment <i>Antonio P Mallarino</i>	43
Metal extraction and analysis <i>Guo Wang and Honghong Li</i>	50
Soil organic matter: Composition <i>Markus Kleber</i>	58
Physical fractionation techniques <i>Sam J Leuthold, Michelle L Haddix, Jocelyn Lavallee, and M Francesca Cotrufo</i>	68
The use of stable isotopes in soil science: Low atomic number elements <i>Chiara Pistocchi, Isabelle Bertrand, Kristýna Kantnerová, Astrid Oberson, Federica Tamburini, and Roland A Werner</i>	81
The use of radioactive isotopes to study soil pools and processes <i>Timothy I McLaren</i>	96
The use of stable isotopes in soil science: Metals <i>Cécile Quantin and Damien Guinoiseau</i>	107

Periodic Table

Alkali metals in soils <i>Guy J Levy and Asher Bar-Tal</i>	116
Alkaline earth metals in soil <i>Justin B Richardson</i>	126
Transition metals in soil <i>Siobhan Staunton</i>	135
Nitrogen and ammonia in soils <i>Nicholas T Girkin and Hannah V Cooper</i>	142
Phosphorus in soil <i>Luisella Celi, Maria Martin, and Elisabetta Barberis</i>	152
Aluminum in soils <i>Syuntaro Hiradate</i>	167
Iron in soils <i>Trolard Fabienne and Guilhem Bourrié</i>	176
Sulfur and selenium in soil <i>Dongli Liang, Fei Zhou, and Mingxing Qi</i>	192
Arsenic in soil <i>Md Imran Ullah Sarkar, Afsana Jahan, Md Tofail Hossain, Mohammad Mahmudur Rahman, and Ravi Naidu</i>	201
Non-transition elements: Metals and metalloids in soils <i>Natasha, Sana Khalid, Muhammad Shahid, Faiz Rabbani, and Camille Dumat</i>	214
Lanthanides, actinides and radon <i>Franco Ajmone Marsan and Mery Malandrino</i>	225
Halogens in soils <i>Svensson Teresia, Thiry Yves, Bueno Maïté, and Oelmann Yvonne</i>	234
Transformation of metals and metalloids by microorganisms <i>Susan M Glasauer, Euan P Burford, and Geoffrey Michael Gadd</i>	243
Dynamics of radionuclides <i>Siobhan Staunton</i>	256
Silicon and other non-metal elements from Group 14 <i>Bruno Delvaux and Zimin Li</i>	263

Soil Physical Chemistry

Chemical equilibria	274
<i>Arthur Paul Schwab</i>	
Dissolution and precipitation processes	284
<i>Zhenqing Shi, Shiwen Hu, Rong Li, and Yang Lu</i>	
Reactions, redox potential and kinetics	297
<i>Trolard Fabienne and Guilhem Bourrié</i>	
Sorption of organic chemicals in soil	315
<i>Joseph J Pignatello</i>	
Sorption—Oxyanions	327
<i>CP Schulthess, H Wijnga, W Yang, Anahita Khosravi, and Baoshan Xing</i>	
Sorption—Metals	336
<i>Daniel G Strawn and Donald L Sparks</i>	
Sorption–desorption kinetics	343
<i>Gautier Landrot and Donald L Sparks</i>	
Sorption-desorption hysteresis	355
<i>Wenhao Wu, Hongxia Zhu, and Kun Yang</i>	
Surface charge and zero-charge points	368
<i>J Chorover</i>	
Flocculation and dispersion	376
<i>Guy J Levy and Isaac Shainberg</i>	

Soil Carbon Chemistry

Soil organic matter dynamics	387
<i>Yu Luo and Jianming Xu</i>	
Soil proteins	396
<i>Siobhan Staunton</i>	
Biochar: Properties, mechanisms, and interactions in the soil	406
<i>Balwant Singh, Shamim Mia, and Marta Camps-Arbestain</i>	
Dissolved organic carbon	421
<i>Jason James and Rob Harrison</i>	
Black carbon, pyrogenic carbon	432
<i>Heike Knicker</i>	
Fire effects on soil	448
<i>Stefan H Doerr, Cristina Santín, and Jorge Mataix-Solera</i>	
Interfacial reactions of microorganisms with minerals and organic matter	458
<i>Rota Wagai, Puu-Tai Yang, and Klaus Kaiser</i>	
Pesticide fate in soils	470
<i>Pierre Benoit, Laure Mamy, Carole Bedos, and Enrique Barriuso</i>	
Soil organic matter in molecular simulations	483
<i>Martin H Gerzabek, Daniel Tunega, Edgar Galicia-Andrés, and Chris Oostenbrink</i>	

Soil Physical Analysis

Raman spectroscopy	497
<i>Anupam Das and Huiyuan Guo</i>	
Fourier transform infrared spectroscopic methods of soil analysis	510
<i>Derek Peak</i>	
Electron paramagnetic resonance spectroscopy: Part I Historic Perspectives	519
<i>Giorgio S Senesi and Nicola Senesi</i>	

Electron paramagnetic resonance spectroscopy: Part II the view forward <i>Robert L Cook, Emmanuel Ampiah, Phillip T Sprunger, and Randall W Hall</i>	533
Fluorescence Spectroscopy: Part I Principles <i>Yong Zhang, Shuai Guo, and Chaoxian Wei</i>	544
Fluorescence spectroscopy: Part II practical applications <i>Yong Zhang, Chaoxian Wei, and Shuai Guo</i>	552
Near and mid infrared soil spectroscopy <i>Wartini Ng, Brendan Malone, Budiman Minasny, and Sangho Jeon</i>	567
Electron microscopic methods (TEM, SEM and energy dispersal spectroscopy) <i>Jian Zhao and Xia Liu</i>	575
Application of atomic force microscopy (AFM) on environmental interfaces <i>Saikat Ghosh, Zhang Di, Bo Pan, and Baoshan Xing</i>	589
Applications of nuclear magnetic resonance for the study of soils <i>Daniel H Lysak, Myrna J Simpson, and Andre J Simpson</i>	605
Neutron-gamma analysis of soil elements <i>Aleksandr Kavetskiy, Galina Yakubova, Nikolay Sargsyan, Stephen A Prior, and H Allen Torbert</i>	625
X-Ray Diffraction <i>Chao Wang and Cheng Gu</i>	642
Soil analysis using laser-induced breakdown spectroscopy <i>Débora Marcondes Bastos Pereira Milori, Ladislau Martin-Neto, Hero T Gollany, Diego Victor de Babos, Luis Carlos Leva Borduchi, and Paulino Ribeiro Villas-Boas</i>	654
Calorimetric analysis <i>Xingmin Rong and Qiaoyun Huang</i>	662
Understanding of soil processes at the microscale—Use of NanoSIMS in soil science <i>Carsten W Mueller, Carmen Hoeschen, and Ingrid Koegel-Knabner</i>	670
Nanoparticles in soil <i>Di Lang, Quan Chen, Yang Liu, and Bo Pan</i>	681
Overview	
Soil chemistry and climate change <i>Jianming Xu, Yong Li, and Zhongmin Dai</i>	692
VOLUME 3: ENVIRONMENT AND MANAGEMENT	
Soil functions, potential uses of soil and ecosystem services	
Soil functions, soil potential uses and ecosystem services <i>Karl Stahr</i>	1
Farming systems to conserve land for nature <i>Rattan Lal</i>	12
Geoarchaeology and soils <i>Carlos E Cordova</i>	20
Civilizations and the role of soil, agriculture, and climate change <i>Jeffrey A Homburg, Eric C Brevik, and Jonathan A Sandor</i>	29
Ecosystem services provided by soils in highly anthropized areas (SUITMAs) <i>Jean Louis Morel, Geoffroy Séré, Viatcheslav Vasenev, and Thomas Nehls</i>	37
Structure and function of forested soils <i>Peter Clinton and Jen Owens</i>	56
Agroforestry <i>NV Thevathasan, AM Gordon, and PKR Nair</i>	68

Rangeland management <i>Alexandra Sandhage-Hofmann</i>	88
Effects on soil of grassland management for pasture, hay and silage <i>Carmen Segura, Laura Cardenas, Richard McDowell, Sarah Morgan, and Martin SA Blackwell</i>	102
Dryland farming: Technological and management options for sustainable agriculture and food systems <i>Ch Srinivasarao, G Ranjith Kumar, Ravula Manasa, Kiran Pilli, Samaresh Sahoo, S Rakesh, Sumanta Kundu, K C Nataraj, KV Rao, JVNS Prasad, and S Malleswari</i>	113
The effect of crop rotations on soil <i>Caio Fernandes Zani, Arlete Simões Barneze, Rogério Peres Soratto, and Charles Andrew Francis</i>	125
The contribution of organic farming systems to soil fertility—A systems perspective <i>Bernhard Freyer, Pierre Ellssel, Jürgen K Friedel, and Kurt Möller</i>	135
Soil functions to support natural plant communities and crops: Soil multifunctionality <i>Leidivan Almeida Frazão, Luiz Arnaldo Fernandes, Adriana Marcela Silva Olaya, and Igor Costa de Freitas</i>	146
Methane and nitrous oxide emissions from land <i>Joanna M Cloy and Keith A Smith</i>	153
Soil and land degradation	
Overview chapter on soil degradation <i>K Müller, Margaret A Oliver, and C Siebe</i>	165
Soil monitoring <i>Ben Marchant and Nicolas Saby</i>	172
Soil quality and health key indicators <i>Edoardo AC Costantini and Simone Priori</i>	181
Water induced soil erosion <i>R Jane Rickson</i>	193
Erosion by wind <i>John Tatarko</i>	208
Human exposure to aeroparticles emitted from unpaved roads due to traffic resuspension and wind erosion <i>Arón Jazcilevich and Ann G Wellens</i>	223
Soil salination processes and management <i>Uri Nachshon and Guy J Levy</i>	236
Degradation of soil structure and soil organic matter <i>Andrew S Gregory</i>	246
Drainage, soil water storage, buffering and filtering <i>Jeffrey Strock and Norm Fausey</i>	257
Environmental effects of irrigation <i>Sevilay Topcu</i>	266
Trade-offs of wastewater irrigation <i>Manzoor Qadir and Christopher A Scott</i>	277
Pollution and risks of trace elements in the soil environment <i>Peng Wang, Fang-Jie Zhao, and Steve P McGrath</i>	288
Pollutants—Persistent organic <i>Rai S Kookana and Divina A Navarro</i>	296
Pollution by microplastic in agricultural areas <i>Nicolas Beriot and Esperanza Huerta-Lwanga</i>	309

Municipal solid waste management: Production, management, and environmental effects <i>Bryant M Serre and Lynda H McCarthy</i>	321
Remediation of metals and organic contaminants in soil <i>Peter Sanderson, Md Mezbaul Bahar, Bhabananda Biswas, and Ravi Naidu</i>	333
Soil change under different scenarios	
Soil change under different scenarios <i>Joann K Whalen</i>	344
The effect of global change on the soil body <i>Sophie Cornu, David Montagne, and Antra Boca</i>	360
The effects of adequate and excessive application of mineral fertilizers on the soil <i>Bijay-Singh and Tek B Sapkota</i>	369
Foliar Application of Nutrients <i>Thomas Eichert and Victoria Fernández</i>	382
Mulches and cover crops part I: Types <i>KK Bandyopadhyay, CL Acharya, and KM Hati</i>	392
Mulches and cover crops part II: Role in soil health and climate resilient agriculture <i>KK Bandyopadhyay, CL Acharya, and KM Hati</i>	401
The potential agronomic and environmental applications of biochar: Prospects and challenges <i>Yu-Jun Wang, Ping Wu, Nanthi S Bolan, and Hai-Long Wang</i>	414
Soil management for carbon sequestration <i>Christopher Brock, Uwe Franko, and Martin Wiesmeier</i>	424
Soil tillage <i>Mathieu Lamandé, Lars Juhl Munkholm, and Trond Børresen</i>	433
Weed and soil management: A balancing act <i>Trevor Kenneth James and Charles Norman Merfield</i>	439
Terraces as traditional agricultural landscapes. Stability amidst change <i>Gerardo Bocco</i>	450
Soil quality protection policies and plans to ensure sustainability <i>Ram Swaroop Meena and Sandeep Kumar</i>	457
Construction of Technosols for green roofs and vertical green system <i>Eleonora Flores-Ramírez, Stefan Abel, Elizabeth Chávez-García, and Geoffroy Séré</i>	473
Management of park areas, sport fields, and school yards, including golf courses and public right of ways <i>Maxine J Levin and Anna Paltseva</i>	488
Soil-society relations	
Overview chapter on soil-society relations <i>C Siebe, Margaret A Oliver, and K Müller</i>	505
Soil art: Sensory and symbolic engagement with soils <i>Alexandra R Toland and Daniel Wolter</i>	509
Soil science in the criminal justice system <i>Lorna Anne Dawson</i>	521
Soil governance at the international, regional and national level <i>Harald Ginzky</i>	532
The economics of soils' contribution to human well-being <i>Bartosz Bartkowski and Julian R Massenberger</i>	547

Soil and human health	555
<i>Margaret A Oliver and Eric C Brevik</i>	
Soil Science Education, Awareness and Outreach in the Modern World	572
<i>Laura Bertha Reyes-Sánchez and Eric C Brevik</i>	
Soils in landscape planning	580
<i>Seth Denizen</i>	
VOLUME 4: PEDOLOGY	
Pedology	
Pedology—Basic principles	1
<i>Rupert Bäumler, Uta Stockmann, and Margaret A Oliver</i>	
Factors of soil formation, pedogenic processes and mineralogy	
Overview chapter on factors of soil formation	7
<i>Rupert Bäumler, Margaret A Oliver, and Uta Stockmann</i>	
Factors of soil formation: Climate	14
<i>Stefaan Dondeyne, Stephan Mantel, and Seppe Deckers</i>	
Factors of soil formation: Soil biota and organic matter as a factor for soil formation and weathering processes	25
<i>Scott Werts</i>	
Soil forming factors: Parent material	34
<i>Stephen Nortcliff</i>	
Parent materials: Loess	42
<i>Tobias Sprafke</i>	
Time	48
<i>Susan E Melzer, Caroline M Yonker, and Eugene F Kelly</i>	
Factors of soil formation—Human impacts	54
<i>Jonathan A Sandor, C Lee Burras, Michael Thompson, and Skye A Wills</i>	
Dynamic soils	69
<i>Markus Egli and Dennis Dahms</i>	
Interactions between soils and climate change	82
<i>Cornelia Rumpel</i>	
Primary minerals	91
<i>Balwant Singh</i>	
Mineral components of the soil clay fraction	109
<i>Darrell G Schulze</i>	
Short-range ordered aluminosilicates	121
<i>Tetsuhiro Watanabe, James B Harsh, and Rota Wagai</i>	
Metal oxides	135
<i>Andreas C Scheinost and Balwant Singh</i>	
Soil morphology	149
<i>Phillip R Owens and Zamir Libohova</i>	
Soil micromorphology	171
<i>Rosa M Poch and Georges Stoops</i>	
Soil Classification and distribution	
History and principles of soil classification	185
<i>Maria Gerasimova and Dmitry Konyushkov</i>	

Soil classification systems: World reference base and soil taxonomy	197
<i>David Badía-Villas</i>	
World reference base for soil resources (WRB)	206
<i>Stephan Mantel, Stefaan Dondeyne, and Seppe Deckers</i>	
World soil map based on soil taxonomy	218
<i>Eswaran Padmanabhan and Paul F Reich</i>	
Soils as archives of city history	232
<i>Hermine Huot and Jean Louis Morel</i>	
Classification of soils by ecozones	
Ecozones and soils – An introduction	246
<i>Rupert Bäumler</i>	
Cold Region Soils: Part I—Distribution, Properties and Processes	250
<i>Nicolas A Jelinski, Chien-Lu Ping, and John CF Tedrow</i>	
Cold region soils: Part II – Genesis and classification	263
<i>Nicolas A Jelinski, Chien-Lu Ping, and John CF Tedrow</i>	
Boreal forest soils	280
<i>Gabriele Broll</i>	
Soils of humid cool temperate regions	289
<i>Nicolas A Jelinski, Justin B Richardson, and Edward A Nater</i>	
Grassland soils in the cool–arid–temperate ecozone (Steppe, grassland)	299
<i>Mengyang You, Xiaobing Liu, Rupert Bäumler, and Margaret A Oliver</i>	
Soils from a warm-temperate semi-tropical ecozone (Mediterranean) with humid winter months	319
<i>J Torrent, Marco Pfeiffer, and Juan José Ibañez</i>	
Arid and semi-arid subtropical and tropical ecozone (II)	332
<i>Bernhard Lucke</i>	
Humid subtropical ecozone with emphasis on Acrisols and Alisols	343
<i>Nilton Curi, Sérgio Henrique Godinho Silva, and Alberto Vasconcellos Inda</i>	
Soils of the humid and sub-humid tropics	353
<i>Karl Stahr and Ludger Herrmann</i>	
Classification of soils by properties and environmental settings	
Sodic soils	368
<i>Guy J Levy and Uri Nachshon</i>	
Organic soils	383
<i>Heike Schimmel and Wulf Amelung</i>	
The pedology of anaerobic soils	398
<i>John Hollis and Allan Lilly</i>	
Alluvium and alluvial soils	417
<i>Janis L Boettinger</i>	
Hydric Soils	424
<i>MJ Vepraskas and JF Berkowitz</i>	
Wetland soils	436
<i>Cornelius Wilhelm van Huyssteen</i>	
Paddy soils: Characteristics, productivity and management	447
<i>Stephan M Haefele</i>	
Volcanic soils and non-volcanic andic soils	458
<i>Rupert Bäumler and G Uehara</i>	

PEDOMETRICS**Pedometrics**

Overview of Pedometrics	471
<i>Alexandre MJ-C Wadoux, Inakwu OA Odeh, and Alex B McBratney</i>	

Statistics and soil data analysis

Statistics in soil science	486
<i>Richard Webster</i>	

Multivariate analysis for soil science	499
<i>Si Yang Han, Patrick Filippi, Mercedes Román Dobarco, Januar Harianto, Mathew S Crowther, and Thomas FA Bishop</i>	

Machine learning and processing of large data	509
<i>Madlene Nussbaum</i>	

The issue of scale and change of support in the spatial analysis of environmental data	521
<i>A Castrignanò, A Belmonte, and N Romano</i>	

Soil measurements in space and time

Sampling for soil survey	535
<i>DJ Brus</i>	

Handling of soil samples in the field, in transport to the laboratory and in the laboratory	551
<i>Claudy Jolivet, Céline Ratié, Déborah Chavrit, Nicolas Soler-Dominguez, and Antonio Bispo</i>	

Digital Soil Morphometrics	568
<i>Yakun Zhang, Alfred E Hartemink, Jingyi Huang, and Budiman Minasny</i>	

Proximal soil sensing in the field	579
<i>Mikaela Tilse, Uta Stockmann, and Patrick Filippi</i>	

Proximal sensing of soil moisture	591
<i>Markus Tuller, Ebrahim Babaeian, and Scott B Jones</i>	

Proximal sensing of land surface temperature	600
<i>Ebrahim Babaeian and Markus Tuller</i>	

Proximal sensing of evapotranspiration	610
<i>Ebrahim Babaeian and Markus Tuller</i>	

Remote sensing of soil moisture	618
<i>Alexander Gruber and Jian Peng</i>	

Remote Sensing of Soil Organic Carbon	631
<i>Bas van Wesemael and Sabine Chabrilat</i>	

Pedotransfer functions and their application to soil water dynamics	642
<i>Marcel G Schaap, Yonggen Zhang, and Attila Nemes</i>	

Spatial and temporal soil variation

Geostatistics: Modelling spatial variation	655
<i>Richard Webster</i>	

Uncertainty assessment of spatial soil information	671
<i>Gerard BM Heuvelink and Richard Webster</i>	

Digital soil mapping: Evolution, current state and future directions of the science	684
<i>Brendan Malone, Dominique Arrouays, Laura Poggio, Budiman Minasny, and Alex B McBratney</i>	

Geographical information systems (GIS) and soils	696
<i>S Grunwald and J Böhner</i>	

Precision agriculture	710
<i>James A Taylor</i>	

Foundations, measurements and trends in pedodiversity <i>Juan José Ibañez and Marco Pfeiffer</i>	726
Soil landscape modelling	
Modeling soil development in a landscape context <i>Peter A Finke, Budiman Minasny, and Arnaud JAM Temme</i>	739
Climate and land surface models: Role of soil <i>Toby Richard Marthews, Holger Lange, Alberto Martínez-de la Torre, Richard J Ellis, Sarah E Chadburn, and Martin G De Kauwe</i>	751
VOLUME 5: SOIL PHYSICS	
Soil structure & mechanics	
Soil structure <i>Steffen Schlüter and John Koestel</i>	1
Aggregation <i>Teamrat A Ghezzehei</i>	8
Porosity and pore-size distribution <i>John R Nimmo, Sheela Katuwal, and Maik Lucas</i>	16
Pore-scale modelling of flow and transport phenomena in soils <i>Kirill M Gerke and Marina V Karsanina</i>	25
Soil texture: Its relevance, measurement and representation internationally <i>Attila Nemes</i>	35
Soil tilth <i>Thomas Keller, Douglas L Karlen, and Paul D Hallett</i>	48
Crusts and seals: Structural <i>Shmuel Assouline and Sally Thompson</i>	57
Soil mechanics <i>Markus Berli and Paul D Hallett</i>	69
Compaction <i>JJH van den Akker, L ten Damme, M Lamandé, and Thomas Keller</i>	85
Bioturbation—Physical processes <i>Siul Ruiz, Paul D Hallett, and Dani Or</i>	100
The reinforcement of soil by plant roots <i>Alexia Stokes and Anil Yildiz</i>	115
Physical effects of soil salinity <i>David Russo</i>	123
Shrink-Swell processes in soils <i>Pascal Boivin</i>	139
Water retention	
Water properties <i>N Lu, C Zhang, and D Hillel</i>	146
Water potential <i>Dani Or, Markus Tuller, and Jon M Wraith</i>	159
Water content and water potential measurement <i>Wolfgang Durner</i>	168
Soil water retention and characteristic curve <i>Markus Tuller and Dani Or</i>	187
Soil water repellency <i>Jörg Bachmann and Marc-Oliver Goebel</i>	203

Flow & transport processes

Aeration	216
<i>Rodney T Venterea and Dennis E Rolston</i>	
Capillarity	225
<i>Dani Or and Markus Tuller</i>	
Darcy's law	240
<i>Conrad Jackisch and Eva Kröner</i>	
Diffusion	247
<i>TKK Chamindu Deepagoda and Shoichiro Hamamoto</i>	
Hydrodynamics in soil	256
<i>Ty Ferre and Art Warrick</i>	
Thermodynamics of soil water	262
<i>Pieter H Groenevelt</i>	
Immiscible Fluids	271
<i>Yukie Tanino and Dubravka Pokrajac</i>	
Overland flow	287
<i>Tammo S Steenhuis, Rajith Mukundan, and Seifu A Tilahun</i>	
Direction-dependency of soil hydraulic and mechanical properties	294
<i>Steffen Beck-Broichsitter, Heiner Fleige, Jose Dörner, and Rainer Horn</i>	
Macropores and macropore flow	306
<i>Nicholas Jarvis and Mats Larsbo</i>	
Unstable preferential flow in porous media	314
<i>Naaran Brindt, J-Yves Parlange, and Tammo S Steenhuis</i>	
Leaching processes	321
<i>Brent Clothier and Steve Green</i>	
Colloid transport	336
<i>Yan Jin, Chongyang Shen, and Volha Lazouskaya</i>	
Scaling of transport properties	353
<i>Robert Ewing and Behzad Ghanbarian</i>	
Soil influence on groundwater	364
<i>David Rudolph and Edwin Cey</i>	

Thermal properties and behaviour

Temperature effects of soil hydraulic properties	379
<i>Jan W Hopmans and Robert Horton</i>	
Heat and vapor transfer	385
<i>Sunny Eng Giap Goh, Joshua L Heitman, and Robert Horton</i>	
Radiation Balance	393
<i>Jerry L Hatfield, Kaitlin Togliatti, Thomas J Sauer, and John H Prueger</i>	
Freezing and thawing cycles	400
<i>Alan W Rempel</i>	
Arctic permafrost	410
<i>Julia Boike, Andrey Abramov, Katrina E Bennett, and Lars Kutzbach</i>	
Permafrost and cold region hydrology	419
<i>Masaki Hayashi and Ylva Sjöberg</i>	

Sensing Techniques

Soil water sensing by neutron scattering	427
<i>Steven R Evett</i>	

Time-domain reflectometry <i>GC Topp and Ty Ferre</i>	436
Geophysical methods for soil applications <i>Sarah Garré, Guillaume Blanchy, David Caterina, Philippe De Smedt, Alejandro Romero-Ruiz, and Nataline Simon</i>	444
Non-invasive imaging of soil processes <i>Craig J Sturrock</i>	459
Hydrology & water balance	
Soil processes in the hydrologic cycle <i>Cosimo Brogi, Harry Vereecken, Heye Reemt Bogena, and Luca Brocca</i>	469
Global water cycle from a soil perspective <i>Michael H Young and Dani Or</i>	482
Water evaporation from bare soil <i>Peter Lehmann</i>	490
Evapotranspiration <i>Yuting Yang</i>	501
Plant available water <i>Quirijn de Jong van Lier, Sally D Logsdon, Everton Alves Rodrigues Pinheiro, and Paulo Ivonir Gubiani</i>	509
Plant-water relations <i>Ian C Dodd, Andrew D Hiron, and Jaime Puértolas</i>	516
Water use efficiency: A review of spatial and temporal variability <i>Simone Fatichi, Athanasios Paschalis, Sara Bonetti, Gabriele Manoli, and Christoforos Pappas</i>	527
Root water uptake and myco-rhizosphere hydraulic properties <i>Andrea Carminati, Maria Marin, Valentin Couvreur, Mohsen Zarebanadkouki, Stéphane Declerck, and Mathieu Javaux</i>	543
Plant-soil-water relations <i>Jan Vanderborght, Quirijn de Jong van Lier, Andrea Schnepf, and Mathieu Javaux</i>	552
Crop and landscape water requirements <i>Luis S Pereira, Isabel Alves, and Paula Paredes</i>	571
Water harvesting <i>Wolde Mekuria and Desalegn Tegegne</i>	593
Irrigation methods <i>Shmulik P Friedman</i>	608
Soil–plant–atmosphere continuum <i>John M Norman, Martha C Anderson, and William P Kustas</i>	624
The Penman-Monteith equation <i>Richard G Allen, Ayse Kilic, Guido D'Urso, and Simona Consoli</i>	634
Infiltration <i>Ty Ferre and Art Warrick</i>	646
Solute transport <i>MC Sukop and M Dentz</i>	652
Lysimetry <i>Thomas Pütz and Jannis Groh</i>	667
Recognition of notable past soil scientists <i>Michael J Goss, Evgenia Blagodatskaya, Ariena HC van Bruggen, Monika Joschko, Stefan Schrader, and Christoph Tebbe</i>	680
<i>Index</i>	691

Acidity and acidification

Siobhan Staunton, UMR Eco&Sols, INRAE, INRAE-IRD-Cirad-Institut Agro-Université de Montpellier, Montpellier, France

© 2023 Elsevier Ltd. All rights reserved.

This text is largely based on the text of the Chapter entitled *Acidity* in the First edition by N.S. Bolan, Massey University, Palmerston North, New Zealand, D. Curtin, New Zealand Institute for Crop and Food Research, Christchurch, New Zealand and D.C. Adriano, University of Georgia, Aiken, SC, USA.

This is an update of NS Bolan, D Curtin, DC Adriano, *Acidity*, Editor(s): DH, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 11–17, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00173-9>.

Introduction	1
Processes generating acidity in soils	2
Acid drainage	2
Acid precipitation	3
Elemental cycling	3
Fertilizer reactions	4
Measurement and effects of soil acidity	4
Amelioration of soil acidity	6
References	6

Abstract

Soil pH is an important property and may determine land-use and quality. Few crops are well adapted to acid soils and so amendments and management are sought to alkalize or limit further acidification. This chapter reviews the various processes that influence soil acidity. The effects of industrial activity and farming are considered. Two important industrial processes are acid drainage following pyrite oxidation and acid precipitation, best known as 'acid rain'. The cycling of C (carbon), N (nitrogen) and S (sulfur) induce changes in acidity and each of these is examined. The measurement of soil pH is briefly discussed. The adverse effects of soil acidity and countermeasures are also considered.

Key points

- Soils may be acidified by industrial and agricultural activity
- Liming and appropriate land use may mitigate acidification

Introduction

A large number of soils worldwide, particularly highly weathered tropical soils, are moderately to extremely acid (Ulrich and Sumner, 1991; Rengel, 2003; Kipling, 2022). Acid soils often have marked vertical pH gradients, depending on the origin of acidity; acid topsoils under forests, but acid subsoils in many Australian soils (Condon et al., 2021). A restricted number of crops are well adapted to very acid soils, but in general, these soils are considered to be of poor quality for agriculture (Hartemink and Barrow, 2023). At acid pH, cations such as calcium (Ca^{2+}) and magnesium (Mg^{2+}) on the exchange complex are replaced by protons and aluminum (Al), leading to aluminum toxicity. Soil acidification is a natural process that can either be accelerated by certain plants and human activities or slowed down by careful management practices (Fageria and Nascente, 2014). Industrial and mining activities lead to soil acidification due to acid produced from pyrite oxidation and from acid precipitation caused by the emission of sulfur (S) and nitrogen (N) gases. Acid precipitation has been significantly reduced in recent decades (Gilliam et al., 2019). In managed ecosystems, soil acidification is mainly caused by the release of protons (H^+) during the transformation and cycling of carbon (C), N, and S, and fertilizer reactions. Soil acidification caused by these processes can have adverse effects where soils are unable to buffer against further decrease in pH. For example, in parts of North America and Europe, soil acidification caused by acid precipitation has resulted in forest decline, and in some parts of Australia, continuous legume cultivation and inappropriate use

of N fertilizer have generated sufficient soil acidity that cereal crop cultivation has had to be abandoned because of aluminum (Al) and manganese (Mn) toxicity.

Historically, liming is the most common management practice used to neutralize soil acidity. Most plants grow well in the pH range 5.5–6.5, and the usual objective of liming programs is to maintain pH in this range. Liming enhances the physical, chemical, and biological properties of soil through its direct effect in ameliorating soil acidity and through its indirect effects in mobilizing plant nutrients, immobilizing toxic 'heavy metals', and improving soil physical conditions. In variable-charge soils, liming can be used as a management tool to manipulate the surface charge, thereby controlling the reactions of nutrient ions and metals. Liming provides optimum conditions for several biological processes, including N₂ fixation and mineralization of N, phosphorus (P), and S. Biochar is now often used to alkalize soils.

Processes generating acidity in soils

Processes generating acidity in soils can be broadly grouped into two categories: (1) those occurring in natural ecosystems through industrial activities, and (2) those occurring in managed ecosystems because of farming activities. The various reactions involved in these processes are given in Table 1.

The two important acid-generating processes resulting from industrial activities in natural ecosystems are acid drainage following pyrite oxidation and deposition of acidity in precipitation. While the first process occurs at a local level, the second process can lead to acidification far from the source of acid production and hence an international effort is needed to combat it.

Acid drainage

Acid drainage has both anthropogenic and natural origins, including mining of coal and sulfide-containing metal ores, land disturbances (e.g., rice (*Oryza sativa* L.) cultivation), and industrial activities (mineral processing, manufacturing or recycling of batteries, electronic equipment, wood pulp, tanneries and textile manufacturing, and food processing).

Pyrite is commonly associated with coal and metal ores, as well as mine deltas, wetlands, and rice fields. Exposure of pyrite to the atmosphere leads to its oxidation and the production of extremely acidic drainage water. Pyrite oxidation includes biological and electrochemical (abiotic) reactions (Table 1). Abiotic oxidation of pyrite is pH-sensitive and it is extremely slow in very acidic conditions. The bacteria *Thiobacillus ferrooxidans* and *Thiobacillus thiooxidans* are mainly responsible for the oxidation of pyrite, especially in acid soils.

Table 1 Proton generation and consumption processes in the soil-plant system, including those associated with acid precipitation, pyrite oxidation, and the C, N, and S biogeochemical cycles.

Process Reaction	Equations	H ⁺ (mol _e mol ⁻¹)	Equation number
<i>Acid precipitation</i>			
Oxidation of sulfur dioxide	2SO ₂ + O ₂ → 2SO ₃	0	1.1
Hydrolysis of sulfur trioxide	SO ₃ + H ₂ O → H ₂ SO ₄ → SO ₄ ²⁻ + 2H ⁺	+2	1.2
Photochemical oxidation of nitric oxide	O ₃ + NO → N ₂ O + O ₂	0	1.3
Hydrolysis of nitrogen dioxide	NO ₂ + H ₂ O → HNO ₃ + HNO ₂ → NO ₃ + H ⁺	+1	1.4
<i>Pyrite oxidation</i>			
Pyrite oxidation by oxygen	2FeS ₂ + O ₂ + H ₂ O → 2Fe ²⁺ + 4SO ₄ ²⁻ + 4H ⁺	+2	2.1
Ferrous iron oxidation	4Fe ²⁺ + O ₂ + 4H ⁺ → 4Fe ³⁺ + 2H ₂ O	-1	2.2
Ferric iron precipitation	Fe ³⁺ + 3H ₂ O → Fe(OH) ₃ + 3H ⁺	+3	2.3
Pyrite oxidation by ferric iron	FeS ₂ + 14Fe ³⁺ + H ₂ O → 15Fe ²⁺ + 2SO ₄ ²⁻ + 16H ⁺		2.4
<i>Carbon cycle</i>			
Dissolution of carbon dioxide	CO ₂ + H ₂ O → H ₂ CO ₃ → H ⁺ + HCO ₃ ⁻	+1	3.1
Synthesis of organic acid	Organic C → RCOOH → RCOO ⁻ + H ⁺	+1	3.2
<i>Nitrogen cycle</i>			
N fixation	2N ₂ + 2H ₂ O + 4R'OH → 4R-NH ₂ + 3O ₂	0	3.3
Mineralization of organic N	R-NH ₂ + H ⁺ + H ₂ O → R-OH + NH ₄ ⁺	-1	3.4
Urea hydrolysis	(NH ₂) ₂ CO + 3H ₂ O → 2NH ₄ ⁺ + 2OH ⁻ + CO ₂	-1	3.5
Ammonium assimilation	NH ₄ ⁺ + R'OH → R-NH ₂ + H ₂ O + H ⁺	+1	3.6
Ammonia volatilization	NH ₄ ⁺ + OH ⁻ → NH ₃ + H ₂ O	+1	3.7
Nitrification	NH ₄ ⁺ + 2O ₂ → NO ₃ ⁻ + H ₂ O + 2H ⁺	+2	3.8
Nitrate assimilation	NO ₃ ⁻ + 8H ⁺ + 8e ⁻ → NH ₃ + 2H ₂ O + OH ⁻	-1	3.9
Denitrification	4NO ₃ ⁻ + 4H ⁺ → N ₂ + 5O ₂ + 2H ₂ O	-1	3.10
<i>Sulfur cycle</i>			
Mineralization of organic S	2Organic S + 3O ₂ + 2H ₂ O → 2SO ₄ ²⁻ + 4H ⁺	+2	3.11
Assimilation of sulfate	2SO ₄ ²⁻ + 8H ⁺ + 8e ⁻ → SH ₂ + 2H ₂ O + 2OH ⁻	-2	3.12
Oxidation of S ⁰	2S ⁰ + 2H ₂ O + 3O ₂ → 2SO ₄ ²⁻ + 4H ⁺	+2	3.13

Acid precipitation

Carbon dioxide (CO_2) combines with water to form a dilute solution of carbonic acid (H_2CO_3) with an equilibrium pH of approximately 5.6. For this reason, acid precipitation is arbitrarily defined as precipitation with a pH value of less than 5.6. Natural sources of acid precipitation include geologic weathering, volcanic eruptions, anaerobic decomposition of organic matter, air-borne sea-salt sprays, and production of N oxides (NO_x) during lightning storms. The increased acid precipitation burden in recent decades has been attributed to anthropogenic sources that include combustion of fossil fuels (especially coal and oil), certain industrial processes (especially smelting of ores), exhausts from internal combustion engines, and N fertilizer application to agricultural and forest lands.

Widespread occurrence of acid precipitation (i.e., both wet and dry deposition) results in large part from industrial emissions of the oxides of S (SO_x) and N (NO_x). These compounds are transformed in the atmosphere to sulfuric and nitric acids (Table 1), which can be transported over great distances before deposition on vegetation, soils, surface waters, and building structures. The average annual ratio of sulfuric to nitric acids in North America is approximately 2:1, but nitric acid is becoming progressively important relative to sulfuric acid because of the installation of flue gas desulfurization (FGD) systems in coal-fired power stations. The major impact of acid precipitation has been on freshwater ecosystems, while its impact on terrestrial ecosystems is more controversial. Forest health may be one casualty of acid precipitation, but increasing tropospheric ozone may also be a factor in forest decline.

Elemental cycling

The most significant proton (H^+)-generating processes in soil are associated with the cycling of C, N, and S (Table 1) and these processes can be grouped into two main categories: plant-induced (i.e., uptake and assimilation) and soil-induced (i.e., transformation) processes.

In the case of the C cycle, the metabolism of photosynthates results in the synthesis of organic acids and, at the cytoplasmic pH of the plants (pH 7.2–7.4), some of the carboxyl groups of these acids dissociate to produce H^+ ions. Cytoplasmic pH regulation is generally achieved by transport of excess H^+ ions out of the cytoplasm. Some terrestrial plant species counteract the change in cytoplasm pH by excreting H^+ ions into the soil solution and, at the same time, taking in a nutrient basic cation to balance the charge.

Plants take up N in three main forms—as an anion (nitrate, NO_3^-), as a cation (ammonium, NH_4^+), or as a neutral N_2 molecule (N_2 fixation). Depending upon the form of N taken up and the mechanism of assimilation in the plant, excessive cation or anion uptake may occur. To maintain charge balance during the uptake process, H^+ , OH^- , or bicarbonate (HCO_3^-) ions must pass out of the root into the surrounding soil. While the uptake of NH_4^+ and N_2 fixation result in a net release of H^+ ions, uptake of NO_3^- can result in a net release of hydroxyl (OH^-) ions. In the case of N_2 fixation, acidity is generated even when no ionic species of N are taken up by the plant. This is because basic cations are imported into the legume in exchange for H^+ ions generated during C assimilation into carboxylic acids. The amount of H^+ ion released during N_2 fixation is a function of C assimilation, and therefore it depends mainly on the nature and amounts of amino and organic acids synthesized by the plant. When ammonium assimilation occurs in the root, deprotonation of NH_4^+ to R-NH_2 releases one mole of H^+ per mole of NH_4^+ (Eq. (3.6) in Table 1). When plants take up NO_3^- , the NO_3^- ion is first reduced to NH_4^+ , which is subsequently assimilated into amino acids. This NO_3^- reduction produces one mole of OH^- ion for every mole of NO_3^- reduced to NH_4^+ (Eq. (3.9) in Table 1).

Sulfate (SO_4^{2-}) is assimilated into sulfur-containing amino acids (cysteine, cystine, and methionine) in the form of sulfhydryl ($-\text{SH}$) groups. This reduction process produces two net moles of OH^- for each mole of S assimilated (Eq. (3.12) in Table 1). On decomposition of sulfhydryl-containing amino acids, two moles of H^+ ions are generated for each mole of $-\text{SH}$ oxidized to sulfate. Since plants require roughly 10 times less S than N, assimilation of SO_4^{2-} will have only a minor effect on the H^+ balance in plants.

As microorganisms decompose soil organic matter, they respire CO_2 , which upon hydrolysis forms H_2CO_3 (Eq. (3.1) in Table 1), eventually dissociating to H^+ and HCO_3^- ions (Eq. (3.2) in Table 1). The continuous production of CO_2 through soil and root respiration increases the concentration of CO_2 in the soil air spaces. Soil microorganisms also produce organic acids when they decompose plant litter that is rich in organic compounds but low in charge-balancing basic cations (Eq. (3.2) in Table 1). In general, conifer litter tends to produce more organic acids than does leaf fall from deciduous woodlands.

Nitrification and ammonia (NH_3) volatilization result in the release of H^+ ions. While the ammonification process (conversion of organic forms of N to $\text{NH}_4^+ - \text{N}$) results in the release of OH^- ions, the subsequent oxidation of NH_4^+ to NO_3^- (nitrification) results in the release of H^+ ions. Combined ammonification (Eq. (3.4) in Table 1) and nitrification (Eq. (3.8) in Table 1) of organic N compounds, including urea, in theory generates one net mole of H^+ for every mole of N transformed. In alkaline media NH_4^+ ions dissociate into gaseous NH_3 , which is subject to volatilization (Eq. (3.7) in Table 1), resulting in a net decrease in pH with the consumption of OH^- ions (or release of H^+ ions) as NH_4^+ is converted to NH_3 .

In aerobic soils, H^+ ions are produced during the mineralization and subsequent oxidation of S in soil organic matter (Eq. (3.11) in Table 1). As soil bacteria and fungi decompose plant litter and soil organic matter rich in C but low in S, soil solution SO_4^{2-} may be immobilized. In this case (Eq. (3.11) in Table 1) is reversed and it becomes an H^+ -consuming reaction, because SO_4^{2-} is assimilated into microbial biomass. Under anaerobic conditions, some bacteria have the capacity to use SO_4 as a terminal electron acceptor, resulting in H^+ consumption as SO_4^{2-} is reduced along a chain of intermediate compounds to hydrogen sulfide (H_2S). The

H₂S reacts with metal ions to precipitate as metal sulfides, which is an H⁺-consuming process. However, when these metal sulfides are reoxidized they generate H⁺, acidifying the soil (acid drainage).

Fertilizer reactions

Fertilizer application in ecosystems managed for agricultural production is a major contributor to soil acidification. The acidifying effects of fertilizers commonly used in agricultural production are presented in Table 2.

Application of N fertilizers such as urea and ammonium sulfate to soils produces H⁺ by two processes: nitrification (Eq. (3.8) in Table 1) and NO₃⁻ leaching. Part of the H⁺ produced is neutralized by OH⁻ released by plants during the subsequent uptake of the NO₃⁻ ions (Eq. (3.9) in Table 1). The depletion of basic cations (Ca, K (potassium), Mg, and Na) during the leaching of NO₃⁻ ions (i.e., as ion pair) and product removal (i.e., silage) accelerates the acidification process. In the case of ammonium sulfate, the concomitant leaching of SO₄²⁻ and NO₃⁻ ions causes greater depletion of basic cations. With urea, the initial conversion of amide N (R-NH₂) to NH₄⁺ (ammonification) releases OH⁻ (Eq. (3.5) in Table 1), which neutralizes part of the H⁺ produced during the subsequent oxidation of NH₄⁺ to NO₃⁻. This explains why urea-based N fertilizers are less acidifying than the NH₄⁺-based fertilizers (Table 2).

Superphosphates, the most common phosphate fertilizers, contain monocalcium phosphate (MCP) as the principal P component. Dissolution of MCP in soils results in the formation of dicalcium phosphate with the release of phosphoric acid, which subsequently dissociates into phosphate anions (H₂PO₄⁻) and H⁺. Part of H⁺ is subsequently neutralized by the OH⁻ released during the adsorption of H₂PO₄⁻ by soil particles. Compared with weakly adsorbed SO₄²⁻ and NO₃⁻ ions, H₂PO₄⁻-induced leaching of basic cations is unlikely to occur, because the H₂PO₄⁻ ions are strongly adsorbed by most soils.

In legume-based pasture and crop-production systems, P fertilizers are added to promote N₂ fixation by the legumes. Regardless of P fertilizer source, application of P to legume-based systems promotes N₂ fixation, thereby indirectly causing soil acidification. The amount of acidity produced indirectly by N₂ fixation depends mainly on the extent of NO₃⁻ leaching and it can be greater than that produced directly by the dissolution of MCP in superphosphate fertilizer granules (Table 2).

Elemental sulfur (S⁰) is used as an acidifying agent, as a slow-release S fertilizer or, in a finely divided form, as a fungicide. When S⁰ is added to soils, it is oxidized to sulfuric acid, which dissociates into SO₄²⁻ and H⁺ ions (Eq. (3.13) in Table 1).

Measurement and effects of soil acidity

The acidity of a soil is assessed from the activity of H⁺ ions in the soil solution. There is a huge range of [H⁺] in soil solution, therefore, a logarithmic scale (known as pH scale; pH = -log [H⁺]) is used to quantify acidity. In soils, acidification is indicated by a decrease in either pH or anion neutralizing capacity. For a given input of acidity, the extent to which pH decreases depends mainly on the pH buffering capacity of the soil. Various soil constituents, such as organic matter, Fe (iron) and Al oxides, and CaCO₃ (in calcareous soils) contribute to the pH buffering of soils at different pH values. Depending on organic matter content, texture, and

Table 2 Acidity production by various fertilizers and its effect on pH of two soils with contrasting buffer capacities.

Fertilizer	Acidity equivalence ^a	Acidity produced (kmol H ⁺ ha ⁻¹) ^b	Number of years required to reduce soil pH by 1 unit ^c	
			Tokomari silt loam	Egmont clay loam
Ammonium sulfate	110	2.60	8	26
Ammonium chloride	93			
Ammonium nitrate	60			
Diammonium phosphate	74	2.06	10	33
Monoammonium phosphate	55			
Urea	79	0.86	25	78
Potassium nitrate	-23			
Calcium nitrate	-50			
Sodium nitrate	-29			
Nitrogen fixation	70–250	–		
Single superphosphate	8	0.48	45	140
Triple superphosphate	15	0.50	43	135
North Carolina phosphate rock	-50			
Calcium sulfate	-57			
Potassium sulfate	-64			
Elemental sulfur (S ⁰)	310	1.55	14	43

^aAcidity equivalence is the number of parts by weight of pure lime (calcium carbonate) required to neutralize the acidity caused by 100 parts of the fertilizer. Negative values indicate the liming value (kilograms CaCO₃ 100 kg⁻¹) of the fertilizer.

^bAmmonium sulfate, diammonium phosphate, and urea added at the rate of 25 kg N ha⁻¹ year⁻¹; single and triple superphosphate at the rate of 30 kg P ha⁻¹ year⁻¹; and S⁰ at the rate of 30 kg S ha⁻¹ year⁻¹.

^cpH-buffering capacity (kilomoles H⁺ per hectare) of 21.7 and 67.5 for the Tokomaru and the Egmont soil, respectively.

the nature of the clay mineralogy, the amount of acidity needed to reduce the pH of topsoil (noncalcareous) by 1 unit is normally in the 1- to 8-cmol (H^+) kg^{-1} range. These values translate to buffer capacities of approximately 10–70 kmol (H^+) ha^{-1} for the top 7.5 cm soil layer. The amounts of acidity produced by selected fertilizers and the number of years required to reduce soil pH by 1 unit are shown in Table 2 for two soils with contrasting pH buffer capacities. It is clear that annual application of ammonium sulfate or ammonium phosphates could acidify soil within a short time (10 years), particularly when soil pH-buffering capacity is low.

Acidification affects the transformation and bio-geochemical cycling of both nutrients and metals through its effect on the physical, chemical, and biological characteristics of soils. Soil pH can be viewed as the master variable of all the controlling factors, because it can affect the surface charge and subsequent adsorption of solutes by variable-charge soil components. In addition to its effect on the sorption of metal cations and anions in soils, it also influences metal speciation, complexation of metals with organic matter, precipitation and dissolution reactions, redox reactions, mobility and leaching, dispersion of colloids and, ultimately, the bioavailability of trace metals.

Soil pH affects the surface charge through the supply of H^+ for adsorption on to the metal oxides and dissociation of the functional groups in the soil organic matter. A decrease in pH decreases the net negative charge (often referred to as cation exchange capacity or CEC) and increases the net positive charge (often referred to as anion exchange capacity or AEC).

Because acidity governs the type, number, and activity of microorganisms in soil, it influences the rate of organic matter mineralization and availability of elements such as N, S, and P. Because of suboptimal microbial activity in highly acid conditions, organic matter accumulates, giving rise to a storehouse of nutrients that can be exploited by liming. Likewise, nitrification is markedly reduced below pH 6, whereas the ammonification reaction is relatively insensitive to acidity, resulting in the accumulation of NH_4^+ in acidic soil where nitrification is inhibited. Thus, for certain crops unable to use NH_4^+ , acidification can result in restricted uptake of N and it may promote NH_4^+ toxicity.

Acidity has a deleterious effect on the symbiotic relationship between *Rhizobia* and legumes, and nodulation and N_2 fixation are generally poor below pH 6. Physiological reasons for this include:

- 1) inhibition of rhizobial infection of legume roots, decreasing nodule formation;
- 2) inhibition of nitrogenase enzyme activity in the nodule due to modification of the nitrogenase iron protein;
- 3) decrease in bacterial membrane potential and the inhibition of the leghemoglobin;
- 4) decrease in the supply of photosynthate to the rhizobium due to the poor supply of major nutrients, such as P; and
- 5) poor supply of Mo (molybdenum) and Ca, which are essential for N_2 fixation.

A decrease in soil pH will increase the concentrations of Fe and Al in soil solution and this may lead to adsorption or precipitation of phosphate. In variable-charge soils, a decrease in pH increases the AEC, which increases the retention of phosphate. However, at very low pH, solubilization of P compounds may result in an increase in the concentration of P in the soil solution. Under acid conditions, weathering liberates K from micaceous and feldspar minerals, enabling it to enter the soluble and exchangeable pools; but with? in variable-charge soils, increasing acidity decreases CEC, reducing the ability of the soil to retain K, and this may cause K to be more prone to leaching.

In many soils, organic matter is the main source of S and, since mineralization of organic matter may be slowed by acidity, the release of S for plant uptake may decline when soil is acidified. Furthermore, in highly weathered acid soils, SO_4^{2-} (sulfate) may be adsorbed by sesquioxide surfaces, precipitated as Al—OH—SO_4^- type minerals, such as alunite and basulminite, and or held as an exchangeable anion on positively charged sites on sesquioxides.

One of the major consequences of acidification is a decline in basic cations such as Ca^{2+} and Mg^{2+} , cations leading to potential deficiency of these cations for plant growth. Furthermore, at low pH, the bioavailability of Ca may be restricted due to antagonistic effects of soluble Al. With increasing soil acidification, smaller amounts of Mg^{2+} remain in exchangeable form because of a reduction in negative charge. Since Mg^{2+} is a poor competitor with Al^{3+} and Ca^{2+} for the exchange sites, it tends to accumulate in the solution phase and is therefore more prone to leaching.

Acidification affects the transformation of metal ions through: (1) changes in surface charge in variable-charge soils; (2) changes in metal speciation; and (3) shifts in the oxidation-reduction reactions of metals. In general, the solubility and phytoavailability of metals such as Cu (copper) and Zn (zinc) are inversely related to soil pH. Since Cu is largely complexed with humic substances, the greater solubility of humic compounds at high pH may result in more soluble Cu when pH is raised, although this organically complexed Cu might not have poor bioavailability. One of the major consequences of soil acidification is the increase in concentrations of soluble Al and Mn, both of which are highly toxic to plant growth. A primary aim of liming soils for agricultural production is to decrease the concentration of these elements. While Mn toxicity is directly related to the metabolic requirements of plants, the effect of Al toxicity appears to be largely manifested as malformation and malfunction of the root system, a syndrome which is exacerbated by small amounts of Ca in solution in acid soils.

Under acid conditions, boron (B) occurs in soil solution as the uncharged H_3BO_3 (boric acid) molecule and thus it is readily available in acid soil. Unlike other anions, Mo is highly insoluble in low pH conditions and it may become limiting when soil is acidified.

Cadmium (Cd) has been identified as a potential human health hazard, which enters the food chain through plant uptake. Adsorption of Cd^{2+} by soil decreases with decreasing pH for the following reasons: (1) in variable-charge soils, a decrease in pH results in a decrease in cation adsorption and (2) a decrease in soil pH is likely to result in an increase in free Cd^{2+} at the expense of the more strongly adsorbed hydroxy-cadmium species. In general, Cd uptake by plants increases with decreasing pH and consequently it is recommended that soil pH be maintained at pH 6.5 or greater in land receiving Cd-rich biosolids.

The effect of soil acidity on the adsorption of metalloids such as arsenic (As) and selenium (Se) is controlled by two interacting factors—pH-related changes in negative potential in the plane of adsorption and pH-induced changes in the amount of negatively charged ionic species present in soil solution. While the first factor results in a decrease in As(V) (arsenate) adsorption when pH is raised, the latter factor is likely to cause the opposite effect. Thus, the effect of pH on As(V) adsorption depends on the nature of the mineral surface. In soils with low oxide content, increasing the pH will have little effect on As(V) adsorption, while in highly oxidized soils, adsorption will decrease with increasing pH.

Amelioration of soil acidity

Three approaches can be taken to minimize the rate and impacts of soil acidification: (1) reduce the amount of H^+ ions generated, (2) minimize the extent to which the H^+ and OH^- ions generating processes are uncoupled, and (3) neutralize the acidity produced. In managed ecosystems, the rate of acid generation can be altered by applying nutrients in forms that produce less acidity, selecting plant species that accumulate less cation, and reduce the loss of C, N, and S from soil.

Traditionally, liming has been the most common practice used to alleviate soil acidification (Holland et al., 2018; Chen et al., 2018; Nguyen, 2023; Fageria and Baligar, 2008). A range of liming materials is available, including calcite ($CaCO_3$), burnt lime (CaO), slaked lime ($Ca(OH)_2$), dolomite ($CaMg(CO_3)_2$), and slag ($CaSiO_3$). Recently, the liming potential of other Ca-containing compounds has been evaluated. Some of these materials include phosphate rocks (PRs), FGD gypsum, fluidized bed boiler ash, fly ash, and lime-stabilized organic composts. The amount of liming material required to rectify soil acidity depends on the neutralizing value of the liming material and pH-buffering capacity of the soil.

Unlike soluble P fertilizers, PRs can have a liming value, because they contain some free $CaCO_3$, which itself can act as a liming agent and, secondly, the dissolution process of the P mineral component (i.e., apatite) consumes H^+ , thereby reducing the soil acidity. While the free $CaCO_3$ in PRs dissolves reasonably fast, providing a small, immediate liming effect, the apatite generally dissolves at a slower (but variable) rate and has liming value over a longer period of time.

Alkaline-stabilized biosolids are increasingly being used as agricultural lime substitutes and soil amendments. Alkaline stabilization of biosolids utilizes a combination of high pH, heat, and composting to kill pathogens and stabilize organic matter. A range of alkaline materials is used for this purpose, including cement-kiln dust, lime-kiln dust, lime, limestone, alkaline coal fly ash, FGD, other coal-burning ashes, and wood ash. To minimize metal mobility and bioavailability in biosolid-amended soils, the US Environmental Protection Agency (USEPA) recommends the application of alkaline-stabilized biosolids and other liming agents to increase the soil pH to 6.5 or more.

The primary purpose of liming arable soils is to overcome the chemical problems associated with soil acidity that include large concentrations of acid ions (H^+ and Al^{3+}) and toxic elements (Mn^{2+}), and small concentrations of basic cations (Ca and Mg), and other nutrient ions such as Mo and P. The hydrolysis of the basic cations in lime produces OH^- ions, which neutralize the H^+ ions, thereby decreasing the activity and bioavailability of Al and Mn. Liming also increases the solubility of Mo and P, thereby increasing their availability. Lime provides the basic nutrient cations (Ca^{2+} and Mg^{2+}) and also reduces the solubility of 'heavy' metals, thereby minimizing their bioavailability and mobility in soils.

Alkalinization can also be achieved by the application of biochars (Bolan et al., 2023). There has been considerable research on their uses in the last decade, often focused on tropical soils (Dai et al., 2017; Awad et al., 2018; Li et al., 2018; Basak et al., 2022). Among the advantages of the use of biochars over traditional liming materials is that they can be produced locally, and they are reputed to have other beneficial effects on soil fertility and physical structure.

In conclusion, there is a rather narrow pH range in fertile agricultural land. Various microbiological and chemical processes contribute to soil acidification and therefore management may be required to avoid excessive acidification.

References

- Awad YM, Wang JY, Igalavithana A, Tsang DCW, Kim KH, Lee SS, and Ok YS (2018) Biochar effects on rice paddy: Meta-analysis. *Advances in Agronomy* 148: 1–32.
- Basak BB, Sarkar B, Saha A, Sarkar A, Mandal S, Biswas JK, Wang H, and Bolan NS (2022) Revamping highly weathered soils in the tropics with biochar application: What we know and what is needed. *Science of the Total Environment* 822: 153461.
- Bolan N, Sarmah AK, Bordoloi S, Bolan S, Padhye LP, Van Zwieten L, Sooriyakumar P, Khan BA, Ahmad M, Solaiman ZM, Rinklebe J, Wang HL, Singh BP, and Siddique KHM (2023) Soil acidification and the liming potential of biochar. *Environmental Pollution* 317: 120632.
- Chen J, Lu SY, Zhang Z, Zhao XX, Li XM, Ning P, and Liu MZ (2018) Environmentally friendly fertilizers: A review of materials used and their effects on the environment. *Science of the Total Environment* 613: 829–839.
- Condon J, Burns H, and Li GD (2021) The extent, significance and amelioration of subsurface acidity in southern New South Wales, Australia. *Soil Research* 59: 1–11.
- Dai ZM, Zhang XJ, Tang C, Muhammad N, Wu J, Brookes PC, and Xu JM (2017) Potential role of biochars in decreasing soil acidification—A critical review. *Science of the Total Environment* 581: 601–611.
- Fageria NK and Baligar VC (2008) Ameliorating soil acidity of tropical oxisols by liming for sustainable crop production. *Advances in Agronomy* 99: 345–399.
- Fageria NK and Nasciente AS (2014) Management of soil acidity of South American Soils for Sustainable Crop Production. *Advances in Agronomy* 128: 221–275.
- Gilliam FS, Burns DA, Driscoll CT, Frey SD, Lovett GM, and Watmough SA (2019) Decreased atmospheric nitrogen deposition in eastern North America: Predicted responses of forest ecosystems. *Environmental Pollution* 244: 560–574.
- Hartemink AE and Barrow NJ (2023) Soil pH-nutrient relationships: The diagram. *Plant and Soil* 486: 209–215.
- Holland JE, Bennett AE, Newton AC, White PJ, McKenzie BM, George TS, Pakeman RJ, Bailey JS, Fornara DA, and Hayes RC (2018) Liming impacts on soils, crops and biodiversity in the UK: A review. *Science of the Total Environment* 610: 316–332.

Kipling C (2022) *Acid Soils*. Excelic Press 323.

Li YF, Hu SD, Chen JH, Muller K, Li YC, Fu WJ, Lin ZW, and Wang HL (2018) Effects of biochar application in forest ecosystems on soil properties and greenhouse gas emissions: A review. *Journal of Soils and Sediments* 18: 546–563.

Nguyen T (2023) Lime requirement for agricultural soils: A review of the current and potential methods for routine use. *Journal of Plant Nutrition and Soil Science* 186: 23–29.

Rengel Z (2003) *Handbook of Soil Acidity*. New York, NY: Marcel Dekker.

Ulrich B and Sumner ME (1991) *Soil Acidity*. New York, NY: Springer-Verlag.

pH

Siobhan Staunton, UMR Eco&Sols, INRAE, INRAE-IRD-Cirad-Institut Agro-Université de Montpellier, Montpellier, France

© 2023 Elsevier Ltd. All rights reserved.

This is an update of NS Bolan, K Kandaswamy, pH, Editor(s): D Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 196–202, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00210-1>.

Introduction	8
Theoretical basis of pH	8
Measurement of pH	10
Potentiometric method	10
Determination of pH using pH-sensitive dyes	11
Measurement of soil pH	11
Factors affecting soil pH	12
Carbon dioxide	12
Salt effect	13
Suspension effect	13
Importance of soil pH	14
Summary	14
References	14

Abstract

pH is undoubtedly the most important chemical soil characteristic that determines the chemical environment of higher plants and soil microbes. There are few reactions involving any component of the soil or of its biological inhabitants that are not sensitive to soil pH. This sensitivity must be recognized in any soil-management system. pH is defined. Some approaches to the measurement of soil pH are briefly described. The factors affecting pH and its measurement are discussed.

Key points

- pH is an important chemical property defining many chemical and biological processes.
- Soil pH measurement is subject to possible artefacts.
- Planar optodes allow spatio-temporal mapping of pH in soil.

Introduction

The acidity-alkalinity status of a soil is referred to as 'soil reaction,' which is often quantified by measuring soil pH. pH is considered to be the 'master variable' of a soil, which determines many of the physical (e.g., flocculation or deflocculation), chemical (e.g., adsorption or desorption), and biological (e.g., mineralization or immobilization) properties of soils. In this chapter, the fundamentals of soil pH will be discussed in relation to its measurement and practical implications.

Theoretical basis of pH

Pure water dissociates into the hydronium ion (H_3O^+), which is associated with acidity, and hydroxyl ion (OH^-), which is connected with alkalinity:



In aqueous media, the free acid cation is mostly present as the hydrated proton (H_3O^+) but for convenience it is often referred to as a hydrogen ion or proton (H^+), and thus the dissociation of water is often represented simply as:



This is a reversible reaction that obeys the law of mass action:

$$[\text{H}^+] \times [\text{OH}^-] / [\text{H}_2\text{O}] = K_w \quad (3)$$

where K_w is a constant and $[H^+]$ is the hydrogen ion concentration. Because the degree of dissociation of pure water is so small, the concentration of undissociated H_2O molecules always remains constant so that the product $[H^+] \times [OH^-]$ gives a constant value (K_w), known as the dissociation or ionization constant of water. The K_w value for pure water at 25 °C is 1×10^{-14} , with equal concentrations of H^+ and OH^- ions (1×10^{-7} M):

$$K_w = [H^+] [OH^-] = 1.0 \times 10^{-14} \quad (4)$$

If a chemical substance is added to water, which either donates or consumes H^+ or OH^- ions, then the concentration balance between H^+ and OH^- ions is disturbed. The numerical product of the concentrations must, however, remain constant so that if $[H^+]$ increases, $[OH^-]$ must decrease accordingly, and vice versa.

To simplify the arithmetical representation of Eq. (4), Sørensen suggested that the equilibrium state could be defined simply by reference to the $[H^+]$ and, to avoid the use of cumbersome decimal fractions, he proposed the use of a term designated 'pH' (French *pouvoir hydrogène* or *puissance d'hydrogène*, or 'hydrogen power') which is defined as the $\log_{10}$ of the reciprocal of the $[H^+]$ measured in Eq. (5):

$$pH = \log_{10} 1/H^+ \text{ or } -\log_{10}(H^+) \quad (5)$$

Expressing Eq. (4) as a negative logarithm yields:

$$pH + pOH = 14 \quad (6)$$

where pH and pOH are the negative logarithms of the H^+ and OH^- ion activities, respectively.

Thus, the pH of an aqueous solution is defined as the negative logarithm of the H^+ ion activity in the solution, which is the same as the $[H^+]$ when the salt concentration in the solution is very small. This concept can be applied to a system in which the individual molecules and ions are uniformly dispersed throughout the solution within the limits set by variations due to molecular and ionic thermal movements.

The pH scale ranges from 0 to 14, with pH 7 as the neutral point. At pH 7, $[H^+]$ equals the $[OH^-]$. From pH 7–0, the solution is increasingly more acidic; from pH 7–14, the solution is increasingly more alkaline (basic; Fig. 1). The $[H^+]$, which is the ion measured when determining pH, has a 10-fold change between each whole pH number. Thus, a solution with pH 5 will have 10 times more $[H^+]$ than a solution with pH 6.

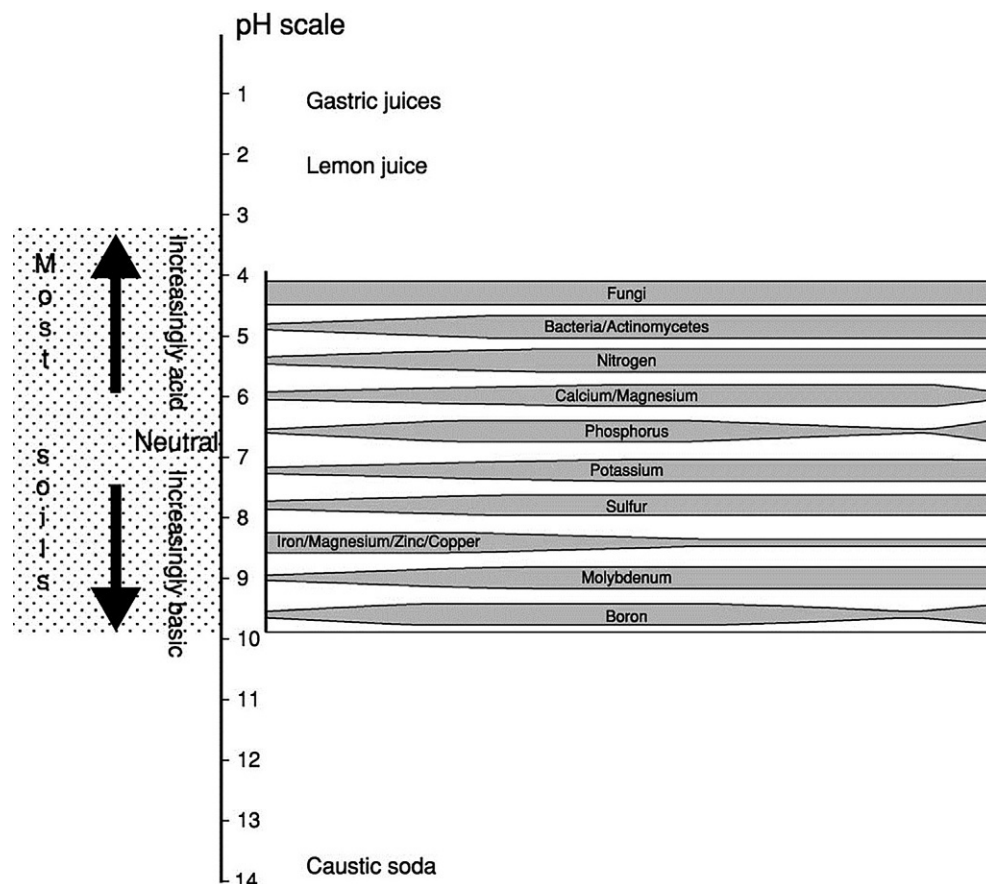


Fig. 1 pH scale and the effect of pH on nutrient availability and microbial activity.

Measurement of pH

pH is commonly measured either electrometrically using potentiometers or colorimetrically using pH-sensitive dyes.

Potentiometric method

This method involves measurement of the potential difference between an electrode whose potential is determined by the $[H^+]$ of the solution and a reference electrode whose potential is independent of $[H^+]$. The normal hydrogen electrode is the glass electrode, and the reference electrode is a calomel electrode in which calomel ($HgCl_2$, mercury chloride) is bathed in a solution of potassium chloride (KCl) (Fig. 2).

The glass electrode consists of a thin glass bulb containing dilute HCl (hydrochloric acid), into which is inserted an Ag—AgCl (silver-silver chloride) wire, serving as the electrode with a fixed voltage. The HCl solution is separated from the test solution by a membrane of special glass, usually a lithium silicate of particular composition. Differences in H^+ activity across this membrane cause a difference in electrical potential, which can be measured by a potentiometer.

When the glass bulb is immersed in a solution, a potential difference develops between the solution in the bulb and that outside the bulb. The electrical potentials developed by this electrode are the membrane potential, plus the potential of the Ag—AgCl—HCl reaction inside the electrode:



This reaction is reversible and the Ag—AgCl potential ($E^0 = 0.222$ V at $25^\circ C$) is likely to remain constant. This potential is cancelled out when the electrode is standardized against a standard pH buffer solution.

The glass electrode develops a second potential at the membrane separating the standard HCl and the test solutions. The minute current flow required by the pH meter develops a double layer in the solutions, causes ion exchange at the inner and outer surfaces, and causes diffusion of ions across the glass membrane. The potentials across the glass membrane can be closely calibrated to the approximate value of the H^+ activity.

A typical reference electrode consists of a wire dipped into the liquid mercury, which makes electrical contact with the pH meter. Current flows from the electrode to the solution phase through the reversible reaction ($E = 0.268$ V at $25^\circ C$):



The Cl^- (chloride) activity is fixed by the KCl concentration (usually saturated KCl). As a result this potential also cancels out when the pH electrode system is standardized in a standard pH buffer. The KCl diffusion through the orifice makes the electrical contact between the reference cell and test solution. This KCl connection forms a 'salt bridge' between the test solution and the reference electrode. The reference electrode is often built in the same body as the glass electrode to form a combination electrode.

When measuring pH, the potential (E) in the glass electrode is formulated by the Nernst equation:

$$E = (RT/nF) \log(K/H^+) \quad (9)$$

where R is the gas constant, T the absolute temperature, n the valence, F the Faraday constant, K a constant, and H^+ the activity of H^+ ions. The E is called the 'half-cell potential' and cannot be measured alone. If the glass electrode is placed against a reference calomel electrode, the potential difference between the two ($E - E_{ref}$) is measurable.

Before any pH measurement, the two electrodes have to be calibrated using solutions of known pH. This is called 'standardizing' the electrodes and the pH meter. The overall potential of the total cell E_0 equals $E - E_{ref}$.

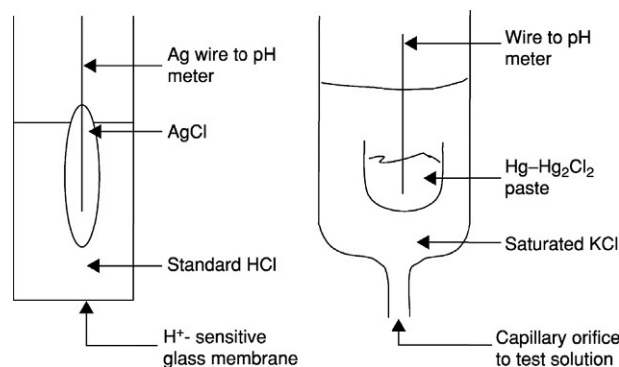


Fig. 2 Hydrogen (glass) and reference (calomel) electrodes. Reproduced from Tan KH (1998) *Principles of Soil Chemistry*. New York: Marcel Dekker.

$$E_0 = [(RT/F) \log(K/H_{\text{std}}^+)] - E_{\text{ref}} \quad (10)$$

If the two electrodes are now placed in the solution with the unknown $[H^+]$, the potential E_c is:

$$E_c = [(RT/F) \log H_{\text{test}}^+] - E_{\text{ref}} \quad (11)$$

Subtracting Eq. (10) from Eq. (11) gives:

$$E_c - E_0 = (RT/F) [\log H_{\text{test}}^+ - \log H_{\text{std}}^+] \quad (12)$$

$$E_c - E_0 = (RT/F) \log (H_{\text{test}}^+/H_{\text{std}}^+) \quad (13)$$

or

$$E_c - E_0 = (RT/F) (\text{pH}_{\text{std}}/\text{pH}_{\text{test}}) \quad (14)$$

where E is the measured potential and is converted to pH units by the scale of the pH meter. The E_{ref} includes all other potentials, which are nullified by standardizing the systems with a standard pH buffer solution.

Since $RT/F \approx 0.0591$ at 25°C :

$$\text{pH} = (E_c - E_0)/0.0591 \quad (15)$$

This means that for a change of 1 pH unit, the potential will change by 0.0591 V.

The glass electrode has a uniform response to a wide range of H^+ activities and is less sensitive to interfering ions. The extent of interference is denoted by the selectivity ratio, the concentration ratio of the test ion to interfering ion at which the interfering ion exerts a significant potential at the membrane. The selectivity ratio of the pH glass membrane for H^+ over Li^+ , the most serious interfering ion, is approximately 10^9 . That is, Li^+ would cause a significant pH error in a pH 9 solution containing $1 \text{ mol L}^{-1} Li^+$ or in a pH 10 solution containing $0.1 \text{ mol L}^{-1} Li^+$. The selectivity ratio for Na^+ , the next most serious interference, is approximately 10^{13} .

The miniaturization of pH electrodes has allowed pH to be measured directly in moist soil or gels containing soil. This can allow pH gradients or dynamics to be monitored. However this approach is limited by the fragility of glass electrodes in contact with abrasive soil particles.

Determination of pH using pH-sensitive dyes

Rapid measurement of pH can also be achieved colorimetrically with pH indicator dyes. Colorimetric methods are based on the change in color that takes place on the dissociation of a weak acid, or weak base organic dye; thus for a weak acid dye:



where raising the pH will cause a change in color as the undissociated dye (HI) changes to the dissociated form (I^-). Typically a mixture of dyes is used to yield color changes that take place over a wide range of pH values. However, the color can only be measured accurately in a solution from which the solid particles are removed; the color also depends on the concentration of various salts present in the solution. A dye solution is mixed with the solid to make a slurry, the dye is decanted and the color is visually compared with a color chart. The pH measured in this manner was considered to be less precise than the results obtained using the glass electrode.

Nevertheless much progress has been made with this approach. Indicators are often dissolved in gels allowing color intensity to be measured non-destructively using portable spectrometers. This makes the technique much more quantitative than a simple visual comparison of color and intensity. A variant on this is to place thin gels containing indicators on top of a layer of soil, in a manner analogous to X-ray film. The indicator is assumed to remain in the gel and be immobile. This gel technique is also used to obtain maps of soil catalytic activity; zymography. The application of optodes can avoid the problem of mobility of either indicator molecules of the proton or hydroxide couple within a porous gel. Optodes may be planar, like the gel film (Faget et al., 2013; Ma et al., 2019; Merl et al., 2022), although the earliest were fiber optics, used in the same way as a glass electrode (Motellier et al., 1995). The use of fluorescent imaging has further improved the sensitivity of this approach (Rudolph et al., 2013).

Measurement of soil pH

Because 'pH' is a term that is only defined for solution, in a strict sense, it cannot be applied to a solid-phase material such as soil. However, the chemical properties of the solid-phase components in soil define the pH of the soil solution (i.e., the solution in the soil pores). Further, the pH of a soil dispersed in water is not a simple concept like that of a solution. When soil particles, which carry ions attached to them, are dispersed in solution, the ions are not uniformly distributed throughout the solution. The concept of the pH of a soil (or soil suspension) can therefore only be discussed in relation to the properties of the ionic atmosphere around the soil particles (i.e., soil-solution interface).

When a negatively charged soil particle is dispersed in water, the negative charge is neutralized by cations, some of which sit firmly on the clay surface, forming the Stern layer, and some dissociate into the dispersion medium to form the Gouy diffuse double layer. The thickness of this double layer depends in part on the ions dissolved in the dispersion medium, and the thickness increases with a decrease in the concentration of the ions, a decrease in the valency of the cations, and an increase in the size of the cation. The $[H^+]$ in the solution surrounding the soil particles is less than that close to the soil particles due to the $[H^+]$ gradient in the double layer. As the double layer is made more compact by adding an electrolyte to the soil-water system, the $[H^+]$ gradient across the double layer is reduced, and the pH in the solution becomes almost equal to that close to the surface of the soil particle. Thus the pH of the bulk solution is greater than the pH just outside the Stern layer unless the salt concentration in the solution is large.

The pH of most soils ranges from 3 to 8.5. Soils do not normally have a pH below 3–4 unless they contain free acids (e.g., acid sulfate soils and acid mine drainage). Similarly soils normally only have a pH >8.45 if they contain enough exchangeable sodium (i.e., alkali soils). Further, in calcareous and alkali soils, the effective pH around plant roots can be determined by the concentration of carbon dioxide (CO_2) produced during respiration by roots and microorganisms.

Strictly, the pH of a soil should refer to the $[H^+]$ in the soil solution. This is very difficult to measure, and hence soil pH is often referred to measurements made in the soil suspension. Three common, standard methods for determining pH in soils involve the suspension in distilled water, 0.01 mol L⁻¹ $CaCl_2$, or 1 mol L⁻¹ KCl solutions at a soil-to-solution (weight-to-volume) ratio of 1:1, 1:2.5 or 1:5 (recommendations vary between countries).

The pH of a solution in equilibrium with a soil varies with the composition and concentration of the salts in the solution. This is because cations in solution displace H^+ and Al^{3+} ions from soil surfaces. The measurement of soil pH in a water extract is expected to be higher than the pH of the soil solution of natural soil water contents. The excess quantity of water added to prepare the suspension results in a dilution of the salts in the soil solution. Since soils always contain some soluble salts, such as nitrates and chlorides, the apparent pH of the soil will depend on the amount of water added to the soil to make the suspension. Further, the amount of soluble salts in a soil varies continuously throughout the season, depending on factors such as the amount of drainage water percolating through the soil.

Soil pH values measured in 0.01 mol L⁻¹ $CaCl_2$ or 1 mol L⁻¹ KCl salt solutions are generally less than those obtained using water, because the cations in these salts tend to displace H^+ and Al^{3+} from soil materials to the soil solution. These salt solutions are sufficiently concentrated to overwhelm variations in salt concentration in soil solution, thereby minimizing the effect of salt concentration on pH in most soils. These solutions also eliminate the junction potential effect. Calcium is used in the determination of pH because it is the predominant soil solution cation in soils of temperate regions. The 0.01 mol L⁻¹ concentration, however, is somewhat higher than that found in most soil solutions, and thus the pH values obtained by this method are usually lower than the values of soil solutions.

For some subsurface horizons of highly weathered soils of the tropics, pH in the salt solution can actually be greater than pH in water, especially in subsoils that are low in permanent-charge silicate clays and organic matter. Ion exchange in these soils is dominated by oxides and hydroxides of Fe and Al, and at acid pH values they are net positively charged, creating an effective sink for anion adsorption. Thus, on salt addition (i.e., $CaCl_2$ or KCl) more OH^- is displaced by Cl^- than H^+ by cations, resulting in an increase in pH.

Factors affecting soil pH

The pH of a soil is affected by the concentration of CO_2 in the soil air, salt concentration (salt effect), and the presence of colloidal particles (suspension effect).

Carbon dioxide

The larger the CO_2 concentration in the soil solution, the lower the pH; the pH of a neutral or calcareous soil is very sensitive to small changes in CO_2 concentration. In calcareous soils, its magnitude can be calculated from the solubility of calcium carbonate and CO_2 in water, because in such soils the pH is determined by the system, calcium carbonate-carbon dioxide-water, which is given approximately by:

$$2pH = K + pCa + pCO_2 \quad (17)$$

where pCO_2 is the negative logarithm of the partial pressure of CO_2 in equilibrium with the solution, pCa the negative logarithm of the activity of the calcium ions, and K is a constant whose value lies between 10 and 10.5, depending on the value used for the solubility constant of calcium carbonate. The decrease in pH is therefore approximately proportional to the logarithm of the partial pressure of CO_2 .

Thus, the pH of a well-aerated calcareous soil containing exchangeable sodium or magnesium ions cannot exceed approximately 8.5 when in equilibrium with the atmosphere; but if the CO_2 concentration rises to 0.1% (i.e., a small value for most soils) the pH drops to 8, and if it rises to 1% (a common value for pasture soils) the pH will be less than 7.5. Many plants suffer from phosphate and minor element deficiencies if the pH rises much above 8, and hence they can only thrive on calcareous soils if they can maintain

an adequate concentration of CO_2 around their roots. This is more easily achieved in soils of low porosity, such as clays, than in soils of high porosity, such as well-drained sands, and in organic matter-rich pasture soils than in organic matter-depleted arable soils.

Salt effect

As discussed above, the pH of a soil depends on the salt concentration in the soil solution. This dependence of measured soil pH on salt concentration can be reduced by making all the measurements in a salt solution that is sufficiently strong to swamp the effects of the changes that occur naturally. But adding a salt solution to the soil will cause cation exchange to take place and, in particular, exchange of H^+ ions from clay surface with the cation of the added salt, which will give a pH that is lower than the soil pH under normal conditions. The added salt solution should therefore cause as little exchange as possible, and the use of 0.01 mol L^{-1} CaCl_2 or 1 mol L^{-1} KCl solution has been proposed for most arable soils on the grounds that they approximate the salt concentration in the soil solution and do not give appreciable junction potential with the calomel electrode. The pH of a soil measured in these solutions is more constant and is much less dependent on the soil-to-solution ratio than the pH measured in a water suspension, and it is also presumably closer to the pH of the solution around the plant roots.

Suspension effect

The difference in pH values between the suspension and its equilibrium solution is referred to as the 'suspension effect,' and this effect is identical to the Donnan e.m.f.

The electrode portion of a pH measurement circuit is presented in Fig. 3, where each vertical bar represents a phase boundary and the double bar represents the liquid-liquid junction. Thus, very appreciable junction potentials can be created between a soil suspension and the calomel electrode so that the measured potential difference between the two electrodes cannot be interpreted solely in terms of the $[\text{H}^+]$ outside the glass electrode.

A junction or diffusion potential always develops when two dissimilar substances or solutions of different composition come into contact. It is inevitable that a diffusion potential is created at the liquid junction between a salt bridge and a solution of different composition. An example in this case is the saturated KCl of the reference electrode diffusing into the test solution. At liquid junctions, differences in the mobilities of the cation and anion create a diffusion potential which is also influenced by any differences in the ion activities that exist across the junction. The potential is minimal when the rates of K^+ and Cl^- diffusion into the test solution are equal. It is therefore appropriate that an electrolyte in which the mobilities of the cation and anion are almost identical is used in the salt bridge. Indeed, KCl is used as the reference-cell electrolyte because of the similarity of K^+ and Cl^- diffusion rates in water. Because the current at the liquid junction is mainly carried by K^+ and Cl^- ions, which have similar diffusion coefficients, the liquid junction potential should be small. As long as the ionic strengths of the standard and unknown solutions are small compared with the concentrated KCl salt bridge solution, any liquid junction potential should, as well as being small, tend to be rather constant in magnitude and hence can usually be ignored.

Because factors involved in the liquid junction potential during pH calibration and subsequent pH measurement are likely to be different, the liquid junction potential contribution to the total e.m.f. of a cell will also be different. In a colloidal suspension, however, the colloid may cause K^+ and Cl^- to diffuse at different rates. Because of attraction and repulsion by the charged colloid, one ion moves ahead of the other. When a salt bridge makes contact through a suspension instead of clear solution, the situation is complicated by a reduced mobility of counter ions which arises from retarding effects created by the oppositely charged colloidal particles. This means that liquid junction potentials (E_j) are likely to be much greater in suspensions than in supernatant liquids. Accurate pH measurements require a negligible E_j , because such potentials are virtually unpredictable. The potentiometer measures all the potentials of the circuit and cannot distinguish between the H^+ ion potential at the glass membrane and the spurious E_j . The value of E_j at the interface between saturated KCl and a colloidal suspension of very small salt concentration can be as high as 240 mV, equivalent to more than 4 pH units (1 pH unit 59 mV). However, values of E_j greater than 30 mV are uncommon in soils, because salt concentrations in even highly leached soils are large enough to suppress E_j . Therefore measuring pH in salt solutions of 0.01 mol L^{-1} or more virtually eliminates E_j .

Suspension effects can be reduced by introducing the liquid junction of the reference electrode into clear supernatant liquid and not into the suspension. The rates of K^+ and Cl^- diffusion are then unaffected by soil colloid particles. The pH glass electrode, on the other hand, can be placed either in the supernatant solution or in the colloidal suspension. The H^+ activity is the same in both phases, and the glass electrode is unaffected by the presence of the colloid. However, it has been suggested that when measuring soil pH the glass electrode should be in the suspension on the basis that the suspension is better buffered than the supernatant liquid.

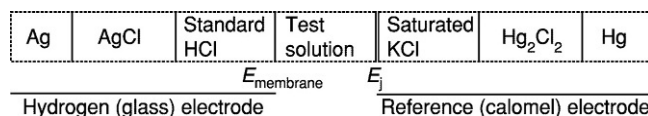


Fig. 3 The electrode portion of a pH measurement circuit.

Importance of soil pH

The soil pH is considered the ‘master variable’ which controls many other soil properties (Fig. 1). It greatly affects the solubility of minerals in soils. Most minerals are more soluble in acid soils than in neutral or slightly basic solutions. A low pH may indicate the deficiency of basic cations such as Ca^{2+} and Mg^{2+} in soils. Under acid soil conditions, the concentration of most ‘heavy’ metals, such as lead, cadmium, and mercury, reach phytotoxic levels. Strongly acidic soils usually have large and toxic concentrations of soluble aluminum and manganese. Azaleas (genus *Rhododendron*), tea (*Camellia sinensis* (L.) Kuntze), rhododendrons (*Rhododendron* L.), cranberries (*Vaccinium macrocarpon* Alton 1789), pineapple (*Ananas comosus* (L.) Merr.), blueberries (*Vaccinium* sect. *Cyanococcus* Rydb.), and several conifer timber species tolerate acidity and grow well in acid soils. In contrast, alfalfa (*Medicago sativa* L.), beans (*Phaseolus vulgaris* L.), barley (*Hordeum vulgare* L.), and sugarbeets (*Beta vulgaris* L.) grow well only in slightly acidic to moderately basic soils because of a high calcium demand to tolerate soluble aluminum.

Soil pH can also influence plant growth through its effect on beneficial microorganisms. For example, most nitrogen-fixing legume bacteria are not very active in strongly acidic soils. Bacteria that decompose soil organic matter and thus release nitrogen and other nutrients for plant use are also hindered by strong acidity. Similarly, earthworms are not active in acid soils, whereas fungi usually tolerate acidity better than do other microbes.

Soil alkalinity, although more difficult to manage than soil acidity, may also be undesirable for plants. Alkaline soils may reach pH values of more than 10 with increasing levels of exchangeable sodium. Plants on soils with $\text{pH} > 9$ usually have reduced growth or even die. Some plants (halophytes) can tolerate large salt concentrations or high pH. Soluble $\text{Al}(\text{OH})_4^-$ may exist at these high pH values and is also toxic to plants. The major effect of basic pH is to reduce the solubility of all micronutrients (except chlorine (Cl), boron (B), and molybdenum (Mo)), especially those of iron (Fe), zinc (Zn), copper (Cu), and manganese (Mn). Also, phosphate is often not readily available to some plants because of its precipitation in the soil solution by calcium or precipitation on solid calcium carbonate. Iron deficiency, associated with soils high in carbonates, has long been known, and is referred to as ‘lime-induced iron chlorosis.’

Summary

Soil pH is one of the most important chemical properties that control many other soil physical, chemical, and biological properties. The pH of a soil depends on the CO_2 concentration in the soil air and the salt concentration in the soil solution, and these are constantly changing. Dynamic measurements using planar optodes allow new insights into pH gradients and dynamics in soils, particularly in zones such as the rhizosphere.

References

- Faget M, Blossfeld S, von Gillhausen P, Schurr U, and Temperton VM (2013) Disentangling who is who during rhizosphere acidification in root interactions: Combining fluorescence with optode techniques. *Frontiers in Plant Science* 4: 392.
- Ma X, Mason-Jones K, Liu Y, Blagodatskaya E, Kuzyakov Y, Guber A, Dippold MA, and Razavi BS (2019) Coupling zymography with pH mapping reveals a shift in lupine phosphorus acquisition strategy driven by cluster roots. *Soil Biology and Biochemistry* 135: 420–428.
- Merl T, Rasmussen MR, Koch LR, Søndergaard JV, Bust FF, and Koren K (2022) Measuring soil pH at in situ like conditions using optical pH sensors (pH-optodes). *Soil Biology and Biochemistry* 175: 108862.
- Motellier S, Noir MH, Pitsch H, and Durdault B (1995) pH determination of clay interstitial water using a fiber-optic sensor. *Sensors and Actuators B* 29: 345–352.
- Rudolph N, Voss S, Moradi AB, Nagl S, and Oswald SE (2013) Spatio-temporal mapping of local soil pH changes induced by roots of lupin and soft-rush. *Plant and Soil* 369: 669–680.

Buffering capacity

Siobhan Staunton^a and Ravi Naidu^b, ^aUMR Eco&Sols, INRAE, INRAE-IRD-Cirad-Institut Agro-Université de Montpellier, Montpellier, France; ^bCentre for Environmental Remediation (GCER), College of Engineering, Science and Environment, The University of Newcastle, Callaghan, NSW, Australia; CRC for Contamination Assessment and Remediation of the Environment (crcCARE), The University of Newcastle, Callaghan, NSW, Australia

© 2023 Elsevier Ltd. All rights reserved.

This is an update of BR James, Buffering Capacity, Editor(s): D Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 142–147, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00180-6>.

Introduction	15
The samovar analogy for buffering in soils	16
Buffering of soil acidity	17
Dissolution and precipitation of CaCO₃ and Al(OH)₃	18
Cation exchange reactions and pH buffering	18
pH-dependent surface charge and pH buffering	19
Buffering of soil oxidation-reduction status	19
Buffering of ion activities via dissolution-precipitation, ion exchange, and ligand complexation	20
Summary	21
References	21

Abstract

Buffering is the capacity to resist changes and involves mineralogical, chemical and biological processes. Some soil ecological concepts are discussed. Most of this contribution focuses on pH-buffering in soils. The roles of minerals and organic matter in maintaining soil pH are discussed. The buffering of oxido-reduction (redox) status of soils is briefly presented.

Key points

- Buffer capacity is the capacity to resist change.
- The term is predominantly used in chemistry to describe pH-buffering.
- In soil science nutrient buffering, especially for potassium and phosphate, is an important concept.
- Buffering has been used to consider the resilience of ecosystems in ecology.

Introduction

Buffering is the resistance of a system to change in response to a perturbation, and it is a key attribute of soils from the molecular level to the landscape scale. Chemical, physical, and biological processes may raise or lower solute concentrations (or activities) as intensive variables in soil solution, thereby temporarily disturbing dynamic equilibria or steady-state conditions between soil water and solid or gas phases of the soil. In response to such a perturbation, one or more processes may release solute to soil solution or remove it to restore wholly or partially the original concentration. The reservoir of such solutes in solution, in solid phases, or in the soil atmosphere for restoration of soil solution chemical composition is known as the ‘capacity factor.’ The measurable change in the capacity factor per unit change in the intensive variable is called the ‘buffer index,’ ‘buffer intensity,’ or, more commonly, ‘buffer capacity.’ Myriad chemical reactions (buffering mechanisms) govern such release or uptake of ions and molecules between soil solution and solid or gas phases, including cation and anion exchange, oxidation-reduction reactions, dissolution-precipitation processes, and metal-organic ligand complexation. Understanding and quantifying buffer capacities and buffering mechanisms at the colloid and molecular level of soils can aid in predicting the sensitivity and resilience of soil-water systems to anthropogenic and natural perturbations of ecosystems.

In soil science buffer capacity of pH (Zhuang et al., 2006; Bloom and Skjellberg, 2011) and some major nutrients (Evangelou et al., 1994; Nair, 2013; Barrow, 2015; Barrow, 2022; Nair, 2022; Certini and Scalenghe, 2023) have been studied for decades. More recently the concept is considered for carbon (C) sequestration (Xu et al., 2020) and ecology (Trabelsi and Mhamdi, 2013; Doody et al., 2016; Certini and Scalenghe, 2023).

Ecological theory related to succession dynamics of disturbed ecosystems identifies that natural systems, including soils, are constantly and naturally recovering from regular and irregular disturbances of various severities. The processes and mechanisms of

recovery determine the biodiversity and stability of ecosystems, and buffering in soils and natural waters is a key control of nutrient availability and pollutant bioavailability. Buffering reactions influence water quality and community regrowth, migration, and recruitment in disturbed patches on the landscape. The success of human efforts to restore disturbed ecosystems and soils is also determined in part by the scale and nature of soil-buffering processes, as related to natural ones that govern element transformations.

The samovar analogy for buffering in soils

Concepts of the intensive variable, capacity factor, buffer capacity, and buffering mechanisms for soils can be modeled and visualized by analogy with old-fashioned Russian samovars used to make large volumes of tea (Fig. 1). Each samovar comprises a large copper reservoir to boil water and a narrow glass tube on the outside of the samovar's tank to indicate how full the tank is. As boiling water is drawn out of the tank through the spigot, the indicator tube empties quickly and temporarily, but is refilled to the new tank level when the spigot is shut off (as indicated by the arrows between the tank and indicator tube in Fig. 1). To decrease the volume of a samovar tank by a given fraction of its total capacity (e.g., from level 1 to level 2 in the indicator tube), a greater volume of water must be withdrawn from the larger samovar than from the smaller one.

Since the level of water in the indicator tube is proportional to how full each tank is, it is independent of the volume of the samovar. The height of the water column is analogous to an intensive variable in soil solution such as pH, a measure of how acidic an aqueous sample is, but not an indicator of its volume. The volume of the samovar represents the capacity factor (e.g., total soil acidity), and the change in the tank volume per unit change in the level of the indicator tube is the buffer capacity. Restoration of the water level in the indicator tube from the samovar after shutting off the spigot is due simply to the flow of water, i.e., the buffering mechanism in this model.

In this simple comparison of two samovars of different volumes, but with identical fractions of their capacities filled with water, as the boiling water is drained completely, the buffer capacity (change in tank volume per change in water level) remains constant. In complex soil colloid-water systems, however, buffer capacities do not usually remain constant as an intensive variable is changed, and the magnitude of the buffer capacity will change accordingly. This is modeled by a peculiar samovar with a narrow middle section and wide sections at the top and bottom (Fig. 2). In this samovar model, the system would be well buffered at the beginning and end of draining the tank, but would be poorly buffered in the middle. In soils, measurements of intensive variables are often made routinely (e.g., pH, Ca concentration, partial pressure of O_2 , and oxidation-reduction potential), and are used to indicate the energy state or 'ability of the system to do work' in affecting other coupled systems such as groundwater or plants rooted in the soil.

The size of the capacity factor and buffering mechanisms linking the intensive and capacity parameters are often not known with the same accuracy as are measurements of intensive variables in soils (i.e., the samovar tank is invisible). Therefore, investigations of the nature and scale of buffering processes in soils are needed to understand controls on the value of an intensive variable and what chemical and biological processes control its resistance to change.

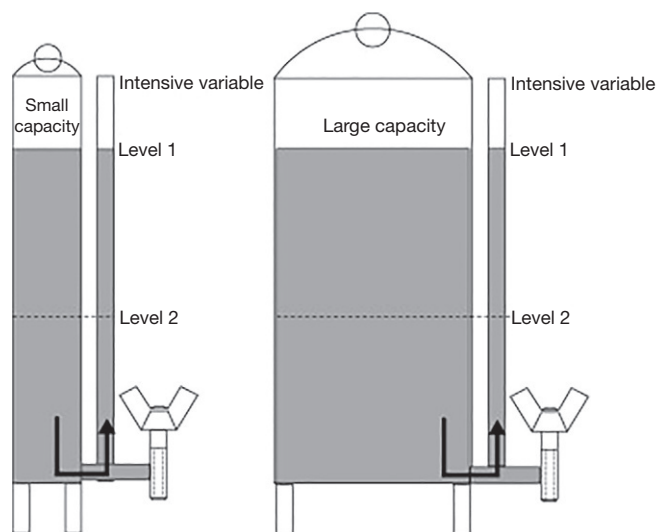


Fig. 1 Samovar models representing small and large buffer capacities (tank sizes) in equilibrium with the same level of intensity variable (represented by the height of liquid in the narrow indicator tube outside the tank). Adapted from Brady NC and Weil RR (2002) *The Nature and Properties of Soils*, 13th edn. Upper Saddle River, NJ: Prentice-Hall.

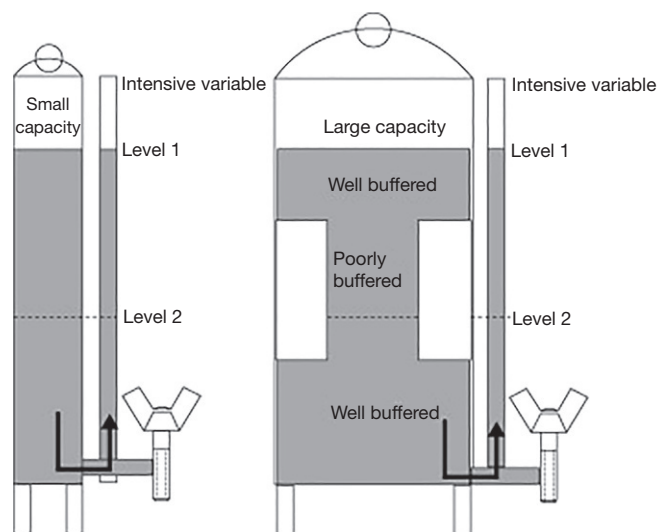


Fig. 2 Samovar models representing small and large buffer capacities (tank sizes) in equilibrium with the same level of intensity variable (represented by the height of liquid in the narrow indicator tube outside the tank). The larger tank with the hourglass shape models a buffered system in which the buffer capacity (change in tank volume per unit change in the height of the liquid in the indicator tube) is not constant and varies during titration (drainage of the tank). Adapted from Brady NC and Weil RR (2002) *The Nature and Properties of Soils*, 13th edn. Upper Saddle River, NJ: Prentice-Hall.

The samovar analogy for soil buffering will be used to explain similarities and differences among several soil chemical processes responsible for the buffering of soil acidity, oxidation-reduction status, and metal ion activities controlled by dissolution-precipitation and organic complexation by ligands. Soil acidity and oxidation-reduction status of soils are considered master variables that control the steady-state chemical composition of soils and kinetics of the approach to chemical equilibrium. As a result, predicting their resistance to change in response to disturbance of soil conditions is relevant to many environmental quality, plant nutrition, and human health effects of soils. Dissolution and precipitation reactions are important in weathering, soil development, and leaching processes, and also in controlling nutrient and pollutant solubility, speciation, bioavailability, and cellular absorption. Organic and inorganic ligand complexation, typically of cations, controls ion activity in equilibrium with soluble and insoluble forms of the element. Each of these processes is a coupling of an ion activity (intensive variable), a capacity factor (reserve of ion), and a buffering capacity and buffering mechanism that govern how the intensity and capacity factors interrelate in a soil system.

Buffering of soil acidity

The intensive variable for soil acidity is conceptually defined as the hydrogen ion activity of soil pore water or 'active acidity,' and is operationally defined by measuring pH in a soil-water or dilute salt suspension with a glass-reference electrode system, where pH is $-\log(H^+)$ and (H^+) is the hydrogen ion or proton activity in units of moles per liter. To neutralize the active acidity (e.g., raising pH from 5.0 to 7.0) in the soil solution of the top 15 cm of a hectare of loamy soil at field capacity moisture (-10 kPa water potential) containing 200 g montmorillonite and 30 g organic matter per kilogram of soil would require approximately 500 g of $CaCO_3$. This is just a handful of lime, compared with the $10,000 \text{ kg } CaCO_3 \text{ ha}^{-1}$ that might be needed to raise the whole soil pH to 7.0. The 20,000:1 ratio of lime needed to neutralize the capacity factor or 'reserve acidity' of the soil to that needed to neutralize the active acidity reflects the relative size of the metaphorical samovar tank and its indicator tube for soil acidity.

Soil properties that determine the buffer capacity for soil pH include initial pH, aluminosilicate clay content, and mineralogy; organic-matter content; and $Al(III)$, $Fe(III)$, and $Mn(III,IV)(hydr)oxide$ contents. Soil pH buffer capacity is determined by adding increasing quantities of acid or base to a given mass of soil and measuring pH after allowing sufficient equilibration time for pH to stabilize. It is often designated:

$$\beta = dC/dpH \quad (1)$$

where β is the buffer index or capacity in units of moles of H^+ or OH^- added per kilogram of soil (dC) per unit pH change, dpH. The buffer capacity is quantified as the reciprocal of the slope of a linear relationship between measured pH plotted on the ordinate versus moles of H^+ or OH^- added per kilogram of soil on the abscissa.

Other, quicker indicator tests for reserve or total acidity as the capacity factor have been developed to estimate 'lime requirement' for acidic agricultural soils used to grow crops that are sensitive to acid soil conditions. Examples of these ways to estimate reserve acidity include chemical extractions of exchangeable and reactive Al^{3+} , measures of cation exchange capacity (CEC) and the fraction

of exchange sites occupied by 'basic' cations (predominantly Ca^{2+} (calcium), Mg^{2+} (magnesium), K^+ (potassium), and Na^+ (sodium)), measurements of decreases in the pH of well-buffered solutions added to the soil sample, and equilibrations of a soil sample with a BaCl_2 (barium chloride) solution buffered at pH 8.2, followed by titration to pH 5 to determine the quantity of acid that reacted with the buffer. Measurements of free CaCO_3 (calcium carbonate) content can be used to estimate the total alkalinity of a soil and its buffer capacity against acidification.

Diverse chemical reactions occurring on colloid surfaces of clay, oxides, and organic matter are responsible for pH buffering. They involve cation exchange reactions, dissolution and precipitation of sparingly soluble compounds, and surface charge changes in response to pH. In most soils, pH buffer capacity and buffering mechanisms are due to combinations of these reactions, and titrations and quick tests for total acidity do not distinguish between them. Nevertheless, understanding the relative importance of these processes and their chemical basis allows predictions of the buffering behavior of diverse soils, and it enables extrapolation to unstudied soil systems and environments, particularly related to ecological processes, soil contamination, and soil remediation.

Dissolution and precipitation of CaCO_3 and $\text{Al}(\text{OH})_3$

Soils containing free CaCO_3 have a pH between 7.0 and 9.5 (typically 8.0–8.5) as governed by the dissolution and precipitation of CaCO_3 in water containing partial pressures of CO_2 as high as 10 times that of atmospheric levels. Acidity produced by the hydration of dissolved CO_2 and by other acid-generating reactions reacts with CaCO_3 in accordance with:



In the presence of large concentrations of soluble or exchangeable Mg^{2+} or Na^+ , soluble MgCO_3 (magnesium carbonate) or Na_2CO_3 (sodium carbonate) will form and allow the soil pH to rise to values higher than that controlled by CaCO_3 – CO_2 equilibria (sometimes to pH >10). Addition of base or acid will be buffered by the lime in the soil, thereby maintaining its pH at approximately 8.2.

Under strongly acid soil conditions, Al^{3+} in soil solution and on cation exchange sites hydrolyzes and generates acidity, while the dissolution of $\text{Al}(\text{OH})_3$ (aluminum hydroxide) neutralizes it, leading to large β values at pHs <4.5:



Soluble and exchangeable Al^{3+} is released from aluminosilicate clays upon weathering of Si-dominated tetrahedral sheets (releasing soluble monosilicic acid, $\text{Si}(\text{OH})_4$), thereby exposing Al-dominated octahedral sheets to H^+ attack. As a result, the buffer capacity of strongly acid soils (pH <5) is dominated by Al chemistry.

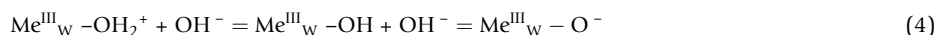
Cation exchange reactions and pH buffering

In the intermediate pH range between 4.5 and 6.5, cation exchange reactions on permanent aluminosilicate cation exchange sites, and on pH-dependent sites of organic matter and (hydr)oxides govern the uptake and release of H^+ and OH^- as controls on the buffering-capacity size and mechanism for pH control. Permanent, negatively charged cation exchange sites on aluminosilicate clay minerals are due to isomorphous substitution of lower charge cations for Si^{4+} (e.g., Al^{3+}) in tetrahedral sheets and for Al^{3+} (e.g., Mg^{2+} and Fe^{2+}) in octahedral positions. Cation exchange sites created in this way during clay mineral formation are dominated by Ca^{2+} at and above pH 7 and become increasingly satisfied by H^+ and Al^{3+} at lower pH values. In this way, the so-called base saturation (an intensive variable reflecting the fullness of the acidity samovar) decreases as exchange sites are occupied by Al^{3+} and H^+ .

Organic matter functional groups (principally carboxylic acids, represented by $\text{R}-\text{COOH}$, and phenolic acid groups, such as $-\text{OH}$ on aromatic rings) are weak acids with association constants (pK_a s) in the range of 2–7 for carboxylic acids and 7–10 for phenolic acid groups. These groups deprotonate at pHs > pK_a , and protonate at pHs < pK_a . When $\text{pH} = \text{pK}_a$, maximal pH buffering is observed. The complexity of humic and fulvic acid molecular aggregates and functional groups leads to overlapping pK_a s that are not independent and which change as adjacent groups are protonated or deprotonated. In addition, conformational changes in the folding patterns of humic acid polymers as the suite of cations changes during titration result in changeable pK_a s for the cation exchange sites responsible for pH buffering. As a result, titration curves of soil organic matter are often linear and do not reflect sharp inflection points characteristic of mono- or poly-protic simple organic acid titration curves. In addition to exchange reactions on organic acids between H^+ , Al^{3+} , and Ca^{2+} , some precipitates, such as calcium oxalate, may exist in organic matter-rich horizons, such as in forest soils. These Ca-organic C precipitates may contribute to pH buffering through dissolution and precipitation reactions similar to those of CaCO_3 , but at soil pHs <4.

pH-dependent surface charge and pH buffering

Oxide and hydroxide coatings of Fe(III), Mn(III,IV), and Al on sand, silt, and clay minerals are dominated by oxygen- and hydroxide-rich planes exposed to soil solution, and changes in pH protonate and deprotonate these functional groups, thereby creating variable or pH-dependent charge colloid surfaces that contribute to buffer capacity:



where Me(III) represents a trivalent metal structurally associated with the mineral surface, and the charges on the surface become increasingly negative as pH is raised. Highly weathered soils that contain few aluminosilicate clay minerals or organic matter are often dominated by these variably-charged colloids, and they contribute to pH buffering through positive- and negative-charge creation.

The relative importance of CaCO₃ dissolution and precipitation, Al hydrolysis, cation exchange on permanently charged aluminosilicate minerals, organic-matter functional group protonation and deprotonation, and variation in negative and positive charges on variably-charged oxide coatings depend on the relative abundance of these colloids in a given soil. In many soils, more than one of these soil chemical processes is responsible for pH buffering, and titrations of whole-soil materials reflect an integration of the several processes. Table 1 summarizes the relative magnitudes of the buffer capacities associated with these processes.

Buffering of soil oxidation-reduction status

The concept of ‘electron activity’ of soils refers to the thermodynamic tendency for electrons to be transferred from reductants (e.g., Fe(II)) to oxidants (e.g., O₂), and it is operationally defined by platinum (Pt) electrode potentials versus a standard reference electrode. When the measured potential is corrected for the reference electrode potential relative to the hydrogen electrode ($E^0 = 0.0$ V), it is designated E_h . This voltage can be converted to pE, a property analogous to pH, by dividing E_h in volts by 0.059 (the Nernstian slope factor relating electrode voltage to electron activity). The E_h or pE is the intensity variable that is a measure of electron activity or ‘electron pressure’ in a soil system. Buffering of electron activity is called poise, and is analogous to proton buffering, except that the processes responsible for resistance to change in pE are due to electron transfer reactions, many of which are microbial and enzymatically catalyzed.

Heterotrophic microbial respiration in soils uses organic C as the electron source (reductant) and an array of electron acceptors as oxidants. Strict aerobes use dissolved oxygen (O₂) as the required oxidant to derive energy from the oxidation of organic C, the most energy-efficient process for cellular respiration. In flooded soils in which all pores are filled with water, the diffusion rate for dissolved O₂ is 10,000 times slower than it is in air. As a result, microbial respiration may deplete available O₂ faster than it is replenished, thereby leading to the onset of anaerobic conditions, with a decrease in the measured pE. Under anaerobic conditions, facultative anaerobes and strict anaerobes use alternative electron acceptors, with decreasing metabolic efficiency in the order (depending on pH): NO₃⁻, Mn(III,IV)(hydr)oxides, Fe(III)(hydr)oxides, SO₄²⁻, CO₂, and H⁺.

The quantities of each of these electron acceptors (the capacity factor for electron activity) in the soil will ‘poise’ the E_h at a given level as governed by the free energy of reaction associated with its reduction, as coupled to the oxidation of organic C. The expected pE and E_h values at which each of these couples would poise the soil are shown at pH 5 and 7 in Table 2. Since the molecular

Table 1 Approximate buffer capacities of soil materials and horizons in the pH range 3.0–10.0.

Soil constituent or horizon	Buffering pH range	Buffering index (mmol kg ⁻¹ pH unit ⁻¹)	pH-buffering mechanism
Aluminosilicate clays	3–10		
Smectite	3–10	178–333	Cation exchange; mineral dissolution/precipitation; Al hydrolysis
Vermiculite	3–10	333–444	Cation exchange; mineral dissolution/precipitation; Al hydrolysis
Illite	3–10	44–88	Cation exchange; mineral dissolution/precipitation; Al hydrolysis
Kaolinite	3–10	2–11	Cation exchange; mineral dissolution/precipitation; Al hydrolysis
Soil organic matter	5–8	360–444	Protonation-deprotonation of weak acid functional groups; conformational changes
Allophane and imogolite	3–10	44–111	Protonation-deprotonation of weak acid functional groups; conformational changes
Fe(III) and Al(III) (hydr) oxides	3–10	11–89	Protonation-deprotonation of weak acid functional groups; conformational changes
CaCO ₃ and MgCO ₃	<7	4444	Dissolution/precipitation
Forest floor organic horizons	3–7	180–360	Metal-ligand complexation/decomplexation; cation exchange; protonation/deprotonation; of weak acid functional groups
Mineral horizons of agricultural and forest soils	4–10	53	Cation exchange, mineral dissolution/precipitation, Al hydrolysis, protonation-deprotonation of weak acid functional groups, conformational changes, and metal-ligand complexation/decomplexation; principally associated with soil organic matter

Table 2 Common reduction half-reactions that poise soils at designated pE values at pH 5 and 7; the log *K* value is the pE for the reaction at pH 0.

Half-reaction (for 1 electron reduction)	log <i>K</i>	Poising pE	
		pH 5	pH 7
$1/4\text{O}_2 + \text{e}^- + \text{H}^+ = 1/2\text{H}_2\text{O}$	20.8	15.6	13.6
$+ \text{e}^- + 6/5\text{H}^+ = 1/10\text{N}_2 + 3/5\text{H}_2\text{O}$	21.1	14.3	11.9
$1/2\text{Mn}_3\text{O}_4 + \text{e}^- + 4\text{H}^+ = 3/2\text{Mn}^{2+} + 2\text{H}_2\text{O}$	30.7	16.7	8.7
$\text{Fe}(\text{OH})_3 + \text{e}^- + 3\text{H}^+ = \text{Fe}^{2+} + 3\text{H}_2\text{O}$	15.8	4.8	-1.2
$+ \text{e}^- + 5/4\text{H}^+ = 1/8\text{H}_2\text{S} + 1/2\text{H}_2\text{O}$	5.2	-1.0	-3.5
$1/8\text{CO}_2 + \text{e}^- + \text{H}^+ = 1/8\text{CH}_4 + 1/4\text{H}_2\text{O}$	2.9	-2.1	-4.1
$\text{e}^- + \text{H}^+ = 1/2\text{H}_2$	0	-5	-7
$1/4\text{CO}_2 + \text{e}^- + \text{H}^+ = 1/24\text{C}_6\text{H}_{12}\text{O}_6 + 1/4\text{H}_2\text{O}$	-0.21	-5.9	-7.9

reaction mechanism for many reduction reactions involves transfer of the electron with a proton to the oxidant (equivalent to the addition of a H atom), the overall reaction raises the soil pH in many cases. As a result, the higher the pH of a soil, the lower the E_h or pE at which a given reduction reaction is expected to occur and at which the soil system will be poised (i.e., at a lower (H^+), a greater (e^-), or 'electron pressure' is needed to effect the given reduction).

Buffering of ion activities via dissolution-precipitation, ion exchange, and ligand complexation

The soil solution activities of ions other than H^+ and OH^- are intensity measurements that may be controlled by dissolution-precipitation, ion exchange, and ligand complexation reactions responsible for buffering in soils. Some processes maintain ion activities in soil solution, for example exchange, and complexation. Other biotic and abiotic processes may increase or decrease, particular ion activities. The disturbance of these equilibria will induce restorative shifts in accordance with the Le Chatelier principle, which states that the balance of products and reactants in a dynamic chemical equilibrium will shift in response to a perturbation to restore the original balance of reactants and products. For example, if sparingly soluble PbCrO_4 (lead chromate) is present in a soil, Pb^{2+} and CrO_4^{2-} concentrations of approximately $5.3 \times 10^{-7} \text{ mol L}^{-1}$ will be maintained in soil solution as controlled by the solubility product, K_{sp} , of 2.8×10^{-13} . Any process that depletes the soluble Pb^{2+} or will induce the dissolution of a small amount of solid PbCrO_4 to restore the equilibrium activities of the Pb^{2+} and CrO_4^{2-} . Similarly, additions of these ions to the soil solution will induce precipitation of more PbCrO_4 . In this way, the ion activities are buffered and maintained in soil solution, as shown in Eq. (5):



Similar reactions occur via cation and anion exchange on charged colloids to buffer ion activities, as discussed above for soil acidity buffering. The suite of exchangeable cations in soils is dominated by Ca^{2+} , Mg^{2+} , K^+ , Na^+ , H^+ , and Al^{3+} , depending on the pH of the soil, the mineralogy of the soil parent material, the aluminosilicate clay mineralogy, plant uptake, microbial processes, and human-induced changes in the soil conditions. Multiple equilibria are established in any soil, but all comprise a capacity factor, represented by the exchangeable cations, and an intensity factor. The capacity factor is sometimes called the 'quantity' term, and the quantity-to-intensity ratio (Q/I) is a measure of the nature of the cation or anion exchange buffering reactions responsible for maintaining ion activities in soil solution. Eq. (6) represents a Ca-K cation exchange reaction for such a Q/I relationship:



in which Ex-Ca represents solid-phase, exchangeable Ca, $\text{K}^+(\text{aq})$ is soluble K^+ , Ex-K₂ is exchangeable K^+ , and $\text{Ca}^{2+}(\text{aq})$ is soluble Ca^{2+} . Any environmental conditions that change the activity ratio of K^+ to Ca^{2+} in soil solution (e.g., changing water content, preferential plant uptake of one cation over the other, precipitation of Ca^{2+} , or addition of Ca- or K-containing materials to the soil) will result in K-for-Ca or Ca-for-K exchange reactions that restore and buffer the ion activities in soil solution. The balance between the ratios of exchangeable Ca-to-K and soluble Ca-to-K in solution can be described by an equilibrium constant, a measure of the buffer capacity between exchangeable and soluble forms of the cations.

The activities of many ions in soil solution are also in equilibrium with complexed or chelated forms of the metals that are also soluble, but in forms in which the positive charge of the cation has been neutralized by negatively charged ligands. Any processes that increase or decrease the activity of the hexaquo form of the 'free' form of the cation, e.g., $\text{Fe}(\text{H}_2\text{O})_6^{3+}$, that is in equilibrium with complexed forms, for example complexed with carboxylic acid groups, hydroxyl ions, or other Lewis bases, will induce a shift in the equilibrium to restore the balance of the complexed and free form of the ion. In this way, the 'free' form of the cation (its activity or intensity) is maintained through equilibrium with the capacity factor, the complexed form. Similar to solubility products and cation

exchange equilibria, a stability constant quantifies the relative thermodynamic stability of the complexed and free forms of a metal in a quantity-intensity relationship.

Summary

To conclude, the capacity of a system to resist changes can be an important property. The pH buffer capacity depends on mineralogy, and the organic and mineral contents, and can be adjusted using amendments such as fresh or pyrolyzed organic matter. Appreciation of the buffer capacity of soils for nutrients leads to reasoned use of fertilizers. The concept of ecosystem buffering, or resilience, underpins strategies to protect ecosystems in the context of climate change and increasing pressures from agriculture and urbanism.

References

- Barrow NJ (2015) Soil phosphate chemistry and the P-sparing effect of previous phosphate applications. *Plant and Soil* 397: 401–409.
- Barrow NJ (2022) How understanding soil chemistry can lead to better phosphate fertilizer practice: A 68 year journey (so far). *Plant and Soil* 476: 117–131.
- Bloom PR and Skjellberg U (2011) Soil pH and pH buffering. In: Huang PM, Li Y, and Sumner ME (eds.) *Handbook of Soil Sciences Properties and Processes*. CRC Press.
- Certini G and Scalenghe R (2023) The crucial interactions between climate and soil. *Science of the Total Environment* 856: 159169.
- Doody DG, Withers PJA, Dils RM, McDowell RW, Smith V, McElarney YR, Dunbar M, and Daly D (2016) Optimizing land use for the delivery of catchment ecosystem services. *Frontiers in Ecology and the Environment* 14: 325–332.
- Evangelou VP, Wang J, and Phillips RE (1994) New developments and perspectives on soil potassium quantity/intensity relationships. *Advances in Agronomy* 52: 173–227.
- Nair KP (2013) The buffer power concept and its relevance in African and Asian soils. *Advances in Agronomy* 121: 447–516.
- Nair KP (2022) Precise quantification of plant nutrient bioavailability: The relevance of the nutrient buffer power concept. *Advances in Agronomy* 174: 189–233.
- Trabelsi D and Mhamdi R (2013) Microbial inoculants and their impact on soil microbial communities: A review. *BioMed Research International* 2013: 863240.
- Xu S, Sheng C, and Tian C (2020) Changing soil carbon: Influencing factors, sequestration strategy and research direction. *Carbon Balance and Management* 15: 1–9.
- Zhuang S-Y, Wang M-K, King H-B, Hwang JL, and Hsu FS (2006) Rain acid buffer capacities of alpine forest soils in central Taiwan. *Geoderma* 137: 174–178.

Soil specific surface area

Emmanuel Arthur^a, Markus Tuller^b, and Kurt D Pennell^c, ^aDepartment of Agroecology, Aarhus University, Tjele, Denmark; ^bDepartment of Environmental Science, The University of Arizona, Tucson, AZ, United States; ^cSchool of Engineering, Brown University, Providence, RI, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	22
Methods for determining SSA	24
Mass-volume estimation methods	24
Sorption-based methods for SSA determination	24
Adsorption from solution	24
Gas adsorption	25
Water vapor sorption	27
EGME retention	29
Indirect methods	30
Proximal sensing methods (spectroscopy)	30
Empirical models or pedotransfer functions	30
Conclusions	30
References	30

Abstract

The specific surface area (SSA) of soil influences important processes in the subsurface, including the retention of water, the adsorption and release of plant nutrients and contaminants, particle attachment, soil swelling, and soil strength. In this chapter, we present state-of-the-art and novel techniques for direct and indirect determination of SSA together with underlying concepts and theories. Both adsorption-based methods and indirect methods such as proximal soil sensing and empirical modeling are presented. The advantages and limitations of each method are discussed with respect to the intended use of the SSA data.

Key points

- State-of-the-art and novel techniques for the determination of the specific surface area (SSA) of soil are presented.
- The factors affecting the accuracy of each technique are elucidated.
- The underlying concepts and theories for SSA estimation are discussed.

Introduction

The specific surface area (SSA) of soils, expressed as the surface area per unit mass ($\text{m}^2 \text{g}^{-1}$) or per unit volume ($\text{m}^2 \text{m}^{-3}$) of soil, plays a crucial role in numerous processes, including the retention of water, infiltration and drainage, ion exchange, adsorption and release of plant nutrients and contaminants, heat transport and storage, structural soil development, microbial processes, soil swelling, plasticity, cohesion, and soil strength. The SSA may be expressed as total, external or internal SSA. The external SSA comprises the area of the solid-air interface of soil in an oven-dry state, where adsorption of polar or non-polar molecules can occur. On the other hand, the internal SSA consists of the surfaces of interlayers of clays or the micropores of organic material that are inaccessible to gas molecules but are accessible to absorption or uptake of polar molecules (Lu and Zhang, 2020). Consequently, the total SSA of soil is the sum of the internal and external SSA. Different methods interrogate different surface areas as explained in subsequent sections.

For soils, total SSA can range from $<0.1 \text{ m}^2 \text{g}^{-1}$ to $800 \text{ m}^2 \text{g}^{-1}$ depending on the particle size distribution, clay mineral type, and soil organic matter content. Sand-sized particles have small SSA values ($<1 \text{ m}^2 \text{g}^{-1}$), while clay-sized particles have large SSA values (e.g., $800 \text{ m}^2 \text{g}^{-1}$ for pure montmorillonite). Consequently, for most natural soils SSA is primarily governed by the clay size fraction.

Apart from the overall clay content of the soil, the dominant clay mineral strongly influences the magnitude of SSA. For example, kaolinite-rich samples usually have much smaller total SSA than soils containing high surface area clays like montmorillonite, which exhibit very large SSA values when probed with polar molecules such as ethylene glycol monoethyl ether (EGME) which can interrogate internal surface area (Table 1). Soil organic matter reportedly has a large SSA ($560\text{--}800 \text{ m}^2 \text{g}^{-1}$), and for soils that have similar fractions of clay, silt and sand, and identical clay mineralogy, the SSA increases as organic matter content increases when measured using a polar molecule such as water or EGME (Table 1, Fig. 1).

Table 1 Comparison of specific surface area values obtained via N_2 gas adsorption (i.e., external SSA) and ethylene glycol monoethyl ether (EGME) retention (i.e., total SSA) for different soil and clay mineral types.

		Specific surface area (m ² g ⁻¹)	
Sample type	Extracted from	N ₂ -BET method	EGME method
Natural Soils			
Sand texture(AZ1, 0.10% OC)	Arthur et al. (2013)	2	7
Loam texture (AZ11, 0.97% OC)	Arthur et al. (2013)	15	61
Fine Loam (Marlette A, 2.14% OC)	Pennell et al. (1995)	10	48
Fine Loam (Marlette B, 0.8% OC)	Pennell et al. (1995)	6	37
Fine Loam (Webster, 3.3% OC)	Pennell et al. (1995)	8	168
Silt texture (AZ16, 0.76% OC)	Arthur et al. (2013)	29	111
Clay texture (AZ18, 1.15% OC)	Arthur et al. (2013)	36	158
Muck (Houghton, 44% OC)	Pennell et al. (1995)	0.8	163
Pure clay minerals			
Kaolinite (KGa-1)	van Olphen and Fripiat (1979)	10	16
Illite (Inter-ILI)	Macht et al. (2011)	41	112
Montmorillonite (SWy-1)	van Olphen and Fripiat (1979)	32	662
Montmorillonite (SAz-1)	van Olphen and Fripiat (1979)	97	820
Wyoming bentonite	Call (1957)	65	372

BET: Brunauer-Emmett-Teller.

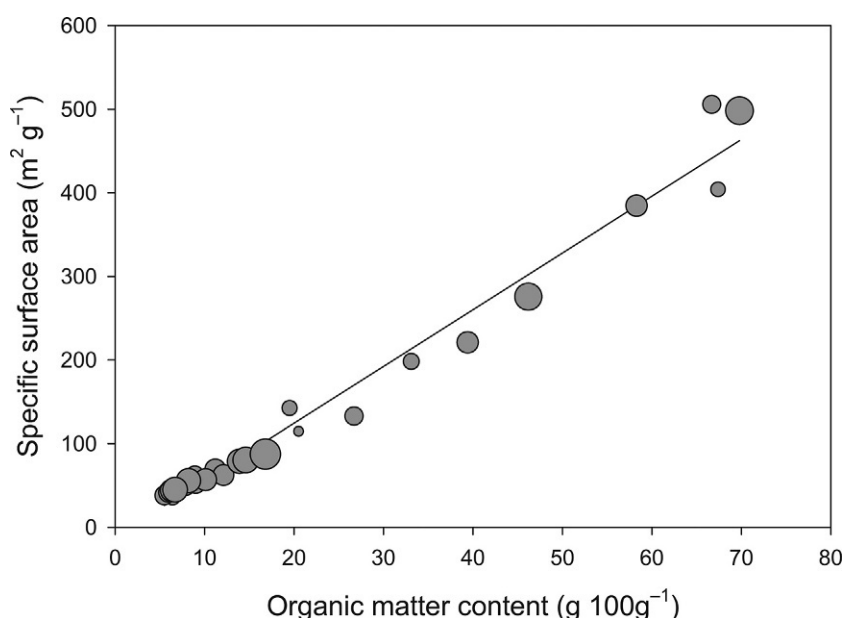


Fig. 1 Water-sorption derived soil specific surface area for samples with different organic matter contents and clay contents ranging from 4% to 13%. Symbol size is indicative of the clay content.

The SSA of porous media can be estimated using various methods. This may involve direct physical methods such as measurement of the particle size and shape with scanning electron microscopy, X-ray micro-computed tomography to estimate SSA from 3D images of particles, and mercury intrusion porosimetry to infer SSA from the effective pore size distribution (e.g., Borkovec et al., 1993; Maroof et al., 2020). Commonly used methods include adsorption or desorption methods where the deposition or retention of a liquid or vapor from the mineral or organic surface is measured. Liquid phase adsorption involves the use of ethylene glycol, EGME and methylene blue. Gas phase adsorption includes the use of water vapor (H_2O), nitrogen (N_2), or carbon dioxide (CO_2). For small SSA media (e.g., $<1 m^2 g^{-1}$), krypton (Kr) or argon (Ar) gas may be used due to their lower vapor pressure at 77 K. Each method presents unique challenges and, in some cases, measures different surface areas (i.e., total or external). Therefore, it is important to recognize the dependence of SSA data on the measurement method and to select one or more methods that are appropriate for the system of interest.

Methods for determining SSA

Mass-volume estimation methods

Direct physical measurement of SSA typically involves the use of light or [electron microscopy](#) to determine the shape and dimensions of individual soil particles. These observations are supplemented by X-ray diffraction measurements to assess the crystallographic structure and interlayer spacing of clay minerals. Provided that a characteristic particle shape and size can be determined, the SSA can be obtained from mass-volume relationships. For example, the SSA of a spherical particle may be calculated with Eq. (1):

$$SSA = \frac{4\pi r^2}{\rho_s V_s} = \frac{4\pi r^2}{\rho_s \frac{4}{3}\pi r^3} = \frac{3}{\rho_s r} \quad (1)$$

where r is the radius of the solid particle, ρ_s is the density of the solid, and V_s is the volume of the solid. Using this approach, the specific surface area of quartz sand with a particle diameter of 1.0×10^{-3} m and a particle density of 2.65×10^3 kg m⁻³ is 2.26 m² kg⁻¹. In practice, SSA values of sands determined by direct observation are often several orders of magnitude smaller than measured values due to the presence of non-spherical particles, [surface roughness](#), and the presence of fine particles ([Table 1](#)).

An analogous mass-volume approach can be used to estimate the SSA of clay minerals, provided that the structural formula and unit cell dimensions are known. For example, [montmorillonite](#) with a nominal structural formula of $K_{0.66}Si_{8.0}(Al_{3.34}Mg_{0.66})O_{20}(OH)_4$ and unit cell dimensions of $a = 5 \times 10^{-10}$ m, $b = 9 \times 10^{-10}$ m, and $c = 9.5 \times 10^{-10}$ m. Assuming that the particle density is approximately 2.8×10^3 kg m⁻³, and that the edge area (i.e., c -dimension) is negligible compared with the area of the basal surfaces (i.e., a - and b -dimensions), the SSA of the montmorillonite can be estimated as:

$$SSA = \frac{2ab}{\rho_s V_s} = \frac{2(5.0 \bullet 10^{-10} \text{ m})(9.0 \bullet 10^{-10} \text{ m})}{(2.8 \bullet 10^3 \text{ kg m}^{-3})(5.0 \bullet 10^{-10} \text{ m})(9.0 \bullet 10^{-10} \text{ m})(9.5 \bullet 10^{-10} \text{ m})} = 7.52 \bullet 10^5 \text{ m}^2 \text{ kg}^{-1} \quad (2)$$

Using a slight variation of Eq. (2), the SSA of the montmorillonite can also be estimated from the molecular weight of the unit cell and Avogadro's number (N_A):

$$SSA = \frac{(2ab)(N_A)}{MW_{\text{mont}}} = \frac{2(5.0 \bullet 10^{-10} \text{ m})(9.0 \bullet 10^{-10} \text{ m})(6.022 \bullet 10^{23} \text{ mol}^{-1})}{7.447 \bullet 10^{-1} \text{ kg mol}^{-1}} = 7.52 \bullet 10^5 \text{ m}^2 \text{ kg}^{-1} \quad (3)$$

The relationships presented in Eqs. (2) and (3) apply to clay minerals that exist as flat, plate-like structures with a thickness corresponding to that of the unit cell. When water is removed from expandable clay minerals, the interlayers collapse. The resulting clay particle will then consist of several unit cells stacked on top of one another. If the number of unit cells contained in a collapsed montmorillonite particle is 10, the resulting surface area would be approximately 7.5×10^4 m² kg⁻¹.

The mass-volume methods described above are generally limited in applicability to clean sands or pure clay mineral samples. Many soil constituents, including [metal oxides](#) and organic matter, exist as irregular or poorly defined amorphous structures, which are challenging to characterize. Furthermore, the SSA of natural soils cannot be treated as a strictly additive property due to surface coatings and mineral-organic matter associations. Except in a few limited cases (e.g., clean sands), the SSA of a whole soil should not be estimated using a summation procedure based on the surface area contributions of individual soil constituents.

Sorption-based methods for SSA determination

Adsorption from solution

The SSA can be derived from the adsorption of dissolved molecules from solution as long as the resulting adsorption isotherm has a limiting or maximum value. Such adsorption isotherms are classified as Type I and can be described by the Langmuir adsorption model ([Fig. 2](#)). The Langmuir model can be derived by assuming that, once equilibrium is reached, the rate of solute adsorption onto open surface sites is equal to the rate of solute desorption from occupied surface sites. Several assumptions are inherent to the Langmuir model, including: (1) adsorption is localized or site-specific; (2) no interactions occur between adsorbed molecules; (3) the energy of adsorption is constant for all adsorption sites; (4) the adsorption capacity of the solid is limited; and (5) the maximum adsorption capacity corresponds to monolayer coverage. For solid-liquid systems the Langmuir equation may be written as:

$$C_s = \frac{C_{s,\text{max}} \bullet \beta \bullet C_a}{1 + \beta \bullet C_a} \quad (4)$$

where C_s is the solid-phase concentration of solute at equilibrium, $C_{s,\text{max}}$ is the maximum solid-phase concentration, β is the ratio of the adsorption and desorption rates, and C_a is the aqueous-phase concentration. A series of Langmuir isotherms is presented in [Fig. 2](#) to illustrate the effect of increasing the value of β from 0.01 to 0.25 L mg⁻¹ with $C_{s,\text{max}}$ fixed at 1.0 g kg⁻¹. As the value of β increases, the rate at which the adsorption isotherm approaches the maximum sorption capacity of the solid ($C_{s,\text{max}}$) increases. However, the shape or steepness of the isotherm has no bearing on the maximum sorption capacity, which is used to calculate SSA.

In practice, one can obtain the β and C_a parameters by fitting Eq. (4) to measured adsorption data using non-linear regression analysis software. Some organic molecules that are used in combination with the Langmuir model to determine SSA include methylene blue or cationic surfactants such as cetylpyridinium bromide (CBP) ([Greenland and Quirk, 1964](#)). On most mineral

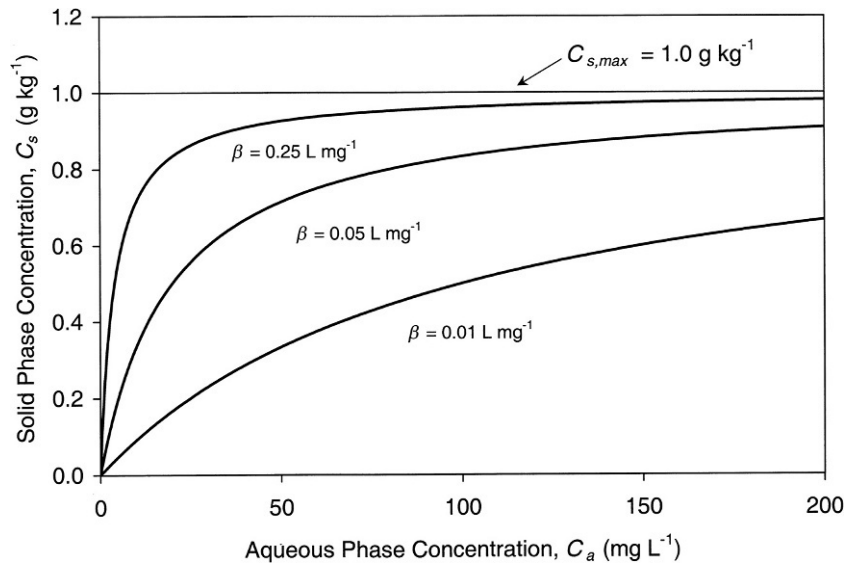


Fig. 2 Effect of changes in β on the Langmuir adsorption isotherms.

surfaces, adsorbed CPB molecules form a double layer, yielding an effective molecular area of $2.7 \times 10^{-17} \text{ m}^2$. However, for expandable clay minerals, the adsorbed CBP bilayer on the interlayer surfaces is shared and yields an effective molecular area of $5.4 \times 10^{-17} \text{ m}^2$. Thus, the external surface area has to be determined separately from the nitrogen sorption isotherm at liquid nitrogen temperatures using the Brunauer-Emmet-Teller equation, the N_2 -BET method, to be able to quantify the contribution of internal surfaces to the overall adsorption of CBP. Furthermore, in samples rich in iron or aluminum oxides, the use of CBP can be challenging, as it may not form complete double layers on such surfaces. Consequently, the sesquioxides present have to be removed before analyses of SSA.

Gas adsorption

The adsorption of gas to determine SSA assumes that one gas molecule covers a defined surface area. When the whole sample surface is covered by one layer of the gas (termed the monolayer capacity) the SSA of the sample can be derived from the monolayer capacity. A gas adsorption isotherm can be obtained by measuring the quantity of gas adsorbed on the sample surface over a range of relative pressures at a constant temperature. Similarly, a desorption isotherm is obtained by measuring the gas remaining on the soil surface when the relative pressure is gradually reduced.

The adsorption isotherms obtained can be combined with an isotherm or physically-based model to quantify SSA. The adsorption process may conform to different isotherm types, depending on the clay mineral type present in the soil. For most cases, solid surface sorption typically exhibits a Type II isotherm (Fig. 3), where molecules are adsorbed in multiple layers.

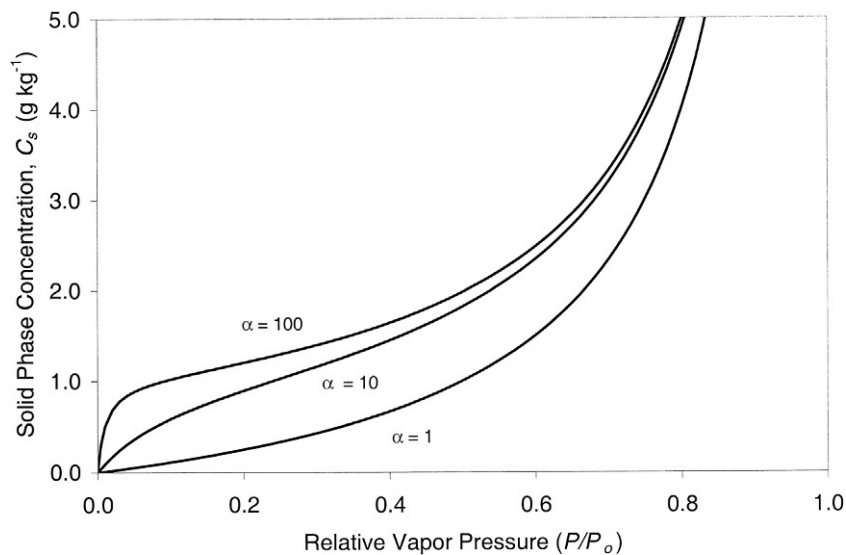


Fig. 3 Effect of changes in α on the Brunauer-Emmet-Teller (BET) adsorption isotherm when $X_m = 1 \text{ g kg}^{-1}$.

The commonly used isotherm model used for SSA measurement in combination with gas adsorption is the Brunauer-Emmet-Teller (BET) equation. This equation is based on the Langmuir model, modified to account for multilayer formation assumes that physical sorption involves the formation of many molecular layers of adsorbate on the adsorbent surface. The BET equation assumes that (a) the heat of adsorption of the first molecular layer is constant, (b) the heat of adsorption for the second and subsequent layers is constant and equivalent to the heat of condensation, (c) adsorption and desorption occur only from exposed layers, and (d) Langmuir theory assumptions apply to each layer. The equation is parameterized using the measured adsorption data to obtain the monolayer adsorbed gas mass:

$$C_s = \frac{X_m \cdot \alpha \cdot P/P_0}{[(1 - P/P_0)(1 - P/P_0 + \alpha \cdot P/P_0)]} \quad (5)$$

where C_s is the mass quantity of sorbate adsorbed (kg kg^{-1}) at a vapor pressure, X_m is the quantity adsorbed (kg kg^{-1}) when the sorbent is covered with a molecular monolayer and α is a constant related to the heat of adsorption, P/P_0 = relative vapor pressure, where P is the vapor pressure and P_0 is the saturated vapor pressure. The dimensionless α parameter can be mathematically defined as:

$$\alpha = \exp \left[\frac{E_1 - E_L}{RT} \right] \quad (6)$$

where E_1 is the heat of adsorption of the first molecular layer of the sorbate; E_L is the heat of condensation of the sorbate; R is the molar gas constant ($8.314 \text{ J mol}^{-1} \text{ K}^{-1}$), and T is the absolute temperature (K). The BET equation is identical to the Langmuir equation when the number of adsorbed layers equals one.

A series of BET isotherms is shown in Fig. 3 for α equal to 1, 10, and 100, with X_m fixed at 1 g kg^{-1} and n equal to infinity. As the value of α increases, the inflection point or 'knee' of the BET adsorption isotherm becomes more evident. The inflection in the BET adsorption isotherm corresponds approximately to the point of monolayer coverage. Below the inflection point, gas adsorption occurs on exposed surfaces. The gradual increase in adsorption above the inflection point corresponds to multilayer formation on the surface, while the steep asymptotic rise in adsorption at relative vapor pressures approaching unity ($P/P_0 = 1$) corresponds to liquid condensation.

Commonly used adsorbates for gas sorption in combination with the BET equation include nitrogen (N_2), CO_2 , krypton and argon (for $\text{SSA} < 0.1 \text{ m}^2 \text{ g}^{-1}$), and water vapor. Regardless of the gas or vapor used, the area covered by one molecule on the surface must be known or estimated. For N_2 this is estimated to be $1.62 \times 10^{-19} \text{ m}^2$, and for water it is $10.8 \times 10^{-20} \text{ m}^2$.

The most commonly used gas for determining SSA is N_2 and measurements are conducted at the boiling temperature of liquid nitrogen at atmospheric pressure (-196°C). To obtain the N_2 -adsorption isotherm, the amount of gas adsorbed is computed for each incremental gas dosage based on the change in vapor pressure at equilibrium. Upon reaching the saturated vapor pressure ($P/P_0 = 1$), the process may be reversed (the vapor pressure is incrementally reduced) to obtain a desorption isotherm, from which pore size analysis can be performed. In general, automated surface area instruments report gas adsorption data as the volume of gas adsorbed per gram of soil at standard temperature and pressure (STP). Volumetric gas adsorption data are converted to a mass basis (grams per kilogram) using the molar volume of an ideal gas at STP ($22.414 \text{ L mol}^{-1}$) and the molecular weight of the gas (e.g., $\text{N}_2 = 28.02 \text{ g mol}^{-1}$).

The second experimental approach used to measure gas or vapor adsorption is based on the continuous introduction of a gas stream at constant vapor pressure. Once equilibrium is attained, the soil sample is either weighed or extracted to determine the

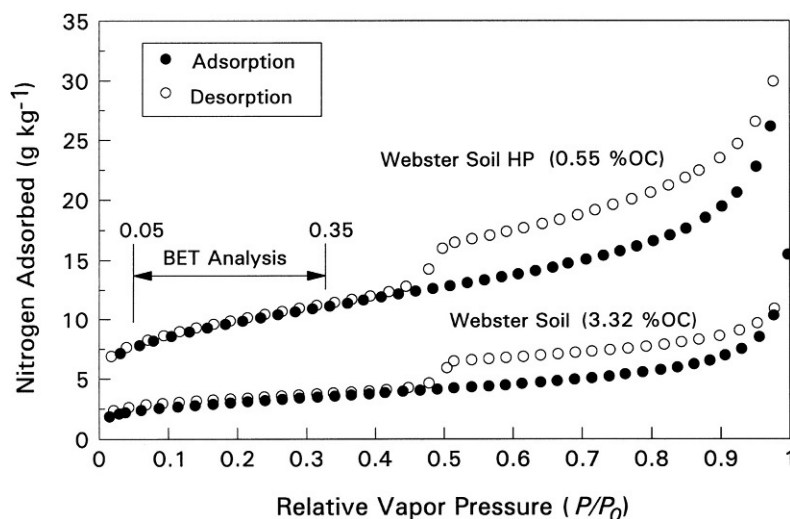


Fig. 4 Adsorption and desorption of N_2 on hydrogen peroxide (HP)-treated and untreated Webster soil.

amount of adsorbed gas. The vapor pressure of the gas flow stream is then incrementally increased over the desired vapor pressure range to obtain the adsorption isotherm. Typical nitrogen adsorption and desorption isotherms for two soils are shown in Fig. 4.

To obtain values for the two unknown parameters in the BET equation, C and X_m , Eq. (5) can be fitted directly to the experimental adsorption data using nonlinear regression.

The N_2 -BET SSA is then calculated as:

$$SSA = \frac{X_m \cdot N_A \cdot A_m}{M_w} \quad (7)$$

where N_A is Avogadro's number ($6.02 \times 10^{23} \text{ mol}^{-1}$), A_m is the area covered by one sorbate molecule ($1.62 \times 10^{-19} \text{ m}^2$ for N_2) and M_w is the molecular weight of the sorbate (28.02 g mol^{-1} for N_2). For example, if X_m was determined as 5 g g^{-1} , the N_2 -BET SSA is calculated as:

$$SSA = \frac{0.005 \text{ g g}^{-1} \cdot (6.022 \cdot 10^{23} \text{ mol}^{-1}) \cdot (1.62 \cdot 10^{-19} \text{ m}^2)}{28.02 \text{ g mol}^{-1}} = 17.4 \text{ m}^2 \text{ g}^{-1} \quad (8)$$

Due to the inability of N_2 to penetrate the interlayer spaces of expandable clay minerals, the SSA determined from N_2 -BET is often considered the external surface area.

Treatment and preparation of soil samples before N_2 adsorption can strongly influence the measured SSA (Figs. 4 and 5). For example, the removal of soil organic matter or carbonates often leads to increased SSA values. This has been attributed to the exposure of mineral surfaces covered by organic matter and the separation of mineral particles held together by organic matter bridging. Data from 10 Arizona soil samples confirm that the N_2 -BET SSA of the samples consistently decreased after the removal of organic matter and carbonates (Fig. 5).

As previously mentioned, the surface of the sample must be cleaned to remove other adsorbed gases (e.g., water vapor) before commencing with the measurement of adsorption of the test gas. The cleaning process often involves a combination of outgassing and high temperatures. The drying or dehydration process can lead to unwanted phase changes in iron hydroxides or collapse of the interlayer space of expandable clay minerals, which are then not accessible to inert gases such as N_2 . In addition, [electron microscopy](#) suggests that air-drying of soil samples results in the collapse and shrinkage of soil [humic acid](#), whereas freeze-drying maintains a complex structural network characteristic of soil organic matter. Thus, freeze-drying of soils to remove the water before N_2 -BET analysis tends to result in larger surface area values, which may be more representative of natural conditions. Alternatively, a low outgassing temperature may be considered to reduce the negative effect on gas adsorption.

Water vapor sorption

Water is a polar covalent molecule and it is attracted to the exchangeable cations in clay minerals present in soils. Consequently, water molecules can penetrate the interlayer space of expandable clay minerals, and measurement of the amount of water sorbed at a given relative humidity (RH) provides a measure of the total SSA. Water vapor can be used to determine SSA with two approaches: (a) a single-point measurement that assumes monolayer coverage (X_m) at a specific RH , for example, 0.20 for soil samples with non-swelling clays and 0.47 for samples containing swelling clays ([Newman, 1983](#))—the SSA can then be determined from Eq. (7), and (b) a combination of water vapor sorption isotherms and associated models.

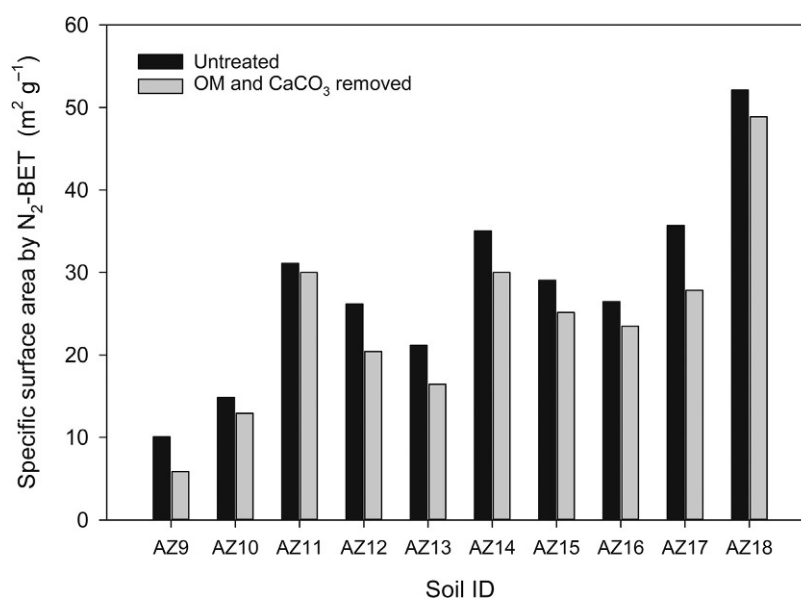


Fig. 5 Comparison of specific surface areas determined with N_2 -BET for ten samples before and after removal of organic matter and calcium carbonates. Partly from Arthur et al. (2013).

Although relatively simple, the single-point measurement approach is strongly affected by the clay minerals present in the sample. Since soil samples are often a mixture of different clay minerals, assuming monolayer coverage using a fixed RH to estimate SSA can lead to errors. The increased availability of instrumentation such as the Aqualab Vapor Sorption Analyzer (METER Group, Inc., Pullman, WA, United States) or Dynamic Vapor Sorption (Surface Measurement Systems, Rosemond Rd., United Kingdom) for rapid measure of water vapor sorption isotherms for clays and soil samples has made the use of the whole isotherm more convenient than in the past. These instruments measure both water vapor adsorption and desorption isotherms within a relatively short time (1–3 days depending on sample composition).

The water vapor sorption isotherms may be parameterized using the BET equation for the RH range between 0.1 and 0.4 since the BET equation does not properly characterize the effect of swelling clay. Consequently, more versatile models have been proposed that can fit the isotherm in the typical RH range from 0.05 to 0.95 (Arthur et al., 2018). Two of these models are the Guggenheim-Anderson-de Boer (GAB) model and the Tuller-Or (TO) model. Examples of water vapor sorption isotherms for three samples and their corresponding TO and GAB model fits to the data are shown in Fig. 6.

The GAB model is similar to the BET equation but has an additional parameter (K) to compensate for the assumption that the sorption state of the water molecules in the layers beyond the first is the same but different from that of the pure liquid state:

$$X = \frac{X_m \cdot C \cdot K \cdot RH}{[(1 - K \cdot RH)(1 - K \cdot RH + C \cdot K \cdot RH)]} \quad (9)$$

where X is the mass quantity of water vapor adsorbed (kg kg^{-1}). When $K = 1$, the GAB model reduces to the BET equation and the X_m and C values will be the same for the two equations. The value of X_m obtained by fitting either the GAB (Eq. (9)) or BET (Eq. (5)) equations to the water sorption isotherm data is then used to determine the SSA using Eq. (9).

The TO model (Tuller and Or, 2005) is a physically based water film sorption model that relates X (kg kg^{-1}) to the matric potential, ψ ($\text{m H}_2\text{O}$) and SSA ($\text{m}^2 \text{kg}^{-1}$) as:

$$X = \sqrt[3]{\frac{A_{svl}}{6 \cdot \pi \cdot \rho_w \cdot g \cdot \psi}} \cdot SA \quad (10)$$

where A_{svl} (J) is the Hamaker constant for solid-vapor interactions through the intervening liquid (-6×10^{-20} J), ρ_w is the density of water (1000 kg m^{-3}) and g is the acceleration due to gravity (9.8 m s^{-2}). Similar to the BET equation, the TO model is parameterized with the water vapor isotherm data using nonlinear regression. The fitting parameter is the SSA which is optimized during the parameterization procedure.

Based on measured data, the water vapor sorption based SSA determined by the GAB model is often slightly higher but correlates well with that obtained by the TO model (Fig. 7).

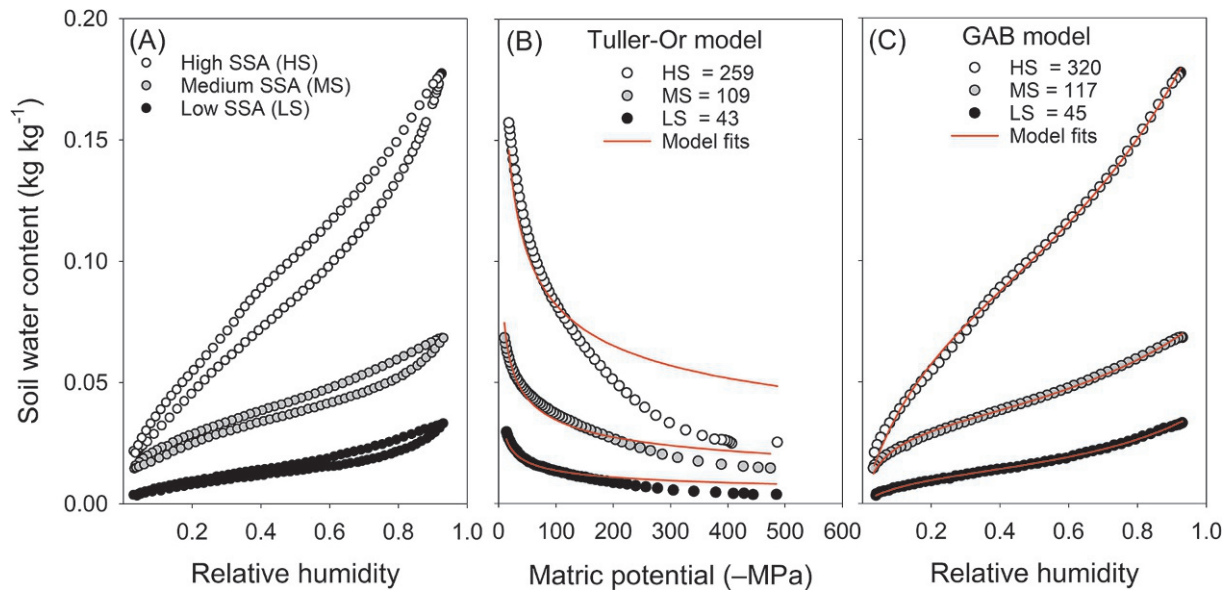


Fig. 6 (A) Example of measured water vapor sorption isotherms for three samples with different soil specific surface areas (SSAs), (B) fit of Tuller and Or (2005) model to the adsorption isotherms, and (C) fit of the Guggenheim-Anderson-de Boer (GAB) model to the desorption isotherms. The numbers in the legend of panels B and C are the SSA estimates in $\text{m}^2 \text{g}^{-1}$ from the two models (Eqs. (9) and (10)). Adapted from Knadel M, de Jonge LW, Tuller M, Rehman HU, Jensen PW, Moldrup P, Greve MH, Arthur E (2020) Combining visible near-infrared spectroscopy and water vapor sorption for soil specific surface area estimation. *Vadose Zone Journal* 19(1): e20007.

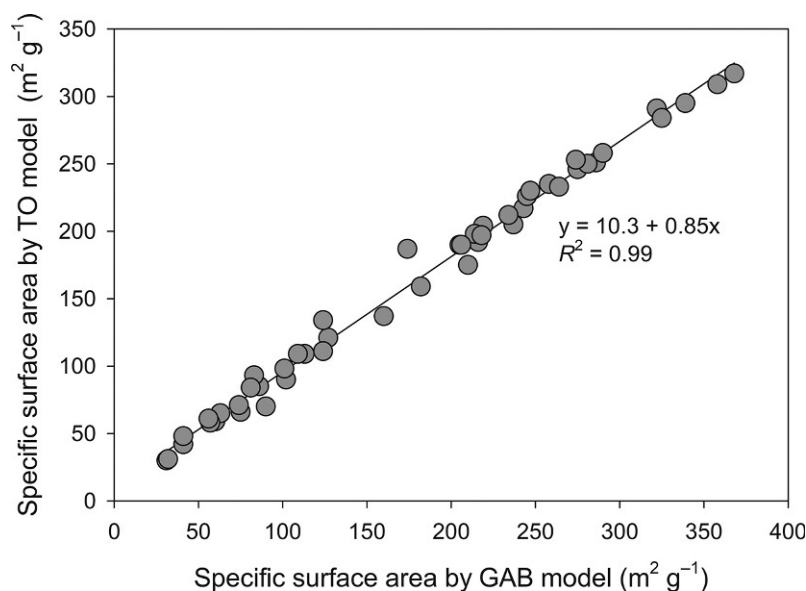


Fig. 7 Comparison of specific surface area ($\text{m}^2 \text{g}^{-1}$) for 48 soil samples determined from the water vapor desorption isotherm combined with the Tuller-Or and GAB models.

Furthermore, the soil water vapor sorption isotherm is hysteretic and thus the SSA derived from water sorption isotherms will depend on whether the adsorption or desorption isotherm is used. It is preferable to use the desorption isotherm to determine SSA because the adsorption isotherm is sensitive to the initial water content, and presence of hydrophobic compounds, and is more affected by intermolecular forces than the desorption isotherm.

Before measurement of the water sorption isotherms, samples can be oven-dried (to avoid the effect of initial water content) or have their organic matter removed to quantify the mineral surface area. These pretreatments affect the final SSA depending on the soil or clay mineral type. For soils that are coarse-textured (sandy soils) and have significant amounts of organic matter content, oven drying can lead to strong hydrophobic effects and a resultant significant decrease in SA. However, oven drying does not seem to affect fine-textured or clay samples.

EGME retention

Similar to water, polar liquids such as EGME or ethylene glycol (EG) can be used to determine SSA. These liquids can also penetrate the interlayer lattice of clay minerals and give an estimate of the total surface area. The method is based on the retention of EG or EGME that forms a monolayer on the surface of the sample, and if the surface area that is occupied by one molecule is known, the SSA can be calculated using Eq. (11). To obtain the monolayer, excess liquid is added to the dry sample, which is allowed to evaporate under vacuum until only one monolayer remains. It is assumed that the monolayer is obtained when the weight of the sample becomes stable after multiple weighing.

Previously, EG was used to determine SA, but the rather long equilibration time between EG and samples has led to a preference for the more volatile EGME.

In practice, an initially air-dried and preferably sieved sample is mixed with laboratory-grade EGME until a slurry is formed. The EGME-sample slurry is subsequently sealed in a desiccator that is connected to a vacuum pump. A solvate of EGME mixed with calcium chloride is placed in the headspace of the desiccator, to maintain a constant EGME vapor pressure and minimize the loss of EGME from the monolayer that forms on the sample surfaces. Vacuum is applied to accelerate the evaporation of excess EGME from the samples. Thereafter, the samples are weighed repeatedly at a resolution of 0.0001 g until a constant mass is obtained. The SSA of the sample is then calculated as follows:

$$\text{SSA} = \frac{W_{\text{EGME}}}{W_{\text{OD}} \bullet CF_{\text{EGME}}} \quad (11)$$

where W_{EGME} is the mass of the EGME retained by the soil at the applied vacuum, W_{OD} is the mass of the oven-dried sample and CF_{EGME} is the mass-surface area conversion factor for EGME ($2.86 \times 10^{-4} \text{ g m}^{-2}$). It is recommended that the SSA of a reference expandable clay mineral, such as Wyoming montmorillonite (SWy-1), be measured to confirm the value of the conversion factor and to ensure that the experimental procedure is operating properly.

Previously, it was considered necessary to remove organic matter from test samples before the determination of EGME SSA to avoid partitioning of EGME into organic matter. Other studies have shown that this may not be necessary as the samples that have the organic matter removed often show similar EGME SSA to those with no pretreatment (Heister, 2014). The exchangeable cations

present in a sample may affect the retention of EGME, through cation solvation that leads to increased SSA. Further, EGME tends to cluster around cation exchange sites before a complete monolayer is formed, and this may lead to erroneous SSA values. Despite these challenges, the EGME method provides a viable alternative for determining the total SSA of most soil and clay samples.

Indirect methods

As described above, the determination of SSA can be a tedious process. Consequently, the SSA may be estimated using indirect methods that involve proximal sensing or the use of empirical models that correlate SSA to easy-to-measure soil properties.

Proximal sensing methods (spectroscopy)

Visible near-infrared (VNIR) spectroscopy (400–2500 nm wavelengths) combined with multivariate data analysis is an indirect method to estimate the SSA of soils. This is possible because light absorption in this region is characteristic of the clay type, clay content and organic matter content of the soil. Since SSA is dependent on these soil properties, the spectra obtained for a given soil can be correlated with its measured SSA. The measurement of a VNIR spectrum is rapid (<2 min per sample). In practice, the approach involves the development of partial least squares (PLS) regression models from VNIR spectra and a reference SSA measured by EGME or water sorption for a group of soils (e.g., Knadel et al., 2020). The developed calibration models are validated with an independent dataset, and the models are afterwards used to estimate SSA for new samples for which VNIR spectra are available. This method is usually applied to air-dry samples since the presence of water interferes with light absorption.

Empirical models or pedotransfer functions

The use of empirical models to estimate SSA is appealing as it is based either on easily measurable soil properties like hygroscopic water content, or readily available soil properties (e.g., clay content, silt content, organic matter content, Atterberg limits). Based on a dataset of 208 soil samples, it is possible to estimate the total SSA measured by EGME from the following empirical model based on clay and organic carbon contents. That had the following variable ranges:

$$\begin{aligned} \text{SSA} &= -1.95 + 2.49 \bullet \text{clay} + 3.16 \bullet \text{OC} \\ N &= 208; R^2 = 0.78; \text{SEE} = 13.8 \end{aligned} \quad (12)$$

where N is the number of samples and SEE is the standard error of the SSA estimate. The data boundaries of the input and predicted variables were clay content [1–56%], organic carbon content [0–8.4%], and SSA [2–161 m² g⁻¹].

Conclusions

Despite the refinement of the methodology for measuring SSA over the years, there can be large differences in the measured SSA depending on the sample characteristics, the sorbate used, and the sample pretreatment. The SSA determined for soil samples, which contain 1:1 layer (i.e., non-expandable) clay minerals, such as kaolinite, and have small contents of organic matter, is often similar regardless of whether the N₂-BET, EGME or water vapor sorption method is used. The EGME and water vapor sorption methods provide reliable estimates of the “total SSA” for most soil and clay samples, but for samples that have large amounts of organic matter, the methods may overestimate the total surface area. This is due to the potential for partitioning or dissolution of the EGME into soil organic matter and the clustering of EGME or water molecules around cation exchange sites. The pretreatment of samples such as drying or removal of organic matter may induce changes in the soil surface structure that will change the available surface area. An optimal approach will be to determine both the “total” and “external” SSA using the EGME or water vapor and the N₂-BET methods, respectively. Based on these two measures, it is possible to quantify the internal surface area and identify sample characteristics or pretreatment effects.

References

- Arthur E, Tuller M, Moldrup P, Resurreccion AC, Meding MS, Kawamoto K, Komatsu T, and de Jonge LW (2013) Soil specific surface area and non-singularity of soil-water retention at low saturations. *Soil Science Society of America Journal* 77(1): 43–53.
- Arthur E, Tuller M, Moldrup P, Greve MH, Knadel M, and de Jonge LW (2018) Applicability of the Guggenheim-Anderson-Boer water vapour sorption model for estimation of soil specific surface area. *European Journal of Soil Science* 69(2): 245–255.
- Borkovec M, Wu Q, Degovics G, Lagner P, and Sticher H (1993) Surface area and size distributions of soil particles. *Colloid Surface A* 73: 65–76.
- Call F (1957) The mechanism of sorption of ethylene dibromide on moist soils. *Journal of Science and Food Agriculture* 8(11): 630–639.
- Greenland DJ and Quirk JP (1964) Determination of the total specific surface areas of soil by adsorption of cetyl pyridinium bromide. *Journal of Soil Science* 15(2): 178–191.
- Heister K (2014) The measurement of the specific surface area of soils by gas and polar liquid adsorption methods-Limitations and potentials. *Geoderma* 216: 75–87.
- Knadel M, de Jonge LW, Tuller M, Rehman HU, Jensen PW, Moldrup P, Greve MH, and Arthur E (2020) Combining visible near-infrared spectroscopy and water vapor sorption for soil specific surface area estimation. *Vadose Zone Journal* 19(1): e20007.

- Lu N and Zhang C (2020) Separating external and internal surface areas of soil particles. *Journal of Geotechnical and Geoenvironmental Engineering* 146(2), 04019126.
- Macht F, Eusterhues K, Pronk GJ, and Totsche KU (2011) Specific surface area of clay minerals: Comparison between atomic force microscopy measurements and bulk-gas (N₂) and -liquid (EGME) adsorption methods. *Applied Clay Science* 53(1): 20–26.
- Maroof MA, Mahboubi A, and Noorzad A (2020) A new method to determine specific surface area and shape coefficient of a cohesionless granular medium. *Advanced Powder Technology* 31(7): 3038–3049.
- Newman ACD (1983) The specific surface of soils determined by water sorption. *Journal of Soil Science* 34(1): 23–32.
- Pennell KD, Boyd SA, and Abriola LM (1995) Surface area of soil organic matter reexamined. *Soil Science Society of America Journal* 59(4): 1012–1018.
- Tuller M and Or D (2005) Water films and scaling of soil characteristic curves at low water contents. *Water Resources Research* 41(9): W09403.
- van Olphen H and Fripiat JJ (1979) *Data Handbook for Clay Minerals and Other Non-metallic Minerals*. New York: Pergamon Press.

Cation exchange capacity and reactions

Donald S Ross^a and Ulf Skjellberg^b, ^aDepartment of Plant & Soil Science, Jeffords Hall, University of Vermont, Burlington, VT, United States;
^bDepartment of Forest Ecology and Management, Swedish University of Agricultural Sciences, Umeå, Sweden

© 2023 Elsevier Ltd. All rights reserved.

Introduction	32
Sources of charge in soils	33
Permanent negatively charged mineral surfaces	33
Variably charged mineral surfaces	34
Charges of soil organic matter functional groups	34
Cation exchange concept and models	34
Cation exchange equations	35
Extension of the cation exchange concept to variably charged soils	37
Exchangeable and non-exchangeable aluminum and hydrogen ions	37
Base saturation and CEC definitions	38
Cation exchange in acidic forest soils	39
Cation exchange in highly weathered tropical soils	41
Summary and outlook	41
References	42

Abstract

In soils with a neutral reaction, the metal cations calcium, magnesium, potassium and sodium are electrostatically attracted to soil particles, described as an electric double layer (EDL) of positively charged cations and negatively charged surfaces. In soils dominated by permanently charged clay minerals, cations in the EDL are rapidly and reversibly exchanged with cations in the bulk solution, as described by selectivity constants. In soils low in clay minerals, the major source of negative charge is soil organic matter and variably charged metal oxides. In acidic organic and forest soils, cation exchange between aluminum and hydrogen ions is an important pH buffering process.

Key points

- Cation exchange is an important chemical reaction in soils that governs the plant availability of nutrients such as calcium, magnesium and potassium.
- The source of negative surface charge in soils is permanently charged clay minerals, variably charged metal oxides and soil organic matter functional groups.
- Different soils may be dominated by different types of cation exchange. Sodium–calcium exchange may be important in high-pH clay soils and aluminum–hydrogen ion exchange in low-pH forest soils.
- Cation exchange can be described quantitatively with reversible, thermodynamic equations. Constants for these equations are conditional because of difficulties in estimating the activity of cations adsorbed by surfaces and occupying the electrical double layer.
- The term ‘base cation’ can easily be misleading for calcium, magnesium, potassium and sodium. In a relatively narrow pH range, the saturation of soil negative charges by these cations (‘base saturation’) is often positively related to pH.
- The pH-dependent chemistry of aluminum hydrolysis and bonding to soil organic matter complicates cation exchange modeling in low-pH soils. This could be overcome by introducing operationally defined exchangeable and non-exchangeable fractions of Al (aluminum) cations.
- In moderately acidic soils, aluminum hydrolysis buffers change in soil pH. In highly acidic soils, exchange between aluminum and hydrogen cations is an important pH buffering process.

Introduction

The study of cation exchange is often credited as the beginning of the discipline of soil chemistry. Early work by H.S. Thompson, followed closely by J.T. Way, was published in 1850 in the Journal of the Royal Agricultural Society of England (Thompson, 1850; Way, 1850). They observed that a cation such as NH_4^+ (ammonium) will displace Ca^{2+} (calcium) from soil and do so rapidly, reversibly and quantitatively. These results still hold essentially true for many soils and combinations of common exchangeable cations, e.g. Na^+ (sodium), K^+ (potassium), NH_4^+ , Mg^{2+} (magnesium), and Ca^{2+} . These cations can be modeled as being held in the

soil by weak, electrostatic interactions with negatively charged surfaces. In this idealized representation, the cations remain hydrated in a diffuse swarm or layer close to the soil surface but do not directly bond. The sum of the negative charges is termed the cation exchange capacity (CEC) of soil. Early work focused on agricultural soils in which the source of the negative charge was largely from clay minerals with isomorphic substitution, e.g. an Al atom with +3 charge substituting in the crystal structure where a Si (silicon) atom with a +4 charge is usually located. This type of charge is termed 'permanent,' because it is independent of pH. Cation exchange processes were relatively straightforward to study in these settings, but are more complex in soils with 'variably' (pH-dependent) charged minerals, such as iron (Fe) oxyhydroxides, and in acidic soils where aluminum is more soluble and reactive. The only finding from Way's early work that has not held up is his assertion that soil organic matter (SOM) had no charge. Subsequent work has shown the importance of SOM, and in fact, it is the primary source of CEC in many soils, especially those with a low clay content. Research on cation exchange processes has continued for the past 170 years with many advances in both our conceptualization and mathematical description of exchange reactions, although much of the focus has remained on clay minerals as the source of negative charge.

The preponderance of cations that are held in an exchangeable form in soils, Ca^{2+} , Mg^{2+} , K^+ , and sometimes NH_4^+ , are important plant nutrients. Their availability for plant uptake is primarily a function of a soil's CEC, although some K^+ and NH_4^+ can be fixed and only slowly released from the interlayers of some 2:1 clay minerals. These cations are relatively small and weakly polarizable (hard) Lewis acids (accept a pair of electrons) and interact electrostatically with negatively charged soil surfaces (similar to hard Lewis bases). Because of its prevalence in most parent material, its divalent charge and its energy of hydration, Ca^{2+} is usually the dominant exchangeable cation in the normal pH range of agricultural soils, e.g. pH 5.5–7.5 (Table 1). The CEC of more acidic soils can be dominated by exchangeable Al^{3+} , whereas more alkaline soils may be sodic and dominated by exchangeable Na^+ . The currently used unit for CEC is centimole of charge per kilogram ($\text{cmol}_c \text{ kg}^{-1}$) and most soils range between 5 and 20 $\text{cmol}_c \text{ kg}^{-1}$. This unit was inherited from the commonly used, and numerically equivalent, pre-SI unit of milliequivalent per 100 g of soil and is likely to be the only instance when one will use the cmol notation. The range of CEC in different soils is quite wide, but higher is not necessarily better (in an agricultural sense) as long as sufficient nutrient cations are retained. Coarse-textured soils, low in clay and SOM, may have very low CEC and not be able to retain sufficient calcium for optimal plant growth. The relative affinity of soil surfaces for Ca^{2+} over Na^+ is of global significance because these two elements have similar abundance in the Earth's crust; most of the Na^+ released from weathering ends up in the oceans while Ca^{2+} is retained in soils.

Sources of charge in soils

Permanent negatively charged mineral surfaces

Early researchers recognized that clay was a source of negative charge in soils, but the structure of clay minerals, a subgroup of phyllosilicates, was not understood until the advent of X-ray diffraction techniques in the 1930s. Based on the work of Linus Pauling and others, the layered structure of some common clay minerals was found to have cations substituted by a different cation of similar size but lower charge. For example, the Si^{4+} cation in a silica sheet could be substituted by Al^{3+} , or an Al^{3+} cation in an aluminum sheet could be substituted by Mg^{2+} or Fe^{2+} . Most clay minerals that develop charge have two silica sheets sandwiching one aluminum sheet (2:1 layer) and this isomorphic substitution results in a net negative layer charge. In some geologically-derived phyllosilicates, such as muscovite, the layer charge is balanced exactly by an interlayer cation, such as K^+ , that is part of the crystal structure, and the overall mineral charge is zero. In many soil-derived clays and weathered inherited phyllosilicates, the layers are no longer held tightly together

Table 1 Median CEC, exchangeable cations, organic carbon, and pH for selected World Reference Base (WRB) soil groups compiled from the World Inventory of Soil Emission Potential (WISE) database (Batjes, 2009).

WRB group	approx. USDA	n	CEC _e <i>cmol_c/kg</i>	CEC _t	Exch-Ca	Exch-Mg	Exch-K	Exch-Na	Exch-Al	Org-C <i>g/kg</i>	pH _w
Acrisol	Ultisol, Oxisol	244	3.9	8.8	1.14	0.60	0.20	0.10	0.80	17.0	4.80
Andosol	Andosol	99	7.6	25.0	4.80	1.90	0.40	0.20	0.10	43.2	6.00
Arenosol	sandy Entisol	158	4.0	4.3	2.75	0.60	0.20	0.10	0.10	3.0	6.90
Calcisol	Aridisol	99	27.9	17.1	23.40	2.80	0.70	0.30	0.00	6.8	8.20
Cambisol	Inceptisol	372	14.9	18.6	10.40	2.20	0.40	0.20	0.10	13.0	6.90
Ferralsol	Oxisol	225	2.7	9.2	0.40	0.30	0.10	0.05	1.04	18.0	4.76
Fluvisol	alluvial Entisol	184	17.8	17.0	12.30	2.90	0.50	0.30	0.00	11.0	7.30
Luvisol	Alfisol	357	11.2	11.7	7.50	1.90	0.40	0.20	0.10	8.0	7.00
Phaeozem	humid Mollisol	168	23.8	23.5	18.00	3.20	0.60	0.20	0.00	14.8	7.05
Regosol	Entosols	107	6.0	9.0	4.10	1.20	0.30	0.10	0.00	9.9	7.00
Solonetz	natric Aridisol	117	15.8	13.5	8.80	2.80	0.60	0.90	0.00	8.0	8.00
Vertisol	Vertisol	436	49.8	46.5	34.10	11.00	0.70	0.66	0.00	9.0	7.60

Podzols were not well represented in this database (see Table 2).

WRB groups are from WRB (2006). The approximate equivalent in USDA soil taxonomy (approx.. USDA) is also provided. The effective CEC (CEC_e) was calculated from the sum of exchangeable (Exch-) cations and the total CEC (CEC_t) was measured at a reference pH of 7.0 or 8.2. The pH (pH_w) is the pH in water.

and water and associated dissolved ions can enter, e.g. illite, a hydrous mica. The exposed clay surfaces have a negative charge that must be balanced by cations in the soil solution. The geometry, distribution and strength of the negative charge is not sufficient to form inner-sphere bonds (strong ionic or covalent associations excluding water molecules) with the cations and a diffuse swarm of hydrated cations balances the charge. This negative charge is considered permanent because it is usually not affected by soil pH, ionic strength, or the type of exchangeable cations present. This permanence may be somewhat less so in acidic soils that foster the precipitation of positively charged polynuclear Al-hydroxides in the clay interlayers, which block some of the permanent charge. The types of soil clay minerals that display this permanent charge tend to dominate the clay-size soil fraction in soils of the temperate regions, such as Europe (including Russia) and the USA, where most of the initial research on soils was performed.

Variably charged mineral surfaces

There are clay minerals, which are a class of phyllosilicates, and there are clay-sized minerals, which include non-silicates such as Al, Fe and Mn (manganese) (oxy)(hydr)oxides. This may be confusing because, in soil science, the word clay is used for two different purposes—particle size and mineralogy. The non-silicates are much more prevalent in acidic and highly weathered soils, such as found in the tropics. The charge on these oxide surfaces is not derived from structural substitution but, instead, from surface functional groups. These are usually -OH groups that can lose or gain an H^+ (hydrogen) and take on a negative or positive charge. The propensity to dissociate and create a negative charge depends on a number of factors that include the nature of the metal, the structure of the mineral and the ionic strength and pH of the solution. Different ions can also affect the charge, but the ion determining primary electrical potential is H^+ and thus pH is usually the controlling factor. Some of the variably charged minerals can have a net positive charge, especially at low pH, and there is a pH at which the overall net charge is zero. These variably charged minerals were not widely researched until the last part of the 20th century, when there was more focus on acidic subtropical and tropical soils. However, Sante Mattson, working with a variety of soil types in Sweden, clearly showed variable charge in his 1931 paper: 'The laws of soil colloidal behavior: VI. Amphoteric behavior' (Mattson, 1931). Most other work during this decade, and for many to follow, focused on phyllosilicates that have only minor variable charge on the edges of the Al and Si sheets.

Charges of soil organic matter functional groups

The importance of SOM for the CEC of soils, similar to variably charged minerals, was documented in the early part of the 20th century by a few researchers, including Mattson, but its general importance for CEC was only recognized decades later. It is now clear that both clay-size minerals (phyllosilicates and metal oxyhydroxides) and SOM contribute to the negative charge of soils and SOM can be the primary source of charge. This is especially true in many forest soils developed from coarse textured parent material, e.g. glacial drift, where surface horizons can be quite low in clay content but quite high in SOM, and higher CEC is strongly correlated with higher SOM (Fig. 1) (Johnson, 2002). The charge on SOM develops from the dissociation of H^+ from primarily carboxylic groups and to a lesser extent phenolic groups. These functional groups have been found to have a wide range of acid strengths (pKa values, which measure the propensity of an acid to release H^+) so that raising or lowering the pH will usually increase or decrease the negative charge of soil linearly, and buffer its pH value. What is not widely recognized, however, is that the negative charge per amount of SOM does not vary much in field soils below a pH of about 5.5 even though adjusting the pH of an individual soil sample will cause a dramatic change (Ross et al., 2008). This is the result of the interplay between Al, pH, and the development of SOM-associated charge, and will be discussed further below. The SOM does not develop detectable positive charge at any pH. Thus, it is not amphoteric like the metal oxyhydroxides. Adding organic residues, manure or compost to soil can bring about many physical and chemical improvements and is an effective means of increasing a soil's CEC. This can be especially important for sustainable agricultural production on low-clay soils where increased SOM may be needed to maintain soil fertility.

Cation exchange concept and models

Conceptual and mathematical models of cation exchange reactions were developed in the early part of the 20th century with the assumption of an idealized, permanently charged surface. An early approximation, the Helmholtz model, describes the particle–soil solution interface as a charged surface with associated, adsorbed cations. Experimental results were however better described by an exponential decrease in the concentration of cations from the negatively charged particle surface into the bulk solution. This so called diffuse double layer (DDL) is explained by the Guoy–Chapman theory. The DDL, later renamed the electrical double layer (EDL) to be more explanatory, extends out into the soil solution to a distance at which there is no longer any electric force from the surface and the cationic and anionic charges cancel out. Thus, two opposing forces act on ions in the soil particle–bulk solution interface—the electrostatic attraction of cations toward the negatively charged surface (together with repulsion of anions) and diffusion forces by which the excess of cations near the surface migrate along the concentration gradient toward the bulk solution (anions migrate in the opposite direction). The thickness of the EDL can be calculated by the Guoy–Chapman theory, and it varies not only with the magnitude of surface charge but also with the valence and concentration of the ions in the bulk solution. Divalent cations will create a thinner double layer than monovalent cations, as will a larger concentration of ions (a higher ionic strength). The EDL model and associated equations help to explain flocculation and dispersion of charged soil particles as influenced by the relative abundance of cations such as Ca^{2+} (promotes thinner EDL that allows flocculation of soil particles) and Na^+ (promotes thicker EDL that keeps particles dispersed). To improve the fit further to experimental data, several modifications of the

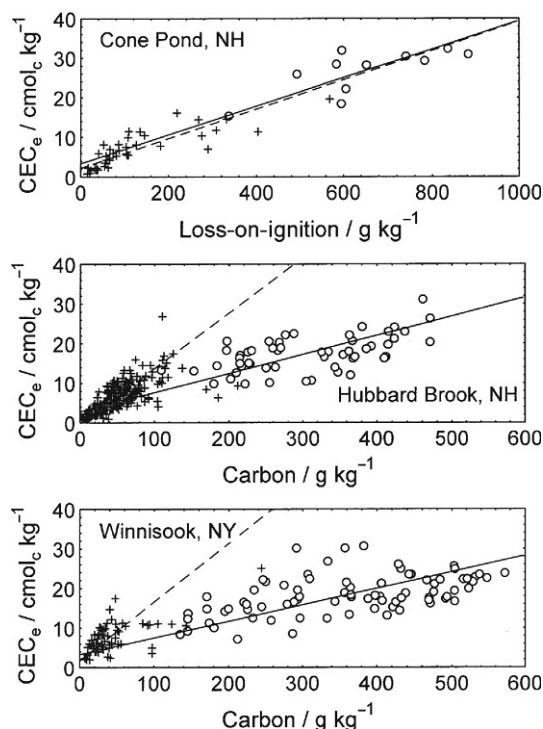


Fig. 1 Relationship between soil organic matter and effective cation exchange capacity (CEC_e) at three different forested watersheds in the northeastern USA (Spodosols and Inceptisols in USDA Taxonomy). Open circles are O horizons and '+' are mineral horizons. Organic matter was measured in the top panel by loss-on-ignition and a direct measurement of carbon ($\sim 58\%$ of SOM) was used in the bottom two panels. The effective CEC (CEC_e) was determined by the sum of exchangeable cations. The source is Johnson CE (2002) Cation exchange properties of acid forest soils of the Northeastern USA. *European Journal of Soil Science* 53: 271–282, part of Fig. 2, which is reproduced with permission.

Guoy–Chapman theory have been proposed. The Stern modification adds a layer of strongly adsorbed cations, also referred to as the Helmholtz layer that effectively decreases surface charge and thus the EDL thickness. In the ideal case, cations adsorbed in the Stern layer are imagined as inner-sphere bonded (non-hydrated), whereas the cations in the EDL are considered fully hydrated (Fig. 2). The propensity of a cation to be adsorbed is a function not only of charge but also the energy of hydration, i.e. the energy released when the cation reacts with water; this reflects the strength of the bond between the cation and the first shell of water molecules. A cation from the same group in the periodic table, e.g. the alkali metals or the alkaline-earth metals, will be preferentially adsorbed in its non-hydrated state if its atomic number and ionic radius are higher and therefore its energy of hydration is lower. Thus K^+ , atomic number 19, will be adsorbed more strongly than Na^+ , atomic number 11, and Ca^{2+} , atomic number 20, will be more strongly adsorbed than Mg^{2+} , atomic number 12. This is the basis for cation selectivity, discussed further in the next section. The theory and equations used to describe the EDL were initially based on the premise of permanently charged surfaces of clay minerals, but have since been adapted for variably charged minerals and functional groups of SOM to cover cation exchange and surface complexation reactions of metals and their ions in soils (Goldberg, 1992; McBride, 1994; Sparks, 1995; Tipping, 2002; Evangelou and Phillips, 2005; Essington, 2015; Strawn et al., 2015).

Cation exchange equations

A variety of equations have been developed to describe exchange reactions between cations in the bulk solution and cations in the so called exchange phase (weakly adsorbed at the surface or in the EDL) of soils. In its simplest form, reactions are restricted to exchange between two cations of the same charge (homovalent, e.g. Na^+ and K^+) or different charge (heterovalent, e.g. Ca^{2+} and K^+) taking place on permanently charged clay mineral surfaces. These reactions take the form of simple chemical equilibria equations, but differ primarily in how the exchange phase is expressed. The difficulty in developing true equilibrium constants, i.e. thermodynamic stability constants, is primarily because of uncertainty in how to calculate the activity of the cations (in principal their effective concentrations) at the surface and in the EDL. It is also difficult to account for electrostatic effects at the surface, i.e. the work required to migrate cations from the EDL into the bulk solution. No suitable method is known. When applied to soils, the heterogeneity of permanent and variably charged surfaces also limits transferability of equilibrium constants. Because the constant in a cation exchange equation incorporates all these factors and uncertainties it is termed a selectivity coefficient and is referred to as conditional rather than being constant. In systems with variably charged particles it also depends strongly on pH and ionic strength. Nonetheless, these equations have been useful in applications such as the soil fertility of the common nutrient cations, effects of

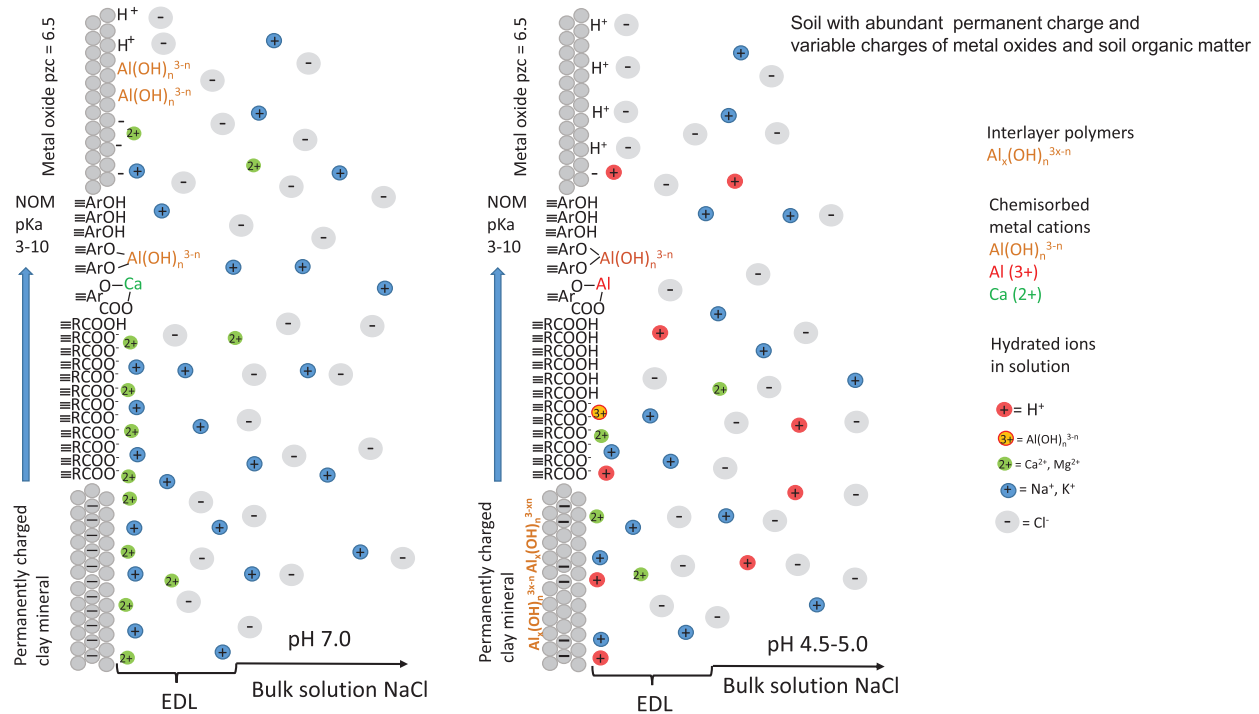


Fig. 2 Conceptual description of the particle surface–soil solution interface of a theoretical soil composed of negatively, permanently charged clay and variably charged natural organic matter (NOM) functional groups and metal oxide surfaces. Anions in solution are for simplicity represented by chloride. The electrical double layer (EDL) has a thickness on the nanometer scale depending on pH and ionic strength. The effective cation exchange capacity (CEC_e) is composed of cations held in the EDL: mono- and di-valent base cations (at pH 7.0, left), and H^+ , $Al(OH)_n^{3-n}$ and BC (at pH 4.5–5.0, right). The total CEC (CEC_t) in addition includes non-exchangeable $RCOOH$ functional groups (H_t-H_e) with a pK_a value of a reference pH (7.0 or 8.2) and below, as well as chemisorbed Ca and Al cations.

liming and in predicting the effect of Na^+ in irrigation water on soil dispersion. Two of the most commonly used equations are presented below and the reader is referred to soil chemistry text books such as [Essington \(2015\)](#) for additional permutations and discussion. Some of the more recently developed equations have a stronger footing in thermodynamics but do not necessarily provide better results.

The Vanselow equation expresses the exchange phase cations as the mole fraction of the total CEC. For a homovalent exchange reaction, this is the same as moles of charge, but this is not the case for heterovalent exchange. In the equation below, the mole fraction of Na^+ , N_{Na^+} , is the moles of Na^+ in the exchange phase divided by the sum of the exchange phase moles of $Ca^{2+} + Na^+$.

Vanselow

$$K_V = N_{Ca}(Na^+)^2 / N_{Na^+}^2 (Ca^{2+}),$$

where K_V is the Vanselow constant, N is the mole fraction in the exchange phase and $()$ is the solution activity.

Gapon

$$K_g = [Ca_{1/2X}] [Na^+] / [NaX] [Ca^{2+}]^{1/2},$$

where K_g is the Gapon constant, X is the exchange phase and the solution cations are expressed as concentrations $[]$.

The Gapon equation expresses the exchange phase in moles of charge instead of a simple mole fraction and uses molar concentration instead of activity in the solution phase. It is probably evident that these two equations will provide different results, but they do not differ dramatically over the normal range of CEC and exchangeable cation concentrations.

What becomes clear when applying these equations is that there is a strong selectivity for different cations, even those of the same valence. For example, at equal concentrations of Ca^{2+} and Na^+ in solution, the equilibrated exchange phase would be over 95% Ca^{2+} . The constant for Ca^{2+}/K^+ exchange tends to be a factor of 10 different from that for Ca^{2+}/Na^+ , and the exchange phase for K^+ is commonly much higher than that of Na^+ . The relative selectivity for common exchangeable cations, also known as the lyotropic series, is $Al^{3+} > Ca^{2+} > Mg^{2+} > K^+ \approx NH_4^+ > Na^+$. The effect of valence is much stronger than that of different hydration energies within a valence, but the latter is also important. Additional cations could be added to this list but these six dominate the exchange phase across a range of soil types and pH. Many other divalent cations found in soils are transition metals, such as Cu^{2+} (copper), Zn^{2+} (zinc), Cd^{2+} (cadmium) and Pb^{2+} (lead), and their interaction with soil surfaces is usually more complex (due to inner-sphere and hydrolysis reactions) than the outer-sphere hydrated ionic behavior assumed for exchange reactions. This complexity is also true for the behavior of Al^{3+} in soils, discussed further below.

Extension of the cation exchange concept to variably charged soils

When the cation exchange concept, originally developed for reactions taking place at the surfaces of permanently charged clays in agricultural soils, was applied to a wider range of soils cations other than Na, K, Mg and Ca had to be considered. In acidic forest soils (e.g. Alfisols, Inceptisols, Spodosols) and highly weathered soils (Oxisols and Ultisols) dominated by variably charged metal oxide minerals and functional groups of SOM, the ‘major cations’ taking part in cation exchange reactions at surface charges of soil components would include H^+ and Al cations (Al^{3+} and its hydrolysis products, collectively written $Al(OH)_n^{(3-n)+}$; $n = 0-3$). During the “acid rain research era” (roughly 1970s to 1990s) simplified models were developed in which the cation exchange between Al^{3+} and the so called base cations (BC: Na^+ , K^+ , Ca^{2+} , Mg^{2+}) was linked to the pH-dependent solubility of the $Al(OH)_3(s)$ mineral gibbsite. This process was described as a major pH buffering mechanism by which soils neutralized H^+ in acid rain and indirectly controlled the nutrient availability and acidity in runoff from soils to surface waters (Christophersen et al., 1982 [ILWAS model]; Cosby et al., 1985 [MAGIC model]; Reuss, 1983). This type of model fits data from B horizons of Spodosols affected by acid rain and with a pH exceeding approximately 4.0–4.5 (Fig. 3), but fails at lower pH as discussed below.

Somewhat earlier, research on clay mineral chemistry conducted at pH values well below neutrality had revealed that the hydrolysis of Al and its products (the cations $Al(OH)_n^{(3-n)+}$) also played a previously overlooked role at permanently charged surfaces (Harwood and Coleman, 1954; Low, 1955). Hydroxy-interlayered vermiculite (HIV) and smectite (HIS) minerals denote permanently charged minerals with part of their negative charge ‘blocked’ by $Al(OH)_n^{(3-n)+}$ cations and/or layers of positively charged polynuclear $Al_x(OH)_n^{(3x-n)+}$ polymers. This effect is largest at pH values in the approximate range of 4.5–6.0 (cf. Fig. 2). When pH is raised further, the hydrolysis of Al and nucleation of non-charged Al-oxyhydroxide minerals, e.g. gibbsite $Al(OH)_3(s)$, reaches completion and covers the interlayers of these minerals, with the permanent negative charge of the clay restored (Bertsch and Bloom, 1996). Similar to Al cations blocking interlayers of HIV and HIS, $Al(OH)_n^{(3-n)+}$ cations may form thermodynamically very stable complexes with functional groups of SOM, e.g. di/tri carboxyls, phenol and quinone (McColl and Pohlman, 1986). These strongly bonded Al cations are not considered readily exchangeable against other cations. When the cation exchange concept was extended to variably charged minerals and soils dominated by such minerals, the CEC concept therefore required some redefinition and extension.

Exchangeable and non-exchangeable aluminum and hydrogen ions

One important extension of the cation exchange concept for variably charged soils is the division of Al into exchangeable and non-exchangeable forms. Aluminum chemistry is markedly different from that of the other major metal cations in soils, which are alkali and alkali-earth metals. It is a primary component of most clay minerals and one of the major elements in the Earth’s crust (third most abundant after oxygen and silicon) but is relatively insoluble at a pH above about 5.5 – 6.0. Below this pH, Al can be solubilized but is also quite reactive and will readily form strong complexes with oxygen-containing functional groups of SOM and oxygen-containing anions such as sulfate and phosphate, and will also hydrolyze to form soluble and insoluble Al-hydroxy polymers. Even though Al participates in numerous reactions in soils, an exchangeable fraction may be operationally defined. A selection of neutral salts of cations that are non-specifically adsorbed to soil particles (e.g. KCl, NH_4Cl , $BaCl_2$) will extract a similar amount of Al from a soil, and this is operationally defined as exchangeable Al. Repeated extractions or leaching with a neutral salt,

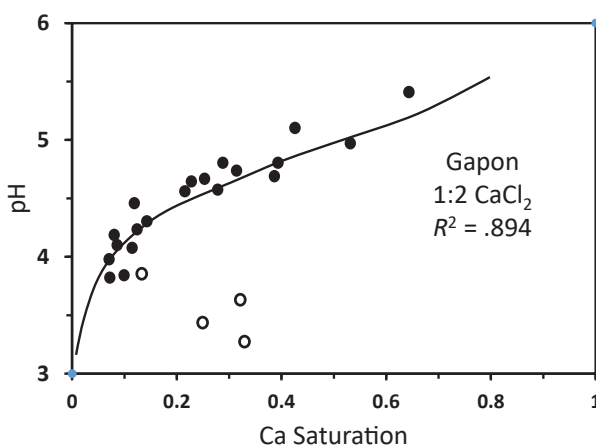


Fig. 3 Relationship between Ca saturation of CEC_e and pH (measured in 0.002 M $CaCl_2$) in mineral horizons of Spodosols. A model based on cation exchange between Al and Ca, depicted by the Gapon equation, linked to pH dependent solubility of gibbsite, $Al(OH)_3(s)$, fits observations in B horizon samples (black squares and black curve). In the E horizon samples (open symbols), the model cannot explain the observations. This is because when pH drops below 4.0 (corresponding to ~ 4.5 at pH in water), Al exchanges with H rather than with Ca (or other BCs), and the model collapses. Printed with permission from Reuss JO, Hopper RWE, Walthall PM, and Roswall EC (1990) Aluminum solubility, calcium-aluminum exchange, and pH in acid forest soils. *Soil Science Society of America Journal* 54: 374–380.

or the use of stronger extractants (i.e. more acidic and/or containing polyvalent metal cations competing with Al for oxygen groups or ligands complexing with Al cations) will however extract more Al. This suggests that there is a continuum of strengths of adsorbed Al in soils, but with a neutral-salt-defined fraction assumed to participate in exchange reactions similar to other exchangeable cations. Notably, this fraction will be strongly pH dependent.

Non-exchangeable Al is represented by Al cations (Al^{3+} and partly hydrolyzed forms collectively captured by the formula $\text{Al}(\text{OH})_n^{(3-n)+}$) strongly complexed by SOM functional groups, as well as $\text{Al}(\text{OH})_n^{(3-n)+}$ cations and/or layers of positively charged $\text{Al}_x(\text{OH})_n^{(3x-n)+}$ polymers captured in interlayers of permanently charged minerals and amorphous Al-oxyhydroxide minerals. Notably, this non-exchangeable fraction is still a minor fraction of the total quantity of Al in soils, dominated by aluminosilicate minerals which can only be extracted by total digestion in very strong acid (e.g. concentrated HF, hydrofluoric acid). In forest surface soils with a naturally low pH (<5.0), conditions at which permanently charged clay minerals are not stable, 0.5 M CuCl_2 (copper chloride) has been successfully used as a quantitative extractant of organically adsorbed Al (Juo and Kamprath, 1979), and the difference between 0.5 M CuCl_2 and 1.0 M KCl extractable Al has been used as a measure of non-exchangeable forms of Al in these types of soil (Skylberg, 1999).

Most productive agricultural soils, including those studied in the first century of soil chemistry research, have a pH above 5.5 and no or very little exchangeable Al^{3+} . More acidic agricultural soils of the tropics and subtropics can have sufficient exchangeable Al^{3+} to cause toxicity to sensitive crops. Because of the hydrolysis reactions of Al (transformation of Al^{3+} to $\text{Al}(\text{OH})_n^{(3-n)+}$ cations and H^+) over the pH range of about 4.2 to 5.5 (even higher pH if complexed by SOM, which suppresses hydrolysis), Al has been considered to be an acid cation. With decreased pH below 5.5, exchangeable Al^{3+} should increase in concentration and become a higher percentage of the CEC, but at pH 4.2 and lower, Al hydrolysis is minimal. In this relatively narrow pH range (4.2–5.5), the Al^{3+} fraction of the CEC can be a good predictor of soil pH (negative correlation).

Similar to Al, hydrogen ions become important in the exchange processes of more acidic soils. This is particularly true when pH decreases below 4.2 and $\text{Al}(\text{OH})_3(\text{s})$ is not a stable mineral. At such a low pH, H^+ is a major cation in soil solutions, readily taking part in cation-exchange reactions. Highly acidic Oa horizons can have exchangeable H^+ greater than $10 \text{ cmol}_c \text{ kg}^{-1}$, occupying over a third of the exchange sites (Table 2). In many forest soils with a natural pH below 4.5 in the organic surface horizons and A/E horizons, cation exchange between H^+ and Al^{3+} is the major pH buffering process (Skylberg, 1999; Ross et al., 2008) and H^+ and Al^{3+} dominate the CEC over Na^+ , K^+ , Mg^{2+} and Ca^{2+} (Figs. 4 and 5).

A non-exchangeable pool of H can also be defined, similar to Al. It is composed of H^+ that is not dissociated from SOM functional groups at the pH of the soil or extraction solution (i.e. these functional groups have a pK_a value well above the pH of the soil, Fig. 2). In SOM-dominated soils, e.g. Oa (H) horizons and Histosols, the pool of non-exchangeable H^+ is large, as represented by the protonated carboxyl and phenol functional groups of SOM. This pool is denoted H_t , an abbreviation for total or titratable H^+ . The latter is understood to be non-exchangeable H^+ released from soil by the reaction with OH^- under formation of water during base titration.

Base saturation and CEC definitions

There are two important CEC definitions: the total (CEC_t) and the effective (CEC_e). Total CEC represents all potential negative charges on soil particles when the soil is raised to a reference pH of 7.0 or 8.2 (the latter often used because it is the pH reached in equilibrium with calcite) and is determined as the sum of total H^+ (H_t), total charges of $\text{Al}(\text{OH})_n^{3-n}$ (Al_t) and the sum of charges of the BC. The effective CEC is the sum of neutral salt (1.0 M KCl or equivalent) extractable charges pertaining to H, Al and BC. The sum of these neutral salt extractable cations is intended to represent the negative charge of soil at ambient conditions. Alternative methods for the determination of a soil's ambient CEC have also been developed, where pH and ionic strength are kept buffered at the ambient conditions during the extraction (Gillman and Sumpter, 1986).

The term base cation does not mean the cations Na^+ , K^+ , Mg^{2+} and Ca^{2+} are bases. The term originates from describing soil components as fully hydrolyzed metal oxides, where Na_2O , K_2O , MgO and CaO are indeed bases giving a pH > 10 in reaction with water, while Al_2O_3 and Fe_2O_3 are considered acidic oxides. The BCs are in fact hard Lewis acids that preferentially adsorb to hard

Table 2 The mean effective CEC (CEC_e), major exchangeable cations, organic carbon, and pH from acidic podzols in the northeastern USA.

Soil horizons	n	CEC_e	Exch- Ca^{2+}	Exch- Al^{3+}	Exch- H^+	Org-C	pH_w
		cmol_c/kg				g/kg	
Oa (H)	142	32.3	9.14	7.77	11.92	46.9	3.55
B	26	9.9	0.37	7.72	1.50	8.2	3.96
Oa and A	48	16.4	4.74	7.99	4.71	35.0	3.74

The CEC_e was calculated by the sum of exchangeable cations for the 142 Oa and 26 B horizons, and by the compulsive exchange method of Gillman and Sumpter (1986) for the 48 Oa and A horizons. The CEC_t was not measured for these soils but would include H_t (titratable acidity) and some non-exchangeable Al (Al_t). The pH (pH_w) is the soil pH measured in water. Data are from two sampling campaigns of regional forest soils described in Ross DS, David MB, Lawrence GB, and Bartlett RJ (1996) Exchangeable hydrogen explains the pH of Spodosol Oa horizons. *Soil Science Society of America Journal* 60: 1926–1932, used with permission.

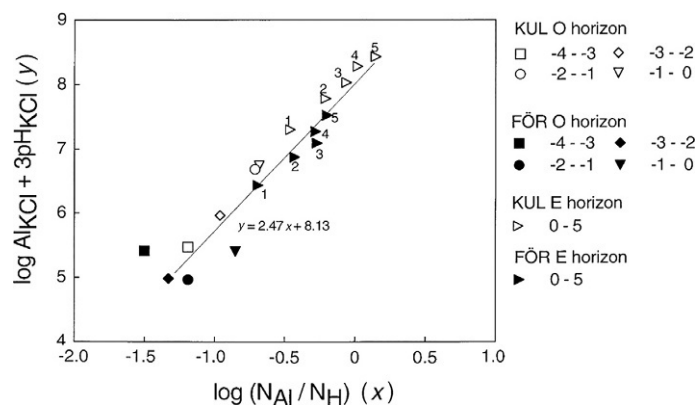


Fig. 4 The positive relationship between total adsorbed $\log (\text{Al}/\text{H}^+)$ on the x-axis and neutral salt (1.0 M KCl) exchangeable $\log (\text{Al}/3\text{H}^+)$ on the y-axis demonstrates that total adsorbed Al (defined by 0.5 M CuCl_2 extraction) and H titrated to pH 7.0 dictates the ratio of Al and H in solution. In the absence of permanent clay minerals and metal oxides (pH is lower than 4.0 in all soil samples where these minerals are not stable), Al^{3+} versus H^+ exchange by soil organic matter functional groups is the process behind the relationship. The 18 soil samples shown in the plot represent 1-cm layers of O and E horizons of two Spodosols. The numbers 1–5 represent 1-cm layers of the E horizon with the zero level set at the interface between O and E. The figure is from Skjellberg U (1999) pH and solubility of aluminium in acidic forest soils: A consequence of reactions between organic acidity and aluminium alkalinity. *European Journal of Soil Science* 50(1): 95–106 and is reproduced with permission.

Lewis bases (i.e. oxygen atoms of minerals or SOM functional groups). While Al^{3+} readily hydrolyzes in most soils, and thus releases 3 H^+ under the formation of $\text{Al}(\text{OH})_3$, the cations Na^+ , K^+ , Mg^{2+} and Ca^{2+} do not hydrolyze below pH values of 10.

The base saturation (BS) is the fraction of CEC accounted for by base cations: BC/CEC_e and BC/CEC_t .

Because of the differences in hydrolysis of BC and Al, the relationship between pH and BS (i.e. BC/CEC_e and BC/CEC_t) in a population of soil samples would be expected to be positively correlated if pH exceeds approximately 4.5 (cf. Fig. 3). Such a positive relationship is a prerequisite for the models developed during the acid rain era (e.g. Christophersen et al., 1982; Reuss, 1983; Cosby et al., 1985) to succeed as pH buffering models in soils.

A related concept is the ideal basic cation saturation ratio (BCSR). Many soil tests will report the saturation of the individual 'base' cations in relation to CEC_t . One can easily find recommendations in the popular literature for attaining ideal saturation ratios of Ca^{2+} charges at 65% (i.e. $\text{exch-Ca}^{2+}/\text{CEC}_t \times 100$), Mg^{2+} at 10% and K^+ at 5%. These ratios are not only proposed to provide proper chemical fertility but also to achieve maximum physical and biological fertility. These ratios are, again, based on historical research that has not withstood the test of time and they may be quite expensive for the agricultural producer to maintain (Kopittke and Menzies, 2007). While it is clear that there can be an imbalance in these exchangeable nutrient cations, numerous studies have shown that the ranges of the BC saturation ratios needed to achieve optimal plant growth are quite wide and encompass those usually found in agricultural soils. Some soils are naturally high or low in Mg, inherited from parent material, and adjusting the Mg^{2+} saturation of CEC_t to 10% could be difficult, expensive and unnecessary. Only excessive additions of one of these BCs can actually cause problems.

Cation exchange in acidic forest soils

In most soils pH is buffered and kept well above pH 4.5 because of kinetically driven weathering (solubilization) reactions with an abundance of aluminosilicate minerals. Exceptions are in the case of inputs of extreme mineral acidity, such as acid mine drainage, or in organic soil horizons such as the Oe (F) and Oa (H) that lack weatherable minerals. It can also be the case in some mineral E and B horizons in forest soils where the abundance, as well as the slow kinetics of aluminosilicate dissolution, is not enough to overcome the instant pH buffering by acidic organic functional groups (carboxylic acids) associated with SOM, contributed by the deposition and turnover of plant litter. In these soils the pH is well below the range for hydrolysis of Al, and the Al^{3+} cation cannot be considered acidic. Under these conditions BS is not positively related to pH (Fig. 5C). It is not difficult to find parallel increases by depth in both pH and exchangeable Al^{3+} . The reason is that H^+ and Al^{3+} ions dominate CEC_e and CEC_t , and the major cation exchange process in these soils is H^+ versus Al^{3+} . At low pH, the minor contribution from BCs and the sites they are associated with, seldom play any role in pH buffering. This is clearly shown by plots of BC, Al_e and H_e vs CEC_e where BC simply increases with increased CEC_e (as controlled by the quantity of SOM) while Al_e increases at the expense of H_e , indicative of $\text{H}^+ - \text{Al}^{3+}$ exchange (c.f. Skjellberg, 1999).

Reuss et al. (1990) demonstrated with a simple model that exchange between Ca and Al is the dominant cation exchange process in the B horizons of acid forest soil with a pH above 4.5–5.0. As expected, in these soils pH is positively correlated to BS when Al–Ca exchange is the dominant cation exchange reaction (Fig. 3). However, in E horizons, where the pH value falls below 4.5, the model applied by Reuss clearly fails, and BS is negatively correlated to pH. This observation was not explained at the time but was later shown to be due to H–Al exchange being the dominant cation exchange process in the O and E horizons of Spodosols

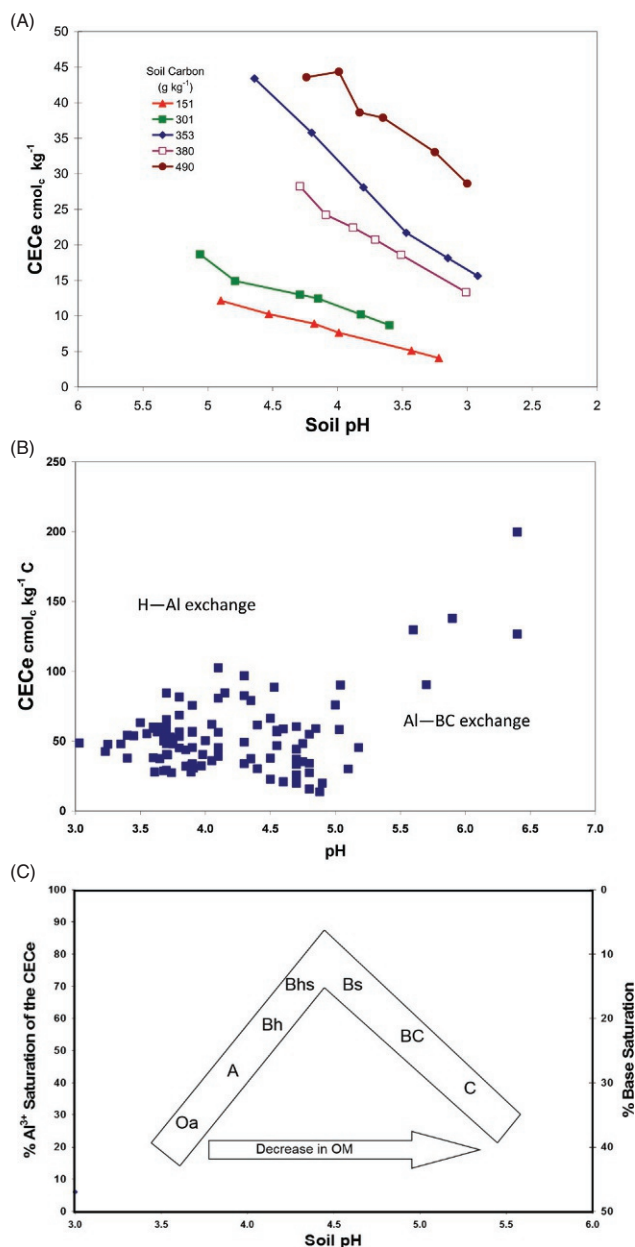


Fig. 5 (A) Relationship between pH and CEC_e in five high-SOM forest soil samples in which pH was adjusted in the laboratory in six different subsamples. (B) Relationship between pH and CEC_e in 105 forest soil samples in which the pH was not adjusted in the laboratory. (C) Change in Al³⁺ saturation of the CEC_e with increasing depth and pH in an idealized Spodosol forest soil profile. Lower Al saturation at lower pH coincides with a higher quantity of SOM. Notably the increase in pH with depth from Oa into Bhs is due to Al—H exchange and the rise in pH above 4.5 is due to Ca—Al exchange. The figures are from Ross DS, Matschonat G, and Skjellberg U (2008) Cation exchange in forest soils: The need for a new perspective. *European Journal of Soil Science* 59(6): 1141–1159, <https://doi.org/10.1111/j.1365-2389.2008.01069.x> and are reproduced with permission.

(e.g. Skjellberg, 1999; Ross et al., 2008). This interpretation was supported experimentally by the work of Hargrove and Thomas (1984) and Walker et al. (1990) on H—Al exchange in acid-washed peats. In Fig. 4, H—Al exchange at SOM sites, with a stoichiometry of 2.47, is the dominant cation exchange process in O and E horizons of two Spodosols. Deviation from the expected value of 3.0 (three H⁺ in exchange for one Al³⁺) is a common observation in studies of Al—H exchange (Hargrove and Thomas, 1984; Walker et al., 1990) and can be explained either by some organically complexed Al retaining a positive charge (R—Al⁺) and/or by electrostatic effects in the EDL, not accounted for in the model. Regardless, it is clear that exchange processes in highly acidic forest soils are better modeled by Al³⁺—H⁺ exchange rather than by exchange involving BC.

At such a low pH of 3.0, the soil solution activity of H⁺ is 1 mmol L⁻¹ which is likely to be higher than any other aqueous phase cation. This H⁺ can be considered exchangeable and, in fact, measurements of CEC in these soils relate better to the sum of exchangeable cations if an estimate of exchangeable H⁺ is included. The source of charge, however, is the dissociation of H⁺ from

organic functional groups and this confounds the view of being truly exchangeable. Nonetheless, due to dissociated acid organic functional groups these organic horizons in forest soils can have quite a large CEC_e (c.f. Table 2) even though BC may not buffer pH. Thus, the CEC_e is also critical for retaining and recycling Ca, Mg and K in these naturally acidic ecosystems.

One other interesting aspect of acidic forest soils is that, unless strongly affected by mineral acidity, the CEC_e per kg of carbon does not appear to vary greatly with pH (Ross et al., 2008). While adjusting pH in the laboratory will cause significant changes in CEC_e , a collection of samples from the field will not usually show an effect of pH below about pH 5 (Fig. 5B). This is, again, related to Al chemistry and pH buffering as discussed above. Highly acidic organic surface horizons will necessarily have lower $\text{Al}_{t,e}$ and thus higher $\text{H}_{e,t}$ because Al is minerogenic and is not recycled by litter to any great extent. Here H^+ will accumulate as a consequence of plant BC uptake and exchange for H^+ . At greater depth in the soil profile the availability of minerogenic Al increases. This will make Al available for reactions. Exchange of H for Al results in an increase in pH but not necessarily in an increase in CEC_e , because most Al bonded to SOM in these horizons is non-exchangeable ($\text{Al}_t > \text{Al}_e$). Deeper into the mineral soil horizons of acid forest soils pH will generally increase because H^+ is consumed by weathering reactions. This results in an increase in BS (Fig. 5B and C) and Al—BC exchange is the dominant cation exchange reaction (Figs. 3 and 5B). Soils receiving large inputs of H^+ may also be depleted of Al at depth in mineral soils, as revealed in some heavily acidified Dutch soils during the acid rain era (Mulder and Stein, 1994). Large inputs of mineral acidity, concentrated by stemflow, were also found to acidify forest soils rapidly in southern Sweden, lowering the CEC_e (Matschonat and Falkengren-Grerup, 2000). Apart from these highly impacted areas, it is difficult to find an effect of pH on CEC in acid forest soils because of the interplay between Al and H.

Cation exchange in highly weathered tropical soils

Soils in the tropics are diverse, possibly more so than at higher latitudes, reflecting a broad range in parent material, climate and age. While many tropical soils contain 2:1 layered clay minerals and permanent CEC, highly weathered soils may develop in humid, older landscapes. In these soils, Oxisols in USDA Taxonomy and primarily Ferralsols in WRB groups (Table 1), the loss of Si from weathering results in the clay size fraction being dominated by 1:1 clay minerals (usually kaolinite) and enriched in Al oxides (primarily gibbsite) and Fe oxides and hydroxides (e.g. goethite and hematite). These minerals have been termed 'low-activity' clays because of the virtual lack of permanent charge (Uehara and Gillman, 1981). This process of desilication (leaching of silicic acid: H_4SiO_4) was historically termed 'laterization' with resulting 'lateritic' soils low in CEC_t ($< 10 \text{ cmol}_c \text{ kg}^{-1}$), low in nutrient cations, high in exchangeable Al, and with a relatively low pH (< 5) (Pincus et al., 2017). The most weathered of these soils may be depleted of 2:1 clays, but any small remnant could contribute disproportionately to the CEC (Nakao et al., 2017).

The 'low activity' mineral fraction means that much of the CEC in the surface horizons can be derived from SOM, accumulated through litter and root turnover or through added amendments. The charge on SOM varies with pH, but always results in a negative charge within the normal pH range of these soils. Subsurface horizons, lower in SOM, may have quite a low CEC_e ($< 2 \text{ cmol}_c \text{ kg}^{-1}$). The pH-dependent variably charged minerals are amphoteric, as discussed above, and Oxisols can therefore develop substantial anion exchange capacity (AEC). The various metal oxides (Mn and Ti (titanium), in addition to the dominant Al and Fe) have a wide range in the pH of their zero point of charge (ZPC). The definitions of ZPC, and the methods for measuring it, have evolved over time and a recent soil chemistry text such as Essington (2015) or Strawn et al. (2015) should be consulted for in depth discussion. The overall net charge of these highly weathered soils evolves toward zero through loss of silica and permanent charge (Chadwick and Chorover, 2001) with SOM being the primary determinant of negative charge by coating the low activity Fe and Al oxyhydroxides minerals (Chorover et al., 2004).

Oxisols (Ferralsols) generally have low native fertility and may be difficult to manage partly because of their low CEC and pH, and therefore have low ability to retain nutrient cations. They occur worldwide throughout the tropics, but are most prevalent in South America and central Africa (Buol and Eswaran, 1999). The dominance of variable charge and the native acidity of these soils results in a wide divergence between the CEC_e and CEC_t . For example, the 225 Ferrasol samples in Table 1 had an average CEC_e of $2.7 \text{ cmol}_c \text{ kg}^{-1}$ and CEC_t of $9.2 \text{ cmol}_c \text{ kg}^{-1}$. Thus additions of lime or other materials to raise the soil pH can increase the retention of nutrient cations by increasing the CEC. However, this may create other soil physical and chemical problems related to flocculation or dispersion and anion (e.g. phosphate) retention due to changes in AEC.

Summary and outlook

Cation exchange reactions are long-studied, fundamental soil chemical processes. In near-neutral soils dominated by clay minerals with permanent negative charge, cation exchange between Ca, Mg, K and Na can be described by relatively simple stoichiometric equations. Cation exchange in soils with variably charged minerals and soil organic matter is more complicated, but can still be modeled by incorporating the effects of pH and ionic strength on surface charge. In acidic soils, H ions and Al cations become important in cation exchange reactions. In forest soils with a pH below 4.5, cation exchange concepts developed for less-acid soils, such as a positive relationship between pH and base saturation, do not apply. In these soils, reactions involving the operationally defined fractions of exchangeable and non-exchangeable H and Al can explain pH buffering through Al—H exchange. Development of geochemical models applicable across the full range of soils found worldwide, and in which all the major ion exchange and pH buffering processes are captured and integrated, remains a challenge in the field of soil science.

References

- Batjes NH (2009) Harmonized soil profile data for applications at global and continental scales: Updates to the WISE database. *Soil Use and Management* 25(2): 124–127. <https://doi.org/10.1111/j.1475-2743.2009.00202.x>.
- Bertsch PM and Bloom PR (1996) Aluminum. In: Sparks DL (ed.) *Methods of Soil Analysis. Part 3. Chemical Methods*, 3rd edn, pp. 517–550. Madison, WI: Soil Science Society of America.
- Buol SW and Eswaran H (1999) Oxisols. In: Sparks DL (ed.) *Advances in Agronomy*, pp. 151–195. Academic Press.
- Chadwick OA and Chorover J (2001) The chemistry of pedogenic thresholds. *Geoderma* 100: 321–353. [https://doi.org/10.1016/S0016-7061\(01\)00027-1](https://doi.org/10.1016/S0016-7061(01)00027-1).
- Chorover J, Amistadi MK, and Chadwick OA (2004) Surface charge evolution of mineral-organic complexes during pedogenesis in Hawaiian basalt. *Geochimica et Cosmochimica Acta* 68: 4859–4876. <https://doi.org/10.1016/j.gca.2004.06.005>.
- Christophersen N, Sepi HM, and Wright RF (1982) A model for streamwater chemistry at Birkenes, Norway. *Water Resources Research* 18: 977–996.
- Cosby BJ, Hornberger GM, Galloway JN, and Wright RF (1985) Modeling the effects of acid deposition: Assessment of a lumped parameter model of soil water and streamwater chemistry. *Water Resources Research* 21: 51–63.
- Essington ME (2015) *Soil and Water Chemistry: An Integrative Approach*, 2nd edn. Boca Raton, FL: CRC Press.
- Evangelou VP and Phillips RE (2005) Cation exchange in soils. In: Tabatabai MA and Sparks DL (eds.) *Chemical Processes in Soils*, pp. 343–410. Madison, WI: Soil Science Society of America.
- Gillman GP and Sumpter EA (1986) Modification to the compulsive exchange method for measuring exchange characteristics of soils. *Australian Journal of Soil Research* 24: 61–66.
- Goldberg S (1992) Use of surface complexation models in soil chemical systems. *Advances in Agronomy* 47: 233–329.
- Hargrove WL and Thomas GW (1984) Extraction of aluminum from aluminum-organic matter in relation to titratable acidity. *Soil Science Society of America Journal* 48: 1458–1460.
- Harwood ME and Coleman NT (1954) Some properties of H- and Al-clays and exchange resins. *Soil Science* 79: 181–188.
- Johnson CE (2002) Cation exchange properties of acid forest soils of the Northeastern USA. *European Journal of Soil Science* 53: 271–282.
- Juo ASR and Kamprath EJ (1979) Copper chloride as an extractant for estimating the potentially reactive aluminum pool in acid soils. *Soil Science Society of America Journal* 43(1): 35–38. <https://doi.org/10.2136/sssaj1979.03615995004300010006x>.
- Kopittke PM and Menzies NW (2007) A review of the use of the Basic Cation Saturation Ratio and the "Ideal" Soil. *Soil Science Society of America Journal* 71(2): 259–265. <https://doi.org/10.2136/sssaj2006.0186>.
- Low PF (1955) The role of aluminum in the titration of bentonite. *Soil Science Society of America Proceedings* 19: 135–139.
- Matschonat G and Falkengren-Grerup U (2000) Recovery of soil pH, cation-exchange capacity and the saturation of exchange sites from stemflow-induced soil acidification in three Swedish beech (*Fagus sylvatica* L.) forests. *Scandinavian Journal of Forest Research* 15(1): 39–48.
- Mattson S (1931) The laws of soil colloidal behavior: VI. Amphoteric behavior. *Soil Science* 32: 343–365.
- McBride MB (1994) *Environmental Chemistry of Soils*. New York: Oxford University Press.
- McColl JG and Pohlman AA (1986) Soluble organic acids and their chelating influence on Al and metal dissolution from forest soils. *Water, Air, and Soil Pollution* 31: 917–927.
- Mulder J and Stein A (1994) The solubility of aluminum in acidic forest soils: long-term changes due to acidic deposition. *Geochimica et Cosmochimica Acta* 58: 85–94.
- Nakao A, Sugihara S, Maejima Y, Tsukada H, and Funakawa S (2017) Ferralsols in the Cameroon plateaus, with a focus on the mineralogical control on their cation exchange capacities. *Geoderma* 285: 206–216. <https://doi.org/10.1016/j.geoderma.2016.10.003>.
- Pincus LN, Ryan PC, Huertas FC, and Alvarado GE (2017) The influence of soil age and regional climate on clay mineralogy and cation exchange capacity of moist tropical soils: A case study from Late Quaternary chronosequences in Costa Rica. *Geoderma* 308: 130–148. <https://doi.org/10.1016/j.geoderma.2017.08.033>.
- Reuss JO (1983) Implications of the calcium-aluminum exchange system for the effect of acid precipitation on soils. *Journal of Environmental Quality* 12: 591–595.
- Reuss JO, Hopper RWE, Walthall PM, and Roswall EC (1990) Aluminum solubility, calcium-aluminum exchange, and pH in acid forest soils. *Soil Science Society of America Journal* 54: 374–380.
- Ross DS, Matschonat G, and Skjellberg U (2008) Cation exchange in forest soils: The need for a new perspective. *European Journal of Soil Science* 59(6): 1141–1159. <https://doi.org/10.1111/j.1365-2389.2008.01069.x>.
- Skjellberg U (1999) pH and solubility of aluminium in acidic forest soils: a consequence of reactions between organic acidity and aluminium alkalinity. *European Journal of Soil Science* 50(1): 95–106.
- Sparks DL (1995) *Environmental Soil Chemistry*. San Diego: Academic Press.
- Strawn DG, Bohn HL, and O'Connor GA (2015) *Soil Chemistry*, 4th edn. Oxford, UK: Wiley Blackwell.
- Thompson HS (1850) On the adsorbent power of soils. *Journal of the Royal Agricultural Society of England* 11: 68–74.
- Tipping E (2002) *Cation Binding by Humic Substances*. Cambridge University Press.
- Uehara G and Gillman GP (1981) *The Mineralogy, Chemistry, and Physics of Tropical Soils with Variable Charge Clays*. Boulder, CO: Westview Press.
- Walker WJ, Cronan CS, and Bloom PR (1990) Aluminum solubility in organic soil horizons from northern and southern forested watersheds. *Soil Science Society of America Journal* 54: 369–374.
- Way JT (1850) On the power of soils to adsorb manure. *Journal of the Royal Agricultural Society of England* 11: 313–379.
- WRB (2006) *World reference base for soil resources – a framework for international classification, correlation and communication*. *World Soil Resources Reports* 103. Rome: International Union of Soil Sciences, ISRIC – World Soil Information and Food and Agriculture Organization of the United Nations.

Testing for nutrients in the soil and the environment

Antonio P Mallarino, Iowa State University, Ames, IA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	43
Soil sampling for soil testing	44
Soil-test extractants and methodologies	45
Phosphorus soil tests	45
Nitrogen soil tests	46
Calibration of soil tests for crop production	47
Quality of soil testing	48
Summary	48
References	49

Abstract

Soil testing is an important diagnostic tool for determining crop nutrient needs and as part of complex tools to assess the risk of loss from fields. This chapter focuses on nitrogen and phosphorus, for which excess loss from fields impairs surface waters. Concepts discussed include taking representative soil samples, using an appropriate sampling depth, using proper calibration with crop yield response to provide a meaning for test values to determine nutrient application rates for crops, and the role of soil testing in tools to assess the risk of nutrient loss from fields to water resources and water quality impairment.

Key points

- Soil testing is an important diagnostic tool for determining crop nutrient needs and as part of complex tools to assess the risk of loss from fields.
- Taking representative soil samples from the area sampled and using both recommended and consistent sampling depth are very important to assure proper interpretation of soil-test results.
- This chapter focuses mainly on testing soils for the plant macronutrients nitrogen and phosphorus, for which excessive losses from agricultural fields are common and impair surface waters or groundwater worldwide.
- Following recommended laboratory assurance and quality by all laboratories to improve accuracy and minimize bias among laboratories is essential for developing reliable soil testing programs.
- Proper calibration of the soil tests with crop yield response is essential to provide a meaning for test values to determine nutrient application rates.
- Calibration of soil tests also is essential when they are used by themselves or as part of complex tools to predict nutrient loss from fields and water quality impairment.

Introduction

Soil testing is and has been for many decades an important diagnostic tool for determining the nutrient needs of plants and for assessments of nutrients in the environment. Some soils are inherently deficient in plant nutrients whereas other soils originally had sufficient concentrations of nutrients in the past, but reserves have been depleted by crop harvest removal. On the other hand, fertilizer or animal manure applications of some nutrients to initially deficient soils over many years have increased topsoil concentrations to optimum or higher than optimum levels for crops. Thus, soil testing is widely accepted and used in most advanced crop production areas of the world to determine fertilizer needs for crops.

Soil testing also can be used to assess excess accumulation in the soil of harmful elements. Materials such as animal manures and industry by-products may provide various plant nutrients. However, large application rates to soils designed to dispose of the material at a low cost may result in nutrient or element (such as heavy metals) loads that are harmful to plant, animal, or human health. Nutrient management regulations are continuously being modified or developed to address land applications of waste materials. Soil testing is required in many regulations and management guidelines to assess environmentally harmful concentrations of certain elements and to determine limits for application rates.

Agricultural soils are tested routinely for soil pH, organic matter, and the primary nutrients phosphorus (P), potassium (K), and nitrogen (N). In some regions, soils are also routinely tested for the often referred to as secondary nutrients such as calcium (Ca), magnesium (Mg), and sulfur (S) and for others referred to as micronutrients because they are required in very small amounts by crops such as boron (B), copper (Cu), iron (Fe), manganese (Mn), molybdenum (Mo), and zinc (Zn). Soils receiving waste biosolid

materials from industries, crops, animal products, or drinking water processing plants are also tested for elements such as arsenic (As), cadmium (Cd), nickel (Ni), lead (Pb), mercury (Hg), and selenium (Se) among others.

Excessive N and P applications to agricultural fields and ineffective nutrient, soil, and water conservation practices are increasing nutrient pollution of water in many regions of the world. Although the basic concepts of soil testing apply to all nutrients and other elements, the dynamics in the soil and mechanisms of loss to groundwater and surface water differ radically for some nutrients. For example, P is included with other elements in a group with relatively low mobility in soils (Ca, Cd, Cu, Fe, Mg, Mn, Mo, Ni, Pb, and Zn among others). Nitrogen, especially in the nitrate form, is included in a group of nutrients and elements with greater mobility (which include B, Cl (chlorine), S, and others).

This chapter focusses mainly on commonly used soil testing for crop-availability of the plant nutrients N and P, which are also nutrients whose excess impairs many surface waters worldwide. Important concepts for these nutrients and many others include taking representative soil samples from fields, use of a correct soil sampling depth, determination of the meaning of a soil-test value for crop production and use in tools for risk loss assessments, and soil-test quality. Soil testing for research, soil quality and health, and elements harmful to human health are beyond the scope of this chapter.

Soil sampling for soil testing

For a soil testing program to be effective, besides laboratory quality control, soil samples should be collected in a cost-effective manner and should accurately represent the nutrient concentrations in the area of interest. Sampling is a critical component of the soil-testing process because there is often considerable spatial variation in the nutrients of most fields and usually represents the largest single source of error in soil testing. Sampling protocols must account for the large diversity in magnitude, structure, and spatial scale of nutrient variation present in agricultural fields and other ecosystems.

Soil nutrient variation within a crop field may be due to soil formation or management factors. Factors such as landscape position and soil parent material can cause great changes in soil texture, organic matter, drainage, and other properties. These properties affect nutrient concentrations directly through their influence on the amount of plant-available nutrient or indirectly through crop yield potential and, thus, the nutrient removal by crops. Variation caused by long-term history of management and land use practices overlays that associated with soil-formation factors. The cumulative effects of nonuniform manure or fertilizer application are sources of potentially considerable variation in soil-test results. Proximity to livestock confinement areas, feed storage areas, and field boundaries are additional examples of historical factors causing great variability in many fields. In addition, significant small-scale variation over distances as small as a few centimeters or meters usually is present in fields with long histories of cropping and fertilizer or manure applications, especially when nutrients have been applied using band methods. The challenge in these situations is to determine effective methods to delineate sampling areas within a field and the number of cores or borings needed for each composite sample to account for small-scale variation appropriately.

A variety of systematic and zone sampling approaches have been developed in different regions. The development of accurate and reliable global positioning technology, affordable geographic information systems, and variable-rate application equipment have led to widespread use of site-specific soil sampling approaches in Canada, the United States, Europe, Australia for example. These sampling approaches typically are used to generate a soil fertility map to serve as an input to computer-controlled equipment for applying varying rates of one or more materials when precision agriculture and variable-rate technologies are used. One such approach is zone sampling, which is essentially a stratified sampling, by which field subregions with more homogeneous properties than an entire field are delineated. Landscape position, soil color, soil mapping unit, crop growth differences, soil electromagnetic inductance, and other soil or crop canopy properties are assessed by direct or remote sensing methods. Aerial images are examples of factors often used to help define sampling zones. The basic assumption for these sampling procedures is that soil-test levels or fertilizer use efficiency are more homogenous within these zones than across an entire field.

This assumption, however, sometimes does not hold especially when land has had a long history of fertilization or manure application and when soils or landscape forms differ little within a field. Therefore, another site-specific approach that has gained adoption in some countries involves grid soil sampling, by which soil-test patterns in a field are determined by a dense, systematic sampling of composite samples. A grid size of approximately 1 ha is a common sampling approach today in the United States, and cores or borings for composite samples are collected from each entire area (cell sampling) or from a very small section (point sampling) positioned systematically or randomly. Small-scale variability of P and K is so extreme in some fields with long histories of fertilizer or manure application that accurate within-field soil fertility mapping is very difficult and sometimes economically unfeasible.

Many articles and book chapters about soil sampling have been written since the 1940s but much uncertainty still exists on how best to perform site-specific soil sampling and generate accurate soil fertility maps. Recent useful research results and concepts for soil sampling for crop production or risk assessments of P loss using precision agriculture technologies have been addressed in recent years by Mueller et al. (2001), Bianchini and Mallarino (2002), Kitchen et al. (2003), Fridgen et al. (2004), Mallarino and Witty (2004), Flowers et al. (2005), Mallarino et al., (2007), Pennock et al. (2007), Sawchik and Mallarino (2008), and Lawrence et al. (2019).

Soil-test values also change significantly with depth. This results from a combination of soil-forming and management factors and from the different mobility of nutrients in soils. The elements with lower mobility (see above) tend to accumulate near the application point (often the shallowest soil layers) because they are more strongly retained by soil constituents than the more mobile nutrients. The tillage system and method of application greatly influence vertical nutrient stratification. For example, the

significant stratification of both P and K in no-till soils is well known, but subsurface nutrient application methods can reduce their concentration near the soil surface. Thus, the proper depth for soil sampling and its consistency are important considerations (James and Wells, 1990; Lawrence et al., 2019). Variation in soil-test results can also result from differences in sampling depth. Soil samples should be collected at the same depth used in the calibration research that serves as the quantitative basis for interpretation of the soil-test and nutrient recommendations. Typically, this is a depth of 15–20 cm for the relatively immobile nutrients (such as P, K, Ca, Mg, Cu, Fe, Mn, Zn) and 30–60 cm for the more mobile nutrients (such as nitrate-N). However, the most appropriate soil sampling depth varies with the objective of the soil testing. For example, a sampling depth of 15–20 cm is usually the best to assess nutrient sufficiency for crops or nutrient loss with subsurface drainage, but a shallower sampling depth is often better to predict loss of P loss with surface runoff especially with no-till management or pastures (Vadas et al., 2005).

Soil-test extractants and methodologies

Phosphorus soil tests

Soil tests for P and other elements have been developed based on an understanding of the chemical forms in which the elements exist in the soil and empirical work. In most soils, P is primarily associated with aluminum (Al), Fe, and Ca. Phosphorus precipitation and (or) adsorption to Fe and Al oxides generally predominates in acid ($\text{pH} < 7.0$) to neutral soils, whereas reactions with Ca predominate in soils with significant amounts of calcium carbonate. The degree and strength to which P is bound in soils are largely determined by the amount and types of Fe and Ca compounds present and by other soil properties such as pH, organic matter, clay mineralogy, and the amount of P currently present in the soil. The influence of Fe increases and that of Ca decreases as rainfall, temperature, and weathering increase. These factors are also important in determining the plant availability and solubility of other immobile elements.

Phosphorus soil-test methods for agronomic use employ dilute strong acids, dilute weak acids, complexing ions, and (or) buffered alkaline solutions (Beegle, 2005). Most tests have been developed to reflect the soil properties related to P sorption that predominate in a region. For example, the Bray P-1 test was developed for use in the acid to neutral soils of the north-central region of the United States, the Olsen test (or sodium bicarbonate) was developed primarily for use on calcareous soils of the same region, and the Mehlich-1 test was developed for the low cation-exchange capacity and highly weathered soils common to southeastern regions. The Mehlich-3 extractant, although developed in the 1980s, has been rapidly adopted in the US and many other countries during the last decade because it is suitable for a wider range of soil properties than other P or K tests and for other nutrients (such as Ca, Mg, Fe, and Zn). Although the results of the different soil P tests are often strongly correlated for soils with similar properties, the actual quantities extracted can differ greatly for soils with specific characteristics. For example, the Mehlich-3 test extracts 1.5–2 times the amount of P as the Mehlich-1 and Olsen tests. These observations highlight the importance of understanding the meaning of a soil-test value and proper calibration.

In the United States and most European countries, regulatory initiatives have been established to mitigate water-quality problems associated with N and P water pollution. Agronomic soil-test methods designed to assess plant-available nutrients are often used for environmental purposes. However, increasing concerns about nonpoint-source P pollution from agriculture has prompted questions about the suitability of agronomic soil P tests for environmental purposes. An effective environmental soil test should measure the nutrient forms most important to eutrophication or other negative environmental impacts. Eutrophication is a process whereby water bodies, such as lakes, estuaries, or slow-moving streams receive excess nutrients that stimulate excessive aquatic plant or algal growth. This enhanced growth reduces dissolved oxygen in the water when dead plant or algal material decomposes, which causes ecosystem imbalances and may promote growth of toxic bacteria and unpleasant odor. Several soil tests have been proposed as environmental tests for soil P because they may provide better estimates of dissolved P delivery to water resources (Kovar and Pierzinsky, 2009). Two of these tests are being used extensively. One test based on FeO-impregnated paper uses a sink approach to extract soil-bound P that is most likely to be available to aquatic organisms. The other test is based on weaker desorption reactions and uses either deionized water or a diluted calcium-chloride solution to extract soil P. Current research indicates that usually there are linear and similar relationships between either agronomic or environmental soil P tests and dissolved P in surface runoff or subsurface drainage. As an example, Fig. 1 shows relationships between total P in the drainage discharge to a lake from five basins and the mean soil-test P of the basins measured by four soil test methods (Klatt et al., 2003). In this case, soil P measured by agronomic tests specifically developed to estimate soil P sufficiency for crops (Mehlich-3 and Olsen) shows approximately similar increasing linear relationships to P measured by so-called environmental P tests Fe-oxide coated filter paper and water-extractable P.

Indices of soil P saturation or environmental soil P tests tend to relate better to dissolved P loss in regions with extremely large soil P concentrations. In these cases, the dissolved P loss sometimes increases exponentially with increasing soil P or shows a discontinuous relationship with an initially flat response and then linear increase after a change point (Maguire and Sims, 2002; Sharpley et al., 2003).

Soil-test P results from agronomic or environmental soil tests are seldom used as a single metric to estimate P losses from fields or loads to water resources. The most important reason is that a certain concentration of an element in the soil may have different impacts on the environment depending on the amount retained by the soil and potential delivery to environmentally critical areas, which can be effected by a variety of mechanisms. For example, soils with high clay content have a greater sorption capacity for P (and for other immobile elements) than coarse-textured soils, and less P is delivered as the distance from the source to channeled water flow increases. Establishment of nutrient loads that can result in unacceptable water-quality degradation will depend on

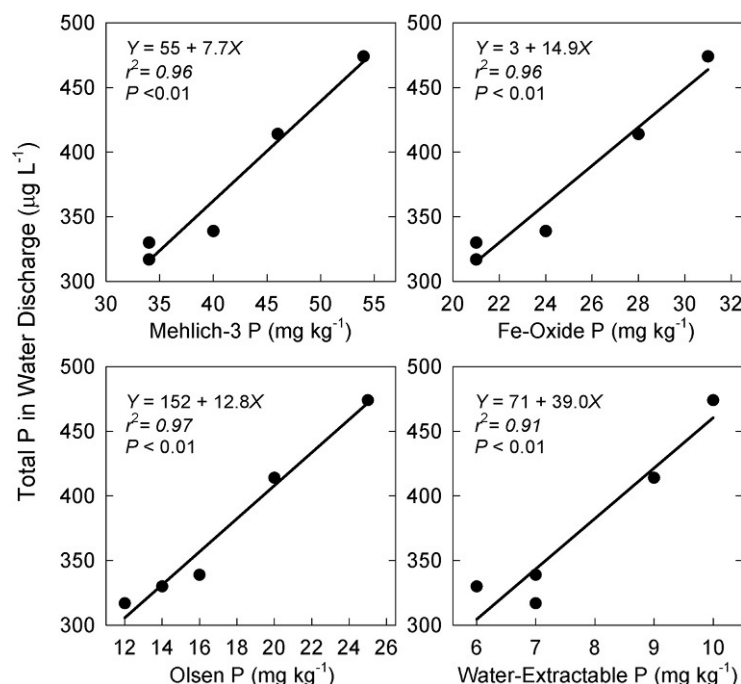


Fig. 1 Relationship between total P concentration in water discharge and soil-test phosphorus measured with four soil P tests for five basins of the Iowa Clear Lake agricultural watershed. Adapted from Klatt JG, Mallarino AP, Downing JA, Kopaska JA, and Wittry DJ (2003) Soil phosphorus, management practices, and their relationship to phosphorus delivery in the Iowa Clear Lake agricultural watershed. *Journal of Environmental Quality* 32: 2140–2149.

factors such as the distance to a sensitive water body, the effectiveness of transport pathways for water and either total or dissolved P fractions, and socioeconomic factors involving land or water use. Therefore, a critical soil-test concentration value or range, often referred to as a threshold, is unlikely to be found for nutrient impacts on the environment and is seldom recommended, whereas risk assessment tools, such as P indices, are used to assess risk of P loss from fields.

A P index is often not a full-blown model of P loss and transport but a more quantitative practical tool that considers soil-test values with other site-specific factors (such as hydrology, nutrient management, water body proximity and sensitivity to nutrient inputs) to determine the potential environmental risk of P concentrations in a field or parts of fields (Sharpley et al., 2003). The Iowa P Index, for example (Mallarino et al., 2002) includes the components total P loss with erosion, dissolved P loss with surface runoff, dissolved P loss through subsurface drainage, and surface source factors (soil-test values as well as fertilizer and manure application rates, methods, and timing) and transport factors (such as soil erosion, surface runoff, subsurface drainage, and distance to streams or water bodies).

Nitrogen soil tests

The cycling of N in soils is different from other nutrients and requires a different soil-testing approach. More than 97% of N in soils is present in organic forms and is converted to inorganic forms available to plants mostly through microbial activity. The rate of N release from organic to inorganic forms is difficult to predict because it depends on temperature, moisture, aeration, type of organic matter, soil pH and many other factors. The predominant inorganic end-result of mineralization is nitrate-N. This form of N is highly mobile in soils, and is subject to loss by leaching and to gaseous losses through denitrification. Because of these factors, N fertilizer recommendations for crops are not based solely on soil organic matter and frequently are made on the basis of crop yield potential coupled with an N-credit system and less frequently on the basis of soil testing. The N-credit system is based on empirical information of the effects of previous crops, several soil properties, and climate on crop response to N fertilization.

Soil organic matter or N fractions, mineralizable soil N, preplant soil nitrate tests, and postemergence or pre-sidedress soil nitrate tests have been investigated and proposed for a long time to estimate maize (*Zea mays* L.) N needs in the United States and other countries. A soil test has been proposed that estimates soil N supply potential for adjustments to N recommendations; it estimates ammonium N plus a labile pool of soil N that could mineralize during the growing season (Khan et al., 2001; Mulvaney et al., 2001). However, evaluations with crop yield response have been inconsistent or shown to be an unreliable diagnostic tool and so is seldom used (Laboski et al., 2008). In low-rainfall areas, where the potential for nitrate leaching is minimal and irrigation is controlled, the amount of nitrate-N in the soil profile before planting maize can be credited directly against the N needs of the crop. In more humid regions, measurements of soil profile nitrate or ammonium before planting have been unreliable for predicting the N needs of maize. However, nitrate-N testing early in the growing season has been shown to provide an index of the soil's N-supplying capability for maize. Soil samples for the pre-sidedress soil nitrate test are collected from a 0–30-cm depth after

planting the crop and before plants reach a height of 15–30 cm. Recent research across several states of the US Midwest region showed that the pre-sidedress soil nitrate test (a 30-cm sample at the V5 to V6 growth stage) was the single best predictor of maize N needs and the economic optimum N rate, and that combinations of this test with preplant profile soil nitrate or potentially mineralizable soil N tests improved the prediction only marginally (Clark et al., 2020). However, all methods produce different measurements of soil nitrate and there is uncertainty for environmental assessments because nitrate from soil samples often does not relate well to the nitrate that moves through the soil profile and is measured with lysimeters (Darrouzet-Nardi and Weintraub, 2014).

Advances in electronics and sensing technologies have resulted in intensive current research for their implementation in N or nitrate soil testing in the field or laboratory. Sensors of different types have been used successfully to measure soil moisture, organic carbon or organic matter, and salts. Recently there were advances in applications for soil nitrate testing but additional research is needed before these methods can substitute for conventional extraction laboratory methods (McBride, 2021; Zhu et al., 2021). Interpretation of the soil nitrate tests is usually complemented by consideration of the previous crop, rainfall during a 2–4 week period before or after collecting the samples, and soil properties that influence nitrate leaching. As with all soil tests, local or regional field calibrations are conducted to determine critical ranges for soil nitrate concentration.

As was discussed for P, N or nitrate, soil tests are seldom used alone as an index of potential losses to water resources and most often regulation is based on rates of N application and or time of application. The use of N indices is less widespread or accepted compared with P indices. For example, the federal Natural Resources Conservation Service (NRCS) of the United States Department of Agriculture (USDA) has proposed a nitrogen leaching index but few states have implemented it or modifications to guide agronomic practices or regulations.

Calibration of soil tests for crop production

Soil tests seldom extract the total amount of nutrients or elements in a soil sample. Soil tests have been developed to measure a fraction of the total soil nutrient concentration that correlates with plant growth. Interpreting a soil-test value requires an understanding of the impacts on test results of the extractant used, method of soil sampling, sample handling, and the intended use for the result. The amount of nutrient measured by various soil tests can vary widely depending on the extractant used. Important extractant properties include the concentration of the chemical compounds and the reaction time with the soil. The method used to measure the nutrient after extraction may be important for some nutrients. For example, colorimetric methods usually measure only orthophosphate P while other methods may also measure other forms of P. The same soil test may also measure widely different amounts of nutrients in soils with contrasting chemical and (or) mineralogical properties. Extractants used and interpretations of results often vary depending on soil properties and the purpose of soil testing (for example, measurement of total, soluble, or plant-available concentrations).

A soil-test method useful for predicting crop response to fertilization should produce values that are well correlated with plant nutrient uptake or yield. Greenhouse and field research is conducted to determine which soil-test extractant may be best suited for a given combination of soil, crop, and growing conditions. Accurate interpretations of soil-test results and appropriate fertilizer recommendations require that the relationship between the amount of a nutrient measured by a given soil test and the crop response to the added nutrient must be known. The process of determining the probability of crop response at a given soil-test value is known as soil-test correlation and the process of determining the amount of nutrients to be applied for desired yield levels is known as calibration; however, sometimes both words are used for either process. The correlation and calibration procedure usually involves growing a crop in soils that are representative of the region on which the test will be used for the application of various fertilization rates. The soils should have a broad range of soil-test values that include deficient, optimum, and above-optimum values.

The crop yield response can be expressed as an absolute value or as a value relative to the yield achieved without nutrient addition. At low soil-test values, crop yield is limited by nutrient deficiency. As soil-test values increase, yield increases at a reduced rate until a maximum yield value. At higher levels, there is no longer a relationship between the soil-test values and yield. The maximum yield value can remain approximately constant for a wide range of soil-test values or can decrease at excessively large nutrient concentrations. The soil-test value at which crop growth or yield reaches a maximum is called the soil-test critical concentration. This is the soil-test value that best separates soils where a positive crop yield response to an added nutrient is likely from those soils where a response is not likely. Various mathematical models can calculate different critical concentration values and some models are asymptotic to a maximum. For example, Fig. 2 shows the relationships between the relative yield response of maize and the amount of soil P extracted by four methods (Mallarino et al., 2005). The figure shows the fitted line for a linear-plateau model and the corresponding estimated critical concentration, but the authors indicated that the critical levels estimated by Cate-Nelson for linear-plateau and quadratic-plateau models ranged from 13 to 28 mg P kg⁻¹ across the four test methods. There is no reasonable agronomic consideration that justifies a single critical concentration. Moreover, economic considerations suggest that maximum economic yield perhaps should be the basis for determining critical concentrations instead of a physical maximum. Thus, critical concentration ranges usually are determined by using a combination of models and a variety of both agronomic and economic considerations. The soil-test response curve often is used to divide soil-test levels into several categories such as very low, low, optimum, high, or excessive.

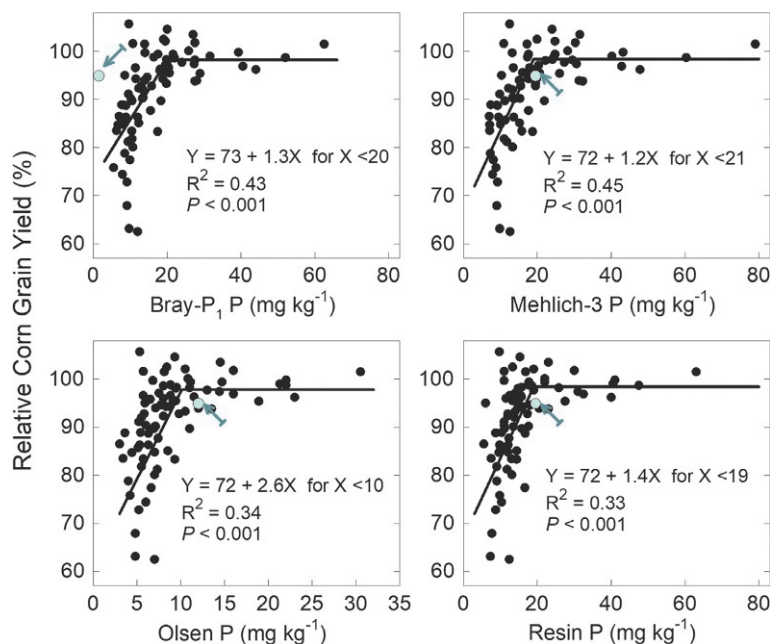


Fig. 2 Relationships between relative crop yield response and soil-test P values by four methods. The gray arrow and point indicate results for a highly calcareous soil. Adapted from Mallarino AP and Atia AM (2005) Correlation of a resin membrane soil phosphorus test with corn yield and routine soil tests. *Soil Science Society of America Journal* 69: 266–272.

A thorough understanding of crop response to nutrients and factors such as sampling date (season of the year), sampling method (mainly sampling depth), method of nutrient application (for example, band or broadcast), and application timing are needed to interpret soil-test results correctly. Other factors can also influence crop growth (climatic factors, plant population, concentrations of other nutrients in the soil, and crop cultivars among others). These factors may influence crop growth without affecting the amount of nutrient needed to produce a certain yield or they may interact with nutrient concentrations in the soil. Climatic factors, especially soil temperature, and both total rainfall and distribution during the year, are especially important for nutrients that are highly mobile in the soil, such as nitrate-N, B, Cl, and S. High nutrient mobility results in more difficult soil-test calibrations and introduces more uncertainty when interpreting soil test results. The temporal variation during the year for the more immobile nutrients can also be significant in coarse textured soils because of leaching from the topsoil to deeper layers in the soil profile or to groundwater. It is possible that advances in the near future from research on in-situ sensor technologies can account better for temporal variation in soil nutrient concentrations.

Quality of soil testing

The quality of a laboratory soil-test result can be assessed on the basis of bias and precision. Bias refers to the deviation of the analytical result from the true value and measures the accuracy of the result. Precision refers to the reproducibility of a given test value. A soil-testing laboratory could have good precision for a specific procedure but have bias in its results. Guidelines for laboratory soil test quality control and quality assurance are provided by institutions in most countries. A primary means of evaluating soil-test quality procedures as well as laboratory performance is through comparisons of the tests performed on well-mixed sub-samples of the same soil at different laboratories. In North America, a voluntary proficiency testing program developed by the Soil Science Society of America was launched in the early 2000s to improve the quality of soil testing for agronomic purposes (the North American Proficiency Testing Program, NAPT, <https://www.naptprogram.org>). Results from this program indicate that, for example, the average variability of P tests is approximately 10–15% and that 65–70% of participating laboratories produced results within this level of variability. Variability is lower within any given laboratory, with precision levels for the P tests discussed are roughly 5–10%.

Summary

Soil testing has been an important diagnostic tool for determining the nutrient needs of plants for decades. Thus, soil testing is widely accepted and used in most advanced crop production areas of the world to determine fertilization needs for crops. Soil testing also has long been used to assess excess accumulation of harmful elements in the soil. More recently, soil testing for N and P

began to be used as part of the complex tools to assess the risk of loss from agricultural fields to water resources. This chapter focused mainly on testing soils for the plant nutrients N and P, which are also nutrients whose excess loss from fields impairs surface waters worldwide. Important concepts discussed, which apply to most nutrients, included taking representative soil samples from fields, use of an appropriate soil sampling depth, which often changes for different nutrients and management systems, use of proper calibration with crop yield response to provide a reliable meaning for the soil-test value to determine recommended nutrient application rates for crop production, and the role of soil testing in complex tools to assess the risk of nutrient loss from fields to water resources and water quality impairment. The applicability of different soil tests for a nutrient varies greatly across soils and regions of the world. In addition, the soil-test values that may indicate sufficient concentrations for crops or excessive levels that are likely to result in excessive nutrient loss to water resources also varies greatly. Therefore, local research is essential for the correct development and implementation of new soil tests and soil testing programs.

References

- Beegle D (2005) Assessing soil phosphorus for crop production by soil testing. In: Sims TS and Sharpley AN (eds.) *Phosphorus, Agriculture and the Environment*. Agronomy Monographs. pp. 123–143. Madison, WI: ASA-CSSA-SSSA.
- Bianchini AA and Mallarino AP (2002) Soil sampling alternatives and variable-rate liming for a soybean-corn rotation. *Agronomy Journal* 94: 1355–1366.
- Clark JD, Fernández FG, Veum KS, Camberato JJ, Carter PR, Ferguson RB, Franzen DW, Kaiser DE, Kitchen NR, Laboski CAM, Nafziger ED, Rosen CJ, Sawyer JE, and Shanahan JF (2020) Soil-nitrogen, potentially mineralizable-nitrogen, and field condition information marginally improves corn nitrogen management. *Agronomy Journal* 112: 4332–4343.
- Darrouzet-Nardi A and Weintraub MN (2014) Evidence for spatially inaccessible labile N from a comparison of soil core extractions and soil pore water lysimetry. *Soil Biology and Biochemistry* 73: 22–32.
- Flowers M, Weisz R, and White JG (2005) Yield-based management zones and grid sampling strategies, describing soil test and nutrient variability. *Agronomy Journal* 97: 968–982.
- Fridgen JJ, Kitchen NR, Sudduth KA, Drummond ST, Wiebold WJ, and Fraisse CW (2004) Management Zone Analyst (MZA). Software for subfield management zone delineation. *Agronomy Journal* 96: 1100–1108.
- James DW and Wells KE (1990) Soil sampling collection and handling: Technique based on source and degree of field variability. In: Westerman RL (ed.) *Soil Testing and Plant Analysis*, 3rd edn, pp. 25–44. Madison, WI: SSSA Inc.
- Khan S, Mulvaney RL, and Hoeft RG (2001) A simple soil test for detecting sites that are nonresponsive to nitrogen fertilization. *Soil Science Society of America Journal* 65: 1751–1760.
- Kitchen NR, Drummond ST, Lund ED, Sudduth KA, and Buchleiter GW (2003) Soil electrical conductivity and topography related to yield for three contrasting soil-crop systems. *Agronomy Journal* 95: 483–495.
- Klatt JG, Mallarino AP, Downing JA, Kopaska JA, and Wittry DJ (2003) Soil phosphorus, management practices, and their relationship to phosphorus delivery in the Iowa Clear Lake agricultural watershed. *Journal of Environmental Quality* 32: 2140–2149.
- Kovar JL and Pierzinsky GM (2009) *Methods of phosphorus analysis for soils, sediments, residuals, and waters*, 2nd edn. Southern Cooperative Series Bulletin No. 408. <https://sera17.wordpress.ncsu.edu/publications/>.
- Laboski CAM, Sawyer JE, Walters DT, Bundy LG, Hoeft RG, Randall GW, and Andraski TW (2008) Evaluation of the Illinois soil nitrogen test in the north central region of the United States. *Agronomy Journal* 100: 1070–1076.
- Lawrence PG, Roper W, Morris TF, and Guillard K (2019) Guiding soil sampling strategies using classical and spatial statistics: A review. *Agronomy Journal* 112: 493–510.
- Maguire RO and Sims TS (2002) Soil testing to predict phosphorus leaching. *Journal of Environmental Quality* 31: 1601–1609.
- Mallarino AP and Wittry DJ (2004) Efficacy of grid and zone soil sampling approaches for site-specific assessment of phosphorus, potassium, pH, and organic matter. *Precision Agriculture* 5: 131–144.
- Mallarino AP, Stewart BM, Baker JL, Downing JA, and Sawyer JE (2002) Phosphorus indexing for cropland: Overview and basic concepts of the Iowa phosphorus index. *Journal of Soil and Water Conservation* 57: 440–447.
- Mallarino AP, Atia AM, and A. M. (2005) Correlation of a resin membrane soil phosphorus test with corn yield and routine soil tests. *Soil Science Society of America Journal* 69: 266–272.
- Mallarino AP, Beegle DB, and Joern BC (2007) Soil sampling methods for phosphorus - Spatial Concerns - A position paper. Minimizing Phosphorus Losses from Agriculture/Pennock, Information Exchange Group (SERA-17/IEG) NIFA (USDA) committee. Available at <https://sera17.wordpress.ncsu.edu/publications>.
- McBride MB (2021) Estimating soil chemical properties by diffuse reflectance spectroscopy, Promise versus reality. *European Journal of Soil Science* 73(1): e13192.
- Mueller TG, Pierce FJ, Schbenberger O, and Warncke DD (2001) Map quality for site-specific fertility management. *Soil Science Society of America Journal* 65: 1547–1558.
- Mulvaney RL, Khan SA, Hoeft RG, and Brown HM (2001) A soil organic nitrogen fraction that reduces the need for nitrogen fertilization. *Soil Science Society of America Journal* 65: 1164–1172.
- Pennock D, Yates T, and Braidek J (2007) Soil sampling designs. In: Carter MR and Gregorich EG (eds.) *Soil Sampling and Methods of Analysis*, 2nd edn, pp. 1–14. Boca Raton, FL: CRC Press.
- Sawchik J and Mallarino AP (2008) Variability of soil properties, early phosphorus and potassium uptake, and incidence of pests and weeds in relation to soybean grain yield. *Agronomy Journal* 100: 1450–1462.
- Sharpley AN, Daniel T, Sims T, Lemunyon J, Stevens R, and Parry R (2003) Agricultural phosphorus and eutrophication. In: *ARS-149*, 2nd edn. Washington, D.C.: USDA/ARS.
- Vadas PA, Mallarino AP, and McFarland A (2005) The importance of sampling depth when testing soils for their potential to supply phosphorus to surface runoff. Position paper. Minimizing Phosphorus Losses from Agriculture/Information Exchange Group (SERA-17/IEG) NIFA (USDA) committee. <https://sera17.wordpress.ncsu.edu/publications>.
- Zhu Y, Chen Y, Ali MA, Dong L, Wang X, Archontoulis SV, Schnable JC, and Castellano MJ (2021) Continuous in situ soil nitrate sensors, The importance of high-resolution measurements across time and a comparison with salt extraction-based methods. *Soil Science Society of America Journal* 85: 677–690.

Metal extraction and analysis

Guo Wang^a and Honghong Li^b, ^aCollege of resources and environment, Fujian Agriculture and Forestry University, Fuzhou, Fujian Province, PR China; ^bSchool of History and Geography, Minnan Normal University, Zhangzhou, Fujian Province, PR China

© 2023 Elsevier Ltd. All rights reserved.

Introduction	51
Metal extraction methods	51
Single extraction	51
Ligands	51
Salts	51
Diluted mineral acid	52
Low molecular weight organic acids (LMWOA)	52
Complex extraction	53
Sequential extraction	53
The Tessier procedure	53
The BCR procedure	54
The diffusive gradients in thin films (DGT)	55
Determination of metals in the extracted solutions	55
Concluding remarks	56
References	57

Abstract

Total content of metals in the soil does not correlate well with the uptake by plants. Therefore, extraction methods have been developed to estimate the so-called bioavailable fractions, including single extraction, sequential extraction and DGT (diffusive gradients in thin films). Each extraction method has its advantages and pitfalls, and is suitable for different situations. However, no extraction method can extract the 'true bioavailable' fraction of metals in the soil because of the complexity of binding forces between the metal ions and soil components. At best, chemically extractible metal content can be a good estimate of the bioavailable content of the metal.

Key points

- Metal extraction methods can provide estimates of the bioavailable content of metals in the soil to a certain extent.
- The 'true bioavailable' fraction of the metals cannot be measured, because uptake and accumulation of the metals are influenced by numerous factors related to the soil, plants and environment.
- Chemical extraction of the metals may be single step or sequential.

Abbreviations

BCR	Community Bureau of Reference
DGT	Diffusive gradients in thin films
DTPA	Diethylene triamine pentaacetic acid
EDTA	Ethylenediamine tetraacetic acid
FAAS	Flame atomic absorption spectrophotometer
GFAAS	Graphite furnace atomic absorption spectrophotometer
ICP-MS	Inductively coupled plasma mass spectrometry technique
LMWOA	Low molecular weight organic acids
SEP	Sequential extraction protocol
SMTP	Standards, Measurements and Testing Program
TCLP	Toxicity Characteristic Leaching Procedure
TEA	Triethylamine

Introduction

Among all soil contaminants, potentially toxic metal(loid)s, often referred to as heavy metals (Duffus, 2002), have become increasingly the focus of urgent environmental concern worldwide. Metal(loid)s in soils are bound with various soil components (such as minerals, organic matter, iron oxides, manganese oxides, etc.) involving different binding forces (such as electrostatic attraction, covalent bond, Van der Waals force, etc.). Consequently metal(loid)s exist in numerous forms in soils, referred to as fractions, with different activity and biological availability. The metals in some fractions, such as water-soluble and exchangeable fractions, are highly bioavailable to plants, but others (such as residual fractions) are much less available. Toxic metal(loid)s with strong bioavailability threaten biota more directly than do the other fractions. Numerous methods have been developed to assess the bioavailability of metals in the soil. Although there is still no unanimity in the adoption of a single method for assessing soil contamination by metals, some metal extraction methods are widely used worldwide.

Metal extraction methods

Chemical extraction is used for investigating the fractions of potentially toxic metals and other elements in the soil. There are two main groups of metal extraction methods, single extraction and sequential extraction; both are used to evaluate the availability of metals in soils.

Single extraction

Single extraction involves a single extraction reagent (generally a ligand, diluted acid or salts) or a compound extraction reagent. These tests are usually conducted to study the eco-toxicity and the bioavailability of metals or other elements in the soils to estimate the potential phytotoxic effects to plants and the nutrient supplying capacity of the metals in soils. The extraction reagents can be classified into several groups, such as ligands, neutral salts, diluted mineral acids and organic acids, etc., based on their chemical properties.

Ligands

The diethylene triamine pentaacetic acid (DTPA) and ethylenediamine tetraacetic acid (EDTA) are two ligands widely used in soils to complex with readily available metals, such as copper (Cu), cadmium (Cd), lead (Pb) and zinc (Zn), thereby removing them from the soil organo-mineral complex.

Diethylene triamine pentaacetic acid, DTPA

The DTPA soil test was developed by Lindsay and Norvell in 1967 to extract the available Zn, iron (Fe), manganese (Mn), or Cu in near-neutral and calcareous soils (Lindsay and Norvell, 1978). This method is now widely used in many soils. The extraction solution consists of 0.005 M DTPA, 0.01 M CaCl_2 , and 0.1 M triethylamine (TEA) at $\text{pH} = 7.3$ (Lindsay and Norvell, 1978). In the extraction solution, DTPA acts as a chelating agent for the metals, triethylamine (TEA) is a buffer (pK_a (negative log of the acid dissociation constant, K_a) = 7.8), Ca^{2+} (calcium) maintains the slight alkalinity with TEA and also replaces metal cations from the soil surface. According to the protocol given by the Standards, Measurements and Testing Program (SMTP, formerly BCR) of the European Commission (1998), the DTPA extraction procedure is as follows: 10 g soil sample is mixed with 20 mL of 0.005 M DTPA solution, and then shaken for 2 h at room temperature ($20 \pm 2^\circ\text{C}$). The DTPA has a strong ability to complex metal cations and to release them into soil solution. The different binding strengths of DTPA with different metal cations largely control the extraction efficiency for these metals. The binding strength between the soil and the metals also affects the extraction efficiency. The DTPA-extraction is suitable for extracting Zn, Fe, Mn, Cu, Pb and Cd from the soil, whereas it is less suitable for extracting nickel (Ni) and chromium (Cr).

Ethylenediamine tetraacetic acid, EDTA

Ethylenediamine tetraacetic acid is widely used as a complexing agent because of its strong complexing ability with metal cations. The EDTA can draw out metals from exchange sites of both organic and inorganic groups, and dissolve calcareous deposits through its complexation with metal ions or Ca^{2+} and Mg^{2+} (magnesium). The 0.05 M EDTA solution ($\text{pH} = 7.0$) is commonly used as an extractant. In 1998, SMTP of the European Commission formulated an EDTA extraction procedure, including preparation of the extractant. According to the SMTP annex, the 0.05 M EDTA extraction solution ($\text{pH} = 7.0$) is prepared by adding 146.12 g of EDTA free acid to 800 mL distilled water and partially dissolving it by stirring in 130 mL of saturated ammonia solution (Quevauviller, 1998). For the extraction, a 5-g soil sample is transferred to an extraction bottle in which 50 mL of 0.05 M EDTA has been added (Sahuquillo et al., 1998). The mixture is shaken for 1 h at room temperature ($20 \pm 2^\circ\text{C}$). The EDTA extracts not only exchangeable fraction of the metals, but also carbonate- and organically-bound fractions. The extraction efficiency of DTPA and EDTA is greater than that of the un-buffered salt solutions because of their strong complexing ability.

Salts

Some salts, such as calcium chloride (CaCl_2), ammonium acetate (NH_4OAc), ammonium nitrate (NH_4NO_3), potassium nitrate (KNO_3), magnesium chloride (MgCl_2), and sodium nitrate (NaNO_3) are used as extractants to evaluate bioavailable metals in soils. In most cases, the extraction solutions of these salts are un-buffered. They are more suitable for predicting the plant available metals than some extractants with stronger extraction capacity and diluted acids. The metals in the fractions extracted with the neutral salts, such as 0.01 M CaCl_2 or 1 M NH_4NO_3 , often show better correlations with plant contents of the metals because of the chemical similarity between the extraction and soil solutions.

Calcium chloride (CaCl_2)

In the 1970s, Mn and aluminum (Al) extracted with 0.01 M CaCl_2 from acidic soils showed strong correlations with their toxicity to plants (Houba et al., 1996). This soil test is applied to extract bioavailable fractions of metals (such as Cd, Ni, Zn) from soils. Other concentrations of CaCl_2 , such as 0.1 M, have also been used to extract metals, which also showed good correlations with their plant contents. The 0.01 M CaCl_2 extraction is both practical and environmentally friendly because this extraction solution matched with the soil solution in terms of the composition and concentration of ions. The metals extracted with 0.01 M CaCl_2 are mainly exchangeable fractions, which are readily bioavailable to plants. The amount of 0.01 M CaCl_2 exchangeable metals is mainly controlled by metal desorption from the soil which is affected by the extraction solution. The electrostatically adsorbed metals can be replaced by Ca^{2+} in the extraction solution, whereas it is generally difficult to desorb chemically adsorbed metals by 0.01 M CaCl_2 . Because 0.01 M CaCl_2 is milder than other extraction solutions, such as EDTA, DTPA or dilute acids, the quantity of metals extracted by 0.01 M CaCl_2 solutions is less than that extracted with complexing solutions or diluted acids.

Ammonium acetate

Ammonium acetate solution (1 M NH_4OAc , pH = 7.0) is commonly used to extract exchangeable cations from the soil (Ure et al., 1993). This solution is usually buffered at pH 7 to avoid the dissolution of carbonate at low pH. For ammonium acetate, NH_4^+ replaces readily exchangeable metals and acetate acts as a ligand to form complexes with the metals. The complexes are then released from the soil into the solution. Specifically sorbed cations, which are less readily exchangeable, are difficult to displace by NH_4^+ but can be replaced by H^+ (hydrogen). Ammonium acetate is also acidified to pH = 5 to dissolve carbonates and release carbonate-bound metals into the soil solution. The large salt concentration (1 M NH_4OAc) in the extracts causes difficulties for determinations using the atomic absorption spectrophotometer or inductively coupled plasma mass spectrometry (ICP-MS). The amount of Cd extracted by the 1 M NH_4OAc (pH = 7.0) procedure is less than that by HCl, EDTA, DTPA, etc.

Diluted mineral acid

The most commonly used diluted acid for metal extraction is diluted hydrochloric acid (HCl). Diluted nitric acid (HNO_3) has a particular advantage because the nitrate anion (NO_3^-) has little complexing capacity for the metal cations. Thus the anion in HNO_3 does not interfere with metal extraction by acid dissolution.

Diluted HCl is one of the most widely used extractants for releasing the acid-soluble fraction of the soil. It can replace exchangeable metal cations by its H^+ ion, dissolve calcareous bound metals through dissolution, and form soluble complexes with metal cations by its Cl^- ion, a strong ligand for metal cations (Leleyter et al., 2012). The 0.1 M HCl is the most commonly used concentration of HCl for metal extraction, although other concentrations such as 0.5 M are also used. It has been reported that bioavailable Pb may be underestimated using a low concentration of HCl (0.1 M) because Pb^{2+} readily precipitates as PbCl_2 (Wei et al., 2005). On the contrary, 0.1 M HCl overestimates bioavailable Cu in the soil because it promotes the release of Cu bound with some Fe or Mn oxides (Yu et al., 2004). The Fe/Mn oxides-bound Cu is not readily available to plants under field conditions, therefore the 0.1 M HCl extractable metal content is generally poorly correlated with the metals taken up by plants. More concentrated solutions of HCl are unsuitable for estimating bioavailable soil metal content because sparingly soluble metals of little significance for plant absorption may be solubilized.

Low molecular weight organic acids (LMWOA)

Low molecular weight organic acids (LMWOA) include several organic acids, such as oxalic, acetic, malic, citric, and succinic acids (Abollino et al., 2011). These acids may be produced by the decomposition of organic matter or plant root exudates. The LMWOA extract metals in soil by a combination of acid dissolution due to their acidity and complexation with the metal cations due to their functional groups, which include carboxyl, phenolic, and amino groups. Ligands of the acids may also substitute the metalloids (such as arsenic) bound with the soil surfaces by ligand exchange. The Fe^{3+} and Mn^{4+} in Fe/Mn oxides may also be complexed by the ligands of LMWOA, resulting in dissolution of the oxides, which leads to the release of the metals bound with the oxides. The gradual disintegration of iron oxides caused by LMWOA may play an important role in releasing arsenic (As). The LMWOA are weaker metal extracting agents than complexing reagents. Furthermore, the extraction capacity differs between acids. Oxalic acid has a greater capacity to mobilize Cu, Zn, Pb, As, Fe and Mn than does acetic acid. The reductive dissolution of LMWOA also plays an important role in the dissolution of iron oxides in the presence of oxalic acid.

Complex extraction

A complex extraction solution, designed to simulate the soil rhizosphere, has been developed to determine the bioavailability of trace metals. The extraction solution based on a mixture of EDTA and NH_4OAc was originally proposed by Lakanen and Ervio (1971). The complexation of EDTA and acetic acid is considered to simulate the complexation behavior of root exudates, and NH_4^+ (ammonium) may desorb the exchangeable soil fraction (Meers et al., 2007).

The Toxicity Characteristic Leaching Procedure (TCLP) is the US EPA's method for testing toxicity of metals in the soil, which involves extraction with diluted acetic acid and sodium hydroxide (Abollino et al., 2011). It is often used as an operational procedure to evaluate the leaching of metals before the reuse of contaminated land. However, some studies have raised the potential limitations of TCLP in evaluating cementitious wastes because of the high final pH of the leaching solution, therefore, the release of metals may be hindered (Halim, 2003; Lu et al., 2019; Malviya and Chaudhary, 2006).

Sequential extraction

Sequential extraction combines several continuous extraction steps and is commonly used to assess the mobility, availability, distribution and potential toxicity of metals in soils, sediments, and other solid materials (Zimmerman and Weindorf, 2010). The first sequential extraction procedure on metal fractions was proposed by Tessier in 1979. In 1987, the Community Bureau of Reference (BCR), Europe Commission, proposed a new sequential extraction procedure, which was a significant advance in sequential extraction following the Tessier procedure. The BCR extraction procedure aimed to harmonize the existing extraction procedures. In 1993 a standardized three-step sequential extraction protocol (SEP) was proposed, and is known as BCR SEP. The theoretical basis for SEP is that the metals are adsorbed or bound with solid surfaces through various binding forces, and can be extracted in the order of their solubility using different extracting agents and under different conditions. The most soluble fraction is the first to be removed from the soil.

The Tessier procedure

Tessier's protocol divides soil metals into five operationally defined fractions in the order of extraction: exchangeable, carbonate bound, Fe and Mn oxide bound, organic matter bound, and residual, respectively (Tessier et al., 1979). The Tessier procedure is the most widely used extraction method for the fractionation of metals in the soil. The exchangeable fraction is contributed by the metals associated with soil particle surfaces by electrostatic mechanisms. The metals in this fraction are easily exchanged with other cations and easily absorbed by plant roots. This fraction is extracted with MgCl_2 (pH = 7.0) solution. One gram of a soil sample is mixed with 8 mL 1 M MgCl_2 (pH = 7.0) in a 50-mL centrifugation tube. The mixture is agitated at room temperature for 1 h. After agitation, the supernatant and the soil residue are separated by centrifugation. The carbonate-bound fraction is extracted with NaOAc solution, whereby 8 mL of 1 M NaOAc (pH = 5.0) is mixed with the soil residue of the exchangeable fraction. The mixture is shaken for 5 h at room temperature, and centrifuged to separate the supernatant and soil residue. Metals bound to Fe and Mn oxides are particularly susceptible to reducing conditions. The Fe—Mn oxides bound fraction is extracted with 0.04 M $\text{NH}_2\text{OH—HCl}$ in 25% (v/v) HOAc (pH = 2) at 96 °C for 6 h with occasional agitation. The organic-bound fraction is extracted with 0.02 M $\text{HNO}_3 + \text{H}_2\text{O}_2$. The soil residue of Fe—Mn oxides bound fraction is mixed with 3 mL of 0.02 M $\text{HNO}_3 + 5$ mL of 30% (m/v) H_2O_2 (pH = 2). The mixture is heated to 85 ± 2 °C for 2 h with intermittent agitation. A second 3-mL aliquot of 30% H_2O_2 (pH = 2 with HNO_3) is then added and the sample heated again to 85 ± 2 °C for 3 h with intermittent agitation. After cooling, 5 mL of 3.2 M NH_4OAc in 20% (v/v) HNO_3 is added to the tube, and the mixture diluted to 20 mL and agitated for another 30 min. The addition of NH_4OAc is to prevent adsorption of extracted metals onto the soil. In the Tessier extraction procedure, after agitation, the mixture is centrifuged to separate the supernatant and the soil residue. The residue from the organic-bound fraction is digested with an HF—HClO_4 mixture. Since the introduction of the Tessier procedure, some modifications have been proposed (Fig. 1).

Although the Tessier SEP has been widely used, there are some pitfalls in the procedure. The first is selectivity of the extracting reagents. The fractions are given names that strongly suggest their chemical form in soil. However, the metals extracted by the various extracting solutions do not represent exactly the fraction bound with the specific components. In the Tessier procedure, MgCl_2 is used to exchange the metal cations adsorbed on soil surfaces with Mg^{2+} . However, more metal cations would be released due to the formation of metal–chloride complexes, displacing the equilibrium in favor of soluble forms. The 1 M NaOAc (pH = 5.0) has been found to dissolve portlandite ($\text{Ca}[\text{OH}]_2$) and brucite ($\text{Mg}_3[\text{OH}]_6$). When the residue of the exchangeable fraction is extracted with 1 M NaOAc (pH = 5.0), a decrease of 2 pH units occurs. This reduction of 2 pH units covers the pH range of the adsorption edge of the majority of the metals, resulting in an overestimate of carbonate-bound metals. The pH range (5–6) is conducive to Mn oxyhydroxide dissolution, so the addition of NaOAc (pH = 5.0) may dissolve manganous and ferrous oxides, leading to metal release from components other than carbonates. Hydrogen peroxide (H_2O_2) is used for the organic-bound fraction. It was found that it only partially dissolved organic materials and was not selective enough. The H_2O_2 solution (pH = 2) also dissolves manganese oxyhydroxides, amorphous iron oxyhydroxides, and sulfides, resulting in an overestimate of the organic-bound fraction (Beckett, 1989; Hass and Fine, 2010).

Another problem is redistribution of the extracted metals. When the exchangeable fraction is extracted, the metals from exchange sites may be adsorbed to specific sorption sites. Soil homogenization (by grinding and sieving), and extraction procedure harmonization (i.e., shaking) can promote specific adsorption of some electrostatically adsorbed metals (Hass and Fine, 2010).

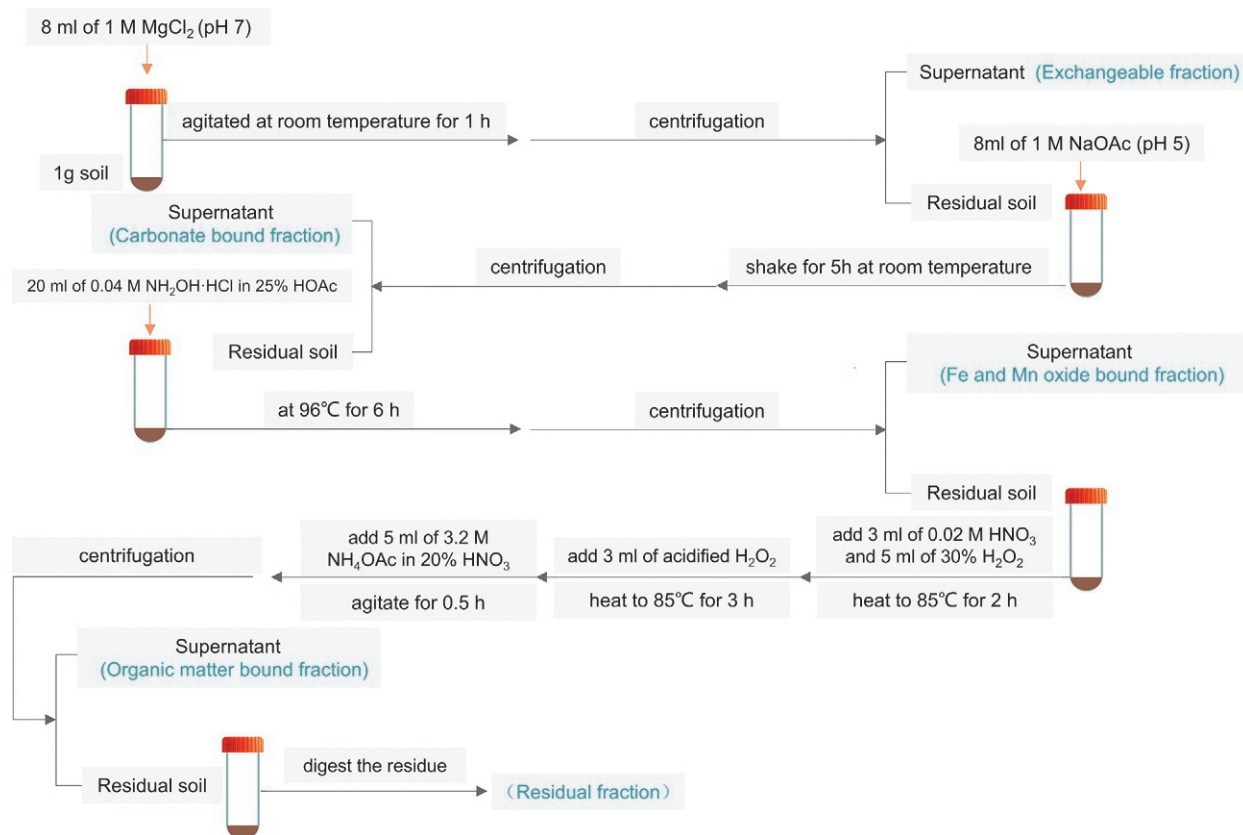


Fig. 1 Tessier sequential extraction procedures used for operational speciation of metals.

Neutral salts that are used to extract the exchangeable fraction result in coagulation and precipitation of organo-metal complexes, leading to an underestimate of the soluble component. During Mn—Fe oxyhydroxide extraction using acid ammonium oxalate, oxalate-metal precipitates may form, resulting in underestimates of the metals in this fraction. Some studies on the Tessier procedure showed that metals in exchangeable, carbonate-bound, and Fe and Mn oxide bound fractions would be underestimated while that for the organic-bound and residual fractions would be overestimated, because metals extracted from the first three fractions were redistributed to the following two fractions.

The above limitations lead to ambiguity of the meaning of the fractions of metals, which makes it difficult to explain them. It must be clear that these fractions are operationally-defined rather than strictly scientifically defined. The complexity and time-consuming aspects of the extraction procedures are also a limitation, which is why it is difficult to compare the results from different laboratories.

The BCR procedure

In 1987, significant changes took place in sequential extraction when the European Commission launched a program within the framework of BCR aimed at harmonizing extraction procedures. In 1993 a standardized three-step SEP (BCR[®] SEP) was proposed (Sutherland, 2010; Ure et al., 1993). In this BCR[®] SEP, metals in the soil are divided into three major fractions: acid extractable, reducible, and oxidizable. The BCR[®] extraction procedure is as follows.

- Step 1. Acetic acid (40 mL 0.11 M) is added to 1 g of soil in a 100 mL centrifuge tube, shaken end-over-end (30 ± 10 rpm) immediately for 16 h at ambient temperature (overnight). The extract and solid residue are separated by centrifugation, and the supernatant solution decanted and stored at 4°C prior to analysis. The residue is washed with water before the next step (20 mL of distilled water, shaken for 15 min then centrifuged).
- Step 2. Hydroxylamine solution (40 mL, 0.5 M $\text{NH}_2\text{OH}\cdot\text{HCl}$ containing 2.5% by volume 2 M HNO_3) is added to the residue, shaken immediately for 16 h at ambient temperature, and centrifuged, decanted and washed as for step 1.
- Step 3. Peroxide solution (10 mL H_2O_2 8.8 M) is added to the residue, heated to $85 \pm 2^\circ\text{C}$ in a water bath and digested for 1 h with occasional manual shaking. The volume is allowed to reduce to <3 mL and a further aliquot of 10 mL H_2O_2 is added and the digestion repeated for 1 h at $85 \pm 2^\circ\text{C}$, and the volume again allowed to reduce to a few mL, then cooled. Finally 50 mL of 1 M

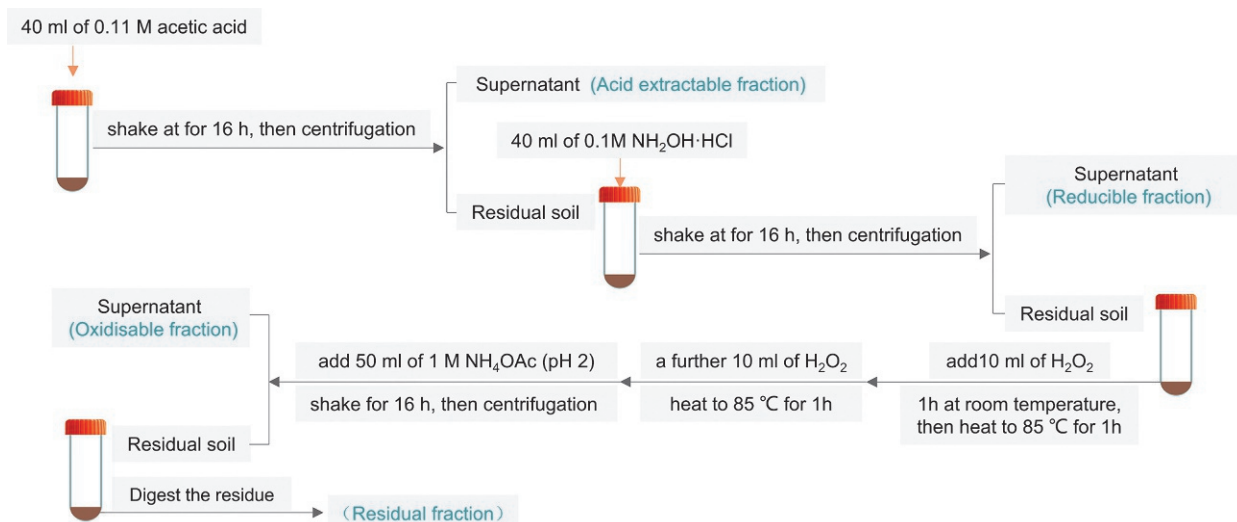


Fig. 2 The modified BCR sequential extraction scheme used for operational speciation of metals.

NH_4OAc (pH 2) solution is added to the residue and the suspension immediately shaken for 16 h at ambient temperature (overnight). Solution and residue are separated by centrifugation and decantation as for step 1 (Fig. 2).

The BCR[®] SEP has been widely used over the last decade since its introduction. However, some suggestions for modification have been proposed. One of the proposals is to add a step to determine the residual fraction. It has been suggested that the residual soil of Step 3 should be digested to determine the total metal content in the residual fraction. This allows the total content to be calculated and thus the metal recovery of each step to be calculated. However, the residual fraction is not an official part of the BCR[®] SEP.

The BCR[®] SEP is similar to the Tessier SEP with some differences. There are the same drawbacks for BCR[®] SEP as for the Tessier SEP. The chief difference is in the first fraction. The BCR[®] SEP does not evaluate the exchangeable fraction and carbonate-bound fraction separately, but combines the two into one fraction, that is, the acid extractable fraction, which leads to more ambiguity in the meaning of the BCR[®] fractions than for the Tessier procedure. The other drawback is the lack of determination of the residual fraction, which prevents any calculation of the recovery of each fraction with respect to the total.

The diffusive gradients in thin films (DGT)

The diffusive gradients in thin films is an in situ device for measuring the fluxes of labile components in water, soil and sediments under undisturbed conditions. The DGT technique was pioneered by William Davison and Hao Zhang of Lancaster University, United Kingdom, in 1993 (Davison and Zhang, 1994). The DGT is a kinetic passive sampler for quantifying trace metals (Guan, 2019). It was initially used to evaluate the concentration of metals in water, and was soon extended to sediments and soils. The DGT device usually consists of a filter membrane, a diffusive gel layer, a binding gel layer and a plastic base and cap. The DGT sampler can be either a piston-type (Fig. 3A) or flat-type (Fig. 3B).

When a DGT sampler is installed in soil slurry (wetted soil), the mobile species of metals from the pore water diffuse through the diffusion layer to the binding layer (chelex gel). The continual removal of metals from the soil pore water to the binding gel induces a concentration gradient within the DGT diffusion layer and immediately adjacent to the soil (Guan, 2019). Unlike extracting reagents, DGT evaluates kinetic processes and the dynamic availability of metals in soils. The DGT has been used to determine the bioavailable metals and phosphorus in soils. It is also widely used to study processes in the rhizosphere by simulating the uptake of metals. Research has shown that DGT can be used to predict plant uptake or the bioavailability of pollutants. The diffusive gradients in thin films is also useful for the selective determination of specific species. When DGT incorporates a high-selectivity agent into the binding gel, it preferentially binds the specific species while excluding other species. It is mainly used to measure trace metals in the medium, whereas its application to organic chemicals is limited. When the DGT sampler is incorporated with other binding agents, such as XAD resin (a macroporous resin) or activated charcoal, it can be used to measure trace organic compounds.

Determination of metals in the extracted solutions

The analytical techniques for determining metals have evolved over time. The most common techniques used for the measurement of metals such as Cd, Pb, Cu, Zn, etc. has traditionally been atomic absorption spectrophotometry, including flame atomic absorption spectrophotometer (FAAS) and graphite furnace atomic absorption spectrophotometer (GFAAS). The FAAS is used for the determination of metals with normal solution concentrations, whereas GFAAS is used for the measurement of metals with small

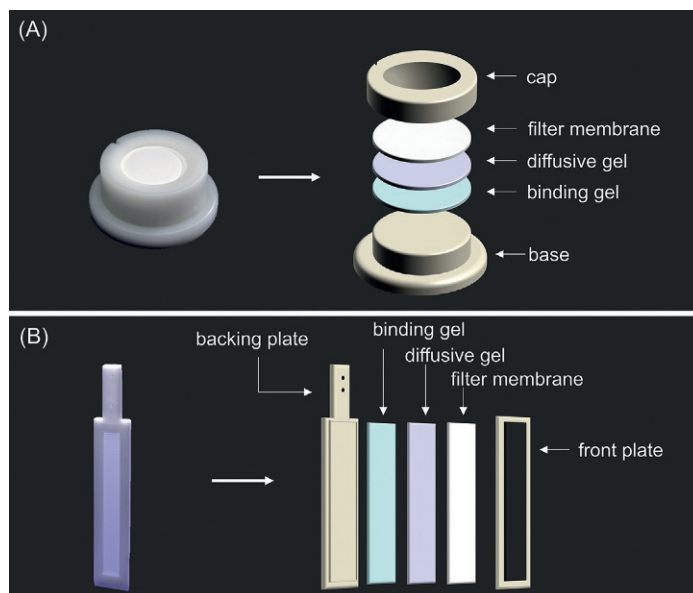


Fig. 3 Simplified setting schematic drawing of a piston-type (A) and a flat-type (B) DGT sampler.

or trace concentration. Atomic fluorescence spectrophotometry is preferred for the determination of the metalloids (As, Hg (mercury), etc.) while cold atomic absorption spectrometry is used to determine Hg.

The ICP-MS was developed more recently than the above instruments. This instrument has remarkable advantages. It can measure several elements in solution simultaneously with good accuracy and cover a wide concentration range. This instrument is rapidly gaining popularity and is gradually replacing the above instruments.

Concluding remarks

It is well accepted that the total content of metals in the soil does not correlate well with the uptake by plants, whereas there is a better correlation with the more easily extractable, so-called bioavailable fractions. Plants absorb only some parts of the fraction of metals in the soil. All the existing extraction methods aim to extract the 'true bioavailable' fraction of the metals in the soil. Unfortunately, this is unattainable because of the complexity of binding forces between the metal ions and the soil components. In addition, the type and species of plant, the growth and nutrient status, root morphology, and environmental conditions (temperature, humidity, wind, etc.) greatly influence the plant uptake of the metals. The soil's supply of metals is not the unique determinant for plant uptake of metals. The extraction methods developed include different extracting reagents and materials under different extracting conditions, static or dynamic, single or sequential extraction. However, it is impossible to extract the real bioavailable metal content in the soil. At best, chemically extractable metal can be a good estimate of the bioavailable content of the metal.

The selection of extracting reagents is the most important factor. The existing extracting reagents are predominantly water, diluted acid, diluted alkali, neutral salts, chelating agents, and surfactants. Water-soluble metals are the fraction that is the most easily absorbed by the plant. However, the water-soluble metals are always at small or trace concentration, and are very sensitive to external conditions (temperature, moisture, wind, growth position of plant, fertilization, agricultural management, etc.). It is not a stable index. Among the extracting solutions of chemical reagents, the diluted solution of neutral salts seems to be the most similar to the soil environment around the roots. The other chemical reagents are too strong to varying degrees to simulate the soil environment around the roots.

The sequential extraction procedures (SEPs) provide useful tools to probe the distribution of metals in a soil. However, the clear distribution of the metals in soil is still difficult to determine because the fractions extracted by the SEPs are operationally defined and do not correspond exactly to the forms suggested by the terms used to describe them. It is probably not realistic to presume that precise fractions can be extracted chemically. The other problems with the SEPs, such as poor selectivity and redistribution, influence the quality of the extraction. There are some suggestions for modification of the Tessier SEP. However, there is still no formal revised version of the Tessier SEP, but the BCR[®] SEP is an improvement on it. The BCR[®] SEP has better reproducibility than the Tessier SEP, partly because certified reference soils have been defined. However, the BCR SEP is still not able to extract the real fractions indicated by their names.

The DGT is essentially a single extraction method but it is quite different from single extractions. It determines dynamically the supply capacity of the soil metal content, and measures the mobile and labile species rather than other inert species of the metals. The DGT shows good predictability for the bioavailability of the metals in the soil.

It should be pointed out that both the conditions of the soil and crop control the uptake, transport of a metal from the soil to the crop, and final accumulation in the edible part. Sometimes, transport of the metal within the crop is more important than the supply capacity of the metal in the soil. The uptake and transport of metals in crops differs from crop to crop, and is influenced by many environmental factors. It is difficult to develop a method that predicts perfectly the accumulation of metals in the soil, especially when a large area (such as a province, a country) is being dealt with.

References

- Abollino O, Malandrino M, Giacomino A, and Mentasti E (2011) The role of chemometrics in single and sequential extraction assays: A review: Part I. Extraction procedures, uni- and bivariate techniques and multivariate variable reduction techniques for pattern recognition. *Analytica Chimica Acta* 688(2): 104–121.
- Beckett PHT (1989) *The Use of Extractants in Studies on Trace Metals in Soils, Sewage Sludges, and Sludge-Treated Soils*. Springer US.
- Davison W and Zhang H (1994) In situ speciation measurements of trace components in natural waters using thin-film gels. *Letters to Nature* 367: 546–548.
- Duffus JH (2002) "Heavy Metals"—A meaningless term? (IUPAC Technical Report). *Pure and Applied Chemistry* 74(5): 793–807.
- Guan DX (2019) Diffusive gradients in thin-films (DGT): An effective and simple tool for assessing contaminant bioavailability in waters, soils and sediments. In: *Encyclopedia of Environmental Health*, pp. 111–124. Elsevier.
- Halim C (2003) Evaluating the applicability of a modified toxicity characteristic leaching procedure (TCLP) for the classification of cementitious wastes containing lead and cadmium. *Journal of Hazardous Materials* 103(1–2): 125–140.
- Hass A and Fine P (2010) Sequential selective extraction procedures for the study of heavy metals in soils, sediments, and waste materials—A critical review. *Critical Reviews in Environmental Science and Technology* 40(5): 365–399.
- Houba VJG, Lexmond TM, Novozamsky I, and Lee JJVD (1996) State of the art and future developments in soil analysis for bioavailability assessment. *Science of the Total Environment* 178: 21–28.
- Lakanen E and Ervio R (1971) A comparison of eight extractants for the determination of plant available micronutrients in soils. *Acta Agricultura Fennica* 123: 223–232.
- Leleyter L, Rousseau C, Biree L, and Baraud F (2012) Comparison of EDTA, HCl and sequential extraction procedures, for selected metals (Cu, Mn, Pb, Zn), in soils, riverine and marine sediments. *Journal of Geochemical Exploration* 116–117: 51–59.
- Lindsay WL and Norvell WA (1978) Development of a DTPA soil test for zinc, iron, manganese, and copper. *Soil Science Society of America Journal* 42: 421–428.
- Lu CC, Hsu MH, and Lin YP (2019) Evaluation of heavy metal leachability of incinerating recycled aggregate and solidification/stabilization products for construction reuse using TCLP, multi-final pH and EDTA-mediated TCLP leaching tests. *Journal of Hazardous Materials* 368: 336–344.
- Malviya R and Chaudhary R (2006) Leaching behavior and immobilization of heavy metals in solidified/stabilized products. *Journal of Hazardous Materials* 137(1): 207–217.
- Meers E, Laing DG, Unamuno V, Ruttens A, Vangronsveld J, Tack FMG, and Verloo MG (2007) Comparison of cadmium extractability from soils by commonly used single extraction protocols. *Geoderma* 141(3–4): 247–259.
- Quevauviller P (1998) Operationally defined extraction procedures for soil and sediment analysis. *TRAC Trends in Analytical Chemistry* 17(5): 289–298.
- Sahuquillo A, Bosch H, Rauret G, and Muntau H (1998) Certified reference materials for extractable trace metals in soils: Effect of the particle size. *Fresenius Journal of Analytical Chemistry* 360: 304–307.
- Sutherland RA (2010) BCR(R)-701: A review of 10-years of sequential extraction analyses. *Analytica Chimica Acta* 680(1–2): 10–20.
- Tessier A, Campbell PGC, and Bisson M (1979) Sequential extraction procedure for the speciation of particulate trace metals. *Analytical Chemistry* 51: 844–851.
- Ure M, Thomas R, and Littlejohn D (1993) Ammonium acetate extracts and their analysis for the speciation of metal ions in soils and sediments. *International Journal of Environmental Analytical Chemistry* 51(1–4): 65–84.
- Wei YL, Yang YW, and Lee JF (2005) Lead speciation in 0.1N HCl-extracted residue of analog of Pb-contaminated soil. *Journal of Electron Spectroscopy and Related Phenomena* 144–147: 299–301.
- Yu S, He ZL, Huang CY, Chen GC, and Calvert DV (2004) Copper fractionation and extractability in two contaminated variable charge soils. *Geoderma* 123(1–2): 163–175.
- Zimmerman AJ and Weindorf DC (2010) Heavy metal and trace metal analysis in soil by sequential extraction: A review of procedures. *International Journal of Analytical Chemistry* 2010: 387803.

Soil organic matter: Composition

Markus Kleber, Department of Crop and Soil Science, Oregon State University, Corvallis, OR, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	58
Organic and inorganic carbon	58
Soil organic matter (SOM) and humus	58
Organismic carbon in soil	60
Xenobiotic organic compounds	64
Thermally altered C	65
Conclusion and outlook	66
References	67

Abstract

This chapter provides an overview of the relevant chemical features of organic carbon in soil. To this end, the text presents three major, functionally and molecularly discernible categories of organic carbon compounds in soil: (1) Organismic, (2) Xenobiotic and (3) Pyrogenic—thermally altered carbon forms. A brief appraisal of the functional relevance of these carbon forms is provided. The definition of “Soil organic matter” continues to evolve, and it is increasingly conceptualized as a dynamic system component rather than as a particular molecular phase. Increasing amounts of organic compounds have been added to the environment over the past 60 years to which natural systems are not adapted.

Key points

- The definition of “Soil organic matter” keeps evolving.
- Soil organic matter is increasingly conceptualized as a dynamic system component rather than as a certain molecular phase.
- The past 60 years have added increasing amounts of organic compounds to which natural systems are not adapted.
- The macro- and micro-scopically structureless organic mass of the soil, also called Humus, encompasses both floral (i.e. plant derived) and microbial necromass at some degree of degradation.

Introduction

To describe the composition of soil organic matter (SOM), it is necessary to define the term. This is not as straightforward a task as one would initially think. To understand the underlying conceptual challenge better, it is useful first to examine the basic convention used to distinguish between major forms of carbon that may occur in soils—the distinction between inorganic C and organic C (Fig. 1).

Organic and inorganic carbon

Fig. 1 shows the long-standing practice to separate the so called “organic carbon” from other carbon containing compounds. The notion that there is such a thing as “organic carbon” dates back to the early days of chemistry as a scientific discipline. During those days, early researchers wanted to recognize such carbon containing materials that were exclusively produced by organisms. But when Wöhler (1828) synthesized the supposedly organic compound Urea [$\text{CO}(\text{NH}_2)_2$] in his laboratory using inorganic precursors, he coincidentally eliminated the basic premise “formed by a living organism” underlying the production and recognition of organic substances.

Accordingly, the separation of soil carbon into organic and inorganic forms (Fig. 1) is a convention from the past, involving numerous inconsistencies. For instance, carbonate minerals and gaseous carbon dioxide are not typically considered “organic compounds” despite the fact that they are more often than not produced by living organisms. Note also that any analytical procedure that measures total carbon in a soil sample will always include carbon from both inorganic and organic forms.

Soil organic matter (SOM) and humus

Having pointed out the problematic nature of distinguishing between organic and inorganic carbon, we find that similar issues must be confronted when attempting to define “soil organic matter”. Here, the problem lies in the fact that scientists realized early

Carbon in soil

Organic carbon	Inorganic carbon
<i>Historic emphasis:</i> "compounds produced by living organisms". <i>Modern emphasis:</i> "a chemical compound in which one or more atoms of carbon are covalently linked to a hydrogen or to atoms of other elements"	Varying conventions arbitrarily exclude CO_2, carbonates (CO_3^{2-}), carbides (CaC_2), cyanides (CN^-) and others from being considered as organic compounds
1. Organismic: - C in living organisms as well as - C in naturally occurring organic residues at various stages of decay, - soluble and insoluble organic metabolites, - involving all forms of biomolecules such as carbohydrates, lignin, proteins, lipids, low molecular weight organic acids, methane and other volatiles	4. Fully oxidized, highly mobile carbon in the - gas phase / soil air, CO_2; and in the - aqueous phase / soil solution, CO_3^{2-}; HCO_3^-; H_2CO_3; also known as dissolved inorganic carbon (DIC)
2. Xenobiotic: anthropogenic origin, not produced by natural systems, can be separated into - monomeric (often halogenated, such as organic pesticides, fungicides, herbicides, fire fighting foams) and - polymeric species (compounds of high molecular mass, often micro-particulate, anthropogenic origin, such as industrial polymers (plastics and microplastics))	5. Carbonate minerals (Calcite, Dolomite, Siderite, Aragonite etc.) either - formed in situ as secondary precipitate (calcrete, caliche) or - inherited from biogenic sedimentary regolith
3. Pyrogenic carbon: organic matter with a history of thermal alteration in the absence of oxygen (= pyrolysis); can be tentatively divided into materials resulting from	
- low temperature pyrolysis (creating volatiles and viscous, tar like compounds)	- high temperature pyrolysis (creating mineral-resembling carbon microcrystallites)

Fig. 1 Organic versus inorganic forms of carbon in soil. Note that, depending on pyrolysis temperature, thermally altered carbon (Box # 3) can exhibit both 'organic' and 'mineral' features, such as considerable crystallinity. Bold face indicates the carbon forms considered to be part of the soil organic matter continuum: (1) Organismic, (2) Xenobiotic, and (3) Thermally altered, pyrogenic carbon.

on that there are forms of organic matter in soil that differ significantly in functionality. Living organisms, for instance, are composed of organic compounds, however, they contribute in other ways than their residues (which are also organic compounds) to soil functionality, especially when those residues have undergone extended oxidation and fragmentation and are then called 'humus'.

As science progressed from the days when humus was considered the most important fraction of the organic matter in soil, definitions of SOM have changed their emphasis. This change involved two major aspects. There has been a progressive abandonment of the historic trend to equate SOM with just the fraction that satisfies the definition of humus towards an all-inclusive-type view (see "Static view" in Table 1). Over the same time frame, and reflecting the difficulty to constrain SOM as a chemical category, there is now a tendency to define SOM no longer as a chemical substance or mixture thereof, but as a continuous flow of organic carbon compounds through the soil (see "Dynamic View" in Table 1). Note that the static and dynamic approaches illustrated in Table 1 are compatible and not mutually exclusive: the complex and varied mixture of Weil and Brady (2016) can be highly dynamic as proposed by Hedges et al. (2000) and Janzen (2006, 2015).

Table 1 A compilation of relatively recent definitions of “SOM”.

<i>Static, ‘chemical’ view. SOM is:</i>		<i>Dynamic, ‘functional’ view. SOM is:</i>	
Trumbore (1997)	“The natural, nonliving component of organic matter in soil”	Hedges et al. (2000)	“A thermodynamic anomaly atop a free energy precipice that drops off on all sides to dispersed, stable ingredients such as carbon dioxide, water, nitrate and phosphate”
		Janzen (2006)	“A relentless flow of C atoms, through a myriad of streams—some fast, some slow—wending their way through the ecosystem, driving biotic processes along the way”
Baldock and Broos (2011)	“The sum of all naturally derived organic materials”	Janzen (2015)	“A dynamic hub in the energy cycles of our ecosystems”
Weil and Brady (2016)	“A complex and varied mixture of organic substances”		

Definitions either attempt to constrain the chemical characteristics of SOM (‘chemical’ view) or to emphasize its highly dynamic nature (‘functional’ view).

An important corollary of these two trends to categorize organic matter is their incompatibility with the view that the ideal state of organic matter in soil should be that of a ‘stable’, long lasting type of material. This is a view that is deeply ingrained in the scientific community and that has historically been the basis of land management recommendations for practitioners.

As science progresses, more challenges and complications have added to the dilemma of constraining the meaning of the term SOM. At issue here are the growing volumes of non-natural organic compounds deposited to soils worldwide. Examples are organic agrochemicals (pesticides), organic industrial contaminants (such as oil spills or fluorinated firefighting foams) and, increasingly recognized, the industrial polymers commonly called plastics. Once these non-natural compounds get into the soil they become substrates, catalysts and reactants for organisms and must be considered as having the potential to interfere with soil functions. Omitting xenobiotic compounds as part of SOM may not be the best way to address the issue.

We realize that for many in the scientific community, there is a conceptual difference between the totality of organic materials that can be found in soil (“organic matter in soil”, including anything that satisfies the definition of an organic compound, such as living organisms, charcoal, microplastics and even xenobiotics) and the subset of these materials that is “the macro- and microscopically structureless organic mass of the soil” (Pauli, 1961) or humus. It is impossible to ignore or evade the widespread informal understanding among practitioners and scientists that the label “SOM” should refer primarily to the insoluble amorphous organic residues (traditionally: humus, also: necromass when of microbial origin) that result from the processing of organic debris by the decomposer community. The combined humus and necromass subfraction often constitutes the largest proportion by mass of total soil carbon and closely relates to soil fertility and other vital soil functions, making it a worthy object of investigation. Fig. 2 illustrates the material categories that produce necromass and humus, including the organisms that actively process organic residues, any molecular fragments that have sufficient polarity to be dissolved or are small enough to be suspended in the soil solution, and organismic detritus as the ultimate source of organic inputs.

Organismic carbon in soil

The chemistry of bulk plant biomass inputs into soils has been comprehensively discussed by Oades (1989) and Kögel-Knabner (2002, 2017) and includes a wide range of compounds such as cellulose, hemicellulose, lignin, protein, DNA, RNA, lipids, waxes, resins and an unlimited number of composites of the above (Fig. 2).

A major insight from these publications is that about 80% of dry plant biomass is macromolecular and or composed of nonpolar organic compounds, and for that reason, is initially not soluble in water (neither plants nor animals dissolve in the rain). Although rarely fully saturated even in humid climates, soils are fundamentally aqueous systems, meaning that their functionality relies on the presence of water as a solvent and means of transport. Accordingly, before further decomposition, organismic debris needs to undergo pretreatment to increase water solubility, and thus mobility, in the soil system. Such pretreatment is also needed to reduce biomacromolecular compounds to molecular sizes small enough for passage through microbial cell walls, because final conversion of biomolecular precursor fragments to the highly oxidized end products CO₂ (carbon dioxide) and HCO₃⁻ (bicarbonate) occurs inside the cells of decomposer organisms. This pretreatment involves incremental breakdown of biomacromolecules by microbial exoenzymes, decreasing the size of primary plant inputs, increasing the polarity of the hydrolyzed or oxidized intermediates, and thus increasing organic matter (OM) solubility in the polar solvent H₂O.

It is necessary here to recognize a process that was not well understood in the early days of SOM research. That is, the carbon, nutrients and energy supplied by primary plant inputs will, after processing and assimilation by decomposer organisms, contribute to the de novo synthesis of bacterial and fungal biomass, primarily generating proteins and cell wall lipids. The efficiency (microbial carbon use efficiency, CUE) at which this occurs appears to be a function of the nominal oxidation state of the carbon in a given compound, moderated by soil temperature regime and nutrient availability.

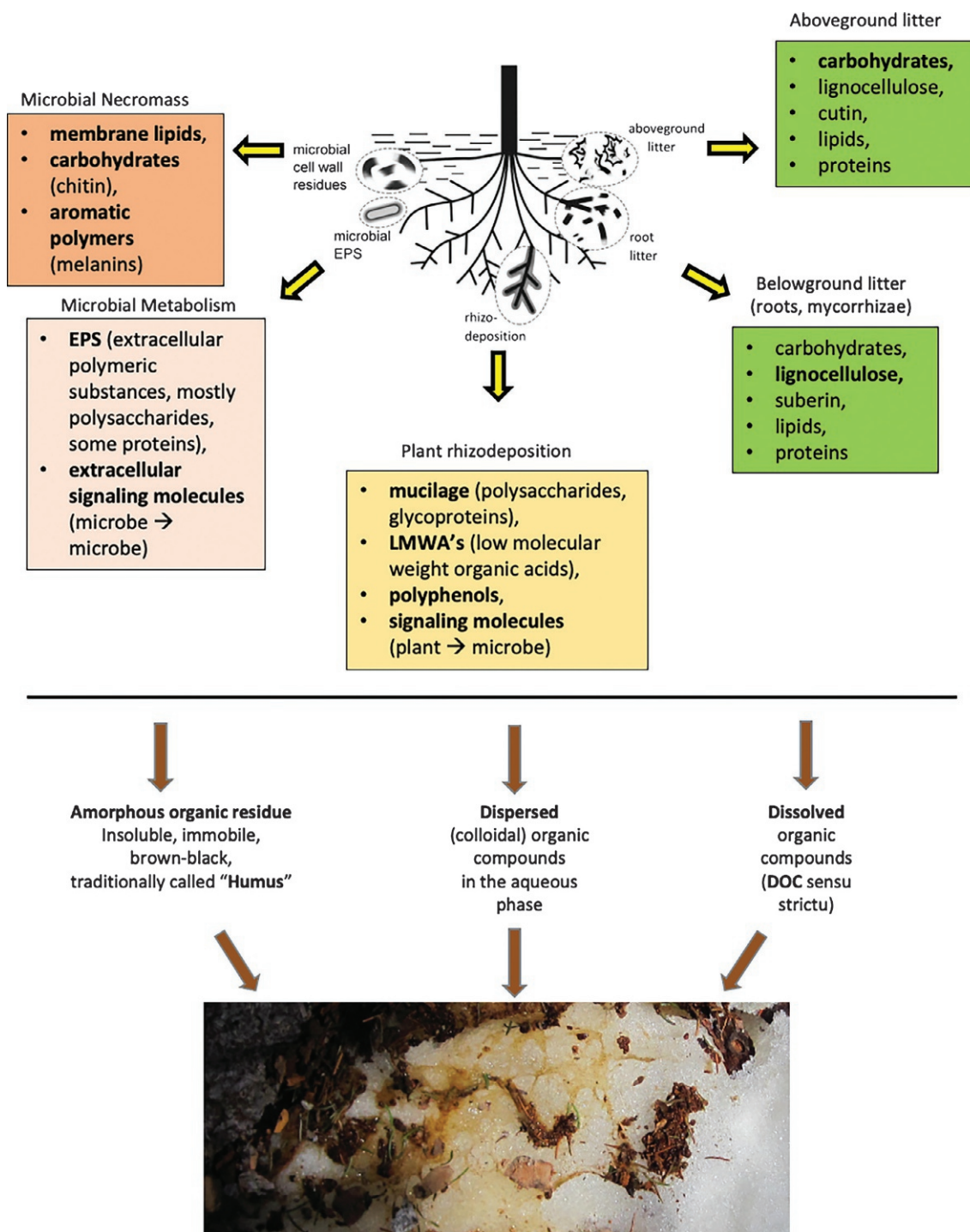


Fig. 2 Five categories of organismic inputs of carbon to the soil. Inputs differ in chemical composition and structure depending on microbial vs plant origin, and whether exudates originate from living organisms or from decomposition of dead organisms. Regardless of organismic input, these materials can eventually be transformed to amorphous, dispersed, or dissolved organic compounds in soil. Modified after Kögel-Knabner I (2017) The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter: Fourteen years on. *Soil Biology & Biochemistry* 105: A3–A8, with permission.

The implications of these transformations for the chemical characteristics of SOM can be visualized by following functional group abundances over the duration of a decomposition experiment (Baldock et al., 2004) (Fig. 3).

As plant residues decay, the carbon that is not respired is synthesized into microbial biomass, leading to an increase over time in proteinaceous and lipid-like materials within the pool of residual materials. Fig. 3 shows that as carbohydrates (O-alkyl carbons) are being consumed, the abundance of carboxylic and amide groups increase, together with the relative proportion of alkyl C

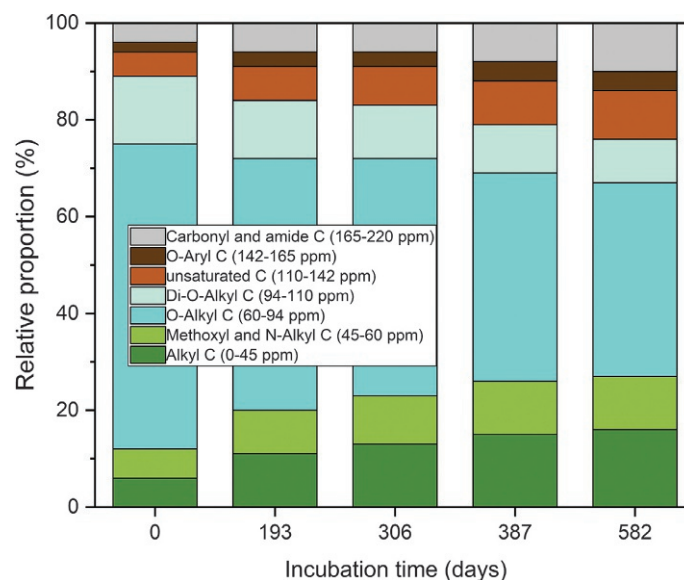


Fig. 3 Chemical alterations (shifts in functional group abundances) of maize (*Zea mays* L.) residue buried at 15-cm depth under deciduous forest over the duration of a 19-month decomposition experiment. Carboxylic groups are found in the 165–220 ppm region; aromatic compounds are between 110 and 165 ppm; carbohydrates range from 60 to 110 and lipids are between 0 and 60 ppm. Data from Baldock JA, Masiello CA, Gelinas Y and Hedges JI (2004) Cycling and composition of organic matter in terrestrial and marine ecosystems. *Marine Chemistry* **92**: 39–64.

representing lipid-like materials. Fig. 3 illustrates that a soil at steady state with regard to the balance between organic matter input and output is similar to the digestive tract of an animal in that it

- constitutes a flow-through system that will
- contain variable proportions of organic input materials at various stages of depolymerization and oxidation, and
- the materials synthesized by the soil microbiome.

Quantitative assessments of the relative contributions of microbially derived materials to SOM have recently been given by two research groups, using comparable methodologies (Angst et al., 2021; Liang et al., 2019; Table 1).

Table 1 suggests that bulk soils under agricultural use have slightly higher proportions of microbial necromass compared with micro- and macro-aggregates from agricultural soils. Two considerations may help in understanding the discrepancy: the database for agricultural bulk soils is significantly larger than that for soil fractions (64 versus 14; Table 2), and the process of separating the physical fractions usually involves considerable exposure to the aqueous phase through wet sieving, flotation and similar procedures, thereby facilitating the loss of soluble compounds. The latter process would lead to a relative enrichment of insoluble plant debris.

Fundamentally, both studies corroborate the notion that the macro- and micro-scopically structureless organic mass of the soil is a mixture of plant debris in various stages of decay and microbial necromass'. The organic matter from grassland soils is particularly rich in microbial contributions while that from forest soils has greater proportions of plant debris.

Residence time of carbon in soil as a function of chemical composition. The appreciation of soil as a flow-through system and of SOM as a relentless flow of carbon means that there is a need to explain why small proportions of organic matter can reside in the

Table 2 Trends in relative contribution of microbial necromass to soil organic matter.

Land use/soil subfraction	# of observations	Microbial necromass %	Bacterial %	Fungal %	Plant residue %	Source
Grassland (0–25 cm)	31	61.8	17.6 ± 1.6	44.2 ± 2.5	38.2	Liang et al. (2019)
Agriculture (Ag) (0–25 cm)	64	55.6	15.5 ± 1.0	40.1 ± 1.8	44.4	Liang et al. (2019)
Micro-aggregates (53–250 µm)	14 (all Ag systems)	49.7	9.1 ± 1.5	40.6 ± 6.5	50.3	Angst et al. (2021)
Macro-aggregates (250–2000 µm)	14 (all Ag systems)	47.2	8.6 ± 1.4	38.6 ± 6.4	52.8	Angst et al. (2021)
Mineral associated	23 (16 Ag, 5 Forest, 2 Grassland)	38.6	15.3 ± 1.6	23.3 ± 4.2	61.4	Angst et al. (2021)
Forest (0–25 cm)	27	32.4	8.7 ± 1.5	23.9 ± 2.4	67.4	Liang et al. (2019)

Data from Angst G, Mueller KE, Nierop KGJ and Simpson MJ (2021) Plant- or microbial-derived? A review on the molecular composition of stabilized soil organic matter. *Soil Biology and Biochemistry* **156**: 108189.; Liang C, Amelung W, Lehmann J and Kastner M (2019) Quantitative assessment of microbial necromass contribution to soil organic matter. *Global Change Biology* **25**: 3578–3590.

soil for millennia. Chemical explanations have been sought to explain these very long residence times. As soils evolve over time from mineral parent materials, their organic matter content typically increases. It is tempting to interpret this observation as an accumulation of such organic phases that have ‘resisted’ decomposition and are thus ‘stable’. This view is supported by numerous decomposition experiments that show certain organic compounds tend to decompose more slowly than others. However, as soils mature, their organic matter concentration always equilibrates (except for peat bogs (histosols) where the biomass produced is preserved through the absence of oxygen). If inherently stable organic compounds were produced, organic matter contents (= the thickness of A horizons) should never equilibrate, they would be expected to increase infinitely, much as pedogenesis can create B horizons of almost unlimited thickness. As early as 1981, this insight led [Jenkinson \(1981\)](#) to postulate that “in the long run, no fraction of organic matter in plants and animals can withstand decomposition to carbon dioxide and water. If this were not so, any completely resistant fraction would by now cover the surface of the earth”.

In contrast to this fundamental consideration (all natural organic matter will eventually decompose unless it is effectively isolated from decomposition processes), the notion that molecular complexity and degree of polymerization should slow down decomposition has received strong support from the work of Martin and coworkers. These scientists introduced ^{14}C into the molecular structure of phenolic acids and artificially prepared phenolic polymers [Haider and Martin \(1975\)](#), allowing them to identify the molecular origin of the carbon in the $^{14}\text{CO}_2$ evolved during decomposition of these molecules and demonstrating that assembling monomers into polymeric structures slowed the decomposition of organic carbon in 12-week incubation experiments. As a result, macromolecularity has long been considered one of the major factors determining the persistence of organic C in geosystems.

Together with earlier suggestions that so-called humic substances have macromolecular structure analogous to seemingly indestructible industrial polymers, these experiments contributed significantly to the view that the chemistry of an organic substrate determines its likelihood of decomposition. Early soil carbon models (Roth C, Century) were based on this premise and assigned intrinsic turnover times to certain molecular categories such as cellulose and lignin.

Residence time of carbon in soil as a function of system architecture, resulting logistics and decomposer specialization. As shown above, the phenomenon of slow decomposition has long been related to the compound in question (such as lignocellulose) being able to ‘resist’ decomposition. However, the notion of an organic compound resisting a chemical reaction is physicochemically incorrect and diverts our attention from the true nature of the underlying mechanisms. Ultimately, decomposition is a sequence of chemical reactions. In chemistry, reactions take place when there is motive and opportunity: the ‘energetic’ requirements for a specific reaction to occur need to be satisfied. The Gibbs free energy, ΔG , of the reaction must be negative, and resource requirements must be met, such as the availability of catalytic tools (enzymes), electron acceptors (O_2), a suitable solvent (H_2O), etc. The reactants in a chemical reaction never ‘resist’; either the requirements for a reaction to proceed are satisfied or they are not. This insight necessarily directs our focus towards the logistic status of a system when we try to rationalize why a certain organic compound is left undecomposed ([Ekschmitt et al., 2008](#)): Is the compound accessible to the decomposer community? Are there decomposers in the system with the appropriate catabolic toolbox? Is there enough oxygen? Is the system too wet or too dry? Are the co-metabolites needed to facilitate the reaction available?

Once the energy and logistic requirements for chemical reactions are in place, decomposition can and will proceed, as demonstrated by [Klotzbuecher et al. \(2011\)](#) for lignin decomposition. Plants have evolved to select and utilize composite molecular compounds (such as cutin and suberin) that require sophisticated logistics to be processed, such that the decomposer community tends to prioritize substrates with easier logistics (such as carbohydrates). Parallel to this, evolution also continues to incentivize decomposer adaptations, resulting in organisms such as termites (that enlist endosymbiotic microorganisms to deal with woody substrates in hours) and white rot fungi (elaborate ‘chemical factories’ able to synthesize both complex enzymes and source the required molecular supplies through their hyphal network). For these reasons, the chemistry of a substrate will influence its turnover dynamics. However, the underlying causality is a matter of resource availability and not a matter of ‘resistance’.

Although organic matter can reside in the soil for millennia, it does not do so because its chemical composition renders it inherently resistant to decomposition. As we strive to understand why a compound or fraction of organic matter decomposes more slowly than others, we need to resist the temptation to classify substrates as inherently labile or inherently stable. A compound such as lignin or even biochar is not left behind because it is ‘recalcitrant’, but where the complicated logistics for its decomposition are not in place (no termites, no white rot fungi, lack of oxygen, lack of co-metabolites, etc.).

Carbon in soil and the humic substances paradigm. The decomposition process produces a cascade of intermediate products with specific properties (structurally amorphous, water insoluble, enriched in ionizable carboxylic groups.) that

- a) visibly distinguish them from other soil constituents and
- b) render those decomposition intermediates highly relevant for the functioning of soils ([Fig. 4](#)).

Examples of such soil functions are the cation exchange capacity contributed by organic matter in advanced decomposition stages and the facilitation of aggregation processes that eventually determine the volume and shape of the pore space needed for life processes within the soil.

The so-called “humic substances paradigm” ([Wershaw, 2000](#)) posits that soil humus is composed of the end products of synthetic reactions that alter the structure of plant degradation products in a way that provides these newly synthesized materials with unique properties that are distinct from non-humified organic matter. At its outset, the humic substances paradigm ([Fig. 5](#)) was conceived to reconcile apparent chemical properties (high molecular mass, large abundance of aromatic moieties and resulting inherent recalcitrance) in a way that would explain the observed dark coloration, slow turnover and resulting old radiocarbon ages of the amorphous, water insoluble organic residues typical for SOM ([Fig. 2](#), bottom left).



Fig. 4 Conversion of plant debris with clearly recognizable macroscopic plant structure (soil surface, far left) into a water-insoluble, seemingly amorphous residue (far right, 40-cm depth). Image taken by the author, showing 5–10 cm depth increments from the surface of a Haplorthod at the Oregon Coast. Note the color change associated with the transformation of brownish plant debris (far left) to amorphous humus (black, far right).

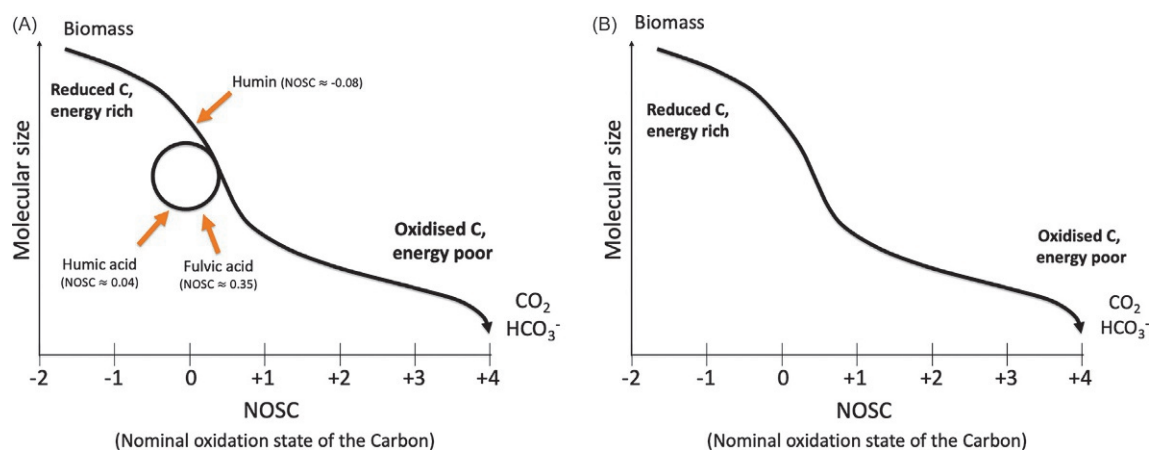


Fig. 5 (A) Trajectory of decomposing organic matter according to the humic substances paradigm, involving initial breakdown followed by subsequent secondary synthesis of decomposition products into intermediate humic and fulvic substances. Approximate NOSC values derived from average elemental compositions of alkali extracted humic substances ($n = 26, 214$ and 410 for humin, fulvic acid and humic acid, respectively) using data from Table 4.3 (all sources) in [Essington \(2015\)](#). (B) Trajectory of decomposing organic matter according to the Continuum Model ([Lehmann and Kleber, 2015](#)), involving progressive deconstruction of organic biomacromolecules to fully oxidized carbon (CO₂; HCO₃⁻).

Evidence has accumulated that the macro- and micro-scopically structureless organic mass of the soil is rarely macromolecular, it is not always highly aromatic, and it may decompose readily if given the opportunity to do so. Together with the lack of a thermodynamic rationale for decomposer organisms to synthesize compounds that deliberately lack any obvious metabolic purpose as proposed by the humic substances paradigm ([Fig. 5A](#)), these findings led to suggestions that this paradigm is an outdated chemical model for SOM. Rather, the evidence available to date suggests that SOM is more accurately represented as a continuum of organic compounds in various stages of decay ([Lehmann and Kleber, 2015](#); [Trumbore, 1997](#)) experiencing progressive oxidation and reduction in molecular size ([Fig. 5B](#)).

Xenobiotic organic compounds

Xenobiotic organic compounds are organic chemicals that are foreign (Greek *xenos* = stranger, guest) to natural living systems. Modern society relies heavily on thousands of chemical compounds that did not exist before the industrial revolution some 200 years ago. Because life was not exposed to these compounds during evolutionary processes of the past, it has had no opportunity to adapt to their presence. This means that the organisms in a natural system have, apart from some rare metabolic coincidence, no adaptations that would allow them to catabolize these compounds or to exclude them deliberately from their metabolism. This can result in toxic effects and to an unwanted accumulation in soil environments and in the organisms inhabiting the soil. Xenobiotic organic compounds can loosely be subdivided into monomeric and polymeric categories.

Monomeric xenobiotics. Monomeric xenobiotics are either deliberately applied to soils, such as agricultural pesticides, or end up in the soil as a consequence of their tolerated or accidental release into the environment, such as veterinary antibiotics or *per*- and poly-fluoroalkyl substances. Quantitatively, these additions may, at first sight, seem insignificant. In 2018, the world average pesticide use per area of cropland was 2.63 kg ha⁻¹ ([Raffa and Chiampo, 2021](#)), which would be equivalent to 0.00263% of the

100,000 kg of organic matter in the top 25 cm of an agricultural field containing 4% of organic matter at a bulk density of 1.0 kg L⁻¹. However, a recent analysis of pesticide effects on soil invertebrates involving 284 different active pesticide ingredients and 275 unique invertebrate species found that 70.5% of properties tested showed negative effects, whereas only 1.4% and 28.1% showed positive or no significant effects, respectively (Gunstone et al., 2021). Attempts to characterize the influence of pesticides on microbial community composition and soil enzyme production are less conclusive (Wolejko et al., 2020). There is some evidence that microbial communities may respond more strongly to the application of different types of natural organic substrates than to the application of a given herbicide (Carpio et al., 2020). More research is needed to clarify pesticide effects on microbial metabolic performance as well as on microbial community composition. At present, however, it appears that soil invertebrates are under greater pressure from pesticide applications than are the more diverse and more adaptable microbial communities.

Apart from concerns for the general functionality of soil organisms, xenobiotics that are introduced in more confined areas warrant attention from a human health standpoint. Manure or sludge applications to the soil may create source zones for veterinary antibiotics (Kemper, 2008; Tasho and Cho, 2016), enabling eventual transfer to human food and water supply systems. Another example of the creation of soil source zones of xenobiotic compounds with potentially harmful effects on human health are the training areas for fire fighters around military bases, airports and other industrial locations (Hu et al., 2016). Poly- and per-fluoroalkyl substances (PFASs) are present in aqueous film-forming foam (AFFF) formulations used to suppress fires caused by aviation fuel and similar nonpolar liquids. These fluorinated compounds are designed to accumulate at the interface between burning fuel and overlying air, thereby sealing the burning liquid off from atmospheric oxygen supply. The PFAS can be initially retained by soil matrices (Barzen-Hanson et al., 2017; Nickerson et al., 2021) and are considered to degrade at very slow rates (Kwiatkowski et al., 2020), however lateral export with groundwater flow of PFAS and their transformation products has been documented (Adamson et al., 2020) and may affect aquifers needed for human water supply.

Polymeric xenobiotics (plastics and microplastics). While 4.1 mt of agricultural pesticides were applied globally in 2015 (Maggi et al., 2019), the global production of industrial organic polymers, also known as plastics (microplastics if smaller than 5 µm), was two orders of magnitude higher at 367 mt in 2020 (Rouch, 2021). Estimates of global emissions of plastic waste to rivers, lakes, and oceans range from 9 to 23 mt per year, with a similar amount emitted into the terrestrial environment, from 13 to 25 mt per year as of 2016 (Borrelle et al., 2020; Lau et al., 2020). According to MacLeod et al. (2021), these rates of emission estimated in 2016 will double by 2025. These authors also suggest that the amount of plastic in the world's agricultural soil may be larger than on the ocean's surface.

Bläsing and Amelung, 2018 identified plastic litter, road runoff (including tire wear particles), and atmospheric deposition of micro- and nano-plastic particles as sources of plastic pollution to urban and rural soils. Other input pathways to soils are plastic mulching with polyethylene films and so-called 'biodegradable' plastic films, compost, and sludge-derived biosolids, as well as polymeric stabilizers against soil erosion. Quantitative estimates for plastic fractions in soils reach up to 0.1% of soil organic carbon (Bläsing and Amelung, 2018) and are at least two orders of magnitude greater than estimates for pesticide applications (Raffa and Chiampo, 2021).

Plastic concentrations in soil are expected to increase because of these ongoing sources of emission combined with extremely slow rates of degradation (MacLeod et al., 2021). Although plastics undergo various fragmentation and disintegration processes, Bläsing and Amelung, 2018 estimate that less than 1% of the mass of conventional plastic will leave the soil on annual time scales, making plastic pollution of soils a poorly reversible phenomenon.

Unfortunately, plastics are not inert admixtures to the soil system. Bandyopadhyay et al. (2018) report long-term changes in soil properties, such as water holding capacity, microbial activity and diversity, nutrient availability, and soil structure associated with increased loads of plastic.

Other implications of plastic accumulation in soils include effects on plant performance and plant diversity and potentially irreversible soil degradation (Steinmetz et al., 2016). Observations from studies in aquatic systems suggest that processes of nitrogen and phosphorus cycling can be affected by microplastics.

Xenobiotic compound summary. For about 60 years, both monomeric and polymeric xenobiotic compounds have been added to soils in ever increasing quantities, and evidence for their role as global stressors in terrestrial ecosystems is mounting. Microplastics <700 nm have now even been found in human blood (Leslie et al., 2022), adding urgency to the need to include these materials in modern concepts of SOM, despite the fact that their quantitative contribution may seem small.

Thermally altered C

Before the advent of the first computer-based models for the turnover of SOM, pyrogenic carbon (carbon compounds formed from organic matter by pyrolysis, i.e. through heating in the absence of oxygen), was not considered a functionally relevant part of SOM. However, when computer models were developed, it was found that they needed to include a supposedly 'inert' conceptual organic matter pool. Inert organic matter is defined as a fraction of SOM that is biologically unreactive and has an equivalent radiocarbon age of more than 50,000 years. The inert organic matter pool is a device to allow the model to represent short-term changes in SOM brought about by changes in land management, and at the same time to account for the great radiocarbon ages measured in surface soils collected before thermonuclear tests. Of all the known forms of carbon, charcoal was seen as one of the most likely candidate materials representing this pool.

Almost parallel to the development of the first carbon turnover models was a report (Seiler and Crutzen, 1980) on the substantial sink of atmospheric CO₂ by the formation of black carbon in vegetation fires. The authors argued that a relatively

unknown (at that time) fraction of the vegetation exposed to fire is converted to an extremely long-lived form of carbon. Because the resulting pyrogenic or 'black' carbon was considered 'refractory', it was seen as a potential explanation for the so-called missing carbon sink, i.e. the discrepancy between the annual release of carbon into the atmosphere from fossil fuel combustion and the actual, observed uptake of carbon by the atmosphere. Scientists are now more reserved about this interpretation, but 30 years ago, the scientific community connected the dots between (i) model requirements for an 'inert' carbon pool, and (ii) recognition of significant black carbon inputs to terrestrial systems, and began to investigate the abundance of pyrogenic carbon in soils more closely.

Abundance of pyrogenic carbon in soil. The first quantitative report of pyrogenic carbon in soils came from Australia, a continent dominated by fire-adapted ecosystems. Skjemstad et al. (1996) estimated that 30% of soil carbon was pyrolytic in the soils investigated. This number attracted wide attention, and numerous studies into other ecosystems followed. It soon became clear that admixtures of pyrogenic carbon to soils are globally ubiquitous, reaching particularly high values in boreal forests and in grassland soils. On a global scale, Reisser et al. (2016) found that pyrogenic carbon contributed on average 13.7% of the soil organic carbon, effectively making it one of the largest groups of chemically distinct carbon compounds in SOM. Reisser et al. (2016) also noted that soil properties and climate variables were statistically correlated with the abundance of pyrogenic carbon, suggesting that factors such as clay content and MAT/MAP (mean annual temperature/mean annual precipitation) were instrumental in the retention of pyrogenic carbon.

Chemistry of pyrogenic carbon. The chemical composition of pyrogenic carbon is a function of (i) the properties of the organic precursor material (wood, grass, fossil fuel etc.) and of (ii) pyrolysis conditions, i.e., the duration and intensity of the exposure to heat. At low temperatures, plant biomass is mainly dehydrated and remains otherwise unaltered. As charring intensity increases, organic macromolecules, e.g., cellulose, lignin and hemicellulose, are lost, and isolated aromatic rings begin to form. Eventually, recognizable biopolymers are lost and isolated aromatic molecules with 2 and 3 rings are formed. When temperature increases further, small and somewhat 'defective' sheets of condensed aromatic rings stack up to form so called turbostratic crystallites. These are small three-dimensional structures that consist of 3 to 5 stacked C sheets with a vertical height of 1–2 nm and a lateral extension of 2–5 nm. These dimensions depend on many factors such as type of organic precursor material (for instance, wood versus grass), peak heating temperature and mineral admixtures. When all amorphous organic C has either been volatilized or converted into aromatic rings, the material is designated as 'carbonized'. Pyrogenic carbon formed in natural environments will not typically fall into this latter category as it either requires prolonged exposure to pyrolysis conditions or heating temperatures in excess of about 700 °C, conditions that are rarely or only briefly met in vegetation fires. However, dependence of the molecular properties of pyrogenic carbon on pyrolysis conditions means that pyrolytic carbon can exhibit a predominance of 'organic' carbon forms when resulting from low intensity pyrolysis, or show decidedly 'mineral' or 'inorganic' features when assembled into graphite-like crystallites able to diffract X-ray radiation. This raises the interesting question of whether highly crystalline chars should still be considered part of the organic carbon pool in soil, or as part of the inorganic mineral matrix (Fig. 1).

With pyrogenic carbon presenting in many diverse molecular forms, identification and quantification of thermally altered carbon in soil remains an analytical challenge that has prompted numerous efforts to standardize and optimize the analytical portfolio.

Impact of pyrogenic carbon on soil functions. Pyrogenic carbon is by no means a chemically homogeneous material. It is no surprise that soil and plant responses to additions of pyrogenic carbon can be negative, positive, or neutral (Joseph et al., 2021). A vibrant research community is presently studying the effects of deliberate additions of well characterized varieties of pyrogenic carbon (so called biochar) to soils. These efforts are expected to contribute to a mechanistic, and potentially generalizable, understanding of the ecological consequences of the contribution of different forms of pyrolytic carbon to soil functionality.

Conclusion and outlook

Humankind has known for millennia that the organic phase is vital for the functioning of soil. Prior to the industrial revolution, a focus on "the macro- and microscopically structureless organic mass of the soil" or humus was understandable and justified as an expression of the recognition of the particular importance of these materials. With the onset of the industrial age, humankind has deliberately applied or tolerated deposition of increasing amounts of xenobiotic organic compounds to which neither plants nor decomposer communities were adapted to. Thermally altered, 'black' carbon was always part of the organic phase in soil, but its ubiquity and quantitative significance has only become known over the past 40 years. In the light of these developments, our concepts of the chemical composition of soil organic matter need to be reevaluated and modernized. A new view is emerging that SOM could be represented more accurately as a continuum of organic compounds derived from both punctuated and continuous processes, natural and anthropogenic sources, and in various stages of decay.

The stage for future research on SOM is set by an acknowledgement of the three major environmental aspects of Global Change: Soil Degradation, Loss of Biodiversity and Climate Change. So far, only Climate Change has received due attention in the media and by the public. This makes it difficult for the scientist to find the resources necessary to develop countermeasures against soil degradation and loss of biodiversity, it also hinders coordinated and timely action by legislators. For this reason, public initiatives such as the Soil Health movement warrant attention and active support from the scientific community.

References

- Adamson DT, Nickerson A, Kulkarni PR, Higgins CP, Popovic J, Field J, Rodowa A, Newell C, DeBlanc P, and Kornuc JJ (2020) Mass-Based, Field-Scale Demonstration of PFAS Retention within AFFF-Associated Source Areas. *Environmental Science & Technology* 54: 15768–15777.
- Angst G, Mueller KE, Nierop KGJ, and Simpson MJ (2021) Plant- or microbial-derived? A review on the molecular composition of stabilized soil organic matter. *Soil Biology and Biochemistry* 156: 108189.
- Baldock JA and Broos K (2011) *Soil Organic Matter, Handbook of Soil Sciences*. CRC Press. 11–11–11–52.
- Baldock JA, Masiello CA, Gelinas Y, and Hedges JI (2004) Cycling and composition of organic matter in terrestrial and marine ecosystems. *Marine Chemistry* 92: 39–64.
- Bandyopadhyay S, Martin-Closas L, Pelacho AM, and DeBruyn JM (2018) Biodegradable plastic mulch films: Impacts on soil microbial communities and ecosystem functions. *Frontiers in Microbiology* 9: 819.
- Barzen-Hanson KA, Davis SE, Kleber M, and Field JA (2017) Sorption of fluorotelomer sulfonates, fluorotelomer sulfonamido betaines, and a fluorotelomer sulfonamido amine in natural foam aqueous film-forming foam to soil. *Environmental Science & Technology* 51: 12394–12404.
- Bläsing M and Amelung W (2018) Plastics in soil: Analytical methods and possible sources. *Science of the Total Environment* 612: 422–435.
- Borrelle SB, Ringma J, Law KL, Monnahan CC, Lebreton L, McGivern A, Murphy E, Jambeck J, Leonard GH, Hilleary MA, Eriksen M, Possingham HP, De Frond H, Gerber LR, Polidoro B, Tahir A, Bernard M, Mallos N, Barnes M, and Rochman CM (2020) Predicted growth in plastic waste exceeds efforts to mitigate plastic pollution. *Science* 369: 1515.
- Carpio MJ, Garcia-Delgado C, Marin-Benito JM, Sanchez-Martin MJ, and Rodriguez-Cruz MS (2020) Soil microbial community changes in a field treatment with chlorotoluron, flufenacet and diflufenican and two organic amendments. *Agronomy* 10: 1166.
- Ekschmitt K, Kandeler E, Poll C, Brune A, Buscot F, Friedrich M, Gleixner G, Hartmann A, Kastner M, Marhan S, Miltner A, Scheu S, and Wolters V (2008) Soil-carbon preservation through habitat constraints and biological limitations on decomposer activity. *Journal of Plant Nutrition and Soil Science-Zeitschrift Fur Pflanzenernahrung Und Bodenkunde* 171: 27–35.
- Essington ME (2015) *Soil and Water Chemistry: An Integrative Approach*. CRC Press.
- Gunstone T, Cornelisse T, Klein K, Dubey A, and Donley N (2021) Pesticides and soil invertebrates: A hazard assessment. *Frontiers in Environmental Science* 9.
- Haider K and Martin JP (1975) Decomposition of specifically C14 labeled benzoic and cinnamic acid derivatives in soil. *Soil Science Society of America Journal* 39: 657–662.
- Hedges JI, Eglington G, Hatcher PG, Kirchman DL, Arnosti C, Derenne S, Evershed RP, Koegel-Knabner I, de Leeuw JW, Littke R, Michaelis W, and Rulkötter J (2000) The molecularly-uncharacterized component of nonliving organic matter in natural environments. *Organic Geochemistry* 31: 945–958.
- Hu XDC, Andrews DQ, Lindstrom AB, Bruton TA, Schaidler LA, Grandjean P, Lohmann R, Carignan CC, Blum A, Balan SA, Higgins CP, and Sunderland EM (2016) Detection of poly- and perfluoroalkyl substances (PFASs) in US drinking water linked to industrial sites, military fire training areas, and wastewater treatment plants. *Environmental Science & Technology Letters* 3: 344–350.
- Janzen HH (2006) The soil carbon dilemma: Shall we hoard it or use it? *Soil Biology & Biochemistry* 38: 419–424.
- Janzen HH (2015) Beyond carbon sequestration: Soil as conduit of solar energy. *European Journal of Soil Science* 66: 19–32.
- Jenkinson DS (1981) The fate of plant and animal residues in soil. In: Greenland DJ and Hayes MHB (eds.) *The Chemistry of Soil Processes*, pp. 505–561. Chichester: Wiley.
- Joseph S, Cowie AL, Van Zwieten L, Bolan N, Budai A, Buss W, Cayuela ML, Graber ER, Ippolito JA, Kuzyakov Y, Luo Y, Ok YS, Palansooriya KN, Shepherd J, Stephens S, Weng Z, and Lehmann J (2021) How biochar works, and when it doesn't: A review of mechanisms controlling soil and plant responses to biochar. *Global Change Biology. Bioenergy* 13: 1731–1764.
- Kemper N (2008) Veterinary antibiotics in the aquatic and terrestrial environment. *Ecological Indicators* 8: 1–13.
- Klotzbuecher T, Kaiser K, Guggenberger G, Gatzek C, and Kalbitz K (2011) A new conceptual model for the fate of lignin in decomposing plant litter. *Ecology* 92: 1052–1062.
- Kögel-Knabner I (2002) The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter. *Soil Biology and Biochemistry* 34: 139–162.
- Kögel-Knabner I (2017) The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter: Fourteen years on. *Soil Biology & Biochemistry* 105: A3–A8.
- Kwiatkowski CF, Andrews DQ, Birnbaum LS, Bruton TA, DeWitt JC, Knappe DRU, Maffini MV, Miller MF, Pelch KE, Reade A, Soehl A, Trier X, Venier M, Wagner CC, Wang ZY, and Blum A (2020) Scientific basis for managing PFAS as a chemical class. *Environmental Science & Technology Letters* 7: 532–543.
- Lau WWY, Shiran Y, Bailey RM, Cook E, Stuchey MR, Koskella J, Velis CA, Godfrey L, Boucher J, Murphy MB, Thompson RC, Jankowska E, Castillo AC, Pilditch TD, Dixon B, Koerselman L, Kosior E, Favoino E, Gutberlet J, Baulch S, Atreya ME, Fischer D, He KK, Petit MM, Sumaila UR, Neil E, Bernhofen MV, Lawrence K, and Palardy JE (2020) Evaluating scenarios toward zero plastic pollution. *Science* 369: 1455.
- Lehmann J and Kleber M (2015) The contentious nature of soil organic matter. *Nature* 528.
- Leslie HA, van Velzen MJM, Brandsma SH, Vethaak AD, Garcia-Vallejo JJ, and Lamoree MH (2022) Discovery and quantification of plastic particle pollution in human blood. *Environment International* 163: 107199.
- Liang C, Amelung W, Lehmann J, and Kastner M (2019) Quantitative assessment of microbial necromass contribution to soil organic matter. *Global Change Biology* 25: 3578–3590.
- MacLeod M, Arp HPH, Tekman MB, and Jahnke A (2021) The global threat from plastic pollution. *Science* 373: 61–65.
- Maggi F, Tang FHM, la Cecilia D, and McBratney A (2019) PEST-CHEMGRIDS, global gridded maps of the top 20 crop-specific pesticide application rates from 2015 to 2025. *Scientific Data* 6: 170.
- Nickerson A, Rodowa AE, Adamson DT, Field JA, Kulkarni PR, Kornuc JJ, and Higgins CP (2021) Spatial trends of anionic, zwitterionic, and cationic PFASs at an AFFF-impacted site. *Environmental Science & Technology* 55: 313–323.
- Oades JM (1989) An introduction to organic matter in mineral soils. In: Dixon JB and Weed SB (eds.) *Minerals in Soil Environments*, 2nd edn, pp. 89–160. Madison, WI: SSSA. 53711.
- Pauli FW (1961) Humus and plant (The direct humus effect). *Science Progress* 49: 427–439.
- Raffa CM and Chiampo F (2021) Bioremediation of agricultural soils polluted with pesticides: A review. *Bioengineering* 8: 92.
- Reisser M, Purves RS, Schmidt MWI, and Abiven S (2016) Pyrogenic carbon in soils: A literature-based inventory and a global estimation of its content in soil organic carbon and SStocks. *Frontiers in Earth Science* 4: 80.
- Rouch DA (2021) *Plastic Future: How to Reduce the Increasing Environmental Footprint of Plastic Packaging*. Melbourne: Clarendon Policy & Strategy Group.
- Seiler W and Crutzen PJ (1980) Estimates of gross and net fluxes of carbon between the biosphere and the atmosphere from biomass burning. *Climatic Change* 2: 207–247.
- Skjemstad JO, Clarke P, Taylor JA, Oades JM, and McClure SG (1996) The chemistry and nature of protected carbon in soil. *Australian Journal of Soil Research* 34: 251–271.
- Steinmetz Z, Wolmann C, Schaefer M, Buchmann C, David J, Troeger J, Munoz K, Fror O, and Schaumann GE (2016) Plastic mulching in agriculture. Trading short-term agronomic benefits for long-term soil degradation? *Science of the Total Environment* 550: 690–705.
- Tasho RP and Cho JY (2016) Veterinary antibiotics in animal waste, its distribution in soil and uptake by plants: A review. *Science of the Total Environment* 563: 366–376.
- Trumbore SE (1997) Potential responses of soil organic carbon to global environmental change. *Proceedings National Academy of Sciences USA* 94: 8284–8291.
- Weil R and Brady NC (2016) *The Nature and Properties of Soils*, 15th edn Harlow, England: Pearson Education Limited.
- Wershaw RL (2000) The study of humic substances - in search of a paradigm. In: Ghabbour EA and Davies G (eds.) *Humic Substances Versatile Components of Plants, Soil and Water*, pp. 1–9. Cambridge: The Royal Society of Chemistry.
- Wöhler F (1828) Über künstliche Bildung des Harnstoffs. *Annalen der Physik und Chemie* 88: 253–256.
- Wolejko E, Jablonska-Trypuc A, Wydro U, Butarewicz A, and Lozowicka B (2020) Soil biological activity as an indicator of soil pollution with pesticides - A review. *Applied Soil Ecology* 147: 103356.

A more extensive list of references is available from the author upon request.

Physical fractionation techniques

Sam J Leuthold^a, Michelle L Haddix^a, Jocelyn Lavallee^{a,b}, and M Francesca Cotrufo^a, ^aDepartment of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, United States; ^bEnvironmental Defense Fund, New York, NY, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	68
Conceptual framework	69
Isolatable SOM fractions	70
Particulate organic matter (POM)	70
Mineral-associated organic matter (MAOM)	71
Other primary SOM fractions	71
Dissolved organic matter	71
Heavy coarse organic matter	71
Soil aggregates	71
Other SOM fractions not isolated by physical fractionation methods	71
Aggregate fractionation methods	72
Primary particle fractionations	72
Dispersal methods	72
Mechanical shaking	72
Ultrasonic vibration	74
Size fractionations	74
Density separations	75
Combined separations	77
Chemical-free physical fractionations	77
Conclusions	79
Acknowledgment	79
References	79

Abstract

Physical fractionation methods enable the separation of soil organic matter fractions that differ in their function, formation pathways, and persistence in soils. Several different physical fractionation methods have been developed to separate different components of soil organic matter by either particle size, particle density, or a combination of these characteristics. Here we summarize these techniques, providing first a conceptual framework for the fractions that are isolatable, and then describing the variety of methods that have been used by researchers to differentiate various fractions.

Key points

- The function and formation of soil organic matter can be better understood by separating it into fractions defined by their physicochemical characteristics.
- Depending on the goals of the researcher, several different fractions can be isolated by separating particles based on size, density, or a combination of the two properties.
- The method of dispersion as well as the dispersing solution used present important considerations for researchers using these methodologies.

Introduction

Soil organic matter (SOM) is a key renewable resource, providing several ecosystem services such as the regulation of nutrient cycling, soil water storage, and carbon sequestration. Soil organic matter is largely composed of diverse, relatively simple molecular compounds that are present in the soil solution, attached to mineral particles, or contained within fragmented organic residues (von Lützow et al., 2007). Over the last 40 years, continued advancements in SOM research have led to an increasing acceptance that to understand best the function, formation, and persistence of SOM, different components must be isolated and examined as critical parts of a greater whole (Cambardella and Elliott, 1992; Christensen, 2001; Lavallee et al., 2020; Lehmann and Kleber, 2015; Sollins et al., 1984). Several methods of separating SOM fractions have been developed, such as the now dismissed use of strong alkaline extractions to isolate organic molecules (Stevenson, 1994), or laboratory incubations that aim to determine turnover times of

different SOM constituents (Paul et al., 1999). In this chapter, we focus our discussion on methods of isolating organic matter fractions by physical means.

Physical fractionation methods enable the isolation of SOM fractions that are less heterogeneous in their turnover time, formation processes, and/or chemical make-up than the bulk SOM. In contrast to other methods of SOM separation, physical fractionation methods reduce the potential for procedural artifacts or modification of the chemical composition of the final samples, and produce samples fit for further analysis (e.g., elemental and isotopic composition, hydrogen pyrolysis, etc.). Most importantly, these physical methods have been developed to separate SOM into fractions based on inherent physiochemical properties, which allows increased insight into the mechanisms of SOM formation, stabilization, and persistence of the different components.

A distinguishing feature of the current paradigm of SOM research has been an inquiry model that exists downstream from the methodology at the researcher's disposal. That is to say, the questions asked are often directly informed by the techniques for separating constituent fractions, rather than fractionation approaches dictated by the differences that exist in the function, formation, and persistence of the substances that comprise the organic component of the soil. The fractions we discuss in this chapter are by nature operationally defined, and the operational definitions will vary given the dogma of a given laboratory, institution, or researcher. While the wealth of high-quality research under this paradigm has brought about a deep understanding of SOM dynamics, the foundation is now strong enough such that we have the opportunity to craft questions using an inductive reasoning framework and to use methods that are informed by the scope of our inquiry, rather than inform it. Here we provide an overview of the different functional components of SOM, the relevance to research studies, and the ways in which researchers can go about separating these components. Our intention is to provide a resource for researchers to develop an understanding of the most typical products of physical fractionation schemes, the information they can provide, and the methods by which they can be isolated such that researchers can make informed decisions concerning the techniques they employ.

Conceptual framework

The molecules within a given organic residue or exudate entering the soil have a variety of potential fates once they begin to decompose and become part of the SOM pool. The secretion of specialized extracellular enzymes by soil microbial communities transforms insoluble molecules such as cellulose, hemicellulose, lignin, suberin and pectin contained within the structural components of plant tissue into low molecular weight compounds by a repeated process of colonization and depolymerization. In contrast, exudates and soluble compounds within residue such as organic acids, amino acids, and sugars, are readily leached from the residue and enter the soil solution directly. Low molecular weight compounds can be assimilated into the microbial biomass or respired by soil microbes, but may also bypass microbial interactions altogether and associate with mineral surfaces, often creating strong and long lasting bonds (Liang et al., 2017). As microbial energy sources become exhausted and communities begin to turn over, the nitrogen-rich molecules contained in microbial byproducts and necromass have a similar range of fates as plant compounds, depending on their polymeric structure, solubility, and affinity for bonding to mineral surfaces.

The different structures, fates, and processing pathways of organic molecules after entering the soil leads to SOM that is diverse in its composition, functional properties, and accessibility to decomposers. Organic matter consisting of structural plant and microbial residues free from mineral association is defined as particulate organic matter (POM), the low molecular weight compounds in the soil solution are defined as dissolved organic matter (DOM), while the organic matter found in fine mineral associations is defined as mineral-associated organic matter (MAOM). While arguably there are no 'uniform' SOM fractions, it is useful to describe them as 'primary' when they are distinct from one another and 'secondary' when they are composite, i.e., formed from the aggregation of primary fractions (Christensen, 2001; Denef et al., 2010). Therefore, POM, MAOM, and DOM can be considered primary SOM fractions, as they occur as discrete conceptual units in the soil, and represent a fundamental level of structural complexity (Christensen, 2001). The secretion of extracellular enzymes or extracellular polymeric substances by microbes onto these primary fractions can facilitate further accumulation of soil particles and OM around them, leading to the formation of soil aggregate structures. As soil aggregates are composed of primary structures, they are considered secondary SOM fractions (Christensen, 2001; Denef et al., 2010). Aggregate formation typically occurs through persistent bonding facilitated by organic molecules, polyvalent cation complexes, and aluminosilicate minerals, often around an organic nucleus (Oades and Waters, 1991). Aggregated structures can protect POM by spatial isolation, and have also been shown to facilitate the formation of MAOM as intra-aggregate breakdown and microbial turnover occur (Witzgall et al., 2021). While all SOM will eventually be utilized by the soil microbial community, stoichiometric complexity and diversity, nutrient deficiencies, and oxygen or temperature limitations can moderate the rate and efficiency with which microbial degradation occurs. Thus, POM fractions are often separated between free and occluded in aggregates of different size fractions as a way to get more insight into their turnover time (Six and Paustian, 2014).

Just as the form and function of different SOM fractions vary, the physical characteristics associated with these different SOM fractions, including their size and density, and the size and density of the mineral particles or aggregates they may be associated with, varies among the different fractions (Fig. 1). Taking advantage of these differences enables the separation of MAOM, POM, DOM, and additional fractions, free or as aggregates, by physical procedures. Despite physical SOM fractions being inherently operationally defined, the covariation in physical characteristics of the fractions with the formation mechanisms, response to environmental changes, and chemical composition of the SOM fractions makes them a useful approach to study SOM dynamics.

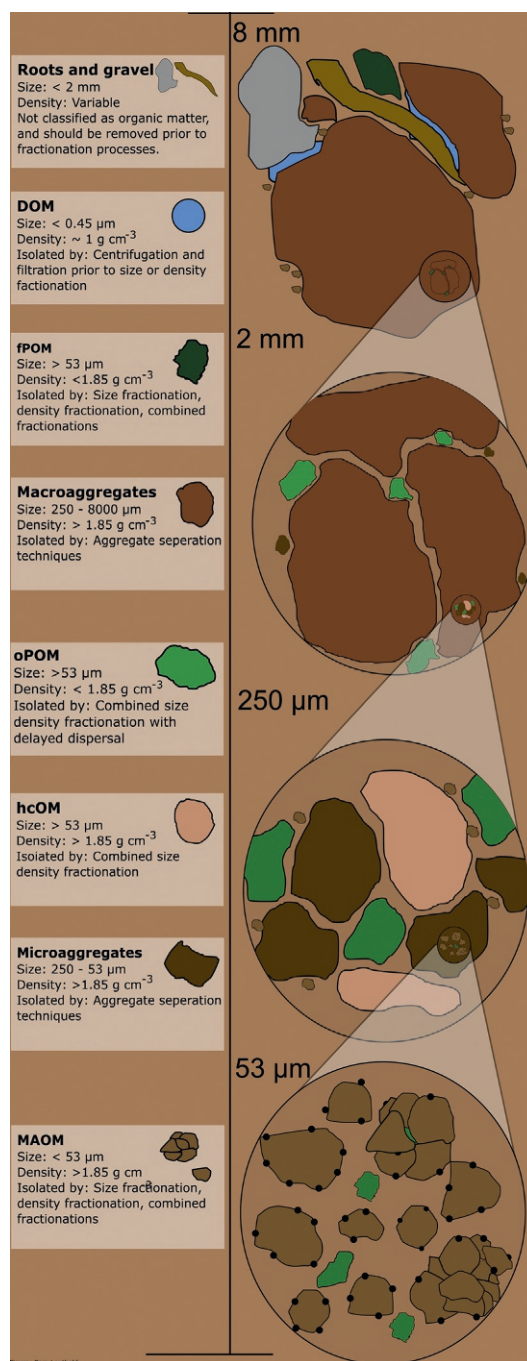


Fig. 1 Conceptual diagram of the fractions that can be isolated by various physical fractionation techniques. Macro- and micro-aggregates can be isolated using aggregate separation techniques, while mineral associated organic matter (MAOM), free and occluded particulate organic matter (fPOM and oPOM, respectively), heavy coarse organic matter (hcOM) and dissolved organic matter (DOM) require size, density, or combined fractionation techniques.

Isolatable SOM fractions

Particulate organic matter (POM)

Particulate organic matter is the term for the fraction of SOM primarily composed of structural materials of both plants and microbes (e.g., cell walls) that have undergone fragmentation and leaching losses but little to no depolymerization. The structural nature of POM makes it relatively light and coarse, leading to an operational definition of organic material with densities <1.6–1.85 g cm⁻³, and or sizes >53 μm. When separated by density fractionation, POM is often referred to as the light fraction, as its lower density allows separation with floatation. Depending on the type of fractionation scheme and the objectives of a given

study, POM can be further fractionated into free and occluded POM. Free POM is uncomplexed POM that is suspended within the soil matrix and is recoverable prior to soil dispersal methods (Fig. 1). In contrast, occluded POM refers to the fraction of POM that is protected within soil aggregates and is only recoverable following the dispersion of soil structures. Despite these differences in protection, both free and occluded POM fractions have similar sizes and densities.

Mineral-associated organic matter (MAOM)

Mineral-associated organic matter refers to the organic molecules that have sorbed to mineral surfaces or are occluded within highly stable fine microaggregate structures. The dominant molecular components of the MAOM fraction are low molecular weight compounds of plant and microbial origin such as amino sugars and microbial polysaccharides, as well as some larger molecules such as extra polymeric substrates, and microbial necromass. The bulk of organo-mineral associations occur with minerals of the silt and clay size, and as such the operational definition of MAOM is the organic material smaller than 60–50 μm and with a density $>1.6\text{--}1.85\text{ g cm}^{-3}$. This fraction is often referred to in the literature as the heavy fraction when separated by density methods, or the silt and clay fraction when separated based on size. The organo-mineral bonds generally protect MAOM from microbial degradation, but recent work has shown that a fraction of MAOM can actively exchange with the soil solution. Previously, chemical methods such as acid hydrolysis have been used to quantify the amount of MAOM that is stable in its bonds and not actively exchanging with the soil solution (Ramnarine et al., 2018), but new methods for separating the exchangeable pool are actively being developed. Recent advances indicate that methods such as the application of oxalic acid or controlled incubations could provide another method of quantifying the ratio of exchangeable vs. stable MAOM (Jilling et al., 2018).

Other primary SOM fractions

Dissolved organic matter

Dissolved organic matter (DOM) accounts for only about 1–2% of the total organic matter within a soil at a given time and is procedurally defined as the organic molecules that are water soluble and smaller than 0.45 μm . Some methodologies enable the direct measurement of DOM by extractions or centrifugation, though dissolved organic carbon (DOC) is often calculated as the difference between bulk soil carbon and the cumulative carbon recovered after fractionation (Zimmermann et al., 2007). It plays an important role as both a microbial energy and nutrient source, and as a precursor to MAOM formation.

Heavy coarse organic matter

A heavy ($>1.6\text{--}1.85\text{ g cm}^{-3}$), coarse ($>60\text{--}50\text{ }\mu\text{m}$) organic matter fraction (hcOM) is isolated as another primary SOM component together with the POM and MAOM fractions when the fractionation scheme used includes both a size and a density separation (Mosier et al., 2021). This fraction has previously been termed both ‘heavy POM’ and ‘coarse MAOM’ and may consist of low molecular weight compounds sorbed to iron (Fe) or manganese (Mn) coatings on the outside of sand grains, microbial biofilms secreted on large primary particles, or POM encrusted with a mineral coating. The composition of this fraction varies depending on site characteristics, dispersion methods, and soil mineralogy. Although this fraction tends to account for a small percentage of the total soil organic carbon, it represents an important consideration when choosing a fractionation method. When size separations are used, it will become part of the POM fraction, but when density separations are used, it becomes part of the MAOM fraction. Ongoing research is attempting to characterize better this fraction and assign its functional role in the greater SOM pool, but currently it requires site-specific understanding to assess (Samson et al., 2020).

Soil aggregates

Cohesion and fragmentation processes in the bulk soil lead to the formation of soil aggregates, groups of primary particles (i.e., POM, hcOM, and MAOM) that are more strongly attached to each other than to other particles in the soil matrix. Aggregates vary in size and are often classified either as microaggregates (53–250 μm) or macroaggregates (250–8000 μm), depending on their diameter, or on the basis of their stability under applied stressors, such as repeated sieving or wetting and redrying. More stable aggregates retain their structure, and less stable ones disintegrate into the constituent primary fractions. Separating soils into different aggregate size and stability fractions helps our understanding of SOM structure and can give important insight into SOM formation and cycling processes. However, because of the composite nature of aggregates, they cannot be quantified together with and compared to the fractions they are composed of. Therefore, aggregates must be dispersed during the fractionation procedure to account for and compare the SOM content of primary fractions. Combined methods have been developed that introduce dispersion at different stages to isolate primary fractions that occur within different aggregate size fractions, which can give insight into primary fraction protection and controls on heterogeneity in primary fraction turnover time.

Other SOM fractions not isolated by physical fractionation methods

In contrast to the physically defined fractions described above, some SOM components have distinct chemical and functional properties, but lack distinct physical characteristics, and can be found distributed across physical fractions. Microbial biomass and pyrogenic organic matter are two such components. A small and ubiquitous SOM component, microbial biomass turns over

quickly, and its methodological quantification is beyond the scope of this chapter (see Horwath and Paul, 1994 for further discussion). Pyrogenic organic matter (PyOM) is present in most soils and results from the transformation of plant material into highly condensed aromatic rings by combustion at high temperatures. These aromatic structures are resistant to decomposition, requiring large amounts of energy to sever the chemical bonds. Even in systems that do not burn regularly, PyOM can be relatively abundant across the primary soil fractions, with potential to influence the calculation of mean residence time. This ubiquity indicates that PyOM should often be considered more fully in SOM studies (Lavallee et al., 2019). Given the strong persistence of PyOM and its distinct properties from POM and MAOM it would be advisable to separate it from these fractions; the combined use of hydrogen pyrolysis or Mid-IR spectroscopy with physical fractionation (Lavallee et al., 2019; Sanderman et al., 2021) hold promise for the feasibility of this approach at scale.

Aggregate fractionation methods

Various methods of aggregate fractionation have been proposed, primarily based on separation by size. Box 1 provides an in-depth overview of the aggregate separation scheme developed by Six et al. (1999), which has been identified as reproducible and consistent during multi-laboratory comparison studies (Poeplau et al., 2018). A key consideration during aggregate fractionation is the method by which aggregates are separated during fractionation. One method of aggregate separation and isolation is 'slaking', the process of rapidly immersing air-dried soil in water before rapidly removing it. Air pressure within the aggregate is increased upon rapid submersion, and low stability aggregates are disrupted. In contrast, rewetting of aggregates prior to fractionation by gradual capillary saturation to a water content of 105% of the bulk soil does not readily disrupt low stability aggregates, as the disruptive force applied to the soil is much less than when slaked. In laboratory comparisons of slaking vs. rewetting methods, more microaggregates were recovered in slaked soils than in rewetted ones, and macroaggregate recovery was increased in the rewetted soils (Six et al., 2000).

The process of separating aggregates is performed by sieving for a set amount of time. Both dry sieving of undispersed soils and wet sieving of air-dried (slaking) or sieving of rewetted soils have been used to separate aggregate size fractions. However, dry sieving has been shown to overestimate enzyme activity, while wet-sieving has been shown to indicate more significant differences between isolated fractions (Nahidan and Nourbakhsh, 2018). Following separation of aggregate size classes, dispersions and further fractionation can be performed to isolate primary fractions and separate the SOM fractions that are occluded within aggregates vs. suspended in the soil matrix (Six et al., 1998).

Primary particle fractionations

Dispersal methods

A key step during all physical primary particle fractionations is aggregate dispersal, for which several methods have been developed. The objective of all soil dispersal techniques is to disrupt the bonds between soil minerals such that the secondary aggregate structures are broken down into the constituent fractions. In practice, full dispersion is rarely if ever achieved; bonds in the microaggregate structures composed of silt and clay particles require tremendous energy to break, and as such some microaggregates will remain following the application of all dispersion methodologies. However, these fine aggregates (<53 µm) function similarly to organic matter bonded to silt and clay-size particles, and the application of enough energy to disrupt these fine aggregates would probably damage the primary minerals, leading to transfer of carbon between fractions. Thus, after conventional aggregate dispersion they are considered part of the MAOM fraction. A common side-effect of dispersal and liquid-mediated fractionation methods is the possibility of carbon transfer between fractions, which should be minimized at all steps of dispersal. Here we review the methodology for and the benefits and disadvantages of two of the most common types of soil dispersal methods: mechanical shaking and ultrasonic dispersion.

Mechanical shaking

Mechanical shaking methods of soil dispersal apply kinetic energy generated by sustained agitation to soils suspended in a liquid to fragment and disperse aggregates into smaller, primary particle fractions. Mechanical shaking can be performed in either a horizontal reciprocal shaker, which transfers energy to the soil solution by consistent movement back and forth along a single axis, or in an end-over-end shaker, which uses the falling of the soil solution to produce the energy necessary for dispersion. For reproducibility, shaking should be reported in terms of oscillations or rotations per minute, however, the shaking intensity is often reported qualitatively in the literature. Though to the authors' knowledge no studies have looked explicitly at shaking intensity, the consensus across studies indicates that shaking should be performed at a low-to-moderate speed (e.g., 110 oscillations per minute). The addition of glass beads increases the energy transfer from the shaker to the soil suspension and better disperses soils. There is a risk of fragmenting POM when using glass beads, however, and of puncturing vacuoles in vegetative tissue causing it to become smaller and denser, possibly leading to a small transfer of POM to the MAOM during fractionation (Diochon et al., 2016).

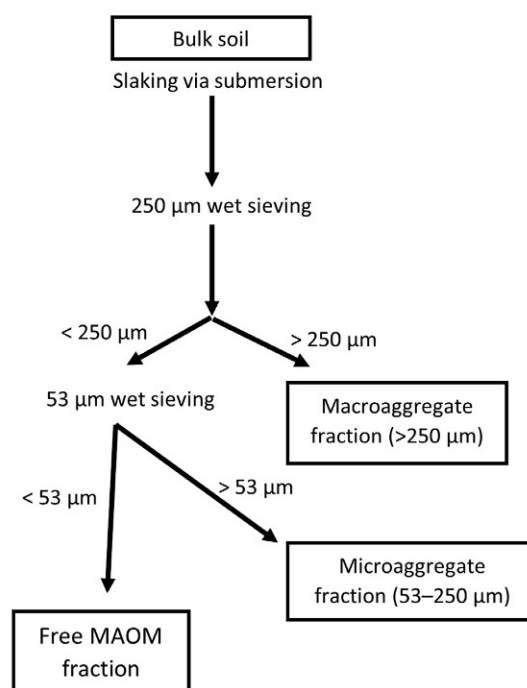
Several liquids have been used as a medium for mechanical dispersal, including water (H₂O), sodium hexametaphosphate ((NaPO₃)₆; SHMP), and sodium polytungstate (Na₆[H₂W₁₂O₄₀]; SPT). Sodium hexametaphosphate is the most commonly used soil dispersant. Its molecular structure allows for the formation of unassociated complexes with cations and reduces the flocculation

Box 1 Methods for physical fractionation of aggregates based on aggregate size.Methodology

- 1) Weigh out approximately 50 g of air dried, bulk soil that has been sieved to pass through a 2-mm sieve and record weight to the 0.01 g.
- 2) Place weighed soil on top of a 250- μ m sieve, and place the sieve into a large, clean container.
- 3) Fill the container with DI water until there is ~2 cm of liquid covering the sieve.
- 4) Allow sample to sit for 5 min in the water.
- 5) After 5 min, gently move the sieve up and down 50 times over the course of 2 min. Aggregates should just come out of the water on the upward stroke, and the sieve should not hit the bottom on the downward stroke.
- 6) Remove sieve from water and backwash aggregates that remain on top of the 250- μ m mesh into a pre-weighed container.
- 7) Place a 53- μ m sieve in a second, large, clean container, and then gently pour the effluent from the first container over the sieve.
- 8) Repeat step 5, and then backwash the aggregates that remain on the 53- μ m sieve into a pre-weighed container.
- 9) Transfer effluent to 1-L centrifuge bottles, and centrifuge for 15 min at 3220 RCF.
- 10) Carefully pour off supernatant, and transfer silt and clay in bottles to pre-weighed container.
- 11) Place tins in oven to dry at 50 °C until all water has evaporated, and then weigh. The first sample collected is the macroaggregate fraction, the second the microaggregate fraction, and the third the silt and clay fraction.

Materials Required

- 40–50 g soil sieved < 2 mm
- 1 250- μ m sieve
- 1 53- μ m sieve
- DI water
- 2 clean large pans (e.g., plastic oil pans)
- 1-L centrifuge bottles
- 3 large, pre-weighed sample tins



of clay minerals. However, the use of SHMP as a dispersant for SOM fractionation techniques can add a level of complexity and error to the results. Therefore, additions of SHMP to the soil must be either rinsed out of the soil by repeated washing with DI water or maintained in the sample throughout the analysis. Rinsing of the soils can lead to SOC leaching from fractions, compromising both total carbon recovery and the accuracy of results. Maintaining SHMP in the organic matter fractions can also lead to artificially inflated recoveries when examined on a mass basis, possibly obscuring loss during the fractionation process that could indicate poor recovery.

Shaking in water avoids the potential to change the chemical character of SOM fractions or add mass that may skew recoveries. A recent study indicated poor reproducibility with water extractions alone, however, as without a complementary dispersal additive water may not have the power to disrupt aggregates fully (Just et al., 2021). Mechanical shaking can also be performed in SPT, which can be convenient when SPT is used in further fractionation steps. However, similar to SHMP, SPT must be rinsed from samples

prior to further analysis. Reviews of rinsing procedures have demonstrated leaching of carbon from all fractions may occur, leading to underestimation of organic carbon across fractions (Plaza et al., 2019). In addition, a review of SPT has shown that it can contaminate samples with non-native nitrogen (N) and tungsten (W), especially in the mineral-associated fraction (Kramer et al., 2009). Mechanical shaking is a good method to use across a variety of soils since its dispersal power is less dependent on soil type than the alternative ultrasonic vibration presented below. For any method, the operator should use it with care and adjust shaking time and intensity, number of beads and rinsing of dispersal agent to minimize drawbacks.

Ultrasonic vibration

The second common method of soil dispersion is sonication, in which ultrasonic vibrations are applied to soil suspensions, disrupting bonds, and separating soil components. The general application of this method introduces ultrasonic waves into a solution containing non-dispersed soils. As the waves travel through the solution, alternating pressure conditions are created. Small bubbles are formed in low pressure areas, which then implode as the bubbles encounter high pressure areas. This implosion, and the force associated with it, disperses the secondary structures as the water jet that is created can have velocities that exceed 100 m sec^{-1} . One of the central benefits of this method is that it provides quantitative values of the amount of force used and is therefore theoretically reproducible. Soils with a high clay content create tightly bound microaggregates, which require a higher dispersion energy than more coarse soils. Therefore, the amount of force needed for proper dispersal varies between soil textures, mineralogies, and structures, and must be calibrated prior to experimentation. The amount of sonication energy used to break aggregates from secondary into primary structures spans a broad range. In a review, Kaiser et al. (2010) report sonication energies used for dispersal ranging from 60 to $10,000 \text{ J cm}^{-3}$. Researchers who work consistently with the same soils may benefit from using the calibrated values provided by sonication methods, large scale projects that encompass a variety of global soils with variable physiochemical characteristics will require approaches that are more consistent across samples.

Carbon transfer between fractions can occur when using sonication dispersal techniques as well. At the POM level, sonication can significantly fragment structural litter particles, leading to redistribution among all size fractions. Using energy levels between 280 and 420 J cm^{-3} , Oorts et al. (2005) found that there was a transfer of 35–37% of the nitrogen and of 20–31% of the carbon from coarse POM into smaller size fractions following dispersal. However, this damage and subsequent carbon transfer has been found at energy levels as low as 90 J cm^{-3} (Amelung and Zech, 1999). This transfer can have important implications depending on ecosystem type and experimental goals. Even in systems where the coarse POM fraction is relatively small, the fragmentation and redistribution of this fraction has large implications for isotopic studies. Primary SOM fractions, because of differences in their formation, degree of microbial processing, and persistence, often have distinct isotopic values. As such, isotopic techniques such as ^{14}C that quantify age and turnover time, or ^{13}C and ^{15}N that quantify origin and degree of transformation can be considerably compromised by the transfer of even small amounts of organic matter among fractions.

For soil particles, excessive sonication energy can potentially damage the crystalline structure of minerals, resulting in bias in the recovery of different SOM fractions based on size. Several studies have documented fragmentation of sand separates into smaller fractions, especially as sonication energy exceeds 700 kJ cm^{-3} . Additional evidence of damage to silt particles has also been observed with scanning electron microscope (SEM) techniques (Kaiser et al., 2012). Currently, there is no evidence that sonication can fragment particles smaller than $20 \mu\text{m}$ (i.e., the clay fraction). However, the energy required to disperse small clay aggregates may exceed the energy that could fragment larger particles, such that these two extremes must be balanced. Calibration of the sonication energy is an important step in the dispersal process for optimizing aggregate dispersal while minimizing the amount of carbon transfer between particle-size classes and SOM fractions.

Size fractionations

Several fractionation methods that use particle size as a differentiating factor between primary fractions have been established. These methods aim to separate primary SOM fractions after dispersal by sieving with different mesh sizes (Fig. 2). Typically, sizes ranging between 53 and $60 \mu\text{m}$ are used to separate POM ($>53 \mu\text{m}$) from MAOM, but a cutoff at $20 \mu\text{m}$ to separate the fine MAOM is also used. In the review of fractionation methods by Poeplau et al. (2018), particle-size fractionation was found to be an effective way of isolating fractions that differed in their turnover times. The methodology of the size fractionation technique deployed by Cotrufo et al. (2019) is shown in Box 2.

Size fractionations are relatively simple and tend to have limited mass loss during the procedure. Furthermore, they require a low level of skill and laboratory infrastructure, are cost effective, and throughput can be relatively high ($25\text{--}30$ samples day^{-1}). Size fractionations have been used in several studies as an effective method of separating two functionally distinct SOM pools, i.e., POM and MAOM. They provide important insight about the formation and persistence of SOM in response to treatment. Those studies that have a large number of samples, or that expect a moderate to large size effect in response to treatment in either the MAOM or POM fraction are well suited for size fractionation methods. However, size fractionations can be limited in scope depending on the research objectives of a given study. Size fractionation methodologies are unable to separate occluded POM from free POM and cannot be used to isolate the hcOM fraction. Further, the DOM fraction will be grouped with the MAOM fraction if not isolated prior to the size fractionation.

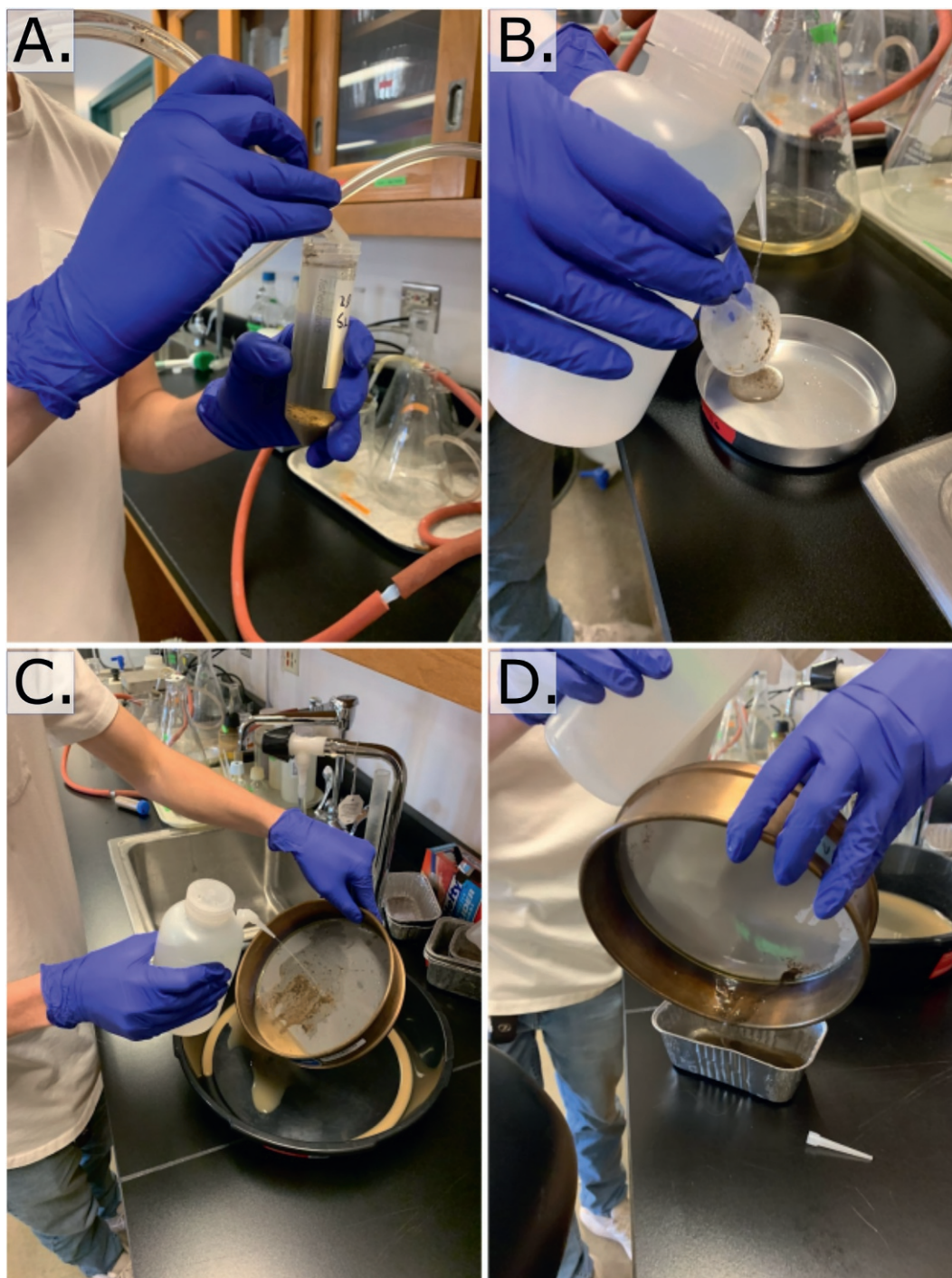


Fig. 2 Demonstrations of fractionation techniques and examples of equipment used. (A) Aspiration of the light fraction floated in SPT using vacuum filtration. (B) Transfer of light fraction from nylon filter to pre-weighed pan. (C) Wet sieving to separate MAOM from hcOM (combined fractionation) or POM (size fractionation). (D) Back washing of hcOM (combined fractionation) or POM (size fractionation) fraction into pre-weighed pans from after completion of sieving.

Density separations

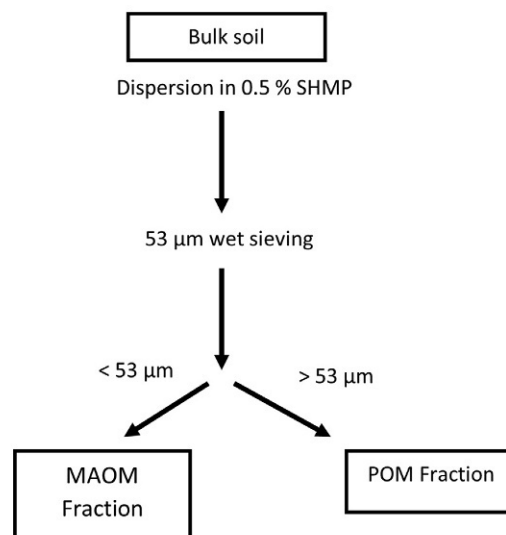
An alternative method to size fractionations is to take advantage of the differences in density of mineral particles and partially decomposed litter fragments. Organic matter has a lower density than minerals, and when submerged and centrifuged in a liquid of a known density, it floats while the mineral-associated portion of the sample sinks. The lighter fraction can then be aspirated and analyzed separately from the heavier fraction (Fig. 2). Sequential density fractionations combined with dispersion can also be used to separate the free POM from the aggregate-protected POM. In this case, the soil sample is initially subjected to little to no

Box 2 Methods for physical fractionation of POM and MAOM based on primary particle size.Methodology

- 1) Weigh out 5.25–5.75 g of air dried, bulk soil that has been sieved to pass through a 2-mm sieve into a pre-weighed centrifuge tube, and place in 60 °C oven overnight.
- 2) Take samples out of oven, cool in desiccator, and weigh samples + tubes to 0.01 g.
- 3) Add 12 small glass beads to the tube, then add 0.5% sodium hexametaphosphate to the 30 mL line and cap samples.
- 4) Shake the samples for 18 hours $\pm$ 15 min at $\sim$ 112 oscillations per minute.
- 5) Place the 2000- μ m sieve on top of the 53- μ m sieve in a clean, large shallow pan. Shake sample to ensure all soil is in solution, then pour sample over the nested sieve.
- 6) Retrieve glass beads from 2000- μ m sieve and wash any soil material onto the 53- μ m sieve.
- 7) Rinse the 53- μ m sieve with DI water until rinsing water runs clear, making sure also to rinse the sides of the sieve.
- 8) Back rinse the material remaining on top of the 53- μ m sieve into a pre-weighed weighing pan and place in the oven at 60 °C.
- 9) Transfer the effluent from the large shallow pan into a pre-weighed weighing pan and place in oven at °C.
- 10) Once samples are completely dry, weigh and ensure that recoveries fall between 95% and 105% of sample mass. If recoveries are large, re-dry and reweigh samples.
- 11) If recoveries are acceptable, samples can be ground using a mortar and pestle and prepared for further analysis.

Materials Required

- 50 mL conical centrifuge tubes
- 0.5% sodium hexametaphosphate
- Pre-weighed weighing pans
- 2000 and 53- μ m sieves.
- DI water
- Large, shallow pan



dispersion prior to density flotation and collection of the free POM. The sample is then more thoroughly dispersed, releasing low-density material that was occluded in aggregate structures, and the occluded POM can be separated by a second density flotation.

The most common solution used in density fractionations is SPT and, as mentioned above, some procedures also call for its use as a dispersing agent. Previous studies have also employed sodium iodide (NaI) as the heavy liquid for density fractionations, but found that the use of NaI may increase the pH of the soil solution, which has the potential to disrupt organo-mineral bonds and facilitate intra-fraction C transfer (Plaza *et al.*, 2019). By changing the amount of water added to heavy liquid, the density can be manipulated to achieve different fractionation outcomes. Previous research by Cerli *et al.* (2012) highlighted the differences in recovery of the light fraction across different densities. Their results suggest that a density of 1.6 g cm⁻³ is the optimal density for achieving maximum floatation of organic material and to minimize the presence of minerals in the light fraction. Most studies use densities of 1.8–1.85 g cm⁻³ to maximize recovery of organic material not associated with minerals and to improve floatation. Even

higher densities have been suggested but the risk of contamination by mineral particles increases as density increases. The ultimate density chosen for a given study should be as close to the maximum density of the organic materials present as possible, while minimizing potential mineral contamination of the light fraction. Calibration for individual sites may be necessary.

Sequential density fractionations, with progressively increasing densities, on dispersed soils have also been used to separate mineral-associated fractions with distinct mineralogical components after the removal of the light fraction, as demonstrated by Sollins et al. (2009). By increasing the density of the suspension liquid with each successive fractionation, it is possible to isolate a range of different MAOM fractions, including short-range order minerals such as allophanes (density: $>2.0 \text{ g cm}^{-3}$), high activity phyllosilicates such as smectite (density: $>2.3 \text{ g cm}^{-3}$), framework silicates such as quartz and feldspars (density: $>2.7 \text{ g cm}^{-3}$), and oxides such as hematite and goethite (density: $<3.0 \text{ g cm}^{-3}$). This increased separation of the MAOM fraction can give further insight into the controlling factors on SOM persistence and formation across different soil environments. However, with each additional density flotation, the risk of decreasing recovery or carbon transfer between fractions becomes greater. Further, density fractionations are time intensive, as SPT must be rinsed from the samples several times during each step, expensive, and require greater laboratory skill to execute compared to size fractionations.

Overall, density fractionations are a useful tool for separating organic and mineral particles that may not differ in size but have other contrasting properties. Studies that are focused on the differences between formation pathways and cycling of POM will benefit from density separations, as it enables explicit isolation of the most plant derived fraction, in contrast with size separations. Studies that concern changes in management or land use in particular may also benefit from density fractionations because the light fraction is the most rapidly altered by shifts in inputs and disturbance.

Combined separations

Aggregate, size, and density fractionations may be combined to separate the key functional organic matter pools and elucidate mechanisms of SOM formation and stabilization. These fractionation procedures allow for increased isolation of pools that may be grouped together in size or density fractions, such as hcOM. In addition, when size and density are combined with aggregate fractionation, they enable a greater degree of differentiation of the organic carbon that is either occluded within aggregates in mineral and particulate forms, or free in the soil matrix (Six et al., 1999). These characteristics of the combined size-density fractionations provide a scheme that can approach the fractions identified more closely in the conceptual framework than size or density fractionations alone. In the comparison study by Poeplau et al. (2018), fractionation methods that combine a size with a density separation following complete dispersion were very effective in differentiating fractions with different turnover times, had average mass recoveries greater than 95%, and were reproducible across multiple fractionations. A typical scheme combining size and density fractionation schemes isolated four fractions: DOM, IPOM, hcOM, and MAOM (Mosier et al., 2021; Box 3). As with the density fractionation methods, further sequential density separations could be done to isolate particular mineralogical constituents. Similarly, further size fractionations could also be performed to isolate particular subclasses of particles (e.g., fine clay vs. silt, etc.). However, adding successive density or size cutoffs can decrease the overall efficacy of the fractionations; the risk of leaching carbon from the samples or transferring between fractions increases with the addition of further washing, shaking, or centrifugation steps (Poeplau et al., 2018).

Chemical-free physical fractionations

There has been some recent research towards the development of chemical free physical fractionation methods because of the complications that can arise with chemical dispersants and density floatation media (Kaiser et al., 2010). Fractionation methods that immerse soils in any liquid can lead to redistribution of carbon between fractions, therefore, these methods present an alternative to the use of SHMP or SPT, possibly reducing the risk of carbon redistribution. This method uses a series of nested sieves, separating particles into six particle classes, ranging from 2 to $<0.315 \text{ mm}$. Soils are sieved prior to analysis, but not otherwise dispersed. A glass petri dish is then charged by rubbing it with cotton, and each particle class is placed separately on a clean stone plate. The charged petri dish is then placed 2–5 cm above the stone plate and moved slowly over the sample. The distance between the charged surface and the size fractions must be calibrated manually for each sample to avoid the attraction of mineral particles. Following the attraction step, organic fractions must be examined visually for mineral contamination, and mineral particles removed. This procedure is then repeated five times for each fraction.

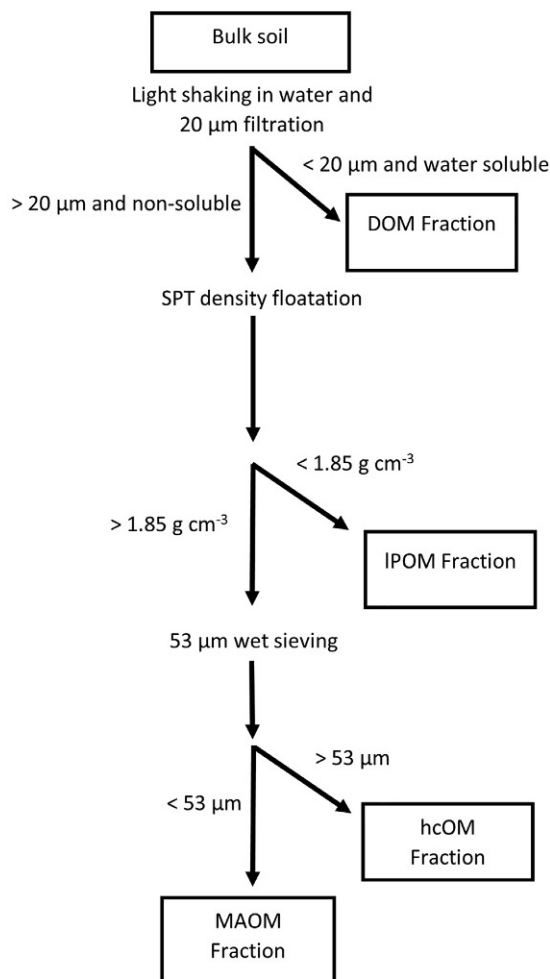
When tested across several arable and forest soils, this method resulted in considerable variation in the amount of organic matter separated from each of the sieved mineral fractions (CV: 0.3–174%; Kaiser et al., 2009). However, when examined across mineral fractions, the variability accorded with the reproducibility reported for size and density fractions. Taken in tandem, these measures indicate a useful methodology for the separation of POM and MAOM, but may be limited in their ability to separate different POM fractions (i.e., occluded vs. free POM). These methods are not as widely tested as those presented in previous sections but represent an interesting and important line of inquiry that should be further explored by research groups. Given the potential drawbacks associated with sonication, shaking, and the use of SPT, non-aqueous non-dispersive methods may provide increased insight into SOM pools moving forward, as well as the role of SOM in facilitating N exchange. However, the time needed per sample, together with the requirement of manual calibration during each successive pass with the charged petri dish makes these methods cumbersome when dealing with large sample sizes.

Box 3 Methods for combined density and size fractionation of primary SOM fractions.Methodology

- 1) Weight out 5.5–6 g air dried, 2-mm sieved soils into a pre-weighed, acid washed centrifuge tube, and place tubes in 60 °C oven overnight.
- 2) The following day, weigh samples on a 4-place scale, and then add DI water to the 35 mL line and shake for 15 min on low (~112 oscillations per min).
- 3) Rinse tubes and caps with DI water, and then balance tubes and centrifuge at 2325 RCF for 15 min.
- 4) Pour supernatant over a 20- μ m nylon filter on a filtration unit, and transfer filtered liquid to a clean, new centrifuge tube.
- 5) With a squirt bottle, rinse any light fraction that accumulated on the filter into a small, pre-weighed pan.
- 6) Fill centrifuge tubes to 25 mL with 1.85 SPT and add 12 glass beads. Resuspend soil pellet and shake on low for 18 h.
- 7) After shaking, rinse tube cap and side with SPT, and then balance tubes and centrifuge at 2325 RCF for 30 min.
- 8) Using a vacuum filter, aspirate the light fraction and SP T over a 20- μ m nylon filter.
- 9) Transfer light fraction from filter to pre-weighed pans using a squirt bottle, and place pans in the 60 °C oven to dry.
- 10) Add DI water to the centrifuge tube up to the 30-mL line and resuspend the pellet. Rinse cap and sides, then balance and centrifuge for 20 min at 2325 RCF.
- 11) Pour supernatant off, and repeat the processes, adding water, suspending the pellet, and then centrifuging for 20 min at 2325 RCF. Repeat rinsing process one more time.
- 12) After rinsing, add water to 20 mL line, suspend pellet, and separate fractions by size, as described in steps 5–11 in **Box 2**.

Materials Required

- Pre-weighed 50-mL conical centrifuge tubes
- 0.5% sodium hexametaphosphate
- Pre-weighed weighing pans
- 2000-and 53- μ m sieves
- DI water
- Large, shallow pan
- Centrifuge
- Oscillating shaker
- 4-mm beads
- Sodium polytungstate
- Filtration unit
- 20- μ m nylon filter



Conclusions

Physical fractionation methods provide a means for researchers to isolate functionally distinct fractions of SOM. Unlike chemical and biological fractionation methods, physical fractionation methods take advantage of the inherent properties of both soil particles and organic material to separate SOM with different formation pathways, mechanisms of stabilization, and thus responsiveness to changes and overall persistence. Here we have reviewed several different possible methods for physical fractionations, including those based on particle size, particle density, and aggregate size. All have their pros and cons, and recovery of fraction mass and carbon (or nitrogen) need to be quantified at the end and reported. The methods examined provide a good foundation for functionally meaningful SOM fractionation, and we hope will continue to be improved and expanded upon by researchers moving forward. There are still important research questions that remain unanswered, such as the relationship between efficacy of the fractionation method and soil physiochemical properties, and if what we separate today as ‘primary’ fractions are indeed unique functional units (e.g., POM and MAOM), or are made of contrasting functional structures requiring further separation (e.g., exchangeable and stable MAOM). Continuation of this research will improve the tools at our collective disposal and advance this field further.

Acknowledgment

The development and writing of this chapter was supported by the United States Department of Agriculture National Institute of Food and Agriculture, Award Number 2021-67019-34241. The authors would also like to thank the members of the Cotrufo Soil Innovation Lab and the Natural Resource Ecology Lab at Colorado State University, whose contributions to the development and implementation of the methods were essential to the information presented here.

References

- Amelung W and Zech W (1999) Minimisation of organic matter disruption during particle-size fractionation of grassland epipedons. *Geoderma* 92: 73–85. [https://doi.org/10.1016/S0016-7061\(99\)00023-3](https://doi.org/10.1016/S0016-7061(99)00023-3).
- Cambardella CA and Elliott ET (1992) Particulate soil organic-matter changes across a grassland cultivation sequence. *Soil Science Society of America Journal* 56: 777–783. <https://doi.org/10.2136/sssaj1992.03615995005600030017x>.
- Cerli C, Celi L, Kalbitz K, Guggenberger G, and Kaiser K (2012) Separation of light and heavy organic matter fractions in soil—Testing for proper density cut-off and dispersion level. *Geoderma* 170: 403–416. <https://doi.org/10.1016/j.geoderma.2011.10.009>.
- Christensen BT (2001) Physical fractionation of soil and structural and functional complexity in organic matter turnover. *European Journal of Soil Science* 52: 345–353. <https://doi.org/10.1046/j.1365-2389.2001.00417.x>.
- Cotrufo MF, Ranalli MG, Haddix ML, Six J, and Lugato E (2019) Soil carbon storage informed by particulate and mineral-associated organic matter. *Nature Geoscience* 12: 989–994. <https://doi.org/10.1038/s41561-019-0484-6>.
- Denef K, Plante AF, and Six J (2010) Characterization of soil organic matter. In: Kutsch WL, Bahn M, and Heinemeyer A (eds.) *Soil Carbon Dynamics: An Integrated Methodology*, pp. 91–126. Cambridge University Press. <https://doi.org/10.1017/CBO9780511711794.007>.
- Dioc'hon A, Gillespie AW, Ellert BH, Janzen HH, and Gregorich EG (2016) Recovery and dynamics of decomposing plant residue in soil: An evaluation of three fractionation methods. *European Journal of Soil Science* 67: 196–205. <https://doi.org/10.1111/ejss.12316>.
- Horwath WR and Paul EA (1994) Microbial biomass. In: Weaver R, Angle S, Bottomley P, Bezdicek D, Smith S, Tabatabai A, and Wollum A (eds.) *Methods of Soil Analysis: Part 2 Microbiological and Biochemical Properties*, pp. 753–773. Madison, WI: Soil Science Society of America. <https://doi.org/10.2136/sssabookser5.2.c36>.
- Jilling A, Keiluweit M, Costanta AR, Frey S, Schimel J, Schneckner J, Smith RG, Tiemann L, and Grandy AS (2018) Minerals in the rhizosphere: Overlooked mediators of soil nitrogen availability to plants and microbes. *Biogeochemistry* 139: 103–122. <https://doi.org/10.1007/s10533-018-0459-5>.
- Just C, Poeplau C, Don A, van Wesemael B, Kögel-Knabner I, and Wiesmeier M (2021) A simple approach to isolate slow and fast cycling organic carbon fractions in central european soils—Importance of dispersion method. *Frontiers in Soil Science* 1: 1–14. <https://doi.org/10.3389/fsoil.2021.692583>.
- Kaiser M, Berhe A, Sommer M, and Kleber M (2012) Application of ultrasound to disperse soil aggregates of high mechanical stability. *Journal of Plant Nutrition and Soil Science* 175(4): 521–526. <https://doi.org/10.1002/jpln.201200077>.
- Kaiser M, Ellerbrock RH, and Sommer M (2009) Separation of coarse organic particles from bulk surface soil samples by electrostatic attraction. *Soil Science Society of America Journal* 73(6): 2118–2130. <https://doi.org/10.2136/sssaj2009.0046>.
- Kaiser M, Wirth S, Ellerbrock RH, and Sommer M (2010) Microbial respiration activities related to sequentially separated, particulate and water-soluble organic matter fractions from arable and forest topsoils. *Soil Biology and Biochemistry* 42: 418–428. <https://doi.org/10.1016/j.soilbio.2009.11.018>.
- Kramer MG, Lajtha K, Thomas G, and Sollins P (2009) Contamination effects on soil density fractions from high N or C content sodium polytungstate. *Biogeochemistry* 92: 177–181. <https://doi.org/10.1007/s10533-008-9268-6>.
- Lavallee JM, Conant RT, Haddix ML, Follett RF, Bird MI, and Paul EA (2019) Selective preservation of pyrogenic carbon across soil organic matter fractions and its influence on calculations of carbon mean residence times. *Geoderma* 354: 113866. <https://doi.org/10.1016/j.geoderma.2019.07.024>.
- Lavallee JM, Soong JL, and Cotrufo MF (2020) Conceptualizing soil organic matter into particulate and mineral-associated forms to address global change in the 21st century. *Global Change Biology* 26: 261–273. <https://doi.org/10.1111/gcb.14859>.
- Lehmann J and Kleber M (2015) The contentious nature of soil organic matter. *Nature* 528: 60–68. <https://doi.org/10.1038/nature16069>.
- Liang C, Schimel JP, and Jastrow JD (2017) The importance of anabolism in microbial control over soil carbon storage. *Nature Microbiology* 2: 1–6. <https://doi.org/10.1038/nmicrobiol.2017.105>.
- Mosier S, Apfelbaum S, Byck P, Calderton F, Teague R, Thompson R, and Cotrufo F (2021) Adaptive multi-paddock grazing increases soil carbon and nitrogen stocks and stabilisation through mineral association in southeastern U.S. grazing lands. *Global Change Biology* 288.
- Nahidan S and Nourbakhsh F (2018) Distribution pattern of amidohydrolase activities among soil aggregates: Effect of soil aggregates isolation methods. *Applied Soil Ecology* 125: 250–256. <https://doi.org/10.1016/j.apsoil.2018.02.004>.
- Oades JM and Waters AG (1991) Aggregate hierarchy in soils. *Australian Journal of Soil Research* 29: 815–825. <https://doi.org/10.1071/SR9910815>.

- Oorts K, Vanlauwe B, Recous S, and Merckx R (2005) Redistribution of particulate organic matter during ultrasonic dispersion of highly weathered soils. *European Journal of Soil Science* 56: 77–91. <https://doi.org/10.1111/j.1351-0754.2004.00654.x>.
- Paul EA, Harris D, Collins HP, Schulthess U, and Robertson GP (1999) Evolution of CO₂ and soil carbon dynamics in biologically managed, row-crop agroecosystems. *Applied Soil Ecology* 11: 53–65. [https://doi.org/10.1016/S0929-1393\(98\)00130-9](https://doi.org/10.1016/S0929-1393(98)00130-9).
- Plaza C, Giannetta B, Benavente I, Vischetti C, and Zaccone C (2019) Density-based fractionation of soil organic matter: Effects of heavy liquid and heavy fraction washing. *Scientific Reports* 9: 1–7. <https://doi.org/10.1038/s41598-019-46577-y>.
- Poeplau C, Don A, Six J, Kaiser M, Benbi D, Chenu C, Cotrufo MF, Derrien D, Gioacchini P, Grand S, Gregorich E, Griepentrog M, Gunina A, Haddix M, Kuzakov Y, Kühnel A, Macdonald LM, Soong J, Trigalet S, Vermeire ML, Rovira P, van Wesemael B, Wiesmeier M, Yeasmin S, Yevdokimov I, and Nieder R (2018) Isolating organic carbon fractions with varying turnover rates in temperate agricultural soils—A comprehensive method comparison. *Soil Biology and Biochemistry* 125: 10–26. <https://doi.org/10.1016/j.soilbio.2018.06.025>.
- Ramnarine R, Voroney RP, Dunfield KE, and Wagner-Riddle C (2018) Characterization of the heavy, hydrolysable and non-hydrolysable fractions of soil organic carbon in conventional and no-tillage soils. *Soil and Tillage Research* 181: 144–151. <https://doi.org/10.1016/j.still.2018.04.010>.
- Samson MÉ, Chantigny MH, Vanasse A, Menasseri-Aubry S, and Angers DA (2020) Coarse mineral-associated organic matter is a pivotal fraction for SOM formation and is sensitive to the quality of organic inputs. *Soil Biology and Biochemistry* 149. <https://doi.org/10.1016/j.soilbio.2020.107935>.
- Sanderman J, Baldock JA, Dangal SRS, Ludwig S, Potter S, Rivard C, and Savage K (2021) Soil organic carbon fractions in the Great Plains of the United States: An application of mid-infrared spectroscopy. *Biogeochemistry* 156: 97–114. <https://doi.org/10.1007/s10533-021-00755-1>.
- Six J and Paustian K (2014) Aggregate-associated soil organic matter as an ecosystem property and a measurement tool. *Soil Biology and Biochemistry* 68: A4. <https://doi.org/10.1016/j.soilbio.2013.06.014>.
- Six J, Elliott ET, Paustian K, and Doran JW (1998) Aggregation and soil organic matter accumulation in cultivated and native grassland soils. *Soil Science Society of America Journal* 62: 1367–1377. <https://doi.org/10.2136/sssaj1998.03615995006200050032x>.
- Six J, Elliott ET, and Paustian K (1999) Aggregate and soil organic matter dynamics under conventional and no-tillage systems. *Soil Science Society of America Journal* 63: 1350–1358. <https://doi.org/10.2136/sssaj1999.6351350x>.
- Six J, Paustian K, Elliott ET, and Combrink C (2000) Soil structure and organic matter I. Distribution of aggregate-size classes and aggregate-associated carbon. *Soil Science Society of America Journal* 64: 681–689. <https://doi.org/10.2136/sssaj2000.642681x>.
- Sollins P, Spycher G, and Glassman CA (1984) Net nitrogen mineralization from light- and heavy-fraction forest soil organic matter. *Soil Biology and Biochemistry* 16: 31–37. [https://doi.org/10.1016/0038-0717\(84\)90122-6](https://doi.org/10.1016/0038-0717(84)90122-6).
- Sollins P, Kramer MG, Swanston C, Lajtha K, Filley T, Aufdenkampe AK, Wagai R, and Bowden RD (2009) Sequential density fractionation across soils of contrasting mineralogy: Evidence for both microbial- and mineral-controlled soil organic matter stabilization. *Biogeochemistry* 96: 209–231. <https://doi.org/10.1007/s10533-009-9359-z>.
- Stevenson FJ (1994) *Humus Chemistry: Genesis, Composition, Reactions*, 2nd edn. New York, NY: John Wiley and Sons Ltd.
- von Lütow M, Kögel-Knabner I, Ekschmitt K, Flessa H, Guggenberger G, Matzner E, and Marschner B (2007) SOM fractionation methods: Relevance to functional pools and to stabilization mechanisms. *Soil Biology and Biochemistry* 39: 2183–2207. <https://doi.org/10.1016/j.soilbio.2007.03.007>.
- Witzgall K, Vidal A, Schubert DI, Höschen C, Schweizer SA, Buegger F, Pouteau V, Chenu C, and Mueller CW (2021) Particulate organic matter as a functional soil component for persistent soil organic carbon. *Nature Communications* 12: 4115. <https://doi.org/10.1038/s41467-021-24192-8>.
- Zimmermann M, Leifeld J, Schmidt MWI, Smith P, and Fuhrer J (2007) Measured soil organic matter fractions can be related to pools in the RothC model. *European Journal of Soil Science* 58: 658–667. <https://doi.org/10.1111/j.1365-2389.2006.00855.x>.

The use of stable isotopes in soil science: Low atomic number elements

Chiara Pistocchi^a, Isabelle Bertrand^a, Kristýna Kantnerová^b, Astrid Oberson^c, Federica Tamburini^c, and Roland A Werner^d, ^aUMR Eco&Sols, INRAE, Montpellier, France; ^bUniversity of Colorado Boulder, Geological Sciences & INSTAAR, Boulder, CO, United States; ^cInstitute of Agricultural Sciences, ETH Zurich, Lindau, Switzerland; ^dDepartment of Environmental Systems Science, ETH Zurich, Zurich, Switzerland

© 2023 Elsevier Ltd. All rights reserved.

Introduction	82
Overview and terminology	82
Why are stable isotopes useful in soil science studies?	82
How to express abundances of isotopes at natural abundance and in enriched samples, and how to express the isotopic fractionation	83
Isotope effect and isotopic fractionation	83
Analytical methods	84
Natural abundance and labelling approaches	86
Application of isotopes of light elements in soil science	86
Carbon	86
Natural ¹³ C abundance to study SOM dynamics in the soil of arable lands	87
Labelling approaches to study SOM in cropped soils	87
Nitrogen	88
The use of ¹⁵ N enriched fertilizers	88
The use of ¹⁵ N at natural abundance	89
Oxygen	89
<i>Oxygen isotopes in the soil water–plant–atmosphere continuum</i>	89
Oxygen in phosphate to study P cycling	90
Oxygen in nitrate and sulfate	90
Sulfur	91
Current trends and future developments	91
Site-specific ¹⁵ N analyses and clumped isotopes to study N ₂ O emissions	92
Isotopic analysis of oxyanions, bulk isotopes composition of total dissolved C and N and of soil carbon pools	92
References	92

Abstract

In this chapter, we focus on low atomic mass stable isotopes that are widely used in soil and environmental studies. These include carbon (C), nitrogen (N), oxygen (O), and sulfur (S). We elucidate how such stable isotopes are used to infer processes, estimate mean residence time of an element in an environmental compartment, and identify sources or relative contribution to a pool. In addition, we describe the basis of the most widespread analytical methods for isotope analyses and report relevant technical advances, which are expected to provide unique insight into biogeochemistry of key elements of life in the near future.

Glossary

Compartment An assemblage of atoms that is not necessarily chemically homogeneous, for example, nitrogen in soil organic matter or in the plant.

Labelling Artificial enrichment or depletion of an isotope in a specific compound or compartment (e.g. the plant).

Moiety In organic chemistry, a part of a molecule.

Pool A compartment containing material that is chemically indistinguishable and equally accessible to plants or to the soil organisms; e.g. soil ammonium pool.

Tracer Substance with atomic or nuclear, physical, chemical or biological properties that enables us to identify, observe or follow various physical, chemical or biological processes.

Key points

- Relevance of stable isotopes of light elements in soil science studies
- Main terminology, notations, and approaches in isotope studies
- The main measurement techniques for isotope ratios
- Relevant examples of application of low atomic mass isotopes (C, N, O, S) in soil science
- Current trends and future development in the application of C, N, O, S isotopes

Introduction

Overview and terminology

Isotopes are atoms of the same element with the same number of protons in the nucleus but different number of neutrons. Therefore, the atomic mass of isotopes of a same element differs by one or more units. The most common isotopes are generally the lightest and are much more abundant (>95 atom percent) than the heavier ones (Table 1).

Isotopes of the same element have qualitatively the same chemical properties, because they have the same number and arrangement of electrons. However, the replacement of an element with one of its isotopes determines slightly different physico-chemical behavior, expressed in changes of chemical reaction rates or equilibrium constants. These differences are called ‘isotope effects’, and are proportional to the relative mass difference between isotopes for most processes. Mass-independent isotope effects of O and S can take place e.g. in the stratosphere, and have proven to be very useful to trace atmospheric inputs to terrestrial ecosystems (see Section “Oxygen in nitrate and sulfate”). For light elements (Table 1), the isotope effects are greater because their difference in mass is proportionally bigger. Such partitioning of isotopes translates into detectable variation in isotopic composition of molecules, chemical species, and phases (e.g. soil water vs water vapor). Naturally occurring or artificially induced variations in isotopic composition of molecules can provide insight into processes affecting environmental compartments or pools.

Molecules that differ only in their isotopic composition or the intramolecular position of the isotopes but having the same chemical formula are generally called isotopocules (isotopically substituted molecules). Molecules that differ only in isotopic composition are called isotopologs (isotopic analogs), and often present distinct masses. Some molecules, such as nitrous oxide (N₂O), include isobaric isotopologs either having the same nominal mass, e.g. ¹⁴N¹⁵N¹⁶O and ¹⁴N¹⁴N¹⁷O, or the same exact mass, e.g. ¹⁴N¹⁵N¹⁶O and ¹⁵N¹⁴N¹⁶O. When isotopocules differ only in the position of the isotope but the masses are the same, they are called isotopomers (isotopic isomers; IUPAC, 1997). An excellent reference for the correct use of terminology is given in Sharp (2017, Table 2.1).

Why are stable isotopes useful in soil science studies?

Low atomic mass stable isotopes of carbon (C), nitrogen (N), oxygen (O), and sulfur (S) are used to assess biotic and abiotic processes that characterize the biogeochemical cycles of such elements. Less commonly used light elements are hydrogen, lithium, boron, silicon, magnesium, and chlorine.

Low atomic mass stable isotopes are used to (i) infer processes or process rates, e.g. heterotrophic nitrification (see Section “Nitrogen”), (ii) estimate mean residence time of an element in an environmental compartment or pool, such as carbon in soil organic matter (see Section “Carbon”), or (iii) identify sources or relative contribution to a pool, as in studies of soil erosion or atmospheric deposition (see Sections “Nitrogen” and “Sulfur”). Isotopes are particularly useful to trace fluxes of an element in the environment when these fluxes are much smaller than the stocks. In some cases, isotopes are an irreplaceable research tool, for example, to estimate gross nutrient fluxes, such as gross mineralization rate. Finally, isotope data are used to constrain biogeochemical models.

We report here relevant although not exhaustive applications of C, N, O, and S stable isotopes in soil science studies.

Table 1 Isotopes widely used in soil science and relative international standards.

Element	Isotope symbol	Atomic mass <i>u</i> , mol g ⁻¹	Nominal abundance atom percent (%) ^a	Relative mass difference %	Standard value <i>R</i> × 10 ⁶ ^b	Standard acronym and identification
Hydrogen	¹ H	1.007825032	99.984426	100	155.76 ± 0.10	V-SMOW, Vienna-Standard Mean Ocean Water
	² H or D	2.014101778	0.015574			
Carbon	¹² C	12	98.8944	8.3	11,237.2 ± 2.9	V-PDB, calcium carbonate found in <i>Belemnitella americana</i> in the PeeDee cretaceous formation in South Carolina, USA
	¹³ C	13.00335484	1.1056			
Nitrogen	¹⁴ N	14.00307401	99.6337	7.1	3676.5 ± 8.1	Air nitrogen
	¹⁵ N	15.00010897	0.3663			
Oxygen	¹⁶ O	15.99491462	99.76206	12.5 (¹⁸ O: ¹⁶ O)	2005.20 ± 0.43 *	V-SMOW, Vienna-Standard Mean Ocean Water
	¹⁷ O	16.9991315	0.0379			
	¹⁸ O	17.9991604	0.2004			
Sulfur	³² S	31.97207073	95.03957	6.3 (³⁴ S: ³² S)	45,005.5 ± 9.3 **	V-CDT, Troilite from the Canyon Diablo iron meteorite
	³³ S	32.97145854	0.74865			
	³⁴ S	33.96786687	4.19719			
	³⁶ S	35.96708081	0.01459			

Data from Coplen, 2011 and Hoefs, 2021.

* Referred to the ¹⁸O/¹⁶O isotope ratio.

** Referred to the ³⁴S/³²S isotope ratio.

^aAtom percent as in the formula in Table 2.

^b± 95% confidence interval.

How to express abundances of isotopes at natural abundance and in enriched samples, and how to express the isotopic fractionation

Stable isotope studies are based on measurements of changes in abundance of different isotopes in a compound, phase, or material. The abundance can be expressed in different ways as summarized in Table 2 (reviewed also in Werner and Cormier, 2022).

In recent years, the use of isotope delta with the per mill notation has been debated (Table 2). This notation is the preferred one in most studies using stable isotopes at natural abundance, even though the ‰ does not represent a unit in the International System of Units (SI; Brand and Coplen, 2012). Coplen (2011) recommended that the multiplicative factor 1000 should be avoided. An alternative option would be to use a new unit milli-Urey, mUr, to replace the ‰ sign (Brand and Coplen, 2012). In this chapter, after considering its still widespread use, we will refer to isotope delta in the per mill notation.

All the notations reported in Table 2 have their distinct application, depending on the study. When working with samples enriched in heavy isotopes (tracing experiments, see Section “Natural abundance and labelling approaches”), use of the delta notation is not advisable, and the atom fraction or atom percent should be chosen.

Isotope delta values behave ‘linearly’ in equations for mass balances, but only when all compounds are at natural abundance. When used for samples or standards enriched in heavy isotopes, delta values will generate errors (Corso and Brenna, 1997). In these instances, the *isotope phi* value can overcome this limitation and is recommended. For an example of calculation, see Box 1. Further details on the use and significance of *phi* are given in Werner and Cormier (2022).

Isotope effect and isotopic fractionation

Fractionation of isotopes is defined by Hoefs (2021) as “the partitioning of isotopes between two substances or two phases of the same substance with different isotope ratios”.

Isotopic fractionation is caused by both physical reactions, including evaporation or diffusion, and chemical reactions. Isotopic fractionation processes can be further divided into: (i) equilibrium or thermodynamic isotopic fractionation, also known as equilibrium exchange reactions, and (ii) kinetic isotopic fractionation.

Table 2 Most common notations for isotopes and measures to express isotopic composition and isotopic fractionation in numbers. In this case, we take the example of carbon stable isotopes, ^{12}C and ^{13}C .

Name	Formula	Notes; References
<i>Isotope notation</i>		
Atom fraction x	$x(^{13}\text{C})_A = \frac{n(^{13}\text{C})_A}{n(^{12}\text{C})_A + n(^{13}\text{C})_A}$	Dimensionless; Coplen (2011)
Atom percent ‰	$\text{‰}(^{13}\text{C})_A = \frac{n(^{13}\text{C})_A}{n(^{12}\text{C})_A + n(^{13}\text{C})_A} \cdot 100$	Old term, according to Coplen (2011) it should be replaced by the atom fraction x ; however, it is still widely used and accepted by authors and reviewers; Coplen (2011)
Atom percent excess	$\text{APE} = \text{‰}(^{13}\text{C})_A - \text{‰}(^{13}\text{C})_{\text{control}}$	Coplen (2011) and references therein
Isotope ratio	$^{13}R_A = \frac{^{13}\text{C}}{^{12}\text{C}}$	Coplen (2011) and references therein
Delta notation	$\delta^{13}\text{C}_A = \frac{^{13}R_A - ^{13}R_{\text{Std}}}{^{13}R_{\text{Std}}}$ $\delta^{13}\text{C}_A = \left(\frac{^{13}R_A}{^{13}R_{\text{Std}}} - 1 \right)$	Until recently, a multiplicative factor 1000 was included in these equations: see discussion in section “How to express abundances of isotopes at natural abundance and in enriched samples, and how to express the isotopic fractionation”; McKinney et al. (1950) and Coplen (2011)
Isotope ratio from delta	$^{13}R_A = (\delta^{13}\text{C}_A + 1) \cdot ^{13}R_{\text{Std}}$	–
Isotope phi	$\varphi^{13}\text{C}_A = \frac{\delta^{13}\text{C}_A}{^{13}R_A + 1}$ $\varphi^{13}\text{C}_A = \frac{x(^{13}\text{C})_A}{^{13}R_{\text{Std}}} - 1$	Corso and Brenna (1997) and Brand and Coplen (2012)
<i>Isotopic fractionation notation</i>		
Alpha ^a	$\alpha^{13}\text{C}_{\frac{B}{A}} = \frac{^{13}R_B}{^{13}R_A} = \frac{\delta^{13}\text{C}_B + 1}{\delta^{13}\text{C}_A + 1}$	Friedman and O’Neil (1977)
Epsilon ^a	$\varepsilon^{13}\text{C}_{\frac{B}{A}} = \alpha^{13}\text{C}_{\frac{B}{A}} - 1$	Can be expressed in 10 ³ or milli-Urey
Per mill isotope fractionation	$10^3 \ln \alpha = 1000 \ln \left(\alpha^{13}\text{C}_{\frac{B}{A}} \right)$	Commonly used in geochemical studies; Friedman and O’Neil (1977)
Relation among isotopic fractionation notations	$\Delta^{13}\text{C}_{\frac{B}{A}} = \delta^{13}\text{C}_B - \delta^{13}\text{C}_A \approx \varepsilon^{13}\text{C}_{\frac{B}{A}} = \alpha^{13}\text{C}_{\frac{B}{A}} - 1 \approx \ln \left(\alpha^{13}\text{C}_{\frac{B}{A}} \right)$	Friedman and O’Neil (1977)

n = number of atoms; A and B = compounds A and B ; Std = standard (see Table 1 for list of standards); control = control compound (see examples about ^{15}N in Section “Nitrogen”); $\ln$ = natural logarithm.

^aIn the following equilibrium oxygen isotope exchange reaction between calcite and water: $1/3 \text{CaC}^{16}\text{O}_3 + \text{H}_2^{18}\text{O} \rightleftharpoons 1/3 \text{CaC}^{18}\text{O}_3 + \text{H}_2^{16}\text{O}$ the $\alpha^{18}\text{O}_{\text{calcite/water}}$ at 33.7 °C is 1.02849 ± 0.00013 and ε is equal to: $\alpha^{18}\text{O}_{\text{calcite/water}} - 1 = 0.02849$ and can be expressed as 28.49‰ (data taken from Coplen, 2011).

Box 1 The use of *isotope phi*.

Let us hypothesize that we need to obtain 100 mL of ^{18}O -labelled water (W_i) with a delta value $\delta^{18}\text{O}$ of +120‰ ($R_i = 0.002245835$) by diluting labelled water (W_{labelled}) with an ^{18}O atom percent of 98% with unlabelled water (W_{unlab}), having a $\delta^{18}\text{O}$ of -11‰ and $R_{\text{unlab}} = 0.001985158$. The $\delta^{18}\text{O}$ values are expressed against the VSMOW standard, which has an $R_{\text{std}} = 0.00200521$. Note that to calculate ϕ , the $\delta^{18}\text{O}$ in per mill, should be divided by 1000.

$$\phi(W_{\text{labelled}}) = (0.98/0.00200521) - 1 = 487.73$$

$$\phi(W_{\text{unlab}}) = (-0.011)/(0.001985158 + 1) = -0.01$$

$$\phi(W_i) = (0.120)/(0.002245835 + 1) = 0.12$$

$$100 \text{ mL } W_i (\phi=0.12) = x \text{ mL } W_{\text{lab}} (\phi=487.73) + (100 \text{ mL} - x) W_{\text{unlabelled}} (\phi = -0.01)$$

Solving the mass balance equation for x , we find that we need 0.03 mL of ^{18}O labelled in 99.97 mL of unlabelled water.

Equilibrium isotopic fractionation involves reactions at chemical equilibrium, i.e. reactions in which forward and backward rates are equal. Note that isotopic equilibrium usually takes longer to be established than chemical equilibrium (Rishavy and Cleland, 1999). During equilibrium reactions, the heavier isotope tends to concentrate in the phase or molecule with stronger covalent bonds (i.e. more energetic). In equilibrium exchange reactions, the compound with the higher oxidation state or more condensed state will also be enriched in the heavier isotope (Coplen, 2011). For example, carbon dioxide is enriched in ^{13}C compared to methane, and liquid water is enriched in ^{18}O and ^2H relative to water vapor. Equilibrium exchange reactions typically occur in closed and well-mixed systems. A typical example is the oxygen isotope exchange between liquid water and CO_2 in a closed container. Once the reaction reaches thermodynamic equilibrium, the reactants and products end up with different isotopic composition, because ^{18}O forms a stronger covalent bond with C than with H (Sharp, 2017 and references therein). This has been suggested as the principal effect controlling the composition of oxygen isotopes in atmospheric CO_2 at the global scale (Francey and Tans, 1987). In equilibrium exchange reactions, the shift in isotopic composition of products is constant and inverse with temperature, i.e. greater isotopic fractionation occurs at lower temperatures.

Kinetic isotopic fractionation processes occur in unidirectional and/or irreversible processes or reactions, that is when a reaction proceeds only in the forward direction, the backward reaction being hindered. These include many biological reactions catalyzed by enzymes. Dissimilatory sulfate reduction by sulfate-reducing bacteria or mineralization of phosphoesters by phosphatase enzymes are examples of kinetic isotopic fractionations mediated by enzymes.

Reversible reactions not reaching chemical equilibrium will display net isotopic fractionations in between the kinetic and equilibrium ones. A classical example is represented by reactions where the products are constantly removed from the system, such as water evaporation from surface waters to the atmosphere (described by the well-known Craig-Gordon model, Horita et al., 2008).

In kinetic isotopic fractionation processes, the lighter isotope usually accumulates in reaction products, as the lighter isotopes usually react faster and statistically form weaker chemical bonds (Coplen, 2011). In the previous example of water evaporation in an open system, the water vapor is progressively depleted in ^{18}O and ^2H relative to the liquid water. Kinetic isotopic fractionation might evince magnitudes similar and often greater than those produced from equilibrium exchange reactions. However, when a kinetically controlled reaction goes to completion, the cumulative product ends up with the same isotopic composition as the initial reactant, in other words quantitative reactions do not show any isotopic fractionation.

The isotopic fractionation between two substances or phases, A and B, is quantified by the isotopic fractionation factor alpha (α , Table 2). The commonly used isotopic fractionation, ϵ , similar to the delta value, represents the deviation of α from 1 (Table 2).

Both α and ϵ can refer to equilibrium exchange and kinetic processes. In complex biogeochemical systems, several often unknown reactions or reaction network might result in a 'net' or 'apparent' isotopic fractionation, which is also expressed as α or ϵ (Coplen, 2011).

Note that some authors define α for a kinetic reaction giving the product B from the reactant A as in Table 2, while others use the reciprocal form ($\alpha = {}^{13}R_A/{}^{13}R_B$, here exemplified for carbon isotopes, see Werner and Cormier, 2022). In soil science, both definitions are used. To avoid confusion, it is recommended that the source reference for the definition of α and ϵ are checked and that the equations authors use to express isotopic fractionation are clearly specified.

If the isotopic fractionation is defined as in Table 2, ϵ is positive when the product is enriched in the heavier isotope with respect to the reactant, whereas ϵ is negative when the reactant is enriched in the heavier isotope.

Analytical methods

For studies in soil sciences, solid, liquid, and also gaseous samples are measured for the isotopic composition of the element of interest.

Isotope-ratio mass spectrometry (IRMS) is the most common method of analysis, although other techniques are now being employed for specific samples. A mass spectrometer can determine the isotope ratio of an element by separating moving ionized atoms and molecules in an electromagnetic field on the basis of their mass and charge. The isotope ratio obtained is then compared to that of a reference gas to determine the delta value.

In general, before entering the mass spectrometer in the form of specific inorganic gases, solid and liquid samples have to be processed by methods of elemental analysis (Fig. 1). An elemental analyzer (EA) quantitatively converts the samples into specific gases. Combustion at temperatures between 950 °C and 1050 °C (used to measure C, N, and S isotopes), in the presence of O₂, produces CO₂, N₂/NO_x, SO₂ and H₂O. To yield N₂ quantitatively, nitrogen oxides NO_x need to be reduced on elemental copper (Cu) at T between 600 °C and 650 °C. In addition, the surplus O₂ from the combustion pulse is trapped with Cu. High-temperature conversion of inorganic compounds (and pyrolysis of organic compounds) takes place in carbon-based reactors under helium (He) flow and in the absence of O₂, and it is used to measure O and H isotope ratios. Samples are converted quantitatively to CO and H₂ at temperatures between 1350 °C and 1450 °C. Water is removed by chemical traps magnesium perchlorate (Mg(ClO₄)₂, Sicapent (P₄O₁₀), or Nafion® tubes flushed with helium). The gases produced are then separated by gas chromatography or temperature-controlled traps (purge and trap) and transferred to the IRMS instrument within a clean He flow.

Nowadays, IRMS systems work in continuous flow mode: the sample gases generated by the EA (or another preparation inlet system see Fig. 1) and the corresponding reference gases are carried to the IRMS instrument within a clean He stream in pulses. The traditional IRMS uses a so-called dual inlet valve system. Pure sample gas (obtained offline) and pure reference gas are alternately and iteratively introduced to the ion source via a changeover valve. This method is now used only when samples need to be measured with greater precision and accuracy. For a thorough comparison of the two systems, please refer to Barrie et al. (1996).

Inside the ion source of an IRMS instrument, the gaseous molecules are first ionized, accelerated, and focused and then enter the flight tube, where the ionized molecules are separated into beams by an (electro)magnetic field on the basis of their mass-to-charge ratio (m/z). The intensity of each beam is then measured by ion collectors (Faraday cups) placed in a specific arrangement. The input from each collector is converted into an electric impulse, amplified, and finally integrated and processed. For further info on mass spectrometer hardware be referred to Brand (2004).

Isotope ratios of liquid samples (e.g. H₂O extracted from soil and plant matter) can also be measured after equilibration with CO₂ and/or H₂ gas, which is then separated by gas chromatography or cryogenic trapping and sent to the IRMS system (Fig. 1). The same is valid for gaseous samples, such as atmospheric carbon dioxide (CO₂) and nitrous oxide (N₂O). Because of possible small amounts of the trace gases, a pre-concentration step by cryo-traps might be needed. Nitrous oxide can be analyzed for bulk $\delta^{15}\text{N}$ and $\delta^{18}\text{O}$ values; in addition, a measurement of the fragment ion at mass 30 ($^{14}\text{NO}^+$) and 31 ($^{15}\text{NO}^+$) produced in the ion source allows N₂O isotopomer analysis (i.e. intramolecular distribution of ^{15}N in N₂O).

Compound-specific isotope ratio analyses can be performed by coupling gas chromatography (GC) or high-performance liquid chromatography (HPLC) online to an IRMS instrument (Jochmann and Schmidt, 2012).

Recent developments in infrared laser spectroscopy (cavity ring-down spectroscopy, CRDS, and quantum cascade laser absorption spectroscopy, QCLAS; Fig. 1, Li et al., 2013) have enabled less costly measurement of isotope ratios of liquid and gaseous

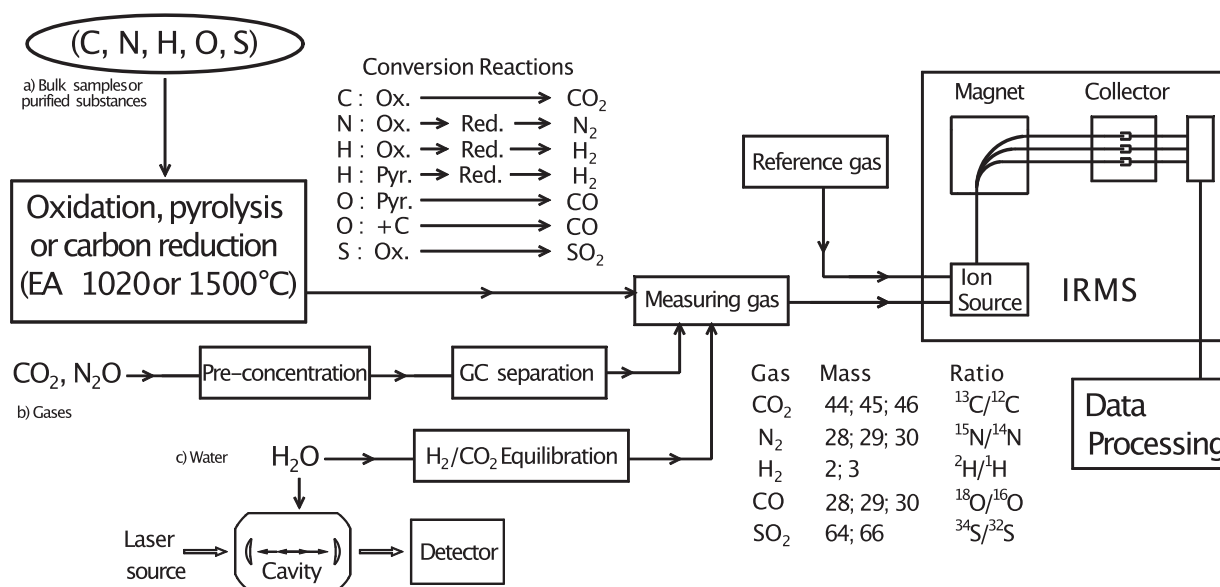


Fig. 1 Workflow for isotope ratio mass spectrometry (IRMS) measurements. (A) For solid bulk samples (e.g. plants, soil) or purified substances, samples are pretreated and compounds are converted into gases prior to measurements. The carbon reduction is normally performed on glassy carbon under He flow; pyrolysis of N containing organic substances for $\delta^2\text{H}$ analysis is performed in chromium reactors (for more information, see text). In case of coupled HPLC–IRMS for $\delta^{13}\text{C}$ determination, wet oxidation is used. (B) For the analysis of gaseous samples (e.g. CO₂ and N₂O), samples are pre-concentrated prior to IRMS analysis. (C) In the routine isotope ratio analysis of water isotope equilibration with gases is often used: for $\delta^{18}\text{O}$ analysis CO₂, for $\delta^2\text{H}$ analysis H₂-gas in the presence of Pt (platinum). More recently, $\delta^{18}\text{O}$ and $\delta^2\text{H}$ in water can be measured with an online method, using cavity ring-down spectroscopy. Modified after Schmidt H-L, Werner RA, Rossmann A, Mosandl A, and Schreier P (2007) Stable isotope ratio analysis in quality control of flavourings. In: Ziegler H (ed.): Flavourings - Production, Composition, Applications, Regulations. 2nd completely revised edition. Weinheim, D: Wiley-VCH.

samples, and with less preparation. The method, based on specific infrared absorption of different gas molecules, enables continuous measuring of isotope ratios. Although laser spectroscopy does not yet provide the same precision as IRMS, these techniques are becoming increasingly popular, especially for their direct application in field studies.

Natural abundance and labelling approaches

Stable isotope studies are based on the measurement of isotopic natural abundance or deliberately induced changes in isotope abundance (labelling¹). In the latter approach, a compound that is deliberately enriched or depleted in the heavier isotope of an element is introduced into the system being studied. A classic example is the application of ¹⁵N enriched fertilizer to the soil.

The advantage of manipulating isotope abundance is that the targeted transformation or flux, e.g. the N uptake of the plant from a given source, is easily detectable. Such studies also allow processes to be investigated over small temporal and spatial scales ranging from days to a few years, and from microcosms to field experiments. On the other hand, these approaches might be subject to experimental artifacts, such as a non-homogeneous distribution of the tracer throughout a substrate.

Natural abundance studies allow larger temporal and spatial scales to be embraced in biogeochemical cycles (Fig. 2), e.g. soil carbon dynamics over decades (see Section “Natural ¹³C abundance to study SOM dynamics in soil of arable lands”) or phosphorus cycling during pedogenesis (see Section “Oxygen in phosphate to study P cycling”). One of the drawbacks of the natural isotope abundance approach is that, in some cases, the relatively small shifts (few delta units) that can be observed, enable only qualitative assessment and might hamper correct interpretations.

Examples of advantages and disadvantages of the two approaches will be presented in the following sections. Detailed comparisons for specific isotopes are given in Robinson (2001) and Chalk et al. (2017).

Application of isotopes of light elements in soil science

Carbon

In soil science, the C isotope ratio is used as a tracer of plant-C to understand its interaction with the soil matrix and the main pathways of soil organic matter (SOM) accumulation. Paleoenvironmental studies use $\delta^{13}\text{C}$ to evaluate past changes in vegetation

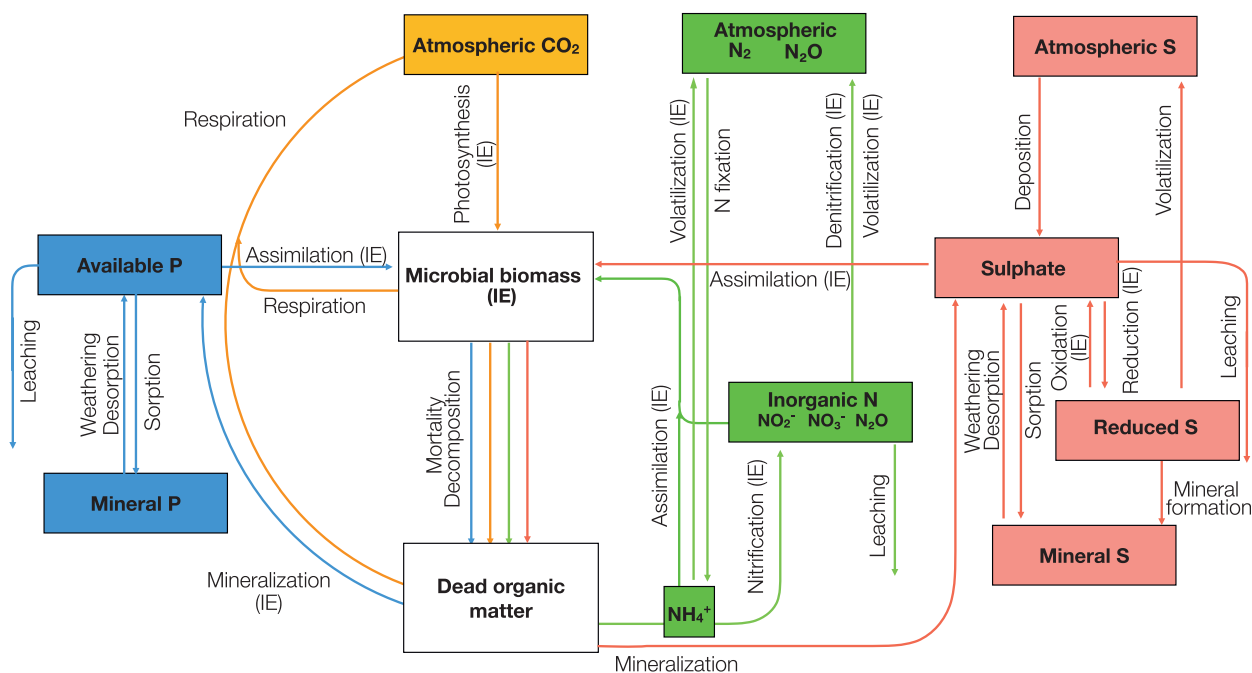


Fig. 2 Schematic representation of C, N, P and S cycles in the soil. A general indication of nutrient pools in soil is given. Abiotic and biotic processes are indicated by arrows, IE: Isotope effect, indicating the presence of known isotopic fractionation of processes. Modified after Janes-Bassett V, Davies J, Rowe Ed C., and Tipping E (2020) Simulating long-term carbon nitrogen and phosphorus biogeochemical cycling in agricultural environments. *Science of The Total Environment*. 714.

¹When the compound is a gas, labelling techniques can be continuous, i.e. continuous exposure of the studied system to the labelled compound or pulsed, i.g. repeated exposure to the labelled compound for short periods of time. An example is provided in Section “Carbon”.

and climate. On the other hand, known changes in vegetation and consequent changes in the $\delta^{13}\text{C}$ of organic matter inputs are used to estimate the turnover time and size of SOM fractions.

Natural ^{13}C abundance to study SOM dynamics in the soil of arable lands

Photosynthesis is the main way how atmospheric CO_2 is captured and transformed into organic compounds. This process occurs with a kinetic isotopic fractionation of C, leading to ^{13}C depletion compared to atmospheric CO_2 , which is different for C_3 and C_4 plants (mean $\delta^{13}\text{C}$ value in C_3 plants is -27‰ , and -13‰ in C_4 plants, Hoefs, 2021). This difference is used to follow the dynamics of plant-C entering the soil through residues left after harvest and their decomposition by soil microorganisms. The use of this technique enables SOM turnover to be followed over medium to long-term scales. For instance, Balesdent and Balabane (1996) grew maize (*Zea mays* L., a C_4 plant) for 4 years in a long-term field experiment that had been previously cultivated with C_3 plants. By following the dynamics of SOM derived from above- and below-ground maize residues, they demonstrated that roots were the major contributor to SOM, while representing only one third of the amount of C inputs to the soil from the residues. These pioneering results stress the importance of roots and associated rhizodeposition mechanisms. They were achieved by following the fate of C_4 plant-C over several years and by using a simple mixing model (Box 2) under the assumptions that the $\delta^{13}\text{C}$ of plant residues is close to that of derived SOM, and that SOM dynamics are close to steady-state.

However, several studies have reported an increase in $\delta^{13}\text{C}$ values with depth in soils under C_3 plants (on average 1–3‰ in the first meter (Balesdent et al., 1993). Different processes were hypothesized to explain this ^{13}C enrichment: (1) roots are enriched in ^{13}C compared to shoots (Werth and Kuzyakov, 2010); (2) isotopic fractionation during microbial respiration leads to an enrichment in ^{13}C during the decomposition process. A recent study concluded that the increase in $\delta^{13}\text{C}$ values with soil depth was due to changes in vegetation $\delta^{13}\text{C}$ over the last 20,000 years (Paul et al., 2020).

Further work based on C_3/C_4 vegetation changes demonstrated that the soil below 30-cm depth (subsoil) contains only 19% of the soil organic carbon (SOC) recently (last 50 years) incorporated, which is mostly concentrated within the first 10 cm of soil. Therefore, cropping seems not to affect the deeper soil horizons greatly in relation to SOM accumulation (> 30 cm, Balesdent et al., 2018).

In cropped ecosystems, natural ^{13}C abundance can also be used to follow the fate of organic fertilizers (waste, digestate or composts as fertilizers, Jamoteau et al., 2021).

Labelling approaches to study SOM in cropped soils

Another way of assessing SOM accumulation and turnover is with ^{13}C labelling of plants in gas-tight growth chambers. The advantage of this technique is the possibility of using a wider range of residues than those from C_4 plants and to apply them to all types of environments regardless of whether C_3 or C_4 plants were present previously. However, these experiments are time consuming and costly, as continuous highly ^{13}C enriched CO_2 injections are required through the whole plant cycle to obtain homogenous labelling of the plant material. An alternative approach is pulse-labelling, where plants are exposed to a $^{13}\text{CO}_2$ flush for short and repeated periods of time. Such methods can be applied in a greenhouse or directly in the field (Leake et al., 2006). More recent methods based on applying dual isotope labelling (^{13}C and ^{15}N) directly onto the leaf surface were reported as being suitable. However, with the leaf surface method the ^{13}C enrichment of the plant tissues remains low ($< 1\%$ atom percent excess or APE, see Table 2) (Putz et al., 2011).

Most studies related to SOM dynamics use labelled crop residues to follow C incorporation into SOM, while those dealing with plant growth or rhizodeposition will need to follow transfer of the tracer from the plants to the soil during the whole life cycle of the plant. The different soil compartments targeted to follow ^{13}C enriched compounds are the rhizosphere, the soil fauna and microorganisms, the SOM and its fractions.

Box 2 Mixing models.

The delta value of a sample (receiver) can be explained by the delta values of the mixing sources (also called endmembers) and their relative contribution (fraction) to the total amount of the element of interest. The equation used is exemplified here for ^{15}N and the simplest case of two contributing sources:

$$\delta^{15}\text{N}_{\text{sample}} = \delta^{15}\text{N}_{\text{sourceA}} * f + \delta^{15}\text{N}_{\text{sourceB}} * (1 - f) \quad (1)$$

where $\delta^{15}\text{N}_{\text{sample}}$ is the delta value of the sample, $\delta^{15}\text{N}_{\text{sourceA/B}}$ are the delta values of the two distinct sources A and B, and f and $1 - f$ are the relative contributions of the two sources.

Rearranging equation 1, the fraction f can be calculated:

$$f = \frac{\delta^{15}\text{N}_{\text{sample}} - \delta^{15}\text{N}_{\text{sourceB}}}{\delta^{15}\text{N}_{\text{sourceA}} - \delta^{15}\text{N}_{\text{sourceB}}} \quad (2)$$

Note that Equation 2 is sometimes written with the numerator and denominator multiplied by -1 , as in the example of the N in a legume derived from atmospheric N_2 (PNdfa, see Section “The use of ^{15}N enriched fertilizers”).

To calculate the mass of N derived from one source in the sample, the total mass of N in the sample must be known. For example, for source A we have:

$$\text{mass}_{\text{sourceA}} = f * \text{mass}_{\text{sample}} \quad (3)$$

- Use of stable isotopes to follow C dynamics from living plants

The amount and quality of C entering the soil through rhizodeposition is still controversial. The interaction between the rhizosphere leads to an increase or decrease of SOC concentration and is termed rhizosphere priming effect. This process could accelerate SOC mineralization by 59% compared to unplanted soil (Huo et al., 2017). The rhizosphere priming effect cannot be studied without isotopes, either stable or radioactive (^{14}C). To distinguish the origin and fate of rhizosphere C, ^{13}C labelling of plants is needed, although several studies use the addition of ^{13}C labelled compounds to mimic root exudates.

- Use of stable isotopes to follow C dynamics from litter residues

Residues enriched with ^{13}C are mostly from plants growing under ^{13}C enriched CO_2 . The level of enrichment ranges around 3 to 5% APE. However, specific studies require strongly labelled residues (i.e. about 99% APE). This is the case of DNA stable isotope probing (DNA-SIP) which, in combination with metagenomics, is dedicated to the identification of the soil microorganism functions in the environment (Bernard et al., 2007).

Plants residues are the main source of organic matter entering the soil. Microbial products resulting from the assimilation of labile components in residues by microorganisms (i.e. microbial immobilization) and microbial turnover were recently shown to contribute significantly to the accumulation of SOM (Cotrufo et al., 2015). These results challenge previous studies, which contend that chemical recalcitrance of residues was the main driver of the rate of decomposition and SOM formation. The study by Cotrufo and collaborators used ^{13}C labelled residues to follow the incorporation of C in active soil microbial biomass (by ^{13}C -PLFA, phospholipid fatty acid analyses) over several years and in SOM fractions for different soil layers (from 0 to 20 cm depth).

The ^{13}C labelled substrates (residues, glucose or other organic compounds) are added to soil in incubation experiments to study microbial response to C addition. Such approaches aim at understanding microbial processes such as stoichiometric constraints (Heuck et al., 2015), enzyme efficiency (Fanin and Bertrand, 2016), carbon use efficiency (Sauvadet et al., 2018) and priming effect (Perveen et al., 2019). In such applications, ^{13}C labelled substrates are an asset for a better understanding of microbial functions and their impact on the environment.

Nitrogen

Human alterations of the global N cycle have dramatically increased emissions of reactive N compounds into terrestrial and aquatic systems and into the atmosphere. Agricultural production plays a key role in these changes because food production depends on the use of reactive N. The ^{15}N enrichment and natural abundance methods have greatly advanced the understanding of N fluxes and dynamics in agricultural systems (arable crops and grassland), which is required to reach more efficient fertilizer N use with fewer losses to the environment.

The use of ^{15}N enriched fertilizers

The use of ^{15}N enriched mineral or organic fertilizers such as animal manures and legume green manures allows fertilizer N to be tracked over time and across different compartments. Often less than half of fertilizer N is taken up by the crops (Douxchamps et al., 2011; Frick et al., 2022). With ^{15}N enriched fertilizers, the fate of fertilizer N not taken up by crops can be tracked into soil N pools and N losses can be quantified. The basic requirement for tracking is that the fertilizer N is homogeneously labelled with ^{15}N . Homogeneity of ^{15}N labelling is usually not a concern for pure, water soluble synthetic N compounds such as urea or ammonium nitrate, which can be either purchased at the required level of ^{15}N enrichment, or produced in a laboratory. The production of homogeneously labelled organic fertilizers is much more challenging. For instance, the production of ^{15}N labelled animal manure requires several days of feeding animal(s) with ^{15}N labelled forage, analysis of the ^{15}N enrichment of urine and feces collected at different times since the start of feeding, and checking on the homogeneity of ^{15}N labelling, e.g. with physicochemical fractionation of fecal N (Frick et al., 2022). The ^{15}N labelled legume residues for green manures have been produced under field and greenhouse conditions (Douxchamps et al., 2011; Traoré et al., 2020), but for mature plant material the ^{15}N enrichment may differ between plant parts, e.g. leaves and stems. The proportion of N in the receiver pool derived from the labelled fertilizer is: $\text{Ndff} = \frac{^{15}\text{N excess}_{\text{receiver}}}{^{15}\text{N excess}_{\text{source}}}$, with excess being the APE as defined in Table 2.

When the size of the receiving pool is known, recovery of the fertilizer in that pool can be calculated. Use of ^{15}N enriched fertilizers has provided information on the:

- recovery of mineral and organic fertilizer N in crops in the year of application and during the following years, delivering information on the residual fertilizer value (Douxchamps et al., 2011; Frick et al., 2022);
- quantification of fertilizer N remaining in the soil over years (see compilation in Smith and Chalk, 2018);
- incorporation of fertilizer N into soil N pools, such as mineral and microbial N (see for example Frick et al., 2022) and into soil aggregate fractions (Bosshard et al., 2008);
- losses of fertilizer N by leaching, denitrification and ammonia volatilization (see for example Thomsen et al., 1997);
- symbiotic N_2 fixation by legumes, through the use of the residual ^{15}N enrichment in soil (Oberson et al., 2007).

Gross N fluxes in soil, namely ammonification and nitrification, have been quantified using ^{15}N pool dilution. This approach has elucidated the significance of gross N fluxes compared to net N mineralization (Braun et al., 2018 and references therein), and their

response to management intervention such as soil tillage or the integration of grasses with nitrification inhibition capacity (Teutschnerová et al., 2022).

The ^{15}N labelling of legumes has been used to estimate their belowground N input (Hammehle et al., 2018) and the transfer of legume N to associated non-fixing plants (Pirhofer-Walzl et al., 2012).

Because of high ^{15}N tracer costs, field applications of ^{15}N enrichment are limited to microplot scales with a maximum of few m^2 per unit.

The use of ^{15}N at natural abundance

The work with ^{15}N natural abundance bases on natural ^{15}N enrichment usually encountered in soils ($\delta^{15}\text{N}$ of 2‰ to 15‰) relative to atmospheric N_2 ($\delta^{15}\text{N}$ of 0‰). This difference has frequently been used to obtain the proportion of N in legumes derived from atmospheric N_2 by symbiotic fixation (see compilation in Chalk et al., 2016).

The great advantage of ^{15}N natural abundance work is that it is not limited in space, and no methodological intervention is needed in the field. However, potential bias induced by isotopic fractionation has to be carefully considered (Robinson, 2001 and references therein), and here it is explained using symbiotic N_2 fixation in legume shoots as an example. Determination of the proportion of N in a legume derived from atmospheric N_2 (PNdfa) is based on the two pools mixing model (mixing of two sources, and a sink or receiver, see Box 2). The equation is:

$$\text{PNdfa} = \frac{\delta^{15}\text{N}_{\text{soil}} - \delta^{15}\text{N}_{\text{legume}}}{\delta^{15}\text{N}_{\text{soil}} - \delta^{15}\text{N}_{\text{atm}}} \quad (4)$$

where $\delta^{15}\text{N}_{\text{soil}}$ is the $\delta^{15}\text{N}$ of a non-fixing reference plant reflecting the $\delta^{15}\text{N}$ of the soil available N pool (first source); $\delta^{15}\text{N}_{\text{atm}}$ is the $\delta^{15}\text{N}$ of the legume relying solely on atmospheric N_2 (second source, in literature often called B), $\delta^{15}\text{N}_{\text{legume}}$ is the $\delta^{15}\text{N}$ of the legume (receiver) (Shearer and Kohl, 1986). This implies that the reference plant takes up N in a similar temporal and spatial pattern to that of the legume, and with the same isotopic fractionation involved. Sampling of different reference plants, from various species, all growing near to the legume, and under the same management as the legume studied, helps to reduce related uncertainty (Obersen et al., 2013). The isotopic fractionation taking place during N_2 fixation and transport of fixed N from nodules to legume shoots is accounted for by the $\delta^{15}\text{N}_{\text{atm}}$ value, which is usually depleted in ^{15}N compared to atmospheric N_2 (e.g., -0.08‰ to -2.6‰ , Büchi et al., 2015). The same principle has been applied to derive the transfer of legume N to associated non fixing plants (Obersen et al., 2013).

Fertilizers differ in their $\delta^{15}\text{N}$ values. The $\delta^{15}\text{N}$ of synthetic fertilizer usually ranges from -3‰ to $+3\text{‰}$, for livestock manure from $+3\text{‰}$ to $+14\text{‰}$, and for composted manure from $+8\text{‰}$ to $+20\text{‰}$ (range indicating 90% of the respective fertilizers, Choi et al., 2017). These differences are used to derive the proportion of fertilizer N derived in the crop (Manandhar et al., 2022). In all these applications, it is necessary to check carefully that the $\delta^{15}\text{N}$ values of the two sources are clearly distinct, and that they account for isotopic fractionation during the transfer of N from these sources to the receiver.

The $\delta^{15}\text{N}$ value of plant shoot tissue has also been used to identify the form of soil N taken up (e.g., preference of ammonium over nitrate, Kahmen et al., 2008). It was the original means of detecting biological N fixation in tropical grasses such as sugar cane (*Saccharum officinarum* L.) or *Urochloa* spp. grasses (Villegas et al., 2020).

To reduce N losses from agricultural fields to other environmental compartments, future studies should explore the mechanisms of N retention in soil at the same time as supplying mineral N to plants. Because the nitrogen cycle is coupled with the carbon cycle, multi-element studies using multiple isotopes will become increasingly relevant.

Oxygen

Oxygen isotopes in the soil water–plant–atmosphere continuum

Stable isotopes in soil water (^2H and ^{18}O) have been used for decades to assess flows in the unsaturated zone and to understand plant water uptake and evapotranspiration patterns. Here we provide some general aspects of O isotopes in the soil–plant–atmosphere continuum. The reader can refer to Gat (2010) and Sprenger et al. (2016) for in-depth accounts.

Precipitation water entering the soil may partially evaporate and the water remaining in the soil will become enriched in heavy isotopes through kinetic isotopic fractionation. Uptake by plants does not produce significant isotopic fractionation. In contrast, leaf water might be considerably enriched in ^{18}O (and ^2H) relative to soil water due to isotopic fractionation during transpiration (Yakir and da Sternberg, 2000). Evapotranspiration can be partitioned by applying mixing models (see Box 2) that combine the isotopic composition of atmospheric water, the transpired water and evaporated soil water (Sprenger et al., 2016 and references therein).

Mixing models are also applied to study the depth or distribution of root water uptake, i.e. by comparing the isotopic composition of the xylem water with the isotopic composition of possible water sources (rainfall, pore water in upper and deeper soil horizons).

In addition to natural isotope abundance, application of isotopic tracers (water artificially enriched in the heavier isotopologs) has also been used to assess sources of water uptake by plants and processes such as the hydraulic lift. For instance, Meunier et al. (2018) used ^{18}O enriched water to label the soil water in deep soil layers and subsequently observed an ^{18}O enrichment of water in

the upper soil layers. They explained these differences by the hydraulic lift operated by the roots of *Lolium multiflorum* (Lam.), Italian ryegrass.

In the plant, carbon dioxide, water and dioxygen (O_2) are the primary sources of O in organic compounds (Schmidt et al., 2001). In contrast to O atoms in hydroxyl groups, O in carbonyl groups of aldehydes and ketones are known to be easily exchangeable with surrounding H_2O molecules (da Silveira Lobo O'Reilly Sternberg and DeNiro, 1983). This leads to an ^{18}O equilibrium enrichment of these functional groups relative to water. Oxygen isotope exchange of intermediates during biosynthesis leads to an ^{18}O enrichment of $+27\text{‰} \pm 3\text{‰}$ of cellulose relative to leaf water (da Silveira Lobo Sternberg, 1989 and literature cited therein). The oxygen isotopic composition of plant carbohydrates is therefore related to that of the water available during biosynthesis and can be used to assess the physiology of plants and obtain information on (paleo)climate (Gessler et al., 2014).

Oxygen in phosphate to study P cycling

In the past two decades, researchers have increasingly used the isotopic composition of O in phosphate ($\delta^{18}O_P$) to study phosphorus cycling in soils, sediments and the soil–plant system.

Phosphorus (P) in the environment is most often associated to four O atoms (PO_4^{3-} , hereafter phosphate), whether in its inorganic or organic forms. The P—O bond is only cleaved by biological processes at temperatures and pressures commonly found on the Earth's surface. In such circumstances, the $\delta^{18}O_P$ retains a trace of the biological processes that the phosphate molecule went through until another biological process intervenes. In the absence of or under limited biological activity, $\delta^{18}O_P$ remains essentially unchanged and can be used with some caveats as a tracer of the origins of P, such as in studies of P transfer from land to surface waters (see as an example Pistocchi et al., 2017).

Larsen et al. (1989) first demonstrated the effect of soil biological activity on the $\delta^{18}O_P$ of soil phosphate by applying $KH_2P^{18}O_4$ to soil treated or untreated with a biocide. Half of the original ^{18}O enrichment was lost in soil phosphate after 3 months of incubation of the biologically active soil due to the cycling of P by soil microorganisms, whereas no label change was observed in the treated soils.

The biological processes causing cleavage of the P—O bond are important enzyme reactions such as the intra- and extra-cellular organic P mineralization through phosphoesterases and the intracellular reactions catalyzed by pyrophosphatase (PPase). The PPase-catalyzed reactions produce an equilibrium isotopic fractionation and complete exchange of the four O atoms of the phosphate with the O atoms of the water (Cohn, 1958). Hence, this equilibrium isotopic fractionation is considered a marker of P cell metabolism. In contrast, the mineralization of organic P compounds by phosphoesterases implies the exchange of one or two O atoms of the phosphate moiety with the surrounding water. The kinetic isotopic fractionation associated with this process is known for a small number of enzymes (ϵ between $+10\text{‰}$ and -20‰ , Liang and Blake, 2006; von Sperber et al., 2015).

The $\delta^{18}O_P$ at natural abundance has been used successfully in chronosequence studies to show how P in primary soil minerals was biologically cycled during pedogenesis (Tamburini et al., 2012; Frkova et al., 2022) or along a weathering gradient in Hawaii (Helfenstein et al., 2018). These studies showed that under conditions that are favorable to microbial activity (e.g. soil humidity), primary P is biologically cycled after a few hundred years, whereas under arid or semi-arid conditions the biological recycling of primary P can take thousands of years.

Experiments with ^{18}O -enriched compounds (either water or phosphate) have been used to infer the prevalent biological processes affecting P in soil solution, the main pool available to plants (Siegenthaler et al., 2020). Pistocchi et al. (2020) have shown that in a P-limited soil horizon, the available P had a $\delta^{18}O_P$ value that was close to that expected from the mineralization of organic P, thus this process was the main contributor to the available P pool. Labelled phosphate compounds have also been used in the field to follow the fate of P fertilizers into soil P pools (Joshi et al., 2016). Incubation experiments with doubly labelled substrate seem promising to study C and P cycles simultaneously in soils (Gross and Angert, 2017).

Oxygen in nitrate and sulfate

The isotope composition of oxygen in nitrate ($\delta^{18}O_{NO_3}$) is used to study nitrate sources and its fate (e.g. nitrate leaching) in terrestrial ecosystems, mostly under natural abundance. The soil nitrate $\delta^{18}O_{NO_3}$ (and $\delta^{15}N_{NO_3}$) may reflect 'natural conditions' or the effects of various anthropogenic impacts such as increased atmospheric N deposition and acidification. Nitrate from atmospheric deposition is strongly enriched in ^{18}O ($\delta^{18}O_{NO_3} > 60\text{‰}$). Synthetic fertilizers, where oxygen is derived from atmospheric O_2 (ca. $+23.5\text{‰}$), have $\delta^{18}O_{NO_3}$ values ranging from $+17\text{‰}$ to $+25\text{‰}$ (Kendall et al., 2007).

Among biological processes, denitrification leaves the residual nitrate pool relatively enriched in both ^{18}O and ^{15}N , with a constant ratio varying between 1:1 and 1:2. In contrast, the isotope effects that affect O and N during nitrification are unrelated because the sources of O and N atoms are different. During nitrification, the O atoms come from water and O_2 , thus the $\delta^{18}O_{NO_3}$ of the nitrate produced reflects the isotopic composition of both sources. Nitrate from nitrification is expected to have $\delta^{18}O_{NO_3}$ values in the range of -10‰ to $+10\text{‰}$. Uncertainties remain about isotopic fractionation associated with O incorporation from O_2 or water, and the degree of equilibrium oxygen isotope exchange with water during reversible steps of nitrification (Kendall et al., 2007; Kool et al., 2011).

The coupled pattern of N and O isotope effects associated with nitrate-consuming processes, such as denitrification and plant uptake, help to constrain the interpretation of isotope data when mixing of multiple nitrate sources can be neglected, such as in non fertilized ecosystems. The deviation from the coupled 1:1 pattern of the $\delta^{18}O_{NO_3}$ and $\delta^{15}N_{NO_3}$ can then be interpreted as the effect of processes inducing unrelated N and O isotope effects, like nitrification.

With the natural isotope abundance approach, the amount of nitrate derived from atmospheric deposition that is leached from soils can be quantified. Recently, Fang et al. (2015) coupled $\delta^{17}\text{O}$, which is an unambiguous tracer of atmospheric nitrate, to $\delta^{18}\text{O}_{\text{NO}_3}$ and $\delta^{15}\text{N}_{\text{NO}_3}$ and successfully quantified gross rates of nitrification and denitrification in a range of forest ecosystems.

Compared to oxygen isotope composition in nitrate, $\delta^{18}\text{O}_{\text{SO}_4}$ is less used in soil studies. As for phosphate, oxygen in sulfate (SO_4^{2-}) does not exchange easily with O atoms in water under common temperature and pressure conditions of terrestrial ecosystems. Sorption to soil particles does not induce isotopic fractionation. Therefore, the $\delta^{18}\text{O}_{\text{SO}_4}$ can be used to infer the extent of biological cycling affecting soil sulfate.

Krouse et al. (1996) observed that $\delta^{18}\text{O}_{\text{SO}_4}$ in aerated forest subsoil with negligible S inputs from parent material is smaller by several units than that in superficial pore water originating from rainwater, while $\delta^{34}\text{S}_{\text{SO}_4}$ is relatively similar. These differences are thought to be the result of S cycling through the biota (plants or microorganisms) by assimilatory reduction and subsequent mineralization of carbon-bound S to sulfate. Mineralization induces the incorporation of 45–75% O atoms into the newly formed sulfate from soil water and atmospheric O_2 (the majority of O incorporation is from water). An isotopic fractionation described by a negative ϵ ranging from -4.3‰ to -11.4‰ is associated with O incorporation from O_2 . Usually, $\delta^{18}\text{O}_{\text{SO}_4}$ values below 5‰ are expected for newly mineralized sulfate (Krouse et al., 1996 and references therein).

Under reducing conditions, such as in peatlands, the $\delta^{18}\text{O}_{\text{SO}_4}$ measured along the soil profile might indicate the extent and depth of dissimilatory sulfate reduction by anaerobic bacteria, which leaves the residual sulfate enriched in heavy isotopes (both O and S) (Novák et al., 2005).

Small values of $\delta^{18}\text{O}_{\text{SO}_4}$ in soil sulfate can also derive from parent material containing reduced S (e.g. pyrite), for example in glacial tills (-5‰ to -20‰ , Krouse et al., 1996 and references therein). On the other hand, sulfate from primary evaporites generally has larger $\delta^{34}\text{S}_{\text{SO}_4}$ and $\delta^{18}\text{O}_{\text{SO}_4}$ ($+9\text{‰}$ to $+18\text{‰}$) values than sulfate derived from oxidation of inorganic or organic S compounds and subsequent reprecipitation (secondary sulfates). Therefore, the study of both S and O isotopes in sulfate provides unique information on sulfate cycling in soils and on its origin from primary or secondary minerals.

As for nitrate, the $\delta^{17}\text{O}$ of sulfate can be used to assess the relative contribution of atmospheric sulfate to terrestrial inputs (Kendall et al., 2007).

Sulfur

Sulfur (S) has four stable isotopes (Table 1). Among these, ^{34}S is the only one used in soil studies, but far less than the stable isotopes of C and N. The majority of earlier studies concern forest ecosystems and understanding of the effects of anthropogenic S emissions on S cycling (Krouse et al., 1996). In recent years, S deficiencies in crops have been increasingly identified because of the worldwide reduction in industrial S emissions and the increase in demand for S by crops with increasing yields. Several recent studies address S dynamics and fertilization in agroecosystems with the use of stable isotopes (Chalk et al., 2017 and references therein).

Atmospheric $\delta^{34}\text{S}$ values are in the range of -5‰ to $+25\text{‰}$. Soil sulfur displays a very large range of $\delta^{34}\text{S}$ values (-30‰ to $+30\text{‰}$), while soil sulfate has a somewhat narrower range (-25‰ to $+20\text{‰}$; Chalk et al., 2017).

Dissimilatory sulfate reduction by microorganisms is known to produce the strongest isotopic fractionation, producing ^{34}S -depleted sulfide (S^{2-}), although the magnitude of the isotope fractionation factor is not known with precision. Isotopic fractionation associated with microbiological sulfide oxidation is less (ϵ up to $+12.5\text{‰}$). Even smaller isotopic fractionations are associated with S mineralization ($<+2\text{‰}$) and plant sulfate uptake ($<\pm 1\text{‰}$) (Chalk et al., 2017; Hoefs, 2021).

Measurements of $\delta^{34}\text{S}$ natural abundance in agroecosystems have been used to discriminate between sources (e.g. atmospheric deposition and fertilizers) that contribute to plant S uptake or to soil S. The difference in $\delta^{34}\text{S}$ between soil and anthropogenic S provides the basis to estimate the relative contribution of S from different sources. In long-term experiments, an increasing and decreasing trend of S concentrations in soil and crops followed the increase and decrease of anthropogenic S emissions during the last century. This trend in S concentrations was mirrored by the $\delta^{34}\text{S}$ values in soil S, which were highest in 1860, they were at a minimum in 1972, and then increased again later. This trend relates to the use of coal, the source of most anthropogenic S emissions, which was depleted in ^{34}S (Chalk et al., 2017).

Analogous to studies with ^{15}N enriched fertilizers, ^{34}S enriched (or depleted) fertilizers have been used to estimate S fertilizer efficiency, mostly in pot studies (Chalk et al., 2017).

Current trends and future developments

Multi-isotope and multi-tracer approaches that also integrate complementary analytical techniques are increasingly used in soil studies (see for example Helfenstein et al., 2018).

Compound-specific isotope probing has opened avenues for tracking pathways and components of the N and C cycles, such as decomposition, assimilation, and de novo synthesis of molecules and microorganisms involved obtained by ^{15}N -DNA probing, ^{13}C - and ^{18}O -DNA and RNA probing, and ^{13}C -PLFA (phospholipid-derived fatty acid) (examples of such approaches are given in Jochmann and Schmidt, 2012). However, these applications are still limited to the laboratory scale.

Such approaches together with analytical advances will probably enable an increasingly wide range of questions to be addressed regarding the cycling and sources of key elements of life in soil and the environment. Here, we report some promising examples and recent technical advances.

Site-specific ^{15}N analyses and clumped isotopes to study N_2O emissions

The biogeochemical cycle of N in soil systems has been studied by determining the isotopic composition of emitted N_2O and quantifying its source and sink processes (Decock and Six, 2013). The studies mostly focus on the four most abundant N_2O isotopocules: $^{14}\text{N}^{14}\text{N}^{16}\text{O}$, and singly substituted isotopocules $^{14}\text{N}^{15}\text{N}^{16}\text{O}$, $^{15}\text{N}^{14}\text{N}^{16}\text{O}$, and $^{14}\text{N}^{14}\text{N}^{18}\text{O}$, with resulting delta values $\delta^{15}\text{N}^{\text{bulk}}$ (the average abundance of $^{14}\text{N}^{15}\text{N}^{16}\text{O}$ and $^{15}\text{N}^{14}\text{N}^{16}\text{O}$) and $\delta^{18}\text{O}$ ($^{14}\text{N}^{14}\text{N}^{18}\text{O}$). A difference between the abundance of the two isotopomers $^{14}\text{N}^{15}\text{N}^{16}\text{O}$ and $^{15}\text{N}^{14}\text{N}^{16}\text{O}$, denoted as ^{15}N site preference (SP), has proved to be a useful tracer for distinguishing different microbial sources (denitrification vs nitrification, Sutka et al., 2006). However, interpretation of N_2O isotopic data from large, plot-scale studies is demanding because N_2O emissions and their isotopic signatures are influenced by many factors, e.g. soil moisture, soil pH, occurrence of large emission events in time (so called 'hot moments') or from particular locations ('hot spots'), or availability of reaction substrates.

Isotopic analysis of N_2O has undergone considerable development in recent years. Classical IRMS has been accompanied by methods based on laser spectroscopy (see Section "Analytical methods") that can be applied directly in the field and supply real-time data. Both methods, however, require a preconcentration technique for focusing the analyte into a smaller volume and removing potential interfering compounds (CO_2 , CH_4 , etc.).

A promising new tracer that will shed more light on N dynamics is the abundance of multiply substituted N_2O isotopocules, so called 'clumped isotopes' ($^{14}\text{N}^{15}\text{N}^{18}\text{O}$, $^{15}\text{N}^{14}\text{N}^{18}\text{O}$, $^{15}\text{N}^{15}\text{N}^{16}\text{O}$, Ostrom and Ostrom, 2017). An analysis of clumped N_2O is already available through new IRMS and laser spectroscopy methods, but has been applied so far only to laboratory-based experiments (N_2O produced by denitrification and degraded by UV photolysis, Kantnerová et al., 2020a; Kantnerová et al., 2020b). For clumped isotope analysis in soil-emitted N_2O , the current preconcentration techniques need to be modified to process larger amounts of N_2O that the new methods currently require.

Isotopic analysis of oxyanions, bulk isotopes composition of total dissolved C and N and of soil carbon pools

Until recently, it has not been possible to analyze oxyanions of the key elements of life directly, e.g. nitrate, phosphate, or sulfate. Prior to the analysis, the analyte has to be converted to a low-molecular weight gas (e.g. for nitrate to N_2O or to N_2 and O_2), which might lead to a loss of the original isotopic information. Over the last 5 years, a new methodology for stable isotope analysis has been under development, using Fourier-transform mass spectrometry with Orbitrap™ mass analyzers (Eiler et al., 2017). Its sample introduction technique does not destroy the oxyanions, therefore they can be analyzed directly and the original information is retained. In addition, because of its sensitivity, the method makes isotope clumping in the oxyanions analytically accessible. So far, it has been shown that the technique can be applied to samples of nitrate, sulfate, phosphate, and even acetate (Mueller et al., 2022 and references therein).

Tools are now available to analyze simultaneously the bulk isotopic composition of total dissolved C and N in liquid samples without offline sample preparation, which was previously impossible. This is achieved by coupling a modified high temperature combustion TOC (total organic carbon) analyzer to an IRMS system (Federherr et al., 2016). Similarly, a TOC analyzer coupled to a Picarro $\delta^{13}\text{C}$ - CO_2 analyzer enables the total dissolved inorganic and organic C isotopes to be measured in liquid samples.

To characterize different soil organic and inorganic C fractions, a novel approach has recently been developed that exploits their distinctive thermal stability (Natali et al., 2018). The thermally based separation is achieved through an automated elemental analyzer (EA), which is coupled with an IRMS. This enables the isotopic analysis of soil organic C pools and soil inorganic C, and avoids the isotopic fractionation of SOM potentially induced by the acid pre-treatment commonly used to strip inorganic C. This analytical approach provides a more in-depth characterization of the composition of SOM fractions and their evolution with physicochemical and biological processes.

Such technical advances will provide so far unobtainable isotope information that can be used to understand further the biogeochemical cycling of many elements and simplify current analytical procedures.

References

Introduction

- Barrie A, Prosser S, and J. (1996) Automated analysis of light-element stable isotope ratio mass spectrometry. In: Button TW and Yamasaki S-i (eds.) *Mass Spectrometry of Soils*. New York, Basel, Hong Kong: Marcel Dekker, Inc.
- Brand WA (2004) Mass spectrometer hardware for analyzing stable isotope ratios. In: de Groot PA (ed.) *Handbook of Stable Isotope Analytical Techniques*. vol. 1. Amsterdam: Elsevier.
- Brand WA and Coplen TB (2012) Stable isotope deltas: Tiny, yet robust signatures in nature. *Isotopes in Environmental and Health Studies* 48(3): 393–409. <https://doi.org/10.1080/10256016.2012.666977>.
- Chalk PM, Inácio CT, and Chen D (2017a) Tracing S dynamics in agro-ecosystems using ^{34}S . *Soil Biology and Biochemistry* 114: 295–308. <https://doi.org/10.1016/j.soilbio.2017.07.001>.
- Coplen TB (2011) Guidelines and recommended terms for expression of stable-isotope-ratio and gas-ratio measurement results. *Rapid Communications in Mass Spectrometry* 25(17): 2538–2560. <https://doi.org/10.1002/rcm.5129>.
- Corso TN and Brenna TJ (1997) High-precision position-specific isotope analysis. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.94.4.1049>.

- Friedman I and O'Neil JR (1977) *Data of Geochemistry: Compilation of stable isotope fractionation factors of geochemical interest. Chapter KK*, vol. 440. US Government Printing Office.
- Francey RJ and Tans PP (1987) Latitudinal variation in oxygen-18 of atmospheric CO₂. *Nature* 327: 495–497.
- Hoefs J (2021) *Stable Isotope Geochemistry*. <https://doi.org/10.1007/978-3-030-77692-3>.
- Horita J, Rozanski K, and Cohen S (2008) Isotope effects in the evaporation of water: A status report of the Craig–Gordon model. *Isotopes in Environmental and Health Studies* 44(1): 23–49. <https://doi.org/10.1080/10256010801887174>.
- IUPAC (1997) In: McNaught AD and Wilkinson A (eds.) *Compendium of Chemical Terminology*. Oxford: Blackwell Scientific Publications. <https://doi.org/10.1351/goldbook>.
- Jochmann MA and Schmidt TC (2012a) *Compound-specific Stable Isotope Analysis*. Royal Society of Chemistry Publishing.
- Li JS, Chen W, and Fischer H (2013) Quantum cascade laser spectrometry techniques: A new trend in atmospheric chemistry. *Applied Spectroscopy Reviews* 48(7): 523–559. <https://doi.org/10.1080/05704928.2012.757232>.
- McKinney CR, McCrea JM, Epstein S, Allen HA, and Urey HC (1950) Improvements in mass spectrometers for the measurement of small differences in isotope abundance ratios. *Review of Scientific Instruments* 21(8): 724–730.
- Rishavy MA and Cleland WW (1999) ¹³C, ¹⁵N, and ¹⁸O equilibrium isotope effects and fractionation factors. *Canadian Journal of Chemistry* 77(5–6): 967–977. <https://doi.org/10.1139/cjc-77-5-6-967>.
- Robinson D (2001a) $\delta^{15}\text{N}$ as an integrator of the nitrogen cycle. *Trends in Ecology & Evolution* 16(3): 153–162. [https://doi.org/10.1016/S0169-5347\(00\)02098-X](https://doi.org/10.1016/S0169-5347(00)02098-X).
- Sharp Z (2017) *Principles of Stable Isotope Geochemistry*, 2nd edn.
- Werner RA and Cormier M-A (2022) Isotopes—Terminology, definitions and properties. In: Siegwolf RTW, et al. (ed.) *Stable Isotopes in Tree Rings: Inferring Physiological, Climatic and Environmental Responses*, pp. 253–289. Cham: Springer International Publishing (Tree Physiology). https://doi.org/10.1007/978-3-030-92698-4_8.

Application of isotopes of light elements in soil science

Carbon

- Balesdent J, et al. (2018) Atmosphere–soil carbon transfer as a function of soil depth. *Nature* 559(7715): 599–602. <https://doi.org/10.1038/s41586-018-0328-3>.
- Balesdent J and Balabane M (1996) Major contribution of roots to soil carbon storage inferred from maize cultivated soils. *Soil Biology and Biochemistry* 28(9): 1261–1263. [https://doi.org/10.1016/0038-0717\(96\)00112-5](https://doi.org/10.1016/0038-0717(96)00112-5).
- Balesdent J, Girardin C, and Mariotti A (1993) Site-related ¹³C of tree leaves and soil organic matter in a temperate forest. *Ecology* 74(6): 1713–1721. <https://doi.org/10.2307/1939930>.
- Bernard L, et al. (2007) Dynamics and identification of soil microbial populations actively assimilating carbon from ¹³C-labelled wheat residue as estimated by DNA- and RNA-SIP techniques. *Environmental Microbiology* 9(3): 752–764. <https://doi.org/10.1111/j.1462-2920.2006.01197.x>.
- Cotrufo MF, et al. (2015) Formation of soil organic matter via biochemical and physical pathways of litter mass loss. *Nature Geoscience* 8: 776–779. <https://doi.org/10.1038/ngeo2520>.
- Fanin N and Bertrand I (2016) Aboveground litter quality is a better predictor than belowground microbial communities when estimating carbon mineralization along a land-use gradient. *Soil Biology and Biochemistry* 94: 48–60. <https://doi.org/10.1016/j.soilbio.2015.11.007>.
- Heuck C, Weig A, and Spohn M (2015) Soil microbial biomass C:N:P stoichiometry and microbial use of organic phosphorus. *Soil Biology and Biochemistry* 85: 119–129. <https://doi.org/10.1016/j.soilbio.2015.02.029>.
- Hoefs J (2021) *Stable Isotope Geochemistry*. <https://doi.org/10.1007/978-3-030-77692-3>.
- Huo C, Luo Y, and Cheng W (2017) Rhizosphere priming effect: A meta-analysis. *Soil Biology and Biochemistry* 111: 78–84. <https://doi.org/10.1016/j.soilbio.2017.04.003>.
- Jamoteau F, et al. (2021) Can stable isotopes quantify soil carbon build-up from organic fertilizers? *Isotopes in Environmental and Health Studies* 57(5): 470–491. <https://doi.org/10.1080/10256016.2021.1946532>.
- Leake JR, et al. (2006) Carbon fluxes from plants through soil organisms determined by field ¹³CO₂ pulse-labelling in an upland grassland. *Applied Soil Ecology* 33(2): 152–175. <https://doi.org/10.1016/j.apsoil.2006.03.001>.
- Paul A, Balesdent J, and Hatté C (2020) ¹³C–¹⁴C relations reveal that soil ¹³C-depth gradient is linked to historical changes in vegetation ¹³C. *Plant and Soil* 447(1): 305–317. <https://doi.org/10.1007/s11104-019-04384-4>.
- Perveen N, et al. (2019) Universality of priming effect: An analysis using thirty five soils with contrasted properties sampled from five continents. *Soil Biology and Biochemistry* 134: 162–171. <https://doi.org/10.1016/j.soilbio.2019.03.027>.
- Putz B, et al. (2011) A simple method for in situ-labelling with ¹⁵N and ¹³C of grassland plant species by foliar brushing. *Methods in Ecology and Evolution* 2(3): 326–332. <https://doi.org/10.1111/j.2041-210X.2010.00072.x>.
- Sauvadet M, et al. (2018) High carbon use efficiency and low priming effect promote soil C stabilization under reduced tillage. *Soil Biology and Biochemistry* 123: 64–73. <https://doi.org/10.1016/j.soilbio.2018.04.026>.
- Villegas DM, et al. (2020) Urochloa grasses swap nitrogen source when grown in association with legumes in tropical pastures. *Diversity* 12(11): 419.
- Werth M and Kuzyakov Y (2010) ¹³C fractionation at the root–microorganisms–soil interface: A review and outlook for partitioning studies. *Soil Biology and Biochemistry* 42(9): 1372–1384. <https://doi.org/10.1016/j.soilbio.2010.04.009>.

Nitrogen

- Bosshard C, et al. (2008) Incorporation of nitrogen-15-labeled amendments into physically separated soil organic matter fractions. *Soil Science Society of America Journal* 72(4): 949–959. <https://doi.org/10.2136/sssaj2006.0376>.
- Braun J, et al. (2018) Full ¹⁵N tracer accounting to revisit major assumptions of ¹⁵N isotope pool dilution approaches for gross nitrogen mineralization. *Soil Biology and Biochemistry* 117: 16–26. <https://doi.org/10.1016/j.soilbio.2017.11.005>.
- Büchi L, et al. (2015) Accumulation of biologically fixed nitrogen by legumes cultivated as cover crops in Switzerland. *Plant and Soil* 393(1): 163–175. <https://doi.org/10.1007/s11104-015-2476-7>.
- Chalk PM, et al. (2016) Do techniques based on ¹⁵N enrichment and ¹⁵N natural abundance give consistent estimates of the symbiotic dependence of N₂-fixing plants? *Plant and Soil* 399(1): 415–426. <https://doi.org/10.1007/s11104-015-2689-9>.
- Choi W-J, et al. (2017) Synthetic fertilizer and livestock manure differently affect $\delta^{15}\text{N}$ in the agricultural landscape: A review. *Agriculture, Ecosystems & Environment* 237: 1–15. <https://doi.org/10.1016/j.agee.2016.12.020>.
- Douxchamps S, et al. (2011) Nitrogen recoveries from organic amendments in crop and soil assessed by isotope techniques under tropical field conditions. *Plant and Soil* 341(1): 179–192. <https://doi.org/10.1007/s11104-010-0633-6>.
- Frick H, et al. (2022) Similar distribution of ¹⁵N labeled cattle slurry and mineral fertilizer in soil after one year. *Nutrient Cycling in Agroecosystems*. <https://doi.org/10.1007/s10705-022-10205-5>.
- Hammelehle A, et al. (2018) Above- and belowground nitrogen distribution of a red clover-perennial ryegrass sward along a soil nutrient availability gradient established by organic and conventional cropping systems. *Plant and Soil* 425(1): 507–525. <https://doi.org/10.1007/s11104-018-3559-z>.
- Kahmen A, Wanek W, and Buchmann N (2008) Foliar $\delta^{15}\text{N}$ values characterize soil N cycling and reflect nitrate or ammonium preference of plants along a temperate grassland gradient. *Oecologia* 156(4): 861–870. <https://doi.org/10.1007/s00442-008-1028-8>.
- Manandhar S, et al. (2022) Stable isotope techniques for quantifying N fertiliser recovery in maize grown in soils with different management histories. *Nutrient Cycling in Agroecosystems* 122(1): 59–72. <https://doi.org/10.1007/s10705-021-10181-2>.

- Oberson A, et al. (2007) Symbiotic N₂ fixation by soybean in organic and conventional cropping systems estimated by ¹⁵N dilution and ¹⁵N natural abundance. *Plant and Soil* 290(1): 69–83. <https://doi.org/10.1007/s11104-006-9122-3>.
- Oberson A, et al. (2013) Nitrogen fixation and transfer in grass-clover leys under organic and conventional cropping systems. *Plant and Soil* 371(1): 237–255. <https://doi.org/10.1007/s11104-013-1666-4>.
- Pirhofer-Walzl K, et al. (2012) Nitrogen transfer from forage legumes to nine neighbouring plants in a multi-species grassland. *Plant and Soil* 350(1): 71–84. <https://doi.org/10.1007/s11104-011-0882-z>.
- Robinson D (2001b) $\delta^{15}\text{N}$ as an integrator of the nitrogen cycle. *Trends in Ecology & Evolution* 16(3): 153–162. [https://doi.org/10.1016/S0169-5347\(00\)02098-X](https://doi.org/10.1016/S0169-5347(00)02098-X).
- Shearer G and Kohl DH (1986) N₂-Fixation in field settings: Estimations based on natural ¹⁵N abundance. *Functional Plant Biology* 13(6): 699–756. <https://doi.org/10.1071/pp9860699>.
- Smith CJ and Chalk PM (2018) The residual value of fertiliser N in crop sequences: An appraisal of 60 years of research using ¹⁵N tracer. *Field Crops Research* 217: 66–74. <https://doi.org/10.1016/j.fcr.2017.12.006>.
- Teutschnerová N, et al. (2022) Gross N transformation rates in soil system with contrasting Urochloa genotypes do not confirm the relevance of BNI as previously assessed in vitro. *Biology and Fertility of Soils* 58(3): 321–331. <https://doi.org/10.1007/s00374-021-01610-z>.
- Thomsen IK, Kjellerup V, and Jensen B (1997) Crop uptake and leaching of ¹⁵N applied in ruminant slurry with selectively labelled faeces and urine fractions. *Plant and Soil* 197(2): 233–239. <https://doi.org/10.1023/A:1004224615075>.
- Traoré O, Y. A., et al. (2020) Nitrogen and phosphorus uptake from isotope-labeled fertilizers by sorghum and soil microorganisms. *Agrosystems, Geosciences & Environment* 3(1): e20111. <https://doi.org/10.1002/agg2.20111>.

Oxygen

- Cohn M (1958) Phosphate-water exchange reaction catalyzed by inorganic pyrophosphatase of yeast. *Journal of Biological Chemistry* 230(1): 369–380.
- Fang Y, et al. (2015) Microbial denitrification dominates nitrate losses from forest ecosystems. *Proceedings of the National Academy of Sciences* 112(5): 1470–1474. <https://doi.org/10.1073/pnas.1416776112>.
- Frkova Z, et al. (2022) Phosphorus dynamics during early soil development in a cold desert: Insights from oxygen isotopes in phosphate. *Soil* 8(1): 1–15. <https://doi.org/10.5194/soil-8-1-2022>.
- Gat JR (2010) *Isotope Hydrology: A Study of the Water Cycle*. World Scientific.
- Gessler A, et al. (2014) Stable isotopes in tree rings: Towards a mechanistic understanding of isotope fractionation and mixing processes from the leaves to the wood. *Tree Physiology* 34: 796–818. <https://doi.org/10.1093/treephys/tpu040>.
- Gross A and Angert A (2017) Use of ¹³C- and phosphate ¹⁸O-labeled substrate for studying phosphorus and carbon cycling in soils: a proof of concept. *Rapid Communications in Mass Spectrometry* 31(11): 969–977. <https://doi.org/10.1002/rcm.7863>.
- Helfenstein J, et al. (2018) Combining spectroscopic and isotopic techniques gives a dynamic view of phosphorus cycling in soil. *Nature Communications* 9(1): 3226. <https://doi.org/10.1038/s41467-018-05731-2>.
- Joshi SR, Li X, and Jaisi DP (2016) Transformation of phosphorus pools in an agricultural soil: An application of oxygen-18 labeling in phosphate. *Soil Science Society of America Journal* 80(1): 69. <https://doi.org/10.2136/sssaj2015.06.0219>.
- Kendall C, Elliott EM, and Wankel SD (2007) Tracing anthropogenic inputs of nitrogen to ecosystems. In: Michener RH and Lajtha K (eds.) *Stable Isotopes in Ecology and Environmental Science*, pp. 375–449. Blackwell Publishing.
- Kool DM, et al. (2011) Oxygen exchange with water alters the oxygen isotopic signature of nitrate in soil ecosystems. *Soil Biology and Biochemistry* 43(6): 1180–1185. <https://doi.org/10.1016/j.soilbio.2011.02.006>.
- Krouse HR, Mayer B, and Schoenau JJ (1996a) Applications of stable isotope techniques to soil sulfur cycling. In: Button Shin-ichi Yamasaki TW (ed.) *Mass Spectrometry of Soils*, pp. 247–284. CRC Press.
- Larsen S, Middelboe V, and Johansen HS (1989) The fate of ¹⁸O labelled phosphate in soil/plant systems. *Plant and Soil* 117(1): 143–145.
- Liang Y and Blake RE (2006) Oxygen isotope composition of phosphate in organic compounds: Isotope effects of extraction methods. *Organic Geochemistry* 37(10): 1263–1277. <https://doi.org/10.1016/j.orggeochem.2006.03.009>.
- Meunier F, et al. (2018) Measuring and modeling hydraulic lift of *Lolium multiflorum* using stable water isotopes. *Vadose Zone Journal* 17(1): 160134. <https://doi.org/10.2136/vzj2016.12.0134>.
- Novák M, et al. (2005) Isotope systematics of sulfate-oxygen and sulfate-sulfur in six European peatlands. *Biogeochemistry* 76(2): 187–213. <https://doi.org/10.1007/s10533-005-4433-7>.
- Pistocchi C, et al. (2017) Tracing the sources and cycling of phosphorus in river sediments using oxygen isotopes: Methodological adaptations and first results from a case study in France. *Water Research* 111: 346–356. <https://doi.org/10.1016/j.watres.2016.12.038>.
- Pistocchi C, et al. (2020) In or out of equilibrium? How microbial activity controls the oxygen isotope signature of phosphate in forest organic horizons with low and high phosphorus availability. *Frontiers in Environmental Science*. <https://doi.org/10.3389/fenvs.2020.564778>.
- da Silveira Lobo Sternberg L (1989) Oxygen and hydrogen isotope ratios in plant cellulose: Mechanisms and applications. In: Rundel PW, Ehleringer JR, and Nagy KA (eds.) *Stable Isotopes in Ecological Research*, pp. 124–141. New York, NY: Springer (Ecological Studies). https://doi.org/10.1007/978-1-4612-3498-2_9.
- da Silveira Lobo O'Reilly Sternberg L and DeNiro MJD (1983) Biogeochemical implications of the isotopic equilibrium fractionation factor between the oxygen atoms of acetone and water. *Geochimica et Cosmochimica Acta* 47(12): 2271–2274. [https://doi.org/10.1016/0016-7037\(83\)90049-2](https://doi.org/10.1016/0016-7037(83)90049-2).
- Schmidt H-L, Werner RA, and Rossmann A (2001) ¹⁸O pattern and biosynthesis of natural plant products. *Phytochemistry* 58: 9–32. [https://doi.org/10.1016/S0031-9422\(01\)00017-6](https://doi.org/10.1016/S0031-9422(01)00017-6).
- Siegenthaler MB, et al. (2020) A dual isotopic (³²P and ¹⁸O) incubation study to disentangle mechanisms controlling phosphorus cycling in soils from a climatic gradient (Kohala, Hawaii). *Soil Biology and Biochemistry* 149: 107920. <https://doi.org/10.1016/j.soilbio.2020.107920>.
- von Sperber C, et al. (2015) The oxygen isotope composition of phosphate released from phytic acid by the activity of wheat and *Aspergillus niger* phytase. *Biogeosciences* 12(13): 4175–4184. <https://doi.org/10.5194/bg-12-4175-2015>.
- Sprenger M, et al. (2016) Illuminating hydrological processes at the soil-vegetation-atmosphere interface with water stable isotopes. *Reviews of Geophysics* 54(3): 674–704. <https://doi.org/10.1002/2015RG000515>.
- Tamburini F, et al. (2012) Oxygen isotopes unravel the role of microorganisms in phosphate cycling in soils. *Environmental Science & Technology* 46(11): 5956–5962. <https://doi.org/10.1021/es300311h>.
- Yakir D and da Sternberg L (2000) The use of stable isotopes to study ecosystem gas exchange. *Oecologia* 123(3): 297–311. <https://doi.org/10.1007/s004420051016>.

Sulfur

- Chalk PM, Inácio CT, and Chen D (2017b) Tracing S dynamics in agro-ecosystems using ³⁴S. *Soil Biology and Biochemistry* 114: 295–308. <https://doi.org/10.1016/j.soilbio.2017.07.001>.
- Krouse HR, Mayer B, and Schoenau JJ (1996b) Applications of stable isotope techniques to soil sulfur cycling. In: Button Shin-ichi Yamasaki TW (ed.) *Mass Spectrometry of Soils*, pp. 247–284. CRC Press.
- Hoefs J (2021) *Stable Isotope Geochemistry*. <https://doi.org/10.1007/978-3-030-77692-3>.

Current trends and future developments

- Decock C and Six J (2013) How reliable is the intramolecular distribution of ^{15}N in N_2O to source partition N_2O emitted from soil? *Soil Biology and Biochemistry* 65: 114–127. <https://doi.org/10.1016/j.soilbio.2013.05.012>.
- Eiler J, et al. (2017) Analysis of molecular isotopic structures at high precision and accuracy by Orbitrap mass spectrometry. *International Journal of Mass Spectrometry* 422: 126–142. <https://doi.org/10.1016/j.ijms.2017.10.002>.
- Federherr E, et al. (2016) A novel tool for stable nitrogen isotope analysis in aqueous samples. *Rapid Communications in Mass Spectrometry* 30(23): 2537–2544. <https://doi.org/10.1002/rcm.7740>.
- Jochmann MA and Schmidt TC (2012b) *Compound-Specific Stable Isotope Analysis*. Royal Society of Chemistry Publishing.
- Kantnerová K, Yu L, et al. (2020a) First investigation and absolute calibration of clumped isotopes in N_2O by mid-infrared laser spectroscopy. *Rapid Communications in Mass Spectrometry* 34(15): e8836. <https://doi.org/10.1002/rcm.8836>.
- Kantnerová K, Jespersen MF, et al. (2020b) Photolytic fractionation of seven singly and doubly substituted nitrous oxide isotopocules measured by quantum cascade laser absorption spectroscopy. *Atmospheric Environment X*, 8: 100094. <https://doi.org/10.1016/j.aeaoa.2020.100094>.
- Mueller EP, et al. (2022) Simultaneous, high-precision measurements of $\delta^2\text{H}$ and $\delta^{13}\text{C}$ in nanomole quantities of acetate using electrospray ionization-quadrupole-Orbitrap mass spectrometry. *Analytical Chemistry* 94(2): 1092–1100. <https://doi.org/10.1021/acs.analchem.1c04141>.
- Natali C, Bianchini G, and Antisari LV (2018) Thermal separation coupled with elemental and isotopic analysis: A method for soil carbon characterisation. *Catena* 164: 150–157. <https://doi.org/10.1016/j.catena.2018.02.022>.
- Ostrom NE and Ostrom PH (2017) Mining the isotopic complexity of nitrous oxide: A review of challenges and opportunities. *Biogeochemistry* 132(3): 359–372. <https://doi.org/10.1007/s10533-017-0301-5>.
- Sutka RL, et al. (2006) Distinguishing nitrous oxide production from nitrification and denitrification on the basis of isotopomer abundances. *Microbial Ecology* 72(1): 638–644. <https://doi.org/10.1128/AEM.72.1.638-644.2006>.

The use of radioactive isotopes to study soil pools and processes

Timothy I McLaren, School of Agriculture and Food Sciences, The University of Queensland, St Lucia, QLD, Australia

© 2023 Elsevier Ltd. All rights reserved.

Introduction	96
Basic principles of radioactive isotopes	97
Units of radioactivity	97
Detection and measures of radioactivity	98
Radioactive isotopes to study soil pools and processes	98
Isotopic dilution in soil-solution systems	98
'Available' phosphorus in soil—Isotopically exchangeable phosphorus (E-value)	99
Soil phosphorus dynamics—Phosphorus isotope exchange kinetics (IEK)	100
Gross phosphorus mineralization	102
Isotopic dilution in soil-plant systems	103
'Available' phosphorus in soil—Isotopically exchangeable phosphorus (L-value)	103
Radioactive isotopes to study the fate of fertilizer	103
Direct labeling approach	104
Indirect labeling approach	104
Limitations of radiotracers	105
Future opportunities for radiotracers	105
Conclusions	106
References	106

Abstract

Radioactive isotopes have been applied in soil science contexts for over 70 years. Their use in soil science is largely based on the isotopic dilution principle. A range of soil pools and processes can be quantified with the use of radioactive isotopes, which would otherwise be difficult to do based on conventional approaches alone. This chapter covers the basic principles and main approaches of using radioactive isotopes in soil science applications. Their continued use will be essential in helping to address major challenges of global concern (e.g., food security and climate change).

Key points

- Radioactive isotopes are an important tool to study pools and processes in soil.
- Approaches in soil science contexts are largely based on the isotopic dilution principle.
- Several methodological factors and assumptions need to be met when applying the isotopic dilution principle.
- Future research will benefit from a holistic view involving multiple isotopes (e.g., C-14, N-15, and P-33) and analytical techniques (e.g., chromatographic and spectroscopic approaches).

Introduction

The soil environment is critical to all life on Earth. Soil is an important source of nutrients for plants and microorganisms, contains a myriad of microorganisms that degrade organic residues and biowastes, and is a vital reservoir of carbon (C) as organic matter. Furthermore, understanding the quantity and turnover of nutrients in soils can lead to greater food security and improved water quality of aquatic or marine ecosystems. Consequently, soil is a key area of concern and new insights on nutrient pools and processes in soil will help address major global challenges such as climate change and food security.

Crop production largely relies on the supply of essential plant nutrients from the soil environment (Frossard et al., 2000). However, soils are often deficient in these nutrients for optimum crop growth (McLaughlin et al., 2011). In contrast, some soils contain an excess of nutrients possibly due to overfertilization which represents a potential risk of transfer to aquatic or marine ecosystems. An important aspect of sustainable crop production involves understanding the quantity and transfer of nutrients in soil. However, new insight on nutrient pools and processes in soil has been limited due to three main factors (Frossard et al., 2011). This includes the difficulty to: (1) distinguish the relative importance of various abiotic and biotic processes in complex systems; (2) assess small changes in nutrient pools compared to that of the total pool of nutrients in soil (differences can be at least one order of magnitude), and; (3) investigate changes in nutrient transfer that can occur at differing timescales (e.g., seconds to years). Addressing these difficulties in agroecosystems can lead to improved crop recovery of fertilizer, reduced fertilizer inputs, and less nutrient transfer to aquatic or marine ecosystems (McLaughlin et al., 2011).

The addition of mineral fertilizers and organic amendments to soils has been an important strategy to improve crop and pasture production. However, direct information on the crop recovery of nutrients released from mineral fertilizers and organic amendments remains limited (McLaughlin et al., 2011). There is also limited information on how agronomic management strategies improve the crop recovery of nutrients from mineral fertilizers and organic amendments. Various approaches have been proposed to track the fate of nutrients from mineral fertilizers and organic amendments in the soil-plant system (Johnston et al., 2014). However, approaches that use radioactive isotopes are perhaps the most important and informative (McLaughlin et al., 2011).

For over 70 years, radioactive isotopes have been an important tool in understanding nutrient pools in soil and their dynamics, and to track the fate of nutrients in the soil-plant system (Frossard et al., 2011). This chapter examines the basic principles and some of the main applications of using radioactive isotopes to probe nutrient pools and their transformation in soil, and as a tracer to assess the fate of fertilizers in soil-plant systems.

Basic principles of radioactive isotopes

At a foundational level, all matter in soil is comprised of atoms from a diverse array of elements. These elements consist of atoms that differ in their atomic number (symbol 'Z'), i.e., the number of protons. For example, all C atoms have six protons ($Z = 6$) whereas all nitrogen (N) atoms have seven protons ($Z = 7$) in the nucleus. Furthermore, nearly all naturally occurring elements comprise multiple isotopes. Isotopes of a particular element consist of atoms that have the same atomic number (protons), but differ in their mass number (symbol 'A'), i.e., the number of protons and neutrons. For example, two important isotopes of N include N-14 with seven protons and seven neutrons ($A = 14$) and N-15 with seven protons and eight neutrons ($A = 15$) in the nucleus. Isotopes of a given element will all exhibit the same chemical behavior, but may undergo various chemical, biological, physical processes at slightly different rates because of small differences in mass. The relative abundance of these naturally occurring isotopes differs among all elements. For example, all naturally occurring samples will consist of N-14 at 99.6% abundance and N-15 at 0.4% abundance.

There are two main classes of isotopes, i.e., 'stable' and 'unstable' isotopes. Stable isotopes refer to atoms with a nucleus of stable configuration which remain constant with time, and do not undergo radioactive decay. In contrast, unstable (or radioactive) isotopes refer to atoms with a nucleus of unstable configuration which will transform over time, and will undergo radioactive decay until a stable configuration is reached. For example, the radioactive isotope of C-14 ($Z = 6$ and $A = 14$) decays to the stable isotope of N-14 ($Z = 7$ and $A = 14$). The process of radioactive decay involves the emission of (radiation) energy.

A major concern with radioactive decay is the emission of ionizing radiation, which potentially has enough energy to remove electrons from atoms and cause the atom to be charged (or ionized). There are several different types of radioactive decay with varying levels of ionizing 'power' and penetration depth. Some important types include: (1) α , alpha particles (helium particles); (2) β^- , beta particles (negatively charged electrons); (3) β^+ , positrons (positively charged electrons), and; (4) γ , gamma rays (photons). Alpha particles are strongly ionizing due to their charge and mass, but their penetration depth is weak and can be easily shielded (e.g., air gap of a few centimeters, skin, or a sheet of paper). Beta particles (and positrons) are moderately ionizing due to their small mass, but can penetrate to greater depths than alpha particles. Beta particles require moderate shielding (e.g., 10 mm thick sheets of acrylic). Gamma rays are of high energy but have no mass or charge, and are the least ionizing form of radiation. However, gamma rays also have the greatest penetration depth and require the most shielding (e.g., thick lead bricks).

All radioactive decay processes follow first-order kinetics, but the rates of decay differ among isotopes. The law of radioactive decay describes the change in nuclei with time (Eq. 1). The 'half-life' ($t_{1/2}$) of a radioactive isotope can then be determined, which refers to the time needed for a given quantity of an unstable isotope to reduce by half. The half-life of a radioactive isotope can vary widely among isotopes (e.g., seconds to millions of years), and is calculated using Eq. (2).

$$N_t = N_0 \times e^{-\lambda t} \quad (1)$$

where N_t refers to the number of nuclei after a given time, t , N_0 refers to the number of nuclei at the start ($t = 0$), and λ refers to the decay constant.

$$t_{1/2} = \frac{0.693}{\lambda} \quad (2)$$

where $t_{1/2}$ refers to the time needed for a given quantity of an unstable isotope to reduce by half (e.g., minutes, days, or years), and λ refers to the decay constant.

There are over 3000 known radioactive isotopes, but the vast majority are made artificially and have half-lives less than 60 min. This makes them impractical for use in most soil science contexts. Furthermore, many radioactive isotopes are not commercially available, but some could potentially be produced in nuclear reactors or cyclotrons for research purposes. Table 1 shows some important radioactive isotopes used in soil science applications.

Units of radioactivity

The becquerel (Bq) is the international unit (SI unit) of radioactivity, which is defined as one nuclear disintegration per second (DPS) (FAO/IAEA, 2016). The non-SI unit of the curie (Ci) is still widely reported in some countries and recently defined as 1 Ci equals 3.7×10^{10} Bq, based on the number of DPS in 1 gram of radium.

Table 1 Some common radioactive isotopes used in soil science contexts.

Parent isotope (radioactive)	Daughter isotope (stable)	Half-life	Main type(s) of energy emitted	Maximum decay energy (MeV)
C-14	N-14	5700.00 y	Beta	0.16
K-42	Ca-42	12.36 h	Beta	3.53
P-32	S-32	14.27 d	Beta	1.71
P-33	S-33	25.35 d	Beta	0.25
S-35	Cl-35	87.37 d	Beta	0.17
Zn-65	Cu-65	243.93 d	Beta (Positron) + Gamma	1.35

Data sourced from the IAEA (2022).

Detection and measures of radioactivity

For most soil science applications, activities in the kilobecquerels (kBq) or megabecquerels (MBq) range will be required. Some of the factors that need to be considered include: (1) the safety regulations on the use, handling, transport, and disposal of radioactive isotopes; (2) the half-life of the radioactive isotope; (3) the analytical method and experimental approach; (4) soil properties, target element pool, or plant selection; (5) the duration of the experiment; (6) the estimated recovery of the radioactive isotope in the sample for measurement, and; (7) the available budget.

Measures of radioactivity from beta emitters primarily occur in solution using a liquid scintillation counter (LSC) (FAO/IAEA, 2016). This typically involves a known aliquot of the unknown solution in a scintillation vial and a known volume of scintillation cocktail, which converts radiation energy to photons of light and is detected by a photomultiplier in the LSC. Measures of radioactivity by the LSC in scintillation vials are generally given as counts per minute (CPM), which is proportional to the concentration of the radioactive isotope in the sample. However, the transfer of radiation energy into photons of light can be reduced due to the inefficient transfer of energy to the solvent and scintillator molecules (termed 'quenching') (Blair and Till, 2003). Two important sources of quenching are chemical quenching and color quenching. To correct for this a chemical or color quench curve needs to be established for each type of matrix to determine counting efficiency. A measure of counting efficiency is then used to convert CPM to disintegrations per minute (DPM) (Eq. 3), which can then be converted into Bq (1 Bq = 60 DPM) (FAO/IAEA, 2016).

$$\text{DPM} = \frac{\text{CPM}}{\text{Counting efficiency (as a decimal)}} \quad (3)$$

where DPM refers to disintegrations per minute and CPM refers to counts per minute.

The isotopic composition of the element is called the specific radioactivity, or the specific activity, which is the ratio of the radioactive isotope compared to the naturally occurring element (Nguyen et al., 2016). Specific activity is the ratio of radioactivity per unit of material (e.g., Bq of P-33 kg⁻¹ of soil) or per unit of the non-labeled element being traced (e.g., Bq of P-33 mg⁻¹ of P).

Radioactive isotopes can be obtained as 'carrier-free', i.e., with a negligible amount of the non-labeled element or with a 'carrier', i.e., with a substantial amount of non-labeled element. Since the number of radioactive atoms is extremely small, the problem with the latter is that the carrier may change the equilibrium of the system. However, the latter may be preferred in some cases for tracer studies.

Since radioactive isotopes decay over time, some approaches require that the measured activity be corrected back to the same reference or 'activation' date, i.e., activity at $t = 0$ (Blair and Till, 2003). This is particularly the case for radionuclides with short half-lives or when multiple measurements are carried out over a relatively long period of time (McLaren et al., 2017).

Radioactive isotopes to study soil pools and processes

Radioactive isotopes have been used as a tool to study nutrient pools and their transformation in soils for over 70 years (Frossard et al., 2011). They are an essential tool to directly assess the relative importance of abiotic and biotic pathways in soils. Here, several approaches are highlighted with some relevant examples to soil science applications. Since a large proportion of studies involving radioactive isotopes are carried out on phosphorus (P), here, P radionuclides (i.e., P-32 and P-33) are discussed as a model. However, many of the concepts can be applied to other radionuclides where applicable. A carrier-free radioactive isotope is the basis for most calculations mentioned below.

Isotopic dilution in soil-solution systems

The isotopic dilution principle involves the addition of a radioactive isotope to the soil system that will distribute uniformly between the soil solution and soil exchangeable pools (Hamon et al., 2008). The isotopic dilution principle is described in Eq. (4) (Hamon et al., 2002).

$$\frac{r_{\text{sol}}}{r_{\text{exch}}} = \frac{c_{\text{sol}}}{c_{\text{exch}}} \quad (4)$$

where r_{sol} refers to the activity of the radioactive isotope in the soil solution, r_{exch} refers to the activity of the radioactive isotope that is exchangeable, c_{sol} is the concentration of the element in the soil solution, and c_{exch} is the concentration of the element that is exchangeable.

Eq. (4) can be rearranged to calculate the concentration of the element that is exchangeable (Eq. 5).

$$c_{\text{exch}} = c_{\text{sol}} \times \frac{r_{\text{exch}}}{r_{\text{sol}}} \quad (5)$$

If a known quantity of a radioactive isotope is added to the soil system, it will distribute itself between the soil solution and soil exchangeable pools (Hamon et al., 2002). Consequently, the activity of the radioactive isotope that is exchangeable can be calculated as the difference between the activity of the radioactive isotope which was added to the soil system and that in the soil solution (Eq. 6).

$$r_{\text{exch}} = R_{\text{added}} - r_{\text{sol}} \quad (6)$$

where R_{added} refers to the activity of the radioactive isotope which was added to the soil system.

The concentration of the element that is exchangeable in soil can be determined when combining Eqs. (5) and (6) (Eq. 7) (Hamon et al., 2002). When converted to a soil weight basis (mg P kg^{-1}), the term 'E-value' is commonly used in the literature, which signifies "exchangeable". However, Hamon et al. (2002) suggests to use the term ' E_e ' to denote the concentration of the element that is exchangeable on the soil solid.

$$E_e = c_{\text{exch}} = c_{\text{sol}} \times \frac{(R_{\text{added}} - r_{\text{sol}})}{r_{\text{sol}}} \quad (7)$$

Eq. (7) can be rearranged so that the concentration of the element in the soil solution and exchangeable phase on the soil solid can be determined (Eq. 8). When converted to a soil weight basis (mg P kg^{-1}), this is considered the most bioavailable pool of the element in soil. Here, Hamon et al. (2002) suggests to use the term ' E_a ' to denote the concentration of the element in the soil solution and exchangeable phase on the soil solid.

$$E_a = (c_{\text{exch}} + c_{\text{sol}}) = c_{\text{sol}} \times \frac{R_{\text{added}}}{r_{\text{sol}}} \quad (8)$$

There are several assumptions that need to be met when applying the isotopic dilution principle in soil science contexts (Hamon et al., 2008; Frossard et al., 2011). These include: (1) the radioactive isotope that is added to the soil system should not alter the equilibrium conditions; (2) the radioactive isotope behaves in the same way as the stable isotope naturally present in the soil system; (3) the target element in the soil solution is isotopically exchangeable, and; (4) the added radioactive isotope is uniformly mixed with the non-labeled element being traced in the soil solution and exchangeable pool. While the isotopic dilution of a radioactive isotope within the soil system occurs in theory at infinite time, in practice an operationally defined time is routinely used (typically 1–3 days).

'Available' phosphorus in soil—Isotopically exchangeable phosphorus (E-value)

Numerous methods have been developed to estimate the pool of 'available' nutrients in soil. Historically, chemical extraction techniques have been the predominant method. The problem with this is that they are operationally defined and have the potential to cause artifacts, which often under- or over-estimate the pool of plant-available nutrients in soil. Measures of isotopically exchangeable P, or 'E-values', can provide an estimate of the 'available' or 'labile' pool of soil P based on the isotopic dilution principle (Frossard et al., 1994).

The determination of E-values typically involves the addition of a radioactive isotope to a soil suspension under equilibrium conditions. The soil to solution ratio, background solution, equilibration times, and periods of exchange time vary in the literature (Kuo, 1996; Di et al., 1997). However, a 1:10 or 1:20 soil to solution ratio, a background solution of H_2O or a dilute electrolyte (e.g., 0.01 M CaCl_2 , calcium chloride), and a 16 h equilibration time appears to be more common (Meyer et al., 2020). The activity of the radioactive isotope in the soil solution (i.e., r_{sol} in Eqs. 7 and 8) is highly dependent on the time of exchange. Therefore, this is typically designated as r_t and refers to the activity of the radioactive isotope in the soil solution after a period of isotopic exchange time t (Hamon et al., 2002).

The E-values are generally obtained on soil suspensions and are highly time dependent, therefore, Eqs. (7) and (8) need to be converted to a soil weight basis (mg P kg^{-1}) and reported at a given period of exchange time (Hamon et al., 2002). For example, Eq. (8) can be converted to take into account the dilution factor (i.e., the ratio of solution volume (L) to soil mass (kg)) and period of exchange time (Eq. 9) (Hamon et al., 2008).

$$E_t = D \times c_{\text{sol}} \times \left(\frac{R_{\text{added}}}{r_t} \right) \quad (9)$$

where D refers to the dilution factor of the soil suspension (i.e., volume of solution (L) divided by the soil mass (kg)), c_{sol} refers to the amount of the nutrient in the soil solution (mg L^{-1}), R_{added} refers to the total amount of activity of the radioactive isotope added to the soil suspension (Bq), and r_t refers to the activity of the radioactive isotope in the soil solution after a period of exchange time t (Bq).

It is not always clear if the reported E-values refer to an E_e or E_a value (Hamon et al., 2002). Although, the latter appears more prevalent and is primarily focused on pools of bioavailable P (Hamon et al., 2002). In general, differences between E_e and E_a values do not differ greatly (Hamon et al., 2008). In soils where r_t is small compared to R_{added} , then differences between E_e and E_a will be negligible. However, where r_t is large compared to R_{added} , then differences between E_e and E_a will be much greater. This is particularly the case in soils with a small P buffering capacity. Furthermore, differences between E_e and E_a will be greater with decreasing periods of exchange time as r_{sol} decreases with time.

Caution is needed when carrying out measures of isotopic exchange in soils with very small concentrations of soil solution P and large P sorption capacity (Kuo, 1996). This is because the concentration of P in the soil solution is often $<10 \mu\text{g P L}^{-1}$, which is difficult to measure accurately by colorimetric approaches (Randriamanantsoa et al., 2013). Nevertheless, various approaches have been developed to measure concentrations as small as $0.8 \mu\text{g P L}^{-1}$ (Randriamanantsoa et al., 2013).

Several studies report E-values in soils under a diverse array of applications (Di et al., 1997; Nguyen et al., 2016). For example, Meyer et al. (2020) determined pools of isotopically exchangeable P within 24 h ($E_{24\text{h}}$) on a range of soils to assess the pool of potentially plant-available P. The authors treated soils with different types and amounts of fertilizer in an incubation experiment. They found differences in $E_{24\text{h}}$ -values among soil types and also that among fertilizer combinations. Most notably, the addition of potassium (K) with P as monoammonium phosphate in a concentrated zone resulted in a decrease in $E_{24\text{h}}$ -values in some soil types compared to that in the absence of K.

Soil phosphorus dynamics—Phosphorus isotope exchange kinetics (IEK)

For over 70 years, the activity of the radioactive isotope (P-32 and P-33) in soil solution was known to decrease with time (Frossard et al., 2011). McAuliffe et al. (1948) investigated the reaction of P-32 in a soil-water suspension under equilibrium conditions. The authors describe two main types of phosphate reaction between the soil solution and soil surface with time. The first was interpreted as a pool of readily exchangeable P and the second as a pool of less exchangeable P. This was later described using a simple power function (Morel et al., 1989) (Eq. 10).

$$\frac{r(t)}{R_{\text{added}}} = \left[\frac{r(1)}{R_{\text{added}}} \right] t^{-n} \quad (10)$$

where t refers to the period of isotopic exchange in minutes, $\frac{r(1)}{R_{\text{added}}}$ is the proportion of activity in the soil solution after 1 min of isotopic exchange relative to that of the total activity added to the soil system, and n is a model parameter that describes the rate of removal of activity in solution.

The simple power equation describing the reaction of P-32 in the soil solution with the soil surface was found to be inadequate because of the long periods of time needed to reach a final isotopic equilibrium. Fardeau (1995) proposed that the exchange of radioactive phosphate could be characterized by its kinetic parameters obtained in an isotopic exchange kinetics (IEK) experiment. Briefly, this involves a soil suspension (10 g of soil in 99 mL of H_2O) under equilibrium conditions. At time $t = 0$, a 1 mL aliquot of a radioactive isotope (P-32 or P-33) solution is added, and the soil suspension sampled with a syringe and filter at increasing periods of isotopic exchange time (e.g., 1 min, 4 min, 10 min, 30 min, 60 min, and 90 min). Consequently, Fardeau et al. (1985) adapted the power function so that a final isotopic equilibrium was no longer necessary (Eq. 11). See Fig. 1 for an example of a P IEK experiment on a soil.

$$\frac{r(t)}{R_{\text{added}}} = m \left(t + m^{\frac{1}{n}} \right)^{-n} + \frac{r_{\infty}}{R_{\text{added}}} \quad (11)$$

where m and n are model parameters derived from the nonlinear regression, r_{∞} refers to the activity of the radioactive isotope in the soil solution at final isotopic equilibrium, $\frac{r_{\infty}}{R_{\text{added}}}$ is the proportion of activity of the radioactive isotope in the soil solution at the final isotopic equilibrium to that of the total activity added to the soil system.

The term $\frac{r_{\infty}}{R_{\text{added}}}$ describes the ratio of r_{sol} to R_{added} at the final isotopic equilibrium or the maximum dilution of the added radioactive isotope by inorganic P in the soil system (Di et al., 1997). It has been operationally defined as the ratio of P in the soil solution or c_{sol} (mg P kg^{-1}) and the pool of soil inorganic P (mg P kg^{-1}) (Eq. 12) (Frossard et al., 2011). The parameter m indicates the immediate physicochemical reactions that occur in soils, and corresponds to the proportion of activity in the soil solution compared to that of the total activity added after a period of isotopic exchange of 1 min (or $\frac{r(1)}{R_{\text{added}}}$) (Achat et al., 2016). The parameter n indicates the slow physicochemical reactions, and accounts for the rate of removal of radioactive P from the soil solution with isotopic exchange times longer than 1 min. In general, phosphate reactivity with the soil solid phase will increase when values of n increase or values of m decrease. Furthermore, the buffering capacity of P of the soil increases with decreasing values of m or increasing values of n .

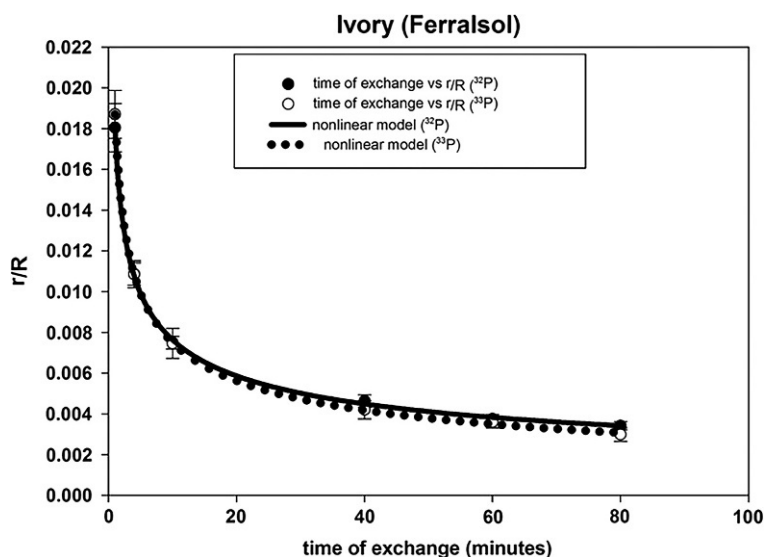


Fig. 1 An example of a P isotopic exchange kinetics experiment (IEK) using P-32 (and P-33) on a Ferralsol soil collected from the Highlands of Madagascar, as reported in [Randriamanantsoa et al. \(2013\)](#). The proportion of activity of P-32 (and P-33) in the soil solution (r) to that of the total activity added (R) decreased with increasing time of exchange (1 min, 4 min, 10 min, 40 min, 60 min, and 80 min). A nonlinear model was fitted to the data (see Eq. 11). Furthermore, the authors found that there was no systematic isotopic fractionation between P-32 and P-33 during the experiment, suggesting these radionuclides have a similar behavior when undergoing isotopic exchange. Reprinted from Randriamanantsoa L, Morel C, Rabeharisoa L, Douzet J, Jansa J, Frossard E (2013) Can the isotopic exchange kinetic method be used in soils with a very low water extractable phosphate content and a high sorbing capacity for phosphate ions? *Geoderma* **200–201**: 120–129, Copyright (2013), with permission from Elsevier.

$$\frac{r_{(\infty)}}{R_{\text{added}}} = \frac{D \times c_{\text{sol}}}{P_{\text{inorg}}} \quad (12)$$

where P_{inorg} refers to the concentration of inorganic P in soil.

The P IEK approach involves homo-ionic exchange of the added radioactive isotope (P-32 or P-33) in the soil solution with the pool of desorbable P on the soil solid phase ([Frossard et al., 2011](#)). It provides information on the factors quantity (Q), intensity (I), and capacity (C) of available P in soil ([Nguyen et al., 2016](#)). The factor Q is the amount of ‘available’ P in soil and is obtained from the calculation of E-values (Eq. 9), the factor I is the concentration of P in the soil solution and is measured directly (i.e., c_{sol}), and factor C is the ability of the soil to resist changes in soil solution P and is obtained from the rate at which the radioactive isotope is removed from the soil solution.

The P IEK approach reveals that inorganic P in soil should be considered a continuum of P from immediately available to slowly available, rather than a distinct number of pools designated as available or non-available ([Nguyen et al., 2016](#)). The model proposed by [Fardeau et al. \(1985\)](#) based on a short-term kinetic (up to 100 min) can derive long-term kinetics (3 months or more). A pluricompartimental model was proposed to describe the quantity of exchangeable P based on plant P uptake ([Fardeau, 1995](#)). This included: (1) the pool of exchangeable P within 1 min, which comprises P in the soil solution and that immediately transferable to the soil solution; (2) the pool of exchangeable P within 1 min and 24 h, which is based on the average time of active plant P uptake by a root part; (3) the pool of exchangeable P within 1 day and 3 months, which is based on the average time of active plant P uptake by a root system, and; (4) the pool of exchangeable P that is greater than 3 months.

The model also enables the calculation of several measures of P dynamics ([Fardeau, 1995](#)). This includes: (1) the average turnover (or exchange) rate between inorganic P on the soil solid phase and in the soil solution (K_m in min^{-1} , also referred to as g_m) (Eq. 13); (2) the average residence time of inorganic P in the soil solution (T_m in min) (Eq. 14), and; (3) the average flux of exchange between inorganic P on the soil solid phase and in the soil solution (F_m in $\text{mg P kg}^{-1} \text{ soil min}^{-1}$) (Eq. 15).

$$K_m = \frac{n}{m^{(1/n)}} \quad (13)$$

where K_m refers to the average turnover (or exchange) rate between inorganic P on the soil solid phase and in the soil solution, and the parameters m and n are obtained from the model.

$$T_m = \frac{1}{K_m} \quad (14)$$

where T_m refers to the average residence time of inorganic P in the soil solution.

$$F_m = D \times c_{\text{sol}} \times K_m \quad (15)$$

where F_m refers to the average flux of exchange between inorganic P on the soil solid phase and in the soil solution.

The P IEK approach provides important information on the amount and transfer of P in soil systems, which are related to various soil properties (Achat et al., 2016) and plant P uptake (Fardeau, 1995). Furthermore, the P IEK approach has provided detailed information on the P status and dynamics of a diverse range of soils, which could predict crop responses, and the likelihood of crop responses to P fertilization (Frossard et al., 2000).

Gross phosphorus mineralization

The turnover of soil organic matter is an important source and sink of available P in soil (Frossard et al., 2000). Quantifying the amount of mineralizable organic P in soil can be difficult (Di et al., 1997), but estimates are possible based on the isotopic dilution principle (Bünemann, 2015). Essentially, the pool of isotopically exchangeable P in soil is labeled with a radioactive isotope (P-32 or P-33). The P in soil organic matter will not be labeled with the radioactive isotope, therefore, any mineralization of organic P will increase the concentration of P in the pool of isotopically exchangeable P and dilute the radioactive isotope (P-32 or P-33). Pools of exchangeable P (E-values) will be greater in the biologically active soil where abiotic and biotic processes occur, relative to that in a sterile soil where only abiotic (physicochemical reactions) processes occur. The difference in E-values between the two soils is considered the amount of inorganic P in soil that has come from the mineralization of organic P within a given period of time (Frossard et al., 2011).

The isotopically exchangeable pool of P (E-values) due to physicochemical processes is typically obtained by a P IEK experiment on the preincubated and unlabeled soil, which is extrapolated and forms the baseline of physicochemical processes (Bünemann, 2015). The isotopically exchangeable pool of P (E-values) due to physicochemical and biological or biochemical processes is typically obtained using a soil incubation experiment that has been labeled with a radioactive isotope (P-32 or P-33) (Oehl et al., 2004). After a given period of time, water extractions are carried out and the specific activity determined to calculate E-values (Bünemann, 2015). The outcome is the difference in relative specific activity and E-values between the baseline and biologically active soil (Fig. 1). Phosphorus immobilization can also be estimated by measuring the soil microbial biomass in the incubation experiment (Bünemann, 2015) (Fig. 2).

The above approach of estimating gross mineralization of organic P in soil has been simplified and there are many other factors to consider (Bünemann, 2015). Nevertheless, it is important that the experiment be carried out at steady-state equilibrium, a low and constant rate of respiration, a constant concentration of soil solution P, and a relatively short incubation period (7–14 days) to minimize any remineralization of organic P that had been labeled subsequently (Frossard et al., 2011; Nguyen et al., 2016).

Studies reporting the gross mineralization of organic P in soil based on the isotopic dilution principle are limited (Bünemann, 2015). Oehl et al. (2004) investigated the rates of mineralization of organic P in an incubation experiment with soils from a long-term field experiment in Switzerland. The authors reported that the rates of basal mineralization of soil organic P ranged from 1.4 to 2.5 mg P kg⁻¹ d⁻¹, which was less than the pool of isotopically exchangeable P due to physicochemical processes. This highlights the importance of physicochemical processes in supplying plant-available P at this site. Furthermore, rates of mineralization of soil organic P were larger for the biodynamic treatment than for the mineral fertilizer treatment.

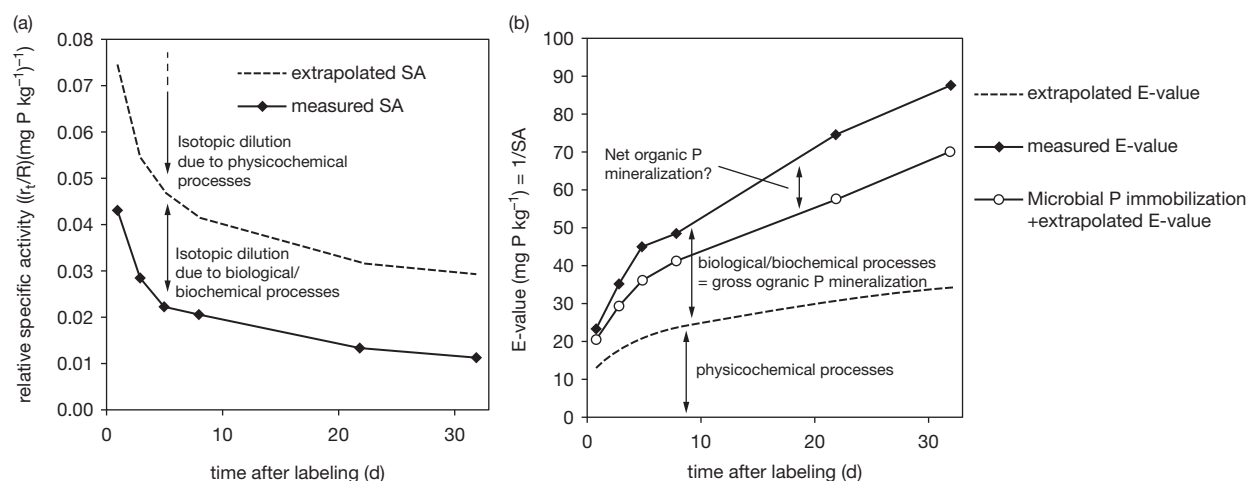


Fig. 2 The isotopic dilution principle used to estimate the gross mineralization rates of organic P in soil, as reported in Bünemann (2015). (a) The relative specific activity (r_t/R) in soil water filtrates that have been obtained using P IEK experiments and extrapolated (physicochemical processes), and also from a soil incubation experiment labeled with P-32 or P-33 (plus biological or biochemical processes). (b) Measures of E-values were obtained as the inverse of the relative specific activity. Reprinted from Bünemann EK, Assessment of gross and net mineralization rates of soil organic phosphorus – A review, Soil Biology and Biochemistry 89: 82–98, Copyright (2015), with permission from Elsevier.

Isotopic dilution in soil-plant systems

'Available' phosphorus in soil—isotopically exchangeable phosphorus (L-value)

Measures of isotopically exchangeable P can also be obtained from soil-plant systems based on the isotopic dilution principle. The term 'L-value' refers to "Larsen" values of soil P that are associated with pools of labile P in the soil (Frossard et al., 2011). It estimates the quantity of isotopically exchangeable P in soil based on the specific activity of P in the plant (Larsen, 1967).

The determination of L-values typically involves the addition of a radioactive isotope to a soil system, which is thoroughly mixed and allowed to reach equilibrium between the added radioactive isotope and the pool of exchangeable P in soil (Nguyen et al., 2016). Nevertheless, equilibrium may not be obtained in practice particularly in soils with recent additions of fertilizer P (Di et al., 1997). Plants are then grown in the radiolabeled soil and the specific activity of shoots measured (Frossard et al., 2011). The L-value can be calculated using Eq. (16).

$$L = (c_{\text{shoot}} - c_{\text{seed}}) / \left(\frac{r_{\text{shoot}}}{R_{\text{added}}} \right) \quad (16)$$

where c_{shoot} refers to the amount of total P in the shoot, c_{seed} is the amount of P from the seed that is contributing to the total P in the shoot, and r_{shoot} refers to the activity of the radioactive isotopes in the shoot.

The L- and E-values should be similar based on a conceptual basis, but they can differ based on several factors (Di et al., 1997). This is partly because measuring L-values involve plants, which have various P acquisition strategies to mobilize soil P. In contrast, measuring E-values involves a soil suspension in the absence of a plant. There are many soil-based processes that occur in the rhizosphere, which may change the specific activity. The measurement of L-values is generally carried out experimentally over a much longer time period than that of E-values. Last, the contribution of seed P to that of shoot P can be difficult to quantify, particularly at early growth stages (Frossard et al., 2011).

Many studies have shown positive relationships between E- and L-values for the majority of crops at similar periods of exchange time (Frossard et al., 2011). For example, Frossard et al. (1994) compared E-values within a 13 week period and that of L-values in common bentgrass (*Agrostis capillaris* L.) across 10 diverse surface soils grown in pots. The authors found that E- and L-values were strongly correlated (see Fig. 3), which highlights the importance of the isotopically exchangeable pool of P for plant P uptake.

Radioactive isotopes to study the fate of fertilizer

There are several approaches to assess the fertilizer-use efficiency of P by crops, and each has its advantages and disadvantages (Johnston et al., 2014). Radioactive isotopes have been an important approach for tracking the fate of fertilizer P in soil-plant systems for over 70 years (Frossard et al., 2011). Furthermore, radioactive isotopes can also be used to assess the fate of fertilizer P into different soil fractions and the extent of their movement over time (McLaren et al., 2017).

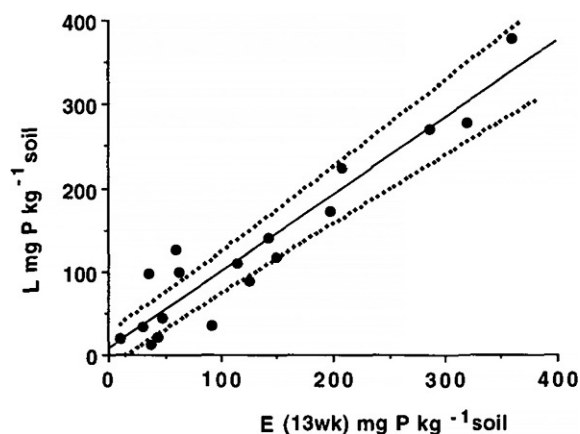


Fig. 3 Measures of L-values obtained on common bentgrass (*Agrostis capillaris* L.) across ten diverse soils, including those that had received phosphate fertilizer after 13 weeks of growth, and that of E-values obtained on the soils based on P isotopic exchange kinetics and extrapolated to 13 weeks of exchange, as reported in Frossard et al. (1994). There was a strong correlation between the reported L- and E-values. Reprinted from the Soil Science Society of America Journal, 58, Frossard E, Morel JL, Fardeau JC, Brossard M, Soil Isotopically Exchangeable Phosphorus: A Comparison between E and L Values, 846–851, Copyright (1994), with permission from John Wiley and Sons.

Direct labeling approach

The direct labeling approach involves labeling the source with a radioactive isotope and provides direct information on the fate of fertilizer P in soil-plant systems (McLaughlin et al., 2011). The approach has been widely used in a variety of agricultural contexts and is based on the homogenous labeling of a fertilizer source (solution or powder) with a radioactive isotope (Frossard et al., 2011). This can involve dissolving a fertilizer (e.g., monocalcium phosphate) and then adding the radioactive isotope (P-32 or P-33) to the solution, which can then be added to the soil in liquid form. Alternatively, the labeled fertilizer can be precipitated out and a dry powder is formed, or radiolabeled granules can be made (McLaren et al., 2017). In addition, radiolabeled plant residues can be made by growing plants in a soil or solution containing the radioactive isotope (P-32 or P-33), and the shoots subsequently harvested and dried for later experiments and analyzes (McLaughlin et al., 1988).

The specific activity of the sample analyzed (e.g., plant shoots or a soil extract) and that of the added fertilizer (e.g., MBq of P-33 mg^{-1} of P) is then used to calculate the amount of P in the sample derived from the fertilizer (Eq. 17) (Frossard et al., 2011). See Fig. 4 for an example of the percentage of plant P derived from the labeled fertilizer ($Pdff$ (%)).

$$Pdff (\%) = 100 \left(\frac{SA_{+P}}{SA_f} \right) \quad (17)$$

where $Pdff$ (%) is the percentage P in the plant derived from the labeled fertilizer, SA_{+P} refers to the specific activity of P in the plant from fertilized treatments, and SA_f refers to the specific activity of the labeled fertilizer.

The ability to label the source with a radioactive isotope is a powerful technique to assess its fate in the soil-plant system. The fate of a variety of fertilizer sources, organic compounds, and plant residues can be assessed. However, there are several limitations to the direct labeling approach. These can include: (1) difficulty to label the material homogeneously; (2) radioactive isotopes with short half-lives favor laboratory- and glasshouse-based experiments rather than field-based experiments, and; (3) tracking the fate of fertilizer P over several growing seasons or assessing the contribution of P in plants from residual pools is either difficult or not possible.

Many studies have investigated the fate of fertilizer P in soil-plant systems using the direct labeling approach, and several have been carried out under glasshouse conditions over relatively short periods of time. Nevertheless, some studies have applied the direct approach under field conditions (McLaughlin et al., 1988). For example, McLaren et al. (2017) investigated the fate of P-33 labeled single superphosphate granules in subterranean clover (*T. subterraneum* L.) pastures grown under field conditions. The authors found that about 50% of the added fertilizer P was recovered in clover shoots, whereas about 20–25% of the added fertilizer P was recovered in the 0–4 cm soil layer and predominantly as inorganic P (Fig. 5). A further 5–15% of the added fertilizer P remained in the granule at the end of the growing season. The authors also reported that early-season applications of fertilizer P to the soil surface was generally the best strategy to improve pasture recovery of fertilizer P.

Indirect labeling approach

There are many contexts where it is not possible to label the source homogeneously (Frossard et al., 2011). These can include animal manures, composts, natural minerals (e.g., rock phosphate), and waste materials. The indirect labeling approach provides an

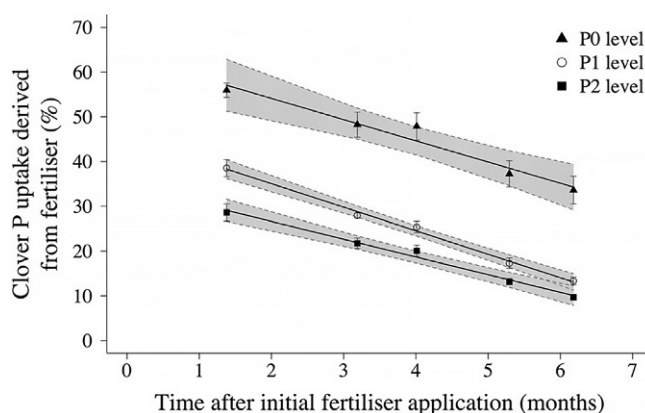


Fig. 4 The percentage of P in subterranean clover (*Trifolium subterraneum* L.) shoots derived from the labeled fertilizer ($Pdff$ (%)) across three concentrations of initial soil P fertility (P0: 4–6 mg Olsen–P kg^{-1} , P1: 10–15 mg Olsen–P kg^{-1} , and P2: 20–25 mg Olsen–P kg^{-1}) on an acidic Luvisol located near Hall, Australia, as reported in McLaren et al. (2017). The direct labeling approach was used in the study. The authors found that the percentage of shoot P derived from the labeled fertilizer was highest in the P0 treatment compared to that in the P1 and P2 treatments, and decreased with time. Reprinted from McLaren TI, McBeath TM, Simpson RJ, Richardson AE, Stefanski A, Guppy CN, Smernik RJ, Rivers C, Johnston C, McLaughlin MJ (2017) Direct recovery of ^{33}P -labelled fertiliser phosphorus in subterranean clover (*Trifolium subterraneum*) pastures under field conditions – The role of agronomic management. *Agriculture, Ecosystems & Environment* **246**: 144–156, Copyright (2017), with permission from Elsevier.

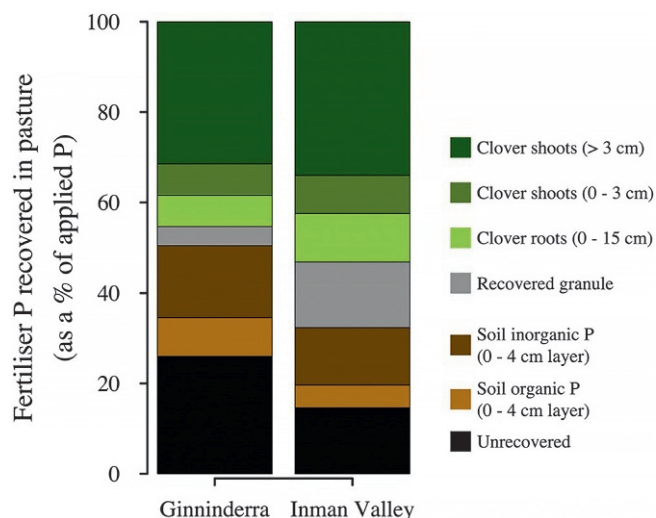


Fig. 5 Recovery of fertilizer P in various components of a subterranean clover (*Trifolium subterraneum* L.) pasture at two sites in the high rainfall zone of south-eastern Australia based on the direct labeling approach (P-33 labeled single superphosphate), as reported in McLaren et al. (2017). Reprinted from McLaren TI, McBeath TM, Simpson RJ, Richardson AE, Stefanski A, Guppy CN, Smernik RJ, Rivers C, Johnston C, McLaughlin MJ (2017) Direct recovery of ^{33}P -labelled fertiliser phosphorus in subterranean clover (*Trifolium subterraneum*) pastures under field conditions – The role of agronomic management. *Agriculture, Ecosystems & Environment* **246**: 144–156, Copyright (2017), with permission from Elsevier.

opportunity for the fate of a complex source to be tracked in the system. This involves labeling the isotopically exchangeable pool of soil P with a radioactive isotope, including the fertilized and non-fertilized treatments. The specific activity of samples (e.g., plant shoots or a soil extract) in the fertilized and non-fertilized treatments is determined, which can then be used to calculate the amount of P in the sample derived from the fertilizer (Eq. 18) (Frossard et al., 2011).

$$Pdff (\%) = 100 \left(1 - \frac{SA_{+P}}{SA_{-P}} \right) \quad (18)$$

where SA_{+P} refers to the specific activity of P in the plant from the fertilized P treatments, and SA_{-P} refers to the specific activity in the plant from the non-fertilized P treatments.

An important assumption of the indirect approach is that the specific activity of the isotopically exchangeable pool of P is similar in the fertilized and non-fertilized soils (Frossard et al., 2011). Any release of inorganic P from the added non-labeled fertilizer will dilute the radioactive isotope (P-32 or P-33). However, a possible issue with this approach is that the addition or presence of some organic material can also affect the specific activity by causing immobilization of P or by stimulating P mineralization of the native organic P present in the soil, which was not labeled. The consequence is that the specific activity of the soil solution of the fertilized treatment will be affected and provide unreliable results on the percentage of plant P derived from the organic fertilizer (called 'pool substitution') (Frossard et al., 2011).

Limitations of radiotracers

Radioactive isotopes (P-32 and P-33) are an important tool to probe nutrient pools and their transformation in soil, and as tracers to assess the fate of fertilizers in soil-plant systems. However, there are several limitations that need to be considered. The first is that radioactive isotopes emit ionizing radiation, which makes their use more restrictive due to various safety and government regulations. The second is that each radioactive isotope has a half-life, which tends to favor short-term and laboratory-based experiments compared to that of long-term and field-based experiments. In general, long half-lives are more favorable particularly for glasshouse or field experiments. However, long half-lives can also be problematic due to issues with their safe disposal. Another issue associated with this is that the activity of radioactive isotopes is decreasing which requires that they are used and analyzed in a timely manner. Consequently, if there is surplus activity to requirements, radioactive isotopes with short half-lives cannot be stored and used at a later date.

Future opportunities for radiotracers

There are several radioactive (and stable) isotopes available to study soil pools and processes. Many studies have been carried out using a single isotope. However, it will be important to combine multiple isotopes and analytical approaches to obtain a holistic

view of nutrient cycling. For example, combining C-14, N-15, and P-33 isotopes to better understand soil organic matter turnover. Furthermore, future research tracking the fate of fertilizers into specific nutrient pools will provide improved insight into the processes that are important in their accumulation or depletion in soil. For example, assessing the relative contribution of fertilizer P into specific forms of organic P (e.g., *myo*-inositol hexakisphosphate) might help develop strategies to reduce their stabilization and improve crop and pasture recovery of the added fertilizer P. In contrast, where the accumulation of soil organic matter is limited by P, radioactive isotopes may help to identify how the P component of organic matter in soil accumulates.

Conclusions

The use of radioactive isotopes has greatly improved our understanding of soil pools and process. The isotopic dilution principle is an important approach when applying radioactive isotopes in soil science contexts. Pools of isotopically exchangeable P are highly relevant for plant availability, and the dynamics of P between the soil solution and soil solid phases are important at assessing P availability and transfer. Furthermore, tracer approaches provide direct information on the fate of fertilizer P in soil-plant systems, and can provide quantitative measures of the crop recovery of fertilizer P and how this is affected by differing agronomic management strategies. Radioactive isotopes will continue to be an important tool to help understand nutrient cycles, which will also help to address major challenges of global concern (e.g., food security and climate change).

References

- Achat DL, Pousse N, Nicolas M, Brédoire F, and Augusto L (2016) Soil properties controlling inorganic phosphorus availability: General results from a national forest network and a global compilation of the literature. *Biogeochemistry* 127: 255–272.
- Blair GJ and Till R (2003) *Guidelines for the Use of Isotopes of Sulfur in Soil-Plant Studies*. Vienna, Austria: Soil and Water Management & Crop Nutrition Section International Atomic Energy Agency.
- Bünemann EK (2015) Assessment of gross and net mineralization rates of soil organic phosphorus—A review. *Soil Biology and Biochemistry* 89: 82–98.
- Di HJ, Condron LM, and Frossard E (1997) Isotope techniques to study phosphorus cycling in agricultural and forest soils: A review. *Biology and Fertility of Soils* 24: 1–12.
- FAO/IAEA (2016) *Use of Phosphorus Isotopes for Improving Phosphorus Management in Agricultural Systems*. Vienna, Austria: Joint FAO/IAEA Division of Nuclear Techniques in Food and Agriculture.
- Fardeau JC (1995) Dynamics of phosphate in soils. An isotopic outlook. *Fertilizer Research* 45: 91–100.
- Fardeau JC, Morel C, and Jappe J (1985) Exchange kinetics of phosphate ions in the soil-solution system. Experimental verification of theoretical equation. *Comptes rendus de l'Académie des Sciences III* 300: 371–376.
- Frossard E, Morel JL, Fardeau JC, and Brossard M (1994) Soil isotopically exchangeable phosphorus: A comparison between E and L values. *Soil Science Society of America Journal* 58: 846–851.
- Frossard E, Condron LM, Oberson A, Sinaj S, and Fardeau JC (2000) Processes governing phosphorus availability in temperate soils. *Journal of Environmental Quality* 29: 15–23.
- Frossard E, Achat DL, Bernasconi SM, Bünemann EK, Fardeau J, Jansa J, Morel C, Rabeharisoa L, Randriamanantsoa L, Sinaj S, Tamburini F, and Oberson A (2011) The use of tracers to investigate phosphate cycling in soil-plant systems. In: Bünemann E, Oberson A, and Frossard E (eds.) *Phosphorus in Action: Biological Processes in Soil Phosphorus Cycling*, pp. 59–91. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Hamon RE, Bertrand I, and McLaughlin MJ (2002) Use and abuse of isotopic exchange data in soil chemistry. *Soil Research* 40: 1371–1381.
- Hamon RE, Parker DR, and Lombi E (2008) Chapter 6: Advances in isotopic dilution techniques in trace element research: A review of methodologies, benefits, and limitations. *Advances in Agronomy*. vol. 99, pp. 289–343. Academic Press.
- IAEA (2022) *Live Chart of Nuclides*. Available at <https://www-nds.iaea.org/relnsd/vcharthtml/VChartHTML.html>.
- Johnston AE, Poulton PR, Fixen PE, and Curtin D (2014) Chapter 5: Phosphorus: Its efficient use in agriculture. In: Sparks DL (ed.) *Advances in Agronomy*. vol. 123, pp. 177–228. Academic Press.
- Kuo S (1996) Phosphorus. In: *Methods of Soil Analysis*, pp. 869–919. Soil Science Society of America. <https://access.onlinelibrary.wiley.com/doi/abs/10.2136/sssabookser5.3.c32>.
- Larsen S (1967) Soil phosphorus. In: Norman AG (ed.) *Advances in Agronomy*. vol. 19, pp. 151–210. Academic Press.
- McAuliffe CD, Hall NS, Dean LA, and Hendricks SB (1948) Exchange reactions between phosphates and soils: Hydroxylic surfaces of soil minerals. *Soil Science Society of America Journal* 12: 119–123.
- McLaren TI, McBeath TM, Simpson RJ, Richardson AE, Stefanski A, Guppy CN, Smernik RJ, Rivers C, Johnston C, and McLaughlin MJ (2017) Direct recovery of 33P-labelled fertiliser phosphorus in subterranean clover (*Trifolium subterraneum*) pastures under field conditions—The role of agronomic management. *Agriculture, Ecosystems & Environment* 246: 144–156.
- McLaughlin MJ, Alston AM, and Martin JK (1988) Phosphorus cycling in wheat pasture rotations. I. The source of phosphorus taken up by wheat. *Soil Research* 26: 323–331.
- McLaughlin MJ, McBeath TM, Smernik R, Stacey SP, Ajiloye B, and Guppy C (2011) The chemical nature of P accumulation in agricultural soils—Implications for fertiliser management and design: An Australian perspective. *Plant and Soil* 349: 69–87.
- Meyer G, Bell MJ, Doolette CL, Brunetti G, Zhang Y, Lombi E, and Kopittke PM (2020) Plant-available phosphorus in highly concentrated fertilizer bands: Effects of soil type, phosphorus form, and coapplied potassium. *Journal of Agricultural and Food Chemistry* 68: 7571–7580.
- Morel JL, Fardeau JC, Béruff MA, and Guckert A (1989) Phosphate fixing capacity of soils: A survey, using the isotopic exchange technique, or soils from north-eastern France. *Fertilizer Research* 19: 103–111.
- Nguyen L, Zapata F, and Adu-Gyamfi J (2016) The use of phosphorus radioisotopes to investigate soil phosphorus dynamics and cycling in soil-plant systems. In: *Soil Phosphorus*, pp. 285–312. CRC Press.
- Oehl F, Frossard E, Fliessbach A, Dubois D, and Oberson A (2004) Basal organic phosphorus mineralization in soils under different farming systems. *Soil Biology and Biochemistry* 36: 667–675.
- Randriamanantsoa L, Morel C, Rabeharisoa L, Douzet J, Jansa J, and Frossard E (2013) Can the isotopic exchange kinetic method be used in soils with a very low water extractable phosphate content and a high sorbing capacity for phosphate ions? *Geoderma* 200–201: 120–129.

The use of stable isotopes in soil science: Metals

Cécile Quantin and Damien Guinoiseau, Université Paris Saclay, UMR GEOPS, Orsay, France

© 2023 Elsevier Ltd. All rights reserved.

Introduction	107
Non-traditional isotope geochemistry	108
Formalism in isotope geochemistry	108
Equilibrium vs kinetic isotope fractionation	108
Isotopes and pedogenetic processes	109
Weathering and formation of secondary clay minerals	109
Influence of organic matter complexation in the soil profile	110
Redox influence on pedogenesis	111
Control of pedological processes on trace metal behavior	111
Tracing the source of metal pollution in soils	113
Conclusions and reflections	114
References	114

Abstract

Metal isotopes can be used to improve the understanding of the dynamics of chemical elements in the soil–plant–water–atmosphere system. Some bio-physicochemical processes induce metal isotope fractionation, which is measurable by mass spectrometry. Therefore, isotope analysis of metals can contribute to the understanding of pedogenetic processes, metal source, i.e. natural or anthropogenic, as well as the processes affecting them if they induce isotope fractionation.

Key points

- Metal isotopes can improve understanding of the dynamics of chemical elements in the soil–plant–water–atmosphere system
- Metal isotopes can contribute to the understanding of pedogenetic processes
- Metal source can be determined by isotope analysis
- Fractionation processes may help to understand weathering processes in soil

Introduction

The Critical Zone extends from the top of the lithosphere to the base of the atmosphere. It is our living environment and supports human activities. The Critical Zone and in particular the ‘soil’ compartment are places of exchanges, interactions and transformation of matter and energy, that further control water and soil quality (Chorover et al., 2007). Many bio-physicochemical interactions that control the biogeochemical cycles of elements, including metals and metalloids, occur in soil. These metals, of which some are essential nutrients, can have a natural or anthropogenic origin. Furthermore, bio-physicochemical processes (e.g. redox changes, sorption, complexation, biological uptake) can affect metal aqueous or solid speciation.

Isotopes, stable or radioactive, can be used to improve understanding of the dynamics of chemical elements in the soil–plant–water–atmosphere system. Isotopes of an element have the same number of protons in their nucleus, but have a different number of neutrons, which gives them different masses. Stable isotope analysis of the ‘light’ elements (or traditional stable isotopes) C (carbon), N (nitrogen), O (oxygen), S (sulfur), and H (hydrogen) has been used for several decades to understand their cycle in the environment and in plants. Some bio-physicochemical processes induce isotopic discrimination and, although weak, this discrimination (or fractionation) is measurable by mass spectrometry. The isotope analysis of metals can also contribute to an understanding of their source, i.e. natural or anthropogenic, as well as the processes affecting them if they induce isotope fractionation. Recent developments in mass spectrometry, in particular the improvement of multi-collector inductively-coupled-plasma mass-spectrometers (MC-ICP-MS) and their installation in many laboratories, now make possible to analyze the stable isotopes of almost the whole periodic table with enough analytical precision (Wiederhold, 2015; Komárek et al., 2021). New perspectives are thus opened to understand the biogeochemical cycles of metals in soils and more broadly in the environment.

Non-traditional isotope geochemistry

Formalism in isotope geochemistry

For elements having more than one isotope, bio-physicochemical processes can result in isotope fractionation, i.e. a subtle modification of isotope abundances between the reactant and the product of the reaction. In this case, enrichment in one isotope in a reservoir will be compensated by its equivalent depletion in the other reservoir, and the isotope balance of the global system remains unchanged. This chapter describes how isotope fractionation can occur during mineral dissolution or precipitation, during mineral adsorption or organic complexation if the reaction is incomplete or follows a unidirectional pathway (see details in [Wiederhold, 2015](#)). The development of non-traditional isotope systems for 20 years (in contrast with C, H, O, N and S isotopes) has been mostly due to the advent of MC-ICP-MS, enabling fast and accurate measurements for numerous isotope systems. High-precision isotope values were also obtained through the strengthened development of efficient and size-reduced chromatographic separations with anionic or cationic exchange resins (e.g. [Conway et al., 2013](#) for Fe (iron), Cu (copper) and Zn (zinc) isotopes). A minimization of procedural blanks, isobaric interferences and/or matrix effects enable scientists to determine isotope ratios at nano- or sub-nanogram scale ([Albarède and Beard, 2004](#)). In addition, methods of post-analysis correction (e.g. double-spike or doping techniques) markedly increased measurement precision and reproducibility, allowing identification of small but resolvable isotope variability and fractionation.

With the ability of mass spectrometry to measure isotope ratios accurately rather than isotope abundances, isotope signatures are reported following a delta notation:

$$\delta^i X = \left[\frac{(^iX)_{\text{sample}}}{(^iX)_{\text{standard}}} - 1 \right] 1000 \quad (1)$$

where i and j are the heavier and the lighter isotopes of the element X , respectively. It must be borne in mind that variation of the isotopic signature relative to the standard, when it exists, is very low, and the reason why the delta value is expressed in ‰.

The isotope fractionation $\alpha_{A/B}$ between two compounds A and B is defined as the ratio of their respective isotope ratio R_A and R_B :

$$\alpha_{A/B} = \frac{^{i/j}R_A}{^{i/j}R_B} = \frac{\delta^i X_A + 1000}{\delta^i X_B + 1000} \quad (2)$$

Since $\alpha_{A/B}$ is close to unity, a convenient approximation can be inferred:

$$\Delta_{A-B} = \delta^i X_A - \delta^i X_B \approx 1000 \ln \alpha_{A/B} \quad (3)$$

Equilibrium vs kinetic isotope fractionation

Stable isotope fractionation in natural samples can be of either a kinetic or equilibrium nature. The only exceptions are mass-independent fractionation and nuclear volume effects that will affect odd and even isotopes differently for some elements (mostly iron and mercury (Hg), ([Wiederhold, 2015](#)) in specific contexts and that are not within the scope of this study.

Kinetic isotope fractionation occurs when isotopes of an element are affected by different rates of reaction, typically because of their mass differences. In other words, lighter isotopes react faster than heavier ones, resulting in a light isotope enrichment in the reaction product. This regime is typically seen for diffusion or evaporation/condensation processes, biological reactions that are enzymatically mediated or during the initiation time of adsorption, precipitation, or complexation reactions. The evolution of a system where kinetic isotope effects are dominant is classically modelled using a Rayleigh fractionation law, describing the instantaneous isotope signature of products which are physically (e.g., evaporation) or chemically (e.g., mineral precipitation) isolated from reactants.

Equilibrium isotope fractionation occurs when forward and backward transformations proceed at similar rates. In this system, enrichment in heavy or light isotopes in the product will be related to the free energy of isotope exchange reactions, which depends on the strength of the molecular bonds ([Bigeleisen and Mayer, 1947](#)). In other words, phases or products which form stiffer bonds will be enriched in heavy isotopes ([Schauble et al., 2009](#)). In the case of adsorption or complexation reactions, a larger bond stiffness is directly related to a lower coordination number and/or to shorter bond lengths between the element of interest and its neighbors. For abiotic or biotically-mediated redox processes, the oxidized species is generally enriched in heavy isotopes. The advantage of equilibrium fractionation is that it can be easily determined through experimental studies in the laboratory (details in [Komárek et al., 2021](#)) or calculated using numerical ab initio models (see [Fujii et al., 2014](#) in [Moynier et al., 2017](#)). In comparison, the magnitude of kinetic isotope fractionation is often difficult to approximate due to the progressive shift observed for most reactions from a kinetically-driven to an equilibrium-driven regime, which is particularly true for metals and metalloids. However, recent studies performed on soils, and more generally in the critical zone report that recorded silicon, Si (and presumably iron) isotope signatures conserve a mixture of both kinetic and equilibrium fractionation; the isotope expression of these two regimes of fractionation depending on the rates of formation of the mineral (see [Oelze et al., 2015](#) in [Poirasson, 2017](#)).

Isotopes and pedogenetic processes

Weathering and formation of secondary clay minerals

In soils, weathering processes control the formation of secondary minerals, such as clay, oxide, or hydroxide minerals. These major pedogenetic processes release the most soluble elements in soil solution such as alkalis (Na (sodium), K (potassium)), alkaline-earths (Ca (calcium), Mg (magnesium)) or silicon for the major ones and affect the elemental dynamics in both soil and soil solution. Clay minerals, i.e. phyllosilicates, precipitated from the weathering of silicate primary minerals, result in the association of an SiO_4 tetrahedral sheet with octahedral layers of either Al (aluminum), Fe or Mg. They are responsible for many of the soil's most important physical and chemical properties, such as cation exchange and shrink-swell properties. To date, Li (lithium), Si and Mg isotope systems (reported as $\delta^7\text{Li}$, $\delta^{30}\text{Si}$ and $\delta^{26}\text{Mg}$, respectively), are the most efficient to track the weathering conditions leading to the formation of clay minerals since they all fractionate during clay formation.

Despite being a trace metal, Li presents three main advantages for use as a tracer of the weathering process in soils (details in Penniston-Dorland et al., 2017). First, with its low atomic mass, considerable isotope fractionation is observed during water-rock interactions (from 38 to 45‰). Second, Li is not involved in plant chemistry or primary productivity cycling, and third, its small content in carbonates compared with silicates, means that variations in the Li isotope are primarily controlled by silicate weathering. During clay formation in soils, light Li isotopes are preferentially incorporated, leaving a soil solution enriched in heavy isotopes (Penniston-Dorland et al., 2017). The extent of the isotope fractionation is mineral-dependent (amorphous allophane fractionates more Li isotopes than clay minerals) and is thus associated with the weathering regime. Interestingly, recent work in Puerto Rico, where the highest rates of silicate weathering have been documented, demonstrates that Li isotopes provide evidence of the decoupling that exists between the soil solution and stream waters drained out of the system. The former traces the secondary mineral dissolution occurring in the higher part of the profile while stream water coincides with the transformation of bedrock to secondary minerals at the bedrock-saprolite interface (Chapela Lara et al., 2022). However, as a trace and soluble element, Li isotope application can be difficult in soils developed under extreme weathering conditions (e.g. tropical soils) or on volcanic islands where high Li depletion is observed and because eolian additions (either from marine aerosols or dust) represent a dominant Li input, making the understanding of Li cycling more complex at the scale of a weathering profile.

In comparison with Li, Si, the main constituent of clays with oxygen, can be more suitable. Its higher mass and one single valence explain the smaller range of terrestrial Si isotopic variation observed in soils (Fig. 1). As for Li, light Si isotopes are preferentially incorporated into secondary clay minerals and also during adsorption onto iron or aluminum oxides while the soil solution and stream waters carry heavier $\delta^{30}\text{Si}$ (details in Poitrasson, 2017). In bulk soils, $\delta^{30}\text{Si}$ decreases linearly with increasing degree of weathering, expressed as the chemical index of alteration (CIA) in Fig. 1:

$$\text{CIA} = (\text{Al}_2\text{O}_3) / (\text{Al}_2\text{O}_3 + \text{CaO} + \text{Na}_2\text{O} + \text{K}_2\text{O}) * 100$$

The CIA, which shows the importance of the leaching of base cations (Ca^{2+} , Na^+ , K^+) during weathering, is compatible with the formation of cation-depleted clay (e.g., kaolinite) with lower $\delta^{30}\text{Si}$ rather than swelling clay (e.g., smectite) at higher CIA, i.e., when

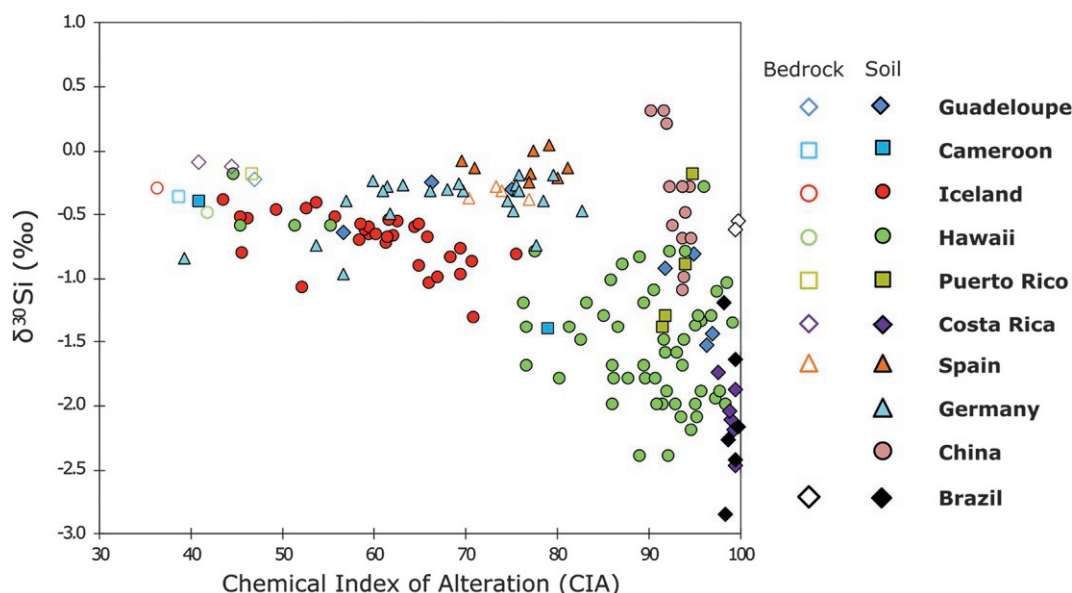


Fig. 1 Evolution of composition of the soil Si isotope singature with the Chemical Index of Alteration (CIA). The figure is adapted with recent data from the original diagram of Opfergelt S and Delmelle P (2012) Silicon isotopes and continental weathering processes: Assessing controls on Si transfer to the ocean. *Comptes Rendus Geoscience* 344: 723–738.

Al content relative to Si is higher (Opfergelt and Delmelle, 2012). The general trend between $\delta^{30}\text{Si}$ and CIA in bulk soils from various regions and contexts shows that weathering intensity rather than climate or lithology seems to influence recorded $\delta^{30}\text{Si}$ in soils. This concept is supported by the variation in $\delta^{30}\text{Si}$ reported in soils for a single clay mineral, but which was precipitated under different weathering conditions (e.g., Cornelis et al., 2019; Guinoiseau et al., 2021). The advantage of the Si isotope system is because during chemical weathering, both kinetic and equilibrium fractionation (see above) can be preserved in clay minerals. The isotope expression of these two fractionation regimes depends on the rate of clay formation. Significant isotope fractionation toward lighter isotope values in minerals is typical of a kinetically driven mechanism and characteristic of high rates of formation during intense weathering activity. Equilibrium-driven fractionation is almost negligible and characteristic of low rates of formation and moderate weathering intensity in soils (details in Guinoiseau et al., 2021). This finding is consistent with experimental and theoretical calculations of isotope fractionation showing that most of fractionation in soils is of kinetic origin (Poitrasson, 2017). The main advantage is that the rates of mineral formation can in turn be linked to environmental conditions, such as fluid residence time at the time of mineral precipitation. For example, Guinoiseau et al. (2021) have shown that the Si isotopes along a lateritic profile depend on the rapidity with which rainwater is percolating downwards through soil porosity, and that in such a context they can reveal the past evolution of climate in the tropics. Under highly weathered environments (right part of the Fig. 1), Si isotopes can also efficiently trace cycles of dissolution–reprecipitation of silicate material (clays but also silcrete, i.e. silicon duricrust) when the weatherable mineral reserve becomes exhausted and that newformed clays start to dissolve (Opfergelt and Delmelle, 2012).

The idealized view of the Si isotope system should be slightly nuanced because Si, as a nutrient, will have its isotope signature in superficial soil horizons impacted by biological cycling, namely through silicon uptake by plants and or the production of phytoliths. Even if the weathering process can be catalyzed by biological activity, the latter may appear as a complicating factor when trying to use Si isotopes to identify silicate weathering, especially in a soil where Si cycling is dominated by biological overprinting (e.g. Riotte et al., 2018).

Finally, Mg is by far the most complex isotope system due to the variability in $\delta^{26}\text{Mg}$ between primary minerals of igneous and sedimentary rocks, to its involvement in carbonate rocks and the effect of biological cycling on its fate in soils (details in Guo et al., 2019). A large variability of $\delta^{26}\text{Mg}$ (from -1.64 to $+0.65\text{‰}$) is observed in soil profiles because of the preferential incorporation of light isotopes in carbonates, the variable $\delta^{26}\text{Mg}$ fractionation associated with Mg transport between biological compartments (root, stem, leaves) during plant growth, or the versatility of $\delta^{26}\text{Mg}$ recorded in clays (details in Teng, 2017). Magnesium can be incorporated in the octahedral sheet of clay minerals with a preference for heavy Mg isotopes (see Wimpenny et al., 2014 in Teng, 2017). A second Mg pool, commonly considered as exchangeable Mg, occurs as adsorbed Mg through electrostatic interactions with the clay exchange complex (interlayers or edge sites of clay minerals) and is, in contrast, enriched in light isotopes. To understand Mg dynamics correctly and how Mg isotopes depend on silicate weathering pathways, a characterization of the different compartments of soils (bulk soil, clay fraction, exchangeable pool, dissolved Mg in soil solutions, vegetation) is necessary. Following this idea, Opfergelt et al. (2014) presented the main drivers of the Mg cycle in Icelandic soils. They highlight the major role played by the soil exchange complex on the export of heavy Mg isotopes out of the soil profile or on the preferential uptake of heavy Mg isotopes by vegetation. This study is a perfect example of the use of Mg isotopes as a proxy to identify the complex interplay between biotic and abiotic mechanisms on the fate of a nutrient in soil.

Influence of organic matter complexation in the soil profile

The formation of metal–organic complexes represents another possibility for metal isotopes to fractionate due to thermodynamic differences in the bonding environment. Isotope effects associated with metal binding to organic matter are of particular interest in soils, particularly for tracing pedogenetic processes, such as podzolization. To illustrate the identification of organo–metallic complexation, we focus on Fe isotopes (as $\delta^{57}\text{Fe}$) since iron mobility can be significantly increased by organo–metallic complexes in soils, namely podzols.

Podzol formation is controlled by acid weathering and vertical translocation of Fe in the soil profile, together with Al, by organic matter complexation. Podzols cover around 485 million ha worldwide, mainly in the temperate and boreal regions of the Northern Hemisphere but are also present in humid temperate climates and in the humid tropics. During podzolization, Fe-bearing minerals are dissolved and Fe is translocated and enriched in the subsoil under the influence of organic ligands. Such iron–aluminum–organic complexes are then transported vertically downward until they re-precipitate in the B horizon. All these processes may induce Fe isotope fractionation. By comparing the isotopic composition of Fe in two soil profiles (Podzol and Cambisol), Wiederhold et al. (2007a) were able to specify the isotope fractionation of iron associated with the process of podzolization (i.e. complexation of Fe by organic acids and translocation of Fe in a complexed form), in generally oxidizing conditions, which therefore suppresses the redox effect. At the scale of the podzolic profile, a significant enrichment of Fe content is observed in Bh and Bs horizons while a depletion characterizes the overlying bleached A/E horizons. Systematic lighter $\delta^{57}\text{Fe}$ in B horizons and heavier $\delta^{57}\text{Fe}$ in A/E horizons are linked to the preferential translocation of light Fe isotopes complexed by dissolved organic matter, but also to the formation of poorly crystalline Fe oxides (Fekiacova et al., 2017). The Fe content in the Cambisol is roughly homogeneous and characterized by slight or moderate weathering of parent material and by the development of brownish colors in the Bw horizon due to the weathering of primary silicate minerals. The release of ferrous iron (Fe^{II}) is further oxidized and precipitated as iron oxide minerals. The variation of $\delta^{57}\text{Fe}$ is limited with depth, in accord with pedogenesis under oxic and well-drained conditions which involves neither transformations of Fe bearing minerals nor Fe translocation.

The previous two sections have shown how the isotopic effects of complexation, sorption, or interactions with plants and microorganisms can potentially overlap in soils. To identify organic complexation in soils, it was thus crucial to quantify its intrinsic isotope fractionation. However, separating free and organic-complexed species without inducing artifacts of isotope fractionation during the separation procedure is experimentally difficult. Among the various methods used to decipher the effect of metal complexation on isotope fractionation, the Donnan-membrane method coupled to thermodynamic modelling was successfully used for Cu and Zn isotopes (see Jouvin et al., 2009 and Ryan et al., 2014 in Moynier et al., 2017). For the latter, a strong isotope fractionation was observed, varying with pH and the type of binding site (low-affinity such as phenolic groups vs high affinity such as carboxylic groups). Heavy isotope enrichments were observed for Zn bound to a purified humic acid in comparison with free Zn^{2+} ions due to their stronger binding environments (e.g. Komárek et al., 2021).

Redox influence on pedogenesis

With silicate weathering and organic complexation, redox processes are responsible for the structure of soils and the redistribution of elements between and within soil horizons and or soil solutions. To determine the magnitude of redox processes in soils, isotope signatures of redox sensitive elements (e.g. Fe or Cr) have now been applied for 15 years. Important investigations were inevitably directed to iron isotope systematics, a major element which can be found as Fe^{II} or Fe^{III} species in soils. Large isotope fractionations (up to 3‰ between Fe^{III} -oxide and aqueous Fe^{II}) were determined for abiotic or biotic redox processes through comprehensive laboratory experiments (see the overview in Wu et al., 2019) with enrichment in light Fe isotopes in Fe^{II} species and heavier ones in Fe^{III} species. However, the variation in iron isotope signatures in bulk soils is surprisingly limited ($\delta^{57}\text{Fe}$ from -0.4 to 0.5‰ , Wu et al., 2019). The occurrence of simultaneous processes (oxidation, reduction, adsorption, organic complexation) and the kinetic vs equilibrium nature of isotope fractionation can explain this difference between laboratory and environmental conditions. Clearly, the use of Fe isotopes as tracers of redox processes depends on the type of soils. In well aerated soils such as Cambisols or Ferralsols, $\delta^{57}\text{Fe}$ is not modified compared with the bedrock value. It is associated with the poor mobility of dissolved iron in solution under these conditions and the quantitative in situ precipitation of iron as oxides and hydroxides (see Poitrasson et al., 2008 in Poitrasson, 2017; Wiederhold et al., 2007a). The variation of $\delta^{57}\text{Fe}$ in podzols was explained in the previous section as the effect of organic ligand control on chemical weathering and on the translocation of Fe species between illuviated and eluviated horizons. In waterlogged or redoximorphic soils (e.g. Gleysols), significant Fe isotope fractionation is associated with alternating oxidizing and reducing conditions, which enhances Fe mobility as Fe^{II} during periods of anoxia (see review of Wu et al., 2019). In these soils, the weathering of primary minerals or iron oxides is redox-controlled and is monitored through the release of light Fe^{2+} in solution and subsequent enrichment of the residual fraction in heavy isotopes. The reduced Fe then migrates and can leave the system, be adsorbed on oxides or be precipitated as oxides or hydroxides in deeper horizons with a depleted $\delta^{57}\text{Fe}$ value compared to that of topsoils (e.g. Wiederhold et al., 2007b). This depleted $\delta^{57}\text{Fe}$ in amorphous or crystalline oxides in comparison with Fe in the silicate phases has also been noted at the centimeter scale within a soil profile with homogeneous bulk $\delta^{57}\text{Fe}$ (Fekiacova et al., 2021). The low $\delta^{57}\text{Fe}$ contained in Fe—Mn oxides and the resulting high $\delta^{57}\text{Fe}$ in clays show that, within a soil profile where Fe mobility seems reasonably static, redox processes can occur efficiently and redistribute Fe between different carriers at the centimeter or micrometer scale.

Due to the complex geochemistry of iron and its limited isotope variation in soils, other tracers such as Cr isotopes ($\delta^{53}\text{Cr}$ hereafter) were recently developed to track the oxidation process in recent soils and paleosols with the goal to date the beginning of oxygenation of the Earth's atmosphere (see review of Wei et al., 2020). Chromium is mostly present in rocks and soils as Cr^{III} (chromite FeCr_2O_4 being the main Cr^{III} mineral), but it can be oxidized as Cr^{VI} by Mn oxides or hydrogen peroxide (details in Liang et al., 2021). Oxidation and partial reduction of Cr induce large isotope fractionation (Ellis et al., 2002), releasing heavy Cr^{VI} in solution and retaining a light $\delta^{53}\text{Cr}$ in residual solids. To date, although the number of studies is scarce, most modern and ancient weathering profiles report negatively fractionated $\delta^{53}\text{Cr}$ in accord with the oxidative mobilization of Cr. The absence of variation in soil $\delta^{53}\text{Cr}$ can also be useful because it should be either a diagnostic of acid dissolution without a redox process if Cr depletion is severe (Berger and Frei, 2014) or of a too-limited weathering intensity if chromium's depletion is minimal (e.g. Wille et al., 2018).

Control of pedological processes on trace metal behavior

In section “Isotopes and pedogenetic processes,” we highlighted the use of isotope ratios to identify the main pedological mechanisms that occur in soil, impacted by variation in mineralogy, redox or climatic conditions. Here, we will identify the main pedological processes controlling the mobility and speciation of trace metals through variation in their isotope signatures in soils. Reference reviews provide excellent summaries of the metal isotope fractionation occurring during chemical (adsorption, oxidation, organic complexation, precipitation) or biological (uptake) reactions (Komárek et al., 2021), which must be clearly identified prior to their use as non-traditional isotope systems for tracers in the environment (Wiederhold, 2015). In this chapter, we illustrate the effect of pedological processes on the dynamics of two trace metals, Cu and Zn, which present two clear advantages. First, experimental isotope fractionations associated with the main processes occurring in soils are well identified and second, the number of studies reporting isotope signatures for Cu (as $\delta^{65}\text{Cu}$) and Zn (as $\delta^{66}\text{Zn}$) in soils have increased considerably during the last 10 years allowing a comprehensive survey of their behavior in soil.

In the vast majority of cases, precipitation of secondary silicate minerals (e.g. clays) from primary minerals does not induce significant isotope fractionation for Zn and Cu incorporated in their crystalline structure (e.g. Little et al., 2019). In soil developed in tropical climates, metal sorption on clays is largely inhibited by acidic pH. This apparent limit is in reality useful to identify the two main mechanisms inducing Zn and Cu redistribution through soil profiles and fractionation of Cu and Zn isotopes: organic complexation and oxide adsorption/coprecipitation (Vance et al., 2016).

During pedogenesis, Cu and Zn depletions (expressed as τ_x values in Fig. 2) occur through their release from bedrock to soil solutions and they are often associated with an enrichment in light isotopes in the secondary minerals. Based on a synthesis of most $\delta^{65}\text{Cu}$ and $\delta^{66}\text{Zn}$ reported in soils, Vance et al. (2016) and more recently Little et al. (2019), explained this evolution by the partitioning of Cu and Zn isotopes between two main reservoirs. With the affinity of both metals to organic ligands, a pool of heavy Cu and slightly heavy Zn isotopes is complexed during or more likely after mineral dissolution and exported out of the system as colloids or organic particles in the soil solution. This process is supported by laboratory experiments (details in Moynier et al., 2017). In addition, the greater reactivity of Cu than Zn to organic ligands is consistent with the most severe depletion of heavy Cu during weathering and soil formation (negative $\Delta^{65}\text{Cu}_{\text{soil-parent_rock}}$ for lower τ_{Cu}) than for Zn isotopes (Fig. 2 and details in Moynier et al., 2017). The residual fraction, de facto enriched in light isotopes, is associated with iron and manganese oxy-hydroxide as adsorbed or internalized species. The process involved between adsorption and coprecipitation is still debated as experimental sorption of Cu and Zn on Fe and Mn oxides enriches the mineral surface in heavy isotopes (details in Komárek et al., 2021), in contradiction with natural observations. The preferential complexation of Cu and Zn by organic ligands can partly explain scavenging of the complementary light pool by oxide surfaces. However, recent studies developed on lateritic tropical profiles, suggest that incorporation of Cu and Zn rather than adsorption could have caused light isotope enrichment in oxides during 'ferrugination' through kinetic isotope control (details in Little et al., 2019), even if experimental isotope fractionation

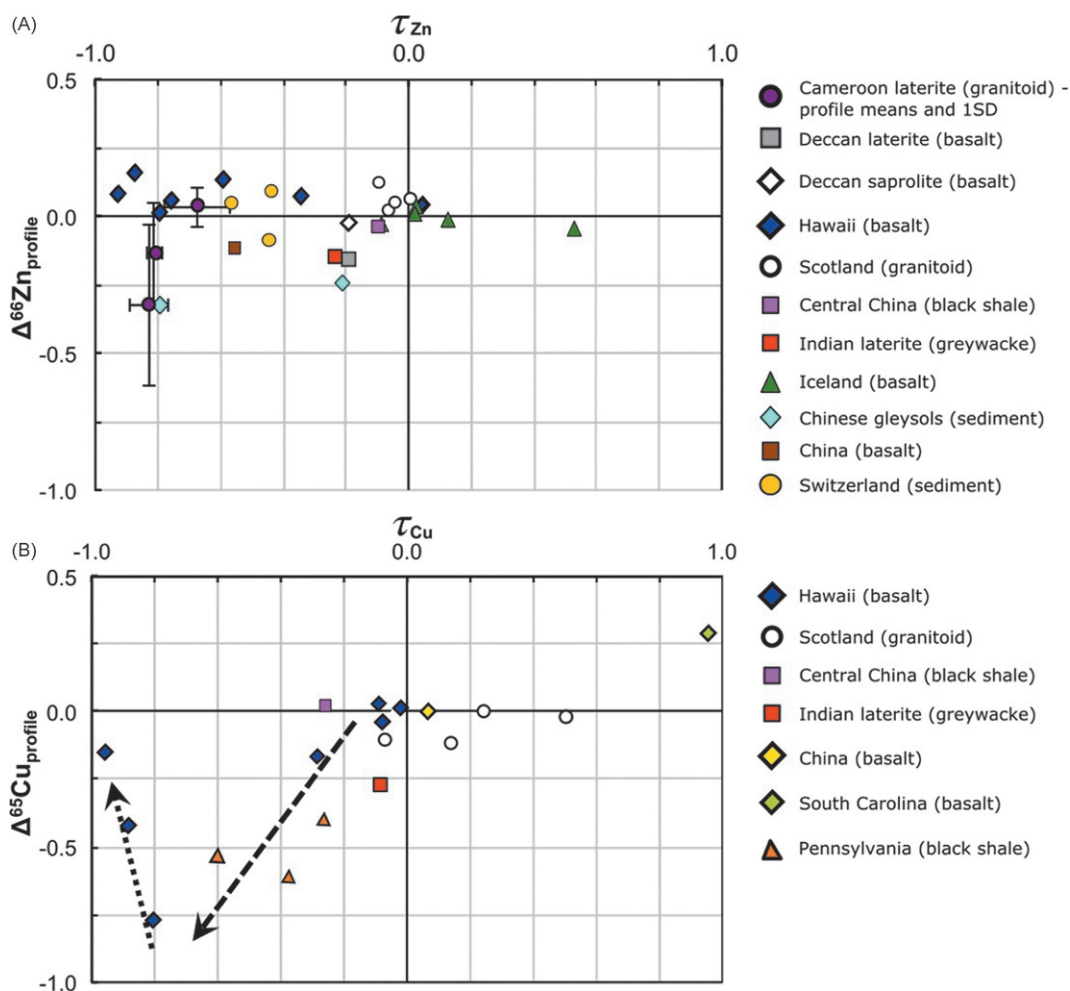


Fig. 2 Evolution of integrated isotope compositions with integrated τ_x for soil profiles. The integrated Cu and Zn isotopes in soils are reported as the deviation between weathered soil profiles and the parent material ($\Delta^{65}\text{Cu}_{\text{profile}} = \delta^{65}\text{Cu}_{\text{int_soil}} - \delta^{65}\text{Cu}_{\text{parent_rock}}$ and $\Delta^{66}\text{Zn}_{\text{profile}} = \delta^{66}\text{Zn}_{\text{int_soil}} - \delta^{66}\text{Zn}_{\text{parent_rock}}$). The integrated τ_x is defined using depth, invariant element (Nb (niobium), Zr (zirconium), Ti (titanium) or Al depending on the study) and soil density when available. This figure updates the previous compilation published in Little SH, Munson S, Prytulak J, Coles BJ, Hammond SJ, and Widdowson M (2019) Cu and Zn isotope fractionation during extreme chemical weathering. *Geochimica et Cosmochimica Acta* 263: 85–107.

during Cu and Zn incorporation in oxides is currently unknown. This complex interplay between organic matter and oxide controls in soils has recently been illustrated for Zn in Icelandic soils (Opfergelt et al., 2017). In this area, the light isotope enrichment observed during pedogenesis is not coupled with the total organic carbon but with the plant-derived organic pool that has been preserved from decomposition through organo–mineral association, typically iron oxides.

To summarize, in soils, the control of Cu and Zn speciation and dynamics by the organic compartment is dominant while immobilization on or in oxides seems secondary. The primary control of organic material is even emphasized in topsoil. In these organic-rich O horizons that can be partially decoupled from the lower part of the soil, the light isotope signatures are mostly associated with the mineralized organic debris of the aboveground vegetation. Enrichment in light Cu and Zn isotopes has been observed in most of higher plants through processes of biological uptake and Cu reduction during assimilation (details in Moynier et al., 2017). Finally, under extreme weathering conditions, demobilization of iron oxides can cause severe Cu and Zn depletion ($\tau_X < -0.8$, Fig. 2). The associated shift toward higher $\delta^{65}\text{Cu}$ and to a lesser extent $\delta^{66}\text{Zn}$ is presumably linked with allochthonous Cu and Zn inputs as atmospheric dust (Vance et al., 2016).

Tracing the source of metal pollution in soils

Soil pollution is one of the relevant threats known to affect soil quality, and in a broader extent human health, food quality, as well as soil biodiversity and water quality. At high content, most metals can be seen as contaminants in soils, mainly due to anthropogenic activities. Metal isotopes can be used to trace the source of the contamination and the number of studies has increased rapidly during the past 15 years. To illustrate the potential that metal isotopes can have to fingerprint the origin of metals in human-impacted environments, we focus on two isotope systems, Pb and Zn.

Lead is one of the main and oldest pollutants of anthropogenic origin, whose origin can be associated with mining, metallurgy, batteries, pigments, ceramics, burning fossil fuels, sludge or treatment plants. Lead has four stable isotopes; three radiogenic: ^{208}Pb (52%), ^{206}Pb (24%), ^{207}Pb (23%) and one non-radiogenic: ^{204}Pb (1%). The isotope composition of Pb is expressed in the form of ratios: e.g. $^{206}\text{Pb}/^{204}\text{Pb}$, $^{206}\text{Pb}/^{207}\text{Pb}$, $^{208}\text{Pb}/^{206}\text{Pb}$. Lead isotopes are commonly used to differentiate anthropogenic Pb inputs from the geochemical background (i.e. geogenic Pb) in soils. They also make it possible to follow the dynamics of Pb in the soil, i.e. its accumulation on the surface of the soil in the litter, or on the contrary, to highlight the pedogenetic control on Pb migration within deeper mineral or organo–mineral horizons (Podzol). By studying the isotopic composition of Pb in two soil profiles (Andosol and Podzol) representative of undisturbed ecosystems and located far from pollution point sources, such as cities, main roads, and industries, Semlali et al. (2001) were able to highlight the strong contribution of anthropogenic Pb in soil litter and A horizons relative to geogenic Pb. With the $^{206}\text{Pb}/^{207}\text{Pb}$ isotopic ratio, they demonstrated that the exogenous/anthropogenic Pb mainly came from historical pollution ($^{206}\text{Pb}/^{207}\text{Pb}$ value 1.16–1.18) linked to the use of Pb ores, rather than pollution linked to automobile activity (leaded gasoline in France had $^{206}\text{Pb}/^{207}\text{Pb}$ ratios from 1.069 to 1.094) or industrial activity ($^{206}\text{Pb}/^{207}\text{Pb}$ ratios varied from 1.143 to 1.155). Andosol pedogenesis is characterized by hydrolysis of volcanic material and neoformation of short-range order minerals such as allophane and imogolite, intimately associated with organic matter. Together, organic matter and short-range order minerals contribute to high Pb retention in the profile, and exogenous Pb accumulates in organic matter rich horizons, i.e. topsoil. Pedogenesis in podzols (i.e. strongly weathered and acid soils, developed on well-drained substrates), is controlled by organic acids produced in the litter that act as strong weathering and metal-complexing agents. Anthropogenic Pb is dominant in litter horizons, but is also found in Bh horizons of the Podzol in humus coatings on mineral grains, due to the migration of Pb associated with colloidal organic matter.

Road activity is also a main source of contamination of soils located near highways, due to the presence of Pb in gasoline until its exclusion in the 2000s. Walraven et al. (2014) clearly demonstrated the effect of leaded gasoline on soils near two highways in The Netherlands, in which Pb accumulated in the top 15 cm of the soil associated with organic matter. This exogenous fraction of Pb in soils is often more mobile and bioavailable than geogenic Pb and can contribute to Pb contamination of groundwater. The migration of Pb in acidic soils with low reactive mineral content reduces Pb retention (Walraven et al., 2014). This high content of available Pb can also be a source of contamination to aboveground plants (Chrastny et al., 2010).

If highly concentrated, zinc can also become a pollutant in soils with various origins: sewage sludge and pig slurry spreading, vehicle emissions and road runoff, industrial and smelter emissions. Zinc isotopes ($\delta^{66}\text{Zn}$), together with Zn solid speciation have been investigated along two soil profiles sampled near one of the largest Pb and Zn processing plants in Europe in northern France (Juillot et al., 2011). The two soils studied show large Zn contents in topsoils (570 to 1380 mg kg⁻¹) which decrease markedly with depth to around 50 mg kg⁻¹, close to the regional Zn geochemical background. Spectroscopic investigations have shown that Zn occurs in the form of franklinite-bearing slag grains originating from processing wastes in the upper soil horizons, while Zn-bearing illite is the main geogenic Zn-containing species. The $\delta^{66}\text{Zn}$ values in the mineral horizons (mean value $0.31 \pm 0.38\text{‰}$) are interpreted as being representative of the local geochemical background, whereas heavier $\delta^{66}\text{Zn}$ values near the surface of the two soils ($0.38 \pm 0.45\text{‰}$ to $0.76 \pm 0.14\text{‰}$) are related to anthropogenic Zn in line with the smelter waste signatures ($\delta^{66}\text{Zn}$ values of $0.81 \pm 0.20\text{‰}$) from the vicinity of the processing plant. It is thus possible to use Zn isotopes to fingerprint Zn sources in smelter-impacted soils. However, the lighter $\delta^{66}\text{Zn}$ measured in the surface horizons compared to that in the slag can also result from an isotope fractionation resulting from pedogenetic bio-physicochemical processes, showing the need to combine isotope approaches with other proxies such as solid speciation analysis.

Conclusions and reflections

The use of non-traditional stable isotope systems in environmental geochemistry is a fast-developing scientific field. Even if the determination of intrinsic isotope fractionation in the laboratory during complexation, biological uptake, adsorption or precipitation reactions has been developed for some isotope systems (“isotope fractionation black box”), efforts must be continued to remove barriers preventing application of isotope systems in the Critical Zone. Theoretical approaches (e.g. ab initio calculations) can also help in understanding mechanisms of fractionation at the molecular scale. This chapter has shown that the use of metal stable isotopes in soils is a powerful tool. Identification of weathering processes and of the main physicochemical processes affecting metals in soil can also be facilitated by tracing the pollution source in human-impacted environments.

Finally, it must be noted that this chapter focusses on abiotic processes. Soil contains a large variety of habitats and nutrients, and many living organisms (plants, bacteria, fungi, invertebrates) are present. They contribute to the recycling of organic matter, mineral weathering, the modification of water flow, and to metal cycling. Studies have shown that metals can be fractionated by biological processes, which are governed by the same physical and chemical principles as abiotic processes, as described above (see for instance Wiederhold, 2015; Poitrasson, 2017; Moynier et al., 2017; Bouchez et al., 2013). Understanding of the complex interplay between abiotic and biotic control on metal dynamics is still at the beginning and it represents an exciting research avenue for the Critical Zone community.

References

- Albarède F and Beard B (2004) Analytical Methods for Non-Traditional Isotopes. *Reviews in Mineralogy and Geochemistry* 55: 113–152.
- Berger A and Frei R (2014) The fate of chromium during tropical weathering: A laterite profile from Central Madagascar. *Geoderma* 213: 521–532.
- Bigeleisen J and Mayer MG (1947) Calculation of equilibrium constants for isotopic exchange reactions. *The Journal of Chemical Physics* 15: 261–267.
- Bouchez J, von Blanckenburg F, and Schuessler JA (2013) Modeling novel stable isotope ratios in the weathering zone. *American Journal Science* 313: 267–308.
- Chapela Lara M, Buss HL, Henehan MJ, Schuessler JA, and McDowell WH (2022) Secondary minerals drive extreme lithium isotope fractionation during tropical weathering. *Journal of Geophysical Research: Earth Surface* e2021JF006366.
- Chorover J, Kretzschmar R, Garcia-Pichel F, and Sparks DL (2007) Soil Biogeochemical Processes within the Critical Zone. *Elements* 3: 321–326.
- Chrastný V, Komárek M, and Hájek T (2010) Lead contamination of an agricultural soil in the vicinity of a shooting range. *Environmental Monitoring and Assessment* 162: 37–46.
- Conway TM, Rosenberg AD, Adkins JF, and John SG (2013) A new method for precise determination of iron, zinc and cadmium stable isotope ratios in seawater by double-spike mass spectrometry. *Analytica Chimica Acta* 793: 44–52.
- Cornelis JT, Weis D, Opfergelt S, Van Ranst E, and Dumon M (2019) Past and current geochemical conditions influence silicon isotope signatures of pedogenic clay minerals at the soil profile scale, Ethiopia. *Chemical Geology* 524: 174–183.
- Ellis AS, Johnson TM, and Bullen TD (2002) Chromium isotopes and the Fate of Hexavalent Chromium in the Environment. *Science* 295: 2060.
- Fekiacova Z, Vermeire ML, Bechou L, Cornelis JT, and Cornu S (2017) Can Fe isotopes fractionations trace the pedogenetic mechanisms involved in podzolization? *Geoderma* 296: 38–46.
- Fekiacova Z, Montagne D, Duvivier A, Guihou A, Deschamps P, and Cornu S (2021) Evolution of Retisol impacted by artificial drainage: What can we learn from stable Fe isotope ratios? *Geoderma* 384: 114771.
- Guinoiseau D, Fekiacova Z, Allard T, Druhan JL, Balan E, and Bouchez J (2021) Tropical weathering history recorded in the silicon isotopes of lateritic weathering profiles. *Geophysical Research Letters*. e2021GL092957.
- Guo B, Zhu X, Dong A, Yan B, Shi G, and Zhao Z (2019) Mg isotopic systematics and geochemical applications: A critical review. *Journal of Asian Earth Sciences* 176: 368–385.
- Juillot F, Maréchal C, Morin G, Jouvín D, Cacaly S, Telouk P, Benedetti MF, Ildefonse P, Sutton S, Guyot F, and Brown GE Jr. (2011) Contrasting isotopic signatures between anthropogenic and geogenic Zn and evidence for post-depositional fractionation processes in smelter-impacted soils from Northern France. *Geochimica et Cosmochimica Acta* 75: 2295–2308.
- Komárek M, Ratič G, Vaňková Z, Šípková A, and Chrastný V (2021) Metal isotope complexation with environmentally relevant surfaces: Opening the isotope fractionation black box. *Critical Reviews in Environmental Science and Technology* 1–31.
- Liang J, Huang X, Yan J, Li Y, Zhao Z, Liu Y, Ye J, and Wei Y (2021) A review of the formation of Cr(VI) via Cr(III) oxidation in soils and groundwater. *Science of The Total Environment* 774: 145762.
- Little SH, Munson S, Prytulak J, Coles BJ, Hammond SJ, and Widdowson M (2019) Cu and Zn isotope fractionation during extreme chemical weathering. *Geochimica et Cosmochimica Acta* 263: 85–107.
- Moynier F, Vance D, Fujii T, and Savage P (2017) The isotope geochemistry of zinc and copper. *Reviews in Mineralogy and Geochemistry* 82: 543–600.
- Opfergelt S and Delmelle P (2012) Silicon isotopes and continental weathering processes: Assessing controls on Si transfer to the ocean. *Comptes Rendus Geoscience* 344: 723–738.
- Opfergelt S, Burton KW, Georg RB, West AJ, Guicharnaud RA, Sigfusson B, Siebert C, Gislason SR, and Halliday AN (2014) Magnesium retention on the soil exchange complex controlling Mg isotope variations in soils, soil solutions and vegetation in volcanic soils, Iceland. *Geochimica et Cosmochimica Acta* 125: 110–130.
- Opfergelt S, Cornelis JT, Houben D, Givron C, Burton KW, and Mattielli N (2017) The influence of weathering and soil organic matter on Zn isotopes in soils. *Chemical Geology* 466: 140–148.
- Penniston-Dorland S, Liu X-M, and Rudnick RL (2017) Lithium Isotope Geochemistry. *Reviews in Mineralogy and Geochemistry* 82: 165–217.
- Poitrasson F (2017) Silicon Isotope Geochemistry. *Reviews in Mineralogy and Geochemistry* 82: 289–344.
- Riotte J, Meunier J-D, Zambardi T, Audry S, Barboni D, Anupama K, Prasad S, Chmieleff J, Poitrasson F, Sekhar M, and Braun J-J (2018) Processes controlling silicon isotopic fractionation in a forested tropical watershed: Mule Hole Critical Zone Observatory (Southern India). *Geochimica et Cosmochimica Acta* 228: 301–319.
- Schauble EA, Méheut M, and Hill PS (2009) Combining Metal Stable Isotope Fractionation Theory with Experiments. *Elements* 5: 369–374.
- Semlali RM, van Oort F, Denaix L, and Loubet M (2001) Estimating distributions of endogenous and exogenous Pb in soils by using Pb isotopic ratios. *Environmental Science and Technology* 35: 4180–4188.
- Teng F-Z (2017) Magnesium Isotope Geochemistry. *Reviews in Mineralogy and Geochemistry* 82: 219–287.
- Vance D, Matthews A, Keech A, Archer C, Hudson G, Pett-Ridge J, and Chadwick OA (2016) The behaviour of Cu and Zn isotopes during soil development: Controls on the dissolved load of rivers. *Chemical Geology* 445: 36–53.
- Walraven N, van Os BJH, Klaver GT, Middelburg JJ, and Davies GR (2014) The lead (Pb) isotope signature, behaviour and fate of traffic-related lead pollution in roadside soils in The Netherlands. *The Science of the Total Environment* 472: 888–900.
- Wei W, Kläbe R, Ling H-F, Huang F, and Frei R (2020) Biogeochemical cycle of chromium isotopes at the modern Earth's surface and its applications as a paleo-environment proxy. *Chemical Geology* 541: 119570.
- Wiederhold JG (2015) Metal stable isotope signatures as tracers in environmental geochemistry. *Environmental Science and Technology* 49: 2606–2624.

- Wiederhold JG, Teutsch N, Kraemer SM, Halliday AN, and Kretzschmar R (2007a) Iron isotope fractionation in oxic soils by mineral weathering and podzolization. *Geochimica et Cosmochimica Acta* 71: 5821–5833.
- Wiederhold JG, Teutsch N, Kraemer SM, Halliday AN, and Kretzschmar R (2007b) Iron isotope fractionation during pedogenesis in redoximorphic soils. *Soil Science Society of America Journal* 71: 1840–1850.
- Wille M, Babechuk MG, Kleinhanns IC, Stegmaier J, Suhr N, Widdowson M, Kamber BS, and Schoenberg R (2018) Silicon and chromium stable isotopic systematics during basalt weathering and lateritisation: A comparison of variably weathered basalt profiles in the Deccan Traps, India. *Geoderma* 314: 190–204.
- Wu B, Amelung W, Xing Y, Bol R, and Berns AE (2019) Iron cycling and isotope fractionation in terrestrial ecosystems. *Earth-Science Reviews* 190: 323–352.

Alkali metals in soils

Guy J Levy and Asher Bar-Tal, Institute of Soil, Water and Environmental Sciences, ARO, Volcani Institute, Rishon LeZion, Israel

© 2023 Elsevier Ltd. All rights reserved.

The alkali metals	116
Potassium in soils and plant nutrition	118
Potassium in soils	118
Potassium effects on soils structure and physical properties	119
Potassium in plant nutrition	119
Potassium availability	120
Fertilizer recommendations and effects on crops and soil	120
Sodium in soils	121
Sodium and soil productivity	122
Sodium and carbon (C) sequestration	124
Concluding comments	124
References	125

Abstract

The current chapter focuses on potassium (K) and sodium (Na) of the alkali metals following their central role in agricultural soils. The former is an essential plant macro-nutrient and accumulation of the latter in soils often results in degradation of soil structure and associated physicochemical, physical and hydraulic properties. In soil, K is present mostly in the structural pool (90%), while the K ion in the soil solution is just 1 to 2% of the soil K pool, yet it is the most available form for plants. Sodium is, in general, the dominant alkali cation in soils. Productivity in Na-affected (sodic) soils is limited because of deterioration in the physical properties, crop tolerance to Na per se and plant nutrition in Na-affected soils.

Key points

- Potassium and Sodium belong to the alkali metals group and play a central role in agricultural soils.
- Potassium ion in the soil solution is the most available form for plants.
- Potassium availability in soils for plants is best estimated based on the concept of quantity (Q) and intensity (I).
- Productivity of Na-affected (sodic) soils is limited.
- Sodic soils rehabilitation may positively affect C cycle and accumulation in soils.

The alkali metals

Group 1A in the Periodic Table comprises the elements lithium (Li), sodium (Na), potassium (K), rubidium (Rb), cesium (Cs) and francium (Fr) that are commonly referred to as Alkali metals. In soils, all these elements, but for Fr, exist in the form of stable isotopes. Conversely, only the radioactive isotope ^{223}Fr occurs naturally but it is rather difficult to find in soils. Readers interested in an in depth discussion on the alkali metals are referred to the review by Scott and Smith (1987) and references cited therein.

In brief (with no reference to Fr), the alkali metals share the same outer electron configuration, and in soils they are reactive and form monovalent cations and soluble salts. Yet, they differ in numerous characteristics, as well as, in their relative abundance and quantity in soils (Table 1).

A chapter containing a detailed discussion on the cation exchange complex of soil solid components (clays, oxides, organic matter) is included in this Encyclopedia. Readers interested in sources of the surface charge of soil particles, and theory of the distribution of ions between the charged surfaces and the soil solution, and the theories of cation exchange are referred to that chapter. The distribution of the alkali metals between the solution and cation exchange complex of the above-mentioned soil solid components has a strong effect on their mobility in soil, on soil structural stability and availability to plants. Therefore, here we present briefly the factors affecting the exchange of the alkali metals with each other and with the dominant divalent cations (Ca^{+2} and Mg^{+2}). The two main factors controlling the binding strength of the charged surface for a specific cation according to the theory of the diffuse double layer charge are the valence and the hydrated radii of the cation. The binding strength of cations increase as a function of the valence, trivalent > divalent > univalent. All the alkali metals are univalent, hence this factor is identical in their binding strength. Consequently, the hydrated radii of the cations is the main factor determining the order of their binding strength, which decreases as the hydrated radii increases. From the values of the hydrated radii (Table 1) we can derive the following order of

Table 1 Characteristics and relative abundance (atoms per 10^6 Si atoms) and quantity of the alkali elements in soils.

	<i>Li</i>	<i>Na</i>	<i>K</i>	<i>Rb</i>	<i>Cs</i>
Atomic number	3	11	19	37	55
Atomic weight	6.939	22.9898	39.102	85.49	132.905
Ionic potential	1.32	1.02	0.75	0.68	0.60
Electronegativity	0.95	0.9	0.8	0.8	0.75
Hydrated radii (nm)	0.340	0.276	0.232	0.228	0.228
Hydration energy (kJ mol^{-1})	519	406	322	293	264
Relative abundance in soil	346	21,900	28,600	141	3
Quantity in soils (mg kg^{-1})	26	6200	14,000	140	5

Reproduced from Scott AD and Smith SJ (1987) Sources, amounts and forms of alkali elements in soils. In Stewart BA (Ed.) *Advances in Soil Science*. vol. 6, 101–147.

binding strength for the alkali metals: $\text{Cs}=\text{Rb}>\text{K}>\text{Na}>\text{Li}$. As the relative abundance of Na and K in soil is of the same order and is much higher (2 to 4 orders of magnitude) than the other alkali metals, this suggests that the exchange complex of soils has a greater affinity for K than for Na. A quantitative estimate of a nearly seven times greater affinity of soils for K over Na is presented in the report of the **US Salinity Laboratory (1954)**.

Naturally occurring lithium appears in the form of two stable isotopes, ^6Li and ^7Li , the latter being the more abundant (92.5% natural abundance) (Krebs, 2006). Lithium has a distinct involvement in soil-forming reactions; stable isotopic analyses of Li help to gain insight into the mechanisms and rates governing the transfer of matter from bedrocks to surface waters. Thus, Li isotopes are a powerful tool for understanding the processes leading to clay mineral formation in soils. Yet, no plant induced Li isotopic fractionation has been recorded, possibly due to its very low concentration in vegetation; it is three orders of magnitude less than that in soil (Schmitt et al., 2012). From the agricultural perspective, Li is not considered an essential plant nutrient, possibly due to its very low concentrations ($<100 \text{ mg kg}^{-1}$) in soils; yet, Li may have toxic effects on plants, especially citrus. Moreover, concentration of $<5 \text{ mg L}^{-1}$ of Li in irrigation water may lead to harmful effects on some plants. However, there are also reports on the beneficial effects of Li on plant growth and subsequently on human health through the food chain (Scott and Smith, 1987 and references cited therein).

Rubidium has two naturally occurring radioactive isotopes, ^{85}Rb (72.2%) and ^{87}Rb (27.8%) with a half-life of 48.8×10^9 years. Rubidium is the 20th most abundant element in the Earth's upper continental crust. The Rb geochemistry is closely related to that of potassium. Rubidium does not form its own minerals but, due to similarity of the ionic radii (Table 1), it can substitute for K in mica and K-feldspar, as well as in rarer minerals, like lepidolite or carnallite. Limestone has small concentrations of Rb (4 mg kg^{-1}), while granite and pegmatite usually have large Rb concentrations (e.g. 120 mg kg^{-1} in granite) as do some sedimentary rocks (140 mg kg^{-1}). Furthermore, it is generally assumed that Rb, after being dissolved during weathering, is adsorbed more strongly by the clay fraction of soils than potassium (Goldschmidt, 1954, in Negrel et al., 2018). Rubidium has very few applications and is mostly used in research, e.g., dating of geological materials using ^{87}Rb , photocells and some pharmaceuticals (e.g., as an antidepressant like Li). Glass dust could be one possible source of Rb transfer to the environment. An additional source of contamination to soils and the environment is the radioactive isotope ^{86}Rb from radioactive fallout; yet its short half-life time (18.6 days) lessens its importance in soils (Scott and Smith, 1987). In agricultural soils, Rubidium is considered as non-essential and non-toxic; it is easily taken up together with K being its associated trace element (Negrel et al., 2018).

Cesium, is the heaviest (nearly 1.6 as heavy as Rb) alkali metal with an estimated small abundance in the upper continental crust of 5 mg kg^{-1} (Table 1). There are however geochemical differences in its spatial distribution; Cs concentration in soils of northern Europe is two times lower than in those of southern Europe. Cesium forms some very rare minerals (e.g., pollucite) and often substitutes for K in mica (especially in muscovite and lepidolite) and in K-feldspar, and is enriched in felsic rocks during magmatic fractionation, with pegmatite having the largest Cs concentrations. With its high volatility, Cs tends to concentrate in the late stages of magmatic differentiation and in sublimates of K-rich volcanic rocks in active volcanic regions. Among sedimentary rocks, Cs concentrations are low in limestone and high in shale and schist. In a silicate environment, most Cs is transferred to the weathering fraction, similar to Rb, and is subsequently enriched in the clay fraction. The mobility of Cs in soil is therefore low, because it strongly adsorbs to the clay material. The strong affinity of Cs for weathering products (i.e., clay minerals) causes considerable fractionation of this element during the chemical weathering of rocks (Negrel et al., 2018 and references cited therein). Cesium is more toxic to plants than Rb, but less so than Li when absorbed at large concentrations, although this rarely occurs under natural environmental conditions.

Compared with the aforementioned alkali metals, K and Na have received far more attention as the former is an essential plant macro-nutrient and accumulation of the latter in soils often results in degradation of soil structure and deterioration of its physico-chemical, physical and hydraulic properties. Following the central role of K and Na in agricultural soils, therefore, each is discussed further in detail below.

Potassium in soils and plant nutrition

Potassium in soils

There are 25 known **isotopes** of potassium, three of which occur naturally: ^{39}K (93.3%), ^{40}K (0.0117%), and ^{41}K (6.7%). Naturally occurring ^{40}K has a **half-life** of 1.250×10^9 years. It decays to stable ^{40}Ar (argon) by **electron capture** or **positron emission** (11.2%) or to stable ^{40}Ca by **beta decay** (88.8%) (Audi et al., 2003). The decay of ^{40}K to ^{40}Ar is the basis of a common method for dating rocks. **Minerals** are dated by measurement of the concentration of K and the amount of radiogenic ^{40}Ar that has accumulated. The minerals best suited for dating include **biotite**, muscovite, metamorphic **hornblende**, and volcanic **feldspar**. Apart from dating, K isotopes have been used as **tracers** in studies of **weathering** and for **nutrient cycling** studies because K is a **macronutrient** required for the complete **life cycle** of the plant (Kirkby, 2012).

Potassium is a univalent cation with a hydrated ionic radius of 0.331 nm and a hydration energy of 314 J mol^{-1} . Potassium is the most abundant alkali metal in soil (Table 1). It is distributed between the following pools in the soil: structural, fixed (non-exchangeable), exchangeable, soluble, organic matter and microbial K (Fig. 1).

The potassium ion in the soil solution is the most available form for plants. However, just 1 to 2% of the soil K is in this pool. Over 90% of soil K is in the structural pool, mainly in micas and feldspars. Its rate of release to the soil solution and the exchangeable and fixed fractions is a very slow process, at a time scale of years; therefore, it is defined as unavailable to plants. However, these minerals affect long term soil fertility serving as large pools of potassium that replenish its depleted labile pools (Barber, 1995). Fixed (non-exchangeable) K is the second largest pool in soils and a considerable amount occurs in the secondary minerals of the mica group, and the vermiculite and Illite clay minerals. Fixed K is different from structural K (Rich, 1968) in that it is not bound within the crystal structures of soil minerals, but is held at the interlayer space of weathered micas, vermiculites, and hydroxyl-interlayered vermiculites (Fig. 2).

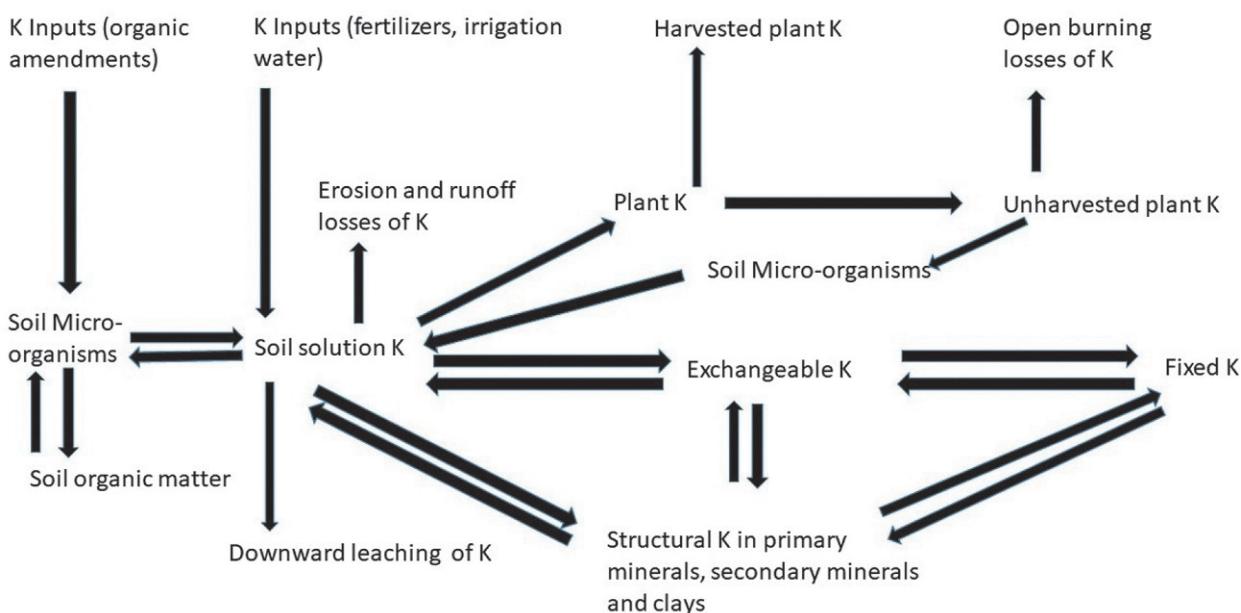


Fig. 1 A schematic diagram of the K cycle in soil including input, output and losses.

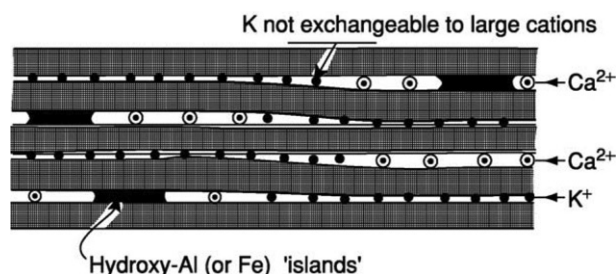


Fig. 2 Proposed model of an expandable layer silicate with interlayers indicating the effect on K fixation. Reproduced with permission from Rich CI (1968) Mineralogy of soil potassium. In: Kilmer VJ, Younts SE, and Brady NC (eds). *The Role of Potassium in Agriculture*. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.

Potassium becomes fixed because the binding forces between K and the mineral surfaces are greater than the hydration forces between individual K ions. Fixed K results in a partial collapse of the crystal structures and the K ions are physically trapped to varying degrees, making K release a slow, diffusion-controlled process. Fixed K can be replaced with ammonium (NH_4) which has a similar hydrated radius, yet larger hydrated cations like Ca, Mg and Na cannot fit into the space of the interlayers.

The release of fixed K to the soil solution is controlled by diffusion. The rate of release of fixed K to the soil solution and the exchangeable pools is much faster than that of the structural K, but much slower than exchangeable K. The contribution of the fixed K pool to K uptake by plants depends on soil properties, environmental conditions, soil water content, transpiration, root size and architecture (Barber, 1995). The exchangeable K pool is much larger than the soluble pool and thus is the main source of K to plants because of the very large exchange rate of K between the soil solution and the cation exchange complex (less than a millisecond) (Bar-Tal et al., 1995). Removal of K from the soil in the root zone includes uptake by plants and removal with harvested plant biomass, and K losses by downward leaching, soil erosion and runoff. A practical implication of the strong preference for K adsorption over Na is the limited mobility of K by leaching and the distribution of K in the soil profile following K fertilizer input and or irrigation water containing K. In clay soils, K accumulates in the top layer with decreasing concentration with depth, while in sandy soils there is a more uniform distribution of K with soil depth.

Addition of K to agricultural soils is mainly by fertilizers, manures and other organic wastes, and irrigation water. Note that when organic wastes are applied to soil most of the K in the organic matter is in inorganic form that becomes available in the soil solution. Nevertheless, a small fraction is consumed by the growing microorganisms that decompose the waste and produce soil organic matter. The dose of applied manure to cultivated lands is determined mainly by its nitrogen content, therefore, when manure rich in K is used, large amounts of K accumulate in the soil. Unharvested plant residues are also decomposed in the soil by microorganisms, which release K ions to the soil solution. The fraction of microorganism-derived K in the whole soil system is <1%, but it plays an important role in the cycling of K between the organic and inorganic pools, although it is much less important than in the cycles of nitrogen, sulfur and phosphorus.

Potassium effects on soils structure and physical properties

The impact of K on soil physical and hydraulic properties has received much less attention than that of Na. Moreover, research on K has yielded varying and at times conflicting results. Soil permeability has been reported to decrease in relation to the different exchangeable cations in the following order $\text{Ca} > \text{Mg} > \text{K} > \text{Na}$ (Arienzo et al., 2012). Some studies have reported that K can be considered equivalent to Na regarding its negative impact on soil physical properties, yet, other studies have noted larger aggregates with greater stability (Bronick and Lal, 2005) and hydraulic conductivity (HC) in K saturated soils than in those saturated with divalent cations. In addition, Levy and Torrento (1995) observed for a loamy sand and a sandy clay, that increasing exchangeable potassium percentage (EPP) up to 15% did not have an adverse effect on soil susceptibility to clay dispersion and aggregate breakdown. Chen et al. (1983) noted that the effect of exchangeable K on soil HC depended on soil type; in a clay and a loamy sand the HC increased up to 20 times with increases in EPP up to 20, and thereafter decreased following an additional increase in EPP. In a sandy clay the HC decreased uninterruptedly with an increase in EPP above 5.8 (Chen et al., 1983). In addition, irrigation with water containing large concentrations of K ($\sim 120 \text{ mg L}^{-1}$) led to a decrease of 80% in the permeability of a sandy loam (Peacock, 2007). Shainberg et al. (1987) suggested that the swelling of K-smectites depends on the charge density of the clay (i.e., the ratio of clay CEC to its specific surface area; the lower the charge density, the greater the swelling). The charge density of montmorillonitic clay can vary between 1.05 and $2.25 \text{ mol}_c \text{ m}^{-2} \times 10^{-6}$. A value greater than 1.3 has been considered to represent a clay with a high charge density (Shainberg et al., 1987). More recently it has been suggested that these conflicting results with respect to the effects of exchangeable K on soil physical and hydraulic properties could, at least in part, be ascribed to differences in soil mineralogy. In predominantly kaolinitic or illitic soils, the effects of K are likely to be either similar to those of Na or intermediate between those of Na and Ca, while in predominantly smectitic soils the effects of K and Ca are likely to be comparable.

Potassium in plant nutrition

Potassium uptake is highly selective and closely coupled to metabolic activity (Hawkesford et al., 2012 and references cited therein). It is characterized by high mobility in plants at all levels – within individual cells and tissues, also in long-distance transport via the xylem and phloem. Uptake and transport of K throughout the plant is facilitated by integral membrane proteins (transporters and cation channels) which enable its movement across the plasma membrane. Potassium is the most abundant cation in the cytosol, and K and its accompanying anions contribute substantially to the osmotic potential of cells and tissues of glycophytic plant species. For various reasons, K has an outstanding role in plant–water relations (Hawkesford et al., 2012 and references cited therein). Potassium is not metabolized and it forms only weak complexes in which it is readily exchangeable (Hawkesford et al., 2012 and references cited therein). Therefore, K does not compete strongly for binding sites of divalent cations (e.g., Mg). On the other hand, due to its large concentrations in the cytosol and chloroplasts, it balances the charge of soluble (e.g., organic acid anions and inorganic anions) and insoluble anions and thus facilitates the stabilizing of pH between 7.0 and 8.0 in these compartments, which is the optimum for most enzyme reactions.

Potassium availability

Estimation of K availability in soils is based on the concept of quantity (Q) and intensity (I). A detailed description of this concept with regard to K availability in soils can be found in numerous reviews (e.g., Huang et al., 2005 and references cited therein). In brief, the quantity (Q) of labile K in soil for plant consumption has been defined as the exchangeable K. A linear relationship between K uptake by plants and the quantity of exchangeable K was obtained in many pot and field studies. However, in other studies this relationship did not hold for some types of soil, in which an intensity (I) factor was found more useful. The intensity (I) of K in a soil at equilibrium with its solution could best be defined by the ratio $a_k/(a_{Ca}+a_{Mg})^{1/2}$ (where a is ion activity) of the soil solution. This equilibrium activity ratio for K, referred to as AR^K , is analogous to the Gapon equation which was developed for Na exchange in soils with the divalent cations, Ca and Mg. The AR^K was suggested by Woodruff (1955) as a measure of K availability. It has been proposed that the activity of K in the soil solution is of primary importance as it is central for K diffusion and K translocation in soils by mass flow to roots (Barber, 1995). Woodruff (1955) suggested a general energy expression of the intensity factor using a natural logarithmic transformation of AR^K multiplied by the Boltzmann constant (R) and temperature ($T, ^\circ K$).

$$\Delta F = RT \ln(AR^K) \quad (1)$$

The larger are the values of labile K ($|\Delta F|$), the greater is the release of K into the soil solution, resulting in a larger pool of labile K.

Strong correlation ($r = 0.80$) between the (I) and (Q) parameters for available K were obtained in pot and field studies when the same soil or soils with similar properties were investigated under different management practices (Fig. 3). However, as can be seen from the data of Feigenbaum and Hagin (1967), soils with the same value of AR^K may have a wide range of capacity (Q) for maintaining AR^K under cultivation (Fig. 3). Therefore, other researchers (e.g., Beckett et al., 1964a,b in Sparks and Huang, 1985) suggested use of the classic Q/I relationships where the activity ratio AR^K value is related to the change in exchangeable K to obtain the effect of the quantity of labile K (exchangeable K) on the intensity of labile K (AR^K). In general, it appears that the Q/I concept is a good measure of the availability of the more labile K in soils to plants. The method used for constructing a typical Q/I curve involves equilibrating a soil with solutions containing a constant concentration of $CaCl_2$ (calcium chloride) and increasing the concentration of KCl (potassium chloride). The AR^K values are then plotted versus the gain or loss of K to form the characteristic Q/I curve (Sparks and Liebhardt, 1981) (Fig. 4).

Several parameters can be obtained from the Q/I plot to characterize soil K status. The value of AR^K when the Q factor (ΔK) equals zero is a measure of the degree of K availability at equilibrium or AR_c^K . The value of ΔK when AR^K equals zero is a measure of labile or exchangeable K in soils or ΔK^0 . The slope of the linear portion of the curve gives the potential buffering capacity of K (PBC^K) and is proportional to the cation exchange capacity of the soil.

The PBC^K indicates the ability of a soil to maintain the intensity of K in the soil solution. A large PBC^K value indicates constant availability of K in the soil over a long period. By contrast, a small PBC^K value would indicate the need for frequent K fertilization. Cropping, fertilization and application of soil amendments can change the PBC^K values of soils (Sparks and Huang, 1985).

Fertilizer recommendations and effects on crops and soil

Fertilizer recommendations for K are based on two groups of methods: soil tests and plants data (Brouder et al., 2021). Various soil tests have been developed for recommendations of K fertilizer based on the estimation of exchangeable K as the Q factor or AR^K as

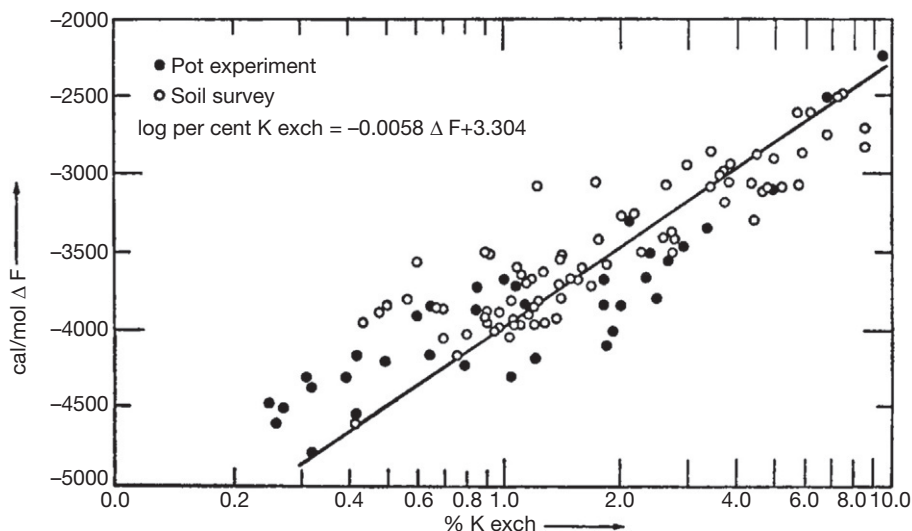


Fig. 3 Relationship between ΔF values and exchangeable K expressed as a percentage of soil cation exchange capacity. From Feigenbaum S and Hagin J (1967) Evaluation of methods for determining available soil potassium based on potassium uptake by plants. *Journal of Soil Science* 18: 197–203.

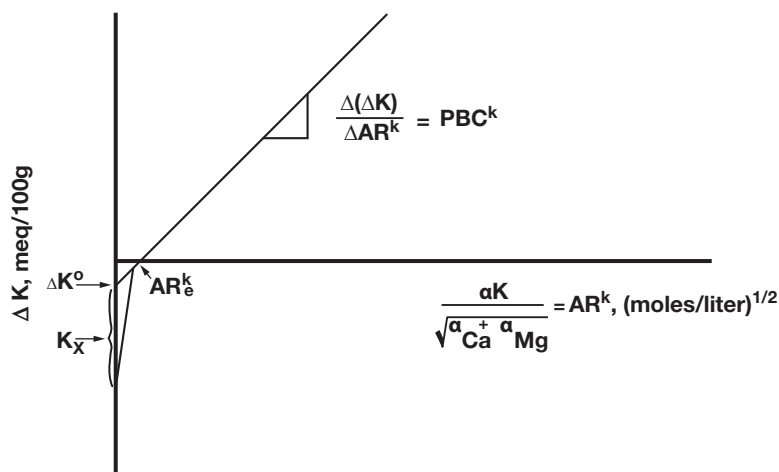


Fig. 4 Typical quantity/intensity (Q/I) plot. Reproduced with permission from Sparks DL and Liebhardt WC (1981) Effect of long-term lime and potassium application on quantity-intensity (Q/I) relationship in sandy soil. *Soil Science Society of America Journal* 45:786–790.

the intensity (I) factor. Although there were strong correlations between both types of parameters and K uptake by plants, in some soils there were many cases where no correlations were obtained. In some soils, in which non-exchangeable K contributes considerably to K uptake by plants, a combined estimate of both sources of K is used. However, the estimate of the relative contribution of non-exchangeable K is not straightforward and depends on complex factors including plant species, management, climate and soil moisture. Another limitation of both Q and I factors is that they signify whether the status of K in the soil, determined at a specific time, is above or below the threshold for fertilization, yet Q and I do not indicate the dose of K fertilizer that should be applied. The Q/I concept described above can be used to predict change in the amount of Q by change in I. A new method based on the Q/I concept was recently developed and tested in pot experiments with various soil types from Israel representing a wide range of properties including clay and lime contents, cation exchange capacity and specific surface area (Shenker and Seth, 2018). However, this method still needs to be tested in field experiments. It requires some additional data such as the optimal K consumption of a given crop in a specific region for an expected yield and the volume of soil supplying K. The main limitation of this method is economic because it requires measurements of exchangeable and soluble K, Ca and Mg at a wide range of K content in soil. Brouder et al. (2021) discuss in depth the gaps in knowledge for improving K fertilizer recommendations. They also discuss potential and required research for the use of mechanistic models of K release, exchange, movement in soil and uptake by plants, and machine learning models based on experimental data.

When reliable soil tests are not available and when there are no well-proved K fertilizer recommendations based on soil tests, there is the option to use recommendations based on plant data. Such recommendations depend on the assumption that K concentrations in soils should be maintained by a K fertilizer rate that replaces K removed by crop harvest. Long-term K fertilization affects clay mineralogy, including the transformation of Illitic clay to montmorillonite under little or no K fertilizer application and vice versa with large K fertilizer application (Sandler et al., 2009 in Bar-Yosef et al., 2015).

Plants' response to K supply or deficiency has been shown in many pot experiments and those of nutrient solutions. An example of the response of olive (*Olea europaea* L.) plants to K fertilizer application in a pot experiment of Erel et al. (2015) is shown in Fig. 5. However, the lack of response to K fertilizer in many field studies and surveys has raised the question of whether fertilization with K is beneficial for crop production (Bar-Yosef et al., 2015). Khan et al. (2014 in Bar-Yosef et al., 2015) presented evidence that soil exchangeable K increases over time without the addition of K to the soil despite its removal by crops in montmorillonitic and illitic soils. The authors termed this behavior 'The potassium paradox'. They also reported a lack of crop response to KCl fertilization. Bar-Yosef et al. (2015) showed that the observations of Khan et al. (2014 in Bar-Yosef et al., 2015) can be interpreted and predicted using current knowledge of K reactions in soil and its uptake by plants, and that there is no paradox in K behavior in the soil–plant system (for more details see Bar-Yosef et al., 2015). Moreover, Mueller et al. (2012) estimated that there should be an increase of 34% in global K fertilization to close global yield gaps to 75% of attainable yields, while eliminating overuse of inputs that may harm the environment.

Sodium in soils

A chapter devoted to Na-affected soils (Sodic Soils) is included in the Encyclopedia (Chapter XX). Readers interested in the effects of sodicity on soil chemo-physical, physical and hydraulic properties are referred to this chapter. The current chapter will focus mostly on the chemical and fertility effects of Na on soils.



Fig. 5 The response of olive plants to 6 months of exposure to irrigation solutions containing: 2.48 mM K and 2.84 mM Na (+K), or 0.08 mM K and 2.48 mM Na (+Na) or 0.09 mM K and 0.13 mM Na (–K –Na). Whole plant response and leaves. Upper row, adopted from Erel R, Ben Gal A, Dag A, Schwartz A, and Yermiyahu U (2014) Sodium replacement of potassium in physiological processes of olive trees (var. Barnea) as affected by drought. *Tree Physiology* 34: 1102–1117; lower row, Erel R, Yermiyahu U, Ben-Gal A, Dag A, Shapira O, and Schwartz A (2015) Modification of non-stomatal limitation and photoprotection due to K and Na nutrition of olive trees. *Journal of Plant Physiology* 177: 1–10.

Sodium is, in general, the dominant alkali cation in soils, especially in salt affected ones. Data compiled by Wansik et al. (2016) indicate that Na-affected soils cover 581 million ha worldwide, with the largest areas of Na-affected soils concentrated in Australia (340 million ha) followed by north-central Asia (120.1 million ha) South America (59.6 million ha), Africa (27 million ha), Europe (22.3 million ha) and North America (9.6 million ha). This spatial distribution of Na-affected soils demonstrates that they occur under a range of climates.

Commonly, the content of Na relative to that of the divalent cations (Ca and Mg) in the soil exchange complex and or the soil solution determines whether a soil is affected by Na or not (see Chapter XX). Yet, soil sensitivity to Na can also be evaluated from Na concentration in the soil solution relative to that of the main anions in the solution (Fig. 6). The data presented in this Figure indicate that once Na concentration is higher than or equal to that of the combined concentration of Cl^- and SO_4^{2-} , then the soil is adversely affected by Na.

Sodium and soil productivity

The productivity of sodic soils is limited, mainly because of the deterioration in their physical properties, which causes poor aeration, restricts root development and enhances root diseases. However, the impact of Na on soil productivity may also depend on crop tolerance to Na per se and plant nutrition in Na-affected soils. A detailed review on nutrient constraints in Na-affected soils is presented in Qadir and Schubert (2002).

Plant species are characterized as ‘natrophilic’ or ‘natrophobic,’ depending on their growth response to Na and their capacity to take up Na by roots and transport it to shoots (e.g., Phillips et al., 2000). The role of Na in plant nutrition in different species is: (i) essential for certain plant species, (ii) replacing K functions in plants at different concentrations, and (iii) a beneficial element enhancing growth (Broadley et al., 2012). Growth of many halophytes, whether C3 or C4 species, is enhanced by large Na concentrations in the substrate (generally, 10–100 mM Na, but up to 510 mM Na in extreme cases; Redondo-Gomez et al., 2010). However, the majority of commercial crops are defined as non-halophytes (glycophytes), therefore such large Na

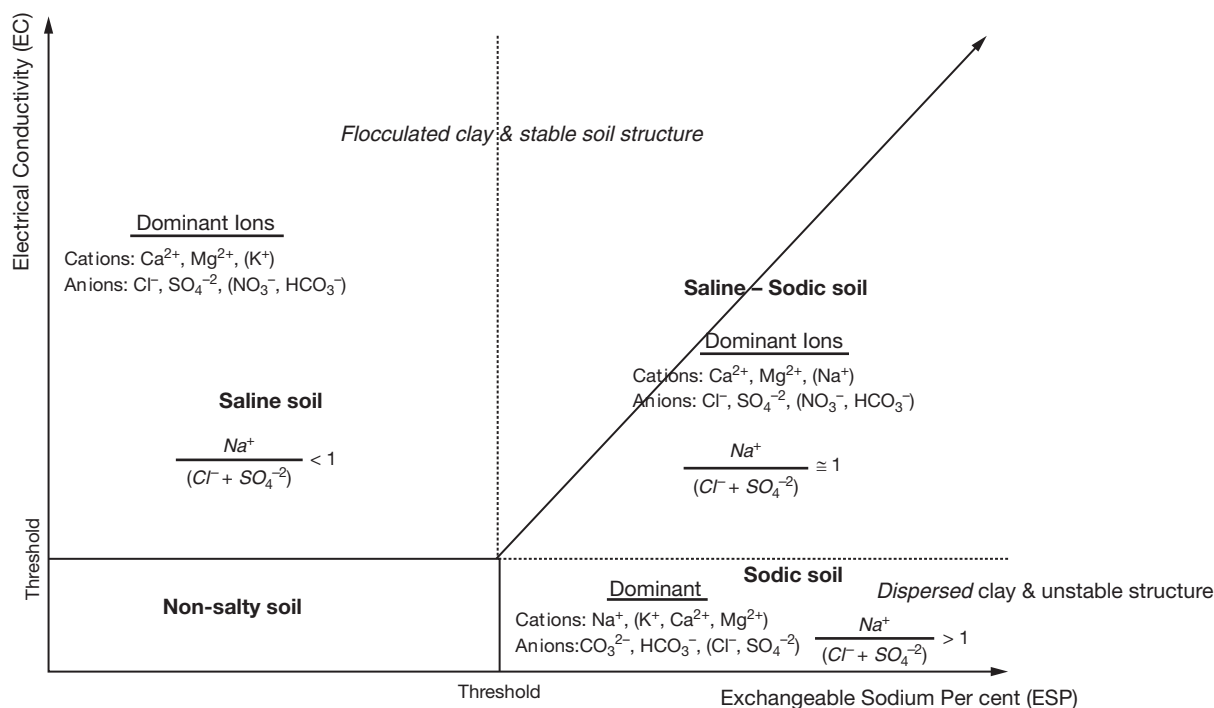


Fig. 6 Graphic illustration of Na-affected soils (below the solid arrow and to the right of the solid part of the vertical and horizontal threshold line) vs. non Na-affected soils (left of the vertical threshold line). From Mapping Salt-affected soils. Technical Manual. FAO. Roma. 2020.

concentrations are not required. Sodium is not essential for C3 species but it is an essential micro-nutrient for many, but not all C4 species. In general, the concentration of Na in the soil solution of the majority of agricultural soils is much higher than that required for C4 species. Sodium was found as a proxy for K and can replace it under low K availability in several herbaceous plants (Kronzucker et al., 2013) and fruit trees (e.g. olives, Erel et al., 2014, Fig. 5).

Large Na concentrations were found to be toxic to many glycophytes plants, where the predominant effect of the co-presence of Na^+ in K^+ -containing media is one of affecting K^+ homeostasis negatively. More detailed description of the various aspects of sodium toxicity are presented in the review by Kronzucker et al. (2013). Typical symptoms of sodium toxicity are necrotic spotting in the leaves, followed by the development of a marginal necrotic band, as shown for an olive tree in Fig. 7.

Some research has indicated that most field crops are classified as moderately tolerant or tolerant to Na; yet some crops such as deciduous fruits, nuts (Juglandaceae), citrus and avocado (*Persea Americana* Mill.) may show signs of Na toxicity (Pearson, 1960). It should be born in mind, however, that ranking of crops for Na tolerance and nutrient constraints strongly depends on the environmental conditions under which the response was determined.



Fig. 7 Sodium toxicity symptoms in the tips of olive leaves in which Na concentration was 4 mg g^{-1} , compared with a threshold value of 2 mg g^{-1} . Courtesy of Dr. Ran Erel and Prof. Uri Yermiyahu.

Chemical problems in sodic soils relating to plant nutrition may range from deficiency of several nutrients to toxicity of Na. Plant growth in sodic soils is usually influenced by the osmotic and specific-ion effects and ionic imbalance leading to deficiency and or toxicity of some nutrients. In sodic soils, concentrations of Na often exceed those of most macronutrients by one or two orders of magnitude, and by even more in the case of micronutrients. Thus, sodic soil solutions may have depressed nutrient-ion activities and extreme ratios of Na/Ca, Na/K, Na/Mg. As a result, the salt- and Na-stressed plants become susceptible to high osmotic stress, specific ion toxicity and nutritional disorders. A collective effect of such stresses and disorders may reduce crop yield and quality (Grattan and Grieve, 1992). This depends upon the ambient sodicity concentrations, composition of the soil solution and exchange complex, soil pH and redox potential, the particular nutrient in question, salinity or sodicity resistance capability and nutrient requirement of the plant species grown, and several environmental factors. In addition, poor physical properties of sodic soils, which directly limit crop growth through poor seedling emergence and root growth, also exhibit indirect effects on plant nutrition by restricting water and nutrient uptake and gaseous exchange. Because sodic soils are usually hard, particularly in the subsoil, restricted root proliferation in such soils reduces water and nutrient supply (Qadir and Schubert, 2002).

Sodium and carbon (C) sequestration

Sodic soils have lost a large fraction of their original C pool. The magnitude of the loss may range between 10 and 30 Mg C ha⁻¹, depending on the antecedent pool and the severity of degradation. It is expected that the significance of C dynamics, as influenced by Na-related soil degradation, is expected to increase in future, as the extent of sodification is projected to expand considerably, especially in dryland areas (NLWRA, 2001).

Sodic soils compound soil organic carbon (SOC) loss by enhancing soil structure degradation and aggregate dispersion that on one hand increase SOC mineralization, yet on the other, restrict access to the substrate for mineralization by increased bulk density. Cycles of wetting and drying in saline-sodic cultivated soils play an important role in soil organic matter (SOM) decomposition. During wetting, aggregate dispersion enhances SOM decomposition, while drying leads to flocculation of aggregates and protection of SOM (Wong et al., 2010). In general, SOM-derived C stocks and fluxes in Na-affected soils are directly related to decreased plant inputs due to low biomass production and, hence, limited accumulation of soil organic matter that in turn affects soil C and nutrient cycling.

Sodic soils commonly occur in arid and semi-arid regions where they generally contain large pools of primary (lithogenic) and secondary (pedogenic) inorganic carbonates. Dissolution of the carbonates in the root zone could result from the accumulation of high $p\text{CO}_2$ (root respiration and decomposition of organic matter) or by acidic root exudates. The resulting HCO_3^- (bicarbonate) can then be leached, especially under irrigation management, down the soil profile and its subsequent precipitation is a possible pathway for C sequestration in an inorganic form. It is estimated that C sequestration through this mechanism ranges between 0.25 and 1.0 Mg C ha⁻¹ year⁻¹. Conversely, the presence of carbonates (CaCO_3) may limit SOM decomposition by bridging Ca ions to SOM particles, thus protecting SOM from microbial degradation (Golchin et al., 1994, in Wong et al., 2010). Yet, the subsequent turnover of the decomposition products under such circumstances is expected to be slower than the initial decomposition, thus leading to larger C concentrations and a longer retention time (Baldock and Nelson, 2000, in Wong et al., 2010). It has been concluded that a comprehensive understanding of the role of inorganic C in SOC dynamics, especially in sodic soils, is yet to be reached (Wong et al., 2010).

Information about the impact of the amelioration of sodic soils on C dynamics and stocks is meager and inconclusive. Moreover, it has been argued that the accumulation of organic matter in sodic soils might be difficult due to the high solubility of Na-organic linkages that are susceptible to removal by runoff or percolating water. Yet, revegetation of scalded lands with pasture has successfully increased SOM to the level of that of a native soil within 10 years (Wong et al., 2010). Use of leguminous trees has also been shown to increase SOM in Na-affected soils (Mishra and Sharma, 2003 in Wong et al., 2010). In addition, non-organic amendments such as gypsum have recently been found to be effective in alleviating sodicity stress, and increasing SOM pools (18% to 35%) and their stability in the Trans Indo-Gangetic Plains of India (Basak et al., 2021).

Finally, following the continued increase in sodicity related degradation of agricultural soils and the subsequent depletion in SOM, Wong et al. (2010) have proposed the following two key issues that are central to C dynamics and stock in soils:

1. The role of SOC dynamics in sodic soils for the assessment of C turnover.
2. Evaluation of the role of rehabilitation of sodic soils in the C cycle and its accumulation.

Concluding comments

The alkali metals group comprises six monovalent cations including K and Na that play a central role in agricultural soils. The importance of K arises from its being an important nutrient for plants. The potassium ion in the soil solution is the most available form for plants, yet just 1 to 2% of soil K is in this pool. Over 90% of soil K is within the crystal structure of soil minerals and in the fixed (non-exchangeable) pool found in the mica group of secondary minerals. Albeit defined as unavailable to plants, these two pools affect long term soil K fertility serving as large pools of K that replenish its depleted labile pools. Determination of its availability for plants is, therefore, central for economic K fertilization, supplemented with both soil tests and plant analyses for

actual fertilizer recommendations. Sodium is, in general, the dominant alkali cation in soils; its accumulation in soils often results in degradation of soil structure and associated physicochemical, physical and hydraulic properties. Productivity in Na-affected (sodic) soils is limited, also by crop tolerance to Na per se and plant nutrition in Na-affected soils. Sodicty related degradation of soil structure and associated properties subsequently leads to soil organic C depletion. Conversely, the rehabilitation of sodic soils may positively affect the C cycle and its accumulation in soils.

References

- Arienzo M, Christen EW, Jayawardane NS, and Quayle WC (2012) The relative effects of sodium and potassium on soil hydraulic conductivity and implications for winery wastewater management. *Geoderma* 173-174: 303–310.
- Audi G, Bersillon O, Blachot J, and Wapstra AH (2003) The NUBASE evaluation of nuclear and decay properties. *Nuclear Physics A* 729: 3–128.
- Barber SA (1995) *Soil Nutrient Bioavailability: a Mechanistic Approach*, 2nd edn. New York: John Wiley.
- Bar-Tal A, Eick MJ, Feigenbaum S, Sparks DL, and Fishman S (1995) Determination of rate coefficients for potassium-calcium exchange on vermiculite using a stirred-flow chamber. *Soil Science Society of America Journal* 59: 760–765.
- Bar-Yosef B, Magen H, Johnston AE, and Kirkby EA (2015) Potassium fertilization: paradox or K management dilemma? *Renewable Agriculture and Food Systems* 30(2): 115–119.
- Basak N, Sheoran P, Sharma R, Yadav RK, Singh RK, Kumar S, Krishnamurthy T, and Sharma PC (2021) Gypsum and pressmud amelioration improve soil organic carbon storage and stability in sodic agroecosystems. *Land Degradation and Development* 32: 4430–4444.
- Broadley M, Brown P, Cakmak I, Ma JF, Rengel Z, and Zhao F (2012) Beneficial elements. In: Marschner P (ed.) *Marschner's Mineral Nutrition of Higher Plants*, 3rd edn. Amsterdam, Netherlands: Academic Press, an imprint of Elsevier.
- Bronick CJ and Lal R (2005) Soil structure and management: A review. *Geoderma* 124: 3–22.
- Brouder SM, Volenec JJ, and Murrell TS (2021) The potassium cycle and its relationship to recommendation development. In: Murrell TS, Mikkelsen RL, Sulewski G, Norton R, and Thompson ML (eds.) *Improving Potassium Recommendations for Agricultural Crops*, pp. 1–46. Cham, Switzerland: Springer. ch. 1.
- Chen Y, Banin A, and Borochoy A (1983) Effect of potassium on soil structure in relation to hydraulic conductivity. *Geoderma* 30: 135–147.
- Erel R, Ben Gal A, Dag A, Schwartz A, and Yermiyahu U (2014) Sodium replacement of potassium in physiological processes of olive trees (var. Barnea) as affected by drought. *Tree Physiology* 34: 1102–1117.
- Erel R, Yermiyahu U, Ben-Gal A, Dag A, Shapira O, and Schwartz A (2015) Modification of non-stomatal limitation and photoprotection due to K and Na nutrition of olive trees. *Journal of Plant Physiology* 177: 1–10.
- Feigenbaum S and Hagin J (1967) Evaluation of methods for determining available soil potassium based on potassium uptake by plants. *Journal of Soil Science* 18: 197–203.
- Grattan SR and Grieve CM (1992) Mineral element acquisition and growth response of plants grown in saline environments. *Agriculture, Ecosystems and Environment* 38: 275–300.
- Hawkesford M, Horst W, Kichey T, Lambers H, Schjoerring J, Möller I, and White P (2012) Chapter 6, Functions of macronutrients. In: Marschner P and 3rd edn (eds.) *Marschner's Mineral Nutrition of Higher Plants*. Amsterdam, Netherlands: Academic Press, an imprint of Elsevier.
- Huang PM, Zhu JM, Xie C, and Wang MK (2005) Potassium in soils. In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*, pp. 303–314. Academic Press.
- Kirkby E (2012) Introduction, definition and classification of nutrients. In: Marschner P (ed.) *Marschner's Mineral Nutrition of Higher Plants*, 3rd edn Amsterdam, Netherlands: Academic Press, an imprint of Elsevier.
- Krebs RE (2006) *The History and Use of Our Earth's Chemical Elements: A Reference Guide*. Westport, Conn: Greenwood Press.
- Kronzucker HJ, Coskun D, Schulze LM, Wong JR, and Britto DT (2013) Sodium as nutrient and toxicant. *Plant and Soil* 369: 1–23.
- Levy GJ and Torrento JR (1995) Clay dispersion and macro-aggregate stability as affected by exchangeable potassium and sodium. *Soil Science* 160: 352–358.
- Mueller ND, Gerber JS, Johnston M, Ray DK, and Ramankutty N (2012) Closing yield gaps through nutrient and water management. *Nature* 490: 254–257.
- Négrel P, Ladenberger A, Reimann C, Birke M, Sadeghi M, and the GEMAS Project Team (2018) Distribution of Rb, Ga and Cs in agricultural land soils at European continental scale (GEMAS): Implications for weathering conditions and provenance. *Chemical Geology* 479: 188–203.
- NLWRA (2001) *Australians and natural resource management*. <http://www.nlwra.gov.au/>. Canberra: National Land and Water Resources Audit.
- Peacock B (2007) Potassium in soils and grapevine nutrition. <http://cetulare.ucdavis.edu/pubgrape/ng999.htm>.
- Pearson GA (1960) Tolerance of crops to exchangeable sodium. USDA Inf. *Bulletin* 216.
- Phillips CJC, Chiy PC, Arney DR, and Kart O (2000) Effects of sodium fertilizers and supplements on milk production and mammary gland health. *The Journal of Dairy Research* 67: 1–12.
- Qadir M and Schubert S (2002) Degradation processes and nutrient constraints in sodic soils. *Land Degradation and Development* 13: 275–294.
- Redondo-Gomez S, Mateos-Naranjo E, Figueroa ME, and Davy AJ (2010) Salt stimulation of growth and photosynthesis in an extreme halophyte, *Arthrocnemum macrostachyum*. *Plant Biology* 12: 79–87.
- Rich CI (1968) Mineralogy of soil potassium. In: Kilmer VJ, Younts SE, and Brady NC (eds.) *The Role of Potassium in Agriculture*. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America.
- Schmitt A-N, Vigier N, Lemarchand D, Millot R, Stille P, and Chabaux F (2012) Processes controlling the stable isotope composition of Li, B, Mg and Ca in plants, soils and water: A review. *Comptes Rendus Geoscience* 344: 704–722.
- Scott AD and Smith SJ (1987) Sources, amounts and forms of Alkali elements in soils. In: Stewart BA (ed.) *Advances in Soil Science*. vol. 6, pp. 101–147. Berlin, Heidelberg, London, Paris, Tokyo: Springer-Verlag. NY.
- Shainberg I, Alperovitch N, and Keren R (1987) Charge density and Na-K-Ca exchange on smectites. *Clays and Clay Minerals* 35: 68–73.
- Shenker M and Seth A (2018) Potassium availability in soils and the use of the Q/I approach – moving from theory to nation-wide realization. *Geophysical Research Abstracts* 20. EGU2018-9716-2.
- Sparks DL and Huang PM (1985) Physical chemistry of soil potassium. In: Munson RD (ed.) *Potassium in Agriculture*, pp. 201–276. Madison, WI: American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America. ch. 9.
- Sparks DL and Liebhart WC (1981) Effect of long-term lime and potassium application on quantity-intensity (Q/I) relationship in sandy soil. *Soil Science Society of America Journal* 45: 786–790.
- USSL Staff (1954) Diagnosis and improvement of saline and alkali soils. *The Handbook of Agriculture*. vol. 60. Washington, DC: USDA.
- Wansik S, Siddikee MA, Joe MM, Benson A, Kim K, Selvakumar G, Kang Y, Jeon S, Samaddar S, Chatterjee P, Walitang D, Chanratana M, and Sa T (2016) Halotolerant plant growth promoting bacteria mediated salinity stress amelioration in plants. *Korean Journal of Soil Science and Fertilizer* 49: 355–367.
- Wong VNL, Greene RSB, Dalal RC, and Murphy BW (2010) Soil carbon dynamics in saline and sodic soils: A review. *Soil Use and Management* 26: 2–11.
- Woodruff CM (1955) The energies of replacement of calcium by potassium in soils. *Soil Science Society of America Journal* 19: 167–171.

Alkaline earth metals in soil

Justin B Richardson, Department of Geosciences, University of Massachusetts Amherst, Amherst, MA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction to alkaline earth metals	126
Speciation, dissolution, and precipitation in solution	127
Sources to soils: Rocks and minerals	128
Sources to soils: Atmospheric deposition	129
Partitioning and measurement of alkaline earth metals in soils	129
Climate, parent material, and time control on alkaline earth metals in soils	130
Plant requirements, uptake, and deficiencies for Ca and Mg	131
Ca and Mg soil amendments	132
Soil beryllium, strontium, barium, and radium toxicity to plants and animals	132
Alkaline earth metals as environmental tracers	133
Conclusion	133
References	134

Abstract

The alkaline earth metals are Group 2 on the Periodic Table of the elements and are abundant in the Earth's crust and soils. They have +2 charge, minimal oxidation–reduction reactions, and moderate to high solubility in soils. Alkaline earth metals are abundant across the aluminosilicate, carbonate, and sulfate minerals. Strontium and barium are present in calcium-bearing minerals due to isomorphic substitution. Radium is predominately sourced from uranium-bearing rocks and minerals. The abundance and biogeochemical cycling of alkaline earth metals in soils is controlled by climate, parent material, and time as they control chemical weathering, precipitation reactions, type of rocks and minerals present, and leaching.

Key points

- Alkaline earth metals are Group 2 elements with +2 charge and minimal oxidation–reduction reactions and moderate to high solubility.
- Alkaline earth metals are released to soils by mineral dissolution and their speciation in soils depends on precipitation and adsorption reactions.
- Magnesium (Mg) and calcium (Ca) are essential for plants and animals while Be (beryllium), Sr (strontium), Ba (barium), and Ra (radium) do not serve essential functions and may be toxic at elevated concentrations.
- Magnesium and Ca can be added to soils through fertilizers to address deficiencies that reduce plant growth.
- Ratios of alkaline earth metals and their stable isotopes can be used as environmental tracers.

Introduction to alkaline earth metals

The alkaline earth metals comprise six metals in Group 2 (previously CAS group IIA) in the Periodic Table of the elements. Atomic symbols and atomic numbers for the alkaline earth metals are beryllium (Be) 4, magnesium (Mg) 12, calcium (Ca) 20, strontium (Sr) 38, barium (Ba) 56, and radium (Ra) 88. Since alkaline earth metals are lithophilic (rock-loving) elements and not siderophiles (iron (Fe)-loving), they are widely abundant in the Earth's crust and the resultant soils (White, 2020). Calcium and Mg are the fifth and eighth most abundant elements in the Earth's crust. Strontium and Ba are the 14th and 15th most abundant elements and are more abundant in the Earth's crust than other elements of similar atomic masses. The overall abundance of Mg, Ca, Sr, and Ba on Earth is due to their formation from nucleosynthesis through exploding massive stars and low mass stars (White, 2020). Conversely, Be and Ra are the 48th and 85th most abundant elements and are less abundant than other elements with similar atomic masses. Radium is rare due to its radioactive decay, with the longest living radioisotope being ^{226}Ra which has a half-life of only 1600 years and is produced from Uranium decay. Beryllium is also comparatively rarer than other light elements and is formed from early galaxy cosmic ray fission to form ^9Be or cosmic ray spallation on Earth to form ^{10}Be .

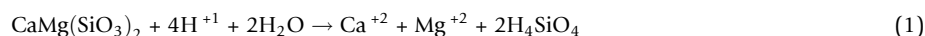
The alkaline earth metals have similar chemical and physical properties, largely due to their empty s-orbital, giving rise to +2 charge and minimal oxidation–reduction reactions at temperatures and pressures in terrestrial environments on Earth. Standard atomic weights increase with atomic number: Be 9.012182 g mol⁻¹, Mg 24.3050 g mol⁻¹, Ca 40.078 g mol⁻¹, Sr 87.62 g mol⁻¹, Ba 137.327 g mol⁻¹, and Ra 226 g mol⁻¹. Densities for the alkaline earth metals decrease from 1.85 g cm⁻³ for Be to 1.54 g cm⁻³ for Ca but increase from 2.64 g cm⁻³ for Sr to 5.5 g cm⁻³ for Ra. Ionic radius increases from 59 pm for Be⁺² to 253 pm for Ba⁺². Because of

this increase in atomic size with similar +2 charge, electronegativity decreases from 1.57 for Be to 0.9 for Ba and Ra on the revised electronegativity Pauling scale. Once hydrated, the Debye–Huckel effective radii range from ~500 pm for Sr^{+2} and Ba^{+2} to 800 pm for Mg^{+2} and Be^{+2} due to their hydration spheres (Garrels and Christ, 1965).

Speciation, dissolution, and precipitation in solution

The alkaline earth metals are characterized by their moderate to high solubility, insensitivity to fluctuations in oxidation–reduction conditions, and limited speciation under the typical pH range of soils. Fig. 1 shows the alkaline earth metals are predominantly at a + 2 charge and do not undergo oxidation or reduction reactions (Takeno, 2005). Of the alkaline earth metals, Mg, Ca, Sr, and Ba typically exist as dissolved ions (Mg^{+2} , Ca^{+2} , Sr^{+2} , Ba^{+2} , Ra^{+2}) in aqueous solution between pH 3 and 8 (Fig. 1). Only Be shows complex speciation and precipitation in the typical pH range of soils. Their large ionic radius and + 2 charge leads to a low charge density that forms weak hydration spheres which do not undergo hydrolysis but allows for ion pairing and complexation (Essington, 2015).

Chemical weathering of primary rocks and minerals is one of the dominant controls on the total abundance, plant availability, and mobility of alkaline earth metals in soils. Chemical weathering and release of alkaline earth metals depends on acid–base reactions, as alkaline earth metals do not undergo oxidation or reduction. As illustrated by Eq. (1), the dissolution of alkaline earth metals from aluminosilicates, in this case Ca and Mg from augite, is driven by the presence of acid. Acidity can be supplied from inorganic sources such as the dissociation of carbonic acid to bicarbonate, but can also be sourced from biotic sources including living organisms (bacteria, fungi, roots) and decomposed organic matter. Aluminosilicate dissolution in soils is considered irreversible to the original mineral due to thermodynamic constraints but, Al (aluminium) and Si (silicon) released can form a new aluminosilicate, such as kaolinite, through incongruent weathering (White, 2020).



Dissolution and precipitation of alkaline earth metals from carbonates is another acid driven reaction with a key difference from aluminosilicate dissolution that the reaction is reversible. Under moist, temperate soil conditions carbonates undergo dissolution while under dry, arid soil conditions carbonates precipitate. For example, Eq. (2) illustrates dolomite containing both Ca and Mg

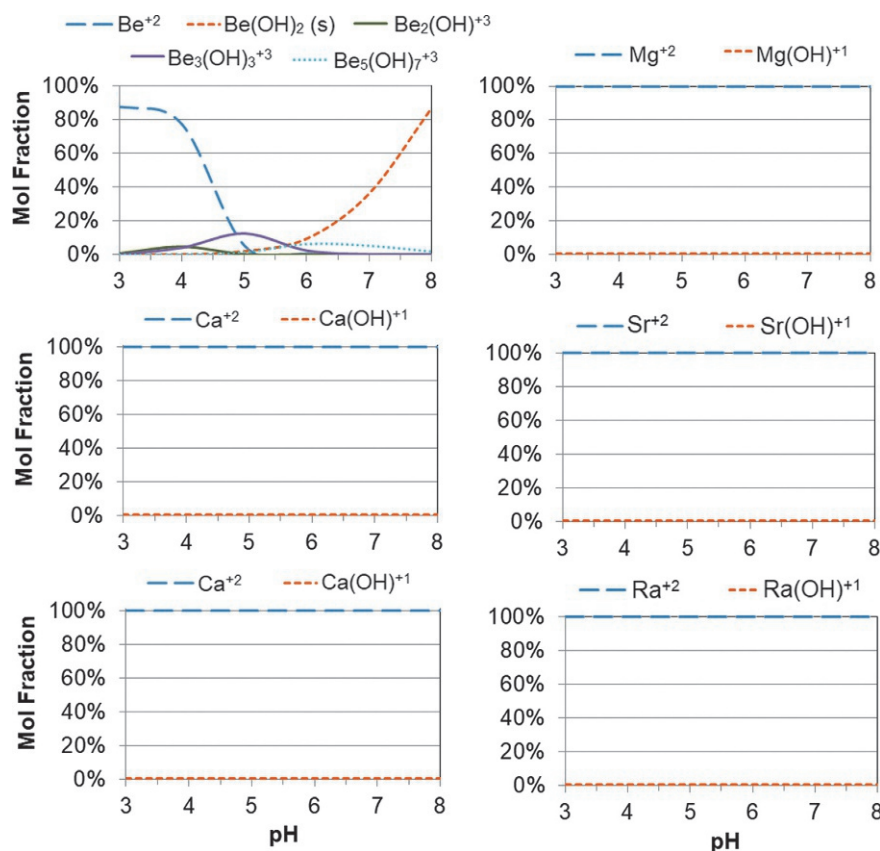
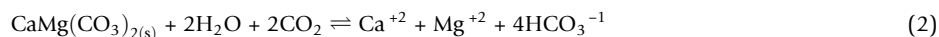


Fig. 1 Speciation of Be, Mg, Ca, Sr, Ba, and Ra under pH range typical for natural waters and soils modeled using Visual MINTEQ at 10 nM 25 °C and 101 kPa.

undergoing dissolution when water and acidity (here carbon dioxide dissolving as carbonic acid) is present. However, the reaction can occur in reverse with precipitation of dolomite under increasing alkalinity in Eq. (2).



The precipitation of alkaline earth metals depends on anions present in solution. While carbonate precipitation is one of the most important at the global scale due to the rising abundance of CO_2 in the atmosphere, other essential anionic compounds and elements should also be considered, particularly: PO_4^{2-} (phosphate), SO_4^{2-} (sulfate), Cl^- (chloride), and F^- (fluoride). The formation of phosphate minerals with alkaline earth metals can be of great importance, as the family of compounds are highly insoluble with solubility constants (K_{sp}) of 1×10^{-24} or smaller. Sulfate compounds, such as gypsum and anhydrite, can also be important in dry environments. Sulfate compounds exhibit some solubility with a K_{sp} of 3.14×10^{-5} (Lide, 2005) and will precipitate if sulfate concentrations are elevated.

Sources to soils: Rocks and minerals

Alkaline earth metals are predominantly sourced to soils from rocks and minerals, particularly as aluminosilicates, carbonate, and sulfate minerals (Table 1). In particular, Mg and Ca are abundant across the five aluminosilicate mineral groups: (1) single tetrahedron nesosilicates (e.g. olivine, forsterite, garnet minerals), (2) double tetrahedron sorosilicates (e.g. epidote), (3) single and double chained inosilicates (e.g. diopside, rhodonite, wollastonite, hornblende), (4) sheet phyllosilicates (e.g. chlorite, talc), and (5) framework tetrahedron tectosilicates (e.g. anorthite). Minerals with the highest concentrations of Ca and Mg are carbonate minerals, specifically calcite. Other high Ca bearing minerals include dolomite and gypsum. Strontium and Ba are also present in Mg- and Ca-bearing aluminosilicate minerals due to isomorphic substitution. There is typically more substitution of Sr and Ba for Ca than for Mg, due to the closer similarity in radii and charge density. Strontium and Ba can be found in high concentrations in sulfate and carbonate minerals, such as celestite (SrSO_4), strontianite (SrCO_3), barite (BaSO_4) and witherite (BaCO_3). Beryllium can be found in Ca- and Mg-bearing aluminosilicate minerals at trace concentrations but is predominately sourced to soils from bertrandite, beryl, and phenakite minerals. Radium exists at trace concentrations in aluminosilicate minerals and is predominately sourced to soils from U-bearing minerals, such as uraninite.

The abundance of alkaline earth metal varies within igneous, metamorphic, and sedimentary rocks due to their geologic provenance (Table 2). Aluminium- and Si-rich felsic igneous rocks (such as granites) and Fe- and Mg-rich igneous rocks (like basalt) can have low to high abundance of alkaline earth metals. Felsic granites and rhyolite igneous rocks can have Ca- and Mg-bearing minerals that include feldspars such as anorthite, amphiboles such as hornblende, and potential inclusion of rhodonite and epidote. Similarly, mafic basalts and gabbro igneous rocks can have Ca- and Mg-bearing minerals such as olivine, forsterite, and augite. Igneous rocks that are dominated by Al, Si, and Na (sodium) bearing minerals can have relatively low amounts of alkaline earth metals while igneous rocks with high proportions of Fe and Mg can have low amounts of Ca and Sr. Sedimentary and metamorphic rocks generally have moderate to high abundance of alkaline earth metals depending on their source material, diagenesis, and the depositional environment they formed in. Most sedimentary rocks that formed in shallow marine environments have a high abundant alkaline earth metals due to biological detritus (e.g. shells of marine invertebrates). Sedimentary rocks formed from clastic sediments (e.g. sand, gravel, clays) can also include Ca- and Mg-bearing minerals in weathering resistant aluminosilicates such as feldspars. Metamorphic rocks that undergo hydrothermal alterations and high temperature and pressure alterations can have mineral assemblages with low to high alkaline earth metals. The abundance of Ra is controlled by the distribution of

Table 1 Examples of concentrations of alkaline earth metals in common Ca- and Mg-bearing minerals.

Minerals	Idealized chemical formula	Be <i>mg kg⁻¹</i>	Ca <i>g kg⁻¹</i>	Mg <i>g kg⁻¹</i>	Sr <i>mg kg⁻¹</i>	Ba <i>mg kg⁻¹</i>
Anorthite ¹	$\text{Ca}(\text{Al}_2\text{Si}_2\text{O}_8)$	0.1	71	48	600	83
Augite ²	$\text{CaMg}(\text{SiO}_3)_2$	—	122	100	15	6
Epidote ³	$\{\text{Ca}_2\{\text{Al}_2\text{Fe}^{3+}\}(\text{Si}_2\text{O}_7)(\text{SiO}_4)\text{O}(\text{OH})$	—	164	1.1	2895	18
Forsterite	Mg_2SiO_4	—	—	—	—	—
Magnesio-Hornblende ²	$\text{Ca}_2(\text{Mg}_4\text{Al})(\text{Si}_7\text{Al})\text{O}_{22}(\text{OH})_2$	—	73	77	53	81
Wollastonite ⁴	CaSiO_3	0.1	328	1.1	1570	5
Chlorite – Clinocllore ³	$\text{Mg}_5\text{Al}(\text{AlSi}_3\text{O}_{10})(\text{OH})_8$	—	0.5	105	33	3
Talc ⁵	$\text{Mg}_3\text{Si}_4\text{O}_{10}(\text{OH})_2$	<0.1	2	181	—	12
Apatite ⁶	$\text{Ca}_5(\text{PO}_4)_3(\text{OH}, \text{F}, \text{Cl})$	—	42	7	350	7
Calcite ⁷	CaCO_3	—	378	13	295	4
Dolomite ⁷	$\text{CaMg}(\text{CO}_3)_2$	—	192	122	33	1
Gypsum ⁸	$\text{CaSO}_4 \cdot 2 \text{H}_2\text{O}$	0.02	162	2	1100	—

Data sources: (1) Bindeman and Bailey (1999); (2) Coint et al. (2013); (3) Xiao et al. (2018); (4) Baudouin and France (2019); Pi-Puig et al. (2020); (6) Sha and Chappell (1999); (7) Vedanand et al. (1993); (8) Maurya and Rai (2020).

Table 2 Examples of alkaline earth metals in common Ca- and Mg-bearing rocks using United State Geological Survey (USGS) rock standards.

<i>Rocks Standard</i>	<i>Be</i> <i>mg kg⁻¹</i>	<i>Ca</i> <i>g kg⁻¹</i>	<i>Mg</i> <i>g kg⁻¹</i>	<i>Sr</i> <i>mg kg⁻¹</i>	<i>Ba</i> <i>mg kg⁻¹</i>
Basalt (BHVO-2)	<1	81	43	389	130
Diabase (W-2A)	1	78	39	190	170
Granodiorite (GSP-2)	2	15	6	240	1340
Mica schist (SDC-1)	3.5	11	10	241	530
Shale (SGR-1)	<1	38	22	610	370
Peridotite (PCC-1)	<1	1	44	0.4	<1

U, which is an incompatible and highly soluble element. Thus, U and Ra are typically associated with coarse-grained granites (e.g. pegmatites) or hydrothermally altered rocks but can also occur in limestone, coarse sedimentary deposits, and faulted volcanic deposits.

Sources to soils: Atmospheric deposition

Alkaline earth metals can be supplied to soils by atmospheric deposition of particulates (e.g. dust), precipitation (rain, snow, fog), and interactions with plants, animals, and humans. The total abundance and composition of alkaline earth metals in atmospheric deposition depends on the source. For example, precipitation in remote areas across the world had Ca and Mg concentrations ranging from <0.1 to 68 $\mu\text{eq L}^{-1}$ and <0.1 to 321 $\mu\text{eq L}^{-1}$, respectively (Galloway et al., 1982). Close proximity to the ocean increases the marine influence and sea spray results in alkaline earth metal enrichment of soils. Volcanic eruptions can generate large deposition events locally during eruptions and be carried by atmospheric winds globally. Further, wind can carry dust and enrich other areas locally and across continents. A key example is the transport of Ca and Mg from the Saharan desert by wind to South America, which serves as an important source to the Ca and Mg poor soils of the Amazonian rainforests (Reichholf, 1986; Boy and Wilcke, 2008).

Plants, animals, and humans can also act as sources of alkaline earth metals to soils. Plants directly affect alkaline earth metals through uptake and cycling back to the soil surface. Plants also can release soluble metals from their leaves by precipitation, known as throughfall and stemflow. In addition, natural and human-caused fires can generate ash and dust that can carry alkaline earth metals to soils. Animals can release alkaline earth metals in their wastes, generating localized hot spots. Humans can also release alkaline earth metals intentionally as fertilizers (discussed later) or unintentionally as pollution from smelting, metallurgy, industrial processing, manufacturing, cement production, construction/demolition, and combustion of fuels (Kabata-Pendias and Mukherjee, 2007). As a prime example, metropolitan areas and industrial areas can have elevated Ca and Mg concentrations, with values ranging from 17 to 1570 $\mu\text{g L}^{-1}$ and 2 to 68 $\mu\text{eq L}^{-1}$, respectively, near the urban center of Chengdu, China and others globally (Wang and Han, 2011).

Partitioning and measurement of alkaline earth metals in soils

Alkaline earth metals are released in soils from primary minerals through chemical and physical weathering. Under physical weathering, alkaline earth metals can be released into soils as small immobile particles or lost from rock outcrops as sediments during mass wasting events. Chemical weathering, however, is one of the most important pathways for the release of alkaline earth metals from primary minerals. Alkaline earth metals are released as dissolved ions that can move with soil water in solution or become temporarily immobile through interaction with solid phase materials. Under dry, arid conditions, alkaline earth metals can precipitate as carbonate minerals, sulfate minerals, or as a salt minerals such as chloride, fluoride, or nitrate. Under moist conditions, alkaline earth metals can adsorb onto mineral surfaces, particularly those with a large surface area, and high charge clay minerals. Adsorption to Al, Fe, and Mn (manganese) oxides is also possible due to their pH-dependent charge, which is negative under moderately acidic to basic conditions. Lastly, alkaline earth metals can adsorb or chelate to solid or dissolved organic matter. Organic matter complexation depends on the functional groups present, molecular size and configuration, and pH-dependent charge. For example, chelation to citric acid can increase overall solubility of Ca and increase rates of Ca weathering into solution while oxalic acid can lead to precipitation of solid Ca-oxalate crystals (Sagoe et al., 1998). Adsorption onto insoluble organic matter can temporarily immobilize and retain alkaline earth metals for later uptake by plants.

The dissolved transport of alkaline earth metals through a soil profile depends on the partitioning coefficient, Kd, which is the ratio of metal adsorbed to soil particles divided by the concentration in the solution phase. The Kd value is an empirical measurement that describes the affinity of a dissolved ion to either adsorb to soil material surfaces or remain in solution and is unique to each soil. The Kd values for alkaline earth metals, as well as other metals, vary with the concentration of metals present in the system. To study the partitioning of metals, typically a single temperature is considered and the changes in Kd with

concentration are described by adsorption isotherms (Essington, 2015). There are several fundamental equations that can describe the slope and shape of adsorption isotherms for the alkaline earth metals for each soil. The samples and slopes rely on the solution pH, types of surfaces present, clay content, organic matter abundance, and temperatures of solutions. Using the K_d and adsorption isotherms, one- and two-dimensional transport rates for alkaline earth metals can be modeled. Alternatively, the production and downward transport of alkaline earth metals can also be modeled with a process-based approach, using thermodynamic and kinetic data for rates of mineral surface weathering, complexation with inorganic surfaces, complexation with organic surfaces, and chelation with ligands. This approach is more robust for application across spatial areas and time scales, but has significant drawbacks such as the demand for boundary conditions, initial conditions, and coefficients for the large number of potential chemical species and interactions.

To quantify alkaline earth metals in the environment, there are several sets of operationally-defined extractions used to estimate the fraction of each alkaline earth metal as a different chemical species. Typical schemes involve one or more of the following: a plant available, exchangeable, or bioavailable fraction, an organic matter bound fraction, a carbonate fraction, a secondary oxide fraction, and a residual fraction. Each fraction is defined with an explicit goal. Plant available, exchangeable, and bioavailable fractions are alkaline earth metals that are readily soluble for uptake by plants and animals. Typical extractions procedures utilize a salt (e.g. NH_4Cl , NH_4 -acetate see Rao et al., 2008) that can exchange for nutrients on cation exchange sites. This is a simple and rudimentary examination of the nutrients most available for plant uptake, but does not necessarily include other fractions that can be taken up or made accessible from rhizosphere processes: acidification, chelate secretion, and microbial symbionts. Measuring exchangeable Ca and Mg is a basic measure needed in agroecosystems. Calcium and Mg saturation of exchange sites contributes to the base cation saturation (which also includes Na and K) of a soil. The base cation saturation is an important measure for the availability of essential plant nutrients, but also a measure of soil acidification and susceptibility to Al toxicity for plants.

The organic matter bound fraction represents alkaline earth metals that may be available for plant uptake or potentially released during further decomposition of the organic matter. Typical extractions utilize a concentrated hydrogen peroxide or other compound to oxidize organic matter to CO_2 (Rao et al., 2008). The carbonate fraction is an estimate of alkaline earth metals precipitated as carbonate phases that may become dissolved under acidification by plant roots or increased amounts of precipitation. The carbonate fraction is measured using a weak acid extraction, specifically acetic acid, to acidify and dissolve carbonates but not dissolve silicate minerals (Rao et al., 2008). The secondary oxide fraction of alkaline earth metals examines adsorption to oxidation–reduction sensitive Fe and Mn surfaces, which can be important in highly weathered soils or soils that undergo periodic flooding. Oxides can be dissolved using reductants such as hydroxylamine, dithionite, or oxalate (Rao et al., 2008). The residual alkaline earth metals represent the remaining concentrations, typically within aluminosilicate minerals. However, alkaline earth metals within silicates such as feldspars or augite are not readily accessible to plants during a single growing season and thus can overestimate the potential uptake by plants. Total alkaline earth metal concentrations can be measured using X-ray fluorescence (Karathanasis and Hajek, 1996) or total digestion by HF-HNO_3 (hydrofluoric and nitric acids) at high temperatures and elevated pressures to dissolve the silicates (See Hossner, 1996). For all the fractions assessed through extractions and digestions, alkaline earth metals can be measured using several analytical instruments, such as the Inductively Coupled Plasma-Optical Emission Spectrometer (ICP-OES), Inductively Coupled Plasma-Mass Spectrometer (ICP-MS) or Atomic Absorption Spectrometer (AAS). Due to the moderate to high solubility of Ca and Mg, home colorimetric tests can be effective qualitative estimates. However, commercial and government laboratories can provide more specific, precise, and accurate measurements.

Climate, parent material, and time control on alkaline earth metals in soils

Soils are the integrated form of many physical, chemical, biological, and anthropogenic actions, but three dominant soil factors are considered to control alkaline earth metals in soils: climate parent material, and time. Climate can control the availability of water, which can drive or limit mineral dissolution and transport of alkaline earth metals from soils (Fig. 2). Under dry climates (aridic, xeric, parts of ustic) where evapotranspiration exceeds precipitation, alkaline earth metals can accumulate as carbonates, largely due to dust inputs and mineral dissolution exceeding leaching losses. For example, dry, aridic conditions of North Africa, Australia, western South America, western North America, and western Asia can lead to soils rich in alkaline earth metals due to limited precipitation and the formation of carbonate subsurface horizons. These dry soils in the US Soil taxonomic system include Aridisols (specifically gypsisols, calcids) and Mollicsols. Under the World Reference Base International soils classification system this includes Calcisols, Gypsisols, Kastanozems, and Chernozems. Excessive alkaline earth metals in soils can drive increases in soil, which lead to immobilization of other essential elements such as P. Under moist climates (parts of ustic, udic) where evapotranspiration does not exceed precipitation, alkaline earth metals are lost from the soil profile because atmospheric inputs and mineral dissolution are exceeded by leaching (Fig. 2). The loss of alkaline earth metals in soils can lead to acidic soil conditions which may cause excessive Al in soils. This is common in tropical and temperate areas of North America, northern Europe, northern Asia, and South America, particularly if intensive weathering occurs without carbonate-bearing parent material.

Parent material can control the abundance of alkaline earth metals present when soil formation begins. Carbonate-rich bedrocks can be found across North America, Africa, Europe, Asia, Australia, and in Brazil. They are typically sedimentary rocks formed in shallow ocean environments and their derivative soils are often rich in alkaline earth metals, particularly Ca. Conversely, soils derived from igneous rocks with high Al and Si content or high Fe and Mg content can have low availability of Ca due to the mineral assemblage present. Soils derived from sedimentary rocks from shallow marine environments typically have high Ca and Mg

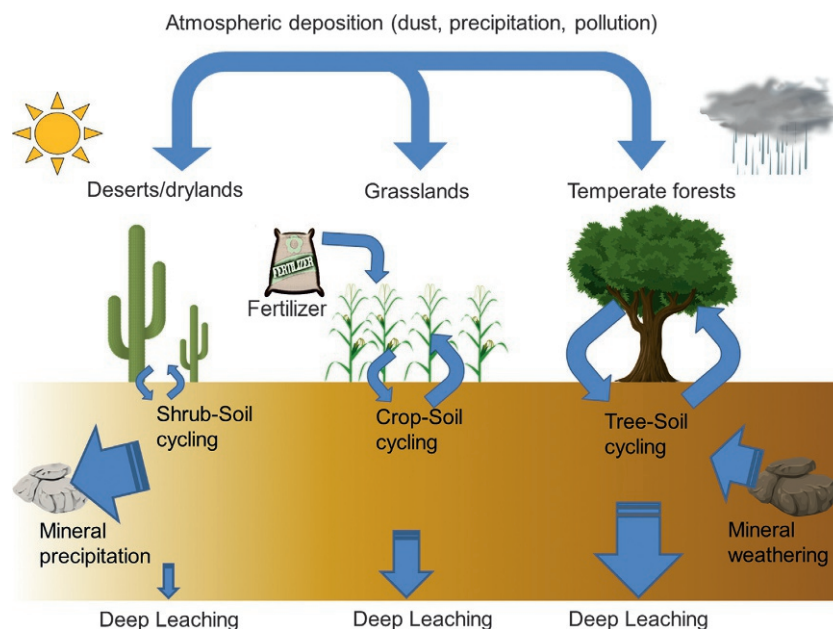


Fig. 2 Conceptual model illustrating soil biogeochemistry of alkaline earth metals across dry ecosystems (desert, drylands), human-altered grassland ecosystems, and forested ecosystems. Arrows are approximations of size of the fluxes; moist soils with trees have greater aboveground cycling and downward leaching rates of alkaline earth metals while dry soils have limited plant-soil cycling and downward leaching rates of alkaline earth metals.

concentrations. Soils derived from metamorphic rocks can have low Ca content, for example soil derived from serpentine, or have low Mg content in soil derived from marble. Other soils with hydrothermally altered metamorphic rocks dominated by chlorite and talc can also affect plant growth due to their low Ca and high Mg concentrations.

Young soils (Entisols, Inceptisols, Mollisols, Alfisols) typically have higher alkaline earth metal contents than older soils (Ultisols, Oxisols). Younger soils still retain many primary minerals that are Ca and Mg bearing while extensive mineral dissolution through time will leave behind only secondary and recalcitrant Al, Si, and Fe bearing minerals. Older soils typically retain alkaline earth metals in their surface soils from biological uplift by vegetation, which is the retention of nutrient or nutrient-like elements through biogeochemical cycling from surface soils to roots to aboveground vegetation. Most soils contain adequate Ca for most plants and only highly leached, low-CEC, acidic soils are likely to cause Ca deficiency in plants. Old soils may be deficient in Ca and Mg, such as Oxisols and Ultisols in tropical regions of South America and Africa, due to extensive chemical weathering and removal by deep leaching. Young sandy soils at higher latitudes under wet, temperate conditions and extremely sandy soils can also lead to Ca deficiency in plants due to limited retention of Ca released during chemical weathering.

Plant requirements, uptake, and deficiencies for Ca and Mg

Of the alkaline earth metals, only Ca and Mg are known to serve as essential nutrients for plants and animals. Other alkaline earth metals, such as Sr and Ba, can substitute for Ca with limited effects on plant health unless elevated to concentrations that substantially shift the Sr/Ca or Ba/Ca ratio (e.g. Myrvang et al., 2016; Burger and Lichtscheidl, 2019; Kabata-Pendias and Mukherjee, 2007). For plants, Ca is a constituent of the plant cell wall and is involved in maintaining the turgidity of plants. Moreover, Ca is an essential aspect of chemical signaling; it is required for cell elongation, cell division, and enzyme activation, particularly those that are membrane-bound. Calcium also serves an important role in the synthesis of cell wall sugars (De Freitas et al., 2016). Magnesium is also an essential element for plants, with its paramount role as the central metal ion of chlorophyll *a* and *b* molecules in chloroplasts in plant leaves. Because of its ionic size and electronegativity, Mg cannot be substituted making it vital for plant health. Limitations in total abundance or an insufficiently bioavailable form of Mg reduces the ability for a plant to obtain enough Mg to synthesize chlorophyll for photosynthesis. Magnesium is also important for starch production during photosynthesis and serves many roles in the enzymes of plants.

Plants predominately obtain Ca and Mg from soils by mass flow with water. However, alkaline earth metals can also be taken up by diffusion, depending on soil solution chemistry and rate of transpiration of the plant. Uptake by the roots can occur through passive and active transport mechanisms. Passive uptake occurs by absorption from higher concentrations in the soil into the root tissues via apoplastic pathways moving between free spaces between cell walls. This is driven by electrochemical potential gradients and requires higher concentrations in the soil solution. Active uptake requires absorption via carrier molecules, such as organic ligands, against electrochemical gradients into cell walls.

Calcium deficiency causes a reduction in growth of meristematic tissues, where the affected tissue becomes soft due to dissolution of the cell walls. Calcium deficiency rarely occurs to the extent of having a negative impact on staple field crops such as maize (*Zea mays*), soybean (*Glycine max*), rice (*Oryza sativa*), and wheat (*Triticum aestivum*), but is more prevalent in fruit crops such as apples, watermelons, tomatoes, and bell peppers. Calcium deficiencies in fruits can be seen as water-soaked tissues from plasma membrane breakdown and dark brown lesions on the fruit surface. In fruits like apples, bitter pit occurs, where the flesh of the fruit under goes cell dehydration and death resulting in dark lesions on the fruit surface (Van der Boon, 1980). In fruits like tomatoes and watermelons, Ca deficiency leads to blossom-end rot, which is the decay of the fruit at the flowering end (De Freitas et al., 2016). For leafy vegetables, Ca deficiency is characterized by stunted growth, curling, and brown spots and necrosis in young leaves, leaf tips, and meristematic cells (De Freitas et al., 2016). Unlike other essential nutrients where deficiencies can be solved through nutrient transfer from other tissues to younger tissues, Ca can only be transferred within the xylem and cannot be transferred from older tissues to the fruit. Since Mg is closely associated with chlorophyll molecules in chloroplasts, Mg deficiency symptoms are chlorosis of the leaves, formation of pale white to yellow spots, and interveinal streaks. At later stages Mg deficiency results in necrosis in the leaves first affected by the deficiency.

Ca and Mg soil amendments

In agroecosystems, Ca is most commonly added to soils to address acidity issues, and secondarily to prevent or address deficiencies. Calcium and Mg are depleted from soils due to chemical weathering, leaching, and nutrient export from crops, while acid generating metals such as Al and Fe remain. Aluminium and Fe can hydrolyze water, creating acidity as water molecules are split to generate hydroxide ligands to neutralize their +3 charge in solution. Furthermore, roots, microbes, and organic matter generate organic acids and commercial fertilizers can also generate acidity (Zoca and Penn, 2017). Soils must be limed with Ca-rich materials to neutralize the acidity of the soil when the base cation saturation is <35%. Lime, a generalized mixture of Ca oxides and hydroxide phases, can be used to increase base cation saturation. There are many types of liming materials for soils, which include calcite (CaCO_3), dolomite ($\text{CaMg}(\text{CO}_3)_2$), slaked lime ($\text{Ca}(\text{OH})_2$), burnt lime (CaO), and wollastonite/slag (CaSiO_3). There are two important factors for considering liming materials. First, is the amount of Ca and acid neutralization of the material, which is typically measured as the calcium carbonate equivalent (CCE) or the acid-neutralizing ability of the material normalized to its weight in pure CaCO_3 . The other major consideration is the duration of the effect; compounds such as slaked lime and burnt lime will act faster than crystalline calcite, limestone, and wollastonite because of their higher solubility and surface area for reaction. Faster acting compounds will help immediately within the growing season while slower acting liming materials may take multiple growing seasons to fully react.

Another common Ca soil amendment is gypsum. Gypsum may be used to reduce acidity when used in combination with other liming agents, but results have been mixed when gypsum is used alone resulting in pH increases, pH decreases, or other small effects (see Zoca and Penn, 2017). More importantly, gypsum has been reported to add Ca^{+2} which reduces Al^{+3} activity and Al on CEC sites. Gypsum is more commonly used as a sulfur (S) source, which is important for several crops that have a large S demand such as brassicaceae (broccoli, cabbage, canola) and fabaceae (soybean, peas). Gypsum also has been used in soil reclamation, as it can help against soil dispersion in sodic soils, leading to greater water infiltration, soil aggregation, and soil porosity (Zoca and Penn, 2017).

Magnesium concentrations in natural ecosystem soils are typically at levels adequate for plant growth from primary mineral weathering, atmospheric deposition, and internal biological cycling. In agroecosystems, Mg concentrations are also typically adequate for crop growth either from primary mineral weathering, direct supplementation from liming, or indirectly through the addition of other fertilizers such as superphosphate. However, Mg can be added as readily soluble forms such as Epsom salt or as a slow released form from dolomite, magnesite, or serpentine. Dolomite is the most commonly used soil amendment because it adds both Ca and Mg and leads to a reduction in soil acidity.

An important aspect when considering the addition of Ca and Mg as soil amendments is to consider the relative abundance of other essential elements. Excess liming can raise the soil pH to neutral or even basic conditions, which will decrease the solubility and availability of other nutrients that are less soluble under higher pH conditions (Essington, 2015). As soil pH increases, micronutrients such as Fe, Mn, B (boron), Cu (copper), and Zn (zinc) become strongly adsorbed to exchange sites and become unavailable for plant uptake. Furthermore excess addition of Ca can limit the availability of P for plants through the precipitation of hydroxyapatite minerals. Thus, important considerations are necessary to avoid micronutrient deficiencies, precipitation of insoluble minerals such as hydroxyapatite, and to avoid pH extremes when adding lime and soil amendments.

Soil beryllium, strontium, barium, and radium toxicity to plants and animals

Of the alkaline earth metals Be, Sr, Ba, and Ra do not serve a known biological function and elevated concentrations in soil pose a potential threat to plants, animals, and humans (Kabata-Pendias and Mukherjee, 2007). However, Be, Sr, Ba, and Ra are present at low concentrations in unmanaged terrestrial ecosystems and agroecosystems with limited to no effect on plants. Quantifying the cause and effect between elevated concentrations of Be, Sr, and Ba is often difficult because of coinciding enrichments of other metals. Most information on the toxicity of these metals is derived from additions of soluble salts, hydroponic solutions, or unique geologic settings with high concentrations of an alkaline earth metal. For example, elevated concentrations of Be $>10 \text{ mg kg}^{-1}$ in

soil or 5 mg L^{-1} in soil solution could reduce crop yields between 30% and 88% (Romney et al., 1962; Kabata-Pendias and Mukherjee, 2007). Soils with elevated Be may reduce root and shoot growth, but the abundance of Ca and Mg, soil pH and clay content can prevent any negative effects of Be on plant growth (Shah et al., 2016; Tanveer and Wang, 2019). Similarly, elevated Sr and Ba concentrations can have potentially negative impacts on plant growth. Mixed effects have been observed with some studies reporting observations of decreased rates of shoot and root growth as well as decreased chlorophyll content, but also positive effects of growth, photosynthesis, and secondary metabolites (Burger and Lichtscheidl, 2019). Regarding Sr toxicity, radioactive Sr from nuclear waste and nuclear disasters are of primary importance. ^{85}Sr , ^{89}Sr , and ^{90}Sr undergo gamma and beta decay, releasing ionizing energy that can have radiotoxic effects resulting in damaged DNA in plants and animals (Burger and Lichtscheidl, 2019). Furthermore, radioactive Sr can generate highly reactive species of ion radicals that can produce reactive oxygen species that can damage cellular structures within cells. Barium has also been observed to have mixed effects of toxicity at elevated concentrations. Solutions containing Ba concentrations at $50 \mu\text{M}$ and less did not have a significant effect on bush bean growth, but solution concentrations of $500 \mu\text{M}$ negatively impacted root and shoot growth as well as leaves (Llugany et al., 2000). The toxicity of Sr and Ba are also controlled by the relative abundance of Ca and Mg and the adsorption or precipitation of Sr and Ba. Most studies have focused on absolute concentrations but identifying adequate ratios of alkaline earth metals in soils to prevent toxicity in important crops and tree species is of future importance to evaluate critically potential negative impacts from Be, Sr, and Ba. Due to its scarcity in typical terrestrial environments, Ra toxicity to plants is limited to areas with point source pollution, such as waste sites and mining areas. As a prime example, Abreu et al. (2014) found that increasing soil ^{226}Ra concentrations from former U mine soils did not generate physiological damage in lettuce (*Lactuca sativa* L.) and maize, and the application of Ca-bearing soil amendments decreased ^{226}Ra uptake.

Exposure to elevated or potentially hazardous concentrations of alkaline earth metals from soils is typically limited because of the small concentrations of Be, Ba, Sr and Ra relative to Ca and Mg. Soil with elevated alkaline earth metals from pollution or unique geologic formations can generate dust. One important exposure route of these metals is the soil to plant to animal or human transfer in areas that may be enriched naturally or from anthropogenic pollution. Agroecosystems in close proximity to current and former alkaline earth metal mines are of most interest. For example, Lu et al. (2019) studied Ba in rice paddy soils adjacent to a Ba mine and found mine water caused elevated Ba concentrations in the rice, but it represented a low risk only to human health if consumed. Urban gardens can contain elevated concentrations of Sr and Ba, which may be transferred to animals or humans that consume the plants. Toxicity of ^{90}Sr from soils to plants to humans from consuming vegetables, mushrooms, or dairy products also poses a threat in areas near nuclear accidents and nuclear waste sites, such as Chernobyl, Ukraine; Fukushima, Japan; and Indian Point, USA. Thus, the current body of literature supports the conclusion that human exposure to hazardous soil concentrations of alkaline earth metals or radioactive forms is very limited.

Alkaline earth metals as environmental tracers

The relative abundance of alkaline earth metals and their specific isotopic ratios have been used to understand natural and anthropogenically altered processes in soils. For example, the relative abundances of Ca, Mg, Sr, and Ba in rainwater, rocks and minerals, dust, plant materials, and soil horizons can be investigated using Ca/Sr, Mg/Sr, Ca/Ba, Mg/Ba, and Sr/Ba ratios to determine relative rates of inputs and internal cycling. As a prime example, the Sr/Ca ratio of seawater and unweathered Pahala Ash deposits were compared to soils to determine the relative importance of sea spray to supply alkaline earth metals to terrestrial ecosystems (Whipkey et al., 2000). Furthermore, the unique Sr/Ca ratios can be applied to forests to determine which minerals are the source of alkaline earth metals for the trees and rates of internal cycling. Furthermore, the ratios of stable isotopes of Sr, specifically the $^{87}\text{Sr}/^{86}\text{Sr}$, and stable isotopes of $^{44}\text{Ca}/^{40}\text{Ca}$ can be used to trace their source and movement in soils. The $^{87}\text{Sr}/^{86}\text{Sr}$ has been effectively used to investigate precipitation inputs, dust inputs, rock and mineral weathering, and human pollution to soils and terrestrial ecosystems. Similarly, native Ca or Ca-amended systems can be studied using the $^{44}\text{Ca}/^{40}\text{Ca}$ ratio to quantify the Ca source, uptake by plants and animals, and rates of biological cycling in terrestrial ecosystems and agroecosystems (e.g. van der Heijden et al., 2013). The application of these non-traditional stable isotope systems has tremendous possibilities to further the understanding of alkaline earth metals in soils, terrestrial ecosystems, and agroecosystems.

Conclusion

The alkaline earth metal group represents a set of elements with many chemical and physical similarities, but with distinct difference for plants and animals. The alkaline earth metals are an important group of metals that are widely abundant on Earth and in soils. They are not sensitive to fluctuations in reducing and oxidizing conditions and have a + 2 charge. Except for Be, alkaline earth metals are +2 ions in the typical pH range for soils. Calcium, Sr, Ba and Ra have similar atomic radii and can substitute for each other within minerals.

Calcium and Mg are the most important of the alkaline earth metals because of their role as major, essential elements within plants, animals, and humans. Calcium is needed by animals for chemical signaling and is required for cell elongation, cell division, and enzyme activation in plants. Magnesium is required as an enzyme cofactor in animals and is a critical component of chlorophyll

molecules in plants. Beryllium, Sr, Ba, and Ra do not serve a known role in the mineral nutrition of animals and plants and can be toxic at elevated concentrations under uncommon conditions.

Alkaline earth metals are sourced to soils predominately from the weathering of rocks and minerals, but can also come from atmospheric deposition and human inputs. The accumulation and retention of alkaline earth metals in soils is governed by their abundance in the parent material, rates of leaching and weathering controlled by climate, and duration for pedogenesis to occur. The study of alkaline earth metals has been primarily through total abundances and elemental ratios. More recently, stable isotopes have greatly advanced our understanding of their biogeochemical cycling at local, regional, and global scales.

References

- Abreu MM, Lopes J, Santos ES, and Magalhães MCF (2014) Ecotoxicity evaluation of an amended soil contaminated with uranium and radium using sensitive plants. *Journal of Geochemical Exploration* 142: 112–121.
- Baudouin C and France L (2019) Trace element partitioning between wollastonite and alkaline silicate magmas. *Chemical Geology* 523: 88–94.
- Bindeman IN and Bailey JC (1999) Trace elements in anorthite megacrysts from the Kurile Island Arc: a window to across-arc geochemical variations in magma compositions. *Earth and Planetary Science Letters* 169(3–4): 209–226.
- Boy J and Wilcke W (2008) Tropical Andean forest derives calcium and magnesium from Saharan dust. *Global Biogeochemical Cycles* 22(1).
- Burger A and Lichtscheidl I (2019) Strontium in the environment: Review about reactions of plants towards stable and radioactive strontium isotopes. *Science of the Total Environment* 653: 1458–1512.
- Coint N, Barnes CG, Yoshinobu AS, Barnes MA, and Buck S (2013) Use of trace element abundances in augite and hornblende to determine the size, connectivity, timing, and evolution of magma batches in a tilted batholith. *Geosphere* 9(6): 1747–1765.
- De Freitas ST, Amarante CD, and Mitcham EJ (2016) Calcium deficiency disorders in plants. *Postharvest Ripening Physiology of Crops* 477–502.
- Essington ME (2015) *Soil and Water Chemistry: an Integrative Approach*. CRC Press.
- Galloway JN, Likens GE, Keene WC, and Miller JM (1982) The composition of precipitation in remote areas of the world. *Journal of Geophysical Research: Oceans* 87(C11): 8771–8786.
- Garrels RM and Christ (1965) *Solutions, Minerals and Equilibria*. San Francisco Freeman Cooper.
- Hossner LR (1996) Dissolution for total elemental analysis. *Methods of Soil Analysis* 49–64. Part 3.
- Kabata-Pendias A and Mukherjee AB (2007) *Trace Elements from Soil to Human*. Springer Science & Business Media.
- Karathanasis AD and Hajek BF (1996) Elemental analysis by X-ray fluorescence spectroscopy. *Methods of Soil Analysis* 161–223. Part 3.
- Lide DR (2005) Solubility product constants. *CRC Handbook of Chemistry and Physics*, 90.
- Llugany M, Poschenrieder C, and Barceló J (2000) Assessment of barium toxicity in bush beans. *Archives of Environmental Contamination and Toxicology* 39(4): 440–444.
- Lu Q, Xu X, Liang L, Xu Z, Shang L, Guo J, Xiao D, and Qiu G (2019) Barium concentration, phytoavailability, and risk assessment in soil-rice systems from an active barium mining region. *Applied Geochemistry* 106: 142–148.
- Maurya S and Rai SK (2020) Geochemical and Isotopic Composition of Gypsum Deposits from Sahastradhara Region of Lesser Himalaya, India. *Journal of the Geological Society of India* 95(2): 205–211.
- Myrvang MB, Bleken MA, Krogstad T, Heim M, and Gjengedal E (2016) Can liming reduce barium uptake by agricultural plants grown on sandy soil? *Journal of Plant Nutrition and Soil Science* 179(4): 557–565.
- Pi-Puig T, Animas-Torices DY, and Solé J (2020) Mineralogical and geochemical characterization of talc from two Mexican ore deposits (Oaxaca and Puebla) and nine talcs marketed in Mexico: Evaluation of its cosmetic uses. *Minerals* 10(5): 388.
- Rao CRM, Sahuquillo A, and Sanchez JL (2008) A review of the different methods applied in environmental geochemistry for single and sequential extraction of trace elements in soils and related materials. *Water, Air, and Soil Pollution* 189(1): 291–333.
- Reichholf JH (1986) Is Saharan dust a major source of nutrients for the Amazonian rain forest? *Studies on Neotropical Fauna and Environment* 21(4): 251–255.
- Romney EM, Childress JD, and Alexander GV (1962) Beryllium and the growth of bush beans. *Science* 135(3506): 786–787.
- Sagoe CI, Ando T, Kouno K, and Nagaoka T (1998) Relative importance of protons and solution calcium concentration in phosphate rock dissolution by organic acids. *Soil Science and Plant Nutrition* 44(4): 617–625.
- Sha LK and Chappell BW (1999) Apatite chemical composition, determined by electron microprobe and laser-ablation inductively coupled plasma mass spectrometry, as a probe into granite petrogenesis. *Geochimica et Cosmochimica Acta* 63(22): 3861–3881.
- Shah AN, Tarveer M, Hussain S, and Yang G (2016) Beryllium in the environment: Whether fatal for plant growth? *Reviews in Environmental Science and Bio/Technology* 15(4): 549–561.
- Takeno N (2005) Atlas of Eh-pH diagrams. In: *Geological survey of Japan open file report*, 419, 102.
- Tarveer M and Wang L (2019) Potential targets to reduce beryllium toxicity in plants: a review. *Plant Physiology and Biochemistry* 139: 691–696.
- Van der Boon J (1980) Prediction and control of bitter pit in apples. I. Prediction based on mineral leaf composition, cropping levels and summer temperatures. *Journal of Horticultural Science* 55(3): 307–312.
- van der Heijden G, Legout A, Midwood AJ, Craig CA, Pollier B, Ranger J, and Dambrine E (2013) Mg and Ca root uptake and vertical transfer in soils assessed by an in situ ecosystem-scale multi-isotopic (26 Mg & 44 Ca) tracing experiment in a beech stand (Breuil-Chenue, France). *Plant and Soil* 369(1): 33–45.
- Vedanand S, Rao PS, and Reddy BJ (1993) Spectroscopic investigations on different grade samples of talc. *Radiation Effects and Defects in Solids* 127(2): 169–176.
- Wang H and Han G (2011) Chemical composition of rainwater and anthropogenic influences in Chengdu. *Southwest China. Atmospheric Research* 99(2): 190–196.
- Whipkey CE, Capo RC, Chadwick OA, and Stewart BW (2000) The importance of sea spray to the cation budget of a coastal Hawaiian soil: a strontium isotope approach. *Chemical Geology* 168(1–2): 37–48.
- White WM (2020) *Geochemistry*. John Wiley & Sons.
- Xiao B, Chen H, Wang Y, Han J, Xu C, and Yang J (2018) Chlorite and epidote chemistry of the Yandong Cu deposit, NW China: Metallogenic and exploration implications for Paleozoic porphyry Cu systems in the Eastern Tianshan. *Ore Geology Reviews* 100: 168–182.
- Zoca SM and Penn C (2017) An important tool with no instruction manual: A review of gypsum use in agriculture. *Advances in Agronomy* 144: 1–44.

Transition metals in soil

Siobhan Staunton, UMR Eco&Sols, INRAE, INRAE-IRD-Cirad-Institut Agro-Université de Montpellier, Montpellier, France

© 2023 Elsevier Ltd. All rights reserved.

Introduction	135
Definitions and terminology	135
Sources and environmental fate	136
Examples of environmentally relevant metals in soil	137
Scandium. Group 3 (formerly IIIB)	137
Titanium. Group 4 (formerly IVB)	137
Vanadium. Group 5 (formerly VB)	137
Chromium. Group 6 (formerly VIB)	137
Manganese. Group 7 (formerly VIIB)	137
Iron. Group 8 (formerly VIII)	138
Cobalt. Group 9 (formerly VIII)	138
Nickel. Group 10 (formerly VIII)	138
Copper. Group 11 (formerly IB)	138
Technetium. Group 7 (formerly VIIB)	139
Palladium. Group 10 (formerly VIII)	139
Silver. Group 11 (formerly IB)	139
Cadmium. Group 12 (formerly IIB)	139
Tantalum. Group 5 (formerly VB)	140
Tungsten. Group 6 (formerly VIB)	140
Osmium. Group 8 (formerly VIII)	140
Iridium. Group 9 (formerly VIII)	140
Platinum. Group 10 (formerly VIII)	140
Gold. Group 11 (formerly IB)	140
Mercury. Group 12 (formerly IIB)	140
Conclusions	140
References	141

Abstract

The chemical definition of transition metals is given, and terminology discussed. The origin, industrial uses and environmental fate of environmentally relevant metals are discussed briefly and references given.

Key points

- Transition metals are defined as metals of the d-block of the periodic table and generally have more than one oxidation state
- They exist in numerous oxidation states in soil and include cationic and anionic forms
- Some have been studied in depth and their fate in the environment is well documented
- They also include emerging pollutants for which more research is required

Introduction

Many of the elements in soil are metals, including transition metals. These elements have various sources, environmental fates and biological activity, making generalizations impossible. Of the 40 transition metals, only a small number are environmentally relevant in soils. Some essential feature of these elements and their dynamics in soil will be given and references for further information are indicated.

Definitions and terminology

The International Union of Pure and Applied Chemistry (IUPAC) define metals as elements which conduct electricity, have a metallic luster, are malleable and ductile, form cations, and have basic oxides. Transition metals are a large subgroup of metals. The

IUPAC has given two definitions. The older, more restrictive definition is an element whose atom has a partially filled d sub-orbital or shell, or which can give rise to cations with an incomplete d sub-orbital. However, since 2011 the preferred, broader definition is an element in groups 3 to 12 of the Mendeleev periodic table, namely all the d-block elements. In this definition, metals in group 12 (zinc (Zn), cadmium (Cd) and mercury (Hg)) with only two oxidation states (0 and +2) whose cations have complete d orbitals, are included, whereas they were formerly defined as post-transition metals. Lanthanides and actinides may be referred to as inner-transition metals.

Transition metals often have several environmentally relevant oxidation states, and may therefore be redox sensitive. An illustrative example of this is given by manganese (group 7) that exists in six oxidation states from +2 to +7. In contrast, metals of group 3 (scandium, yttrium and lutetium), although of little environmental importance, exist only as +3, in addition to the elemental form. At the other end of the transition series, group 12 (zinc, cadmium and mercury) exist as +2, with +1 in addition only for mercury. Another feature of transition elements is that due to d–d electronic transitions their salts are often colored. While not directly of importance in soil science, this property was important in elemental analysis of extracted metals by atomic spectroscopy. Many transition metals also form compounds that are paramagnetic, and this has important consequences for the use of nuclear magnetic resonance (NMR) spectroscopy in soil science. Transition metals also have richer complex, or coordination chemistry than s-block and p-block metals. This is important for their chemistry and dynamics in soils. The Pearson Hard–Soft Acid–Base theory may be useful to predict some aspects of the coordination theory of transition metals. This theory allows a qualitative prediction of the strength of association between Lewis acids (electron pair-acceptor) and Lewis base-donors; soft-soft and hard-hard interactions being favored over hard-soft. The smaller, high charge weakly polarizable cations to the left of the periodic table are hard, and thus associate with oxygen ligands. Heavier (second and third row) metal cations tend to be soft, and associate with hydrides and sulfur-containing ligands. However most of the environmentally relevant transition metals are borderline between hard and soft.

The terminology of transition metals may be confusing. The term was first coined by the chemist Charles Rugeley Bury in 1921. As stated above, the IUPAC definition of transition elements has varied over time, to exclude or include elements of groups 3 and 12 that are d-block elements, but do not have incomplete d orbitals in all oxidation states. However despite many criticisms, the term ‘heavy metal’ is still widely used (Duffus, 2002). This loosely defined term often encompasses metals and metalloids with density greater than 5 kg dm⁻³. The term implies contaminants with toxic properties. But elemental density has no relation with chemistry or biological properties. Some metals have no known biological role, while others may be micronutrients, required at small levels for example in catalysts, but can become toxic at excess levels. The terms trace element or metal make no assumptions about the biological role of the element, simply excluding major elements such as iron (Fe), aluminum (Al) and manganese (Mn). The terms ‘potentially toxic’ or ‘nutrient’ have biological meaning, but no chemical implications; transition element or metal has chemical meaning, but includes elements with a wide range of biological properties. Heavy metal, although used in toxicological literature and by legislative authorities in many countries, is neither well-defined nor informative, and should be avoided.

Sources and environmental fate

The sources of transition metals in the soil environment are variable. In most ecosystems, most of the metals present are of natural, geological origin. For example, iron and manganese are major components of soils; they are carrier phases for other metals, which influence their dynamics. In restricted geographic areas a metal may be present naturally with contents that are considered excessive. The best-known example is nickel (Ni) in soils developed on serpentine that may contain up to 100 times the world average Ni content (20 mg kg⁻¹). There are cases of deficiency in some metals in agricultural systems, particularly for copper (Cu), zinc, nickel and molybdenum (Mo). Soils may be contaminated by transition metals due to human activity, notably agricultural and industrial activity. Source of metal addition to soils may be atmospheric fallout from industry, and transport, and agricultural amendments, including phosphate fertilizers containing cadmium, manure and sludge containing copper and zinc. New technology leads to emerging pollutants, with different chemical and biological properties. The dynamics of these metals, including titanium, yttrium, zirconium, tantalum, tungsten, and zirconium are as yet poorly understood and documented (Souza et al., 2021).

In general native, or legacy, metals in soil, largely of geochemical origin, are not labile and contribute little to the bioavailable and mobile fractions of the metals in soil. More recently added metal, whatever the origin, but depending on the chemical form, tends to be more mobile, labile and bioavailable. Added metal salts are rapidly precipitated, adsorbed on soil minerals, and complexed with soil organic matter. In general, only a tiny proportion remains in solution, and this proportion is also influenced by the formation of soluble or colloidal complexes in solution. With time, recently added metals tend to become less labile, more fixed on the soil organo-mineral complex; a process known as attenuation.

The dynamics of transition metals in soil depend on soil redox conditions. There are both direct and indirect influences: most transition metals have redox chemistry and so their oxidation state, and thus mobility, reactions with soil minerals and biological properties, depend on the redox state. In addition, as for other elements in soil, the redox-sensitive solubility of carrier phases such as iron and manganese (oxy)hydroxides influences the immobilization or solubility of other metals.

When soils are contaminated by transition metals, there are various strategies to manage these soils and these have been treated in other chapters. Briefly, these include

- phytoextraction, where plants are used to extract metal thus decreasing the amount in soil, assuming that labile metal is removed which leaves progressively less bioavailable metal in the soil, therefore effectively cleaned-up to some extent

- use of non-food crops or non agricultural land use
- covering very contaminated soil with other soil and growing crops to stabilize the soil physically against erosion
- use of soil amendments to provide carrier phases to immobilize the contaminant, to change soil pH, or to improve the nutrient status of soil to compete with contaminant metal cations and thus reduce their uptake.

Examples of environmentally relevant metals in soil

Of the 40 transition metals, only a small number are environmentally relevant in soils. Essential information on these metals is considered below. Compendia of reviews of the chemistry and biological properties of metals in soils can be found in text books, notably *Kabata-Pendias and Mukhrjee (2007)* and *Coughtrey and Thorne (1983)*, although the latter focuses on elements whose radioactive isotopes may be found in the environment. Trace element deficiencies and their causes, with emphasis on soils of England, are reviewed by *Thornton and Alloway (1974)*, and recent advances in the understanding of micronutrients has been reviewed by *Singhal et al. (2022)*. References are given in subsections only if recent reviews are available for elements with relatively little published data.

Scandium. Group 3 (formerly IIIB)

Scandium (Sc) is in the first row and first group of the transition metals, atomic mass 44.96, present as Sc^{3+} , that can substitute for Al and Fe. It is present naturally in soils, often associated with biotite and igneous rocks and released by uranium processing and some industrial processes. There is little information on this element, and it has no known beneficial biological role.

Titanium. Group 4 (formerly IVB)

Titanium (Ti) is in the first row and second group of the transition metals, atomic mass 47.87, and often present in the tetravalent form in titanates and silicates. It is relatively abundant in the Earth's crust, $0.4\text{--}0.6\text{ mg kg}^{-1}$, and so is present naturally in soils, with a mean of about 0.3 mg kg^{-1} . Although it may be released into the environment from some industrial processes, notably paint production, it has been little studied. Its oxide is however an emerging contaminant as a nanoparticle. It has various industrial uses, including as a component of steel alloys, in catalysts, the textile industry.

Vanadium. Group 5 (formerly VB)

Vanadium (V) is in the first row and third group of the transition metals, atomic mass 50.9; it exists in various oxidation states, from -3 to $+5$, forming anions and cations. Its abundance in both the Earth's crust and soils is about 60 mg kg^{-1} , which makes it one of the most abundant elements. It is widely used in many industries, and anthropogenic activity thus creates a potential pollution risk. It appears to be an essential micronutrient, and small additions may promote plant growth, however it may become toxic in excess. Its complex redox chemistry makes the study of its dynamics complicated, and therefore rather little information is currently available (*Chen et al., 2021*).

Chromium. Group 6 (formerly VIB)

Chromium (Cr) is in the first row and fourth group of the transition metals, atomic mass 51.9; it exists in various oxidation states, the most common being $+3$ and $+6$, in addition to the elemental form, and occurs as both anions and cations. Background contents of Cr in soil vary, but an average may be about 50 mg kg^{-1} with larger contents at depth than in surface soils. Very large sources of Cr have been reported in serpentine soils, and there are numerous anthropogenic sources of Cr in the environment because of its industrial uses, notably in the tannery industry, steel production, electroplating and the chemical industry. Although there is no known beneficial role for Cr in plants, it is an essential micronutrient for animals, although toxic or dermo-irritant at larger concentrations. The speciation of chromium depends strongly on redox potential and pH. The $+3$ form is the most stable under most environmental conditions. This form is strongly sorbed on soil, and thus is very immobile, except in very acid soils, with correspondingly low bioavailability. It also interacts with organic and inorganic ligands to form complexes however this form is also much less toxic to living organisms than is the more oxidized form. The hexavalent form is considered to be among the most harmful of pollutants and is readily absorbed by inhalation, ingestion and absorption. The mechanism of its toxicity is largely due to its strong oxidizing capacity, thus forming reactive oxygen species. More information is reviewed by *Ertani et al. (2017)*.

Manganese. Group 7 (formerly VIIB)

Manganese is in the first row and fifth group of the transition metals, atomic mass 54.9, with variable oxidation states, from $+2$ to $+7$. Manganese is an abundant element in both the Earth's crust and in soils. The manganese contents of soils worldwide vary by more than 3 orders of magnitude from about 7 to over 9000 mg kg^{-1} , depending on geological origin and texture. Manganese (hydr)oxides, often amorphous, are important minerals in soil that are formed, solubilized and reformed under varying redox

conditions, contributing to the dynamics of many other elements. The most important oxidation states are +2, +4 and +7, with both cationic and anionic forms, and numerous organic and inorganic complexes in solution, including many with hydroxyl groups (e.g. MnOH^+ , $\text{Mn}_2(\text{OH})_2^{2+}$, Mn_2OH^+).

In addition to naturally occurring manganese, the numerous industrial uses of manganese including metallurgy, pigments, and transport, contribute to its addition to the environment. Manganese is an essential micronutrient contained in many enzymes for example. However while excess exposure can be toxic, there is no evidence of manganese toxicity from environmental sources. Most soil science research on manganese focuses on the dynamics of manganese oxides under fluctuating redox conditions and their roles as carrier phases (Berg et al., 2015; Li et al., 2021).

Iron. Group 8 (formerly VIII)

Iron is in the first row and sixth group of the transition metals, atomic mass 55.8, with variable oxidation states, from +2 to +6, predominantly +3. Iron is the most abundant metal in the lithosphere, accounting for about 4.5% of the Earth's crust and between, 0.1 and 10% of soil. It occurs as oxides and hydroxides, both amorphous and crystalline minerals, and forms coatings on clay minerals. Amorphous iron oxides may make an important contribution to soil specific surface area, particularly in tropical soils where the major clay mineral is kaolinite. It is an important carrier phase for other metals, and for organic matter and phosphate. It has a complex geochemistry that depends strongly on the Eh–pH system. Like manganese, the dynamics of iron minerals, especially under fluctuating redox and pH conditions, determine the dynamics of many other metals in soil. Transport and transformation of Fe minerals determine much of soil weathering and development. Iron is often a major contributor to soil color, with red or yellow colored minerals including hematite, goethite and magnetite.

Iron is an essential nutrient for plants and animals, being an important component of many organic compounds, including proteins. Iron toxicity exists, but the cause in humans is generally by poisoning, not environmental exposure. Iron deficiency in humans and other animals may be caused by both physiological disorders and inadequate dietary intake. In plants, iron deficiency (chlorosis) may occur in alkaline or calcareous soils, or soils with excess amounts of cations of similar size such as Mn, Ni or Co. Some plants have mechanisms to cope with iron deficiency, including the release of phytosiderophores and reduction of Fe^{3+} .

Cobalt. Group 9 (formerly VIII)

Cobalt is in the first row and seventh group of the transition metals, atomic mass 58.93, with oxidation states, of +2, +3 and +4, predominantly +2. Cobalt is present with similar contents in both the Earth's crust and soil, namely about 10 mg kg^{-1} . It is contained in various minerals often associated with sulfur (S), selenium (Se) and arsenic (As), and its geochemical cycle resembles those of Fe and Mn. Its major industrial use is in metallurgy, with smaller requirements for the pharmaceutical industry, fuel production, pigments and plastics. There is little evidence of soil contamination by Co, except near some metal processing plants, and so soil contents are usually determined by parent materials and soil-forming processes. Soil contents tend to be greatest in serpentine soils and soils in arid and semiarid regions, and lowest in colder climates. Cobalt is often associated with iron and manganese oxy-hydroxides, and thus is sensitive to redox conditions.

Cobalt is an essential nutrient for plants and animals. It is a component of vitamin B12 and can be deficient for grazing animals on many soils (Hu et al., 2021; Singhal et al., 2022). It is also essential for symbiotic nitrogen (N) fixation. Deficiency is generally addressed by nutrient supplements rather than soil amendments.

Nickel. Group 10 (formerly VIII)

Nickel is in the first row and eighth group of the transition metals, atomic mass 58.69, with oxidation states, of +1, +2, +3 and +4, predominantly +2, with little environmentally relevant redox chemistry. Nickel is present in similar amounts in both the Earth's crust and soil, namely about 20 mg kg^{-1} , but with a large range and up to 450 in serpentine soils. The major industrial use of Ni is in the steel industry, but it is also used in ceramics, food technology and catalysis. In soil, Ni is associated with Fe and Mn (hydr)oxides and organic matter.

Nickel is an essential element, but there is little evidence of deficiency. However it is potentially toxic at elevated concentrations; its mobility, particularly in sandy soils (Luo et al., 2017), and phytoavailability may be problematic when plants are grown on contaminated soils. Plants growing on serpentine soils have developed the ability to hyperaccumulate Ni and then to sequester it with chelating agents as a tolerance mechanism (Brady et al., 2005). Topsoil contamination has been reported near smelters, and after the use of contaminated sludges. In a few cases there is evidence of atmospheric deposition on soil after long distance transport of atmospheric Ni.

Copper. Group 11 (formerly IB)

Copper is in the first row and ninth group of the transition metals, atomic mass 63.54, with oxidation states, of +1 and +2, predominantly +2, with little environmentally relevant redox chemistry. Copper is present in similar abundance in both the Earth's crust and soil, namely about 20 mg kg^{-1} , but with a large range from 1 to over 100. Copper was one of the first metals known to man and has been used for millennia. Copper minerals are usually sulfides, oxides or more rarely carbonates. Industrial uses of

copper are numerous and include metallurgy, art, ammunition and telecommunications, and even in modern times, its earliest uses in jewelry and coinage remain. Another environmentally significant use of copper is in agriculture. It is a feed additive for livestock, and a common pesticide, particularly in vineyards and in organic agriculture.

Copper is associated with clay minerals, organic matter, carbonates and Fe and Mn oxyhydroxides. It has little environmentally relevant redox chemistry.

Copper is an essential nutrient and deficiency has been reported in many soils and agricultural uses. The treatment of deficiency tends to be foliar sprays for crops and feed amendments for livestock, rather than soil amendments. Copper contamination of soils has been the object of numerous studies. The sources are both industrial and agricultural, in particular for vineyard soils and soils subjected to sewage sludge amendments. The low mobility of Cu in soils means that although the bioavailability of Cu is quite poor, it remains present in soils for a long time. Since vines are deep-rooting and often suffer no ill-effects from surface applied Cu, even after years of accumulation (Ruyters et al., 2013). Problems of phytotoxicity occur after land-use changes. There is also much evidence of Cu-related deterioration of soil microbial activity.

Technetium. Group 7 (formerly VIIB)

Technetium (Tc) is in the second row and fifth group of the transition metals, below manganese, atomic mass 98.9, with oxidation states, of +4, +6 and +7, and it occurs mainly as the heptavalent oxyanion, TcO_4^- . All the isotopes of Tc are radioactive. Tiny amounts of Tc are found in uranium ore. The sources of Tc in the environment are from fallout and nuclear processing. All the research on this element has been in the field of radioecology and some of its properties and environmental fate have been discussed in the chapter on radionuclides in the environment.

Palladium. Group 10 (formerly VIII)

Palladium (Pd) is in the second row and eighth group of the transition metals, below nickel, atomic mass 106.42 with oxidation states of +1, +2, +3 and +4, but predominantly +2. Its abundance in the Earth's crust is about $10 \mu\text{g kg}^{-1}$, with similar, slightly larger amounts in soil. It is a noble metal and a member of the platinum group metals (PGM), namely iridium, osmium, palladium, platinum, rhodium and ruthenium, recently reviewed by (Savignan et al., 2021). These metals are exceptionally rare, are often found together and have similar chemical properties. All are increasingly used in new technology and are therefore potential emerging pollutants. In particular, their use in catalysts leads to contamination from automobile exhaust catalysts. Almost all the soil science research on palladium to date involves monitoring of soil content and analytical techniques. It is known to be toxic to humans when inhaled, but no data are available on the consequences of environmental contamination.

Silver. Group 11 (formerly IB)

Silver (Ag) is in the second row and ninth group of the transition metals, below copper, atomic mass 107.87 with oxidation states of +1, +2 and +3, but predominantly +1. It is a noble metal, known and appreciated by man in jewelry for millennia. It is scarce, and occurs in the Earth's crust at contents of about $70 \mu\text{g kg}^{-1}$, and in the range $0.06\text{--}0.4 \text{ mg kg}^{-1}$ in soils, with larger amounts in organic soils, up to 5 mg kg^{-1} . Jewelry remains one of the major uses of silver. Its use in photography and in X-ray films has decreased dramatically, while various industrial applications and medical uses remain. In soil, silver is associated with sulfides and organic matter. There are few data on silver mobility and bioavailability in soil, and these have been reviewed by (Wang et al., 2018). Silver nanoparticles may be an emerging pollutant, because of their antimicrobial properties that make them attractive in textiles, food, medical and pharmaceutical industries. However their toxicity would be related to their size rather than the chemical and biological properties of the element. Silver has no known biological role.

Cadmium. Group 12 (formerly IIB)

Cadmium is in the second row and tenth group of the transition metals, below zinc, atomic mass 112.41 with oxidation states of +2. Its abundance in soil is about 0.5 mg kg^{-1} , it has a similar chemistry to Zn, with a greater affinity for sulfide. It is often a byproduct of mining for Zn. It is widely used in batteries and in pigments. In soil it is associated with colloids, sulfides and organic matter. Its dynamics are strongly related to pH, partly due to its complex formation. It may precipitate at alkaline pH and under reducing conditions due to sulfide formation.

There are many anthropogenic sources of Cd in soils, the most important being atmospheric deposition and the use of phosphate fertilizers (Liu et al., 2019). Plant uptake may be reduced by soil liming, although other techniques such as land-use change, phytoremediation and inoculation with microorganisms have been tested. There is no known beneficial role for cadmium in plants or animals and it is one of the most toxic elements to man. The mechanisms of toxicity involve its binding with sulfur groups of proteins, and enhancing the production of reactive oxygen species. However one of the major sources of human intake of cadmium is from smoking tobacco.

Tantalum. Group 5 (formerly VB)

Tantalum (Ta) is in the third row and third group of the transition metals, below vanadium, atomic mass 180.09 with oxidation states of +3, +4 and +5, mainly +5. Its abundance in soil is about 1 mg kg^{-1} , about half that in the Earth's crust. Although there have been few studies of Ta, its increasing use in both surgery and electronic devices makes it a potentially emerging pollutant.

Tungsten. Group 6 (formerly VIB)

Tungsten (W) is in the third row and fourth group of the transition metals, below chromium, atomic mass 183.8 with oxidation states from +2 to +6, mainly +6. Its abundance in soil is about 1 mg kg^{-1} , similar to that in the Earth's crust. There have been few studies of W in the environment, although elevated soil contents have been observed around industrial plants, including ore processing plants. There is some evidence that W may have a biological role, and that it may compete with Mo in reductase reactions in plants. Its increasing use in metal alloys, in pigments and automobile exhaust catalysts make it a potentially emerging pollutant.

Osmium. Group 8 (formerly VIII)

Osmium (Os) is in the third row and sixth group of the transition metals, below iron, atomic mass 190.2 with oxidation states of +3, +4, +6 and +8, mainly +3. Its abundance in the Earth's crust is about $1 \text{ } \mu\text{g kg}^{-1}$. It is a member of the platinum metal group. It is increasingly used in electronic devices. Non-elemental forms are thought to be mobile and are known to be highly toxic.

Iridium. Group 9 (formerly VIII)

Iridium (Ir) is in the third row and seventh group of the transition metals, below cobalt, atomic mass 192.2 with oxidation states of +2, +3, +4 and +6, mainly +4. Its abundance in the Earth's crust is about $0.5 \text{ } \mu\text{g kg}^{-1}$. It is a member of the platinum metal group. It is increasingly used in specialist alloys and electronic devices. It appears to be mobile and bioavailable, although no biological roles or adverse effects are known.

Platinum. Group 10 (formerly VIII)

Platinum (Pt) is in the third row and eighth group of the transition metals, below nickel, atomic mass 195.07 with oxidation states of +2, +3, +4 and +5, mainly +4. Its abundance in the Earth's crust is variable, but low, in the range of $\mu\text{g kg}^{-1}$ and is associated with ores containing other PGM. It is a noble metal and gives its name to the platinum metal group. It is well known for its use in jewelry and is increasingly used in specialist alloys, catalysts and medical devices. Its major source in the environment appears to be from road traffic. It appears to have some mobility and bioavailability, but there is currently very little information on toxicity. The distribution, fate and possible toxicity of PGM have recently been reviewed (Savignan et al., 2021).

Gold. Group 11 (formerly IB)

Gold (Au) is in the third row and ninth group of the transition metals, below copper, atomic mass 196.97 with oxidation states of +1 and +3, mainly +1, in addition to the elemental state. Its abundance in the Earth's crust is about $3 \text{ } \mu\text{g kg}^{-1}$. It is a noble metal, and often found in the elemental state. It was one of the earliest metals known to man and is still well known for its use in jewelry. Although gold may be found as large, easily identified nodules, the more finely divided metal may be extracted using flocculation with mercury or cyanidation; both environmentally hazardous processes. Despite its chemical inertia and dominance of the elemental form, oxidized forms of gold have some mobility in the environment and bioavailability. There is some evidence of toxicity, but due to professional exposure rather than dietary intake for humans.

Mercury. Group 12 (formerly IIB)

Mercury (Hg) is in the third row and tenth group of the transition metals, below zinc, atomic mass 200.59 with oxidation states of +1, +2 and +3, mainly +2, in addition to the elemental state. Its abundance in the Earth's crust is in the range $20\text{--}60 \text{ } \mu\text{g kg}^{-1}$, with a similar content in soil. It is a liquid in the elemental state, and is probably best known for its use in thermometers. Its dispersal in the environment is facilitated by the volatilization of elemental Hg and to a lesser extent methylated forms. Its production and industrial use tends to decrease, largely because of its toxic properties. It is highly toxic to microbes, plants and animals. A considerable volume of literature exists on its toxic properties, particularly methylated species (Natasha et al., 2020).

Conclusions

Transition metals are a chemically and biologically heterogeneous group, and many of these elements have significant environmental impact. Research will continue on the dynamics and fate of traditional metal pollutants and micronutrients. However the most exciting aspect of transition metal research will focus on emerging pollutants released by innovative technology and the unravelling of the biological processes involving trace amounts of these metals as essential nutrients or toxic substances (Fig. 1).

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
H																	He
Li	Ba											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	La - Lu	Hf	Ta	W	Ru	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	Ac-Lr	Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg	Cn	Uut	Fl	Uup	Lv	Uus	Uuo

Lanthanides	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
Actinides	Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr

Fig. 1 Simplified periodic table showing the transition metals with blue shading and intensity of color giving a crude indication of their importance in the soil environment.

References

- Berg B, Erhagen B, Johansson M-B, Nilsson M, Stendahl J, Trum F, and Vesterdal L (2015) Manganese in the litter fall-forest floor continuum of boreal and temperate pine and spruce forest ecosystems – A review. *Forest Ecology and Management* 358: 248–260.
- Brady KU, Kruckeberg AR, and Bradshaw HD Jr. (2005) Evolutionary ecology of plant adaptation to serpentine soils. *Annual Review of Ecology, Evolution, and Systematics* 36: 243–266.
- Chen L, Liu JR, Hu WF, Gao J, and Yang JY (2021) Vanadium in soil-plant system: Source, fate, toxicity, and bioremediation. *Journal of Hazardous Materials* 405: 124200.
- Coughtrey PJ and Thorne MC (1983) *Radionuclide Distribution and Transport in Terrestrial and Aquatic Ecosystems. A Critical Review of Data*. vol. 2. Netherlands: A.A. Balkema.
- Duffus JH (2002) "Heavy Metal" – A meaningless term? *Pure and Applied Chemistry* 74: 793–807.
- Ertani A, Miletto A, Borin M, and Nardi S (2017) Chromium in Agricultural Soils and Crops: A Review. *Water, Air, & Soil Pollution* 228.
- Hu X, Wei X, Ling J, and Chen J (2021) Cobalt: An essential micronutrient for plant growth? *Frontiers in Plant Science* 12: 768523.
- Kabata-Pendias A and Mukhrjee AB (eds.) (2007) *Trace Elements from Soil to Human*. Germany: Springer, Berlin.
- Li H, Santos F, Butler K, and Herndon E (2021) A critical review on the multiple roles of manganese in stabilizing and destabilizing soil organic matter. *Environmental Science & Technology* 55: 12136–12152.
- Liu X, Tang A, Li H, Fangmeier A, Ma X, Zhuang Z, and Niño-Savala AG (2019) Cadmium pollution from phosphate fertilizers in arable soils and crops: An overview. *Frontiers of Agricultural Science and Engineering* 6: 419.
- Luo D, Zheng H, Chen Y, Deleporte P, Xie T, Staunton S, and Wang G (2017) Influence of soil properties on Ni accumulation in food crops and corresponding dietary health risk with a typical Chinese diet. *Soil Use and Management* 33: 653–662.
- Natasha, Shahid M, Khalid S, Bibi I, Bundschuh J, Khan Niazi N, and Dumat C (2020) A critical review of mercury speciation, bioavailability, toxicity and detoxification in soil-plant environment: Ecotoxicology and health risk assessment. *Science of the Total Environment* 711: 134749.
- Ruyters S, Salaets P, Oorts K, and Smolders E (2013) Copper toxicity in soils under established vineyards in Europe: A survey. *Science of the Total Environment* 443: 470–477.
- Savignan L, Faucher S, Chery P, and Lespes G (2021) Platinum group elements contamination in soils: Review of the current state. *Chemosphere* 271: 129517.
- Singhal RK, Fahad S, Kumar P, Choyal P, Javed T, Jinger D, Singh P, Saha D, Md P, Bose B, Akash H, Gupta NK, Sodani R, Dev D, Suthar DL, Liu K, Harrison MT, Saud S, Shah AN, and Nawaz T (2022) Beneficial elements: New Players in improving nutrient use efficiency and abiotic stress tolerance. *Plant Growth Regulation* 100: 237–265.
- Souza IC, Morozesk M, Azevedo VC, Mendes VAS, Duarte ID, Rocha LD, Matsumoto ST, Elliott M, Baroni MV, Wunderlin DA, Monferran MV, and Fernandes MN (2021) Trophic transfer of emerging metallic contaminants in a neotropical mangrove ecosystem food web. *Journal of Hazardous Materials* 408: 124424.
- Thornton I and Alloway BJ (1974) Geochemical aspects of the soil-plant-animal relationship in the development of trace element deficiency and excess. *The Proceedings of the Nutrition Society* 257–266. England.
- Wang R, Du H, Wang Y, Wang D, Sun Q, and Zhou D (2018) Retention of silver nanoparticles and silver ion to natural soils: Effects of soil physicochemical properties. *Journal of Soils and Sediments* 18: 2491–2499.

Nitrogen and ammonia in soils

Nicholas T Girkin and Hannah V Cooper, School of Water, Energy and Environment, Cranfield University, Cranfield, United Kingdom; Agricultural and Environmental Sciences, School of Biosciences, Sutton Bonington Campus, University of Nottingham, Nottingham, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction	143
The need for nitrogen	143
Soil nitrogen transformations	143
Nitrogen compounds in soils	144
Nitrogen fixation and assimilation	144
Nitrogen mineralization and immobilization	145
Ammonia and ammonium	145
Nitrification	146
Denitrification	146
Codenitrification	147
Chemodenitrification	147
Dissimilatory nitrate reduction to ammonium	148
Anammox	148
Measuring and monitoring human impacts on nitrogen cycles	148
Anthropogenic impacts on the soil nitrogen cycle	148
Quantification of soil nitrogen transformations and emissions	149
Future research directions and conclusions	150
Research priorities for soil nitrogen	150
Conclusions	150
References	150

Abstract

Soil nitrogen transformations underpin plant growth and are fundamental to healthy ecosystem functioning. Excess nitrogen applications, particularly in the form of high fertilizer inputs, adversely affect soil health and cause substantial nitrous oxide and ammonia emissions, contributing to climate change and atmospheric pollution. Understanding nitrogen transformations in soil is therefore an important first step to devising mitigation strategies. Dominant nitrogen pathways include mineralization of organic nitrogen into plant available forms (ammonium and nitrate), and subsequent nitrification and denitrification, which ultimately produce nitrous oxide and/or dinitrogen. A range of techniques is now available for quantifying soil nitrogen turnover and partitioning pathways, including the application of stable isotopes and micro-meteorological methods.

Glossary

Aminization The first stage of nitrogen mineralization, and the process by which complex proteins are broken down into smaller amino acids, amides and amines.

Ammonification The second stage of nitrogen mineralization, and the process by which organic forms of nitrogen, including amino acids, amides and amines, are converted into ammonia or ammonium.

Denitrification The reduction of nitrate and nitrite to gaseous forms of nitrogen, including nitrous oxide and nitrogen, occurring under anaerobic conditions.

Immobilization The uptake of nitrate and ammonium by the soil microbial community.

Nitrification The oxidation of ammonia to nitrite, occurring under aerobic conditions.

Nitrogen fixation The process by which atmospheric nitrogen is assimilated into organic compounds.

Mineralization The conversion of organic nitrogen into inorganic, plant available forms of nitrogen.

Key points

- Nitrogen is an essential element for plant growth and is often added in agricultural systems to maintain yields.
- Nitrogen applications are often in excess of crop demand driving substantial gaseous emissions and ultimately causing adverse environmental impacts.

- Soil nitrogen turnover is dominated by the processes of nitrification and denitrification, but other processes (for example codenitrification) may make important contributions under specific circumstances.
- A range of methodologies is now available for quantifying the scale of nitrogen losses and tracing soil nitrogen dynamics, and collectively creating new opportunities for improving nitrogen use efficiency and mitigating emissions.

Introduction

Nitrogen is an essential element for plant growth, and ultimately underpins food security. High rates of application of nitrogen fertilizer, often in excess of crop demands, alongside pollution, have substantially increased the transfer of nitrogen into soil systems, and the atmosphere. This has led to a range of adverse environmental impacts across a range of ecosystems. For example, substantial nitrous oxide (N_2O) emissions are an important contributor to climate change. Moreover, the impacts of climate change are likely to affect soil nitrogen transformation pathways, potentially leading to climate feedbacks. Balancing the need for nitrogen to maintain agricultural production, and concurrently mitigating gaseous emissions of N_2O and ammonia (NH_3) requires a thorough understanding of pathways, and the mechanisms which regulate them. Equally a similar understanding is also needed for many other ecosystems, from peatlands to tropical forests, to understand potential climate responses.

The need for nitrogen

Nitrogen plays a vital role in plant growth; a deficiency in nitrogen and plants cannot thrive, leading to low crop yields; but an excess of nitrogen can be toxic to plants and have detrimental effects on the environment. Nitrogen is a key component of plant proteins (as a constituent of amino acids, the building blocks of proteins) and chlorophyll (the most important pigment needed for photosynthesis), which significantly impacts plant productivity. Nitrogen is a naturally occurring gaseous element, accounting for approximately 78% of the Earth's atmosphere. However, plants are unable to utilize the element in this form (N_2). To be available to plants, nitrogen must be transformed through the nitrogen fixation process to the reduced forms of this element.

Nitrogen can be produced naturally through three main pathways:

- I) Lightning - A small amount of nitrogen is fixed when lightning provides the energy required for atmospheric nitrogen (N_2) to react with oxygen, producing nitrogen oxide (NO) and nitrogen dioxide (NO_2). These forms of nitrogen then enter the soil system through atmospheric deposition in rain or snow.
- II) Biological - Nitrogen fixation also occurs naturally in the soil, carried out by a specialized group of prokaryotes. These organisms utilize the enzyme nitrogenase to convert atmospheric nitrogen to NH_3 (NH_3). Another important process occurs when legumes are present. Legumes are able to form symbiotic relationships with nitrogen fixing bacteria called rhizobia. This symbiosis results in nodule formation on the plant root, within which the bacteria can convert atmospheric nitrogen into NH_3 which can be used by the plant.
- III) Organic matter decomposition - Dead organic matter on and/or in the soil profile is decomposed by a wide range of microorganisms, called decomposers. These organisms consume and break down the organic matter and the nitrogen that is subsequently released is converted to plant accessible inorganic forms.

However, even with natural nitrogen fixation, low nitrogen concentrations in soils still often limit plant growth. When this happens, plant accessible forms of nitrogen must be added manually to ensure successful plant growth. Artificial nitrogen fixation, through the Haber-Bosch process, is the main industrial method for the production of NH_3 . This is where atmospheric nitrogen can be combined with hydrogen (H^+) to make anhydrous NH_3 and other nitrogen fertilizer products which are manually added to the soil. The use of nitrogen fertilizer has risen from 11 million tonnes in 1961 to 137 million tonnes in 2013, supporting substantial increases in crop production to feed a growing, global population. However, despite the clear successes of modern agriculture, an estimated 50% of applied nitrogen fertilizer is lost to the environment in water run-off from fields, animal waste and gaseous emissions from soil-microbe metabolism. These trade-offs have diverse and far-reaching impacts on public health, the environment and the economy. For example, excess nitrogen in the water causes algae to grow faster than ecosystems can handle (eutrophication). Significant increases in algae harm water quality, food resources and habitats, and decrease the oxygen that fish and other aquatic life need to survive. Excessive application of synthetic fertilizers has also been shown to acidify soils, damaging soil health and reducing the productivity of soils. Surplus nitrogen in the soil can also be oxidized and lost to the atmosphere as N_2O , which is a long-lived greenhouse gas that contributes to global warming. It persists in the atmosphere for an average of 114 years and is 300 times more potent than carbon dioxide. Mitigating these adverse effects requires an understanding of the dominant processes of soil nitrogen turnover.

Soil nitrogen transformations

Nitrogen compounds in soils

Soil nitrogen has three dominant forms: organic nitrogen, ammonium (NH_4^+) and nitrate (NO_3^-), the latter of which are the only ones accessible to plants for growth, although other forms of nitrogen appear as intermediaries in transformation pathways (for example nitrite, NO_2^-), or else are released as gaseous end products (for example N_2O) (Table 1). Examples of organic nitrogen include the nitrogen present in plant material (for example proteins and DNA), and waste material (for example manure and urea).

Nitrogen transformations are key to making nitrogen available for plant growth and include mineralization (the conversion of organic nitrogen to inorganic forms), nitrification (the conversion of NH_4^+ to NO_2^- and NO_3^-), denitrification (the conversion of NO_3^- to N_2O and nitrogen gas) (Fig. 1). However, a range of additional, and frequently less well characterized pathways, can also make potentially significant contributions to nitrogen turnover under certain prevailing soil conditions. Examples of such processes include Anammox, chemodenitrification and codenitrification. Most nitrogen transformations are microbially driven and are strongly influenced by a range of environmental variables including temperature, pH, and substrate availability. Understanding the precise controls over nitrogen turnover is important as it may ultimately allow the development of specific management practices that can help to mitigate emissions.

Nitrogen fixation and assimilation

Nitrogen fixation is the process by which gaseous nitrogen is converted to NH_3 or NH_4^+ through biological fixation, or NO_3^- through high energy physical processes (for example lightning, volcanic action and combustion). Atmospheric nitrogen is extremely stable, with a strong triple covalent bond, and a large amount of energy is required to break the bonds that hold the two N atoms. Lightning burns at 20,000 °C and is powerful enough to break the bonds of the nitrogen molecule in the atmosphere. Once split, the nitrogen atoms are able to bond with oxygen in the atmosphere, forming NO_2 and NO . These dissolve in precipitation, forming

Table 1 Dominant forms of nitrogen in soils.

Name	Chemical formula
Ammonia	NH_3
Ammonium	NH_4^+
Dinitrogen	N_2
Nitrate	NO_3^-
Nitric oxide	NO
Nitrite	NO_2^-
Nitrogen dioxide	NO_2
Nitrous oxide	N_2O
Organic nitrogen	R-NH_2

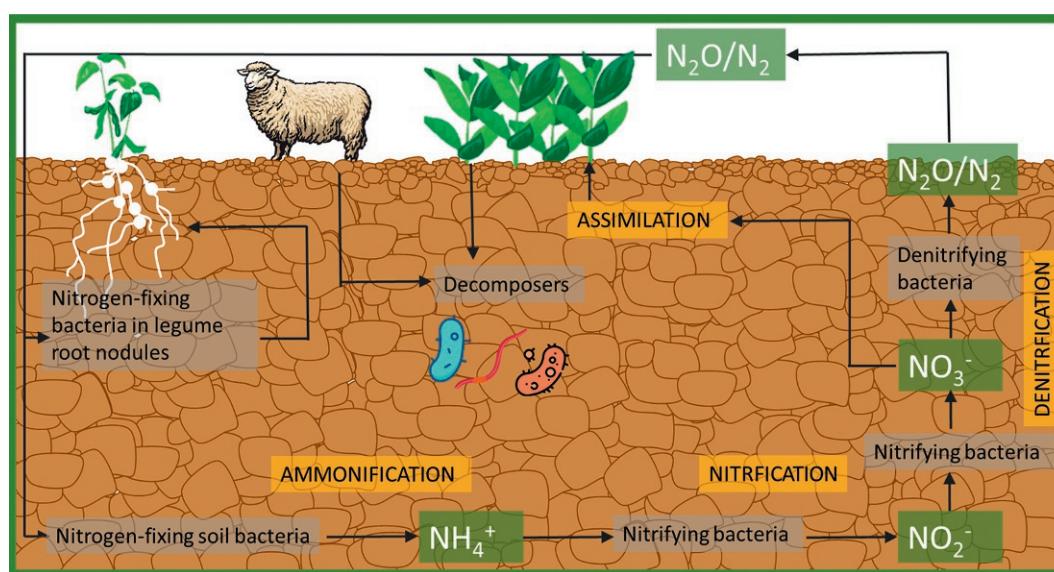


Fig. 1 Dominant soil nitrogen transformation pathways.

NO_3^- , which are carried down to the soil surface by atmospheric deposition (for example rain and snowfall). Atmospheric fixation contributes to an estimated 5–8% of the total nitrogen fixed.

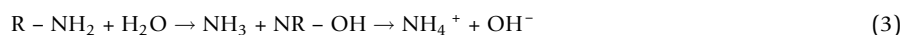
A greater amount of biologically available nitrogen, however, is naturally generated through the biological conversion of atmospheric nitrogen to $\text{NH}_3/\text{NH}_4^+$. Biological nitrogen fixation (BNF) results in nitrogen gaining electrons and is termed a reduction reaction (Eq. 1). The capability to fix nitrogen is found only in certain bacteria and archaea, and more than 90% of all nitrogen fixation is affected by these organisms. There are two main groups of nitrogen-fixing bacteria. The first are the free-living (non-symbiotic) bacteria, which includes the cyanobacteria and genera such as *Azotobacter* and *Clostridium*. The second group includes the mutualistic (symbiotic) bacteria, such as *Rhizobium*, which are associated with leguminous plants and *Azospirillum*, which are associated with cereal grasses. The *Rhizobia* interact with leguminous plants to form root nodules within which conditions are favorable for bacterial nitrogen fixation which the host plant can use. However, successful symbiosis requires the partners (plant and bacteria) to be compatible with each other throughout the process of symbiotic growth. Nonetheless, incompatibility regularly occurs, whereby a bacterial strain is unable to nodulate a certain host plant or form nodules that are able to fix atmospheric nitrogen.



The NO_3^- and NH_3 resulting from the nitrogen fixation process are assimilated into the specific tissue compounds of algae and plants. Animals then eat these algae and plants, incorporating them into their own body compounds.

Nitrogen mineralization and immobilization

Nitrogen mineralization is the process by which organic nitrogen present in materials ranging from manures to crop residues are transformed to inorganic (or mineral) forms, including NH_3 , NH_4^+ and NO_3^- . During amination, the first step in nitrogen mineralization, larger macromolecules of organic nitrogen compounds are degraded into smaller simpler organic compounds (for example amino acids, sugars and nucleic acids; Eq. 2). During subsequent ammonification, amino groups (R-NH_2) are converted to NH_3 , in a reaction carried out exclusively by heterotrophic microorganisms (Eq. 3). Rates of mineralization depend on soil temperature, moisture content, and aeration. Mineralization is generally coupled with immobilization which operates in the reverse direction. Immobilization can be achieved through either biotic or abiotic processes. During biotic immobilization (the dominant form of immobilization particularly in agricultural systems), NO_3^- and NH_4^+ are taken up by soil microorganisms making them unavailable for plant growth, and thereby returning them to organic compounds. This process, however, is temporary as when microorganisms die, organic nitrogen is gradually converted back to plant-available forms through mineralization and nitrification. Mean turnover times for microbial biomass nitrogen have been estimated at up to 2 months (Davidson et al., 1992). The availability of organic carbon to soil microorganisms for decomposition is generally considered to determine the immobilization capacity of soil nitrogen. Substrates with high carbon to nitrogen ratios (for example straw) can increase biological activity resulting in a greater microbial nitrogen demand, and thereby enhance rates of nitrogen immobilization. This underlines the close coupling between nitrogen and carbon cycling in soils. During abiotic immobilization, inorganic nitrogen can fix to clay minerals in a relatively rapid process. Abiotic immobilization is particularly common in forests and semi-arid grasslands, among other ecosystems (Moritsuka et al., 2004).



Ammonia and ammonium

The NH_3 and NH_4^+ forms exist in equilibrium in the soil, with the balance of each largely determined by soil pH. The relative concentration of NH_3 increases with increasing pH above 7, while NH_4^+ concentration decreases. Volatilization is the loss of nitrogen through the conversion of NH_4^+ to NH_3 gas. It occurs when NH_4^+ at the soil surface is converted to NH_3 gas at high soil pH. Rates of NH_3 volatilization increase at higher soil pH, as elevated H^+ concentrations promote the conversion of NO_3^- to NH_4^+ . Major factors that regulate NH_3 emissions include soil properties (primarily pH), temperature and soil moisture, and the timing and type of fertilizer application (primarily by affecting NH_4^+ concentrations). For example, NH_3 volatilization is greater from ammonium sulfate ($(\text{NH}_4)_2\text{SO}_4$), and urea applications than from NH_4NO_3 . Of these, urea is the most widely-used fertilizer, accounting for nearly 30% of all global fertilizer sales. When applied to soils, up to 50% can be volatilized. Globally, NH_3 volatilization is a major pathway of nitrogen loss and may account for up to approximately two-thirds of applied nitrogen from synthetic fertilizers. The NH_3 emissions from agricultural fields can be deposited in nearby ecosystems, driving eutrophication. Some NH_3 can also be oxidized and converted to nitric acid which contributes to acid rain. Alongside the adverse environmental impacts, ammonia volatilization represents a potentially substantial economic loss for farmers. As a result, a range of potential mitigation options have been proposed, including the use of non-urea based fertilizers, fertilization of lower soil horizons, combining applications with amendments, and the use of urease inhibitors (Pan et al., 2016).

Urease inhibitors are often applied with additions of urea to reduce ammonia emissions. These are generally structural analogs of urea and therefore act as competitive inhibitors for hydrolysis by soil urease. This delays the formation of NH_4^+ from urea, and thereby reduces the number of co-occurring areas of high pH and large NH_4^+ concentrations which are optimal for NH_3 volatilization (Harty et al., 2016).

Nitrification

Nitrification is the biological oxidation of NH_3 or NH_4^+ to NO_2^- and subsequently NO_3^- , occurring under aerobic conditions. In agricultural soils, NH_4^+ is rapidly converted to NO_3^- resulting in relatively low concentrations of NO_2^- . In general, NO_2^- does not accumulate in soils except under specific, temporary conditions where the activity of NO_2^- oxidizers is inhibited. This can occur following the application of high concentrations of urea or anhydrous NH_3 fertilizers that can inhibit NO_2^- oxidation due to their toxicity, or by lowering pH (Somers et al., 2019). Nitrification ultimately can lead to N_2O production, and NO_3^- leaching, and can result in the loss of up to 50% of plant available nitrogen in soils (Beeckman et al., 2018). Nitrogen losses of this magnitude result in agriculture being a major contributor to climate change, and the eutrophication of water bodies.

In autotrophic nitrification, ammonia oxidizing bacteria (AOB) convert NH_4^+ to hydroxylamine and then to NO_2^- (Eq. 4). Examples of AOB include *Nitrosomonas* and *Nitrospira*. Ammonia oxidizing archaea (AOA), examples of which include *Nitroso-sphaera*, can also oxidize NH_4^+ to NO_2^- although by a distinct metabolic pathway. The NO_2^- oxidizing bacteria, such as *Nitrobacter* and *Nitrospira* subsequently convert NO_2^- to NO_3^- (Eq. 5). Some *Nitrospira* bacteria have also been found to undertake complete NH_3 oxidation to NO_3^- , in a process known as comammox (Van Kessel et al., 2015). Some heterotrophs have also been reported to carry out nitrification, albeit at lower rates, and primarily in acidic soils (Sahrawat, 2008).



The rate of nitrification in soils is determined by a combination of substrate availability, environmental conditions (for example pH and oxygen availability), and the abundance of specific microbial communities. The availability of NH_3 and/or NH_4^+ often limits the rate of nitrification. Similarly, the absence of oxygen (for example in highly waterlogged soils) can limit the abundance of nitrifier communities. Nitrification occurs within a narrow range of pH (6.6–8.0) as the process is restricted to a limited range of microbial species, with a strong decline below pH 5.5. High rates of nitrification, particularly due to fertilizer applications and leaching of NO_3^- can result in the acidification of soils, necessitating the management of agricultural soils through liming. Nitrification is sensitive to temperature changes across a range of soil types, with optimum temperatures specific to a given environment (Norton and Ouyang, 2019).

Nitrification inhibitors are frequently combined with fertilizer applications to delay the oxidation of NH_4^+ by the soil nitrifier community. This results in a temporary accumulation of NH_4^+ and reduces losses of nitrogen through denitrification and leaching, thereby increasing the efficiency of fertilizer applications. Dicyandiamide (DCD) is one of the most widely used nitrification inhibitors in agriculture because it is relatively cheap, and water-soluble thereby allowing easy application (Harty et al., 2016). The benefits of nitrification inhibitors are partially dependent on the environment (for example pH and texture) and other additional management factors (for example nitrogen fertilization rate).

Denitrification

Denitrification involves the reduction of NO_3^- and NO_2^- to nitric oxide, N_2O and N_2 , under anaerobic conditions. Alongside annamox, it is one of the few pathways by which reactive forms of nitrogen can be converted back into nitrogen gas. In general, denitrification requires available electron donors, such as organic carbon compounds, and available nitrogen oxides (for example NO_3^-) as terminal electron acceptors. The rate of denitrification, and the ratios of the main products (N_2O and nitrogen gas) are largely controlled by interactions between soil properties (particularly availability of mineral nitrogen and labile carbon) and microbial communities, and further mediated by management and climate (particularly temperate which affects microbial activity). Complete denitrification to nitrogen gas is driven by high soil temperatures and water contents, neutral to slightly basic soil pH, low rates of O_2 diffusion and the presence of a labile carbon source (Saggar et al., 2013).

During denitrification, nitrogen oxides are reduced following the addition of two electrons per nitrogen atom, with each step regulated by distinct reductase enzymes coded for by specific genes. The NO_3^- is reduced to NO_2^- by nitrate reductase (Nar) before further reduction by nitrite reductase (Nir) (Eqs. 6 and 7). Nitric oxide is then reduced by nitric reductase (Nor) encoded by the *cnorB* or *qnorB* gene, resulting in N_2O production (Eq. 8). Nitric oxide reductase (Nos) further reduces this to nitrogen gas (Eq. 9). Denitrification genes are named for their enzyme products. Nar is regulated by *narG* or *napA* genes; Nir is regulated by *nirK* or *nirS* genes; Nor is regulated by *cnorB* or *qnorB* gene; and Nos is regulated by *nosZ*. The abundance and activity of *nirS*, *nirK* and *nosZ* genes has previously been proposed as a potential indicator for overall denitrification activity in soils, with high ratios of (*nirK* + *nirS*)/*nosZ* gene copies associated with increased potential N_2O production. Genes involved in denitrification are activated sequentially, with Nar activated within 2–3 h, Nir between 2 and 12 h, and Nor between 24 and 48 h, in response to the addition of carbon substrate under anaerobic conditions. Each enzyme in denitrification requires various metal cofactors. For example, Nor requires copper (Cu), highlighting the important role of the element in denitrification.



Denitrifiers are diverse and include bacteria, archaea, and eukarya, with over 60 genera of bacteria and archaea alone identified as capable of denitrification. All eukaryotic denitrifiers, and the majority of prokaryotic denitrifiers are heterotrophs. However, many lack one or more enzymes required for complete denitrification. Most fungi and a third of bacteria lack Nor, resulting in substantial N_2O production versus nitrogen gas (Philippot et al., 2011).

A range of environmental factors has been identified as regulating soil denitrification, including NO_3^- concentrations, oxygen availability (through determining redox conditions), carbon availability, and temperature (O'Neill et al., 2020). Soil NO_3^- concentrations depend on mineralization and nitrification rates, the presence of plants (i.e. plant nitrogen uptake), immobilization, and NO_3^- losses through leaching and diffusion. Soil NO_3^- is also an important factor in regulating $\text{N}_2\text{O}:\text{N}_2$ ratios. Higher NO_3^- concentrations usually result in increased $\text{N}_2\text{O}:\text{N}_2$ ratios due to the suppression of Nos at higher concentrations (Saggar et al., 2013).

As denitrification can only occur under anoxic conditions, oxygen availability, and by extension soil moisture, is a key factor in regulating rates. Below certain thresholds, largely dependent on soil type, denitrification rates appear unrelated to soil water content, but thereafter rapidly increase with increasing soil water. Low rates with increased oxygen concentrations arise from the inhibition of the reductase enzymes. The ratio of $\text{N}_2\text{O}:\text{N}_2$ also generally decreases with increasing soil water content/decreasing soil oxygen, with 75% water-filled pore space identified as an important threshold. Management practices that impact soil oxygen supply and/or soil moisture can therefore have a significant impact on rates. Compaction from grazing or the use of heavy farm machinery can restrict soil oxygen diffusion, thereby enhancing denitrification. Similarly, rainfall events can affect rates of soil denitrification directly, through altering soil moisture content, but also indirectly by increasing soil carbon mineralization and releasing trapped soil gases.

The optimum pH for denitrification is between 7 and 8 but can still occur at more acidic pH. Denitrification has also been reported at pH values as low as 3.5 and can therefore still account for significant nitrogen losses in naturally acidic soils. At lower soil pH the $\text{N}_2\text{O}:\text{N}_2$ ratio is increased, although the underlying mechanisms remain to be clarified. Nos activity has been found to be reduced at lower pHs, and possibly reflects increased fungal dominance. Increasing soil pH may therefore represent an important mechanism to reduce field emissions of N_2O , although Nor is increasingly inhibited, resulting in increased production of N_2O versus N_2 (i.e. increased $\text{N}_2\text{O}:\text{N}_2$ ratios). Heterotrophic denitrification is frequently limited by the availability of labile carbon. Consequently, processes that influence the rate of carbon mineralization and/or immobilization in soils (such as drying and rewetting cycles, residue or fertilizer additions or root exudates) can significantly affect rates of denitrification. Generally, increasing soil organic carbon decreases $\text{N}_2\text{O}:\text{N}_2$, although the effect of carbon availability on fluxes has also been shown to depend on both NO_3^- concentration and soil moisture content. Rhizosphere soils have been suggested as having reduced $\text{N}_2\text{O}:\text{N}_2$, most likely resulting from increased available carbon from root exudate input, although other factors including soil pH and oxygen availability may be involved.

Denitrification occurs across a range of temperatures, with optimal temperatures of approximately 30 °C. However, as the denitrifying microbial communities can adapt to changes in soil temperature, different regions have been reported to have varying optima. In addition, directly regulating rates of microbial activity, temperature can indirectly affect denitrification through its influence on gas solubility and diffusion rates, and availability of the required substrates. The activation energy for N_2O reduction is greater than that for N_2O production, and therefore some studies have reported decreasing $\text{N}_2\text{O}:\text{N}_2$ ratios with increasing soil temperatures, although this is not a universal finding (Wang et al., 2021).

Codenitrification

Stable isotope experiments with ^{15}N tracers have indicated the potential importance of codenitrification as an additional, although understudied, nitrogen transformation pathway (Clough et al., 2017). During codenitrification (also referred to as biotic nitrogen-nitrosation), organic nitrogen compounds (for example amines) are co-metabolized with an inorganic nitrogen source to produce N_2O and gaseous nitrogen under neutral to alkaline conditions. During codenitrification, the hydrogen atom in the organic compound is replaced with a nitroso group ($-\text{N}=\text{O}$) and is catalyzed by enzymatic nitrosyl compounds which attract a variety of nucleophilic co-metabolites including hydroxylamine, NH_4^+ , hydrazine, amino compounds, and NH_3 . The products contain one nitrogen atom from both the organic and inorganic nitrogen sources. Significant rates of codenitrification are only likely to occur when nucleophile concentrations are at least one order of magnitude greater than that of NO_2^- or nitric oxide (Spott et al., 2011). Theoretically, the environmental controls over denitrification also apply to codenitrification, although there have been relatively few studies conducted to date which demonstrate this. The contribution of codenitrification to soil N_2O fluxes has been estimated at between 13% and 90% (Clough et al., 2017; Selbie et al., 2015).

Chemodenitrification

Chemical denitrification, termed chemodenitrification, is the abiotic reduction of inorganic nitrogen under anaerobic conditions. During the process, NO_2^- reacts with ferrous iron (Fe^{2+}) and water to produce N_2O and Fe^{3+} . The presence of Fe^{2+} is driven by heterotrophic Fe^{3+} reducing microorganisms, as well as by the availability of NO_2^- , produced during the reduction of NO_3^- during denitrification, thereby underlying the close relationship between iron redox cycling and the nitrogen cycle. The extent of N_2O

production by chemodenitrification compared to denitrification is relatively poorly understood although it has been suggested that it may account for between 31% and 75% of total N_2O production in agricultural soils (Venterea, 2007).

Dissimilatory nitrate reduction to ammonium

Dissimilatory nitrate reduction to ammonium (DNRA) is an aerobic form of respiration by a number of groups of heterotrophic bacteria and some fungi. These bacteria use NO_3^- as a terminal electron acceptor when oxidizing soil organic matter. In the wider nitrogen cycle, this pathway therefore bypasses denitrification and nitrogen fixation by producing NH_4^+ . However, the process is often overlooked in descriptions of the soil nitrogen cycle, primarily because it is generally considered to occur only under highly reducing soil conditions (for example in flooded environments). The ^{15}N stable isotope tracing, however, suggests that DNRA can be the dominant NO_3^- consumption process under certain circumstances.

Anammox

While the majority of nitrification is carried out under aerobic conditions, anaerobic NH_4^+ oxidation (anammox) can also occur, mediated by a limited number of bacteria within the phylum *Planctomycetes* (Strous et al., 1999). Although initially identified in a wastewater treatment process, anammox bacteria have subsequently been reported across a range of soil types and soil conditions, including rice paddies and arable upland soils. The NH_4^+ and nitrogen dioxide are converted to nitrogen gas and water under anoxic conditions. First, NO_2^- is reduced to nitric oxide by nitrite oxidoreductase (nir), before being combined with NH_4^+ to form hydrazine by hydrazine synthase (hzs). This is oxidized by hydrazine oxidoreductase (hzo) to dinitrogen gas. Relationships between rates of nitrogen loss through anammox and specific management interventions (for example fertilizer applications) remain poorly understood compared to other nitrogen transformation pathways. Anammox bacteria can tolerate large soil NH_4^+ concentrations and may be an important nitrogen loss pathway under certain circumstances, for example ranging from 3.15–9.62% in a rice–wheat rotation (Gu et al., 2017).

Measuring and monitoring human impacts on nitrogen cycles

Anthropogenic impacts on the soil nitrogen cycle

Anthropogenic activities have dramatically increased concentrations of N_2O in the atmosphere, from 291 ppb in 1960 to 330 ppb in 2017, and are greatly increasing the amount of nitrogen cycling between soil, water, and the atmosphere (Fig. 2). In many ecosystems, the supply of nitrogen is a key factor controlling the nature and diversity of plant life, the population dynamics of grazing animals and their predators and essential ecological activities such as plant productivity and the cycling of carbon and soil minerals. Excessive nitrogen can pollute ecosystems (for example eutrophication), drive soil acidification and result in substantial in gaseous emissions, collectively resulting in a change ecological functioning and potentially the communities they support.

Soil acidification is a major problem identified with the excessive addition of soil nitrogen. Acidification can result in the leaching of cationic nutrients including potassium (K^+), sodium (Na^+), calcium (Ca^+) and magnesium (Mg^+). Collectively, this leads to a reduction in plant productivity, and can affect the balance of nitrogen transformation pathways and therefore gaseous emissions. The NO_3^- , as a highly mobile compound that can leach downwards into the groundwater, particularly if the groundwater is shallow and the underlying soil type is sandy. Leaching into larger water bodies can result in eutrophication resulting in excessive growth of weeds and algae, and can contaminate drinking water supplies.

Most anthropogenic effects on the nitrogen cycle are highly local in scale, although can have global effects. For example, studies in pasture systems have frequently reported that most nitrogen losses through denitrification occur following fertilizer additions, primarily because nitrogen availability represents the most important limitation on denitrification. There is therefore significant

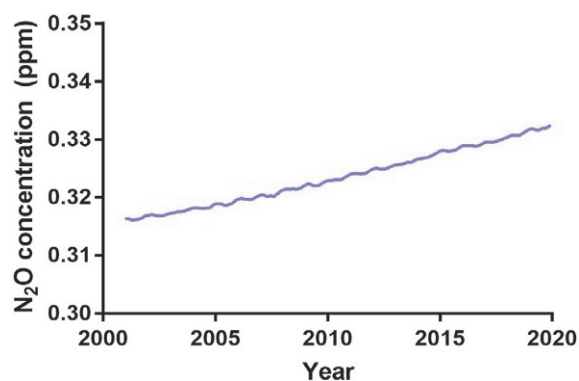


Fig. 2 Rise in atmospheric N_2O concentrations (ppm). Data from Dlugokencky (2020).

spatial variation in N_2O emissions, particularly under grazing where urine and manure inputs lead to emission hotspots. The extent of losses are in large part determined by the quantity of nitrogen applied (for example if it is being applied in excess of crop demand), the type (for example urea or NH_4NO_3), and the inclusion of any inhibitors of nitrogen turnover. An example of a widely used nitrification inhibitor is dicyandiamide (DCD).

In addition to agriculture, other local activities that include burning of fossil fuels, power generation plants, and industries can also be a significant source of atmospheric nitrogen. Large applications, often in excess of agronomic need, have resulted in a surplus of nitrogen with major and long-term environmental consequences for large regions of the Earth.

Quantification of soil nitrogen transformations and emissions

A range of methodologies has been proposed for quantifying key soil nitrogen transformation pathways, but key challenges remain in terms of distinguishing dominant transformation pathways (Fig. 3). Denitrification is particularly difficult to quantify because of considerable spatial and temporal variation, and challenges in quantifying its end-product (N_2). In particular, the 'hotspots' and 'hot moments' concept, whereby relatively small areas become substantial $\text{N}_2\text{O}/\text{N}_2$ sources for brief periods of time can make capturing emissions challenging without continual measurements. Difficulties in quantifying processes and emissions result in subsequent problems in modelling and scaling fluxes.

The use of ^{15}N stable isotope labelling has been widely adopted in the literature as a standard technique for assessing soil nitrogen turnover. Across two studies in 1954–1955, Kirkham and Bartholomew developed the isotope pool dilution technique allowing the quantification of nitrogen transformation rates. When combined with modern ^{15}N tracing models, the approach can be crucial for disentangling previously poorly understood pathways (for example codenitrification). In addition to quantifying nitrogen transformation rates, ^{15}N labelling is also useful for partitioning N_2O emission sources (for example the balance of nitrification versus denitrification). The ^{15}N labelling also allows the quantification of soil-derived N_2 , against high atmospheric background concentrations. This is particularly important in the context of mitigating substantial GHG emissions from agriculture, as it allows the assessment of the impact of different practices on $\text{N}_2\text{O}:\text{N}_2$, and therefore the identification of management approaches that can divert N_2O to N_2 (Van Groenigen et al., 2015). Alongside ^{15}N tracing, the use of oxygen isotopes (^{18}O) as an additional tracer allows the further separation of pathways, specifically NH_4^+ derived N_2O emissions from NH_4^+ oxidation versus nitrifier–denitrification (Kool et al., 2011). Acetylene (C_2H_2) has also been used to quantify gaseous nitrogen emissions because it can inhibit denitrification. This allows the indirect estimation of N_2 production without the use of stable isotope techniques which can be prohibitively expensive (Felber et al., 2012). However, there is a range of limitations including incomplete inhibition due to diffusion barriers in soils, suppressed microbial respiration (Groffman et al., 2006). Such techniques can be particularly powerful when coupled with characterizations of soil microbial community structure and function. For example, quantification of denitrifiers can be achieved using real-time quantitative PCR, focussing on nirK, nirS, and nosZ (Jones et al., 2014).

Several studies have undertaken direct field measurements of gaseous nitrogen emissions. Many have used static chambers because of the ease of measurement and potential for large scale replication. However, static chambers are generally poor at capturing the temporal and spatial variation of emissions, resulting in over- or under-estimates of N_2O emissions. Alternative approaches are predominantly based on micro-meteorological techniques, the most well-known and widely-used example of which is eddy covariance. This approach allows the continuous measurement of N_2O fluxes at ecosystem scales, which theoretically provides additional insights into factors that regulate N_2O emissions, and the underlying soil nitrogen transformation processes (Liang et al., 2018). Satellite derived data are becoming increasingly available, including for nitrogen emissions, offering an important opportunity for helping to close emissions budgets.

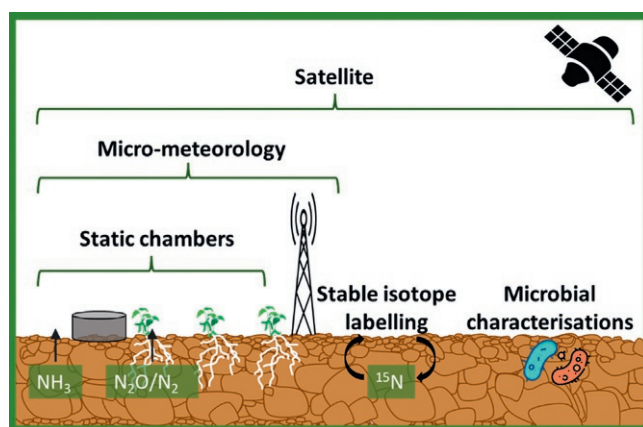


Fig. 3 Methodologies for quantifying soil nitrogen transformations and emissions.

Future research directions and conclusions

Research priorities for soil nitrogen

Ongoing environmental issues continue to arise from large inputs of nitrogen in excess of crop and soil demand, and potential feedbacks from climate change may substantially impact the terrestrial nitrogen cycle. Continued research focussing on developing more effective nitrogen management, particularly in agriculture, is therefore urgently required. Potential practices of note include those often grouped as regenerative agricultural management, encompassing organic nitrogen fertilizers, biofertilizers, the incorporation of crop residues and the use of biochar. Many terrestrial systems also remain relatively understudied in terms of the nitrogen cycle. Examples include tropical wetlands, which evidence suggests may be a substantial N_2O source (D'Amelio et al., 2009). While the majority of research on soil nitrogen transformation pathways continues to focus on nitrification and denitrification, the importance of other processes (for example codenitrification) remain relatively understudied, despite evidence suggesting they may make a substantial contribution to nitrogen turnover (Spott et al., 2011). Moreover, our understanding of species and individual plant-level controls over nitrogen cycling (for example the contributions of root turnover and root exudation of carbon and nitrogen to nitrous oxide emissions) remains relatively limited in a number of ecosystems. Finally, potential climate feedbacks must also be considered. This includes understanding the combined effects of warming and changes in precipitation, and alterations in soil moisture regimes across a wide range of ecosystems, from agriculture to forests (Wang et al., 2021).

Conclusions

Nitrogen is among the most important nutrient for plant growth and as such is heavily applied to agricultural soils to maintain crop yields. Application rates, however, are frequently in excess of crop demand, resulting in large rates of loss and negative environmental impacts. Mitigating these adverse effects requires a detailed understanding of the nitrogen transformation processes that occur in soil systems. Soil nitrogen turnover is dominated by a series of processes beginning with mineralization whereby organic forms of nitrogen are made available for plant uptake. Nitrification and denitrification convert NH_4^+ and NH_3 to NO_2^- , NO_3^- and ultimately N_2O and N_2 gases. During immobilization, microbial communities can take inorganic forms and fix them in the microbial biomass. A growing number of techniques, often coupled to new modelling approaches, are now available and are increasingly being used to understand both the scale of nitrogen turnover in soils, and opportunities for mitigation.

References

- Beeckman F, Motte H, and Beeckman T (2018) Nitrification in agricultural soils: Impact, actors and mitigation. *Current Opinion in Biotechnology* 50: 166–173.
- Clough TJ, Lanigan GJ, De Klein CA, Samad MS, Morales SE, Rex D, Bakken LR, Johns C, Condon LM, Grant J, and Richards KG (2017) Influence of soil moisture on codenitrification fluxes from a urea-affected pasture soil. *Scientific Reports* 7(1): 1–12.
- D'Amelio MTS, Gatti LV, Miller JB, and Tans P (2009) Regional N_2O fluxes in Amazonia derived from aircraft vertical profiles. *Atmospheric Chemistry and Physics* 9: 8785–8797. <https://doi.org/10.5194/acp-9-8785-2009>.
- Davidson EA, Hart SC, and Firestone MK (1992) Internal cycling of nitrate in soils of a mature coniferous forest. *Ecology* 73(4): 1148–1156.
- Dlugokencky E (2020) NOAA/GML Trends in Atmospheric Nitrous Oxide. gml.noaa.gov/ccgg/trends_n2o. Accessed 20th May 2020.
- Felber R, Conen F, Flechard CR, and Neftel A (2012) Theoretical and practical limitations of the acetylene inhibition technique to determine total denitrification losses. *Biogeochemistry* 9(10): 4125–4138.
- Groffman PM, Altabet MA, Bohike JK, Butterbach-Bahl K, David MB, Firestone MK, Giblin AE, Kana TM, Nielsen LP, and Voytek MA (2006) Methods for measuring denitrification, Diverse approaches to a difficult problem. *Ecological Applications* 16: 2091–2122.
- Gu C, Zhou HF, Zhang QH, et al. (2017) Effects of various fertilization regimes on abundance and activity of anaerobic ammonium oxidation bacteria in rice-wheat cropping systems in China. *Science of the Total Environment* 599: 1064–1072.
- Harty MA, Forrestal PJ, Watson CJ, McGeough KL, Carolan R, Elliot C, Krol D, Laughlin RJ, Richards KG, and Lanigan GJ (2016) Reducing nitrous oxide emissions by changing N fertiliser use from calcium ammonium nitrate (CAN) to urea based formulations. *Science of the Total Environment* 563: 576–586.
- Jones CM, Spor A, Brennan FP, Breuil M-C, Bru D, Lemanceau P, Griffiths B, Hallin S, and Philippot L (2014) Recently identified microbial guild mediates soil N_2O sink capacity. *Nature Climate Change* 4: 801–805.
- Kool DM, Dolfing J, Wrage N, and Van Groenigen JW (2011) Nitrifier denitrification as a distinct and significant source of nitrous oxide from soil. *Soil Biology and Biochemistry* 43: 174–178. <https://doi.org/10.1016/j.soilbio.2010.09.030>.
- Liang LL, Campbell DI, Wall AM, and Schipper LA (2018) Nitrous oxide fluxes determined by continuous eddy covariance measurements from intensively grazed pastures: Temporal patterns and environmental controls. *Agriculture, Ecosystems & Environment* 268: 171–180.
- Moritsuka N, Yanai J, Mori K, and Kosaki T (2004) Biotic and abiotic processes of nitrogen immobilization in the soil-residue interface. *Soil Biology and Biochemistry* 36(7): 1141–1148.
- Norton J and Ouyang Y (2019) Controls and adaptive management of nitrification in agricultural soils. *Frontiers in Microbiology* 10: 1931.
- O'Neill RM, Girkinn NT, Krol DJ, et al. (2020) The effect of carbon availability on N_2O emissions is moderated by soil phosphorus. *Soil Biology and Biochemistry* 142: 107726.
- Pan B, Lam SK, Mosier A, Luo Y, and Chen D (2016) Ammonia volatilization from synthetic fertilizers and its mitigation strategies: A global synthesis. *Agriculture, Ecosystems & Environment* 232: 283–289.
- Philippot L, Andert J, Jones CM, Bru D, and Hallin S (2011) Importance of denitrifiers lacking the genes encoding the nitrous oxide reductase for N_2O emissions from soil. *Global Change Biology* 17(3): 1497–1504.
- Saggar S, Jha N, Deslippe J, et al. (2013) Denitrification and N_2O : N_2 production in temperate grasslands: Processes, measurements, modelling and mitigating negative impacts. *Science of the Total Environment* 465: 173–195.
- Sahrawat KL (2008) Factors affecting nitrification in soils. *Communications in Soil Science and Plant Analysis* 39(9–10): 1436–1446.
- Selbie DR, Lanigan GJ, Laughlin RJ, et al. (2015) Confirmation of co-denitrification in grazed grassland. *Scientific Reports* 5(1): 1–9.

- Somers C, Girkin NT, Rippey B, Lanigan GJ, and Richards KG (2019) The effects of urine nitrogen application rate on nitrogen transformations in grassland soils. *The Journal of Agricultural Science* 157(6): 515–522.
- Spott O, Russow R, and Stange CF (2011) Formation of hybrid N_2O and hybrid N_2 due to codenitrification: First review of a barely considered process of microbially mediated N-nitrosation. *Soil Biology and Biochemistry* 43: 1995–2011.
- Strous M, Fuerst JA, Kramer EHM, et al. (1999) Missing lithotroph identified as new planctomycete. *Nature* 400: 446–449.
- Van Groenigen JW, Huygens D, Boeckx P, Kuyper TW, Lubbers IM, Rütting T, and Groffman PM (2015) The soil N cycle: New insights and key challenges. *The Soil* 1(1): 235–256.
- van Kessel MAHJ, Speth DR, Albertsen M, Nielsen PH, Op den Camp HJM, Kartal B, et al. (2015) Complete nitrification by a single microorganism. *Nature* 528: 555–559. <https://doi.org/10.1038/nature16459>.
- Venterea RT (2007) Nitrite-driven nitrous oxide production under aerobic soil conditions: Kinetics and biochemical controls. *Global Change Biology* 13(8): 1798–1809.
- Wang C, Amon B, Schulz K, and Mehdi B (2021) Factors that influence nitrous oxide emissions from agricultural soils as well as their representation in simulation models: A review. *Agronomy* 11(4): 770.

Phosphorus in soil

Luisella Celi, Maria Martin, and Elisabetta Barberis, Soil and Biogeochemistry, Department of Agricultural, Forest and Food Sciences, Grugliasco, Torino, Italy

© 2023 Elsevier Ltd. All rights reserved.

Introduction	153
Phosphorus forms in soils	153
Abiotic processes controlling P availability	154
Adsorption and desorption reactions	156
Solubilization and precipitation reactions	157
Biotic processes	158
Microbial immobilization and mineralization	158
Plant nutrition and strategies of plants to acquire P	159
Losses of P from soil	162
Phosphorus resources and global fluxes	163
Concluding remarks	164
References	165

Abstract

Phosphorus (P) in soil derives from weathering of primary minerals, such as apatites, present in almost all types of rocks, globally ensuring its availability for plants. In natural systems, P availability depends on a concerted interaction between abiotic processes – sorption/desorption, precipitation/dissolution, leaching and runoff – and biological processes – microbial mineralization/immobilization, plant uptake and organic P recycling–. However, the P cycle has been largely modified by anthropic actions, leading to an uneven P distribution in global soils that can increase the global imbalance in P-cycling and food production. The sustainability of future agricultural production requires that P recovery from wastes is improved, losses to water systems are limited and the availability of legacy soil P is enhanced.

Glossary

1:1 Al silicates Secondary phyllosilicates whose crystals comprise an alternating tetrahedral sheet (silicon and oxygen ions) bonded to an octahedral sheet (oxygen, aluminum, and hydroxyl ions) with a 1:1 ratio.

Fe and Al oxi(hydr)oxides Different pools of minerals including oxides, hydroxides and oxy-hydroxides of iron (Fe) and aluminum (Al).

Inner-sphere complex Type of surface complex with direct bonding between the surface and the adsorbed molecule, without interposition of water molecules.

Mycorrhizae Symbiotic association between fungal hyphae and plant roots. Ectomycorrhizae, the fungal hyphae do not enter the root cells; arbuscular mycorrhizae, the fungal hyphae enter plant cells forming vesicles and ramified structures for nutrient exchange; ericoid mycorrhizae, intraradical coils of hyphae in the outermost layer of root cells.

Myo- and scyllo-inositol hexaphosphate Two isomers of phytic acid, general formula $C_6H_{18}O_{24}P_6$, a P reservoir in bran and seeds.

Outer-sphere complex Type of surface complex where the adsorbate retains its hydration shell, water molecules are interposed between the surface and the adsorbate which is retained through electrostatic interaction.

Point of zero charge The pH value at which the net total charge of the surface of a colloidal particle is equal to zero.

Key points

- Soil P is essential for life
- Abiotic processes dominate in P rich soils while biological processes become essential in P poor soils
- The P cycle has been largely modified by anthropic actions
- Global imbalance in P stocks, cycling and food production need to find new integrated solutions that improve P recovery from wastes, limit losses to water systems and enhance the availability of legacy soil P

Introduction

The cycle of phosphorus (P) in ecosystems differs from those of carbon (C) and nitrogen (N), which have much faster cycles. The dynamics of P can be treated as one-way traffic from rock deposits to waters and sediments, as a full cycle would take millions of years to complete. Phosphorus in soil derives from weathering of primary minerals, such as apatites, present in almost all types of rocks, globally ensuring its availability for plants. Phosphorus is essential to all forms of life and constitutes about 0.2–0.5% of plant dry weight, where it is primarily a component of tissue molecules such as nucleic acids, phospholipids, and inositol phosphates. After nitrogen, P is the second most limiting nutrient: it can reduce plant growth and development and potentially limit crop yield. In natural systems, P availability depends on a concerted interaction between abiotic processes such as mineral weathering, sorption/desorption, complexation, precipitation-coprecipitation/dissolution, leaching and runoff of P compounds bound to soil colloidal surfaces, and biological processes including microbial mineralization/immobilization, mobilization of organic P, and plant uptake and recycling. However, the P cycle has been largely modified by anthropic actions, including mineral and organic fertilization. In many agricultural soils this has caused an excess of P, which can be detrimental to the environment because P can enter freshwater bodies through surface runoff and cause algal bloom reducing water quality. On the other hand, the large use of P fertilizers implies a flux of P from countries rich in P ores or deposits to developed areas, increasing the global imbalance in P stocks, cycling and in crop productivity. This highlights the need to find new integrated solutions that can improve not only P cycling in anthropic systems, but must also consider the mining and industrial processes of P fertilizer, the food/feed chain production, and management of P-containing wastes.

Phosphorus forms in soils

In the soil–water–plant–atmosphere continuum of terrestrial ecosystems over 90% of the P is stored in soil, while freshwaters and the atmosphere contain only very small P concentrations. In plant tissues, P concentrations are less than those of nitrogen (N), potassium (K) and calcium (Ca).

The weathering of phosphate minerals represents the primary source of P in soils. This source is only slowly renewable, since the slow abiotic cycle closes with soil renewal as a consequence of pedogenetic factors, or with tectonic uplift, that bring new minerals to the soil in geological timescales. In contrast to the slow geochemical cycle, internal biotic cycling is much faster, leading to the buildup of sparingly available organic P forms. Therefore, the disruption of P cycling in agricultural soils can lead to long-term P depletion of the system (Fig. 1).

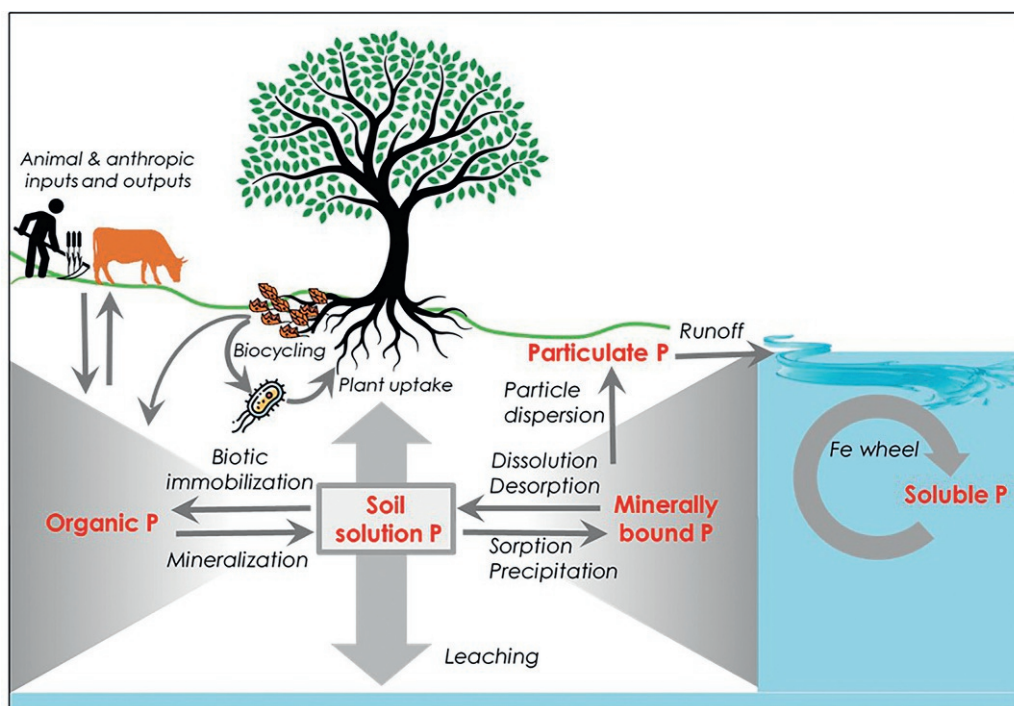


Fig. 1 Phosphorus cycle showing the flux of P from primary minerals into the soil solution; abiotic processes including adsorption/desorption and precipitation/dissolution; biotic processes, including microbial immobilization/mineralization and plant uptake; anthropic actions, including crops and livestock, that change P inputs/outputs; runoff and leaching with transfer of P from soil to waters.

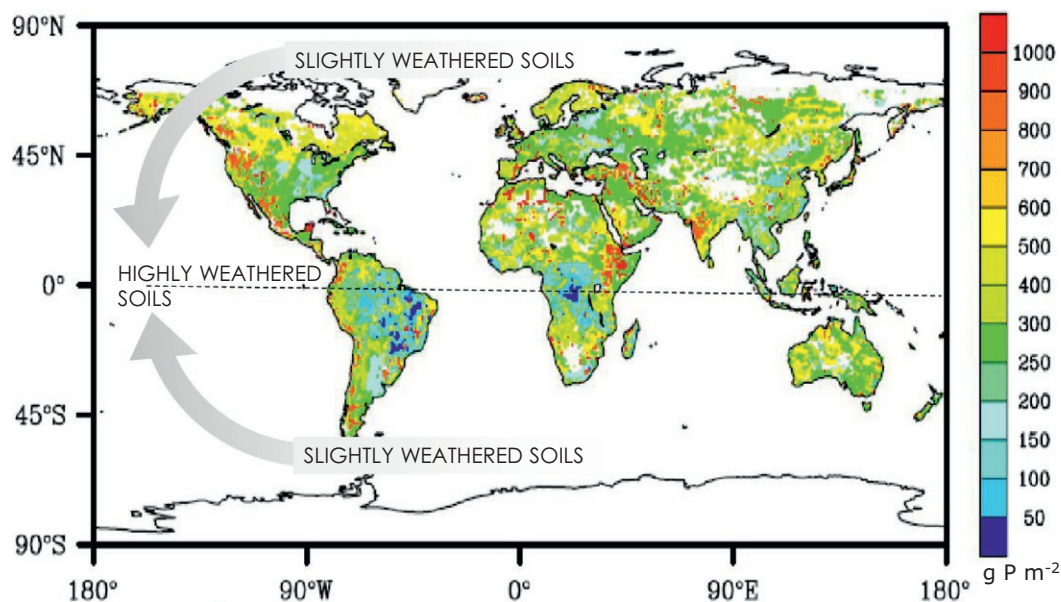


Fig. 2 The global distribution of total P in soil showing a latitudinal gradient with decreasing content from high to low latitudes (unit: g P m^{-2}). Adapted from Yang X, Post WM, Thornton PE, and Jain A (2013) The distribution of soil phosphorus for global biogeochemical modeling. *Biogeosciences* 10: 2525–2537.

Soil phosphorus content can vary widely, in the range $50\text{--}3000 \text{ mg kg}^{-1}$, depending on the type of parent material, pedogenesis, and land use. In general, soil P content is lowest in tropical regions, where soils (i.e. Ultisols) have gone through millions of years of soil development and have lost most of their original P through leaching or erosion, and highest in Inceptisols and Entisols of colder regions (Yang et al., 2013; Fig. 2).

The proportion of organic and inorganic P forms varies with time and with soil environments; the organic fraction ranges between 20 and 80% of the total. It commonly increases in surface soil layers compared with deeper ones, together with the concentration of organic matter. It can become the dominant form in soils of moist, cool climates with thick organic layers, before the total P content begins to decline, and strongly fixed, poorly available forms become prevalent (Fig. 3). Conversely, in deeper soil layers, particularly in young soils of temperate and dry regions, inorganic P forms largely prevail.

Inorganic P mainly derives from the weathering of primary minerals, mostly in the form of apatite, a mineralogically complex compound with the empirical formula $\text{Ca}_{10}(\text{PO}_4)_X\text{X}_2$ ($X = \text{F}^-, \text{Cl}^-, \text{OH}^-, \text{or } \text{CO}_3^{2-}$). The content of Ca-phosphates in soil parent material can vary depending on the kind of bedrock. Within igneous rocks, total P content generally increases from silica-rich rocks, such as rhyolite and granite (median, $310\text{--}440 \text{ mg kg}^{-1}$), to iron-rich ones, such as basalts and alkali basalts (median, 920 to $>2000 \text{ mg kg}^{-1}$). However, across all rock types, ultramafic igneous rocks, such as peridotite (median, 120 mg kg^{-1}) have the lowest P content. Sedimentary rocks have P contents that decrease with increasing grain size ($1100\text{--}500 \text{ mg kg}^{-1}$), while carbonate rocks have generally lower P concentrations (Porder and Ramachandran, 2013). As mineral weathering and soil development proceeds, the P in primary apatites dissolves and, if not taken up by soil biota or leached to the water system, it is adsorbed or (co) precipitated with Fe and Al compounds, becoming in general sparingly soluble.

The fraction of organic P increases during the development of natural soil systems to a maximum, where it can encompass 80% of the total soil P, then it begins to decline, while changing its composition (Turner et al., 2007). The organic forms of P found in soil reflect the composition of the soil biome and of the P-bearing biomolecules derived from the decomposition of plant and microbial debris. However, their proportion varies compared with those found in living cells, depending on their lability and rate of degradation (George et al., 2018). Nucleic acids and phospholipids represent the most abundant forms of P in living cells, whereas in soils they form only about 2 and 5%, respectively, of the total organic P, and monoesters can represent over 50%. Phosphate monoesters, in particular inositol phosphates, such as *myo*- and *scyllo*-inositol hexaphosphate, accumulate in soils since they are protected against degradation by forming stable surface complexes with Fe and Al ox(hydr)oxides and 1:1 Al-silicates, or insoluble precipitates with iron, aluminum and Ca. The main organic forms of P that can be found in living organisms and soil are shown in Fig. 4.

Abiotic processes controlling P availability

Phosphorus is known to be relatively immobile in the soil environment because, differently from N, it forms stable complexes with soil solid phases and or sparingly soluble precipitates with Fe, Al or Ca. For this reason, its concentration in the soil solution is generally lower than those of other macronutrients, ranging between 10^{-8} and 10^{-5} M (Frossard et al., 1995). The partition of soil P

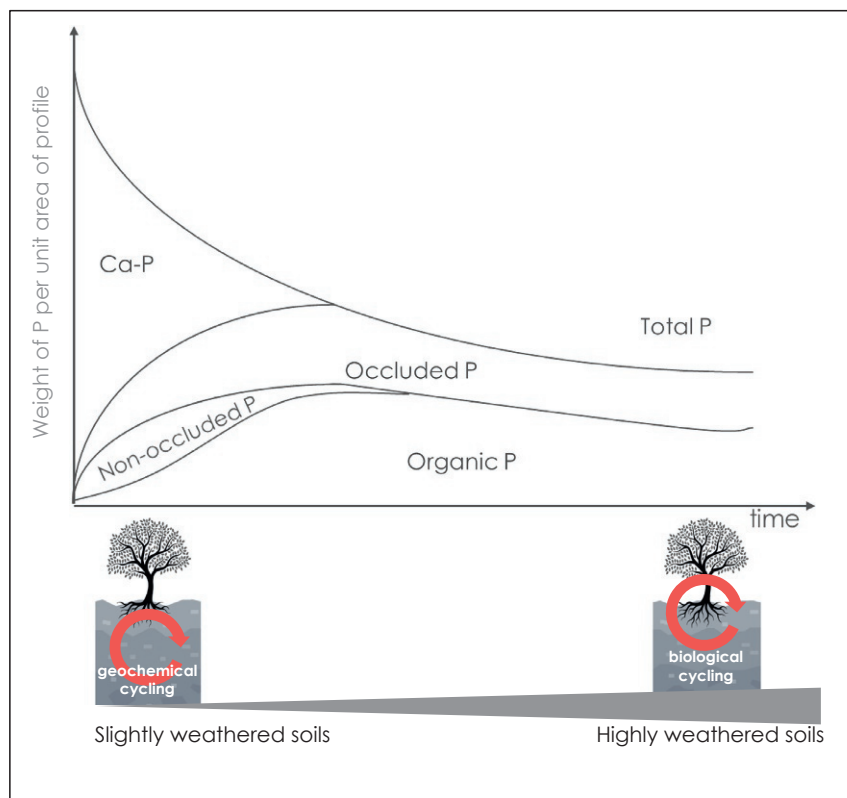


Fig. 3 Conceptual model modified from Walker and Syers (1976) showing variation in soil P pools with time: total soil P; Ca-P: calcium phosphates; organic P; occluded P; nonoccluded P. The model suggests that highly weathered soils have less total P, higher relative proportions of occluded and organic P compared to less developed soils. Modified from Walker TW and Syers JK (1976) The fate of phosphorus during pedogenesis. *Geoderma* 15: 1–19.

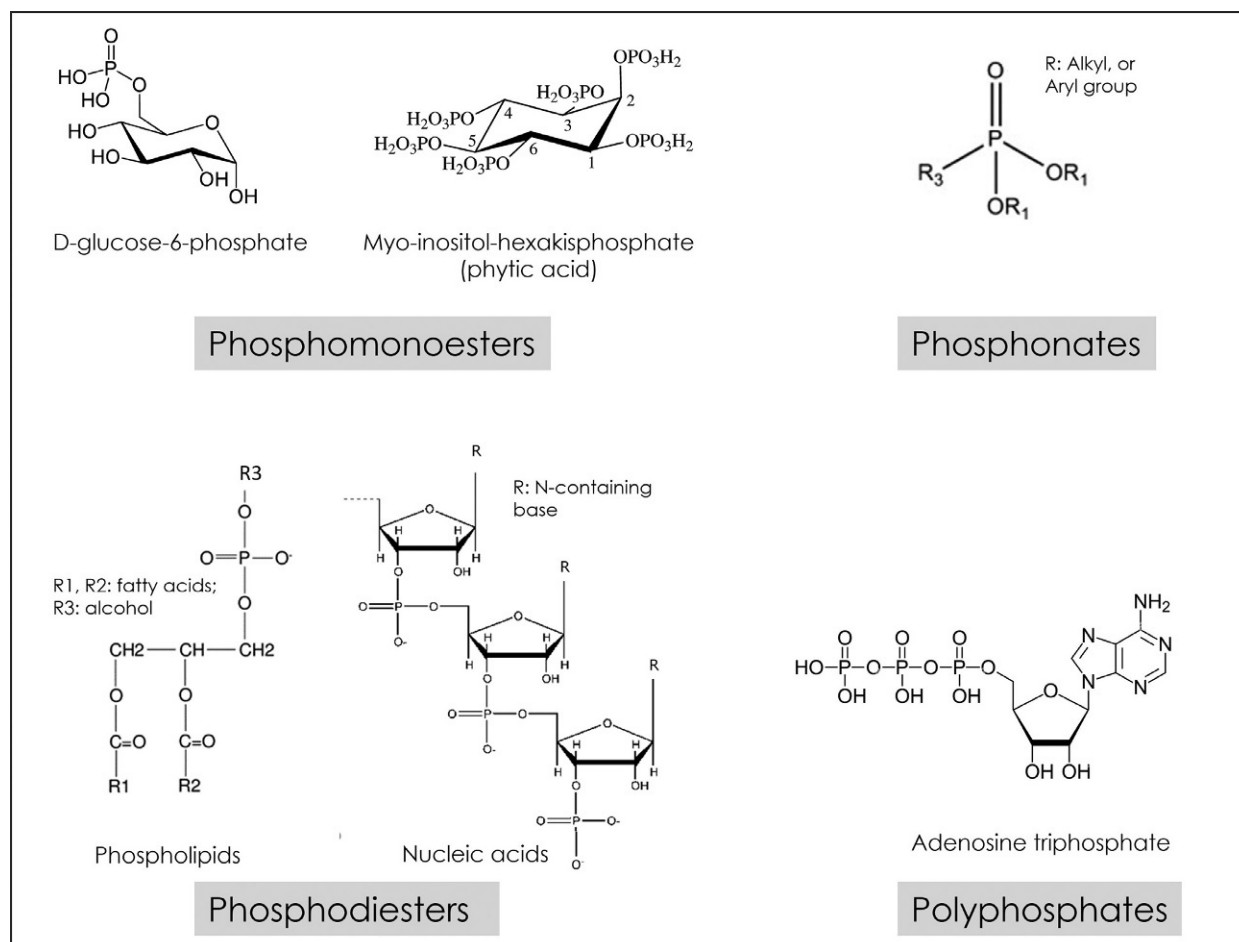


Fig. 4 Main organic P forms present in living organisms and in soil.

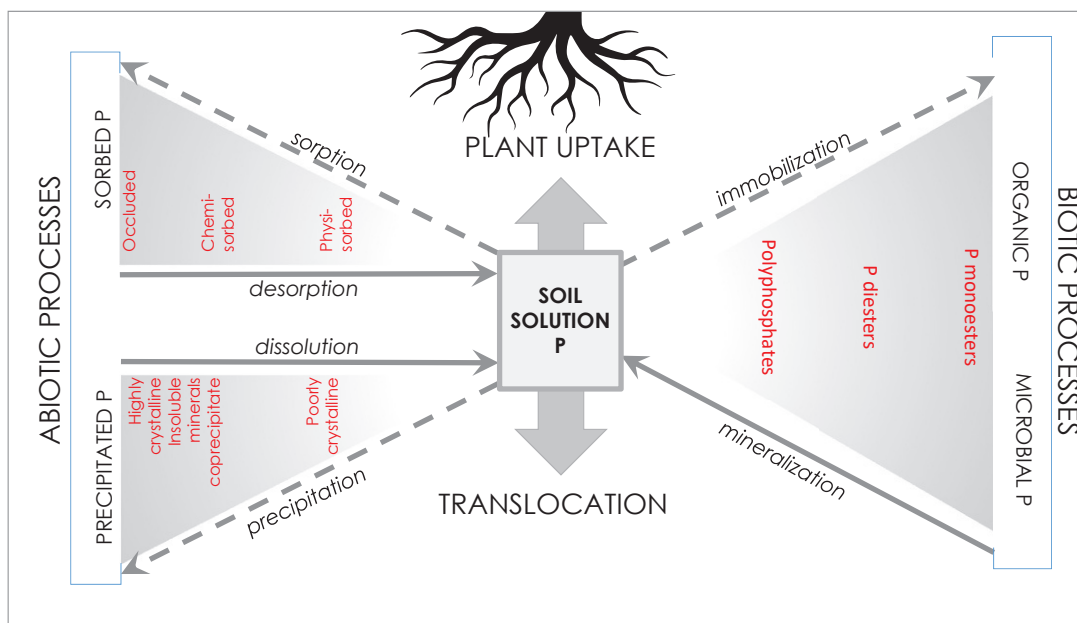


Fig. 5 Abiotic and biotic processes controlling the availability of P for plant uptake. Sorption processes are reported in order of increasing retention from the physi- and chemi-sorbed pools to occluded ones and from precipitated as poorly crystalline hydroxides, crystalline oxides, to coprecipitates and insoluble minerals. Biotic processes include immobilization of P distributed in different organic forms.

among more or less mobile fractions, which can be taken up by plants or move to different environmental compartments (such as ground and surface waters) depends on the mechanisms controlling the equilibria between the forward and backward direction of three couples of reactions, i.e., adsorption/desorption, precipitation/dissolution, and immobilization/mineralization by soil biota. The major factors controlling such equilibria are soil pH, redox conditions, contents of Fe and Al oxy(hydr)oxides and 1:1 Al-silicates, presence of competing solutes, organic matter and other environmental factors affecting soil microbial activity (Fig. 5).

Adsorption and desorption reactions

In developed soils, phosphorus limitations to plant growth can largely derive from the build-up of the inorganic and organic P pools adsorbed on the surfaces of Fe and Al oxy(hydr)oxides and clay minerals. In principle, phosphate anions can adsorb as outer-sphere complexes to any positive site of soil solid phases. However, phosphate can replace structural singly coordinated surface hydroxyls on Fe and Al mineral surfaces forming inner-sphere complexes, and this binding mechanism is considered dominant in acid soils (Arai and Sparks, 2007). Phosphorus species adsorbing with the inner-sphere mechanism can form monodentate, bidentate binuclear, or bidentate mononuclear complexes in different proportions depending on the chemical species of P involved, pH and surface coverage. Inorganic phosphate seems preferentially to form bidentate, binuclear complexes with most Fe oxides (Fig. 6), although monodentate chemi-sorption and physi-sorption are also proposed as potentially important mechanisms (Fig. 5).

Adsorption is the dominant mechanism for P retention operating in acid and subacid soil environments, where variably charged minerals mostly display positively charged surfaces, which favors anion adsorption. The ability of such minerals to deplete the soil solution of inorganic and organic P species depends on their surface area and point of zero charge (e.g., Celi et al., 2020). Thus, poorly ordered minerals with a large specific surface area can sequester more P than well crystallized oxides, for which the specific surface can decrease over one order of magnitude during crystal growth. Different Fe minerals with similar specific surface areas, but with different surficial hydroxyl configurations, such as goethite and hematite, can display different P adsorption capacity. Phosphate adsorption on variably charged surfaces decreases as pH increases because of the reduction in positively charged sites, and with a marked decrease above their point of zero charge. In soils with pH close to, or above neutrality, the adsorption of low concentrations of P on Ca and magnesium (Mg) carbonates can occur, whereas P sequestration from the solution of calcareous soils following large P additions mainly depends on other mechanisms such as its precipitation as sparingly soluble Ca salts. Soil organic molecules can offer positively charged sites for P adsorption, such as protonated amino groups, however, the dominant net negative charge displayed by soil organic matter hampers the adsorption of large amounts of anionic species. In contrast, low molecular weight organic acid anions are known to promote phosphate desorption, competing for the same binding sites on positively charged surfaces and displacing the adsorbed P species. Thus, the input of organic amendment can improve phosphate availability in soils with high P-fixing capacity, and the production of low molecular weight organic acids (e.g., oxalic, citric, malic, etc.) by plant roots and soil microorganisms is an important tool for improving P availability. However, this depends on the forms of P, with inositol phosphates being retained more than P diesters and inorganic phosphate (Fig. 6). It has been observed that the reversibility of P adsorption on Fe and Al minerals diminishes with age of the P-surface bonding, resulting in substantial hysteresis of P

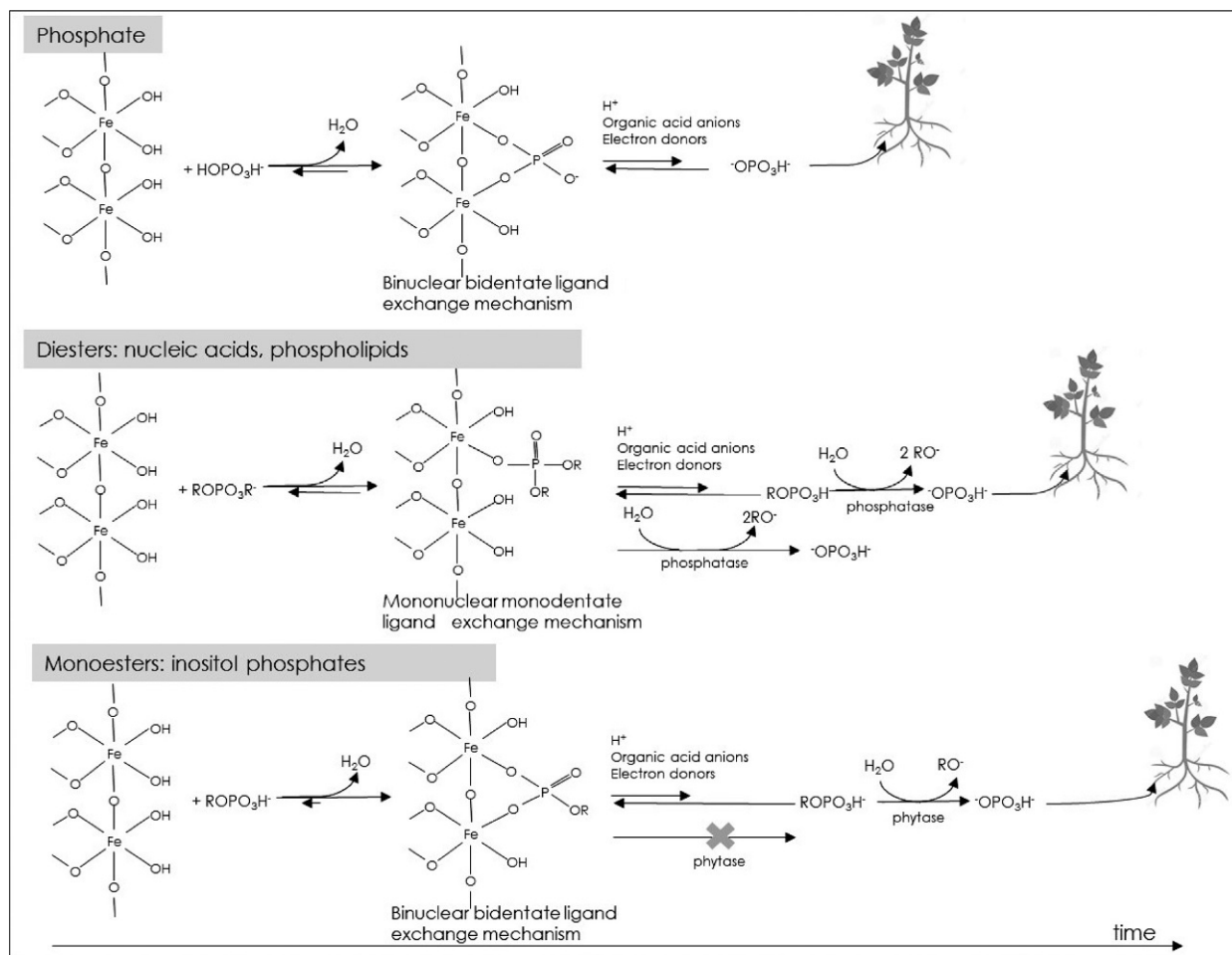


Fig. 6 Adsorption mechanisms of P on Fe oxy(hydr)oxides via (i) a bidentate binuclear ligand exchange mechanism (inorganic phosphate and organic P monoesters), and (ii) a monodentate mononuclear ligand exchange mechanism (organic P diesters). The retention of organic P compounds on mineral surfaces reduces phosphatase activity and inhibits phytase activity. Plant and microorganisms can produce protons, organic acid anions and electron donors to release the organic P compound from minerals for enzyme hydrolysis and favor plant P uptake.

adsorption/desorption curves. This is possibly explained by slow diffusion processes of the adsorbed P into the solid phase, with increasingly scarce availability of legacy phosphorus in agricultural soil (Smolders et al., 2021).

In soils of arid environments, P is commonly less mobile than in those of moist climates. In addition to the slower dissolution of primary phosphate minerals, this depends on the increased pore volume occupied by the soil solution and consequent dilution effect, which enhances P transfer from the solid to liquid phase. Soil moisture also stimulates the activity of soil organisms, accelerating the solubilization of P that derives from the decomposition of organic moieties, as well as the formation of soluble organic molecules. The release into solution of the P adsorbed on Fe minerals is markedly enhanced in submerged soils. The anoxic conditions are accompanied by a decrease in soil redox potential, which causes the reductive dissolution of Fe oxides, starting from the scarcely crystalline forms that have a large active surface. In such soils, the fate of P is closely related to the Fe redox wheel, and conditions that favor a rapid Fe dissolution enhance a correspondingly fast release of P in solution. Conversely, soil drainage and the establishment of oxidizing conditions determine the precipitation of Fe oxy(hydr)oxides and the re-immobilization of P on these solid phases. In periodically submerged soils, such as paddy rice fields, the availability of soil P is closely related to water management practices, with a marked increase in P availability to plants following soil submersion. When reducing conditions persist for a long time, the formerly dissolved P can adsorb on the residual surfaces of non redox-sensitive minerals, as well as on newly formed minerals, including mixed-valence Fe compounds such as green rusts, thus limiting in time the effects of Fe reductive dissolution on P concentration in soil solution.

Solubilization and precipitation reactions

Available soil P in principle originates from the dissolution of primary P-bearing minerals, mainly represented by apatites. Apatite dissolution is favored by moist climates, accumulation of organic matter and sub-acidic pH conditions. In addition to the dissolving effects of a sufficient water content in soil, the presence of organic matter promotes the activity of the soil biome and production of

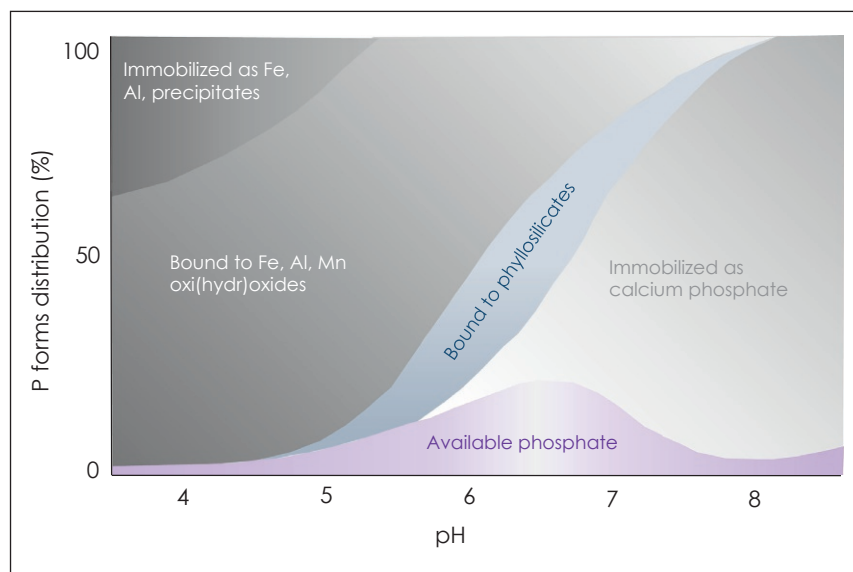


Fig. 7 Retention of inorganic P as a function of soil pH: at low pH P is precipitated as Fe or Al salts, and bound to Fe, Al and Mn oxy(hydr)oxides. As pH increases, P can be bound to phyllosilicates, whereas at pH > 7 P is immobilized as calcium phosphate. The actual proportion of P that remains in an available form is in the narrow pH range between 5 and 7.

organic acids. The latter may decrease the activity of metal cations in solution (including Ca^{2+} , Mg^{2+} and other ions), enhancing the dissolution of primary apatites. Moreover, the release of carbon dioxide with microbial respiration increases the acidity of soil porewater (Frossard et al., 1995; Hinsinger, 2001). In alkaline, carbonate-rich soils, dissolved P forms can be subtracted from solution by the precipitation of secondary Ca-phosphates. This mechanism is governed by soil pH, the solubility of Ca-phosphates increasing as pH decreases, while their precipitation is enhanced with increasing P concentrations (e.g., following P fertilization) and metal ion activity in the soil solution.

In acid soils, P can precipitate with Fe and Al, forming minerals such as vivianite, strengite, variscite; in contrast to Ca-phosphate, the solubility of Fe- and Al-phosphates tends to decrease with increasing soil acidity. The contrasting effect of pH on the solubilization of Ca-phosphates vs Al- and Fe-phosphates, together with the maximum adsorption of P on mineral surfaces when pH is acidic, explains why the optimum soil pH range for P availability to plants is around neutral (pH 6 to 7, Fig. 7).

In soil, secondary P precipitates often display a weak crystalline order because of the presence of dissolved organic molecules and other solutes that hamper the crystallization. However, with time, they tend to evolve to more crystalline forms and become progressively less soluble, which further reduces P availability for plant uptake (Hinsinger, 2001). Organic P forms, such as inositol phosphates, can form secondary precipitates with metal oxides as for phosphate ions. Considering the large concentration that dissolved Fe^{2+} ions can reach in anoxic soils, the precipitation of Fe—P compounds can represent a significant P sink in submerged environments (Santoro et al., 2019). Surface precipitation of organic and inorganic P compounds can also occur after P adsorption on calcite, as well as on Fe and Al oxy(hydr)oxides, particularly with large concentrations of dissolved P (Arai and Sparks, 2007). This phenomenon involves multilayer adsorption and tridimensional growth of the adsorbate on the adsorbent surfaces. While bulk precipitation occurs when the adsorbate is oversaturated with respect to a solid phase, surface precipitation may occur even if the bulk solution is undersaturated with respect to the solubility product of sparingly soluble adsorbate compounds. The formation of surface precipitates, compared with adsorption, further diminishes the reversibility of P bonding to soil solid surfaces, particularly at high P loadings and prolonged residence times. This has to be considered in planning phosphate fertilization management in agricultural soils.

Biotic processes

Microbial immobilization and mineralization

Microorganisms compete with plants for available P, assimilating and immobilizing the element in their cells to synthesize vital biochemical compounds. This process is particularly important in preventing P leaching or fixation of P on soil surfaces provided the microbes remain alive. With microbial death and senescence, microbial organic P becomes part of soil organic P that can be mineralized and thereby supply the bioavailable inorganic P (Pi) pool.

Microorganisms and plants may also cooperate by modifying the rhizosphere characteristics (e.g. pH, redox conditions, enzyme activity), weakening the strength of bonds between P and the soil matrix, and consequently enhancing P availability. In general, the microbes are an advantage for plants because microorganisms often mineralize organic P compounds for their C

needs rather than P (Spohn and Kuzyakov, 2013). This may first lead to a net P immobilization when plant residues show a $C/P > 300$ (i.e. maize (*Zea mays* L.), rice (*Oryza sativa* L.), pruning residues), which in turn, is followed by a net mineralization when C/P decreases to ≤ 100 , due to decomposition or in species such as alfalfa (*Medicago sativa* L.), sugar beet (*Beta vulgaris* L. var. *saccharifera*), potato (*Solanum tuberosum* L.), which have a large N content. Net P mineralization is also influenced by the degree of maturity of the residue itself, contact time with soil, and by the amount of inorganic P initially available for microbial immobilization. Net P mineralization can create an extra P pool for plants although physicochemical processes may mask or dominate microbial P mineralization process. This occurs especially in P-rich soils where rates of gross P mineralization represent a small percentage of P exchanged by physicochemical processes. Conversely, in P-poor soils with a low P-sorbing capacity, microbial P mineralization may dominate over other processes and primarily contribute to the available P pool (Fig. 8).

Input of organic C into soil by litter or root exudates strongly shapes P immobilization/mineralization processes by microorganisms, acting directly on microbial activity or changing the environmental conditions for enzyme reactions. Root exudates can stimulate microbial gross P mineralization and phosphatase activity. The mineralization of organic P compounds may be determined by the C requirements of the microbial community (Spohn and Kuzyakov, 2013), decoupling P mineralization rates from those of C. In P-poor soils, it becomes important to add organic matrices, such as compost, manure, and biowastes to supply nutrients and reduce competition for P reserves between plants and microorganisms. As for nitrogen, synchronization between crop residue or compost addition and fertilization can optimize P availability to plants, keeping in mind that in many biowastes P exceeds crop needs (Fig. 8).

The equilibrium between P immobilization and mineralization can be affected by changes in temperature and soil moisture. In cool climates, low temperatures constrain biological activity, resulting in more organic P but less microbial P than in temperate regions. In temperate regions, seasonality strongly affects microbial P cycling with a larger microbial uptake during the spring–summer time (Bol et al., 2016). In warmer regions a fast biocycling of microbial P is essential to prevent strong limitation of P availability due to the dominance of retention processes, as described above. With climate change, soils subject to progressive drying may exhibit slower microbial P mineralization because water limitation may reduce enzyme activity. Conversely, drastic short-term drying and re-wetting events may cause the release of P through microbial cell lysis.

Plant nutrition and strategies of plants to acquire P

Phosphorus accounts for 0.2–0.5% of the plant dry weight and is an essential macronutrient, playing a fundamental role in the plant cell, as a constituent of DNA molecules at the genetic level, by entering into the composition of membrane phospholipids at the structural level, and in the form of adenosine triphosphate (ATP) at the energetic level. In plant metabolism, P promotes various functions such as cell division, flowering, fruit and seed formation, ripening, root development, and resistance to cold and pest attacks. In cases of P deficiency, which occurs when P in tissues is $<0.1\%$, the plant suffers a reduction in both root and shoot

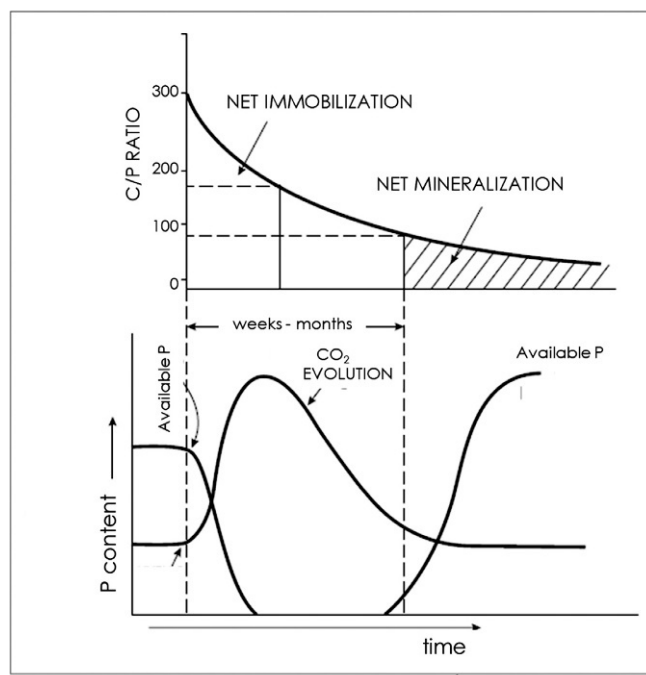


Fig. 8 Microbial immobilization/mineralization as a function of C needs: net P immobilization occurs when plant residues show a $C/P > 300$ causing a decrease of available P in the soil system; net P mineralization occurs with $C/P \leq 100$ resulting in an increase of available P.

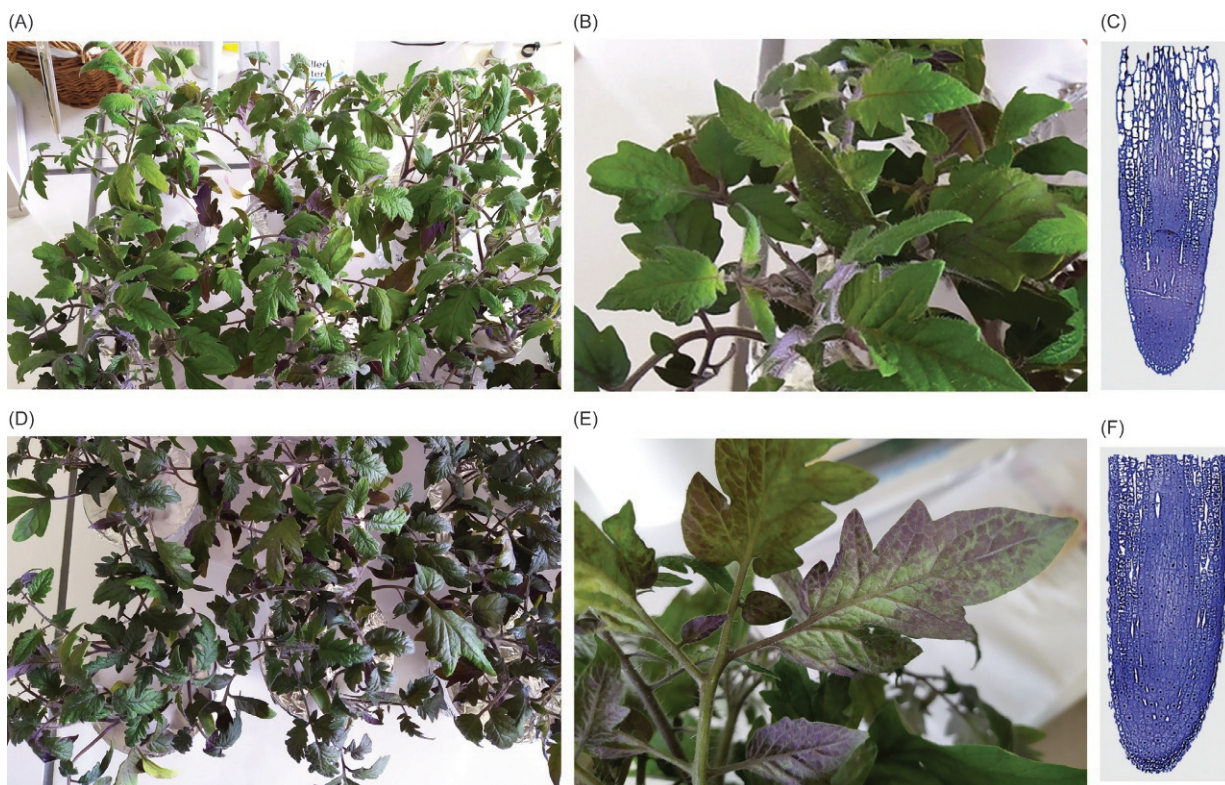


Fig. 9 Shoots and root details of tomato plants grown in conditions of P sufficiency (A, B, C) and P deficiency (D, E, F). The leaves of P-deficient plants show a dark-bluish coloration with subsequent reddening of marginal areas (D and E) and the root (F) shows ultrastructural disturbance, such as the elongation of the meristematic zone along the longitudinal axis. Photos by Dr. Veronica Santoro.

growth, and delays in flowering and fruiting that affect the quantity and quality of production. The shoot/root ratio decreases and leaves, especially the older ones, turn into a dark-bluish-green color with subsequent reddening and necrosis of marginal areas. Excess P has no direct consequences, but it causes lower uptake of some microelements such as zinc (Zn), copper (Cu) and Fe. Therefore, a proper balance between deficiency and excess must be maintained (Fig. 9).

In contrast to the diversity of P pools present in soil, the only form taken up in significant amounts by plants is orthophosphate, which is mediated by dedicated P_i transporter proteins. However, because of the low solubility of sorbed and mineral-bound P, phosphate concentrations in the soil solution are generally relatively small. Available P in the soil solution is taken up by the plant by mass flow, which is a mechanism that creates a flux of water and ions to the roots, but is negligible for phosphate nutrition. More important is the diffusion mechanism: the plant absorbs phosphate because of the decrease in P concentration around the root, creating a concentration gradient that allows P to flux from more concentrated areas toward the root (Fig. 10). The plant can absorb P from its surroundings through an active process that occurs at the expense of respiration and is more pronounced in acidic pH soils. Transport of P across membranes occurs mainly by specific transporter proteins, mostly as co-transport with protons. Translocation of P_i in the xylem is mainly controlled by the difference in water potential between leaves and the soil. Relocation in the phloem can occur as P_i or organic P compounds such as ATP or hexose phosphate, all are highly mobile in this vascular tissue.

To avoid P starvation plants have developed various strategies, which can be subdivided into foraging (moving toward nutrients) and mining (mobilization of poorly available resources) (Richardson et al., 2011). The first strategy is based on the modification of the root architecture, investing a greater proportion of plant biomass in the root system and enlarging the surfaces by root hairs. Adaptations of root architecture that enable greater exploration of the soil volume and or the mining of localized patches of relatively high P availability include inhibition of primary root growth, promotion of lateral root growth, enhancement of root hair development and cluster root formation. An additional major reaction of plants to low nutrient availability is the formation of arbuscular-, ericoid- or ecto-mycorrhizae with respective fungal partners. These lead to increased plant development and vigor due to a high efficiency in P uptake even where available forms are scarce. The second strategy includes root-induced changes in the rhizosphere (Hinsinger, 2001). Roots can acidify the rhizosphere and promote P release from soil surfaces and plant uptake. In addition, plants can improve their capacity to scavenge for P through the release of organic acid anions, particularly citrate, oxalate and succinate, that can compete with phosphate for the same sorption sites. These anions can also increase P availability in acid soils by complexing Fe and Al. In neutral or calcareous soils, acidification increases the solubility of Ca phosphates (Fig. 10).

Roots may also exude low-molecular weight phenolic acids that can contribute indirectly to mobilize P by reducing Fe-containing minerals to which P is associated.

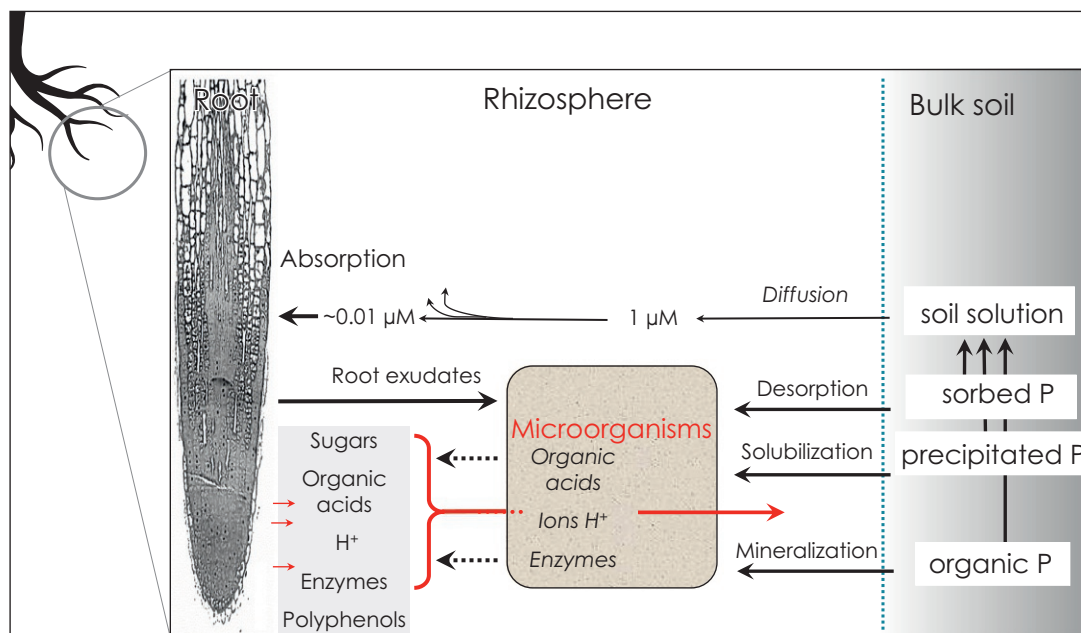


Fig. 10 Processes controlling P dynamics at the soil–plant interface: by desorption, solubilization and mineralization, phosphate can be released from bulk soil and diffuse toward the depletion zone. Plants and microorganisms can exude organic acid anions, protons, enzymes and electron donors to favor the release of P from bulk soil.

Enzyme concentration is generally greater in the rhizosphere than in the bulk soil (Nannipieri et al., 2011). Enzymes are often assumed to be produced mainly by root-associated microorganisms, but plant roots can also contribute (Richardson et al., 2011). Phosphorus bound in organic compounds can be mineralized by phosphatases, comprising phospho-monoesterases (including phytases), phospho-diesterases, triphospho-monoesterases and enzymes acting on phosphoryl-containing anhydrides and on P–N bonds (Nannipieri et al., 2011). An important difference concerns the generally most abundant group of phospho-monoesterases, which can be acid or alkaline according to their pH-optimum. While the former are produced by all relevant organisms including plant roots and ectomycorrhizal hyphae, the latter are mainly produced by free living microorganisms. The pH value may also affect the distribution of other P hydrolyzing enzymes.

Although phosphatase activity in soils is governed to a large degree by the homeostatic behavior of microorganisms and the related strongly constrained enzyme acquisition ratios, there is also clear evidence for the strong effects of P and N availability. While an increase of P availability by P fertilization leads to a reduction in phosphatase activity, aggravation of P limitation by N fertilization or N deposition leads to an increase in phosphatase.

Soil phosphatase activity varies strongly spatially in soil. Spohn and Kuzyakov (2013) showed that phosphatase release at the rhizoplane can differ from that in the extended rhizosphere in terms of pH optimum and dependence on P availability. Although mobilization of P through the abovementioned mechanisms increases P availability for plants and microorganisms, it rarely increases phosphate concentration in the soil solution. Thus, the depletion zones around roots that are created by P_i uptake may not be replenished sufficiently fast by P_i because of its low mobility and strong affinity to the solid phase. Thus, organisms need to take up mobilized P directly to avoid losses to high affinity sorption sites of the soil solid phase. Therefore, the generally more mobile dissolved organic P may be critical for replenishment of the rhizosphere, where it is mineralized by phosphatases in the rhizoplane. Phosphatase may further assist plant nutrition through enhanced recycling of internal P and efficient capture of organic P compounds that may be lost from roots (Richardson et al., 2011).

To what degree are the interactions between plant and microorganisms mutualistic or competitive with regard to P? There is some evidence that plants might be able to stabilize mutualism. For example, under P deficient conditions where the roots are colonized simultaneously by several species and/or genotypes of mycorrhizal fungi, plants can direct C flow to more beneficial mycorrhizal symbionts. While the role of mycorrhizal hyphae in mobilizing sorbed and bound P_i is well established, their involvement in the mineralization of organic P and bioweathering of more recalcitrant P from minerals is, in most cases, poorly quantified. The role of P mobilizing compounds released by non-symbiotic rhizosphere microorganisms for plant nutrition and how this depends on interactions with roots and hyphae is also poorly understood. Because microorganisms require homeostasis, the status of microbial P may be more important for P mobilization in the rhizosphere than the plant P status. The activity of free-living rhizosphere microorganisms may therefore support plant nutrient uptake or compete with it. The search for supporting plant-growth promoting microorganisms is of great current interest for improving P availability for plants.

Losses of P from soil

Phosphorus continues to be a significant threat to global water quality. Rosemarin (2004) estimated that about 25% of the 250×10^6 tons of P extracted from phosphate ores since 1950 is immobilized in the soil or has been delivered to waters. This has led to widespread eutrophication problems because P is the limiting nutrient for algal growth. Eutrophication, that is, the progressive increase of aquatic organisms due to nutrient enrichment, is one of the most important phenomena of environmental degradation related to anthropogenic activities. In a eutrophic waterbody, as a result of the accelerated and abnormal growth of algal and plant biomass, most of the balances in the trophic chain are modified by a significant impoverishment of biodiversity and the impairment of those services and uses (drinking, bathing and tourist use, fishing) associated with surface water quality.

The poor availability of P for crops, due to biotic and abiotic immobilization processes that occur in the soil, and the low cost of phosphate fertilizers has led, since the Second World War, to an excessive use of fertilizers with the result that many soils of industrialized countries are overfertilized. The diffusion of intensive and often specialized farming systems with a growing imbalance between livestock load and cultivated areas where manure could have been distributed is leading farmers to add more P than the crop needs. Imbalance in the manure N/P ratio with P in excess compared to plant needs, has focused mainly on N in the context of legislation, and further contributes to P overfertilization.

Nowadays, point-source nutrient emissions have been reduced with the enlargement of sewage systems, P removal from municipal and industrial effluent, and changes in industrial processes. Thus, agriculture is now considered the main source of P pollution for waters.

Phosphorus in soil is vulnerable to mobilization during storm flow events or during irrigation, and is subsequently delivered to the watercourses. Transfer into waters also occurs through erosion, runoff or leaching as particulate or soluble P. Losses occur more by runoff than by leaching and particulate P is often the major form of P transported from cultivated soils, representing up to 80% of total transported P. Conversely, runoff from uncultivated soils generally moves little sediment, and P losses are generally dominated by dissolved P. Runoff involves mainly the surface soil layer down to 5 cm, and the subsurface if impermeable layers are present. In a catchment there are only a few areas, the so-called 'critical source areas' (CSA), that provide the vast majority of P (Gburek and Sharpley, 1998). These critical areas are influenced by two interdependent factors: (i) source factors, indicating areas that undergo large P accumulation and that consequently represent a great potential for contributing to P export, and (ii) transport factors that depend on the interaction between landscape and climate.

The processes that mobilize added and native P in soils are solubilization and detachment. Solubilization indicates P transfer from a solid phase: at the normal concentration of soil solution P, desorption is generally more important than dissolution or mineralization of organic matter in regulating soluble P concentration. Detachment produces particulate P: this process brings into suspension soil aggregates, primary particles and P-associated colloids (Fig. 11). The controlling mechanisms are mechanical forces exerted by moving water and repulsive forces between soil particles having the same electrical charge.

Soils with a low P sorption capacity, i.e. sandy or organic soils poor in sorbing constituents, can be easily saturated. Under these conditions P concentration of the soil solution could be quite large and is ready for transport by both leaching and runoff. However, P leaching is effective only in sandy, well drained soils and with large volumes of incoming flow. Under these conditions, the risk of P losses can be estimated by the degree of P sorption saturation (DPSS), i.e. the ratio between the previously sorbed P and the P sorption capacity of the soil:

$$DPSS = P_{ox} / (Al_{ox} + Fe_{ox}),$$

where DPSS is the degree of P sorption saturation, P_{ox} , Al_{ox} and Fe_{ox} are the amounts of P, Al and Fe extractable with acid oxalate buffer.

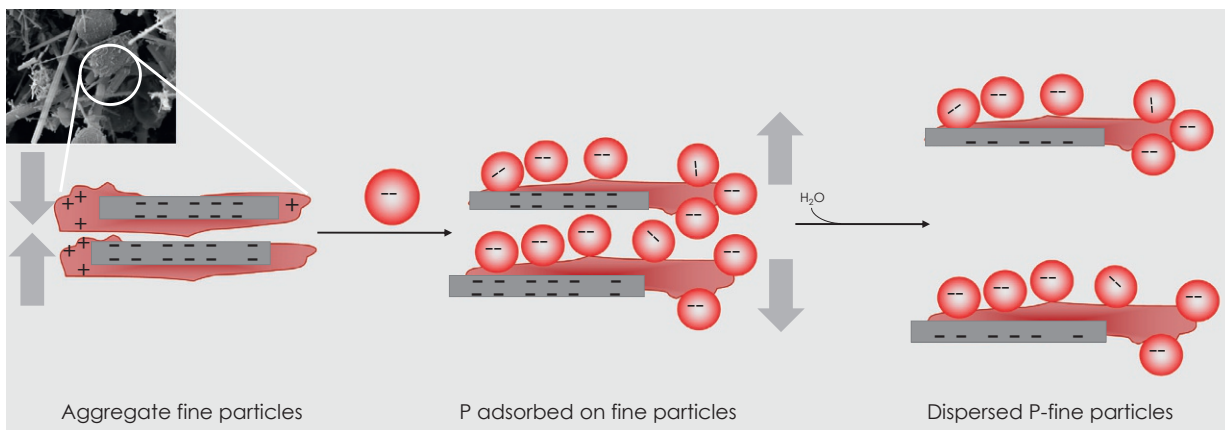


Fig. 11 Dispersion of clay particles because of P retention: clay particles are in general aggregated in most soils; P sorption can cause a reversal of charge producing coulombic repulsive forces that favor particle dispersion and mobilization during runoff with consequent transfer of particulate P from soil to waterbodies.

This index is based on the assumptions that P-sorption capacity is determined by poorly crystalline Fe and Al oxi(hydr)oxides and that P extracted by oxalate represents the P reversibly sorbed by the soil. This approach is reasonably accurate for acid and coarse-to-medium textured soils in humid regions, but it fails to explain the behavior of soils with different mineralogical characteristics. For example, it fails in calcareous soils or in soils where P sorption is determined mainly by crystalline Fe oxides, carbonates and silicate clays; it requires modified equations to calculate the DPSS as a function of the different soil mineralogy.

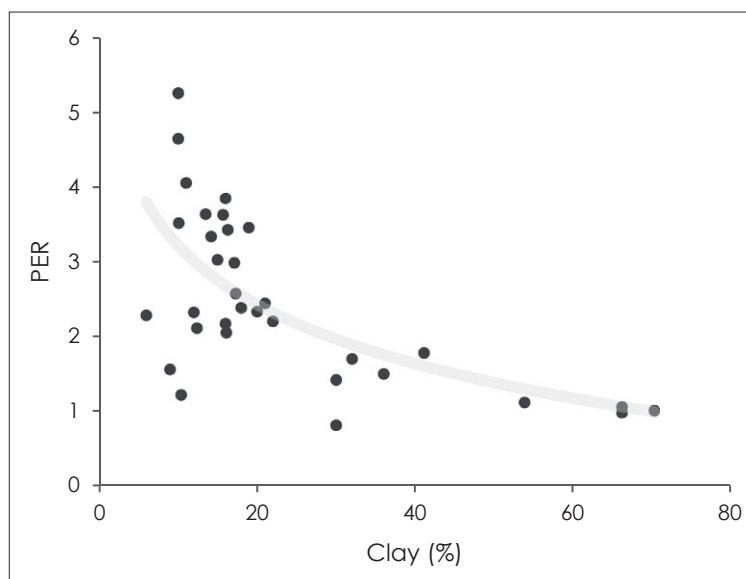
Particle detachment, responsible for losses of particulate forms, occurs through a combination of physical and chemical factors. The main physical factors are rain splash energy, slaking forces of wetting fronts, and the shear forces of overland flow. The chemical factors are related to the accumulation of P in the finest soil particles and to the equilibrium between particle dispersion and aggregation. Factors that affect soil P sorption capacity such as clay and organic matter content, control its accumulation in the finest particles. In turn, P sorption influences the surface charge of the sorbent particles and may even cause a reversal of the positive charge of Fe and Al oxi(hydr)oxides, increasing the electrostatic repulsive forces and enhancing soil particle dispersion (Fig. 11). The detachment and subsequent settling of soil particles results in selective transport of the finest microaggregates and primary particles within water flows. This indicates that total soil P content is not a good index for predicting P losses because the P accumulated in fine particles depends on soil texture rather than on soil total P content, with the highest values in sandy soils (Fig. 12). Thus, tests based on soil dispersibility are useful for predicting P transfer in particulate form and for comparing the vulnerability to P losses of different soils.

As P losses depend on the combination of source and transport factors, all management practices that limit P input, favor particle aggregation or reduce surface runoff can be effective in limiting P losses. Because soil P surpluses are caused by unbalanced fertilization plans, all new integrated multifunctional solutions that combine plant needs and soil properties with a correct fertilization plan and better utilization of legacy P are efficient ways to protect water resources.

Phosphorus resources and global fluxes

Phosphorus is the thirteenth most abundant element in the Earth crust and is a structural and functional component of all organisms. It is an essential nutrient, necessary for growth and development of plants and animals on which our food supply depends. There are no substitutes for P as a plant nutrient, therefore, P fertilizers are required to sustain high crop productivity and quality to provide sufficient food for a rapidly growing world population.

Phosphorus fertilizers are produced from finite, irreplaceable, non-renewable resources, estimates of which are largely uncertain. Rock phosphates can be classified as (i) reserves, i.e. a resource base that is likely to exist, and with production costs that allow current economic extraction, (ii) reserve base: based on probability of occurrence and production costs, including those resources that are currently economic, marginally economic or sub-economic; and (iii) additional resources, i.e. undiscovered resources and unconventional occurrences that can be used in case of future scarcity (Van Vuuren et al., 2010). The U.S. Geological Survey (USGS, 2008) reported a world reserve of almost 18 Gt phosphate rock and a reserve base (including reserves) of 51 Gt.



High-grade ores are concentrated in a few countries: Morocco and Western Sahara control 71% of the world's phosphate, and China, Egypt, Algeria, Brazil, and South Africa 17% (USGS, 2022). while China (53.43%), United States (11.13%), Morocco and Western Sahara, Russia and Jordan are responsible for the 83% of the world's annual production of phosphate rock (USGS, 2019), leaving importing countries vulnerable to geopolitically induced scarcity.

The risk of depletion of P resources has been studied by numerous authors (i.e. Cordell et al., 2009) who, with various mathematical models, have predicted a risk of phosphate rocks exhaustion within a few decades or, at most, within 1–2 centuries. The main difficulties in estimating when phosphate rocks will be depleted is largely due to uncertainties in quantifying the amount of phosphate rocks available in North Africa and China, the increasing fertilizer demand from countries that do not currently use fertilizers, and changes in food that significantly increase the meat-based diet in developing countries. In 2014, P was included in the European Union's list of critical raw materials (CRMs, 2014).

In the pre-industrial period, societal production, processing and consumption of food, feed and fiber were closely related in space and time. Phosphorus, which was removed from the soil with crops from agricultural production areas, was compensated for by regular flooding or shifting cultivation, or was supplemented by P in manure from livestock grazing on surrounding rangeland. Waste products containing P (manure, crop residues, human waste) were sufficient to improve or maintain soil fertility to produce reasonable yields, closing the P cycle. The introduction of fertilizers led to agricultural intensification and specialization, which allowed the human population to grow and afforded humankind a more affluent diet. However, it also led to less recycling of waste streams back to the sites of production, causing imbalances between oversupplied regions and regions where soils are mostly P-deficient. Phosphate fertilization started around the 1850s and rapidly increased after World War II with the Green Revolution. World fertilizer production reached $22.4 \cdot 10^6$ tons in 1988 and then sharply declined by 30% in the 1990s due to both political-economic problems in former USSR countries and to environmental policies, especially in Europe and US to reduce eutrophication problems. Since 2001, however, fertilizer production began to increase again, and today about $18.5 \cdot 10^6$ tons of phosphate rock (about 89% of the amount mined) are used annually for fertilizer production or animal feed.

The global cropland area amounts to $1.6 \cdot 10^9$ ha (FAO, 2021), with an estimate of average P fertilizer application of 10 kg ha^{-1} but very large differences among continents (e.g. 3 vs 25 kg ha^{-1} in Africa and Europe, respectively) and even within the same continent (e.g. 34 and 7 kg ha^{-1} in Ireland and Sweden, respectively). This means that, in contrast to areas that continue to be severely overfertilized, there are large areas, mainly in Africa, South America and some parts of Asia, where soils are often naturally P-poor and already suffer from fertilizer shortages, resulting in poor agricultural productivity and further depletion of soil nutrient reservoirs. Farmers in these countries will be unable to cope with further unavoidable cost increases when phosphate resources begin to run out.

An adult consumes $300\text{--}450 \text{ g P year}^{-1}$ depending on whether his or her diet is vegetarian or includes foods of animal origin: thus, about $3.1 \cdot 10^6$ tons P are required for human nutrition every year, a figure that continues to rise as the world population increases. Comparison of this number with the $18.5 \cdot 10^6$ tons of phosphate rocks designated for agricultural production shows clearly the enormous inefficiency of P use in the food production chain: only about a sixth of the P contained in mined phosphate rocks is consumed in food. The main causes of P inefficiency are due to P transfer from soil to waterbodies by erosion or leaching, accumulation in soil, losses throughout the food chain, and during extraction from mines and processing (Cordell et al., 2009). In addition, close to 100% of P eaten in food is excreted, about 50% of which is in urine and feces, with an average recovery of only about 10% on a worldwide basis; again with enormous differences among countries (Fig. 13).

Concluding remarks

From this analysis, we can infer that significant losses occur throughout the system, from mine to field, to fork. This shows that to address simultaneously phosphate scarcity and water pollution from P leakage, we must adopt an integrated approach that cannot involve only agricultural production and the food supply chain, but must also consider the mining and industrial processes of fertilizer and feed production, and the management of P-containing wastes. To reach this aim, we can work in at least four directions: (i) improving the efficiency of converting phosphate rock into fertilizer; (ii) minimizing phosphorus losses from the farm; (iii) minimizing losses in the food commodity chain; (iv) upcycling alternative renewable P sources, like manure, human excreta and food residues, v) find new mechanisms to reduce overall P demand (Cordell and White, 2011). Implementing these measures will inevitably require an integrated approach that looks beyond the current focus on reducing agricultural P leakage into waterways. This suggests: (i) calibrate soil P inputs; (ii) develop innovative agronomic technologies, including genetic selection; (iii) integrate agricultural and livestock systems in efficient circular systems; (iv) minimize the transfer of P from soil to water by favoring crop P uptake; (v) recover P from urban wastewater with the emphasis to urine; (vi) decrease P losses and improve P recovery throughout the food chain; (vii) influence diet choices; (viii) improve techniques for P extraction; and (xi) develop new technologies for the production of biowaste-derived P biofertilizers.

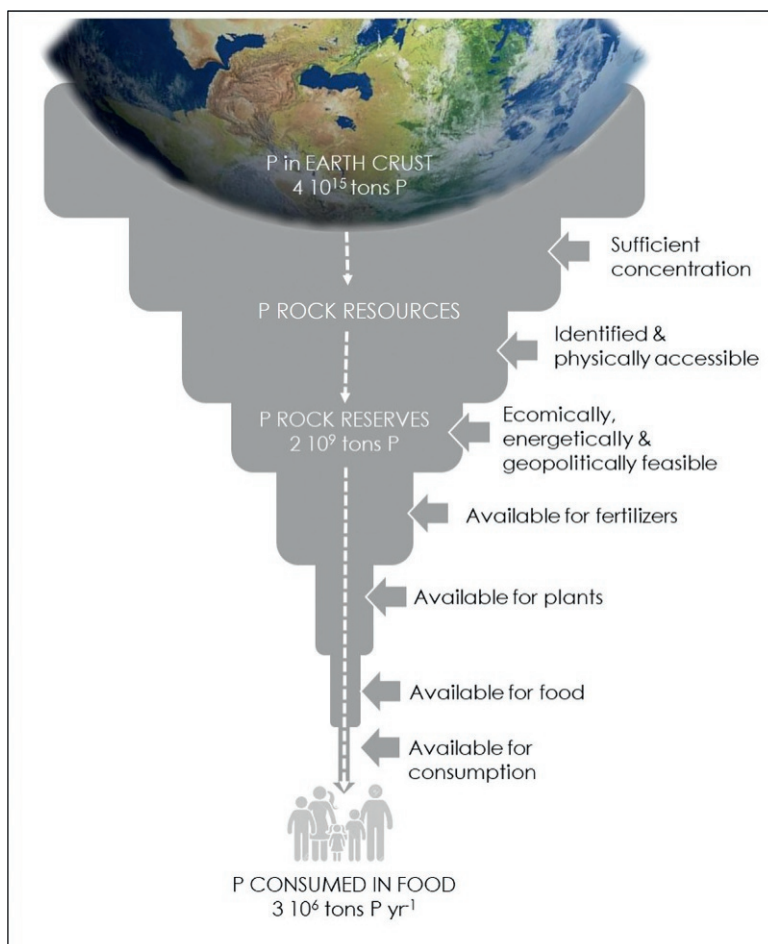


Fig. 13 Phosphorus availability bottlenecks: physical, economic, social and ecological factors limit the availability of P for productive use by humans for fertilizers and hence food production. Units are in P. Adapted from Cordell D and White S (2011) Peak phosphorus. Clarifying the key issues of a vigorous debate on long-term phosphorus security. *Sustainability* 3: 2027–2049.

References

- Arai Y and Sparks DL (2007) Phosphate reaction dynamics in soils and soil minerals: A multiscale approach. *Advances in Agronomy* 94: 135–179.
- Barberis E and Withers P (2002) Influence of soil processes on detachment of P forms: A review of experimental data. In: Chardon WJ and Schoumans OF (eds.) *Phosphorus Losses from Agricultural Soils: Processes at the Field Scale*, pp. 53–60. Wageningen, The Netherlands: Alterra.
- Bol R, Julich D, Brödlén D, Siemens J, Kaiser K, Dippold MA, Spielvogel S, Zilla T, Mewes D, von Blanckenburg F, Puhlmann H, Holzmann S, Weiler M, Amelung W, Lang F, Kuzyakov Y, Feger K-H, Gottselig N, Klumpp E, Missong A, Winkelmann C, Uhlir D, Sohr J, von Wilpert K, Wu B, and Hagedorn F (2016) Dissolved and colloidal phosphorus fluxes in forest ecosystems—an almost blind spot in ecosystem research. *Journal of Plant Nutrition and Soil Science* 179: 425–438. [MM1].
- Celi L, Prati M, Magnacca G, Santoro V, and Martin M (2020) Role of crystalline iron oxides on stabilization of inositol phosphates in soil. *Geoderma* 374: , 114442.
- Cordell D and White S (2011) Peak phosphorus. Clarifying the key issues of a vigorous debate on long-term phosphorus security. *Sustainability* 3: 2027–2049.
- Cordell D, Drangert J-O, and White S (2009) The story of phosphorus: Global food security and food for thought. *Global Environmental Change* 19: 292–305.
- CRMs (2014) <https://phosphorusplatform.eu/scope-in-print/news/1534-white-phosphorus-added-eu-crm>.
- FAO (2021) Land use statistics and indicators statistics. Global, regional and country trends 1990–2019. In: *FAOSTAT Analytical Brief Series No 28*. Rome.
- Frossard E, Brossard M, Hedley M, and Metherell A (1995) Reactions controlling the cycling of P in soils. In: Tiessen H (ed.) *Phosphorus in the Global Environment*, pp. 107–137. Chichester, England: John Wiley & Sons.
- Gburek WJ and Sharpley AN (1998) Hydrologic controls on phosphorus loss from upland agricultural watersheds. *Journal of Environmental Quality* 27: 267–277.
- George TS, et al. (2018) Organic phosphorus in the terrestrial environment: A perspective on the state of the art and future priorities. *Plant and Soil* 427: 191–208.
- Hinsinger P (2001) Bioavailability of soil inorganic P in the rhizosphere as affected by root-induced chemical changes: A review. *Plant and Soil* 237: 173–195.
- Nannipieri P, Giagnoni L, Landi L, and Renella G (2011) Role of phosphatase enzymes in soil. In: Bütemann EK, Oberson A, and Frossard E (eds.) *Phosphorus in Action. Soil Biology*, vol. 26, pp. 215–243. Berlin Heidelberg: Springer.
- Porder S and Ramachandran S (2013) The phosphorus concentration of common rocks—a potential driver of ecosystem P status. *Plant and Soil* 367: 41–55.
- Richardson AE, Lynch JP, Ryan PR, Delhaize E, Smith FA, Smith SE, Harvey PR, Ryan MH, Veneklaas EJ, Lambers H, Oberson A, Culvenor RA, and Simpson RJ (2011) Plant and microbial strategies to improve the phosphorus efficiency of agriculture. *Plant and Soil* 349: 121–156.
- Rosemarin A (2004) In a fix: The Precarious Geopolitics of Phosphorus. In: *Down to Earth*, pp. 27–31. Delhi: Center for Science and Environm.
- Santoro V, Martin M, Persson P, Lerda C, Said-Pullicino D, Magnacca G, and Celi L (2019) Inorganic and organic P retention by coprecipitation during ferrous iron oxidation. *Geoderma* 348: 168–180.

- Smolders E, Nawara S, De Cooman E, Merckx R, Martens S, Elsen A, Odeurs W, Vandendriessche H, Santner J, and Amery F (2021) The phosphate desorption rate in soil limits phosphorus bioavailability to crops. *European Journal of Soil Science* 72: 221–233.
- Spohn M and Kuzyakov Y (2013) Phosphorus mineralization can be driven by microbial need for carbon. *Soil Biology and Biochemistry* 61: 69–75.
- Turner BL, Condron LM, Richardson SJ, Peltzer DA, and Allison VJ (2007) Soil organic phosphorus transformations during pedogenesis. *Ecosystems* 10: 1166–1181.
- USGS (2008) <https://www.usgs.gov/centers/nmic/phosphate-rock-statistics-and-information>.
- USGS (2019) <https://www.usgs.gov/centers/nmic/phosphate-rock-statistics-and-information>.
- USGS (2022) <https://www.usgs.gov/centers/nmic/phosphate-rock-statistics-and-information>.
- Van Vuuren DP, Bouwman AF, and Beusen AHW (2010) Phosphorus demand for the 1970–2100 period: A scenario analysis of resource depletion. *Global Environmental Change* 20: 428–439.
- Walker TW and Syers JK (1976) The fate of phosphorus during pedogenesis. *Geoderma* 15: 1–19.
- Yang X, Post WM, Thornton PE, and Jain A (2013) The distribution of soil phosphorus for global biogeochemical modeling. *Biogeosciences* 10: 2525–2537.

Aluminum in soils

Syuntaro Hiradate, Laboratory of Soil Science, Division of Bioproduction Environmental Sciences, Department of Agro-environmental Sciences, Faculty of Agriculture, Kyushu University, Fukuoka, Japan

© 2023 Elsevier Ltd. All rights reserved.

This is an update of D.R. Parker, ALUMINUM SPECIATION, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 50–56, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00199-5>.

Introduction	168
Elemental properties of aluminum	168
Aluminum species in the solid phase of soils	169
Aluminum species complexed with inorganic ligands in the solution phase of soils	169
Aluminum species complexed with organic ligands in the solution and solid phases of soils	171
Soil components controlling aluminum concentration in soil solution	172
Technical problems in Al speciation in soil environments	174
Summary	175
References	175

Abstract

Soil aluminum (Al) fractions are reviewed in relation to their representative constituents, characteristics, and their extractants. For the Al species in soil solution, the formation constants of Al with ligands, including OH[−] (hydroxyl ion), F[−] (fluoride ion), phosphate ions, and organic ligands, are summarized. For the Al species in soil solids, the dissolution constants of Al-containing minerals are also summarized, and dissolution diagrams based on the constants are shown to explain a general method to identify soil components responsible for controlling Al concentration in soil solution. In addition, technical problems on Al speciation in soil environments are referred to with possible clues for improvement in future research.

Glossary

Mononuclear Al species A molecule containing one Al atom in its chemical structure.

Polynuclear Al species A molecule containing two or more Al atoms in its chemical structure with Al—O(H)—Al bond.

Key points

This article provides information on:

- (1) elemental properties of Al
- (2) Al species in the solid phase of soils
- (3) Al species complexed with inorganic ligands in the solution phase of soils
- (4) Al species complexed with organic ligands in the solution and solid phases of soils
- (5) soil components controlling Al concentration in the soil solution, and
- (6) technical problems in Al speciation in soil environments.

Nomenclature

AAS	Atomic absorption spectroscopy
ICP-MS	Induced coupled plasma mass spectrometry
ICP-OES	Induced coupled plasma optical emission spectroscopy
NMR	Nuclear magnetic resonance
MAS	Magic angle spinning

Introduction

Aluminum (Al) is the third most abundant element in the Earth's crust (mean: 8.2%) and in soils (median: 7.1 to 7.2%) after oxygen (O) and silicon (Si). Because the solubility of Al in water is very low, most soil Al exists in the solid phase, i.e., primary minerals, secondary minerals, and Al-organic matter complexes (Table 1). The relative mobility of Al in landscapes is lowest among common elements in the Earth's crust and in the following order: Cl (chloride) > SO₄ (sulfate) > Na (sodium) > Ca (calcium) > Mg (magnesium) > K (potassium) > Si (silicon) > Fe (iron) > Al (Hudson, 1995). Therefore, the content of Al in soils will increase with the progress of soil formation and weathering.

When Al dissolves in solution and exists as Al³⁺, it exerts an inhibitory effect on the growth and development of many organisms, including higher plants, soil microorganisms, phytoplankton, fish, and other aquatic animals (Parker, 2005). The solubility of Al in water is strongly affected by solution pH. In acidic soils with soil pH(H₂O) range between 4 and 5, the reduction of plant production can be mostly ascribed to Al toxicity rather than H⁺ (hydrogen ion) toxicity. Therefore, it is important to understand the chemical forms of Al in solution in relation to its solubility, pH, origin of the dissolved Al, and interactions of dissolved Al species with other ligands to form complexes. Basic knowledge of the elemental properties of Al will be of help to understand the principle of Al behavior in soils.

Elemental properties of aluminum

In nature, Al has only one stable oxidation number of +III and two stable coordination numbers of 4 and 6. Here, the coordination number stands for the number of anionic atoms (mostly O atoms in nature) bonded to the Al atom. The stereo structures of Al with the coordination numbers of 4 and 6 are the tetrahedron and octahedron, respectively (Fig. 1). In aqueous solution, tetrahedral Al is possible when all of the surrounding O atoms form the hydroxo ion (OH⁻), forming [Al(OH)₄]⁻. When one or more of the surrounding O atoms form(s) the aquo ion (H₂O), Al takes the octahedral structure, forming [Al(H₂O)₆]³⁺, [Al(H₂O)₅(OH)]²⁺, [Al(H₂O)₄(OH)₂]⁺, [Al(H₂O)₃(OH)₃]⁰, and polynuclear Al species. Therefore, the stereo structure and the chemical form of Al in aqueous solution depend strongly on the pH of the solution.

Table 1 Soil Al fractions related with representative constituents, characteristics, and extractants.

<i>Al fractions</i>	<i>Representative constituents</i>	<i>Characteristics</i>	<i>Representative extractants</i>
Al in solution	Monomeric Al species such as [Al(H ₂ O) ₆] ³⁺ , polymeric Al species	Directly toxic to organisms	Water
Al attracted electrostatically	[Al(H ₂ O) ₆] ³⁺ adsorbed electrostatically on negatively charged sites	Toxic after released into solution	Salt solutions (e.g., 1 M KCl)
Al-organic matter complexes	Al coordinated with carboxylic (COOH) groups with ligand exchange reaction	Stabilize soil organic matter	0.1 M pyrophosphate solution
Al in hydroxides and poorly crystalline minerals	Gibbsite, poorly-ordered Al hydroxides, allophane, imogolite	Secondary minerals	0.2 M acid oxalate solution (pH 3.0)
Al intercalated in the interlayer spaces of layer silicates	Hydroxyaluminum polymer ions, hydroxyaluminosilicate polymer ions	Exist as stable pillared structure	0.3 M sodium citrate at 100 °C
Al incorporated in Fe minerals	Al substituted in the crystal structure of goethite, hematite, and ferrihydrite	Al substituted in secondary Fe minerals	Dithionite-citrate-bicarbonate solution
Al composing 1:1 and 2:1 aluminosilicate minerals	Kaolinite, halloysite, smectite, vermiculite, chlorite	Secondary clay minerals	HF-HCl solution
Al composing primary minerals	Feldspars, pyroxenes, amphibole, volcanic glass	Stabilized in crystal structure	HF-HCl solution

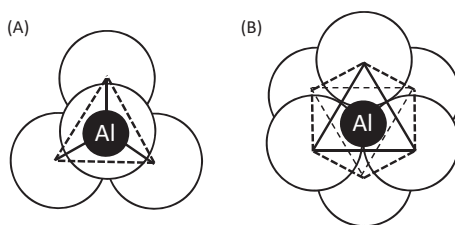


Fig. 1 Stereo structures of (A) tetrahedral Al with coordination number of 4 and (B) octahedral Al with coordination number of 6.

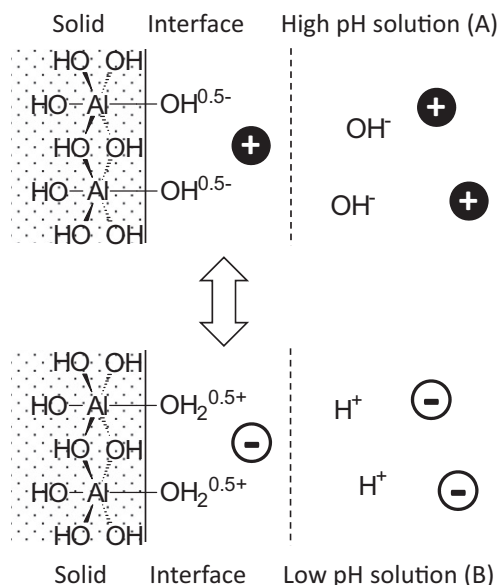


Fig. 2 Amphoteric interface on the surface of gibbsite under (A) high pH and (B) low pH conditions. The interface can possess negative or positive charge depending on solution pH and electrostatically hold ion(s) with opposite charge.

According to the 'electrostatic valency principle' of Pauling's Rules, each of six coordination sites on the octahedral Al should possess a positive charge of 0.5+, because a positive charge of 3+ on the Al atom should be divided equally and distributed at each of the coordination sites ($[3+]/6 = 0.5+$). Therefore, when OH^- and H_2O occupy the coordination site on octahedral Al, the site should possess a negative charge of 0.5- and positive charge of 0.5+, respectively. This phenomenon explains the amphoteric properties of aluminol sites ($\text{Al}-\text{OH}$) coordinated on octahedral Al depending on surrounding pH (Fig. 2). The aluminol site on octahedral Al exists on allophane, imogolite, and the broken edge of 1:1 and 2:1 aluminosilicate minerals and gibbsite. In the case of tetrahedral Al, each of four coordination sites on the tetrahedral Al should possess a positive charge of 0.75+ ($[3+]/4 = 0.75+$). Therefore, each coordination site on the tetrahedral Al possessing OH^- should have a negative charge of 0.25-.

Aluminum species in the solid phase of soils

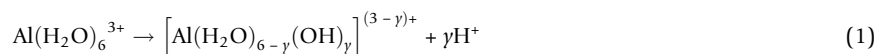
In most primary minerals in igneous rocks, such as alkali feldspars, plagioclase feldspars, feldspathoids, pyroxenes, amphiboles, and volcanic glass, tetrahedral Al constitutes the basic skeleton of those minerals together with tetrahedral Si by forming an $\text{Al}-\text{O}-\text{Si}$ bond, and octahedral Al is only a minor component if it exists. In magma, which generates all igneous rocks, tetrahedral Al will be much more stable than octahedral Al under high pressure, alkaline and anhydrous conditions. In micaceous minerals, which are primary aluminosilicate minerals incorporating a large amount of OH (hydrous minerals), tetrahedral Al and octahedral Al can dominate the tetrahedral sheet and octahedral sheet, respectively. On the other hand, octahedral Al is the major form of Al in secondary soil minerals, such as gibbsite, kaolinite, halloysite, allophane, imogolite, and Al-organic matter complexes. Under soil forming conditions with normal pressure and temperature in the presence of water, tetrahedral Al in primary minerals will be converted into octahedral Al by chemical weathering. The ratio of octahedral Al to tetrahedral Al can be used as an index of mineral weathering in soils.

Soil Al can be operationally classed into several fractions by chemical extractions (Table 1). Most of the reagents and procedures for dissolving the soil Al fractions are identical to those for dissolving soil Fe fractions. Many types of soil organic matter (SOM) are stabilized and sequestered in soils as Al- or Fe-organic matter complexes, and the proportion of the Al-organic matter complexes is generally higher than that of the Fe-organic matter complexes. The aquo ion (H_2O) group on the octahedral Al on the surface of hydroxides and poorly crystalline minerals (Table 1) is very reactive and easily replaced with ligands, such as organic matter with carboxyl/carboxylate group, hydroxyl group, phosphate/phosphoric acid, and HF/F^- .

Aluminum species complexed with inorganic ligands in the solution phase of soils

The hexaquo Al ion, $[\text{Al}(\text{H}_2\text{O})_6]^{3+}$, is a mononuclear Al species and dominant in simple acidic solutions with $\text{pH} < 4.3$. Here, the mononuclear Al species stands for a molecule containing one Al atom in its chemical structure. The $[\text{Al}(\text{H}_2\text{O})_6]^{3+}$ species is toxic to many organisms, and the activity of this species is often best correlated with Al toxicity; no toxic effects of Al have been demonstrated at alkaline pH. The $[\text{Al}(\text{H}_2\text{O})_6]^{3+}$ undergoes hydrolysis when solution pH increases, and it generates four

hydroxylated mononuclear species, i.e., $[\text{Al}(\text{H}_2\text{O})_5(\text{OH})]^{2+}$, $[\text{Al}(\text{H}_2\text{O})_4(\text{OH})_2]^+$, $[\text{Al}(\text{H}_2\text{O})_3(\text{OH})_3]^0$, and $[\text{Al}(\text{OH})_4]^-$. The hydrolysis reaction is described as follows:



with overall (cumulative) formation constants (β) of the form:

$$\beta_y = \left([\text{Al}(\text{H}_2\text{O})_{6-y}(\text{OH})_y]^{(3-y)+} (\text{H}^+)^y / (\text{Al}(\text{H}_2\text{O})_6^{3+}) \right) \quad (2)$$

where the parentheses indicate ionic activities. The overall formation constants for the hydrolysis reactions of the mononuclear Al species are listed in Table 2. The overall formation constants are valuable when the total activity of the mononuclear Al species in the solution is given or no polymerization reaction occurs. The polymerization reaction forms Al—O(H)—Al bond, resulting in polynuclear Al species, which is defined as a molecule containing two or more Al atoms in its chemical structure with Al—O(H)—Al bond.

When the OH^- ion is added to a pure Al solution with the OH/Al ratio < 3, a polynuclear Al species having a chemical formula of $\text{AlO}_4\text{Al}_{12}(\text{OH})_{24}(\text{H}_2\text{O})_{12}^{7+}$ (Fig. 3) is easily formed in the solution. This molecule is comprised of 12 octahedral Al forming a hollow ball and one tetrahedral Al placed inside of the ball; it is called an Al tridecamer (hereafter Al_{13}) or Keggin polymer of Al. The Al_{13} is soluble in water, toxic to plants, and stable for months or even years in solution (Parker, 2005), but its existence in soils and soil solution are rare (Sparks, 2003; Parker, 2005) because the formation and persistence of Al_{13} are suppressed by ubiquitous components of soils, such as silicic acid, sulfate, phosphate, and organic ligands (Hiradate and Yamaguchi, 2003; Yamaguchi et al., 2004).

Table 2 Hydrolysis reaction of hexaaquo aluminum ion ($[\text{Al}(\text{H}_2\text{O})_6]^{3+}$) and its overall formation constants at infinite dilution at 25°C^a.

Reaction	$\log \beta$
$[\text{Al}(\text{H}_2\text{O})_6]^{3+} \rightarrow [\text{Al}(\text{H}_2\text{O})_5(\text{OH})]^{2+} + \text{H}^+$	-5.0
$[\text{Al}(\text{H}_2\text{O})_6]^{3+} \rightarrow [\text{Al}(\text{H}_2\text{O})_4(\text{OH})_2]^+ + 2\text{H}^+$	-10.1
$[\text{Al}(\text{H}_2\text{O})_6]^{3+} \rightarrow [\text{Al}(\text{H}_2\text{O})_3(\text{OH})_3]^0 + 3\text{H}^+$	-16.8
$[\text{Al}(\text{H}_2\text{O})_6]^{3+} \rightarrow [\text{Al}(\text{OH})_4]^- + 2\text{H}_2\text{O} + 4\text{H}^+$	-23.0
$[\text{Al}(\text{H}_2\text{O})_6]^{3+} + \text{SO}_4^{2-} \rightarrow [\text{Al}(\text{OH})_5(\text{SO}_4)]^+ + \text{H}_2\text{O}$	3.5
$[\text{Al}(\text{H}_2\text{O})_6]^{3+} + 2\text{SO}_4^{2-} \rightarrow [\text{Al}(\text{OH})_4(\text{SO}_4)_2]^- + 2\text{H}_2\text{O}$	5.0
$[\text{Al}(\text{H}_2\text{O})_6]^{3+} + \text{F}^- \rightarrow [\text{Al}(\text{H}_2\text{O})_5\text{F}]^{2+} + \text{H}_2\text{O}$	7.0
$[\text{Al}(\text{H}_2\text{O})_6]^{3+} + 2\text{F}^- \rightarrow [\text{Al}(\text{H}_2\text{O})_4\text{F}_2]^+ + 2\text{H}_2\text{O}$	12.7
$[\text{Al}(\text{H}_2\text{O})_6]^{3+} + 3\text{F}^- \rightarrow [\text{Al}(\text{H}_2\text{O})_3\text{F}_3]^0 + 3\text{H}_2\text{O}$	16.8
$[\text{Al}(\text{H}_2\text{O})_6]^{3+} + 4\text{F}^- \rightarrow [\text{Al}(\text{H}_2\text{O})_2\text{F}_4]^- + 4\text{H}_2\text{O}$	19.4
$[\text{Al}(\text{H}_2\text{O})_6]^{3+} + 5\text{F}^- \rightarrow [\text{Al}(\text{H}_2\text{O})\text{F}_5]^{2-} + 5\text{H}_2\text{O}$	20.6
$[\text{Al}(\text{H}_2\text{O})_6]^{3+} + 6\text{F}^- \rightarrow \text{AlF}_6^{3-} + 6\text{H}_2\text{O}$	20.6
$[\text{Al}(\text{H}_2\text{O})_6]^{3+} + \text{H}_2\text{PO}_4^- \rightarrow [\text{Al}(\text{H}_2\text{O})_5(\text{H}_2\text{PO}_4)]^{2+} + \text{H}_2\text{O}$	ca. 3
$[\text{Al}(\text{H}_2\text{O})_6]^{3+} + \text{HPO}_4^{2-} \rightarrow [\text{Al}(\text{H}_2\text{O})_5(\text{HPO}_4)]^+ + \text{H}_2\text{O}$	ca. 7

^aCited from Parker (2005) with modification.

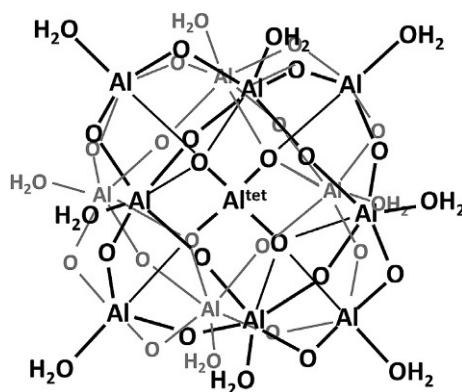


Fig. 3 Chemical structure of $\text{AlO}_4\text{Al}_{12}(\text{OH})_{24}(\text{H}_2\text{O})_{12}^{7+}$ (Al tridecamer, Al_{13} , Keggin polymer of Al). Al^{tet} , tetrahedral Al; Al, octahedral Al; O, OH.

Sulfate, fluoride, and phosphate ions have a high affinity to Al, and their overall formation constants with Al are given in Table 2. On the other hand, the Al ion has a low affinity to form complexes with Cl^- , NO_3^- (nitrate ion), and HCO_3^- (bicarbonate ion) (Parker, 2005). On the basis of the high stability of Al—O—Si bond in aluminosilicate minerals in soils, it would be possible to postulate the presence of a molecule containing Al and Si atoms in its structure with the Al—O—Si bond (i.e., hydroxyalumino-silicate ions) in the soil solution (e.g., Inoue and Yoshida, 1990), but the evidence for its existence is not strong. In the interlayer spaces of 2:1 clay minerals, the hydroxyaluminosilicate polymer ions have been reported to exist as pillared components to form 1.4 nm (14 Å) minerals together with hydroxyaluminum polymer ions in acidic soils (e.g., Bautista-Tulin and Inoue, 1997).

Aluminum species complexed with organic ligands in the solution and solid phases of soils

Aluminum has high affinities to form complexes with organic ligands, such as low-molecular-weight organic acids (LMWOA), high-molecular-weight dissolved organic matter, and high-molecular-weight precipitates of SOM. In most cases, carboxyl groups are responsible for the high affinity, and the contribution of alcoholic hydroxyls and amines to the stability is less important in general.

Many LMWOA, such as propionate, succinate, oxalate, malate, and citrate, can form soluble complexes with Al in solution. They are derived from microbes and plants and are reported to reduce the toxicity of Al by forming complexes, because most Al toxicity is ascribed to Al^{3+} and Al complexed with LMWOA is less toxic (Hiradate et al., 2007). The formation constants of the Al-LMWOA complexes are listed in Table 3. However, some of these constants are under discussion and different values have been proposed under the assumption of different chemical reactions and different experimental settings (Parker, 2005).

Table 3 Complexation reactions of Al with Low-Molecular-Weight Organic Acids (LMWOA) and its overall formation constants at $I = 1.0 \text{ mol dm}^{-3}$ ^a.

Reaction	$\log \beta$	Ionic strength	Temperature
Monocarboxylate			
$\text{Al} + (\text{formate}) \rightarrow \text{Al}(\text{formate})$	0.6–1.4	1.0 mol dm^{-3}	25 °C
$\text{Al} + 2(\text{formate}) \rightarrow \text{Al}(\text{formate})_2$	1.8–2.0	1.0 mol dm^{-3}	25 °C
$\text{Al} + (\text{acetate}) \rightarrow \text{Al}(\text{acetate})$	1.5	1.0 mol dm^{-3}	25 °C
$\text{Al} + 2(\text{acetate}) \rightarrow \text{Al}(\text{acetate})_2$	7.4	0 mol dm^{-3}	25 °C
$\text{Al} + (\text{propionate}) \rightarrow \text{Al}(\text{propionate})$	1.7–1.8	1.0 mol dm^{-3}	25 °C
$\text{Al} + 2(\text{propionate}) \rightarrow \text{Al}(\text{propionate})_2$	3.4	1.0 mol dm^{-3}	25 °C
Dicarboxylate			
$\text{Al} + (\text{oxalate}) \rightarrow \text{Al}(\text{oxalate})$	4.9–6.1	1.0 mol dm^{-3}	25 °C
$\text{Al} + 2(\text{oxalate}) \rightarrow \text{Al}(\text{oxalate})_2$	11.1	1.0 mol dm^{-3}	25 °C
$\text{Al} + 3(\text{oxalate}) \rightarrow \text{Al}(\text{oxalate})_3$	15.1	1.0 mol dm^{-3}	25 °C
$\text{Al} + (\text{succinate}) \rightarrow \text{Al}(\text{succinate})$	3.2	0.5 mol dm^{-3}	25 °C
$\text{Al} + (\text{malate}) \rightarrow \text{Al}(\text{malate})$	5.3	0.2 mol dm^{-3}	20 °C
$\text{Al} + 2(\text{malate}) \rightarrow \text{Al}(\text{malate})_2$	9.3	0.2 mol dm^{-3}	20 °C
$\text{Al} + (\text{D,L-tartarate}) \rightarrow \text{Al}(\text{D,L-tartarate})$	5.2	0.1 mol dm^{-3}	25 °C
$\text{Al} + 2(\text{D,L-tartarate}) \rightarrow \text{Al}(\text{D,L-tartarate})_2$	7.7	0.2 mol dm^{-3}	25 °C
Tricarboxylate			
$\text{Al} + (\text{citrate}) \rightarrow \text{Al}(\text{citrate})$	8.0–8.7	0.1 mol dm^{-3}	25 °C
$\text{Al} + 2(\text{citrate}) \rightarrow \text{Al}(\text{citrate})_2$	12.9	0.1 mol dm^{-3}	25 °C
Benzoate and Polyphenol			
$\text{Al} + (\text{salicylate}) \rightarrow \text{Al}(\text{salicylate})$	12.9	0.1 mol dm^{-3}	20 °C
$\text{Al} + 2(\text{salicylate}) \rightarrow \text{Al}(\text{salicylate})_2$	23.2	0.1 mol dm^{-3}	20 °C
$\text{Al} + 3(\text{salicylate}) \rightarrow \text{Al}(\text{salicylate})_3$	29.8	0.1 mol dm^{-3}	20 °C
$\text{Al} + (\text{catechol}) \rightarrow \text{Al}(\text{catechol})$	16.1–16.9	0.1 mol dm^{-3}	25 °C
$\text{Al} + 2(\text{catechol}) \rightarrow \text{Al}(\text{catechol})_2$	29.1–30.6	0.1 mol dm^{-3}	25 °C
$\text{Al} + 3(\text{catechol}) \rightarrow \text{Al}(\text{catechol})_3$	37.8–39.5	0.1 mol dm^{-3}	25 °C
$\text{Al} + (\text{pyrogallol}) \rightarrow \text{Al}(\text{pyrogallol})$	24.5	0.2 mol dm^{-3}	22 °C
$\text{Al} + 2(\text{pyrogallol}) \rightarrow \text{Al}(\text{pyrogallol})_2$	44.6	0.2 mol dm^{-3}	22 °C
$\text{Al} + 3(\text{pyrogallol}) \rightarrow \text{Al}(\text{pyrogallol})_3$	58.8	0.2 mol dm^{-3}	22 °C
$\text{Al} + (\text{gallate}) \rightarrow \text{Al}(\text{gallate})$	14.2	0.6 mol dm^{-3}	25 °C
$\text{Al} + 2(\text{gallate}) \rightarrow \text{Al}(\text{gallate})_2$	25.4	0.6 mol dm^{-3}	25 °C
$\text{Al} + 3(\text{gallate}) \rightarrow \text{Al}(\text{gallate})_3$	33.3	0.6 mol dm^{-3}	25 °C
Synthetic Reference			
$\text{Al} + (\text{EDTA}) \rightarrow \text{Al}(\text{EDTA})$	16.5	0.1 mol dm^{-3}	25 °C

Water molecules on hydration shell and charges are omitted for simplicity.

^aCited from Nordstrom and May (1996) with modification.

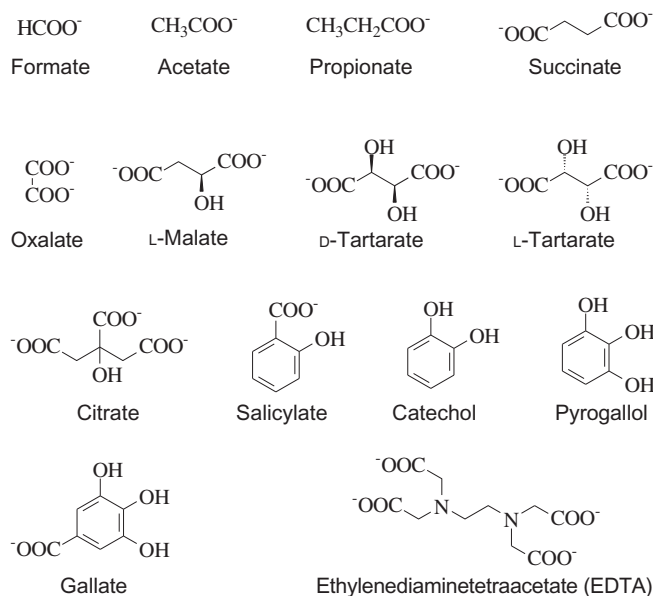


Fig. 4 Chemical structures of deprotonated forms of Low-Molecular-Weight Organic Acids (LMWOA).

In general, the binding affinities of LMWOA to form complexes with Al are in the following order: tricarboxylates > dicarboxylates > monocarboxylates. Because carboxyl C is one of the major forms of soil organic C and therefore SOM can be regarded as a polycarboxylate, the Al ion is fixed on SOM in the solid phase and also acts as the glue to polymerize and stabilize SOM by forming carboxylate–Al–carboxylate bond. The Al complexed with SOM is called Al-SOM or Al-humus complexes (solid components), and they are the major forms of C sequestered in acidic soils, together with Fe-SOM or Fe-humus complexes. For benzoates and polyphenols, such as salicylate, catechol, pyrogallol, and gallate, much greater affinities to Al have been reported (Table 3 and Fig. 4). All of these organic molecules possess 1-carboxyl-2-hydroxy or 1,2-diol structure on a benzene ring, and these partial structures are responsible for the high affinity to Al. The high affinities reported for these compounds are for fully deprotonated forms and thus their actual affinities in soil environments are much less. Nevertheless, these compounds have been reported to have the ability to form complexes with Al in soils (e.g., Furubayashi et al., 2007). Therefore, when SOM contains 1-carboxyl-2-hydroxy or 1,2-diol partial structure on a benzene ring, the Al ions would be fixed through them.

Soil components controlling aluminum concentration in soil solution

In general, Al concentration (activity) in the soil solution is regulated by the dissolution–precipitation reactions of the solid phase. Soils contain various solids that may supply Al to the soil solution (Table 1). Essentially, the solid with the greatest Al solubility determines the Al activity in the soil solution. In many soils, Al-SOM complexes, Al hydroxides and oxyhydroxides, hollow aluminosilicates, 1:1 layer aluminosilicates, or 2:1 layer aluminosilicates supply Al into the soil solution, and their dissolution reactions are shown with dissolution constants ($\log K_{\text{dis}}^\circ$) in Table 4.

In the cases of Al hydroxides and oxyhydroxides, the Al activity in solution at equilibrium can be estimated with the solution pH. For example, the dissolution reaction of gibbsite in Table 4 can be rewritten as below:

$$\log(\text{Al}^{3+}) = 8.04 - 3\text{pH} \quad (3)$$

Therefore, the relationship between $\log(\text{Al}^{3+})$ and pH is linear as depicted in Fig. 5. The cases of equilibria with amorphous $\text{Al}(\text{OH})_3$ and $\gamma\text{-AlOOH}$ (boehmite) are also shown in Fig. 5.

For the aluminosilicate minerals, in addition to pH, the activities of $\text{Si}(\text{OH})_4^0$ and other chemical species should be taken into account for estimating Al activity in solution. For example, the dissolution reaction of montmorillonite in Table 4 can be rewritten as below:

$$0.29 \log(\text{Mg}^{2+}) + 1.71 \log(\text{Al}^{3+}) + 0.22 \log(\text{Fe}^{3+}) + 3.81 \log(\text{Si}(\text{OH})_4) + 6.76\text{pH} = 2.68 \quad (4)$$

$$0.17 \log(\text{Mg}^{2+}) + \log(\text{Al}^{3+}) + 0.13 \log(\text{Fe}^{3+}) + 2.23 \log(\text{Si}(\text{OH})_4) + 3.95\text{pH} = 1.57 \quad (5)$$

$$\log(\text{Al}^{3+}) + 3\text{pH} = 1.57 - 0.95\text{pH} - 0.17 \log(\text{Mg}^{2+}) - 0.13 \log(\text{Fe}^{3+}) - 2.23 \log(\text{Si}(\text{OH})_4) \quad (6)$$

Therefore, when the pH and the activities of Mg^{2+} and Fe^{3+} of the solution are given, the relationship between $(\log(\text{Al}^{3+}) + 3\text{pH})$ and $\log(\text{Si}(\text{OH})_4^0)$ is linear as shown in Fig. 6. Similarly, the dissolution of the other aluminosilicate minerals together with that of Al hydroxides are depicted in Fig. 6.

Table 4 Dissolution constants of aluminum-containing minerals in soils.

Reaction	$\log K_{dis}^0$
Aluminum Hydroxide and Oxyhydroxide	
$\text{Al}(\text{OH})_3$ (amorphous) + $3\text{H}^+ \rightarrow \text{Al}^{3+} + 3\text{H}_2\text{O}$	9.66 ^a
$\alpha\text{-Al}(\text{OH})_3$ (bayerite) + $3\text{H}^+ \rightarrow \text{Al}^{3+} + 3\text{H}_2\text{O}$	8.51 ^a
$\gamma\text{-AlOOH}$ (boehmite) + $3\text{H}^+ \rightarrow \text{Al}^{3+} + 2\text{H}_2\text{O}$	8.13 ^a
$\text{Al}(\text{OH})_3$ (norstrandite) + $3\text{H}^+ \rightarrow \text{Al}^{3+} + 3\text{H}_2\text{O}$	8.13 ^a
$\gamma\text{-Al}(\text{OH})_3$ (gibbsite) + $3\text{H}^+ \rightarrow \text{Al}^{3+} + 3\text{H}_2\text{O}$	8.04 ^a
$\alpha\text{-AlOOH}$ (diaspore) + $3\text{H}^+ \rightarrow \text{Al}^{3+} + 2\text{H}_2\text{O}$	7.92 ^a
Hollow Aluminosilicate	
$\text{Al}_2\text{O}_3\text{SiOH}(\text{OH})_3$ (imogolite) + $6\text{H}^+ \rightarrow 2\text{Al}^{3+} + \text{Si}(\text{OH})_4^0 + 3\text{H}_2\text{O}$	12.1 ^b
1:1 Layer Aluminosilicate	
$\text{Al}_2\text{Si}_2\text{O}_5(\text{OH})_4$ (halloysite) + $6\text{H}^+ \rightarrow 2\text{Al}^{3+} + 2\text{Si}(\text{OH})_4^0 + \text{H}_2\text{O}$	8.72 ^a
$\text{Al}_2\text{Si}_2\text{O}_5(\text{OH})_4$ (kaolinite) + $6\text{H}^+ \rightarrow 2\text{Al}^{3+} + 2\text{Si}(\text{OH})_4^0 + \text{H}_2\text{O}$	5.45 ^a
2:1 Layer Aluminosilicate	
$\text{K}(\text{Si}_3\text{Al})\text{Al}_2\text{O}_{10}(\text{OH})_2$ (muscovite) + $10\text{H}^+ \rightarrow \text{K}^+ + 3\text{Al}^{3+} + 3\text{Si}(\text{OH})_4^0$	13.44 ^a
$(\text{Si}_{2.97}\text{Al}_{1.03})(\text{Al}_{1.44}\text{Fe}^{\text{III}}_{0.07}\text{Fe}^{\text{II}}_{0.99}\text{Mg}_{3.24})\text{O}_{10}(\text{OH})_8$ (Vermont chlorite) + $16.08\text{H}^+ \rightarrow 2.97\text{Si}(\text{OH})_4^0 + 2.47\text{Al}^{3+} + 3.24\text{Mg}^{2+} + 0.99\text{Fe}^{2+} + 0.07\text{Fe}^{3+} + 6\text{H}_2\text{O}$	48.2 ^a
$\text{K}_{0.6}\text{Mg}_{0.05}(\text{Si}_{3.5}\text{Al}_{0.5})(\text{Al}_{1.8}\text{Mg}_{0.2})\text{O}_{10}(\text{OH})_2$ (illite) + $8\text{H}^+ + 2\text{H}_2\text{O} \rightarrow 0.6\text{K}^+ + 0.25\text{Mg}^{2+} + 2.3\text{Al}^{3+} + 3.5\text{Si}(\text{OH})_4^0$	10.35 ^a
$0.4^+(\text{Si}_{3.81}\text{Al}_{1.71}\text{Fe}^{\text{II}}_{0.22}\text{Mg}_{0.29})\text{O}_{10}(\text{OH})_2$ (montmorillonite) + $6.76\text{H}^+ + 3.24\text{H}_2\text{O} \rightarrow 0.29\text{Mg}^{2+} + 1.71\text{Al}^{3+} + 0.22\text{Fe}^{3+} + 3.81\text{Si}(\text{OH})_4^0$	2.68 ^a
$0.51^+(\text{Si}_{3.7}\text{Al}_{0.3})(\text{Al}_{1.345}\text{Fe}^{\text{II}}_{0.405}\text{Mg}_{0.270})\text{O}_{10}(\text{OH})_2$ (Houston black smectite) + $7.2\text{H}^+ + 2.8\text{H}_2\text{O} \rightarrow 0.27\text{Mg}^{2+} + 0.405\text{Fe}^{3+} + 1.645\text{Al}^{3+} + 3.7\text{Si}(\text{OH})_4^0$	1.605 ^a

Water molecules on hydration shell and charges are omitted for simplicity.

^aCited from Lindsay and Walthall (1996).

^bCited from Wada and Kakuto (1999).

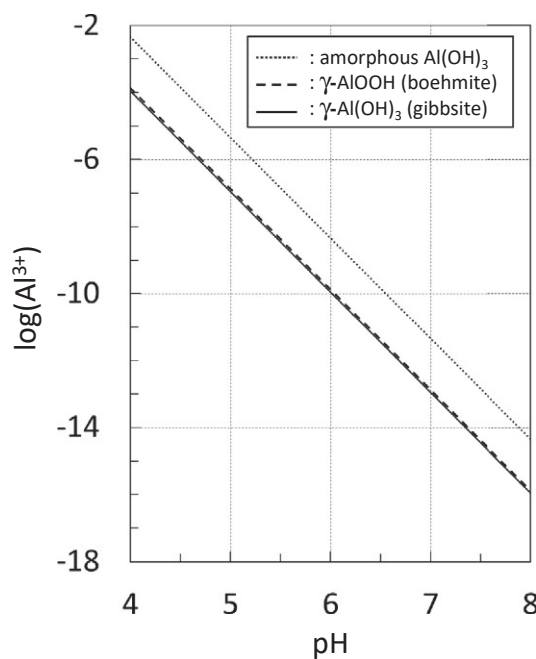


Fig. 5 Solubility diagram for amorphous $\text{Al}(\text{OH})_3$, $\gamma\text{-AlOOH}$ (boehmite), and $\gamma\text{-Al}(\text{OH})_3$ (gibbsite) drawn based on the data in Table 4. (Al_{3+}): activity of Al_{3+} in mol L^{-1} . Modified and recalculated from Nordstrom DK and May HM (1996) Aqueous equilibrium data for mononuclear aluminum species. In: Sposito G (ed.) The Environmental Chemistry of Aluminum, 2nd edn., pp. 39–80. Boca Raton, FL: CRC Press; Sparks DL (2003) Environmental Soil Chemistry, 2nd edn. Amsterdam: Academic Press.

This kind of figure is used to estimate the solid phase regulating Al activity in the soil solution and identify the most stable solid phase in soils. The solid phase with the highest $\log(\text{Al}^{3+}) + 3\text{pH}$ value (y-axis) is likely to supply Al to the soil solution, and that with the lowest y-axis value corresponds to the most stable solid phase. Fig. 6 suggests that amorphous $\text{Al}(\text{OH})_3$ will supply Al to the soil solution when the activity of $\text{Si}(\text{OH})_4^0$ is high, but when $\text{Si}(\text{OH})_4^0$ activity decreases and is depleted, montmorillonite will supply Al to the solution. Similarly, Houston black smectite is the most stable solid phase when the activity of $\text{Si}(\text{OH})_4^0$ is high, but when the activity of $\text{Si}(\text{OH})_4^0$ is decreased, kaolinite and gibbsite become the most stable solid phase in the soil system. It is also effective

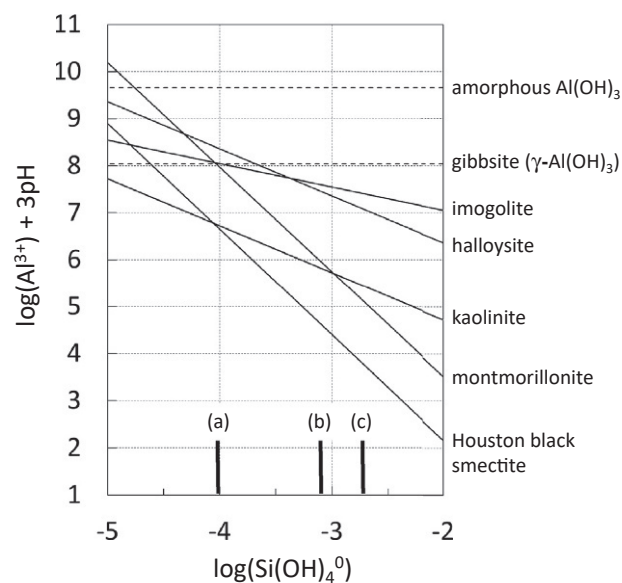


Fig. 6 Solubility diagram for amorphous $\text{Al}(\text{OH})_3$, $\gamma\text{-Al}(\text{OH})_3$ (gibbsite), imogolite, halloysite, kaolinite, montmorillonite, and Houston black smectite at pH 7 under Fe activity equilibrated with soils ($\text{Fe}(\text{OH})_3(\text{soil}) + 3\text{H}^+ = \text{Fe}^{3+} + 3\text{H}_2\text{O}$, $\log K_{\text{dis}}^0 = 2.70$) and K^+ and Mg^{2+} activities of $10^{-4} \text{ mol L}^{-1}$. Dissolution constants were referred from Table 4. a: $\text{Si}(\text{OH})_4^0$ equilibrated with quartz. b: $\text{Si}(\text{OH})_4^0$ equilibrated with soil SiO_2 . c: $\text{Si}(\text{OH})_4^0$ equilibrated with amorphous SiO_2 . (Al^{3+}): activity of Al^{3+} in mol L^{-1} , ($\text{Si}(\text{OH})_4^0$): activity of $\text{Si}(\text{OH})_4^0$ in mol L^{-1} . Modified and recalculated from Nordstrom DK and May HM (1996) Aqueous equilibrium data for mononuclear aluminum species. In: Sposito G (ed.) *The Environmental Chemistry of Aluminum*, 2nd edn., pp. 39–80. Boca Raton, FL: CRC Press; Sparks DL (2003) *Environmental Soil Chemistry*, 2nd edn. Amsterdam: Academic Press.

in identifying the Al supplier to plot the soil solution data of solution pH and chemical composition on this kind of diagram (Lindsay and Walthall, 1996; Sparks, 2003).

It should be noted that the dissolution constant is obtained under equilibrium conditions, and actual soils do not reach true equilibrium. Thus various minerals coexist in a soil. Furthermore, even if a reactive solid phase exists, the surface cover of the solid will prevent the subsequent reactions to supply Al in soil solution. Surface precipitation and sorption are the possible mechanisms that protect the reactive surface from the subsequent reactions. In surface soil horizons, in particular, SOM would be the important agent acting as the surface protector. The complexing ability and surface protecting characteristics of SOM have not been well understood to date because of the complexity of SOM.

Technical problems in Al speciation in soil environments

In general, the concentrations of Al species in soil solution are low ($\mu\text{mol L}^{-1}$ order or less), and the reactivity of the Al species with other chemical species is high. This situation complicates the quantitative determination of the concentration of each Al species in the actual soil solution. For example, many speciation methods of Al rely on reaction-based mechanisms; chromatography and electrophoresis to separate Al species in a mobile phase, resulting in the alteration of pH and composition of the chemical species of the original solutions. Flow-injection analysis and other kinetic analyses employ reactive ligands, such as ferron, 8-hydroxyquinoline, chrome azurol S, and pyrocatechol violet. It is assumed that certain Al species are measured quantitatively during the reaction time with the reactive ligands and that other less labile species are excluded from the reaction, although the relevance of this assumption is not always guaranteed. Atomic absorption spectroscopy (AAS), induced coupled plasma mass spectrometry (ICP-MS), and induced coupled plasma optical emission spectroscopy (ICP-OES) can determine the total Al concentration with high sensitivity, but they cannot determine the concentration of each Al species separately without pre-treatments. The degree of Al polymerization is also very difficult to determine; chemical analyses and computer simulation models are not very effective for this purpose. It is also difficult to separate dissolved Al from colloidal Al distinctly, because a small colloidal particle can pass through a submicron filter membrane. Information on the interaction between Al species and soil organic matters is still necessary to clarify the stability, solubility, degradability, and chemical form of the complexes. These problems were also raised by Parker (2005) in detail.

It would be valuable to combine several analyses and confirm the results with thermodynamic data by employing computer simulation models (e.g., GEOCHEM-PC, MINEQL+, MINTEQA2, PHREEQC, etc.) for obtaining reliable results. In addition, ^{27}Al nuclear magnetic resonance (NMR) spectrometry, a direct and non-destructive analytical method, is a promising tool for the speciation of Al in soil environments (Hiradate, 2004). For the analysis of solutions in particular, the signals originating from the nucleus of ^{27}Al can be obtained non-destructively by simply introducing the sample solution into the spectrometer and irradiating

radio-frequency pulses without the addition of any chemical reagents at ambient temperatures. The problem in analyzing Al species in solution with liquid-state ^{27}Al NMR is the sensitivity and detection limit; liquid-state ^{27}Al NMR can detect 10^{-4} mol L $^{-1}$ or even lower concentrations of $[\text{Al}(\text{H}_2\text{O})_6]^{3+}$ and $[\text{Al}(\text{OH})_4]^-$, but the limit of detection increases to 10^{-3} to 10^{-2} mol L $^{-1}$, and Al may even become undetectable, when ^{27}Al NMR signals are broadened by distortion of the magnetic symmetry of the Al nucleus. This typically happens when $[\text{Al}(\text{H}_2\text{O})_6]^{3+}$ is hydroxylated to form $[\text{Al}(\text{H}_2\text{O})_5(\text{OH})]^{2+}$, $[\text{Al}(\text{H}_2\text{O})_4(\text{OH})_2]^+$, and $[\text{Al}(\text{H}_2\text{O})_3(\text{OH})_3]^0$, and also occurs when Al forms complexes with organic ligands. Therefore, one ^{27}Al NMR spectrum alone may not provide strong evidence, but comparing many ^{27}Al NMR spectra would give stronger evidence to support discussion in some cases. For example, determining changes in Al species with time by liquid-state ^{27}Al NMR multi-spectra was effective to clarify the effect of humic substances on the presence of Al_{13} in solution (Yamaguchi et al., 2004). For the speciation of Al in the solid phase, solid-state ^{27}Al magic angle spinning (MAS) NMR can be applicable. Although the signals of ^{27}Al determined by the solid-state MAS NMR are broader than those obtained by the liquid-state NMR, octahedral Al can be clearly differentiated from tetrahedral Al. It should also be noted that NMR signals are suppressed by the coexistence of paramagnetic materials, such as Fe and Mn. These materials accelerate the free induced decay of NMR signals (shortened relaxation), resulting in extreme broadening of NMR signals.

Summary

Soil Al fractions were reviewed in relation to their representative constituents, characteristics, and extractants. For the Al species in soil solution, the formation constants of Al with ligands, including OH^- , F^- , phosphate, and organic ligands, were summarized. For the Al species in soil solids, the dissolution constants of Al-containing minerals were also summarized, and a dissolution diagrams based on the constants were shown to explain a general method to identify soil components responsible for controlling Al concentration in soil solution. In addition, technical problems in Al speciation in soil environments were pointed out with possible clue of improvement for future study. During the transformation processes of rocks into soils (soil genesis), the chemistry of Al changes dynamically, i.e., coordination number, solubility, chemical form, etc. Because Al has a major impact on chemical, biological, and physical properties of soils, further detailed studies on soil Al need to be performed to understand and manage soils well.

References

- Bautista-Tulin AT and Inoue K (1997) Hydroxy-interlayered minerals in Japanese soils influenced by eolian deposition. *Soil Science Society of America Journal* 61: 631–640.
- Furubayashi A, Hiradate S, and Fujii Y (2007) Role of catechol structure in the adsorption and transformation reactions of L-DOPA in soils. *Journal of Chemical Ecology* 33: 239–350.
- Hiradate S (2004) Speciation of aluminum in soil environments: Application of NMR technique. *Soil Science and Plant Nutrition* 50: 303–314.
- Hiradate S and Yamaguchi N (2003) Chemical species of Al reacting with soil humic acids. *Journal of Inorganic Biochemistry* 97: 26–31.
- Hiradate S, Ma JF, and Matsumoto H (2007) Strategies of plants to adapt to mineral stresses in problem soils. *Advances in Agronomy* 96: 65–132.
- Hudson BD (1995) Reassessment of Polynov's ion mobility series. *Soil Science Society of America Journal* 59: 1101–1103.
- Inoue K and Yoshida M (1990) Composition and behavior of aluminum ions and colloidal aluminosilicates in acidified terrestrial waters. *Soil Science and Plant Nutrition* 36: 461–468.
- Lindsay WL and Walthall PM (1996) The solubility of aluminum in soils. In: Sposito G (ed.) *The Environmental Chemistry of Aluminum*, 2nd edn., pp. 333–361. Boca Raton, FL: CRC Press.
- Nordstrom DK and May HM (1996) Aqueous equilibrium data for mononuclear aluminum species. In: Sposito G (ed.) *The Environmental Chemistry of Aluminum*, 2nd edn, pp. 39–80. Boca Raton, FL: CRC Press.
- Parker DR (2005) Aluminum speciation. In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*, pp. 50–56. Oxford: Elsevier B.V.
- Sparks DL (2003) *Environmental Soil Chemistry*, 2nd edn Amsterdam: Academic Press.
- Wada S-I and Kakuto Y (1999) Solubility and standard Gibbs free energy of formation of natural imogolite at 25°C and 1 atm. *Soil Science and Plant Nutrition* 45: 947–953.
- Yamaguchi N, Hiradate S, Mizoguchi M, and Miyazaki T (2004) Disappearance of aluminum tridecamer from hydroxyaluminum solution in the presence of humic acid. *Soil Science Society of America Journal* 68: 1838–1843.

Iron in soils

Trolard Fabienne^{a,b} and Guilhem Bourrié^b, ^aINRAE—UMR 1114 Environnement Méditerranéen et Modélisation des Agro-hydrosystèmes—Bâtiment Climat—Agroparc, Paris, France; ^bAcadémie d'Agriculture de France, Paris, France

© 2023 Elsevier Ltd. All rights reserved.

This is an update of K. Mengel, H. Kosegarten, Iron nutrition, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, pp. 260–267, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00478-1>.

Introduction	176
Properties of the chemical element iron (Fe)	177
Iron oxide formations	177
Zonal soils: Example of tropical/equatorial situations	177
Typology of lateritic formations	177
Paragenesis of Fe oxides as an indicator of aridity/humidity	179
Azonal soils: Example of hydromorphic soils	180
Typology of hydromorphic soils	180
Iron as the main biogeochemical marker of hydromorphic soils	181
Pathways of iron mineral formation	182
Natural forms of iron oxides	182
The master variables that control the stability of iron oxides	183
In aqueous media: The hydrogen potential, pH and the electron potential, pe	183
In the solid phase: Size of iron mineral	184
Surface properties of iron “oxides”	185
Evidence of in situ dynamics of iron minerals	186
Influence of Fe on landscape evolution	186
Interaction iron minerals—Living organisms	187
Key role of siderophores	188
Pathways for iron in the rhizosphere—Plant system	188
Strategies of iron up-take by organisms	188
Conclusion	190
References	190

Abstract

Iron in the soil is an essential chemical element that contributes both to the structure of soils and to their organization in landscapes. It plays an essential role as a nutrient for all organisms that feed on soil constituents. The duality of insoluble/soluble and oxidized/reduced redox states of iron is a cornerstone of soil formation both zonally and azonally in all surficial Earth formations. These properties are also essential in the mechanisms put in place by organisms to obtain the quantity of iron they need to live from soils where a priori the bioavailability of iron is low.

Key points

- The duality of insoluble/soluble and oxidized/reduced redox states of iron is a cornerstone of soil formation.
- Iron zonal formation laterites cover about one-third of Earth's continental land area.
- Iron azonal formations are dominated by $\text{Fe}^{3+}/\text{Fe}^{2+}$ seasonal dynamics.
- Fougérite is a mixed FeII–FeIII hydroxide formed in moderate reductive conditions as a layered structure.
- There are two iron uptake strategies: (i) Strategy I in all dicots and non-grass monocots, and (ii) strategy II in other organisms, such as grass monocots, bacteria and fungi.

Introduction

When we look at the soil map on a global scale, we observe that the number of major soil types is limited to about a dozen and that the distribution of these soils is above all latitudinal. This means that the main factor in their evolution is essentially the bioclimatic context (temperature, humidity). In a given climatic zone, influence of the substratum on the nature of the soil constitutes the second factor, and at a very local level landform constitutes the third factor of soil evolution.

In all soil classification systems, iron is used as an indicator of horizon differentiation. Thus, classically, brown, red or yellow colors are associated with iron (III) oxides (*sensu lato*), while gray, blue, green or black colors are associated with iron (II) minerals (sulfides and oxides); the white color is interpreted as an indication of the absence of iron (III) oxides.

Iron also plays an essential role in the life of soils, and more generally in all biological cycles interacting with the soil.

Properties of the chemical element iron (Fe)

Iron, Fe ($Z = 26$, A.M. = 55.85 g mol^{-1}) is a chemical element that is part of the first series of so-called transition elements. Due to the increase in number of protons in the nucleus, the filling of the $3d$ subshell, after the filling of the $4s$ subshell, occurs before the filling of the $3p$ subshell. The five $3d$ orbitals of the transition elements are divided, according to group theory, into two families designated by t_{2g} or t_2 and e_g or e . From the energy point of view, these five orbitals are strictly identical in a metal isolated from any influence of electric and or magnetic fields. But in a crystalline structure, the various d orbitals of a transition atom or ion do not react in the same way simply because of their position in space with respect to the anions located at the vertices of the polyhedra. This leads to changes in the energy level of the orbitals, i.e., lifting of degeneracy in a crystal field (Fig. 1).

Iron is chemically active and forms two major series of chemical compounds: the trivalent iron (III), or ferric compounds and the bivalent iron (II), or ferrous compounds. In these compounds, the iron is mainly in octahedral coordination and sometimes in tetrahedral coordination. When iron is in the form of an Fe^{3+} cation, the d orbitals contain five electrons and, in the form of an Fe^{2+} cation, the d orbitals contain six electrons. Depending on the nature of the crystalline field in which the iron is found, several configurations of distribution of the electrons, called high spin and low spin, on the d orbitals are possible.

Stability of the cation depends both on the crystal field to which it is subjected and on its electronic structure. In the mineral, a compromise is thus established between, on the one hand, the fact that the energy of certain d orbitals is decreased and therefore fills in priority and, on the other hand, that in each orbital, the electrons must remain unique as much as possible because pairing consumes energy (of the same order of magnitude as the break in degeneracy). For example, in the mineral goethite, iron can only occupy octahedral sites, *a priori* ideal sites for the stabilization of Fe^{3+} with a crystal field splitting energy (CFSE) of zero. Because the octahedra have common edges leads to local distortions linked repelling iron cations. These distortions are compensated for either partially by changes in the spin state of the iron or by acceptance in the structure of other metals, such as aluminum. In mixed hydroxides like fougérite, stability of the cations in the octahedra is obtained by each cation of Fe^{3+} being surrounded by only divalent cations, either Fe^{2+} or Mg^{2+} .

Iron oxide formations

By order of abundance in the Earth's surface crust, iron ranks fourth and iron oxides constitute a mineral fraction widely represented in soils and weathering horizons. They are ubiquitous and have significant effects on many chemical processes in soils. They can be distributed diffusely in the profile, or concentrated preferentially in specific horizons in proportions that can exceed 50% of the minerals present; e.g., laterites. They form fine crystals, even nanocrystals, such as those observed in coatings on the surface of large minerals or along old root channels or even dispersed in clays; they can also exist as concretions in cemented aggregates, the size of which can easily reach a few centimeters. The crystal lattice of certain oxides allows the substitution of iron by another metal; 10 metals have affinities with iron and the rates of substitution are often significant, i.e., larger than 10%. Due to their small size, iron oxides have large specific surface areas, thus facilitating their interactions with other soil constituents, for example by adsorption with organic complexes, which are sometimes chelating, and with clay minerals, etc.

In the ferric state, iron is soluble only in extremely acidic environments, pH 2–3. In neutral or alkaline environments, Fe^{3+} is not very mobile and is mainly retained in soil minerals. In the ferrous state, iron also forms oxides, but behaves as a water-soluble cation when anaerobiosis becomes established. For this reason, the $\text{Fe}^{2+}/\text{Fe}^{3+}$ redox couple acts as a powerful driving force in the genesis and structure of soils, by cementation/dissolution of aggregates, by the mobility of Fe^{2+} in landscapes.

We focus here on two types of soils where the main mechanisms involved in the dynamics of iron stand out: (i) an example of a zonal soil, the evolution of which depends initially on climatic conditions, and (ii) an azonal soil observed at all latitudes, the evolution of which depends on local waterlogging conditions.

Zonal soils: Example of tropical/equatorial situations

The characteristic of zonal soils is to develop and change under the action of the climate. This is particularly visible when these soils are subjected to extreme climates (precipitation, temperature) over long periods. These soils are currently observed in very low latitude (equatorial and tropical) or very high latitude (boreal) regions.

Typology of lateritic formations

The word "laterite" is derived from the Latin word "later" because in situ, when moist, laterites can be easily cut into bricks. Then later with drying it becomes a very robust material, widely used in the construction of traditional houses in Mesopotamia, Africa, Brazil or India. In the large intertropical zones, due to intensive leaching, laterites are the major end-product of the weathering

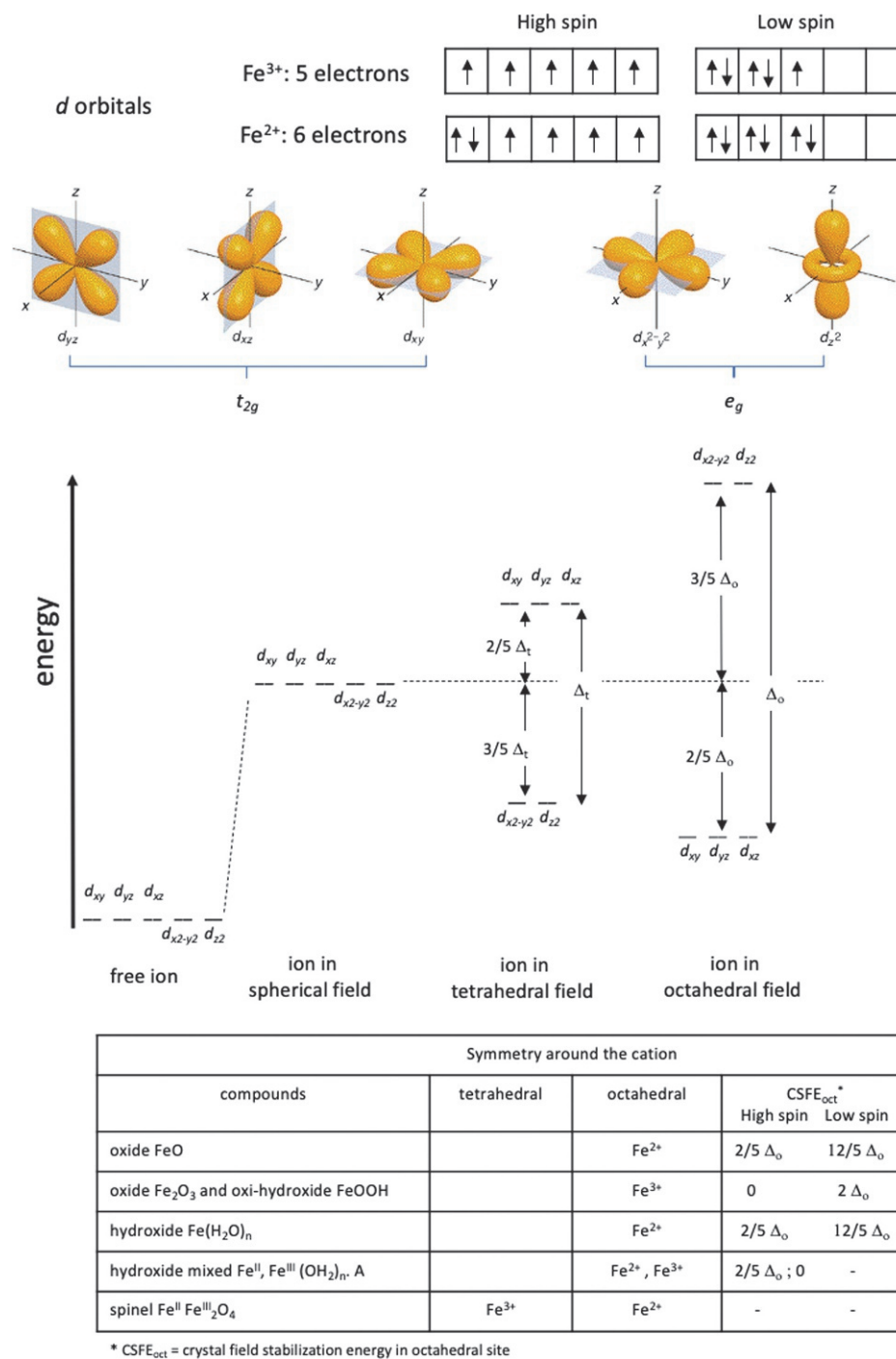


Fig. 1 The geometry of *d* orbitals of the element Fe, breaking of degeneracy of the crystal field in octahedral or tetrahedral coordination of Fe²⁺ and Fe³⁺, and symmetry around the cations in different compounds.

system, which favors the massive accumulation of Fe oxides, relative or absolute and by epigenesis (molecular replacement of one mineral by another). These soils are poor in organic matter, nitrogen, phosphate and calcium, while iron oxides are in excess.

Tardy (1997) estimated that laterites cover about one-third of Earth's continental land area. Lateritic soils form the subsoils of the equatorial forest, savannas and Sahelian steppes that cover most of the land area between the tropics of Cancer and Capricorn. However, laterites can also reflect past weathering conditions, particularly when they occur in present-day non-tropical areas. They can also be formed in geological epochs when the area was near the equator. Thus, present-day laterites that occur outside the humid tropics can be considered to be indicators of climate change, continental drift or a combination of both (Bourman, 1993).

In laterites, two main phenomena participate in the accumulation of iron: (i) a relative accumulation by dissolution and loss of the other elements; (ii) an absolute accumulation, due in particular to the formation and growth of nodules from the matrix (Leprun, 1979). Re-examining the various solums that the latter author observed (Fig. 2) the following remarks can be made:

- In un-weathered bedrocks, whether acidic, metamorphic, basic or ultrabasic, iron is associated with primary minerals and represents a variable percentage depending on the nature of the source rock (1–20% expressed as Fe_2O_3).
- In the context of variegated clays, the plagioclases, K-feldspars, biotites, amphiboles... are increasingly weathered toward the top of the profile, in association with smectites and kaolinites, then kaolinites. Iron is expressed mainly in the form of goethite and hematite. The isovolumetric balance shows strict conservation of the iron content (Millot and Bonifas, 1955).
- In the mottled clay horizon, which represents the next stage of weathering, a first mobilization of iron occurs by redistribution. It is concentrated in sites particularly rich in kaolinite and locally forms reddish to ochre spots. In this horizon, the iron isovolumetric balance is also preserved.
- In the ferricretes, iron accumulates absolutely in the form of nodular “oxides,” which increase in size toward the surface of the solum. The Fe_2O_3 contents then vary between 20% and 60%.

Paragenesis of Fe oxides as an indicator of aridity/humidity

In the context of paleo-climatic reconstructions, the relationships between pedogenesis and climate can be explored by thermodynamic modeling because the durations for establishment of these surficial formations are long enough for equilibria between minerals and climatic conditions to be reached. This is the case, for the parageneses of iron and aluminum oxides observed in laterites in Africa where it is possible to calculate a relationship between the two dominant Fe^{III} oxides (goethite and hematite) in well-drained soils, and pedo-climatic indicators such as climate aridity.

The geochemical system to be considered is the Fe_2O_3 — Al_2O_3 (aluminum oxide)— SiO_2 (silica)— H_2O (water) system for which the domains of stability of the cardinal minerals of lateritic soils can be calculated and represented, namely iron and aluminum oxides and hydroxides. The four main factors that govern the balances between the different minerals are: (i) the water potential (“water activity”), (ii) the temperature, (iii) the chemical composition of the solution (“silica activity”), and (iv) the particle size (differences in solubility products K_{sp}) with the following results (Trolard, 1997):

- At a given temperature, the decrease in water activity (or increase in aridity) leads to the eventual transformation of hydrated minerals into dehydrated minerals, e.g., goethite into hematite, gibbsite into boehmite then gibbsite to corundum or halloysite to kaolinite;
- A decrease in water activity or an increase in temperature have a similar effect on the equilibria leading to the same successions of minerals or parageneses, e.g., in systems with a low aluminum content: $\text{Al-goethite} \rightarrow \text{Al-goethite} + \text{Al-hematite} \rightarrow \text{Al-hematite}$; in systems with high aluminum content: $\text{Al-goethite} + \text{gibbsite} \rightarrow \text{Al-goethite} + \text{boehmite} \rightarrow \text{Al-hematite} + \text{boehmite} \rightarrow \text{Al-hematite} \rightarrow \text{corundum}$.
- In the association (Al-goethite + gibbsite) or (Al-goethite + Al-hematite), the Al concentrations both in goethite and hematite depend only of water activity and temperature: they increase as water activity decreases or as temperature increases;
- The association of Al-goethite and Al-hematite with kaolinite induces significant decreases of Al/Fe ratios in the Fe-Al oxides; these correlate with both increases of water and silica activities.

Thus, the diagrams of the domains of stability of the different oxide parageneses are consistent with the pedoclimatic zoning observed in the natural laterites (Fig. 3), thus establishing the link between the nature of the minerals and the aridity of the climate.

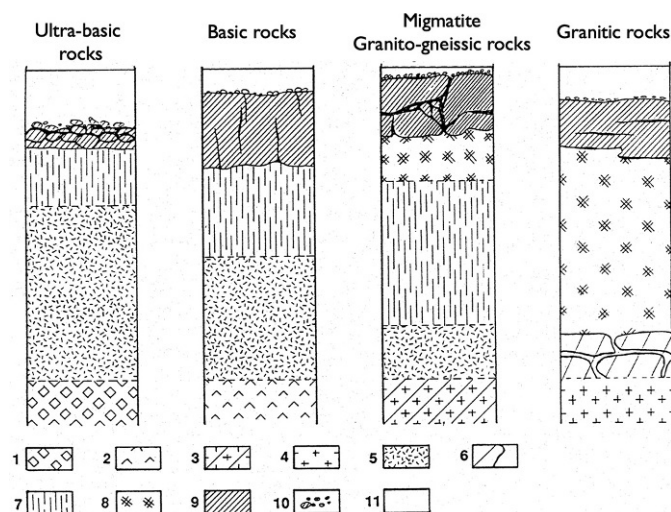


Fig. 2 Summary of the various lateritic solums according to the bedrock (constituted from Leprun's observations, 1979).

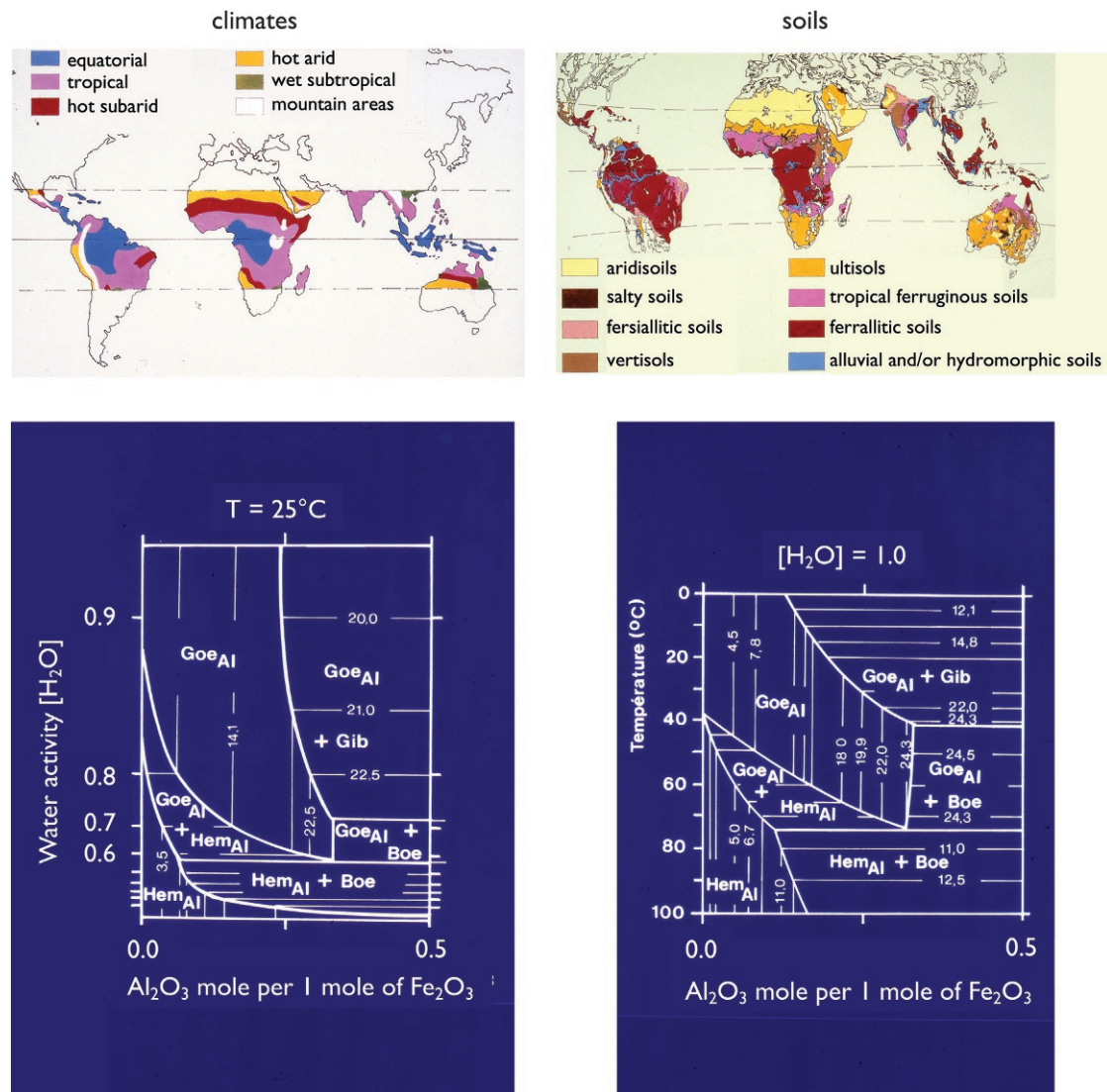


Fig. 3 (a) Pedoclimatic zoning observed in natural laterites and (b) numerical simulations of the stability domains of the main iron and aluminum oxide minerals as a function of climatic variables: humidity (i.e., water activity) and temperature. From Trolard F (1997) *Les "oxydes" de fer des latérites et des sols hydromorphes—géochimie, minéralogie et modélisations thermodynamiques*. Agrégation de l'Enseignement Supérieur, Université Catholique de Louvain (Belgium), 283 pp.

The position of these domains of stability can be modified depending on the state of division of the iron oxides. The influence of the state of division of the minerals is a very sensitive phenomenon, more specifically, in the case of small crystals because a small variation in size leads to significant differences in the stability of the minerals (see also below). Thus, an increase in grain size causes a decrease in the value of the solubility product, and therefore an increase in the stability of the mineral concerned (Trolard, 1988). Thus, increasing the size of the aluminous hematite particles, while keeping that of the goethite constant, leads to a progression of the hematite domains associated with the Al-goethite or boehmite toward increasing water activities.

Azonal soils: Example of hydromorphic soils

Some soils present characteristics, which are independent of climate and the substratum; this is the case for hydromorphic soils. They develop under particular conditions: oxidation–reduction reactions that produce specific features, which are easily recognizable in soils.

Typology of hydromorphic soils

A soil is said to be hydromorphic when it bears characteristics that are attributable to an excess of water. However, several conditions concerning the water regime and the biogeochemical properties of the environment must be fulfilled simultaneously (Table 1). These are: (1) an excess of water, (2) restriction of oxygen sources, (3) presence of metabolizable substrates, (4) at least temporarily

Table 1 Summary of relationships between processes and soil properties.

Processes	Soil properties
Water saturation	Occupation of pore spaces
Waterlogging	High water residence time
Anaerobiosis	As a result of oxygen restriction, the microflora uses other electron acceptors
Oxido-reduction	Electron transfer between donors and acceptors
Hydromorphy	Occurrence of morphological features attributable to excess water

thermal conditions favorable to microbial activity, (5) the presence of chemical elements likely to change the oxidation–reduction state and to record these changes more or less reversibly, called geochemical markers.

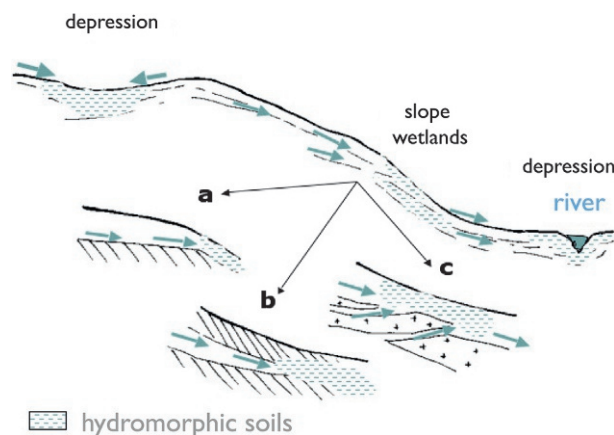
In landscapes, seasonally waterlogged soils are often located in areas with low or zero slope where water flows converge toward local depressions (Fig. 4). These areas are observed both on plateaus and in valleys. Whether Stagnosols (Soil Survey Staff, 1999), located on the plateaus or Gleysols observed in the valleys, they are also called topomorphic hydromorphic soils. This is suggested by the coincidence on the map of these soil types with that of the Kirkby index that links the morphometric index to the probability of soil water saturation (Mérot and Bruneau, 1993). In other cases, hydromorphic soils are observed on the slopes (Fig. 4) in association with the lithological discontinuities. Stagnosols are also often called hydromorphic lithomorph soils (Bourri  et al., 1994). In a third situation, hydromorphy is induced by the soil processes themselves, due to changes in the nature and or organization of the constituents drastically modifying the permeability of the soil. They are observed both in Stagnosols and Gleysols and are called automorphic hydromorphic soils (Bourri  et al., 1994).

An example of a well-developed hydromorphic soil is given Fig. 5. The solum presents from top to bottom, an organo–mineral horizon, a bleached horizon where the iron has migrated then an essentially mineral mottled horizon, marked by more or less diffuse precipitation of iron oxides. Such precipitates can become concretions and crusts where the water table fluctuates. Finally, a blue-green horizon, whose color turns to ochre when the soils are open to the outer atmosphere. This color has often been ascribed to the occurrence, in moderately reduced waterlogged soils, of mixed Fe(II)–Fe(III) compounds with the likely structure of green rust (GR) and of Fe^{2+} in the soil solution. This assumption was formulated in the 1960s and has been widely discussed in the literature. This characteristic blue-green color has consequently been used in all soil classifications since the 1960s and adopted in the World Reference Database (WRB) (IUSS—Working Group, WRB, 2022) as the universal criterion for the recognition of oximorphic or redoximorphic features, used to characterize gleyic properties, reducing conditions and stagnic properties.

Iron as the main biogeochemical marker of hydromorphic soils

Among those elements for which the oxidation state can change under present Earth surface conditions, iron is the most abundant, and the switching between di- and tri-valent redox states of iron initiates several significant geochemical reactions. Moreover, the large difference in mobility between Fe^{2+} , soluble, and Fe^{3+} , insoluble (except at very acidic pH) induces segregations of iron in the horizons and or in soil sequences. These differences are ordered in the landscape and are considered responsible for processes that produce concretions (see ref. included in Trolard and Bourri , 2008).

In hydromorphic soils, iron is mobile, sometimes in significant proportions following seasonal dynamics. These dynamics can be of two types: (i) the first is linked to the succession of processes of waterlogging, reduction then aeration, oxidation, classically observed in the neighborhood where the water table fluctuates and (ii) the second is observed when the mobility of the iron is controlled by the degree of water oxygenation, linked to the rapid infiltration of rainwater in a permanently saturated system. These dynamics, which essentially concern Fe(II) in water, condition the mineralogy of iron oxides that are observed in soils. Thus, during

**Fig. 4** Typology of hydromorphic soils.

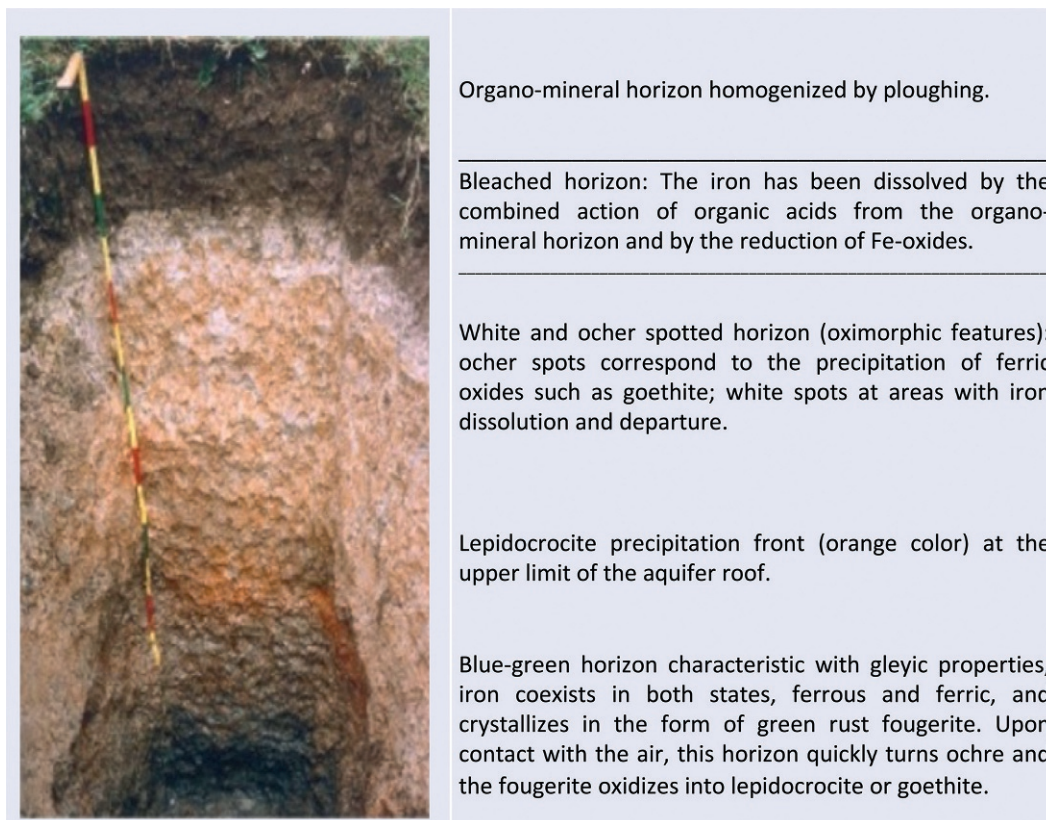


Fig. 5 An example of a hydromorphic soil developed in the granitic area of Brittany, France. Modified from Trolard F, Bourrié G (2008) Chapter 5: Geochemistry of green rusts and fougerite: A reevaluation of Fe cycle in soils. *Advances in Agronomy* 99: 227–289. [https://doi.org/10.1016/S00652113\(08\)00405-7](https://doi.org/10.1016/S00652113(08)00405-7).

the period of oxidation, the dominant Fe oxides are lepidocrocite, in environments where the water is poor in other soluble elements, and ferrihydrite and goethite in other cases. During the period of reduction, if the medium is moderately reduced, green rust, fougerite, becomes the main mineral (soil becomes a blue-green color), and if the medium becomes strongly reducing Fe-sulfides (black color) precipitate.

Pathways of iron mineral formation

Study of the formation of iron oxide pathways in natural environments has been a widely developed topic in the literature for more than a century. It has also been supported by numerous laboratory studies because iron oxides are compounds that can be easily synthesized at ambient temperature and pressure, which establishes knowledge of their structure, their thermodynamic stability and reactivity. This topic has experienced renewed interest over the past 15 years, in particular through research on iron oxides in nanosized particle form.

Natural forms of iron oxides

Iron oxides exist in a bewildering variety of polymorphs (Cornell and Schwertmann, 2003) and the general term “oxide” refers to metal hydroxides, oxyhydroxides and hydroxides. Anhydrous oxides include hematite, α -Fe₂O₃ and maghemite, γ -Fe₂O₃, which both contain only ferric iron and, magnetite, Fe₃O₄, which contains both ferric and ferrous iron. Maghemite and magnetite, with their spinel structure, can form a continuous solid solution. The main oxyhydroxides observed in soils and surface environments include goethite, α -FeOOH, lepidocrocite, γ -FeOOH, and akaganeite, δ -FeOOH. More hydrated forms such as ferrihydrite, nominally Fe(OH)₃, but whose proposed structures are also approximatively Fe₅HO_{8.4} H₂O, better recast as FeOOH.0.4 H₂O, and schwertmannite, ideally Fe₈O₈(OH, SO₄)₁₂₋₁₃ · 10–12 H₂O, has even more variable water content. Hydrated phases containing both ferrous and ferric iron include green rust fougerite, $[(Fe^{II},Mg)_{1-x}Fe_x^{III}(OH)_2]^{+x}[x/n A^{-n}, mH_2O]^x$, with a layered structure where A is the interlayer anion, n is its valency and $1/4 \leq x \leq 1/3$ (Trolard et al., 2007). The octahedron is the basic structural unit for all Fe oxides. Each Fe atom is surrounded by six oxygens either both O₂²⁻ and OH⁻ ions or only OH⁻ in the case of fougerite. In this latter mineral, the presence of Fe(III) gives a positive charge to the hydroxide sheet, which is compensated by interlayer anions with numerous possible types (Trolard et al., 2022). The various Fe oxides differ primarily in their ratios of O/O + OH and Fe^{III}/Fe_{total} (Fig. 6), the octahedra arrangements and how they are linked.

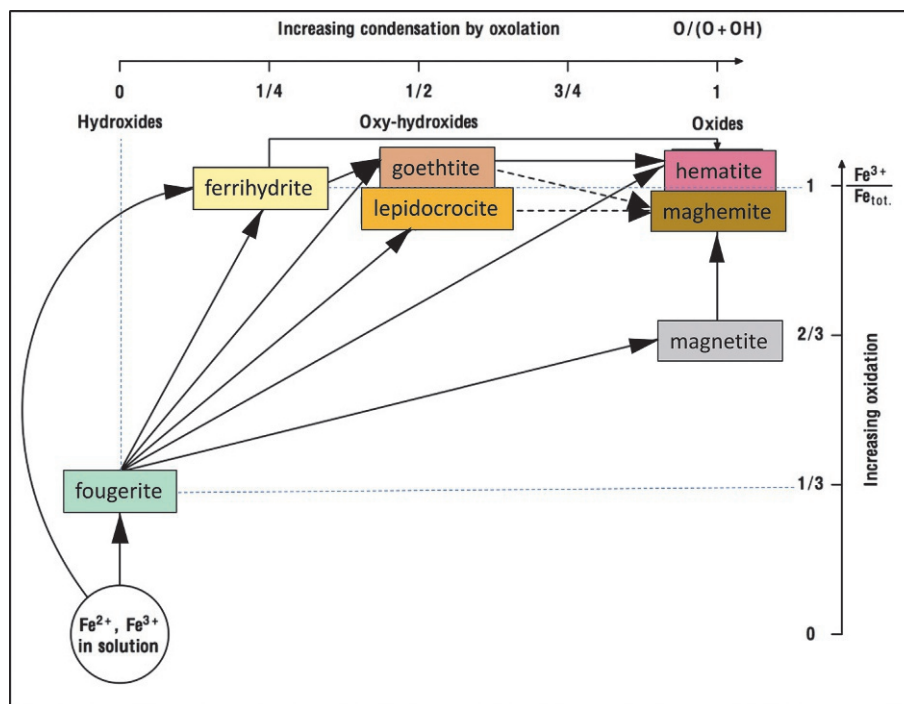


Fig. 6 Pathways between the major iron oxides. Modified from Trolard F, Bourrié G (2008) Chapter 5: Geochemistry of green rusts and fougerite: A reevaluation of Fe cycle in soils. *Advances in Agronomy* 99: 227–289. [https://doi.org/10.1016/S00652113\(08\)00405-7](https://doi.org/10.1016/S00652113(08)00405-7).

The master variables that control the stability of iron oxides

The three main variables that control the reactions in the soil are: (i) the hydrogen potential or pH, (ii) the electron potential or pe and (iii) the size of the iron minerals. The first two variables can be measured only in aqueous medium; our comment below applies only to chemistry in the domain of water stability. The third variable concerns the solid phase when the size of the polymorphs has a significant impact on the stability and reactivity of iron oxides.

In aqueous media: The hydrogen potential, pH and the electron potential, pe

The hydrogen potential or pH is related to the activity of the hydronium ions H⁺ in solution by the relationship:

$$\text{pH} = -\log[\text{H}^+], \quad (1)$$

where [H⁺] defines the activity of hydrated protons. The pH is determined by the relative abundances of acids (proton donors) and bases (proton acceptors), according to the Henderson–Hasselbalch equation:

$$\text{pH} = \text{pK}_a + \log \left(\frac{[\text{Don}_\text{H}]}{[\text{Acc}_\text{H}]} \right), \quad (2)$$

where pK_a is the decimal cologarithm of the acidity constant of the acid–base couple, [Don_H] is the activity of the proton donor and [Acc_H] is the activity of the proton acceptor of the acid–base couple considered.

In the neutral to moderately alkaline or acid pH range, between 5 and 8, iron oxides are very stable and very sparingly soluble. At very acidic pH, below 3, the Fe³⁺ ions can be free in solution and the iron oxides dissolved. At the other extreme, at very alkaline pH values above 10, iron oxides can be destabilized by releasing Fe(OH)[−] or Fe(OH)₂[−] anions into solution.

The electron potential, pe is similarly defined by:

$$\text{pe} = -\log[\text{e}^-], \quad (3)$$

where [e[−]] is the electron activity. The pe is determined by the relative abundances of electron donors and electron acceptors, according the following equation:

$$\text{pe} = \text{pe}^0 + \log \left(\frac{[\text{Don}_\text{e}]}{[\text{Acc}_\text{e}]} \right), \quad (4)$$

where pe⁰ is the value of pe under standard conditions ($P = 10^5 \text{ Pa}$, $T = 298.15 \text{ K}$), [Don_e] is the activity of the electron donor and [Acc_e] is the activity of the electron acceptor of the redox-couple considered.

The pe is related to the oxidation–reduction potential, E_h, by the relationship:

$$\text{pe} = nF E_\text{h} / \ln(10) RT, \quad (5)$$

where n is the number of electrons involved in the half-reactions, F is the Faraday constant ($=96,485 \text{ C mol}^{-1}$), R the ideal gas constant ($8.3144 \text{ J mol}^{-1} \cdot \text{K}^{-1}$) and T the absolute temperature (K).

However, there are fundamental differences between pH and pe in water because water is a protolytic solvent on the one hand and an insulator on the other. The hydrated protons, of which H_3O^+ is the first, are constituents of the solvent and when there is a reaction equilibrium is established almost instantaneously (a few pico-seconds to a few seconds) with the acids and bases. This is also true for the hydroxylated surfaces of many solid soil compounds, including iron oxides. Therefore, even in the absence of strong acids or bases, pH is a well-defined variable, directly measurable in the soil solution with a pH electrode, for establishing the intensity of proton transfer reactions.

On the contrary, the redox equilibria are referenced against the normal hydrogen electrode, but the electron cannot flow freely in the aqueous medium because water is an insulator. Therefore, the pe of an aqueous solution is not related to a free chemical entity, but describes the tendency to accept or donate electrons from reactive compounds (Schüring et al., 2000). It is therefore necessary that an electron donor and an electron acceptor become close neighbors for a reaction to proceed and these reactions are relatively slow (a few seconds to several minutes or even an hour). Thus, measurement by a redox electrode, possible only in an aqueous medium, provides information on the redox state of the soil only through measurement of the ratio of the activities of certain electroactive redox couples such as $\text{Fe}^{2+}/\text{Fe}^{3+}$, $\text{H}_2/\text{H}_2\text{O}$ or $\text{H}_2\text{S}/\text{S}^0$. Another difficulty lies in the fact that in soils, some oxidation–reduction reactions are homogeneous in aqueous solution, but many others are heterogeneous and involve both solids and the solution. Thus Sposito (1981) defined a critical electron potential, $\text{pe}_{\text{critical}}$, computed at a given pH and for a given ratio of the activities of Don_e and Acc_e . Accordingly, for homogeneous reactions, the $[\text{Don}_e]/[\text{Acc}_e]$ ratio is taken conventionally to be 10^6 . For heterogeneous reactions, where the oxidized species is a solid and dissolves during reduction, $[\text{Don}_e]$ is taken as 10^{-7} , with the assumption that the solid phase is pure and its activity $[\text{Acc}_e]$ is equal to 1. For pe smaller than the “critical” value, the reduction reaction is considered as quasi-complete. In the particular case of iron, it is then possible to observe that over the entire water stability domain, there are redox couples that can control the dynamics of iron at the solid–solution interface (Fig. 7).

In the solid phase: Size of iron mineral

Iron oxides in soils are commonly fine-grained and poorly crystalline. Thermodynamically, the competition between surface enthalpy and the energetics of phase transformation leads to the general conclusion that polymorphs, which are metastable as micrometer-sized or larger crystals can often be stabilized at the nanoscale. This question is important because each material has unique physical properties, chemical reactivity and bioavailability. Furthermore, physical and chemical properties change with particle size and degree of hydration (Navrotsky et al., 2008). Advances in nanosciences, new structural spectroscopies and

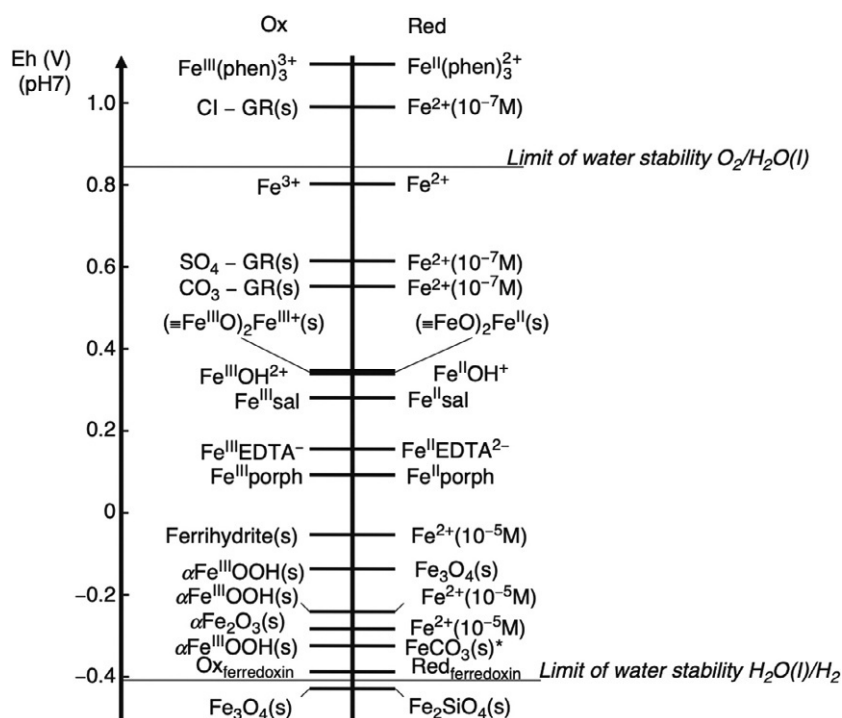


Fig. 7 Scale of “critical” values of E_h for major Fe redox couples, computed at pH = 7 and valid for $[\text{HCO}_3^-] = 10^{-3} \text{ M}$; phen, phenanthroline; sal, salicylate; porph, porphyrine and GR, green rust. From Trolard F, Bourrié G (2008) Chapter 5: Geochemistry of green rusts and fougérite: A reevaluation of Fe cycle in soils. *Advances in Agronomy* 99: 227–289. [https://doi.org/10.1016/S00652113\(08\)00405-7](https://doi.org/10.1016/S00652113(08)00405-7).

thermodynamic approaches are now enabling knowledge of structural details, thermodynamics and reactivity, particularly of nano-scaled iron oxides, to be acquired.

The structure of nanoparticles often varies as a function of their size and the surrounding medium. For example, hematite nanoparticles may vary in structure showing tetrahedral defects near the surface with a maghemite-like structure. As another example, it has been shown that the surface regions in poorly crystalline lepidocrocite are sufficiently different from the bulk for their Mössbauer signals to be distinguished and referenced to a lepidocrocite–hematite mixture (De Grave et al., 1986); or when the crystal size of akaganeite becomes sufficiently small, akaganeite may contain goethite-like structure possibly related to the collapse of exposed tunnels (Deore, 2007). The structure of ferrihydrite is still in debate: is it a single phase with iron both octahedrally and tetrahedrally coordinated, or a mixture of phases with variable structure and crystallinity (Carabello et al., 2022)? Finally, GR and fougérite with a layered double hydroxide structure crystallize mainly in the form of nanoparticles due to the relative stability of the hydroxide sheet compared to other polymorphs of iron oxides. Moreover, its anionic clay structure gives unique properties to GRs and fougérite that allows them to accommodate or interact with more than 30 chemical elements and molecules, including organic ones. Their sheet structure also represents a configuration conducive to the realization of certain oxidation–reduction reactions (see references included in Trolard et al., 2022).

Small particles are expected to have higher enthalpies and free energies than large crystals because of a positive surface energy. In contact with water, the nanoparticles hold on to their water tenaciously. As metastability of the coarse phase increases, its surface enthalpy decreases. The increasing metastability of the bulk polymorph leads to crossovers in enthalpy (and also in free energy) of polymorphs at the nanoscale. With thermodynamic data (Table 2), values of surface enthalpies also correlate with the heat of surface hydration and the fraction of strongly bound water (Navrotsky et al., 2008).

For example, the most stable iron oxide, hematite, absorbs water the most strongly; the water fraction that is strongly bound on goethite is 60%, 40% for akaganeite and lepidocrocite. Thus, maghemite becomes stable with respect to hematite and there are complex crossovers for the FeOOH polymorphs.

Surface properties of iron “oxides”

“Oxides” (s.l. including oxides, hydroxides and oxyhydroxides) can contribute, under certain conditions, to the characteristic surface charge of a material. Their hydroxylated surfaces take part in the reactions of dissociation or association of protons. The surface potential is thus governed by the adsorption of H^+ and OH^- ions, called potential determining ions (PDI). The properties of such minerals are different from those of clay minerals. In the case of clays, the surface properties are governed by the surface charge due to isomorphic substitutions in the layers by ions having a valence generally lower than those of the replaced ions, thus creating a charge imbalance. The surface charge is thus fixed by the rate of substitution in the mineral, independently of the pH and ionic strength of the electrolyte (Jolivet, 2015).

In the case of iron oxides, iron does not participate directly in the surface charge (Fig. 8). Surface charge is variable and is determined by the chemical composition of the solution with which the oxide equilibrates. Depending on the pH, iron oxides can trap either A^- anions, or A^+ cations and also neutral species.

Table 2 Thermodynamic data for some iron oxides.

Oxide	ΔH_f° (kJ.M ⁻¹)	S° (J.M ⁻¹ .K ⁻¹)	ΔS° (J.M ⁻¹ .K ⁻¹)	ΔG_f° (kJ.M ⁻¹)	ΔH_s^h (J.m ⁻²)	ΔH_s (J.m ⁻²)
Hematite, α -Fe ₂ O ₃	-826.2 ± 1.3	87.4 ± 0.2	-274.5 ± 0.3	-744.4 ± 1.3	0.75 ± 0.16	1.90 ± 0.3
Maghemite, γ -Fe ₂ O ₃	-811.6 ± 2.2	93.0 ± 0.2	-268.9 ± 0.3	-731.4 ± 2.0	0.57 ± 0.10	0.71 ± 0.13
Goethite, α -FeOOH	-561.5 ± 1.5	59.7 ± 0.2	-237.9 ± 0.2	-490.6 ± 1.5	0.60 ± 0.10	0.91 ± 0.09
Lepidocrocite, γ -FeOOH	-552.0 ± 1.6	65.1 ± 0.2	-232.5 ± 0.2	-482.7 ± 3.1	0.40 ± 0.16	0.62 ± 0.14
Akaganeite, β -FeOOH	-554.7 ± 1.9	53.8 ± 3.3	-246.2 ± 3.3	-481.7 ± 1.9	0.34 ± 0.04	0.44 ± 0.04

Enthalpies of formation (ΔH_f°) and Gibbs free energies of formation (ΔG_f°) are for conditions of 298.15 K and 1 bar. Surface enthalpies are given for hydrated (ΔH_s^h) surfaces.

From Navrotsky A, Mazeina L, and Majzlan J (2008) Size-driven structural and thermodynamic complexity in iron oxides. *Science* 319(5870): 1635–1638. <https://doi.org/10.1126/science.1148614>.

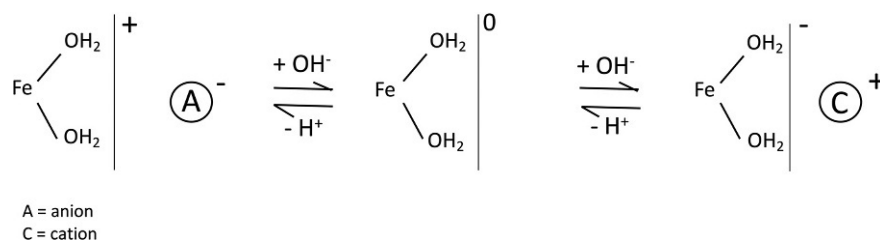


Fig. 8 Schematic representation of the surface charge of iron “oxides.” H^+ and OH^- are the potential determining ions (PDI).

At acidic pH, iron oxides are positively charged and therefore have an appreciable anion exchange capacity, when present in sufficient quantity in soils. In an alkaline medium, the capacity of iron oxides to absorb OH^- gives them, when they are in sufficient quantity, a buffering power, which can be significant.

Organic matter is negatively charged over a wide range of pH, giving soils a negative net surface charge in the superficial horizons. However, if the amounts of soil organic matter have been greatly reduced by agricultural practices or deforestation, for example, the positive surface charge of iron oxides can take over. In this situation, the soil has a low capacity for the retention of cationic nutrients (e.g., Ca, K, Na, . . .), which are then easily leached. This is an important problem in tropical and equatorial zones where soils develop in a hot climate marked by heavy rainfall. One of the methods proposed to increase the cation exchange capacity is amendment with anionic silicates which, by adsorbing preferentially on the iron oxides, relatively increase the net negative surface charge of the soil by reducing the positive charge of the “oxide.”

Evidence of in situ dynamics of iron minerals

Fast mineralogical transformations can be observed in situ at well-defined points in a soil, at different times and depths by a portable Mössbauer spectrometer. Observations were made in hydromorphic soils of Fougères's forest in Brittany (France). During alternating aerobic and anaerobic conditions, the soil color changes from ochre to blue in the horizon deeper than 70 cm. Fougérite is the almost exclusive iron mineral of the system, except during the summer when lepidocrocite can occur. Seasonal variations in the $\text{Fe(III)}/\text{Fe}_{\text{total}}$ ratio at a given point and depth are observed; the variations remain in the interval $[1/3; 2/3]$ of the stability domain of fougérite. The more ferric minerals are observed in the upper oximorphic soils, whereas the more ferrous forms are present in the lower reductimorphic horizons. At a given depth, rapid temporal variations of this ratio are correlated with fluctuation of the water table, causing alternating aerobic and anaerobic conditions. Monitoring by Mössbauer spectrometry in situ, at a given depth, also highlighted the crystallization of fougérite in less than a week in sufficient concentrations for a signal to be recorded and compatible with the Fe^{2+} contents and the volume of soil solution available at this location (Féder et al., 2005). This demonstrates that fast and reversible mineralogical transformations of iron oxides can exist in these soils. Fast geochemical phenomena in soils are thus not restricted to ion exchange or precipitation or dissolution of salts, as classically considered.

Another way to show that iron oxide mineralogy can be a good indicator of soil evolution is by making measurements of iron isotopes. Iron isotopes are geochemically robust and retain their environmental signature even under extreme metamorphic conditions, which makes them a useful tool in deciphering the biogeochemical history of highly deformed and metamorphosed rocks (Peacock et al., 2016). Iron has four naturally occurring stable isotopes ^{54}Fe (5.84%), ^{56}Fe (91.7%), ^{57}Fe (2.12%) and ^{58}Fe (0.28%), where stable isotope data are commonly reported as $^{56}\text{Fe}/^{54}\text{Fe}$. In soils, metabolically processed Fe may also retain a “vital effect” due to the redox related changes that control Fe isotope fractionation during iron mineral transformations. Thus, a soil that has undergone or has been subjected to in its history alternations of anaerobiosis–aerobiosis processes resulting in the reduction or the oxidation of iron will show a distribution of iron isotopes at the level of the reaction fronts with:

- an enrichment in heavy isotopes (i.e., ^{56}Fe) in the solid fraction, which undergoes dissolution;
- an enrichment in light isotopes (i.e., ^{54}Fe) in the solid fraction, which precipitates from the Fe(II) of the soil solution.

In an upstream–downstream soil sequence with a hydromorphic front, the $\delta^{56}\text{Fe}$ signature will tend to be high in the upstream part of the sequence and in the superficial horizons and lower in the downstream part of the sequence and in the deeper horizons.

Influence of Fe on landscape evolution

Iron accumulates as Fe(III) oxides, first because most other elements are more mobile, this is a relative accumulation; second, because under reducing conditions Fe(II) is as soluble as Mg(II) . At short distances, in lower parts of the landscape, it precipitates; this is absolute accumulation. Ferricretes form in both cases. As ferricretes are more resistant to erosion, this leads to the formation of relief where former deep horizons or former lowland formations are now present at high elevations, such as in Burkina Faso (Fig. 9).

These high-level terraces are the relics of paleosurfaces that have undergone a long duration weathering of different bedrocks: gabbros, micascists, granites, under a rainforest. They are thus indicators of changes in climate. However, in some other cases, as in NW Hoggar (Mouydir) ferricretes are specifically located on pyrite-rich schists either in alluvial fans or directly on rock outcrops at mid slope positions. In this case, Fe mobility is due to sulfide oxidation into sulfuric acid that results in a very acidic pH (ca. 2). Iron mobility is due, therefore, to local conditions and these ferricretes are not indicators of a past humid paleoclimate.

The Fe(III) oxides appear as nanoparticles that are positively charged in acid conditions, as their pH of zero point of charge is large. These particles are linked with clay minerals, and this gives rise to stable aggregates, and favorable physical properties (large porosity and hydraulic conductivity), typical of ferrallitic and fersiallitic topsoils. These soils typically occur on plateaus. Hydromorphic conditions may develop in depressions at the centers of these plateaus. This results in Fe reduction and destruction of the links between clay minerals and Fe oxides. Soil structure collapses, porosity decreases and also hydraulic conductivity. This, in turn, results in poor drainage and a further decrease of redox potential. Progressively, hydromorphic features extend from the center of the plateau to the entire plateau, with soil color changes (transformation from red soils to beige soils) and a geomorphological transformation of slopes, as observed in West Africa (Chauvel and Pedro, 1978) (Fig. 10).



Fig. 9 Inversion in the landscape of iron formed in an old valley in Burkina Faso (Ruellan).

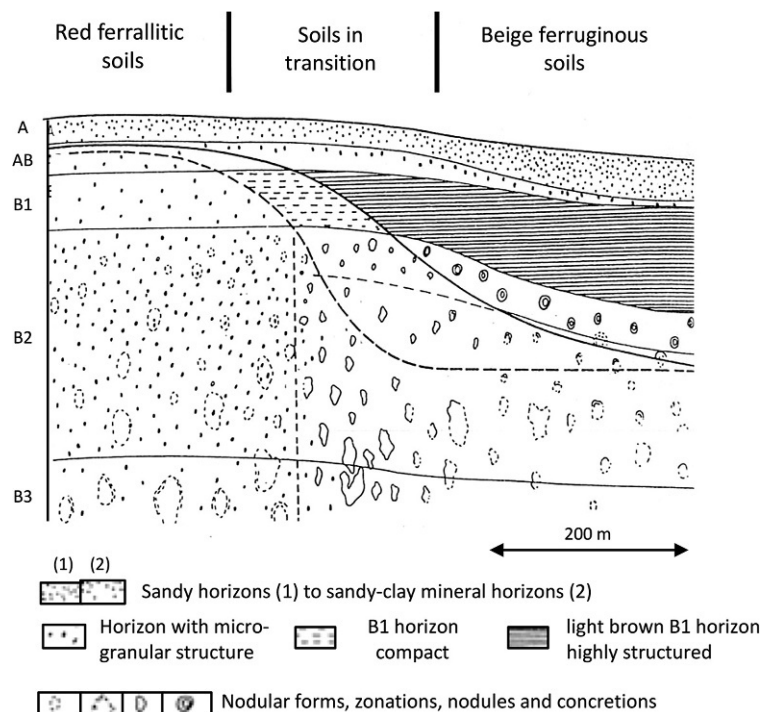


Fig. 10 Transformation of the mineralogy of the iron oxides by weathering in a landscape in Senegal. Modified from Chauvel A and Pédro G (1978) *Genèse de sols beiges (ferrugineux tropicaux lessivés) par transformation des sols rouges (ferrallitiques) de Casamance*. Cahiers Orstom, série Pédologie, XVI, 3: 231–249.

Interaction iron minerals—Living organisms

It is hard to find a biological process or environment in which iron do not participate. From the Earth's surface to the deep oceans, from pigeon brains to magnetotactic bacteria, anhydrous and hydrated iron oxides are ubiquitous. Iron oxides and oxyhydroxides in rocks and soils, and products of bacterial processes are also an essential source of nutrients for life. Iron is present in all electron transfer complexes, photosystems PSI and PSII, cytochrome *b*/*f*6 complex and ferredoxins. Respiration and photosynthesis rely on

Fe(II)-Fe(III) electron transfer. The role of Fe is also implied in the biogenesis of cofactors, such as heme (iron protoporphyrin XI), a tetrapyrrole cofactor of hemoglobin, and iron–sulfur clusters. Iron deficiency results in chlorosis in plants, because of the lack of chlorophyll synthesis, but this is modulated by P deficiency through complex regulations, which indicate that signals are exchanged between chloroplasts and the cell nucleus (Nam et al., 2021).

The biological importance of iron is a result of its electronic structure, which can reverse changes in oxidation state over a wide range of redox potentials (see Fig. 7). The ease of reversibility of the $\text{Fe}^{2+}/\text{Fe}^{3+}$ couple makes iron a cofactor that is frequently incorporated into protein complexes to catalyze oxidation–reduction reactions. Iron is thus an essential element present in all biological tissues and in all cellular compartments. Iron proteins have essential functions in plant metabolism such as photosynthesis, respiration, DNA synthesis, nitrogen and sulfur metabolism, etc. When incorporated into proteins, iron can be coordinated in different ways: either it is linked to sulfur atoms (iron–sulfur center), or it is linked to nitrogen atoms in the form of heme or directly linked to amino acids.

Key role of siderophores

In general, the availability of iron in the soil solution for the nutrition of living organisms is low or even very low, far below that required for optimal growth: 10^{-5} to 10^{-7} M for microbes and 10^{-4} to 10^{-9} M for plants (Robin et al., 2008). In the soil, iron is mainly present in the form of ferric oxides and oxyhydroxides, which are very stable and poorly soluble at circumneutral pH. In an aerobic aqueous solution, the concentration of free ferric iron (Fe^{3+}) is often less than 10^{-10} M. Many bacteria and fungi, but also the plants, particularly in iron poor soils, get around this impasse by excreting low molecular weight, Fe(III)-specific ligands known as siderophores (from Greek: “iron carrier”). In soils, siderophores or their breakdown products can be so abundant that they dominate ferric iron components. Siderophores have several properties that make them ideal Fe(III) chelators. They usually form a stable, hexadentate, octahedral complex preferentially with Fe^{3+} compared to other metal ions in soil, although they have fewer than six donor atoms. They can also coordinate with water. The most effective siderophores are those that have three bidentate ligands per molecule, forming a hexadentate complex around the central Fe^{3+} cation resulting in a smaller entropic change than that caused by chelating a single ferric ion with separate ligands. The Fe^{3+} is a strong Lewis acid, preferring to coordinate with strong Lewis bases such as anionic or neutral oxygen atoms.

The major groups of siderophores include the catecholates (phenolates), hydroxamates and carboxylates (e.g., derived from citric acid, which can also act directly as a siderophore). They form especially strong 1:1 surface complexes, and their association constant for Fe(III) can be several times higher than common soil organic acids, e.g., oxalic acid. This important property maintains ferric iron in a soluble form that minimizes its loss from the aqueous environment by precipitation of ferric solid oxide and oxyhydroxide.

The wide diversity of siderophores biosynthesized by microorganisms and plants seems to have two objectives:

- produce structurally different siderophores, which cannot be fixed by other microbes' or plants' specific active transport systems, or in the case of pathogens cannot be deactivated by the host organism;
- by a specific signature ensures an exclusive access to a given source of ferric iron-containing mineral.

Pathways for iron in the rhizosphere—Plant system

In the rhizosphere, in particular, the volume of soil surrounding roots where microorganisms are influenced by the living roots the competition for iron is very high, and the organisms have evolved active uptake strategies to increase the availability of iron (Fig. 11) by:

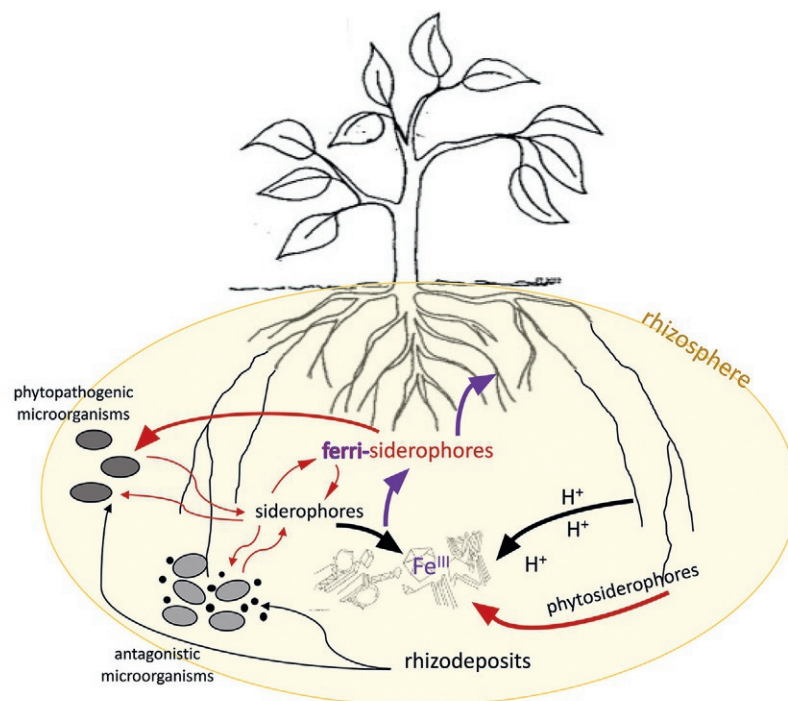
- acidification through proton extrusion (proton pump), organic acid secretions and an increase of pCO_2 by respiration;
- chelation through secretion of siderophores, whose biosynthesis is tightly controlled by available iron concentrations and;
- reduction through secretion of molecules with reducing properties or through the expression of a membrane-bound reductase activity.

These organic and inorganic substances, named rhizodeposits, exert a direct influence on soilborne microorganisms including antagonistic and phytopathogenic populations. Rhizodeposition affects soil microorganisms leading to (i) an increase in their density, biomass and activity; and (ii) variation and diversity of their populations. Energetic metabolism of heterotrophic microorganism is based on electron donors (organic compounds) and electron acceptors (oxygen, nitrogen oxidized species, e.g., nitrate, or ferric oxides). Microbial siderophores can also trigger defense responses in the host plant and promote plant iron nutrition.

Strategies of iron up-take by organisms

In the soil, iron is mainly found in the form of insoluble ferric iron oxide or oxy-hydroxides. Before absorbing iron from the soil, plants and soil organisms must solubilize it. There are two iron uptake strategies: (i) strategy I is found in all dicots and non-grass monocots, and (ii) strategy II in other organisms, such as grass monocots, bacteria and fungi (Robe, 2021).

(a)



(b)

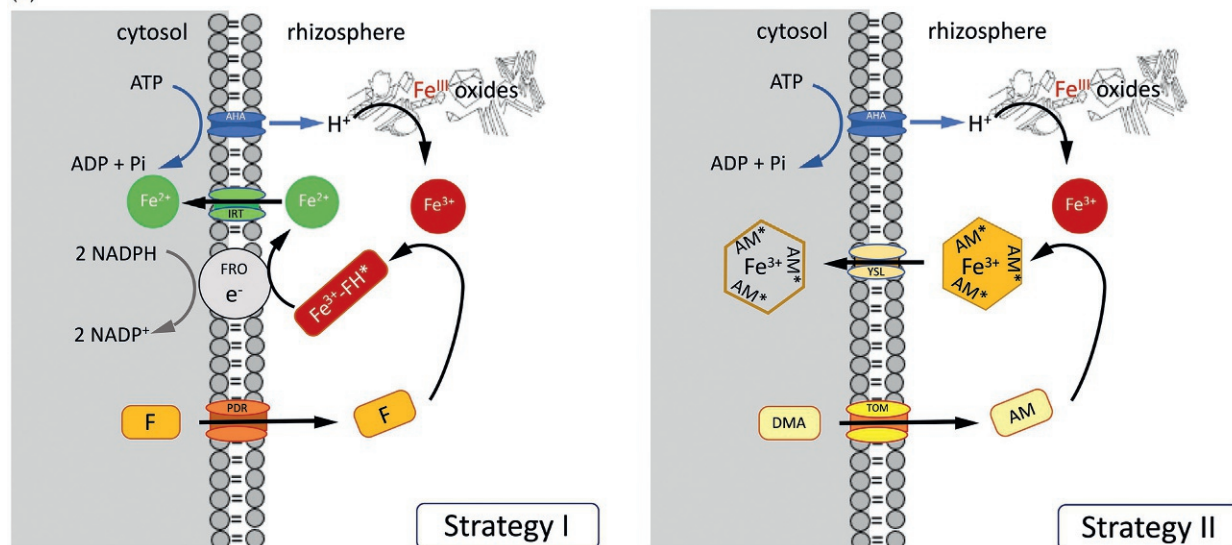


Fig. 11 Schematic representation of (a) iron-mediated interactions between plants and microbes promoting plant health and iron nutrition (b) the two strategies for iron uptake by organisms. Panel (a): Modified from Robin A, Vansuyt G, Hinsinger P, Meyer JM, Briat JF, Lemanceau P (2008) Chapter 4: Iron dynamics in the rhizosphere: Consequences for plant health and nutrition. *Advances in Agronomy* 99: 183–213. [https://doi.org/10.1016/S00652113\(08\)00404-5](https://doi.org/10.1016/S00652113(08)00404-5). Panel (b): Modified from Robe K (2021) Localisation, dynamique et rôle des coumarines dans la nutrition en fer chez *Arabidopsis thaliana*, Thèse de Doctorat, Montpellier SupAgro, Université de Montpellier, 407 pp.

Strategy I consists of removing iron from the soil by acidifying the external environment and thus solubilizing the iron so that it can be transported into the cell. Excretion of protons by H^+ -ATPases (AHA) is quite rapid. Roots can lower soil pH to around 3 within hours, increasing iron solubility by a factor of 1000.

In a situation of iron deficiency, the plant can emit phytosiderophores (e.g., coumarins, flavins, etc.) capable of chelating the Fe^{3+} released by acidification. The iron Fe^{3+} is subsequently reduced to ferrous iron (Fe^{2+}) by membrane reductases FRO (ferric reductase oxidase). Finally, ferrous iron is transported into the root by membrane transporters with a high affinity for Fe^{2+} , and IRT (iron regulated transporter) located on the plasma membrane of the root epidermis (Fig. 11b).

Recent work has shown that IRT, FRO and AHA interact physically at the plasma membrane, that their physical proximity ensures optimal iron acquisition and that their activities are proportional to the availability of iron in the rhizosphere, i.e., the more that iron from the soil is scarce, the higher are their activities.

Strategy II does not require the reduction of Fe^{3+} to Fe^{2+} (Fig. 11b). In this case the organisms secrete siderophores capable of dissolving iron oxides and oxyhydroxides and of directly chelating the released Fe^{3+} . The production of siderophores is all the more important because iron is rare in the living environment of the organisms. Among the phytosiderophores, mugineic acids (MA) are produced by plants and are secreted by the transporter TOM ("transporter of mugineic" acid family phytosiderophore). The ferric iron complex thus chelated is then assimilated by the plant using the YS/YSL transporter (Yellow Stripe/Yellow Stripe Like).

Among plants, rice (*Oryza sativa* L.) has both Strategy I and Strategy II iron acquisition components. Although rice mainly uses the Strategy II pathway, an Fe^{2+} transporter was identified and a slight ferric reductase activity was measured. These specificities can be explained by the fact that rice often grows in anoxic flooded environment where Fe^{2+} can be present; the presence of an Fe^{2+} -transporter under these conditions is therefore an advantage for the plant.

Iron is an essential nutrient for all living organisms but it can lead to cytotoxicity when present in excess. In plants, classic symptoms of Fe toxicity are leaf discoloration (bronzing) and a stunted root system.

Excessive Fe generates oxidative stress inside two important cell compartments, i.e., mitochondria and plastids, in organisms including plants. Surplus free Fe produces hydroxyl radicals through the Fenton reaction, displays enhanced production of ROS and stimulation of oxidative stress-detoxifying enzymes. The unusually elevated Fe content and production of ROS in mitochondria lead to increased production of NO in frataxin-deficient plants. This enriched production of NO has been observed to be a crucial feature to reduce high levels of oxidative damage in the root cells of plants. Interestingly, NO accumulation may, perhaps, protect plants from oxidative stress either by direct scavenging of O_2^- or indirectly by NO-induced expression of ferritin genes (AtFer1 and AtFer4). Thus, NO certainly plays an essential role in reducing the levels of free Fe within the plant cell organelles; further, it maintains a check on the ROS production with subsequent protection of plants from oxidative stress and apoptosis (see references included in Zaid et al., 2020).

Iron toxicity often occurs in rice crops in the context of submerged paddy fields where redox potential decreases, leading to increases in the concentration of ferrous ions, disrupting cell homeostasis and impairing leaf and root growth and yield.

Conclusion

Iron in the soil is an essential chemical element that contributes both to the structure of soils and to their organization in landscapes. It plays an essential role as a trace element and nutrient for all organisms that feed on soil constituents. It is virtually insoluble in the ferric state at pH close to neutral, and it is expressed in the mineral form of oxides, hydroxides and oxy-hydroxides. Iron can also become as soluble as an alkaline or alkaline-earth cation in its reduced form Fe^{2+} , when the milieu becomes moderately reducing. This duality of insoluble/soluble and oxidized/reduced redox state of iron is a cornerstone of soil formation both zonally and azonally in all surficial Earth formations. These properties are also essential in the mechanisms put in place by organisms to obtain the quantity of iron they need to live from soils where a priori the bioavailability of iron is low.

References

- Bourman RP (1993) Perennial problems in the study of laterites: A review. *Australian Journal of Earth Sciences* 40: 387–401.
- Bourrié G, Maitre V, and Curmi P (1994) *Mise en évidence de deux dynamiques saisonnières du fer dans les sols hydromorphes en climat tempéré*. Paris: Comptes Rendus de l'Académie des Sciences 87–92. 318–II.
- Carabello MA, Asta MP, Perez JPH, and Hochella MF (2022) Past, present and future global influence and technological applications of iron-bearing metastable nanominerals. *Gondwana Research*. <https://doi.org/10.1016/j.gr.2021.11.009>.
- Chauvel A and Pédro G (1978) *Genèse de sols beiges (ferrugineux tropicaux lessivés) par transformation des sols rouges (ferrallitiques) de Casamance*. *Cahiers Orstom, série Pédologie*, XVI 3: 231–249.
- Cornell RM and Schwertmann U (2003) *The Iron Oxides: Structure, Properties, Reactions, Occurrences and Uses*, 2nd edn. vol. 664. Weinheim, Germany: Wiley. VCH edition, 148 pp.
- De Grave E, Persoons RM, Chambaere DG, Vandenberghe RE, and Bowen LH (1986) An ^{57}Fe Mössbauer effect study of poorly crystalline $\gamma\text{-FeOOH}$. *Physics and Chemistry of Minerals* 13: 61–67.
- Deore SW (2007) *A thermodynamic and structural study of atomistic, nano and bulk systems*. PhD-Thesis Davis: University of California. publication number: AAI3261148; source: dissertation, abstracts international, volume 68–04, section B, 238 pp.
- Féder F, Trolard F, Klingelhöfer G, and Bourrié G (2005) In situ Mössbauer spectroscopy: evidence for green rust (fougerite) in a gleysol and its mineralogical transformation with time and depth. *Geochimica et Cosmochimica Acta* 69(18): 4463–4483. <https://doi.org/10.1016/j.gca.2005.03.042>.
- Jolivet JP (2015) De la solution à l'oxyde—chimie aqueuse des cations métalliques, synthèse des nanostructures. 2ème édition In: *Edp Sciences—Savoirs Actuels*. Sciences et Techniques. 400 pp.
- Leprun JC (1979) *Les cuirasses ferrugineuses des pays cristallins de l'Afrique Occidentale sèche : genèse, transformation et dégradation*. *Sciences Géologiques Mémoire* 58: Strasbourg: University of Strasbourg. 224 pp.
- Mérot P and Bruneau P (1993) Sensitivity of bocage landscapes to surfaces run-off: Application of Kirby index. *Hydrological Processes* 7: 167–176.
- Millot G and Bonifas M (1955) Transformations isovolumétriques dans les phénomènes de latéritisation et de bauxitisation. *Bulletin Service de la Carte Géologique d'Alsace-Lorraine* 8: 3–10.
- Nam HI, Shahzad Z, Dorone Y, Clovez S, Zhao K, Bouain N, Lay-Pruitt KS, Cho H, Rhee Y, and Rouached H (2021) Interdependent iron and phosphorus availability controls photosynthesis through retrograde signaling. *Nature Communications* 12(7211). <https://doi.org/10.1038/s41467-021-27548-2>.
- Navrotsky A, Mazina L, and Majlan J (2008) Size-driven structural and thermodynamic complexity in iron oxides. *Sciences* 319: 1635–1638.

- Peacock CL, Lalonde SV, and Konhauser KO (2016) Chapter 6: Iron minerals as archives of Earth's redox and biochemical evolution. In: Ahmed IAM and Hudson-Edwards KA (eds.) *Redox-Reactive Minerals: Properties, Reactions and Applications in Natural Systems and Clean Technologies*, EMU Notes in Mineralogy 17: 121–172. <https://doi.org/10.1180/EMU-notes.17.14>.
- Robe K (2021) *Localisation, dynamique et rôle des coumarines dans la nutrition en fer chez Arabidopsis thaliana*. Thèse de Doctorat Montpellier SupAgro: Université de Montpellier. 407 pp.
- Robin A, Vansuyt G, Hinsinger P, Meyer JM, Briat JF, and Lemanceau P (2008) Chapter 4: Iron dynamics in the rhizosphere: Consequences for plant health and nutrition. *Advances in Agronomy* 99: 183–213. [https://doi.org/10.1016/S00652113\(08\)00404-5](https://doi.org/10.1016/S00652113(08)00404-5).
- Schüring J, Schulz HD, Fischer WR, Böttcher J, and Duijnvisveld WHM (eds.) (2000) *Redox—Fundamentals, Processes and Applications*, p. 250. Berlin-Heidelberg: Springer.
- Soil Survey Staff (1999) Soil Taxonomy. A basic system of classification for marking and interpreting soil surveys. In: *Agric. Handboock 436*, 2nd edn. Washington, DC: Natural Resources Conservation Service, United States Department of Agriculture.
- Sposito G (1981) *The Thermodynamics of Soil Solutions*. Oxford: Oxford University Press. 223 pp.
- Tardy Y (1997) *Petrology of Laterites and Tropical Soils*. Taylor & Francis. ISBN: 978-90-5410-678-4. 419 pp.
- Trolard F (1988) Physico-chimie des cuirasses latéritiques. Domaine de stabilité des oxydes et hydroxydes de fer et d'aluminium. In: *Sciences Géologiques Mémoire*, 81. 131 pp.
- Trolard F (1997) Les "oxydes" de fer des latérites et des sols hydromorphes—géochimie, minéralogie et modélisations thermodynamiques. In: *Agrégation de l'Enseignement Supérieur*. Université Catholique de Louvain (Belgium). 283 pp.
- Trolard F and Bourrié G (2008) Chapter 5: Geochemistry of green rusts and fougérite: A reevaluation of Fe cycle in soils. *Advances in Agronomy* 99: 227–289. [https://doi.org/10.1016/S00652113\(08\)00405-7](https://doi.org/10.1016/S00652113(08)00405-7).
- Trolard F, Bourrié G, Abdelmoula M, Refait Ph, and Feder F (2007) Fougérite, a new mineral of the pyroaurite – iowaite group TO7: Description and crystal structure. *Clays and Clay Minerals* 3: 323–334.
- Trolard F, Duval S, Nitschke W, Menez B, Pisapia C, Ben Nacib J, Andréani M, and Bourrié G (2022) Mineralogy, geochemistry and occurrence of fougérite in a modern hydrothermal system and its implications for the origin of life. *Earth Science Reviews* 225, 103910.
- WRB (2022) *World Reference Base for Soil Resources—International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*, 4th edn. Vienne, Austria: International Union for Soil Sciences (IUSS).
- Zaid A, Ahmad B, Jalleel H, Wani SH, and Hasanuzzaman M (2020) Chapter 4: A critical review on iron toxicity and tolerance in plants: Role of exogenous phytoprotectants. In: Aftab T and Hakemm KR (eds.) *Plant Micronutrients—Deficiency and Toxicity Management*, pp. 83–100. Springer. ISBN: 978-3-030-49855-9. <https://doi.org/10.1007/978-3-030-49855-9>.

Sulfur and selenium in soil

Dongli Liang^{a,b}, Fei Zhou^a, and Mingxing Qi^a, ^aCollege of Natural Resources and Environment, Northwest A&F University, Yangling, Shaanxi, China; ^bKey Laboratory of Plant Nutrition and the Agri-environment in Northwest China, Ministry of Agriculture and Rural Affairs, Yangling, Shaanxi, China

© 2023 Elsevier Ltd. All rights reserved.

Introduction	192
Sulfur in the soil	194
Sulfur sources in soil	194
Sulfur contents in soil	194
Sulfur speciation in soil	194
Transformation processes of sulfur in soil	194
Mineralization of organic sulfur	195
Immobilization of inorganic sulfur	195
Oxidation and reduction of sulfur in soil	195
Adsorption and desorption of sulfate by soil	195
Availability of sulfur in soil	195
Selenium in soil	196
Selenium sources in soil	196
Selenium contents in soil	196
Selenium speciation in soil	196
Selenium species in soil	196
Distribution of selenium fractions in soil	196
Transformation processes of selenium in soil	197
Adsorption-desorption reaction	197
Dissolution-precipitation reaction	197
Oxidation-reduction reaction	198
Complexation reaction	198
Volatilization reaction	198
Bioavailability of selenium in soil	198
Interaction between sulfur and selenium in soil	198
Conclusion	199
References	199

Abstract

Sulfur (S) plays an irreplaceable role in the biological functions of plants. Selenium (Se) is an essential micronutrient for humans and a beneficial element for plant growth. Because S and Se have similar chemical properties, exploring their behavior and interactions in soil are of great importance to clarify the S and Se supply potential of soil and guide possible Se biofortification in agricultural production. The physicochemical properties, sources, content, speciation distribution, chemical transformation, and bioavailability of S and Se in soil are summarized in this chapter to show their similarities and differences and corresponding mechanisms in soil.

Key points

- Summarize the physicochemical properties, sources, content, speciation distribution, chemical transformation, and bioavailability of S and Se in soil.
- Clarify the interaction mechanisms between Se and S and the regulatory measures of their bioavailability in soil.

Introduction

Sulfur (S) and selenium (Se) are essential nutrients for humans and animals. Although Se is not necessarily an essential nutrient for plant growth like S, plants are the main source of Se intake for humans and animals. Soil is an important pathway for Se and S entry into the food chain through plants. Since Se and S belong to the same group in the periodic table of elements, they have similar chemical properties. Therefore, understanding the similarities and differences in the chemical properties and behavior of S and Se in soil is of great significance to regulate their bioavailability and promote crop growth.

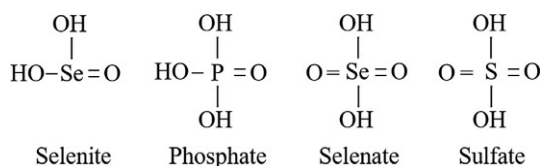
Table 1 Physicochemical properties of S and Se.

Element properties	S	Se
Periodic position	Group VIA, Period 3	Group VIA, Period 4
Atomic number	16	34
Atomic weight	32.065	78.96
Atomic radius	102 pm	120 pm
Stable isotopes	³² S, ³³ S, ³⁴ S, ³⁶ S	⁷⁴ Se, ⁷⁶ Se, ⁷⁷ Se, ⁷⁸ Se, ⁸⁰ Se, ⁸² Se
Melting point	385.95 K	494 K
Boiling point	717.75 K	958 K
Oxidation states	-2, 0, +4, +6	-2, -1, 0, +4, +6
Mohs hardness	2	2
Density	2.36 g/cm ³	4.81 g/cm ³
Electronic configuration	[Ne] 3s ² 3p ⁴	[Ar] 3d ¹⁰ 4s ² 4p ⁴
Electronegativity	2.58	2.54
Ionization potential	10.413 eV	9.752 eV
CAS#	7704-34-9	7782-49-2

Sulfur is a non-metallic element with atomic number 16 and CAS (Chemical Abstracts Service) accession number 7704-34-9. It is a group VIA element and lies in the third period of the periodic table of elements. Sulfur is active in nature and has many valence states; it usually exists in the form of sulfide (HS⁻), sulfate (SO₄²⁻), or elemental sulfur (S⁰). In general, S⁰ presents as yellow crystals and has many allotropes, such as rhombic, monoclinic, and elastic sulfur. Moreover, S⁰ is insoluble in water, slightly soluble in ethanol, and easily soluble in carbon disulfide. It is mainly used to produce fertilizer, gunpowder, lubricants, pesticides, and antifungal agents. Sulfur plays an irreplaceable role in the life activities of animals and plants; it is not only a medium-element for plants to synthesize important substances but also a participant in several metabolic processes of plants, such as respiration, protein synthesis, lipid synthesis, and sugar metabolism. However, S is also a chemical pollutant. Acid deposition, the greenhouse effect, and ozone layer depletion, for instance, are all related to S pollution.

Selenium is between arsenic (As) and bromine (Br) in the fourth period and between S and tellurium (Te) in group VIA of the periodic table of elements. The atomic number and mass of Se are 34 and 78.96, respectively, and its electronic configuration is [Ar] 3d¹⁰ 4s² 4p⁴ (Table 1). Elemental Se (Se⁰) is a red or gray powder that is insoluble in water and ethanol, soluble in concentrated sulfuric acid and chloroform, and slightly soluble in carbon disulfide. Although Se is a non-metallic element, it is conductive and, thus, often used as semiconductor and photosensitive material. In nature, Se mainly exists in two forms: inorganic and organic Se. Inorganic Se mainly includes selenate (SeO₄²⁻) and selenite (SeO₃²⁻), while organic Se mainly includes selenomethionine (SeMet), selenocysteine (SeCys) and methyl-SeCys (MeSeCys). As the active center of various selenoenzymes and selenoproteins, Se plays important physiological functions in regulating hormone metabolism in the thyroid gland, antagonizing heavy metals, and maintaining oxidation-reduction homeostasis. Therefore, Se is considered an essential micronutrient for animals and humans, and a beneficial element for plant growth. However, excessive intake of Se also has negative effects on human health, and the range between the necessary (40 µg day⁻¹) and toxic (>400 µg day⁻¹) doses of Se is narrow. Consequently, ensuring that the Se content of crops reaches an appropriate amount to meet human daily needs and avoid Se toxicity is crucial with the Se biofortification of crops.

The chemical properties of Se are very similar to those of S in terms of periodic position, oxidation states, atomic size, electronegativity, and ionization potential (Table 1, Fig. 1). Therefore, selenate can pass through the root cell membrane through SO₄²⁻ transport channels and compete with the latter for crop absorption. Because of defects in biological identification, Se absorbed by plants can replace S in S-containing proteins (e.g., methionine (Met)), thus nonspecifically incorporating it into a Se-containing protein (e.g., SeMet), or specifically incorporating it into selenoprotein as the active site of SeCys, thereby interfering with the normal functions of the protein (B'Hymer and Caruso, 2006). Selenite is mainly absorbed by crops through the transport channels of phosphate (PO₄³⁻) or other ions. The major difference between Se and S is that Se exists in the reduced quadrivalent form, while S exists in the oxidized quadrivalent form. In addition, the acid strength of Se is higher than that of S (i.e., coefficient of acidity [pKa]: hydrogen selenide (H₂Se) > hydrogen sulfide (H₂S)). Therefore, in view of the similarities of the chemical and physical properties of Se and S, understanding their distribution and transformation in soil and exploring the similarities and differences in their environmental chemical behavior in this matrix further are necessary to clarify the S supply potential of soil and guide the biofortification of Se in agricultural production.

**Fig. 1** Chemical structure of selenite, phosphate, selenate and sulfate.

Sulfur in the soil

Sulfur sources in soil

The inputs of S into soil include parent material, atmospheric deposition, irrigation water, and fertilization, etc. After the natural weathering of rock or minerals, the main S-containing minerals, such as gypsum, are decomposed by microorganisms in the soil and transformed into SO_4^{2-} . Under the comprehensive action of biology, physics and chemistry, SO_4^{2-} undergoes a series of migration and transformation processes and enters the soil through the absorption and decomposition of plants or runoff and tidal migration. Atmospheric S sources can be divided into natural S sources and anthropogenic S sources. Natural S sources are mainly gaseous S oxides and H_2S emitted from volcanic activities, crustal hydrothermal activities, and processes in waterlogged areas. Anthropogenic sources of S are gases from coal and crude oil combustion. Agricultural irrigation and fertilizer also contain some S to meet the needs of crop growth.

Sulfur contents in soil

Sulfur is the fourteenth most abundant element in the Earth's crust. The total S content in agricultural soil is usually between 0.03 g kg^{-1} and 10 g kg^{-1} , with the average is approximately 0.7 g kg^{-1} (Alloway, 2013). The total S in soil correlates strongly with soil organic C and total N, thereby suggesting that S is an integral part of soil organic matter. The content of S in soil varies with the original content of the soil parent material, changes in the soil formation process, climatic conditions, and land use types. The content of S in magmatic rocks is lower than that in sedimentary rocks, and its content in mineral soils is generally between 0.1 g kg^{-1} and 0.5 g kg^{-1} . In addition, the S content of mineral soils increases with increasing organic matter content. The total S content in subtropical cultivated soil is usually lower than that in temperate soil because of the leaching effect accompanied by strong weathering and lower organic matter content. The content of soil available S varies with the land utilization, which is in the order of paddy field > orchard > upland soil > forest land for these four land use types. Because organic matter accumulates easily in paddy soil and irrigation water can carry considerable amounts of S into soil, the available S in paddy fields is generally higher than that in upland soil. The abundance and deficiency of available S in soil can be divided into four levels: less than 10 mg kg^{-1} which is deficient, $10\text{--}16 \text{ mg kg}^{-1}$ which is potentially deficient, $16\text{--}30 \text{ mg kg}^{-1}$ is medium, and more than 30 mg kg^{-1} is abundant (Liu et al., 1990).

Sulfur speciation in soil

Sulfur in soil mainly exists in different oxidation states, namely, hexavalent (e.g., sulfate, S(VI)), tetravalent (e.g., sulfite, S(IV)), negative bivalent (S(-II)), and nonvalent (S(0)). In soil, S occurs in inorganic and organic forms.

In general, the majority of S in soil is bound to organic molecules, usually accounting for over 60% of the total S; it occurs mainly in soil organic matter (SOM) and animal or plant residues. At present, the specific chemical composition of organic S in soil is not well understood. According to the reduction method of hydroiodic acid (HI) proposed by Johnson and Nishita (1952), organic S can be divided into three forms: carbon-bonded S (C—S), sulfate esters (C—O—S), and unknown organic S (UO—S). Carbon-bonded S is the fraction of organic S bound directly to carbon (C), which can be reduced to H_2S by Raney nickel. This fraction consists of S-containing amino acids, thiols, sulfoxides, sulfinic acids, and sulfonic acid linked to an aromatic nuclei. The C—O—S is relatively stable and leads to lower S availability than C—S for plants. Organic S not directly bound to C, such as sulfonic acid esters, some sulfamic acids, and S-cysteine sulfonic acid, among others, is composed primarily of C—O—S , which is reduced to H_2S by HI (Lucheta and Lambais, 2012). This fraction accounts for 30–70% of the organic S and is an active component of organic S in soil. The C—O—S is easier to convert into inorganic S than C—S , and can be utilized by plants (Blair et al., 1991). The third fraction of organic S, UO—S or residual S, and its availability to plants require further study.

The species of S in soil mainly includes SO_4^{2-} , sulfite (SO_3^{2-}), sulfide (S^{2-}), and S^0 . Sulfur also can be divided into water-soluble, adsorbed, hydrochloric acid-soluble, and zinc-hydrochloric acid-reducing S according to the sequential extraction method. The main existing valence state of water-soluble S in soil is SO_4^{2-} , which mainly exists in soil solution. A small amount of reducing inorganic S, such as H_2S , HS^- , S^{2-} , and S , has also been noted, and the contents of these substances is usually less than 10%. Adsorbed S is retained on the surface of soil colloids by anion exchange adsorption and coordination adsorption (Wilhelm, 2009). Hydrochloric acid-soluble S is precipitated by the combination of calcium carbonate and magnesium carbonate in soil and can be dissolved under acidic conditions. Acid-volatile sulfide, a volatile sulfide that can be decomposed under acidic conditions, is the least common form of inorganic S in soil.

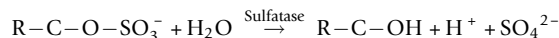
Transformation processes of sulfur in soil

The S pools in soils are dynamically balanced and controlled by complex mechanisms, including the immobilization of inorganic S forms in organic and mineral soil layers; these inorganic forms are usually incorporated into organic matter by covalent bonding. The mobilization of large organic S molecules into smaller ones and mineralization to yield plant-available inorganic S are other mechanisms controlling the S pools in soil.

Mineralization of organic sulfur

The S in the surface layer is mainly organic, but the S absorbed directly by plants from the soil is almost completely inorganic S. Organic S must be converted by biochemical (e.g., hydrolyzation of sulfate esters by different sulfatases) or microbiological (e.g., mineralization of C-bonded S) means to obtain inorganic SO_4^{2-} prior to plant uptake. The mineralization of organic S links organic and inorganic S in the soil and is an important means through which plants obtain available S; this process is also an essential link in the biogeochemical cycle of soil S. Studying the rate of mineralization and capacity of organic S in the soil helps in estimating the S supply capacity and evaluating the potential productivity of the soil.

The mineralization of organic S in soil can be divided into biochemical and biological mineralization. Soil microorganisms and plant roots produce a type of sulfatase when the content of SO_4^{2-} in soil is low. Biochemical mineralization refers to the release of SO_4^{2-} from the C—O—S pool through enzymic hydrolysis (Fitzgerald, 1978). The reaction process is as follows:



Biological mineralization occurs when microorganisms use compounds containing C-bonded S as C sources to provide energy. In this process, S is released as a by-product of the oxidation of C to CO_2 . Under aerobic conditions, organic matter is decomposed by microorganisms and organic S is oxidized into SO_4^{2-} . However, organic S is reduced to H_2S under anaerobic conditions. Many studies have shown that biological mineralization is the main process by which organic S is mineralized in soil, and the amount of mineralization at any time depends on the soil physicochemical properties, such as temperature, humidity, pH, nutrient status, addition of organic or inorganic substances, and plant growth.

Immobilization of inorganic sulfur

Immobilization of inorganic S and mineralization of organic S occur simultaneously in soil. The immobilization of inorganic S is affected by SO_4^{2-} and organic C contents of the soil. Most S is converted into the C—O—S when the content of SO_4^{2-} is high in the soil solution. By contrast, the immobilization of inorganic S is reduced. The extent of immobilization depends on the C:S ratio in the soil. When the C:S ratio is below 200, mineralization of organic S compounds is promoted; by contrast, C:S ratios between 200 and 400 result in no net changes of SO_4^{2-} concentration in the soil solution. Organic matter with a C:S ratio above 400 induces immobilization of SO_4^{2-} and renders it readily available to plants (Wilhelm, 2009; Lucheta and Lambais, 2012). Therefore, the addition of organic C can increase the amount of SO_4^{2-} immobilized in soil (Saggar et al., 1981).

Oxidation and reduction of sulfur in soil

The biochemical reactions of S in soil are characterized by oxidation and reduction, both of which are of great significance to the material recycling of the soil system. Reduced S^0 must be oxidized to SO_4^{2-} before it can be absorbed by plants. The oxidation of S in soil includes abiotic and biotic oxidation. Abiotic oxidation is mainly controlled by redox potentials, and low-valence S (e.g., S^{2-} , S^0 , SO_3^{2-}) is easily oxidized to SO_4^{2-} because of the low standard potential of S. However, S oxidation in soil is mainly achieved through biochemical oxidation, which differs from chemical oxidation through its involvement of microorganisms. These microorganisms are mainly S-oxidizing bacteria that can completely or partially oxidize low-valence reduced sulfides (e.g., H_2S , S, SO_3^{2-} , thiosulfate ($\text{S}_2\text{O}_3^{2-}$)) to higher-valence sulfides (e.g., final product SO_4^{2-}). The reduction of S in soil is also very important, and oxidation and reduction usually occur simultaneously. In highly reducible soil and sediments, SO_4^{2-} functions as an important electron acceptor and can be reduced to S^{2-} with the participation of microorganisms. The SO_4^{2-} -reducing bacteria exist widely in the environment and reduce high-valence S compounds (such as SO_4^{2-} , SO_3^{2-} , and $\text{S}_2\text{O}_3^{2-}$) and S^0 to low-valence sulfide (H_2S).

Adsorption and desorption of sulfate by soil

Sulfate (SO_4^{2-}) is the main acid-inducing component in acid deposition, and the ability of soil to adsorb and desorb this ion can effectively reflect the sensitivity and tolerance capacity of the soil to acid deposition. Adsorption of SO_4^{2-} in soil can result in the release of OH^- ions from the soil surface, thereby improving the soil's ability to neutralize acid, preventing cation leaching, and delaying the acidification of soil and surface water (Dail and Fitzgerald, 1999). The adsorption mechanisms of SO_4^{2-} are as follows: (1) binding with hydrated oxide, which is mainly characterized by specific adsorption; (2) ligand exchange between SO_4^{2-} and other surface groups of clay, during which SO_4^{2-} can replace hydrated or hydroxyl groups; (3) molecular adsorption; and (4) electrostatic adsorption. The latter two adsorption mechanisms may generally be described as nonspecific adsorption and follow the law of mass conservation (Singh, 1984). Desorption can be regarded as the reverse process of adsorption. Some studies show that soil properties, including iron oxide content, organic matter content, pH, degree of base saturation, aluminum saturation, soluble SO_4^{2-} content, and temperature are the main factors affecting the soil adsorption of SO_4^{2-} . The comprehensive effects of various physicochemical properties in soil could lead to changes in the adsorption capacity of this matrix for SO_4^{2-} .

Availability of sulfur in soil

Many chemical fractions of S exist in soil, and analyzing the fractions distribution of S in soil is of great significance in understanding the chemical cycle of S and its influence on the environmental behavior of other substances. Water-soluble S is a readily available fraction for plant uptake and utilization. Adsorbed S is stable and difficult to leach. The availability of hydrochloric acid-soluble S is less than that of water-soluble and adsorbed S. Available S is the S component in soil that can be absorbed directly

and utilized by plants and usually includes water-soluble S, adsorbed S, and some organic S. Soil available S is the main index used to measure soil S supply capacity. In general, when the available S content in soil is less than 10–16 mg kg⁻¹, crops may face the risk of S deficiency.

Selenium in soil

Selenium sources in soil

Weathering of parent materials and rocks is the major source of Se in soil. The Se content of different rocks is quite different, and the Se content in soil formed by sedimentary rocks, igneous rocks, or black shale is consistently higher than that of basalt and metamorphic rocks. Atmospheric deposition is another important natural source of soil Se. The Se in particles produced by volcanic eruptions and forest fires accounts for approximately 10% of the natural sources of Se in the atmosphere. Selenium released by bio-mediated volatilization accounts for ~30–50% of the annual total Se emissions (Zoller et al., 1974) and is another important source of Se in the atmosphere. Selenium entering the atmosphere eventually re-enters the soil through deposition. Coal mining, thermal power generation, combustion of fossil fuels, and the use of mineral fertilizers or Se-enriched organic materials is a key factor leading to Se in rock minerals or Se-containing agricultural wastes entering the soil, which are important anthropogenic sources of soil Se.

Selenium contents in soil

Selenium is a rare and dispersed element that lies between the oxygen-group element S and the metal element Te (tellurium). The Se content of the Earth's crust is very small. Considering the narrow range between the necessary and toxic doses of Se for humans, Tan (1989) defined a grading standard of total Se content in various soil areas as follows: Se < 0.125 mg kg⁻¹ refers to an Se-deficient area, Se = 0.125–0.175 mg kg⁻¹ refers to an Se-marginal area, Se = 0.175–0.45 mg kg⁻¹ refers to an Se-sufficient area, Se = 0.45–2.0 mg kg⁻¹ refers to an Se-rich area, and Se > 3.0 mg kg⁻¹ refers to a seleniferous area. Because it is influenced by soil parent materials, climate (e.g., temperature and precipitation), and topographic conditions (e.g., slope and altitude), the global distribution of soil Se is extremely uneven and seleniferous and Se-deficient areas exist side by side. The Se content in most soils in the world ranges from 0.01 to 2.00 mg kg⁻¹, with an average of 0.40 mg kg⁻¹, although the soil Se content in some seleniferous areas can reach 1200 mg kg⁻¹. The Se-rich areas are mainly in Wyoming and South Dakota in the United States (20–40 mg kg⁻¹; Gerla et al., 2011), Enshi and Ziyang in China (36.69–79.08 mg kg⁻¹; Dinh et al., 2018), Punjab in India (Dhillon and Dhillon, 2003), and several other regions. However, approximately 40 countries and regions in the world are Se-deficient; these countries and regions are mainly in the middle latitudes and latitudes above 30 degrees in both the southern and northern hemispheres (e.g., Finland, central Serbia, and parts of the Congo). Overall, the global area of Se-deficient soil is larger than that of Se-rich soil (Dhillon and Dhillon, 2003).

Selenium speciation in soil

The chemical forms and valence of soil Se affect whether it can be absorbed and utilized by plants. Under natural conditions, Se in soil often exists in different chemical forms and valence states, and the fractions or valence states of Se could transform with changes in environmental conditions.

Selenium species in soil

Selenium is a valence-variable element that often exists in four valence states, namely, hexavalent (e.g., selenate, Se(VI)), tetravalent (e.g., selenite, Se(IV)), nonvalent (e.g., elemental, Se⁰), and negative divalent (e.g., selenide, Se(–II)). In well-ventilated alkaline soil, Se mainly exists in soil solution in the form of Se(VI) or combined with soil particles by electrostatic adsorption (Wang et al., 2012). Selenate is often reduced to selenite in acidic and reducing soils; thus, Se mainly exists in such soil solutions in the form of Se(IV), adsorbed on the surface of soil particles or combined with solid components of the soil (e.g., iron (Fe) and manganese (Mn) oxides, hydrated oxides, and organic matter). Elemental Se (Se⁰) is often enveloped by organic matter or exists in soil in the form of a residue. Inorganic-bound Se(–II) mainly exists as a precipitate, while organic-bound Se(–II) is often in the soil solution, and leaves soil as a volatile component (e.g., dimethyl Se), or exists in organic matter combined with soil solids (Zhang and Moore, 1996).

Distribution of selenium fractions in soil

Soil Se can be divided into different forms according to the degree of binding of Se with soil components determined by different extraction and classification procedures. A widely used method for grading soil Se forms is five-step sequential extraction (Wang et al., 2012). In this method, soil Se can be divided according to the difficulty of extraction into water-soluble (SOL-), exchangeable and carbonate-bound (EXC-), iron-manganese oxide-bound (FMO-), organic-bound (OM-), and residual (RES-) Se. Organic-bound Se can be transformed into inorganic Se, which can be directly absorbed and utilized by crops through mineralization and it plays an important role in the geochemical cycle of Se. Therefore, organic-bound Se can be further divided into humic acid-bound Se (HA-Se) and fulvic acid-bound Se (FA-Se) according to its acid solubility; FA-Se, in turn, can be further subdivided into hydrophilic Se (Hy-Se) and hydrophobic Se (HON-Se).

Different types of soils have different compositions and properties, and the proportion of Se in each form is different. The distribution of Se fractions in natural Se-rich soil is different from that in other soils. In natural Se-rich soil, the stable organic-bound and residual fractions are the main fractions of Se in soil (~80–90% of the total Se in soil), while the water-soluble and exchangeable Se fraction (i.e., bioavailable Se) account for less than 5% of total Se (Wang et al., 2012). Although the sum of residual and organic-bound Se fractions account for over 80% of the total soil Se in low-Se soil, the proportion of bioavailable Se in the total Se varies from only 1.1% to 20% depending on the soil properties and environmental conditions (Zhang and Moore, 1996).

Transformation processes of selenium in soil

With changes of environmental conditions, Se in soil undergoes a series of transformations such as adsorption-desorption, dissolution-precipitation, oxidation-reduction, complexation, and volatilization reactions in terms of its form and valence state (Fig. 2). Therefore, understanding the chemical transformation of Se in soil is of great importance in evaluating the supply capacity or bioavailability of Se in this matrix.

Adsorption-desorption reaction

Selenate and selenite are the main forms of Se absorbed by crops. After entering the soil, exogenous selenate (Se(VI)) forms an outer-sphere complex with the soil through electrostatic interaction and physical combination. Selenite (Se(IV)) diffuses into the inner lattice of soil mainly through micropores and forms a relatively tight inner-sphere complex with iron and manganese oxides or organic matter in soil. Because the adsorption strength of the inner sphere is stronger than that of the outer sphere, the mobility and bioavailability of selenite in soil are significantly less than those of selenate (Sharma et al., 2015). Moreover, competitive anions co-existing in the soil solution, such as PO_4^{3-} , SO_4^{2-} , and nitrate (NO_3^-), have a remarkable impact on the adsorption of Se (especially Se(IV)) in soil; PO_4^{3-} (with low affinity), for example, has been observed to have a greater effect on the adsorption of Se than SO_4^{2-} (Goh and Lim, 2004). Desorption also plays an important role in the fractional transformation of Se in soil. The rate of desorption of Se in soil is negatively correlated with the content of metal(hydroxide) oxides, but positively correlated with soil pH; indicating that the desorption of Se in soil is also affected by soil properties.

Dissolution-precipitation reaction

The dissolution-precipitation reaction directly affects the fraction of Se in the soil and the difficulty of crop utilization. Soil Se can be transformed into residual fraction and difficult for plants to uptake and utilize. At low pH, Se(IV) can be reduced to Se^0 or Se—S precipitates in sulfide solution; such precipitation is more significant at lower pH than that at higher pH. However, the precipitation of Se^0 or metal selenides (e.g., FeSe and FeSe_2) also limits the dissolution of Se in soil under reducing conditions. The microbial degradation of organic matter (Eswayah et al., 2016) and reductive dissolution of soil minerals (e.g., ferric hydroxide)

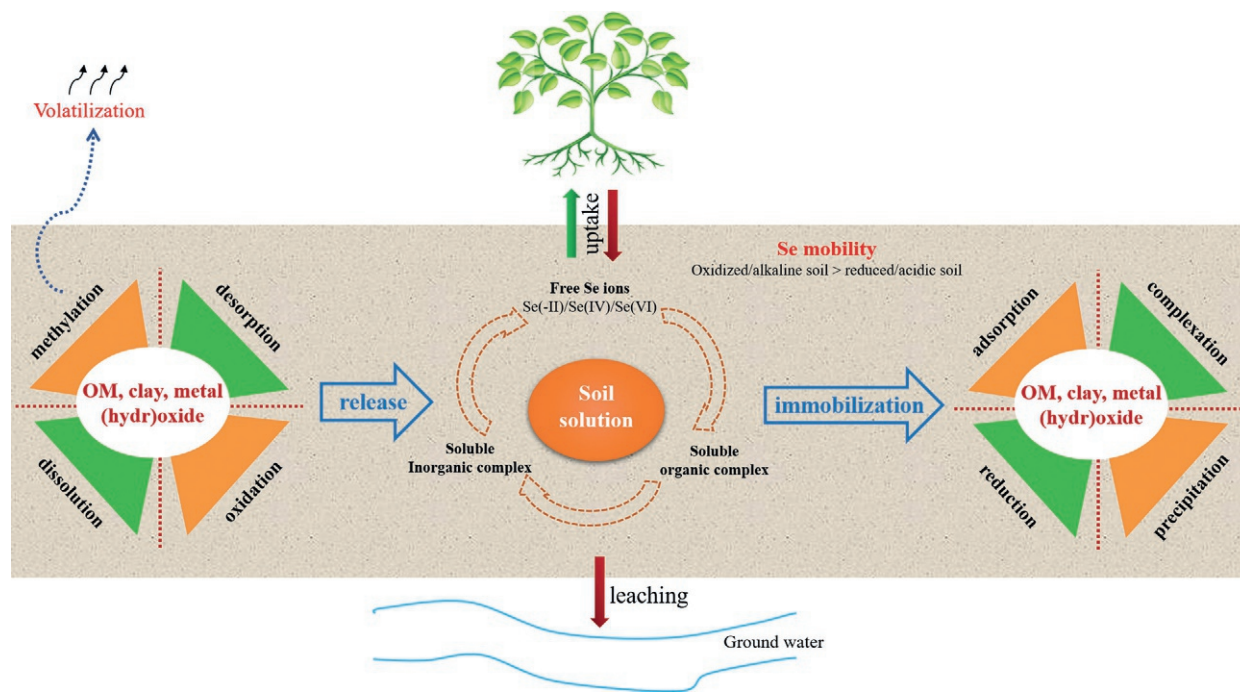


Fig. 2 Chemical behavior of Se in soil (Dinh et al., 2019).

(Shaheen et al., 2017) lead to the re-release of bound Se into the soil solution, thereby improving its bioavailability. Small-molecule organic acids produced by returning organic materials to the field or secretion of plant roots can also dissolve the Se fixed by soil and prevent it from being adsorbed and coprecipitated by iron/aluminum oxide/hydroxides (Dinh et al., 2017).

Oxidation-reduction reaction

Under anaerobic conditions, Se(VI) can be gradually reduced to Se⁰ by the action of some bacteria, and further reduced to Se(–II) by the assimilation of microorganisms (Eswayah et al., 2016). Moreover, Se(VI) can be reduced to Se⁰ by the dissimilation of chemical/biochemical reducing agents. The existence of SO₄^{2–} and iron oxide contributes to the reduction of Se(VI) (Olegario et al., 2010), but NO₃[–] has an inhibitory effect on this process (Bassil et al., 2018). Although the oxidation of Se is mainly mediated by microorganisms, it can also be promoted by abiotic factors (Bassil et al., 2018). The Se⁰ and Se(IV) forms can be oxidized by soil microorganisms under aerobic conditions, and the abiotic oxidation of Se takes place in the presence of peroxodisulfate ions and synthetic manganese oxide and birnessite (Scott and Morgan, 1996). Nevertheless, the rate of biological oxidation of Se(IV) is generally lower than the biological reduction of Se(VI) (Bassil et al., 2018), which means Se tends to be reduced rather than oxidized.

Complexation reaction

Selenium can be complexed with organic matter and iron and/or manganese oxides in soil, and the complexation strength of soil for Se(IV) is greater than that for Se(VI). Therefore, selenite is more easily fixed by soil components than selenate (Dinh et al., 2017). Soil OM increases sorption sites, which directly facilitate complexation with Se (Dinh et al., 2017). Soil microorganisms can promote the fixation of Se(IV) in soil and promote reduction of the fixed Se(IV). In addition, because the functional groups of organic compounds (e.g., hydroxyl, carboxyl, and carbonyl) and Se oxyanions are negatively charged, soil Se and organic matter can be fixed by forming a ternary complex (e.g., Se–HM–OM) through positively charged metal bridge bonds (Dinh et al., 2017).

Volatilization reaction

Methylation is the key process of the volatilization and removal of Se in soil. Two mechanisms of soil Se methylation have been identified, namely, chemical and biological; of these the biological process is dominant. The existing forms of Se, microbial activity, organic matter content, temperature, and water content in the soil affect the methylation process of soil Se. In addition to bacteria, microalgae, and fungi, the biomethylation of soil Se can also be mediated by plants (Terry et al., 2000). The demethylation process of Se is opposite to its methylation process, and can occur under both aerobic and anaerobic conditions. Under anaerobic conditions, dimethyl selenide is rapidly demethylated by microorganisms to produce H₂Se (Eswayah et al., 2016). By comparison, Se(VI) is the final product of demethylation under aerobic conditions (Bolan et al., 2014).

Bioavailability of selenium in soil

The concept of ‘bioavailability’ was first proposed in the context of the water environment and is used to describe the degree of utilization of pollutants during bio-transmission or bio-reaction. Bioavailable Se refers to Se in soil that can be directly absorbed and utilized by plants, which is the main factor that determines the Se content of the plants (Terry et al., 2000). The total amount of Se does not indicate and evaluate the bioavailability of Se in soil, but its form, especially the content of bioavailable Se, is the key to determining its mobility and toxicity.

Selenium in soil often exists in different chemical forms and valence states, and its solubility, mobility, and bioavailability in different forms and valences are quite different. Elemental Se and Se(–II), for example, are insoluble in water and difficult to absorb and utilize by plants; by contrast, Se(IV) and Se(VI) are water-soluble and are the main forms of Se absorbed by crops (Dinh et al., 2018). Moreover, because Se(IV) is easily adsorbed by soil components (e.g., clay minerals and organic matter), the mobility and bioavailability of Se(VI) in soil are higher than those of Se(IV) (Wang et al., 2012).

Among the five fractions of soil Se, SOL-Se has the strongest mobility and is most easily absorbed by plants. The EXC-Se refers to Se(IV) combined with hydrated oxides or adsorbed on the surfaces of clay minerals and humus. The bioavailability of EXC-Se is less than that of SOL-Se, but the former can be converted into the latter and absorbed by plants. Therefore, the sum of SOL-Se and EXC-Se content is often regarded as the plant-available Se content in soil. The OM-Se in soil mainly includes FA-Se and HA-Se. Whereas HA-Se is stable in structure and difficult to decompose and utilize, FA-Se has a low molecular weight and can easily be mineralized into inorganic Se and small-molecule organic Se, which can be absorbed and utilized by crops (Dinh et al., 2017). In addition to the form and valence of soil Se, soil physicochemical properties (e.g., pH, redox potential (Eh), texture, organic matter, co-existing ions), agronomic measures, and climatic conditions also indirectly affect the bioavailability of Se in soil. These should all be considered in evaluating the bioavailability of Se in soil either in phytoremediation or biofortification.

Interaction between sulfur and selenium in soil

In anaerobic conditions, S usually co-exists with Se; in natural aerobic environments, low-valence S is usually oxidized into SO₃^{2–} and SO₄^{2–} and then enters the ocean through the action of surface runoff in the water system. Selenium can easily be oxidized to

Se⁰ and Se(IV) but not Se(VI). Therefore, Se exist mostly in stable fractions in soil. Furthermore, the production of volatile Se compounds is also affected by S, and different forms of S exert different influences on this process. For example, large concentrations of SO₃²⁻ or SO₄²⁻ can almost completely inhibit the formation of dimethyl selenide (DMSe) from selenite, while sulfide can inhibit DMSe formation to a small extent only.

The amount and strength of S and Se adsorption on soil components are different. Iron or aluminum oxides in soil are the main adsorbents of SO₄²⁻, and the adsorbed SO₄²⁻ can easily be replaced by other anions. By comparison, the adsorption of Se by soil depends on its oxidation state but it can also easily be achieved by iron or aluminum oxide. The order of adsorption capacity of iron oxide for some inorganic anions is as follows: orthophosphate (H₂PO₄⁻) > arsenate (H₂AsO₄⁻) = SeO₃²⁻ > silicate (H₄SiO₄) > molybdate (MoO₄²⁻) > SO₄²⁻ > SeO₄²⁻ > chloride (Cl⁻) = NO₃⁻ (Ryden et al., 1987). This order indicates that the affinity of adsorption sites on iron oxide for SO₄²⁻ is between those of SeO₃²⁻ and SeO₄²⁻. Moreover, Se adsorbed in soil can be exchanged by SO₄²⁻, which can significantly weaken the adsorption of Se by iron oxides. Consequently, there must be competition for adsorption sites between SO₄²⁻ and SeO₃²⁻ or SeO₄²⁻, thus affecting the adsorption of selenite and selenate on the soil solid surface and their bioavailability. However, the interactions between S and Se in soil are related to the soil types, plant species, S and Se concentrations, and so on. A systematic understanding of the mechanism of competition between S and Se in soil is currently lacking.

Conclusion

Sulfur and Se undergo a series of chemical processes, such as oxidation-reduction, precipitation-dissolution, and adsorption-desorption, in soil, which leads to continuous changes in their speciation and valence state. These variations are the main challenges in research on the interactions of these two elements. The competition mechanism of S and Se in soil remains unclear, and most previous studies focused only on the interaction between SO₄²⁻ and SeO₄²⁻ during plant uptake. Moreover, the influence of interactions between S and Se on transformation of their speciation in soil has rarely been reported. Scanning electron microscopy, polymerase chain reaction-denaturing gradient gel electrophoresis, 16SrDNA, synchrotron radiation, and other new research methods or technical means should be used to explore the enzymology, microbiology, and soil chemistry of the interactions between S and Se in future research. Such efforts will be of great significance in advancing the understanding of the interactions between S and Se in soil systems and guiding the rational application of inorganic fertilizers and biofortification of Se in crops.

References

- Alloway BJ (2013) *Bioavailability of Elements in Soil. Essentials of Medical Geology*. Dordrecht: Springer, pp. 351–373.
- Bassil J, Naveau A, Bueno M, et al. (2018) Leaching behavior of selenium from the karst infillings of the hydrogeological experimental site of poitiers. *Chemical Geology* 483: 141–150.
- B'Hymer C and Caruso JA (2006) Selenium speciation analysis using inductively coupled plasma-mass spectrometry. *Journal of Chromatography A* 1114(1): 1–20.
- Blair GJ, Lefroy RDB, Chaitop W, et al. (1991) Matching sulfur fertilizers to plant production systems. In: *International Symposium on the Role of Sulphur*.
- Bolan N, Kunhikrishnan A, Thangarajan R, et al. (2014) Remediation of heavy metal(loid)s contaminated soils-to mobilize or to immobilize? *Journal of Hazardous Materials* 266: 141–166.
- Dail DB and Fitzgerald JWS (1999) Cycling in soil and stream sediment: Influence of season and in situ concentrations of carbon, nitrogen and sulfur. *Soil Biology and Biochemistry* 31(10): 1395–1404.
- Dhillon KS and Dhillon SK (2003) *Distribution and Management of Seleniferous Soils, Advances in Agronomy*. Academic Press, pp. 119–184.
- Dinh QT, Li Z, Tran TAT, et al. (2017) Role of organic acids on the bioavailability of selenium in soil: A review. *Chemosphere* 184: 618–635.
- Dinh QT, Cui ZW, Huang J, et al. (2018) Selenium distribution in the Chinese environment and its relationship with human health: A review. *Environment International* 112: 294–309.
- Dinh QT, Wang M, Tran TAT, et al. (2019) Bioavailability of selenium in soil-plant system and a regulatory approach. *Critical Reviews in Environmental Science and Technology* 49(6): 443–517.
- Eswayah AS, Smith TJ, and Gardiner PHE (2016) Microbial transformations of selenium species of relevance to bioremediation. *Applied and Environmental Microbiology* 82(16): 4848–4859.
- Fitzgerald JW (1978) Naturally occurring organosulfur compounds in soil. *Sulfur in the Environment* 2: 391–443.
- Gerla PJ, Sharif MU, and Korom SF (2011) Geochemical process controlling the spatial distribution of selenium in soil and water, west central South Dakota. *Environment and Earth Science* 62: 1551–1560.
- Goh KH and Lim TT (2004) Geochemistry of inorganic arsenic and selenium in a tropical soil: Effect of reaction time, Ph, and competitive anions on arsenic and selenium adsorption. *Chemosphere* 55: 849–859.
- Johnson CM and Nishita H (1952) Microestimation of sulfur in plant materials, soils, and irrigation waters. *Analytical Chemistry* 24(4): 1525.
- Liu CQ, Cao SQ, Chen GA, et al. (1990) Sulphur in the agriculture of China. *Acta Pedologica Sinica* 27(4): 398–404. (In Chinese).
- Lucheta AR and Lambais MR (2012) Sulfur in agriculture. *Revista Brasileira de Ciência do Solo* 36(5): 1369–1379.
- Olegario JT, Yee N, Miller M, et al. (2010) Reduction of Se(VI) to Se(–II) by zerovalent iron nanoparticle suspensions. *Journal of Nanoparticle Research* 12(6): 2057–2068.
- Ryden GJ, Syers JK, and Tillman RW (1987) Inorganic anion sorption and interactions with phosphate sorption by hydrous ferric oxide gel. *Journal of Soil Science* 38(2): 211–217.
- Saggar S, Bettany JR, and Stewart JWB (1981) Sulphur transformations in relation to carbon and nitrogen in incubated soils. *Soil Biology and Biochemistry* 13: 499–511.
- Scott MJ and Morgan JJ (1996) Reactions at oxide surfaces. 2. Oxidation of Se(IV) by synthetic birnessite. *Environmental Science and Technology* 30(6): 1990–1996.
- Shaheen SM, Frohne T, White JR, et al. (2017) Redox-induced mobilization of copper, selenium, and zinc in deltaic soils originating from Mississippi (U.S.A.) and Nile (Egypt) River Deltas: A better understanding of biogeochemical processes for safe environmental management. *Journal of Environmental Management* 186: 131–140.
- Sharma VK, McDonald TJ, Sohn M, et al. (2015) Biogeochemistry of selenium. A review. *Environmental Chemistry Letters* 13(1): 49–58.
- Singh BR (1984) Sulfate sorption by acid forest soils. *Soil Science* 138(3): 189–197.
- Tan JA (1989) *The Atlas of Endemic Diseases and Their Environments in the People's Republic of China*. China: Science Press.

- Terry N, Zayed AM, de Souza MP, et al. (2000) Selenium in higher plants. *Annual Review of Plant Physiology and Plant Molecular Biology* 51: 401–432.
- Wang SS, Liang DL, Wang D, et al. (2012) Selenium fractionation and speciation in agriculture soils and accumulation in corn (*Zea mays* L.) under field conditions in Shaanxi Province, China. *Science of the Total Environment* 427: 159–164.
- Wilhelm SH (2009) Sulfur in soils. *Journal of Plant Nutrition and Soil Science* 172(3): 326–335.
- Zhang Y and Moore JN (1996) Selenium fractionation and speciation in a wetland system. *Environmental Science and Technology* 30: 2613–2619.
- Zoller WH, Gladney ES, and Duce RA (1974) Atmospheric concentrations and sources of trace metals apt the South Pole. *Science* 183: 198–200.

Arsenic in soil

Md Imran Ullah Sarkar^{a,b,c}, Afsana Jahan^c, Md Tofail Hossain^{a,b,d}, Mohammad Mahmudur Rahman^{a,b}, and Ravi Naidu^{a,b}, ^aCentre for Environmental Remediation (GCER), College of Engineering, Science and Environment, The University of Newcastle, Callaghan, NSW, Australia; ^bCRC for Contamination Assessment and Remediation of the Environment (crcCARE), The University of Newcastle, Callaghan, NSW, Australia; ^cSoil Science Division, Bangladesh Rice Research Institute, Gazipur, Bangladesh; ^dDepartment of Agricultural Botany, Sher-e-Bangla Agricultural University, Dhaka, Bangladesh

© 2023 Elsevier Ltd. All rights reserved.

Introduction	202
Occurrence, distribution, and geochemistry of arsenic	202
Arsenic in agricultural soil	203
Transport of As from soil to rice grain	204
Factors affecting As transformation and transport in the soil-plant system	204
Soil pH and redox potential	204
Soil organic matter	204
Microorganisms	204
Clay mineralogy	205
Interaction with soil nutrients	205
Phosphorus	205
Silicon	205
Nitrate	205
Sulfur	205
Arsenic phytotoxicity in rice plants	206
Arsenic in rice grains	206
Arsenic exposure and human health risk	206
Remediation strategies for As	207
Water management	207
Organic amendments	207
Biochar	207
Other organic amendments	207
Inorganic amendments	208
Nitrogen fertilizer	208
Phosphorus fertilizer	208
Sulfur fertilizer	208
Iron supplementation	208
Manganese supplementation	209
Silicon supplementation	209
Seed priming	209
Phytoremediation	210
Bioremediation	210
Cultivars that exclude As or breeding of low As accumulating cultivars	210
Conclusion and future research	211
References	211

Abstract

Arsenic (As) is a toxic metalloid for all living organisms and is a global concern due to its adverse effect on human health. Humans are mainly exposed to As through the consumption of As-contaminated drinking water and food crops. Apart from drinking water, As-contaminated food crops have become a major route of As entrance into the human body due to As contamination of the agricultural soil from both natural and anthropogenic sources. The transformative behavior of As in the soil, especially in paddy soil, is a complex process that needs better understanding for efficient removal of As from the contaminated paddy soils. This chapter provides current knowledge on As behavior in farmland soil, and special emphasis has been given to different remediation approaches for mitigating As accumulation in rice.

Key points

- Some farmland soils in different parts of the world exceeded the European Union recommended level (20 mg kg^{-1}) of As in agricultural soils.
- Soil pH, redox potential, organic matter and nutrient elements like nitrogen, phosphorus, silicon and sulfur affect As transformation and translocation to plants.
- Among the cereal crops rice is more susceptible to As contamination because of increased As mobilization in the submerged paddy soil.
- Alternate wetting and drying irrigation, Fe- or Mn-modified biochar and silicon fertilizer are some efficient mitigation strategies of As contamination in farmland soil.

Introduction

Arsenic (As) is poisonous and carcinogenic to living organisms yet is ubiquitous in the geosphere. Humans are potentially exposed to As mainly through drinking water and food crops contaminated by both natural and anthropogenic activities. Increased contamination of As in groundwater has been a worldwide concern in the last few years with more than 70 countries reporting As contamination. This means an As-induced potential health risk for an estimated 150 million people globally (Brammer and Ravenscroft, 2009). The situation is particularly alarming in South-East Asian countries including Bangladesh, in areas of India, Nepal, Myanmar, Pakistan, Vietnam, Laos, Cambodia, Indonesia and China (Rahman et al., 2009). Apart from increased concentrations in drinking water, As contamination has become a global human health issue due to elevated levels in the agro-ecosystems and subsequent accumulation in food crops. For example, As-polluted irrigation water has resulted in extensive contamination in paddy soils of many Asian countries. The anaerobic environments in flooded paddy soils promote As mobilization, causing increased bioavailability to rice (*Oryza sativa* L.) plants compared to other cereal crops. About half of the world's population consume rice as their staple food making contaminated rice grains a serious health hazard globally for humans. Increased As in paddy soil is not only a threat to food safety but also induces phytotoxicity in rice plants and considerable loss in grain yield, threatening food security. Thus, remediation of As polluted soil has been a global research focus over the last few decades.

The transformative behavior of As in soil involves complex processes regulated by several soil properties such as soil pH, redox potential (Eh), soil texture, cation exchange capacity (CEC), organic matter (OM), oxidation state of iron (Fe) and microbial activity (Nath et al., 2014; Chen et al., 2016). Precise understanding of the soil factors that determine bioavailability of As in farmland soil is crucial for developing successful mitigation strategies. Among the farmland soils, paddy soil needs special attention as rice is grown throughout the world and feeds more than half of the world population. Moreover, the submerged condition which exists only in paddy soil favors As mobilization and increased accumulation in rice grains. This chapter addresses current knowledge on the remediation of As from farmland soil, and in particular paddy soil. In addition, As biogeochemical cycling, the transport mechanism in rice and factors associated with As mobilization in paddy soil are also discussed to provide a better understanding of As remediation in paddy soil.

Occurrence, distribution, and geochemistry of arsenic

Arsenic is a toxic metalloid which belongs to the group VA in the periodic table and exists naturally in the environment. Arsenic ranks 20th among the elements in terms of abundance and its concentration varies between 1.5 and 5 mg kg^{-1} in the Earth's crust (Rahaman et al., 2021). Although, this is a small amount and it is rarely in elemental form, it is a constituent of more than 200 minerals (Upadhyaya and Roychoudhury, 2022) and exists as inorganic forms. Sulfur (S), iron and phosphate minerals are major sources of As in the form of arsenates (60%), sulfides and sulfosalts (20%), and other forms such as arseniurates, arsenites, arsenides, silicates and oxides (20%) (Mandal and Suzuki, 2002; Upadhyaya and Roychoudhury, 2022). The major As bearing primary minerals are arsenopyrite (FeAsS), realgar (As_4S_4), orpiment (As_2S_3), cobaltite (CoAsS), niccolite (NiAs), arsenolite (As_2O_3), tennantite ($(\text{Cu,Fe})_{12}\text{As}_4\text{S}_{13}$), haematolite ($(\text{Mn,Mg})_4\text{Al}(\text{AsO}_4)(\text{OH})_8$), and scorodite ($\text{FeAsO}_4 \cdot 2\text{H}_2\text{O}$). In the Earth's crust, arsenopyrite is the dominant form of mineral As under elevated temperature conditions (Rahaman et al., 2021). Igneous, metamorphic, and sedimentary rocks are another natural source of As. Metamorphic ($\sim 5 \text{ mg kg}^{-1}$) and sedimentary rocks ($5\text{--}10 \text{ mg kg}^{-1}$) have larger concentrations of As than igneous rocks (1.5 mg kg^{-1}).

In the environment, As contamination primarily occurs from the three sources: hydrothermal activity, ore deposits and Cenozoic sediments (Masuda, 2018). In some countries of Latin America, United States and New Zealand, the major source of As is hydrothermal activity like volcanic eruptions and hot springs. Ore deposits are the second major source of As contamination and mainly result from mining activities, leading to As contamination in both the hydrosphere and pedosphere. This has caused As contamination in many parts of South Africa, Zimbabwe, Thailand and China. The third major cause of As contamination is Cenozoic sediments that cause As contamination in the aquifer. This is the origin of As contamination in many South Asian countries including Bangladesh, India and Pakistan (Masuda, 2018). In addition, other natural sources that contribute to As

contamination include forest fires, wind-blown dust, and sea salt. Other than these natural sources, human activities also cause As contamination in the environment. For example, use of As bearing pesticides in the twentieth century caused severe As contamination in many countries of the world. Although, the use of these pesticides has decreased, especially in the developing countries, but alloys containing As are still being used in semiconductors (Masuda, 2018). Other anthropogenic sources include mining activities, metallurgy, combustion of fossil fuels, urban waste, and cattle farming. It is estimated that the global annual contribution of anthropogenic sources to As contamination is about 75–100 metric tonnes.

The mobilization processes involved in natural As pollution of the soil and water are reductive dissolution, alkali desorption, sulfide oxidation and geothermal activity (Ravenscroft, 2007). Reductive dissolution is the most common mode which occurs when As is adsorbed to iron oxyhydroxides in sediments. Reductive dissolution mainly occurs in humid areas, while alkali desorption occurs in arid regions (Ravenscroft, 2007). In arid soils the presence of small clay, organic matter and free Fe contents, and large negative surface charges are attributed to the desorption of arsenic, mainly As(V) (Kandakji et al., 2015; Rahman et al., 2019).

In organic and inorganic compounds As mainly exists in four oxidation states: arsenate (+5), arsenite (+3), arsenic (0) and arsenide (−3). Arsenic in −3 (e.g. arsines and metal arsines) and 0 oxidation states are unstable in the soil. Arsenate is the dominant form under aerobic condition which exists as arsenic acid (H_3AsO_4) and occurs at pH 2–3. With an increase in pH (3–11), H_3AsO_4 dissociates to $\text{H}_2\text{AsO}_4^{4-}$ (dihydrogen arsenate) and HAsO_4^{2-} (hydrogen arsenate). In anaerobic conditions arsenite is the predominant form of As found as H_3AsO_3 , arsenous acid. Arsenic in soil can undergo a methylation process resulting in organic As species mainly monomethylarsonic acid (MMA) and dimethylarsinic acid (DMA).

Arsenic in agricultural soil

In the pedosphere, As concentration ranges between 10 and 40 mg kg^{-1} from uncontaminated natural sources with an average concentration 5–8 mg kg^{-1} (Mandal and Suzuki, 2002). The concentration of As in soil varies with the soil horizon, soil type and lithology. In rare cases, some soils have very large concentrations of As (0.1–2%) from either geological or anthropogenic sources (Ongley et al., 2007). These soils may have detrimental effects on human health and the wider ecosystem. Geographically there is a wide variation in the magnitude of As contamination in agricultural soils. Table 1 represents a global scenario of As contamination in agricultural soils. In farmland soil As contamination comes from both natural and anthropogenic sources. Weathered rocks and alluvial sediments are major natural sources of As contamination. Other than natural sources, use of As contaminated irrigation water largely contributes to the increased As contamination in paddy soil. Anthropogenic activities such as disposal of As-containing effluents, tailings, deposition of metallurgical fumes and metal mining have resulted in As accumulation in the soil and subsequent transport to rice grains in several areas of the world.

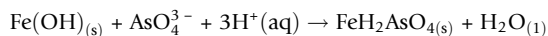
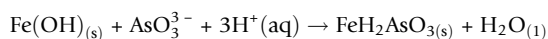
Table 1 Arsenic concentrations in agricultural soils across the world.

Country	Region	As (mg kg^{-1})	References
Bangladesh		46	Chowdhury et al. (2001)
		2.06	Islam et al. (2012)
Cambodia	Kandal and Prey Veng Province	0.8–18	Seyfferth et al. (2014)
China	Zhejiang Province	20.87	Tang et al. (2020)
China	Jiangsu Province	47	Zhu et al. (2019)
China	Zhejiang Province,	105.1	Zhai et al. (2020)
China	Hunan province	288.6	Irshad et al. (2020)
India	Uttar Pradesh	5.4–14.10	Srivastava and Sharma (2013)
India	Central India (Chhattisgarh)	9–390	Patel et al. (2005)
Pakistan	Sindh	46.2	Samrana et al. (2017)
Vietnam	Backan province	3.8–39.3	Bui et al. (2016)
Chile	Valparaíso	11–30	Mondaca et al. (2022)
Ecuador	Azuay Province	24–422	Romero-Crespo et al. (2023)
	Napo province	43	Jimenez et al. (2023)
Mexico	Zacatecas	39–165	Salas-Muñoz et al. (2022)
England	Cornwall, Devon	6–546	Williams et al. (2007)
France	Camargue	5–10	Williams et al. (2007)
Scotland	East coast	6–10	Williams et al. (2007)
Spain	Doñana	4–11	Williams et al. (2007)
	Cádiz	1–2	Williams et al. (2007)
United States	Louisiana	2.8–73	Ori et al. (1993)
	Texas,	1.76	Ma et al. (2020)
	Northeastern	0.91–6.34	Hu et al. (2021)
	California	2–4	Williams et al. (2007)
	Arkansas	4–7	Williams et al. (2007)

In paddy soil, As occurs mostly in two inorganic forms: As(V) and As(III). Under flooded conditions As(III) is the dominant form. Small amounts of organic species like MMA, DMA and trimethylarsine (TMA) also occur in paddy soil due to microbial biomethylation of inorganic As. The behavior and fate of As in soil solutions are mainly regulated by its speciation. Inorganic As species are more toxic than organic species; in particular the toxicity of As(III) is greater than As(V). The latter is attributed to the reaction of As(III) with sulfhydryl (–SH) groups of proteins and enzymes which causes a disturbance in cell activities and eventually cell mortality. It is possible that As(III) has a greater mobility than As(V) which causes decreased adsorption on soil particles.

Transport of As from soil to rice grain

Paddy soil remains submerged most of the time during the rice growing period and this environment favors the mobilization of As. In paddy soil under submerged conditions, As(V) is also reduced to As(III) which is more toxic. Under submerged conditions, more than 95% of the As in paddy soil is As(III). The increased bioavailability of As in flooded conditions is attributed to the reductive dissolution of iron oxyhydroxides. Moreover, continuously submerged conditions promotes the abundance of the Fe-reducing bacteria population which further enhances the reductive dissolution of iron hydroxides and related As reduction. The As mobilization from Fe-oxide surfaces can be explained by three mechanisms: (a) site transfer of adsorbed As between liquid and solid phase and subsequent desorption, (b) reduction of Fe(II) to Fe(III), and (c) reduction of As(V) to As(III) from the Fe-oxide-As complex. The crystal structure of As-oxide can also affect As mobilization in the soil solution, where poorly crystallized Fe-oxide restricts As mobilization. Arsenic mobilization forming Fe-oxides in the soil solution can be explained by the following equations.



In paddy soil, As is also transformed into organic forms mainly MMA and DMA through a microbially-induced methylation process. The bioavailability of As species to plants is in the order As(III) > MMA > As(v) > DMA, while the order of toxicity is As(III) > As(V) > MMA > DMA. In plants, As(V) is transferred through phosphate transport channels, whereas As(III) and organic As species are transported by aquaporin channels. The toxicity, solubility and bioavailability of As is largely governed by various soil properties such as OM, pH, mineral oxides, clay content and CEC. Arsenic in the rice plant accumulates in the roots and is then distributed to the different plant parts in the order of root > leaves > shoots > husk > grain.

Factors affecting As transformation and transport in the soil-plant system

Soil pH and redox potential

Soil pH and redox potential largely regulate the mobilization of different forms of As. In soil with high pH, the availability of As is increased by the desorption of As(III) from the negatively charged soil minerals. In acidic soil, As bound with Fe and Al hydroxides is released into the soil solution. Moreover, at low pH, phosphorus (P) and As(V) both compete for the same sorption sites, increasing the mobility of As in the soil matrix.

The submerged conditions of paddy soil decrease redox potential (Eh) to –200 mV, because of depleted oxygen. This soil condition favors the microbially-mediated transformation of Fe oxyhydroxides, which is the key sink of As, and releases the sorbed As into the soil solution. Under negative Eh in flooded conditions, As(V) is reduced to As(III) and the ratio of As(III)/As(V) is high.

Soil organic matter

Soil organic matter (OM) can affect As mobilization both positively and negatively. In the presence of OM, As forms a complex with different organic compounds, reducing the solubility of As and therefore uptake by plants (Kumarathilaka et al., 2018b). On the other hand, negatively charged dissolved organic carbon (DOC) present in OM competes with As for sorption on the surface of Fe-oxyhydroxides, which increases As in the soil solution and plant uptake. Increased DOC also promotes Fe(III)-reducing bacterial activity, which enhances the release of As in soil pore water through the reduction of Fe-oxyhydroxides (Chen et al., 2016). Increased concentration of soluble P in the presence of OM is also responsible for increased As mobilization in the soil.

Microorganisms

Soil microbes are involved in the transformation of As by regulating the oxidation and reduction reactions in soil and methylation of As. The soil microbes involved in this transformation are resistant to As at large concentrations. The microbially-facilitated oxidation of As(III) to As(V) is a detoxification process as the pentavalent form is less toxic than the trivalent form. During the microbial respiration process As(III) acts as an electron donor and is oxidized to As(V). The microbial reduction of NO_3^- or SO_4^{2-} in anerobic conditions also causes the oxidation of As(III). Alternatively, microbial reduction of As(V) occurs when As(V) is used as an electron acceptor which releases As(III) in the soil matrix. Apart from oxidation-reduction, As methylation is caused by different microorganisms including bacteria, cyanobacteria and algae (Kuehnelt and Goessler, 2003). Microbes also affect the mobilization

of As in the root zone by manipulating the oxidation-reduction of Fe. The microbially-mediated reduction of Fe oxides (ferric) releases the adsorbed As(V) in the soil solution and increases its availability. The released As(V) is reduced to As(III), which is subjected to methylation by an abundance of *arsC* (arsenate reductase) and *arsM* (arsenite methyltransferase) genes in the flooded paddy soil. Alternatively, Fe-oxidizing microbes cause the oxidation of ferrous Fe in the presence of NO_3^- which promotes the precipitation or adsorption of As with ferric minerals in the soil and reduces As availability in the rice rhizosphere, and therefore the uptake by rice. This means the microbial transformation of As could be manipulated for the bioremediation of As in paddy soil.

Clay mineralogy

The low redox potential in paddy soil increases the flocculation and dispersion of clay particles which leads to their downward movement. The deposition of clay particles below the plow zone, and other mechanical activities during the rice growing period, leads to the formation of hard plow pans. Clay minerals are negatively charged fine particles with a large surface area, where As bearing Fe-oxyhydroxides precipitate. The precipitation of Fe-oxyhydroxides reduces As availability in soil solutions and consequently its uptake by the rice plant. The sorption and speciation of As in paddy soil can vary with the crystal structure and type of clay minerals. Deposition of different secondary minerals on the surface of clay minerals can alter their physico-chemical properties, which subsequently affect their ability to adsorb As (Uddin, 2017). However, the long-term interaction of As with various clay minerals such as kaolinite, montmorillonite, and illite are not fully understood and requires further study.

Interaction with soil nutrients

Phosphorus

Phosphorus (P) has a strong relationship with soil As availability, uptake and translocation in rice plants. Arsenate is analogous to phosphate in chemical form, and both adsorb onto the surface of Fe-oxide through a ligand exchange mechanism. The As(V) is absorbed by the rice plant through the same pathway as phosphate. Therefore, the competition for sorption sites between these two ions can result in two different effects. Phosphorus can increase or decrease As availability in soil and subsequently increase or decrease its uptake by plants. Deng et al. (2020) reported different effects from phosphate on As transformation in paddy soil under suboxic and anoxic conditions. Under suboxic conditions, phosphate replaces As(V) from the adsorption sites and increases the concentration of As(V) in the soil solution. In anoxic conditions, phosphate promotes the reduction of Fe and As through increased microbial activity, increasing As(III) concentration in the soil solution at both acidic and neutral or alkaline pH. However, some studies have reported that phosphate fertilizer can suppress As uptake by rice grains when the available P concentration is large in the soil solution and by forming amorphous Fe minerals where As(V) coprecipitates. The interaction of phosphate and As is governed by soil factors such as soil Eh, pH, P availability, organic ligands and Fe minerals (Deng et al., 2020). Therefore, further study is required for a clear and quantitative understanding of the effects of phosphate on As mobility in the soil and its transfer to rice grains in a complex soil system.

Silicon

Silicon (Si) is naturally present in the paddy soil. It is also applied to reduce the effect of different abiotic stresses, such as drought, on the rice plant. In an aqueous solution of paddy soil, Si is released as silicic acid (H_4SiO_4) which has a strong affinity to Fe oxyhydroxides. Silicic acid (H_4SiO_4 , $\text{pK}_a = 9.8$) and arsenite (H_3AsO_3 , $\text{pK}_a = 9.2$) are translocated into the rice plant by the same transporters (Ma et al., 2008). Thus, H_4SiO_4 competes with As (III) which reduces the As(III) uptake by the rice plant. The transporters regulate As translocation within the rice tissue parts only and have no ability to regulate root As absorption, whereas Si in the soil solution affects As(III) uptake by the root (Zhao et al., 2010). Both H_4SiO_4 and As(III) compete for the same adsorption sites which releases As(III) into the soil solution and subsequently results in the increased uptake of As(III) by rice plants (Kumarathilaka et al., 2018a). Silicic acid may also increase the organic As species (DMA) in the soil solution which is less toxic through the release of sorbed DMA and increase in As(III) methylation (Limmer et al., 2018). This warrants further investigation to understand the probable mechanisms of As immobilization when Si is applied to paddy soil.

Nitrate

Nitrate (NO_3^-) plays an important role in As speciation in the soil solution by regulating the reduction of Fe. Nitrate promotes the oxidation of Fe^{2+} and inhibits the reduction of Fe^{3+} in the soil solution. This process promotes As sequestration by binding As in the secondary minerals of Fe and reduces the As availability in paddy soil. In anaerobic conditions nitrate creates the oxidation of As(III) to As(V) followed by adsorption of As(V) on the surface of Fe-oxyhydroxides.

Sulfur

Sulfur is an essential nutrient element for plants that can immobilize As in the soil. In soil solutions, S is present as SO_4^{2-} which is transformed to sulfide (S^{2-}) with the decrease in Eh under anoxic conditions. The sulfide reacts with As(III) and precipitates as As_2S_3 (arsenic trisulfide) or AsS (arsenic monosulfide). Sulfates also compete with As(V) in metabolism and transport, which reduces the translocation of As in cell membranes. In the rhizosphere, As binds with organosulfur compounds through a chelation process that immobilizes As in the soil solution. There is evidence of S-induced Fe plaque formation on the root surface which restricts As uptake by the rice plant. Thus, S fertilizer in paddy soil can play an important role in As immobilization and uptake by rice plants.

Arsenic phytotoxicity in rice plants

Arsenic is not only toxic to human health, but also has an adverse effect on crop plants. Under As toxicity, plants suffer from oxidative stress due to excessive production and accumulation of reactive oxygen species (ROS) and hydrogen peroxide (H_2O_2). It also inhibits adenosine triphosphate (ATP) production and carbohydrate assimilation in rice grains which result in poor grain yield. Arsenic toxicity may cause reduced seed germination, stunted plant growth, brown spot and scorched leaves (Moulick et al., 2018).

Arsenic in rice grains

Arsenic content of rice grains varies largely with rice cultivars and locations as rice is cultivated in about 100 countries with more than 120,000 cultivars. Table 2 shows a global scenario of As concentration in rice grains where it varies from as low as 0.03 mg kg^{-1} to as high as 0.77 mg kg^{-1} . The concentration of As in rice grains differs widely with geographic location, cultivars and growing conditions. Zavala and Duxbury (2008) reported that in Bangladesh, increased As concentration in rice grains resulted from As-contaminated irrigation water, rather than soil As, whereas in the southern United States it was caused by the excessive use of As-containing agrochemicals and subsequent accumulation in the soil. Williams et al. (2005) conducted a global level investigation on As speciation and reported that inorganic As was the dominant fraction in rice grains of Bangladesh ($64 \pm 1\%$), India ($80 \pm 3\%$) and Europe ($81 \pm 4\%$), whereas in the United States, the dominant fraction was organic As species (DMA) and inorganic fractions were only $42 \pm 5\%$ of the total. The color and size of the rice grains also contribute to the variation in As concentration. For example, one study reported higher As concentration in the extra-long slender Indian basmati rice than in short-bold brown rice. In rice grains most As accumulates in the external layer of the grain called bran which is removed in white rice during polishing yet stays intact in brown rice. Consequently, As concentration in brown rice is higher than in white rice. One study reported that 10% polishing of the rice grain can decrease the total As by 61–66% and inorganic As by 51–70%. However, polishing can affect other essential nutrient content in rice grains. Arsenic translocation in rice grains is also affected by the growing season. Dry season rice contains more As than that grown in the wet monsoon season. This seasonal variability is attributed to the use of As-polluted groundwater for irrigation in the dry season, which increases As availability, followed by increased accumulation in the rice grain. Together with environmental factors, As content of rice grains is also governed by genetic variation. One study reported that environment (E), genotype (G) and $G \times E$ interaction caused 69–80, 9–10 and 10–21% variation, respectively in rice grain As.

Arsenic exposure and human health risk

Arsenic contamination in groundwater and agricultural soil has been a global concern due to its subsequent translocation in food grains, especially rice. In many parts of the world, including Bangladesh and India, the groundwater As concentration has exceeded the maximum tolerable limit of As (0.01 mg kg^{-1}) in drinking water (Rahman et al., 2009). In China the background value of soil As (11.2 mg kg^{-1}) exceeded the global average As concentration by (5–10 mg kg^{-1}). The increased contamination of As in soil and groundwater has resulted in increased contamination in food crops, especially rice. Dietary intake of rice has been a major pathway of As exposure in many countries where rice is the main diet. A study from China reported that 2.5% of 160 polished rice samples surpassed the upper limit of inorganic As (0.1 mg kg^{-1}) recommended for infants and children by the European Union.

Table 2 Global scenario of As concentration in rice grain.

Country	Sampling source	Sampling number	Total As (mg kg^{-1})	References
Bangladesh	Market	144	0.17	Sarkar et al. (2023)
Australia	Market	21	0.27	Rahman et al. (2014)
Brazil	Market (White rice)	17	0.11	Batista et al. (2010)
Brazil	Market (White parboiled rice)	6	0.09	Batista et al. (2010)
Jiangsu, China	Field	70	0.13	Li et al. (2018)
China	Field	446	0.12	Yao et al. (2020)
West Bengal, India	Household	56	0.34	Halder et al. (2020)
Jamaica	Market	16	0.20	Antoine et al. (2012)
Nigeria	Market	8	0.10	Mwale et al. (2018)
Nigeria	Field	300	0.03	Mandal et al. (2022)
Punjab, Pakistan	Field	204	0.44	Natasha et al. (2022)
Portugal and Spain	Market (white rice)	56	0.17	Pinto et al. (2016)
Portugal and Spain	Market (parboiled rice)	13	0.16	Pinto et al. (2016)
Sri Lanka	Market	11	0.14	Mwale et al. (2018)
South Korea	Contaminated sites	40	0.25	Kwon et al. (2017)

Arsenic contamination in rice grains poses a potential health risk for humans through dietary intake. The International Agency for Research on Cancer (IARC) has categorized arsenic as a group I carcinogen. Arsenic-induced skin, lung, bladder, kidney, liver, and prostate cancers in humans have been widely reported. Many cell and organ activities in the human body are disrupted by As toxicity. Arsenic is also related to ischemic heart disease in humans.

Remediation strategies for As

Water management

Arsenic contaminated irrigation water has caused As accumulation in paddy soils of many countries, especially in Southeast Asia. The water requirement of rice crops is much greater than other grain crops, and rice is normally cultivated under flooded conditions. However, rice crops do not require continuous standing water throughout their life cycle. Therefore, alternate wetting and drying (AWD) is an efficient irrigation practice that not only saves water but also reduces As mobilization in paddy soil and uptake by the rice. One study found that AWD reduced As availability by 86%, 55% and 41% in pore water, root plaque and rice grains, respectively. The AWD technique promotes soil aeration which favors the oxidation of As(III) to As(V), the less toxic form (Takahashi et al., 2004). Increased aeration also increases Fe plaque formation in the root zone of rice which reduces As uptake by the crop.

Organic amendments

Biochar

Biochar is the stable, carbon rich material produced from a wide range of biomass created by pyrolysis, which has several agronomic, environmental, and economic advantages as a soil amendment. A further benefit to soil health is biochar's capacity for metal adsorption due to its large surface area, high porosity, and diverse functional groups. However, the efficacy of biochar on metal remediation varies widely due to the variation in feedstock and pyrolysis process, biochar properties, soil type, microbial activity and rate of application. Recently, biochar has been modified by various chemical and physical treatments to improve the immobilization efficiency of target metals. Biochar performs better for the immobilization and adsorption of cationic metals such as cadmium (Cd), lead (Pb), zinc (Zn), and chromium (Cr), while this effect is not as prominent on anions like phosphate and As. In some cases, biochar even increases the mobilization of anionic metals. Biochar application causes several chemical changes in the soil, for example an increase in soil pH, DOC, pore water phosphate, microbial activity and the reduction of Fe and Mn, which favors the mobilization of soil As. Biochar rich in P, increases the soil solution phosphate which desorbs As(V) from the solid phase into the soil solution from competition between phosphate and As(V) for adsorption sites. Biochar releases DOC which also contributes to As mobilization through reductive dissolution of Fe oxyhydroxides and subsequent desorption of the As bound to them. Biochar induced As mobilization is attributed to the increased phosphate and DOC concentration in the soil solution and increased reduction of Fe due to increase in Fe-reducing bacteria population. However, some studies have found that biochar can substantially immobilize soil As by forming Fe—Mn plaque on the rice root. Moreover, the As immobilization efficiency of biochar can be increased by modifying the biochar. Surface modification of biochar with Fe- and manganese-based materials appears efficient for As remediation in paddy soils with their greater affinity for As (Zhang et al., 2022). Among Fe-based materials, iron oxide and nano zero valent iron (nZVI) appeared as the most promising, compared to other iron-based materials like FeCl₃, ferric chloride (Wu et al., 2018). Application of biochar, modified with different iron-based materials such as Fe-oxyhydroxy sulfate, FeCl₃ and ZVI, reduced the NaHCO₃-extractable As in the soil solution by 14–30, 101–28, and 17–35%, respectively (Wu et al., 2018). Biochar modified with manganese oxide is also an efficient As immobilization agent in soil by increasing the hydrogen-carbon and oxygen-carbon ratios, hydrophilicity and adsorption capacity of biochar (Yu et al., 2015). One study reported that the application of 2% manganese oxide (MnO) modified biochar significantly reduced As in different parts of the rice plant through oxidation of As(III) to As(V) and subsequent adsorption of As(V) with MnO. The efficacy of biochar on soil As immobilization is also largely affected by the biomass used for biochar production, biochar properties, application rate and soil physicochemical characteristics. Therefore, the interaction of As species with biochar produced from different biomass, needs careful investigation to identify the best rate for field application.

Other organic amendments

Inorganic fertilizers and different organic manures are often applied to improve rice yield and soil health. Organic amendments, especially those rich in cellulose, can immobilize certain metals because of their various functional groups that act as adsorption sites for them. One study revealed that application of cow dung and sugarcane (*Saccharum officinarum* L.) bagasse significantly reduced As contents in rice roots, shoots, grain and husk. In a field study, application of vermicompost with AWD irrigation reduced As accumulation in rice grains by 66% compared to no amendments under flooded conditions. Alternatively, several studies have indicated that organic amendments increased As translocation in rice grains by increasing As mobilization in the soil solution, which was attributed to decreased Eh and increased DOC (Syu et al., 2019). Mobilization of As with organic amendments also varies with the source of amendments. For example, in one example the solubilization of As was greater with livestock droppings than plant-based compost and cow dung. Addition of organic amendments also promotes methylation in paddy soil which acts as an As detoxification mechanism. For example, in a microcosm study soil treated with 2% rice straw showed increased translocation

of methylated As in rice tissue by increasing the microbially-mediated methylation of soil As. Another study also reported increased organic As content in rice grains with a single application of 80 ton ha⁻¹ of olive (*Olea europaea* L.) mill waste compost under field conditions. The use of organic amendments, however, is often hindered by its availability, the large amount required and associated extra cost. The effect of organic amendments on As immobilization is a complex process that requires more study.

Inorganic amendments

Nitrogen fertilizer

Nitrogen (N) is one of the most limiting essential plant nutrients for plant growth and yield and is required in large amounts. The effects of N fertilizer on As mobilization depends on the form of N applied. When N is applied as NO₃⁻ it immobilizes As by inhibiting Fe reduction and increasing Fe oxidation (Liu et al., 2022; Chen et al., 2023). Under anaerobic conditions application of NO₃⁻ fertilizers immobilizes As by inhibiting the reduction of Fe minerals and oxidizing the more mobile and toxic As(III) into the less mobile and toxic As(V) in pore water. However, in wetland soil NO₃⁻ reduction occurs spontaneously due to microbial activity and NO₃⁻ is depleted rapidly. Liu et al. (2022) revealed that NO₃⁻ application reduces the pore As concentration in the early stages (1–3 days) with no major change at the later stages of the incubation period (14 days). Therefore, the benefits of NO₃⁻ on As immobilization in wetland soil particularly in paddy soil is short term and might not be feasible for remediation of paddy soil. In flooded soil, when ammoniacal N fertilizer is applied, the NH₃ is subjected to anaerobic oxidation which occurs simultaneously with Fe₃⁺ reduction. This process is termed Feammox and it releases As adsorbed with Fe oxides. In a field study, Chen et al. (2023) reported that NH₃—N application from urea and ammonium bicarbonate promoted the reduction of As(V) to As(III) in paddy soil and subsequently increased the As concentration in rice grains. The NH₃—N decreased soil pH and Eh which supports Fe³⁺ reduction and increases As mobility. However, in an incubation study, Liu et al. (2022) showed that the application of NH₄Cl decreased As availability in paddy soil with long-term effects. This was explained by the decrease in soil pH which overcomes the Feammox effects on As mobilization. Application of NH₄Cl decreases soil pH with the release of H⁺ through hydrolysis and subsequently increases the positive charge on the surface of Fe minerals which adsorbs As(V). The effect of N fertilizer application on soil As mobilization and speciation is complex, and is affected by the source of N and soil properties such as pH, Eh, organic carbon (OC) and Fe-oxide content. More studies are needed under field conditions in a complex soil system to obtain better insight on the effect of N fertilizer application on mitigating As availability in paddy soil.

Phosphorus fertilizer

Phosphorus is an essential plant nutrient for plant growth and yield. Depending on soil properties and degree of As contamination, phosphate fertilizer application can control As uptake by the rice plant. Because As(V) and PO₄⁻ share the same translocation pathway, the application of P fertilizer can reduce As uptake by the rice plant. Efficacy of P fertilizer on reducing As mobilization in soil and subsequent uptake by the rice plant is governed by three major factors: (i) the competitive interaction between PO₄⁻ and As for adsorption sites in the soil matrix, (ii) the uptake competition between available P and As by the rice root and (iii) the translocation of P from root to shoot (Lee et al., 2016). Research has shown that increased P in the soil solution from external P application reduced As(V) uptake by the rice plant. Alternatively, P application can significantly increase As uptake by the rice plant by increasing As concentration in the soil solution from competition between As(V) and PO₄⁻ for the adsorption sites. Moreover, in flooded rice cultivation systems where the dominant species of As is As(III), P application may show no significant effect on reducing As uptake because PO₄⁻ is not analog of As(III). However, there is a possibility of As(III) oxidation to As(V) even under flooded conditions because of the oxidized rice root zone. Thus, the effect of P amendment on As transformation in flooded paddy soil cannot be disregarded and requires more study. Research is also needed to observe the effect of P fertilizer application on As transformation and uptake by the rice plant under AWD water management, which promotes aeration in the soil.

Sulfur fertilizer

Application of sulfur fertilizer can reduce As mobilization in soil and uptake by the rice plant by changing the chemistry of the rhizosphere soil and inducing As precipitation. In anaerobic conditions, with low Eh and microbial activity, Fe³⁺ and SO₄²⁻ are reduced into Fe²⁺ and S²⁻, respectively, which interact to form nanoparticles of FeS that act as binding sites for As in the soil solution. Moreover, sulfide can immobilize As(III) as As(III)—S phases e.g. As₂S₃. Alternatively, microbes can oxidize S to SO₄²⁻ which prevents the reduction of Mn oxide and Fe hydroxide, ultimately reducing the dissolution of As bound with these compounds (Herath et al., 2018). Sulfur-containing thiol compounds like glutathione and phytochelatin can also immobilize As in the soil solution, forming stable complexes and subsequently reducing As uptake by the rice plant. Some studies have shown that application of S-containing gypsum fertilizer has increased the immobilization of As in paddy soil. The potential of S fertilizer to mitigate As in paddy soil and uptake by rice grains requires long-term field studies for better understanding and large scale application.

Iron supplementation

Application of additional Fe to paddy soil can enhance As immobilization and decrease As uptake by rice plants. Research suggests that Fe supplements can promote As immobilization by stimulating Fe-plaque formation in the rhizosphere. Various iron-based amendments such as Fe sulfate, zero valent Fe, red mud and steel shot can immobilize soil As and reduce As translocation in rice grains. One study reported that application of poorly crystalline iron oxides, FeCl₂·NaNO₃ and FeCl₂ at grain filling stage, decreased As content of rice grains by 54%, 52% and 46%, respectively, in a control sample. The efficacy of Fe amendments on As

immobilization is regulated by the oxidation-reduction reaction in the soil solution. The prolonged anaerobic conditions in paddy soil results in the reduction and dissolution of Fe(III) which eventually releases As(III) into the soil solution. Excessive Fe shows a toxic effect on plant growth, particularly in wetland rice soil. In flooded paddy soil, iron toxicity results from excessive concentration of reduced iron (Fe^{2+}) in the soil solution and its translocation to rice leaves. The excess iron (Fe^{2+}) in rice tissue causes cell damage and physiological disorder due to the production of excessive oxygen radicals. Therefore, the existing amount of Fe in soil should be considered before any additional Fe application.

Manganese supplementation

Manganese (Mn)-based amendments, particularly manganese oxides (MnO_2), can immobilize As by transforming As forms in the soil and successively reducing plant uptake. With their strong reactivity, poor crystal structure and large surface area, manganese oxides are strong oxidants which catalyze the oxidation of As(III) to As(V). The oxidized As(V) becomes immobilized by adsorbing onto the surface of Mn oxides. However, the oxidation of As(III) by manganese oxides is largely influenced by their structure, surface charge properties, mineral crystallinity, abundance and soil pH. The best immobilization effect of manganese oxides has been observed at pH 7–8. At low pH, especially at 4, the effect decreased with the dissolution of manganese oxides. Some synthetic Mn oxides-based amendments such as hausmannite and α - MnO_2 nanorods can significantly reduce As concentration in pore water, and subsequently reduce the accumulation of As in rice plant parts through enhanced oxidation of As(III) to As(V).

Silicon supplementation

Silicon (Si) is second among the elements in the Earth's crust and plays a beneficial role in plant growth, especially when plants suffer from different biotic and abiotic stresses. Research suggests that Si-based amendments reduced As uptake, mainly As(III), by rice plants grown in As-polluted soil. The decreased concentration of As(III) in rice tissues was due to the increased Si concentration downregulating the As(III) transporters, which are the same as salicylic acid transporters (Ma et al., 2008). Silicon also immobilizes As by forming iron plaque in the root zone and changing the Fe- and As-transforming microbial community (Limmer et al., 2018). The effectiveness of Si-based amendments on As varies with the source of Si, application dose, method, time and water management. There are several available inorganic and organic sources of Si such as salicylic acid, silicates of Na, K, Ca and Mg, wollastonite, quartz sand, Si- rich crop residues (e.g. rice and wheat (*Triticum aestivum* L.) straw) and slag-based silicate fertilizers. In a study, the effect of various Si-based organic and inorganic amendments (rice husk, charred rice husk and silicate fertilizers) was examined which showed no significant effect of the amendments on total As concentration in rice grains. They did, however, significantly reduce the inorganic As concentration by increasing organic As concentration, which is less toxic, and charred rice husk performed better than the other amendments. In another study, application of silica gel increased the concentration of organic As (DMA) in rice grains, which was attributed to the increased concentration of DMA in the soil solution from desorption by salicylic acid. The solubility of Si in the amendments also regulates the amelioration effect, because highly soluble Si sources, e.g. silica gel and rice husk decrease As concentration in rice tissues, and less soluble sources like diatomaceous clay increase As concentration. A study on the effect of different Si doses and water management on As accumulation in rice plants showed that a 3 mM Si supplement significantly reduced As uptake by rice plants in soil treated with 300 μM As in pot cultures and the effect was better under aerobic than anaerobic conditions. A field experiment reported that top dressing with Si fertilizer (K_2SiO_3) at booting stage of the rice crop, significantly reduced inorganic As contents in brown rice by increasing Fe-plaque formation in the root zone and retaining As on the Fe plaque. However, some studies showed that Si amendments increased pore water As(III) concentration because of competition for absorption sites in soil minerals with salicylic acid and subsequently increased the grain As concentration. In that case, foliar application of Si could be an alternate method for reducing As translocation in rice. A study on the effect of foliar Si application at different growth stages revealed that the application of Si at tillering reduced inorganic As by 61.4% in the polished rice. The main mechanism of silicon amendments on decreasing As(III) concentration in rice tissues are: i. increased Si concentration in the rice tissue from competitive uptake between Si and As(III), ii. reduced expression of the salicylic acid and As(III) transporters (Lsi1 and Lsi2), and increased methylation of As(III) to organic As. Moreover, in situ immobilization of Si in the rice rhizosphere with formation of amorphous crystalline Fe oxides in the soil and root Fe plaques that act as adsorption sites for As. Silicon fertilization also increases the more As-resistant and As(III)-oxidizing bacteria populations in the rhizosphere. The effects of Si-based amendments on reducing As immobilization in soil solution and uptake by rice involves complex mechanisms that are influenced by several factors with both positive and negative findings. Therefore, more research is required for better understanding of the mechanisms involved before large scale application in the real field conditions.

Seed priming

Plants suffering from severe As stress have reduced plant growth and yield because of changes in physiological processes. Some studies showed that seed priming is effective in reducing As stress on the vegetative growth of rice. One study investigated the effect of potassium humate as a priming agent on rice seed germination and seedling growth in a 50 μM solution of As(III) and As(V) for 12 days and found improved seed germination and seedling growth. These improvements were attributed to increased chlorophyll content, relative water content, antioxidant defense mechanism and nutrient translocation in the rice seedlings. Another study reported that rice seed priming with moringa leaf extract resulted in better germination and improved plant growth, attributed to the increased relative water content in the rice seedlings and reduced electrolyte leakage with a reduction in malondialdehyde content. Seed priming with selenium showed promising effects on rice seed germination, seedling growth, nutrient uptake, physiological

activity and reducing As accumulation in rice tissue parts (Moulick et al., 2016; Moulick et al., 2018). The beneficial effects of selenium for seed priming were possibly from the reduced translocation of As from soil-root-shoot, decreased oxidative stress through increased antioxidant activities and increased uptake of micronutrients. Study results suggest seed priming is promising for alleviating As stress on rice plant growth, particularly at an early stage, and reducing As translocation in rice grains. However, there is limited research on this issue and most of the results are based on the early growth stage of rice in confined conditions which may not reflect a total growth stage study under field conditions. More research is vital to exploit the benefits of seed priming in mitigating As stress in rice plants for practical application.

Phytoremediation

Some aquatic plants (e.g. *Ipomoea aquatic* Forssk, *Marsilea minuta* L. 1771, *Phragmites australis* (Cav.) Trin. ex Steud., *Pteris vittata* L., *Pteris ryukyuensis* Tagawa, *Pteridium Aquilinum* (L.) Kuhn, *Vetiveria zizanioides* (L.) Roberty) have the hyper accumulation ability of metals that can be used for As phytoremediation. Ma et al. (2001) first reported *P. vittata* can translocate As to its aboveground biomass and therefore can be used for soil As removal. A study on the efficacy of *P. vittata* on mitigating As uptake by rice in five different soils showed that rice cultivation with *P. vittata* removed 4–11% total soil As and reduced the phosphate-extractable As by 11–38%, soil pore water As by 18–77%, total As in straw by 17–82%, total As in grain by 22–58% and grain inorganic As by 50–58%. Another study found that growing rice with *M. minuta*, a water fern, significantly reduced As accumulation in rice roots and subsequent accumulation in the above ground biomass. Intercropping of aquatic vegetables (*Ipomoea aquatica* Forssk, *Oenanthe javanica* (Blume) DC, *Sagittaria sagittifolia* L.) with rice showed reduced As concentration, a bioaccumulation factor and translocation factor in brown rice (Huang et al., 2021). However, phytoremediation also affects rice plant growth and yield because of competition between the rice crop and accumulator plant for nutrients, light and space. The disposal of the phytoremediator after harvesting is also an issue of concern. The effectiveness of the phytoremediation technique varies with the growth of hyperaccumulating plants, As accumulation potential, tolerance to As toxicity, concentration and bioavailability of the contaminants to be remediated. The As accumulation potential and tolerance of the accumulator plant to As can be increased by several amendments like plant growth promoting chemicals and fertilizers, biochar and certain microbes (Franchi et al., 2017). Phytoremediation could be an eco-friendly technique for As remediation compared to chemical amendments, although there are limitations and drawbacks that require more research to exploit the benefit of phytoremediation in mitigating As in rice grains.

Bioremediation

Various microorganisms including bacteria, fungi and algae in the rhizosphere are resistant to As stress through different mechanisms like oxidation/complexation, solubilization, biosorption, methylation, sequestration etc. which can immobilize As in the soil solution. Successful inoculation of these microbes into the rhizosphere could alleviate As stress in contaminated paddy soil. The microbial immobilization and decontamination of As is achieved by: (i). oxidation and methylation of toxic As forms to less toxic forms, (ii). accumulation of As species at the extracellular surface of the microorganisms, (iii). formation of amorphous $\text{Fe}(\text{OH})_2$ on the cell surface and (iv). inhibition of the mRNA expression of As(III) transport pathways. Several As-resistant bacterial strains such as *Kocuria flava* AB402, *Bacillus vietnamensis* AB403, *Staphylococcus arlettae* (NBRIEAG-6), *Staphylococcus* sp. (NBRIEAG-6), and *Brevibacillus* sp. (NBRIEAG-6) have been identified with capacity to reduce As accumulation in rice crop. Some studies also reported arbuscular mycorrhizal fungi can decrease As translocation in rice tissues by suppressing the expression of As (III) transporters and ameliorate As stress to the rice plant by increasing antioxidant enzyme activity. The efficiency of bioremediation varies with the abundance of the microbial population, which is governed by several factors such as OC, pH, redox potential, soil nutrients, etc. (Jia et al., 2015). One study reported that among different soil factors, soil pH is the most limiting factor for microbial population abundance related to As transformation. The application of NO_3^- in flooded paddy soils substantially increased the As(III) oxidizing bacteria population and consequently reduced the bioavailable As (Zhang et al., 2017). Inoculation of Fe(II)/Mn(II) oxidizing bacteria immobilized As by promoting Fe-plaque formation in the root surface. In another study it was reported that inoculation of SO_4^{2-} -reducing bacteria and SO_4^{2-} fertilizer reduced the total As content in rice roots by about 23%. This was attributed to the immobilization of As(III) as As_2S_3 and FeS complex (Jia et al., 2015). Application of microbial inoculants could be an effective and environmentally friendly technique for mitigating As stress in rice. However, successful application of this practice may be limited by how the inoculated microbes will adapt to natural field conditions and compete with the indigenous microbial population.

Cultivars that exclude As or breeding of low As accumulating cultivars

Varietal differences in crop plants are associated with differences in translocation and uptake of As. For example, some rice cultivars have reduced As uptake because of their greater root porosity, radial oxygen loss and iron-plaque formation on the root surface (Suriyagoda et al., 2018b). Rice plants can also detoxify As through glutathione- or glutaredoxin-induced reduction of As(V) to As(III), the latter is transported out of the cell or sequestered into the vacuole (Suriyagoda et al., 2018a). Cultivation of rice varieties that accumulate small amounts of As could be an effective As mitigation option in contaminated paddy soils. A study showed that the total and inorganic As contents were larger in improved high yielding varieties than in local varieties. The varietal resistance of

rice plants to As is largely influenced by management practices and growing environment (Suriyagoda et al., 2018a). Therefore, remediation of As through varietal selection is a complex process and the performance is likely to vary in relation to growing conditions, soil types and cultural practices across the rice growing areas.

Conclusion and future research

Arsenic in agricultural soil is a great threat for food safety and human health. Thus, innovative mitigation strategies to diminish As accumulation in rice are urgently required to decrease these risks. Such solutions to reduce the As content in crop plant particularly in rice can be achieved by: exploiting genetic variations or the identification of existing rice cultivars that accumulate less inorganic As, different water management regimes and or the use of mineral nutrients as soil amendments to offset As uptake in grains. Limited resources have been directed towards mitigation strategies that minimize As accumulation in paddy rice using single techniques alone. Therefore, more research is needed on combined techniques, including organic and inorganic amendments, to immobilize As in paddy soils and decrease As in rice grains.

References

- Antoine JMR, Hoo Fung LA, Grant CN, Dennis HT, and Lalor GC (2012) Dietary intake of minerals and trace elements in rice on the Jamaican market. *Journal of Food Composition and Analysis* 26: 111–121.
- Batista B, De Oliveira Souza V, Da Silva F, and Barbosa JF (2010) Survey of 13 trace elements of toxic and nutritional significance in rice from Brazil and exposure assessment. *Food Additives and Contaminants* 3: 253–262.
- Brammer H and Ravenscroft P (2009) Arsenic in groundwater: A threat to sustainable agriculture in South and South-east Asia. *Environment International* 35: 647–654.
- Bui AT, Nguyen HT, Nguyen MN, Tran T-HT, Vu TV, Nguyen CH, and Reynolds HL (2016) Accumulation and potential health risks of cadmium, lead and arsenic in vegetables grown near mining sites in Northern Vietnam. *Environmental Monitoring and Assessment* 188: 1–11.
- Chen Z, Wang Y, Xia D, Jiang X, Fu D, Shen L, Wang H, and Li QB (2016) Enhanced bioreduction of iron and arsenic in sediment by biochar amendment influencing microbial community composition and dissolved organic matter content and composition. *Journal of Hazardous Materials* 311: 20–29.
- Chen G, Du Y, Fang L, Wang X, Liu C, Yu H, Feng M, Chen X, and Li F (2023) Distinct arsenic uptake feature in rice reveals the importance of N fertilization strategies. *Science of the Total Environment* 854: 158801.
- Chowdhury U, Rahman M, Mondal B, Paul K, Lodh D, Biswas B, Basu G, Chanda C, Saha K, and Mukherjee S (2001) Groundwater arsenic concentration and human suffering in West Bengal, India and Bangladesh. *Environmental Sciences* 8: 393–415.
- Deng Y, Weng L, Li Y, Chen Y, and Ma J (2020) Redox-dependent effects of phosphate on arsenic speciation in paddy soils. *Environmental Pollution* 264: 114783.
- Franchi E, Rolli E, Marasco R, Agazzi G, Borin S, Cosmina P, Pedron F, Rosellini I, Barbaferi M, and Petruzzelli G (2017) Phytoremediation of a multi contaminated soil: Mercury and arsenic phytoextraction assisted by mobilizing agent and plant growth promoting bacteria. *Journal of Soils and Sediments* 17: 1224–1236.
- Halder D, Saha JK, and Biswas A (2020) Accumulation of essential and non-essential trace elements in rice grain: Possible health impacts on rice consumers in West Bengal, India. *Science of the Total Environment* 706: 135944.
- Herath I, Vithanage M, Seneweera S, and Bundschuh J (2018) Thiolated arsenic in natural systems: What is current, what is new and what needs to be known. *Environment International* 115: 370–386.
- Hu R, Teasley WA, and Seyfferth AL (2021) Paired soil and rice arsenic and cadmium from northeastern US rice farms. *Agricultural & Environmental Letters* 6: e20040.
- Huang S, Zhuo C, Du X, and Li H (2021) Remediation of arsenic-contaminated paddy soil by intercropping aquatic vegetables and rice. *International Journal of Phytoremediation* 23: 1021–1029.
- Irshad MK, Noman A, Alhaithloul HA, Adeel M, Rui Y, Shah T, Zhu S, and Shang J (2020) Goethite-modified biochar ameliorates the growth of rice (*Oryza sativa* L.) plants by suppressing Cd and As-induced oxidative stress in Cd and As co-contaminated paddy soil. *Science of the Total Environment* 717: 137086.
- Islam M, Brammer H, Mustafizur Rahman G, Raab A, Jahiruddin M, Solaiman A, Meharg AA, and Norton GJ (2012) Arsenic in rice grown in low-arsenic environments in Bangladesh. *Water Quality Exposure and Health* 4: 197–208.
- Jia Y, Bao P, and Zhu Y-G (2015) Arsenic bioavailability to rice plant in paddy soil: Influence of microbial sulfate reduction. *Journal of Soils and Sediments* 15: 1960–1967.
- Jimenez PAJ, Díaz X, Silva MLN, Vega A, and Curi N (2023) Assessing and Understanding Arsenic Contamination in Agricultural Soils and Lake Sediments from Papallacta Rural Parish, Northeastern Ecuador, via ecotoxicology factors, for environmental embasement. *Sustainability* 15: 3951.
- Kandakji T, Udeigwe TK, Athanasiou D, and Pappas S (2015) Chemistry of arsenic in semi-arid alkaline soils of the Southern High Plains, USA: Sorption characteristics and interactions with soil constituents. *Water, Air, & Soil Pollution* 226: 1–11.
- Kuehnelt D and Goessler W (2003) Organoarsenic compounds in the terrestrial environment. In: Craig PJ (ed.) *Organometallic Compounds in the Environment*, 2nd, pp. 223–275. John Wiley & Sons Inc.
- Kumarathilaka P, Seneweera S, Meharg A, and Bundschuh J (2018a) Arsenic accumulation in rice (*Oryza sativa* L.) is influenced by environment and genetic factors. *Science of the Total Environment* 642: 485–496.
- Kumarathilaka P, Seneweera S, Meharg A, and Bundschuh J (2018b) Arsenic speciation dynamics in paddy rice soil-water environment: Sources, physico-chemical, and biological factors-a review. *Water Research* 140: 403–414.
- Kwon JC, Nejad ZD, and Jung MC (2017) Arsenic and heavy metals in paddy soil and polished rice contaminated by mining activities in Korea. *Catena* 148: 92–100.
- Lee C-H, Wu C-H, Syu C-H, Jiang P-Y, Huang C-C, and Lee D-Y (2016) Effects of phosphorous application on arsenic toxicity to and uptake by rice seedlings in As-contaminated paddy soils. *Geoderma* 270: 60–67.
- Li T, Song Y, Yuan X, Li J, Ji J, Fu X, Zhang Q, and Guo S (2018) Incorporating bioaccessibility into human health risk assessment of heavy metals in rice (*Oryza sativa* L.): A probabilistic-based analysis. *Journal of Agricultural and Food Chemistry* 66: 5683–5690.
- Limmer MA, Mann J, Amaral DC, Vargas R, and Seyfferth AL (2018) Silicon-rich amendments in rice paddies: Effects on arsenic uptake and biogeochemistry. *Science of the Total Environment* 624: 1360–1368.
- Liu L, Shen R-L, Zhao Z-Q, Ding L-J, Cui H-L, Li G, Yang Y-P, Duan G-L, and Zhu Y-G (2022) How different nitrogen fertilizers affect arsenic mobility in paddy soil after straw incorporation? *Journal of Hazardous Materials* 436: 129135.
- Ma LQ, Komar KM, Tu C, Zhang W, Cai Y, and Kennelley ED (2001) A fern that hyperaccumulates arsenic. *Nature* 409: 579.
- Ma JF, Yamaji N, Mitani N, Xu X-Y, Su Y-H, McGrath SP, and Zhao F-J (2008) Transporters of arsenite in rice and their role in arsenic accumulation in rice grain. *Proceedings of the National Academy of Sciences* 105: 9931–9935.

- Ma X, Sharifan H, Dou F, and Sun W (2020) Simultaneous reduction of arsenic (As) and cadmium (Cd) accumulation in rice by zinc oxide nanoparticles. *Chemical Engineering Journal* 384: 123802.
- Mandal BK and Suzuki KT (2002) Arsenic round the world: A review. *Talanta* 58: 201–235.
- Mandal J, Bakare WA, Rahman MM, Rahman MA, Siddique AB, Oku E, Wood MD, Hutchinson SM, and Mondal D (2022) Varietal differences influence arsenic and lead contamination of rice grown in mining impacted agricultural fields of Zamfara State, Nigeria. *Chemosphere* 305: 135339.
- Masuda H (2018) Arsenic cycling in the Earth's crust and hydrosphere: Interaction between naturally occurring arsenic and human activities. *Progress in Earth and Planetary Science* 5: 1–11.
- Mondaca P, Valenzuela P, Roldán N, Quiroz W, Valdenegro M, and Celis-Díez JL (2022) Remediation of agricultural soils with long-term contamination of arsenic and copper in two Chilean Mediterranean areas. *Agronomy* 12: 221.
- Moullick D, Ghosh D, and Santra SC (2016) Evaluation of effectiveness of seed priming with selenium in rice during germination under arsenic stress. *Plant Physiology and Biochemistry* 109: 571–578.
- Moullick D, Santra SC, and Ghosh D (2018) Effect of selenium induced seed priming on arsenic accumulation in rice plant and subsequent transmission in human food chain. *Ecotoxicology and Environmental Safety* 152: 67–77.
- Mwale T, Rahman MM, and Mondal D (2018) Risk and benefit of different cooking methods on essential elements and arsenic in rice. *International Journal of Environmental Research and Public Health* 15: 1056.
- Natasha, Bibi I, Niazi NK, Shahid M, Ali F, Masood Ul Hasan I, Rahman MM, Younas F, Hussain MM, Mehmood T, Shaheen SM, Naidu R, and Rinklebe J (2022) Distribution and ecological risk assessment of trace elements in the paddy soil-rice ecosystem of Punjab, Pakistan. *Environmental Pollution* 307: 119492.
- Nath S, Panda P, Mishra S, Dey M, Choudhury S, Sahoo L, and Panda SK (2014) Arsenic stress in rice: Redox consequences and regulation by iron. *Plant Physiology and Biochemistry* 80: 203–210.
- Ongley LK, Sherman L, Armienta A, Concilio A, and Salinas CF (2007) Arsenic in the soils of Zimapán, Mexico. *Environmental Pollution* 145: 793–799.
- Ori L, Amacher M, and Sedberry J Jr. (1993) Survey of the total arsenic content in soils in Louisiana. *Communications in Soil Science and Plant Analysis* 24: 2321–2332.
- Patel K, Shrivastava K, Brandt R, Jakubowski N, Corns W, and Hoffmann P (2005) Arsenic contamination in water, soil, sediment and rice of central India. *Environmental Geochemistry and Health* 27: 131–145.
- Pinto E, Almeida A, and Ferreira IMPLVO (2016) Essential and non-essential/toxic elements in rice available in the Portuguese and Spanish markets. *Journal of Food Composition and Analysis* 48: 81–87.
- Rahaman MS, Rahman MM, Mise N, Sikder MT, Ichihara G, Uddin MK, Kurasaki M, and Ichihara S (2021) Environmental arsenic exposure and its contribution to human diseases, toxicity mechanism and management. *Environmental Pollution* 289: 117940.
- Rahman MM, Naidu R, and Bhattacharya P (2009) Arsenic contamination in groundwater in the Southeast Asia region. *Environmental Geochemistry and Health* 31(supplement 1): 9–21.
- Rahman MA, Rahman MM, Reichman SM, Lim RP, and Naidu R (2014) Arsenic speciation in Australian-grown and imported rice on sale in Australia: Implications for human health risk. *Journal of Agricultural and Food Chemistry* 62: 6016–6024.
- Rahman M, Clark M, Yee L, Comarmond M, Payne T, and Burton E (2019) Effects of pH, competing ions and aging on arsenic (V) sorption and isotopic exchange in contaminated soils. *Applied Geochemistry* 105: 114–124.
- Ravenscroft P (2007) Predicting the global distribution of natural arsenic contamination of groundwater. In: *Symposium on Arsenic: The Geography of a Global Problem*, Royal Geographical Society, London.
- Romero-Crespo P, Jiménez-Oyola S, Salgado-Almeida B, Zambrano-Anchundia J, Goyburo-Chávez C, González-Valoys A, and Higuera P (2023) Trace elements in farmland soils and crops, and probabilistic health risk assessment in areas influenced by mining activity in Ecuador. *Environmental Geochemistry and Health*. <https://doi.org/10.1007/s10653-023-01514-x>.
- Salas-Muñoz S, Valdez-Valdez E, Mauricio-Castillo JA, Salazar-Badillo FB, Vega-Carrillo HR, and Salas-Luevano MA (2022) Accumulation of As and Pb in vegetables grown in agricultural soils polluted by historical mining in Zacatecas, Mexico. *Environmental Earth Sciences* 81: 374.
- Samrana S, Ali I, Azizullah A, Daud M, and Gan Y (2017) Arsenic-based pollution status in Pakistan. *Annals of Agricultural & Crop Sciences* 2: 1027.
- Sarkar MIU, Shahriar S, Naidu R, and Rahman MM (2023) Concentrations of potentially toxic and essential trace elements in marketed rice of Bangladesh: Exposure and health risks. *Journal of Food Composition and Analysis* 117: 105109.
- Seyfferth AL, McCurdy S, Schaefer MV, and Fendorf S (2014) Arsenic concentrations in paddy soil and rice and health implications for major rice-growing regions of Cambodia. *Environmental Science & Technology* 48: 4699–4706.
- Srivastava S and Sharma Y (2013) Arsenic phytotoxicity in black gram (*Vigna mungo* L. Var. PU19) and its possible amelioration by phosphate application. *Journal of Plant Physiology & Pathology* 3: 2.
- Suriyagoda LD, Dittert K, and Lambers H (2018a) Arsenic in Rice soils and potential agronomic mitigation strategies to reduce arsenic bioavailability: A review. *Pedosphere* 28: 363–382.
- Suriyagoda LD, Dittert K, and Lambers H (2018b) Mechanism of arsenic uptake, translocation and plant resistance to accumulate arsenic in rice grains. *Agriculture, Ecosystems & Environment* 253: 23–37.
- Syu C-H, Wu P-R, Lee C-H, Juang K-W, and Lee D-Y (2019) Arsenic phytotoxicity and accumulation in rice seedlings grown in arsenic-contaminated soils as influenced by the characteristics of organic matter amendments and soils. *Journal of Plant Nutrition and Soil Science* 182: 60–71.
- Takahashi Y, Minamikawa R, Hattori KH, Kurishima K, Kihou N, and Yuita K (2004) Arsenic behavior in paddy fields during the cycle of flooded and non-flooded periods. *Environmental Science & Technology* 38: 1038–1044.
- Tang X, Shen H, Chen M, Yang X, Yang D, Wang F, Chen Z, Liu X, Wang H, and Xu J (2020) Achieving the safe use of Cd- and As-contaminated agricultural land with an Fe-based biochar: A field study. *Science of the Total Environment* 706: 135898.
- Uddin MK (2017) A review on the adsorption of heavy metals by clay minerals, with special focus on the past decade. *Chemical Engineering Journal* 308: 438–462.
- Upadhyaya G and Roychoudhury A (2022) Arsenic-toxicity and tolerance: Phytochelatin-mediated detoxification and genetic engineering-based remediation. In: *Global Arsenic Hazard: Ecotoxicology and Remediation*, pp. 481–508. Springer.
- Williams P, Price A, Raab A, Hossain S, Feldmann J, and Meharg AA (2005) Variation in arsenic speciation and concentration in paddy rice related to dietary exposure. *Environmental Science & Technology* 39: 5531–5540.
- Williams PN, Villada A, Deacon C, Raab A, Figuerola J, Green AJ, Feldmann J, and Meharg AA (2007) Greatly enhanced arsenic shoot assimilation in rice leads to elevated grain levels compared to wheat and barley. *Environmental Science & Technology* 41: 6854–6859.
- Wu C, Cui M, Xue S, Li W, Huang L, Jiang X, and Qian Z (2018) Remediation of arsenic-contaminated paddy soil by iron-modified biochar. *Environmental Science and Pollution Research* 25: 20792–20801.
- Yao BM, Chen P, and Sun GX (2020) Distribution of elements and their correlation in bran, polished rice, and whole grain. *Food Science & Nutrition* 8: 982–992.
- Yu Z, Zhou L, Huang Y, Song Z, and Qiu W (2015) Effects of a manganese oxide-modified biochar composite on adsorption of arsenic in red soil. *Journal of Environmental Management* 163: 155–162.
- Zavala YJ and Duxbury JM (2008) Arsenic in rice: I. Estimating normal levels of total arsenic in rice grain. *Environmental Science & Technology* 42: 3856–3860.
- Zhai W, Dai Y, Zhao W, Yuan H, Qiu D, Chen J, Gustave W, Maguffin SC, Chen Z, and Liu X (2020) Simultaneous immobilization of the cadmium, lead and arsenic in paddy soils amended with titanium gypsum. *Environmental Pollution* 258: 113790.

- Zhang J, Zhao S, Xu Y, Zhou W, Huang K, Tang Z, and Zhao F-J (2017) Nitrate stimulates anaerobic microbial arsenite oxidation in paddy soils. *Environmental Science & Technology* 51: 4377–4386.
- Zhang W, Cho Y, Vithanage M, Shaheen SM, Rinklebe J, Alessi DS, Hou C-H, Hashimoto Y, Withana PA, and Ok YS (2022) Arsenic removal from water and soils using pristine and modified biochars. *Biochar* 4: 55.
- Zhao FJ, Ago Y, Mitani N, Li RY, Su YH, Yamaji N, McGrath SP, and Ma JF (2010) The role of the rice aquaporin Lsi1 in arsenite efflux from roots. *New Phytologist* 186: 392–399.
- Zhu N, Qiao J, and Yan T (2019) Arsenic immobilization through regulated ferrolysis in paddy field amendment with bismuth impregnated biochar. *Science of the Total Environment* 648: 993–1001.

Non-transition elements: Metals and metalloids in soils

Natasha^a, Sana Khalid^a, Muhammad Shahid^a, Faiz Rabbani^a, and Camille Dumat^b, ^aDepartment of Environmental Sciences, COMSATS University Islamabad, Islamabad, Pakistan; ^bDYNAFOR, Dynamiques et écologie des paysages agriforestiers, UMR INRAE, INP, Toulouse, France

© 2023 Elsevier Ltd. All rights reserved.

Introduction	214
Arsenic	215
General information	215
Arsenic contamination of soil	215
Arsenic bioavailability and speciation in soil	216
Selenium	216
General information	216
Selenium contamination of soil	218
Selenium bioavailability and speciation in soil	218
Antimony	218
General information	218
Antimony contamination of soil	219
Antimony bioavailability and speciation in soil	219
Lead	219
General information	219
Lead contamination of soil	220
Lead bioavailability and speciation in soil	220
Barium	220
General information	220
Barium contamination of soil	220
Barium bioavailability and speciation in soil	220
Conclusions and perspectives	222
References	222
Further reading	223

Abstract

The non-transition elements are the main group s- and p-block elements. The p-block elements that possess the hybrid characteristics of both metals and nonmetals are called metalloid elements. These elements exist naturally in the Earth's crust at varying concentrations; moreover they are used in many industrial processes and applications. In soil, they play key roles in several biochemical processes, which ultimately affect plant growth, food production and water quality. A review of the available data on the biogeochemical behavior of some non-transition metalloids (arsenic, barium, antimony, lead and selenium) in the soil system illustrates changes concerning their amounts and behavior (speciation, mobility or bioavailability) with implications for environmental health.

Key points

- The five non-transition elements (arsenic, barium, antimony, lead and selenium) have markedly different concentrations in soils, illustrating global scale pollution from anthropogenic activities.
- Soil pH, redox potential, microbial activities, mainly influence the speciation, behavior and phytoavailability of these elements in the soil ecosystem.
- A sustained effort is needed to minimize pollution risks, both by reducing vulnerability, use of (eco)toxic substances and the strengthening of health-environment regulations.

Introduction

The non-transition chemical elements are in the s- and p-block (depending upon the entry of last electrons in the s- or p-orbital) from Mendeleev's classification, including noble gasses. These elements include: (i) alkali metals such as potassium (K) and sodium (Na), (ii) alkaline earth metals such as magnesium (Mg) and strontium (Sr), (iii) chalcogens such as oxygen (O) and sulfur

(S), (iv) halogens such as chlorine (Cl) and bromine (Br), (v) actinoid series such as protactinium (Pa) and thorium (Th), and (vi) lanthanoid series such as praseodymium (Pr) and cerium (Ce), and (vii) metalloids such as arsenic (As) and silicon (Si). Metalloid elements are known to possess the characteristics of both metals and nonmetals; hybrid or intermediate properties. Owing to their hybrid or intermediate characteristics, it is not possible to classify them as either metals or non-metals. There is no commonly accepted definition of metalloid. Notwithstanding, the term 'metalloid' is consistently used in scientific literature.

Silicon (Si), tellurium (Te), boron (B), arsenic (As), germanium (Ge), and antimony (Sb) are currently known as the metalloids. In addition, selenium (Se), aluminum (Al), astatine (At), carbon (C), and polonium (Po) have been less frequently classified as metalloids. These 11 metalloids are present in a diagonal area of the p-block of a standard periodic table. Some elements such as beryllium (Be), bismuth (Bi), gallium (Ga), hydrogen (H) or zinc (Zn), occasionally classify as metalloids. Similarly, the p-block metals, and non-metals (nitrogen (N) and carbon) have also occasionally been referred to as metalloids as these elements have the potential to modify their characteristics by forming alloys with metals.

The characteristics of different metalloids vary from each other as well as from (non)metals. Overall, metalloids possess a metallic appearance. However, these elements are moderately good conductor of electricity and are brittle and shiny. The metalloids primarily refer to nonmetals because their chemical characteristics are very similar to nonmetals. Metalloids can form alloys with metals. However, the majority of their other physicochemical characteristics are between those of metals and non-metals. Depending on the method of classification, the number of elements characterized as metalloid varies: four, seven and 12 elements may be classified as metalloids. In this chapter, five non-transition metals and metalloids (arsenic, barium, antimony, lead and selenium) with contrasting properties, uses and behavior in the environment have been selected for further description (Kabata-Pendias, 2011). Table 1 and Fig. 1 respectively present the main characteristics of these non-transition elements and their positioning in the periodic table.

The five selected elements are used in numerous anthropogenic activities at the global scale, and of these lead, arsenic and antimony are associated with soil pollution. Table 2 gives the major producers of non-transition metals and metalloids with a major contribution by China (80000, 25000, 2400, and 1100 metric tons respectively of Sb, As, Pb and Se).

Arsenic

General information

Arsenic is a naturally occurring toxic metalloid and has become an element of prevailing concern (Natasha et al., 2021). It is the 20th most abundant element in the Earth's crust with 0.0001% crustal abundance (Shahid et al., 2018). The concentration of As in soil ranges $<1\text{--}250,000\text{ mg kg}^{-1}$, which varies with the nature of mineral sources. Arsenic and its compounds are classified among Group 1 carcinogens by various health- and environment-related organizations. Furthermore, it is categorized as the foremost among priority hazardous elements by the Agency for Toxic Substances and Disease Registry (ATSDR: Substance Priority List). To date, As contamination of the environment has been described widely, mostly in Asian and south Asian regions (Nagra et al., 2022). The major sources of arsenic released into the environment are geogenic (rock weathering, mineral decomposition and geothermal water). The anthropogenic sources include those from agriculture such as pesticides and fertilizers, from mining and smelting, and industrial activities. With the overexploitation of As-rich minerals, the discharge of As to farmland soil from irrigation is increasing. Large As concentrations also occur in the environment from the application of wastewaters and As-rich groundwater (Natasha et al., 2021).

Arsenic contamination of soil

Major As containing minerals are presented in Table 3. The natural or background As levels in the soil ranges from $0.5\text{ to }80\text{ mg kg}^{-1}$, with a mean of 10 mg kg^{-1} . But large soil As concentrations have been described in various areas worldwide (Table 4). Depending

Table 1 Comparison of characteristics of selected non-transition metals and metalloids.

Characteristic	Arsenic	Selenium	Lead	Barium	Antimony
Atomic number	33	34	82	56	51
Atomic mass	74.9216	78.971	207.2	137.327	121.76
Group	pnictogens	chalcogens	Carbon	Alkaline earth metals	pnictogens
Group #	15	16	14	2	15
Period	4	4	6	6	5
Block	p-block	p-block	p-block	s-block	p-block
Melting point (°C)	816.8	221	327.46	727	630.63
Electronegativity (Pauling scale)	2.18	2.55	2.33	0.89	2.05
Van der Waals radius (pm)	185	190	202	268	206
Oxidation states	-3, +3, +5	-2, -1, +1, [2] +2, +3, +4, +5, +6	-4, -2, -1, +1, +2, +3, +4	+1, +2	-3, -2, -1, 0, [2] +1, +2, +3, +4, +5
Density (g cm ⁻³)	5.727	4.81	11.34	3.51	6.697

58 Ce 140.1	59 Pr 140.9	60 Nd 144.2	61 Pm	62 Sm 150.4	63 Eu 152.0	64 Gd 157.3	65 Tb 158.9	66 Dy 162.5	67 Ho 164.9	68 Er 167.3	69 Tm 168.9	70 Yb 173.0	71 Lu 175.0
90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

Fig. 1 Position of selected non-transition metals and metalloids in the periodic table.

Table 2 Annual world mine production (metric tons for 2016–2020 period) of selected non-transition metals and metalloids (United States Geological Survey, 2022, <https://www.usgs.gov/centers/national-minerals-information-center/commodity-statistics-and-information>).

<i>Metals</i>	<i>As</i>	<i>Pb</i>	<i>Sb</i>	<i>Se</i>
2016	36,500	4820	130,000	2200
2017	37,000	4700	150,000	3300
2018	35,000	4400	140,000	2800
2019	33,000	4500	160,000	2800
2020	32,000	4400	153,000	2900

on parent rocks, arsenic exists in more than 200 natural minerals and arsenic concentration in igneous rocks may reach 100 mg kg^{-1} , in limestone (20 mg kg^{-1}), in Mn ores ($15,000 \text{ mg kg}^{-1}$), and in S ores (8000 mg kg^{-1}). The reductive dissolution of these minerals releases arsenic in natural waters and sediments (Natasha et al., 2021).

Arsenic bioavailability and speciation in soil

Arsenic bioavailability depends on dynamic and complex biogeochemical processes such as speciation, bioavailability, bioaccumulation, etc. These soil extractable or bioavailable arsenic fractions (less than the total in the soil) give insight into the mobility or phytoavailability of As in soil. It depends on the soil organic matter content, pH, redox potential (Eh), soil type, microbial functioning, and competing cations. The greater mobility and bioavailability of As in soil results in its large phyto-uptake. Arsenic in soil can exist in free ionic, exchangeable forms, adsorbed, precipitated or present as various minerals. Its bioavailability or potential toxicity is now well understood to depend on its chemical speciation rather than its overall concentrations. The species of As present in soil are listed in [Table 3](#).

Selenium

General information

Selenium is a common element in the Earth's crust (categorized 67th among the most abundant elements). Selenium distribution over the Earth's surface varies and ranges from near zero to 1250 mg kg⁻¹. Selenium is an essential element for living beings. It performs a dual role in the ecosystem: at large concentrations (and depending on speciation, and plant type) it can be phytotoxic, while at low-optimum concentrations, it is beneficial to plants (Natasha et al., 2018). Naturally, Se is released by forest fires,

Table 3 Main minerals of selected non-transition elements with different speciation.

<i>Minerals</i>	<i>Chemical formula</i>
Arsenic	
Aktashite	$\text{Cu}_6\text{Hg}_3\text{As}_4\text{S}_{12}$
Alacránite	(As_8S_9)
Argentobaumhauerite	$\text{Ag}_{1.5}\text{Pb}_{22}\text{As}_{33.5}\text{S}_{72}$
Arsenolite	As_4O_6
Arsenopyrite	FeAsS
Baumhauerite	$\text{Pb}_3\text{As}_4\text{S}_9$
Bellite	$\text{PbCrO}_4, \text{AsO}_4, \text{SiO}_2$
Cervandonite	$(\text{Ce}, \text{Nd}, \text{La})(\text{Fe}^{3+}, \text{Fe}^{2+}, \text{Ti}^{4+}, \text{Al})_3\text{SiAs}(\text{Si}, \text{As})\text{O}_{13}$
Christite	TiHgAsS_3
Claudetite	As_2O_3
Cobaltite	CoAsS
Dimorphite	As_4S_3
Enargite	Cu_3AsS_4
Enneasartorite	$\text{Ti}_6\text{Pb}_{32}\text{As}_{70}\text{S}_{140}$
Erythrite	$(\text{Co}, \text{Ni})_3(\text{AsO}_4)_2 \cdot 8\text{H}_2\text{O}$
Fettelite	$\text{Ag}_{16}\text{HgAs}_4\text{S}_{15}$
Freibergite	$(\text{Ag}, \text{Cu}, \text{Fe})_{12}(\text{Sb}, \text{As})_4\text{S}_{13}$
Gabrielite	$\text{Ti}_6\text{Ag}_3\text{Cu}_6(\text{As}, \text{Sb})_9\text{S}_{21}$ or $\text{Ti}_2\text{AgCu}_2\text{As}_3\text{S}_7$
Galkhaite	$(\text{Cs}, \text{Ti})(\text{Hg}, \text{Cu}, \text{Zn})_6(\text{As}, \text{Sb})_4\text{S}_{12}$
Geocronite	$\text{Pb}_{14}(\text{Sb}, \text{As})_6\text{S}_{23}$
Gersdorffite	NiAsS
Glaucodot	$(\text{Co}, \text{Fe})\text{AsS}$
Gratonite	$\text{Pb}_9\text{As}_4\text{S}_{15}$
Hendekasartorite	$\text{Ti}_2\text{Pb}_{48}\text{As}_{82}\text{S}_{172}$
Heptasartorite	$\text{Ti}_7\text{Pb}_{22}\text{As}_{55}\text{S}_{108}$
Hutchinsonite	$(\text{Ti}, \text{Pb})_2\text{As}_5\text{S}_9$
Jolliffeite	NiAsSe or $(\text{Ni}, \text{Co})\text{AsSe}$
Jordanite	$\text{Pb}_{14}(\text{As}, \text{Sb})_6\text{S}_{23}$
Lautite	CuAsS
Lorándite	TiAsS_2
Lucabindiite	$(\text{K}, \text{NH}_4)\text{As}_4\text{O}_6(\text{Cl}, \text{Br})$
Madocite	$\text{Pb}_{17}(\text{Sb}, \text{As})_{16}\text{S}_{41}$
Marrite	PbAgAsS_3
Minetite	$\text{Pb}_5(\text{AsO}_4)_3\text{Cl}$
Orpiment	As_2S_3
Pääkkönenite	Sb_2AsS_2
Pararealgar	As_4S_4
Pearceite	$\text{Cu}(\text{Ag}, \text{Cu})_6\text{Ag}_9\text{As}_2\text{S}_{11}$
Polybasite	$[(\text{Ag}, \text{Cu})_6(\text{Sb}, \text{As})_2\text{S}_7][\text{Ag}_9\text{CuS}_4]$
Proustite	Ag_3AsS_3
Realgar	$\alpha\text{-As}_4\text{S}_4$
Reinerite	$\text{Zn}_3(\text{AsO}_3)_2$
Routhierite	$\text{Ti}(\text{Cu}, \text{Ag})(\text{Hg}, \text{Zn})_2(\text{As}, \text{Sb})_2\text{S}_6$
Sartorite	PbAs_2S_4
Scorodite	$\text{FeAsO}_4 \cdot 2\text{H}_2\text{O}$
Seligmannite	PbCuAsS_3
Stibarsen	AsSb
Tennantite	$\text{Cu}_{12}\text{As}_4\text{S}_{13}$
Teruggite	$\text{Ca}_4\text{MgAs}_2\text{B}_{12}\text{O}_{22}(\text{OH})_{12} \cdot 12\text{H}_2\text{O}$
Wakabayashilite	$[(\text{As}, \text{Sb})_6\text{S}_9][\text{As}_4\text{S}_5]$
Zimbabweite	$(\text{Na}, \text{K})_2\text{PbAs}_4(\text{Nb}, \text{Ta}, \text{Ti})_4\text{O}_{18}$
Zykaite	$\text{Fe}^{3+}_4(\text{AsO}_4)_3(\text{SO}_4)(\text{OH}) \cdot 15(\text{H}_2\text{O})$
Barium	
Baryte (barium sulfate)	BaSO_4
Witherite (barium carbonate)	BaCO_3
Lead	
Galena	PbS
Boulangerite	$\text{Pb}_5\text{Sb}_4\text{S}_{11}$
Anglesite	PbSO_4
Cerussite	PbCO_3
Selenium	
Mercury selenide	HgSe
Lead selenide	PbSe
Zinc selenide	ZnSe
Copper-indium-gallium diselenide	$\text{Cu}(\text{Ga}, \text{In})\text{Se}_2$
Antimony	
Sulfide mineral stibnite	Sb_2S_3

Table 4 Arsenic concentration (mg kg^{-1}) in soil around the globe.

Level (mg kg^{-1})	Country	Source	References
34.84	Hamadan, Iran	Agriculture	Kharazi et al. (2021)
443	China	Waste Reservoir	Yang et al. (2021)
6.4	Shandong Province, China	Agriculture	Wang et al. (2021)
10.8	Pakistan	Agriculture	Natasha et al. (2021)
6000	Shimen realgar mine, China	Mining	Li et al. (2021)
56	China	Agriculture	Yañez et al. (2018)
1985	Hong Kong, China	Geogenic	Li et al. (2017)
3.81	Pakistan	Agriculture	Rehman et al. (2016)
5240.8	Shimen realgar mine, China	Mining	Tang et al. (2016)
533	Humboldt Smelter, Arizona	Mining	Ramirez-Andreotta et al. (2013)

volcanic eruptions, soil erosion and environmental oxidation reactions. Overall, much larger Se concentrations have been reported in sedimentary rocks (such as coal and shales) than in igneous rocks. Selenium occurs in coal deposits, for example in China, where the Se concentration in coal gangue varies from 0.06 to 17 mg kg^{-1} (Long et al., 2019). The sulphones of Fe, Co, and Ni, however, are detached by magmatic processes, and Se is emitted to the atmosphere. Selenium released to the atmosphere may contaminate soils by atmospheric deposition. For example, long-term Se deposition from rainwater has caused large Se concentrations in coastal areas of Ireland. About 50–65% of total Se global emissions are released by human activities.

Selenium contamination of soil

The major source of Se in soil is from weathering of rocks containing Se. The overall Se quantity occurring in rocks is about 40% of the total Se in soil. Natural Se concentrations in soil vary from 0.01 to 2 mg kg^{-1} , with a world mean of 0.4 mg kg^{-1} . Soils with a Se concentration of less than 0.5 mg kg^{-1} are said to be deficient in this element (Mirlean et al., 2017). The greatest concentrations of Se are found in shales (0.6 mg kg^{-1}) and volcanic rocks (0.35 mg kg^{-1}) from various areas of the world. However, Se concentrations may reach 8000 mg kg^{-1} in areas of its toxicity (Mora et al., 2015). Large Se concentrations in soil occur in North and South Dakota in the United States, and parts of China, Ireland, Wyoming, Venezuela and Colombia (Haug et al. 2007).

Selenium bioavailability and speciation in soil

Selenium occurs naturally in six isotopic forms that have atomic masses of 74, 76, 77, 78, 80 and 82, with ^{78}Se and ^{80}Se accounting for around 73.3% of the total mass. Selenium mostly occurs in ecosystems as sulfides (Table 3) or associated with other inorganic elements such as Cu, Ag, Ni or Pb. Some other minerals containing Se are: berzelianite, eucairite, umangite, penroseite, tiemannite, clausenthalite, klockmannite, palladseite, hakite, and selen-tellurium (Table 3). In soil, Se is present in insoluble forms, as Se^0 and Se^{2-} , but under aerobic conditions these are readily converted into their soluble forms i.e., selenate (SeO_4^{2-}) and selenite (SeO_3^{2-}).

The water-soluble Se fraction or Se in soil-pore water determines Se bioavailability in soil. Small concentrations of bioavailable Se in soil have been described in several studies (Natasha et al., 2018). Selenium mobility and bioavailability depend on the total Se concentration, and also on its oxidation state and speciation in soil. In general, three mechanisms govern Se bioavailability in soil i.e., oxidation vs. reduction, mineralization vs. immobilization, and volatilization. Soil redox conditions and pH, in particular, can affect Se speciation, and thereby its availability and phyto-uptake. In general, Se^0 and Se (II-) predominate in a reducing environment, whereas oxidizing conditions favor Se (IV) and Se (VI). The oxidized chemical forms (SeO_4^{2-} and SeO_3^{2-}) of Se have greater mobility than the reduced forms (Se^0 and Se^{2-}). The bioavailable forms of Se are Se(IV), Se (VI) and organic Se (Natasha et al., 2018) (Table 5).

Antimony

General information

Antimony does not have any biological function, and is categorized as a non-essential element (Pierart et al., 2015). Both natural (volcanic eruptions, rock weathering, wild fires, etc.) and human activities release it into the soil and water ecosystems (Natasha et al., 2019), which can lead to concentrations above the permissible limits. Consequently, it may be a harmful element to the environment. Antimony is the 62nd most abundant element in the Earth's crust (concentrations vary from to 0.2 mg kg^{-1}) and occurs in the Earth's crust as various ores (Table 3). It is ranked at 244th among hazardous contaminants by the ATSDR and the EPA ranked it at 114 in the US Environmental Protection Agency (USEPA) priority list of pollutants.

Table 5 Selenium concentration in soil (mg kg^{-1}) around the globe.

Level (mg kg^{-1})	Country	Source	References
0.315	Sichuan Basin, China	Agriculture	Liu et al. (2021)
0.02–1.67	Shaanxi province, China	Agriculture	Liu et al. (2021)
0.87	China	Agriculture	Xiao et al. (2020)
0.85–11.46	Enshi, Central China	Agriculture	Chang et al. (2019)
3	Mumbai, India	–	Paikaray (2016)
0.31–0.65	Dashan Region, China	Mountain region	Xing et al. (2015)
13.1	Punjab India	–	Eiche et al. (2015)
0.9	Gumuskoy (Kutahya), Turkey	Mining area	Sasmaz et al. (2015)
0.62	Netherlands	Grassland field	Supriatin et al. (2015)
0.53	Netherlands	Arable land	Supriatin et al. (2015)

Antimony contamination of soil

Table 6 shows that Sb concentration can vary widely in soil. The background level is 0.2 mg kg^{-1} , however, as a result of mining and smelting activities, and vehicle pollution, soil Sb concentrations might become many times greater (Pierart et al., 2015). In some areas, Sb concentration in soil exceeds the threshold level of 36 mg kg^{-1} . Antimony concentrations may reach 5045 mg kg^{-1} in the sediments and soil of contaminated areas.

Antimony bioavailability and speciation in soil

In soil, Sb occurs in different oxidation forms such as Sb (III) and Sb (V). However, under reducing conditions, Sb (III) is dominant, especially in paddy soils. The oxidative dissolution of antimony trisulfide (Sb_2S_3) in rocks naturally releases Sb in soil (Hu et al., 2017). Antimony may form complexes with different soil constituents such as Fe and Mn oxides. Numerous soil factors govern Sb mobility and phytoavailability such as its concentration, soil characteristics, redox conditions, and the competing ions. These factors can affect Sb speciation and hence its biogeochemical behavior in the soil-plant nexus. In particular, soil pH and redox conditions greatly affect soil-plant transfers of Sb in soil (Cui et al., 2015; Natasha et al., 2019).

Lead

General information

Lead (Pb) is a ubiquitous and versatile element, with a background concentration of 27 mg kg^{-1} (Kabata-Pendias, 2011; Natasha et al., 2020). It is the 36th most abundant element in the Earth's crust (with an overall value of 14 mg kg^{-1}). Lead is one of the major inorganic persistent global pollutants; its increased use in various industries and numerous other anthropogenic activities has enhanced its concentration in soil. With modern urbanization and industrialization, and with the recent past uses of Pb (material such as paints, petroleum, etc.), large quantities of Pb have been discharged into the soil (Fang et al., 2019). Exposure to lead is a health concern, especially for young children and pregnant women. Lead can affect almost every organ and system in your body. The nervous system is the main target for lead poisoning in children and adults. Exposure to lead can cause developmental effects in children, including but not limited to reduced IQ and attention span, hyperactivity, impaired growth, and learning disability.

Table 6 Antimony concentration in soil (mg kg^{-1}) around the globe.

Level (mg kg^{-1})	Country	Source	References
300	Shimen realgar mine, China	Mining	Li et al. (2021)
0.47	Agricultural soil	Australia	Egodawatta et al. (2020)
7900	Mining site	Australia	Ngo et al. (2020)
1080	Mining area	China	Zhao et al. (2020)
2.5	Farmland	China	Geng et al. (2020)
24,345	Mining area	China	Deng et al. (2020)
0.23	Playgrounds	China	Peng et al. (2019)
1.7	Waste incinerator	China	Li et al. (2019)
218–292	Shimen realgar mine, China	Mining	Tang et al. (2016)
11,798	Xikuangshan, Hunan, China	Mining	Okkenhaug et al. (2011)
7316	Xikuangshan, Hunan, China	Mining	Wang et al. (2011)
472	Taipu River basin	–	Wang et al. (2019)
0.02–26.4	Murcia Region, Spain		García-Lorenzo et al. (2015)

Lead contamination of soil

In soil, large concentrations of lead are found around: roads as a result of emissions into the atmosphere from vehicles that used leaded gasoline for decades, buildings that used lead paint, toxic waste sites and other areas close to industrial sites such as the recycling of lead from car batteries (Natasha et al., 2020). Children and adults can be exposed to lead in soil through various leisure activities (gardening, eating contaminated fruits and vegetables or soil ingestion).

Interpreting soil lead results can be challenging as there is no single international threshold that defines acceptable levels of lead in soil. For example, the US-EPA (United States-Environment Protection Agency) sets a threshold for a child-occupied facility, such as a play area, of 0.4 mg g^{-1} of total lead. Lead generally forms stable complexes with soil components, therefore the observed environmental and toxic effects can be reduced (Zhang et al., 2019).

Lead bioavailability and speciation in soil

In general, Pb is distributed in different forms in soil for which the bioavailability differs significantly. Small concentrations only of soluble Pb are observed in soil because of its strong binding with organic matter and clay minerals (Kushwaha et al., 2018). Lead generally concentrates in the upper soil due to its low mobility and potential to form complexes with different soil particles.

It is now well-established that Pb speciation in soil mainly controls its geochemistry as well as phyto-uptake. Different chemical forms of lead have different adsorption and desorption potential that affect its behavior in the soil-plant nexus (Shahid et al., 2014). It is known that Pb exists in soil in different forms such as (i) Pb ions, (ii) complexed by HCO_3^- (bicarbonate), CO_3^{2-} (carbonate), SO_4^{2-} (sulfate) and Cl^- (chloride), or attached to fulvic, amino, and humic acids. In addition, Pb can also exist attached to organic matter, Fe-oxides, clay particles and biological materials.

According to Kushwaha et al. (2018), lead has no biological function and induces ecotoxicity in soil microflora and plants. Certain plants and microbes have evolved to develop detoxification mechanisms to counter the ecotoxic effect of lead and can be used for bioremediation of lead contaminated sites. These are mainly based on changes in chemical speciation. There are numerous options to reduce the bioavailability of lead in polluted soil including changes in soil pH, organic matter content, presence of various amendments, clay minerals and presence of organic colloids and iron oxides (Pourrut et al., 2011).

Barium

General information

Barium (Ba) is a silvery-white alkaline earth metal that occurs naturally in different compounds. It is the 16th most abundant element (0.04%) in Earth's crust, but it is not present in free ionic form because of its high chemical reactivity (Lu et al., 2019). It is also present naturally in surface water aquifers. Natural sources of Ba result from the weathering of minerals and rocks. Anthropogenically, Ba is mainly released by industrial activities. In the atmosphere, the residence time of Ba might be up to several days.

Barium contamination of soil

The major minerals of Ba are witherite (BaCO_3) and baryte (BaSO_4), often found in underground ore deposits (Krishna et al., 2020) (Table 3). Barium and its compounds have a variety of uses including as getter material in electronic tubes, rodenticides, color in paints and as an X-ray contrast medium. Barium is relatively abundant in the Earth's crust, with mean values ranging between 265 and $835 \text{ } \mu\text{g g}^{-1}$ dry weight (dw) depending on the soil type. Natural concentrations of Ba in soils range between 5 and $21 \text{ } \mu\text{g g}^{-1}$. In most soil systems, Ba is not a very mobile element due to the formation of water-insoluble salts and their partition into soils and sediments.

Barium bioavailability and speciation in soil

Witherite and baryte are both insoluble in water. i.e., 0.02 g L^{-1} (at 20°C) for BaCO_3 and 0.001 g L^{-1} (at 0°C) for barium sulfate. In minerals containing these elements, Ba largely substitutes Pb and Sr cations, but less so for the somewhat smaller Ca and K ions. Replacement for K shows a linked substitution to maintain charge balance. Barium occurs as a trace or minor element in potassium feldspar and mica in common igneous rocks, where it substitutes K. There is less substitution of Ca in amphiboles, apatite, calcite, plagioclase feldspar, and pyroxenes. Barium is usually found as baryte or as an organic complex in sedimentary rocks and hydrothermal deposits. The solubility, mobility and phytoavailability of Ba also vary under varying soil conditions such as pH and competing cations. Barium has been reported to precipitate in the presence of carbonate, sulfate, or chromate. For example, at $\text{pH} < 9.3$ with $\geq 2 \text{ mg L}^{-1}$ of SO_4 , Ba may precipitate as baryte (BaSO_4). Similarly, Ba precipitates as witherite (BaCO_3) with a large CO_3 concentration.

The water-soluble compounds of Ba are toxic and are often used as rodenticides. Sulfate is the stable form of sulfur in the Earth's crust and soil, therefore barium sulfate (BaSO_4), is fairly widespread, which explains the existence of trace quantities of baryte in many sedimentary rocks. Baryte has low solubility in oxidizing (sulfate-stable) conditions, but dissolves easily in reducing conditions with hydrogen sulfide (H_2S) (Lu et al., 2019). Barium solubility is known to be influenced by its species. Barium salts such as chloride, acetate, perchlorate, and nitrate are water soluble, whereas baryte (BaSO_4) is a very insoluble mineral (Krishna et al., 2020) whatever the solvent (water, acids, and bases) and is therefore has low (eco)toxicity (Menzie et al., 2008) (Fig. 2) (Tables 7 and 8).

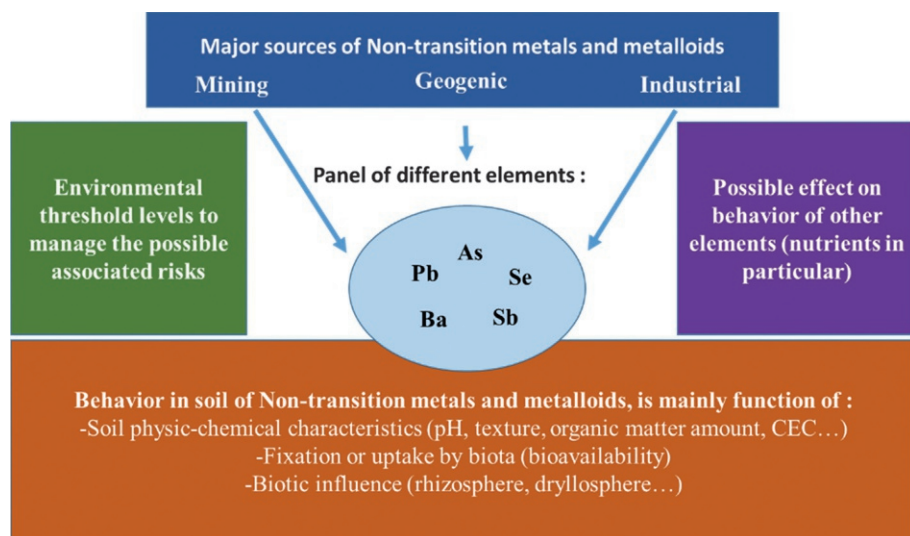


Fig. 2 Non-transition metals and metalloids in soils.

Table 7 Lead concentration (mg kg^{-1}) in soil around the globe.

Level (mg kg^{-1})	Country	Source	References
16.83	Bandel thermal power station, Bengal, India	Agriculture	Mandal et al. (2022)
124	Henan, China	Agriculture	Yang et al. (2022)
20.6	Abha City, Netherlands	Agriculture	Eid et al. (2021)
45.18	Erbil City	Agriculture	Tariq (2021)
36.9	Azerbaijan Province, Iran	Industrial	Shojaei et al. (2021)
38.79	Tehran	Industrial	Mirzaei et al. (2021)
400	Chicago, United States	Multiple historical sources	Watson et al. (2021)
20	Copper mining towns of Zambia	Mining	Chileshe et al. (2020)
49	Vehari	Wastewater	Sarwar et al. (2020)
>400	18 Chinese provinces	Geogenic	Wu et al. (2020)
>800	Urban soil of China	Geogenic	Liu et al. (2016)

Table 8 Barium concentration (mg kg^{-1}) in soil around the globe.

Level (mg kg^{-1})	Country	Source	References
144	New York	Garden soil	McBride et al. (2014)
518–65,760	China	Mining area	Lu et al. (2019)
108.9	Turkey	Orchards	Sungur et al. (2019)
490–5300	Norway	Mineral soils	Myrvang et al. (2016)
39.3	Egypt	Vertisols	Elbehiry et al. (2021)
222.08	Brazil	Agriculture	Reis et al. (2021)
3.6	Brazil	Agriculture	Rabêlo et al. (2021)
261.8	Brazil	Coastal area	Lima et al. (2021)
400	China	Industrial area	Ding et al. (2022)

Conclusions and perspectives

The five non-transition elements (arsenic, barium, antimony, lead and selenium) considered in this chapter show a wide range of different concentrations in the soils, illustrating global scale pollution by natural and anthropogenic activities such as metal(loid) mining. The main soil factors influencing the behavior of the five elements in soil-plant nexus are the soil pH, redox potential and biological activity, which govern their speciation and bioavailability in soils.

A sustained effort is needed to minimize further pollution risk induced by these non-transition elements, both by reducing vulnerability through remediation and mitigation strategies and by increasing law enforcement for minimizing the use of chemical fertilizers and pesticides on arable land. In 2022, with the reinforcement of environmental regulation and mobilization of public space for health, a reduction of chemicals released to the environment is being progressively observed in various sectors. For example, Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) regulation in Europe aims to reduce the number of chemicals currently used. For agriculture, biological practices are being developed increasingly at the global scale to provide alternatives to pesticides. In addition, evaluating the potential of low-cost and eco-friendly amendments (such as green manure plants, biochar, etc.) combined with proper management of fertilizers and soil conditioners to govern the phyto-uptake of metal(loid)s should lead to practical and appropriate recommendations to farmers, gardeners and other relevant stakeholders.

References

- Chang C, Yin R, Wang X, Shao S, Chen C, and Zhang H (2019) Selenium translocation in the soil-rice system in the Enshi seleniferous area, Central China. *Science of the Total Environment* 669: 83–90.
- Cui X-D, Wang Y-J, Hockmann K, and Zhou D-M (2015) Effect of iron plaque on antimony uptake by rice (*Oryza sativa* L.). *Environmental Pollution* 204: 133–140.
- Deng R, Tang Z, Hou B, Ren B, Wang Z, Zhu C, Kelly S, and Hursthouse A (2020) Microbial diversity in soils from antimony mining sites: geochemical control promotes species enrichment. *Environmental Chemistry Letters* 18: 911–922.
- Eiche E, Bardelli F, Notthstein AK, Charlet L, Göttlicher J, Steininger R, Dhillon KS, and Sadana US (2015) Selenium distribution and speciation in plant parts of wheat (*Triticum aestivum*) and Indian mustard (*Brassica juncea*) from a seleniferous area of Punjab, India. *Science of The Total Environment* 505: 952–961.
- Fang T, Yang K, Lu W, Cui K, Li J, Liang Y, Hou G, Zhao X, and Li H (2019) An overview of heavy metal pollution in Chaohu Lake, China: Enrichment, distribution, speciation, and associated risk under natural and anthropogenic changes. *Environmental Science and Pollution Research* 26: 29585–29596.
- Hu X, He M, Li S, and Guo X (2017) The leaching characteristics and changes in the leached layer of antimony-bearing ores from China. *Journal of Geochemical Exploration* 176: 76–84.
- Haug A, Graham RD, Christophersen OA, and Lyons GH (2007) How to use the world's scarce selenium resources efficiently to increase the selenium concentration in food. *Microbial Ecology in Health and Disease* 19: 209–228.
- Kabata-Pendias A (2011) *Trace Elements in Soils and Plants*, 3rd edn. CRC Press.
- Krishna S, Jaiswal A, Gupta M, Sharma D, and Ali Z (2020) Barium poisoning with analytical aspects and its management. *International Journal of Advanced Research in Medicinal Chemistry* 2: 20–27.
- Kushwaha A, Hans N, Kumar S, and Rani R (2018) A critical review on speciation, mobilization and toxicity of lead in soil-microbe-plant system and bioremediation strategies. *Ecotoxicology and Environmental Safety* 147: 1035–1045.
- Liu X, Wang M, Zhou F, Zhai H, Qi M, Liu Y, Li Y, Zhang N, Ma Y, Huang J, Ren R, and Liang D (2021) Selenium bioavailability in soil-wheat system and its dominant influential factors: A field study in Shaanxi province, China. *Science of the Total Environment* 770: 144664.
- Long J, Zhang S, and Luo K (2019) Selenium in Chinese coal gangue: Distribution, availability, and recommendations. *Resources, Conservation and Recycling* 149: 140–150.
- Lu Q, Xu X, Liang L, Xu Z, Shang L, Guo J, Xiao D, and Qiu G (2019) Barium concentration, phytoavailability, and risk assessment in soil-rice systems from an active barium mining region. *Applied Geochemistry* 106: 142–148.
- Menzie CA, Southworth B, Stephenson G, and Feisthauer N (2008) The importance of understanding the chemical form of a metal in the environment: The case of barium sulfate (barite). *Human and Ecological Risk Assessment* 14: 974–991.
- Mirlean N, Seus-Arrache E, and Vlasova O (2017) Selenium deficiency in subtropical littoral pampas: Environmental and dietary aspects. *Environmental Geochemistry and Health* 40(1): 543–556.
- Mora M, Durán P, Acuña J, Cartes P, Demanet R, and Gianfreda L (2015) Improving selenium status in plant nutrition and quality. *Journal of Soil Science and Plant Nutrition* 15: 486–503.
- Nagra MA, Natasha N, Bibi I, Tariq TZ, Naz R, Ansar S, Shahid M, Murtaza B, Imran M, Khalid MS, Masood N, Shah GM, Niazi NK, and Dumat C (2022) Biowaste-based sorbents for arsenic removal from aqueous medium and risk assessment. *Environmental Geochemistry and Health*. <https://doi.org/10.1007/s10653-022-01402-w>.
- Natasha SM, Niazi NK, Khalid S, Murtaza B, Bibi I, and Rashid MI (2018) A critical review of selenium biogeochemical behavior in soil-plant system with an inference to human health. *Environmental Pollution* 234: 915–934.
- Natasha SM, Khalid S, Dumat C, Pierart A, and Niazi NK (2019) Biogeochemistry of antimony in soil-plant system: Ecotoxicology and human health. *Applied Geochemistry* 106: 45–59.
- Natasha DC, Shahid M, Khalid S, and Murtaza B (2020) Lead Pollution and Human Exposure: Forewarned is Forearmed, and the Question Now Becomes How to Respond to the Threat!. In: *Lead in Plants and the Environment*, pp. 33–65. Springer.
- Natasha SM, Khalid S, Niazi NK, Murtaza B, Ahmad N, Farooq A, Zakir A, Imran M, and Abbas G (2021) Health risks of arsenic buildup in soil and food crops after wastewater irrigation. *Science of the Total Environment* 772: 145266.
- Pierart A, Shahid M, Séjalon-Delmas N, and Dumat C (2015) Antimony bioavailability: Knowledge and research perspectives for sustainable agricultures. *Journal of Hazardous Materials* 289: 219–234.
- Pourrut B, Shahid M, Dumat C, Winterton P, and Pinelli E (2011) Lead uptake, toxicity, and detoxification in plants. *Reviews of Environmental Contamination and Toxicology* 213: 113–136.
- Shahid M, Dumat C, Pourrut B, Silvestre J, Laplanche C, and Pinelli E (2014) Influence of EDTA and citric acid on lead-induced oxidative stress to *Vicia faba* roots. *Journal of Soils and Sediments* 14: 835–843.
- Shahid M, Khalid M, Dumat C, Khalid S, Niazi NK, Imran M, Bibi I, Ahmad I, Hammad HM, and Tabassum RA (2018) Arsenic level and risk assessment of groundwater in Vehari, Punjab Province, Pakistan. *Exposure and Health* 10: 229–239.
- Wang X, He M, Xi J, and Lu X (2011) Antimony distribution and mobility in rivers around the world's largest antimony mine of Xikuangshan, Hunan Province, China. *Microchemical Journal* 97(1): 4–11.
- Zhang Y, Hou D, O'Connor D, Shen Z, Shi P, Ok YS, Tsang DC, Wen Y, and Luo M (2019) Lead contamination in Chinese surface soils: Source identification, spatial-temporal distribution and associated health risks. *Critical Reviews in Environmental Science and Technology* 49: 1386–1423.

Further reading

- Chileshe MN, Syampungani S, Festin ES, Tigabu M, Daneshvar A, and Odén PC (2020) Physico-chemical characteristics and heavy metal concentrations of copper mine wastes in Zambia: Implications for pollution risk and restoration. *Journal of Forestry Research* 31: 1283–1293.
- Ding D, Kong L, Jiang D, Wei J, Cao S, Li X, Zheng L, and Deng S (2022) Source apportionment and health risk assessment of chemicals of concern in soil, water and sediment at a large strontium slag pile area. *Journal of Environmental Management* 304: 114–228.
- Eid EM, Shaltout KH, Alamri SA, Alrumman SA, Hussain AA, Sewelam N, El-Bebany AF, Alfharhan AH, Picó Y, and Barcelo D (2021) Prediction models based on soil properties for evaluating the uptake of eight heavy metals by tomato plant (*Lycopersicon esculentum* Mill.) grown in agricultural soils amended with sewage sludge. *Journal of Environmental Chemical Engineering* 9: 105977.
- Egodawatta LP, Holland A, Koppel D, and Jolley DF (2020) Influence of soil phosphate on the accumulation and toxicity of arsenic and antimony in choy sum cultivated in individually and co-contaminated soils. *Environmental Toxicology and Chemistry* 39: 1233–1243.
- Elbehiry F, Elbasiouny H, Cappuyns V, and Brevik EC (2021) Available concentrations of some potentially toxic and emerging contaminants in different soil orders in Egypt and assessment of soil pollution. *Journal of Soils and Sediments* 21: 3645–3662.
- García-Lorenzo ML, Martínez-Sánchez MJ, Pérez-Sirvent C, López-Sánchez J, Molina-Ruiz J, and Tudela ML (2015) Geogenic distribution of arsenic (As) and antimony (Sb) in soils of the Murcia Region in Spain. *Environmental Forensics* 16: 88–95.
- Geng L, Yang Z, and Xu Z (2020) Effects of antimony contamination in soil on the nutrient composition of three green leafy vegetables. *Journal of Soils and Sediments* 20: 2217–2224.
- Kharazi A, Leili M, Khazaei M, Alkhani MY, and Shokoohi R (2021) Human health risk assessment of heavy metals in agricultural soil and food crops in Hamadan, Iran. *Journal of Food Composition and Analysis* 100: 103890.
- Li J, Beiyuan J, Tsang DCW, Wang L, Poon CS, Li X-D, and Fendorf S (2017) Arsenic-containing soil from geogenic source in Hong Kong: Leaching characteristics and stabilization/solidification. *Chemosphere* 182: 31–39.
- Li Y, Zhang H, Shao L, Zhou X, and He P (2019) Impact of municipal solid waste incineration on heavy metals in the surrounding soils by multivariate analysis and lead isotope analysis. *Journal of Environmental Sciences* 82: 47–56.
- Li Y, Zhang M, Xu R, Lin H, Sun X, Xu F, Gao P, Kong T, Xiao E, Yang N, and Sun W (2021) Arsenic and antimony co-contamination influences on soil microbial community composition and functions: Relevance to arsenic resistance and carbon, nitrogen, and sulfur cycling. *Environment International* 153: 106522.
- Lima LS, Oliveira UC, Rangel CM, Barreto CF, Aguiar VMDC, Neto JAB, and Fonseca EMD (2021) Assessment of pollutants in soil and underground water samples collected from the north coastal area of Rio de Janeiro State—Brazil. *Journal of Integrated Coastal Zone Management* 20: 219–231.
- Liu N, Wang M, Zhou F, Zhai H, Qi M, Liu Y, Li Y, Zhang N, Ma Y, Huang J, Ren R, and Liang D (2021) Selenium bioavailability in soil-wheat system and its dominant influential factors: A field study in Shaanxi province, China. *Science of the Total Environment* 770: 144664.
- Liu L, Zhang X, and Zhong T (2016) Pollution and health risk assessment of heavy metals in urban soil in China. *Human and Ecological Risk Assessment: An International Journal* 22: 424–434.
- Lu Q, Xu X, Liang L, Xu Z, Shang L, Guo J, Xiao D, and Qiu G (2019) Barium concentration, phytoavailability, and risk assessment in soil-rice systems from an active barium mining region. *Applied Geochemistry* 106: 142–148.
- Mandal S, Bhattacharya S, and Paul S (2022) Assessing the level of contamination of metals in surface soils at thermal power area: Evidence from developing country (India). *Environmental Chemistry and Ecotoxicology* 4: 37–49.
- McBride MB, Shaylor HA, Spliethoff HM, Mitchell RG, Marquez-Bravo LG, Ferenz GS, Russell-Anelli JM, Casey L, and Bachman S (2014) Concentrations of lead, cadmium and barium in urban garden-grown vegetables: Impact of soil variables. *Environmental Pollution* 194: 254–261.
- Mirzaei F, Abbasi Y, and Sohrabi T (2021) Modeling the distribution of heavy metals in lands irrigated by wastewater using satellite images of Sentinel-2. *The Egyptian Journal of Remote Sensing and Space Science* 24: 537–546.
- Myrvang MB, Hillersøy MH, Heim M, Bleken MA, and Gjengedal E (2016) Uptake of macro nutrients, barium, and strontium by vegetation from mineral soils on carbonatite and pyroxenite bedrock at the Lillebukt Alkaline Complex on Stjernøy, Northern Norway. *Journal of Plant Nutrition and Soil Science* 179: 705–716.
- Natasha BI, Shahid M, Niazi NK, Younas F, Naqvi SR, Shaheen SM, Imran M, Wang H, Hussaini KM, Zhang H, and Rinklebe J (2021) Hydrogeochemical and health risk evaluation of arsenic in shallow and deep aquifers along the different floodplains of Punjab, Pakistan. *Journal of Hazardous Materials* 402: 124074.
- Ngo LK, Price HL, Bennett WW, Teasdale PR, and Jolley DF (2020) DGT and selective extractions reveal differences in arsenic and antimony uptake by the white icicle radish (*Raphanus sativus*). *Environmental Pollution* 259: 113815.
- Okkenhaug G, Zhu Y-G, Luo L, Lei M, Li X, and Mulder J (2011) Distribution, speciation and availability of antimony (Sb) in soils and terrestrial plants from an active Sb mining area. *Environmental Pollution* 159: 2427–2434.
- Paikaray S (2016) Origin, mobilization and distribution of selenium in a Soil/Water/Air System: A Global perspective with special reference to the Indian Scenario. *CLEAN – Soil, Air, Water* 44: 474–487.
- Peng T, O'Connor D, Zhao B, Jin Y, Zhang Y, Tian L, Zheng N, Li X, and Hou D (2019) Spatial distribution of lead contamination in soil and equipment dust at children's playgrounds in Beijing, China. *Environmental Pollution* 245: 363–370.
- Rabêlo FHS, Lavres J, Pinto FA, and Alleoni LRF (2021) Photosynthetic Parameters and Growth of Rice, Lettuce, Sunflower and Tomato in an Entisol as Affected by Soil Acidity and Bioaccumulation of Ba, Cd, Cu, Ni, and Zn. *Archives of Environmental Contamination and Toxicology* 81: 91–106.
- Ramirez-Andreotta MD, Brusseau ML, Artioli JF, and Maier RM (2013) A greenhouse and field-based study to determine the accumulation of As in common homegrown vegetables grown in mining-affected soils. *Science of the Total Environment* 443: 299–306.
- Rehman ZU, Khan S, Qin K, Brusseau ML, and Din I (2016) Quantification of inorganic As exposure and cancer risk via consumption of vegetables in southern selected districts of Pakistan. *Science of the Total Environment* 550: 321–329.
- Reis IMS, de Melo WJ, Lima ESA, Pereira MG, da Conceição Silva US, de Sousa MA, Júnior JM, Camargo LA, de Melo GMP, Araújo ASF (2021) Barium and cadmium in tropical soils cropped and under native forest. Research Square. <https://doi.org/10.21203/rs.3.rs-813516/v1>.
- Sarwar T, Shahid M, Khalid S, Shah AH, Ahmad N, Naeem MA, Ul Haq Z, Murtaza B, and Bakhat HF (2020) Quantification and risk assessment of metal build-up in soil-plant system after irrigation with untreated city wastewater in Vehari, Pakistan. *Environmental Geochemistry and Health* 42(12): 4281–4297.
- Sasmaz M, Akgül B, and Sasmaz A (2015) Distribution and Accumulation of Selenium in Wild Plants Growing Naturally in the Gumuskoy (Kutahya) Mining Area, Turkey. *Bulletin of Environmental Contamination and Toxicology* 94: 598–603.
- Shojaei S, Jafarpour A, Shojaei S, Gyasi-Agyei Y, and Rodrigo-Comino J (2021) Heavy metal uptake by plants from wastewater of different pulp concentrations and contaminated soils. *Journal of Cleaner Production* 296: 126–345.
- Sungur A, Gur E, Everest T, Soylak M, and Ozcan H (2019) Assessment of relationship between geochemical fractions of Ba in soil of cherry orchards and plant barium uptake and determination by inductively coupled plasma optical emission spectrometry. *Atomic Spectroscopy* 40: 173.
- Supriatin S, Weng L, and Comans RN (2015) Selenium speciation and extractability in Dutch agricultural soils. *Science of the Total Environment* 532: 368–382.
- Tang J, Liao Y, Yang Z, Chai L, and Yang W (2016) Characterization of arsenic serious-contaminated soils from Shimen realgar mine area, the Asian largest realgar deposit in China. *Journal of Soils and Sediments* 16: 1519–1528.
- Tariq F (2021) Heavy metals concentration in vegetables irrigated with municipal wastewater and their human daily intake in Erbil city. *Environmental Nanotechnology, Monitoring & Management* 16: 100–475.
- Wang Y, Yang Z, Wang T, et al. (2019) Analysis and risk assessment of heavy metals in soils of Taipu river basin. *Environmental Engineering* 37: 18–22.

- Wang S, Gao Z, Zhang Y, Zhang H, Wu Z, Jiang B, Liu Y, and Dong H (2021) Source and Health Risk Assessment of Heavy Metals in Soil–Ginger System in the Jing River Basin of Shandong Province, North China. *International Journal of Environmental Research and Public Health* 18: 6749.
- Watson GP, Martin NF, Grant ZB, Batka SC, and Margenot AJ (2021) Soil lead distribution in Chicago, USA. *Geoderma Regional* 28: e00480.
- Wu X, Cai Q, Xu Q, Zhou Z, and Shi J (2020) Wheat (*Triticum aestivum* L.) grains uptake of lead (Pb), transfer factors and prediction models for various types of soils from China. *Ecotoxicology and Environmental Safety* 206: 111387.
- Xiao K, Lu L, Tang J, Chen H, Li D, and Liu Y (2020) Parent material modulates land use effects on soil selenium bioavailability in a selenium-enriched region of southwest China. *Geoderma* 376: 114554.
- Xing K, Zhou S, Wu X, Zhu Y, Kong J, Shao T, and Tao X (2015) Concentrations and characteristics of selenium in soil samples from Dashan Region, a selenium-enriched area in China. *Soil Science and Plant Nutrition* 61: 889–897.
- Yañez LM, Alfaro JA, and Bovi Mitre G (2018) Absorption of arsenic from soil and water by two chard (*Beta vulgaris* L.) varieties: A potential risk to human health. *Journal of Environmental Management* 218: 23–30.
- Yang S, Feng W, Wang S, Chen L, Zheng X, Li X, Zhou D (2021) Farmland heavy metals can migrate to deep soil at a regional scale: A case study on a wastewater-irrigated area in China. *Environmental Pollution* 281:116977.
- Yang Y, Li Y, Wang T, Chen W, Wang M, and Dai Y (2022) Derivation of human health risk-based thresholds for lead in soils promote the production of safer wheat and rice. *Ecotoxicology and Environmental Safety* 230: 113–131.
- Zhao L, Shangguan Y, Yao N, Sun Z, Ma J, and Hou H (2020) Soil migration of antimony and arsenic facilitated by colloids in lysimeter studies. *Science of the Total Environment* 728: 138874.

Lanthanides, actinides and radon

Franco Ajmone Marsan^a and Mery Malandrino^b, ^aUniversità di Torino, DISAFA—Chimica agraria e pedologia, Grugliasco, Torino, Italy; ^bDepartment of Chemistry, Università di Torino, Torino, Italy

© 2023 Elsevier Ltd. All rights reserved.

Introduction	225
Lanthanides	225
Actinides	228
Radon	231
Conclusions	232
References	232

Abstract

The relentless technological evolution of the world has resulted in the diffusion of new contaminants. The concurrent advances of analytical techniques have facilitated the recognition of actual and potential threats to human health and of the environment. Radon is a radioactive gas generated in rock and soil which is potentially mobile and can diffuse into the indoor and outdoor atmosphere. Lanthanides are elements classified as 'emerging contaminants' as their intense use in technological devices is resulting in their dispersion in the environment. Among the actinides, only thorium and uranium are widespread in nature.

Key points

- Technology demand for rare elements
- Dispersion of elements in soils
- Soil properties control their availability and mobility
- Improvement of analytical techniques
- Experimental gaps

Introduction

Lanthanides and actinides are elements that have gained recognition in recent decades for being identified as dangerous for the health of humans and of the environment. Radon, being a decay product of the actinide radium is also included here. Lanthanides, although they belong to the group of 'rare earth elements', are not rare in the lithosphere and are being intensively mined for applications in many so-called 'green' technologies. Actinides and radon radionuclides are mainly of natural origin as they derive from rocks, but the former have also been added to the environment through the exploitation of nuclear reactions. The consequent dispersion of these elements in the environment has been of increasing concern in recent years. This is particularly significant in an age when the green and ecological transition is regarded as vitally important.

The concern about these elements rises from a certain lack of knowledge about their reactivity, and consequently their fate, in the ecosystems. Little is known so far about the role of these elements in the soil–plant system and research is required to understand the possible threat deriving from the increased use of lanthanides and actinides.

Lanthanides

The lanthanides are a chemically uniform group of substances of Group IIIb in the Periodic Table that includes 15 metallic elements with atomic numbers 57–71 from lanthanum (La) through to lutetium (Lu), Fig. 1. Together with the elements scandium (Sc) and yttrium (Y) they constitute the group commonly known as the rare-earth elements (REE) or rare-earth metals.

In lanthanide atoms, the configuration of the valence electrons of the outermost shell is the same for all the species while the 4f orbitals are progressively filled with the increase in atomic number, while the 5s and 5p subshells are already filled and screen the 4f orbitals; this makes the physical and chemical properties of these elements remarkably similar. Another consequence is the so-called 'lanthanide contraction' in which the ionic radius decreases progressively from La³⁺ to Lu³⁺ as inefficient shielding of the increased nuclear charge by the extra 4f electrons causes the 5s and 5p electrons to contract in toward the nucleus. The progressive decrease of the ionic dimension changes the charge/radius ratio influencing their reactivity, especially in aqueous solutions. In natural environments, most lanthanides occur as trivalent ions except for Ce (cerium) that may also be +4 in oxic conditions (>300 mV at pH 7). Samarium (Sm), europium (Eu) and ytterbium (Yb) may reach the +2 state under extremely low

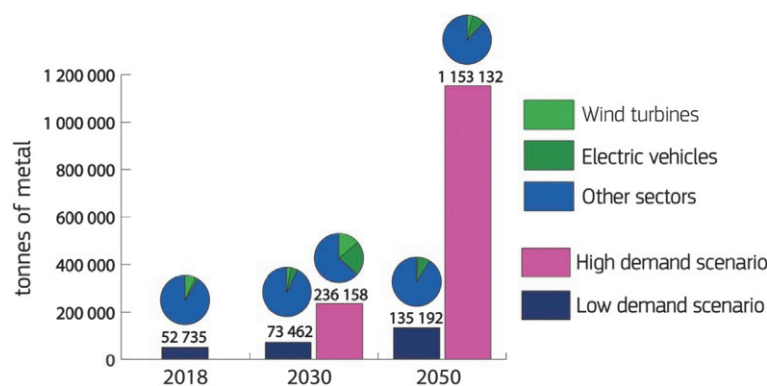


Element	La	Ce	Pr	Nd	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
Abundance	30	60	6.7	27	5.3	1.3	4	0.65	3.8	0.8	2.1	0.3	2	0.35

After Wedepohl KH (1995) The composition of the continental crust. *Geochimica et Cosmochimica Acta* 59: 1217–1232.

Lanthanides, although they belong to the ‘rare’ elements group, are not actually rare on Earth. Their estimated average concentration in the Earth’s crust ranges from around 130 mg kg⁻¹ to 240 mg kg⁻¹, which is greater than other commonly exploited elements. Cerium is more abundant in the Earth’s crust than copper (Cu); neodymium (Nd) and La are more abundant than lead (Pb), and praseodymium (Pr), Sm and gadolinium (Gd) are more abundant than tin (Sn). [Table 1](#) presents the average abundances (in mg kg⁻¹) of lanthanides in Earth’s crust ([from Wedepohl, 1995](#)). Although not rare, they are seldom concentrated into economically-significant amounts for extraction. Despite this fact, there were only 184,000 t of all 17 Rare Earth Oxides produced globally in 2018 compared with 21 million t of copper, 4.4 million t of lead and 310,000 t of tin in the same year ([Castilloux, 2019](#)).

Lanthanides are non-essential elements for life but, due to their unique physical and chemical properties, they have become extremely important as a component of present-day life because of their widespread application in a variety of technologies. These elements have distinctive magnetic, phosphorescent, and catalytic properties (e.g., Ce, La, Nd, and Gd). Lanthanides are used in



Source: Joint Research Centre (JRC).

Fig. 2 Estimated global demand for Nd, Pr, Dy, and Tb for wind turbine generators, motors for electric vehicles and other sectors, according to low- and high-demand scenarios. From: Alves Dias P, Bobba S, Carrara S, Plazzotta B (2020) The role of rare earth elements in wind energy and electric mobility, EUR 30488 EN. Luxembourg: Publication Office of the European Union, ISBN 978-92-79-27016-4, doi:10.2760/303258, JRC122671.

many domestic, medical, industrial, and strategic applications because of their unique catalytic, metallurgical, nuclear, electrical, magnetic, and luminescent properties. Examples of the many applications for REE are their use in magnets and super magnets, electric motors, metal alloys, electronic, and computing equipment, batteries, catalytic converters, petroleum refining, medical imaging, coloring agents in glass and ceramics, phosphors, lasers, and glass. A stable provision of lanthanides for the next 20–30 years is therefore important for the success of transitioning into a carbon-neutral society (Fig. 2).

The elements of the lanthanide group (and rare earth elements in general) have recently become ‘contaminants of emerging concern’ (CEC). The concern arises because (i) they are currently unregulated, (ii) they are not routinely monitored in environmental and public health programs, (iii) they are micropollutants detectable with analytical equipment with very low detection limits, and (iv) their mechanisms of environmental and human toxicity are poorly understood.

The concentration of lanthanides in the soil is very variable, with a mean concentration of REE in the topsoil of Europe of 125.6 mg kg^{-1} , which is close to that of the upper continental crust (146.4 mg kg^{-1}) (Gwenzi et al., 2018). However, typical contents vary from a few tens of mg kg^{-1} for Nd to a few hundreds of mg kg^{-1} for the sum of the rare earth elements, of which lanthanides are a large part (Laveuf and Cornu, 2009). This large variation depends mainly on the type of parent material from which the soil originated. In some tropical soils of Brazil, Silva et al. (2020) found that the background concentrations of lanthanides and their spatial variation were dependent on the concentration of iron (Fe) oxides. In Germany, on the contrary, Mihajlovic and Rinklebe (2018) found that the dependence of lanthanide content on the soil parent material decreased in the order: carbonatite > basalt > orthogneiss > clay slate > loess > sandstone > Pleistocene and Holocene sediments > organic material.

The majority of all mineral resources that contain lanthanide are concentrated in the minerals bastnasite, monazite, and xenotime. Other minerals, typically heavy minerals such as anatase, ilmenite, sphene, rutile, and zircon, can also contain lanthanides. These minerals are relatively resistant to weathering; therefore, they do not contribute significantly to the mobility of the lanthanides. Phosphatic minerals can contain large REE concentrations—in the order of thousands of mg kg^{-1} of lanthanides—depending on the origin and genesis of the phosphate. Apatite is the main mineral used for the production of phosphate fertilizers, and lanthanide in apatite comes from the substitution of Ca by ions of trivalent lanthanide. The P-fertilizers are thus an indirect but relevant anthropogenic source of lanthanides in soils. It has been calculated that with a P-fertilizer application of $300 \text{ kg ha}^{-1} \text{ year}^{-1}$, the input of lanthanides can range from 30 to $170 \text{ g ha}^{-1} \text{ year}^{-1}$ (Laveuf and Cornu, 2009). In addition to this indirect input, in recent years lanthanides have been used as a complement to fertilizers to increase seed germination, root growth, chlorophyll content, plant resistance, and agricultural productivity (Gwenzi et al., 2018).

The increasing use of these elements in industrial applications and their high content in the waste from electrical and electronic equipment (e-waste) could represent another contribution to the concentration of lanthanides in soils and in the environment in general (Brewer et al., 2022). Unregulated disposal of technological items and uncontrolled recycling attempts may become sources of lanthanides in soils, especially in urban areas, the cores of the Technosphere.

The ever-growing demand for lanthanides for technological applications implies an increase in mining and production activities that may lead to adverse environmental and health effects. Mining operations can release dust containing REEs, other toxic metals and chemicals into the surrounding soil, water bodies and impact wildlife and vegetation in addition to humans. In general, the concentrations of lanthanides in bulk rock or ore are up to four orders of magnitude higher than those observed in soils. Therefore, the latter may be comparable to, or even higher, than those of rocks if the soils are in the vicinity of mining sites (not necessarily lanthanides mining sites) or plants where rocks or ores are processed such as phosphorites processing plants. In these areas, weathering of dumps and tailings accumulated during excavation or mineral processing (and containing trace amounts of

lanthanides) can be responsible for the critical accumulation of lanthanides in soil and the spread of contamination to other environmental compartments. This is the case concerning the largest known deposit of lanthanides in China, where the elements are dispersed in soils, sediments and waters of the adjacent area (Zhou et al., 2020). Similarly, the disposal of tailings in coal mining operations in Brazil favored the dispersal of lanthanides and their accumulation in soybean (*Glycine max* (L.) Merrill) plants cultivated in the surrounding areas (Galhardi et al., 2020).

The mobility of lanthanides in soil depends on the properties of the elements and the properties of the soil. Their chemistry in soil is similar but there are slight differences in hydrolysis and complexation constants, and in the resistance to weathering of the lanthanide-bearing minerals that result in fractionation and differential migration potential. In general lanthanide mobility is low, apart from some leaching of the light elements (lanthanum through gadolinium) and a relative enrichment of heavy elements (terbium to lutetium).

Lanthanides in the soil profile would show relations with the reactive components such as clay, carbonates, organic matter, aluminum (Al), iron, and manganese (Mn) and chemical properties such as pH, redox potential and cation exchange capacity. Similar to other soil metallic elements, their solubility increases with decreasing pH and with increasing redox potential (Mihajlovic and Rinklebe, 2018).

Lanthanides seem to have a particular affinity for organic substances and iron–manganese oxides. They form complexes with various organic (e.g., humic and fulvic acids) ligands often have coordination numbers exceeding 6, and sometimes as high as 12 (Gwenzi et al., 2018). Sorption of lanthanide on the surface of Fe oxides increases with increasing pH as the charge of the oxide becomes increasingly negative. Their affinity with Fe–Mn oxides makes them prone to being released when the oxides are solubilized under anoxic soil conditions. It has been reported that the availability of La, Ce, Gd, and Y increases with a decrease in pH and redox potential (Gwenzi et al., 2018).

In an experiment with two soils of different mineralogy and texture, Dinali et al. (2019) observed that sorption of lanthanides was similar in both soils at their original pH but increased with increasing pH two-fold in the Mollisol compared to the Oxisol. The differences in sorption were explained in terms of mineralogy, 2:1 clays vs oxide respectively, and organic matter content. Sorption was particularly influenced by soil pH for Oxisols, in which minerals with pH-dependent charge dominate.

Analogous results have been obtained by Rogova et al. (2020). They spiked soddy-podzolic and Chernozem soils, both rich in organic matter, with La, Ce, and Nd and found that fixation of the elements in the form of metal–organic complexes was predominant for both soil types (35–38% in podzolic soil and 50–79% in Chernozem) with respect to the exchangeable, specifically sorbed or Mn bound fractions.

Liu et al. (2022) undertook to remediate mine tailings by the addition of amendments and cultivation of plants, and observed a significant increase in pH and nutrients, which led to a decrease of 80–90% in soil-extractable lanthanides. They identified pH, and total C, N, and P among the most important factors controlling the availability of soil lanthanides.

Pathways by which lanthanides migrate in soil include plant uptake, erosion, leaching of lanthanide–inorganic complexes, organic complexation, eluviation of clay, precipitation reactions, and adsorption by phyllosilicates and oxyhydroxides of Fe and Mn.

Geochemical fractionation studies suggest that the bioavailability of lanthanides is relatively low while the residual fraction is relatively high which mean that their translocation in plants is difficult to establish. Gwenzi et al. (2018) reported that several studies indicated a preferential accumulation of lanthanides in the roots of plants. Results collated by Khan et al. (2017) showed an inconsistent picture where some plants may accumulate lanthanides while others preferentially translocate other elements. Some plants were observed to be hyperaccumulators of lanthanides such as *Carya tomentosa* Nutt. (Mockernut hickory) *Pronephrium simplex* Holtt., and *Dicranopteris dichotoma* Bernh. (Old World forked fern). Carpenter et al. (2015) reported large concentrations of Dy (dysprosium), Er, Nd, Pr, Sm, and Tb in roots of native plants *Asclepias syriaca* L. (common milkweed), *Desmodium canadense* (L.) DC. (showy tick-trefoil), *Panicum virgatum* L. (switchgrass) growing on contaminated soils. More recently *Phytolacca americana* L. (pokeweed) was found to accumulate lanthanides and has been proposed for the phytoremediation of contaminated sites and even the recycling of lanthanides (Yuan et al., 2018). The specific pattern of lanthanides uptake by plants has also been applied to establish the possible traceability of some food products, wine in particular (Bronzi et al., 2020; Barbera et al., 2021).

Actinides

Actinides are elements whose isotopes are all radioactive and are listed in Table 2. The presence of actinium (Ac), protactinium (Pa), uranium (U) and thorium (Th) in nature is because the semi-transformation periods of the ^{235}U , ^{238}U and ^{232}Th isotopes are long enough to ensure that these species are still present today. However, only thorium and uranium are widespread in nature.

It is essential to take into account the electronic configuration of the ground state of the actinides (Table 2) and the competition between the configurations $5f^n 7s^2$ and $5f^{n-1} 6d 7s^2$ to understand their chemical behavior and to compare it with that of lanthanides. Actinides, especially the first elements of the series, namely Th, Pa, U, Np, Pu, and Am, exhibit a greater tendency to provide more bonding electrons and, consequently, to give higher oxidation states than lanthanides. This is because the promotion of an electron from the $5f$ to $6d$ orbital requires less energy than that required for the promotion of an electron from the $4f$ to $5d$ orbital. The second half of the actinide series, namely Cf, Es, Fm, Md, No, and Lr, have chemical characteristics more similar to the lanthanides. The energies of the orbitals in the actinides are similar over a certain range of atomic numbers (especially from U to americium (Am)) and, because the orbitals overlap spatially they are all available to form bonds. This situation is highlighted

Table 2 Actinides and their properties.

Z	Symbol	Name	Electronic configuration	Oxidation ^a states
89	Ac	Actinium	6d7s ²	+3
90	Th	Thorium	6d ² 7s ²	+4
91	Pa	Protactinium	5f ² 6d7s ² or 5f ¹ 6d ² 7s ²	+3; +4; +5
92	U	Uranium	5f ³ 6d7s ²	+3; +4; +5 ; +6
93	Np	Neptunium	5f ⁵ 7s ²	+3; +4; +5 ; +6; +7
94	Pu	Plutonium	5f ⁶ 7s ²	+3; +4 ; +5; +6; +7
95	Am	Americium	5f ⁷ s ²	+2; +3 ; +4 ; +5; +6
96	Cm	Curium	5f ⁷ 6d7s ²	+3 ; +4
97	Bk	Berkelium	5f ⁸ 6d7s ² or 5f ⁹ 7s ²	+3 ; +4
98	Cf	Californium	5f ¹⁰ 7s ²	+2; +3 ; +4
99	Es	Einsteinium	5f ¹¹ 7s ²	+2; +3
100	Fm	Fermium	5f ¹² 7s ²	+2; +3
101	Md	Mendelevium	5f ¹³ 7s ²	+2; +3
102	No	Nobelium	5f ¹⁴ 7s ²	+2; +3
103	Lr	Lawrencium	5f ¹⁴ 6d7s ²	+3

^aBold numbers represent the most stable oxidation state.

in the chemical properties by the fact that actinides have a greater aptitude to form complexes than lanthanides with their almost exclusively ionic bonds.

Uranium and thorium are primordial radioactive elements and are ubiquitous in the Earth's crust. The uranium concentration in the Earth's crust is about 2.4 mg kg⁻¹. Typical concentrations of uranium range from 1 to 10 mg kg⁻¹ in sandstone, shale or limestone, whereas granite contains up to 15 mg kg⁻¹ uranium. Uranium exists in the environment in the form of tetravalent (U⁴⁺) or hexavalent (U⁶⁺) oxidation states. The solubility of U in water, and hence its mobility in the lithosphere and potential for water pollution, depends strongly on its oxidation state and the presence of organic chelators. The U(VI) is the most common form in soil (80–90%) as the uranyl (UO₂²⁺) cation, and it is also the most mobile form of U. It exists in solution predominantly as the stable ion UO₂²⁺ and as soluble carbonate complexes. The uranyl cation is the main species responsible for U toxicity in aquatic organisms. Within the pH range of 4.0 and 7.5 and in the absence of dissolved inorganic ligands (carbonate, fluoride, sulfate, and phosphate), the hydrolysis species UO₂OH⁺, UO₂(OH)₂, and (UO₂)₂(OH)₂²⁺ dominate the U(VI) speciation. However, uranium in soil and water forms complexes with sulfate and phosphate as well as with carbonate and hydroxide. These complexes increase the total solubility of U. The U(IV), which is present under anoxic conditions, has a much lower solubility than U(VI) (Merian et al., 2004).

Uranium mining, processing of uranium ores, application of fertilizers, production and disposal of radioactive materials and recycling of spent nuclear fuel are the main sources of ²³⁵U. Emissions of uranium during mining and processing are generally low. Significant increases above natural levels may occur in local aquatic systems. Uranium in the hexavalent state is easily susceptible to migration in water while tetravalent uranium is immobile and stays in the soil itself. Infiltration of rainwater and subsurface flow may carry significant amount of uranium from the rocks and soils, which is then redistributed in the surrounding subsurface environment (Fig. 3). In general, groundwater was found to have a larger uranium concentration than the other fresh water sources because of prolonged residence time in the aquifers with rocks and soils containing uranium. However, the presence of uranium in water systems could also be due to the anthropogenic activities listed above. Intake of drinking water with a high concentration of uranium for a long duration can cause health effects such as kidney damage, cancer, increase in defective births, early graying of hair and cardiovascular diseases. Military uses (fission bombs based on ²³⁵U) and accidental releases of radionuclides such as during the Chernobyl accident may lead to the contamination of large areas. However, the contribution of uranium and its decay products to the radiotoxicity of nuclear fallout is marginal. Most of the acute and protracted dose is caused by relatively volatile and soluble fission or activation products such as ¹³¹I, ⁹⁰Sr, ¹³⁷Cs, or ¹⁴C with short, intermediate and long half-lives of 8 days, 29 years, 30 years, and 5730 years, respectively.

The Earth's crust contains about 13 mg kg⁻¹ thorium, as a constituent of more than 100 minerals. Abundance is comparable to that of beryllium and cobalt, and three times more than that of uranium. Typical concentrations range from 1 to 10 mg kg⁻¹ in sandstone, shale or limestone, whereas granite contains up to 80 mg kg⁻¹ thorium. Monazite, an orthophosphate of rare earth and other elements, may contain up to 28% Th oxide (Alloway, 2013).

These elements were produced in supernova synthesis and, therefore, they are indicators of geochemical processes and can be used to identify the origin of soil or rocks. Anthropogenic sources of these elements include mainly waste generated in coal-fired power plants and nuclear energy processing plants, where natural radionuclides of U and Th are used as a fuel. These elements may leach from rocks, soils and waste generated over time by industrial processing of naturally occurring radioactive materials (NORM), and transfer to humans through the food chain. Hazardous effects due to U and Th radionuclides in biological systems are well

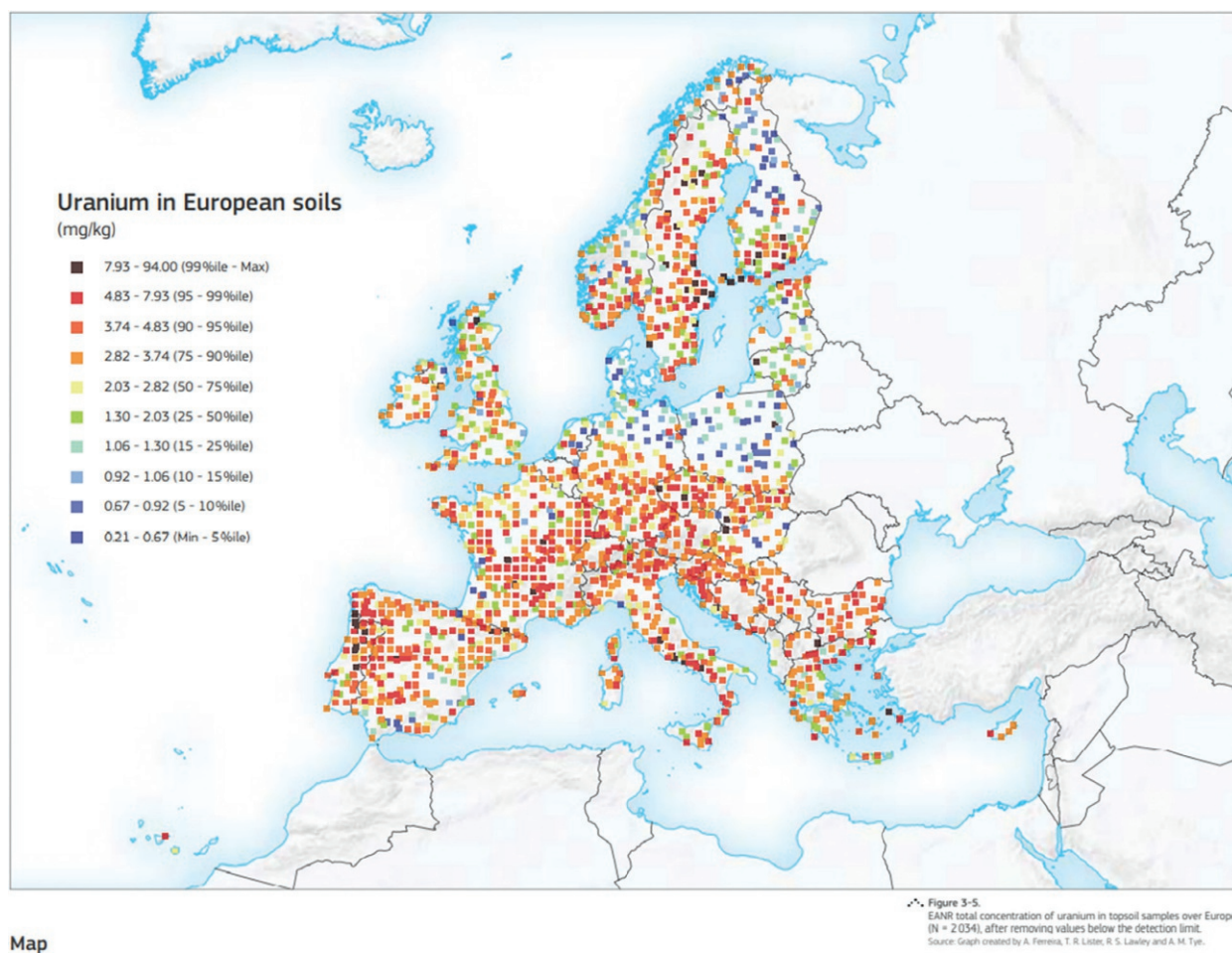


Fig. 3 Uranium in European soils. Source: European Atlas of Natural Radiation Online version. ISBN 978-92-76-08258-3. doi:10.2760/46388, Catalogue number KJ-02-19-425-EN-N.

documented (IAEA, 2010). In particular, ^{232}Th and long-lived uranium isotopes ^{235}U and ^{238}U are both chemotoxic and radiotoxic, whereas all other isotopes and decay products are dominated by their radiotoxicity due to the ionizing radiation emitted during their radioactive decay. This is the result of the inverse relationship between the specific radioactivity (Bq kg^{-1}) and the half-life of a radionuclide. Since ionizing radiation has the potential to damage the genome, either directly or through ions and radicals generated in the vicinity of DNA, all substances emitting ionizing radiation must be considered mutagens and carcinogens. Moreover, for humans and animals, uranium and its salts are also highly chemotoxic. Dermatitis, kidney damage, acute arterial lesions, cancer, and cardiovascular diseases may occur. Finally, uranyl compounds readily complex with the phosphate-containing mineral matrix of bone.

The determination of concentrations of U and Th in environmental samples, mainly in soils, is required as background data to estimate their contamination level compared to natural abundances. Moreover, it is important to quantify the transfer of radionuclides from soil to plants by the transfer factor (TF), which is defined as the ratio of activity concentration of a radionuclide in dry plants to that of the same radionuclide in dry soil. The TF varies with the type of plant and radionuclide, from site to site and by season. These variations depend on several factors such as physical and chemical properties of the soil, environmental conditions, and the chemical forms of the radionuclide in soil (Saenboonruang et al., 2018). Soil properties that affect the absorption of radionuclides can include mineral and soil structure, organic matter content, pH and fertility.

The distribution coefficient (K_d) of radionuclides in the soil-solution system is an important parameter that expresses the ability of these elements to migrate in groundwater (Sheppard et al., 2006). It is defined as the ratio of the concentration of the element in the solid (soil) phase divided its concentration in the liquid phase (water) at equilibrium, usually expressed as L kg^{-1} . The K_d value is influenced by various physicochemical properties of soils such as soil texture, pH, carbonates (CaCO_3), cation exchange capacity (CEC) and chemistry of the transporting groundwater. For this reason, it is difficult to derive generic K_d values for radionuclides. The K_d values are usually estimated for site-specific conditions. Indeed, TF and K_d of artificial and natural radionuclides were calculated for different soils. However, the large variation in U K_d values reflects a strong dependence on soil properties, especially pH.

Numerous studies have attempted chemical fractionation of U and Th using sequential extractions. Each fraction corresponds to chemical forms of the elements, namely exchangeable (extractable with MgCl_2), which bond to the carbonates (extractable with CH_3COONa with CH_3COOH), bond to organic matter (extractable with $\text{CH}_3\text{COONH}_4$), bond to oxides (extractable with $\text{NH}_4\text{C}_2\text{O}_4$ and $\text{H}_2\text{C}_2\text{O}_4$) and residual (representing the radionuclides strongly bound to soil particles). Typically, the so-called 'exchangeable' and 'carbonate-bound' fractions are recognized as readily available (Rout et al., 2016). As a complement to chemical extractions, isotopic dilution assays have the potential to add new insights into time-dependent changes in U speciation. The technique involves adding a known amount of an isotopic tracer to a soil suspension in equilibrium; the resulting change in isotopic ratios gives an indication of the isotopically exchangeable concentration, which provides a more mechanistically based estimate of the reactive pool (Young et al., 2005).

All this information (TF, K_d , chemical fractionation and speciation) is of paramount relevance in the context of radioecological environmental contamination assessment and management. Is this a basis for better management of contamination?

Radon

Radon (^{222}Rn) is a radioactive gas produced by the decay of radium (^{226}Ra), which is a decay product of uranium and thorium in the uranium series (The series begins with naturally occurring ^{238}U and includes astatine, bismuth, lead, polonium, protactinium, radium, radon, thallium, and thorium). Radon is potentially mobile and can diffuse through rock and soil to escape into the indoor and outdoor atmosphere. Uranium is ubiquitous in rocks and soils, and Rn emissions vary according to the specific mineralogical composition.

Radon atoms are short lived, having a half-life of 3.8 days but are considered to be the most significant contributors to human exposure from natural sources. Inhalation of the gaseous radioactive isotope ^{222}Rn causes concern because it has been connected to lung cancer (Abrahams, 2002). Epidemiological studies on humans pointed out that up to 14% of lung cancers are induced by exposure to low to moderate concentrations of radon (Ćujić et al., 2021).

The gas emitted from fractured or porous rocks can easily move through the soil into the air before decaying. Radon can be a component of soil gas and enter and accumulate in houses by diffusion, determined by the difference in pressure between the house and soil beneath.

Not all radon produced in this process moves into soil pores. The number of radon atoms released per the number of radon atoms generated is known as the emanation fraction. The amount of Rn escaping to the atmosphere varies greatly as a function of the geology (e.g., U content and chemical form, degree of jointing or faulting), soil characteristics (e.g., permeability, moisture content) and climatic variables (e.g., temperature, humidity) (Oliver and Khayrat, 2001). The highest concentrations of uranium are found in granites (Ćujić et al., 2021). The emanation fraction is inversely proportional to the sizes of the soil particles in relation to their associated surface areas. Furthermore, soil moisture influences the emanation fraction by reducing the speed of radon atoms and preventing them from embedding in another particle.

Averages of the radon emanation fractions determined from minerals, rock, soil (1025 samples), mill tailings (77), and fly ash (46) were found to be 0.03 (range 0.00–0.25), 0.13 (0.00–0.40), 0.20 (0.00–0.83), 0.17 (0.04–0.35) and 0.03 (0.00–0.35) respectively. The Rn emanation fractions vary considerably even for the same materials (Sakoda et al., 2011).

Sometimes local soils can contribute to elevated outdoor ^{222}Rn concentrations, and such exposure needs to be included in epidemiologic studies.

Recent studies in the tectonically active Himalaya area (Kumar et al., 2022) found an average radon concentration in soils of up to $4969 \pm 561 \text{ Bq m}^{-3}$ and a radium content of 51.04 Bq kg^{-1} . Similarly, Walat and Khairi (2021) reported a mean soil depth activity of ^{222}Rn and ^{226}Ra in the Duhok district of Iraq of $305 \pm 44 \text{ Bq m}^{-3}$ and $33 \pm 5 \text{ Bq kg}^{-1}$ respectively.

The contribution of uranium mining to contamination of the surrounding environment is depicted in a study by Huang et al. (2021). They reported a maximum activity concentration of ^{238}U and ^{226}Ra near a mine in China of 5600 Bq kg^{-1} and 4791 Bq kg^{-1} , respectively. Similar results were also obtained by Komati et al. (2021) in soils around a gold mine in South Africa.

Radium may be taken up by plants, as observed by Parekh et al. (2008) in nuts grown in various localities of South America. Radium concentrations in nuts grown in the four regions were similar: ^{226}Ra ($17\text{--}27 \text{ mBq g}^{-1}$) and ^{228}Ra ($18\text{--}31 \text{ mBq g}^{-1}$), suggesting that the amounts in the nuts were influenced by bioaccumulation rather than soil content.

Several countries around the world have adopted regulations to control exposure to radon in indoor environments. In Europe (Fig. 4), a guideline value for ^{222}Rn indoor concentration of 300 Bq m^{-3} has been recommended for the European Union Member States (Directive 2013/59/Euratom).

Analogously, the World Health Organization indicates a reference level possibly lower than 100 Bq m^{-3} and not higher than 300 Bq m^{-3} . In Italy, a survey by the National Institute for Environmental Protection and Research (ISPRA) and the Italian National Institute of Health (ISS) provided an estimate of the average national value of indoor ^{222}Rn of about 70 Bq m^{-3} which is well above the worldwide mean of about 40 Bq m^{-3} (Guarino et al., 2022). In the UK, Cornwall and Devon are particularly affected by radon; they have 53% of UK homes that are estimated to contain a concentration of ^{222}Rn above the 'action level' of 200 Bq m^{-3} (Abrahams, 2002). This is possibly related to the areas with granite exposed at the surface or close to it.

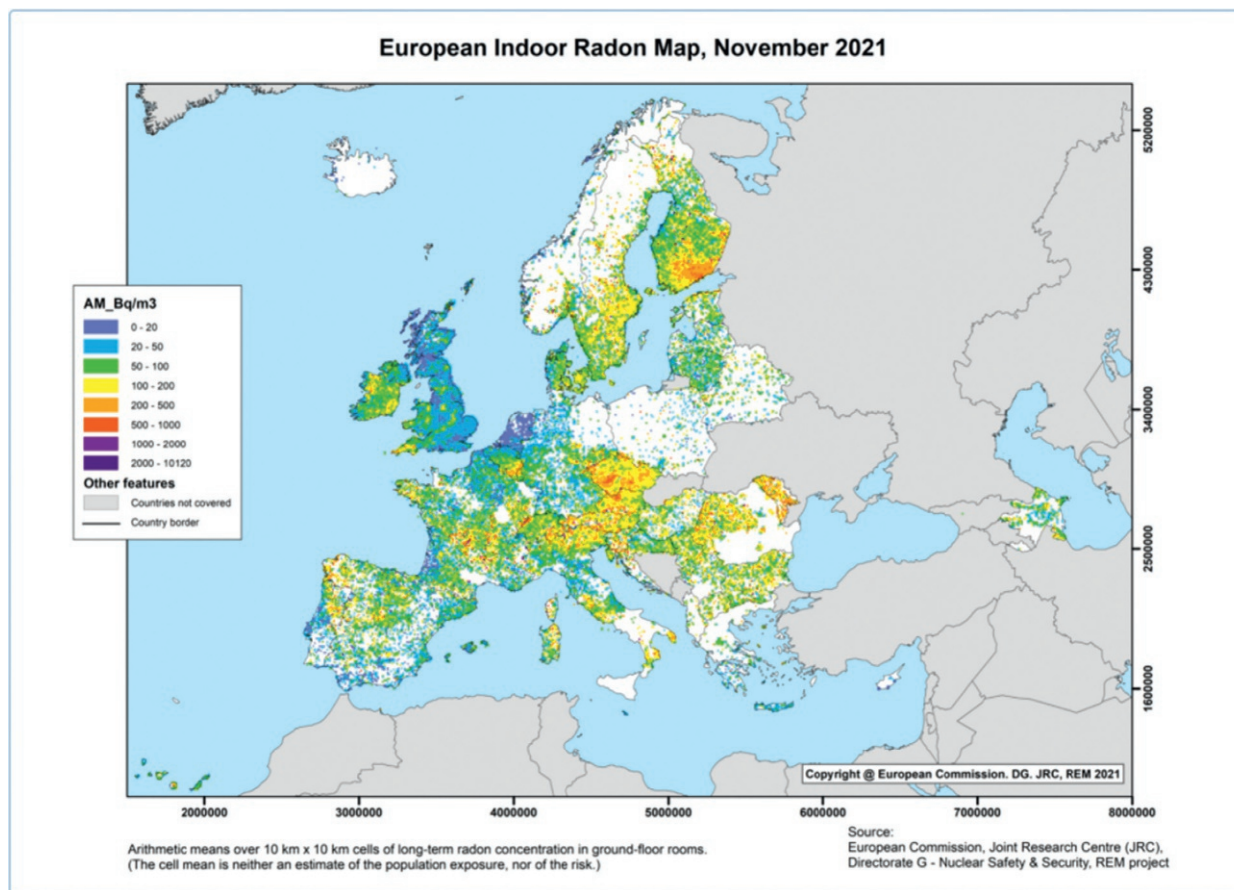


Fig. 4 Indoor Rn map of Europe. Source: European Atlas of Natural Radiation Online version. ISBN 978-92-76-08258-3. doi:10.2760/46388. Catalogue number KJ-02-19-425-EN-N.

Conclusions

The technological evolution of human activities has brought about this trail of new contaminants. At the same time the advances of analytical techniques have facilitated the recognition of actual and potential threats to the health of humans and of the environment.

In general, it is evident that the research effort in this field has progressed, but it is far from obtaining a clear picture of the environmental significance of radon, lanthanides and actinides. Monitoring of these contaminants becomes indispensable if the ecosystem is to be protected.

References

- Abrahams PW (2002) Soils: Their implications to human health. *The Science of the Total Environment* 291: 1–32.
- Alloway BJ (2013) *Heavy Metals in Soils: Trace Metals and Metalloids in Soils and Their Bioavailability*. Springer Science & Business Media.
- Barbera M, Zuddas P, Palazzolo E, and Saiano F (2021) The distribution of Rare Earth Elements discriminates the growth substrate of *Vitis vinifera* L. *Chemosphere* 266: 128993.
- Brewer A, Dror I, and Berkowitz B (2022) Electronic waste as a source of rare earth element pollution: Leaching, transport in porous media, and the effects of nanoparticles. *Chemosphere* 287: 132217.
- Bronzi B, Brilli C, Beone GM, Fontanella MC, Ballabio D, Todeschini R, Consonni V, Grisoni F, Parri F, and Buscema M (2020) Geographical identification of Chianti red wine based on ICP-MS element composition. *Food Chemistry* 315: 126248.
- Carpenter D, Boutin C, Allison JE, Parsons JL, and Ellis DM (2015) Uptake and effects of six rare earth elements (REEs) on selected native and crop species growing in contaminated soils. *PLoS One* 10: e0129936. <https://doi.org/10.1371/journal.pone.0129936>.
- Castilloux R (2019) *Rare earth elements: Small market, big necessities*. Adamas intelligence report.
- Čujić M, Janković ML, Petrović J, Dragović R, Đorđević M, Đokić M, and Dragović S (2021) Radon-222: Environmental behavior and impact to (human and non-human) biota. *International Journal of Biometeorology* 65: 69–83.
- Dinali GS, Root RA, Amistadi MK, Chorover J, Lopes G, and Guilherme LRG (2019) Rare earth elements (REY) sorption on soils of contrasting mineralogy and texture. *Environment International* 128: 279–291.
- Galhardi JA, Leles BP, de Mello JWV, and Wilkinson KJ (2020) Bioavailability of trace metals and rare earth elements (REE) from the tropical soils of a coal mining area. *Science of the Total Environment* 717: 134484.

- Guarino A, Cicchella D, Lima A, and Albanese S (2022) Radon flux estimates, from both gamma radiation and geochemical data, to determine sources, migration pathways, and related health risk: The Campania region (Italy) case study. *Chemosphere* 287: 132233.
- Gwenzi W, Mangori L, Danha C, Chaukura N, Dunjana N, and Sanganyado E (2018) Sources, behaviour, and environmental and human health risks of high technology rare earth elements as emerging contaminants. *The Science of the Total Environment* 636: 299–313. <https://doi.org/10.1016/j.scitotenv.2018.04.235>.
- Huang Y, Yang L, Zhao F, Guo G, and Wu L (2021) Spatial distribution and characteristic of radiological hazard of the paddy field around a decommissioned uranium mine in eastern China. *Journal of Radioanalytical and Nuclear Chemistry* 327: 789–799.
- International Atomic Energy Agency (2010) *Handbook of Parameter Values for the Prediction of Radionuclide Transfer in Terrestrial and Freshwater Environment, Technical Reports Series N° . 472*. Vienna: IAEA.
- Khan AM, Abu Bakar NK, Abu Bakar AF, and Ashraf MA (2017) Chemical speciation and bioavailability of rare earth elements (REEs) in the ecosystem: A review. *Environmental Science and Pollution Research* 24: 22764–22789. <https://doi.org/10.1007/s11356-016-7427-1>.
- Komati FS, Strydom R, and Ntwaeaborwa OM (2021) Measurements of radon exhalation from a South African gold mine tailings using sealed tube method. *Radioprotection* 56: 327–336.
- Kumar G, Bhadwal R, Kumar M, Kumari P, Kumar A, Walia V, Mehra R, and Goyal A (2022) Radioactivity monitoring in the vicinity of Jawalamukhi thrust NW Himalaya, India for tectonic study. *Natural Hazards*. <https://doi.org/10.1007/s11069-021-05134-5>.
- Laveuf C and Cornu S (2009) A review on the potentiality of rare earth elements to trace pedogenetic processes. *Geoderma* 154: 1–12.
- Liu C, Liu WS, Huot H, Guo MN, Zhu SC, Zheng HX, Morel JL, Tang YT, and Qiu RL (2022) Biogeochemical cycles of nutrients, rare earth elements (REEs) and Al in soil-plant system in ion-adsorption REE mine tailings remediated with amendment and ramie (*Boehmeria nivea* L.). *Science of the Total Environment* 809: 152075.
- Merian E, Anke M, Ihnat M, and Stoepler M (2004) *Elements and Their Compounds in the Environment*, 2nd edn. Weinheim: Wiley-VCH Verlag.
- Oliver MA and Khayrat AL (2001) A geostatistical investigation of the spatial variation of radon in soil. *Computers & Geosciences* 27(2001): 939–957.
- Mihajlovic J and Rinklebe J (2018) Rare earth elements in German soils-A review. *Chemosphere* 205: 514–523.
- Parekh PP, Khan AR, Torres MA, and Kitto ME (2008) Concentrations of selenium, barium, and radium in Brazil nuts. *Journal of Food Composition and Analysis* 21(2008): 332–335.
- Rogova OB, Fedotov PS, Dzheloda RK, and Karandashev VK (2020) Fractionation and fixation of rare earths elements in soils: Effect of spiking with lanthanum, cerium, and neodymium chlorides. *Journal of Rare Earths* 40: 143–152. <https://doi.org/10.1016/j.jre.2020.12.006>.
- Rout S, Kumar A, Ravi PM, and Tripathi RM (2016) Understanding the solid phase chemical fractionation of uranium in soil and effect of ageing. *Journal of Hazardous Materials* 317: 457–465.
- Saenboonruang K, Phonchanthuek E, and Prasandee K (2018) Soil-to-plant transfer factors of natural radionuclides (^{226}Ra and ^{40}K) in selected Thai medicinal plants. *Journal of Environmental Radioactivity* 184–185: 1–5.
- Sakoda A, Ishimori Y, and Yamaoka K (2011) A comprehensive review of radon emanation measurements for mineral, rock, soil, mill tailing and fly ash. *Applied Radiation and Isotopes* 69: 1422–1435.
- Sheppard SC, Sheppard MI, Tait JC, and Sanipelli BL (2006) Revision and metaanalysis of selected biosphere parameter values for chlorine, iodine, neptunium, radium, radon and uranium. *Journal of Environmental Radioactivity* 89: 115–137.
- Silva CMCAC, Nascimento RC, da Silva YJAB, Barbosa RS, da Silva YJAB, do Nascimento CWA, and van Straaten P (2020) Combining geospatial analyses to optimize quality reference values of rare earth elements in soils. *Environmental Monitoring and Assessment* 192: 453. <https://doi.org/10.1007/s10661-020-08406-y>.
- Walat AA and Khairi MSA (2021) Determination of radium and radon exhalation rate as a function of soil depth of Duhok Province—Iraq. *Journal of Radiation Research and Applied Science* 14: 486–494.
- Wedepohl KH (1995) The composition of the continental crust. *Geochimica et Cosmochimica Acta* 59: 1217–1232.
- Young SD, Zhang H, Tye AM, Maxted A, Thums C, and Thornton I (2005) Characterizing the availability of metals in contaminated soils. I. The solid phase: Sequential extraction and isotopic dilution. *Soil Use and Management* 21: 450–458.
- Yuan M, Liu C, Liu WS, Guo MN, Morel JL, Huot H, Yu HJ, Tang YT, and Qiu RL (2018) Accumulation and fractionation of rare earth elements (REEs) in the naturally grown *Phytolacca americana* L. in southern China. *International Journal of Phytoremediation* 20: 415–423.
- Zhou J, Wang X, Nie L, McKinley JM, Liu H, Zhang B, and Han Z (2020) Geochemical background and dispersion pattern of the world's largest REE deposit of Bayan Obo, China. *Journal of Geochemical Exploration* 215: 106545.

Halogens in soils

Svensson Teresia^a, Thiry Yves^b, Bueno Maïté^c, and Oelmann Yvonne^d, ^aDepartment of Thematic Studies—Environmental Change, Linköping University, Linköping, Sweden; ^bAndra, Research and Development Division, Châtenay-Malabry, France; ^cUniversite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, Institut des Sciences Analytiques et de Physico-Chimie pour l'Environnement et les Matériaux, UMR 5254, Pau, France; ^dGeoecology, University of Tübingen, Tübingen, Germany

© 2023 Elsevier Ltd. All rights reserved.

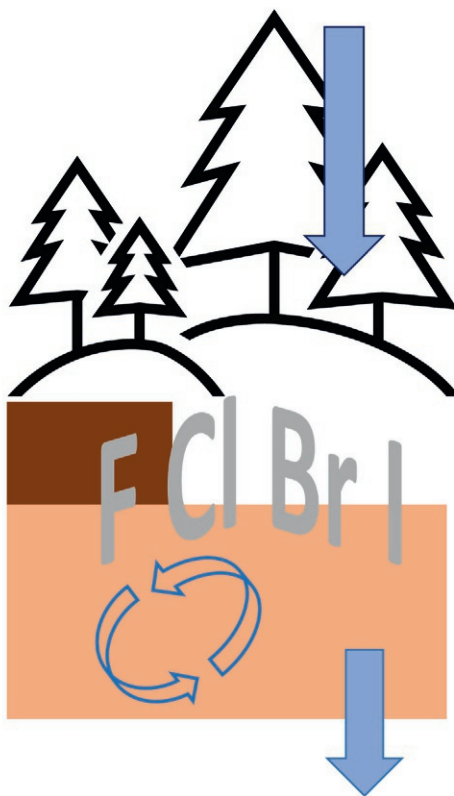
Introduction	235
Fluorine in soils	236
Chlorine in soils	237
Bromine in soils	239
Iodine in soils	239
Lessons learned from comparison of the soil halogens	241
References	241

Abstract

Among all halogens fluorine (F) and chlorine (Cl) have the highest concentrations in soil followed by bromine (Br) and iodine (I). The variability of halogen contents in soils is high, which depends to a major degree on atmospheric deposition (Cl, Br, I), parent material (F), organic matter (Cl, Br, I), secondary minerals (F, Br), and local pollution (F, Cl). The enzyme-mediated reaction between halide ions and organic matter known as halogenation is considerable in some soils for Cl, Br and I, and should be accounted for in future ecosystem halogen budgets.

Key points

- Br, Cl and I act similarly in soil through the coexistence of two main pools: inorganic and organic. However, each pool seems regulated by cycles operating on different timescales.
- The main pool of fluorine is minerals in soil. The loss of dissolved F from soil is constrained by the strong sorption to pedogenic oxides probably via inner-sphere substitution reactions between OH^- and F^- .
- Chloride dominates import and export from soil while organically bound Cl can dominate the standing stock Cl.
- Among the four halogens, bromine takes an intermediate position in terms of mineral and organic pools, and also of the fluxes of sorption, halogenation and leaching.
- The strong biophile nature of iodine promotes its association with soil organic matter compared with the other halogens.



Introduction

The halogens are a family of elements in group 17 (VIIA) of the periodic table, consisting of six elements: fluorine (F), chlorine (Cl), bromine (Br), iodine (I), astatine (At), and tennessine (Ts). Out of all halogens, there are no stable (non-radioactive) isotopes of At and Ts and therefore, At and Ts are not present in soils. The name 'halogen' was proposed by Jöns Jakob Berzelius in 1842 based upon the Greek words 'hal' (salts) and 'gen' ("to produce"). The halogen atoms have seven electrons in their outer shells, so that the addition of only one electron gives the corresponding anions (halide ions) a stable electron configuration and an oxidation number of -1 (Table 1). The specific chemical properties of halogens are often referred to as reactive, and thus they are never found in their monatomic form, but rather as part of other compounds. Therefore, all halogens are oxidizing elements and F, followed by Cl, is the most electronegative.

Halogen geochemistry has been well documented in general (Fuge, 2019; Schlesinger et al., 2020) (F), (Graedel and Keene, 1996) (Cl), (Fuge and Johnson, 1986) (I). Contrary to old reviews stating that halogens are more or less inert in soils, i.e., few halogens interact in any biological processes, it is now clear that several biogeochemical processes control the fate of the halogens (Svensson et al., 2021) (Cl), (Fuge and Johnson, 2015) (I), (Epp et al., 2020) (Br).

There is scattered information on the abundance of halogens in soils, of which F and Cl are the most reported. Average contents in soil are approximately 321 and 300 mg kg^{-1} for F and Cl, respectively, and 10 and 2.8 mg kg^{-1} for Br and I, respectively

Table 1 Selected properties of halogens.

Property	F (Fluorine)	Cl (Chlorine)	Br (Bromine)	I (Iodine)
Atomic number	9	17	35	53
Atomic mass	18.99	35.42	79.90	126.90
Atomic radius	57	97	122	132
Valence	-1	-1	-1	-1
Electronegativity (eV)	3.9	3.0	2.8	2.5

(Kabata-Pendias, 2011). The variability of halogen contents in soils is considerable and is to a major part due to distance to the coast, parent material, organic matter, secondary minerals, and local pollution. The most commonly used methods for estimating soil contents rely on measuring the soluble inorganic forms or the total content of halogens only, while ignoring the transformation of halogens between inorganic and organic forms which is considerable in some soils for Cl, Br and I.

Fluorine in soils mostly originates from natural processes such as atmospheric deposition and weathering resulting in their release from F-bearing minerals in rocks (Fuge, 2019; Schlesinger et al., 2020). Nevertheless, atmospheric fluxes of F are dominated by anthropogenic sources (55%), such as brick manufacturing, and the mining of P-containing rocks for phosphate fertilizer production. Anthropogenic atmospheric fluxes have tripled since preindustrial time (3.2–9.7 Tg F yr⁻¹; Schlesinger et al., 2020). Though the uptake of F by plants is reported to be small, it is important to monitor the concentrations of fluorinated compounds in plants and groundwater to be able to maintain safe levels for human and livestock intake. As the fluxes to soils are dominated by anthropogenic sources, the main interest in F for soils has been on pollution control and remediation (environmental chemistry, food production).

Chlorine is in general as common as F in soils and its major source in soils is by atmospheric deposition. Thus, there is a strong gradient of decreasing soil Cl concentrations from coast to inland. The general view is that chloride is the dominant form of Cl in ecosystems, but in soils, about 50% could be organically bound, and the storage of Cl in the pedosphere has been estimated to contain 48 Pg Cl (Graedel and Keene, 1996; Svensson et al., 2021). Chlorine in soils has been of interest from multiple perspectives, including the use of chloride as a tracer for sub-surface water flow and contaminant transport (hydrology), organochlorine pollutants and salinization (environmental chemistry), Cl cycling (biogeochemistry, soil sciences), radioactive waste (radioecology; ³⁶Cl is of great concern), and biology (plant science; Cl as an essential element for functions of living plants). During the past decades there has been a rapid development towards improved understanding of the terrestrial Cl cycle (Lovett et al., 2005; Bastviken et al., 2009; Tanaka and Thiry, 2020).

Bromine has received little attention compared to F and Cl due to its small background concentrations in soil. For a long time, Br had been assumed to sorb weakly to soil particles and thus to be less reactive than F and Cl. This conservative behavior enabled its application as a tracer (Levy and Chambers, 1987). However, the assumption of its conservative behavior was challenged because, for example, the correlation of Br and organic matter suggests a retention mechanism in soil (Leri and Myneni, 2012). As a consequence of the artificial synthesis, organobromine compounds have been released into the environment. Thus, a major focus has been on anthropogenic sources although natural bromination processes are also occurring (Leri and Myneni, 2012). Whereas Cl, I and probably F are essential elements for human and animal nutrition, no function has been identified for Br. However, in view of its ubiquitous occurrence in the biosphere, Br may well have some function in animal health and more research is needed to explore this possibility.

¹²⁷Iodine, the only stable isotope of I, is an essential element for humans and higher animals while ¹³¹I and ¹²⁹I are of potential concern since they may be contaminants from accidental nuclear release and the storage of radioactive waste, respectively. Seawater is rich in stable I which can be volatilized and transported to the atmosphere resulting in marked enrichment in near-coastal soils. The content of I in soils is thus in large part due to its atmospheric input, but also due to geological and edaphic conditions (Fuge and Johnson, 2015). Human intake of stable I is mainly from food, but some populations also obtain substantial quantities of I from drinking water. Most agricultural topsoils and irrigation water contain very small concentrations of I. Moreover, the bioavailable forms IO₃⁻ and I⁻ in soil may be limited by leaching, strong sorption or their volatilization. Thus, plant-derived dietary I may be insufficient, and deficiency still affects many people around the world. Methods to improve the uptake of I by plants and its accumulation in edible parts are still at an early development stage (e.g. Izydorczyk et al., 2021).

Fluorine in soils

Fluorine (F) is the lightest element of the halogen group and has the largest electronegativity compared to all other elements. Accordingly, -1 is the only stable oxidation state of F. It is a monoisotopic and mononuclidic element with ¹⁹F being the only isotope that is present beyond trace quantities.

In the crust, F is the most abundant halogen with average concentrations of 575 mg kg⁻¹. Because of the similarity in the ionic radii, hydroxyl groups (OH) can be readily substituted by F in the crystal lattice of silicates and apatite (fluorapatite: Ca₅(PO₄)₃F). In addition, F is present in the crystal lattice of fluorite (CaF₂) and topaz (Al₂SiO₄(F, OH)₂). Most of these F-containing minerals are sparingly soluble. Igneous rocks containing CaF₂ and shales have the highest F concentration while sedimentary rocks mostly contain <300 mg kg⁻¹ F (Fuge, 2019; Schlesinger et al., 2020). The concentrations of F in sedimentary phosphorites is two orders of magnitude higher because they contain carbonate fluorapatite (Schlesinger et al., 2020).

Globally, soils contain on average 321 mg F kg⁻¹ (Kabata-Pendias, 2011). The lowest concentrations occur in Arenosols, Podzols and Histosols and the highest in Cambisols (Kabata-Pendias, 2011). This difference can be attributed to clay and organic matter contents (Pickering, 1985). Furthermore, weathering and leaching during other or paleo climates may have influenced local soil F concentrations. The F concentrations in non-contaminated soils are commonly less in top- than in sub-soil (Epp et al., 2020; Neidhardt et al., 2022). In general, F concentrations are lower in soil than in the parent material suggesting loss of F during soil formation.

In contrast to the other halogens, the loss of F during soil formation cannot be explained by gaseous emissions. The emission of gaseous F compounds is restricted to anthropogenic manufacturing processes (see below) or natural processes involving high temperatures (Schlesinger et al., 2020). Therefore, there is no loss of gaseous F compounds from soil. The leaching of F is supposed

to be low because most primary and secondary F-bearing minerals in soil have a low solubility (Pickering, 1985). Accordingly, only ca 1% of total F in soil is water-soluble (Epp et al., 2020; Neidhardt et al., 2022) and F concentrations in the soil solution are mostly $\ll 1 \text{ mg L}^{-1}$ (Epp et al., 2020). Nevertheless, leaching was inferred from the smaller F concentrations in topsoil compared to the subsoil and, from the decrease of F concentrations in soil with increasing precipitation. Particularly in acidic soils, leaching is fostered by the formation of labile fluoro-iron (Fe(III)) or fluoro-aluminum (Al) complexes that increase the dissolution of F-bearing minerals and the desorption of F^- by the concomitant dissolution of the sorbing agents, namely Fe/Al (oxy)hydroxides (Pickering, 1985). In calcareous soils, slow rates of adsorption can result in rapid leaching. Leaching might also comprise water-soluble organofluorine compounds (Pickering, 1985).

Leaching loss from soil might be counterbalanced by external input and retention of F. Depending on the distance to the sea, marine-derived wet deposition of F might be important although the marine origin is controversial (recently reviewed by Fuge, 2019). In Norway, wet deposition of F ranges between 0.89 and $168 \text{ mg m}^{-2} \text{ yr}^{-1}$. Assuming an average F concentration of 0.02 mg L^{-1} in rainfall of remote regions (Schlesinger et al., 2020), wet deposition would result in a range of $1\text{--}30 \text{ mg F m}^{-2} \text{ yr}^{-1}$ for sites in southwestern Germany and Chile (Epp et al., 2020; Neidhardt et al., 2022). In addition, wind-blown dust from soil that forms part of the dry deposition needs to be included as an atmospheric input. However, the dry deposition flux of F to ecosystems has not yet been quantified. As F^- is the smallest ion among the halogens, it is strongly sorbed to charged surfaces in the soil matrix. In line with this, F has the fastest kinetics and the greatest sorption capacity compared with the other halide ions on model minerals (Fe and Al oxides, kaolinite) and in soil (Pickering, 1985; Chubar et al., 2005). Sorption of F^- comprises substitution reactions between OH and F in secondary minerals such as pedogenic oxides (metal (oxy)hydroxides) and clay minerals (Pickering, 1985). Metal hydroxides in soil e.g., $\text{Al}(\text{OH})_3$, and amorphous, poorly ordered hydrous oxides of Al were shown to be more important than clay minerals for such substitution reactions (Pickering, 1985). The increase in sorption of F^- with soil depth was attributed to an increase in sorption capacity and a decrease in desorbability. The fluorination of organic matter is another retention process. However, the large electronegativity of F results in a very large reduction potential ($E^0 = +3.06 \text{ V}$) and thus, constrains enzyme-catalyzed reactions (Carvalho and Oliveira, 2017). For a long time, no enzyme capable of fluorination had been known. Only recently, a natural fluorinase enzyme, the related organofluorine biosynthetic gene cluster and the fluorination mechanism were revealed in the soil-dwelling bacterium *Streptomyces cattleya* (Carvalho and Oliveira, 2017). Although more than 5000 biogenic organohalogens have been identified to date, only about 30 fluorinated organic compounds of natural origin e.g., fluoroacetate (FA), ω -fluoro-fatty acids and nucleocidin, have been discovered in the environment (Carvalho and Oliveira, 2017). Biologically produced, natural organofluorine compounds contain only one F atom (Carvalho and Oliveira, 2017). Abiotic processes linked to volcanic activity might also produce organofluorine compounds such as fluorotrichloromethane, trifluoropropene or fluorobenzene (Carvalho and Oliveira, 2017). Although the C—F bond is extremely stable (Schlesinger et al., 2020), defluorination might occur.

Anthropogenic activities influence the abundance of F in soil. For example, the manufacturing of bricks from clay-rich raw materials and the mining and processing of phosphate rock (fluorapatite deposits) used as fertilizers are the two largest sources of emission of gaseous F-containing compounds (e.g., HF and SiF_4) in the atmosphere (Pickering, 1985; Fuge, 2019; Schlesinger et al., 2020). Another emission source is Al smelters (Fuge, 2019; Schlesinger et al., 2020). Gaseous emissions of HF and SiF_4 affect F concentrations in soil as illustrated by greater soil F concentrations in the proximity compared to more distant locations from the Al smelters. In addition, phosphate rock contains 1.5–4 mass% fluoride and is applied as a fertilizer directly to the soil (Loganathan et al., 2001). With larger inputs, this might (Loganathan et al., 2001) or might not result in increased F concentrations in soil in the long run. For example, F applied as fertilizer was reported to be retained in topsoil (Loganathan et al., 2001) but did not pose a risk for surface water contamination. Furthermore, man-made organofluorine compounds are now other ubiquitous contaminants in the environment including soils, but biodegradation of many of these compounds is unknown (Carvalho and Oliveira, 2017).

The sorption of organofluorine compounds in soil requires further research given its influence on the persistence of natural and man-made organofluorines in the environment. Because of the large anthropogenic contribution (>55%) to atmospheric F fluxes (Schlesinger et al., 2020) combined with long-distance transport of F in the atmosphere, the anthropogenic impact on the fate of F in soils means that it probably reaches remote regions in the world. In addition, the application of rock-phosphate derived fertilizers around the globe and the transport of F in particulate form from arable fields to surface waters by soil erosion (Schlesinger et al., 2020) complicates the study of the natural biogeochemical cycle of F.

Chlorine in soils

Chlorine (Cl) is one of the 20 most common elements on earth. Its atomic number is 17 and it has a very high electron affinity and electronegativity, thereby molecular Cl is a strong oxidant. Consequently, molecular Cl is rare in nature and instead the inorganic ion form, chloride, which is highly soluble in water, typically dominates in the hydrosphere and in minerals. Chlorine occurs in nature as primarily two stable isotopes, ^{35}Cl (ca 76%) and ^{37}Cl (ca 24%). In addition to these, seven radioactive isotopes exist of which ^{36}Cl has a very long half-life, 3.01×10^5 years. Chloride in soil has traditionally been associated with soil salinity and toxicity for plants, though there are many studies on Cl that suggest it is essential for life for various reasons.

The largest Cl reservoirs on the earth surface are the crust and the ocean (Graedel and Keene, 1996). Inorganic Cl (chloride) by far dominates these reservoirs. Estimates for the reservoirs are largely based on chloride concentration measurements. This assumption of a general dominance of chloride is problematic for the pedosphere because organically bound Cl has been shown to range from 11 to almost 100% of the total Cl pool in a large range of soil types (Redon et al., 2013). In deeper mineral soils >80% of the Cl seems to be organic. This means that the pedosphere Cl pool may be at least twice as large as was proposed by Graedel and Keene (1996).

Total Cl contents typically range from 20 to >1000 mg Cl kg⁻¹ in non-saline soils (Svensson et al., 2021). The concentrations of organically bound Cl in surface soil layers are greater than chloride levels in most cases. The dry weight fraction of organically bound Cl in surface soils (0.01–0.5%) is comparable to naturally occurring concentrations of several important nutrients, including phosphorus (0.03–0.2%), nitrogen (1–5%) and sulfur (0.1–1.5%). In forests, the total storage of organically bound Cl is usually larger in the mineral soil layer than on the forest floor because of smaller Cl concentrations in mineral soil, but larger thickness and bulk density of mineral soil (Redon et al., 2013). The available data, primarily from dry forest soils but also a few agricultural soils, indicate that total Cl concentrations are highest in soils with more organic matter and the percentage Cl_{org} is frequently above 80% in humus (Svensson et al., 2021 and references therein).

It should be emphasized that the available information is based on a limited number of studies and general patterns as well as the chemical complexity of the organically bound Cl are scarce. In a sandy loam soil (organic matter content of approximately 4%), the organically bound Cl was associated with organic matter with a molecular mass <10,000 Da, while most organic matter in the soil studied had a molecular mass >10,000 Da. The results suggest that most of the newly formed chlorinated organic matter in the soil studied was of low molecular weight.

More than 5000 chlorinated organic substances have been identified (Gribble, 2015). Naturally produced organically bound Cl is a diverse group, which contains simple alkane compounds, such as chloromethane, but also more complex forms as a result of, for instance, decomposition of plant material. Many of these have known anti-competitor functions, used by organisms to secure resources, and as antibiotics. For most of the compounds, the ecological function remains unknown. In addition, the concentrations of organically bound Cl in most soils are typically as large as, or exceed, the amounts of chloride, which indicates the likelihood of additional ecological functions for organically bound Cl.

Most of the studies on Cl biogeochemical cycling have hitherto focused on bulk soil in which plant roots were removed by sieving, even though vegetation and associated organisms have a strong influence on element turnover. Increased availability of labile organic matter and plant roots seems to increase the rates of formation of organically bound Cl. This indicates that increased rates of chlorination are to be expected when more organic matter is recycled and corresponds to the idea that chlorination is controlled by biotic activity and is highest in the surface soil layers with greater root density and more input of fresh organic matter.

A consistent pattern has shown that the concentration of organically bound Cl is higher in soils under coniferous forest than those covered by deciduous forest. In experimental forests, plots with Norway spruce (*Picea abies* (L.) H. Karst.) had the highest net accumulation of chloride and organically bound Cl over the experimental period of 30 years and showed 7- and 9-times greater storage of chloride and organically bound Cl, respectively, in the soil humus layer compared to plots with oak (*Quercus petraea* (Matt.) Liebl.) (Montelius et al., 2015). Similarly, Neidhardt et al. (2022) found the largest total Cl concentrations in the mineral soil of a coniferous forest comprising *Araucaria araucana* (Molina) K. Koch in Chile. Thus, it seems likely that vegetation characteristics may be one important factor to explain soil Cl concentrations and why the spatial variation may be independent of atmospheric deposition.

Within soils, many dynamic processes affect the cycling and leaching of Cl. These processes are primarily controlled by precipitation, but are also affected by vegetation, gaseous emission from the soil, and biochemical transformation in soil. It is generally assumed that most of the deposited chloride is transported through soil unaffected, with only a minor influence of soil adsorption by e.g. ion-exchange processes. Traditionally chloride is thought to be bound only electrostatically and therefore the binding is easily reversible, which results in rapid equilibration with the soil solution. As mentioned previously, chloride is easily dissolved in water and could percolate to groundwater and surface waters. However, significant percentages of organically bound Cl, 16–54% of the total Cl leached from topsoil. Though the estimates are based on laboratory studies, considerable Cl is probably transported as organically bound from topsoil. However, on a catchment scale the annual transport of Cl is dominated by chloride (Svensson et al., 2021). Using a modeling approach, Tanaka and Thiry (2020) estimated that fluxes in the annual organic cycle leading to net retention in soil concerned only small amounts of Cl circulating in the system. This also explains why imbalance between inflow and outflow from Cl transformation in soil is not easy to detect at large scales.

The output of approximately 2% of chloride from a small stream at Hubbard Brook experimental forest (NY, USA) has been estimated to originate from weathering bedrock consisting mainly of granite; this amount is small compared with the atmospheric contribution (Lovett et al., 2005). Montelius et al. (2015) measured the mineral-bound Cl in soil developed from granite, collected in Breuil-Chenue (Bourgogne, France); it was less than 12% of the total soil Cl.

There are indications that volatilization of Cl could be a significant export pathway of Cl from the soils. Volatile organochlorines (VOCl_s) are produced in a wide variety of ecosystems such as wetlands, salt marshes and forests and in different climatic regions. Budget and transport estimates on various scales are highly uncertain, partly because the small concentrations of each specific VOCl make sampling and analyzes challenging. Emissions are considered to be small compared to the wet and dry deposition of Cl. However, a ³⁶Cl radiotracer study indirectly indicated the substantial release of VOCl_s in soils corresponding to 0.18 g Cl m⁻² yr⁻¹ or 44% of the local annual wet deposition (Bastviken et al., 2009). This is a high number that needs validation, but interestingly it includes all possible VOCl_s in contrast to other estimates that measure specific VOCl compounds only. Previous studies showed average emissions of chloroform and chloromethane corresponding to 0.13 and 0.04 g Cl m⁻² yr⁻¹, respectively, from a coniferous forest soil (Dimmer et al., 2001). This suggests that the formation of VOCl would represent a significant part of Cl transformation in soil.

Studies in terrestrial environments have revealed that Cl is very reactive, which contradicts the previous view that chloride deposited on soil has a short residence time there (no interaction with soil minerals, microorganisms or roots and vegetation), and is quickly flushed to groundwater and eventually into the oceans. It is now evident that there is ubiquitous and extensive natural chlorination of organic matter in terrestrial ecosystems. Experimental research with radioactive ³⁶Cl as a tracer has confirmed extensive natural rates of chlorination corresponding to as much as 50–300% of the annual wet deposition of Cl in several types of

soils (Bastviken et al., 2009). Many organisms display an active metabolism involving halogens (Gribble, 2015), but the fundamental reasons for the halogenation are still unclear. Several alternative hypotheses have been suggested for chlorination processes including allelopathic processes, e.g., (i) defense against pathogens/competitors/parasites, (ii) formation of HOCl could either be a defense against reactive oxygen species that are formed as by-products in cellular metabolism, (iii) or a way for microorganisms to access refractory organic matter. Many of the haloperoxidase producing organisms are microorganisms associated with plants and animals (e.g., rhizosphere associated microorganisms, endophytes/endosymbionts and pathogens/parasites). Thus, dominating tree species, soil organic matter (SOM) content, humus type and soil pH seem to be important for the accumulation of Cl in soils. The new views of Cl have implications for research on Cl storage and dynamics in terrestrial ecosystems and landscapes, raising issues as to what is controlling the processes and what ecological functions are involved.

Bromine in soils

In the environment, bromine (Br) is mainly present in the most stable oxidation state of -1 but other oxidation states also exist (0 , $+1$, $+3$, $+5$ and $+7$). Apart from several radioisotopes, Br has two stable isotopes (^{79}Br , ^{81}Br). In the Earth's crust, Br is present in the crystal lattice of minerals of the apatite group, biotites and amphiboles at concentrations ranging from 0.6 to 15 mg kg^{-1} (mean: 1 mg kg^{-1} ; Kabata-Pendias, 2011). Probably because of adsorption to charged mineral surfaces, sedimentary rocks contain more Br than igneous rocks, though extrusive igneous rocks might also have large concentrations.

Globally, soils contain on average 10 mg Br kg^{-1} (Kabata-Pendias, 2011). There are a few compilations of Br concentrations in soil that cover different lithologies or climate zones (Gerzabek et al., 1999; Yamasaki et al., 2015; Neidhardt et al., 2022). For topsoils under agricultural use in Austria, Gerzabek et al. (1999) found a range of 2 – 12 mg Br kg^{-1} (mean 6 mg kg^{-1}) and soils on calcareous parent material had larger Br concentrations than those on non-calcareous parent material. Similarly, large Br concentrations were reported for soils developed on extrusive igneous rocks (Yamasaki et al., 2015). Br concentrations in Andosols (mean 37 mg kg^{-1}) were at least three times higher than for other soil types (range of means: 0.5 – 13 mg kg^{-1} for Podzols, Cambisols, Acrisols, Gleysols and Regosols; Yamasaki et al., 2015). Soils developed on intrusive igneous rock along a climate gradient from the arid Atacama Desert to the (super) humid temperate zone in Chile showed a range in Br concentrations between 2 and 64 mg kg^{-1} (Neidhardt et al., 2022). By far the largest Br concentrations were found in the humid zone (mean of 43 mg kg^{-1}) where soil formation was most advanced (soil type: Umbrisol; Neidhardt et al., 2022). Topsoils had larger Br concentrations than subsoils (23 and 6 mg kg^{-1} , respectively; Yamasaki et al., 2015). However, this pattern does not seem to be universally valid since the opposite was also observed (Neidhardt et al., 2022). In general, Br contents are larger in soil than in the parent material suggesting retention of Br during soil formation.

Soil enrichment in Br seems to be a combination of different biogeochemical processes comprising atmospheric deposition of Br, sorption of Br to charged surfaces of the soil matrix and bromination of organic matter. Bromide ions (Br^-) are the main component of atmospheric deposition of Br (Biester et al., 2004) ranging from 3 (Gilfedder et al., 2011; Epp et al., 2020) to $1.56\text{ g Br m}^{-2}\text{ yr}^{-1}$ (Biester et al., 2004). Organobromine species were found in rain and snow, but their contribution is commonly less than one third of Br in precipitation. After deposition, Br^- might be sorbed to charged surfaces in the soil matrix. The sorption capacity tended to be larger for Br^- than for Cl^- (Chubar et al., 2005) while similar sorption of Br^- and Cl^- was also reported. Bromide might also serve as a reaction educt for bromination of SOM which was shown to be favored over other halogenation reactions (Leri and Myneni, 2012). Similarly, Biester et al. (2004) found that Br enrichment was largest during peat decomposition among the halogens studied. Bromination can be catalyzed by the same enzymes as those for chlorination (e.g., chloroperoxidase; Leri and Myneni, 2012), but can also result from abiotic processes (Leri and Ravel, 2015). The majority of organobromine species in soil are monobrominated (Tang et al., 2022). Bromination might be the main factor for retention of Br in soils because organic matter concentrations are strongly correlated with Br concentrations in Japanese soils (Yamasaki et al., 2015), and organobromine species dominated over inorganic Br species in Andosol samples (Takeda et al., 2018).

Bromine can be released from soil by the emission of gaseous Br species, or the leaching of Br. Plants and soil microorganisms are known to emit methylbromide (CH_3Br) (Horst et al., 2014). Abiotic formation of CH_3Br was also suggested (Horst et al., 2014). The leaching of Br has not yet been systematically quantified. Assuming that about half of the precipitation is lost as seepage fluxes, Br^- leaching would amount to $15\text{ mg m}^{-2}\text{ yr}^{-1}$ in the temperate forest studied by Epp et al. (2020). The output of Br^- would be three times larger than the atmospheric deposition of Br^- at the site investigated (Epp et al., 2020).

Compared to the other halogens, fewer studies are available on the fate of Br in soil. Therefore, basic knowledge gaps exist regarding sorption mechanisms of Br^- , production of organically bound Br, and the release of dissolved organic Br. Similarly, we are not aware of any study that set up Br budgets to evaluate the accumulation versus the loss of Br in soil. Furthermore, the role of redox dynamics that potentially influence the speciation and mobility of Br has not been fully explored.

Iodine in soils

With an atomic number of 53 and an atomic mass of 127 , iodine (I) is the heaviest of the stable halogens. It has several main oxidation states (-1 , $+1$, $+3$, $+5$ and $+7$), and only one naturally occurring stable isotope (^{127}I). The admission of I into the crystal lattice of rock-forming minerals is hampered by the large size ($2.20\text{ \AA}$) of its anionic radius. That explains why I is less abundant than the other halogen elements in the Earth's crust, its average concentration is estimated to be about 0.3 mg kg^{-1}

(Fuge and Johnson, 2015) and 1.4 mg kg^{-1} for the upper crust, where sedimentary rocks commonly contain larger amounts than metamorphic and igneous rocks.

From a database of over 1500 soil samples collected at the European continent scale (De Vos et al., 2006), the calculated median I content is 3.9 mg kg^{-1} in topsoil (0–25 cm) and 3.4 mg kg^{-1} in subsoil (a 25-cm thick section sampled from 50 to 200-cm deep); values range from <2 to 70.8 mg kg^{-1} in topsoils and up to 92.5 mg kg^{-1} in subsoils. For French forest soils (50 sites), Roulier et al. (2019) reported a median concentration of 4.2 mg kg^{-1} , a value slightly higher than the median European concentration of 3.9 mg kg^{-1} for topsoil, but the concentration range (0.4 – 35.7 mg kg^{-1}) was quite similar. Of note is that in general median values for soils are much higher than the average of 1.4 mg kg^{-1} for the upper crust, suggesting that I enrichment in most soils does not originate from weathering of the rocks on which they developed but rather from retention of atmospheric deposits. Though, at local scale in regions characterized by limestone and chalk terrains, calcareous parent-rock represents a key reservoir of weathered I for soil or drainage water enrichment.

Iodine is a biophilic element, therefore, the behavior of both stable and radioactive isotopes in the environment is of great concern (Thiry et al., 2022; Neeway et al., 2019). The understanding of I behavior in terrestrial environments is complicated by intricate microbial, photochemical, physicochemical, and organic interactions, in vegetation and soil, leading to its release through volatilization and drainage, or to its accumulation through SOM iodination or sorption by various biological and mineral components; see Fuge and Johnson (2015), Kaplan et al. (2014) and Neeway et al. (2019) for recent reviews, and references therein.

The properties that most affect I retention in soils involve soil organic matter (SOM), Fe–Al oxides and pH (Fuge and Johnson, 2015). Iodine incorporation into SOM by the covalent aromatic C–I bond, usually at aromatic carbon sites, is a central mechanism controlling I retention; and this transformation may greatly limit the transport of I in soil on the time scale of SOM persistence. The adsorption of I anions by bonding to localized positive charges on the surface of metal oxides is likely to be higher at low pH.

Soil fractionation and extraction help in dividing soil I into operationally defined pools of different mobility. However, controls of individual soil properties on I retention in soils remain difficult to determine. For example, Roulier et al. (2019) reported larger I concentrations in soils with higher organic matter, total iron and aluminum concentrations, and with these three soil constituents being correlated. That observation prevented the identification of soil constituents that had the greatest statistical influence on the I concentrations in soils. This result is not surprising since efficient sorption of organic matter on metal oxides surfaces is well known; moreover, soils with higher organic matter content also tend to have lower pH. Inorganic I sorption onto oxide phases would have a minor impact on the rate of I incorporation into soil humic materials. The semi-mechanistic model proposed by Chang et al. (2014) supported the assertion that the organic I fraction is an important ultimate reservoir in soil; the I uptake coefficient for SOM is an order-of-magnitude greater than that for mineral surfaces.

The reaction of I with SOM resulting in the formation of non-soluble organically bound I is quite rapid and largely influenced by microbial activity. Iodine accumulation in the non-labile pool of a forest floor is believed to be controlled by a process of sorptive accumulation in addition to organic matter iodination, since soluble organic I represent a major fraction in rain and throughfall (Pisarek et al., 2022 and references therein). In addition to iodide (I^-), dissolved organic I also represents a major form of I in soil solution, its vertical mobility in the soil column is further influenced by the solubility and partitioning of organic carbon. Both the addition of I at the soil surface and its interaction with SOM normally lead to a gradual decrease of I content with soil depth that is concomitant with the organic matter content. In some cases, a peak of I can occur at soil depths of 15–45 cm. Pisarek et al. (2022) ascribed that peak to the occurrence of an I retention front resulting from the reaction of dissolved organoiodine with a layer enriched in secondary Al precipitates that are common in acid soils.

The sequestration of soil-bound organoiodine is only transitional; its remobilization, notably in the organic colloidal form, largely depends on the local SOM turnover. In addition, microbial deiodination of organoiodine may occur naturally in soil, usually under anoxic conditions. Overall, there is still a lack of information on the mechanisms and kinetics controlling the degradation and release of soil-bound (i.e. non-labile) organoiodine. It is not clear what influence microbes, redox processes, or the nature of the SOM have on long-term I retention in the soil column (Kaplan et al., 2014). Based on a modeling approach for forest sites, Thiry et al. (2022) proposed that, under steady-state conditions, the non-labile I cycle in soil can lead to the storage of about 20% of the total atmospheric deposits, with a mean residence time of 900 years, while labile I is recycled with a mean residence time of 90 years. In that context, the production of volatile I from soil appeared to be a more significant export pathway than drainage of soluble I for most sites.

Both plants and soils are potentially important terrestrial sources of atmospheric methyl iodide (CH_3I) (Sive et al., 2007; Fuge and Johnson, 2015 and references therein). Bacteria seem to play a principal role in volatilization of I from aerated soils and especially in the rhizosphere, although the formation of various volatile organoiodides including CH_3I is also possible through abiotic processes. Knowledge of the rate of I emission from soil remains limited. A few experimental studies using different soil–plant systems reported a possible volatilization loss of 0.01–5% of radioiodine input. A mean release of $0.16 \text{ mg I m}^{-2} \text{ yr}^{-1}$ from a forest soil has also been measured at Duke Forest, USA (Sive et al., 2007). That flux was consistent with other measurements made at two coastal coniferous bog sites in Ireland.

In general, developing a budget including volatile I in soil is complicated by the poor understanding of the spatial and temporal variability in its production, consumption, and diffusion within the soil column. While rainfall transfers large amounts of I with an obvious atmospheric marine source in coastal soils, it is believed that ‘land hopping’ involving volatilization from the terrestrial environment and further re-deposition may greatly influence I inputs to inland soils. In any geographic situation, variation in local edaphic conditions is expected to influence the capacity of soil greatly to retain the atmospheric inputs of I, and finally the spatial variability of I content in soils.

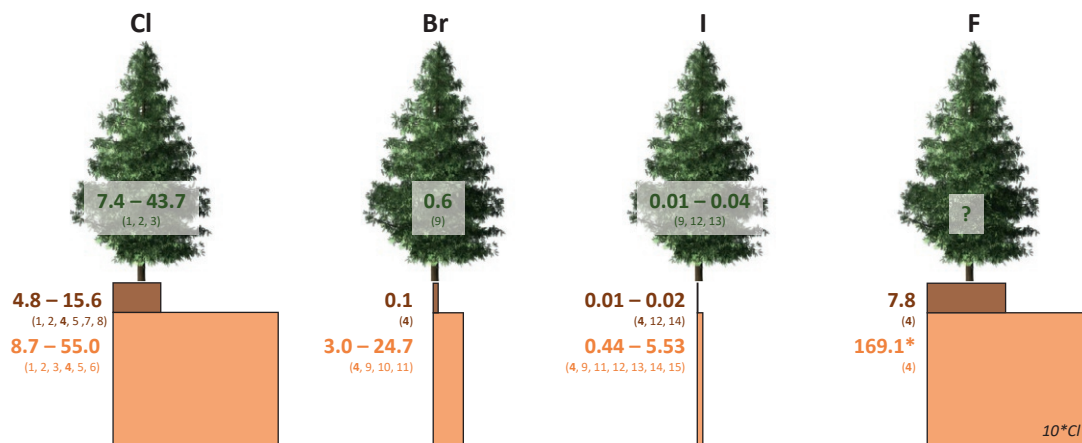


Fig. 1 Total halogen stocks in foliar biomass (mol ha⁻¹), organic layers (kmol ha⁻¹) and mineral soil (kmol ha⁻¹). 1: Montelius et al. (2015), 2: Svensson et al. (2021), 3: Öberg et al. (2005), 4: Epp et al. (2020), 5: Redon et al. (2013), 6: Redon et al. (2013), 7: Johansson et al. (2001), 8: Zlamal et al. (2017), 9: Yuita, (1983), 10: Cortizas et al. (2016), 11: Takeda et al. (2018), 12: Roulier et al. (2019), 13: Bowley et al. (2019), 14: Roulier et al. (2019). Note that these publications include a wide variety of soil types (e.g., Cambisols, Umbrisols, Andosols). Because variables required to calculate the mass of each compartment were published in two studies only, we used these for all other studies as well (reference no. 12: foliar biomass: 5100 kg ha⁻¹; reference no. 4: organic layer: 10.5 cm thickness, 0.7 g cm⁻³ bulk density; mineral soil: 60 cm thickness, 1.3 g cm⁻³ bulk density). The resulting masses were then multiplied by the published total halogen concentrations. The box width reflects the stock size. * Note that the box width of F stocks in mineral soil could not be displayed proportionally: the width is the same as for Cl stocks, but F stocks are 10 times Cl stocks.

Lessons learned from comparison of the soil halogens

In most soils, a large fraction of chlorine, bromine and iodine derives from atmospheric deposits, which is mirrored in their large soil contents in areas close to the oceans, whereas soil fluorine largely depends on concentration in its parent material. When comparing the halogens on the forest floor, fluorine and chlorine have the largest pools (Fig. 1) primarily due to their greater atmospheric deposition, and because SOM halogenation is particularly active in the organic surface layer (Svensson et al., 2021). Mean wet atmospheric deposition of Cl and I in mol ha⁻¹ yr⁻¹ (96–2208 mol ha⁻¹ yr⁻¹) for chlorine (Redon et al., 2013) and 0.06–0.41 mol ha⁻¹ yr⁻¹ for iodine (Roulier et al., 2019) in French forested areas is about 1000–5000 times less for iodine compared with chlorine. This difference is reduced to about 10 times in soil because of the strong biophilic nature of iodine which promotes its association with SOM compared with chlorine. It is likely that SOM tends to be enriched in bromine, chlorine and iodine through similar halogenation processes. Although SOM chlorination is a well-known source of chlorine accumulation in soil and fundamentally affects long-term Cl dynamics (Svensson et al., 2021), a major part of chlorine (Cl⁻ ions) is highly soluble and rapidly leached from soils. The modeled organic Cl cycle in forest soils showed an accumulation of <1% of the annual atmospheric deposition of Cl in the form of organic chlorine (Tanaka and Thiry, 2020), while the stabilization of organically bound atmospheric I input was about 20% (Thiry et al., 2022). Interestingly, the residence time of both organic pools in soil is similar, but the major export pathways are different. Though the estimates of Cl volatilization are large, drainage of Cl would probably be a relatively more significant export pathway than volatilization while volatile I production from soil would represent the major losses for most sites.

The comparison of the fate of the different halogens in soil is constrained by a predominance of halogen-specific studies with different methodologies, whereas a few studies only measured several halogens at the same site (e.g., Suess et al., 2019). Therefore, multi-halogen studies are essential and should include sources, speciation, measurements of soil processes, volatilization and multi-isotope approaches. Previous studies of halogens have revealed differences between soil types, and local variation within sample plots that can be being partly explained by environmental variables. Thus, knowledge on the variation of the stocks and fluxes in soil of different ecosystems are of fundamental importance for future knowledge on halogens biogeochemistry.

References

- Bastviken D, Svensson T, Karlsson S, Sandén P, and Öberg G (2009) Temperature sensitivity indicates that chlorination of organic matter in forest soil is primarily biotic. *Environmental Science & Technology* 43: 3569–3573.
- Biester H, Keppler F, Putschew A, Martinez-Cortizas A, and Petri M (2004) Halogen retention, organohalogens, and the role of organic matter decomposition on halogen enrichment in two Chilean peat bogs. *Environmental Science & Technology* 38: 1984–1991.
- Bowley HE, Young SD, Ander EL, Crout NMJ, Watts MJ, and Bailey EH (2019) Iodine bioavailability in acidic soils of Northern Ireland. *Geoderma* 348: 97–106. <https://doi.org/10.1016/j.geoderma.2019.04.020>.
- Carvalho MF and Oliveira RS (2017) Natural production of fluorinated compounds and biotechnological prospects of the fluorinase enzyme. *Critical Reviews in Biotechnology* 37: 880–897.

- Chang HS, Xu C, Schwehr KA, et al. (2014) Model of radioiodine speciation and partitioning in organic-rich and organic-poor soils from the Savannah River Site. *Journal of Environmental Chemical Engineering* 2: 1321–1330.
- Chubar NI, Samanidou VF, Kouts VS, et al. (2005) Adsorption of fluoride, chloride, bromide, and bromate ions on a novel ion exchanger. *Journal of Colloid and Interface Science* 291: 67–74.
- Cortizas AM, Vazquez CF, Kaal J, Biester H, Casais MC, Rodriguez TT, and Lado LR (2016) Bromine accumulation in acidic black colluvial soils. *Geochim Cosmochim Acta* 174: 143–155. <https://doi.org/10.1016/j.gca.2015.11.013>.
- De Vos W, Tarvainen T, Salminen R, et al. (2006) *Geochemical Atlas of Europe. Part 2. Interpretation of Geochemical Maps, Additional Tables, Figures, Maps and Related Publications*. Espoo: Geological Survey of Finland.
- Dimmer CH, Simmonds PG, Nickless G, and Bassford MR (2001) Biogenic fluxes of halomethanes from Irish peatland ecosystems. *Atmospheric Environment* 35: 321–330.
- Epp T, Neidhardt H, Pagano N, Marks MA, Markl G, et al. (2020) Vegetation canopy effects on total and dissolved Cl, Br, F and I concentrations in soil and their fate along the hydrological flow path. *Science of the Total Environment* 712: 135473.
- Fuge R (2019) Fluorine in the environment, a review of its sources and geochemistry. *Applied Geochemistry* 100: 393–406.
- Fuge R and Johnson CC (1986) The geochemistry of iodine—A review. *Environmental Geochemistry and Health* 8: 31–54.
- Fuge R and Johnson CC (2015) Iodine and human health, the role of environmental geochemistry and diet, a review. *Applied Geochemistry* 63: 282–302.
- Gerzabek MH, Muramatsu Y, Strebl F, and Yoshida S (1999) Iodine and bromine contents of some Austrian soils and relations to soil characteristics. *Journal of Plant Nutrition and Soil Science* 162: 415–419.
- Gilfedder BS, Petri M, Wessels M, and Biester H (2011) Bromine species fluxes from Lake Constance's catchment, and a preliminary lake mass balance. *Geochimica et Cosmochimica Acta* 75: 3385–3401.
- Graedel TE and Keene WC (1996) The budget and cycle of Earth's natural chlorine. *Pure and Applied Chemistry* 68: 1689–1697.
- Gribble G (2015) A recent survey of naturally occurring organohalogen compounds. *Environmental Chemistry* 12: 396–405.
- Horst A, Holmstrand H, Andersson P, et al. (2014) Stable bromine isotopic composition of methyl bromide released from plant matter. *Geochimica et Cosmochimica Acta* 125: 86–195.
- Izydorczyk G, Ligas B, Mikula K, et al. (2021) Biofortification of edible plants with selenium and iodine—A systematic literature review. *Science of the Total Environment* 754: 141983.
- Johansson E, Ebenå G, Sandén P, Svensson T, and Öberg G (2001) Organic and inorganic chlorine in Swedish spruce forest soil: Influence of nitrogen. *Geoderma* 101: 1–13.
- Kabata-Pendias A (2011) *Trace Elements in Soils and Plants*. Boca Raton: Taylor & Francis.
- Kaplan DI, Denham ME, Zhang S, et al. (2014) Radioiodine biogeochemistry and prevalence in groundwater. *Critical Reviews in Environmental Science and Technology* 44: 2287–2335.
- Leri AC and Myneni SCB (2012) Natural organobromine in terrestrial ecosystems. *Geochimica et Cosmochimica Acta* 77: 1–10.
- Leri AC and Ravel B (2015) Abiotic bromination of soil organic matter. *Environmental Science & Technology* 49: 13350–13359.
- Levy B and Chambers R (1987) Bromide as a conservative tracer for soil-water studies. *Hydrological Processes* 1: 385–389.
- Loganathan P, Hedley MJ, Wallace GC, and Roberts AHC (2001) Fluoride accumulation in pasture forages and soils following long-term applications of phosphorus fertilisers. *Environmental Pollution* 115: 275–282.
- Lovett GM, Likens GE, Buso DC, Driscoll CT, and Bailey SW (2005) The biogeochemistry of chlorine at Hubbard Brook, New Hampshire, USA. *Biogeochemistry* 72: 191–232.
- Montelius M, Thiry Y, Marang L, et al. (2015) Experimental evidence of large changes in terrestrial chlorine cycling following altered tree species composition. *Environmental Science & Technology* 49: 4921–4928.
- Neeway JJ, Kaplan DI, Bagwell CE, et al. (2019) A review of the behavior of radioiodine in the subsurface at two DOE sites. *Science of the Total Environment* 691: 466–475.
- Neidhardt H, Lemke E, Epp T, et al. (2022) Impact of abiotic and biogeochemical processes on halogen concentrations (Cl, Br, F, I) in mineral soil along a climatic gradient. *Environmental Science: Processes & Impacts* 13.
- Öberg G, Holm M, Sandén P, Svensson T, and Parikka M (2005) The role of organic-matter-bound chlorine in the chlorine cycle: A case study of the Stubbetorp catchment, Sweden. *Biogeochemistry* 75: 241–269.
- Pickering WF (1985) The mobility of soluble fluoride in soil. *Environmental Pollution* 9: 281–308.
- Pisarek P, Bueno M, Thiry Y, et al. (2022) Influence of tree species on selenium and iodine partitioning in an experimental forest ecosystem. *Science of the Total Environment* 809: 151174.
- Redon PO, Jolivet C, Saby N, Abdelouas A, and Thiry Y (2013) Occurrence of natural organic chlorine in soils for different land uses. *Biogeochemistry* 114: 413–419.
- Roullet M, Coppin F, Bueno M, et al. (2019) Iodine budget in forest soils: Influence of environmental conditions and soil physicochemical properties. *Chemosphere* 224: 20–28.
- Schlesinger W, Klein EM, and Vengosh A (2020) Global biogeochemical cycle of fluorine. *Global Biogeochemical Cycles* 34: 18.
- Sive BC, Varner RK, Mao H, et al. (2007) A large terrestrial source of methyl iodide. *Geophysical Research Letters* 34: 17.
- Suess E, Aemissegger F, Sonke JE, et al. (2019) Marine versus continental sources of iodine and selenium in rainfall at two European high-altitude locations. *Environmental Science & Technology* 53: 1905–1917.
- Svensson T, Kylin H, Montelius M, Sandén P, and Bastviken D (2021) Chlorine cycling and the fate of Cl in terrestrial environments. *Environmental Science and Pollution Research* 28: 7691–7709.
- Takeda A, Nakao A, Yamasaki S, and Tsuchiya N (2018) Distribution and speciation of bromine and iodine in volcanic ash soil profiles. *Soil Science Society of America Journal* 82: 815–825.
- Tanaka T and Thiry Y (2020) Assessing the recycling of chlorine and its long-lived ³⁶Cl isotope in terrestrial ecosystems through dynamic modelling. *Science of the Total Environment* 700: 134482.
- Tang CM, Chen GS, Jiang B, et al. (2022) High-performance nontarget analysis of halogenated organic compounds in tap water, fly ash, soil and sediment using ultrahigh resolution mass spectrometry and scripting approaches based on Cl/Br-specific search algorithms. *Analytica Chimica Acta* 1204: 10.
- Thiry Y, Tanaka T, Bueno M, et al. (2022) Recycling and persistence of iodine 127 and 129 in forested environments: A modelling approach. *Science of the Total Environment* 831: 154901.
- Yamasaki S, Takeda A, Watanabe T, et al. (2015) Bromine and iodine in Japanese soils determined with polarizing energy dispersive X-ray fluorescence spectrometry. *Soil Science and Plant Nutrition* 61: 751–760.
- Yuita K (1983) Iodine, bromine and chlorine contents in soils and plants of Japan. *Soil Science and Plant Nutrition* 29: 403–428. <https://doi.org/10.1080/00380768.1983.10434645>.
- Zlamal J, Raab T, Little M, Edwards R, and Lipson D (2017) Biological chlorine cycling in the Arctic Coastal Plain. *Biogeochemistry* 134: 243–260.

Transformation of metals and metalloids by microorganisms

Susan M Glasauer^a, Euan P Burford^b, and Geoffrey Michael Gadd^b, ^aSchool of Environmental Sciences, University of Guelph, Guelph, ON, Canada; ^bGeomicrobiology Group, School of Life Sciences, University of Dundee, Dundee, Scotland, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction	243
Microbially-induced metal transformations	244
Mobilization	245
Acidification	245
Complexation by ligands	245
Siderophores	245
Biomethylation	245
Redox reactions	246
Immobilization	248
Biosorption	248
Precipitation of metals and minerals	248
Partitioning in the cellular environment	248
Oxalates	251
Carbonates	251
Redox reactions	251
Bacterial and fungal oxidation	251
Dissimilatory metal-reduction by bacteria	252
Biotechnological applications of microbial metal transformations	253
Ore leaching	253
Bioremediation	253
Biorecovery	254
Summary	254
Acknowledgments	254
References	254

Abstract

Microorganisms exhibit a variety of direct and indirect mechanisms that can alter the chemical speciation of metals and metalloids, and their release from minerals and other solid substrates. Such mechanisms are important components of biogeochemical cycles for metal(loid)s, and may be important determinants of microbial metal tolerance or resistance. Those mechanisms that alter metal mobility through, e.g., solubilization or bioprecipitation, are of significance in transformations of environmental pollutants, as well in environmental biotechnologies such as bioleaching, bioremediation and element biorecovery.

Key points

- To survey the main mechanisms by which prokaryotes and fungi interact with and chemically transform metals and metalloids
- To emphasize the importance of metal and mineral transformations in biogeochemical processes and in environmental biotechnology

Introduction

Fungi and prokaryotes from all major taxonomic groups are found in metal-rich habitats such as [mineral ores](#) and leachates, acidic soils, or polluted environments. The survival of these [microorganisms](#) may depend on a number of biochemical, structural, and physiological properties that modify metal [bioavailability](#) and toxicity, which are constitutive to a species and/or have been selected by adaptation.

All microorganisms have essential metabolic and structural requirements for specific metals in metabolism, including potassium (K), sodium (Na), magnesium (Mg), calcium (Ca), manganese (Mn), iron (Fe), copper (Cu), zinc (Zn), cobalt (Co), and nickel (Ni), while some metals, e.g., rubidium (Rb), caesium (Cs), aluminium (Al), cadmium (Cd), silver (Ag), gold (Au), mercury (Hg),

and lead (Pb), and **metalloids**, e.g., arsenic (As) and tellurium (Te), have no apparent essential function. Intense competition for some metals, coupled with low bioavailability, has driven the evolution of tightly regulated mechanisms for metal acquisition. These include highly specialized, metal-specific **membrane transport** systems and metal chelates to scavenge metals effectively from the environment. All metals and metalloids, regardless of their metabolic function, are, however, toxic in excess and can inhibit growth, affect **sporulation** and adhesion, induce morphological changes, and affect biochemical activities. This potential toxicity has necessitated organisms to develop diverse mechanisms of defense. Both genetically-coded mechanisms of resistance and non-specific mechanisms such as oxidation or reduction of the metals to produce less toxic chemical species, or **sorption** and precipitation on cell surfaces and exterior to cells, can protect cells from toxic metals.

The chemical form, or species, of metals and metalloids is also affected by microbial processes other than those related to nutrition and toxicity. Metal speciation is important because different metal species can have widely differing chemical and biological properties, including mobility in the environment and toxicity. Excreted cellular metabolites transform metal compounds by non-specific reactions such as dissolution by protons and organic acids. Some bacteria use metals as electron donors and acceptors during respiration, a process that is a major pathway for the cycling of metals in soils and sediments. Such microbial transformations both mobilize and precipitate metals, leading to alteration and modification of soils and sediments. Many of these processes also have potential for the **bioremediation** of metal-contaminated sites as well as the biorecovery of valuable elements (Lloyd et al., 2008; Gadd, 2010; Johnson, 2014).

Microbially-induced metal transformations

Interactions between **microorganisms** with metals and soil components take many forms: unravelling their complexity is integral to understanding metal speciation in the soil. The relative roles of fungi, archaea and bacteria depend on associated biota (other microorganisms, plants, and animals) as well as on **abiotic factors** such as physicochemical components of the environmental matrix, e.g., pH, water, inorganic and organic ions, molecules, compounds, colloids, and particulates (Sullivan and Gadd, 2019) (Fig. 1).

The ability of microorganisms to affect metal speciation in the soil environment stems from two opposing basic phenomena: mobilization and immobilization, which influence the distribution of metal species between soluble and insoluble phases. **Solubilization** of metals can be achieved by various mechanisms: protonation, chelation, redox and chemical transformations. Metal immobilization can occur by precipitation or crystallization of insoluble organic or **inorganic compounds**, or by **sorption**, uptake, and intracellular sequestration. In addition, redox reactions can mobilize or immobilize metals, depending on the environment, metal and microbial species (Banfield and Nealson, 1997; Lovley, 2000; Hernandez and Newman, 2001; Korenevsky et al., 2002; Gadd, 2010).

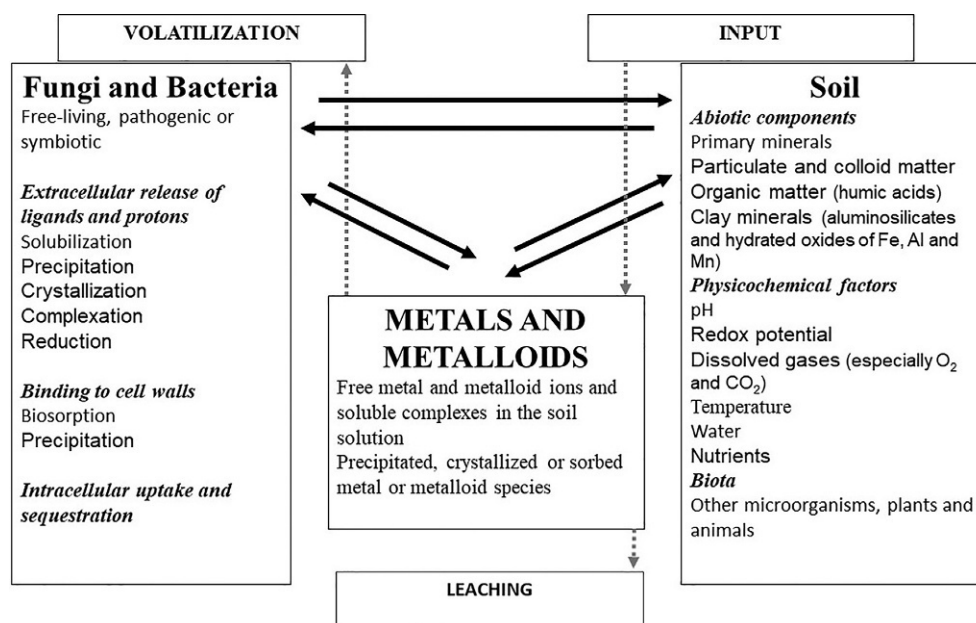


Fig. 1 Fungal and bacterial interactions with metals and metalloids in the soil environment. Metal(loid) species lost from the soil environment are indicated by broken arrows. Interactions mediated by **microorganisms** are indicated by solid arrows.

Mobilization

Fungi and bacteria can facilitate the transformation of metals into soluble forms via the dissolution of metal compounds, including oxides, phosphates, [sulfides](#), and more complex [mineral ores](#), or by desorption from exchange sites on, for example, clay minerals, colloids and organic matter. Metal leaching by Fe- and S-oxidizing bacteria is discussed further on. A number of other pathways by which [microbes](#) can leach [metal ions](#) have been proposed.

Acidification

Competition between protons and the metal (in a metal-anion complex) leads to protonation of the anion under conditions of increased acidity, and results in the release of free metal cations. Fungi and bacteria can acidify their environment in a number of ways; for example, proton [efflux](#) via [plasma membrane](#) H^+ -ATPases, maintenance of charge balance, or as a result of respiratory carbon dioxide accumulation.

Complexation by ligands

Microorganisms produce a number of extracellular metabolites that can complex metals in solution, including [polysaccharides](#), pigments, organic acids, and [siderophores](#). Organic acids are important in energy production as intermediates in the tricarboxylic (TCA) cycle and in cellular pH [homeostasis](#). They can correct internal perturbations in charge balance and pH resulting from [nutrient uptake](#) by their ability to produce and consume protons via the formation and removal of carboxylic groups, and by efflux from cells.

Citric and [oxalic acids](#) are the most commonly reported organic acids in soils and are released into the soil by fungal [hyphae](#), [lichens](#) (a symbiotic relationship between a fungus and a photosynthetic [cyanobacterium](#) and/or alga), bacteria, and plant roots and symbiotic mycorrhizas. Organic acids are believed to play an essential role in nutrient acquisition as a result of their [acidification](#) and metal-binding properties. For example, phosphate is released from common soil compounds such as phosphates of calcium, iron, and aluminium ($CaHPO_4$, $Ca_3(PO_4)_2$, $FePO_4 \cdot 4H_2O$, and $AlPO_4$, respectively) either by direct exchange of chemically binding species, or ligands, between phosphate and the [organic anion](#) or by binding of the metal to the anion. Metal complexation depends on metal-to-anion availability, the stability of the particular metal-ligand complex, and pH. Anion concentration (which will depend on the pK_a of the acid) is important, because fully protonated organic acids chelate metals poorly. The affinity depends on the stability constant of the organic acid-metal complex. For example, the efficiency of organic acids to solubilize phosphate from soils often follows the order: citrate > [oxalate](#) > malate > acetate. The [bioavailability](#) of a solubilized metal-ligand complex will depend ultimately on its long-term fate in the environment, including [sorption](#) to components of the soil matrix or potential utilization by biota. Metal citrates can be highly resistant to microbial degradation and this can result in effective leaching of metals from soil. In contrast, metal oxalates may crystallize depending on metal and oxalate concentrations, pH, and the organisms involved (Gadd, 2007; Gadd et al., 2014).

Many soil bacteria, fungi, and lichens are particularly efficient at solubilizing metal phosphates and other metal-bearing compounds, which has been attributed to their ability to produce protons and organic ligands. This can be demonstrated in simple laboratory experiments: if insoluble metal compounds are incorporated into solid growth medium, clear zones may appear around microbial colonies if they are capable of mineral dissolution.

In addition, organic acid production by microbes is an important agent of rock and mineral biodeterioration, playing a role in both biogenic chemical weathering and soil formation. Fungi can also physically weather minerals by etching and fracturing mineral grains during hyphal growth (Burford et al., 2003). Such effects are also evident in the built environment where fungi are the prevalent microbial form causing biodeterioration of rocks, minerals and mineral-based building materials.

Siderophores

Siderophores are the largest class of known compounds that can bind and transport, or shuttle, Fe. They are highly specific Fe(III) ligands with formation constants often greater than 10^{30} . These [low-molecular-weight](#) coordination molecules are excreted by a wide variety of fungi and bacteria to aid Fe assimilation. Although the mechanism could be used to acquire other metals, Fe is the only known essential element for which these specific organic shuttles operate. This is most likely because Fe is needed in larger amounts by cells than other poorly soluble metals, and, given the low solubility-product constant of ferric hydroxide (less than 10^{-38}), the concentration of free Fe is too low to support [microbial growth](#) at pH values where most life exists.

In a typical soil, many different species, including plants, compete for Fe via siderophores. Organisms have most likely evolved mechanisms to ensure that the demand for essential Fe is met through the production of species-specific siderophores, or by attachment to solid Fe-containing minerals (e.g., Fe oxides), to shorten the pathway between the Fe substrate and the site of uptake.

Biomethylation

[Methylation](#) of Hg, As, Se, Sn, Te, and Pb can be mediated by a range of microorganisms, including [clostridia](#), [methanogens](#), and [sulfate-reducing bacteria](#) under [anaerobic conditions](#), and principally fungi under [aerobic conditions](#), such as *Penicillium*,

Cephalosporium, and *Alternaria* spp. as well as a variety of bacteria, including pseudomonads. Methyl groups are enzymatically transferred to the metal, and a given species may transform a number of different metal(loid)s. Methylated metal compounds formed by these processes differ in their solubility, volatility, and toxicity. Volatile methylated species, e.g., dimethyl selenide $((\text{CH}_3)_2\text{Se})$ and dimethyl diselenide $((\text{CH}_3)_2\text{Se}_2)$, can be lost from the soil, reducing potential selenium toxicity to the biota. Methylation can also increase element toxicity; e.g., methylmercury (CH_3Hg^+) produced by bacteria in anaerobic sediments is highly toxic.

Redox reactions

Microorganisms can mobilize or immobilize metals, metalloids, and organometallic compounds by reduction and oxidation processes (Lovley, 2000; Korenevsky et al., 2002; Lloyd et al., 2008; Gadd, 2000, 2010; Johnson and du Plessis, 2015) (Fig. 2). For example, oxidation of metal-complexing dimethylsulfide, dimethylsulfoxide, or thiosulfate increases metal availability if metal sulfates are formed, and the solubilities of Fe and Mn increase upon reduction of Fe(III) to Fe(II) and Mn(IV) to Mn(II). While many microorganisms reduce small amounts of Fe during metabolism, most Fe reduction is carried out by facultative and strictly anaerobic bacteria such as *Shewanella* spp. and *Geobacter* spp. that use Fe(III) for respiration, and there is evidence that reduction of Fe(III) was one of the first forms of microbial respiration. In this process, oxidized forms of metals are used as the final electron acceptor in place of oxygen that is consumed during aerobic respiration. Dissimilatory metal-reducing bacteria respire and thereby reduce oxidized metal species at or near the outer cell membrane. The membrane-bound electron transport system provides energy to cells by cycling electrons through an electrochemical cascade of metalloenzymes, expelling electrons (e^-) and protons (H^+) from the cell. Energy is generated by the flux of H^+ into the cell through specific membrane-bound channels (Fig. 3). Dissimilatory metal-reducing bacteria can use a variety of metal(loid)s with an appropriate redox couple, including Fe(III), Mn(IV), Se(IV, VI), chromium(VI) (Cr(VI)), and uranium(VI) (U(VI)). While Fe(III) and Mn(IV) increase their solubility upon reduction, the solubility of other metals such as U(VI) to U(IV) and Cr(VI) to Cr(III) decreases, resulting in immobilization (see section “Precipitation of metals and minerals,” below). Reduction of Hg(II) to Hg(0) by bacteria and fungi results in diffusion of elemental Hg out of cells, and conferment of Hg resistance on organisms that possess this property since Hg(0) is less toxic than Hg(II).

Dissimilatory metal reduction requires that bacteria access relatively large amounts of insoluble metals. Bacteria can use redox-active organic compounds to shuttle electrons between cells and the metal being respired under anaerobic conditions. Many quinoid compounds, common components of humic substances, can function as electron shuttles. These compounds move electrons from the cell membrane to the target metal, such as Fe(III) contained in ferric oxides. The shuttle molecule regenerates when the metal is reduced and the cycle continues (Fig. 4). While the use of electron shuttle mechanisms appears to be widespread among bacteria, the impact and ecologic significance of the process in natural subsurface environments is still not fully understood.

In the absence of processes that mix water columns, sediments, and soils, redox potentials tend to decrease with depth, reflecting a change in the dominant electron acceptors from O_2 to NO_3^- , Mn(IV), Fe(III), and SO_4^{2-} . The interplay between microbial and abiotic metal reduction forms characteristic element profiles (Fig. 5). *Shewanella*, *Geobacter*, and *Desulfovibrio* spp. were among the first bacteria to be intensively studied for their metal-reducing capability, and species continue to be added to the group of bacteria

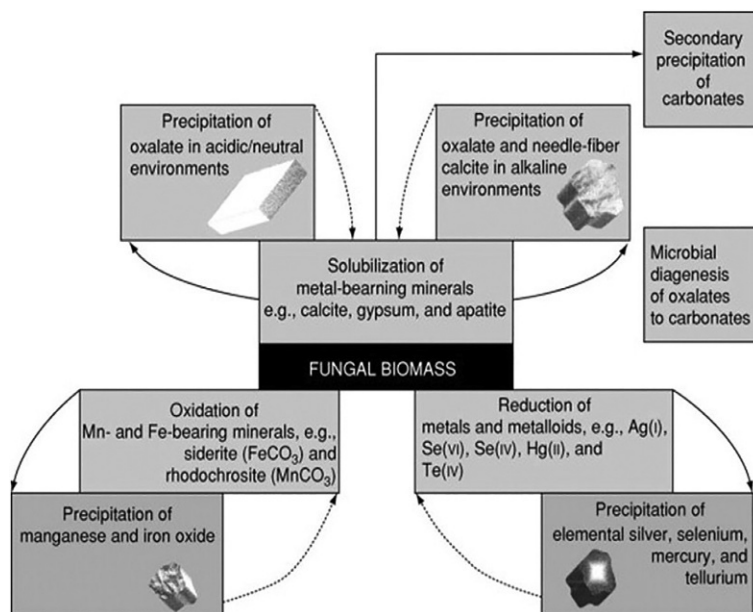


Fig. 2 Metal precipitation by microorganisms. Solid arrows represent precipitation by fungi and bacteria. Broken arrows represent precipitation and association with microbial biomass (living or dead).

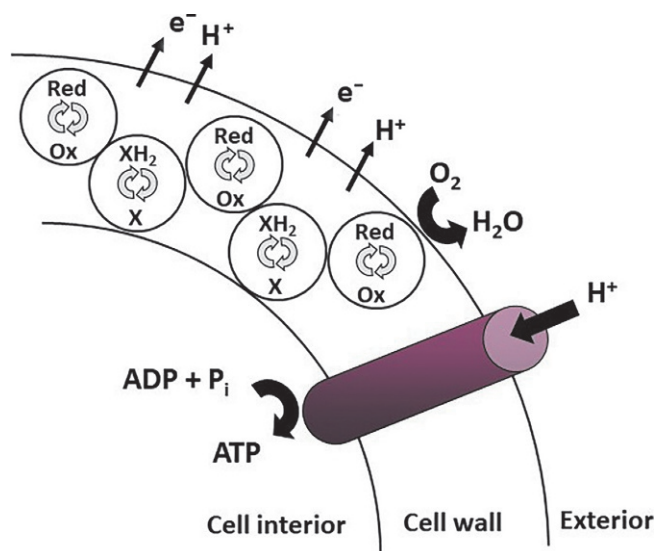


Fig. 3 Simplified bacterial electron transport chain, showing linked carriers in the cell wall that cycle electrons down an electrochemical gradient to the final electron acceptor (in this case O_2). Energy is generated by the flux of protons through specific channels across the membrane. Red, reduction; Ox, oxidation; P_i , inorganic phosphorus; ADP, adenosine diphosphate; ATP, adenosine triphosphate.

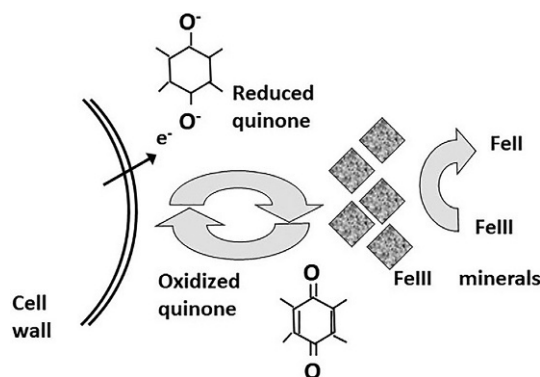


Fig. 4 Proposed mechanism for electron transfer from metal-reducing cells to oxidized Fe minerals via chemical reduction of a quinone shuttle. The shuttle regenerates when Fe(III) is reduced to Fe(II).

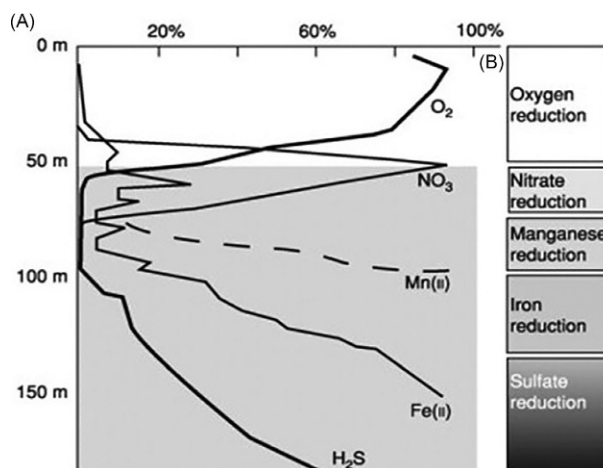


Fig. 5 Reductive processes in the environment: (A) representative electron acceptor profiles (percentage of maximum) with water depth in meters. Data are from the Black Sea; (B) major processes occurring in the redox zone with depth. Reproduced with permission from Nealson KH and Saffarini D (1994) Iron and manganese in anaerobic respiration: environmental significance, physiology and regulation. *Annual Review of Microbiology* 48: 311–343. © Annual Reviews, Inc.

with this capability. Molecular studies reveal that many as yet uncultured species exist in the subsurface which can also contribute, and improved culturing methods are likely to show novel dissimilatory metal reducers among these organisms. The poorly crystalline iron fraction is generally considered to be the 'reactive' fraction for these bacteria; however, iron in some clay minerals and crystalline Fe oxides could also be available for reduction.

Immobilization

There are a number of processes whereby microorganisms and **microbial biomass** can immobilize metals in the soil environment. Although immobilization reduces the external free metal activity, it may also shift the equilibrium to release more metal into the soil solution.

Biosorption

The strong interaction of metals with cell surfaces stems from cell wall characteristics, including component pigments and **polysaccharides**, and extracellular polymeric substances (EPS). Bacterial cells depend on the diffusion or transport of nutrients for survival, and their small size, shape, and characteristic **membrane structures** have evolved to maximize their interaction with nutrient solutes (Beveridge and Doyle, 1989; McLean et al., 2002; Korenevsky et al., 2002). There is remarkable similarity in the general format of bacterial wall structure, which falls into two broad classes based on reaction with the Gram stain, which utilizes **crystal violet**. Gram-positive walls consist of a thick matrix of **peptidoglycan** and secondary polymers, which provide structural support and give the cell its shape and form (Fig. 6). The phosphate and **carboxyl groups** within the wall fabric and exposed at the surface confer pH-dependent negative charge to the cell in the types of environments where **microbes** generally exist. The point of zero charge (PZC), or pH where positive charge equals negative charge, is approximately pH 2–3 for bacterial cell walls. This charge provides the basis for metal **sorption**. At pH values above the PZC, the net charge on the cell wall is negative, which facilitates sorption of positively charged metal species. Gram-negative cells are also negatively charged, although structurally more complex. In this case, a thin layer of peptidoglycan floats in a gel-like periplasm, sandwiched between the **plasma membrane** and the outer membrane (Fig. 6). The outer membrane has characteristic surface **lipopolysaccharides** (LPS) containing ionizable phosphoryl, carboxyl, hydroxyl, and amine groups; the phosphoryl groups are the most reactive to metals. In addition to these basic structures, surface capsules, sheaths, EPS and S-layers provide additional sites for metal interaction and the heterogeneous nucleation of minerals (Beveridge and Doyle, 1989).

Fungal cell walls are predominantly composed of polysaccharides (80–90%) with **glycoproteins** and some lipids. The polysaccharide component includes microcrystalline fibrils of β -linked polysaccharides (chitin, chitosan, and β -glucans) that are responsible for wall strength and rigidity. As for bacteria, the chemical properties of the functional groups associated with fungal walls, including carboxyl, amine, hydroxyl, phosphoryl, and sulfhydryl groups, provide the basis for the sorption of metals to cell walls. In addition, pigments such as **melanin** contribute to rigidity and have significant metal-binding properties, preventing metal entry into melanized structures and contributing to survival in the presence of potentially toxic metals as well as other stress factors like desiccation and ultraviolet (UV) irradiation (Gadd, 2007; Fomina and Gadd, 2014).

Precipitation of metals and minerals

Partitioning in the cellular environment

Metals are found in all parts of bacterial and fungal cells: in the **cytoplasm**, the cell wall, or sorbed to cell surfaces and extracellular materials (Beveridge and Doyle, 1989; Gadd, 2007, 2010; French et al., 2013). At sufficient concentrations, metal precipitates can

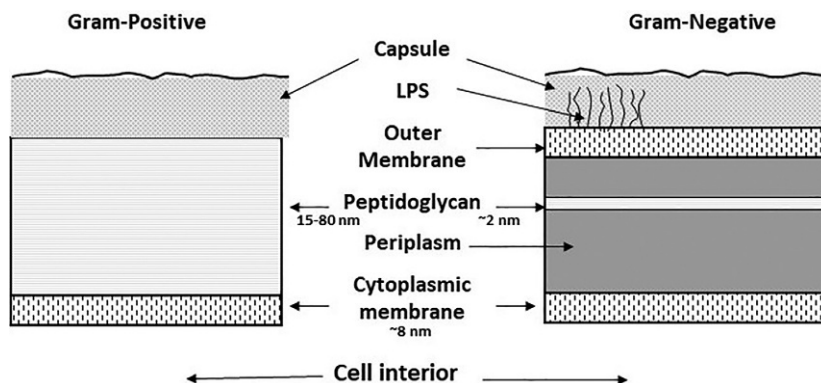


Fig. 6 Cross-sectional view showing the basic cell wall structures characteristic for Gram-positive (left) and Gram-negative (right) bacteria. LPS, lipopolysaccharide.

develop or nucleate on or within cells, or in the immediate extracellular environment (demonstrated for a bacterial cell in Fig. 7). In addition to immobilizing metals, this also reduces the external free metal activity and may shift the equilibrium to release more soluble metal species to the soil solution. Precipitation and crystallization can restrict metal leaching and **bioavailability**, and can also release nutrients such as **sulfate** and phosphate.

The formation of extracellular and sorbed precipitates depends on the chemical composition of the extracellular environment and can be mediated or influenced by cellular processes (French et al., 2013). The precipitates can be copious and may completely entomb the cells, effectively entombing them (Fig. 8).

Environmental anions, such as hydroxyl (OH^-), sulfate (SO_4^{2-}), sulfide (S^{2-}), phosphate (PO_4^{3-}), carbonate (CO_3^{2-}) and oxalate ($\text{C}_2\text{O}_4^{2-}$), provide counterions for mineral hydroxides, sulfate and **sulfides**, phosphates, carbonates, and **oxalates** (Gadd, 2010). It is likely that cellular metabolism and excretion of such anions also contributes to the formation of these minerals, e.g., sulfide, oxalate, as observed for the formation of framboidal pyrite (FeS_2) (Fig. 9).

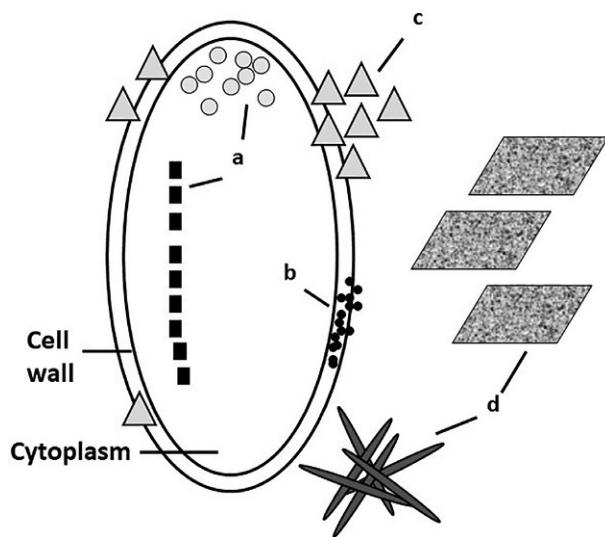


Fig. 7 Possible locations of metal precipitates and minerals associated with a bacterial cell: (A) biogenic precipitates within the cell **cytoplasm**; (B) metal precipitates formed by nucleation in the cell wall, e.g., periplasmic; (C) mineral **nanocrystals** formed independently of cell activities and subsequently sorbed to the cell wall; (D) minerals formed outside cells due to the influence of cellular activities.

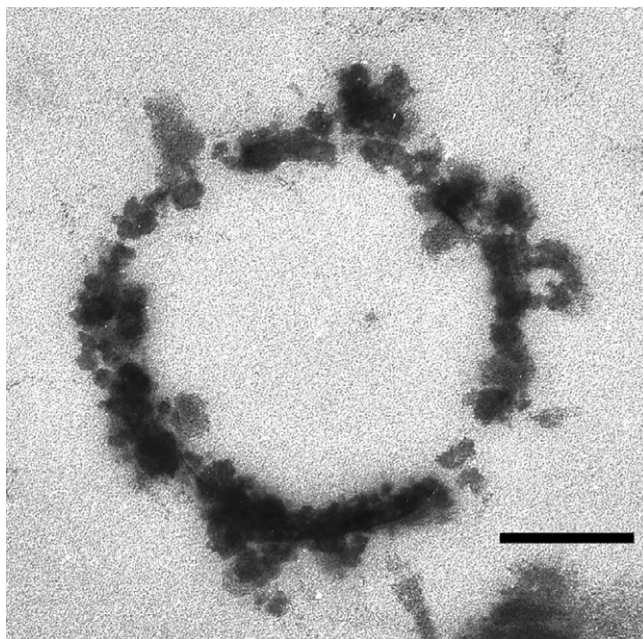


Fig. 8 Transmission electron micrograph of a thin-section sample taken from a Yellowstone **hot spring** (United States), showing an unknown bacterial cell embedded in metal-rich precipitates. Scale bar 150 nm.

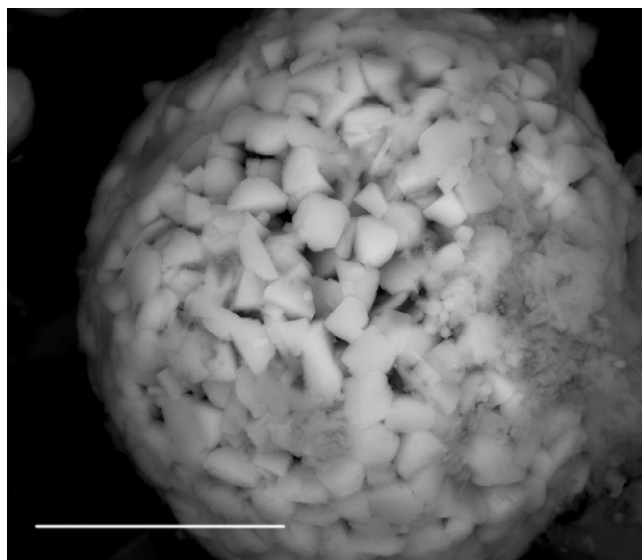


Fig. 9 Framboidal pyrite in marsh soil that receives acid mine drainage water containing dissolved sulfate and iron. Bar marker 10 μm (scanning electron micrograph courtesy of K. Mizutani).

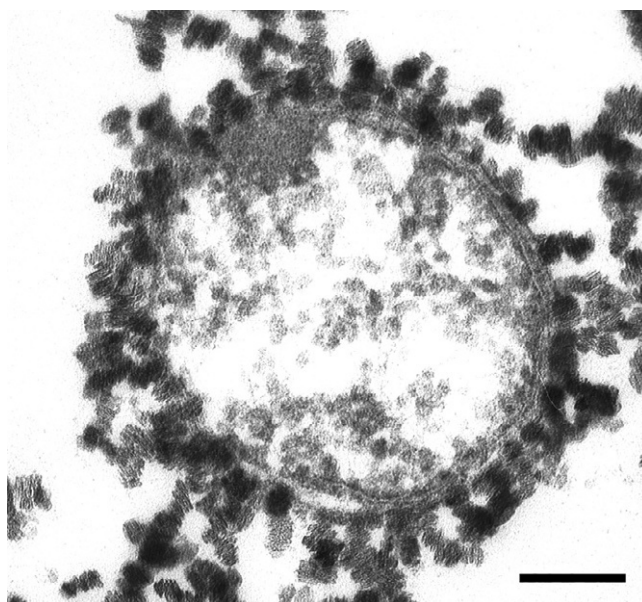


Fig. 10 Transmission electron micrograph of a thin section showing abiotically precipitated goethite nanocrystals sorbed to the cell wall of *Shewanella putrefaciens*. Scale bar 150 nm.

Metal oxide and hydroxide minerals formed by abiotic reactions can also attach to cells, and it can be difficult to distinguish these phenomena (Fig. 10). This is, however, another important mechanism by which cells concentrate metals as well as by anions released through solubilization of, for example, mineral phosphates.

Inside cells, metals can be incorporated into cellular constituents. Excess or toxic metals may be effluxed, form intracellular precipitates, bind to metal-complexing molecules, and, in eukaryotes, partition into vacuoles. The formation of internal precipitates may result from a change in metal oxidation state, which does not require an active cell; for example, gold is highly toxic to cells, yet intracellular precipitates of elemental gold form in bacteria after cells are killed. The formation of oriented chains of magnetite (Fe_3O_4) crystals by magnetotactic bacteria, found in sediments and hydric soils, is a unique example of directed, intracellular biomineralization in a prokaryote. In this case, the formation of magnetite may enable the bacteria to orient in the Earth's magnetic field, although other explanations have been proposed.

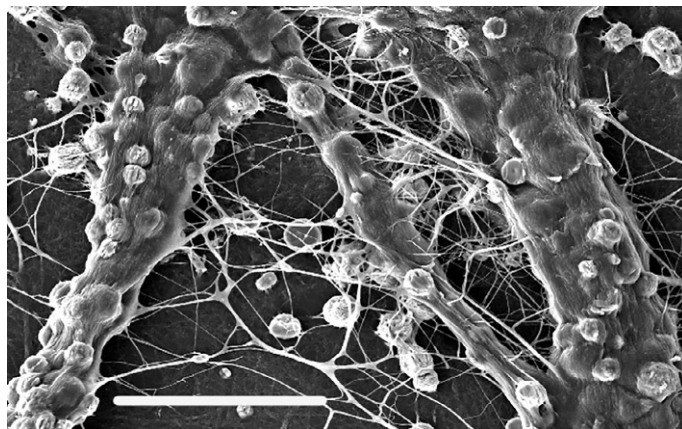


Fig. 11 Copper oxalate precipitated by *Beauveria caledonica*. The spherical crystals are associated with extracellular polymeric substances (EPS) produced by the filamentous hyphae and hyphal strands. Scale bar 200 μm (scanning electron micrograph produced by Marina Fomina).

Oxalates

Calcium oxalate is the most common form of oxalate associated with soils and leaf litter, occurring as dihydrate (weddellite) or the more stable monohydrate (whewellite). Calcium oxalate crystals are commonly associated with free-living, pathogenic, and plant-symbiotic fungi and are formed by the precipitation of soluble calcium as the oxalate. Fungal-derived calcium oxalate can exhibit a variety of crystalline forms (tetragonal, bipyramidal, plate-like, rhombohedral, or needles), which often associate with the outer surfaces of hyphae. The formation of calcium oxalate by fungi has an important influence on biological and geochemical processes in soils, acting as a reservoir for calcium in the ecosystem, and also influencing phosphate availability (Burford et al., 2003; Gadd, 2007; Gadd et al., 2014).

Fungi can also produce other metal oxalates with a variety of different metals and metal-bearing minerals, e.g., Cd, Co, Cu, Mn, Sr, and Zn (Fig. 11). The formation of metal oxalates may provide a mechanism whereby fungi can tolerate environments containing potentially high concentrations of toxic metals. A similar mechanism occurs in lichens growing on copper sulfide-bearing rocks, where precipitation of copper oxalate occurs within the thallus (see Gadd et al., 2014).

Carbonates

In many arid and semiarid regions, calcareous soils and near-surface limestones (calcretes) are often secondarily cemented with calcite (CaCO_3). This phenomenon has been partly attributed to physicochemical processes, although the abundance of calcified fungal filaments in weathered profiles of chalky limestone and Quaternary calcretes indicates fungal activity (Verrecchia, 2000). Calcite precipitation on fungal hyphae has been observed with the ectomycorrhizal fungus *Pisolithus tinctorius* in calcareous soil, as well as many other fungal species.

It has been proposed that calcite forms indirectly through fungal excretion of oxalic acid and precipitation of calcium oxalate, which results in dissolution of the internal pore walls of the limestone matrix. This enriches the matrix solution in free carbonate ions. During passage of the solution through the pore walls, calcium carbonate recrystallizes as a result of a decrease in CO_2 , hardening the limestone. Calcite crystals act as sites of further secondary calcite precipitation when the hyphae decompose. Through these processes, fungal activity is important in lithification (hardening), alteration, and diagenesis of subsurface limestone formations (Verrecchia, 2000).

Biomined carbonate precipitates are also found in association with bacterial biofilms. One striking example of microbial carbonate precipitation is the towering microbialites at Lake Van (Turkey), formed by thick biofilms of cyanobacteria (*Pleurocapsa* spp.) that colonize crusts of inorganic calcite precipitated in the alkaline soda lake water. Groundwater forced up and through the carbonate matrix raises the towers of aragonite and calcite to heights up to 40 m.

Redox reactions

Bacterial and fungal oxidation

The range of bacterial species with the capability to oxidize Fe and Mn spans sedimentary and soil environments. For bacteria to use Fe(II) specifically for growth, however, they must recover the energy produced in the oxidative process. Demonstration of a respiration-linked, energy-yielding reaction is particularly difficult for iron due to its rapid chemical oxidation at neutral pH and in the presence of atmospheric O_2 . It has been more readily shown for bacteria that grow at low pH and/or low O_2 , conditions that drastically slow chemical oxidation of Fe(II). Bacterial Fe oxidation is probably ubiquitous in environments with sufficient Fe^{2+} and conditions to support the growth of bacteria that can live by oxidative respiration of reduced metal species. Specific environments include drainage waters and tailing piles in mined areas, pyritic and hydric soils (bogs and sediments), drain pipes and irrigation ditches, and plant rhizosphere zones.

Iron-oxidizers found in acidic soil environments are acidophilic chemolithotrophs, organisms that use inorganic sources of chemical energy, such as *Acidithiobacillus ferrooxidans*, which is important for its role in generating acid mine drainage. Facultative chemolithotrophic microaerophiles such as *Leptothrix ochracea*, *Sphaerotilus natans*, and *Gallionella ferruginea* are common in mildly acidic to neutral environments, and accumulate abundant Fe hydroxide precipitates on their filaments or stalk structures that can age to more crystalline minerals.

After Fe, Mn is the second most abundant transition metal in the Earth's crust and is actively cycled in the same environments as Fe. Manganese oxidation is carried out by many of the organisms that are implicated in Fe oxidation, including *Bacillus*, *Pseudomonas*, and *Leptothrix* spp. In contrast to Fe^{2+} , substantial chemical oxidation of Mn^{2+} does not occur below pH 8, and solutions containing Mn^{2+} can be stable for weeks. This fact has allowed bacterial Mn oxidation to be identified more readily than its Fe counterpart.

It is not yet clear why cells carry out metal oxidation if not for gaining energy. Oxidation may enable the cell to rid itself of toxic intracellular concentrations of the metal, or the precipitates that coat the bacterial surface subsequent to oxidation may protect the cells from external stresses including predation by other microbes. Protection from UV radiation would have been especially crucial during the early evolution of life on Earth.

Fungi can also oxidize metals in their environment. Desert varnish is an oxidized metal layer (patina) a few millimeters thick found on rocks and in soils of arid and semiarid regions, and is commonly believed to involve fungal and bacterial processes. *Lichenothelia* (a fungus) can oxidize manganese and iron from metal-bearing minerals such as siderite (FeCO_3) and rhodochrosite (MnCO_3), and metals from rainfall or windblown dust. Several other fungal species can also promote Mn(II) oxidation to Mn(IV) O_2 through indirect mechanisms, such as interaction with a metabolite or structural component, and/or direct enzymic mechanisms mediated by, e.g., multicopper oxidases.

Dissimilatory metal-reduction by bacteria

When dissimilatory metal-reducing bacteria utilize relatively soluble metals such as Cr(VI) and U(VI) as electron acceptors, less-soluble Cr(III) and U(IV) species may result. Once reduced, soluble metals sorb to the cells and to extracellular polymers or partition to the bulk extracellular solution, where new minerals can form from the metals and suitable counterions. Post-reduction partitioning of metals depends on the element. For example, U(IV) accumulates as fine precipitates in the periplasm of Gram-negative metal reducers (Fig. 12). The Fe^{2+} and Mn^{2+} , generated by dissimilatory reduction, can form copious extracellular precipitates and nodules. The fate of the reduced metal depends on the environment, and it is feasible that a metal such as Fe, together with any associated redox-active metals, could be cycled hundreds of times between reducing and oxidizing zones before being incorporated into refractory minerals (Fig. 13).

Fungi and other microorganisms also precipitate reduced forms of metal species, e.g.: Ag(I) to Ag(0); selenate (Se(VI)) and selenite (Se(IV)) to Se(0); and tellurite (Te(IV)) to Te(0), by a variety of mechanisms (Gadd, 2007, 2010; Nanchaiaiah et al., 2016).

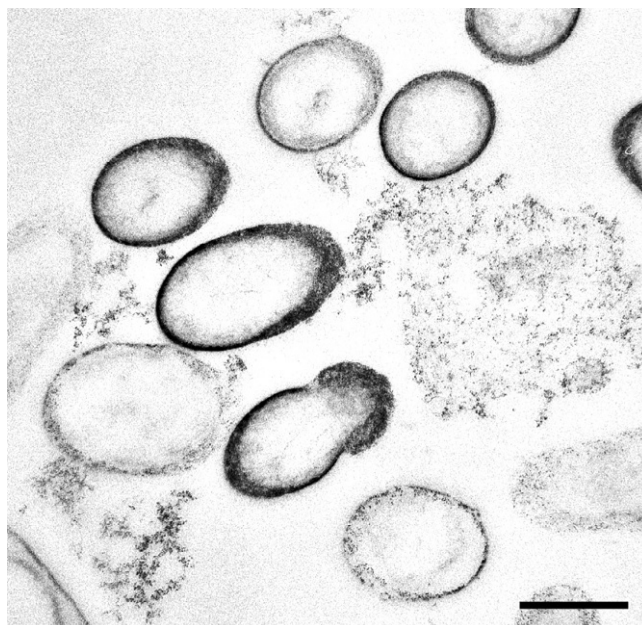


Fig. 12 Transmission electron micrograph of a thin section showing uranium precipitates formed in the periplasmic space of *Geobacter metallireducens* cells during dissimilatory reduction of U(VI). No metal stains were used, so that the contrast in the image is from the uranium precipitates only. Scale bar 150 nm.

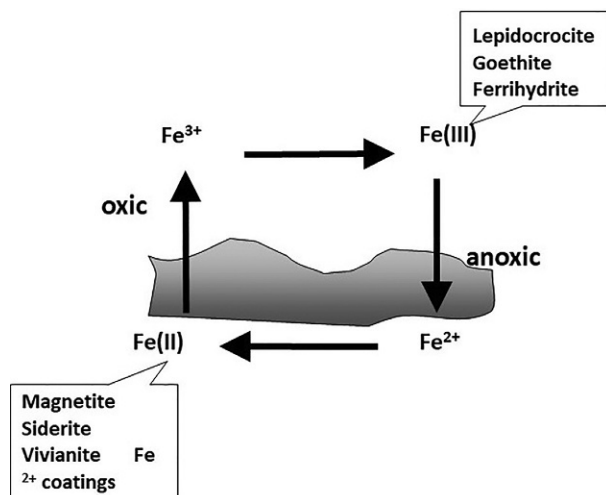
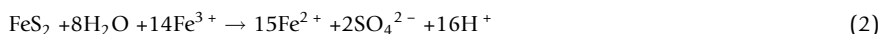


Fig. 13 Cycling of Fe between soluble and insoluble mineral, oxidized and reduced forms mediated by **microorganisms** in soils and sediments. Although by quantity Fe is the major element to undergo redox transformations, other metals, e.g., Mn, can be cycled by the same processes and associate with the precipitating minerals.

Biotechnological applications of microbial metal transformations

Ore leaching

Metals in ores that are not worth recovering by smelting can sometimes be effectively leached by bacteria and fungi. Important species of bacteria in such bioleaching processes are Fe and S **oxidizers** such as *A. ferrooxidans*, *Acidithiobacillus thiooxidans*, and *Leptospirillum ferrooxidans*. Their activities solubilize Fe and S contained in metal **sulfides**, with the concomitant release of associated metals such as Cu and Zn (Johnson, 2014). Large amounts of Fe^{2+} and **sulfuric acid** are generated in the process. It is most commonly accepted that the bacteria catalyze the oxidation of Fe^{2+} to regenerate Fe^{3+} , but do not directly attack the mineral:



where Eq. (1) gives the reaction for bacteria, and Eq. (2) is the abiotic reaction. There is, however, increasing evidence that *A. ferrooxidans* may directly enhance dissolution of metal sulfides via an enzymatic mechanism. Approximately 90 mol Fe^{2+} is oxidized to assimilate 1 mol C, which explains the tremendous amount of acidity—‘acid mine drainage’—that is generated. Industry can take advantage of the acidic leachate by recirculating it through the **ore body** to dissolve and harvest more metal. Here, the acidic leachate is recirculated through the ore body to solubilize more metal, and the soluble metals are then separated by sedimentation, solvents, or **electrodialysis**. When uncontrolled, such as at poorly managed or abandoned mining sites, the same process can have disastrous environmental consequences.

Extracellular ligands produced by fungi, including *Aspergillus* and *Penicillium* spp., have also been used to leach metals such as Zn, Cu, Ni, and Co from a variety of solid materials, including low-grade **mineral ore**, **solid waste**, and **electronic scrap** materials.

Bioremediation

Toxic metals in natural, industrial, and agricultural soils may be a risk to biota at all **trophic levels** if they exist at potentially toxic levels and are, or become, bioavailable. Some of the processes outlined above have potential for application to treat contaminated land. **Solubilization** processes provide a route for removal of metals from soil matrices, whereas immobilization processes enable metals to be transformed to insoluble chemically inert forms (Gadd, 2000, 2001). Although some processes can be used in situ (e.g., leaching using S-oxidizing bacteria), many are probably most suitable for ex situ use in **bioreactors**, where the mobilized or immobilized metal can be separated from soil components. The industrial development and operation of large-scale bioreactors containing sulfate-reducing bacteria for metal decontamination of effluents and process streams is now well known. Living or dead fungal and bacterial **biomass** and metabolites have been used to remove metals and **metalloids** from solution by biosorption or chelation, but sustained commercial application has not been achieved (Fomina and Gadd, 2014).

There has been interest in the ability of some mycorrhizal fungal **symbionts** to improve plant metal tolerance and increase metal translocation to the aboveground parts of plants. Such associations could be used to enhance the **phytoextraction** capabilities of plants grown for removing toxic metals from soil by accumulation in their shoots. Amelioration of metal **phytotoxicity** by **mycorrhizal fungi** has been widely demonstrated, particularly when inoculated with mycorrhizas isolated from metalliferous or **acid soils** (Gadd, 2007).

The increased attention to [microbial activities](#) in anaerobic, subsurface environments in the past few years has opened up new possibilities for the [bioremediation](#) of metal contaminants (Lovley, 2000). Metal(loid)s that form insoluble precipitates when reduced may be of particular interest for in situ treatment, such as Se(0), Cr(III), Tc(IV), and U(IV). Natural processes in subsurface soil may immobilize contaminants otherwise predicted to leach. Thorough knowledge and monitoring of the system are essential for such treatment schemes, because reoxidation can result in mobilization.

Biorecovery

There is growing concern about the security of supply of metal and mineral resources, particularly for those metals important in new digital technologies, electronics, and sustainable energy generation because of depletion of natural resources and inadequate reclamation and recycling. Many microbial metal and mineral transformations are relevant to biorecovery of critical metals, and bioprocessing is viewed as a useful contributor to solving this problem (Liang and Gadd, 2017). Reductive dissolution of oxidized ores can be used for recovery of cobalt and nickel, as well as other metals, in a more environmentally-friendly and cost-effective approach in comparison to traditional biomining (Johnson and du Plessis, 2015; Smith et al., 2017). Other potential approaches with apparent environmental benefits include biorecovery through bioprecipitation from solution using microbial metabolites or ligands released into solution through microbial activity. These include phosphate, liberated from organic or inorganic sources by phosphate-solubilizing organisms, fungal oxalate excretion, and microbially-induced carbonate precipitation (MICP). All such mechanisms have been examined for metals such as cobalt and nickel (Li and Gadd, 2017; Liang and Gadd, 2017; Yang et al., 2020), as well as biorecovery of rare earth elements (REE), e.g., cerium (Ce), lanthanum (La), neodymium (Nd), praseodymium (Pr), from leachates or monazite resources (Corbett et al., 2017; Kang et al., 2020), often exhibiting high efficiency in laboratory or pilot-scale demonstrations.

Summary

Dissolution, precipitation, oxidation, and reduction processes are vital to metal cycling in aerobic and anaerobic surface and subsurface environments. As active agents in cycling, bacteria, archaea and fungi are fundamental biotic components of natural [biogeochemical cycles](#) for metals and [metalloids](#). Some important processes catalyzed by [microorganisms](#) also have important applications in environmental [biotechnology](#) in the areas of ore leaching, [bioremediation](#) and element biorecovery.

Acknowledgments

This article is dedicated to the late Terry Beveridge, our colleague, friend and mentor, who co-authored the previous version of this article. S.M.G. acknowledges support from the US Department of Energy Natural and Accelerated Bioremediation program. G.M.G. gratefully acknowledges financial support from the [Biotechnology](#) and Biological Sciences Research Council (94/MAF12243; 94/BRE13640), British [Nuclear Fuels](#) plc. and the Natural Environment Research Council (NE/M010910/1 (TeaSe); NE/M011275/1 (COG³)).

References

- Banfield JH and Nealson KH (eds.) (1997) Geomicrobiology: Interactions between microbes and minerals. *Reviews in Mineralogy* 35.
- Beveridge TJ and Doyle RJ (eds.) (1989) *Metal Ions and Bacteria*. Wiley: New York.
- Burford EP, Fomina M, and Gadd GM (2003) Fungal involvement in bioweathering and biotransformation of rocks and minerals. *Mineralogical Magazine* 67: 1127–1155.
- Corbett MK, Eksteen JJ, Niu XZ, Croue JP, and Watkin ELJ (2017) Interactions of phosphate solubilising microorganisms with natural rare-earth phosphate minerals: A study utilizing Western Australian monazite. *Bioprocess and Biosystems Engineering* 40: 929–942.
- Fomina M and Gadd GM (2014) Biosorption: Current perspectives on concept, definition and application. *Bioresource Technology* 160: 3–14.
- French S, Puddephatt D, Habash M, and Glasauer S (2013) The dynamic nature of bacterial surfaces: Implications for metal-membrane interaction. *Critical Reviews in Microbiology* 39: 196–217.
- Gadd GM (2000) Bioremediation potential of microbial mechanisms of metal mobilization and immobilization. *Current Opinion in Biotechnology* 11: 271–279.
- Gadd GM (ed.) (2001) *Fungi in Bioremediation*. Cambridge University Press: Cambridge.
- Gadd GM (2007) Geomycology: Biogeochemical transformations of rocks, minerals, metals and radionuclides by fungi, bioweathering and bioremediation. *Mycological Research* 111: 3–49.
- Gadd GM (2010) Metals, minerals and microbes: Geomicrobiology and bioremediation. *Microbiology* 156: 609–643.
- Gadd GM, Bahri-Esfahani J, Li Q, Rhee YJ, Wei Z, Fomina M, and Liang X (2014) Oxalate production by fungi: Significance in geomycology, biodeterioration and bioremediation. *Fungal Biology Reviews* 28: 36–55.
- Hernandez ME and Newman DK (2001) Extracellular electron transfer. *Cellular and Molecular Life Sciences* 58: 1562–1571.
- Johnson DB (2014) Biomining—Biotechnologies for extracting and recovering metals from ores and waste materials. *Current Opinion in Biotechnology* 30: 24–31.
- Johnson DB and du Plessis CA (2015) Biomining in reverse gear: Using bacteria to extract metals from oxidised ores. *Minerals Engineering* 75: 2–5.
- Kang X, Csetenyi L, and Gadd GM (2020) Monazite transformation into Ce- and La-containing oxalates by *Aspergillus niger*. *Environmental Microbiology* 22: 1635–1648.
- Korenevsky A, Glasauer S, and Beveridge TJ (2002) Biomineralization by bacteria. In: Bitton G (ed.) *Encyclopedia of Environmental Microbiology*. New York: Wiley.

- Li Q and Gadd GM (2017) Fungal nanoscale metal carbonates and production of electrochemical materials. *Microbial Biotechnology* 10: 1131–1136.
- Liang X and Gadd GM (2017) Metal and metalloid biorecovery using fungi. *Microbial Biotechnology* 10: 1199–1205.
- Lloyd JR, Pearce CI, Coker VS, Pattrick RADP, van der Laan G, Cutting R, et al. (2008) Biomineralization: Linking the fossil record to the production of high value functional materials. *Geobiology* 6: 285–297.
- Lovley DR (ed.) (2000) *Environmental Microbe–Metal Interactions*. Washington, DC: ASM Press.
- McLean JS, Lee JU, and Beveridge TJ (2002) Interactions of bacteria and environmental metals, fine-grained mineral development, and bioremediation strategies. In: Huang PM, Bollag J-M, and Senesi N (eds.) *Interactions Between Soil Particles and Microorganisms*, pp. 228–261. New York: Wiley.
- Nanchaiah YV, Mohan SV, and Lens PNL (2016) Biological and bioelectrochemical recovery of critical and scarce metals. *Trends in Biotechnology* 34: 137–155.
- Smith SL, Grail BM, and Johnson DB (2017) Reductive bioprocessing of cobalt-bearing limonitic laterites. *Minerals Engineering* 106: 86–90.
- Sullivan TS and Gadd GM (2019) Metal bioavailability and the soil microbiome. *Advances in Agronomy* 155: 79–120.
- Verrecchia EP (2000) Fungi and sediments. In: Riding RE, Awramik SM, and S.M. (eds.) *Microbial Sediments*, pp. 69–75. Berlin: Springer-Verlag.
- Yang Y, Song W, Ferrier J, Liu F, Csetenyi L, and Gadd GM (2020) Biorecovery of cobalt and nickel using biomass-free culture supernatants from *Aspergillus niger*. *Applied Microbiology and Biotechnology* 104: 417–425.

Dynamics of radionuclides

Siobhan Staunton, UMR Eco&Sols, INRAE, INRAE-IRD-Cirad-Institut Agro-Université de Montpellier, Montpellier, France

© 2023 Elsevier Ltd. All rights reserved.

Introduction	256
Definitions	256
Radionuclide dynamics: Similarities and particularities	257
The roles of soil in the bioavailability of radionuclides and the exposure of living organisms to radioactivity	258
Transfer from soil to the human food chain	259
Remediation of contaminated soils	259
Soil microbiological activity	260
Case study: Cesium	260
Case study: Strontium	261
Case study: Radon	261
Case study: Halogens	262
Conclusions	262
References	262

Abstract

The chemistry of radioactive isotopes in soil does not differ from that of their stable isotopes. There are however three important differences: radioisotopes are often present in trace amounts; isotopes may change chemical and physical form when they decay; and adverse effects on biological organisms may occur without physical contact or absorption due to ionizing radiation. Case studies are given for the most environmentally relevant isotopes.

Key points

- Concepts, definitions and units of radioactivity
- Similarities and differences between radionuclides and various pollutants and nutrients
- Effects of soil on mobility of radionuclides and exposure of living organisms to radiation
- Examples of most relevant radionuclides in soil

Introduction

The objective of this chapter is to give an overview of the role of soil in the bioavailability of radionuclides and the exposure of living organisms to radioactivity. The essential concepts and units are presented, and the dynamics of radionuclides compared to stable elements, whether they are nutrients or potentially toxic. More detail can be found in standard texts on nuclear physics and radioprotection and in the information sheets produced for the general public by the International Atomic Energy Authority (IAEA), UN (UNSCEAR, 2000) and many national authorities (for example Desmet et al., 1990; Keith-Roach and Livens, 2002; Tykva and Berg, 2004). Detailed reviews of the dynamics of radionuclides are given in the three volumes of the books by Coughtrey and Thorne (Coughtrey et al., 1983; Coughtrey and Thorne, 1983a, 1983b).

Definitions

Radionuclides or radioisotopes are isotopes with unstable nuclei that lose energy by decaying and releasing ionizing radiation; decay may result in a radioactive or a stable isotope that may be the same element or a different element with different chemical properties. Natural radionuclides may be primordial, namely present at the origin of the Earth and so long-lived that they are still present, (Ishii et al., 2004) or secondary, namely formed as part of the decay chain of primordial isotopes, or (iii) formed by cosmic radiation, as in the case of ^{14}C . There are also man-made radionuclides produced during nuclear energy production, military activities and for therapeutic reasons. The three classes of radioactive decay that are relevant to environmental science, and the ionizing radiations produced, have different physical properties.

- Gamma (γ) radiation is electromagnetic radiation released from the nucleus of atoms. In common with X-rays, produced by electronic transitions, outside the nucleus, this radiation penetrates matter, losing energy, with the potential to harm living organisms. It is attenuated by physical barriers, but not completely stopped.
- Beta (β) radiation is formed by electrons (e^-) or positrons (e^+) created within the nucleus by the transformation of a neutron to a proton and an electron, or of a proton to a neutron and a positron. In both cases the number of protons in the nucleus changes and so the radioactive decay results in a chemical change. The energy of β radiation depends on the decaying element. Radiation is referred to as hard or soft; hard radiation is energetic, energizing radiation such as that emitted by ^{32}P , and soft radiation is weakly ionizing such as ^{14}C . This radiation can be completely stopped by a physical barrier. The thickness of the barrier required to stop this radiation depends on both the energy of the radiation and the nature of the barrier.
- Alpha (α) radiation consists of helium nuclei, known as alpha particles, with an atomic mass of 4 and a charge of +2. After emission of alpha particles, the nucleus changes in both atomic mass and atomic number resulting in a chemical change. These particles interact strongly with matter, losing energy, which makes them easy to stop with physical barriers, but very dangerous to living organisms.

Radioactive decay is stochastic, thus the probability of a radioactive nucleus decaying is random, and described by a constant denoted λ , which is inversely related to the half-life of the radioisotope. Decay rate, or activity, A , which can be expressed mathematically as the rate of decrease in the number of nuclei present, depends on the number of radioactive nuclei present at any time t , N_t , thus.

$$\text{Rate of decay} = A = -dN/dt = \lambda N_t \quad (1)$$

The half-life, $\tau_{1/2}$, is defined as the time for the number of nuclei initially present, N_0 , to decrease by decay to half that number. It is inversely related to λ , as seen by rearranging equation 1, and integrating.

$$\ln 2 = \lambda \tau_{1/2} \quad (2)$$

The units of radioactivity are the Becquerel (Bq), defined as one disintegration per second. The older, non-SI unit is also still used, the Curie (1 Ci = 3.7×10^{10} Bq). Half-life should be expressed in SI units, namely seconds, but is often converted to years and multiples thereof. The biological danger associated with radioactivity depends not only on the radioactivity, but also on the energy transferred, and this is denoted by the absorbed dose, Gray (Gy) defined as one Joule transferred to a material with respect to its mass (J kg^{-1}). The older, non-SI unit is also used, the rad = 0.01 Gy. Finally, since the nature of the radioactivity also influences the danger associated with exposure to radioactivity, a quality factor, Q , is also used to define the effective dose, or Sievert (Sv). The values of Q are 1 for β , γ and X-ray radiations, and 10 and 20, respectively for proton, and alpha radiation respectively. The same quality factors are used to derive the effective dose in rem (roentgen equivalent man) from the absorbed dose in rem.

Radionuclide dynamics: Similarities and particularities

Although in many respects radionuclide dynamics are not different from those of other elements of interest in soils and the environment, some points that merit special attention are summarized below and will be treated in greater detail in the text.

- Environmentally relevant radionuclides may not be the focus of interest as stable isotopes. The most studied elements in soil and environmental sciences are nutrients or potentially toxic. Radionuclides of environmental interest may have no or little known biological role, they may also be present in negligible amounts in the environment. Examples of elements that are predominantly studied as radionuclides include artificial isotopes such as neptunium and technetium, but also elements present in nuclear waste or released from nuclear weapons testing or related to nuclear power production, including cesium, strontium and chlorine. A few exceptions are studied both as radionuclides and stable elements, such as nickel and selenium.
- The chemistry of radionuclides and stable isotopes of the same elements is largely similar. The chemistry of an element is determined by its position in the Mendeleev Periodic Table, and thus its atomic number. Thus to a large extent all isotopes of an element have identical chemistry, regardless of atomic mass. Nevertheless covalent bond strength varies with the mass of atoms and so the strength of covalent bonds and the rates of reactions involving these bonds differ between isotopes, leading to an "isotope effect."
- Concentration effect. Another feature of the chemistry of radionuclides is that they are generally present in tiny amounts and their environmental chemistry may be concentration dependent. An important example of an element where the concentration effect is relevant is radiocesium. This element will be considered later, given its presence in the environment due to atmospheric fallout and accidental releases.
- The isotope ratios of light elements, notably C and N, and more recently heavier elements, arise from either radioactive decay, or from small isotopic discrimination repeated numerous times. These ratios are used either to date material or to study food web ecology. They may also provide information on the source of an element in the environment. In radioecology the isotopic ratio is a particularly useful tool to establish the origin of contamination.
- Decay. Radioactive decay is a stochastic process, and thus completely independent of chemical and physical conditions. This is an important contrast with organic contaminants.

- Change of state. Radioactive decay often leads to a change in the element and thus its chemistry. This may include a change in physical state. The best known example of this is for radon. Radon is formed in the decay chain of uranium and radium. All its isotopes are radioactive. Radon is a gas, but all its progeny (or daughters) are nongaseous. Therefore, once radon has been inhaled, any radon not almost immediately exhaled will remain in the lungs and will decay to give elements that are precipitated, thereby increasing the biological residence period.
- External and internal exposure. A major difference between radioactive isotopes and other potentially toxic species is the risk associated with external exposure to radiation. Some radionuclides, including the actinide elements, may also have chemical toxicity. Radiation may be harmful with no direct contact, ingestion or inhalation, between the isotope and living organisms. The dangers associated with internal exposure are increased because exposure time depends on the residence time in the living organism. A biological or effective half-life, the time required for half the isotope or the radioactivity to be eliminated, is thus defined.
- Soil as source, regulator and shield. The functions of source and shield of soil for radionuclides are similar to those for nutrients or other pollutants. However the shield function is related to the risk of external exposure and thus particular to radionuclides, both natural or introduced by human activity.

The roles of soil in the bioavailability of radionuclides and the exposure of living organisms to radioactivity

Fig. 1 is a schematic representation of the sources of external and internal exposure to radiation. Data are approximate, as exposure varies according to many factors including place of residence, building regulations to limit radon accumulation, occupation, medical procedures, and diet. Man-made exposure is less than that of natural exposure, and even food intake is largely due to natural radioisotopes, for example foods high in potassium due to K-40. Soil may be a source of radionuclides and associated radioactivity. This is the case for natural, primordial or secondary radionuclides, of geological origin. Soil acts as a sink for many radionuclides of anthropogenic origin, in common with many pollutants. As both a source and sink, soil regulates the mobility and biological exposure to radioactive elements. The major sources of man-made radionuclides in the terrestrial environment are from accidental releases, the most notorious being the accidents at the Chernobyl and Fukushima nuclear power plants in 1986 and 2011, respectively. Other, often legal and tolerated, releases have occurred around nuclear power plants. The disposal of radioactive waste is also a potential source of release, probably released into the geosphere with the potential for slow mobility toward the pedosphere. Although radioactive sources are used for medical purposes, these tend to be short-lived. Nuclear weapons testing began after the Second World War and caused atmospheric fallout of radionuclides, the most studied has been ^{137}Cs , which peaked in the early 1960s. Cesium-137 is widely used as a tracer of soil erosion because of the relatively homogenous deposition of fall-out containing this isotope in the Northern hemisphere (Zapata, 2003). The multidisciplinary scientific field of radioecology developed from increasing concern about the impact of radionuclides in the environment, and soil science figures prominently within this field. An illustration of this research focus was the foundation of the Journal of Environmental Radioactivity in 1984. Radioecology is a discipline that has experienced marked fluctuations with peaks in activity following the major accidents at Chernobyl and Fukushima. Although exposure to ionizing radiation must be limited to protect human and animal health, it should be remembered that, except in highly contaminated areas, the radiation exposure from contaminated soil and food is very small compared to

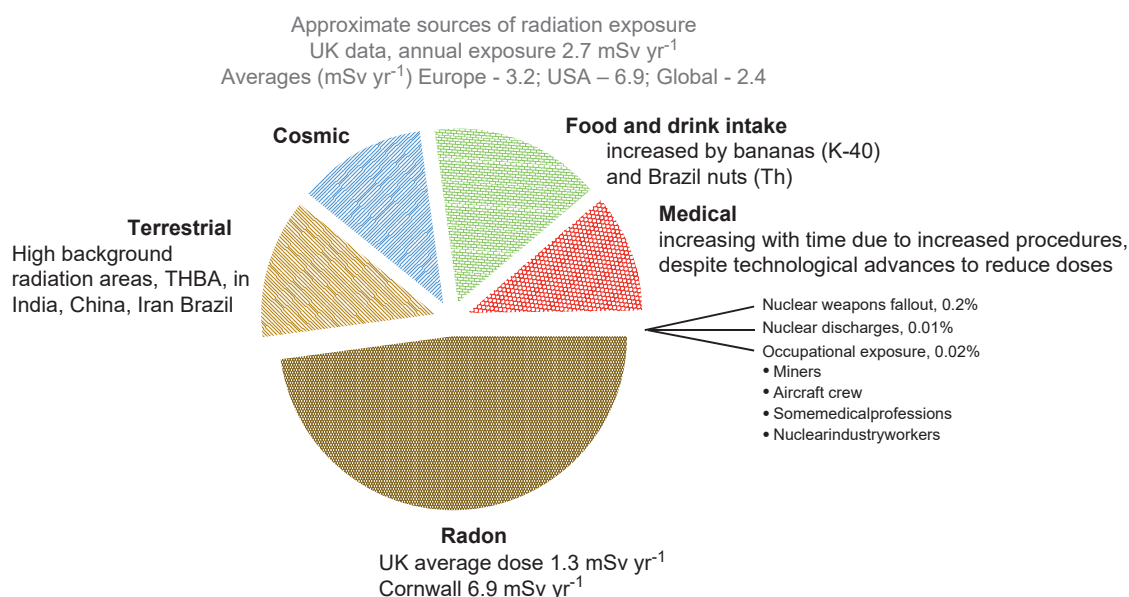


Fig. 1 Relative sources of exposure to ionizing radiation.

that from natural and medical sources. Large variations in terrestrial (including radon) radiation exist, with “high background radiation” areas often associated with black sand containing monazite.

In common with other pollutant substances, the interactions between radionuclides and soil determine their mobility and hence bioavailability. A major feature of these interactions is adsorption, which is often described by the distribution coefficient, K_d , defined as the ratio of content in the adsorbed phase and concentration in solution. The principles underlying the adsorption and desorption of radionuclides are no different from those for non-radioactive pollutants or nutrient elements. When sorption is dominated by electrostatic interactions, there is little distinction between radioactive and stable isotopes of a given element. However, the experimental measurement of adsorption and its sensitivity to various parameters is complicated by the effects of concentration. Environmental concentrations of a radioisotope may differ by orders of magnitude from convenient concentrations for laboratory experiments; this is particularly well documented for radiocesium. Isotopic discrimination is generally assumed to be negligible at the soil solution–biological membrane interface and for transport within living organisms. Nevertheless, the repetition of reactions with a small discrimination leads to isotopic signatures or isotopic fingerprints as observed and interpreted for stable isotopes. Isotopic discrimination is of greater importance for reactions involving covalent bonds, often involving catalytic reactions.

Transfer from soil to the human food chain

Soil–plant transfer is an important step in the contamination of the human food chain. The bioavailability of radionuclides is often described using the soil–plant transfer factor, FT . Although contaminated food and water usually make minor contributions to the exposure of humans to radiation (Fig. 1), protecting the food chain is essential following a serious contamination. In the short term, immediately after an accident, the only solution is to monitor and remove contaminated food from the market, or ban removal of plants and animals from the environment for consumption. Such restrictions were imposed on fishing, hunting and mushroom gathering after the accidents at Chernobyl and Fukushima. In the longer term soil amendments, land-use changes from food crops to fiber or construction-material crops may be recommended. Clay and even silver thiourea ($AgTU$) were tested as amendments to reduce radiocesium uptake, particularly on extensive grazing land, but were usually found not to be cost-effective, or with serious disadvantages. Soils may be fertilized, for example with potassium additions, which may limit the uptake of cesium. The discrimination between radionuclides and other elements is often studied in the context of soil-to-plant transfer. This is a feature of all contaminant availability, more usually referred to as competition, often in the context of pollution. Some radionuclides have no biological role and the elements may be present only in trace amounts in the environment. For example, cesium has similar chemistry to potassium, and strontium to calcium.

Soil ingestion is a significant mode of transfer to the human food chain. Considerable amounts of soil may be ingested by grazing animals, particularly sheep when grass is short. Soil may adhere to vegetation after splash-back and thus be consumed if food is not washed. This mode of entry into the food chain exists for other contaminants, but is most important when the contaminant has been deposited on the soil surface, and is sparingly mobile because it is strongly adsorbed. Inhalation of contaminated soil is another mechanism for contamination, particularly for agricultural workers. Cesium is an example of an element where soil ingestion plays an important role in radioecology.

An important feature of the dynamics of radioisotopes in soil, compared with other potentially harmful substances, is the barrier effect of soil. As described above, radiation may be attenuated or completely stopped by material. There is little risk of external exposure to α -emitters that are stopped by less than 1 mm of a dense material such as soil. For soil contaminated by a β - or γ -emitter, the addition of a layer of soil above them may be an effective solution to limit external exposure to radiation, however the effect on bioavailability depends on many factors.

Remediation of contaminated soils

Various remediation measures have been proposed to limit human exposure to radiation from radiocesium and its the entry into the human food chain. In extreme situations near accident sites, the only feasible solution is to create exclusion zones where access is strictly limited and monitored. Soil may be planted to minimize erosion and contamination of surface and ground water. Land is not used for food or crop production and measures are taken to limit the risk of wild-fires that would cause the release of contaminants into the atmosphere. In less extreme situations, soil may be covered with uncontaminated soil to limit external radiation, and wind erosion. But this is unlikely to reduce plant uptake because of root distribution at depth. Soil fertilization or amendments may reduce soil-to-plant transfer. When agriculture is extensive, as for sheep grazing on upland soils, amendments are unlikely to be cost effective. The solution used is to transfer animals to uncontaminated pasture before slaughter, thus reducing animal contamination to acceptable levels. The application of silver thiourea was proposed to complex radiocesium and thus immobilize it. However this solution was not only expensive, but carried the risk of excessive consumption of potentially toxic silver and thiourea by animals. In many cases, as for non-radioactive contaminants, a cost-effective solution is to use land for non-food crops, but a full safety analysis would include the risk of external exposure from radioactivity in the finished fiber or building material. Common approaches to safe use of contaminated land include the choice of crops that do not accumulate the elements, sometimes due to rooting distribution, and the use of amendments or fertilizers. Potassium fertilization has been used as a remediation measure in Cs-contaminated soils, but the overall effect may be complicated by potentially conflicting biological and chemical effects.

Soil microbiological activity

Radioactivity commonly found in the environment is not thought to have adverse effects on soil microbiological activity. To illustrate this the doses used to sterilize soil or medical equipment may be compared to environmental levels of radioactivity. Sterilization uses radiation of several tens of kGy, typically up to 40 kGy, whereas annual doses from all sources are of the order of mSv, and even in the exclusion area around Chernobyl a day visit for a tourist (avoiding the most dangerous areas) would only result in a dose of the order of μSv , equivalent to about 10 mSv year^{-1} . Another example is that the maximum doses recorded after the accident at Fukushima in 2011 were 400 mSv h^{-1} , equivalent to $3.5 \text{ kSv year}^{-1}$. These figures can be compared to 2 Sv for a session of radiotherapy.

The dynamics of some radionuclides, however, may be sensitive to soil microbial activity. Some radionuclides may be bioaccumulated in microorganisms, including adsorption on biological membranes. Obviously this depends on their adsorption properties and bioavailability. These effects are reviewed by (Lloyd, 2003). Important carrier phases that are sensitive to microbial activity include iron and manganese oxides. Other effects of microbial activity are indirect, and related to the chemical and mineralogical effects of microbial activity, including pH changes, and weathering of soil minerals thereby changing adsorption properties. These effects have not been widely studied. One well known example is the weathering of mica-like minerals to yield frayed-edge sites that have a strong affinity for cesium. In addition, many radionuclides are sensitive to the redox changes that may be caused by microbial activity. These radionuclides include both nonmetals, such as iodine, selenium and technetium, and metals, in particular the actinides and lanthanides, thus many artificial isotopes and the isotopes created in the decay chain of uranium and radon. In many cases the adsorption, and thus the mobility, of the elements differ markedly as a function of their oxidation state. For example, selenium and technetium are sparingly adsorbed by soil minerals and are mobile in soil in their oxidized state (SeO_3^{2-} , SeO_4^{2-} and TcO_4^-). Both abiotic and microbially induced biotic reduction of these species yield less mobile and precipitated species (elementary Se, selenides, $\text{TeO}_2 \cdot 2\text{H}_2\text{O}$ and insoluble sulfides of Tc(IV)). Bacterial respiration has been reviewed with emphasis on Se and Tc by Stolz and Oremland (1999) and Lloyd et al. (1998, 2000). Selenium has also been reported to undergo transformations by bacteria and fungi to yield volatile methylated compounds. Similar reactions are known for iodine, but have not yet been conclusively demonstrated for Tc.

Case study: Cesium

Cesium is an alkali metal (Group 1 of the Periodic Table). Cesium-133 is the only natural isotope (100% natural abundance). Cesium has no known biological role, but may substitute for potassium and so is relatively easily assimilated by biological systems. Three of the numerous artificial radioisotopes may be present in the environment: ^{135}Cs , ^{134}Cs and ^{137}Cs . The first of these is a component of long-lived fission product ($\tau_{1/2} = 3 \times 10^6$ years) that is a low-energy β -emitter. As a component of nuclear waste, there is a risk of its release into the environment. The other isotopes have been released by atmospheric fallout caused by nuclear weapons testing in the 1950s and 1960s and by accidental releases particularly at Chernobyl in 1986 and Fukushima in 2011. They are both γ - and β -emitters, with half-lives of 2.07 years for ^{134}Cs and 30.2 years for ^{137}Cs .

Cesium is the most widely studied radionuclide because of the presence of ^{137}Cs (and to a lesser extent ^{134}Cs) in both atmospheric fallout from weapons testing in the 1950s and 1960s and from the notorious accidents at Chernobyl and Fukushima. Atmospheric fallout from weapons testing has made ^{137}Cs an important and widely used marker of soil erosion. By comparison most fallout from Chernobyl and Fukushima occurred over much shorter times shortly after the events and therefore the distribution was more heterogeneous and dependent on rainfall. The relation between climate and soil type had important consequences on the subsequent dynamics of radiocesium. Many high-rainfall areas affected by fallout were upland and peat dominated areas with small clay content, and this led to low fixation of Cs. This was highlighted after the Chernobyl accident, when some models predicted rapid strong fixation of Cs on soil and therefore low mobility and soil-to-plant transfer.

Cesium is reversibly adsorbed on negatively charged clay minerals. In general, clay minerals have a greater affinity for Cs than other cations present on the exchange complex, particularly divalent cations. This is coherent with cation exchange theory and arises from its valency and hydration energy. There is a marked concentration dependence of Cs adsorption on soils and clay minerals. Adsorption isotherms are non-linear, often being fitted to the Freundlich isotherm, with affinity decreasing with increasing concentration. The strong affinity at trace concentrations of Cs is ascribed to the marked affinity of a small proportion of adsorption sites on illites and weathered mica. These are known as frayed edge sites that collapse following adsorption of Cs with loss of the hydration water molecules (Cremers et al., 1988). This reaction is energetically favorable and thus essentially irreversible. These adsorption sites have been named "frayed edge sites" (FES). The concentration dependence of affinity for Cs is an indication of the heterogeneity of surfaces. Some authors assert that a small number of FES with distinct affinities exist, although there may be a continuum of sites with differing affinities. The most important consequence of this is that measurements made with stable Cs (thus at much greater concentrations than found in the environment) are not relevant to understanding or predicting environmental dynamics. Since radiocesium is present in trace amounts in soils, the amounts of micaceous minerals required to dominate the overall affinity are not directly quantifiable by direct methods. Cesium adsorption is thus not well predicted by clay content. Clay minerals in soils are associated with organic matter coatings. Organic matter has been shown to decrease the strong affinity of micaceous minerals for Cs (Staunton et al., 2002). This is thought to reflect a physical blockage of the FES, preventing their collapse and reducing the specific adsorption of Cs. The strong adsorption of Cs on soil limits its bioavailability, but conversely leads to a

significant contribution to contamination of the food chain by soil ingestion of grazing animals, particularly sheep in poor grazing conditions.

Soil potassium plays an important role in Cs dynamics in soil. Potassium competes with Cs both at biological membranes, inhibiting uptake, but also inversely at the soil exchange complex, limiting adsorption. Potassium fertilization is recommended to reduce radiocesium uptake by crops. In addition, excess soil potassium would reduce the depletion in the rhizosphere where K concentrations may be markedly smaller than in bulk soil, thus enhancing Cs uptake. The “radiocesium interception parameter” (RIP) is important for understanding Cs dynamics in the soil–plant system (Delvaux et al., 2000). It combines the limiting Cs-K selectivity coefficient and the abundance of FES. This parameter has been reported to enable a good prediction of the Cs distribution coefficient in soils.

Case study: Strontium

Strontium is an alkaline earth metal (Group II) that exists as a stable isotope, with a simple chemistry, similar to that of calcium. It has numerous radioactive isotopes, the most environmentally significant being ^{85}Sr , ^{89}Sr , and ^{90}Sr , with half-lives of 64.8 days, 50.5 days and 29.1 years respectively. Strontium-90, a beta-emitter in common with Cs, has been introduced into the environment by fallout from atmospheric testing and by accidental releases. Soil contamination of the two elements are often reported together. However Sr contents are considerably smaller than those of Cs and cause little environmental concern. Strontium is adsorbed by soil components and this adsorption is easily reversible. There is little biological discrimination between Sr and Ca, therefore Sr is found in milk and other animal products, and once absorbed may be accumulated in bone. There are no effective countermeasures against radiostrontium contamination of soil.

Case study: Radon

Radon is the most widely studied natural radioelement in soil science and radioecology. Soil acts as both a source, a barrier and a brake to the diffusion of radon. Unlike other radioelements of environmental concern radon is a gas, the heaviest element in the noble gas group of the periodic table. Fig. 2 shows the decay chain of uranium, highlighting the decay chain of radon, the only gaseous element in the chain. All its isotopes are radioactive and the most commonly studied is ^{222}Rn , an alpha-emitter, with a half-life of 3.8 days. It is formed as one of the progeny of the uranium decay chain originating from the disintegration of uranium-238, known as the uranium or radium series. To a large extent, the risk of exposure to radon depends on the geology of an area, with large global and national variations. Greatest amounts of radon are found in granitic or volcanic areas. Large regional differences are reported in human exposure to radiation from radon, for example there is factor of 2 between the radiation dose due to radon in Brittany compared with the French national average dose, and similarly a factor of 5 between doses in Cornwall and on average in the UK. Radon may accumulate in mines and other enclosed underground places such as caves and basements. Although the cause was not understood, the illness due to high radon exposure in mines has been known for centuries: the wasting disease of miners was named *mala metallorum* by Paracelsus in 1530. As a noble gas there is little chemical interaction between soil and radon. Nevertheless, soil, in particular soil texture, plays an important role in attenuating exposure to radon. The porosity and permeability

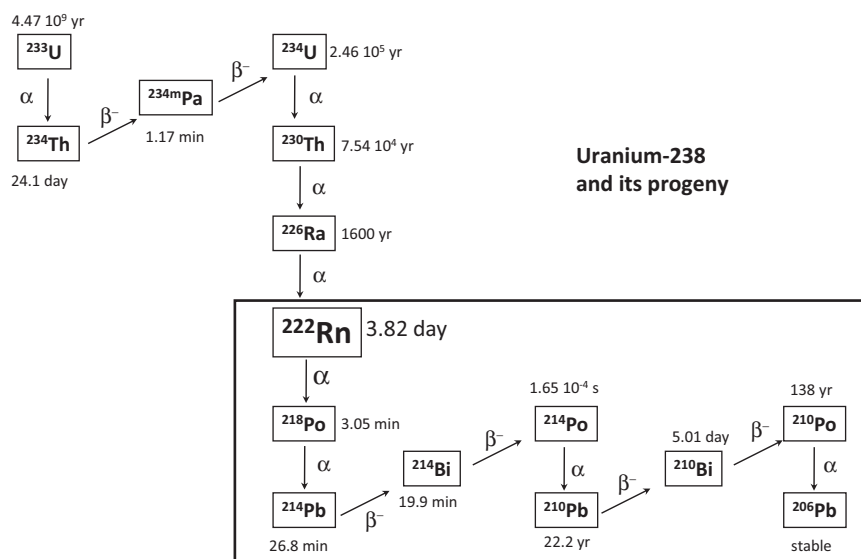


Fig. 2 Decay chain of ^{238}U , giving half-lives and major radiation, highlighting the decay chain of one of its progeny, ^{222}Rn .

of soil determine the emanation of radon at the surface. Clay rich soils and subsoils slow radon diffusion, and so a greater proportion decays to form non-gaseous progeny before reaching the surface. Emanation of radon also varies with soil humidity, because of the volume fraction and tortuosity of air-filled pores. These may result in a bell shaped relation between radon emanation and soil humidity.

Case study: Halogens

Two elements of the halogen group exist as radioisotopes of environmental concern; chlorine and iodine. Chlorine has two stable isotopes and 14 radioactive isotopes with atomic masses between 32 and 43. One of these, ^{36}Cl , is a predominantly beta-emitter with a half-life of 3×10^5 years. There are both natural and artificial sources of this isotope in the environment. The natural sources include interactions of argon, potassium and calcium with cosmic radiation and underground neutron activation of ^{36}Cl with neutrons produced in the uranium decay series. Chlorine-36 has been used as a geological dating technique for rocks. It is produced within nuclear reactors and thus may be released during retreatment and waste disposal. It was also present in atmospheric fallout following weapons testing during the 1950s and 1960s. Amounts of ^{36}Cl from natural and artificial sources are of approximately the same order, and it is therefore difficult to quantify contamination with certainty. The chemistry of radiochlorine is largely identical to that of chloride, a very mobile, sparingly adsorbed, anion. Iodine is a heavier element of the halogen group. It has 25 isotopes, with atomic masses between 117 and 141, and only ^{127}I is stable. Two of the radioisotopes have half-lives sufficiently long to be of environmental concern; ^{131}I and ^{129}I with half-lives of 8 days and 1.57×10^7 years respectively. Both emit both β^- and electromagnetic radiation. Some ^{129}I is produced naturally by some fission reactions and persists because of its long half-life. Both isotopes are produced in nuclear reactors, considerably more ^{131}I than ^{129}I and have been introduced into the environment as tolerated discharges, waste disposal, fallout from atmospheric weapons testing in the 1950s and 1960s and from accidents. Given the difference in half-lives only ^{129}I remains in discharges or from fallout. However, the presence of ^{131}I in accidental releases such as following the accidents at Chernobyl and Fukushima presented risks of human exposure with fixation of radioiodine to the thyroid. This may be counteracted by administration of stable iodine to saturate the thyroid gland. The dominant pathway of human contamination by radioiodine following the Chernobyl accident was by milk consumption, and particularly affected children. In contrast, the main pathway of contamination after Fukushima was air inhalation. The chemistry of iodine is more complex than that of chlorine, with a more extensive organic and redox chemistry. It is therefore less mobile in the environment and its dynamics less simple to predict and model.

Conclusions

In conclusion, the dynamics of radionuclides in soil are controlled by similar processes as for other natural or artificial contaminants. The major differences arise from the risk of external exposure, in addition to toxicity following absorption. Soil may be a source of radioactivity, as well as a physical barrier to radiation and to the mobility of radionuclides. Radioactivity may be of natural or man-made origin, and in many countries the exposure to natural radiation accounts for a large proportion of exposure, although this includes radon against which preventive measures can be taken.

References

- Chaboche P-A, Saby NPA, Laceby JP, Minella JPG, Tiecher T, Ramon R, Tassano M, Cabral P, Cabrera M, da Silva YJAB, Lefevre I, and Evrard O (2021) Mapping the spatial distribution of global ^{137}Cs fallout in soils of South America as a baseline for Earth Science studies. *Earth-Science Reviews* 214: 103542.
- Coughtrey PJ and Thorne MC (eds.) (1983a) *Radionuclide Distribution and Transport in Terrestrial and Aquatic Ecosystems*. In: vol. 1. Netherlands, Rotterdam: A.A. Balkema.
- Coughtrey PJ and Thorne MC (eds.) (1983b) *Radionuclide Distribution and Transport in Terrestrial and Aquatic Ecosystems*. In: vol. 2. Netherlands, Rotterdam: A.A. Balkema.
- Coughtrey PJ, Jackson D, and Thorne MC (eds.) (1983) *Radionuclide Distribution and Transport in Terrestrial and Aquatic Ecosystems*. In: vol. 3. Netherlands, Rotterdam: A.A. Balkema.
- Cremers A, Elsen A, de Preter P, and Maes A (1988) Quantitative analysis of radiocesium retention in soils. *Nature* 335: 247–249.
- Delvaux B, Kruyts N, and Cremers A (2000) Rhizospheric mobilization of radiocesium in soils. *Environmental Science and Technology* 34: 1489–1493.
- Desmet G, Nassimbeni P, and Belli M (eds.) (1990) *Transfer of Radionuclides in Natural and Semi-Natural Environments*. UK, London: Springer Dordrecht.
- Ishii N, Tagami K, Enomoto S, and Uchida S (2004) Influence of microorganisms on the behavior of technetium and other elements in paddy soil surface water. *Journal of Environmental Radioactivity* 369–380. England.
- Keith-Roach MJ and Livens FR (eds.) (2002) *Interactions of Microorganisms With Radionuclides*. The Netherlands, Amsterdam: Elsevier Science.
- Lloyd JR (2003) Microbial reduction of metals and radionuclides. *FEMS Microbiology Reviews* 27: 411–425.
- Lloyd JR, Nolting HF, Sole VA, and Bosecker K (1998) Technetium reduction and precipitation by sulfate-reducing bacteria. *Geomicrobiology Journal* 15: 45–58.
- Lloyd JR, Sole VA, van Praagh CVG, and Lovley DR (2000) Direct and Fe(II)-mediated reduction of technetium by Fe(III)-reducing bacteria. *Applied and Environmental Microbiology* 66: 3743–3749.
- Staunton S, Dumat C, and Zsolnay A (2002) Possible role of organic matter in radiocaesium adsorption in soils. *Journal of Environmental Radioactivity* 58: 163–173.
- Stolz JF and Oremland RS (1999) Bacterial respiration of arsenic and selenium. *FEMS Microbiology Reviews* 23: 615–627.
- Tykva R and Berg D (eds.) (2004) *Man-Made and Natural Radioactivity in Environmental Pollution and Radiochronology*. The Netherlands, Dordrecht: Springer Science.
- UNSCEAR (ed.) (2000) *Sources and Effects of Ionizing Radiation*, In: vol. 1. Sources. USA, United Nations, New York.
- Zapata F (2003) The use of environmental radionuclides as tracers in soil erosion and sedimentation investigations: Recent advances and future developments. *Soil & Tillage Research* 69: 3–13.

Silicon and other non-metal elements from Group 14

Bruno Delvaux and Zimin Li, Université catholique de Louvain (UCLouvain), Earth and Life Institute, Soil Science, Louvain-La-Neuve, Belgium

© 2023 Elsevier Ltd. All rights reserved.

Introduction	263
Inorganic C forms in soil	263
Carbon–silicon interactions	265
Silicon in the soil–plant system	265
Distribution of silicon	265
Mineral weathering and clay formation	267
Adsorption of H_4SiO_4 onto oxide surfaces	268
Silicon in soil impacts pH buffering	270
Terrestrial biological cycle of silicon	271
Germanium	271
Germanium is analogous to silicon	271
Germanium in the soil–plant system	272
The atomic Ge/Si ratio as a powerful biogeochemical proxy	272
Conclusion	273
References	273

Abstract

Carbon (C), silicon (Si) and germanium (Ge) are members of the Group 14 elements; C is a non-metal while Si and Ge are metalloids. All have a strong propensity for covalent bonding. Their oxides (CO_2 , SiO_2 , GeO_2) dissolve to form weak acids in aqueous media. Carbon is abundant in the atmosphere and essential to life. Silicon is ubiquitous in the Earth's crust and beneficial to plants, while Ge is a non-essential microelement. Inorganic C, Si and Ge constitute a trilogy of great interest in soil science and biogeochemistry: C and Si interact closely in terrestrial, marine and global biogeochemical cycles, while Ge is a relevant tracer for Si in these cycles.

Key points

- Inorganic C forms in soils
- Role of carbonate and silicate minerals in soil buffering and acid neutralizing capacity
- Fate of silicon in the soil–plant system (weathering, clay formation, silica sorption, silica uptake)
- Biogeochemical cycle of silicon and carbon–silicon interactions
- Germanium as a tracer of silicon cycles

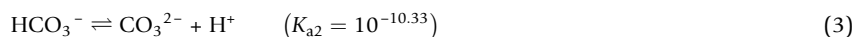
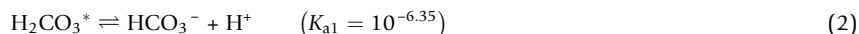
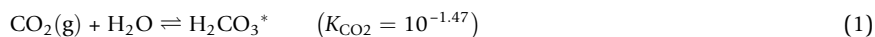
Introduction

The Group 14 elements have a ns^2np^2 electronic configuration, and they show an increasing metallic character from the lightest to heaviest. Carbon (C) is a non-metal while silicon (Si) and germanium (Ge) are metalloids, and all have a strong propensity for covalent bonding. Their oxides (CO_2 , SiO_2 , GeO_2) dissolve to form weak acids in aqueous media ($\text{pK}_a \geq 6.4$). Their exceptional properties make both natural and artificial intelligence possible: C is at the center of natural intelligence and life, while Si and Ge are at the core of artificial intelligence and electronic technology. Carbon is abundant globally and is essential for life itself; it forms so many natural compounds that it makes up the domain of organic chemistry. Silicon is ubiquitous in the Earth's crust, is beneficial to plants, essential to diatoms, and crucial for smart agriculture, while Ge is a microelement used to trace Si fluxes in ecosystems. This chapter deals with the inorganic chemistry of these elements in soil, with a focus on solid–liquid interactions in the soil–plant system.

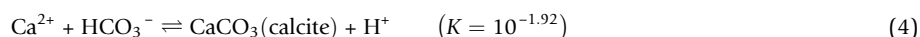
Inorganic C forms in soil

The addition of C substances to soil occurs chiefly through the deposition of dead biomass from plant photosynthetic production. In soil, C substances are consumed by living organisms whose respiration produces CO_2 and CH_4 in aerobic and anaerobic conditions, respectively. Carbon dioxide (CO_2) is widespread in the gaseous phase of well-drained soils, but it can give way to methane (CH_4) in poorly-drained, reduced soils. Wetland soils constitute the largest source of CH_4 entering the atmosphere, emitting about one-third of the global total; rice paddy soils are the major emitters (Sposito, 2016).

The presence of CO_2 in soil air can have a significant effect on soil acidity and mineral weathering by releasing H^+ ions through the following reactions:



where non-volatile carbonic acid H_2CO_3^* designates the sum of H_2CO_3 and aqueous CO_2 , and K is the equilibrium constant at 25 °C (Lindsay, 1979). Reactions (1) to (3) show that H_2CO_3^* is a major source of protons in the soil solution, making HCO_3^- (bicarbonate) an ubiquitous soil solution species under aerobic conditions because it is the major aqueous carbonate species in dilute solution (Stumm and Morgan, 1996). The degree of CO_2 impact on proton release, however, depends on (i) temperature, (ii) CO_2 partial pressure (P_{CO_2}), (iii) proton consumption and or HCO_3^- and CO_3^{2-} leaching, (iv) precipitation of carbonates and or evaporites (Stumm and Morgan, 1996). These four points are briefly discussed hereafter. (i) A decrease in temperature shifts reaction (1) to the right ($K_{\text{CO}_2} = 10^{-1.2}$ at 5 °C) (Stumm and Morgan, 1996). (ii) At atmospheric air pressure, $P_{\text{CO}_2} \approx 10^{-3.52}$ atm ($10^{-1.51}$ kPa), pH = 5.7 in deionized water, but once P_{CO_2} reaches $\approx 10^{-2}$ atm (1.01 kPa) from, for example, intense biological activity in soil, pH = 4.9. (iii) Proton (H^+) consumption by mineral weathering or H^+ adsorption on the soil exchange complex, and leaching of HCO_3^- and or CO_3^{2-} (carbonate) anions, move reactions (2) and (3) to the right, favoring additional H^+ release. Proton adsorption by clay minerals induces their dissolution and aluminum (Al) release. (iv) The precipitation of CaCO_3 (calcium carbonate) occurs as follows:



In semiarid or arid environments, where evaporation exceeds rainfall and accounts for major soil water loss, oversaturation of the soil solution with CO_3^{2-} and or HCO_3^- ions (depending on pH) results in the precipitation of secondary calcium (Ca) and or magnesium (Mg) carbonates (Sposito, 2016), e.g., calcite (CaCO_3) as shown by reaction (4), dolomite ($\text{CaMg}(\text{CO}_3)_2$), magnesite (MgCO_3), and or evaporites, e.g., nahcolite (NaHCO_3), natrite (Na_2CO_3). Reactions (1), (2), and (4) indicate that the precipitation of CaCO_3 (calcite) depends on P_{CO_2} , $[\text{Ca}^{2+}]$ and pH, where [solute] means solute concentration. At common P_{CO_2} and $[\text{Ca}^{2+}]$ values, calcite precipitates at pH around 7.8–8.5. Oversaturation of the soil solution enhances the formation of carbonates and evaporites, increasing the acid neutralizing capacity (ANC) of soils (van Breemen et al., 1983), which then undergo net alkalization (Fig. 1a), which arises in some Solonchaks, Gypsisols, Calcisols and Durisols (IUSS, 2015).

Conversely, Ca and Mg carbonates can be inherited from the soil parent material and dissolve, as described in the reverse reaction (4), in soils under wetter conditions. In environments where evaporation balances rainfall, both dissolution and precipitation processes may occur within the same annual cycle. An exemplary case concerns the seasonal dynamics of CaCO_3 in some soils under semiarid or subhumid continental environments, e.g., certain Calcisols, Kastanozems, Chernozems and Vertisols (IUSS, 2015). In these environments, spring moisture and increased temperature stimulate biological activity in the topsoil, resulting in an increase in P_{CO_2} , which moves reactions (1) and (2) to the right, leading to CaCO_3 dissolution as described in the reverse reaction (4). The solutions enriched in HCO_3^- and or CO_3^{2-} ions migrate to the subsoil horizons, but they are driven

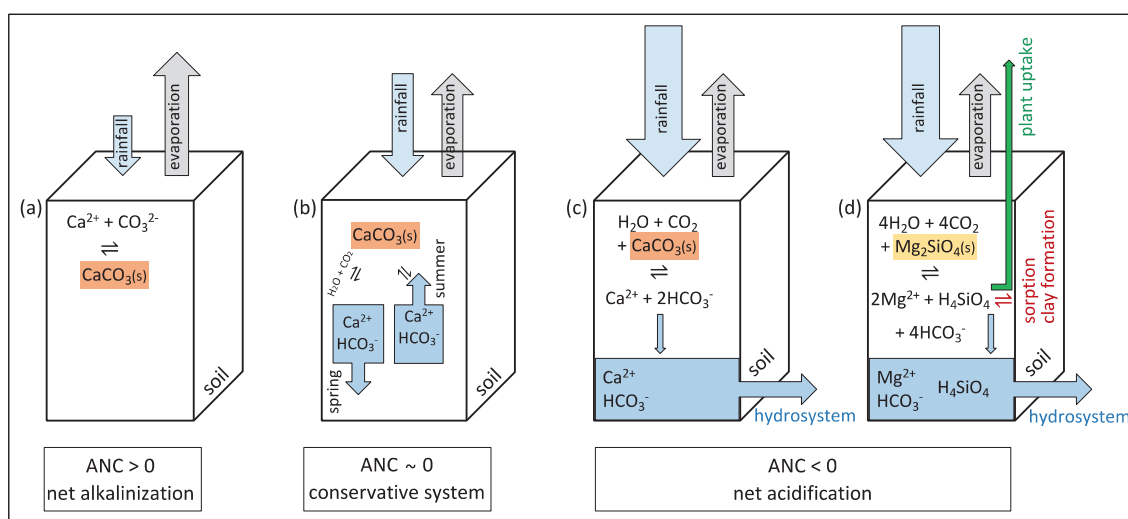


Fig. 1 (a–c) Schematic representation of the impact of the annual rainfall–evaporation balance on soil carbonate equilibria, illustrating the role of CaCO_3 as a privileged proton consumer that can protect other weatherable minerals of lower solubility and/or dissolution rate in well-drained soils (c). (d) Same representation illustrating silicate dissolution (forsterite): aqueous H_4SiO_4 can be leached and involved in precipitation or sorption processes, and is taken up by plants. The Acid Neutralizing Capacity (ANC) is here evaluated on an annual basis as the difference $[\text{ANC}_{\text{yr}+1} - \text{ANC}_{\text{yr}}]$ (van Breemen et al., 1983).

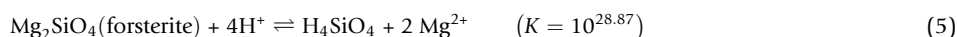
upwards by capillarity to upper horizons with water loss in the dry and hot summer. Then, oversaturation results in CaCO_3 precipitation along pore walls, giving rise to particular forms of deposits (soft powdery lime, pseudomycelium) and partial pore clogging (IUSS, 2015). This process keeps the ANC constant on an annual basis (Fig. 1b), making these soils geochemically conservative, where calcite may appear as a primary mineral as well as pedogenic CaCO_3 .

In well-drained soils under sub-humid and humid environments where rainfall exceeds evaporation, the aqueous alkalinity, which is largely attributed to HCO_3^- activity, is leached out with base cations and further transferred to watersheds and rivers, thus into the hydrosystem (Fig. 1c). This results in a decrease in ANC and thus soil net acidification, while soil solution pH remains constant as long as H^+ ions are consumed by CaCO_3 (van Breemen et al., 1983).

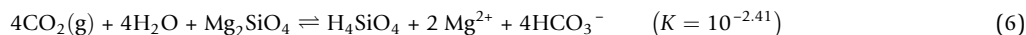
Thus, considering Dokoutchaïev's global zonality, the latitudinal distribution of soils derived from calcareous parent materials depends strongly on the fate of carbonate in both its precipitated and aqueous forms, which are principally controlled by water balance and biological activity. Consequently, from arid and semi-arid environments to subhumid and humid environments, the soil cover changes globally from e.g., Solonchak, Calcisol to Kastanozem, Chernozem, and eventually to decarbonated Phaeozem, Umbrisol, Luvisol, as illustrated schematically in Fig. 1a–c (IUSS, 2015). In carbonated soils, CaCO_3 can be seen as a privileged proton consumer, the protective effect of which can retard the dissolution of weatherable silicate minerals, which have very slow dissolution kinetics (Fig. 1c–d).

Carbon–silicon interactions

Inorganic carbon forms in the soil solid, liquid and gaseous phases are thus clearly pivotal in the processes of net acidification and alkalization of soil, and in pH buffering (Fig. 1a–c). Atmospheric CO_2 concentration can be regulated by silicate weathering that transforms CO_2 into carbonate alkalinity and releases monosilicic acid (H_4SiO_4), as illustrated for forsterite, a primary mineral representing the Mg endmember of olivine (Fig. 1d), following the reactions below:



Combining reactions (1), (2) and (5) gives:



Thus, 4 mol $\text{CO}_2(\text{g})$ react with 1 mol forsterite to form 1 mol H_4SiO_4 and release aqueous bicarbonate and base cation, the leaching of which moves reaction (6) to the right, which emphasizes the contribution of weatherable silicate minerals to ANC (Fig. 1d). The global biogeochemical cycles of C and Si thus interplay through silicate weathering on continents and C fixation by diatoms in marine ecosystems (Erhart, 1963). Silicate weathering consumes CO_2 if soil conditions enhance the dissociation of H_2CO_3^* and solute leaching, exporting aqueous H_4SiO_4 toward watersheds, rivers and oceans (Fig. 1d). In marine ecosystems, siliceous diatoms take up Si as a nutrient, precipitating approximately $6.7 \cdot 10^{12}$ kg Si yr^{-1} , while fixing CO_2 , which contributes roughly half of net primary production of C ($26 \cdot 10^{12}$ kg C yr^{-1}) (Conley, 2002). Thus, the annual mobilization of Si and C by marine diatom biomass corresponds to a Si:C molar ratio of ~ 0.11 mol mol^{-1} . In terrestrial ecosystems, the global net primary production by photosynthesis is about $60 \cdot 10^{12}$ kg C yr^{-1} while plants accumulate Si in the range of ~ 1.6 – $5.6 \cdot 10^{12}$ kg Si yr^{-1} , noting that roughly 55% of the photosynthetic production is attributed to plants accumulating Si actively (Conley, 2002). The annual mobilization of C and Si by plants corresponds to a Si:C molar ratio ranging between ~ 0.012 and ~ 0.04 mol mol^{-1} .

Given that the biological cycle of Si is critical in primary productivity and C cycling in both terrestrial and marine ecosystems (Conley and Carey, 2015), and that Si uptake by plants improves their vigor and health (Coskun et al., 2019), the fate of inorganic C and Si forms in soil is at the core of major challenges facing global change and sustainable agriculture. Fig. 2 shows that the C and Si cycles have a connecting point in soil once protons released by H_2CO_3^* dissociation are consumed by silicate weathering. This connection also holds for other proton donors released during the aerobic decomposition of dead biomass, such as organic and nitric acids, and also, but more indirectly, by the extrusion of protons from plant roots (van Breemen et al., 1983). While inorganic C can be present in all three soil phases (Figs. 1 and 2), forms of Si are distributed exclusively between the liquid and solid phases (Fig. 2). The latter hosts an impressive diversity of Si-bearing minerals, whereas H_4SiO_4 is the quasi-unique Si species in the soil solution. The large $\text{pK}_{\text{a}1}$ value of the first ionization of H_4SiO_4 ($\text{pK}_{\text{a}1} = 9.8$ – 9.9) largely restricts its dissociation: about 90% of dissolved monosilicic acid is undissociated at pH ~ 9 . This suggests that, unlike H_2CO_3^* and its dissociated products, H_4SiO_4 does not contribute to ANC at the common pH values of soil solutions (van Breemen et al., 1983). Another major fact illustrated in Fig. 2 is that, unlike inorganic C, Si is substantially taken up from soil solution by vegetation in amounts that are similar to macronutrient concentrations in plants, but without being substantially incorporated into organic substances (Coskun et al., 2019). Instead, silica mostly precipitates in plant tissues as phytoliths, which are amorphous biogenic opal-A ($\text{SiO}_2 \cdot n\text{H}_2\text{O}$).

Silicon in the soil–plant system

Distribution of silicon

Silicon is the second most mass-abundant element ($\sim 31\%$) in the Earth's crust after oxygen ($\sim 49\%$) and before aluminum (Al) ($\sim 8\%$) and iron (Fe) ($\sim 5\%$). By volume, over 90% of the Earth's crust is made up of SiO_2 and silicates. Silicon is thus omnipresent

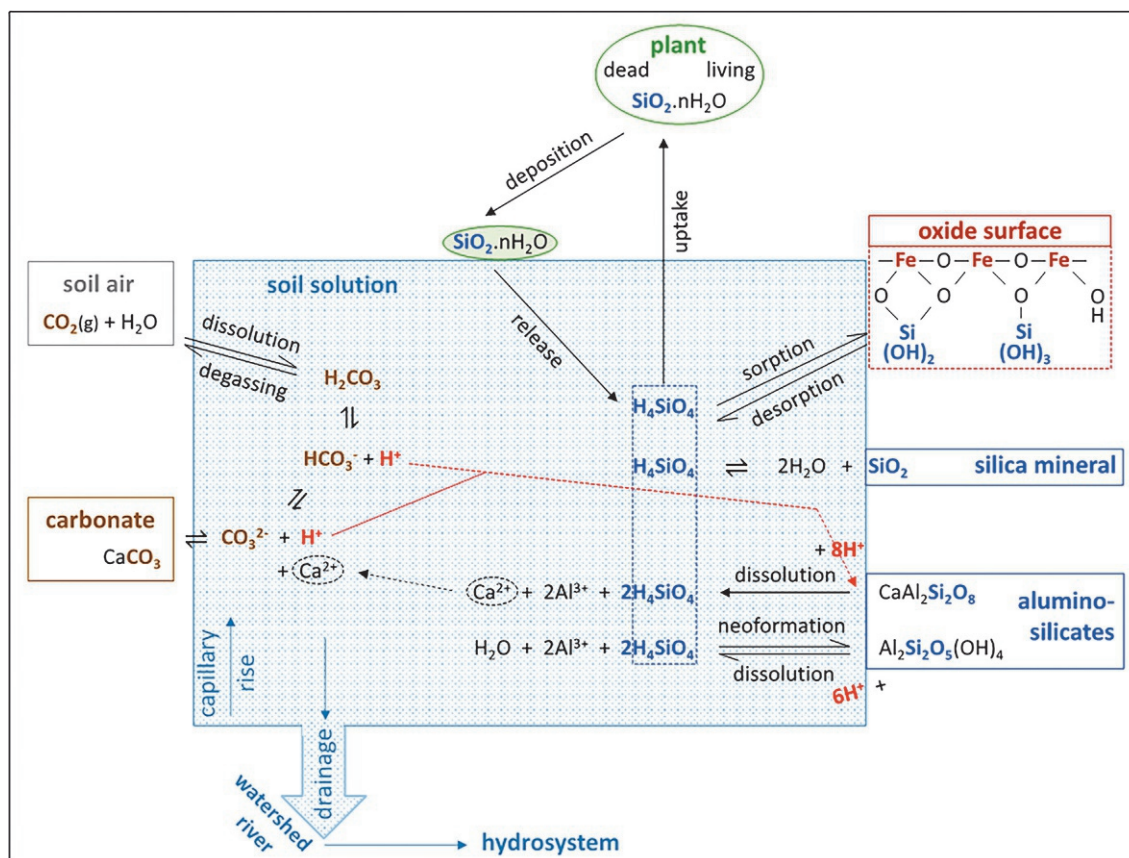


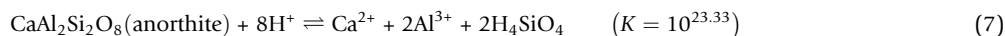
Fig. 2 Schematic view illustrating some major aspects of the fate of inorganic forms of C and Si in a freely drained soil undergoing net acidification ($\text{ANC} < 0$). Inorganic C (left) occurs in the three soil phases while Si (right) is present exclusively in the liquid and solid phases. Aqueous H_4SiO_4 can follow four routes: mineral neoformation, sorption by oxides, uptake by biota, transfer to the hydrosystem. The Al-Si minerals encompass the aluminosilicates, here anorthite ($\text{CaAl}_2\text{Si}_2\text{O}_8$) and kaolinite ($\text{Al}_2\text{Si}_2\text{O}_5(\text{OH})_4$). Plant $\text{SiO}_2 \cdot n\text{H}_2\text{O}$ refers here to phytolith (opal-A), a biogenic Si form. The reactive surfaces that sorb aqueous H_4SiO_4 are figured here by an iron oxide, for example goethite ($\alpha\text{-FeOOH}$).

in nature and ubiquitous in rocks, sediments and soils, where it occurs in a huge number of silica, silicate and aluminosilicate minerals (Monger and Kelly, 2002). In these minerals, Si is exclusively in the IV valence state, and it exists only as a silicate tetrahedron, SiO_4^{4-} . In all these minerals, Si^{IV} bonds to the oxygen (O) atom, through which it links with other metals, the most common one being Al. Therefore, Si^{IV} does not bond directly with other Si^{IV} atoms or with other metals. This binding behavior differs greatly from that of C, which can form direct bonds with other C atoms. In soil, Si occurs exclusively in its solid and liquid phases. The total Si content depends on soil mineralogy and ranges from 5 to 470 g kg^{-1} whereas the content of aqueous Si varies from $3 \cdot 10^{-3}$ to $4.5 \cdot 10^{-1} \text{ g kg}^{-1}$ (Cornelis and Delvaux, 2016, and references therein). Soil Si-bearing minerals encompass silicate and silica minerals, both crystalline and amorphous. Silicate minerals involve various primary and secondary minerals, among which aluminosilicates predominate globally, particularly as clay minerals either inherited from soil parent material, transformed or formed during pedogenesis. Silica minerals include the SiO_2 polymorphs (e.g., cristobalite, tridymite), quartz varieties and some opals (Monger and Kelly, 2002). Non-mineral forms of silica include biogenic opals (opal-A, $\text{SiO}_2 \cdot n\text{H}_2\text{O}$), which are widespread in soils (phytoliths) and riverine and marine ecosystems (diatoms, sponge spicules, radiolaria) (Monger and Kelly, 2002). In particular, phytoliths form in plant tissues following the root uptake of aqueous H_4SiO_4 from soil solution, its transfer to transpiration termini where it precipitates as biogenic opal (Coskun et al., 2019). The content of phytolithic Si in soils is estimated to range from 1 to 10 g kg^{-1} worldwide, but may be higher locally, although rarely exceeding 70 g kg^{-1} (Meunier et al., 2014).

Fig. 2 illustrates schematically, in its right part, the fate and distribution of Si in soils. This graph shows that the primary source of Si in the soil–plant system is the pool of weatherable Si-bearing primary minerals, represented here by anorthite, the Ca-plagioclase ($\text{CaAl}_2\text{Si}_2\text{O}_8$), the dissolution of which releases H_4SiO_4 . Fig. 2 shows that aqueous H_4SiO_4 , generally designated as dissolved silicon (DSi), can follow four routes: (i) coprecipitation with Al to form clay minerals (e.g., kaolinite ($\text{Si}_2\text{Al}_2\text{O}_5(\text{OH})_4$) in Fig. 2), (ii) adsorption onto the surfaces of Al and Fe (hydr)oxides, (iii) absorption by plants (or by other biota), where it forms biogenic silica, (iv) transfer to watersheds and rivers. These routes are complex pathways characterized by feedback and/or reverse processes.

Mineral weathering and clay formation

In the early stages of soil formation, the weathering of primary silicate minerals is the mandatory initial step that releases DSI into the soil solution. It begins with mineral dissolution, which is induced by protons or ligands that form strong complexes with metals (Sposito, 2016). Primary mineral dissolution is illustrated in Fig. 2 by the acid dissolution of anorthite following the reaction:



This reaction is promoted by the leaching or biological uptake of Ca^{2+} , the hydrolysis of detached Al, which produces H^+ ions, the complexation of Al with inorganic or organic ligands, the removal of Si through leaching or formation of secondary minerals, i.e., smectite, kaolinite and gibbsite through the respective reactions below:

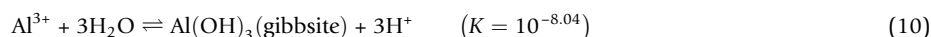
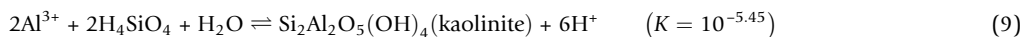
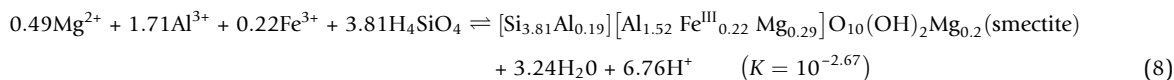


Fig. 3 illustrates the solubility lines of anorthite and the above-mentioned secondary minerals in a stability diagram using gibbsite ($\text{Al}(\text{OH})_3$) as the reference stable mineral ($\log(\text{Al}^{3+}) + 3\text{pH} = 8.04$). The concentration of aqueous H_4SiO_4 , designated as $[\text{H}_4\text{SiO}_4]$, is kinetically controlled by the dissolution of primary silicate minerals (e.g., anorthite) and the precipitation–dissolution of secondary minerals that are more soluble than quartz ($\text{SiO}_2(\text{quartz})$; $[\text{H}_4\text{SiO}_4] = 10^{-4} \text{ M}$), and less soluble than amorphous silica ($\text{SiO}_2 \cdot n\text{H}_2\text{O}$; $[\text{H}_4\text{SiO}_4] = 10^{-2.7} \text{ M}$). Reaction (8) requires the presence of aqueous Mg^{2+} and ferric (Fe^{3+}) ions as reactants that integrate the octahedral sheet of the smectite formed. Magnesium and Fe^{3+} ions can be released by the dissolution of primary minerals, e.g., ferromagnesian silicates such as olivine, pyroxene, amphibole or biotite. Reactions (8) to (10) show that $[\text{H}_4\text{SiO}_4]$ controls the nature of the secondary mineral formed (Fig. 3). The highest H_4SiO_4 concentration promotes the formation of smectite (reaction (8)), while its decrease favors the synthesis of kaolinite (reaction (9)) and finally gibbsite in strongly desilicated environments (reaction (10)), where H_4SiO_4 is strongly leached out and transferred to the hydrosystem (Figs. 1d and 2). Consequently, H_4SiO_4 concentration, which is rainfall-dependent, is one of the properties controlling the zonal distribution of soils notably in intertropical regions: in the wet tropics, highly weathered soils containing kaolinite, gibbsite and Fe (hydr)oxides predominate, while smectite-rich soils occur in subhumid environments, characterized by a marked dry season, which is conducive

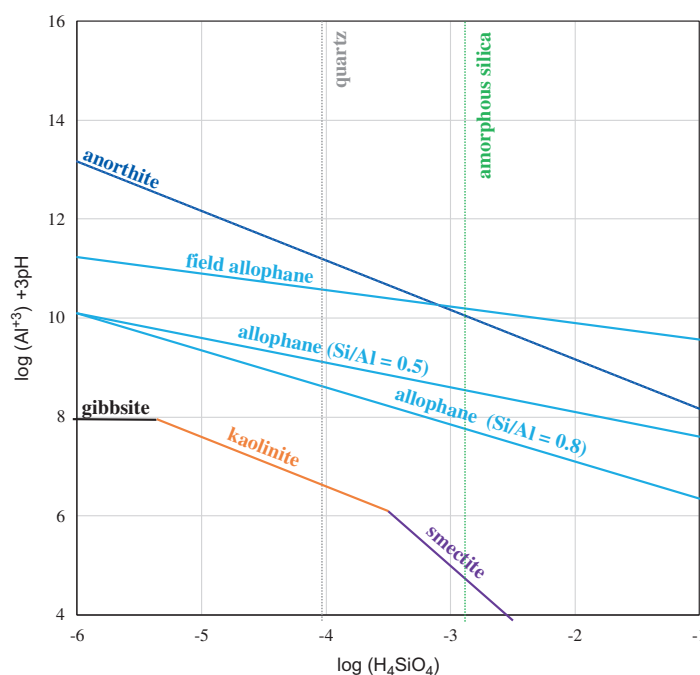
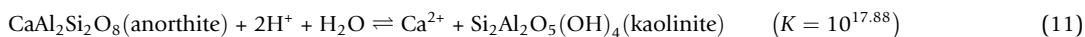
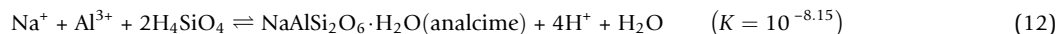


Fig. 3 Stability diagram for reference anorthite, gibbsite, kaolinite, smectite and allophanes in the $\text{Al}_2\text{O}_3\text{--SiO}_2\text{--H}_2\text{O}$ system (Li et al., 2022). The solubility lines for anorthite, gibbsite, kaolinite, smectite, quartz and amorphous silica refer to Lindsay (1979). The allophane solubility lines were compiled for field allophane and synthetic allophanes with Si/Al ratio of 0.5 and 0.8 mol mol⁻¹, respectively, using the *K* values of reactions (15)–(17) (Dobrzyński, 2006). Following parameters were fixed when required: pH = 6.0; $[\text{Ca}^{2+}] = [\text{Mg}^{2+}] = 10^{-3} \text{ M}$; $\log[\text{Fe}^{3+}] = 2.7\text{--}3\text{pH}$.

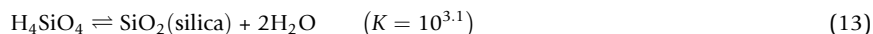
to the concentration of DSi in the soil solution. Combining reactions (7) and (9) formulates both the dissolution of anorthite and the synthesis of kaolinite in a single reaction:



In soils with a constant annual ANC and/or undergoing net alkalization ($\text{ANC} \geq 0$), hydrolysis reactions can be reversible without leading to precipitation of the original primary mineral. Other secondary silicates can be formed such as pedogenic zeolites in salt-affected, alkaline soils, e.g.:

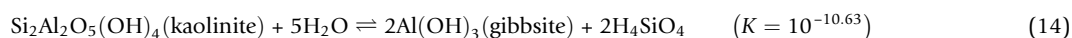


Pedogenic zeolites are, however, rare in soils. Under Mediterranean and/or semiarid conditions with enough water for weathering but not enough for intensive leaching ($\text{ANC} \sim 0$), silica may precipitate:



As with pedogenic carbonate, pedogenic silica precipitates in soil voids and contributes to pore clogging as observed in Durisols (IUSS, 2015), where its accumulation leads to cementation.

In soils undergoing net acidification, the dissolution–precipitation process (e.g., reaction (11)) is irreversible. Furthermore, any decrease in H_4SiO_4 concentration and pH makes the secondary minerals formed unstable. In such conditions, they can dissolve. Their respective dissolutions are formalized through the reverse reactions (8) and (9), which release aqueous H_4SiO_4 . Combining the reverse reaction (9) with reaction (10) further illustrates the dissolution of kaolinite and subsequent formation of gibbsite in humid, desilicated environments:



Thus, secondary crystalline clay minerals can act as sink and source of DSi in soil, as shown in Fig. 2 for kaolinite. The same is true for amorphous aluminosilicates such as allophanic substances following the reactions formulated below, for field allophane (reaction (15)) and synthetic allophanes with Si/Al atomic ratio 0.5 and 0.8, respectively (reactions (16) and (17)) (Dobrzyński, 2006):

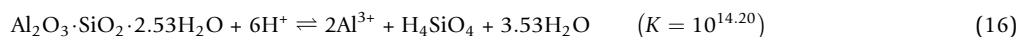


Fig. 3 shows that, within usual ranges of pH values and $[\text{H}_4\text{SiO}_4]$ (10^{-5} – 10^{-3} M), allophanes are metastable with respect to smectite and kaolinite (Dobrzyński, 2006). In particular, at H_4SiO_4 concentration around 10^{-3} M, the solubility lines are narrowly aligned for both the anorthite and field allophane reported here. Experimental data suggest that allophanes do not have a true thermodynamic stability field in a phase diagram. Overall, allophanic substances are considered transient aluminosilicate phases in soil weathering processes (Dobrzyński, 2006). They are therefore an important potential source of DSi (Li et al., 2022) due to their metastable character and rapid dissolution, especially under acid conditions (reactions (15)–(17)). Their rate of dissolution is one order of magnitude higher than that of crystalline phases of similar composition. These transient phases are also considered to be the highly labile hydroxyaluminosilicate (HAS), which may dissolve and precipitate following seasonal dynamics; they play a central role in controlling H_4SiO_4 concentration as an alternative source and sink for DSi in soil, as observed in some grassland soils within the same year (White et al., 2012).

Adsorption of H_4SiO_4 onto oxide surfaces

The concentration of H_4SiO_4 in soil solution is primarily determined by the precipitation and dissolution of silicate minerals. However, it may be affected by its sorption onto secondary Al and Fe (hydr)oxides, here referred to as (Al, Fe) oxides. Fig. 2 shows that in addition to clay formation and uptake by biota, H_4SiO_4 may be withdrawn from the soil solution by adsorption onto (Al, Fe) oxides. Gibbsite formation requires the dissolution of aluminosilicates in wet, desilicated environments (reaction (14)), which are widespread in the humid tropics. Iron oxides are ubiquitous in soils at all latitudes, as their formation can follow several routes: mineral weathering, oxidation–reduction, hydrolysis, chelation, protolysis and biomineralization (Bigham et al., 2002). Available data from the literature on the adsorption of H_4SiO_4 concerns mainly gibbsite, goethite and ferrihydrite. The sorption process occurs through the interaction between H_4SiO_4 with the oxide surface OH groups through ligand exchange and the subsequent formation of a covalent Si–O–metal bond, resulting in inner-sphere complexation (Bowden et al., 1973; Sposito, 2016). This process may also occur with kaolinite, but to a considerably lesser extent.

Gibbsite and Fe oxides possess a variable surface charge depending on pH, electrolyte concentration and counterion valence. They can generate both cation and anion exchange capacity, i.e., CEC and AEC, as a consequence of H^+ and OH^- sorption from soil solution following the schematic, simplified reaction (Uehara and Gillman, 1985):

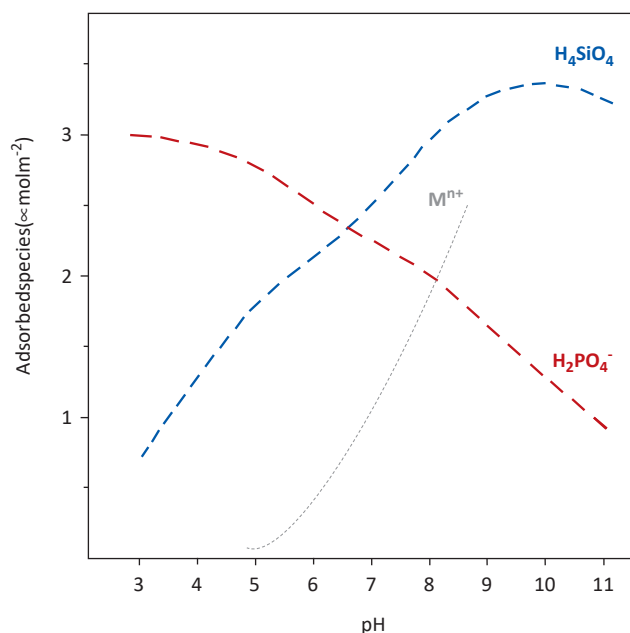
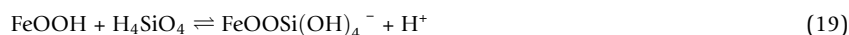


Fig. 4 Illustration of the specific sorption of aqueous H_4SiO_4 , H_2PO_4^- and metal cations M^{n+} on metal oxide surfaces that bear reactive OH groups. As adapted from Bowden JW, Bolland MDA, Posner, AM and Quirk, JP (1973) Generalized model for anion and cation adsorption at oxide surfaces. *Nature Physical Science* 245(145), 81–83.



The pH value at which the total net surface charge is zero is the point of zero charge (p.z.c.) (Sposito, 2016), also termed pH at the isoelectric point (pH_{IEP}). Gibbsite and Fe oxides have a p.z.c. that usually ranges between 7 and 10 (Uehara and Gillman, 1985). Therefore, they develop a positive surface charge, thus AEC, at common pH values in the soil solution. On (Al, Fe) oxide surfaces, the sorption of oxyanions of polyprotic acids, such as phosphate, molybdate, and arsenate, is maximum at low pH and decreases with increasing pH, unlike metal cations as illustrated in Fig. 4 (Bowden et al., 1973) because of the progressive pH-induced reversal of the net surface charge from AEC to CEC.

Fig. 4 shows that the adsorption of H_4SiO_4 peaks at $\text{pH} \sim 9$, and decreases at lower and higher pH values (Bowden et al., 1973; Hiemstra et al., 2007). Unlike phosphate oxyanions, H_4SiO_4 remains undissociated at pH below 9, meaning that its sorption is due only to the interaction of Si with the surface OH groups. Illustrating this process schematically predicts that a decrease in pH leads to a decrease in H_4SiO_4 sorption (Hiemstra et al., 2007):



The inner-sphere complexation of monomeric Si results in the formation of either a monodentate $\equiv\text{FeOH}-\text{FeOSi}(\text{OH})_3$ or a bidentate $\equiv\text{Fe}_2\text{O}_2\text{Si}(\text{OH})_2$ surface complex (Hiemstra et al., 2007; Hiemstra, 2018) (Fig. 2). It involves monomeric Si as the main adsorbed species at a low loading of aqueous Si ($< \sim 0.9$ mM). At higher loading of aqueous Si ($> \sim 0.9$ mM), Si-tetramer can form on goethite surfaces, whereas Si-oligomers (di, tri, or tetramer) form on ferrihydrite surfaces at H_4SiO_4 concentrations below those required for polymerization in solution (Hiemstra et al., 2007; Hiemstra, 2018). The siliceous nature of natural ferrihydrites is likely because of the formation of Si-oligomers at relatively high Si loadings, which can occur in weathering solutions (Hiemstra, 2018). At pH above 9, H_4SiO_4 dissociation releases anionic H_3SiO_4^- as the oxide surface becomes increasingly negatively charged, generating coulombic repulsion and a subsequent decrease in H_4SiO_4 sorption (Fig. 4).

The specific adsorption of H_4SiO_4 onto (Al, Fe) oxides is poorly reversible. As a result of Si sorption, the release of protons from the Si-enriched (Al, Fe) oxides generates a downward shift of the p.z.c., thus increasing the CEC, which also occurs for phosphate oxyanions and other specifically adsorbed solutes (Uehara and Gillman, 1985). Desorption of Si may occur in the presence of competitive solutes that are also specifically adsorbed onto (Al, Fe) oxides. The specific sorption sites of these soil oxides are subject to competition between different adsorptive aqueous species. In addition to H_4SiO_4 , some solutes of environmental and agronomic interest are, for example, aqueous organic anions, arsenate, fluoride, phosphate and molybdate. Arsenate and fluoride fixation onto (Al, Fe) oxides detoxifies soils and waters. Molybdenum is a plant micronutrient and is crucial for the optimal development of rhizobial symbiosis, a bioprocess that improves the nitrogen status of tropical legume crops. The competition between solutes for absorbing sites suggests that preloading the oxide surface with silicate may depress the sorption of competitive anions, depending on adsorbent selectivity and the surrounding conditions, notably pH and electrolyte concentration (Uehara and Gillman, 1985; Sposito, 2016). Conversely, organic and inorganic anions that are selectively adsorbed may displace Si and increase its mobility.

Silicate sorption can be exploited to amend oxide-rich soils (Ferralsols) to increase CEC and better regulate the bioavailability of phosphorus (P), P fixation being a major constraint in these soils (Uehara and Gillman, 1985), particularly at low pH (Fig. 4).

Aluminum and Fe oxides have been claimed to affect the fate of DSi strongly in soils because their surface OH groups interact specifically with H_4SiO_4 . This statement must be moderated by the three processes discussed above: (i) the weak reversibility of the process may limit subsequent Si sorption, (ii) the change of oxide composition due to inner-sphere complexation by the Si–O–metal bond shifts the value of the p.z.c. downwards, which also reduces further Si sorption, (iii) various solutes strongly compete with H_4SiO_4 for adsorption, which may significantly hamper the Si sorption process. On the other hand, the magnitude of the process depends on both the nature and content of (Al, Fe) oxides in soils. The oxide surface reactivity for oxyanions, weak acids and water is well known to decrease with increasing oxide crystallinity (Bigham et al., 2002). As crystallinity increases, oxide crystals become larger and surface area decreases. Thus, poorly crystalline, fine-sized Al and Fe oxides have a much larger specific surface area than their crystalline counterparts and adsorb larger amounts of H_4SiO_4 . In the range of $[\text{H}_4\text{SiO}_4]$ of 10^{-4} – $10^{-2.7}$ M, pure, poorly crystalline, fine-sized Fe oxide such as ferrihydrite would adsorb ~ 1 – $25 \mu\text{mol Si kg}^{-1}$ at pH 5.5. For instance, within a range of 1–10% ferrihydrite content in the soil, Si adsorption would theoretically amount to 0.01 – $2.5 \mu\text{mol Si kg}^{-1}$ soil, yielding a small amount of adsorption despite the absence of competitive adsorptive species. However, given the affinity of the surface OH groups of (Al, Fe) oxides for Si, the soil Si sorption capacity has been proposed as an index revealing both the nature and magnitude of soil surface properties, linked to their stage of weathering (Gallez et al., 1977), since end-stage weathering results in the relative accumulation of (Al, Fe) oxides for given parent materials. More generally, weathering promotes the relative accumulation of quartz and secondary (Al, Fe) oxides depending on the composition of soil parent material. Therefore, it is worth considering the effect of Si chemistry in soil on pH buffering.

Silicon in soil impacts pH buffering

Unlike (Al, Fe) oxides (p.z.c. ~ 7 – 10), Si oxides have a very low p.z.c. (2 – 2.5) (Monger and Kelly, 2002; Uehara and Gillman, 1985). These huge differences explain why highly weathered soils such as Acrisols and Ferralsols (IUSS, 2015), although both are strongly desaturated in base cations, differ widely in pH. The Nernst equation predicts that the p.z.c. value is the pH at which the surface potential is zero, corresponding to the point of maximal chemical stability (Uehara and Gillman, 1985). Thus, highly weathered, variable-charge soils tend to have pH values around their p.z.c., the value of which results from their contents of organic matter, kaolinite, (Al, Fe) oxides and residual quartz. This tendency for soil pH to drift toward the p.z.c. (pH_{IEP}) has been termed ‘isoelectric weathering’ by Mattson in 1932 (Uehara and Gillman, 1985). In well-drained soils undergoing net acidification ($\text{ANC} < 0$), weathering processes ultimately lead to the relative accumulation of residual quartz or secondary oxides (Al, Fe) depending on whether they are derived from siliceous felsic or mafic rocks. In the first case, highly weathered siliceous soils (Acrisols, Podzols) with low p.z.c. tend to be strongly acid ($\text{pH} \leq 4.5$), whereas highly leached soils rich in (Al, Fe) oxides with higher p.z.c. tend to be sub-neutral ($\text{pH} \sim 5.5$ – 6.5), as exemplified in some Brazilian Ferralsols or Hawaiian perhydrated Andosols (Uehara and Gillman, 1985). This is illustrated in Fig. 5, which schematically represents a hypothetical titration curve of theoretical soil materials illustrating the two routes described above: a decrease and an increase in pH in highly weathered soils depending on the nature of their dominant components. This graph also illustrates the importance of silicates and aluminosilicates, both primary and secondary, in buffering soil pH through their contribution to ANC.

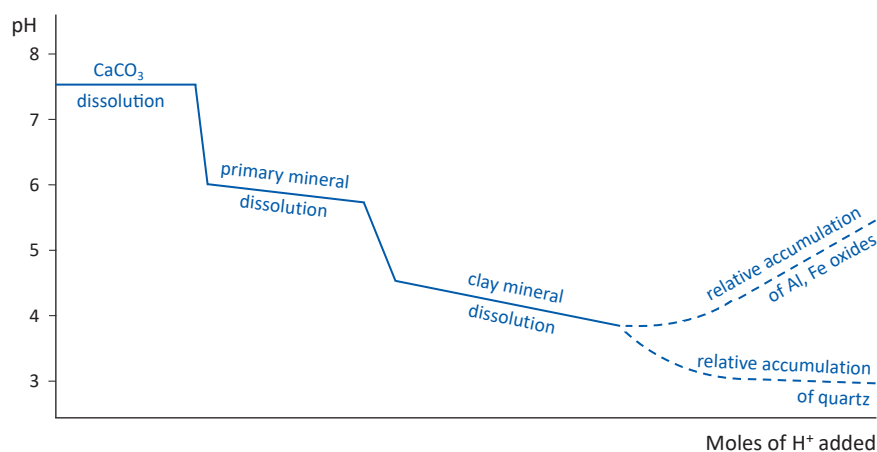


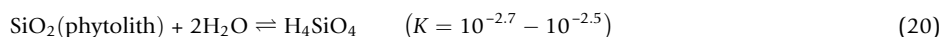
Fig. 5 Schematic illustration of a hypothetical titration curve of theoretical well-drained calcareous soil materials by a strong acid. Progressive proton consumption by soil minerals leads to the dissolution of CaCO_3 , primary minerals, secondary clay minerals, and eventually the relative accumulation of oxides either quartz or (Al, Fe) oxides depending on the nature of soil parent siliceous/felsic or mafic, respectively. The weathering of siliceous/felsic and mafic rocks would indeed lead to the relative accumulation of quartz and (Al, Fe) oxides, respectively, making pH tending to p.z.c. according to the principle of isoelectric weathering (Uehara and Gillman, 1985). As adapted from van Breemen N, Mulder J and Driscoll CT (1983) Acidification and alkalization of soils. *Plant and Soil* 75, 283–308.

Once carbonates have been dissolved (reverse reaction (4)), pH buffering depends mostly on the weathering of primary and secondary silicate minerals (reactions (5), (7) and reverse reactions (8), (9), respectively) together with the adsorption of H^+ ions by clay minerals and organic matter (van Breemen et al., 1983). Reaction (11), which combines anorthite dissolution and kaolinite formation, shows the proton consumption through the transfer of H^+ ions from the weathering solution into the octahedral sheet of the clay mineral formed. At the final stages of weathering, the fate of Si is also critical for pH buffering. The relative accumulation of quartz gives rise to strongly acid soils such as Acrisols and Podzols (IUSS, 2015), whereas the heavy leaching of H_4SiO_4 limits the formation of kaolinite and promotes the relative accumulation of (Al, Fe) oxides. This contributes to the increase in soil pH in some Ferralsols and Andosols (IUSS, 2015), according to the principle of isoelectric weathering (Uehara and Gillman, 1985). In both cases, the soils are highly weathered and depleted in base cations and nutrients, but they largely differ in pH: the former are strongly acid ($pH \leq 4.5$), the latter are sub-neutral ($pH \approx 5.5$ – 6.5). Since the mineral source for mobile Al is lacking, Al toxicity does not occur readily in these highly weathered soils, contrary to a popular myth. Instead, this constraint is associated with clay mineral dissolution (Fig. 5).

Terrestrial biological cycle of silicon

As discussed above and illustrated in Fig. 4, dissolution and precipitation of soil silicate minerals first govern the concentration of aqueous H_4SiO_4 , and therefore the bioavailability of Si. However, plants also exert a control on $[H_4SiO_4]$ and Si fluxes through mineral weathering induced by root uptake, as well as phytolith formation, return and dissolution in soil. Thus, soil processes and Si biocycling interact considerably in soil–plant systems (Cornelis and Delvaux, 2016).

Plants play a crucial role in soil evolution by promoting silicate weathering through cation uptake and proton extrusion by roots, and by forming phytoliths that can dissolve in soils:



Reaction (20) indicates that H_4SiO_4 undersaturation is the driving force for dissolving amorphous silica. But temperature and electrolyte concentration also strongly affect the dissolution kinetics, as observed for quartz (Dove et al., 2008). At 25 °C, the rate of dissolution of phytoliths depends strongly on pH; it increases by two orders of magnitude from pH 3 to 8 (Frayse et al., 2009). At common pH values of soil solutions, the rate of dissolution of phytoliths is 10^2 – 10^4 higher than that of crystalline clay minerals and primary mafic silicates and feldspars (Frayse et al., 2009), but is similar to that of amorphous minerals in soils (Li et al., 2022). However, their dissolution in soils can be affected by both phytolith properties and soil conditions or processes, among others: phytolith size, morphology and specific surface area, entrapment in soil aggregates, protection by Al and Fe coatings (Li et al., 2022, and references therein).

Fig. 2 indicates that plants act as Si sinks and sources in soil–plant systems. However, their quantitative contribution to the biological cycle of silicon in terrestrial ecosystems depends on the mineral reserve of the soil (illustrated in Fig. 2 by anorthite), and therefore on its degree of weathering. The net acidification of soil results in the progressive depletion of primary and secondary silicate minerals (Fig. 5). Therefore, with increasing weathering and soil development, advanced desilication depletes the mineral source of DSi, and hence phytoliths progressively become the major source of bioavailable Si in soil (Lucas et al., 1993; Meunier et al., 1999). In other words, the biological Si feedback loop progressively takes over the Si plant uptake from weatherable primary and secondary minerals.

Paradoxically, Si biocycling is more intense in soils depleted in weatherable minerals, as illustrated in Fig. 6. There are two major reasons for this observation: (i) the depletion of silicate minerals, (ii) the rate of dissolution of phytoliths, which is much greater than that of most soil silicate minerals at common pH values of the soil solution. In this respect, the restitution of phytolithic Si to soil through the deposition of dead biomass may alleviate desilication through the control exerted by (H_4SiO_4) on mineral stability (see for example, Eqs. (7)–(10) and (14)–(17)). Plants may therefore have a dual role in Si cycling in soil–plant systems: (i) they enhance mineral dissolution through root uptake, (ii) they build up the pool of phytolithic Si in soil, which favors the stability of clay minerals through phytolith control on aqueous H_4SiO_4 (Lucas et al., 1993). By converting mineral Si into phytolithic Si, plants increase the mobility of Si in terrestrial ecosystems and alleviate desilication in topsoils (Lucas et al., 1993; Meunier et al., 1999). Although not essential to plants, silicon is therefore an essential link between the mineral and living worlds. Moreover, its assimilation by plants alleviates biotic and abiotic constraints that reduce crop yields in several agrosystems where the management of Si fluxes from soil to plant can be crucial for sustainable farming and smart agriculture.

Germanium

Germanium is analogous to silicon

Germanium (Ge) is widespread in the Earth's crust where it ranks as the 54th most abundant element with a concentration estimated at 1.3–1.6 $\mu g\ g^{-1}$. From the position of Ge in the periodic table, Mendeleev had predicted its existence and some of its properties, and called it 'ekasilicon.' The similarities between Ge and Si are remarkable. They are due to their identical outer electron structures, similar ionic radii (Si^{4+} : 0.04 nm, Ge^{4+} : 0.053 nm), and tetrahedral bond lengths (Ge–O: 0.175 nm; Si–O: 0.164 nm).

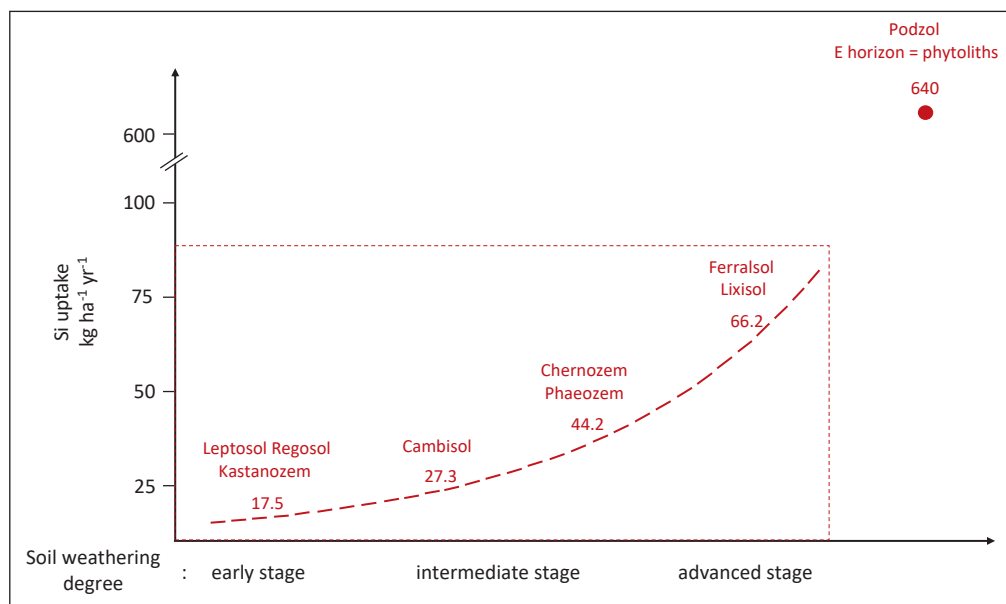


Fig. 6 Schematic illustration of the variation of Si uptake rate by natural vegetation as influenced by soil weathering stage (arbitrary units). The Si uptake rate was measured in the following ecosystems: forest and prairie in Leptosol, Regosol, Kastanozem; forest in Cambisol, prairie in Chernozem and Phaeozem; rainforest in Ferralsol and savanna in Lixisol; bamboo forest in Podzol. The respective figures associated with the respective reference soil groups (IUSS, 2015) are the average values of Si uptake from literature data. The hatched rectangle represents the set of collected data, excluding the exceptional, unusual Si uptake rate measured in a Podzol, in which the eluvial E horizon is a 15-cm-thick phytolithic material developed at the expense of highly weathered trachytic ashes (Meunier et al., 1999). Adapted from Table 1 in Cornelis JT and Delvaux B (2016) Soil processes drive the biological silicon feedback loop. *Functional Ecology* 30(8), 1298–1310. and references cited therein.

Germanates and silicates thus form isostructural compounds in the Earth's crust (Wiche et al., 2018, and references therein). Most importantly, however, trace Ge readily substitutes for Si in the silicate tetrahedron SiO_4^{4-} .

Germanium in the soil–plant system

Being analog to Si, most Ge in the Earth's crust occurs in primary silicate minerals in rocks and eventually in secondary silicates in soils. The variation in Ge concentration ($1.6\text{--}0.9\ \mu\text{g g}^{-1}$) follows that of Si and thus in the following sequence: granites, siliceous rocks > mafic rocks > ultramafic rocks (Wiche et al., 2018). With increasing weathering and soil development, fractionation and sequestration of Ge relative to Si may occur which eventually leads to the formation of Ge-containing secondary minerals, as predicted by reaction (11) for Si. The dissolution of both primary and secondary minerals releases aqueous H_4SiO_4 (Fig. 2; reactions (5), (7), (15), (16), (17); reverse reactions (8), (9)), and H_4GeO_4 , which is thus ubiquitous in trace concentrations in both weathering and soil solutions. Germanic acid (H_4GeO_4) is a weak acid, with a large $\text{p}K_{a1}$ value for the first ionization of H_4GeO_4 ($\text{p}K_{a1} \sim 9.3$). In aqueous solutions, monomeric undissociated H_4GeO_4 is, by far, the dominant inorganic species of Ge, and has a first dissociation constant similar to that of monomeric silicic acid H_4SiO_4 . However, the behavior of aqueous Si and Ge may differ because Ge can form stable complexes with carboxylic acids and phenolic groups in acidic and neutral or basic solutions, respectively (Wiche et al., 2018), likely due to the increased metallic character of Ge compared to Si. Depending on the nature of mineral phases that control $[\text{H}_4\text{SiO}_4]$ in the liquid phase (Fig. 3), the orders of magnitude of $[\text{H}_4\text{SiO}_4]$ and $[\text{H}_4\text{GeO}_4]$ in weathering and soil solutions would be mostly within the ranges $10^{-3}\text{--}10^{-5}$ and $10^{-6}\text{--}10^{-8}$ M, respectively. The root uptake of Ge and Si are similar in hydroponically grown plants. Germanium is not a plant nutrient, nor a strongly toxic element in the environment. However, at concentrations around $1\ \text{mg l}^{-1}$ in the nutrient solution, Ge induces toxic symptoms in Si accumulators such as rice (*Oryza sativa* L.). Plant Ge in biomass is likely to be associated with biogenic Si in phytoliths. Similar to H_4SiO_4 , H_4GeO_4 may be withdrawn from the soil solution by being adsorbed onto secondary (Al, Fe) oxides. Similar to Si, the sorption of Ge onto Fe oxide strongly increases with increasing pH to a maximum at pH 9. In some highly weathered soils in the tropics, the upper iron-rich horizons can be significantly enriched in Ge indicating adsorption on Fe oxides.

The atomic Ge/Si ratio as a powerful biogeochemical proxy

Germanium closely follows the properties and behavior of Si, and is thus considered a tracer of Si in biogeochemical cycles. Overall, Fig. 2 describes the routes followed by both Si and trace Ge in the soil–plant system. Data in the literature extensively illustrate the use of the atomic Ge/Si ratio ($\mu\text{mol mol}^{-1}$) to trace Si in continental, oceanic and global cycles, but also to determine changes in

weathering over time. In this respect, the monitoring of the Ge/Si ratio in the different Si reservoirs can substitute the use of stable Si isotopes (^{28}Si , ^{29}Si , ^{30}Si) to study Si fluxes and cycles. Particularly in terrestrial ecosystems, the Ge/Si ratio gives a sharp distinction between plant uptake and clay formation (Fig. 2).

The Ge/Si ratio is around $2\ \mu\text{mol mol}^{-1}$ in most silicate rocks, but this value covers a substantial range, from 0.5 in quartz to $5\ \mu\text{mol mol}^{-1}$ in some primary aluminosilicates. Compared to this average value, the Ge/Si ratio increases significantly in the secondary soil silicates ($>4\ \mu\text{mol mol}^{-1}$), revealing the preferential partitioning of Ge relative to Si into clay minerals. As a result, weathering and soil solutions, and drainage waters and rivers have typically lower values ($\text{Ge/Si} < 1\ \mu\text{mol mol}^{-1}$). Most plant Ge/Si measurements are performed on extracted phytoliths rather than on bulk tissues. Phytoliths typically have Ge/Si values below $1\ \mu\text{mol mol}^{-1}$; most available Ge/Si values range between 0.003 and $0.5\ \mu\text{mol mol}^{-1}$. The Ge/Si atomic ratio is a highly relevant biogeochemical proxy for tracing Si fluxes and cycles.

Conclusion

The inorganic forms of carbon, silicon, and germanium, all non-metallic elements of Group 14, constitute a trilogy of vital interest in soil science and biogeochemistry because they are at the crossroads of the living and mineral worlds. Carbon and silicon interact closely in terrestrial, marine and global biogeochemical cycles, while germanium is a relevant tracer for silicon in these same cycles. In terrestrial ecosystems, together with soil organic matter, inorganic forms of carbon and silicon exert a major control over the buffering capacity of soils.

The future of research on inorganic C, Si, and Ge dynamics in soil is a challenge beyond the scope of this contribution because of the specificity of each of these elements. Nevertheless, substantial progress could be made by integrating advances in soil science, including biotic and abiotic processes, plant physiology, biogeochemistry, and ecology to improve our overview of C–Si interactions at the ecosystem level while using Ge as a tracer for Si fluxes. Such advances will help to define better the ecosystem services provided by silicon and carbon–silicon interactions in order to propose appropriate management practices to improve the sustainability of agroecosystems.

References

- Bigham JM, Fitzpatrick RW, and Schulze DG (2002) Iron oxides. In: Dixon JB and Schulze DG (eds.) *Soil Mineralogy With Environmental Applications*, pp. 323–366. Madison, Wisconsin, USA: Soil Science Society of America Inc.
- Bowden JW, Bolland MDA, Posner AM, and Quirk JP (1973) Generalized model for anion and cation adsorption at oxide surfaces. *Nature Physical Science* 245(145): 81–83.
- Conley DJ (2002) Terrestrial ecosystems and the global biogeochemical silica cycle. *Global Biogeochemical Cycles* 16: 68–71.
- Conley DJ and Carey JC (2015) Silica cycling over geologic times. *Nature Geoscience* 8: 431–432.
- Cornelis JT and Delvaux B (2016) Soil processes drive the biological silicon feedback loop. *Functional Ecology* 30(8): 1298–1310.
- Coskun D, Deshmukh R, Sonah H, Menzies JG, Reynolds O, Ma JF, Kronzucker HJ, and Bélanger RR (2019) The controversies of silicon's role in plant biology. *New Phytologist* 221(1): 67–85.
- Dobrzyński D (2006) Silica solubility in groundwater from Permian volcanogenic rocks (the Sudetes Mts., SW Poland)—The role of reversible aluminosilicate solids. *Geological Quarterly* 50(4): 407–417.
- Dove PM, Han N, Wallace AF, and De Yoreo JJ (2008) Kinetics of amorphous silica dissolution and the paradox of the silica polymorphs. *Proceedings of the National Academy of Sciences of the United States of America* 105: 9903–9908.
- Erhart H (1963) Sur le cycle de la silice hydratée dans la biosphère. *Comptes Rendus de l'Académie des sciences de Paris* 256: 3731–3734.
- Frayse F, Pokrovsky OS, Schott J, and Meunier JD (2009) Surface chemistry and reactivity of plant phytoliths in aqueous solutions. *Chemical Geology* 258: 197–206.
- Gallez A, Herbillon AJ, Juo ASR, et al. (1977) Characteristics of silica sorption and solubility as parameters to evaluate the surface properties of tropical soils: I. The index of silica reactivity. *Soil Science Society of America Journal* 41(6): 1146–1150.
- Hiemstra T (2018) Ferrihydrite interaction with silicate and competing oxyanions: Geometry and Hydrogen bonding of surface species. *Geochimica et Cosmochimica Acta* 238: 453–476.
- Hiemstra T, Barnett MO, and van Riemsdijk WH (2007) Interaction of silicic acid with goethite. *Journal of Colloid and Interface Science* 310(1): 8–17.
- IUSS Working Group WRB (2015) *World Reference Base for Soil Resources 2014. International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports*. vol. 106. Rome: FAO.
- Li Z, Meunier JD, and Delvaux B (2022) Aggregation reduces the release of bioavailable silicon from allophane and phytolith. *Geochimica et Cosmochimica Acta* 325: 87–105.
- Lindsay WL (1979) *Chemical equilibria in soils*. Chichester, Sussex: John Wiley and Sons Ltd.
- Lucas Y, Luizao FJ, Chauvel A, Rouiller J, and Nahon D (1993) The relation between biological activity of the rain forest and mineral composition of soils. *Science* 260: 521–523.
- Meunier JD, Colin F, and Alarcon C (1999) Biogenic silica storage in soils. *Geology* 27: 835–838.
- Meunier JD, Keller C, Guntzer F, Riotte J, Braun JJ, and Anupama K (2014) Assessment of the $1\% \text{Na}_2\text{CO}_3$ technique to quantify the phytolith pool. *Geoderma* 216: 30–35.
- Monger HC and Kelly EF (2002) Silica minerals. In: Dixon JB and Schulze DG (eds.) *Soil Mineralogy With Environmental Applications*, pp. 611–636. Madison, Wisconsin, USA: Soil Science Society of America Inc.
- Spósito G (2016) *The Chemistry of Soils*, 3rd ed, New York, NY: Oxford University Press.
- Stumm W and Morgan JJ (1996) *Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters*. New York: John Wiley & Sons, Inc.
- Uehara G and Gillman G (1985) *The Mineralogy, Chemistry, and Physics of Tropical Soils With Variable Charge Clays. West View Tropical Agriculture Series*. Colorado: Boulder.
- van Breemen N, Mulder J, and Driscoll CT (1983) Acidification and alkalization of soils. *Plant and Soil* 75: 283–308.
- White AF, Vivit DV, Schulz MS, Bullen TD, Evett RR, and Agarwal J (2012) Biogenic and pedogenic controls on Si distributions and cycling in grasslands of the Santa Cruz soil chronosequence, California. *Geochimica et Cosmochimica Acta* 94: 72–94.
- Wiche O, Székely B, Moschner C, and Heilmeyer H (2018) Germanium in the soil-plant system—A review. *Environmental Science and Pollution Research* 25(32): 31938–31956.

Chemical equilibria

Arthur Paul Schwab, Soil and Crop Sciences, Texas A&M University, College Station, TX, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of A.P. Schwab, CHEMICAL EQUILIBRIA, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 189–194, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00188-0>.

Introduction	275
Definitions of chemical equilibrium	275
Kinetic definition	275
Free-energy definition	275
Le Chatelier's principle and chemical equilibrium in soils	276
Applicability of chemical equilibrium principles to soils	277
Limitations to equilibrium in soils	277
Soil systems in which equilibrium concepts may apply	278
Experimental approach to equilibrium in soils	278
System variables	279
Reactions and associated equilibrium constants	279
Equilibrium modeling of soil systems	279
Application of chemical equilibrium to soils: Examples	279
Interactions of lead and phosphate in contaminated soils	280
Manganese(III) phosphate	281
Calcite in soils: Equilibrium, nonequilibrium, supersaturation	281
The outlook for studying chemical equilibrium in soils	281
References	282

Abstract

The application of chemical equilibrium to soils and soil solutions is a potentially flexible tool, allowing the user to interpret experimental data and predict chemical changes after perturbations. Geochemical models are widely available to assist in this practice but require an understanding of fundamental concepts and an appreciation of the limitations to assumptions of equilibrium. This chapter establishes thermodynamic principles, discusses the importance of kinetics of reactions, examines shortcomings of applying these concepts in soils, and presents published applications of the approach.

Key points

- Basic concepts of chemical equilibria in soils are developed
- Chemical equilibria in soils have kinetic and thermodynamic requirements
- Equilibrium concepts can be applied to soils but challenges exist
- Geochemical modeling can be a powerful tool when the limitations are understood

Nomenclature

μ_i	Chemical potential of species i
μ_i°	Standard chemical potential of species i
DTPA	Diethylenetriamine pentaacetic acid
Equilibrium	A state of balance in a chemical reaction in which the concentrations of the reactants and products are no longer changing and in which minimum free energy has been attained
IAP	Ion activity product. The mathematical result of inserting measured or calculated activities of species into an equilibrium expression, including stoichiometric coefficients.
k	Rate constant
K_{eq}^c	Conditional equilibrium constant, based on concentrations of species
K_{eq}°	Thermodynamic equilibrium constant, based on activities of species
Redox	Oxidation/reduction
Solubility constant	Thermodynamic equilibrium constant for the dissolution of a solid phase

Steady state	A balance in an experimentally observed chemical reaction in which the concentrations of the reactants and products appear to be unchanging with time. Steady state differs from equilibrium in that the minimum free energy for a given reaction has not been achieved
Stoichiometry	The relative quantities (in moles) of products and reactants in a chemical reaction are indicated by the coefficients in the balanced chemical equation
ν_x	Rate of reaction. Forward reaction: $x = f$; reverse reaction $x = r$

Introduction

Although soils are dynamic, complex, and living systems, they are composed of identifiable minerals, amorphous compounds, and discrete ionic species that can be quantified. Solubilities, solid-phase transformations, dissolution of gases, and changes in oxidation state can be predicted by chemical thermodynamics. However, implicit in the application of these principles to soils is that the soil system (or, at the very least, the constituents in question) are in chemical equilibrium. If the assumption of equilibrium is valid, then we have the means at our disposal to describe the chemical behavior of the solid, liquid, and gas phases in soil. The challenges in applying chemical equilibrium to soils include: (1) identifying phases in the soil involved in the reactions; (2) developing the equilibrium concepts; (3) determining if equilibrium can be applied to systems of interest.

Definitions of chemical equilibrium

Before discussing chemical equilibrium in soils, we must clearly understand the concept. Two basic definitions of chemical equilibrium will be discussed: kinetic and free energy. The two definitions express the same phenomenon in different ways and allow us to understand better the constraints on equilibria in soils.

Kinetic definition

Consider a generalized chemical reaction in which A and B react to form products C and D with stoichiometric coefficients a, b, c, and d:



The rate of the forward reaction (from left to right) is given by:

$$v_f = k_f[A]^a[B]^b \quad (2)$$

in which $[x]$ represents the solution concentration of component x , and k_f is the forward rate constant. The rate expression for the reverse reaction is given by:

$$v_r = k_r[C]^c[D]^d, \quad (3)$$

where k_r is the rate constant for the reverse reaction. When only the reactants are present, the forward reaction proceeds relatively quickly. As products begin to accumulate, the reverse reaction occurs simultaneously, and the progress of the reaction slows. Eventually, the point will be reached in which the forward and reverse reactions are proceeding at the same rate, and the concentrations of the reactants and products will not change with time. This is the point of a dynamic, chemical equilibrium. When the forward and reverse rates are equal:

$$k_f[A]^a[B]^b = k_r[C]^c[D]^d. \quad (4)$$

Rearranging

$$K_{eq}^c = \frac{k_f}{k_r} = \frac{[C]^c[D]^d}{[A]^a[B]^b} \quad (5)$$

in which K_{eq}^c is the conditional equilibrium constant based on species concentrations.

Free-energy definition

The chemical potential of any species in solution (μ_i) is given as:

$$\mu_i = \mu_i^0 + RT \ln(i) \quad (6)$$

in which μ_i^0 is the standard chemical potential of species i , R is the ideal gas constant, T is the temperature in Kelvin, and (i) is the activity of species i . The expression for the free-energy change of any chemical reaction is:

$$\Delta G = \sum v_i \mu_i \quad (7)$$

in which ΔG is the change in Gibbs free energy during the course of the reaction, and v_i is the stoichiometric coefficient for the components of the equilibrium expression (v_i is positive for products, negative for reactants). Substituting:

$$\Delta G = \sum_i v_i \mu_i^0 + RT \sum_i v_i \ln(i) \quad (8)$$

or

$$\Delta G = \Delta G^0 + RT \ln \prod_i (i)^{v_i} \quad (9)$$

and ΔG^0 is the standard Gibbs free energy change of the reaction. The reaction quotient, Q , can be represented by:

$$Q = \ln \prod_i (i)^{v_i}, \quad (10)$$

such that:

$$\Delta G = \Delta G^0 + RT \ln Q. \quad (11)$$

By definition, the free-energy change of reaction equals zero ($\Delta G = 0$) at equilibrium, and the value of Q at equilibrium becomes the equilibrium constant, K_{eq}^0 :

$$K_{eq}^0 = Q_{eq} = \prod_i (i)_{eq}^{v_i} \quad (12)$$

and:

$$\Delta G^0 = -RT \ln K_{eq}^0. \quad (13)$$

Thus, we have the free-energy definition of equilibrium and the thermodynamic derivation of the equilibrium constant in which K_{eq}^0 is the equilibrium constant based on activities of the species. The expression for the K_{eq}^0 for the generalized equilibrium (Eq. 1) is given by

$$K_{eq}^0 = \frac{(C)^c (D)^d}{(A)^a (B)^b} \quad (14)$$

where the activity of species X is defined as $(X) = \gamma_x [X]$ and γ_x is the activity of species X . Substituting into Eq. (5),

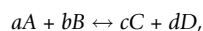
$$K_{eq}^0 = \frac{\gamma_c [C]^c \gamma_d [D]^d}{\gamma_a [A]^a \gamma_b [B]^b} = K_{eq}^c \frac{\gamma_c \gamma_d}{\gamma_a \gamma_b}. \quad (15)$$

Le Chatelier's principle and chemical equilibrium in soils

Le Chatelier's principle is a fundamental concept that predicts the impact of change on chemical equilibria. Le Chatelier's original work is still available (Le Chatelier, 1884) and can be paraphrased:

If a system at equilibrium receives a stress (change in pressure, volume concentration, temperature), then the equilibrium will be disrupted. The system will experience internal modifications that will induce changes that are opposite in sign to that of the stress, resulting in a relief of that stress and, ultimately, a new state of equilibrium.

Referring to the generalized chemical equilibrium (Eq. 1):



if one of the products, D , is added to the system to increase the concentration of D , then according to Le Chatelier's principle, the equilibrium would shift to the left to reduce the concentration of D to "relieve the stress." Product D will react with product C , decreasing both D and C , and increasing the concentrations of A and B until a new equilibrium is reached. In the context of free energy, adding D to the system induces an initial change in free energy such that $\Delta G \neq 0$ and $Q \neq K_{eq}^0$. Concentrations will change until the inequalities are removed: $\Delta G = 0$ and $Q = K_{eq}^0$. From a kinetics perspective, increasing the concentration of D will increase v_r , and $v_f \neq v_r$. Rates of consumption of C and D will increase as will the rate of production of A and B . Eventually, the forward and reverse reactions will begin to balance until $v_f = v_r$ once more.

Applicability of chemical equilibrium principles to soils

Invoking the assumption of chemical equilibrium in soils, particularly under natural conditions, is difficult because soils are highly complex and forever changing. Soils continually have inputs of external energy, water, gases, and dissolved constituents. Losses of these same components are equally possible through the radiation of heat, leaching, and biological activities. These changes perturb the soil system and upset any equilibrium that may be present. With this in mind, a reasonable question is, “Can the principles of chemical equilibrium be applied to soils?”

Despite the dynamic nature of soils and continuous inputs and removal from the system, chemical equilibrium can be applied if care is taken, the limitations are clearly understood, and expectations are reasonable.

Limitations to equilibrium in soils

Most soils are open systems with frequent inputs and removal of mass and energy. Even the simple act of rain falling on a soil results in the dilution of dissolved constituents and disruption of previous equilibria. Changes in temperature, plant assimilation of nutrients, evaporation of water, and diffusion of gases into or out of soil have a similar disruptive effect. These perturbations in equilibrium may be temporary or may have long-term implications, depending upon the kinetics of these reactions. In aqueous solutions, reactions tend to establish equilibrium relatively quickly. For example, the half-life of complexation and dissociation reactions varies from 10^{-8} to 10^4 s (Table 1). Thus, outside disturbances to equilibria will have a small impact on most complexation reactions. The slowest reaction in Table 1 is the oxidation of Fe(II) to Fe(III) at pH 6, with a half-life of approximately 3 h.

Solid-phase dissolution and precipitation reactions can require a long time to achieve equilibrium relative to aqueous phase complexes. A classic example of this is the precipitation and dissolution of gibbsite (γ -Al(OH)₃) which has been studied frequently. The chemistry of aluminum in soils is complex, in both the solution and solid phases. Nevertheless, apparent chemical equilibrium with gibbsite has been observed. Fig. 1 illustrates the time trend as an acidic solution of 0.025 mol L^{-1} Al was slowly leached

Table 1 Reaction orders and half-lives of selected complexation and dissociation reactions in aqueous solutions. All data from Pankow and Morgan (1981) unless otherwise stated.

Reaction	Order	Half-life (s)
$\text{MnSO}_4(\text{aq}) \leftrightarrow \text{Mn}^{2+} + \text{SO}_4^{2-}$	1	10^{-8}
$\text{H}^+ + \text{HS}^- \leftrightarrow \text{H}_2\text{S}$	2	10^{-8}
$\text{FeOH}^{2+} + \text{H}^+ \leftrightarrow \text{Fe}^{3+} + 3\text{H}_2\text{O}$	2	10^{-8}
$\text{Fe}^{3+} + 3\text{H}_2\text{O} \leftrightarrow \text{FeOH}^{2+} + \text{H}^+$	1	10^{-7}
$\text{Mn}^{2+} + \text{SO}_4^{2-} \leftrightarrow \text{MnSO}_4(\text{aq})$	2	10^{-6}
$\text{Cr}^{3+} + \text{H}_2\text{O} \leftrightarrow \text{CrOH}^{2+} + \text{H}^+$	1	10^{-6}
$\text{FeOH}^{2+} + \text{H}_2\text{O} \leftrightarrow \text{Fe(OH)}_2^+ + \text{H}^+$	1	10^{-5}
$\text{H}_2\text{S} \leftrightarrow \text{H}^+ + \text{HS}^-$	1	10^{-4}
$\text{CO}_2 + \text{H}_2\text{O} \leftrightarrow \text{H}_2\text{CO}_3$	1	10^1
$\text{CO}_2 + \text{OH}^- \leftrightarrow \text{HCO}_3^-$	2	10^{-1}
$\text{Fe}^{2+} + \text{O}_2 \leftrightarrow \text{Fe(III)} \text{ (pH 8)}$	2	10^0
$\text{Fe}^{2+} + \text{O}_2 \leftrightarrow \text{Fe(III)} \text{ (pH 7)}$	2	10^2
$\text{Fe}^{3+} + \text{F}^- \leftrightarrow \text{FeF}^{2+}$	2	10^{2a}
$\text{Fe}^{2+} + \text{O}_2 \leftrightarrow \text{Fe(III)} \text{ (pH 6)}$	2	10^4

^aPlankey and Patterson (1986).

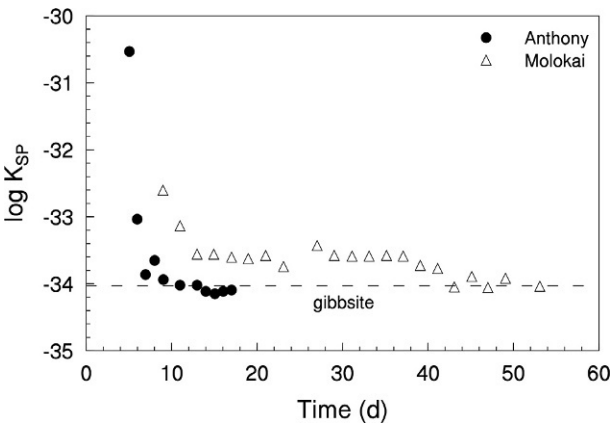


Fig. 1 The approach to equilibrium with gibbsite over time for soluble Al in two soils. In this case, $K_{\text{SP}} = (\text{Al}^{3+})(\text{OH}^-)^3$. The dashed line represents equilibrium with γ -Al(OH)₃(gibbsite). Data from Marion et al. (1976).

through soils (Marion et al., 1976). Equilibrium was approached from supersaturated conditions, with aluminum solids precipitating as the solution passed through the soils. In the Anthony soil (entisol from Arizona), equilibrium with gibbsite was approached in approximately 10 days. The exudate solution from the Molokai soil (oxisol from Hawaii) was still 10 times above gibbsite saturation at 10 days and did not approach apparent equilibrium until 40 days. Slight shifts in pH, temperature, or moisture content would result in nonequilibrium conditions for $\gamma\text{-Al(OH)}_3$ (gibbsite) that would again require several days to correct.

Some reactions are thermodynamically favorable but are associated with a large energy barrier (e.g., energy of activation). Unless that energy barrier can be overcome, equilibrium will not be realized. The conversion of $\text{N}_2(\text{g})$ to nitrate should be a spontaneous reaction, but the presence of approximately 78% $\text{N}_2(\text{g})$ in the atmosphere is clear evidence that this reaction does not proceed as predicted. Redox couples with oxygen-containing anions (e.g., arsenite/arsenate ($\text{AsO}_3^{3-}/\text{AsO}_4^{3-}$)) also attain equilibrium slowly in the absence of a catalyst.

Measurement of critical system variables can be subject to limitations that restrict rigorous application of thermodynamics to natural soils. Detection limits of soluble constituents may preclude quantification, redox-sensitive electrodes have shortcomings, and distinguishing oxidation states of soluble elements can be troublesome. Redox is a crucial system parameter, but chemical limitations of redox electrodes are well documented. Not all oxidation–reduction couples establish reversible equilibria at the surface of the sensing electrode.

Thermodynamic data and/or equilibrium constants are available for hundreds of minerals found in soils, but many of these minerals can only be formed under conditions of high temperatures and pressures. This is particularly true of the primary minerals in soils. The chemical reaction can be written, and thermodynamics might predict that these minerals form in soils. However, considerations of equilibrium have less impact when one or more of the constituents cannot form.

Organic solid phases and soluble organic matter participate in soil chemical reactions, and equilibrium can be established. However, organic compounds in soils can be dynamic and transform relatively rapidly. Many elements in soil react strongly with organic matter, including Fe (iron), Al, Cu (copper), Zn (zinc), As (arsenic), and P (phosphorus). These interactions are difficult to predict because of the structural, temporal, and spatial variability of soil organic matter but should not be ignored when assessing the equilibrium status of soils.

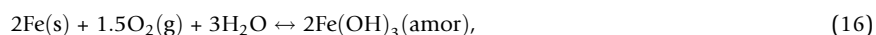
Soil systems in which equilibrium concepts may apply

Chemical equilibrium principles can be applied with reasonable confidence to most reactions occurring exclusively in the solution phase. As illustrated in Table 1, most reactions attain equilibrium in a few seconds or less; therefore, typical perturbations to equilibrium relax quickly for these soluble components. Attempts to apply equilibrium to the solid phase in the soil require more consideration. The kinetics of dissolution of typical soil minerals are slow. Calcium phosphate and calcium sulfate minerals can achieve equilibrium in hours; aluminum phosphates, carbonates, and most oxides in a matter of days or weeks; and complex aluminosilicates in years. If the soil solution has been in contact with soil minerals for the time necessary to establish equilibrium while experiencing minimal disturbances, application of equilibrium principles is theoretically possible. Complete equilibrium among all species in the soil solution and all soil minerals is not essential; however, these principles cannot be applied with confidence if even one component ion of the mineral being studied is out of equilibrium.

Consider the example of variscite, $\text{AlPO}_4 \cdot 2\text{H}_2\text{O}$. Phosphate ions may reach constant concentrations in a matter of hours in the soil solution, but Al^{3+} may require days or weeks due to the slow reaction kinetics of aluminum oxides. Therefore, the system will be in equilibrium only when both Al^{3+} and PO_4^{3-} have reached time-independent concentrations.

Universal adherence to chemical equilibrium in soils is unrealistic. However, employing thermodynamics and associated solubility and complexation constants in soil systems can be fruitful if one has reasonable expectations.

Predict endpoints of reactions Thermodynamics can be useful in determining possible endpoints versus impossible outcomes of chemical reactions in soils. For example, we can predict that metallic iron will readily convert to ferric oxides under surface soil conditions:



but the reverse reaction would be highly unfavorable.

Determine ionic activities in soil solution The activities of individual species in the soil solution are often the driving forces behind many chemical and biological reactions in soils and can be calculated using chemical equilibrium constants.

Calculate final conditions for systems that attain equilibrium rapidly When changes to the soil system occur infrequently relative to the timescale required for equilibrium, then equilibrium concepts can be applied.

Experimental approach to equilibrium in soils

Thermodynamics may be used to describe chemical reactions in soils, interpret experimental data, or predict changes in soil chemistry induced by environmental perturbations. In all cases, the user must be aware of the limitations of chemical equilibrium in soils and must have knowledge of the systems to which these principles may be applied. The steps taken are generally the same: critical system variables must be identified; key reactions of all the contributing and interacting components are established; reliable equilibrium constants are located and assigned to each reaction; and, finally, the system of equilibrium equations is set up as a mathematical array and solved.

System variables

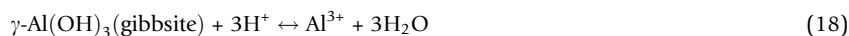
The most important variables in a soil system are similar to those in any other geochemical environment. The list of parameters to be considered is driven in part by the interest and focus of the user and is strongly influenced by other variables that are intimately linked to the systems of interest. For example, if one is considering only the solubility of aluminum oxides in soils with low pH, then the obvious system parameters are pH and soluble Al. However, Al^{3+} can form important solution complexes with Cl^- (chloride), SO_4^{2-} (sulfate), and F^- (fluoride). Therefore, one must account for these ancillary anions before the Al system can be fully defined.

Reactions and associated equilibrium constants

For any given system, all relevant reactions and the corresponding constants must be identified. This is an arduous task and requires databases of reliable information. Much of this work was done in the development of geochemical models, as will be discussed below. For a very simple system such as $\gamma\text{-Al}(\text{OH})_3$ (gibbsite) in equilibrium in soil, the following solution species should be considered: Al^{3+} , AlF^{2+} , AlF_2^+ , AlSO_4^+ , AlOH^{2+} , $\text{Al}(\text{OH})_2^+$, and $\text{Al}(\text{OH})_3^\circ$. Other complexes that could be important include polynuclear Al species, $\text{Al}_x(\text{OH})_y(\text{H}_2\text{O})_z$ (Bertsch and Parker, 1995) and soluble organic complexes (Nordstrom and May, 1995). A typical equation and constant would be:



for which $\log K_{\text{eq}}^\circ = -5.00$. Equations and constants are assembled for all species, including the dissolution reaction for the controlling solid phase:



and $\log K_{\text{eq}}^\circ = 7.74$. After all species have been associated with an independent equation or an actual measurement of a system parameter (e.g., total soluble Fe or pH), they are assembled in a mathematical array of simultaneous equations; the array is solved, and the system can be defined exactly. For small systems, this can be done by hand, but larger systems require a full geochemical model to provide a solution.

Equilibrium modeling of soil systems

Solving the array of simultaneous equations that results from defining a soil system in equilibrium provides a serious challenge. Hand solutions are time-consuming as is writing small computer programs for each new system. Fortunately, this problem was recognized in the mid-1970s, and several research groups set about the task of writing generalized computer models for geochemical systems.

Geochemical models Perhaps the earliest published computer-based geochemical model was designed to map the reaction pathways involving minerals and aqueous solutions (Helgeson, 1968). The WATEQ model was a relatively comprehensive one with a carefully reviewed database of thermodynamic constants (Truesdell and Jones, 1974). The MINEQL model was designed for rapid execution times with the ability for ‘mass transfer’, quantitatively transferring constituents into and out of the solid phase as required by equilibrium constraints (Westall et al., 1976). Since the early beginnings, geochemical modeling has evolved in many directions with many new models: GEOCHEM; PHREEQC; SOILCHEM; visual MINTEQ; GEMS-PSI; Geochemist’s Workbench; and many others. Some models have transport features, and others build in graphic capabilities for visualizing results.

Model inputs Each model requires that the user define the systems under consideration. The computer models can accept experimental data for calculating ionic activities and comparing them with mineral solubilities. The models also may be used in a predictive mode in which assumptions are made concerning equilibrium with solid phases, pH, temperature, gas partial pressures, etc. The models then will report resulting activities, total elemental solubilities, residual masses of solid phases, and the distribution of mass of a given element across all phases.

Continuing with the aluminum example, let us assume equilibrium with $\gamma\text{-Al}(\text{OH})_3$ (gibbsite) at pH 6.5 of pH 5 with total soluble fluoride of $2.0 \times 10^{-5} \text{ mol L}^{-1}$, and total soluble sulfate 0.001 mol L^{-1} . This information was used as input for the visual MINTEQ model (Gustafsson, 2022), and the results are given in Table 2. Total Al solubility is strongly affected by pH as is the distribution of soluble species. Although polynuclear Al species are known to be important in aqueous systems, the modeling results predict that these species will account for $<<1\%$ of soluble Al, agreeing with the notion that these species are metastable (Bertsch and Parker, 1995).

Application of chemical equilibrium to soils: Examples

This section will explore several instances of the application of equilibrium to soils in which soil solution measurements were used as inputs into a geochemical model. The output of the model was then used to determine the ion activity products (IAP) and to compare them with corresponding equilibrium constants. The first examples will illustrate apparent equilibrium followed by apparent nonequilibrium examples.

Table 2 Modeling results with visual MINTEQ with imposed gibbsite equilibrium, total SO_4^{2-} 0.001 mol L^{-1} , total F $2.00 \times 10^{-5} \text{ mol L}^{-1}$, and either pH 6.5 or pH 5.

Component	Species	pH 6.5		pH 5	
		log (activity)	% of total	log (activity)	% of total
SO_4^{2-}	SO_4^{2-}	-3.10	100	-3.10	99.9
F	F	-4.27	99.9	-5.33	24.7
	HF°	-8.04	<0.1	-7.15	0.4
	AlF^{2+}	-9.47	<0.1	-5.58	16.4
	AlF_2^+	-8.58	<0.1	-5.29	54.3
Al^{3+}	AlF_3°	-9.23	<0.1	-6.55	4.2
	Al^{3+}	-11.76	<0.1	-7.26	0.9
	AlOH^{2+}	-10.26	0.9	-7.26	0.7
	Al(OH)_2^+	-9.05	11.9	-7.55	0.3
	Al(OH)_3°	-8.95	14.4	-8.95	<1
	Al(OH)_4^-	-8.76	23.6	-10.26	<1
	AlSO_4^+	-11.02	0.1	-6.52	3.4
	AlF^{2+}	-9.47	5.4	-5.58	34.5
	AlF_2^+	-8.58	36.0	-5.29	57.0
	AlF_3°	-9.23	7.6	-6.55	3.0
	Soluble Al ^a	-8.11	100.0	-5.02	100

^aSoluble Al is the sum of the concentrations of all aqueous Al species in mol L^{-1} .

Interactions of lead and phosphate in contaminated soils

Soils contaminated with lead (Pb) can be amended with soluble phosphate to reduce the environmental and health risks of Pb. The reaction between Pb^{2+} and PO_4^{3-} can result in the formation of sparingly soluble minerals such as hydroxyphosphomorphite ($\text{Pb}_5(\text{PO}_4)_3\text{OH}$). A contaminated soil was treated with combinations of CaCO_3 (calcite), rock phosphate, or potassium phosphate; aged for approximately 70 days; and samples of the soils were suspended in water for 7 days (Zwonitzer et al., 2003). Soluble constituents were measured, and activities were determined with a geochemical model. Despite the relatively short equilibration times, correspondence between the measured activities and theoretical solubility suggests equilibrium with hydroxyphosphomorphite for those soils amended with KH_2PO_4 (with or without CaCO_3) and for soils amended with rock phosphate without CaCO_3 (Fig. 2). The soils treated with CaCO_3 only, rock phosphate plus CaCO_3 , and the untreated control soil were apparently in equilibrium with PbCO_3 (cerussite). Those soils amended with rock phosphate only may have been in simultaneous equilibrium with cerussite and hydroxyphosphomorphite.

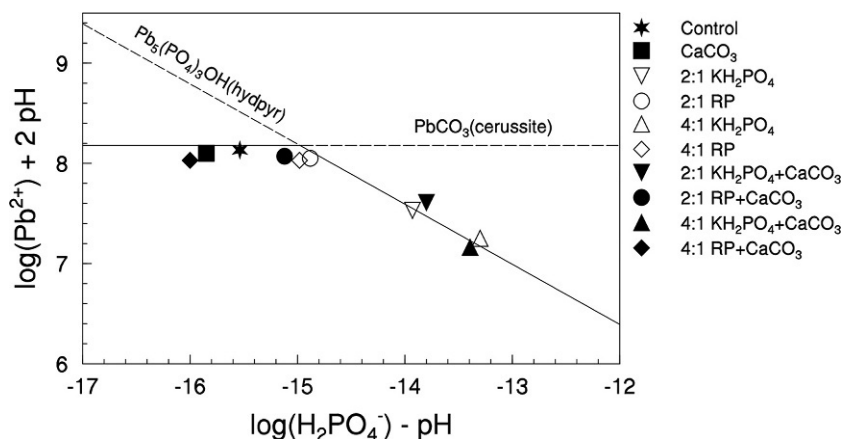


Fig. 2 The impact of fertilization of Pb-contaminated soil in apparent equilibrium with PbCO_3 (cerussite) or $\text{Pb}_5(\text{PO}_4)_3\text{OH}$ (hydroxyphosphomorphite). Soil suspensions were equilibrated for 5 days (Zwonitzer et al., 2003).

Manganese(III) phosphate

Manganese is a critical element in soils for plant and animal nutrition but can be a pollutant if Mn concentrations become elevated. Manganese phosphates are known to form in natural environments but are too soluble to persist, with the exception of Mn(III) $\text{PO}_4 \cdot 1.5\text{H}_2\text{O}$. Although the equilibrium constant for this mineral ($\log K_{\text{eq}}^0 = -34.4$) suggests very low solubilities, Mn(III) is present in vanishingly small concentrations in soil ($10^{-15} \text{ mol L}^{-1}$ as Mn^{3+}). Nevertheless, two studies seem to suggest that the mineral can form in P-fertilized soils (Boyle and Lindsay, 1985; Schwab, 1989).

Activities of PO_4^{3-} and Mn^{3+} were determined in calcareous soils (Boyle and Lindsay, 1985), and the correspondence between these activities and the theoretical solubility of $\text{Mn(III)PO}_4 \cdot 1.5\text{H}_2\text{O}$ was apparent (Fig. 3). A similar study was conducted in nonalkaline soils (Schwab, 1989), and two trends were observed in solubility, depending upon whether the soil had been recently fertilized. In soils fertilized annually with soluble phosphate, activities were slightly elevated relative to $\text{Mn(III)PO}_4 \cdot 1.5\text{H}_2\text{O}$, suggesting the presence of a less-crystalline, more soluble form of the mineral. However, in those soils in which P applications had ceased several years prior to soil sampling, the correspondence between measured and theoretical activities for $\text{Mn(III)PO}_4 \cdot 1.5\text{H}_2\text{O}$ was closer. A gradual approach to solid-phase equilibrium is expected in all soils (Fig. 1), and apparent steady state with a more soluble variant of a solid phase is observed frequently.

Calcite in soils: Equilibrium, nonequilibrium, supersaturation

Precipitation in soils of calcium carbonate, presumed to be calcite, is considered easily reversible and quickly establishes equilibrium. Although many studies examining this mineral have been conducted and support the notion of calcite equilibrium, others have found varying degrees of nonequilibrium. When at equilibrium, the solubility of calcite may be expressed by the following equation:



for which $\log K_{\text{eq}}^0 = -8.48$. Determining the single ion activities of Ca^{2+} and CO_3^{2-} in soil solution makes possible the comparison of measured IAP to the thermodynamic constant. Theoretically, if care is taken in determining the ionic activities, then correspondence between measured and predicted activities can be observed. A series of experiments were run to determine the time needed to reach equilibrium and the state of calcite saturation in groundwaters and calcareous soils (Suarez et al., 1992; Amrhein et al., 1993). Soil solutions were equilibrated for up to 20 days, and activities were determined by various methods. Both the groundwater and soil solutions were found to be in apparent supersaturation with respect to calcite (Fig. 4). The cause of the supersaturation appeared to be the dynamic nature of the soils in which organic matter mineralized rapidly enough to maintain slightly elevated ion activity products for Ca^{2+} and CO_3^{2-} relative to the thermodynamic solubility constant. For soil, the $\log \text{IAP} = -8.2$ was slightly supersaturated compared to $\log K_{\text{eq}}^0 = -8.48$. For groundwater, $\log \text{IAP} = -8.01$. In another study of subsoils with higher pH, the result was similar ($\log \text{IAP} = -8.2$) suggesting that achieving a steady state with elevated Ca^{2+} and CO_3^{2-} activities is a common occurrence. Determining whether this is a function of continual generation of ions from organic matter or the formation of a transient solid phase with greater solubility would need to be determined in each case.

The outlook for studying chemical equilibrium in soils

Advances in geochemical modeling and computing capacity have established new approaches to studying soils. Many applications are possible and include plant nutrition, fate of fertilizers, leaching of soluble constituents, examining the mobility of contaminant

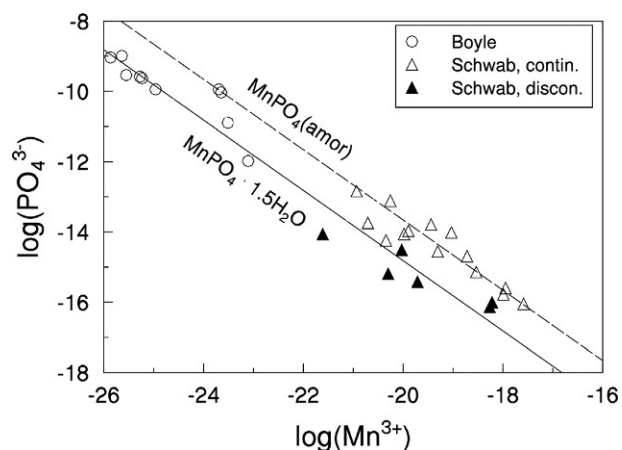


Fig. 3 Comparison of Mn^{3+} and PO_4^{3-} activities to $\text{MnPO}_4 \cdot 1.5\text{H}_2\text{O}$ solubility. Activities were calculated from experimentally determined solution parameters (Boyle and Lindsay, 1985; Schwab, 1989).

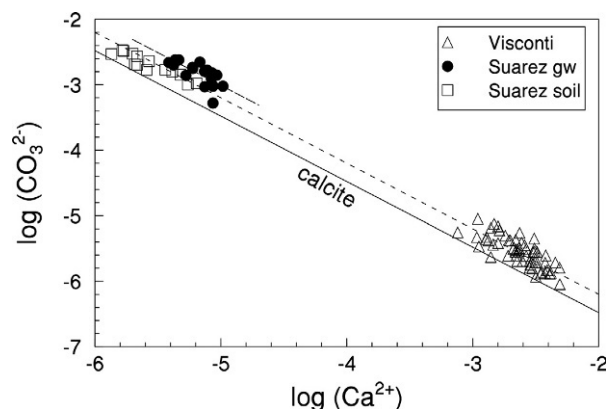


Fig. 4 Measured Ca^{2+} and CO_3^{2-} activities in field soil solutions (open squares; Suarez et al., 1992), groundwater samples (filled circles; Suarez et al., 1992), and saturated extracts from soils (open triangles; Visconti et al., 2010). The solid line is the theoretical solubility of CaCO_3 (calcite) ($\log K_{\text{sp}} = -8.48$). The dashed lines correspond to regression equations for the groundwater samples ($\log \text{IAP} = -8.01$) and the combined soil samples from Suarez et al. (1992) and Visconti et al. (2010) ($\log \text{IAP} = -8.20$). Solution concentration data were used to calculate activities with visual MINTEQ (Gustaffson, 2022).

metals, and volatilization of organics from soils. Examples can be found in the literature that illustrate the relevance of chemical equilibrium in living, dynamic systems.

The DTPA soil test for micronutrients was developed by relating mechanisms of plant uptake to soil chemical equilibrium principles (Lindsay and Norvell, 1978). A mathematical simulation of soil solid phases in the presence of the chelate, DTPA, maximized the correlation between extractable concentrations and phytoavailability of Zn, Mn, Fe, and Cu. The method has been cited thousands of times and is used internationally.

To offset ever dwindling supplies of high-quality phosphate rock used in the production of P fertilizers, efforts are being made to recover phosphate from waste streams. Geochemical modeling demonstrated that struvite ($\text{NH}_4\text{MgPO}_4 \cdot 6\text{H}_2\text{O}$) can be produced and recovered from sludge, and geochemical modeling was employed to optimize the process (Daneshgar et al., 2019; Morita et al., 2019).

Complex cement materials have been modeled thermodynamically to provide a better understanding of the basic reactions in the formation of cement and to avoid unwanted side reactions (Kulik et al., 2021). With this approach, the potential for magnesium phosphates to act as cements was demonstrated (Lothenbach et al., 2019). Chemical equilibrium modeling explained how acid treatments degrade smectite clays and prevent the formation of ettringite in soil cements (Akula et al., 2020).

Human and environmental risk assessments nearly always include knowledge of water-soluble concentrations for pollutants of interest. Integrating geochemical modeling into the assessment can provide concentrations of specific toxic species (e.g., CrO_4^{2-} vs Cr^{3+}) or predict changes in risk in response to contamination or pollutant abatement measures. In Pakistan and Thailand, deep aquifers (>30 m depth) were found to exceed regulatory limits for arsenic in drinking water, and geochemical modeling indicated that Fe-oxides may be responsible for the excessive arsenic concentrations (Shakoor et al., 2018; Santha et al., 2022). Geochemical modeling was interfaced directly with risk assessment models by providing concentrations of the Cd^{2+} species, assumed to be the toxic agent in high Cd groundwater (Miao et al., 2021).

Powerful geochemical models and knowledge of the chemical and physical limitations of applying chemical equilibria in soils can prepare soil scientists to tackle the most challenging and interesting problems of chemical behavior in soil systems.

References

- Akula P, Little D, and Schwab P (2020) Thermodynamic evaluation of smectite treated with hydrogen ion stabilizer. *Journal of Materials in Civil Engineering* 32: 04020098.
- Amrhein C, Zahow MF, and Suarez DL (1993) Calcite supersaturation in soil suspensions. *Soil Science* 156: 163–169.
- Bertsch PM and Parker DR (1995) Aqueous polynuclear species. In: Sposito G (ed.) *The Environmental Chemistry of Aluminum*, 2nd edn, pp. 117–168. Boca Raton, Florida: CRC Publishers.
- Boyle FW and Lindsay WL (1985) Preparation, X-ray diffraction pattern, and solubility product of manganese (III) phosphate hydrate. *Soil Science Society of America Journal* 49: 758–760.
- Daneshgar S, Vanrolleghem PA, Vaneekhaute C, Buttafava A, and Capodaglio AG (2019) Optimization of P compounds recovery from aerobic sludge by chemical modeling and response surface methodology combination. *Science of the Total Environment* 668: 668–677.
- Gustaffson JP (2022) Visual MINTEQ 3.1. <https://vminteq.lwr.kth.se/>.
- Helgeson HC (1968) Evaluation of irreversible reactions in geochemical processes involving minerals and aqueous solutions-I. Thermodynamic relations. *Geochimica et Cosmochimica Acta* 32: 853–877.
- Kulik DA, Winnefeld F, Kulik A, Miron GD, and Lotenbach B (2021) CemGEMS – an easy-to-use web application for thermodynamic modelling of cementitious materials. *RILEM Technical Letters* 6: 36–52.
- Le Chatelier H (1884) Sur un énoncé générale des lois des équilibres chimiques. *Comptes Rendus Académie des Sciences* 99: 786–789.

- Lindsay WL and Norvell WA (1978) Development of a DTPA soil test for zinc, iron, manganese, and copper. *Soil Science Society of America Journal* 42: 421–428.
- Lothenbach B, Xu B, and Winnefeld F (2019) Thermodynamic data for magnesium (potassium) phosphates. *Applied Geochemistry* 111: 104450.
- Marion GM, Hendricks DM, Dutt GR, and Fuller WH (1976) Aluminum and silica solubility in soils. *Soil Science* 121: 76–85.
- Miao F, Zhang Y, Li Y, Liang X, Lin Q, and Zhou Y (2021) Establishing a weighted methodology for human health risk assessment of cadmium based on its equilibrium speciation in groundwater. *Journal of Cleaner Production* 322: 1–12.
- Morita DM, de Luca Avila R, and Aida FN (2019) Nucleation in the formation of struvite: State of art. *Engenharia Sanitaria e Ambiental* 24: 637–654.
- Nordstrom DK and May HM (1995) Aqueous equilibrium data for mononuclear aluminum species. In: Sposito G (ed.) *The Environmental Chemistry of Aluminum*, 2nd edn, pp. 39–80. Boca Raton, Florida: CRC Publishers.
- Pankow JF and Morgan JJ (1981) Kinetics for the aquatic environment. *Environmental Science and Technology* 15: 1155–1164.
- Plankey BJ and Patterson HH (1986) Kinetics of aluminum fluoride complexation in acidic waters. *Environmental Science and Technology* 20: 160–165.
- Santha N, Sangkajan S, and Saenton S (2022) Arsenic contamination in groundwater and potential health risk in Western Lampang Basin, Northern Thailand. *Water* 14: 465.
- Schwab AP (1989) Manganese–phosphate solubility relationships in an acid soil. *Soil Science Society of America Journal* 53: 1654–1660.
- Shakoor MB, Bibi I, Niazi NK, Shahid M, Nawaz MF, Faorqu A, Naidu R, Rahman MM, Murtaza G, and Luttge A (2018) The evaluation of arsenic contamination potential, speciation and hydrogeochemical behaviour in aquifers of Punjab, Pakistan. *Chemosphere* 199: 737–746.
- Suarez DL, Wood JD, and Ibrahim I (1992) Reevaluation of calcite supersaturation in soils. *Soil Science Society of America Journal* 56: 1776–1784.
- Truesdell AH and Jones BF (1974) WATEQ, a computer program for calculating chemical equilibrium of natural waters. *U.S. Geological Survey* 2: 233–248.
- Visconti F, DePaz JM, and Robio JL (2010) Calcite and gypsum solubility products in water-saturated salt-affected soil samples at 25°C and at least up to 14 dS m⁻¹. *European Journal of Soil Science* 61: 255–270.
- Westall JC, Zachary JL, and Morel FMM (1976) *MINEQL, a computer program for the calculation of chemical equilibrium composition of aqueous systems*. Massachusetts Institute of Technology, Technical Note 18 Cambridge, MA: MIT Press.
- Zwonitzer JC, Pierzynski GM, and Hettiarachchi GM (2003) Effects of phosphorus additions on lead, cadmium, and zinc bioavailabilities in a metal-contaminated soil. *Water, Air, and Soil Pollution* 143: 193–209.

Dissolution and precipitation processes

Zhenqing Shi^a, Shiwen Hu^b, Rong Li^a, and Yang Lu^c, ^aThe Key Lab of Pollution Control and Ecosystem Restoration in Industry Clusters, Ministry of Education, School of Environment and Energy, South China University of Technology, Guangzhou, People's Republic of China; ^bState Environmental Protection Key Laboratory of Integrated Surface Water-Groundwater Pollution Control, School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, China; ^cSouth China Institute of Environmental Sciences, Ministry of Ecology and Environment (MEE), Guangzhou, People's Republic of China

© 2023 Elsevier Ltd. All rights reserved.

Introduction	284
Theories for dissolution and precipitation of soil minerals	285
Abiotic dissolution of minerals	287
Biotic dissolution of minerals	289
Precipitation of soil minerals	291
Transformation of precipitated soil minerals	292
Sequestration of heavy metals and organic carbon in soil precipitates	293
Summary	295
References	295

Abstract

The dissolution and precipitation of minerals control the physical properties and chemical composition of soil, speciation and reactivity of contaminants, and preservation of soil organic matter. In this chapter, the basic theories of mineral dissolution and precipitation are first introduced conceptually. Then the biotic and abiotic mineral dissolution processes and the formation and transformation processes of the precipitated soil minerals are discussed with examples of common soil minerals. Lastly, the environmental significance of soil mineral dissolution and precipitation is addressed from the perspective of the sequestration of heavy metals and soil organic matter.

Key points

The objectives of this chapter are:

- To describe briefly the basic theories for dissolution and precipitation of soil minerals;
- To review both abiotic and biotic dissolution of typical soil minerals including aluminosilicates, metal-oxides, and sulfides;
- To review the precipitation and transformation mechanisms of minerals in the soil environment;
- To demonstrate the effects of soil mineral dissolution and precipitation processes on the sequestration of heavy metals and soil organic matter.

Introduction

Soil is the top layer of the Earth's crust which is composed of soil minerals, soil organic matter (SOM), and soil animals and microbes. Originally, the rocks break down during a long period of time by various physical, chemical, and biological processes, which is commonly called weathering. The fine particles after weathering aggregate together and form the physical structure of soil. Natural rainfall and wind erode and transport soil particles, often causing the further decrease of their particle size. The soil contains the remains of dead plants and animals, which gradually forms SOM, together with the input from soil microbes. The mineral nutrients present in the soil are determined by the parent material from which the soil has formed from. The abovementioned soil formation (pedogenic) processes make the soil porous and aerated, and as a result the quality of the soil and the soil structure is improved.

During pedogenic processes, the dissolution of primary minerals and the precipitation of secondary minerals often play a critical role. The primary minerals that are less resistance to chemical weathering (i.e., calcite, olivine) can be preferentially dissolved by the carbonic acid from rainfall or the organic acids and acidic substances derived from the decomposition of organic matter in soil, while other less soluble minerals (quartz, micas, and some feldspars) can persist in soil for a long period of time. Some of the dissolved ions can reprecipitate and crystallize to secondary minerals through chemical processes including hydrolysis and oxidation. Among the many soil minerals, clay and Fe (iron), Al (aluminium), and Mn (manganese) (hydr)oxides are the most reactive ones because of their large specific surface areas and abundant surface functional groups. These minerals assist in the preservation of SOM

and provide a habitat for microbial reproduction. On the other hand, soil minerals interact with soil organic or inorganic contaminants and affect the environmental behavior of soil pollutants. Overall, the dissolution and precipitation of minerals are significant (bio)geochemical processes in soil, which can have a remarkable influence on the fate of elements and pollutants.

Theories for dissolution and precipitation of soil minerals

In general, the dissolution and precipitation of soil minerals can be considered as reversible reactions,



where M represents the mineral phase that composes of soluble components A_i , and s_i represents the stoichiometric coefficient of soluble component A_i .

In chemical thermodynamics, saturation index (SI) is an indicator for the direction of dissolution and precipitation reactions,

$$SI = \log_{10} \left(\frac{IAP}{K_{sp}} \right) \quad (2)$$

$$K_{sp} = \prod_i a_{eq,A_i}^{s_i} \quad (3)$$

$$IAP = \prod_i a_{A_i}^{s_i} \quad (4)$$

In above equations, IAP is the ion activation product of dissolved ions in the aqueous phase, K_{sp} is the solubility product of the mineral phase, a_{eq,A_i} which represents the activity of solute species A_i under dissolution–precipitation equilibrium, and a_{A_i} represents the activity of the same solute in the solutions for calculating SI. When $SI > 0$, the aqueous phase is over-saturated, and the ions may precipitate to approach the equilibrium status. When $SI < 0$, the aqueous phase is undersaturated and the mineral phase may dissolve and release the soluble ions to increase the IAP, until chemical equilibrium is reached. The minerals in soil include a variety of primary and secondary minerals such as quartz, ferrihydrite, hematite, goethite, gibbsite, calcite, dolomite, siderite, gypsum, sulfates, etc. An important measure that describes mineral solubility is pK_{sp} (the negative logarithm of mineral solubility K_{sp}) ($pK_{sp} = -\log_{10} K_{sp}$). Table 1 give the pK_{sp} values of some common minerals (Ball and Nordstrom, 1991).

Although the direction of mineral dissolution–precipitation reactions can be determined from the thermodynamic index SI, the rates of reactions are related to the properties of mineral surfaces in addition to the saturation states of the aqueous phase. The transition-state-theory (TST)-based rate is a widely used theoretical model that can correlate the rates of mineral dissolution with the factors mentioned above. A simplified form of TST-based rate can be expressed as,

$$R = Ak \left(1 - e^{-\frac{n\Delta G}{R_g T}} \right) = Ak \left(1 - \left(\frac{IAP}{K_{sp}} \right)^n \right) \quad (5)$$

where R represents the rate of reaction, A is the reactive surface area of the dissolving mineral, k is the intrinsic dissolution rate constant for the given reaction, ΔG is the free energy change of the dissolution reaction, R_g is the gas constant, T is the absolute temperature, and n is the constant related to the reaction. If we assume $n = 1$ (has been widely applied in dissolution–precipitation of several minerals, i.e., quartz and calcite), the term $\left(1 - \frac{IAP}{K_{sp}} \right)$ represents the degree of disequilibrium for the aqueous phase solutes, which is considered to be the thermodynamic driving force of mineral dissolution.

Mineral dissolution may involve different reaction mechanisms such as proton-promoted dissolution, ligand-promoted dissolution, reductive dissolution, and so on. In some complicated conditions, several parallel reaction pathways may simultaneously determine the kinetics of mineral dissolution. For example, calcite dissolution has three parallel pathways as shown in following equations,



In this case, the overall rate of reaction of calcite dissolution is the sum of the rates for the three parallel reactions,

$$R = A(k_1 a_{H^+} + k_2 a_{H_2CO_3} + k_3) \left(1 - \frac{IAP}{K_{sp}} \right) \quad (9)$$

where a_{H^+} and $a_{H_2CO_3}$ are the activities of H^+ and H_2CO_3 in the aqueous phase, respectively, and k_1 , k_2 and k_3 represent the intrinsic rate constants of reactions (6), (7) and (8), respectively.

The rate limiting steps in mineral dissolution may vary in different environmental conditions. There are three typical rate-limiting steps including transport control, surface reaction control, and mixed transport and surface reaction control. Fig. 1

Table 1 The pK_{sp} values of some common minerals.

Mineral	pK_{sp}	Formula
Al(OH) ₃ (a)	31.17	Al(OH) ₃
Gibbsite	33.86	Al(OH) ₃
Boehmite	33.39	AlOOH
Diaspore	35.09	AlOOH
Witherite	8.56	BaCO ₃
Barite	9.97	BaSO ₄
Portlandite	5.18	Ca(OH) ₂
Hydroxyapatite	54.45	Ca ₅ (PO ₄) ₃ OH
Aragonite	8.34	CaCO ₃
Calcite	8.48	CaCO ₃
Dolomite	17.09	CaMg(CO ₃) ₂
Anhydrite	4.36	CaSO ₄
Fe(OH) ₃ (a)	37.08	Fe(OH) ₃
Hematite	87.95	Fe ₂ O ₃
Magnetite	108.18	Fe ₃ O ₄
Greigite	101.00	Fe ₃ S ₄
Greenalite	63.13	Fe ₃ Si ₂ O ₅ (OH) ₄
Siderite	10.89	FeCO ₃
Mackinawite	18.64	FeS
Pyrite	26.89	FeS ₂
Jarosite(ss)	90.35	K _{0.77} Na _{0.03} H _{0.2} Fe ₃ (SO ₄) ₂ (OH) ₆
Brucite	11.14	Mg(OH) ₂
Magnesite	8.03	MgCO ₃
Rhodochrosite(d)	10.39	MnCO ₃
Birnessite	15.62	MnO ₂
Pyrolusite	17.84	MnO ₂
Manganite	18.26	MnOOH
Quartz	3.98	SiO ₂
SiO ₂ (a)	2.71	SiO ₂

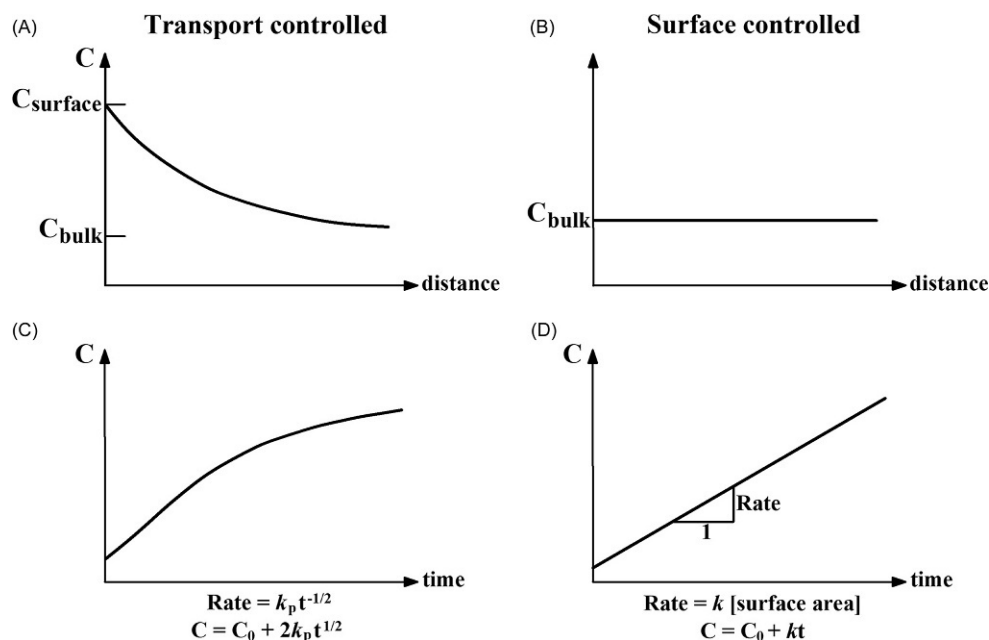


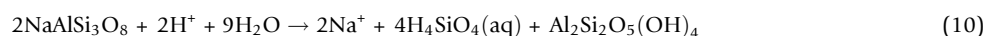
Fig. 1 Schematic representation of concentration in solution, C , as a function of distance from the surfaces of the dissolving mineral (A, B) and reaction time (C, D) for (A, C) transport-controlled and (B, D) surface reaction-controlled dissolution processes. Reproduced with permission from Stumm W (1992) *Chemistry of the Solid-Water Interface: Processes at the Mineral-Water and Particle-Water Interface in Natural Systems*. John Wiley & Son Inc.

shows different concentration profiles of dissolved ions as a function of distance from the surfaces of dissolving minerals for both transport and surface reaction-controlled dissolution (Stumm, 1992). For the transport-controlled dissolution process, the overall rates of reaction are mainly limited by the physical transport of dissolved products from the mineral surfaces to the bulk solution. As a result, the concentrations of dissolved products decay slowly with distance from the mineral surfaces. In comparison, for the surface reaction-controlled dissolution process, there is no limit to the transport of dissolved products and their concentrations close to the mineral surfaces are almost the same as that in the bulk solution. These two systems also result in different concentration profiles as a function of reaction time as shown in Fig. 1. In certain conditions, a mixed transport and surface reaction-controlled dissolution may develop, and the concentrations profile will vary between the above two scenarios.

The kinetics of mineral precipitation are usually complicated because of the multi-step reaction mechanisms. To form the precipitates, the supersaturated ions in the aqueous phase need to nucleate first and the morphology of precipitate is highly dependent on the degree of solution supersaturation. The small nuclei then gradually grow to larger particles accompanied by the consumption of supersaturated solute ions (decreasing IAP). With greater aqueous phase supersaturation, the solid phase can nucleate with a larger number and finer size of nuclei. At the crystal growth stage, a large number of nuclei can further limit the size of particles and make the precipitates more dispersed or poorly crystallized. After the supersaturation has been fully depleted ($SI \leq 0$), the morphology of the poorly crystallized minerals may change further through the Ostwald ripening mechanism or recrystallize to form the chemically more stable mineral phases. Therefore, the precipitated minerals may vary in morphology and crystal structure depending on the stepwise kinetic behavior of nucleation, crystal growth, and Ostwald ripening and recrystallization. The comprehensive review by Fritz and Noguera (2009) provides detailed information on the theoretical background of the kinetics for mineral precipitation.

Abiotic dissolution of minerals

Silicate or aluminosilicate minerals usually have slow rates of dissolution and a lifetime over many years, but they may eventually dissolve in typical soil environments. During the dissolution processes, both congruent and incongruent pathways may occur. In congruent dissolution, the dissolution of a mineral produces only soluble species, such that the stoichiometry of components in the mineral and solution phases after equilibrium are identical, while in the incongruent pathway, the dissolution of a mineral produces one or more mineral species. Feldspar is an abundant mineral component in soils derived from granites and their parent materials, and its incongruent dissolution can release nutrient elements such as K (potassium), Na (sodium), Ca (calcium), and Si (silicon) to the ecosystem, and promotes the long-term alteration of soil mineral composition. The albite (also known as sodium-feldspar) dissolution process is often accompanied by the precipitation of secondary clay minerals, such as kaolinite. The overall reaction can be expressed simply as Eq. (10),



When the albite is dissolved, Na^+ ions and silicic acid only are released into the aqueous phase while Al is enriched in the newly precipitated clay mineral (kaolinite). Because the dissolution of feldspar in typical soil environmental conditions is known to be very slow, the rate of reaction is more likely to be controlled by the surface reactions rather than the transport processes (Fig. 1).

In general, the dissolution kinetics of silicate, aluminosilicate and other oxide type minerals depends strongly on the solution pH (Chou and Wollast, 1985). Fig. 2 shows that the rates of dissolution for albite are the lowest in the neutral pH range and much higher under both acid or basic conditions.

Dissolution of oxide-type minerals is catalyzed by H^+ or OH^- which results from producing the charged surface complexes (Eqs. 11 and 12) that can polarize and weaken the metal(loid)-oxygen binding with the underlying crystal lattice (Stumm, 1992; Brantley, 2008).



With increasing concentrations of charged surface complexes of SiOH_2^+ under low pH or SiO^- under high pH, the surface reaction-controlled dissolution rates increase. Moreover, the uncharged surface sites also dissolve simultaneously without the catalytic contribution of H^+ and OH^- , and contribute to the dissolution kinetics under neutral conditions. Therefore, the pH dependent dissolution rates of silicate can be calculated quantitatively as follows,

$$R = A(k_{\text{acid}}a_{\text{H}^+}^{n_{\text{H}^+}} + k_{\text{neutral}} + k_{\text{base}}a_{\text{OH}^-}^{n_{\text{OH}^-}}), \quad (13)$$

where k_{acid} , k_{neutral} , and k_{base} represent the dissolution rate coefficients of the three dissolution pathways mentioned above, a_{H^+} and a_{OH^-} are the activities of H^+ and OH^- ions, and n_{H^+} and n_{OH^-} are the partial orders of reactions promoted by H^+ and OH^- , respectively.

It is worth mentioning that the rate of dissolution of silicate and aluminosilicate minerals may change significantly with time due to the dynamic alteration of mineral surfaces during the incongruent dissolution or secondary precipitation processes (Yuan et al., 2019). Fig. 3A shows that during the incongruent dissolution of feldspar, Na^+ ions are preferentially leached and Si and Al are enriched on the surfaces. In this case, protons need to diffuse through the Na-depleted layer to contact with the fresh albite for dissolution, and the overall reaction will be slowed down by the diffusion process. Similarly, the precipitated secondary minerals coated on the surfaces may also increase the diffusion barrier and passivate the dissolution process (Fig. 3B).

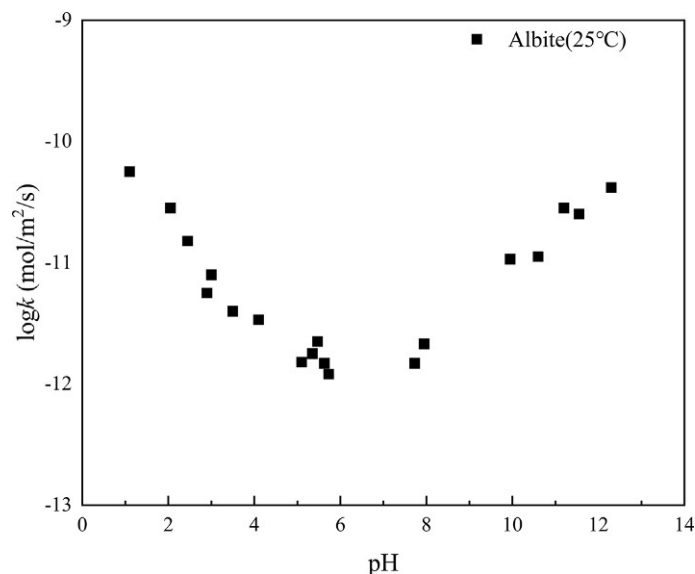


Fig. 2 Rate of albite dissolution as a function of reaction pH at 25 °C. k ($\text{mol m}^{-2} \text{s}^{-1}$) represents the overall rate of mineral dissolution that is normalized by the surface area. Original data are from Chou L and Wollast R (1985) Steady-state kinetics and dissolution mechanisms of albite. *American Journal of Science* 285(10): 963.

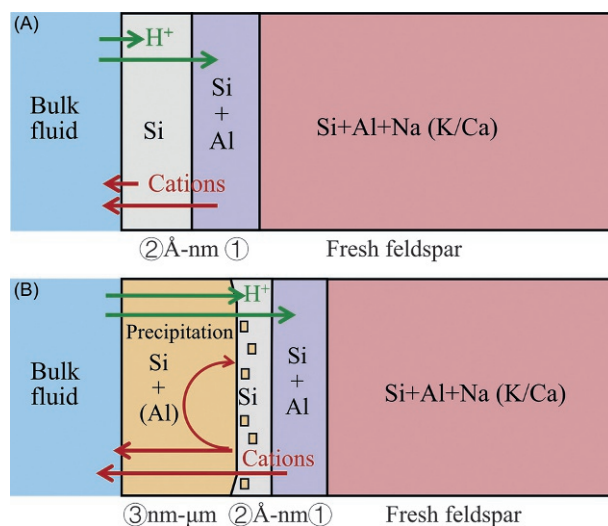


Fig. 3 Schematic diagram of the diffusion control of feldspar dissolution after surface alteration. A, preferential cation leaching mechanism during incongruent dissolution and B, coating of secondary precipitation. ① is the cation-depleted residual layer, ② is the Al-depleted residual layer, and ③ is the reprecipitated coating layer. Green and red arrows represent the diffusion pathways of H^+ and dissolved cations. Reproduced with permission from Yuan GH, Cao YC, Hao F, Jon, G, Liu HY, et al. (2019) A review of feldspar alteration and its geological significance in sedimentary basins: From shallow aquifers to deep hydrocarbon reservoirs. *Earth-Science Reviews*, 191: 114–140.

Iron (hydr)oxides are among the most investigated metal oxides due to their wide distribution. Because of the low solubility of Fe(III) minerals and changeable valence between Fe(III) and Fe(II), reductive dissolution is a key process to regulate the solubilization of most Fe(III) minerals in the environment. Reductive dissolution usually involves two different mechanisms, abiotic reaction of a reductant with the mineral's surfaces and microbially mediated mineral reduction. Organic matter, a redox active substance ubiquitous in the environment, can attach directly to the surfaces of minerals and may also act as electron shuttles to accelerate reduction of the mineral. Typical reductive dissolution processes consist of three steps, diffusion of the reductant molecules to the mineral surfaces, surface chemical reactions, and release of reaction products and their diffusion from the mineral surfaces. The surface chemical reactions, which are often the rate-limiting step, usually include precursor complex formation, electron transfer, and breakdown of the successor complex. The rate of reduction and extent of Fe minerals depend on the electron-donating ability of organic matter. Since the redox potentials of Fe(III)-Fe(II) are between those of reduced and oxidized nitrogen, carbon, sulfur, and oxygen redox species, reductive dissolution of Fe(III) minerals is vital for numerous environmental processes, including carbon storage, ocean productivity, greenhouse gas emissions, and the fate of toxic metal/metalloids.

Organic acids, including citric, ascorbic, oxalate, tartaric acid, and fumaric acid, have been shown to be capable of dissolving Fe minerals. Ascorbate, a commonly used antioxidant, is found widely in natural waters and can induce the reductive dissolution of Fe minerals. The interaction between ascorbate and Fe minerals can be described by following three basic steps (Huang et al., 2017). First, surface inner-sphere complexes of ascorbate (Asc) and Fe are formed on the Fe mineral surfaces,



Second, electron transfer occurs from ascorbate to the surface structural Fe(III), forming surface-bound Fe(II) species and ascorbate radicals ($\cdot\text{Asc}$),



Finally, the adsorbed Fe(II) is gradually released from Fe mineral surfaces into the aqueous phase,



The above organic acid-mediated reductive dissolution processes are influenced by various factors. The most important factor is the pH of the initial solution, and temperature also promotes the reductive dissolution of Fe minerals. Moreover, the chelating capacity of polycarboxylic acids can also affect the reductive dissolution of Fe minerals and more electrons can be transferred in the presence of organic substances with stronger chelating capacity.

Biotic dissolution of minerals

Microbially-mediated mineral dissolution also plays a significant role in regulating the geochemical cycling of elements and transformation/formation of various minerals such as clay minerals and Fe and Mn (hydr)oxides. The Fe(III) reducing bacteria couple the oxidation of inorganic or organic electron donors with the reduction of structural Fe(III) in secondary Fe minerals, which is a common process in an anoxic environment. The electron donors used by Fe(III)-reducing bacteria include carbohydrates, aromatic substances, dihydrogen, amino, and fatty acids. Generally, Fe(III) appears as either crystalline Fe oxides (e.g., hematite and goethite) or short-range-ordered Fe oxides (e.g., ferrihydrite). In almost all anaerobic environments, various bacteria such as *Shewanella* spp., *Geobacter* spp., *Albidiferax ferrireducens*, and hyperthermophilic *Archaea* have been shown to be capable of coupling the oxidation of inorganic or organic electron donors with the reduction of Fe minerals.

There are different pathways of electron transfer from bacteria to Fe(III) minerals as summarized in Fig. 4 (Melton et al., 2014). One of the most common ways of dissimilatory Fe reduction involves direct contact between the Fe(III) mineral surfaces and

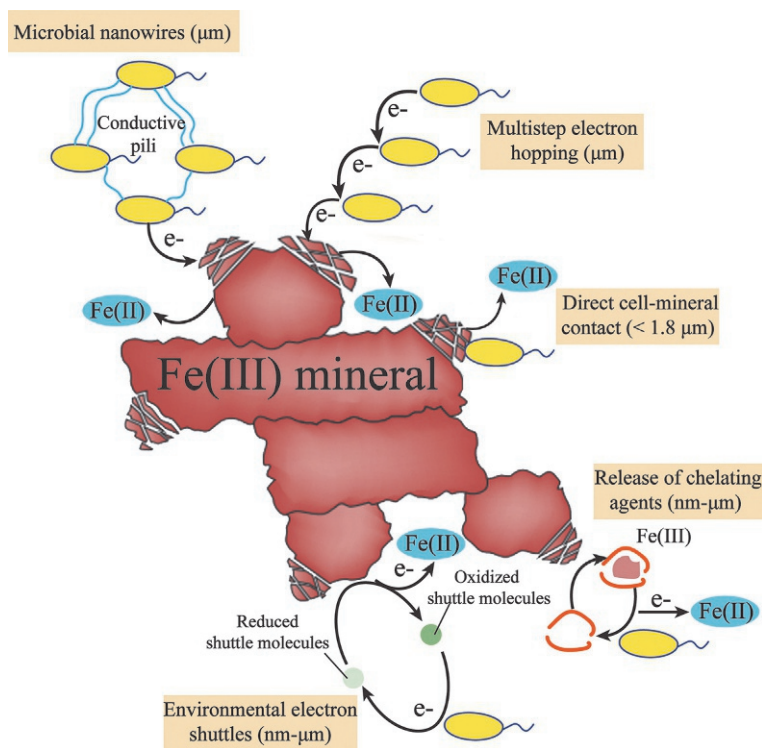
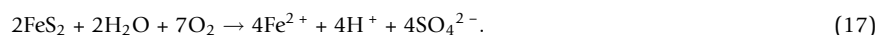


Fig. 4 Pathways of electron transfer from bacteria to Fe(III) minerals. Reproduced with permission from Melton ED, Swanner ED, Behrens S, Schmidt C, Kappler A, (2014) The interplay of microbially mediated and abiotic reactions in the biogeochemical Fe cycle. *Nature Review Microbiology*, 12(12): 797–808.

proteins attached to outer cell wall. However, due to the poor solubility of Fe(III) minerals and the limited transition distance of an electron, several mechanisms have been put forward for the extracellular transfer of electrons from bacteria to Fe minerals. Bacteria may utilize chelating ligands (e.g., humic substances), which can form strong complexes with Fe and promote solubilization of Fe and rapid transfer of electrons. Redox-active Fe minerals (e.g., magnetite), reductive substances secreted by microbes (e.g., riboflavin), and redox-active electron shuttles (e.g., dissolved organic matter) can also be used as an intermediate for electron transfer between distant Fe minerals and intracellular electron transfer chain. In addition, bacteria may transfer electrons to Fe minerals through strategies that include redox-active pili (known as bacteria nanowires) and multistep electron hopping.

The physicochemical properties of Fe oxides, including crystallinity, shape, and particle size, could influence the Fe(III) bio-reduction. Iron minerals with a low degree of crystallinity usually are more likely to be reduced than highly crystalline ones. The exposed facets of Fe minerals during the reactions may also affect the reductive dissolution of the Fe mineral (Hu et al., 2020b). Fig. 5 shows that the rate of reduction of hematite {100} is less than that of hematite {001}, which can be explained by both the attachment of cells onto hematite surfaces and surface electrochemistry of hematite. Specifically, the resistance of hematite {001} is less than that of hematite {100}, therefore, hematite {100} is less favorable for accepting the extracellular electrons from bacteria. Moreover, hematite {001} has a high density of adsorption sites, resulting in more biomass adsorbed on its surfaces, which may reduce the distance of electron transport and thereby increase the rate of Fe(III) reduction.

In addition to the secondary minerals, soil microbes can also directly mediate or indirectly accelerate the dissolution of primary minerals, known as the microbial weathering process. One of the widely known microbially-mediated dissolution processes is the biotic oxidation of sulfide minerals in soil. The latter are present as either discrete sulfide phases (e.g., CuS (copper sulfide), Cu₂S (copper(I) sulfide), ZnS and PbS (lead sulfide)) or coprecipitates with major Fe-sulfide phases such as FeS₂ (e.g., chalcopyrite, CuFeS₂). When exposed with dissolved oxygen (O₂), the sulfide minerals can be oxidized through both abiotic or biotic pathways; the biotic pathways can significantly accelerate the mineral dissolution (Roden, 2008). For example, pyrite (FeS₂) the major sulfide mineral can be abiotically oxidized by the dissolved O₂, which releases the protons, ferrous ion, and sulfate ions,



The released ferrous Fe ions can be further oxidized biotically by *A. ferrooxidans* that are known as the Fe oxidizing acidophilic bacteria,



Although reaction (18) can also occur chemically, the rate is extremely slow at pH values below 4, whereas the rate of microbially catalyzed oxidation is several orders higher at pH < 4 (Fig. 6). In the sulfide oxidizing environment, solutions are often very acid due

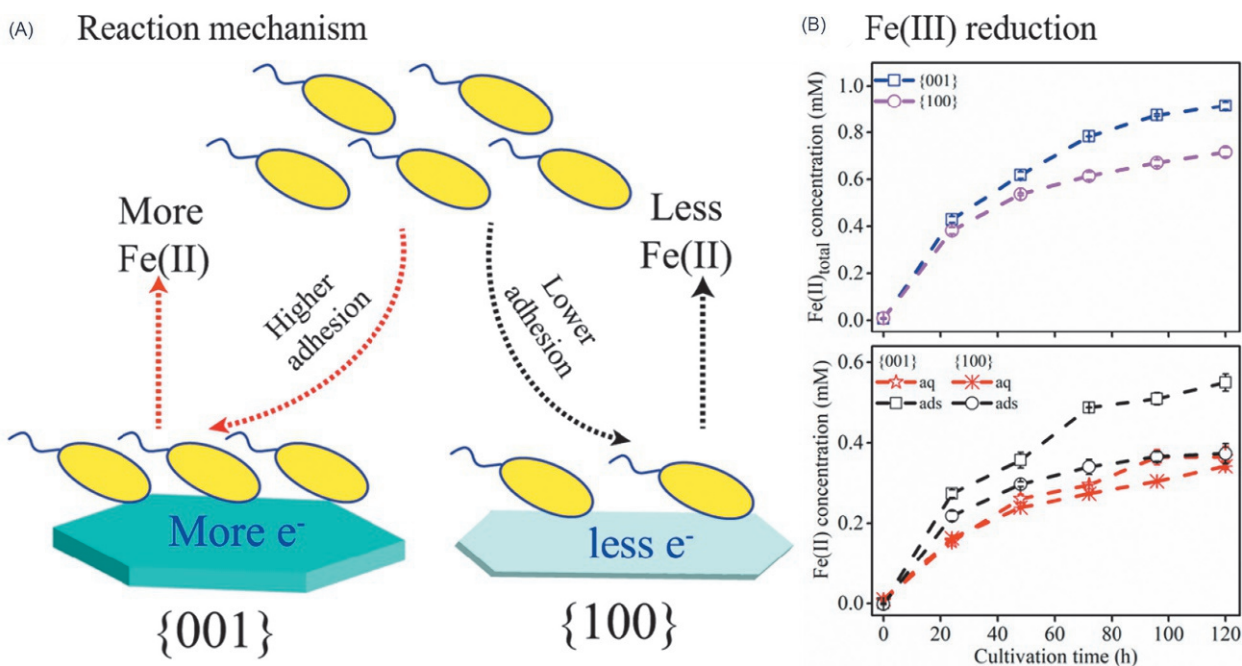


Fig. 5 (A) Diagram of facet-dependent reductive dissolution of hematite by *Shewanella putrefaciens* CN-32, and (B) temporal changes of total Fe(II), adsorbed Fe(II), and aqueous Fe(II) concentrations during the bioreduction process. The aq and ads in (B) represent the aqueous and adsorbed Fe(II), respectively. Reproduced with permission from Hu SW, Wu YD, Ding ZC, Shi ZQ, Liu TX et al. (2020) Facet-dependent reductive dissolution of hematite nanoparticles by *Shewanella putrefaciens* CN-32. *Environmental Science: Nano*, 7(9): 2522–2531.

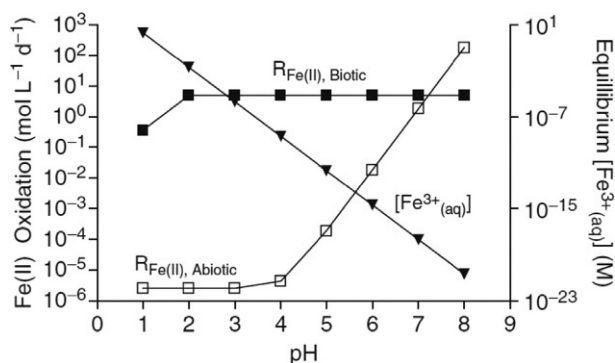


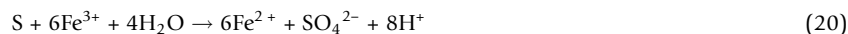
Fig. 6 Rates of abiotic and biotic Fe(II) oxidation and the equilibrium Fe(III) ion concentrations $[\text{Fe}^{3+}_{(\text{aq})}]$ as a function of solution pH. Reproduced with permission from Roden EE (2008) Microbiological controls on geochemical kinetics 2: case study on microbial oxidation of metal sulfide minerals and future prospects. In *Kinetics of Water-Rock Interaction*. Springer, New York, NY, pp. 417–467.

to the release of protons from reactions (17) and (20), therefore, a considerable fraction of Fe^{3+} produced in reaction (18) is expected to be soluble (Fig. 6).

Subsequently, pyrite can be further oxidized by Fe^{3+} chemically,



The elemental sulfur from reaction (19) can also serve as the electron donor and chemical energy source for the sulfur-oxidizing bacteria or archaea, i.e., *A. thiooxidans* and *A. ferrooxidans*,



During the stepwise oxidizing process, the biotic and abiotic effects are coupled tightly, and jointly accelerate the overall dissolution rate of pyrite.

Precipitation of soil minerals

The precipitation process of soil minerals forms amorphous soil minerals with small particle size, poor crystallinity, and large specific surface area, and also causes the dissolved Fe, Al, Si (silicon), Ca (calcium), and Mn (the main mineralized elements in soil) to return to the solid phases from soil solutions. This dynamic process is an important part of the geochemical cycling of soil minerals, which has far-reaching implications for the behavior of soil pollutants and organic matter.

Newly precipitated Fe minerals are highly reactive and can interact with various elements/compounds due to their large specific surface areas, changeable surface charges, and surface activities. At the redox interfaces of the soil solution system, Fe^{2+} can be oxidized to Fe^{3+} , after which Fe^{3+} would further be precipitated to form amorphous Fe oxides such as ferrihydrite. Ferrihydrite has a small crystal size and is usually identified according to the number of its XRD peaks. In natural environments, all forms of ferrihydrite are widespread usually as young Fe oxides and play an important role as an active sorbent. Ferrihydrite is not stable and may transform to thermodynamically stable Fe oxides such as goethite and hematite, depending on the environmental conditions. Manganese (Mn) (oxyhydr)oxides with various mineral phases and oxidation states are also widespread in soils, and are usually formed through the oxidation of Mn(II) to higher valenced Mn. Birnessite ($\delta\text{-MnO}_2$) is the most common Mn oxide, and abiotic oxidation of Mn(II) by dissolved O_2 is an important pathway for birnessite formation, whereas biological oxidation of Mn(II) is regarded as the dominant pathway at certain pHs.

Clay minerals have abundant Al and Si, and usually have a layered structure. In addition to direct precipitation, metal sorption may induce the dissolution of cations held within the lattice structure of clay minerals, and the released cations can re-precipitate to multinuclear surface precipitates with the pre-adsorbing cations. For example, adsorbed Ni (nickel) on soil minerals can promote the dissolution of Si and Al in the crystal structure of soil minerals. The Al and Si released are then sequestered by the newly formed Ni surface precipitates, forming a mixed Ni-Al layered double hydroxide (LDH) precipitate (Fig. 7) (Shi et al., 2012). A significant Ni-LDH phase can occur within a few hours in soil environments and this process may last for days and months. During the aging process, Ni adsorbed on SOM may transfer slowly to the Ni-LDH phase. In general, LDHs are a class of clay minerals with brucite ($\text{Mg}(\text{OH})_2$)-like cationic layers containing anions in the hydrated interlayer for charge balance. The LDH phases are usually thermodynamically preferred to pure metal hydroxides, and might have significant implications for metal sequestration in soils.

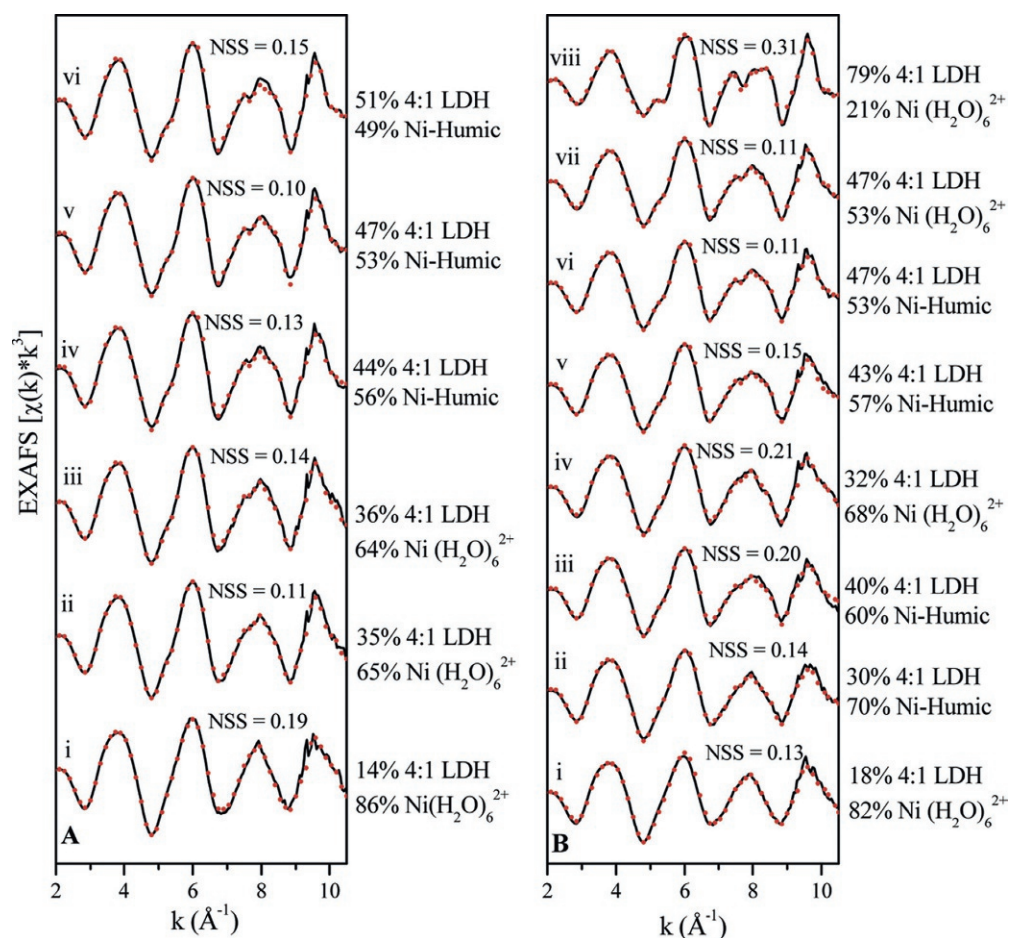


Fig. 7 The EXAFS spectra for Ni speciation in soil showing EXAFS data (—) and best fit lines (●) from linear-least-squares fitting: (A) Soils spiked at pH 7.0 after (i) 30 min, (ii) 4 h, (iii) 8 h, (iv) 12 h, (v) 18 h, and (vi) 24 h; (B) Soils spiked at pH 7.5 after (i) 30 min, (ii) 2 h, (iii) 5 h, (iv) 9 h, (v) 12 h, (vi) 19 h, (vii) 24 h, and (viii) 720 h. Reproduced with permission from Shi ZQ, Peltier E, Sparks DL (2012) Kinetics of Ni sorption in soils: Roles of soil organic matter and Ni precipitation. *Environmental Science & Technology* 46(4): 2212–2219.

Transformation of precipitated soil minerals

The newly precipitated soil minerals are usually meta-stable and have high Gibbs free energy to transform to soil minerals with greater crystallinity and smaller specific surface areas. Ferric Fe ions will precipitate with increasing pH, which forms ferrihydrite nanoparticles after the nucleation process, and ferrihydrite can transform further to well crystallized Fe oxides, for example, hematite and goethite. The formation of hematite begins with ferrihydrite aggregation, mainly through the dehydration of ferrihydrite. The hematite formation process also includes deprotonation of hydroxyl groups, creation of oxy-linkages, and redistribution of cation vacancies. While goethite formation is mainly through the dissolution and recrystallization of ferrihydrite, 2-line ferrihydrite can transform to hematite by a two-stage crystallization process, with goethite as an intermediate product. Hematite is a common iron oxide with the formula Fe₂O₃ and is widespread in rocks and soils. It has different morphologies at the nano scale, including ellipsoidal, cubic, plate-like, or rod nano particles, depending on the environmental conditions. Goethite (FeOOH) is one of the most widespread Fe oxides in natural environments and always has a needle-like morphology. Generally, hematite formation is favored under near-neutral conditions, and a higher temperature and ionic strength, whereas goethite formation is preferred under extreme pH conditions (e.g., less than 4, or greater than 12) and at a lower temperature and ionic strength.

Ferrihydrite can also transform to other iron oxides, for example, lepidocrocite and magnetite. During the microbially mediated Fe mineral reductive dissolution and transformation process (Shi et al., 2020), transmission electron microscopy (TEM) images have shown that ferrihydrite gradually transforms into biogenic magnetite (Fig. 8). The TEM images have also revealed that the newly formed magnetite nanoparticles have a size of 40–60 nm after 120-h transformation experiments. Moreover, with the presence of anions such as arsenic (As), the rates of Fe oxide transformation generally decrease. Fig. 8 shows that with the increase in As concentrations, the transformation of ferrihydrite to magnetite has slowed.

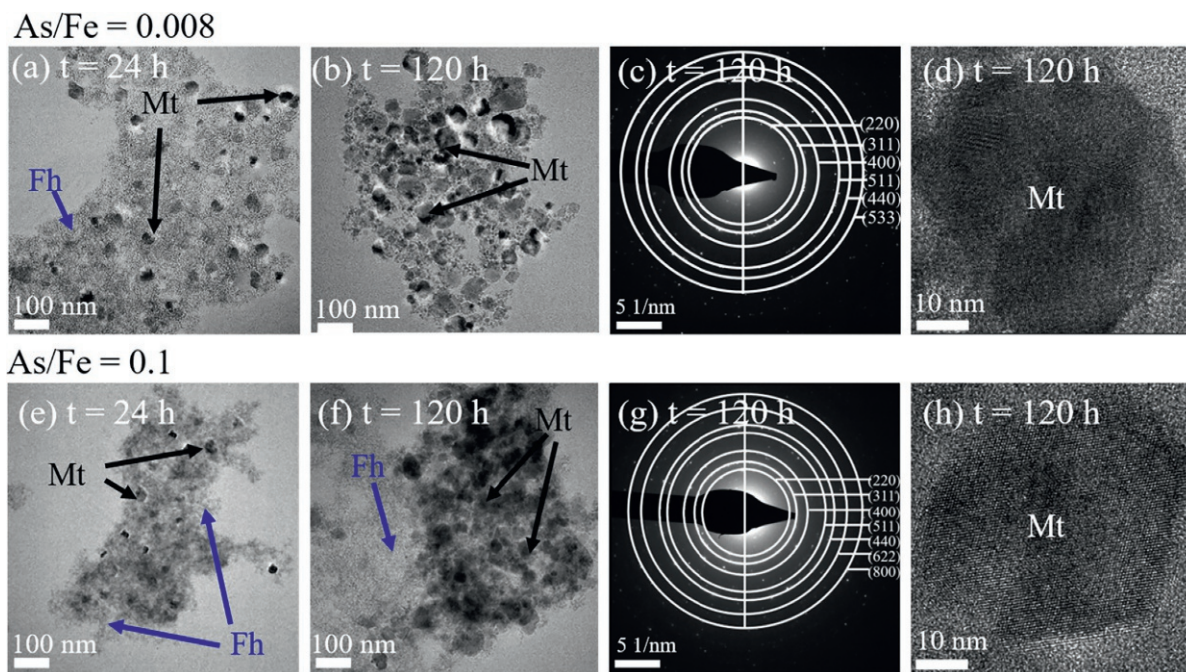


Fig. 8 The TEM characterization of the transformation products of Fe oxides during the *Shewanella putrefaciens* CN-32 mediated reductive dissolution process with two different As/Fe molar ratios. Reproduced with permission from Shi ZQ, Hu SW, Lin JY, Liu TX, Li XM et al. (2020) Quantifying microbially mediated kinetics of ferrihydrite transformation and arsenic reduction: role of the arsenate-reducing gene expression pattern. *Environmental Science & Technology*, 54(11): 6621–6631.

Sequestration of heavy metals and organic carbon in soil precipitates

Heavy metals and SOM can be adsorbed on or coprecipitated with minerals because of the abundant binding sites and large specific surface areas of freshly precipitated soil minerals. During the transformation process of precipitated minerals, adsorbed heavy metals have the potential to be incorporated into the crystal structure of minerals. Some elements, such as Co (cobalt), Cu, Ni, and Zn, might have a strong possibility of being incorporated into Fe oxides during the ferrihydrite transformation process. These elements usually have a similar atomic radius to Fe atoms. Some other elements, for example, Pb ions, cannot be incorporated easily into Fe oxide structures because of the larger atomic radius of Pb than that of Fe. However, under certain extreme environmental conditions (e.g., high pH and Pb concentrations), Pb ions might be able to enter the nano pores or defects within Fe oxides as revealed by the high-angle annular dark-field scanning transmission electron microscopy (HAADF-STEM) imaging (Lu et al., 2020). Some Pb ions can even enlarge the lattice spacing of hematite and distribute along a specific zone axis of hematite during the ferrihydrite transformation process, indicating that some Pb ions might have amorphous substitution with Fe atoms (Fig. 9). Heavy metals within defects or with substitution into the crystal structure of minerals usually have low mobility, which has significant implications for the fate and transport of heavy metals in the environment.

During the formation and transformation processes of soil minerals, the interactions between SOM and minerals protect SOM from degradation and mineralization. At the global scale, mineral protection plays a key role in regulating the long-term stability of natural organic matter in the environment. Soil organic matter (SOM) is a complex mixture of organic macromolecules containing various reactive functional groups such as carboxylic, phenolic, amino, and thiol groups, which can either adsorb on or precipitate with soil minerals. Compared with the surface adsorption process, coprecipitation of SOM with minerals can usually sequester more organic matter. Fig. 10 shows that surface adsorption usually results in partial organic matter cover on dense ferrihydrite aggregates but little organic matter can penetrate into the aggregates (Kleber et al., 2015). For coprecipitates of organic matter with ferrihydrite, ferrihydrite nanoparticles are less aggregated and organic matter is more homogeneously distributed within ferrihydrite aggregates. Depending on the initial organic matter concentrations, some organic matter may be occluded within the pore spaces of ferrihydrite aggregates, providing a significant pathway for organic matter sequestration by ferrihydrite.

The SOM may slow down the Fe oxide transformation process and thus affect the interactions between organic matter and Fe oxides. The presence of organic matter may also affect the formation of the transformation products. During the Fe(II)-induced transformation of ferrihydrite coprecipitated with fulvic acids (FA) and arsenate ions (As), ferrihydrite (Fh) gradually transforms to a mixture of lepidocrocite (Lep) and goethite (Goe) as shown in Fig. 11 (Hu et al., 2020a). For ferrihydrite coprecipitates with FA and As, the line scans of STEM coupled with electron energy loss spectroscopy (EELS) show that the signals of C generally correlate well with both Fe and O, consistent with the conceptual model of organic matter distribution in ferrihydrite coprecipitates shown in Fig. 10. For the transformation products, there is a significant difference between lepidocrocite and goethite for organic C

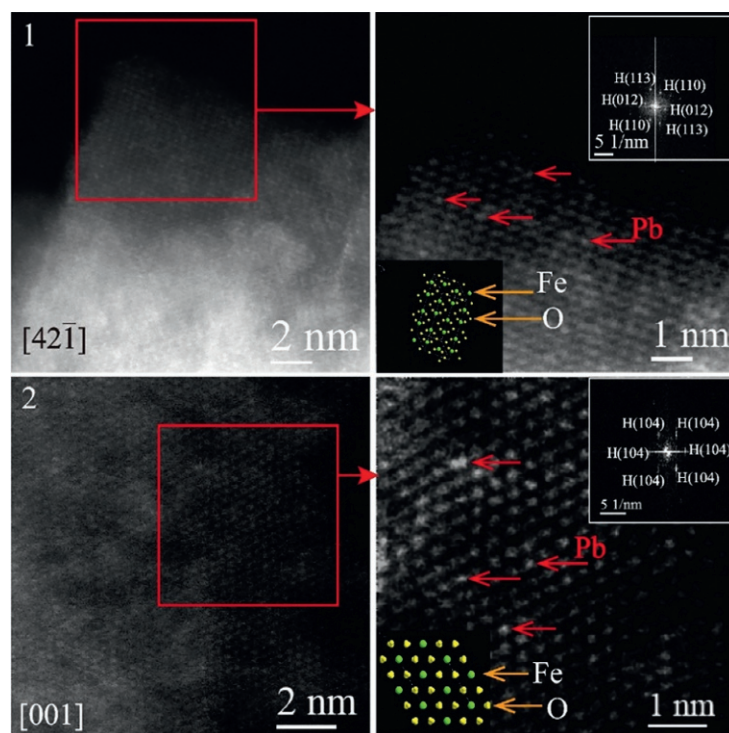


Fig. 9 The HAADF-STEM images of hematite samples during the ferrihydrite transformation process with the presence of Pb ions. The brighter atoms pointed by the red arrows were considered to be the incorporated Pb, and the darker atoms were considered to be iron. The crystal structures of hematite in different zone axis are shown in the lower left corner of zoomed in images. Reproduced with permission from Lu Y, Hu SW, Liang Z, Zhu MQ, Wang ZM et al. (2020) Incorporation of Pb(II) into hematite during ferrihydrite transformation. *Environmental Science: Nano*, 7(3): 829–841.

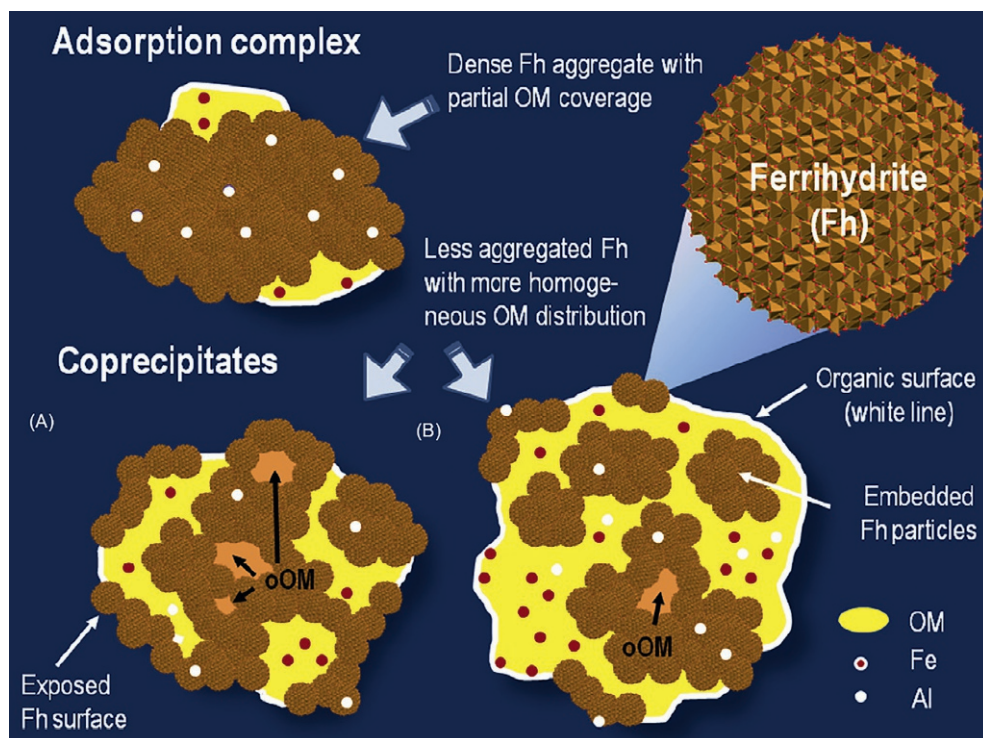


Fig. 10 Possible arrangement of poorly crystalline minerals (e.g., ferrihydrite; Fh) and organic matter (OM) on adsorption complexes and coprecipitates. Reproduced with permission from Kleber M, Eusterhues K, Keilueit M, Mikutta C, Mikutta R, et al. (2015) Chapter one-mineral-organic associations: formation, properties, and relevance in soil environments. In: Sparks, D.L. (Ed.) *Advances in Agronomy*. Academic Press, pp. 1–140.

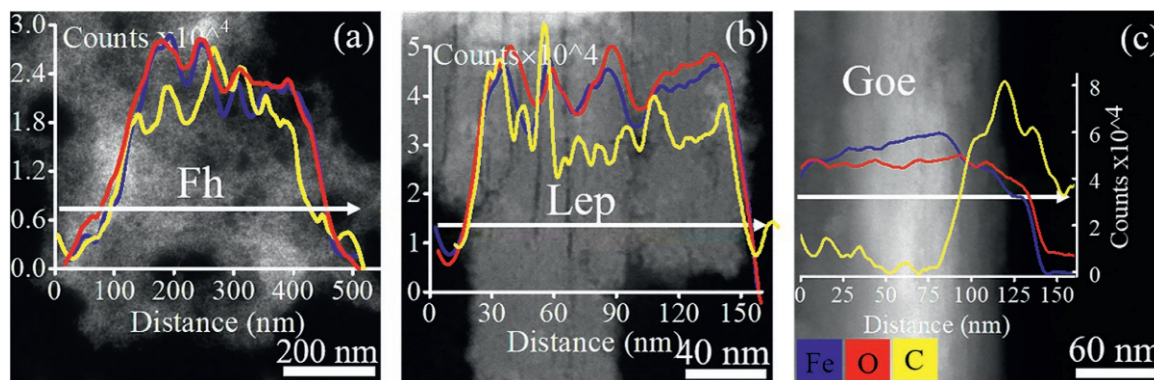


Fig. 11 The STEM-EELS line scans of Fe (blue), O (red), and C (yellow) distribution on Fe oxide particles. Lep denotes lepidocrocite and Goe denotes goethite. Reproduced with permission from Hu SW, Liang YZ, Liu TX, Li FB, and Shi ZQ (2020a) Kinetics of As(V) and carbon sequestration during Fe(II)-induced transformation of ferrihydrite-As(V)-fulvic acid coprecipitates. *Geochimica et Cosmochimica Acta* 272: 160–176.

distribution. For lepidocrocite, there is no accumulation of organic C on the edges of lepidocrocite nanoparticles, but organic matter molecules may be incorporated into its defects or pore spaces of lepidocrocite. In comparison, EELS line scans have clearly shown that organic C has accumulated on the edges of goethite nanoparticles, indicating that organic C is mainly adsorbed on the surfaces of goethite. Therefore, in addition to adsorption, the defects or nano pore spaces formed in the structures of Fe oxides may provide an effective way to sequester C through physical isolation, which has significant implications for predicting mineral transformation and C sequestration in the environment.

Summary

Dissolution and precipitation of minerals play critical roles in determining the physical properties and chemical composition of soil, and consequently control the speciation and reactivity of contaminants and determine the preservation capacity of SOM. In this chapter, we have provided an introduction to the knowledge of soil mineral dissolution and precipitation processes obtained during past decades. The saturation index (SI) that considers the mineral's specific solubility product (K_{sp}) and ion activity product in aqueous solution (IAP) can be referred to as the thermodynamic driving force of mineral dissolution and precipitation. As specific examples for the abiotic mineral dissolution process, proton promoted aluminosilicate dissolution and redox controlled iron oxide dissolution are described. The solution pH, organic ligands, and surface properties of minerals may jointly determine the rates of reaction and end products of mineral dissolution by various mechanisms. Moreover, the biotic mineral dissolution process is also described by pyrite oxidation and the reductive dissolution of Fe oxides. In the precipitation subsections, the nucleation, growth, and transformation processes of precipitated minerals are discussed. Moreover, the transformation of Fe oxides is discussed and the environmental significance of soil mineral dissolution and precipitation is addressed from the perspective of heavy metal and SOM sequestration. Because of the complex, heterogeneous, and highly dynamic nature of soil systems, further comprehensive studies that couple mineral dissolution/precipitation with other critical soil hydro-biogeochemical processes under broad spatial and temporal scales are expected to improve the mechanistic understanding and capability of model predictions of dynamic soil systems.

References

- Ball JW and Nordstrom DK (1991) *User's manual for WATEQ4F, with revised thermodynamic data base and text cases for calculating speciation of major, trace, and redox elements in natural waters*. 91–183.
- Brantley SL (2008) Kinetics of mineral dissolution. In: *Kinetics of Water-Rock Interaction*, pp. 151–210. New York: Springer.
- Chou L and Wollast R (1985) Steady-state kinetics and dissolution mechanisms of albite. *American Journal of Science* 285(10): 963.
- Fritz B and Noguera C (2009) Mineral precipitation kinetics. *Reviews in Mineralogy and Geochemistry* 70(1): 371–410.
- Hu SW, Liang YZ, Liu TX, Li FB, and Shi ZQ (2020a) Kinetics of As(V) and carbon sequestration during Fe(II)-induced transformation of ferrihydrite-As(V)-fulvic acid coprecipitates. *Geochimica et Cosmochimica Acta* 272: 160–176.
- Hu SW, Wu YD, Ding ZC, Shi ZQ, Liu TX, et al. (2020b) Facet-dependent reductive dissolution of hematite nanoparticles by *Shewanella putrefaciens* CN-32. *Environmental Science: Nano* 7(9): 2522–2531.
- Huang XP, Hou XJ, Song FH, Zhao JC, and Zhang LZ (2017) Ascorbate induced facet dependent reductive dissolution of hematite nanocrystals. *The Journal of Physical Chemistry C* 121(2): 1113–1121.
- Kleber M, Eusterhues K, Keilueit M, Mikutta C, Mikutta R, et al. (2015) Chapter one-mineral–Organic associations: Formation, properties, and relevance in soil environments. In: Sparks DL (ed.) *Advances in Agronomy*, pp. 1–140. Academic Press.
- Lu Y, Hu SW, Liang Z, Zhu MQ, Wang ZM, et al. (2020) Incorporation of Pb(II) into hematite during ferrihydrite transformation. *Environmental Science: Nano* 7(3): 829–841.
- Melton ED, Swanner ED, Behrens S, Schmidt C, and Kappler A (2014) The interplay of microbially mediated and abiotic reactions in the biogeochemical Fe cycle. *Nature Review Microbiology* 12(12): 797–808.

- Roden EE (2008) Microbiological controls on geochemical kinetics 2: Case study on microbial oxidation of metal sulfide minerals and future prospects. In: *Kinetics of Water-Rock Interaction*, pp. 417–467. New York, NY: Springer.
- Shi ZQ, Peltier E, and Sparks DL (2012) Kinetics of Ni sorption in soils: Roles of soil organic matter and ni precipitation. *Environmental Science & Technology* 46(4): 2212–2219.
- Shi ZQ, Hu SW, Lin JY, Liu TX, Li XM, et al. (2020) Quantifying microbially mediated kinetics of ferrihydrite transformation and arsenic reduction: Role of the arsenate-reducing gene expression pattern. *Environmental Science & Technology* 54(11): 6621–6631.
- Stumm W (1992) *Chemistry of the Solid-Water Interface: Processes at the Mineral-Water and Particle-Water Interface in Natural Systems*. New York: John Wiley & Son Inc.
- Yuan GH, Cao YC, Hao F, Jon G, Liu HY, et al. (2019) A review of feldspar alteration and its geological significance in sedimentary basins: From shallow aquifers to deep hydrocarbon reservoirs. *Earth-Science Reviews* 191: 114–140.

Reactions, redox potential and kinetics

Trolard Fabienne^{a,b} and Guilhem Bourrié^b, ^aINRAE—UMR 1114 Environnement Méditerranéen et Modélisation des Agro-hydrosystèmes—Bâtiment Climat—Agroparc, Paris, France; ^bAcadémie d'Agriculture de France, Paris, France

© 2023 Elsevier Ltd. All rights reserved.

This is an update of R.D. DeLaune, K.R. Reddy, Redox Potential, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, pp. 366–371, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00212-5>.

Introduction	297
Drivers of reactivity in soils	298
Heterogeneity, a need	298
Phase transitions	299
Oxido-reduction in soils	300
Oxido-reduction: The chemical point of view	301
Oxido-reduction: Application to natural environments	302
The rH measurements in the soils	303
Evidence for biotic–abiotic reactions	304
Layered double hydroxides: A solution for auto-organization of the reactions	307
Structure of layered double hydroxides (LDHs)	307
The ordering of cations within a sheet along the (a, b) plane	307
The ordering of anions within the interlayer along the (a, b) plane	307
Ordering along the c-axis	309
Spontaneous abiotic reduction of nitrate by fougérite green rusts	309
The LDHs as supports of the auto-organization of reactions	310
Conclusion	313
References	313

Abstract

Redox reactions are fundamental to life, and local heterogeneities in the soil create gradients in pH, redox and chemical concentrations that support chemical reactions. In situ monitoring of waters can distinguish abiotic and biotic reactions, and the analysis of temporal data makes it possible to establish the kinetics of these reactions. Field data show that the oxidation–reduction potential can vary over a wide range of pe in a few hours, and that fougérite can be a competitor of denitrification. Layered double hydroxides can drive self-organizing reactions, including redox reactions, usable for storage and replication of chemical information.

Key points

- redox reactions are fundamental to life.
- local heterogeneities in the soil create gradients in pH, redox and chemical concentrations that support chemical reactions.
- fougérite is a denitrification competitor.
- in situ monitoring of waters can distinguish abiotic and biotic reactions.
- the structure of the double layered hydroxides allows spontaneous redox reactions.
- auto-organization reactions are needed for information storage and replication.

Introduction

The soil, the fragile interface between the atmosphere, the hydrosphere and the lithosphere, is an environment and habitat for more than half, by mass, of the organisms living on Earth. Soils differ from mere weathered rock, because they show an approximately vertical differentiation, i.e., the soil horizon, that has been produced by the continual influence of percolating water and living organisms. Among the chemical reactions that act on the transformation of soils, oxidation–reduction reactions have a special place because by acting on the environment, they enable living organisms to recover the energy necessary for their metabolism. By studying both the solutions and the mineral and or organic constituents of the soil, redox reactions can be observed and measured in situ. The heterogeneity of the organization of soil constituents promotes the emergence of pH, redox and chemical

concentration gradients that control chemical reactions. Recent research shows that the mineralogical structure of double layered hydroxides can create steric environments that favor certain reactions, including oxidation–reduction reactions, information storage and replication.

Drivers of reactivity in soils

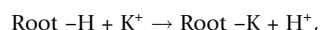
Heterogeneity, a need

Soils are porous media formed at the land surface by weathering processes mediated by biological, geological and hydrological phenomena. From the point of view of chemistry, the soil is a multi-component open environment, with biogeochemical systems containing solids, liquids and gases. That it is an open system means the soil exchanges both matter and energy with the surrounding atmosphere, biosphere and hydrosphere. These flows of matter and energy to, or from soils are highly variable in time and space, but they are the essential fluxes that cause the development of soils and govern the patterns of their functionalities, e.g., fertility... (Sposito, 2008).

Soil is a mixture of constituents in all three states, i.e., solid, liquid and gaseous (Fig. 1). We call “phase” a portion of matter bearing the same properties at any point in space (e.g., refraction index, density...). Geochemical and biochemical reactions redistribute matter and create local heterogeneities, sources of new possibilities for reaction. The gases in the soil mix very quickly, and despite the variability of the molecules that can constitute them, they form only one phase. Soil solutions do not mix as rapidly as gases.

Depending on the distribution and geometry of the solid constituents, pores of varying sizes, connected or not, define spaces where solutions of different chemical composition can co-exist. Finally, minerals with different crystal structures do not mix and each will be considered a separate phase.

For there to be a reaction between constituents in a soil, thermodynamic and or chemical, temperature, water state and or pressure imbalances must be present. There are two types of gradients: (i) gradients through an interface, either mineral or biological; (ii) gradients from one site to another, due to small diffusivity in the soil solution. Because of the small particle sizes, the number of atoms constituting the surface of the solid is proportionally larger than the number of atoms in the volume of the solid. For example, a solid in the form of a 20-nm diameter sphere exposes 5% of its atoms on its surface, while in the form of a 2-nm sphere, it exposes 60%. These imbalances are favored when physical and or chemical gradients are established between the constituents of the soil and the organisms that generate and maintain them throughout their life. This is observed, particularly in the rhizosphere where the mineral nutrition of plants takes place. In this process, there is a biological control, in addition to passive absorption. Plants take up nutrients mainly in the form of soluble cations, thus creating a charge deficit in the root environment. To maintain electrical neutrality, the plant emits protons toward the soil solution to compensate following a reaction such as:



where H is hydrogen and K is potassium.

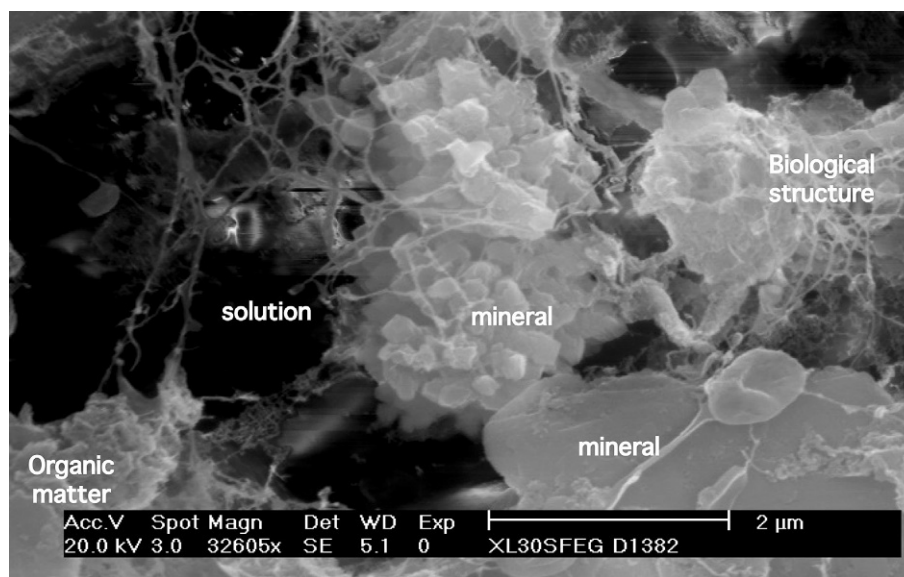


Fig. 1 Heterogeneities and diversity of constituents in a soil observed at the micrometric scale.

Phase transitions

Biogeochemical reactions in soils lead to phase transitions of several types. Changes of state (i.e., solid/liquid, liquid/gas, etc.) are always characterized by sharp variations in volume, V , and therefore in enthalpy, H , entropy, S , and specific heat, C_p , at the transition point T_1 . Thus, we distinguish the first order cooperative transition phases, which can be illustrated by the gypsum–anhydrite transition in gypsum and saline soils under arid conditions. This transition is marked by the departure of two water molecules per cell and a change of crystal system. At the transition temperature (at constant pressure) the chemical potential, μ , of the constituent in the two states is the same, but the curve of μ as a function of T presents a discontinuity, which is explained by the fact that the two constituents do not have the same density (Fig. 2a). There are other transitions where no discontinuity of volume, enthalpy or entropy is observed (Fig. 2b). We then speak of second-order cooperative transitions, or transition λ . The transition by change of conformation of a protein by reorganization of its main chain is an example.

In many situations in soils, phase transitions, which are neither first order nor second order, can be observed and are called intermediate transitions or co-operative phenomena (Fig. 3) (Williams and Frausto da Silva, 1997).

The difference between first- and second-order changes is the degree of co-operation (see some examples in Table 1). For a first-order change it is sharp and global, applying to all atoms in the sample. For a second-order change it is local but extensive, statistically distributed and fluctuating. It is more likely that second-order changes will occur in weakly interacting systems with many energy states. We need to note that for many polymers, for example proteins and polynucleotides, second-order changes are usual; this is fundamental in biology and has important consequences in interactions between living organisms and the components of soils.

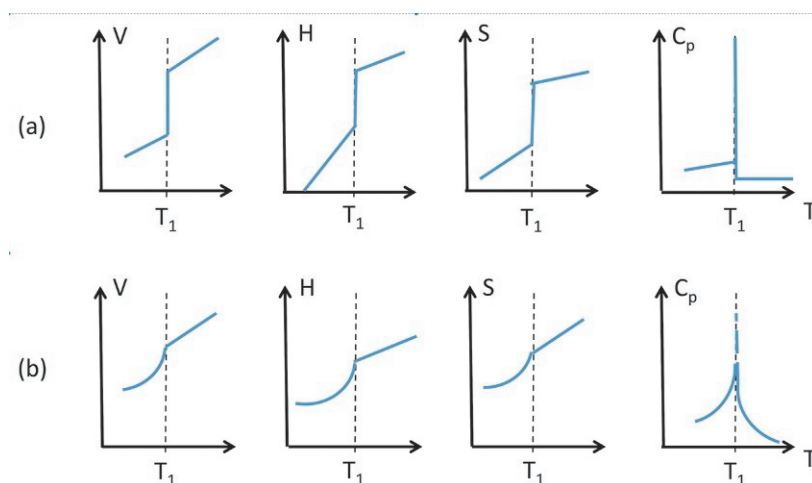


Fig. 2 Changes in thermodynamic properties accompanying first-order (a) and second-order (b) phase transitions at constant pressure. T is the temperature, T_1 the transition temperature, V is the volume, H the enthalpy, S the entropy and C_p the specific heat. From Trolard F, Bourrié G (2018) Chapter 4: Soils as bio-physicochemical reactors. In: Berthelin J, Munch JC, Coord Valentin C, Mariotti A Eds., Volume 1: Soils as a Key Component of the Critical Zone—Functions and Services, Geosciences Series, Soil Sets, ISTE, Wiley ed., pp. 81–115.

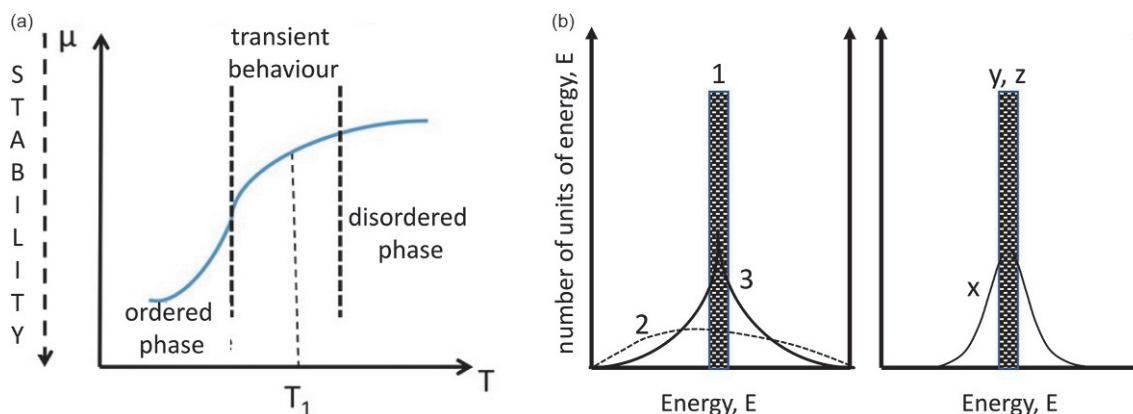


Fig. 3 (a) Variation of the chemical potential, μ , close to the temperature of a second-order transition process; (b) the distribution of energy over molecules or monomer units of a polymer in (1) a perfect crystal, (2) in solution, and (3) a weakly co-operative material such as some proteins; (c) the distribution of energy (E) within an energized polymer such as a protein that can oscillate their component parts along the x axis but not along the y or z directions. In each case the absolute value of E differs.

Table 1 Some examples of first- and second-order phase transitions.

	Transition	Comment
<i>First-order transition</i>		
S	Rhombic–monoclinic	Change in packing of S_8 molecules
$CaCO_3$	Aragonite–calcite	6- to 8- co-ordinate of Ca ions
<i>Second-order transition</i>		
MnO	Magnetic transformation	Spins aligned below Curie temperature
protein	Conformation change	Main chain reorganization
cell membrane + cholesterol	Fluidity change	Lipid packing altered

Modified from Williams RJP, Fraústo da Silva JJR (1997) *The Natural Selection of the Chemical Elements*, Oxford University Press Inc., New York, 646 pp., ISBN 0-19-855842-2.

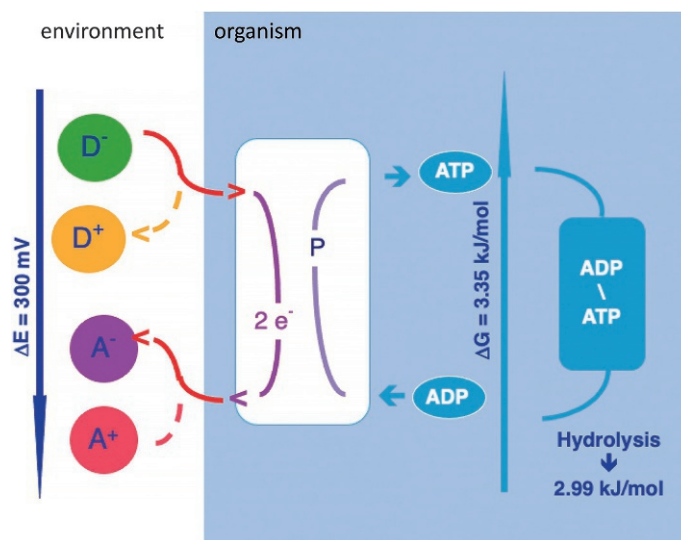
Oxido-reduction in soils

In soils, the geochemical systems where oxidation–reduction processes occur are non-equilibrium systems, and where thermodynamic equilibrium will often never be reached (Stumm and Stumm-Zollinger, 1971). For example, organic matter should not exist in the presence of most electron acceptors (e.g., SO_4^{2-} (sulfate), $FeOOH$ (ferric oxyhydroxide), NO_3^- (nitrate) . . .). However, organic matter exists in soils and can remain there for long periods (several thousand years). This is because to produce this organic matter, living organisms can decrease the significant kinetic barriers between redox couples and recover the energy of these relatively slow reactions (from milliseconds to minutes) to maintain the thermodynamic imbalances that allow them to live.

All manifestations of life are based on a series of chemical and physical processes, where all the transformations of matter are directly linked to changes in the energy situation of the cell. The system that allows the cell to have an energy reserve, available and for non-exclusive use, is oxidative phosphorylation (Fig. 4).

The phosphorylation of adenosine diphosphate (ADP) ($\Delta G = 3.35 \text{ kJ mol}^{-1}$) is coupled to a potential decrease in energy (about 300 mV) established between an electron donor (D) and an electron acceptor (A). The energy accumulator is adenosine triphosphate (ATP) and its hydrolysis releases 2.99 kJ mol^{-1} , 64% of which can be used freely by the organism. The ATP allows the cell to perform, independently and with a certain inertia all of the chemical, mechanical, osmotic and electrical work it needs. All the ingenuity of the living lies in the remarkable exploitation of this flow of electrons. Electrons do not circulate freely in the cell, they are exchanged by a series of discontinuous steps between an electron donor, which is in the reduced state, and an electron acceptor, which is in the oxidized state. The electrons thus culminate from one protein carrier to another to the final acceptor.

Non-equilibrium conditions suggest that the nature and extent of the redox reaction is controlled by several variables, such as the free energy gain produced by the reaction or the availability of suitable catalysts. In many cases, the catalytic activity is due to the enzymes of the microorganisms. But there are also several abiotic mediators of electron transport, such as organic compounds in a reduced state in an oxidizing medium or the surfaces of minerals containing reduced elements likely to oxidize, e.g., iron, manganese, sulfide.

**Fig. 4** Principle of the oxidative phosphorylation system.

When organic matter is no longer synthesized or repaired by the living organism and is not fossilized (e.g., coal, oil), it is transformed by oxidation into elementary mineral compounds; this oxidation is ambiguously called the “mineralization” of organic matter.

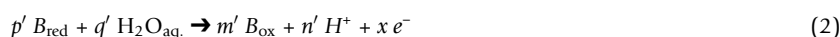
Redox-reactive minerals can serve as electron donors or acceptors both for abiotic reactants and or microbial metabolic processes.

Oxido-reduction: The chemical point of view

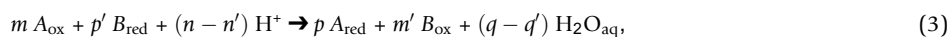
In chemistry, an oxidation–reduction reaction, or redox reaction, is a chemical reaction in which one or more electrons are transferred from one chemical species to another. Since the electron is not considered as an independent chemical species, demonstration of the transfer of electrons is obtained by splitting the oxidation–reduction reaction into two half-reactions: an oxidation reaction in which a chemical species acts as an electron donor and a reduction reaction in which another chemical species acts as an electron acceptor. Usually, the reduction reaction is written as:



the oxidation reaction is written as:



and the oxido-reduction reaction:



where A and B are the chemical species, whose index specifies the reduced or oxidized state of the component, $m, m', n, n', x, p, p', q$ and q' are the stoichiometric coefficients, e^- is the electron and H^+ is the proton in the solution.

A theory equivalent to that defining pH is proposed for “free electrons” of the solution. The Jorgensen’s oxidizability is defined by the negative value of the logarithm (base 10) of the activity of the electron in the solution, i.e., $pe = -\log[e^-]$. High positive values of pe preferentially point to the existence of oxidized compounds, i.e., electron acceptors while strongly negative values favor reduced compounds, i.e., free electron donors. Thus, by definition at equilibrium for the reaction:



the electron potential is:

$$pe = 1/x \{ -\log K_R + \log([B_{\text{ox}}]^{m'}/[B_{\text{red}}]^{p'}) - n' pH \} \quad (5)$$

with $[B_{\text{ox}}]$ the activity of compound B in oxidized state; $[B_{\text{red}}]$ the activity of compound B in reduced state; $pe = -\log[e^-]$ and $pH = -\log[H^+]$.

These equations enable us to predict the activity ratio of the proton acceptor and proton donor at any acidity state of an aqueous solution. In this way, pe - and pH -values are used to calculate the stability fields of redox species (e.g., Pourbaix’s diagrams) in a chemical system if it approaches equilibrium, independently of the time required and the reaction pathways necessary.

There are, however, several fundamental differences between the pH and pe values (Pfeifer, 2000). Acid–base equilibria always refer to the protolysis of the solvent, i.e., a proton acceptor reacts with water and de-protonates it, whereas a proton donor transfers protons to water to form H_3O^+ (hydronium ion). Hydrated protons are components of the solvent and quickly establish an equilibrium with dissolved acids or bases (this also holds true for the hydroxylated surfaces of many solid phases, Stumm, 1992). The pH is a well-defined variable for the intensity of proton transfer reactions, because a pH -electrode, i.e., the hydroxylated surface of a glass membrane, also establishes a rapid equilibrium with water. The pH can be measured and reproduced repeatedly in this way.

In contrast, redox equilibria refer to the standard hydrogen electrode. Consequently, pe describes the tendency to accept electrons from reactants, or donate them to the reactants and does not describe the extent of electron transfer from or to the solvent (Pfeifer, 2000). The life span of hydrated electrons is extremely small (of the order of 10^{-12} – 10^{-17} s). Thus, characterization of the redox state of an aqueous solution can be obtained only by the measurement of the activity ratios of certain redox couples, e.g., Fe^{2+}/Fe^{3+} , H_2/H_2O or H_2S/S^0 , that are not necessarily in equilibrium with the same value of pe .

In chemistry, the standard potential E^0 of the redox couples given in the tables is defined at $pH = 0$ and with the convention that the activity of the chemical species and the $[A_{\text{ox}}]/[A_{\text{red}}]$ ratio are equal to 1.

These conventions enable the classification of redox couples according to the value of the standard potential E^0 and thus prediction of the direction of the reaction (see for example, Fig. 5).

The direct application of this formalism, however, to a heterogeneous milieu such as soil is unsatisfactory because:

- the pH of a soil is never equal to 0;
- the activity of species in solution is rarely equal to 1;
- the activity of solids equal to 1 is an assumption that cannot be applied to minerals of variable chemical composition, or of small sizes; which excludes almost all minerals in the soil, that is 90 to 95% of its constituents.

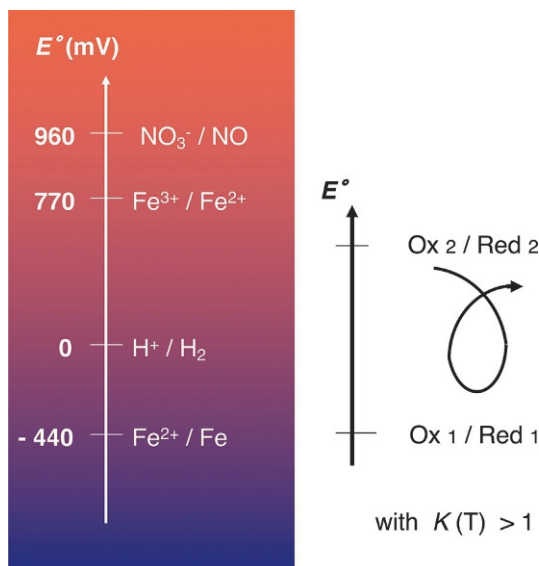


Fig. 5 Order of the redox couples defined according to the standard potential E^0 at pH = 0, $[A_{\text{ox}}]/[A_{\text{red}}] = 1$. The arrow indicates the direction of the possible oxidation–reduction reaction between two couples.

Oxido-reduction: Application to natural environments

The various oxidation–reduction reactions possible in soils cannot be classified simply by order of increasing standard potential E^0 (value of the potential of the electron, pe , when $[A_{\text{ox}}]/[A_{\text{red}}] = 1$ and given in tables). On the one hand, the stoichiometric coefficients of the reduced and oxidized species and of H^+ vary from one reaction to another and, on the other hand, some reactions are homogeneous in solution, and others are heterogeneous, solid and solution. It is possible, however, to order the various reduction reactions by establishing, by convention, at a given pH by calculating the value of the electron potential, pe , for which the ratio of the activities of the reduced and oxidized species is greater than a threshold convention such that the reduced species is largely dominant. This leads to defining of “ pe_{critical} ” and “ Eh_{critical} ” ($Eh_{\text{critical}} = 0.05916 \text{ } pe_{\text{critical}}$, where 0.05916 is the Nernst factor at 25 °C).

For homogeneous reactions, the conventional $[A_{\text{red}}]/[A_{\text{ox}}]$ ratio is taken as 10^6 . For example, the reduction of nitrate to nitrite is practically complete when pe exceeds the critical value: $pe_{\text{critical}} = 4.35$ ($Eh_{\text{critical}} = 256 \text{ mV}$). This value is derived from the equilibrium equation:

$$\log ([\text{NO}_2^-]/[\text{NO}_3^-]) + 2\text{pH} + 2pe = \log K^0 \quad (6)$$

With $\log K^0 = 28.64$ at 298.15 K and 10^5 Pa , pH = 7 and $[\text{NO}_2^-]/[\text{NO}_3^-] = 10^6$.

For heterogeneous reactions, where the oxidized species is solid and dissolves during reduction, A_{red} is taken as 10^{-7} , assuming that the solid phase is pure and its activity A_{ox} is 1. For example, for the reaction of reductive dissolution of goethite, $\alpha\text{-FeOOH}$:



One obtains $pe_{\text{critical}} = 1.0$ ($Eh_{\text{critical}} = 59.2 \text{ mV}$), from the relationship:

$$\log ([\text{Fe}^{2+}]/[\text{FeOOH}_s]) + 3\text{pH} + pe = \log K^0, \quad (8)$$

where $\log K^0 \approx 15$ at 25 °C, 10^5 Pa , with pH = 7, $[\text{FeOOH}_s] = 1$ and $[\text{Fe}^{2+}] = 10^{-7}$.

By confusing, as a first approximation, the activity of Fe^{2+} with its concentration in the solution, this corresponds to an analytical measurement limit of $10^{-7} \text{ mol. L}^{-1}$ in the solution.

From thermodynamic data of the most stable species (Table 2), it is possible to establish a scale of sequential reduction potentials of N(V)/N(IV), Fe(III)/Fe(II), S(VI)/S(-II)... The classical order is obtained and steps larger than 100 mV separate the successive redox couples. However, the outcomes are quite different when labile or metastable species are considered, e.g., fougérite green rusts (GRs) for the iron system (Trolard and Bourrié, 1999). By placing the different iron pairs on the same pe_{critical} diagram as a function of pH, Fig. 6 shows that the values are distributed over a large area within the area of water stability. This suggests that: (i) the transfer of electrons in an iron-containing system at a given pH can occur over the entire domain of water stability (cf “Iron in soils” in this encyclopedia), because the $\text{Fe}^{\text{III}}/\text{Fe}^{2+}$ couple can respond to different levels of activation; (ii) metastable species can significantly alter redox potential values and thus mineral–solution equilibria and (iii) these species should not be overlooked, in particular, when seeking to identify the mineral which controls the activity of iron in solution.

Table 2 Thermodynamic data used to compute the pe_{critical} values.

Réactions	$\log K^0$	$pH = 6.0$	pe_{critical} $pH = 7.0$	$pH = 8.5$
$1/4 \text{ O}_2(\text{g}) + \text{H}^+ + \text{e}^- \rightleftharpoons 1/2 \text{ H}_2\text{O}(\text{l})$	20.78	14.6	13.6	12.1
$1/8 \text{ NO}_3^- + 5/4 \text{ H}^+ + \text{e}^- \rightleftharpoons 1/8 \text{ NH}_4^+ + 3/8 \text{ H}_2\text{O}$	14.88	6.7	5.4	3.5
$\text{NO}_3^- + 2 \text{ H}^+ + 2 \text{ e}^- \rightleftharpoons \text{NO}_2^- + \text{H}_2\text{O}$	28.57	5.3	4.3	2.8
$\text{Fe}(\text{OH})_3(\text{s}) + 3 \text{ H}^+ + \text{e}^- \rightleftharpoons \text{Fe}^{2+} + 3 \text{ H}_2\text{O}$	15.87	4.9	1.9	-2.6
$1/8 \text{ SO}_4^{2-} + \text{H}^+ + \text{e}^- \rightleftharpoons 1/8 \text{ S}^{2-} + 1/2 \text{ H}_2\text{O}$	2.52	-4.2	-5.2	-6.7
$\text{CR}(\text{Cl}^-) + 2 \text{ H}^+ + 1/4 \text{ e}^- \rightleftharpoons \text{Fe}^{2+} + 1/4 \text{ Cl}^- + 2 \text{ H}_2\text{O}$	10.92	(20.7) ^a	(16.7)	4.7
$\text{CR}(\text{SO}_4^{2-}) + 2 \text{ H}^+ + 1/3 \text{ e}^- \rightleftharpoons \text{Fe}^{2+} + 1/6 \text{ SO}_4^{2-} + 2 \text{ H}_2\text{O}$	10.34	(16.5)	10.5	1.5
$\text{CR}(\text{CO}_3^{2-}) + 2 \text{ H}^+ + 1/3 \text{ e}^- \rightleftharpoons \text{Fe}^{2+} + 1/6 \text{ CO}_3^{2-} + 2 \text{ H}_2\text{O}$	9.76	(14.8)	8.8	-0.2
$\text{CR}(\text{OH})^- + 8/3 \text{ H}^+ + 2/3 \text{ e}^- \rightleftharpoons \text{Fe}^{2+} + 8/3 \text{ H}_2\text{O}$	15.47	9.7	5.7	-0.3
$\gamma\text{FeOOH} + 3 \text{ H}^+ + \text{e}^- \rightleftharpoons \text{Fe}^{2+} + 2 \text{ H}_2\text{O}$	16.65	5.7	2.7	-1.9
$\alpha\text{FeOOH} + 3 \text{ H}^+ + \text{e}^- \rightleftharpoons \text{Fe}^{2+} + 2 \text{ H}_2\text{O}$	14.97	4.0	1.0	-3.5

^aValues in parentheses are not valid in the stability domain of water.

Modified from Trolard F, Bourrié G (1999) Influence des oxydes de fer de type « rouilles vertes » sur les séquences d'oxydo-réduction dans les sols. *Comptes-rendus de l'Académie des Sciences, Paris, Sciences de la terre et des planètes/Earth & Planetary Sciences* 329: 801-806.

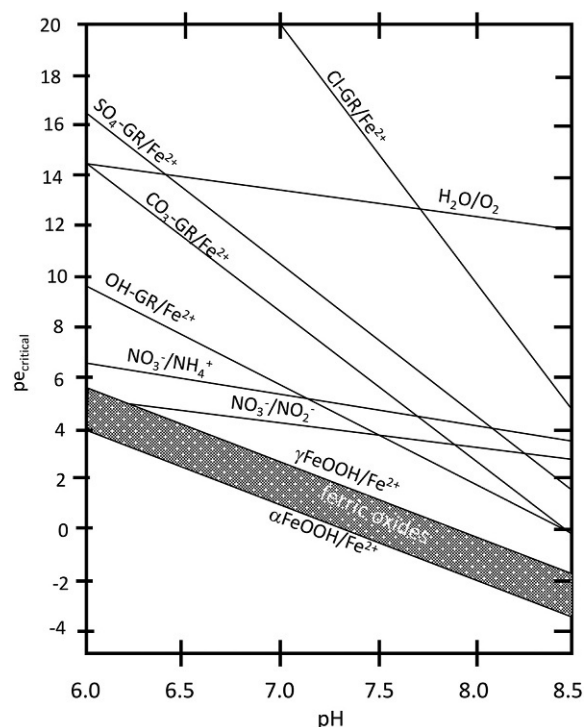


Fig. 6 Variation in the critical potential, pe_{critical} in a system where fougurite green rusts may be present. Cl-GR = chloride fougurite, $\text{SO}_4\text{-GR}$ = sulfate fougurite; $\text{CO}_3\text{-GR}$ = carbonate green rust; OH-GR = fougurite, αFeOOH = goethite, γFeOOH = lepidocrocite. Modified from Trolard F, Bourrié G (1999) Influence des oxydes de fer de type « rouilles vertes » sur les séquences d'oxydo-réduction dans les sols. *Comptes-rendus de l'Académie des Sciences, Paris, Sciences de la terre et des planètes/Earth & Planetary Sciences* 329: 801–806.

The critical potential scales must therefore be constructed according to: (i) the labile, metastable and stable species that control the redox potential in solution, therefore the mineralogy, including nano-minerals and soil geochemistry; (ii) the pH of the environment and taking into account concentrations of the elements present in the soil solution, therefore according to the water geochemistry.

The rH measurements in the soils

The diagnostic criterion of reductic or gleyic properties is an rH value < 20 , with $rH \equiv -\log p\text{H}_2$ where $p\text{H}_2$ is the partial pressure of hydrogen, and rH is related to the redox potential, Eh, and pH by the relationship:

$$rH = Eh/0.029 + 2 \text{ pH}, \quad (9)$$

where Eh is expressed in mV and 0.029 is the approximate value of $\ln_{10}RT/2F$ at 298.15 K, $R (=8.3144598 \pm 0.0000048 \text{ J mol}^{-1} \text{ K}^{-1})$ is the ideal gas constant, $F (=96,485.33289 \pm 0.00059 \text{ C mol}^{-1})$ is the Faraday constant and T is the absolute temperature in kelvin.

It involves measuring both Eh and pH in the soil solution and the sample temperature; rH is related to pe and pH by:

$$rH = 2(pe + \text{pH}) \quad (10)$$

$$pe = \frac{F \text{ Eh}}{\ln_{10}RT}, \quad (11)$$

where Eh is the redox potential relative to the normal hydrogen electrode in volts.

The pH, Eh and temperature measurements recorded with a multiparametric probe in Fougères' forest in Brittany (France) and a rice (*Oryza sativa* L.) field in the Camargue, Provence (France) show large variations covering the whole domain of existence of aqueous Fe(II) and Fe(III) oxides (Fig. 7).

In this example, although the geological, pedo-climatic and soil formation contexts are very different between the forest of Fougères in Brittany and the rice fields in Camargue, Fig. 7 shows similarities between the recordings of Eh–pH dynamics between the two sites. In Brittany, the soils result from the alteration of granitic rocks in an oceanic climate and the study site is in a multisecular beech (*Fagus* L.) forest, while the soils of the Camargue result from carbonated fluvial sedimentary deposits in a Mediterranean climate, and the study site is in cultivated and irrigated rice fields. In both situations, hydromorphic soils are observed with more or less pronounced waterlogging depending on the season. The data on the free water of the soils were acquired in Fougères during winter and spring, and in the Camargue during the irrigation season, from April to October, at the rate of one measurement every hour for several months. The variation in pe–pH shows a definite pattern, with a fast decrease in pe, stabilization of pe between -3 and -4 , and oscillations of the system between two opposite states, linked to infiltration of oxygenated rainwater: an oxidizing and reducing environment with very few points between.

According to equation 10, the rH values in Fougères vary between 4 and 14.4, and in the Camargue between 4 and 8.5; these values are much lower than 20, which is the upper bound set for reducing media. The evolving patterns can almost be superimposed despite differences in the soil and climate, which suggests that the biogeochemical processes responsible for these dynamics are the same.

Evidence for biotic–abiotic reactions

The dynamics of rice growth in waterlogged rice field soils of the Camargue, which start with a few kilograms of seed per hectare to become 8–10 tons of dry matter per hectare at harvest in less than 6 months also involves strong biogeochemical dynamics, in particular pe–pH. Alternate flooding and drying conditions at the beginning of the crop season aim to get rid of the natural variety of rice that loses its grains (domesticated rice keeps its grain) and of some other weeds. Reducing conditions are maintained by flooding during all the cultivation period until just before the harvest. Two plots were instrumented and monitored during a complete rice growing season: in plot R178 the organic matter content in soil was very small (less than 3%) because the straw from the previous year's harvest was burned; in plot R179 the organic matter content was larger because the straw had been incorporated into the soil for several years during post-harvest tillage (Nawaz et al., 2013).

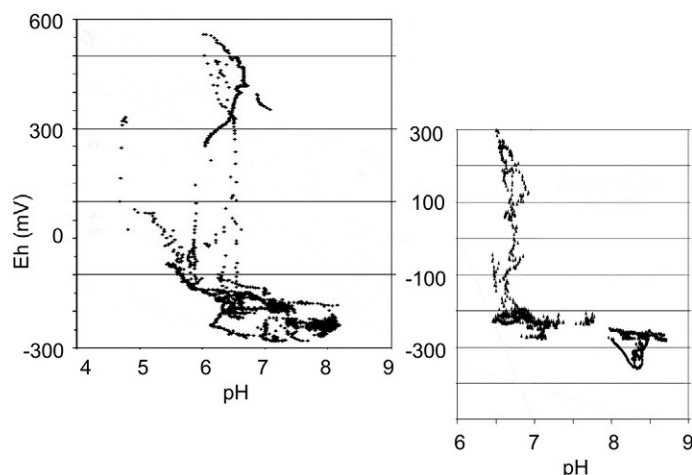


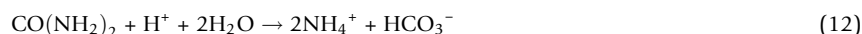
Fig. 7 Eh–pH dynamics registered in soil-free water using a multiparametric probe, in Fougères (a) and Camargue (b). From Bourrié G and Trolard F (2018) Chapter 1: Water quality in soils. In: Bourrié G (ed.), Soils as a Key Component of the Critical Zone, Vol. 4, Soils and Water Quality. Geosciences Series, Soils Set, Fig. 1.21, pp. 1–72. ISTE and Wiley Editions.

Fig. 8 shows that pe decreases very rapidly (points every hour are clearly distinct), then have low and quasi-constant values close to $pe = -4$. After the initial decrease in pe at quasi-constant pH, differences in pH occur at quasi-constant pe .

When compared with “critical” pe values, it seems that nitrate is largely unstable with respect to ammonium. When pe decreases, the system successively crosses the thermodynamic stability lines of lepidocrocite, goethite, hematite, and ultimately that of natural green rust, fougérite. The steady pe value is in the range of -3 to -4 . This is higher than the “critical” pe for methane formation and sulfate reduction, which are almost the same ~ -5 . This means that methane emission is possible, but should be limited by the abundance of sulfate due to the proximity of the sea.

A series of pH values were recorded during the first 45 days after flooding, just before pH returns to pH_{stat} as shown in Fig. 9.

The differences in pH are partly due to water management practices, the pH of irrigation water (Rhône river) is 0.5 of a unit higher than the pH of the soil solution, but they mostly result from urea applications applied 2 weeks before flooding. Urea is applied as the N fertilizer in both R178 and R179 plots, which results in a pH increase, according to the reaction:



Maximum pH occurs about 45 days after flooding. Then, pH relaxes to equilibrium, according to an exponential law. The different straw management techniques (burned straw versus straw incorporation) reflect in different characteristic times of relaxation: relaxation time with straw burning is almost twice as long (966 h) as that for straw incorporation (469 h). This correlated with the smaller dissolved organic carbon content and alkalinity of the soil solution, and to the diminution of surface acid–base sites in the organic matter of crop residues. Straw burning, therefore, decreases the buffering capacity of pH. The differences between measured and predicted pH by the relaxation exponential law show definite, regular oscillations (Fig. 10). The period of these oscillations is 24 h, as shown by the Fourier transforms (Nawaz et al., 2013).

Fig. 10 shows a sharp increase in pH after midnight and before sunrise, and a relatively slow decrease in pH in the afternoon. Temperature and pH are in opposite phase: pH is maximum when temperature is minimum and vice versa. During the night, in the absence of root absorption, the abiotic phenomenon of urea hydrolysis is dominant and results in an increase in pH; maximum pH is observed at night, with minimum temperatures. During the day, the biotic process of ammonium absorption by roots is dominant with the simultaneous release of H^+ that results in a decrease in pH due to the proton pump, and the classical rhizosphere effect (Nye, 1981); minimum pH is occurs during the day, with maximum temperatures.

Redox potential is poised by the dissolution or precipitation of iron oxides involving Fe(III) oxides (*sensu lato*), lepidocrocite, goethite, hematite and mixed Fe(II)—Fe(III) hydroxide, fougérite. This limits pe to the narrow range of -4 to -3 , which favors NH_4^+ (ammonium), although non negligible concentrations of nitrate are observed. Moreover, it prevents reduction from proceeding

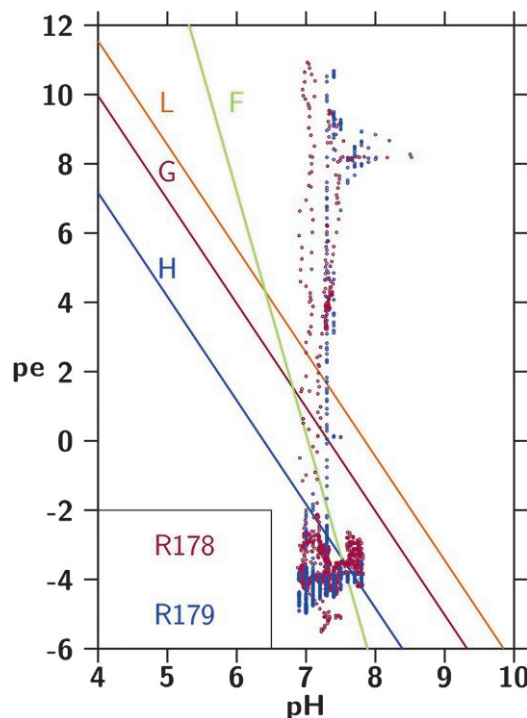


Fig. 8 (pe , pH) variations in the Camargue with straw burned (R178) and straw incorporated (R179). Stability lines are drawn for $pFe = -7$ and the solubility products of the following iron oxides: G = goethite, H = hematite, L = lepidocrocite, F = fougérite (here computed with the formula $Fe_3(OH)_7$). Modified from Nawaz MF, Bourrié G, Trolard F, Mouret J-Cl, Henry P (2013) Effects of agronomic practices on the physico-chemical properties of soil waters in rice culture. *Turkish Journal of Agriculture and Forestry* 37: 195–202.

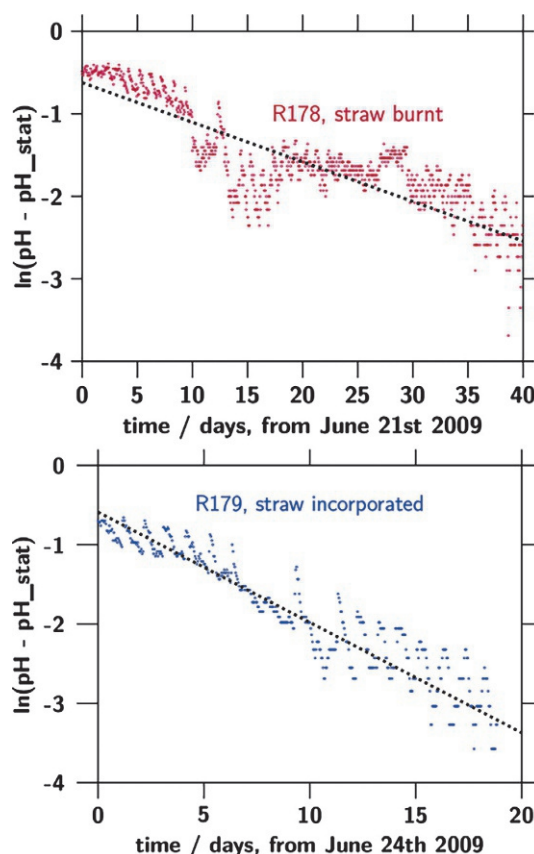


Fig. 9 Differences in average pH_{stat} and pH values over time in the soil solution of the different rice plots. For R178 (a), from 21/06/2009 to 09/08/2009 and for R179 (b); from 24/06/2009 to 14/07/2009. Points are measured data and the dotted curves are the fitted exponential equations: $\ln |pH - pH_{stat}| = -t/\tau$, where τ is the time of relaxation. Modified from Nawaz MF, Bourrié G, Trolard F, Mouret J-Cl, Henry P (2013) Effects of agronomic practices on the physico-chemical properties of soil waters in rice culture. *Turkish Journal of Agriculture and Forestry* 37: 195–202.

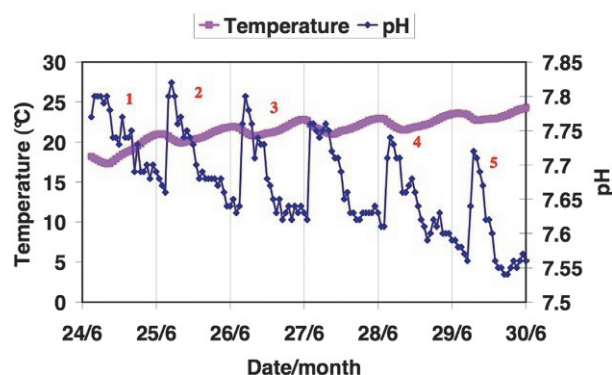


Fig. 10 Temperature and pH daily variations for R179.

toward extensive sulfate reduction and methanogenesis, though small amounts of sulfide were detected but probably limited by FeS (mackinawite) formation. With the large amount of sulfate from the Mediterranean-sea sulfate reduction is likely to prevent methanogenesis effectively in anaerobic pathways.

Redox reactions in soils are both biotic and abiotic. Transitions from the reduced to oxidized state may be controlled by differences in pe, but also by variation in pH at constant pe. These transitions are influenced by availability of oxygen from the air or as dissolved oxygen in rain. They are also affected by the variation in soil moisture, by waterlogging and fluctuations of the water table, by the input of organic matter, nitrogen fertilizers, such as urea, and by the proton pump. The variation in the controlling variables, pe and pH, in turn affect metabolic activities of aerobic and anaerobic microorganisms.

Layered double hydroxides: A solution for auto-organization of the reactions

If most of the oxido-reduction reactions in soils are controlled by living organisms, some minerals such as layered double hydroxides (LDHs) can, by their architecture, create abiotic environments favorable to redox reactions, store information and orientate the transfer of compounds.

Structure of layered double hydroxides (LDHs)

In soils, the clay fraction, defined by particles less than 2 μm in size, contains several families of minerals including phylites, in particular. When the structural layers are based on silica tetrahedra, they form clay minerals whose negative surface charge is compensated for by interlayer cations, which can be exchanged depending on the environmental conditions, i.e., cation exchange capacity. Other structural layers based on hydroxides form the layered double hydroxides (LDHs), whose positive surface charge is compensated by interlayered anions. These clays, improperly called “anionic clays,” whose structure allows anion exchange, also have an electrochemical capacity and magnetic properties.

The LDH compounds form a large group of minerals (Table 3) including green rusts, the natural form of which is fougérite.

The LDH structure is formed of stacked layers of edge-sharing metal octahedra containing divalent and trivalent metal ions separated by anions within the interlayer spaces. The range of composition of LDHs is very narrow and they differ mainly by the nature of the interlayer anion. The number of anions that may be intercalated per unit weight of LDH, known as the anion-exchange capacity depends on the charge density of the layer, i.e., the number of M^{3+} cations within the layer. The general structural formula can be expressed as:

$$[M^{II}_{1-x} M^{III}_x(\text{OH})_2] [x/nA^{-n}, m\text{H}_2\text{O}], \quad (13)$$

where, for example, M^{II} can be Mg (magnesium), Ni (nickel), Co (cobalt), Zn (zinc), Mn (manganese) and Fe (iron) and M^{III} can be Al (aluminum), Fe and Cr (chromium) (Sparks, 2003). All sites of the layers $[M^{II}_{1-x} M^{III}_x(\text{OH})_2]$ are occupied, therefore, the layer has a net positive charge x per formula unit, which is balanced by an equal negative charge from an interlayer anion A^{-n} , such as Cl^- (chloride), Br^- (bromide), I^- (iodide), NO_3^- (nitrate), OH^- (hydroxide), ClO_4^- (perchlorate) and CO_3^{2-} (carbonate). Water molecules are also present in the interlayer.

When considering order within the structure of LDH minerals there are three aspects to consider: (i) cations within a sheet along the (a, b) plane, (ii) anions within the interlayer along the (a, b) plane and, (iii) the order imparted when the LDH layers stack along the c [001] direction. The LDHs are the only known mineral structures that have both great plasticity in the (a, b) plane and fidelity along the c axis. For example, the structure of green rusts, which are particular LDHs, can react both in the structural brucite/amakinite layer, and in the interlayer space with more than 35 different chemical elements (Trolard et al., 2022).

The ordering of cations within a sheet along the (a, b) plane

In the hydroxide sheet, the octahedra of hydroxyl ligands are occupied by cations M with charges of +2 or +3. However, with Fe-LDH, to maintain the stability of the sheet, the close neighbors of an Fe^{3+} cation can be only divalent cations M^{2+} , otherwise spontaneous electrostatic repulsion of two neighboring Fe^{3+} cations leads to an oxolation reaction with the loss of an H^+ (Bourrié et al., 2004). This imposes an M^{2+}/Fe^{3+} stoichiometric ratio of 2 (Fig. 11a). However, for the structure of LDH to be stable, a minimum of Fe^{3+} must be present in the structure of the hydroxide sheet. Experimentally, it has been shown that a quarter of the sites must be occupied by a trivalent cation and the distance between two trivalent cations should be at least, $a\sqrt{3}$.

The ordering of anions within the interlayer along the (a, b) plane

Anions in LDH interlayers are often considered dynamic, precluding the possibility of long-range order, but the geometry of the anion leads to different modes of their organization in the interlayer. For a given LDH, interlayered anions compensate for the excess positive charge of the layer. However, steric hindrance limits the number of anions the interlayer can accommodate, which depends on the charge and size of the anion. For monovalent spherical anions such as OH^- , Cl^- , F^- (fluoride), Br^- , ... the first-nearest neighbor sites of the anion site cannot be occupied by another anion. The minimal distance between two spherical

Table 3 Structural formula of some natural layered double hydroxides.

Mineral	Structural formula	Interlayer anion
Fougérite	$[\text{Mg}_y \text{Fe}^{II}_{1-x} \text{Fe}^{III}_x (\text{OH})_2] [x/n A^{-n}, m\text{H}_2\text{O}]$	Possible: OH^- , Cl^- , CO_3^{2-} ...
Meixnerite	$[\text{Mg}_6 \text{Al}_2 (\text{OH})_{16}] [(\text{OH})_2, 4 \text{H}_2\text{O}]$	OH^-
Woodallite	$[\text{Mg}_6 \text{Cr}_2 (\text{OH})_{16}] [(\text{Cl}^-)_2, 4 \text{H}_2\text{O}]$	Cl^-
Iowaite	$[\text{Mg}_4 \text{Fe}^{III} (\text{OH})_{10}] [(\text{Cl}^-), 2 \text{H}_2\text{O}]$	Cl^-
Takovite	$[\text{Ni}_6 \text{Al}_2 (\text{OH})_{16}] [(\text{CO}_3^{2-}, \text{OH}^-), 4 \text{H}_2\text{O}]$	OH^- , CO_3^{2-}
Hydrotalcite	$[\text{Mg}_6 \text{Al}_2 (\text{OH})_{16}] [(\text{CO}_3^{2-}), 4 \text{H}_2\text{O}]$	CO_3^{2-}
Pyroaurite	$[\text{Mg}_6 \text{Fe}^{III} (\text{OH})_{16}] [(\text{CO}_3^{2-}), 4 \text{H}_2\text{O}]$	CO_3^{2-}

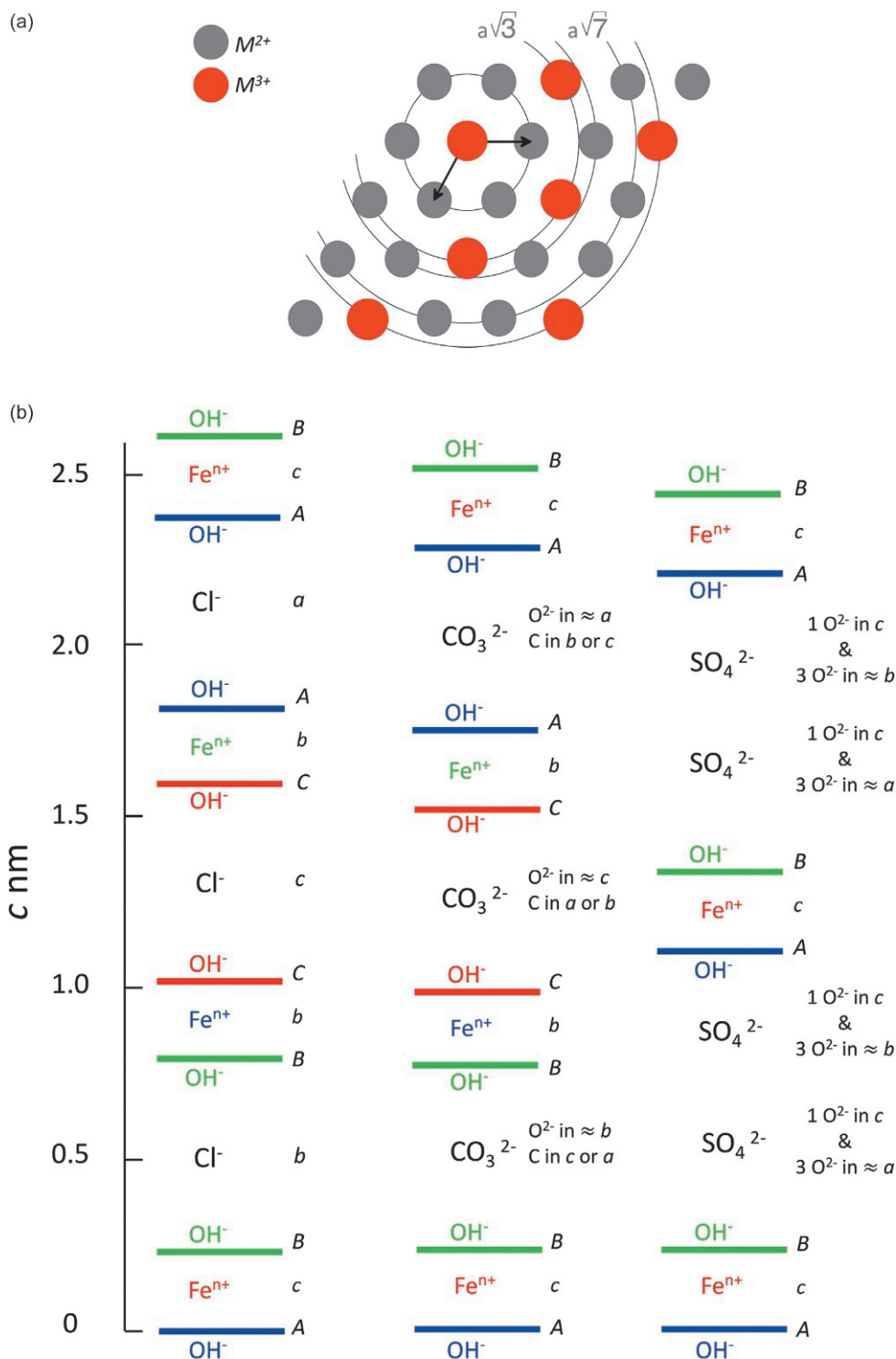


Fig. 11 (a) Possible positions of Fe^{3+} in the hydroxide sheet; (b) Stacking sequence at the scale of M cation hydroxide layers, anion interlayers and OH^- layers, along the threefold axis of Fe-LDH (Cl^-), Fe-LDH (CO_3^{2-}) and Fe-LDH(SO_4^{2-}) structures. A, B, C or a, b, c positions represent the sites in the hexagonal pavement, i.e., geometric assemblage of OH^- ions. Panel A: Modified from Bourrié G, Trolard F, Refait Ph, Feder F (2004) A solid solution model for Fe(II) - Fe(III) - Mg(II) green rusts and fougurite and estimation of their Gibbs free energies of formation. *Clays and Clay Minerals* 52: 383–395. Panel B: Modified from Génin JMR, Ruby C (2004) Anion and cation distributions in Fe(II)-Fe(III) hydroxysalt green rusts from Mössbauer analysis (carbonate, chloride, sulphate, ...); the “fougurite” mineral. *Solid State Sciences* 6: 705–718. Doi: 10.1016/j.solidstatesciences.2004.03.021.

anions is $a\sqrt{3}$ and the maximum number of anions, on the basis of 2 OH per mole formula, would reach 1/3 (Génin and Ruby, 2004). There is only one layer of water molecules in the interlayer and the structure is rhombohedral, with space group R-3m. For divalent anions with a three-dimensional geometry such as sulfate, oxalate, selenate. . . , there will be two layers of water molecules and the structure is a trigonal type with a P-3m1 space group (Fig. 11b). The screening effect of the two water layers allows more freedom for the lateral position of anions. Large organic anions like benzene sulfonate, terephthalate or larger monovalent aliphatic carboxylate anions, Cx, . . . may undergo rotations and site-hopping motions at high frequencies, as evidenced by using combined Nuclear Magnetic Resonance (NMR), Infrared (IR) Spectroscopy and digital modeling (Greenwell and Coveney, 2006). Thus, a long-range order of anions in the interlayer is not needed in the (*a*, *b*) plane, but the correlated distribution of both cations and anions is propagated through the crystal structure along the threefold *c*-axis.

Ordering along the *c*-axis

The existence of super-periodicity (propagation of cation ordering along the *c*-axis is a result of not only cation ordering, but also of cation–anion ordering; the existence of cation ordering is only likely at M^{II}/M^{III} ratios of 2 as only one arrangement can exist where the positive M^{III} centers are not adjacent (Drits and Bookin, 2006). Without being exhaustive, the existence of a long-range order along the crystallographic *c*-axis is important for information storage within LDH crystals and subsequent transfer of information at their surface.

Spontaneous abiotic reduction of nitrate by fougérite green rusts

Experimentally, while in solution NO_3^- cannot be reduced by Fe^{2+} , but chloride (Cl-GR1) or sulfated (SO_4^- GR2) green rusts reduce nitrates stoichiometrically into ammonium in several hours under abiotic conditions (Hansen et al., 2001). The rates of reduction increase with increasing nitrate concentrations up to some threshold concentrations, depending on the initial conditions of the reaction, above which no further increase in reduction rates takes place. The kinetics of the reaction depends on the type of interlayer anion, the layer charge and relative content of Fe^{II} in the hydroxide layers. In addition, pH modifies the rate of reduction by a direct effect of H^+ in the redox reaction following first-order kinetics and the proton affects nitrate adsorption onto reactive sites of iron grains (Huang and Zhang, 2004).

Thermodynamic calculations show that the reduction of nitrate proceeds largely before the reduction of ferric oxides, following the classical sequence when fougérite green rusts are absent (Trolard and Bourrié, 1999). In the field when fougérite is present, and irrespective of the nature of the compensating anion, the redox sequence is determined by the pH of the milieu (Fig. 12). This suggests that nitrate and fougérite can compete as electron acceptors for microflora in anaerobic conditions. At pH 6, redox buffering by fougérite protects nitrate and thus bypasses denitrification. At pH 8.5, denitrification is possible in presence of fougérite. However, in this case, nitrate can be spontaneously reduced by fougérite to NH_4^+ .

The layered structure of the mineral explains the particular reactivity of fougérite in abiotic conditions. For example, Fig. 13 summarizes the steps of reduction of nitrate by SO_4 -GR2: the first step is an anion exchange between nitrate and sulfate in the interlayer of the GR, and SO_4 -GR2 transforms into unstable NO_3 -GR1.

Under these conditions, a minimum of 8 divalent Fe^{2+} are immediate neighbors of NO_3^- . The reduction of NO_3^- to NH_4^+ can thus proceed spontaneously with simultaneous transfer of 8 electrons from Fe^{2+} to N. In the layer, the $\text{Fe}^{II}\text{—OH—Fe}^{II}$ bonds transform into $\text{Fe}^{III}\text{—O—Fe}^{III}$ by oxolation, releasing $8e^-$ and 8H^+ ; four additional H^+ equilibrate the two sulfate anions released; $8e^-$ and 10H^+ are necessary for each NH_4^+ , so that there is a net release of 2 protons. The GR oxidizes into hematite.

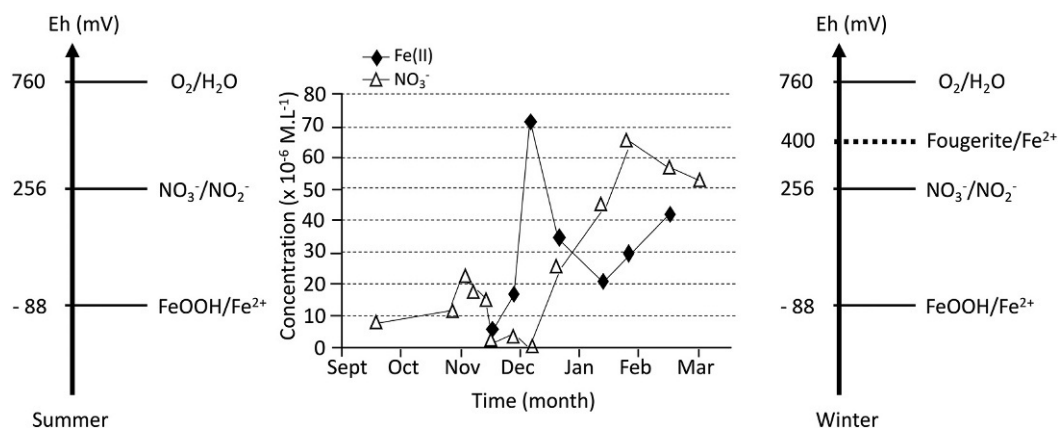


Fig. 12 Field dynamics of Fe(II) and NO_3^- in soil solution. At the end of summer, no Fe(II) is present in solution. At the beginning of autumn, nitrate concentration decreases to zero before Fe(II) is released in solution. This is the classical scheme (a). Then, in winter, fluctuations of Fe(II) are observed due to partial entry of oxygen from rainwater. Fougérite precipitates and competes with nitrate inducing the coexistence of N(V) and Fe(II) in solution, which is well explained by the occurrence of the redox couple fougérite/ Fe^{2+} in the system (b). From Trolard F, Bourrié G (2008) Geochemistry of green rusts and fougérite: A reevaluation of Fe cycle in soils. *Advances in Agronomy* 99: chapter 5, 227–287.

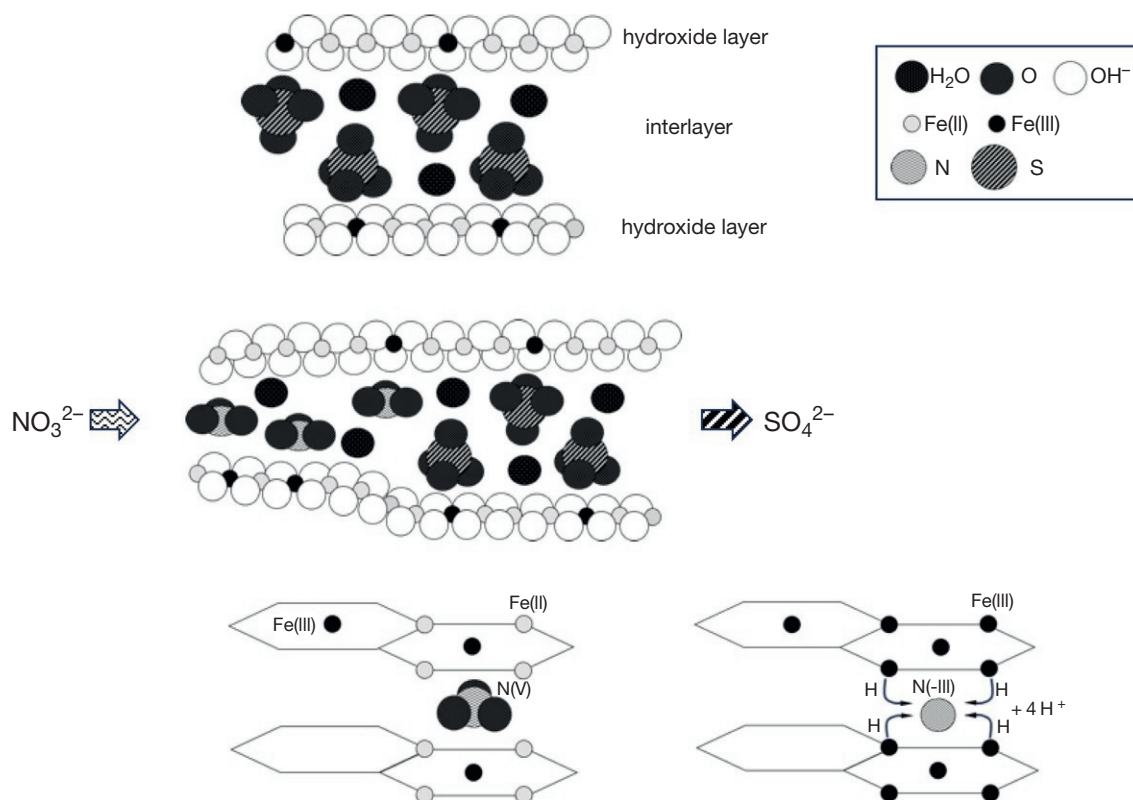
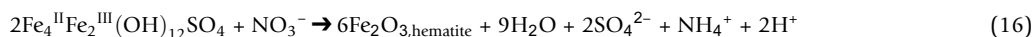
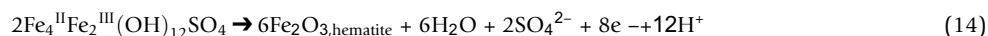


Fig. 13 Reaction scheme for the abiotic reduction of nitrate by $\text{SO}_4\text{-GR2}$: (a) initial state of the interlayer in $\text{SO}_4\text{-GR2}$; (b) exchange of sulfate by nitrate and collapse of the interlayer; (c) initiation of reduction of nitrate by eight neighboring Fe^{2+} ions, and final state of ammonium production before ammonium release to solution. Modified from Trolard F, Bourrié G (2008) Geochemistry of green rusts and fougérite: A reevaluation of Fe cycle in soils. *Advances in Agronomy* 99: chapter 5, 227–287.

The half-reactions and the net reaction can be written as follows:



The LDHs as supports of the auto-organization of reactions

The relationship between information theory and the evolution of life has become a research area in its own right (Avery, 2003) that requires exploration of abiotic mineral–organic combinations favorable to the self-organization of reactions, which are essential for the storage, replication and transfer of information. The layered double hydroxide compounds, in addition to concentrating and reacting with or catalyzing reactions of important prebiotic molecules, provide a superior coding environment relative to the cationic clays and could have served as prebiotic “ribozymes,” which would eventually be the first RNA materials to be produced and, ultimately, genetic takeover by nucleic acid systems occurred (Greenwell and Coveney, 2006).

The replication of an LDH is constrained by the distribution of interlayer anions, which reproduces the charge distribution of the octahedral layers. If in the plane (a , b) the distribution is variable within certain limits (see above), in the direction perpendicular to the sheets along the c -axis the copy is constrained. Thus, periodic boundary conditions are imposed in all three space directions (Fig. 14). The ideal dianions would have some directional nature with respect to the two negative charges, and the charge would neutralize one positive site on the LDH sheet above, and one on the LDH sheet below. If we consider starting with one LDH sheet and placing the ideal dianions at positive sites, the dianion has one negative charge to satisfy (Fig. 14a). When a second LDH sheet is considered (Fig. 14b), the LDH sheet charge sites not neutralized from the dianions in the previous layer are filled to give a string of numbers opposite to the previous one. A different arrangement of anions would have a different information content.

This conceptual model can be extended further by considering the nature of the binary information that is displayed at the surface of LDH crystals (Fig. 15). If the same mineral has polytypes with a two-layer repeat or three-layer repeat (Fig. 15b), this has important implications for the information encoded at the external surfaces of the LDH minerals. In a two-layer repeat system (Fig. 15a), the arrangement of sites on either side of the crystal will be complementary, rather than mirror images.

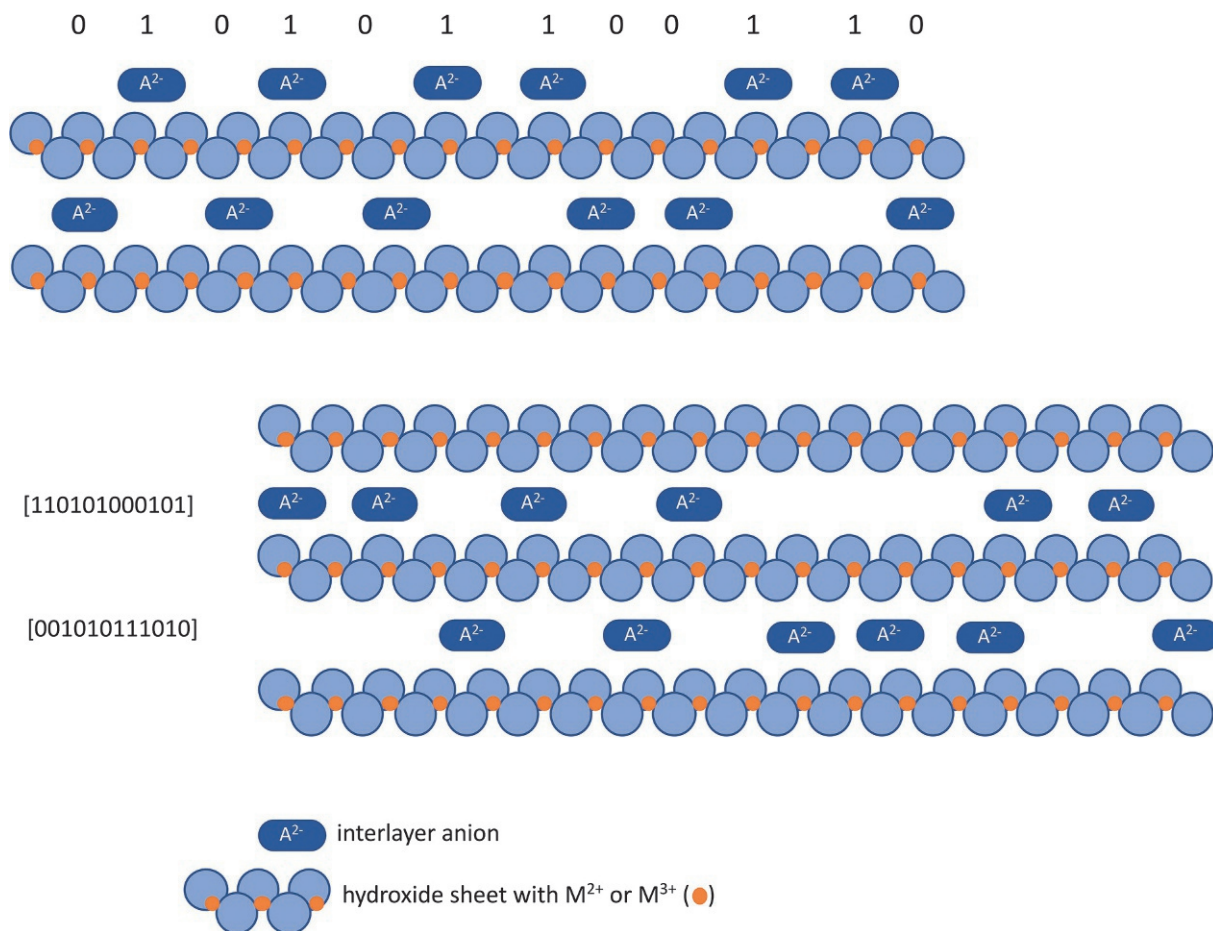


Fig. 14 Schematic representation of information storage in LDH interlayers for different arrangements of dianion: (a) with 2 hydroxide layers and (b) with 3 hydroxide layers. Filled LDH sheet sites can be denoted “1,” and non-charge balanced sites as “0.” Modified from Greenwell HC, Coveney PV (2006) Layered double hydroxide minerals as possible prebiotic information storage and transfer compounds. *Origins of Life and Evolution of Biospheres* 36: 13–37.

This complementarity can be considered a rudimentary form of “base pairing.” In a three-layer repeat cell (Fig. 15b), the arrangement of sites at the top face of the crystal will be the same as that of sites at the bottom face, so far as the information content is concerned. In the suggested mechanism (Fig. 15c) for information transfer between LDH crystals, the surface adsorbed anions are electrostatic sites on the LDH surface H-bond with other anions, setting up an inverse bonding surface (c1), or a complementary surface (c2). We can note that in these idealized systems where electrostatic interactions dominate, the local sites of occupancy in the lattice are not differentiated from those of the interlayer species (modified from Greenwell and Coveney, 2006). In three dimensions the opposing external surfaces of a three-layer polytype LDH are enantiomers, i.e., non-superimposable mirror images. Under such conditions, the adsorption of prochiral anionic monomers and their subsequent oligomerization might conceivably result in the formation of molecules that are enantiomers—molecules that are non-superimposable mirror images of each other (Greenwell and Coveney, 2006).

An area of interest in the study of LDHs is the biomimetic nature of LDH catalysts. Computer simulation shows that the enhanced reactivity of some organo-LDH catalysts may be due to the variable layer separation and amphiphilic nature of hosted bilayers. It results in increased selectivity and controlled organization of substrate reactant molecules that effectively provide an inorganic or organic-hybrid pseudo-enzyme active site.

In the particular case of LDH green rusts, such as fougérite, Branscomb and Russel (2019) suggest that pyrophosphate formation, as a form of energy storage, can occur. The mechanism would involve a progression of pyrophosphate accompanied by magnesium following an undulation of the sheet, of about 1 nm, facilitated by the conduction of electrons in the hydroxide sheet. In fougérite, unlike other LDHs, the bivalent and trivalent ions are of the same nature, and an electron can pass from an Fe^{2+} ion to the Fe^{3+} ion, i.e., by the propagation of polaron holes of about 1.2 nm.

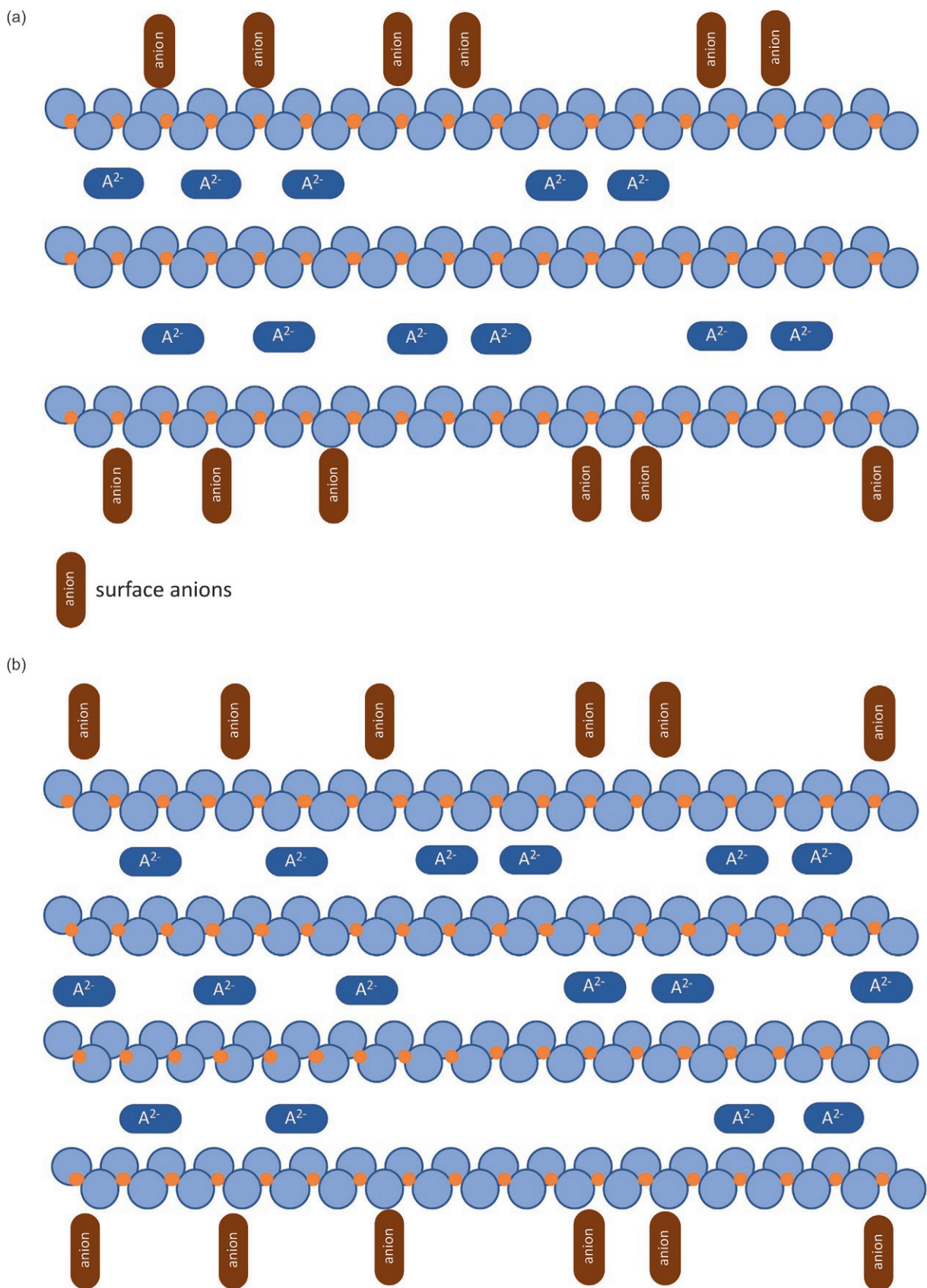


Fig. 15 Schematic information encoded at the external surfaces of LDH polytypes: (a) with a two-layer repeat structure and (b) three-layer repeat structure and (c) a suggested mechanism for information transfer between LDH crystals. Modified from Greenwell HC, Coveney PV (2006) Layered double hydroxide minerals as possible prebiotic information storage and transfer compounds. *Origins of Life and Evolution of Biospheres* 36: 13–37.

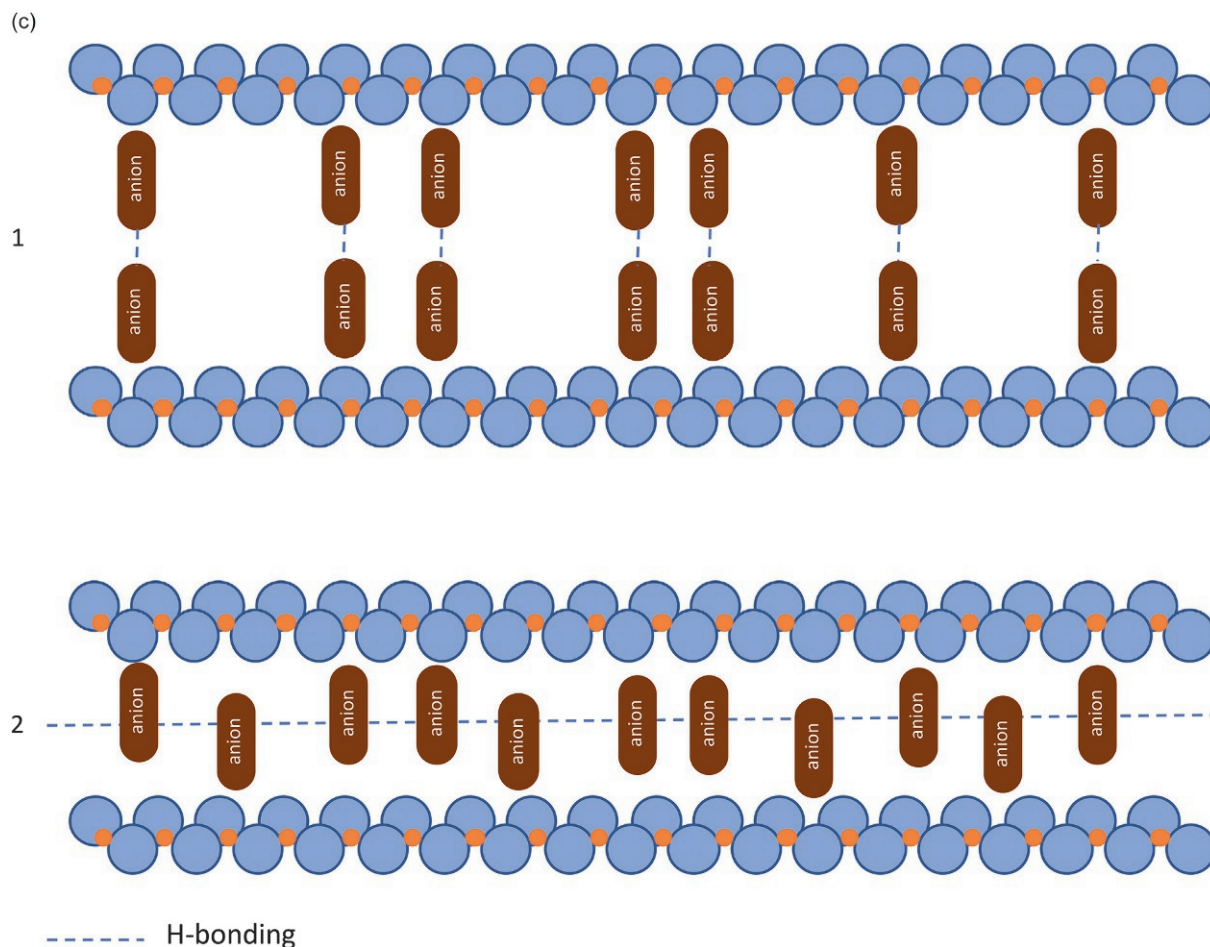


Fig. 15—Cont'd

Conclusion

Local heterogeneities in the soil create gradients in pH, redox and chemical composition that support chemical reactions. Redox measurement requires the identification of electroactive redox couples at the platinum electrode and speciation of chemical species in the soil solution... Redox reactions couple different biogeochemical cycles through both biotic (aerobic or anaerobic) and abiotic processes. Specifically, the nitrogen cycle involves competition between biotic reduction (denitrification) and direct abiotic reduction from nitrate to ammonium. In this process, fougérite plays a major role. In addition, LDHs and especially fougérite can accommodate different anions, which results in variation of the interlayer length. The reactants can move in a restricted space, compared to the 3-D space of the free solution. Thus LDHs, by their 2-D structure can support auto-organization of reactions that facilitate chemical reactions, including redox reactions. They are also good mineral candidates for the storage and replication of information over long-range order both in the plane (a , b) and along the c -axis, perpendicular to the sheets.

References

- Avery J (2003) *Information Theory and Evolution*. London: World Scientific Publishing Co. Pte Ltd. ISBN: 978-981-238-399-0. 232 pp.
- Bourrié G, Trolard F, Refait P, and Feder F (2004) A solid solution model for Fe(II) - Fe(III) - Mg(II) green rusts and fougérite and estimation of their Gibbs free energies of formation. *Clays and Clay Minerals* 52: 383–395.
- Branscomb E and Russel M (2019) On the beneficent thickness of water. *Interface Focus* 9(6): 20190061.
- Drits VA and Bookin AS (2006) Crystal structure and X-ray identification of layered double hydroxides. In: Rives V (ed.) *Layered Double Hydroxides: Present and Future*, pp. 39–92. New-York: Nova Science publishers. ISBN: 1-59033-060-9.
- Génin JMR and Ruby C (2004) Anion and cation distributions in Fe(II)-Fe(III) hydroxysalt green rusts from Mössbauer analysis (carbonate, chloride, sulphate,...); the "fougérite" mineral. *Solid State Sciences* 6: 705–718. <https://doi.org/10.1016/j.solidstatesciences.2004.03.021>.

- Greenwell HC and Coveney PV (2006) Layered double hydroxide minerals as possible prebiotic information storage and transfer compounds. *Origins of Life and Evolution of Biospheres* 36: 13–37.
- Hansen HCB, Gulberg S, Erbs M, and Koch CB (2001) Kinetics of nitrate reduction by green rusts: Effects of interlayer anion and Fe(II):Fe(III) ratio. *Applied Clay Science* 18: 81–91.
- Huang YH and Zhang TC (2004) Effects of low pH on nitrate reduction by iron powder. *Water Research* 38(11): 2631–2642.
- Nawaz MF, Bourrié G, Trolard F, Mouret J-C, and Henry P (2013) Effects of agronomic practices on the physico-chemical properties of soil waters in rice culture. *Turkish Journal of Agriculture and Forestry* 37: 195–202.
- Nye PH (1981) Changes of pH across the rhizosphere induced by roots. *Plant and Soil* 61(1): 7–26.
- Pfeifer S (2000) Chapter 3: Characterisation of redox state of aqueous systems: Towards a problem-oriented approach. In: Schüring JS, Schulz HD, Fischer WR, Böttner J, and Duijnisveld WHM (eds.) *Redox: Fundamentals, Processes and Applications*, pp. 24–41. Berlin, Heidelberg, New York: Springer-Verlag. ISBN: 3-540-66528-5.
- Sparks D (2003) *Environmental Soil Chemistry*, 2nd edn. USA: Academic Press, Elsevier Science. ISBN: 0-12-656446-9/352.
- Sposito G (2008) *The Chemistry of Soils*, 2nd edn. Oxford University Press. ISBN: 978-0-19-531369-7. 328 pp.
- Stumm W (1992) *Chemistry of the Solid Water Interface: Processes at the Mineral-Water and Particle-Water Interface in Natural Systems*. New York: Wiley. ISBN: 978-0-471-57672-3. 448 pp.
- Stumm W and Stumm-Zollinger E (1971) Chemiostasis and homeostasis in aquatic ecosystems; principles of water pollution control. In: Hem JD (ed.) *Non Equilibrium Systems in Natural Water Chemistry. Advances in Chemistry*, vol. 106, pp. 1–29. Washington: American Chemical Society.
- Trolard F and Bourrié G (1999) Influence des oxydes de fer de type « rouilles vertes » sur les séquences d'oxydo-réduction dans les sols. *Comptes-rendus de l'Académie des Sciences, Paris, Sciences de la terre et des planètes/Earth & Planetary Sciences* 329: 801–806.
- Trolard F, Duval S, Nitschke W, Ménez B, Pisapia C, Ben Nacib J, Andréani M, and Bourrié G (2022) Mineralogy, geochemistry and occurrences of fougérite in a modern hydrothermal system and its implications for the origin of life. *Earth-Science Reviews* 225: , 103910. <https://doi.org/10.1016/j.earscirev.2021.103910>.
- Williams RJP and Fraústo da Silva JJR (1997) *The Natural Selection of the Chemical Elements*. New York: Oxford University Press Inc. ISBN: 0-19-855842-2. 646 pp.

Sorption of organic chemicals in soil

Joseph J Pignatello, Department of Environmental Sciences, The Connecticut Agricultural Experiment Station, New Haven, CT, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of B. Xing, J.J. Pignatello, SORPTION I Organic Chemicals, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 537-548, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00555-5>.

Introduction	316
General considerations	316
Categories	316
Thermodynamics	316
Forces involved	317
Sorption of nonionic compounds	318
Sorption of nonionic compounds to mineral solids	318
Sorption of nonionic compounds to soil organic matter	318
General considerations	318
Nature of soil organic matter as a sorbent	320
Sorption to humic substances and kerogen	321
Sorption of nonionic compounds to pyrogenic carbonaceous matter (PCM)	323
Kinetics of sorption	324
Sorption of organic ions	324
Concluding remarks	326
References	326

Abstract

Sorption of organic chemicals to soil particles governs their mobility, reactivity, and bioavailability in soil. This chapter deals with physisorption of both nonionic and ionic species. The various types of sorption that may occur are discussed, together with the intermolecular forces involved in the context of the nature of different particles that occur in soil, including oxide minerals, clays, hard and soft organic matter, and pyrogenic organic materials. The chapter also focuses on sorption equilibrium processes and touches on sorption kinetics.

Key points

- Sorption is a key underlying process in soils regulating transport, bioavailability and reactivity of organic compounds.
- Sorption includes all steps from extraction of the molecular entity from the aqueous phase, rearrangement of the solid phase to accept the entity, and establishment of intermolecular forces between the entity and the solid surface or matrix.
- Depending on the sorbent and sorbing molecule, sorption can occur as solid-phase dissolution (partitioning), binding to discrete surface sites, condensation in pores, or dissolution in or on water films associated with particles.
- Important driving forces for sorption of nonionic compounds are the hydrophobic effect and combinations of the weak intermolecular forces, dispersion, induction, electrostatic, and hydrogen bonding, depending on solute molecular structure.
- In addition to the above, organic ions can sorb by anion or cation exchange, ion-pair sorption, formation of charge-assisted hydrogen bonds with weak acid groups, or ligation with metal ions underlying mineral surfaces.
- Free energy relationships between sorption and solute structure and sorbent parameters have been proposed but are still unsatisfactory for polar and ionic compounds and lack translatability among soils.
- Sorption and desorption can exhibit slow stages due to the presence of highly viscous organic matter absorptive phases, tortuous pore networks within individual particles or particle aggregates, and steric effects as intra-particle pore dimensions approach molecular dimensions.
- Sorption to organic matter is primarily by partitioning but the softness or hardness of the organic matter can strongly affect the magnitude and linearity of sorption.
- Sorption to pyrogenic carbonaceous matter occurs by adsorption and pore-filling.

Introduction

Organic chemicals enter soils through the intentional use of chemicals, such as pesticides or fire-fighting products; the application of biosolids to farmlands; and the disposal or accidental release of petrochemicals, industrial wastes, and consumer wastes. Some compounds have negative effects on the environment and public health. A key process in soils that regulates the transport, bioavailability and reactivity of organic compounds is sorption. 'Sorption' and the reverse, 'desorption,' refer to the process of molecular exchange between the gas or solution phase and an immobile phase (the sorbent). The potentially sorbing molecule is called the solute (sometimes the sorptive) and the sorbed component is referred to as the 'sorbate.'

Sorption may be categorized by the type of intermolecular attractive forces between sorbate and sorbent. Ordinary physical sorption ('physisorption') involves relatively weak intermolecular forces, while chemisorption involves strong bonding of a covalent nature. Physisorption is, by far, the most common type encountered in the context of soil and environmental sciences, and will be the exclusive focus here. Chemisorption occurs for only a few types of environmental contaminants—e.g., amines reacting with ketone, aldehyde or quinone functional groups in SOM (soil organic matter) giving the imine or Michael addition product—and is irreversible or quasi-reversible (Li et al., 2000). Chemisorption may also result following prior thermochemical, photochemical, or enzymatic conversion (Dec and Bollag, 2000) of a compound to a reactive intermediate that then couples covalently with SOM. Such coupling reactions are technically outside the definition of sorption, since the molecules undergoing coupling are transformed species, not the original ones. Physisorption is a necessary first step to chemisorption and any other transformations that occur on or within soil particles.

The reactive mineral components in soils are silicate clays and oxyhydroxide minerals of Si (silicon), Fe (iron), Al (aluminium), Mg (magnesium), Mn (manganese), etc. Soil organic matter comprises the totality of substances derived from biomass, mainly of plant and microbial origin, in various stages of decomposition and diagenesis. A fraction (typically small) of SOM may be fire-derived matter. Soil organic matter is the predominant sorbent of most neutral organic compounds except at very low SOM content or very low moisture content. Ionic species can interact strongly with both SOM and mineral components.

General considerations

Categories

Physisorption processes are often broadly classified as 'absorption' or 'adsorption.' Absorption (also called 'solid-phase dissolution' and 'partitioning') occurs when molecules penetrate the solid–fluid interface and are able to intermingle with the internal molecular or atomic matrix of the sorbent. Soil organic matter is the only natural component of soils that is penetrable in this way, but polluted soils may contain additional penetrable absorptive phases such as solvents, oils, tars, plastics, etc.

Adsorption is a term used in the phenomenological sense to imply association at a solid–fluid interface. Mechanistically, however, adsorption may encompass several qualitatively different processes, including binding of molecules independently to discrete interfacial sites, partitioning of molecules into an ordered hydration phase near the surface, condensation of the molecules into a liquid-like or solid-like state in small pores, and adsorption of molecules on the surfaces of water films associated with the particles. These processes will be discussed in more detail as this chapter progresses.

Thermodynamics

Sorption from water may be written as a series of steps (Eqs. 1a–1d) summing to the overall net reaction in Eq. (1e):



Step 1a represents the removal of solute x from the water to the gas phase, where $\{ \}$ symbolizes the solvation (hydration) shell around x , and $\cdots$ represents the cavity left behind by the vacated solute molecule. Step 1b represents collapse of the cavity and reconfiguration of solvation-shell water molecules to the bulk water state represented by $()$. Step 1c is reorganization of the sorbent matrix ($\blacksquare$) to some altered state ($\blacksquare'$) that may be necessary to accommodate the sorbate, such as displacement of water or other

sorbed molecules into solution and structural changes such as swelling and pore flexing. Little is known about matrix changes and their effects on sorption equilibrium. Finally, step 1d represents the formation of bonds between sorbate and the altered sorbent.

When sorption occurs from the vapor state steps 1a and 1b do not apply and net sorption is the sum of steps 1c and 1d:

$$x_g + \blacksquare = x-\blacksquare' \quad (1f)$$

The curve defining the equilibrium relationship at a given temperature between the sorbed-phase concentration S and fluid-phase concentration C is the 'sorption isotherm.' Various models have been used to describe the isotherm depending on the mechanism(s) postulated.

It is convenient to define the 'sorption distribution ratio,' K_d , as:

$$K_d = \frac{S}{C}, \quad (2)$$

where C is expressed in units of pressure or moles per unit volume. The units of S depend on convention and whether sorption occurs by adsorption or absorption. For adsorption they may be based on surface area or pore volume, and for absorption on solid volume. However, these properties are often not known accurately and so dry weight is commonly used as a surrogate. The K_d is concentration dependent unless the isotherm is linear.

The K_d is related to the Gibb's free energy of sorption, ΔG_{sorp}^0 by

$$\ln K_d + \text{const} = -\frac{\Delta G_{\text{sorp}}^0}{RT}, \quad (3)$$

where R is the ideal gas constant and T is temperature. A constant 'const' is needed to convert the practical units of K_d to thermodynamic units, since K_d must be dimensionless and expressed as a mole fraction relative to a reference state. While the reference state for the liquid phase is well defined, that for the solid phase is not; thus, the value of 'const' is generally unknown. The ΔG_{sorp}^0 , which is concentration dependent when K_d is concentration dependent, encompasses all enthalpy and entropy changes that occur during sorption. That is, the position of equilibrium depends on i) the difference in enthalpy between all sorbate-sorbent and all solute-water interactions, and ii) the difference in entropy associated with each of the steps 1a-1d. There are virtually no examples where the contributions of these factors have been fully delineated.

Forces involved

The fundamental forces between solute and surface are represented by Eq. (1d) Physisorption involves the so-called 'weak' intermolecular forces: coulombic, dispersion, induction, electrostatic, and hydrogen (H) bonding (Israelachvili, 1992). These forces act simultaneously in combinations permitted by the structures of the interacting species. Coulombic force is the attractive or repulsive force between ions of opposite or like charge, and it is proportional to the distance separating the ions. Dispersion is the mutual attraction of momentary dipoles produced by the synchronization of electronic motion in interacting species. Induction is the attraction of a permanent dipole or charge in one species with the dipole it induces in the other. Electrostatic (distinguished from coulombic) includes dipole-dipole and charge-dipole forces, as well as analogous multiple forces. The strengths of dispersion, induction, and electrostatic (often lumped as 'London-van der Waals forces') depend on the molecular area of contact, separation distance (to the inverse sixth power), orientation, and applicable molecular properties: charge, dipole moment, ionization potential and polarizability.

The H-bond can be written $(\text{DH} \cdots \text{A})$, where D and :A are proton donor and proton acceptor groups. Hydrogen bonds can vary tremendously in strength. Most ordinary H-bonds contain O (oxygen), N (nitrogen), or S (sulfur) as terminal atoms, range in strength from 2 to 10 kcal mol⁻¹, and are essentially of dipole-dipole character. Much weaker H-bonds are known, such as between a donor and the π system of an aromatic ring or the halogen atom of an electrophilic carbon-halogen bond. Hydrogen bonding between a contaminant and sites on soil particles must compete with H-bonding between water molecules and the same sites. In a water-rich environment it is questionable whether compounds forming ordinary H-bonds, and especially weaker H-bonds, can out-compete water molecules for H-bonding sites.

In general, the strength of an H bond is inversely related to the difference in the acid dissociation constants of DH and AH⁺, ΔpK_a . As the ΔpK_a approaches zero, the bond strengthens, the proton is more equally shared, and the bond takes on covalent character (Gilli and Gilli, 2009). A family of very strong H-bonds exists, but the category most relevant to aqueous systems in the environmentally relevant pH range (~3 to ~9) is the 'charge-assisted H-bond' (CAHB), which forms between weak acids or bases and where one or both terminal atoms bear a charge. The CAHB category includes the (-)CAHB where both donor and acceptor atoms are negative: $(-\text{O}^- \cdots \text{H}^+ \cdots \text{O}^-)$ or $(-\text{N}^- \cdots \text{H}^+ \cdots \text{O}^-)$, and the (+/-)CAHB, when one is negative and the other neutral: $(-\text{O}^- \cdots \text{H}^+ \cdots \text{N}-)$. The CAHB reaches strengths of as much as 35 kcal mol⁻¹. The H-bond between a carboxylic acid and its conjugate anion, $\text{RCO}_2^- \cdots \text{H}^+ \cdots \text{O}_2\text{CR}$, for which ΔpK_a is zero, is among the strongest of H-bonds known among organic compounds. The bond is akin to the CAHB in hydrogen difluoride, $\text{F}^- \cdots \text{H}^+ \cdots \text{F}$, the strongest of all H-bonds known. There are many examples of weak acid and weak base compounds of environmental concern with pK_a values in the environmental pH range, including carboxylic acids, phenols, amines, heterocyclic amines, sulfonamides, and others. Soil organic matter is rich in weak acid functional groups with intrinsic pK_a values close to those of weak organic acids or base contaminants and therefore can form a CAHB with them. The CAHB equilibrium is rapid and reversible. Several examples of CAHB formation in soil-relevant systems have been documented.

Another important driving force for sorption from aqueous solution is the hydrophobic effect, which is a solution-centered effect. The hydrophobic effect is due to entropy and enthalpy costs associated with the more ordered alignment of water molecules near the apolar regions of the dissolved solute, as well as the consequent disruption of water–water hydrogen bonds there. (Apolar molecules or regions of molecules are those not strongly polarized and lacking appreciable H-bonding capability). For a series of compounds of related structure (e.g., aromatic hydrocarbons), K_d is found to correlate with the molar refractive index, the *n*-octanol–water partition coefficient (K_{ow}), or the sub-cooled water solubility, often used as surrogates for a compound's net hydrophobic character. This is true whether the sorbent domain is the polar surface of a mineral, or the comparatively apolar surface or interior of organic matter.

Sorption of nonionic compounds

Sorption of nonionic compounds to mineral solids

Minerals present two kinds of environments for sorption: (i) the hydroxylated surface found on hydrous oxide minerals and the edges of aluminosilicate clays, and (ii) the interlayers of layer silicate clays, which contain siloxane surfaces of varying charge density.

Hydroxylated surfaces have –OH groups protruding into solution from the topmost layer of lattice metal ions. Sorption of apolar compounds from dilute solution to low-porosity hydrous oxide surfaces is proportional to mineral surface area and is not very sensitive to surface charge, solution pH, or solution ionic strength (although high ionic strength may reduce water solubility and increase the tendency for sorption). Sorption of apolar compounds is also proportional to the hydrophobicity indices mentioned above. While data are sparse, sorption tends to be endothermic and sorption isotherms tend to be linear to slightly nonlinear. Sorption to minerals has been covered well by [Schwarzenbach et al. \(2017\)](#).

There is good reason to believe that sorption of apolar compounds to hydroxylated surfaces occurs by partitioning (i.e., absorption) of the sorbate into the 'vicinal water layer'—the semi-ordered multilayer of water under the influence of the surface. This is depicted as 'B' in [Fig. 1A](#). However, apolar compounds are thought unable to penetrate the closest 1–2 layers of water molecules, and therefore do not contact the surface. This is inferred from the observation that sorption of apolar organic vapors is strongly suppressed by water vapor, and that, per unit surface area, the heat of adsorption of organic molecules is much smaller than that of water. Compounds capable of H-bonding to surface hydroxyl groups may be able to penetrate the vicinal water layer and compete with strongly bound water molecules for surface sites ('A' in [Fig. 1A](#)).

The siloxane surface exists on the faces of many layer silicate clays and consists of O atoms bridging underlying Si^{IV} atoms ([Fig. 1B](#)). The O atoms may have a permanent negative charge, depending on whether they lie near a point of isomorphic metal ion substitution in the underlying lattice. The charged sites are closely associated with metal cations, and the surface in the vicinity of the charge is strongly hydrophilic. Polar sorbates may hydrogen bond to metal ion-coordinated water molecules. If it has an appropriate functional group, a sorbate may coordinate directly with a metal ion. The neutral regions between surface charges are hydrophobic or only weakly hydrophilic; hence, apolar molecules or parts of molecules may compete favorably with water for the bare surface in those regions. Sorption in clay interlayers is affected by interlayer accessibility to the solute. The charge density and the exchangeable cation with its hydration sphere control the interlayer spacing and the size of the adsorption surface available for incoming molecules.

The sorbate may also condense on surfaces or in small pores, depending on the abundance of water and the partial pressure or concentration of the compound. The condensate is regarded as a liquid-like or disordered-crystalline state depending on its freezing point. Due to capillary forces condensation in pores progresses from the smallest to the largest of pores (or narrowest to widest part of an individual pore) as the partial pressure or concentration increases ([Fig. 1C](#)).

Water has a strong negative effect on vapor adsorption to soils. Sorption of organic vapors decreases by an order of magnitude or more as the soil moisture content increases to approximately the wilting point, which corresponds to an average of 3–5 molecular layers of water ([Poulsen et al., 2000](#)) ([Fig. 2](#)). This inhibition is due to competition by water molecules for pore volume space and polar surface sites on inorganic surfaces. At moisture contents above the wilting point sorption approaches what it would be in water-saturated soil.

When dealing with sorption of vapors to unsaturated soils it is often necessary to include the amount of dissolved in liquid water coating surfaces and in pores, apart from the vicinal water. That water is technically considered part of the 'immobile phase.' Sorption may occur by a combination of aqueous-phase dissolution ('C' in [Fig. 1A](#)) and vapor-phase adsorption at the gas-water interface (D in [Fig. 1A](#)) ([Costanza and Brusseau, 2000](#)). The latter is especially important for compounds like alkanes that are highly insoluble in water. Finally, sorption of organic chemicals to minerals can be strongly affected by the presence of initially dissolved organic matter on their surfaces.

Sorption of nonionic compounds to soil organic matter

General considerations

Soil organic matter is usually the preferred sorbent of neutral compounds in soils because it offers a relatively hydrophobic domain for hydrophobic molecules. Whether sorption to SOM dominates total sorption in the soil depends on its content and the relative sorbate affinity for it. A 'rule of thumb' is that sorption to SOM becomes dominant when the soil organic carbon (OC) concentration exceeds 0.1% dry weight.

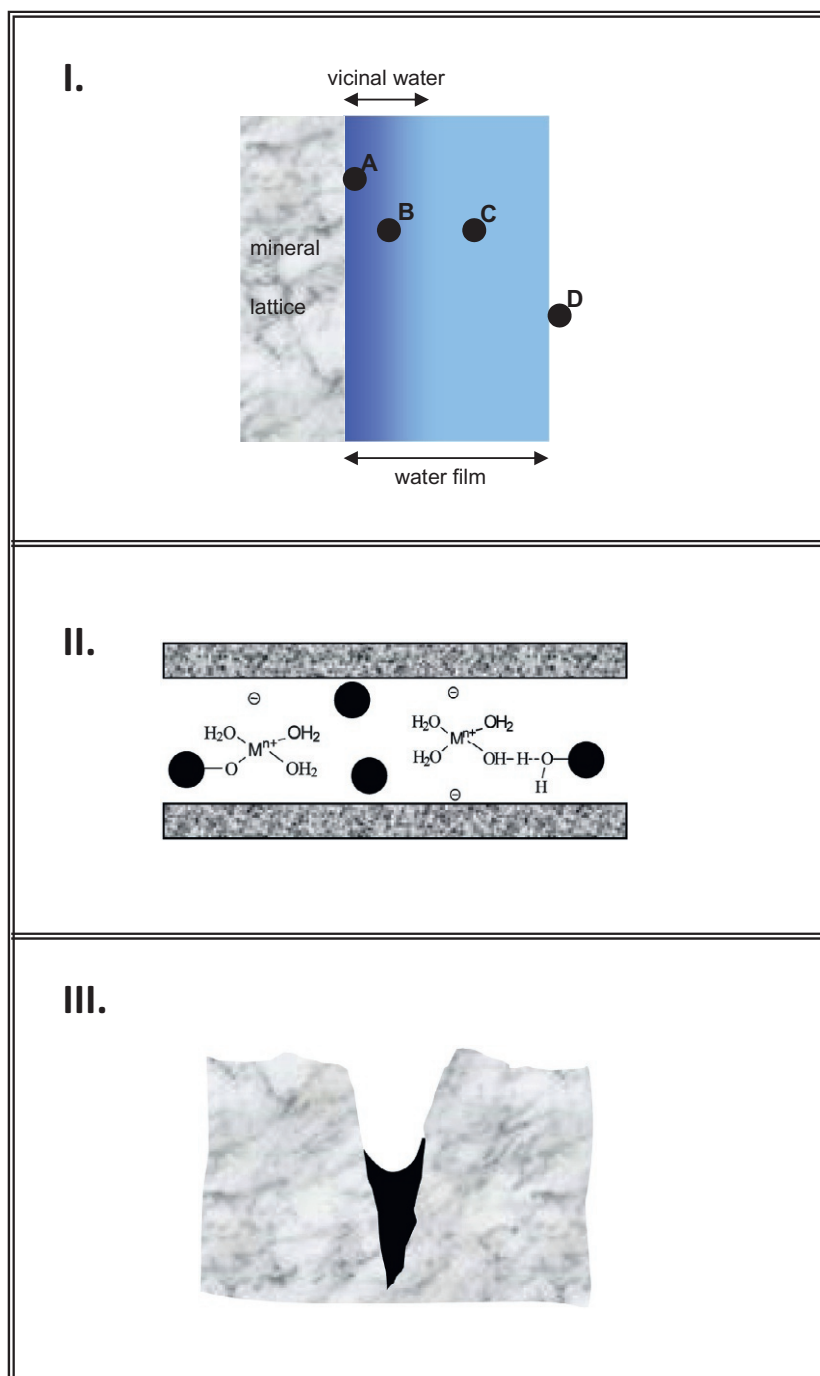


Fig. 1 Depiction of different types of sorption in mineral particles. Box I. Interaction with discrete sites on the hydroxylated surface (A); partitioning in the vicinal water layer on the hydroxylated surface (B); dissolution in water films (C); and adsorption on the surface of water films (D). Box II. Sorption in a clay interlayer: left to right, direct coordination to interlayer metal ion; association with the uncharged siloxane surface; and hydrogen bonding to a metal-coordinated water molecule. Box III. Condensation in a small pore.

Sorption isotherms in SOM and SOM-rich soils are often fitted by a power law known as the Freundlich equation,

$$S = K_F C^N \quad (4)$$

where K_F is an affinity-capacity coefficient and N is an exponent reflecting deviation from linearity of the relationship. The value of N is typically ≤ 1 , reflecting a constant ($N = 1$) to diminishing ($N < 1$) affinity of a solute for the solid with increasing loading, and may vary depending on the range of concentrations over which it is measured. The Freundlich model can be derived from the classical thermodynamic Langmuir model, which postulates competition between solute and water molecules for sites on a surface

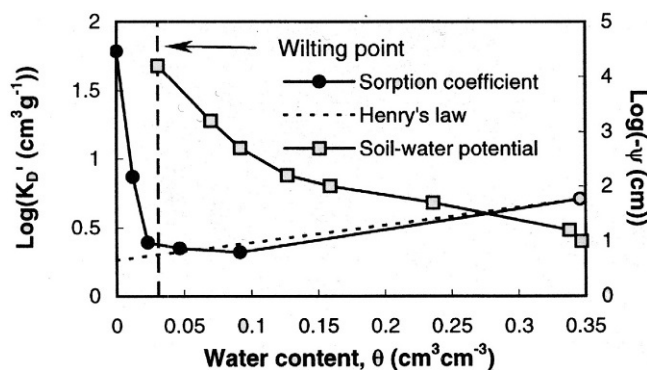


Fig. 2 Toluene sorption coefficient (K_D') and soil–water potential (ψ) as functions of volumetric water content in Lundgaard soil (1.1% organic carbon). The 'Henry's law' line is the calculated sorption assuming dissolution into water and aqueous–solid partitioning only (i.e., zero vapor–solid sorption). From Poulsen TG, Yamaguchi T, Moldrup P, de Jonge LW, and Rolston DE (2000) Predicting volatile organic vapor sorption from soil specific surface area and texture. *Journal of Environmental Quality* 29: 1642–1649.

of a specific energy, by assuming a population of sites of different energy exists that varies from zero to infinity. It should be noted that all sorption tends toward linearity below a certain concentration (the so-called 'Henry's law' region), a result that the Freundlich equation fails to predict.

When $N = 1$, the isotherm is given simply by

$$S = K_p C, \quad (5)$$

where K_p is equal to K_d and is referred to as the 'partition coefficient.' Many researchers have taken N to be 1 to simplify sorption terms in solute transport and bioavailability models.

On the assumption that sorption occurs exclusively to the SOM fraction, the distribution, Freundlich, or partition coefficient is often normalized to the fraction of OC (f_{oc}) in the soil; i.e.,

$$K_{oc} = K_d (\text{or } K_F \text{ or } K_p) f_{oc}^{-1} \quad (6)$$

Note that K_{oc} cannot be regarded as a universal constant since SOM is not uniform from place to place—in fact, K_{oc} may vary by an order of magnitude depending on its composition.

Linear free energy relationships between K_{oc} and parameters that represent hydrophobic properties, such as the octanol–water partition coefficient K_{ow} , water solubility, parachor (quantity related to surface tension), connectivity index, or UNIFAC functional-group activity coefficients, have been tested extensively. Such relationships work best for a set of closely-related compounds in a given soil, the reason being that the enthalpy and entropy of Eqs. (1a)–(1d) depend on sorbate molecular structure. Polyparameter linear free-energy relationships have also attracted interest (Endo and Goss, 2014). Such models use various molecular descriptors and system parameters to describe sorption, and fit experimental isotherms better than the single-parameter models for a given sorbent. However, none was designed to take into account nonlinearity, and system parameters found in complex, highly heterogeneous solids like soil and carbonaceous sorbents are difficult to define.

Solution pH and ionic strength have only modest (and inconsistent) effects on sorption of neutral compounds to SOM.

Nature of soil organic matter as a sorbent

Soil organic matter is a complex mixture of fresh to highly refractory components. The refractory components (often called 'humic substances') may also contain recognizable fragments of carbohydrates and proteins covalently attached to, or sequestered in, the solid phase. Humic substances are operationally defined as consisting of fulvic acid (soluble at all pH), humic acid (soluble in alkaline solution), and humin (insoluble at all pH and typically strongly associated with inorganic particles). The SOM fraction may also contain ancient materials such as kerogens and coaly particles. Kerogens are biogenic, macromolecular substances that have undergone low-temperature chemical and physical diagenesis over geologic time. 'Kerogens' are abundant in soft coals and shale and may be found in soils due to re-working processes. Hard coals (coaly type particles) are biogenic macromolecular substances that have been subjected to higher temperature and pressure (catagenesis) over geologic time and are found in some soils. Pyrogenic carbonaceous matter (PCM; sometimes called 'black carbon') is the carbonaceous byproduct of biomass burning (char) or fossil-fuel combustion (soot). The PCM in soil may comprise a trace to a small fraction of total organic C, but can be at larger concentrations in disturbed areas. The soil water also contains 'dissolved' organic matter (DOM). This ranges from small, truly-dissolved molecules to colloidal particles. The DOM entities are regarded as supramolecular aggregates of smaller molecules. Organic compounds can partition into these entities, although somewhat less effectively than into solid SOM.

Sorption to humic substances and kerogen

Humic substances and kerogens are considered to be random-network macromolecular solids; that is, they are composed of large molecules (a few hundred to many thousands of Daltons) of heterogeneous primary structure that aggregate randomly to form a three-dimensional phase. This phase is variously hydrated depending on the humidity. Charged groups such as carboxylate may coordinate with free metal ions to the top layer lattice metals of minerals. From fulvic acid to humic acid to humin to kerogen, there is a trend of increasing average molecular weight, apolar character, C content, mean fused aromatic ring size, density, crosslinking, and 'hardness.'

For humics and kerogens, absorption rather than adsorption is a better description of what occurs (Chiou, 2002). Surface areas are too small to support an adsorption interpretation. Organic molecules may penetrate the surface due to the solid's flexible nature and intermingle with the strands of macromolecules in the interior. A similar, 'solid-phase dissolution' paradigm for small molecules sorbed by synthetic organic polymers is well established. Sorption to humic substances appears to be slightly exothermic, although few data exist.

However, sorption to humics and kerogens is more complex than just simple partitioning (Huang et al., 1997; Pignatello, 2011; Xia and Pignatello, 2001). According to established polymer theory, amorphous macromolecular solids may exist in 'rubbery' (soft) or 'glassy' (hard) states, depending on temperature. These states differ in physical properties and sorbent behavior. The rubbery state is more expanded and the macromolecular segments have greater mobility. Sorption in rubbery solids is similar to dissolution in a viscous liquid in that specific sorption sites are fleeting and sorbed molecules experience an average chemical environment. This results in a linear isotherm like Eq. (5) and noncompetitive sorption behavior when multiple solutes are present. The glassy state is more dense and rigid. Because it is incompletely relaxed, the glassy state contains internal voids that are not filled by macromolecules and are poorly interconnected. These voids (or 'holes') are of molecular dimensions and may accommodate one or a few sorbate molecules. Hence, sorption in the glassy state exhibits dual-mode sorption behavior. One mode is 'dissolution' in the bulk solid which is linear and noncompetitive. The other mode is 'hole-filling' in the voids that is best modeled by a Langmuir-type equation. The dual-mode isotherm is given by (Xing and Pignatello, 1997):

$$S = K_D C + \frac{S_H^{\max} K_H C}{1 + K_H C}, \quad (7)$$

where K_D is the dissolution domain partition coefficient, and $S_H^{\max}$ and K_H are the hole domain capacity and affinity parameter, respectively. Eq. (7) predicts nonlinear and competitive sorption in multi-solute systems, both of which have been observed for many soils, humic acids extracted from soil then reconstituted in particulate form, and kerogens.

In the context of SOM, holes may be regarded as semi-permanent voids within the folds of individual macromolecules, between macromolecules, or between macromolecules and the mineral surface to which they are anchored, as illustrated in Fig. 3A. When Eq. (7) is applied to natural materials, (a) the K_D term represents the sum of partition coefficients for all rubbery domains present and the dissolution domains of all glassy substances present; and (b) the Langmuir term represents a weighted average of holes of different sorption potential. The affinity for a given solute is somewhat greater in holes than in the dissolution domain because in the latter a cavity has to be created to accommodate the sorbing molecule, which costs free energy.

As loading of the compound increases the isotherm may become more complex. The glassy state may become softened ('plasticized') by the sorbate, causing it to convert to the rubbery state and the holes to disappear. At high loadings the isotherm may begin to curve upwards relative to the solution concentration axis due to swelling (Xia and Pignatello, 2001). The glass-to-rubber transition temperature of some natural substances has been measured, including wet and dry humic acids (43–72 °C), coals (>300 °C), and wet lignin (60–90 °C) (Huang et al., 1997). Lignin is an amorphous, random macromolecular polymer from the woody parts of plants that is regarded as a major precursor of terrestrial humic substances. Whole soil SOM may exhibit multiple glass transitions due to its heterogeneous nature. Synthetic polymers, humic acids, and coals all have been shown to have a large internal surface area associated with the holes.

Fig. 4 shows sorption of the hydrophobic compound phenanthrene in SOM. Phenanthrene sorption is nonlinear ($N = 0.772$) and the dual-mode model fits well. Langmuir hole-filling type of sorption dominates overall sorption until the aqueous concentration reaches $0.14 \mu\text{g mL}^{-1}$, where partitioning then predominates. The overall dissolution coefficient is $4330 \mu\text{L g}^{-1}$ and Langmuir (hole) sorption capacity is $961 \mu\text{g g}^{-1}$. If the surface of this SOM sample (specific surface area, $0.88 \text{ m}^2 \text{ g}^{-1}$) were as rigid as a mineral, the maximal (external) surface coverage would be only $286 \mu\text{g g}^{-1}$ calculated from the molecular volume of phenanthrene as $182 \text{ cm}^3 \text{ mole}^{-1}$. Clearly, phenanthrene molecules have diffused into and sorbed within the SOM matrix, as opposed to being adsorbed on the surface.

Nonlinearity and competitive effects tend to increase in the expected order of increasing glassy character of the solid: dissolved humic acid < particulate humic acid < whole soil SOM < humin < kerogen. Artificially cross-linking the humic acid by flocculating it with polyvalent metal ions (e.g., Al^{3+}) decreases isotherm linearity and enhances competition by a second compound.

In general, diagenetic alteration of SOM increases aromaticity and promotes higher sorption, greater nonlinearity, and more intense competitive behavior than geologically younger samples. Many investigations report positive correlations between a sorption coefficient and the percent aromatic C of SOM, deviation from linearity, and intensity of competition. Recent research indicates that aliphatic components of SOM, particularly the paraffinic groups, can also significantly sorb apolar compounds and give rise to nonlinearity (Chefetz et al., 2000). Both amorphous and crystalline paraffinic regions have been identified by solid state NMR (nuclear magnetic resonance) spectroscopy (Xing and Chen, 1999). It has been suggested that the sorbate partitions into the flexible, amorphous paraffinic regions, but not in the crystalline regions. To account for the nonlinearity of paraffinic regions, it is

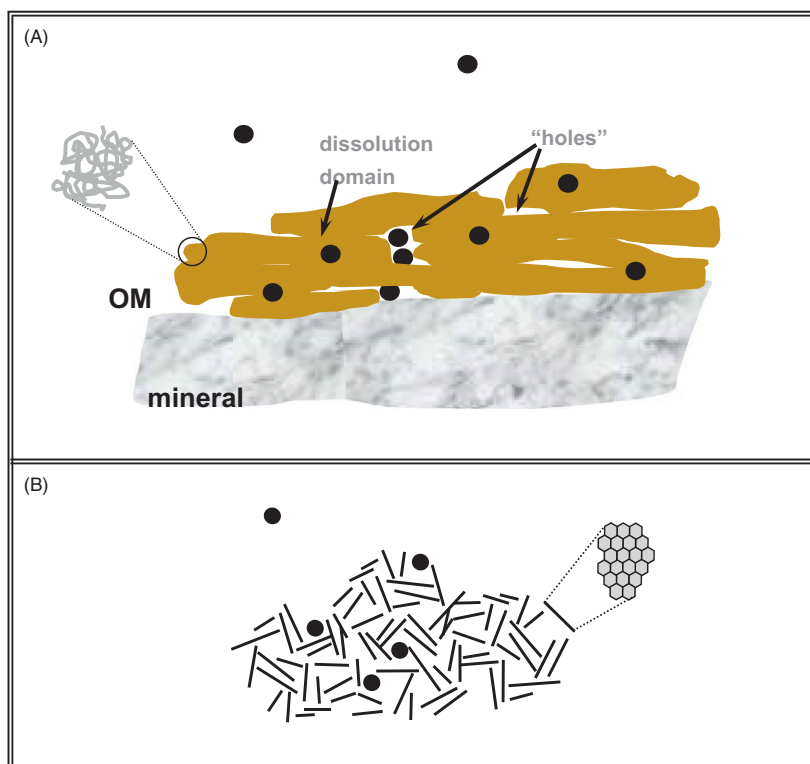


Fig. 3 Sorption in soil organic matter fractions. (A) Macromolecular solids such as humic substances and kerogens: sorbate may undergo solid-phase dissolution in domains composed of macromolecules in a relatively fluid state and hole-filling in voids within the organic phase and between the organic phase and the mineral surface. (B) Micrographitic substances such as PCM: sorbate may adsorb in shallow micropores and some may penetrate into deep micropores between polyaromatic platelets.

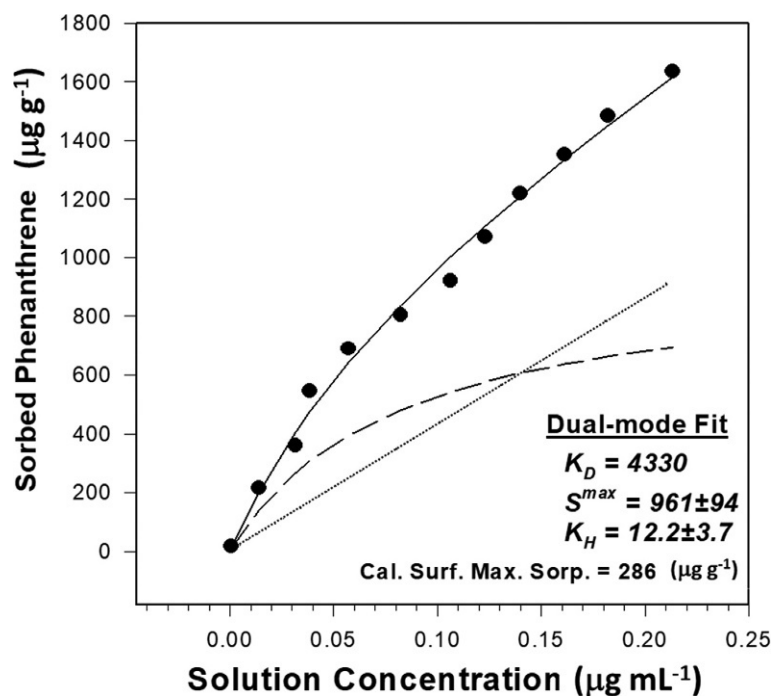


Fig. 4 Phenanthrene sorption by a soil with a large organic content. Solid line is the dual-mode fit assuming only one type of Langmuir (hole) site and solid circles are actual data points; dotted and dashed lines are contributions from partitioning and Langmuir (hole) type sorption, respectively. The maximum surface sorption capacity calculated from the surface area and molecular size is $286 \mu\text{g g}^{-1}$. The Freundlich model parameters (Eq. 4) are $K_F = 5610 (\mu\text{g g}^{-1}) (\mu\text{g mL}^{-1})^{-N}$ and $N = 0.772$ (curve not shown).

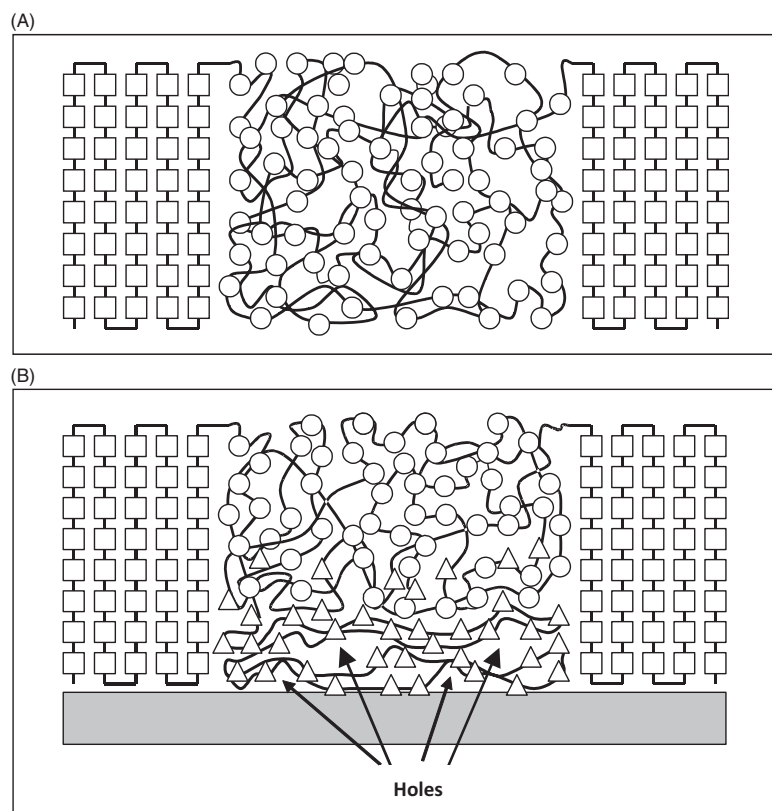


Fig. 5 Schematic configuration showing the distribution of crystalline ($\square$ -) and amorphous ($\circ$ -) aliphatic organic carbons (A). Interaction of the aliphatic carbons with a clay mineral surface ($\blacksquare$) rearranges part of the amorphous domain into a condensed configuration ($\triangle$ -) having nm-sized holes (B). These holes are the locus of adsorption sites responsible for nonlinear and competitive sorption. From Gunasekara AS and Xing B (2003) Sorption and desorption of naphthalene by soil organic matter: Importance of aromatic and aliphatic components. *Journal of Environmental Quality* 32: 240–246.

proposed that amorphous layers located at mineral interfaces become less flexible (harder), providing holes that act as adsorption sites (Gunasekara and Xing, 2003) (Fig. 5). This proposal is consistent with reports that nonlinearity and competitive sorption observed for humic acids adsorbed on minerals are greater compared with non-sorbed humic acids.

Sorption of nonionic compounds to pyrogenic carbonaceous matter (PCM)

With increasing temperature and heating time, the anoxic pyrolysis of lignocellulosic biomass evolves through several distinct physical–chemical states of increasing polyaromatic character, culminating in a turbostratic state of disordered graphitic microcrystallites. The polyaromatic sheets may be functionalized along the edges with hydroxyl, carboxyl, keto or other groups. Aliphatic C is converted to aromatic C at $\sim 450^\circ\text{C}$. The structure is very rigid and highly porous. The specific surface area depends on the heating conditions and can reach $600\text{ m}^2\text{ g}^{-1}$ or higher. Most of the surface area lies in mesopores (2–50 nm) and micropores ($<2\text{ nm}$ effective diameter), which are considered to be slit-shaped. Sorption occurs mainly by adsorption and pore-filling mechanisms (Figs. 1–III and 3). Pore filling at large concentrations can cause measurable swelling because small pores are expandable due to the local pressure of the sorbate.

Sorption isotherms of neutral compounds on PCM tend not to fit well to either the Freundlich or Langmuir models. For example, the isotherm of benzene in a wood charcoal is a reverse S shape, with a linear region from 10^{-7} to 10^{-4} times its water solubility (Braidia et al., 2003). Sorption of neutral compounds from water to graphite or activated carbon is exothermic. Sorption to PCM depends on surface area and pore-size distribution. Compared to non-pyrogenic forms of SOM, PCM has a greater affinity for organic compounds due primarily to their higher surface area and porosity. Another important property contributing to greater sorbent affinity is their aromatic character. The aromatic interface can be more closely approached than the alkane surface for steric reasons, permitting greater dispersion with sorbing molecules. (This is one reason why solvent–water partition coefficients of nonionic compounds are greater for aromatic solvents like benzene than aliphatic solvents like n-octanol and n-heptane.) It has been found that adsorption energy to wood chars correlates with fused ring size up to a certain size (~ 55 carbons) because polarizability increases with ring size up to that point (Yang et al., 2021). Chemical and biological oxidation processes such as those that occur during weathering of PCM in soil introduce O functionality into PCM surfaces that attract water molecules. This is disadvantageous to sorption because the bound water molecules compete for available surface sites. Lastly, it must be emphasized

that sorption of chemicals to PCM is strongly affected by the presence of dissolved organic matter. Such substances can bind to the surfaces of PCM where they can influence sorption by competing for sorption sites and blocking access to pore networks.

Proposed models for soils or sediments that contain both the ordinary softer matter (humics plus kerogen) and PCM take the form:

$$S = (1 - r_{\text{PC}}) \cdot K_{\text{p}} C + r_{\text{PC}} \cdot f_{\text{PC}}(C), \quad (8)$$

where r_{PC} is the ratio of pyrogenic carbon (PC) relative to the total organic carbon, and $f_{\text{PC}}(C)$ is a nonlinear function of dissolved concentration C describing sorption to PC. Predicting sorption with such models is problematic. Since PCM is a broad continuum of materials it is unclear how best to quantify r_{PC} (chemothermal or wet oxidation method?), which $f_{\text{PC}}(C)$ should be used, and what reference material is appropriate. As already mentioned, K_{p} is not strictly concentration-independent for a given soil and is not a universal constant. Dual sorbent models are able to predict pore water concentrations of highly hydrophobic compounds only to within about 2 orders of magnitude (Arp et al., 2009).

Kinetics of sorption

The rates of uptake and release of chemicals from soil particles play a role in controlling their mobility and bioavailability in soil (Pignatello, 2000; Pignatello, 2011; Alexander, 1995). In well-mixed soil–water systems it been shown experimentally that sorbate uptake is biphasic, with most occurring within a few hours, whereas days, weeks, or even months may be required to reach full equilibrium, depending on the compound and soil. Desorption that is induced by dilution of the liquid phase or reducing the pressure is almost always considerably slower than uptake for reasons that are unclear. Moreover, exhaustive desorption to water or air may reveal a minor fraction that seems permanently bound, yet is recoverable by solvent extraction or other vigorous means. Fractions that become highly resistant to desorption have been generated in purely inorganic as well as purely organic reference materials.

In general, soil particles have structural features that can hinder molecular diffusion of molecules to and from available sites. These features include highly viscous absorptive phases, tortuous pore networks within individual particles and the interstices of particle aggregates, and narrow pores. Models based on pore diffusion and intra-organic matter diffusion as the rate-limiting have been presented in the literature. Diffusion rates in soils are frequently independent of nominal particle radius, implying that the controlling diffusion length scale is much smaller—perhaps as small as 10–100 nm. In general, intraparticle sorption or desorption rate coefficients decrease with increasing molecular size due to steric hindrance of diffusion.

Diffusion depends strongly on the nature and geometry of the diffusing medium, the boundary conditions, and the local concentration (more correctly, chemical potential) gradient across boundaries. Classical mathematical diffusion laws are difficult to apply to soils due to their heterogeneous nature both compositionally and geometrically. Evidence suggests that diffusivity follows the order: humic acid > humin > kerogen or coal. Diffusion coefficients of toluene, *n*-hexane, and acetone in pressed humic acid disks are in the range 10^{-8} – 10^{-9} cm² s⁻¹, about the same as in rubbery polymers and are 3–4 orders of magnitude smaller than those in water. Diffusion within PCM has received little attention. From studies on homogeneous reference materials (Rouquerol et al., 1999) it is known that diffusion in fixed-pore networks is retarded by sorption to pore walls and tortuosity (deviation from straight-line paths and degree of pore connectivity). Diffusion in micropores, where pore openings are not much larger than the diffusant diameter, is also slowed by steric effects.

To avoid difficulties inherent in the application of diffusion models to soils, many people have employed non-mechanistic rate laws based on simpler mathematics. Recently published ones include the two-compartment, multi-compartment, and stochastic models. Although meeting with some success, they lack translatability to new systems. The two-compartment model, popular in soil column and bioavailability experiments, partitions the sorbate into instantaneous and kinetic compartments. The kinetic compartment has first-order terms for the forward and reverse steps. The multi-compartment model is popular in batch desorption experiments in which molecules are instantly removed as they desorb either by a stream of gas or by including a large excess of a polymeric sorbent in the suspension. The multi-compartment model assumes that desorption occurs concurrently from ‘fast,’ ‘slow,’ and sometimes ‘very slow’ compartments, each described by a first-order term. Stochastic models assume that the rate can be described with an array of equilibrium sorption coefficients and or first-order rate constants that are distributed continuously according to a probability density function. Slow kinetics play an important role in sorption isotherm hysteresis, in which the desorption branch does not coincide with the sorption branch. Hysteresis can be due to insufficient time allowed for equilibrium or irreversible alterations of the sorbent structure caused by uptake of the solute (Pignatello, 2011). Further discussion of sorption kinetics in the context of hysteresis will be covered in another chapter.

Sorption of organic ions

Ionic and ionizable organic compounds (IOC) constitute an important category of compounds of environmental concern found in soil and sediments. It includes many pesticides, pharmaceuticals, personal care compounds, surfactants, endocrine disrupting compounds, and per- and poly-fluoroalkyl (PFAS) substances. Examples of permanently charged functional groups or groups that are always charged under normal pH conditions in soil–water systems (pH ~3 to ~9) include R_4N^+ , sulfur heterocyclic ions, RSO_3^- , R_fCO_2^- , where R is any organic group and R_f is a perfluoroalkyl group. Examples of functional groups with pH-dependent charge

include primary, secondary and tertiary amines, heterocyclic amines, carboxylic acids, arylsulfonamides, and phenols. The sorption of ionizable compounds can vary tremendously with pH as the measured pH transitions through the pK_a value.

In the case of weak acids the neutral form of the molecule usually sorbs more intensely than the anionic form due to its greater hydrophobicity. In the case of weak bases, the contrast is not so clear because, although the cationic form is less hydrophobic than the neutral form, it can associate strongly with anionic sites on soil particles.

Organic ions undergo the same weak physisorption forces with soil particles as do nonionic molecules, plus other driving forces owing to their charge. One is ion exchange sorption, in which the organic ion replaces the 'natural' anion (e.g., $X^- = Cl^-$) or cation (e.g., $M^+ = Na^+$) initially associated with the solid (■) (Eqs. 9a and 9b).



A second driving force is ion-pair sorption, in which the organic ion pairs up with a counterion from solution before associating with the solid (Eqs. 10a and 10b).



Charge neutrality must be preserved in the transfers of reactions 9 and 10. Similar equations, properly balanced, can be written for polyvalent organic ions and counterions (e.g., Mg^{2+}). Eqs (9a) and (9b) are driven to the right by physisorptive interactions of the part of the organic ion not bearing the charge, and to the left by large concentrations of counterions. Cation exchange is more important than anion exchange in most cases due to the prevalence of negative charges on soil particles at near-neutral pH.

Ion pairs can partition into OM phases and in small pores of PCM and minerals. In contrast to ion exchange, large counterion concentrations control the equilibrium of reaction 10 to the right. Sorption is often found to exceed the measured cation or anion exchange capacity of the soil, which are based on sorption of inorganic ions (usually, NH_4^+ (ammonium) or Br^- (bromide)); ion-pair sorption is a likely explanation. Zwitterions, which contain sites of opposing charge within the same molecule, have not been studied thoroughly.

Carboxylate and some other weak acid anions sorb to PCM more strongly than expected considering that pH at the point of zero charge of these materials is typically above pH 4—i.e., their surfaces have an excess of negative charge. One explanation is the formation of a charge-assisted hydrogen bond (–)CAHB, as discussed above. The reaction between a weak acid anion D^- and a weak acid group on the sorbent, such as the carboxylate group, can be written:



This reaction has been conclusively demonstrated for reactions between simple weak acids and PCM, such as biochar and carbon nanotubes (e.g., Teixeira et al., 2011; Li et al., 2013). Reaction (11) is most favorable at a pH near the mean pK_a of DH and PCM anionic group. Upon sorption of D^- species such as carboxylate, sulfonaminate ($R-SO_2NH^-$), and other organic anions, the pH of the solution has been observed to increase, which is consistent with hydroxide ion generation in reaction 13. Moreover, competitive displacement of bound D^- has been shown to be much more effective when the competitor contains a (–)CAHB-capable functional group, compared to one that does not. Potential (–)CAHB reactions with non-pyrogenic SOM are likely but have not yet been demonstrated conclusively.

A third driving force for ion sorption, which is open only to organic anions, is ligand exchange with metals. Ligand–metal bonds may form between carboxylate, phenolate, thiolate, sulfonate, or phosphate groups on the sorbate, and metal ions—especially Fe^{3+} and Al^{3+} —on the surface of minerals or within SOM, usually in exchange for hydroxide (or water) ligands, e.g.:



where $\equiv$ represents the surface, M is a metal ion, and L^- is the sorbing ligand. Reaction (12) is more favorable when L^- is bidentate—for example, salicylate, oxalate, catecholate or *o*-phthalate and their derivatives. In that case the ligand can coordinate through both functional groups to a single metal ion or to adjacent metal ions. It is also possible for L^- to coordinate to a metal ion attached to SOM through an appropriate functional group on the humic structure. In that case, L is considered linked to SOM through a cation bridge. The energetics and kinetics of surface ligand exchange reactions are not well characterized. The intrinsic equilibrium constant seems to depend on structure in the same way as the corresponding reaction in solution between the ligand and free metal ion.

A fourth category of ion sorption occurs at the unhindered interface between water and oxide minerals (Korotky, 2000). The surface can be described as the flat plane of a mixture of positive $\equiv S-OH_2^+$ or negative $\equiv S-O^-$ sites depending on pH and ionic strength. The surface creates an electric double layer, where the charge at the surface is balanced by a 'diffuse layer' of counterions in solution near the charged mineral surface. Depending on the model used, the electric potential decreases linearly or exponentially with distance from the surface. These models have been applied predominantly to inorganic electrolytes and have been poorly tested with organic ions.

Concluding remarks

Sorption is a critical underlying process governing all types of behavior of chemicals in soil. With rare exceptions organic molecules or ions in soil spend the vast majority of their time in the sorbed state compared to the gaseous or dissolved states. Mobility from one soil parcel to the next is possible only if sorptive bonds are broken in the first stage and made in the second. Soil dwelling organisms—apparently, even single-cell organisms—lack access to molecules in the sorbed state. Thus, contaminant mobility and bioavailability of chemicals in soil depend not only on the position of the sorption equilibrium but also on the rates of sorption and desorption. Sorption is a highly complex process because it involves a variety of different types of chemicals, multiple fundamental forces acting in combination, multiple sorbent materials whose contributions are not necessarily additive, and constantly changing conditions in the soil aqueous phase. Soils are not uniform around the world. Although tens of thousands of studies have been devoted to sorption in soils, no model has yet been put forward that accurately predicts sorption in a given soil under a given set of conditions. This is especially true for ionic and polar molecules. Future research should be devoted to developing such models that require a minimum of measurement inputs.

References

- Alexander M (1995) How toxic are toxic chemicals in soil? *Environ. Sci. Technol.* 29: 2713–2717.
- Arp HP, Breedveld GD, and Cornelissen G (2009) Estimating the in situ sediment-porewater distribution of PAHs and chlorinated aromatic hydrocarbons in anthropogenic impacted sediments. *Environ. Sci. Technol.* 43(15): 5576–5585.
- Braida W, Pignatello J, Lu Y, Ravikovitch P, Neimark A, and Xing B (2003) Sorption hysteresis of benzene in charcoal particles. *Environ. Sci. Technol.* 37: 409–417.
- Chefetz B, Deshmukh A, Guthrie EA, and Hatcher PG (2000) Pyrene sorption by natural organic matter. *Environ. Sci. Technol.* 34: 2925–2930.
- Chiou CT (2002) *Partition and Adsorption of Organic Contaminants in Environmental Systems*. Hoboken, NJ: John Wiley & Sons.
- Costanza MS and Brusseau ML (2000) Contaminant vapor adsorption at the gas-water interface in soils. *Environ. Sci. Technol.* 34: 1–11.
- Dec J and Bollag J-M (2000) Phenoloxidase-mediated interactions of phenols and anilines with humic materials. *J. Environ. Qual.* 29: 665–676.
- Endo S and Goss K-U (2014) Applications of polyparameter linear free energy relationships in environmental chemistry. *Environ. Sci. Technol.* 48(21): 12477–12491.
- Gilli G and Gilli P (2009) *The Nature of the Hydrogen Bond: Outline of a Comprehensive Hydrogen Bond Theory*. Oxford, UK: Oxford University Press.
- Gunasekara AS and Xing B (2003) Sorption and Desorption of naphthalene by soil organic matter: Importance of aromatic and aliphatic components. *J. Environ. Qual.* 32: 240–246.
- Huang WL, Young TM, Schlautman MA, Yu H, and Weber WJ Jr. (1997) A distributed reactivity model for sorption by soils and sediments. 9. General isotherm nonlinearity and applicability of the dual reactive domain model. *Environ. Sci. Technol.* 31: 1703–1710.
- Israelachvili JN (1992) *Intermolecular and Surface Forces*, 2nd edn. London, UK: Academic Press.
- Koretsky C (2000) The significance of surface complexation reactions in hydrologic systems: A geochemist's perspective. *Journal of Hydrology* 230(3): 127–171.
- Li H, Lee LS, Jafvert CT, and Gravelle JG (2000) Effect of substitution on irreversible binding and transformation of aromatic amines with soils in aqueous systems. *Environ. Sci. Technol.* 34: 3674–3680.
- Li X, Pignatello JJ, Wang Y, and Xing B (2013) New insight into the mechanism of adsorption of ionizable compounds on carbon nanotubes. *Environ. Sci. Technol.* 47: 8334–8341.
- Pignatello JJ (2000) The measurement and interpretation of sorption and desorption rates for organic compounds in soil media. *Adv. Agronomy*. 69: 1–73.
- Pignatello JJ (2011) *Interactions of Anthropogenic Organic Chemicals with Natural Organic Matter and Black Carbon in Environmental Particles*. Hoboken, N.J: Wiley. pp. 3–50.
- Poulsen TG, Yamaguchi T, Moldrup P, de Jonge LW, and Rolston DE (2000) Predicting volatile organic vapor sorption from soil specific surface area and texture. *J. Environ. Qual.* 29: 1642–1649.
- Rouquerol F, Rouquerol J, and Sing K (1999) *Adsorption by Powders and Porous Solids*. San Diego, CA: Academic Press.
- Schwarzenbach RP, Gschwend PM, and Imboden DM (2017) *Environmental Organic Chemistry*, 3rd edn. Hoboken, NJ: Wiley-Interscience. Chapters 11 and 14.
- Teixido M, Pignatello JJ, Beltran JL, Grenados M, and Peccia J (2011) Speciation of the ionizable antibiotic sulfamethazine on black carbon (Biochar). *Environmental Science & Technology* 45: 10020–10027.
- Xia G and Pignatello JJ (2001) Detailed sorption isotherms of polar and apolar compounds in a high-organic Soil. *Environ. Sci. Technol.* 35: 84–94.
- Xing B and Chen Z (1999) Spectroscopic evidence for condensed domains in soil organic matter. *Soil Sci.* 164: 40–47.
- Xing B and Pignatello JJ (1997) Dual-mode sorption of low-polarity compounds in glassy polyvinylchloride and soil organic matter. *Environ. Sci. Technol.* 31: 792–799.
- Yang J, Pignatello JJ, Yang K, Wu W, Lu G, Zhang L, Yang C, and Dang Z (2021) Adsorption of organic compounds by biomass chars: Direct role of aromatic condensation (ring cluster size) revealed by experimental and theoretical studies. *Environ. Sci. Technol.* 55(3): 1594–1603.

Sorption—Oxanyons

CP Schulthess^a, H Wijnja^b, W Yang^c, Anahita Khosravi^d, and Baoshan Xing^d, ^aUniversity of Connecticut, Storrs, CT, United States; ^bMassachusetts Department of Agricultural Resources, Southborough, MA, United States; ^cCalifornia Air Resources Board, Sacramento, CA, United State; ^dUniversity of Massachusetts, Amherst, Massachusetts, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of C.P. Schulthess, H. Wijnja, W. Yang, SORPTION I Oxanyons, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 548–555, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00556-7>.

Introduction	327
Techniques for detection and quantification of oxanyons in solution	328
Techniques for investigation of oxanyons adsorbed at mineral–water interfaces	328
Vibrational spectroscopies	328
X-Ray absorption spectroscopy	330
Sorption strength of oxanyons	330
Effect of ions on adsorption of oxanyons	332
Sorption mechanisms of oxanyons	333
Site preference and site reactivity	334
Conclusions	335
References	335

Abstract

Oxanyons are negatively charged compounds composed of a central cation and oxygen ions. These species are ubiquitous in nature and are essential nutrients for plants but can also be harmful to organisms and the environment when their concentrations exceed critical levels. The sorption of oxanyons to solid soil components plays a crucial role in their fates in the environment, as it influences their mobility and availability. If oxanyons are retained weakly by soil particles, they can easily be desorbed and transported by the soil solution. Techniques such as ion-selective electrodes, colorimetry, and inductively coupled plasma-atomic emission spectrometry are essential for predicting their behavior in soil, which is critical for assessing environmental quality and developing effective management strategies.

Key points

- Oxanyons are prevalent in nature.
- Oxanyons can be toxic to organisms when their concentrations exceed critical levels.
- Sorption of oxanyons to solid soil components plays a critical role in their fate in the environment.
- The use of spectroscopic techniques is essential for understanding and predicting the behavior of oxanyons in soil.

Introduction

An oxanyon is a net negatively charged compound consisting of a cation at its center surrounded by oxygen ions. There are many different oxanyons present in soils. The most common ones of interest to soil and environmental scientists include phosphates (PO_4^{3-}), arsenite (AsO_3^{3-}), arsenate (AsO_4^{3-}), selenite (SeO_3^{2-}), selenate (SeO_4^{2-}), chromate (CrO_4^{2-}), molybdate (MoO_4^{2-}), carbonate (CO_3^{2-}), sulfate (SO_4^{2-}), nitrate (NO_3^-) and perchlorate (ClO_4^-). Oxanyons are prevalent in nature, and most of these are essential nutrients for plants. Almost all of them, however, are also toxic to plants and the environment when their concentrations rise above a certain critical level. For example, excessive amounts of some oxanyons, namely nitrates and phosphates, are directly responsible for the eutrophication of our ponds and lakes (Priya et al., 2022).

There are numerous other oxanyons that are discussed in the environmental literature, particularly in connection to industrial and nuclear waste disposal, or spills and remediation. Industrial oxanyon wastes include many of those already mentioned above. Radioactive waste and environmental studies may include oxanyons such as pertechnetate (TcO_4^-). No stable isotopes of Tc exist (which are prepared in very small amounts in cyclotrons) and because this radionuclide is rather dangerous to work with, a surrogate oxanyon is typically studied instead, namely perhenate (ReO_4^-).

The fate of oxanyons in soils in the environment depends strongly on their sorption to soil mineral components (Xinxin et al., 2021). That is, an oxanyon that is adsorbed to an immobile solid phase will not be transported to a new location, because movement of oxanyons in the environment is through the liquid phase. Retention by solids results in a longer availability of oxanyons in a defined area. For example, leaching from the soil's upper horizons down toward groundwater does not occur as long

as it is sorbed to the immobile solid phase. Lateral liquid movement to nearby streams is also minimized, as is its subsequent transport to larger bodies of water.

This assumes that the solid phase to which the oxyanion is sorbed is itself immobile. Keep in mind, however, that some solids (such as colloidal-sized particles) are very small and can be translocated by the moving water. This translocation of mobile solids together with the ions sorbed on to them can be vertically through soil horizons, horizontally with the groundwater flow, or as soil erosion moving horizontally over the soil surface (Hussain et al., 2021).

An understanding of the sorption of oxyanions to solid soil components is important for an evaluation of the quality of the environment. If the oxyanions are retained only weakly, then the moving water will easily remove them from the soil. If the retention of oxyanions by the immobile solid phase is moderate in strength, then the moving water will remove the oxyanions from the area slowly. Conversely, if the oxyanions are retained strongly by immobile solids, then the moving water will remove them from the area very slowly. If the solid phases are small enough to be mobilized by moving water, then attention to the turbidity of the liquid phase or the presence of surface soil erosion must be included in any evaluation of environmental quality. This is because the mobile solids typically will have sorbed contaminants on them.

Techniques for detection and quantification of oxyanions in solution

Quantitative studies of the adsorption of oxyanions on soils and minerals require determination of the oxyanion concentration remaining in the aqueous phase (Liang et al., 2021). That is, the amount adsorbed equals the amount of anion added to the solid-liquid suspension minus the amount remaining in the liquid phase at equilibria. Many different instruments can be used for the liquid-phase analysis, such as ion-selective electrodes, colorimetric methods, ion chromatography, atomic absorption, and inductively coupled plasma-atomic emission spectrometry. With some of these methods (namely, the first three mentioned), one must be particularly cautious to avoid problems attributed to speciation and complexation interferences. These complications can usually be avoided by measuring the liquid phase concentration with a high background ionic strength and at a specified pH condition. One does not generally quantify the number of oxyanions adsorbed by direct measurement of the solid-phase concentration. This is because many of the solid phase measuring methods require expensive instruments and/or the quantitative precision is not satisfactory.

Techniques for investigation of oxyanions adsorbed at mineral–water interfaces

Various surface-sensitive techniques have been used to obtain structural information of surfaces and surface species that are needed to identify adsorption mechanisms and reactions. An important aspect of the application of spectroscopic techniques to the investigation of adsorbed species is the experimental conditions under which the spectra of the sample are collected. Invasive or ex situ methods commonly require extensive sample desiccation in a vacuum, which have the potential to modify or destroy sorption complexes usually present at the solid–liquid interface in an aqueous environment. Conversely, noninvasive or in situ methods require little or no alteration of a sample from its natural state. Clearly, to resolve the structure of surface species present at a mineral–water interface of natural particles, in situ spectroscopic techniques are required that can be applied in the presence of liquid water (Savoye et al., 2021). Three examples of in situ techniques that have been used to investigate oxyanions on mineral–water interfaces are attenuated total reflectance–Fourier-transformed infrared (ATR–FTIR), Raman, and X-ray absorption spectroscopy (XAS). A brief overview of these techniques is given below.

The predominant aim of spectroscopic analysis of adsorbed oxyanions at the solid–liquid interface is to understand the symmetry and binding mechanism of the anion on the surface. Quantitative measurements (“How many ions are adsorbed?”) are not the primary focus because, as was noted above, this is much more easily determined based on mass balance of the elements added and recovered from the liquid phase. Spectroscopy is the only method, however, that can determine how the ions are held on the surface. Fig. 1 shows five commonly encountered possibilities. Using selenite as the example oxyanion, it can be held as an outer sphere (Fig. 1A), where at least one molecule of water exists between the anion and the surface. It may be embedded in the internal regions of the mineral (Fig. 1B) due to solid diffusion of the anion into the mineral or to some coprecipitation of the anion with a cation onto the surface of the mineral. Fig. 1C–E shows the anion held as an inner sphere, where the anion bonds directly with the mineral surface. An inner-sphere anion can be monodentate (Fig. 1D) or bidentate (Fig. 1C and E). ‘Monodentate’ means that only one bond is formed between the anion and the surface, whereas ‘bidentate’ means that two bonds are formed between the anion and the surface. The bidentate inner-sphere bonds can be mononuclear (Fig. 1C) or binuclear (Fig. 1E), which refer to the minimum number of mineral cations that are required to link the anion’s oxygen ions to the mineral.

Vibrational spectroscopies

Vibrational spectroscopic techniques such as FTIR and Raman spectroscopy probe the vibrational properties of molecules (Sun and Doner, 1996). Characteristic vibrational properties of oxyanions originate from X–O(H) stretching and bending vibrations. Each oxyanion has its characteristic pattern of vibrations that relates to its symmetry state. The symmetry, and consequently the vibration pattern, changes when an oxyanion complexes with other ions such as protons or metal cations. Changes in the number and

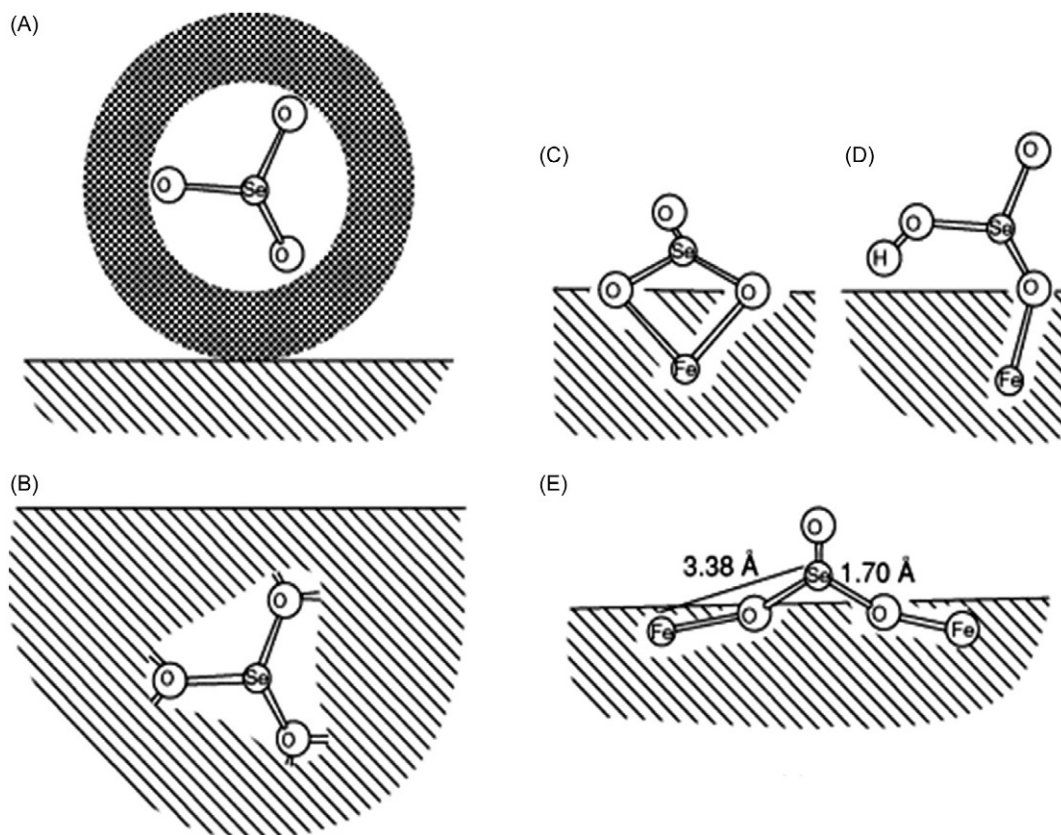


Fig. 1 Schematic diagram of several possible geometric arrangements of an oxyanion at the solid–water interface. Selenite is the adsorbate oxyanion and an iron oxide the adsorbent mineral. The area above each surface (indicated by solid horizontal lines) is water, and the solid mineral phases are below the solid horizontal lines. Reprinted from Brown GE Jr, Parks GA, and Chisholm-Brause CJ (1989) In-situ X-ray absorption spectroscopy studies of ions at oxide–water interfaces. *Chimia* 43: 248–256.

position of the vibrational bands provide information about the speciation and coordination state of the oxyanion. This behavior is the underlying principle of the determination of the types of surface species using vibrational spectroscopy.

Spectra of outer-sphere complexes are similar to the spectra of the free anion in solution. Differences in position and number of adsorption bands between the spectra of solution and adsorbed species indicate coordination with the surface, that is, inner-sphere complexation. Determination of the coordination state of an inner-sphere complex (mono- or bi-dentate and mono- or bi-nuclear) is usually based on the comparison of spectra of adsorbed species with data from the literature of spectra of coordinated compounds with known structures. Extensive infrared and Raman spectral data of inorganic and coordination compounds are available.

Attenuated total reflectance–Fourier-transformed infrared (ATR–FTIR) spectroscopy is an in situ, infrared sampling technique that can provide information about surface species at the mineral–water interface. Note that, since water strongly adsorbs infrared, it was the introduction of ATR (and other similar detection modifications, such as cylindrical internal reflectance, CIR) that has given us the necessary breakthrough to use FTIR effectively in aqueous (i.e., natural) media. This technique has been applied successfully to investigate various anions (PO_4^{3-} , $\text{B}(\text{OH})_4^-$, SO_4^{2-} , CO_3^{2-} , NO_3^-) at metal oxide–solution interfaces (Bowmaker et al., 2005).

The use of ATR–FTIR does not fully avoid water absorption of the infrared signal. The spectral window in metal oxide aqueous suspensions is limited by the strong adsorption by water and metal oxide lattice vibrations in the region below 1000 cm^{-1} and the strong adsorption by water in the region of $3000\text{--}4000\text{ cm}^{-1}$. For example, vibrations of the oxyanions of Se and As are expected below 1000 cm^{-1} and, consequently, are difficult to observe in these systems with the ATR–FTIR technique.

Diffuse reflectance FTIR (DRIFT) is an infrared sampling technique that requires air-drying of the sample. In general, this does not greatly affect the surface speciation and coordination of the adsorbed oxyanion because a very thin layer of water should remain on the mineral surface. Because there is less interference with the strong absorbing of water, this sampling technique enables the investigation of OH-stretching in the spectral range of $3400\text{--}3800\text{ cm}^{-1}$. The OH-stretching bands can provide information about the effect of oxyanion adsorption on the surface OH groups of the mineral surface.

Raman spectroscopy can provide a window in the spectral region below 1000 cm^{-1} to investigate aqueous minerals and, accordingly, can provide complementary data to IR spectroscopy (Dobrea et al., 2015; Nakamoto, 2009). One of the attractive aspects of Raman spectroscopy to minerals in aqueous suspension is that it is transparent to water, which makes in situ investigation of surface species possible. The recent advances in Raman instrumentation have brought new opportunities for applications to investigate surface species at mineral–water interfaces. Raman spectroscopy has been used to study the interfacial species on metal oxides of SeO_4^{2-} , SO_4^{2-} , AsO_4^{3-} and AsO_3^{3-} .

X-Ray absorption spectroscopy

X-ray absorption spectroscopy (XAS) is another technique that can be applied to in situ studies of adsorbates at mineral–water interfaces (Fendorf et al., 1997). It is an element-specific, short-range probe that provides information about the local structural and compositional environment of an adsorbing atom. X-ray absorption near-edge structure (XANES) spectra provide information about the bonding and oxidation state of the adsorbing atom. Extended X-ray absorption fine structure (EXAFS) can provide information about the composition and structure in the vicinity of the element. The absence of back-scattering beyond the first shell indicates an outer-sphere complex. In the case of inner-sphere complexes, the number, distance, and type of back-scatterers provide information for identifying the type of inner-sphere complex (mono-versus bidentate). X-ray absorption spectroscopy has been used to identify the adsorption mechanisms of the oxyanions of Se, As, and Cr (Bullen et al., 2022). The technique requires high-energy X-rays, which are only available at synchrotron facilities.

Sorption strength of oxyanions

The strength of the bond of oxyanions with mineral surfaces strongly affects the fate of the oxyanion in the environment. Weakly held anions leach easily and are lost from the upper soil horizons down toward the groundwater or are readily carried in surface runoff to nearby streams. There are a few ways to determine the relative strength of these bonds. One common method involves adsorption isotherms, and another involves adsorption edges.

An adsorption isotherm analysis is based on measuring the amount of anion adsorbed by a solid phase as a function of the equilibrium concentration present in the liquid phase, as shown in Fig. 2 for sulfate adsorption onto an aluminum oxide (Wijnja and Schulthess, 2000). The curvature of the isotherm, which is particularly visible at low aqueous concentrations, is a function of the adsorption strength. Qualitatively, the sharper are the rate of increase and amount of curvature observed in the adsorption isotherm, the stronger is the adsorption strength of the compound with the surface. It is the curvature of the adsorption isotherm, not the adsorption maximum that indicates the strength of adsorption.

Efforts to quantify adsorption strength have been the focus of much debate in the literature for almost a century. For example, if we assume a simple adsorption mechanism:



then the amount adsorbed can be predicted mathematically using data-fitting techniques with the Langmuir isotherm equation:

$$\text{Amount Adsorbed} = MKc/(1 + Kc) \quad (2)$$

where M is the maximum adsorption, which can be obtained from the isotherm at large aqueous concentrations; K is the Langmuir constant, which is equal to the ratio P/R , with P the adsorbed anion concentration, R the (concentration of unoccupied surface

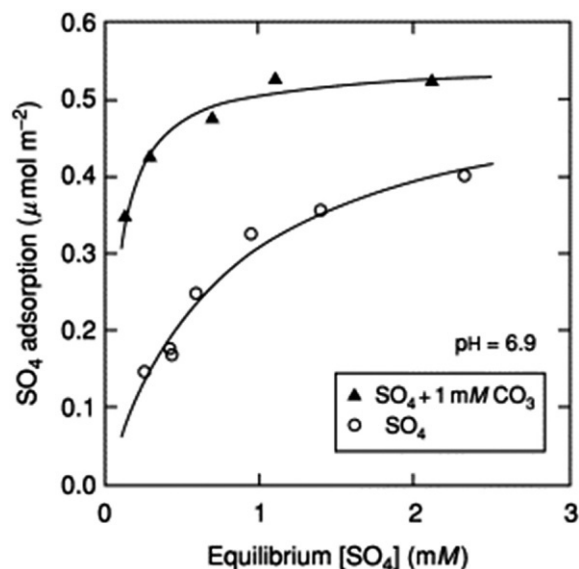


Fig. 2 Adsorption isotherms of sulfate on an aluminum oxide (bayerite). Notice that the presence of carbonates increased the adsorption strength of sulfate, which is contrary to the competitive adsorption behavior typically observed when two adsorbing anions are in solution. Reprinted from Wijnja H and Schulthess CP (2000) Interaction of carbonate and organic anions with sulfate and selenate adsorption on an aluminum oxide. *Soil Science Society of America Journal* 64: 898–908.

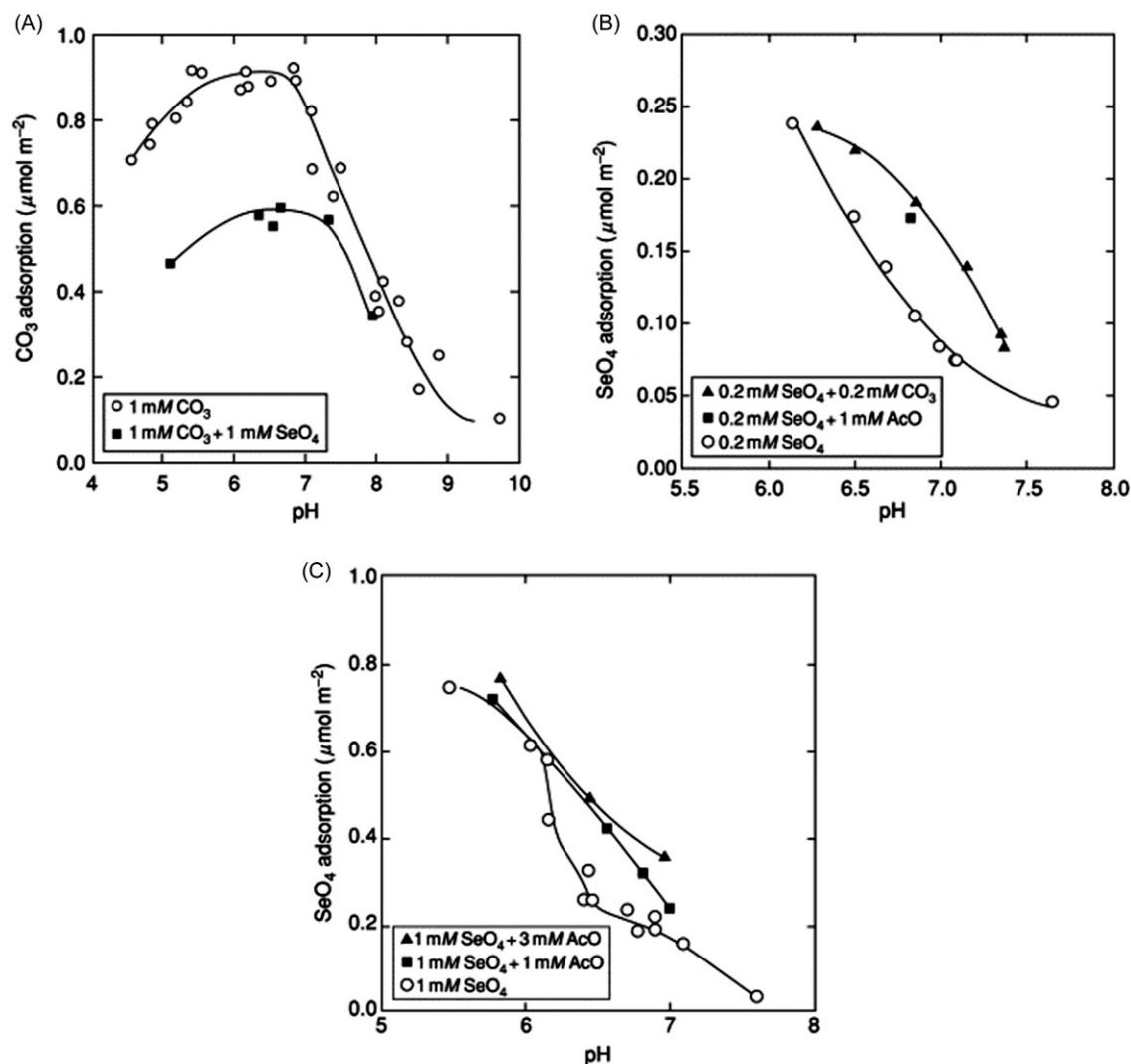
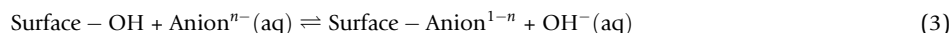


Fig. 3 Adsorption envelopes (adsorption as a function of pH) for various oxyanions on goethite: (A) adsorption of carbonate. Notice the competitive behavior of selenate on carbonate adsorption; (B, C) adsorption of selenate in the presence of carbonate and acetate. Notice the enhancement by these oxyanions of selenate adsorption. Reprinted from Wijnja H and Schulthess CP (2002) Effect of carbonate on the adsorption of selenate and sulfate on goethite. *Soil Science Society of America Journal* 66:1190–1197.

sites) $\times$ (aqueous anion concentration); and c is the aqueous anion concentration at reaction equilibria. The adsorption strength is quantified here by the Langmuir constant, K . There is considerable debate, however, about how the adsorption mechanism should be modeled; that is, modeling the adsorption reaction by Eq. (1) and other similar equations is not yet fully validated. Note that a model that does not fit the data is clearly invalid. However, a model that does fit the data is not necessarily automatically adequately validated.

One can also infer the relative strength of an adsorbing anion based on the pH value of the adsorption edge, which is shown in Fig. 3 (Wijnja and Schulthess, 2002). For a given mixture, the amount of aqueous anion adsorbed by the solid phase depends strongly on the equilibrium pH of the system. Working through the pH scale, first there is no adsorption of anions at high pH values. At intermediate pH values, there is an abrupt increase in adsorption. The pH value measured at the midpoint of this abrupt increase in adsorption is the pH value of the adsorption edge. At much lower pH values, there is a decrease in the number of anions adsorbed. The entire bell-shaped curve is known as an adsorption envelope.

If we assume that the adsorption mechanism involves the co-adsorption of hydroxyl anions (OH^-), for example:



then we can use the pH value of the adsorption edge to compare the relative adsorption strength of the anion. At equimolar concentrations, the anion that has a larger adsorption edge value is the one that binds more strongly with the surface. Notice in Eq. (3) that a strong surface–anion bond will require a greater concentration of OH^- anions to reverse the reaction. The midpoint of

equilibria for Eq. (3) occurs at the adsorption edge. Accordingly, if the adsorption of an anion is strong, then the number of OH^- anions needed to reverse the reaction will be large, and the pH value of the adsorption edge will also be high.

Efforts to quantify the adsorption strength based on the pH value of the adsorption edge are not yet fully resolved. The problem arises because we are not sure that Eq. (3) or other similar equations fully describe the adsorption process. Consequently, the mathematical interpretation of pH-dependent data is still discussed extensively in soil science research articles. In addition, it is noted that while some progress has been made in comprehending the adsorption of oxanions on mineral surfaces, there are still gaps that are not fully addressed. Ongoing research is needed to perceive fully the underlying mechanisms and to develop accurate models that can predict the adsorption behavior of different oxanions under different conditions. We do know, however, that if the pH is kept constant, then Eq. (3) will simplify to Eq. (1), and equations similar to Eq. (3) will simplify to equations similar to Eq. (1). We also know that anion adsorption involves the release of hydroxyl anions. Note that it is difficult to distinguish hydroxyl release (which chemically removes protons from solution) from proton coadsorption (which physically removes protons from solution). Accordingly, essentially identical concepts can be presented using equations that illustrate proton co-adsorption to the surface rather than hydroxyl release from the surface.

With the tools outlined above, a general review of the published data suggests the following trend: the larger is the negative charge present on the oxyanion, the stronger is the affinity of the oxyanion toward the surface. Within each group, there is much uncertainty about which is more strongly adsorbed, and this is probably due to numerous reasons (such as, presence of surface contaminants). In general, the affinity of trivalent oxanions follows the sequence: $\text{AsO}_3^{3-} > \text{PO}_4^{3-} \approx \text{AsO}_4^{3-}$. Divalent oxanions adsorb more weakly, and in general follow the affinity sequence $\text{SeO}_4^{2-} > \text{CrO}_4^{2-} > \text{CO}_3^{2-} > \text{SeO}_3^{2-} \approx \text{SO}_4^{2-} > \text{MoO}_4^{2-}$. Monovalent oxanions such as ReO_4^- , NO_3^- and ClO_4^- , are very weakly adsorbed. Another, more useful observation is the close correlation of adsorption affinities of oxanions with the stability constants (pK_a values) of their conjugate acids; noteworthy is the high adsorption affinity of $\text{B}(\text{OH})_4^-$ by minerals. For these reasons, nitrates or perchlorates are sometimes used as background electrolytes in adsorption studies and are often assumed not to adsorb at all or to cause no significant effect on the adsorption process of the other compounds present.

Effect of ions on adsorption of oxanions

The presence of dissolved electrolytes in the mixture affects the adsorption of oxanions (Li et al., 2022; Rietra et al., 2000). Adsorption studies, therefore, are pursued under constant ionic-strength conditions. These background electrolytes consist of dissolved cations and anions such as Na^+Cl^- or Na^+NO_3^- . It is critical to the adsorption study that the reactivity of these ions with the surface be minimal, preferably completely unreactive, or at least constant in their reactivity. Sometimes, an aqueous phase complexation reaction occurs between the oxyanion being studied and the components of the background electrolyte. This will complicate the analysis and should be avoided as much as possible, preferably by using a different background electrolyte.

The presence of dissolved electrolytes affects the adsorption process in numerous ways. First, it raises the ionic strength of the solution. This tends to decrease the electrostatic attraction of the adsorbate due to the lowering of surface potential or the masking of the surface charge. This effect on the adsorption process is not serious provided that the ionic strength remains constant throughout the study.

Second, the background anion can compete with the oxyanion for adsorption sites, which is very likely at high ionic-strength concentrations. Although adsorption of the background anion (e.g., Cl^-) is typically extremely weak, it is rarely zero and its competitive effect increases significantly as its concentration in solution increases (Schulthess and Hu, 2001). This effect on the adsorption process is not serious as long as the concentration of background ions is kept constant. It will, however, cause some complications to attempt to understand or quantify the adsorption strength of the oxyanion. When comparing adsorption isotherms or adsorption edges of adsorption envelopes, it is beneficial if the oxanions are compared at the same ionic strength and initial conditions. This effect is shown in Fig. 3A at low pH, where the concentration of weakly competitive aqueous Cl^- anions has increased to the point of rendering it significantly competitive with the carbonate oxyanion. The source of the Cl^- anion at low pH is from the addition of acids (namely, HCl) needed to achieve the low-pH conditions (Schulthess and Sparks, 1991).

Competition between anions is also shown in Fig. 3A, where the presence of selenate significantly reduces the adsorption of carbonate. Competition between oxanions is commonly observed. Competitive behavior should be expected because of the structural similarities of the oxanions and the similarity of their adsorption sites (Tahmasebi, 2020). Large Cl^- concentrations will also decrease the amount adsorbed over a broad pH range due to competitive adsorption.

Third, reactivity of the background anion (e.g., Cl^-) with the surface may be pH-dependent. Note that the first two problems mentioned above assume, for simplicity, that the reactivity of the background anion is pH independent. However, this is perhaps one of the most serious problems that can arise in the study of oxyanion-adsorption processes. In these situations, reactivity of the background anion must be monitored if the adsorption reactivity of the oxyanion with the solid surface is to be properly understood.

Fourth, some anions may be present in the mixture that enhance the adsorption reaction of the aqueous oxyanion with the solid surface. Note that this is a counterintuitive process because traditional views do not include an anion that causes an increase in adsorption of another similarly charged ion. Fig. 2 shows a sharp increase in adsorption of sulfate on an aluminum oxide in the presence of carbonate. Fig. 3B and C shows that this enhancing effect is also pH-dependent (Wijnja and Schulthess, 2002). The enhancement of oxyanion adsorption, such as sulfate and selenate adsorption on aluminum and iron oxides, has been observed in

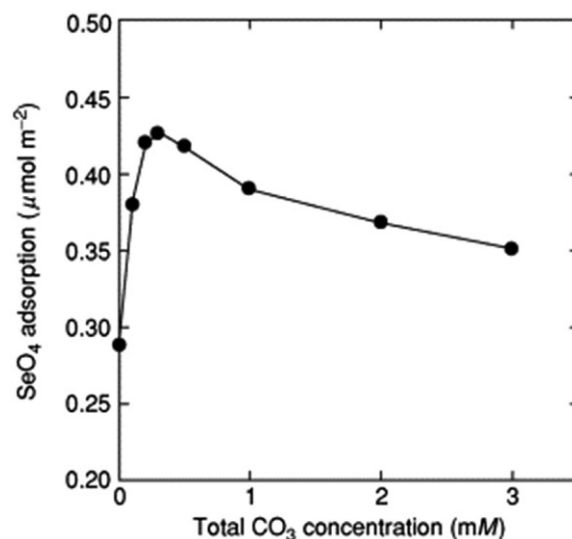


Fig. 4 Adsorption of 1 mM selenate at pH 6.8 on an aluminum oxide (bayerite) as a function of the amount of carbonate added. Notice that the presence of carbonates increases the adsorption of selenate at low carbonate concentrations. Reprinted from Wijnja H and Schulthess CP (2000) Interaction of carbonate and organic anions with sulfate and selenate adsorption on an aluminum oxide. *Soil Science Society of America Journal* 64: 898–908.

the presence of carbonates, acetate, and formate anions. At large concentrations, these anions seem to become competitive. Hence, an enhancing effect does not exclude the existence of a concurrent competitive effect. The enhancing effect is most pronounced at small concentrations, where any possible competitive effect would also be at its lowest level (Fig. 4) (Wijnja and Schulthess, 2000). It is not clear how prevalent the enhancing effect is in nature or in laboratory studies. The co-adsorption effect, where the presence of certain anions enhances the adsorption of other oxyanions to solid surfaces, is a topic of ongoing research. Studies have demonstrated this effect in various systems, but the underlying mechanisms and prevalence in natural systems are not fully understood. Further research is needed to comprehend fully the implications of the co-adsorption effect on the fate and transport of oxyanions in the environment. We do know, however, that carbonates are nearly always present in aquatic systems. Their concentration in laboratory solutions varies widely depending on the purity of water sources used or on the techniques used to minimize or control the carbonate concentrations. In other words, aqueous carbonate impurities can easily enter sample preparations unbeknown to the experimenters and cause a wide range in the variance of the data collected. This creates a serious problem for those that wish to understand and quantify the adsorption of oxyanions by solid surfaces in aqueous media.

Sorption mechanisms of oxyanions

The adsorption reaction of oxyanions at a mineral surface is described as a surface complexation reaction between the oxyanion and a surface functional group forming a surface complex. With respect to the types of surface complexes or adsorption mechanisms, there are commonly two major types distinguished: inner-sphere and outer-sphere complexes. An inner-sphere complex is formed by a ligand-exchange reaction with a surface-OH or surface -OH^+_2 group and is directly coordinated with the underlying surface metal ion. In a monodentate inner-sphere complex, one bond is formed with a surface group. In a bidentate complex, two bonds are formed, either with one surface metal ion (mononuclear) or with two surface metal ions (binuclear). In contrast, outer-sphere complexes form mainly by electrostatic interactions and at least one water molecule is interposed between the oxyanion and the mineral surface. Diffuse-swarm ions do not form a complex with a surface functional group, but are bound only through electrostatic interactions that neutralize the surface charge (Zheng et al., 2020). Diffuse-swarm is a specific type of adsorption mechanism where the anion does not form a complex with the surface functional group, but instead is bound only through electrostatic interactions that neutralize the surface charge (McBride, 1994). Spectroscopic tools are essential for the development of detailed descriptions of the adsorbed oxyanion. A mechanistic description of the adsorption process requires spectroscopic information of the adsorbed species, information on the stoichiometry of the reaction and rates of reaction, and a measure of the strength of adsorption.

Oxyanions with a relatively strong adsorption affinity such as PO_4^{3-} , AsO_4^{3-} , SeO_4^{2-} and B(OH)_4^- form predominantly inner-sphere complexes on metal oxide surfaces. Both monodentate and bidentate complexes have been found on metal oxides for several of these oxyanions (Lepore et al., 2022). For example, it has been shown that the relative distribution of monodentate and bidentate inner-sphere complexes of AsO_4^{3-} and CrO_4^{2-} varies with the surface coverage.

The CO_3^{2-} anion forms monodentate inner-sphere complexes at metal oxide–water interfaces. Oxyanions with weak affinity, such as NO_3^- and ClO_4^- adsorb mainly as outer-sphere and diffuse-swarm ions.

Several anions that have high or moderate sorption affinity, such as AsO_4^{3-} , SeO_4^{2-} and SO_4^{2-} , show intermediate behavior, where both inner- and outer-sphere surface complexes coexist at metal oxide–water interfaces. For these moderately adsorbing anions, the relative distribution of inner- versus outer-sphere adsorption depends on pH, ionic strength, and type of mineral.

When an anion is held as an inner-sphere compound, then it has probably been preceded by an initial outer-sphere attraction to the mineral surface (Tahmasebi, 2020). The mechanism of forming an inner-sphere bond must involve the release of a molecule of water from the solid–liquid interface into the bulk solution. A net volume change occurs to the system whenever this process occurs, and this can be measured with a dilatometer. Based on measured volume changes, there is approximately one molecule of H_2O released into the bulk solution for every aqueous molecule of SO_4^{2-} , SeO_4^{2-} , or H_2PO_4^- adsorbed onto an amorphous iron hydroxide at pH 4.5, which should be expected for inner-sphere adsorption of these oxyanions (Yamaguchi et al., 1999).

Anion adsorption is accompanied by proton uptake. The net proton uptake is the result of a ligand exchange reaction (OH release), surface protonation, oxyanion protonation on the surface, or a combination of these. Quantifying the number of protons co-adsorbed (or OH released) by the surface per molecule of oxyanion adsorbed is an important component of our understanding of the adsorption process. This stoichiometry of the reaction directly affects how the surface reaction will be described, such as the 1:1 stoichiometry described in Eq. (3). This, in turn, also affects how the surface complexation model will be developed. The best way to obtain data on the adsorption or co-adsorption of protons is by an acid–base titration method known as the backtitration technique (Schulthess and Sparks, 1986).

Anion adsorption affects the electrostatic properties of mineral surfaces. The net surface charge of a mineral surface is the result of proton-transfer reactions and surface-complexation reactions. Electrophoretic mobility measurements can reveal the net changes in surface charge due to oxyanion adsorption. In general, the adsorption of an oxyanion tends to lower the positive charge present on a surface. The positively charged surface does not necessarily become negatively charged following the adsorption of oxyanions. Rather, it tends to approach a net neutral surface charge. Taking these data in conjunction with the proton co-adsorption data mentioned above, the adsorption of oxyanions and proton co-adsorption tends to form net neutral products on the mineral surfaces.

The adsorption of oxyanions generally reaches a maximum at around the pK_a value for the conjugate Bronsted acid, as shown in Fig. 3A for carbonate adsorption ($\text{pK}_a = 6.3$). One can generally infer that the higher is the value of the oxyanion's acidity constant (pK_a), the stronger will be its adsorption by a hydrophilic surface. This further reflects the importance of the co-adsorption of protons on the adsorption mechanism of oxyanions (Figueiredo and Quintelas, 2014). The mechanism of bringing the co-adsorbed protons to the mineral surface does not necessarily follow a separate, independent path from that taken by the adsorbing oxyanion. The proton can be brought toward the mineral surface by the protonated oxyanions. The fate of the co-adsorbed proton may be: (1) to remain complexed with the adsorbed oxyanion (hence, the oxyanion is held as a Bronsted acid on the mineral surface); (2) to remain complexed with the surface (hence, the oxyanion's negative charge is neutralized by a positively charged surface–surface, which may or may not be directly below or adjacent to the adsorbing oxyanion); or (3) it may react with a surface-OH group and form an H_2O product that diffuses back to the bulk solution (this may also be described schematically as a hydroxyl exchange mechanism, where the hydroxyl is neutralized by free protons in the aqueous phase).

Site preference and site reactivity

Surface hydroxyl groups constitute the complexation sites on metal oxide surfaces suspended in aqueous solutions. The surface groups behave as weak Bronsted acids or bases, capable of ionizing to yield a pH-dependent charged surface or serving as a ligand in surface complexation reactions. Analysis of the crystal structure of metal oxides indicates the existence of various types of surface hydroxyl groups with different reactivities. Their reactivity depends, among other factors, on their coordination environment (singly, doubly, and triply metal-coordinated). The different crystal planes of a metal oxide mineral have a different composition with respect to the types of surface groups and densities of these groups.

With respect to complexation of oxyanions, the singly coordinated surface hydroxyls are known to be most reactive. Recent studies have used advanced techniques such as XAS and FTIR to provide more detailed information on the structure of adsorbed oxyanions at mineral–water interfaces, which can provide new insights into the complexation mechanisms (Tamijani et al., 2020). An FTIR study on the effect of As oxyanions on the various types of surface groups has indicated that the singly coordinated surface groups are involved in the ligand exchange reaction to form inner-sphere complexes (Manning et al., 1998).

There are several things that the adsorbed oxyanion can do after it has adsorbed to the mineral surface. It can desorb from the surface and diffuse back to the aqueous bulk solution. Given the high energy of hydration of many of these oxyanions, this desorption step is very likely and is the basis for establishing the equilibrium condition where only a specific density of anions remain adsorbed to the surface for a given concentration of anions in solution.

The adsorbed oxyanion can also diffuse into the mineral through a process known as solid diffusion (product is shown in Fig. 1B). Solid diffusion at ambient temperatures is a very slow process. However, the oxyanion may also become entrapped inside the mineral through the process of oxyanion-enhanced dissolution of the mineral followed by co-precipitation of the oxyanion–metal complex.

A final option for the adsorbed oxyanion is to migrate laterally, traveling horizontally on the mineral surface along a two-dimensional matrix. This depends on the strength of the bond of the oxyanion with the surface. The lateral, two-dimensional

mobility of the oxyanion allows it to migrate toward a more stable adsorption site (more stable, that is, relative to its point of initial collision and reaction with the surface). Lateral mobility of adsorbed ions is known to occur on solid–gas interfaces, but data are missing that quantify accurately this mobility on solid–liquid interfaces.

Conclusions

Adsorption of oxyanions to mineral–water interfaces plays a crucial role in determining their fate in the environment. A variety of techniques such as ATR–FTIR, Raman and XAS have been used to study the structural information of surfaces and surface species, and in situ methods are preferred because they require little or no alteration of the sample from its natural state. It has been determined that oxyanions with a high adsorption affinity, such as PO_3^{4-} , AsO_3^{3-} , SeO_2^{3-} and $\text{B}(\text{OH})_4^-$, form predominantly inner-sphere complexes on metal oxide surfaces. In contrast, oxyanions with low affinity, such as NO_3^- and ClO_4^- , mainly form outer-sphere and diffuse-swarm ions. The adsorption strength of oxyanions can be quantified by methods such as the Langmuir isotherm equation and pH value of the adsorption edge. However, the mechanism of adsorption is still debated, and further research is needed to understand fully the process. Overall, understanding the adsorption of oxyanions to mineral–water interfaces is crucial for assessing soil and environmental quality, and for developing effective management strategies.

References

- Bowmaker GA, Marfua S, Skelton BW, and White AH (2005) Syntheses, structures and vibrational spectroscopy of some 1:1 and 1:2 adducts of silver (I) oxyanion salts with 2, 2'-bis (pyridine) chelates. *Inorganica Chimica Acta* 358: 4371–4388.
- Bullen JC, Lapinee C, Miller LA, Bullough F, Berry AJ, Najorka J, Cibin G, Vilar R, and Weiss DJ (2022) Spectroscopic (Xas, FTIR) investigations into arsenic adsorption onto TiO₂/Fe₂O₃ composites: Evaluation of the surface complexes, speciation and precipitation predicted by modelling. *Results in Surfaces and Interfaces* 9: 100084.
- Dobrea ID, Ciocan CE, Dumitriu E, Popa MI, Petit E, and Hulea V (2015) Raman spectroscopy—Useful tool for studying the catalysts derived from Mo and V-oxyanion-intercalated layered double hydroxides. *Applied Clay Science* 104: 205–210.
- Fendorf S, Eick MJ, Grossl P, and Sparks DL (1997) Arsenate and chromate retention mechanisms on goethite. 1. Surface structure. *Environmental Science & Technology* 31: 315–320.
- Figueiredo H and Quintelas C (2014) Tailored zeolites for the removal of metal oxyanions: Overcoming intrinsic limitations of zeolites. *Journal of Hazardous Materials* 274: 287–299.
- Hussain MM, Bibi I, Niazi NK, Shahid M, Iqbal J, Shakoor MB, Ahmad A, Shah NS, Bhattacharya P, and Mao K (2021) Arsenic biogeochemical cycling in paddy soil-rice system: Interaction with various factors, amendments and mineral nutrients. *Science of the Total Environment* 773: 145040.
- Lepore GO, Schingaro E, Mesto E, Lacalamita M, Cristiani C, Stampino PG, Dotelli G, Finocchio E, and D'ACAPITO, F. & Giuli, G. (2022) Lanthanum captured in montmorillonite: Evidence of inner-sphere complexes from X-ray Absorption Spectroscopy investigations. *Applied Clay Science* 230: 106676.
- Li M, Sugimoto T, Yamashita Y, and Kobayashi M (2022) Aggregation and charging of natural allophane particles in the presence of oxyanions. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* 649: 129413.
- Liang X, Lin X, Wei G, Ma L, He H, Santos-Carballal D, Zhu J, Zhu R, and De Leeuw NH (2021) Competitive adsorption geometries for the arsenate As (V) and phosphate P (V) oxyanions on magnetite surfaces: Experiments and theory. *American Mineralogist: Journal of Earth and Planetary Materials* 106: 374–388.
- Manning BA, Fendorf SE, and Goldberg S (1998) Surface structures and stability of arsenic (III) on goethite: Spectroscopic evidence for inner-sphere complexes. *Environmental Science & Technology* 32: 2383–2388.
- McBride MB (1994) *Environmental Chemistry of Soils*, Illustrated ed New York, NY: Oxford University Press.
- Nakamoto K (2009) *Infrared and Raman Spectra of Inorganic and Coordination Compounds, Part B: Applications in Coordination, Organometallic, and Bioinorganic Chemistry*. John Wiley & Sons.
- Priya E, Kumar S, Verma C, Sarkar S, and Maji PK (2022) A comprehensive review on technological advances of adsorption for removing nitrate and phosphate from waste water. *Journal of Water Process Engineering* 49: 103159.
- Rietra R, Hiemstra T, and Van Riemsdijk W (2000) Electrolyte anion affinity and its effect on oxyanion adsorption on goethite. *Journal of Colloid and Interface Science* 229: 199–206.
- Savoye S, Schlegel M, and Frasca B (2021) Mobility of selenium oxyanions in clay-rich media: A combined batch and diffusion experiments and synchrotron-based spectroscopic investigation. *Applied Geochemistry* 128: 104932.
- Schulthess C and Hu Z (2001) Impact of chloride anions on proton and selenium adsorption by an aluminum oxide. *Soil Science Society of America Journal* 65: 710–718.
- Schulthess C and Sparks D (1986) Backtitration technique for proton isotherm modeling of oxide surfaces. *Soil Science Society of America Journal* 50: 1406–1411.
- Schulthess C and Sparks D (1991) Equilibrium-based modeling of chemical sorption on soils and soil constituents. In: *Advances in Soil Science*. Springer.
- Sun X and Doner HE (1996) An investigation of arsenate and arsenite bonding structures on goethite by FTIR. *Soil Science* 161: 865–872.
- Tahmasebi E (2020) Insights into the adsorption mechanism of Al₃₀ polyoxocations-modified graphene oxide nanosheets for efficient removal of phosphate, chromate and selenate oxyanions: A comparative study. *Journal of Molecular Liquids* 299: 112111.
- Tamijani AA, Bjorklund JL, Augustine LJ, Catalano JG, and Mason SE (2020) Density functional theory and thermodynamics modeling of inner-sphere oxyanion adsorption on the hydroxylated α -Al₂O₃ (001) surface. *Langmuir* 36: 13166–13180.
- Wijnja H and Schulthess C (2000) Interaction of carbonate and organic anions with sulfate and selenate adsorption on an aluminum oxide. *Soil Science Society of America Journal* 64: 898–908.
- Wijnja H and Schulthess C (2002) Effect of carbonate on the adsorption of selenate and sulfate on goethite. *Soil Science Society of America Journal* 66: 1190–1197.
- Xinxin M, Siebecker MG, Wenxian G, Ling L, and Wei L (2021) A review of cadmium sorption mechanisms on soil mineral surfaces revealed from synchrotron-based X-ray absorption fine structure spectroscopy: Implications for soil remediation. *Pedosphere* 31: 11–27.
- Yamaguchi NU, Okazaki M, and Hashitani T (1999) Volume changes due to SO₂–4, SeO₂–4, and H₂PO₄–4 adsorption on amorphous iron (III) hydroxide in an aqueous suspension. *Journal of Colloid and Interface Science* 209: 386–391.
- Zheng H, Zhang C, Liu B, Liu G, Zhao M, Xu G, Luo X, Li F, and Xing B (2020) *Biochar for Water and Soil Remediation: Production, Characterization, and Application. A New Paradigm for Environmental Chemistry and Toxicology*. Springer.

Sorption—Metals

Daniel G Strawn^a and Donald L Sparks^b, ^aUniversity of Idaho, Moscow, ID, United States; ^bDepartment of Plant and Soil Sciences, University of Delaware, Interdisciplinary Science and Engineering (ISE) Laboratory, Newark, DE, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	336
Terminology and role of sorption processes in soils	336
Properties of metals	337
Surface characteristics	337
Permanent charge minerals	337
Surface functional groups	339
Surface complexation	339
pH and ion selectivity effects on metal sorption	339
Metal surface precipitation	340
Summary and outlook	342
References	342

Abstract

Sorption of metals at the solid–solution interfaces of soil minerals and soil organic matter is an important process for metal retention and release in soils. Sorption reactions include inner-sphere and outer-sphere adsorption complexation, and surface precipitation. This chapter discusses different metal sorption reaction processes and the soil properties that influence them.

Key points

- Metals adsorb on soil mineral and organic matter surfaces via outer-sphere and inner-sphere complexation.
- Metals can form multinuclear precipitates on soil mineral surfaces.
- Multiple metal sorption mechanisms can occur within the same soil.
- Sorption mechanism depends on soil pH, ionic strength, and soil composition.

Introduction

The sorption (retention) of metals, including alkali, alkaline earth, transition, and other metals (e.g., Pb^{2+}) on soil mineral and organic constituents is one of the most important processes controlling their fate, transport, and bioavailability in soil and water environments (McKenzie, 1980; Sumner, 1998; Sauve et al., 2000; Sparks, 2002; Strawn et al., 2020). This chapter discusses the terminology for sorption processes, the role of surface functional groups on mineral and organic soil components in metal sorption, the types of surface complexes in soils that metals form, and the importance of surface precipitation reactions in sequestering metals in soils.

Terminology and role of sorption processes in soils

Adsorption and surface precipitation are examples of sorption—a general term that is used when the retention mechanism of a chemical on a surface is not specified. Adsorption is a type of sorption process describing the accumulation of a substance or material at an interface between the solid surface and the bathing solution. Adsorption describes the removal of chemicals from solution (solute) and attachment to surfaces as a separate phase from the solid. Adsorption does not include surface precipitation, which is the formation of a three-dimensional phase on a surface, or polymerization of small multinuclear inorganic species on mineral surfaces.

There are various metal sorption mechanisms that can occur at soil mineral surfaces that involve both physical and chemical processes (Fig. 1) (Sposito, 1987; Hiemstra et al., 1989; Goldberg, 1992). Some general terms that describe adsorption reaction processes are as follows: the solid surface is referred to as the adsorbent; the ion in solution that has the potential of being adsorbed is the adsorptive; and the material accumulated on a surface is the adsorbate. If the general term *sorption* is used, the material that accumulates at the surface, the solid surface, and the ion in solution are referred to as sorbate, sorbent, and sorptive, respectively.

Adsorption is one of the most important chemical processes in soils. It determines the quantity of plant nutrients, metals, pesticides, and other organic chemicals that are retained on soil surfaces, and therefore is one of the primary processes affecting

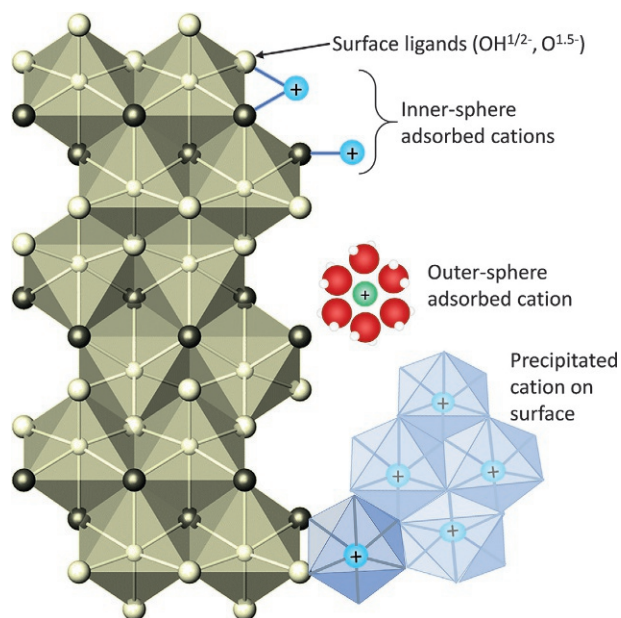


Fig. 1 Illustrations of sorption mechanisms on a metal oxide mineral surface.

transport of both nutrients and contaminants in soils. Adsorption also affects the electrostatic properties of particles in soils, e.g., coagulation and settling of suspended colloids.

Adsorption involves three types of interactions: (1) electrostatic interactions, including van der Waals forces (e.g., adsorption partitioning of uncharged solutes), (2) electrostatic outer-sphere complexes (e.g., ion exchange), and (3) formation of chemical bonds (Sposito, 1981; Bolt and Van Riemsdijk, 1979; Evangelou and Phillips, 2005). Chemical bonds are short-range molecular interactions, referred to as inner-sphere complexation, which involves covalent and ionic bonding between the ligand functional groups on soil mineral or organic matter surfaces and ions or uncharged chemicals that occur on the mineral surfaces (sorbate) (Bolt and Van Riemsdijk, 1979; Forbes et al., 1976; Goldberg, 1992).

Properties of metals

Metals occur in the environment in many species, including oxyanions (e.g., CrO_4^{2-}), hydrolysis complexes (e.g., NiOH^+), and hydrated cations (e.g., $\text{Na}^+(\text{H}_2\text{O})_6$ and $\text{Zn}^{2+}(\text{H}_2\text{O})_6$). Table 1 lists metals that are often of concern in the environment and their typical concentrations in soils. Other metals not listed in Table 1, such as rare earth elements and lanthanides and actinides, are of concern in some environments; for example, areas impacted by nuclear waste disposal. The sorption characteristics of metals depend on their ability to share electrons, their oxidation state, and their ionic size. Valence and ionic radius can be combined to calculate the ionic potential. For example, Mg^{2+} has a charge of two plus and a relatively small ionic radius, so has a large ionic potential that strongly attracts the partial negatively charged oxygen in water molecules. The result is the Mg^{2+} cation has a relatively strong or large hydration sphere that is maintained during adsorption as an outer-sphere complex. Other metals, such as Cu^{2+} or Pb^{2+} , have electron orbital vacancies that can be shared with the electrons of ligands on mineral and organic matter surfaces forming covalent bonds, which are called inner-sphere complexes (Strawn and Sparks, 1999).

Surface characteristics

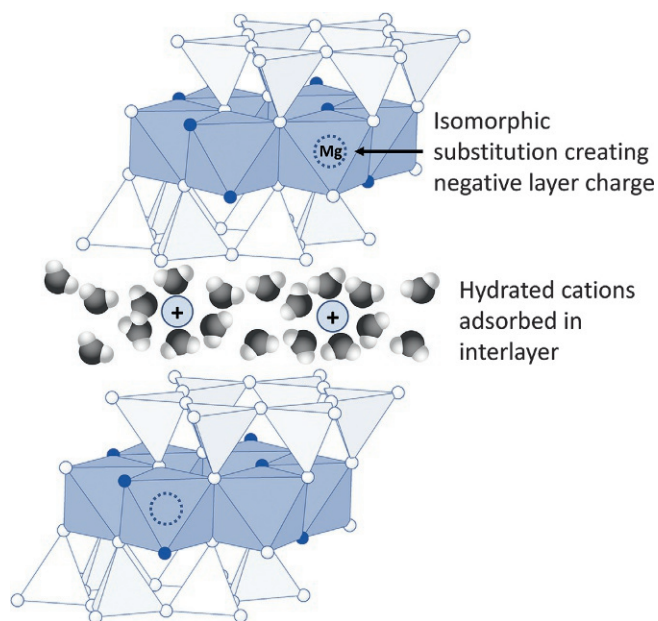
Another important consideration in metal sorption mechanisms are the properties of the soil particle surfaces. Some minerals, such as clays, have permanently charged surfaces that do not change charge magnitude or sign with changes in pH. Variable charged mineral surfaces have functional groups in which the charge magnitude and sign change as pH changes because of protonation and deprotonation of the acidic surface functional groups. The type of interaction of functional groups with adsorptive ions depends on the characteristics of the surface functional groups (i.e., acidity and molecular coordination to the mineral cations) (Kubicki et al., 2008), and the properties of the adsorptive ions.

Permanent charge minerals

Clay minerals in soils have permanent negative charges on their interlayer surfaces created from isomorphous substitution (Sumner, 1998; Strawn, 2021). Cations are attracted to the negatively charged interlayers through electrostatic forces. Most cations maintain

Table 1 Oxidation states and concentration of important metals in soils. Values are from Bowen (1979).

Element	Oxidation states	Range (mg kg ⁻¹)	Median
<i>Alkali metals</i>			
Sodium	Na ⁺	150–25,000	5000
Potassium	K ⁺	80–37,000	14,000
Cesium	Cs ⁺	0.3–20	4
<i>Alkaline earth metals</i>			
Magnesium	Mg ²⁺	400–9000	5000
Calcium	Ca ²⁺	700–500,000	15,000
Strontium	Sr ²⁺	4–2000	250
Barium	Ba ²⁺	100–3000	500
<i>Transition metals</i>			
Vanadium	V ^{5+a}	3–500	90
Cobalt	Co ²⁺	0.05–65	8
Chromium	Cr ^{6+a} , Cr ³⁺	5–1500	70
Molybdenum	Mo ^{6+a}	0.1–40	1.2
Manganese	Mn ⁴⁺ , Mn ²⁺	20–10,000	1000
Iron	Fe ³⁺ , Fe ²⁺	2000–550,000	40,000
Nickel	Ni ²⁺	2–750	50
Copper	Cu ²⁺	2–250	30
Zinc	Zn ²⁺	1–900	90
Cadmium	Cd ²⁺	0.01–2	0.4
Mercury	Hg ²⁺ , Hg ¹⁺ , Hg ⁰	0.01–0.5	0.06
<i>Other metals</i>			
Aluminum	Al ³⁺	10,000–300,000	71,000
Lead	Pb ⁴⁺ , Pb ²⁺	2–300	35

^aSpecies occurs as oxyanion.**Fig. 2** Metal adsorption in the interlayer of a clay mineral as a hydrated exchangeable cation. Adapted from Strawn D, Bohn HL and O'Connor GA (2020) *Soil Chemistry*, John Wiley & Sons: Hoboken, NJ, with permission.

their hydration sphere in the interlayer of the clays; K⁺ and Cs⁺ are exceptions because they have large ionic radii with low ionic potential and thus dehydrate and fit perfectly in the clay mineral structure becoming fixed in the interlayer. Hydrated metals adsorbed in the interlayer of clays are exchangeable, and referred to as exchangeable cations, which comprises the soil's cation exchange capacity (CEC) (Fig. 2) (Sposito, 1981).

Surface functional groups

Surface functional groups in soils play a significant role in adsorption processes (Anderson and Sposito, 1992; Chorover and Sposito, 1995; Goldberg, 1992; Hiemstra et al., 1989; Tournassat et al., 2016; Uehara and Gillman, 1980). A surface functional group is a chemically reactive molecular unit attached at the boundary of a solid with the reactive groups exposed to the soil solution. Surface functional groups can be organic (e.g., carboxyl, carbonyl, phenolic) or inorganic molecular units. The major inorganic surface functional groups in soils are hydroxyl groups that are associated with the edges of minerals such as kaolinite, amorphous materials such as allophane and ferrihydrite, and metal oxides, oxyhydroxides, and hydroxides such as goethite, hematite, and gibbsite. Minerals with reactive surface functional groups occur in all soils, but are especially important in highly-weathered soils.

Surface complexation

When a metal is adsorbed on a mineral or organic matter surface functional group, it is called a surface complex. The overall reaction is referred to as surface complexation. There are two types of surface complexes that can form: outer-sphere and inner-sphere (Fig. 1) (Strawn and Sparks, 1999). If a water molecule is present between the surface functional group and the bound ion or molecule, the surface complex is termed outer-sphere. If there is not a water molecule present between the ion or molecule and the surface functional group to which it is bound, this is an inner-sphere complex. Inner-sphere complexes can be monodentate (metal is bonded to only one oxygen) and bidentate (metal is bonded to two oxygens).

Outer-sphere complexes involve electrostatic coulombic interactions and are thus weak compared to inner-sphere complexes, where the metals are bound through covalent or ionic bonds. Outer-sphere complexation, also called cation exchange, is usually a rapid process that is reversible. Outer-sphere complexation occurs on surfaces that are of opposite charge to the adsorbate.

Inner-sphere complexation may be a slower reaction than outer-sphere complexation, and is sometimes irreversible (Selim and Sparks, 2001). Inner-sphere complexation changes the overall surface charge because the metal charge is added to the intrinsic surface charge. Surface charge affects the sorption process through electrostatic attraction and repulsion forces between the surface and the metal cations in solutions. Outer- and inner-sphere complexation can occur simultaneously on some mineral surfaces.

In recent decades, molecular-scale analytical techniques such as X-ray absorption fine-structure spectroscopy (XAFS) and Fourier transform infrared spectroscopy (FTIR) have been employed to determine the type of metal surface complexes and products that form on soil surfaces. These techniques are advantageous because they can analyze systems under natural conditions where water is present. Based on these studies, it is observed that many metals, such as transition metals and Pb primarily form inner-sphere complexes, while strongly hydrated alkali and alkaline earth metals form outer-sphere complexes. At higher metal loadings and pHs, sorption of metals such as Co, Cr, Ni, and Zn on phyllosilicates and metal hydr(oxides) can result in the formation of surface precipitates (Corey, 1980; Siebecker et al., 2018). Thus, environmental factors such as pH, surface loading, ionic strength, type of sorbent, and time all affect the type of sorption complex or product. An example of this is shown for Pb sorption on montmorillonite at ionic strengths (I) of 0.006 and 0.1 and a pH range of 4–8 (Fig. 3) (Strawn and Sparks, 1999). With XAFS analysis, it was determined that at a pH of 4.48 and an I of 0.006 outer-sphere complexation on basal planes in the interlayer regions of the montmorillonite predominated. At a pH of 6.77 and I of 0.1, inner-sphere complexation on edge sites of montmorillonite was most prominent, and at pH of 6.76, I of 0.006 and pH of 6.31, I of 0.1, both inner- and outer-sphere complexation occurred. These data show that there is a continuum of adsorption complexes that can exist in soils.

pH and ion selectivity effects on metal sorption

Sorption of metal cations on variable charged surfaces is pH-dependent and is characterized by a narrow pH range where sorption increases to nearly 100%, traditionally known as an adsorption edge (Fig. 4) (Harter, 1983; Sauve et al., 2000). The pH position of the adsorption edge for a particular metal cation is related to its hydrolysis or acid–base characteristics. In addition to pH, sorption of metals depends on sorptive concentration, surface coverage, and the type of the sorbent. For oxyanion metals, such as chromate, molybdate and vanadate (Table 1), adsorption on variable charged surfaces is greater at low solution pH than occurs at high pH (i.e., opposite adsorption trend from cations).

One can measure the relative affinity or selectivity of a cation for a sorbent. The properties of the metal cation, sorbent, and the solvent all affect the selectivity. With monovalent alkali metal cations, electrostatic interactions predominate, and the general order of selectivity is $\text{Li}^+ < \text{Na}^+ < \text{K}^+ < \text{Rb}^+ < \text{Cs}^+$ (Xu and Harsh, 1990). This order is related to the size of the hydrated radius. The cation with the smallest hydrated radius, Cs^+ , can approach the surface the closest and be held the most tightly. However, this order is highly variable, depending on the solid surface properties and solution conditions (i.e., pH, ionic strength, and presence of competing or interfering ions).

With divalent metal ions, there is little consistency in the selectivity order. The type of surface and the pH both appear to have major effects on the selectivity sequence. Divalent transition and heavy metal cations, both of which are often sorbed as inner-sphere complexes, are more strongly sorbed than alkali and alkaline earth metals.

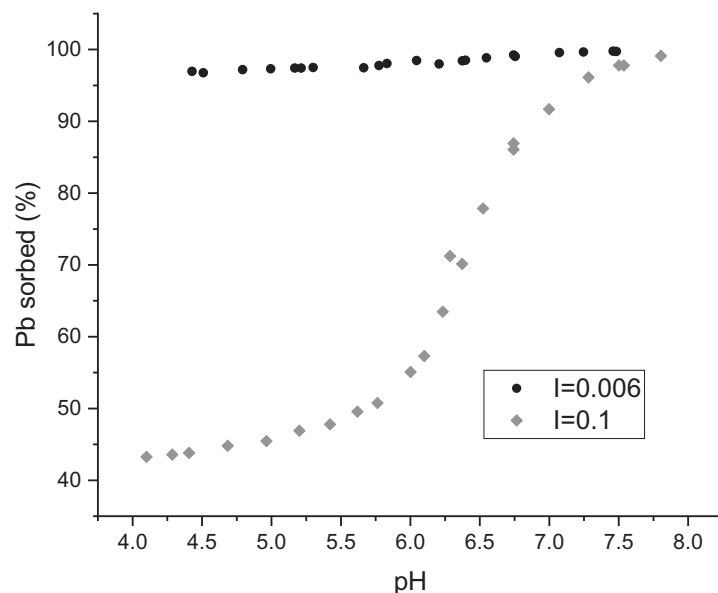


Fig. 3 The pH adsorption edges for Pb^{2+} adsorption on montmorillonite at two ionic strengths (I). Adapted from Strawn DG and Sparks DL (1999) The use of XAFS to distinguish between inner- and outer-sphere lead adsorption complexes on montmorillonite. *Journal of Colloid and Interface Science* 216: 257–269. with permission.

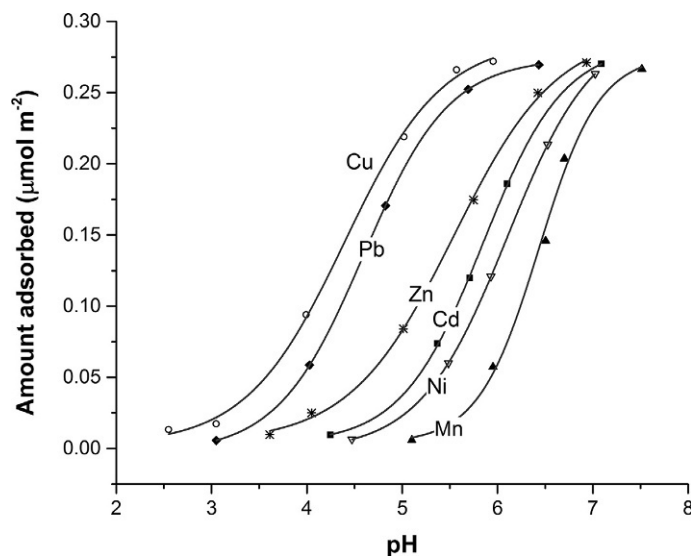


Fig. 4 The pH-dependent adsorption of metals on goethite. Reproduced with permission from McKenzie R (1980) The adsorption of lead and other heavy metals on oxides of manganese and iron. *Soil Research* 18: 61–73.

Metal surface precipitation

As the amount of metal cation sorbed on a surface increases, sorption can proceed from mononuclear adsorption to surface precipitation (Scheidegger et al., 1996). If the formation of a precipitate occurs under solution conditions that would, in the absence of a sorbent, be undersaturated with respect to any known solid phase, this is referred to as surface-induced precipitation. There are several thermodynamic reasons for surface precipitate formation: (1) the solid surface may lower the energy of nucleation by providing sterically similar sites; (2) the activity of the surface precipitate is <1 ; and (3) the solubility of the surface precipitate is lowered because the dielectric constant of the solution near the surface is less than that of the bulk solution. Surface precipitates can arise via polymeric metal complexes (dimers, trimers, etc.) that form on mineral surfaces, and from the sorption of aqueous polymers.

Using in situ molecular-scale analysis (X-ray absorption fine structure spectroscopy (XAFS)) it has been shown that multinuclear metal hydroxide complexes and surface precipitates of Co, Cr, Cu, Ni, Pb, and Zn can form on metal oxides, phyllosilicates, soil clays, and in soils (Fig. 5) (Siebecker et al., 2018; Scheidegger et al., 1996). These metal hydroxide phases occur at metal loadings below a theoretical monolayer coverage and in a pH range well below the pH where the formation of metal hydroxide precipitates would be expected according to the thermodynamic solubility product.

When a surface precipitate consists of chemical species derived from both the aqueous solution and dissolution of the mineral, it is referred to as a coprecipitate. The composition of the coprecipitate varies between that of the original solid and a pure precipitate of the sorbing metal. The ionic radius of the sorbing metal and sorbent ions must be similar for coprecipitates to form. Thus Co, Mn, Ni, and Zn, for example, form coprecipitates on sorbents containing Al and Si, but Pb does not form a coprecipitate because it is considerably larger (ionic radius = 0.12 nm) than Al and Si (0.054 and 0.04 nm, respectively). Coprecipitate formation is most limited by the rate of mineral dissolution that releases Si and Al ions (Scheidegger et al., 1996).

Sorption of Co, Ni, and Zn on phyllosilicates and aluminum oxides can result in the formation of mixed metal-Al hydroxide surface precipitates, which are coprecipitates (Siebecker et al., 2018). The precipitate phase shares structural features common to the hydrotalcite group of minerals and layered double hydroxides (LDH) minerals. The LDH structure comprises stacked sheets of edge-sharing metal octahedra containing divalent and trivalent metal ions separated by anions between the interlayer spaces. Reaction processes controlling Al release from minerals is an important aspect of LDH surface precipitation because the Al slowly diffuses into the octahedral layer of the surface precipitate. For example, the Al^{3+} partially replaces the Ni^{2+} in the octahedral sites of Ni-Al (LDH), which is thermodynamically favored over Ni hydroxide (Scheidegger et al., 1996).

The formation of metal hydroxide surface precipitates is an important sorption process for metal sequestration and immobilization. As the surface precipitates age, metal release is greatly reduced (Scheidegger and Sparks, 1996). Thus, the metals are less prone to leaching into ground and surface waters, and less available for uptake by plants, microbes, and animals.

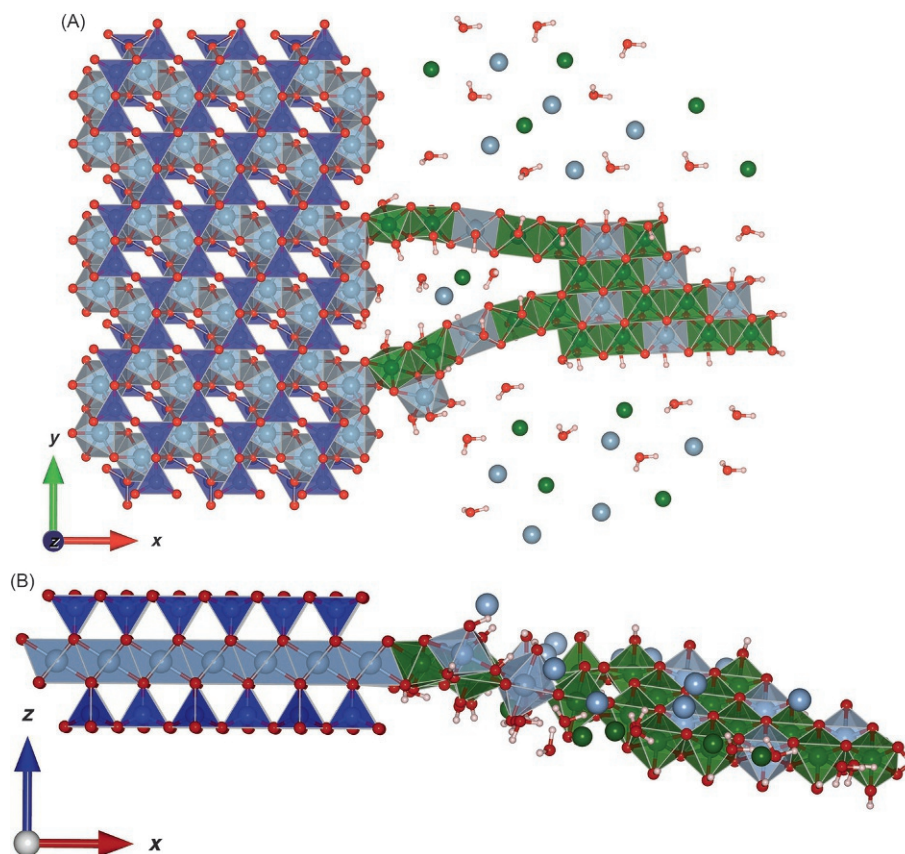


Fig. 5 Two views (top of (001) plane in panel (A) and side view in panel (B)) of layered double hydroxide mineral surface precipitates (green, blue, and red) on a clay mineral (blue and red). Reproduced from Siebecker MG, Li W and Sparks DL (2018) The important role of layered double hydroxides in soil chemical processes and remediation: What we have learned over the past 20 years. In: Sparks DL (ed.) *Advances in Agronomy*. San Diego: Elsevier Academic Press Inc., with permission.

Summary and outlook

Sorption of metals is an important process controlling their transport through the soil and availability to biological organisms. Metals react with soil minerals or soil organic matter through inner-sphere complexation, outer-sphere complexation, and surface precipitation. Continued research to understand sorption processes in soils needs to account for the highly heterogeneous sorbents in soils and how the complexity impacts metal sorption under dynamic conditions. This includes biological influences on sorption reactions (e.g., rhizosphere or biofilm processes) and competing reactions between different sorbent surfaces.

References

- Anderson SJ and Sposito G (1992) Proton surface-charge density in soils with structural and pH-dependent charge. *Soil Science Society of America Journal* 56: 1437–1443.
- Bolt GH and Van Riemsdijk WH (1979) Chapter 13: Ion adsorption on inorganic variable charge constituents. In: Bolt GH (ed.) *Developments in Soil Science*. Elsevier.
- Bowen HJM (1979) *Environmental Chemistry of the Elements*. London: Academic Press.
- Chorover J and Sposito G (1995) Surface-charge characteristics of kaolinitic tropical soils. *Geochimica et Cosmochimica Acta* 59: 875–884.
- Corey RB (1980) Sorption vs precipitation. In: Anderson MA and Rubin AJ (eds.) *Adsorption of Inorganics at Solid-Liquid Interfaces*. Ann Arbor, Michigan: Ann Arbor Science.
- Evangelou VP and Phillips RE (2005) Cation exchange in soils. In: *Chemical Processes in Soils*. Madison: SSSA.
- Forbes EA, Posner AM, and Quirk JP (1976) The specific adsorption of divalent Cd, Co, Cu, Pb, and Zn on goethite. *Journal of Soil Science* 27: 154–166.
- Goldberg S (1992) Use of surface complexation models in soil chemical-systems. In: *Advances in Agronomy*. Academic Press.
- Harter RD (1983) Effect of soil pH on adsorption of lead, copper, zinc, and nickel. *Soil Science Society of America Journal* 47: 47–51.
- Hiemstra T, Vanriemsdijk WH, and Bolt GH (1989) Multisite proton adsorption modeling at the solid-solution interface of (hydr)oxides—A new approach. 1. Model description and evaluation of intrinsic reaction constants. *Journal of Colloid and Interface Science* 133: 91–104.
- Kubicki JD, Paul KW, and Sparks DL (2008) Periodic density functional theory calculations of bulk and the (010) surface of goethite. *Geochemical Transactions* 9: 4.
- McKenzie R (1980) The adsorption of lead and other heavy metals on oxides of manganese and iron. *Soil Research* 18: 61–73.
- Sauve S, Hendershot W, and Allen HE (2000) Solid-solution partitioning of metals in contaminated soils: Dependence on pH, total metal burden, and organic matter. *Environmental Science & Technology* 34: 1125–1131.
- Scheidegger AM and Sparks DL (1996) Kinetics of the formation and the dissolution of nickel surface precipitates on pyrophyllite. *Chemical Geology* 132: 157–164.
- Scheidegger AM, Fendorf M, and Sparks DL (1996) Mechanisms of nickel sorption on pyrophyllite: Macroscopic and microscopic approaches. *Soil Science Society of America Journal* 60: 1763–1772.
- Selim HM and Sparks DL (2001) *Heavy Metals Release in Soils*. CRC Press.
- Siebecker MG, Li W, and Sparks DL (2018) The important role of layered double hydroxides in soil chemical processes and remediation: What we have learned over the past 20 years. In: Sparks DL (ed.) *Advances in Agronomy*. San Diego: Elsevier Academic Press Inc.
- Sparks DL (2002) *Environmental Soil Chemistry*. San Diego, CA: Academic Press.
- Sposito G (1981) Cation exchange in soils: A historical and theoretical perspective. In: *Chemistry in the Soil Environment*.
- Sposito G (1987) Distinguishing adsorption from surface precipitation. In: *Geochemical Processes at Mineral Surfaces*. American Chemical Society.
- Strawn DG (2021) Sorption mechanisms of chemicals in soils. *Soil Systems* 5: 13.
- Strawn DG and Sparks DL (1999) The use of XAFS to distinguish between inner- and outer-sphere lead adsorption complexes on montmorillonite. *Journal of Colloid and Interface Science* 216: 257–269.
- Strawn D, Bohn HL, and O'Connor GA (2020) *Soil Chemistry*. Hoboken, NJ: John Wiley & Sons.
- Sumner ME (1998) Soil chemistry: Past, present and future. In: Huang PM, Sparks DL, and Boyd SA (eds.) *Future Prospects for Soil Chemistry*. Madison, Wisconsin: Soil Science Society of America.
- Tournassat C, Davis JA, Chiaberge C, Grangeon S, and Bourg IC (2016) Modeling the acid–base properties of montmorillonite edge surfaces. *Environmental Science & Technology* 50: 13436–13445.
- Uehara G and Gillman GP (1980) Charge characteristics of soils with variable and permanent charge minerals. 1. Theory. *Soil Science Society of America Journal* 44: 250–252.
- Xu S and Harsh JB (1990) Monovalent cation selectivity quantitatively modeled according to hard/soft acid/base theory. *Soil Science Society of America Journal* 54: 357–363.

Sorption–desorption kinetics

Gautier Landrot^a and Donald L Sparks^b, ^aSynchrotron SOLEIL, L'Orme des Merisiers, Saint-Aubin, France; ^bDepartment of Plant and Soil Sciences, University of Delaware, Interdisciplinary Science and Engineering (ISE) Laboratory, Newark, DE, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	343
Experimental approaches to measure kinetics relevant to soil systems	344
Sorption behavior in soils	347
Desorption behavior in soils	348
Application of sorption-desorption kinetics to conceptual models	350
Conclusions and future directions	353
References	354

Abstract

Sorption–desorption kinetics have a major impact on the fate of chemicals in soils. Analytical advances enable one to determine better elementary reactions occurring at the molecular level that are part of the overall reaction mechanism, and the extent of multiple sorption processes taking place simultaneously on soil surfaces. Thus, compared to rate parameters previously reported in the literature, those measured with new approaches may be less apparent and more applicable to predict accurately the fate of chemicals in soils using conceptual models.

Key points

- Sorption may result in multiple processes simultaneously occurring on soil component surfaces even after apparent equilibrium is reached in the soil solution.
- The amount of contact time can dramatically affect the extent of desorption
- Probing the chemical forms of adsorbates helps constrain the overall sorption–desorption mechanism.
- Sorption–desorption kinetics are used in mechanistic models to elucidate reaction pathways.
- A predictive model should account for all chemical and transport processes that are involved for a given sorption–desorption mechanism.
- Rate parameters corresponding to chemical kinetics may be employed as input parameters in models to predict chemical fate.
- Chemical kinetics corresponding to elementary reactions may be measured using XAFS combined with PCA and MCR-ALS data-treatment procedures

Introduction

Sorption is a general term that describes the retention of a chemical species by a solid component of a soil. This retention can occur at the liquid–solid interface or within the solid component. The retention mechanism can be adsorption, surface precipitation, or polymerization. Adsorption can refer to an ion exchange mechanism in the diffuse layer as a result of electrostatic forces (nonspecific sorption) or surface complexation between a chemical species and a specific sorption site (specific sorption). Desorption refers to the inverse process, i.e., the release of the chemical species from the surface to the soil solution. Sorption–desorption reactions largely control the fate of chemicals in soils, including plant nutrients, heavy metals, and organic contaminants (Goldberg et al., 2007). These reactions occur in soils over a wide time scale, from milliseconds to days or even years (Sparks, 2003). Since soils are complex and dynamic, a chemical species present in soil solution is likely to be involved in a continuum of multiple reversible and irreversible biogeochemical reactions, including sorption and desorption processes. Among these multiple reactions, those that are irreversible, such as the transformation of sorbed species into secondary phases that remain permanently in a soil, may occur on timescales longer than the longest time period typically studied in the laboratory (i.e. usually less than a year). Those that are reversible hardly reach equilibrium. For example, transport of the soil solution within the soil porous network can specifically affect sorption–desorption reactions as it locally modifies the chemical composition of the soil solution next to sorption sites. This includes the nature and amounts of dissolved species that can potentially compete with sorbed species for sorption sites. Therefore, a reversible sorption or desorption reaction may never reach local equilibrium if it occurs at a time scale longer than the hydraulic residence time of the soil solution (Zhang and Selim, 2011). It is possible that equilibrium is not reached even when the concentration of a dissolved chemical species, experimentally measured in soil solution at different time intervals, does not vary with time. This case can correspond to a steady-state that is out of equilibrium, where the difference between

the total amounts of the chemical appearing in and disappearing from soil solution reaches a constant value (Lasaga, 1998). To quantify these two opposite chemical fluxes occurring in the soil solution at steady-state, including their fractions that correspond to sorption or desorption processes, employing a kinetic approach can be more suitable than a thermodynamic one. Indeed, the former enables one to infer chemical concentration at any time before reaction equilibrium is reached, based on rate parameters that are experimentally determined. The latter enables one to determine chemical activities based on an equilibrium constant that are valid only when true equilibrium is reached. Therefore, the topic of sorption–desorption kinetics is crucial in soil science as it represents a key factor in predicting content and speciation of chemicals present in soils. However, it still represents a field needing additional research, despite the plethora of studies that have been conducted on this topic over recent decades. Most experimental approaches available until now to measure reaction kinetics in soil systems have inherent limitations, which are described in the first part of this chapter. Some emerging kinetic approaches are also discussed in the first part. Typical kinetic behavior of sorption and desorption processes reported in the literature are presented in the second and third parts of this chapter, respectively. It will be demonstrated that the characterization of the solid phases, notably via spectroscopic tools, can help to constrain better the mechanisms involved during sorption or desorption processes. Some models recently developed, where sorption–desorption kinetics were employed to identify the nature of chemical reaction pathways, will be presented in the last part of this chapter. Some of the remaining challenges to apply sorption–desorption kinetics successfully in conceptual models to predict chemical fate in soils accurately will be also discussed in the last part of this chapter.

Experimental approaches to measure kinetics relevant to soil systems

The kinetics of reactions relevant to soil systems, including sorption–desorption reactions, have been studied using various experimental approaches. These approaches could differ from each other by the type of method employed (e.g. batch, stirred-flow, column) or nature of the solid material investigated (i.e. a well-mixed suspension of an individual soil component, or whole soil studied as a mixed suspension or saturated static soil). Details of the apparatuses that have been used to measure kinetics in soil systems can be found elsewhere (Sparks, 1996). This first part of the review mainly focuses on how the rates can depend on the chosen experimental approach employed to measure them. This aspect must be understood to fully comprehend the typical kinetic behaviors of sorption–desorption reactions that have been reported in the literature, which are presented in the next parts of this chapter.

Traditional batch, stirred-flow, and column methods are employed in approaches A, B, C described in Fig. 1, respectively. All transport processes that a chemical species may undergo between the bulk soil solution and a sorption site, specifically when approach A, B, or C is employed, are schematically represented in Fig. 1. Multiple transport processes may occur when a chemical interacts with an individual soil component (Fig. 1, approach A). This includes: (1) transport in the solution phase, which is rapid, and in the laboratory, can be eliminated by rapid mixing; (2) transport across a liquid film at the particle–liquid interface (film diffusion); (3) transport in liquid-filled pores larger than 2 nm; (4) diffusion of a sorbate along pore wall surfaces (surface diffusion); (5) diffusion of sorbate occluded in pores smaller than 2 nm (pore diffusion); and (6) diffusion processes in the bulk of the solid (Sparks, 2003). Soil particles comprise multiple individual soil components that are mixed together as aggregates. In a soil aggregate, transport processes only occur in the region of each individual soil component that is exposed to the soil solution (Fig. 1, approach B). Therefore, an individual soil component inside a soil aggregate may be partially or not at all exposed to the soil solution. In a static saturated soil (e.g. a saturated soil packed in a column, as opposed to a well-mixed soil suspension in a batch or stirred-flow reactor), transport processes can occur between solid particles in the empty spaces of the soil that are parts of the connected pore network (Fig. 1, approach C). Additionally, transport processes can occur inside solid particles of the soil, in regions that are in contact with the connected pore network. Therefore, a soil aggregate inside a static saturated soil may be partially or not at all exposed to the soil solution.

A sorption process leads to the formation of an adsorbate, which is a chemical species retained at the surface of a solid component. The solid surface where sorption–desorption takes place is referred as the adsorbent. The chemical species sorbing to and desorbing from the adsorbent are the adsorptive and desorptive, respectively (Sparks, 2003). The amounts of adsorbate appearing or disappearing at the location where the sorption or desorption process takes place are not determined when a traditional batch, stirred-flow, or column method is employed. Instead, when any of these methods is used, the rates are measured by quantifying the variation in concentration of adsorptive or desorptive in the bulk soil solution. Consequently, the overall rates observed using a batch, stirred-flow, or column method can be apparent. They can include, in addition to the chemical rates (CR) corresponding to the actual sorption or desorption mechanism, those corresponding to all transport processes that occur specifically between sorption sites and the bulk solution when approach A, B, or C is employed (Fig. 1). When multiple processes occur successively or in parallel, the process that has the slowest or fastest rate is considered as the rate-determining step of the overall kinetics, respectively (Lasaga, 1998). Therefore, the kinetics determined using approach A, B, and C described in Fig. 1 may not be representative of the CR if the latter does not represent the rate-determining step of the process observed experimentally. Additionally, none of approaches A, B, and C is ideal to determine how rapidly a given sorption or desorption process, involving a chemical and a specific adsorbent, truly occurs in a saturated soil. With approach A, only the adsorbent being studied is present in the system. Therefore, the possibility that the adsorbent inside a soil aggregate may be only partially or not all available to soil solution, depending on the physical organization of the aggregate, is not considered. Similarly, approaches A and B do not consider that the soil aggregate itself, which contains the specific adsorbent, may be partially or not all available to the solution in the soil,

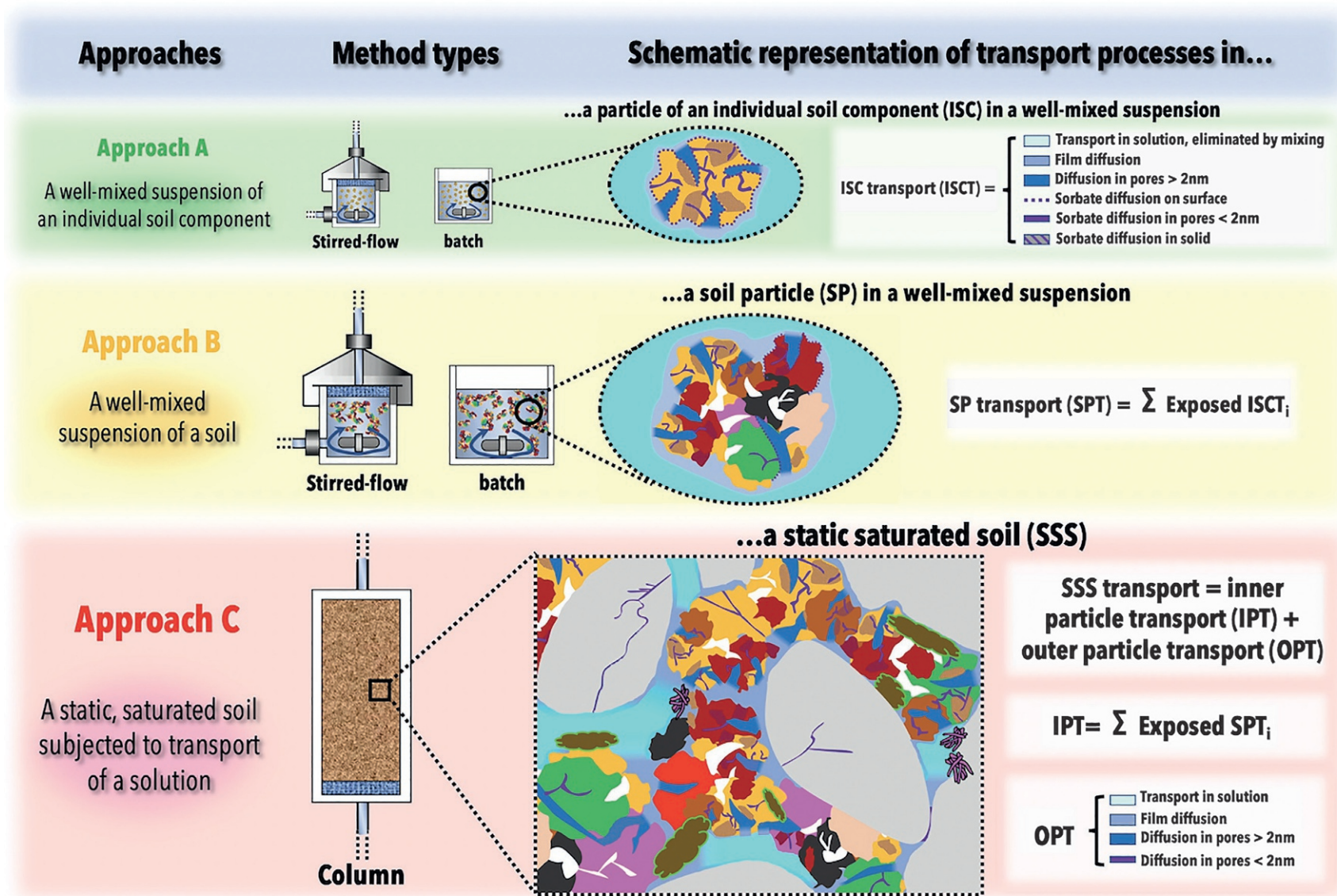


Fig. 1 Examples of approaches that can be employed to measure kinetics of sorption-desorption reactions relevant to soil systems, and schematic representation of all transport processes occurring in the system when each of these approaches is used. The overall transport occurring in a soil particle (SPT) composed of aggregated individual soil components is equal to the sum of all transport processes occurring in the parts of each individual soil component that are exposed to the soil solution. Similarly, the inner particle transport (IPT) occurring in a static saturated soil is equal to the sum of all transport processes occurring in the parts of each soil particle that are exposed to the soil solution.

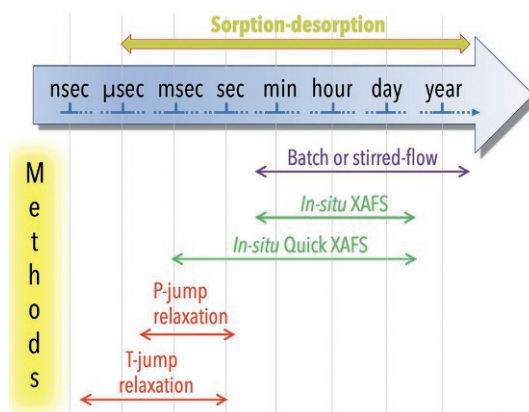


Fig. 2 Time-scales of sorption–desorption reactions in soils and various kinetic methods.

depending on the proximity of the aggregate to the connected pore network. In contrast, all transport processes that may affect the overall sorption–desorption kinetics in a static saturated soil can be considered using approach C. However, an adsorptive or desorptive is likely to interfere simultaneously with multiple adsorbents in a soil. Therefore, approaches B and C cannot be employed to determine the kinetics of a sorption–desorption reaction involving a specific adsorbent and occurring in a saturated soil if the rates are measured by quantifying the concentration of the adsorptive or desorptive in the soil solution.

Some techniques, such as P- or T-jump relaxation methods, allow the rates to be determined by varying a physical property of the system, e.g. temperature or pressure. With these techniques, rate constants are determined indirectly as they are measured from linearized rate equations that include parameters determined from equilibrium and modeling studies (Sparks, 2015). The application of an analytical technique, such as batch, stirred-flow, XAFS (X-ray absorption fine structure spectroscopy), and relaxation methods, is limited to a specific time scale (Fig. 2). Batch and stirred-flow techniques do not allow rapid kinetics to be observed as they require at least half a minute to record the first data point.

In contrast, P- and T-jump relaxation techniques can only measure rapid kinetics that are shorter than one minute. The maximum time length covered by XAFS techniques cannot exceed the total allocated beam time, which is typically a few days or less. The minimum time length corresponds to the time required to complete one XAFS scan, which is more than a millisecond and second for Quick and regular XAFS techniques, respectively.

Batch and stirred flow techniques have been so far the most commonly used methods to measure sorption–desorption kinetics in soil systems (approach A or B in Fig. 1, to study a specific soil component or whole soil, respectively). Interpretation of the apparent rates constrained with batch or stirred-flow techniques can be less ambiguous using complementary data, notably the ex-situ characterization of the solid phase, as demonstrated in the next two parts of this chapter. Another powerful approach combines a column, batch, or stirred-flow method with an analytical tool that allows one to determine in-situ the nature and quantity of all main forms of a chemical species, including the forms that are possibly sorbed. For example, Quick X-ray absorption fine structure (Quick-XAFS) spectroscopy was employed with a flow cell (Siebecker et al., 2014) or batch method (Landrot et al., 2010; Vantelon et al., 2019) to follow in-situ the kinetics of rapid environmentally-relevant reactions occurring in soils. This approach allowed one to determine in-situ the speciation of all principal chemical forms of an element, regardless of the fact that these forms were dissolved or sorbed species. For example, Quick-XAFS was employed to measure in-situ the kinetics of nickel that simultaneously adsorbed and precipitated into a layer double hydroxide (LDH) on the surface of a clay, which was mixed with glass beads inside a flow cell (Siebecker et al., 2014). Additionally, XAFS data can be analyzed by a principal component analysis (PCA) followed by a multivariate curve alternative least-square fitting (MCR-ALS) method. This unique data analysis procedure can be done in Fastosh, which is a new XAFS data treatment freeware (Landrot, 2018). This approach allows spectra corresponding to the principal components of an element and the relative quantities of these components to be obtained. This can be notably valuable in identifying reaction intermediates that form during a chemical reaction. For example, PCA & MCR-ALS methods were applied to Quick-XAFS spectra collected at the Fe K-edge during an in-situ experiment where the formation of Fe oxide–organic matter aggregates was followed in a batch reactor under N_2 (Vantelon et al., 2019). This chemometric method provided quantification of the relative amounts (Fig. 3A) and corresponding XAFS spectra (Fig. 3B) of all principal components of Fe that were present throughout the kinetic experiment. These components included an intermediate species that formed temporally during the experiment (Fig. 3A). The results then suggested that the overall mechanism featured two elementary reactions, where the intermediate species was the product and the reactant of the first and second elementary reactions, respectively. Therefore, combining Q-XAFS and the PCA & MCR-ALS procedure is promising as it can potentially single out elementary reactions occurring at the molecular level that are part of the overall reaction mechanism. Measuring rate parameters corresponding to an elementary reaction instead of a group of combined elementary reactions can be beneficial for developing conceptual models that can predict chemical fate in environmental systems, as discussed in the last part of this chapter.

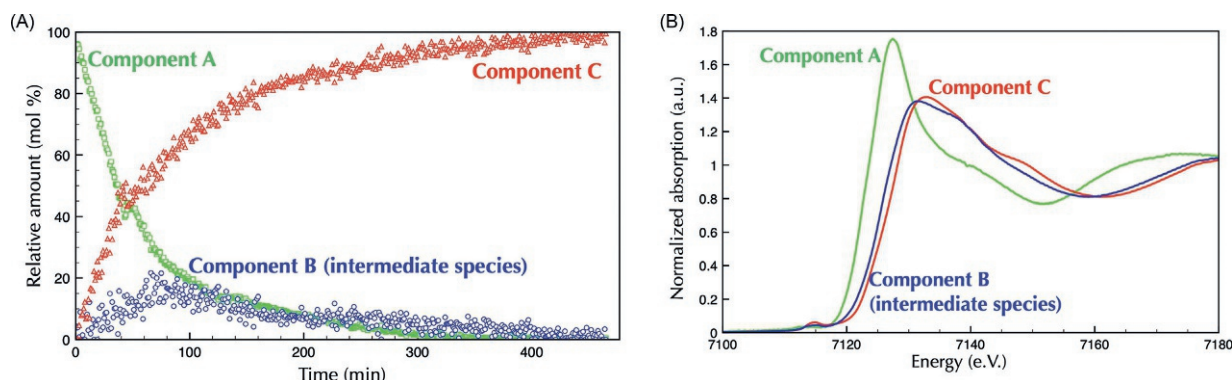


Fig. 3 (A) Relative amount of the three principal chemical forms of iron (Fe) present in a batch reactor and (B) their corresponding pure spectra determined using a PCA & MCR-ALS method applied to XAFS spectra at the Fe K-edge, collected during an in-situ experiment. Shell-by-shell analyses of the pure spectra (not shown), subsequent to their extraction by MCR-ALS, suggested that components A, B, and C corresponded to Fe(II)aqueous, Fe(II) sorbed to humic acids, and Fe–organic matter nano aggregates, respectively. Reproduced with permission from Vantelon D, Davranche M, Marsac R, Fontaine C, Guenet H, Jestin J, Campaore G, Beauvois A, and Briois V (2019) Iron speciation in iron-organic matter nanoaggregates: A kinetic approach coupling Quick-EXAFS and MCR-ALS chemometry. *Environmental Science: Nano*, 6.

Sorption behavior in soils

The kinetics of heavy-metal sorption, determined by a traditional batch or stirred-flow method, on an individual soil component (approach A, Fig. 1) or a soil (approach B, Fig. 1), is typically characterized by a biphasic process in which sorption is initially rapid followed by slow reactions. The rapid step, which occurs over milliseconds to hours, can be ascribed to CR and film diffusion processes. During this rapid reaction process, a large portion of the sorption may occur. For example, a study demonstrated that Ni could sorb rapidly and precipitate into an LDH phase on the surface of the clay in less than 40 min, based on in-situ analyses of the solid–water interface using Quick-XAFS (Siebecker et al., 2014). For Ni sorption on a soil, a study showed that a large portion of total Ni could sorb during the first 24 h and continue to sorb for at least 500 h (Fig. 4). Post-mortem XAFS analyses of the solid phase revealed that Ni was mostly adsorbed onto soil surfaces during the first 24 h. In contrast, it was mostly present as LDH at 650 h (Fig. 4). Similarly, a study demonstrated that molybdenum (Mo) sorption on three different types of soils followed a biphasic behavior. In all three soils, Mo sorption was fastest during the first 8 h, and was mostly associated with the exchangeable fraction. The Mo sorption on each soil continued for 1 year, but at a slower pace than that of the first 8 h. After 1 year, Mo was mostly associated with Fe/Al oxides in two soils. An irreversible sorption site had to be considered to model Mo retention successfully over time in these soils (Sun and Selim, 2019). The biphasic sorption behavior can occur regardless of the affinity of a metal to associate with an absorbent and type of sorption mechanism involved. For example, it was shown that selenate sorption on aluminum hydroxide or goethite was fastest during the first 24 h, although the extent of selenate sorption was significantly higher on aluminum hydroxide than on goethite. Additionally, an inner-sphere and outer-sphere could form on the former and latter absorbent, respectively (Loffredo et al., 2011).

Sorption by metals and oxyanions is characterized by a rapid and reversible stage followed by a much slower non-reversible process or biophasic kinetics. Many investigations have shown that organic contaminants behave similarly. The rapid phase has been ascribed to retention of the organic chemical in a labile form that is easily desorbed. The labile form of the chemical is also available for microbial attack. However, the much slower reaction phase involves the entrapment of the chemical in a nonlabile form that is difficult to desorb and is resistant to biodegradation. This slower sorption–desorption reaction has been ascribed to diffusion of the chemical into micropores of the organic matter and inorganic soil components (Sparks, 2003). The slow diffusion of organic chemicals in soil organic matter (SOM) can be explained by considering SOM as a combination of ‘rubbery’ and ‘glassy’ polymers. The rubbery-like phases are characterized by an expanded, flexible, and highly solvated structure with pores of subnanometer dimensions (holes). Sorption in the rubbery phase results in linear, noncompetitive, and reversible behavior. The glassy phases have pores that are of subnanometer size and sorption in this phase is characterized by nonlinearity and is competitive (Sparks, 2003).

The kinetics of a sorption reaction involving an organic chemical found ubiquitously in the environment is shown as an example in Fig. 5 (Yan et al., 2014). In this study, the adsorptive and adsorbent were an inositol hexakisphosphate (IHP), which represents the most abundant type of organic phosphate (OP) in soils and sediments, and amorphous aluminum hydroxide (AAH), respectively. The [IHP] in solution progressively decreased during the first 6 h and then remained stable after 6 h (Fig. 5A). Between 3 h and 4 days, a fraction of IHP was retained as an inner sphere complex on the surface of AAH, based on the presence of a peak at 6.5 ppm in all solid-state ^{31}P single-pulse magic angle spinning nuclear magnetic resonance spectroscopic (SP/MAS NMR) spectra, which corresponded to AAH reacted with IHP for different time periods (Fig. 5B). Additionally, a fraction of IHP sorbed to AAH progressively transformed, between 6 h and 2 days, into a surface precipitate analogous to aluminum phytate. This was based on attenuated total reflectance Fourier transform infrared spectroscopy (ATR-FTIR) analyses of AAH reacted with IHP for different time periods (Fig. 5C). Therefore, sorption of an organic species can result in more than one sorption mechanism, like the example

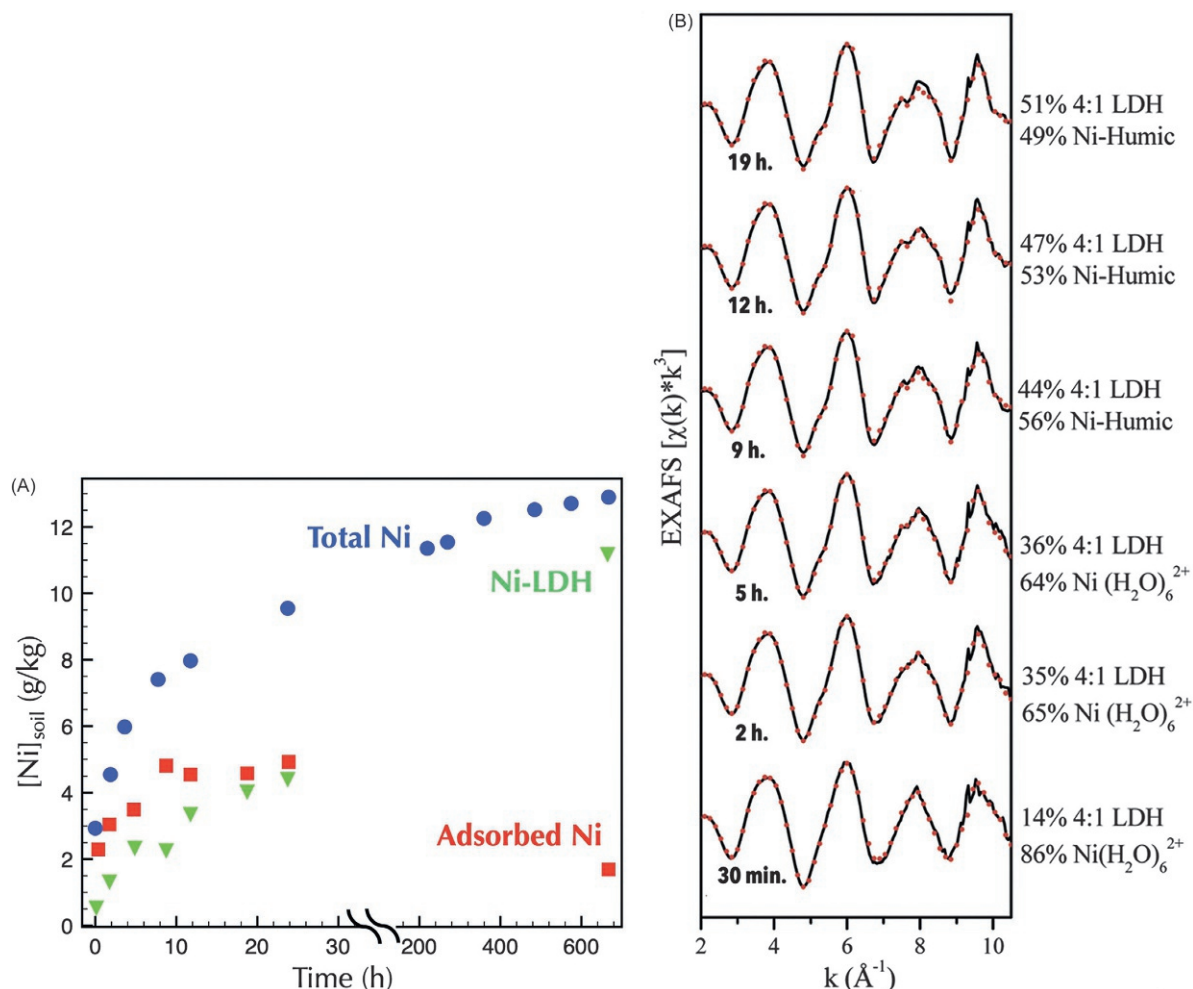


Fig. 4 (A) Kinetics of Ni sorption onto a soil determined using a batch method (blue dots) and amounts of Ni-LDH and adsorbed Ni measured using post-mortem XAFS analyses of the Ni-reacted soil's solid phase at different time intervals in green and red colors, respectively; (B) linear combination fitting of the EXAFS corresponding to the Ni-reacted soil's solid phase at different time intervals. Reproduced with permission from Shi Z, Peltier E, and Sparks DL (2012) Kinetics of Ni Sorption in Soils: Roles of Soil Organic Matter and Ni Precipitation. *Environmental Science & Technology* 46: 2212–2219.

involving the sorption of an inorganic species shown in Fig. 4. In the case of IHP sorption to AAH, the first sorption mechanism was rapid and occurred within the first 3 h, the second one was slower and started at 6 h although the concentration of IHP in solution did not vary further. This suggested that determining the sorption kinetics by measuring only the concentration of adsorptive in solution can be misleading as sorption processes can continue after the concentration of the adsorptive seems to have reached apparent equilibrium. In both examples shown in Figs. 4 and 5, analyses of the reacted solid phase using spectroscopic tools, i.e. XAFS, ATR-FTIR, and SP/MAS NMR, allowed one to identify more than one sorption process occurring on the surface and to understand the overall sorption behavior better.

Desorption behavior in soils

The amount of contact time between metals and soil sorbents (residence time) can dramatically affect the degree of desorption, depending on the metal and sorbent. This behavior is due to aging mechanisms that can occur when the metal is associated with the sorbent, such as surface precipitation or diffusion into the solid phase. For example, it was mentioned above that the amount of Ni adsorbed in a soil (Fig. 3) or on a clay surface (Siebecker et al., 2014) could progressively precipitate into an LDH phase. This transformation can affect metal desorption kinetics. For example, it was shown that Zn sorption to soils excavated from Brazilian mining areas could result in the formation of Zn insoluble secondary phases, including Zn-LDH, which could lead to a significant decrease in the exchangeable/mobile fraction of Zn in the soils. The desorption kinetics of Zn from the soils where these secondary phases formed was then much slower than from the mine tailings where Zn was mostly present in the exchangeable/mobile fraction (Lopes et al., 2021).

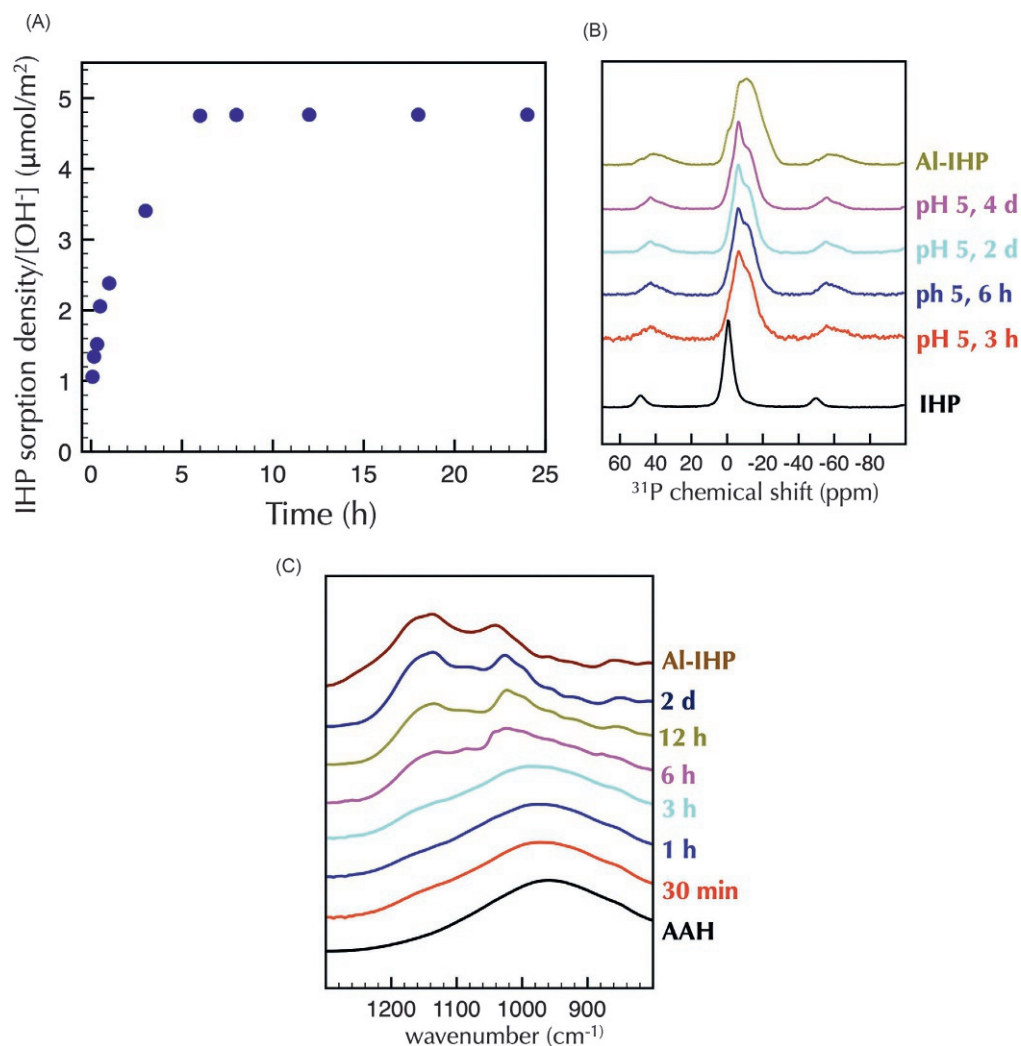


Fig. 5 (A) Kinetics of inositol hexakisphosphate (IHP) sorption onto amorphous aluminum hydroxide (AAH) normalized by $[\text{OH}^-]$ released to soil solution, (B) solid-state ^{31}P SP/MAS NMR spectra of IHP sorbed on AAH in 0.1 M KCl at pH 5 as a function of time, and (C) ATR-FTIR spectra of AAH in the presence of 1305 μM IHP in 0.1 M KCl at pH 5 as a function of time. Reproduced with permission from Yan Y, Li W, Yang J, Zheng A, Liu F, Feng X, and Sparks D (2014) Mechanism of myo-inositol hexakisphosphate sorption on amorphous aluminum hydroxide: Spectroscopic evidence for rapid surface precipitation. *Environmental Science & Technology* 48.

It was demonstrated recently that the retention of antimony (Sb) in soils can increase over time, by a factor of 2–4 between about 2 weeks and 6 months (Verbeeck et al., 2021). Results obtained from XAFS analyses suggested that this increased retention was not due to precipitation as with Ni, but rather micropore diffusion with Sb substitution in surface Fe octahedra of ferrihydrite, and/or changes in surface adsorption geometry (Verbeeck et al., 2021). Another recent study demonstrated that the amount of Pb that could be released from iron oxide decreased with increasing length of aging period, until an aging period of 50 h was reached (Fig. 6A). With an aging period >50 h, the amounts of Pb released were constant. Chemical mapping at the microscopic scale indicated that much more Pb was present inside a Pb-reacted hematite particle aged for 168 h compared to another particle aged for 24 h (Fig. 6B). This suggested that Pb progressively diffused inside the solid particle, which made it less likely to be released to solution (Lu et al., 2020).

It has been demonstrated that as for some metals, with increased residence time the nonlabile portion of the organic chemical in the soil or sediment becomes more resistant to release and is less available for microbial degradation. Soils reacted with desorption agents in the laboratory over a short period of time often show more rapid organic chemical release than field soils that have been contaminated for many years. The difference in release has been ascribed to greater diffusion into micropores of clay minerals and humic components that occurred over longer times in field settings (Sparks, 2003). Some studies, such as Cheng et al. (2019), have reported that the aging effect could also occur over a time period of less than a year. For example, Fig. 7 shows the desorption kinetics of Dechlorane Plus (DP), which is a toxic additive flame retardant and a common soil pollutant, from soils that were aged

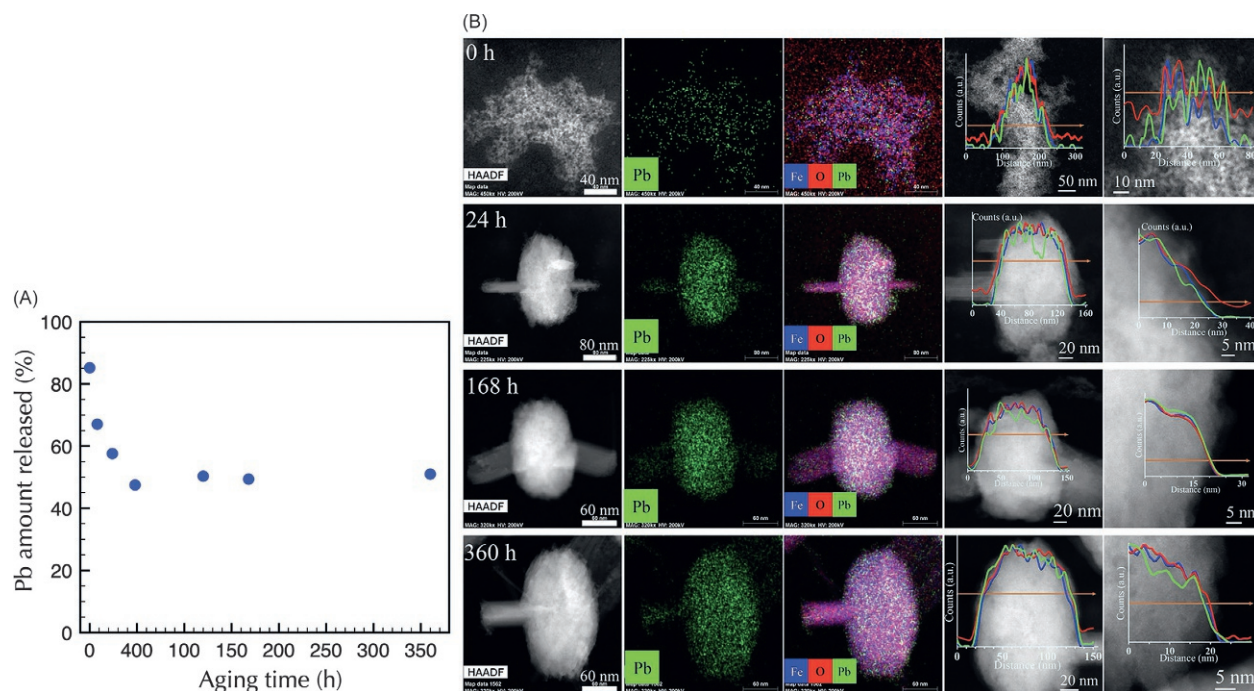


Fig. 6 (A) Amounts of Pb released from Pb-reacted iron oxide particles aged for different time periods, using an acid solution at pH 3.0; (B) STEM-EDS mapping and line scan results of elemental distribution (Pb, Fe, and O) on Pb-reacted iron oxide samples aged for various time periods. Blue, red, and green colors refer to Fe, O, and Pb, respectively. Reproduced with permission from Lu Y, Hu S, Liang Z, Zhu M, Wang Z, Wang X, Liang Y, Dang Z, and Shi Z (2020) Incorporation of Pb(II) into hematite during ferrihydrite transformation. *Environmental Science: Nano* 7: 829–841.

for various time periods of 0–260 days. The desorption kinetics systematically followed a biphasic behavior, regardless of the aging effect, as they were rapid during the first 50 h, and then slowly decreased from 50 to 400 h (Fig. 7A). However, the total amount of DP that could be released to solution during these desorption experiments decreased significantly with increasing aging time (Fig. 7A). Results obtained from a new sequential chemical extraction procedure indicated that the amounts corresponding to the labile DP fraction present in the soils decreased with an increasing aging period, while the amounts corresponding to the residual fraction of DP increased with an increasing aging period (Fig. 7B). Therefore, as for some metals in soils, this organic pollutant that adhered to soil components could progressively transform into secondary forms that were more environmentally persistent.

Application of sorption–desorption kinetics to conceptual models

Many studies have used theoretical rate laws to reproduce mathematically the sorption–desorption kinetics measured experimentally. Details of these rate laws are available elsewhere (Selim and Zhang, 2013). Applying rate laws to kinetic data can be potentially useful in achieving two goals. First, it may allow us to identify reaction pathways and infer an overall reaction mechanism involving a chemical present in a soil, thus improving a fundamental understanding of the behavior of a chemical in the soil. Second, it can help to develop conceptual models that can accurately predict the fate of a chemical in environmental systems other than those where the rates have been determined experimentally. Unfortunately, examples of studies where either of these two goals was unambiguously achieved are still rare. In practice, the characterization of an overall reaction mechanism, describing all elementary reactions involved in the reaction, is often not possible when rate laws are applied to kinetic data measured using the methods currently available. The inherent limitations of most kinetic approaches discussed in the first part of this chapter also suggest that the rate parameters derived from rate laws can be apparent, as they may be affected by transport phenomena, and thus cannot be easily implemented in predictive models. These limitations will be described further in the last part of the chapter. This part will also describe some recent conceptual models where sorption–desorption kinetics were characterized to constrain reaction pathways as well as some of the remaining challenges to predicting chemical fate in soil systems successfully, including chemical retention behavior.

Sorption–desorption kinetics may depend on multiple biological, physical, chemical phenomena (Sparks, 2003). Some rate laws are purely empirical as they do not explicitly consider any of these phenomena in their mathematical expression. Consequently, the parameters derived from these rate laws cannot be easily applied to systems other than the ones for which they are determined. This includes the Elovich equation, power function, as well as first or second-order rate laws. Some rate laws can feature some parameters in their mathematical expression that account for chemical or physical aspects of the system. This includes

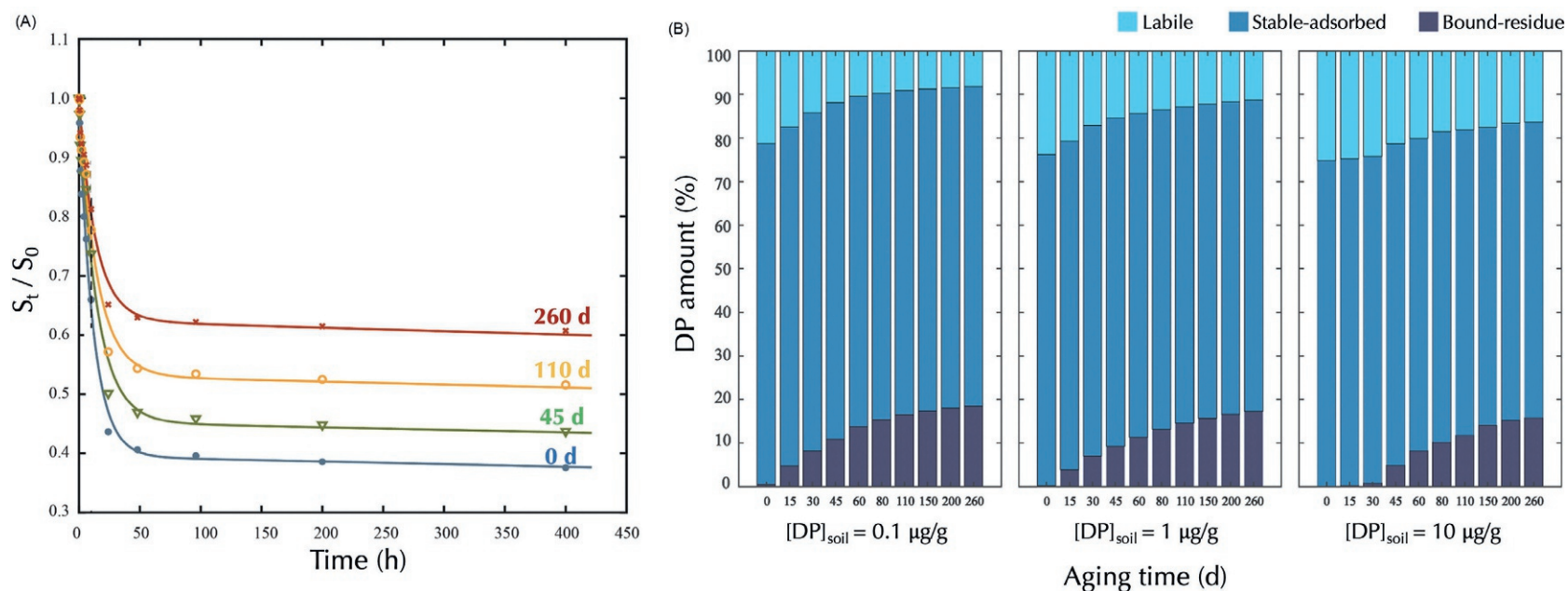


Fig. 7 (A) Desorption kinetics of Dechlorane Plus (DP) released from soils reacted with DP for specific aging time periods; and (B) labile, stable-adsorbed, and bound-residue fractions determined using a sequential extraction procedure in soils spiked with 0.1, 1, or 10 $\mu\text{g g}^{-1}$ DP. Reproduced with permission from Cheng Y, Ding J, Xie X, Ji X, and Zhang Y (2019) Validation and application of a 3-step sequential extraction method to investigate the fraction transformation of organic pollutants in aging soils: A case study of dechlorane plus. *Environmental Science & Technology* 53: 1325–1333.

chemical nonequilibrium rate laws based on a two-site model, where two reactions are assumed to reproduce the observed biphasic behavior: one fast reaction that quickly reaches equilibrium and a slower one that lasts for a long period of time (Sparks, 2003). Similarly, when a physical nonequilibrium rate law based on a two-site model is used, it is considered that the slow reaction is due to a transport rate-limiting step occurring during the reaction. A stop–flow approach can be employed together with a stirred-flow method to determine whether a process is chemically or physically rate-limited (Sparks, 1996). Alternatively, the energy of activation (E_a), whose value can be measured experimentally by employing the Arrhenius equation, can be used to determine whether a specific elementary reaction, or a set of combined elementary reactions (in this case, E_a is apparent) is chemically or physically rate-limited (Lasaga, 1998). Therefore, the stop flow approach, or E_a , can help determine whether a chemical or physical non-equilibrium rate law should preferentially be used for a specific system. However, the applicability of a given chemical or physical non-equilibrium rate law to multiple soil systems can be limited, as for pure empirical rate laws. Multiple types of sorption sites can be considered in a chemical non-equilibrium rate law. However, the nature of the chemical reactions occurring at these sites is not identified. Physical non-equilibrium rate laws often rely on multiple parameters that are measured empirically in the laboratory (Sparks, 2003).

A theoretical model can be considered mechanistic when it describes all types of retention processes affecting the rates that are observed experimentally. Three mechanistic models, developed over the last decade to reproduce metal sorption–desorption kinetics in soils, are summarized in Table 1. In these models, retention processes could correspond to a specific sorption mechanism such as Ni surface precipitation forming an LDH, or the association between a metal and a major soil sorption pool, e.g. clays, SOM, dissolved organic matter (DOM), which were assumed to reach equilibrium instantaneously. The sorption–desorption rate coefficients corresponding to these processes were inferred by equilibrium partition coefficients, which were calculated using the Windermere Humic Aqueous Model (WHAM) or the Charge Distribution-Multisite Complexation (CD-MUSIC) model. The former can describe the speciation of metals and metalloids associated with soil organic matter (Tipping, 1994). The latter is a surface complexation model that is particularly efficient in predicting metal complexation with some iron oxides (Gustafsson et al., 2011). The kinetics of Ni surface precipitation to LDH was mathematically reproduced using a simple rate law and was not assumed to reach equilibrium rapidly (Shi et al., 2012). However, the rate order of this rate law was empirically set to 0.77 so that the LDH amounts predicted by the model and those determined in-situ by XAFS analyses could match each other. All three kinetic models described in Table 1 could reproduce all retention behaviors well that were measured experimentally using a stirred-flow method. This suggested that the overall retention mechanisms assumed in these models could be valid. However, in a saturated static soil subjected to reactive transport (e.g. a static soil under field conditions, not a mixed soil suspension in the laboratory), the assumption that sorption–desorption processes reach equilibrium may be challenged when these reactions occur at a time scale longer than the hydraulic residence time of the soil solution (Molins and Knabner, 2019).

Some rate constants, corresponding to sorption or desorption processes considered in geochemical models, such as those described in Table 1, were empirically determined by employing a least-squares fitting method. The values of these rate constants were floated freely during the fitting procedure to minimize the difference between the experimental and model results. For example, all rate constants considered in the competitive multi-reaction models employed in Sun and Selim (2019) were empirically determined by a least-square fitting procedure. This was also the case for the rate constants corresponding to all desorption processes considered in the mechanistic models developed in Shi et al. (2012), Shi et al. (2013), and Peng et al. (2018). Therefore, these models, which can help in elucidating retention mechanisms, cannot be employed as purely predictive tools to estimate sorption–desorption kinetics in soil systems. Predictive models should enable these rates to be determined theoretically without the need to measure them in the laboratory, and the model parameters to be scaled based on the experimental data. Moreover, tools that can predict chemical retention behavior in specific regions of a soil cannot be models that consider only the overall geochemistry of the system. They must be essentially reactive transport models that can describe locally the coupled geochemistry, hydrology, biogeochemistry, soil physics, and fluid dynamic processes, which may affect chemical retention behavior simultaneously in different parts of a soil (Steefel et al., 2005). A description and comparison of various reactive transport models currently available can be found in Steefel et al. (2015). Lastly, multiple transport processes within soil particles can occur when a stirred-flow method is employed, even when the soil suspension is well stirred (approach B in Fig. 1). The effects of such

Table 1 Examples of mechanistic models where metal sorption–desorption kinetics were employed to constrain reaction pathways involving various metals.

Study	Adsorptive/desorptive	Predicted kinetics in soil samples	Method	Process	Specificity of rate formulation
Shi et al. (2012)	Ni	Sorption	Stirred-flow	Association with SOM Association with clays Ni surface precipitation to LDH	Equilibrium-based (WHAM VI) Equilibrium-based (WHAM VI) Rate order empirically defined
Shi et al. (2013)	Cu, Pb, Zn	Adsorption & desorption	Stirred-flow	Association with SOM Association with DOM	Equilibrium-based (WHAM VI) Equilibrium-based (WHAM VI)
Peng et al. (2018)	Cu, Pb, Zn, Ni, Cd	Adsorption & desorption	Stirred-flow	Association with SOM Association with Al oxides Association with Fe-(hydr) oxides	Equilibrium-based (WHAM VI) Equilibrium-based (WHAM VI) Equilibrium-based (CD-MUSIC)

phenomena on the rates are not considered by the mechanistic models described in Table 1, or other types of models recently applied to reproduce metal sorption–desorption kinetics, such as competitive multi-reaction models (Sun and Selim, 2019). Competitive multi-reaction models consider different sorption sites that a chemical can sorb to. This includes a site at equilibrium where sorption is assumed to occur instantaneously, a site where irreversible sorption occurs, and a site where reversible sorption occurs. A sorbed species associated with the latter site can either reversibly desorb from this site or irreversibly transfer to an additional pool. This pool can represent, for example, the fraction of a sorbed species on a surface that transforms irreversibly into a secondary mineral phase. Therefore, competitive multi-reaction models can potentially reproduce the complex sorption behaviors well that were described in previous parts of this chapter. However, these models may be considered less mechanistic than those described in Table 1 because they do not enable determination of the nature of the chemical reactions occurring at the sorption sites.

Transport kinetics can potentially represent the rate-determining step of a sorption or desorption mechanism and can depend on multiple factors that are specific to a given soil system, including porosity, permeability, tortuosity. In addition, diffusion kinetics can be affected by the electrostatic potential in the diffuse double layer on soil component surfaces, which, in turn, depends on pH, PZC (point of zero charge), and electrolyte concentration (Goldberg, 1992; Tournassat and Steefel, 2021). A well-constrained predictive model should then account for all chemical and transport processes that are involved in a given sorption–desorption process. Chemical kinetics are not dependent on system conditions (e.g. soil physical properties, chemical concentrations, surface loading, pH, or mixing) except for the temperature (Sparks, 2003). Therefore, rate parameters corresponding to chemical kinetics of a specific sorption–desorption mechanism may be used as model input parameters to predict retention behavior in any soil system where this mechanism occurs. This is possible since chemical kinetics essentially correspond to rates of elementary chemical reactions occurring at the molecular scale. The rate parameters corresponding to an elementary reaction can be used in any other process where this reaction is part of the overall reaction mechanism (Lasaga, 1998). In contrast, the rates corresponding to an overall reaction that is essentially a combination of elementary reactions cannot be used in systems other than the one where they are measured. Such an overall reaction can potentially be obtained by combining different types of elementary reactions whose nature may depend on the conditions of the system (Lasaga, 1998). Therefore, a predictive model that could be applied to multiple soil systems, using the same chemical rate parameters, should ideally describe all elementary reactions occurring during a given sorption–desorption process, including those where intermediate species are involved. The model would then constrain the overall kinetics of the process by only predicting transport kinetics, which can depend on the conditions and nature of the soil system. This would still require, however, all soil physicochemical properties that can potentially affect transport to be specified as model input parameters. One of the impediments to developing such an ideal predictive model is the difficulty in quantifying accurately all mass transport processes that occur between the bulk soil solution and sorption sites in a static saturated soil. Old and new approaches to model diffusion processes can be found elsewhere (Molins and Knabner, 2019; Tournassat and Steefel, 2021).

Advances in analytical methods allow for better characterization of mineral distribution in soils and solution transport in porous materials. For example, a method was recently developed to map mineral distribution in rocks in 2-D (Kim et al., 2021). It is based on a machine learning approach where an artificial neural network (ANN) is employed to treat micro X-ray fluorescence (μ XRF) and micro X-ray diffraction (μ XRD) data. Another method, coupling scanning electron microscopy–energy dispersive X-ray spectroscopy (SEM–EDX) with an image segmentation technique, was developed to map mineral distribution in a sandstone rock (Landrot et al., 2012). This method also allowed one to quantify the reactive surface area exposed to the connected pore network of each mineral phase present in the sandstone. In addition, phase-shift tomographic methods, which have emerged over the last two decades, enable one to discriminate better between materials that have similar absorption and refraction properties compared to absorption-based tomographic techniques (Prade et al., 2016). Therefore, these methods can potentially discriminate better between solid, liquid, and gas phases and visualize micro-cracks or pores compared to conventional absorption-based tomographic techniques (Teixidó et al., 2021). Phase-shift tomographic methods particularly benefit from fourth generation synchrotron facilities, which are currently emerging in different parts of the world; these new machines deliver X-ray beams with increased beam coherence and brightness. These methods have been used to observe in-situ the transport of liquid into solid materials containing pores or fractures, such as mortar (Prade et al., 2016), glass, carbon and flax fabrics, as well as a 3-D printed structures (Teixidó et al., 2021). Therefore, these new analytical approaches may be potentially valuable in determining first, the parts of the soil pore network where adsorptives or desorptives are transported in the soil solution and second, the specific regions of solid components that are exposed to the soil solution where sorption–desorption processes occur. Understanding such aspects may represent a prerequisite for developing models that can successfully predict mass transport and chemical retention behaviors.

Conclusions and future directions

The kinetics of some relevant sorption–desorption reactions that occur in soils were presented in this chapter. It was shown that the amount of contact time between some inorganic or organic chemicals and soil sorbents can dramatically affect the extent of desorption of these chemicals, and can progressively transform them into secondary forms that are more environmentally persistent. Sorption of inorganic or organic chemicals in soils may result in multiple processes that occur simultaneously on the surfaces of soil components. Some of these mechanisms can take place even after apparent equilibrium has been reached in the soil solution. Therefore, determining the kinetics of these processes by quantifying only the amounts of adsorptive or desorptive components in solution, using traditional batch or stirred-flow methods, may not allow one to elucidate the full retention behaviors. This chapter has also demonstrated that the use of analytical techniques enabling the characterization of the adsorbates'

chemical forms directly on the soil components where sorption–desorption mechanisms occur, can be crucial to constrain the overall process.

Mechanistic models to study sorption–desorption kinetics is useful for elucidating the main reaction pathways that a given chemical may follow in a soil. In addition, parameters of chemical rates corresponding to a given sorption–desorption process may be employed in conceptual models to predict chemical fate in soil systems where this mechanism occurs. These parameters cannot be determined using any of the experimental approaches described in Fig. 1 because the kinetics measured with these methods may be affected by multiple transport phenomena. Chemical rates must be measured using techniques that can quantify adsorbate concentrations and follow the kinetics of all elementary reactions that are part of a sorption–desorption chemical mechanism, including those involving intermediate species. This may potentially be possible using the emerging approach that combines XAFS with PCA and MCR-ALS data-treatment. Frontiers in the field include the development of theoretical approaches that can quantitatively predict all transport processes that a chemical species may undergo between the bulk soil solution and specific types of sorption sites, including in non-saturated soils.

References

- Cheng Y, Ding J, Xie X, Ji X, and Zhang Y (2019) Validation and application of a 3-step sequential extraction method to investigate the fraction transformation of organic pollutants in aging soils: A case study of dechlorane plus. *Environmental Science & Technology* 53: 1325–1333.
- Goldberg S (1992) Use of surface complexation models in soil chemical systems. *Advances in Agronomy* 47: 233–329.
- Goldberg S, Criscenti L, Turner D, Davis J, and Cantrell K (2007) Adsorption–desorption processes in subsurface reactive transport modeling. *Vadose Zone Journal* 6.
- Gustafsson JP, Tiber G, Edkymish A, and Kleja DB (2011) Modelling lead(II) sorption to ferrihydrite and soil organic matter. *Environmental Chemistry* 8: 485–492.
- Kim JJ, Ling FT, Plattenberger DA, Clarens AF, Lanzarotti A, Newville M, and Peters CA (2021) SMART mineral mapping: Synchrotron-based machine learning approach for 2D characterization with coupled micro XRF-XRD. *Computers & Geosciences* 156: 104898.
- Landrot G (2018) FASTOSH: a software to process XAFS data for geochemical & environmental applications. In: *Goldschmidt*. Boston: Goldschmidt Abstracts.
- Landrot G, Ginder-Vogel M, and Sparks DL (2010) Kinetics of chromium(III) oxidation by manganese(IV) oxides using quick scanning X-ray absorption fine structure spectroscopy (Q-XAFS). *Environmental Science & Technology* 44: 143–149.
- Landrot G, Ajo-Franklin J, Yang L, Cabrini S, and Steefel C (2012) Measurement of accessible reactive surface area in a sandstone, with application to CO₂ mineralization. *Chemical Geology* 318–319: 113–125.
- Lasaga AC (1998) *Kinetic Theory in the Earth Sciences*. Princeton University Press.
- Loffredo N, Mounier S, Thiry Y, and Coppin F (2011) Sorption of selenate on soils and pure phases: Kinetic parameters and stabilisation. *Journal of Environmental Radioactivity* 102: 843–851.
- Lopes G, Li W, Siebecker MG, Sparks DL, and Guilherme LRG (2021) Combining zinc desorption with EXAFS speciation analysis to understand Zn mobility in mining and smelting affected soils in Minas Gerais, Brazil. *Science of the Total Environment* 754: 142450.
- Lu Y, Hu S, Liang Z, Zhu M, Wang Z, Wang X, Liang Y, Dang Z, and Shi Z (2020) Incorporation of Pb(II) into hematite during ferrihydrite transformation. *Environmental Science: Nano* 7: 829–841.
- Molins S and Knabner P (2019) Multiscale approaches in reactive transport modeling. *Reviews in Mineralogy and Geochemistry* 85(1): 27–48.
- Peng L, Liu P, Feng X, Wang Z, Cheng T, Liang Y, Lin Z, and Shi Z (2018) Kinetics of heavy metal adsorption and desorption in soil: Developing a unified model based on chemical speciation. *Geochimica et Cosmochimica Acta* 224: 282–300.
- Prade F, Fischer K, Heinz D, Meyer P, Mohr J, and Pfeiffer F (2016) Time resolved X-ray dark-field tomography revealing water transport in a fresh cement sample. *Scientific Reports* 6: 29108.
- Selim HM and Zhang H (2013) Modeling approaches of competitive sorption and transport of trace metals and metalloids in soils: A review. *Journal of Environmental Quality* 42: 640–653.
- Shi Z, Peltier E, and Sparks DL (2012) Kinetics of Ni sorption in soils: Roles of soil organic matter and Ni precipitation. *Environmental Science & Technology* 46: 2212–2219.
- Shi Z, Di Toro DM, Allen HE, and Sparks DL (2013) A general model for kinetics of heavy metal adsorption and desorption on soils. *Environmental Science & Technology* 47: 3761–3767.
- Siebecker M, Li W, Khalid S, and Sparks D (2014) Real-time QEXAFS spectroscopy measures rapid precipitate formation at the mineral–water interface. *Nature Communications* 5: 5003.
- Sparks DL (1996) Chemical Equilibria and Kinetics in Soils. *Soil Science* 161(260): 261.
- Sparks DL (2003) *Environmental Soil Chemistry*. Elsevier Science.
- Sparks DL (2015) Advances in coupling of kinetics and molecular scale tools to shed light on soil biogeochemical processes. *Plant and Soil* 387: 1–19.
- Steefel CI, Depaolo DJ, and Lichtner PC (2005) Reactive transport modeling: An essential tool and a new research approach for the Earth sciences. *Earth and Planetary Science Letters* 240: 539–558.
- Steefel CI, Appelo CAJ, Arora B, Jacques D, Kalbacher T, Kolditz O, Lagneau V, Lichtner PC, Mayer KU, Meeussen JCL, Molins S, Moulton D, Shao H, Šimůnek J, Spycher N, Yabusaki SB, and Yeh GT (2015) Reactive transport codes for subsurface environmental simulation. *Computational Geosciences* 19: 445–478.
- Sun W and Selim HM (2019) Transport and retention of molybdenum(VI) in soils: Kinetic modeling. *Soil Science Society of America Journal* 83: 86–96.
- Teixidó H, Caglar B, Revol V, and Michaud V (2021) In-operando dynamic visualization of flow through porous preforms based on X-ray phase contrast imaging. *Composites Part A: Applied Science and Manufacturing* 149: 106560.
- Tipping E (1994) WHAMC—A chemical equilibrium model and computer code for waters, sediments, and soils incorporating a discrete site/electrostatic model of ion-binding by humic substances. *Computers & Geosciences* 20: 973–1023.
- Tourmassat C and Steefel CI (2021) Modeling diffusion processes in the presence of a diffuse layer at charged mineral surfaces: A benchmark exercise. *Computational Geosciences* 25: 1319–1336.
- Vantelon D, Davranche M, Marsac R, Fontaine C, Guenet H, Jestin J, Campaore G, Beauvois A, and Briois V (2019) Iron speciation in iron-organic matter nanoaggregates: A kinetic approach coupling Quick-EXAFS and MCR-ALS chemometry. *Environmental Science: Nano* 6.
- Verbeeck M, Moens C, and Gustafsson JP (2021) Mechanisms of antimony ageing in soils: An XAS study. *Applied Geochemistry* 128: 104936.
- Yan Y, Li W, Yang J, Zheng A, Liu F, Feng X, and Sparks D (2014) Mechanism of Myo-inositol hexakisphosphate sorption on amorphous aluminum hydroxide: Spectroscopic evidence for rapid surface precipitation. *Environmental Science & Technology* 48.
- Zhang H and Selim H (2011) Non-equilibrium transport of heavy metals in soils: Physical and chemical processes. In: *Dynamics and Bioavailability of Heavy Metals in the Rootzone*, pp. 37–63. Boca Raton, FL: CRC Press.

Sorption-desorption hysteresis

Wenhao Wu, Hongxia Zhu, and Kun Yang, Department of Environmental Science, Zhejiang University, Hangzhou, China

© 2023 Elsevier Ltd. All rights reserved.

Introduction	355
Sorption-desorption hysteresis	356
Hysteresis indices	357
Possible mechanisms of sorption-desorption hysteresis	358
Effects of properties of SOM on sorption-desorption hysteresis	359
Humic substances	359
Rubbery and glassy domains of humic substances	359
Plasticization of humic substances	360
Functional groups of humic substances	361
Black carbon	361
Structures of black carbon	361
Pyrolysis temperatures of black carbon	362
Functional groups of black carbon	363
Effects of physiochemical properties of anthropogenic organic compounds on sorption-desorption hysteresis	364
Molecular sizes of anthropogenic organic compounds	364
Functional groups of anthropogenic organic compounds	364
Effects of environmental factors on sorption-desorption hysteresis	364
Temperature	364
Ionic strength	365
pH	365
Co-solutes	365
Conclusion	366
References	366

Abstract

Sorption-desorption hysteresis of anthropogenic organic compounds (AOCs) in soils is essential to the migration, transformation, bioavailability and fate of AOCs in the environment. In this chapter, several topics related to sorption-desorption hysteresis are elaborated, including a definition of hysteresis, distinction between true and artificial hysteresis, quantification methods and possible mechanisms for hysteresis, effects of the properties of soil organic matter (SOM) (i.e., rubbery/glassy states and functional groups), the physiochemical properties of AOCs (i.e., molecular sizes and functional groups) and the environmental factors (i.e., temperature, ionic strength, pH and competitive co-solutes) on hysteresis.

Key points

- Hysteresis is attributed to irreversible pore deformation and sorptive interactions.
- SOM (soil organic matter) with glassy structure can induce significant sorption-desorption hysteresis.
- SOM with rubbery structure or fixed-porous structure cannot induce hysteresis.
- Hysteresis was observed for anthropogenic organic compounds with functional groups and larger molecular sizes.
- Temperature, ionic strength, pH and co-solutes can significantly affect the hysteresis.

Introduction

Nowadays, thousands of tons of anthropogenic organic compounds (AOCs) are produced in factories and used everywhere in our modern life. It is inevitable that some of these AOCs are released into the soil, probably through the uncontrolled disposal of sewage sludge or industrial waste, the applications of pesticides in agriculture, and the atmospheric precipitation of emitted AOCs from the combustion of fuels or biomasses. The contamination of soils by AOCs could cause high risks to the environment and public health. Sorption and desorption of AOCs by soils are the key processes that determine the migration, transformation, bioavailability and fate of AOCs in the environment. Sorption can cause the immobilization of AOCs in soils and generally impede the migration and bioavailability of AOCs in the environment. Desorption, as the reverse process of sorption, could cause the remobilization of sorbed AOCs from soil and enhance the migration and bioavailability of AOCs in the environment (Qiu et al., 2016). Therefore,

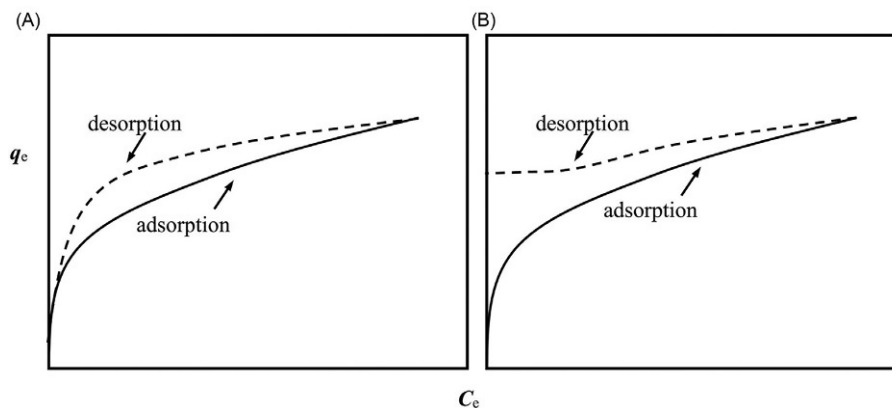


Fig. 1 Sorption-desorption isotherm shows (A) reversible hysteresis with closed hysteresis loop and (B) irreversible hysteresis with open hysteresis loop. Reproduced with permission from Yang K and Xing BS (2010) Sorption of organic compounds by carbon nanomaterials in aqueous phase: Polanyi theory and its application. *Chemical Review* 110: 5989–6008.

knowledge of sorption and desorption behaviors of AOCs in soils are essential for assessing the environmental and health risks of released AOCs. Sorption of AOCs by soils can be reversible or irreversible. The ‘reversible’ means that the pathways for sorption and desorption are the same, while the ‘irreversible’ means that their pathways are different. The ‘hysteresis’ in the isotherm (i.e., the desorption branch is higher than the sorption branch) can be used to identify the irreversible sorption. Sorption-desorption hysteresis has been widely observed for AOCs in soils and sediments, which can be categorized into two types (Yang and Xing, 2010a), i.e., reversible hysteresis and irreversible hysteresis. Reversible hysteresis means that complete desorption can be achieved without any intervention (e.g., vigorous solvent extraction), showing a closed hysteresis loop in desorption and sorption branches of an isotherm (Fig. 1A). Irreversible hysteresis means that complete desorption cannot be achieved, showing an open hysteresis loop in the desorption and sorption branches of an isotherm (Fig. 1B) and an extrapolated nonzero sorbed concentration at zero solution-phase concentration.

Soils are mainly composed of minerals and soil organic matter (SOM). The minerals in soils are silicate clays and oxyhydroxide minerals of Si (silicon), Fe (iron), Al (aluminum), Mg (magnesium), Mn (manganese), etc., while the SOM mainly consists of humic substances that have formed by biogeochemical degradation of plant or animal debris and carbonaceous materials such as soot and charcoal (Yang and Xing, 2010b). Generally, sorption of AOCs by soil minerals is several magnitudes lower than that by SOM, and the sorption by soil minerals may become insignificant if soils contain more than 0.1% SOM (Yang and Xing, 2010b). Since the contents of organic matter are within 0.5–3.0% for most soils in the natural environment, the role of soil minerals in sorption-desorption hysteresis of AOCs is expected to be negligible. Therefore, this chapter focuses on the sorption-desorption hysteresis caused by SOM.

Sorption-desorption hysteresis

Sorption-desorption hysteresis may be true or artificial. The sorbed amounts of sorbate after the sorption process are calculated by the mass difference of sorbate in solution before and after sorption (Eq. 1). Desorption experiments are generally conducted by a batch decant-refill method following the sorption experiment, i.e., replacing the aqueous supernatant of sorbate after sorption equilibrium with a sorbate-free aqueous solution, which allows the sorbed-sorbate to desorb into the fresh solution until a new equilibrium is reached. The sorbed amounts of sorbates after the desorption process are calculated by subtracting the amounts of sorbed-sorbate released into the fresh solution from the sorbed amounts at the sorption equilibrium (Eq. 2).

$$q_e = \frac{m_0 - m_e}{M} = \frac{(C_0 - C_e) \times V}{M} \quad (1)$$

$$q'_e = q_e - \frac{m_{Re}}{M} = q_e - \frac{(C'_0 - C'_e) \times V'}{M} \quad (2)$$

where q_e is the sorbed amount of sorbate in sorption process; q'_e is the sorbed amount of sorbate in desorption process; m_0 and m_e are the masses of sorbate before and after sorption, respectively; m_{Re} is the mass of sorbed-sorbate released into the fresh solution; M is the mass of sorbent; C_0 and C_e are the aqueous concentrations of sorbate before and after sorption, respectively; C'_0 and C'_e are the aqueous concentrations of sorbate before and after desorption, respectively; V and V' are the solution volumes in sorption and desorption, respectively. Artificial hysteresis may be caused by the artificial decrease of sorbed amounts in the sorption process (q_e) or by the artificial increase of sorbed amounts in the desorption process (q'_e), which could mainly result from the artificial increase of C_e or the artificial decrease of C'_e . Four main categories of circumstances can be responsible for the artificial change in sorbate concentration (C_e or C'_e) and subsequently the apparent sorption-desorption hysteresis, i.e., (i) the mass loss of

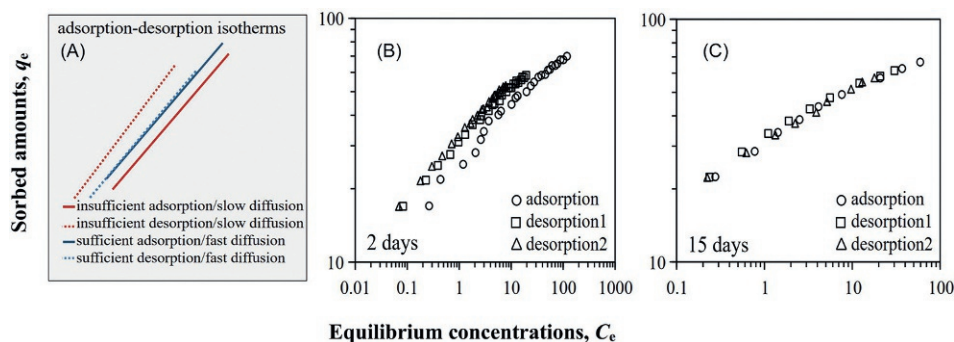


Fig. 2 Schematic diagram of sorption-desorption isotherms with insufficient and sufficient equilibrium times (A) and the example of sorption-desorption isotherms of 1,3,5-trinitrobenzene by a chitin-derived biochar with equilibrium times of 2 days (B) and 15 days (C). Reproduced with permission from Zhu HX, Liu XY, Jiang Y, Zhang M, Lin DH, Yang K (2021) Time-dependent desorption of anilines, phenols, and nitrobenzenes from biochar produced at 700 °C: Insight into desorption hysteresis. *Chemical Engineering Journal* 422: 130584.

sorbent and sorbate from the vials, (ii) the biological degradation of sorbate, (iii) failing to allow the sorption and desorption processes to reach their respective equilibrium, and (iv) the concentration and property change of a third phase (e.g., colloids) after the dilution in desorption process.

The mass loss of sorbent and sorbate from vials can be avoided by careful management of the experiments and proper selection of vials. The biological degradation of sorbate can be addressed by adding biocides. However, nonequilibrium sorption and desorption can be the most serious circumstance for artificial hysteresis. It is commonly acknowledged that contact time of weeks or months are required for approaching the true sorption and desorption equilibrium in the AOC-soil system (Huang et al., 1998). Hence, short contact time such as several hours to several days for sorbate and sorbent could be responsible for this type of artificial hysteresis. In this circumstance, the sorption process of sorbate continues during the desorption experiment, resulting in the underestimating of q_e , while the desorption process of sorbate is stopped before the true equilibrium, leading to the overestimating of q_e' . Accordingly, an unrealistically larger difference between the apparent q_e and q_e' is observed, i.e., the artificial hysteresis (Fig. 2A). For example, significant hysteresis was observed for the sorption-desorption of 1,3,5-trinitrobenzene by a chitin-derived biochar with a contact time of 2 days (Fig. 2B), while much less hysteresis was observed with a longer contact time (Fig. 2C) (Zhu et al., 2021). The failure to allow the sorption and desorption of 1,3,5-trinitrobenzene by the biochar to reach the true equilibrium within 2 days could be responsible for the artificial hysteresis. Usually, nonequilibrium desorption contributes more than nonequilibrium sorption to this type of artificial hysteresis, because the desorption process is frequently slower than the corresponding sorption process due to the slower diffusion process in desorption.

The fourth category of circumstance for artificial hysteresis may occur when there is an 'invisible' partition phase (e.g., dissolved organic matter or macromolecule) for AOCs that coexisted in the sorption system. The partition phase could significantly enhance the apparent solubility of AOCs in aqueous solution and thus lower their q_e values to soils. The refilling solution used in the following desorption process is generally free of both the sorbate and the organic matter/macromolecule. The decrease in the partition phase of the desorption process can generate a higher q_e' value. Hence, an unrealistically larger difference between the apparent q_e and q_e' for AOCs to soils is observed, i.e., the artificial hysteresis.

Hysteresis indices

Although the sorption-desorption hysteresis could be artificial, true hysteresis for the sorption and desorption of organic compounds by soils is commonly observed where the separation of the sorption and desorption isotherms of organic compounds by soils is not attributed to the experimental artifacts. Quantification of the extent of sorption-desorption hysteresis is of great importance for evaluating the sorption and desorption behaviors of organic compounds in soils and further for assessing their potential environmental and health risks. Several empirical hysteresis indices (HIs) have been used to quantify the sorption-desorption hysteresis (Huang et al., 1998; Sander et al., 2005), which have a similar form representing the relative difference between the characteristics of sorption and desorption isotherms (Eq. 3).

$$HI = \frac{|W_{\text{sorb}} - W_{\text{desorb}}|}{W_{\text{sorb}}} \quad (3)$$

where W could be one of the following: sorbed amount of sorbate (q_e), distribution coefficient ($K_d = q_e/C_e$), the slope of isotherm (S), the area under isotherm (A) and the exponent of Freundlich equation fitted to the isotherm (N); W_{sorb} and W_{desorb} are those parameters in the sorption and desorption isotherms, respectively. Hysteresis usually occurs for the nonlinear sorption and desorption of organic compounds by soils, where the values of parameters q_e , K_d and S vary with the aqueous concentrations of organic compounds. Hence, the sorbate concentration should be fixed when comparing or discussing the HI based on the parameters above. Conversely, the HI based on parameters A and N are independent of sorbate concentration, because the values of

A and N do not vary for the nonlinear isotherms. It is not necessary to fix the sorbate concentration when analyzing the HI based on A or N . Although the empirical HIs mentioned above are convenient for identifying the extent of sorption-desorption hysteresis of organic compounds by soils, one major drawback of them is that they depend arbitrarily on the dilution ratio in the desorption process. For true sorption-desorption hysteresis, a higher dilution ratio could decrease the value of C_0' and increase the value of q_e' in Eq. (2), and subsequently lead to a greater hysteresis. Thus, the dilution ratio should be fixed when comparing HIs for different sorption-desorption of AOCs by soils. The greater hysteresis could be identified by the larger HI value when the dilution ratio remains constant.

Possible mechanisms of sorption-desorption hysteresis

Sorption-desorption hysteresis (i.e., irreversible sorption) of AOCs by soils could be caused by the irreversible pore-filling and pore-emptying. The possible mechanisms of true hysteresis include the irreversible pore deformation of soils during the sorption-desorption processes (Sander and Pignatello, 2005; Braida et al., 2003) and the irreversible sorptive interaction (i.e., chemical bonds) between AOCs and soils (Hameed et al., 2020). There are two main categories of irreversible pore deformation that are responsible for sorption-desorption hysteresis, i.e., pore dilation (or pore creation) and formation of closed interstitial space. First, the pore volume of sorbent material with flexible structure (e.g., glassy polymer) is postulated to increase as a result of pore creation or dilation caused by the sorption of organic molecules. However, the created or dilated pores do not relax freely when the fresh solution is added during the desorption of organic molecules (Sander and Pignatello, 2005). As the pore volume of the sorbent is larger in the desorption process than that in the sorption process, the pathways for sorption and desorption of organic compounds are different, resulting in the irreversible sorption and desorption (Fig. 3).

Second, pore deformation of the sorbent during the sorption-desorption process is likely to induce the formation of closed interstitial space, where the sorbed molecules could no longer exchange freely with molecules in the aqueous phase (Braida et al., 2003). Some pores of the sorbent may be swollen and open as a result of the sorption of organic molecules at relatively high concentrations, and then collapse when the external concentration decreases abruptly with the addition of fresh solution in the desorption process. The sorbed organic molecules in the collapsed pores are likely to become entrapped and unable to re-equilibrate with the organic molecules at a decreased concentration in the aqueous phase (Fig. 4). The mechanism of irreversible pore deformation has been successfully used to explain the sorption-desorption hysteresis of organic compounds by soils or the components of soils, such as naphthalene and 1,4-dichlorobenzene by kerogenous organic matter, 1,3-dichlorobenzene and 1,2,4-trichlorobenzene by humic acids and a peat soil (Sander and Pignatello, 2005; Sander et al., 2005).

The sorption and desorption behaviors of AOCs by soils is largely determined by the sorptive interactions (e.g., Van de Waals forces, π - π interactions, hydrogen bonds and chemical bonds), in addition to the pore structure of soils mentioned above. The AOCs sorbed by soils through physical interactions (e.g., Van de Waals forces, π - π interactions and hydrogen bonds) are easily desorbed with the addition of fresh solution in the desorption process, due to the relatively lower energy of those physical interactions. However, the desorption process of AOCs sorbed through chemical interactions (e.g., covalent bonds) is more difficult because of the much higher energy of chemical interactions. Chemical interactions (e.g., covalent bonds) have been identified in the sorption

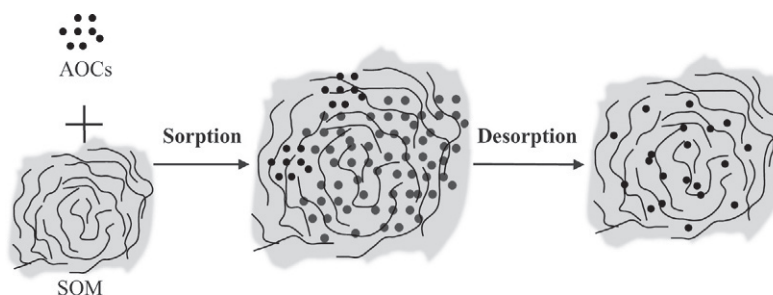


Fig. 3 Schematic diagram of matrix expansion of soil organic matter (SOM) by the sorption of anthropogenic organic compounds (AOCs) that are responsible for sorption-desorption hysteresis.

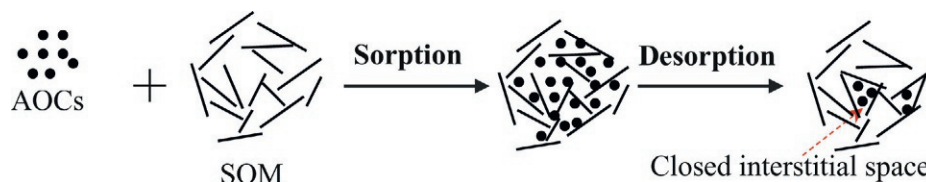


Fig. 4 Schematic diagram of the formation of closed interstitial space and entrapment of anthropogenic organic compounds (AOCs) in soil organic matter (SOM) that is responsible for sorption-desorption hysteresis.

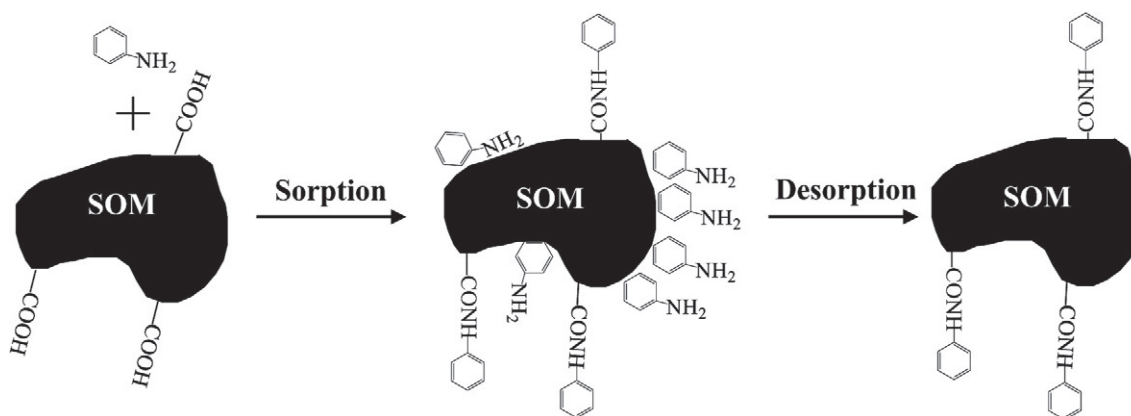


Fig. 5 Schematic diagram of irreversible sorptive interactions (i.e., covalent bonds) between anthropogenic organic compounds (AOCs) and soil organic matter (SOM) responsible for sorption-desorption hysteresis.

of aromatic amines, phenols, pesticides, explosive compounds, and polycyclic aromatic hydrocarbons by soils or the components of soils (Thorn et al., 1996; Thorn and Kennedy, 2002; Hameed et al., 2020). The irreversible sorption interactions could be another possible mechanism that is responsible for the sorption-desorption hysteresis of AOCs by soils (Fig. 5).

Effects of properties of SOM on sorption-desorption hysteresis

Although minerals and SOM are the two main components of soils, SOM is dominantly responsible for the sorption and desorption behaviors of organic compounds in soils. Soil organic matter consists mainly of humic substances (e.g., fulvic acids, humic acids and humins) and carbonaceous materials (e.g., black carbon). Humic substances are complex and heterogeneous mixtures of polydispersed materials formed by biochemical and chemical reactions during the decay and transformation of plant and microbial remains (a process called humification). Black carbon is formed either by carbonization (charring) of organic matter during combustion (charcoal particles) or by condensation from the gaseous phase in reducing flames (soot particles). Both humic substances and black carbons are important geo-sorbents for organic compounds. The effects of their properties on the sorption-desorption hysteresis of organic compounds will be elaborated below.

Humic substances

Rubbery and glassy domains of humic substances

Humic substances are macromolecular organic solids, which are suggested to comprise two principal reactive domains: a highly amorphous, 'soft' or 'rubbery' domain and a condensed, 'hard' or 'glassy' domain (Sander et al., 2006). Generally, the two different domains exhibit distinctly different sorption behaviors for AOCs. For example, sorption of AOCs by the rubbery domain appears to be linear, fast and noncompetitive, whereas sorption by the glassy domain appears to be nonlinear, slow and competitive. Sorption of AOCs by the rubbery domain is a bulk process (i.e., partition), similar to the dissolving of AOCs in organic solvents, while sorption by the glassy domain is a surface process (i.e., adsorption) in which the AOCs adhere to the surface of an adsorbent (Sander et al., 2006). Moreover, it was observed that the two different domains exhibit distinctly different sorption-desorption hysteresis for organic compounds. For example, there was no sorption-desorption hysteresis for phenanthrene by a topsoil (Chelsea II) with a mainly rubbery domain, whereas significant sorption-desorption hysteresis was observed for phenanthrene by the subsurface sandy soil (Wagner II) with a mainly glassy domain (Weber Jr et al., 1998). Weber et al. suggested that the sorption-desorption hysteresis was attributed to the irreversible entrapment and/or slow desorption rate of sorbed molecules within porous structures of the glassy domain. The larger glassy domain a soil contains, the greater extent of sorption-desorption hysteresis it exhibits (Weber Jr et al., 1998).

Sander et al. conducted a series of studies to further investigate the mechanisms of sorption-desorption hysteresis of AOCs by a glassy organic matter (Sander and Pignatello, 2005; Sander et al., 2006). They observed that sorption and desorption of naphthalene by the glassy organic matter were reversible at low solute concentration, but were irreversible at high solute concentration. The glassy organic matter could be referred to as a particle that consists of many macromolecules with a microporous network (Sander and Pignatello, 2005). It was proposed that naphthalene at low concentration merely filled the existing voids of glassy organic matter without inducing significant deformation. However, naphthalene at high concentration could dilate the existing voids of glassy organic matter or create new voids, resulting in the significant swelling of organic matter. Matrix relaxation of the glassy organic matter was incomplete, therefore, the structure of organic matter after the sorption-desorption cycle was different from the original one, i.e., the deformation of glassy organic matter is irreversible (Sander and Pignatello, 2005). The swollen organic matter could provide more sorption sites for AOCs and induce higher sorbed amounts in desorption cycle than that in sorption cycle.

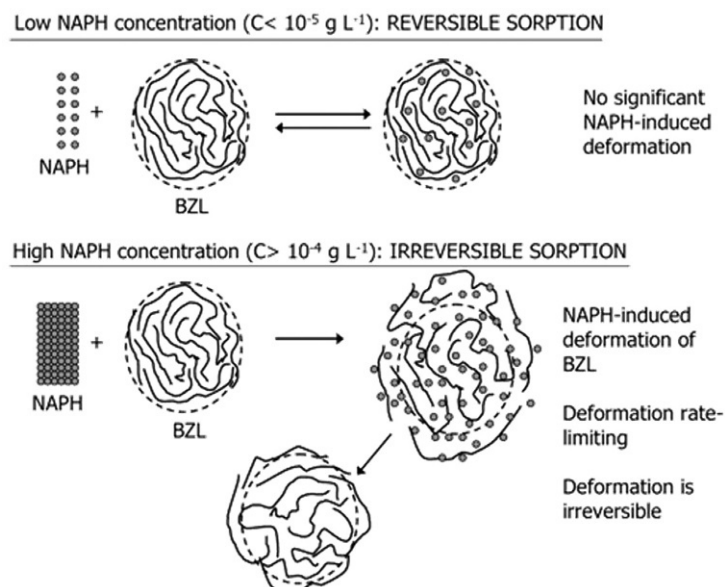


Fig. 6 Proposed mechanism for naphthalene (NAPH) sorption and desorption by a glassy organic matter. Reproduced with permission from Sander M, Pignatello JJ (2005) An isotope exchange technique to assess mechanisms of sorption hysteresis applied to naphthalene in kerogenous (glassy) organic matter. *Environmental Science & Technology* 39: 7476–7484.

Therefore, it was proposed that the sorption-desorption hysteresis of naphthalene at high concentration was attributed to the irreversible deformation of glassy organic matter (Fig. 6).

The enhanced sorption of a solute by a sorbent pre-exposed to the same solute or a similar solute with high concentrations is similar to the 'condition effect' (i.e., enhanced repeat sorption) of glassy polymers, which supports a pore deformation mechanism for the sorption-desorption hysteresis of a solute by glassy polymers (Sander et al., 2006). Additionally, the sorption-desorption of AOCs by rubbery organic matter and rubbery polymers are both reversible. Thus, it is possible to hypothesize that the sorption-desorption behaviors of organic compounds by humic substances are similar to those by polymers. Rubbery and glassy are the two main states of macromolecule polymers, which could be interconverted at a glassy transition temperature (T_g). The rubbery state ($T > T_g$) could be regarded as the segmental flexibility of the macromolecules, with the structure being readily recovered after exposure to an applied external condition such as sorption. With $T < T_g$, the segments of macromolecules become inflexible (i.e., glassy state), with the structure being slowly rearranged and not recoverable after sorption. Hence, excess free voids for sorption were created. Sorption and desorption in the rubbery state of macromolecules is reversible because the structure is readily recoverable during the sorption-desorption cycle. However, the sorbed amounts in the glassy state of macromolecules during the desorption cycle are larger than that during the sorption cycle, due to the increase in free voids of macromolecules after irreversible deformation for sorption. Since the rubbery and glassy states of macromolecules can be interconverted by changing temperature, the sorption-desorption hysteresis can be changed simultaneously. Conditioning-annealing studies using chlorobenzene as a conditioning agent and polychlorinated benzenes as test compounds in a second sorption step have been conducted to further confirm the above hypothesis (Sander et al., 2006). The conditioning effects and sorption-desorption hysteresis were observed for a peat soil, a soil humic acid, and poly-vinyl chloride (a model glassy polymer), but not for poly-ethylene (a model rubbery polymer). The conditioning effects and sorption-desorption hysteresis for the peat soil and the humic acid disappeared after annealing between 45 °C and 91 °C. However, at environmentally relevant temperatures, it is expected that the conditioning effects could not completely relax and the sorption-desorption hysteresis would not disappear.

Plasticization of humic substances

It was observed that the concentration-dependent sorption-desorption hysteresis of 1,4-dichlorobenzene and naphthalene by a Beulah-Zap lignite was similar, i.e., the hysteresis index increased with the increase in solute concentration. However, the concentration-dependent sorption-desorption hysteresis of 1,4-dichlorobenzene by a Pahokee peat soil was reverse, i.e., the hysteresis index decreased with the increase in solute concentration (Sander and Pignatello, 2009). Clearly, it is the properties of sorbents that are responsible for the differences in concentration-dependent hysteresis. This can be explained by the influence of plasticization in addition to the matrix flexibility of sorbents. Semi-permanent holes in sorbents are affected not only by temperature but also by solute concentration. With increasing sorbate concentration, the sorbate-polymer interactions become dominant over the polymer-polymer interactions, leading to an increase in polymer segmental flexibility and a reduction in the glassy transition temperature. This is regarded as the plasticization (Sander et al., 2006). The transition of glass to rubbery state is induced at the isothermal glass transition concentration (S_g) of the sorbate. Sorption-desorption hysteresis of AOCs by glassy sorbents is significant due to irreversible matrix deformation of the sorbent, whereas hysteresis by both the rubbery sorbent (highly

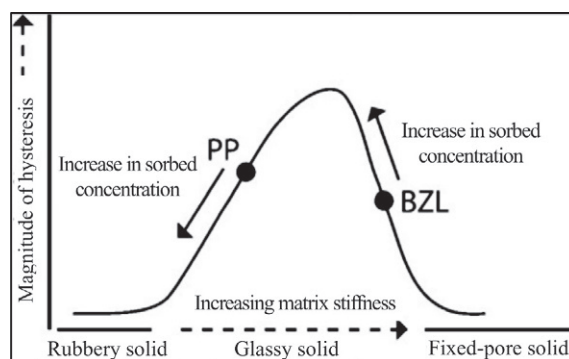


Fig. 7 Schematic diagram of hysteresis as a function of matrix stiffness. The arrows point in the direction of increasing plasticization of Pahokee peat (PP) or Beulah-Zap lignite (BZL) with the loading of a sorbate. Reproduced with permission from Sander M, Pignatello JJ (2009) Sorption irreversibility of 1,4-dichlorobenzene in two natural organic matter-rich geosorbents. *Environmental Toxicology Chemistry* 28: 447–457.

flexible) and completely fixed-pore sorbents (inflexible) are negligible. Sander and Pignatello proposed that there should be an apex of sorption-desorption hysteresis of sorbate with the increase in matrix stiffness of sorbents from rubbery sorbents to completely fixed-pore sorbents (Sander and Pignatello, 2009). Hysteresis increases with the increase in matrix stiffness of sorbents from zero (for rubbery sorbents) to the apex, and then decreases with the increase in matrix stiffness of sorbents from the apex back to zero (for completely fixed-pore sorbents) (Fig. 7). The change in hysteresis corresponds to the matrix stiffness of glassy sorbents. For a glassy sorbent with matrix stiffness below the value that corresponds to the apex of hysteresis (i.e., the rubbery side of the apex), hysteresis would decrease with the increase in degree of plasticization, i.e., the increase in sorbate concentrations. By contrast, for a glassy solid with matrix stiffness above the value that corresponds to the apex of hysteresis (i.e., the fixed-pore side of apex), hysteresis would increase with the increase in degree of plasticization. The T_g of Pahokee peat and Beulah-Zap lignite are 52 °C and 115 °C, respectively, where the Pahokee peat was suggested to lie on the rubbery side of the apex and the Beulah-Zap lignite on the fixed-pore side of it. The 4-dichlorobenzene at low concentration was favorably sorbed in the holes of relatively rigid domains of Pahokee peat where it dilated the existing holes or created new holes, but this was not enough to substantially soften the Pahokee peat. Thus, sorption-desorption hysteresis was observed. However, with increasing 1,4-dichlorobenzene concentration, the plasticization of Pahokee peat became more significant, leading to enhanced reversibility of hole expansion and creation and subsequently the weaker sorption-desorption hysteresis of 4-dichlorobenzene at relatively high concentration than that at relatively low concentration.

Functional groups of humic substances

Sorption and desorption behaviors of AOCs by humic substances are determined not only by the structure of humic substances mentioned above, but also by the interactions between AOCs and humic substances, including both physical and chemical interactions. Physical interactions between organic compounds and humic substances, e.g., hydrophobic interaction, π - π interaction and hydrogen-bonding interaction, are common with relatively low sorption energy and are reversible during the sorption-desorption cycle. For example, sorption and desorption of naphthalene, nitrobenzene, 4-methylnitrobenzene, 2,4-dichlorophenol, 4-nitrophenol by a humus-like substance humified from lignin were all observed to be reversible in our previous study, because the sorptive interactions of the humus-like substance for naphthalene, nitrobenzene, 4-methylnitrobenzene (i.e., hydrophobic interaction and π - π interaction) and for 2,4-dichlorophenol, 4-nitrophenol (i.e., hydrophobic interaction, π - π interaction and hydrogen-bonding interactions) are reversible (Ren et al., 2019). Chemical interactions between AOCs and humic substances such as covalent bonds are common with much higher sorption energy and are irreversible during the sorption-desorption cycle. Amino groups of aromatic amines could interact with certain carbonyl or carboxyl groups of humic substances to form covalent bonds (i.e., imines or amides, Fig. 5), generating unextractable residue and significant sorption-desorption hysteresis. For example, Thorn et al. observed the irreversible combination (i.e., imine bond) of aniline with humic and fulvic acids through the characterization of ^{15}N nuclear magnetic resonance (NMR) spectra (Thorn et al., 1996). Thorn et al. also observed the covalent binding of five reduced TNT amines to soil humic acids through the ^{15}N NMR investigation (Thorn and Kennedy, 2002). Coupling reaction is another kind of chemical interactions between AOCs and humic substances that also could generate unextractable residue and sorption-desorption hysteresis. For example, oxidative couplings of fungicide cyprodinil and herbicide 2,4-dichlorophenoxyacetic acid with humic acids were observed through NMR characterization. Other examples include explosive compounds and phenols.

Black carbon

Structures of black carbon

Black carbon is recognized as pyrogenic carbonaceous matter (e.g., combustion soot, charcoal and biochar) that is produced by natural processes (e.g., natural vegetation fires) or anthropogenic processes (e.g., anthropogenic combustion of fossil fuels, wood

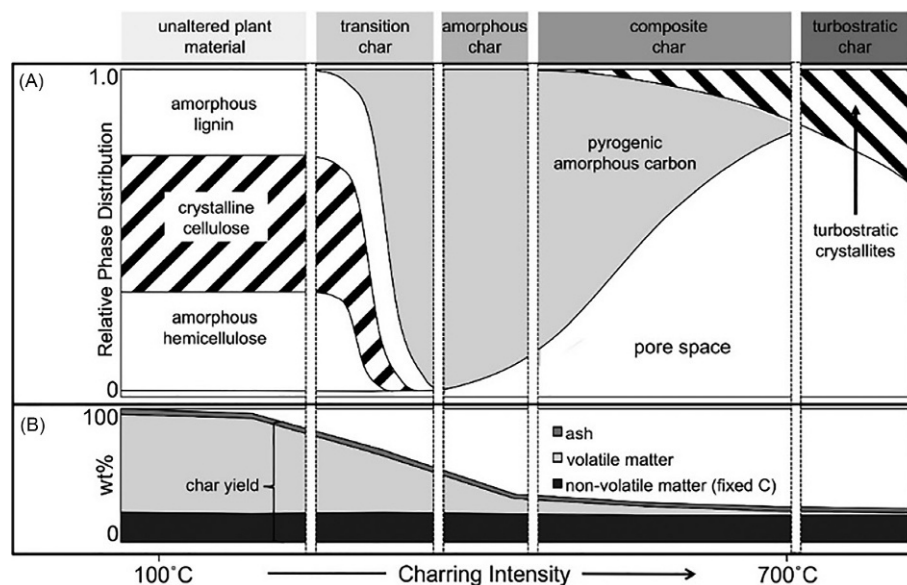


Fig. 8 Dynamic molecular structure of plant biomass-derived black carbon (biochar) across a charring gradient and schematic representation of the five proposed char categories and their individual phases. Reproduced with permission from Keiluweit M, Nico PS, Johnson MG, Kleber M (2010) Dynamic molecular structure of plant biomass-derived black carbon (Biochar). *Environmental Science & Technology* 44: 1247–1253.

and other biomass) and subsequently discharged into the environment. Charcoal and biochar are usually produced by pyrolyzing plant biomass at temperatures typically between 100 °C and 700 °C under limited-oxygen conditions. The chemical composition and structural properties of biochar are determined mainly by the pyrolysis temperature and physiochemical properties of the biomass precursors. After detailed characterization (e.g., surface area, XRD, NEXAFS, and FTIR) of wood- and grass-derived biochars prepared with pyrolysis temperatures from 100 °C to 700 °C, Keiluweit et al. suggested that there are five distinct categories of biochar (Keiluweit et al., 2010): (i) *unaltered plant materials*: at the initial stage of charring (100–200 °C), loss of water and initial dehydration reactions of plant precursors occur and the native structure of the plant biopolymers (e.g., cellulose, hemicellulose, and lignin) are preserved; (ii) *transition char*: plant precursors increasingly undergo dehydration and depolymerization of biopolymers at this charring stage (200–350 °C), creating volatile dissociation products with small molecular weight, but the cellulose retains notable portions of its crystallinity; (iii) *amorphous char*: cellulose is almost completely depolymerized; condensation reactions occur at this charring stage (350–500 °C), forming small aromatic units arranged in random order and some oxygen-containing intermediates; the aliphatic components are ‘fixed’ in the aromatic matrices resisting volatilization and degradation, suggesting a disintegrated and randomly disordered C phase; (iv) *composite char*: turbostratic crystallites are formed at this charring stage (500–700 °C); aromatic, aliphatic, and O-containing components are preserved, forming the amorphous matrix surrounding turbostratic crystallites, i.e., turbostratic crystallites embedded in a low-density amorphous phase; (v) *turbostratic char*: at temperatures higher than 700 °C, the nanoporous phase of turbostratic crystallites with increasing long-range order are formed (Fig. 8).

Pyrolysis temperatures of black carbon

The black carbon in soils and sediments is regarded as an effective sorbent for AOCs due to its large surface area and high reactivity. The sorption and desorption behaviors of AOCs by biochars with different chemical compositions and structural properties are quite different. According to the characteristics (e.g., aromaticity, aliphaticity and porosity) of biochars mentioned above, those produced at low pyrolysis temperatures (i.e., the unaltered plant materials) could be regarded as a rubbery solid; those produced at moderate pyrolysis temperatures (i.e., the transition char and amorphous char) could be regarded as a glassy solid; whereas those from pyrolysis at high temperatures (i.e., the composite char and turbostratic char) could be regarded as fixed-pore solids. According to the significant sorption-desorption hysteresis of organic compounds by glassy SOM and the insignificant hysteresis by rubbery SOM and SOM with fixed-pore structure as discussed in Section “The plasticization of humic substances”, sorption of AOCs is likely to be reversible for biochars from low and high pyrolysis temperatures, but potentially irreversible for biochars from moderate pyrolysis temperatures. Ren et al. observed that sorption of nitrobenzene, 4-methylnitrobenzene, 4-nitroaniline, 2,4-dichlorophenol, 4-nitrophenol and naphthalene by a lignin material treated at 200 °C were all reversible (Ren et al., 2019). The sorption mechanism of AOCs by biochars with low pyrolysis temperature is by partitioning, whereby the biochars can be regarded as partition media with rubbery structure. Sorption of AOC by rubbery solids is commonly reversible. Zhu et al. reported that sorption of aniline, 4-nitroaniline, phenol, 4-nitrophenol by a wood-derived biochar pyrolyzed at 700 °C were all reversible, and sorption of nitrobenzene by five biochars (i.e., wood-, chitin-, cellulose-, lignin- and rice straw-derived biochars) that pyrolyzed at 700 °C were also reversible (Zhu et al., 2021). Braidia et al. investigated the sorption hysteresis of benzene by a wood-derived

charcoal produced at 400 °C and attributed the hysteresis to the glassy structure of charcoal and the pore deformation of charcoal by sorption of benzene (Braida et al., 2003). The swelling of charcoal by benzene was confirmed by liquid displacement and sedimentation measurements. Three scenarios of pore deformation were proposed as being responsible for the sorption-desorption hysteresis of benzene by charcoal, i.e., (i) leakage of sorbate into closed sectors where it is no longer in contact with solute in bulk solution; (ii) irreversible pore expansion; and (iii) contraction of pores and trapping of sorbate within the collapsed structure. The biochars with moderate pyrolysis temperature should have a glassy rather than the rigid, fixed-pore structure. Wang et al. also observed the sorption-desorption hysteresis of AOCs (i.e., nitrobenzene, naphthalene, atrazine) by a rice straw-derived biochar with moderate pyrolysis temperature (400 °C), where the hysteresis and the desorption were weakened when the leachable pyrogenic organic carbon was removed from the biochar by alkaline solution (Wang et al., 2017). The leachable and non-leachable pyrogenic organic carbons could be regarded as the sources of glassy and fixed porous structures of biochar, respectively. The removal of leachable pyrogenic organic carbon changed biochar from a glassy to a fixed-pore solid, reducing the sorption-desorption hysteresis of AOCs.

Therefore, it could be concluded that (i) sorption and desorption of organic compounds by biochars with low pyrolysis temperatures (100–300 °C) should usually be reversible, mainly due to the rubbery structure of biochars and the reversible partition mechanism; (ii) sorption and desorption of organic compounds by biochars with moderate pyrolysis temperatures (300–500 °C) could be irreversible due to the glassy structure of biochars and the irreversible pore deformation mechanism or the irreversible interactions; (iii) sorption and desorption of organic compounds by biochars with high pyrolysis temperatures (500–700 °C) should usually be reversible due to the fixed-nanoporous structure of biochars.

Functional groups of black carbon

With increasing pyrolysis temperature, the amount of oxygen-containing functional groups of black carbon would normally decrease, in addition to the structure change. Irreversible interactions (e.g., covalent bonds) are possibly formed between the functional groups of black carbon and AOCs, and could subsequently be responsible for the sorption-desorption hysteresis. For example, significant sorption-desorption hysteresis of oxytetracycline was observed for rice straw-derived biochars pyrolyzed at 300 °C, 400 °C and 500 °C, but not for biochars pyrolyzed at 200 °C, 600 °C and 700 °C (Fig. 9) (Hameed et al., 2020). The biochars with moderate pyrolysis temperatures (300–500 °C) commonly have plenty of residue carboxyl groups (—COOH) which could form amide bonds with the amino groups (—NH₂) of oxytetracycline. Although biochars with low pyrolysis temperatures (100–200 °C) also contain abundant carboxyl groups from the plant precursors, sorption of AOCs to biochars is dominated by the partition mechanism, where irreversible chemical interactions are probably not essential and the sorption of oxytetracycline is reversible. Biochars, however, with high pyrolysis temperatures (500–700 °C) usually have far fewer carboxyl groups because of decomposition reactions and probably could not interact with oxytetracycline through the irreversible chemical bonds.

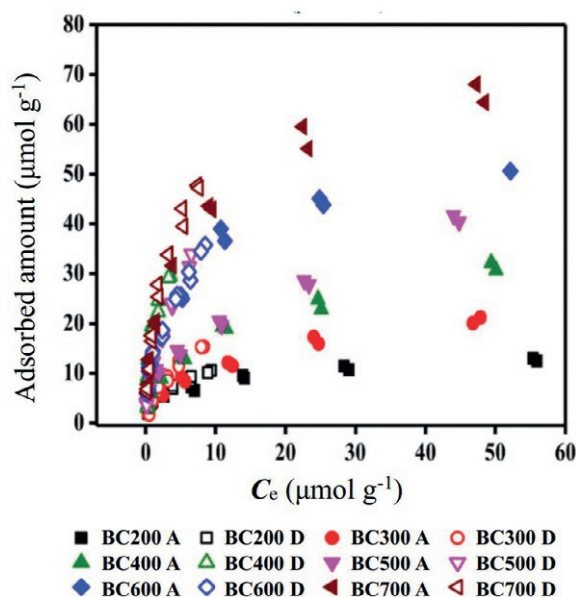


Fig. 9 Sorption and desorption isotherms of oxytetracycline on a rice straw-derived biochar with different pyrolysis temperatures. The letters A and D after BC200–BC700 represent sorption and desorption, respectively. Reproduced with permission from Hameed R, Lei C, Lin DH (2020) Adsorption of organic contaminants on biochar colloids: Effects of pyrolysis temperature and particle size. *Environmental Science and Pollution Research* 27: 18412–18422.

Effects of physiochemical properties of anthropogenic organic compounds on sorption-desorption hysteresis

Molecular sizes of anthropogenic organic compounds

For humic substances and low temperature-pyrolyzed black carbon with rubbery structure, sorption and desorption of AOCs are mainly dominated by the reversible partition mechanism. Generally, sorption-desorption hysteresis is absent no matter what the difference is in molecule sizes of AOCs. For example, sorption of nitrobenzene, 4-methylnitrobenzene, 4-nitroaniline, 2,4-dichlorophenol, 4-nitrophenol and naphthalene, which have different molecular sizes, by a lignin-derived humus-like substance were all reversible. However, for humic substances and moderate-temperature pyrolyzed black carbon with glassy structure or high temperature-pyrolyzed black carbon with fixed-porous structure, sorption and desorption of AOCs are mainly dominated by pore-filling and pore-emptying mechanisms; there is a steric hindrance effect and the role of molecular size is important. Normally, the rates of sorption and desorption of AOCs should depend on their diffusion rate through the pores of glassy or fixed-porous solids, which is negatively correlated with the molecular sizes of the AOCs. Therefore, sorption and desorption rates of AOCs with larger molecular sizes are likely to be slower. The slow rate of sorption-desorption could be an important cause of artificial sorption-desorption hysteresis. The AOC with larger molecular size and slower rate of sorption or desorption by glassy and fixed-porous SOM is prone to be associated with apparent and artificial sorption-desorption hysteresis. For example, it was reported that the sorption of aniline, 4-nitroaniline, phenol, 4-nitrophenol by a wood-derived biochar pyrolyzed at 700 °C was all reversible within 2 days. Sorption of nitrobenzene by five biochars (i.e., wood-, chitin-, cellulose-, lignin- and rice straw-derived) pyrolyzed at 700 °C was also reversible within 2 days. However, significant sorption-desorption hysteresis was observed for 1,3,5-trinitrobenzene with a relatively larger molecular size by the five biochars. It was confirmed that 2 days were required for the sorption or desorption equilibrium of aniline, 4-nitroaniline, phenol, 4-nitrophenol and nitrobenzene by the biochars, while 5 days and 15 days were needed for the equilibrium of 1,3,5-trinitrobenzene by the wood-derived and the chitin-derived biochars, respectively, and more than 30 days for the other three biochars (Zhu et al., 2021).

Apart from the slow sorption and desorption rates, AOCs with larger molecular sizes were considered to induce irreversible pore deformation of SOM to a greater extent or to combine with SOM through stronger interactions, resulting in a more significant sorption-desorption hysteresis. Zhao et al. observed that sorption-desorption hysteresis of perfluorocarboxylates by sediments increased with the increase in molecular chain length (e.g., C6, C8, C9 and C10), and proposed that the perfluorocarboxylates with longer chain length could form stronger interactions with sediments and could also cause irreversible structural rearrangement of the sediment and entrapment of molecules in the sediment (Zhao et al., 2012). Bhandari and Xu observed that sorption-desorption hysteresis of phenolic compounds by two surface soils increased with the increase in molecular size, e.g., phenol < cresol < naphthol (Bhandari and Xu, 2001). Ding and Rice observed that sorption-desorption hysteresis of phenanthrene was more significant than that of naphthalene by two natural organic matters (Ding and Rice, 2011). Higher sorption affinity for SOM and AOC with larger molecular size was probably responsible for the more significant hysteresis.

Functional groups of anthropogenic organic compounds

The functional group of AOCs is another important factor that determines the sorption-desorption hysteresis by soils. AOCs containing functional groups that can form irreversible interactions (e.g., covalent bonds) with soils could induce the hysteresis. For example, significant sorption-desorption hysteresis was observed for oxytetracycline by rice straw-derived biochars with moderate pyrolysis temperatures (e.g., 300–500 °C) due to the formation of amide bonds between the amino group ($-\text{NH}_2$) of oxytetracycline and the carboxyl groups ($-\text{COOH}$) of biochars. However, no hysteresis was observed for phenanthrene due to the lack of a polar functional group (Hameed et al., 2020). Similarly, the observed imine bonds of aniline reacted with soil humic and fulvic acids and five reduced TNT amines with soil humic acids could also induce sorption-desorption hysteresis, due to the irreversible combination of amino groups of amines and carbonyl or carboxyl groups of humic substances. Fungicide cyprodinil and herbicide 2,4-dichlorophenoxyacetic acid with polar functional groups could react with humic substances through oxidative couplings, which is also a type of irreversible combination and could cause significant sorption-desorption hysteresis. Zhao et al. observed that the sorption-desorption hysteresis of perfluoroalkane sulfonates by sediments was more significant than that of perfluorocarboxylates with the same chain length, and proposed that the sulfonate group of perfluoroalkane sulfonates could strongly interact with the sediment and be responsible for the hysteresis (Zhao et al., 2012).

Effects of environmental factors on sorption-desorption hysteresis

Temperature

The sorption-desorption hysteresis of AOCs by soils could be influenced by temperature from two aspects, i.e., changing the rate of sorption or desorption and the structure of SOM. Higher temperatures could accelerate the movement of molecules and was confirmed to increase the rates of sorption and desorption of AOCs by soils, which consequently could shorten the time required to reach the sorption-desorption equilibrium and reduce artificial hysteresis. Generally, sorption through physical interactions is an exothermic process, while the corresponding desorption is endothermic. In addition to increased rates of sorption and desorption, higher temperatures could enhance the amounts of AOCs desorbed by soils, but reduce those sorbed, changing the sorption-

desorption equilibrium to be desorption-favorable. For example, Carroll et al. observed that desorption of polychlorinated biphenyls from soils and sediments was significantly enhanced by increasing temperature (Carroll et al., 1994). Thermal desorption technology (i.e., treating the organic contaminated-soils at high temperatures) has widely been used for removing organic contaminants in soil remediation. However, the enhancement of desorption and the reduction in sorption by the increase in temperature could not be responsible for sorption-desorption hysteresis because the isotherms of sorption and desorption decrease simultaneously and may still overlap. Conversely, sorption by chemical interactions (e.g., amidation reaction) is an endothermic process, while the corresponding desorption is an exothermic process. The increase in temperature could enhance chemical sorption and reduce the desorption of organic compounds by soils, changing the sorption-desorption equilibrium to be sorption-favorable. For example, the sorption of aniline by carbonaceous sorbents was significantly enhanced by increasing temperature. The increase of chemical sorption by increasing temperature probably could enhance sorption-desorption hysteresis, because sorption through chemical bonds could not be released in the desorption process. For example, the sorption-desorption hysteresis of aniline by carbonaceous sorbents increased with the increase in temperature from 10 °C to 40 °C (Wu et al., 2013). The rubbery state of SOM can be changed to the glassy state by increasing the temperature once it exceeds the glassy transition temperature of SOM. Sorption of AOCs by SOM in the rubbery state is reversible, while sorption by SOM in glassy state is irreversible. Therefore, the irreversible sorption of AOCs by SOM with glassy structure could be changed to reversible sorption by an increase in temperature that exceeds the glassy transition temperature. For example, the significant sorption-desorption hysteresis of 1,2,4-trichlorobenzene by a Pahokee peat disappeared as the temperature increased from room temperature to 91 °C (Sander and Pignatello, 2007).

Ionic strength

The influences of ionic strength on sorption and desorption of AOCs by soils mainly relate to three aspects: (i) effect on the structure of SOM; (ii) salting-out effects for AOCs; (iii) competition with AOCs for sorption. Wu and Sun (2010) reported that the sorption-desorption hysteresis of phenanthrene by a soil and a sediment were both weakened as the ionic strength increased from 100 $\mu\text{g L}^{-1}$ to 500 $\mu\text{g L}^{-1}$. They proposed that a higher ionic strength could compress humic substances to a more condensed form and decrease the size of voids in humic materials by electrostriction which is responsible for the sorption-desorption hysteresis of organic compounds (Wu and Sun, 2010). Kan et al. observed that the irreversible fraction of sorbed naphthalene increased with the increase of ionic strength from 0.05 M to 1.0 M, due to the blocking of micropores in soils and sediments by the flocculation of clay (Kan et al., 1994). The flocculation tendency of clay was positively correlated with ionic strength. Qiu et al. reported that the sorption-desorption hysteresis of tetrachloroethylene by soils was weakened as the ionic strength increased from 0.01 M to 0.5 M, and proposed that higher ionic strength could induce salting-out effects and transform more tetrachloroethylene from the aqueous to solid phase in the rubbery state, enhancing reversible sorption and reducing sorption-desorption hysteresis (Qiu et al., 2016). Sorption of ionic AOCs by soils through electrostatic interactions could be reduced by the increase of ionic strength due to competition in ion exchange, while desorption could be enhanced. However, the change of sorption-desorption hysteresis depends on the reversibility of sorption derived from electrostatic interactions. Hysteresis could be reduced if the weakened sorption by ionic strength is irreversible, whereas it could be unchanged or enhanced if the weakened sorption by ionic strength is reversible.

pH

The influences of solution pH on sorption and desorption of AOCs by soils could be mainly through altering the packed structure of SOM, the physiochemical properties of AOCs, and subsequently the sorptive interactions of AOCs with SOM. Humic substances could be regarded as supramolecular associations of self-assembling small molecules, which are stabilized by weak interactions such as hydrophobic interactions, electrostatic interactions and hydrogen-bonding interactions (Piccolo, 2001). The dissociation of functional groups (e.g., $-\text{COOH}$, $-\text{OH}$, $-\text{NH}_2$) of humic substances largely depends on the solution pH. At low pH, humic substances are tightly packed by the strong intermolecular hydrogen-bonding interactions. With increasing pH, the formation of hydrogen-bonding interactions is prevented by the dissociation of functional groups, leaving the surface molecules of humic substances detached. At high pH, the functional groups of humic substances are fully negatively charged by deprotonation, which try to locate as far as possible from each other because of strong electrostatic repulsion and make the humic substances loosely packed. Similar to humic substances, the dissociation of functional groups of ionizable AOCs (e.g., $-\text{COOH}$, $-\text{OH}$, $-\text{NH}_2$) is also largely determined by the solution pH, which could influence the irreversible interactions between AOCs and SOM and subsequently the sorption-desorption hysteresis. For example, at neutral or high pH, the amino groups of organic compounds are not dissociated and could form amide bonds with carboxyl groups of SOM, inducing significant sorption-desorption hysteresis. However, at low pH, the amino groups are dissociated and can no longer form the amide bonds, resulting in the disappearance of sorption-desorption hysteresis. Therefore, solution pH could affect irreversible pore deformation through altering the structure of SOM and irreversible sorption interactions through changing the degree of dissociation of SOM and AOCs.

Co-solutes

In the real environment of soils, different kinds of AOCs commonly co-exist, resulting in competitive sorption of different AOCs by soils. When the sorption system was diluted by a fresh solution to initiate desorption, sorption-desorption hysteresis was probably enhanced due to the 'competitor dilution effect'. The concentration of co-solute in the desorption process of a target solute is less

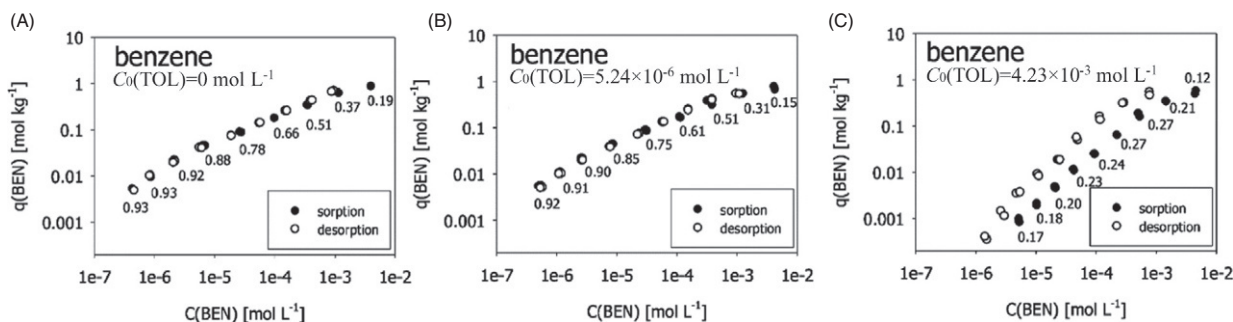


Fig. 10 Sorption-desorption isotherms for benzene by a wood charcoal with the coexistence of toluene at various initial concentrations (C_0). Reproduced with permission from Sander M and Pignatello JJ (2007) On the reversibility of sorption to black carbon: Distinguishing true hysteresis from artificial hysteresis caused by dilution of a competing adsorbate. *Environmental Science & Technology* 41: 843–849.

than that in the sorption process, resulting in the different pathways for sorption and desorption. In the sorption system, the co-solute with a higher concentration could occupy more sorption sites and reduce the sorption of a target solute to a greater extent. In the desorption system, the co-solute with a lower concentration could weaken the sorption of a target solute to a lesser extent. Therefore, the desorption branch of isotherms for a target solute should be higher than the sorption branch, i.e., significant sorption-desorption hysteresis. Moreover, the hysteresis should be positively correlated with the initial concentration of the co-solute, because dilution could induce a greater difference in concentration for the co-solute with a higher initial concentration in the sorption and desorption systems, i.e., the greater difference between sorption and desorption pathways for a target solute. For example, Sander and Pignatello reported that hysteresis was not observed for the single-solute sorption-desorption of benzene, toluene and nitrobenzene by a wood charcoal, but hysteresis was observed in the bi-solute sorption-desorption system and enhanced with the increasing concentration of the co-solute (Fig. 10) (Sander and Pignatello, 2007).

Conclusion

Soil organic matter (SOM) including humic substances and black carbon are mainly responsible for the sorption-desorption hysteresis of AOCs in soils. There are three possible mechanisms that could explain the sorption-desorption hysteresis, i.e., (i) irreversible pore dilation or pore creation, (ii) formation of closed interstitial space and entrapment of organic molecules, and (iii) formation of irreversible sorptive interactions. The SOM with glassy structure (e.g., humic substances in a glassy state and black carbon with moderate pyrolysis temperature) could induce significant sorption-desorption hysteresis, due to limited flexibility of the glassy structure and irreversible pore deformation of glassy structure. However, the SOM with rubbery structure (e.g., humic substances in rubbery state and black carbon with low pyrolysis temperature) did not result in hysteresis because of the highly flexible rubbery structure and reversible pore deformation. The SOM with fixed-porous structure (e.g., black carbon with high pyrolysis temperature) did not lead to hysteresis either because of inflexibility of the fixed-porous structure and the absence of pore deformation. Sorption-desorption hysteresis was frequently observed for AOCs with larger molecular sizes, which could induce longer equilibrium time and more irreversible pore deformation of SOM, and also for AOCs with functional groups due to the formation of irreversible sorptive interactions with SOM. Environmental effects such as temperature, ionic strength, solution pH and competitive co-solute could significantly affect sorption-desorption hysteresis, mainly through creating different pathways for sorption and desorption. This knowledge is crucial for predicting the migration, transformation and bioavailability of AOCs in soils and assessing the potential risks to the environment and public health.

References

- Bhandari A and Xu FX (2001) Impact of peroxidase addition on the sorption-desorption behavior of phenolic contaminants in surface soils. *Environmental Science & Technology* 35: 3163–3168.
- Braida WJ, Pignatello JJ, Lu YF, et al. (2003) Sorption hysteresis of benzene in charcoal particles. *Environmental Science & Technology* 37: 409–417.
- Carroll KM, Harkness MR, Bracco AA, and Balcarce RR (1994) Application of a permeant/polymer diffusional model to the desorption of polychlorinated biphenyls from hudson river sediments. *Environmental Science & Technology* 28: 253–258.
- Ding GW and Rice JA (2011) Effect of lipids on sorption/desorption hysteresis in natural organic matter. *Chemosphere* 84: 519–526.
- Hameed R, Cheng L, and Lin DH (2020) Adsorption of organic contaminants on biochar colloids: Effects of pyrolysis temperature and particle size. *Environmental Science and Pollution Research* 27: 18412–18422.
- Huang WL, Yu H, and Weber WL Jr. (1998) Hysteresis in the sorption and desorption of hydrophobic organic contaminants by soils and sediments 1. A comparative analysis of experimental protocols. *Journal of Contaminant Hydrology* 31: 129–148.
- Kan AT, Fu GM, and Tomson MB (1994) Adsorption/desorption hysteresis in organic pollutant and soil/sediment interaction. *Environmental Science & Technology* 28: 859–867.
- Keiluweit M, Nico PS, Johnson MG, and Kleber M (2010) Dynamic molecular structure of plant biomass-derived black carbon (biochar). *Environmental Science & Technology* 44: 1247–1253.
- Piccolo A (2001) The supramolecular structure of humic substances. *Soil Science* 166: 810–832.

- Qiu ZF, Yang WW, He L, et al. (2016) Characteristics and influencing factors of tetrachloroethylene sorption-desorption on soil and its components. *Chemosphere* 144: 895–901.
- Ren LF, Lin DH, and Yang K (2019) Correlations and nonlinear partition of nonionic organic compounds by humus-like substances humified from rice straw. *Scientific Reports* 9: 15131.
- Sander M and Pignatello JJ (2005) An isotope exchange technique to assess mechanisms of sorption hysteresis applied to naphthalene in kerogenous organic matter. *Environmental Science & Technology* 39: 7476–7484.
- Sander M and Pignatello JJ (2007) On the reversibility of sorption to black carbon: Distinguishing true hysteresis from artificial hysteresis caused by dilution of a competing adsorbate. *Environmental Science & Technology* 41: 843–849.
- Sander M and Pignatello JJ (2009) Sorption irreversibility of 1,4-dichlorobenzene in two natural organic matter-rich geosorbents. *Environmental Toxicology and Chemistry* 28: 447–457.
- Sander M, Lu YF, and Pignatello JJ (2005) A thermodynamically based method to quantify true sorption hysteresis. *Journal of Environmental Quality* 34: 1063–1072.
- Sander M, Lu YF, and Pignatello JJ (2006) Conditioning-annealing studies of natural organic matter solids linking irreversible sorption to irreversible structural expansion. *Environmental Science & Technology* 40: 170–178.
- Thorn KA and Kennedy KR (2002) ^{15}N NMR investigation of the covalent binding of reduced TNT amines to soil humic acid, model compounds, and lignocellulose. *Environmental Science & Technology* 36: 3787–3796.
- Thorn KA, Pettigrew PJ, and Goldenberg WS (1996) Covalent binding of aniline to humic substances. 2. ^{15}N NMR studies of nucleophilic addition reactions. *Environmental Science & Technology* 30: 2764–2775.
- Wang BY, Zhang W, Li H, et al. (2017) Micropore clogging by leachable pyrogenic organic carbon: A new perspective on sorption irreversibility and kinetics of hydrophobic organic contaminants to black carbon. *Environmental Pollution* 220: 1349–1358.
- Weber WL Jr., Huang WL, and Yu H (1998) Hysteresis in the sorption and desorption of hydrophobic organic contaminants by soils and sediments 2. Effects of soil organic matter heterogeneity. *Journal of Contaminant Hydrology* 31: 149–165.
- Wu WL and Sun HW (2010) Sorption-desorption hysteresis of phenanthrene-effect of nanopores, solute concentration, and salinity. *Chemosphere* 81: 961–967.
- Wu WH, Jiang W, Zhang WD, Lin DH, and Yang K (2013) Influence of functional groups on desorption of organic compounds from carbon nanotubes into water: Insight into desorption hysteresis. *Environmental Science & Technology* 47: 8373–8382.
- Yang K and Xing BS (2010a) Adsorption of organic compounds by carbon nanomaterials in aqueous phase: Polanyi theory and its application. *Chemical Reviews* 110: 5989–6008.
- Yang K and Xing BS (2010b) Soil physicochemical and biological interfacial processes governing the fate of anthropogenic organic pollutants. In: Huang PM (ed.) *Handbook of Soil Sciences*, 2nd edn., pp. 4103–4198. CRC Press.
- Zhao LX, Zhu LY, Yang LP, Liu ZT, and Zhang YH (2012) Distribution and desorption of perfluorinated compounds in fractionated sediments. *Chemosphere* 88: 1390–1397.
- Zhu HX, Liu XY, Jiang Y, et al. (2021) Time-dependent desorption of anilines, phenols, and nitrobenzenes from biochar produced at 700°C: Insight into desorption hysteresis. *Chemical Engineering Journal* 422: 130584.

Surface charge and zero-charge points

J Chorover, Department of Environmental Science, The University of Arizona, Tucson, AZ, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of J. Chorover, ZERO-CHARGE POINTS, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 367-373, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00154-5>.

Introduction	369
Components of surface charge	369
Surface charge balance	369
Points of zero charge	372
Point of zero charge	372
Point of zero net proton charge	373
Point of zero net charge	373
Conclusions	375
References	375

Abstract

Soil surface charge exerts strong control over the adsorption of ions from the soil solution, colloidal stability and particle aggregation, and rates of mineral dissolution. There are multiple methods for quantifying surface charge including proton titrations and measurements of surface excess of charged ions. These methods can be used to determine the pH values where one or more components of surface charge vanishes (i.e., points of zero charge).

Glossary

Adsorbate Matter accumulating at the interface between adsorbent and the aqueous phase.

Adsorbed ion charge density (Δq) The net charge density created by adsorption of ions, other than H^+ and OH^- , into inner-sphere complexes, outer-sphere complexes, and the diffuse ion swarm.

Adsorbent The solid surface on which matter accumulates.

Adsorption The net accumulation of matter at the adsorbent surface.

Diffuse layer surface-charge density (σ_d) The net surface charge density created by adsorbed ions in the diffuse layer.

Inner-sphere complex surface-charge density (σ_{is}) The net charge density created by ions adsorbed via direct coordination to surface functional groups (i.e., with no water molecules interposed between adsorbate and surface).

Intrinsic surface-charge density (σ_{in}) The sum of structural and proton surface charge densities.

Outer-sphere complex surface-charge density (σ_{os}) The net charge density created by adsorption of hydrated ions that form stable complexes with adsorbent surface functional groups.

Point of zero charge The pH value at which particle surface charge is equal to zero.

Point of zero net charge The pH value at which adsorbed ion charge density is equal to zero.

Point of zero net proton charge The pH value at which proton surface charge density is equal to zero.

Proton surface-charge density (σ_H) Surface-charge density created by the formation of adsorption complexes involving proton or hydroxide ions as adsorbate; the difference between surface excess of protons and hydroxide ions ($\sigma_H = q_H - q_{OH}$).

Structural surface-charge density (σ_0) Surface-charge density created by isomorphic substitutions among ions of differing valence in the crystal structure of an adsorbent.

Key points

- Inorganic and organic soil particles exhibit “a balance of surface charge” wherein charge resulting from isomorphic substitutions within mineral structures and from complexation of ions in inner- or outer-sphere surface complexes is balanced by ions adsorbed in a diffuse ion swarm.
- Both organic and inorganic particles contain ionizable functional groups that undergo protonation and proton dissociation as a function of pH.
- In addition to protons, other dissolved ions can also form surface complexes, which, in turn, alters particle surface charge.
- On the basis of surface charge balance, we can denote several “points of zero charge” that are quite useful to predicting the behavior of adsorption–desorption reactions in soil.
- The objective of this chapter is to utilize the principal of surface charge balance to review reactions governing particle surface chemistry and associated points of zero charge.

Introduction

Many soil chemical reactions occur at the interface between solid particles and the aqueous phase. Fundamental to the reactivity of this interface is the surface charge on soil particles, which affects a range of chemical phenomena including adsorption of charged solutes, the nature and extent of the electrical double layer on soil particles, kinetics of colloid aggregation, and mineral dissolution rates. **Cation and anion exchange capacities**—which influence the retention of plant-available nutrients, as well as the immobilization of metal(loid) contaminants, such as arsenic (As) and lead (Pb) (Karna et al., 2018)—are directly affected by the nature of soil surface charge and its response to changes in solution of chemical properties such as pH, ionic strength, and ion composition.

Charge properties have traditionally been used to characterize soils as “permanent” or “variable” charge media. These characterizations reflect the relative predominance of one or more soil solid-phase constituents. The surface chemistry of permanent charge soils is governed by isomorphically substituted mineral particles (i.e., 2:1 layer-type clays such as **smectite** and vermiculite), whereas **variable-charge soils** are dominated by amphoteric hydroxylated surfaces (i.e., **metal oxides**, hydroxides or oxyhydroxides, and organic matter). As a result of the ubiquity of hydroxylated solids in soil environments, all soils have some portion of their total surface charge that is conditional upon aqueous chemical conditions.

The law of surface-charge balance states that, irrespective of the source of particle surface charge, it must be balanced by the net accumulation of ions in the surrounding diffuse **swarm** (Sposito, 2016). The balance of surface charge and relationships among components may be used to constrain the application of molecular models to the particle–water interface and to verify internal consistency of adsorption measurements. Particularly useful in this respect are points of zero charge, which define the physico-chemical conditions wherein one or several components of surface charge sum to zero.

Components of surface charge

Electric charge develops on soil particle surfaces as a result of: (1) structural disorder or isomorphous substitutions among ions of differing valence within **soil minerals**, and (2) reactions of surface functional groups with ionic species in **aqueous solution**. These mechanisms of charge development are incorporated into four “components of surface charge” that together contribute to the “total particle surface-charge density,” σ_P . Surface-charge density is conventionally expressed in coulombs per square meter, consistent with the **International System of Units** (SI). For complex solids, including soils, charge is likely to be distributed unequally among the many phases present, owing to differences in structural charge, surface site density, and specific surface area. Furthermore, measurement of total surface area alone is nontrivial and method-dependent. For these reasons, soil charge is often measured on a mass basis and expressed as moles of charge per kilogram. Values presented in each of these sets of units differ by the factor F/a_s , where F is the Faraday constant and a_s is specific surface area.

The “net structural surface-charge density,” σ_0 , is created by isomorphous substitutions in soil minerals. These substitutions occur in both primary and secondary minerals, but they produce significant surface charge only in the 2:1 layer-type **aluminosilicates** (e.g., Fig. 1). The “net proton surface-charge density,” σ_H , results from the difference between the moles of protons and the moles of hydroxide ions complexed by surface functional groups (Fig. 2). The “net inner-sphere complex surface-charge density,” σ_{IS} , results from the net total charge of ions, other than H^+ (hydrogen) and OH^- (hydroxyl), which are bound into inner-sphere surface coordination. Inner-sphere complexation involves the formation of one or more direct bonds between the adsorbate molecule and adsorbent surface functional group, with no water molecules interposed (Fig. 3a). The “net outer-sphere complex surface-charge density,” σ_{OS} , results from the net total charge of ions, other than H^+ and OH^- , which are bound into outer-sphere surface coordination. Outer-sphere complexes involve hydration water interposed between the adsorbate and the adsorbent surface (Fig. 3b). The net total particle surface charge is the sum of these four components:

$$\sigma_P = \sigma_0 + \sigma_H + \sigma_{IS} + \sigma_{OS} \quad (1)$$

The right side of Eq. (1) indicates contributions to particle surface charge from both permanent (σ_0) and conditional (σ_H , σ_{IS} , σ_{OS}) components; the value of σ_0 depends only upon the composition and structure of the solid phase, while the values of the remaining components depend upon the nature of both the solid and solution phases. The last two terms on the right side of Eq. (1) together define adsorbed ions that are immobilized in surface complexes, also known as the “Stern-layer charge,” σ_S .

Surface charge balance

Aqueous particle suspensions are electrically neutral. Therefore, if σ_P is nonzero, it must be balanced by a diffuse **swarm** of **ion charge** equal in magnitude but opposite in sign to σ_P . These ions differ from those bound into the Stern layer because their much greater mobility relative to the adsorbent surface is analogous to diffusive motions of free ions in **aqueous solution**. The net charge of such ions defines the “diffuse layer surface-charge density,” σ_d . The charge balance equation for soil particles may then be written:

$$\sigma_P = -\sigma_d \quad (2)$$

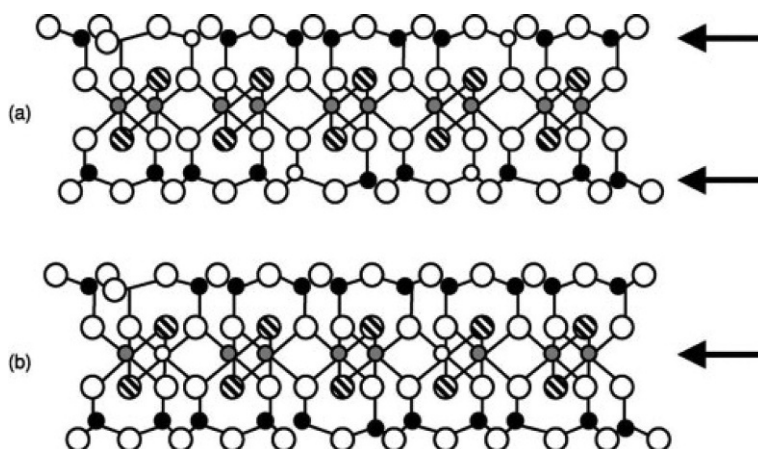


Fig. 1 Isomorphic substitution of ions of differing valence gives rise to structural charge density (σ_0) that is independent of solution chemistry. Model structures of vermiculite (a) and smectite (b) show that substitutions (small white spheres) can give rise to charge in the (a) tetrahedral (Al^{3+} (aluminum) for Si^{4+} (silicon) substitution) or (b) octahedral (e.g., Mg^{2+} (magnesium) for Al^{3+} substitution) sheets of 2:1 layer-type silicates. The location of substitution, particularly in regard to proximity to the surface, affects site reactivity. Large white spheres, O; black spheres, Si; gray spheres, Al; hatched spheres, OH.

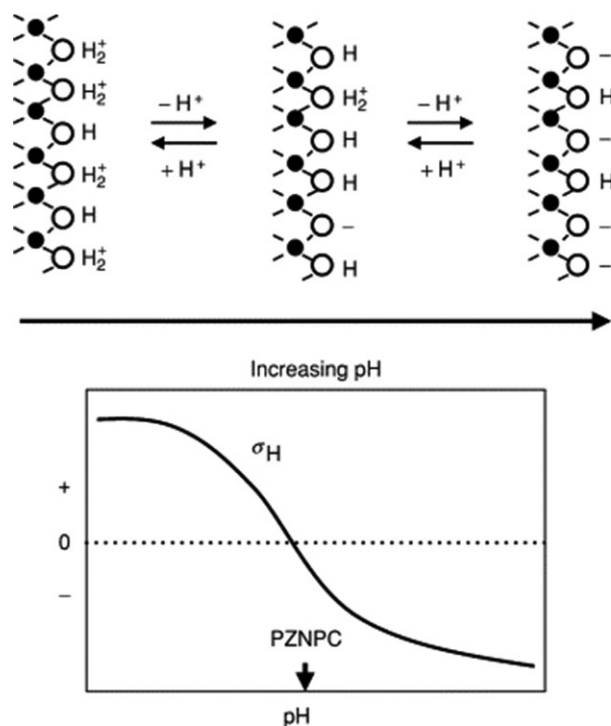


Fig. 2 Development of proton surface-charge density (σ_H) derives from the net adsorption of protons on particle surfaces and reflects the Brønsted acid–base chemistry of surface functional groups. PZNPC, point of zero net proton charge.

and substituting the left side of Eq. (1) into Eq. (2) gives:

$$\sigma_0 + \sigma_H + \sigma_{\text{IS}} + \sigma_{\text{OS}} + \sigma_d = 0 \quad (3)$$

The terms σ_{IS} , σ_{OS} , and σ_d correspond to different molecular mechanisms of ion adsorption that are difficult to distinguish on the basis of quantitative macroscopic measurements. These three terms in Eq. (3) together give the “adsorbed ion charge density,” Δq , which is the net charge of ions, other than H^+ and OH^- , adsorbed at the soil–water interface (Fig. 4):

$$\Delta q = \sigma_{\text{IS}} + \sigma_{\text{OS}} + \sigma_d = q_+ - q_- \quad (4)$$

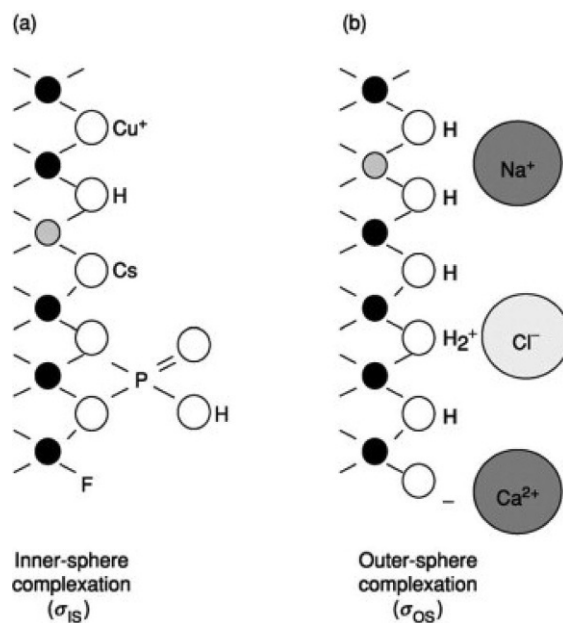


Fig. 3 Surface coordination of ions in stable complexes results in the development of (a) inner-sphere complex surface-charge density (σ_{IS}) and (b) outer-sphere complex surface charge density (σ_{OS}), both of which contribute to total particle surface charge. Isomorphous substitutions, are shown schematically as gray spheres. (Cu, copper; Cs, cesium; P, phosphorus; F, fluorine; Na, sodium; Cl, chloride; Ca, calcium).

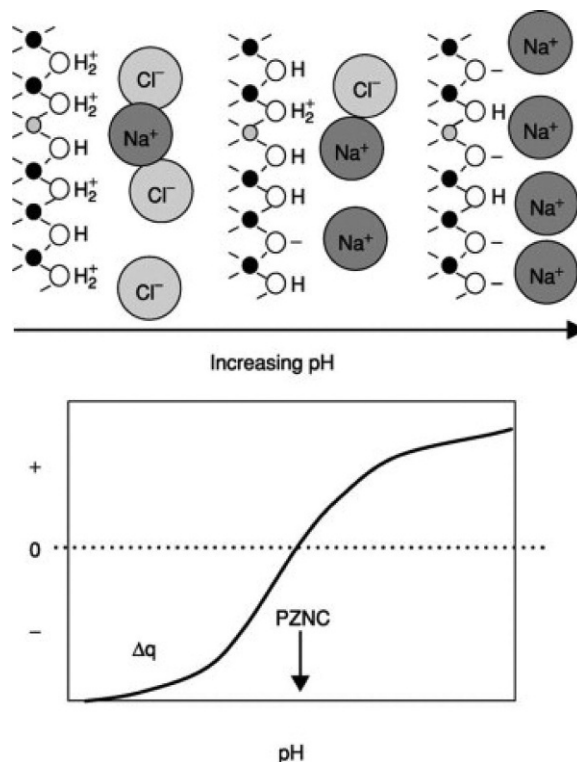


Fig. 4 The net charge deriving from adsorption of ions (other than H^+ and OH^-) in surface complexes ($\sigma_{IS} + \sigma_{OS}$) plus the diffuse ion swarm (σ_d) constitute the net adsorbed ion charge density (Δq). PZNC, point of zero net charge.

The value of Δq in Eq. (4) can be determined by measuring the difference between “surface excess” of cation charge (q_+ corresponding to all cations other than H^+) and “surface excess” of anion charge (q_- corresponding to all anions other than OH^-). The charge balance may then be rewritten as:

$$\sigma_0 + \sigma_H + \Delta q = 0 \quad (5)$$

Points of zero charge

Points of zero charge are defined most commonly as pH values for which one or more of the surface-charge components is equal to zero at a given temperature, pressure, and aqueous solution composition. The emphasis on pH derives from the wide and dynamic range of proton concentrations in soil solutions, and the fact that complexation of H^+ and OH^- at the surfaces of soil particles strongly affects the adsorption of other cations and anions (e.g., see Fig. 4 and Eq. (5)). Although pH will be used to illustrate points of zero charge here, it is important to note that one can also define them in terms of aqueous concentrations of other surface complexing ions, such as $H_2PO_4^-$ (phosphoric acid) or Pb^{2+} (lead) (Fig. 3a), which can affect the value of σ_P directly. Three principal points of zero charge are named conventionally as: (1) the “point of zero charge” (PZC), (2) the “point of zero net proton charge” (PZNPC), and (3) the “point of zero net charge” (PZNC).

Point of zero charge

The PZC ($\sigma_P = 0$) is defined as the pH value at which the net total particle surface charge, σ_P , is equal to zero. According to Eq. (2), σ_d also vanishes at the PZC, so that all adsorbed ions (other than H^+ and OH^-) must be bound into surface complexes. Traditionally, the PZC has been assessed by measuring the pH value at which soil particles exhibit zero electrophoretic mobility (i.e., the isoelectric point or IEP), assuming that a negligible portion of the diffuse ion swarm is advected during the electrophoresis measurement. Model calculations suggest that surface charge heterogeneity or patchiness can also result in a nonzero electrophoretic mobility for a particle whose overall charge is equal to zero. Values of PZC for soil minerals derived dominantly from electrophoresis measurements are presented in Table 1.

According to the DLVO (Derjugin–Landau–Verwey–Overbeek) theory, a suspension of colloidal particles should undergo rapid flocculation when net charge in the diffuse ion swarm (σ_d) is reduced to zero. Therefore, the PZC may also be determined by measuring the pH where a colloidal suspension of particles undergoes rapid coagulation (flocculation). However, the influence of

Table 1 Points of zero charge for selected soil mineral constituents*

<i>Solid</i>	<i>Chemical formula</i>	<i>PZNPC</i>	<i>PZNC</i>	<i>PZC</i>
Albite	$NaAlSi_3O_8$	3.9	–	–
Allophane	$SiAl_2O_5 \cdots 0.5H_2O$	8.2	7.9	10
Alumina	$\gamma-Al_2O_3$	7.0	6.7	8.6
Birnessite	$(Na,Ca)Mn_7O_{14} \cdots 2.8H_2O$	2.4	1.9	1.7
Boehmite	$\gamma-AlOOH$	–	–	8.6
Calcite	$CaCO_3$	–	–	8.5
Corundum	$\alpha-Al_2O_3$	–	–	5–6
Dolomite	$CaMg(CO_3)_2$	–	–	8.0
Forsterite	Mg_2SiO_4	–	–	4.5
Gibbsite	$\gamma-Al(OH)_3$	8.5	8.5	8.1
Goethite	$\alpha-FeOOH$	7.3	7.5	8.2
Hematite	$\alpha-Fe_2O_3$	8.2	8.3	7.1
Illite	$K_{1.75}Si_7Al(Al_3Fe_{0.25}(III)Mg_{0.75})O_{20}(OH)_4$	2.7	–	–
Imogolite	$SiAl_2O_3(OH)_4$	6.8	8.4	11.5
Kaolinite	$Si_4Al_4O_{10}(OH)_8$	5.0	3.6	4.7
Magnetite	Fe_3O_4	6.8	–	8
Quartz	SiO_2	–	2.5	2.0
Rutile	TiO_2	6.0	–	5.9
Silica	SiO_2	2.0	2.4	2.4
Smectite	$Na_2Si_{7.5}Al_{0.5}(Al_{3.5}Mg_{0.5})O_{20}(OH)_4$	8.5	–	<2.3

PZNPC, point of zero net proton charge; PZNC, point of zero net charge; PZC, point of zero charge.

* Sources of data: *Albite*: Mukhopadhyay and Walther (2001). *Allophane*: Harsh et al. (2002). *Alumina*: Goynes et al. (2002). *Birnessite*: Tripathy et al. (2001); Sposito (2004). *Boehmite*: Ermakova et al. (2001). *Calcite*: Wolthers et al. (2008); Al Mahrouqi et al. (2017). *Corundum*: Stack et al. (2001); Fitts et al. (2005). *Dolomite*: Pokrovsky et al. (1999). *Forsterite*: Pokrovsky and Schott (2000). *Gibbsite*: Jodin et al. (2005). *Goethite*: Sigg and Stumm (1981); Zhu et al. (2020). *Hematite*: Chibowski and Janusz (2002); Walsch and Dultz (2010). *Illite*: Sinisyn et al. (2000). *Imogolite*: Harsh et al. (2002). *Kaolinite* (KGa-1): Schroth and Sposito (1997); Sposito (2004). *Magnetite*: Regazzoni et al. (1983); Tombácz et al. (2013). *Quartz*: Eggleston and Jordan (1998). *Rutile*: Ermakova et al. (2001). *Silica*: Ermakova et al. (2001); Goynes et al. (2002). *Smectite*: Tombácz and Szekeres (2006).

surface-charge heterogeneity probably affects collision efficiency as well. Methods involving the use of scanning force microscopy (e.g., [Stack et al., 2001](#)) are useful for elucidating patchwise heterogeneity of surface charge and providing direct force measurements of the PZC (e.g., as used to obtain the value of PZC for [quartz](#) in [Table 1](#)). Such methods have also been applied to characterize the surface charge heterogeneity of clay minerals ([Guo and Yu, 2017](#)).

Point of zero net proton charge

The pH value at which the proton surface-charge density, σ_H , is zero defines the PZNPC. Proton surface-charge density is given by:

$$\sigma_H = q_H - q_{OH} \quad (6)$$

where q_H and q_{OH} are the surface excess (e.g., in moles of charge per kilogram) of H^+ and OH^- , respectively. Therefore, at the PZNPC the moles of surface-adsorbed H^+ and OH^- are balanced. Proton and hydroxide adsorption reactions can be represented by the mass action expressions:



where SOH is a neutral surface hydroxyl group on a soil particle, and s, aq, and l represent the solid, aqueous, and liquid phases, respectively. If proton adsorption–desorption is assumed to occur dominantly at hydroxylated sites (e.g., as in the case of pure metal oxides), then the PZNPC defines the condition where the density or concentration of protonated sites is equal to that of proton-dissociated sites [i.e., $(SOH_2^+) = (SO^-)$]. Thus, both protonated and proton-dissociated sites can coexist at the PZNPC.

In “indifferent” electrolyte solutions (i.e., those comprising monovalent cations and anions that do not form stable complexes at particle surfaces), proton adsorption behavior, as depicted in [Fig. 2](#), is a function of the Brönsted acidity of surface functional groups. Soil particles exhibit a wide range in surface acidity, as indicated by the PZNPC values shown in [Table 1](#). Low PZNPC values are typical of more strongly acidic soil constituents such as silica, quartz, Mn (manganese) oxides, and natural organic matter, whereas higher values are observed for (hydr)oxides of Fe (iron) and Al. [Table 1](#) shows PZNPC values of soil minerals suspended in indifferent electrolyte solutions. However, the presence of cations (anions) that form inner- or outer-sphere surface complexes ([Fig. 3a](#)) will tend to decrease (increase) the PZNPC because of adsorptive competition with H^+ (OH^-).

Proton- and hydroxide-promoted dissolution of soil minerals is found to increase at pH values both below and above the PZNPC. Correlations are observed between dissolution rate constants and the absolute magnitude of σ_H , suggesting that metal–oxygen bonds at the particle surface are weakened as reactions in [\(7a\)](#), [\(7b\)](#) proceed to the right.

Point of zero net charge

The first two terms in [Eq. \(5\)](#) are often combined to provide a measure of “intrinsic surface-charge density,” σ_{in} , which derives from the crystal and surface structure of the adsorbent:

$$\sigma_{in} = \sigma_0 + \sigma_H \quad (8)$$

The PZNC is the pH value at which σ_{in} is equal to zero. The intrinsic surface charge is balanced by the adsorption of ions. Since σ_{in} results from both proton adsorption–desorption reactions (pH-dependent) and isomorphic substitutions, it comprises conditional and permanent components ([Fig. 4](#)). Since $\sigma_{in} = -\Delta q$, the PZNC may be determined: (1) by measuring the pH where surface excess values of cation and anion charge are equal (i.e., $\Delta q = 0$), or (2) by measuring the pH where $\sigma_0 = -\sigma_H$.

Charge balance dictates that, as σ_H decreases with increasing pH (affected by mass action as indicated in [Eqs. \(7a\)](#) and [\(7b\)](#)), ion adsorption–desorption reactions must balance the change in intrinsic charge density. The dependence of σ_H on pH is shown schematically in [Fig. 2](#), and corresponding effects on Δq are illustrated in [Fig. 4](#). For a soil adsorbent that is devoid of structural charge (i.e., $\sigma_0 \approx 0$), such as a highly weathered [Oxisol](#) that contains a negligible quantity of 2:1 layer-type [silicates](#), $\Delta q = -\sigma_H$, and the PZNPC equals the PZNC ([Fig. 5](#), top). This is also the case for pure oxides and hydroxides that are free of structural defects. The presence of negative structural charge ($\sigma_0 < 0$), which occurs because of isomorphic substitutions in many [temperate-zone](#) soils, results in the PZNC being less than PZNPC, and the value of Δq (or σ_H) at the PZNPC (PZNC) is equal to $-\sigma_0$ ([Fig. 5](#), middle). Positive structural charge ($\sigma_0 > 0$) is rare in soils, whereas it is detectable when PZNPC is less than PZNC, and its magnitude is given by the value of Δq (or σ_H) at the PZNPC (PZNC) ([Fig. 5](#), bottom).

In addition to the effects of σ_0 , the nature of variable-charge behavior, such as that depicted in [Fig. 5](#), is highly dependent on the Brönsted acidity—i.e., the proton affinity and adsorption–desorption behavior of the soil particles. For example, the magnitude and pH range of greatest local slope (i.e., $\delta\sigma_H/\delta pH$ or $\delta\Delta q/\delta pH$) depends on the identity and quantity of surface sites undergoing proton adsorption–desorption reactions. Whereas oxides and hydroxides of Fe and Al are weakly acidic, with PZNPC values near neutrality, more strongly acidic groups such as those residing on natural organic matter functional groups (e.g., carboxylic acids), silica or quartz surfaces, and Mn oxide surfaces all tend to decrease PZNPC values for whole soils.

The surface-charge balance of [Eq. \(5\)](#) is summarized in [Fig. 6](#); a plot of Δq vs σ_H should be linear with a slope equal to -1 , and x and y intercepts equal to $-\sigma_0$. The x and y scales in [Fig. 6](#) are arbitrary and actual values depend on surface chemistry of the

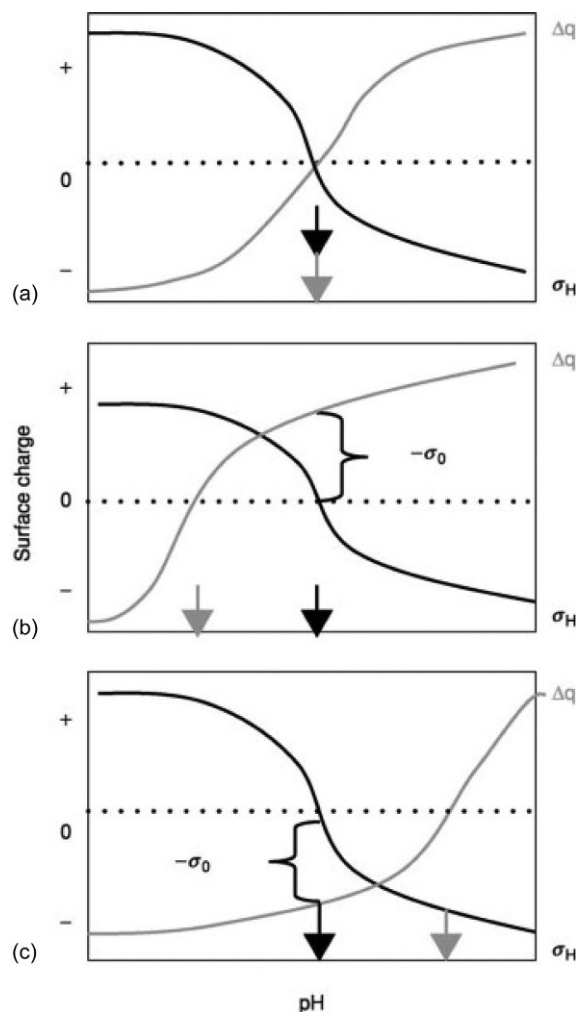


Fig. 5 The inverse relation between σ_H and Δq is depicted schematically for (a) $\sigma_0 = 0$ (PZNC = PZNPC), (b) $\sigma_0 < 0$ (PZNC < PZNPC), and (c) $\sigma_0 > 0$ (PZNC > PZNPC). The PZNC is indicated by gray arrows and the PZNPC is indicated by black arrows. PZNC, point of zero net charge; PZNPC, point of zero net proton charge.

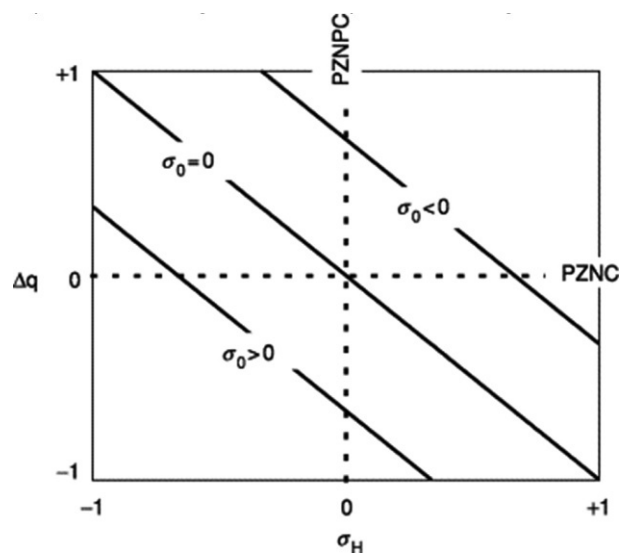


Fig. 6 Surface-charge balance dictates that a plot of Δq vs σ_H must be linear, with a slope equal to -1 and x and y intercepts equal to $-\sigma_0$. PZNPC, point of zero net proton charge; PZNC, point of zero net charge.

adsorbent and the units selected for plotting surface-charge data (e.g., moles of charge per kilogram or coulombs per square meter). Curves such as those depicted in Fig. 5 (with pH as independent variable) are strongly affected by ionic strength and electrolyte composition for a given adsorbent, whereas the relation depicted in Fig. 6 must be reproducible for a given soil, irrespective of changes in solution chemistry.

Conclusions

When conducted rigorously, wet chemical methods, such as proton titration and ion adsorption, can be used to quantify the surface charge behavior of heterogeneous soil suspensions. In so far as natural soils are a multi-component mixture of minerals and organic matter, the charge properties characterized in this way are necessarily an average over all constituents. Hence, under a given set of solution conditions (pH, ionic strength and aqueous ion composition), different types of particles (with different extents of isomorphic substitutions and Bronsted acidities) exhibit different surface charge characteristics, for which the bulk charge is being measured. Conversely, surface force and atomic force microscopies have been employed to quantify the distribution of surface charge across individual particles, thereby greatly increasing the spatial resolution of the method. Future work should focus on interpreting the macroscopic behavior of soils, e.g., bulk surface charge and colloid aggregation behavior, as influenced by the charge properties of the individual constituent particles.

References

- Al Mahrouqi D, Vinogradov J, and Jackson MD (2017) Zeta potential of artificial and natural calcite in aqueous solution. *Advances in Colloid and Interface Science* 240: 60–76.
- Chibowski S and Janusz W (2002) Specific adsorption of Zn(II) and Cd(II) ions at the α -Fe₂O₃/electrolyte interface structure of the electrical double layer. *Applied Surface Science* 178: 1–13.
- Eggleston CM and Jordan G (1998) A new approach to pH of point of zero charge measurement: Crystal-face specificity by scanning force microscopy (SFM). *Geochimica et Cosmochimica Acta* 62: 1919–1923.
- Ermakova L, Sidorova M, Bogdanova N, and Klebanov A (2001) Electrokinetic and adsorption characteristics of (hydr)oxides and oxide nanostructures in 1:1 electrolytes. *Colloids and Surfaces* 192: 337–348.
- Fitts JP, Shang X, Flynn GW, Heinz TF, and Eisenthal KB (2005) Electrostatic surface charge at aqueous/ α -Al₂O₃ single-crystal interfaces as probed by optical second-harmonic generation. *Journal of Physical Chemistry B* 109: 7981–7986.
- Goyne KW, Zimmerman A, Newalkar BL, Komarneni S, Brantley SL, and Chorover J (2002) Surface charge of variable porosity Al₂O₃ (s) and SiO₂ (s) adsorbents. *Journal of Porous Materials* 9: 243–256.
- Guo Y and Yu X (2017) Characterizing the surface charge of clay minerals with Atomic Force Microscope (AFM). *AIMS Materials Science* 4: 582–593.
- Harsh J, Chorover J, and Nizeyimana E (2002) Allophane and imogolite, Chapter 9. In: Dixon JB and Schulze DG (eds.) *Soil Mineralogy With Environmental Applications, No. 7 in the Soil Science Society of America Book Series*. Madison, WI: Soil Sci. Soc. Am.
- Jodin M-C, Gaboriaud F, and Humbert B (2005) Limitations of potentiometric studies to determine the surface charge of gibbsite γ -Al(OH)₃ particles. *Journal of Colloid and Interface Science* 287: 581–591.
- Karna RR, Noerpel MR, Luxton TP, and Scheckel KG (2018) Point of zero charge: Role in pyromorphite formation and bioaccessibility of lead and arsenic in phosphate-amended soils. *Soil Systems* 2: 22.
- Mukhopadhyay B and Walther JV (2001) Acid–base chemistry of albite surfaces in aqueous solutions at standard temperature and pressure. *Chemical Geology* 174: 415–443.
- Pokrovsky OS, Schott J, and Thomas F (1999) Dolomite surface speciation and reactivity in aquatic systems. *Geochimica et Cosmochimica Acta* 63: 3133–3143.
- Pokrovsky OS and Schott J (2000) Forsterite surface composition in aqueous solutions: a combined potentiometric, electrokinetic, and spectroscopic approach. *Geochimica et Cosmochimica Acta* 64: 3299–3312.
- Regazzoni AE, Blesa MA, and Maroto AJG (1983) Interfacial properties of zirconium dioxide and magnetite in water. *Journal of Colloid and Interface Science* 91: 560–570.
- Schroth BK and Sposito G (1997) Surface charge properties of kaolinite. *Clays and Clay Minerals* 45: 85–91.
- Sigg L and Stumm W (1981) The interaction of anions and weak acids with the hydrous goethite (α -FeOOH) surface. *Colloids and Surfaces* 2: 101–117.
- Sinisyn VA, Aja SU, Kulik DA, and Wood SA (2000) Acid–base surface chemistry and sorption of some lanthanides on K⁺-saturated Marblehead illite: 1. Results of an experimental investigation. *Geochimica et Cosmochimica Acta* 64: 185–194.
- Sposito G (2004) *The Surface Chemistry of Natural Particles*. New York: Oxford University Press.
- Sposito G (2016) *The Chemistry of Soils*. New York: Oxford University Press.
- Stack AG, Higgins SR, and Eggleston CM (2001) Point of zero charge of a corundum–water interface probed with optical Second Harmonic Generation (SHG) and Atomic Force Microscopy (AFM): New approaches to surface charge. *Geochimica et Cosmochimica Acta* 65: 3055–3063.
- Tombácz E and Szekeres M (2006) Surface charge heterogeneity of kaolinite in aqueous suspension in comparison with montmorillonite. *Applied Clay Science* 34: 105–124.
- Tombácz E, Tóth IY, Nesztór D, Illés E, Hajdu A, Szekeres M, and Vekas L (2013) Adsorption of organic acids on magnetite nanoparticles, pH-dependent colloidal stability and salt tolerance. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* 435: 91–96.
- Tripathy SS, Kanungo SB, and Mishra SK (2001) The electric double layer at hydrous manganese dioxide/electrolyte interface. *Journal of Colloid and Interface Science* 241: 112–119.
- Walsch J and Dultz S (2010) Effects of pH, Ca- and SO₄ concentration on surface charge and colloidal stability of goethite and hematite—Consequences for the adsorption of anionic organic substances. *Clay Minerals* 45: 1–13.
- Wolthers M, Charlet L, and Van Cappellen P (2008) The surface chemistry of divalent metal carbonate minerals: A critical assessment of surface charge and potential data using the charge distribution multi-site ion complexation model. *American Journal of Science* 308: 905–941.
- Zhu S, Zhang P, Liang Y, Wang M, Xiong J, and Tan W (2020) Effects of aluminum substitution on the surface charge of colloidal goethite particles: Experiments and MUSIC modeling. *Environmental Science and Pollution Research* 27: 38397–38406.

Flocculation and dispersion

Guy J Levy and Isaac Shainberg, Institute of Soil, Water and Environmental Sciences, ARO, Volcani Institute, Rishon LeZion, Israel

© 2023 Elsevier Ltd. All rights reserved.

This is an update of I. Shainberg, G.J. Levy, Flocculation and dispersion, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 27–34, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348,530-4/00363-5>.

Introduction	376
Shape and charge of soil clays	377
Forces between clay platelets	377
Repulsion forces	377
The electrical double layer	377
Hydration repulsion forces	378
Flocculation and attraction forces	378
Van der Waals force of attraction	380
Electrostatic EF attraction forces	380
Clay charge density and electrostatic attraction forces	380
Spontaneous and mechanical dispersion	380
Flocculation and dispersion in Na or Ca clays	380
Effect of pH on the FV of sodium montmorillonite	381
Dispersion and flocculation in Na or Ca clay systems	382
The effect of peptizing agents on dispersion	382
Soil clay systems	382
Dispersion and the hydraulic properties of soils	384
Summary	386
References	386

Abstract

Clay dispersion (deflocculation) and flocculation are fundamental phenomena in soils. Clay dispersion or flocculation depends mainly on clay mineralogy, soil solution ionic strength and pH, and the composition of the exchangeable cations. Various attraction and repulsion forces take place concurrently between clay particles and the balance between them dictates whether dispersion or flocculation occurs. In the case of the latter, different modes for clay particles association may take place. Monovalent cations (e.g., Na⁺) enhance clay dispersion whereas divalent ones (e.g., Ca⁺⁺) enhance flocculation. The sensitivity of soil clay systems to dispersive behavior is usually significantly greater than that of pure clay systems. Dispersion of soil clay particles alter soil pore size distribution and adversely affect soil hydraulic properties.

Key points

- Clay dispersion (deflocculation) and flocculation play a central role in determining soil physical and hydraulic properties.
- The balance between the repulsion and attraction forces acting simultaneously on clay particles determines the stability of the clay suspension.
- The electrical double layer thickness depends on the counter ion valence and the ionic strength of the bulk solution.
- Critical flocculation concentrations depend on clay mineralogy, pH, type of electrolyte and ionic strength.
- Spontaneous clay dispersion often takes place when sodic clay is placed in water of very low electrolyte concentration, yet upon increasing the latter, external mechanical force is needed to force dispersion.

Introduction

The clay particles of soils are of colloid size (less than 2 μm equivalent spherical diameter—Stokes' law) and usually carry a net negative charge. This charge is neutralized by exchangeable cations. The interaction between soil solution and soil clay and the phenomena of clay dispersion and flocculation depends on the type and amount of the soil clays, the nature of the exchangeable cations, and the electrolyte concentration in the soil solution (Shainberg and Letey, 1984). Clay dispersion (deflocculation) and flocculation are very important in determining the physical properties of soils such as hydraulic conductivity, aggregate stability, crusting, infiltration rate, and erosion (e.g., Quirk, 1994).

In a stable clay suspension, dispersed particles collide frequently because of their Brownian motion but separate again because of diffuse double-layer repulsion forces. When salt is added to the clay suspension, the particles may stick together upon collision, forming flocs that settle. The stability of the colloidal suspension is a balance between repulsive and attractive forces acting simultaneously among the suspended particles (Rengasamy and Sumner, 1998; Goldberg et al., 2012).

Shape and charge of soil clays

The clay minerals of most soils are dominated by various layer silicates. Layer silicate clay minerals are classified as 1:1 where each layer consists of one tetrahedral silica sheet and one octahedral alumina sheet (e.g., kaolinite); 2:1 where each layer consists of one octahedral sheet sandwiched between two tetrahedral sheets (e.g., montmorillonite and vermiculite); or 2:1:1 where a metal hydroxide sheet is sandwiched between the 2:1 layers (chlorite).

The clay minerals are characterized by isomorphic substitution of lower-valence cations in either the tetrahedral (Al replacing Si) or octahedral (Mg replacing Al) sheets or both. This excess of permanent negative charge inside the crystal is balanced by exchangeable cations on the external surface. Layer silicate clay minerals also possess variable charge located at the broken edges of clay minerals. At the edges of the 2:1 clay minerals, hydroxyl Al or hydroxyl Si is present. Negative charge may develop on the edges of the octahedral (Al—OH) layer at high pH and positive charge at low pH (Stul and Mortier, 1974).

The colloidal activities of the clay minerals depend on their specific surface area (with units of meters squared per gram) and the charge density expressed by their cation exchange capacity (CEC, in centimoles of cation charge per kilogram) divided by their specific surface area. Kaolinite is a widespread, 1:1 clay mineral with low CEC ($1\text{--}10\text{ cmol}_c\text{ kg}^{-1}$), small specific surface area ($20\text{--}50\text{ m}^2\text{ g}^{-1}$), and a low level of colloidal activity. Conversely, montmorillonite, the dominant clay in soils from semiarid regions, belongs to the 2:1 clay mineral category, has a large CEC ($80\text{--}120\text{ cmol}_c\text{ kg}^{-1}$) and specific surface area ($800\text{ m}^2\text{ g}^{-1}$) and a high level of colloidal activity. Vermiculite with large CEC ($120\text{--}150\text{ cmol}_c\text{ kg}^{-1}$) but intermediate surface area and illite with intermediate CEC ($20\text{--}40\text{ cmol}_c\text{ kg}^{-1}$) and surface area are clay minerals with low-to-intermediate colloidal activity (Goldberg et al., 2012). The mineralogy, charge characteristics, and CEC of several clay minerals, reported by Goldberg et al. (1999), are presented in Table 1.

Forces between clay platelets

Repulsion forces

The electrical double layer

According to the Gouy–Chapman theory, the diffuse double layer consists of the lattice charge and the compensating counter ions (Babcock, 1963). The counter ions are subject to two opposing tendencies: the cations are attracted electrostatically to the negatively charged clay surface; and the cations tend to diffuse from the surface of the particle, where their concentration is high, into the bulk of the solution, where their concentration is lower (Van Olphen, 1977). The two opposing tendencies result in decreasing counterion concentration from the clay surface to the bulk solution (Fig. 1). Divalent cations are attracted to the surface with a force twice as large as that of monovalent cations. Thus, the diffuse double layer in the divalent ion system is more compressed toward the surface. With an increase in the electrolyte concentration in the bulk solution, the tendency of the counterions to diffuse away from the surface is diminished and the diffuse double layer compresses toward the surface. The distribution of mono- and divalent cations in the diffuse double layer at high and low electrolyte concentration is also presented in Fig. 1.

In Gouy–Chapman's diffuse double-layer theory, the counter ions are considered as point charges without dimensions. As long as the ions are far from the clay surface, their concentration is small and their size has little influence on the statistical calculations of the diffuse double layer. However, in the vicinity of the clay surface, where the concentration of the cations is relatively large, the distance of closest approach to the surface is important, and the size of the ions and the energy of hydration of the cations must be considered to give an accurate description of the distribution of counterions near the clay surface. Consequently, the double layer is

Table 1 Charge characteristics and cation exchange capacities of clay minerals.

Solid	Charge per unit half-cell		Cation exchange capacity ($\text{cmol}_c\text{ kg}^{-1}$)
	Tetrahedral	Octahedral	
Kaolinite	0	0	1–10
Smectite			80–120
Montmorillonite	0	–0.33	
Vermiculite	–0.85	0.23	120–150
Mica			20–40
Muscovite	–0.89	–0.05	
Chlorite			10–40

Reproduced with permission from Goldberg S, Lebron I, and Suarez DL (1999) Soil colloidal behavior. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. B195–B240. Boca Raton, FL: CRC Press.

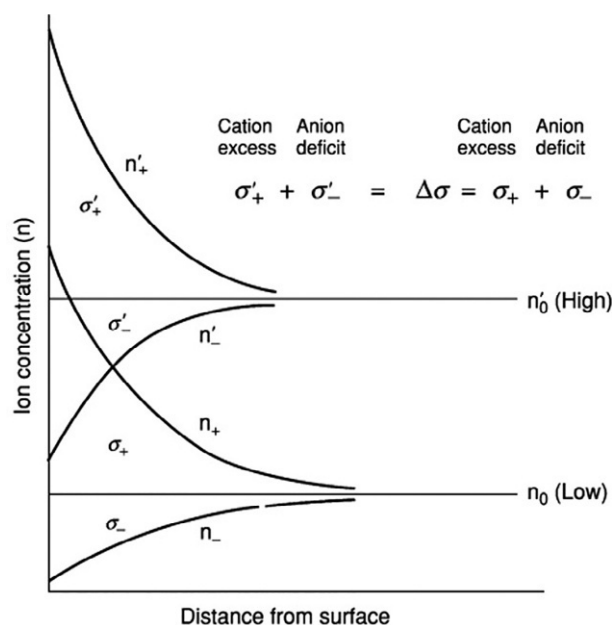


Fig. 1 Distribution of counter ions (n'_+ and n_+) and co-ions (n'_- and n_-) in the electrical double layer for a permanent charge surface (σ) at two electrolyte concentrations. After Van Olphen H (1977) *An Introduction to Clay Colloid Chemistry*, 2nd edn. New York, NY: John Wiley.

divided into the Stern layer, i.e., the first layer of the counterions immediately next to the clay surface, and the Gouy layer, which is the diffuse layer. Adsorbed cations are classified as being in the diffuse or Stern layer depending on their dimension and hydration energy (Van Olphen, 1977). Calculations using the amended theory predict differences in the adsorption characteristics and flocculation values of Li^+ , Na^+ , and K^+ montmorillonite in general agreement with experimental observations.

When two clay platelets approach each other, their diffuse counterion atmospheres overlap. Work must be performed to overcome the electrical repulsion forces between the two positively charged ionic atmospheres. The electrical double-layer repulsion force, also called the 'swelling pressure,' can be calculated by means of the diffuse double-layer theory (Goldberg et al., 2012). The greater the compression of the ionic atmosphere toward the clay surface, the smaller the overlap of atmospheres for a given distance between particles. Consequently, the repulsion forces between the particles decrease with increases in the salt concentration and in the valence of the adsorbed ions. Thus, a high level of repulsion and swelling persist in dilute solutions of sodium clay (Rengasamy and Sumner, 1998; Van Olphen, 1977).

Hydration repulsion forces

The adsorbed cations and the clay surfaces must lose some of their hydration water, because the clay particles approach each other very closely in the process of flocculation. The work that is needed to remove these water molecules manifests itself as the repulsion energy. Relative to the double-layer repulsion forces, the range of these forces is short. These forces partially negate van der Waals forces of attraction, resulting in particles not being trapped at a true primary minimum. This makes it easier for particles to disperse with little input of energy as the electrolyte concentration is reduced (Van Olphen, 1977; Shainberg and Kemper, 1996).

Flocculation and attraction forces

A stable clay suspension flocculates upon the addition of salt. Flocculation takes place only when a threshold electrolyte concentration, the 'flocculation value' (FV), is applied (Hunter, 1987). A suspension of clay platelets flocculates in three possible modes of particle association (Fig. 2): (1) association between flat oxygen planes of two parallel platelets (face to face, FF); (2) association between edge surface of neighboring particles (edge to edge, EE); and (3) association between edge and oxygen planar surface (edge to surface, EF) (Van Olphen, 1977; Goldberg et al., 2012). In the presence of small concentrations of NaCl, the diffuse double-layer repulsion prevents particle association and a stable suspension is obtained. As the concentration of NaCl increases, double layers at soil surfaces become compressed and, at the FV, attraction forces may prevail and EF, EE, and FF association may occur (Van Olphen, 1977). Similarly, when polyvalent cations are adsorbed (e.g., Ca and Mg) on the clay surface, the diffuse double layer is compressed and attraction forces may prevail (Quirk and Schofield, 1955). Although the three types of association are really three modes of flocculation, only the EE and the EF types of association were discussed in the past and were labeled clay flocculation with high gel volume. The thicker particles, which result from simple FF association, were not called 'flocs,' because they were condensed with low gel volume. Special cases of FF association, which lead to high-volume flocs, have now been described (Fig. 3) and it has been concluded that FF association may also lead to high-volume flocculation (Keren et al., 1988).

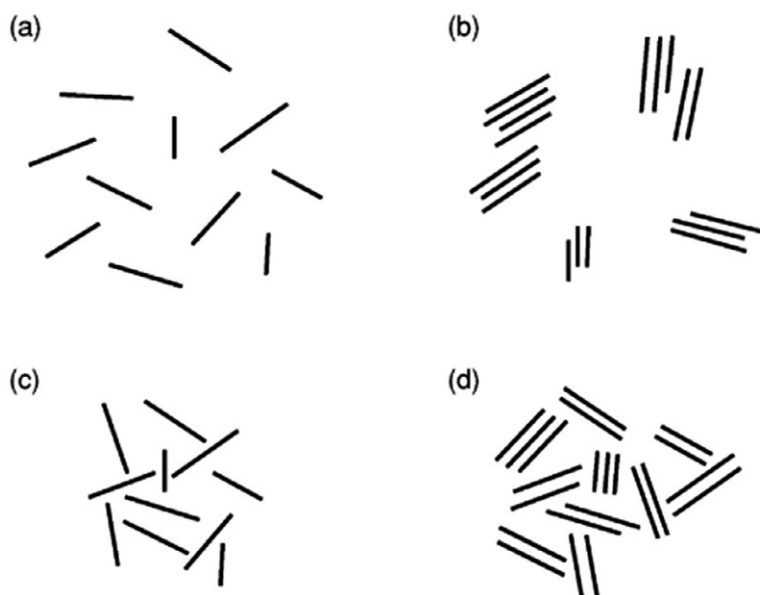


Fig. 2 Schematic representation of modes of particle flocculation: (a) dispersed; (b) face-to-face flocculation; (c) edge-to-face flocculation; (d) face-to-face and edge-to-face flocculation. After Van Olphen H (1977) *An Introduction to Clay Colloid Chemistry*, 2nd edn. New York, NY: John Wiley.

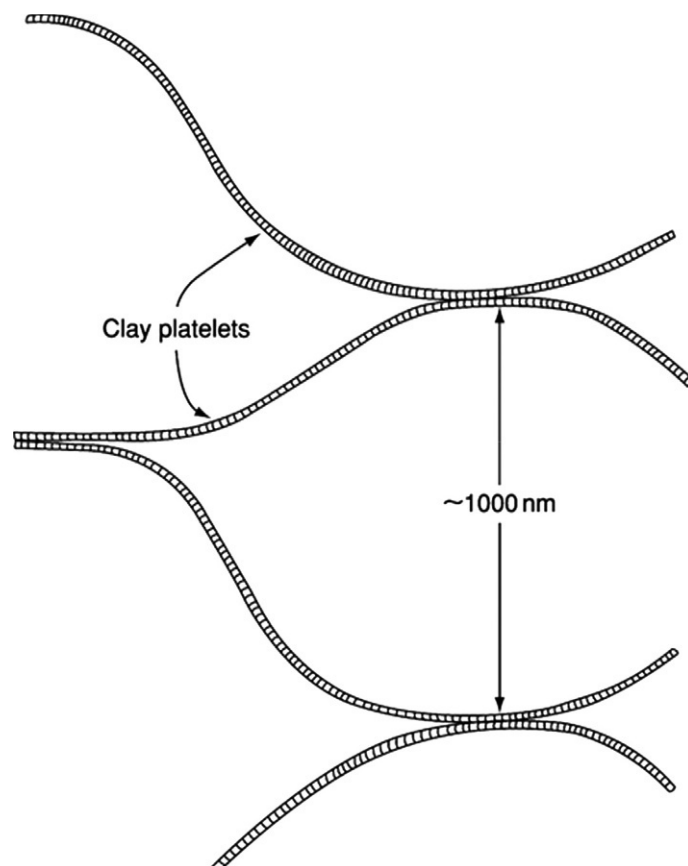


Fig. 3 Mode of platelet association (face-to-face) in sodium montmorillonite suspension at moderate electrolyte concentration (e.g., 44 mmol dm^{-3} NaCl concentration). Reproduced with permission from Keren R, Shainberg L, and Klein E (1988) Settling and flocculation value of Na-montmorillonite particles in aqueous media. *Soil Science Society of America Journal* 52: 76–80.

Van der Waals force of attraction

The van der Waals attractive force between colloidal particles arises from the fluctuating dipole of one atom that polarizes another, and the two atoms attract each other. This attraction between atom pairs is additive and therefore the energy of interaction between particles decreases much more slowly with distance than that between individual atoms. The van der Waals attraction force leads to FF clay association (Israelachvili, 1992).

Electrostatic EF attraction forces

The flat surfaces are not the only surfaces of the plate-like clay particles; they also expose an edge surface area. At the edges, the octahedral alumina or iron sheets are disrupted and primary bonds are broken. This situation is analogous to that on the surface of alumina or iron particles in alumina (Al_2O_3), and iron sols. On such surfaces an electric double layer is created by the adsorption of potential-determining ions. The hydroxyl-aluminum edge carries a positive charge in acid solution and a negative charge in high-pH solution. The point of zero charge (pzc) does not necessarily occur at neutral pH. The charge becomes more positive with decreasing pH, and its sign may be reversed with increasing pH. When the charge on the edges is positive and the repulsive diffuse double-layer forces are diminished by a high concentration of electrolytes, a positive edge to negative-face surface predominates and EF flocculation occurs (Keren et al., 1988; Quirk, 1994).

Clay charge density and electrostatic attraction forces

The colloid properties of clays depend to a large extent on the density of charge on the clay surface, that in turn affects the electrostatic attractions forces between the clay colloids. Numerous detailed discussions on this issue can be found in the literature (e.g., Keren et al., 1988; Rengasamy and Sumner, 1998; Goldberg et al., 2012 and references cited therein). Clay surfaces charge density can be derived from the exchange capacity per gram and the surface area per gram (Table 1). Sodium montmorillonite in dilute suspensions exists as single platelets. Calcium montmorillonite, even in distilled water, exists in packets (tactoids or quasicrystals), each consisting of several (4–9) clay platelets with a film of water 0.45 nm on each internal surface and c-spacing of less than 1.9 nm. The swelling pressure of calcium montmorillonite can be predicted from the diffuse double-layer theory if one assumes that the number of platelets in a tactoid is 5 and that only the Ca ions on the external surfaces of the tactoids form a diffuse double layer. Sodium, magnesium, and calcium vermiculite do not swell beyond 1.5 nm (two molecular layers of water between the platelets), even in distilled water. The illites exist in quasicrystals with c-spacing of less than 1 nm regardless of the adsorbed cations. It has been suggested that electrostatic attraction forces, which increase with the increase in surface charge density and with increase in cation valence, prevent swelling in clays with high surface charge density.

The charge density at clay surfaces is an average value, and heterogeneity of the charge density due to isomorphous substitution has been determined for montmorillonites ranging between 0.7×10^{-3} and $1.5 \times 10^{-3} \text{ mmol}_c \text{ m}^{-2}$. Wyoming bentonite is the most heterogeneous among six various clay types. It is possible, therefore, that at specific sites on the planar surfaces the charge density is very low or even zero (the surface area of these particular sites may not be detected). Thus, in the extreme situation where the charge density is very low (or even zero, such as in pyrophyllite), there are no repulsion forces (both double layer and hydration), and the van der Waals attraction forces exceed the repulsion forces and an FF association may occur (Fig. 3). Similarly, in sites where the charge density is large (such as in vermiculite), the electrostatic attraction forces between the cations and the clay platelets are also very high, swelling decreases, and an FF association may occur (Fig. 3). Since the montmorillonite platelets are long and flexible and able to bend, the platelets may form cohesive junctions at several locations of low or high surface charge densities to form a network structure, as illustrated in Fig. 3. This results in a high gel volume, as demonstrated for Na-montmorillonite at pH 9.8 (Fig. 4).

Spontaneous and mechanical dispersion

Spontaneous dispersion often takes place without energy inputs when sodic clay is placed in water at very small electrolyte concentration. The proportion of clay particles separated in this way increases with an increase in the sodium adsorption ratio (SAR) of the soil solution and with a decrease in the electrolyte concentration (Rengasamy and Sumner, 1998).

Hydrated clay particles in equilibrium with low-SAR solution, which have undergone limited separation, can be pushed further apart by applying external mechanical pressure such as stirring or raindrop impact. When the SAR of the clay solution is low to moderate and the electrolyte concentration is large enough to prevent spontaneous dispersion, the introduction of a mechanical force, such as mixing the suspension, may cause clay dispersion (Shainberg and Levy, 2005). The clay at the soil surface is especially susceptible to dispersion because of the mechanical energy applied by raindrops (Agassi et al., 1981). This is one of the reasons why the soil surface is more susceptible to clay dispersion and aggregate disintegration by low sodic conditions, compared with deeper in the soil profile.

Flocculation and dispersion in Na or Ca clays

Adsorbed Na ions form a diffuse double layer at the clay surface, and, because of the high swelling between the platelets, single platelets tend to persist in dilute solution. When salt is added to the clay suspension, the particles stick together upon collision, forming flocs that settle (Oster et al., 1980; Goldberg et al., 2012). The suspension is then separated into the bottom sediment and a particle-free supernatant liquid. The FVs of clays depend on counterion valency. The FVs of sodium and calcium montmorillonite

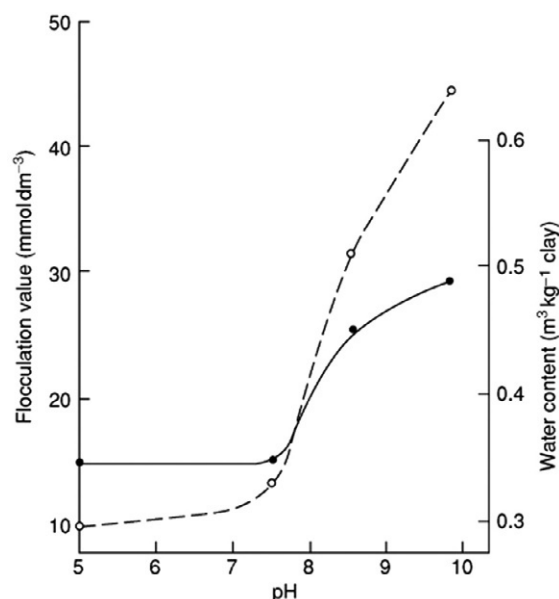


Fig. 4 The flocculation value (dashed line) and gel volume (solid line) of sodium montmorillonite in NaCl solution as affected by suspension pH (the clay concentration was 0.1% and the apparent size of the platelets was less than 2 μm). Reproduced with permission from Keren R, Shainberg L, and Klein E (1988) Settling and flocculation value of Na-montmorillonite particles in aqueous media. *Soil Science Society of America Journal* 52: 76–80.

Table 2 Selected critical flocculation concentrations for some clay minerals.

Mineral	CFC (mol m^{-3})	Background electrolyte	pH	Suspension strength (g kg^{-1})
Georgia kaolinite	5	NaCl	7	0.6
	245	NaHCO_3	8.3	0.6
	0.4	CaCl_2	7	0.6
Montmorillonite	12	NaCl	?	1
	0.13	CaCl_2	?	1
Vermiculite	38	NaCl	7	0.6
	58	NaHCO_3	8.3	0.6
	0.8	CaCl_2	7	0.6
Grundy illite	7.24	NaCl	6	1
	0.2	CaCl_2	6	1
Illite 36	9	NaCl	7	0.6
	185	NaHCO_3	8.3	0.6
	0.13	CaCl_2	7	0.6

CFC, critical flocculation concentration.

Based on data from Goldberg S, Lebron I, and Suarez DL (1999) Soil colloidal behavior. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. B195–B240. Boca Raton, FL: CRC Press; Oster JD, Shainberg I, and Wood JD (1980) Flocculation value and gel structure of Na/Ca montmorillonite and illite suspensions. *Soil Science Society of America Journal* 44: 955–959.

are 12 $\text{mmol}_c \text{L}^{-1}$ NaCl and 0.25 $\text{mmol}_c \text{L}^{-1}$ CaCl_2 , respectively (Table 2). The sodium montmorillonite gel has approximately 330 g of water per gram of clay, which corresponds to a film thickness of 440 nm per surface (Fig. 4). This thickness exceeds the range of diffuse double-layer or van der Waals forces. The open structure, typical of sodium montmorillonite gels, suggests that EF attraction forces or the flexible platelets mechanisms (Fig. 3) are operating (Keren et al., 1988).

Effect of pH on the FV of sodium montmorillonite

The FV of sodium montmorillonite by NaCl is pH dependent and its value is 12.6 $\text{mmol}_c \text{L}^{-1}$ at pH 7, and 10, 13, 31, and 44 $\text{mmol}_c \text{L}^{-1}$ of NaCl at pH 5, 7.5, 8.5, and 9.8, respectively (Fig. 4). The final gel volume of sodium montmorillonite samples at the above pH values is large (350–490 g of water per gram of clay) (Fig. 4). The large gel volume indicates that a three-dimensional ‘card house’ structure is formed, even at pH 9.8, in spite of the fact that at pH 9.8 a negative charge predominates at the broken edges. This considerable gel volume at pH 9.8 suggests that an open structure with FF association between the platelets (at sites of low and high charge density) predominates rather than the EF card house structure (Keren et al., 1988; Goldberg et al., 2012).

Dispersion and flocculation in Na or Ca clay systems

Whereas single platelets tend to persist in a dilute solution of sodium montmorillonite, calcium montmorillonite platelets aggregate into tactoids or quasicrystals. Each tactoid consists of several (4–9) clay platelets in parallel array, with interplatelet distances of 0.9 nm. A diffuse double layer exists only on the external surfaces of the tactoids. In a suspension containing a mixture of Na and Ca ions, 'demixing' of the cations occurs so that some interlayer spaces contain mainly Na ions and others mainly Ca ions (Shainberg and Letey, 1984). An indication of the location of the adsorbed ions in montmorillonite clay saturated with a mixture of mono- and divalent cations is obtained by measurements of the electrophoretic mobility of clay particles in suspension (Bar-On et al., 1970). A slight addition of exchangeable Na (e.g., 10%) to Ca-saturated clay has a considerable effect on the electrophoretic mobility of the clay (Fig. 5). When the exchangeable sodium percentage (ESP) in a clay suspension is less than 40%, the Na ions concentrate on the external surfaces of the tactoids and the Ca ions concentrate on the internal surfaces (Shainberg and Letey, 1984).

Results of studies of the FVs of montmorillonite and illite suspensions saturated with mixtures of Na and Ca ions in the exchange phase (Fig. 6) indicate that a small addition of exchangeable Na to Ca-saturated montmorillonite can have a considerable effect on the FV (Oster et al., 1980). Similarly, when the ESP of the clay suspension reaches 50, the electrophoretic mobility of the Na-Ca mixture reached that of pure sodium montmorillonite (Fig. 5). Also, the gel volume increases with an increase in SAR until it reaches the gel volume of pure sodium montmorillonite at SAR 50 (Fig. 7). The demixing model, which suggests that Na ions concentrate on the external surfaces of the tactoids and increase the flocculation values of the clay considerably, can be used to explain these results (Shainberg and Letey, 1984). Yet in clay suspensions with ESP of more than 50, disintegration of the tactoids is complete and single platelets predominate (Shainberg and Letey, 1984).

The effect of peptizing agents on dispersion

Sodium polymetaphosphate $((\text{NaPO}_3)_{13})$ is a very effective peptizing agent. Organic anions may also act as peptizing agents. The FV of peptizer-treated clay suspensions increases rapidly with increasing peptizer concentration. The anions are probably chemisorbed at the edges of the clay particles by reacting with the exposed Al. Thus, the charge of the edges becomes negative, which prevents the EF bond from taking place (Frenkel et al., 1992). If the EF bond is eliminated by the adsorption of polymetaphosphate, then one might expect that the open, card-house structure of the gel is also eliminated. However, an open structure with large gel volume has been obtained for the sodium gel (Table 3). This also indicates that an FF bond may lead to open gel structure (Van Olphen, 1977; Shainberg and Letey, 1984).

Soil clay systems

Flocculation and dispersion behavior of soil clays differs significantly from that of pure clay systems, possibly because soil clays usually occur as mixtures and because of their association with other minerals and organic matter present in the soil. A few studies demonstrate that the FV values of soil clays is two- to tenfold higher than for pure clay systems (Frenkel et al., 1992; Fig. 8). Conversely, in the presence of hydrous oxides or sparingly soluble minerals such as CaCO_3 , clay dispersivity is reported to be less severe (Oster and Shainberg, 1979). Hence, extrapolation from pure clay systems is of limited practical value.

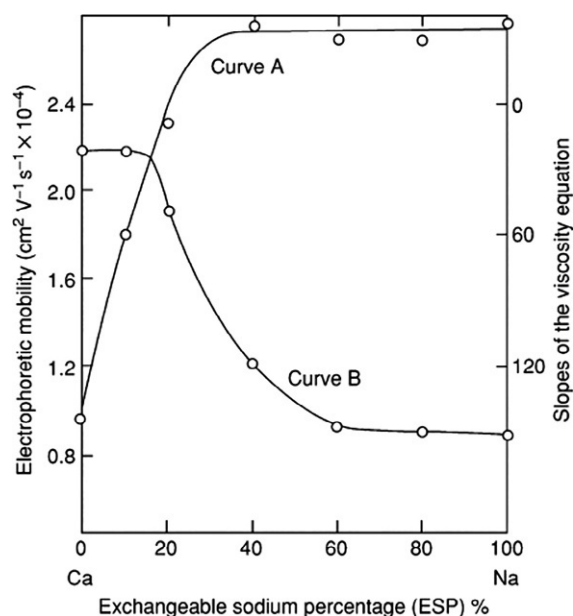


Fig. 5 Dependence of electrophoretic mobility (curve A) and the relative size (curve B) of montmorillonite particles on the exchangeable sodium percentage (ESP). The relative size is expressed in units of the slope in Einstein's equation for the viscosity of suspension. Reproduced with permission from Shainberg I and Letey J (1984) Response of soils to sodic and saline conditions. *Hilgardia* 52: 1–57.

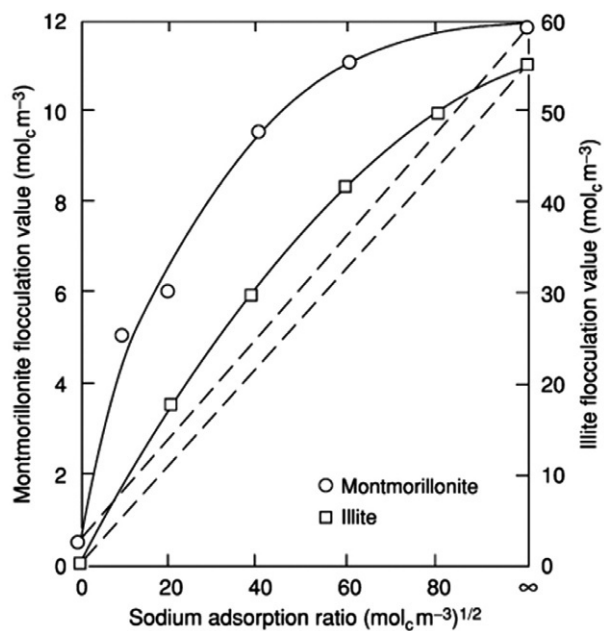


Fig. 6 Flocculation value for Wyoming montmorillonite (API 25) and Fithian illite (API 35) as a function of the sodium adsorption ratio. Reproduced with permission from Shainberg I and Letey J (1984) Response of soils to sodic and saline conditions. *Hilgardia* 52: 1–57.

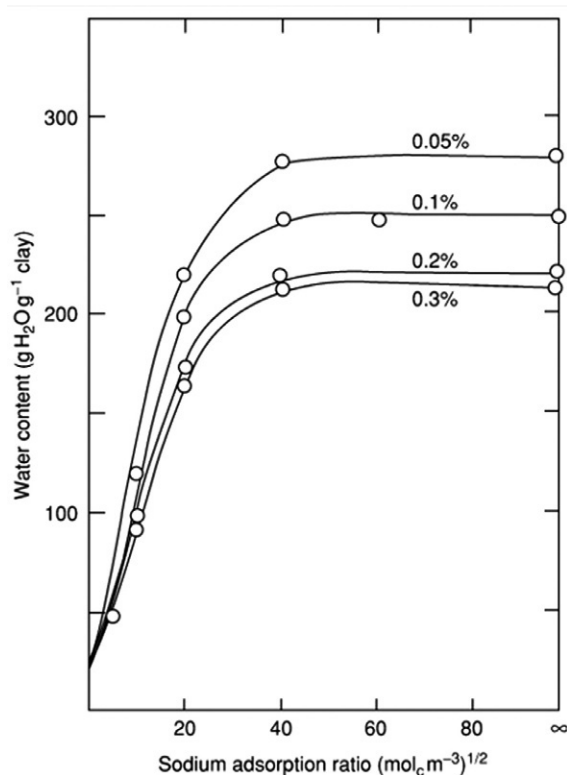


Fig. 7 Water content of Wyoming montmorillonite gel as a function of Na adsorption ratio and suspension clay content. Reproduced with permission from Oster JD, Shainberg I, and Wood JD (1980) Flocculation value and gel structure of Na/Ca montmorillonite and illite suspensions. *Soil Science Society of America Journal* 44: 955–959.

Table 3 The effect of sodium polymetaphosphate ($(\text{NaPO}_3)_{13}$ $\frac{1}{4}$ 102×13) on the flocculation value and the water content of the montmorillonite gel prepared from suspensions that contained 0.1% and 0.2% clays.

Peptizer concentration		NaCl flocculation value ($\text{mol}_c \text{ m}^{-3}$)	Water content of gel ($\text{g H}_2\text{O g}^{-1}$ clay)			
			With peptizer		Without peptizer	
			0.1%	0.2%	0.1%	0.2%
ppm	$\text{mol}_c \text{ m}^{-3}$					
0	0	12	240	220	240	220
1	0.01	20	260	235	200	190
10	0.1	80	170	180	150	140
100	1.0	120	160	165	130	130

Reproduced with permission from Oster JD, Shainberg I, and Wood JD (1980) Flocculation value and gel structure of Na/Ca montmorillonite and illite suspensions. *Soil Science Society of America Journal* 44: 955–959.

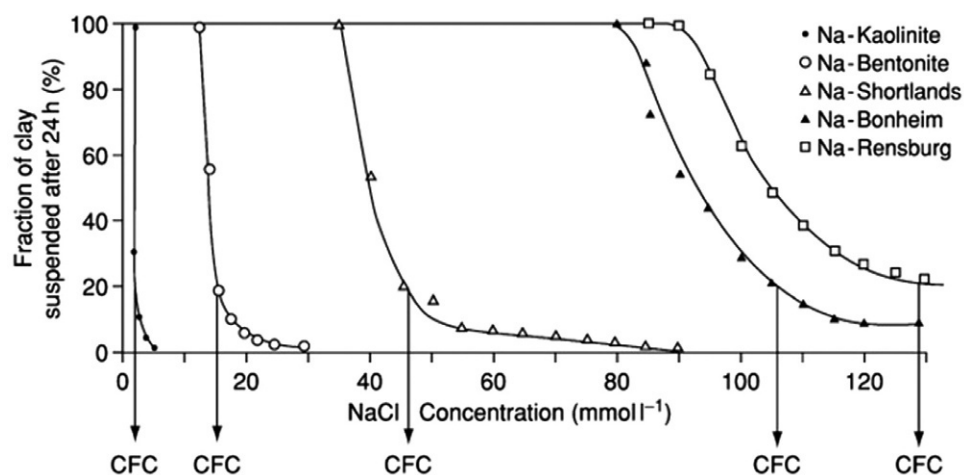


Fig. 8 Flocculation behavior of Na clays suspended in NaCl solution. Reproduced with permission from Frenkel H, Fey MV, and Levy GJ (1992) Organic and inorganic anion effects on reference and soil clay critical flocculation concentration (CFC). *Soil Science. Society of America Journal* 56: 1762–1765.

Dispersion and the hydraulic properties of soils

Shainberg and Oster (1978) highlighted the importance of clay dispersion and movement in determining the hydraulic conductivity of soils. In this experiment a noncalcareous loam with 20% clay (montmorillonite) and a calcareous clay soil (50%, montmorillonitic) were equilibrated with a solution of $20 \text{ mmol}_c \text{ L}^{-1}$ with a SAR of 10 to give soils with an ESP of 10. Following equilibration, the soils were leached with distilled water. The loam was completely sealed and the clay soil maintained a relative hydraulic conductivity of about 20% (Fig. 9). This observation contradicts the generalization that soils having a greater content of expansive clays are more susceptible to sodic conditions because of more swelling. In the calcareous clay soil, salt concentration in the soil solution was maintained at $4 \text{ mmol}_c \text{ L}^{-1}$ by dissolution of carbonates. At ESP 10, this concentration prevents clay dispersion, and the hydraulic conductivity decrease is due only to swelling. Conversely, in the noncalcareous loam, soil solution concentration was below $1 \text{ mmol}_c \text{ L}^{-1}$, thus clay dispersion occurred and the dispersed clay clogged the conducting pores (Shainberg and Oster, 1978).

The infiltration rate of the soil surface, exposed to rain, is more susceptible to dispersion and sodic conditions than the hydraulic conductivity of the soils (Fig. 10). Furthermore, the infiltration rates of calcareous and noncalcareous soils are equally sensitive to low levels of exchangeable sodium (Kazman et al., 1983), whereas the hydraulic conductivity of the calcareous soils is less sensitive to low ESP than that of the noncalcareous soils (Shainberg and Oster, 1978). Both phenomena are explained by the clay-dispersion mechanism. At the soil surface, the concentration of the soil solution is determined solely by rain water and the low concentration of the salt enhances clay dispersion. Also, the impact of raindrops enhances clay dispersion. Both mechanisms enhance clay dispersion, aggregate disintegration, low infiltration rate, and high runoff (Shainberg and Letey, 1984).

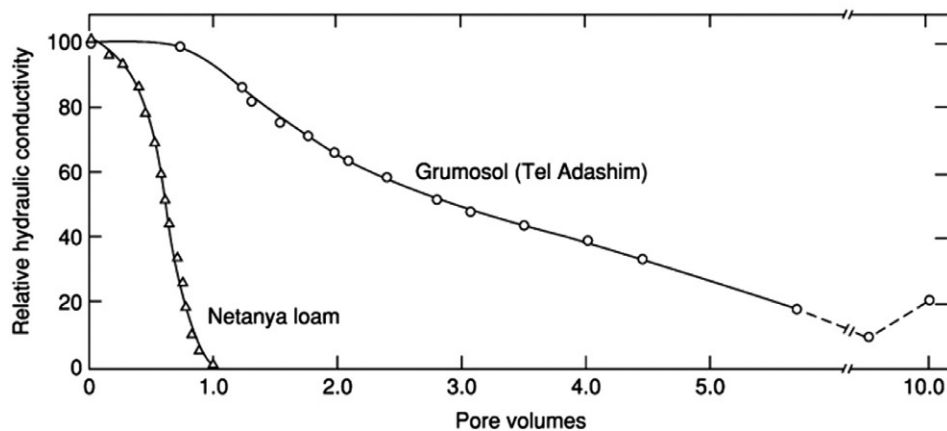


Fig. 9 Hydraulic conductivity of two montmorillonite soils having an ESP of 10 and leached with distilled water (Shainberg and Oster, 1978). Reproduced with permission from Shainberg I and Lety J (1984) Response of soils to sodic and saline conditions. *Hilgardia* 52: 1–57.

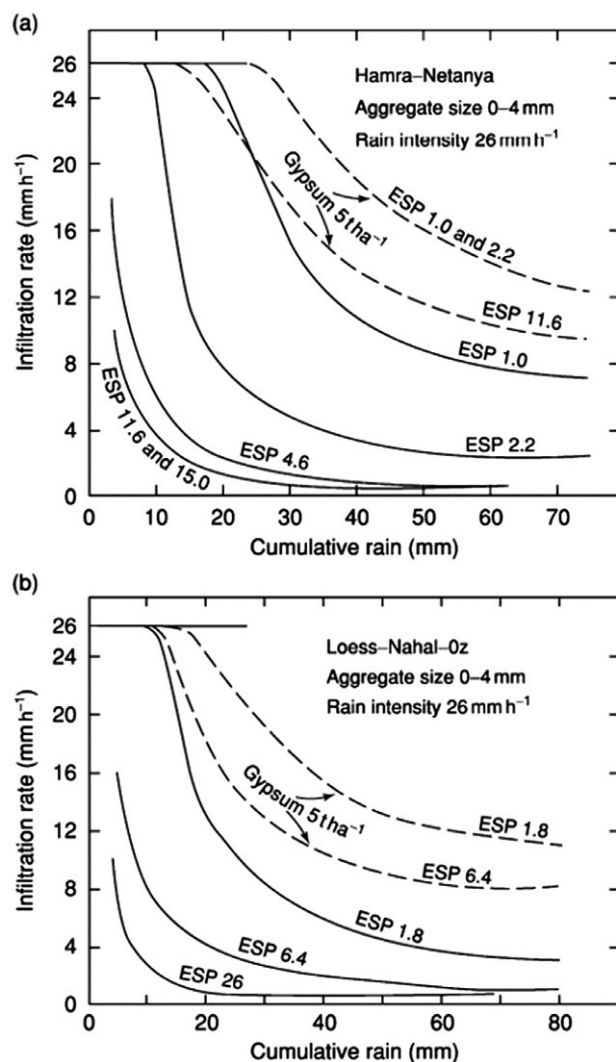


Fig. 10 Effect of soil ESP (and phosphogypsum) on the infiltration rate of the (a) Hamra (Netanya) and (b) loess soil. Adapted from Shainberg I and Lety J (1984) Response of soils to sodic and saline conditions. *Hilgardia* 52: 1–57, with permission.

Summary

The dispersion and flocculation of clays depend on the mineralogy of the clay and are affected by the composition of exchangeable cations, by the concentration of salt, by pH, by the presence of peptizing agents, and by the application of a stirring mechanism. In a mixed Na-Ca suspension, the flocculation value of the clay increases sharply with the introduction of a small amount of sodium and with an increase in pH. Dispersion, movement, and deposition of soil colloids alter the geometry of soil pores and affect the hydraulic properties of soils. Thus, introducing a small amount of Na to soils exposed to leaching with rainwater may cause clay dispersion and a deterioration of soil physical properties.

References

- Agassi M, Shainberg I, and Morin J (1981) Effect of electrolyte concentration and soil sodicity on infiltration rate and crust formation. *Soil Science Society of America Journal* 45: 848–851.
- Babcock KL (1963) Theory of chemical properties of soil colloidal systems at equilibrium. *Hilgardia* 34: 417–542.
- Bar-On P, Shainberg I, and Michaeli I (1970) The electrophoretic mobility of Na/Ca montmorillonite particles. *Journal of Colloid and Interface Sciences* 33: 471–472.
- Frenkel H, Fey MV, and Levy GJ (1992) Organic and inorganic anion effects on reference and soil clay critical flocculation concentration. *Soil Science Society of America Journal* 56: 1762–1766.
- Goldberg S, Lebron I, and Suarez DL (1999) Soil colloidal behavior. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. B195–B240. Boca Raton, FL: CRC Press.
- Goldberg S, Lebron I, Seaman JC, and Suarez DL (2012) Soil colloidal behavior. In: Huang PM, Li Y, and Sumner ME (eds.) *Handbook of Soil Science, Properties and processes*, pp. 15–1–15–39. Boca Raton, FL: CRC Press.
- Hunter RJ (1987) *Foundations of Colloid Science*. London, UK: Academic Press.
- Israelachvili JN (1992) *Intermolecular and Surface Forces*, 2nd ed. San Diego, CA: Academic Press.
- Kazman Z and Shainberg I (1983) Effect of low levels of exchangeable Na and applied phosphogypsum on the infiltration rate of various soils. *Soil Science* 35: 184–192.
- Keren R, Shainberg I, and Klein E (1988) Settling and flocculation value of Na-montmorillonite particles in aqueous media. *Soil Science Society of America Journal* 52: 76–80.
- Oster JD and Shainberg I (1979) Exchangeable cation hydrolysis and soil weathering as affected by exchangeable sodium. *Soil Science Society of America Journal* 44: 429–433.
- Oster JD, Shainberg I, and Wood JD (1980) Flocculation value and gel structure of Na/Ca montmorillonite and illite suspensions. *Soil Science Society of America Journal* 44: 955–959.
- Quirk JP (1994) Interparticle forces: A basis for the interpretation of soil physical behavior. *Advances in Agronomy* 53: 121–183.
- Quirk JP and Schofield RK (1955) The effect of electrolyte concentration of soil permeability. *Journal of Soil Science* 6: 163–178.
- Rengasamy P and Sumner ME (1998) Processes involved in sodic behavior. In: Sumner ME and Naidu R (eds.) *Sodic Soils*. New York: Oxford University Press.
- Shainberg I and Kemper WD (1996) Hydration status of absorbed ions. *Soil Science of Society of America Proceedings* 30: 707–713.
- Shainberg I and Letey J (1984) Response of soils to sodic and saline conditions. *Hilgardia* 52: 1–57.
- Shainberg I and Levy GJ (2005) Flocculation and dispersion. In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*. vol. 2, pp. 27–34. Oxford, UK: Elsevier Ltd., (ed-in-chief).
- Shainberg I and Oster JD (1978) *Quality of Irrigation Water*. IIC Publication No. 2. P.O. Box 6, Bet Dagan, Israel: Volcani Center.
- Stul MS and Mortier WJ (1974) The heterogeneity of the charge density in montmorillonite. *Clays and Clay Minerals* 22: 391–396.
- Van Olphen H (1977) *An Introduction to Clay Colloid Chemistry*, 2nd edn. New York, NY: John Wiley.

Soil organic matter dynamics

Yu Luo and Jianming Xu, Institute of Soil & Water Resources and Environmental Science, College of Environmental and Resource Sciences, Zhejiang Provincial Key Laboratory of Agricultural Resources and Environment, Zhejiang University, Hangzhou, China

© 2023 Elsevier Ltd. All rights reserved.

Introduction	388
Decomposition and factors	388
Recalcitrance of chemical compounds	388
Low-molecular-weight organic substances	389
Recalcitrant compounds (cellulose, lignin)	389
Pyrogenic organic matter	389
Role of organisms in decomposition	389
Bacteria and fungi	390
Fauna (protist, nematodes, collembola, earthworms)	390
Formation of SOM	390
Genesis and pathways	390
Plant-derived vs microbially-derived organic matter	390
Microbial necromass and its accumulation in soils	390
Carbon flow through the atmosphere-plant-microbe-soil continuum	390
Abiotic drivers	391
Climate	391
Soil characteristics	391
Agricultural practices	391
Particulate organic matter and mineral associated organic matter	391
Fungal aggregate–mineral interactions underlying carbon stabilization	392
Approaches applied in SOC studies	392
Biomarkers	392
Isotopic approaches	393
Tracing carbon flows and nutrient uptake isotopes	393
Partitioning soil CO ₂ emissions among various C sources	393
Outlook: New paradigms for SOC research	393
Recalcitrant nature of organic matter and microbial accessibility	393
Traits-based strategy	394
Conclusion	394
References	395

Abstract

Soil organic carbon (SOC) constitutes the largest carbon (C) stock in terrestrial ecosystems, and a rigorous understanding of the key processes involved in SOC dynamics is vital for soil fertility, soil health and a climate change mitigation strategy. Here, we review (i) the balance between C input and output, i.e., stabilization vs mineralization, which are fundamental processes upon which C content in soils ultimately depends; (ii) the controlling factors of SOC formation and decomposition, e.g., the climate, and physical, chemical and biological soil properties. Finally, classic and cutting-edge methods such as biomarker and isotope tools that are applied in SOC studies are introduced, and, the outlook for future research related to SOC is provided.

Key points

- Carbon (C) enters the soil from autotrophic microbes and plant photosynthesis.
- Gain of plant derived C in soil is by physical transfer and microbial necromass.
- C is lost to the atmosphere by microbial mineralization.
- New substrate-C induces native SOC priming by co-metabolism.
- SOC (Soil organic C) content is determined by the C balance between sequestration and losses (i.e., stabilization vs mineralization).
- Changes in C content and composition affect multiple physical, chemical and biological soil properties.
- The processes of mineralization involve both abiotic and biotic factors, including substrate recalcitrance, the soil matrix and the microbial community.

- The role of organisms, including bacteria, fungi, protists, nematodes, collembola, and earthworms, in decomposition are discussed.
- Classification of SOC into POM (particulate OM) and MAOM (mineral associated OM) advances our conceptual understanding of the biophysical mechanisms underpinning SOM turnover,
- Fungal aggregate–mineral interactions underlying carbon stabilization.
- Approaches of biomarkers and isotopes applied in SOC studies are introduced.

Introduction

Soil organic matter (SOM) is a key indicator of soil fertility and health, and changes in it will inevitably affect multiple physical, chemical and biological soil properties, including soil structure, nutrient availability, and biological activity and diversity. Soil organic carbon (SOC) constitutes the largest carbon (C) stock in terrestrial ecosystems, accounting for more than 1500 Pg C globally. Global SOC storage to a soil depth of 1 m is double the atmospheric C pool and treble the vegetation C pool. A critical component of global C sequestration is to increase SOC storage. Increasing SOC accumulation in terrestrial ecosystems is of vital importance for addressing the major challenge of global climate change resulting from increasing atmospheric CO₂ concentrations.

Soil organic C stocks involve the balance between two processes: (i) C enters the soil from plant-derived photosynthesis, and (ii) C is lost to the atmosphere by microbial mineralization (Fig. 1). The gain and loss of C are fundamental processes upon which all maintenance of life and C sequestration in soils ultimately depends. The SOC content is determined by the C balance between sequestration and losses (i.e., stabilization vs mineralization). A rigorous understanding of the key processes involved in SOC stabilization (e.g., entombing effects) and mineralization (e.g., priming effects) is vital for managing SOC pools as a strategy for climate change mitigation.

Decomposition and factors

The processes of mineralization involve both abiotic and biotic factors, including substrate recalcitrance, the soil matrix and microbial community (Lehmann and Kleber, 2015).

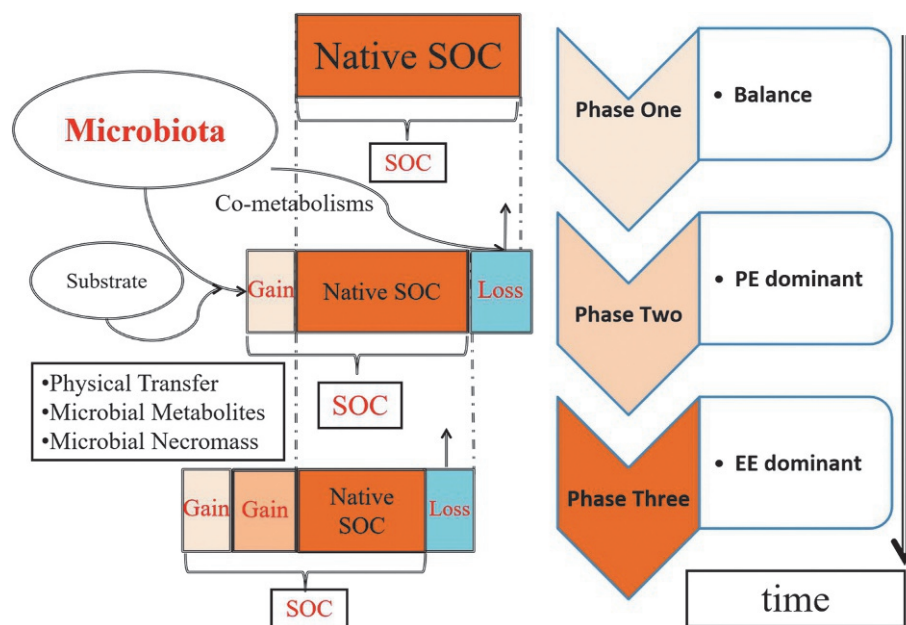


Fig. 1 Soil organic carbon stocks depend on the balance between carbon input (EE: entombing effects) and output (PE: priming effects). The SOC dynamics involves three phases, (i) balanced stage: C enters the soil from autotrophic microbes and plant photosynthesis equal to C loss via plant and microbial respiration; (ii) priming effects dominant stage: larger loss of new C and native SOC through microbial utilization and priming due to co-metabolisms; and (iii) entombing dominant stage: larger gain of C in soil via physical transfer, excreted metabolites, and microbial necromass.

Recalcitrance of chemical compounds

Soil organic C is largely a structurally unquantified pool, comprising random condensation products of microbial metabolites. While it has no clearly defined chemical structure, it can be characterized operationally as labile (available for microorganism metabolism) and recalcitrant (resistant to microbial degradation) components. For example, litter decomposition depends on the complexity and bioavailability of organic substances as well as the functional groups. As decomposition proceeds, soluble components are lost followed by the depolymerization of free cellulose and hemicellulose. Finally, the acid unhydrolyzable (AUR) fraction, considered a proxy for lignin and lignin-encrusted cellulose, starts to mineralize. Simple molecules, such as glucose can be degraded within minutes, whereas complex macromolecules containing covalent bonds between hemicellulose and lignin are slowly mineralized over months to years.

Low-molecular-weight organic substances

Numerous studies have investigated the microbial community response to plant-derived low-molecular-weight organic substances (LMWOS). Microorganisms in the rhizosphere have distinct substrate preferences in their responses and utilization of contrasting LMWOS as characterized by enzyme activities. Although the incorporation of plant-derived compounds (sugar, lipids, etc.) in soil has been widely studied, the identity of microorganisms that are stimulated by LMWOS and their metabolic processes remain poorly described. The stable-isotope probing combined with DNA is an effective tool for identifying microbial activities and functions in tracing the fate of LMWOS, such as glucose, methanol, p-hydroxybenzoic acid and propionate.

Recalcitrant compounds (cellulose, lignin)

Lignin is highly resistant to degradation due to its structural complexity. Yet, it can still be utilized by selected microbial communities, particularly some specialized fungi. The phylum Basidiomycota is associated with the decomposition of lignin. Complex and recalcitrant compounds in plant residues, such as lignin and cellulose, are not readily available for bacterial decomposition, whereas fungi release extracellular hydrolytic and oxidative enzymes that target specific structural polymers resulting in smaller, more “bacterially accessible” products. Specifically, the cellulolytic and ligninolytic capacities of the microbial community are responsible for the decomposition of cellulose and lignin, respectively. Crawford et al. (1977) showed that ^{14}C -labeled cellulose decomposed four times as fast as lignin. Real-time PCR was used to follow the abundance of genes and coding for cellobiohydrolases (*cbhI*) and lignobiohydrolases (*lcc*) identified that the phylum Ascomycota was gradually replaced by Basidiomycota during the decomposition process; the latter being able to synthesize enzymes to degrade polymers of lignin.

Pyrogenic organic matter

Pyrogenic organic matter (PyOM), the product of pyrolyzed organic materials, has drawn extensive attention as a promising material to enhance C sequestration because of their resistance to degradation. In the case of wildfires, alteration of the physico-chemical structure of plant-derived C to highly condensed polyaromatic organic matter promotes C sequestration in post-fire forest soils. Even though PyOM is relatively stable, it can still be partially mineralized. The decomposition of biochar was relatively rapid during the first 3 months following its incorporation into soils, but slowed considerably during the following 3.2 years. The initial soil CO_2 immediately following PyOM is often considered to be derived from the mineralization of labile components and often short-lived. The intensity of the CO_2 efflux is closely related to the physicochemical characteristics of PyOM, which were largely dependent on feedstock type and pyrolysis temperature. Differences in pyrolysis temperatures result in variation in PyOM chemical structure and composition, which largely explains biochar stability and rates of C mineralization in soil. The PyOM often has a greater abundance of recalcitrant components (aromatic density) with increasing pyrolysis temperature, making them more recalcitrant to microbial decomposition.

Although PyOM increases the SOC content, it promotes contradictory effects on non-biochar C dynamics, such as (i) interacting with plant-derived C (rhizodeposits;), and (ii) causing loss or gain of native SOC. Following PyOM amendment to soil, both negative priming effects (PEs, i.e., decreased CO_2 release) and positive PEs (i.e., increased CO_2 release) have been reported. Positive PEs in particular are of concern because they offset the C sequestration effects of PyOM additions. The mechanisms involved in PyOM-induced SOC mineralization are attributed to shifts in the soil microbial community through changes in abiotic variables, such as acid neutralization via increasing soil pH (Luo et al., 2017). Microbial biomass, activity and enzyme production were increased in the charosphere, the interface between the PyOM and bulk soil, and can be largely ascribed to (i) improved microbial habitat afforded by the large porosity and surface area of PyOM, and (ii) increased supply of labile organic components or other nutrients contained within biochar resulting in stimulation of the microbial community.

Role of organisms in decomposition

Organic substances in soils annually release an estimated 58 Pg C into the atmosphere through microbial decomposition. To underpin microbial community dynamics during litter or straw decomposition, it is crucial to elucidate the underlying mechanisms by which organic residues affect soil C sequestration. Understanding microbial communities and their utilization of rhizo-C and litter and straw is crucial for optimizing the sequestration of photosynthesized-C inputs into soils.

Bacteria and fungi

Straw and litter are model plant-derived substrates containing lignin, cellulose and hemicellulose. Decomposition of the primary constituents of straw is a complex process that depends on a progression of decomposers as decomposition proceeds. The diversity of bacteria and fungi, often regarded as the major drivers for decomposition of the labile and recalcitrant fractions of plant residues, have been frequently monitored throughout the decomposition process. Complex and recalcitrant compounds in plant residues, such as lignin and cellulose are not readily available for bacterial decomposition, whereas fungi release extracellular hydrolytic and oxidative enzymes that target specific structural polymers, thereby creating smaller, more “bacterially accessible” products. Microbial succession with a shifting enzymatic production exerted a positive effect on litter decomposition. For example, soft-rot fungi are common in the early stages of decomposition, whereas white- and brown-rot fungi dominate in the later stages. Endophytic fungi were regarded as weak decomposers that preferentially utilized the relatively labile organic matter fraction at the early stage of decomposition, whereas ectomycorrhizal fungi became dominant at later stages.

Fungi are among the most widely distributed organisms on earth and are of great environmental importance. Fungi are ubiquitous in soils, and dominate in soils with large organic matter content and acidic pH values, but constitute a smaller proportion of the microbial community in intensively managed mineral soils. Many fungi are free-living in soil, whereas others form parasitic or symbiotic relationships with plants or animals. Fungi are involved in a multitude of functional roles, such as decomposition, nutrient acquisition and mineral weathering. Previous studies confirmed the important role of fungi in the biodegradation of amended organic matter, e.g., plant litter.

Mycorrhizal fungi are known to be associated with plants, as it helps plant to obtain nutrients, while benefits from root-released organic substrates for C and energy. Arbuscular mycorrhiza fungi (AMF) are widely reported to play an important role in plant nutrient uptake, as well as the mineralization of root derived C. The AMF also increase the mineralization of SOC through a rhizosphere priming effect, which can be attributed to the widespread spatial distribution of rhizodeposits and extended area to obtain nutrients. For example, AMF help plants and themselves to exploit N through SOM mineralization (i.e., N-mining), when soils are nutrient deficient that cannot meet nutrient stoichiometric requirements (Jeewani et al., 2021b).

Fauna (protist, nematodes, collembola, earthworms)

Soil meso- and macro-fauna are major drivers in the early decomposition phases, with physical fragmentation of materials being a critical first step prior to mineralization by fungi and bacteria. Protists are increasingly recognized as key participants in multiple ecological processes, and important contributors to C turnover and transfer within soil food webs (Geisen et al., 2018). Soil protists can use rhizodeposit-C either through direct assimilation or by indirect predation. Soils harbor an enormously diverse and abundant protist population, most of which are considered major consumers of soil bacteria and fungi. These protist consumers influence C and nutrient flows across several trophic levels of soil food webs by preying on a wide range of bacteria and fungi. The diversity of trophic modes makes protists major participants of the soil microbial loop, which influences microbial biomass, activity and community structure in soil ecosystems. Protist predation, therefore, plays a significant role in microbial-driven decomposition.

Soil invertebrates (e.g., collembolan and earthworms) form a crucial part of the soil process related to the rates of decomposition of SOC and litter C. Soil mesofauna collembola alter the availability of organic matter, thereby indirectly mediating the primary decomposing microflora through trophic interactions. For instance, Yu et al. (2021) revealed the significant role of collembola in the degradation of maize (*Zea mays* L.) litter by increasing the C availability and shaping the community of decomposers, e.g., Ascomycota. Similarly, it has been suggested that earthworms modulate maize C mineralization by (i) changing the availability of C sources for primary decomposers and (ii) fragmenting and mixing these organic substances within the upper soil horizons. Earthworms also affect antagonistic interactions within the microbial community, particularly fungal groups.

Formation of SOM

Genesis and pathways

Plant-derived vs microbially-derived organic matter

Plant root-derived C dynamics in the rhizosphere are crucial for SOM formation and thus mitigation of greenhouse effects from increasing atmospheric CO₂ concentrations (Lehmann and Kleber, 2015). In general, C is transferred from the atmosphere to soil through plant photosynthesis and rhizodeposition in the form of water-soluble sugars and organic acids, and insoluble high-molecular-weight compounds.

Microbial necromass and its accumulation in soils

Microbial necromass (i.e., dead microbial biomass) is a significant source of SOM mainly because of recalcitrance of the cell walls, which have relatively long residence times in soil. The contribution of necromass to SOC content is far greater than previously considered, and was estimated to account for up to 40–80% of SOC (Liang et al., 2017).

Carbon flow through the atmosphere-plant-microbe-soil continuum

The microbial pathway through microbial biomass is poorly reflected in current climate models that incorporate SOC sequestration. An average of 0.54% of the ¹³C was found in microbial biomass carbon (MBC) immediately after ¹³CO₂ pulse labeling of flooded rice (*Oryza sativa* L.), and up to 0.41% remained in the MBC pool at the end of the growing season (Liu et al., 2019). The

$^{13}\text{CO}_2$ -pulse labeling and phospholipid fatty acids-stable isotope probing (PLFA-SIP) were used to quantify the allocation of photosynthetic C flows through plants shoots, roots, soil aggregates and the microbial community during ryegrass (*Lolium perenne* L.) elongation. Soil amended with biochar had decreased rhizodeposited- ^{13}C assimilation by fungi from 141 ng C g^{-1} soil to 33.6 ng C g^{-1} soil. The microbial pathway through necromass, however, has largely been ignored (Liang et al., 2017). Inclusion of microbial necromass into SOC sequestration models can help predictions of C storage in soil. Quite recently, a continuous $^{13}\text{CO}_2$ labeling study was conducted to trace C flows in the plant-microbe-soil necromass continuum, and the authors found that fungal necromass associated with soil mineral plays a significant role in soil C sequestration (Luo et al., 2021).

Abiotic drivers

Climate

Soil C cycling is regulated by soil microorganisms that in turn are strongly linked to climate variables. A temperature increase can enhance C loss in temperate forest soils by accelerating carbohydrate decomposition. Soil C cycling is further controlled by moisture conditions as regulated by precipitation, with CO_2 efflux and the metabolic respiration quotient increasing with increasing precipitation. Numerous investigations have reported rhizodeposit-driven C sequestration under various climate change scenarios. In general, soil C sequestration increases under elevated CO_2 (e CO_2) conditions. Rhizodeposits increased because e CO_2 -stimulated net primary production, shoot and root biomass, and increased the root/shoots ratio leading to increased rhizodeposited C inputs. By using continuous $^{13}\text{CO}_2$ labeling method, (Luo et al., 2018) found that rhizodeposits promote the accumulation of SOC under elevated CO_2 .

Soil characteristics

Herein, the broadly described soil abiotic factors that affect soil biotic properties and thus SOC dynamics include soil properties such as pH, mineral composition, nutrient composition. It has been widely reported that mineralization of root exudates, plant residues and pyrogenic C is affected by soil pH, mineral content, SOC and nutrient concentrations. This is mainly through changes in soil biotic properties and the regulation of the microbial community composition. For example, a meta-analysis based on 107 datasets from 64 long-term mineral fertilizer trials worldwide reported an average 15.1% increase in MBC mainly due to an increase in pH and nutrient contents (Geisseler and Scow, 2014). Numerous studies have confirmed that soil pH is one of the most influential factors affecting the biomass, activity and composition of the microbial community. Soil pH has a less profound effect on fungi than bacteria. Pietri and Brookes (2009) observed that fungal PLFAs were more dominant at lower pH values in the Hoosfield acid strip experiment at Rothamsted Research Station (across a uniform pH gradient due to a single chalk application during the 19th century).

Agricultural practices

The addition of organic matter amendments (e.g., straw, manure) has been demonstrated to be an effective practice in agro-ecosystems to build up SOC stocks to mitigate global CO_2 emissions. With large amounts of straw being incorporated into soils annually, the build-up of soil organic matter reached $0.35 \text{ t C ha}^{-1} \text{ yr}^{-1}$. This equates to a globally-scaled C sequestration potential for straw return of $34.4 \text{ Tg C yr}^{-1}$ (Minasny et al., 2017). A recent global meta-analysis based on 363 publications concluded that SOC content could be increased significantly on average by 14.9% through residue incorporation (Xia et al., 2018). Furthermore, a century-long field study at Rothamsted demonstrated the efficacy of building up SOM stocks through organic manure application.

Application of fertilizers affects substrate decomposition and may induce SOM mineralization (priming effects) as a result of a shift in the size of microbial biomass and structure of the microbial community. Nitrogen fertilizer application affected the accumulation of rhizodeposits in soil because of changes in the soil aggregate size fractions and shifts in the microbial community (Luo et al., 2018). Research on microbial responses to fertilization and consequent C dynamics are warranted.

Particulate organic matter and mineral associated organic matter

The formation of SOM is complex and largely depends on the input of plant-derived carbon and its transformation in the soil matrix. It is manifested by emerging evidence that SOM persistence mostly relies not only on C recalcitrance but also soil biophysical structures. These structures affect the microbial accessibility of organic matter that adheres on mineral surfaces or is protected by aggregates, as opposed to inherent chemical recalcitrance. Soil aggregates and mineral surfaces are often considered as hotspots for the production of biofilms by C-microbe-mineral interactions, leading to greater sequestration potential when additional C is added (Cotrufo and Lavelle, 2022). Classifying SOC into particulate organic matter (POM, mostly of plant origin) and mineral-associated organic matter (MAOM, mainly secondary minerals such as Al and Fe (hydr)oxides and layer silicate clays) advances our conceptual understanding of the biophysical mechanisms underpinning SOM turnover, namely stabilization via physical occlusion and chemical bonds, and mineralization through microbial utilization.

The efficiencies of POM and MAOM formation (% of C-inputs of rhizo-C, root and straw incorporated into each fraction) were quantified by isotopic techniques with rhizodeposits having the highest MAOM formation efficiency (46%). According to Six et al. (2002), both POM and MAOM are spatially linked from microaggregates to macroaggregates; While the fresh litter surfaces of POM concurrently promote the occlusion of organic matter into aggregates, the surfaces of MAOM acts as a nucleus in the formation of organo-mineral associations. Organo-mineral complexes are the basic components of aggregation and key to the persistence of organic matter. Less than 20% of the mineral surface area supported organic matter attachment sites, and most new inputs of organic

matter were attached to organo–mineral clusters already existing on the soil surface (Vogel et al., 2014). Therefore, the MAOM fraction is controlled by adsorption onto mineral surfaces with subsequent physical occlusion into microaggregates, whereas POM occurs either as a free (non-attached) fraction or is occluded into macroaggregates.

This physical hierarchy shapes niche differentiation for microbial assembly and subsequent organic matter transformations. The abundance and activity of microorganisms determine the transformation pathways for plant-derived C evolution into SOM. The microbiome is controlled by soil microsites within POM and MOAM, where biological, chemical and physical factors determine microbial community composition and metabolism. The microsites within POM and MOAC provide physical support for these biochemical processes through adsorption and desorption or complexation and dissolution making C sources available to colonized microorganisms. Mineral, such as kaolinite, ferrihydrite and quartz, host microbial communities by controlling microbial colonization on their surfaces. These minerals serve a foundational role in shaping the microbial community and thereby transformations of SOM. The adhesion of microorganisms to mineral and organic–mineral surfaces provides a fundamental basis for the underlying soil biogeochemical processes. Biotic factors mainly refer to the growth, activity, development, reproduction and metabolism of microorganisms, which are dependent on soil structure and C nutrient availability. The way that microorganisms, the key players in transforming plant-derived C into small molecules and subsequent humification, are controlled by the content of organic components (POM) and inorganic components (MAOM) offers a perspective about the mechanistic understanding of C stability.

Soil C sequestration potential in both POM and MAOM fractions is affected by vegetation (e.g., quality of C inputs), climate and soil properties including texture, mineralogy, pH, and microbial community structure. Chemical properties of plant-derived C inputs determine its distribution and transformation efficiency within these fractions. While the more recalcitrant compounds play a larger role in POM formation, the relatively simple compounds have a larger contribution to MAOM formation. For example, it was estimated that the formation efficiency by rhizodeposits compared with straw in agro-ecosystems to SOC formation was five times greater. This reflects the significance of MAOM in the accumulation of SOM from the input of low-molecular-weight compounds from plant rhizodeposits. In addition, compared to POM formation, MAOM formation requires nitrogen (N) and has a limited storage capacity when N availability is deficient. For instance, nitrogen-containing compounds that remain after the microbial metabolism of straw are preferentially fixed as mineral–organic complexes. This might reflect that (i) POM components from straw are usually lignin-like substances, and or (ii) the mineral-associated small molecules produced by microbial depolymerization are stoichiometrically limited.

Fungal aggregate–mineral interactions underlying carbon stabilization

The AMF promote rhizo-C stabilization by hyphae which induces soil aggregation and thus increases C storage (Averill et al., 2014). The AMF contribute to soil aggregate stability and aggregate formation by extra-radical hyphae, e.g., the genus *Glomus* transports plant-derived organic matter from roots to act as glue in forming soil aggregates by mineral–hyphal–aggregate interactions and or indirectly by altering biochemical properties of the soil–root environment.

Minerals protect SOC mainly through adsorption and or precipitation, with short-range-order phases and metal–organic complexes (Jeewani et al., 2020). Mineral such as goethite can promote Fe–organic complexes, which accelerate the formation of aggregates and influence both the processes and fate of rhizo-C and SOC (Jeewani et al., 2020, 2021a). These processes of SOC stabilization and mineralization driven by biotic–abiotic (hyphae–mineral) interactions occur simultaneously. The AMF and minerals can promote SOC sequestration through hyphae–aggregate–mineral interactions. Jeewani et al. (2021a) highlighted the contributions of both goethite (abiotic attributes) and AMF (biotic contributor) to C accrual in soil–plant systems. They found amendments of AMF and goethite resulted in a 0.6-fold decrease of rhizodeposit-derived CO₂, and an 11.3-fold larger allocation of rhizodeposits into macroaggregates. This was most likely due to SOC precipitation with goethite and macroaggregate formation stimulated by AMF hyphae. Further analyses by micron-scale Fourier transform infrared spectroscopy (μ -FTIR) showed that the spatial distribution of polysaccharides overlapped with Fe–O minerals within macroaggregates, supporting the concomitant processes of rhizodeposit stabilization and aggregate formation through hyphal–aggregate–mineral interactions.

Yet, these abiotic and biotic interactions via AMF, minerals and aggregates might have conflicting consequences on the SOC content. On the one hand, AMF might enhance microbial access to mineral-protected organic matter. The AMF have shown the potential to enhance the mobilization and mineralization of organic compounds associated with minerals. Further, it was reported that AMF help plants to obtain phosphorus from organic matter that is bound to goethite. This causes C loss through nutrient mining by microorganisms. On the other hand, AMF promote C stabilization by the (i) transfer of rhizodeposits to mineral surfaces (Averill et al., 2014), and (ii) creation of aggregates that promotes physical protection of organic matter by encapsulation or association with minerals (Six et al., 2002). These processes of SOC stabilization and mineralization driven by hyphae–mineral–aggregate interactions still require further investigation.

Approaches applied in SOC studies

Biomarkers

Soil organic matter inputs consist of partially-decomposed litter, root residues, rhizodeposits and microbial metabolites (Lehmann and Kleber, 2015). Establishing effective biomarkers that distinguish among these C sources and separate living microorganisms from dead microbial necromass is very challenging (Liang et al., 2017). Phospholipid fatty acid analysis combined with stable isotope probing can be used to trace plant-fixed C flows to the living microbial community in soil (Chen et al., 2021).

The accumulation of amino sugars in soil is a good indicator of dead microbial components, considering that plants do not produce amino sugars. Although more than 20 amino sugars have been identified in microorganisms, glucosamine (GlcN), galactosamine (GalN) and muramic acid (MurN) are the most widely used biomarkers. Glucosamine in soil is mainly derived from chitin originating in fungal cell walls, while MurN originates exclusively from bacterial cell walls. Thus, the relative amounts of glucosamine and muramic acid provide an indication of fungal (GlcN) and bacterial quantities (MurN), respectively (Liang et al., 2017).

Isotopic approaches

Tracing carbon flows and nutrient uptake isotopes

Carbon inputs through rhizodeposition are difficult to quantify because most measurements focus on the C pool size, which is relatively stable. Isotopic labeling techniques provide a tracer to track the flow and fate of photosynthesized-C through the plant-microbe-soil continuum. With isotopic labeling, it was estimated that ~45% of the photo-assimilates were distributed in plant shoots, 1–5% in roots and 0.2–3% were recovered in the soil organic C (SOC) pool, with 2–8% lost via belowground respiration. A dual labeling approach with ^{13}C and ^{15}N allows elucidation of C-N tradeoffs between the plant root and soil (Fu et al., 2021). For instance, biochar application increased ^{15}N uptake in plant shoots and leaves, but decreased rhizodeposited- ^{13}C recovery from large and small macroaggregates by 12–57% and 57–72%, respectively.

Partitioning soil CO₂ emissions among various C sources

To partition total CO₂ evolved from a single added substrate, a classic two-pool isotope mixing model has been widely used to partition among the two C sources, the added substrate and SOM. For instance, stable isotope techniques can distinguish C decomposition between substrate and SOC. However, quantitative partitioning between C sources using two stable isotopes (^{13}C and ^{12}C) has only been applied to two C systems. Separating among three C sources precisely, e.g., SOM, glucose and crop straw, is not possible using the two-pool model. It remains challenging to partition among three C sources, which are ubiquitous in natural systems, e.g., plant-soil contains rhizodeposits, litter, and SOC.

Recently, a dual radioactive- ^{14}C plus stable- ^{13}C labeling enabled the separation of three C sources (Luo et al., 2017). Radioactive ^{14}C is sensitive and precise, providing highly accurate measurements of partitioning among C pools and fluxes, and therefore, CO₂ from ^{14}C labeled substrate can be excluded first from the system. By using a mixed model twice the remaining C sources (^{13}C , ^{12}C) can be separated. In addition, Whitman and Lehmann (2015) introduced a dual-isotope approach which separate soil total CO₂ to SOC, biochar and root respiration. The development of more complex methodological approaches to discern among three C sources will provide several important advances to future research.

Outlook: New paradigms for SOC research

Recalcitrant nature of organic matter and microbial accessibility

A central paradigm of SOC mineralization is that the rate is controlled by either the nature of organic matter recalcitrance or microbial activity, abundance and community composition. It is posited that SOM mineralization mostly depends on the accessibility of soil microorganisms to SOM. Soil organic matter turnover is mainly governed by microbial accessibility rather than the recalcitrance of SOC. Thus, if SOC is not continuously converted to bioavailable and water soluble forms, or, microorganisms cannot move to the C sources, the mineralization of SOC would be at a very slow rate. With the importance of SOC mineralization to the overall C sequestration process, it is crucial to understand the dynamics of microbial biomass better in response to C source accessibility, which is most likely to be protected by abiotic factors such as physical occlusion or chemical precipitation.

Kemmitt et al. (2008) noted that even though 90% of microorganisms are destroyed by CHCl₃ (chloroform) fumigation of soil, following fumigant removal, SOC mineralization continued at the same rate as in the nonfumigated soil. They posited that mineralization of soil organic C is a two-stage process: non-bioavailable, humified forms are converted by abiotic processes to bioavailable forms which, only then, can undergo mineralization by the soil microbial biomass. Following this rationale, the former describes abiotic transformations of non-bioavailable SOC to bioaccessible and bioavailable forms, whereas the subsequent process is the biological mineralization of bioavailable SOC. They termed this concept the “regulatory gate hypothesis,” which explains why the mineralization of SOC proceeds at the same rate in both nonfumigated and fumigated soil, even though the size of the microbial population was decimated by fumigation, and both the microbial activity and diversity were drastically changed by fumigation. Brookes et al. (2017) demonstrated that the spatial arrangement of aggregate structure largely determined SOC bioaccessibility to consumers, providing additional support for the regulatory gate hypothesis. Their data indicated that SOC had first to undergo abiotic processes to become bioaccessible or bioavailable before it could be mineralized by the microbial biomass.

The precise nature of factors controlling the first may include processes such as adsorption and desorption, chemical oxidation and free radical activity. The biogeochemical cycling of organic substances depends on their interactions with the mineral soil matrix. The SOM can be stabilized on mineral surfaces and or inside aggregates during their formation, which protects the organic compounds from microbial degradation. Further, Jeewani et al. (2020) suggest that the sequestration of rhizodeposits in soil is via abiotic “rusty sink” (i.e., precipitation by iron oxide) and biotic mechanisms, e.g., microorganisms with branched filaments contributing to the development of aggregation.

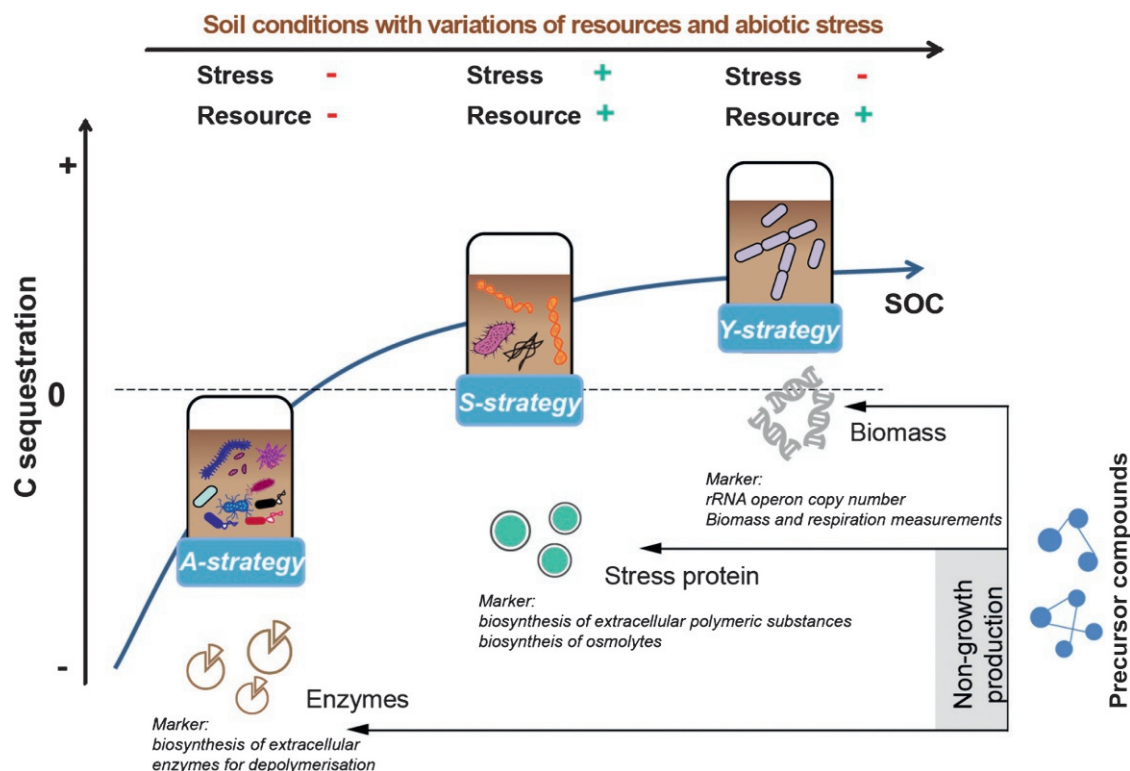


Fig. 2 Changes in trait-based strategy by fertilizer regimes affecting carbon sequestration. Control signifies soil without fertilizer, NPK signifies soils with fertilizer application of nitrogen, phosphorus and potassium, and NPKM signifies soils with fertilizer application of NPK and manure.

Traits-based strategy

Soil organic carbon content is determined by the carbon balance between sequestration and losses (i.e., stabilization vs mineralization). These opposing processes respond to organic C additions by a microbial tradeoff between anabolic and catabolic metabolisms, which is central to understanding the role of environmental biology in C exchange. This requires a strategy shift from a solely taxon-based approach toward an assessment of metabolic strategies. This provides a new paradigm for research on microbially driven processes, with a particular focus on soil C dynamics. Yet, the link between microbial metabolism (i.e., tradeoff between catabolism and anabolism) and C sequestration (stabilization vs mineralization) remains elusive.

Microorganisms compete with one another to acquire C, energy and nutrients during C mineralization. Fierer et al. (2007) demonstrated that SOC mineralization depends on a microbial community shift between two microbial functional groups that differ in terms of their C-use traits: r-strategists and K-strategists. The r-strategists respond quickly to simple, labile C substances in resource abundant environments, whereas K-strategists breakdown complex biopolymers in resource limited environments. Therefore, SOC mineralization is a function of the balance between r- and K-strategists, which is underpinned by substrate recalcitrance, nutrient status and the prevailing environmental niche. Further, Malik et al. (2019) classified microorganisms into three main strategy groups according to combinations of resource availability and environmental stress response: Y-strategists (flourish in soils without stress and an abundance of resources), A-strategists (adapt to ambient stress and resource-limited environments) and S-strategists (adapt to stress in resource-abundant environments). Moreover, Anthony et al. (2020) demonstrated how microbial communities and their eco-physiology (e.g., microbial growth, metabolism, etc.) as shaped by environmental stress (e.g., acidity, salinity) influence C decomposition. Future research to elucidate how changes in soil physicochemical properties by climatic changes (i.e., eCO_2) or agriculture practices (i.e., fertilization) that lead to shifts in the dominance of Y, A and S-strategists will help decipher the integrated mechanisms underlying C decomposition and its sequestration potential in soils (Fig. 2).

Conclusion

Soil organic C constitutes the largest C stock in terrestrial ecosystems, and its accumulation in soil pool is of vital importance for addressing the challenge of increasing atmospheric CO_2 concentrations. Soil organic carbon stocks depend on the balance between carbon input and output. The factors controlling SOC formation and decomposition include soil structure, nutrient availability, and biological biomass, and community diversity. Specifically, the processes of mineralization involve both abiotic and biotic factors,

including substrate recalcitrance, the soil matrix and microbial community, while the microbe (necromass)–mineral–aggregate interactions addressed the underlying mechanisms underpinning SOC formation. The methods used in SOC studies and the outlook for future research were given to advance our understanding of the key processes involved in SOC dynamics, which is vital for soil fertility, C sequestration, soil health and a climate change mitigation strategy.

References

- Anthony MA, Crowther TW, Maynard DS, van den Hoogen J, and Averill C (2020) Distinct assembly processes and microbial communities constrain soil organic carbon formation. *One Earth* 2: 349–360.
- Averill C, Turner BL, and Finzi AC (2014) Mycorrhiza-mediated competition between plants and decomposers drives soil carbon storage. *Nature* 505: 543–545.
- Brookes PC, Chen YF, Chen L, Qiu GY, Luo Y, and Xu JM (2017) Is the rate of mineralization of soil organic carbon under microbiological control? *Soil Biology and Biochemistry* 112: 127–139.
- Chen Z, Kumar A, Fu Y, Singh BP, Ge T, Tu H, Luo Y, and Xu JM (2021) Biochar decreased rhizodeposits stabilization via opposite effects on bacteria and fungi: diminished fungi-promoted aggregation and enhanced bacterial mineralization. *Biology and Fertility of Soils* 57: 533–546.
- Cotrufo MF and Lavalley JM (2022) Soil organic matter formation, persistence, and functioning: A synthesis of current understanding to inform its conservation and regeneration. *Advances in Agronomy* 172: 1–66.
- Crawford DL, Crawford RL, and Pometto L (1977) Preparation of specifically labeled ^{14}C -(lignin)- and ^{14}C -(cellulose)-lignocelluloses and their decomposition by the microflora of soil. *Applied and Environmental Microbiology* 33: 1247.
- Fierer N, Bradford MA, and Jackson RB (2007) Toward an ecological classification of soil bacteria. *Ecology* 88: 1354–1364.
- Fu Y, Kumar A, Chen L, Jiang Y, Ling N, Wang R, Pan Q, Singh BP, Redmile-Gordon M, Luan L, Li Q, Shi Q, Reid BJ, Fang Y, Kuzyakov Y, Luo Y, and Xu J (2021) Rhizosphere microbiome modulated effects of biochar on ryegrass ^{15}N uptake and rhizodeposited ^{13}C allocation in soil. *Plant and Soil* 463: 359–377.
- Geisen S, Mitchell EAD, Adl S, Bonkowski M, Dunthorn M, Ekelund F, Fernández LD, Jousset A, Krashevska V, Singer D, Spiegel FW, Walochnik J, and Lara E (2018) Soil protists: A fertile frontier in soil biology research. *FEMS Microbiology Reviews* 42: 293–323.
- Geisseler D and Scow KM (2014) Long-term effects of mineral fertilizers on soil microorganisms—A review. *Soil Biology and Biochemistry* 75: 54–63.
- Jeewani PH, Gunina A, Tao L, Zhu Z, Kuzyakov Y, Van Zwieten L, Guggenberger G, Shen C, Yu G, Singh BP, Pan S, Luo Y, and Xu JM (2020) Rusty sink of rhizodeposits and associated keystone microbiomes. *Soil Biology and Biochemistry* 147: 107840.
- Jeewani PH, Luo Y, Yu GH, Fu YY, He XH, Van Zwieten L, Liang C, Kumar A, He Y, Kuzyakov Y, Qin H, Guggenberger G, and Xu JM (2021a) Arbuscular mycorrhizal fungi and goethite promote carbon sequestration via hyphal-aggregate mineral interactions. *Soil Biology and Biochemistry* 162: 108417.
- Jeewani PH, Van Zwieten L, Zhu Z, Ge T, Guggenberger G, Luo Y, and Xu JM (2021b) Abiotic and biotic regulation on carbon mineralization and stabilization in paddy soils along iron oxide gradients. *Soil Biology and Biochemistry* 160: 108312.
- Kemmitt SJ, Lanyon CV, Waite IS, Wen Q, Addiscott TM, Bird NRA, O'Donnell T, and Brookes PC (2008) Mineralization of native soil organic matter is not regulated by the size, activity or composition of the soil microbial biomass. *Soil Biology and Biochemistry* 40: 61–73.
- Lehmann J and Kleber M (2015) The contentious nature of soil organic matter. *Nature* 528: 60–68.
- Liang C, Schimel JP, and Jastrow JD (2017) The importance of anabolism in microbial control over soil carbon storage. *Nature Microbiology* 2: 17105.
- Luo Y, Zang HD, Yu ZY, Chen ZY, Gunina A, Kuzyakov Y, Xu JM, Zhang KL, and Brookes PC (2017) Priming effects in biochar enriched soils using a three-source-partitioning approach: ^{14}C labelling and ^{13}C natural abundance. *Soil Biology and Biochemistry* 106: 28–35.
- Luo Y, Zhu Z, Liu S, Peng P, Xu J, Brookes P, Ge T, and Wu J (2018) Nitrogen fertilization increases rice rhizodeposition and its stabilization in soil aggregates and the humus fraction. *Plant and Soil* 445: 125–135.
- Liu Y, Ge T, Zhu Z, Liu S, Luo Y, Li Y, Wang P, Gavrichkova O, Xu X, Wang J, Wu J, Guggenberger G, and Kuzyakov Y (2019) Carbon input and allocation by rice into paddy soils: A review. *Soil Biology and Biochemistry* 133: 97–107.
- Luo Y, Xiao M, Yuan H, Liang C, Zhu Z, Xu JM, Kuzyakov Y, Wu JS, Ge TD, and Tang C (2021) Rice rhizodeposition promotes the build-up of organic carbon in soil via fungal necromass. *Soil Biology and Biochemistry* 160: 108345.
- Malik AA, Martiny JBH, Brodie EL, Martiny AC, Treseder KK, and Allison SD (2019) Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. *The ISME Journal* 14: 1–9.
- Minasny B, Malone BP, McBratney AB, Angers DA, Arrouays D, Chambers A, Chaplot V, Chen Z-S, Cheng K, Das BS, Field DJ, Gimona A, Hedley CB, Hong SY, Mandal B, Marchant BP, Martin M, McConkey BG, Mulder VL, O'Rourke S, Richer-de-Forges AC, Odeh I, Padarian J, Paustian K, Pan G, Poggio L, Savin I, Stolbovoy V, Stockmann U, Sulaeman Y, Tsui C-C, Vågen T-G, van Wesemael B, and Winowiecki L (2017) Soil carbon 4 per mille. *Geoderma* 292: 59–86.
- Pietri JCA and Brookes PC (2009) Substrate inputs and pH as factors controlling microbial biomass, activity and community structure in an arable soil. *Soil Biology and Biochemistry* 41: 1396–1405.
- Six J, Conant RT, Paul EA, and Paustian K (2002) Stabilization mechanisms of soil organic matter: Implications for C-saturation of soils. *Plant and Soil* 241: 155–176.
- Vogel C, Mueller CW, Hoschen C, Buegger F, Heister K, Schulz S, Schlöter M, and Kogel-Knabner I (2014) Submicron structures provide preferential spots for carbon and nitrogen sequestration in soils. *Nature Communications* 5: 2947.
- Whitman T and Lehmann J (2015) A dual-isotope approach to allow conclusive partitioning between three sources. *Nature Communications* 6: 8708.
- Xia L, Lam SK, Wolf B, Kiese R, Chen D, and Butterbach-Bahl K (2018) Trade-offs between soil carbon sequestration and reactive nitrogen losses under straw return in global agroecosystems. *Global Change Biology* 24: 5919–5932.
- Yu Z, Schmidt O, Zhao Y, Liu M, Kumar A, Luo Y, Xu J, and Luo Y (2021) Dinotefuran alters Collembola-fungi-bacteria interactions that control mineralization of maize and soil organic carbon. *Journal of Hazardous Materials* 418: 126391.

Soil proteins

Siobhan Staunton, UMR Eco&Sols, INRAE, INRAE-IRD-Cirad-Institut Agro-Université de Montpellier, Montpellier, France

© 2023 Elsevier Ltd. All rights reserved.

Introduction	396
Basic information on protein chemistry	396
Proteins in soil	397
Protein adsorption	397
Study of adsorbed proteins	399
Extraction, characterization and quantification of soil proteins	399
Soil enzymes	400
Glomalin-related soil protein (GRSP)	401
Potentially hazardous proteins in soil: Prions and Bt (<i>Bacillus thuringiensis</i>) proteins	403
Soil metaproteomics	404
Conclusions	404
References	404

Abstract

Proteins are complex macromolecules and their three-dimensional structure determines their activity. They play essential roles in soil; they act as reserves of essential nutrients particularly C and N, as enzymes they determine the biogeochemical cycles of essential nutrients, as important components of soil organic matter, they contribute to soil physical stability. After a brief presentation of basic protein chemistry and a description of proteins in soil, including adsorption properties, extraction and characterization, case studies of enzymes, two potentially toxic proteins and a controversial protein known as glomalin-related soil protein are described. A final section presents the pitfalls of soil metaproteomics.

Key points

- Soil proteins are important components of soil organic matter
- They are strongly, quasi-irreversibly, adsorbed.
- Enzyme activity is strongly dependant on protein properties
- The persistence and biological properties of two potentially toxic proteins (insecticidal Bt proteins and prion proteins) are determined by interaction with soil organo–mineral surfaces.
- Soil metaproteomics has been hindered by the absence of robust methods to extract and purify proteins

Introduction

Proteins are macromolecules that play many roles in soil. Their properties depend on their three-dimensional structure that is strongly influenced by their interactions with soil organo–mineral, and organic components. An overview of the most important features of protein chemistry in soil is given, with emphasis on commonly encountered problems in the study and interpretation of protein interactions and biological activity in soil. Special attention is given to enzymes that control the biogeochemical cycles of major elements, to the controversial group of proteins known as glomalin-related soil protein and to two classes of proteins with potentially adverse properties, namely prion proteins and the insecticidal Cry or Bt-proteins. Finally the features of protein chemistry in soil that have to date limited the expansion of soil metaproteomics are considered.

Basic information on protein chemistry

Proteins are complex three-dimensional organic molecules composed of amino acids connected with peptide links ((Norde et al., 2012), and many standard text books). Typically they have molecular masses of tens of thousands of grams; smaller chains of amino acids comprising only tens of amino acids are generally referred to as peptides or oligopeptides. Protein mass is given with the unit Dalton (Da), defined as 1/12 of the mass of a carbon-12 atom, and thus almost equivalent to the gram. Proteins are the building blocks of living organisms and are therefore abundant in soils. The precise composition of a protein depends on the component amino acids but in general they contain about 53% C and 16% N. The other major component is oxygen, with much smaller amounts of phosphorus (P), sulfur (S) and other elements. The chemistry of proteins is determined not only by their amino acid

composition, but also by their 3-dimensional structure. The structure of proteins is defined at three levels: the primary structure is the sequence of amino acids; the secondary structure is the local organization of these amino acids forming, for example alpha helices and beta-sheets; the tertiary structure is the overall shape and size of the molecule, depending on salt bridges, hydrogen bonds and disulfide bonds. The native conformation of proteins is the most energetically favorable tertiary structure and depends not only upon the amino acid composition but also external factors such as temperature, pH, hydration and salt concentration. In aqueous solution proteins often have a hydrophobic core leaving hydrophilic groups exposed. Protein structure varies to minimize the energetic state depending on external conditions. The propensity to conformation change leads to the classification as hard or soft proteins: hard proteins resist conformational changes whereas soft proteins change conformation readily as a function of external conditions, including adsorption, which will be described later. Proteins are amphoteric, namely they contain both acidic and basic groups and so their charge is sensitive to pH. Even at their point of zero charge there will be positively and negatively charged patches. Finally, proteins may form oligomers and polymers, and the propensity to do so differs between proteins and depends on external conditions such as pH, ionic strength and temperature. If a protein solution becomes cloudy, then the presence of polymers is likely and the solution may be filtered or centrifuged to obtain a clear solution. However, clear solutions may contain a mixture of monomers and oligomers with different properties, including adsorption properties.

Proteins in soil

Proteins are ubiquitous in soils (Quiquampoix and Burns, 2007). They may be intracellular, existing within living or formerly living organisms, or may be extracellular after release by cell lysis or active exudation. Extracellular proteins are predominantly rapidly and strongly adsorbed on solid surfaces, including the organo-mineral surfaces of soils, therefore free extracellular proteins are uncommon in soil. Proteins play many roles in soil. They control the biogeochemical cycles of major elements such as C (carbon), N (nitrogen), P and S, both as enzymes and sources of these nutrient elements. Proteins, together with their constituent amino acids are reported to be the major chemical form of soil organic nitrogen (Schulten and Schnitzer, 1998; Rillig et al., 2007) and may comprise a non-negligible fraction of soil organic carbon. As with other organic molecules, proteins contribute to the stock of organic matter and the physical structure of soils. These aspects, however, have received little research attention, with the exception of glomalin-related soil proteins, which will be discussed later. By far the greatest research on soil proteins has been on the catalytic activity of soil enzymes. The understanding and prediction of the persistence of proteins in soil and the catalytic activity of enzymes in soil require consideration of the extent and consequences of adsorption.

Protein adsorption

The adsorption of proteins on many surfaces including soils and soil minerals has been widely studied (Norde et al., 2012 #27; Quiquampoix and Burns, 2007 #32). The most detailed studies have been carried out using well defined, clean surfaces and a single protein. Adsorption mechanisms and driving forces are more difficult to investigate in detail in soils, because of the complex heterogeneous surfaces and large number of proteins present. However studies of model systems are nevertheless pertinent to soil science.

Soils and their components have a large capacity for the adsorption of proteins. To some extent the capacity is determined by surface area. Adsorption capacity of proteins on various clean mineral and organic surfaces is of the order of mg m^{-2} . This means that adsorption capacity of up to g g^{-1} could be expected on clay-textured soils, as has been reported for Bt and prion proteins, for example (Helassa et al., 2011). Proteins appear to adsorb to form monolayers on surfaces and there is no evidence that proteins adsorb in multilayers. In theory proteins could form suprastructures on surfaces, but this has not been demonstrated. Adsorption isotherms tend to take the form of Langmuir isotherms, namely increasing adsorption with increasing concentration in solution until a maximum or plateau is reached. The mathematical form of the Langmuir isotherm was developed from the adsorption of ideal gases with assumptions about sorbent and surface interactions, however goodness of fit to a model does not imply that the underlying assumptions apply. Contrary to the assumptions of perfect gas adsorption, adsorption sites in soil are not identical; adsorbed protein molecules interact, and adsorption of proteins is not a dynamic reversible process. While many studies of protein adsorption in chemistry and colloid science focus on maximum adsorption, in soil science it is often not pertinent to consider maximum adsorption. In the case of enzyme adsorption, concentration and surface coverage are often orders of magnitude below the maximum surface coverage. It is therefore more precise to refer to affinity rather than capacity. Adsorption affinity is often assessed using the distribution coefficient, K_d , the ratio of adsorbed content to concentration in solution. Since the adsorption isotherm is non-linear, the average K_d value may be very different from the tangent to the adsorption isotherm, which reflects the affinity of the surface for an additional small amount of sorbent.

Adsorption at surfaces is very rapid, and within the limits of experimental observation, it may be considered to be instantaneous. Empirically observed adsorption kinetics can thus be ascribed to rate limiting access to adsorption sites, the break-up of aggregates exposing new surfaces, and possibly to a lesser extent conformational changes of the adsorbed protein favoring further adsorption. It has also been shown with clean, homogeneous surfaces that the conformation of molecules depends on surface coverage; as a surface is increasingly covered, additional proteins adsorbed later will tend to have smaller contact area with the surface. Proteins are

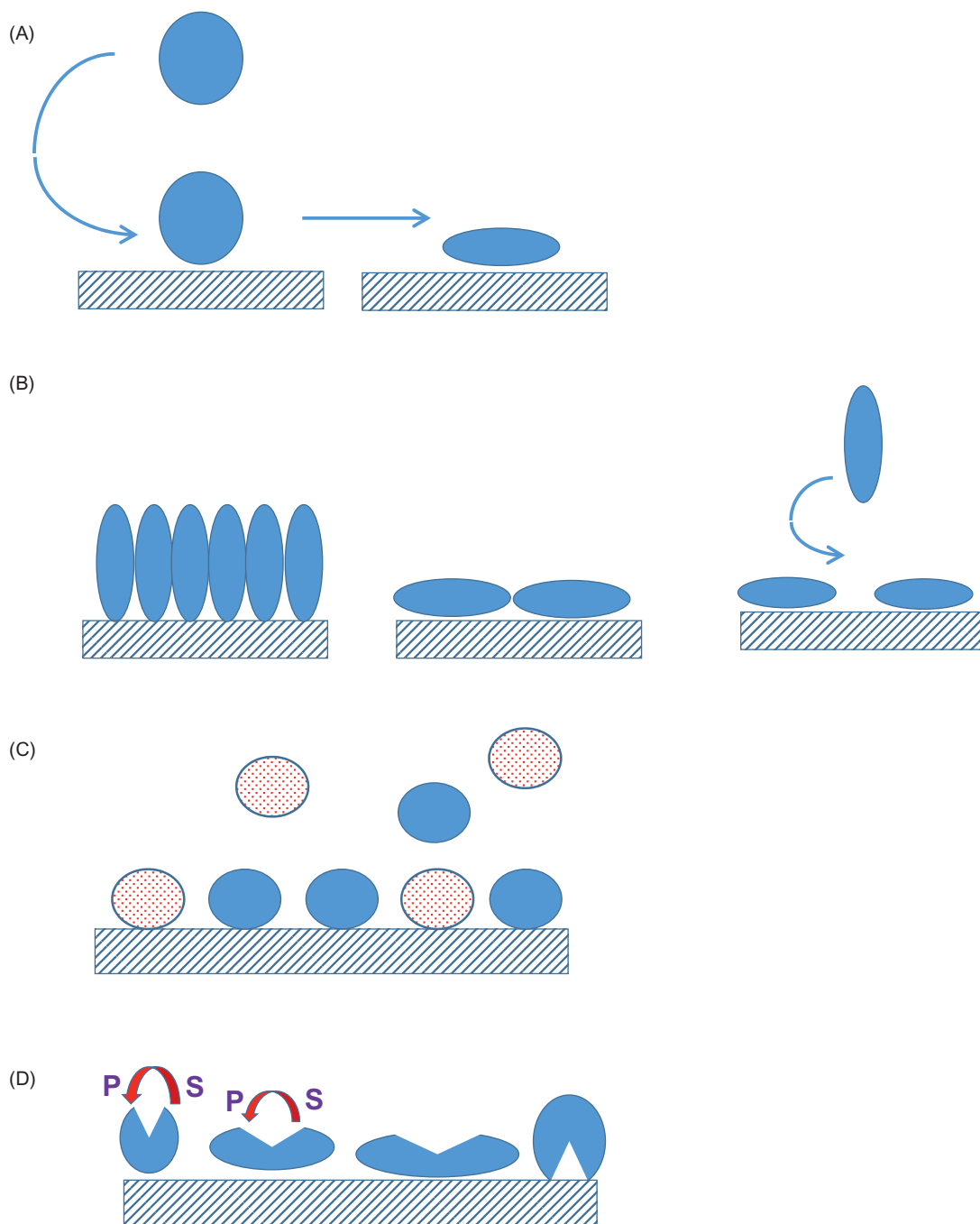


Fig. 1 Schematic representation of protein adsorption. (A) Protein adsorption, followed by (time-dependent) conformational change with spreading on the surface, thereby increasing contact area. (B) End-on packing on a surface; side-on packing with greater contact area and lower surface coverage; protein arriving at a nearly saturated surface forced to adopt an end-on contact. (C) Preferential adsorption of one protein (in blue) leaving a greater proportion of the other (red, hatched) in solution (D) Enzyme adsorbed with catalytic activity (substrate-to-product) conserved; catalytic activity reduced due to conformational change; catalytic activity lost due to conformational changes; active site oriented toward surface, thus not available for catalytic reaction.

not spherical, and adsorption may be end-on or side-on, depending on whether the longer axis is perpendicular or parallel to the surface (Fig. 1).

The driving forces for protein adsorption are complex; they include electrostatic interactions, hydrophobic interactions and hydrogen bonding. Conformational changes following adsorption involve changes in entropy. Protein structure may rearrange to present hydrophilic moieties to the solution, keeping hydrophobic moieties within the core of the protein or facing the surface. This can lead to adsorption isotherms that depend on the rate of addition of protein, namely the initial concentration in solution. A simple representation of this is to consider a large amount of protein brought in contact with a surface over a short period of time:

much protein would adsorb, with no time to spread out, whereas if protein is added more slowly, or in a step-wise fashion, the earliest adsorbed molecules spread, leaving less space for newer molecules to adsorb. Molecules reaching the surface later may thus either not adsorb, or adsorb with different conformation or orientation (Fig. 1). Conformational changes after adsorption also contribute to the apparently irreversible adsorption of proteins. These processes can also lead to a decrease with time of extractability or biological properties, that is difficult to distinguish from degradation (Fig. 1).

Protein adsorption on all surfaces is strongly pH dependant. This is true for charged surfaces as well as uncharged surfaces. In most cases a bell shaped curve is observed (Quiquampoix and Burns, 2007), with the largest adsorption capacity being near to the isoelectric point (iep), the pH at which net charge on the protein is zero, although both positively and negatively charged amino groups may be involved in adsorption. When net charge on the protein is opposite to that of the surface (positive at pH below the iep, in contact with electronegative surfaces) electrostatic interactions are strong, but other interactions cannot be ignored. At more alkaline pH there is nevertheless adsorption showing that some amino acids remain positively charged, and also that other interactions contribute to adsorption. Adsorption tends to decrease with increasing ionic strength, or salt concentration, in solution. This is in part due to a common ion effect, but also reflects the importance of electrostatic interactions on protein adsorption to charged organo-mineral surfaces.

Protein adsorption is usually calculated from the loss of protein from solution. When non-specific techniques are used for the quantification of protein, this may require unrealistically large concentrations of protein. Smaller amounts of adsorption can be quantified using a quartz crystal microbalance, that can also give information on the hydration properties of the adsorbed molecules, but this method cannot be used for soils. An alternative may be to label the protein, if a single protein is to be studied, with either a radioactive or fluorescent label. In both cases it is assumed that the presence of the label has no effect on the affinity of the protein for the surface, nor the subsequent conformational changes. This assumption is difficult to verify. In the case of enzymes, the change in catalytic activity in solution is measured, but the possible adsorption of inhibitors or activators from solution is usually overlooked.

Proteins also interact with colloidal or soluble organic molecules in soil. While this may not strictly be adsorption, the consequences may be similar on organo-mineral surfaces. The biological and biochemical properties may be modified, as observed for example with the loss of catalytic activity for enzymes in solutions of tannin or humic substances. It is often difficult to separate proteins and other soluble organic matter, as discussed later for the groups of proteins known as glomalin-related soil protein.

Study of adsorbed proteins

There is an abundant literature on the study of proteins adsorbed on surfaces (Norde and Lyklema, 2012). The practical applications of such research include the technological importance of immobilized enzymes, the fouling of medical materials and antigen binding to adsorbed antibodies. Methods are based on various techniques, a non exhaustive list includes vibrational spectroscopy, fluorescence spectroscopy, microscopy, optical methods, atomic composition obtained from X-ray photoelectron spectroscopy, hydrolysis followed by amino acid analysis and low temperature ashing, thermal analysis and surface physical properties such as hydrophobicity and wettability. However the large majority of these studies involve clean, uniform, often planar, surfaces, in marked contrast to the 'dirty' heterogeneous surfaces of soils. There are huge differences in the detail of information on protein structure for purified crystalline proteins, proteins adsorbed on clean well-defined surfaces, and adsorbed on soils. Mica, glass and quartz may be considered to be models of soil minerals, and are used for the methodologies that require planar surfaces. Vibrational spectroscopies, Raman and Fourier transform infra-red (FTIR) spectroscopies, have also been used to investigate the conformational changes of well-characterized proteins following adsorption on model clay minerals, such as montmorillonite. The investigation of adsorbed proteins in soil is therefore restricted to a limited range of experimental techniques, and much information is obtained on indirect evidence. These include the quantification of adsorption, kinetics of adsorption (although the rate limiting factors are certainly related to access to surfaces not the adsorption step), the conservation of biological and biochemical properties, pH-dependence of adsorption, and the persistence of the adsorbed protein. It is generally accepted that adsorption confers some degree of protection against microbial and catalytic breakdown. However some of this evidence is based on observations of loss of biological activity, whereas in reality the decline may be due to changes in conformation and orientation of adsorbed molecules. It is widely reported that immobilization changes the temperature dependence of catalytic reactions and the temperature stability of enzymes. This is important for the technological use of immobilized enzymes, and is also thought to be significant for soil enzymes. However in soil there is a more complex interplay between the synthesis, activity and persistence of enzymes produced by soil (micro)organisms.

Extraction, characterization and quantification of soil proteins

Extraction from soil is often the limiting step of the characterization and quantification of molecules and elements in soil. This is a huge obstacle to the study of soil proteins, and no methods are universally accepted (Greenfield et al., 2018). Extraction yields vary between soils, and depend on the protein and the protein-surface interaction, although this has never been investigated in depth. Extraction solutions are often adapted empirically from solutions used to extract proteins from biological material or used to clean surfaces. In general, alkaline pH and large salt concentration promote protein desorption. In addition, surfactants, or detergents,

may improve extraction yield, but may cause problems for the subsequent characterization or quantification of the proteins. Urea and guanidinium may also be used to disrupt hydrogen bonds. Proteins may also be removed from soil surfaces by heterogeneous exchange, namely the addition of other proteins. This is only appropriate when the added protein can be efficiently and specifically removed before further analysis, or if detection of the target protein is highly specific, for example when immune-assays are available. Extraction may alter protein structure and hence their biological or biochemical properties.

As described below, an operationally defined soil protein or group of soil proteins, known as glomalin-related soil protein, GRSP, or glomalin, is extracted by autoclaving in neutral or alkaline citrate solution. This technique extracts large amounts of protein, far in excess of the extraction yield at ambient temperature. However a drawback to the large extraction yield is the co-extraction of other substances. It has been suggested that strong, possibly covalent bonds are formed between proteins and other substances during the autoclave-extraction, making purification difficult.

Extraction yield may be calculated after controlled addition of a model protein to soil or to model minerals. However calculation of the extraction yield of native proteins from soil is limited by the lack of independent measurement of total protein present. It is only possible to compare the amounts of proteins removed by the use of different techniques. In general the amount of protein recovered is small compared with the probable content, given that proteins and amino acids are thought to be the primary forms of soil nitrogen. There is currently no evidence of the proportion of soil protein extracted, nor what relation there is between extraction efficiency and protein origin or composition. Studies of model proteins, including enzymes, Cry proteins or prion proteins, indicate that extractability decreases with time. It is therefore difficult to distinguish between extractability and decay.

After extraction, a range of techniques is available to quantify soil proteins. In a small number of cases highly specific, immune-tests are available and are subject to little interference. However in most cases, quantification is obtained using non-specific methods that may be subject to interference. Common protein quantification methods include UV-spectrometry (aromatic amino acids absorb at around 260 nm) and colorimetric assays. The most common colorimetric tests are Bradford, Lowry and the related bicinchoninic acid (BCA). They differ in the ease of experimental operation, their sensitivity to interference and their detection limits and working ranges. All these tests are based on chemical reactions with amino functional groups; the apparent protein concentration detected depends therefore on protein structure, the exposure of component amino acids to the solution containing chromophores, and the proteins used to calibrate the assay. The most convenient and therefore widely used test is the Bradford assay. However, this has a narrow working range and is sensitive to many of the reagents contained in common protein extraction cocktails, such as surfactants. Protocols for these methods are widely available in standard texts, research papers and in commercial documentation. However potential artifacts are widely overlooked by soil scientists; these include interference from non-protein solutes and the color of soil extracts.

Protein chemists and biochemists routinely clean up protein-containing solutions prior to quantification. The approaches routinely used are not always well adapted to soil science laboratories or to the impurities found in soil extracts. Acid precipitation is often reported for soil extracts even though it does not separate humic substances from proteins. Phenol extraction is often used by protein chemists but is unusual in soil science, and therefore is often found to be inconvenient.

Gel quantification is generally used to separate proteins to count the number of distinct proteins present in solution and may provide semi-quantitative assays. However inadequate sample purification can prevent the use of gel electrophoresis, with smeared gels that cannot be exploited. In the few published studies of soil proteins, the number of proteins separated is unreasonably small.

It is often not necessary to quantify soil proteins by chemical methods, since the biological or biochemical properties are of more importance than the absolute amount of protein present. It is therefore reasonable to study these properties without extraction. However when proteins are removed from the soil matrix before investigation, it must be remembered that they may have been denatured by the physical or chemical extraction method. There is never a simple, constant relation between biological, or biochemical, activity and amount of protein present.

Once in solution, chromatography may be used to separate proteins according to their size, hydrophobicity or charge. Amino acid composition, obtained by mass spectrometry after acid hydrolysis may give some information on the protein composition of soil organic matter. Although spectroscopic approaches are available to investigate protein composition and conformation in solution, the complexity of soil extracts limits their usefulness.

Soil enzymes

Many soil proteins are enzymes, and they probably compose the most studied class of soil proteins. These proteins dominate the biogeochemical cycles of important nutrients such as C, N, S and P. Many aspects of soil enzymes are treated in other chapters. However the protein nature of enzymes has important consequences on the performance and dynamics of enzymes. Extracellular enzymes are usually adsorbed on soil organo-mineral surfaces, and after adsorption they undergo conformational changes leading to changes in catalytic activity. Both the extent of adsorption and consequent conformational changes depend on both the enzyme and the surface. The complex, heterogeneous nature of soil organo-mineral surfaces and of the soil solution prevents precise prediction of the extent of adsorption and conformational changes of soil enzymes. Soil solution typically contains components that may activate or inhibit catalytic activity (Gianfreda and Bollag, 1994; Allison, 2006). The soil chemical property that has the greatest influence on catalytic activity is pH.

Soil catalytic activity is usually measured by adding model substrates and assaying the amount of a single product after a relatively short period. Since enzyme activity is strongly substrate dependent, the reported activity can at best be considered to be

potential activity (Nannipieri et al., 2012, 2018). For this reason there is often weak correlation between given catalytic activity for different substrates, for example fluorescent and colorimetric chromophores. In addition there is also weak correlation between measured catalytic activity and longer term assessment of the targeted biogeochemical function, such as organic residue turnover. It is therefore not correct to assume a direct positive relation between the amount of enzyme present and measured catalytic activity.

Nevertheless, because of the important roles of catalytic activity in soil, they are widely studied. In general, catalytic activity is measured with no attempt to desorb enzymes from soil. This is because of the difficulty of removing enzymes, like other proteins, from the soil organo–mineral surfaces. It may be reasonable to assume that all soil enzymes remain adsorbed and are not desorbed into solution. This choice introduces other potential sources of artifacts. Typically, soil is suspended in a solution, model substrate added and the concentration of product monitored or assayed after a suitable reaction period, often after stopping the reaction and separating phases by centrifugation. A well-known, but sometimes overlooked example of this is the adsorption of phosphate, the product of phosphatase and phytase reactions. One strategy to circumvent this problem is to estimate the proportion of phosphate product adsorbed from a separate adsorption isotherm and to correct the measured concentration. The use of fluorescent substrate-product pairs has enabled methodological advances, notably with the improvement of detection limits. However, if the reaction product is assayed without separation of the solution and solid phases, there may be artifacts due to absorbance of the fluorescence of the product. Another crucial choice is the composition of the solution used as a background. When solutions are buffered, often to the supposed optimal pH of the enzyme, the dominant cation is often sodium (Na), whereas in soil, calcium (Ca) is more often the dominant cation in both solution and the exchange complex. The catalytic activity of some enzymes requires the presence of calcium to form cation bridges. Ionic strength may also influence catalytic activity. However the major influence is pH. Since there is often a bell-shaped relation between catalytic activity and pH, rather small differences in pH around the optimal pH may have a large effect on measured activity. More importantly, there may be a continuum of pH optima of soil enzymes and this information will not be available from measurement at a single value of pH. Some studies prefer to measure catalytic activity in unbuffered soil suspension, but the pH of such suspensions is liable to be unstable, and pH may vary with the soil:solution ratio of the suspension.

Despite the poor yield of protein desorption, and the absence of a reliable prediction of the bias in desorption and the risk of protein denaturation by chemical extraction conditions, some studies attempt to desorb proteins prior to activity measurements (Tabatabai and Fu, 1992; Fornasier et al., 2015). Although it is well-known that salt solutions have very poor yields in protein extraction, simple salt solutions may be used to avoid protein denaturation. These approaches may avoid some of the experimental artifacts of direct measurement of catalytic activity, such as adsorption of the product, but others remain, including the relevance of model substrates and choice of pH.

Yet another approach to elucidate the interactions of enzymes with soil components and the consequences on enzyme activity and persistence, is to add controlled amounts of model enzymes to soil, or model soil components. The limit to this approach is that both enzymes and substrates are model molecules, often chosen because they are commercially available, not because of their relevance to soil conditions. It has been suggested that extracellular soil enzymes are better adapted to the presence of soil mineral surfaces than enzymes that are not native to soil. Therefore extracellular enzymes of native soils may be more likely to retain catalytic activity following adsorption.

Conformational changes following adsorption are inferred from changes in activity. This is an approximation. Observed activity may result from an average of various conformations, or even of various orientations, some of which may make the active site inaccessible. Decreases of activity with time may be interpreted as the microbial breakdown of enzymes by proteases, but may be due to changes in conformation.

Despite limitations to the interpretation of the study of controlled adsorption of enzymes to soils and soil components, important conclusions have been made. Adsorption is variable, depending on both the enzyme, the surface and external conditions. It is also strongly pH-dependent, and activity is generally observed to be reduced after adsorption. The reduction varies both with the enzyme and the surface, and varies with pH. Asymmetric reduction in activity may lead to a shift in the pH optimum of activity of the adsorbed enzyme, often toward alkaline pH. Every combination of pH-dependence of adsorption and resulting changes in activity have been reported from no adsorption to complete adsorption, and the subsequent reduction in activity being complete, pH-dependent or negligible (Quiquampoix and Burns, 2007). There are rare cases where activation of activity in the adsorbed state have been observed (Allison, 2006). Various explanations of such enhancement of activation include changes in oligomerization, conformation forced by surface shape, such as curved surfaces of nanoparticles, or by the adsorption of inhibitors, allowing a fuller potential of catalytic activity to be expressed.

Soil catalytic activity has been studied as an indicator of soil quality for many decades (Dick, 1994). Originally this arose from the assumption that soil enzymes were of biological origin and thus their presence and thus activity indicated good soil biological health. Usually a combination of different catalytic activities is measured to construct a compound index, avoiding too much emphasis on a single enzyme. Nevertheless, all the above criticisms of soil catalytic activity can be levelled against the quest for a universal index of soil quality.

Glomalin-related soil protein (GRSP)

One protein, or group of proteins, that has received much attention since its first description in 1996 is glomalin, or glomalin-related soil protein (GRSP) (Wright and Upadhyaya, 1996). Although the number of publications and citations

concerning GRSP has increased continuously, this fraction is highly controversial. This is an operationally defined fraction of soil organic matter, initially assumed to be a protein of arbuscular mycorrhizal fungal origin, obtained by the autoclaving of soil in citrate solution. This technique had been used to extract protein from fungal hyphae, and was transposed to soil. It was assumed that other soil proteins did not withstand the harsh extraction procedure. By extension, this fraction is thought to be particularly stable in soil. It is now understood that humic substances and other complex polyphenolic components are co-extracted with soil protein, and that the proteins are not exclusively of fungal origin. A few reports have proposed GRSP as a measure of soil protein in general, however there is no foundation for this assumption. An alternative terminology, autoclaved- citrate-extractible (ACE) protein has been suggested to avoid the assumption of fungal origin, but is not widely used.

Two extraction techniques are used: either a single extraction in neutral (pH 7) 20 mM sodium citrate solution to give easily extractable GRSP, GRSP_{EE}, or repeated extractions in alkaline (pH 8) 50 mM sodium citrate solution to give total GRSP, GRSP_T. It was recommended that the total extraction should be repeated up to 8 times, or until a characteristic reddish-brown color, thought to be the protein, was no longer visible. However there is no independent evidence that the color is caused by protein, since humic substances are known to be co-extracted. After extraction, the protein is usually quantified by the non-specific colorimetric assay known as the Bradford assay, and calibrated by a model protein, bovine serum albumin (BSA). Because of the use of the Bradford assay, some authors recommend the use of the term Bradford-reactive soil protein (BRSP), however this terminology underestimates the importance of the extraction process. This assay is convenient and commercial kits are available, but it has several shortcomings. First, it is non-specific and is subject to both negative and positive interferences, including sample color – authors rarely report any attempt to correct for color. Second, it has a rather small working range and so samples often require dilution. Furthermore, the spectral overlap between the absorbances of the reagent and the reactive product formed between protein and reagent make the calibration curve non-linear. Finally, although BSA is often used to calibrate protein assays; the sensitivity of any assay depends on the amino-acid composition of the protein and on the exposure of the amino acids to the reagents, thus the protein's tertiary structure. The data should therefore be expressed as equivalent BSA and not absolute protein content. One unusual consequence of these artifacts is that the amount of GRSP calculated in a soil extract often appears to depend on dilution. Various solutions have been suggested, including controlled dilution to a target sample absorbance (color intensity), the use of a modified Lowry method in place of the Bradford assay, controlled addition of a protein (BSA) to take account of sample matrix effects, or purification. Attempts have been made to purify the GRSP extract prior to quantification and characterization. Purification often consists of acid precipitation followed by re-solubilization and sometimes dialysis. However, the 'purified' extract usually does not have a typical protein composition: reported C and N contents are unrealistic, indicating that humic-like components are still present. The failure (often unreported) of other approaches to purify these extracts, for example by gel separation or chromatography, suggests that stable bonds have been formed between proteins and other components during the autoclave extraction. An alternative approach to quantification is to use an ELISA (enzyme linked immunosorbent assay) assay. However these assays are not commercially available and so must be developed by each laboratory. The antibody used for these tests, MAb 32B11, was selected because it was found to react against fungal proteins from various hyphae. Immuno-reactive soil protein (IRSP) is generally found to be considerably less abundant than Bradford reactive protein, indicating that non-fungal proteins are extracted, or that the protein is modified by extraction.

Despite the controversy surrounding the nature and origin of GRSP, it is widely studied. It is a minor but significant component of SOM with a possible role in C sequestration. Agricultural soils typically contain a few g kg⁻¹ GRSP, accounting for about 1–5% of SOC, depending on the supposed C content of GRSP. Forest soils, particularly tropical forest soils, have larger GRSP contents. The GRSP increases with increasing SOC, leading to doubts that it is a distinct fraction of SOM. Similarly it is positively correlated with soil nitrogen, which is consistent with the large N content of proteins (around 16%). When changes in land use lead to declines in SOM, GRSP often shows better resistance, leading to an enrichment of SOM in GRSP (Preger et al., 2007; Cissé et al., 2020). This observation has reinforced the hypothesis that GRSP is a stable fraction of SOM. An alternative hypothesis is that GRSP input to soil depends on plant cover (Cissé et al., 2021). The ratio GRSP:SOC tends to be largest under intensively farmed cereal crops and smallest in organic-rich forest soils (Staunton et al., 2020). A study of archived soils from long-term bare fallow soils showed that both GRSP and SOM declined with time over a period of decades; the decline in GRSP was not always slower than that of SOM.

The GRSP is widely considered to be a convenient, low-cost marker of fungal activity and turnover. However weak and inconsistent correlations are often found between various GRSP contents (EE and T, Bradford reactive and immune-reactive) and other fungal markers. This may be due in part to differences in turnover times and persistence of fungal hyphae and proteins in soil, but also reflects the reality that GRSP is not exclusively of fungal origin.

Glomalin-related soil protein (GRSP) is assumed to confer physical stability to soils and this is supported by the correlations observed between GRSP content and physical stability (Rillig et al., 2002). However in the majority of these studies, the possible relation between SOC and GRSP is untested, therefore begging the question as to the specific role of GRSP. The same criticism may be levelled against the affirmation that GRSP sequesters metals in soils and sediments. Originally the link between GRSP and metal (contamination) arose from the assumptions that GRSP is a marker of fungal activity and that fungal activity will be negatively impacted by metal contamination (Cornejo et al., 2008). This was generally deduced from the metal content of extracted GRSP, however when metal contents of GRSP and SOM are compared, the capacity of GRSP to sequester metals is not conclusive.

Glomalin-related soil protein (GRSP) is thought to be a stable fraction of SOM, with an average age estimated as being years to decades. This age is considerably longer than that of fungal hyphae, but in accord with SOM age, although this is often cited as evidence of the recalcitrance of GRSP in soil. Small seasonal variations of GRSP have been reported by some studies (Gispert et al., 2018), and this tends to contradict the assumption of recalcitrance of GRSP. The GRSP_{EE} is often assumed to be more recent than the

total fraction. The underlying logic to this assumption has not been explicitly described nor has the hypothesis been independently tested. One explanation would be that with time proteins become increasingly strongly bound to soil, a process known as fixation, and thus more difficult to remove. However this has never been tested for GRSP and rarely for other proteins at appropriate time scales. The often reported strong correlation between the two fractions contradicts a marked difference in age.

Potentially hazardous proteins in soil: Prions and Bt (*Bacillus thuringiensis*) proteins

Two groups of proteins potentially present in soil have attracted concern since the late 20th century. Prion proteins are the infectious agents responsible for transmissible spongiform encephalopathies (TSEs), or prion diseases (Saunders et al., 2008; Smith et al., 2011). One of these diseases, scrapie, has been known to affect sheep for centuries, and the role of soil as a reservoir has been suspected for decades. However it was the cross species transmission with the appearance of bovine spongiform encephalitis, BSE (mad cow disease) and a new variant of Creutzfeldt-Jakob disease in humans, linked to the use of contaminated meat and bone meal as animal feed that sparked interest in the role of soil in the transmission of these diseases. The current research focus has moved from Europe to North America where wild deer (cervids) suffer from chronic wasting disease (CWD). Prion proteins enter soil from contaminated carcasses, animal excreta and inadequately inactivated meat and bone meal. They may pass from soil to animals by soil ingestion. They are known to conserve biological properties after adsorption on soils, and possibly for years. This implies that the proteins conserve their conformation after adsorption, and that adsorption protects them against microbial and protease breakdown. It appears that prion proteins may be biologically active and cross biological membranes while remaining in the adsorbed state. Pathogenicity is assessed using bioassays, however these require large numbers of sensitive animals, and so are time consuming, expensive and variable. Most studies therefore prefer assays of soluble protein, thus requiring prior extraction of the proteins. On the basis of the few available studies, the infectivity appears to decline more slowly with time than does chemical extractability. No universal extraction technique exists, but as for other proteins, high pH, large salt concentration, detergents and the addition of an indifferent protein improve the extraction yield. Quantification methods for the soluble protein are usually ELISA tests, with many commercially available kits. An interesting alternative to extraction-detection is protein misfolding cyclic amplification (PMCA) that appears to give good detection efficiency.

Another group of proteins of environmental concern are known as Cry or Bt proteins. They have insecticidal properties and are introduced into soil by the endemic bacterium *Bacillus thuringiensis* or by a class of genetically modified (GM) (Spohn et al., 2016) crops, known as Bt-crops, that have been engineered to produce one or more of the proteins. The insecticidal properties of these proteins results from a complex series of reactions involving protein oligomerization and attachment to specific receptors in insect guts (Clark et al., 2005; Helassa et al., 2013). There is no known mechanism for these proteins to affect mammals adversely, but their persistence in soil has other important implications. These include the possible exposure to non target insects, the low-level exposure to target insects that would accelerate the appearance of resistance and soil monitoring for the surveillance of GM use or for organic farming certification. A very large number of proteins are known to be produced by *B. thuringiensis*, although only a small number have been characterized. The bacterium produces large amounts of protein, up to 30% of its mass during sporulation. It is the crystal-like appearance of the proteins at this stage that has caused the name Cry to be given to them. Each strain produces a small number of proteins, usually between 1 and 5, and the proteins have a very narrow, highly specific range of target insects that must ingest the protein to be exposed to its toxic properties. Toxins and spores are widely used as biopesticides. A very small number of proteins is produced by GM crops, and the protein they synthesize may enter soil by root exudation, or the decomposition of crop residues. An important difference between the proteins produced by the bacterium and those produced by GM crops, is that the former are protoxins, with molecular masses of about 73–130 kDa. The first step in the process of toxicity is the enzymatic cleavage of these protoxins in the alkaline mid-gut of insects, whereas GM plants synthesize the toxin directly, with a molecular mass of about 60 kDa.

All Cry proteins are strongly adsorbed on soil organo-mineral surfaces. Adsorption isotherms conform to the Freundlich isotherm, with maxima in accord with the specific surface area of the mineral. However, at lower concentrations there is no consensus on a relation between affinity and the nature of the mineral surface. In solution, many have a strong tendency to oligomerize, and thus must be stored at high pH in concentrated salt solutions. These properties complicate the experimental investigation of their properties and adsorption behavior, and can lead to artifacts and incorrect interpretation of observations. It is probable that monomers and oligomers have different adsorption properties, and there is spectroscopic evidence from FTIR that oligomerization is impeded after adsorption (Helassa et al., 2011). There are conflicting data on the persistence of Cry proteins in soil, with some studies reporting the rapid disappearance of the protein, and others claim it can be detected for months. The cause of these discrepancies may lie in the difficulty of chemical extraction of the protein prior to quantification. As for other proteins, extraction is promoted by large pH, high salt concentration, the addition of detergent and another protein. The insecticidal properties can be monitored using a large, relatively easily manipulated insect, *Manduca sexta*. The insecticidal properties are conserved after adsorption, which may appear surprising since the mechanism of toxicity requires the proteins to be in solution. However, the composition of the insect mid-gut resembles that of many efficient extraction solutions; high pH and presence of surfactants. Toxicity of soil-adsorbed Cry protein, measured with *Manduca sexta* bioassays, and quantification using commercial ELISA kits following chemical extraction, follow the same trends. The temperature effects of the decline in toxicity and extractability suggest that declining toxicity may be explained by time-dependent fixation of the protein on mineral surfaces, rather than microbial breakdown (Hung et al., 2016).

Soil metaproteomics

Metaproteomics is the large-scale characterization of the entire protein complement of an ecosystem. This is made feasible by large throughput technology. The term is a parallel to meta-genomics, the term coined in the early 21st century as an alternative to microbial environmental genomics, to avoid the bias toward culturable bacteria (Rodríguez-Valera, 2004). The underlying driver for such research is that improving knowledge of a system must lead to progress, even if the precise nature of the progress is not fully predictable. Meta proteomics of many environmental systems has made remarkable progress. There are three steps in such analysis: extraction; purification; analysis by mass spectrometry. There was a flurry of enthusiasm for soil metaproteomics, but this has petered out (Bastida et al., 2009; Keiblinger et al., 2012). There are two major reasons for this. The greatest barrier to any systematic study of extracellular soil proteins is the absence of a robust method to extract and purify proteins efficiently and without bias. As has been discussed in the previous section, proteins are intrinsically difficult to desorb from surfaces. The situation in soils is further complicated by the diversity of organo–mineral surfaces in soils and of extracellular proteins. To date there is no widely accepted method to extract proteins from soils. As has been described above, detergents, alkaline pH, large salt concentration and indifferent proteins have been found to be effective to extract some proteins, but there have been no systematic studies. Furthermore extracted proteins must be separated from the extractant that contains detergent, for example, and co-extracted organic compounds (see the example of GRSP above) before analysis. It has proved difficult to separate protein from other soluble soil organic components. The contamination of extracts with tannin and various humic-like components preclude the use of separation by gel electrophoresis. In the few published studies of soil proteome, the number of proteins identified by gel electrophoresis is astonishingly small; tens or hundreds. Finally, the identification of proteins and their attribution to soil biota will be limited by the paucity of knowledge of soil biota. The study of proteins associated with soil biota, namely intracellular proteins, has been somewhat more successful. By first removing free-living biota from soil, the extraction process is simpler, and less subject to subsequent interference in the quantification and characterization of proteins (Maron et al., 2007). Soil metaproteomics is an exciting field of research that requires a technological leap to allow significant progress in the future.

Conclusions

The chemistry of proteins has important consequences on their role and dynamics in soil. Interpretation of the quantification of proteins and measurements of their persistence and reactions may be erroneous if essential features of their chemistry are overlooked.

References

- Allison SD (2006) Soil minerals and humic acids alter enzyme stability: Implications for ecosystem processes. *Biogeochemistry* 84: 361–373.
- Bastida F, Moreno JL, Nicolás C, Hernández T, and García C (2009) Soil metaproteomics: A review of an emerging environmental science. Significance, methodology and perspectives. *European Journal of Soil Science* 60: 845–859.
- Cissé G, van Oort F, Chenu C, Essi M, and Staunton S (2020) Is the operationally defined fraction of soil organic matter, GRSP (glomalin-related soil protein), stable in soils? Evidence from trends in long-term bare fallow soil. *European Journal of Soil Science* 72: 1101–1112.
- Cissé G, Essi M, Kedi B, Mollier A, and Staunton S (2021) Contrasting effects of long term phosphorus fertilization on glomalin-related soil protein (GRSP). *European Journal of Soil Biology* 107: 103363.
- Clark BW, Philips TA, and Coats JR (2005) Environmental fate and effects of *Bacillus thuringiensis* (Bt) proteins from transgenic crops: A review. *Agricultural and Food Chemistry* 53: 4643–4653.
- Cornejo P, Meier S, Borie G, Rillig MC, and Borie F (2008) Glomalin-related soil protein in a Mediterranean ecosystem affected by a copper smelter and its contribution to Cu and Zn sequestration. *Science of the Total Environment* 154–160. Netherlands.
- Dick RP (1994) Soil enzyme activities as indicators of soil quality. In: *Defining Soil Quality for a Sustainable Environment*, pp. 107–124. USA: Madison, WI.
- Fornasier F, Dudal Y, and Quiquampoix H (2015) Enzyme extraction from soil. In: *Methods of Soil Enzymology*, 371–383.
- Gianfreda L and Bollag J-M (1994) Effect of soils on the behavior of immobilized enzymes. *Soil Science Society of America Journal* 58: 1672.
- Gispert M, Pardini G, Emran M, Doni S, and Masciandaro G (2018) Seasonal evolution of soil organic matter, glomalin and enzymes and potential for C storage after land abandonment and renaturalization processes in soils of NE Spain. *Catena* 162: 402–413.
- Greenfield LM, Hill PW, Paterson E, Baggs EM, and Jones DL (2018) Methodological bias associated with soluble protein recovery from soil. *Scientific Reports* 11186. England.
- Helassa N, Revault M, Quiquampoix H, Dejardin P, Staunton S, and Noirville S (2011) Adsorption on montmorillonite prevents oligomerization of Bt Cry1Aa toxin. *Journal of Colloid and Interface Science* 718–725. United States.
- Helassa N, Quiquampoix H, and Staunton S (2013) Structure, biological activity and environmental fate of insecticidal Bt (*Bacillus thuringiensis*) cry proteins of bacterial and genetically modified plant origin. In: *Molecular Environmental Soil Science. Progress in Soil Science*. Dordrecht: Springer.
- Hung TP, Truong LV, Binh ND, Frutos R, Quiquampoix H, and Staunton S (2016) Persistence of detectable insecticidal proteins from *Bacillus thuringiensis* (Cry) and toxicity after adsorption on contrasting soils. *Environmental Pollution* 208: 318–325.
- Keiblinger KM, Wilhartz IC, Schneider T, Roschitzki B, Schmid E, Eberl L, Riedel K, and Zechmeister-Boltenstern S (2012) Soil metaproteomics - Comparative evaluation of protein extraction protocols. *Soil Biology and Biochemistry* 14–24. England.
- Maron PA, Ranjard L, Mougel C, and Lemanceau P (2007) Metaproteomics: a new approach for studying functional microbial ecology. *Microbial Ecology* 53: 486–493.
- Nannipieri P, Giagnoni L, Renella G, Ceccanti B, Masciandaro G, Fornasier F, Moscatelli MC, and Marinari S (2012) Soil enzymology: classical and molecular approaches. *Biology and Fertility of Soils* 48: 743–762.
- Nannipieri P, Trasar-Cepeda C, and Dick RP (2018) Soil enzyme activity: A brief history and biochemistry as a basis for appropriate interpretations and meta-analysis. *Biology and Fertility of Soils* 54: 11–19.

- Norde W and Lyklema J (2012) Interfacial behaviour of proteins, with special reference to immunoglobulins. A physicochemical study. *Advances in Colloid and Interface Science* 5–13. Netherlands.
- Norde W, Horbett TA, and Brash JL (2012) Proteins at interfaces III: Introductory overview. *Proteins at Interfaces III State of the Art* 1–34. USA, Washington, DC.
- Preger AC, Rillig MC, Johns AR, Du Preez CC, Lobe I, and Amelung W (2007) Losses of glomalin-related soil protein under prolonged arable cropping: A chronosequence study in sandy soils of the South African Highveld. *Soil Biology and Biochemistry* 39: 445–453.
- Quiquampoix H and Burns RG (2007) Interactions between proteins and soil mineral surfaces: Environmental and health consequences. *Elements* 3: 401–406.
- Rillig MC, Wright SF, and Eviner VT (2002) The role of arbuscular mycorrhizal fungi and glomalin in soil aggregation: Comparing effects of five plant species. *Plant and Soil* 238: 325–333.
- Rillig MC, Caldwell BA, Wösten HAB, and Sollins P (2007) Role of proteins in soil carbon and nitrogen storage: Controls on persistence. *Biogeochemistry* 85: 25–44.
- Rodríguez-Valera F (2004) Environmental genomics, the big picture? *FEMS Microbiology Letters* 231: 153–158.
- Saunders SE, Bartelt-Hunt SL, and Bartz JC (2008) Prions in the environment: Occurrence, fate and mitigation. In: *Prion*, pp. 162–169. United States.
- Schulten HR and Schnitzer M (1998) The chemistry of soil organic nitrogen: A review. *Biology and Fertility of Soils* 26: 1–15.
- Smith CB, Booth CJ, and Pedersen JA (2011) Fate of prions in soil: A review. *Journal of Environmental Quality* 40: 449–461.
- Spohn M, Pötsch EM, Eichorst SA, Wöbken D, Wanek W, and Richter A (2016) Soil microbial carbon use efficiency and biomass turnover in a long-term fertilization experiment in a temperate grassland. *Soil Biology and Biochemistry* 97: 168–175.
- Staunton S, Saby NPA, Arrouays D, and Quiquampoix H (2020) Can soil properties and land use explain glomalin-related soil protein (GRSP) accumulation? A nationwide survey in France. *Catena* 193: 104620.
- Tabatabai MA and Fu M (1992) Extraction of enzymes from soils. In: Bollag JM and Stotzky G (eds.) *Soil Biochemistry*. USA, Boca Raton: CRC Press. <https://doi.org/10.1201/9781003210207>.
- Wright SF and Upadhyaya A (1996) Extractaion of an abundant and unusual protein from soil and comparison with hyphal proteis of arbuscular mycorrhizal fungi. *Soil Science* 161: 575–586.

Biochar: Properties, mechanisms, and interactions in the soil

Balwant Singh^a, Shamim Mia^b, and Marta Camps-Arbestain^c, ^aSydney Institute of Agriculture, School of Life and Environmental Sciences, The University of Sydney, NSW, Australia; ^bDepartment of Agronomy, Patuakhali Science and Technology University, Patuakhali, Bangladesh; ^cSchool of Agriculture and Environment, Massey University, Palmerston, New Zealand

© 2023 Elsevier Ltd. All rights reserved.

Introduction	406
Properties of biochar	407
Production	407
Soluble salts, ash, pH, and liming potential	407
Total carbon, nitrogen, oxygen, and hydrogen	408
Surface charge characteristics of biochars	411
Bulk density, pore space, and specific surface area	411
Biochar applications	413
Biochar for long-term carbon storage in the soil	413
Fertilizer benefits of biochar in soil	414
Influence of biochar on soil physical properties	414
Effects of biochar application to soil on crop yield	415
Biochar and the soil microbial community	416
Biochar standardization and quality assurance for potential contaminants	416
Remediation of contaminated soil	417
Biochar in composting of biological wastes	418
Conclusions and the future	419
References	419

Abstract

Biochar is a carbon-rich solid made by the pyrolysis of organic wastes and other feedstocks for soil and other environmental applications. Its chemical and physical properties are highly variable, depending strongly on the feedstock, pyrolysis temperature and conditions used for its production. This chapter reviews important chemical and physical properties of biochars arising from particular types of feedstock and pyrolysis temperatures. It reviews the effects of additions of biochar on the chemical, physical and microbial properties of soil, and on crop yields. It includes novel applications for the remediation of contaminated soil and its use in complex fertilizers. Last, it identifies topics for future research.

Key points

- Biochars are made from various feedstocks and under various pyrolysis conditions.
- Biochar added to soil, has important effects on properties and can increase crop yields.
- Novel uses of biochar are its addition to fertilizers and for remediation of contaminated soil.
- Further research on use of biochar in agriculture is suggested.

Introduction

Biochar is a stable solid material, rich in carbon (C), produced by thermal decomposition of organic material in the absence of oxygen or with too little oxygen for combustion. It has a predominant condensed-aromatic structure that makes it chemically and biologically persistent and able to remain in soil for several decades and even for millennia. When added to soil, it remains as a long-term store of C, and it improves soil conditions for plant growth. The idea of adding biochar to soil arose from observations made on the Amazonian dark earths ('Terra Preta do Indio' of Brazil), which contain more stable organic C than the surrounding soil and are more productive (Glaser, 2007). Natural wildfires also produce charred material and soot which are commonly referred to as either black carbon or pyrogenic C. According to Reisser et al. (2016) pyrogenic C constitutes on average 13.7% of the soil's organic C globally and as much as 60% in some places.

In the 1990s, scientists saw the deliberate conversion of waste biomass into biochar and its addition to the soil as a means of storing C and mitigating global warming; the soil could serve as a carbon sink. Lehmann et al. (2021) estimated that application of biochar to the soil could diminish emissions of greenhouse gases and remove CO₂ by 3.4 to 6.3 Pg CO₂ equivalent per year: a significant effect. Biochar can also help to preserve organic matter in the soil through negative priming. Addition of biochar to soil

improves the soil's physical, chemical and biological properties associated with plant nutrition thereby leading to enhanced growth and larger yields of crops. These benefits can vary significantly, depending on the feedstock used for biochar, the pyrolysis temperature, and on the soil and its environment. Here, we review the properties of biochar, the effects of biochar on the soil's chemical, physical and microbial properties and implications for agronomy and the environment. The research on biochar done in the last 20 years has resulted in a huge literature recording the results, some of which conflict. We cite only a small proportion of the sources but have attempted to list the most significant and informative.

Properties of biochar

Production

The properties of biochar vary, depending on the conditions of pyrolysis and on the feedstock and other ways in which it is treated, including hydrothermal carbonization, gasification, and flash carbonization. Pyrolysis is a thermochemical decomposition of biomass under anoxic conditions at temperatures ranging from 350 °C to 800 °C or more; it is the commonest means of biochar production. The product is a solid material, biochar, plus liquid tar and a mixture of gases including H₂ (hydrogen), CH₄ (methane), CO₂ (carbon dioxide) and CO (carbon monoxide). The relative yields of these products vary depending on the nature of the biomass, the temperature and the rate of heating. Increases in temperature and residence time reduce the yields of biochar. Low temperatures lead to greater production of liquids, whereas at high temperatures more gases are produced. Increases in temperature also lead to a decrease in acidic functional groups and an increase in the salts of alkali and alkaline earth elements in the biochar.

As the temperature increases there is a progressive loss of oxygen (O) and hydrogen (H), both of which are lost rapidly between 300 °C and 500 °C. The organic compounds in the feedstock, such as cellulose, hemicellulose, and lignin, decompose through the cleavage and cracking of weak bonds, and condensed aromatic structures are generated during heating. The main sources of feedstock are biological wastes: wood, crop residues, weed grass, and manure. Characteristics of biochars produced from different feedstocks differ somewhat from one another and are described below.

Soluble salts, ash, pH, and liming potential

Biochars contain readily soluble salts, in particular chlorides of potassium (K), sodium (Na) and calcium (Ca), carbonates, and bicarbonates of Ca, Na, and magnesium (Mg), and sulfates and phosphates of Ca, K, and Mg. These salts contribute to the alkalinity of biochars, although the particular mixtures of the salts depend on the feedstock and production process. They confer a pH on the biochar that is generally neutral to alkaline, consequently giving it a value for liming.

Biochars made from woody biomass (median pH 8.1) are generally less alkaline than those made from grasses (median pH 9.0), crop residues (median pH 9.3) and manures (median pH 9.4) (Fig. 1). Exceptions to this are biochars from some species of trees, in particular of pine (*Pinus*), eucalyptus (*Eucalyptus*) and oak (*Quercus* L.), produced at low temperatures (≤ 350 °C), which can have

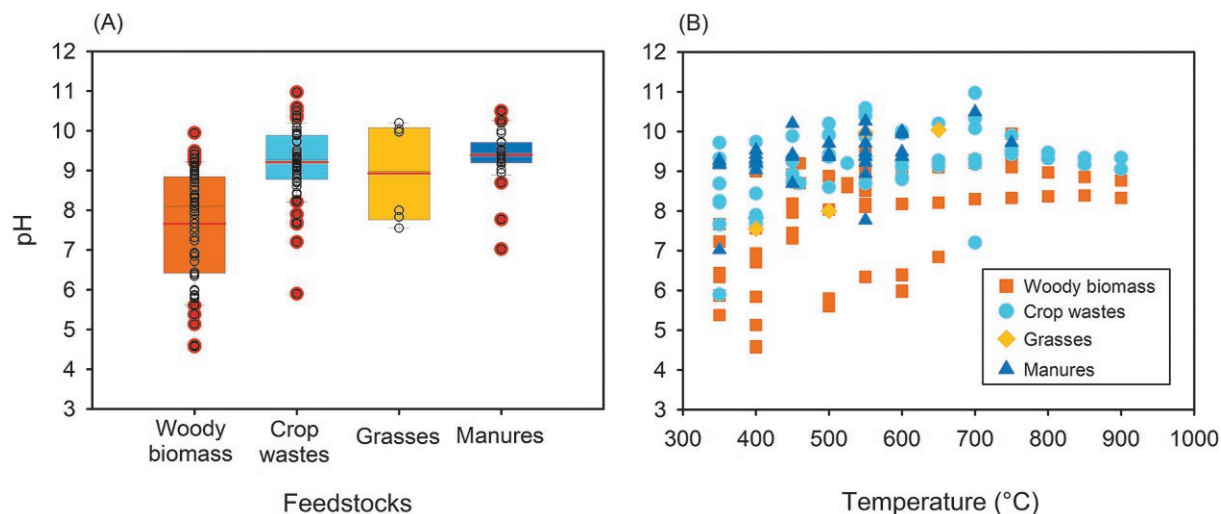


Fig. 1 Distributions of pH of biochars: (A) box-plots for four feedstocks; (B) individual values classified by pyrolysis temperature and feedstock. The number of observations for woody biomass, crop waste, grass and manure biochars is 74, 8, 40 and 78, respectively. In Fig. 1A the red and black lines indicate median and mean values, respectively while the circles in the boxplot indicate outliers. Data were taken from Campos et al. (2020); Domingues et al. (2017); Enders et al. (2012), Mukherjee et al. (2011) and Novak et al. (2009) cited in Singh et al. (2017); Kloss et al. (2012), Rajkovich et al. (2012) and Zhao et al. (2013) cited in Ippolito et al. (2020); Singh et al. (2010); and Zhang et al. (2017).

pH < 5. The pH of biochar generally increases with increasing temperature of the pyrolysis from 350 °C to 550 °C with a maximum of about pH 10. Further increases in temperature tend to lead to small decreases in pH (Fig. 1B).

It is the ash in biochar that can make it alkaline and hence effective as a liming agent, particularly on acid soils. Biochar's liming potential is measured as CaCO_3 equivalent, which is closely related to its ash content, and is greater in biochars made from manures, grasses, and crop wastes than from the woody biomass biochars (Fig. 3). The combination of plant nutrients and base cations in the ash make biochar valuable as both a fertilizer and a counter to the soil's acidity. In general, the richer are biochars in plant nutrients, the greater are crop yields resulting from their application.

The presence of salts in biochars often increases electrical conductivity (EC) of soils, and the increase in EC depends to some extent on the feedstock and pyrolysis temperature (Fig. 2), and the rate of biochar application to soil. Biochars from wood and grass have median ECs of 0.24 and 0.30 dS m^{-1} , respectively, whereas the median EC of biochar from crop residues is 0.57 dS m^{-1} but for biochar from manure it is much larger at 4.33 dS m^{-1} . The EC tends to increase with increasing pyrolysis temperature to between 500 °C and 600 °C, which is attributed to an increase in the residual ash during pyrolysis. Biochars produced at low temperatures from nutrient-rich feedstocks can still have large ECs, however, because de-protonated volatile molecules and small organic molecules are present.

Salts are part of the ash in biochars, which also tend to increase with increases in the temperature of pyrolysis, to a maximum about 600 °C, and with the duration of heating (Fig. 3). The composition of the ash also varies widely with differences in feedstock. In general, the richer is the feedstock in mineral elements, the more ash there is in the biochar. Fig. 3 shows the proportions of ash in biochar from grass (median 14.6%) to be the largest and that from woody biomass biochars (median 2.4%) to be least. We do not have precise estimates of the liming value of the ash, but we can expect that biochars from wood are likely to be less effective than those produced from the other three feedstocks (Fig. 3C and D). Nevertheless, strongly alkaline biochar is valuable for ameliorating acid soils.

Total carbon, nitrogen, oxygen, and hydrogen

The content of total C in biochar is important. It varies considerably from one feedstock to another. It is the greatest in wood biochars (mean 77.8%) and the smallest in biochars from manure (mean 62.1%) (Table 1). The proportions of C in biochars made from grass (mean 64.2%) and crop wastes (67.7%) are only slightly larger. For biochars that contain carbonates the total C concentration should be taken into account when the proportions of organic C are estimated. The proportions of nitrogen (N) in biochars rank in reverse order to those of C; the largest are in manure biochars and smallest in wood biochars (Table 1). The concentrations of O and H are much the same for all feedstocks, but both decrease with increasing temperature of pyrolysis as O- and H-containing compounds, such as ethers and lactones and the functional groups hydroxyl, carboxyl carbonyl, are lost. So too are the easily decomposable substances and volatile compounds containing N and S. In particular, the atomic ratios H/C_{org} and O/C_{org} decrease as the pyrolysis temperature increases. The net result is increased concentration of other major elements C, Ca, Mg, and K.

The nutrients P, K, Ca, and Mg and the trace elements copper (Cu), zinc (Zn), manganese (Mn) and iron (Fe) are generally most concentrated in manure biochars and are least in wood biochars (Table 2). The concentrations of most nutrient elements are closely related to the proportions of ash in the biochars.

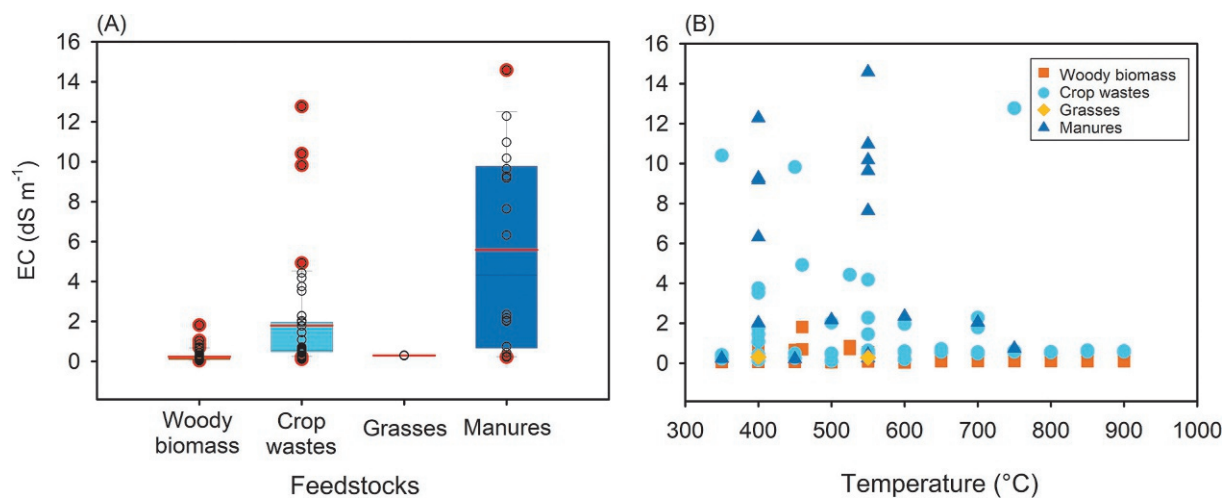


Fig. 2 Distributions of ash and CaCO_3 equivalent (liming value) of biochars: (A) box-plots of ash for four feedstocks; (B) individual values of ash classified by pyrolysis temperature and feedstock. (C) box-plots of CaCO_3 equivalent of biochars for four feedstocks; (D) individual values of CaCO_3 equivalent classified by pyrolysis temperature and feedstock. The number of observations for woody biomass, crop waste, grass and manure biochars is 68, 6, 36, and 72, respectively for ash content while these values are 10, 2, 8, and 16, respectively for the liming value. The red and black lines indicate median and mean values, respectively while the circles in the boxplot indicate outliers. Data sources are the same as given in Fig. 1.

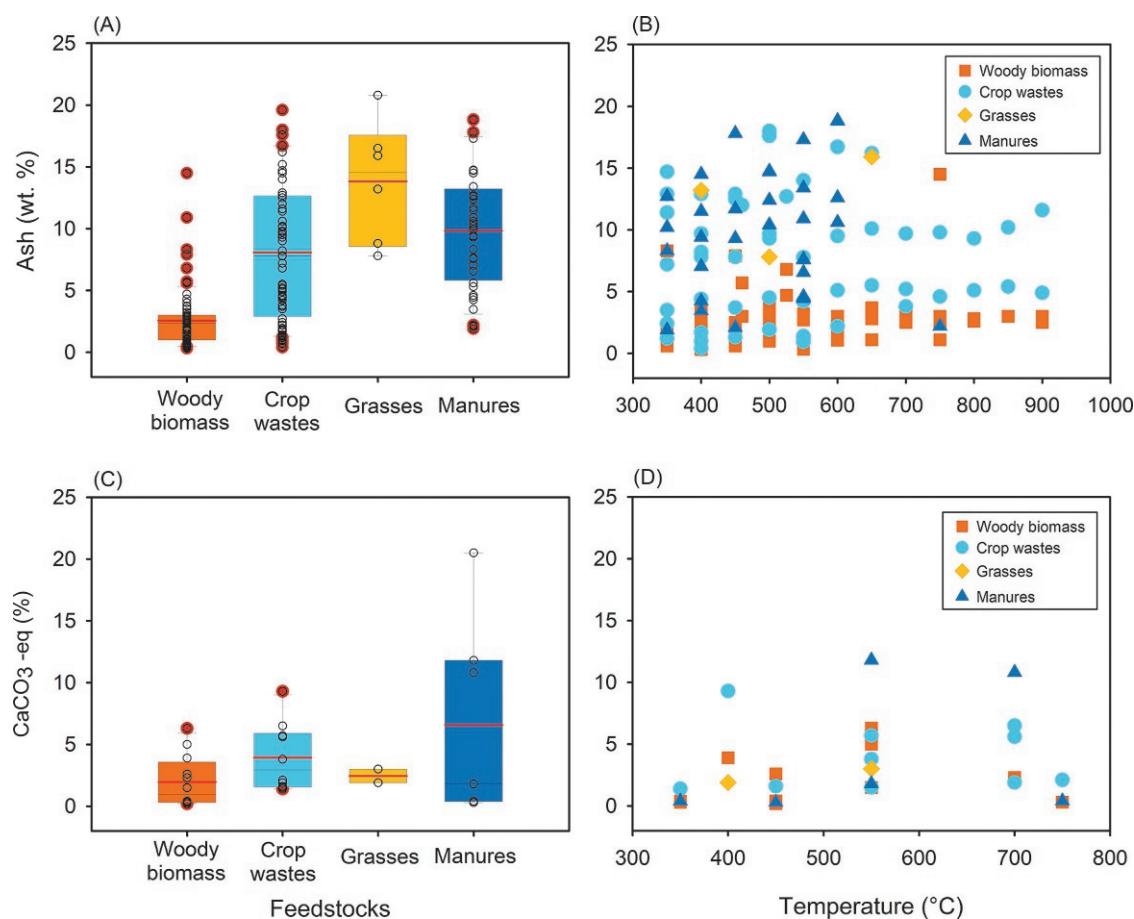


Fig. 3 Distributions of electrical conductivity (EC) of biochars: (A) box-plots for four feedstocks; (B) individual values classified by pyrolysis temperature and feedstock. The number of observations for woody biomass, crop waste, grass and manure biochars is 47, 2, 18, and 56, respectively. The red and black lines indicate median and mean values, respectively while the circles in the boxplot indicate outliers. Data sources are the same as given in Fig. 1.

Table 1 Elemental composition of biochars produced from different feedstocks (mean $\pm$ SE).

Feedstock	N ^a	C %	H %	O %	N %
Crop wastes	91	67.7 $\pm$ 1.3 b	2.8 $\pm$ 0.1	15.9 $\pm$ 0.9	1.3 $\pm$ 0.1 ab
Grasses	4	64.2 $\pm$ 6.8 ab	2.2 $\pm$ 0.1	14.0 $\pm$ 3.2	1.6 $\pm$ 0.2 a
Manures	25	62.1 $\pm$ 3.4 b	2.8 $\pm$ 0.2	14.2 $\pm$ 1.2	1.9 $\pm$ 0.4 a
Woody biomass	56	77.8 $\pm$ 0.1 a	2.6 $\pm$ 0.1	14.5 $\pm$ 1.0	0.4 $\pm$ 0.0 b
P value		<0.01	0.53	0.65	<0.01

Different letters represent statistically significant different in the same column, *p*-values are given for each of the elements.

^aN is the number of observations in each feedstock category.

Data were used from Břendová K, Száková J, Lhotka M, Krulíková T, Punčochář M, Tlustoš P (2017) Biochar physicochemical parameters as a result of feedstock material and pyrolysis temperature: predictable for the fate of biochar in soil? *Environmental Geochemistry and Health* 39: 1381–1395; Campos P, Miller, AZ, Knicker, H, Costa-Pereira, MF, Merino, A and De la Rosa, JM (2020). Chemical, physical and morphological properties of biochars produced from agricultural residues: Implications for their use as soil amendment. *Waste Management* 105: 256–267; Domingues, RR, Trugilho, PF, Silva, CA, de Melo, I, Melo, LCA, Magriotis, ZM, and Sanchez-Monedero, MA (2017). Properties of biochar derived from wood and high-nutrient biomasses with the aim of agronomic and environmental benefits. *PLOS ONE* 12(5): e0176884; Enders et al. (2012), Kloss et al. (2012), Novak et al. (2009), and Rajkovich et al. (2012) cited in Singh B, Dolk MM, Shen Q, Camps-Arbestain M (2017) Biochar pH electrical conductivity and liming potential. In: Singh B, Camps-Arbestain M and Lehmann J (Eds.) *Biochar: A Guide to Analytical Methods*. pp. 23–38. Boca Raton: CRC Press/Taylor Francis Group; Singh, B., Singh, B.P. and Cowie, A.L. (2010). Characterisation and evaluation of biochars for their application as a soil amendment. *Australian Journal of Soil Research* 48: 516–525; Tag AT, Duman G, Ucar S, Yanik J (2016) Effects of feedstock type and pyrolysis temperature on potential applications of biochar. *Journal of Analytical and Applied Pyrolysis* 120: 200–206; Zhang HZ, Chen CR, Gray EM, and Boyd SE (2017) Effect of feedstock and pyrolysis temperature on properties of biochar governing end use efficacy. *Biomass & Bioenergy* 105: 136–146; Zhao L, Cao X, Mašek O, and Zimmerman A (2013) Heterogeneity of biochar properties as a function of feedstock sources and production temperatures. *Journal of Hazardous Materials* 256–257: 1–9.

Table 2 Concentrations of major and micro-nutrient elements in biochars produced from different feedstocks (mean $\pm$ SE).

Feedstocks	N ^a	P	K	Ca	Mg	S	Cu	Zn	Mn	Fe
Crop wastes	23–60	1627 $\pm$ 130b	12,326 $\pm$ 1375b	6005 $\pm$ 1114b	2479 $\pm$ 310b	795 $\pm$ 97b	7.3 $\pm$ 2.5b	80 $\pm$ 18b	93 $\pm$ 13b	681 $\pm$ 132b
Grasses	2–3	1415 $\pm$ 528bc	7976 $\pm$ 2825abc	7418 $\pm$ 2182ab	1780 $\pm$ 1250bc	633 $\pm$ 33ab	3.0 $\pm$ 0.0ab	90 $\pm$ 67ab	56 $\pm$ 45b	5325 $\pm$ 5173a
Manures	11–29	4009 $\pm$ 402a	19,668 $\pm$ 1668a	24,297 $\pm$ 5811a	7038 $\pm$ 568a	2326 $\pm$ 303a	64.7 $\pm$ 12.3a	455 $\pm$ 102ab	118 $\pm$ 12b	737 $\pm$ 185b
Woody biomass	23–55	523 $\pm$ 76c	2405 $\pm$ 173c	6461 $\pm$ 727b	942 $\pm$ 86c	213 $\pm$ 48 b	16.3 $\pm$ 1.1b	544 $\pm$ 148a	334 $\pm$ 29a	341 $\pm$ 115b
P value		<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01

Different letters represent statistically significant different ($p < 0.01$) in the same column.

^aN is the number of observations in each feedstock category.

Data were used from Břendová K, Száková J, Lhotka M, Krulíková T, Punčochář M, Tlustoš P (2017) Biochar physicochemical parameters as a result of feedstock material and pyrolysis temperature: predictable for the fate of biochar in soil? *Environmental Geochemistry and Health* 39: 1381–1395; Campos, P, Miller, AZ, Knicker, H, Costa-Pereira, MF, Merino, A and De la Rosa, JM (2020). Chemical, physical and morphological properties of biochars produced from agricultural residues: Implications for their use as soil amendment. *Waste Management* 105: 256–267; Domingues, RR, Trugilho, PF, Silva, CA, de Melo, I, Melo, LCA, Magriotis, ZM, and Sanchez-Monedero, MA (2017). Properties of biochar derived from wood and high-nutrient biomasses with the aim of agronomic and environmental benefits. *PLOS ONE* 12(5): e0176884; Enders et al. (2012); Kloss et al. (2012); Novak et al. (2009); Rajkovich et al. (2012) cited in Singh B, Dolk MM, Shen Q, Camps-Arbestain M (2017) Biochar pH electrical conductivity and liming potential. In: Singh B, Camps-Arbestain M and Lehmann J (Eds.) *Biochar: A Guide to Analytical Methods*. pp. 23–38. Boca Raton: CRC Press/Taylor Francis Group; Singh, B., Singh, B.P. and Cowie, A.L. (2010). Characterisation and evaluation of biochars for their application as a soil amendment. *Australian Journal of Soil Research* 48: 516–525; Tag, A.T., Duman, G., Ucar, S., and Yanik, J. Effects of feedstock type and pyrolysis temperature on potential applications of biochar (2016) *Journal of Analytical and Applied Pyrolysis* 120: 200–206; Zhang, H.Z., Chen, C.R., Gray, E.M., and Boyd, S.E. (2017). Effect of feedstock and pyrolysis temperature on properties of biochar governing end use efficacy. *Biomass & Bioenergy* 105: 136–146; Zhao L, Cao X, Mašek O, and Zimmerman A (2013) Heterogeneity of biochar properties as a function of feedstock sources and production temperatures. *Journal of Hazardous Materials* **256–257**: 1–9.

Surface charge characteristics of biochars

Surface charge on biochar particles is generated mainly by the de-protonation of hydroxyl, carboxylic, lactonic and phenolic functional groups. These groups differ in their pK_a values; those for most carboxylic groups fall between 2 and 4, those for lactonic groups between 4 and 7, and for phenolic groups from 7 to 10 and more. Within the normal range of pH for most agricultural soil, carboxylic and lactonic groups of biochar predominantly de-protonate and thereby contribute to the cation exchange capacity (CEC). Given that native soil organic matter has larger CECs (average $280 \text{ cmol}_c \text{ kg}^{-1}\text{C}$) at pH 7 than fresh biochar (usually $<100 \text{ cmol}_c \text{ kg}^{-1}\text{C}$), the latter only contributes significantly to the soil's CEC in the short-term in soils with a small CEC, such as sandy soils and those poor in organic matter. Over time, however, biochar is oxidized, and functional groups of C form on biochar surfaces and increase the CEC of most soils. Such data as we have suggest that CECs are broadly similar for biochars made from different feedstocks and at different pyrolysis temperatures (Fig. 4): median CECs are $9.7 \text{ cmol}_c \text{ kg}^{-1}$ for biomass from wood and crop residues, $9.8 \text{ cmol}_c \text{ kg}^{-1}$ from grass, and $10.7 \text{ cmol}_c \text{ kg}^{-1}$ from manures. No clear relationship between CEC and pyrolysis temperature is evident, although we might expect that cleavage and loss of acidic functional groups caused by increasing temperature would result in a decrease in CEC.

Fresh biochars have some anion exchange capacity (AEC), but this is generally smaller than their CEC, and it decreases rapidly over time. Singh et al. (2010) analyzed 10 biochars made from various feedstocks and found values of AEC ranging from 31 to $1030 \text{ cmol}_c \text{ kg}^{-1}\text{C}$, with generally smaller values for wood biochars and larger ones for biochars made from paper sludge and manure.

Bulk density, pore space, and specific surface area

Biochars are generally highly porous, particularly those produced from grasses and woody materials. Feedstock type, pyrolysis conditions and/or pre- and post-pyrolysis processing influence the particle size, density, porosity, and hydrophobic or hydrophilic properties of biochars. Fig. 5 summarizes the bulk density, pore volume, and specific surface area of biochars produced from the principal feedstocks. Biochars from grass and wood have median bulk densities $<0.2 \text{ g cm}^{-3}$, while the medians for crop waste and manure biochars are 0.23 and 0.39 g cm^{-3} , respectively. Pore volume is greatest in woody biochars (median 57.2%) with much smaller values for manure biochars (median 4.9%) and even smaller values for grasses and crop waste biochars (median 1.1% and 0.04% , respectively). The specific surface area of woody biochars is the greatest (median $108.3 \text{ m}^2 \text{ g}^{-1}$), followed by manure (median $45.0 \text{ m}^2 \text{ g}^{-1}$), and crop waste (median $28.1 \text{ m}^2 \text{ g}^{-1}$) biochars, and the minimum is for grass biochars (median $8.7 \text{ m}^2 \text{ g}^{-1}$). There is no consistent trend in physical properties with pyrolysis temperature for different feedstock biochars.

Chia et al. (2015) showed that micropores contribute significantly to the surface area of biochars. Based on the data from several published studies, they showed a linear relation between the micropore volume and BET (Brunauer–Emmett–Teller) surface area of many biochars. They found that both the micropore volume and BET surface area were much greater for activated biochar (range 0.056 to $0.72 \text{ cm}^3 \text{ g}^{-1}$ and 156 to $2436 \text{ cm}^3 \text{ g}^{-1}$, respectively, whereas the corresponding values for non-activated biochars were 0.002 – $0.2 \text{ cm}^3 \text{ g}^{-1}$ and 4 – $448 \text{ cm}^3 \text{ g}^{-1}$).

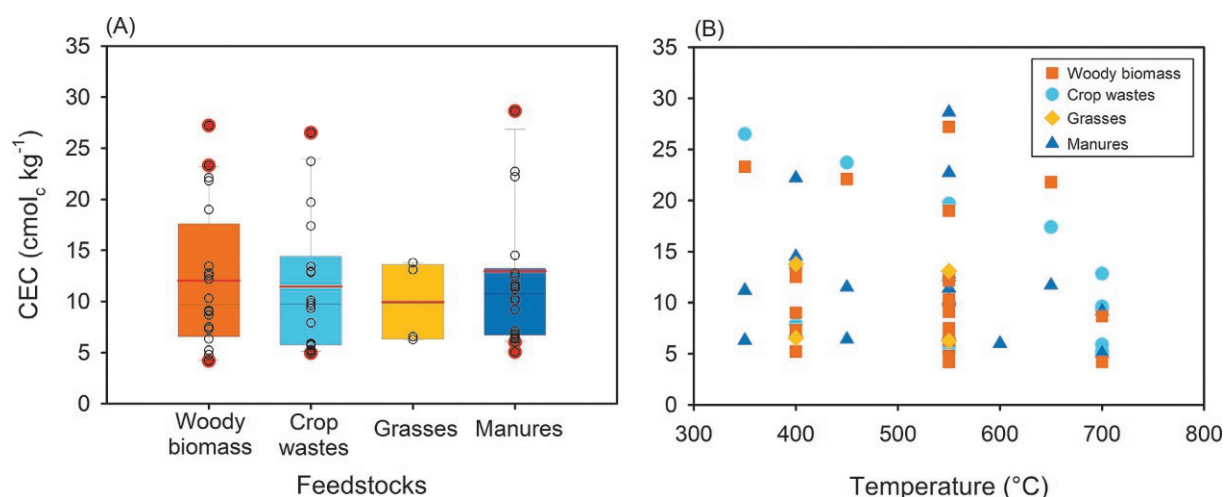


Fig. 4 Distributions of cation exchange capacity (CEC) of biochars: (A) box-plots for four feedstocks; (B) individual values classified by pyrolysis temperature and feedstock. The number of observations for woody biomass, crop waste, grass, and manure biochars is 20, 18, 4, and 22, respectively. The red and black lines indicate median and mean values, respectively while the circles in the boxplot indicate outliers. Data were taken from Ippolito et al. (2020); Souza et al. (2021) and Zama et al. (2017).

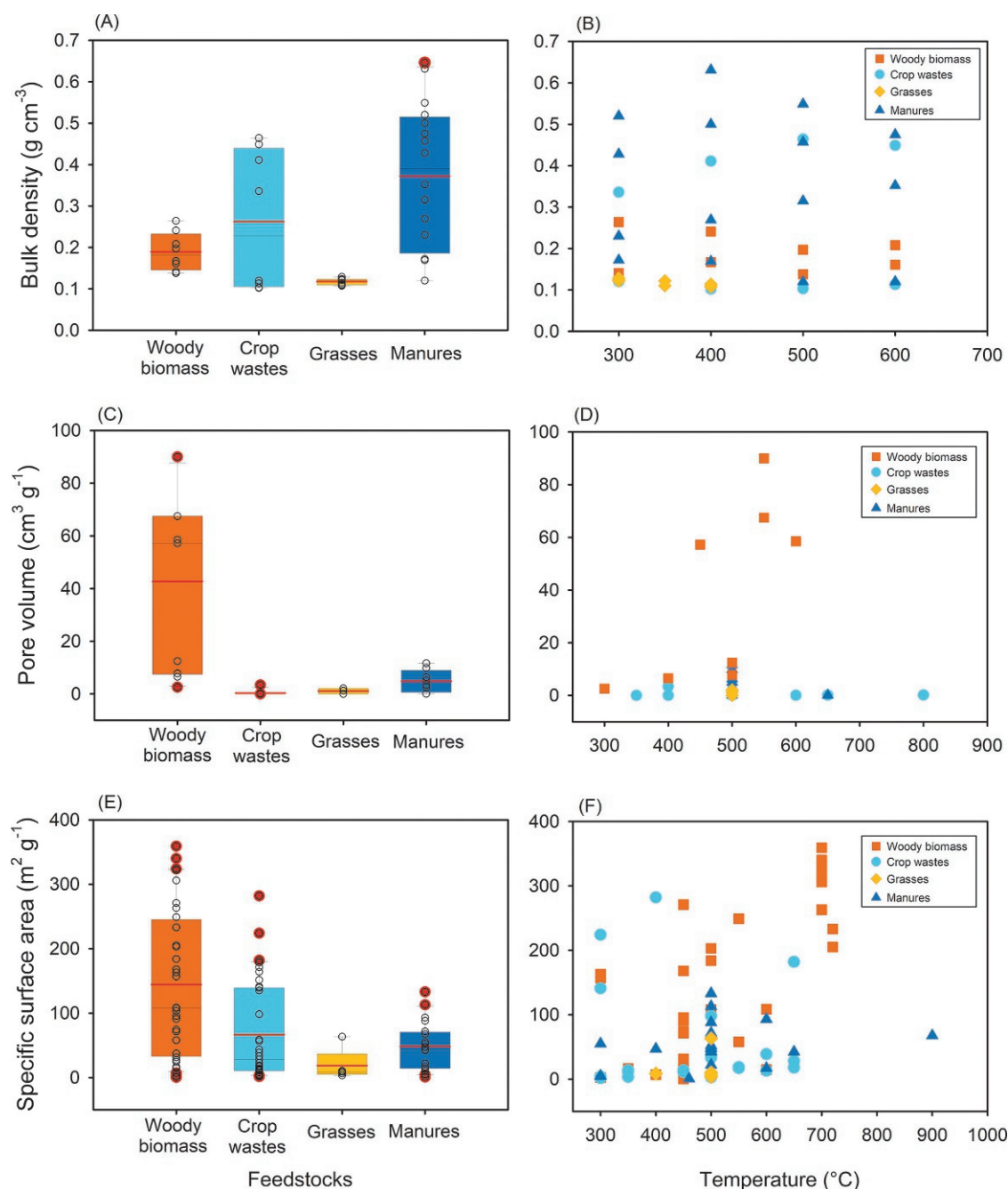


Fig. 5 Distributions of bulk density, pore volume and specific surface area of biochars: (A) box-plots of bulk density for four feedstocks; (B) individual values of bulk density classified by pyrolysis temperature and feedstock. (C) box-plots of pore volume for four feedstocks; (D) individual values of pore volume classified by pyrolysis temperature and feedstock. (E) box-plots of specific surface area for four feedstocks; (F) individual values of specific surface area classified by pyrolysis temperature and feedstock. The number of observations for woody biomass, crop waste, grass, and manure biochars was 12, 3, 8, and 10, respectively for bulk density while these values 8, 9, 16, and 8, respectively for pore volume. For specific surface area, the values are 36, 6, 20, and 32 respectively. Data were taken from Baltreñaitė-Gedienė et al. (2020); Behazin et al. (2016), Chen et al. (2011), Jeong et al. (2015), and Rajkovich et al. (2012) are cited in Ippolito et al. (2020); Esfandbod et al. (2017); Keiluweit et al. (2010) cited in Ye et al. (2020); Mia et al. (2017); Nzediegwu et al. (2021); Yuan et al. (2018); Zama et al. (2017); Zhang et al. (2019); Zhao et al. (2017); Zhao et al. (2019).

Biochar applications

Biochar for long-term carbon storage in the soil

Biochar remains in the soil for much longer than uncharred biomass; that is its advantage for C retention. To maximize that advantage for soil acting as a C sink, the biochar production system should be tuned so that the output contains large concentrations of C (Table 1) with large and condensed aromatic moieties or structures, while minimizing the C used to collect, transport and store the feedstock, in the pyrolysis, post-processing, and during application to the land.

To maintain that advantage, biochar must be produced only from organic wastes such as crop residues and prunings or from trees grown in short rotation for the purpose. In the latter case the next generation of trees has the further advantage of capturing C from the atmosphere. This advantage can be expressed as carbon debt (Lehmann et al., 2021); that is the difference in the amount of CO₂ released through pyrolysis and that released during the decomposition of the same amount of un-pyrolyzed feedstock. It needs to be small. Most crop residues have a small C debt; purpose-grown trees tend to have a larger debt. In the latter situation planners must show that there is or will be a net reduction in the amount of C lost to the atmosphere over the life-cycle of a rotation and thereby reduce the damage to the environment.

Biochar contributes to the labile fraction of C in soil, especially if produced at low temperature. This fraction together with plant nutrients and alkalinity can stimulate microbial activity and trigger the mineralization of native soil organic matter in the short term. This is also referred to as positive priming. The increase in alkalinity might also weaken the bonds of organic ligands to mineral surfaces, and thereby increase the labile pool of C. In the long term, however, biochar has a net negative priming effect because its surfaces provide new sites on which organic ligands can be preserved. The overall effect is to sequester C and reduce the amount of CO₂ released into the atmosphere (Lehmann et al., 2021).

The overall CO₂ removal created by biochar production relies heavily on the C sink potential of the biochar. The rate of degradation of the biochar C must be several orders of magnitude slower than that of un-pyrolyzed feedstock C when that is left to decompose naturally. The 'Stable C Protocol' was developed by the International Biochar Initiative (Budai et al., 2013) and is based on the relation between the H/C_{org} ratio of biochars and their rates of degradation determined in incubation experiments with biochar in soil of 3–5 years duration. The methodology has been further refined first by Lehmann et al. (2015) and, recently, by Woolf et al. (2021) by the provision of a regression equation as shown in Fig. 6, which provides estimates of the fraction of organic C (C_{org}) in biochar, referred to as F_{perm}, that will remain in the soil for more than 100 years at a soil temperature of 14.9 °C (the mean annual air temperature in croplands globally). The equation is—

Here H/C_{org} is the molar ratio of H and C in organic forms. The choice of 100 years arises from the time frame used by the Intergovernmental Panel on Climate Change (IPCC) for assessing the potential of greenhouse gases to warm the planet (IPCC, 2007). The authors also provided permanence coefficients for biochar C sequestration as a function of other soil temperatures (ranging from 5 °C to 25 °C) for different time frames (100–1000 years).

The atomic H/C_{org} and therefore F_{perm} are strongly associated with the degree of aromatic condensation of C in biochar. The IPCC (2019) has recently proposed a method of assessment that would allow biochar to be included in national inventories as a soil

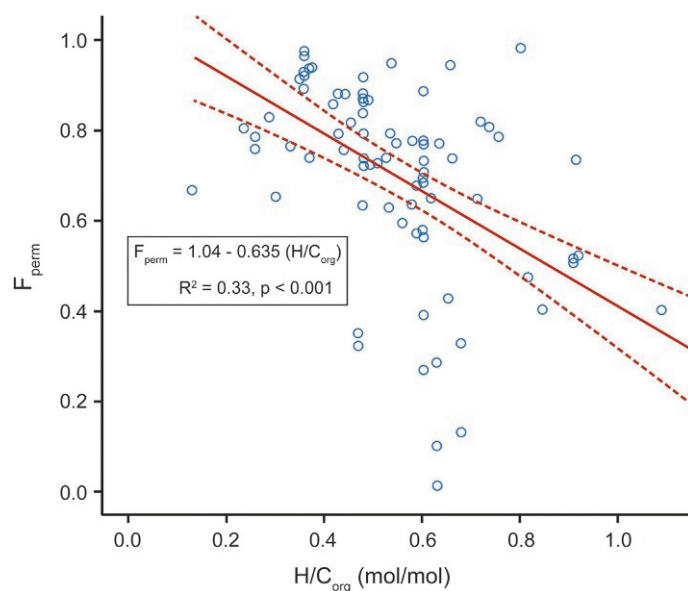


Fig. 6 Relationship between F_{perm} (the fraction of biochar carbon remaining in soil after 100 years) and the molar ratio of hydrogen to organic carbon (H/C_{org}) in biochar. Biochar pyrolysis temperatures below 350 °C were excluded in the data. In all cases, F_{perm} values were calculated for a soil temperature of 14.9 °C (the mean annual air temperature in croplands globally). The dashed lines indicate 95% confidence intervals. From Woolf, D, Lehmann, J, Ogle, S, Kishimoto-Mo, AW, McConkey, B, and Baldock, J (2021). Greenhouse gas inventory model for biochar Additions to soil. *Environmental Science & Technology* 55(21):14795–14805.

amendment. It would use predicted values of F_{perm} (also referred to as BC_{+100}) based on the pyrolysis temperature, because that information is more readily available than the $\text{H}/\text{C}_{\text{org}}$ ratio.

Fertilizer benefits of biochar in soil

The concentration of nutrient elements in a feedstock is the single most important factor governing the potential fertilizer value of a biochar. Those concentrations vary widely, and not all biochars have value as fertilizer. The International Biochar Initiative (IBI) has set certification standards for the fertilizer values of biochars; five classes are recognized, based on their total and available concentrations of P (phosphorus), K (potassium), S (sulfur) and Mg (magnesium). Biochars made from wood generally have the smallest nutrient concentrations, and those from manures (and food wastes) have the largest. Nutrient concentrations in biochars from crop residues and grasses are intermediate (Table 2). The proportion of N (nitrogen) retained from the original feedstock depends on the temperature of the pyrolysis; for example, Lang et al. (2005) found that the proportion in biochar from three woody and four herbaceous feedstocks started to decline at about 400 °C, and nearly half of the N was lost at about 750 °C. Other investigators have reported similar effects, in particular for manures. The N available for plant nutrition also declines with the increase in pyrolysis temperature because aromatic and heterocyclic N–ring structures are formed, and these are slow to disintegrate in the soil. Phosphorus, in contrast, has been found to increase in proportion as the temperature of the pyrolysis increases to 600 °C, but it can decrease at higher temperatures.

The available P (i.e., that extracted in 2% formic acid) of biochars, rich in ash, ranges between 26 and 93% of the total P content, and this value largely depends on the type of feedstock and pyrolysis temperature (Wang et al., 2012). The concentration of total K has been found to increase from 3.7% at 300 °C to 5.0% at 600 °C, and that which is available (i.e., water-soluble K) to increase from 37% to 47% of the total K (Zheng et al., 2013). Other major nutrients, namely S, Ca, and Mg also remain in the ash fraction, and the amounts therefore depend on the type of feedstock (Table 2). The proportion of extractable S has been found to increase from 10 to 76% with an increase in temperature, and 35% to 100% of the Ca and 18% to 95% of the Mg are readily soluble.

Influence of biochar on soil physical properties

Addition of biochar to soil can change the soil's pore-size distribution, water-holding capacity, hydraulic conductivity and total porosity (and so decrease the bulk density). The extent of the changes depends on the properties of the biochar (including its particle-size distribution) and of the soil. Decreases in bulk density of 7.6% to 12.0% on average have been reported (Table 3), the largest being for fine-textured soil, and less for medium- and coarse-textured soil. Recently, Blanco-Canqui (2017) found the

Table 3 The effects of biochar application on important soil physical properties.

<i>Soil property</i>	<i>Biochar's effect on soil property</i>	<i>Number of studies</i>	<i>References</i>
Bulk density	Decreased—		
	3–31%; average = 12%	19	Blanco-Canqui (2017)
	6.5–8.7; average = 7.6%	128	Omondi et al. (2016)
Total porosity	average = 9%	111	Razzaghi et al. (2020)
	Increased—		
	2–41%	14	Blanco-Canqui (2017)
Plant-available water	5.8–11.1; average = 8.4%	38	Omondi et al. (2016)
	Increased—		
	4–130%; average = 72%	29	Blanco-Canqui (2017)
Water retention	5.8–11.1; average = 8.4%	38	Omondi et al. (2016)
	average 45% in coarse textured soils average 21% in coarse textured soils	44	Razzaghi et al. (2020)
	average 14% in coarse textured soils	68	
Saturated hydraulic conductivity	Increased—average = 90% in most soils but not in clayey soils	31	
	Decreased—	19	Blanco-Canqui (2017)
	7–2270%; average = 392% in coarse textured soils	28	Blanco-Canqui (2017)
Mean weight diameter of water stable aggregates	Increased—25–328%; average = 98% in fine textured soils	13	
	Increased—		
	average = 17.8% in fine textured soils	18	Omondi et al. (2016)
	average = 27.3% in medium textured soils	30	
	average = 36.5% in coarse textured soils	28	
	Increased—		
	4–58%	21	Blanco-Canqui (2017)
	1.1–15.7%; average = 8.2%	10	(Omondi et al., 2016)

opposite trend, however, with the addition of biochar increasing the porosity of coarse-textured soil most. The effect was attributed to several factors, in particular increased interaction with mineral particles favoring their aggregation and resistance to packing.

Results from both field and laboratory suggest that addition of biochar can alleviate compaction in modern agriculture. The increased porosity of the soil enhances the movement of water, heat and gases, though these effects do not necessarily benefit plant growth. The effects on hydraulic conductivity depend strongly on the soil's texture. In particular, they decrease with increasing clay content (Table 3). Blanco-Canqui (2017) reported a decrease in coarse-textured soils, an increase in fine-textured and compacted soil, and no effect on medium-textured soil (Table 3). The variation in response has been explained largely by big differences in the pore-size distribution of both soil and biochar and by biochar particles' filling of the existing pore space. Most studies have shown that biochar increases water retention and available water in soil, with coarse-textured soils being more responsive than fine-textured ones.

Herath et al. (2013) showed that in two silty loam soils the change in soil water availability was quite different with the application of identical types and amounts of biochar. The increase in plant available water was greater in the Alfisol than Andisol with the application of maize stover biochars produced at 350 °C and 550 °C (Fig. 7). The increase in water retention and available water in coarse-textured soil has been attributed to a decrease in the average pore size and an increase in the proportion of micropores and specific surface area. Not all applications of biochar have increased the soil's water retention, however; the hydrophobicity of fresh biochar is the likely reason. As biochar ages in the soil, O-containing surface functional groups form on biochar surfaces so that they become hydrophilic and both infiltration and retention increase. These observations are based mostly on short-term laboratory studies. Long-term studies in the field are needed to confirm the effects, which also depend on the rate of application.

Effects of biochar application to soil on crop yield

Experimental results on the effects on plant production and crop yield published since 2010 report estimated average increases of between 9 and 17% in the yields of field crops (Fig. 8). In most, yield responses have been largest on acidic and sandy soil, on soil

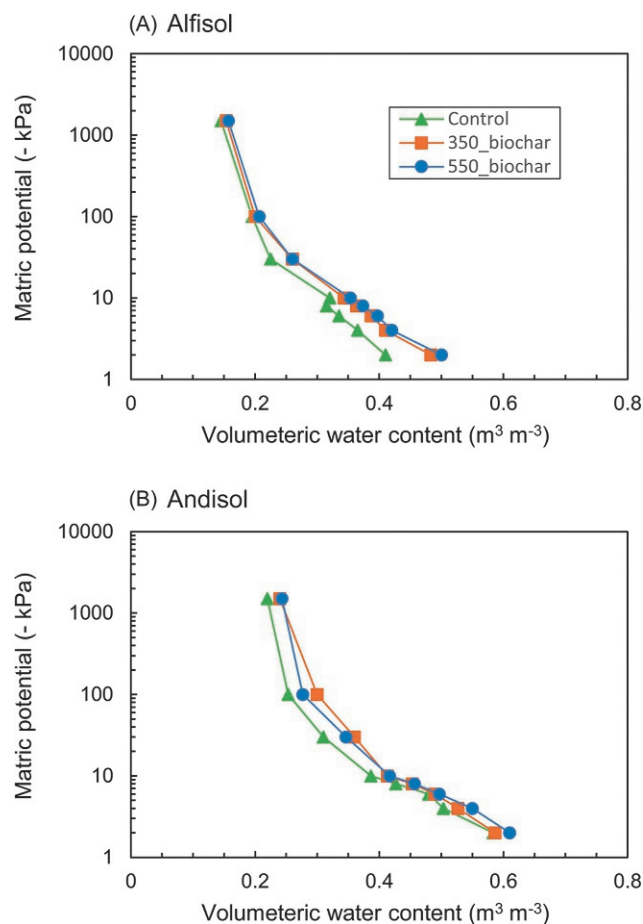


Fig. 7 Matric potentials of soil mixed with maize stover biochar for two soil types: (A) an Alfisol and (B) an Andisol. The plotted symbols are means of measurements on three samples. From Herath, HMSK, Camps-Arbestain, M, Hedley, M, et al. (2013). Effect of biochar on soil physical properties in two contrasting soils: An alfisol and an andisol. *Geoderma* 209–210: 188–197.

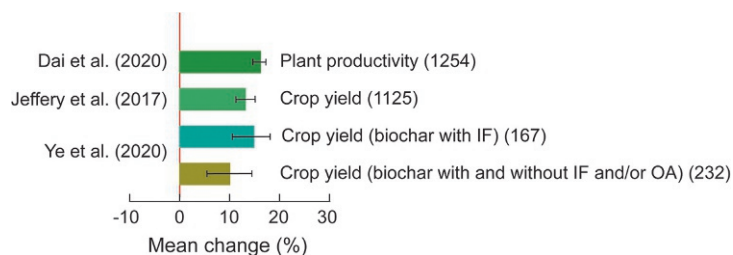


Fig. 8 The mean effects of biochar application on plant productivity or crop yield in soils (redrawn after Jeffery et al. (2017); Dai et al. (2020) and Ye et al. (2020)). In the first two studies the effects of biochar application are compared to undefined controls (Jeffery et al., 2017; and Dai et al., 2020), whereas Ye et al. (2020) evaluated (i) the effects of biochar application together with inorganic fertilizer (IF) compared to 'fertilized control' and (ii) the effects of biochar application with and without IF and or organic amendment (OA) relative to the fertilized control. The error bars represent 95% confidence intervals in the three datasets. The number of pair-wise observations for each study are given in the parentheses.

with small CEC and little organic matter, particularly in the tropics. As one might expect, addition of nutrient-rich biochar has enhanced yields more than less rich biochar.

There is a growing interest in the potential benefits of fertilizers augmented with biochar. Ye et al. (2020) showed that biochar in combination with inorganic fertilizers, increased crop yield by 14.5% compared with fertilized controls. The overall increase in yield with biochar (with and without inorganic fertilizer and or organic amendment) relative to the fertilized control was 9.9%. The authors only considered field studies, with application rates $\leq 20 \text{ t ha}^{-1}$, and yield data within first year of application. They found that application rates of $>10 \text{ t ha}^{-1}$ did not contribute to further increases in crop yield.

Combined biochar and inorganic fertilizer (biochar-based fertilizers) can locally increase the CEC around the fertilizer granules, water retention and microbial activity. From a recent meta-analysis, Melo et al. (2022) reported that when C-rich biochars produced at a high pyrolysis temperature were part of these composite fertilizers, plant growth was greater than that achieved with conventional fertilizers. The authors suggested the potential role that this type of biochar may have is as electron shuttle in the rhizosphere that improves nutrient uptake.

Biochar and the soil microbial community

Biochar can profoundly influence the soil's biome. The following are reported to have been increased in soil after the addition of biochar (Zhang et al., 2018), with the mean increase in parentheses: activity of all microorganisms (total phospholipid-derived fatty acids (PLFA), 8.3%), bacteria (20%), fungi (19%), actinomycetes (9.1%), Gram-positive bacteria (11%), and Gram-negative bacteria (13%). These effects of biochar can be attributed directly to a safer habitat, better nutrition, more pore space and more available water. They might also have been brought about less directly by a more favorable pH and greater CEC. Addition of biochar could therefore change the composition and activity of the microbial community, and the soil food web, which might affect the cycling of soil organic matter and major nutrients, and consequently affect plant growth. The interactions of biochars with soil minerals enhance water and nutrient retention, which in turn would increase microbial biomass over time.

Biochar can increase the abundance of bacteria involved in N cycling including denitrifiers and thereby reduce emissions of N_2O from the soil (Borchard et al., 2019). It can increase biological fixation of by promoting the activity of *Rhizobia*, too. These effects depend to some extent on the nature of both the soil and the biochar, and on the rate and method by which the biochar is applied (Borchard et al., 2019). Biochars with liming potential when added to acid soil decrease the exchangeable acidity and concentration of free aluminum. This causes a shift in the soil's microbiome. Biochars produced at low temperature when added to soil containing little organic C ($<1\%$) were found to increase the fungi-to-bacteria ratio by 38%, and this response was larger in acidic soil than in neutral and alkaline soil. Acid conditions favor fungi over bacteria.

Some biochars carry toxic constituents, such as polycyclic aromatic hydrocarbons (PAHs) that can be produced during the thermal carbonization process. Pine biochars produced at low temperature ($\leq 400^\circ\text{C}$) can contain phenolic compounds, such as retene, in their water-extractable fractions. These can harm soil microorganisms or inhibit their growth. In some instances, the total microbial biomass and mycorrhizae decreased after the addition of biochar, seemingly because of nutrient limitations. Increasing the dominance of certain microbial species, following biochar application, can reduce the diversity in the soil's whole microbial community.

Biochar standardization and quality assurance for potential contaminants

Given the diverse qualities of biochar, several agencies, including the IBI, the European Biochar Certificate and the Australia New Zealand Biochar Initiative (ANZBI) have established voluntary standards for the material; they have sought a legislated system of certification (Meyer et al., 2017). This is important because some biochars contain contaminants, either present in the feedstock or formed during the pyrolysis.

Potentially toxic elements (PTE) include cadmium (Cd), copper (Cu), nickel (Ni), chromium (Cr), zinc (Zn), lead (Pb) and arsenic (As). Their concentrations depend primarily on the feedstock, and they can become concentrated during pyrolysis. If their initial concentrations in the feedstocks are small, however, they are unlikely to exceed limits set by the regulatory agencies. Biochars from the redwood (*Sequoia sempervirens* (D. Don) Endl.), rice straw, maize, bamboo, and softwood were found to have the following concentrations (in mg kg^{-1}): As 0.03–0.27, Cd 0.02–0.94, Cr 0.12–6.48, Cu 0.04–13.2, Ni 0.1–1.37, Pb 0.06–3.87, and Zn 0.94–207. The concentrations varied substantially from one feedstock to another; bamboo biochar had the largest concentrations of Ni (1.37 mg kg^{-1}), Pb (3.87 mg kg^{-1}), Zn (207 mg kg^{-1}), and As (0.29 mg kg^{-1}); redwood biochar had the largest concentration of Cd (0.94 mg kg^{-1}); maize biochar had the largest concentrations of Cr (6.48 mg kg^{-1}) and Cu (10.6 mg kg^{-1}). The concentrations of Pb and Zn increased almost two-fold with increasing pyrolysis temperature from 300°C to 600°C . All metals, except for the concentration of Zn in the bamboo biochar pyrolyzed at 600°C , were less than the ambient background concentration as defined by the International Organization for Standardization, ISO (ISO, 2005).

Organic contaminants can also be generated during pyrolysis. The benzo(a)pyrene group and other PAHs are persistent; they form as a result of incomplete combustion. Two fundamental mechanisms lead to their formation during slow pyrolysis: (i) condensation, carbonization, and aromatization of solid materials at temperatures $<600^\circ\text{C}$, with a peak at 500°C , and (ii) pyrosynthesis, during which gaseous hydrocarbon radicals generated at temperatures $>500^\circ\text{C}$ react to form polyaromatic ring structures, with the dominance of PAHs in tar and gas over the solid phase. Low-molecular-weight PAHs (with two or three rings) are usually formed at temperatures $<500^\circ\text{C}$, whereas the high-molecular-weight PAHs (five and six rings) are produced at temperatures exceeding 500°C . Naphthalene and phenanthrene usually comprise the largest contributions to total PAHs in biochars.

Recorded concentrations of PAHs vary widely. Small concentrations ($\Sigma 16$ EPA: Environmental Protection Agency has identified the 16 most frequently occurring and/or dangerous PAHs as priority pollutants) of $0.07\text{--}3.27 \text{ mg kg}^{-1}$ in 50 biochars produced at temperatures between 250°C and 900°C from manure, paper-mill waste, pine wood, and maize straw, and $<0.5 \text{ mg kg}^{-1}$ in 11 biochars from five different feedstocks. In contrast, large concentrations have been reported: $1.2\text{--}19 \text{ mg kg}^{-1}$ in 11 biochars produced from sawdust, maize stover, and wood waste at temperatures ranging from 350°C to 850°C , and 64.7 mg kg^{-1} in rice-husk biochar and 145 mg kg^{-1} in a pinewood charcoal.

Wood gasification gives rise to the largest concentrations of PAHs because the compounds undergo condensation in solid residues at the high temperatures prevailing. The amounts of PAHs in biochar can be minimized during production by separation of the gases, and the process can be controlled at high temperatures to produce biochars with small concentrations of PAHs. A wide range of waste feedstocks can be pyrolyzed at temperatures $>500^\circ\text{C}$ to produce biochars that (i) have concentrations of PAHs less than soil background concentrations, and (ii) have low bioavailability of PAHs. Furthermore, although some biochars can have large concentrations of PAHs that exceed established limits in the standards, the PAHs are largely unavailable to biological organisms because they are strongly adsorbed on biochar surfaces. The biochars act more as a sink than as a source of PAHs, and hence they can be used to sequester PAHs that happen to be in the soil.

Polychlorinated dibenzo-p-dioxins and -furans (PCDD and Fs) can form when a biomass containing large concentrations of chlorine is pyrolyzed. For example, pyrolysis of food wastes containing salt (2.9% Cl) were found to produce a toxic dioxin concentration of 1.2 pg g^{-1} TEQ (toxic equivalency). Plants growing in salty soil or irrigated with saline 'gray' water with large concentrations of chlorine ($>1\%$ Cl) can produce biochar with large concentrations of PCDD and Fs. If feedstocks with only small concentrations of Cl are selected, then the biochar itself produced should contain small PCDD and F concentrations and comply with the standards for organic contaminants.

Remediation of contaminated soil

Biochars can sorb many contaminants, both inorganic such as heavy metals, ammonium, nitrate, phosphate, and organic including agricultural chemicals, antibiotics, PAHs, polychlorinated biphenyls (PCBs), volatile organic compounds (VOCs) and aromatic dyes. They immobilize cationic trace metals, primarily by adsorption and precipitation. At soil pH more acid than their points of zero charge, biochars are expected to be positively charged and to adsorb anionic pollutants, such as nitrate (NO_3^-), phosphate (PO_4^{3-}), arsenate (AsO_3^-) and chromate (CrO_4^{2-}). Organic contaminants are retained by biochars with chemical interactions on surfaces and in pores, and highly porous biochars with abundant aromatic structures and functional groups adsorb organic compounds in various ways. The interactions of organic contaminants with biochar can involve inner- or outer-sphere complexation and other processes such as pore diffusion, van der Waals forces, H-bonding, π - π electron donor-acceptance, universal electrostatic attraction between non-polar molecules (London dispersion forces) and attraction between polar molecules (dipole-dipole forces) involving COOH and OH functional groups. Organic contaminants sorbed on biochars are eventually degraded by microorganisms.

The capacity of biochars to sorb and immobilize contaminants in soil and water depends on their characteristics, which, as above governed by pyrolysis temperature, feedstock and other treatments. In general, biochars pyrolyzed at high temperatures ($>500^\circ\text{C}$) have larger specific surface areas, microporosity, condensed aromaticity and hydrophobicity, and therefore a greater affinity for organic contaminants. Specific surface areas of biochars made from rice straw are reported as $19.7 \text{ m}^2 \text{ g}^{-1}$ pyrolyzed at 400°C , $44.0 \text{ m}^2 \text{ g}^{-1}$ at 500°C and $193.2 \text{ m}^2 \text{ g}^{-1}$ at 600°C . The adsorption of the insecticide cyromazine (N-cyclopropyl-1,3,5-triazine-2,4,6-tri-amine) on biochars also increased with increasing temperature; it was greatest for the biochar produced at 600°C .

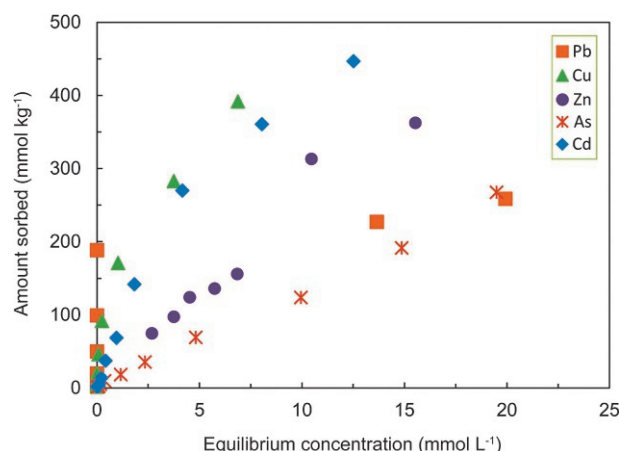


Fig. 9 Sorption of As, Cd, Cu, Pb and Zn on a wood biochar (*Eucalyptus saligna* Sm.) pyrolyzed at 550 °C from initial loadings of the trace elements between 2 and 800 mmol at an initial solution pH = 7 and in the presence of 0.01 m Ca(NO₃)₂ as a background electrolyte. From Namgay T, Singh, B and Singh BP (2010) Influence of biochar application to soil on the availability of As, Cd, Cu, Pb, and Zn to maize (*Zea mays* L.). Australian Journal of Soil Research 48: 638–647.

Not all consequences of the addition of biochar to soil are desirable. Fresh biochars, due to their hydrophobic surfaces and large internal surface area, have a strong affinity for organic herbicides and pesticides. These agrochemicals can be less effective than intended where biochar has been applied at large rates (>20 t ha⁻¹), the amounts of the chemicals need to be adjusted accordingly in soil amended with biochar. In time the surfaces of the biochar become oxidized and hydrophilic, and its effect attenuates or even reverses.

Biochars produced at low pyrolysis temperatures (<500 °C) are generally preferred for the adsorption of inorganic contaminants because they have functional groups with which to react. A biochar produced at 350 °C was found to be more effective than a high temperature biochar in retaining trace metals, Pb²⁺ and Cu²⁺, in an acid soil. It seems that functional groups on the surfaces of the biochars retain cationic trace metals in the soil. The following sequence of sorption of trace elements on biochar made from eucalyptus (*Eucalyptus saligna* Sm.) wood at 550 °C has been reported as Pb > Cu > Cd > Zn > As (Fig. 9; Namgay et al., 2010). The pH of the equilibrium solution increased with increasing sorption of the cations which suggested a cation exchange reaction involving metallic cations and H⁺ ions. There was an increase in the solution's pH in the presence of As, possibly because of exchange between arsenate (AsO₄⁻) and hydroxyl (OH⁻) ions on biochar surfaces.

Producers can now customize biochars for specific applications. They can make biochars from high-quality woody feedstocks and activate the surfaces after pyrolysis so that the product will remove both cationic and anionic contaminants from water and contaminated soil at a moderate cost. In one case, biochar made from willow feedstock sequestered organochlorine pesticides at old sheep dip sites that were also contaminated with As.

Biochar in composting of biological wastes

Producing biochar from waste, instead of composting it, can significantly reduce emissions of greenhouse gases. Composting produces CH₄ and N₂O. The IPCC (2006) has set default emissions for the composting of biological wastes; they are 10 g CH₄ and 0.6 g N₂O per kg dry waste. The wetness of many of these wastes, such as municipal green waste, waste food, sewage sludge and animal manure, make them unsuitable for biochar production, however. Wastes suitable for biochar can be separated from those better suited for composting. The biochar can then be used as a bulking agent during composting, and the combination produces less greenhouse gas and recycles nutrients for agricultural crops. Reduction in emissions of both CH₄ and N₂O during composting are estimated to be 80% under the most favorable conditions.

Porous and reactive surfaces of biochar provide an aerobic environment and improve the structure of the composting pile, and that change stimulates aerobic microorganisms (e.g., methanotrophs and nitrifiers), increases the oxidation of C to CO₂ and N to NO₃⁻, favors the retention of nitrogenated molecules on biochar surfaces, and decreases emissions of CH₄ and N₂O. It has also been found that addition of biochar to composting substrate reduced losses of total N by 30.2%, NH₃ by 52.6% and N₂ by 66.2%.

The amount of CH₄ produced depends strongly on the proportion of biochar in the compost and characteristics of the organic residue in the composting process. Reported examples include a 55% reduction in CH₄ emission when chicken manure was composted with 20% biochar, and reductions of >70% and >80% were reported for cow manure and municipal solids, respectively, when composted with 10% biochar. Small (3%) proportions of biochar in the mixture of poultry manure (78%) and barley straw (22%) had no significant effect on the emission of CH₄. Evidently, some larger proportion is needed to have a significant effect.

Conclusions and the future

Biochar is rich in carbon (C) and long-lasting in soil. When produced from organic wastes or other feedstock and added to soil, it represents a site of sequestered C. To have a net benefit requires that production of CO₂ is minimized, it must be produced from renewable feedstock grown purposefully for it or from waste materials close to the place where it is made and used. The benefit can be quantified by a Life Cycle Assessment.

The chemical composition, pH, liming potential, pore volume, bulk density and specific surface area, which determine the main characteristics of biochars, are highly variable. They depend on the feedstock, the conditions of pyrolysis, in particular the temperature during which it is made, and treatment thereafter. These should be standardized, and production planned to achieve the desired qualities for particular applications. Addition of biochar to soil increases the soil's porosity (and decreases its bulk density) and hydraulic conductivity.

Biochar contains plant nutrients and carbonates in its ash. It is therefore valuable as a fertilizer and a liming agent, especially in light textured, acid soil poor in organic matter. Long-term field studies are still needed to evaluate the durability of these effects and resolve conflicting evidence from existing trials.

Most field experiments on biochar have shown that its addition to soil has resulted in increased crop yields, especially when applied with mineral fertilizers (compared with fertilizer only application). The increases depend on the properties of both the soil and the biochar and its rate of application, and weather.

Some biochars contain large concentrations of toxic elements and organic compounds, which might enter the food chain and could affect soil biota. Both the feedstock and the pyrolysis process must be properly controlled so that concentrations of toxic elements, in particular heavy metals, and organic toxins are kept within specified bounds for intended applications of the biochar.

Biochar can bind contaminants, particularly toxic organic contaminants from industrial and waste water. This makes them valuable for restoring polluted lands. Trials at the field scale are still needed to confirm the results of many laboratory and glasshouse studies.

As biochar-based fertilizers are being introduced in some countries, such guidelines need to be developed to ensure that the fertilizers are as effective as claimed for increasing crop yields and cutting emissions of greenhouse gases.

References

- Baltrėnaitė-Gedienė E, Marčiulaitienė E, Pranskevičius M, Titova J, Bhatnagar A, and Abu-Danso E (2020) Physicochemical properties of pyrogenic carbonaceous product, biochar, syngenetically modified for its use in adsorption systems. *Journal of Environmental Engineering* 146: 4020078.
- Blanco-Canqui H (2017) Biochar and soil physical properties. *Soil Science Society of America Journal* 81: 687–711.
- Borchard N, Schirrmann M, Cayuela ML, Kammann C, Wrage-Mönning N, Estavillo JM, Fuertes-Mendizábal T, Sigua G, Spokas K, Ippolito JA, and Novak J (2019) Biochar, soil and land- use interactions that reduce nitrate leaching and N₂O emissions: A meta- analysis. *Science of the Total Environment* 651: 2354–2364.
- Budai A, Zimmerman AR, Cowie AL, Webber JBW, Singh BP, Glaser B, Masiello CA, Andersson D, Shields F, Lehmann J, Camps AM, Williams M, Sohi S, and Joseph S (2013) Biochar Carbon Stability Test Method: An Assessment of Methods to Determine Biochar Carbon Stability. https://www.biochar-international.org/wp-content/uploads/2018/04/BI_Report_Biochar_Stability_Test_Method_Final.pdf.
- Campos P, Miller AZ, Knicker H, Costa-Pereira MF, Merino A, and De la Rosa JM (2020) Chemical, physical and morphological properties of biochars produced from agricultural residues: Implications for their use as soil amendment. *Waste Management* 105: 256–267.
- Chia CH, Downie A, and Munroe P (2015) Characteristics of biochar: Physical and structural properties. In: Lehmann J and Joseph S (eds.) *Biochar for Environmental Management: Science, Technology and Implementation*, pp. 89–109. London: Earthscan.
- Dai Y, Zheng H, Jiang Z, and Xing B (2020) Combined effects of biochar properties and soil conditions on plant growth: A meta-analysis. *Science of the Total Environment* 713: 136635.
- Domingues RR, Trugilho PF, Silva CA, de Melo I, Melo LCA, Magriotis ZM, and Sanchez-Monedero MA (2017) Properties of biochar derived from wood and high-nutrient biomasses with the aim of agronomic and environmental benefits. *PLoS One* 12(5): e0176884.
- Esfandbod M, Phillips IR, Miller B, Rashti MR, Lan ZM, Srivastava P, Singh B, and Chen CR (2017) Aged acidic biochar increases nitrogen retention and decreases ammonia volatilization in alkaline bauxite residue sand. *Ecological Engineering* 98: 157–165.
- Glaser B (2007) Prehistorically modified soils of central Amazonia: A model for sustainable agriculture in the twenty-first century. *Philosophical Transactions of the Royal Society, B: Biological Sciences* 362: 187–196.
- Herath HMSK, Camps-Arbestain M, and Hedley M (2013) Effect of biochar on soil physical properties in two contrasting soils: An alfisol and an andisol. *Geoderma* 209–210: 188–197.
- IPCC (2006) 2006 Guidelines for National Greenhouse Gas Inventories. Vol. 5, Chapter 4: Biological Treatment of Solid Waste. https://www.ipcc-nggip.iges.or.jp/public/2006gl/pdf/5_Volume5/V5_4_Ch4_Bio_Treat.pdf.
- IPCC (2007) *Climate Change: The Physical Science Basis. Working Group I Contribution to the Fourth Assessment Report of the IPCC*. Cambridge, UK: Cambridge Univ Press.
- IPCC (2019) *Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories*. Switzerland: IPCC. Chapter 2 Generic methodologies applicable to multiple land use categories. Appendix4: Method for Estimating the Change in Mineral Soil Organic Carbon Stocks from Biochar Amendments: Basis for Future Methodological Development https://www.ipcc-nggip.iges.or.jp/public/2019rf/pdf/4_Volume4/19R_V4_Ch02_Ap4_Biochar.pdf.
- Ippolito JA, Cui L, Kammann C, Wrage-Mönning N, Estavillo JM, Fuertes-Mendizábal, Cayuela ML, Sigua G, Novak J, Spokas J, and Borchard N (2020) Feedstock choice, pyrolysis temperature and type influence biochar characteristics: A comprehensive meta-data analysis review. *Biochar* 2: 421–438.
- ISO (2005) *International Organisation for Standardisation (2005) Soil Quality: Guidance on the Determination of Background Values. ISO 19258: 2005*. International Organisation for Standardisation.
- Jeffery S, Abalos D, Prodana M, Bastos AC, van Groenigen JW, Hungate BA, and Verheijen F (2017) Biochar boosts tropical but not temperate crop yields. *Environmental Research Letters* 12: 053001. <https://doi.org/10.1088/1748-9326/aa67bd>.
- Lang T, Jensen AD, and Jensen PA (2005) Retention of organic elements during solid fuel pyrolysis with emphasis on the peculiar behavior of nitrogen. *Energy & Fuels* 19: 1631–1643.
- Lehmann J, Abiven S, Kleber M, Pan G, Singh BP, Sohi SP, and Zimmerman AR (2015) Persistence of biochar in soil. In: Lehmann J and Joseph S (eds.) *Biochar for Environmental Management: Science, Technology and Implementation*, pp. 235–299. London: Earthscan.

- Lehmann J, Cowie A, Masiello CA, Kammann C, Woolf D, Amonette JE, Cayuela ML, Camps-Arbestain M, and Whitman T (2021) Biochar in climate change mitigation. *Nature Geoscience* 14: 883–892.
- Melo LCA, Lehmann J, Carneiro JSS, and Camps-Arbestain M (2022) Biochar-based fertilizers and their effect on crop yield: A meta-analysis. *Plant and Soil* 36(1): 2–18.
- Meyer S, Genesio L, Vogel I, Schmidt HP, Soja G, Someus E, Shackley S, Verheijen FGA, and Glaser B (2017) Biochar standardization and legislation harmonization. *Journal of Environmental Engineering and Landscape Management* 25: 175–191.
- Mia S, Dijkstra FA, and Singh B (2017) Aging induced changes in biochar's functionality and adsorption behavior for phosphate and ammonium. *Environmental Science and Technology* 51(15): 8359–8367.
- Namgay T, Singh B, and Singh BP (2010) Influence of biochar application to soil on the availability of As, Cd, Cu, Pb, and Zn to maize (*Zea mays* L.). *Australian Journal of Soil Research* 48: 638–647.
- Nzediegwu C, Naeth MA, and Chang SX (2021) Carbonization temperature and feedstock type interactively affect chemical, fuel, and surface properties of hydrochars. *Bioresource Technology* 330: 124976.
- Omondi MO, Xia X, Nahayo A, Liu XY, Korai PK, and Pan GX (2016) Quantification of biochar effects on soil hydrological properties using meta-analysis of literature data. *Geoderma* 274: 28–34.
- Razzaghi F, Obour PB, and Arthur E (2020) Does biochar improve soil water retention? A systematic review and meta-analysis. *Geoderma* 361: 114055.
- Reisser M, Purves RS, Schmidt MWI, and Abiven S (2016) Pyrogenic carbon in soils: A literature-based inventory and a global estimation of its content in soil organic carbon and stocks. *Frontiers in Earth Science* 4: 1–14.
- Singh B, Dolk MM, Shen Q, and Camps-Arbestain M (2017) Biochar pH electrical conductivity and liming potential. In: Singh B, Camps-Arbestain M, and Lehmann J (eds.) *Biochar: A Guide to Analytical Methods*, pp. 23–38. Boca Raton: CRC Press/Taylor Francis Group.
- Singh B, Singh BP, and Cowie AL (2010) Characterisation and evaluation of biochars for their application as a soil amendment. *Australian Journal of Soil Research* 48: 516–525.
- Souza CD, Bomfim MR, de Almeida MD, Alves LD, de Santana VN, Amorim ICD, and Santos JAG (2021) Induced changes of pyrolysis temperature on the physicochemical traits of sewage sludge and on the potential ecological risks. *Scientific Reports* 11: 974.
- Wang T, Camps-Arbestain M, Hedley MJ, and Bishop PA (2012) Predicting phosphorus bioavailability from high-ash biochars. *Plant and Soil* 357: 173–187.
- Woolf D, Lehmann J, Ogle S, Kishimoto-Mo AW, McConkey B, and Baldock J (2021) Greenhouse gas inventory model for biochar Additions to soil. *Environmental Science & Technology* 55(21): 14795–14805.
- Ye L, Camps-Arbestain M, Shen Q, Lehmann J, Singh B, and Sabir M (2020) Biochar effects on crop yields with and without fertilizer: A meta-analysis of field studies using separate controls. *Soil Use and Management* 36(1): 2–18.
- Yuan HY, Ding LJ, Zama EF, Liu PP, Hozzein WN, and Zhu YG (2018) Biochar modulates methanogenesis through electron syntrophy of microorganisms with ethanol as a substrate. *Environmental Science & Technology* 52: 12198–12207.
- Zama EF, Zhu YG, Reid BJ, and Sun GX (2017) The role of biochar properties in influencing the sorption and desorption of Pb(II), Cd(II) and As(III) in aqueous solution. *Journal of Cleaner Production* 148: 127–136.
- Zhang HZ, Chen CR, Gray EM, and Boyd SE (2017) Effect of feedstock and pyrolysis temperature on properties of biochar governing end use efficacy. *Biomass & Bioenergy* 105: 136–146.
- Zhang L, Jing, Xiang Y, Zhang R, and Lu H (2018) Responses of soil microbial community structure changes and activities to biochar addition: A meta-analysis. *Science of the Total Environment* 643: 926–935.
- Zhang M, Shan SD, Chen YG, Wang F, Yang DY, Ren JK, Lu HY, Ping LF, and Chai YJ (2019) Biochar reduces cadmium accumulation in rice grains in a tungsten mining area-field experiment: Effects of biochar type and dosage, rice variety, and pollution level. *Environmental Geochemistry and Health* 41: 43–52.
- Zhao L, Nan H, Kan Y, Xu X, Qiu H, and Cao X (2019) Infiltration behavior of heavy metals in runoff through soil amended with biochar as bulking agent. *Environmental Pollution* 254: 113114.
- Zhao SX, Ta N, and Wang S D. (2017) Effect of temperature on the structural and physicochemical properties of biochar with apple tree branches as feedstock material. *Energies* 10(9): 1293.
- Zheng H, Wang Z, Deng X, Zhao J, Luo Y, Novak J, Herbert S, and Xing B (2013) Characteristics and nutrient values of biochars produced from giant reed at different temperatures. *Bioresource Technology* 130: 463–471.

Dissolved organic carbon

Jason James^a and Rob Harrison^b, ^aExponent Inc., Environmental and Earth Science Practice, Bellevue, WA, United States; ^bSchool of Environmental and Forest Sciences, University of Washington, Seattle, WA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	421
Definition	422
Origins of DOC	422
Factors associated with DOC inputs and production	422
Climate	422
Soil type	423
pH	424
Quantity and quality of organic material	424
Soil use and management	424
Composition and transport of dissolved organic matter	426
Composition	426
Transport	426
Interactions with minerals and microbes	427
Effect of DOM composition on organo-mineral interactions	427
Adsorption	428
Microbial processing, degradation, and communication	428
Importance of DOC in the global carbon cycle	430
Conclusion	430
References	430

Abstract

Soil dissolved organic carbon (DOC) is a small but important fraction of total soil carbon that is involved in an enormous variety of important interactions with plants, microbes, and minerals. DOC is composed of thousands of individual compounds with differing sizes, reactivities, and functions, and is fundamental to soil forming processes. The reactivity of DOC with mineral surfaces changes depending on each molecule's functional groups, shape, and the composition of the soil solution. Adsorption is a dominant fate of DOC, but adsorption reactions are reversible and aged DOC may be released and exported from soils to groundwater and fluvial networks.

Introduction

Soil dissolved organic carbon (DOC) is generally a small fraction of the total soil C pool at any given time but is of considerable relative importance in shaping soil ecosystem function and properties. The impact of DOC lies not in its total amount, but that it is reactive, dynamic and mobile, and over time helps shape the distribution and nature of total soil C. Most current global studies of C focus on atmospheric and plant C pools and dynamics. However, total soil C (to 1 m depth) exceeds the total amount of C in plants and the atmosphere globally. When soil depth is extended further, the amount of soil C generally increases, forming a huge global C pool little studied, or even quantified, and much less well understood, especially in terms of its properties and dynamics.

Soil DOC has high mobility and reactivity and can act as a carbon and energy source for soil heterotrophic organisms deeper in the soil profile. Though soil water, and thus DOC, can and does move in all directions in soil, movement downward into the soil profile is the prevalent DOC transportation mechanism in humid regions, and one of the primary mechanisms resulting in the formation of unique soil profiles. Soil DOC both affects and is affected by the soil solid phase as it moves through it, adding to chemical and biological weathering of soil parent materials, acting as a carbon and energy source for soil microorganisms, and stripping or depositing byproducts of soil weathering as it moves through a soil profile. Soil leaching can further move C and soluble products of soil weathering completely out of the soil profile, potentially to groundwater, or into aquatic ecosystems such as streams and lakes.

DOC is the most dynamic and mobile pool of total soil C. Much of soil C has formed and accumulated over relatively long periods of time. Under conditions of climate change, expected impacts include shifting human, plant and animal populations and associated land-use changes, invasive species spread, extinctions, and altered temperature and precipitation. Consequently, assuming soil C and especially soil DOC will remain constant in the future is risky, particularly in assessing feedbacks between soil C and the global climate. Changes in the rates of formation of DOC in soil and its movement into the soil profile could greatly impact total ecosystem C and global soil C.

Definition

Soil DOC is typically defined empirically, with the specific definition in each instance depending on methods of collection and determination. Soil DOC is by nature highly variable chemically, biologically, and physically, which can result in some ambiguity in definition. For example, to collect soil DOC, a sample of soil solution is taken, either by extraction or centrifuging, and soil solution C which will pass through a filter with a pore size typically 0.2–0.7 μm is then defined as “soil DOC”. As such, soil properties can significantly affect what will remain in solution. The most important soil properties determining soil DOC — usually soil surface charges (cation exchange capacity, CEC, and anion exchange capacity, AEC), pore size, soil solution composition, and other properties, particularly those that enhance hydrophobic adsorption (Jardine et al., 1989; Kothawala et al., 2008) — bind C molecules to soil surfaces that might otherwise remain in solution. Aging of soil extracts can also affect operational soil DOC, since C molecules are typically chemically active, and can undergo precipitation or fractionation that can change a molecule’s ability to pass through a filter.

The method of DOC collection in the field or extraction in the laboratory can substantially affect the quality and quantity of measured values, and consequently care is necessary in interpreting the literature. For example, DOC measured from soil lysimeters in the field is typically less concentrated than DOC extracted by centrifuging soils in the laboratory, which in turn are less concentrated than soil extracted with mild salt solutions or hot water under vacuum. Due to these differences, laboratory extracted DOC is frequently identified as water extracted organic carbon (WEOC) to avoid confusion with field-collected DOC. When soil solution is collected by extraction, usually with tension soil lysimeters, it will already be filtered depending on lysimeter pore size, and any carbon in the solution may be defined as dissolved. In common use, the terms soil DOC and soil dissolved organic matter (DOM) are sometimes used interchangeably, with soil DOC typically reported as the mass of C only and DOM as the total mass of dissolved material. The fate and reactivity of DOC depends not just on the structure and type of carbon-to-carbon bonds, but on the kind, quantity, and bonding orientation of other elements within the molecular structure, particularly hydrogen, oxygen, phosphorous, and nitrogen. Thus, discussions in this chapter about mineral-organic interactions, as well as fate and transport will focus on soil DOM rather than just DOC.

Origins of DOC

Dissolved organic carbon enters soil systems through rainfall, plant residues, root exudates, desorption or degradation of soil organic matter (SOM), and microbial biomass (Gmach et al., 2020). Plants contribute to soil DOC through the breakdown of plant debris by soil organisms both large and small as well as deliberate release of soluble organic molecules by plant roots, including facilitation of mutualistic relationships with microbes. Once this plant derived material enters solution, it becomes food for a complex ecosystem of fungi, bacteria, and archaea (Sylvia et al., 2005). These microbes consume and transform plant-derived DOC and release a variety of organic substances including communication molecules and extracellular enzymes (while alive) and a host of compounds after cell death and breakdown of cellular structures. In mineral soils, where plant residues mostly accumulate at the surface, DOC most closely resembles plant-derived molecules near the surface and gradually transforms with depth to resemble microbially-derived DOC (Kaiser and Kalbitz, 2012). A portion of DOC observed in soil solution is derived from old, solid-phase SOM which is released through desorption or partial microbial degradation of surface-bound SOM (Leinemann et al., 2018).

The relative importance of each input of DOC varies between soils and ecosystems, as well as the depth within the soil profile. For example, surface litterfall in a coniferous forest resulted in considerable DOC movement from surface litter into mineral soil, accounting for 22% of annual carbon inputs below 40 cm (Sanderman et al., 2008). In a contrasting grassland, surface litter accounted for only 2% of carbon inputs below 20 cm, which instead was dominated by inputs from root exudates and turnover. In both soils, field collected DOC was significantly aged (based on radiocarbon dating), indicating that DOC in solution was not simply the fraction of surface-derived DOC remaining in solution but that there must be exchange with older SOM with residence times of decades to centuries (Sanderman et al., 2008). Adsorption was the dominant fate of DOC once it entered the soil, with mean residence times of 90–150 years.

Rainfall provides a small but important input of DOC to soils. Approximately 80% of carbon in rainfall is organic, and rainfall deposits approximately 0.5 Pg C y^{-1} , of which 70% is over land (Willey et al., 2000). Beyond carbon input, the duration, quantity, and intensity of rainfall affects DOC movement and flux in soil by affecting the hydrologic dynamics of soil solution.

Factors associated with DOC inputs and production

Climate

A compilation of DOC from 365 soils in different ecosystems around the globe found that temperate climates — generally found between 35 and 60°N latitude — had, on average, the largest DOC concentrations (Fig. 1; Camino-Serrano et al., 2014). Despite the enormous productivity of tropical ecosystems, tropical soils (<35°N latitude) had the smallest average DOC concentration. Boreal soils had similar DOC concentrations in surface soil compared with temperate soils, but smaller subsurface concentrations (Camino-Serrano et al., 2014). The differences between these broad climate classes reflect offsetting, climate-dependent processes: DOC production and DOC decomposition (Kalbitz et al., 2000). The rate of DOC production is weakly related to temperature, but

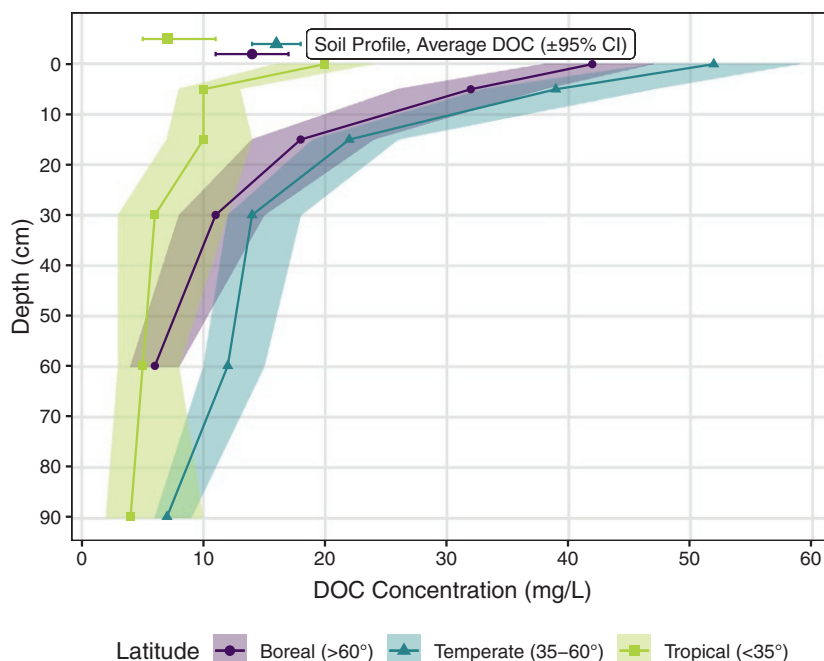


Fig. 1 Distribution of DOC in mineral soils by latitude and climate classification. Whole soil average with 95% confidence intervals are shown at the top for each group. Redrawn with data from Camino-Serrano et al. (2014).

also depends upon the nutrition available to soil microbes as well as the composition of decomposing plant matter, such as lignin content (Moore et al., 2008). In contrast, the ratio of CO_2 :DOC production increases steeply with temperature (Moore et al., 2008), which suggests that high DOC production in the tropics is offset by high decomposition rates. In boreal soils, low temperatures reduce both DOC production and decomposition. Temperate soils thus maintain larger DOC concentrations due to the balance of greater rates of DOC production than boreal and arctic soils but smaller rates of decomposition than tropical soils (Camino-Serrano et al., 2014).

DOC concentrations exhibit seasonal fluctuations. This effect is more pronounced in surface soil and muted in the subsurface (B and C horizons) due to greater microbial activity in surface soil, which is partly controlled by temperature (Kalbitz et al., 2000). Due to elevated microbial activity in the growing season, larger concentrations of hydrophilic compounds derived from microbes and carbohydrates are released into soil solution (Guggenberger and Zech, 1994). Soil moisture is also a significant factor in seasonal fluctuations in soil DOC. After dry periods, studies have consistently identified a flush of DOC upon rewetting, likely from mobilization of microbial products that accumulated on soil surfaces or in mineral-associated pore water (Kalbitz et al., 2000).

Soil type

Organic peatland soils typically have the greatest concentrations of DOC because they are composed of a large reservoir of incompletely decomposed plant material (Fig. 2, Camino-Serrano et al., 2014). Unlike mineral soils, DOC concentrations increased with depth in non-forest organic soils. Despite the great quantity of plant material and DOC available for decomposition in organic soils, complete water saturation and low rates of oxygen recharge greatly slow microbial activity. Decomposition is an oxidation-reduction (redox) process where organic matter is oxidized (loses electrons) and oxygen atoms are reduced (accept electrons). As redox conditions decline with depth in organic soils, alternative electron acceptors are used by different varieties of microbes. Nitrate, manganese oxide, iron, sulfate, and carbon dioxide are used as alternative electron acceptors in decreasing order of energetic yield and rates of decomposition (Sylvia et al., 2005). The residence time of water also increases with depth in organic soils, which allows DOC to build up even though DOC production rates are slow due to highly reducing and anaerobic conditions (Camino-Serrano et al., 2014).

The effect on DOC concentrations of different mineral soil types depends on their texture and mineralogy, particularly the capacity to adsorb DOC on mineral surfaces. Camino-Serrano et al. (2014) observed that volcanic soils had the smallest DOC concentrations, on average, and attributed the difference to widespread sorption of DOC to minerals (iron and aluminum oxides) derived from volcanic ash. In comparison, weakly developed soils with coarse or sandy textures such as Podzols, Arenosols, and Regosols (FAO soil taxonomy) (Spodosols and Entisols in USDA taxonomy) had the greatest concentrations of DOC among mineral soils (Camino-Serrano et al., 2014). Sand particles have small surface area relative to their volume and little capacity to adsorb organic molecules. The effects of differing mineralogy on DOC dynamics will be detailed below in the discussion of adsorption.

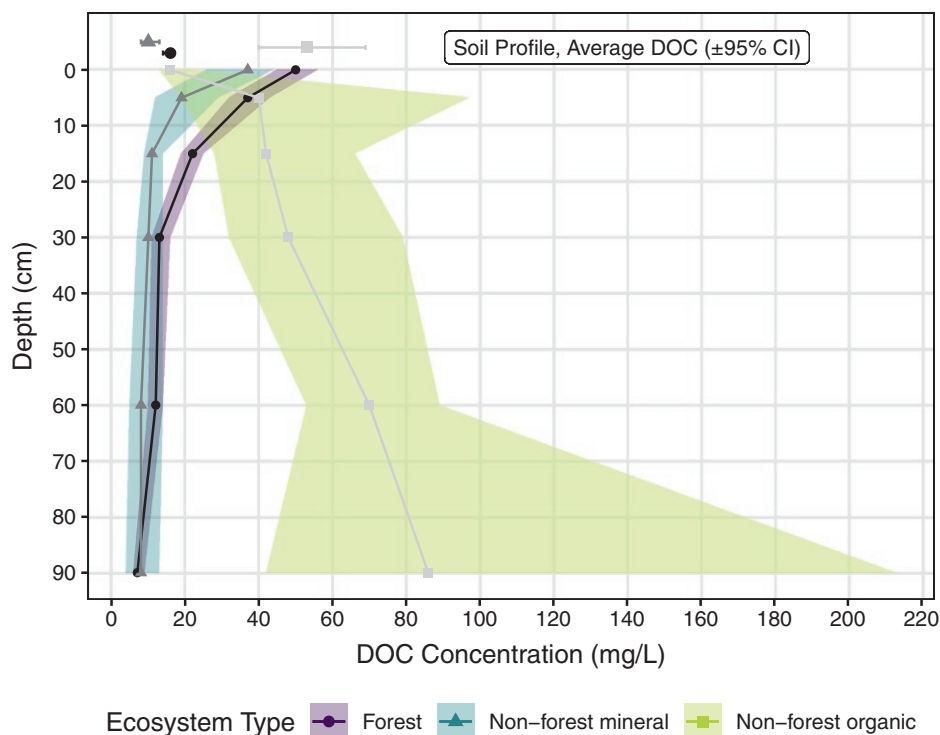


Fig. 2 Distribution of DOC by organic or mineral soil type. Whole soil average with 95% confidence intervals are shown at the top for each group. Redrawn with data from Camino-Serrano et al. (2014).

pH

pH is a dominant control over the electrical charge on both organic molecules and soil minerals, including clays and iron and aluminum (hydr)oxides. DOC release from organic soils is positively correlated with pH, likely due to protonation of organic functional groups at low pH which reduces solubility (Tipping and Woof, 1991). In mineral soils, by contrast, soil solution DOC is enhanced in more acidic soils, especially in subsurface layers below 10 cm (Fig. 3; Camino-Serrano et al., 2014). Low pH in mineral soils encourages saturation of exchange sites (e.g. with protons and exchangeable aluminum) and dissolution of organic complexes containing polyvalent cations (Kalbitz et al., 2000), both of which reduce adsorption.

Quantity and quality of organic material

The quantity and quality DOC differs between coniferous and broadleaf forests (Fig. 4), although the reasons may not be directly due to the forest type. Broadleaved trees tend to produce less acidic litter, and changes in soil acidity have a profound effect on the composition and activity of soil microbes. Consequently, conifer forests may have slightly greater soil DOC because of altered soil metabolism. However, forest composition is not independent of the soil and climate conditions in which the forest grows. Conifers, especially pines, are more frequently located on sandy soils and in cold climates. In sandy soils, the flow regime and fast transport through macropores dominates DOC dynamics, possibly reducing adsorption and microbial processing (Kalbitz et al., 2000). Cold temperatures likewise inhibit microbial activity, resulting in greater concentrations of DOC remaining in solution. All three mechanisms – inherent differences in microbial processing of litter from different trees, hydrological controls on DOC transport, and temperature dependence of both microbial processes and tree species growth ranges – co-occur and are difficult to disentangle (Camino-Serrano et al., 2014).

Soil use and management

The enormous extent and intensity of human use of land and soil is readily visible on any airline flight or through satellite imagery. The UN FAO estimates that currently 38% of global land area is used for agriculture, about a third of which is cropland and two thirds of which is permanent meadow and pasture (UN FAO, 2021). In many parts of the world, forestland is also increasingly utilized. Much of the rest is either urban or not considered arable. As world population has increased since 1960, land use for agriculture has not increased greatly in actual area, but intensity (humans fed per hectare) has nearly doubled since 1960, primarily through genetic improvement of plants, but also increased use of fertilizers and pesticides (UN FAO, 2021). This increased agricultural intensity strongly affects the amount and molecular composition of DOM, which constitutes the main vector of carbon

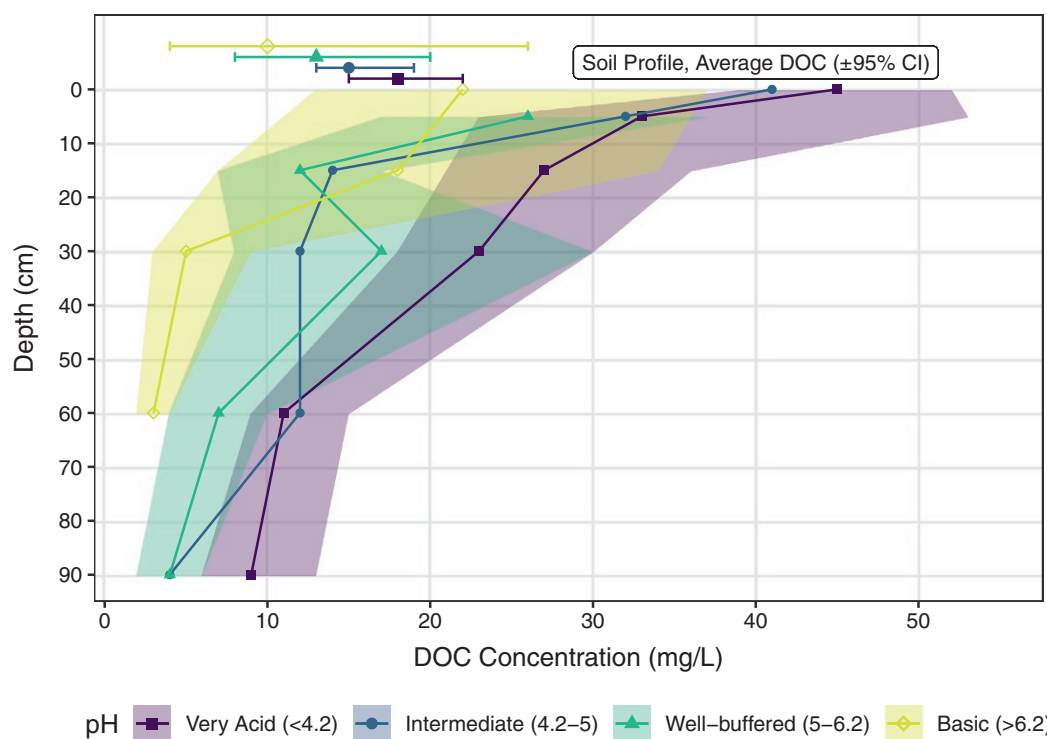


Fig. 3 Distribution of soil DOC by pH classification, excluding Histosols (organic soils). Whole soil average with 95% confidence intervals are shown at the top for each group. Redrawn with data from Camino-Serrano et al. (2014).

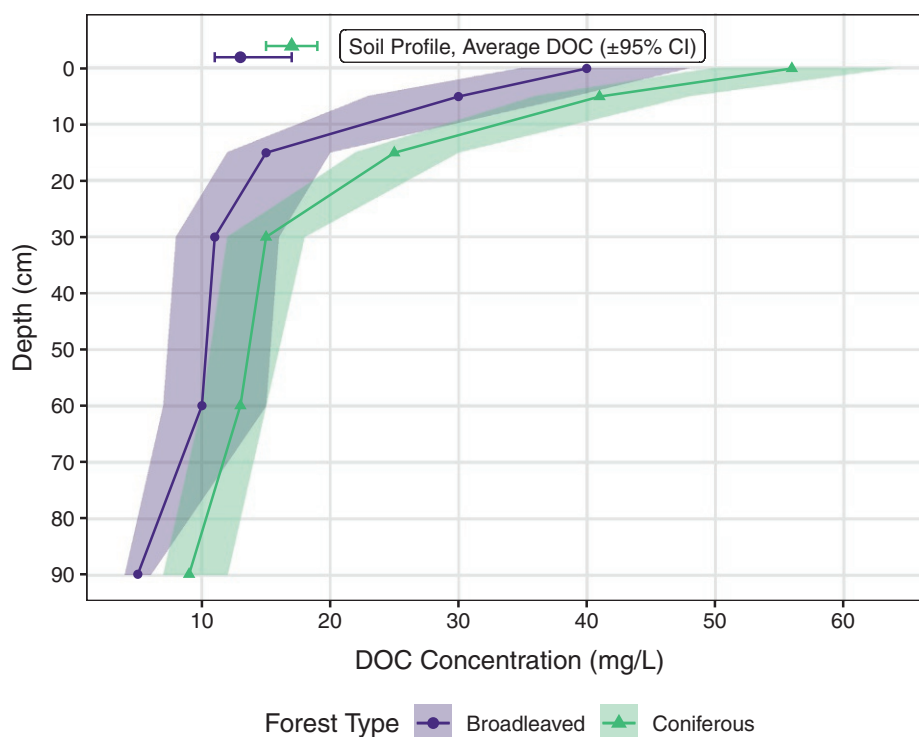


Fig. 4 Distribution of soil DOC by forest type (excluding organic soils). Whole soil average with 95% confidence intervals are shown at the top for each group. Redrawn with data from Camino-Serrano et al. (2014).

transport from soils to freshwater systems and to the ocean and is involved in a large variety of biogeochemical processes. For example, Graeber et al. (2015) found that DOM exported from agricultural catchments had a more microbial and less plant-derived composition, and was accompanied by greater dissolved organic nitrogen (DON), which together could drive greater CO₂ outgassing in freshwater systems.

There is some additional arable land available, but the trend of increased production per hectare needs to continue so that much of that area can be preserved, and hopefully also to allow some land currently being utilized to be allocated for conservation and preservation. Understanding the implications for soil C in land areas under more intensive agriculture is essential to any future management. If greater intensity of management causes shifts from solid-phase to soluble soil C, the implications could be enormous both for future productivity and global C (Paustian et al., 2016).

Composition and transport of dissolved organic matter

Composition

DOM consists of hundreds to many thousands of individual compounds, each of which have different sizes, reactivities, functional configurations, and solubilities. Due to the complexity of the mixture in solution, a wide variety of methods have been developed to probe DOM composition. For example, fluorescence spectroscopy generates a signature of the subset of organic compounds that can absorb a photon of light and re-emit a photon at a different wavelength. Other techniques, such as mass spectrometry, differentiate distinct compounds by weight, with some advanced methods that can tentatively identify each of thousands of molecular formulas contained within a DOM mixture (for example, Fourier-transform ion cyclotron resonance mass spectrometry, or FTICR-MS). Yet, knowing many of the molecular formulas for compounds does not reveal critical features of molecules that would be necessary to completely understand their reactivity, such as the structure and conformation of the molecule. Due to the diversity of methods old and new, there are many classification schemes for DOM compounds based on perceived, inferred, or measured properties, including recalcitrance and lability (how quickly each compound will biodegrade), bioavailability, aromaticity, size, and even inferred origin (Xenopoulos et al., 2021).

Many organic molecules have been extensively studied, particularly toxic and potentially hazardous chemicals such as benzene, naphthalene, polycyclic aromatic hydrocarbons (PAHs), or other organic pollutants monitored by the U.S. EPA, European Chemicals Agency (ECHA) and other regulatory agencies. Understanding the chemistry of these molecules is critical for designing engineering systems and remediation plans to safeguard the public following historical or contemporary industrial releases. In contrast, much less is known about the many thousands of natural organic compounds found in soil DOM, including the biodegradation products of hydrocarbons and other contaminants.

However, there are unifying features of many organic molecules in solution that permit a general understanding of the controls and interactions that DOM undergoes with soil minerals. For example, natural oils and waxes contain long, linear or branching chains of carbon atoms that do not have an electrical charge (called aliphatic). Regardless of the length of the chain or what other organic structure is attached, that portion of the organic molecule is not attracted to polar water molecules and will preferentially separate itself from solution. Other organic molecules, such as those containing specific functional groups like carboxylic acids, aldehydes, or alcohols, can have negative charges depending upon soil solution pH, which will cause them to interact with ions in solution and charged mineral surfaces. Likewise, dissolved organic molecules containing amine groups (e.g., amino acids) can maintain positive charges.

Transport

The movement of DOM depends upon both the soil type and mineralogy within the soil profile, as well as the hydrologic connections, location, slope, and aspect of soils in the landscape. In most soils, rainwater percolates through the surface and is transported downward by gravity until reaching the water table. In some soils (both mineral and organic) the topography is flat and water table is near or at the soil surface, in which case the rate of movement of water within the profile can be slow. The transport of DOM in soil water is also affected by capillary forces (the attraction of water molecules to soil surfaces), which can keep small pockets of water (and DOM) in place under draining conditions. Above a water table, capillary action in soil draws water upwards, together with the DOM and dissolved inorganic carbon (DIC, such as carbonic acid and bicarbonate) contained in it. In areas where evapotranspiration exceeds precipitation and the pH is alkaline, inorganic carbon rather than DOC may be the dominant form of carbon in solution, which allows carbonate minerals to precipitate and accumulate in the soil profile.

The movement of DOM within the soil profile does not simply follow the flow of water. DOM released from plant materials or roots at or near the surface interact with minerals and microbes, resulting in slowed movement, partial decomposition, and transformation (Fig. 5). Near the soil surface, particularly the organic rich O and A horizons, DOM has modern or near-modern radiocarbon signatures, and thus is likely the product of recent photosynthetic products (Sanderman and Amundson, 2008; Froberg et al., 2009). In contrast, studies around the globe across diverse ecosystems have observed that subsoil DOM is depleted in radiocarbon. Thus, subsoil DOM is composed of older compounds or compounds containing pieces of aged carbon, and little DOC derived from the surface reaches mineral subsoil (Froberg et al., 2009; Hagedorn et al., 2015). In fact, labeled DOC leached through

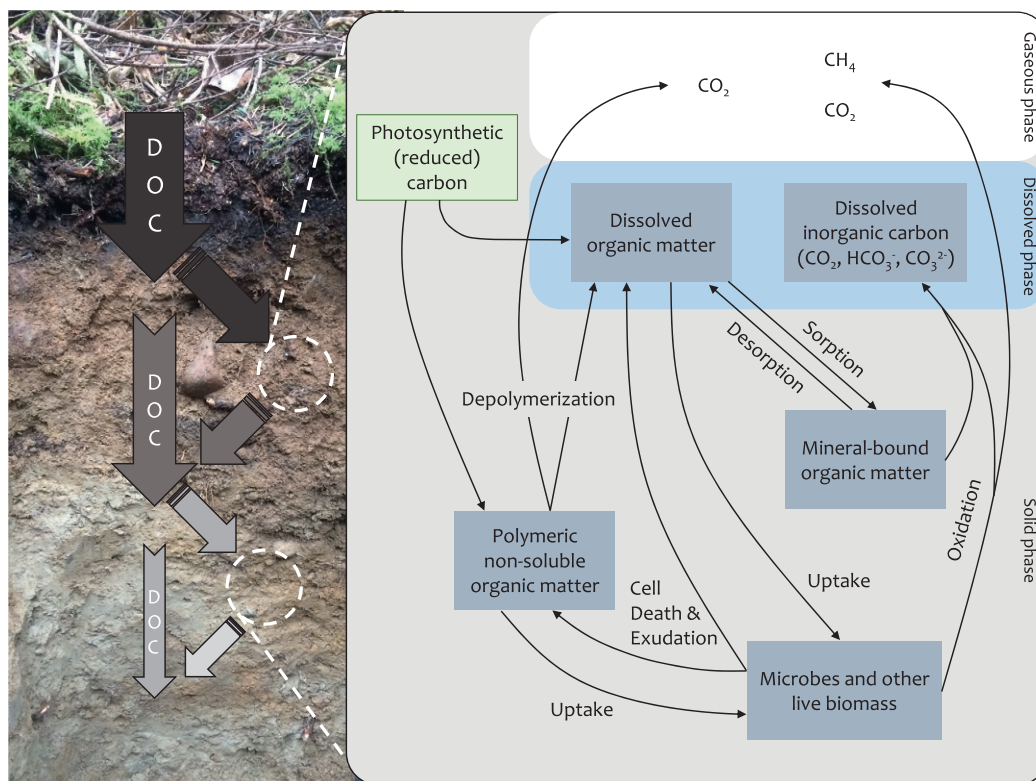


Fig. 5 Cycling of dissolved organic matter in soils through a cascade of interactions. The processes depicted take place in a continuum throughout the soil profile. Photosynthetic carbon comprises organic products of plant photosynthesis released to the soil through plant litter or roots. Polymeric non-soluble organic matter is composed of larger particles or molecules of organic matter that are not in soil solution. Under most circumstances, CO_2 is the primary respiration product of microbial consumption and oxidation of organic matter; CH_4 may be created and emitted only under highly water-logged and reducing conditions. CH_4 generated in soil may not be released into the atmosphere as microbes that produce and consume CH_4 frequently co-occur (Kaiser and Kalbitz, 2012).

soil columns revealed that almost all was retained in soil due to mineral interactions, with the label concentrated along preferential flow paths in deeper soil (Hagedorn et al., 2015). DOM moving through the soil profile is subject to a stepwise cascade of reactions and interactions, including adsorption to minerals, microbial transformation and degradation, and re-release of DOM into soil solution during transport down the soil profile (Kaiser and Kalbitz, 2012; Leinemann et al., 2018).

Interactions with minerals and microbes

Effect of DOM composition on organo-mineral interactions

Advances in analytical techniques over the last two decades have allowed identification, within individual DOM samples, of tens of thousands of different organic compounds with widely varying molecular structures, sizes, and functional groups. The differences in the structure, size, and functionality of different molecules within DOM result in diverse modes of interaction with mineral surfaces in soil (von Luetzow et al., 2006). Aliphatic organic molecules, which are chains of single or double-bonded C atoms without other functional groups, such as hexane (C_6H_{14}), as well as aromatic molecules, which contain one or more 6 carbon rings, such as benzene (C_6H_6) and naphthalene (C_{10}H_8) can bind to mineral surfaces through hydrophobic exclusion: non-polar molecules will associate with one another away from water, which is inherently polar. This interaction is pH independent, whereas most other modes of organic attachment to minerals are pH dependent (Jardine et al., 1989; Kleber et al., 2007). Some aliphatic and aromatic molecules contain minor functional groups, such as hexanol ($\text{C}_6\text{H}_{13}\text{OH}$), which adds an alcohol (OH) group to hexane, and benzoic acid ($\text{C}_6\text{H}_5\text{OOH}$), which adds a carboxyl group to benzene. In addition to hydrophobic interactions, the polar ends of these molecules can attach to minerals through hydrogen bonding. If these molecules lose a hydrogen atom, they become negatively charged and will seek to satisfy this charge with positive charges on mineral surfaces or metal cations. Some organic molecules such as glucose ($\text{C}_6\text{H}_{12}\text{O}_6$) are inherently polar with no aliphatic or aromatic structures and will primarily interact with minerals through hydrogen bonding (Kleber et al., 2021). However, the additional functionalization of these kinds of molecules, as in glucosamine ($\text{C}_6\text{H}_{13}\text{NO}_5$) or gluconic acid ($\text{C}_6\text{H}_{12}\text{O}_7$), will allow the molecule to become negatively charged (for gluconic acid) or positively charged (for glucosamine) and thereby interact with minerals through electrical forces of attraction. Positive charges in DOM mostly

occur with N-containing groups due to the ability for -NH_2 groups to attract an additional proton to become -NH_3^+ . Finally, some DOM molecules, such as the amino acid phenylalanine ($\text{C}_9\text{H}_{11}\text{NO}_2$) as well as many proteins and polypeptides, contain all major functionalities and may interact with minerals in numerous ways depending upon the pH of the soil solution (Kleber et al., 2021).

The quantity and composition of soil DOM also effects the mobility and bioavailability of metals in soil. DOM molecules containing multiple negative charges can completely satisfy the positive charges of multivalent cations (e.g., Al^{3+} , Cd^{2+} , Fe^{2+} , Ni^{2+} , Zn^{2+}) through a process called chelation. Chelated metals are subsequently transported with the soluble organic molecule until such time as the molecule is degraded by microbes, immobilized by organo-mineral adsorption, or environmental conditions change, such as increasing solution pH. Organic interactions with and translocation of Fe and Al are essential reactions during pedogenesis. For example, the soil forming process called podzolization, which results in some of the most visibly striking and photogenic soils across the globe, is a direct result of metal chelation, transport, and deposition through DOM interactions (Lundström et al., 2000). During podzolization, dissolved organic acids from low-pH plant litter at the soil surface gradually strip away Fe and Al from the surface mineral soil, forming a gray, eluviated (E) horizon. As the organic acids leach downward, buffering of soil solution gradually raises the pH, which reduces the quantity of pH-dependent negative charges present on DOM. Subsequently, chelated iron is released and precipitates as ferrihydrite or goethite minerals, and chelated aluminum is released to form allophane or imogolite minerals through interactions with silicon (Ugolini and Sletten, 1991). The quantity and identity of metals mobilized by DOM in any specific soil or watershed depends upon the full suite of soil forming factors (climate, parent material, topography, organisms, and time) and the composition of DOM in solution.

Adsorption

Adsorption encapsulates a series of kinetic and electrochemical processes, namely diffusion, competition, and exchange of organic molecules at mineral surfaces. Adsorption processes are dynamic rather than static or irreversible (Kleber et al., 2021). The interface between DOM and minerals continuously evolves, based on changes in the composition of DOM in soil solution and the chemical composition of soil solution over seasons and years, as well as being mediated by biota, including both plant roots and microbes.

DOM adsorption processes are inherently diverse due to convolution in the spatial arrangement of the reactants and complexity in the composition of DOM and clay minerals (Fig. 6). At the large scale, soil minerals compartmentalize and structure the physical environment where organo-mineral interactions take place. Mineral soils contain a continuum of pore spaces with different sizes, morphology and connectivity depending on soil texture and the degree of soil development (e.g., age) (Six and Paustian, 2014). Larger particles, such as sand, provide fewer reactive surfaces than clays, but can provide macropores where water, DOM, and microbes can move relatively unhindered. On the other hand, the smallest micropores may restrict microbial access to DOM that would otherwise be consumable. Although gravity is the primary force driving water migration in macropores, water (and DOM contained in that water) near charged mineral surfaces will be increasingly bound and kept from transmission through the soil. Together these effects constrain the kinetics of mineral-DOM interactions by preventing or limiting spatial orientations that would be required for reactions to proceed unhindered and by limiting diffusion of organic molecules (Kleber et al., 2021).

Many interactions are possible between DOM and charged surfaces in soil. The observed diversity in adsorption processes is driven by the inherent variability of both the organic and inorganic reactants (Kleber et al., 2021). The strongest interactions involve ligand exchange, in which an organic adsorbate shares an oxygen with the crystal lattice of a clay molecule, most frequently an Fe- or Al-oxide. Negatively charged organic molecules in solution can interact with negatively charged clays by a cation 'bridge', wherein a cation with multiple positive charges (e.g., Ca^{+2} , Mg^{+2} , Al^{+3}) is attracted to both the clay and organic molecule (Fig. 6). Long-chain organic molecules without electrical charge may interact with charged surfaces by cation bridging as well, except involving monovalent cations (e.g., K^+) which have their positive charges mostly satisfied by the clay. Van der Waals forces attract uncharged, long-chain molecules to uncharged clay or other mineral surfaces. Hydrogen bonding, in which a polar bond between hydrogen and an electronegative atom (frequently oxygen) causes electrostatic attraction, can form bonds between an organic molecule and a clay surface or another organic molecule. Hydrogen bonding is typically stronger than van der Waals forces but weaker than cation bridging or ligand exchange reactions. Finally, long-chain, uncharged (aliphatic) molecules are frequently hydrophobic and will resist attraction to polar water molecules, which can cause them to form hydrophobic zones where these molecules partition together away from polar water or charged organic or inorganic structures.

DOM can partition out of solution to form multilayer assemblages with complex, layered structures at varying distances from a mineral surface (Kleber et al., 2007). The precise structure, distribution, and persistence of multilayer or zonal structures remains an area of inquiry. DOM molecules compete among each other and with inorganic ions for sorption sites. Even once a molecule has formed an adsorptive bond, they may desorb, although often with conformational changes to the structure of the organic molecule. Changes in soil pH, electron availability, or modifications to sorbent surfaces can induce desorption (Thompson et al., 2006). Plant roots can also exude organic molecules specifically to release bound organic matter as a means of making organic-bound nutrients more available (Keiluweit et al., 2015). Consequently, although adsorption to mineral surfaces is an important fate for DOM as it moves through soil, this fate is not irreversible (Leinemann et al., 2018).

Microbial processing, degradation, and communication

Microbial processes are critical to the release, processing, and composition of DOM. Microbes are present on and in plant litter when it drops to the soil surface, as well as intimately associated with plant roots. As soil macrofauna (such as millipedes, worms, etc.)

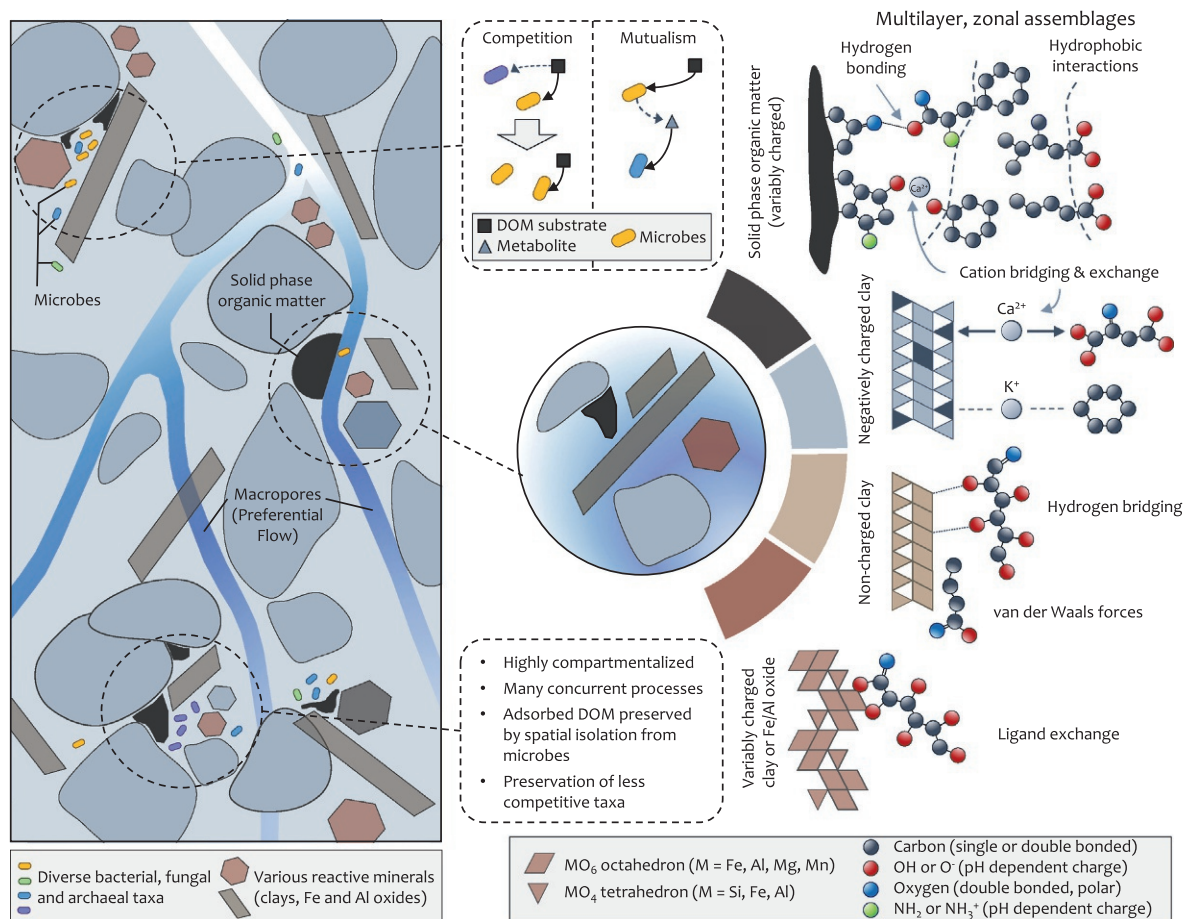


Fig. 6 Interactions between dissolved organic matter, minerals, and microbes. At large scales, the mineral matrix compartmentalizes the spaces where physiochemical processes can take place. Compartmentalization creates static and dynamic constraints on smaller-scale processes such as sorption and preferential pathways and transport pathways for DOM and microbes. Physical structures in soil (aggregates) can isolate adsorbed or polymeric organic matter from microbes, thus preventing decomposition. Less competitive microbial taxa may also be preserved in compartmentalized spaces. Mineral interfaces primarily interact with a diverse array of low-molecular-weight organic ligands that are produced through oxidative depolymerization of larger biomolecules (often plant derived). Near mineral surfaces, organic molecules can bind through one or multiple complexation mechanisms depending on the surface charge distribution, structure of the molecule, and competition with inorganic ions (such as PO_4^{2-}). Desorption of organic molecules from mineral surfaces is common, although with conformational changes to the structure of the molecule or mineral. There remain many unknowns with these processes, including the frequency and distribution of multilayer assemblages, the controls on susceptibility of adsorbed molecules to microbial degradation, and the selectivity and kinetics of desorption.

break down larger plant tissues and structures, microbial enzymes break down specific organic bonds to release smaller molecules that can be taken up by both microbes and plants. There is a complex interplay of both competitive and mutualistic relationships among microbes. In coarse sandy and gravelly soils, microbial species competing for the same compounds will lead to the dominance of the most successful species. However, in fine grained soils, species less well adapted to compete can survive in areas inaccessible to more competitive species, increasing overall soil microbial diversity (Kleber et al., 2021). Yet other microbes may be adapted to consume the degradation products of other microbes including the products of dead microbial cells. Typically, there is great functional diversity and redundancy in the soil microbial community, wherein enzymes from multiple microbes are needed to fully decompose organic matter, yet many different microbes use similar metabolic pathways and produce the same enzymes.

The ability of microbes to take up carbon from DOM or SOM and transform it into growth (rather than respiration) is called carbon use efficiency (CUE). Soils with high CUE have small respiratory losses and thus potentially can keep more C in the soil. Due to the importance of mineral-organic interactions in retaining DOM in soil, the microbial efficiency-matrix stabilization (MEMS) framework hypothesized that microbes filter DOM and particulate carbon inputs, delivering the organic products to the mineral matrix or respiring it to the atmosphere, thereby acting as the critical link between plant litter and stabilized SOM (Cotrufo et al., 2013). Because more recalcitrant plant litter is difficult to break down, it results in reduced microbial CUE, whereas litter that is prone to decomposition may enhance long-term SOM storage (Cotrufo et al., 2013).

In addition to breakdown of plant residues deposited at the soil surface, microbes associated with plant roots consume sugars and other molecules released from roots in exchange for nutrients. In forests, both arbuscular and ectomycorrhizal fungi are

intimately connected to plant roots. These symbionts fetch specific nutrients that would otherwise be unavailable to plants (e.g., in larger soil organic molecules or in minerals) in exchange for carbohydrates, amino acids, and other products of photosynthesis. The composition of microbial communities associated with roots, called the root microbiome, may be considered a secondary genome that provides host plants with microbe-derived compounds and traits. Plants can modify the rhizosphere bacterial community by modulating and adjusting the composition of root exudates, which in turn stimulate or inhibit different community members (Lareen et al., 2016).

The enriched soil around plant roots is a highly competitive environment for microbes. Consequently, microbes produce secondary metabolites (dissolved organic molecules of different types), including antibiotics that suppress pathogenic species (Lareen et al., 2016) and that can inhibit competitors occupying similar niches. There is also communication between fungal and bacterial cells and eukaryotic domains through the synthesis and diffusion of signal molecules that are subsequently perceived by other community members. Communication of this type is critical for both community coordination, such as forming biofilms, as well as regulating individual gene expression in response to changes in population density (Lareen et al., 2016). Thus, not only is soil DOM critical to soil C dynamics, but it is the primary medium of communication for the vast majority of living organisms on the planet.

Importance of DOC in the global carbon cycle

In addition to its role within soils, DOC is one of the primary conduits connecting terrestrial and aquatic ecosystems. The Intergovernmental Panel on Climate Change (IPCC) estimates that 2.5 Pg C y^{-1} is exported from the terrestrial biosphere to rivers (Canadell et al., 2021). While this estimate includes inorganic and particulate organics in addition to DOC, the total export is similar in magnitude to the present-day net uptake of atmospheric CO_2 by the oceans. Estimates of DOC export to oceans via freshwaters is approximately 0.20 to 0.8 Pg C y^{-1} and makes up 25–50% of C inputs to the oceans, depending on the river size and geographic region (Canadell et al., 2021; Williamson et al., 2021). An even greater quantity of DOC is processed within rivers without being delivered to oceans, either by burial in sediment or emitted as CO_2 to the atmosphere. The percentage of terrestrial gross primary production that is shunted to streams and rivers, where it is ultimately released as CO_2 to the atmosphere or buried in sediments, increases with river discharge, and emphasizes the importance of hydrology in the land-to-water transfer of carbon (Liu et al., 2022).

Organic peatland soils contain one third to one half of all SOC, despite occupying only about 3% of land surface area (Nichols and Peteet, 2019). They likewise contain the greatest concentrations of DOC (Fig. 2; Camino-Serrano et al., 2014) and yield the most DOC to rivers among soil types (Aitkenhead and McDowell, 2000). When wetlands or peatlands are artificially drained, the soil C and DOC, which were protected from degradation by the lack of oxygen, become much more susceptible to microbial breakdown and loss to the atmosphere. Consequently, the conservation, restoration, and protection of peatland and wetland soils is of special significance to the global carbon cycle.

Anthropogenic activities also impact mineral soils. While DOC and DIC are naturally exported from soils to some degree, human activities such as agriculture, urbanization, deforestation, and changes to drainage systems alter rates of organic matter oxidation and weathering, which can shift both the quantity and composition of riverine C exports (Butman et al., 2015; Xenopoulos et al., 2021). Anthropogenic activities have increased DOC and DIC export from soils to river by 0.2 to 1 Pg C y^{-1} , of which 80% is derived from soil (Regnier et al., 2013). The age of DOC in rivers has also been shown to increase with human population density and the proportion of human-dominated landscapes (e.g., agricultural, urban) within a watershed, with between 3% and 9% of DOC in rivers being aged carbon mobilized by anthropogenic activity (Butman et al., 2015).

Conclusion

Dissolved organic C (DOC or DOM) represents a relatively small but critically important pool of total soil C due to its highly reactive and mobile nature, relative to the bulk of soil C. Reactions of soil with DOC moving through the soil profile often defines the very nature of that soil profile, and is highly variable in location, climate, vegetation type, and season. Human use of soil, particularly for agriculture and urban development, is likely to change originally natural processes significantly, and in ways which we also do not currently well understand. Effects of human land use on DOC within soil profiles and exported from impacted catchments may serve as an early warning system of threats to total soil C, which is greater than all C in the atmosphere and all C in terrestrial plants, combined. Increasing study and understanding of DOC's importance as a dynamic determinant of total soil C (and thus global C) is critical to improving soil health and sequestering carbon from the atmosphere.

References

- Aitkenhead JA and McDowell WH (2000) Soil C:N ratio as a predictor of annual riverine DOC flux at local and global scales. *Global Biogeochemical Cycles* 14(1): 127–138. <https://doi.org/10.1029/1999GB900083>.
- Butman DE, Wilson HF, Barnes RT, Xenopoulos MA, and Raymond PA (2015) Increased mobilization of aged carbon to rivers by human disturbance. *Nature Geoscience* 8(2): 112–116. <https://doi.org/10.1038/ngeo2322>.

- Camino-Serrano M, Gielen B, Luyssaert S, Ciais P, Vicca S, Guenet B, De Vos B, Cools N, Ahrens B, Arain MA, Borken W, Clarke N, Clarkson B, Cummins T, Don A, Pannatier EG, Laudon H, Moore T, Nieminen T, Nilsson M, Peichl M, Schwendenmann L, Siemens J, and Janssens I (2014) Linking variability in soil solution dissolved organic carbon to climate, soil type, and vegetation type. *Global Biogeochemical Cycles* 1199–1214. <https://doi.org/10.1002/2013GB004726>.
- Canadell JG, Monteiro PMS, Costa MH, Cotrim da Cunha L, Cox PM, Eliseev AV, Henson S, Ishii M, Jaccard S, Koven C, Lohila A, Patra PK, Piao S, Rogelj J, Syampungani S, Zaehle S, and Zickfeld K (2021) Global carbon and other biogeochemical cycles and feedbacks. In: Masson-Delmotte V, Zhai P, Pirani A, Connors SL, Pean C, Berger S, ... Zhou B (eds.) *Climate Change 2021: The Physical Basis. Contribution of the Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pp. 673–816. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.
- Cotrufo MF, Wallenstein MD, Boot CM, Denef K, and Paul E (2013) The Microbial Efficiency-Matrix Stabilization (MEMS) framework integrates plant litter decomposition with soil organic matter stabilization: Do labile plant inputs form stable soil organic matter? *Global Change Biology* 19(4): 988–995. <https://doi.org/10.1111/gcb.12113>.
- Froberg M, Hanson PJ, Trumbore SE, Swanston CW, and Todd DE (2009) Flux of carbon from C-14-enriched leaf litter throughout a forest soil mesocosm. *Geoderma* 149(3–4): 181–188. <https://doi.org/10.1016/j.geoderma.2008.11.029>.
- Gmach MR, Cherubin MR, Kaiser K, and Cerri CEP (2020) Processes that influence dissolved organic matter in the soil: A review. *Scientia Agricola* 77(3): 1–10.
- Graeber D, Boëchat IG, Encina-Montoya F, Esse C, Gelbrecht J, Goyenola G, Gucker B, Heinz M, Kronvang B, Meerhoff M, Nimptsch J, Pusch MT, Silva RCS, Von Schiller D, and Zimmmermann E (2015) Global effects of agriculture on fluvial dissolved organic matter. *Scientific Reports* 5: 1–8. <https://doi.org/10.1038/srep16328>.
- Guggenberger G and Zech W (1994) Composition and dynamics of dissolved carbohydrates and lignin-degradation products in two coniferous forests, N.E. Bavaria, Germany. *Soil Biology and Biochemistry* 26: 19–27.
- Hagedorn F, Bruderhofer N, Ferrari A, and Niklaus PA (2015) Tracking litter-derived dissolved organic matter along a soil chronosequence using C-14 imaging: Biodegradation, physico-chemical retention or preferential flow? *Soil Biology and Biochemistry* 88: 333–343. <https://doi.org/10.1016/j.soilbio.2015.06.014>.
- Jardine PM, Weber NL, and McCarthy JF (1989) Mechanisms of dissolved organic carbon adsorption on soil. *Soil Science Society of America Journal* 58: 1378–1385.
- Kaiser K and Kalbitz K (2012) Cycling downwards - dissolved organic matter in soils. *Soil Biology and Biochemistry* 52: 29–32. <https://doi.org/10.1016/j.soilbio.2012.04.002>.
- Kalbitz K, Solinger S, Park JH, Michalzik B, and Matzner E (2000) Controls on the dynamics of dissolved organic matter in soils: A review. *Soil Science* 165(4): 277–304. <https://doi.org/10.1097/00010694-200004000-00001>.
- Keiluweit M, Bougoure JJ, Nico PS, Pett-Ridge J, Weber PK, and Kleber M (2015) Mineral protection of soil carbon counteracted by root exudates. *Nature Climate Change* 5(6): 588–595. <https://doi.org/10.1038/nclimate2580>.
- Kleber M, Sollins P, and Sutton R (2007) A conceptual model of organo-mineral interactions in soils: Self-assembly of organic molecular fragments into zonal structures on mineral surfaces. *Biogeochemistry* 85(1): 9–24. <https://doi.org/10.1007/s10533-007-9103-5>.
- Kleber M, Bourg IC, Coward EK, Hansel CM, Myneni SCB, and Nunan N (2021) Dynamic interactions at the mineral–organic matter interface. *Nature Reviews Earth & Environment* 2(6): 402–421. <https://doi.org/10.1038/s43017-021-00162-y>.
- Kothawala DN, Moore TR, and Hendershot WH (2008) Adsorption of dissolved organic carbon to mineral soils: A comparison of four isotherm approaches. *Geoderma* 148(1): 43–50. <https://doi.org/10.1016/j.geoderma.2008.09.004>.
- Lareen A, Burton F, and Schäfer P (2016) Plant root-microbe communication in shaping root microbiomes. *Plant Molecular Biology* 90(6): 575–587. <https://doi.org/10.1007/s11103-015-0417-8>.
- Leinemann T, Preusser S, Mikutta R, Kalbitz K, Cerli C, Höschen C, Mueller CW, Kandeler E, and Guggenberger G (2018) Multiple exchange processes on mineral surfaces control the transport of dissolved organic matter through soil profiles. *Soil Biology and Biochemistry*. <https://doi.org/10.1016/j.soilbio.2017.12.006>.
- Liu S, Kuhn C, Amatulli G, Aho K, Butman DE, Allen GH, Lin P, Pan M, Yamazaki D, Brinkerhoff C, Gleason C, Xia X, and Raymond PA (2022) The importance of hydrology in routing terrestrial carbon to the atmosphere via global streams and rivers. *Proceedings of the National Academy of Sciences of the United States of America* 119(11): 1–9. <https://doi.org/10.1073/pnas.2106322119>.
- Lundström US, van Breeman N, and Bain D (2000) The podzolization process. A review. *Geoderma* 94(2–4): 91–107. [https://doi.org/10.1016/S0016-7061\(99\)00036-1](https://doi.org/10.1016/S0016-7061(99)00036-1).
- Moore TR, Paré D, and Boutin R (2008) Production of dissolved organic carbon in Canadian forest soils. *Ecosystems* 11(5): 740–751. <https://doi.org/10.1007/s10021-008-9156-x>.
- Nichols JE and Peteet DM (2019) Rapid expansion of northern peatlands and doubled estimate of carbon storage. *Nature Geoscience* 12(11): 917–921. <https://doi.org/10.1038/s41561-019-0454-z>.
- Paustian K, Lehmann J, Ogle S, Reay D, Robertson GP, and Smith P (2016) Climate-smart soils. *Nature* 532(7597): 49–57. <https://doi.org/10.1038/nature17174>.
- Regnier P, Friedlingstein P, Ciais P, Mackenzie FT, Gruber N, Janssens IA, Laruelle GG, Lauerwald R, Luyssaert S, Andersson AJ, Arndt S, Arnott C, Borges AV, Dale AW, Gallego-Sala A, Godderis Y, Goossens N, Hartmann J, Heinze C, Ilyina T, Joos F, LaRowe DE, Leifeld J, Meysman FJR, Munhoven G, Raymond PA, Spahni R, Suntharalingam P, and Thullner M (2013) Anthropogenic perturbation of the carbon fluxes from land to ocean. *Nature Geoscience* 6(8): 597–607. <https://doi.org/10.1038/ngeo1830>.
- Sanderman J and Amundson R (2008) A comparative study of dissolved organic carbon transport and stabilization in California forest and grassland soils. *Biogeochemistry* 89(3): 309–327. <https://doi.org/10.1007/s10533-008-9221-8>.
- Sanderman J, Baldock JA, and Amundson R (2008) Dissolved organic carbon chemistry and dynamics in contrasting forest and grassland soils. *Biogeochemistry* 89(2): 181–198. <https://doi.org/10.1007/s10533-008-9211-x>.
- Six J and Paustian K (2014) Aggregate-associated soil organic matter as an ecosystem property and a measurement tool. *Soil Biology and Biochemistry* 68: A4–A9. <https://doi.org/10.1016/j.soilbio.2013.06.014>.
- Sylvia DM, Hartel PG, Fuhrmann JJ, and Zuberer DA (2005) *Principles and Applications of Soil Microbiology*, 2nd edn Upper Saddle River, New Jersey: Pearson Education.
- Thompson A, Chadwick OA, Boman S, and Chorover J (2006) Colloid mobilization during soil iron redox oscillations. *Environmental Science & Technology* 40: 5743–5749.
- Tipping E and Woof C (1991) The distribution of humic substances between the solid and aqueous phases of acid organic soils: A description based on heterogeneity and charge-dependent sorption equilibria. *Journal of Soil Science* 42: 437–448.
- Ugolini FC and Sletten RS (1991) The role of proton donors in pedogenesis as revealed by soil solution studies. *Soil Science* 151(1): 59–75. <https://doi.org/10.1097/00010694-199101000-00009>.
- UN FAO (2021) *Statistical Yearbook—World Food and Agriculture*. Rome, Italy: United Nations Food and Agriculture Organization.
- von Luetzow M, Koegel-Knabner I, Ekschmitt K, Matzner E, Guggenberger G, Marschner B, and Flessa H (2006) Stabilization of organic matter in temperate soils: Mechanisms and their relevance under different soil conditions—A review. *European Journal of Soil Science* 57(4): 426–445. <https://doi.org/10.1111/j.1365-2389.2006.00809.x>.
- Wiley D, Kieber RJ, Eymann MS, and Brooks AG (2000) Rainwater dissolved organic carbon: Concentrations and global flux. *Global Biogeochemical Cycles* 14(1): 139–148.
- Williamson JL, Tye A, Lapworth DJ, Monteith D, Sanders R, Mayor DJ, Barry C, Bowes M, Bowes M, Burden A, Callaghan N, Farr G, Felgate S, Fitch A, Gibb S, Gilbert P, Hargreaves G, Keenan P, Kitidis V, Juergens M, Martin A, Mounteney I, Nightingale PD, Pereira MG, Olszewska J, Pickard A, Rees AP, Spears B, Stinchcombe M, White D, Williams P, Worrall F, and Evans C (2021) Landscape controls on riverine export of dissolved organic carbon from Great Britain. *Biogeochemistry* 2. <https://doi.org/10.1007/s10533-021-00762-z>.
- Xenopoulos MA, Barnes RT, Boodoo KS, Butman D, Catalán N, D'Amaro SC, Fasching C, Kothawala DN, Pisani O, Solomon CT, Spencer RGM, Williams CJ, and Wilson HF (2021) How humans alter dissolved organic matter composition in freshwater: Relevance for the Earth's biogeochemistry. *Biogeochemistry* 154(2): 323–348. <https://doi.org/10.1007/s10533-021-00753-3>.

Black carbon, pyrogenic carbon

Heike Knicker, Department of Biogeochemistry, Vegetal and Microbial Ecology, Instituto de Recursos Naturales y Agrobiología de Sevilla (IRNAS-CSIC), Sevilla, Spain

© 2023 Elsevier Ltd. All rights reserved.

Introduction	433
Definition of black carbon and the black carbon continuum	433
Origin of PyC	433
Methods for the characterization of the chemical structure of PyC	434
Elemental analysis and Van Krevelen diagrams	434
Degradative techniques	435
Gas chromatography and mass spectrometry (GC, MS)	435
Thermogravimetric analysis with differential scanning calorimetry	435
Spectroscopic techniques	436
Fourier transform infrared spectroscopy	436
Solid-state ¹³ C nuclear magnetic resonance spectroscopy	436
Alteration of the chemical structure of PyC during heating	437
Decomposition temperature of major biopolymers	437
Chemical structure of PyC	438
Transformation of cellulose and hemicellulose during heat treatment	438
Heat-induced alteration of lignin	439
Fate of Lipids during heating	439
Formation of black nitrogen	439
Conceptual model for the chemical structure of PyC	439
Effect of PyC on soil properties	440
Quality of SOM and soil color	440
Soil pH	440
Nitrogen and phosphorus availability for plants	440
Porosity	441
Identification and quantification of PyC in soils	441
Degradation pattern of PyC in soils	442
Preservation of PyC in soils	443
Conclusion	444
References	444
Further reading	445

Abstract

Black carbon or pyrogenic carbon (PyC) represents a ubiquitous entity of soil organic matter (SOM). Presently, it is best described as a continuum of thermally altered biomolecules. The degree of their alteration depends on heat severity, heating conditions (oxic, pyrolysis), and the nature of the feedstock. Comparably, a paradigm shift occurred with respect to the biochemical recalcitrance of PyC taking into consideration that the latter depends not only on the degree of charring of the respective PyC but also relies on soil conditions. There is a clear effect on PyC properties (increasing pH, darkness, hydrophobicity, and plant nutrient contents), which can survive on a long-term scale. Thus, fire and biochar amendment may play an important role in pedogenetic processes.

Key points

- Definition and description of the nature of black carbon or pyrogenic carbon (PyC)
- Description of the processes involved in the formation of PyC
- Overview of the impact of PyC on soil properties and their distribution in soils
- Introducing a concept describing the interaction between soil conditions and the occurrence/survival of PyC in soils

Introduction

Black carbon or pyrogenic carbon (PyC) represents a ubiquitous entity of soil organic matter (SOM) and as such, it has a considerable effect on soil ecological functions but also on the role of soils within biogeochemical cycles. It has been calculated that it amounts to 54–109 Pg of PyC within the first meter of soil (Bird et al., 2015). For U.S. agricultural soils about 35% of the SOM (Skjemstad et al., 2002) and between 4% and 18% of organic carbon in soils under native grassland are reported to be PyC (Glaser and Amelung, 2003). An estimate of the global distribution of PyC in the soil, atmosphere, and marine sediments can be found in Santin et al. (2016). In soils close to highly industrialized or coal mining areas, contributions from residues of fossil fuels and soot cannot be excluded. A further important PyC fraction in soils derives from ancient charcoal inputs. The effect of ancient charcoal from human activity during Neolithic times has been evaluated by Eckmeier et al. (2007), but there is also increasing interest in the impact of PyC on soil properties around former charcoal kilns. Indeed, PyC has been considered a defining property of the Terra Preta de Índios soils in South America. Its positive effect on soil fertility and the high biochemical recalcitrance of PyC initiated the idea of amending artificial biochar produced from biowaste as a soil improver and tool for increasing the C sequestration potential of soils. Based on this trend, it has been estimated that biochar production may add an additional 110–220 Tg of C per year to the soil (Woolf et al., 2010).

The dynamics of PyC in soils are largely influenced by the physical and chemical structure of the charred material. It can change soil properties such as nutrient availability, pH and the quality and quantity of SOM. Today, post-harvest burning and application of biochars as soil conditioners take advantage of the beneficial effect of PyC on soil fertility. Current scientific interest in PyC produced during vegetation fires is directed more to its role within the global carbon cycle, whereas much of what we know about the chemical structure and physical properties of PyC derives from research on biochar production and its amendment to soils.

Definition of black carbon and the black carbon continuum

Chemically, black carbon is defined as a component of fine particulate matter ($PM \leq 2.5$ in aerodynamic diameter) and consists of almost pure carbon with some oxygen and hydrogen bound into layered, hexagonal structures. It is a soot-like by-product formed during incomplete thermal decomposition of fossil fuels, biofuel, and biomass or by recondensation of volatile compounds formed during this process. The thermal degradation can be done by combustion, which means in the presence of oxygen or by pyrolysis during which oxygen is absent or nearly absent.

In soil and environmental sciences, the term black carbon is not limited to describing soot and soot-like compounds, but is also a continuum of pyrogenic organic materials (PyOM) from slightly to heavily charred organic residues. Thus, this so-called black carbon continuum comprises a wide range of charred biomass formed during natural and man-made vegetation fires, and also charcoal produced primarily for soil amendments and carbon sequestration. The nature of this material depends on the feedstock, production temperature and conditions. Man-made black carbon is often known as biochar. The latter is commonly abbreviated to BC, therefore, the abbreviation PyC is used for the term black carbon throughout this chapter to avoid confusion. Biochar is further distinguished as hydrochar or pyrochar. Hydrochar is obtained by pyrolysis via hydrothermal carbonization (HTC) in the presence of water and under self-generated pressure at temperatures between 180 °C and 350 °C, whereas pyrochar is the solid product formed under dry pyrolysis conditions at temperatures between 300 °C and 700 °C and higher. If the organic carbon content of the charcoal is above 50%, its atomic H/C ratio is <0.7 and if it derives from sustainable organic resources, it can be considered as biochar according to the European Biochar Certification (EBC, 2019). Formed by both anthropogenic activities and during natural fires, smaller particles of PyC can be transported in aerosols by wind reaching even to Arctic sites.

Combustion and pyrolysis lead to the transformation of organic matter to carbon dioxide (CO₂), carbon monoxide (CO), nitrogen gas (N₂), and volatile organic compounds (VOC) that are released, and the concentration of the organic fraction in the char decreases with heat intensity leading to relative preservation of the mineral phase. Complete combustion removes all organic matter leaving the mineral phase as a white powder commonly referred to as ash. Composition of the ash depends on the mineral content of the feedstock and is the main contributor to the strong alkalinity of many PyCs. In agriculture, it is often used for liming in particular of acid soils and as fertilizer.

Origin of PyC

In natural settings, vegetation fires are the main sources of PyC in soils. Fires have affected the Earth for at least 325 million years, thus fire occurred long before recent soils developed, and the respective PyC residues or markers can still be found in ancient sediments.

Most of the charred residues accumulating on the soil surface were produced by vegetation fires after partial combustion, thus in the presence of oxygen. Here, combustion is the chemical reaction where any fuel is oxidized and heat is produced. Under an unlimited supply of oxygen, clean combustion takes place and turns all organic matter into CO₂ and water. Under those conditions, PyC can only accumulate on the soil if the heat intensity affecting the feedstock was below the threshold needed for complete combustion. However, during burning of wood stems and branches, air supply may become restricted leading to pyrolysis

conditions that considerably increase the temperature needed for a complete degradation of the fuel. Furthermore, underground smoldering fires, occur under oxygen-depleted thus pyrolysis-like conditions.

Air transport of fine particles and volatile compounds of PyC released during a fire via aerosols or as fine particles can affect the health of ecosystems by their deposits on plant leaves and soils. The aerosols or fine particles contribute to increases in temperature, reduce the sunlight reaching the Earth, and modify rainfall patterns. Changing rain patterns can have far-reaching consequences for both ecosystems and human livelihoods, for example by disrupting monsoons, which are critical for agriculture in large parts of Asia and Africa. However, most of the PyC reaches the soil after vegetation fires or as a soil amendment in the case of biochar. Wildfires of medium intensity, in particular, produce large amounts of PyC (Certini, 2005), a considerable portion of which suffers erosion and is transported from terrestrial to aquatic systems either as particles or as dissolved PyC. These can contribute to the organic matter in river, lake or ocean sediments. After having been partially oxidized at the soil surface, the increased solubility of PyC in water allows vertical transport either with the leaching water or adsorbed on soil particles into deeper soil regions where it can accumulate (Velasco-Molina et al., 2016). However, transport within the soil can also be caused by bioturbation. Wagner et al. (2017) suggested that other components of dissolved organic matter may act as mediators in the solubilization of condensed aromatic molecules and serve to increase the solubility of PyC.

The extent of PyC production during the burning of organic matter depends largely on combustion intensity, i.e., fire intensity. Prescribed (controlled) fires and post-harvest fires are initiated at moderate soil moisture and generally have low severity. Fires in grassland with lower fuel loadings usually have ground-level temperatures of $<25^{\circ}\text{C}$, although higher temperatures have been measured. Uncontrolled wildfires in areas with abundant dry fuel can be severe, although the severity can be very heterogeneous due to variation in fuel distribution and other local factors. Intense forest fires can lead to temperatures of at least 1500°C , but soil surface maximum temperatures are usually around $500\text{--}700^{\circ}\text{C}$. Rates of spread are commonly fast (i.e., as much as 7 km h^{-1}). Fast moving fires on grass may be intense in terms of energy release per unit area, but do not transfer the same amounts of heat to mineral soil as do slow-moving fires involving moderate to heavy fuel loads. However, the longer is the duration of the fire, the greater is the belowground damage to ecosystems. With depth, the severity of heating decreases depending on factors such as intensity and duration of heat transfer, heat conductivity of the mineral phase, soil porosity and soil moisture. In general, in dry soils, below-ground temperatures will rise very slowly because they are a very good insulator.

In peat and swamp areas smoldering fires spread at low speed, as low as 0.5 m week^{-1} . Lowering the water table is a considerable hazard because drying the peat creates a significant amount of high-energy fuel. Fires that can burn deep into the peat may even be able to survive several centimeters of rainfall. The respective loss of organic material can be dramatic such that the peat surface sinks clearly with respect to its surroundings.

Methods for the characterization of the chemical structure of PyC

Several models that describe char on a molecular level are reported, but a real understanding of the chemical nature of PyC is still lacking, most probably due to the heterogeneous nature of its constituents and the strong chemical resistance of some that reduces its accessibility to extraction and subsequent analyses. Although HTC, pyrolysis and incomplete combustion result in blackish material with large contents of aromatic C, the chemical reactions involved differ and determine the chemical composition of the final product. In addition, the charring intensity expressed in heat intensity, length of time of thermal treatment, and the nature and composition of the feedstock affect the nature and properties of PyC.

Elemental analysis and Van Krevelen diagrams

A first approach to reveal the chemical structure of PyC represents determination of the elemental composition (contents of carbon, nitrogen, oxygen and hydrogen), which is mainly performed by dry combustion. Techniques that use chemical oxidation such as the Walkley Black method are of limited use since they may miss inert forms of carbon such as highly polyaromatic structures.

The elemental composition of the PyC can be plotted as a Van Krevelen diagram, originally developed for assessing the origin and maturity of kerogen and petroleum. It cross-plots the hydrogen:carbon (H/C) atomic ratio as a function of the oxygen:carbon (O/C) atomic ratio (Fig. 1). The H/C ratio separates compounds according to their degree of saturation, whereas the O/C ratios display the O classes. Specific regions of the Van Krevelen diagram are associated with specific compound classes, such as carbohydrates, lipids, lignin or peat. Following the chemical alteration of biomass compared with heat severity via a Van Krevelen plot, shifts assigned to decarboxylation, dehydration or demethylation can be followed by the location in the plot. Maturation of natural coals or increasing heat intensity during PyC formation is associated with dehydration, condensation and decarboxylation which moves the cross plots to the lower left corner of the diagram. Accordingly, the location of the cross plots provides valuable information about the degree of aromatization and coalification of the PyC.

For soot, atomic H/C ratios of <0.2 were determined, which accord with a highly condensed aromatic network. Graphenic structures have also been confirmed for PyC produced under pyrolysis conditions (without oxygen) at temperatures $>700^{\circ}\text{C}$ (Freitas et al., 1999). However, PyC produced at lower temperatures shows atomic H/C ratios >0.5 . Paneque et al. (2017) and Hoffmann et al. (2019) found atomic H/C ratios of hydrochars between 0.7 and 1.1. Such values demonstrate that, on average, at least every second C is protonated and consequently the chemical structure of such PyC differs greatly from that expected for soot.

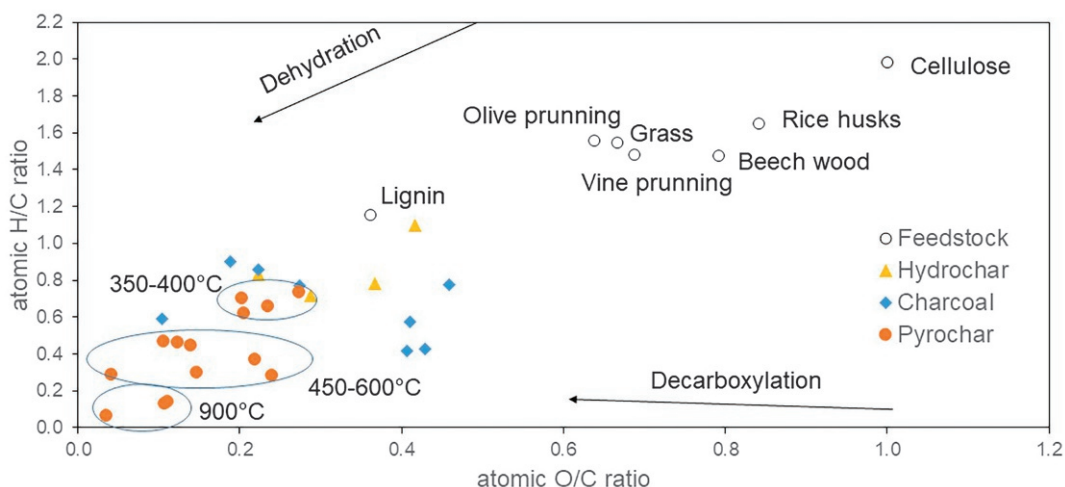


Fig. 1 Van Krevelen Plot correlating the atomic H/C and O/C ratios of different feedstocks, hydrochars, pyrochars and chars produced under oxidic conditions. The circles embrace data derived from pyrochars produced within the indicated range of pyrolysis temperature. Data are obtained from Knicker H, et al. (2008) A new conceptual model for the structural properties of char produced during vegetation fires. *Organic Geochemistry* 39(8): 935–939. <https://doi.org/10.1016/j.orggeochem.2008.03.021>. Knicker H (2010) “Black nitrogen”—An important fraction in determining the recalcitrance of charcoal. *Organic Geochemistry* Pergamon. 41(9): 947–950. <https://doi.org/10.1016/j.orggeochem.2010.04.007>. Paneque M, et al. (2017) Hydrothermal carbonization and pyrolysis of sewage sludges: What happen to carbon and nitrogen? *Journal of Analytical and Applied Pyrolysis* 128: 314–323. <https://doi.org/10.1016/j.jaap.2017.09.019>. Hoffmann V, et al. (2019) Conductive carbon materials from the hydrothermal carbonization of vineyard residues for the application in electrochemical double-layer capacitors (EDLCs) and direct carbon fuel cells (DCFCs). *Materials Switzerland*. 12(10): 1703. <https://doi.org/10.3390/ma12101703>. Campos P, et al. (2020) Chemical, physical and morphological properties of biochars produced from agricultural residues: Implications for their use as soil amendment. *Waste Management* 105: 256–267. <https://doi.org/10.1016/j.wasman.2020.02.013>.

The concept of differing chemical structures for soot and charcoal PyC is also supported by the atomic O/C ratios. Soot exhibits typical values of <0.3 , whereas PyC originating from plants and SOM have average atomic O/C ratios of up to 0.5. Such relatively large O concentrations are inconsistent with a large and graphitic network of condensed polyaromatics. They tend to confirm the participation of anhydrosugars, furans, pyranones, and 5 hydroxymethylfurfural (HMF) formed by the thermal alteration of cellulose and other carbohydrates.

Degradative techniques

For a more detailed analysis of the nature of PyC, more sophisticated tools are required. Many of these require solubility of the sample or at least of different fractions. This is rarely the case for PyC and for most of its fractions, degradative pre-treatments based on chemical and thermal degradation with subsequent derivatization of fragments produced have been applied to transform at least part of the material into structures that can be identified.

Gas chromatography and mass spectrometry (GC, MS)

Analytical pyrolysis combined with gas chromatography mass spectrometry (GC/MS) is used because it does not require solvent use or pre-concentration of the sample, and is not excessively time-consuming. However, its application for the analysis of heterogeneous mixtures may result in secondary rearrangements of reactive intermediates which can yield components that were not present in the original sample. Evidently, this can bias the interpretation of the data obtained. Correspondingly, proteins may lose their N-containing functional groups and become heterocyclic structures. Polysaccharides can form levoglucans or furans and pyrans by dehydration. Fatty acids may decarboxylate producing n-alkane or n-alkene pairs of which a small proportion may cycle to alkylbenzenes and alkyl-naphthalenes or lignins may be demethylated. In addition, one must also bear in mind that comparable reactions occur during analytical pyrolysis and PyC formation, which reduces the distinction between the origin of the compounds identified.

A further problem is that only compounds that have not been volatilized during previous PyC formation can be released during analytical pyrolysis. Consequently, compounds of PyC can be identified only if the temperature during analytical pyrolysis is higher than that during PyC formation.

Thermogravimetric analysis with differential scanning calorimetry

Thermogravimetric analysis (TGA) probes the thermal stability of materials and relates it to specific compound classes present in the sample. The sample is heated to high temperatures at a constant rate while monitoring heat-induced weight changes. The mass loss at a certain temperature can be associated with the decomposition of different sorts of OM. By combining TGA with differential scanning calorimetry (DSC) and a pack bed, together with online gas measurement using Fourier-transform infrared (FTIR)

spectroscopy enables the identification of hemicellulose, cellulose, and lignin in biomass together with the gas products from the biomass pyrolysis.

The TGA–DSC is already applied routinely to characterize chemical changes in organic matter fractions of soils and sediments, degraded plant tissue and composts, but has also been used to characterize forms of PyC. Observations using this technique confirmed the idea of PyC as a continuum in terms of thermal stability. Addition of charcoal and soot to soils revealed a strong linear relationship between the amount of added and detected PyC, from which it was concluded that this method may be suitable to quantify thermally stable PyC in soil. Coupling TGA–DSC to an isotope ratio mass spectrometer (IRMS) enables better discrimination of the organic matter and PyC components of different sources and varying thermal stability.

Spectroscopic techniques

Spectroscopic techniques such as infrared (IR) or solid-state nuclear magnetic resonance (NMR) spectroscopy may be applied as alternatives to degradative analytical tools. Infrared measures the interactions of infrared radiation with organic matter and identifies the vibrations of different bonds, whereas NMR takes advantage of the magnetic properties of a nucleus in different chemical and physical environments and probes its interactions with a radio-frequency pulse in a static magnetic field.

Fourier transform infrared spectroscopy

A typical Fourier transform (FT) IR spectrum of PyC reveals adsorption bands around 3400 cm^{-1} assignable to H-bonded O–H stretching vibrations of hydroxyl groups from alcohols, phenols, and organic acids around $2900\text{--}2800\text{ cm}^{-1}$ indicating the presence of CH groups. Absorption at a wavelength of $1700\text{--}1400\text{ cm}^{-1}$ is evidence of aromatic and olefinic C=C vibrations and C=O in amide (I), ketone and quinone groups. A peak around 1680 cm^{-1} , corresponds to carbonyl groups, a band at 1460 cm^{-1} to C–H deformation of CH_3 groups and adsorption at wave number $900\text{--}700\text{ cm}^{-1}$ to aromatic CH. Important advantages of this technique include its low costs, small sample size required and ease of use, but it fails to provide quantitative data. If this is needed, the more expensive but highly sophisticated solid-state NMR spectroscopy is recommended.

Solid-state ^{13}C nuclear magnetic resonance spectroscopy

Most solid-state ^{13}C NMR spectra of PyC are defined by an intense but broad signal assigned to aromatic C (Fig. 2), which distinguishes them from spectra of common soils with no impact of fire or biochar addition. Depending on the carbonization

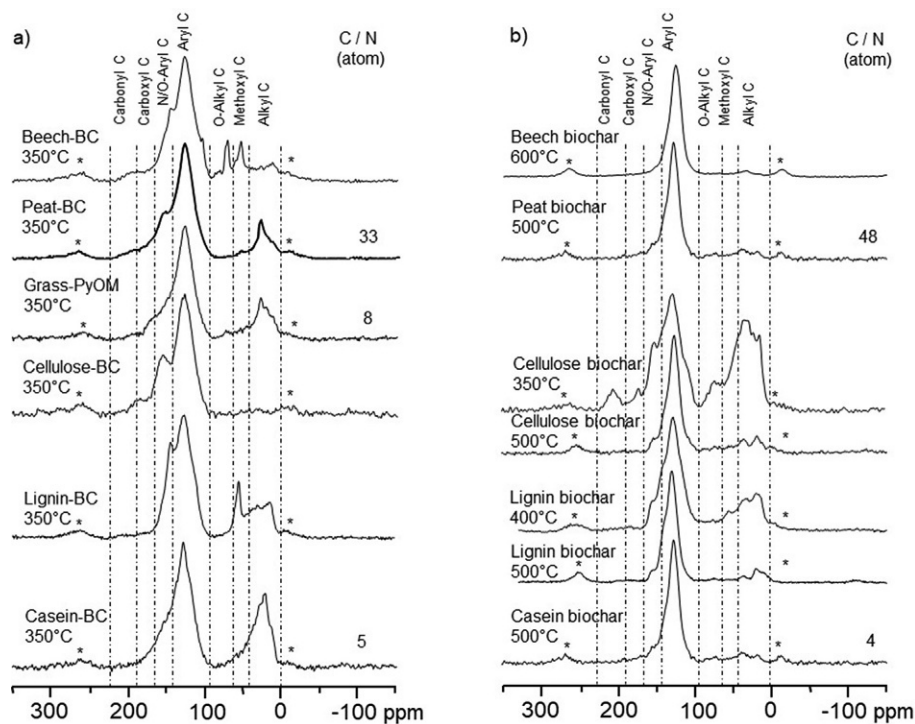


Fig. 2 Solid-state ^{13}C NMR spectra of pyrogenic C (PyC) from different feedstocks and produced under (a) oxidic conditions and (b) pyrolysis.

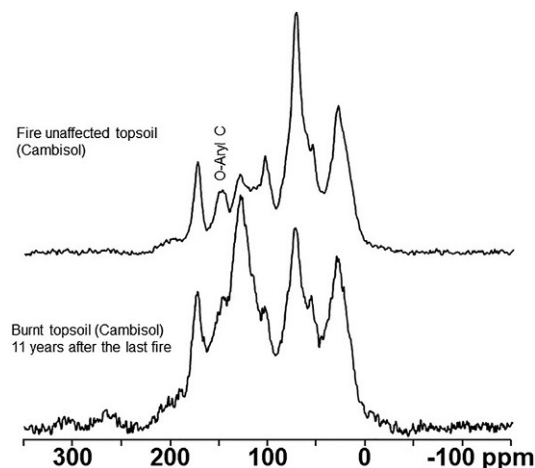


Fig. 3 Solid-state ^{13}C NMR spectra of soil organic matter (SOM) in topsoils of Cambisols from Central Spain which were not affected by fire and were taken 11 years after a severe fire event.

degree of the PyC smaller signals attributed to alkyl C are also visible in the spectrum. A more detailed analysis of such spectra often allows the identification of a small shoulder assignable to O-substituted aromatic C (O-aryl C) on the low field side of the peak of aromatic C resonance. In an NMR spectrum of a typical soil such a signal is attributed to phenol C in lignin or tannins (Fig. 3). Those, however, suffer considerable modification during heat treatment at elevated temperatures. Although the contribution of this phenol to the total aromaticity represents an important parameter not only for a better understanding of the chemical structure but also for elucidating the reactivity and adsorption capacity of a biochar, it is rarely considered, most probably because a highly condensed aromatic network is generally assumed.

Solid-state ^{13}C NMR spectra of soil and PyC are commonly acquired with the cross polarization technique. This is based on the transfer of magnetization from the highly abundant ^1H spin system to the rare ^{13}C spins (cross polarization; CP) which are finely detected and quantified. In brief, the time needed for a complete transfer, thus for a full magnetization of the ^{13}C spins depends upon the strength of the interactions or coupling between ^1H and ^{13}C . If ^1H couples strongly with ^{13}C , for example for directly bound ^1H in carbohydrates, the CP is efficient and complete within a short CP time (T_{CH}), whereas a considerable increase of this parameter is expected in condensed polyaromatic structures. This behavior can create a problem because during CP, the ^1H spin system loses energy to the environment (relaxation) and can only partially cross-polarize the ^{13}C . However, if the relaxation is sufficiently long to allow almost complete cross polarization, acquisition parameters can be applied that circumvent biased quantification of such spectra.

To obtain more detailed information about the structure of the aromatic network in PyC, modern solid-state NMR spectroscopy offers special pulse sequences which allow an improved assignment of signal intensity to C groups. One of these is the two-dimensional (2-D) ^1H - ^{13}C heteronuclear correlation (HETCOR) NMR experiment in which ^{13}C and ^1H are correlated through their spin-spin interactions. This allows identification of the connection between specific proton and carbon groups and has also found its way into PyC research contributing to a better understanding of the processes that occur during the formation of PyC.

Alteration of the chemical structure of PyC during heating

Decomposition temperature of major biopolymers

The thermal stability of biomass depends strongly on the reaction medium in which the process is carried out. Under pyrolytic conditions, hemicellulose and cellulose will decompose at lower temperatures than lignin. Hemicellulose decomposes at between 220 °C and 315 °C, followed by cellulose (315–400 °C). The preferential loss of carbohydrates is also indicated by an increase of $\delta^{13}\text{C}/^{12}\text{C}$ of PyC relative to the feedstock (1–2‰) attributable to loss of the isotopically heavier cellulose.

The decomposition of lignin occurs over a temperature range from 160 °C to 900 °C. Sharma et al. (2004) reported that approximately 40% and 60% were volatilized at 450 °C and 750 °C, respectively. Condensation reactions among the volatile products, leading to the formation of coke, were absent under the pyrolysis conditions used. Repeating the experiment under oxidic conditions led to comparable char yields at temperatures up to 350 °C, but at 550 °C only 20% was sequestered from complete combustion that remained as char.

Experiments during which lignin, cellulose and grass residues were burnt (under oxidic conditions) at 350 °C for 8 min yielded 67%, 11% and 34% recovery, respectively, of the initial weight as solid residue (Knicker et al., 2008). Increasing the temperature to 450 °C decreased the yield to 23%, 3% and 18%, respectively (Knicker, 2010). After pyrolysis of cellulose at 350 °C for 1 h, 33% of the total weight was recovered (Knicker et al., 2021), indicating that under oxidic and anoxic conditions different reactions are responsible for the thermal degradation.

With casein as a reference for proteins and peptides in PyC precursors, burning in an oxic environment at 350 °C and 450 °C for 4 min resulted in a recovery of PyC of 48% and 23%, respectively (Knicker et al., 2008). Applying TGA showed that thermal degradation of casein follows a complex mechanism with chemical degradation starting at 176 °C. The main degradation was observed between 248 °C and 634 °C leaving a stable residue of 20% of the initial sample (Mocanu et al., 2012).

During HTC, the degradation and depolymerization of biomass occur at considerably lower temperatures than during dry pyrolysis. Accordingly, hemicellulose and cellulose decompose around 160–180 °C, whereas most of the lignin remains stable until near or above the critical point of water. With the presence of hot compressed water in the HTC process, the degradation of biomass is primarily initiated by hydrolysis, resulting in cleavage of ether and ester bonds.

Further factors that have to be considered if heating temperature and organic matter degradation are correlated, are moisture of the feedstock and heating time. Burning moist feedstock is associated with heavy smoke that condenses to tar and acidic residues on cold surfaces. It produces less heat and leads to incomplete combustion. Thus, considerably higher temperatures than those reported for complete combustion from laboratory experiments are needed to allow charring of fresh wood or moist litter layers. Furthermore, heating time plays an important role, since chemical reactions can only occur if the respective activation energy has been reached and even surpassed. This requires that the heat needed has completely and homogeneously penetrated through the heated feedstock. Campos et al. (2020) showed that pyrolysis of different feedstocks at 500 °C for 30 min was not long enough to remove alkyl C; it required heating for 1 h for an almost complete removal of such a structure. Increasing pyrolysis time to 2 and 4 h, however, did not cause major additional chemical alterations. Charring grass residues under oxic conditions at 350 °C, a burning time of 75 s, was needed to initiate noticeable heat-induced alteration of the chemical composition of solid residue, although weight losses were identified with less heating time (Knicker et al., 1996). With this in mind, the discrepancy between the relation between temperature and chemical changes of organic matter during wildfires and organic matter pyrolyzed or combusted in laboratory experiment as described by Santín et al. (2016) can be explained. They observed that under typical wildfire conditions substantial chemical changes in the organic matter of the forest floor occurred at temperatures >600–700 °C rather than 30–400 °C commonly assumed from laboratory experiments. Charred organic residues were still detected after the fire reached temperatures of 950 °C, a temperature at which under oxic conditions all organic matter should have experienced complete combustion. However, the rate of spread of the observed fire was around 6–7 m min⁻¹ resulting in average heating durations of the forest floor of 180 s at 300 °C, 81 s at 500 °C and 21 s at 700 °C. Such short heating times would certainly not allow efficient heat and energy penetration below the outer surface of the forest floor.

Chemical structure of PyC

Most of what we know about thermal decomposition of biomass comes from experimental studies carried out under pyrolytic conditions. A few studies only considered thermal degradation in the presence of oxygen.

Older models describing the chemical nature of PyC assume a polycondensed aromatic network in which the cluster size increases with temperature. Ultimately, micrographitic sheets are formed. However, newer research, in particular in fuel and material sciences, confirmed that such a model is oversimplified if a description of PyC is intended.

Transformation of cellulose and hemicellulose during heat treatment

The solid-state ¹³C NMR spectra in Fig. 2 indicate that the organic composition of PyC from different feedstocks produced under different conditions can vary considerably, although all spectra reveal elevated aromatic C concentration if compared to the original feedstock. Thermal decomposition starts with dehydration and condensation reactions which accords with decreasing atomic H/C and O/C ratios. At mild temperatures, this affects mainly carbohydrates and leads to their depolymerization and fragmentation, and also the formation of volatile products together with char. Major products are levoglucosan and furanose derivatives, furfural, HMF and hydroxybenzene. In an ¹³C NMR spectrum, this can be verified through the relative decrease of the O-alkyl C signal intensity by the concomitant increase of signal intensity in the aryl C and alkyl C regions. Based on this observation the aromatic C/alkyl C ratio was suggested as an index to describe the degree of charring of vegetation residues and SOM. However, here one has to keep in mind that feedstocks with a large content of lipids or proteins will shift this ratio toward higher alkyl contents because the alkyl chain of these biomolecules has higher thermal stability than carbohydrates.

As indicated by X-ray diffraction, heating of biomass goes together with a decrease in crystallinity, and the augmentation of amorphous structures by the destruction of hydrogen bonds and cleavage of molecular chains. Above 550 °C graphene sheets are formed, although the overall structure remains substantially disordered. With increasing temperature progressive stacking of the graphene sheets and lateral growth of graphene planes resulting in the formation of crystallites were deduced. These observations accord with those of Tang and Bacon (1964) who reported a large weight loss and breakdown of crystalline structures after pyrolysis of cellulose at 400 °C, which was associated with the thermal scission of 1–4 glycosidic linkages and dehydration reactions. They identified aliphatic aldehydes, ketones and acids which seemed to be an integral stable part of the solid char at these temperatures. With higher temperatures their contribution declined due to thermal decarbonylation. There is evidence that most of the aromatic structures of the cellulose char pyrolyzed at 350 °C can be assigned to benzofuran-like units rather than to polyaromatic clusters. Their contribution decreases with pyrolysis temperature probably because of bond cleavage and conversion of furan groups to hydroxyl-substituted aromatic C. However, Fig. 2 shows that different reactions occur during thermal degradation of cellulose under oxic and anoxic conditions from the missing signals assignable to aliphatic aldehydes, ketones and acids in the spectrum of cellulose PyC produced under oxic conditions.

The different thermal stability of hemicellulose and cellulose shows that the mechanism involved in their thermal degradation also differ. Falco et al. (2011) suggested that glucose units are dehydrated to HMF that condense to a polyfuranic network and subsequently condensate to polycondensed aromatic units with occasional substitution of aromatic rim carbon.

Heat-induced alteration of lignin

Lignin is a polymer of phenylpropane units containing three different aromatic ring substitution patterns: p-hydroxyphenyl, guaiacyl (4-hydroxy-3-methoxyphenyl), and syringyl (3,5-dimethoxy-4-hydroxyphenyl), depending on the wood species. Its thermal degradation starts at a higher temperature than hemicellulose and cellulose, therefore, its solid residue is relatively enriched in plant-derived PyC formed at temperatures $>350^{\circ}\text{C}$.

The first change in the thermally induced alteration of the solid residue is comparable to the pyrolysis of cellulose and concerns the loss of O-alkyl C, which is best explained with dehydroxylation and partial loss of propanyl side chains. These reactions are the main cause of the weight loss both during pyrolysis and partial combustion, although some removal of aromatic C occurs. The relative increase of the aromatic C during thermal treatment of lignin is different from that of cellulose and is caused by selective preservation of the lignin-derived aromatic structures rather than by their neoformation both under oxic and pyrolytic conditions (Sharma et al., 2004). Knicker et al. (2008) concluded that demethylation leaving hydroxylated aryl C rather than demethoxylation occurs at temperatures up to 350°C . For this char, they revealed that at least every second aryl C is protonated which corresponds to small cluster units such as anthracene or benzanthracene. Higher temperatures, however, decreased the contribution of phenol C concomitantly with alkyl C, leaving a small signal attributed to terminal methyl groups. Sharma et al. (2004) suggested that the weight loss observed during the heating of lignin from 400°C to 600°C is caused by the elimination of phenol C which is likely to result in the condensation of smaller aromatic rings into larger polycondensed aromatic networks.

Fate of Lipids during heating

Prolonged charring reduces the average chain length of n-alkanes and shifts the characteristic odd/even predominance of fresh plants toward a balanced odd/even distribution. Unsaturated fatty acids are more affected by heat than their saturated counterparts. Although it was suggested that the composition of aliphatic and aromatic hydrocarbons of fire-affected recent and fossil soils offers specific indicators, "fingerprints," for the diagnosis of vegetation burning and burning conditions, one has to consider that their microbial decomposition and biosynthesis within a soil matrix can rapidly modify n-alkanes and fatty acid series. Thus, their instability against biodegradation may hamper the identification of charcoal residues in such environments.

Formation of black nitrogen

Originally, black nitrogen (BN) was not a common concept included in the description of PyOM, since most PyOM was expected to derive from wood residues commonly containing negligible amounts of nitrogen (N). However, charring protein-rich plant residues often leads to decreasing C/N ratios which is best interpreted by a selective enrichment of N-containing compounds. With the introduction of solid-state ^{15}N NMR spectroscopy as a routine analytical tool for the characterization of soil organic matter, a door opened that allowed a more detailed analysis of the N-functionality of charred residues and coal.

Applying this technique revealed that heteroaromatic N represents an integral part of the aromatic structure of PyC. Based on this, the term BN was introduced with the intention to enhance attention that aside from pure organic C structures, PyOM can contain organic N moieties if it derives from N-containing feedstock. The heat-resistant N compounds were identified as pyrrole or indole-type N with a minor contribution of pyridine N derives mainly from proteins and peptides that decompose thermally by systematic and random depolymerization reactions. Note that many proteins have higher thermal recalcitrance than cellulose.

Another important reaction that occurs during charring of biomaterial is the Maillard reaction. During this process, ammonia or free amino groups of amino acids and amino sugars react with carbonyl groups of sugars to form Schiff bases that subsequently undergo rearrangements via Amadori compounds to produce dark-colored melanoidins. The major compounds found after pyrolysis of four Amadori compounds derived from glucose and that reacted with asparagine, valine, leucine and threonine were ketones, aldehydes, pyridines, pyrazines, pyrroles and carboxylic acids.

Taking the narrow atomic C/N ratios for pyrrole ($\text{C/N} = 4$), pyridine ($\text{C/N} = 5$), and carbazole ($\text{C/N} = 12$), the approximate contribution of such heterocyclic compounds to PyOM can be estimated from the C/N ratio of PyOM. In particular for N-rich plant residues, BN can comprise a considerable fraction of the charred residue. After having been incorporated into soils, the bound N will become part of the refractory N pool which is not directly available for biomass production. The expected high biochemical recalcitrance of this nitrogen initiated the idea that N-rich PyC could be used as slow-release fertilizer.

Conceptual model for the chemical structure of PyC

According to results obtained from charring experiments and reports from the pyrolysis literature, PyCs are considerably heterogeneous which will determine their physical properties and behavior in soil systems. Based on this it was proposed to replace the former model of highly polycondensed aromatic structures with a concept that reflects better the strong variability of feedstock and production conditions. According to this concept (Knicker et al., 2008), PyC is a heterogeneous mixture of thermally altered biomolecules. The degree of their alteration depends upon the severity of the heat treatment and the chemistry and heating conditions. In particular, in plant-derived material heated to 400°C furans and ketones and pyrroles occur together with sparsely altered lignin residues and some lipids. At this stage, the degree of condensation of the aromatic structures is relatively low and does not exceed an average aromatic cluster size of six rings. Note that for cellulose, in particular, different chemical reactions occur

under oxic and pyrolysis conditions. Intensively charred material (>500–600 °C) reflects a greater degree of condensation and has an entirely aromatic structure with little alkyl C residues. With respect to elemental analysis N, O and probably also S (sulfur) substitutions are common features of the charring products.

Effect of PyC on soil properties

Quality of SOM and soil color

Incorporation of PyC into soils increases the concentration of aromatic C and N and to a lower extent that of alkyl C, which alters not only the quantity but also the quality of SOM (Fig. 3).

During natural humification, acidic functional groups are formed, whereas the addition of PyC decreases the relative contribution of O-groups and increases hydrophobicity. The latter may affect soil water distribution and adsorption properties. After wildfires, the content of water-soluble organic compounds declines although there is a gradual restoration over time. A further major alteration of soil properties induced by the addition of PyC is darkening of the soil, which is commonly related to the increasing aromaticity. However, there was no correlation between the aromatic C content of model chars and darkness (Knicker, 2011) indicating that the latter is also determined by the specific structure of the aromatic network. Other factors such as calcium carbonate content, soil texture and interaction with metals may contribute to the darkness of a soil. Long-term effects of PyC addition can be evaluated by characterizing soils around former charcoal kilns. In addition to larger SOM contents, greater aromaticity and a lower percentage of dissolved organic matter, the large and continuous input of PyC into the soils surrounding such ancient kilns resulted in higher pH values, cation exchange capacity and nutrient contents than in comparable soils that were unaffected by the former charcoal production. Based on such observations, it was suggested that the natural weathering and biochemical reworking of charcoal in soils, together with ash input, is sufficient to generate fertile and resilient soils with unusual SOM and properties commonly associated with the high fertility and C sequestration potential of the typical Terra preta de Índios.

Soil pH

Loss of CO₂ and volatile organic compounds during partial combustion and pyrolysis of vegetal feedstocks results in an accumulation of mineral matter with a large content of cations causing an increase of pH to alkaline values, together with large electrical conductivity. This provides not only nutrients previously immobilized in the growing plant to the soils but can also contribute to an increase of the soil pH; although significant effects are only expected for PyC produced at temperatures >450 °C. In contrast to soils with frequent inputs of PyC, however, this effect, known as the liming effect, does not persist on a long-term scale in soils occasionally affected by a wildfire. Here, the nutrients are rapidly used for the build-up of the re-growing vegetation or they are lost by leaching or removal of PyC by erosion. In particular, on steeply sloping soils, the latter represents a major danger for nutrient depletion of fire-affected sites. The low density and high hydrophobicity of PyC and the fact that it is accumulating on the soil surface make this material particularly susceptible to transport by water flows following heavy rains into catchments, river systems and eventually into oceans. This will not only remove nutrients but will also dislocate considerable amounts of organic carbon associated with PyC.

Nitrogen and phosphorus availability for plants

Production of PyC is accompanied by N losses to the atmosphere, which in the case of vegetation fires is lost from the respective ecosystem. On the other hand, ammonium as a direct product of combustion can accumulate in the burnt soil. Nitrate on the other hand is not directly formed by the fire but derives from the transformation of ammonium (nitrification). In undisturbed soils, the use of ammonium by nitrifiers is in strong competition with the vegetation. Destruction of the vegetation and combustion of the allelopathic compounds, such as terpenes and phenolics promotes the increase of nitrifiers and the transformation of available ammonium to nitrate. However, without any potential plant uptake nitrate will be lost from the ecosystem either by denitrification or leaching. If prompt recolonization by plants does not occur, this loss can have subsequent effects on plant succession and ecosystem restoration.

A considerable fraction of the N of feedstock is sequestered as BN, which can be mobilized and subsequently used by microorganisms and plants for build-up of their biomass. Its rate of incorporation into microbial and plant biomass is considerably less than that from inorganic N or unburned organic N. In particular, in fire-affected areas, the lower bioavailability of N in BN prevents complete N-loss from the fire-affected soil layer by leaching after the first post-fire rains. Therefore, BN accumulation after fires may play a more important role in the recovery of a fire-affected ecosystem than formerly thought.

A comparable conclusion may be drawn for phosphorous (P) which accumulates in the solid residue after heating of vegetal feedstocks. Burning liberates organic P from SOM as plant-available orthophosphate which is not lost by volatilization and can accumulate on the soil surface. The bioavailability of this P may be promoted by the liming effect since the greatest P bioavailability is around pH 6.5. However, binding of P to Ca-minerals in neutral and alkaline soils and to Al (aluminum), Fe (iron), and Mn (manganese) in acid environments cancels out the positive nutrient effect. Similar to N, heating of biomass leads to a relative

enrichment of P in the PyC. Its plant availability declines with reaction temperature (Dai et al., 2016). After PyC has been amended to soil covered with plants, part of the P that was immobilized into the PyC turns into mobile plant-available forms (Dai et al., 2016). Future research needs to show if this is due to processes initiated by development of the root and vegetative system of the plants or by microbial activities in the rhizosphere.

Porosity

The destruction of cell structure and the chemical alteration of biomolecules during heating of vegetal biomass increases the porosity during PyC formation. Pore size and distribution in PyC represent an important factor defining its ecological role in soils. In addition to providing niches for microbes, adsorption sites for nutrients, and also metals, this property can have a major positive impact on the water holding capacity of the amended soil.

Measuring the porosity of PyC is challenging because pore size ranges over at least five orders of magnitude, from sub-nanometer sizes to pores of tens of micrometers that remain from partially destroyed cellular structures. Macropores and mesopores can be revealed by microscopic techniques such as X-ray microtomographic imaging or scanning electron microscopy.

Smaller pores in the range of nm (micropores and mesopores) are commonly analyzed by gas adsorption techniques accompanied by the Brunauer–Emmett–Teller (BET) modeling to determine the specific surface area (SSA). However, pore space analysis based on these methods and models is limited to pores smaller than 300 nm, which gives little information about its water holding properties once introduced into soils. The most common probe for the determination of SSA is N₂ (at 77 K). However, its value for PyC is questionable because of deformation and swelling processes during adsorption and the fact that N₂ adsorption is severely restricted by diffusion limitation in narrow micropores <0.7 nm. Alternatively, CO₂ adsorption is applied for the characterization of pores <2 nm. Mercury porosimetry can probe pores from a few nanometers to a few hundred micrometers, thus it can detect more of the biochar pore volume than gas adsorption. Apart from safety issues, however, the drawbacks of this technique include the risk of crushing the sample with the high pressures required for analysis.

Identification and quantification of PyC in soils

The chemical composition of PyC does not represent a single structure but a wide range of thermally altered materials, therefore, its quantification in soil is challenging. This is particularly so if the quantification includes PyC produced at low temperatures since their properties can overlap with those of uncharred SOM.

Currently, optical, thermal, chemical and photo-oxidation approaches, and also spectroscopic and molecular marker methods are applied for the identification and quantification of charcoal. Each of these techniques relies on operational definitions and covers different regions within the combustion continuum.

Visual or microscopic techniques identify charcoal pieces visually and quantify their amount by subsequent particle counting or physical separation with weighing. This method, however, misses soot and charcoal degradation products. Physical techniques are mostly non-destructive and separate PyC from soil by differences in density or size.

Chemical and photo oxidation, and also thermal methods rely on the refractory nature of PyC, assuming that the latter is selectively enriched whereas the more labile unburned material is destroyed. The widely used technique of chemo-thermal oxidation was calibrated for soot and excludes the identification of PyC at <850 °C. Alternatively, TGA–DSC allows scanning over a wide range of temperatures and thus can provide information about the entire PyC continuum. Assuming that PyC is systematically more stable than uncharred SOM, the first can be distinguished from the latter by a typical thermal signature. However, only slightly charred residues produced below the degradation temperature of lignin may still be underestimated.

Chemical oxidation with peroxide also induces substantial degradation of PyC. Alternatively, oxidation with dichromate is used to separate labile organic matter from highly condensed PyC. Grass and wood are efficiently degraded after 4–6 h, but paraffin structures from plant waxes survive even harsher oxidation scenarios. A considerable part of those structures is removed by subsequent Soxhlet extraction, indicating that their persistence is due to their hydrophobic nature rather than an inherent chemical recalcitrance caused by charring. Persistence of paraffin structures has also been reported in residues after dichromate treatment of some Australian Mollisols and Oxisols (Krull et al., 2006), in sediment samples (De La Rosa et al., 2008) and a peat humic acid oxidized with nitric acid (Simpson and Hatcher, 2004). Consequently, the resistance of such structures to strong acids can be described as a common feature, which can bias PyC quantification with techniques based on chemical oxidation in particular if soils with higher lipid contents are analyzed. A further drawback of this method is that aged, thus carboxylated PyC, can be chemically oxidized and therefore will escape identification by such methods.

Biomarker essays are also destructive and rely on the identification of typical charring products as molecular biomarkers such as benzenepolycarboxylic acids (BPCAs) released after oxidation with nitric acid or levoglucans which are a typical pyrolysis product of cellulose. The latter is used for identification rather than quantification. The BPCAs method is also applied for quantification but it is highly sensitive to operating conditions and to the factor used to convert BPCA yields into PyC amounts. The less condensed fraction of chars is expected to be completely decomposed by the strong oxidative attack, whereas the most refractory, graphitized forms of PyC may resist digestion preceding BPCA analysis. Further, BPCAs derive from C in larger aromatic clusters, which are not major constituents in plant-derived PyC and coals up to the bituminous maturation stage.

Spectroscopic techniques such as X-ray diffraction, FTIR or NMR spectroscopy are non-destructive but are hampered by the difficulty of unbiasedly distinguishing signals from PyC from those of natural SOM. To overcome this problem, one seeks to identify specific spectroscopic patterns that are typical for PyC and enable its identification and quantification within the heterogeneous mixture of SOM.

Degradation pattern of PyC in soils

For a long time, the paradigm that PyC represents an inert and environmentally recalcitrant form of organic carbon was well accepted. Residence times of thousands of years were generally assumed. Table 1 gives a short overview about mean residence times of PyC in soils (Knicker, 2011).

However, this view has changed over the past few decades with the knowledge that PyC comprises a range of materials with different biochemical stability and degradation pathways. Biochemical stability of PyC is commonly related to the temperature during its production and the aromaticity obtained. In addition to the chemical properties of the PyC, the degradability of such residues depends on the environmental conditions under which degradation takes place. These can vary considerably. Oxidic conditions may favor degradation over anoxic conditions as PyC abundance decreases in oxidic compared with anoxic marine sediments (Masiello and Druffel, 2003). Warm and humid conditions are likely to enhance microbial degradation. Interaction with minerals on the other hand may hinder the ability of microorganisms to access and decompose PyC residues. Additional labile litter input may induce priming and stimulate PyC degradation, whereas the addition of N can decrease PyC cycling.

Under the alkaline conditions of calcaric soils or ash layers, chemically mediated physical interactions can result in fragmentation of the charcoal which enables its transport by wind and water both vertically to deeper soil regions and horizontally by erosion. Cheng et al. (2006) suggested abiotic processes such as chemisorption of oxygen on unsaturated ring carbon with its subsequent oxidation play a major role in the PyC degradation process. There is evidence for microbial degradation although not much is known about the participating organisms. Organisms that degrade lignin are most probably involved. The use of solid-state ^{13}C NMR spectroscopy to study the chemical alteration during biotic degradation of plant chars showed that within 2 months the chemical structure of the remaining PyC was strongly modified. This included the formation of O-containing functional groups and the loss of aryl C structures by mineralization and conversion to other carbon groups, mostly carboxylic C (Hilscher and Knicker, 2011). A two-step mechanism was suggested: the first step involves the formation of catechol-like structures from microbially accessible aryl rings, and second, the oxygenated O-aryl rings are cleaved, which increases the amount of carboxyl and carbonyl C. This will not only increase the accessibility of the ring system for further microbial degradation but also enhance its solubility, facilitating subsequent transport within the soil column as observed in soils of the Brazilian Cerrado (Velasco-Molina et al., 2016) or in aquatic systems (Wagner et al., 2017). Modeling the degradability of PyC in aquatic systems revealed comparable bioavailability to dissolved organic matter (DOM). It was concluded that PyC may be actively transformed within the river corridor.

Table 1 Mean residence times of selected PyC published in the literature and the method of their determination.

Material and production condition	Degradation condition and its determination	Mean residence time	Literature
Rye-grass PyC mixed with soil from the B horizon of a Cambisol (350 °C, 4 min, oxidic)	Laboratory incubation (48 days)	19 years	Hilscher et al. (2009)
Pine wood PyC mixed with material from the B horizon of a Cambisol (350 °C, 4 min, oxidic)	Laboratory incubation (48 days)	56 years	Hilscher et al. (2009)
Straw PyC mixed with sand (350 °C, 2 h, pyrolysis)	Laboratory incubation (60 days)	39 years	Hamer et al. (2004)
Oak wood PyC mixed with sand (800 °C, 20–24 h, pyrolysis)	Laboratory incubation (60 days)	76 years	Hamer et al. (2004)
Sapwood PyC mixed with sand (200–350 °C, 24–72 h, pyrolysis)	Laboratory incubation (120 days)	<2% of total C after 120 days	Baldock and Smernik (2002)
Natural PyC in soils (Savannah)	Natural soils, isotopic ratio determination	<100 years (half-life time)	Bird et al. (1999)
PyC containing light fraction	long-term field trial	21–23.4 years	Murage et al. (2007)
^{14}C -labeled rye grass PyC mixed with material from the Ap horizon of an Haplic Luvisol (13 h, 400 °C, pyrolysis)	Laboratory incubation (1181 days)	200 years	Kuzyakov et al. (2009)
^{14}C -labeled rye grass PyC (13 h, 400 °C, pyrolysis)	Extrapolation from laboratory data to natural soils	2000 years	Kuzyakov et al. (2009)
PyC in Chernozems	Natural soils, isotopic ratio determination	1160–5040 years	Schmidt et al. (2002)
PyC in sediments	Chronosequence, chemical oxidation	2400–13,900 years	Masiello and Druffel (1998)

Modified from Knicker H (2011) Pyrogenic organic matter in soil: Its origin and occurrence, its chemistry and survival in soil environments. *Quaternary International* 243(2): 251–263. <https://doi.org/10.1016/j.quaint.2011.02.037>.

Preservation of PyC in soils

Pyrogenic carbon is one of the most resistant forms of reduced organic matter and comprises a major proportion of the slow-cycling carbon pools in terrestrial systems and aquatic sediments. In some fire-affected surface soils, the typical PyC pattern of the SOM remains visible for centuries. In others, however, the pattern is lost after short recovery times either because it is masked by the input of fresh litter or by the loss of PyC due to degradation or erosion. To explain these divergent results a concept was developed to explain, at least in part, the dependence of PyC survival on environmental conditions.

According to the simplified concept shown in Fig. 4, after vegetation fires, a considerable part of the aboveground vegetation is affected or destroyed (1). The PyC formed accumulates on the soil surface (2) and enters the SOM (3) giving it the typical PyC pattern with high aromaticity and hydrophobicity. The extent of this pattern depends upon the severity of the fire and the type of vegetation cover. Increased fire intensity decreases the amount of produced charred residues. Less and finer PyC is expected to accumulate after burning of grassland than after forest fires.

Vegetation fires increase the SOM pool not only by the production of PyC but also by the augmentation of decaying fire-affected but non-combusted above-ground vegetation and roots of dead vegetation (4). Additional fresh litter input is expected from the recovering vegetation (5) which takes advantage of the liming effect and accumulation of plant-available P and N. This can lead to the situation where the amount of SOM developed from this input may exceed the PyC content leading to “masking” (I) of the former PyC pattern (6), although the PyC contents may have changed only slightly. Evidently, increased litter production goes together with an increase of the organic matter content of the soil.

If the newly produced plant biomass is constantly removed which is the case for burned agricultural soils, the PyC is selectively enriched (II) in the soil (7). Management of such soils with frequent post-harvest burning will increase PyC concentration with SOM formed from decaying roots and harvest residues.

Under the oxic conditions of the soil surface, microbial reworking of the PyC (8) or even mineralization (9) starts to leave residues with increased contents of carboxylic functional groups. These groups increase the water solubility of the PyC residues and their adsorption capacity to soil minerals. The latter (10) may increase their stability against further microbial attack, their physical removal into aggregates or their translocation via leaching with clay minerals into dark illuvial horizons with PyC accumulation (11). Alternatively, transport may occur with leaching soil water (12). Here, different conditions such as changes in the soil pH or mineralogy can support immobilization of the dissolved PyC. Oxygen depletion deeper in the soil reduces the biodegradability of transported PyC (13), and compared to sediments the residues are selectively preserved and can survive for decades and even millennia.

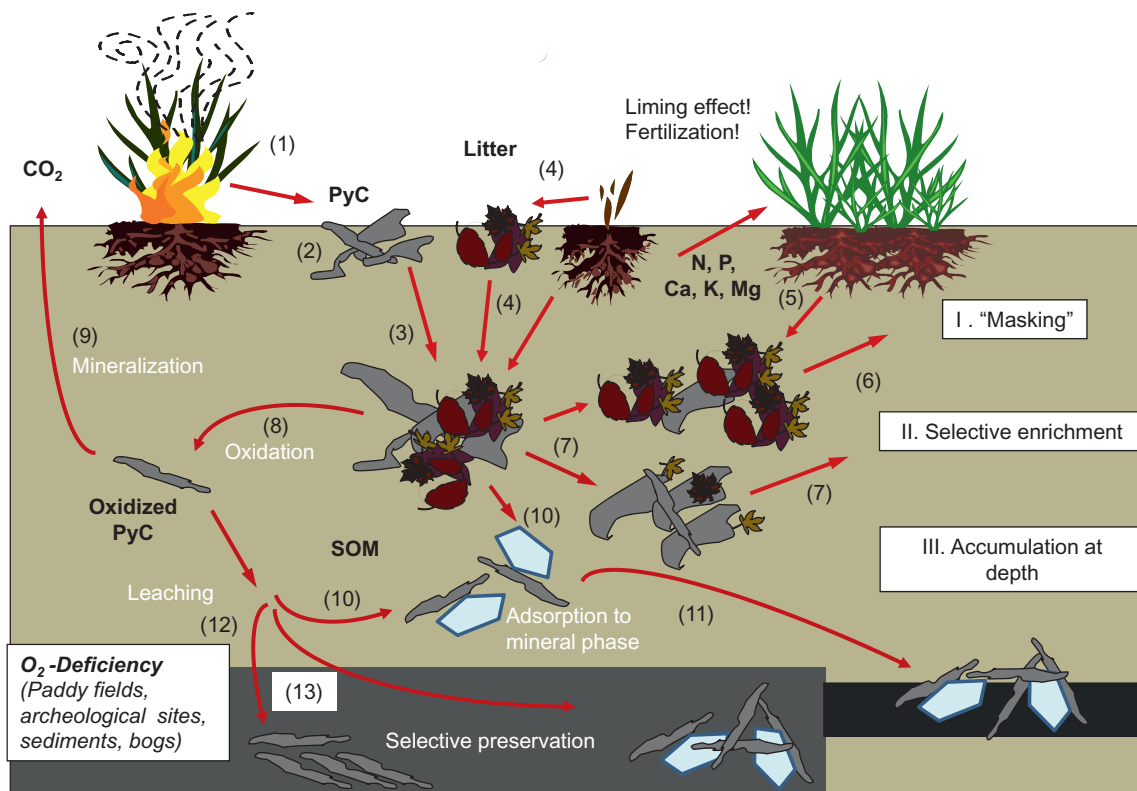


Fig. 4 Concept of the fate of pyrogenic Carbon (PyC) in different soil environments. Modified and adapted from Knicker H (2011) Pyrogenic organic matter in soil: Its origin and occurrence, its chemistry and survival in soil environments. *Quaternary International* 243(2): 251–263. <https://doi.org/10.1016/j.quaint.2011.02.037>.

Although its chemical nature differs from that of SOM, in soils, the survival of PyC in soils is determined by similar factors to those of SOM that have been unaffected by fire. Similar to fire-unaffected and humified (stabilized) SOM with residence times of several thousand years, PyC is involved in pedogenic processes. Thus, PyC input by wildfires or human activity of Neolithic and earlier times has contributed to the formation of soils with typical features (i.e., some Chernozems, black illuvial horizon in fire-prone regions, Terra Preta de Índios, soils around ancient kilns and black soils at archeological sites). Consequently, one has to be aware that using fire and PyC amendment as agricultural practices can alter soil environments and properties not only in the short term but also on a long-term scale.

Conclusion

Apart from its impact on soil properties, PyC represents an important contributor to the slow pool of SOM. As such it plays an important role as a C sink within the global elemental cycles. There is increasing awareness that although high aromaticity is certainly a common feature of most PyC, it cannot necessarily be assumed that the chemical nature of the aromatic structures of PyC with different origins is comparable. It should rather be seen as a continuum of thermally altered biomolecules. The degree of these alterations depends on the heat severity, heating conditions (oxic, pyrolysis) and nature of the feedstock. A paradigm shift occurred with respect to the biochemical recalcitrance of PyC that there are microorganisms that can attack and degrade highly aromatic structures. Consequently, the recalcitrance of PyC depends not only on its chemical structure but relies also on soil conditions. Under conditions that support microbial and fungal activity, PyC can be oxidized and eventually mineralized, whereas conditions in dry (i.e., deserts) or oxygen-depleted environments (sediments, deep soils) lead to its preservation and accumulation. Note that partial oxidation of PyC enables its interaction with minerals and its transport as dissolved pyrogenic matter either vertically within the soils or horizontally with superficial waters. Little is still known about the nature of the latter, although it represents a considerable source of PyC in aquatic systems.

This shift of views increased the awareness that there is neither a common structure nor a common behavior that applies to all PyC. This heterogeneity affects not only the predictability of global carbon cycling models that assign PyC to a single stable pool, but also the reliability of methods developed for the identification and quantification of this material by biomarkers, chemical or thermal stability. Thus, there is an urgent need to improve our understanding of the relationship between the structure of PyC and its chemical, physical or biochemical stability. In addition, its thermal resistance has to be considered in particular in environments with frequent fire events. This knowledge will not only increase our general understanding of PyC and its ecological role but will also enable better prediction of the amount and distribution of this material in soils, rivers, oceans and sediments. However, this also requires that analytical tools are developed that can separately determine the proportion and turnover rates of labile, assimilable and stable components of PyC.

References

- Baldock JA and Smernik RJ (2002) Chemical composition and bioavailability of thermally altered *Pinus resinosa* (Red pine) wood. *Organic Geochemistry* 33(9): 1093–1109. [https://doi.org/10.1016/S0146-6380\(02\)00062-1](https://doi.org/10.1016/S0146-6380(02)00062-1).
- Bird MI, et al. (1999) Stability of elemental carbon in a savanna soil. *Global Biogeochemical Cycles* 13(4): 923–932. <https://doi.org/10.1029/1999GB900067>.
- Bird MI, et al. (2015) The pyrogenic carbon cycle. *Annual Review of Earth and Planetary Sciences* 43(43): 273–298. <https://doi.org/10.1146/annurev-earth-060614-105038>.
- Campos P, et al. (2020) Chemical, physical and morphological properties of biochars produced from agricultural residues: Implications for their use as soil amendment. *Waste Management* 105: 256–267. <https://doi.org/10.1016/j.wasman.2020.02.013>.
- Certini G (2005) Effects of fire on properties of forest soils: A review. *Oecologia* 143(1): 1–10. <https://doi.org/10.1007/s00442-004-1788-8>.
- Cheng C-H, et al. (2006) Oxidation of black carbon by biotic and abiotic processes. *Organic Geochemistry* 37(11): 1477–1488. <https://doi.org/10.1016/j.orggeochem.2006.06.022>.
- Dai L, et al. (2016) Biochar: a potential route for recycling of phosphorus in agricultural residues. *GCB Bioenergy* 8(5): 852–858. <https://doi.org/10.1111/gcbb.12365>. John Wiley & Sons, Ltd.
- De La Rosa JM, et al. (2008) Determination of refractory organic matter in marine sediments by chemical oxidation, analytical pyrolysis and solid-state ¹³C nuclear magnetic resonance spectroscopy. *European Journal of Soil Science* 59(3): 430–438. <https://doi.org/10.1111/j.1365-2389.2007.00979.x>.
- EBC (2019) *European Biochar Certificate, Guidelines for a Sustainable Production of Biochar. Version 8.2E of 19th April 2019, European Biochar Foundation (EBC), Arbaz, Switzerland. Version 6.1 of 19th June 2015.* <http://www.european-biochar.org/en/download>: Arbaz, Switzerland: European Biochar Foundation (EBC). <https://doi.org/10.13140/RG.2.1.4658.7043>.
- Eckmeier E, et al. (2007) Pedogenesis of chernozems in Central Europe—A review. *Geoderma* 139(3–4): 288–299. Available at: <http://www.sciencedirect.com/science/article/B6V67-4NDDT4N-1/2/f5d75b7f1daef8fe90aac992a5aa2907>.
- Falco C, et al. (2011) Hydrothermal carbon from biomass: Structural differences between hydrothermal and pyrolyzed carbons via ¹³C solid state NMR. *Langmuir* 27(23): 14460–14471. <https://doi.org/10.1021/la202361p>. American Chemical Society.
- Freitas JCC, Bonagamba TJ, and Emmerich FG (1999) ¹³C high-resolution solid-state NMR study of peat carbonization. *Energy & Fuels* 13(1): 53–59. <https://doi.org/10.1021/ef980075c>.
- Glaser B and Amelung W (2003) Pyrogenic carbon in native grassland soils along a climosequence in North America. *Global Biogeochemical Cycles* 17(2). <https://doi.org/10.1029/2002GB002019>. p. n/a-n/a.
- Hamer U, et al. (2004) Interactive priming of black carbon and glucose mineralisation. *Organic Geochemistry* 35: 823–830.
- Hilscher A and Knicker H (2011) Degradation of grass-derived pyrogenic organic material, transport of the residues within a soil column and distribution in soil organic matter fractions during a 28 month microcosm experiment. *Organic Geochemistry* 42(1): 42–54. <https://doi.org/10.1016/j.orggeochem.2010.10.005>.
- Hilscher A, et al. (2009) Mineralisation and structural changes during the initial phase of microbial degradation of pyrogenic plant residues in soil. *Organic Geochemistry* 40(3): 332–342. <https://doi.org/10.1016/j.orggeochem.2008.12.004>.

- Hoffmann V, et al. (2019) Conductive carbon materials from the hydrothermal carbonization of vineyard residues for the application in electrochemical double-layer capacitors (EDLCs) and direct carbon fuel cells (DCFCs). *Materials* 12(10): 1703. <https://doi.org/10.3390/ma12101703>. Switzerland.
- Knicker H (2010) "Black nitrogen"—An important fraction in determining the recalcitrance of charcoal. *Organic Geochemistry* 41(9): 947–950. <https://doi.org/10.1016/j.orggeochem.2010.04.007>. Pergamon.
- Knicker H (2011) Pyrogenic organic matter in soil: Its origin and occurrence, its chemistry and survival in soil environments. *Quaternary International* 243(2): 251–263. <https://doi.org/10.1016/j.quaint.2011.02.037>.
- Knicker H, et al. (1996) 13C- and 15N-NMR spectroscopic examination of the transformation of organic nitrogen in plant biomass during thermal treatment. *Soil Biology and Biochemistry* 28(8): 1053–1060. [https://doi.org/10.1016/0038-0717\(96\)00078-8](https://doi.org/10.1016/0038-0717(96)00078-8).
- Knicker H, et al. (2008) A new conceptual model for the structural properties of char produced during vegetation fires. *Organic Geochemistry* 39(8): 935–939. <https://doi.org/10.1016/j.orggeochem.2008.03.021>.
- Knicker H, Velasco-Molina M, and Knicker M (2021) 2D solid-state HETCOR 1H-13C NMR experiments with variable cross polarization times as a tool for a better understanding of the chemistry of cellulose-based pyrochars—A tutorial. *Applied Sciences* 11(18): 8569. <https://doi.org/10.3390/app11188569>.
- Krull ES, et al. (2006) Importance of charcoal in determining the age and chemistry of organic carbon in surface soils. *Journal of Geophysical Research* 111. <https://doi.org/10.1029/2006JG000194>.
- Kuzaykov Y, et al. (2009) Black carbon decomposition and incorporation into soil microbial biomass estimated by 14C labeling. *Soil Biology and Biochemistry* 41(2): 210–219. Available at: <http://www.sciencedirect.com/science/article/B6TC7-4TVR73D-1/2/41ac7d801ec184cdf09a1937e21ad9f4>.
- Masiello CA and Druffel ERM (1998) Black carbon in deep-sea sediments. *Science* 280: 1911–1913.
- Masiello CA and Druffel ERM (2003) Organic and black carbon 13C and 14C through the Santa Monica Basin oxic-anoxic transition. *Geophysical Research Letters* 30: 34/1–34/4.
- Mocanu AM, et al. (2012) Study on the thermal behavior of casein under nitrogen and air atmosphere by means of the TG-FTIR technique. *Thermochimica Acta* 546: 120–126. <https://doi.org/10.1016/j.tca.2012.07.031>.
- Murage EW, Voroney P, and Beyaert RP (2007) Turnover of carbon in the free light fraction with and without charcoal as determined using the 13C natural abundance method. *Geoderma* 138(1–2): 133–143. Available at: <http://www.sciencedirect.com/science/article/B6V67-4MKTY01-1/2/ed4995727a88d8ffff8bf034e1171b4a>.
- Panque M, et al. (2017) Hydrothermal carbonization and pyrolysis of sewage sludges: What happen to carbon and nitrogen? *Journal of Analytical and Applied Pyrolysis* 128: 314–323. <https://doi.org/10.1016/j.jaap.2017.09.019>.
- Santin C, et al. (2016) Towards a global assessment of pyrogenic carbon from vegetation fires. *Global Change Biology* 22(1): 76–91. <https://doi.org/10.1111/gcb.12985>.
- Santin C, et al. (2016) Forest floor chemical transformations in a boreal forest fire and their correlations with temperature and heating duration. *Geoderma* 264: 71–80. <https://doi.org/10.1016/j.geoderma.2015.09.021>.
- Schmidt MWI, Skjemstad JO, and Jäger C (2002) Carbon isotope geochemistry and nanomorphology of soil black carbon: Black chernozemic soils in central Europe originate from ancient biomass burning. *Global Biogeochemical Cycles* 16: 1123–1130. <https://doi.org/10.1029/2002gb001939>.
- Sharma RK, et al. (2004) Characterization of chars from pyrolysis of lignin. *Fuel* 83: 1469–1482.
- Simpson M and Hatcher P (2004) Overestimates of black carbon in soils and sediments. *Naturwissenschaften* 91(9). <https://doi.org/10.1007/s00114-004-0550-8>.
- Skjemstad JO, et al. (2002) Charcoal carbon in U.S. agricultural soils. *Soil Science Society of America Journal* 66: 1249–1255.
- Tang M and Bacon R (1964) Carbonization of cellulose fibers—I. Low temperature pyrolysis. *Carbon* 2(3): 211–220. [https://doi.org/10.1016/0008-6223\(64\)90035-1](https://doi.org/10.1016/0008-6223(64)90035-1).
- Velasco-Molina M, et al. (2016) Biochemically altered charcoal residues as an important source of soil organic matter in subsoils of fire-affected subtropical regions. *Geoderma* 262: 62–70. <https://doi.org/10.1016/j.geoderma.2015.08.016>.
- Wagner S, Ding Y, and Jaffé R (2017) A new perspective on the apparent solubility of dissolved black carbon. *Frontiers in Earth Science* 5. <https://doi.org/10.3389/feart.2017.00075>.
- Woolf D, et al. (2010) Sustainable biochar to mitigate global climate change. *Nature Communications* 1(1): 56. <https://doi.org/10.1038/ncomms1053>.

Further reading

Role of BC in Global Carbon Cycling

- Bird MI, Wynn JG, Saiz G, Wurster CM, and McBeath A (2015) The pyrogenic carbon cycle. *Annual Review of Earth and Planetary Sciences* 43(43): 273–298. <https://doi.org/10.1146/annurev-earth-060614-105038>.
- Pingree MRA and DeLuca TH (2017) Function of wildfire-deposited pyrogenic carbon in terrestrial ecosystems. *Frontiers in Environmental Science* 5. <https://doi.org/10.3389/fev.2017.00053>.
- Ribeiro-Kumara C, Köster E, Aaltonen H, and Köster K (2020) How do forest fires affect soil greenhouse gas emissions in upland boreal forests? A review. *Environmental Research* 184: 109328. <https://doi.org/10.1016/j.envres.2020.109328>.
- Santin C, Doerr SH, Kane ES, Masiello CA, Ohlson M, Maria de la Rosa J, Preston CM, and Dittmar T (2016) Towards a global assessment of pyrogenic carbon from vegetation fires. *Global Change Biology* 22: 76–91. <https://doi.org/10.1111/gcb.12985>.

Erosion and BC

- Doerr SH, Santin C, Merino A, Belcher CM, and Baxter G (2018) Fire as a removal mechanism of pyrogenic carbon from the environment: Effects of fire and pyrogenic carbon characteristics. *Frontiers in Earth Science* 6. <https://doi.org/10.3389/feart.2018.00127>.
- Guearena DT, Lehmann J, Walter T, Enders A, Neufeldt H, Odiwour H, Biwott H, Recha J, Shepherd K, Barrios E, and Wurster C (2015) Terrestrial pyrogenic carbon export to fluvial ecosystems: Lessons learned from the White Nile watershed of East Africa. *Global Biogeochemical Cycles* 29: 1911–1928. <https://doi.org/10.1002/2015gb005095>.
- Rumpel C, Chaplot V, Planchon O, Bernadou J, Valentin C, and Mariotti A (2006) Preferential erosion of black carbon on steep slopes with slash and burn agriculture. *Catena* 65: 30–40. <https://doi.org/10.1016/j.catena.2005.09.005>.
- Wagner S, Ding Y, and Jaffé R (2017) A New perspective on the apparent solubility of dissolved Black Carbon. *Frontiers in Earth Science* 5. <https://doi.org/10.3389/feart.2017.00075>.

Chemical Nature of BC and its Persistence in Soils

- Abney RB and Berhe AA (2018) Pyrogenic carbon erosion: Implications for stock and persistence of pyrogenic carbon in soil. *Frontiers in Earth Science* 6. <https://doi.org/10.3389/feart.2018.00026>.
- Abney RB, Jin L, and Berhe AA (2019) Soil properties and combustion temperature: Controls on the decomposition rate of pyrogenic organic matter. *Catena* 182: 104127. <https://doi.org/10.1016/j.catena.2019.104127>.
- Bird MI, McBeath AV, Ascough PL, Levchenko VA, Wurster CM, Munksgaard NC, Smernik RJ, and Williams A (2017) Loss and gain of carbon during char degradation. *Soil Biology & Biochemistry* 106: 80–89. <https://doi.org/10.1016/j.soilbio.2016.12.012>.
- Bird MI, Veenendaal EM, Moyo C, Lloyd J, and Frost P (1999) Stability of elemental carbon in a savanna soil. *Global Biogeochemical Cycles* 13: 923–932. <https://doi.org/10.1029/1999GB900067>.
- Braadbaart F and Poole I (2008) Morphological, chemical and physical changes during charcoalification of wood and its relevance to archaeological contexts. *Journal of Archaeological Science* 35: 2434–2445. <https://doi.org/10.1016/j.jas.2008.03.016>.

- Brodowski S, Amelung W, Haumaier L, Abetz C, and Zech W (2005) Morphological and chemical properties of black carbon in physical soil fractions as revealed by scanning electron microscopy and energy-dispersive X-ray spectroscopy. *Geoderma* 128: 116–129. <https://doi.org/10.1016/j.geoderma.2004.12.019>.
- De la Rosa JM and Knicker H (2011) Bioavailability of N released from N-rich pyrogenic organic matter: An incubation study. *Soil Biology and Biochemistry* 43: 2368–2373. <https://doi.org/10.1016/j.soilbio.2011.08.008>.
- De la Rosa JM, Miller AZ, and Knicker H (2018) Soil-borne fungi challenge the concept of long-term biochemical recalcitrance of pyrochar. *Scientific Reports* 8: 2896. <https://doi.org/10.1038/s41598-018-21257-5>.
- De la Rosa JM, Rosado M, Paneque M, Miller AZ, and Knicker H (2018) Effects of aging under field conditions on biochar structure and composition: Implications for biochar stability in soils. *Science of the Total Environment* 613–614: 969–976. <https://doi.org/10.1016/j.scitotenv.2017.09.124>.
- Flessa H, Amelung W, Helfrich M, Wiesenberger GLB, Gleixner G, Brodowski S, Rethemeyer J, Kramer C, and Grootes PM (2008) Storage and stability of organic matter and fossil carbon in a Luvisol and Phaeozem with continuous maize cropping: A synthesis. *Journal of Plant Nutrition and Soil Science* 171: 36–51. <https://doi.org/10.1002/jpln.200700050>.
- Gibson C, Berry TD, Wang R, Spencer JA, Johnston CT, Jiang Y, Bird JA, and Filley TR (2016) Weathering of pyrogenic organic matter induces fungal oxidative enzyme response in single culture inoculation experiments. *Organic Geochemistry* 92: 32–41. <https://doi.org/10.1016/j.orggeochem.2015.12.003>.
- Gibson CD, Hatton P-J, Bird JA, Nadelhoffer K, Ward CP, Stark RE, and Filley TR (2018) Interacting controls of pyrolysis temperature and plant taxa on the degradability of PyOM in fire-prone northern temperate forest soil. *Soil Systems* 2. <https://doi.org/10.3390/soilsystems2030048>.
- Golchin A, Clarke P, Baldock JA, Higashi T, Skjemstad JO, and Oades JM (1997) The effects of vegetation and burning on the chemical composition of soil organic matter in a volcanic ash soil shown by ¹³C NMR spectroscopy. I. Whole soil and humic acid fraction. *Geoderma* 76: 155–174. [https://doi.org/10.1016/S0016-7061\(96\)00103-6](https://doi.org/10.1016/S0016-7061(96)00103-6).
- Graham EB, Song H-S, Grieger S, Garayburu-Caruso V, Stegen J, Bladon KD, and Myers-Pigg A (2022) Potential bioavailability of pyrogenic organic matter resembles natural dissolved organic matter pools. *EGU sphere* 2022: 1–22. <https://doi.org/10.5194/egusphere-2022-194>.
- Lehmann J, Gaunt J, and Rondon M (2006) Bio-char sequestration in terrestrial ecosystems—A review. *Mitigation and Adaptation Strategies for Global Change* 11: 403–427. <https://doi.org/10.1007/s11027-005-9006-5>.
- Pellegrini AFA, Harden J, Georgiou K, Hemes KS, Malhotra A, Nolan CJ, and Jackson RB (2022) Fire effects on the persistence of soil organic matter and long-term carbon storage. *Nature Geoscience* 15: 5–13. <https://doi.org/10.1038/s41561-021-00867-1>.
- Zimmerman AR (2010) Abiotic and microbial oxidation of laboratory-produced black carbon (biochar). *Environmental Science & Technology* 44: 1295–1301. <https://doi.org/10.1021/es903140c>.

Distribution of BC in Soils

- Carcaillet C and Talon B (2001) Soil carbon sequestration by Holocene fires inferred from soil charcoal in the dry French Alps. *Arctic, Antarctic, and Alpine Research* 33: 282–288. <https://doi.org/10.1080/15230430.2001.12003432>.
- Hobley E (2019) Vertical distribution of soil pyrogenic matter: A review. *Pedosphere* 29: 137–149. [https://doi.org/10.1016/S1002-0160\(19\)60795-2](https://doi.org/10.1016/S1002-0160(19)60795-2).
- López-Martín M, González-Vila FJ, and Knicker H (2018) Distribution of black carbon and black nitrogen in physical soil fractions from soils seven years after an intense forest fire and their role as C sink. *Science of the Total Environment* 637–638: 1187–1196. <https://doi.org/10.1016/j.scitotenv.2018.05.084>.
- Maie N, Parish KJ, Watanabe A, Knicker H, Benner R, Abe T, Kaiser K, and Jaffé R (2006) Chemical characteristics of dissolved organic nitrogen in an oligotrophic subtropical coastal ecosystem. *Geochimica et Cosmochimica Acta* 70: 4491–4506. <https://doi.org/10.1016/j.gca.2006.06.1554>.
- Maie N, Knicker H, Watanabe A, and Kimura M (2006) Heterocyclic N in the highly humified humic acids extracted from the subsoil of paddy fields and surface ando soils. *Organic Geochemistry* 37: 12–19. <https://doi.org/10.1016/j.orggeochem.2005.08.020>.
- Rumpel C, Kögel-Knabner I, and Bruhn F (2002) Vertical distribution, age, and chemical composition of organic carbon in two forest soils of different pedogenesis. *Organic Geochemistry* 33: 1131–1142.

BC in Archeological Soils and Ancient Kilns

- Abdelrahman H, Hofmann D, Berns AE, Meyer R, Bol R, and Borchard N (2018) Historical charcoal additions alter water extractable, particulate and bulk soil C composition and stabilization. *Journal of Plant Nutrition and Soil Science* 181: 809–817. <https://doi.org/10.1002/jpln.201800261>.
- Bird MI and Ascough PL (2012) Isotopes in pyrogenic carbon: A review. *Organic Geochemistry* 42: 1529–1539. <https://doi.org/10.1016/j.orggeochem.2010.09.005>.
- Cao ZH, Ding JL, Hu ZY, Knicker H, Kögel-Knabner I, Yang LZ, Yin R, Lin XG, Dong YH, Cao ZH, Hu ZY, Yang LZ, Yin R, Lin XG, and Dong YH (2006) Ancient paddy soils from the Neolithic age in China's Yangtze river delta. *Naturwissenschaften* 93: 232–236. <https://doi.org/10.1007/s00114-006-0083-4>.
- Downie AE, Van Zwieten L, Smernik RJ, Morris S, and Munroe PR (2011) Terra Preta Australis: Reassessing the carbon storage capacity of temperate soils. *Agriculture, Ecosystems & Environment* 140: 137–147. <https://doi.org/10.1016/j.agee.2010.11.020>.
- Glaser B, Lehmann J, and Zech W (2002) Ameliorating physical and chemical properties of highly weathered soils in the tropics with charcoal—a review. *Biological Fertility of Soils* 35: 219–230. <https://doi.org/10.1007/s00374-002-0466-4>.
- Hardy B, Leifeld J, Knicker H, Dufey JE, Deforce K, and Cornélis JT (2017) Long term change in chemical properties of preindustrial charcoal particles aged in forest and agricultural temperate soil. *Organic Geochemistry* 107: 33–45. <https://doi.org/10.1016/j.orggeochem.2017.02.008>.
- Kaal J, Martínez Cortizas A, Eckmeier E, Costa Casais M, Santos Estévez M, and Criado Boado F (2008) Holocene fire history of black colluvial soils revealed by pyrolysis-GC/MS: A case study from Campo Lameiro (NW Spain). *Journal of Archaeological Science* 35: 2133–2143. <https://doi.org/10.1016/j.jas.2008.01.013>.
- Lima HN, Schaefer CER, Mello JWW, Gilkes RJ, and Ker JC (2002) Pedogenesis and pre-Colombian land use of “Terra Preta Anthrosols” (“Indian black earth”) of Western Amazonia. *Geoderma* 110: 1–17. [https://doi.org/10.1016/S0016-7061\(02\)00141-6](https://doi.org/10.1016/S0016-7061(02)00141-6).
- Mastrolonardo G and Certini G (2020) Viewpoint. Charcoal hearth soils should be better accounted for by the WRB and the soil taxonomy. *Journal of Plant Nutrition and Soil Science* 183: 633–636. <https://doi.org/10.1016/j.catena.2011.08.006>.

Impact of BC on Soil Biota and Properties

- Alexis M, Rasse DP, Rumpel C, Bardoux G, Péchot N, Schmalzer P, Drake B, and Mariotti A (2007) Fire impact on C and N losses and charcoal production in a scrub oak ecosystem. *Biogeochemistry* 82: 201–216. <https://doi.org/10.1007/s10533-006-9063-1>.
- Certini G (2005) Effects of fire on properties of forest soils: A review. *Oecologia* 143: 1–10. <https://doi.org/10.1007/s00442-004-1788-8>.
- Certini G, Moya D, Lucas-Borja ME, and Mastrolonardo G (2021) The impact of fire on soil-dwelling biota: A review. *Forest Ecology and Management* 488: 118989. <https://doi.org/10.1016/j.foreco.2021.118989>.
- Christel W, Bruun S, Magid J, and Jensen LS (2014) Phosphorus availability from the solid fraction of pig slurry is altered by composting or thermal treatment. *Bioresource Technology* 169: 543–551. <https://doi.org/10.1016/j.biortech.2014.07.030>.
- Czimczik CI, Preston CM, Schmidt MWI, Werner RA, and Schulze ED (2002) Effects of charring on mass, organic carbon, and stable carbon isotope composition of wood. *Organic Geochemistry* 33: 1207–1223. [https://doi.org/10.1016/S0146-6380\(02\)00137-7](https://doi.org/10.1016/S0146-6380(02)00137-7).
- Dymov AA, Abakumov EV, Bezkorovaynaya IN, Prokushkin AS, Kuzyakov YV, and Milanovsky EY (2018) Impact of forest fire on soil properties (review). *Theoretical and Applied Ecology* 2018: 13–23. <https://doi.org/10.25750/1995-4301-2018-4-013-023>.
- Maestrini B, Abiven S, Singh N, Bird J, Torn MS, and Schmidt MWI (2014) Carbon losses from pyrolysed and original wood in a forest soil under natural and increased N deposition. *Biogeosciences* 11: 5199–5213. <https://doi.org/10.5194/bg-11-5199-2014>.
- Schaller J, Tischer A, Struyf E, Bremer M, Belmonte DU, and Potthast K (2015) Fire enhances phosphorus availability in topsoils depending on binding properties. *Ecology* 96: 1598–1606. <https://doi.org/10.1890/14-1311.1>.

Chemistry and BC

- Le Brech Y, Delmotte L, Raya J, Brosse N, Gadiou R, and Dufour A (2015) High resolution solid state 2D NMR analysis of biomass and biochar. *Analytical Chemistry* 87: 843–847. <https://doi.org/10.1021/ac504237c>.
- Kawamoto H (2017) Lignin pyrolysis reactions. *Journal of Wood Science* 63: 117–132. <https://doi.org/10.1007/s10086-016-1606-z>.
- Leifeld J (2007) Thermal stability of black carbon characterised by oxidative differential scanning calorimetry. *Organic Geochemistry* 38: 112–127. <https://doi.org/10.1016/j.orggeochem.2006.08.004>.
- McBeath AV, Smernik RJ, Krull ES, and Lehmann J (2014) The influence of feedstock and production temperature on biochar carbon chemistry: A solid-state ^{13}C NMR study. *Biomass and Bioenergy* 60: 121–129.
- Sergides CA, Jassim JA, Chughtai AR, and Smith DM (1987) The structure of hexane soot, part II, ozonation studies. *Applied Spectroscopy* 41: 482–492.
- Sharma RK, Chan WG, Wang J, Waymack BE, Wooten JB, Seeman JI, and Hajaligol MR (2004) On the role of peptides in the pyrolysis of amino acids. *Journal of Analytical and Applied Pyrolysis* 72: 153–163. <https://doi.org/10.1016/j.jaap.2004.03.009>.
- Xin S, Yang H, Chen Y, Yang M, Chen L, Wang X, and Chen H (2015) Chemical structure evolution of char during the pyrolysis of cellulose. *Journal of Analytical and Applied Pyrolysis* 116: 263–271. <https://doi.org/10.1016/j.jaap.2015.09.002>.
- Yang H, Yan R, Chen H, Lee DH, and Zheng C (2007) Characteristics of hemicellulose, cellulose and lignin pyrolysis. *Fuel* 86: 1781–1788. <https://doi.org/10.1016/j.fuel.2006.12.013>.

Quantification of BC in Soils

- Brodowski S, Rodionov A, Haumaier L, Glaser B, and Amelung W (2005) Revised black carbon assessment using benzene polycarboxylic acids. *Organic Geochemistry* 36: 1299–1310. <https://doi.org/10.1016/j.orggeochem.2005.03.011>.
- Chatterjee S, Santos F, Aiven S, Itin B, Stark RE, and Bird JA (2012) Elucidating the chemical structure of pyrogenic organic matter by combining magnetic resonance, mid-infrared spectroscopy and mass spectrometry. *Organic Geochemistry* 51: 35–44. <https://doi.org/10.1016/j.orggeochem.2012.07.006>.
- de la Rosa Arranz JM, González-Vila FJ, González-Pérez JA, Martín GA, Hernández Z, Martín ML, and Knicker H (2013) How useful is the mid-infrared spectroscopy in the assessment of black carbon in soils. *Flamma* 4: 147–151.
- Elias VO, Simoneit BRT, Cordeiro RC, and Turcq B (2001) Evaluating levoglucosan as an indicator of biomass burning in Carajás, Amazonia: A comparison to the charcoal record. *Geochimica et Cosmochimica Acta* 65: 267–272. [https://doi.org/10.1016/S0016-7037\(00\)00522-6](https://doi.org/10.1016/S0016-7037(00)00522-6).
- Glaser B, Haumaier L, Guggenberger G, and Zech W (1998) Black carbon in soils: The use of benzenecarboxylic acids as specific markers. *Organic Geochemistry* 29: 811–819. [https://doi.org/10.1016/S0146-6380\(98\)00194-6](https://doi.org/10.1016/S0146-6380(98)00194-6).
- Hardy B, Borchard N, and Leifeld J (2022) Identification of thermal signature and quantification of charcoal in soil using differential scanning calorimetry and benzene polycarboxylic acid (BPCA) markers. *The Soil* 8: 451–466. <https://doi.org/10.5194/soil-8-451-2022>.
- Kaal J and Rumpel C (2009) Can pyrolysis-GC/MS be used to estimate the degree of thermal alteration of black carbon? *Organic Geochemistry* 40: 1179–1187. <https://doi.org/10.1016/j.orggeochem.2009.09.002>.
- Knicker H, Wiesmeier M, and Dick DP (2008) A simplified method for the quantification of pyrogenic organic matter in grassland soils via chemical oxidation. *Geoderma* 147: 69–74. <https://doi.org/10.1016/j.geoderma.2008.07.008>.
- Kuhlbusch TAJ (1995) Method for determining black carbon in residues of vegetation fires. *Environmental Science and Technology* 29: 2695–2702. <https://doi.org/10.1021/es00010a034>.
- López-Capel E, Sohi S, Gaunt JL, and Manning DAC (2005) Use of thermogravimetry-differential scanning calorimetry to characterize modelable soil organic matter fractions. *American Journal of Soil Science Society* 69: 3192–3198. <https://doi.org/10.2136/sssaj2005.0136a>.
- Roth PJ, Lehnendorff E, Brodowski S, Bornemann L, Sanchez-García L, Gustafsson Ö, and Amelung W (2012) Differentiation of charcoal, soot and diagenetic carbon in soil: Method comparison and perspectives. *Organic Geochemistry* 46: 66–75. <https://doi.org/10.1016/j.orggeochem.2012.01.012>.
- Sanders EB, Goldsmith AI, and Seeman JI (2003) A model that distinguishes the pyrolysis of D-glucose, D-fructose, and sucrose from that of cellulose. Application to the understanding of cigarette smoke formation. *Journal of Analytical and Applied Pyrolysis* 66: 29–50. [https://doi.org/10.1016/S0165-2370\(02\)00104-3](https://doi.org/10.1016/S0165-2370(02)00104-3).
- Schmidt MWI, Skjemstad JO, Czimczik CI, Glaser B, Prentice KM, Gélinas Y, and Kuhlbusch TAJ (2001) Comparative analysis of black carbon in soils. *Global Biogeochemical Cycles* 15: 163–167. <https://doi.org/10.1029/2000GB001284>.
- Skjemstad JO, Taylor JA, and Smernik RJ (1999) Estimation of charcoal (char) in soils. *Communications in Soil Science and Plant Analysis* 30: 2283–2298. <https://doi.org/10.1080/00103629909370372>.
- Song J, Peng P, and Huang W (2002) Black carbon and kerogen in soils and sediments. 1. Quantification and characterization. *Environmental Science and Technology* 36: 3960–3967. <https://doi.org/10.1021/es025502m>.
- Wolbach WS and Anders E (1989) Elemental carbon in sediments: Determination and isotopic analysis in the presence of kerogen. *Geochimica et Cosmochimica Acta* 53: 1637–1647. [https://doi.org/10.1016/0016-7037\(89\)90245-7](https://doi.org/10.1016/0016-7037(89)90245-7).

Fire effects on soil

Stefan H Doerr^a, Cristina Santín^{b,c}, and Jorge Mataix-Solera^d, ^aDepartment of Geography, Swansea University, Swansea, United Kingdom; ^bResearch Institute of Biodiversity (IMIB), Mieres, Spain; ^cDepartment of Biosciences, Swansea University, Swansea, United Kingdom; ^dDepartment of Agrochemistry and Environment, Environmental Soil Science Group, Miguel Hernández University, Alicante, Spain

© 2023 Elsevier Ltd. All rights reserved.

Introduction	448
Fire, combustion and soil heating	449
Direct effects of fire	450
Soil chemical and mineralogical characteristics	450
Soil organic matter content and soil carbon stocks	450
Soil organic matter composition	451
Nutrients	452
pH	452
Soil mineralogical composition	452
Soil physical characteristics	452
Water repellency and other hydrological characteristics	452
Aggregate stability and bulk density	453
Soil biological characteristics	454
Microorganisms	454
Invertebrates	454
Soil seed bank	454
Indirect effects of fire	454
Conclusions	456
References	456

Abstract

Fires affect many landscapes whether ignited by lightning or humans. They can affect soil physical, chemical and biological characteristics directly through heating and combustion, or indirectly through modified biological, pedological and hydrological processes after the fire. Direct effects include changes in organic matter, nutrients, biota, water repellency, aggregate stability, and an increased susceptibility to erosion. Indirect effects arise from incorporation of ash and from changes in vegetation cover, the water balance, organic matter inputs or protection from erosion. These effects can last from months to years and, in fire-adapted ecosystems, fire can be considered a key factor in soil formation.

Key points

- Fire effects on soils are highly variable, from negligible to long-lasting. Fire can be a key soil forming factor.
- Fire impacts on soils are both direct, from heating, and indirect, by, for example, ash addition or changes in vegetation cover.
- Fire impacts on soils include changes in their physical (e.g., stability, water repellency), chemical (e.g., pH, organic matter and nutrient contents) and biological (e.g., microbial activity and composition) properties.

Introduction

Vegetation fires (i.e., landscape fires including wildfires and fires ignited for land management purposes) currently burn over 3–5 million km² every year, which corresponds to approximately 4% of Earth's vegetated land surface—roughly the size of Europe (Jones et al., 2019). They have occurred naturally for over 420 million years and humans have used fires to modify the landscape for over 125,000 years (Santín and Doerr, 2016; MacDonald et al., 2021; Roebroeks et al., 2021). Fire can modify physical, chemical and biological characteristics of soils (Fig. 1). This can occur directly through soil heating from the burning of the vegetation, litter and woody debris, and the combustion of soil organic matter itself, or indirectly through modified landscape processes in the post-fire period. Indirect effects mostly arise from the change in vegetation cover, for example, alterations in the water balance, amount and type of organic matter inputs or protection from erosion. Depending on the severity¹ of the fire, these effects can last for

¹Fire severity is a broad term used to describe the degree to which an area has been altered directly by a fire (i.e., burn severity). It most commonly concerns the degree of vegetation destruction. To describe the effect of fire on soil, the term soil burn severity is commonly used (Keeley, 2009).

as little as a few months to many years and can be permanent where temperatures reached result in the modification of soil minerals or where fire leads to land cover changes. In fire-adapted ecosystems fire can be considered as a key factor in soil formation (Certini, 2014).

Fire, combustion and soil heating

Whilst the temperature at the hottest part of a flame in a vegetation fire can exceed $>1000^{\circ}\text{C}$, most of the thermal energy released by the flaming combustion process is directed upwards with only a small proportion available to heat the soil. Thus, the intensity of a fire (i.e., the rate of energy release per meter of flaming front expressed in kW m^{-1}), which is broadly related to the length of the flames, is not necessarily related to the impacts of a fire on soil. For example, a fast-moving intense fire with several meter-long flames rapidly consuming most of the aboveground vegetation may only lead to limited heating of the soil (Stoof et al., 2013), whereas a slow-moving low-intensity fire with small flames (i.e., a long residence time) consuming ground fuels (such as a thick litter layer) can lead to much greater soil heating (Hartford and Frandsen, 1992) (Fig. 1).

In addition, soils are poor conductors of heat due to their porous structure. Dry soils have a thermal conductivity typically below $0.5 \text{ W m}^{-1} \text{ K}^{-1}$, which is low compared to, e.g., amorphous quartz ($>1 \text{ W m}^{-1} \text{ K}^{-1}$) or steel ($>16 \text{ W m}^{-1} \text{ K}^{-1}$). Thermal conductivity increases with soil moisture; however, a substantial amount of energy is used to evaporate this moisture and only once this is driven off can the soil temperature exceed $>100^{\circ}\text{C}$ (Campbell et al., 1994). During flaming fires that consume litter and aboveground vegetation, there is therefore typically a steep temperature gradient in the soil profile with soil a few centimeters below the surface experiencing little if any heating. The highest temperatures and deepest heat penetration typically occur under wood pile burns or smoldering logs. Broad temperature ranges typically reached in mineral surface soils for different types of fires are given in Fig. 1. In relating the impact of a specific temperature on soil characteristics, it is important to consider that it is not only the maximum temperature reached that is important, but also the duration of heating (i.e., the residence time). Temperature thresholds at which changes in soils occur have often been established using prolonged heating (tens of minutes) in laboratory settings, whereas the heat peaks in wildfires often last only seconds (Fig. 2). In general, slow-moving wildfires where large amounts of ground fuel are entirely consumed and deep heat penetration occurs often have the greatest impact on soil. In contrast, fast moving wildfires in low fuel vegetation (e.g., grasses) or prescribed burns carried out at higher soil and fuel moisture contents and where the ground fuel may be only partially consumed have little if any direct effect on the soil (Meira-Castro et al., 2015; Santín and Doerr, 2016).

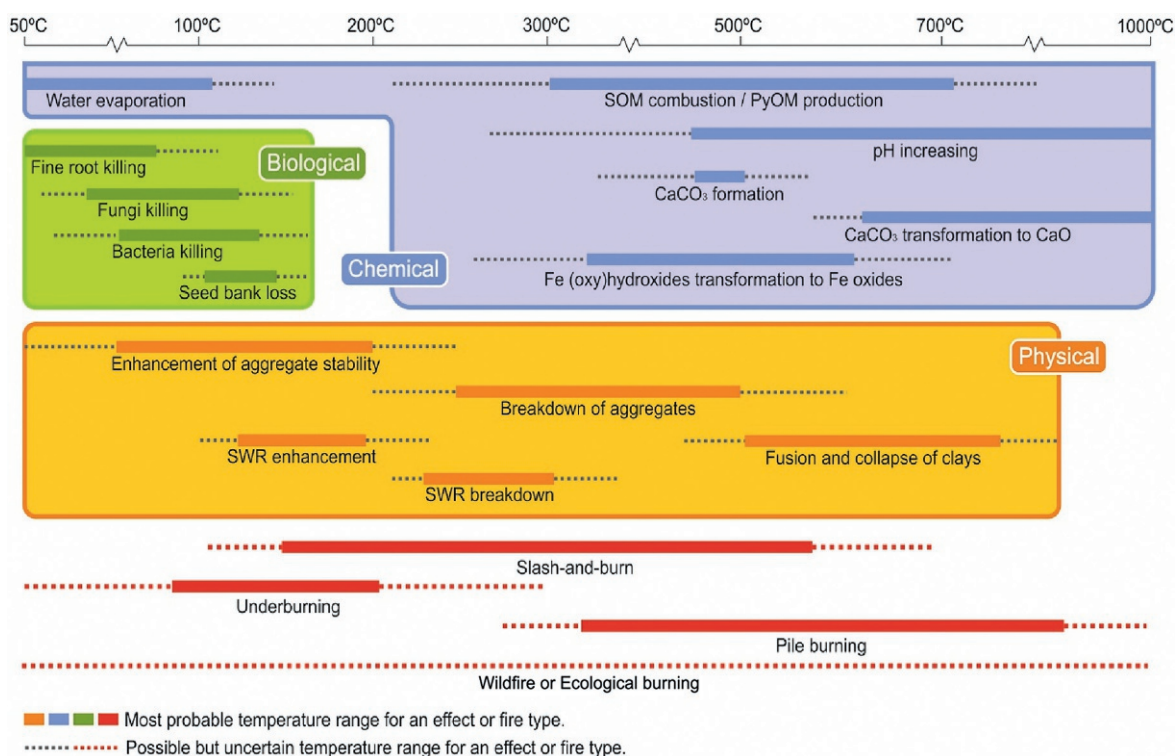


Fig. 1 Effects of fire on soil biological, chemical and physical characteristics and the temperature range at which these effects commonly occur. Red lines indicate the typical temperature ranges near the soil surface for different types of fires. Specific temperatures and associated soil effects will depend on the characteristics of each fire and soil. PyOM: Pyrogenic organic matter; SOM: Soil organic matter; SWR: Soil water repellency. Modified from Santín C and Doerr SH (2016) Fire effects on soils: The human dimension. *Proceedings of the Royal Society B: Biological Sciences* 371: 20150171.



Fig. 2 Soil before and after a moderate- to high intensity prescribed fire in a eucalypt forest (NSW, Australia). The litter and Oa layers (A) were burned by the fire and transformed into ash (B), whereas the Ah horizon experienced little direct impact. Maximum temperatures recorded using thermocouple sensors: litter: 577 °C; Oa: 400 °C; Ah: 59 °C. Photos S. Doerr; From Santín C, Otero XL, Doerr SH and Chafer CJ (2018) Impact of a moderate/high-severity prescribed eucalypt forest fire on soil phosphorous stocks and partitioning. *Science of the Total Environment* 621: 1103–1114.

Where the soil organic matter (SOM) content is sufficiently high (i.e., peats and other organic-rich soils), combustion of the soil itself may occur. This typically involves slow, smoldering combustion at temperatures in the region of ~600 °C that can consume soil layers to many decimeters depth and lead to extensive heating of any adjacent mineral soil (Fig. 3). Smoldering peat fires are very difficult to extinguish and in extreme cases can continue throughout a winter season even under snow cover (also known as overwintering or ‘zombie’ fires; Scholten et al., 2021).

Direct effects of fire

Soil chemical and mineralogical characteristics

Apart from the loss of water from evaporation, the key effects from heating and combustion concern SOM, nutrients, pH and, especially at higher temperatures, soil mineralogical transformations. Given that SOM is typically most abundant in the uppermost layers of soil where the heating during fire is also at its maximum, SOM is especially affected by fire.

Soil organic matter content and soil carbon stocks

Reported effects of fire on SOM stocks vary from negligible to substantial decreases or increases. This is a reflection of the wide range of fires studied. For mineral soils, low-severity fires have little impact on the SOM pool and associated carbon (C) stocks. In more severe fires where the soil can experience substantial heating (>200–300 °C), decreases in soil organic carbon (SOC) have been observed. For example, for a moderate-high severity fire in eucalyptus forest over sandy soils near Sydney, complete combustion of the Oa horizon (maximum temperature: ~400 °C), was observed, whereas the C concentration in the Ah horizon (maximum temperature: ~50 °C) was not altered (Fig. 2; Santín et al., 2018). A study examining SOC at 0–5-cm soil depth after different severities of soil burn in pine forest and shrublands in NW Spain showed that the lowest severities of soil burn did not translate into a significant decrease of SOC, but the highest resulted in SOC losses of up to 86% (Vega et al., 2013). Fire can also increase SOC stocks, however, due to inputs of organic constituents from ash² and charred organic matter, which can become incorporated into the soil (Fig. 3; Francos et al., 2018).

Fires in areas with organic types of soils, such as peatlands, often consume only the aboveground vegetation such as grasses, mosses or shrubs due to the high water content of the soil horizons. Following drought, and in some cases exacerbated by anthropogenic deforestation or drainage, however, organic soils can be very susceptible to combustion. For example, the recurring

²Here we refer to ash as the particulate post-fire residue consisting of mineral materials and charred organic components (Bodí et al., 2014).

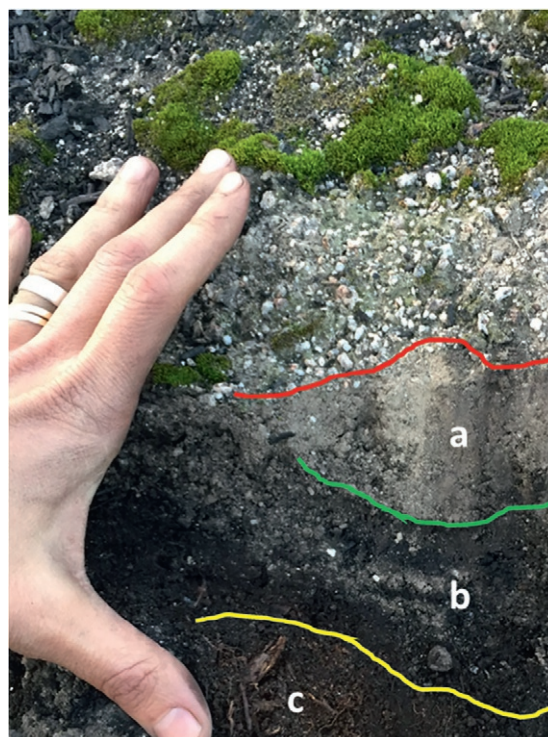


Fig. 3 Umbric Leptosol developed over granite 1 year after fire (Galicia, NW Spain). Three distinct layers can be identified: (a) has suffered near-complete SOC combustion with loss of soil structure; (b) a substantial amount of SOC remains in the form of PyC materials formed by pyrolysis; (c) SOC-rich layer of largely unaffected by fire (photo C. Santín).

peatland fires in parts of Indonesia, which have affected large areas of degraded peatland soils over the last two decades, have led to substantial losses of peat, large emissions of carbon to the atmosphere (e.g., in 2015 more than the annual fossil fuel C emissions of Japan in 2013; Field et al., 2016) and regional air quality problems. The increasing extent, frequency and severity of boreal peatland fires observed in recent years associated with climate change are thought to have caused an overall decline in peatland C stocks with SOC losses from peat combustion exceeding the rate of peat accumulation in some regions in the globally important SOC store (Walker et al., 2019).

Soil organic matter composition

The combustion of SOM leads to mineralization and loss of organic components through volatilization and oxidation. When soil temperatures exceed 200–300 °C, the formation of new pyrogenic organic compounds in zones of incomplete combustion also begins to occur, in addition to any input of charred organic constituents from the partially combusted vegetation, litter and overlying organic layers (Santín and Doerr, 2016). Charring of SOM reduces oxygen-rich functional groups through dehydration and dehydrogenation reaction. In the lipid fraction, a reduction of the chain lengths is observed by thermal cracking. Cellulose undergoes aromatization and, at higher temperatures, polycyclic aromatic hydrocarbons begin to form (Knicker, 2007).

The term ‘pyrogenic carbon’ (PyC) is widely used to refer to the whole range of pyrogenic organic materials from partially charred vegetal biomass through to charcoal and soot (Santín et al., 2016). PyC is a broad continuum of C materials ranging from biolabile depolymerization products to highly degradation-resistant condensation products. Charring induces mainly aromatization and condensation reactions, which result in increased resistance against degradation of PyC. Whilst not inert (as assumed in the past) it has become clear that PyC exhibits a resistance to biochemical degradation in the environment that is orders of magnitude higher than its unburned precursors (Santín et al., 2016). It therefore plays an important role in the overall carbon budget of fires and the global carbon cycle. Its annual production is equivalent to ~12% annual C emissions from landscape fires (Jones et al., 2019), leading to a net C accumulation in some fire affected regions (Bowring et al., 2022). Soils are one of the largest pools of PyC in the Earth system; it accounts for ~14% of the global SOC stock (Reisser et al., 2016).

The formation of PyC is a medium- to long-term effect of fire on soil chemical properties, whereas, the overall effects of fire on SOM quantity and composition are many and highly site specific. There is therefore little that can be predicted in terms of universal effects and trends of fire on SOM, except that where the soil system is resilient to fire overall, a recovery towards a pre-disturbance equilibrium SOM quantity and quality can be expected.

Nutrients

The direct impact of heating and combustion processes in the soil is usually the volatilization of nutrients such as organic nitrogen (N) and phosphorus (P), which, in principle, leads to an overall loss. The loss of the different elements depends on their volatilization temperatures, with N and P being lost at lower temperatures ($\sim 500^\circ\text{C}$ and 700°C , respectively) than elements such as Ca (calcium), Mn (manganese) or Mg (magnesium) that need very high temperatures ($>1000^\circ\text{C}$) and, therefore, usually remain in the burned residues (Bodí et al., 2014). The overall direct effect of burning, however, is often an increase in nutrients such as phosphorus or calcium in surface soils (Pellegrini and Jackson, 2020). This is mostly due to leaching from, or the direct incorporation of, nutrient rich ash from vegetation deposited at the soil surface (Caon et al., 2014). During heating, conversion of organic to inorganic forms occurs; the latter being more accessible to plants and microorganisms. This fertilizing effect has long been taken advantage of in agriculture, such as slash and burn in shifting cultivation in tropical forest regions or stubble burning following harvesting in croplands (Maass, 1995). This fertilization effect, however, is rather transient and postfire erosion processes can lead to a substantial net loss of nutrients in the longer term (Caon et al., 2014; Santín et al., 2018).

pH

Following fire, an increase in soil pH is commonly observed, especially in acidic soils immediately after burning, with up to 3 units having been reported (Ulery and Graham, 1993). This is associated with the denaturation of organic acids and the release of bases from the combustion of organic matter (Certini, 2005). Much of the increase in pH near the soil surface can arise from the ash layer that is commonly deposited on the soil surface during fire. Vegetation fire ash is rich in alkaline cations (e.g., Ca^{2+} , Mg^{2+} , K^+ , potassium) and can exceed pH 9–11 in solution (Bodí et al., 2014). Its effect on soil, together with any increases in pH from in situ changes in the soil, are relatively short-lived when soils are exposed to prolonged or repeated periods of rainfall.

Soil mineralogical composition

Most fires have little impact on soil mineralogy except where sustained heating can make soil temperature exceeds $300\text{--}500^\circ\text{C}$, for example under smoldering logs or slash burns. Dehydroxylation, converting iron oxyhydroxides such as goethite, $[\text{FeO}(\text{OH})]$ to iron oxides such as hematite $[\text{Fe}_2\text{O}_3]$ (Fig. 1), is often evident by the red color of strongly heated soils (Certini, 2005). In the presence of organic matter or some other reducing agent, magnetic minerals such as magnetite and maghemite can also be formed in Fe-rich soil, which are used in archaeological studies for evidence of fire (Oldfield, 1999). Carbonates such as calcium carbonate (CaCO_3 —a key mineral in calcareous soils), dissociate to oxides between 580°C and 1100°C depending on the carbonate compound in question. Perhaps more important, however, is the addition of CaCO_3 to the soil from ash produced in severe fires (Bodí et al., 2014).

Soil physical characteristics

Water repellency and other hydrological characteristics

A common observation following fire is the presence of soil water repellency (SWR; also termed water repellence; hydrophobicity; Fig. 4). It is caused by hydrophobic compounds such as aliphatic or amphiphilic organic molecules covering soil particle surfaces or accumulating in soil pore spaces. These compounds can be produced in situ in the soil by, for example, fungal or microbial activity



Fig. 4 Strong water repellency following a severe wildfire in a eucalyptus forest near Sydney 2020, as indicating by a water droplet resting on sandy soil (photo S. Doerr).

or added from plant litter (Doerr et al., 2000) or hydrophobic ash (Bodí et al., 2014). The non-polar areas of these molecules have a low affinity to water, leading to an overall reduction in surface tension of the soil to below that of water ($72.75 \times 10^{-3} \text{ N m}^{-1}$ at 20°C). This means that water is more attracted to itself than to the surface it is in contact with, resulting in no spreading and infiltration of water (Fig. 4). Depending on the severity and persistence of SWR, this can cause a delay in infiltration or non-wetting of the most water repellent area of a soil. Many soils exhibit natural levels of water repellency at low soil moisture contents, however, fire can modify and often enhance its presence by (i) removing the litter layer and making it more apparent, (ii) drying the soil, (iii) adding hydrophobic compounds to the soil released during the burning of organic matter, and (iv) in situ enhancement of SWR by heating and modifying organic matter in the soil (Doerr et al., 2000). The temperatures and associated parameters at which changes in SWR occur have been the subject of numerous studies. Common observations are that SWR begins to be enhanced below 100°C , can be strongly enhanced up to $\sim 250\text{--}270^\circ\text{C}$, but is generally destroyed at higher temperatures. Oxidation of the causal compounds seems to be key in the elimination of SWR as the threshold destruction temperature can be increased to $>500^\circ\text{C}$ in the absence of oxygen. In fires that result in a high severity of soil burn, heating can be sufficient to eliminate SWR in the top few millimeters or centimeters of soil, which are then underlain by a layer of enhanced water repellency (Shakesby and Doerr, 2006).

Soil water repellency can not only lead to reduced or impeded infiltration, but its often highly variable spatial nature also leads to the formation of preferential flow pathways and enhanced spatial variation in soil moisture. The net effect at the landscape scale is lower soil water holding capacity, and enhanced overland flow, preferential flow and associated stream runoff (Fig. 5; Doerr et al., 2000).

The loss of organic matter, a breakdown of aggregates or the clogging of pores from ash ingress or the formation of an ash crust (Bodí et al., 2014) can also lead to a reduction in soil infiltration capacity and enhanced overland flow irrespective of any SWR. However, the permeability and water holding capacity of ash is generally greater than that of the underlying soil and, thus, the ash layer typically reduces overland flow (Bodí et al., 2014).

Aggregate stability and bulk density

Moderate soil heating can enhance aggregate stability associated with an increase in particle surface water repellency. At higher temperatures, effects can vary from disaggregation to strong aggregation at high temperatures if recrystallization of minerals (e.g., Fe and Al (aluminium) oxyhydroxides) occurs (Fig. 1; Mataix-Solera et al., 2011). Bulk density can increase following fire where aggregate collapse and the ingress of fine ash leads to a reduction of pore space (Certini, 2005).



Fig. 5 Overland flow transporting burned soil, ash and charred debris during intense rain following a wildfire in eucalypt forest in the Victorian Alps, south-east Australia in 2003. From Shakesby RA and Doerr SH (2006) Wildfire as a hydrological and geomorphological agent. *Earth-Science Reviews* 74: 269–307.

Soil biological characteristics

Microorganisms

The direct, short-term effects are often a decrease and modification of the soil microbial community through heat induced mortality. This begins to occur above 60–80 °C, a temperature that can easily be reached in the top few centimeters of soil especially when the soil is moist, which facilitates heat transmission for temperatures to reach 100 °C. In general, fungi, especially ectomycorrhizal types, appear to be more heat sensitive than bacteria (Pellegrini and Jackson, 2020). In ecosystems where fires are rare they may cause major shifts in, for example, fungal community that last years or even decades. In ecosystems adapted to frequent fires, fire may also be necessary to maintain fungal communities (Fox et al., 2022). In extreme cases, soil heating can lead to sterilization of the topsoil layer, however, unless heat penetration is very deep and extensive (e.g., under large pile burns), rapid recolonization from subsurface layers or adjacent areas that experienced less soil heating will normally occur (Certini et al., 2021). Immediately after fire, alterations in soil pH, nutrients and SOM (including the formation of toxic compounds) can affect the activity and composition of soil microbiota. For example, an increase in heterotrophic bacteria activity associated with an increase in soluble carbon and nutrients is commonly observed in the post-fire period. This initial increase is generally followed by a decrease once the easily accessible carbon and nitrogen forms have been depleted. An increased ratio of bacteria to fungi is thought to be associated with the increase in pH from ash. Although an overall reduction in soil microbial activity in the top few centimeters is commonly found in the first few years after fire, modification of the microbial and fungal soil biome with recolonization from adjacent soil areas and aerial deposition of microbes and spores may be the more important medium-term fire impact (Certini et al., 2021). The proliferation of edible pyrophilic fungi such as those of the *Morchella* genus after fire can be of substantial local economic importance (Larson et al., 2016).

Invertebrates

Apart from some flying insects, most soil dwelling invertebrates easily succumb to fire due to their relative immobility and sensitivity to heat. For example, the lethal temperature for many soil arthropods can be as low as 40 °C resulting in very high mortality for those present in the litter and topsoil layers. Mortality is reduced for earthworms and other burrowing animals when they escape the flames or the heat shock by migrating to deeper layers. However, rather than direct mortality, the indirect effects of fire on food resources may have a greater impact overall, causing the invertebrate population to decline. This can especially affect herbivore species such as litter-dwelling detritivores following extensive high-severity fires. On the other hand, other types of organism may even increase in abundance after the fire due to, for example, stimulation of emergence of adults from increases in soil temperature or greater availability of dead wood (Molinas-González et al., 2019).

Where invertebrate mortality has been high, recovery proceeds gradually, firstly by highly mobile species, from nearby unburned areas or patches of low fire severity. Thus, for example, recolonization of carabids occurs sooner than that of soil dwelling macrofauna, while arachnids, earthworms, molluscs and myriapods are rather slow colonizers (Certini et al., 2021). Complete recovery of the invertebrate community after fire may take years, or even decades, depending on fire severity, its patchiness and shape and extent of the burned area. It is important to consider, however, that fire also creates new habitats and, beyond the immediate area burned, also increases habitat diversity over larger spatial and temporal dimensions. Burned areas can attract pioneer species that take advantage of dead vegetation or new herbaceous growth, and some invertebrates, such as ants or termites have dwellings that can be resistant to fire. Ultimately most invertebrates (and other animal species) recover best from a single fire with patchy severity, followed by a sufficiently long recovery period within the normal interval for fire return of a given ecosystem (Certini et al., 2021).

Soil seed bank

The soil seed bank is an important source of resilience in ecosystems to perturbations. The direct effect of fire on the seedbank is closely related with the temperature reached in the soil. Thus, fire behavior, such as its intensity and residence time together with soil thermal properties, determine the amount of soil heating. The associated effects on the seedbank tend to be spatially variable within a burned area. The ability of seeds to respond to fire also differs between plant species and with the distribution of seeds in the soil profile (Kwiatkowska-Falinska et al., 2014) given that soil temperatures decrease markedly with depth during burning (Bradstock and Auld, 1995). In some fire-adapted species, seed germination is also influenced by the indirect effects of fire like smoke or charred wood (Read et al., 2000). Thus, if temperatures in the soil profile remain below that at which damage to seeds occurs, fire can promote their germination through heating, smoke stimulus, the change in soil nutrient levels, and by providing greater access to water and light due to temporary removal of plant cover (Snyman, 2005). Much research has been conducted on the effects of fires on soil seed banks in fire-adapted ecosystems, especially in Mediterranean regions. Other ecosystems where fires have been rare, have seen limited attention. This is an important research gap given that fire is becoming more prevalent in some of these.

Indirect effects of fire

Some of the most severe and lasting effects of fire on soil can be indirect and often occur gradually over the post-fire period; an important consideration when evaluating any effects on the soil of a given fire (Santín and Doerr, 2016). Major impacts can arise from the loss of vegetation and litter cover, and, where relevant, the immediate or delayed mortality of plants and roots. This has

implications for soil biogeochemical processes, soil thermal regime and runoff and erosion. The reduction or complete loss of soil cover, combined in some cases with reduction in soil stability and enhanced water repellency, allow erosive rainfall to affect the soil surface directly, leading to splash erosion, enhanced particle entrainment, and accelerated overland flow and erosion (Fig. 5). These processes can also trigger large-scale gully erosion and mass movement events. Especially in steep terrain, substantial soil- and associated carbon and nutrient losses can occur. Published values of soil losses range between 0.1 and 41 Mg ha⁻¹ per year after moderate to severe fires compared to 0.003–0.1 Mg ha⁻¹ in comparable long unburned terrain (Shakesby and Doerr, 2006). The ‘window of disturbance’ during which the hydrogeomorphological response remains enhanced will vary with fire severity, terrain and rates of vegetation recovery and can last from a few weeks to many years. Given that surface soil holds the greatest amount of soil organic matter, nutrients and microorganisms, fire-triggered erosion of surface soil is often viewed as ‘soil destruction’, but it is important to remember that wildfires and soil erosion are often natural processes. While on-site losses may be detrimental, the redistribution of organic- and nutrient-rich sediment often promotes the formation of fertile soils in depositional areas, in riparian zones and floodplains, both within and well beyond a given burned area. Only material that is deposited in lacustrine or marine sediments can therefore be considered as being removed from the pedosphere in the longer term (Santín and Doerr, 2016).

Of relevance are also the inputs of new material to the soil triggered by fire in addition to sediment redistribution discussed above. For example, some of the immediate effects of the deposition of wildfire ash have already been highlighted under the direct effects of fire. While some ash is derived directly from charred topsoil, much of it typically originates from the burned aboveground biomass with production values up to 150 Mg ha⁻¹ having been reported. Some of the ash will be redistributed by wind or water erosion, but some will become incorporated into the soil via infiltrating water or bioturbation, adding nutrients and carbon gradually to the soil. Further gradual inputs of organic materials also occur in the form of decaying unburned vegetation killed by the fire, and the gradual incorporation of PyC deposited on the soil surface. The potential for PyC (or its human-made ‘cousin’: biochar) to increase the fertility of soils has been the focus of much research since the discovery of the fertile Terra preta soils in Amazonia (Santín and Doerr, 2016) and it is the basis of biochar technology (Jeffery et al., 2011).

A further process leading to the addition to soil is fire-induced breakdown of parent material (i.e., rock and regolith) into smaller fragments. The high temperature gradient in exposed rock surfaces caused by fire leads to spalling, the flaking of typically mm–cm thick small rock fragments (Fig. 6). This weathering process has, for example, been estimated to be similar or orders of magnitude more effective over repeated fire cycles than other weathering processes generating sediment (Buckman et al., 2021). In addition, especially where fires cause high tree mortality (stand replacing fires), the accelerated uprooting of trees brings regolith or deep soil to the surface and exposes it to accelerated weathering. Thus, while the accelerated losses of soil and associated fertility in the post-fire period are often of concern in addition to the direct impact of fire, the inputs of new material triggered by fire also need to be considered.



Fig. 6 Fire-induced spalling on rock surface, and rock fragments derived from it on the ground, after the unprecedented 2019–2020 Australian bushfire season (photo S. Doerr).

Finally, during the post-fire vegetation recovery process, changes in vegetation composition, for example, from pioneer to climax species composition, are accompanied by changes in micro-climate, hydrology and biogeochemical processes, which in turn affect soil processes. Where alterations in fire regimes or human-driven deforestation and degradation fires occur, these can result in long-term or even permanent change in the vegetation cover, which will also have implications for soil processes.

Conclusions

Fire is a common perturbation in many vegetated regions and can have profound effects on soil physical, chemical and biological process. These arise from direct impacts through heating and combustion during the fire and indirect processes in the postfire period such as loss of vegetation cover and accelerated erosion. The effect of these processes may diverge. For example, the addition of nutrients from ash in the short-term may be more than offset on steep slopes by nutrient losses from accelerated erosion in the post-fire period. In most cases, wildfires or management burns of low severity will have little lasting impact on soil, whereas severe fires, especially when they occur more frequently than is the norm for a given ecosystem, can lead to very detrimental and long-lasting effects on the soil. The soil degradational impacts of fire are therefore often a focus after severe fires, however, it is important to consider that fire has been a recurring land surface process for over 420 million years affecting pedogenesis. On balance, and excluding areas where the occurrence of extremely severe or frequent fires prevent medium-term recovery or de novo formation of soil, fire has probably been an important factor of soil formation in many fire-affected regions around of the globe.

References

- Bodi MB, Martin D, Balfour VN, Santín C, Doerr SH, Pereira P, Cerdà A, and Mataix-Solera J (2014) Wildland fire ash: Production, composition and eco-hydro-geomorphic effects. *Earth-Science Reviews* 30: 103–127.
- Bowring SPK, Jones MW, Clais P, Guenet B, and Abiven S (2022) Pyrogenic carbon decomposition critical to resolving fire's role in the Earth system. *Nature Geoscience* 15: 135–142.
- Bradstock RA and Auld TD (1995) Soil temperatures during experimental bushfires in relation to fire intensity: Consequences for legume germination and fire management in southeastern Australia. *Journal of Applied Ecology* 32: 76–84.
- Buckman S, Morris RH, and Bourman RP (2021) Fire-induced rock spalling as a mechanism of weathering responsible for flared slope and inselberg development. *Nature Communications* 12: 2150.
- Campbell G, Jungbauer JJ, Bidlake W, and Hungerford R (1994) Predicting the effect of temperature on soil thermal conductivity. *Soil Science* 158: 307–313.
- Caon L, Vallejo VR, Ritsema CJ, and Geissen V (2014) Effects of wildfire on soil nutrients in Mediterranean ecosystems. *Earth-Science Reviews* 139: 47–58.
- Certini G (2005) Effects of fire on properties of forest soils: A review. *Oecologia* 143: 1–10.
- Certini G (2014) Fire as a soil-forming factor. *Ambio* 43: 191–195.
- Certini G, Moya D, Lucas-Borja ME, and Mastrodonato G (2021) The impact of fire on soil-dwelling biota: A review. *Forest Ecology and Management* 488: 118989.
- Doerr SH, Shakesby RA, and Walsh RPD (2000) Soil water repellency: Its causes, characteristics and hydro-geomorphological significance. *Earth-Science Reviews* 51: 33–65.
- Field RD, Van Der Werf GR, Fanin T, Fetzer EJ, Fuller R, Jethva H, Levy R, Livesey NJ, Luo M, Torres O, and Worden HM (2016) Indonesian fire activity and smoke pollution in 2015 show persistent nonlinear sensitivity to El Niño-induced drought. *Proceedings of the National Academy of Sciences of the United States of America* 113(33): 9204. doi: 10.1073/pnas.1524888113.
- Fox S, Sikes BA, Brown SP, Cripps CL, Glassman SI, Hughes K, Semenova-Nelsen T, and Jumpponen A (2022) Fire as a driver of fungal diversity—A synthesis of current knowledge. *Mycologia*. <https://doi.org/10.1080/00275514.2021.2024422>.
- Francois M, Úbeda X, Pereira P, and Alcañiz M (2018) Long-term impact of wildfire on soils exposed to different fire severities. A case study in Cadiretes Massif (NE Iberian Peninsula). *Science of the Total Environment* 615: 664–671.
- Hartford R and Frandsen W (1992) When It's Hot, It's Hot... Or Maybe It's Not! (Surface Flaming May Not Portend Extensive Soil Heating). *International Journal of Wildland Fire* 2: 139–144.
- Jeffery S, Verheijen FG, van der Velde M, and Bastos AC (2011) A quantitative review of the effects of biochar application to soils on crop productivity using meta-analysis. *Agriculture, Ecosystems & Environment* 144: 175–187.
- Jones MW, Santín C, van der Werf GR, and Doerr SH (2019) Global fire emissions buffered by the production of recalcitrant pyrogenic carbon. *Nature Geoscience* 12: 742–747.
- Keeley JE (2009) Fire intensity, fire severity and burn severity: A brief review and suggested usage. *International Journal of Wildland Fire* 18: 116–126.
- Knicker H (2007) How does fire affect the nature and stability of soil organic nitrogen and carbon? A review. *Biogeochemistry* 85: 91–118.
- Kwiatkowska-Falinska A, Jankowska-Błaszczuk M, and Jaroszewicz B (2014) Post-fire changes of soil seed banks in the early successional stage of pine forest. *Polish Journal of Ecology* 62(3): 455–466.
- Larson JA, Cansler AC, Cowdery SG, Hiebert D, Furniss TJ, Swanson ME, and Lutz JA (2016) Post-fire morel (*Morchella*) mushroom abundance, spatial structure, and harvest sustainability. *Forest Ecology and Management* 377: 16–25.
- Maass J (1995) Conversion of tropical dry forest to pasture and agriculture. In: Bullock SH, Mooney HA, and Medina E (eds.) *Seasonally Dry Tropical Forests*, pp. 399–422. Cambridge: Cambridge University Press.
- MacDonald K, Scherjon F, van Veen E, Vaesen K, and Roebroeks W (2021) Middle Pleistocene fire use: The first signal of widespread cultural diffusion in human evolution. *Proceedings of the National Academy of Sciences* 118: e2101108118.
- Mataix-Solera J, Cerdà A, Arcenegui V, Jordán A, and Zavala LM (2011) Fire effects on soil aggregation: A review. *Earth-Science Reviews* 109: 44–60.
- Meira-Castro A, Shakesby RA, Marques JE, Doerr SH, Meixedo JP, Teixeira J, and Chaminé HI (2015) Effects of prescribed fire on surface soil in a *Pinus pinaster* plantation, northern Portugal. *Environmental Earth Sciences* 73(6): 3011–3018.
- Molinas-González CR, Castro J, González-Megías A, and Leverkus AB (2019) Effects of post-fire deadwood management on soil macroarthropod communities. *Forests* 10: 1046.
- Oldfield F (1999) The rock magnetic identification of magnetic mineral and grain size assemblages. In: Walden J, Oldfield F, Smith J (Eds.) *Environmental Magnetism: A Practical Guide*. Quaternary Research Association Technical Guide 6.
- Pellegrini AF and Jackson RB (2020) The long and short of it: A review of the timescales of how fire affects soils using the pulse-press framework. *Advances in Ecological Research* 62: 147–171.
- Read TR, Bellairs SM, Mulligan DR, and Lamb D (2000) Smoke and heat effects on soil seed bank germination for the re-establishment of a native forest community in New South Wales. *Austral Ecology* 25(1): 48–57.

- Reisser M, Purves RS, Schmidt MWI, and Abiven S (2016) Pyrogenic carbon in soils: A literature-based inventory and a global estimation of its content in soil organic carbon and stocks. *Frontiers in Earth Science* 4: 1–14.
- Roebroeks W, MacDonald K, Scherjon, et al. (2021) Landscape modification by Last Interglacial Neanderthals. *Science Advances* 7: eabj5567.
- Santín C and Doerr SH (2016) Fire effects on soils: The human dimension. *Proceedings of the Royal Society B: Biological Sciences* 371: 20150171.
- Santín C, Doerr SH, Kane ES, Maseillo CA, Ohlson M, de la Rosa JM, Preston CM, and Dittmar T (2016) Towards a global assessment of pyrogenic carbon from vegetation fires. *Global Change Biology* 22: 76–91. <https://doi.org/10.1111/gbc.12985>.
- Santín C, Otero XL, Doerr SH, and Chafer CJ (2018) Impact of a moderate/high-severity prescribed eucalypt forest fire on soil phosphorous stocks and partitioning. *Science of the Total Environment* 621: 1103–1114.
- Scholten RC, Jandt R, Miller EA, Rogers BM, and Veraverbeke S (2021) Overwintering fires in boreal forests. *Nature* 593: 399–404.
- Shakesby RA and Doerr SH (2006) Wildfire as a hydrological and geomorphological agent. *Earth-Science Reviews* 74: 269–307.
- Snyman HA (2005) The effect of fire on the soil seed bank of a semi-arid grassland in South Africa. *South African Journal of Botany* 71: 53–60.
- Stoof CR, Moore D, Fernandes PM, Stoorvogel JJ, Fernandes RES, Ferreira AJD, and Ritsema CJ (2013) Hot fire, cool soil. *Geophysical Research Letters* 40: 1534–1539.
- Ulery AL and Graham RC (1993) Forest fire effects on soil color and texture. *Soil Science Society of America Journal* 57: 135–140.
- Vega JA, Fontúrbel T, Merino A, Fernández C, Ferreiro A, and Jiménez E (2013) Testing the ability of visual indicators of soil burn severity to reflect changes in soil chemical and microbial properties in pine forests and shrubland. *Plant and Soil* 369: 73–91.
- Walker XJ, Baltzer JL, Cumming SG, Day NJ, Ebert C, Goetz S, Johnstone JF, Potter S, Rogers BM, Schuur EAG, Turetsky MR, and Mack MC (2019) Increasing wildfires threaten historic carbon sink of boreal forest soils. *Nature* 572(7770): 520–523.

Interfacial reactions of microorganisms with minerals and organic matter

Rota Wagai^a, Puu-Tai Yang^a, and Klaus Kaiser^b, ^aInstitute for Agro-Environmental Science, National Agriculture and Food Research Organization, Tsukuba, Ibaraki, Japan; ^bSoil Science and Soil Protection, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	458
Microbe attachment at mineral interfaces	459
Microbial cell surface structure and chemistry	459
Cell surface structure	459
Extracellular compounds and extracellular polymeric substances (EPS)	460
Functional group chemistry of cell surfaces	461
Cell-to-mineral attachment mechanisms	461
Coulombic interaction between cell and mineral	461
Non-coulombic interaction between cell and mineral	462
Interplay of multiple binding mechanisms at the cell–mineral interface	462
Forces involved in a single bacterium on a mineral surface	462
Role of EPS at the cell–mineral interface	463
Consequences of microbial–mineral interfacial reactions	463
Microbially-mediated weathering and formation of minerals	464
Physical aspects	464
Chemical aspects	464
Element cycling	465
Contribution to soil aggregate formation	466
Outlook	467
References	468

Abstract

Microbial–mineral interfacial reactions are crucial in the development and functioning of soil. We present basic mechanisms by which microorganisms and the polymers they secrete attach to mineral surfaces, followed by discussion on three fundamental consequences of these reactions: weathering (including secondary mineral formation), element cycling (including organic matter turnover and accumulation), and formation of physical structure in soil. We also point out potentially important questions arising from the microbe–organic matter–mineral interactions.

Key points

- Microbe–mineral interfacial reactions result from an interplay of microbial cells and mineral surface through a range of bonding mechanisms.
- Microbial interface reactions drive important soil processes (e.g., weathering, elemental cycling, and soil structure formation), provide nutrients and more extensive and heterogeneous habitats to support microbial and plant life.
- Therefore, microbial interface reactions are essential components of pedogenesis and soil functioning.

Introduction

Microorganisms (microbes) and the organic matter (OM) they process, synthesize, and reside in, are often in contact with, or in the vicinity of, mineral and rock surfaces in soils. Microbe–OM–mineral interactions are ubiquitous on the Earth's surface, including marine sediments and even aerosols, and their history is presumably as old as the origin of life. As microbes are typical first colonizers in most terrestrial settings, soil genesis is largely induced and driven by microbe–OM–mineral interactions. As a result, many soil processes and functions, such as nutrient cycling, water holding and filtration, structural formation, and carbon (C) storage, are directly or indirectly linked to the reactions taking place at microbe–OM–mineral interfaces (Fig. 1). In this chapter, we focus primarily on direct mineral–microbe interactions, since other aspects of the mineral–OM–microbe continuum (e.g., OM decomposition, organic–mineral interactions, rhizosphere processes) are covered in other chapters.

First, we summarize and explain the fundamental factors of microbe–mineral interfacial reactions, starting from microbial cell surface characteristics to the bonding mechanisms. We then discuss the consequences of mineral–OM–microbial interfacial reactions by focusing on microbially-mediated weathering, elemental cycling, and soil aggregate formation.

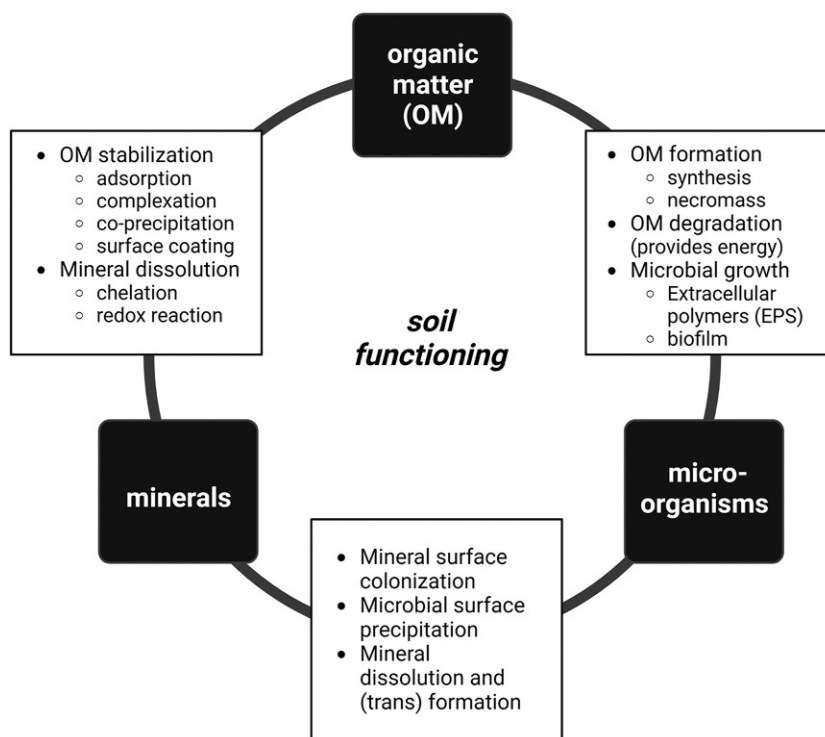


Fig. 1 Examples of mineral–organic matter–microbe interactions that primarily take place at the solid–aqueous interface. Solid phases in soil systems can be minerals, insoluble biomolecules (plant and microbial detritus and metabolites), and living and dead organisms.

Microbe attachment at mineral interfaces

Microbes can directly attach or adhere to mineral surfaces or indirectly via the extracellular compounds they secrete (Wingender et al., 1999). A range of binding mechanisms facilitates the attachment of microbes to mineral surfaces. We use the term ‘adhesion’ when multiple bonding forces render cell attachment hardly reversible. We provide basic information on microbial cell surface structure, including extracellular polymers that often surround cells, and the surface charge characteristics of those materials. Then we discuss the different types of forces involved in cell–mineral attachment. With charges arising from the functional groups on cell surfaces, coulombic interactions control the attachment of cells to charged mineral surfaces. Non-coulombic interactions may additionally come into play, especially on uncharged surfaces. Finally, we discuss the consequences of multiple forces being involved simultaneously or in sequence during surface interactions of cells and minerals.

Microbial cell surface structure and chemistry

Cell surface structure

Cell surface structure varies strongly among bacteria (Gram-positive vs. Gram-negative bacteria) and between bacteria and fungi, although all these organisms having plasma membranes (composed of lipid bilayers) underneath their cell walls. The cell wall of Gram-positive bacteria (Fig. 2) is characterized by a thick layer (15–80 nm) of peptidoglycan, a porous cross-linked polymer consisting of a glycan backbone (a repeat polymer of two amino sugars) and peptide cross-linkages, giving it a semi-rigid structure. Fibers embedded in the peptidoglycan layer and sometimes extending away from the cell surface are composed of teichoic acids, glycopolymers having phosphodiester-linked polyol repeat units. Their main function is to provide flexibility and attract cations due to their strong negative charges. In contrast, the cell wall of Gram-negative bacteria consists of an outer membrane (made up of lipids, proteins, lipopolysaccharides) above a thin layer (1–2 nm) of peptidoglycan (Fig. 2). Lipopolysaccharides, composed of lipid-A (phosphorylated glucosamine disaccharide) and polysaccharide are a major source of negative charge in Gram-negative bacterial cell walls and can extend their branches away from the cell surface for tens of nanometers. In addition to cell walls, some bacteria also have appendages that enable active surface attachment, including flagella, fimbriae, and pili. While flagella are primarily for swimming and attachment, fimbriae and pili play crucial roles in surface attachment. These thin hairlike external protein appendages are 1 to 20 μm in length and often occur in large numbers. At the end of each fimbria are special proteins, so-called adhesins. These surface structures have the potential to mediate bacterial attachment directly to solids.

The fungal cell wall is mainly composed of polysaccharides, proteins, and lipids, and occasionally contains melanin, polyphosphate, or inorganic acids (Fig. 3). The major polysaccharides are glucans (β -1,3- (30–80%), β -1,6-, and α -1,3-), chitin (1–15%),

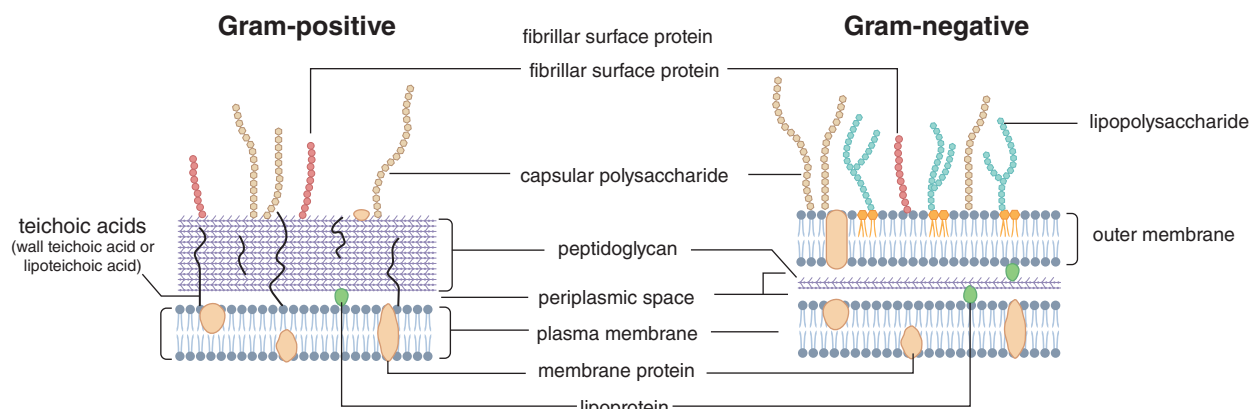


Fig. 2 Structure of bacterial cell wall. Plasma membrane, peptidoglycan, and proteins present in both Gram-positive and Gram-negative bacteria. Lipoprotein covalently binds with peptidoglycan on one side, and embeds the other polar side into the outer membrane, strengthening the integrity. Fibrillar surface protein (e.g. fibrillar adhesins) and capsular polysaccharide may also be present on the cell wall, controlling the cell attachment to solid surfaces. Modified from Chorover J (2022) 5. Microbe–biomolecule–Mineral interfacial reactions. In: Yang Y, Keiluweit M, Senesi N, and Xing B (eds.) *Multi-Scale Biogeochemical Processes in Soil Ecosystems: Critical Reactions and Resilience to Climate Changes*. pp. 117–140. John Wiley & Sons, Inc.

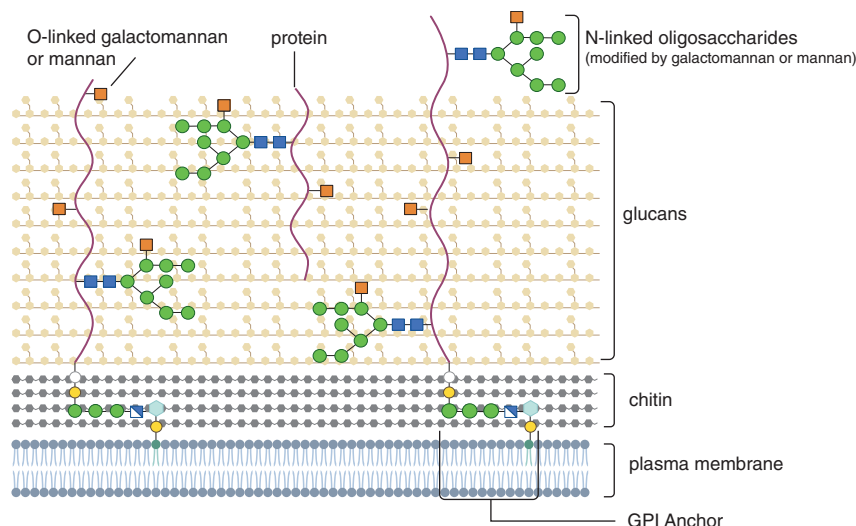


Fig. 3 Structure of fungal cell wall. Polysaccharides (i.e., glucans, chitin, mannan, galactomannan) are the major components of fungal cell walls. The structure is strengthened by protein and or other components such as pigments, lipids, or polysaccharides. Proteins contain O-linked galactomannans or mannans, or N-linked oligosaccharides. These proteins can covalently link to the cell wall either without or with a glycosylphosphatidylinositol (GPI) anchor. Over half of the cell wall proteins are GPI-anchored proteins. Modified from Free SJ (2013) Chapter Two - Fungal cell wall organization and biosynthesis. In: Friedmann T, Dunlap JC, and Goodwin SF (eds.) *Advances in Genetics*. pp. 33–82. Academic Press.

chitosan, mannans, and galactomannans (Free, 2013). Proteins are relatively more abundant in the outer part of the fungal cell wall, while carbohydrates occur primarily in the inner part (Fig. 3). The hyphal cell wall contains S-glucan, hydrophobins, and melanins (Feofilova, 2010). Structurally, glucans, chitin, and glycoproteins form three-dimensional linkages within the cell wall. Some of the glycoproteins are anchored to the cell membrane through glycosylphosphatidylinositol, GPI (Fig. 3). Hydrophobins are small, cysteine-rich proteins, secreted only from the aerial hyphae of filamentous fungi. Because of their amphiphilic property (Bromley et al., 2015), hydrophobins mediate attachment to hydrophobic surfaces (Feofilova, 2010). The composition of the functional groups on cell walls, including proteins (or amino acids) and some polysaccharides (mannans and galactomannans), determines the attachment to minerals.

Extracellular compounds and extracellular polymeric substances (EPS)

Microbes secrete various compounds, including organic acids and polymers such as extracellular enzymes. Perhaps most important for cell–mineral interface interactions are extracellular polymeric substances (EPS), which often cover microbial cell surfaces. The EPS mainly consist of polysaccharides and proteins, with minor amounts of enzymes, lipids, and nucleic acids (Flemming and Wingender, 2010).

The EPS promote cell-to-cell attachment and adhesion to mineral surfaces. In addition, EPS can protect microbes from predators and abiotic stress (i.e., extreme pH, drought, toxic chemicals, oxidative damage), and mediate enzyme activity, nutrient availability, and electron transport (Op De Beeck et al., 2021; Kaur et al., 2022). The role of EPS in soil processes, especially soil aggregation, has been studied extensively (see Section “Contribution to soil aggregate formation”). The EPS in the environment are grouped into three chemically and functionally different forms: capsular EPS on the surface of bacteria, biofilm EPS as the matrix of biofilms, and free EPS in the aqueous phase (Wingender et al., 1999).

Production of EPS is widespread across wide taxonomic groups of bacteria and fungi. The large investment of energy, C, and nutrients (in particular nitrogen (N)) for its synthesis suggests an evolutionary advantage of EPS production (Chenu and Cosentino, 2011). There are several reviews on the functions of EPS synthesized by bacteria (Tourney and Ngwenya, 2014) and fungi (Op De Beeck et al., 2021). The production and chemical composition of EPS varies with the microbial type (functions) and growth stage, together with environmental conditions (i.e., pH, temperature, water, and oxygen availability) (Kaur et al., 2022). The EPS could also provide a gradient of pH or ionic strength, controlling the diffusion of compounds both into and out of cells. With similar molecular composition, fungal EPS is expected to have similar physicochemical properties and, therefore, play similar roles as bacterial EPS for microbial life (Op De Beeck et al., 2021).

Functional group chemistry of cell surfaces

The surfaces of microorganisms feature carboxyl, phosphoryl, sulfuryl, amine, and hydroxyl groups that derive mainly from polysaccharides, proteins, lipids, and pigments in the cell wall (Chorover, 2022). These functional groups can either dissociate or become protonated in response to the changes in local pH conditions. Dissociation constants (pKa) of the functional groups on Gram-positive and Gram-negative bacteria, measured by potentiometric titration, are similar: 2–6 (carboxyl), 0.2–2.91 and 5.65–7.20 (pKa₁ and pKa₂ of phosphoryl), 9.6–10.8 (phenolic hydroxyl), and 9.1–10.6 (amine) (Hong and Brown, 2006). Fewer data are available for the functional groups on the fungal cell wall, but pKa ranges are likely to be comparable. Three types of dissociated protons were identified for *Rhizopus arrhizus* by Naja et al. (1999). At pH <4.5, released protons derived from phosphoryl or carboxyl groups. Between pH 4.5 and 7, carboxyl groups played a major part in proton dissociation. At pH >8, phenolic and amine groups contributed to the acidity. The EPS have similar functional groups as cell walls, with carboxyl, phosphate, amine, and hydroxyl groups dominating, and smaller contributions of sulfate, amide, and phosphodiester groups (Tourney and Ngwenya, 2014) (and the references therein). Therefore, the acidity of EPS is similar to that of bacterial and fungal cells, and generally ranges between 3 and 11 (Tourney and Ngwenya, 2014).

Carboxyl or phosphoryl groups dissociate and become negatively charged under acidic to neutral conditions. Amine and hydroxyl groups dissociate only under alkaline conditions, and thus, are positively- and non-charged, respectively, at neutral and acidic pH. Despite the wide variation in pKa values, soil microbial cells usually have a net negative charge within the common soil pH range (<8) because of the much larger abundances of carboxyl and phosphate groups than amine groups (Chorover, 2022).

Functional groups with varying pKa are heterogeneously distributed on cell walls and within EPS. Thus, despite the overall net negative charge of cells, positively and negatively charged groups may co-occur, allowing for electrostatic interactions with both negatively and positively charged ions and surfaces. The heterogeneity of surface charge can decrease electrostatic repulsion at the local scale and support irreversible particle attachment.

Cell-to-mineral attachment mechanisms

Coulombic interaction between cell and mineral

The charges arising from the functional groups on cell surfaces can attract solutes and facilitate attachment to solids with opposing charges. The two major types of coulombic interaction between cells and minerals are reversible electrostatic attraction and less reversible surface complexation.

Electrostatic interaction between microbe and mineral is controlled by the charges of cells and minerals, moderated by the solution pH. As mentioned, the net surface charge of microorganisms is usually negative within the common soil pH range (<8) due to the dominance of acidic groups (see Section “Functional group chemistry of cell surfaces”). The surface charge of minerals is dictated by their point of zero charges (PZC). The average PZC of common minerals in soil is 8.9 for boehmite, 9.9 for gibbsite, 7.6–9.1 for hematite, 8.1–8.6 for goethite, 7.1 for lepidocrocite, 7.9–8.1 for ferrihydrite, 4.9 for manganese dioxide (MnO₂), and 3.0 for quartz (Kosmulski, 2009). Consequently, within the common soil pH range (<8), iron (Fe) and aluminum (Al) hydrous oxides are mainly positively charged, while manganese (Mn) oxides and silicates are negatively charged. Therefore, microbes are likely to bind electrostatically to Fe or Al hydrous oxides via their negatively-charged functional surface groups (carboxyl, phosphoryl, sulfuryl, phenol), and to Mn oxides and silicates via amino groups.

Cation bridging is another type of coulombic interaction. Multivalent cations (e.g., Ca²⁺, Mg²⁺, Fe³⁺, Al³⁺, Si⁴⁺) can act as bridges between negatively charged minerals and negatively charged cell surfaces or EPS, thus overcoming the electrostatic repulsion between the two constituents. Both direct electrostatic interactions and cation bridging are reversible and sensitive to the changes in solution chemistry, such as pH and ionic strength.

Surface complexation can occur between specific functional groups and certain mineral surface sites. As both electrostatic attraction and covalent bonds are involved, the reaction results in much stronger, and thus, less reversible attachment. For instance, initial bacterial attachment to Fe oxides involves inner-sphere complexation of bacterial phosphate groups at the surfaces of Fe oxides (Parikh and Chorover, 2006). Bacterial EPS, containing both polysaccharide-rich and protein-rich components, often exhibit

strong metal complexation capabilities, which also makes them prone to forming strong surface complexes on Al and Fe hydrous oxides (Chorover, 2022). In addition, EPS can co-precipitate with metals when their solution concentrations are high enough, which may also promote cell–mineral interactions.

Non-coulombic interaction between cell and mineral

Uncharged organic surfaces can associate with uncharged mineral surfaces by non-coulombic interactions (e.g., dipole interactions, i.e., due to van der Waals and London forces, or hydrogen (H) bonds). The EPS contain large uncharged components, such as non-ionic polysaccharides and proteins. Uncharged portions are also present on microbial surfaces, including, teichoic acids, and lipopolysaccharides (see Section “Cell surface structure”). These interactions may be mediated via water molecules, which are dipoles, forming so-called water bridges. Thus, non-coulombic interaction can play a role in cell and or EPS attachment to mineral surfaces. Short-distance forces may come into play when cells bind through other mechanisms, they then add to the strength of adhesion.

Hydrophobic interaction can take place at cell/EPS–mineral interfaces. While the majority of metal oxides and clay minerals are hydrophilic, hydrophobicity differs among minerals. The relative hydrophobicity of some common soil minerals is, in decreasing order, birnessite, montmorillonite, kaolinite, quartz, goethite, and mica (Hong et al., 2012). Hematite is moderately hydrophobic, while goethite and aluminum hydroxide are hydrophilic. Other factors, such as particle size and crystallinity and environmental conditions, can also affect the hydrophobicity of minerals. On the other hand, the hydrophobicity of microbial cells depends on the relative abundance of hydrophobic and hydrophilic surface structures. Lipids and proteins on microbial cell walls tend to be relatively hydrophobic; polysaccharides are hydrophilic. For EPS, hydrated polysaccharides contribute to hydrophilicity, and DNA, proteins, and lipopolysaccharides tend to be relatively hydrophobic (More et al., 2014). The methyl- and acetyl-functional groups on polysaccharides are also attributed to the hydrophobicity of EPS (Flemming and Wingender, 2010). Cells or excreted polymers can bind on particles, and in that, can change their surface charge or hydrophobicity. The binding of intact bacterial cells or cell fragments enhances the hydrophobicity of minerals such as hematite (Poorni and Natarajan, 2013). The EPS tend to decrease the hydrophobicity of mineral surfaces, whereas extracellular proteins increase the hydrophobicity (Poorni and Natarajan, 2013). Environmental conditions, such as ionic strength and pH, affect the electrostatic repulsion, functional group protonation, and hydrophobicity of cells, EPS, and minerals (Li et al., 2019; Tournay and Ngwenya, 2014; Guhra et al., 2022). For example, the carboxyl groups of fatty acids on the outer cell wall of *Thiobacillus ferrooxidans* (*T. ferrooxidans*) protonate under acidic conditions, which enhances hydrophobicity and cell adhesion to mineral surfaces (Solari et al., 1992).

Filamentous fungi synthesize and secrete hydrophobins (see Section “Cell surface structure”) that confer hydrophobicity to fungal surfaces in contact with air, and thus, mediate attachment of hyphae to hydrophobic surfaces. Some bacterial colonies embedded in EPS (i.e., biofilms) also exhibit hydrophobicity caused by surface-active proteins having amphiphilic structures (bacterial hydrophobins) (Bromley et al., 2015).

Interplay of multiple binding mechanisms at the cell–mineral interface

Microbe–mineral interactions usually involve multiple forces, including both long-range forces (mostly electrostatic interactions) and short-range ones (van der Waals forces, hydrophobic effects, chemical bonding). The relative importance of different forces changes with the environmental conditions. Given its complexity, we focus on two topics and discuss multiple binding mechanisms that take place either in sequence or simultaneously.

Forces involved in a single bacterium on a mineral surface

We start with a simple scenario where a typical bacterium approaches and attaches to a negatively-charged mineral surface (Fig. 4, adapted from Liao et al. (2015)). A bacterium equipped with flagella can enhance the attachment by moving toward the surface and overcoming the surface water barrier. While only a weak van der Waals force may be present at >50 nm distance, the bacterium closer to the surface may face a hydrodynamic boundary layer and repulsive electrostatic forces. When the distance becomes approximately 2–10 nm, both repulsive and attractive electrostatic interactions may occur depending on the range of functional groups distributed non-uniformly on the cell surface (section 2.1.3). Hydrophobic groups on the cell surface (e.g., fimbriae and ‘hairy’ polymers, see Section “Cell surface structure”) may expel water molecules adsorbed on the surface at this distance. The exclusion of surface water molecules during the hydrophobic interaction between cells and surfaces would increase the entropy and lower the Gibbs free energy (Rosenberg, 1991). At <1 nm distances, a variety of chemical interactions (ligand exchange, ion bridging, van der Waals) between the cell and the surface enable and reinforce the attachment (i.e., adhesion).

Upon drying, cells can approach the mineral surface due to evaporative loss of water films? Thus, under dry conditions or in soil undergoing wet–dry cycles, non-coulombic forces may become increasingly important. Water stress (as well as predation pressure) could induce increased EPS production by bacteria (Kaur et al., 2022) and, therefore, promote EPS-induced complex interactions (next section).

Limited detachment of adhering cells (i.e., reduced reversibility of cell–mineral attachment) has been observed frequently. Spectroscopic studies examining cell–mineral interfaces revealed time-dependent changes in bonds formed at mineral surfaces. On Fe oxide (hematite), the bacterial attachment started by inner-sphere complexation with phosphate groups (formation of surface phosphorus (P)–oxygen (O)–Fe bonds), followed by bonding of carboxylates (Chorover, 2022). The importance of non-coulombic interactions at the cell–mineral interface increases in the presence of EPS (see next section).

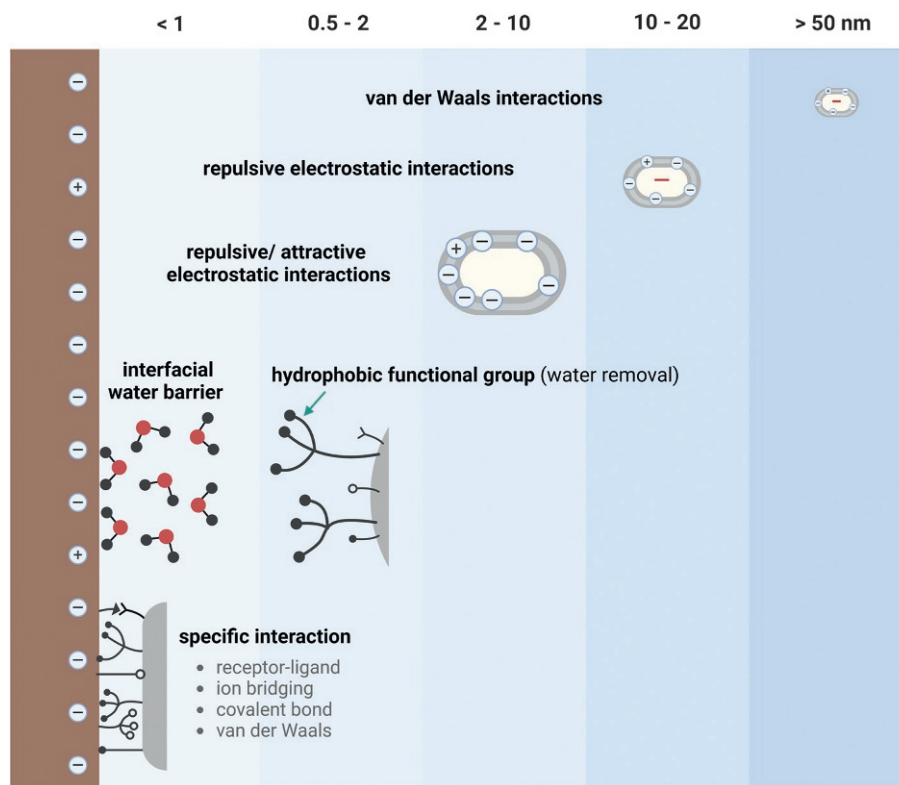


Fig. 4 Forces involved as a cell approaches a net negatively charged mineral surface. Modified from Liao C, Liang X, Soupir ML, and Jarboe LR (2015) Cellular, particle and environmental parameters influencing attachment in surface waters: A review. *Journal of Applied Microbiology* 119: 315–330.

Role of EPS at the cell-mineral interface

The EPS contain multiple functional groups (see Section “Functional group chemistry of cell surfaces”) as well as uncharged components and moieties, allowing their interaction with mineral surfaces through electrostatic interaction (at distance) and by van der Waals interactions, H bonds, cation bridges, cation–anion exchange, or hydrophobic interactions at proximity (Chorover, 2022). The co-occurrence of coulombic and non-coulombic interactions makes them act as effective glues, promoting adhesion among cells and of cells to minerals.

Bacteria in natural environments can be present in the form of biofilm (clusters of microbial communities embedded in EPS), although the relative abundance of biofilm-forming bacteria in the soil system is largely unknown. The formation of biofilms starts with reversible surface attachment, followed by irreversible attachment, colonization, biofilm maturation, and eventually dispersion (Sauer, 2003). The transformation from reversible to irreversible attachment is facilitated by EPS excretion.

The EPS contain both hydrophobic (DNA, proteins, lipopolysaccharides) and hydrophilic (polysaccharides) components (More et al., 2014). Due to the variation in EPS composition as well as mineral surface chemistry, EPS can have both positive and negative effects on cell adhesion (Tourney and Ngwenya, 2014). Binding of EPS has also been reported to decrease the hydrophobicity of minerals (Poorni and Natarajan, 2013). Coverage of minerals by negatively charged EPS can decrease the surface’s isoelectric point and either promote aggregation or, depending on the extent of coverage, generate electrostatic repulsion among colloidal mineral particles (Guhra et al., 2022, reference therein). In addition, changes in pH can alter the protonation or deprotonation of functional groups and/or induce EPS structural changes, which in turn affects EPS–cell–mineral interactions.

Consequences of microbial–mineral interfacial reactions

Microbe–mineral interactions have significant impacts on soil processes, such as biogeochemical cycling, at the pedon, ecosystem, catchment, and even global scale. The initial and one of the most profound consequences of the interactions is rock weathering, which includes the physical breakdown and chemical dissolution of rock minerals as well as the neoformation of secondary minerals that are often more chemically reactive than primary minerals. Driven by their needs for nutrients and energy, microbial activities in soil strongly control the cycling of elements that are important for biota. Microbial–mineral interaction also contributes to soil aggregation and the development of porous soil structure, which supports water storage and gas diffusion, ensuring terrestrial ecosystem functioning.

Microbially-mediated weathering and formation of minerals

Rock weathering starts with the water percolating into pores and cracks, largely created by abiotic effects, such as fluctuation in temperature and moisture (e.g., freeze–thaw and wet–dry cycles). By colonizing and dissolving the surface of minerals, microbes acquire nutrients or energy, sequester toxic substrates, or create environments suited for their life (Chorover, 2022). Initial colonizers include (a) chemolithoautotrophs that obtain energy from the oxidation of reduced elements in minerals (e.g., Fe, Mn, sulfur (S)) while fixing CO₂, (b) photoautotrophic bacteria, cyanobacteria, and algae (particularly in nutrient-poor systems exposed to sunlight), and (c) heterotrophic microbes that acquire energy by oxidizing organic C fixed by autotrophs (Chorover, 2022). Their successive colonization and proliferation result in partial or full destruction of minerals by both physical and chemical effects as well as the formation of secondary, often highly reactive, minerals. These neoformed mineral phases, in turn, enhance the accumulation of microbial products (extracellular polymers, necromass) by sorption, complexation, and aggregation reactions.

Microbially-mediated mineral or rock weathering and transformation have been studied for fungi (Smits and Wallander, 2017; Martino and Perotto, 2010) and bacteria (Uroz et al., 2009). Here, we briefly summarize the physical and chemical mineral weathering due to microbe–mineral interfacial interactions.

Physical aspects

Fungi and lichens are among the first colonizers. Compared to bacteria, fungi have the advantage of actively extending hyphae even to the hydrophobic and dry surfaces of minerals (sections “Cell surface structure” and “Non-coulombic interaction between cell and mineral”). Fungi can sense the size and type of minerals to maximize the benefit of foraging, and physically destruct minerals by widening existing pores and cracks as their hyphae grow (Smits and Wallander, 2017). Fungi can generate adhesive forces at their hyphal tips on contact with the mineral surface, producing significant osmotic pressure. The forces at the mineral surface together with the large internal hyphal pressure (0.4 to 1 MPa) (Martino and Perotto, 2010) can mechanically break down minerals. Mycorrhiza may cause the formation of tubular pores on minerals by either dissolution processes, and the newly created pores may weaken the minerals’ physical strength and promote further penetration of hyphae (Martino and Perotto, 2010). Bacteria also promote physical weathering but their contribution has been studied much less than that of fungi.

In general, physical weathering processes, either abiotically or biotically induced, increase the specific surface area of minerals by creating fractures and increasing pores, cracks, and surface roughness. Increased mineral surface area, in turn, promotes accelerated abiotic and biotic chemical dissolution and microbial colonization.

Chemical aspects

Modes of microbially-mediated chemical weathering can be placed into three categories: change in pH, chelation (ligand-promoted dissolution), and redox reactions (Samuels et al., 2020). Together with physical weathering, they are depicted in Fig. 4. Microbially mediated weathering can (i) cause structural changes in minerals or dissolve them, (ii) provide nutrients to microbes and plants, (iii) sustain microbial energy demands (electron transfer reactions), and (iv) promote secondary mineral formation.

Protons and negatively charged ligands released either during microbial processing of organic substrates or direct secretion, in particular of low-molecular-weight organic acids (LMWOA), can cause the dissolution of minerals (Fig. 5). Microbially mediated proton production can be induced by ammonium oxidizers (e.g., after N fertilizer application) and Fe oxidizers (e.g., at redox fluctuating zones). Ligand-promoted dissolution starts with metal complexation or chelation by various biomolecules (chelators). In addition to the compounds produced during incomplete oxidation of organic substrates, microbes (and plant roots) secrete numerous LMWOAs, such as oxalic acids, citric acids, amino acids, and phenolic acids, which all are effective complexing agents (chelators). Siderophores are specific chelators that selectively complex ferric ions (Fe(III)). Complexing agents can bind directly to the metals on the mineral surface, causing mineral dissolution. Complexation of metal ions by LMWOAs in solution reduces their activity and thereby accelerates mineral dissolution (Berenjian and Seifan, 2022).

The EPS and biofilms that cover mineral surfaces can drive chemical and physical mineral weathering. Due to their diverse composition, EPS can hinder or accelerate the dissolution of minerals. They mediate metal transport and availability by sorption, complexation, redox cycling, particle trapping, and biomineralization. While neutral EPS components (e.g., sugars, starch, cellulose, xanthan) have no effects on mineral dissolution (Tourney and Ngwenya, 2014, the references therein), acidic EPS components dissolve minerals by complexation of metals or by lowering pH (Samuels et al., 2020). Since EPS is mostly negatively charged, it binds with and accumulates cations, which might result in the neoformation of minerals (Berenjian and Seifan, 2022). In addition, EPS can act as an electron acceptor or donor (Flemming and Wingender, 2010).

Microbes on the surfaces or in the vicinity of minerals may utilize minerals as electron sinks or sources in redox reactions. The mineral redox reactions are coupled with half reactions that provide nutrients, generate energy, or prevent toxic effects for microorganisms (Chorover, 2022). Oxidation of reduced metals (e.g., Fe(II), Mn(II)) present in mineral structures may serve as energy sources in C- and nutrient-poor environments, such as regolith and at the soil–rock interface. Microbially oxidized metals can undergo nucleation and subsequently secondary mineral precipitation on minerals, microbial cells, and EPS (Chorover, 2022). In addition, redox reactions either produce (oxidation) or consume (reduction) protons, and thereby affect proton-driven weathering.

Microbially-mediated mineral (trans)formations can alter soil physical and chemical properties as well as soil development pathways. For example, mineral weathering by ectomycorrhizal (EM) fungi contributes to podzolization by dissolving Al in upper horizons (van Schöll et al., 2008). Bacteria may help EM fungi or even dominate the weathering process by producing LMWOAs

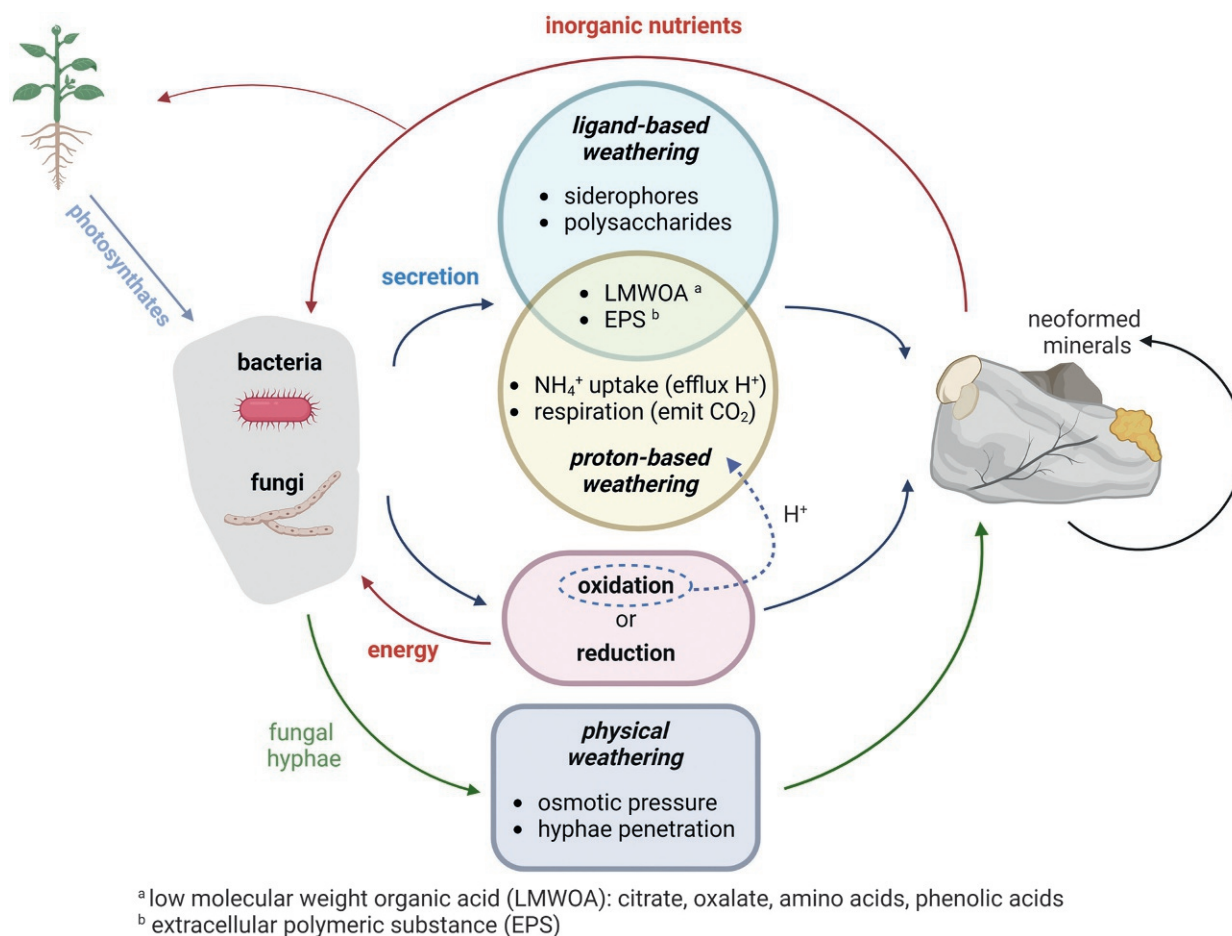


Fig. 5 Microbially-induced mineral weathering pathways. At the microbial–mineral interface, elements within the mineral structure are released by different microbially-mediated processes: proton-induced dissolution, complexation with ligands, and oxidation/reduction. Physical weathering may support the chemical weathering. The elements released are either taken up by microorganisms and plants or precipitate to form secondary minerals such as phyllosilicate clays as well as Fe, Mn, and Al hydrous oxides.

(Lambers et al., 2009). However, only certain types of bacteria and fungi control the weathering of specific minerals. Removal of K^+ from primary phyllosilicate minerals (i.e. mica and chlorite) by bacteria and fungi support their transformation into expandable clay minerals such as vermiculite, and interstratified forms, and eventually the genesis of Vertisols. Fungi can alter the biotite interlayer spacing and promote both K^+ depletion and Fe(II) oxidation, causing the formation of vermiculite and Fe(III) oxide clusters (Bonneville et al., 2009). Iron-reducing bacteria induce the transformation of smectite into illite (Kim et al., 2004), while Fe-oxidizing bacteria can oxidize Fe(II) and dissolve Al and Si from illite, thus, promoting the neoformation of smectite (Zhao et al., 2017). Microbially-formed secondary Fe(III) and Mn(IV) minerals can serve as electron acceptors for heterotrophic respiration in oxygen-limited environments (Chorover, 2022). The alteration and neoformation of clay minerals and hydrous metal oxides can modify soil physical properties (e.g., porosity, water and gas permeability) and control soil fertility by adsorption–desorption reactions and stabilization of soil OM. Microbially-mediated transformation of minerals is therefore a crucial component of soil and ecosystem development.

Element cycling

Microbially-mediated weathering makes rock-derived elements available to biota, and thereby fuels biogeochemical cycles. All essential elements for microbes and plants, except for those present dominantly in water and air (i.e., H, C, N, O), derive from rock (Table 1). Nutrient elements released from rock minerals will form precipitates, leach downward, or become incorporated in plant and microbial biomass, and are re-released during microbial breakdown of plant debris and microbial necromass. The re-cycling of OM-bound nutrients is of increasing importance as soils deplete in rock mineral sources with progressing ecosystem development and pedogenesis. Since OM interacts with minerals, much of the OM turnover and related element cycling in soils is linked to microbial–mineral as well as mineral–OM–microbial interactions.

Table 1 Examples of rock mineral sources of essential nutrients to microbes in soils.

Elements	Minerals
K	feldspar (microcline, orthoclase), mica (biotite, muscovite), illite, smectite (saponite, montmorillonite)
Ca	feldspar (anorthite), apatite, smectite (saponite, Ca-montmorillonite), calcite, pyroxene
Mg	mica (biotite), smectite (saponite), chlorite, sepiolite, phlogopite, palygorskite, dolomite, hornblende, olivine, pyroxene, amphibole
Fe	mica (biotite), chlorite, sepiolite, lizardite, olivine, pyroxene, amphibole, pyrite
Mn	pyroxene, amphibole
P	apatite, hydroxyapatite, strengite, variscite
S	pyrite, chalcopyrite

Si is not classified an essential nutrient.

The vast majority of soil microbes gain energy from organic substrates that ultimately originate from plant photosynthates. The entry of plant-derived OM into the soil system involves OM–microbe interfacial reactions, such as microbial colonization on plant litter surfaces. Depolymerization of plant-derived macromolecules that requires the binding of extracellular enzymes to their target biomolecules (e.g., cellulase to cellulose), is often the rate-limiting step in OM decomposition processes. Despite this, our understanding of how heterotrophic bacteria and fungi colonize and degrade the surface of plant-derived OM in soil is limited to the fungal degradation of wood. Microbial metabolites and residues, especially bacterial and fungal cell wall fragments (Miltner et al., 2012), together with plant-derived compounds, are the major components of OM in soils (Sokol et al., 2019). Due to microbial decomposition, the OM entering soil decreases in molecular size and becomes increasingly oxidized. The increasing number of oxidized functional groups (e.g., carboxylic and phenolic groups) allows for various sorptive interactions with mineral surfaces. These interactions with minerals tend to slow down microbial decay and thus enhance the persistence of soil OM and retention of the OM-bound elements (i.e., C, N, P, S).

While mineral-bound OM is relatively stable, it can be re-mobilized by plant- and microbially-secreted organic ligands or by electron shuttling (see “Chemical aspects” above), allowing C, N, and redox-sensitive metals to cycle in the ecosystem. The oxidation of reduced elements including organic C is coupled with the reduction of other elements. Under O₂-limited conditions, as in waterlogged soils, oxidized metals (i.e., Mn(IV) and Fe(III)) in minerals may then serve as electron acceptors of heterotrophic respiration. The reduction of these metals promotes the dissolution of their host minerals. Organic and inorganic forms of elements bound to or co-precipitated with Mn(IV) and Fe(III) hydrous oxides would be re-dissolved and available for biota by the reduction of these secondary minerals.

Contribution to soil aggregate formation

Another consequence of mineral–OM–microbial interaction is the clustering or aggregation of soil organic and inorganic particles. Microbial cells and extracellular polymers can bind strongly to mineral surfaces, and microbially-mediated mineral weathering promotes the formation of nano-sized secondary minerals featuring large specific surface areas and reactivities (see above). These organic and mineral particles interact strongly with each other in non-random ways to form composite particles, so-called aggregates.

An intriguing example of the formation of such structural units at the micro scale is the intimate association of bacteria with their metabolites and exudates, and nano-sized minerals, as revealed by transmission microscopic observation of soil thin sections (Fig. 6A and B). Such bacterial microaggregates encrusted by EPS-clay coatings may form within weeks at the suspension–solid interface (Lünsdorf et al., 2000). Nano-sized, secondary Fe hydrous oxides as well as short-range-ordered aluminosilicates can precipitate on negatively charged surfaces of microbes and EPS (Chorover, 2022). Clay coatings on fungal hyphae, bacterial cells, or microcolonies have been observed in soil thin sections, often embedded in EPS. Their thickness varies between 0.1 and 1 µm (thicker on hyphae) and the clays tend to be reoriented in parallel to the microbial surfaces (Chenu and Cosentino, 2011). Such coatings may be advantageous against desiccation and predation with the tradeoff of reduced diffusion of oxygen, enzymes, and substrate (Chenu and Cosentino, 2011).

At a slightly larger spatial scale, particulate OM (POM), representing plant debris, is often covered with nano- to micron-sized minerals (Fig. 5C), which can result from colonization by heterotrophic microbes that degrade POM. Such POM–microbe–clay interaction may serve as starting point for the occlusion of POM into larger aggregates (Golchin et al., 1997). Then the degradation of occluded POM is strongly controlled by the aggregate pore structure (Schlüter et al., 2022).

Microbial interfacial interactions can contribute to the formation and stability of soil aggregates by several mechanisms. Extracellular polymers (e.g., EPS) and microbial biomass itself (e.g., hyphae, necromass) can cause and enhance inter-particle cohesion. Extracellular polysaccharides with high molecular weight (10⁵ to >10⁶ Da) and the capability to form intermolecular linkages are most effective for inter-particle cohesion (Chenu and Cosentino, 2011). The hyphal network of fungi and actinobacteria can effectively entangle and interlink soil particles. Controlling factors include the strength of adhesion at the hyphae–particle contact point, hyphae tensile strength, hyphae spatial distribution, and fungal taxonomic group (Chenu and Cosentino, 2011). Fungal hydrophobin (section “Cell surface structure”) enhances water repellency and, thus, aggregate stability against water dispersion.

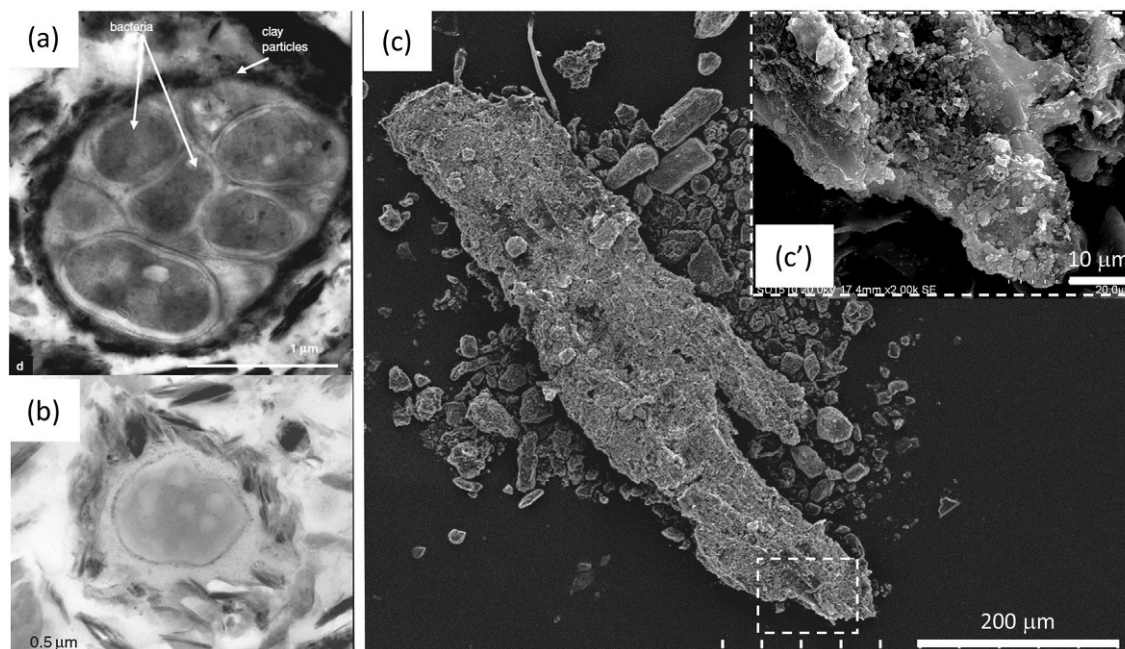


Fig. 6 Examples of aggregation as consequences of microbe–organic matter–mineral interfacial reactions. Transmission electron microscope images of bacterial microaggregates observed in thin sections of (A) an intact soil (Robert and Chenu, 1992) and (B) water-dispersible clay (Chenu and Plante, 2006). (C) SEM image of decaying plant detritus with clay coatings on its surface (C').

On the other hand, soil microbes promote disaggregation as most soil microbes actively degrade OM to gain energy and to obtain nutrients. Even the enzymes present in EPS can reduce aggregate stability (Kaur et al., 2022, and references therein). Organic compounds, therefore, can serve as substrates or aggregate binding agents. At the macro-aggregate scale, the input of plant-derived OM into soil appears to control macroaggregate formation and the microbial contribution is rather passive. This is evident from the frequently observed breakdown of macroaggregates with the reduction of OM input in agricultural soils.

A fundamental soil feature arising from the adhesion and clustering of organic and inorganic particles is the formation of pores, the void spaces between those particles. It is these pores or along their walls where solid–water–gas interfacial reactions take place. Connected pores (i.e., pore network) provide the flow path for water and gases, as well as providing habitats for soil microbes, micro- and meso-fauna, and plant roots. The chemical and physical changes that take place with the interfacial interactions of OM, minerals, and microbes, therefore, have profound consequences on biogeochemical cycles and pedogenesis (Chorover, 2022).

Outlook

Starting from the different mechanisms of microbial attachment to mineral surfaces, we have elaborated three fundamental consequences of microbe–OM–mineral interfacial reactions taking place in the soil, namely mineral weathering and transformation, elemental cycling, and formation of structure. Once a rock is exposed to the surface of the Earth, mineral surfaces are prone to initial microbial attachment and colonization. With the activities of these first colonizers (e.g., Fe oxidation), secondary minerals start forming, and rock-derived elements begin to be available via weathering, providing more extensive surfaces and nutrients to support microbial and plant life. The secondary minerals then bind strongly to OM, forming metastable OM-mineral agglomerates that feature both mineral and organic surfaces. Due to the heterogeneity of the surface chemistry (and physical structure), the neoformed mineral and organo–mineral phases provide habitats for more diverse microbial communities and plant roots. With these feedback loops (microbes create new surfaces and these surfaces allow more and other microbes to colonize), microbe–mineral surface interactions drive pedogenesis and ecosystem development toward greater diversity and complexity. Therefore microbe-mediated weathering, element cycling, and porous structure formation are essential components of pedogenesis and soil functioning.

Despite the knowledge accumulated so far, our current understanding of microbe–mineral interfacial interactions still appear incomplete. Open research questions that might be worth exploring further in the future include

- Colonization of mineral surfaces: Are biofilms common in soils? Is their occurrence constrained by certain physical conditions and microenvironments? How do they form and are they confined to specific microbial communities? How stable is the attachment of biofilms on the surfaces?

- Weathering: Are there antagonistic, synergistic, and successive effects of bacteria and fungi in the destruction of primary and the formation of secondary minerals? Is it mostly mycorrhizal fungi that initiate the weathering process, or is the contribution of bacteria rather underestimated? What are the individual controlling factors of bacterial and fungal weathering?
- Element cycling: How does microbe–OM–mineral attachment relate to stabilization and destabilization of OM? Or more precisely, what controls the balance between stabilization and destabilization? Is it linked to the type of mineral, chemical conditions, and localization (i.e. rhizosphere vs. bulk soil)?
- Structure formation: While the role of microbe–mineral interfacial interactions in the formation of nano- and micro-scale structural soil units is well recognized, it remains unclear if these effects extend to the macrostructure of soil. To what extent and how does the porous, hierarchical soil structure account for soil microbial diversity? What are the dominant factors controlling the formation of micro- and macro- structures?
- Consequences of microbe–mineral surface interactions: To what extent, do microbemineral interactions control the rate and trajectory of pedogenesis and ecosystem development? Does the contribution change with the stages of pedogenesis and ecosystem development? What is the contribution of microbe–mineral interactions to soil processes relative to those induced by plants?

Answering these questions will not only add to the knowledge on surface interactions in soil but might also add much to the emerging understanding that soils form, develop, and function in more direct responses to microbial life.

References

- Berenjian A and Seifan M (2022) *Mineral Formation by Microorganisms*. Springer Cham.
- Bonneville S, Smits MM, Brown A, et al. (2009) Plant-driven fungal weathering: Early stages of mineral alteration at the nanometer scale. *Geology* 37: 615–618.
- Bromley KM, Morris RJ, Hobbey L, et al. (2015) Interfacial self-assembly of a bacterial hydrophobin. *Proceedings of the National Academy of Sciences of the United States of America* 112: 5419–5424.
- Chenu C and Cosentino D (2011) *Microbial Regulation of Soil Structural Dynamics*. CABI International.
- Chenu C and Plante AF (2006) Clay-sized organo-mineral complexes in a cultivation chronosequence: Revisiting the concept of the 'primary organo-mineral complex'. *European Journal of Soil Science* 57: 596–607.
- Chorover J (2022) 5. Microbe–biomolecule–Mineral interfacial reactions. In: Yang Y, Keilueit M, Senesi N, and Xing B (eds.) *Multi-Scale Biogeochemical Processes in Soil Ecosystems: Critical Reactions and Resilience to Climate Changes*, pp. 117–140. John Wiley & Sons, Inc.
- Feofilova EP (2010) The fungal cell wall: Modern concepts of its composition and biological function. *Microbiology* 79: 711–720.
- Flemming H-C and Wingender J (2010) The biofilm matrix. *Nature Reviews Microbiology* 8: 623–633.
- Free SJ (2013) Chapter Two - Fungal cell wall organization and biosynthesis. In: Friedmann T, Dunlap JC, and Goodwin SF (eds.) *Advances in Genetics*, pp. 33–82. Academic Press.
- Golchin A, Baldock J, and Oades J (1997) *A Model Linking Organic Matter Decomposition, Chemistry, and Aggregate Dynamics. Soil Processes and the Carbon Cycle*. Boca Raton: CRC Press 245–266.
- Guhra T, Stolze K, and Totsche KU (2022) Pathways of biogenically excreted organic matter into soil aggregates. *Soil Biology and Biochemistry* 164: , 108483.
- Hong Y and Brown DG (2006) Cell surface acid–base properties of *Escherichia coli* and *Bacillus brevis* and variation as a function of growth phase, nitrogen source and C:N ratio. *Colloids and Surfaces B: Biointerfaces* 50: 112–119.
- Hong Z, Rong X, Cai P, et al. (2012) Initial adhesion of *Bacillus subtilis* on soil minerals as related to their surface properties. *European Journal of Soil Science* 63: 457–466.
- Kaur R, Page KL, Dalal RC, et al. (2022) Can molecular genetic techniques improve our understanding of the role the microbial biomass plays in soil aggregation? *European Journal of Soil Science* 73: , e13307.
- Kim J, Dong H, Seabaugh J, et al. (2004) Role of microbes in the smectite-to-illite reaction. *Science* 303: 830–832.
- Kosmulski M (2009) Compilation of PZC and IEP of sparingly soluble metal oxides and hydroxides from literature. *Advances in Colloid and Interface Science* 152: 14–25.
- Lambers H, Mougél C, Jaillard B, and Hinsinger P (2009) Plant-microbe-soil interactions in the rhizosphere: An evolutionary perspective. *Plant and Soil* 321: 83–115.
- Li GL, Zhou CH, Fiore S, and Yu WH (2019) Interactions between microorganisms and clay minerals: New insights and broader applications. *Applied Clay Science* 177: 91–113.
- Liao C, Liang X, Soupir ML, and Jarboe LR (2015) Cellular, particle and environmental parameters influencing attachment in surface waters: A review. *Journal of Applied Microbiology* 119: 315–330.
- Lünsdorf H, Erb RW, Abraham W-R, and Timmis KN (2000) 'Clay hutchies': A novel interaction between bacteria and clay minerals. *Environmental Microbiology* 2: 161–168.
- Martino E and Chorover J (2010) Mineral transformations by mycorrhizal fungi. *Geomicrobiology Journal* 27: 609–623.
- Miltner A, Bombach P, Schmidt-Brücken B, and Kästner M (2012) SOM genesis: Microbial biomass as a significant source. *Biogeochemistry* 111: 41–55.
- More TT, Yadav JSS, Yan S, et al. (2014) Extracellular polymeric substances of bacteria and their potential environmental applications. *Journal of Environmental Management* 144: 1–25.
- Naja G, Deneux-Mustin S, Mustin C, et al. (1999) Potentiometric titration: A dynamic method to study the metal binding-mechanism of microbial biomass. In: Amils R and Ballester A (eds.) *Process Metallurgy*, pp. 201–210. Elsevier.
- Op De Beeck M, Persson P, and Tunlid A (2021) Fungal extracellular polymeric substance matrices – Highly specialized microenvironments that allow fungi to control soil organic matter decomposition reactions. *Soil Biology and Biochemistry* 159: , 108304.
- Parikh SJ and Chorover J (2006) ATR-FTIR Spectroscopy reveals bond formation during bacterial adhesion to iron oxide. *Langmuir* 22: 8492–8500.
- Poorni S and Natarajan KA (2013) Microbially induced selective flocculation of hematite from kaolinite. *International Journal of Mineral Processing* 125: 92–100.
- Robert M and Chenu C (1992) Interactions between soil minerals and microorganisms. In: Stotzky G and Bollag JM (eds.) *Soil Biochemistry*, pp. 307–404. New York: CRC Press.
- Rosenberg M (1991) Basic and applied aspects of microbial adhesion at the hydrocarbon: Water interface. *Critical Reviews in Microbiology* 18: 159–173.
- Samuels T, Bryce C, Landenmark H, et al. (2020) Microbial weathering of minerals and rocks in natural environment. In: Dontsova K, Balogh-Brunstad Z, and Roux GL (eds.) *Biogeochemical Cycles: Ecological Drivers and Environmental Impact*. American Geophysical Union.
- Sauer K (2003) The genomics and proteomics of biofilm formation. *Genome Biology* 4: 219.
- Schlüter S, Leuther F, Albrecht L, et al. (2022) Microscale carbon distribution around pores and particulate organic matter varies with soil moisture regime. *Nature Communications* 13: 2098.
- Smits MM and Wallander H (2017) Chapter 3 - Role of mycorrhizal symbiosis in mineral weathering and nutrient mining from soil parent material. In: Johnson NC, Gehring C, and Jansa J (eds.) *Mycorrhizal Mediation of Soil*, pp. 35–46. Elsevier.
- Sokol NW, Sanderman J, and Bradford MA (2019) Pathways of mineral-associated soil organic matter formation: Integrating the role of plant carbon source, chemistry, and point of entry. *Global Change Biology* 25: 12–24.

- Solari JA, Huerta G, Escobar B, et al. (1992) Interfacial phenomena affecting the adhesion of *Thiobacillus ferrooxidans* to sulphide mineral surface. *Colloids and Surfaces* 69: 159–166.
- Tourney J and Ngwenya BT (2014) The role of bacterial extracellular polymeric substances in geomicrobiology. *Chemical Geology* 386: 115–132.
- Uroz S, Calvaruso C, Turpault M-P, and Frey-Klett P (2009) Mineral weathering by bacteria: Ecology, actors and mechanisms. *Trends in Microbiology* 17: 378–387.
- Van Schöll L, Kuyper TW, Smits MM, et al. (2008) Rock-eating mycorrhizas: Their role in plant nutrition and biogeochemical cycles. *Plant and Soil* 303: 35–47.
- Wingender J, Neu TR, and Flemming H-C (1999) *Microbial Extracellular Polymeric Substances*. Berlin, Heidelberg: Springer.
- Zhao L, Dong H, Edelmann RE, et al. (2017) Coupling of Fe(II) oxidation in illite with nitrate reduction and its role in clay mineral transformation. *Geochimica et Cosmochimica Acta* 200: 356–366.

Pesticide fate in soils

Pierre Benoit, Laure Mamy, Carole Bedos, and Enrique Barriuso, Université Paris-Saclay, INRAE, AgroParisTech, UMR ECOSYS, Palaiseau, France

© 2023 Elsevier Ltd. All rights reserved.

Introduction	471
Physicochemical properties of pesticides	471
Sorption and mobility in soil	472
Sorption	472
Mobility and transport	472
Transfer toward surface water	472
Transfer toward groundwater	474
Transfer toward the atmosphere	474
Persistence in soil	475
Abiotic degradation	475
Biotic degradation	475
Pesticide non-extractable (bound) residues	476
Effects on soil biota and uptake by plants	478
Uptake of pesticides by plants	478
Effects on soil organisms (microflora and macrofauna)	479
Integrating process-based knowledge in risk assessment	479
Pesticide fate models	479
Concluding remarks and the future	481
References	481

Abstract

Pesticide molecules in soils undergo different processes of biological, physical and physico-chemical origin, and occurring over different time (minutes to decades) and spatial (nm to km) scales. Pesticides are represented by a diversity of molecules having contrasting physico-chemical properties inducing various kinds of behavior in terms of persistence, sorption, transport with the potential for contamination of water, uptake by plants, and possible impacts on soil biota and microbial processes. This chapter proposes an update of current available knowledge on the main processes and current state of the art of knowledge integration in pesticide fate models at the plot and watershed scales for environmental risk assessment.

Key points

- The behavior of most pesticide is controlled by their molecular properties and the soil characteristics.
- Individual processes are interdependent due to their coupling in time and space.
- This coupling is driven by local pedoclimatic conditions related to composition of the soil solution and soil particles, soil porosity determining microhabitats for degrading microorganisms and accessibility to reactive sites.
- Significant progress has been made to represent spatially the main soil processes at the landscape scale and to integrate the different routes for environmental contamination or exposure.
- Although used for regulatory risk assessment, process-based models still have difficulties in describing some processes such as particulate and colloidal transport, the formation or release of non-extractable residues.
- Other limits of currently used models are related to the poor representation of diversified cropping practices and cropping successions (pluri-annual) and the feedback loops on soil processes.

Nomenclature

c	Concentration
DT_{50} , DT_{90}	Disappearance time (50%, 90%) (days)
K_H	Henry's Law constant (dimensionless)
k	First-order degradation rate constant (day^{-1})
K_d	Sorption coefficient to soil (liters per kilogram)
K_{oc}	Sorption coefficient to soil organic carbon (liters per kilogram)

K_{om}	Sorption coefficient to soil organic matter (liters per kilogram)
K_{ow}	Octanol/water partition coefficient (dimensionless)
T	Temperature (Kelvin)
t	Time
$t_{1/2}$	Half-life (days)
VP	Vapor pressure (Pascal)
WS	Water solubility (moles per liter)

Introduction

Synthetic organic pesticides have been used widely since the late 1940s to increase crop yields in agriculture. Though strict registration criteria have led to a ban on the use of many older compounds, such as organochlorine insecticides, several hundred active ingredients still remain available. With different regulations according to countries and continents, several compounds, now banned in the European Union (EU), are still used in USA, Latin-America, China or Africa. In the EU, for example, more than 450 of approximately 850 formerly approved compounds are still marketed. Pesticide contamination occurs in most environmental compartments (water, air, soil, living organisms) and has potential impacts on the environment and human health. Despite restrictions on use and some product withdrawals, pesticide residues can be detected for several years in the environment, and soils play a key role in the long-term fate of these agrochemicals.

The soil is the ultimate sink for a large part of the pesticide that is applied to crops. Non-agricultural uses may still exist but in other cases have been banned, depending on countries and current regulations. Nevertheless, agricultural uses are the major sources of pesticides in the environment. Compounds such as nematicides (for control of plant-parasitic nematode worms) are usually incorporated into the topsoil. Many herbicides used in arable crops are applied either to bare soil prior to emergence of the sown crop or to the seed bed immediately post-germination, and so most of the compounds reach the soil surface directly. In contrast, insecticides and fungicides are usually applied later in the season when the crop may have closed its canopy, and so most compounds will be intercepted by the crop foliage; even so, some spray will drift downward, and wash-off by rain may subsequently move more compound to the soil. The rates at which pesticides are applied vary widely: according to the FAO, European agriculture consumed an average of 3.1 kg ha^{-1} in 2018, with considerable variation from more than 8 kg ha^{-1} in the Netherlands and Belgium, to less than 1 kg ha^{-1} in Scandinavia and the Baltic countries.

To understand the behavior of so many diverse compounds in the environment, it is necessary to develop an overarching framework of understanding based on the physicochemical properties and behavior of each compound. Important aspects of pesticide behavior in soil to be considered are persistence, sorption, transport with its potential to contaminate water, uptake by plants, and possible impacts on soil biota and microbial processes.

Physicochemical properties of pesticides

Water solubility, WS, plays an important role in the fate of pesticides in the environment as one of the routes of pesticide transport within the soil or between water and other environmental compartments is water. The water solubility of pesticides was shown to be mainly related to the nature of the surface and volume of molecules, to their branching, and to their polarizability (which can be related to the volume) (Mamy et al., 2015).

The octanol-water partition coefficient, K_{ow} , is the most frequently used parameter to characterize the hydrophobicity (or lipophilicity) of chemicals, which is a very important property in the environmental sciences. It was demonstrated that the K_{ow} of pesticides depends on the volume of the molecule, its branching, the polarizability, the dipole moment, the energy of the highest occupied molecular orbital, and energy of the lowest unoccupied molecular orbital (Mamy et al., 2015).

The dissociation constant, pK_a , which describes the extent to which a compound dissociates in solution, is a fundamental physical property of a chemical. It is a key feature that governs the chemical reactivity of the pesticides with other compounds in any solvent, and also interaction with the solvent itself, and thus its hydrophobicity and water solubility. Mamy et al. (2015) showed that the pK_a of organic compounds were related to the charges of the molecule and to the energy of the highest occupied molecular orbital.

The vapor pressure, VP, and the Henry's law constant, K_H , are involved in the fit between physico-chemical properties and volatilization potential of compounds when considering pesticide volatilization from leaves and from soil, respectively. The VP is an indicator of the volatility of a pure chemical compound, and depends exponentially on the temperature. The K_H measures the partitioning of a chemical between two phases, air and water. Chemicals with low K_H will tend to stay in the aqueous phase, while those with high K_H will partition more into the gaseous phase. Air and water are major compartments of the environment, and water is considered to act as a vector among air, soil, sediment, and biota, therefore, knowledge of K_H is important in assessing environmental risks associated with a chemical. Polarizability and connectivity indices are good molecular properties to explain the vapor pressure range of organic compounds, but no particular property appeared to be more relevant than another to estimate K_H (Mamy et al., 2015).

Sorption and mobility in soil

Sorption

The sorption of pesticides in soils is a combination of the adsorption process, which is the passage of a solute from an aqueous phase to the surface of a solid adsorbent, and of desorption, which is the reverse process. According to the properties of solutes and substrates, several interaction mechanisms with the surface are possible: ion exchanges, interactions with metallic cations, polar interactions, charge transfers, and London-van der Waals forces/hydrophobic effect (Calvet, 1989). The greater is the pesticide sorption, the lower is its mobility and hence the risks of groundwater contamination. However, transport of adsorbed solute may also occur when adsorbing particles are displaced with the vertical migration of dispersed clays in the soil profile or through horizontal migration toward surface waters because of runoff or erosion (see below). The sorption of pesticides in soils is usually characterized by distribution coefficients between soil and soil solution according to various models such as the linear (K_d , L kg⁻¹) or the Freundlich ones (K_f). A high value for these coefficients indicates strong sorption. The sorption of pesticides depends on the molecular properties of the compounds, the soil properties, environmental conditions, and residence time in soils.

The sorption of pesticides in soils increases with the molecular volume of the compounds and with their branching (Mamy et al., 2015). It also depends on the ionization of the molecule (for example, cation sorption is high on negatively charged surfaces such as clay or humic substances), and on the hydrophilic or hydrophobic properties of the compound (sorption decreases when WS increases because the pesticide will have a greater affinity for the liquid phase). However, glyphosate for example, is a strongly adsorbed compound despite its strong water solubility because sorption of this herbicide (having four pK_a) involves high energy ionic bindings that counterbalance the effect of WS. Finally, the electronic structure of the molecule plays an important role in the sorption through polarity, the polarizability and charge delocalization (Calvet, 1989; Chaplain et al., 2011). It has to be emphasized that the surfactants of commercial products may change the sorption of active substances in soils (Beigel et al., 1998).

Organic matter (OM), and especially humic substances have strong chemical reactivity and are often the main adsorbents of pesticides. Clay minerals, oxides and hydroxides, however, may also be involved in adsorption (Calvet, 1989). In general, sorption increases with soil organic matter (OM) content, except for ionic compounds for which soil pH is the main determining property (Fig. 1, Barriuso and Calvet, 1992). Soil structure (bulk density, pore geometry), which also depends on agricultural practices and climate, influences pesticide sorption: for example, in conservation agriculture (low or no-till practices), OM accumulation on the soil surface favors the retention of pesticides (Alletto et al., 2010). Finally, pesticide sorption shows horizontal and vertical spatial variabilities due to the variation in soil properties (Chaplain et al., 2011).

Soil water content determines the retention of pesticides as it favors their transport toward the adsorption sites. In addition, when the soil water content increases, the OM becomes more hydrophilic which increases the sorption of hydrophilic pesticides. It was also demonstrated that the sorption of pesticides decreases when the soil temperature increases (Calvet, 1989).

The residence time of pesticide residues in soils changes their sorption properties through diffusion in the microporosity, physical entrapment and degradation (Barriuso et al., 2008). In general, a large increase in retention with time is observed for weakly adsorbed pesticides while for strongly adsorbed and persistent pesticides, sorption decreased with time or remained stable, and K_d does not vary with time when the pesticide's half-life is higher than 30 days (Fig. 2, Mamy and Barriuso, 2007).

Mobility and transport

Pesticides that arrive in soils can be transferred toward water and air. The transfers result from transport processes (molecular diffusion, advection/dispersion) coupled to sorption and degradation processes (Calvet, 1989). The main processes of transfer are runoff and erosion toward surface water, leaching toward groundwater, and volatilization toward the atmosphere. The relative extent of each process depends on the pesticide and soil properties, on the application methods, climatic conditions, and agricultural practices.

Transfer toward surface water

At the watershed scale, surface waters are streams, ponds, and ditches. Pesticide transfers occur in solute and or particulate phases. They are mainly controlled by climate, soil and pesticide properties, and agricultural practices.

Among the various climate characteristics, the intensity and duration of precipitation will strongly determine pesticide transfer toward surface water. Rain intensity can cause runoff when the speed of water supply exceeds the soil infiltrability, which is aggravated by long rain events. The date of the rain event compared to the date of pesticide application will also be decisive for transfer.

The soil slope is one of the key factors governing pesticide transfer toward surface water. When the slope increases, runoff speed increases so that erosion and the thickness of soil affected by erosion increase. The controlling soil characteristics are the presence of a weakly permeable soil layer, structural stability of the soils, and OM content which influences the structural stability. Finally, the soil water content is also crucial: if a rain event occurs when the soil is water saturated, water accumulates at the soil surface leading to runoff or to standing water depending on the slope.

Pesticide transfer depends on the physicochemical properties of the substance (WS, adsorption), and on the formulation of the plant protection product (for example, wettable powder and liquid formulations favor transport). They also depend on the

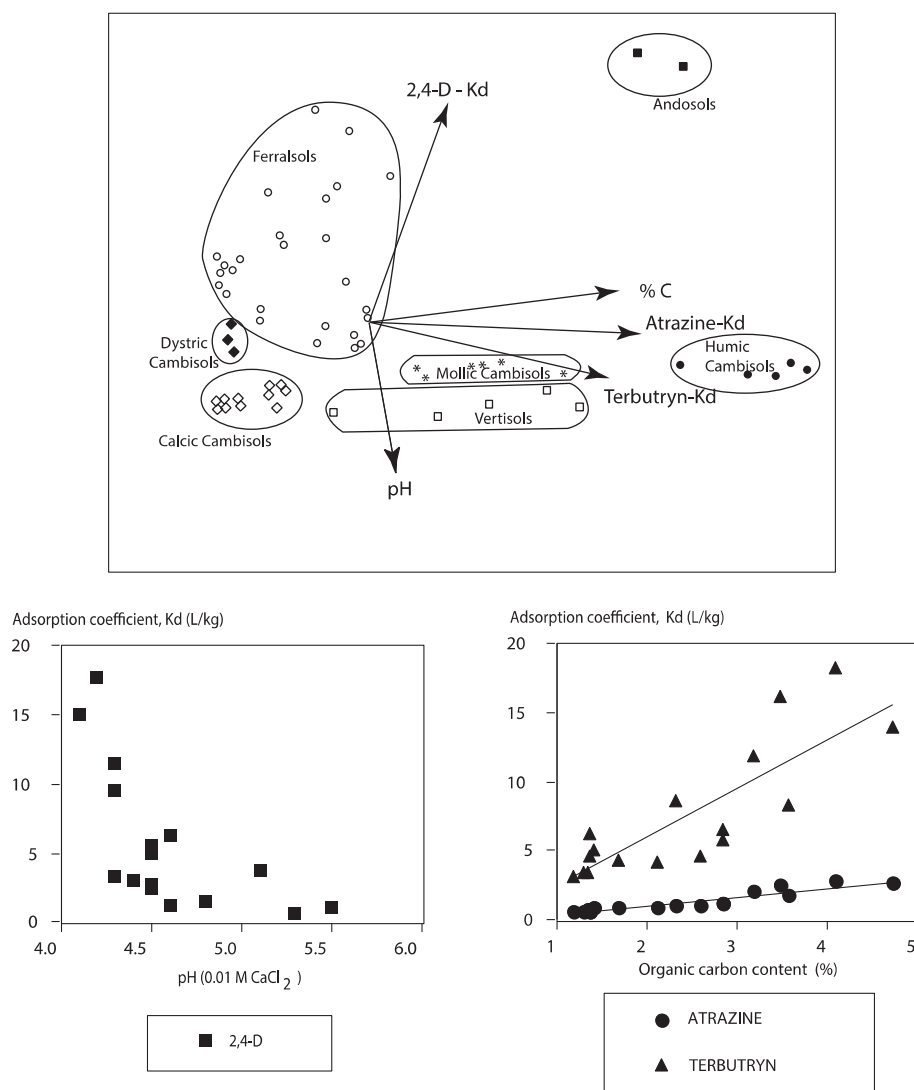


Fig. 1 Principal component analysis showing correlations between sorption coefficients K_d of three herbicides (atrazine, terbutryn and 2,4-D), organic carbon content and soil pH for surface horizons from different soil types; the two figures at the bottom highlight the relations between K_d and soil pH for a weak acid (2,4-D), and soil organic C for non-ionic and weak bases (atrazine, terbutryn). Adapted from Barriuso E and Calvet R (1992) Soil type and herbicides adsorption. *International Journal of Environmental Analytical Chemistry* 46: 117–128.

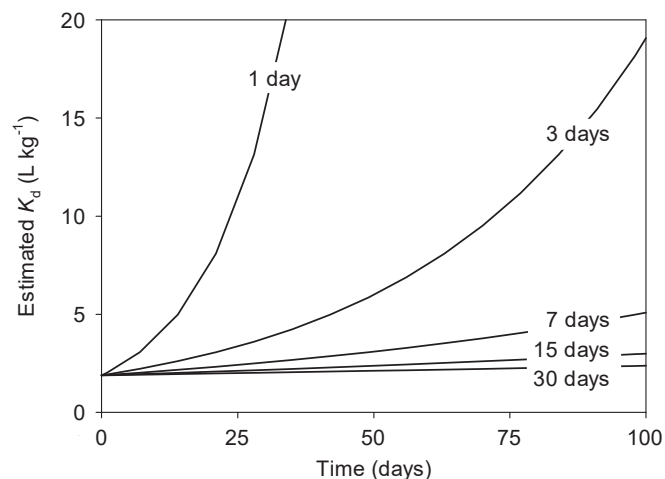


Fig. 2 Sensitivity analysis of the relationship between K_d and time as a function of pesticide half-life (days). From Mamy L and Barriuso E (2007) Desorption and time-dependent sorption of herbicides in soils. *European Journal of Soil Science* 58: 174–187.

amounts applied, and on the location of the pesticide in the soil (the risk of transfer strongly decreases if the pesticide is incorporated into the soil).

Maintaining plant cover or cropping residues (conservation agriculture) at the soil surface increases OM content leading to better soil structural stability and aggregate cohesion and a strong decrease in soil erosion. The presence of grass strips at the edges of fields favors pesticides adsorption and limits the amounts that could be transferred by runoff and erosion. Finally, irrigation may help to prevent runoff by maintaining an acceptable soil water content.

Transfer toward groundwater

Pesticides can reach groundwater through leaching, by vertical transfer in the dissolved phase, or through particulate transfer. The transfer mainly depends on pesticide sorption, soil properties and structure, and on the agricultural practices.

Numerous results have shown a close relationship between the depth of pesticide leaching and their sorption coefficients (Calvet, 1989). In general, the greater is the sorption, the less is the leaching. However, adsorbed pesticide may move vertically in the particulate phase (colloid facilitated transport).

The leaching of pesticides depends strongly on the soil structure which is heterogeneous and may show preferential pathways due to cracks (clay soils) or earthworms burrows, for example. Infiltrated water moves rapidly through macropores leading to limited contact with the adsorbing surfaces and to pesticide transfers at depth.

In conservation agriculture, the decrease in soil tillage favors the presence of macropores leading to a high risk of pesticides leaching (Alletto et al., 2010). On the contrary, the addition of organic amendments to compensate for the loss of soil fertility results in a decrease in vertical pesticide transfers.

Transfer toward the atmosphere

It is commonly accepted that volatilization, ranging from 0.1% of the applied pesticide dose to several tens of % under certain conditions, occurs after application from a few days to a few weeks for some compounds. However, the state of knowledge on the volatilization process is based on a relatively small number of datasets (fewer than 50 datasets have been acquired in the field on volatilization from soil) from studies developed since the 1990s, but that have been carried out with a diversity of methods. Moreover, the existing studies concern some compounds which are now banned or not widely used.

Volatilization depends on temperature and soil water content. Although an increase in temperature generally leads to an increase in volatilization, this effect may be limited by soil drying conditions, a common phenomenon at the soil surface, which may lead to an adsorption of the compound from the gaseous phase to the soil for certain compounds, thus limiting their volatilization. This sorption process occurs in addition to those presented in the previous section. García et al. (2014) took this process into account when modeling pesticide volatilization from bare soil to improve the description of diurnal dynamics. This process also explains the increase in S-metolachlor volatilized after 5 days with the increase in soil moisture at the time of application. This was observed by Prueger et al. (2017) through experiments conducted over 13 consecutive years. It should be noted that this effect cannot be predicted for all pesticides because the gas/soil adsorption coefficient is not available for all compounds, and the prediction of the soil water transfer under dry conditions remains a challenge.

The vapor pressure of a compound is the first indicator of its volatilization potential (see previous section). However, as compounds interact with the soil matrix, other characteristics have an indirect effect on the volatilization process, such as WS, K_{oc} and DT50. The estimated values of VP, K_H or WS can vary considerably depending on the method of determination used. This variability is important because it leads to uncertainty in the volatilization flux predictions.

In addition, the commercial formulation of the pesticide may also have an effect on the volatilization process. Houbraken et al. (2018) stated that effective vapor pressures of formulated compounds are the most appropriate values to be used in pesticide fate models. Das and Hageman (2020) estimated the soil-air partition coefficients of chlorpyrifos, pyrimethanil and trifluralin, formulated or not, and with the addition of adjuvants. They found that, under the conditions tested, formulated compounds may undergo greater volatilization than the active ingredient alone but they also indicated that these results can be contradictory depending on temperature.

The presence of a mulch can have a strong effect on volatilization by (1) increasing the exchange surface with the atmosphere (which tends to increase transfers), (2) modifying the temperature and humidity conditions and (3) modifying the availability of the pesticide to transfers, depending on whether or not it is adsorbed onto the mulch or degraded. However, prediction of the effect of mulches on volatilization is still uncertain (Bedos et al., 2017). Few data sets exist and the corresponding studies present contradictory results: an increase in volatilization in the presence of mulch but also the reverse effect, i.e. a reduced volatilization after rainfall as reported by Bedos et al. (2017). These authors based their study on the effect of the presence of a mulch on the volatilization dynamics of S-metolachlor and recommended to characterize better interception of the compound by the mulch at the time of application, modification of the surface conditions in the presence of a mulch, and finally its sorption/degradation in the mulch.

Incorporation of the pesticide into soil has been identified as a way to limit volatilization and photodegradation. Quantification of the effect of this practice on the volatilization of trifluralin applied to bare soil and incorporated into the soil 24 h after application, showed a decrease in the rate of volatilization by two orders of magnitude (Bedos et al., 2019).

Persistence in soil

The degradation of pesticides is one of the key processes of their fate in soil because it controls their removal from the environment. Degradation is due to numerous chemical transformations which change the composition and structure of the molecules. Degradation can be biotic or abiotic and several intermediate compounds (degradation products or metabolites) can be produced between the initial molecule and the final mineral ones. When the pesticide is transformed to a mineral molecule, such as CO_2 , it is fully eliminated (this is mineralization). Most of the time, degradation follows first-order kinetics, allowing the degradation half-time (DT_{50} ¹ days) to be calculated. The higher is the DT_{50} , the higher is the pesticide persistence in the soil. However, the degradation kinetics can also be bi-phasic or present a lag phase.

Abiotic degradation

Abiotic transformations are due to chemical reactions that are not catalyzed by enzyme systems. They include hydrolysis, oxidation–reduction, and photochemical reactions.

Hydrolysis and oxidation-reduction reactions occur in solution. The reaction kinetic depends on the ionic composition of the medium, in particular pH and metal cations, on the nature of adsorbing phases, and on the temperature. The most favorable physicochemical conditions for reduction reactions are found in hydromorphic soils. The electrons sources come from sulfides, reduced metals, or organic compounds originating from OM or microorganism activity. Oxidation reactions are less known, but they seem to be favored by adsorption on clay minerals and metal oxides (Chaplain et al., 2011). The hydrolysis and reduction reactions of pesticides depend on pesticide quantum molecular properties such as the energy of the lowest unoccupied molecular orbital, while the highest occupied molecular orbital is the most useful pesticides molecular property to estimate their oxidation in the environment (Mamy et al., 2015).

Pesticide molecules can absorb a part of the light energy (UV and visible) they receive. The formation of an excited state through the absorption of a photon allows the molecule to be transformed or to interact with other molecules. Photochemical reactions may degrade pesticides by two types of processes: direct photolysis and indirect photodegradation. In direct photolysis, the pesticide itself absorbs light energy, becomes excited and, depending on the reaction activation energy, may undergo a transformation reaction. Indirect photodegradation is defined as the reaction of a ground-state pesticide with another photochemically produced species. Both reactions occur mainly in the dissolved phase when the pesticide is in the liquid phase of the soil, but also in the adsorbed phase if it is exposed to light. Consequently, photolysis or photodegradation reactions depend on the soil characteristics (especially porosity) but also on the climate, time of day, season, latitude, altitude, and atmosphere. In the liquid phase, degradation kinetics vary with pH, the presence of a solid phase or the presence of dissolved or suspended substances (Chaplain et al., 2011). Indirect photodegradation is favored when the soil water content is high while direct photolysis is more important in dry soil because light penetrates more easily. Agricultural practices also play a role, for example through the addition of organic amendments that add humic substances to the soil: humic substances tend to decrease direct photolysis through light screening, static and dynamic quenching. The most relevant pesticide molecular properties to explain the variability of photochemical reactions of organic compounds are the energy of the lowest unoccupied molecular orbital and energy of the highest occupied molecular orbital, molecular weight, polarizability and dipole moment (Mamy et al., 2015). Finally, it should be noted that pesticides that are intercepted by plants are on the leaf surfaces, and may also be subjected to photochemical reactions.

Biotic degradation

Biotic degradation is the main degradation mechanism of pesticides. It is due to the action of various organisms (mainly bacteria and fungi but also algae and protozoa) and results from chemical transformations by enzyme systems, which require the pesticides to be dissolved in the soil liquid phase. Pesticide biotic degradation can result from three actions by microorganisms: direct metabolism (the pesticide is used as an energy source by microorganisms), co-metabolism (the organisms grow at the expense of a co-substrate to transform the pesticide without deriving any nutrient or energy for growth from the process), and synthesis (conjugation and oligomerization). The final step of pesticide biodegradation is mineralization. If the microorganisms do not possess all the enzymatic equipment to mineralize pesticides, metabolites are excreted into the soil and can be used by other microorganisms. The biotic degradation of pesticides in soils depends on physical, physicochemical and biological factors, together with pesticide properties and agricultural practices.

The main physical factors influencing the biodegradation of pesticides are soil temperature and water content which depend on the climate. In general, biodegradation increases with soil temperature to a limit that is compatible with microbial activity. The optimal water content for aerobic microbial activity is close to field capacity. Soil structure also plays an important role by favoring, or restricting, the circulation of water and air. For example, compaction from farm traffic and soil tillage leads to changes in soil physical properties such as a decrease in porosity, oxygen content and water infiltration that can change the biological functioning of soil and create anaerobic conditions. Soil texture, pH, OM content, and redox potential have an effect on pesticide degradation because they influence their sorption (pesticides have to be in the liquid phase to be degraded) and the activity of microorganisms.

¹ DT_{50} values correspond to degradation but often include other processes involved in the disappearance of the compounds and in that case DT_{50} is the dissipation half-time.

The biodegradation of pesticides depends on the microorganisms' specificity (some pesticides can be degraded by a high diversity of microorganisms in contrast to others), on their adaptation (see below), and on their spatial heterogeneity (microbial biomass is often higher at the soil surface than in the deeper soil layers, and degradation tends to increase when the sizes of soil aggregates increase) (Vieublé-Gonod et al., 2003). The spatial variability of mineralization at the aggregate scale is very large (Fig. 3).

The molecular properties of pesticides control the ability of microorganisms to degrade a compound. In general, degradation decreases when the molecular weight, the branching and or the number of halogen atoms of the molecule increase (Mamy et al., 2015).

Agricultural practices based on monoculture favor repeated applications of the same pesticides that induce adaptation of the microflora resulting in the accelerated degradation of pesticides. The decrease in soil tillage in conservation agriculture, leading to the accumulation of OM at the soil surface, has contrasting effects on pesticide degradation (Alletto et al., 2010): competition between pesticide retention on the mulch (decreasing the bioavailability) and microbial degradation can decrease degradation but, conversely, the presence of OM can stimulate the microbial activity which favors pesticide degradation. During pesticide degradation, the formation of non-extractable residues is frequently observed. It can be considered a stabilization process that is responsible for pesticide immobilization in the soil (Fig. 4).

Pesticide non-extractable (bound) residues

All pesticides arriving to the soil form non-extractable (bound) residues (NER) (Gevao et al., 2000). The NER represent compounds in soils which persist as the parent substance or its metabolite(s) after extraction; the extraction method must not substantially change the compounds themselves or the soil structure. The proportion of NER is operationally defined by the extraction procedure employed, and this can also indicate the mechanisms involved in their formation: residues strongly sorbed or entrapped, covalently or biogenically bound residues derived from biotic degradation and incorporated into the microbial biomass (Kästner et al., 2014). For most pesticides, the formation of NER is usually correlated to the soil's biological activity and to the amount of OM in the soil.

A comprehensive review of the pesticide registration data showed that about 50% of the pesticides exhibit a low proportion of NER (less than 30% of the initial amount) while only 12% of pesticides have a proportion of NER that exceeds 70% (Barriuso et al., 2008). The largest proportion of NER was found for carbamates, and in particular for dithiocarbamates. Organophosphates are the pesticides that form the least NER (Fig. 5). In general, pesticides or metabolites supporting 'free' reactive chemical groups, such as

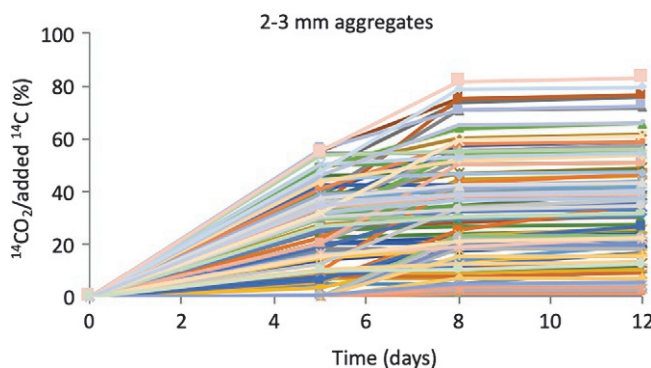


Fig. 3 Spatial variability of 2,4-D mineralization at the mm aggregate scale. Adapted from Vieublé-Gonod L, Chenu C, Soulas G (2003) Spatial variability of 2,4-dichlorophenoxyacetic acid (2,4-D) mineralisation potential at a millimetre scale in soil. *Soil Biology & Biochemistry* 35: 373–382.

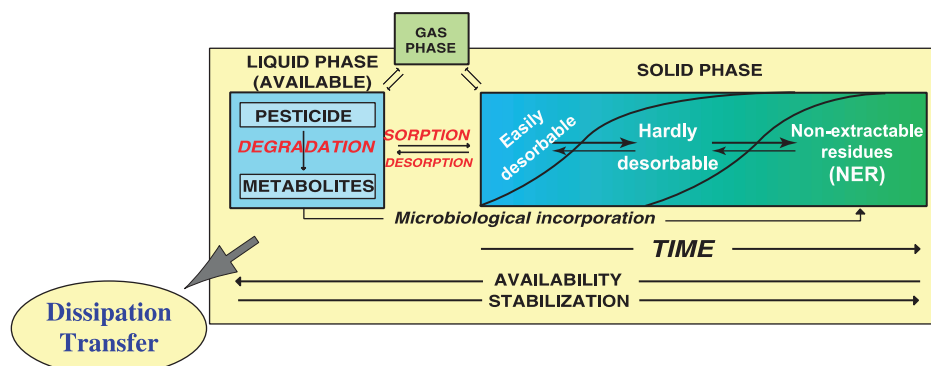


Fig. 4 A dynamic view of the main processes involved in pesticide sorption and degradation in soils.

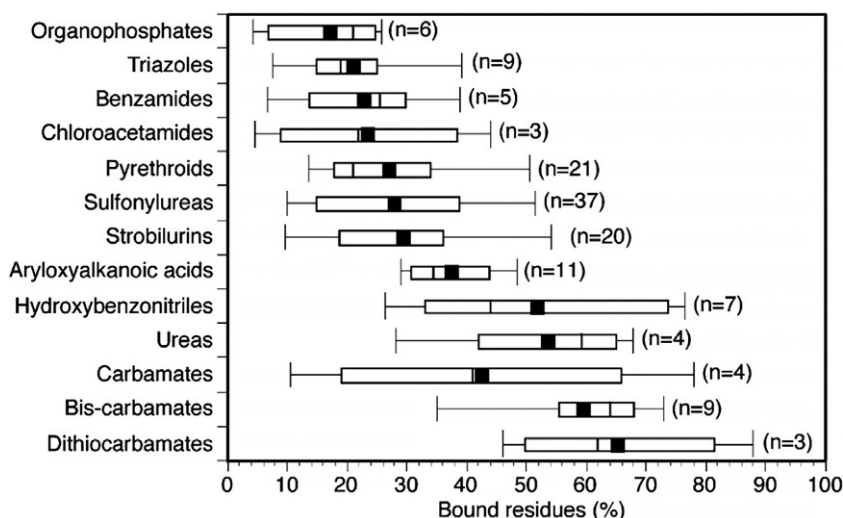


Fig. 5 Classification of pesticide families according to their capacities to form non-extractable (bound) residues. Plotted data correspond to the proportion of NER formation reported in the EU assessment reports and concern experiments with a target duration of ca. 100 days. From Barriuso E, Benoit P, Dubus I. (2008) Formation of pesticide nonextractable (bound) residues in soil: Magnitude, controlling factors and reversibility. *Environmental Science & Technology* 42: 1845–1854.

aniline or phenol, tend to produce a larger proportion of NER. The degradation of pesticides that produce metabolites with hydroxyl or amine groups leads to an increase in NER formation by covalent bounding because their metabolites are more reactive than the parent compound (Benoit et al., 1999). The result of these different reactivities between the parent and its metabolites is a competition between degradation and the formation of NER. If the parent compound is involved in NER formation, then rapid degradation competes with NER formation. If, on the other hand, the metabolite is involved in NER formation, then rapid degradation may lead to considerable formation of NER. Pesticides with a large number of electronegative substituents, such as the halogens, tend to form fewer NER than compounds with a smaller number of substituents. This is an illustration of the importance of degradation pathways and intermediate degradation products on NER formation.

Most of environmental factors affecting biological activity, such as temperature or soil moisture content, are likely to have an influence on NER formation. There is therefore an apparent competition between pesticide mineralization and NER formation. In general, NER amounts increase with increasing water content until the soil reaches saturation, and also increase with temperature. The nature and the amount of soil OM also influence NER formation (Fig. 6). In general, an increase in OM content leads to an increase in NER, as demonstrated for compost additions. Frequently pesticides incubated on mulch or fresh organic matter resulted in a higher proportion of NER than those incubated directly in soils, and the increase in the humification degree of the plant residues increased the proportion of NER (Benoit et al., 1999).

Bioavailability and uptake by plants or earthworms represent only a small percentage of the total amounts of NER; the formation of NER leads to a decrease in the toxicity and bioavailability of pesticides. However, a proportion of NER could be released by different natural mechanisms: ionic modifications and the addition of N-ammoniacal fertilizers to the soil can induce a partial release of some NER.

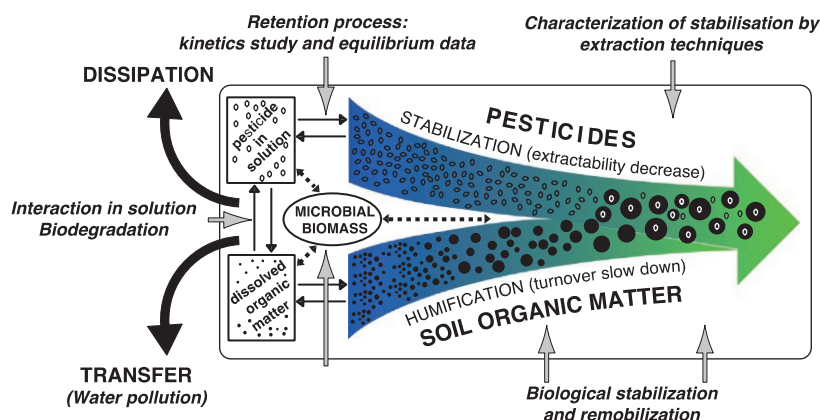


Fig. 6 Hypotheses of formation of pesticide non-extractable residues by incorporation into soil organic matter during the humification process.

Soil column leaching experiments have demonstrated that NER are usually concentrated at the top of the column, indicating the poor leaching capacity of these residues. Although the bulk of residues was confined to the first few centimeters of the soil columns, it should be noted that their extractability decreased with depth and that significant proportions of NER remained at the bottom of the columns (Benoit et al., 2000). The exact origin of these NER deep in the column is difficult to establish. They could result from the leaching of NER formed near the soil surface, indicating an intrinsic or facilitated mobility, or from the formation of NER at various depths following leaching of the parent pesticide and its metabolites down the column.

In summary, gaining detailed information of how NER are formed and subsequently released is essential if they are to be included in environmental risk assessments. However, the explicit consideration of NER in environmental risk assessment would require more intensive long-term experiments. The formation of NER can be interpreted conceptually by flow mass using an approach based on kinetic compartment models including compartments involving different mechanisms of NER formation (Fig. 7, Kästner et al., 2014). The concepts of 'slow sorption' can be easily implemented in pesticide fate models through a kinetics approach. This slow sorption could evolve to an irreversible sorption pool by suppressing desorption from the slow sorption sites, or by adjusting equilibrium constants of these sites to high values, thereby creating 'restrictive sites' or sites with irreversible retention. The main issue with this kind of approach is the lack of information regarding the exact chemical nature of NER.

Effects on soil biota and uptake by plants

Uptake of pesticides by plants

Pesticides can be transferred to vegetation from the soil and or water by uptake through the roots (i.e., symplastic way), from the atmosphere by their aboveground parts after wet or dry deposition, or following rain-splash in which soil particles are dispersed on to leaf surfaces. For atmospheric contaminants, the predominant initial site of interception is the cuticle, where adsorption is the first step in the transfer of volatile and non-volatile organic compounds to the plant. The absorption of pesticides by plants was shown to be mainly driven by molecular properties related to the size of the compounds (branching and steric descriptors) (Mamy et al., 2015).

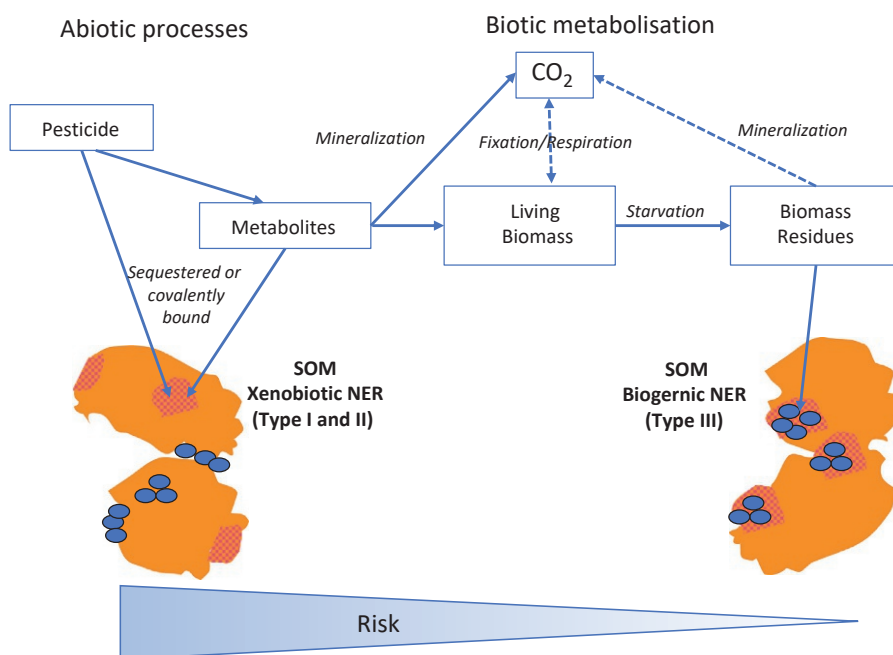


Fig. 7 Schematic diagram of abiotic and biotic non-extractable residues (NER) formation including microbial degradation of organic compounds in soil: Type I (sequestered) and Type II (covalently bound) NER are formed mostly by abiotic processes such as oxidation with reactive oxygen on catalytic surfaces, non-enzymatic hydrolysis or photodegradation on the soil surface, and finally sorption, entrapment, or covalent bonding to soil organic matter (SOM). In addition, these reactions may also be triggered by a microbial radical forming exoenzymes (e.g., laccases and peroxidases). With increasingly productive (biomass formation) biodegradation of the compounds, the relevance of bioNER formation also increases. BioNER formation by heterotrophic CO₂ fixation has to be taken into account if labeled CO₂ is produced during microbial degradation processes. Adapted from Kästner M., Novack KM, Miltner A, Trapp S, Schäffer A. (2014) Classification and modelling of nonextractable residue (NER) formation of xenobiotics in soil—A synthesis. *Critical Reviews in Environmental Science and Technology* 44: 2107–2171.

Effects on soil organisms (microflora and macrofauna)

The persistence of pesticide residues in soil is identified as a major threat for soil dwelling organisms that support many ecosystem services. Leading reviews of the literature have already addressed the effect of pesticides on soil organisms (Datta et al., 2016), including bacterial populations (Johnsen et al., 2001; Thiour-Mauprivez et al., 2019). Relevant ecotoxicological endpoints for terrestrial organisms (e.g., soil microorganisms, invertebrates such as earthworms and enchytraeids, and also other groups) that support a range of ecosystem services are lacking compared to aquatic organisms. The availability of ecotoxicological parameters is also less for chronic than for acute ecotoxicity endpoints.

Although authorities release pesticides to the market only after their careful and thorough evaluation, the risk assessment for soil organisms is unsatisfactory, particularly for microorganisms for which pesticide toxicity is considered solely by one global test measuring N mineralization (Thiour-Mauprivez et al., 2019). Literature on the effects of pesticides on soil microorganisms suggests that they have only minor or transient effects when they are applied at the recommended doses, but the corresponding processes and mechanisms are still poorly understood. The effects of pesticides on the soil microflora are of importance because many microbial functions are critical to crop production, soil sustainability, and environmental quality. Therefore, the structural and functional characteristics of microbial communities, when they can be measured, will represent indicators of choice for monitoring the impact of pesticides on soil ecosystems and for assessing their biological status. The relationships of the different structures of pesticides on the growth of various groups of soil microorganisms are not easily predicted. Some pesticides stimulate the growth of microorganisms, but others have depressive effects or no effect. For example, diflubenzuron can stimulate the activity of nitrogen fixation bacteria in soil, but methylpyrimitos decreases this activity. Information on microbial populations is obtained by culture-dependent methods. Many microorganisms cannot be cultured, therefore, progress in molecular methods and omics tools have opened up large possibilities.

Pesticides may harm soil organisms such as earthworms and enchytraeids, but there is a lack of knowledge on their relative sensitivity to these chemicals and the consequences for soil functions (Pelosi et al., 2021). Although there are several reports on the effects of pesticides on earthworms and enchytraeids (Datta et al., 2016; Römbke et al., 2017), traditionally ecotoxicology has paid little attention to the direct or indirect effects of pesticides on soil functions such as soil OM mineralization. Pelosi et al. (2021) have demonstrated that earthworms no longer promoted soil OM mineralization when fungicide was present at three-fold the recommended field rate. The effects of enchytraeids on soil OM mineralization were similar with and without fungicide exposure.

The treatment of crop seeds with insecticides and or fungicides, aimed at protecting seeds from pests and diseases, is widely used in conventional agriculture. During the growing season, fields often receive additional broadband herbicide applications. Despite this widespread use, little is known about potential side effects or interactions between these different pesticide classes on soil organisms. In a greenhouse pot experiment, Van Hoesel et al. (2017) studied single and interactive effects of seed dressing of winter wheat (*Triticum aestivum* L. var. Capo) with neonicotinoid insecticides and or strobilurin and triazolinthione fungicides, and an additional one-time application of a glyphosate-based herbicide on the activity of earthworms, soil microorganisms, litter decomposition, and crop growth. They showed that seed dressing significantly reduced the surface activity of earthworms with no difference between the use of insecticides or fungicides. Moreover, the effects of seed dressing on earthworm activity were intensified by herbicides (significant herbicide $\times$ seed dressing interaction). They concluded that the interactive effects on soil biota and processes of different pesticide classes should receive more attention in ecotoxicological research.

Integrating process-based knowledge in risk assessment

Pesticide fate models

Numerous models have been developed to simulate the fate of pesticides in the environment following their application to agricultural fields, both at the field and catchment scales (Mottes et al., 2014). The pesticide fate models are considered extensive, economic and efficient tools to predict the fate of pesticides in addition to field and laboratory data. They take into account most of the major processes involved in the fate of pesticides in the environment such as sorption, degradation, leaching, volatilization, plant uptake and wash-off, erosion and runoff, but the processes considered depend on the model (Table 1). The model structure will therefore influence how pesticides are subjected to dissipation transfer processes and how they are distributed into different environmental compartments or into different spatial units.

Among the various existing models, some of the most used are 1D field-scale models that are used for risk assessment of groundwater contamination for pesticides registration at the European level (FOCUS, 2000): MACRO (Model of Water Flow and Solute Transport in Macroporous Soil), PEARL (Pesticide Emission model At the Regional and Local scales), PELMO (Pesticide Leaching Model), and PRZM (Pesticide Root Zone Model). These models differ in their description and consideration of the various processes. For example, in PRZM and PELMO, the soil hydrology is based on a 'tipping-bucket' approach where water will only percolate to the deeper soil layer if field capacity is exceeded, while MACRO and PEARL implement the Richards equation to simulate the water flow. Solute transport used to be described by the advection-dispersion equation.

The soil is generally considered as homogeneous with one porosity form, except in MACRO which describes preferential flow processes by dividing the soil pore system into two flow domains, micropores and macropores. Sorption is considered an instantaneous process and is simulated through linear, and Freundlich formalisms. First-order degradation kinetics of pesticides is assumed by the four models, although PRZM also enables the use of a bi-phasic equation for this process. MACRO simulates

Table 1 Summary of the most important processes simulated by PELMO, PRZM, PEARL, MACRO (FOCUS, 2000) and HYDRUS (Šimůnek et al., 2008) models at the local (soil profile or plot) scale.

	<i>PELMO</i>	<i>PRZM</i>	<i>PEARL</i>	<i>MACRO</i>	<i>HYDRUS</i>
<i>Dimension</i>	1D	1D	1D	1D	1D, 2D or 3D
<i>Water</i>					
Hydrology	Tipping-bucket approach	Tipping-bucket approach	Richards equation Capillarity	Micropores: Richards equation Macropores: Gravity flow Capillarity Preferential flows	Richards equation Dual porosity: non-equilibrium mobile-immobile model (MIM)
<i>Pesticides</i>					
Sorption	Linear or Freundlich Instantaneous, f(pH)	Linear or Freundlich Instantaneous, reversible or f(time)	Linear or Freundlich Instantaneous or non-equilibrium	Linear or Freundlich, instantaneous or f(time), in micro- and macropores	Linear, instantaneous Two-site sorption in the mobile zone
Degradation	1st order	1st order or bi-phasic	1st order	1st order in micro- and macropores	1st order
Volatilization	From soil and plant surface	From soil and plant surface	From soil and plant surface	Empirical relationship or global coefficient (leaves)	Equivalent stagnant boundary Layer depth
Plant uptake	f(transpiration)	f(transpiration)	f(transpiration)	f(transpiration)	f(transpiration)
Solute transport	Advection-dispersion	Advection-dispersion	Advection-dispersion	Micropores: Advection-dispersion Macropores: Convection	Advection-dispersion or MIM
Erosion	Modified Universal Soil Loss Equation	Modified Universal Soil Loss Equation	–	–	–
Runoff	Curve number approach	Curve number approach	–	–	–
<i>Temperature</i>					
Soil temperature	Empirical model using air temperatures	Heat conduction equation	Heat conduction equation	Heat conduction equation	Heat conduction-advection equation

degradation and sorption processes in both micro- and macro-pore domains. All four models can take into account the effect of soil moisture content and soil temperature on the rate of pesticide degradation. PEARL, PELMO and PRZM can simulate the pesticide volatilization, while MACRO does not include a comprehensive description of this process. PRZM and PELMO simulate soil erosion and surface runoff in contrast to MACRO and PEARL. The efficiency of these models was tested for different pesticides under several crop and pesticide conditions; for example, MACRO was the most efficient model to simulate the leaching of pesticides to groundwater and their dissipation in a maize (*Zea mays* L.) crop (Marín-Benito et al., 2014).

The 2D and 3D models, such as HYDRUS-2D and HYDRUS-3D (Šimůnek et al., 2008) can simulate water, heat, and solute movement in two- and three-dimensional variably saturated media. In HYDRUS, water flow is solved by the Richards equation while solute transport is modeled with the advection-dispersion equation assuming first-order degradation kinetics of the solute in the liquid and solid phases, and an instantaneous and linear sorption of the solute onto soil solid surfaces. Degradation depends on temperature and water content. The soil hydraulic functions (water content and hydraulic conductivity) are described with the van Genuchten-Mualem model (Šimůnek et al., 2008). HYDRUS-2D can simulate the fate of some pesticides and also the soil water content and fluxes at the plot scale over a pluriannual crop rotation (wheat, maize) accounting for the effects of lateral heterogeneities created by soil tillage and the incorporation of organic amendments (Filipović et al., 2016).

Overall, existing models could be improved by considering, among others, (1) more detailed description of agricultural practices; (2) particulate transfer; (3) refined climatic time scale. Accounting for spatial and temporal scales in pesticide fate modeling remains an important challenge for research. From the point of view of users, i.e. for water resources management, models can be potential tools for three types of action: (1) diagnostic tool to assess the risk of contamination and (2) to identify vulnerable areas for priority protection and or (3) decision tool for choosing adapted measures to limit pesticide transfer at the catchment or regional scale. For example, many studies have been concerned with either vulnerability diagnostics or measures to protect and restore catchments used for drinking water. Nevertheless, there is a critical need for methodological research to evaluate action programs based on the use of spatialized models able to simulate the fate and transfer of pesticides over pluriannual periods. Moreover, such tools have to be able to describe horizontal transfers between plots or landscape elements, which requires the development of territory-scale models that take into account the effect of landscape elements on these transfers, whether hydrological or atmospheric. This remains one of the major challenges of current developments. Treating the question of uncertainties is another important challenge considering both variation in the soil, climate, plants, agricultural practices and other related properties, and the simplification inherent in modeling and the coupling between models which is necessary for integrating spatial and time-dependent processes.

Concluding remarks and the future

There is considerable literature on pesticides, on the processes involved in their fate in soils and on their environmental impacts, mainly concerning ecotoxicological effects and water contamination. The impacts on air and soil contamination have been less well studied. Part of this knowledge has been used to design numerical models describing the fate of pesticides and for risk assessments in the context of registration procedures. The available information on pesticide fate in soils is limited at certain levels of detail because of the large number of different pesticide molecules, and agronomic and environmental situations. It is often difficult to go beyond case studies for a given pesticide and a given situation and to obtain a generic understanding for different situations.

Knowledge gaps and data deficiencies remain, for example, in the consideration of real situations that involve not only active substances, but also formulated commercial products. Standard methods for characterizing adsorption and degradation take little account of the heterogeneity of soils, and also fluctuations in experimental conditions (temperature, water content). Spatial micro-heterogeneities linked to soil structure will condition the accessibility of pesticides to adsorption sites, but also to the microorganisms responsible for their degradation. The behavior of pesticides under extreme conditions, e.g., at very low water contents on the surface of dry soils, are considered very little in the literature, although they may have an important effect on the processes of adsorption, degradation and volatilization.

Finally, the soil processes involved in the fate of pesticides have their own kinetics. Standard methods of characterization tend to focus on short-term or fast processes, which do not describe slow processes adequately, e.g., long-term adsorption, formation of non-extractable residues, or ecotoxicological impacts on biodiversity.

References

- Alletto L, Coquet Y, Benoit P, Heddadj D, and Barriuso E (2010) Tillage management effects on pesticide fate in soils: A review. *Agronomy for Sustainable Development* 30: 367–400.
- Barriuso E and Calvet R (1992) Soil type and herbicides adsorption. *International Journal of Environmental Analytical Chemistry* 46: 117–128.
- Barriuso E, Benoit P, and Dubus I (2008) Formation of pesticide nonextractable (bound) residues in soil: Magnitude, controlling factors and reversibility. *Environmental Science & Technology* 42: 1845–1854.
- Bedos C, Alletto L, Durand B, Fanucci O, Brut A, Bourdat-Deschamps M, Giuliano S, Loubet B, Ceschia E, and Benoit P (2017) Observed volatilization fluxes of S-metolachlor applied on a bare soil and on soil covered by crop residues. *Environmental Science and Pollution Research* 24: 3985–3996.
- Bedos C, Générmont S, Castell JF, and Cellier P (2019) Agriculture and Air Quality. In: *Investigating, Assessing and Managing*. Springer. QUAE.
- Beigel C, Barriuso E, and Calvet R (1998) Sorption of low levels of nonionic and anionic surfactants on soil: Effects on sorption of triticonazole fungicide. *Pesticide Science* 54: 52–60.

- Benoit P, Barriuso E, and Soulas G (1999) Degradation of 2,4-D, 2,4-dichlorophenol, and 4-chlorophenol in soil after sorption on humified and non-humified organic matter. *Journal of Environmental Quality* 28: 1127–1135.
- Benoit P, Barriuso E, Vidon P, and Real B (2000) Isoproturon movement and dissipation in undisturbed soil cores from a grassed buffer strip. *Agronomie* 20: 297–307.
- Calvet R (1989) Adsorption of organic chemicals in soils. *Environmental Health Perspectives* 83: 145–177.
- Chaplain V, Mamy L, Vieublé-Gonod L, Mougin P, Benoit P, Barriuso E, and Nélieu S (2011) Fate of pesticides in soils: Toward an integrated approach of influential factors. In: Stoytcheva M (ed.) *Pesticides in Modern World—Risks and Benefits*, p. 26. Intech. ISBN: 978-953-307-458-0. Chapter 29.
- Das S and Hageman KJ (2020) Influence of adjuvants on pesticide soil–air partition coefficients: Laboratory measurements and predicted effects on volatilization. *Environmental Science & Technology* 54: 7302–7308.
- Datta S, Singh J, Singh S, and Singh J (2016) Earthworms, pesticides and sustainable agriculture: A review. *Environmental Science & Pollution Research* 23: 8227–8243.
- Filipović V, Coquet Y, Pot V, Houot S, and Benoit P (2016) Modeling water and isoproturon dynamics in a heterogeneous soil profile under different urban waste compost applications. *Geoderma* 268: 29–40.
- FOCUS (2000) *FOCUS Groundwater Scenarios in the EU Review of Active Substances*. Report of the FOCUS Groundwater Scenarios Workgroup, EC Document Reference Sanco/321/2000 rev.2, p. 202.
- Garcia L, Bedos C, Générmont S, Benoit P, Barriuso E, and Cellier P (2014) Modeling pesticide volatilization: Testing the additional effect of gaseous adsorption on soil solid surfaces. *Environmental Science & Technology* 48: 4991–4998.
- Gevao B, Semple KT, and Jones KC (2000) Bound pesticide residues in soils: A review. *Environmental Pollution* 108: 3–14.
- Houbraken M, Senaev D, Dávila EL, Habimana V, De Cauwer B, and Spanoghe P (2018) Formulation approaches to reduce post-application pesticide volatilisation from glass surfaces. *The Science of the Total Environment* 633: 728–737.
- Johnsen K, Jacobsen CS, Torsvik V, and Sørensen J (2001) Pesticide effects on bacterial in agricultural soils—A review. *Biology and Fertility of Soils* 33: 443–453.
- Kästner M, Novack KM, Miltner A, Trapp S, and Schäffer A (2014) Classification and modelling of nonextractable residue (NER) formation of xenobiotics in soil—A synthesis. *Critical Reviews in Environmental Science and Technology* 44: 2107–2171.
- Mamy L and Barriuso E (2007) Desorption and time-dependent sorption of herbicides in soils. *European Journal of Soil Science* 58: 174–187.
- Mamy L, Patureau D, Barriuso E, Bedos C, Bessac F, Louchart X, Martin-Laurent F, Miegé C, and Benoit P (2015) Prediction of the fate of organic compounds in the environment from their molecular properties: A review. *Critical Reviews in Environmental Science and Technology* 45: 1277–1377.
- Marín-Benito JM, Pot V, Alletto L, Mamy L, Bedos C, Barriuso E, and Benoit P (2014) Comparison of three pesticide fate models with respect to the leaching of two herbicides under field conditions in an irrigated maize cropping system. *Science of Total Environment* 499: 533–545.
- Mottes C, Lesueur-Jannoyer M, Le Bail M, and Malézieux E (2014) Pesticide transfer models in crop and watershed systems: A review. *Agronomy for Sustainable Development* 34: 229–250.
- Pelosi C, Thiel P, Bart S, Amossé J, Jean-Jacques J, Thoisy JC, and Crouzet O (2021) The contributions of enchytraeids and earthworms to the soil mineralization process in soils with fungicide. *Ecotoxicology* 3: 1910–1921.
- Prueger JH, Alfieri J, Gish TJ, Kustas WP, Daughtry CST, Hatfield JL, and McKee LG (2017) Multi-year measurements of field-scale metolachlor volatilization. *Water, Air, & Soil Pollution* 228: 84.
- Römbke J, Schmelz RM, and Pelosi C (2017) Effects of organic pesticides on enchytraeids (Oligochaeta) in agroecosystems: Laboratory and higher-tier tests. *Frontiers in Environmental Science* 5: 20.
- Šimůnek J, van Genuchten MT, and Šejna M (2008) Development and applications of the HYDRUS and STANMOD software packages and related codes. *Vadose Zone Journal* 7: 587–600.
- Thiour-Mauprivez C, Martin-Laurent F, Calvayra C, and Barthelmebs L (2019) Effects of herbicide on non-target microorganisms: Towards a new class of biomarkers? *The Science of the Total Environment* 684: 314–325.
- Van Hoesel W, Tiefenbacher A, König N, Dorn VM, Hagenguth JF, Prah U, Widhalm T, Wiklicky V, Koller R, Bonkowski M, Lagerlöf J, Ratzenböck A, and Zaller JG (2017) Single and combined effects of pesticide seed dressings and herbicides on earthworms, soil microorganisms, and litter decomposition. *Frontiers in Plant Science* 8: 215.
- Vieublé-Gonod L, Chenu C, and Soulas G (2003) Spatial variability of 2,4-dichlorophenoxyacetic acid (2,4-D) mineralisation potential at a millimetre scale in soil. *Soil Biology & Biochemistry* 35: 373–382.

Soil organic matter in molecular simulations

Martin H Gerzabek^a, Daniel Tunega^{a,b}, Edgar Galicia-Andrés^{a,c}, and Chris Oostenbrink^c, ^aDepartment of Forest and Soil Sciences, Institute of Soil Research, University of Natural Resources and Life Sciences Vienna, Vienna, Austria; ^bSchool of Pharmaceutical Science and Technology, Tianjin University, Tianjin, China; ^cDepartment of Material Sciences and Process Engineering, Institute of Molecular Modeling and Simulation, University of Natural Resources and Life Sciences Vienna, Vienna, Austria

© 2023 Elsevier Ltd. All rights reserved.

Introduction	483
Computer simulation of reality	484
Molecular modeling methods	485
Quantum chemistry	486
Wave function methods	487
Density functional theory—Basic principles	487
Force field methods	488
Force field parameterization	490
Examples for applications in soil science	490
Modeling of soil organic matter—The Vienna soil organic matter modeler	491
Interaction of SOM with reactive mineral surfaces	493
Organic contaminant interactions with mineral surfaces and organic matter	494
Conclusions and outlook	495
Acknowledgments	495
References	496

Abstract

Molecular modeling (MM) methods achieved an enormous success in fields such as chemistry or pharmacy. The first attempts to use MM methods in soil science are about 25 years old; however, their wider applicability in soil research is still limited mainly due to extreme soil complexity and diversity. This chapter introduces the most important methodological approaches of quantum mechanics and force field methods and provides several examples evidencing the usefulness and applicability of the MM methods in the soil science. The examples comprise modeling of soil organic matter, the simulation of organo-mineral interactions and adsorption of organic pollutants to soil components.

Glossary

Ab initio modeling Computational chemistry methods based on quantum chemistry, aiming at the approximate solution of the Schrödinger equation.

Force-field methods A group of methods aiming at calculating forces between atoms in a molecule or between molecules. The required parameters can be derived from physical or chemical measurements or from quantum chemical calculations.

Key points

- Discuss why detailed modeling on the molecular level is useful in soil science.
- To show which methods of molecular modeling (MM) can be applied in soil research.
- Potentials and limitations of MM.
- Modeling of the highly complex soil organic matter (SOM).
- Examples on using MM in modeling of interactions of SOM with reactive mineral surfaces and adsorption of organic species in soil.

Introduction

Soil is a highly complex system. Mineral and organic soil components form a unit with soil organisms, plant roots, soil water and soil air. The physical, chemical, and biological processes taking place in the soil are strongly modified by environmental conditions and are thus dynamic, i.e., subject to continuous change. Viewing the soil as a black box permits the derivation of statistical models about the relationships between soil conditions, but it is not possible to understand the fundamental processes. Many of these

processes, such as the binding of pollutants and nutrients in the soil, the stabilization of soil organic matter (SOM), the wettability of soil with water, or also changes in the chemistry due to the activity of soil organisms and plant root secretions, take place on the molecular scale. Analytical methods have made enormous progress in soil research over the past decades. Today, many of the latest physical, chemical and biological investigation methods can be used. They have been adapted for use in soil research and can describe reactions and processes in great detail. For example, X-ray photoelectron spectroscopy (XPS) methods can now be used to investigate elements found in soil surfaces (layer thickness of approx. 10 nm) on a microscopic scale. Atomic Force Microscopy can elucidate adsorption processes by measuring the interaction energy under some conditions. However, even today's cutting-edge analytical methods only provide limited information about probable conformations of organic molecules and their stability in the soil. Also, the detailed interaction mechanisms of inorganic or organic substances with reactive soil surfaces and especially their stability are still not fully accessible to highly specialized analytics—and therefore, not entirely understood. This is where molecular modeling (MM) and simulation can come in. The MM methods, due to the enormous developments in the power of computers, are now able to describe in detail systems that range from the very small atomic scale, by producing approximate solutions to the Schrödinger equation (SE), to larger ones with a few thousand atoms, with methods that require much less computer resource in relation to the size of the systems. The great challenge here is to bridge the spatial and temporal scale boundaries, and to use the results of molecular modeling to explain phenomena observed microscopically and macroscopically in soil.

The use of MM methods in soil research is now much wider than 20 years ago. Since then, many successful results have been achieved, as had been predicted by Gerzabek et al. (2001). By providing a detailed molecular insight, MM has contributed to the understanding of soil experimental measurements correlating the structure, physicochemical properties, and functions of the soil and of soil components (Schaumann, 2006). The importance of MM lies in the processing of interactions between soil components, their environment, and other substances present (Devarajan et al., 2020). As an example, modeling of SOM in the presence of soil minerals has contributed to elucidate the stabilization and protection mechanisms of organic matter (Loganathan et al., 2020; Zhang et al., 2020). Likewise, modeling of SOM with different metal ions has led to the understanding of processes that control speciation, solubility and toxicity at the molecular scale (Liang et al., 2019). A particular subject of interest was prediction of the bioavailability of pollutants in soil such as agrochemicals (Haberhauer et al., 2001), drugs (Aristilde and Sposito, 2010) and waste products (Lan et al., 2019).

In this chapter, we show how MM methods can be used and of what they are capable, based on the methodological background and individual case studies. Soil organic matter is well studied in terms of its elemental composition, functional groups, chemical and physical properties, as well as its functions and processes. However, there are still large gaps regarding the structure and conformation of SOM, which is a mixture of an incredibly large number of molecules that are probably different from each other. This is where MM can make a difference. In particular, based on the existing state of knowledge, computer methods can be used to calculate models of molecules that correspond to the observations in terms of their properties. Particularly important questions that MM can help to solve are, for example, (i) the stability of organic matter and the formation of supramolecular structures, (ii) the interaction of organic molecules and associations with reactive mineral surfaces, such as clay minerals and pedogenic oxides, and (iii) the interaction of organic contaminants with mineral surfaces and soil organic matter.

Computer simulation of reality

Long before the advent of computers, scientists have been making models of reality. In this context, the word model may refer to physical models of objects that are either too small or too large to study in reality. One can think of a physical molecular model, like the structures of DNA that were built by Watson and Crick, or a planetarium describing the motions of the planet. A model can also take the form of a philosophical model, that describes the current thinking about complex processes or the composition of complex mixtures, like soil. Our view of the real world is quite limited but becomes accessible through models. Fig. 1 shows schematically how experiments may be performed in the real world, leading to experimental data (van Gunsteren et al., 2006). These data enable

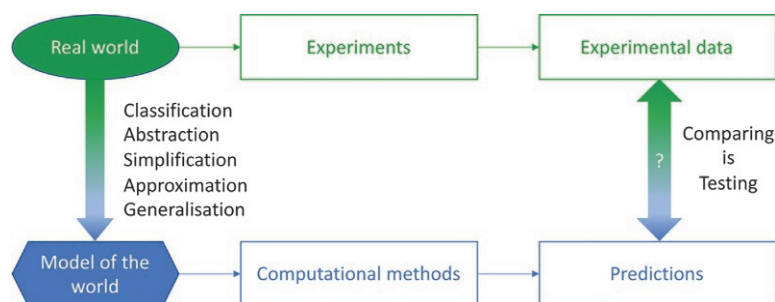


Fig. 1 Computational soil science involves the formulation and testing of mathematical models of the real world. Adapted from van Gunsteren WF, Bakowies D, Baron R, Chandrasekhar I, Christen M, Daura X, Gee P, Geerke DP, Glättli A, Hünenberger PH, Kastenholz MA, Oostenbrink C, Schenk M, Trzesniak D, Van der Vegt NFA and Yu HB (2006) Biomolecular modeling: Goals, problems, perspectives. *Angewandte Chemie International Edition* 45: 4064–4092. (At: <https://onlinelibrary.wiley.com/doi/10.1002/anie.200502655>).

classifications and abstractions, which together with simplifications, approximations and generalizations lead to a model of the real world. If such models can be digitized, we obtain a computer model. Mathematical methods can subsequently be applied to obtain predictions, which can once more be compared to (independent) experimental data to validate or falsify the model.

The main reason to create molecular models of soil components is because molecules are the smallest entities that are, in the end, responsible for the behavior of virtually all physical processes. A true understanding of properties and processes of soil will only be possible at the molecular level. However, the experimental data that are obtained in the real world, contain very little information about the behavior of individual molecules. Rather, these data are representative of averages over many complex molecules and over time scales that are long compared to typical molecular motion. To obtain a visualization and an in-depth understanding of the relevant processes, one can turn to molecular models. It is important to note, therefore, that computational modeling and real-world experiments are complimentary in their description of the real world. Macroscopic properties may be described phenomenologically, with the physical basis being understood from much simplified models. One of the main challenges is to bridge the space and time-scale gap between the experiments in the real world and the models. In other words, to ensure that the simplifications and approximations made in the model building are appropriate for the aspects of physical processes that one aims to understand.

Molecular modeling methods

This section briefly summarizes molecular modeling methods of computational chemistry that are applicable in soil sciences. Details can be found in many excellent text books and reviews available in the literature (Cramer, 2004; Frenkel and Smit, 2002; Leach, 2001; Young, 2001).

Depending on the approximations and simplifications one is willing to make, various models at different levels of resolution can be constructed. Such models will differ in the elementary particles that are being treated. Three relevant examples for molecular models of a single molecule are given in Fig. 2.

Generally, molecular modeling comprises two basic groups of methods built on principles of (a) quantum chemistry (QC) and (b) force field (FF) methods, respectively. The QC methods describe systems as a set of mutually interacting nuclei and electrons, and complete information about the system can be achieved by an approximate solution of the SE. In the FF methods, classical particles represent a model system consisting of interacting particles (e.g., atoms, or larger beads representing a functional group or set of atoms) and the main task is to find proper, relatively simple analytical expressions able to describe a potential energy hypersurface of the system under study. The smallest degree of freedom that is taken into account in FF is no longer an electron, but we describe the system in terms of entire atoms that are connected to each other in some way. This means that we are no longer able to break chemical bonds. In a way, we are trying to capture the average effect of the electrons in the force field equations to come to a coarser description of the model. We can continue to make our description coarser and try to average over additional degrees of freedom (see Fig. 2). We would need to capture the mean effect of the neglected degrees of freedom within specific FF contributions. Examples of this are FFs that describe the presence of a solvent in an implicit way, or FFs in which the elementary particles are beads that are representative of multiple atoms, e.g., an entire functional group, or a set of 4–5 water molecules. The advantage could be that fewer degrees of freedom are explicitly calculated, reducing the computational effort at the cost of a loss of detail in the

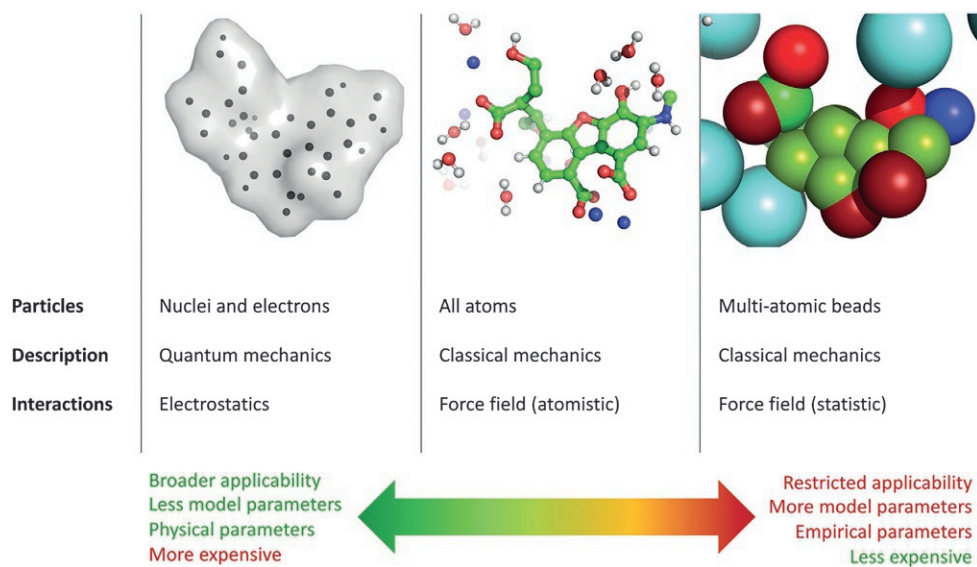


Fig. 2 Three representations of the same organic molecule.

description. Such FFs typically use very similar functions to those that will be described below, but will not be discussed explicitly. Nevertheless, the decision as to what level of description will be most appropriate is a crucial one in any modeling effort, posing such questions as follow. Are the electronic degrees of freedom necessary, or is an atomistic description of the system sufficient? Can the degrees of freedom be reduced even further, or would we lose relevant details if we make our description even coarser? The sections below discuss QC and atomistic FF methods in more detail.

Quantum chemistry

Fig. 3 shows a basic classification of QC methods. The core of quantum chemistry is the SE that fully describes the space- and time-evolution of a system consisting of interacting electrons and nuclei. Mathematically, the SE is a second-order partial differential equation. Its non-relativistic time-independent form is

$$\hat{H}\Psi = E\Psi \quad (1)$$

where $\hat{H}$ is the differential Hamiltonian operator, E is the total energy, and Ψ is the many-body wave function of the system. The Hamiltonian operator is the sum of the kinetic and potential energy operators

$$\hat{H} = -\sum_i \frac{1}{2m_i} \nabla^2 + V(\mathbf{R}, \mathbf{r}) \quad (2)$$

where m_i is the mass of the i th particle (electron or nucleus), ∇^2 is the Laplacian operator and $V(\mathbf{R}, \mathbf{r})$ is the operator describing the potential energy representing the Coulombic interactions of all n_e electrons and all n_N nuclei. $\mathbf{R}$ and $\mathbf{r}$ are position vectors of nuclei and electrons, respectively. Thus, the total energy E in Eq. 1 is the sum of the kinetic energy of all electrons and nuclei, and interaction energies between all electrons ($e-e$), nuclei ($N-N$), and electron and nuclei ($e-N$) (Eq. 3).

$$E = T_e + T_N + V_{ee} + V_{eN} + V_{NN} \quad (3)$$

The existence of a wave function Ψ is one of the postulates in quantum chemistry. The square of the wave function is interpreted as the probability of finding an electron in a configurational space. Generally, an exact solution of the SE for a many-body system is not possible and practical solutions require approximations. The most common is the *Born–Oppenheimer (BO) approximation* that assumes the SE solution separately for nuclei and electrons because electronic motion is much faster than nuclear motion (nuclei are much heavier than the electrons by a factor ~ 1830). Thus, the electronic wave function only depends parametrically on the positions of the nuclei ($\hat{H}(\mathbf{r}, \mathbf{R}) \rightarrow \hat{H}_{BO}(\mathbf{r} | \mathbf{R})$; $T_N = 0$ and V_{NN} is a parameter in Eq. 3). The next two approximations leading to the practical solution are (a) application of the *variational principle* and (b) expression of the many-electron wave function as a product of one-electron wave functions (*Hartree–Fock* method, HF). The variational principle supposes we select a trial wave function that depends on one or more parameters, $\mathbf{p}$, $\Psi(\mathbf{r} | \mathbf{R}, \mathbf{p})$. By varying these parameters, it is possible to find a minimum for $E[\Psi]$ (eigenvalue). In other words, to fulfill the condition for the stationary point, the following should be true.

$$\frac{\partial E[\Psi]}{\partial p_i} = 0 \quad (\text{for all } i) \quad (4)$$

If Ψ is exact, then the solution is the exact energy for any other Ψ , $E[\Psi] \geq E_{\text{exact}}$.

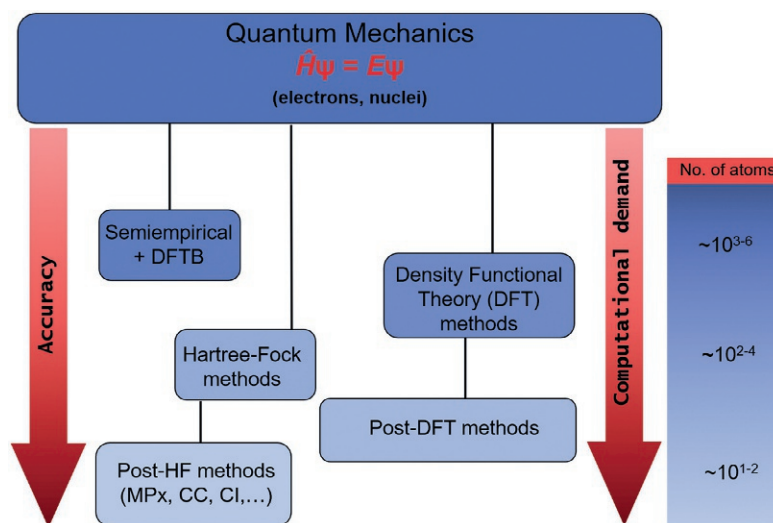


Fig. 3 Schematic categorization of the quantum chemical methods.

Wave function methods

The concept of the HF method is based on a specific approach of “independent particle model” meaning that the trial wave function is expressed as a product of one-electron wave functions, ψ_i , (called molecular orbitals).

$$\Psi_{\text{HF}} = |\psi_1 \psi_2 \cdots \psi_i \cdots \psi_n| \quad (5)$$

The Ψ_{HF} should fulfill the Pauli exclusion principle (sign is changed if two electrons are permuted). The HF solution means that each individual electron moves in an averaged field of all the remaining electrons and a mutual correlation of electron movement is neglected (mean field model).

For a practical solution, the orbital wave functions are expressed in a set of auxiliary functions (called atomic orbitals—basis set) as

$$\psi_i(\mathbf{r}) = \sum_{\mu} c_{\mu i} \phi_{\mu}(\mathbf{r}) \quad (6)$$

where a molecular orbital ψ_i is represented as the sum of n atomic orbitals ϕ_{μ} (μ runs from 1 to n) and $c_{\mu i}$ are unknown linear expansion coefficients. ϕ_{μ} represents the basis set of the i th molecular orbital. The variation of wave function with respect to these coefficients leads to the set of Hartree-Fock Roothaan nonlinear equations

$$F\psi_i = \varepsilon_i \psi_i \quad (7)$$

where F is a Fock operator (FO) and ε_i is the expectation value over the Fock operator representing the average energy of electrons in the i th orbital and the total energy of the system is the sum of all orbital energies $E_{\text{tot}} = \sum \varepsilon_i$.

The FO depends on its own eigenfunctions; thus, the solution of Eq. (7) requires an iterative approach—(a) estimate initial orbital coefficients → (b) calculate FO → (c) diagonalize FO to obtain new orbitals → (d) calculate E_{tot} → (e) check energy convergence—if E_{tot} is converged, stop, else go back to step (b).

Although the HF solution is approximative, it covers >99% of the exact energy. Unfortunately, the missing part (correlation energy, CE) can lead to large errors because in chemistry we require accuracy of $\sim 1 \text{ kcal mol}^{-1}$. Therefore, to improve accuracy of the HF solution required the development of methods able to include correlation energy in the calculation. These methods (post-HF methods) are categorized into several groups such as (i) Configuration Interaction (CI) methods, (ii) Møller–Plesset perturbation theory (MPx) methods, and (iii) Coupled Cluster (CC) theory methods. The form of CC, which includes singles and doubles and non-iterative triples (CCSD(T)), is considered a very accurate method frequently called the “gold standard” in quantum chemistry.

A decision to select a particular ab initio method is affected mainly by the size of the system (number of electrons, n), property(ies) to be calculated, computational time required and available computational resources. The wave function-based methods discussed above are computationally scaling with n^4 and more; thus, the size of the system is an important factor. A special category of the QC methods is represented by semiempirical and density functional tight binding (DFTB) methods suitable for systems of thousands of atoms. In semiempirical methods some complicated integrals are replaced by empirical parameters reducing the computational time substantially with scaling n^{2-3} ; however, the price for that in many cases is a decrease in accuracy.

Density functional theory—Basic principles

This section briefly describes principles of the second principal group of quantum chemistry based on the density functional theory (DFT). The DFT-based methods are very popular due to their economy and efficiency compared to the post-HF methods because they scale with $\sim n^3$. Thus, these methods are applicable for large molecular systems consisting of hundreds of atoms. Apart from HF and post-HF methods, energy of the system, $E(\rho)$, is expressed as a function of the three-dimensional electronic density of the system, ρ . The method, appeared in the 1960s and is based on two Hohenberg–Kohn theorems and the practical Kohn–Sham approach.

The first theorem supposes that the electron density completely determines the Hamiltonian operator and, consequently, all the properties (including energy) of the system. The total energy functional in the Hohenberg–Kohn–Sham (HKS) approach can be expressed as

$$E[\rho] = T[\rho] + E_{\text{ee}}[\rho] + \int \hat{V}_{\text{ext}} \rho(\mathbf{r}) d\mathbf{r} = F[\rho] + \int \hat{V}_{\text{ext}} \rho(\mathbf{r}) d\mathbf{r} \quad (8)$$

where $T[\rho]$ and $E_{\text{ee}}[\rho]$ are unknown functionals of the kinetic energy and electron–electron interaction, and $F[\rho]$ is the universal DFT functional. The last term in Eq. (8) represents the electron interactions with the effective external potential.

The second theorem claims that the energy density functional is variational meaning that it has a minimum with respect to variations in the electron density, $\delta\rho(\mathbf{r})$.

These two theorems mean that there is a universal energy density functional $E[\rho]$, which, if it is known, can be varied with respect to the electron density to compute the exact ground state density and energy.

For a practical solution, the universal functional was reformulated as

$$F[\rho] = T[\rho] + J[\rho] + E_{\text{XC}}[\rho] \quad (9)$$

where $J[\rho]$ is a Hartree (classical electrostatic repulsion term) and $E_{xc}[\rho]$ is the exchange-correlation energy containing everything that is unknown. The success of the DFT solution lies in finding proper expressions for kinetic energy and exchange-correlation energy terms.

For example, for a fictitious system of non-interacting electrons the kinetic energy functional can be expressed as a single determinant of n one-electron wave functions (Kohn–Sham orbitals), ϕ_i :

$$T_s[\rho] = -\frac{1}{2} \sum_i^n \langle \phi_i | \nabla^2 | \phi_i \rangle \quad (10)$$

Using this approximation, the electron density can be expressed as

$$\rho(\mathbf{r}) = \sum_i^n |\phi_i(\mathbf{r})|^2 \quad (11)$$

Similarly, as in the HF approach, such HKS reformulation leads to non-linear equations describing non-interacting particles in an effective local potential

$$\left(-\frac{1}{2} \nabla^2 + V_s(\mathbf{r}_i) \right) \phi_i = \varepsilon_i \phi_i \quad (12)$$

where ε_i are energies of corresponding KS orbitals ϕ_i . The practical solution is done in an iterative manner, similar to the HF solution.

The major problem in DFT theory is with the expression of functionals for exchange and correlation energies. This led to the development of numerous solutions starting with the simplest local density approximation (LDA), through the generalized gradient approximation (GGA) method that includes gradients of the electron density to the functional forms, to the hybrid and meta-GGA approaches. From numerous functionals of the GGA theory, we can mention popular functionals such as PBE (Perdew–Burke–Ernzerhof), PW91 (Perdew–Wang 1991), or BLYP (Becke–Lee–Yang–Parr). Details on DFT functionals can be found in many textbooks. From the hybrid functionals, the most popular in the chemical community is the B3LYP that mixes the HF exact exchange with the DFT exchange–correlation using adjustable parameters. Its popularity arises from provision of relatively good accuracy in predicted properties of many chemical systems. From the beginning of DFT theory it was found that DFT functionals can frequently fail in a prediction of properties such as charge–transfer, excited states, or weak dispersion interactions. The main reason is the inability to account for a nonlocal electron correlation effect. Thus, there is a systematic development of DFT functionals dealing with this problem. This effort brings new methods at several levels of theory—from purely empirical corrections to the standard DFT functionals to rigorous dispersion functionals derived from first principles. Thus, the main message for the practical application of DFT methods is in the understanding of basic principles, limiting accuracy of the particular DFT functionals and their applicability and suitability with respect to the chemical nature of the modeled system.

Force field methods

We mentioned above the Born–Oppenheimer approximation, which states that from the electron’s point of view, the nuclei seem to be standing still and the electron cloud around them will be adjusted instantaneously. The positions of the nuclei are therefore parameters of the Hamilton operator, and the resulting electronic energy is a function of the nucleic coordinates. Given a certain configuration of the nuclei (their positions), the electrons will adjust instantaneously and this leads to a certain energy. This gives us a direct link between the nucleic coordinates and the energy of the system. If the nuclei are positioned in an unfavorable configuration, the electrons cannot adjust around them and this will lead to relatively high energy, while a molecule that is in a favorable configuration, will lead to a lower energy. From the point of view of the nuclei, the Born–Oppenheimer approximation implies that they move according to a potential energy that is given by the electrons and is a function of the atomic coordinates.

Calculation of the electronic wavefunctions and associated energies may be computationally quite demanding, but they are crucial in cases where the electronic degrees of freedom are important. For instance, if one is interested in chemical reactions in excited electronic states (e.g., to study fluorescence) or in electron transfer reactions, the electrons should be taken explicitly into account. However, in many systems, we are interested in how the molecules behave, but not particularly in the electronic degrees of freedom. We can then try to approximate the potential energy of the molecular system with a simplified equation that captures the mean effect of the electronic degrees of freedom. Such an approximation is what we call a force field (FF), and it typically involves various energy terms that describe certain aspects of the potential energy the nuclei experience. The calculation of FF energy is usually orders of magnitude faster than solving the SE (Eq. 1), therefore it can be done for a significantly larger number of particles.

A force field defines the total potential energy, $V^{\text{pot}}(\mathbf{r})$, as a function of the atomic coordinates, $\mathbf{r}$. The position of a single particle i is denoted with the three-dimensional vector $\mathbf{r}_i$. $V^{\text{pot}}(\mathbf{r})$ can be written as the sum of various contributions, for example

$$V^{\text{pot}}(\mathbf{r}) = V^{\text{bonded}}(\mathbf{r}) + V^{\text{nonbonded}}(\mathbf{r}) \quad (13)$$

It is common to distinguish the bonded interactions, $V^{\text{bonded}}(\mathbf{r})$, and the nonbonded interactions, $V^{\text{nonbonded}}(\mathbf{r})$. As the name implies, the bonded interactions come from the interactions of atoms that are bound to each other in some way, while the

nonbonded interactions represent the interactions between atoms through space, rather than through some set of chemical bonds. The bonded and nonbonded energy terms are themselves again the sum of various more specific terms that will be outlined below. Additional FF terms may be added to model specific aspects. The definition of a specific force field consists of everything we need to approximate the molecular potential energy. This means that a force field consists of a detailed description of the functional form that is used, together with all the parameters that are needed to compute the energy from the atomic coordinates.

The bonded interactions are typically split further into contributions related to bond stretching, bond-angle bending, torsional dihedral angle rotations and out-of-plane or improper dihedral angle terms,

$$V^{\text{bonded}}(\mathbf{r}) = V^{\text{bond}}(\mathbf{r}) + V^{\text{angle}}(\mathbf{r}) + V^{\text{torsion}}(\mathbf{r}) + V^{\text{improper}}(\mathbf{r}) \quad (14)$$

Most force fields, approximate the potential energy associated with a single bond by a harmonic potential, as in

$$V_i^{\text{bond}}(\mathbf{r}) = \frac{1}{2} K_i^b [b_i(\mathbf{r}) - b_i^0]^2 \quad (15)$$

where K_i^b is the harmonic force constant and b_i^0 is the optimal bond length. These are the parameters of this force field term, and $b_i(\mathbf{r})$ is the actual bond length that can be computed from the atomic coordinates. Bonds are stiff degrees of freedom, allowing for little fluctuation. Assuming an average energy of RT (accounting $\sim 2.5 \text{ kJ mol}^{-1}$ at room temperature, R is an ideal gas constant of $8.31446261815324 \text{ J K mol}^{-1}$), and a force constant of $0.33 \times 10^6 \text{ kJ mol}^{-1} \text{ nm}^{-2}$, for example, an aliphatic carbon-carbon bond, the harmonic function in Eq. (15) can fluctuate by approximately 0.004 nm:

$$2.5 = \frac{1}{2} \times 0.33 \times 10^6 [b_i(\mathbf{r}) - b_i^0]^2 \Rightarrow [b_i(\mathbf{r}) - b_i^0] \approx 0.004 \text{ nm} \quad (16)$$

So, the typical fluctuations that can be expected are in the thousandths of a nanometer. The associated period of a single fluctuation is on the order of 10 fs.

For the potential energy of an individual bond angle, most force fields use, again, a harmonic potential

$$V_i^{\text{angle}}(\mathbf{r}) = \frac{1}{2} K_i^\theta [\theta_i(\mathbf{r}) - \theta_i^0]^2 \quad (17)$$

where K_i^θ is the harmonic angle-bending force constant, θ_i^0 is the minimum energy angle value and $\theta_i(\mathbf{r})$ is the actual angle, which can be computed from the molecular configuration. At room temperature, for typical angles in organic molecules, one expects fluctuations of $5\text{--}10^\circ$.

Four atoms that are sequentially connected by three bonds, $i\text{--}j\text{--}k\text{--}l$, will give rise to a torsional profile, corresponding to a rotation around the $j\text{--}k$ bond. The potential energy as a function of the torsional angle ϕ , will be a periodic function, so a functional form that is based on a cosine seems to be a sensible choice. One common form would be

$$V_i^{\text{torsion}}(\mathbf{r}) = K_i^\phi [1 + \cos(m_i \phi_i(\mathbf{r}) + \delta_i)] \quad (18)$$

where K_i^ϕ is the force constant, or half barrier height, m_i is the multiplicity of the torsional term, which determines the number of maxima and minima that is observed along a full turn, δ_i is a phase shift, with which we can determine, if a zero-degree angle should be a maximum or rather a minimum of the potential energy and $\phi_i(\mathbf{r})$ is the actual value of the dihedral angle, which can be computed from the atomic coordinates. It can be defined as the angle between the plain through atoms i , j and k and the plain through atoms j , k and l .

To ensure a proper coordination of four atoms that should remain in a plain or in a tetrahedral arrangement an out-of-plane or improper dihedral angle term may be included. The coordinate ξ is used to define the spatial arrangement, which can then be enforced using a harmonic potential energy term:

$$V_i^{\text{improper}}(\mathbf{r}) = \frac{1}{2} K_i^\xi [\xi_i(\mathbf{r}) - \xi_i^0]^2 \quad (19)$$

The nonbonded interactions in Eq. (13), theoretically run over all pairs of atoms and consist of nonpolar interactions, $V^{\text{nonpolar}}(\mathbf{r})$, and the electrostatic interactions, $V^{\text{elec}}(\mathbf{r})$, such that we can write

$$V^{\text{nonbonded}}(\mathbf{r}) = \sum_{i=1}^N \sum_{j>i}^N [V_{ij}^{\text{nonpolar}}(\mathbf{r}) + V_{ij}^{\text{elec}}(\mathbf{r})] \quad (20)$$

with the nonpolar (dispersion) interactions typically approximated by the Lennard-Jones potential:

$$V_{ij}^{\text{nonpolar}}(\mathbf{r}) = V_{ij}^{\text{LJ}}(\mathbf{r}) = 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \quad (21)$$

where ϵ_{ij} is the well-depth of the interaction at the minimum energy distance, and σ_{ij} is the distance at which the potential energy passes through zero, i.e., switches from attractive to repulsive, and r_{ij} is the distance between particles i and j .

The electrostatic interactions are given by Coulomb's law

$$V_{ij}^{\text{elec}}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0\epsilon} \frac{q_i q_j}{r_{ij}} \quad (22)$$

where ϵ_0 is the dielectric constant of a vacuum and ϵ is the relative dielectric constant of the medium, q_i and q_j are the partial charges of particles i and j . The electrostatic interaction energy decays as $1/r_{ij}$, which means that this interaction acts at a considerably longer range than the nonpolar interactions. Usually, the nonpolar interactions beyond a distance of, for example, 1.4 nm are typically ignored, while for the electrostatic interactions this would be too rough an approximation. Long range effects can be included either in a mean-field way, using a reaction-field contribution,

$$\Phi_i(r_i) = \frac{1}{4\pi\epsilon_0\epsilon} \sum_{j=1}^N q_j \left[\frac{1}{r_{ij}} - \frac{C_{\text{rf}} r_{ij}^2}{2R_{\text{rf}}^3} - \frac{1 - \frac{1}{2} C_{\text{rf}}}{R_{\text{rf}}} \right] \quad (23)$$

where R_{rf} is the reaction-field cut-off that we apply and C_{rf} depends on the relative dielectric constant, ϵ_{rf} , and the inverse Debye screening length, κ_{rf} , of the medium outside the cutoff,

$$C_{\text{rf}} = \frac{(2\epsilon - 2\epsilon_{\text{rf}})(1 + \kappa_{\text{rf}} R_{\text{rf}}) - \epsilon_{\text{rf}}(\kappa_{\text{rf}} R_{\text{rf}})^2}{(\epsilon + 2\epsilon_{\text{rf}})(1 + \kappa_{\text{rf}} R_{\text{rf}}) + \epsilon_{\text{rf}}(\kappa_{\text{rf}} R_{\text{rf}})^2} \quad (24)$$

Alternatively, the long-range interactions can be taken into account explicitly for condensed phase systems, by applying periodic boundary conditions. The electrostatic potential on an infinite, periodic lattice can then be approximated by lattice-sum methods such as Ewald summation or its variants like particle-mesh Ewald (PME) or particle-particle-particle-mesh (P3M) methods. However, these approaches may introduce an artificial periodicity that could affect the outcome of the calculations.

Force field parameterization

For all of the potential energy terms described in the previous section, there are FF parameters involved. These parameters may be specific for a given molecule, in a given environment and under given thermodynamic boundary conditions. Because it is not feasible to determine all FF parameters for a given system and conditions, we tend to rely on the transferability of many of the FF parameters. It is likely that a carbon-carbon bond in a butane molecule does not vibrate very differently from a carbon-carbon bond in pentane. The aim is to obtain a set of parameters that can be used for as large a set of molecular systems as possible. For this reason, FF parameters are typically derived from smaller molecules, to then apply in larger molecular systems. To facilitate the choice of parameters, this set should be kept as small as possible and also to avoid the risk of overfitting, since there may not be sufficient data to parameterize, for example, the nonpolar parameters for all possible pairs of atom types in different chemical environments.

Furthermore, the choice of parameters for one potential energy term, may strongly influence the choice of parameters for another potential energy term. For example, the nonpolar interactions and the electrostatic interactions are tightly coupled. They both derive from a model in which the electron cloud around some nuclei governs the nonbonded interactions, and it is to some extent a matter of choice how much of the nonbonded interaction energy is represented by the nonpolar and how much by the electrostatic interactions. This strong interdependence of the FF parameters poses a considerable challenge to the parameterization of a force field: all parameters need to be determined self-consistently and preferably lead to the description of transferable functional groups.

There are many different types of force fields described in the literature. For biomolecular systems, there is a choice of AMBER, CHARMM, GROMOS, OPLS, etc. and each of these come in different parameter sets or releases (Riniker, 2018). For mineral systems (specifically aluminosilicates), the ClayFF has been derived (Cygan et al., 2004). When one needs to model molecules that have never been described in the force field of choice yet, it is tempting to “borrow” the parameters from another force field. The strong interdependency of the parameters will usually lead to there being a poor mix of parameters that do not work together. It could result in distorted molecules, or in the absence of properties that should emerge from the interactions, like hydrogen bonds.

As a final remark, the simulation settings that were used in the parameterization of the force field (e.g., electrostatic parameters, cutoffs) should be considered an integral part of the force field and may further limit the transferability of parameters between force fields and parameter sets.

Examples for applications in soil science

Most soils consist of macroaggregates, which in turn are composed of smaller microaggregate structures that comprise mineral, organic and biotic materials linked during pedogenesis. It is known that SOM-SOM, SOM-mineral and mineral-mineral interactions initiate the build-up of composite building units and thus the processes of microaggregate formation. The vast diversity of components in a given aggregate (Fig. 4, Lehmann et al., 2008) provides an architecture intimately linked to the functions of the soil such as the C-storage or the fate of nutrients and pollutants in the environment.

Humic substances (HSs) can be considered a major component of SOM. Thus, they can play a key role in important soil functions. A variety of models has been proposed aimed to represent HS. These models consider either the existence of hypothetical molecules whose composition resembles the functional groups observed in the organic fraction of the soil matrix, or active sites of interaction, such as hydrophilic groups as part of long hydrocarbon chains. These models have served as the basis to approach possible structural arrangements or describe mechanisms of interactions such as aggregation of SOM. Here, cations, particularly those with high ionic potentials (z/r) such as Mg^{2+} and Ca^{2+} , contribute significantly to aggregation through electrostatic interactions with negatively charged sites represented mainly by carboxylate groups. Similarly, modeling also allowed description of the formation of complexes of SOM and dissolved organic matter (DOM) with neutral xenobiotic compounds (e.g.,

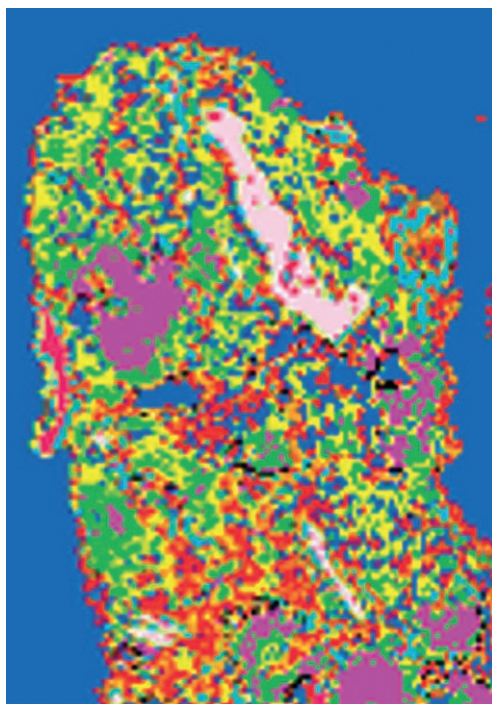


Fig. 4 Distribution of carbon contents in a soil micro assemblage from Nandi Forest (Kenya) determined by NEXAFS (Near-edge X-ray absorption fine structure) in combination with STXM (Scanning transmission X-ray microscopy). The colors represent different functional groups. Modified from Figure 2 of Lehmann J, Solomon D, Kinyangi J, Dathe L, Wirick S and Jacobsen C (2008) Spatial complexity of soil organic matter forms at nanometre scales. *Nature Geoscience* 1: 238–242. (At: <https://www.nature.com/articles/ngeo155>).

pentachlorophenol, atrazine and dichlorodiphenyltrichloroethane (DDT)) having an origin in dispersive forces and hydrogen bonds (h-bonds) detected then as main factors for their temporary retention in soil.

Soil minerals determine transport, diffusion, runoff, leaching or volatilization process in the environment. On the other hand, soil minerals are also important in material science with cosmetic and pharmaceutical applications. Currently, there are models capable of describing the hydrated crystalline structure of soil minerals and their interface with aqueous solutions and biomolecular compounds. For example, many efforts have been made to characterize clay mineral surfaces with FF-MD, which are now also starting points for modeling organo-mineral complexes that are of utmost importance for the stabilization of SOM in soils.

Modeling of soil organic matter—The Vienna soil organic matter modeler

Humic substances (HSs) stand out as being among the most stable compounds in soil. Their composition and three-dimensional structure are decisive both in sorption processes in the soil matrix and in its functions. The Vienna Soil Organic Matter Modeler (VSOMM) (Sündermann et al., 2015; Escalona et al., 2021) can generate condensed-phase models that are diverse in composition, functional groups and structure according to solid-state and liquid-state ^{13}C nuclear magnetic resonance (NMR) carbon distributions. The Modeler uses the concept of building blocks as the basic units to create HSs samples (Fig. 5), similar to amino acids in biochemistry to assemble more complex structures. Currently, the Modeler utilizes a database of 54 building blocks which can have one or more of the following functional groups: carbonyl-C, carboxyl-C, O-aryl-C, aryl-C, di-O-alkyl-C, O-alkyl-C, methoxyl-C and alkyl-C.

Within the natural organic matter, the intensively studied Leonardite humic acid (LHA) is often used as a model because it is characterized by a high content of HSs, especially humic acids (HAs). The resulting models showed that the hydration level of LHA (Fig. 6, Petrov et al., 2017) affects the mobility of molecules within the soil matrix, as well as the viscosity and density of the samples due to the complex network of h-bonds formed.

The hydrophilic sites of HSs contain carboxylic and phenol functional groups, which give them an acid character, as they can release protons. Because most of this acidity is reactive below pH 7, HSs have a net negative charge in all but the most acidic soils. The electrostatic attraction between the HSs negative charges, HSs and the common soil cations give rise to the formation of cation bridges (Fig. 7, Galicia-Andrés et al., 2021a). These coordination complexes are classified as: inner-sphere complex (ISC) if the ligand (HS) and the metal cation interact directly; or outer-sphere complex (OSC) if they interact through the solvent (e.g., water molecules).

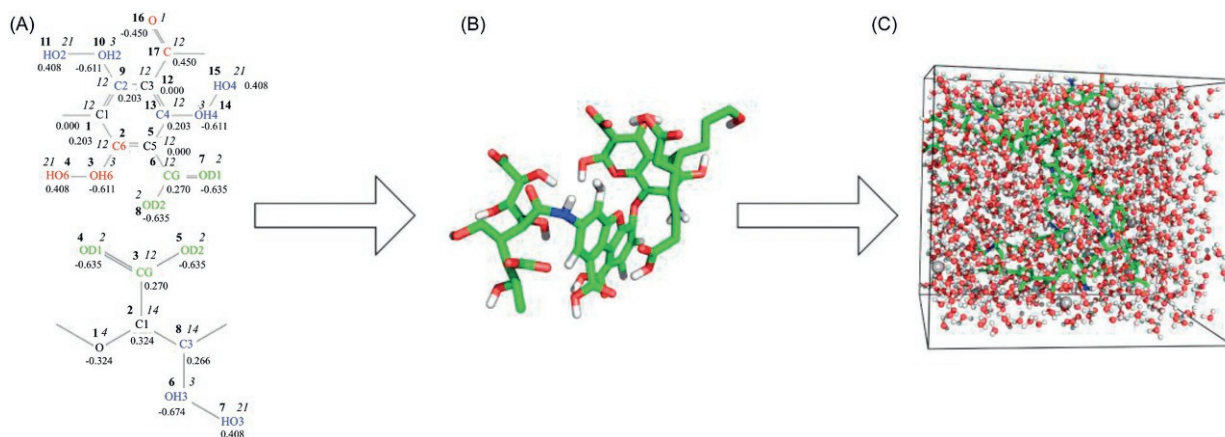


Fig. 5 Assembling process of two building blocks linked together to form HS molecules. From Sündermann A, Solc R, Tunega D, Haberhauer G, Gerzabek MH and Oostenbrink C (2015) Vienna Soil–Organic–Matter Modeler—Generating condensed-phase models of humic substances. *Journal of Molecular Graphics and Modelling* 62: 253–261. (At: <https://www.sciencedirect.com/science/article/pii/S109332631530067X>).

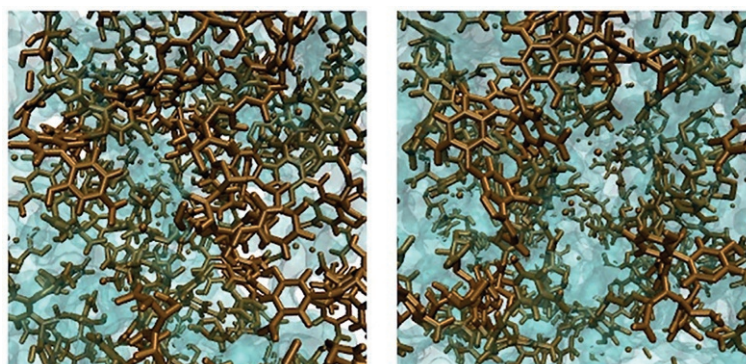


Fig. 6 Soil matrix scheme with two hydration levels. Leonardite as orange sticks and water as the blue surface. Modified from Figure 1 of Petrov D, Tunega D, Gerzabek MH and Oostenbrink C (2017) Molecular dynamics simulations of the standard Leonardite humic acid: Microscopic analysis of the structure and dynamics. *Environmental Science and Technology* 51: 5414–5424. (At: <http://pubs.acs.org/doi/10.1021/acs.est.7b00266>).

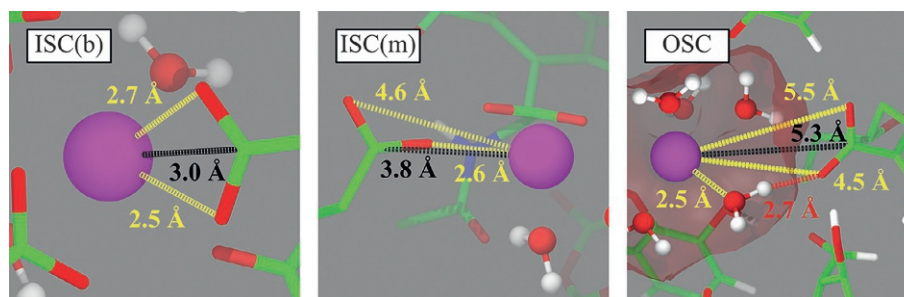


Fig. 7 Representation of the inner-sphere complex (ISC), bidentate (b), monodentate (m), and outer-sphere complex (OSC) between carboxylate groups and Ca^{2+} cations. HAs as sticks, cations as balls (violet), water molecules as balls and sticks (red and white) and distances are shown in Angstroms. Modified from Figure 5 of Galicia-Andrés E, Escalona Y, Oostenbrink C, Tunega, D and Gerzabek MH (2021a) Soil organic matter stabilization at molecular scale: The role of metal cations and hydrogen bonds. *Geoderma* 401: 115237. (At: <https://www.sciencedirect.com/science/article/pii/S0016706121003177>).

The stability of the coordinated pair depends on the complex formed. The bidentate ISCs are the most stable with two coordinated oxygen atoms, unlike monodentate ISCs in which only one atom is coordinated. Conversely, the OSCs exhibit the least stability by being separated by the solvent and so increasing the probability of dispersing the aggregate. This suggests that the formation of supramolecular aggregates is the result of the interaction between cations and HSs. The aggregation mechanism promoted by the presence of cations is related to the DLVO (Derjaguin–Landau–Verwey–Overbeek) theory formulated for emulsions and colloids.

Interaction of SOM with reactive mineral surfaces

For decades, it was assumed that the stable organic matter consisted mainly of the so-called HSs, the larger complex molecules. It was believed that the microbial degradability of organic substances in the soil varied greatly due to their chemical structures (chemical recalcitrance). At the beginning of this millennium (Schmidt et al., 2011), however, it was increasingly recognized that the rates of degradation of complex molecules, such as lignin, and much smaller molecules, such as sugars or proteins, are very similar. At the same time, interaction with reactive mineral surfaces has been established as a new explanatory model for the stabilization of organic molecules in soil. The decomposition of plant residues to carbon dioxide is thus slowed down by binding of the decomposition products to clay minerals and oxides; stable pools of organic matter are created precisely by this mechanism (Lehmann and Kleber, 2015).

The formation of microaggregates in the soil through physical, chemical, and biological processes takes place in the course of soil development. During the process, additional surfaces are created and aggregation has significant effects on soil functions such as water-holding capacity or the retention of nutrients and pollutants. An essential part of the process is the formation of organo-mineral complexes (Totsche et al., 2018).

These organo-mineral aggregates are stabilized mainly by hydrophobic and h-bond interactions; however, ion-exchange, cation bridges and the dielectric environment also play an important role. Clay mineral surfaces such as the montmorillonite (100) basal surface have a negative charge due to the isomorphous substitutions of cations with lower charges than the primordial cations induced on their basal surface structures. These negatively charged layers promote the adsorption of positively charged species, e.g., counterions, creating a positive surface charge effect. Other positively charged species can compete with counterions on the surface, leading to ion exchange (Grančič and Tunega, 2020). If counterion layers attract negatively charged species, such as HSs, these are adsorbed on the surface via cation bridges (ISCs and OSCs).

On the other hand, surfaces that do not exhibit a net charge, such as the basal surface of kaolinite, the dielectric environment takes precedence. Simulations enable two types of surrounding environments to be modeled with different dielectric properties which influence the SOM adsorption mechanism. Nanopore cavity (periodic) models can polarize the immersed environment (e.g., water molecules) by replicating the model layer an infinite number of times in the direction normal to the exposed surface (Fig. 8A), whereas slab models assume a single layer of infinite surface with the environment non-polarized (Fig. 8B) (Galicía-Andrés et al., 2019).

Both models show SOM adsorbing on the basal surface of kaolinite via h-bonds; however, the polarized environment promotes the breakdown of the supramolecular aggregate and subsequently the dispersion of individual molecules on the surface (Fig. 9). The detached HS molecules are forced to the surface of the alumina octahedral sheet and the cations to the siloxane surface of the kaolinite.

Unlike the aggregate, detached molecules have greater mobility, allowing them to disperse effectively over the surface until its saturation. In contrast, non-polarized models favor the aggregate structures by clustering dispersed molecules through cation bridges. These observations may explain the behavior of SOM as supramolecular aggregates clustered by cation bridges and h-bonds in solution. Once SOM comes into contact with soil minerals, it is adsorbed. When the environment is polarized, as inside a nanopore, the SOM breaks down into its individual components and the molecules are dispersed, trapping SOM inside the nanopore.

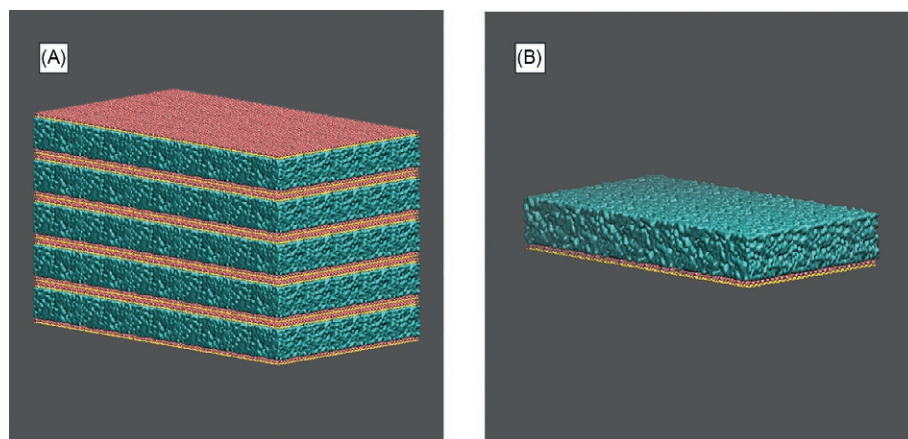


Fig. 8 Kaolinite–water interface of periodic model (A); and slab model (B). Reprinted with permission from Galicía-Andrés E, Petrov D, Gerzabek MH, Oostenbrink C and Tunega D (2019) Polarization effects in simulations of kaolinite–Water interfaces. *Langmuir* 35: 14975–15432. (At: <https://pubs.acs.org/doi/10.1021/acs.langmuir.9b02945>). Copyright (2019) American Chemical Society.

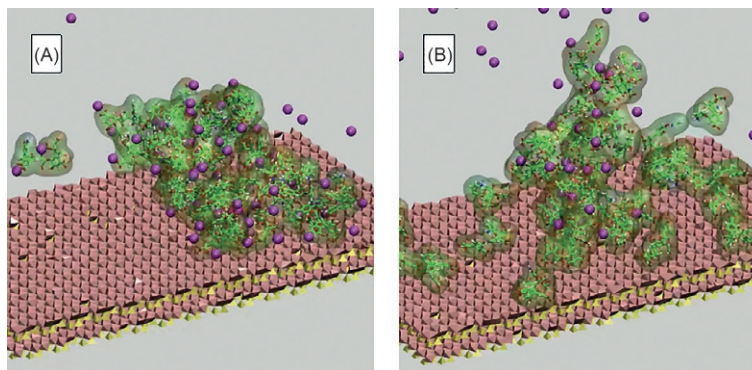


Fig. 9 Snapshots of the adsorbed supramolecular organic aggregate (A) and the individual HA molecules dispersed on the kaolinite basal surface (Al–OH octahedral layer) (B). Based on Figure 14 of Galicia-Andrés E, Grančić P, Gerzabek MH, Oostenbrink C and Tunega D (2021b) Modeling of interactions in natural and synthetic organoclays. In: Sainz Diaz IC (ed.). *Computational Modeling in Clay Mineralogy*, pp. 211–254. Bari, Italy: Digilabs.

Organic contaminant interactions with mineral surfaces and organic matter

Special attention is paid to the fate of pollutants, such as pesticides, antibiotics, and pharmaceuticals, polycyclic aromatic hydrocarbons and oil components, because of their potential effects on the ecosystem (Tunega et al., 2016; Churchman and Velde, 2019). Understanding the mechanisms involved in the sorption processes and the interactions of organic molecules with mineral surfaces and organic matter is a key point in the application of remediation processes.

Glyphosate is one of the most widely used herbicides due to its broad spectrum of selectivity against weeds. However, long-term application may have adverse effects due to the toxicity of breakdown products, leaching to water sources as well as its retention in soil. The main mechanism of interaction between glyphosate and 1:1 clay minerals, such as kaolinite, is through h-bonds, which represent a precursor stage to the formation of complexes. Glyphosate interacts with the surface of the alumina octahedral sheet forming as many h-bonds as possible, i.e., parallel to it (Fig. 10, Galicia-Andrés et al., 2020); although, it is also capable of interacting through its COOH and PO(OH)₂ sites.

Furthermore, the pH solution influences the adsorption strength. In the neutral state glyphosate (pH < 2.3) is weakly adsorbed on the kaolinite ($\Delta G = -5 \text{ kJ mol}^{-1}$) compared to anionic glyphosate ($2.3 < \text{pH} < 6.0$) which exhibits a stronger adsorption ($\Delta G = -14 \text{ kJ mol}^{-1}$) due to its ionic nature. This behavior is strongly enhanced when the glyphosate is inside a nanopore cavity. In contrast, the adsorption strength of the neutral glyphosate is not influenced by the polarized environment, however some degrees of freedom are restricted.

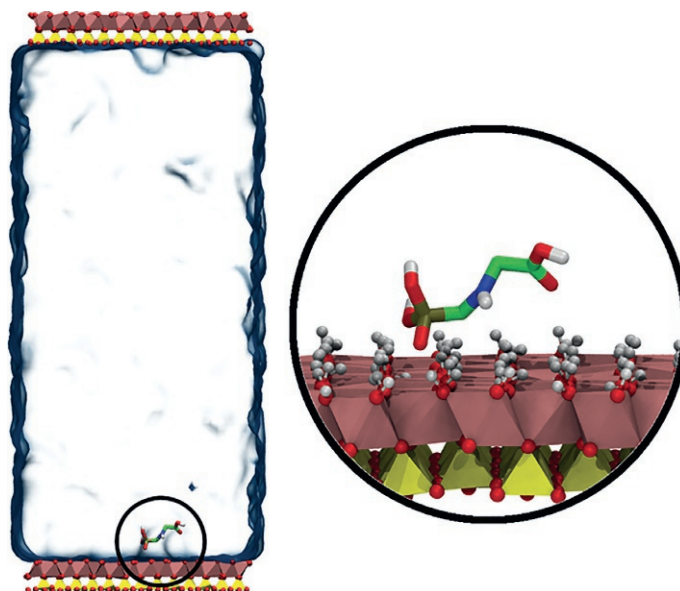


Fig. 10 Representation of neutral glyphosate in solution in contact with a kaolinite Al-octahedron layer. Modified from Figure 1 of Galicia-Andrés E, Tunega D, Gerzabek MH and Oostenbrink C (2020) On glyphosate–kaolinite surface interactions. A molecular dynamic study. *European Journal of Soil Science* 72: 1231–1242. (At: <https://doi.org/10.1111/ejss.12971>).

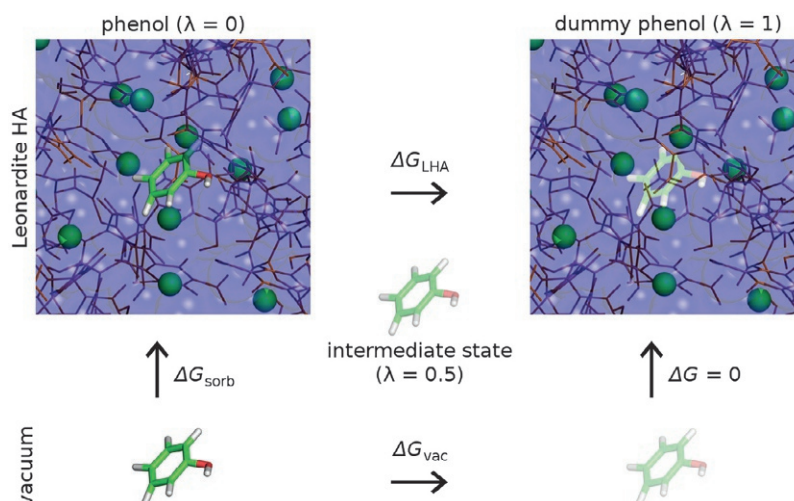


Fig. 11 Thermodynamic cycle of sorption free energy of phenol in the Leonardite humic acid (LHA) medium and vacuum. The model represents a solute-medium completely interacting form ($\lambda = 0$) and a non-interacting ($\lambda = 1$). Modified from Figure 1 of Petrov D, Tunega D, Gerzabek MH and Oostenbrink C (2020) Molecular modelling of sorption processes of a range of diverse small organic molecules in Leonardite humic acid. *European Journal of Soil Science* 71: 831–844. (At: <https://onlinelibrary.wiley.com/doi/10.1111/ejss.12868>).

Similar to soil minerals, SOM can incorporate pollutants into its matrix. This sorption capacity depends on the composition and three-dimensional structure of SOM, as well as environmental variables such as water content, pH solution and cation content (Fig. 11, Petrov et al., 2020). It is known that water associated with SOM enhances or weakens the adsorption potential, however, small compounds can influence SOM–water interactions as well affecting its hydration state. These interactions are mainly controlled by the polar interactions (i.e., electrostatic interactions and h–bonds) and hydrophobic moieties between sorbent and water molecules.

Sorption free energies of sorbate molecules on hydrated LHA exhibited similar values to hydration free energies, attributed to the preferential location of sorbate molecules to water patches. This behavior explains why adsorption of sorbents depends strongly on the location of the soil particle.

Conclusions and outlook

Molecular modeling and simulation have been used frequently in the chemical sciences for several decades. For example, it is hard to imagine current day drug development without it. The enormous advances in computer performance during the past few decades have enabled the simulation of relatively large systems in a comparatively short time. This strong performance of the methods makes it increasingly possible to model very complex systems, such as soil organic matter or the soil matrix in general. Today, soil science benefits in several ways from the methodological developments in modeling. The ab initio methods are able to calculate systems with an increasing number of atoms and thus can obtain results independently of experiments. The FF methods can handle large systems with a multitude of molecules and are an important bridge to the experimental systems. The results of MM can be used to explain and discuss experimental data, as many interactions that are important in soil cannot be reproduced in the necessary detail experimentally. In the present chapter some relatively complex systems up to organo-mineral complexes in soil have been presented as examples. Tools such as the VSOMM now provide completely new views, consistent with known experimental knowledge, of the complex structures and interactions that characterize soil organic matter. The organo-mineral interactions presented here point to the great future potential of MM. Numerous “hot” topics in soil research, such as SOM stabilization, the binding mechanisms of non-extractable organic contaminants, or the molecular interaction of pollutants in soil in general, are among those research topics that will benefit greatly from MM methods today and in the future.

Acknowledgments

The authors thank the Austrian Science Fund (FWF) for financial support through the projects I4876–N (and DFG BO 5388/1–1), P30224–N34 and I3263–N34. The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC) under the projects “Application of computer simulation for soil bioremediation” No. 71170 and “Modeling of sorption processes at surfaces and interfaces of minerals and organoclays” No. 70544. We thank Wilfred van Gunsteren for designing the first versions of Figs. 1 and 2.

References

- Aristilde, L., & Sposito, G. (2010). Binding of ciprofloxacin by humic substances: A molecular dynamics study. *Environmental Toxicology and Chemistry* 29, 90–98. (At: <https://setac.onlinelibrary.wiley.com/doi/10.1002/etc.19>).
- Churchman GJ and Velde B (2019) *Soil Clays: Linking Geology, Biology, Agriculture and the Environment*, 1st edn. Boca Raton: CRC Press.
- Cramer CJ (2004) *Essentials of Computational Chemistry: Theories and Models*, 2nd edn. West Sussex, England: John Wiley and Sons, Ltd.
- Cygan RT, Liang J-J, and Kalinichev AG (2004) Molecular models of hydroxide, oxyhydroxide, and clay phases and the development of a general force field. *Journal of Physical Chemistry B* 108: 1255–1266.
- Devarajan, D., Liang, L., Gu, B., Brooks, S.C., Parks, J.M., and Smith, J.C. (2020). Molecular dynamics simulation of the structures, dynamics, and aggregation of dissolved organic matter. *Environmental Science and Technology* 54, 13527–13537. (At: <https://pubs.acs.org/doi/abs/10.1021/acs.est.0c01176>).
- Escalona, Y., Petrov, D. and Oostenbrink, C. (2021). Vienna soil organic matter modeler 2 (VSOMM2). *Journal of Molecular Graphics and Modelling* 103, 107817, (At: <http://www.sciencedirect.com/science/article/pii/S1093326320306069>).
- Frenkel D and Smit B (2002) *Understanding Molecular Simulation*, 2nd edn. San Diego, CA: Academic Press.
- García-Andrés, E., Petrov, D., Gerzabek, M.H., Oostenbrink, C. and Tunega, D. (2019). Polarization effects in simulations of kaolinite–Water interfaces. *Langmuir* 35, 14975–15432. (At: <https://pubs.acs.org/doi/10.1021/acs.langmuir.9b02945>).
- García-Andrés, E., Tunega, D., Gerzabek, M.H. and Oostenbrink, C. (2020). On glyphosate–kaolinite surface interactions. A molecular dynamic study. *European Journal of Soil Science* 72, 1231–1242. (At: <https://doi.org/10.1111/ejss.12971>).
- García-Andrés, E., Escalona, Y., Oostenbrink, C., Tunega, D. and Gerzabek, M.H. (2021a). Soil organic matter stabilization at molecular scale: The role of metal cations and hydrogen bonds. *Geoderma* 401, 115237. (At: <https://www.sciencedirect.com/science/article/pii/S0016706121003177>).
- García-Andrés E, Grančić P, Gerzabek MH, Oostenbrink C, and Tunega D (2021b) Modeling of interactions in natural and synthetic organoclays. In: Sainz Diaz IC (ed.) *Computational Modeling in Clay Mineralogy*, pp. 211–254. Bari, Italy: Digilabs.
- Gerzabek MH, Aquino AJA, Haberhauer G, Tunega D, and Lischka H (2001) Molecular modelling—Opportunities for soil research. *Die Bodenkultur* 52: 133–146.
- Grančić, P. and Tunega, D. (2020). Geometry and molecular arrangement of phosphatidylcholine-montmorillonite bioclays via classical molecular dynamics simulation. *Applied Clay Science* 198, 105815. (At: <https://www.sciencedirect.com/science/article/abs/pii/S016913172030380X>).
- Haberhauer, G., Pfeiffer, L., Gerzabek, M.H., Kirchmann, H., Aquino, A.J.A., Tunega, D., and Lischka, H. (2001). Response of sorption processes of MCPA to the amount and origin of organic matter in a long-term field experiment. *European Journal of Soil Science* 52, 279–286. (At: <https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2389.2001.00382.x>).
- Lan, T., Liao, J., Yang, Y., Chai, Z., Liu, N., & Wang, D. (2019). Competition/Cooperation between humic acid and graphene oxide in uranyl adsorption implicated by molecular dynamics simulations. *Environmental Science & Technology* 53, 5102–5110. (At: <https://pubs.acs.org/doi/10.1021/acs.est.9b00656>).
- Leach A (2001) *Molecular Modelling: Principles and Applications*, 2nd edn. Essex, England: Pearson Education Limited.
- Lehmann, J. and Kleber, M. (2015). The contentious nature of soil organic matter. *Nature* 528, 60–68. (At: <https://www.nature.com/articles/nature16069>).
- Lehmann, J., Solomon, D., Kinyangi, J., Dathe, L., Wirick, S. and Jacobsen, C. (2008). Spatial complexity of soil organic matter forms at nanometre scales. *Nature Geoscience* 1, 238–242. (At: <https://www.nature.com/articles/ngeo155>).
- Liang, Y., Ding, Y., Wang, P., Lu, G., Dang, Z., and Shi, Z. (2019). Molecular characteristics, proton dissociation properties, and metal binding properties of soil organic matter: A theoretical study. *Science of the Total Environment* 656, 521–530. (At: <https://www.sciencedirect.com/science/article/pii/S0048969718347430?via%3Dihub>).
- Loganathan, N., Ferguson, B.O., Arey, B., Argersinger, H.E., and Bowers, G.M. (2020). A mechanistic exploration of natural organic matter aggregation and surface complexation in smectite mesopores. *The Journal of Physical Chemistry A* 124, 9832–9843. (At: <https://pubs.acs.org/doi/10.1021/acs.jpca.0c08244>).
- Petrov, D., Tunega, D., Gerzabek, M.H., and Oostenbrink, C. (2017). Molecular dynamics simulations of the standard Leonardite humic acid: Microscopic analysis of the structure and dynamics. *Environmental Science and Technology* 51, 5414–5424. (At: <http://pubs.acs.org/doi/10.1021/acs.est.7b00266>).
- Petrov, D., Tunega, D., Gerzabek, M.H., and Oostenbrink, C. (2020). Molecular modelling of sorption processes of a range of diverse small organic molecules in Leonardite humic acid. *European Journal of Soil Science* 71, 831–844. (At: <https://onlinelibrary.wiley.com/doi/10.1111/ejss.12868>).
- Riniker, S. (2018). Fixed-charge atomistic force fields for molecular dynamics simulations in the condensed phase: An overview. *Journal of Chemical Information and Modeling* 58, 565–578. (At: <https://pubs.acs.org/doi/10.1021/acs.jcim.8b00042>).
- Schaumann, G. E. (2006). Soil organic matter beyond molecular structure Part I: Macromolecular and supramolecular characteristics. *Journal of Plant Nutrition and Soil Science* 169, 145–156. (At: <https://onlinelibrary.wiley.com/doi/10.1002/jpln.200521785>).
- Schmidt, M.W., Torn, M.S., Abiven, S., Dittmar, T., Guggenberger, G., Janssens, I.A., Kleber, M., Kögel-Knabner, I., Lehmann, J., Manning, D.A.C., Nannipieri, P., Rasse, D.P., Weiner, W. and Trumbore, S.E. (2011). Persistence of soil organic matter as an ecosystem property. *Nature* 478, 49–56. (At: <https://www.nature.com/articles/nature10386>).
- Sündermann, A., Solc, R., Tunega, D., Haberhauer, G., Gerzabek, M.H., and Oostenbrink, C. (2015). Vienna Soil–Organic–Matter Modeler—Generating condensed-phase models of humic substances. *Journal of Molecular Graphics and Modelling* 62, 253–261. (At: <https://www.sciencedirect.com/science/article/pii/S109332631530067X>).
- Totsche, K.U., Amelung, W., Gerzabek, M.H., Guggenberger, G., Klumpp, E., Knief, C., Lehdorff, E., Mikutta, R., Peth, S., Prechtel, A., Ray, N. and Kögel-Knabner, I. (2018). Microaggregates in soils. *Journal of Plant Nutrition and Soil Science* 181, 104–136. (At: <https://onlinelibrary.wiley.com/doi/full/10.1002/jpln.201600451>).
- Tunega D, Gerzabek MH, Haberhauer G, Lischka H, and Aquino AJA (2016) Organic and contaminant geochemistry. In: *Molecular Modeling of Geochemical Reactions*, pp. 177–243. West Sussex, UK: John Wiley and Sons, Ltd.
- van Gunsteren, W.F., Bakowies, D., Baron, R., Chandrasekhar, I., Christen, M., Daura, X., Gee, P., Geerke, D.P., Glättli, A., Hünenberger, P.H., Kastenholz, M.A., Oostenbrink, C., Schenk, M., Trzesniak, D., Van der Vegt, N.F.A., and Yu, H.B. (2006). Biomolecular modeling: Goals, problems, perspectives. *Angewandte Chemie International Edition* 45, 4064–4092. (At: <https://onlinelibrary.wiley.com/doi/10.1002/anie.200502655>).
- Young D (2001) *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*. New York: John Wiley and Sons, Inc.
- Zhang, Y., Liu, X., Zhang, C., and Lu, X. (2020). A combined first principles and classical molecular dynamics study of clay-soil organic matters (SOMs) interactions. *Geochimica et Cosmochimica Acta* 291, 110–125. (At: <https://www.sciencedirect.com/science/article/pii/S0016703719307793>).

Relevant Websites

<https://somm.boku.ac.at/>—BOKU-Soil Organic Matter Modeler.

Raman spectroscopy

Anupam Das and Huiyuan Guo, Department of Chemistry, The State University of New York at Binghamton, Binghamton, NY, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	497
Raman spectroscopy	498
Surface-enhanced Raman spectroscopy (SERS)	499
Application of Raman spectroscopy and SERS in the analysis of soil	500
Application of Raman spectroscopy and SERS in the analysis of soil components	500
Classification of soil samples	501
Analysis of soil organic matter	501
Analysis of soil phosphorus	503
Application of Raman spectroscopy and SERS in the analysis of soil contaminants	503
Analysis of organic contaminants in soil	503
Detection of thiram in soil	503
Detection of organophosphorus pesticides in soil	503
Detection of chlorpyrifos in soil	504
Detection of thiabendazole in soil	504
Detection of persistent organic pollutants in soil	506
Analysis of inorganic contaminants in soil	506
Detection of vivianite arsenates in soil	506
Detection of harmful metalloids in soil	506
Detection of lead in roadside soil	506
Detection of Zn(II) in soil	507
Conclusion and outlook	507
Acknowledgment	508
References	508

Abstract

The chapter focuses on Raman spectroscopy and its applications in soil analysis. Raman spectroscopy detects vibrational modes of molecules and finds wide applications in the detection of organic and inorganic components in soils. An upgraded version of Raman spectroscopy is surface-enhanced Raman spectroscopy (SERS), which plays a crucial role in the highly sensitive detection of trace elements present in soil samples due to its significant signal enhancement compared to Raman spectroscopy. Combined with other analytical tools and database algorithms, Raman spectroscopy and SERS provide specific and sensitive analysis for soil components and contaminants.

Key points

- The theory, principle, and mechanism of Raman spectroscopy and SERS are explained in detail.
- The differences between Raman spectroscopy and SERS are discussed.
- Important applications of Raman spectroscopy and SERS in the analysis of soil components and contaminants are reviewed.

Introduction

Raman spectroscopy studies the interactions between matter and light (electromagnetic radiation) as a function of the wavelength or frequency of the radiation. A fraction of incident radiation gets scattered upon interaction with molecules, atoms, electrons or phonons. There are different types of scattering, such as, Compton scattering, Brillouin scattering, Rayleigh scattering, and Raman scattering. Among them, Raman scattering plays an important role in the field of analytical applications due to its molecular fingerprinting capability. Raman scattering of electromagnetic radiation by molecules was predicted by Adolf Smekal using classical quantum theory in 1923 (Smekal, 1923). Later, C. V. Raman along with his co-researcher V. R. Krishnan experimentally proved the inelastic scattering phenomena in 1928 (Raman and Krishnan, 1928). When molecules are irradiated with monochromatic light, the radiation scatters in all directions. The majority of scattered radiation does not exhibit any change in frequency ($\Delta\nu$), i.e., the frequency of incident light (ν_i) is equal to that of the scattered light (ν_s), $\Delta\nu = \nu_i - \nu_s = 0$. This type of scattering is called Rayleigh scattering (Fig. 1). A tiny fraction (usually 1 part per 10 million) of the scattered radiation exhibits different frequencies or energy

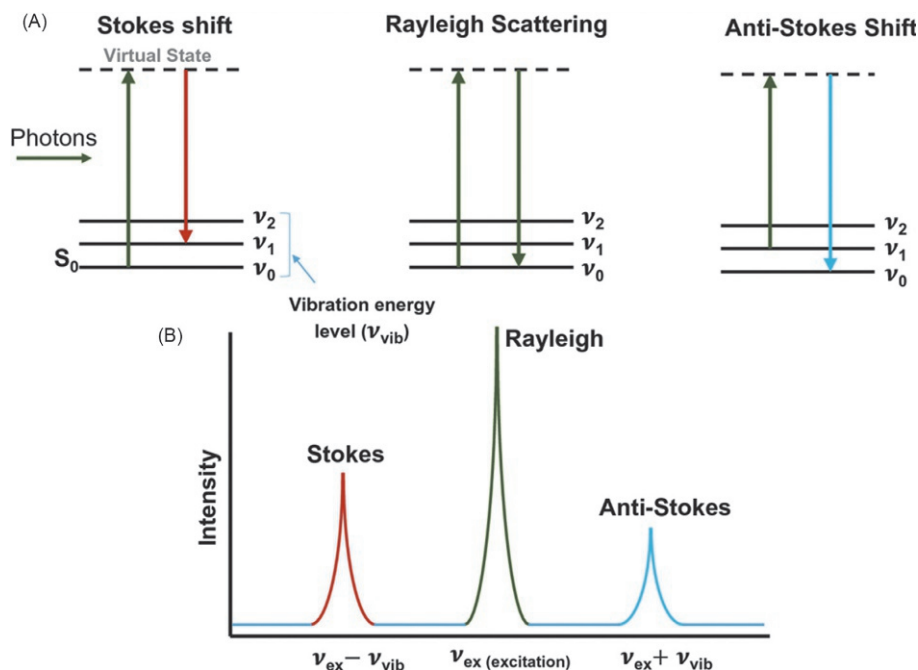


Fig. 1 Jablonski diagram showing the origin of Rayleigh, Stokes and Anti-Stokes Raman scattering (A) and their distribution of frequency (B). Adapted with permission from Cho YC and Ahn SI (2020) Fabricating a Raman spectrometer using an optical pickup unit and pulsed power. *Scientific Reports* 10: 11692–11699.

levels compared to that of the incident radiation; this fraction is called Raman scattering which consists of two types of scattering (Cho and Ahn, 2020). When the molecules gain energy from the interacting radiation photon, i.e., the frequency of the scattered photon is smaller than that of the incident one ($\nu_s < \nu_i$). It is called Stokes Raman scattering and the change in frequency ($\Delta\nu_s = \nu_i - \nu_s$) is called the Stokes shift (Fig. 1). On the other hand, when the molecules lose energy to the interacting photon, the frequency of the incident photon is smaller than that of the scattered one ($\nu_i < \nu_{AS}$). It is called anti-Stokes Raman scattering and the change in frequency ($\Delta\nu_{AS} = \nu_{AS} - \nu_i$) is called the anti-Stokes shift (Das and Agrawal, 2011).

Raman spectroscopy

In general, Raman shifts are represented in terms of wavenumber ($\bar{\nu}$), i. e., the inverse of wavelength (λ) and directly proportional to energy (E). It has a unit of inverse length, i.e., cm^{-1} . The following formula can be used to establish the relationship between the spectral wavenumber and wavelength of a Raman spectrum. Typically, wavelength is expressed in nm. In this case, the final form of the relationship is as follows, where, λ_i is the wavelength of the incident radiation and λ_R is the wavelength of the Raman scattering.

$$\Delta\bar{\nu}(\text{cm}^{-1}) = \left(\frac{1}{\lambda_i(\text{nm})} - \frac{1}{\lambda_R(\text{nm})} \right) \times \frac{10^7 \text{ nm}}{\text{cm}}.$$

The following energy-level diagram (Fig. 1) explains different types of scattering derived from the interactions between incident radiation photons and sample molecules.

One of the most important criteria of a molecule to show Raman shift or to be Raman-active is its electromagnetic field-induced polarizability. Briefly, polarizability is defined as the tendency or ability of electron clouds of molecules to be distorted from their normal shape upon interaction with an external electromagnetic field. When a molecule is placed in the electromagnetic field, separation of charges, i.e., electrons and positive charges, happens, which is called induced dipole moment or polarizability (μ_i) (Gardiner, 1989). This induced dipole moment is directly proportional to the strength of the electromagnetic field applied (E). Therefore,

$$\mu_i = \alpha E$$

where α is the polarizability of a molecule, i.e., induced dipole moment per unit of the electromagnetic field applied (Gardiner, 1989). The change in polarizability is caused by the molecular vibrations induced by external interacting electronic vibration. The point to be noted is that, though large molecules have more vibrational modes ($3N - 6$), where N is the number of atoms, not all of the modes are equally Raman active. Some functional groups, for example, $-\text{S}-\text{S}-$, $\equiv\text{C}-\text{NO}_2$, $\equiv\text{C}-\text{S}-$, $\equiv\text{C}-\text{X}$ ($\text{X} = \text{F}, \text{Cl}, \text{Br}$

or I), $\text{—C}\equiv\text{N}$, and $\text{—C}\equiv\text{C—}$ exhibit Raman-active vibrational modes and strong changes in polarizability, leading to strong Raman signals (Hadjiivanov et al., 2021). In addition, symmetric stretches are more intense in Raman spectra. Therefore, molecular symmetry is considered as one of the major criteria for a molecule to be Raman-active. For example, $\text{N}\equiv\text{N}$ (N_2) is a symmetric molecule and thus Raman active. Therefore, a molecule having certain chemical bonding and symmetry exhibits characteristic vibrational frequencies, leading to distinct Raman peaks, which enables Raman spectroscopy to identify molecules (Dijkstra et al., 2005).

Raman spectroscopy has many advantages: (1) It provides both qualitative and quantitative information about the samples analyzed. Raman bands are the result of molecular vibrations after interactions with radiation. These vibrational modes are sensitive to even minor changes in chemical structure. Therefore, subtle differences in molecular environments are detectable. (2) Due to the noninvasive nature and fingerprinting capability, Raman spectroscopy has found applications involving non-destructive chemical analysis and microscopic imaging. Most Raman spectrometers use the radiation of visible or near-visible ranges, which can pass through transparent objects. Therefore, it is possible to scan samples even when they are sealed in transparent containers (i.e., glass or plastic vials and capillary tubes) or have viewing windows (i.e., temperature or pressure cells). (3) It requires small sample size and minimal sample preparation. For liquid samples, only a small volume (microliter or even nanoliter) is required for analysis. For solid samples, as long as the target regions are reachable to the microscope, Raman spectroscopy can be used to record the spectra on the samples directly. (4) Unlike infra-red (IR), Raman spectroscopy faces little to no interference by water molecules present in samples. This is because longer bonds scatter more photons based on the theory of light scattering. For example, molecules with long π -bonds (e.g., benzene) scatter a larger number of photons. On the other hand, water has small single bonds and is recognized as a weak Raman scattering molecule. (5) Raman detection is rapid, allowing spectral acquisition in seconds or even faster. These advantages helped Raman spectroscopy gain popularity in the detection and imaging of target molecules in different fields, such as pharmaceuticals, cosmetics, forensic science, geology, material science, and life sciences (Kuhar et al., 2018).

Though there are several advantages of using Raman spectroscopy as a tool to detect target analytes, it has several drawbacks that limit its wide application. First, Raman scattering is weak and detection cannot attain great sensitivity. Second, Raman spectroscopy suffers from fluorescence interference if the samples or sometimes impurities have background fluorescence signals that overshadow the Raman fingerprint region. To overcome these inherent problems, an improved version of Raman spectroscopy, known as surface-enhanced Raman spectroscopy (SERS) was developed.

Surface-enhanced Raman spectroscopy (SERS)

In 1974, Fleischman et al. observed for the first time a strong amplification of Raman signals derived from pyridine molecules adsorbed on roughened silver electrodes (Fleischmann et al., 1974). Subsequently, in 1977, van Duyne and co-workers coined the term surface-enhanced Raman spectroscopy (SERS) and carried out further experiments and validation (Jeanmaire and Van Duyne, 1977). Briefly, SERS is a phenomenon whereby Raman scattering is enhanced by many folds when the molecules are in the vicinity of plasmonic metal surfaces. In more technical terms, SERS is a phenomenon that involves both the plasmonic process (light–metal interaction) and vibrational spectroscopy (light–molecule interaction). When the incident light (laser) shines on roughened metal surfaces, especially noble metal nanostructures, the electromagnetic radiation drives the cloud of delocalized conduction electrons on the surface into collective oscillation. When the inherent oscillation frequency of the free electrons on the nanostructures matches with the frequency of the incident light, surface plasmon resonance (SPR) takes place. The SPR frequency varies depending upon the shape, size, electron-density, dielectric nature, and effective electron mass of the nanostructures. When the size of the nanostructures is smaller than the wavelength of the excitation light, confinement of surface plasmons happens and leads to localization of SPRs called localized surface plasmon resonance (LSPR) (Panikar et al., 2021). The nanostructures able to generate strong LSPR effects are called plasmonic nanostructures. They are usually made of noble metals, e.g., Au (gold), Ag (silver), or Cu (copper) (Das et al., 2021a, b). These noble metal nanostructures show brilliant colors in their colloidal solutions as the LSPR occurs at visible wavelengths. The LSPR helps resonant absorption as well as scattering of the incident radiation. Therefore, the incident light frequency is effectively coupled with the LSPR of noble metal nanoparticles, which largely enhances the magnitude of local electromagnetic field intensity at the nanoparticle surface. As a result, the original low intensity of Raman scattering of a molecule is enhanced (Fig. 2) upon adsorption at the vicinity of plasmonic nanostructures, leading to SERS (Jeanmaire and Van Duyne, 1977).

Though the exact working mechanism underlying the huge signal enhancement is often debated, principally there are two major driving forces behind the SERS signal enhancement. One is the electromagnetic field (EM) theory proposed by van Duyne and co-workers (Jeanmaire and Van Duyne, 1977), the other is a chemical enhancement (CE) theory suggested by Albrecht and Creighton (Albrecht and Creighton, 1977). The EM mechanism occurs when the incident light is exposed to the metal surface. If the frequency of the incident light is in resonance with surface conduction electrons, the collective oscillation enhances the strength of electromagnetic field in a specific area, called the plasmonic ‘hot spot’ (Stiles et al., 2008). The enhancement of Raman signals from molecules by the EM mechanism occurs when the molecules reside close to the hot spots and fall off rapidly with increasing distance from the surface (Jeanmaire and Van Duyne, 1977). The CE mechanism is related to the interactions between metal surfaces and chemisorbed molecules. The molecules adsorbed onto the metal surface generate charge transfer intermediates that can produce Raman excitation photons (Stiles et al., 2008). Overall, the EM mechanism is considered to be the dominant force, contributing 4–10 orders of magnitude over the CE mechanism, which contributes only 1–3 orders of magnitude in total

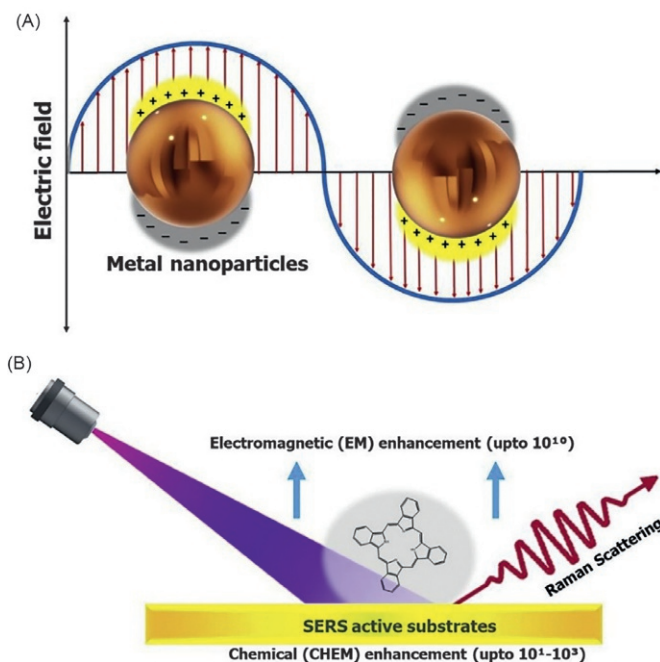


Fig. 2 A schematic illustration of the (A) collective oscillation of the free electrons after illuminating the nanoparticles with monochromatic light with resonance wavelength, (B) role of the electromagnetic and chemical enhancement factor in SERS. Adapted with permission from Panikar SS, Cialla-May D, De la Rosa E, Salas P, and Popp J (2021) Towards translation of surface-enhanced Raman spectroscopy (SERS) to clinical practice: Progress and trends. *Trends in Analytical Chemistry* 134: 116122–116142.

enhancement (Panikar et al., 2021; Stiles et al., 2008). In recent decades, SERS has evolved as an ultrasensitive vibrational spectroscopy technique to detect low-concentration analytes located on or near the surface of plasmonic nanostructures. It has found extensive uses in sensing various analytes directly attached to SERS substrates, such as biomolecules (e.g., proteins, nucleic acids, and pathogens), environmental contaminants, drugs, explosives, and adulterated foods (Das and Agrawal, 2011).

Conventionally, the analysis of soil components requires on-site sample collection followed by multi-step, complex, and time-consuming laboratory analyses. This process is labor intensive, time-consuming, and often costly because samples are collected and packaged on-site in the field, transported to the laboratory, and then processed following specialized protocols for analysis by specific instruments. To make the whole process of analysis simpler and less time consuming, molecular spectroscopic methods, especially Raman spectroscopy, could play beneficial roles. It enables rapid on-site analysis of samples with minimal sample collection and preparation, and provides quick, specific, and highly sensitive results. Conventional Raman spectroscopy can deliver information and data on the physico-chemical composition of components present in soil samples. Furthermore, it is possible to characterize multiple sample components simultaneously because of the narrow Raman spectral bands. One potential drawback of Raman spectroscopy when analyzing multiphase complex samples like soils, is interference of the intense background fluorescence signals coming from soil organic matter (e.g., humic substances), with the weak Raman signals. The SERS method can be used to address this problem because of the amplification of vibrational scattering of sample molecules by up to $\sim 10^{10}$ fold. Typically, Raman spectroscopy is most suitable for analyzing soil samples with high analyte concentration and low organic content due to the weak Raman scattering and strong fluorescence signals from soil organic matter (SOM). The SERS method can enhance Raman scattering many fold, which quenches fluorescence interference and allows the detection of low-concentration analytes with high sensitivity. Therefore, SERS methods have the potential to reduce the sample refinement process prior to analysis with its huge enhancement of the Raman signal and the ability to quench fluorescence. These are the reasons why regular Raman spectroscopy and SERS have found uses in many examples of analyzing soil components and contaminants.

Previous studies on the application of Raman spectroscopy and SERS in soil analysis can be classified into two categories: Analysis of soil components, and Analysis of soil contaminants. We will discuss these studies in detail in the following sections.

Application of Raman spectroscopy and SERS in the analysis of soil

Application of Raman spectroscopy and SERS in the analysis of soil components

In this section, we discuss potential applications of Raman spectroscopy and SERS to analyze different soil components, including soil organic matter, humic substances, phenolic compounds, and phosphorus.

Classification of soil samples

Soil classification is a crucial step in preparing for high quality cultivation. Luna et al. used Raman spectroscopy to differentiate soil types in Brazil (Luna et al., 2014). Five soil samples from different fields were collected and oven-dried followed by disaggregation through sieving. Soils were homogenized and compacted into tablet forms before analyses. For representative sampling, Raman scans were performed at 25 different points on each tablet-shaped sample within an area of $200\ \mu\text{m} \times 200\ \mu\text{m}$. Soil identification was achieved based on the fingerprinting of Raman spectra of the soils. They designed a robust characterization method for the soil samples utilizing Raman spectroscopy coupled with chemometric tools. The chemometric tools involve a genetic algorithm (GA) method and support vector machine (SVM). The combination of Raman spectroscopy and chemometric tools resulted in high specificity, sensitivity, and accuracy in the validation steps. Their method provides rapid differentiation of soil samples from different regions. High frequency regions of different Raman spectra indicate the presence of compounds with aromatic rings while the low frequency region of the Raman spectra reveals the presence of minerals and metal oxides (Fig. 3). The results obtained from their study detected elements that constitute oxides, humic compounds, and organic matter in metal complexes (Ca and Fe) in the soil samples tested. By integrating with chemometric tools, their method offers comprehensive information on the compositional features of the samples analyzed.

Analysis of soil organic matter

Soil organic matter (SOM) is one of the important indicators of soil quality and fertility. They play vital roles in water retention, nutrients cycling, and aggregate formation. Xing and coworkers combined Raman spectroscopy with Fourier-transform infrared-photoacoustic spectroscopy (FTIR-PAS) to predict SOM with the aid of partial least squares regression (PLSR) analysis (Xing et al., 2016). Composite samples were prepared by mixing the pure regional samples made of black soil, fluvo-aquic soils, paddy soils, and red soils from different regions of China. The composite soil was dried, sieved, and sampled in polythene bags prior to analysis. The soil samples were scanned from 180 to $3200\ \text{cm}^{-1}$ at room temperature for acquisition of Raman spectra with a portable Raman spectrometer (Fig. 4). They determined the optimal number of analytical factors by cross-validation, namely leave-one-out. They detected SOM with Raman spectroscopy and FTIR-PAS independently and obtained statistically significant values of R^2 and relative percentage difference (RPD) values; 0.75 and 1.96 (Raman) and 0.74 and 1.9 (FTIR-PAS) respectively for validation sets. The IR spectroscopy is often considered complementary to Raman spectroscopy as bonds that are Raman-inactive may be IR-active and vice versa. Therefore, they combined both for a more accurate analysis of the SOM. Compared to single application of Raman spectroscopy or FTIR-PAS, their combination made prediction of SOM more accurate with R^2 and RPD values of 0.81 and 2.18 , respectively for validation sets. A good fit for model analysis technically shows higher R^2 and RPD values. In conclusion, with the integration of Raman spectroscopy and FTIR-PAS an efficient and affordable detection of SOM is possible.

Humic substances and phenolic compounds are among the main components and active elements of the SOM. To determine the biological effects, it is important to analyze the chemical composition. The chemical structure and properties of the humic substances are not well-understood yet and are sometimes controversial. One thing is clear that there are different chemical

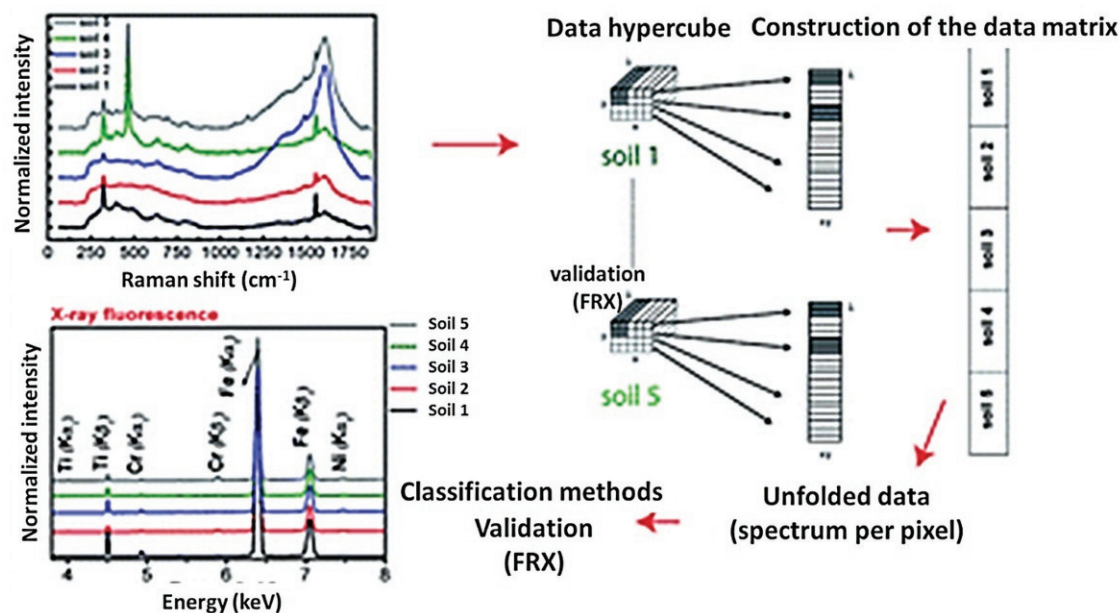


Fig. 3 Schematic diagram of the classification of soil samples based on Raman spectroscopy and X-ray fluorescence spectrometry combined with chemometric methods and the selection of variables. Adapted with permission from Luna AS, Lima ICA, Rocha WFC, Araújo JR, Kuznetsov A, Ferreira EHM, Boqué R, and Ferré J (2014) Classification of soil samples based on Raman spectroscopy and X-ray fluorescence spectrometry combined with chemometric methods and variable selection. *Analytical Methods* 6: 8930–8939.

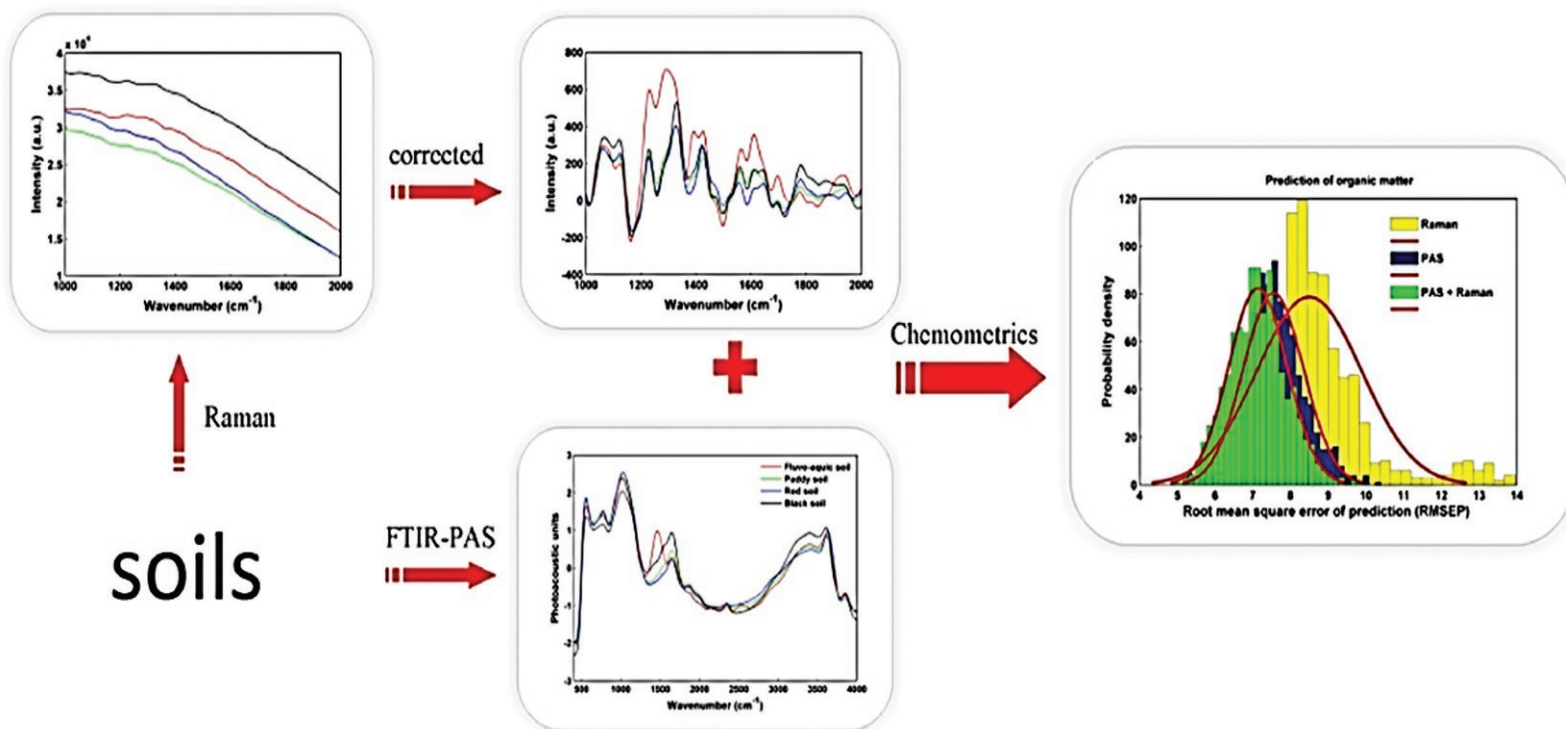


Fig. 4 Chemometrics of extracted data from FTIR-PAS and Raman spectra for detection and analysis of organic matter in farmland soil samples. Adapted with permission from Xing Z, Du C, Tian K, Ma F, Shen Y, and Zhou J (2016) Application of FTIR-PAS and Raman spectroscopies for the determination of organic matter in farmland soils. *Talanta* 158: 262–269.

components comprising humic substances in soil. Humic substances play important roles in crop productivity, therefore research to determine their chemical composition and properties is much needed. With the help of SERS and Fourier transform infrared (FT-IR), Muscolo et al. observed the biological effects of humic substances, and water-soluble soil phenols on antioxidant activities, growth, proteins, carbohydrates, vitamins, and phenols in roots and shoots of *Pinus laricio* (Muscolo et al., 2020). Soil samples were collected under *Pinus laricio* trees on Mount Peripoli, San Lorenzo, Italy. KOH solution (0.1 M) was used to extract humic substances from the dried soil samples. The extract was incubated with colloidal silver for SERS measurements. The FT-IR and Raman spectroscopies were used in a complementary way to determine the chemical structure and composition of the humic substances. In general, the Raman scattering and FTIR absorption provide complementary information of molecular vibrations. The former is active for symmetric vibrations that alter the polarizability while the latter is functional for anti-symmetric vibrations that alter the dipole moment. The IR absorption spectroscopy, which is particularly sensitive to polar bonds such as O—H or N—H, is often used for identification of functional groups of molecules. On the other hand, Raman scattering spectroscopy, which is active for bonds such as C=C, C—S, or S—S, is applied to identify skeletal or framework structures of molecules (Hashimoto et al., 2019).

Analysis of soil phosphorus

Phosphorus content in soil is an important factor in promoting root growth, winter hardiness, maturity, and tillering in plants. Zheng et al. used Raman spectroscopy to detect phosphorus concentration in 15 sandy soil samples (Zheng et al., 2012). In this research, to achieve their objectives, they designed their project involving three sub-objectives, i.e., preparation of different types of sandy soil samples containing varying concentrations of phosphate, measurement of highly precise Raman spectra for soil samples in the same condition, and effective data processing to establish an accurate model to predict phosphorus concentration based on Raman spectra. Briefly, pure sandy soil (base soil) was leached to remove organic matter and nutrients. Calcium phosphate and mixed phosphates were added to the base soil. The mixture was dried to prepare final soil samples containing different phosphate concentrations. They measured the signature Raman spectra of phosphorous content in the soil sample and derived a relationship between the concentration of phosphorus and their corresponding Raman spectra. All Raman spectra of soil samples were acquired within the wavenumber range of 239 cm^{-1} to 4045 cm^{-1} with a resolution of 4 cm^{-1} and integration time of 20s. Furthermore, they established a partial least squares (PLS) model to predict phosphorus concentration in the soil. It was also found that different phosphate compounds showed different Raman intensities at a fixed concentration of phosphate. The order of Raman intensity was aluminum phosphate > mixture of phosphates > magnesium phosphate > calcium phosphate. The phosphate Raman bands ranged from 345 cm^{-1} to 545 cm^{-1} and from 1311 cm^{-1} to 1402 cm^{-1} . Their analysis showed a good coefficient of determination (R^2) both for calibration ($R^2 = 1$) and validation ($R^2 = 0.937$), indicating that the method is promising for detecting varying concentrations of phosphorus in soil samples.

Application of Raman spectroscopy and SERS in the analysis of soil contaminants

In this section, we discuss major applications of Raman spectroscopy and SERS-based techniques for the analysis of soil contaminants, including various organic and inorganic ones such as thiram, organophosphorus pesticides, chlorpyrifos, thiabendazole, persistent organic pollutants (POPs), and various heavy metals.

Analysis of organic contaminants in soil

Detection of thiram in soil

Organic contaminants can be released into soils due to agricultural activities, such as pesticide applications, or industrial and municipal waste disposal, such as bio solids from wastewater treatment plants. Thiram is one of the major pesticides used to prevent crop damage and deterioration during storage or transport. Due to potential environmental issues caused by off-target exposure, detection of thiram in soil samples is crucial to evaluate its concentration and impact on soil safety. Zhang et al. reported a novel SERS technique to detect thiram selectively and sensitively in soil (Zhang et al., 2017). They developed a SERS substrate using triangular silver nanoparticles (TSNPs) of suitable nano size (29 nm) and with sharp corners to detect thiram residue in soil samples (Fig. 5). The soil samples collected were powdered, dried, and mixed with different concentrations of thiram. Ethanol was used to extract thiram from the soil and centrifuged to obtain the thiram rich supernatant. Finally, soil extracts containing different concentrations of thiram were mixed with TSNP colloids. The thiram-containing TSNPs were scanned under a Raman spectroscopy. Due to their strong SERS signal enhancement, it was found that the limit of detection (LOD) value was lower than the LODs obtained by other SERS substrates, including nanocubes and nanospheres. Their method exhibited good linearity in the concentration range of 0.12 to $4.8\text{ }\mu\text{g g}^{-1}$ and a low LOD of 90 ng g^{-1} . The results show great potential and applicability of TSNPs as novel SERS substrates in the detection of pesticides in soil samples.

Detection of organophosphorus pesticides in soil

Guselnikova and co-workers developed a surface plasmon polariton (SPP)-supported and a metal–organic framework compound (MOF-5)-coated gold nanosubstrate for selective, reproducible, and highly sensitive SERS detection of organophosphorus pesticides (Guselnikova et al., 2019). The MOF-5 is a class of metal–organic framework compounds with the formula $\text{Zn}_4\text{O}(\text{BDC})_3$, where BDC stands for 1,4-benzodicarboxylate. Due to the unique porous structure, MOFs were used to absorb pesticide molecules

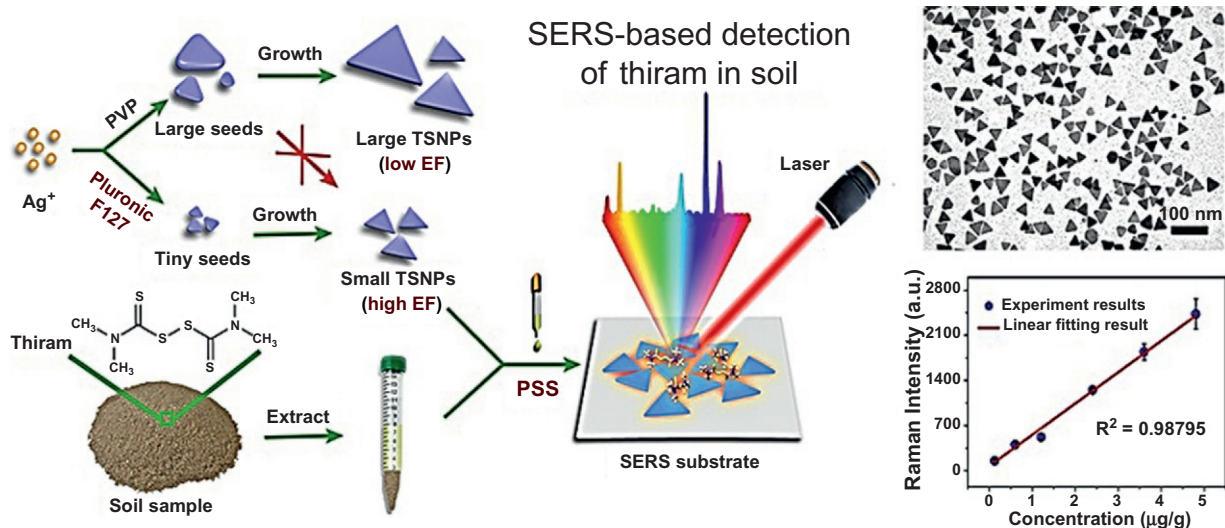


Fig. 5 Synthesis of highly plasmonic small and sharp triangular silver nanoplates utilizing tiny triangular nuclei with their excellent SERS activity for selective detection of thiram residue in soil. Adapted with permission from Zhang C-H, Zhu J, Li J-J, and Zhao J-W (2017) Small and sharp triangular silver nanoplates synthesized utilizing tiny triangular nuclei and their excellent SERS activity for selective detection of thiram residue in soil. *ACS Applied Materials & Interfaces* 9: 17387–17398.

efficiently and enhance their analytical signals. The immobilization of MOFs on the plasmon-active surface can promote selective entrapping of the analyte molecules, increase utilization of plasmonic hot-spots as well as enhance molecular resonance phenomena, through the charge transfer between the MOFs and analyte molecules. The SPP is a modified version of surface-plasmon resonance (SPR) and it involves both charge motion on the metal surface (surface plasmon) and electromagnetic propagation in air or a dielectric medium (polariton). In particular, SPPs are electromagnetic waves that originate at and travel along the interfaces of plasmonic materials and dielectric materials, practically in the range of infrared or the visible frequency. As the soil analysis was performed in a gold–air interface, a large number of SPPs were generated and enhanced SERS signals were observed when organophosphorus pesticides were detected. They used simulated soil samples composed of homogenized soil, organophosphate pesticides, and water. The mixture was vortex-mixed and dried. The dried soil samples were again suspended in water before centrifugation. The supernatants were collected and incubated with the MOF-based SERS substrate for spectroscopic studies. High reproducibility of SERS signals was achieved due to the homogeneous distribution of surface plasmons along the Au surface (Fig. 6). The developed gold and MOF-based SERS substrate performed well to detect organophosphorus pesticides with a low limit of detection (up to 10^{-12} M). Apart from organophosphorus pesticides, other organic contaminants, e.g., mycotoxins, and azo dye from spiked soil samples were also detected by the SERS substrate. The method is suitable for on-site detection since the detection was carried out by a handheld Raman spectrometer.

Detection of chlorpyrifos in soil

Chlorpyrifos (CPF) is a well-known pesticide, widely used to prevent and control crop-pests and diseases, though its excessive use causes environmental pollution. Compared to the traditional methods to analyze pesticides in soil samples, SERS-based techniques have gained credibility due to their ultrasensitive and specific detection ability. He et al. used gold nanoparticles of different sizes as SERS substrates to determine the concentration of CPF residue in soil (Fig. 7) (He et al., 2019). In their study, first, CPF solutions of different concentrations were mixed with the soil samples. The CPF residue was extracted from the soil samples following the QuEChERS method. Briefly, the soil sample was homogenized, dissolved, and centrifuged. Then the supernatants were put through a solid phase extraction process. The extracts were then incubated with the gold nanoparticles before SERS measurements. It was found that with increasing CPF concentration from 0.025 to 9.54 mg kg⁻¹, the characteristic SERS peaks of CPF at 674 cm⁻¹ (benzene ring), 610 cm⁻¹ (C-Cl), and 529 cm⁻¹ (P=S) increased gradually. Moreover, the AuNPs prepared enabled ultra-sensitive and reproducible SERS analysis yielding a limit of detection (LOD) as low as 10 μg L⁻¹ in the spiked soil sample. With real soil samples and using the partial least squares regression (PLSR) model, the LOD of CPF residue was as low as 0.025 mg kg⁻¹ with a coefficient of determination (R^2) of 0.977. With a simple method to synthesize gold nanoparticles as SERS substrates with ultra-sensitivity and high reproducibility, the SERS method is very promising to determine pesticide residues in soil.

Detection of thiabendazole in soil

Thiabendazole (TBZ) is widely used as an antifungal, antiparasitic, and preserving agent to treat downy mildew, sclerotium blight, as well as root rot diseases in plants. Nonselective and excessive use of TBZ creates excess accumulation of pesticide residues in soil, leading to environmental pollution and soil compaction. Therefore, it is an important task to monitor TBZ concentrations in soils. Nie et al., used gold nanoparticles as a SERS substrate to enhance Raman signal intensity and determined the presence of TBZ

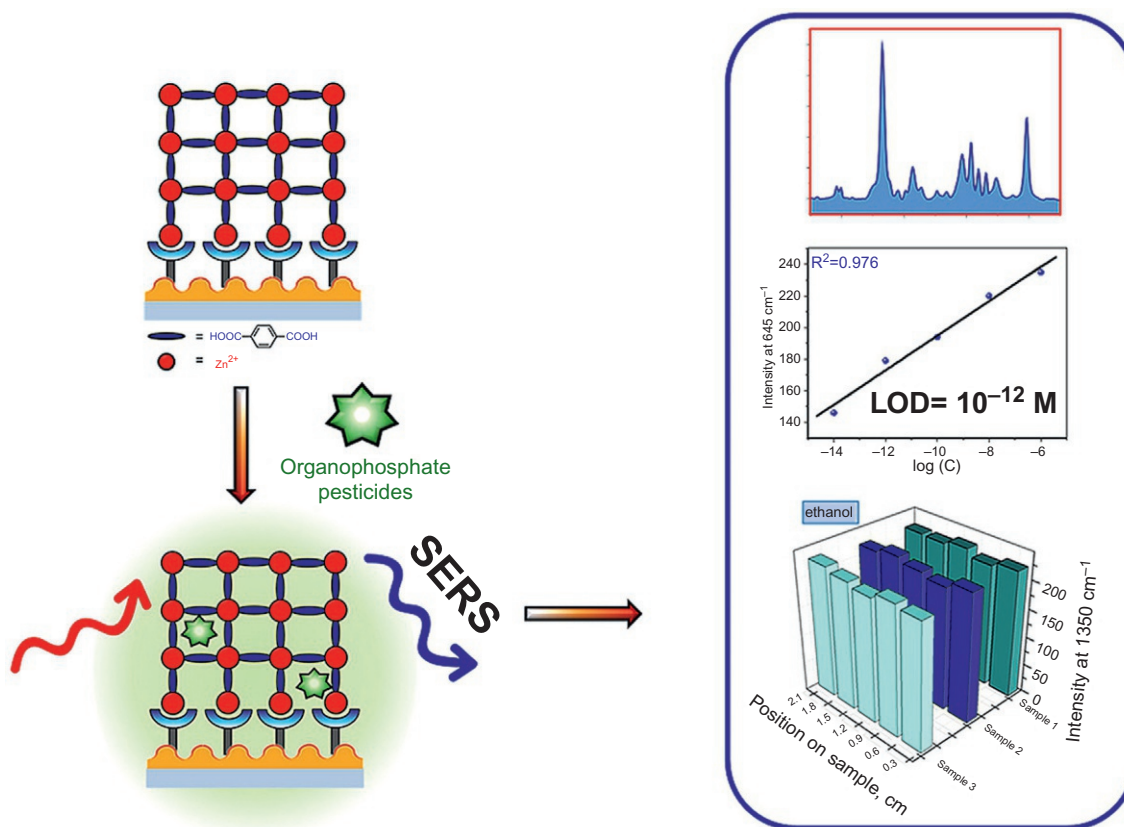


Fig. 6 A layer of MOF-5 is synthesized on the plasmonic gold grating. Porous MOF-5 absorbs organophosphate pesticides and they are detected by their SERS signal. Adapted with permission from Guselnikova O, Postnikov P, Elashnikov R, Miliutina E, Svorcik V, and Lyutakov O (2019) Metal-organic framework (MOF-5) coated SERS active gold gratings: a platform for the selective detection of organic contaminants in soil. *Analytica Chimica Acta* 1068: 70–79.

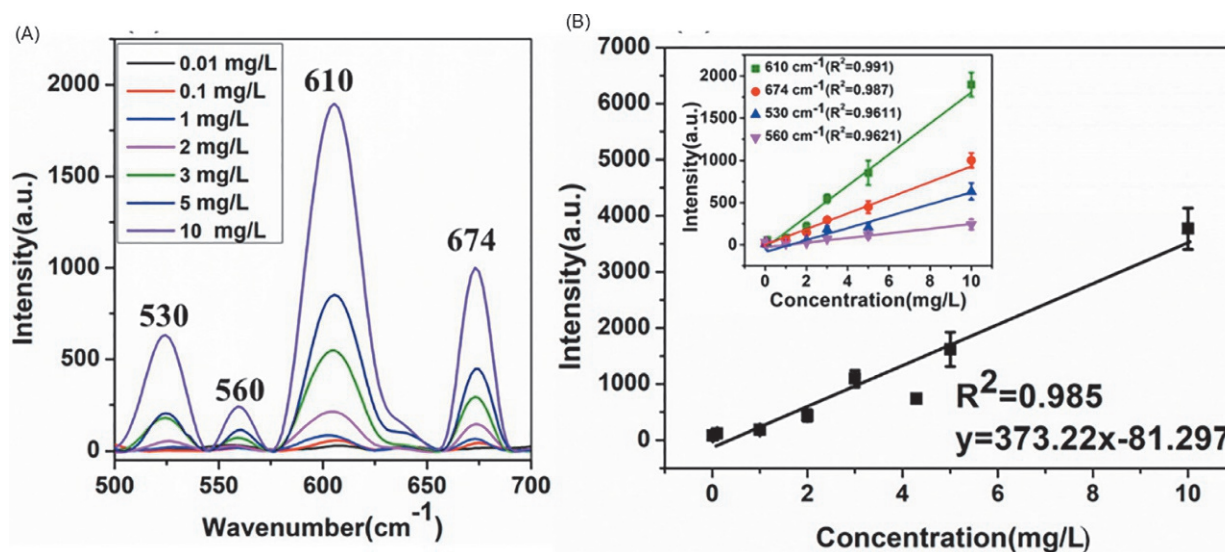


Fig. 7 The Raman spectra of CPF with gold nanoparticles at different concentrations (A) ranging from 0.01 to 10 mg L⁻¹, linear equation of Raman characteristic peak intensity and its concentration (B) at 530, 560, 610, 674 cm⁻¹ of the CPF molecule. Adapted with permission from He Y, Xiao S, Dong T, and Nie P (2019) Gold nanoparticles with different particle sizes for the quantitative determination of chlorpyrifos residues in soil by SERS. *International Journal of Molecular Sciences* 20: 2817–2828.

quantitatively in red soil extracts (Nie et al., 2018). The target analyte TBZ was extracted using solvent dissolution followed by soil matrix removal by centrifugation and membrane filtration. The TBZ-treated soil extracts were further incubated with plasmonic gold nanoparticles for SERS measurements. Their analysis showed the characteristic Raman peaks of TBZ in soil are at 784, 1008, 1270, 1328, 1406 and 1576 cm^{-1} . They calculated a limit of detection (LOD) value that reached 0.1 mg L^{-1} based on the peak at 784 cm^{-1} characteristics of the $\delta(\text{C—H})$ bond. A linear correlation between TBZ concentration in soil and Raman peak intensity was observed ($R^2 = \sim 0.995$) with a limit of quantification (LOQ) up to 1 mg L^{-1} . Overall, their technique was able to detect TBZ quantitatively in the soil extracts.

Detection of persistent organic pollutants in soil

Persistent organic pollutants (POPs), e.g., polychlorinated biphenyls (PCBs), polychlorinated benzene, and dioxins are major environmental pollutants and hazards to human health. Jency et al. reported a simple method to detect PCBs in soil samples using SERS with the help of β -cyclodextrine (β -CD) functionalized gold nanoparticles as the SERS substrate (Arockia Jency et al., 2015). The POP polluted soil samples collected from agricultural fields were ground into powder and sun-dried. Acetone was used to extract POPs from soil samples. The transparent supernatant was taken and incubated with the β -CD conjugated gold nanoparticles for SERS experiments. The β -CD is widely used in chemistry and pharmaceutical research for their ability to improve the solubility and stability of compounds with very low solubility in water. The β -CD molecules have a cavity at the center of the molecular arrangement and can thus form stable inclusion complexes with the analytes. The SERS spectral analysis showed that the functionalized gold nanoparticles of small size (9 nm) and large surface area exhibited strong SERS activity as collectively they produce more plasmonic junctions or SERS 'hot spots.' The C—H in-plane bending and C—C stretching modes suggested that the PCBs were adsorbed on the β -CD gold nanoparticle surface in a stand-on orientation. When the contaminated soil samples were added to AuNP- β -CD, the binding of PCB molecules with β -CD resulted in the aggregation of gold nanoparticles due to change in the medium hydrophobicity. This phenomenon produced excellent Raman signal enhancement. This method has the potential to detect PCBs, and to facilitate risk assessment and the removal of these hazardous materials from the environment.

Analysis of inorganic contaminants in soil

Detection of vivianite arsenates in soil

Soil contamination by inorganic chemical species, such as heavy metals, is of great concern due to their considerable effects on soil quality and safety. Arsenic compounds are often prevalent in soils because of their presence in mine tailings, cattle dips, or wood preservatives. Long-term persistence of arsenic compounds in soils together with their leachates can sometimes be strongly reactive and mobile, leading to the formation of highly stable metal arsenate compounds. One of such arsenate minerals is vivianite. Frost et al. used Raman spectroscopy to characterize vivianite arsenates in soil (Frost et al., 2003). Soil samples contaminated with vivianite arsenates of different divalent metal ions were ground, dried and placed on a polished metal surface on the stage of a Raman microscope. Raman spectroscopy enabled the identification of spectral bands corresponding to the mineral phosphate and arsenate units. Each mineral exhibits its own characteristic Raman spectrum that enables ready and rapid identification of the mineral type. For example, as the most common class of arsenic compounds available in soil as contaminants, vivianite arsenate has a characteristic Raman spectral band at 951 cm^{-1} .

Detection of harmful metalloids in soil

Soil water contaminated by mining and acid mine drainage often possesses high levels of harmful metalloids, some of them, e.g., Pb (lead), Cd (cadmium), and As (arsenic), are potentially dangerous due to their toxicity. Experts often use limestone fillers to immobilize the elements through adsorption and precipitation owing to its large content of carbonate and low cost. Pérez-Sirvent et al. analyzed fillers with the help of Raman spectroscopy and X-ray diffraction and reported that the former technique was more beneficial for assessing the physicochemical behavior of the fillers (Pérez-Sirvent et al., 2019). First, the heavy metal ions were absorbed into the porous limestone fillers which worked as a matrix for the metal ions. Then, the Raman spectra of the absorbed ions were collected and compared to the crystalline ones as references. After the absorption process, the heavy metal salts formed new bonds with the filler materials, which generated greater Raman intensity. This is how the presence of heavy metals and metalloids were detected by Raman spectroscopy. For example, As formed an As—O—H bond on absorption by filler material and showed a more intense band at 1088 cm^{-1} . Raman spectroscopy was found to be a very useful tool to monitor the process of sorption of potentially harmful metalloids (precipitation for Cd^{2+} , Pb^{2+} , Zn^{2+} , Cu^{2+} and adsorption for As^{5+}) on limestone filler surfaces because the percent increment was calculated from the characteristic Raman spectral bands. One of the drawbacks of the method is that it is qualitative.

Detection of lead in roadside soil

One interesting application of Raman spectroscopy is the detection of vehicle-released pollutants in roadside soils and plants. Carrero et al. carried out a study to assess the effects of urban pollutants on roadside soils and to understand the release mechanism of the metals (Carrero et al., 2012). They selected soils, guardrails, and plants near to busy roads in the urban area of Bilbao, Spain. The contaminated soil samples were powdered, dried, and placed on a plate for Raman measurements. For example, Spain banned leaded gasoline in 2001, but they found accumulated lead (Pb) in roadside soils during the study in 2011. Raman and energy dispersive X-ray fluorescence (m-ED-XRF) analysis of the selected soil sample indicated the presence of litharge (PbO). This type of

lead compound is not easily removed and accumulates in the top layer of soil. Therefore even after phasing out leaded gasoline more than a decade before, a large concentration of Pb was detected. Their analysis was a qualitative one, but it was able to detect litharge in the roadside soils.

Detection of Zn(II) in soil

Soil contamination by heavy metals is a common hazard faced by many industrially developed and developing countries. Szabó et al. reported a novel method for the detection of heavy metals with the help of SERS techniques (Szabó et al., 2011). As metal ions do not exhibit Raman signals by themselves, a chelating agent 4-(2-pyridylazo) resorcinol (PAR) for Zn(II) was used to determine SERS spectra of the Zn(II)-PAR complex in contaminated soil samples. To enhance the detection limit, they attached the chelate complex to a SERS substrate, silver colloid. The Zn(II) ions were extracted from the soil by the acid–water solvent extraction method. When PAR was added to the extracted solution, the color of the solution turned red indicating the formation of a metal-PAR complex. Then the Zn(II)-PAR extract solution was added to the silver colloid. The color of the silver colloid changed to green, indicating adsorption of the Zn(II)-PAR on the silver surface. The PAR on complex formation with Zn(II) gives rise to a new Raman peak at 1335 cm^{-1} . The Raman spectra of PAR alone, the Zn(II)-PAR chelate complex, and PAR–soil mixture were recorded. The SERS spectra of Zn-contaminated soil extracts and the Zn(II)-PAR standard complex were in good accord in terms of spectral profile and band position. They used SERS of the Zn(II)-PAR complex to detect Zn(II) in the contaminated soil samples and later on quantified its concentration with UV–Vis spectroscopy. The Zn(II) concentration in the contaminated soil was determined (390 mg kg^{-1}) which is in good agreement with the standard values reported by the Romanian Environmental Protection Agency. In the current paper, the authors reported that the UV–Vis spectra cannot distinguish between two different heavy metals of a PAR complex, e.g., Zn(II) and Cu (II). They did not provide any explanation as to why SERS was not used for the quantitative detection of Zn(II), especially given that it is a more sensitive and molecular-selective tool than UV–Vis.

Conclusion and outlook

This chapter covers the fundamental principles and mechanisms of Raman spectroscopy and SERS, and their applications in soil analysis. Raman spectroscopy is a spectroscopic technique that detects vibrational modes of molecules. In general, it is a non-invasive chemical analysis technique providing in-depth data and information about the chemical composition, concentration, and structure. Raman spectra are used to extract unique ‘fingerprint’ vibrational modes of the sample molecules and provide chemical, physical, and structural properties of the molecules. Raman spectroscopy has several advantages, including its suitability for analyzing a wide range of inorganic and organic molecules, non-invasive analysis, no water interference, minimal sample preparation, and small sample size requirement. These are the reasons why Raman spectroscopy has been used in many instances for analyzing soil components and contaminants. Despite the above-mentioned advantages, there are also several major disadvantages, i.e., weak signal intensity leading to lower sensitivity and the background fluorescence signals of impurities that overshadow the Raman peak of target molecules. To overcome the limitations, an advanced technique based on surface plasmon resonance of the plasmonic nanostructures was introduced, i.e., surface-enhanced Raman spectroscopy (SERS). The SERS enabled by plasmonic nanostructures (typically noble metal nanoparticles) demonstrates strong selectivity, specificity, and sensitivity, allowing detection of a wide range of soil components and contaminants.

Although a variety of methods based on regular Raman spectroscopy and SERS have been reported for the detection of soil components and contaminants, there are several voids to be filled in the future.

1. Given that two major limitations of Raman spectroscopic methods, i.e., weak sensitivity and background fluorescence signals, SERS has emerged to be a better alternative to Raman spectroscopy. Therefore, development of novel SERS substrates with a greater signal enhancement factor (EF) should be a major point of focus in future. High EF could overshadow background signals coming from impurities. Moreover, SERS substrates with higher EF values (beyond $\sim 10^7$) could achieve limits of detection (LODs) below nanomolar concentrations of the target analytes.
2. The confocal Raman imaging technique, which has not been used for soil analysis previously, could be applied as a potential technique to investigate the spatial distribution of soil components. The technique was used successfully for analyzing the spatial distribution of proteins and phenolic compounds in grains (Piot et al., 2000) as well as silver nanoparticles adsorbed by hydrated minerals (Brittle et al., 2018). It is rapid and noninvasive without complicated sample preparation. The technique uses Raman bands as spectral markers to construct distribution maps of chemical species. High spatial resolution of the Raman mapping technique reveals chemical and molecular heterogeneity within 3-D samples. Thus, different spatially distributed components of soil samples can be chemically as well as structurally analyzed by the technique.
3. Raman spectroscopy and SERS-based techniques acquire a large number of vibrational spectra from a mixture of complex components during analysis or imaging. Classical linear methods of data processing are no longer adequate for analyzing data of complex systems. Advanced multitasking machine learning methods are required. In recent years, machine learning has found potential applications in analytical science by extracting information from big and complex datasets in spectroscopy. Various machine learning methods have been used to analyze food colors, additives, meat types in complex food samples, biomolecules (proteins, lipids) present in cancer tissue samples, and illicit drugs present in forensic samples (Lussier et al., 2020). Soil samples are extremely complex in nature containing organic, inorganic, biotic, abiotic, solid, gaseous, and liquid components. Therefore,

suitable machine learning techniques in combination with Raman spectroscopy and SERS tools are promising to provide specific analysis of the components in complex soil samples in the future.

4. A Raman spectroscopy and SERS based 'soil-on-chip' microfluidic platform is expected to be a breakthrough technology in soil analysis. Though still in its infancy, microfluidic soil chip or 'soil-on-chip' techniques allow direct spectroscopic analysis of soil components and contaminants in well-controlled microfluidic channels (Stanley et al., 2016). Raman spectroscopy and SERS can play a major role as a technique in 'soil-on-chip' application due to their strong signal enhancement, small sample volume requirement, and noninvasive nature of analysis. The proposed 'soil-on-chip' sensing platform has the following potential applications: (i) determination of spatial and time resolved distribution, transport, and transformation of organic and inorganic pollutants in soil, (ii) analysis of SOM dynamics to understand soil carbon storage more comprehensively, (iii) identification and classification of soil microbes, and (iv) understanding chemo taxis of soil microbes at the micro scale spatial resolution.
5. Another important future aspect is the integration of SoilCards with Raman spectroscopy/SERS techniques. SoilCards are paper-based test-strips that analyze plant nutrients in a quick, easy, and affordable fashion, especially for underprivileged farmers without access to soil testing laboratories. These paper strips detect basic soil nutrients (N, P, K) and pH of the soil samples through visible color changes for corresponding elements. A combination of SERS nanosubstrates with the conventional strips could enhance their sensitivity and detection ability many fold, which should be explored in the future.

Acknowledgment

Funding information: Startup fund by the Research Foundation for the State University of New York.

References

- Albrecht MG and Creighton JA (1977) Anomalous intense Raman spectra of pyridine at a silver electrode. *Journal of the American Chemical Society* 99: 5215–5217.
- Arockia Jency D, Umadevi M, and Sathe GV (2015) SERS detection of polychlorinated biphenyls using β -cyclodextrin functionalized gold nanoparticles on agriculture land soil. *Journal of Raman Spectroscopy* 46: 377–383.
- Brittle SW, Foose DP, O'Neil KA, Sikon JM, Johnson JK, Stahler AC, Ryan J, Higgins SR, and Sizemore IE (2018) A Raman-based imaging method for characterizing the molecular adsorption and spatial distribution of silver nanoparticles on hydrated mineral surfaces. *Environmental Science & Technology* 52: 2854–2862.
- Carrero JA, Arana G, and Madariaga JM (2012) An important use of Raman spectroscopy to help understand the impact of traffic on roadside soils and plants. *Spectroscopy Europe* 24: 5–12.
- Cho YC and Ahn SI (2020) Fabricating a Raman spectrometer using an optical pickup unit and pulsed power. *Scientific Reports* 10: 11692–11699.
- Das RS and Agrawal YK (2011) Raman spectroscopy: Recent advancements, techniques and applications. *Vibrational Spectroscopy* 57: 163–176.
- Das A, Choi N, Moon JI, and Choo J (2021a) Determination of total iron-binding capacity of transferrin using metal organic framework-based surface-enhanced Raman scattering spectroscopy. *Journal of Raman Spectroscopy* 52: 506–515.
- Das A, Kim K, Park SG, Choi N, and Choo J (2021b) SERS-based serodiagnosis of acute febrile diseases using plasmonic nanopopcorn microarray platforms. *Biosensors & Bioelectronics* 192: 113525.
- Dijkstra RJ, Ariese F, Gooijer C, and Brinkman UAT (2005) Raman spectroscopy as a detection method for liquid-separation techniques. *Trends in Analytical Chemistry* 24: 304–323.
- Flieschmann M, Hendra PJ, and McQuillan AJ (1974) Raman spectra of pyridine adsorbed at a silver electrode. *Chemical Physics Letters* 26: 163–166.
- Frost RL, Klopogge T, Weier ML, Martens WN, Ding Z, and Edwards HGH (2003) Raman spectroscopy of selected arsenates - implications for soil remediation. *Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy* 59: 2241–2246.
- Gardiner DJ (1989) Introduction to Raman scattering. In: Gardiner DJ and Graves PR (eds.) *Practical Raman spectroscopy*, pp. 1–12. Berlin, Heidelberg: Springer.
- Guselnikova O, Postnikov P, Elashnikov R, Miliutina E, Svorcik V, and Lyutakov O (2019) Metal-organic framework (MOF-5) coated SERS active gold gratings: a platform for the selective detection of organic contaminants in soil. *Analytica Chimica Acta* 1068: 70–79.
- Hadjilivanov KI, Panayotov DA, Mihaylov MY, Ivanova EZ, Chakarova KK, Andonova SM, and Drenchev NL (2021) Power of infrared and Raman spectroscopies to characterize metal-organic frameworks and investigate their interaction with guest molecules. *Chemical Reviews* 121: 1286–1424.
- Hashimoto K, Badarla VR, Kawai A, and Ideguchi T (2019) Complementary vibrational spectroscopy. *Nature Communications* 10: 4411–4416.
- He Y, Xiao S, Dong T, and Nie P (2019) Gold nanoparticles with different particle sizes for the quantitative determination of chlorpyrifos residues in soil by SERS. *International Journal of Molecular Sciences* 20: 2817–2828.
- Jeanmaire DL and Van Duyne RP (1977) Surface Raman spectroelectrochemistry: Part I. Heterocyclic, aromatic, and aliphatic amines adsorbed on the anodized silver electrode. *Journal of Electroanalytical Chemistry* 84: 1–20.
- Kuhar N, Sil S, Verma T, and Umapathy S (2018) Challenges in application of Raman spectroscopy to biology and materials. *RSC Advances* 8: 25888–25908.
- Luna AS, Lima ICA, Rocha WFC, Araújo JR, Kuznetsov A, Ferreira EHM, Boqué R, and Ferré J (2014) Classification of soil samples based on Raman spectroscopy and X-ray fluorescence spectrometry combined with chemometric methods and variable selection. *Analytical Methods* 6: 8930–8939.
- Lussier F, Thibault V, Charron B, Wallace GQ, and Masson J-F (2020) Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering. *Trends in Analytical Chemistry* 124: 115796–115810.
- Muscolo A, Pizzeghello D, Francioso O, Sanchez Cortes S, and Nardi S (2020) Effectiveness of humic substances and phenolic compounds in regulating plant-biological functionality. *Agronomy* 10: 1553–1566.
- Nie P, Dong T, Xiao S, Lin L, He Y, and Qu F (2018) Quantitative determination of thiabendazole in soil extracts by surface-enhanced Raman spectroscopy. *Molecules* 23: 1949–1963.
- Panikar SS, Cialla-May D, De la Rosa E, Salas P, and Popp J (2021) Towards translation of surface-enhanced Raman spectroscopy (SERS) to clinical practice: Progress and trends. *Trends in Analytical Chemistry* 134: 116122–116142.
- Pérez-Sirvent C, Martínez-Sánchez MJ, Veiga JM, Bech J, and García-Lorenzo ML (2019) In situ chemical immobilisation by limestone filler of potentially harmful metal(loid) in contaminated soils: Monitoring by Raman spectroscopy. *Applied Geochemistry* 111: 104441–104447.
- Piot O, Autran J-C, and Manfait M (2000) Spatial distribution of protein and phenolic constituents in wheat grain as probed by confocal Raman Microspectroscopy. *Journal of Cereal Science* 32: 57–71.
- Raman CV and Krishnan KS (1928) A new type of secondary radiation. *Nature* 121: 501–502.
- Smekal A (1923) Zur Quantentheorie der Dispersion. *Naturwissenschaften* 11: 873–875.

- Stanley CE, Grossmann G, Solvas XC, and DeMello AJ (2016) Soil-on-a-Chip: Microfluidic platforms for environmental organismal studies. *Lab on a Chip* 16: 228–241.
- Stiles PL, Dieringer JA, Shah NC, and Duyn RPV (2008) Surface-enhanced Raman spectroscopy. *Annual Review of Analytical Chemistry* 1: 601–626.
- Szabó L, Leopold LF, Cozar BI, Leopold N, Herman K, and Chiş V (2011) SERS approach for Zn(II) detection in contaminated soil. *Central European Journal of Chemistry* 9: 410–414.
- Xing Z, Du C, Tian K, Ma F, Shen Y, and Zhou J (2016) Application of FTIR-PAS and Raman spectroscopies for the determination of organic matter in farmland soils. *Talanta* 158: 262–269.
- Zhang C-H, Zhu J, Li J-J, and Zhao J-W (2017) Small and sharp triangular silver nanoplates synthesized utilizing tiny triangular nuclei and their excellent SERS activity for selective detection of thiram residue in soil. *ACS Applied Materials & Interfaces* 9: 17387–17398.
- Zheng L, Lee WS, Li M, Katti A, Yang C, Li H, and Sun H (2012) Analysis of soil phosphorus concentration based on Raman spectroscopy. In: Larar AM, Chung H-S, Suzuki M, and Wang J-Y (eds.) *SPIE Asia-Pacific Remote Sensing*, pp. 852718–852725. Tokyo: SPIE.

Fourier transform infrared spectroscopic methods of soil analysis

Derek Peak, Department of Soil Science, College of Agriculture and Bioresources, University of Saskatchewan, Saskatoon, SK, Canada

© 2023 Elsevier Ltd. All rights reserved.

Introduction	510
The nature of molecular vibrations	510
Infrared peaks of interest	511
Infrared spectral analysis	511
FTIR as a quantitative tool	513
Instrumentation	513
Common experimental techniques	513
Transmission-mode measurements	513
Example of transmission experiments	514
Diffuse reflectance infrared Fourier transform experiments	514
Attenuated total reflectance experiments	514
Example ATR experiments	515
The ATR analysis of soil organic matter (SOM)	517
Conclusions	517
References	517

Abstract

Fourier transform infrared (FTIR) spectroscopy is a vibrational spectroscopy that is useful in the study of a wide range of soil chemical processes. Topics included in this chapter include a brief explanation of the theory of molecular vibrations, a survey of useful vibrations for soil scientists, a description of different experimental approaches to measuring FTIR spectra, and current examples of using FTIR spectra to answer field and laboratory based soil chemical research questions.

Key points

- FTIR spectroscopy is useful in studying soil minerals, organic matter, and adsorption processes on soil surfaces.
- A variety of FTIR sampling approaches (diffuse reflectance and Attenuated Total Reflectance) are popular for soil samples.

Introduction

Fourier transform infrared (FTIR) spectroscopy is a form of vibrational spectroscopy that is useful in the study of a variety of soil chemical processes. In the mid-infrared (mid-IR) range, vibrations arise from many environmentally important molecules such as organic acids, soil organic matter, mineral phases, and oxyanions (Johnston et al., 1996). It is possible to utilize FTIR spectroscopy as a quantitative analytical method and also as a tool to determine bonding mechanisms in solids and on surfaces. Molecular vibrations can be related directly to the symmetry of molecules, and so it is often possible to determine precisely how a molecule is bonding on surfaces or as a component in a solid phase from its infrared spectrum. Many experimental methods exist for probing samples in various states and at different spectral regions.

The nature of molecular vibrations

The infrared region of the electromagnetic spectrum lies between 10 and 10,000 cm^{-1} , which can be expressed as a frequency of 10^{12} – 10^{14} Hz. This frequency overlaps with the frequency of molecular vibrations. When infrared radiation is at the same frequency as a molecular vibration, it can induce a transition to a higher energy state. This transition will only occur if it induces a dipole. A true molecular vibration is defined as a movement of the atoms of a molecule that do not change the center of mass and that involve no rotation. Vibrations that obey this definition are known as the ‘normal modes.’ For a nonlinear molecule, the normal modes may be due to symmetric or asymmetric stretching, and in-plane or out-of-plane bending. Not all of a molecule’s normal-mode vibrations will be active when subjected to infrared energy. Rather, the symmetry of a molecule will affect the position and number of infrared-active peaks.

Infrared peaks of interest

In the mid-IR range ($400\text{--}4000\text{ cm}^{-1}$), there are several extremely important and useful classes of bonds that can be studied. The FTIR spectroscopy has been used extensively by both geologists and soil scientists to study clay minerals and other inorganic soil components. The structure of minerals can substantially affect their infrared spectra, and entire textbooks are available with tables of peak positions and figures containing spectra of reference minerals (e.g. Farmer, 1974). Water has an intense infrared absorbance in the mid-IR region, which typically is considered a limitation of the technique for environmental research. However, the strong absorbance of water bands has been exploited by researchers who investigated the effects of humidity changes on self-supported films of clay minerals.

Many organic compounds also have diagnostic peaks in the infrared region, and FTIR has been used to study the nature of soil organic matter as well as the adsorption of organic acids. Many metals and metalloids form negatively charged oxyanion molecules in aqueous solution, and FTIR spectroscopy has been used extensively to study their adsorption and precipitation reactions. Fig. 1 contains infrared spectra from selected experiments using FTIR to study reactions of: (1) natural organic matter (NOM), and (2) inorganic minerals in soils. Infrared absorbance bands in Fig. 1a have been labeled to show how changes in functional-group makeup can give rise to new infrared peaks. Also notice how many of the peaks in the NOM disappear or shift in position when NOM is adsorbed on to a mineral surface. This is due to chemical bonding with carboxylic and phenolic NOM groups and the iron oxide surface. In Fig. 1b, a mineral transformation is monitored using FTIR spectroscopy. The initial $\gamma\text{-Al}_2\text{O}_3$ solid phase is not stable in water, and over the 2.5-month experimental period it transforms into a crystalline gibbsite $\text{Al}(\text{OH})_3$ phase.

Infrared spectral analysis

A detailed analysis of the theoretical infrared spectrum from first principles is not always required. It is possible simply to compare unknown samples to those of spectra with known molecular configurations to determine the structure of an

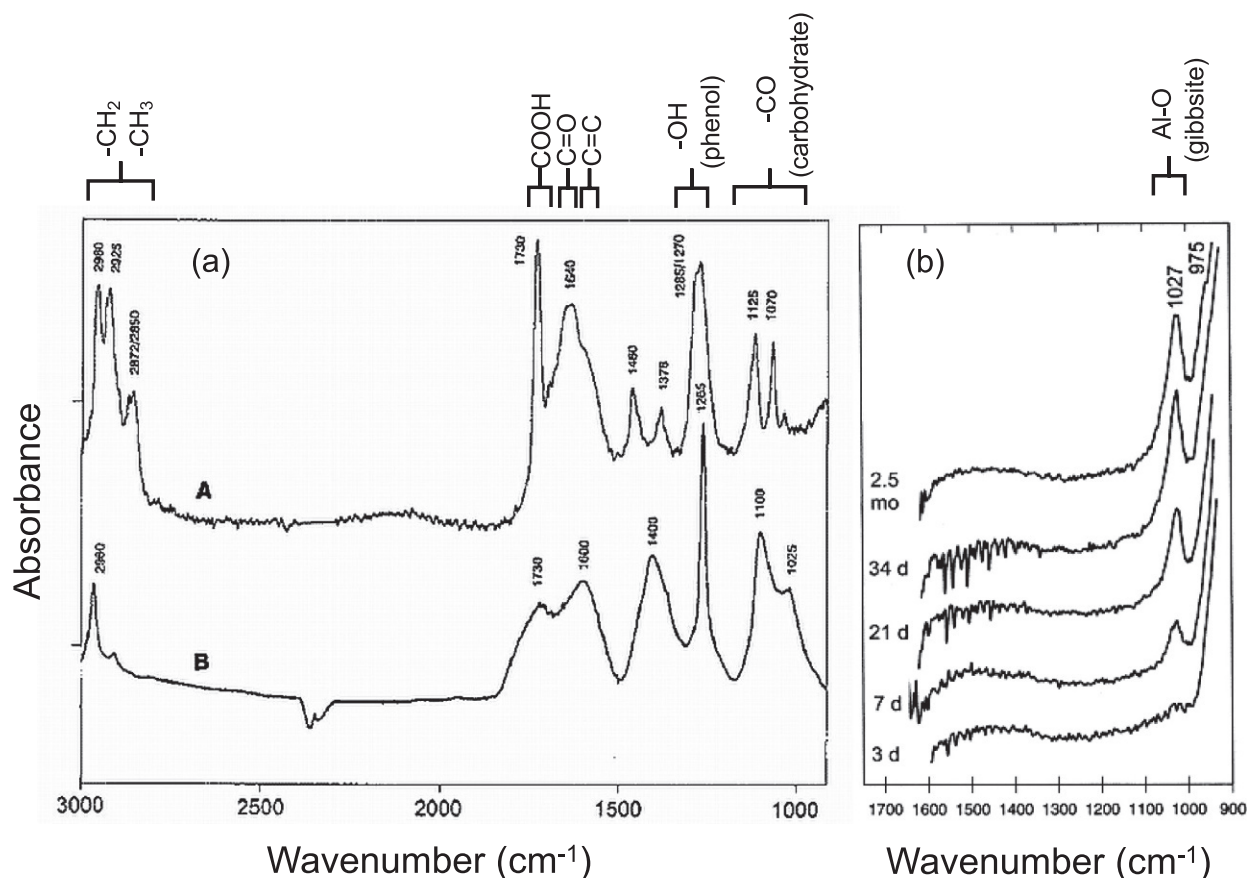


Fig. 1 (a) Infrared spectra of natural organic matter in water compared with natural organic matter (NOM) adsorbed on an iron oxide. Reproduced with permission from Gu B, Schmitt J, Chen J, Liang L, and McCarthy JF (1994) Adsorption and desorption of natural organic matter on iron oxide: Mechanisms and models. *Environmental Science and Technology* **28**: 28–46. (b) Infrared spectra of $\gamma\text{-Al}_2\text{O}_3$ aged from 3 to 75 days. The peak at 1027 cm^{-1} is due to transformation of the initial Al_2O_3 into a crystalline gibbsite phase. Adapted from Wijnja H and Schulthess CP (1999) ATR-FTIR and DRIFT spectroscopy of carbonate species at the aged $\gamma\text{-Al}_2\text{O}_3$ – water interface. *Spectrochimica Acta. Part A Molecular and Biomolecular Spectroscopy* **55**(4): 861–872, with permission.

unknown. This is referred to as spectral fingerprinting. This technique relies on one's ability to synthesize reliably reference compounds of known chemical structure or obtain their spectrum from a library. This approach is reasonably straightforward in the case of organic molecules, as large commercial software spectral libraries are available. In the case of unknown inorganic spectral identification, however, suitable standards can often be difficult to synthesize, and aqueous standards are often present as mixtures of species. For these reasons, assigning peaks based on rules from molecular symmetry is often necessary.

Symmetry rules assign a molecule to a unique point group based upon the arrangement of its atoms in space (Nakamoto, 1986). A common example of how infrared spectra may be related to molecular symmetry is the case of sulfate. The FTIR spectroscopy has been used by many researchers to study sulfate bonding in mineral structures and as surface adsorption complexes. The clear relationship between the molecular symmetry of sulfate species, its point group, and its infrared spectrum is an example of the utility of this approach. In the case of aqueous sulfate, there are two normal modes that are accessible to spectroscopic investigation. They are the nondegenerate symmetric stretching ν_1 , and the triply degenerate asymmetric stretching ν_3 bands. As a free anion in solution, sulfate has tetrahedral symmetry and belongs to the point group T_d . For this point group, only one broad peak, due to the ν_3 stretch, is observed at approximately 1100 cm^{-1} . If sulfate is covalently bonded to one or more cations, the symmetry is lowered, a ν_1 peak becomes infrared active, and the ν_3 band splits into more than one peak. When sulfate is bound as a monodentate complex, the ν_3 band splits into two peaks, while the ν_1 band becomes fully active at approximately 975 cm^{-1} , and C_{3v} point group symmetry results. If sulfate forms a bidentate binuclear (bridging) complex, the complex's symmetry is decreased to C_{2v} , and the ν_3 band splits into three bands between 1050 and 1250 cm^{-1} , while the ν_1 band is usually shifted to a higher wave number. The relationship between molecular symmetry and the number and position of infrared peaks is detailed in Fig. 2. Sulfate is typical of tetrahedral oxyanions, and so the relationship also may be readily extended to other tetrahedral oxyanions of environmental relevance (such as chromate, selenate, phosphate, and arsenate).

Similar rules may be derived for molecules of other point groups also, making the symmetry approach very powerful in the assignment of molecular configuration from FTIR spectra. However, one must be aware that infrared peaks may arise from many potentially occurring molecular vibrations. Of particular concern in environmental science is the fact that interactions with protons and water molecules will substantially affect the infrared spectrum of sulfate, not only when covalently bound to sulfate (as bisulfate or sulfuric acid), but also when hydrogen bonds are formed with waters of hydration. To improve peak assignment in such cases, experiments are usually repeated in D_2O (heavy water) as a solvent. Substitution of D (deuterium) for H (hydrogen) in surface complexes or mineral structures will shift the position of infrared peaks, while peaks that do not involve protons are unaffected. It is also possible to calculate theoretical infrared spectra using molecular modeling software. The absolute position of peaks may not be correct, but trends in how species are affected by the presence of protons or metals are still quite helpful in peak assignment.

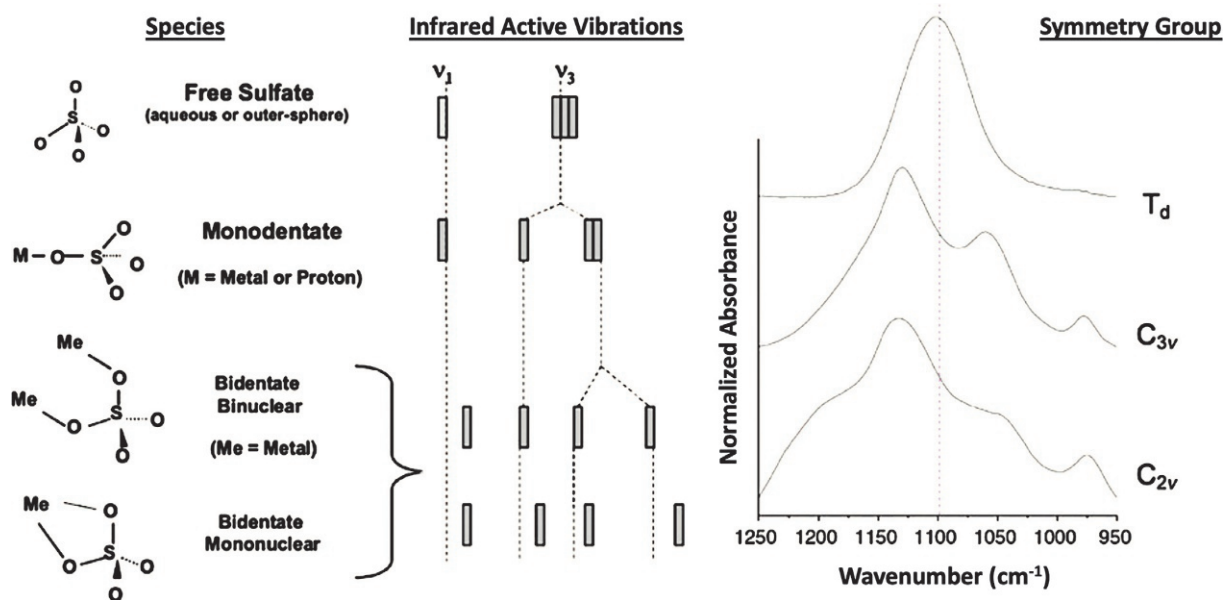


Fig. 2 Relationship between molecular symmetry and infrared spectra for aqueous, monodentate, and bidentate sulfate species. Adapted from Peak D, Ford RG, and Sparks DL (1999) An in situ ATR-FTIR investigation of sulfate bonding mechanisms on goethite. *Journal of Colloid Interface Science* **218**: 289–299.

FTIR as a quantitative tool

The infrared absorbance of a molecular vibration obeys the Beer-Lambert law. This means that FTIR spectroscopy, if properly calibrated, is a quantitative technique. The relationship is as follows:

$$A = \epsilon bC, \quad (1)$$

where A is absorbance, ϵ is the molar absorptivity coefficient, b is the thickness of the sample (or pathlength), and C is the concentration of the molecule in the sample. Note that ϵ is specific to a given vibration, and that A is typically reported in terms of peak height or integrated absorbance from a fitted peak. Changes in sample configuration (transmission vs. reflectance) for different types of FTIR experiments can quantification require spectral transformations or correction factors.

Instrumentation

FTIR refers to a class of spectrometers that are distinguished from older, double-beam infrared units by their use of a single beam and an interferometer and detector to collect simultaneously a spectrum over the entire wavelength of interest. These raw data, called an interferogram, are an oscillation over time that can then be converted to a function of frequency, as in a traditional infrared spectrum, by a mathematical procedure called Fourier transformation. The FTIR instruments have largely replaced older double-beam units because they have better spectral resolution, allow for much more rapid data collection, and have an improved signal-to-noise ratio.

Common experimental techniques

Several different FTIR experimental techniques are commonly employed by environmental researchers. These are transmission mode experiments with pressed disks, diffuse reflectance experiments with powder samples, and attenuated total reflectance studies with aqueous solutions and hydrated suspensions. The primary differences in the commonly used techniques are: (1) the amount of sample preparation that is required, (2) the spectral range that is accessible to the technique, and (3) the detection limits. Each technique may be particularly well suited to certain research questions, and more than one technique is sometimes required.

Transmission-mode measurements

Transmission-mode experiments are both the simplest and most common type of FTIR experiment. Samples are diluted in some medium and placed perpendicular to the infrared beam. A detector measures the amount of infrared beam that is transmitted through the sample. To obtain the spectrum of only the sample, the spectrum of the pure medium is first measured and used as a reference. A spectrum of the sample in the medium is then collected and the ratio to the reference determined to obtain the spectrum of the sample. A typical experimental configuration for transmission-mode experiments is shown in Fig. 3.

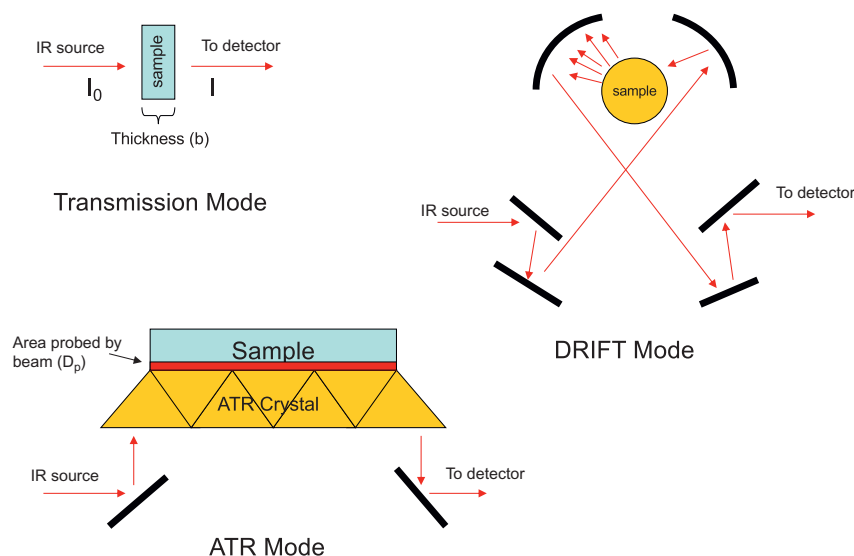


Fig. 3 Common sampling methodologies for transmission-mode, diffuse reflectance (DRIFT), and attenuated total reflectance (ATR) Fourier transform infrared experiments. I_0 , energy prior to the sample; I , energy after the sample; D_p , depth of penetration through the sample.

Solid samples are typically diluted to 1–2% wt in KBr (potassium bromide) salt and then made into a thin disk with a press. The final product is a thin, clear disk that is placed in a sample holder and analyzed. For transmission mode experiments to be successful, water must be removed from the system, so samples are heated to remove as much water as possible prior to dilution. Care must be taken with amorphous and unstable solids to avoid transformations that may occur at elevated temperatures. Organic materials can also often be studied in transmission mode by simply placing them in a thin liquid cell. The solvent should have a relatively low infrared absorbance and should not have infrared peaks that overlap with those of the organic compounds of interest.

Transmission-mode experiments can provide quantitative information more easily than some other experiments because Beer's Law is directly applicable. The path length through the sample must be known, and molar absorptivity can be determined by collecting spectra at several known concentrations and constructing a calibration curve. Once the path length and molar absorptivity are known, concentration and absorbance can then be determined.

Example of transmission experiments

This technique has been quite useful historically for soil scientists who are studying mineral transformations or who need to determine quantitatively the mineral composition in unknown samples. However, the technique involves substantial sample modification (removal of water by drying and vacuum; physical mixing with KBr) that may change the structure of both solid phases and adsorbed surface species. Particularly in complex systems, such as soils, the peaks of minor components such as clay minerals may be masked by other species of higher concentration such as quartz.

Recent advances in analytical capabilities have led to a resurgence in transmission mode FTIR spectroscopy as an important part of field-based soils research. Improved software packages designed to handle large spectral datasets and statistical approaches that determine correlation coefficients and principal components in those datasets now make it possible to deconvolute large (thousands and greater) datasets of transmission mode infrared spectra. In addition, advances in high throughput FTIR have made it possible to collect this quantity of spectra in a reasonable time frame. The limitation of these approaches is typically the preparation of KBr pellets for measurement.

Diffuse reflectance infrared Fourier transform experiments

The diffuse reflectance technique was developed partly to address some of the shortcomings of the pressed-disk method of sampling, and also to extend FTIR spectroscopy to additional types of samples. Many substances, particularly powders, scatter light in all directions in a diffuse manner. The diffuse reflectance technique isolates the portion of the infrared light that is scattered diffusely from that which is either transmitted or specularly (regularly) reflected and collects it as a diffuse reflectance spectrum. A typical experimental configuration for transmission-mode experiments is shown in Fig. 3. The DRIFT spectroscopy is conducted only on powder samples. Samples may be used neat (with no dilution), or diluted with a non-absorbing matrix such as KBr. Samples are typically ground with a mortar and pestle to avoid large particles with flat surfaces that may distort the spectrum by exhibiting specular reflectance. Samples are not pressed into disks as with transmission experiments, but are simply packed into a sample holder. This makes the collection of DRIFT spectra somewhat simpler than transmission-mode experiments. It also avoids changes in sample structure that may occur when subjected to elevated pressure. However, quantitative DRIFT studies are more difficult than transmission-mode IR, because steps in sample preparation can substantially affect a diffuse reflectance spectrum. Specifically, the particle size, the sample homogeneity, and the degree of packing will influence the results.

Regardless of the care taken in sample preparation, DRIFT data differ from data obtained in a transmission-mode experiment. Peak intensities at higher wave number are typically compressed, and all peak shapes tend to be rather rounded compared with the sharp peaks from pressed disks. A mathematical correction can be applied to the raw spectra to compensate for these anomalies and make DRIFT data directly comparable with transmission-mode references. This conversion is known as the Kubelka-Munk equation, and the FTIR control software typically can perform this conversion automatically.

The primary advantage of DRIFT over transmission-mode studies is that sample preparation does not require the pressing of disks. This means samples can be prepared more quickly and with less effort. However, samples are still dried and diluted in salts, therefore, sample preparation may still affect the chemical species in an environmental sample. The primary drawback of DRIFT is that it trades ease of sample preparation and ability to observe weaker bands by the analysis of neat samples for additional complexity in quantitative analysis.

Attenuated total reflectance experiments

The attenuated total reflectance (ATR) method for FTIR measures the reflectance spectrum of a sample placed upon the surface of a high refractive index optical crystal. The ATR mode of FTIR measurement allows the spectra of solids to be measured without dilution in KBr, and for materials to be measured without grinding or other alterations. It also enables FTIR spectra to be collected in the presence of water. Water is an extremely strong infrared absorber, and traditional sampling methods require drying to remove interference from H₂O. The ATR technique circumvents this limitation by directing the infrared beam through a long crystal. The beam is redirected through the crystal at an angle and bounces multiple times internally and then exits to the detector. The infrared beam produces an evanescent wave on the surface of the crystal, and hydrated samples in close contact with the crystal can be analyzed by comparing the reflectance of the sample with that of the bare crystal. A typical experimental configuration for ATR-mode experiments is shown in Fig. 3.

The ATR technique presents a substantial deviation from a transmission-mode experiment, therefore, experimental complications that affect the adherence of ATR data to the Beer-Lambert law are to be expected. For example, the effective path length of the beam is required for quantitative research and is related to the number of reflections through the crystal and the depth of penetration into the sample. The number of reflections is related to the ATR crystal's physical properties such as angle of incidence of the crystal, and the crystal's physical length and thickness. The depth of penetration into the sample is also related to the physical dimensions of the crystal, but is also a function of both the crystal and the sample's refractive index and the wavelength of light that is passed through the crystal. The ATR crystals may have 30°, 45°, or 60° effective angles of incidence, and may be made of a variety of materials. The most common and widely used material for ATR crystals is ZnSe (zinc selenide), but diamond, amorphous material transmitting IR (AMTIR), germanium, and silicon are also available. Different crystal materials may be preferred due to their refractive index, their useful spectral range, or their chemical sensitivity.

Modern ATR accessories typically feature a 1–3 reflection diamond coated ZnSe crystal for better chemical and physical resistance and ease of maintenance; the decreased surface contact area is compensated for by slightly more efficient reflection of IR light such that a 1-reflection diamond ATR signal is often 30–40% as strong as a traditional horizontal ATR crystal. In addition to ease of use improvements for the diamond ATR designs, fewer reflections mean that water does not fully absorb the infrared light in single reflection experiments and so the spectral range of study is expanded and vibrations of the water of hydration and surface functional groups can often be resolved in aqueous experiments.

The fact that sample penetration depth changes with wavelength presents a serious problem in attempting quantitative experiments with the ATR method. For this reason, ATR data are properly reported as percentage reflectance (%R) rather than percentage transmittance (%T), as the intensity of light transmitted is not constant throughout the spectrum with ATR, but instead is a function of wavelength. Fortunately, a mathematical correction can be made to ATR data that accounts for the wavelength dependence on penetration depth. This correction is typically available in the instrument's data-collection software.

With the ATR technique, it is possible to analyze a wide range of samples. Solids can be pressed into contact with the ATR crystal, while liquid samples and hydrated suspensions can be placed into a sample holder that has a trough with a volume of a few milliliters. This flexibility in sampling has made it possible to conduct adsorption studies of organic acids and oxyanions in situ with FTIR. The ability to monitor the mechanism of adsorption on mineral surfaces is the primary application of ATR-FTIR in environmental science (Suarez et al., 1998).

The primary advantage of the ATR method for environmental science research is that it allows environmental samples to be analyzed with normal amounts of water present and at ambient pressures. The results from ATR are therefore much more representative of chemistry in natural systems. However, ATR experiments are more difficult to compare with traditional transmission studies, difficult to conduct reliably in a quantitative manner, and the experimental procedure is more time-consuming than other methods.

Example ATR experiments

The most common uses for ATR-FTIR in the environmental sciences are in adsorption experiments between an IR-active adsorbate and a large surface area mineral adsorbent. Two experimental approaches are typically utilized in adsorption studies: analysis of reacted suspensions of soil solids, and analysis of adsorption on to solids deposited on to the ATR crystal. Both experimental approaches have merits and drawbacks. It is faster and easier to conduct macroscopic adsorption experiments in the laboratory and then use ATR-FTIR spectroscopy to analyze representative samples than it is to conduct all experiments in situ with deposited solids. This also provides the opportunity to determine the exact surface loading and reaction conditions of the sample prior to analysis. However, using wet pastes or slurries presents some experimental difficulties. Specifically, to obtain a spectrum of the adsorbed species, one must separate the solid from the solution by filtration, resuspend the solid in the supernatant at a constant solid-to-solution ratio for all samples, and then measure a spectrum of the supernatant, the unreacted solid, and the reacted solid. By subtracting both the supernatant and the unreacted solid spectra from the reacted spectrum, the result is the spectrum of the adsorbed species.

In contrast, the deposition method places a thin coating of a mineral on to the ATR crystal that is constant throughout the experiment. A reference spectrum of the ATR crystal, the mineral deposit, background electrolyte, and water is measured. Then, the reactant is added to the solution at a low (micromolar) concentration. The reactant is concentrated at the mineral-crystal interface so that adsorbed species are above detection limits. Since the solid concentration at the surface is constant and the solution-phase reactant is not detected, a high-quality spectrum of the absorbed ion can be obtained. With improved spectral quality, results from deposition studies are typically much easier to fit and interpret. The use of a flow cell and separate controlled environment reaction vessel also enables adsorption (and desorption) experiments to be carried out in real time. This can produce adsorption isotherms by systematically adjusting reactant concentration (see Fig. 4), pH envelopes or edges by adjusting reaction pH, or kinetics (Fig. 5) by collecting measurements repeatedly and rapidly as a function of time.

In addition to equilibrium-based adsorption studies detailed above, improvements in rates of data acquisition and FTIR instrumental capabilities make performing adsorption kinetics experiments at a time resolution of seconds possible (Parikh et al., 2008). Fig. 5 details a representative kinetics ATR-FTIR study of sulfate adsorption on FeOOH (ferric oxyhydroxide) at pH 4.0. This experimental approach makes it possible to follow reaction mechanisms of many soil reactions in situ and extract both rate parameters and complexation information simultaneously. For more details of experimental methodology, see Schmidt et al. (2020).

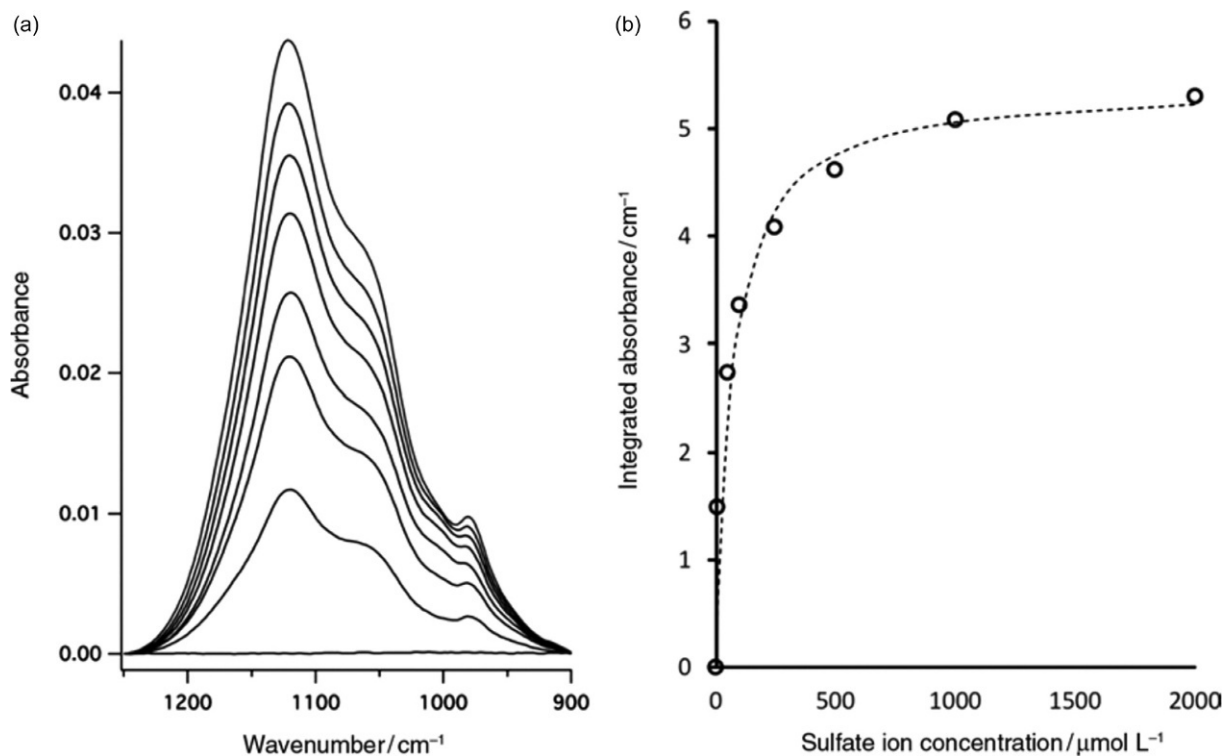


Fig. 4 (a) The ATR spectra recorded on a thin film of Fe_2O_3 (ferric oxide) at pH 4 with increasing sulfate concentration (0–2 mmol L^{-1}) measured after enough time for sulfate to reach steady state adsorption using a flow cell. (b) Integrated absorbance from 900 to 1250 cm^{-1} from (a) spectra plotted against sulfate concentration to produce an isotherm that has been modeled with the Langmuir equation (dotted line). From McQuillan AJ, Osawa M, Peak D, Ren B, Tian ZQ (2019) Experiments on adsorption at hydrous metal oxide surfaces using attenuated total reflection infrared spectroscopy (ATRIRS). *Pure and Applied Chemistry* **91**: 2043–2061.

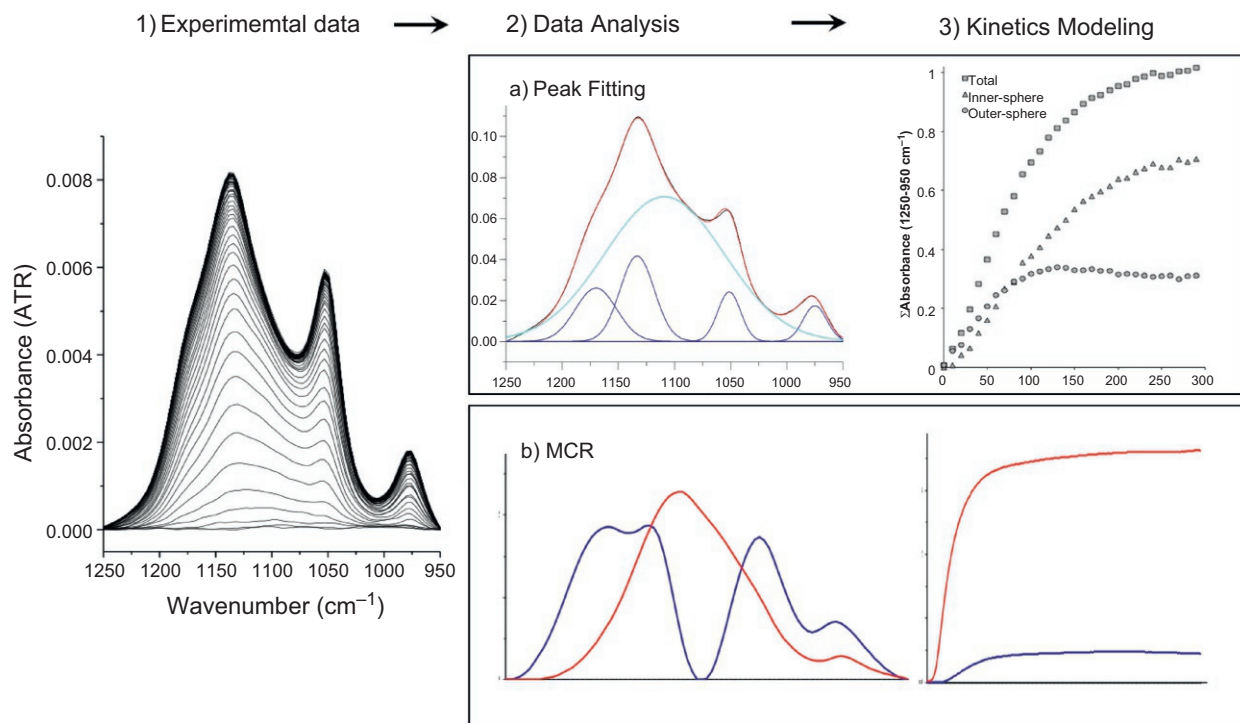


Fig. 5 Sulfate adsorption kinetics on goethite measured at 8s resolution and pH 4.0 for 5 min. (left) shows the background-subtracted ATR-FTIR spectra. (middle) Results from peak fitting (a) used to extract an outer-sphere and inner-sphere monodentate surface complex (as shown in Fig. 2) to every spectrum to perform time-resolved speciation on the adsorption reaction compared to the two species identified by multiple curve resolution (MCR). (right) Kinetics modeling utilizing either the first 5 min of the reaction (top) or 60 min of data (bottom) with MCR.

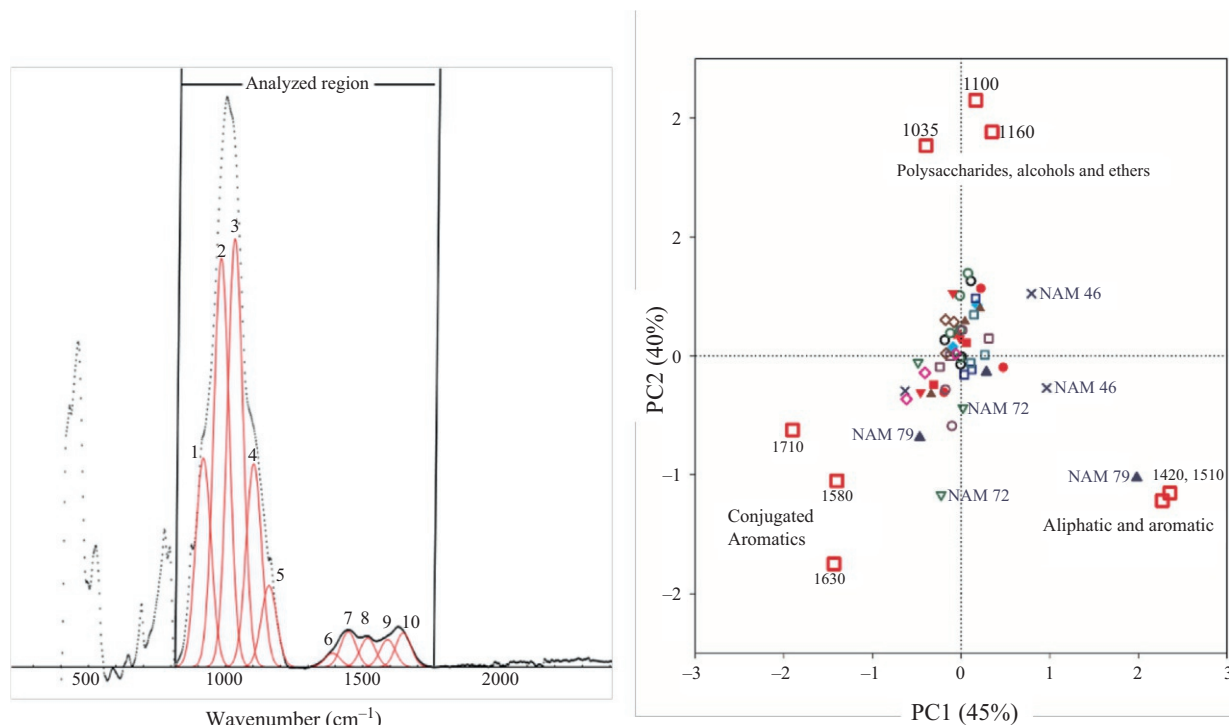


Fig. 6 A representative ATR-FTIR spectrum of whole soil showing the fitted Gaussian peaks representing major C functional groups within the wavenumber range of 1800–900 cm^{-1} . Peaks are identified as follows: 1–917 cm^{-1} ; 2–986 cm^{-1} ; 3–1037 cm^{-1} ; 4–1103 cm^{-1} ; 5–1162 cm^{-1} ; 6–1370 cm^{-1} ; 7–1434 cm^{-1} ; 8–1509 cm^{-1} ; 9–1584 cm^{-1} ; 10–1644 cm^{-1} .

The ATR analysis of soil organic matter (SOM)

The ATR-FTIR spectra of soil samples can be used to characterize the major absorbance bands representing different functional groups in SOM. Due to interference from clay mineral and carbonate vibrations, the wavenumber range of 1800–900 cm^{-1} is often chosen for analysis of C functional groups. Because there is strong overlap among the bands of organic functional groups within this wavenumber range, the individual bands must be resolved by spectral deconvolution into individual bands by Gaussian peak fitting. A typical soil spectrum with functional group peaks fitted is shown in Fig. 6.

Polysaccharides, alcohols, and ether C appears at 1000–1160 cm^{-1} (peaks 3–5); aliphatic and aromatic C bands appear at 1420–30 and 1510 cm^{-1} (peaks 7 and 8); and aromatic and conjugated carboxylic functional groups appear at 1580–1650 cm^{-1} (peaks 9 and 10) (Dhillon et al., 2017). When many soil spectra are analyzed and compared, the results are useful in studying trends and changes in the relative amounts of SOM functional groups, such as in the PCA (principal component analysis) in Fig. 6.

Conclusions

Fourier Transform Infrared spectroscopy (FTIR) has been used for decades as a method to identify and quantify minerals in soil samples. More recently, improvements in sampling methods such as Attenuated Total Reflectance have extended the utility of FTIR to samples with water present. This makes it far more useful for the analysis of adsorption samples either as slurries or via thin film studies in situ. Fourier Transform Infrared spectroscopy is also increasingly used for characterizing soil organic matter functionality. As advances in instrument capabilities reduce signal to noise and increase throughput, and multivariate data analysis approaches such as MCR and PCA are feasible, it is expected that FTIR will increasingly be integrated with ecosystem and field scale studies as well as kinetic experiments.

References

- Dhillon GS, Gillespie A, Peak D, and Van Rees KCJ (2017) Spectroscopic investigation of soil organic matter composition for shelterbelt agroforestry systems. *Geoderma* 298: 1–13.
- Farmer VC (1974) *Infrared Spectra of Minerals*. London, UK: Mineralogical Society.
- Johnston CT, Farmer W, and Aochi Y (1996) Fourier transform infrared and Raman spectroscopy. In: Sparks DL (ed.) *Methods of Soil Analysis, Part 3. Chemical Methods*, 3rd edn, Soil Science Society of America, Book Series No. 53rd edn, 269–321. Madison, WI.

- Nakamoto K (1986) *Infrared and Raman Spectra of Inorganic and Coordination Compounds*. New York: John Wiley.
- Parikh SJ, Lafferty BJ, and Sparks DL (2008) An ATR-FTIR spectroscopic approach for measuring rapid kinetics at the mineral/water interface. *Journal of Colloid and Interface Science* 320(1): 177–185.
- Schmidt MP, Siciliano SD, and Peak D (2020) Spectroscopic quantification of inner- and outer-sphere oxyanion complexation kinetics: Ionic strength and background cation effect on sulfate adsorption to hematite. *ACS Earth and Space Chemistry* 4(10): 1765–1776.
- Suarez DL, Goldberg S, and Su C (1998) Evaluation of oxyanion adsorption mechanisms on oxides using FTIR spectroscopy and electrophoretic mobility, ACS Symposium Series. *American Chemical Society* 715: 136–178.

Electron paramagnetic resonance spectroscopy: Part I Historical Perspectives

Giorgio S Senesi^a and Nicola Senesi^b, ^aConsiglio Nazionale delle Ricerche (CNR), Istituto per la Scienza e Tecnologia dei Plasmi (ISTP), Bari, Italy; ^bDipartimento di Scienza del Suolo, della Pianta e degli Alimenti, Università di Bari “Aldo Moro”, Bari, Italy

© 2023 Elsevier Ltd. All rights reserved.

This is an update of N. Senesi, G.S. Senesi, ELECTRON-SPIN RESONANCE SPECTROSCOPY, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 426–437, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00209-5>.

Introduction	520
Basic principles	521
The resonance phenomenon	521
Spectral parameters	521
Instrumentation and methodology	522
The instrument and the experiment	522
Sample preparation	523
Sensitivity and resolution	523
Determination and interpretation of spectral parameters	523
Whole soil	524
Organic free radicals in soil humic substances	524
The EPR spectra and parameters	524
Interpretation of EPR parameters	525
Factors affecting EPR parameters of humic substances	525
Structural implications	526
Interactions of humic free radicals with organic chemicals and metal ions	526
Paramagnetic transition metal ions in soil	526
The EPR spectra and parameters	527
Ferric iron	527
Divalent copper	528
Vanadyl ion	528
Divalent manganese	529
Molybdenum (V) and (III)	529
Spin derivatization studies	529
Spin labeling and spin trapping	529
Metal spin probes	530
Complexation chemistry of humic substances	530
Ion exchange on layer silicates	530
Chemisorption on mineral surfaces	530
Thermodynamic constants from EPR parameters	531
Conclusions and perspectives	531
References	531

Abstract

Electron paramagnetic resonance (EPR) spectroscopy is a nondestructive, noninvasive, highly sensitive and accurate analytical technique applicable to chemical species that possess one or more unpaired electrons. The first part of this chapter summarizes the basic principles, instrumentation and methodology of EPR spectroscopy. In the second part, the main applications of EPR to the study of indigenous and transient organic free radicals present in soil humic substances and their interactions with organic chemicals, together with paramagnetic transition metal ions in soil constituents, are described. The chapter ends with a brief overview of spin derivatization approaches and conclusion and perspectives.

Key points

- Electron paramagnetic resonance spectroscopy (EPR) provides basic principles of the resonance phenomenon, instrumentation and methodology, sample preparation and determination and interpretation of spectral parameters.
- Application of EPR to whole soil samples.
- Application of EPR to the study of organic free radicals in soil humic substances, interpretation and factors affecting EPR spectra and parameters, structural implications.
- EPR to study the interactions of humic free radicals with organic chemicals and metal ions.

- Application of EPR to the study of paramagnetic metal ions in soil organic and mineral constituents, EPR spectra and parameters, ferric iron, divalent copper, vanadyl ion, divalent manganese.
- Spin derivatization, spin-labeling, spin-trapping, metal spin probes, complexation chemistry of humic substances, ion exchange on layer silicates, chemisorption on mineral surfaces, thermodynamic constants from EPR parameters.

Nomenclature

EPR	Electron Paramagnetic Resonance
ESR	Electron Spin Resonance
ENDOR	Electron Nuclear Double Resonance
ESEEM	Electron Spin Echo Envelope Modulation
NMR	Nuclear Magnetic Resonance
HS	Humic Substances
HA	Humic Acid
FA	Fulvic Acid
DPPH	<i>N,N</i> -diphenylpicrylhydrazyl
Q	Quinone
H ₂ Q	Hydroquinone
HQ [•]	Semiquinone radical
Q ^{•-}	Semiquinone radical anion
5-SASL	Stearic Acid Spin-Label with nitroxide free radical in position 5 of the hydrocarbon chain
DMPO	5,5-dimethylpyrroline 1-oxide
PBN	phenyl- <i>N</i> - <i>t</i> -butylnitrone
RT	Room Temperature
DC	Direct Current
H, H ₀	Strength of the static magnetic field
ΔH	Width of the resonance line, or linewidth
E	Energy for the unpaired electron
ΔE	Absorption of energy by the electron
h	Planck's constant
μ _z	Electron magnetic moment
β	Electron Bohr magneton
M _z	Electron spin angular momentum
S	Electron spin
ν ₀	Frequency of the alternating magnetic field
g	Electronic splitting factor, or g-value, or g-tensor
g _{xx} , g _{yy} , and g _{zz}	Components of the g-tensor along the axes x, y, z
g _{iso} or g ₀	Isotropic g-factor
g	Component of the g-tensor parallel to the symmetry axis
g _⊥	Component of the g-tensor perpendicular to the z-axis
A	Hyperfine coupling constant
I	Nuclear spin
T	Tesla
q _u	Weight of the sample
K _c	Weighted average equilibrium constant

Introduction

Electron paramagnetic resonance (EPR) spectroscopy, otherwise known as electron spin resonance (ESR) spectroscopy, is a nondestructive, noninvasive, highly sensitive and accurate analytical technique that can detect and characterize paramagnetic chemical species possessing unpaired electrons (among several others, [Weil et al., 1994](#); [McBride, 1986](#)). Soil species able to produce EPR spectra include organic free radical moieties in humic substances (HS) and organic- and mineral-associated paramagnetic metal ions ([Senesi and Steelink, 1989](#); [Senesi and Chen, 1989](#); [Senesi, 1990a, 1990b, 1996](#); [Cheshire and Senesi, 1998](#); [Senesi and Loffredo, 2005, 2008](#)). Further, EPR labeling, trapping and metal probing methods are used to study the dynamics of macromolecules in soil solution, HS complexation chemistry, and ion sorption on mineral surfaces (among several others, [Berliner, 1976](#)).

Basic principles

The resonance phenomenon

The basic physical phenomenon underlying EPR spectroscopy is referred to as the Zeeman effect, which consists of the interaction between the magnetic moment of an unpaired electron and an external magnetic field that produces the splitting of the energy levels of the unpaired electron. If a static magnetic field of strength H is applied in the z -direction, i.e., parallel to the z -axis, the Zeeman interaction leads to an energy, E , for an unpaired electron given by:

$$E = -\mu_z H = -g \beta H M_z \quad (1)$$

where μ_z is the electron magnetic moment in the direction of the conventional z -axis; g is a dimensionless constant called the spectroscopic splitting factor or g -value; β is the electron Bohr magneton; and M_z is the component of the electron spin angular momentum in the direction of the applied magnetic field H (z -axis). The values that M_z may assume are $+1/2$ or $-1/2$ depending on the alignment of the electron spin, S , either parallel (high energy) or antiparallel (low energy) to the magnetic field direction. Thus, two energy levels exist with an energy difference that increases linearly with the magnitude of H (Fig. 1).

In a sample containing unpaired electrons in thermodynamic equilibrium in a magnetic field of value H_0 , a population difference exists between the two energy levels with an excess population in the lower level. If an incident electromagnetic radiation is supplied to the sample, e.g., by applying an alternating magnetic field of frequency ν_0 perpendicular to the static magnetic field H_0 , an absorption of energy, ΔE , occurs provided that ν_0 satisfies the following equation:

$$h\nu_0 = \Delta E = g \beta H_0 \quad (2)$$

where h is the Planck's constant. This is known as the 'resonance condition'. The measurement of this energy absorption, recorded as its first derivative, is the EPR signal (Fig. 1), i.e., the basis of EPR spectroscopy.

Spectral parameters

The position and shape of the EPR signal depend on the environmental conditions in the vicinity of the electron, which may result in spectral patterns more complicated (Fig. 1b) than the single-line spectrum (Fig. 1a). The most important effects that influence the position and pattern of the EPR spectrum are the 'electron Zeeman', 'nuclear hyperfine', 'ligand superhyperfine', and 'nuclear quadrupole' interactions. These effects are related to spectral parameters that include the g -factor and nuclear hyperfine and superhyperfine coupling constants.

For a given magnetic field and a particular microwave frequency, the 'electron Zeeman' effect shifts the resonance position of the free electron from a value of $g = 2.00232$ to one that depends on the molecular properties of the paramagnetic species (Eq. 2). The g -value of organic free radicals is generally close to 2.00, and does not distinguish well between different radicals, whereas for paramagnetic metal ions g is often typical of a particular ion and its valence state. For most paramagnetic ions g is anisotropic, i.e., dependent on the orientation of the molecule relative to the external magnetic field, and is completely described by the three components, g_{xx} , g_{yy} , and g_{zz} , along the three axes x , y , z that are generally coincident with molecular symmetry axes. These three components differ from each other ($g_{xx} \neq g_{yy} \neq g_{zz}$) for paramagnetic species that have no principal axis of symmetry, whereas a

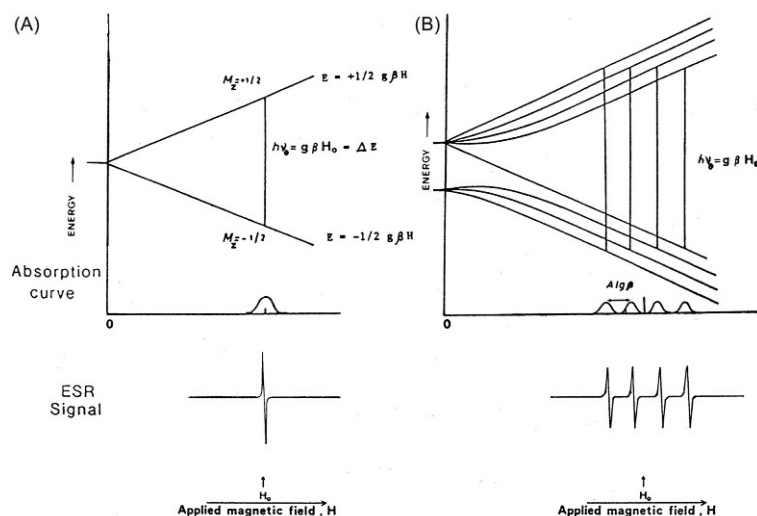


Fig. 1 Effect of an applied magnetic field, H , on the energy levels (E) of the two spin states of an electron ($M_z = \pm 1/2$). EPR transition(s) at $\nu_0 = g \beta H_0/h$, actual absorption curve, and experimental first-derivative EPR spectrum for the cases of no nuclear interaction (a) and interaction with a nucleus having $I = 3/2$ (b), e.g., Cu^{2+} . The nuclear hyperfine splitting is given by $A/g\beta$.

single isotropic g -factor, g_{iso} or $g_0 = 1/3$ ($g_{xx} = g_{yy} = g_{zz}$) is exhibited by octahedral, tetrahedral, or cubic symmetries. Species with 'axial symmetry', such as Cu^{2+} (copper) and V^{4+} (vanadium), have one principal or a threefold axis of symmetry, conventionally the z -axis, and equivalent x - and y -axes. These species feature so-called 'rigid-limit' spectra and have two g -values, usually labeled $g_{\parallel}$ ($= g_{zz}$, the g -value parallel to the symmetry axis, or z -axis), and $g_{\perp}$ ($= g_{xx} = g_{yy}$, the g -value perpendicular to the z -axis in the x,y -plane). The g -values can provide information on the nature of the paramagnetic species and the symmetry of the metal ion in the sample matrix.

The 'hyperfine splitting or structure' arises from the 'nuclear hyperfine' interaction of the unpaired electron with its nucleus if it features a nonzero spin ($I \neq 0$), such as Cu ($I = 3/2$), Mn (manganese, $I = 5/2$) or V ($I = 7/2$), and is independent of the applied field, H_0 . Permanent local fields arising from magnetic nuclei split each electron spin level into $2I + 1$ components, which results in a set of $2I + 1$ equally spaced hyperfine-structure lines replacing the single line resonance in the EPR spectrum. For example, four lines appear for Cu (Fig. 1b), six for Mn, and eight for V. The hyperfine splitting is generally approximated by $A/g\beta$, where A is the magnitude of the nuclear hyperfine interaction, the so-called 'hyperfine coupling constant'. The parameter A , like g , may also exhibit an orientation-dependence (i.e., anisotropy), and may provide useful information on the nature and molecular symmetry of paramagnetic species possessing magnetic nuclei.

The 'superhyperfine splitting or structure' arises when an interaction occurs between the unpaired electron and ligand nuclei having nonzero nuclear spin in paramagnetic metal-ligand complexes. The most common nucleus giving rise to superhyperfine structures in EPR spectra is ^{14}N (nitrogen), which has a nuclear spin, $I = 1$. Thus, each hyperfine line is split into three approximately equally spaced components of equal intensity. Very complex spectra are obtained, however, when more than one N-ligand, especially if not equivalent, is involved in metal complexation.

The 'nuclear quadrupole' interaction arises from the nuclear quadrupole moment (for example, of the Mn nucleus, $I = 5/2$) and the electric field gradient. The interaction, besides modifying the hyperfine splitting, also 'mixes' the hyperfine levels, so that 'forbidden' transitions can occur. In general, quadrupole effects are very small and rarely observed in powder and frozen solution spectra.

Instrumentation and methodology

The instrument and the experiment

The basic components of a typical EPR spectrometer operating in the X-band (microwave) frequency region of the electromagnetic spectrum are (Fig. 2): (a) an electromagnet supplying a static, or direct current (DC), continuous magnetic field that can be varied in strength, usually in the range 0–1 T (1 T (T) = 10^4 gauss); (b) the source of microwaves, usually a vacuum tube oscillator named 'klystron' that provides a monochromatic coherent source of electromagnetic radiation with a frequency typically around 9.5 GHz that is held constant to within 1 part in 10^6 ; (c) a resonant or 'microwave' cavity, which is a hollow-conducting box generally cuboid where the sample is placed and irradiated; (d) a waveguide that couples the klystron to the resonant cavity that is locked electronically to the resonance frequency; (e) a circulator incorporated between the klystron and the cavity, which ensures optimum transfer of the microwave power from the klystron to the cavity and from the cavity to the detector; (f) a small iris diaphragm that interfaces the cavity to the waveguide, and whose effective diameter can be varied until all the incident microwaves are absorbed by the cavity so that it is said to be 'critically coupled'; (g) a magnetic field modulator that increases the intensity of the reflected

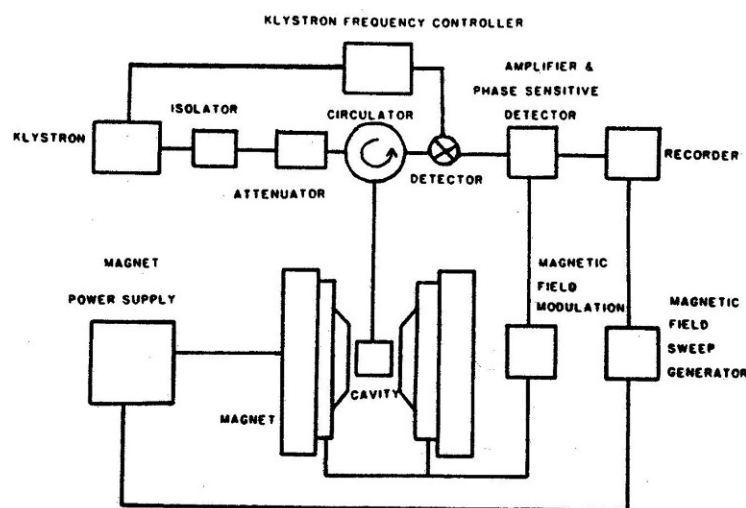


Fig. 2 Block diagram of a typical X-band EPR spectrometer.

microwave at low frequency (usually 100 KHz) by superimposing upon the external magnetic field H_0 a second magnetic field of a few tens of milliT; and (h) a phase-sensitive detector locked to the magnetic field modulation frequency.

In the practical EPR experiment, the microwave frequency is usually fixed and the applied external magnetic field is continuously swept until the resonance condition (Eq. 2) is met. Most EPR spectrometers enable repetitive sweeping of the magnetic field through the resonance, and employ a computer to average transients, thus improving sensitivity. The output of the detector-amplifier system, that is, the EPR spectrum, is displayed as the first derivative of the absorption peak, and recorded on a chart recorder as a function of the applied external magnetic field, H (Fig. 1). Occasionally, the second derivative is plotted to improve the resolution of lines that are close together.

Sample preparation

The EPR technique can analyze samples in any form, that is, liquid, solid powder, amorphous polymer, frozen glass, or single crystal. In general, solid powders are packed into an EPR tube with an inside diameter of a few millimeters. Aqueous solutions or suspensions are placed in capillary tubes or specially designed flat cells to reduce dielectric absorption of microwave radiation by water molecules. Tubes and cells are made of very pure quartz with no defects, which gives no EPR signal and has a very small dielectric loss. Clay films also can be used by mounting them on flat quartz 'tissue cells'.

Removal of paramagnetic O_2 (oxygen) molecules that can broaden EPR spectra by dipole-dipole interaction is sometimes required before EPR analysis. If the concentration of paramagnetic species in a sample is too large, the sample must be diluted magnetically in a diamagnetic matrix to minimize spin-spin coupling that may broaden the spectrum.

Sensitivity and resolution

The sensitivity of the technique depends primarily upon the number of spins present in the sample. Typically, a conventional X-band spectrometer can detect 10^{15} to 10^{16} spins, which implies a sensitivity of 10^{-7} to 10^{-8} mol. The sample must be placed correctly in the cavity that generally has a detection region about 2-cm high, with the greatest sensitivity at its center. Sensitivity is generally improved by setting the modulation amplitude at larger values that, however, should not exceed a fraction of the peak-to-peak resonance linewidth, otherwise distortion of the spectral line shape can occur. According to Curie's law, maximum sensitivity for paramagnetic species is attained at the smallest possible sample temperature. Although organic free radical species can generally be measured easily at room temperature (RT), for most paramagnetic metal ions cooling at liquid N_2 (boiling point = 77 K) or liquid He (boiling point = 4.2 K) temperature is often required. The extremely large dielectric loss of water at microwave frequencies prevents analysis at RT with aqueous systems, and requires EPR measurements to be made at liquid N_2 or liquid He temperature on frozen aqueous solutions. However, large metal-organic molecules cannot tumble rapidly in solution, thus the EPR spectrum at RT is often similar to that observed for powders or frozen solutions.

The major limitation of an EPR experiment is the resolution of the signal linewidths that may overlap to such an extent that information is lost. The resonance linewidth is affected by two mechanisms, namely, homogeneous and inhomogeneous broadening. The most important homogeneous broadening effects are: (a) 'spin-lattice relaxation' by which an electron at a higher energy level returns to the ground state by losing energy to its environment (i.e., the 'lattice'), either in a short time, thus producing linewidth broadening, or over a long time, which results in narrower linewidths; and (b) 'microwave-power saturation' that may arise if a sufficiently large microwave power is applied that tends to equalize the electron populations of the two levels, thus decreasing signal intensity and increasing linewidth. In the absence of saturation, the EPR signal intensity should increase as the square root of the microwave power. Inhomogeneous broadening arises from non-uniformities in the magnetic field throughout the sample resulting from other neighboring paramagnetic species or from neighboring magnetic nuclei, or from dipolar interactions between unlike spins. These effects, often referred to as 'spin-spin interactions', are random in direction and result in the merging of individual resonant lines or spin packets into a single overall line or envelope, with loss of line resolution and related information.

Two techniques are potentially useful to overcome the limitation related to linewidth resolution: (i) electron nuclear double magnetic resonance (ENDOR) spectroscopy, that is a combination of EPR and nuclear magnetic resonance (NMR) techniques; and (ii) electron spin echo envelope modulation (ESEEM) spectroscopy, that is a time-domain electron magnetic resonance technique (Kevan and Kispert, 1976; Kevan and Schwartz, 1979). Both techniques can extend the resolution of hyperfine and superhyperfine contributions to inhomogeneously broadened single-line EPR spectra.

Determination and interpretation of spectral parameters

Once the EPR spectrum has been measured, values of spectral parameters can be determined accurately and rigorously by comparing the experimental spectrum to a computer-simulated spectrum for the species being studied, which is analyzed using trial parameters and expressions for the magnetic fields at which transitions occur. Then, the procedure is repeated until satisfactory agreement is found between the two spectra. In practice, elaborate computer simulation may be avoided by the relatively simple, even though not rigorous, computation of spectral parameters directly from the experimental EPR spectrum and from spectrometer measurement settings using the equations described in detail later.

Once the EPR parameters have been obtained, they can be related to the nature of paramagnetic species, metal oxidation state, and type and site symmetry of metal binding by using rigorous physicochemical approaches and complex mathematical tools. This procedure can be avoided by empirical comparison and correlation of experimentally obtained EPR spectral parameters with EPR data available on similar model, synthetic, or natural systems, which can provide the information of chemical and structural interest to the soil scientist.

Whole soil

The EPR spectra of whole soil samples commonly show a very broad and intense ferromagnetic resonance peak centered at approximately $g = 2.00$, which can be attributed to strongly magnetic materials consisting of randomly oriented crystals and or polycrystalline powdered particles of different shapes and random orientation, such as iron oxihydroxides and minerals rich in Mn, Ti (titanium) and other transition metals (Senesi, 1996). High porosity and crystal defects and or impurities can further increase EPR signal linewidth. As a result of these effects, EPR of a whole soil sample provides only very limited chemical and structural information, thus soil components, i.e., HS and their most important fractions, humic acid (HA) and fulvic acid (FA), metal-HS complexes, clay minerals and metal-clay associations must be isolated from the sample and studied separately.

Organic free radicals in soil humic substances

The EPR spectra and parameters

Humic substances are the most abundant, and most chemically and biologically active fraction of the nonliving organic matter in soil, whose macromolecules typically contain indigenous free radical moieties. The EPR spectrum of HS free radicals is usually obtained at RT on solid or water-dissolved samples of unfractionated HS, and HA and FA fractions and subfractions (Cheshire and Senesi, 1998; Senesi, 1996). The magnetic field is swept over a relatively narrow scan range (generally 5–10 mT) through the field at which the free electron resonates until the resonance condition is met, which generally occurs at a field of about 340 mT and corresponds to a g -value close to that of the free electron. Typical spectrometer settings and operating conditions used in the measurement of EPR spectra of HS free radicals are: microwave frequency close to 9.5 GHz; microwave attenuation, 13 dB, corresponding to a microwave power of about 10 mW, at which the signal is generally least saturated; and modulation amplitude, 0.63 mT (Cheshire and Senesi, 1998; Senesi, 1996). A typical EPR spectrum of HA and FA free radicals features a single-line resonance, devoid of any structure (Fig. 3a), while a partially resolved hyperfine structure is rarely observed (Atherton et al., 1967) (Fig. 3b).

Traditionally three spectral parameters of importance to characterize the nature, origin, and properties of HS free radicals can be derived from EPR spectra; they are the g -value, the width of the absorption line, i.e., the linewidth, and, rarely, the hyperfine structure (Cheshire and Senesi, 1998; Senesi and Steelink, 1989). The g -value can be accurately approximated from the magnitude of the magnetic field at which the resonance occurs for the sample and for a standard of known g -value, usually *N,N*-diphenylpicrylhydrazyl (DPPH) diluted in powdered KCl ($g_{\text{DPPH}} = 2.0036$). A small amount of the standard material in a capillary tube can be taped to the sample tube and its signal used as a field 'marker.' Since ν_0 is identical for both, the g -values of the sample and standard are readily calculated from the ratio of field positions using the relationship derived from Eq. (2) (Cheshire and Senesi, 1998; Senesi, 1996):

$$g_u = g_k H_k / H_u \quad (3)$$

where u (unknown) and k (known) refer to the sample and standard, respectively, and H_k can be calculated from Eq. (2) using ν_0 and g_k . The width of the resonance line, or linewidth, ΔH , is generally measured in Tesla or in gauss (1 gauss = 10^{-4} T), as the

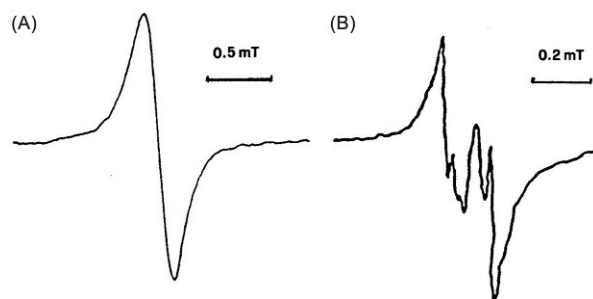


Fig. 3 Typical single-line EPR spectrum of organic free radicals in soil humic acids and fulvic acids (a) (Reproduced with permission from Senesi N (1996) Electron spin (or paramagnetic) resonance spectroscopy. In Sparks DL (ed.) *Methods of Soil Analysis. Part 3. Chemical Methods*, Book Series No. 5, pp. 323–356, Madison, WI: Soil Sci. Soc. of America and American Soc. of Agronomy), and four-line EPR spectrum observed for some acid-boiled humic acids (b) (Reproduced with permission from Atherton NM, Cranwell PA, Floyd AJ and Haworth RD (1967) Humic acid I. ESR spectra of humic acid. *Tetrahedron* **23**: 1653–1667).

peak-to-peak separation of the first derivative EPR signal. The hyperfine splitting is measured as the separation (in gauss or in Tesla) between the hyperfine lines.

The concentration of unpaired electrons, i.e., of organic free radicals in HS can be determined by comparing its signal area with that of a standard char containing a known content of paramagnetic centers, e.g., the 'strong pitch' supplied by the manufacturer, or DPPH (Senesi, 1996; 1998). A double integration of the first derivative curves should be performed to obtain the corresponding areas rigorously. The line shape of the sample and standard are generally of the Lorentzian type, therefore, the double integration can be avoided by simply calculating each area as the product of the height (h) and square of the width (w) of the first derivative signal. If the sample undergoes hyperfine splitting, the area under the hyperfine lines must be summed to obtain the total area to be considered. The spin concentration, expressed in spins g^{-1} , can then be calculated by:

$$(\text{spin } g^{-1})_u = \left[(\text{spin } g^{-1})_k (hw^2)_u (\text{gain})_k \right] / \left[(hw^2)_k (\text{gain})_u q_u \right] \quad (4)$$

where u and k are as above Eq. (3), gain is the signal amplifier gain used and q_u is the weight (in grams, moisture and ash free) of the sample.

This procedure requires that the sample and standard are measured in an identical environment, i.e., the same matrix, geometry, volume, temperature, and spectrometer settings, therefore, either a double cavity instrument or concentric sample tubes should be used. More often, however, separate tubes for the sample and standard are used, placing them alternately in exactly the same position in the resonant cavity.

Interpretation of EPR parameters

The g -values commonly measured for organic free radicals in soil HAs and FAs range between 2.0030 and 2.0050, which are consistent with indigenous semiquinone radical moieties conjugated to aromatic rings (e.g., $g = 2.0041$ for 9,10-anthraquinone), although a contribution from methoxybenzene radicals (g -values range, 2.0035–2.0040) and N- or S-associated radicals ($g = 2.0031$ –2.0037) cannot be excluded (Senesi and Steelink, 1989; Cheshire and Senesi, 1998; Senesi, 1990a,b;).

The linewidths of the EPR signal of HA and FA free radicals generally range between 3.5 and 7.5×10^{-4} T for solid samples and between 2.0 and 3.0×10^{-4} T for the corresponding water solutions (Senesi, 1996). The EPR linewidths do not show any particular dependence on the nature and origin of FAs and HAs, but they are generally slightly greater for FA than for HA from the same source. Factors affecting EPR signal linewidth are free radical concentration and aggregation state of the sample, interactions with solvent and/or metal ions, power-saturation effects, and temperature.

The EPR spectra of HA and FA free radicals rarely show a hyperfine structure. For example, some acid-boiled HAs feature a four-line hyperfine structure (Atherton et al., 1967) (Fig. 3b) attributed to interaction of the unpaired electron with two nonequivalent H (hydrogen) nuclei, whereas an oxidized soil FA exhibits a three-line spectrum ascribed to the interaction of the unpaired electron of a semiquinone O atom with two adjacent equivalent H nuclei (Senesi et al., 1977a).

The concentration of organic free radicals in soil HAs and FAs depends on their nature and origin, and generally ranges between 10^{16} and 10^{18} spins g^{-1} (Senesi and Steelink, 1989). The FAs usually show one-third to one-fifth of the spin content of HAs from the same source. The spin concentration of FAs measured in solution at neutral and alkaline pHs is always greater than that measured in the solid state. Several factors affect the spin content measurement, including power saturation effects, the presence of pronounced shoulders associated with a broad signal, and, in the case of solution samples, the solvent, pH, and time (Senesi and Steelink, 1989; Senesi, 1990a, 1990b; Senesi, 1996). Caution should be used in evaluating and comparing the spin contents of HS. Reasonable reliability and comparability can be expected in the evaluation of changes in spin concentrations that occur in a certain sample subjected to differences in various physical and chemical conditions.

Factors affecting EPR parameters of humic substances

Changes of concentration, g -value and linewidth of free radicals in HS subjected to differences in pH, ionic strength, state of aggregation, hydrolysis, alkylation, redox potential, irradiation, temperature, and humidity, can provide valuable insight into their molecular features (Senesi and Steelink, 1989; Cheshire and Senesi, 1998; Senesi, 1990a,b, 1996).

Increase of pH, chemical reduction, acid hydrolysis, or visible- and ultraviolet (UV)-light irradiation enhances the free radical concentration of HS, while leaving the g -value and signal linewidth almost unaltered (Senesi, 1990a, 1990b, 1996; Cheshire and Senesi, 1998; Senesi and Schnitzer, 1977; Steelink and Tollin, 1967; Riffaldi and Schnitzer, 1972). In all cases, except acid hydrolysis, the increase in spin content is not sustained over time, i.e., it returns gradually, and almost reversibly, to a value similar to that before the treatment. These results suggest that short-lived, 'transient' free radicals chemically similar to stable native radicals are produced in HS with any of the mentioned treatments.

An increase in temperature to 450°C or exposure to high-energy irradiation (gamma rays) causes a marked increase in spin content, broadening of linewidths, and decrease of g -value in solid HAs (Senesi, 1990a, 1990b; Senesi et al., 1977b; Schnitzer and Skinner, 1969). These changes, which are accompanied by the loss of O and increase in C (carbon) content, can be ascribed to homolytic bond-cleavage and subsequent delocalization and stabilization of newly-produced free electrons from semiquinonic O atoms to aromatic C atoms.

Chemical or electrochemical oxidation generally produces a time- and pH-dependent decrease of the free radical content of water-dissolved HAs and FAs, which is reversed by a subsequent reduction treatment (Senesi, 1990a, 1990b). A decrease in spin content is also observed with an increase of neutral electrolyte concentration in FA and HA solutions at a pH close to neutrality (Senesi and Schnitzer, 1977). However, no changes in g-values or linewidths are observed with any of these treatments. A decrease of both organic free radical concentration and EPR signal linewidth is observed with increasing humidity of solid HA samples, which is almost reversible upon re-drying of the samples. Contrasting results have been obtained for the effect of methylation on the free radical content of HS, i.e., it decreases dramatically for the HAs and FAs of podzolic soil, whereas it increases for several other soil HAs and FAs (Senesi and Steelink, 1989; Steelink and Tollin, 1967; Schnitzer and Skinner, 1969).

Structural implications

The free radical concentration is correlated to other compositional and structural parameters of HS, either positively, e.g., to the E_4/E_6 ratio (ratio of absorbance at 465 and 665 nm), absorbance at 465 nm, atomic C/H and O/C ratios, O percentage, phenolic OH content, and aromaticity, or negatively, e.g., to H percentage and aliphaticity (Senesi, 1990a, 1990b; Steelink and Tollin, 1967; Schnitzer and Skinner, 1969; Chen et al., 1977). These results suggest that the free radical content is directly related to the dark color, aromatic or aliphatic character, molecular complexity and particle size of HS. A strong correlation also exists between the free radical concentration and degree of humification of HS in peat soils and in organically amended soils (Cheshire and Senesi, 1998).

Accumulated evidence from EPR (Senesi and Steelink, 1989; Cheshire and Senesi, 1998; Senesi, 1990a,b) strongly supports an indigenous quinone (Q)-hydroquinone (H_2Q) system as that principally responsible for the generation and stabilization of semiquinone radicals ($HQ^\bullet$) and radical anions ($Q^{\bullet-}$) in HS macromolecules, which is based on the simple electron donor-acceptor (or charge-transfer) reversible mechanism:



The large decrease in spin content observed in podzolic HAs and FAs from the selective blocking of phenolic OH groups by methylation confirms these groups as the most important electron donors responsible for the formation and existence of organic free radicals in HS (Schnitzer and Skinner, 1969). A quinone content of 0.5 mmole g^{-1} , typical of most HAs, would theoretically yield up to 6×10^{20} spins g^{-1} of the semiquinone anion at alkaline pH and in the presence of a donor group such as hydroquinone. This effect can increase the amount of stable free radicals in HS from two to five orders of magnitude, with relevant implications in their chemical and biochemical reactivity. Further, semiquinone radicals can be generated by reduction or photoirradiation of quinones in solution in the presence of electron donors (Senesi, 1990a, 1990b).

Interactions of humic free radicals with organic chemicals and metal ions

The EPR technique can be applied usefully for evaluating the role of organic free radicals in the interaction of HS with organic chemicals, such as pesticides (Cheshire and Senesi, 1998; Senesi, 1990a,b). For example, the increase in free radical concentration measured in the interaction products between HAs and s-triazine or substituted urea herbicides provides evidence of the occurrence of an electron donor-acceptor mechanism, with formation of charge-transfer complexes (Senesi and Chen, 1989). In contrast, the products of interaction of chlorophenoxyalkanoic herbicides with HA show a considerable decrease in free radical concentration which suggests the occurrence of homolytic cross-coupling reactions between HA free radicals and free radical intermediates generated in the preliminary chemical, photochemical, and/or biological degradation of the herbicide (Senesi and Chen, 1989).

The EPR spectroscopy can also be used to study the effect of some metal ions on HS free radicals (Cheshire and Senesi, 1998). In general, the addition of diamagnetic metal ions such as Na^+ (sodium), Zn^{2+} (zinc) and Al^{3+} (aluminium) to HA or FA solutions does not affect their free radical content, whereas addition of paramagnetic metal ions such as Fe^{3+} (iron), Cu^{2+} , Mn^{2+} , VO^{2+} (vanadium oxide), Ni^{2+} (nickel) and Co^{2+} (cobalt) causes a marked decrease of HS free radical content as a function of the metal species and HS origin (Cheshire and Senesi, 1998). In studies of podzolization processes, EPR data suggest that recently formed, low-molecular weight HS cannot undergo further polymerization because their free radicals combine with Fe^{3+} and move to the Bh horizon (Cheshire and Senesi, 1998). In another study, EPR spectroscopy revealed the formation of semiquinone radicals by single electron transfers occurring at goethite and manganese oxide surfaces during the oxidation of hydroquinone (Cheshire and Senesi, 1998). Further, EPR spectroscopy can be used to characterize the nature of HS formed by oxidative polymerization of phenolic and other organic compounds in the presence of natural clays, soils, soil oxides, and other metal ion catalysts (Cheshire and Senesi, 1998).

Paramagnetic transition metal ions in soil

The EPR technique can be successfully applied to study natural complexes of paramagnetic metal ions of importance in soil, such as Fe^{3+} , Cu^{2+} , Mn^{2+} , and V^{4+} , with HAs, FAs and plant litter, and structural paramagnetic metal ions, such as Fe^{3+} , in soil minerals (Senesi, 1990a, 1990b, 1992, 1996; Cheshire and Senesi, 1998; Senesi and Loffredo, 2005, 2008). In particular, EPR analysis can provide unique information on the identity and oxidation states of metal ions, metal binding site(s) including ligand types, coordination and symmetry, and stability of metal-organic complexes.

The EPR spectra and parameters

The EPR spectra of paramagnetic metals in soil organic and mineral constituents are usually obtained at either RT on powder samples or at liquid N₂ temperature (77 K) on powders or frozen (77 K) solution samples. The magnetic field is usually scanned tentatively at first over a wide range of strengths (from 0 T to 1 T), then an enlarged spectrum is recorded over a narrower field (0.2 T or 0.1 T) comprising the region where resonances appear to enable their detailed analysis and interpretation. Measurement conditions commonly employed are: microwave frequency, approximately 9.2 GHz (spectra at 77 K) or 9.8 GHz (spectra at RT); microwave attenuation, 13 dB (microwave power oscillates between 9.0 and 9.2 mW); and modulation amplitude, between 0.63 and 4 mT, according to the metal ion being measured (Senesi, 1992). Four examples of typical EPR spectra of metal-organic complexes that occur naturally in isolated soil organic fractions are shown in Figs. 4 and 5.

Three types of spectral parameters can be obtained from EPR spectra of metal-organic complexes: (i) the *g*-value(s) of the metal(s) in the sample; (ii) the hyperfine coupling constant(s), *A*, if the metal nucleus has a nonzero spin; and (iii) the ligand superhyperfine splitting(s), if the unpaired electron of the metal ion is delocalized onto magnetic nuclei of surrounding ligands. Resonance lines due to 'forbidden transitions' are rarely observed, e.g., for Mn²⁺ (Senesi, 1990a, 1990b, 1992, 1996; Cheshire and Senesi, 1998; Senesi and Loffredo, 2005, 2008).

The *g*-values and hyperfine and superhyperfine constants, *A*, can be calculated from experimental spectral data and the values of spectrometer settings used according to the standard equations (Senesi and Loffredo, 2005, 2008):

$$g = h \nu_0 / \beta H_0 = 0.714484 \nu_0 / H_0 \quad (6)$$

$$A \text{ (cm}^{-1}\text{)} = A \text{ (MHz)} / c = 2.80247 (a \text{ g}) / (c g_e) = 0.469766 \cdot 10^{-4} a \text{ g} \quad (7)$$

where ν_0 (MHz) is the microwave frequency at the resonance condition; H_0 is the value of the magnetic field at which the resonance is centered (on calibrated chart paper); g_e (=2.00232) is the *g*-value of the free electron calculated by Eq. (6); *a* is the hyperfine splitting measured as the peak-to-peak separation (in 10⁻⁴ T) between the hyperfine lines in the experimental spectrum; and *c* is the speed of light in a vacuum. Accuracy of the magnetic field calibration of chart paper should be checked using a suitable standard.

Ferric iron

The EPR spectra of HAs, FAs and decomposing litter layers generally exhibit an asymmetrical isotropic resonance line at about *g* = 4.2, which is consistent with high spin (five unpaired d-electrons) Fe³⁺ ions held in tetrahedral or octahedral sites of a low-symmetry (rhombic) ligand field possibly by carboxylic and/or phenolic hydroxyl groups (Fig. 4a) (Senesi, 1992; Senesi and Loffredo, 2005, 2008; Senesi et al., 1986; Senesi et al., 1991). This form of Fe exhibits considerable resistance to proton and metal exchange, and to chemical reduction, which suggests that Fe³⁺ is strongly bound and protected in inner-sphere complexes in HS.

A very broad signal near *g* = 2 is often exhibited by HS (Fig. 4a'), which possibly arises from extended spin-spin coupling of various neighboring Fe³⁺ ions (Senesi, 1992; Senesi and Loffredo, 2005, 2008; Senesi et al., 1986; Senesi et al., 1991). Iron in such

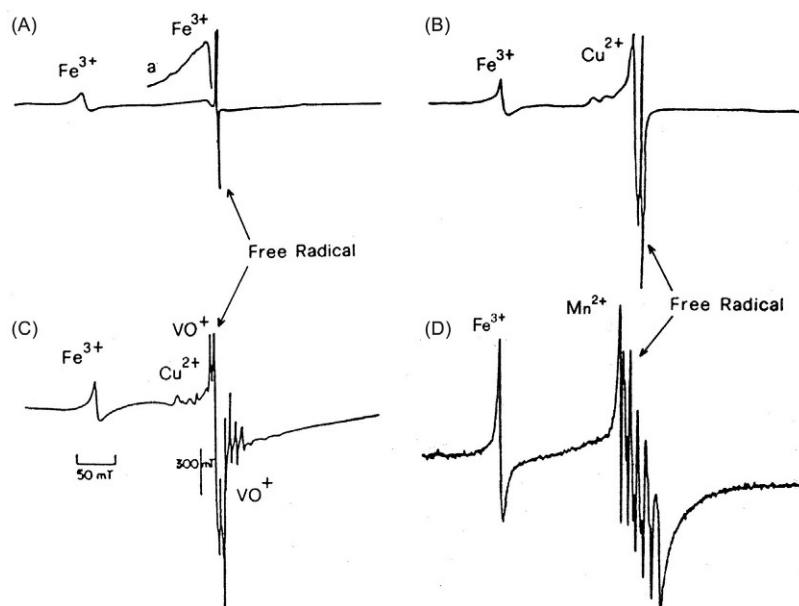


Fig. 4 Representative wide scan range (800 mT) EPR spectra at 77 K of a Mollisol humic acid from the IHSS Reference and Standard collection (a; inset, higher gain), a Paleosol humic acid (b), a loam soil humic acid (c), and an aqueous extract of decomposing leaf litter from a forest soil (d). (Reproduced with permission from Senesi N (1996). Electron spin (or paramagnetic) resonance spectroscopy. In Sparks DL (ed.) *Methods of Soil Analysis. Part 3. Chemical Methods*, Book Series No. 5, pp. 323–356, Madison, WI: Soil Sci. Soc. of America and American Soc. of Agronomy).

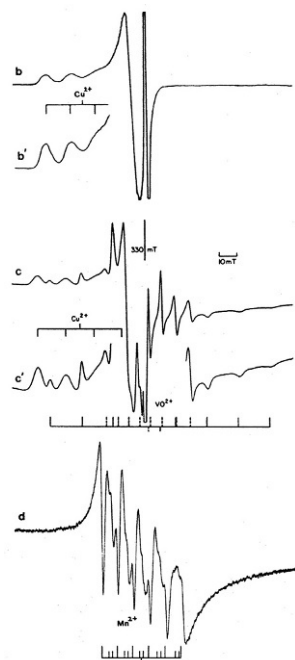


Fig. 5 Same EPR spectra as in Fig. 4b, c, d, but recorded on an enlarged scan range (200 mT).

sites is easily reduced by chemical agents and easily extracted by complexing agents, suggesting that it is weakly bound on the external surfaces of HS. Two weak resonances near $g = 9$ and $g = 6$ are sometimes observed in EPR spectra of HS, which possibly arise, respectively, from Fe^{3+} at sites with near orthorhombic symmetry and from high-spin Fe^{3+} in largely distorted, axially symmetric crystal fields (Senesi et al., 1991).

An EPR signal at $g = 4.3$ can be observed in layer silicates such as kaolinites, vermiculites, micas and smectites, which have anisotropic and isotropic g -factors that possibly arise from structural octahedral $\text{FeO}_4(\text{OH})_2$ groups (McBride, 1986, 1989). Soil clays such as vermiculites and weathered phlogopite often exhibit strong ferrimagnetic resonance at $g = 2$ that can be attributed to ferrimagnetic clusters of structural Fe^{3+} and/or ferric oxides and tends to obscure other spectral details (McBride, 1986, 1989). No direct EPR evidence of Fe^{2+} species is obtained in soil mineral and organic components.

Divalent copper

Soil HS often show an anisotropic rigid-limit spectrum of the 'axial' type in the $g = 2$ region (Figs. 4b, c, and 5b, c), which is ascribed to Cu^{2+} ions (Senesi, 1992; Senesi and Loffredo, 2005, 2008; Senesi et al., 1986; Senesi et al., 1991). The nuclear spin of Cu is $I = 3/2$, thus its EPR spectrum should be split into four (i.e., $2I + 1$) features both at $g_{\parallel}$ and $g_{\perp}$. However, only the component at $g_{\parallel}$ is generally resolved partially into a quadruplet (Figs. 5b, b¹, c, c¹), while splitting of the $g_{\perp}$ component is not resolved. The EPR parameters of this spectrum are consistent with a $d_{x^2-y^2}$ ground state for Cu^{2+} ions held in square planar (distorted octahedral) coordination sites (tetragonal symmetry) as inner-sphere complexes with either only O ligands, such as carboxyls, phenolic hydroxyls, carbonyls, and, often, water molecules, or both O and N ligands (Fig. 6a), or even only N ligands (i.e., a tetraporphyrin site). The resolved pattern observed in some cases at $g_{\perp}$ is ascribed to superhyperfine coupling of the Cu^{2+} unpaired electron to N ligand nuclei ($I = 1$) (Senesi et al., 1991).

The values of spectral parameters also provide evidence of the contribution of a variable covalent bond (i.e., delocalization of the unpaired electron towards the ligands) for Cu^{2+} in HS. At small Cu^{2+} loading, covalent bonding is favored and complexation to amine-N groups is preferred to O ligands, whereas large Cu^{2+} loading in HS determines the formation of small covalency binding of Cu^{2+} largely to O-containing ligands. The accurate analysis of Cu^{2+} EPR patterns at $g_{\parallel}$, possibly with the aid of a computer simulated spectrum, often reveals the presence of two (or more) superimposed quadruplets, which proves the existence of different binding sites for Cu^{2+} in HS (Senesi, 1990a, 1990b, 1992, 1996).

Vanadyl ion

Soil HAs and FAs may feature a richly structured but relatively well-resolved pattern at about $g = 2$ (Figs. 4c and 5c) comprising two distinct overlapping axial rigid-limit spectra (Senesi, 1990a, 1990b, 1992, 1996; Senesi et al., 1991). One spectrum is the typical pattern of Cu^{2+} described previously, and the other comprises two superimposed hyperfine octuplets corresponding to the $g_{\parallel}$ and $g_{\perp}$ components of VO^{2+} ions (nuclear spin of V, $I = 7/2$) rigidly bound in square planar coordination sites as inner-sphere complexes with phenolate and possibly N ligands (large covalency, tightly bound forms) and/or surface carboxylate groups and water molecules (small covalency, relatively labile and exchangeable forms).

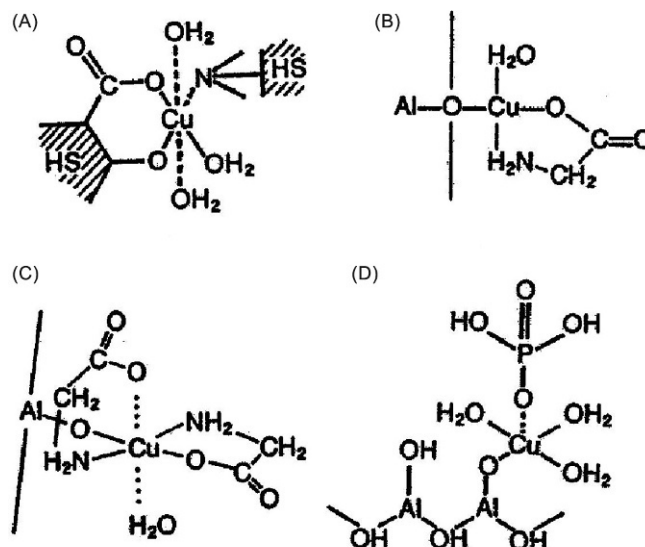


Fig. 6 Models for Cu^{2+} complexes in humic substances (a), on surfaces of gibbsite (b) and boehmite (c) in the presence of glycine, and on aluminium oxides and allophanes in the presence of phosphate (d).

Divalent manganese

The EPR spectra obtained for decomposing leaf litters (Figs. 4d and 5d) and some soil HAs and FAs have a well-resolved isotropic pattern in the region $g = 2$, which comprises six almost equally spaced principal lines and, possibly, 10 secondary lines (corresponding to forbidden transitions) of lesser intensity (Senesi, 1990a, 1990b, 1992, 1996; Senesi et al., 1991). The EPR parameters of such spectra are consistent with high-spin hexahydrated Mn^{2+} ($I = 5/2$) ions in outer-sphere complexes, bound by electrostatic forces to six O atoms of carboxylate and phenolate groups in a distorted octahedral environment.

Molybdenum (V) and (III)

The EPR spectrum of a peat soil HA complex with Mo(V) enriched in ^{95}Mo (nuclear spin, $I = 5/2$) features two distinct components, each split into two six-line hyperfine patterns at $g_{\parallel}$ and $g_{\perp}$, which are consistent with two different axially symmetric Mo(V)-HA complexes (Goodman and Cheshire, 1982; Lakatos et al., 1977). Treatment with 0.1 M HCl of these complexes produced a low-intensity EPR spectrum probably arising from Mo(III) species.

Spin derivatization studies

Spin labeling and spin trapping

The spin-labeling technique, which is the attachment of a simple and stable paramagnetic species such as the nitroxide radical to the compound of interest, is a powerful EPR method for evaluating the dynamics of HS macromolecules in solution, and, in turn, for providing information on the aggregation state, molecular conformation and micellar character of HS (Berliner, 1976; Senesi, 1990a, 1990b). Spin labels can also be used to investigate anisotropic molecular motion on mineral surfaces, which provides information on the nature of adsorption processes not obtained by other methods (Berliner, 1976; Senesi, 1990a, 1990b).

An interesting illustration of this technique is the use of two organic anionic nitroxide spin labels, provided with either a carboxylate or an organophosphate functional group, to study the nature of bonding to non-crystalline aluminium oxide, boehmite and gibbsite in aqueous suspensions (McBride, 1986, 1989). The EPR analysis revealed that both species were adsorbed rapidly onto the large surface area of aluminium oxide and boehmite, whereas only the organophosphate was adsorbed on gibbsite, with a loss in rotational motion, especially of the carboxylate. The values of motional restriction of organic radical anions on surfaces suggest that boehmite adsorbs carboxylate largely by ligand exchange with a single surface OH and organophosphate by bidentate binding, whereas non-crystalline aluminium oxide binds weakly to carboxylate by nonspecific electrostatic forces in addition to ligand exchange of surface OH.

The EPR spectra of the spin-label probe 5-SASL (stearic acid spin-label with nitroxide free radical in position 5 of the hydrocarbon chain) in HA suspensions suggest its bonding with surface hydrophobic sites of HA below pH 5, whereas these sites appear to be unavailable for bonding above pH 5. Kinetics of release of the nitroxide spin probes, neutral Tempol and cationic

Tempamine, from Ca-hectorite aggregates and or pastes measured by EPR suggest that the probes are not strongly sequestered. Nowadays, spin-labeled xenobiotics, e.g., pesticides, represent a promising application of EPR spectroscopy to study their distribution, fate and availability in soils.

The technique of spin trapping involves the reaction of the 'spin trap', such as nitrosobenzene, 2-methyl-nitrosopropene, 5,5-dimethylpyrroline 1-oxide (DMPO) or phenyl-N-t-butyl nitron (PBN), with a short-lived radical, for example, produced in HS by light irradiation, reduction, or raising the pH, to yield a relatively stable radical with an EPR spectrum that enables identification and quantification of the trapped short-lived radical (Thompson, 1990).

Metal spin probes

Complexation chemistry of humic substances

The EPR analysis of synthetic complexes obtained by the reaction of a paramagnetic metal 'probe' such as Cu^{2+} , Mn^{2+} , or VO^{2+} with natural soil HS can provide useful information on the chemistry of the 'residual' complexing capacity of HS towards metal ions and on the stability of the complexes formed (Senesi and Loffredo, 2005, 2008; Senesi et al., 1991). For example, when an FA is doped with 5.5–50.1% Fe^{3+} , the $g = 2$ signal is enhanced relative to the $g = 4$ signal, which suggests that Fe^{3+} is preferentially bound in easily exchangeable forms to the surface sites of FA (Senesi, 1990b). Evidence of inner-sphere complexes of Mn^{2+} coordinated octahedrally, possibly with carboxyl, phenolic hydroxyl and or carbonyl groups, is obtained for some soil HAs doped with Mn^{2+} at high pH (>8) or temperature ($>50^\circ\text{C}$) (McBride, 1986, 1989). Rotational correlation times obtained by EPR analysis combined with gel filtration chromatography data enable measurement of the dynamics of motion and stoichiometry of FA- VO^{2+} complexes in solution, and to obtain information on their molecular conformation and aggregation properties (Templeton and Chasteen, 1980). Quantitative EPR spectroscopy can also be used as a more sensitive, convenient, and faster technique, to determine weighted average equilibrium constants (K_c) for water soluble Mn^{2+} -FA complexes, which accord well with K_c values determined by ion-exchange methods (Gamble et al., 1977).

Ion exchange on layer silicates

The EPR spectroscopy confirms that metal ions such as Cu^{2+} , Mn^{2+} , VO^{2+} , and Cr^{3+} (chromium) retain their inner hydration sphere and considerable rotational mobility on exchange sites of layer silicate clays (McBride, 1989). For example, EPR spectra of exchangeable Cu^{2+} and Mn^{2+} on kaolinite indicate the presence of planar $\text{Cu}(\text{H}_2\text{O})_4^{2+}$ ions oriented parallel to the surface and with strong mobility (McBride, 1989; McBride, 1986). In contrast, an orientation-dependent rigid-limit EPR spectrum of Cu^{2+} is obtained for hydrated Cu^{2+} ions adsorbed on hectorite when the interlamellar spacing on the clay is limited to the equivalent of one or two molecular layers of water, which suggests considerable motional restriction for Cu^{2+} ions (McBride, 1986).

Chemisorption on mineral surfaces

The EPR analysis proves that paramagnetic metal ions at trace levels are bound rigidly in chemisorbed, nonexchangeable forms at isolated surface sites of soil oxides, hydroxides and aluminosilicates, whereas they are sorbed in exchangeable forms to layer silicate clays (Senesi, 1996). For example, Cu^{2+} has been proven to chemisorb at isolated sites on non-crystalline $\text{Al}(\text{OH})_3$ and microcrystalline AlOOH , possibly by the formation of one or two direct bonds between surface $\text{Al}-\text{O}$ groups and Cu^{2+} , which is favored at high pH (McBride, 1986, 1989). At low pH gibbsite can chemisorb small amounts (<0.5 mmole 100 g^{-1}) of monomeric Cu^{2+} that is oriented with its z -axis perpendicular to the (001) planes of the mineral. At pH >5 , a broad featureless resonance and reduction of the rigid-limit spectrum of Cu^{2+} appear suggesting the presence of different, possibly hydrolyzed and polymerized forms of Cu^{2+} on the surface. The large decrease in g -values and increase in $|A_{\parallel}|$ -values of the low-pH, gibbsite-chemisorbed Cu^{2+} with exposure to NH_3 vapors suggest that Cu^{2+} may form a complex with NH_3 while remaining rigidly bound to the oxide surface. The EPR data indicate that Cu^{2+} ions in the presence of the chelating ligand glycine absorb on microcrystalline gibbsite and boehmite in the form of ternary complexes in which Cu^{2+} coordinates simultaneously with one surface hydroxyl and one (on gibbsite) or two (on boehmite) glycine molecules, with orientation of the Cu^{2+} z -axis normal to the (001) sheets of the mineral (Fig. 6b and c) (McBride, 1986, 1989).

The EPR studies have shown that Cu^{2+} can be adsorbed by a direct bond (chemisorption) in nonexchangeable forms on allophanes and imogolite by both a preferential binuclear mechanism involving two adjacent AlOH groups and a weaker type of binding site probably involving isolated AlOH or SiOH groups (McBride, 1986, 1989). Furthermore, NH_3 can readily displace H_2O (water) and OH -ligands from chemisorbed Cu^{2+} , leading to the formation of ternary Cu^{2+} -ammonia-surface complexes. At low pH, Cu^{2+} has also been shown to chemisorb strongly to bidentate sites of titanium dioxide, whereas at higher pH a weaker complex is formed involving single $\text{Ti}-\text{OH}$ groups (McBride, 1986, 1989).

The EPR technique shows that Cu^{2+} is adsorbed on aluminium oxides and allophanes in the presence of phosphate forming a ternary complex with the phosphate coordinated to the axial position of a surface-bound Cu^{2+} (Fig. 6d) (McBride, 1986, 1989). Large amounts of phosphate suppress Cu^{2+} adsorption, apparently by blocking the coordination of Cu^{2+} to surface AlOH groups. Evidence from EPR has also shown that VO^{2+} can be co-adsorbed with phosphate on boehmite and aluminosilicates as an inner-sphere complex (McBride, 1986, 1989). Finally, EPR has proved that Cr^{3+} in the presence of selenite, phosphate, and fluoride is co-adsorbed on hectorite and montmorillonite with formation of ternary surface complexes (Charlet and Karthein, 1990).

Thermodynamic constants from EPR parameters

Metal spin probes can be used to estimate thermodynamic stability constants of metal-surface complexes from the EPR parameter g_{II} (Senesi, 1996). For example, there is a linear relationship between the g_{II} -value and corresponding formation constants of square planar Cu^{2+} complexes with hydrous- Al_2O_3 , TiO_2 and some silicas in the presence of bidentate ligands (Motschi, 1984). A decrease in the g_{II} value by 0.1 units corresponds to an increase in thermodynamic stability of about eight orders of magnitude. Based on such results, a major revision of current concepts of cation adsorption on layer silicates is possible.

Conclusions and perspectives

The EPR technique can provide unique information on the features of permanent organic free radicals indigenous to soil HS and in elucidating their interaction and bonding mechanisms with metal ions and organic chemicals. Furthermore, EPR enables measurement of the formation or disappearance of transient radical species in HS promoted by various environmental factors, which may help to elucidate several naturally occurring processes affected by HS, such as photosensitized reactions involving organic chemicals. Finally, EPR can provide unique insight into the oxidation states of metal ions and their mechanisms and sites of binding to HS, the identity of HS ligands involved, and the stability of metal complexes under different conditions.

The EPR shows promise for future more extensive and in depth research on the role of organic free radicals as potential initiators/activators/catalysts/mediators/antagonizers in one-electron redox processes occurring in soil and the direct physiological effects exerted by HS on seed germination, root initiation and plant growth in general. Furthermore, major advancement is expected by the application of spin derivatization techniques such as spin-labeling and spin-trapping, whereas ENDOR and ESEEM are expected to overcome some intrinsic difficulties of EPR such as improving its resolution in studies of randomly oriented soil components.

References

- Atherton NM, Cranwell PA, Floyd AJ, and Haworth RD (1967) Humic acid I. ESR spectra of humic acids. *Tetrahedron* 23: 1653–1667.
- Berliner LJ (1976) *Spin Labelling: Theory and Applications*. New York: Academic Press.
- Charlet L and Karthein R (1990) Study of inorganic-ligand chromium (III)-surface ternary complexes by ESR spectroscopy. *Aquatic Science* 52: 517–527.
- Chen Y, Senesi N, and Schnitzer M (1977) Information provided on Humic Substances by E_d/E_g ratios. *Soil Science Society of America Journal* 41: 352–358.
- Cheshire MV and Senesi N (1998) Electron spin resonance spectroscopy of organic and mineral soil particles. In: Huang PM, Senesi N, and Buffle J (eds.) *Structure and Surface Reactions of Soil Particles. IUPAC Series on Analytical and Physical Chemistry of Environmental Systems*. pp. 325–373. Chichester, England: John Wiley and Sons Ltd.
- Gamble DS, Schnitzer M, and Skinner DS (1977) Mn (II)-fulvic acid complexing equilibrium measurements by electron spin resonance spectrometry. *Canadian Journal of Soil Science* 57: 47–53.
- Goodman BA and Cheshire MV (1982) Reduction of molybdate by soil organic matter: EPR evidence for formation of both Mo (V) and Mo (III). *Nature (London)* 299: 618–620.
- Kevan L and Kispert L (1976) *Electron Spin Double Resonance Spectroscopy*. New York: Wiley-Interscience.
- Kevan L and Schwartz RN (1979) *Time Domain Electron Spin Resonance*. New York: Wiley-Interscience.
- Lakatos B, Tibai T, and Meisel J (1977) ESR spectra of humic acids and their metal complexes. *Geoderma* 19: 319–338.
- McBride MB (1986) Magnetic methods. In: Klute A (ed.) *Methods of Soil Analysis*. 2nd edn, *Book Series No. 5, Part 12nd edn*, pp. 219–270. Madison, WI: Soil Science Society of America and American Society of Agronomy.
- McBride MB (1989) Reactions controlling heavy metal solubility in soils. In: Stewart BA (ed.) *Advances in Soil Science*. vol. 10, pp. 1–56. New York: Springer-Verlag.
- Motschi H (1984) Correlation of EPR parameters with thermodynamic stability constants for copper (II) complexes. Cu (II) EPR as a probe for surface complexation at the water/oxide interface. *Colloids and Surfaces* 9: 333–347.
- Riffaldi R and Schnitzer M (1972) Effects of diverse experimental conditions on ESR spectra of humic substances. *Geoderma* 8: 1–10.
- Schnitzer M and Skinner SIM (1969) Free radicals in soil humic compounds. *Soil Science* 108: 383–390.
- Senesi N (1990a) Application of electron spin resonance (ESR) spectroscopy in soil chemistry. In: Stewart BA (ed.) *Advances in Soil Science*. vol. 14, pp. 77–130. New York: Springer-Verlag, Inc.
- Senesi N (1990b) Molecular and quantitative aspects of the chemistry of fulvic acid and its interactions with metal ions and organic chemicals. Part I. The electron spin resonance approach. *Analytica Chimica Acta* 232: 51–75.
- Senesi N (1992) Metal-Humic substances complexes in the environment. Molecular and mechanistic aspects by multiple spectroscopic approach. In: Adriano DC (ed.) *Biogeochemistry of Trace Metals*, pp. 429–496. Boca Raton: CRC Press.
- Senesi N (1996) Electron spin (or paramagnetic) resonance spectroscopy. In: Sparks DL (ed.) *Methods of Soil Analysis. Part 3. Chemical Methods*, vol. 5, pp. 323–356. Madison, WI: Soil Science Society of America and American Society of Agronomy.
- Senesi N and Chen Y (1989) Interactions of toxic organic chemicals with humic substances. In: Gerstl Z, Chen Y, Mingelgrin V, and Yaron B (eds.) *Toxic Organic Chemicals in Porous Media*, pp. 37–90. Berlin: Springer-Verlag.
- Senesi N and Loffredo E (2005) Metal ion complexation by soil humic substances. In: Tabatabai MA and Sparks DL (eds.) *Chemical Processes in Soil*, pp. 563–617. Madison, WI: Soil Science Society of America and American Society of Agronomy. *Book Series No. 8*.
- Senesi N and Loffredo E (2008) Spectroscopic techniques for studying metal-Humic complexes in soil. In: Violante A, Huang PM, and Gadd GM (eds.) *Biophysico-Chemical Processes of Heavy Metals and Metalloids in Soil Environments*, pp. 125–168. Hoboken, NJ: John Wiley and Sons Inc.
- Senesi N and Schnitzer M (1977) Effects of pH, reaction time, chemical reduction and irradiation on ESR spectra of fulvic acid. *Soil Science* 123: 224–234.
- Senesi N and Steelink C (1989) Application of ESR spectroscopy to the study of Humic Substances. In: Hayes MHB, MacCarthy P, Malcolm RL, and Swift RS (eds.) *Humic Substances II. Search of Structure*, vol. 2, pp. 373–407. Chichester, England: John Wiley & Sons Ltd.
- Senesi N, Chen Y, and Schnitzer M (1977a) Hyperfine splitting in electron spin resonance spectra of fulvic acid. *Soil Biology and Biochemistry* 9: 371–372.
- Senesi N, Chen Y, and Schnitzer M (1977b) The Electron and gamma Irradiation of Humic Substances. *Fuel* 56: 171–176.
- Senesi N, Sposito G, and Martin JP (1986) Copper (II) and Iron (III) complexation by soil humic acids: An IR and ESR study. *The Science of the Total Environment* 55: 351–362.
- Senesi N, Sposito G, Bradford GR, and Holtzclaw KM (1991) Residual metal reactivity of humic acids extracted from soil amended with sewage sludge. *Water, Air, and Soil Pollution* 55: 409–425.
- Steelink C and Tollin G (1967) Free radicals in soil. In: McLaren AD and Peterson GM (eds.) *Soil Biochemistry*, pp. 147–169. New York: Dekker.

- Templeton GD and Chasteen ND (1980) Vanadium-fulvic acid chemistry: Conformational and binding studies by electron spin probe techniques. *Geochimica et Cosmochimica Acta* 44: 741–752.
- Thompson AJ (1990) Electron paramagnetic resonance and electron nuclear double resonance spectroscopy. In: Andrews DL (ed.) *Perspectives in Modern Chemical Spectroscopy*, pp. 295–320. Berlin: Springer-Verlag.
- Weil JA, Bolton JR, and Wertz JE (1994) *Electron Paramagnetic Resonance: Elementary Theory and Practical Applications*. New York: Wiley-Interscience.

Electron paramagnetic resonance spectroscopy: Part II the view forward

Robert L Cook^{a,d}, Emmanuel Ampiah^{a,d}, Phillip T Sprunger^{b,d}, and Randall W Hall^{c,d}, ^aDepartment of Chemistry, Louisiana State University, Baton Rouge, LA, United States; ^bDepartment of Physics and Astronomy, Louisiana State University, Baton Rouge, LA, United States; ^cDepartment of Natural Sciences and Mathematics, Dominican University of California, San Rafael, CA, United States; ^dLouisiana State University Superfund Research Center, Baton Rouge, LA, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of N. Senesi, G.S. Senesi, Electron-Spin Resonance Spectroscopy, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 426–437, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00209-5>.

Introduction	533
Basic principles	533
Fundamentals	534
Instrumentation	535
Applications	536
Soils and soil organic matter	536
Metal binding	538
Spin labels	539
Environmentally persistent free radicals (EPFRs)	541
Summary	542
Acknowledgment	542
References	542

Abstract

This chapter presents the application of electron paramagnetic resonance (EPR) to soils, soil systems, and natural organic matter. The basic principles of EPR are introduced with a focus on the examples discussed within the chapter. The power and promise of EPR is illustrated using present-day applications (i) of high frequency EPR in environmental soil sciences, such as natural organic matter characterization and its binding of metals, (ii) of spin labels to probe pollutant sorption, and (iii) in the detection and characterization of a new class of compounds of interest known as environmentally persistent free radicals.

Key points

- EPR is a highly sensitive non-destructive technique
- EPR has seen some major recent advances, especially in terms of high field EPR and associated computational methods, that are allowing for detailed molecular level EPR studies on natural organic matter
- EPR can be used to monitor the association of metals and organic molecules with natural organic matter and whole soils
- EPR is an essential tool for studying environmentally persistent free radical

Introduction

Electron pair resonance (EPR, also known, especially in the past, as electron spin resonance [ESR]) can be viewed as the electron complement to nuclear magnetic resonance. The EPR has a long history of application in soil science, as illustrated by Lagerkrantz and Yhland (1963) and Atherton et al. (1967), primarily in the early studies on humic acids. Electron pair resonance spectroscopy has been used to study native and anthropogenic organic and inorganic components within soils, with studies ranging from humic acid and mineral characterization, to metal binding, to the production and fate of environmentally persistent free radicals. For a brief introduction and review of EPR applied to soils and soil components, please refer to Part I of this topic, *Historical perspective*, in the Encyclopedia of Soil Science. Advances in EPR and the resulting new applications of EPR to soils and soil-related components will be the focus of this chapter, with illustrative examples from the literature, preceded by a brief presentation of the basic principles of this technique.

Basic principles

This section presents the fundamentals of modern EPR with a focus on those that are particularly important to its application in the field of soil science. This is not meant to be a comprehensive introduction to the underlying principles or theory of EPR, but rather a short introduction based on a number of sources (Corvaja, 2009; Eaton et al., 2010; Weil and Bolton, 2007: the reader is referred to these references for further exploration).

Fundamentals

Spin: An intrinsic characteristic of electrons, and all elementary particles, is a quantum mechanical angular momentum known as spin, S . The term 'spin' is more customary than factual and refers to a behavior of an electron as if it were spinning (rotating about an axis), generating two distinct behaviors in the external magnetic field. An electron is a quantum particle and its behavior is controlled by quantum mechanics. For our purposes we will simply view electron spin as being in one of two states, usually indicated by α ($1/2$, spin-up, clockwise rotation) and β ($-1/2$, spin-down, counterclockwise rotation). Because of quantum mechanical considerations, we also need only to consider, S_z , the z -axis Cartesian component of S . Electrons are charged particles and when an electron spins, a magnetic moment will be created. The magnitude, S , of a magnetic moment expressed in $\hbar$ units, is given by:

$$|S| = \sqrt{S(S+1)}. \quad (1)$$

If the electron spin quantum number, S , is equal to $\pm 1/2$; then $|S| = \sqrt{3/4}$. For simplicity, we will only consider electrons, and hence only $+1/2$ (α) and $-1/2$ (β) states, for the remainder of this chapter.

Zeeman effect: As an electron has an associated spin, it will have an associated magnetic moment μ_e , proportional to S , as follows:

$$\mu_e = g \mu_B S, \quad (2)$$

where μ_B is the Bohr magneton and g is the Landé, or more commonly called the g -factor. If an external magnetic field, B , is applied, then there is an interaction between the external magnetic field and the magnetic moment. The energy (E) associated with this interaction can be calculated (to make this calculation straight forward and informative for how EPR experiments are executed, we will assume a magnetic field along the z -axis with a magnitude of B_0) as follows:

$$E = g|\mu_B|B_0S_z. \quad (3)$$

For half spin systems this means that we will have electrons with two different energies, as follows:

$$E_{\pm} = \pm(1/2)g|\mu_B|B_0. \quad (4)$$

With the energy difference between the two spin states given by:

$$h\nu = E_{1/2} - E_{-1/2} = \Delta E = g|\mu_B|B_0. \quad (5)$$

Eq. (5) shows that the energy difference between the two states scales linearly with the magnitude of the field applied.

A collection of electrons: For real samples, one is dealing with a large collection of spins, in the order of trillions or larger, which means that in a magnetic field there will be a population in a lower and higher energy state Eq. (4). There will a population difference between the lower and higher energy states, given by the Boltzmann distribution law:

$$N_{\alpha}/N_{\beta} = \exp\left(-\frac{g|\mu_B|B_0}{k_B T}\right), \quad (6)$$

where T is the absolute temperature of the lattice and k_B is the Boltzmann constant. Under typical conditions, the population excess between the two states would be 1/1000. Although this may seem exceedingly small, it is more than enough for a detectable EPR absorption signal as a microwave induced transition from a low to high energy state; the signal being proportional to the number of spins in the initial state.

Why is g not constant: Above, we have viewed electrons as free electrons; however, for those in a molecule the situation is more complex. To help to provide insight into this complexity, first consider an atom with a fully occupied shell and one additional unpaired electron. Here, all the electron spins are coupled except that for the unpaired electron. In this case, there are two contributors to the electron angular momentum: (i) the spin of the electron and (ii) the electron's orbital motion around the nucleus. A layer of complexity is added when we lose the spherical symmetry of an atom and replace it with the low (far from ideal) symmetry of a molecule. In the case of a molecule, the orbital angular momentum tends to zero, leaving only the electron spin component of the electron's angular momentum. In other words, we are back to where we started with a free electron, even though the electron is not free. The fact that the electron in question is not free means that when it is excited, it obtains an orbital angular momentum contribution. Therefore, when ground state and excited electrons mix, a small amount of orbital angular momentum is transferred to the ground state electron; this is known as the spin-orbit contribution.

At this point it is useful to consider spin-orbital coupling from a chemical point of view, which may be accomplished by treating the nucleus (or nuclei) simply as an orbiting positive charge that produces a second magnetic field, dB , at the electron. This means that rather than B_0 the field experienced is B_{eff} . With this in mind, we return to and expand on some of the previous mathematics, so that:

$$h\nu = \Delta E = g|\mu_B|B_{eff} = g_e|\mu_B|(B_0 + dB) = (g_e + dg)|\mu_B|B_{eff}. \quad (7)$$

Therefore, the g -value ($g_e + dg$) contains the all-important chemical information of the molecule's electronic structure.

The magnitude of the spin-orbital contribution depends on the size of the nucleus that possesses the unpaired electron. This means that metal complexes have a substantial spin-orbital contribution with their large nuclei. For organic radicals, on the other

hand, their atoms are light, i.e., small, therefore the spin-orbital contribution is small. Nevertheless, even small dg values provide important information, such as the identity of an atom that the unpaired electron is associated with (centered on), as presented below:

Carbon centered: $g \sim 2.002$ to 2.003 .

Carbon centered next to an oxygen: $g \sim 2003$ to 2004 .

O/N (oxygen/nitrogen) centered: $g \sim 2.004$ to 2.006 .

S (sulfur) centered: $g \sim 2.007$ to 2.010 .

External magnetic field: Although not discussed directly, there are two ways in which one can collect EPR data, namely by applying a fixed (constant) magnetic field (B_0) and varying the applied microwave radiation or by applying microwave radiation of fixed frequency (ν_0) and varying the applied magnetic field. Due to a range of technical issues, it has been easier to vary the frequency of the applied microwave radiation compared to doing that with the magnetic field. Thus, the standard is to have a fixed external magnetic field when carrying out EPR measurements. The applied magnetic field does not always have to be the same and there are various technical reasons as to why one may wish to have different fixed magnetic fields depending on the experiment. These different fields (in units of Tesla, T) are broken down into a series of ranges known as bands (wavebands) based on the external magnetic field and the associated frequency (or wavelength) of the microwave source, with the common bands summarized in Table 1.

Although traditionally EPR has been carried out at lower frequencies (low frequency EPR, LFEPR), for example within the X-band, higher frequency EPR, akin to NMR, has advantages in terms of resolution and sensitivity as well as increased orientation selectivity for distorted systems. However, until recently, with hardware limitations and the much higher frequencies utilized in EPR than NMR, there have been several technical barriers to high frequency EPR and its application beyond highly specialized EPR research centers. With the recent introduction of commercial high frequency EPR (HFEPR) systems, HFEPR should become more widespread in a range of scientific fields, including soil science.

Local magnetic fields: Within a real system, such as a molecule, nuclei that possess spin angular momentum are present and hence an associated magnetic moment may be at play and be affected by the magnetic field in an analogous manner to electrons, resulting in the nuclear Zeeman effect.¹ A system with spin angular momentum will produce its own magnetic field, meaning that nuclei with a spin angular momentum will cause electron spins to experience an additional magnetic field. In other words, there will be an electron–nucleus spin interaction, known as a hyperfine interaction, leading to hyperfine splitting within the EPR signal into two (for protium, normal hydrogen, ^1H , carbon-13, ^{13}C , phosphorus-31, ^{31}P , and nitrogen-14 ^{14}N) or three (for deuterium, heavy hydrogen, ^2H , and nitrogen-14, ^{14}N) components. Importantly, carbon-12, ^{12}C , and oxygen-16, ^{16}O , do not possess spin angular momentum.

If one takes a step back and reflects on the above paragraph, the essence is that, within a molecule with nuclei possessing spin angular momentum both electron spin components, S (EPR), and nuclear spin components, I (NMR components will be present, and can be extracted. This is the principle behind such experiments as EPR-electron nuclear double resonance, ENDOR; such experiments can be viewed, in a very simple way, as EPR detected NMR with the sensitivity of EPR.

Instrumentation

An electron paramagnetic spectroscopy (EPR) instrument consists of a number of components, as illustrated in Fig. 1. The most important components are the microwave bridge, consisting of the microwave source as well as a detector, the sample cavity in which the sample is inserted, the electromagnet to induce the Zeeman effect, and associated electronics. The microwave bridge is the heart of an EPR spectrometer as this is where the microwaves are generated and detected. Most of the time, the difference between

Table 1 The EPR bands in terms of approximate required magnetic field and frequency.

Band	L	S	C	X	P	K	Q	U	V	E	W	F	D	J
B_0/T	0.03	0.11	0.14	0.33	0.54	0.86	1.25	1.8	2.3	2.7	3.5	3.9	4.9	10.2
ν/GHz	1	3	4	10	15	24	35	50	65	75	95	111	140	285

¹For a nucleus to have a spin angular momentum, it must have spin. A nucleus of an atom can be viewed as being made of protons and neutrons, collectively called nucleons, whose spins are then paired up in a similar manner to how electrons are paired in the different electron spin orbitals. The spin numbers of the protons and neutrons are then summed up, resulting in the nuclear spin. As is the case with electrons, up ($\uparrow$) and down ($\downarrow$) spins cancel each other, so the $I = 0$. Nuclear spin is more complex than electron spin, and follows these general rules: i) if the nucleus contains an even number of both types of nucleons, then I (nuclear spin) = 0 (e.g., the ^{12}C isotope (6 protons and 6 neutrons) and ^{16}O isotope, ii) if the nucleus consists of an odd number of nucleons, then I is a half integer, e.g., $\frac{1}{2}$ in ^1H , ^{13}C , ^{15}N , and ^{31}P , and (iii) if the nucleus contains an odd number of neutrons and an odd number of protons, then I is an integer e.g., 1 in ^2H or ^{14}N . It is the smaller particles called quarks that comprise neutrons (and protons) that give neutrons their spin. Quarks have charge in the form of $-\frac{e}{3}$ or $+\frac{2e}{3}$. A neutron consists of two $-\frac{e}{3}$ quarks and one $+\frac{2e}{3}$ quark, no combination of quark spins may result in a zero spin, explaining why neutrons have spin.

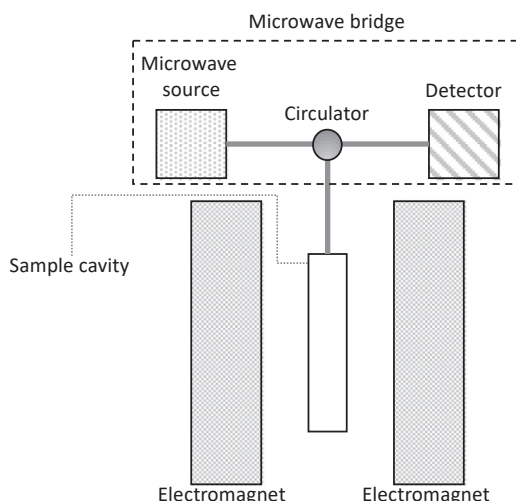


Fig. 1 A simplified block diagram of an EPR spectrometer.

the microwave energy emitted versus that reflected back by the sample is monitored. Hence, in addition to the microwave source and the detector, more electronics are needed. These additional electronics can best be explained by following the microwaves themselves. As the microwaves leave the source, they are split into two signals. The first signal is sent to the sample while the second signal is sent to the detector and acts as the reference signal. Both the sample and reference signal go through an attenuator. This attenuation step is essential to avoid saturation and allows the microwave power reaching the sample to be both accurately and precisely controlled. As the microwave energy makes its way to the sample, it goes through a circulator that directs the emitted microwaves to the sample and the reflected microwaves to the detector diode.

The sample of interest itself is placed in a cavity. The purpose of the cavity is to act as a resonator that amplifies the weak sample signal by storing this signal, so that it is not reflected back. A factor known as Q (proportional to the amount of energy stored versus energy dissipated) is used to determine the efficiency of the cavity, where the higher the Q the higher the sensitivity of the spectrometer.

Applications

Soils and soil organic matter

To obtain an EPR signal, unpaired electrons that give rise to paramagnetic systems must be present; for soils and soil organic matter this is made possible mainly by the presence of either the metal centers or organic radicals. For this chapter, we focus primarily on the organic radicals; the role of metal ions is only briefly mentioned (for a more in-depth review of the metal centers in EPR soil research, the reader is referred to the EPR chapter entitled *Historical perspectives* in the current edition). It should be noted that organic radicals within soils and soil organic matter can be divided into two classes, namely long-lived (persistent) and short-lived (transient). In real world soils and soil organic matter, both the naturally occurring persistent free radicals and transient species exist; the latter have been discussed more frequently in the scientific literature.

The majority of EPR studies on soils and soil organic matter² have utilized X-band or low frequency EPR that yielded broad and rather featureless EPR spectra. From these spectra three parameters have been derived, namely radical concentration (usually reported as spins gram^{-1} , typically in the 10^{16} – 10^{17} range), chemical information from the g -value (typically ranging from 2.0032 to 2.0043), and line width that may be informative in relation to radical heterogeneity. Linewidths are difficult to interpret as they are influenced by a range of factors such as the mobility of the radical, small to large variations within local environments surrounding the radicals, as well as chemical heterogeneity of the detected radicals themselves.

How much?; Or, spin counting: This is the quantitative aspect of EPR, and is, in some ways, technically the most difficult measurement. This is because of the nature of the EPR measurement and the fact that the spin content and its counting are usually obtained by comparison with a standard stable radical reference, e.g., diphenylpicrylhydrazyl (DPPH). This approach dictates that in order to compare samples quantitatively, the two samples must be measured in identical environments. Of utmost importance is sample placement within the cavity (a resonator in which microwave power is focused on the sample) because even the smallest deviations in the exact location can result in substantial variations in signal intensity, and hence, measured spins per gram. Approaches to alleviate this concern are to use a double cavity set-up or have the reference in a smaller concentric tube (a similar practice to that used in NMR). Other considerations when doing such an absolute comparison must be given to the power

²Due to the complex nomenclature within the soil community when it comes to soil organic matter, humic substances, humic acid, fulvic acid, humin, etc. we will utilize natural organic matter, or NOM.

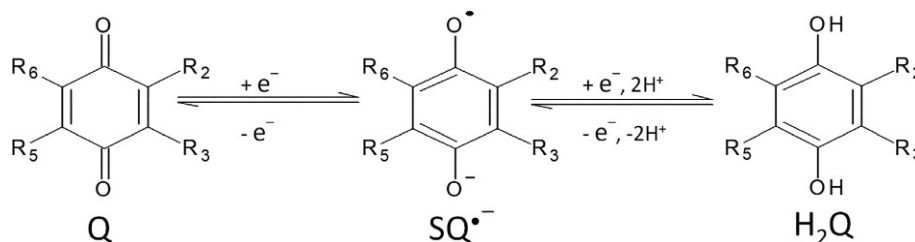


Fig. 2 Equilibrium between quinone (Q), semiquinone ($\text{SQ}^{\bullet-}$) and hydroquinone (H_2Q) compounds.

saturation effects and background levels of the instrument, sample holder, etc. Relative spins gram^{-1} , especially if the sample tube is not moved during the experiments are, as a whole, more reliable. For an in-depth discussion of quantitative EPR the reader is referred to the excellent monograph by Eaton et al. (2010).

The nature of radicals within soils and NOM (natural organic matter) sample conditions also play a major role in how many spins gram^{-1} are detected. The majority of the radicals within soils and NOM are believed to be semiquinoid radicals which arise from the equilibria between phenolic and quinoid units (Giannakopoulos et al., 2005, and references there in), an example of this is the equilibrium between quinone (Q), semiquinone ($\text{SQ}^{\bullet-}$) and hydroquinone (H_2Q) compounds illustrated in Fig. 2. These equilibria are strongly influenced by sample conditions, especially pH, bound metal ions, and the isolation technique (Paul et al., 2006; Jerzykiewicz et al., 2002). The quinoidal radical model was particularly well illustrated by Ariese et al. (2004) in which a hydroquinone/quinone ($\text{H}_2\text{Q}/\text{Q}$) was compared to a very rare well resolved EPR spectrum of an NOM sample, Laurentian fulvic acid (LFA), in which most of the LFA EPR spectrum was modelled by the five line $\text{H}_2\text{Q}/\text{Q}$ EPR spectrum. The work by Paul et al. (2006) illustrates well the influence of pH on radical content for a large number of NOM samples; the pH of the isolation condition strongly influenced the spins gram^{-1} detected (as pH increased, so did the spins gram^{-1} and vice versa). This pH effect was much stronger than that of metal complexation, which has also been shown to influence the number of detected spins; the presence of Mn(II) (manganese), Ni(II) (nickel), Co(I) (cobalt), Fe(II) (iron), and Cu(II) (copper) metal ions caused a decrease in the number of spins detected (Jerzykiewicz et al., 2002). This spin quenching is believed to be due to an antiferromagnetic interaction of the metal ion's *d*-subshell with the unpaired electrons of the $\text{H}_2\text{Q}/\text{Q}$ complex. The reversibility of this effect has also been shown by Keeler and Maciel (2003) in which samples demonstrated higher spin counts after the removal of iron and other metal ions by treatment with aqueous hydrofluoride (HF) acid. However, for such metal ions as Zn(II) (zinc), Cd(II) (cadmium), Ca(II) (calcium), and Pb(II) (lead), the opposite was shown to be true, with spin counts increasing with metal binding (Jerzykiewicz et al., 1999).

In terms of spin counts (spins gram^{-1}), especially in absolute terms, i.e., not relative values, one must use caution when reporting them as they are based on observable EPR signals which, in turn, depend on how the data were collected and on the composition of the sample, including the conditions under which the sample was isolated. This means that there is a need for as much information as possible about how the sample was collected and the conditions under which the data were collected. If possible, metal ion content should also be presented.

What is it (g-value and more)?: While spins per gram can be viewed as the quantitative (*how much?*) side of EPR analysis, *g*-values can be viewed as the qualitative (*what is it?*) side of EPR analysis. As stated above, *g*-values for soils and NOM tend typically to range from 2.0032 to 2.0043, with a reported range of 2.003–2.005 (Christoforidis et al., 2007). These numbers raise the question as to how different 2.003 is from 2.005, or more precisely, are these numbers significantly different? The short answer is yes. The longer answer is that one should not focus on the whole number per se when comparing similar classes of radicals, such as organic radicals. The issue is that, unlike NMR, EPR values are not broken down to specific elemental ranges (e.g., carbon chemical shifts), therefore, changes within organic radical signals are in a seemingly narrow region. In reality, however, the difference between 2.003 and 2.005 and reported differences within this range are significant and highly reproducible. Nevertheless, when compared with NMR data, the *g*-value for NOMs has been of limited use for deriving detailed information on chemical composition. There are two main reasons for this: (1) the general use of X-band EPR spectrometers results in an isotropic *g*-value (g_{iso}) because there is relatively little anisotropy of *g* for organic radicals and (2) the heterogeneous nature of NOM leads to a broad statistical distribution of local environments and *g*-values that tend to average into a single broad peak. Yet, there are exceptions reported in the literature, as exemplified by Ariese et al. (2004) in which a reported well-resolved five-line spectrum was similar to that of a model $\text{H}_2\text{Q}/\text{Q}$ system and could be explained by coupling³ to the four protons within an aromatic ring. Fluorescence studies in addition strengthened and supported the EPR-based findings within this study. A closer examination of the EPR spectrum of the NOM used in the study revealed a small shift in the *g*-values that is consistent with electron withdrawing ring substitutions. Thus, in certain and rather rare circumstances with the aid of simulated data, *g*-values obtained from X-band EPR can yield structural information beyond a simple single *g*-value.

Higher magnetic fields in EPR (e.g., HF-EPR), as with NMR, allow for a greater Zeeman splitting, and therefore a wider frequency range. This, in turn, leads to a larger separation of the signal in the frequency domain, and higher resolution within the obtained

³Coupling arises due to the influence of the magnetic field of one spinning system within a magnetic field on another. This type of interaction is distance-dependent and it is short range.

spectrum. Consequently, at large enough magnetic fields, the g_{iso} signal will be deconvoluted into its three g tensor components, $g_{\text{iso}} = (g_x + g_y + g_z)/3$ (Christoforidis et al., 2007), as presented theoretically in Fig. 3. In particular, this is the case for the application of HFEPR to NOM. The power of HFEPR in the analysis of NOM was illustrated by Christoforidis et al. (2007) on four different humic acids (three from coal and one from soil), in which all four yielded well-resolved g_x , g_y , and g_z tensor components. Fig. 4 illustrates the advantage of HFEPR versus LFERP on a real humic acid sample, HA1. In addition, it was shown with spectral simulation that the EPR signal could be broken down into two radical types with different collective g_x , g_y , and g_z values that were prevalent at different pHs (the observed pH behavior was simple and reversible). These higher resolution data enable a more detailed structural analysis, especially with the aid of model compounds (semiquinone, β -carotene, and pheophytin radicals), to provide insights such as the unpaired electron being in a p -orbital, most likely conjugated p -orbitals, and the radical centers being most probably p -type, with carbon-centered radicals dominant at low pH and oxygen- or nitrogen-centered radicals dominant at high pH. Since this pioneering study, HFEPR has been used to study natural organic matter, especially in relation to metal binding.

Metal binding

Electron paramagnetic resonance has a long history in the literature in relation to metal binding. For reviews of those studies the reader is referred to the following work of Senesi (Senesi, 1990). As a whole, the majority of the metal EPR research focused either on the metal ion itself or some form of quantitative analysis (as discussed above), but did not focus directly on the NOM, with one notable

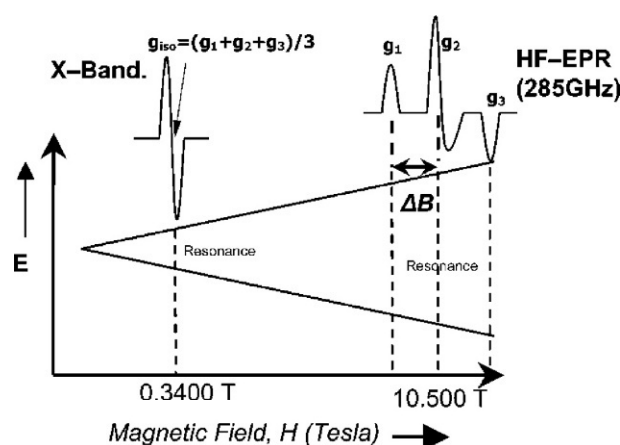


Fig. 3 A theoretical illustration of the increased resolution provided by HFEPR over LFERP as the g_{iso} is separated into g_1 , g_2 , and g_3 (e.g., g_x , g_y , and g_z) as the magnetic field and associated frequency increase. Reproduced with permission from Christoforidis KC, Un S, and Deligiannakis Y (2007) High-field 285 GHz electron paramagnetic resonance study of indigenous radicals of humic acids. *Journal of Physical Chemistry A* 22: 11860–11866, doi: 10.1021/jp0717692.

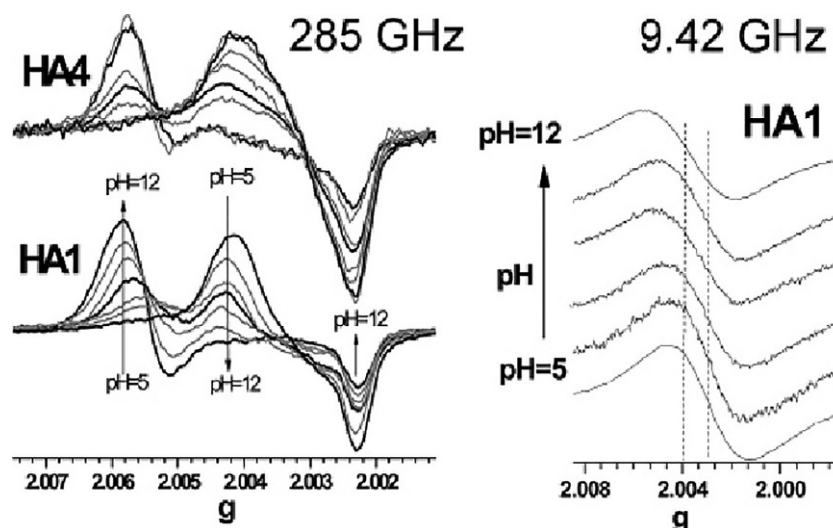


Fig. 4 Resolution advantage of HFEPR (left) over LFERP (right) for a humic acid (HA1) due to the ability to deconvolute the g_{iso} signal into its g_1 , g_2 , and g_3 contributions. Reproduced with permission from Christoforidis KC, Un S, and Deligiannakis Y (2007) High-field 285 GHz electron paramagnetic resonance study of indigenous radicals of humic acids. *Journal of Physical Chemistry A* 22: 11860–11866, doi: 10.1021/jp0717692.

exception being that on Pb(II). The development and application of HFEPR has allowed a re-examination of metal binding to NOM, especially from the NOM viewpoint through the resolution of g_x , g_y , and g_z tensors. A multi-metal study by Christoforidis et al. (2010) utilized the high resolution of HFEP to show that Mg(II) (magnesium) interacts with the NOM under study by purely electrostatic interactions and, in doing so, dramatically alters the H-dissociation of the type I and II radicals, as shown by pH-dependent studies. The same study used HFEPR to show that Mg (II), Cd(II), Sn(II) (tin), and Pb(II) have very different binding modes with NOM. For instance, Cd(II) binding is shown to alter the g tensors of both type I and II radicals, resulting in a less anisotropic spectrum, which suggests covalent binding. It was also found that Sn(II) bonds covalently to the NOM studied, and with the resolution of HFEPR it was also possible to observe (by subtracting spectra of the NOM studied after 2 h from that after 5 days of incubation with Sn(II)) the formation of a new type of radical within the NOM due to its binding with Sn(II). This new type of NOM radical formation with metal binding had been reported previously for Pb(II) when studied by X-band EPR; however, HFEPR has allowed new and significant insights into this phenomenon (Jerzykiewicz, 2004; Christoforidis et al., 2010), with exemplarily data presented in Fig. 5.

The new type of radical formed by Pb(II) binding to NOM has an unusually low g -value (~ 2.001 : Jerzykiewicz, 2004; Giannakopoulos et al., 2005), which is less than that of a free electron, and was initially explained by a localization of spin density on the Pb(II). This very unusual finding has continued to be investigated by Jerzykiewicz and co-workers using mainly a combination of EPR on NOM and model compounds in concert with density functional theory (DFT)-based calculations (Jerzykiewicz, 2004; Giannakopoulos et al., 2005) to gain a detailed molecular understanding of the situation. This is significant given its environmental importance in terms of Pb(II) speciation as well as radical formation within NOM. With the higher resolution provided by HFEPR, two complexes were found for the model compounds with different g tensor components. The DFT calculations were then able to model the experimental data using a two-step mechanism, in which a diamagnetic Pb(II) complex is formed and subsequently oxidized. From the point of view of NOM, this translates into Pb(II) being complexed by carboxyl oxygens and the hydroxyl oxygens then being protonated.

Spin labels

Not only can metal binding be monitored by EPR, but so can organic molecule (contaminant) sorption (binding/association) using organic spin labels (SLs). The inherent sensitivity of EPR is a major advantage because it allows for low concentrations of a pollutant to be monitored not just quantitatively but also qualitatively and non-destructively. There have been two major approaches to the use of SLs. The first is direct monitoring of the SL itself in the presence and absence of NOM under different conditions, while the second is monitoring of the effect of SL-induced paramagnetic relaxation on the NMR signal of NOM.

SL monitoring: In Ferreira et al. (2001), EPR monitoring of an SL and its sorption to a soil-extracted humic acid (HA) showed that the EPR spectral signal of a N-oxyl-4',4'-dimethyloxazolidine derivative of 5-ketostearic acid (5-SASL) was clearly visible over the inherent (radicals native to NOM) organic radical signal of the NOM sample under the conditions used in the study, including the hyperfine splitting within the 5-SASL EPR spectrum. By varying the pH, and hence conformation of the NOM and monitoring

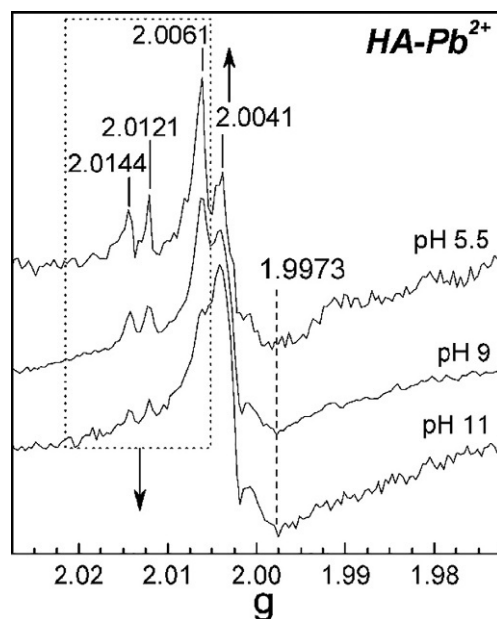


Fig. 5 A set of HFEPR spectra for Pb(II) bound to a soil humic acid (HA) at three different pHs illustrate how the resolution of the HFEPR technique enables the identification of two unique radical species within this system. Reproduced with permission from Christoforidis KC, Un S, and Deligiannakis Y (2010) Effect of metal ions on the indigenous radicals of humic acids: high field electron paramagnetic resonance study. *Environmental Science and Technology* 44: 7011–7016, doi: 10.1021/es101708f.

the linewidths of the 5-SASL EPR signal, the authors proposed a two-binding site (interior and exterior) model, in which the interior sites depended strongly on NOM conformation (determined by pH dependence) and the lower affinity exterior hydrophobic sites became more important at higher pH (>5). Aleksandrova (2013) then extended this research to soils and SLs differing by a single functional group on the opposite side of the ring to the oxygen-centered nitroxide radical, utilizing a base spin label called tetramethylpiperidin-1-oxyl (TEMPO), with an example of the information-rich spectra presented in Fig. 6. This research, once again, used the hyperfine EPR spectral features within the EPR spectrum of the SL, in which there was a shift of the outer lines inwards showing a change in the polarity of the SL as it interacted with the soils. Furthermore, changes in the linewidths of the EPR signal, and hence, mobility of the SLs were noted with an increase and decrease in mobility for the 4-hydroxyl TEMPO and 4-amino TEMPO SLs, respectively, with increasing incubation time. In addition, the signal of the native soil radicals decreased as the SL interacted with the soil over time. This study showed that EPR can be used to probe the chemical interaction of SL mimicking pollutants in soils and allows one to gain insight into the mechanism of binding with soil organic matter, including the role of pollutant functional groups. Matthies et al. (2016) expanded the work of Aleksandrova (2013) by utilizing amino-functionalized SLs in the form of 4-amino TEMPO and 2,5,5-trimethyl-2-(3-aminophenyl)pyrrolidine-1-oxyl, anilino-NO, to compare aliphatic versus aromatic functionality in terms of NOM binding with Leonardite humic acid. This study, once again, revealed the importance of functionality when a pollutant binds to NOM; anilino-NO showed two binding motifs compared with a single motif for 4-amino TEMPO. The fitting of simulated spectra of the experimentally-determined spectra yielded a rotational correlation of 95 ps for 4-amino TEMPO, which indicated an ionic type interaction with the NOM. For the anilino-NO, two rotational correlation times were determined and assigned to weak and strong interactions within the NOM. Further kinetic studies with laccase and N-tert-butyl- α -phenylnitrone (radical scavenger) strongly indicated that a covalent bond forms between the anilino-NO and the NOM. Matthies and co-workers (Ricke et al., 2021) extended this approach further with a model humic-clay system to demonstrate that spin labeled sulfonamides (SL-SDZ) were covalently bound to quinone moieties within NOM and that the clay mineral increased immobilization of the SL-SDZ.

Paramagnetic Relaxation Probes: The electron magnetic moment that is at the core of EPR is about 650 times greater than the magnetic moment of a proton. This magnetic moment perturbs the local magnetic field, and induces relaxation pathways through dipolar coupling between the electron and nucleus (r^{-6} distance dependence, leading to a strongly localized effect), which results in fast relaxation of the nuclear magnetic moment and, in turn, line broadening and eventual disappearance of the NMR signals. The longer the relaxation time of the electron magnetic moment, the more it perturbs the NMR signal (this is due to the much longer nucleus relaxation times in its shielded local environment than that of an electron). This phenomenon is known as paramagnetic relaxation. Thus, a radical not only reveals its local environment by its EPR signal but also by the way it eliminates the NMR signal. The research of Bleam and co-workers (Chien et al., 1997) and that of Pignatello and co-workers (Lattao et al., 2012) are two examples of this phenomenon at work. Chien et al. (1997) used EPR to reveal that TEMPO, via its hyperfine coupling, interacts with the surface of the NOM studied. The work then used gadolinium ethylenediaminetetraacetic acid, Gd-EDTA, (hydrophilic paramagnetic probe) and TEMPO (hydrophobic paramagnetic probe) to show that TEMPO and not Gd-EDTA induces paramagnetic relaxation (via ^{19}F NMR) within an ^{19}F -labeled atrazine sorbed to the solubilized NOM. This provided strong evidence that atrazine sorbed to the hydrophobic surface domains within the NOM studied. Lattao et al. (2012) utilized TEMPO and 4-hydroxyl TEMPO paramagnetic probes as pollutant surrogates. Electron paramagnetic resonance was used to show, by line broadening, that TEMPO was confined within the Pahokee peat humic acid NOM. The research by Lattao et al. (2012) did not reveal any specific functional NOM group binding by either of the two paramagnetic probes examined by paramagnetic relaxation.

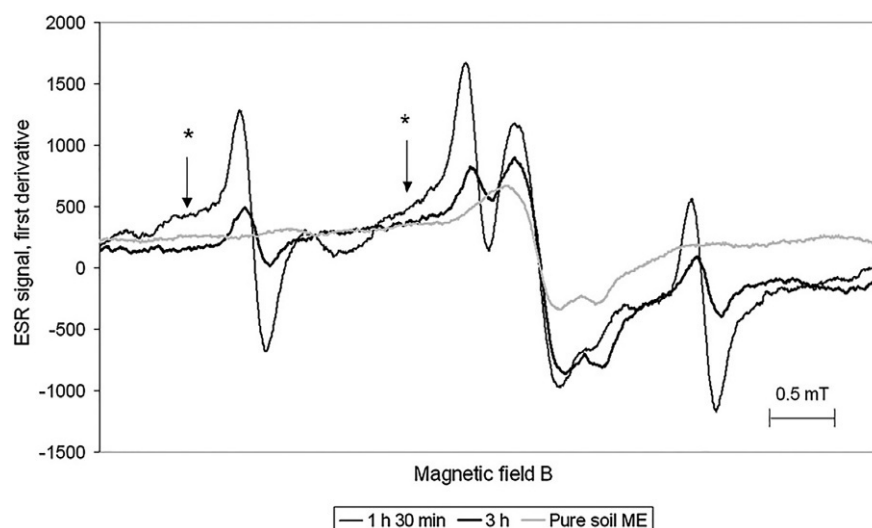


Fig. 6 A set of EPR spectra illustrating the interaction of tetramethylpiperidin-1-oxyl (TEMPO) with a whole soil at increasing incubation times. Reproduced with permission from Aleksandrova (2013) Spin labeling ESR investigation of the molecular environment of soil interacting with chemical organic contaminants. *Journal of Geochemical Exploration* 129: 6–13, doi: 10.1016/j.gexplo.2012.12.016.

Environmentally persistent free radicals (EPFRs)

In the past decade or so, a new class of organic compounds of concern have come to the forefront of environmental research, namely environmentally persistent free radicals (EPFRs), and EPR has been the central tool in their identification, classification, and study. The EPFRs result from the formation of an organic radical within an environmental matrix, including soils. Although naturally occurring environmentally persistent free radicals, nEPFRs, have been known for decades as illustrated by the early application of EPR to NOM samples, including soil organic matter within whole soils, a second class of EPFRs has been found to be the product of organic pollutant transformation. In the work of de la Cruz et al. (2011, 2014) EPR was essential in showing that EPFRs are at much higher concentrations within soils at a range of polluted sites, and the soil fraction(s) that contained the EPFRs. Further research by others has shown just how prevalent these soil-associated EPFRs are within the environment. Electron paramagnetic resonance has been central in determining that these soil EPFRs arise from two main mechanisms, namely (i) atmospheric deposition, in situ burning, or biochar incorporation and (ii) redox reaction with a metal center.

Metal Center Driven: With the use of EPR and soil fractionation, Fig. 7, de la Cruz et al. showed that ~90% of the EPFRs at a polluted wood treatment site were associated with the mineral/humin fraction of the soil and were oxygenic carbon-centered (g -factor, 2.0031) in nature. Further, by combining XAFS and EPR, Nwosu et al. (2016a, b) showed that Fe(III)- and Cu(II)-exchanged montmorillonite clays, when dosed with phenol, produced oxygenic-carbon centered (g -factor of 2.0034) EPFRs under environmentally relevant conditions, and provided clear evidence that this effect was due to the transfer of an electron from phenol to the Fe(III) or Cu(II) center. Based on the long lifetime (persistence) and resistance to oxygen degradation determined by EPR (as well as XRD) data, the EPFRs were thought to be formed interstitially. In addition, based on the loss of intensity of the EPR signal, humidity was found to be a major EPFR deactivator, while the presence of a synthetic surrogate soil organic matter was shown to induce even longer periods of EPFR persistence. Furthermore, EPR spin trapping using 5,5-dimethyl-1-pyrroline-N-oxide (DMPO), revealed that EPFRs formed in metal exchange clay systems could induce the formation of biologically relevant reactive oxygen species. Other work on montmorillonite clay systems has also shown with EPR analysis that EPFRs form under environmentally relevant conditions, with the transformation of anthracene into either a carbon, g -factor: 2.0028–2.0030, or an oxygenic carbon-centered radical, g -factor: 2.0032–2.0038 respectively (Jia et al., 2016).

Reactive Oxygen Species: Electron Paramagnetic Resonance has also been key in understanding the role of EPFRs in the decomposition of organic pollutants. The generally accepted mechanism involves the EPFR-generated reactive oxygen species (monitored by EPR in DMPO trapping experiments) which, in turn, react with the pollutant of interest (along the line of Fenton chemistry). However, through detailed EPR analysis Yang et al. (2016) showed that EPFRs within biochars decomposed *p*-nitrophenol by an electrophilic attack.

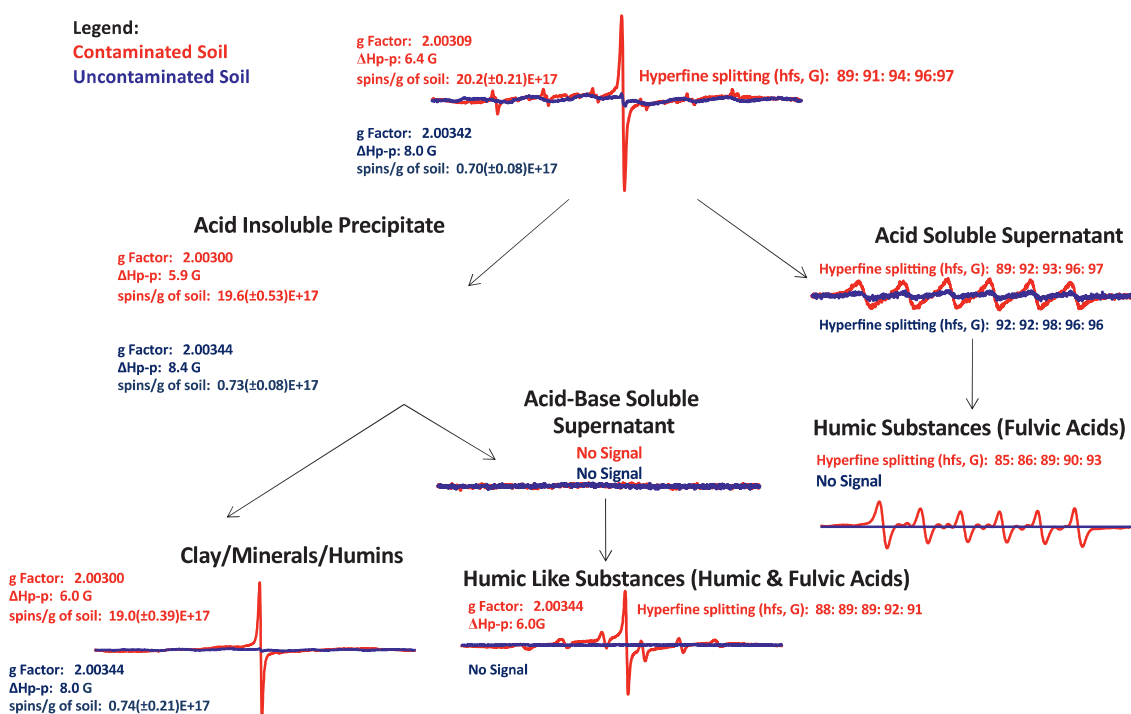


Fig. 7 Tracking the radical, including environmentally persistent free radicals (EPFRs), containing soil fraction(s). Adapted from Fig. 5 and reproduced with permission from de la Cruz AL, Gehling W, Lomnicki S, Cook R, and Dellinger B (2011) Detection of environmentally persistent free radicals at a superfund wood treating site. *Environmental Science and Technology* 45: 6356–6365, doi: 10.1021/es2012947.

Summary

Within soil science the most popular magnetic resonance technique is nuclear magnetic resonance, NMR, due to its high resolution and ability to analyze samples selectively by nuclear spin manipulation. Its electron counterpart, electron paramagnetic resonance, or EPR, spectroscopy, has been much less popular. There are several reasons for this, including hardware accessibility and difficulty in spectral interpretation; however, the major reason has been the low resolution of the majority of soil-focused EPR spectra reported within the literature, especially for organic-focused research. However, as the discussion and selected examples above demonstrate, in the world of soil science EPR research has an important place in the modern era; therefore, the much greater sensitivity of EPR compared with NMR holds great promise for the technique. A major factor for delivering on this promise is how EPR experiments and EPR data are analyzed. The advancement of EPR and the unique insights it enables in soil research results from careful experimental design, taking the limitations of the technique into consideration and, in some cases, turning them into an advantage. An example of this is the use of spin labels as surrogate pollutants with the sensitivity of the hyperfine structure within the EPR spectrum of these spin labels serving as a probe for soil sorption environments. The spectral interpretation required in such experiments has been greatly aided by modern computational methods. It is also clear that EPR complements and is complemented by other techniques to allow for a fuller interpretation of the data obtained. Finally, the introduction of commercial HF-EPR instruments and the associated higher resolution spectral data enable acquisition of higher quality and more information-rich EPR spectra for soils and soil systems, and a better understanding of radical entities within these systems.

Acknowledgment

The authors acknowledge support from the National Institute of Environmental Health Science Superfund Research Program through Grant P42 ES013648-03. The authors thank Elzbieta Cook for her editing skills.

References

- Aleksandrova (2013) Spin labeling ESR investigation of the molecular environment of soil interacting with chemical organic contaminants. *Journal of Geochemical Exploration* 129: 6–13. <https://doi.org/10.1016/j.gexplo.2012.12.016>.
- Ariese F, van Assema S, Gooijer C, Bruccoleri AG, and Langford CH (2004) Comparison of Laurentian Fulvic Acid luminescence with that of the hydroquinone/quinone model system: Evidence from low temperature fluorescence studies and EPR spectroscopy. *Aquatic Science* 66: 86–94. <https://doi.org/10.1007/s00027-003-0647-8>.
- Atherton NM, Cranwell PA, Floyd AJ, and Haworth RD (1967) Humic acid I. ESR spectra of humic acids. *Tetrahedron* 23: 1653–1667. [https://doi.org/10.1016/S0040-4020\(01\)82564-3](https://doi.org/10.1016/S0040-4020(01)82564-3).
- Chien Y-Y, Kim E-G, and Bleam WF (1997) Paramagnetic relaxation of atrazine solubilized by humic micellar solutions. *Environmental Science and Technology* 31: 3204–3208. <https://doi.org/10.1021/es9701927>.
- Christoforidis KC, Un S, and Deligiannakis Y (2007) High-field 285 GHz electron paramagnetic resonance study of indigenous radicals of humic acids. *Journal of Physical Chemistry A* 111: 11860–11866. <https://doi.org/10.1021/jp0717692>.
- Christoforidis KC, Un S, and Deligiannakis Y (2010) Effect of metal ions on the indigenous radicals of humic acids: High field electron paramagnetic resonance study. *Environmental Science and Technology* 44: 7011–7016. <https://doi.org/10.1021/es101708f>.
- Corvaja C (2009) Introduction to Electron paramagnetic resonance. In: Brustolon M and Giamlo E (eds.) *Electron Paramagnetic Resonance: A Practitioner's Toolkit*, pp. 3–36. Hoboken, New Jersey: John Wiley & Sons Inc.
- de la Cruz AL, Gehling W, Lomnicki S, Cook R, and Dellinger B (2011) Detection of environmentally persistent free radicals at a superfund wood treating site. *Environmental Science and Technology* 45: 6356–6365. <https://doi.org/10.1021/es2012947>.
- de la Cruz ALN, Cook RL, Dellinger B, Lomnicki SM, Donnelly KC, Kelly MA, and Cosgriff D (2014) Assessment of environmentally persistent free radicals in soils and sediments from three superfund sites. *Environmental Science: Processes and Impacts* 16: 44–52. <https://doi.org/10.1039/c3em00428g>.
- Eaton GR, Eaton SS, Barr DP, and Weber RT (2010) *Quantitative EPR*. New York: Springer-Verlag/Wien.
- Ferreira JA, Nascimmento OR, and Martin-Neto L (2001) Hydrophobic interactions between spin-label 5-SASL and humic acid as revealed by ESR spectroscopy. *Environmental Science and Technology* 35: 761–765. <https://doi.org/10.1021/es0010251>.
- Giannakopoulos E, Christoforidis KC, Tsipis A, Jerzykiewicz M, and Deligiannakis Y (2005) Influence of Pb(II) on the radical properties of humic substances and model compounds. *Journal of Physical Chemistry A* 109: 2223–2232. <https://doi.org/10.1021/jp045121q>.
- Jerzykiewicz M (2004) Formation of new radicals in humic acids upon interaction Pb(II) ions. *Geoderma* 122: 305–309. <https://doi.org/10.1016/j.geoderma.2004.01.017>.
- Jerzykiewicz M, Czechowski F, Jerzierski A, and Drozd J (1999) Organic radicals and paramagnetic metal complexes in municipal solid waste composts. An EPR and chemical study. *Chemosphere* 92: 253–268. [https://doi.org/10.1016/S0045-6535\(99\)00107-1](https://doi.org/10.1016/S0045-6535(99)00107-1).
- Jerzykiewicz M, Jerzierski A, Czechowski F, and Drozd J (2002) Influence of metal ions binding on free radical concentration in humic acids. A quantitative electron paramagnetic resonance study. *Organic Geochemistry* 33: 265–268. [https://doi.org/10.1016/S0146-6380\(01\)00158-9](https://doi.org/10.1016/S0146-6380(01)00158-9).
- Jia H, Nulaji G, Gao H, Wang F, Zhu Y, and Wang C (2016) Formation and stabilization of environmentally persistent free radicals induced by the interaction of anthracene with Fe(III)-modified clays. *Environmental Science and Technology* 50: 6310–6319. <https://doi.org/10.1021/acs.est.6b00527>.
- Keeler C and Maciel GE (2003) Quantitation in the solid-state ¹³C NMR analysis of soil and organic soil fractions. *Analytical Chemistry* 75: 2421–2431. <https://doi.org/10.1021/ac020679>.
- Lagerkrantz C and Yhland M (1963) Photo-induced free radical reactions in the solution of some tars and humic acids. *Acta Chemica Scandinavica* 17: 1299–1306. <https://doi.org/10.3891/acta.chem.scand.17-1299>.
- Lattao C, Cao X, Li Y, Mao J, Schmidt-Rohr K, Chappell MA, Miller LF, dela Cruz AL, and Pignatello JJ (2012) Sorption selectivity in natural organic matter studied with nitroxyl paramagnetic relaxation probes. *Environmental Science and Technology* 46: 12814–12822. <https://doi.org/10.1021/es302157j>.
- Matthies M, Glirk K, Thelling M, Hildeg K, and Steinhoff H-J (2016) Kinetics of rapid covalent bond formation of aniline with humic acid: ESR investigations with nitroxide spin labels. *Applied Magnetic Resonance* 47: 627–641. <https://doi.org/10.1007/s00723-016-0782-8>.

- Nwosu UG, Khachatryan L, Youm SG, Roy A, de la Cruz ALN, Nestrov EE, Dellinger B, Cook RL, Roy A, de la Cruz ALN, Dellinger B, and Cook RL (2016a) Model system study of environmentally persistent free radicals formation in a semiconducting polymer modified copper clay system at ambient temperature. *RSC Advances* 6: 43453–43462. <https://doi.org/10.1039/C6RA08051K>.
- Nwosu UG, Roy A, dela Cruz ALN, Dellinger B, and Cook R (2016b) Formation of environmentally persistent free radical (EPFR) in iron(III) cation-exchanged smectite clay. *Environmental Science: Processes and Impacts* 18: 42–50. <https://doi.org/10.1039/C5EM00554J>.
- Paul A, Stosser R, Zehl A, Zwirnmann E, and Voght and Steinberg, C. E. (2006) Nature and abundance of organic radicals in natural organic matter: effect of pH and irradiation. *Environmental Science and Technology* 40: 5897–5903. <https://doi.org/10.1021/es060742d>.
- Ricke A, Kalai T, Steinhoff H-J, and Matthies M (2021) Interaction kinetics and accessibility of sulfadiazine in model clay-humic acid suspension: electron spin resonance investigations with nitroxide spin label. *Science of the Total Environment* 796: 149042. <https://doi.org/10.1016/j.scitotenv.2021.149042>.
- Senesi N (1990) Applications of ESR spectroscopy in soil chemistry. In: Stewart BA (ed.) *Advances in Soil Science*. vol. 14, pp. 77–130. New York: Springer-Verlag.
- Weil JA and Bolton JR (2007) *Electron paramagnetic resonance: elementary theory and practical applications*, 2nd edn, Hoboken, New Jersey: John Wiley & Sons Inc.
- Yang J, Pan B, Li H, Liao S, Zhang D, Wu M, and Xing B (2016) Degradation of *p*-nitrophenol on biochars: role of Persistent free radicals. *Environmental Science and Technology* 50: 694–700. <https://doi.org/10.1021/acs.est.5b04042>.

Fluorescence Spectroscopy: Part I Principles

Yong Zhang, Shuai Guo, and Chaoxian Wei, State Key Laboratory of Marine Environmental Science of China, Xiamen University, Xiamen, PR China; College of Environment and Ecology, Xiamen University, Xiamen, PR China

© 2023 Elsevier Ltd. All rights reserved.

Introduction	544
Basic principles, methodology, and instrumentation	544
The fluorescence phenomenon	544
Instruments and experimental procedures	545
Fluorescence spectra	546
Fluorescence efficiency, lifetime, and intensity	547
Limitations and problems	547
Molecular and environmental factors affecting fluorescence	548
Fluorescence quenching	548
Fluorescence polarization	549
Fluorescence correlation spectroscopy	549
Laser induced fluorescence spectroscopy	549
X-ray fluorescence spectroscopy	550
Time-resolved fluorescence spectroscopy	550
Conclusion	551
References	551

Abstract

Fluorescence spectroscopy is a noninvasive, nondestructive and highly sensitive technology that requires minimal sample pretreatment. Therefore, it is widely used in the study of soil components. This chapter describes the principles of luminescence of fluorescent substances and the fluorescence instrument used. In addition, the basic concepts of fluorescence efficiency, lifetime, intensity, polarization, and quenching, and environmental factors that affect the luminescence of fluorescent substances are also elaborated. Finally, recently developed fluorescence spectroscopy techniques commonly used in the detection of soil organic components are introduced, including fluorescence correlation spectroscopy, laser induced fluorescence spectroscopy, X-ray fluorescence spectroscopy and time-resolved fluorescence spectroscopy.

Key points

- The luminescence principles of fluorescent substances are elaborated.
- Important fluorescence characteristic parameters are introduced.
- Some fluorescence spectroscopy techniques commonly used in the detection of soil organic components are introduced.

Introduction

Classical monodimensional fluorescence spectroscopy in the emission, excitation, and synchronous scan modes and total luminescence or tridimensional (3D) fluorescence are noninvasive, nondestructive, and highly sensitive techniques that require minimal sample pretreatment. Therefore, fluorescence spectroscopy is widely used in the study of organic and inorganic soil components. With the rapid development of fluorescence technology, fluorescence spectroscopy has proven to be a very promising and useful, especially as an in-situ method for determining soil organic components. The diversity of inherent fluorescence structures in soil organic components has provided invaluable information on their structural and functional chemistry and reactivity as a function of relevant environmental factors such as pH and ionic strength.

Basic principles, methodology, and instrumentation

The fluorescence phenomenon

Every molecule has a series of closely spaced energy levels (Fig. 1). After a molecule absorbs energy, such as light, an electron is promoted from the ground state, S_0 , to an excited singlet state, S_1 , S_2 , etc. Typical transitions of this kind are very likely in complex molecular systems containing atoms with lone pairs of electrons, such as oxygen or nitrogen, and aromatic or aliphatic conjugated unsaturated systems capable of considerable electron delocalization. Any excited state electrons then relax to the lowest vibrational

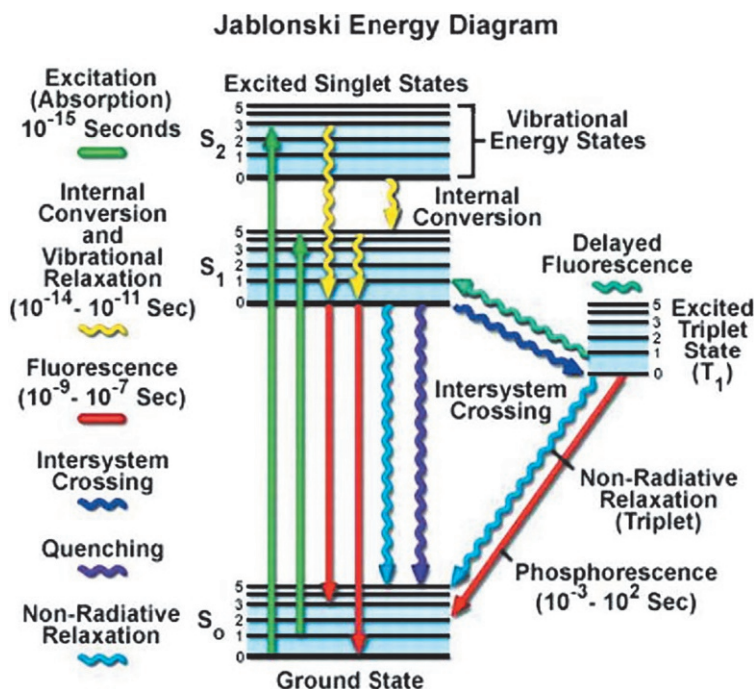


Fig. 1 Jablonski energy diagram. Adapted from Davidson MW (2009) Alexander Jablonski, PhD: Fluorescence Spectroscopy. *Laboratory Medicine* 40: 694–695.

level of that state, and, subsequently, energy is released by ‘internal conversion,’ dropping the electron to the lowest vibrational level of the lowest excited state, S₁. Relaxation from the S₁ state, which has a lifetime of 10⁻⁹ s, to the ground state S₀ can then occur by different mechanisms (Fig. 1). Recently, (Kim et al., 2021) found that the lifetime of S₁ is greatly increased to 1.80 ns at the zero-point level when catechol (a toxic organic compound, C₆H₄(OH)₂) is complexed with water. The above relaxation processes include: (1) the radiative photo process of fluorescence, which involves the emission of photons less energetic than those absorbed because of the energy loss incurred during internal conversion processes; and (2) various nonradiative processes, such as: (a) internal conversion processes from relaxation through internal vibrations or through collisions with solvent and or solute molecules; (b) transfer of energy to another compound with similar wavelengths and close to each other, resulting in a photochemical reaction or the formation of an excited-state dimer (excimer) or excited-state complex (exciplex) which then emits photons at longer wavelengths than fluorescence; and (c) intersystem crossing to a triplet state, T₁, that has a lifetime much longer (often >10⁻⁵ s) than its S₁ precursor, which is followed by either a delayed release of energy known as phosphorescence, or by relaxation through internal conversion processes.

Instruments and experimental procedures

In general, fluorescence measurements require only a small amount of sample preparation. Solutions and diluted colloidal suspensions (<1 g solid⁻¹) can be placed directly into quartz cells, whereas the use of dense colloidal suspensions should be avoided because these result in increases in light scattering and absorption, and a subsequent decrease in signal-to-noise ratio. Furthermore, the inner filter effect caused by excessive fluorophore concentration or coexistence with other absorbable species cannot be disregarded. The inner filter effect can be divided into two types when the solution concentration is too high: (1) after the incident light is strongly absorbed by the fluorescent substance in the front part of the sample cell, the incident light received by the fluorescent substances in the back of the sample cell will be greatly weakened, resulting in a decrease in the fluorescence intensity; (2) the absorption of incident light by impurities in the solution reduces the excitation light intensity of the fluorescent substance.

The principal components of a fluorescence spectrophotometer used for steady-state measurements are shown schematically in Fig. 2. The excitation energy is provided by a high-energy continuous light source, usually a xenon discharge lamp. Other light sources, such as a pulse source, a light-emitting diode source and long-lived xenon lamps, have been widely used in recent years. Here a xenon arc lamp is used as an example (Fig. 2). The excitation monochromator contains two concave gratings of reduced hetero astigmatism. Both monochromators are motorized and can scan wavelengths automatically. Fluorescence is obtained with a detector, such as photomultiplier tube and quantified with an appropriate electronic device. The output component is typically displayed graphically and stored digitally. The desired excitation wavelength is selected by passing incident radiation through a monochromator before it enters the sample cell, preferably a quartz cell. Radiation emitted by the sample is generally monitored at an angle of 90 degrees to the excitation radiation to minimize signal distortion from light scattering, or by front-surface detection for a solid sample. Wavelength selection of the emitted radiation is obtained with another monochromator. The light emitted is then

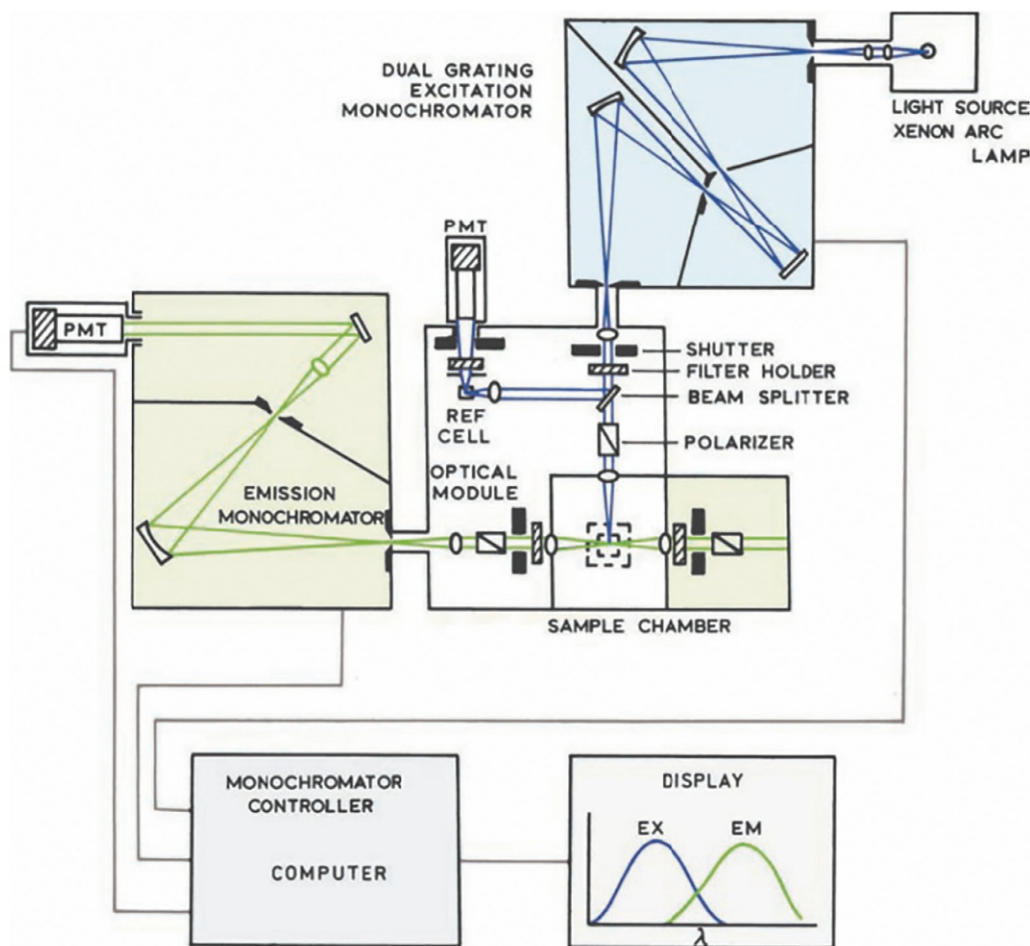


Fig. 2 Schematic diagram of the fluorescence spectrophotometer using a xenon arc lamp as an example. Adapted from Lakowicz JR (2006) *Principles of Fluorescence Spectroscopy*, Third Edition. New York: Springer. 28–29.

measured by the detector, usually a photomultiplier tube. The detector is usually coupled with a device camera (for example, charged-couple device (CCD) camera), high speed camera, or other imaging devices. The instrument is completed by a readout device, usually a computer, or, preferably, a microprocessor for data analysis.

Temperature control devices such as thermostatic cell holders, circulating water temperature controllers, and a semiconductor thermostatic device to control sample temperature and cell stirrers have been added to most commercial spectrofluorometers. Laser-induced fluorescence spectroscopy (LIFS) has been used widely for exploring the properties and environmental behavior of soil organic components. The entire instrument can be put under computer control with several benefits in precision, calculation of Φ , and corrections for factors affecting fluorescence, such as the inner-filter effect and photodecomposition. Time-resolved fluorescence (TRF) spectra or fluorescence lifetime measurements, measures the time a molecule is in the excited state. It involves the use of pulsed or modulated light sources and a rapid-detecting photomultiplier, a photodiode array, or a time-correlated single-photon counter. Fluorescence lifetime, a widely used method for the time domain, includes time-dependent single photon counting (TCSPC), generating a time of pulsed photon arrival by pulse sampling and a time-gated histogram to quantify the fluorescence decay curve directly (Bitton et al., 2021). The frequency-domain method measures the amplitude and phase shift of the fluorescent signal in response to modulated excitation light (Alfonso-Garcia et al., 2021). Polarized fluorescence measurements can be obtained by adapting a conventional instrument. Briefly, the appropriate oriented polarizers are placed between the excitation monochromator and the sample, and between the sample and the emission monochromator.

Fluorescence spectra

Conventional monodimensional fluorescence spectra can be obtained in three modes: emission, excitation, and synchronous-scan excitation on samples dissolved or suspended in liquid. The emission spectra were recorded by measuring the relative intensity of radiation emitted as a function of wavelength at a constant excitation wavelength. The excitation spectra were recorded by measuring the emission intensity at a fixed wavelength when changing the excitation wavelength. According to the relationship

that the excitation and emission wavelengths have with each other during simultaneous scanning, synchronous fluorescence spectrometry is subdivided into four types: (1) constant wavelength; (2) constant energy; (3) variable angle; and (4) matrix isopotential synchronous fluorescence spectrometry. The synchronous-scan technique can selectively increase the intensity of specific peaks, thus improving the resolution of spectral structure by decreasing the intrinsic broadness and overlapping of bands that occur in emission and excitation spectra of molecularly complex and heterogeneous multicomponent samples.

Total luminescence or 3D fluorescence spectra provide a complete representation of the fluorescence spectral features of a sample in the form of an excitation-emission matrix (EEM), in which fluorescence intensity is measured as a function of the excitation wavelength on one axis and emission wavelength on the other. The EEM spectra can be compared using the uncorrected matrix correlation (UMC) method. The UMC value equals 1 where the two matrices are identical and decreases towards zero with decreasing similarity between them. For the fluorescence zone integration (FZI) method, it could quantify differences in EEM among various samples, and measure the continuum of binding sites in metal and organic complexes of a sample. However, the relatively poor data processing efficiency of the FZI method limits its application for analyzing a large number of environmental samples. Recently, EEM combined with parallel factor (EEM-PARAFAC) analysis provides detailed spectral information from the fluorescence spectrum of the soil organic component (Qin et al., 2020). The accuracy and efficiency of multi-component fluorescence spectrum analysis is improved by separating the fluorescence signal into relatively independent components, especially for a large number of samples (Ravansari et al., 2020).

Fluorescence efficiency, lifetime, and intensity

Fluorescence efficiency, or quantum yield, Φ_f , measures the effectiveness with which the adsorbed energy is reemitted, and is defined as the ratio of total energy emitted as fluorescence to the total energy adsorbed. Absolute measurements of efficiency are difficult because they require calibrated sources and detectors. However, the relative fluorescence efficiency of a sample, Φ_s , can be obtained easily by measuring the fluorescence intensity of a dilute solution (I_k) of a standard of known efficiency, Φ_k , such as quinine sulfate, and that of the sample (I_s):

$$\Phi_s = \Phi_k (I_s q_k A_k / I_k q_s A_s), \quad (1)$$

where the I values are determined by integrating the area beneath the corrected fluorescence spectrum, q is the relative photon output of the source at the excitation wavelength taken directly from the curve, and A is the absorbance.

Fluorescence lifetime depends on the molecular species and is sensitive to environmental factors such as conformation and binding state of the absorbing molecule, temperature, pH, or solvent viscosity. Fluorescence lifetime (τ), refers to the probability of finding a given molecule in the excited state at time t after the excitation light source is turned off, and represents the mean lifetime of the excited state. The value of τ can be calculated by measuring the fluorescence intensity I at time t , $I(t)$, according to:

$$I(t) = I_M \exp(-t/\tau) \quad (2)$$

where I_M is the maximum intensity with the excitation light source on, i.e., during excitation, and t is the time elapsed after the excitation light source is turned off.

Fluorescence intensity, I , depends on the molar concentration of the adsorbing species in very dilute solutions, C , and the fluorescence efficiency, Φ_f , according to:

$$I = \Phi_f I_0 [1 - \exp(-\epsilon bC)] \quad (3)$$

where I_0 is the intensity of incident radiation, ϵ is the molar absorptivity at the excitation wavelength, and b is the path length of the cell. For very dilute solutions, where ϵbC is sufficiently small, Eq. (3) reduces to a linear relationship:

$$I = \Phi_f I_0 \epsilon bC \quad (4)$$

and the fluorescence intensity is essentially homogeneous throughout the sample. When ϵbC increases, the fluorescence intensity is no longer homogeneous within the cell but is increasingly localized at the front surface of the cell, thus the fraction of fluorescence measured by the common detection system is smaller (inner-cell effect). To avoid this effect, instruments are designed to examine an area at the front of the cell rather than at the center.

Limitations and problems

Fluctuations in intensity of the exciting light source, variation in instrument sensitivity of the detection system, and other factors related to the instrument may cause distortion and variation of instrument's response. The shape of the excitation spectrum should be identical to that of the absorption spectrum of the molecule, but differences may be observed due to artifacts arising from variation in the light source intensity with wavelength and other causes. The shapes of emission and excitation spectra should be independent of the wavelength of the exciting radiation and the wavelength at which fluorescence is measured, but spectra are often wavelength dependent. In practice, however, the variations expected are likely to be small when the fixed excitation or emission wavelengths used are varied within a limited range of values, i.e., between 320 nm and 370 nm and 520 nm and 570 nm,

respectively. Light-scattering effects such as Rayleigh, Tyndall, and Raman effects may also influence the fluorescence spectra recorded to a varying extent.

Corrections accounting for instrument related artifacts and for scattering effects should be applied to fluorescence spectra, particularly when quantitative comparisons are made between spectra measured on a variety of spectrofluorimeters by different operators. However, comparisons of uncorrected fluorescence spectra may be allowed on a qualitative basis when they are recorded with the same instrument using the same experimental conditions.

Recently, the following problems have usually occurred with in situ detection and the investigation of soil organic components. First, the environmental matrix is complex, and thus, there is usually fluorescence interference caused by other complex components when detecting the target compounds. Second, the complex and changeable environment and other environmental factors may lead to the distortion and change of response by the instrument during field in-situ investigation and research. Stability of the instrument is required to be good because of fluctuation in the intensity of the excitation light source and the change in sensitivity of the detection system in the complex in-situ field measurement environment. In addition, the in-situ field investigation has certain requirements for the portability of the instrument, the adaptation of type of power supply, and the accuracy of data.

Molecular and environmental factors affecting fluorescence

Several molecular properties can affect fluorescence intensity and wavelength. In particular, atoms and ions of large atomic number, such as halogens and paramagnetic metals, and the decrease in fluorescence efficiency of a fluorescent organic compound. Increasing the extension of the π -electron system increases fluorescence and shifts emission wavelengths towards longer values. Steric hindrance generally reduces fluorescence, whereas increasing rigidity of the molecular structure increases it. Electron-withdrawing groups, such as carbonyls and carboxyls, largely reduce fluorescence intensity, whereas the opposite is true for electron-donating substances such as hydroxide, alkoxide, and amino groups. Further, all these groups shift fluorescence to longer wavelengths.

Several environmental factors can also affect the fluorescence response of a sample. For example, fluorescence efficiency is decreased by temperature increase, or by a heavy atom in the solvent molecule or co-existing solute molecules containing heavy atoms, whereas it is increased by increasing solvent viscosity. Fluorescence wavelength is usually shifted to longer values with an increase in solvent polarity. Fluorescence intensity is reduced by increasing solute-solvent interaction, e.g., by hydrogen bonding. Therefore, heavy-atom solvents and hydrogen-bonding solvents should be avoided when possible. However, compared with non-polar solvents, wavelength of the fluorescence signal in aqueous solution is usually red shifted. For example, shift of the fluorescence spectrum when measuring polarized fluorescence is caused by the organic solvent effect.

Light absorption by the solvent, large concentrations of a fluorescent solute, the presence in solution of a solute that absorbs radiation together with the fluorescent molecule, or a solute that absorbs the fluorescence of the sample molecule, give rise to the so-called 'inner-filter' effect that causes quenching of the fluorescence of the sample. This effect can be reduced by selecting a suitable excitation wavelength, diluting the sample, viewing fluorescence closer to the front surface of the cell, and or using the method of standard additions or a correction factor.

The pH of the medium affects fluorescence spectra of acidic or basic aromatic compounds by rapid proton transfer reactions that occur to various extents during the lifetime of the excited state. Dissolved oxygen in organic solvents is an important reason for the quenching of organic molecules' fluorescence. For example, the fluorescence of polycyclic aromatic hydrocarbons (PAHs) is easily and dynamically quenched by dissolved oxygen in nonpolar solvents, resulting in a decrease in the quantum yield of PAHs. This dynamic quenching mechanism results from inactivation of the fluorophore in the absence of photon emission through collision with molecular oxygen during its excited state lifetime. This effect can be easily eliminated by bubbling nitrogen through the solution or by the addition of chemical oxygen scavengers, such as sodium sulfite (Na_2SO_3), calcium chloride (CaCl_2) and zinc (Zn) reduction for oxygen scavenging. Metal ions, especially paramagnetic ions, are also fluorescence quenchers, even where they do not form complexes with the fluorescent molecule. Metal quenching can be reduced markedly, e.g., by diluting the sample.

Fluorescence quenching

The process of physical or chemical interaction between fluorescent substance molecules and solvent or solute molecules, which lead to the decrease in fluorescence intensity, is called fluorescence quenching. The quenching process competes with the luminescence process to shorten the lifetime of the excited state of the luminescent molecule. Quenching may occur in the interaction between the quencher and excited state molecules of the fluorescent substance. Interactions between the quencher and ground state molecules of the fluorescent substance may also occur. The former process is called dynamic quenching, and the latter is called static quenching. In the process of dynamic quenching, the excited state molecules of the fluorescent substance lose their excitation energy and return to the ground state by energy or charge transfer through collision with quencher molecules. The efficiency of dynamic quenching is controlled by the lifetime of the excited state molecules of the fluorescent substance and concentration of the quencher. In static quenching, the quencher reacts with fluorescent substance molecules in the ground state, and the resulting complex does not usually emit light. Dissociation of the complex may be slower, so that decay of the exciplex to the ground state by a nonradiative pathway is more efficient. On the other hand, formation of the ground-state complex also reduces fluorescence intensity of the fluorescent substance because of competition with ground-state molecules of the fluorescent substance to absorb the excitation light. In general, fluorescence quenching methods are more sensitive and selective than direct

fluorescence assays. The emission intensity of a fluorescent species can be quenched by its association and or adsorption to organic and inorganic substrates. Fluorescence quenching is widely used in the titration of humus and quantitative determination of metals in soil.

Fluorescence polarization

After the fluorescent substance has been irradiated with polarized light in a single plane, it can absorb light energy and jump into an excited state. When it returns to the ground-state, it releases energy and emits polarized fluorescence in a single plane. The degree of polarization of polarized fluorescence is inversely proportional to the speed at which the molecule rotates when the fluorescent substance is excited. Fluorescence polarization methods can be applied to study the conformation of a fluorescent organic molecule and the sorption or binding of a fluorescent organic pollutant to a variety of environmental sorbents. Any process that causes a decrease in the rate of rotation of a fluorophore, i.e., a variation in its size, during the lifetime of the excited state, including change of conformation, adsorption on suspended particulate matter, or binding to dissolved materials, decreases the polarization. For example, fluorescence polarization can be used to measure changes in the rate of rotation of humic acids in soil, thereby inferring configurational changes in them.

Fluorescence correlation spectroscopy

Fluorescence correlation spectroscopy (FCS) is a powerful tool for detecting molecular dynamics by analyzing the intensity of fluctuations emitted by biomolecules as they diffuse and diffuse in a focused light. Compared with other kinetic oriented methods, FCS has a much wider measurable time frame, from nanoseconds (ns) to seconds (s), with greater sensitivity achievable at the single-molecule level. At present, FCS technology is widely used to characterize kinetic mechanisms, concentration dependent diffusion, and clustering of cell surface receptors in solution. Specifically, the correlation FCS technique with laser light (excitation at 488, 514 nm or 671 nm) is focused on the sample using confocal optics, thus obtaining an open, illuminated volume element that is called 'confocal volume'. The volume obtained by conventional diffraction-limited FCS is about 200 nm in size laterally and 600 nm axially. In the absence of chemical processes, temporal variation in measured fluorescence intensity in the confocal volume can be attributed uniquely to translational diffusion of the fluorophore. These variations can be analyzed with an autocorrelation function that includes a parameter, τ_D , which is the diffusion time characteristic of the particle under study in the confocal volume. The system can be calibrated with a Rhodamine 6G (R6G) standard that has a known diffusion coefficient. The values of diffusion times for R6G and the sample in each condition are then determined from the best fitting autocorrelation function, and is finally used to calculate the diffusion coefficient of the fluorescent sample.

The emergence of new technologies has expanded the development of FCS technology. Fluorescence cross-correlation spectroscopy (FCCS) is a variant of FCS based on the principle that the fluctuations in fluorescence of two fluorescent objects are correlated temporally (Fig. 3). With a laser of a desired wavelength, fluorescence is collected by the objective, separated from the excitation light by the dichroic mirror and focused by the tube lens in a pinhole. The pinhole collects only the fluorescence from the focal plane that goes to the detectors for correlation analysis. In the case of FCCS, only if the two species interact and diffuse together in the confocal volume (II), will the correlation between both signals result in a cross-correlation (III). Otherwise, the count rate graphs do not match (IV) and the correlation function will be zero (V). Once the signal has been obtained, different software packages can be used to analyze the data to obtain the diffusion coefficient or diffusion time of the measurements, as well as the number of particles inside the confocal volume.

Two-dimensional correlation spectroscopy (2D-COS) is an emerging spectral analytical technique that has been extensively used to investigate the environmental interactions between dissolved organic matter (DOM) and metals that can be toxic. The 2D-COS analysis provides better spectral resolution, specifically by obtaining a series of dynamic spectra under external perturbations such as temperature, concentration and changes in pH prior to analysis. After the spectral data have been obtained, the order of the spectral peak changes during metal-DOM binding (i.e., synchronous and asynchronous maps) can be shown by two 2-D plots (synchronous and asynchronous plots) (Cui et al., 2020). The 2D-COS offers the capability to perform heterospectral correlation analysis between different spectroscopic probes, and it complements the power of each spectroscopic probe and overcomes the limitations of using a single spectroscopic technique alone. For example, Zhu et al. (2018) found that humic-like components played an important role in the binding process of DOM and atrazine under four types of land use. The 2D-COS further confirmed that humic-like components combined first with atrazine followed by the protein-like components.

Laser induced fluorescence spectroscopy

Laser induced fluorescence spectroscopy (LIFS) formed by the absorption of fluorescence emitted when a molecule of a fluorescent substance absorbs laser energy to transition from the ground state to an excited state and back to the ground state. The main advantages of laser excited samples over lamp induced fluorescence spectra are the brightness, strong direction, monochromaticity and coherence of the laser excited radiation, resulting in a better signal-to-noise ratio and greater excitation selectivity, reducing factors that interfere with the fluorescence signal. Moreover, the use of a high- power laser source improves the sensitivity and can be applied to the determination of whole soil organic components (Tadini et al., 2021).

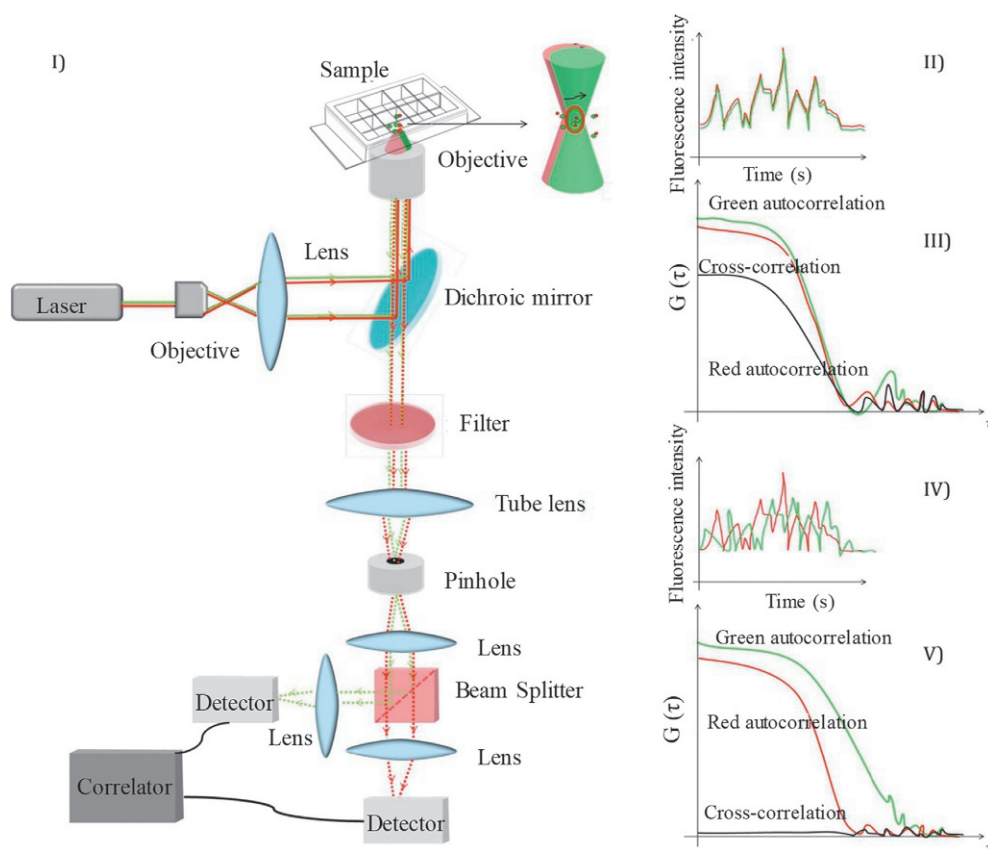


Fig. 3 Schematic diagram of a fluorescence correlation spectroscopy device and principle. Adapted from Martínez-Moro M, Di Silvio D, Moya SE (2019) Fluorescence correlation spectroscopy as a tool for the study of the intracellular dynamics and biological fate of protein corona. *Biophysical Chemistry* **253**: 106218.

Given that LIFS enables a faster evaluation of the quality and stability of soil organic matter (SOM), this technique has the advantage that it can be applied to whole soil samples without the need for laborious pretreatments. Furthermore, LIFS can be used to investigate in situ the degree of humification of intact SOM in untreated solid soil samples, which is feasible for the following reasons. First, excitation by near UV blue radiation is more likely to resonate with rigid systems containing condensed and/or substituted aromatic rings, which are conjugated to large π -electron systems that can emit fluorescent signals even when the sample is in a solid state. Second, a novel humification index (H_{LIFS}), corresponding to the ratio of the area under the broad emission-mode LIFS and the corresponding total soil C concentration measured by a conventional method, was proposed to evaluate the humification status of SOM assuming that the concentration of these systems increases with progressing humification.

X-ray fluorescence spectroscopy

In general, X-ray fluorescence spectroscopy is a versatile analytical technique that provides information on both qualitative and quantitative chemical elements. Electrons exist in atoms at discrete energy levels that are unique for each element. X-ray fluorescence occurs when a sample is illuminated by X-ray resulting in the ionization of its elements by the ejection of inner shell electrons. Subsequently, higher energy electrons transition to lower energy states and fill the voids in inner electron shells yielding either X-ray fluorescence or auger electrons. Auger electrons are generated when the X-ray fluorescence of an atom is re-absorbed by the outer electrons of the emitting element itself, and this effect dominates for lighter elements with lower atomic numbers. Due to the conservation of energy, electrons undergoing transition emit X-ray fluorescence that is unique to the atom. The X-ray fluorescence measurements are based on the entire contents of the analyte and volume of the matrix dependent fluorescence regression.

Time-resolved fluorescence spectroscopy

Time correlated single photon counting (TCSPC) was developed in the 1960s by Bollinger et al. The fluorescence intensity of molecules is directly proportional to the number of molecules in the excited state, and the decay rate of fluorescence intensity

reflects the residence time of molecules in the excited state. When molecules in the excited state have undergone processes such as chemical reactions or intermolecular energy transfer, the decay rate of the fluorescence intensity of molecules will change. Therefore, the study of fluorescence intensity with time can provide information about the reaction kinetics such as the rate at which the reactant molecules react chemically in the excited state, the interaction between the reactant and quencher molecules, and the effect of solvent on the reaction. Since the advent of time-dependent single photon techniques, TRF spectroscopy has been widely applied to the mechanistic study of biochemical reactions in solution. Because of its excellent temporal resolution, good measurement accuracy and sensitivity, large dynamic range, and easy storage and processing of data, this technique has been widely used in many fields. Central to time correlated single photon counting is photon counting technology, which can be used with the detection of weak signals. Recently, many TRF spectroscopic methods, such as laser-induced nanosecond TRF spectroscopy, have been derived on this basis. It can be used to study the interaction between SOM and organic pollutants or metal ions.

Conclusion

Fluorescence spectroscopy, including the emission, excitation, synchronous scan modes, total luminescence and 3D fluorescence, are noninvasive, nondestructive, and highly sensitive techniques that require minimal sample pretreatment. In addition, the unique properties of laser light sources have made LIFS a widely used method in SOM research. Therefore, fluorescence spectroscopy has broad application potential in the study of organic and inorganic soil components. For example, the diversity of inherent fluorescence structures in soil organic components, especially in their humified fractions, such as humic substances, mainly humic acids and fulvic acids, has provided invaluable information on their structural and functional chemistry and reactivity as a function of relevant environmental factors such as pH and ionic strength.

References

- Alfonso-Garcia A, Bec J, Weyers B, et al. (2021) Mesoscopic fluorescence lifetime imaging: Fundamental principles, clinical applications and future directions. *Journal of Biophotonics* 14: e202000472.
- Bitton A, Sambrano J, Valentino S, and Houston JP (2021) A review of new high-throughput methods designed for fluorescence lifetime sensing from cells and tissues. *Frontiers in Physics* 9: 13.
- Cui HY, Zhang SB, Zhao MY, Zhao Y, and Wei ZM (2020) Parallel fraction analysis combined with two-dimensional correlation spectroscopy reveal the characteristics of mercury-composting-derived dissolved organic matter interactions. *Journal of Hazardous Materials* 384: 10.
- Kim KK, Kim J, Woo KC, and Kim SK (2021) S1-State Decay Dynamics of Benzenediols (Catechol, Resorcinol, and Hydroquinone) and Their 1:1 Water Clusters. *The Journal of Physical Chemistry A* 125(35): 7655–7661.
- Qin XQ, Yao B, Jin L, et al. (2020) Characterizing soil dissolved organic matter in typical soils from china using fluorescence EEM-PARAFAC and UV-visible absorption. *Aquatic Geochemistry* 26: 71–88.
- Ravansari R, Wilson S, and Tighe M (2020) Portable X-ray fluorescence for environmental assessment of soils: Not just a point and shoot method. *Environment International* 134: 105250.
- Tadini AM, Xavier AAP, Milori D, et al. (2021) Evaluation of soil organic matter from integrated production systems using laser-induced fluorescence spectroscopy. *Soil & Tillage Research* 211: 9.
- Zhu LJ, Zhao Y, Yan-Ni Chen YN, et al. (2018) Characterization of atrazine binding to dissolved organic matter of soil under different types of land use. *Ecotoxicology and Environmental Safety* 147: 1065–1072.

Fluorescence spectroscopy: Part II practical applications

Yong Zhang^{a,b}, Chaoxian Wei^{a,b}, and Shuai Guo^{a,b}, ^aState Key Laboratory of Marine Environmental Science of China, Xiamen University, Xiamen, PR China; ^bCollege of Environment and Ecology, Xiamen University, Xiamen, PR China

© 2023 Elsevier Ltd. All rights reserved.

Introduction	552
Soil organic components	553
Fluorescence spectra	553
Nature of fluorophores	553
The effects of molecular and environmental factors on SOM	554
Fluorescent properties of soil dissolved organic components	556
Fluorescence lifetime of soil dissolved organic components	556
Conformational studies by polarization	556
Diffusion coefficients determined by fluorescence correlation spectroscopy	556
Fluorescence quenching of soil organic components by metal ions	557
Molecular and mechanistic aspects	557
Quantitative aspects	558
Sorption or binding of fluorescent organic pollutants to soil organic components	559
Fluorescence polarization	559
Fluorescence quenching	559
Soil mineral components	561
Structural quenching	562
Adsorption of organic chemicals to mineral surfaces	563
Surface acidity	563
Polarity and charge adsorption of mineral surfaces	564
Molecular distribution, conformation, aggregation and mobility	565
Biochemical transformation of clay surfaces	565
Ligand-promoted mineral dissolution	566
Conclusion	566
References	566

Abstract

Soil organic matter comprises typical light-absorbing or emitting structures and functional groups, and the source, optical properties and rotary motion of its molecules can be determined by the fluorescence method. A large variety of fluorescence technologies is applied to investigate the interaction process and related mechanism among soil organic components and chemicals from quantitative and molecular aspects. Static and dynamic fluorescence spectroscopy of organic probes sorbed on soil clay minerals can provide important and unique insights into the specific interface and surface behavior of soil mineral components, such as their molecular distribution, conformation, aggregation and mobility.

Key points

- The practical applications of fluorescence spectroscopy for soil organic matter and soil clay minerals are introduced.
- The interaction process and related mechanism among soil organic components and chemicals are investigated by fluorescence technologies.
- Specific interface and surface behavior of soil mineral components are explored by static and dynamic fluorescence spectroscopy.

Introduction

Fluorescence occurs when a singlet electron in the excited state is coupled with an electron in the ground state by opposing spins. Part I of this topic in the Encyclopedia gives the principles of the method. In this chapter, the practical applications of fluorescence spectroscopy for soil organic matter and soil clay minerals are introduced. For soil organic components and clay minerals, fluorescence spectroscopy can be applied to determine their structures, properties and the variation induced by environmental factors. To explore the interaction among chemicals and soil components, fluorescence quenching, polarization, correlation spectroscopy, and the excitation–emission matrix (EEM) and time resolved fluorescence are discussed.

Soil organic components

Fluorescence spectra

Soil organic matter (SOM) generally includes dissolved organic matter (DOM) and humic substances (HS). Its components are grouped according to their solubility into humic acids (HAs), fulvic acids (FAs) and humin. Humic acids are insoluble at pH < 2, whereas FA is soluble under both acidic and alkaline conditions. Humin is insoluble at any pH values. Soil organic matter comprises typical light-absorbing or emitting structures and functional groups which are susceptible to variations in ambient biochemical and geochemical conditions. Absorption and emission occur mostly from SOM with the lowest vibrational energy when the electronic energy levels of the fluorophores endure changes, which can be captured by optical spectra. Fluorescence spectroscopy has been widely exploited to explore the chemical composition and biogeochemical cycling of SOM because it is fast and sensitive. Therefore, it is an efficient tool to trace the sources, features and environmental behavior of SOM.

Steady-state and time resolved fluorescence (TRF) are the two main types of fluorescence method to measure HS. For example, steady-state fluorescence, the most common type, is performed with excitation, emission and synchronous fluorescence spectra and EEM fluorescence (Fig. 1) (Chen and Yu, 2021).

The SOM consists of various fluorescent structures, therefore, their fluorescence spectra contain diverse and interwoven fluorescence information. In the last two decades, many approaches have been developed to explicate the fluorescence data to decode the structural information of SOM. The EEM and synchronous fluorescence spectra are the main prevailing fluorescence approaches for investigating SOM. Synchronous fluorescence spectra are generally coupled with two-dimensional correlation spectroscopy (2D-COS) rather than EEM to simplify complex spectra with multiple overlapping peaks and improve spectral resolution. The EEM can be processed separately with sophisticated applications, such as fluorescence regional integration and parallel factor analysis, for a thorough separation of components and correction of data.

Steady-state fluorescence measurements are frequently applied to explore various spatial and temporal scales of DOM concentration, composition, and biogeochemical dynamics, and the time-resolved measurements are gaining attention because of their unique advantage. In general, most of the molecular information from fluorescence is lost by the time averaging process. For example, the shape of the anisotropic decay includes information on the conformation of the macromolecule and its flexibility, but part of that information is lost as the intensity decays over the decay time. Typically, macromolecules can have multiple conformations and the decay time of a binding probe may depend on the conformation. The intensity of decay can indicate different decay times as well as the existence of more than one conformational state. The steady-state intensity can only indicate a mean intensity based on a weighted mean of the multiple decay times. Furthermore, the intensity of decay could reveal how acceptors are distributed in space around the donors in the presence of energy transfer. Fluorescence quenching also expresses diffusion or complex formation with the ground-state fluorophores by time-resolved measurements. Thus, time-resolved measurement is essential for revealing molecular information, including the intermolecular clustering, conformational rearrangement of adsorbed polymers at the solid-liquid interface and changes in the advanced structure of proteins.

Nature of fluorophores

The fluorophores of SOM comprise assorted chemical groups, including carboxyl, phenol, quinonyl, aldehyde, ester, ketone, hydroxyl, amino, glycosyl and other functional groups. Because of the molecular complexity and chemical heterogeneity, fluorescence spectra of SOM probably represent the sum of the spectra of several different fluorophores. There is an urgent need for mathematical approaches that will separate and reasonably attribute each component of SOM due to its integral role in natural and engineered systems.

The 2D-COS method, first proposed by Noda in 1986, has become a routine analytical technique in spectroscopic studies. It has two main advantages, including enhancing spectral resolution and discerning relative directions and specific order of structural variations. The subtle key information can be evaluated from a variety of probes under various forms of external perturbations.

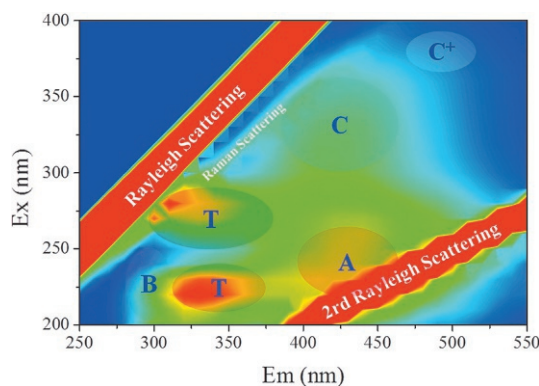


Fig. 1 Representative Ex/Em peaks in natural organic matter. Background EEM spectrum was obtained from a sediment sample. Adapted from Chen W and Yu HQ (2021) Advances in the characterization and monitoring of natural organic matter using spectroscopic approaches. *Water Research* **190**: 116759.

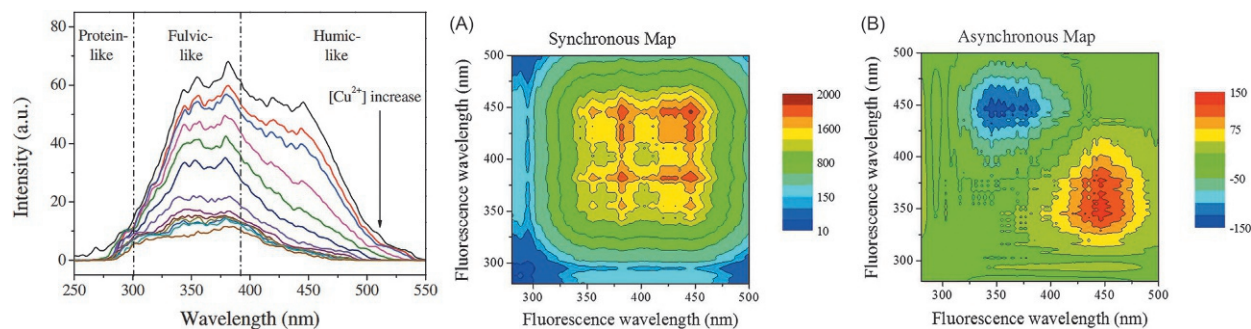


Fig. 2 Changes in the synchronous fluorescence spectra of 10 mg L^{-1} HA at pH 7.0 with the addition of copper at concentrations of 0, 1, 2, 5, 10, 20, 40, 60, 80, 100, 120 and $160 \times 10^{-6} \text{ mol L}^{-1}$ (a). Synchronous (b) and asynchronous (c) 2-D correlation maps generated from the synchronous fluorescence spectra of HA with increasing copper ion concentration in the 280–500 nm region. Adapted from Chen W, Habibul N, Liu XY, Sheng GP and Yu HQ (2015) FTIR and Synchronous fluorescence heterospectral two-dimensional correlation analyses on the binding characteristics of copper onto dissolved organic matter. *Environmental Science & Technology* **49** (4): 2052–2058.

In Fig. 2, HAs are regarded as a typical HS, and copper concentration was used as the external perturbation (Chen et al., 2015). The changes in the synchronous fluorescence spectra of HAs with the addition of copper, synchronous and asynchronous 2-D correlation maps generated from the synchronous fluorescence spectra of HAs with increasing copper ion were revealed in Fig. 2a–c, respectively. The 2D-COS, synchronous fluorescence and infrared absorption spectroscopy have been successfully combined to interpret the copper–DOM interaction mechanisms. It could reveal a sequence of heterogeneous binding sites in HAs and their subsequent subtle changes within the molecular interactions. The results showed that fluorescence quenching of fluorescent humic-like moieties occurred with a vibrational change of the related functionalities (i.e., phenolic and aryl carboxylic groups), which further induced the fluorescence quenching of fulvic-like fractions. Furthermore, small amounts of amide and aliphatic groups participated in the copper binding after fluorescence of the protein-like fraction decreased.

The EEM coupled with fluorescence regional integration (EEM-FRI) has been authenticated as a suitable method to integrate the volumes beneath defined EEM regions, making a semi-quantitative analysis of the configuration and heterogeneity of SOM. Fluorophores in SOM can be separated into five regions by FRI. Table 1 gives the excitation and emission wavelengths characteristic of each region (Chen et al., 2003). The EEM-FRI analysis can be used to trace the fate of SOM. But the application of EEM-FRI to evaluate SOM components is still controversial because the component concentration is not always linearly related to fluorescence intensity. In addition, interference might result in fluorescence quenching or enhancement of specific fluorophores in SOM, which might not be revealed by FRI.

The EEM combined with parallel factor (EEM-PARAFAC) analysis can convolute complex EEM fluorescence spectrum into independent fluorescent components that represent groups of similar fluorophores, which could help changes in SOM components to be comprehensively tracked. The EEM-PARAFAC analysis is non-destructive, but is valid only when the EEM data are trilinear (Table 2) (Ishii and Boyer, 2012). In addition, the EEM-PARAFAC technique provides only the information on finite components of SOM. Compared with EEM-PARAFAC analysis, EEM-FRI is a quantitative technique associated with all wavelength-dependent paired fluorescence intensity data in an EEM.

The effects of molecular and environmental factors on SOM

Changes in soil pH resulting from soil acidification can affect both the concentration and chemical characteristics of SOM. The extensive interactions with most inorganic and organic pollutants and SOM depend strongly on the protonation and complexation properties of SOM. Biological assimilation and degradation of organic carbon play a crucial role in regulating both the production

Table 1 Excitation and emission wavelength ranges of the five integrated regions identified by FRI.

FRI region	Excitation (nm)	Emission (nm)	Source
R1	200–250	250–330	Tyrosine-like protein
R2	200–250	330–350	Tryptophan-like protein
R3	200–250	350–500	Fulvic acid-like
R4	250–280	250–380	Microbial-like
R5	250–400	380–500	Humic-like

Adapted from Chen W, Westerhoff P, Leenheer JA and Booksh K (2003) Fluorescence excitation-emission matrix regional integration to quantify spectra for dissolved organic matter. *Environmental Science & Technology* **37** (24): 5701–5710.

Table 2 Traditional peak classifications of components and EEM location information.

Component	Traditional classification	Description based on EEM location
Component 1	A (UVC humic-like)	Absorbs light primarily in the UVC region Terrestrial UV light is lacking in UVC Expected to be photoresistant due to limited exposure to UVC light Short excitation wavelengths
Component 2	A + C (UVC humic-like +UVA humic-like)	Expected to consist of small molecular size compounds Absorbs light in the UVC and UVA region UVA light constitutes the majority of terrestrial UV radiation Expected to be photodegraded by UVA light Long excitation and emission wavelengths
Component 3	A + M (UVC humic-like +UVA marine humic-like)	Expected to consist of large molecular size, hydrophobic compounds Absorbs light in the UVC, UVB, and UVA region Absorption band extends less into the UVA light region than that of Component 2 Expected to be photodegraded by UVA light, but to a lesser extent than Component 2 Intermediate excitation and emission wavelengths Expected to consist of compounds with molecular sizes that fall between Component 1 and Component 2

Adapted from Ishii SKL and Boyer TH (2012) Behavior of Reoccurring PARAFAC Components in fluorescent dissolved organic matter in natural and engineered systems: A critical review. *Environmental Science & Technology* **46** (4): 2006–2017.

of greenhouse gases and accumulation and stabilization of SOM (Fig. 3). The DOM originates primarily from the decomposition of SOM that had accumulated from vegetation, the addition of biological materials (e.g., biosolids and livestock manures), the release of root exudates and the lysis of microorganisms (Bolan et al., 2011). Microbial degradation of SOM, followed by desorption of organic substances from the soil matrix are regarded as the most important processes that release DOM, and microbial assimilation of readily available carbon in DOM results in the ultimate degradation of DOM to carbon dioxide.

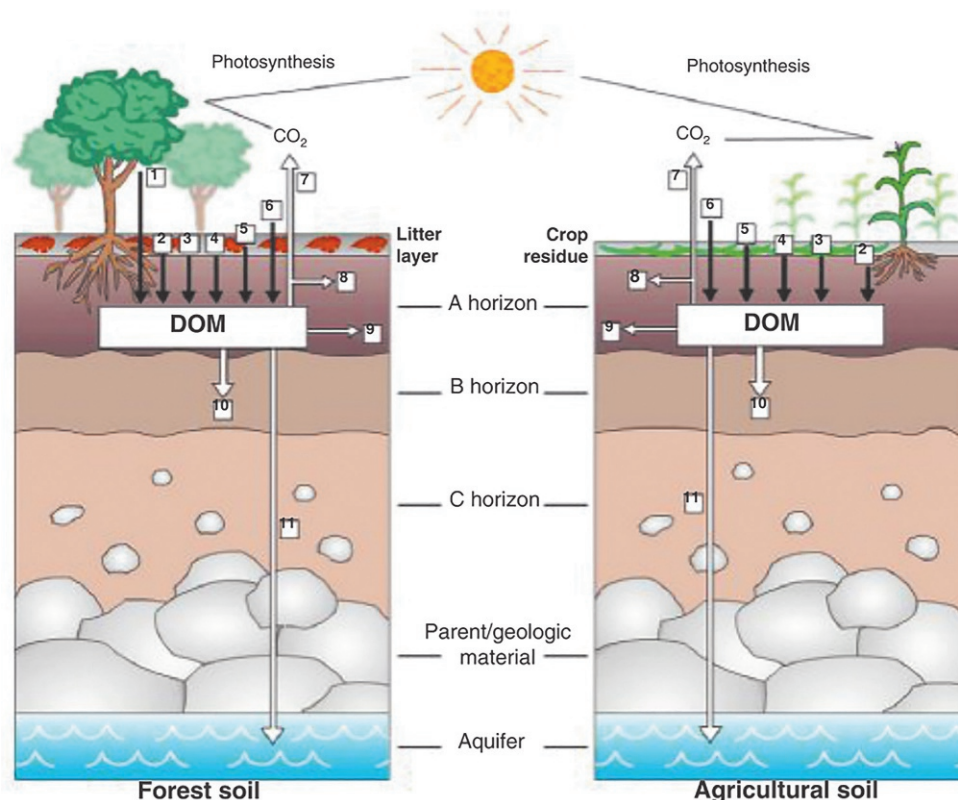


Fig. 3 Pathways of inputs and outputs of DOM in forest and agricultural soils. Inputs: 1, throughfall and stemflow; 2, root exudates; 3, microbial lysis; 4, humification; 5, litter or crop residue decomposition; 6, organic amendments; 7, microbial degradation; 8, microbial assimilation; 9, lateral flow; 10, sorption; 11, leaching. Adapted from Bolan NS, Adriano DC, Kunhikrishnan A, James T, McDowell R and Senesi N (2011) Dissolved organic matter: Biogeochemistry, dynamics, and environmental significance in soils. *Advances in Agronomy* **110**: 1–75.

Fluorescent properties of soil dissolved organic components

Laboratory measurements and characterization of DOM for source identification and dynamic tracking have received much attention over the past two decades. The in-situ techniques providing optical properties to assess the quality of dissolved organic carbon (DOC) are sensitive, simple to use, precise, and non-invasive compared to chemical analyses. In general, fluorescent DOM signatures interpret specific spectral characteristics using a quantitative measurement of an ever-expanding number of optical parameters. These measurements are surrogates for the reactivity and composition of DOM. The intrinsic fluorescence optical parameters of DOM generally consisted of fluorescence index, humification index, freshness index, biological index and etc.

Fluorescence lifetime of soil dissolved organic components

Fluorescence lifetime can distinguish between static and dynamic quenching of fluorophores. For static quenching, the ratio, τ_0/τ , of fluorophore fluorescence lifetime in the absence (τ_0) and presence (τ) of a quencher is one. In contrast, the fluorescence lifetime of fluorophores for dynamic quenching depends on quencher concentrations; in this case, τ_0/τ is equal to the ratio, F_0/F , of fluorescence intensity in the absence (F_0) and presence (F) of a quencher (Table 3) (Fu et al., 2016).

Conformational studies by polarization

Fluorescence anisotropy was used to explore conformational variations in DOM and molecular interactions between it and environmental contaminants by evaluating the rotary motion of DOM molecules. Fluorescence intensity increased with rising HA concentrations between 20 mg L⁻¹ and 100 mg L⁻¹, while fluorescence polarization decreased. However, fluorescence intensity decreased and fluorescence polarization increased with HA concentrations >100 mg L⁻¹. This indicated self-absorption and collision related to fluorescence quenching and energy transfer. The HA was coiled in shape at concentrations of 20–100 mg L⁻¹ because of the variation in fluorescence polarization. But HA aggregated under higher concentrations (above 100 mg L⁻¹), resulting in prolonged planar conformation from strong repulsion between molecules. With increasing ionic strength of the solution, peak B of the HA decreased considerably, suggesting that conformational coiling occurred at Peak B (Fig. 4a) (Mei et al., 2009).

Fig. 4b shows that fluorescence polarization varied with pH. Neutralization of the molecular charge of the macromolecules of the HA resulted in agglomeration under acidic conditions. Nevertheless, for an individual molecule, the shortened molecular distance with agglomeration ultimately results in mutual repulsion between molecules and coiled molecular conformation. The increasing fluorescence polarization of HAs at lower pH (≤ 4) suggested that the formation of elongated linear conformation occurred in the HA macromolecules. As a result of protonation of carboxyl, molecular aggregation and repulsion led to elongated linear conformation of HAs.

Diffusion coefficients determined by fluorescence correlation spectroscopy

The properties of DOM are complex, heterogeneous and dynamic, and thus, it is important to reveal their effects on the spatial distribution of the dynamic information, such as concentrations, mobility, equilibrium, rate constants, and interactions related to the molecular size and weight of DOM. Fluorescence correlation spectroscopy (FCS) is one of few ultrasensitive and noninvasive single-molecule detection techniques for in-situ quantitative analysis. The average diffusion coefficients for Suwannee River FAs and Suwannee River HAs were determined by FCS in the range of $2\text{--}3 \times 10^{-10} \text{ m}^2 \text{ s}^{-1}$ under different concentrations, hydration times, pH, and ionic strength (Ca^{2+} (calcium)) conditions, which corresponded to hydrodynamic diameters of approximately 1.5–2.1 nm. With a decrease of pH from 10 to 3, the slight but significant decrease in diffusion coefficients suggested that a little

Table 3 The average fluorescence lifetime of fluorophores in the DOM and roxarsone—DOM systems.

Peaks	Position Ex (nm)/Em (nm)	The average fluorescence lifetime (ns)		τ_0/τ	F_0/F
		τ^a	τ_0^b		
Peak A	235/427	2.85	2.79	1.02	1.61
Peak I	288	1.72	1.83	0.94	1.42
Peak II	339	2.05	2.15	0.95	1.58

^aAverage fluorescence lifetime for DOM.

^bAverage fluorescence lifetime for roxarsone-DOM system.

Adapted from Fu QL, He JZ, Blaney L and Zhou DM (2016) Roxarsone binding to soil-derived dissolved organic matter: Insights from multi-spectroscopic techniques. *Chemosphere* **155**: 225–233.

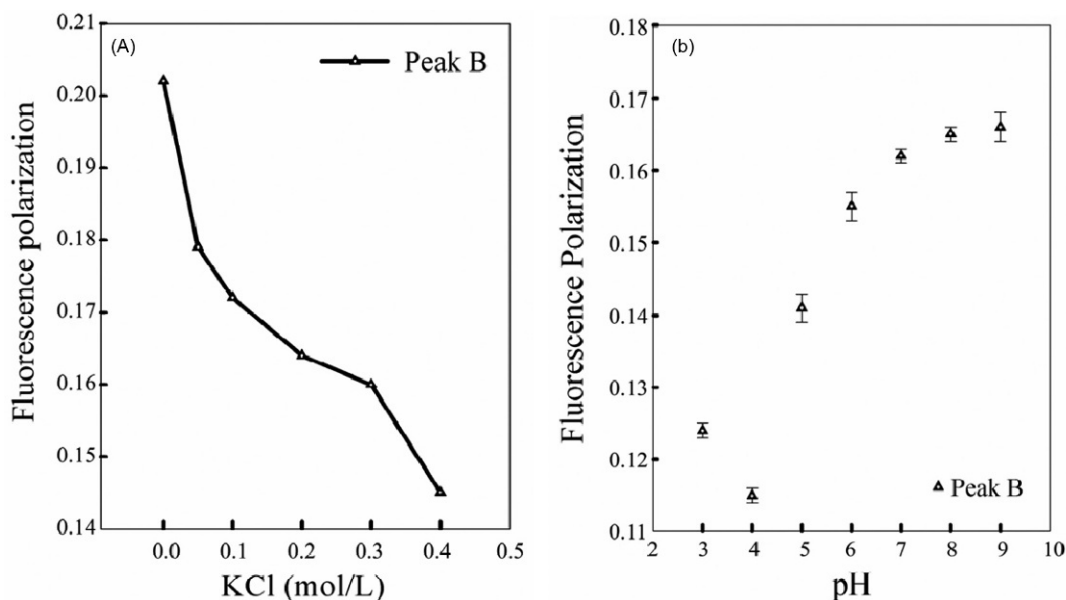


Fig. 4 Effect of ionic strength on decreasing fluorescence polarization of the HA Peak B (a). Effects of pH on the fluorescence polarization of changes in the HA Peak B (b). Adapted from Mei Y, Wang L and Wu F (2009) Effects of water chemistry and concentrations of dissolved organic matter on its fluorescence characteristics and molecular conformation. *Chinese Journal of Geochemistry* 28(4): 413–420.

aggregation occurs with the formation of dimers and trimers. Reduction in the diffusion coefficient was either small or insignificant with ionic strength changes up to 0.1 mol L^{-1} . There was no effect on FA and HA concentrations ($1\text{--}50 \text{ mg L}^{-1}$) or hydration time ($1\text{--}14$ days).

Fluorescence quenching of soil organic components by metal ions

Molecular and mechanistic aspects

Soil organic matter interacts with other chemicals, such as ‘heavy metal’ cations, in the natural environment to form complexes. Paramagnetic transition metal ions, such as Cu^{2+} (copper), Fe^{3+} (iron), Fe^{2+} , Ni^{2+} (nickel), Cr^{3+} (chromium) and VO_2^{2+} (vanadium), can effectively quench the fluorescence of HS by enhancing the rate of some nonradiative processes that compete with fluorescence, such as intersystem crossing. Cations such as Cd^{2+} (cadmium), K^+ (potassium), Na^+ (sodium), Ca^{2+} , and Ba^{2+} (barium) show no quenching effect. Metal ions and HS can form two main types of complexes: (1) inner-sphere and (2) outer-sphere. The quenching ability of Cu^{2+} on FAs is excellent, because the FAs formed strong inner-sphere complexes. This phenomenon binds several weak sites causing conformational changes that make available additional internal binding sites of FAs. On the contrary, outer-sphere complexes are formed by Mn^{2+} (manganese), and FAs that are not involved in weakly acidic phenolic sites where the ions are further away from the fluorophore.

The shifts in the optimal fluorescence emission or excitation peak often occur with interaction between HS and paramagnetic metal ions. The molecular nature or quality of soil organic components can influence their binding with cations. The fluorescence quenching of FAs by Mn^{2+} decreased with the increase in FA molecular weight. Synchronous fluorescence analysis of three molecular weight (MW) fractions of FAs shows that affinity of the large-MW fraction to Al^{3+} (aluminium) and UO_2^{2+} (uranyl) is larger than that of the small-MW fraction. The complexing ability of SOM and metal ions is highly dependent on pH values. The conditional stability constants (K_c) between lanthanides (14 elements) and HAs increased with increasing pH. Fluorescence quenching of HAs and FAs from various origins by Cu^{2+} , Co^{2+} (cobalt), Pb^{2+} (plumbum), Ni^{2+} and Mn^{2+} increases as pH increases.

Synchronous-scan fluorescence spectra, EEM-PARAFAC analysis and TRF have been used successfully to study the fluorescence quenching of FAs by metal ions. Principal component analysis of synchronous quenching spectra obtained by the difference between before-and-after additions of Cu^{2+} , Pb^{2+} , Ni^{2+} , Co^{2+} , Mn^{2+} , Mg^{2+} (magnesium), and Al^{3+} at various total concentrations indicates that multiple fluorescent binding sites of FAs are involved in divalent metal complexation in solutions at pH 7.5, but not pH 5.0. The Al^{3+} and Cu^{2+} ions compete for FA binding sites and displace Mg^{2+} bound to FAs. Synchronous fluorescence spectra usually have overlapping fluorophores, therefore, a method with better resolution is required to obtain accurate binding parameters. The EEM-PARAFAC analysis can be a reliable tool for determining binding parameters between metal ions and individual components within SOM.

The fluorescence score of all components in SOM decreased with Pb^{2+} more significantly than with Cd^{2+} , indicating the stronger interaction of SOM with Pb^{2+} than Cd^{2+} . It is possible to study organic matter–metal interaction with TRF because different chemicals have different fluorescence lifetimes, especially in complex samples like HAs. The results of TRF for the interaction between Cu^{2+} and HAs have shown that the average lifetime of HAs appears to be independent of Cu^{2+} addition, suggesting that the main interaction mechanism is static quenching with a non-fluorescent ground-state complex formation. Dynamic quenching, caused by the collision of a quencher that occurs during the lifetime of the excited state, affects the fluorescence lifetime of SOM. Static quenching occurs when a ground-state non-fluorescent complex is formed, and the fluorescence lifetime of a fluorescent ligand and complex are unaffected.

Quantitative aspects

Fluorescence quenching titrations of a soil FA with Cu^{2+} showed a complexing capacity (C_L) value of FA $\sim 20 \text{ mol L}^{-1} \text{ Cu}^{2+}$ at pH 6.0 and 7.0, and a K_c of the complex FA– Cu^{2+} two to three times greater at pH 7.0 than at pH 5.0 and 6.0. Average C_L and conditional stability constant (K_b) values determined by fluorescence quenching are of the same order of magnitude as those determined by polarographic titration. At pH 7.0 and 6.0 the final fluorescence intensity is approximately 20% of the initial value for FAs, and is 40% at pH 5.0, indicating that most FA fluorophores are directly involved in Cu^{2+} complexation.

The integral standard volume of the total fluorescence area ($\Phi_{T,n}$, $\text{au}\cdot\text{nm}^2$) values caused by different metals followed the sequence $\text{Cu}^{2+} > \text{Fe}^{3+} > \text{Cd}^{2+} > \text{Cr}^{3+} > \text{Zn}^{2+}$ (Fig. 5), indicating that they had different binding ability or reactivity with fluorescence SOM (Liu and Lang, 2021). The $\Phi_{T,n}$ of fluorescence DOM induced by Cr^{3+} and Zn^{2+} were much lower than those by Cu^{2+} and Fe^{3+} , indicating that fluorescent DOM has low binding ability to Cr^{3+} and Zn^{2+} . In addition, different fluorescence regions have different responses to metals (Fig. 5b). Specifically, about 33% humic-like and 21% aromatic protein II are involved in the binding between Fe^{3+} and DOM, indicating that humic- and protein-like substances play a key role in DOM– Fe^{3+} binding. In the binding of DOM and Cu^{2+} , humic-, fulvic-, and soluble microbial by product-like substances all exhibit a strong reactivity, accounting for about 35%, 24% and 21% of Cu^{2+} binding sites, respectively. However, Cr^{3+} showed different binding preferences, with fulvic-like substances contributing up to 35% of binding sites. In addition, humic-like substances are also important in Cr^{3+} binding, accounting for 26% of binding sites. In contrast, Cd^{2+} and Zn^{2+} mainly bind to protein-like components of DOM; protein-like I and II components accounted for 22% and 45% in DOM– Cd^{2+} binding and 38% and 22% in the DOM– Zn^{2+} binding respectively.

Synchronous-scan fluorescence spectra of FAs with increasing amounts of added Be^{2+} (beryllium) at pH ranged from 4.0 to 7.0, indicating a fluorescent component unaffected by Be^{2+} and another component resulting from the formation of relatively strong fluorescent complexes with Be^{2+} by different FA binding sites. The C_L values calculated are smaller at pH 4.0 than at higher pH values because of protonation of some FA binding sites that become unavailable to complexation at low pH. The K_b values were also increased from pH 4.0 to 6.0, but decreased slightly at pH 7.0. A similar analysis was applied to complexation of Cu^{2+} , UO_2^{2+} and Al^{3+} by FAs at various pH and concentrations. For Al^{3+} , one component corresponded to the formation of a fluorescent complex and another to a non-fluorescent complex. The K_b values indicated that the stability of FA– Al^{3+} complexes decrease markedly at pH < 4.5.

With increasing Al^{3+} concentration, fluorescence polarization increases for the smallest MW fraction of a soil FA and decreases for the largest MW fractions, suggesting that the former FA fraction does not change, whereas the latter fractions change conformation on complexation with Al^{3+} ions. The K_b values for Al^{3+} –FA complexes can be calculated with anisotropy values obtained from polarization data. Time-resolved laser fluorescence spectroscopy has also been used to calculate K_b values of complexes formed by soil HAs with trivalent metal ions such as Cm^{3+} (curium) and Eu^{3+} (europium).

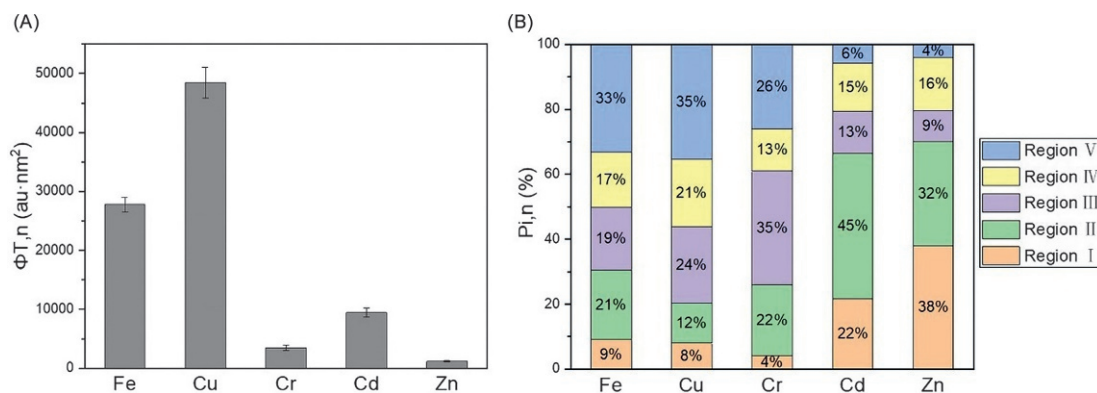


Fig. 5 Integrated standard volume of the total fluorescence area ($\text{au}\cdot\text{nm}^2$) of ∇EEM (a), and the percentage contribution of different regions in ∇EEM (b). The concentration of all the metal ions is $10 \mu\text{mol L}^{-1}$ and that of dissolved organic carbon is 8.52 mg L^{-1} . Adapted from Liu MX and Lang YC (2021) Differences in the spectroscopic characteristics of wetland dissolved organic matter binding with Fe^{3+} , Cu^{2+} , Cd^{2+} , Cr^{3+} and Zn^{2+} . *Science of the Total Environment* **800**: 149476.

In general, fluorescence spectroscopy based on metal–HS complexation fluorescence quenching is more sensitive than anodic stripping voltammetry (ASV) and ion-selective electrode potentiometry (ISE) is sufficient for application to unmodified natural organic ligands without pre-concentration. Although several limitations and assumptions are inherent to fluorescence data analysis, this method has several advantages because it is relatively rapid and less sample is needed. Moreover, there is no need to add supporting electrolyte or buffer or adsorbing material to the sample, neither CO₂ nor O₂ interferes with the measurement, and its sensitivity is adequate. Thus, the errors associated with the separation step in most speciation methods are avoided. It also allows direct measurement of the complexing capacity of the ligand by determining the concentration of free ligands to distinguish between free and bound ligands.

Sorption or binding of fluorescent organic pollutants to soil organic components

Both fluorescence polarization and fluorescence quenching methods can be applied to molecular and quantitative studies of the binding or sorption of polycyclic aromatic hydrocarbons (PAHs) to soil organic components, and especially to HS. The major advantage of these methods is that no separation step is required, which eliminates possible errors due to incomplete separation of the free or dissolved from bound or sorbed pollutant with potential improvements in reproducibility. These methods are also relatively rapid and convenient, as well as offering good precision and inherent sensitivity.

Fluorescence polarization

Fluorescence polarization is preferentially applied where the pollutant molecule is an efficient fluorophore with respect to the adsorbent, i.e., the HS molecule. Alternatively, fluorescence quenching can be applied to systems where fluorescence of the pollutant is affected by that of HS. The fluorescence of perylene (Pery) in glycerol-water (75/25, w/w) solutions at pH values of 2.0, 5.0, and 11.0 becomes more polarized as a function of added FAs, which indicates that Pery binds to FAs, forming a larger species that rotates more slowly in solution than does Pery alone. Fluorescence polarization can be extended to non-fluorescent compounds by measuring either changes in intrinsic fluorescence polarization of the adsorbent HS molecule or with competitive binding experiments using an adequate fluorophore probe. In the first case, the polarization changes are likely to be small and difficult to measure with precision, thus good-quality instrumentation is required, whereas in the second case, irreversible binding and slow kinetics can complicate the experiment.

Fluorescence quenching

Fluorescence quenching is a sensitive and fast method to quantify the interactions between PAHs and a quencher, such as HS. The PAHs possess strong intrinsic fluorescence due to their conjugated structure, but its fluorescence intensity would be quenched when they are bound to HS. Thus, the binding characteristics of PAHs and HS can be obtained. Stationary combined with time-resolved measurements could be used to distinguish between static and dynamic quenching. Disregarding dynamic quenching may lead to an erroneous estimate of K_{oc} (adsorption coefficients normalized with respect to organic carbon content) values of the pollutants investigated, because of large contributions of dynamic quenching in some of the systems that require attention. For example, the HAs are chemically modified by bleaching, acid hydrolysis, and decarboxylation, and then used in fluorescence quenching experiments for phenanthrene (Phe), 9-phenanthrol (Ptr) and naphthalene (Nap). The quenching mechanisms of these chemicals depend on HA properties, aromatic components exhibit static quenching for Phe and Nap, while carboxyl groups dominantly display dynamic quenching. Aromatic components and carbohydrates in HAs primarily bind (static quenching) rather than collide (dynamic quenching) with Ptr. Carboxyl groups interact with Ptr by dynamic quenching only when carboxyl groups are on the benzene ring (Fig. 6) (Wang et al., 2017). Combining fluorescence quenching and dialysis equilibrium systems, enables exact quantification of the contribution of dynamic quenching. Dynamic quenching plays an essential role, its contribution is generally over 20%, and up to 95% for Nap. Interactions between Ptr and HAs have linear Stern–Volmer plots, and the contribution of dynamic quenching is over 60%. The quenching process should not be judged only according to Stern–Volmer equation (F_0/F vs. [DOC]) because a linear Stern–Volmer plot reflects a solely dynamic or static quenching, without any combination.

Chrysene (Chry) can bind to HAs by quenching the intrinsic fluorescence, and consequently the fluorescence intensity of HAs decreases regularly with the addition of Chry (Fig. 7) (Wu et al., 2020). The quenching rate constant (K_q) is $4.12 \times 10^{12} \text{ L mol}^{-1} \text{ s}^{-1}$ is greater than the collision quenching constant of $2.0 \times 10^{10} \text{ L mol}^{-1} \text{ s}^{-1}$, showing that the process is static quenching.

The 2D-COS can be applied to analyze further the interaction between HAs and benzo[ghi]perylene (B[ghi]P). Although the synchronous spectrum cannot extract information on the HAs from the mixture, their fluorescence characteristics with smaller concentrations can be obtained from a sliced asynchronous spectrum at the signature peak of B[ghi]P (Fig. 8) (Xia et al., 2020). It demonstrates a novel approach for the separation of fluorescence intensities of HAs from PAHs, providing corrected spectra of PAHs for quantitative analysis of the mixture.

The conventional fluorescence quenching method depends on the emission spectral data of PAHs with single wavelength excitation. Nevertheless, in a natural environment where a variety of PAHs coexist, this technique cannot differentiate the interaction of a single PAH with HS from the overall fluorescence spectral information when measuring the binding characteristics

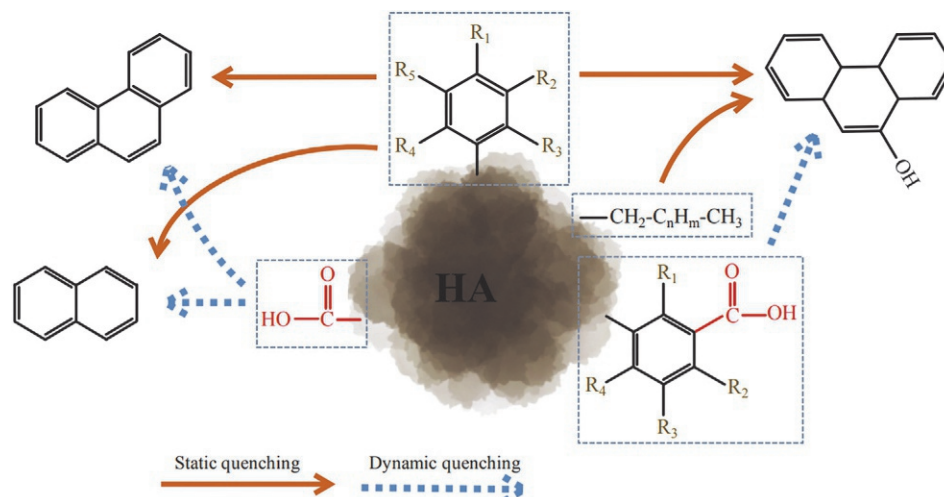


Fig. 6 Schematic diagram of HA-functional groups interaction with different chemicals. Adapted from Wang L, Li H and Yang Y (2017) Identifying structural characteristics of humic acid to static and dynamic fluorescence quenching of phenanthrene, 9-phenanthrol, and naphthalene. *Water Research* **122**: 337–344.

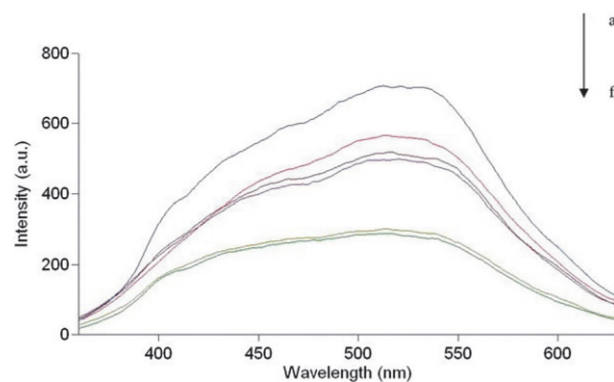


Fig. 7 The changes of fluorescence spectra of Chry-HAs intensity with chrysene concentration (a–f: 0, 0.2, 0.4, 0.6, 0.8, $1.0 \times 10^{-5} \text{ mol} \cdot \text{L}^{-1}$). Adapted from Wu JB, Geng B, and Sun XK (2020) Study on the interaction between chrysene with humic acids using fluorescence spectroscopy. *Polycyclic Aromatic Compounds* **40**: 960–966.

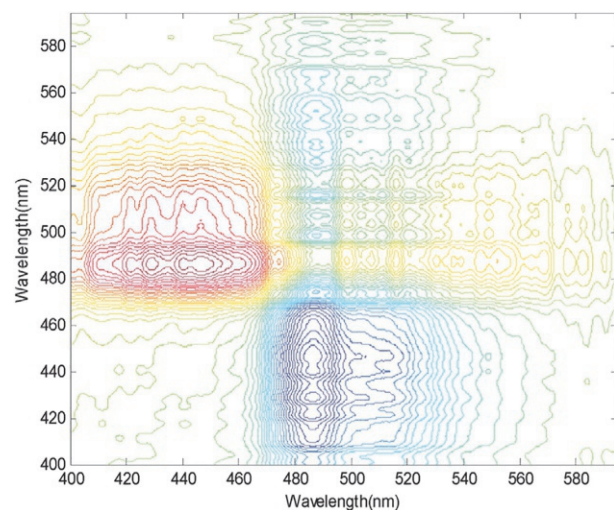


Fig. 8 Asynchronous 2–D correlation fluorescence spectrum of the low-concentration HAs and B[ghi]P mixture. Adapted from Xia MM, Dong GM, and Yang RJ (2020) Study on fluorescence interaction between humic acid and PAHs based on two-dimensional correlation spectroscopy. *Journal of Molecular Structure* **1217**: 128428.

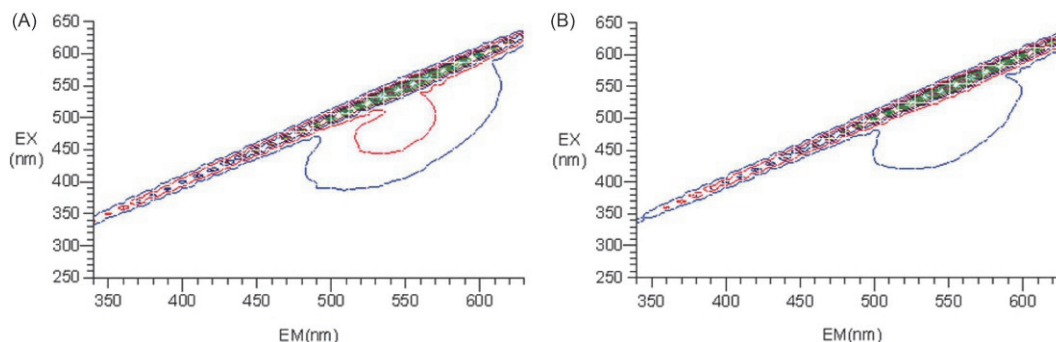


Fig. 9 Three-dimensional fluorescence contour map of HA (a), and the Chry-HA system (b). Adapted from Wu JB, Geng B, and Sun XK (2020) Study on the interaction between chrysene with humic acids using fluorescence spectroscopy. *Polycyclic Aromatic Compounds* **40**: 960–966.

of HS to PAHs. Fortunately, EEM fluorescence spectroscopy provides this potential. For example, with the addition of Chry, the EEM fluorescence intensity of HAs clearly decreases (Fig. 9), suggesting that Chry could enter hydrophobic cavities in some domains of HAs, and the binding of Chry would induce some micro-environmental changes in HAs (Wu et al., 2020).

The EEM-PARAFAC analysis provides a detailed characterization of the binding process of PAHs and FAs (Fig. 10) (Lu et al., 2013). The K_c and binding capacities (L_t) of each PAH with FAs in a mixture can be calculated simultaneously by (Eq. 1). Log K_c is 2.78 for Flu and 2.44 for Ant, and L_t $8.65 \times 10^{-3} \text{ g L}^{-1}$ for Flu and $12.67 \times 10^{-3} \text{ g L}^{-1}$ for Ant, suggesting both PAHs bind tightly to FAs, thus increasing their solubility in water.

$$\frac{I}{I_{\text{ref}}} = 1 + \left(\frac{I_{\text{PF}}}{I_{\text{ref}}} - 1 \right) \frac{1}{2K_c L_t} \left[1 + K_c L_t + K_c C_F - \sqrt{(1 + K_c L_t + K_c C_F)^2 - K_c^2 L_t C_F} \right] \quad (1)$$

where I is the fluorescence intensity at a given C_F , I_{ref} is the initial fluorescence intensity of the PAHs without the addition of FAs, I_{PF} is the relative fluorescence intensity of the FAs-PAHs complex and K_c is the conditional stability constant.

Fluorescence quenching and synchronous-scan fluorescence spectroscopy has been used to obtain K_c for the binding of carbamate pesticides to soil DOM and HAs. For example, the binding constant of carbofuran is greater than that of carbaryl and aldicarb at pH 6.0. These results can be used to predict the potential transport of pesticides down the soil profile.

Soil mineral components

Static and dynamic fluorescence of organic probes can yield important and unique information about specific interface and surface behavior of clay particles in diluted colloidal dispersions or suspensions. The use of fluorescent probes is based mainly on their fluorophores, and the wavelength and time resolution are determined by the spectral properties of the fluorophores. Therefore, the lifetime of the fluorophores must be compared with the timescale of interest in the experiment. Intrinsic and external fluorophores constitute the two main categories of fluorophores. To obtain the relevant information required by the experiment, appropriate fluorescent probes should be selected. For example, the measurement of pH can be carried out with a pH sensitive fluorescent probe. Moreover, the influence of background fluorescence must also be considered when determining the target substances in soil.

To measure properties of soil mineral material, the commonly used fluorescent probes include some organic cationic dyes, such as acridine orange (AO), methylene blue, proflavine and rhodamine 6G, and various PAHs, such as Ant, 1,2,5,6-dibenzoanthracene, Pery, B[a]P, pyrene (Pyr) and various cationic pyrenyl derivatives. Advantages of the fluorescence methods to determine mineral components in soil are as follows: (1) it has great potential for non-invasive research of colloidal suspensions because there is no need for a complex pretreatment process for solid-liquid separation; (2) the specific fluorescent probe can detect very small concentration targets in hydrated clay suspensions under natural environmental conditions because of the analytical sensitivity of the fluorescence method; (3) the fluorescent probe detection signal not only provides information about target detection substances, but can also provide multi angle information of the colloid solution surface; (4) reaction rates can be determined continuously across wide time scales, ranging from minutes to femtoseconds. In addition to the reaction rates, events occurring at femtoseconds can be measured by TRF decay.

There are still some disadvantages, however, in using fluorescence spectroscopy to study mineral systems that need to be solved. First, dense colloidal solutions can lead to excessive light scattering, which is not conducive to the excitation and recording of the fluorescence signal. Second, fluorescence quenching also affects the collection of fluorescence signals. For example, the Fe^{3+} ion is located within the crystal structure or on the surface of clay minerals, and it can exchange electrons with the fluorescence probe to cause fluorescence quenching. Therefore, the use of fluorescence is restricted by the concentration of the suspended colloidal solution and the cations that may produce a quenching effect.

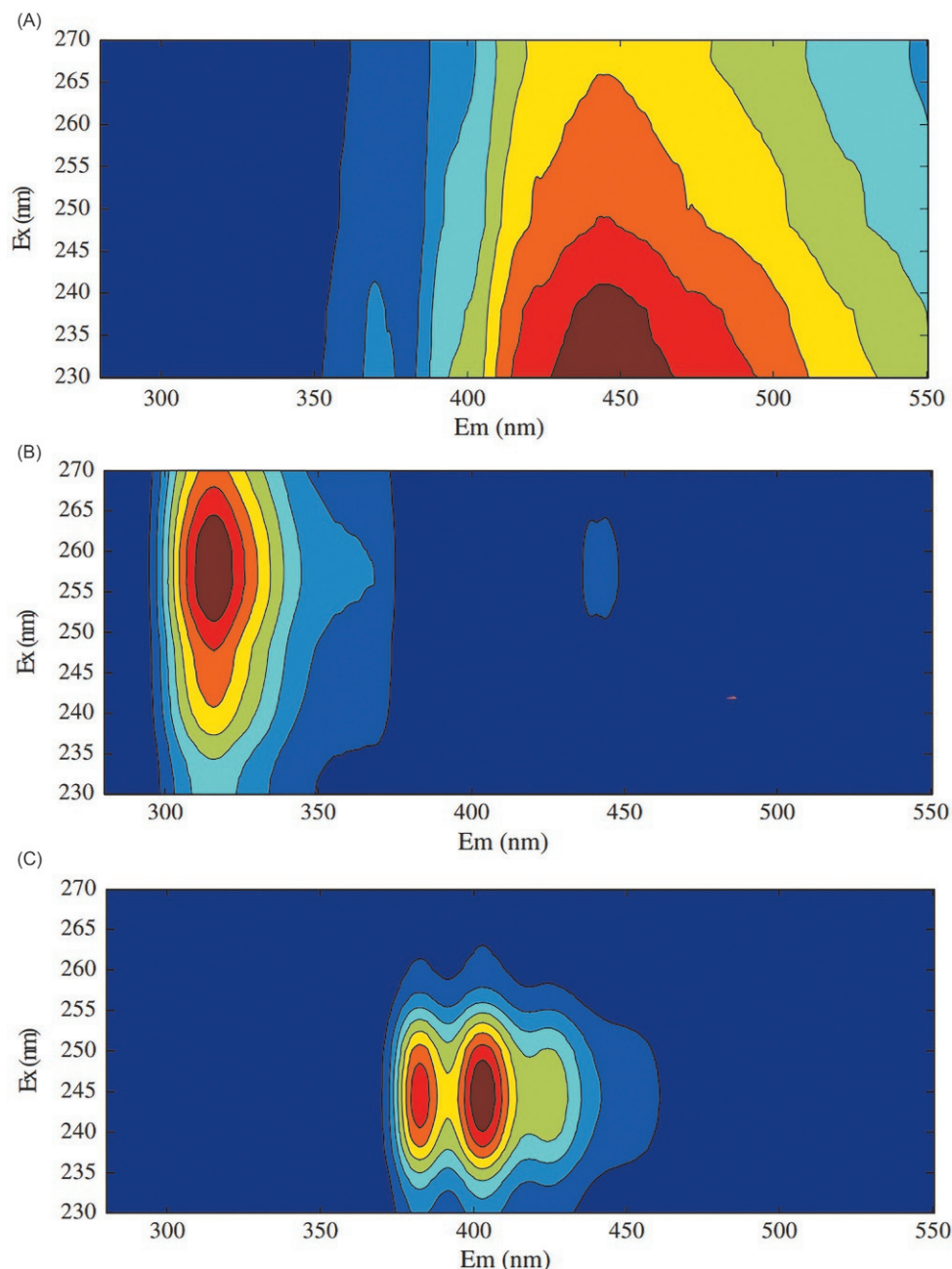


Fig. 10 The EEM fluorescence spectra of FA (a), fluorene (b), and anthracene (c) generated by PARAFAC analysis. Adapted from Lu R, Sheng GP, and Liang Y (2013) Characterizing the interactions between polycyclic aromatic hydrocarbons and fulvic acids in water. *Environmental Science and Pollution Research* **20**: 2220–2225.

Structural quenching

The quenching effect of fluorescent probes tris (2,2''-bipyridyl) Ru^{3+} (ruthenium) and tris (2,2''-bipyridyl) Cr^{3+} on sodium saturated Wyoming montmorillonite is better than that on synthetic montmorillonite containing Cr^{3+} in the octahedral layer. Furthermore, compared with the Fe^{3+} and Cr^{3+} in complex states, which means they are in octahedral positions, the Fe^{3+} and Cr^{3+} at the interlayer exchange sites are more likely to have electron exchange reactions with the fluorescence probe, that is, fluorescence quenching is more likely to occur. Similarly, the quenching efficiency of octahedral Cu^{2+} for the adsorbed fluorophore [4-(1-pyrenyl) butyl] trimethylammonium (PN4) is higher than that of exchangeable Cu^{2+} . These effects can be attributed to the close combination of adsorbed fluorine groups with siloxane dihedral cavities on the clay surface.

The accessibility of fluorophores to quenchers can be used to determine the position of fluorescent species at the mineral-water interface and its distribution in mineral micelles. For example, the inorganic fluorophore Ce^{3+} (cerium) preferentially occupies ion exchange sites in the tubules of imogolite. The site of solubilization of Pyr determines the mechanism of quenching by polar and non-polar quenchers. When the fluorescent probe Pyr is at the water-clay micelle interface, it is easily quenched by polar and nonpolar quenchers, but when it is inside the nonpolar micelle it will not be quenched by polar quenchers (such as Cu^{2+}). However, in both cases, the fluorescence of Pyr is easily quenched by non-polar quenchers, and the Pyr can also be dissolved within the interior of micelles.

Adsorption of organic chemicals to mineral surfaces

Measurements of anisotropy can reveal the average angular displacement of the fluorophore between photon absorption and subsequent emission. This angular displacement depends on the rate and degree of rotational diffusion in the lifetime of the excited state. The rate of rotational diffusion depends on the viscosity of the solvent and the size and shape of the rotating molecules. The rotational rate of fluorophores in solution depends on the viscous drag imposed by the solvent. A change in solvent viscosity will result in a change in fluorescence anisotropy. Therefore, the adsorption of Ant, Pyr, Pery, 1,2,5,6-dibenzo Ant, B[a]P and Chry on a hydrosol kaolinite was studied by the fluorescence polarization method. In all cases, fluorescence anisotropy decreased significantly as a function of PAH concentration, which means that the rotational relaxation time at low concentration is much larger than that at high concentration. The change in fluorescence anisotropy is clearly related to the concentration of colloidal kaolinite. At small concentrations, most PAH molecules are adsorbed on colloidal kaolinite, and their motion in solution is hindered. With increasing concentration, surface saturation of the colloid is reached, and additional molecules are no longer adsorbed and remain 'free' to rotate, leading to the decrease in fluorescence anisotropy.

Electro fluorescence polarization spectroscopy has also been used to measure the polarization of aqueous suspensions of the clay minerals sepiolite, attapulgite and saponite particles. These aqueous suspensions are labeled with dye molecules of different sizes and charges, and oriented via the application of an electric field. The data obtained make it possible to infer different possible binding sites between the dyes and clay. For example, cationic dyes are mainly confined to the channel of sepiolite rods, while attapulgite or Hector stone has no effect on them. To analyze the fluorescence information obtained fully, EEM-PARAFAC can be used to examine the chemical interactions between soil mineral surfaces and DOM. For example, the PARAFAC of multidimensional fluorescence spectra enables direct determination of the different affinities of the modeled DOM components (i.e., fluorophores) to soil mineral surfaces. This information may be a valuable management tool in assessing the potential sequestration of organic matter derived from a variety of carbon-rich soil amendments in both managed and unmanaged ecosystems.

The fluorescence results showed that DOC desorbed from mineral soil surfaces had a substantially lower fluorescence index than the equilibrium DOC solution. This finding also emphasizes the need to be aware that while overall DOC concentration may reduce substantially, the resulting composition of DOC in the solution phase can reflect not only adsorption, but also desorption processes. Based on the fluorescence index, soils with the greatest capacity to adsorb DOC had the greatest decrease in molecular weight distribution and aromatic content. From an ecological perspective, the Fe status of mineral horizons may be of primary importance when attempting to predict the quantity and quality of DOC exported from soils into aquatic ecosystems. In summary, fluorescence polarization can be used to study the interaction between organic pollutants and natural minerals without any separation. The dependence of fluorescence anisotropy on fluorophore motion has resulted in applications of this technique to dynamic processes such as temperature effect and clay adsorption on the properties of fluorophore.

Surface acidity

In addition to the interaction between a fluorophore and the surrounding environment, the fluorophore can also undergo chemical excitation, because the distribution of electrons in the fluorophore is changed by light absorption, thus changing its chemical or physical properties. For example, phenol and naphthol can lose phenolic protons in an excited state in a neutral solution. Because deprotonation is more likely to occur in the excited state, electrons from the hydroxyl group are likely to be transferred to the aromatic ring, making the solution more acidic. Therefore, fluorophores with pH-dependent fluorescence can be used to detect the acidity of a clay surface and the protonation of organic molecules at the clay-solution interface. For example, the emission spectrum of quinoline on sodium saponite shows that quinoline ions are adsorbed on the surface in the range of pH 3.0–9.0. The protonation of quinoline can be attributed to the specific acidic sites at the mineral-water interface and the strong ion exchange capacity of quinoline ions. In addition, the characteristic emission spectra of diprotonated protoflavine were observed when sodium saturated montmorillonite, Wyoming bentonite or saponite suspension was added in a monoprotonated form.

Whether a fluorophore loses or gains a photon, in the excited-state reaction is determined from the direction of change in the acid dissociation constant (pK_A). If the pK_A decreases ($\text{pK}_\text{A}^* < \text{pK}_\text{A}$), where the asterisk denotes the excited state, then the fluorophore will tend to lose a proton. If the pK_A increases ($\text{pK}_\text{A}^* > \text{pK}_\text{A}$), then the fluorophore may obtain a proton in the excited state. Based on this process, fluorescence spectra can be used to infer the presence of both protonated and unprotonated species on the clay surface. For example, aminopyrene sorbed to colloidal silica exhibits both the largely structured emission spectrum of protonated aminopyrene and the broad, featureless emission of neutral aminopyrene, which indicate the presence of both species on the surface; the protonated one on the more acidic sites, and the neutral one on the less acidic sites.

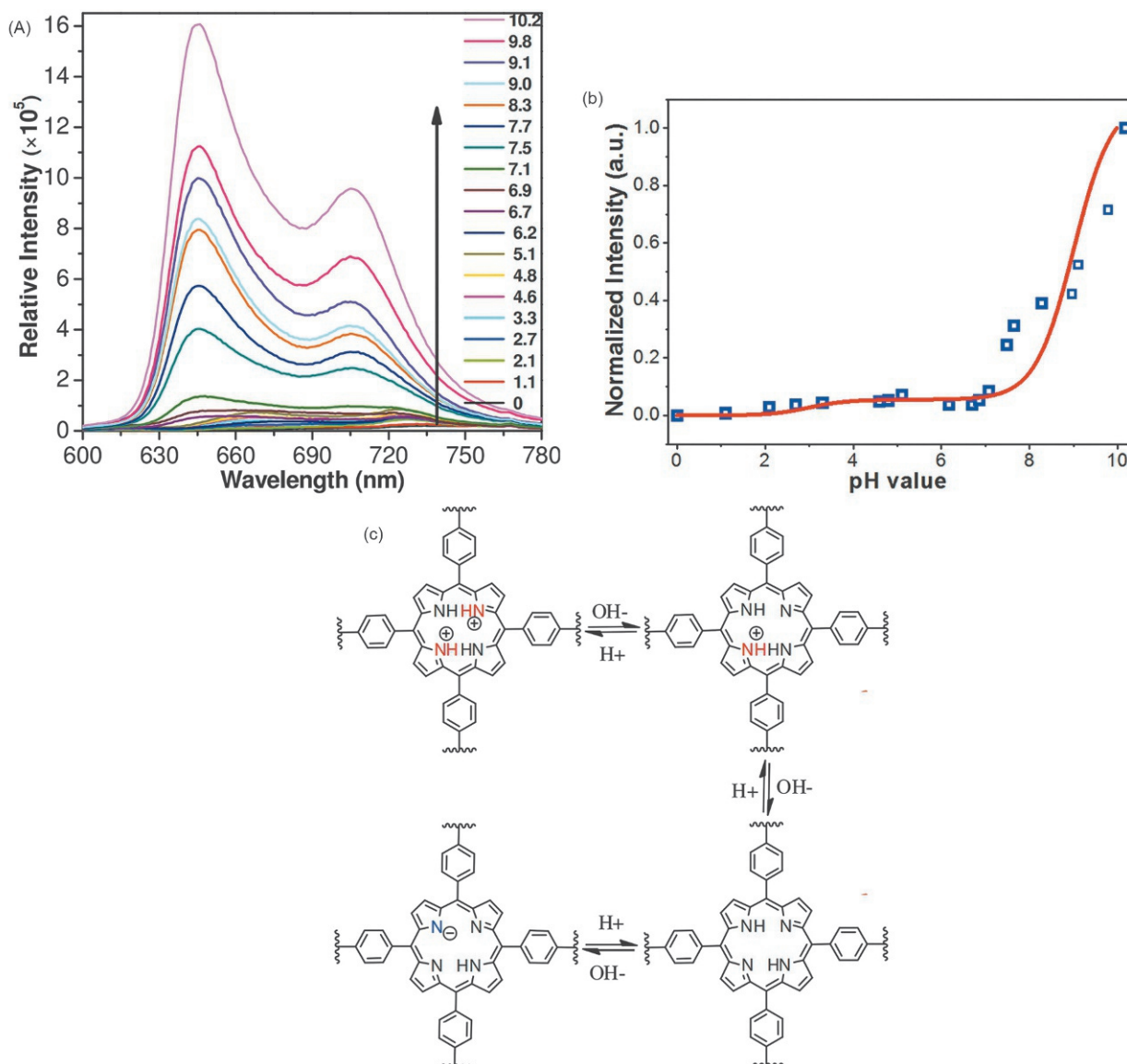


Fig. 11 (a) pH dependent fluorescence of PCN-225 in aqueous solutions with pH ranging from 0 to 10.2 measured under excitation of 415 nm. (b) Blue points showed fluorescence emission at 725 nm at different pH values, matching the simulation fluorescence intensity (red line), and (c) protonation and deprotonation processes of porphyrin involved in the PCN-225 framework of experimental acidic and basic media (pH = 0 to 10.2). Adapted from Jiang HL et al. (2013) An exceptionally stable, porphyrinic Zr metal-organic framework exhibiting pH-dependent fluorescence. *Journal of the American Chemical Society* **135**(37): 13934–13938.

Based on these pH-related properties, relevant fluorescent sensor materials can be developed, zirconium–porphyrin frameworks, PCN-225 and PCN-225(Zn), have been developed as the first metal–organic framework (MOF) with a (4,8)-connected SQC net.

They exhibit strong chemical stability in aqueous solutions with pH values from 1.0 to 11.0 (Fig. 11), and as far as we know, this is the most extensive pH range at which porphyrin MOF can exist (Jiang et al., 2013). It is worth noting that PCN-225 has stability at a wide range of pH. With the change of pH, the central core region of a porphyrin free base is in the equilibrium state of protonation and deprotonation, and its fluorescence intensity is closely related to the change of pH; $7.0 < \text{pH} < 10.0$ is the most sensitive range for intensity response. The results show that PCN-225 has broad application potential for pH sensing. In addition, the ready-made installation of the metal sites exposed by the porphyrin center in PCN-225, coupled with its stability, should improve its potential in various application fields, especially catalysis, light acquisition, sensors, and so on.

Polarity and charge adsorption of mineral surfaces

Research has shown that some fluorescent probes have great sensitivity to the polarity of the surrounding solvent, therefore, these probes can provide specific information about the charged regions on colloidal mineral surfaces. The large value of the ratio of

fluorescence emission intensities at $\lambda = 372$ to $\lambda = 391$ nm (I/III ratio) of Pyr, [3-(1-pyrenyl) propyl] trimethyl ammonium (PN3), and 8-(1-pyrenyl) octyl ammonium (PN8), adsorbed on to laponite from water and methanol solutions, confirms that the charged regions of the interlamellar surfaces of clay particles are largely polar. In contrast, the very small I/III ratios observed when fluorescent probes are adsorbed on to clays coated with alkylammonium surfactants, suggests that they are in less polar environments in forms strongly associated with the adsorbed surfactants.

Organic dye molecules can also alter the apparent polarity of clay mineral surfaces. Cationic dye molecules can exchange ions at permanent charge sites on the internal and external surfaces of mineral as well as at acidic sites on the mineral edges. For the fluorescence measurements, powdered samples were placed on quartz substrates. Fluorescence was excited by a continuous wave He–Cd laser at 325 nm. Klika et al. (2009) have confirmed that a small layer charge and small amount of adsorbed methyl blue (MB) induce large fluorescence intensities in MB after laser excitation. They found that the reduced charge of montmorillonite with loading a small amount of MB had large fluorescence intensity after laser excitation. A single sharp fluorescence maximum of these complexes matches the presence of dye monomers attached to the montmorillonite surface. The charge of the silicate layer had a strong effect on the peak shift of the crystalline methyl red (MR) fluorescence band. When MR is adsorbed on montmorillonite and vermiculite, the maximum shift of the fluorescence band is transferred from 800 nm to 565 nm and 645 nm, respectively.

Molecular distribution, conformation, aggregation and mobility

Excitation spectra and fluorescence lifetime measurements show that adsorbed cationic probes, including PN4, PN3, and PN8, tend to cluster on the surfaces of hydrated clay minerals, and then form fluorescent dimers like excimers. By analyzing the ratio of monomer emission to excimer emission of Pyr-labeled polyacrylic acid (Pyr-PAA) in solution and the ratio of adsorption to aluminium oxide in aqueous suspension, the effect of pH on the interfacial conformations of polymeric and polyelectrolyte macromolecules (e.g., Pyr-PAA) can be examined. No change in excimer emission intensity is observed: when Pyr-PAA is adsorbed on to aluminium oxide at pH 8.0, the solid is separated from the aqueous phase and then resuspended in a solution at pH 4.8. This result shows that the polymer is adsorbed in the uncoiled, linear conformation and remains uncoiled independent of the pH of the surrounding medium. In other words, no conformation change occurs as a function of pH. In contrast, excimer emission intensity is observed to decrease when Pyr-PAA is adsorbed on to aluminium oxide at pH 4.7; the particles are separated and then resuspended at pH 8.0.

The formation of dye aggregates is affected by the charge density. The higher aggregates (H-aggregates) with a face-to-face association of dye cations would be formed on high charge-density surfaces, while monomers, dimers, and less densely packed conjugated polymers were favored over H-aggregates with a moderate reduction in layer charge. The dye is positively charged and minerals overall are negatively charged. Although the lumen of halloysite is positively charged, it is unlikely that hydrophobic interactions between acridine orange (AO) and the charged mineral surface played a role. The oxygen on the surface of the tetrahedral sheet has strong electronegativity and thus, could interact with AO^+ . Substitution of Al^{3+} for Si^{4+} (silicon) at tetrahedral sites could also contribute to the net negative surface charge, seat of charge, and charge density of the minerals, which could cause the sorption and aggregation of AO on their surface. In addition, at maximum AO adsorption, the tight packing of AO leads to stronger intermolecular interactions, resulting in the formation of dimers and other aggregates, ultimately leading to significant fluorescence quenching. Interestingly, for the same mineral class, the adsorption, absorption, and strength of Cm^{3+} (curium) increase with surface roughness around the surface holes or grain boundaries.

Surface roughness becomes the main retention property of minerals with low adsorption capability when competitive adsorption occurs between multi-mineral phases. In high surface roughness areas primarily, Cm^{3+} inner-sphere sorption complexation and surface incorporation are prominent and at selected sites formation of stable Cm^{3+} ternary complexes is observed. The adsorption of organic matter by clay minerals in lake sediments plays an important role in organic matter burial. The clay surfaces selectively preserve algal organic matter (AOM) in eutrophic lakes, and the fraction of various AOM components preserved varies. Fluorescence emission spectroscopy has a promising future in studying the molecular structure of other polymer interfaces in the mineral–solution region. However, confirmation of fluorescence results usually requires the use of one or more additional techniques, such as molecular simulation.

Biochemical transformation of clay surfaces

Degradation of quinoline in a hectorite suspension by bacteria was studied by fluorescence spectroscopy. At pH values well above the logarithmic base dissociation constant (pK_b) of quinoline (4.9), the only fluorescent species was the protonated quinolinium ion adsorbed on to hectorite surfaces. During microbial degradation, the reduction in concentration of adsorbed quinolinium was monitored together with production of the metabolite 2-hydroxyquinoline. Desorption of the adsorbed quinoline ions was determined as the rate-limiting step in the bacterial degradation of quinoline by comparing the degradation rates of saponite with and without hectorite.

The interactions between microorganisms and clay minerals involve the formation, transformation, and preservation of organic matter formation. In addition, the bioavailability of 'heavy metals' is strongly influenced by the formation of clay mineral–bacteria complexes and their derivatives. Therefore, bacterial biomineralization is increasingly being considered as a good alternative for the remediation of contaminated soils, to recover the fertility of abandoned mining lands, and to precipitate or immobilize toxic metals. Bacterial biomineralization can also be used for the treatment of sewage, and industrial and domestic wastewater.

Ligand-promoted mineral dissolution

Direct evidence can be provided for the formation of fluorescent inner-sphere complexes between Al^{3+} ions and 8-hydroxyquinoline-5-sulfonate at the Al oxide-water interface. Fluorescence measurements show that concentration of the surface complex is constant during long-term dissolution of the oxide, which indicates that the oxide surface is continuously regenerated during dissolution; this confirms a surface-controlled process of mineral dissolution.

Previous studies have shown that organic ligands could significantly affect the kinetics and reaction mechanisms of mineral dissolution through surface complexation with metal ions. Fiore et al. (2011) found that kaolinite can be formed through a two-step biological induction process. The first step is precipitation of aluminosilicates in the presence of oxalates and other organic compounds, including extracellular polymers substances, which are metabolites produced by bacteria. The second step is due to micro-environmental changes caused by the metabolic activity of bacteria, and the gel may dissolve locally or rearrange into a kind of solid state, forming kaolinite crystals.

Conclusion

The practical applications of fluorescence quenching, fluorescence polarization, fluorescence correlation spectroscopy, EEM and time-resolved fluorescence can be applied to explore the interaction among soil components and chemicals. For example, fluorescence anisotropy was used to explore conformational variations in DOM and molecular interactions between it and environmental contaminants by evaluating the rotary motion of DOM molecules. Moreover, synchronous-scan fluorescence spectra, EEM-PARAFAC analysis and TRF have been used successfully to study the fluorescence quenching of FAs by metal ions. For the micro-environmental behavior of chemicals of the soil mineral components, the above techniques could not obtain valid information. To address this issue, further study should focus on the application of fluorescence imaging technology.

References

- Bolan NS, Adriano DC, Kunhikrishnan A, James T, McDowell R, et al. (2011) Dissolved organic matter: Biogeochemistry, dynamics, and environmental significance in soils. *Advances in Agronomy* 110: 1–75.
- Chen W and Yu HQ (2021) Advances in the characterization and monitoring of natural organic matter using spectroscopic approaches. *Water Research* 190: 116759.
- Chen W, Westerhoff P, Leenheer JA, and Booksh K (2003) Fluorescence excitation—Emission matrix regional integration to quantify spectra for dissolved organic matter. *Environmental Science & Technology* 37(24): 5701–5710.
- Chen W, Habibul N, Liu XY, Sheng GP, and Yu HQ (2015) FTIR and synchronous fluorescence heterospectral two-dimensional correlation analyses on the binding characteristics of copper onto dissolved organic matter. *Environmental Science & Technology* 49(4): 2052–2058.
- Fiore S, Dumontet S, Huertas FJ, and Pasquale V (2011) Bacteria-induced crystallization of kaolinite. *Applied Clay Science* 53(4): 566–571.
- Fu QL, He JZ, Blaney L, and Zhou DM (2016) Roxarsone binding to soil-derived dissolved organic matter: Insights from multi-spectroscopic techniques. *Chemosphere* 155: 225–233.
- Ishii SKL and Boyer TH (2012) Behavior of reoccurring PARAFAC components in fluorescent dissolved organic matter in natural and engineered systems: A critical review. *Environmental Science & Technology* 46(4): 2006–2017.
- Jiang HL, Feng DW, Wang KC, Gu ZY, Wei ZW, et al. (2013) An exceptionally stable, porphyrinic Zr metal-organic framework exhibiting pH-dependent fluorescence. *Journal of the American Chemical Society* 135(37): 13934–13938.
- Klika Z, Pustková P, Praus P, Kovář P, Pospíšil M, et al. (2009) Fluorescence of reduced charge montmorillonite complexes with methylene blue: Experiments and molecular modeling. *Journal of Colloid and Interface Science* 339(2): 416–423.
- Liu MX and Lang YC (2021) Differences in the spectroscopic characteristics of wetland dissolved organic matter binding with Fe^{3+} , Cu^{2+} , Cd^{2+} , Cr^{3+} and Zn^{2+} . *Science of the Total Environment* 800: 149476.
- Lu R, Sheng GP, Liang Y, Li WH, Tong ZH, et al. (2013) Characterizing the interactions between polycyclic aromatic hydrocarbons and fulvic acids in water. *Environmental Science and Pollution Research* 20(4): 2220–2225.
- Mei Y, Wang L, and Wu F (2009) Effects of water chemistry and concentrations of dissolved organic matter on its fluorescence characteristics and molecular conformation. *Chinese Journal of Geochemistry* 28(4): 413–420.
- Wang L, Li H, Yang Y, Zhang D, Wu M, et al. (2017) Identifying structural characteristics of humic acid to static and dynamic fluorescence quenching of phenanthrene, 9-phenanthrol, and naphthalene. *Water Research* 122: 337–344.
- Wu JB, Geng B, and Sun XK (2020) Study on the interaction between chrysene with humic acids using fluorescence spectroscopy. *Polycyclic Aromatic Compounds* 40(4): 960–966.
- Xia MM, Dong GM, Yang RJ, Li XC, and Chen Q (2020) Study on fluorescence interaction between humic acid and PAHs based on two-dimensional correlation spectroscopy. *Journal of Molecular Structure* 1217: 128428.

Near and mid infrared soil spectroscopy

Wartini Ng^a, Brendan Malone^b, Budiman Minasny^a, and Sangho Jeon^c, ^aSydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Eveleigh, NSW, Australia; ^bCSIRO Agriculture and Food, Black Mountain, ACT, Australia; ^cNational Institute of Agricultural Sciences, Rural Development Administration, Wanju, Republic of Korea

© 2023 Elsevier Ltd. All rights reserved.

Introduction	567
Spectroscopy in soil science	567
Fundamental absorption of clay minerals	569
Soil chemometrics	569
Soil properties	571
Soil physical properties	571
Soil chemical properties	571
Soil biological properties	572
Soil contaminants	572
Lower cost - miniaturized spectrometer	573
Conclusions	573
References	573

Abstract

Soil Infrared spectroscopy uses electromagnetic energy to characterize the soil. In the past decades, visible, near to mid infrared spectroscopy has gained considerable interest as a viable alternative for in-field and laboratory measurements, especially where a high spatial density of measurements is needed, because of its speed and cost-effectiveness. Once chemometric models have been calibrated, a single spectrum can be used to predict a range of soil properties simultaneously. This chapter provides an overview of spectroscopy in soil research, including predicting physical, chemical, and biological properties, as well as determination of soil contaminants.

Key points

- Infrared spectroscopy has been used to predict a range of physical, chemical and biological soil properties successfully.
- Soil contaminants have also been successfully quantified with infrared spectroscopy.
- Developments in artificial intelligence, computational power and big data have made more complex machine learning algorithms possible.
- Smaller, portable and lower-cost spectrometers are now available.

Introduction

Obtaining quantitative soil information rapidly and objectively is essential for monitoring current soil conditions and maintaining if not improving them, to ensure the current management practice does not lead to soil degradation. Quantitative soil data, which vary in time and space, are conventionally obtained from laboratory analysis. However, this is expensive, time-consuming, and requires the use of chemical reagents.

Research has been conducted worldwide to investigate proximal sensors that can measure soil properties from a close distance to the soil surface using a various range of wavelengths of the electromagnetic spectrum as shown in Fig. 1. The electromagnetic energy ranges from gamma rays which have higher frequencies to radio waves with the lowest frequencies. The interaction of energy with matter depends on its molecular composition. A material can reflect, scatter, or emit electromagnetic radiation, which results in a unique spectral signature. These sensors measure the response of the material at certain wavelengths, providing information about its constituents. Soil responds uniquely to the infrared spectrum, and the technology has made infrared spectrometers suitable for the analysis of soil samples as they can measure rapidly, cost-effectively, and non-destructively.

Spectroscopy in soil science

Optical sensors utilize various wavelengths from the electromagnetic spectrum (Fig. 1) to characterize properties (visible (vis): 400–700 nm; near-infrared (NIR): 700–2500 nm; mid-infrared (MIR): 2500–25,000 nm). The absorption in the MIR region can be readily identified as it is related to the fundamental vibrations of molecules. The absorbance in the vis–NIR is more difficult to

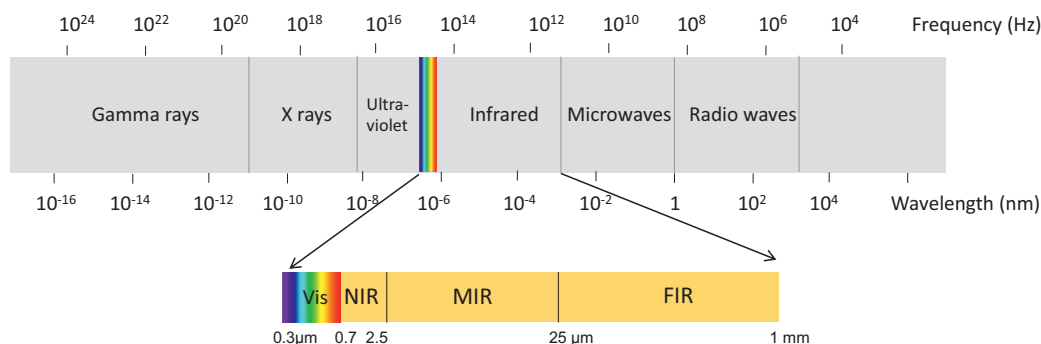


Fig. 1 The electromagnetic spectrum showing the boundaries between different regions, including the visible region (380–700 nm), the near infrared (NIR; 700–2500 nm), the mid-infrared (MIR: 2500–25,000 nm), and the far infrared (FIR 25,000–1,000,000 nm).

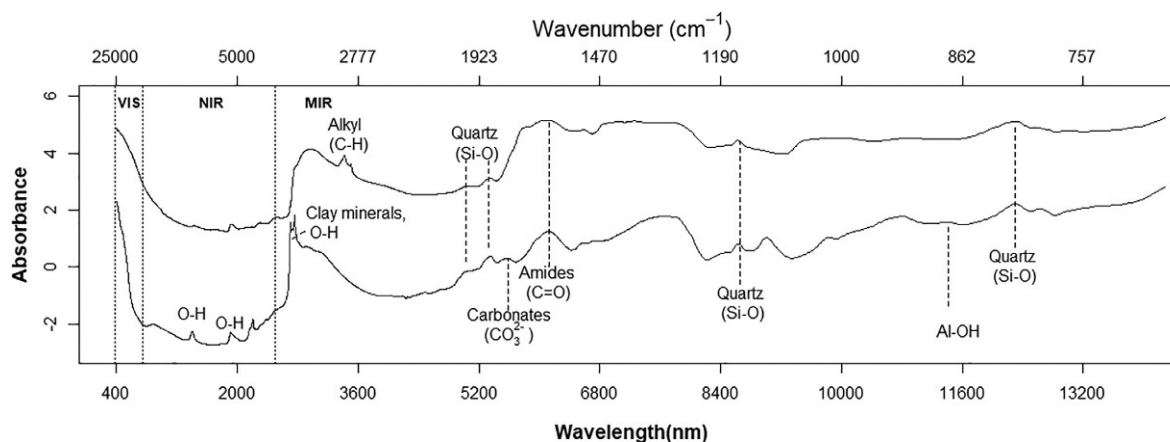


Fig. 2 The vis-NIR-MIR spectra of two soil samples showing fundamental peaks on MIR and overtones on NIR.

interpret as there are fewer and broader absorptions related to overtones and combinations of the fundamental vibrations in the MIR region. Absorptions in the visible region are associated with minerals that contain iron and organic matter, while absorptions in the NIR region (780–2500 nm) result from the overtones of hydroxyl (OH^-), sulfate (SO_4^{2-}), and carbonate (CO_3^{2-}) groups, and combinations of fundamental features of water (H_2O) and carbon dioxide (CO_2) (Stenberg et al., 2010). An illustration of spectra from the vis-NIR-MIR region is included in Fig. 2.

Discovery of the infrared region dates back to 1800 by Herschel (1800). However, the practical applications did not occur until 150 years later. In 1950s, researchers explored the use of infrared to characterize soil minerals (Hunt et al., 1950). In the 1980s, although NIR was mainly used in agriculture and the food industries, it became popular the decade after when practical applications started to expand to other industries, including pharmaceutical, earth science and environmental analyses. In soil science, early use of near infrared spectroscopy was by Bowers and Hanks (1965) where the influence of moisture, organic matter and particle size on the soil absorbance spectrum were observed.

Both vis-NIR and MIR have been used to determine successfully various chemical (pH, cation exchange capacity, carbon contents (organic, inorganic, and total) and nutrient contents), physical (texture (sand, silt and clay) and minerals (kaolinite, gibbsite, montmorillonite and iron oxides)), and biological (microbial biomass and enzyme activities) soil properties (Soriano-Disla et al., 2014).

The use of infrared spectroscopy is gaining interest for applications in various settings due to its possible application on different platforms, ranging from proximal sensing in the field, within the laboratory setting to remote sensing platforms with hyperspectral capabilities. Improvement in instrumentation also leads to the development of portable and handheld spectrometers. Recently, lower-cost portable and miniaturized NIR spectrometers with limited wavelength ranges have also become available, making it a cost-effective method of analysis for farmers, agricultural consultants, and researchers. It has been demonstrated that the miniaturized spectrometer provided comparable accuracy to the standard vis-NIR spectrometer (Tang et al., 2020).

Over the past two decades, the application of spectroscopy to complement soil analyses has increased rapidly. This increase is correlated with progress in the multivariate statistical data analysis and the development of both infrared and computer technologies.

Fundamental absorption of clay minerals

The vis–NIR and MIR spectra can be used for detecting clay mineral species and iron oxides. The most common soil forming clay minerals include kaolinite, montmorillonite (also commonly referred to as smectite) and illite. The most common iron oxide species include hematite and goethite. These minerals are distinct in composition, structural arrangements, and physicochemical characteristics. Consequently, each mineral has specific spectral responses at specific wavelengths in the electromagnetic spectrum (Clark et al., 1990). Earlier contributions from Gadsden (1975) comprehensively summarized diagnostic spectral features of minerals. Some diagnostic wavelengths in both the vis–NIR and MIR spectral regions for the key clay minerals and iron oxides are provided in Table 1.

The relatively simple way to assess the mineral composition of a soil sample directly from infrared spectra is to compare the reflectance of the diagnostic wavelengths from the reference spectra with those at the same wavelengths of the soil samples. As established, the wavelength ranges specifically diagnostic to each clay mineral and iron oxide specimen are known and they can be isolated and standardized by fitting a convex hull to them. The convex hull fits a straight line that connects the local maxima of the spectrum of a specific wavelength range and represents the background information. Fig. 3 shows an example of the fitted convex hulls to the diagnostic reflectance ranges from the reference spectra sourced from the United States Geological Survey (USGS) spectral library for end member specimens of smectite and kaolinite, and the associated spectra once the continuum is removed.

Soil infrared spectra are similarly standardized followed by the application of comparative metrics between reference and soil spectra to derive relative abundances. For example, shape match algorithms can be used and are implemented in the USGS Tetracorder decision making framework for mineral detection (Clark et al., 2003). Other metrics include relative area and depth of the absorbance feature and slope of the convex hull. Relative abundance of the minerals can then be inferred. For example, Fig. 4 provides statistical summaries of the relative abundances of kaolinite and smectite minerals derived from soil survey samples from the Hunter valley region, NSW Australia used in Malone et al. (2014). Here, there is a clear relationship between estimated abundance and its corresponding spectral response feature relative to the reference specimen.

The alternative to direct inference on mineral composition from infrared spectra is with soil spectral calibration models, which are described in more detail further on. Robinson and Kitching (2016) demonstrated this where partial least squares regression (PLSR) and support vector machine models were trialled and constructed to estimate soil clay mineral compositions (kaolinite, smectite and illite) based on associated soil MIR spectra. These models were extended to a larger spectral library to enable digital soil mapping of these variables in their study area in Victoria, Australia. The authors reported that their model can have an accuracy of 70% or better for each of the mineral species.

Soil chemometrics

As soil is a complex mixture of materials, it is difficult to assign particular features of the spectra to specific physical, chemical, or biological components. Multivariate calibration techniques are used to predict soil properties, relating the spectra to observed soil properties (properties that have been measured using conventional laboratory analysis).

Digital soil spectra usually contain hundreds or thousands of reflectance values as a function of wavelength. Since there are typically more predictor variables than the observed variables, methods that reduce the dimensions of the predictors are required. Principal component regression and PLSR are commonly utilized methods. Partial least squares regression extracts successive linear combinations of the predictors, optimizing the combined goals of explaining response variation and predictor variation.

Another way to handle large dimensional data is to use variable selection techniques. When the high dimensional data have been reduced to several components or important variables have been selected, they are used for prediction using either linear regression or data-mining tools. Regression trees, neural networks and support vector machines have been used for such predictions.

There are also machine learning algorithms that can handle a large dimension of inputs and extract information from many samples. Models such as Cubist models, random forest, neural networks, support vector machines have been trialled successfully for predicting soil properties from vis–NIR and MIR data. While such models have been used successfully for prediction, each model

Table 1 Diagnostic wavelength ranges of clay minerals and iron oxides as reported by Gadsden (1975).

Mineral	Diagnostic wavelength range(s) in wavenumber (cm^{-1})	
	Visible and near infrared	Mid infrared
Kaolinite	4812–4411	3636, 2710, 1923–1818, 1587, 1064, 855
Smectite	4721–4372	3636, 1961, 1852, 1639, 1111, 1064
Illite	4640–4413, 4336–4192	3226, 2874, 2141, 2604–2513, 1799, 1639
Goethite	21,881–17,761; 12,887–7899	912–882, 812–793
Hematite	21,978–16,340; 13,072–9524	560–550

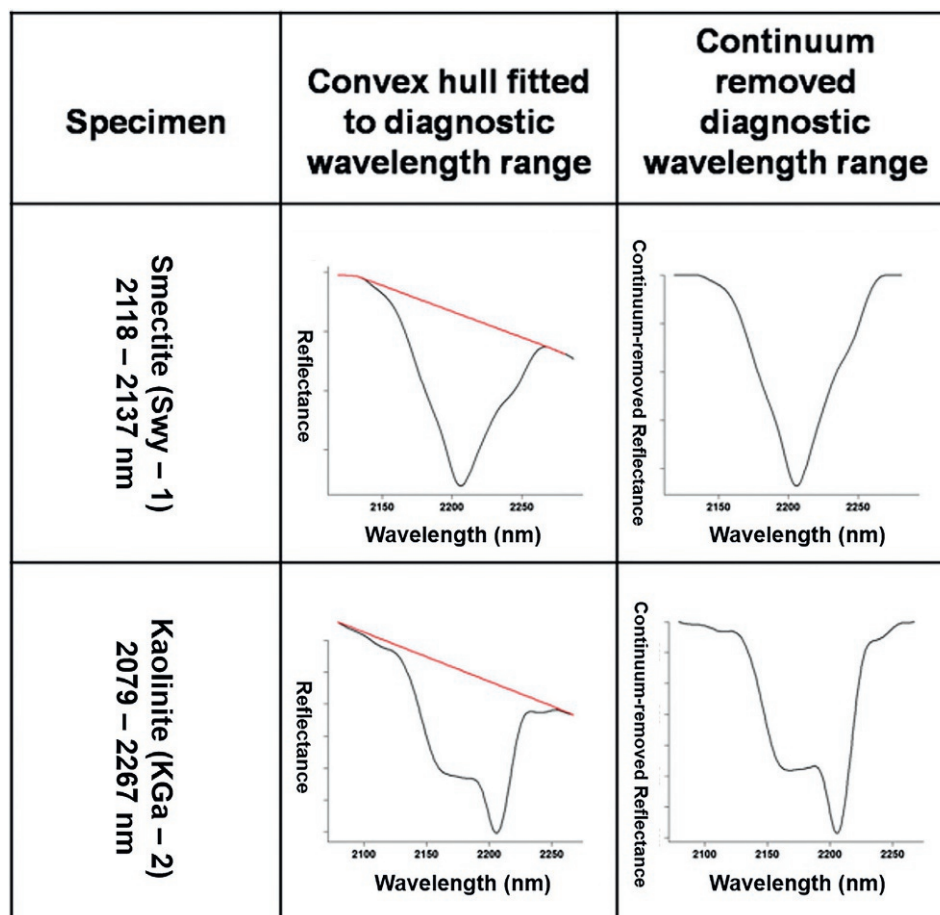


Fig. 3 Diagnostic wavelength ranges for smectite and kaolinite. To prepare for analyses, spectra need to be standardized by fitting a convex hull (red line). Soil spectra and reference spectra are then compared based on the continuum removed spectra. Sourced from Malone BP, Hughes P, McBratney AB, Minasny B (2014) A model for the identification of terrons in the Lower Hunter Valley, Australia. *Geoderma Regional* 1: 31–47.

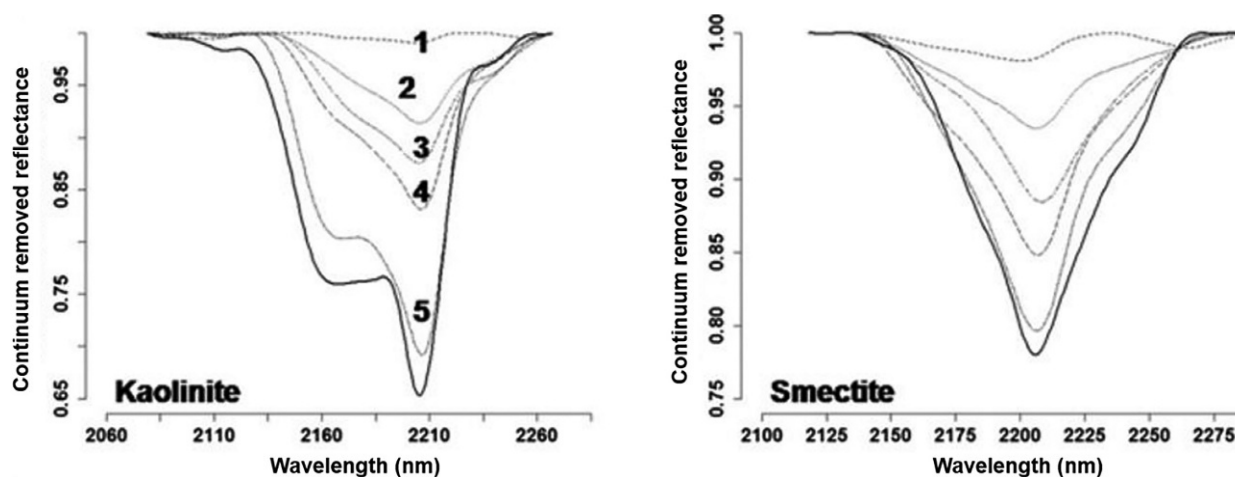


Fig. 4 Continuum removed soil spectra (dashed lines) with comparison to reference spectra at their given diagnostic wavelength ranges. The dashed lines represent (as numbered for kaolinite) the spectra for soils where relative mineral abundance is (1) the minimum, (2) first quartile, (3) second quartile, (4) third quartile, and (5) maximum. Partial figure sourced from Malone BP, Hughes P, McBratney AB, Minasny B (2014) A model for the identification of terrons in the Lower Hunter Valley, Australia. *Geoderma Regional* 1: 31–47.

form varies in complexity and ability of interpretation, and these features should be considered in detail in the exploratory stages of soil spectral model development.

Large soil spectral libraries are expanding as more spectral data are collected at a regional or national scale. An emerging method used for large data and the processing complex images, and which is capable of handling high-dimensional data is the deep learning model. One form is the convolutional neural network (CNN) which is widely used in handling high dimensional data, such as images. Ng et al. (2019) explored the use of the CNN model for predicting soil properties on a large vis-NIR and MIR spectral database. The architecture of the CNN model is quite flexible and can accept multi-dimensional input and provide multi-task predictions. The study reported that deep learning on an extensive MIR spectra library could predict fundamental soil properties (texture, organic and inorganic C, CEC, and pH) very accurately with $R^2 > 0.95$. Furthermore, only one model needed to be trained to provide predictions of multiple properties. Data fusion can also improve the prediction of soil properties.

Soil properties

Soil physical properties

Soil physical properties regulate the movement, retention and availability of water and nutrients to plants, and determine the flow of heat and air into, within and out of the soil. Despite the importance of these properties, a common feature in their measurement is that they often require considerable time to collect the data, are expensive to acquire, and require very specialized equipment. Consequently, alternative and less costly approaches with pedotransfer functions, which exploit easier to acquire attributes coupled with statistical modelling, have been proposed and have demonstrated their utility in numerous applications. Similarly, soil spectral modelling of these attributes has been demonstrated. Soriano-Disla et al. (2014) provide an all-encompassing review of the applicability of vis-NIR and MIR spectroscopy for the prediction of various soil properties such as texture, aggregation and soil water content. Stenberg et al. (2010) provide a similar account but focus specifically on vis-NIR soil spectroscopy. It is recommended that readers review those contributions and the references cited therein to appreciate not only the breadth and variation of studies that have been undertaken, but also the comparative assessment of their model performances.

In general, MIR spectroscopy outperforms vis-NIR for most soil physical properties. Soil spectral models based on properties like clay, sand and water content using either vis-NIR or MIR spectra are generally comparable ($R^2 = 0.7$ – 0.8). Soil aggregation as measured by wet sieving also been reported as being well-predicted, for example, mean aggregate diameter and mean weight diameter with R^2 values between 0.66 and 0.79. Despite seemingly good modelling outcomes reported for soil physical properties mentioned above, there are also corresponding examples where outcomes have not been so good such as bulk density and water retention at field capacity. There are many potential causes for this variability, with some stemming from the data used and the extent of soil variability captured, to differences in models and how the data are processed and handled. However, the more probable reason is that soil spectral calibrations of soil physical properties are really about the utilization of variables related to these properties that are known to be well-predicted, such as clay, soil carbon, and mineralogy. This co-relationship is acknowledged in the reporting of very good results in the prediction of COLE and Atterberg limits with vis-NIR spectroscopy and reported R^2 of 0.84 and 0.48–0.71, respectively (Rehman et al., 2020). Utilizing both soil spectroscopy and pedotransfer functions together, i.e. soil spectral inference (McBratney et al., 2006), is seen as a more appropriate way to estimate some soil physical properties, particularly those based on soil pore-space relationships, e.g. volumetric water retention, hydraulic conductivity and penetration resistance, and bulk density. This somewhat echoes the conclusions from Minasny et al. (2008) that soil physical properties based on soil pore space relations are not in general well predicted directly with infrared spectroscopy, therefore more focus should be given to soil spectral inference systems to improve what can be predicted well and use those estimates in existing or, newly calibrated pedotransfer functions.

Soil chemical properties

Near- and mid-infrared spectroscopy is mostly used to predict soil chemical attributes. These properties affect soil biological activity and nutrient dynamics in the soil and have been used as soil quality assessment indicators. Soil chemical properties include pH, organic carbon, nitrogen, nutrients, as well as cation exchange capacity.

Numerous studies had been carried out to measure soil carbon. This interest is due to the recognition that storing additional carbon in the soil serves as one of the options to mitigate anthropogenic climate change. Soil organic carbon is a component of soil organic matter that affects the physical, chemical, and biological properties of soil, influencing soil structure, water and nutrient retention. In addition to total carbon, organic and inorganic soil carbon predictions were also found to be successful, with R^2 ranging from 0.94–0.98 (McCarty et al., 2002). The use of infrared spectroscopy for the prediction of various carbon fractions, including labile and recalcitrant carbon fractions ($R^2 = 0.89$ – 0.97), had been proven to be successful (Zimmermann et al., 2007). In the MIR range, organic carbon may also be identified by absorption peaks at approximately 2920 and 1230 cm^{-1} (Zimmermann et al., 2007), while carbonates may be identified by strong and numerous absorption bands at 1450, 1060, 880 and 700 cm^{-1} (Clark et al., 1990) (as shown in Fig. 2). In the NIR range, organic carbon can be related to absorption peaks at approximately 600, 1910 and 2100 nm, while carbonates may be identified by peaks at 2300–2350 nm (Clark et al., 1990).

Although soil pH is non spectrally active, it is one of the soil properties that is predicted quite well ($R^2 = 0.57$ – 0.79) because of its relationship with exchange sites related to clay minerals and organic matter (Soriano-Disla et al., 2014).

Infrared spectroscopy has been used successfully to predict exchangeable cation related properties. While cation exchange capacity (CEC) is not a spectrally active component, it has a strong correlation with clay mineralogy and soil organic matter content (Stenberg et al., 2010), particularly absorbance of peaks near 2926–2850 cm^{-1} , and carboxylate species 1633–1413 cm^{-1} . Greater prediction accuracy for certain cations (Ca^{2+} , Mg^{2+} , and K^{+}) with R^2 of 0.70–0.84 compared to Na^{+} ($R^2 = 0.32$) was most likely because those cations are dominant on the exchange sites.

Infrared spectroscopy has also been used to predict nutrients in the soil, particularly the nitrogen (N), phosphorus (P), and potassium (K) concentrations in the soil. However, most studies have shown that only the total amount of nutrients can be well predicted as this relates to soil minerals and organic matter. Available nutrients cannot be well-predicted because they are determined by extraction and solution from soil. Integration of the prediction of plant nutrients and fertilizer recommendations maximizes the nutrient management in the agroecosystems.

Soil biological properties

The correlation of soil biological properties and organic carbon enables the biological properties to be successfully predicted using spectral features. Infrared spectroscopy has been used for the rapid assessment of a range of biological properties, including microbial biomass, soil respiration, microbial groups, and soil enzymatic activities. No spectral contribution is expected from the microbial biomass because it represents roughly less than 5% of the total organic carbon content (Janik et al., 1998). Successful predictions of biological properties were most likely due to the correlation of biological properties with organic matter or other soil properties that have spectral signatures.

Chang et al. (2001) obtained R^2 of 0.6 for biomass C using NIR. Meanwhile, Janik et al. (1998) and Ludwig et al. (2015) reported greater accuracy for the microbial biomass C predictions with MIR ($R^2 = 0.69$ and 0.79 respectively). In addition to biomass C, biomass N ($R^2 = 0.79$) and P ($R^2 = 0.62$) were also determined with moderate accuracy by Ludwig et al. (2015).

A better prediction was reported by Chang et al. (2001) using NIR for microbial respiration than the soil microbial biomass with R^2 of 0.72, 0.66, and 0.82 for mineralizable N, total respiration rate and basal respiration rate (BSR), respectively. van Groenigen et al. (2003) reported greater accuracy for mineralizable N using NIR ($R^2 = 0.46$) compared to MIR ($R^2 = 0.21$) in flooded soils. Similarly, Ludwig et al. (2015) reported an acceptable accuracy of $R^2 = 0.63$ for BSR using MIR.

Microbial biomarkers have also been utilized as a biological indicator of change in the microbial community because the size of microbial biomass itself is not sufficient to explain the effect of certain management regimes. Zornoza et al. (2008) successfully predicted several variables related to soil microbial groups based on PLFA biomarkers: bacteria ($R^2 = 0.93$), G+ bacteria ($R^2 = 0.91$), actinomycetes ($R^2 = 0.92$), VAM fungi ($R^2 = 0.91$) and PLFAs ($R^2 = 0.91$).

The presence of soil enzymes can be linked to soil physico-chemical properties. Although 500 enzymes might be present, only a handful have been identified and measured to date. Reeves et al. (2000) achieved an R^2 within the range of 0.8 with NIR to predict four enzymes (dehydrogenase, phosphatase, arylsulfatase and urease).

Soil contaminants

Contaminants include a wide range of natural and synthetic materials that can be found in soil. Given that exposure to contaminants leads to detrimental effects on human health and ecosystems, there is a need to develop rapid, low-cost methods to assess the spatial and temporal variation of soil contaminants. The application of spectroscopy in detecting contaminants is beneficial for rapid screening to determine whether contamination exceeds a certain threshold and if remediation is required. The prediction of organic contaminants can be directly related to the absorption peaks in the spectra because of sensitivity of the IR method to organic materials. Nonetheless, these absorbances of the organic contaminants overlapped with other organic and inorganic components that exist in the soil throughout the electromagnetic spectrum. Several examples of organic pollutants include petroleum hydrocarbons, microplastics and pesticides.

Absorbance peaks at 2222–2439 nm and 3000–2600 cm^{-1} in the NIR and MIR region have been used to predict the total petroleum hydrocarbon concentration (Forrester et al., 2013) with high accuracy. The peaks in the MIR region are related to two $-\text{CH}_3$ peaks (2950 and 2730 cm^{-1}), which were absent in the natural soil organic matter absorption $-\text{CH}_2$ region at 2930–2850 cm^{-1} . A similar range in the NIR region at 2250–2350 nm was reported by Chakraborty et al. (2010) for the prediction of total petroleum hydrocarbon, which achieved good accuracy in moist intact soil. The absorption at 2200 nm is due to the O—H bonds in the clay minerals, such as kaolinite, illite, and smectite, while absorptions between 2250 and 2450 nm were associated with the C—H bonds in the organic matter and carbonates (Clark et al., 1990).

Absorptions between 1130 and 1720 nm in the NIR range have been reported in various studies (e.g. Sato et al., 2003) in the identification of plastic polymers, such as polyethylene terephthalate (PET) and low density polyethylene (LDPE). Absorbance at 1210, 1420 and 1730 nm is due to C—H (CH_2) first and second overtones and C—H deformation (Sato et al., 2003), while absorbance at 1660 nm is related to overtone stretching of aromatic C—H groups which is present in PET but not LDPE.

Hermansen et al. (2020) successfully utilized NIR for predicting glyphosate (an active ingredient in herbicides) sorption in soil ($R^2 = 0.93$). Bengtsson et al. (2007) also successfully utilized NIR as a tool to predict other pesticide sorption in soil, particularly lindane ($R^2 = 0.85$) and linuron ($R^2 = 0.84$), despite no specific wavelengths being identified.

Although inorganic contaminants, such as heavy metals are not spectrally active, diagnostic screening is made possible due to the relationship with iron oxides, clay, and organic matter. Gholizadeh et al. (2015) using the vis–NIR range detected heavy metal

concentrations (Cu, Pb, Zn and Cd), which had low to moderate to correlations with clay and/or organic matter content, whereas Mn did not have a dependence on clay and organic matter. Similarly, Kemper and Sommer (2002) achieved good R^2 accuracy for the prediction of As (0.84), Fe (0.72), Hg (0.96), Pb (0.95) and Sb (0.93).

Lower cost - miniaturized spectrometer

Lower cost and miniaturized spectrometers with a limited NIR spectral range are now widely available and can potentially enable NIR measurements of soil properties to be used more broadly. For example, a miniaturized Michelson interferometer can be built using monolithic micro-electro-mechanical systems (MEMS). A case study in Australia demonstrates that a miniaturized spectrometer operating at 1250–2500 nm provides comparable accuracy with the complete vis–NIR spectrometer (400–2500 nm) in predicting soil pH, CEC and exchangeable Ca and Mg ($R^2 > 0.63$ –0.78), and slightly less accurate predictions of total carbon, sand, and clay ($R^2 > 0.70$ –0.80) (Tang et al., 2020).

Conclusions

Soil spectroscopy is a viable measurement technology with untapped possibilities. Soil reflectance in the visible-, near- and mid-infrared offers rapid, cheap, and relatively accurate predictions for several soil properties. A single infrared spectrum can offer the prediction of numerous soil properties. The MIR spectra have been shown to provide signatures that can be used to derive many physical, chemical and biological properties. The mechanisms for the prediction of soil properties are due to the direct and indirect interaction between the soil properties and the organic matter and mineral components of soil. Infrared can predict fundamental soil properties related to the soil matrix or solid constituents and soil solution concentrations in equilibrium with the solid phase. Soil properties that do not have active spectral components can be predicted if they are correlated to the mineral or organic components of the soil. However, soil spectroscopy has been mainly used in research studies and has not been widely used as a routine analysis method in agriculture.

References

- Bengtsson S, Berglöf T, and Kylin H (2007) Near infrared reflectance spectroscopy as a tool to predict pesticide sorption in soil. *Bulletin of Environmental Contamination and Toxicology* 78(5): 295–298.
- Bowers SA and Hanks RJ (1965) Reflection of radiant energy from soils. *Soil Science* 100(2).
- Chakraborty S, Weindorf DC, Morgan CLS, Ge YF, Galbraith JM, Li B, and Kahlon CS (2010) Rapid identification of oil-contaminated soils using visible near-infrared diffuse reflectance spectroscopy. *Journal of Environmental Quality* 39(4): 1378–1387.
- Chang CW, Laird DA, Mausbach MJ, and Hurburgh CR (2001) Near-infrared reflectance spectroscopy-principal components regression analyses of soil properties. *Soil Science Society of America Journal* 65(2): 480–490.
- Clark RN, King TV, Klejwa M, Swayze GA, and Vergo N (1990) High spectral resolution reflectance spectroscopy of minerals. *Journal of Geophysical Research - Solid Earth and Planets* 95(B8): 12653–12680.
- Clark RN, Swayze GA, Livo KE, Kokaly RF, Sutley SJ, Dalton JB, McDougal RR, and Gent CA (2003) Imaging spectroscopy: Earth and planetary remote sensing with the USGS Tetracorder and expert systems. *Journal of Geophysical Research, Planets* 108(E12).
- Forrester ST, Janik LJ, McLaughlin MJ, Soriano-Disla JM, Stewart R, and Dearman B (2013) Total petroleum hydrocarbon concentration prediction in soils using diffuse reflectance infrared spectroscopy. *Soil Science Society of America Journal* 77(2): 450–460.
- Gadsden JA (1975) *Infrared Spectra of Minerals and Related Inorganic Compounds*. London: Butterworths.
- Gholizadeh A, Borůvka L, Saberioon MM, Kozák J, Vašát R, and Němeček K (2015) Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features. *Soil & Water Research* 10(4): 10.
- Hermansen C, Norgaard T, Wollesen de Jonge L, Moldrup P, Müller K, and Knadel M (2020) Predicting glyphosate sorption across New Zealand pastoral soils using basic soil properties or vis–NIR spectroscopy. *Geoderma* 360: 114009.
- Herschel W (1800) Observations of the powers of the prismatic colours to heat and illuminate objects. *Philosophical Transactions of the Royal Society of London* 90: 225.
- Hunt JM, Wisherd MP, and Bonham LC (1950) Infrared absorption spectra of minerals and other inorganic compounds. *Analytical Chemistry* 22(12): 1478–1497.
- Janik LJ, Merry RH, and Skjemstad JO (1998) Can mid infrared diffuse reflectance analysis replace soil extractions? *Australian Journal of Experimental Agriculture* 38(7): 681–696.
- Kemper T and Sommer S (2002) Estimate of heavy metal contamination in soils after a mining accident using reflectance spectroscopy. *Environmental Science & Technology* 36(12): 2742–2747.
- Ludwig B, Sawallisch A, Heinze S, Joergensen RG, and Vohland M (2015) Usefulness of middle infrared spectroscopy for an estimation of chemical and biological soil properties – Underlying principles and comparison of different software packages. *Soil Biology and Biochemistry* 86: 116–125.
- Malone BP, Hughes P, McBratney AB, and Minasny B (2014) A model for the identification of terrons in the Lower Hunter Valley, Australia. *Geoderma Regional* 1: 31–47.
- McBratney AB, Minasny B, and Viscarra Rossel R (2006) Spectral soil analysis and inference systems: A powerful combination for solving the soil data crisis. *Geoderma* 136(1): 272–278.
- McCarty GW, Reeves JB, Reeves VB, Follett RF, and Kimble JM (2002) Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. *Soil Science Society of America Journal* 66(2): 640–646.
- Minasny B, McBratney AB, Tranter G, and Murphy BW (2008) Using soil knowledge for the evaluation of mid-infrared diffuse reflectance spectroscopy for predicting soil physical and mechanical properties. *European Journal of Soil Science* 59(5): 960–971.
- Ng W, Minasny B, Montazerolghaem M, Padarian J, Ferguson R, Bailey S, and McBratney AB (2019) Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma* 352: 251–267.
- Reeves JB, McCarty GW, and Meisinger JJ (2000) Near infrared reflectance spectroscopy for the determination of biological activity in agricultural soils. *Journal of Near Infrared Spectroscopy* 8(3): 161–170.

- Rehman HU, Arthur E, Tall A, and Knadel M (2020) Estimating coefficient of linear extensibility using vis-NIR reflectance spectral data: Comparison of model validation approaches. *Vadose Zone Journal* 19(1): e20057.
- Robinson NJ and Kitching M (2016) The 3D distribution of phyllosilicate clay minerals in western Victoria. *Geoderma* 284: 152–177.
- Sato H, Shimoyama M, Kamiya T, Amari T, Šašić S, Ninomiya T, Siesler HW, and Ozaki Y (2003) Near infrared spectra of pellets and thin films of high-density, low-density and linear low-density polyethylenes and prediction of their physical properties by multivariate data analysis. *Journal of Near Infrared Spectroscopy* 11(4): 309–321.
- Soriano-Disla JM, Janik LJ, Rossel RAV, Macdonald LM, and McLaughlin MJ (2014) The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied Spectroscopy Reviews* 49(2): 139–186.
- Stenberg B, Viscarra Rossel RA, Mouazen AM, and Wetterlind J (2010) Visible and near infrared spectroscopy in soil science. In: Sparks DL (ed.) *Advances in Agronomy*. vol. 107, pp. 163–215. Burlington, MA, USA: Academic Press.
- Tang Y, Jones E, and Minasny B (2020) Evaluating low-cost portable near infrared sensors for rapid analysis of soils from South Eastern Australia. *Geoderma Regional* 20.
- van Groenigen JW, Muters CS, Horwath WR, and van Kessel C (2003) NIR and DRIFT-MIR spectrometry of soils for predicting soil and crop parameters in a flooded field. *Plant and Soil* 250(1): 155–165.
- Zimmermann M, Leifeld J, and Fuhrer J (2007) Quantifying soil organic carbon fractions by infrared-spectroscopy. *Soil Biology and Biochemistry* 39(1): 224–231.
- Zornoza R, Guerrero C, Mataix-Solera J, Scow KM, Arcenegui V, and Mataix-Beneyto J (2008) Near infrared spectroscopy for determination of various physical, chemical and biochemical properties in Mediterranean soils. *Soil Biology and Biochemistry* 40(7): 1923–1930.

Electron microscopic methods (TEM, SEM and energy dispersal spectroscopy)

Jian Zhao^{a,b} and Xia Liu^a, ^aInstitute of Coastal Environmental Pollution Control, Key Laboratory of Marine Environment and Ecology, Ministry of Education, Frontiers Science Center for Deep Ocean Multispheres and Earth System, Ocean University of China, Qingdao, P. R. China;

^bLaboratory for Marine Ecology and Environmental Science, Qingdao National Laboratory for Marine Science and Technology, Qingdao, China

© 2023 Elsevier Ltd. All rights reserved.

Introduction	576
Fundamental principles and instrumentation	576
Fundamental principles	576
Fundamental principles	576
Transmission electron microscope	577
Scanning electron microscope	578
Energy dispersal spectroscopy	578
Instrumentation	578
TEM	578
SEM	578
Strengths and limitations of TEM and SEM	579
Applications of electron microscopes in soil chemistry	580
Morphology of soil particles	580
Minerals	580
Soil organic matter	581
Elemental composition and crystalline phases of soil particles	581
SEM-EDS	581
TEM-EDS and SAED	582
HR-TEM	582
Soil organic matter-mineral interaction	583
Soil contamination and remediation	584
Electron microscope-based new technologies in soil chemistry	584
Scanning transmission electron microscope (STEM) and cryogenic electron microscopy (cryo-EM)	584
Environmental scanning electron microscope (ESEM)	586
SEM–Raman system	586
Aberration-corrected electron microscope (AC-EM)	587
Summary	587
References	587

Abstract

Electron microscopic methods are powerful tools for soil chemistry investigations. This chapter introduces the fundamental principles and instrumentation of the transmission electron microscope (TEM), scanning electron microscope (SEM), and energy dispersal spectroscopy (EDS). These techniques are important for the morphological characterization and analysis of the chemical composition of soil particles. The applications of these techniques in organic matter–mineral interaction and soil remediation are addressed. Moreover, electron microscope-based new technologies (e.g., scanning transmission electron microscope, cryo-electron microscope, environmental scanning electron microscope) are briefly introduced. Finally, an outlook for the application potentials of aberration corrected-electron microscope in soil chemistry is given.

Key points

- Fundamental principles and instrumentation of TEM and SEM are introduced;
- TEM and SEM are important for soil particle characterization and soil chemical composition analysis;
- Applications of TEM and SEM in SOM–mineral interaction and soil contamination/remediation are overviewed;
- Electron microscope-based new technologies could bring new breakthroughs in soil chemistry.

Nomenclature

AC-EM	Aberration-corrected electron microscope
ACF	Auto-correlation function
BSE	Backscattered electrons
Cryo-EM	Cryogenic electron microscopy
Cs	Spherical aberration
EDS	Energy dispersive X-ray analysis
EELS	Electron energy loss spectrometry
ESEM	Environment scanning electron microscope
FFT	Fast Fourier transform
HA	Humic acid
HNTs	Halloysite nanotubes
HR-TEM	High-resolution TEM
nZVI	Zero-valent iron
PCP	Pentachlorophenol
SAED	Selected area electron diffraction
SE	Secondary electrons
SEM	Scanning electron microscope
SOM	Soil organic matter
STEM	Scanning transmission electron microscope
TEM	Transmission electron microscope

Introduction

The nanoscale surface and structure of soil particles are rarely resolved by light microscopes due to their limited resolution (~200 nm). Electron microscopes were first developed in the 1930s, opening up the world of particles and materials at the nano-scale level. There are two main types of electron microscopes: transmission electron microscope (TEM) and scanning electron microscope (SEM). Energy dispersive X-ray analysis (EDS) is an important accessory to the electron microscope, which can provide additional information (e.g., elemental composition) of the detected particles. Nanoparticles are widely present in soils, including nano-scaled clay minerals, Fe (iron), Al (aluminium) and Mn (manganese hydroxide) oxides, and soil organic matter (SOM). These nanoparticles play important roles in maintaining soil functions. Therefore, it is necessary to have a thorough understanding of these soil nanoparticles by characterizing their physical properties, elemental composition, chemical reactivity, and environmental processes. In addition, soil is a main sink for environmental contaminants, and the interactions between soil particles and contaminants are important for understanding the fate of these contaminants. For example, iron oxides, such as ferrihydrite, goethite, and hematite in soil can adsorb heavy metals (e.g., Pb^{2+}) by surface complexation (Liu et al., 2018); SOM is an important sink for nonpolar organic pollutants (e.g., phenanthrene) (Wen et al., 2007). All the above properties and interactions can be understood better with the assistance of electron microscopes.

Therefore, this chapter aims to introduce electron microscopy (SEM, TEM, and EDS) applications in soil chemistry, including (i) fundamental principles and instrumentation; (ii) application for soil particle characterization; (iii) soil chemical composition analysis; (iv) SOM–mineral interaction; and (v) soil contamination and remediation. Finally, electron microscope-based new technologies that could be potentially used in soil chemistry are briefly introduced.

Fundamental principles and instrumentation**Fundamental principles****Fundamental principles**

Electron microscopes are different from light microscopes because they collect the signals from the interaction between electrons and the soil sample to generate an image. The wavelength (λ) of moving electrons (ignoring relativistic correction) is shown in Eq. (1):

$$\lambda = h / \sqrt{2meV_{acc}} \quad (1)$$

where h and V_{acc} are the Plank constant and the acceleration voltage for the electrons, respectively; m is the mass; and e is the charge of an electron. Based on Eq. (1), electron microscopes have much higher resolution (0.1 nm in theory) than optical microscopes which are limited by the wavelength of visible light. However, the resolution of electron microscopes (e.g., TEM) cannot achieve 1 Å because of a certain extent of aberration on the magnetic lenses. The resolution (δ) of a TEM is roughly calculated as Eq. (2):

$$\delta = 1.2 \sqrt[4]{C_s \lambda^3} \quad (2)$$

Thus, resolution is related to V_{acc} , that is a larger V_{acc} is still preferable for higher resolution. It is noted that higher resolution does not always provide true information for the soil and other samples. For example, for a specific clay mineral particle that consists of relatively light elements (e.g., O (oxygen), Mg (magnesium), Si (silica)), high-energy electrons could penetrate the particles and re-enter the substrate or other particles. The image would be ambiguous under this condition.

Transmission electron microscope

As shown in Fig. 1, transmitted electrons consist of directly transmitted electrons, elastically scattered electrons, and inelastically scattered electrons. The TEM mainly relies on the collection of directly transmitted electrons and inelastically scattered electrons through the sample to create an image. Elastically scattered electrons are used to analyze crystal structure. The direction of elastically scattered electrons changes but their wavelengths do not change. Inelastic scattering has a longer wavelength because an energy loss of the incident primary electron occurs. Currently, TEM is widely used for collecting morphological and structural information about soil particles.

Crystalline structure and chemical information are critical in soil chemistry. High-resolution TEM (HR-TEM) responds perfectly to these demands. The HR-TEM developed in the 1960s has been successfully used to image the atomic structure of samples directly. Different from TEM, the HR-TEM is a phase contrast image rather than amplitude imaging, because more than one electron beam is used (Deepak et al., 2015) (Fig. 2). There are two stages in the HR-TEM imaging process: (i) the formation of an object surface wave after the interaction between electron beams and soil sample; (ii) two Fourier transforms (Fig. 2). In detail, the object surface wave obtains the crystal structural information when the electron beam passes through the sample, and the diffraction wave forms on the back focal plane. The diffraction waves at all levels are recombined to obtain the image wave which retains the original phase, and the lattice fringe pattern is obtained at the phase surface.

In addition to high image resolution, another advantage of TEM is the combination of imaging and diffractometry, which can provide the crystal lattice structure of soil samples. Selected area electron diffraction (SAED) with TEM is a powerful method for characterizing the structure of soil samples, including perfect crystals and defect structures. The main principle of SAED is shown in Fig. 3. Diffracted electrons can be focused into a regular arrangement of diffraction spots on the back focal plane by modifying the electromagnetic lenses; they are recorded as electron diffraction patterns. The advantages of SAED include (i) extremely short wavelength (about 2 pm), (ii) strong atomic scattering, and (iii) the ability to examine tiny particles (about 10 nm^3) (Bendersky and Gayle, 2001).

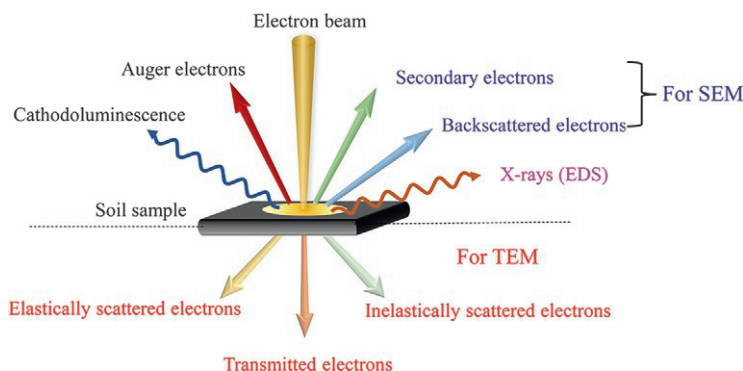


Fig. 1 Fundamental principles of electron microscopes for soil sample observation. The TEM mainly relies on the collection of transmitted electrons through the sample, while SEM collects the response of reflected electrons and photons to generate an image.

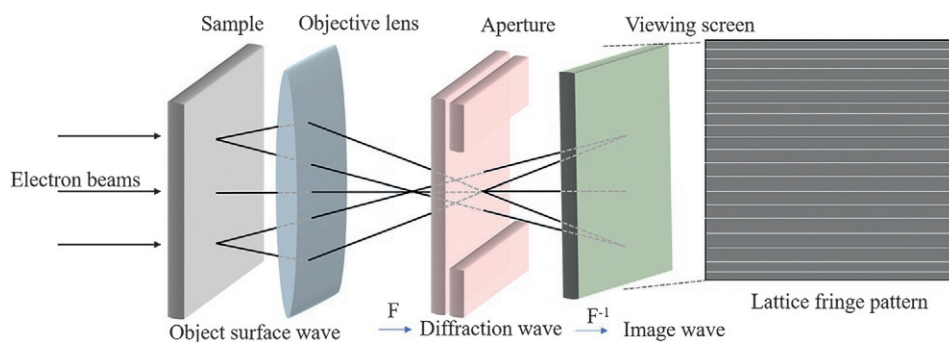


Fig. 2 Two Fourier transforms in the HR-TEM imaging process.

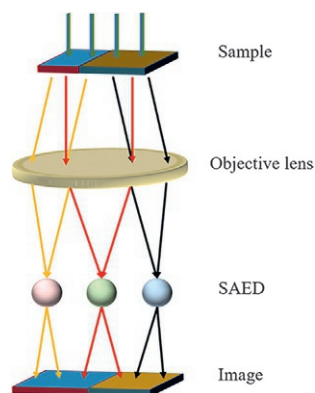


Fig. 3 Fundamental principles of the SAED imaging process.

Scanning electron microscope

The SEM collects reflected electrons and photons to generate an image (Fig. 1). It mainly uses secondary electrons (SE), back-scattered electrons (BSE), and characteristic X-rays for sample characterization. The SE with low energy can provide topographic and chemical information on the surface layer at 50 nm. The energy of BSE is equivalent to the primary incident electron, and thus originates from a depth of micrometers from the soil surface (Kwiecińska et al., 2019), which is beneficial for characterizing porosity, and local crystallographic structure of soil samples and other materials.

Energy dispersal spectroscopy

The EDS further widens the application of electron microscopes in soil characterization. Characteristic X-rays are produced when a beam of accelerated electrons bombard the sample surface. The rays can be collected at characteristic wavelengths and their intensities can be measured by EDS. Therefore, it can be used to identify chemical elements of the soil sample qualitatively and quantitatively (Girão et al., 2017). Electron energy loss spectrometry (EELS) is complementary to EDS. The EELS analyzes the loss of energy of electrons that pass through the thin sample (Fig. 1), thus providing information on atomic composition, bonding and oxidation state, surface properties, and distribution of near-neighbor atoms. It is noted that EELS is a much more sensitive technique for light elements, whereas EDS is more suitable for heavy elements (von Harrach et al., 2010).

Instrumentation

TEM

Fig. 4 shows the core components of TEM, including electron gun, anode, electrostatic lenses (condenser, objective, intermediate and project lenses), and a transmitted electron detection system (fluorescent screen, and charge coupled device camera). The electron gun produces a stream of electrons, and these dispersive electrons are focused on to a small, thin, parallel electron beam by condenser lenses. When it strikes the sample or its specimen, the transmitted electron beam is focused by an objective lens, and an image of the sample is created. The image is passed down the column and enlarged through the intermediate and projector lenses. The enlarged image is then projected onto a fluorescent screen or CCD cameras. After further post processing, the final TEM image can be saved (Kwiecińska et al., 2019).

SEM

Fig. 5 shows the core components of SEM, which consists of an electron gun, anode, condenser lenses, scan coil, objective lens, detectors (BSE, SE and X-ray), and amplifier. An electron gun produces a stable electron beam, which will diverge after passing through the anode plate. The first SEM system generally used tungsten "hairpin" or lanthanum hexaboride cathodes, whereas field emission sources are applied widely in the modern SEM; field emission sources can provide enhanced current and lower energy dispersion. A condenser lens is used to converge and collimate the electron beam into a relatively parallel stream. Scanning coils deflect the electron beam so that it can scan the sample surface. The SE and BSE detectors are used to detect different signals. The most frequent type of SE detector is the Everhart Thornley detector, which comprises a biased grid to attract the electrons, a scintillator to convert the SE signal, and a photomultiplier tube to amplify the signal. Semiconductor or scintillator plates usually determine BSE detector types. To enhance chemical or topographic contrast, two to four BSE detectors can be arranged at different solid angles. The X-ray detector counts the number of X-rays at a selected location on the sample. The collected signals are appropriately amplified and digitized. To generate a real-time image viewed on a display screen, the strict correspondence between scanned point and image point of the sample is needed. Finally, the acquired information is stored in a computer.

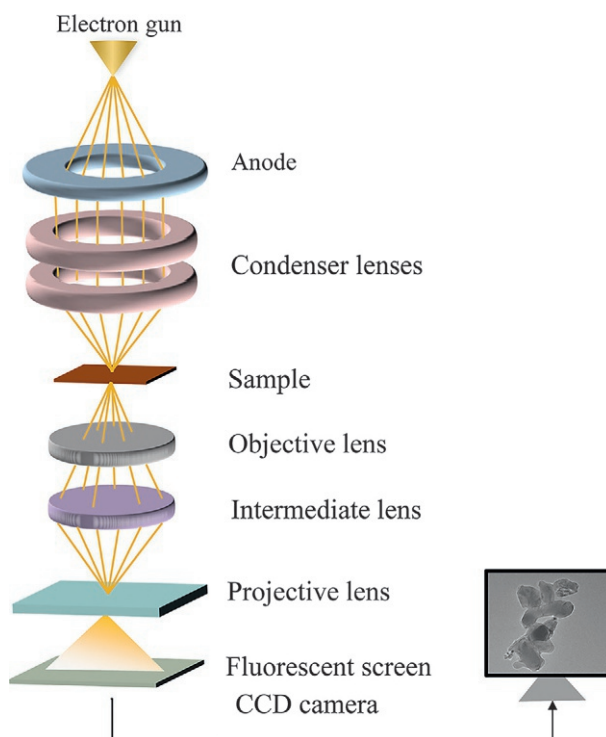


Fig. 4 Schematic diagram of the core components of TEM.

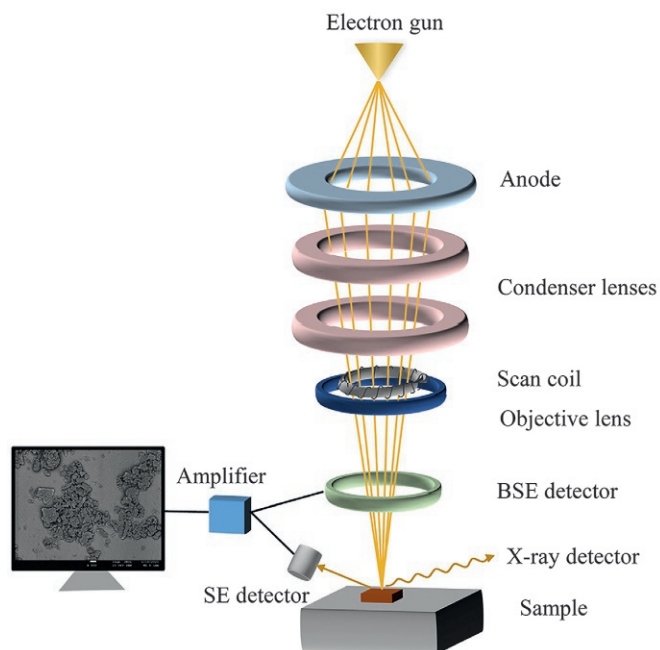


Fig. 5 Schematic diagram of the core components of SEM.

Strengths and limitations of TEM and SEM

Both TEM and SEM are powerful tools for characterizing soil samples due to their versatility and high spatial resolution compared with light microscope resolution of $\sim 0.2 \mu\text{m}$. As described above, electrons are used to acquire images of samples for both TEM and SEM. The SEM uses a specific set of coils to scan the beam in a raster-like pattern and collect the scattered electrons while TEM uses the transmitted electrons. Some core components of TEM and SEM are the same, including electron guns, electromagnetic lenses and electron apertures. In addition, both TEM and SEM can characterize the morphology of soil samples at the nanoscale. Despite the similarities, TEM and SEM have their unique strengths and limitations.

The TEM offers higher resolution (~ 0.2 nm) than SEM, and can even reach near-atomic resolution for HR-TEM. Additionally, TEM provides information on both element and compound structure (i.e., crystal structure, elemental information and morphology). However, despite the advantages of magnification and versatility, TEM has some limitations. First, TEM samples must be thin enough for electrons to pass through, generally below 150 nm, and in cases that require high-resolution imaging even below 30 nm. Therefore, sectioning is required for the thick soil and biological samples. Then, the samples must be analyzed in a vacuum because the electrons are easily scattered by other molecules in the air. This means that TEM cannot characterize wet samples or biological samples (e.g., proteins, dissolved organic matter); but this limitation can be overcome by the cryo-electron microscope. Furthermore, TEM is sensitive to vibration and electromagnetic fields and must be housed in an area that isolates them from possible exposure.

The SEM provides a 3-D image of the sample by imaging the original morphology of solid surfaces, whereas TEM images are 2-D projections of the sample from the inner structure. The resolution of SEM is about 1 nm which is less than for TEM, and SEM does not need the sample to be sectioned and it can also characterize samples as large as 100 mm. Similar to TEM, SEM cannot characterize wet or biological samples either, which can be overcome by the environmental SEM.

Clearly, there is no “better” technique; it depends on the type of sample that has been prepared and the information required. By combining the advantages of SEM and TEM, the morphological and structural information of soil samples can be understood better. Some limitations of SEM and TEM have been overcome at present by developing electron microscopy-based new technologies (see below). With these improvements, electron microscopes will continue to increase our understanding in soil chemistry.

Applications of electron microscopes in soil chemistry

Morphology of soil particles

The morphology of soil particle provides important information on understanding the shear strength, water-retention ability, and compressibility of soil. Electron microscopes can provide a variety of morphological information (e.g., shape, pore, roughness, and size) for soil particles including minerals present and SOM.

Minerals

Soil minerals occupy almost half of a soil sample. Iron and aluminum oxides are ubiquitously present in soil. These minerals play important roles in the fate, transport, and cycling of nutrients or contaminants in soil. Therefore, it is necessary to characterize these minerals in soil comprehensively. The SEM and TEM are very effective tools for morphological characterization. Information on the morphology of minerals using SEM and TEM can be found in numerous studies. Fig. 6 shows the SEM image for Fe- and Al-rich chlorites from the Strzegom pegmatite (Poland). The SEM provides clear topographic information on the chlorite surface. Most of the chlorites are triangular, prism-like shapes and commonly have a lateral size less than 50 μm . The thickness of each particle is about several microns, which can also be observed from this SEM image. The SEM is also effective in resolving the topographic information for large-size minerals, whereas TEM would be more suitable for the nanoscale minerals. Fig. 7 shows that halloysite nanotubes (HNTs) consist of a transparent central area that is tubular (Cheng et al., 2018). The internal diameter and length of HNTs range from 20 to 30 nm and 0.2–1.7 μm , respectively, and their shell thickness is about 15–20 nm, which could be clearly obtained from TEM imaging.

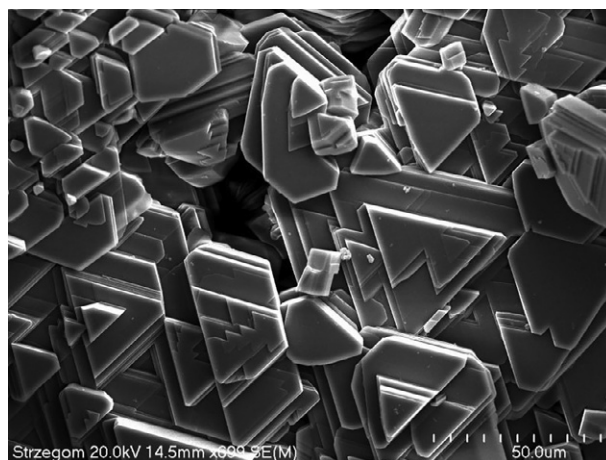


Fig. 6 The SEM image for Fe- and Al-rich chlorite from the Strzegom pegmatite. Image reproduced from the “Images of Clay Archive” of the Mineralogical Society of Great Britain & Ireland and The Clay Minerals Society (<https://www.minersoc.org/images-of-clay.html>).

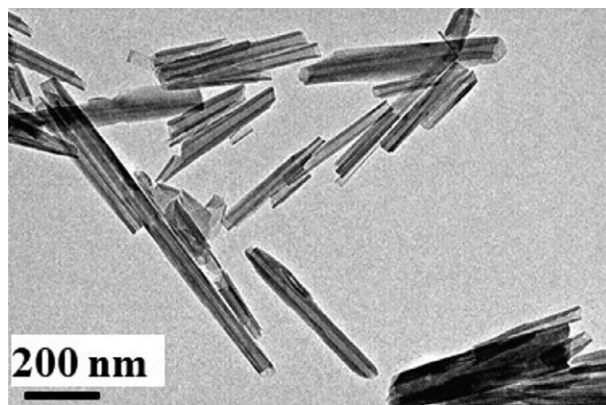


Fig. 7 Morphology of halloysite nanotubes as imaged by TEM. Reproduced with permission from Cheng R, Li H, Liu Z and Du C (2018) Halloysite nanotubes as an effective and recyclable adsorbent for removal of low-concentration antibiotics ciprofloxacin. *Minerals* 8: 387.

Soil organic matter

The SOM occupies around 30% (by mass) of typical surface soils; it has a great influence on soil properties (e.g., water retention, the ability to chelate and complex ions, pH) and environmental processes of various elements in soil. Commonly, SOM molecules (e.g., humic acid (HA)) are in the form of long-chain aliphatic structure in water. Fig. 8A shows the aggregation of HA molecules, and a closely-woven flake network structure and a thickened sheet structure with linear-like protrusions can also be observed (Qi et al., 2004), probably from the drying pre-treatment before SEM imaging. The TEM observation in the work of Monreal et al. (2010) showed that SOM molecules and aggregates were attached to the surface of minerals (Fig. 8B), which may be caused by electrostatic attraction and complexation due to abundant oxygen-containing groups in SOM.

Elemental composition and crystalline phases of soil particles

The EDS is a highly specialized technique that can be combined with both TEM and SEM; it is widely used in the analysis of elemental composition of soil particles including both minerals and SOM. In addition, HR-TEM and SAED are powerful for characterizing structural properties of soil particles, including perfect crystals and defects.

SEM-EDS

The SEM images of quartz, sphalerite and magnetite are shown in Fig. 9. The lateral sizes of these particles were about 40, 25, and 200 μm , respectively. After EDS mapping, the main elements on the surface of these minerals were then detected (Drif et al., 2018). Clearly, Si and O were present on the surface of quartz, while Zn and S were on sphalerite (Fig. 9A). For magnetite, Fig. 9B shows that O and Fe were distributed homogeneously.

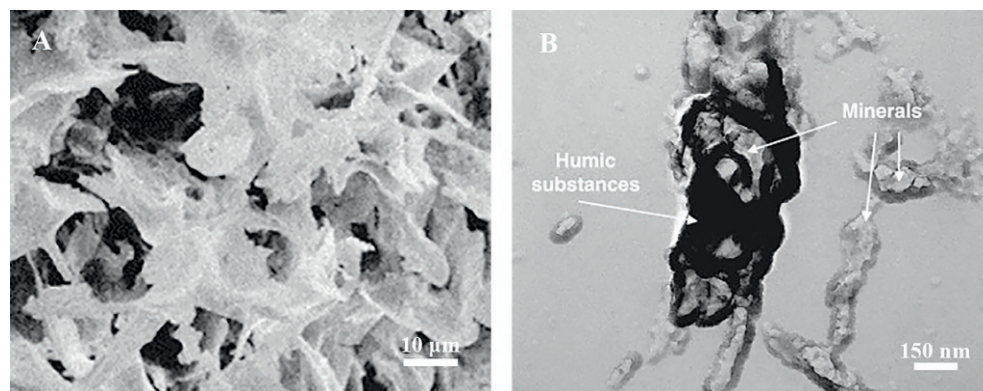


Fig. 8 The SEM and TEM images of SOM. (A) SEM of SOM aggregates; and (B): TEM for the SOM–minerals interaction. (A) Reproduced with permission from Qi BC, Aldrich C and Lorenzen L (2004). Effect of ultrasonication on the humic acids extracted from lignocellulose substrate decomposed by anaerobic digestion. *Chemical Engineering Journal* 98: 153–163; (B) reproduced with permission from Monreal CM, Sultan Y and Schnitzer M (2010) Soil organic matter in nano-scale structures of a cultivated Black Chernozem. *Geoderma* 159: 237–242.

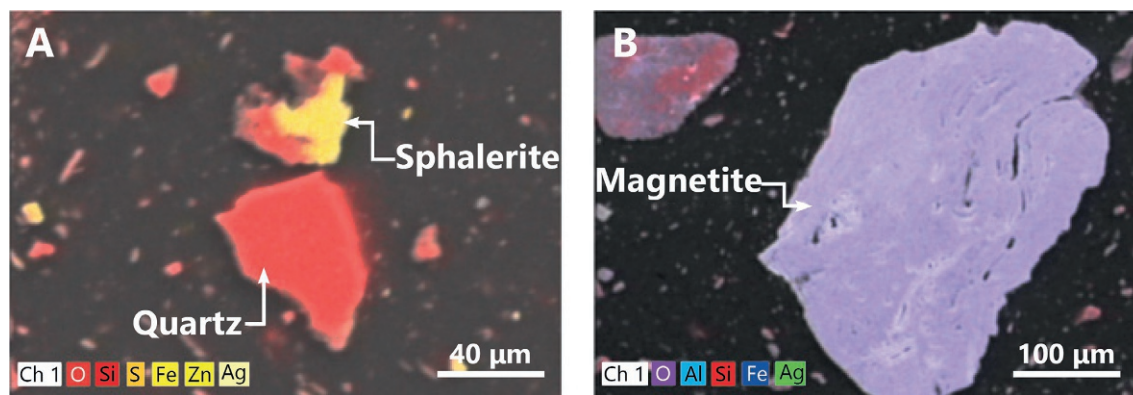


Fig. 9 The SEM-EDS of minerals. The SEM and EDS-mapping images of specific minerals in the Zgounder mine tailings composite sample. (A) Quartz and sphalerite; (B) magnetite. Reproduced from Drif B, Taha Y, Hakkou R and Benzaazoua M (2018) Recovery of residual silver-bearing minerals from low-grade tailings by froth flotation: The case of Zgounder mine Morocco, *Minerals* 8: 273.

TEM-EDS and SAED

Kaolinite is a typical mineral in soil. From TEM imaging, the observed kaolinite particles are at the nanoscale, in the form of a laminar structure of pseudo-hexagonal platelets (Suzuki et al., 2013) (Fig. 10A). The EDS-mapping analysis shows that the ratio of the atomic concentration of Al/Si is close to 1.0, which is consistent with the theoretical chemical composition of kaolinite ($\text{Al}_2\text{Si}_2\text{O}_5(\text{OH})_4$) (Fig. 10B). Some minerals (e.g., maghemite and magnetite) have the same elemental composition but different crystal structures, which could be further identified by SAED. For kaolinite, the well crystallized structure was demonstrated based on the its electron diffraction pattern using SAED (Fig. 10C). Therefore, it is concluded that an unknown mineral could be identified by combining SEAD with EDS.

HR-TEM

In addition to SAED, HR-TEM is widely used for crystalline phase analysis. To distinguish the specific minerals of soil nanomagnets, Ahmed and Maher (2018) applied HR-TEM to identify magnetic nanoparticles in fossil soil (paleosol) (Fig. 11). Fig. 11A shows the visible lattice fringe of magnetic nanoparticle using HR-TEM. A structural fingerprinting approach was used to characterize magnetic nanoparticles in the soil further. The structural fingerprinting approach utilizes crystallographic processing of HR-TEM images in which crystal structure factors were obtained using Fourier analysis. The auto-correlation function (ACF) can be used to detect the periodicity of a single ferrite crystal and to reveal composite crystals. As described in the HR-TEM imaging principle (Fig. 2), the

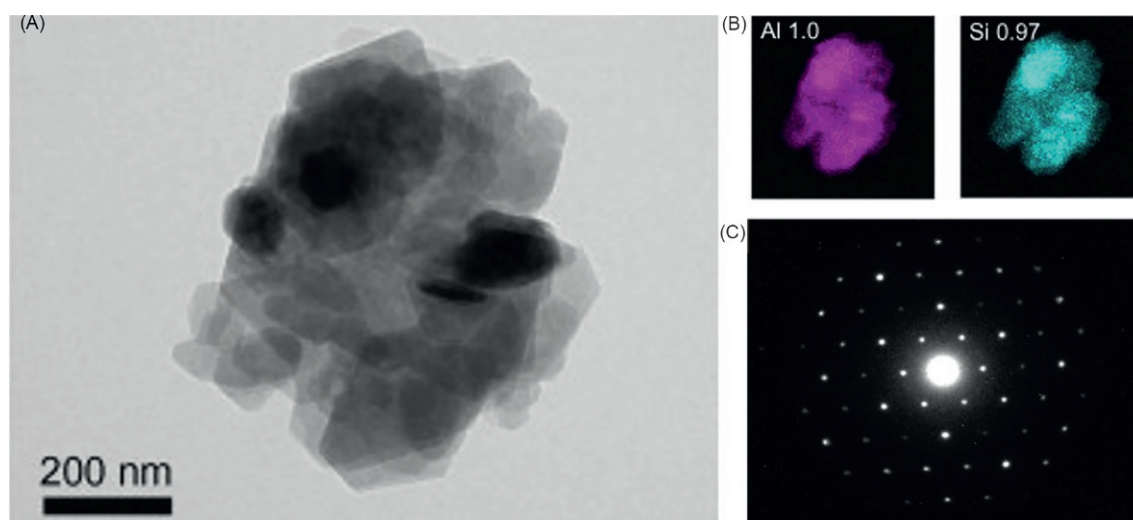


Fig. 10 The TEM analysis of kaolinite. (A) TEM image; (B) EDS mapping images; and (C) SAED pattern. The numbers next to chemical symbols represent the relative atomic concentration standardized to Al. Reproduced with permission from Suzuki T, Nakamura A, Niinae M, et al. (2013) Lead immobilization in artificially contaminated kaolinite using magnesium oxide-based materials: Immobilization mechanisms and long-term evaluation. *Chemical Engineering Journal* 232: 380–387.

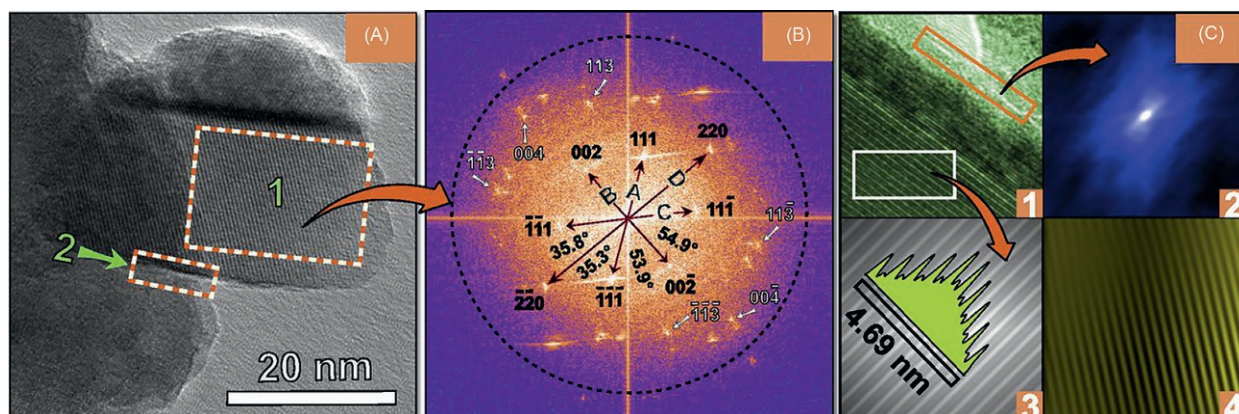


Fig. 11 The HR-TEM observation of magnetite particles in a paleosol. (A) HR-TEM image; (B) FFT power spectra and lattice indexing; (C) ACF analysis. Reproduced with permission from Ahmed IA and Maher BA (2018) Identification and paleoclimatic significance of magnetite nanoparticles in soils. *Proceedings of the National Academy of Sciences* 115: 1736–1741.

imaging process of a TEM contained twice the FT, Fourier transform (F and F^{-1}). Fast Fourier transform (FFT) was performed by a discrete finite sum in a computer, which greatly decreased the computation time. After FFT and ACF analysis of the HR-TEM image (Fig. 11A), FFT power spectra and lattice plane distances of the magnetic nanoparticles could be acquired (Fig. 11B and C), and the magnetite crystals were successfully identified in this soil sample.

Soil organic matter-mineral interaction

The SOM commonly serves as a redox mediator in mineral transformation in the natural environment. For example, abundant oxygen-containing functional groups (e.g., -OH, -COOH) on HA can promote its interaction with minerals, such as adsorption and transformation. TEM-EDS could be the first choice for the investigation of HA-mineral interaction at the nano-scale level. Here we given an example of the role of HA in siderite transformation (Xing et al., 2020). Fig. 12 shows the TEM image of siderite after incubation with HA, in which spherical particles (area A), clusters (area B), and rod-like structures (area C) were observed. Combined with EDS-mapping, these nanoparticles were identified as siderite, ferrihydrite, and goethite (FeOOH), respectively. The elements Fe and C were the major elements for siderite (area A) and ferrihydrite (area B), whereas element C was rare in goethite (area C). This result implied that the newly formed goethite did not adsorb HA. The expulsion of HA indicated that the bonds between HA and Fe atoms were broken during the transformation of siderite to goethite. Moreover, phosphate can be adsorbed on both siderite and the newly-formed goethite, which is clearly shown in Fig. 12.

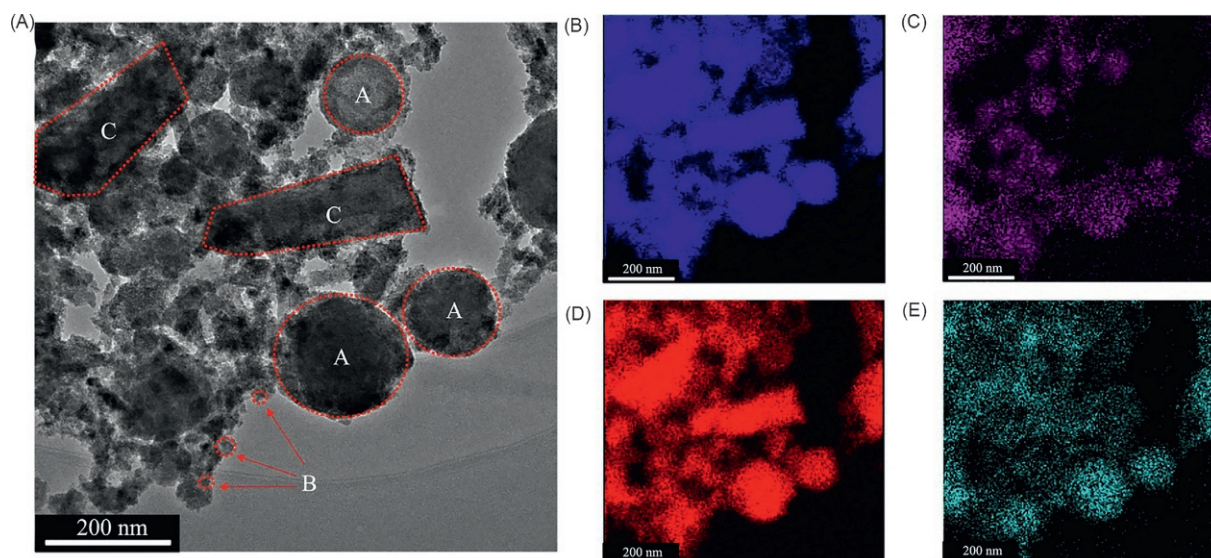


Fig. 12 The TEM-EDS mapping of siderite after incubating with HA and phosphate. Before observation, the samples were lyophilized for 12 h after aging for 7 days. Reproduced with permission from Xing B, Graham N and Yu W (2020) Transformation of siderite to goethite by humic acid in the natural environment. *Communications Chemistry* 3(1): 1–11.

Soil contamination and remediation

Soil contamination by heavy metals raises increasing concerns worldwide. Quick and reliable detection of the species and abundance of heavy metals are crucial for soil monitoring and remediation. The EDS is a valid method to analyze the distribution of heavy metals on both minerals and SOM. Lime was applied by Hamid et al. (2020) to remediate Cd and Pb that co-contaminated soil. From SEM imaging, lime showed irregular structure after amendment (Fig. 13). The EDS mapping demonstrated the distribution of Cd and Pb on the lime, indicating that Cd and Pb were removed by lime.

Moreover, TEM-EDS analysis was performed to investigate the underlying mechanisms of soil remediation by zero-valent iron (nZVI) coupled with the increase in rice production (Liu et al., 2020). The nZVI could be taken up by rice roots, and TEM is more suitable for observing the distribution of nZVI in root tissues than SEM. Fig. 14B shows that numerous nanoparticles are present in rice root intercellular spaces after amendment with nZVI for 30 days in pentachlorophenol (PCP) contaminated soil, whereas no particles can be detected in rice roots without the amendment of nZVI (Fig. 14A). The EDS result demonstrates that these nanoparticles are rich in Fe (Fig. 14B), which indicates the uptake of nZVI or other Fe-based species in roots. It is noted that the specific types of Fe-based minerals cannot be confirmed by only TEM-EDS because nZVI may undergo biotransformation during uptake and transport in plants. The HR-TEM is suitable to confirm further the specific mineral types (e.g., nZVI or other produced Fe-based minerals) in soil. For the case in Fig. 13, if the heavy metals (e.g., Cd, Pb) were mineralized during remediation, the minerals formed could also be identified by HR-TEM.

Electron microscope-based new technologies in soil chemistry

Over the last few years, electron microscope-based new technologies have been developed, mainly in material and biological fields. Although they have not been widely applied in soil chemistry, potential uses have been observed based on currently limited applications. Hence, we provide typical examples on the applications of some important electron microscope-based new technologies in soil chemistry.

Scanning transmission electron microscope (STEM) and cryogenic electron microscopy (cryo-EM)

The STEM, which combines the advantages of TEM and SEM, is a very powerful instrument for the characterization of soil samples. Similar to TEM, STEM requires very thin samples or their ultrathin sections. The principal advantage over TEM is that STEM is equipped with scanning coils that can collect simultaneously signals from both reflected and transmitted electrons. Fig. 15A and B shows the topological structure and nanometer scale spatial resolution images provided by STEM of the organic–organic interface of

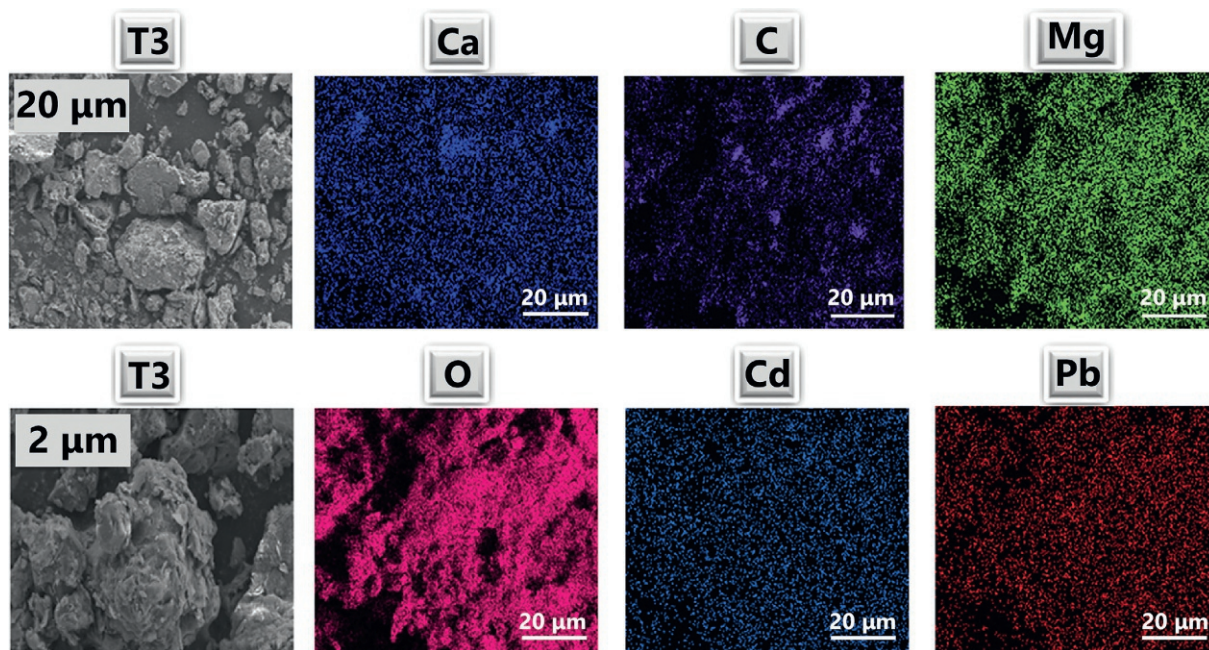


Fig. 13 The SEM images and EDS analysis of lime after application to a heavy metal-contaminated soil. Reproduced with permission from Hamid Y, Tang L, Hussain B, et al. (2020). Efficiency of lime, biochar, Fe containing biochar and composite amendments for Cd and Pb immobilization in a co-contaminated alluvial soil. *Environmental Pollution* 257: 113609.

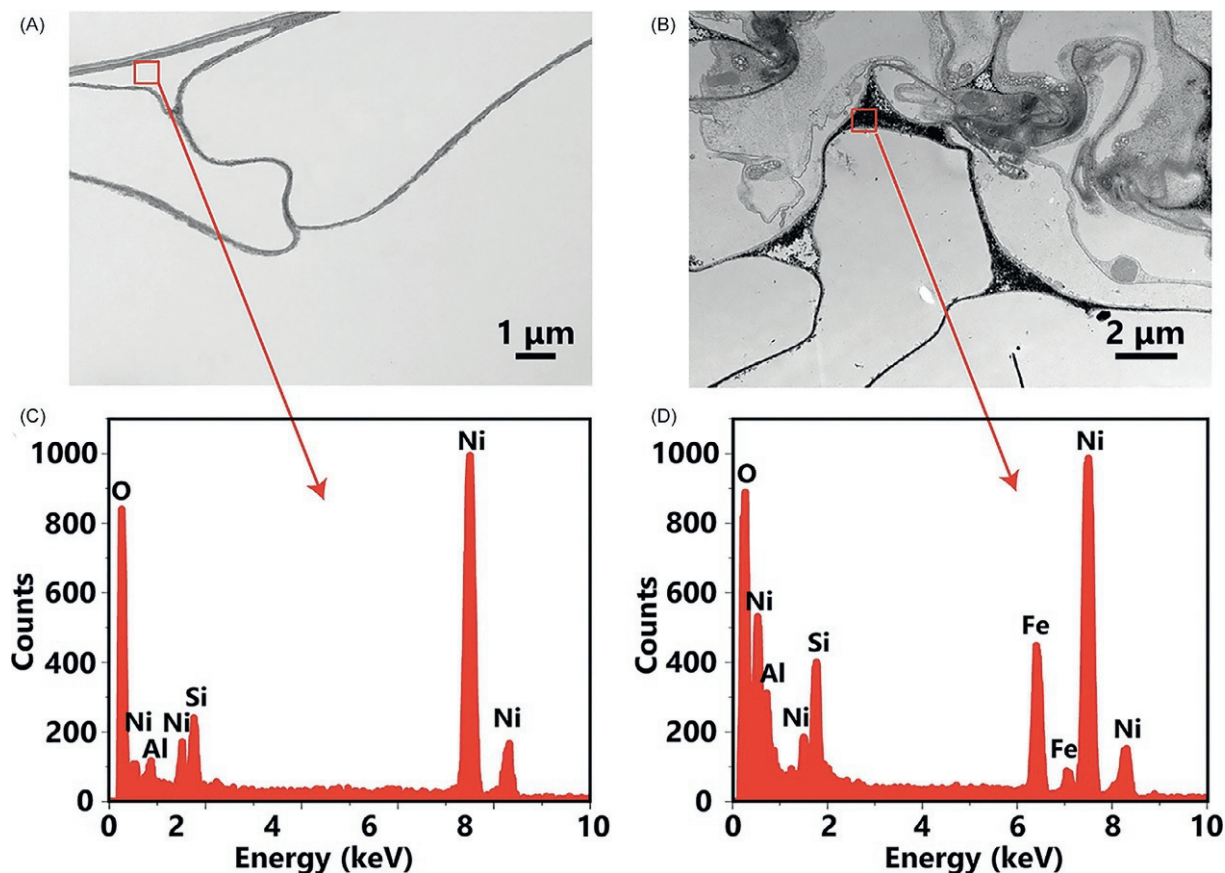


Fig. 14 Application of nZVI nanomaterials on the remediation of PCP-contaminated soil as analyzed by TEM-EDS. (A), (B) TEM images of root intercellular spaces without and with the treatment of nZVI nanomaterials in soil (100 mg kg^{-1}), respectively. (C), (D) EDS analysis of the rectangular zones highlighted in (A) and (B), respectively. Reproduced with permission from Liu Y, Wu T, White JC and Lin D (2020) A new strategy using nanoscale zero-valent iron to simultaneously promote remediation and safe crop production in contaminated soil. *Nature Nanotechnology* 16: 197–205.

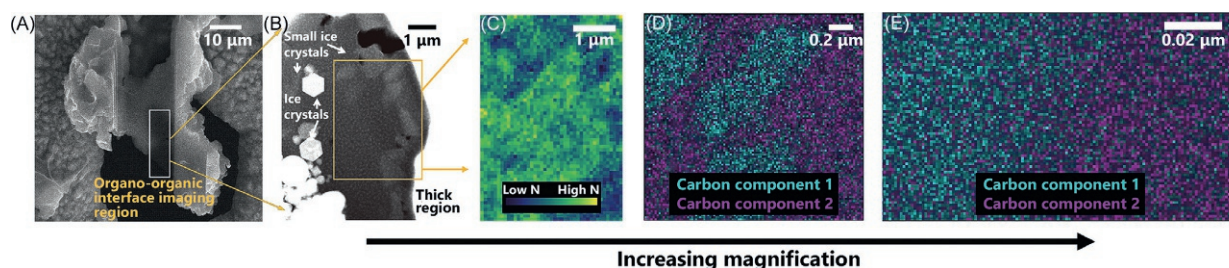


Fig. 15 Cryo-STEM and EELS of the organic–organic interface in a volcanic soil sample. (A) STEM image of volcanic soil sample with additional thinning by cryogenic focused ion beam; (B) Annular dark field STEM image of the enlarged region from a white box in (A); (C–E) EELS mapping of nitrogen (C) and carbon (D and E) for the region marked with a yellow box in (B), and (D and E) the EELS elemental map of carbon at micrometer- and nanometer-scale, respectively. Reproduced with permission from Possinger AR, Zachman MJ, Enders A, et al. (2020). Organo–organic and organo–mineral interfaces in soil at the nanometer scale. *Nature Communications* 11: 1–11.

the sampled minerals (Possinger et al., 2020). Compared with the surrounding matrix, the organo–organic interface can be clearly identified (the region marked with the white square in Fig. 15A).

Cryo-EM is an advanced electron microscopy technique for the observation of wet or biological samples, and these samples should be cooled by liquid nitrogen to cryogenic temperatures and embedded in an environment of vitreous water. Currently, cryo-EM is used mainly in structural biology and materials science, especially for the high-resolution structural determination of biomolecules in solution. In soil chemistry, this technique can keep the soil sample in its original state by avoiding possible morphological and structural changes during drying pre-treatment. Therefore, cryo-EM could be further used to investigate the organo–organic interface by preventing possible conformation changes or the formation of large crystals of ice in the soil sample

(Possinger et al., 2020). In Fig. 15B, only small ice crystals could be found, which is because the samples were rapidly cooled by immersion in liquid nitrogen (average temperature -207°C) to decrease the size of ice crystals relative to a standard freezer (e.g., -20 to -80°C). Furthermore, STEM combined with EELS can provide organic elemental information. The EELS elemental map shows the patchy distribution of nitrogen and carbon in the organic–organic interface region, and two distinct carbon bonding environments (Fig. 15C). After further analysis by carbon K-edge multivariate curve resolution, the two types of carbon bonds were assigned to $\text{C}=\text{C}$ bonds associated with aromatic structures, and alkyl $\text{C}-\text{H}$ bonds with a variety of transitions, respectively (Fig. 15D and E). This work combines STEM, cryo-EM and EELS, successfully providing comprehensive information on the organic–organic interface and enabling direct visualization and interface analysis.

Environmental scanning electron microscope (ESEM)

The ESEM has been defined as a SEM capable of maintaining a minimum water vapor pressure of at least 609 Pa in its specimen chamber. It has a simple procedure for sample preparation, a good depth of field, and a higher permitted gas pressure. Therefore, ESEM is extremely important for the direct characterization of biological samples by controlling the relative humidity, which is a considerable advantage compared with SEM. Li et al. (2020) used ESEM to investigate the internalization pathway of microplastics ($\sim 2\text{ }\mu\text{m}$) in a wheat seedling. By avoiding drying, the large gap between the epidermal cells at the site of the lateral root was observed in the wheat seedling (Fig. 16). The observed gaps were further confirmed as the main pathway for microplastics uptake. This result indicates that ESEM has great application potential in investigating the original sample morphology of plants as well as other organisms in soil. In addition, ESEM is also suitable for investigating the pore structures of minerals which are sensitive to water (e.g., bentonite).

SEM–Raman system

Although SEM is effective for characterizing the surface morphology of mineral samples, the structural and electronic information with diffraction were unknown. The Raman spectrometer can offer fundamental spectral information on the molecular structure of a given mineral, while the morphology can be obtained only by the light microscope which has a much lower resolution than SEM. A combination of these two techniques in a single system (SEM–Raman) could fulfill the above requirements during mineral characterization. For example, the SEM–Raman system was applied to track in-situ the morphological and spectral information of a newly-formed Pb-based mineral (Wu et al., 2020). The SEM image in Fig. 17 shows that Pb-based minerals are mainly scale-like, which cannot be observed clearly by the light microscope. Specific regions were further analyzed in-situ by the Raman spectrometer. The Raman spectra obtained display peaks at 1440 , 1480 , and 1590 cm^{-1} , confirming that the secondary mineral was Pb oxalate. Of note, the small and large Pb oxalate particles show distinct intensities of specific peaks in Raman mode, indicating the variable orientation of the secondary Pb mineral. To the best of our knowledge, this is the only current research that has focused on soil/mineral characterization by SEM–Raman. With its excellent performance, SEM–Raman may become widely applied in soil science in the future.

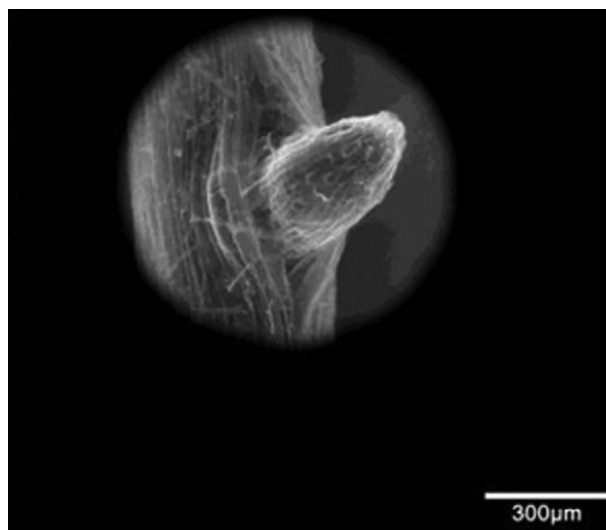


Fig. 16 ESEM of a 7-day-old wheat seedling at the lateral root formation stage. Large gap was observed between the epidermal cells at the site of the lateral root. Reproduced with permission from Li L, Luo Y, Li R, et al. (2020). Effective uptake of submicrometre plastics by crop plants via a crack-entry mode. *Nature Sustainability* 3: 929–937.

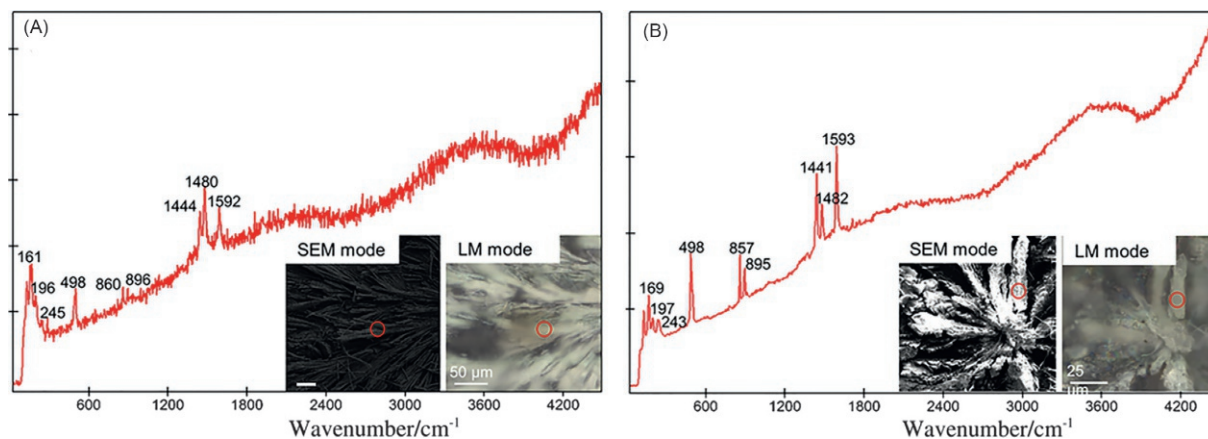


Fig. 17 The SEM-Raman analysis on the secondary Pb-based mineral with two different morphologies. (A) Small particles; (B) large particles. The detected regions for Raman analysis were indicated with red circles. Scar bar is 40 μm for the two SEM images. Reproduced with permission from Wu Y, Shao X, Jiao H, et al. (2020). Tracking the fungus-assisted biocorrosion of lead metal by Raman imaging and scanning electron microscopy technique. *Journal of Raman Spectroscopy* 51: 508–513.

Aberration-corrected electron microscope (AC-EM)

To correct the aberration, a corrector is installed in the objective lens and/or the condenser of an electron microscope, and this kind of electron microscope is called an AC-EM. Resolution of the AC-EM is as low as 0.05 nm due to the elimination of spherical aberration (Cs). The Cs corrector of TEM is installed in objective lenses, whereas it is installed in condenser lenses for AC-STEM. Sometimes, Cs correctors can be installed in both condenser and objective lens, which is called the double spherical aberration electron microscope. With its extremely high resolution, AC-EM has been widely applied in the characterization of some crystalline solids (e.g., Au (gold), GaN (gallium nitride), and other crystals) to resolve individual atomic structures in the field of material science. In addition to atomic structure, the chemical bonds (e.g., C—C bond) could be observed in graphene sheets. However, the application of AC-EM in soil chemistry has not been reported to our knowledge.

Summary

The SEM and TEM with high spatial resolution in both morphological and structural characterization have been confirmed to be extremely useful, highly versatile techniques for soil particle analysis over the past few years. In this chapter, we have summarized the foundational principles and key instrumentation of SEM, TEM and EDS. The SEM uses mainly SE, BSE while TEM collects transmitted electrons to create images. As a result, TEM offers valuable information on the inner structure of a sample while SEM provides information on the surface. The EDS mainly utilizes characteristic X-rays to display qualitative and quantitative information, and the spatial distribution of chemical elements in soil samples. In addition, applications of TEM and SEM in the characterization of soil minerals and SOM are introduced and show that SEM provides 3-D information, whereas TEM images are 2-D projections of the soil from the inner structure. The TEM is more effective for collecting structural and atomic information (crystalline/amorphous, lattice finger pattern, crystal face) of soil particles than SEM. By combining with SEM or TEM, EDS can obtain elemental species and their distribution in the soil sample. In addition to soil particle characterization, TEM, SEM and EDS play an important role in the investigation of SOM–mineral interactions and soil contamination/remediation; some examples are provided in this chapter. Moreover, we introduced some electron microscope-based new techniques including STEM, cryo-EM, ESEM, SEM–Raman and AC-EM, together with their applications in soil chemistry. The STEM takes advantage of both SEM and TEM techniques in soil particle analysis. Cryo-EM and ESEM overcome vacuum limitations, and are suitable for the observation of biological samples (e.g., plant tissues). The SEM–Raman system can give both morphological and structural information of minerals during in-situ observations. The resolution of AC-EM can achieve up to 0.05 nm, and information on the atomic or chemical bonds can be collected. Electron microscopes will continue to improve our understanding of micro-nanoscale physical and chemical characteristics of soil particles.

References

- Ahmed IA and Maher BA (2018) Identification and paleoclimatic significance of magnetite nanoparticles in soils. *Proceedings of the National Academy of Sciences* 115: 1736–1741.
Bendersky LA and Gayle FW (2001) Electron diffraction using transmission electron microscopy. *Journal of Research of the National Institute of Standards and Technology* 106: 997.

- Cheng R, Li H, Liu Z, and Du C (2018) Halloysite nanotubes as an effective and recyclable adsorbent for removal of low-concentration antibiotics ciprofloxacin. *Minerals* 8: 387.
- Deepak FL, Mayoral A, and Arenal R (eds.) (2015) *Advanced Transmission Electron Microscopy: Applications to Nanomaterials*. Heidelberg, Switzerland: Springer.
- Drif B, Taha Y, Hakkou R, and Benzaazoua M (2018) Recovery of residual silver-bearing minerals from low-grade tailings by froth flotation: The case of Zgounder mine, Morocco. *Minerals* 8: 273.
- Girão AV, Caputo G, and Ferro MC (2017) Application of scanning electron microscopy–energy dispersive X-ray spectroscopy (SEM-EDS). In: Rocha-Santos TAP and Duarte AC (eds.) *Comprehensive Analytical Chemistry*, 6th edn, pp. 153–168. The Netherlands: Elsevier, Amsterdam.
- Hamid Y, Tang L, Hussain B, et al. (2020) Efficiency of lime, biochar, Fe containing biochar and composite amendments for Cd and Pb immobilization in a co-contaminated alluvial soil. *Environmental Pollution* 257: 113609.
- Kwiecińska B, Pusz S, and Valentine BJ (2019) Application of electron microscopy TEM and SEM for analysis of coals, organic-rich shales and carbonaceous matter. *International Journal of Coal Geology* 211: 103203.
- Li L, Luo Y, Li R, et al. (2020) Effective uptake of submicrometre plastics by crop plants via a crack-entry mode. *Nature Sustainability* 3: 929–937.
- Liu H, Lu X, Li M, et al. (2018) Structural incorporation of manganese into goethite and its enhancement of Pb (II) adsorption. *Environmental Science & Technology* 52: 4719–4727.
- Liu Y, Wu T, White JC, and Lin D (2020) A new strategy using nanoscale zero-valent iron to simultaneously promote remediation and safe crop production in contaminated soil. *Nature Nanotechnology* 16: 197–205.
- Monreal CM, Sultan Y, and Schnitzer M (2010) Soil organic matter in nano-scale structures of a cultivated Black Chernozem. *Geoderma* 159: 237–242.
- Possinger AR, Zachman MJ, Enders A, et al. (2020) Organo–organic and organo–mineral interfaces in soil at the nanometer scale. *Nature Communications* 11: 1–11.
- Qi BC, Aldrich C, and Lorenzen L (2004) Effect of ultrasonication on the humic acids extracted from lignocellulose substrate decomposed by anaerobic digestion. *Chemical Engineering Journal* 98: 153–163.
- Suzuki T, Nakamura A, Niinae M, et al. (2013) Lead immobilization in artificially contaminated kaolinite using magnesium oxide-based materials: Immobilization mechanisms and long-term evaluation. *Chemical Engineering Journal* 232: 380–387.
- von Harrach HS, Klenov D, Freitag B, et al. (2010) Comparison of the detection limits of EDS and EELS in S/TEM. *Microscopy and Microanalysis* 16: 1312–1313.
- Wen B, Zhang JJ, Zhang SZ, et al. (2007) Phenanthrene sorption to soil humic acid and different humin fractions. *Environmental Science & Technology* 41: 3165–3171.
- Wu Y, Shao X, Jiao H, et al. (2020) Tracking the fungus-assisted biocorrosion of lead metal by Raman imaging and scanning electron microscopy technique. *Journal of Raman Spectroscopy* 51: 508–513.
- Xing B, Graham N, and Yu W (2020) Transformation of siderite to goethite by humic acid in the natural environment. *Communications Chemistry* 3(1): 1–11.

Relevant websites

www.minersoc.org/images-of-clay—Mineralogical Society of Great Britain & Ireland and The Clay Minerals Society.

Application of atomic force microscopy (AFM) on environmental interfaces

Saikat Ghosh^a, Zhang Di^a, Bo Pan^a, and Baoshan Xing^b, ^aYunnan Provincial Key Laboratory of Soil Carbon Sequestration and Pollution Control, Faculty of Environmental Science & Engineering, Kunming University of Science & Technology, Kunming, P. R. China; ^bUniversity of Massachusetts, Amherst, Massachusetts, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	590
Modes of AFM operation	591
Contact mode AFM	591
Application to mineral dissolution	591
Non-contact (NC)- and tapping mode (TM)-AFM imaging	591
Application of TMAFM imaging in air and aqueous medium	592
Role of substrate symmetry and electric dipole on the epitaxial growth	592
Electrical and magnetic force microscopy (EFM/MFM)	595
Enhanced electrical force microscopy or Kelvin probe force microscopy (KPFM)	595
Application of phase imaging and KPFM and MFM	595
Colloidal probe force microscopy (CPFM) and DLVO interactions	598
Experimental setup	599
Application of CPM techniques	599
Chemical force microscopy (CFM) and dynamic force spectroscopy (DFS)	600
Concluding remarks	603
References	603

Abstract

Atomic force microscopy (AFM) provides a unique opportunity to identify the nanoscale processes on environmental interfaces. The NOM-mediated transformation of engineered nanomaterials and the role of substrate symmetry and epitaxial growth of ordered aggregates can be detected from AFM imaging. Kelvin probe force microscopy (KPFM) helps identification of the asymmetric surface charge of clay particles as well as screw dislocation-assisted growth of colloidal heteroaggregates prevalent on interfaces. Colloidal probe force microscopic (CPFM) evaluation of nanoscale forces determines the relative efficiencies of organic and inorganic cations, structural heterogeneity of NOM (natural organic matter), and hematite nanominerals in stable soil aggregate formations between silica and clay.

Glossary

- μ** An elastic modulus is a quantity that measures an object or substance's resistance to being deformed elastically when stress is applied.
- B** Burgers vector named after the Dutch physicist Jan Burgers is a vector, that represents the magnitude and direction of the lattice distortion resulting from a dislocation in a crystal lattice.
- H** Hamaker Constant is defined for a van der Waals body-body interaction.
- k** Spring constant defines the stiffness of the Hookean spring.
- r** Separation distance between the centers of the interacting particles.
- γ** Surface energy is defined as the work per unit area produced by the force that creates the new surface.
- ϵ** The dielectric constant of a solvent is a measure of its polarity.
- ϵ_0** Permittivity of the free space. A physical constant is used in electromagnetism and represents the capability of a vacuum to permit electric fields.
- κ^{-1}** Debye length represents the thickness of the electrical double layer.
- σ** Distance at which the intermolecular potential between the two particles is zero.
- σ_1 and σ_2** The surface charge densities of the two surfaces. The amount of charge per unit area.
- ϕ** Phase angle refers to the angular component of a periodic wave.
- ϕ** The work function is defined as the minimum amount of energy required to remove an electron from a solid to a point in the vacuum immediately outside the solid surface.
- ζ** Raleigh damping.

Key points

- A brief introduction of AFM-based experimental techniques to elucidate the mechanism of nanoscale environmental phenomena.
- The role of NOM-mediated environmental transformation of anthropogenic nanomaterials on an air–water interface and subsequent heteroaggregation with a natural clay colloid. A screw dislocation-driven growth of a colloid cluster.
- Role of hematite nanominerals, NOM, and cationic species on the adhesion of silica and clay platelets. Role of nanoscale forces on primary soil aggregate formation.

Introduction

Discovery of the generic scanning probe microscope in the early 1980s has opened up the potential application of both scanning tunneling microscopy (STM) and atomic force microscopy (AFM)-based techniques. This discovery by G. Binnig and H. Rohrer of the IBM Zurich Research laboratories, Switzerland brought them the Noble Prize for Physics in 1986. The development of AFM applications in soil science and geochemistry enriches our understanding of several fundamental nanoscale processes occurring on environmental interfaces starting from mineral dissolution to bacterial adhesion. Nanometer-scale resolution of AFM (up to 0.5 nm), non-invasive sample measurement techniques with minimal sample damage, and its operational flexibility in a wide range of environmental conditions have furthered the scope of identifying nanoscale processes and the characterization of soil colloids on environmental interfaces. The working principle of AFM relies on the fact that a sharp tip attached to the end of a microcantilever with a spring constant (k_c) raster scans a portion of the sample surface resulting in an attractive force between the tip and the surface. The cantilever deflection (q) produced from the tip-substrate interaction corresponds to a harmonic oscillator with upward restoring force following Hooke's law as shown schematically (Fig. 1) (Reifenberger, 2016). At equilibrium, this restoring force must be equal and opposite to the interaction force that caused the deflection. During the movement of the tip across the substrate, the surface contour controls the tip movement and subsequently alters the cantilever deflections and maps the substrate surface.

Early generation AFM assembly involves the determination of interaction forces between the probe and sample holder, which undergoes upward and downward movement to control probe–sample distance. However, in the new generation AFM machines, the cantilever is attached to the piezoelectric sensor fixed to the piezoelectric tube to maintain constant interaction force. The probe scans the surface and a topographic image is obtained after acquiring the vertical control signals with upward and downward movement of the cantilever according to the surface morphology, while the interaction forces remain constant during scanning by supplying the current through a feedback loop. Piezoelectric ceramics are materials that expand or contract in the presence of a voltage gradient and facilitate the creation of a three-dimensional positioning device with high precision. The precision of vertical and lateral deflections of the cantilever is controlled by the optical lever with the reflection of a laser beam from the back of the cantilever to a photodiode. The differences between the segments of the photo-detector signals indicate the position of the laser spot on the detector and thus the angular deflections of the cantilever. The image data are sampled digitally at equally spaced intervals, generally from 64 to 2048 points per line.

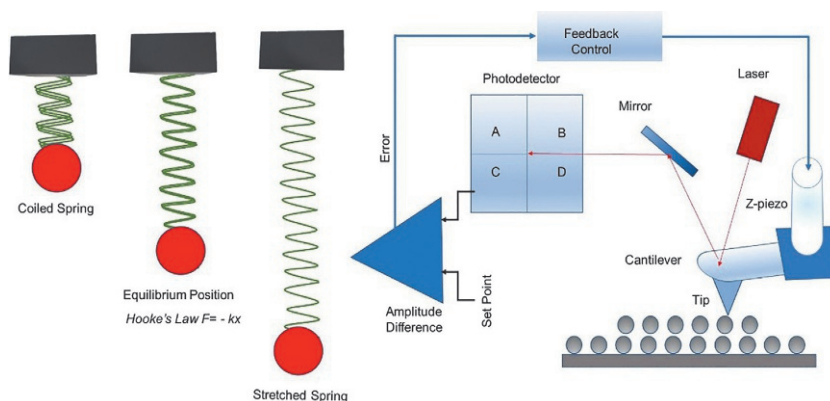


Fig. 1 Hooke's law of harmonic oscillator and schematic diagram of the AFM set up.

Modes of AFM operation

Contact mode AFM

The AFM not only measures the force on the sample but also regulates it, allowing the acquisition of images at very low forces. In contact mode, AFM uses the feedback loop voltage to regulate the force on the sample. The assembly of the feedback loop consists of a tube scanner that controls the height of the tip, the cantilever, and the optical lever, which measures the local height of the sample; and a feedback circuit that attempts to keep the cantilever deflection constant by adjusting the voltage applied to the scanner. A well-constructed feedback loop is essential to performance of the microscope. During scanning, a constant bend in the cantilever is maintained. This bend corresponds to a displacement of the tip representing the height of the sample. The major limitations of the contact mode AFM operations include contamination and deformation of the samples because a significant amount of force has been required to move the probe across the substrate (<https://www.azonano.com/article>).

Application to mineral dissolution

The application of contact mode AFM imaging equipped with a fluid cell has facilitated our understanding of the mineral–water interface governing several key environmental issues, ranging from nutrient cycling to biomineralization processes. Monitoring of in situ mineral dissolution, nucleation and growth of the crystal, and epitaxial precipitation reactions of metal ions has improved our understanding of the role of biogeochemical cycling. The capability of AFM fluid cells relies on recording the dynamic evolution of mineral topographies, formation, and migration of steps. A comprehensive review of the application of contact mode AFM is given by Maurice (1996) and Hochella et al. (1999).

Non-contact (NC)- and tapping mode (TM)-AFM imaging

The limitations of contact mode AFM imaging resulted in the evolution of non-contact mode AFM imaging by Martin et al. (1987). In this mode, an oscillating probe hovers $\sim 50\text{--}150\text{ \AA}$ above the sample surface to sense the attractive van der Waal's force resulting in a shift in the resonance frequency of the cantilever. The alteration in both the amplitude and phase of the cantilever has been monitored by the feedback loop to control tip–sample distance.

The non-contact mode provides very good resolution of the height of the sample but with reduced lateral resolution. The application of non-contact mode AFM in ambient conditions requires maintenance of a constant tip–sample distance and prevention of the tip contacting the sample surface. The major limitation of non-contact mode imaging is that once the tip makes accidental contact with the sample the tip can get stuck and perturb the imaging process. The operation of non-contact AFM in the fluid is extremely difficult since the fluid layer is substantially thicker than the range of the van der Waals force gradient. Therefore, imaging of the true surface with non-contact AFM fails because the oscillating probe becomes trapped in the fluid layer.

The advent of tapping mode AFM (TMAFM) has substantially improved the limitations of the non-contact mode. To prevent the AFM tip from sticking to the substrate surface in non-contact mode a deliberate 'tapping' is introduced in which the tip strikes against the substrate surface and detaches on each oscillation cycle by using a large vibrational amplitude. In regular TMAFM, the probe oscillates at its fundamental resonance frequency with constant adjustment of the tip's vertical position to maintain a constant oscillation amplitude while scanning a surface. The higher amplitude of oscillation that makes intermittent contact with the substrate in each cycle provides improved vertical and lateral resolution and the capability of imaging in a liquid environment (<https://www.azonano.com/article>). In TMAFM, dynamic vibration of the cantilever can be divided into three zones including free, transition, and contact vibration zones (Chan and Green, 2006). Therefore, TMAFM can be applied for imaging soft biological samples as well as having operational capability in a liquid medium including imaging of soil colloid particles in their natural hydrated state. The TMAFM phase of imaging has significant application in identifying the properties of materials. The phase shift is interpreted as being the phase lag between the driving voltage signal and the actual displacement response of the cantilever. In TMAFM generally, silicon (Si) tips with cantilever lengths of 125 and 225 μm and resonant frequencies of 298 to 355 kHz and 67 to 94 kHz are used (Namjesnik-Dejanovic and Maurice, 2000). However, in a liquid environment, a silicon nitride (Si_3N_4) tip with a shorter cantilever length (100 μm), resonance frequency 0–50 kHz, and 0.032 Nm^{-1} spring constant were used. Except for free vibration amplitude and resonance frequency of the cantilever, the quality (Q) factor depends significantly on the experimental environment and on the damping ratio (ζ): $Q = \frac{1}{2\zeta}$, ζ is one of the parameters determining the Raleigh damping as shown schematically in Fig. 2 (Deng et al., 2015). Damping of the cantilever while imaging in solution necessitated enhancement of the drive amplitude with stronger vibration of the probes; however, this may damage soft materials adsorbed on the substrate surfaces. The damping-induced limitation of the cantilever in TMAFM has been improved by the introduction of Peak force tapping technology. In this process, a constant feedback loop is maintained by applying a sinusoidal ramping instead of linear ramping used in regular tapping mode at forces of 100 pN or less. In this process, as the tip approaches the surface the velocity tends to be zero enabling more accurate and direct control of the tip–sample interaction force. Recently high-speed frequency modulation AFM (FM-AFM) has enabled true atomic-resolution of calcite in liquid at $\sim 1\text{ frame s}^{-1}$, which is ~ 50 times faster than the conventional FM-AFM (Miyata et al., 2017). The surface features of the moving step edges of calcite were detected during dissolution in water at the subnanometer scale.

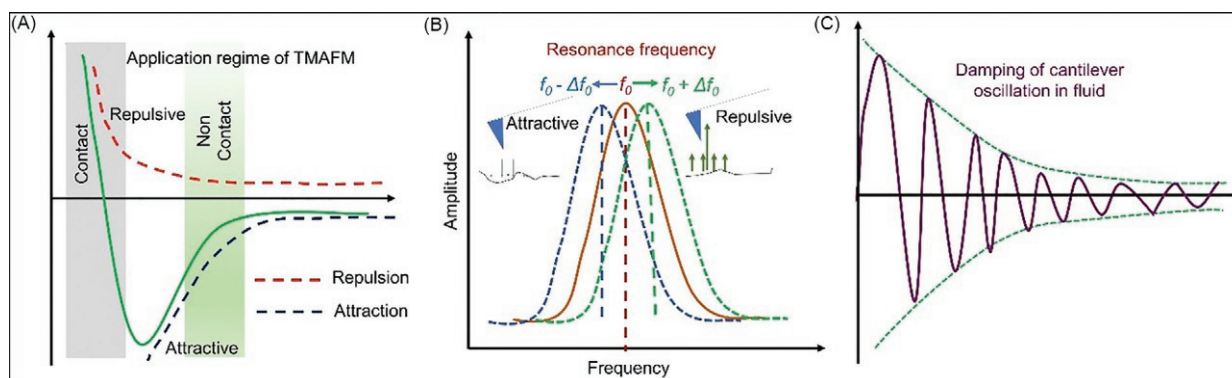


Fig. 2 (A) The range of AFM application regime. (B) resonance frequency of the cantilever in air and frequency shift due to tip–substrate attractive and repulsive forces and (C) damping of the cantilever oscillation in a fluid medium (<https://www.azonano.com/article>).

Application of TMAFM imaging in air and aqueous medium

The molecular conformation of the natural organic matter (NOM) molecules has a profound role in the entrapment of contaminants including hydrophobic organic compounds (HOCs). A correlation between NOM conformation and sorption behavior of heavy metal ions was determined by Liu et al. (2001) with ex-situ AFM imaging. Namjesnik-Dejanovic and Maurice (2000) applied in situ AFM to identify the role of solution properties including pH, ionic strength, concentration, and nature of the exchangeable cations on NOM conformation on mica and hematite substrates. At relatively low NOM concentrations (20–50 mg C L⁻¹), and intermediate pH (5–6), NOM molecules isolated from groundwater produced small micellar aggregates on mica due to cation-assisted charge screening. However, at 100 mg C L⁻¹ NOM concentration ring and toroidal structures were detected. The phase-contrast imaging of NOM on a hematite substrate revealed the development of flattened disk, ring-shaped and spherical structures and accords with the conformation of the Fe-NOM complex solutions viewed from X-ray microscopy (Myneni et al., 1999). Pan et al. (2008) have shown a correlation between HOC sorption and NOM conformation as a function of solution variables including pH, ionic strength, and concentration. The phase-contrast TMAFM imaging provides the opportunity to identify the interaction of NOM with mineral colloids including anthropogenic NPs as well as adhesion of NOM molecules to mineral particles isolated from the groundwater (Plasche et al., 2002).

Our previous investigations have elucidated the mechanistic difference in suspension instability of aluminum oxide NPs in the presence of low molecular weight Aldrich humic acid (HA) and the high molecular weight soil-derived HA7 fraction from Amherst peat (Fig. 3A–C). The TMAFM imaging and the dynamic light scattering (DLS) measurements revealed that Aldrich HA induced charge neutralization of an Al₂O₃ NPs suspension, and specifically at a higher concentration of the former it facilitated stronger van der Waals attraction resulting in destabilization of the colloidal suspension. Nonetheless, even at a lower concentration of high molecular weight HA7, Al₂O₃ NPs undergo significant flocculation. The TMAFM imaging of the HA7-Al₂O₃ NPs mixture revealed that the long-chain fractions of HA7 with a large radius of gyration (the root-mean-square distance between each atom in the molecule and the center of mass of the molecule itself, normalized for the number of atoms) anchors on several Al₂O₃ NPs facilitated bridging flocculation, specifically at low ionic strength (Ghosh et al., 2008) (Fig. 3). The TMAFM image of single Al₂O₃ NP showed a positive phase shift due to repulsive interaction between NPs and the hard probe while a negative phase shift was detected from the viscoelastic drag in HA7-modified Al₂O₃ NPs.

Unlike at the low concentration of HA7, the complete surface modification of Al₂O₃ NPs by HA7 at higher ionic strength resulted in the formation of an HA7 corona with the enhancement of colloidal stability (Fig. 3D–F). The TMAFM image of bare Al₂O₃ NPs at pH 5 on Ca²⁺-adsorbed mica showed significant aggregation as seen from the AFM height and large repulsive phase shift (Fig. 3D and inset). Therefore, it follows the classical Derjaguin–Landau–Veerwey–Overbeek (DLVO) model of colloidal aggregation. In contrast, HA7-coated Al₂O₃ NPs were predominantly separated from each other with reduced aggregation evident from the height image and it complies with the dynamic light scattering (DLS) data (Ghosh et al., 2010). The AFM phase image of the HA7-coated NPs at pH 5 exhibited a dark corona of adsorbed HA7 molecules around the core NP due to the viscoelastic drag between the soft HA 7 molecules and the hard tip (Fig. 3E and F). The enhanced suspension stability of HA7-coated Al₂O₃ NPs arises from the overlapping of excluded volumes of the coated NPs resulting in steric stabilization. Flory–Krigbaum’s statistical thermodynamic theory of polymer solvation clearly explained that as the coils of the adsorbed macromolecules overlap, the concentration of the polymer doubles in that overlapping region, and the system becomes thermodynamically unstable ($\Delta G > 0$). This causes the instantaneous movement of the water molecules within the lens-like overlapping zone with an overall rise in colloidal stability due to entropic depletion repulsion.

Role of substrate symmetry and electric dipole on the epitaxial growth

The air–water interface prevalent in unsaturated soils is the hotspot of environmental interactions of anthropogenic nanomaterials (NM) with mineral particles and NOM molecules. The nanoscale resolution of AFM and the capability of imaging with a fluid cell bolsters the possibility of identifying the interfacial processes regarding dissolution, transformation, and heteroassociation of anthropogenic colloids in diverse porewater compositions. Numerous applications of Ag (silver) NPs including antibacterial and

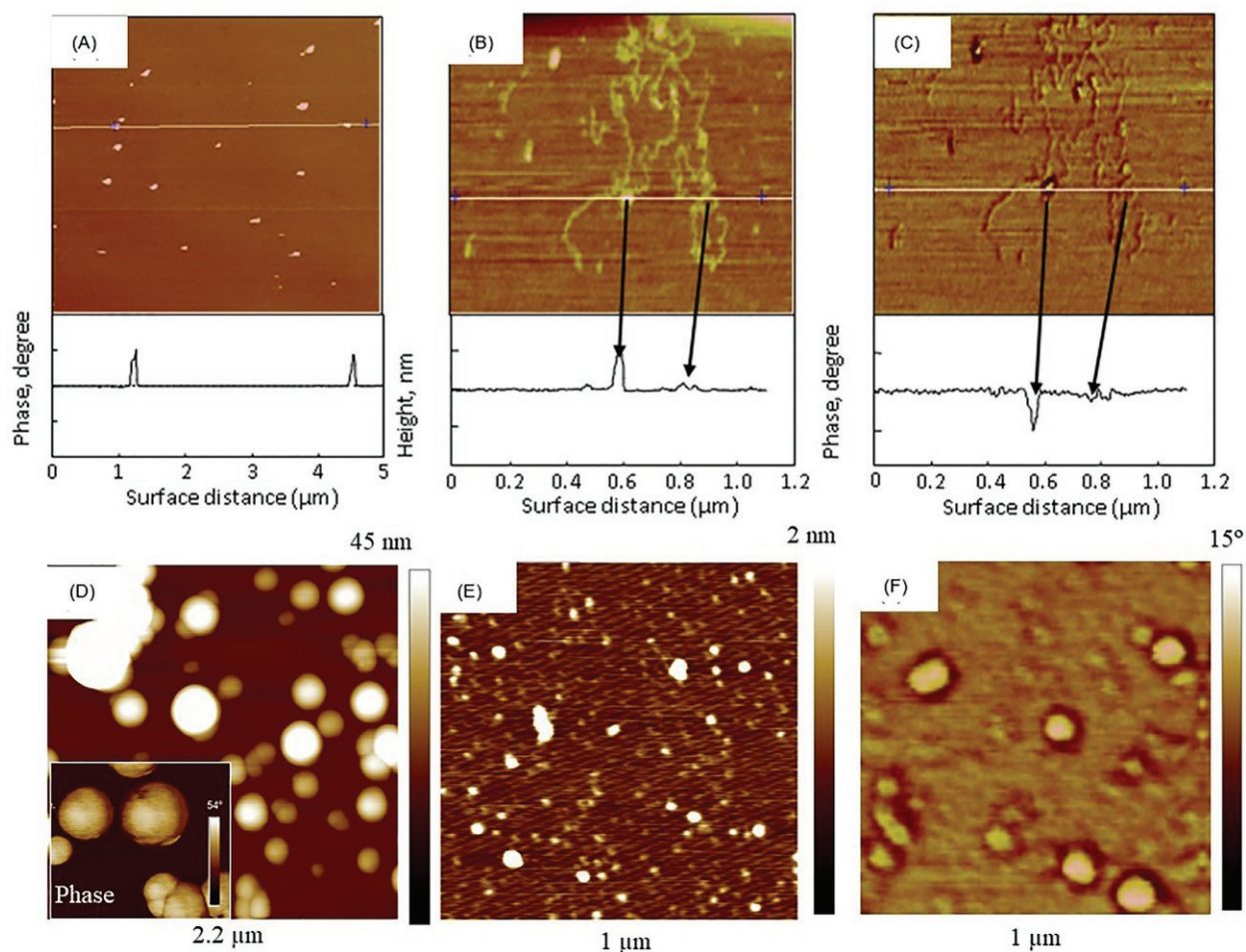


Fig. 3 (A) The TMAFM phase image produced from repulsive interactions between the hard silicon cantilever and aluminum oxide NPs with a positive phase shift, (B and C) TMAFM height and phase contrast images of aluminum oxide NPs and HA7 mixture. The presence of long-chain biomolecules in high molecular weight HA7 contributed to bridging flocculation, (D) TMAFM height and phase images (inset) of aluminum oxide NPs in Ca²⁺-saturated mica substrate, reflecting significant aggregation according to the DLVO model. The inset shows a large positive phase shift due to probe-cluster repulsion. However, HA7-coated aluminum oxide showed enhanced colloidal stability, (E) height, and (F) phase images of HA7-coated aluminum oxide revealing steric stabilization (Ghosh et al., 2008, 2010).

catalytic properties and greater production facilitated its environmental incorporation into soil. The Ag⁺ produced from the particle dissolution and formation of reactive oxygen species (ROS) caused toxicity to organisms. The AFM height imaging coupled with nanolithography techniques provides a convenient way of measuring the dissolution kinetics of AgNP in an air-saturated salt solution (Liu et al., 2018). Except for the dissolution of metallic NPs, our study has shown the interfacial transformation of carbonaceous fullerene water suspension (FWS) i.e. nC₆₀ in the presence of tannic acid (TA), a surrogate of NOM (Ghosh et al., 2019).

The TMAFM image of FWS showed the formation of spherical 3-D clusters (height above 10–15 nm) due to their inherent hydrophobicity specifically on the hydrophilic mica surface (Fig. 4A–D). However, π electron transfer from TA molecules to electron-deficient fullerene and consequent surface hydroxylation produce oligomeric 1-D arrangements evident from AFM height images on the mica surface. The interfacial heteroassociation of FWS with Na-saturated kaolinite (Na-KI) produced random aggregates. However, in the presence of TA and Ca²⁺ (added through the inlet port of the AFM fluid cell), heteroassociation of FWS with Na-KI produced 1-D needle structures on the air–water interface. Reduced sorption of TA on negatively charged Na-KI and oligomerization of TA-functionalized fullerene produced an amphiphilic helical backbone and subsequent entrapment of Na-KI particles at the nodes.

We also observed in a muscovite mica substrate symmetry-directed organization of the needles into a 2-D rectangular herringbone motif. The presence of a mirror plane on the dioctahedral muscovite (001) (2-D space group) with almost equal energy particulate attachment sites resulted in V-shaped and 180°-rotated twinned growth of the herringbone motif (Fig. 4E–H, Ghosh et al., 2014, 2019). The minute elevation due to the difference in orientation of the –OH groups below the quasi-hexagonal silica tetrahedral layer on mica in the α (001) and β (001) cleavage planes has resulted in an almost 60° orientation between the arms of the V-shaped structures (indicated by the black lines), equivalent to a 120° mutual rotation. However, the more complex geometry of the colloid cluster would probably occur on a trioctahedral phlogopite mica (2-D space group $p31m$) with three mirror axes and higher rotational degrees of freedom. The lower dimension of the FWS due to TA-assisted functionalization and the electric

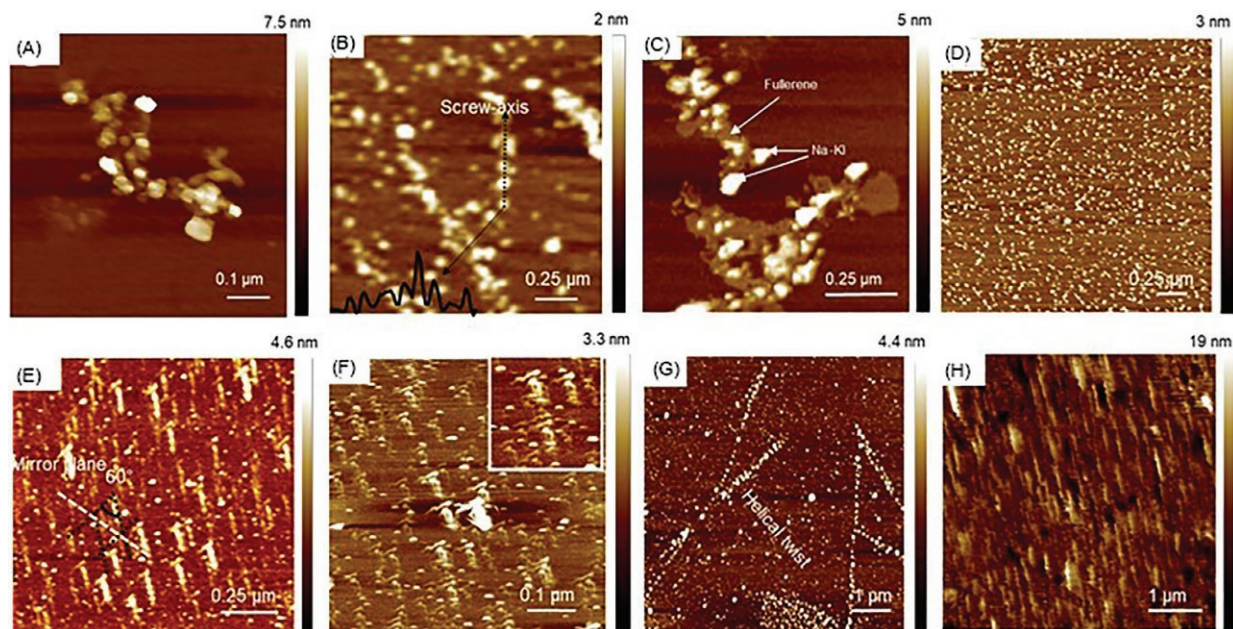


Fig. 4 The AFM height image of (A) FWS and (B) TA-assisted oligomerization of FWS and amphiphilic arrangement of the functionalized fullerene at pH 7, (C) hetero-association of FWS and Na-KI produced random aggregates in the absence of TA and any cationic species at pH 7, (D) Na-KI and FWS mixture in the presence of TA and (E and F) in 10 mM Ca^{2+} on muscovite mica producing the herringbone pattern of aggregates and screw dislocation-driven growth of tubular aggregates, (G) mica substrate preadsorbed with 10 mM K^+ produced helical wire-like aggregates, and (H) at higher particle concentration and on mica substrate treated with K^+ , the ternary mixture produced liquid crystalline aggregates (Ghosh et al., 2019).

field produced by the exchangeable K^+ on mica facilitated the growth of the needle. The role of exchangeable K^+ was evident from the liquid crystalline arrangement of particles on the air–water interface. Low-energy electron diffraction (LEED) investigation on freshly cleaved mica has shown that a half monolayer of K^+ facilitates the surface-dipole-directed linear growth. However, Ca^{2+} -mediated enhanced nucleation of the functionalized FWS and carbon enrichment on the muscovite surface subdued the K^+ -induced electric field and transformed the aggregate geometry.

In situ AFM images of the ternary mixture after equilibration with the mica surface (not treated with KCl) followed by the addition of 10 mM Ca^{2+} resulted in the growth of mound-like aggregates. The larger diffusion-limited mass transfer of particulates across the Ehrlich–substrate adhesion.

Schwoebel barrier (ESB) or step edge barrier contributed to the Z-directed growth of spiral mounds. Quadratic polynomial fitting of the clusters obtained at different time intervals after the introduction of Ca^{2+} demonstrated an increase in the sharpness of the parabolic curve and reduced curve planarity (Fig. 5). The undulating geometry of the parabolic curves indicated at the top, specifically approximately 30 min after Ca^{2+} addition, emphasized screw dislocation and transportation of particulates around the spiral core to higher steps.

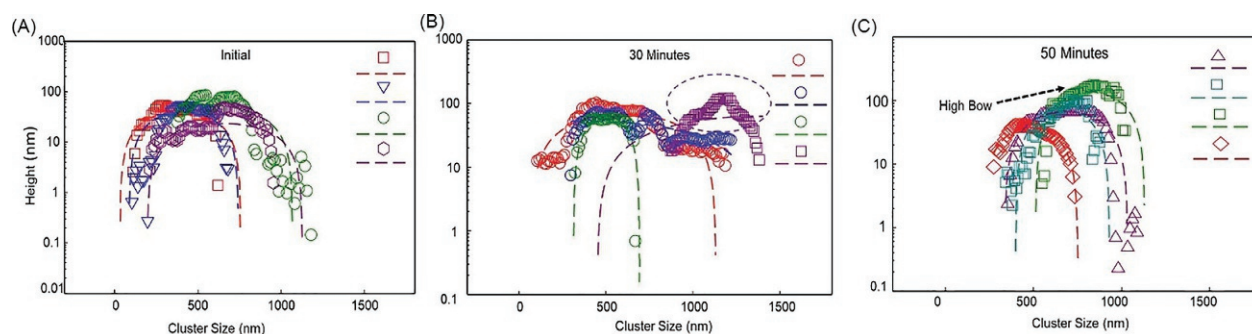


Fig. 5 Quadratic polynomial fitting of the aggregate heights obtained from the ternary mixture of Na-KI, FWS, and TA at different time intervals after the injection of 10 mM CaCl_2 . The ternary mixture was equilibrated with mica for 120 min before CaCl_2 addition. Reduced 2-D nucleation, skinnier and sharper parabolic curves with high bow indicated deviation from the fitting with time. The screw-dislocation-driven spiral motion of particulates around the dislocation core facilitated the mass transfer of particles to avoid the steep Ehrlich–Schwoebel barrier (ESB); hence prompting the growth of a binary colloid crystal (Ghosh et al., 2019).

Electrical and magnetic force microscopy (EFM/MFM)

The versatility of scanning probe force microscopy (SFM) has application in mapping the surface potential or work function, and visualization of magnetic forces emanating from the materials. In these two-pass force microscopic systems, in the first pass, a topographic image is recorded from the short-range van der Waals force (in the order of r^{-6}) due to a non-contact mode rastering of the probe over the substrate surface at its resonance frequency. In the second pass, the probe rests at a specific height away from the surface and maps the long-range electric and magnetic fields (in the order of r^{-2}) with a change in phase angle. Briefly, the upward (repulsive) or downward (attractive) bending of the magnetic/electrical probe under the magnetic and electric fields emanating from the colloidal particles and nanostructures adhered to the substrate surface and consequent phase shift is interpreted as the deviation from the driving voltage signal and the actual displacement response of the cantilever (www.parksystems.com).

The evolution of MFM, EFM, and KPFM (electrical or Kelvin probe force microscopy) techniques has provided nanoscale-resolutions of materials that are unlikely to be obtained by any other means. Let us consider the equation of motion of an oscillating cantilever in the absence of any interaction with the substrate,

$$y_0 = A_0 \sin(\omega t \pm \varphi) = A_0 \sin(2\pi f_0 t \pm \varphi_1) \quad (1)$$

above a surface with a free amplitude A_0 , f_0 is the resonance frequency of the cantilever, ω is the angular velocity, t is the time, and φ is the phase angle. However, during interactions with the substrate, the free amplitude of the cantilever changes. In the case of attractive interactions, $f_0 > f$, while in the case of repulsive interactions $f_0 < f$ and the modified equation:

$$y_1 = A \sin(\omega t \pm \varphi_2) = A_0 \sin(2\pi f t \pm \varphi_2) \quad (2)$$

The phase difference is shown schematically in Fig. 6. The frequency shift allows direct measurement of the force changes, phase lag, and amplitude signals which are then combined to calculate the dissipation of energy in each oscillation cycle due to tip-sample interaction.

Enhanced electrical force microscopy or Kelvin probe force microscopy (KPFM)

The surface charge behavior of soil mineral particles plays a significant role in diverse geochemical processes including sorption of NOM on the mineral matrix, adsorption of contaminants, and colloidal transport through environmental media. Among the naturally abundant mineral particles, clay showed significantly complex surface charge behavior due to the isomorphous substitution and pH-dependent ionization of the edge groups. Several experimental techniques including electrokinetic measurements have been used to determine the electrical double layer (EDL) and surface charge behavior of kaolinite particles at a macro level. However, the asymmetric distribution of surface charge on clay particles requires the application of KPFM. The Kelvin probe technique, named after Lord Kelvin, who provided the concept of built-in contact potential difference in metals first observed by Alessandro Volta in the early 19th century. Let ϕ be the work function i.e. the minimum energy required to remove an electron from the ground state or in other words the difference in energy between electrons in a free state to that of those at the Fermi level (Palermo et al., 2006). When two capacitors composed of two different metals, with different work functions (WF) are electrically connected, electrons will flow from lower WF (ϕ_1) material to the material with higher WF (ϕ_2) because of greater electron binding capacity. The opposite charges on the capacitor plates and a contact potential difference (CPD) between the two materials can be identified. The presence of an electric field in the capacitor due to these charges can then be easily detected. An external potential V_c can be applied to nullify this field. At equilibrium, the field is zero, and the externally applied potential equals the contact potential difference, which corresponds to the difference between the two WFs (Fig. 6). Therefore, knowledge of ϕ_1 has shown the possibility of calculating ϕ_2 : $\phi_2 = \phi_1 - qV_c$ where q is the elementary charge. Lord Kelvin first realized this electron tunneling phenomenon while studying the interaction phenomenon between polished Cu and Zn surfaces with an electroscope measurement. However, Zisman (1932) improved the method by using a vibrating capacitor setup with mechanical oscillation at a specific frequency which induces a change in the capacitance of the system with the formation of a small alternating current. Therefore, with the discovery of AFM in the early 1980s, Nonnenmacher et al. (1991) applied the KPFM technique to determine the CPD between a sharp conducting probe and the surface. The CPD between the tip and sample is defined as $V_{CPD} = (\phi_{tip} - \phi_{sample})/-e$, where ϕ_{tip} and ϕ_{sample} are the work functions of the sample and the tip. For the actual measurement of CPD, an additional voltage is applied between the two plates until space in-between becomes field-free and the current goes to zero. The KPFM measures the potential difference by applying an AC voltage on the tip. The electrostatic force ($F_{ES}(z)$) is given by the equation:

$$F_{ES}(z) = -\frac{1}{2} \Delta v^2 \frac{dc(z)}{dz} \quad (3)$$

that originated from CPD between the tip and sample surface, and where z is the direction normal to the sample surface, Δv is the potential difference between V_{CPD} and the voltage applied to the AFM tip, and $\frac{dc(z)}{dz}$ is the gradient of the capacitance between tip and sample surface.

Application of phase imaging and KPFM and MFM

The electrical and magnetic properties of soil mineral particles provide a semi-quantitative estimate of material properties. The interfacial arrangement of clay particles on the air–water interface in the presence and absence of NOM depends upon the surface

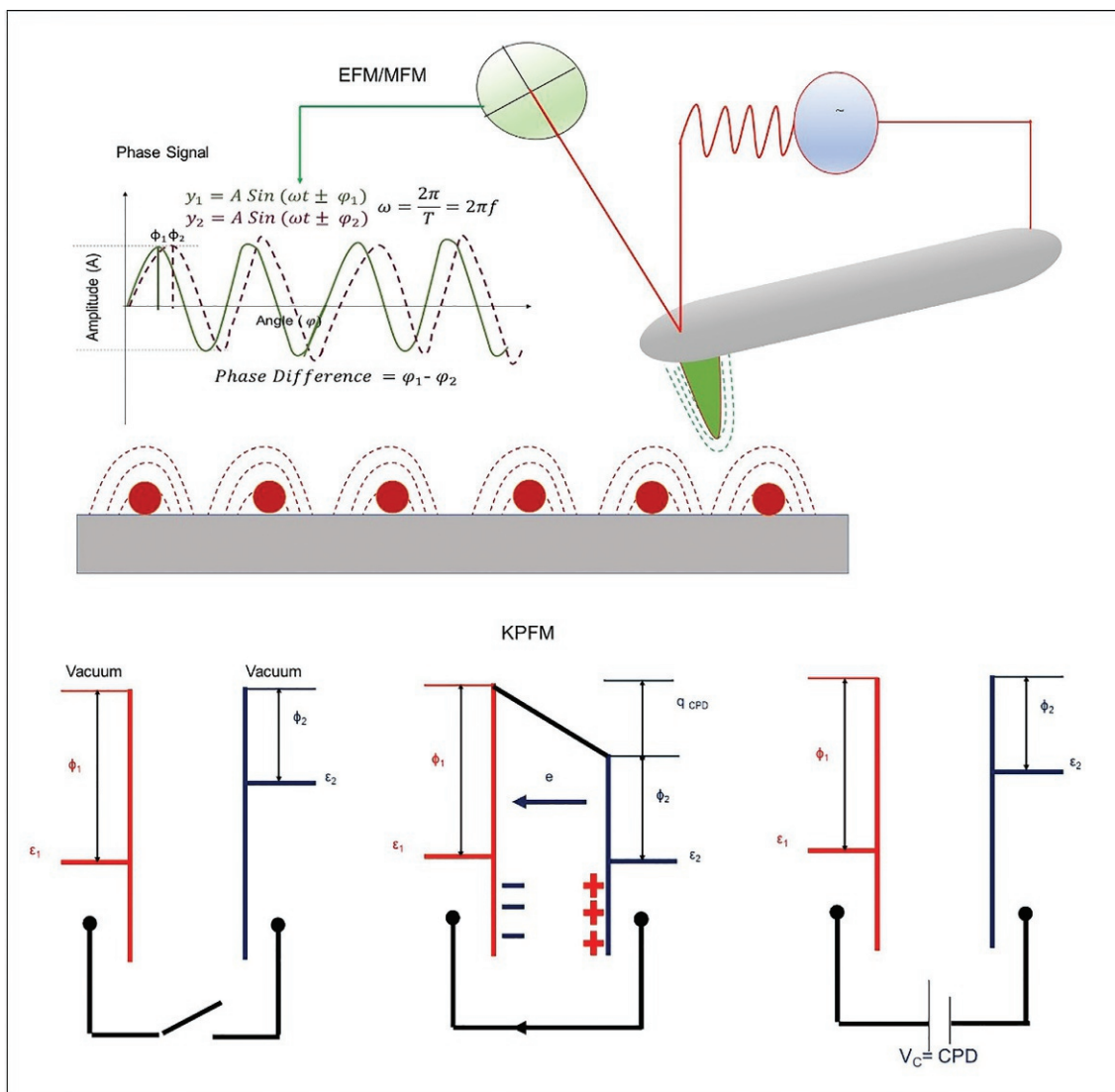


Fig. 6 The detection of stray electric and magnetic fields in a two-pass mode. In the first pass, topographic images are recorded, and in the second pass at a specific lift height probes detect the electric and magnetic fields emanating from the particles or materials that caused a change in phase angle. Schematic diagram of the electron tunneling effect proposed by Lord Kelvin the foundation of the KPFM process. Based on Palermo V, Palma M, and Samori, P (2006) Electronic characterization of organic thin films by Kelvin probe force microscopy. *Advanced Materials* 18(2): pp. 145–164.

charge asymmetry of kaolinite particles. In situ AFM images of an Na–Kl and TA mixture at pH 7 in the absence of added cations showed the wire-like arrangements of particles as indicated in Fig. 7. The attainment of polarizability maxima and EDL repulsion minima facilitated such an arrangement. The surface potentials of Na–Kl and the heteroaggregate were determined by KPFM in a Bruker Multimode 8 AFM with an MESP-V2 KPFM cantilever (0.01–0.025 Ohm-cm antimony doped Si) with a spring constant of 3 N m^{-1} and a resonance frequency of 75 kHz in lift mode. Ex situ AFM height and corresponding KPFM images of Na–Kl at pH 7 also showed the tip-to-tip arrangement of kaolinite. Enhanced ionization of Si–OH groups on the tetrahedral facet of silicate (Si–Tr) and asymmetric surface charge at pH 7 and stronger EDL repulsion prevented the sideways approach of Na–Kl particles but led to a head-to-tail arrangement. A potential energy minimum can be attained at minimum EDL repulsions with maximum polarizability. Except for the interfacial arrangement of particles, the heteroassociation mechanism between Na–Kl and fullerene NPs in the presence of TA and Ca^{2+} showed 1-D screw dislocation-mediated growth of tubular structure. The surface potential map of the tubular structure revealed the generation of crystallographically oriented nearly linear nanopores with an average pore radius of $\approx 30\text{--}50 \text{ nm}$, produced from the screw dislocation-mediated shear with a concomitant rise in the Burgers vector. The KPFM height and corresponding potential contrast revealed the formation of a hollow core and movement of the mass from the tubular core i.e. dislocation motion. This is in line with the formation of GaN nanopipes. Frank (1951) explained that the strain around the

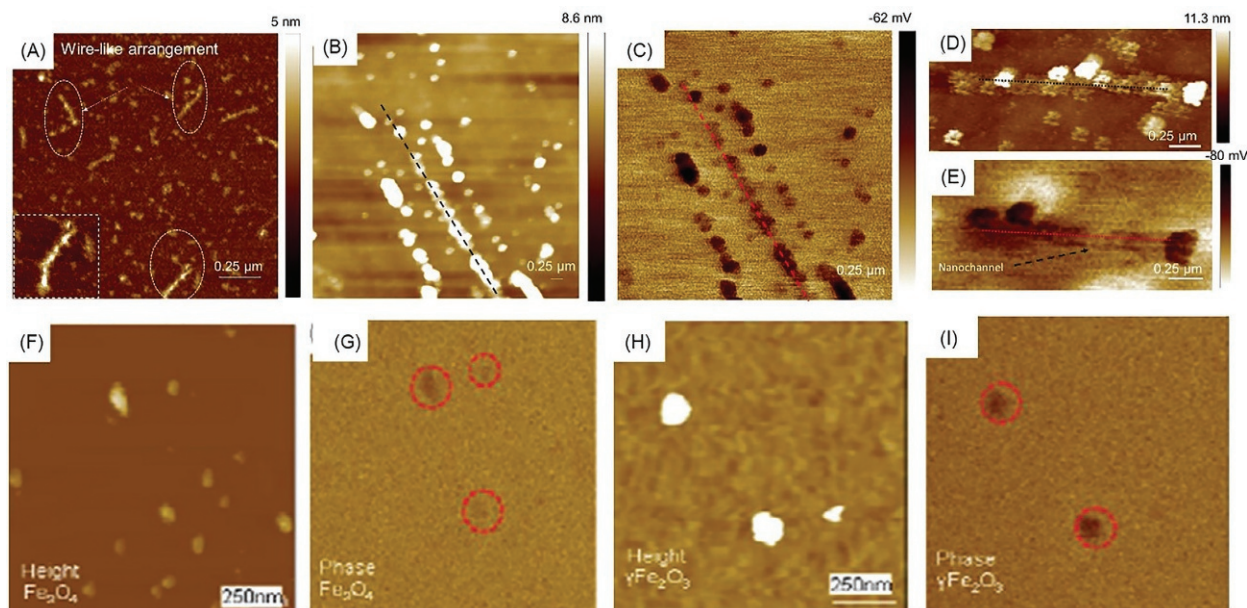


Fig. 7 (A) In situ AFM height images of tip-to-tip wire-like assembly of Na-KI in the presence of TA at pH 7 due to asymmetric EDL repulsion. Ex situ (B) AFM height and (C) KPFM surface potential of pure Na-KI particles at pH 7. (D–E) AFM height and corresponding KPFM images of Na-KI and FWS mixture in the presence of 2 mg L^{-1} TA and 10 mM Ca^{2+} at pH 7 revealed screw dislocation driven tubular growth. Ex situ AFM (F) height and corresponding (G) MFM image of Fe_3O_4 nanocluster. (H) AFM height and (I) MFM phase images of $\gamma\text{-Fe}_2\text{O}_3$ NPs. It is evident from these images that the magnetic phase shift is lower from the larger Fe_3O_4 NPs clusters relative to smaller clusters of $\gamma\text{-Fe}_2\text{O}_3$ NPs (darker color). Notably in both the MFM images the magnetic field emanating from the smaller clusters was much weaker than larger clusters. The images were taken at a constant lift height of 30 nm (Ghosh et al., 2014, 2019; Guo et al., 2020).

screw-dislocation and high Burgers vector enable particulate movement next to dislocation sites to reduce strain and showed that the Burgers vector (B) for the formation of hollow cores is:

$$B > \frac{66\pi\gamma}{\mu} \quad (4)$$

where the surface energy (γ) is 0.4 J m^{-2} and the elastic modulus (μ) $\approx 4 \text{ G Pa}$ for kaolinite. Adding in these values showed that the right-hand side term must be $>20 \text{ nm}$ for a hollow core to be formed in the tubular aggregates. The average estimated pore radius (r) for the tubular cores determined from the following equation was found to be $\approx 50 \text{ nm}$:

$$B^2 = \frac{8\pi^2\gamma r}{\mu} \quad (5)$$

The AFM images of the sample showed pore radii within the range of $30\text{--}50 \text{ nm}$, which are consistent with the calculated value.

The natural abundance of magnetic NPs and synthesized nanoscale zero-valent iron (nZVI), which are effective in the remediation of contaminated groundwater requires their delivery to contamination sites. The magnetic dipolar force-mediated aggregation of the particles and consequent loss of reactive surface area reduces their efficacy toward the degradation of contaminants including chlorinated compounds. The application of MFM, a semi-quantitative technique has shown strong efficiency in determining the stray magnetic field emanating from magnetite NPs (Magnetic Force Microscopy (MFM), n.d.). The MFM imaging of these nanoclusters has determined the 'Z'-directed stray magnetic field with a magnetized 45 nm Co—Cr (cobalt—chromium) alloy-coated silicon MFM tip (MFM, Asylum Research, USA). The application of MFM has shown enormous potential in identifying the magnetic domains in NPs or nanostructures. The TMAFM height image of magnetite NPs and head-to-tail alignment of the magnetic dipole showed the formation of magnetic nanowire on a mica substrate. To identify the difference in magnetic field that emerged from Fe_3O_4 (magnetite) and $\gamma\text{-Fe}_2\text{O}_3$ NPs, MFM measurements at a constant lift height of 30 nm indicated a weaker magnetic phase shift from the larger clusters of Fe_3O_4 NPs relative to the smaller dimensional $\gamma\text{-Fe}_2\text{O}_3$ NPs maintained at pH 6 (Fig. 7F–I). It is likely that despite the larger dimension of the Fe_3O_4 NPs cluster, the size of the individual magnetite NPs was less than the individual $\gamma\text{-Fe}_2\text{O}_3$ NPs (darker color, stronger phase shift) referring to greater thermal fluctuations and loss of magnetic moments. The darker phase-contrast corresponding to the larger negative magnetic phase shift is attributed to the stronger attractive magnetic dipolar force between the magnetic probe and the NPs. But Fe_3O_4 NPs with a larger surface to core ratio and enhanced surface defects, and oxygen-mediated loss of surface magneto-crystalline anisotropy that contributed to the magnetization loss. The advantage of MFM relies on the fact that instead of measuring the actual magnetic moments it can provide a semi-quantitative estimate of the nanoparticle's magnetic strength. Because the determination of magnetic moments in a superconducting quantum interference device (SQUID) is very expensive because of the need for liquid helium. Therefore, MFM can be useful in estimating the magnetic moments of nanostructures as well as deciphering the mobility of magnetic domains.

Colloidal probe force microscopy (CPFM) and DLVO interactions

The DLVO forces or surface forces, a summation of van der Waals attraction and electrostatic double layer (EDL) repulsion, determine the interparticle interaction forces specifically for inorganic colloidal particles. In a high dielectric aqueous medium, the mutual interactions between the EDLs result in strong repulsion due to ionization of the surface functional groups. Mathematically, the electrostatic repulsive force (F_{el}) is defined as (Butt et al., 2005):

$$F_{el} = \frac{4\pi R \sigma_1 \sigma_2}{\epsilon \epsilon_0 \kappa} e^{-\kappa D}, \quad (6)$$

$$\kappa = \sqrt{\sum_i \frac{n_{i\infty} e^2 z_i^2}{\epsilon \epsilon_0 kT}}, \quad (7)$$

where F_{el} as a function of distance D for a probe and has been normalized by the sphere radius (R). Since $F/2\pi R$ is equal to the energy per unit area between two equivalent flat surfaces, σ_1 and σ_2 are the surface charge densities of the two surfaces, ϵ_0 is the permittivity of free space, ϵ is the dielectric constant of the solvent, κ^{-1} is the Debye length determined from the salt concentration; $n_{i\infty}$ is the bulk concentration of the ion species i ions per unit volume, z_i denotes the valency of that ion species. The sum runs over all ionic species present, e is the unit charge. Typical force curve measurement data were obtained for a silica probe and silica surface at different ionic strengths, and curve-fitting determined the Debye length from the F vs D curves. However, at higher ionic strengths and pH near the point of zero charges (PZCs), the van der Waals attraction (VdW) between the colloidal probe and substrate originated from the dipole–dipole, and dipole induced dipole interactions become dominant. The van der Waals force is calculated from Lifshitz theory based on quantum electrodynamics and spectroscopic measurements. For a sphere of radius R interacting with a plane, the VdW force is given by:

$$V_a = -\frac{HR}{6D^2}, \quad (8)$$

where H is the Hamaker constant and depends on the dielectric properties of the interacting materials and the medium in between. The van der Waals force is usually attractive with very few exceptions. In CPFM, between inorganic surfaces except for DLVO interaction, colloidal probes experience an extra hydration repulsion force at smaller probe–substrate separation with a probable steric or entropic origin. However, the evaluation of hydration forces is often misconstrued with the roughness of the substrate because the range of surface forces has the same order of magnitude as the surface roughness. Conversely, at long-range probe–substrate separations, a hydrodynamic drag force also originates from the speed of colloidal probe movement during the approach toward the substrate surface (Fig. 8).

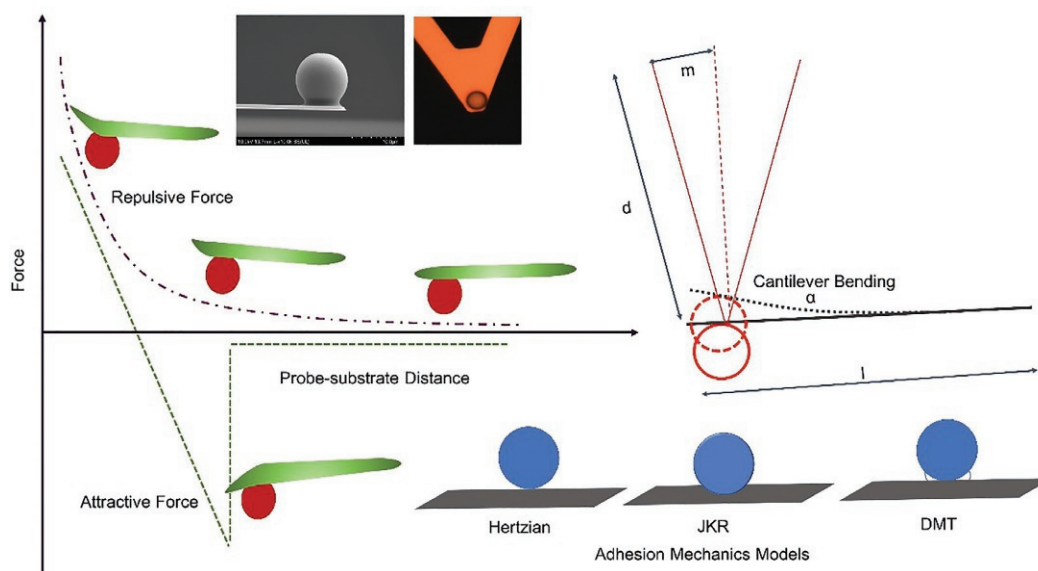


Fig. 8 A schematic diagram of probe–substrate interactions and deflection of the cantilever in repulsive and attractive regimes. Three different models of probe–substrate adhesion mechanics. SEM and optical microscopy images of the colloidal silica probe. Based on Butt H-J, Cappella B, Kappl M (2005) Force measurements with the atomic force microscope: Technique, interpretation and applications. *Surface Science Reports* 59(1–6): 1–152 and Israelachvili J (2011) *Intermolecular and Surface Forces*. Amsterdam: Academic Press.

Experimental setup

The experimental setup for CPM was first introduced by [Ducker et al. \(1991\)](#), where the sharp AFM tip is replaced by a spherical or quasi-spherical colloid particle of microscale dimensions including silica, glass, iron oxide, and polystyrene beads. The bead is fixed to the tipless cantilever with epoxy glue aided by a micromanipulator followed by measurement of the spring constant with thermal tuning of the probe. At large separations where the interactions between the probe and substrate are minimal, the cantilever deflection is small. However, at a specific probe–substrate separation, when cantilever deflection results from an interaction between the probe and sample, the change in separation is a combination of static deflection and piezo displacement. At smaller distances, a repulsive force acts between the tip and sample, and the cantilever bends upward. The tip and sample are not in contact, therefore, this region is often referred to as the noncontact region. In the case of attractive interaction energy, the force gradient exceeds the spring constant resulting in spring instability and the two surfaces accelerate toward each other. Cantilever deflection is measured by the optical lever technique, first a beam from a laser diode is focused onto the end of the cantilever and the position of the reflected beam is monitored by a position-sensitive detector (PSD). The reverse-side of the cantilever is usually covered with a thin layer of gold to enhance its reflectivity. When a force is applied to the probe, the cantilever bends and the reflected light beam moves through an angle equal to twice the change in the end slope $\tan \alpha$ ([Butt et al., 2005](#)). For a cantilever with a rectangular cross-section of width w length l and thickness h the change in the endslope is

$$\tan \alpha = \frac{Fl^2}{2E_c I}, \quad (9)$$

where E_c is the elastic modulus of the cantilever, and $I = \frac{wh^3}{12}$ is the moment of inertia.

If the detector is at distance d from the cantilever and with a small deflection of the cantilever if the laser moves a distance of m , then

$$m = 2d \tan \alpha = \frac{Fl^2}{E_c I}, \quad (10)$$

the distance d away from the cantilever, for a small cantilever its deflection is given by:

$$\Delta z_c = \frac{Fl^3}{3E_c I} = m \frac{l}{3d} \quad (11)$$

Except for the DLVO forces, the abundance of NOM in the soil can significantly affect interparticle interactions. The molecular conformation of the biomolecules adheres to either substrate or the probe under compression between the approaching surfaces, and can induce long-range electrosteric interactions. The confinement of the adsorbed polymer between the probe and the substrate, and their large radius of gyration of the adsorbed polymeric chains provide an entropic contribution or steric force ([Biggs, 1995](#)).

The underlying principle of measuring adhesion force by AFM is based on the principles of 'contact mechanics' first proposed by Hertz, where the interaction force is purely repulsive on contact. However, Johnson, Kendall, and Roberts (JKR) accounted for the theory of adhesion between two elastic bodies where experimentally measured contact areas were larger than that proposed by Hertzian interactions, specifically at low loads or zero applied load. Derjaguin–Muller–Toporov (DMT) also derived a separate expression to include adhesion in the contact of elastic bodies, where the contact profile remains Hertzian and the attractive forces act like an additional external load ([Israelachvili, 2011](#)). The contact mechanics in the JKR model is a short-range adhesion force, whereas the DMT model represents a long-range surface force. The differences between these two models were resolved by Tabor who had pointed out that both the JKR and DMT models are at opposite ends of the same spectrum ([Johnson and Greenwood, 1997](#)). Experimentally, the actual adhesion mechanism between the elastic probes and the substrates follows the Lennard Jones potential (V_{LJ}) a weak form of van der Waals attraction according to the following equation ([Israelachvili, 2011](#)):

$$V_{LJ} = 4\varepsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right], \quad (12)$$

where V_{LJ} is the intermolecular potential between the two atoms or molecules. ε is the potential well depth and a measure of how strongly the two particles attract each other, σ is the distance at which the intermolecular potential between the two particles is zero, and r is the separation distance between the centers of the interacting particles.

Application of CPM techniques

The environmental mobility and groundwater transport of colloids, adhesion of structurally different particulates, attachment of bacteria to mineral substrates, and stable soil aggregate formation depend on surface forces. The pore water chemistry including pH, ionic strength, and molecular makeup of the NOM molecules play a key role in interparticle interaction prevalent in unsaturated soil. The CPM was successfully applied to evaluate the variation in EDL repulsion between a colloidal silica probe and silica substrates with variation in surface density of the silanol groups ([Butt et al., 2005](#)). Stronger EDL repulsion with a spherical silica probe was recorded for a substrate with a 10° water contact angle relative to the substrate surface with a 45° contact angle. Therefore, this emphasizes the role of nanoscale interactions in the macroscopic properties of mineral substrates.

The interaction forces between a colloidal silica probe with silica and mica substrates have significant implications the coherence of soil mineral particles, essential in primary soil aggregate formation ([Wang et al., 2021](#)). Previous macrolevel studies have shown the importance of sesquioxide minerals as cementing agents in stable soil aggregate formation. In this context, interaction of the

colloidal silica probe with naturally occurring amorphous iron oxyhydroxide (FeOOH), and hematite-modified silica and mica substrates showed strong pH-dependence regarding probe–substrate adhesion. Our recent investigation on colloidal forces between a silica probe and small-sized hematite NPs layer-coated mica (Hm-mica) substrates in the presence of 5 and 50 mg C L⁻¹ with carboxylate-rich HA and phenol-rich TA, and alkaline earth metal ions and the organic guanidinium cation (Gd⁺[(NH₂)₃C⁺]) revealed the relative adhesion efficiencies regarding the coherence of soil mineral particles. In the absence of cationic species, the high surface charge density of the ionized low pK_a carboxylates in HA, isolated from the Amherst peat soil encapsulating the HA-modified Hm-mica contributed to long-range EDL repulsion with the silica probe as seen from the approach curve. However, the lack of ionized carboxylate moieties in TA showed reduced probe–substrate repulsion. The weakly hydrated alkaline earth metal ions induced charge screening and minimized the long-range EDL repulsions between the probe and the substrate and eventual ‘jump-to-contact’ due to attractive van der Waals force (Fig. 9). Screening of the EDL charge in HA-modified Hm-mica followed the order Mg²⁺ < Ca²⁺ < Sr²⁺ < Ba²⁺, in the reverse order of hydrated ionic radius. The retract curves exhibited ion-specific interactions-mediated charge screening by the weakly hydrated cations and possible formation of a ‘pinned micelle’ with exposed hydrophobic domains, which contributed minimally to adhesion with the silica probe (Wang et al., 2021).

Unlike carboxylate-rich HA, weakly hydrated alkaline earth metal ion-mediated polarizability of the π -electron cloud in phenol-rich TA and the formation of a porous hydrogel-like structure that maintains interfacial water holding capacity contributed to the strong adhesion with the silica probe. The weak adhesion between silica probe and HA-modified Hm-mica in the presence of alkaline earth metal ions showed significant improvement in adhesion in the presence of the organic Gd⁺ cation. Barely hydrated Gd⁺ interactions with ionized carboxylate and the formation of salt bridges with hydrogen bond donor-acceptor interactions and the formation of water-enriched hydrogel-like structures contributed to the probe–substrate adhesion (Fig. 10).

The interfacial porous nanotubular structures produced from HA interactions with Gd⁺ determined from the AFM height imaging also supplemented the AFM force data. Therefore, ion specificity of the cationic species, the concentration of NOM molecules, and the formation of novel nanostructures with enhanced water holding capacity may contribute substantially to soil aggregate stability against destabilizing shear forces during slaking.

Biologically active CPM has been used to measure the interactions quantitatively between *Shewanella oneidensis* (a dissimilatory metal-reducing bacterium) and goethite (α -FeOOH), both commonly found on Earth in near-surface environments (Lower et al., 2001). Biologically active force probes are prepared by attaching bacterial cells to a colloidal bead to the end of a silicon nitride cantilever. The force–distance curves between the probe and mineral surface under aerobic and anaerobic conditions have shown that the affinity between *S. oneidensis* and goethite rapidly increased under anaerobic conditions due to electron transfer from the bacterium to the mineral surface. The worm-like chain (WLC) model fitting in the adhesion force data revealed that the outer membrane of *S. oneidensis* specifically interacts with the goethite surface facilitating electron transfer.

The results of the adhesion forces data using CPM need to be judged carefully. Both the roughness of the probes and substrate surfaces specifically at higher loading caused sticking of the probe to the substrate surface. The first contact of the probe with any roughness on the hard substrate and topographic inhomogeneity determines the ‘zero distance’ and creates a potential wall during retraction of the probe; therefore, this affected the plausibility of the colloidal probe data. Butt et al. (2005) suggested reporting the root mean square (RMS) roughness of the substrate surfaces, but also cautioned that it may not be an absolute measure and could sometimes be misleading.

Chemical force microscopy (CFM) and dynamic force spectroscopy (DFS)

Intermolecular interactions are critical in the molecular conformation, adhesion, and sorption behavior of soil organic matter (SOM) on the mineral–water interfaces. The potential energy landscapes of intermolecular interactions can be effectively determined from the interactions of the harmonic oscillator available in tiny springs attached to the AFM cantilevers together with other techniques including bioforce probes and or optical traps. The seminal work by Bell on the kinetics of forced bond rupture by applying the external force or loads based on the exponential amplification theory of Kramers (1940), which states that the shuttling action of Brownian motion can set free a particle from the potential energy barrier and create a transition state, thereby a method for calculating the rate of a chemical reaction (Friddle et al., 2008 and references therein). Evans and others had proven that the force required to rupture an adhesive bond depends upon the loading rate and shape of the interaction potential. Therefore, measuring the bond rupture forces over a range of loading rates is termed dynamic force spectroscopy (DFS). With the absolute force resolution in AFM, it is feasible to probe individual molecular interactions.

In DFS spectroscopy, a functionalized AFM probe is allowed to interact with a substrate until a bond is formed followed by pulling out of the cantilever from the surface until the bond is broken. In DFS, probes rupture forces under two different regimes specifically: (a) at a large range of pulling velocities, which explores bond rupture dynamics in a near-equilibrium regime i.e. stochastic, and (b) in a non-equilibrium regime with strong rates of pulling. Both DFS and CFM have been useful to probe nanoscale forces at the pN level. Nevertheless, CFM differs from DFS because in the former rupture forces occur at a single pulling velocity. A general perception about these techniques can be viewed from the difference between a harmonic and anharmonic potential energy diagram defined from the Morse potential (Fig. 11) (Friddle et al., 2008). The ratio of the rates of inherent binding to unbinding in equilibrium states provides information about the efficacy of bond formation against thermal fluctuations-driven bond breaking. In contrast, in a non-equilibrium regime bond-breaking becomes deterministic. Sand et al. (2017) indicated that a systematic variation in the rate of loading ($\dot{r}$) could be used as an analytical function for mean rupture force vs rates of loading to

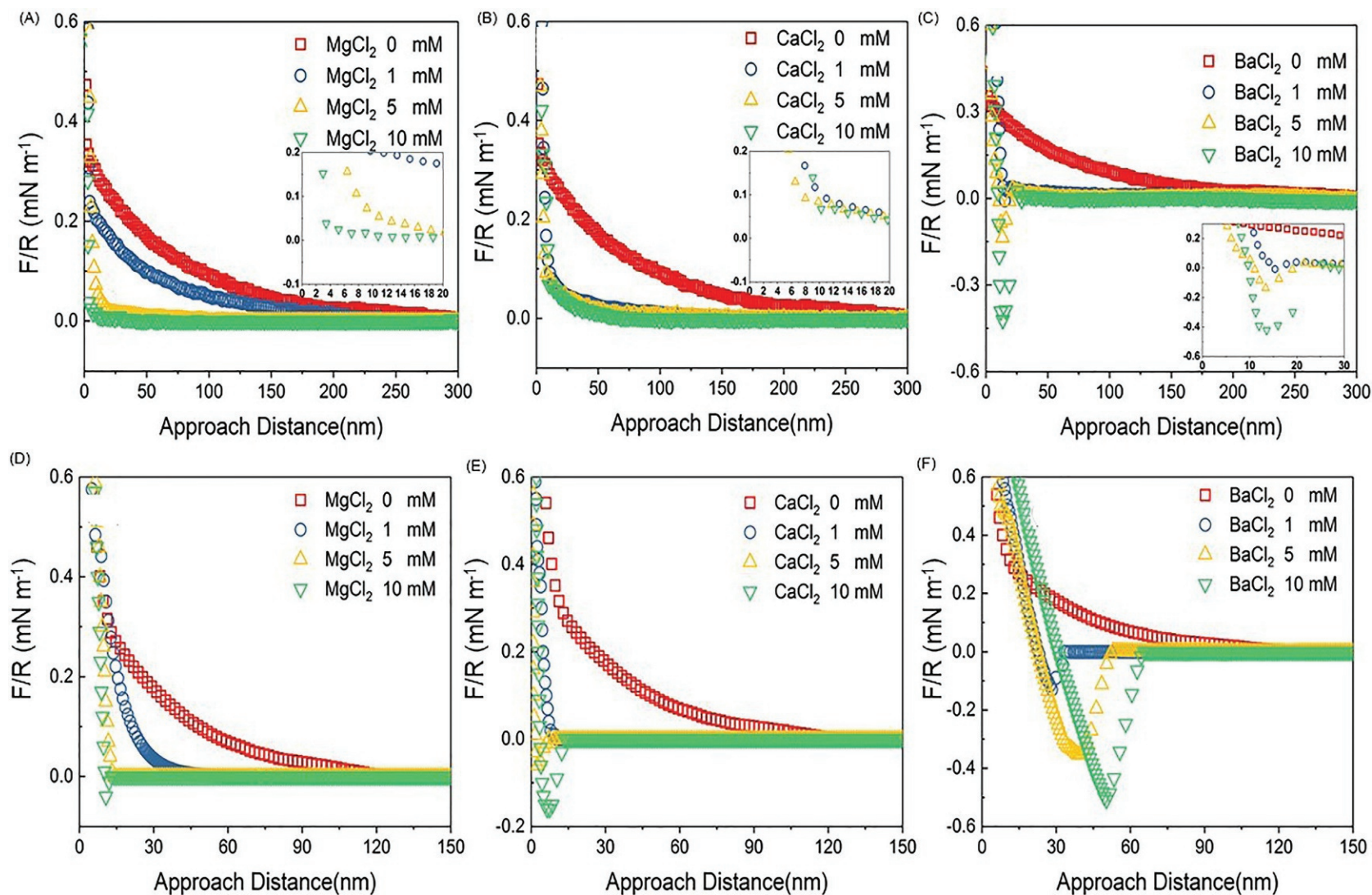


Fig. 9 The approach curves of a silica probe and (A–C) HA-modified, (D–F) TA-modified hematite-coated mica (Hm-mica) substrates in the presence of different concentrations of Mg^{2+} , Ca^{2+} , and Ba^{2+} . The large carboxylic content in HA contributed to stronger repulsive interactions between the silica probe and substrate relative to TA-modified Hm-mica in the absence of any added electrolyte (Wang et al., 2021).

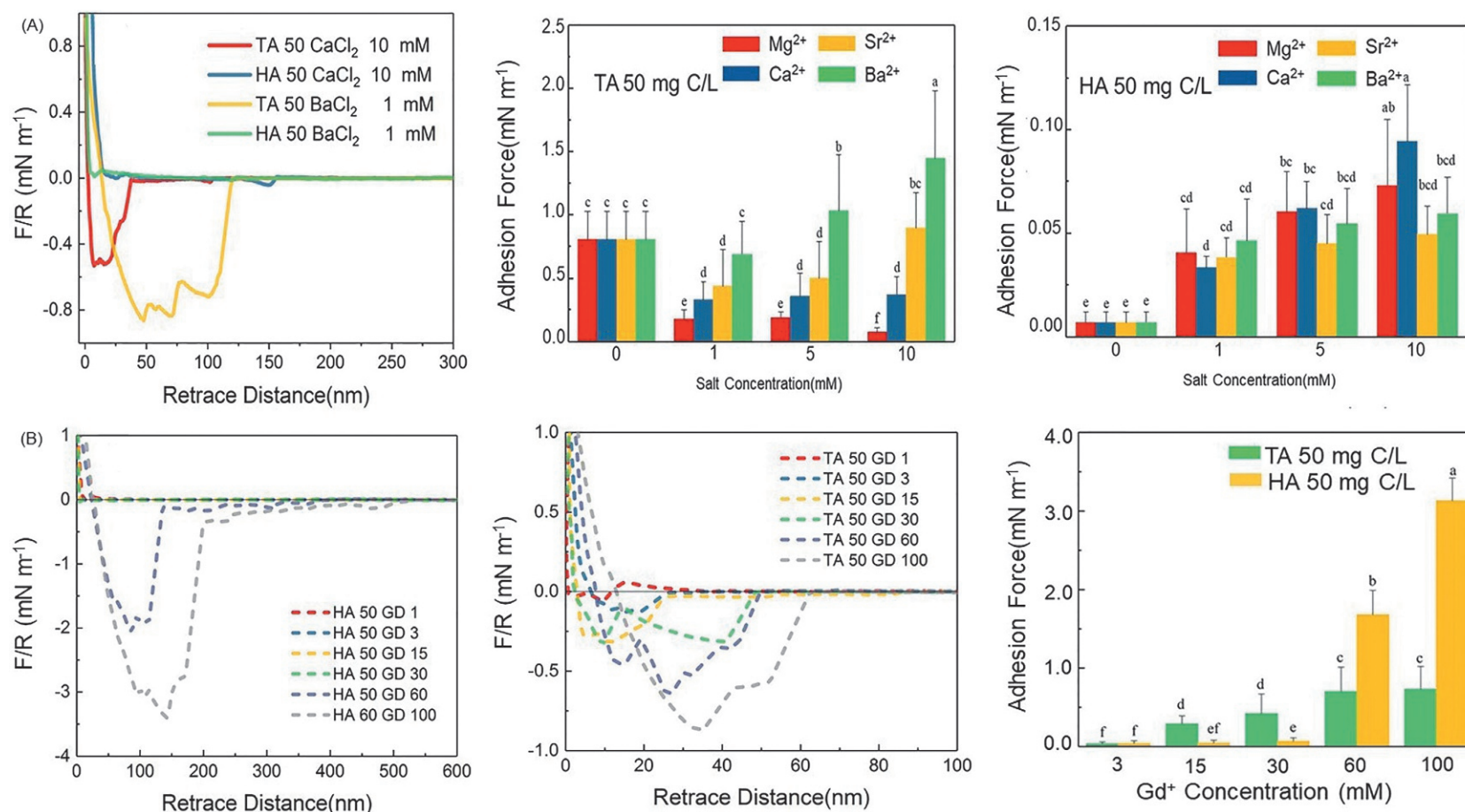


Fig. 10 (A) Retrace curves of the silica probe and 50 mg C L⁻¹ HA- and TA-modified Hm-mica substrates in the presence of 1 and 10 mM Ca²⁺. Statistical distribution of the adhesion force data of TA- and HA-modified Hm-mica in the presence of inorganic cations. The retrace curves of (B) HA-modified and TA-modified Hm-mica substrates in the presence of Gd³⁺ and their statistical distribution of the adhesion force data in HA- and TA-modified Hm-mica with increasing Gd³⁺ concentrations (Wang et al., 2021).

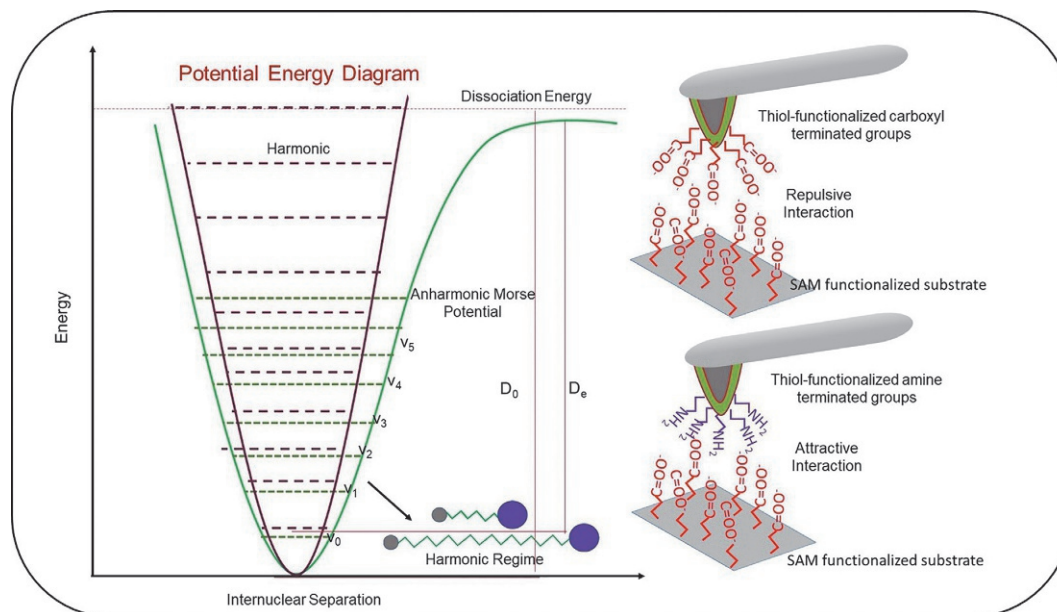


Fig. 11 Morse potential energy diagram and fundamental mechanism of CFM/DFS.

determine the behavior of biopolymer adhesion with mineral substrates. A molecular picture of the soil was obtained from quantification of the organo–mineral binding by Newcomb et al. (2017) using DFS spectroscopy. Briefly, DFS measurements were carried out between Au-coated AFM probes functionalized with -COO^- , PO_4^{3-} , NH_3^+ , and CH_3 groups and 001 faces of muscovite mica, and the 010 faces of goethite at varying rates of loading to estimate the binding strength between model organics and minerals, (Newcomb et al., 2017). A similar approach to force measurement was applied earlier in determining the adhesion behavior of biomineralized nacre, mother of pearl, where long titin protein molecules were attached to the AFM probe forming the molecular spring (Smith et al., 1999).

Concluding remarks

To understand the natural soil environmental processes requires an evaluation of the nanoscale phenomena occurring on environmental interfaces. The generality of AFM applications and its nanoscale resolution has shown enormous benefits in understanding the geochemical process to contaminant dynamics. The AFM has shown significant utility in identifying the bacterial adhesion to mineral substrates to NOM-assisted interfacial transformation of engineered nanomaterials and heteroassociation with soil mineral colloids. In addition, the role of solution chemistry and subsequent alteration in biomolecular conformation on the contaminant dynamics were determined unequivocally with AFM. Furthermore, the effect of mineral substrate surface symmetry and exchangeable cation-assisted electric dipole on the growth geometry of the self-assembled colloid crystals can be elucidated with AFM. In addition, the dislocation-driven epitaxial growth of colloidal crystals ascertained from force microscopic techniques can be extended to identifying the complex biomineralization processes occurring in soils and sediments. Furthermore, the determination of nanoscale forces with AFM to the coherence of soil mineral particles enriches our understanding of soil aggregate formation and empowers us to meet the imminent environmental challenges.

References

- Biggs S (1995) Steric and bridging forces between surfaces bearing adsorbed polymer: An atomic force microscopy study. *Langmuir* 11(1): 156–162.
- Butt H-J, Cappella B, and Kappel M (2005) Force measurements with the atomic force microscope: technique, interpretation and applications. *Surface Science Reports* 59(1–6): 1–152.
- Chan SSF and Green J-BD (2006) Tapping mode imaging and measurements with an inverted atomic force microscope. *Langmuir* 22(15): 6701–6706.
- Deng W, Zhang G-M, Murphy MF, Lilley F, Harvey DM, and Burton DR (2015) Analysis of dynamic cantilever behavior in tapping mode atomic force microscopy. *Microscopy Research and Technique* 78(10): 935–946.
- Ducker WA, Senden TJ, and Pashley RM (1991) Direct measurement of colloidal forces using an atomic force microscope. *Nature* 353(6341): 239–241.
- Frank FC (1951) Capillary equilibria of dislocated crystals. *Acta Crystallographica* 4: 497–501.
- Friddle RW, Podsiadlo P, Artyukhin AB, and Noy A (2008) Near-equilibrium chemical force microscopy. *Journal of Physical Chemistry C* 112: 4986–4990.

- Ghosh S, Mashayekhi H, Pan B, Bhowmik P, and Xing B (2008) Colloidal behavior of aluminum oxide nanoparticles as affected by pH and natural organic matter. *Langmuir* 24(21): 12385–12391.
- Ghosh S, Mashayekhi H, Bhowmik P, and Xing B (2010) Colloidal stability of Al_2O_3 nanoparticles as affected by coating of structurally different humic acids. *Langmuir* 26(2): 873–879.
- Ghosh S, Pradhan NR, Mashayekhi H, Dickert S, Thantrige R, Tuominen MT, Tao S, and Xing B (2014) Binary short-range colloidal assembly of magnetic iron oxides nanoparticles and fullerene (nC_{60}) in environmental media. *Environmental Science & Technology* 48(20): 12285–12291.
- Ghosh S, Guo Q, Wang Z-Q, Zhang D, Pradhan NR, Pan B, and Xing B (2019) Tannic acid-and cation-mediated interfacial self-assembly and epitaxial growth of fullerene (nC_{60}) and kaolinite binary graphitic aggregates. *Journal of Colloid and Interface Science* 556: 717–725.
- Guo Q, Wang Z-Q, Xu Q, Mao H, Zhang D, Ghosh S, Pradhan NR, Pan B, and Xing B (2020) Suspended state heteroaggregation kinetics of kaolinite and fullerene (nC_{60}) in the presence of tannic acid: effect of π - π interactions. *Science of the Total Environment* 713: 136559.
- Hochella MF Jr., Rakovan JF, Rosso KM, Bickmore BR, and Rufe E (1999) 'New directions in mineral surface geochemical research using scanning probe microscopes', ACS Symposium Series. *American Chemical Society: Washington, DC* 715: 37–56.
- Israelachvili J (2011) *Intermolecular and Surface Forces*. Amsterdam: Academic Press.
- Johnson KL and Greenwood JA (1997) An adhesion map for the contact of elastic spheres. *Journal of Colloid and Interface Science* 192: 326–333.
- Kramers HA (1940) Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* 7: 284–304.
- Liu C, Frenkel AI, Vairavamurthy A, and Huang PM (2001) Sorption of cadmium on humic acid: mechanistic and kinetic studies with atomic force microscopy and X-ray absorption fine structure spectroscopy. *Canadian Journal of Soil Science* 81(3): 337–348.
- Liu C, Leng W, and Vikesland PJ (2018) Controlled evaluation of the impacts of surface coatings on silver nanoparticle dissolution rates. *Environmental Science & Technology* 52(5): 2726–2734.
- Lower SK, Hochella MF Jr., and Beveridge TJ (2001) Bacterial recognition of mineral surfaces: nanoscale interactions between *Shewanella* and α -FeOOH. *Science* 292(5520): 1360–1363.
- Magnetic Force Microscopy (MFM) <https://www.parksystems.com>.
- Martin Y, Williams CC, and Wickramasinghe HK (1987) Atomic force microscope—force mapping and profiling on a sub 100-Å scale. *Journal of Applied Physics* 61(10): 4723–4729.
- Maurice PA (1996) Applications of atomic-force microscopy in environmental colloid and surface chemistry. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* 107: 57–75.
- Miyata K, Tracey J, Miyazawa K, Haapasilta V, Spijker P, Kawagoe Y, Foster AS, Tsukamoto K, and Fukuma T (2017) Dissolution processes at step edges of calcite in water investigated by high-speed frequency modulation atomic force microscopy and simulation. *Nano Letters* 17(7): 4083–4089.
- Myneni SCB, Brown JT, Martinez GA, and Meyer-Illse W (1999) Imaging of humic substance macromolecular structures in water and soils. *Science* 286(12): 1335–1337.
- Namjesnik-Dejanovic K and Maurice PA (2000) Conformations and aggregate structures of sorbed natural organic matter on muscovite and hematite. *Geochimica et Cosmochimica Acta* 65(7): 1047–1057.
- Newcomb CJ, Qafoku NP, Grate JW, Bailey VL, and De Yoreo JJ (2017) Developing a molecular picture of soil organic matter–mineral interactions by quantifying organo–mineral binding. *Nature Communications* 8: 396.
- Nonnenmacher M, O'Boyle MP, and Wickramasinghe HK (1991) Kelvin probe force microscopy. *Applied Physics Letters* 58: 2921–2923.
- Palermo V, Palma M, and Samori P (2006) Electronic characterization of organic thin films by Kelvin probe force microscopy. *Advanced Materials* 18(2): 145–164.
- Pan B, Ghosh S, and Xing B (2008) Dissolved organic matter conformation and its interaction with pyrene as affected by water chemistry and concentration. *Environmental Science & Technology* 42(5): 1594–1599.
- Plasche M, Römer J, and Kim JI (2002) Characterization of Grolle groundwater colloids by atomic force microscopy. *Environmental Science & Technology* 36(21): 4483–4488.
- Reifenberger R (2016) *Fundamentals of Atomic Force Microscopy Part I (Volume 4): Foundations*. Singapore: World Scientific Publishing Co. Pte. Ltd.
- Sand KK, Friddle RW, and DeYoreo JJ (2017) Quantifying the free energy landscape between polymers and minerals. *Scientific Reports* 7: 8663.
- Smith BL, Schäffer TE, Viani M, Thompson JB, Frederick NA, Kindt J, Belcher A, Stucky GD, Morse DE, and Hansma PK (1999) Molecular mechanistic origin of the toughness of natural adhesives, fibers and composites. *Nature* 399: 761–763.
- Wang Z-Q, Feng B, Zhang D, Ghosh S, Pan B, and Xing B (2021) Role of NOM-Hematite nanoparticle complexes and organic and inorganic cations in the coherence of silica and clay particles: Evaluation based on nanoscale forces and molecular self assembly. *Environmental Science: Nano* 8(3): 822–836.
- Zisman WA (1932) A new method of measuring contact potential differences in metals. *Review of Scientific Instruments* 3: 367–370.

Applications of nuclear magnetic resonance for the study of soils

Daniel H Lysak, Myrna J Simpson, and Andre J Simpson, Environmental NMR Centre, University of Toronto Scarborough, Toronto, ON, Canada

© 2023 Elsevier Ltd. All rights reserved.

Introduction	605
A brief explanation of nuclear magnetic resonance	606
Common nuclei in NMR	606
Choosing a nucleus	607
An overview of different NMR techniques	607
Preparing the NMR experiment	607
Solution-state NMR	609
Solid-state, HR-MAS and CMP NMR	610
Which NMR technique should I use and when?	611
Low field NMR, hyphenated NMR and magnetic resonance imaging (MRI)	611
NMR for structural elucidation	612
Multidimensional NMR	615
Quantitative NMR	616
Beyond structure—Studying soil interactions via NMR	618
Chemical shift and relaxation changes	618
Diffusion-based techniques	618
Nuclear overhauser effect-based techniques	619
Saturation transfer difference (STD) techniques	620
Studying interactions via CMP NMR	620
Studying interactions via solid-state NMR	621
Sensitivity improvements in NMR	622
Conclusion	623
References	623

Abstract

Nuclear Magnetic Resonance (NMR) is a powerful analytical tool with the ability to not only identify chemical structures, but to examine non-covalent interactions in soils. Here, a range of modern NMR techniques is examined and their potential for soil research is summarized. In addition, a brief description of the pertinent theory and sample preparation, along with examples from the literature and practical experimental considerations is presented. The overall goal is to provide an overview and showcase of the strengths and capabilities of NMR, which is accessible to specialists and non-specialists alike.

Key points

- Provide a brief background of NMR Spectroscopy
- Summarize the wide range of NMR experiments available
- Provide an overview of sample preparation
- Address NMR in solid, gel, liquid and multi-phase samples
- Cover NMR in terms of both structure elucidation and non-covalent interactions

Introduction

Whether used to identify unknown compounds in a soil extract, examine the carbon profile of a dried soil, or study interactions between soil and pesticides, nuclear magnetic resonance (NMR) is a powerful analytical tool for the study of soils. When applied to its full potential, NMR can not only give an unparalleled understanding of the composition and structure of soil, but is arguably the most powerful method for studying non-covalent interactions such as aggregation and both covalent/non-covalent soil-contaminant binding. This is an increasingly important area of study, as an understanding of contaminant binding is essential to the understanding of long-term soil fertility, contaminant fate/transport, as well as developing the most effective remediation strategies. There have been numerous reviews addressing the application of NMR to soil research (Kögel-Knabner, 1997; Simpson et al., 2011; Cardoza et al., 2004; Mao et al., 2017; Simpson and Simpson, 2014), as well as some introductory feature articles (Simpson et al., 2012, 2018). Here we do not aim to provide a comprehensive literature review, but to explore and describe the

power of NMR for soil research from a practical perspective. We will focus on the wealth of NMR techniques available and their applications to soil research through the use of literature examples, with the goal of making the NMR of soil more accessible.

A brief explanation of nuclear magnetic resonance

Prior to a discussion of the potential and specifics of NMR as a tool for the study of soils, it is helpful to lay a groundwork for understanding NMR. A rigorous explanation of the theory behind NMR cannot be done justice in a few paragraphs, nor is it fitting with the ethos of making NMR accessible and thus will not be attempted here. Instead, we aim to provide a baseline explanation of NMR phenomena, which may help in understanding the requirements for a successful NMR experiment.

In contrast to most common spectroscopy techniques which involve electrons, NMR is based on the difference in energy between spin states of the nucleus in the presence of a magnetic field. A prerequisite for any NMR experiment is that the desired nucleus of study must have a nuclear spin. Fortunately, elements such as hydrogen, carbon, nitrogen, phosphorous and fluorine which are common in soil samples (and contaminants/agrochemicals) all have spin active stable isotopes (specifically ^1H , ^{13}C , ^{15}N , ^{31}P and ^{19}F). When a magnetic field is present, a nucleus with non-zero spin can align with the magnetic field, resulting in a lower energy state, or against the magnetic field, resulting in a higher energy state. This energy splitting is known as the Zeeman effect and the energy difference between these two states is exploited in the NMR experiment. On average, slightly more nuclear spins are aligned with the field than against it, creating a net magnetization, which can be manipulated by pulsing the sample with radiofrequency (RF) energy. Manipulation of the sample's magnetization forms the basis for all NMR experiments, ranging from simple 1-D experiments to complex multinuclear multidimensional NMR. Using RF pulses, the magnetization can be "flipped" from its equilibrium position (aligned with the external magnetic field), into a plane perpendicular to the magnetic field, where it can be detected. After this flip, the magnetization, like a spinning top that is pushed off axis, begins to precess around the magnetic field. The frequency of this precession is known as the resonant (or Larmor) frequency and is unique to the chemical environment of the nucleus. The sensitivity of the resonant frequency with respect to changing chemical environments is precisely what makes NMR a powerful tool for both identifying chemical structures and examining non-covalent interactions.

Common nuclei in NMR

Prior to examining the differences between the types of NMR, the nucleus under study is important to consider as it determines the type of information resulting from the experiment as well as the difficulty and length of the experiment. In general, ^1H is by far the most commonly studied nucleus due to its excellent NMR sensitivity (of note, ^1H is used in the vast majority of studies in the solution-state, but is rarely used for solid-state analysis as dipolar couplings lead to very broad and often unresolvable spectra). In the solid-state, ^{13}C is the most commonly studied nucleus. When considering the use of NMR to study a specific nucleus, there are many factors to consider: the gyromagnetic ratio, natural isotopic abundance, nuclear spin (i.e., spin $\frac{1}{2}$ or quadrupolar), relaxation time and spectral dispersion. The gyromagnetic ratio is an inherent property of a nucleus and the NMR signal scales with this value to the third power (Webb, 2012). Thus, considering only this factor, ^{13}C already gives ~ 64 times less signal than ^1H . However, natural abundance, which describes the proportion of the specific nucleus found in natural samples, is also an important factor. For ^1H , the natural abundance is $\sim 99.99\%$, whereas for ^{13}C it is only $\sim 1.1\%$, further making ^{13}C ~ 2 orders of magnitude more difficult to study. Overall, the receptivity of ^{13}C at natural abundance relative to ^1H is $\sim 0.02\%$ (Harris et al., 2001). That said, difficult does not mean impossible and modern NMR techniques are commonly used to study carbon in whole soils, though these experiments are much longer than proton studies. The ^1H , ^{13}C and ^{15}N nuclei make up the vast majority of soil NMR studies, but research has been done on other nuclei such as ^{19}F , ^{31}P and other more niche nuclei including various metals, silicon and deuterium (Simpson et al., 2011). With regard to the nuclear spin, nearly all nuclei studied in soils have a spin of $\frac{1}{2}$. This is beneficial as compounds that have a spin greater than $\frac{1}{2}$ (known as quadrupoles) possess electric quadrupole moments which interact with nuclear spins, leading to significant spectral broadening and overlap. This can often make spectra very difficult to interpret and is the main reason that ^{15}N is studied (0.137% abundance) compared to the much more abundant, but quadrupolar ^{14}N ($\sim 99.6\%$ abundance). The values for gyromagnetic ratios and abundances quoted here are sourced from the IUPAC recommended values, as outlined in the work by Harris et al. (2001).

Another factor that affects the time required for analysis is the relaxation time, which changes with both the nucleus studied and the sample itself. In simple terms, after each RF pulse and data acquisition, some time must be allotted to allow the spins to relax to the equilibrium state. This time is known as the recycle delay and limits the rate at which experiments can be repeated. Since NMR is inherently insensitive, signal averaging is central to NMR experiments, and, crucially, is linked to the recycle delay. The greater the number of "scans" (or transients) that can be acquired per unit time, the greater the signal to noise ratio (SNR). Specifically, the SNR scales with a factor of $n^{1/2}$, where n is the number of scans. Especially in environmental mixtures such as soil, many scans (for example 50,000 for ^{13}C analysis) can be required to analyze samples, and thus shorter relaxation times are typically beneficial. Properties of the sample such as the solvent or presence of paramagnetic species (e.g., iron, oxygen) have significant effects on relaxation time. That said, relaxation times in environmental samples (e.g., soils, water samples) tend to be substantially lower than for small molecules (e.g., organic chemistry samples), and it is important to ensure that the recycle delay matches the relaxation time of the sample to ensure the optimal SNR. The nucleus studied also has a significant impact on the recycle delay and ^1H tends to have

much shorter relaxation times compared to ^{13}C and ^{15}N , though the value of the latter can be decreased by the addition of a relaxation agent such as chromium(III) acetylacetonate.

Choosing a nucleus

Given the factors discussed above, it seems easy to conclude that ^1H NMR is simply better than ^{13}C or ^{15}N NMR. While in terms of ease of experiment, this is probably the case, it is not the whole story. First, it is important to note that these analyses give different information. For example, quaternary carbons cannot be detected directly by ^1H NMR. Further, both ^{13}C and ^{15}N benefit from significantly improved dispersion, as the chemical shift is spread over >200 ppm, whereas ^1H chemical shifts typically do not extend past ~ 12 ppm. In ^{13}C spectra, due to the rarity of these nuclei at natural abundance (as well as ^1H decoupling), there is virtually no splitting, leading to less spectral overlap, and less complex line shapes than ^1H experiments. This of course begs the question: “which nucleus should be used to study soils?” While the answer is unique to the research question at hand, a few general recommendations can be made.

For solution-state NMR, ^1H is the easiest to implement and thus its use would be prudent, providing it addresses the research question. In addition, a range of “inverse detected” 2-D solution-state experiments exist (see section on multidimensional NMR) that retain sensitive ^1H detection, but, if used to their full potential, can provide information on all carbons indirectly (including quaternary carbons). However, if the analysis is in the solid-state, ^{13}C NMR is recommended, given that ^1H in the true solid-state is relatively uninformative. For an excellent review on solid-state ^{13}C NMR of soils, see the work by Preston (Preston, 2001). Due to its difficulty, ($\sim 2\%$ of the receptivity of the already challenging ^{13}C) ^{15}N NMR is recommended only when information on nitrogen is the major focus of the research work.

Two other nuclei are of note: ^{19}F , which possesses excellent sensitivity ($\sim 80\%$ of ^1H) and ^{31}P which has moderate sensitivity ($\sim 7\%$ of ^1H). These are recommended if the research question specifically involves compounds that contain these atoms. For example, ^{19}F has been used to study the interactions of fluorinated compounds with soil (Masoom et al., 2015), and ^{31}P can be used to examine phosphorus speciation and cycling (Vestergren et al., 2012). As a final note, there are several multidimensional NMR techniques which can involve multiple nuclei and can be very useful to enhance dispersion and elucidate chemical structures. These are discussed in the multidimensional NMR section.

An overview of different NMR techniques

Though classically NMR analysis could only be carried out in either the solid (i.e., a dry soil) or solution-state (e.g., a soil extract), recent developments in NMR technology have allowed much greater latitude in sample preparation and analysis. A general outline of the broad types of NMR techniques is given in Table 1. In soil research, the most commonly used forms of NMR are solid-state NMR, followed by solution-state NMR. Here we will discuss and provide examples of the various NMR techniques used to study soils. Sample preparation for solution-state NMR is very different from the other major techniques (solid-state, high-resolution magic angle spinning (HR-MAS) and comprehensive multiphase (CMP) NMR) and these will be discussed separately.

Preparing the NMR experiment

In common with any analytical technique, there are a number of factors that must be addressed to obtain high quality spectra, and will be discussed here. For more detail on experimental preparation, see Simpson et al. (2011). Though most of these processes were once done by hand, modern spectrometers tend to have automated programs to perform these steps, and, if available, this is generally recommended. After placing the sample in the spectrometer, it is critical to wait for the temperature to stabilize prior to starting experiments (ideally before shimming), as this can help reduce temperature induced spectral drift and helps ensure reproducibility. Once a sample is stable in the spectrometer, a spectrometer lock (if available) should be performed, typically on a deuterated solvent. The spectrometer lock helps to adjust for magnetic field fluctuations, giving more reproducible spectra with sharper line shapes. This is especially important for long experiments since the risk of field drift becomes larger with time. After locking, the sample should be shimmed to obtain improved line shape. Shimming involves adjusting the current through smaller electromagnets around the central magnet to homogenize the magnetic field in the sample region as much as possible, which gives rise to optimal line shape. This is a crucial step and proper shimming can be the difference between a broad, relatively uninformative profile (i.e., badly shimmed) and an information rich profile potentially containing 1000 s of partially resolved peaks (well shimmed). Over the years, shimming (especially for solution-state experiments) has improved dramatically and modern gradient-based shimming methods are highly recommended, if available.

Calibrating the pulse length is an important factor and one that greatly influences the SNR. If the pulse length is not set correctly, the SNR of the entire experiment can suffer significantly. In natural samples it is not uncommon to see differences up to 100% in pulse length between samples, (Simpson et al., 2011) and not calibrating for such a change could result in cutting the SNR in half. Thus, it is imperative to calibrate the pulse length on a per sample basis. This can be done via an automated process on most modern spectrometers or by incrementing increasing pulse lengths and selecting the one that gives the largest SNR (or $1/4$ of the null at the

Table 1 A brief description of modern NMR techniques.

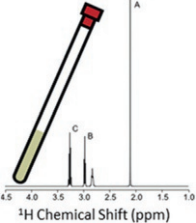
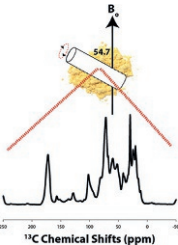
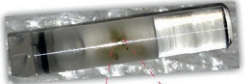
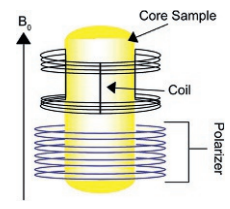
Technique	Description	Comments
Solution-State NMR 	Provides highest resolution but only for soluble components. The approach is fully quantitative if collected under appropriate conditions. For samples with low salt cryogenically cooled probes can increase signal to noise. Both helium and nitrogen cooled versions are available. For samples with high salts using a small diameter room temperature probe (1 mm or 1.7 mm) instead of a cryoprobe is a good option.	Probes range from 1 mm (5 μL) to 10 mm (4 mL) in size with 5 mm (600 μL) the most common. For a 5 mm cryoprobe, the detection limit for individual compounds is in the ng range, but as environmental mixtures can contain a large number of components 50–100 mg is recommended (1D and 2D NMR). Smaller probes are ideal when sample is mass limited. For mass-limited samples 1–3 mg is possible with a 1.7 mm probe. Only use a 1 mm probe when there is <1 mg sample.
Solid-State NMR 	Important for environmental analysis as it provides an overview of the types of carbon in a sample. This sample is spun at magic angle (54.7°) to narrow line shape and high power ^1H decoupling is required to improve carbon lineshape. The technique can be quantitative but considerable care must be taken if absolute quantification is required.	4 mm rotors are the most common and spin up to 15kHz. 4 mm rotors often require 80–120 mg of sample. 7 mm rotors hold ~400 mg of sample and thus can give more signal. However, RF homogeneity reduced and max spinning speed only ~7 kHz. Sidebands appear in solid-state NMR, they can be either moved out of the spectral window by spinning fast enough or can be suppressed using various methods such as TOSS.
High Resolution Magic Angle Spinning NMR 	This technique invented in 1996 combines magic angle spinning with a lock and magic angle gradient. Sometimes called “gel-phase” NMR, the approach allows swollen components to be studied with resolution similar to that in the solution state. When using ^1H both true solutions and swollen components are detected but true crystalline solids are not (the approach uses low power electronics and therefore cannot decouple the strong ^1H—^1H dipolar interactions in crystalline solids).	The approach has huge potential for study of soils, plants, air-particles and even small organisms in their swollen state. Samples can be either swollen in the rotor (for example soil + water) or put in whole with the natural water as acting as the solvent (i.e., a piece of carrot). Provides information on moieties swollen by water. In soil for example the approach highlights the all-important soil-water interface. As with solid-state NMR rotors are used, but with additional liquid seals. In general solution state NMR experiments are used in HR-MAS studies.
Comprehensive Multiphase NMR 	Introduced in 2012, CMP combines solids and solution state NMR. The approach can be thought of as a HR-MAS probe with the added ability to handle high power and therefore perform solid-state experiments. As such the probe can access <i>all components</i>, liquids, gels and solids in a sample. When applied to samples in their swollen state it provides a unique insight into conformation and structure in the natural state.	Key technique to follow processes that involve solids, being converted into liquids or vice versa. Examples include, flocculation, swelling, drying, growth, reactions, conformation change or sequestration of a liquid contaminant. For studies looking at non-dynamic multiphase systems combining a HR-MAS and solids probes will give similar results to CMP. However for dynamic systems (components dynamically changing phase) a CMP probe is ideal.
Hyphenated NMR 	Becoming commercially available in the 1990's, the approach combines separation with NMR. The most common setup being HPLC-NMR. More advanced setups also integrate SPE (for concentration) and MS. In principle, one can acquire NMR spectra (provides connectivity) and mass spectra (provide molecular formulae) to solve specific unknowns de-novo.	In practice for complex environmental samples, neither MS or NMR are the limitation, with the separation being the bottleneck. For example, in a recent 2D HPLC-NMR study of DOM even after fractionation into 10,000 fractions each still contained 100's of molecules making them too complex for the assignment of exact structures by MS or NMR. Hyphenated NMR is currently best used for non-targeted screening of less complex matrices, such as waste water or identification of targeted unknowns within mixtures.

Table 1 (Continued)

Technique	Description	Comments
Low field NMR 	Low field NMR encompasses many areas including mobile NMR and field cycling. Field cycling maps the relaxation times across a range of field strengths and provides novel information on dynamics. Portable NMR includes low field instruments that largely reproduce higher field NMR experiments but at low field and in a greatly reduced size and cost format. Specific instruments such as well loggers provide relaxation information can then identify the quality and existence of crude oil deposits.	Low field NMR has considerable potential for environmental research. Once purchased, low field NMR units based around permanent magnets require only a power supply to operate. Currently monitoring sites under remediation is extremely costly and can involve daily, weekly or monthly sampling combined with GC-MS and/or LC-MS analysis. If mobile NMR could be used in some of these applications it could greatly reduce cost as well as providing complimentary and real-time monitoring. The drawback is the reduced resolution and sensitivity over higher field instruments.
Magnetic Resonance Imaging	While spectroscopy focuses on chemical environments of nuclei within a sample, imaging produces visual images based on the water distribution within the sample. MRI is appealing as it provides a visual overview and internal structure that is inaccessible to most techniques.	MRI has huge potential to monitor and quantify larger scale processes such as contaminant transport, water swelling and provide key information on pore sizes and dynamics. It is also possible to perform localized spectroscopy in MRI.

Reproduced with permission from Simpson AJ, Simpson, MJ, Soong R. (2018). Environmental nuclear magnetic resonance spectroscopy: An overview and a primer. *Analytical Chemistry* 90: 628–639. Copyright 2018 American Chemical Society.

360° angle). Of note, in experiments that use multiple pulses, for example 2-D NMR or more complex 1-D experiments, pulse calibrations are even more essential. In such cases, a mis-calibrated pulse does not just produce smaller SNRs, but incomplete spin flips, meaning that the desired interactions (coherences) do not form and the experiments essentially fail to produce the desired signals.

The receiver gain, which determines the “sensitivity” of the receiver is another important factor. For challenging and dilute samples, the gain must be set to the maximum for optimal SNR. That said, a particularly large peak (usually due to a non-deuterated solvent) may “overflow” the receiver, degrading the spectrum and this can be avoided by decreasing the gain until the overflow no longer occurs. The final factor that will be discussed here is the recycle delay. This is the time required between scans to allow the spins to relax to equilibrium. The relaxation time can be found from an inversion recovery experiment (see the quantitative NMR section), but for qualitative or semi-quantitative spectra it is common to use a shorter recycle delay.

Solution-state NMR

As the name implies, solution-state NMR requires a liquid sample. In soil research, this typically involves a soil extract such as the important and often studied soil humic and fulvic acid extracts. In solution-state NMR ^1H is the most commonly studied nucleus (since it is the most sensitive nucleus at natural abundance), but there has also been research on ^{13}C , ^{15}N , ^{19}F and other less common nuclei. Soil extracts are typically concentrated via lyophilization and redissolved in a deuterated solvent for solution-state analysis. The use of a deuterated solvent is important for two main reasons: it allows spectrometer locking (which compensates for instrumental drift, leading to more reproducible spectra) and prevents receiver dynamic range issues (i.e., receiver overload) due to the solvent which would otherwise be present in much greater concentrations than the analyte. The choice of solvent is typically dependant on the solubility of the sample. Sodium deuteroxide/deuterium oxide ($\text{NaOD}/\text{D}_2\text{O}$) or deuterated dimethyl sulfoxide ($\text{DMSO}-d_6$) are commonly used solvents in soil analysis, but other solvents such as deuterated methanol or dichloromethane (DCM) can also be employed. Deuterated aqueous solvents are often used to study dissolved organic matter, but have the disadvantage of suppressing signals from exchangeable protons. $\text{DMSO}-d_6$ on the other hand, is a very effective solvent; not only does it help dissolve many compounds that are water insoluble, but it retains exchangeable protons (providing the solvent is dry), and gives sharp line shapes (helping to prevent spectral overlap). However, it is also important to note that $\text{DMSO}-d_6$ will give a signal in ^{13}C spectra, while $\text{NaOD}/\text{D}_2\text{O}$ will not (Simpson et al., 2011). Further, DMSO is an incredible solvent but is only effective when the soil is in the protonated form, which can be achieved by H^+ -ion exchange or by adding acid.

The mass of sample required for NMR analysis depends on the type of NMR probe used. In general, a greater sample mass means that there are more spins to be analyzed, and thus more signal, providing that the sample is not limited by solubility. A standard solution-state probe is 5-mm in diameter with $\sim 600\ \mu\text{L}$ of sample, and is excellent for most applications. Some ~ 50 – $200\ \text{mg}$ of sample are recommended for a 5-mm probe. While the NMR detection limits for individual compounds can be on the order of nanograms (ng), soil is a complex mixture of hundreds of thousands of compounds, and thus much larger masses are required for analysis. Larger probes up to 8 or 10-mm (a few mL) are also available if there is sufficient sample, however, RF homogeneity in larger probes is often reduced and pulse field gradients (required for diffusion, complex 2D NMR and advanced water suppression)

are not always available. On the other hand, if the sample is mass limited (i.e., only a very small mass is available) smaller diameter probes such as 1.7 or 1-mm probes can be used. The benefit is that smaller diameter probes show a greater mass sensitivity (i.e., signal per unit mass). These probes allow analyses of samples in the 1–3 mg range. In general, if sufficient mass is available, the use of a 5-mm probe is recommended and these have been the most commonly used for the study of soil extracts.

Solid-state, HR-MAS and CMP NMR

All of these techniques involve packing a sample into a small rotor, which is then rapidly spun. The sample is spun at the “magic angle” ($\sim 54.74^\circ$) relative to the magnetic field, which helps reduce broadening due to anisotropic interactions in solid and semi-solid components. Despite the spinning, the anisotropic interactions are still strong enough so that solid-state spectra have broader line shape than solution-state spectra. Further, pure solid probes typically do not have a lock channel (leading to potential chemical shift drift) or pulsed field gradient (useful for many 2-D experiments and diffusion-based spectral editing). Pure solid NMR uses dry samples, for example soils that were heated to remove moisture. The sample should be packed as tightly as possible to increase the mass of sample to be analyzed. The size of the rotor determines the maximum sample that can be added, with 4-mm rotors being the most common and holding ~ 100 mg of sample. Larger 7-mm (~ 300 – 400 mg sample mass) rotors are available, but they cannot be spun as fast, leading to potential issues with line broadening and spinning side bands (a form of spectral artefact). For an excellent review on solid-state NMR analysis of natural organic matter (including soil) see the treatise by Mao et al. (2017).

The HR-MAS and CMP NMR methods also employ rotors, but can additionally analyze solution and gel-like samples, allowing the study of interactions between the physical phases. Both techniques typically use a gradient and a lock, allowing for more reproducible and gradient-based experiments. HR-MAS can be used to analyze the solution-state and semi-solid/gel components of a sample. In a “hydrated” soil, this would include any compounds present in the solution-state, and the gel-like, hydrated edges of the soil. Further, the gel-like domains do not exhibit such strong anisotropic interactions compared to the solid-state, leading to line shapes that are better than solid-state (though still not as good as the solution-state). The CMP method takes the benefits of HR-MAS a step further and allows analysis of all phases in the sample resulting in liquid, gel and solid sub-spectra from complex heterogeneous samples. An example of the use of CMP for soil structural elucidation is shown in Fig. 1 (Masoom et al., 2016). Here, three spectra resulting from different components of the swollen soil are present along with assigned peak regions and a 2-D NMR spectrum of the swollen soil shows the additional spectral dispersion and information that can be elucidated through CMP analysis of unaltered soil.

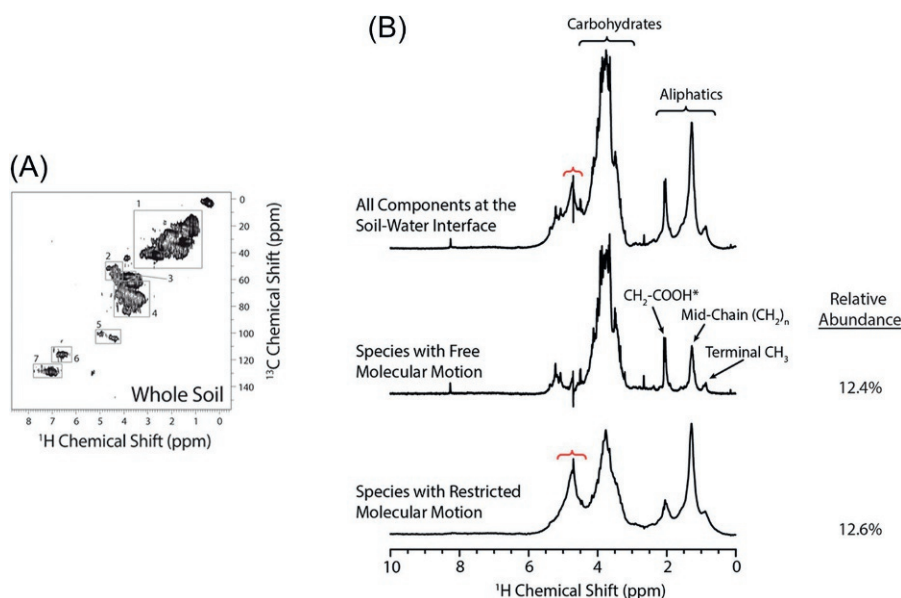


Fig. 1 (A) The 2-D ^1H – ^{13}C Heteronuclear Single Quantum Coherence (HSQC) spectrum (see Table 2) and (B) ^1H spectra of a whole soil recorded using a comprehensive multiphase NMR probe. The numbered boxes in the 2-D spectrum indicate the peak regions corresponding to the resonances as follows: (1) Aliphatic single bonds and amino acid side chains, (2) Alpha proton on amino acids, (3) Methoxy peak from lignin, (4) CH_2 groups from carbohydrates, (5) Anomeric protons from carbohydrates, (6) Aromatic lignin and phenylalanine, (7) Aromatic signals from tyrosine and tryptophan. The three 1-D spectra indicate different components of the soil, separated based on their molecular motion. Adapted with permission from Masoom H, Courtier-Murias D, Farooq H, et al. (2016). Soil organic matter in its native state: Unravelling the most complex biomaterial on earth. *Environmental Science and Technology* 50: 1670–1680. Copyright 2016 American Chemical Society.

Which NMR technique should I use and when?

This is best addressed by a series of queries

1) Is it important to preserve all interactions or the exact structural profile of the sample as close as possible to its natural state?

If the answer is yes, and the material is a solution, then “direct NMR” may be the best approach. Direct NMR is simply the use of advanced water suppression that allows an aqueous based medium to be studied as it is (i.e., without sample preconcentration or alteration) (Lam and Simpson, 2008).

If the answer is yes, but the material is for example a swollen soil, then CMP NMR can be employed to study the solutions, swollen components and solids in the soil in their natural state (Masoom et al., 2016). The HR-MAS technique is also an option if only the swellable components are of interest.

2) Do I just need a compositional overview to compare samples across a series?

If the answer is yes, and the material is a solution, then traditional extraction or preconcentration and then solution-state NMR is recommended. Screening with 2-D NMR greatly increases resolving power, in addition to giving enhanced connectivity information, and is generally encouraged, if sufficient spectrometer time is available.

If the answer is yes, and the material is not a solution, then drying and studying the sample by solid-state ^{13}C NMR is an effective way to compare carbon distribution across samples.

3) Is there anything I should avoid?

Generally, it is not ideal to take a true solution, for example dissolved organic matter, dry it and study it as a solid. For example, consider the example of a solid NMR spectrum collected at 75 MHz ^{13}C (300 MHz ^1H) if an exponential function corresponding to 30 Hz line broadening is applied during processing this leads to a resolving power of ~ 400 –500 peak capacity across the ^{13}C spectral window. However, if the same sample were freeze dried redissolved and a 2-D solution-state NMR spectrum was collected, the resolving power could approach 2,000,000 (Hertkorn et al., 2007) providing a much deeper insight into a dissolved sample. Conversely, if a sample is not in solution in nature, then it is important to balance the potential fractionation and sample alteration that can occur during extraction.

Second, direct NMR and CMP-NMR are not always the “best.” As an example, for direct NMR even on a highly sensitive cryoprobe, it can take >24 h to profile a sample of water with only ~ 2 ppm dissolved organic matter. However, if a litre of the sample was freeze dried and the sample redissolved in 600 μL of solvent, a better SNR would be obtained in minutes. Thus, unless there is an overwhelming reason to study a sample without preconcentration (for example loss of volatiles on freeze drying, or the natural aggregated state being critical) then traditional concentration is recommended. A similar argument can be made for CMP-NMR. Consider, for example, a researcher who wants to see how carbon distribution changes across vegetation-rich horizons. If the plant material is taken as it is (i.e., no drying) it may be $\sim 80\%$ water. Of the remaining 20% perhaps only 40% is a “true” solid. This means that if CMP-NMR is applied, the solid fraction will give only $\sim 8\%$ of the signal compared to drying the entire sample, packing it into a rotor and analyzing it as a pure solid. The resultant information is, of course, different: CMP-NMR would give information on solids as they occur in the true natural state, but if the study aims only to study carbon distribution across different samples, then such information may not be required. As SNR in NMR is proportional to the square root of the number of scans, then in our example, the solid-state NMR of the dried sample could be collected ~ 150 times faster than studying the true solid fraction in the natural state using CMP-NMR.

Ultimately, the type of NMR most ideal for a particular study depends on the research questions being posed. The important point is that modern NMR now offers many new opportunities, techniques and experiments that were not available 20 years ago and it is critical that the next generation of soil researchers embrace all possibilities to propel the field forward efficiently.

Low field NMR, hyphenated NMR and magnetic resonance imaging (MRI)

While the focus of this chapter is mainly on high-field NMR spectroscopy, these three additional techniques can also be useful in the study of soils. Low field NMR refers to the use of much weaker (often permanent) magnets for sample analysis. This of course comes with two major disadvantages: lower resolution and sensitivity compared to high field spectrometers. However, there are also major advantages: low field NMR spectrometers are significantly cheaper, in addition to being portable. This means that, in theory, a low field spectrometer could be taken into the field to perform analyses. Although the lower resolution and sensitivity would pose a challenge, NMR technology is constantly improving and the potential for in-field measurements is increasing. Though standard spectroscopy techniques may be more challenging at low-field, fast field cycling (FFC) and relaxometry techniques are very feasible (Conte and Meo, 2020). The FFC method works by examining the changes in longitudinal relaxation time (T_1) at various field strengths. The strength of this technique lies in the fact that it can explore slow dynamic motions, such as those occurring at the surface of porous media, and that are often missed in high-field relaxometry. For example, soil pore size can be correlated with the relaxation time, helping to understand the underlying water–soil dynamics.

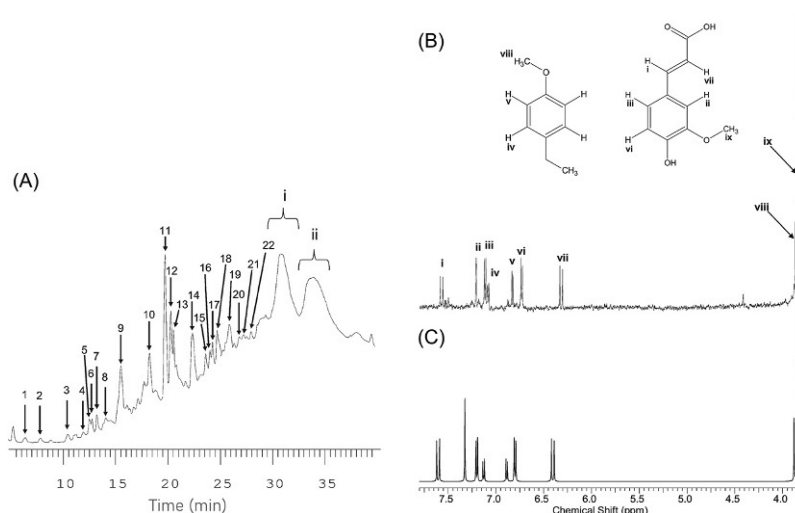


Fig. 2 (A) A chromatogram of a Harvard Forest fulvic acid sample. (B) The structures, NMR spectrum and peak assignments of the compounds corresponding to peak 19 in the chromatogram. (C) A computer predicted NMR spectrum for the compounds identified, used to increase confidence in the structural assignments. Adapted from Simpson AJ, Tseng LH, Simpson MJ, et al. (2004). The application of LC-NMR and LC-SPE-NMR to compositional studies of natural organic matter. *Analyst* 129: 1216–1222 with permission from the Royal Society of Chemistry.

Hyphenated NMR refers to the combination of NMR with additional analytical techniques. For example, liquid chromatography (LC) could be used to help fractionate the sample (decrease spectral overlap), followed by a solid phase extraction (to help concentrate the sample) and then split to NMR and mass spectrometry (MS) for detection. This would give the ultimate set of information about a sample: the NMR spectra could be used to elucidate the structures of the individual compounds that elute in the column and mass spectrometry could help detect compounds whose concentrations are insufficient for NMR. One example of this is a 2004 publication that demonstrated the preliminary use of both LC-NMR and LC-solid-phase extraction (SPE)-NMR to study natural organic matter in both soils and water samples (Simpson et al., 2004). Fig. 2 shows a chromatogram of a soil fulvic acid sample, along with the ^1H NMR spectrum from resulting peak 19.

Based on the ^1H spectrum and comparison to a predicted spectrum, peak 19 was identified to be a mixture of (2E)-3-(4-hydroxy-3-methoxyphenyl)acrylic acid and 1-ethyl-4-methoxybenzene. The study noted that although MS was not employed, it would further improve the ease of compound identification. Though LC-NMR-MS is a very powerful combination of techniques, it requires multiple instruments, and thus expertise in both MS and NMR to set up properly, and due to the complexity of soil, separation may not always be sufficient for unequivocal structural assignment. Nonetheless, hyphenated NMR retains considerable potential for the study of soil organics.

Magnetic resonance imaging utilizes similar principles to NMR, but results in an “image” of the sample as opposed to a spectrum. Commonly applied in the medical field, MRI uses a factor such as relaxation time to differentiate between different components of a sample. This has seen significant use in the study of aggregation and water transport processes in porous media. For example, the study by Pohlmeier et al. examined the change in water content of the soil root interface of a castor bean plant (Pohlmeier et al., 2010). Fig. 3 shows the setup of the plant growth itself, along with a relaxometry-based 3-D difference spectrum, showing the difference in water content after 6 days. It was found that there was a significantly greater change in water content near the upper central part of the soil, whereas the lower region did not have as much uptake. As another example, a water-soluble tracer could be used to see where or how quickly water can pass through a soil, giving an understanding of the soil’s permeability/water retention.

NMR for structural elucidation

A major benefit of NMR is the ability to elucidate unknown structures accurately. To do this, a variety of NMR experiments, outlined in Table 2 can be employed. In brief, by taking advantage of the long- and short-range coupling interactions between nuclei, it becomes possible to elucidate virtually any chemical structure. In fact, NMR is the only technique (except X-ray crystallography, which is limited to crystalline structures) that can solve chemical structures de-novo and without the use of libraries, making it an essential discovery tool.

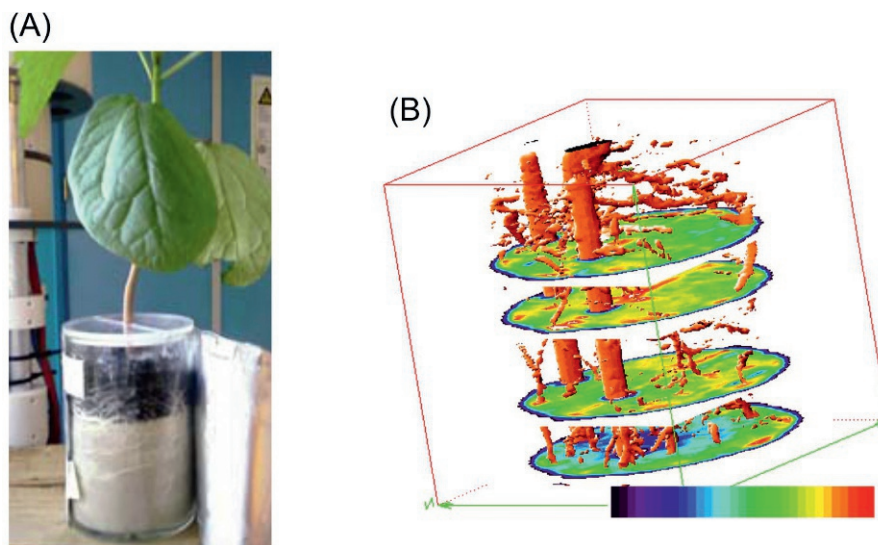
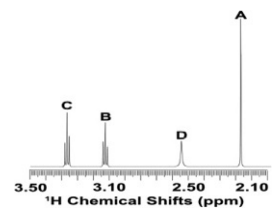
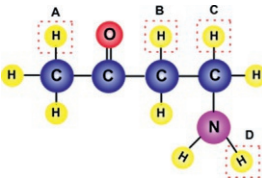
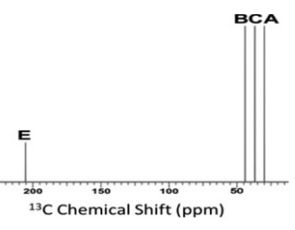
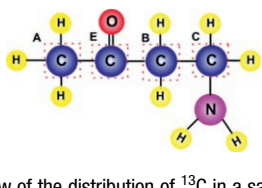
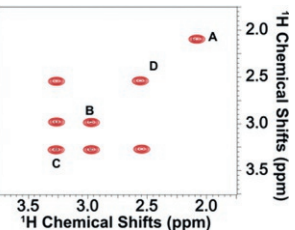
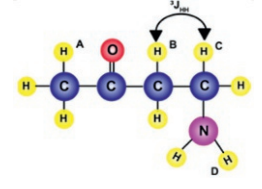


Fig. 3 (A) Image of a *Ricinus* plant prior to MRI analysis of the soil-root system. (B) The 3-D-MRI Rapid Imaging with Refocused Echoes (RARE) difference spectrum of the soil root system (right). A darker orange color indicates a larger change, and the two long cylinders are marker tubes. Adapted from Pohlmeier A, Vergeldt F, Gerkema E, et al. (2010). MRI in soils: Determination of water content changes due to root water uptake by means of a multi-slice-multi-echo sequence (MSME). *The Open Magnetic Resonance Journal* 3: 69–74.

Table 2 A brief summary of the main NMR experiments of use in structural elucidation in environmental analysis.

Experiment	Information	Interpretation
¹ H NMR 	 An overview of the distribution of ¹H in a sample.	¹ H chemical shift spans ~0–12 ppm. Multiplet patterns can be used to distinguish between neighboring protons through J _{HH} coupling constants. ¹ H in the D position is exchangeable and would not be present in D ₂ O.
¹³ C NMR 	 An overview of the distribution of ¹³C in a sample	Unlike ¹ H NMR, ¹³ C NMR has a much wider chemical shift range of ~250 ppm. Therefore, it exhibits much higher spectral resolution. ¹³ C suffers from lower sensitivity in comparison to ¹ H. ¹ H decoupling is required to remove ¹ J _{CH} couplings.
2-D ¹ H— ¹ H COSY 	 A liquid-state experiment identifying ¹H—¹H correlations between neighboring protons.	Correlation Spectroscopy (COSY) is a 2D NMR technique that shows the connectivity between neighboring protons. In this example, B couples to C so cross peaks are at the corners of a rectangle connecting the 2 protons. As COSY is only between neighbors, proton B does not correlate with proton D.

(Continued)

Table 2 (Continued)

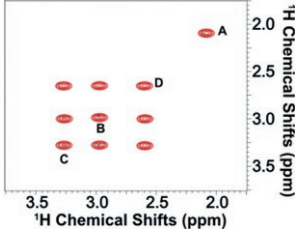
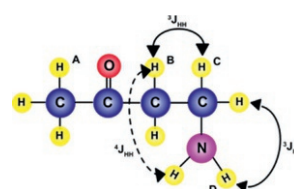
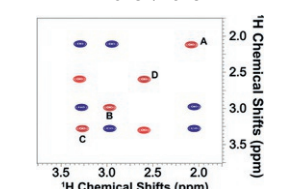
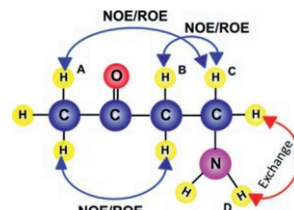
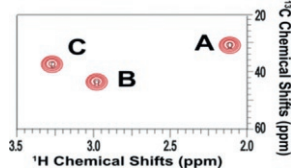
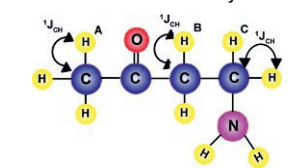
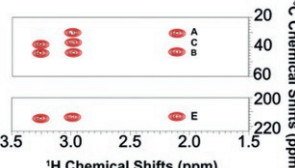
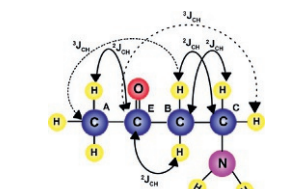
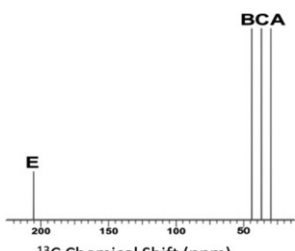
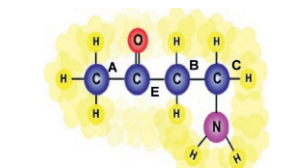
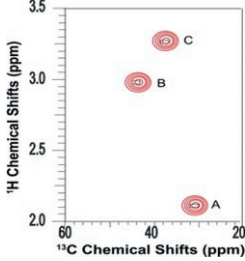
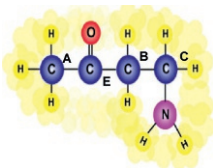
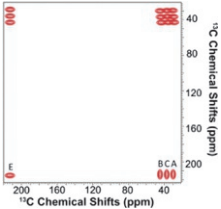
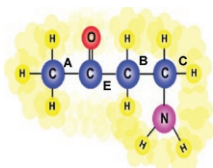
Experiment	Information	Interpretation
2-D ^1H—^1H TOCSY 	 A liquid-state experiment showing long range ^1H—^1H correlations.	Total Correlation Spectroscopy (TOCSY) 2D NMR technique that shows the connectivity of all protons in a spin system. Unlike COSY, TOCSY transfers through an entire ^1H—^1H network. Relays do not pass through quaternary positions so A cannot “see” B. However, B “sees” C and D as there is an unbroken network of ^1H spins connecting them.
2-D ^1H—^1H NOESY/ROESY 	 2D NMR technique that shows connectivity through space rather than bonds, in the liquid-state. Key experiment for conformational analysis.	Nuclear Overhauser Effect Spectroscopy (NOESY) and Rotating-Frame NOE Spectroscopy (ROESY). For small molecules in NOESY dipolar interactions through space are opposite phase to the diagonal and chemical exchange peaks the same phase. For large molecules both spatial interactions and the exchange have the same phase as the diagonal.
2-D ^1H—^{13}C HSQC 	 A liquid-state one bond ^1H—^{13}C correlation experiment.	Heteronuclear Single Quantum Coherence (HSQC) is a 2D NMR technique that correlates ^1H chemical shift and ^{13}C chemical shifts for directly bonded C and H. High spectral dispersion helps reduce overlap in mixtures. NOTE: Carbonyl is absent in an HSQC spectrum since there is no directly attached ^1H
2-D ^1H—^{13}C HMBC 	 A liquid-state long-range ^1H—^{13}C correlation.	Heteronuclear Multiple-Bond Correlation (HMBC) identifies connectivity between ^1H and ^{13}C up to 2 and 3 bonds away. Critical for piecing together molecular structures de-novo. Importantly, this technique is particularly useful as it shows connectivity with quaternary groups such as carbonyl.
^{13}C CP-MAS 	 A solid state NMR experiments that uses the polarization of neighboring protons to enhance ^{13}C sensitivity under MAS.	Cross polarization magic angle spinning (CP-MAS) is a solid-state experiment commonly used for characterizing environmental samples. CP increases sensitivity by up to 4 times and the recycle delay is determined by the ^1H relaxation time rather than ^{13}C making acquisition faster. If absolute quantification is required Direct Polarization (DP)-MAS (i.e., no CP enhancement) is the gold standard. However recent improvements in CP-MAS are also providing more quantitative results.

Table 2 (Continued)

Experiment	Information	Interpretation
2-D ^1H—^{13}C HETCOR 	 A one bond solid state ^1H—^{13}C correlation NMR experiment. (Note: Polarization transfer can occur between neighboring protons and a carbonyl group.)	Heteronuclear Shift Correlation (HETCOR) is a ^1H—^{13}C correlation NMR experiment often used in solid-state NMR. In this case, this is a ^{13}C detect experiment with CP as a transfer step. In solids interactions are via the stronger ^1H—^1H dipolar couplings rather than J-couplings. CP conditions have to be optimized carefully to observe only single bond ^1H—^{13}C correlations. Longer cross polarization times give rise to long range as well as short range correlations.
2-D ^{13}C—^{13}C DARR 	 2D solid-state ^{13}C—^{13}C correlation via dipolar couplings between ^{13}C spins.	Dipolar Assisted Rotational Resonance (DARR) is a solid-state ^{13}C—^{13}C correlation experiment. In this case, the correlations are facilitated via ^1H—^{13}C dipolar couplings under MAS conditions. Such experiments are likely too insensitive at natural abundance but are ideal for following the conversion of ^{13}C labelled biomass in the environment.

Reproduced with permission from Simpson AJ, Simpson MJ, Soong R. (2018). Environmental nuclear magnetic resonance spectroscopy: An overview and a primer. *Analytical Chemistry* 90: 628–639. Copyright 2018 American Chemical Society.

Multidimensional NMR

Multidimensional NMR offers much greater spectral dispersion, which enables distinction between peaks that would completely overlap in a 1-D spectrum. While the peak capacity (that is, the number of compounds that can be resolved under perfect conditions) in a 1-D ^1H spectrum is estimated to be a few thousand (and $\sim 30,000$ for ^{13}C), the value for a 2-D HSQC is $\sim 2,000,000$ and can jump to the order of 100,000,000 for 3-D spectra (Hertkorn et al., 2007). As such, multidimensional NMR has an unprecedented ability to help understand the chemical makeup of complex soil mixtures. An example of a spectrum from a commonly applied 2-D experiment (HSQC) of a soil is shown in Fig. 4 (Soucémariadin et al., 2017). It shows the wealth of information that can be gained from the additional dispersion onto two dimensions, as even individual compounds can be identified in certain cases.

It is also noteworthy that although Table 2 focuses mainly on ^1H and ^{13}C interactions, it is possible to utilize these experiments on any coupled nuclei (though sensitivity constraints must be considered). For example, an interesting recent publication studied the transformation of phosphorus in soil through the use of a ^1H — ^{31}P HSQC experiment, which can be found in Fig. 5 (Vestergren et al., 2012). Here, it was possible to differentiate between the various phosphorus species in the hydroxide extract of a Swedish soil. The authors noted that the signals “linked” by lines indicate resonances that were completely unresolvable in the 1-D ^{31}P spectra, showing the enhanced dispersion afforded by multidimensional techniques.

In addition to 2-D techniques, higher dimensional techniques have also been explored. The first application of 3-D NMR was possibly a 2003 study that used total correlation spectroscopy (TOCSY)-HSQC to assigned plant derived structures in soil (Simpson et al., 2003). However, the most impressive to date is arguably a 2015 study that utilized 4-D isotope filtered NMR to identify the substitution patterns of phenolic compounds from the fulvic acid of a peat soil. The 4-D spectrum is in Fig. 6 (Bell et al., 2015). Essentially, each point on the ^{13}C 1-D spectrum shown has an associated cube that further disperses the signal over two ^1H dimensions and another ^{13}C dimension.

Though these techniques provide excellent spectral dispersion, the major drawback compared to 1-D NMR is the length of experiment. To acquire a multidimensional spectrum, a specific delay in the experimental pulse sequence must be incremented. What this means is that the experiment must be repeated under slightly different conditions multiple times. These repetitions are known as increments and the number of increments determines the resolution in the second dimension. As such, it is common to need more than 100 increments to gather good quality 2-D data, leading to an experiment time that is orders of magnitude longer than for 1-D. With higher dimensional spectra this problem is further exacerbated, and many 3+ dimensional spectra, such as the one in Fig. 6 require costly isotopic labelling (i.e., increasing the amounts of ^{13}C or ^{15}N in the sample) to yield results in reasonable time frames. That said, certain techniques, such as non-uniform sampling (NUS) exist, and help decrease the acquisition time for multidimensional experiments in environmental samples (Farooq et al., 2013). Modern NMR processing involves a Fourier transform of the acquired signal, giving the well-known frequency absorption spectrum. However, the Fourier transform requires

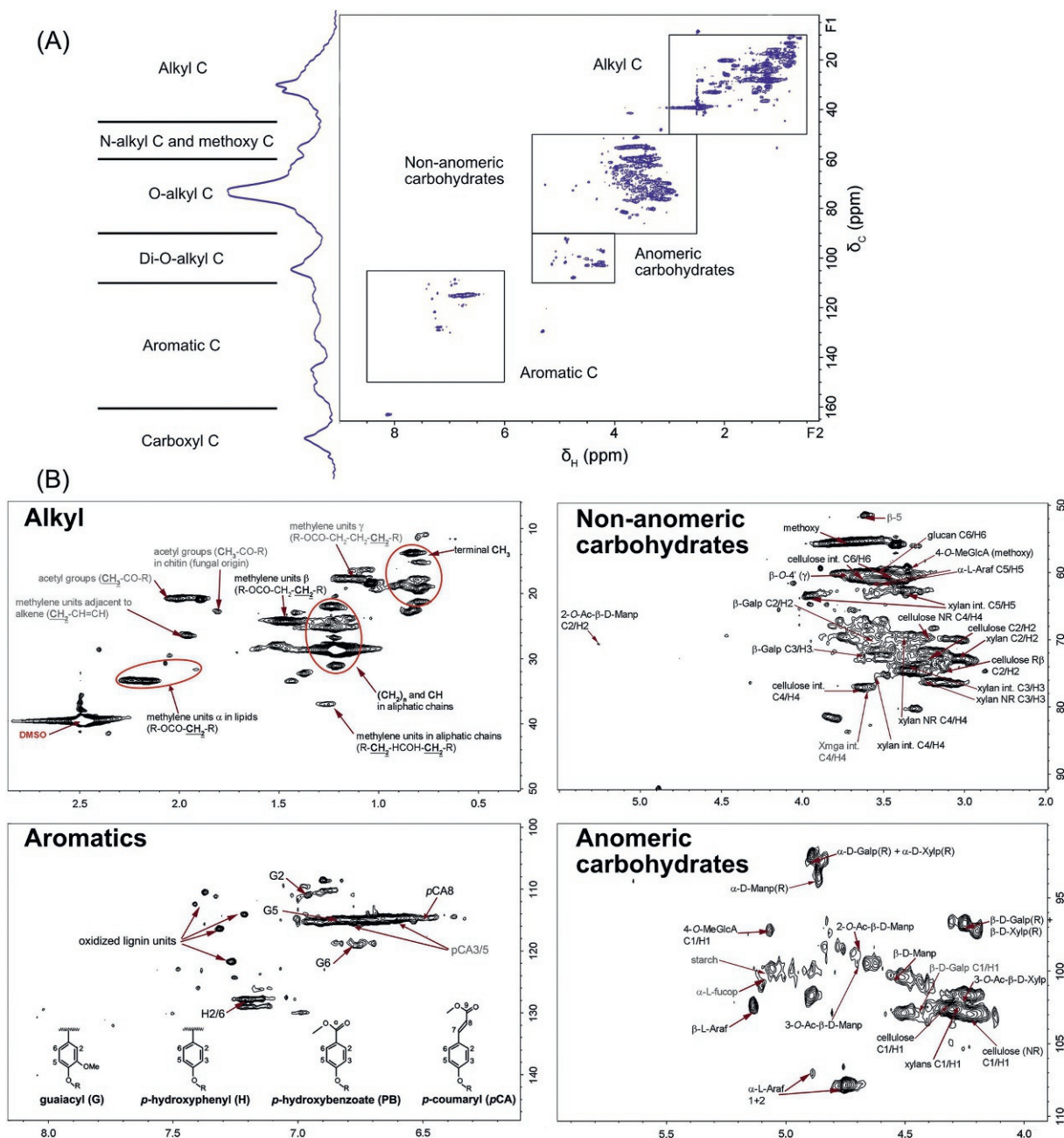


Fig. 4 (A) Overall 2-D ^1H - ^{13}C HSQC spectrum of a spruce soil DMSO extract, and (B) selected regions with identified compounds/moieties labelled. Adapted from Soucémariadin LN, Erhagen B, Nilsson MB, et al. (2017). Two dimensional NMR spectroscopy for molecular characterization of soil organic matter: Application to boreal soils and litter. *Organic Geochemistry* 113: 184–195 with permission from Elsevier.

linearly spaced points; more points give better resolution, but also take longer to acquire. With fast relaxing samples, such as of soil, there is much less signal in the later data points and using NUS (with a weighted bias towards earlier data points) enables acquisition with fewer points, allowing for more scans and thus a larger SNR during the same analysis time.

Quantitative NMR

Though the nature of NMR is inherently quantitative, there are some factors that must be considered to obtain reliable quantitative results. The main consideration is that all components for which quantification is desired must be given sufficient time to relax fully (i.e., return to equilibrium after pulsing), and this condition is considered met if the delay between scans is larger than $5 \times T_1$. T_1 is the spin-lattice relaxation time and can be determined from an inversion relaxation NMR experiment. Simpson *et al.* include a

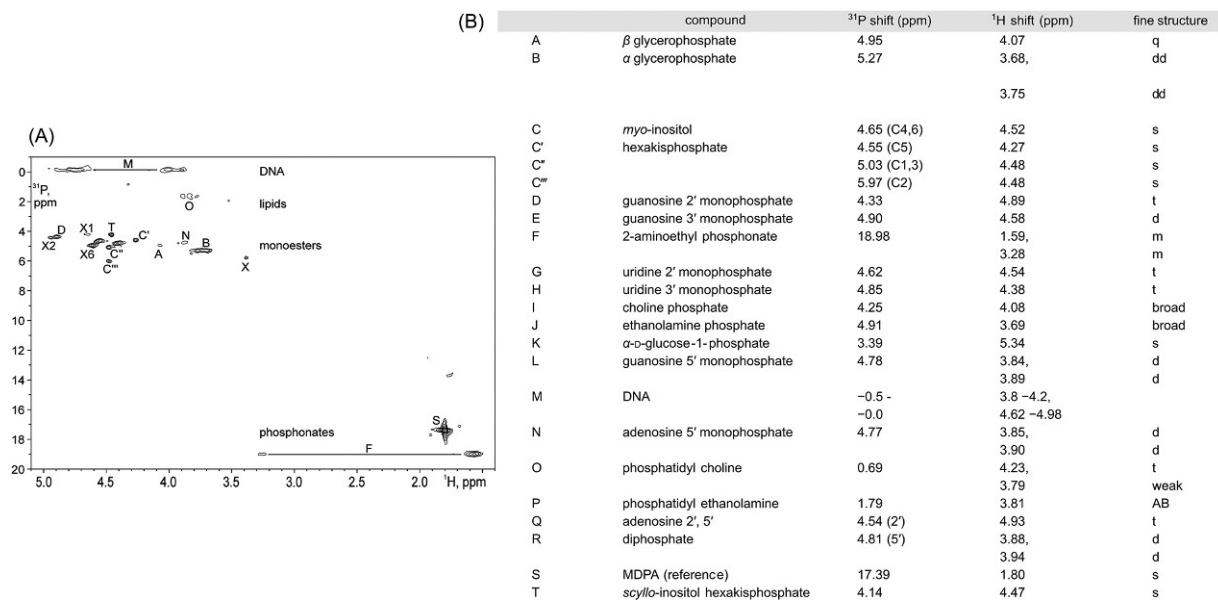


Fig. 5 (A) The ^1H – ^{31}P HSQC spectrum of a typical haplocryod soil. (B) The regions corresponding to different resonances are labelled, and some individual compounds are identified. X denotes unidentified compounds. The resonances linked by lines indicate overlapping peaks in the 1-D ^{31}P spectrum. Adapted with permission from Vestergren J, Vincent AG, Jansson M, et al. (2012). High-resolution characterization of organic phosphorus in soil extracts using 2D ^1H – ^{31}P NMR correlation spectroscopy. *Environmental Science and Technology* 46: 3950–3956. Copyright 2012 American Chemical Society.

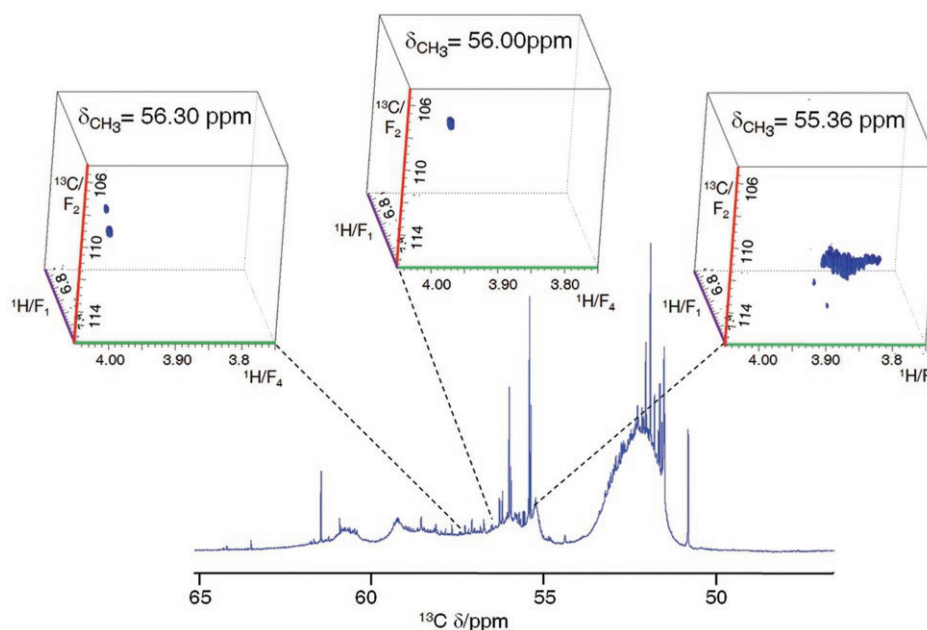


Fig. 6 The 4-D NMR spectrum of a ^{13}C labelled peat soil fulvic acid sample. This experiment was designed to characterize the phenolic fraction of the sample. Reproduced with permission from Bell NGA, Michalchuk AAL, Blackburn JWT, et al. (2015). Isotope-filtered 4D-NMR spectroscopy for structure determination of humic substances. *Angewandte Chemie, International Edition* 54: 8382–8385. Copyright 2015 John Wiley and Sons.

practical guide that explains how to estimate T_1 quickly for environmental samples (Simpson et al., 2011). At $5 \times T_1$, more than 99% of the magnetization has returned to equilibrium and the sample is ready to be pulsed again. The T_1 time depends on the sample solvent, presence of paramagnetic compounds and the individual chemical environments themselves. For example, aromatic or ketone compounds tend to relax more slowly as they do not have other protons in close proximity to which they can pass their magnetization. Most NMR experiments are not run under fully quantitative conditions. The reason for this is that an improved SNR in the same unit time can typically be achieved by using delay times $< 5 \times T_1$, often as low as $1\text{--}3 \times T_1$. Though the data cannot be directly quantified, if a qualitative spectrum is desired, the higher SNR is clearly advantageous, especially for longer, multidimensional experiments. Further, data resulting from this are still “semi-quantitative” in the sense that integrals from a spectral region, run under identical conditions, can still be used to compare relative amounts across samples (Simpson et al., 2011).

The other major consideration for quantification is the use of an analytical standard. NMR has the major benefit of needing only a single internal standard, which can be any detectable compound. This is especially useful for quantifying compounds with unavailable/expensive standards (Bharti and Roy, 2012). Further, external standards have also been shown to work very well for NMR quantification. In a method known as pulse length–based concentration determination (PULCON), an external standard sample with accurately known concentration can be used to “calibrate” the NMR, and the calibration obtained from that sample can be used for quantification of any other sample. Given an apt choice and proper storage of a primary standard sample, it can be used as a calibration for extended periods of time, recalibrating as necessary (for example, if the probe is changed, or after a duration of time). Within the detection range of NMR, it typically performs very well with regard to the accuracy of quantification, and shows unparalleled reproducibility (Bharti and Roy, 2012).

In the study of soils, it is common to employ cross polarization (CP) experiments for solid-state ^{13}C analysis. Cross polarization works by transferring magnetization from ^1H to ^{13}C via the dipolar interactions that dominate in the solid-state. This has two major benefits: by transferring magnetization from the proton, which has a higher gyromagnetic ratio, there is an up to ~4-fold improvement in SNR, and since the magnetization originates on the proton, the shorter proton relaxation time can be used for the recycle delay, allowing many more scans per unit time. Thus, CP results in both faster and more sensitive ^{13}C spectral acquisition. Similar polarization transfer can be done in solution-state NMR by insensitive nuclei enhancement by polarization transfer (INEPT) or distortionless enhancement by polarization transfer (DEPT). However, the disadvantage of these techniques is that quaternary carbons (i.e., that are not strongly proton coupled) will not be enhanced as much. This means that not only can some aspects of the sample be under or overrepresented, but that typical CP cannot be truly quantitative. Direct polarization (DP) is the gold standard for quantitative solid-state NMR, though recent work has demonstrated marked improvements in the quantitative ability of CP (Mao et al., 2017). In the solution-state, there is an additional consideration: quantitative carbon spectra require inverse gated decoupling (that is, decoupling only during the acquisition), as opposed to decoupling throughout the whole experiment, to prevent the (non-quantitative) Nuclear Overhauser Effect (NOE) enhancement of proton coupled carbons. This is of particular importance as most chemistry departments use NOE enhancement to improve SNR for routine chemical analysis. Thus, if quantitative data are required on soils, researchers should work with the NMR facility personnel to ensure the experiment and acquisition parameters are optimized rather than use a standard template on a “walk-up” system.

Beyond structure—Studying soil interactions via NMR

Studying non-covalent interactions is arguably the most unique ability of NMR. Examining aspects such as the aggregation of humic substances, or contaminant binding to soil are key aspects of soil research. Multiple methods have been developed which allow NMR to study such interactions and will be explored here. Though these mainly originate from protein research, they have since been successfully applied to the study of soils, and have significant potential in this field. It is also important to note that these techniques have been developed for solution-state studies, albeit comprehensive multi-phase and solid-state interactions are considered briefly at the end of this section.

Chemical shift and relaxation changes

One of the simplest methods for studying interactions is to examine changes in chemical shift. By comparing a spectrum of just one compound, with a spectrum of the compound plus its binding partner, a change in chemical shift may indicate binding. Fig. 7 shows an example of the changing chemical shift of 4'-Fluoro-1'-acetophenone, as the concentration of fulvic acid increases (Dixon et al., 1999).

Of note, the chemical shift changes are not very large, and could easily be overlooked because pH and ionic strength changes between samples if enough care is not taken. Although by far the simplest, this is not the most versatile technique, especially since not all binding interactions will result in significant changes in chemical shift.

On the other hand, studying the relaxation time of a compound can be more reliable, though more time-consuming. The relaxation time depends on several factors, but key here is the size and shape of the compound. Macromolecules such as proteins, or humic colloids have restricted molecular motion, leading to much shorter relaxation times compared to small molecules. If a humic colloid interacts with a small molecule, the molecule's motions are restricted and the relaxation time decreases. As such, by examining the relaxation time of the molecule under various humic concentrations (or vice versa), binding can be studied. Table 3, from the same study by Dixon et al., demonstrates the changing relaxation time (Dixon et al., 1999).

The change in relaxation time is much more dramatic than the chemical shift, and can be used to estimate quantitatively the proportion of bound vs free molecules and thus binding coefficients. Though easy to perform, this technique has a few disadvantages: it is time consuming and rarely gives information on where the binding occurs, which are factors that can be addressed by additional NMR techniques.

Diffusion-based techniques

Diffusion-based NMR, namely the 2-D technique known as diffusion ordered spectroscopy (DOSY) can also be used to study interactions. This technique differentiates molecules based both on the chemical shift and the diffusion coefficient. Thus, the basis

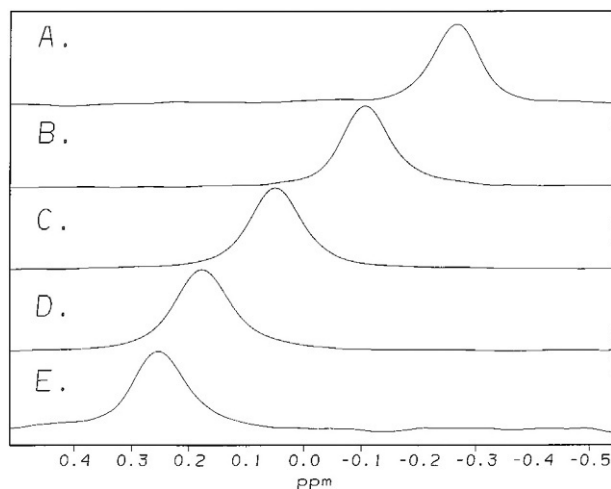


Fig. 7 The change in ^{19}F chemical shift of 4'-fluoro-1'-acetonaphthone as the concentration of fulvic acid increases. The fulvic acid concentrations are as follows: (A) 0 mg mL^{-1} , (B) 4.5 mg mL^{-1} , (C) 6.75 mg mL^{-1} , (D) 9 mg mL^{-1} , (E) 13.5 mg mL^{-1} . Reproduced with permission from Dixon AM, Mai MA, Larive CK. (1999). NMR investigation of the interactions between 4'-fluoro-1'-acetonaphthone and the Suwannee River fulvic acid. *Environmental Science and Technology* 33: 958–964. Copyright 1999 American Chemical Society.

Table 3 The T_1 and T_2 ^{19}F relaxation times (s) of 4'-fluoro-1'-acetonaphthone with increasing fulvic acid concentrations (mg mL^{-1}).

Fulvic acid concentration	T_1 relaxation time	T_2 relaxation time
0.0	1.781 ± 0.016	1.720 ± 0.013
4.5	0.555 ± 0.006	0.083 ± 0.006
6.8	0.396 ± 0.011	0.056 ± 0.002
9.0	0.469 ± 0.005	0.060 ± 0.004
13.5	0.288 ± 0.014	0.036 ± 0.000

Reproduced with permission from Dixon AM, Mai MA, Larive CK. (1999). NMR investigation of the interactions between 4'-fluoro-1'-acetonaphthone and the Suwannee River fulvic acid. *Environmental Science and Technology* 33: 958–964. Copyright 1999 American Chemical Society.

of the technique is similar to that of the relaxation measurements: small molecules have fast diffusion rates, and when bound to a macromolecule (or aggregate) the diffusion rates slow down significantly. A benefit of DOSY over relaxation measurements is that it is quite feasible to study multiple small molecules at the same time. This results in much faster throughput, and similar information regarding binding coefficients can be extracted, in addition to molecular weight information. As an example, Šmejkalová and Piccolo utilized DOSY to study the aggregation and disaggregation of soil humic substances, and a humic acid DOSY spectrum, as well as 1-D ^1H slices are shown in Fig. 8 (Šmejkalová and Piccolo, 2008). Here, the resonances near the top represent the more slowly diffusing (that is, larger) components. In this study, increasing humic concentrations were correlated with larger molecular sizes, supporting the fact that humic samples were aggregating into supramolecular associations. Though DOSY can provide higher throughput than relaxation measurements, it is slightly more difficult to set up, and provides rather little information on the binding orientation.

Nuclear overhauser effect-based techniques

Nuclear Overhauser effect spectroscopy (NOESY) has long been a gold standard for examining interactions in proteins, and can be useful in studying soils. NOESY and the closely related technique known as rotational nuclear Overhauser effect spectroscopy (ROESY) both rely on dipolar coupling interactions. What this means is that these techniques can look at interactions through physical space, as opposed to coupling through bonds employed by techniques such as HSQC. This is very powerful as it can help to elucidate the exact binding conformations of compounds with soil humic substances. NOESY and ROESY can be used in combination to distinguish between exchangeable protons (OH or NH groups) and other interactions. If a small compound binds to a large macromolecule, a cross peak will appear in the NOESY spectrum, which can identify precisely where binding occurs. Further, NOESY can help elucidate the 3-D structures of large molecular assemblies, by analyzing whether or not certain sites are in

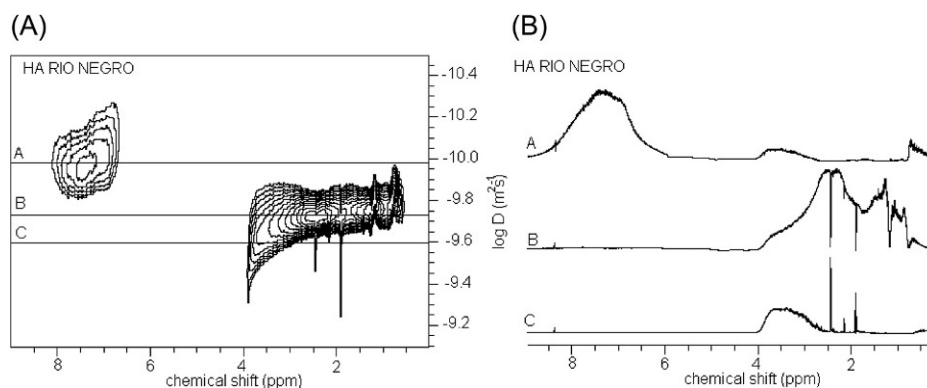


Fig. 8 (A) The ^1H DOSY spectrum of humic acid extracted from a Brazilian hydromorphic soil. (B) The 1D— ^1H spectral “slices” through the diffusion axis of the DOSY spectrum. Reproduced with permission from Šmejkalová D, Piccolo A. (2008). Aggregation and disaggregation of humic supramolecular assemblies by NMR diffusion ordered spectroscopy (DOSY-NMR). *Environmental Science and Technology* 42: 699–706. Copyright 2008 American Chemical Society.

close spatial proximity. Although the information gained from NOESY is quite powerful and the experiment is fairly easy to setup, NOESY can suffer from significant overlap and tends to have lower sensitivity than saturation transfer techniques, which are discussed next.

Saturation transfer difference (STD) techniques

The STD techniques are the most complex, (and arguably the most powerful) interaction studying techniques, and were specifically developed for this purpose. The STD technique operates by saturating (irradiating with radiofrequency energy) the peaks that are specific to a macromolecule (or set of macromolecules) and examining the changes in intensity of the NMR spectrum of the small molecule (ligand). If the ligand is sufficiently close to the macromolecule ($\sim 4\text{--}6\text{ \AA}$), it will acquire some of the saturation of the macromolecule, resulting in a decreased NMR signal intensity. The major benefit of this technique is that, since the STD effect is sensitive to the distance from the macromolecule, NMR signals from areas of the compound that do not bind are not attenuated. By gathering a series of spectra and subtracting them as seen in the Fig. 9, contributions from the nonbinding components are eliminated. This is known as saturation transfer double difference (STDD) and helps create flat baselines that allow changes from binding to be quantified accurately. The result is that STD NMR can be used to create an “epitope map”—a representation of the exact areas on the ligand that bind the macromolecule. An example of an epitope map in Fig. 10 (Shirzadi et al., 2008) shows the molecular level mechanistic information obtained through STD. Overall, STD NMR gives excellent mechanistic binding information, but though more sensitive than NOESY, may still require relatively long experiment times.

Reverse Heteronuclear Saturation Transfer Difference (RH-STD), utilizes similar principles to STD, but relies on saturating the ligand, as opposed to the macromolecule. This can help to determine which specific macromolecular components bind to the ligand. For example, a study of soil interactions with perfluorinated contaminants found that perfluorooctanoic acid had a great preference for the protein derived domains in a soil (Longstaffe et al., 2010). A significant drawback of RH-STD is that the ligand must be saturated without saturating the macromolecules. This typically requires a heteronucleus, the most common of which is ^{19}F , and is thus limited to samples that contain such a nucleus. Nonetheless, this is a powerful technique and since ^2H can be incorporated into nearly any organic molecule without changing its chemistry, it holds great potential for future interaction studies.

Studying interactions via CMP NMR

Though the previously discussed techniques are mainly used in the solution-state, CMP NMR can study interactions between phases, which is crucially important to the understanding of contaminant mobility and sequestration. An example of soil CMP analysis is shown in Fig. 11 (Masoom et al., 2015). Here, the molecular journeys of perfluorooctanoic acid (PFOA) and perfluorophenol (PFP) through the various phases of the soil were examined using ^{19}F CMP NMR. There are two things to note in the NMR spectra of this figure: the line shapes become significantly broader going from the solution to solid-state and PFP and PFOA have very different sequestration behaviors, with the majority of PFOA remaining in the solution-state, while most PFP enters the gel/solid-state. The right side of the figure shows a kinetic curve of the fate of the fluorinated contaminants, showing that NMR can be used to obtain time resolved interaction data. Overall, CMP NMR, as well as HR-MAS have excellent potential for the study of soil contaminant interactions.

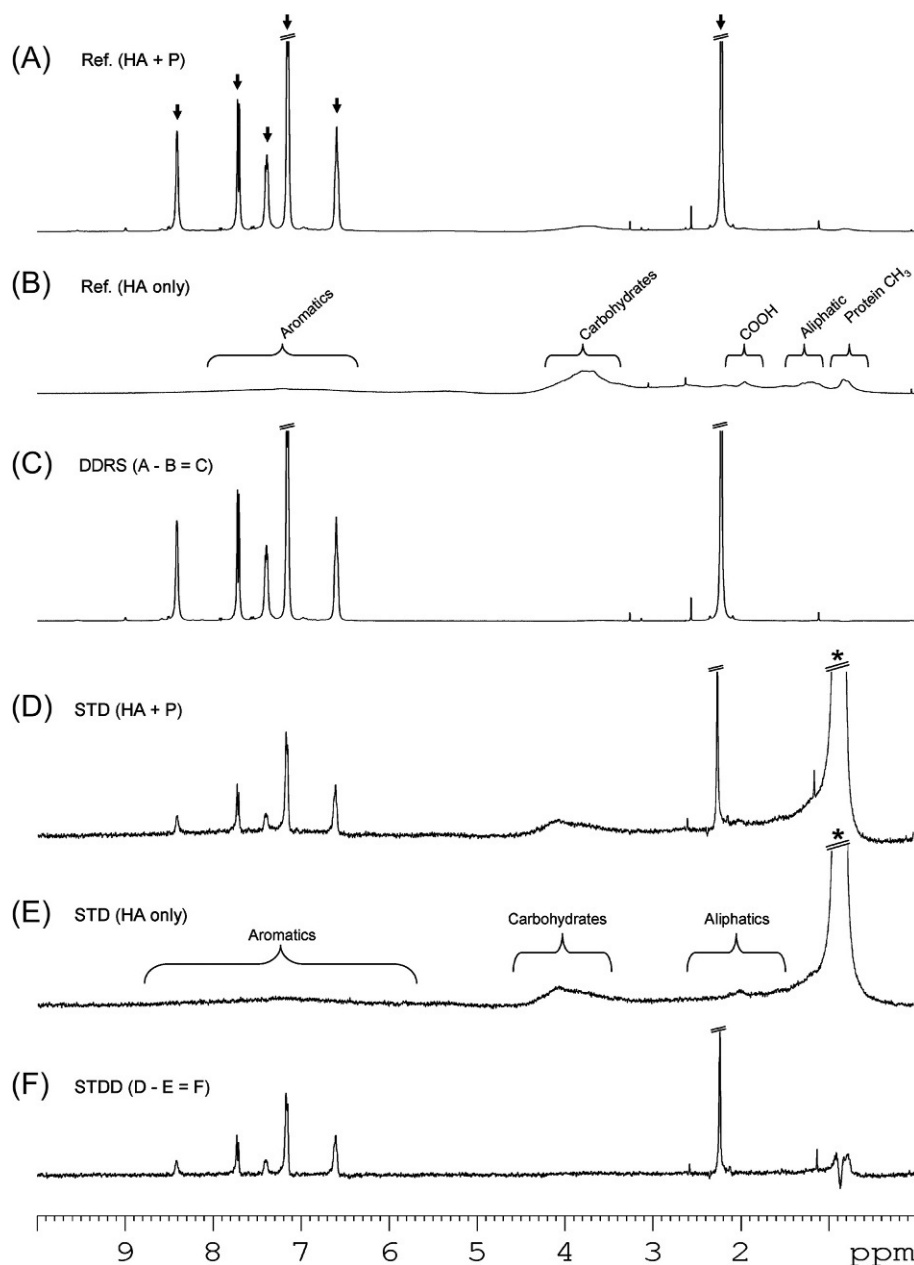


Fig. 9 An illustration of the process involved in generating an STD epitope map. First, two reference spectra (A) which has both the ligand and humic acids and (B) which contains only the humic acids are gathered. Spectrum (C) is the double difference reference spectrum and is generated by subtracting (B) from (A). Following this, the STD spectra (D) (humic acid and ligand) and (E) (only humic acid) are gathered and subtracted to make spectrum (F), the STDD spectrum. By comparing spectra (F) and (C), an epitope map can be created as shown in Fig. 10. Reproduced with permission from Shirzadi A, Simpson MJ, Xu Y, Simpson AJ. (2008). Application of saturation transfer double difference NMR to elucidate the mechanistic interactions of pesticides with humic acid. *Environmental Science and Technology* 42: 1084–1090. Copyright 2008 American Chemical Society.

Studying interactions via solid-state NMR

To close this section, we briefly consider the work by Sachleben et al. which is an example of studying interactions in the solid-state (Sachleben et al., 2004). Here, the interactions of a hydrophobic pollutant (pyrene), with plant derived cuticular matter (which contributes aliphatic compounds to soil organic matter) was examined through the use of both one-dimensional and multi-dimensional solid-state NMR. Based on analyses of the relaxation time and chemical shift anisotropy of pyrene, it was found that pyrene shows anisotropic motion under sorption to cutan, whereas under sorption to cutin the motion is initially isotropic but become anisotropic after 22 months. The study noted that this type of work provides an improved understanding of pollutant sorption to soil organic matter, which can help develop improved sorption models and soil remediation techniques.

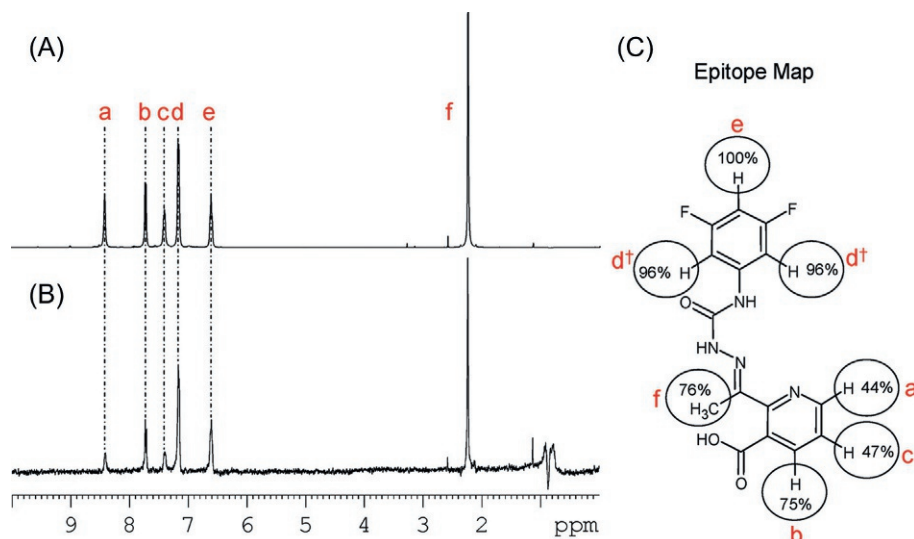


Fig. 10 A comparison of the (A) double difference reference and (B) STDD spectra of diflufenzopyr and humic acid (left). The resultant epitope map (C) shows the degree of interaction of the moieties of the molecule with humic acids, the largest of which is the proton labelled e. Reproduced with permission from Shirzadi A, Simpson MJ, Xu Y, Simpson AJ. (2008). Application of saturation transfer double difference NMR to elucidate the mechanistic interactions of pesticides with humic acid. *Environmental Science and Technology* 42: 1084–1090. Copyright 2008 American Chemical Society.

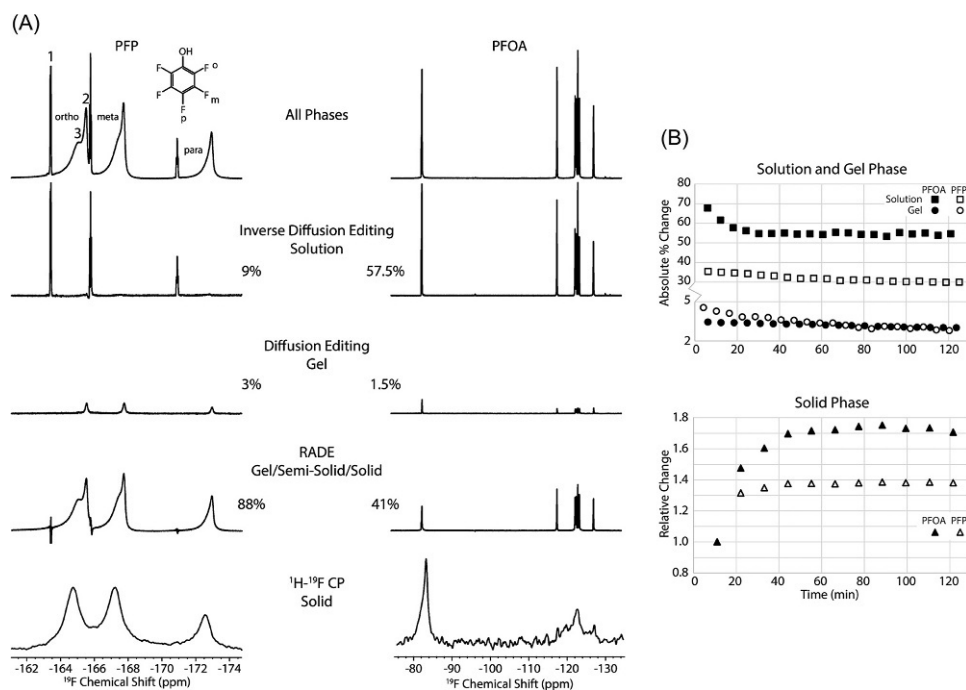


Fig. 11 (A) The ^{19}F NMR spectra detailing the transfer of fluorinated contaminants in a soil. (B) Kinetic transfer curves for the movement of fluorinated contaminants between phases. Adapted with permission from Masoom H, Courtier-Murias D, Soong R, et al. (2015). From spill to sequestration: The molecular journey of contamination via comprehensive multiphase NMR. *Environmental Science and Technology* 49: 13983–13991. Copyright 2015 American Chemical Society.

Sensitivity improvements in NMR

It is without doubt that the greatest weakness of NMR is its relatively low sensitivity, and therefore there has been substantial effort spent improving NMR sensitivity. As was alluded to in the introduction, the greater the difference in energy between the two spin states, the larger the net magnetization produced, which gives rise to a larger detectable signal. Two major factors that affect this energy difference are the strength of the magnetic field and the gyromagnetic ratio of the nucleus of study. The gyromagnetic ratio is fixed by the nucleus of study (see the section on choosing a nucleus for more information), while the strength of the magnetic field

depends on the NMR spectrometer available. Modern NMR spectrometers typically use very strong superconducting magnets that require cooling by liquid helium and nitrogen. The field strengths of modern high-field NMR spectrometers vary widely from <300 MHz to ~1.2 GHz, though spectrometers in the upper range are more rare; they are considerably more expensive both to purchase and in access fees. In the field of NMR, the field strength is commonly denoted in MHz (as opposed to the SI unit of Tesla), which is the resonant frequency of ^1H at that field strength. Higher field strength brings two major advantages: increased spectral resolution, which scales proportionally with the magnetic field, and increased SNR, which typically scales with the field to the power of $3/2$ (Webb, 2012). Overall, stronger magnets typically give much improved spectra and given just the improvements in NMR field strength, experiments that were not possible on spectrometers from the 1990s are now much more feasible.

Another commonly touted example of sensitivity improvement in NMR are cryoprobes. In this case, by cryogenically cooling the electronic components involved in the NMR circuitry, the amount of thermal noise (the main source of noise in NMR) can be reduced, resulting in up to 3–4-fold SNR improvements (Styles et al., 1984). However, cryoprobes are typically not as effective for samples with high salt contents, and thus the benefit for soil analysis is quite variable and unlikely to reach 3–4-fold improvement (Simpson et al., 2011). That said, probes with varying detection coil architectures can still provide sensitivity (and other) improvements. For example, an inverse probe, in which the proton coil is closer to the sample, will result in better ^1H spectra, while a broadband probe (optimized for the heteronucleus) will give better ^{13}C spectra by a factor of ~2 (Simpson et al., 2011).

Conclusion

Nuclear Magnetic Resonance is a powerful analytical tool for the study of soil constituents and interactions. Despite the large offering of NMR experiments, many of which can yield excellent information and insights, the vast majority of soil NMR is currently restricted to solid-state ^{13}C NMR, in large part due to the historical use of solid-state NMR in early soil research in the 1960s–1990s. However, it is critical for researchers to explore all branches of NMR and the authors intend this article to be a small step towards making the various types of NMR more accessible in soil research. For example, saturation transfer difference techniques have great potential to help understand the critically important processes of aggregation, sequestration and other non-covalent interactions that are extremely difficult to study by MS, but only a few studies involving these techniques have been published. The HR-MAS and CMP NMR techniques open the door to the analysis of soils in their natural state, giving information about multi-phase processes that occur in the real world. With there being an increasing shift from questions such as “What are humic substances?” to “How does the soil aggregate structure form” or “Why does herbicide X bind more than herbicide Y to a soil,” it will become more and more important to study soils in their natural, hydrated state. Thus, preparations that involve drying to generate “pure” solids or extractions to generate “humic” fractions are no longer the only ways forward for soil research.

MRI is still the most effective approach to look at human anatomy and NMR is a close cousin with the ability to look inside intact soils with bond-level specificity. In many scenarios, especially those concerning novel structures and non-covalent interactions, NMR is unique in its abilities among other modern instrumentation. As such, it holds the key to answer some of the most challenging and complex questions facing the field. It is equally important that the next generation of soil researchers fully embrace cutting edge NMR technologies, and that current and future NMR experts (and manufacturers) strive to make NMR research more accessible to those from “non-traditional” disciplines such as soil and environmental science. In summary, NMR has a critical and central role to play in the arsenal of analytical technologies required to understand the role of anthropogenic impacts on the environment and their subsequent effects on ecosystems and human health.

References

- Bell NGA, Michalchuk AAL, Blackburn JWT, et al. (2015) Isotope-filtered 4D-NMR spectroscopy for structure determination of humic substances. *Angewandte Chemie, International Edition* 54: 8382–8385.
- Bharti SK and Roy R (2012) Quantitative ^1H NMR spectroscopy. *Trends in Analytical Chemistry* 35: 5–26.
- Cardozo LA, Korir AK, Otto WH, et al. (2004) Applications of NMR spectroscopy in environmental science. *Progress in Nuclear Magnetic Resonance Spectroscopy* 45: 209–238.
- Conte P and Meo PL (2020) Nuclear magnetic resonance with fast field-cycling setup: A valid tool for soil quality investigation. *Agronomy* 10: 1–33.
- Dixon AM, Mai MA, and Larive CK (1999) NMR investigation of the interactions between 4'-fluoro-1'-acetonephthone and the Suwannee River fulvic acid. *Environmental Science and Technology* 33: 958–964.
- Farooq H, Courtier-Murias D, Soong R, et al. (2013) HR-MAS NMR spectroscopy: A practical guide for natural samples. *Current Organic Chemistry* 17: 3013–3031.
- Harris RK, Becker ED, Cabral de Menezes SM, et al. (2001) NMR nomenclature. Nuclear spin properties and conventions for chemical shifts. *Pure and Applied Chemistry* 73: 1795–1818.
- Hertkorn N, Ruecker C, Meringer M, et al. (2007) High-precision frequency measurements: Indispensable tools at the core of the molecular-level analysis of complex systems. *Analytical and Bioanalytical Chemistry* 389: 1311–1327.
- Kögel-Knabner I (1997) ^{13}C and ^{15}N NMR spectroscopy as a tool in soil organic matter studies. *Geoderma* 80: 243–270.
- Lam B and Simpson AJ (2008) Direct ^1H NMR spectroscopy of dissolved organic matter in natural waters. *Analyst* 133: 263–269.
- Longstaffe JG, Simpson MJ, Maas W, and Simpson AJ (2010) Identifying components in dissolved humic acid that bind organofluorine contaminants using $^1\text{H}\{^{19}\text{F}\}$ reverse heteronuclear saturation transfer difference NMR spectroscopy. *Environmental Science and Technology* 44: 5476–5482.
- Mao J, Cao X, Oik DC, et al. (2017) Advanced solid-state NMR spectroscopy of natural organic matter. *Progress in Nuclear Magnetic Resonance Spectroscopy* 100: 17–51.
- Masoom H, Courtier-Murias D, Soong R, et al. (2015) From spill to sequestration: The molecular journey of contamination via comprehensive multiphase NMR. *Environmental Science and Technology* 49: 13983–13991.

- Masoom H, Courtier-Murias D, Farooq H, et al. (2016) Soil organic matter in its native state: Unravelling the most complex biomaterial on earth. *Environmental Science and Technology* 50: 1670–1680.
- Pohlmeier A, Vergeldt F, Gerkema E, et al. (2010) MRI in soils: Determination of water content changes due to root water uptake by means of a multi-slice-multi-echo sequence (MSME). *The Open Magnetic Resonance Journal* 3: 69–74.
- Preston CM (2001) Carbon-13 solid-state NMR of soil organic matter - using the technique effectively. *Canadian Journal of Soil Science* 81: 255–270.
- Sachleben JR, Chefetz B, Deshmukh A, and Hatcher PG (2004) Solid-state NMR characterization of pyrene-cuticular matter interactions. *Environmental Science and Technology* 38: 4369–4376.
- Shirzadi A, Simpson MJ, Xu Y, and Simpson AJ (2008) Application of saturation transfer double difference NMR to elucidate the mechanistic interactions of pesticides with humic acid. *Environmental Science and Technology* 42: 1084–1090.
- Simpson AJ, Tseng LH, Simpson MJ, et al. (2004) The application of LC-NMR and LC-SPE-NMR to compositional studies of natural organic matter. *Analyst* 129: 1216–1222.
- Simpson AJ, Kingery WL, and Hatcher PG (2003) The identification of plant derived structures in humic materials using three-dimensional NMR spectroscopy. *Environmental Science and Technology* 37: 337–342.
- Simpson AJ, McNally DJ, and Simpson MJ (2011) NMR spectroscopy in environmental research: From molecular interactions to global processes. *Progress in Nuclear Magnetic Resonance Spectroscopy* 58: 97–175.
- Simpson AJ, Simpson MJ, and Soong R (2012) Nuclear magnetic resonance spectroscopy and its key role in environmental research. *Environmental Science and Technology* 46: 11488–11496.
- Simpson AJ, Simpson MJ, and Soong R (2018) Environmental nuclear magnetic resonance spectroscopy: An overview and a primer. *Analytical Chemistry* 90: 628–639.
- Simpson MJ and Simpson AJ (eds.) (2014) *NMR Spectroscopy: A Versatile Tool for Environmental Research*. Chichester: John Wiley and Sons Publishing.
- Šmejkalová D and Piccolo A (2008) Aggregation and disaggregation of humic supramolecular assemblies by NMR diffusion ordered spectroscopy (DOSY-NMR). *Environmental Science and Technology* 42: 699–706.
- Soucémariadin LN, Erhagen B, Nilsson MB, et al. (2017) Two dimensional NMR spectroscopy for molecular characterization of soil organic matter: Application to boreal soils and litter. *Organic Geochemistry* 113: 184–195.
- Styles P, Soffe NF, Scott CA, et al. (1984) A high-resolution NMR probe in which the coil and preamplifier are cooled with liquid helium. *Journal of Magnetic Resonance* 60: 397–404.
- Vestergren J, Vincent AG, Jansson M, et al. (2012) High-resolution characterization of organic phosphorus in soil extracts using 2D ^1H - ^{31}P NMR correlation spectroscopy. *Environmental Science and Technology* 46: 3950–3956.
- Webb A (2012) Increasing the sensitivity of magnetic resonance spectroscopy and imaging. *Analytical Chemistry* 84: 9–16.

Neutron-gamma analysis of soil elements

Aleksandr Kavetskiy^a, Galina Yakubova^a, Nikolay Sargsyan^b, Stephen A Prior^a, and H Allen Torbert^a, ^aUSDA-Agricultural Research Service, National Soil Dynamics Laboratory, Auburn, AL, United States; ^bDepartment of Computer Science and Software Engineering, Auburn University, Auburn, AL, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	626
General principles of neutron-gamma analysis	627
Modifications of neutron-gamma analysis	627
Pulsed fast/thermal neutron analysis (PFTNA)	627
Tagged neutron method (or associated particle imaging, API, method)	628
Measurement system components and their configuration for field application	628
Neutron sources	628
Gamma detectors	628
Shielding and construction materials	629
Operational electronics and data acquisition software	629
The PFTNA field mobile system configuration	629
Results of PFTNA measurements and soil elemental content	630
Raw data	630
System background spectra	631
Peak area determinations	631
Calibration dependencies	631
Carbon calibration dependency	634
Determination of soil carbon content (in weight %)	634
Moisture (hydrogen) calibration dependency	635
Comparison of PFTNA and chemical analysis (dry combustion) data	636
Methodology of field PFTNA measurements	637
Soil volume associated with PFTNA measurements	637
The PFTNA measurement time	637
Acquiring data for elemental mapping	637
Determination of total field soil carbon content for 10- and 30-cm layers	638
Soil elemental field distribution maps	638
Soil carbon content distribution map	638
Map of moisture distribution	640
Conclusion	641
References	641

Abstract

This chapter is devoted to the application of neutron-gamma methods, mainly pulsed fast thermal neutron analysis (PFTNA), for soil elemental measurement and mapping. Herein, we provide a description of general PFTNA principles, main components and configurations of a PFTNA field system, and procedures for primary data (gamma spectra) acquisition for determining soil elemental content. In addition, methodology of field PFTNA measurements and the creation of soil elemental maps will be discussed. Soil elemental maps plotted from PFTNA measurements are useful in soil science and for assessing modern agricultural practices. Results of PFTNA can provide soil carbon content in the 30-cm depth soil layer and such information can be useful for carbon credit estimates.

Key points

- Description of general principles of neutron-gamma analysis, modifications of neutron-gamma analysis, measurement system components and their configuration for field applications are provided.
- Assessing soil elemental content from pulsed fast/thermal neutron analysis results, soil element datasets required for mapping and mapping procedures are discussed.
- Examples of agricultural field elemental distribution maps and data on total field soil carbon content suitable for carbon credit estimates relevant to the climate change debate are presented.

Introduction

Soil science plays an importance role in optimizing food production including the understanding of environmental processes such as soil carbon sequestration and soil health. State-of-the-art soil science and agricultural practices for increasing soil productivity and profitability require precise knowledge of the chemical composition of large volumes (tens, hundreds, thousands of cubic meters) of soil, composts, and fertilizers. Traditional chemical analysis is labor intensive and time consuming since they require a large number of samples to be collected for laboratory processing. To date, various physical methods for soil element determination (mainly carbon, C) have been developed, which include *in situ* determinations of some soil elements. Among these, different modifications of neutron-gamma analysis [Wielopolski \(2011\)](#) are a more promising alternative to traditional chemical analysis procedures. The neutron-stimulated gamma analysis method can be used for *in situ* measurements of primary elements in agricultural soils (e.g., C, Si, O, H, Fe, and K, see [Table 1](#) for element names), C:N ratio of composts ([Yakubova et al., 2019](#)), and

Table 1 Primary gamma ray energies of main soil element nuclei appearing upon neutron irradiation.

Element	Nucleus	Neutron Type	Neutron induced process			
			Type	Reaction	Cross-section, mb ^a	Gamma ray energy, MeV ^a
Hydrogen	¹ H	Thermal	TNC	¹ H(n _{th}) ² H* → ² H + γ	332.7	2.223
Carbon	¹² C	Fast (14 MeV)	INS	¹² C(n,n') ¹² C* → ¹² C + γ	210	4.439
Nitrogen	¹⁴ N	Fast (14 MeV)	INS	¹⁴ N(n,n') ¹⁴ N* → ¹⁴ N + γ	22.8	1.635
					56	2.313
					44	5.106
				¹⁴ N(n,t) ¹² C* → ¹² C + γ	22.1	4.439
				¹⁴ N(n,α) ¹¹ B* → ¹¹ B + γ	19.1	4.444
				¹⁴ N(n _{th}) ¹⁵ N* → ¹⁵ N + γ	79.8	5.269
Oxygen	¹⁶ O	Fast (14 MeV)	INS			10.829
					145	6.129
					43.3	2.742
					31.7	6.917
					62.5	7.117
				¹⁶ O(n, n'α) ¹² C* → ¹² C + γ	16.6	4.439
				¹⁶ O(n,α) ¹³ C* → ¹³ C + γ	74.5	3.684
				¹⁶ O(n,α) ¹³ C* → ¹³ C + γ	42	3.854
				¹⁶ O(n,α) ¹³ C* → ¹³ C + γ	22	3.089
				¹⁶ O(n,p) ¹⁶ N(β ⁻ 7.1 s) → ¹⁶ O* → ¹⁶ O + γ	42	6.129
Silicon	²⁸ Si	Fast (14 MeV)	DA	²⁸ Si(n,n') ²⁸ Si* → ²⁸ Si + γ	120	1.778
			DA	²⁸ Si(n,p) ²⁸ Al(β ⁻ 2.24 m) → ²⁸ Si* → ²⁸ Si + γ	280	1.778
			TNC	²⁸ Si(n _{th}) ²⁹ Si* → ²⁹ Si + γ	177	3.539
		Thermal	TNC			4.934
						0.847
Iron	⁵⁶ Fe	Fast (14 MeV)	INS	⁵⁶ Fe(n,n') ⁵⁶ Fe* → ⁵⁶ Fe + γ	74	1.238
						0.847
						1.811
						2.113
						7.631
Aluminum	²⁷ Al	Thermal	TNC	⁵⁶ Fe(n _{th}) ⁵⁷ Fe* → ⁵⁶ Fe + γ	2490	9.298
		Thermal	TNC	⁵⁴ Fe(n _{th}) ⁵⁵ Fe* → ⁵⁵ Fe + γ	2440	0.844
		Fast (14 MeV)	INS	²⁷ Al(n,n') ²⁷ Al* → ²⁷ Al + γ	9 (2 MeV—100)	1.014
					17 (2 MeV—230)	2.211
					32 (3 MeV—240)	3.004
					23.5 (4.5 MeV—180)	1.810
				²⁷ Al(n,d) ²⁶ Mg* → ²⁶ Mg + γ	15.6	0.844
				²⁷ Al(n,p) ²⁷ Mg(β ⁻ 9.46 m) → ²⁷ Al* → ²⁷ Al + γ	73	1.014
						1.367
				²⁷ Al(n,α) ²⁴ Na(β ⁻ 15 h) → ²⁴ Mg* → ²⁴ Mg + γ	122	2.754
						4.122
		Thermal	DA	²⁷ Al(n _{th}) ²⁸ Al(β ⁻ 2.24 m) → ²⁸ Si* → ²⁸ Si + γ	232	1.778

*, Indicates excited state that relaxes issued gamma ray.

^a(IAEA, 2014; NNDC, 2013).

Modified from: Yakubova G, Kavetskiy A, Prior SA and Torbert HA (2019) Application of neutron-gamma analysis for determining compost C/N ratio. *Compost Science & Utilization* **27**: 146–160.

Cl (chlorine) contamination of soils. This method is a non-destructive analysis that requires no sample preparation and can perform multi-elemental analyses of large volumes that are negligibly affected by local sharp changes in elemental content.

Neutron-gamma analysis is based on detecting gamma lines that appear due to neutron nuclei interactions. Many nuclei can be detected and quantified by the presence of these characteristic gamma lines. State-of-the-art nuclear physics methodologies and instrumentation combined with commercial availability of portable pulse neutron generators, high-efficiency gamma detectors, reliable electronics, and measurement and data processing software have currently made the application of neutron-gamma analysis possible for routine measurements in various fields of study. These are the detection of hazardous materials (explosives, drugs, and dangerous chemicals), diamond detection, planetary science applications for obtaining information on bulk elemental composition, archaeological site surveying and provenance studies, elemental composition of human and animal bodies, real-time elemental analysis of bulk coal on conveyor belts, chloride content of reinforced concrete, and in oil well logging [Valkovic \(2016\)](#). Neutron-stimulated gamma ray analysis can also be used in soil science for in situ measurements of soil carbon and other soil elements. Soil carbon measurement is based on detecting 4.44 MeV gamma rays issued from carbon nuclei excited by fast neutrons promptly after the interaction ([Wielopolski, 2011](#)). Note that accurate quantification of soil carbon is important since it is an indicator of soil quality that can affect soil carbon sequestration, fertility, erosion, and greenhouse gas fluxes ([Smith et al., 2012](#)).

The general principles of neutron-gamma analysis, modifications of neutron-gamma analysis applicable to soil science, measurement system components and its configuration for field application, neutron-gamma methodology for determination of soil elements, and examples of an application of this method in agricultural settings will be described in this chapter.

General principles of neutron-gamma analysis

Neutron-gamma analysis is based on the registration of gamma rays issued by nuclei upon interaction with neutrons due to different neutron-nuclei interaction processes. Fast neutrons can penetrate a surveyed volume of soil, impact and excite nuclei, and scatter. Neutrons lose some energy (moderating process) due to this inelastic scattering while excited nuclei relax into a ground state issuing gamma rays. For example, the inelastic scattering of fast neutrons on ^{12}C nuclei elevate them to the 4.44 MeV excited energy state ([NNDC, 2013](#)). Excited $^{12}\text{C}^*$ promptly returns to the ground state issuing the 4.44 MeV gamma ray. Finally, the neutron is moderated until thermal energy can be captured by some nuclei. Nuclei issue gamma rays promptly after interaction due to both inelastic neutron scattering (INS) and thermal neutron capture (TNC). In addition, in some cases the interaction of neutrons (both fast and thermal) can lead to the creation of new radioactive isotopes due to nuclear reaction type of (n (neutron), p (proton)), (n, α), (n, t (tritium)), etc. Decay of these isotopes is accompanied by delay activation (DA) gamma rays. Each kind of nucleus and process produces gamma-rays of particular energy. This characteristic gamma line of particular energy can serve as an analytical line for elemental determinations. The list of main soil elements, their characteristic gamma lines, particular energy and processes responsible for line appearance are shown in Table 1. As can be seen, many gamma ray energy lines of interest are in the 1–12 MeV range. The cross-sections of neutron (both fast, 14 MeV, and thermal) interactions with nuclei are shown in [Table 1](#).

The net peak area of an element of interest can be defined from measuring the gamma spectral response. However, for determining the content (or concentration) for the element of interest in soil, the calibration dependence that associates the content of this element with a gamma response peak value is required. This calibration dependence can be created from measurements of gamma spectra of neutron irradiation of reference soils with well-known elemental contents. The elemental content of reference soil can be determined by chemical analysis, or these soils can be prepared by mixing known amounts of substances (e.g., sand and carbon grains).

Modifications of neutron-gamma analysis

Pulsed fast/thermal neutron analysis (PFTNA)

The main difficulty in conducting soil neutron-gamma analysis and determining net peak values of interest (e.g., soil carbon peak) is the overlapping of different gamma lines from nuclei of soil and construction components of the measurement system. Thus, it can be problematic to extract net peaks of interest from the gamma response spectra due to soil irradiation. However, it is possible to separate the gamma ray spectrum due to INS reactions from the total complicated gamma ray spectrum caused by all nuclear reactions using a pulsed neutron generator and certain measurement techniques. By these means, 14.1 MeV neutron flux pulses (with 5–20 μs width and 5–20 kHz frequency) can be used to irradiate studied objects. The INS reactions will only occur during the neutron pulse (5–20 μs range), while TNC and DA processes occur during and between neutron pulses due to relationships between time of neutron moderation, lifetime of thermal neutrons, and lifetime of excited nuclei. The data acquisition system can collect data during neutron pulses at one memory address and switch to another memory address between pulses. The average intensity of TNC spectra under these measurement conditions have practically the same value between pulses and during pulses ([Kavetskiy et al., 2017](#)). Thus, the INS spectrum can be restored by extracting the (TNC + DA) spectrum count rate (measured between pulses) from the (INS + TNC + DA) spectrum measured during pulses.

The measured spectra consist of two parts: spectra of the studied object and spectra of the measurement system (system background spectra). The last spectra can be determined by conducting PFTNA measurements by moving the measurement system away from the ground, floor, walls of buildings, and other large nearby objects. This can be accomplished by lifting the system

above the ground. The (INS + TNC + DA) and (TNC + DA) gamma spectra do not change at heights greater than 3–4 m, and readings in this range can be equated to measurement of the system background. With this background information, the net (INS + TNC + DA) and (TNC + DA) spectra of the object studied can be determined by subtracting the system background spectra from the corresponding measured spectra of the object (e.g., soil). The net spectra will represent the spectra of gamma rays for only the soil (sample) due to INS or TNC + DA processes. After this determination, the net INS spectrum of soil can be restored by extracting the (TNC + DA) soil spectrum from the (INS + TNC + DA) soil spectrum.

Peak values from these spectra can be attributed to nuclei present only in the soil and can be used to calculate elemental content. Carbon, silicon, oxygen, and aluminum concentrations in soil can be determined from the net INS spectrum, while the hydrogen, iron, and silicon can be determined from the net (TNC + DA) spectrum.

Following the classification of neutron-gamma analysis methods given in (Barzilov et al., 2012), the method of neutron-gamma analysis described can be called Pulsed Fast/Thermal Neutron Analysis (PFTNA). This method is currently used for soil element mapping of agricultural fields and has potential use for determinations of total field soil carbon for carbon credit markets.

Tagged neutron method (or associated particle imaging, API, method)

There is another possible means of using neutron-gamma technology for soil elemental analysis. Neutron generators that produce neutrons due to (d (deuterium), t) nuclear reactions are widely used in other gamma-neutron measurement systems (see further). At the (d, t) reaction, both neutrons and alpha particles are generated at the same time, but move in opposite directions. The alpha detector and sample are installed on opposite sides of the neutron generator target. The alpha particle pulse (in the alpha detector) can be used to 'tag' the neutron, and only the gamma rays stimulated by this 'tagged' neutron will be registered. The tagged neutron method is based on measurements of gamma spectra in the alpha-gamma coincidence mode. Several nanoseconds are required for neutrons to reach the sample before gamma rays are produced and registered. The geometry of the alpha-detector setup defines the measurement cone, and only gamma rays (due to INS processes only) from neutron-sample interactions (in this cone) that coincide with alpha particles are acquired.

The gamma spectra measured by the tagged neutron method has two main advantages. First, it produces practically clean INS spectra that are suitable for INS active soil element (C, Si, O) measurements. Second, this gamma spectrum relates primarily to the object studied.

Developing the tagged neutron method for soil elemental analysis is now in progress. Laboratory experiments demonstrated (Kavetskiy et al., 2019) that the carbon minimal detectible level (MDL) in sand-carbon mixtures for the tagged neutron method is better than other neutron gamma methods, in particular PFTNA. Thus, the API method of soil elemental analysis appears to be a very promising direction for the development of a field mobile system with improved metrological parameters.

Measurement system components and their configuration for field application

The configuration of a measurement system for neutron-gamma analysis consists of a neutron source, gamma detector, shielding and construction materials, operational electronics, and data acquisition software. Below we briefly consider the main features of these components.

Neutron sources

Portable neutron generators are primarily used in neutron-gamma analysis systems for field applications. From a radiation safety point of view, use of a neutron generator is preferred because no radiation is produced when the generator is turned 'off'.

Neutrons are produced in the portable neutron generator by (d, t) or (d, d) nuclear reactions when the target is hit (saturated by tritium or deuterium) by accelerated deuterons. Generators can work in continuous and pulsed modes. Generators for applications of the tagged neutron (or API) system have a built-in alpha detector.

Gamma detectors

Among the different types of gamma detectors, scintillation detectors with inorganic scintillators of large volumes (more than 1 dm³, preferably 5–10 dm³) are more suitable for neutron-gamma analysis systems as noted by Valkovic (2016). They have a high Z value and effectively detect a relatively low flux of characteristic gamma rays with energies up to ~12 MeV and provide an energy resolution that enables peaks of interest to be resolved. The interaction of neutrons with nuclei in the material of these detectors does not produce gamma rays that overlap useful signals emitted from samples. Detector materials are void of isotopes that are anticipated in samples analyzed and neutron activation of isotopes inside the detector volume is minimal. Shifts in detector sensitivity due to changing environmental conditions (temperature, etc.) is minimal. Thus, these detectors are suited for field applications of neutron-gamma analysis. Detectors based on sodium iodide NaI(Tl), bismuth germanate Bi₄Ge₃O₁₂ (BGO), and lanthanum bromide LaBr₃(Ce) inorganic scintillators are commonly used in neutron-gamma analysis systems.

Shielding and construction materials

Direct fast neutron flux on the gamma detector and gamma radiation that appear from neutron interactions with detector nuclei produce a large gamma spectral background signal. Large background signals affect minimal detection limits of the measurement system, thereby increasing measurement errors. Shielding used between the neutron generator and gamma detector can help to eliminate this problem in the measurement system. Fast neutron shielding consists of two or three components that slow down fast neutrons to thermal energy (moderator), absorbs thermal neutrons (absorber), and then attenuates the gamma rays produced by different neutrons-nuclei interactions in the moderator and absorber (attenuator). Light materials like water, heavy water, and polyethylene are usually used as neutron moderators, and boric acid is a possible absorber. Lead, tungsten, and iron are usually used for attenuation of gamma radiation. The shielding should be of reasonable thickness and should not produce gamma lines within the energy range of interest.

Construction materials should have minimal susceptibility to neutron activation by fast or thermal neutrons, issue few gamma rays in the energy range of interest, and have minimum high energy gamma rays that increase the system background spectrum. Duralumin can be recommended as construction material for neutron-gamma analysis systems.

Operational electronics and data acquisition software

Operational electronics and data acquisition software used in neutron stimulated gamma analysis systems can be built on the basis of standard electronic blocks paired with a pulsed neutron generator or can be custom-made electronic schemes and data acquisition software. Electronics of the PFTNA system should be capable of acquiring gamma spectra during and between neutron generator pulses separately (split electronics). Electronic boards and ProSpect-Vega software from XIA LLC (Hayward, CA) demonstrated reliable workability in PFTNA systems (Yakubova et al., 2014). The data acquisition system for API equipment should work in the alpha-gamma coincidence mode. This can be built using standard electronic blocks produced by Ortec or Canberra companies.

The PFTNA field mobile system configuration

The field mobile system should be mounted on a mobile platform and as previously mentioned, consists of a neutron generator, gamma detector, shielding, operational electronics, and data acquisition software. This system can be towed by tractors or all-terrain vehicles over various field terrains. This system should have an autonomous power system that provides an uninterrupted power supply for at least one working day.

The following is one example of the system used to conduct research at the USDA-ARS National Soil Dynamic Laboratory in Auburn, AL. This system works in the PFTNA mode and has three separate construction blocks (Fig. 1). Components of the first block are an MP320 pulsed neutron generator (Thermo Fisher Scientific, Colorado Springs, CO), an R2D-410 neutron detector (Bridgeport Instruments, LLC, Austin, TX), and a power system (Fig. 1A). The neutron generator has a neutron pulsed output of

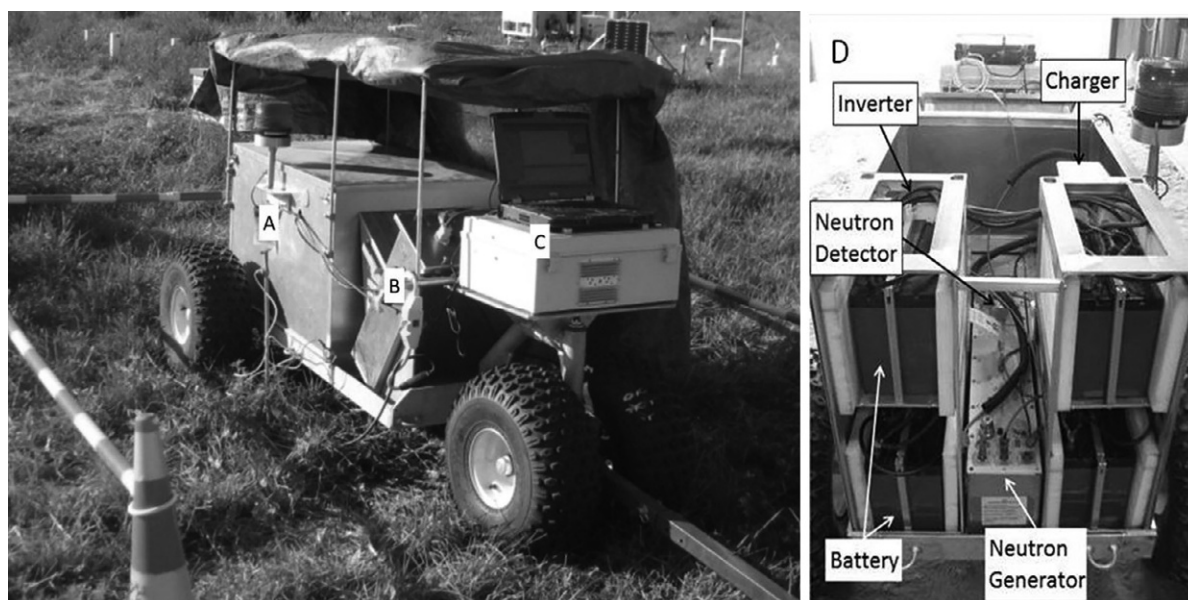


Fig. 1 Overview of the PFTNA system: (A) neutron generator, neutron detector, and power system; (B) three NaI (Tl) detectors; (C) equipment operation; (D) general view of a showing individual components. Modified from: Kavetskiy A, Yakubova G, Torbert HA and Prior SA (2017) Neutron-stimulated gamma ray analysis of soil. In: Maghraby AM (ed.) *New Insights on Gamma Rays*, pp. 133–178. INTECH: <https://www.intechopen.com/books>.

10^7 – 10^8 n s⁻¹ (depending on parameter settings) and a neutron energy of 14.1 MeV. Components of the autonomous power system are four DC105–12 batteries (12 V, 105 Ah) and a DC-AC inverter (CGL 600 W series; Nova Electric, Bergenfield, NJ). This inverter transformed 12 VDC battery power to 110 VAC. A Quad Pro Charger model PS4 (PRO Charging Systems, LLC, LaVergne, TN) is used for charging batteries during nighttime garage storage of the system. The first block also contains water, iron, and boric acid shielding for isolating the detectors from the neutron beam and focusing the beam on the soil area of interest. The second block has the gamma ray measuring equipment (Fig. 1B) and contain three 12.7 cm × 12.7 cm × 15.2 cm scintillation NaI(Tl) detectors (Scionix USA, Orlando, FL). A Cockcroft–Walton type high voltage power supply (C1248 series Hamamatsu Photonics K.K.) ensures high output linearity of the gamma detectors photomultiplier tubes (PMT) while maintaining low power consumption by detectors. Electronic boards (XIA LLC, Hayward, CA) connected to the PMT included fixed analog gain, 14-bit resolution and 100 MHz sampling rate analog-to-digital converter, and a 2048-channel multichannel analyzer (for each detector) to acquire and transfer gamma spectra to the operational computer. A GPS system (Trimble R2 receiver, Trimble Inc.) with an accuracy of ±0.5 m connected to the operational computer by Bluetooth technology identifies and records the geographical position of the acquired gamma spectra. For equipment operation, the third block (Fig. 1C) housed a laptop computer for controlling the neutron generator, detectors, and data acquisition system (ProSpect 0.1; XIA LLC) (Fig. 1C).

The dimensions of this mobile system are 75 cm × 23 cm × 95 cm and weighs ~300 kg. While the frame's primary construction material is aluminum, the shielding (mainly iron) contributes more weight. This system can be towed over the field at speeds up to 10 km h⁻¹ during measurements in the scanning mode. An Android tablet (Samsung Galaxy Tab S5e) utilizing Google Maps [Google \(2019\)](#) and specially developed software is connected by Bluetooth technology to the GPS. The tablet is in the towing vehicle (within the driver's view) to provide operational information. Information provided on the tablet includes the current system position, traversed path, total spectra acquisition time, etc., and the driver utilizes this information for selecting pathways for scanning field measurements.

Results of PFTNA measurements and soil elemental content

Raw data

Two gamma spectra are acquired from PFTNA measurements: (INS + TNC + DA) spectra acquired during the neutron pulse and (TNC + DA) spectra acquired between neutron pulses. The PFTNA system being towed across the field is shown in Fig. 2. An example of raw (INS + TNC + DA) and (TNC + DA) gamma spectra of soil measured by the PFTNA system described above is shown in Fig. 3.

The net (INS + TNC + DA) and (TNC + DA) spectra of soil can be found by subtracting the system background spectra from the corresponding measured spectra. The net INS spectrum can be restored by extracting the net (TNC + DA) spectrum in count rate unit from the (INS + TNC + DA) spectrum of soil. Net INS spectral peaks with centroids of 1.78 MeV and 6.13 MeV can be directly attributed to silicon-28 (²⁸Si) and oxygen-16 (¹⁶O) nuclei in soil, respectively; the peak with a centroid of 4.44 MeV is complex, but peak area attributed to the carbon-12 (¹²C) nuclei can be determined from this peak. Peaks with centroids at 2.22 MeV and 7.63 MeV in the net (TNC + DA) spectrum of soil can be attributed to hydrogen and iron-56 (⁵⁶Fe) nuclei, respectively.



Fig. 2 The PFTNA system towed over the field.

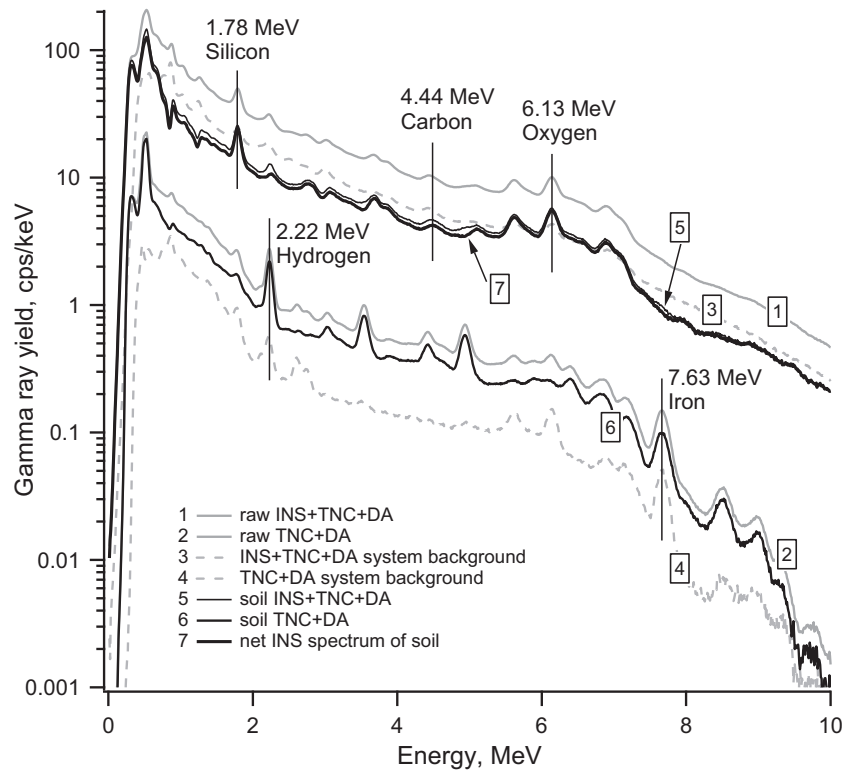


Fig. 3 Raw (INS + TNC + DA) and (TNC + DA) gamma spectra of soil, system background spectra, and soil net INS and (TNC + DA) spectra.

System background spectra

The system background spectra are acquired when the PFTNA system is raised above the ground (and away from buildings and large objects) by use of a crane (Fig. 4). Measured (INS + TNC + DA) spectra and (TNC + DA) spectra of the PFTNA system at different heights above the ground are shown in Fig. 5.

Relations between different peak areas with height are shown in Fig. 6. The data represented demonstrate that spectra measured at heights greater than 3–4 m can be taken as the system background spectra (see Fig. 3) and can be used in calculations of net INS and (TNC + DA) spectra of soil.

Peak area determinations

After determining the net INS and net (TNC + DA) spectra of soil, areas of corresponding soil element peaks can be determined. Considering that peaks have a Gaussian shape, approximation of peak regions by Gaussian(s) on linear (or non-linear) backgrounds can be done. Peak area is defined as a Gaussian area. Approximation can be done using available software such as IGOR (WaveMetrics, 2018). The Gaussian approximation method is preferable since it can be used for area determinations of non-resolved neighboring peaks, and accuracy of this method of peak area determination is better than the traditional trapezoidal method (Gilmore, 2008).

Examples of peak approximations by Gaussian functions are shown in Fig. 7. The 4.45 MeV peak in the spectra of different sand-carbon mixtures can be approximated by summation of two Gaussian functions with centroids at 4.44 MeV and 4.50 MeV on a linear background. These spectra show that the peak in the sand spectrum can be approximated by one Gaussian function, while two Gaussian functions are needed for good fitting in sand-carbon mixtures. The peak with a centroid at 4.50 MeV can be attributed to the silicon cascade transition peak, while the peak with a centroid at 4.44 MeV can be attributed to carbon. The silicon transition peak varies slightly from spectrum to spectrum, while the carbon peak increases with increasing carbon content in the sample. The value of the carbon peak area can be used for carbon content determinations.

Calibration dependencies

To assess soil elemental content from the net peak area of interest, calibration dependencies should be created. To accomplish this task, PFTNA measurements of field or synthetic soil with well-known elemental contents of interest should be conducted. The following is a discussion of creating the calibration dependence for carbon.



Fig. 4 The PFTNA system is lifted by the crane for measuring background spectra.

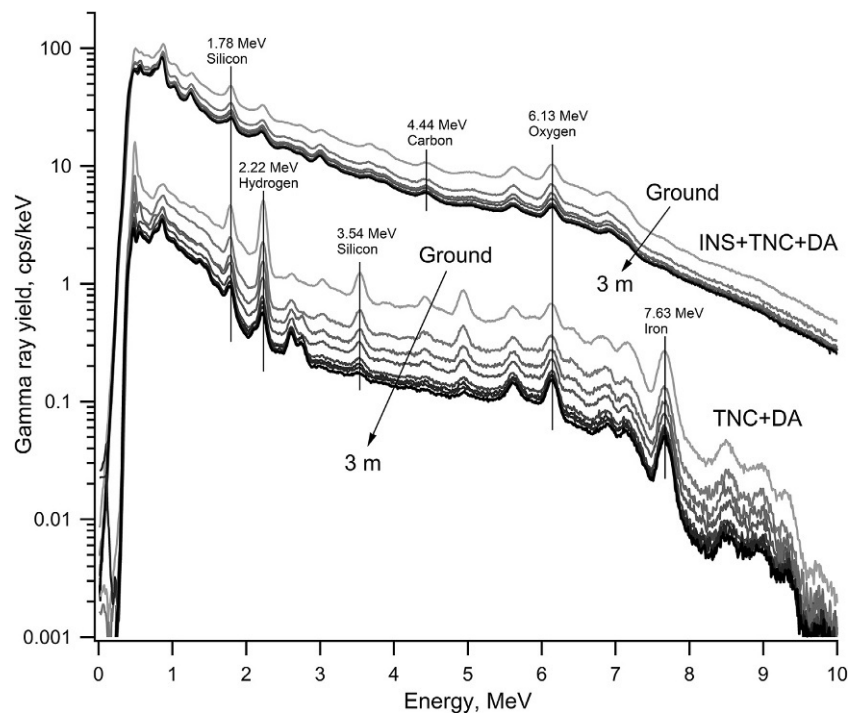


Fig. 5 Measured (INS + TNC + DA) and (TNC + DA) spectra of the PFTNA system at different heights above the ground.

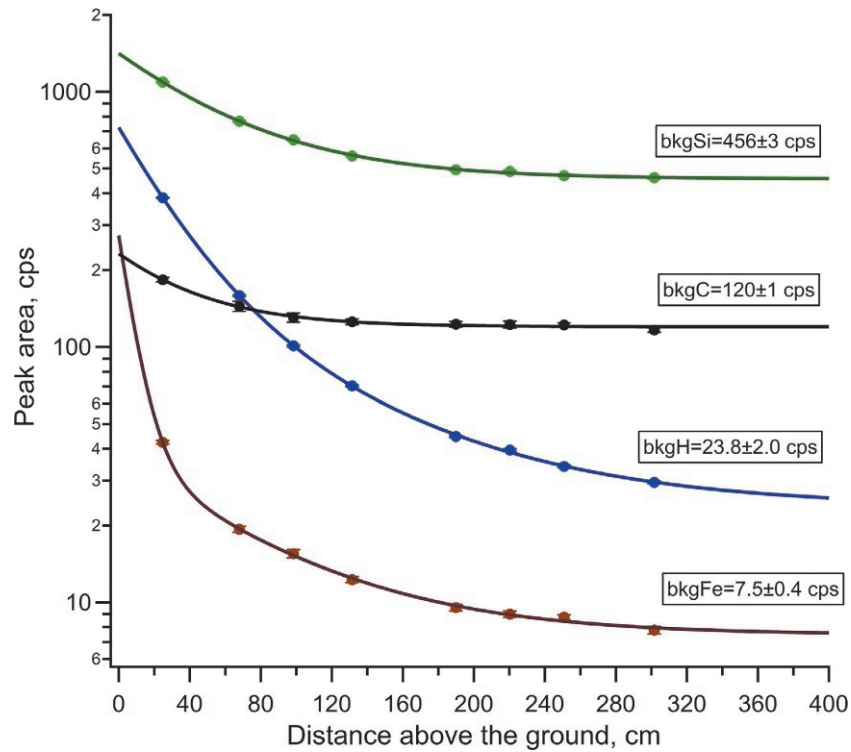


Fig. 6 Relations between different peak areas with height.

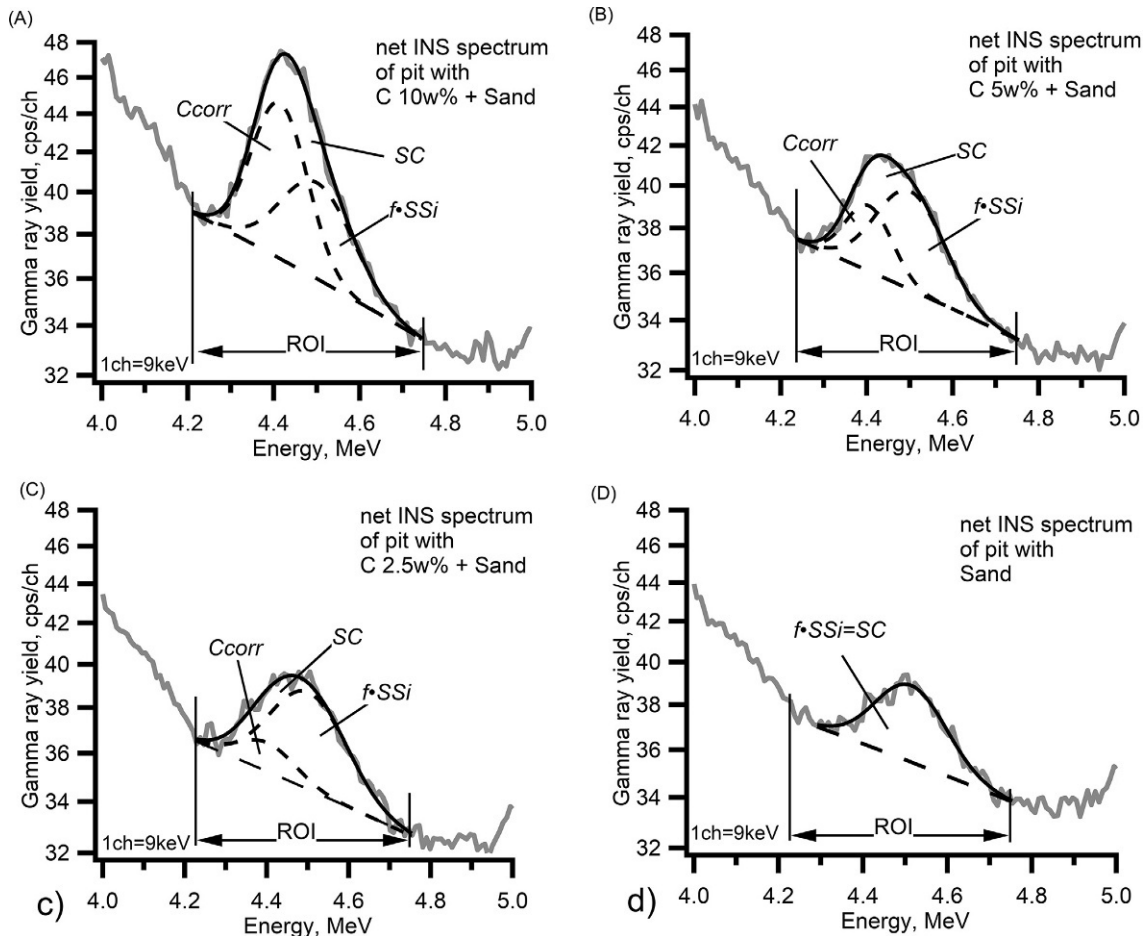


Fig. 7 Fragments of net INS spectra of sand-carbon mixtures (solid gray line) measured by the PFTNA system described and their approximation in the region of interest (ROI, solid black line around 4.44 MeV) by summing of two Gaussian functions (dashed lines) on the linear background (dashed line, long strokes). (A) sand + 10 w% of carbon; (B) sand + 5 w% of carbon; (C) sand + 2.5 w% of carbon; (D) sand. Modified from: Yakubova G, Kavetskiy A, Prior SA and Torbert HA (2017a) Applying Monte-Carlo simulations to optimize an inelastic neutron scattering system for soil carbon analysis. *Applied Radiation and Isotopes* **128**, 237–248.

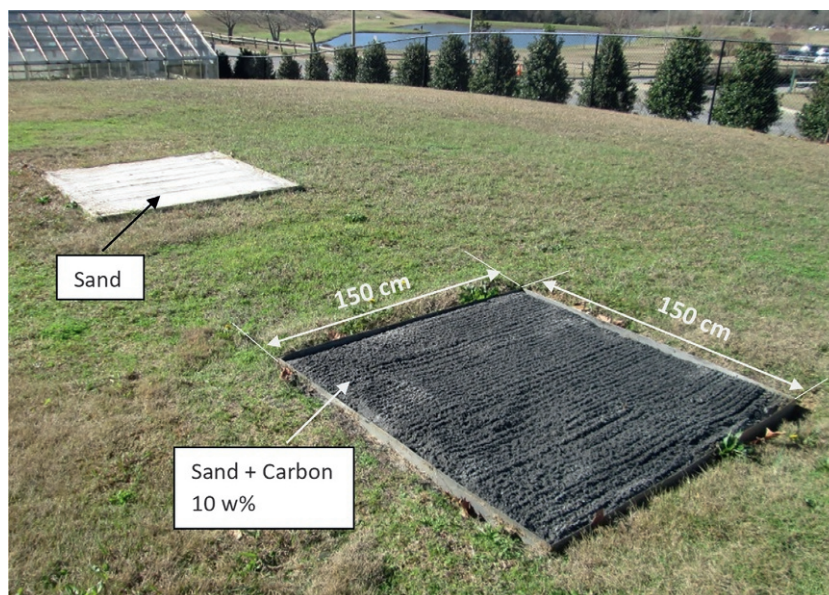


Fig. 8 View of calibration pit with sand and pit with 10 w% sand-carbon mixture. Modified from: Yakubova G, Kavetskiy A, Prior SA and Torbert HA (2017b) Measurements of soil carbon by neutron-gamma analysis in static and scanning modes. *Journal of Visualized Experiments* **126**: e56270.

Carbon calibration dependency

Four calibration pits with carbon-sand mixtures of known carbon content were prepared. The size of each pit was $1.5 \text{ m} \times 1.5 \text{ m} \times 0.6 \text{ m}$, and these pits can be seen in Fig. 8. The PFTNA system was positioned over each pit in such a manner that the neutron source was located over the geometric center of each pit. Net INS spectra were calculated for each pit while accounting for the system background spectra as described above. Fragments of peaks around 4.45 MeV in the experimental net INS spectra and their Gaussian approximations are shown in Fig. 7A–D.

It is important to note that peak fitting by summation of two Gaussian functions gives approximately the same value for different component parameters; for this reason, the sum was used in the analysis while areas of the components were not. The silicon cascade transition component in the 4.45 MeV peak was defined as follows. The silicon net peak area in the net INS spectrum of the i -th pit was denoted as SSi_i (centroid at 1.78 MeV) while the 4.45 MeV net peak area in the spectra of i -th pit was denoted as SC_i . Both peak areas in each spectrum were found by Gaussian approximation. The net carbon peak area (centroid at 4.44 MeV) in the i -th spectrum was denoted as $Ccorr_i$, and the carbon content in the i -th mixture was denoted as $Cont_i$. The assumption was that:

$$Ccorr_i = SC_i - f \cdot SSi_i \quad (1)$$

where f is the ^{28}Si nucleus cascade transition coefficient. These designations are shown in Fig. 7 A–D for clarity. $Ccorr_i$ was considered to be directly proportional to the carbon content of the mixture as $s \cdot Cont_i$. Coefficient s gives an idea of the sensitivity to system carbon (cps per unit of the amount of carbon in the sample). Using SC_i and SSi_i data, the values of f and s can be determined by minimizing the following expression:

$$\sum_i (SC_i - f \cdot SSi_i - s \cdot Cont_i)^2 \rightarrow \min \quad (2)$$

The f and s values were found by equating the derivatives of this sum with respect to f and s set to zero. The standard mathematical software MathCAD was used for these calculations. The defined value of f (0.056) for the example discussed was very close to the theoretically calculated value of the cascade transition coefficient for silicon (0.0547; Herman et al., 2007). This agreement supports the use of this method for determining carbon calibration dependence.

The dependence of the net carbon peak area ($Ccorr_i$) with pit carbon contents ($Cont_i$, expressed in weight percent) is shown in Fig. 9. As can be seen, this is a directly proportional dependence that passes through the 'zero-zero' point (in frames of accuracy of $Ccorr_i$ determinations). The slope of this dependence is the coefficient s (defined as equal to $13.5 \pm 0.6 \text{ cps w\%}^{-1}$) which represents the system sensitivity to carbon.

Determination of soil carbon content (in weight %)

Thus, to determine soil carbon content from the PFTNA measurement for a given area, we need the following:

- defined net INS spectra;
- defined silicon net peak area, SSi_{soil} , and 4.45 MeV net peak area, SC_{soil} , using Gaussian approximations;
- calculated soil carbon content $Cont_{\text{soil}}$ (in w%) using calibration coefficients f and s from the following equation:

$$Cont_{\text{soil}} = \frac{SC_{\text{soil}} - f \cdot SSi_{\text{soil}}}{s} \quad (3)$$

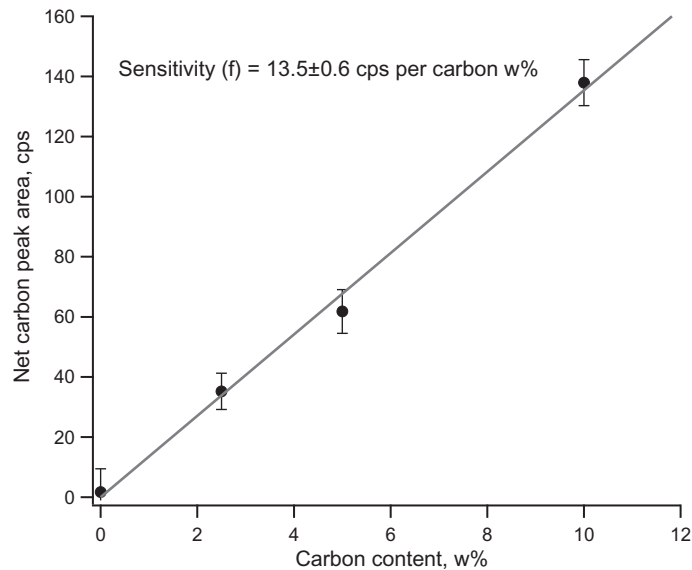


Fig. 9 Carbon calibration dependence. Points with an error bar are data calculated from measurements; straight line is an approximation.

Moisture (hydrogen) calibration dependency

Generally, calibration dependence can be created by parallel measurements of some element content by PFTNA and another method. In the case of hydrogen, the parallel method can be moisture measurements using a nuclear moisture-density gauge. Moisture by this gauge is determined through the detection of thermalized neutrons ('fast' neutrons which have been slowed down by hydrogen present in the material, normally in the form of water). The ^{241}Am —Be radioisotope is used in this gauge as a neutron source. The net hydrogen peak area measured by the PFTNA method compared to moisture measured by the nuclear moisture-density gauge is shown in Fig. 10. Measurements at several locations in the field plot are collected. As can be seen, the dependence of the net hydrogen peak area with moisture is not linear and can be approximated by a parabola. Super linear increasing of the hydrogen peak area can be explained by the increasing number of hydrogen nuclei and the simultaneous increase in thermalization of fast neutrons with increasing moisture. The reversible function of the parabolic fitting function can be taken as the calibration dependence for calculating soil moisture from the net hydrogen peak area.

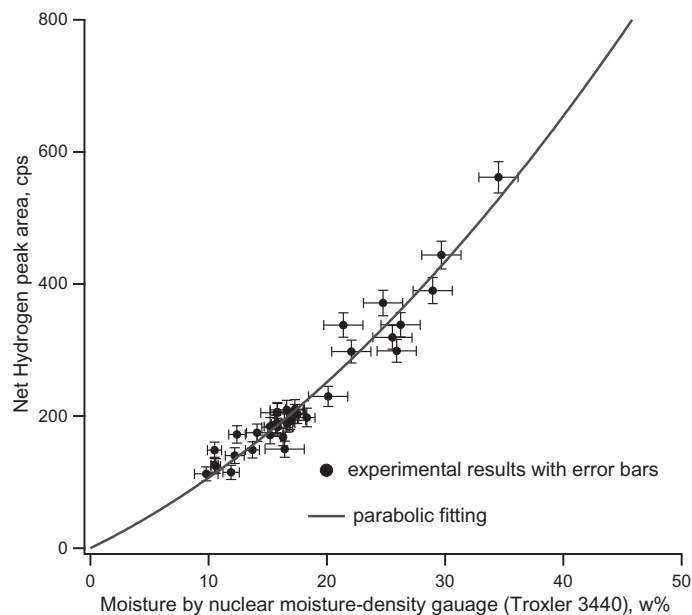


Fig. 10 The net hydrogen peak area vs moisture measured by Troxler 3440 moisture gauge.

Comparison of PFTNA and chemical analysis (dry combustion) data

Verification of determinations of soil elemental content by the PFTNA method was done by assessing carbon levels by traditional methods. For this purpose, three to five soil cores (4 cm in diameter $\times$ 30 cm in length) were taken around the center of each position in the field where PFTNA measurements (in static mode) were conducted. All cores were divided into 10-cm increments, sieved (2 mm), and oven-dried at 55 °C until constant weight. Subsamples were ground with a roller grinder and analyzed by the dry combustion technique (LECO TruSpec CN analyzer, LECO Corp., Saint Joseph, MI) to determine soil carbon concentration in percent by weight. Results of carbon content measurements by the two methods (dry combustion and PFTNA; average weight percent in the upper 10-cm soil layer and its standard deviation) are shown in Table 2. For clarity, the data are also shown in Fig. 11. These data demonstrate good agreement between both methods, especially for average values for the whole field. It should be noted that the accuracy of the carbon concentration measurement using PFTNA is no worse than that of the dry combustion method.

Table 2 Comparison of 10 cm layer soil carbon content (in w%) measurement results by PFTNA and chemical analysis (dry combustion).

Location site#	PFTNA measurements			Dry combustion measurements		
	Carbon Content, w%	Standard Deviation, w%	Mean $\pm$ error by Field, w%	Carbon Content, w%	Standard Deviation, w%	Mean $\pm$ error by field, w%
OF1	2.20	0.29	2.24 $\pm$ 0.23(95%)	2.85	0.25	2.25 $\pm$ 0.26(95%)
OF2	2.51	0.29		2.54	0.31	
OF3	1.76	0.22		1.91	0.13	
OF4	1.88	0.23		2.99	0.94	
OF5	2.82	0.25		3.03	0.37	
OF6	2.15	0.21		1.99	0.26	
OF7	2.77	0.32		1.92	0.41	
OF8	2.52	0.25		2.44	0.15	
OF9	2.06	0.26		1.79	0.27	
OF10	2.17	0.27		2.25	0.45	
OF11	2.39	0.22		2.23	0.30	
OF12	3.11	0.31		2.91	0.47	
OF13	1.44	0.25		1.49	0.42	
OF14	1.93	0.29		1.80	0.19	
OF15	1.86	0.27		1.67	0.25	

Modified from: Kavetskiy A, Yakubova G, Torbert HA and Prior SA (2017) Neutron-stimulated gamma ray analysis of soil. In: Maghraby AM (ed.) *New Insights on Gamma Rays*, pp. 133–178. INTECH: <https://www.intechopen.com/books>.

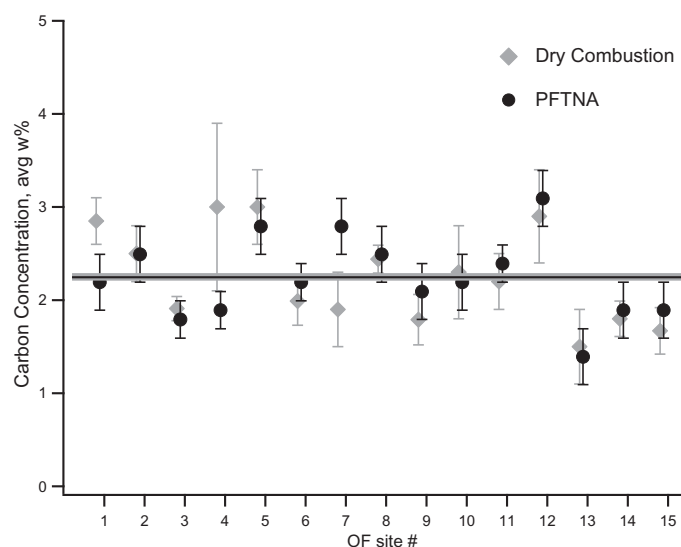


Fig. 11 Comparison of results of soil carbon content (in w%) measurements in the 10-cm layer by PFTNA and chemical analysis (dry combustion). Modified from: Kavetskiy A, Yakubova G, Torbert HA and Prior SA (2017) Neutron-stimulated gamma ray analysis of soil. In: Maghraby AM (ed.) *New Insights on Gamma Rays*, pp. 133–178. INTECH: <https://www.intechopen.com/books>.

Methodology of field PFTNA measurements

Soil volume associated with PFTNA measurements

Soil irradiation by fast neutrons excites soil nuclei in some given volume. Relaxation of these nuclei produces the gamma response. The measurement and analysis of these gamma rays provides information on the elemental content in the excited volume. Computer simulations of the 14.1 MeV neutron propagation in soil using MCNP code (MCNP 5, 2007) showed that 90% of the gamma rays hit the detector in an approximately ellipsoidal shaped volume of $\sim 0.22 \text{ m}^3$ with a footprint of 2.1 m^2 ($r \approx 0.8 \text{ m}$) and a depth of 24 cm, while 99% of gamma rays hit the detector from a soil volume that is 3.5 times larger (Wielopolski, 2011). In this simulation, the soil element (nuclei) distribution was taken as uniform. Thus, in real measurements of soil with uniformly distributed elements (nuclei), the gamma response (peak area of some nuclei in gamma spectra) can be attributed to the corresponding average elemental content (i.e., weight percent) in the indicated volume. In some cases, element distribution with soil depth is not uniform. For instance, a carbon depth profile can be described as exponential with a two-fold decrease at 10–15 cm (Wielopolski, 2011). In such cases, the result of measuring the gamma peak area of some nuclei (for instance peak with centroid 4.44 MeV for ^{12}C) can be attributed, as was demonstrated by Yakubova et al. (2017a, 2017b), to the average carbon content in the upper 10-cm soil layer (in weight percent) independent of the depth distribution shape. Thus, PFTNA carbon content measurements are attributed to the average carbon content (in weight percent) in the upper 10-cm soil layer.

The PFTNA measurement time

Two main factors have an effect on the accuracy of determinations of average soil element content for some given volume. There are statistical errors associated with gamma counts acquired in the peak of interest, and random variation in soil elemental content across short distances (decimeters to meters). The average elemental content in soil and cross-section values of neutron interaction affects the statistical error. A suitable value of this error component can be adjusted by selecting the measurement time. To eliminate measurement uncertainty associated with short-range elemental distribution in non-homogeneous soil, measurements can be conducted in the scanning mode with the measurement system moving slowly (4–8 km per hour) across the field. Error analysis of static and scanning measurement results indicates (Kavetskiy et al., 2019) that scanning measurements can be conducted with the same accuracy as static measurements in approximately a quarter of the time. For example, to obtain results analogous to traditional chemical analyses in the case of carbon [i.e., ± 0.26 (95%) w%], scanning time over a given site should be ca. 15 min using the current measurement system configuration. For other elements with higher content and neutron interaction cross-sections, measurement time can be significantly less.

Acquiring data for elemental mapping

In addition to short distance variations in elemental contents, long-range (hundred to thousand meters) variations can also occur. Long-range trends of changing elemental content are of interest when mapping soil elemental distribution for agricultural purposes. Elemental distribution on maps can be represented as isolines overlaying a geographical map of the studied field. Data suitable for mapping should be prepared as datasets of geographical coordinates and corresponding elemental contents. Isolines across the field can be created from such datasets by different methods (e.g., kriging, local polynomial interpolation) using available software e.g., ArcMap (2018).

The mapping procedure can be as follows. The measurement system moves across the field while simultaneously measuring the gamma spectra. Acquired gamma spectra and geographical coordinates of the measurement system's position are saved every 30 s ($\sim 50 \text{ m}$ of travel). The travel track across the surveyed field should approximately ensure even coverage of the field. To accomplish this, the field is virtually divided into sites of approximately equal area. The size of each of these sites should be about 10 times more than the area covered for one measurement (30 s or $\sim 50 \text{ m}$ of travel covered an area $\sim 150 \text{ m}^2$). The system should measure each site for a defined time. For example, if a 5 min defined time is used, the 10 spectra measured within each site will be saved and used for the elemental measurement in that area. During data processing, the difference between two sequentially recorded spectra and geographical coordinate midpoints of where they were recorded are determined and the differential spectral measurements (midpoints spectra) are assigned to these midpoints.

In addition, there are two possibilities that may occur. If the error of determination of the net peak area of interest from the midpoint spectra is more than 10% (like in case of carbon), these midpoints spectra should be sorted and summarized by site. The net elemental peak area and elemental content should be determined from this summary site spectrum. Geographical coordinates of points (attributed to the determined value) are defined as a weighted position of midpoint coordinates attributed to this site. This database (contents-coordinates) will be used for mapping. The number of records in the data base will be equal to the number of sites. On the other hand, if elemental content and neutron interaction cross-section is such that errors of determination of net peak area from midpoint spectra are no more than 5–10% (cases of hydrogen, silicon), site sorting is not required. The net peak area and determination of elemental content can be done using the midpoint spectra. In this data base, the number of records will be equal to the number of midpoints.

If the purpose of soil elemental analysis is to determine the average elemental content of a given surveyed field (e.g., average carbon content for carbon credits), this value can be defined from spectra of the last record of the field scan. In this case, the total

time of the field survey should be no less than 2 h. In this case, the scanning path should eventually cover the whole field. This can be done by virtually dividing the field into equal site areas and scanning each site for the same amount of time. Under this condition, the carbon value returned from the last recorded spectra gives the average carbon content for the whole field.

Determination of total field soil carbon content for 10- and 30-cm layers

The total carbon content in the upper 10 or 30 cm soil layer of a surveyed field can be defined from PFTNA measurement results. Knowledge of these can be useful in carbon credit programs. In addition to PFTNA data of carbon content (in w%), field soil density (d in kg m^{-3}) is required. Total field soil carbon in the 10 cm layer (TC10, t) can be determined according to following equation:

$$TC10 = \sum_{i=1}^n \frac{Cont_{\text{soil } i}}{100} \cdot d_i \cdot S_i \cdot \frac{0.1}{1000}, \quad (4)$$

where n is the number of sites in a divided field for PFTNA measurements, $Cont_{\text{soil } i}$ and d_i , S_i are soil carbon content (w%), soil density (kg m^{-3}) and area (m^2) of the i -th site, respectively. Area can be taken from the computer software that divided the field into sites. Given that the PFTNA measurement result is an average soil carbon content for the field, $\overline{Cont_{\text{soil}}}$, then:

$$TC10 = \frac{\overline{Cont_{\text{soil}}}}{100} \cdot \overline{d_{\text{field}}} \cdot S_{\text{field}} \cdot \frac{0.1}{1000}, \quad (5)$$

where $\overline{d_{\text{field}}}$, S_{field} are average field soil density (kg m^{-3}) and field area (m^2), respectively.

As can be seen from these equations, the soil density value is required for determining the mass of carbon in soil. The nuclear moisture-density gauge is a suitable instrument for soil density determinations under field conditions. This instrument does not require soil sampling and laboratory processing. With this device, a radioactive ^{137}Cs source is inserted to a desired depth based on a previously prepared narrow channel in the soil, and the soil density is defined from the gamma flux attenuation as it propagates through the soil. The measurement procedure for the nuclear moisture-density gauge (e.g., Troxler Model 3440) is relatively easy, and one measurement that takes 1 min can be done in parallel with PFTNA scanning. For these reasons, the nuclear moisture-density gauge was used for soil density determinations. Also note that by using available computer codes (e.g., Google Earth), the area of sites, S_i , or area of field, S_{field} , can be easily defined with good accuracy.

Total carbon content in the upper 30-cm soil layer of the surveyed field (TC30, t) can be defined as

$$TC30 = \frac{TC10}{0.65}, \quad (6)$$

where the coefficient 0.65 is the ratio of the carbon surface density, g cm^{-2} , in the 10-cm layer to the carbon surface density in the 30-cm layer with an error ± 0.10 . This coefficient was found to be an average value for different carbon depth profiles for several agricultural fields in Alabama. As previously discussed, carbon depth profile data were determined using traditional chemical and mass analysis.

Soil elemental field distribution maps

As mentioned above, the main purpose of the PFTNA agricultural field survey was to create a soil elemental distribution map that would be useful in assessing modern agricultural practices, or for defining total carbon content in the 30-cm soil layer for carbon credit purposes. Let us consider examples of such maps.

Soil carbon content distribution map

A soil carbon distribution map of an agricultural field in Alabama (Lee County) is shown in Fig. 12. A scanning survey of this field was conducted to create this map. Movement tracks of the measuring system are illustrated by thin black lines (Fig. 12). Small black points mark places where spectra were recorded at 30-s intervals. A total of 867 records were acquired during this scanning, and the total survey time was 7.2 h. The white polygons are the virtual sites with approximately equal areas. The measurement system spent about the same amount of time at each site. After determining the midpoint spectra and midpoint coordinates between neighboring recorded spectra, data were sorted by sites and summary spectra for each site were determined. At sorting, 33–35 midpoints were attributed to each site. Associated carbon contents in upper 10-cm soil layer were calculated from summary spectra and attributed to one point in each site. These points are weighted centers of midpoints attributed to sites. They are marked by black triangles with yellow values being carbon content. Using this data base (carbon contents and their weighted centers), the carbon distribution map was created. A local polynomial interpolation method with a kernel function polynomial (of the order of 5) in ArcMap software was used. Black lines are isolines of carbon distribution and digits on these black lines denote carbon content in weight percent. Isolines are plotted using increments of 0.25 w%. As can be seen from this map, the variation in carbon content across this field is small. Some decrease in carbon content from northeast to southwest can be observed. This map can be used for assessing agricultural practices (e.g., conventional tillage vs no-tillage practices, row crop sequence, cover crop usage, pasture management, etc.).

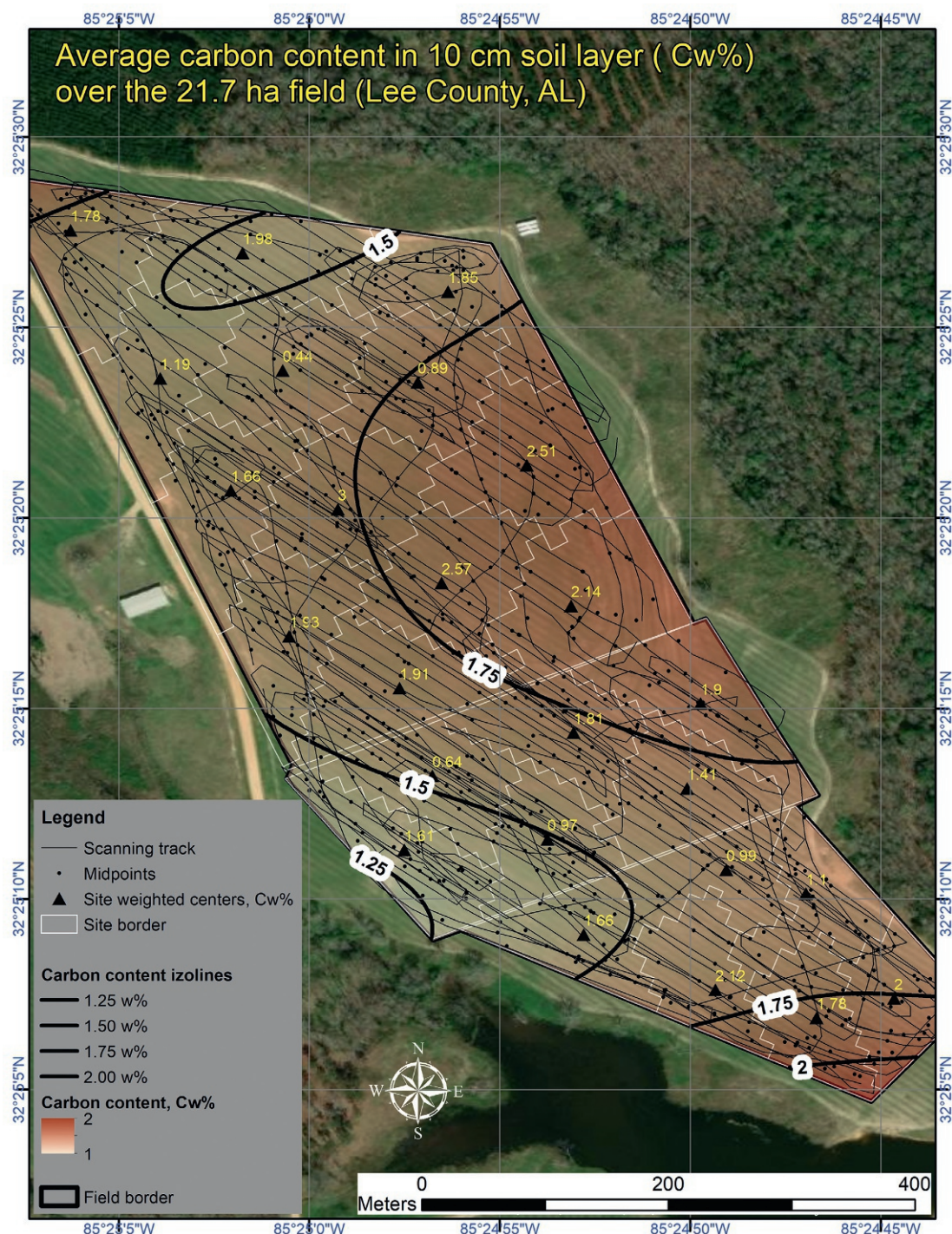


Fig. 12 Map of soil carbon distribution in an agricultural field in Alabama.

Using the defined average weight percent of carbon at each site at 10-cm depth, the total carbon content in the 30-cm soil layer over this field can be calculated using Eq. (6). This value was equal to 860 ± 100 (95%) t. From the last record, the average carbon weight percent for the whole field was 1.75 ± 0.28 (95%) w%, and the total carbon content in the 30-cm soil layer was 900 ± 300 (95%) t. Considering limits of error, values of total carbon content determined by each method are in accord, but the accuracy of the first method is much better than that of the second method. Thus, the first method for determining total field carbon content is preferable while the second method can be used for approximate estimates.

Map of moisture distribution

A map of moisture distribution for an agricultural field in Alabama is shown in Fig. 13. This map was created using a dataset consisting of midpoint coordinates and moisture defined from midpoints spectra acquired from each 30-s measurement. To acquire information for this dataset, the field was divided into virtual sites of equal area which provided uniform coverage of the field from scanning paths. Again, the local polynomial interpolation method with a kernel function polynomial of the order of 5 in ArcMap software was used for mapping. The black points on the map mark midpoint positions to which were attributed defined values of moisture. The black lines with digits are isolines showing constant levels of moisture with digits denoting moisture in percent.

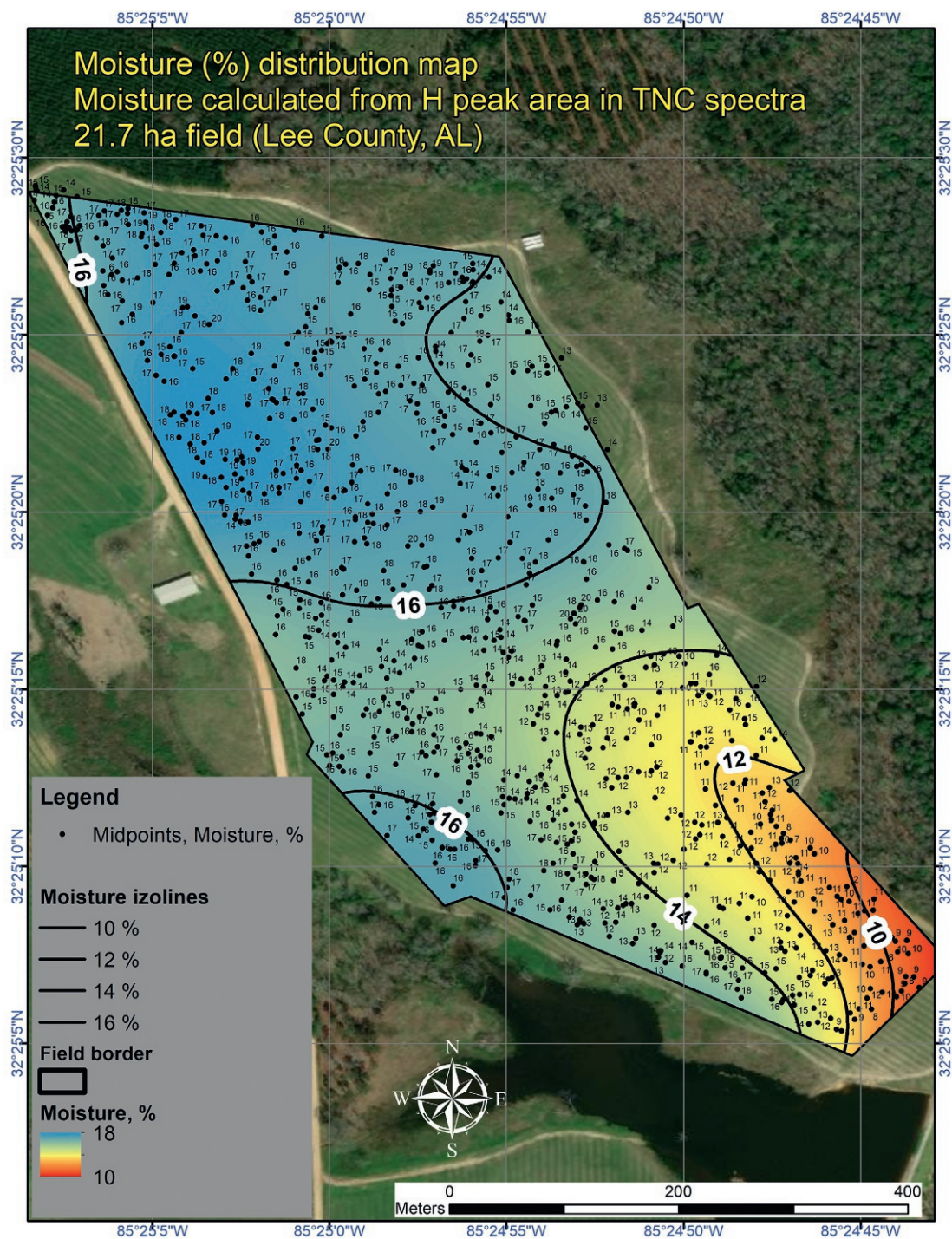


Fig. 13 Map of moisture distribution for an agricultural field in Alabama.

Isolines were plotted with moisture steps of 2%. As can be seen from this map, practically all field values of moisture are ~16% and there is only a slight decrease in moisture in the south-east corner. This map can be used for assessing agricultural practices. Naturally, this moisture map is representative for the date of survey and will change as weather conditions change over time.

Conclusion

Applications of neutron-gamma analysis both in soil science and in assessing agricultural practices in terms of soil elemental determinations can serve as an alternative to traditional chemical analysis. This method is based on measuring the gamma ray response that appears by irradiating objects of interest, in particular soil, by fast neutrons. This method of soil elemental analysis has advantages over traditional laboratory chemical methods. This is a nondestructive *in situ* method of analysis that requires no sample preparation and can perform simultaneous multi-elemental analyses of primary soil elements (Si, C, O, H—moisture, Fe) in large soil volumes. Importantly, measurements are negligibly affected by local sharp changes in elemental content. The accuracy of soil elemental content determinations by the neutron-gamma method is no worse than chemical analysis as was shown in the soil carbon example. These advantages support use of the neutron-gamma method in soil science and for assessing agricultural practices.

The Pulsed Fast/Thermal Neutron Analysis (PFTNA) is a modification of neutron-gamma analysis of soil that applies both static and scanning modes. This method, paired with GPS, presents the possibility of acquiring a data set suitable for creating *in situ* maps of elemental distribution. Together with the nuclear method of the determination of field soil density (e.g., Troxler 3440), the PFTNA can be applied to define total soil carbon in fields. Despite some error (10 or even 30% at 95% confidence level), acquired data can be used for carbon credit estimates.

Modifications of neutron-gamma analysis such as the tagged neutron method or associated particle imaging (API) method are promising directions for *in situ* determination of soil elements. The better signal-to-noise ratio of the API method (compared to PFTNA) can decrease the measurement time required to reach desirable accuracy. Further development of this method for applications in soil science and agriculture is currently underway.

References

- ArcMap (2018) ArcMap Desktop. <https://support.esri.com/en/products/desktop/arcgis-desktop/arcmap/>
- Barzilov AP, Novikov IS, and Womble PC (2012) Material analysis using characteristic gamma rays induced by neutrons, gamma radiation. In: Adrovic F (ed.) *Gamma Radiation*, pp. 17–40. INTECH. <https://www.intechopen.com/books>.
- Gilmore G (2008) *Practical Gamma-ray Spectrometry*, 2nd edn, Chichester, England: John Wiley & Sons, Ltd.
- Google (2019) Google Maps. <https://www.google.com/maps>
- Herman M, Capote R, Carlson BV, Oblozinsky P, Sin M, Trkov A, Wienke H, and Zerkov V (2007) EMPIRE: Nuclear reaction model code system for data evaluation. *Nuclear Data Sheets* 108: 2655–2715.
- IAEA (2014) *Prompt Gamma-ray Neutron Activation Analysis*. <https://www-nds.iaea.org/pgaa>
- Kavetskiy A, Yakubova G, Torbert HA, and Prior SA (2017) Neutron-stimulated gamma ray analysis of soil. In: Maghraby AM (ed.) *New Insights on Gamma Rays*, pp. 133–178. INTECH. <https://www.intechopen.com/books>.
- Kavetskiy A, Yakubova G, Sargsyan N, Wikle C, Prior SA, Torbert HA, and Chin BA (2019) Scanning mode application of neutron-gamma analysis for soil carbon mapping. *Pedosphere* 29: 334–343.
- MCNP 5 (2007) <https://rsicc.ornl.gov/>
- NNDC (2013) *Upton*. NY: Brookhaven National Laboratory. Available at: <http://www.nndc.bnl.gov/chart>.
- Smith KE, Watts DB, Way TR, Torbert HA, and Prior SA (2012) Impact of tillage and fertilizer application method on gas emissions (CO₂, CH₄, N₂O) in a corn cropping system. *Pedosphere* 22: 604–615.
- Valkovic V (2016) *14 MeV Neutrons: Physics and Applications*. Boca Raton, FL: CRC Press/Taylor & Francis Group.
- WaveMetrics 2018 Igor Pro Standard Software. <https://www.wavemetrics.com/>.
- Wielopolski L (2011) Nuclear methodology for non-destructive multi-elemental analysis of large volumes of soil. In: Carayannis E (ed.) *Planet Earth: Global Warming Challenges and Opportunities for Policy and Practice*, pp. 467–492. INTECH. <https://books.google.com>.
- Yakubova G, Wielopolski L, Kavetskiy A, Torbert HA, and Prior SA (2014) Field testing a mobile inelastic neutron scattering system to measure soil carbon. *Soil Science* 179: 529–535.
- Yakubova G, Kavetskiy A, Prior SA, and Torbert HA (2017a) Applying Monte-Carlo simulations to optimize an inelastic neutron scattering system for soil carbon analysis. *Applied Radiation and Isotopes* 128: 237–248.
- Yakubova G, Kavetskiy A, Prior SA, and Torbert HA (2017b) Measurements of soil carbon by neutron-gamma analysis in static and scanning modes. *Journal of Visualized Experiments* 126: e56270.
- Yakubova G, Kavetskiy A, Prior SA, and Torbert HA (2019) Application of neutron-gamma analysis for determining compost C/N ratio. *Compost Science & Utilization* 27: 146–160.

X-Ray Diffraction

Chao Wang and Cheng Gu, State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University, Nanjing, Jiangsu, P. R. China

© 2023 Elsevier Ltd. All rights reserved.

Introduction	642
Basic principles, methodology and instrumentation	643
X-ray	643
X-ray diffraction	643
The Bragg equation	643
Instrumentation	644
Intensity of the diffraction line	645
Factors affecting the XRD signal strength	646
The XRD analysis for soil mineral qualification	646
The XRD-based quantification methods for soil minerals	646
Method of known addition	646
Absorption-diffraction method	646
Mineral intensity factor method	648
The application of XRD for soil evolution	648
Conclusion	651
References	653

Abstract

Since the discovery of single-crystal X-ray diffraction (XRD), various X-ray diffractometers have been developed and utilized for the identification, characterization and quantification of soil components because of their convenience, efficiency, accuracy and non-destruction. Amongst soil components, soil minerals often act as an indicator of diagenesis in soil, which can be identified and quantified by XRD techniques. Basic principles, methodology and instrumentation of XRD and factors affecting the XRD signal strength are described in this chapter. Moreover, the application of XRD in soil analysis is revealed, especially for soil mineral qualification, soil mineral quantification and soil evolution.

Key points

- Basic principles, methodology and instrumentation of XRD are introduced.
- Factors affecting the XRD signal strength are summarized.
- Soil mineral qualification is studied using XRD.
- XRD-based technique is employed to quantify soil minerals.
- XRD is applied in the investigation of Soil Evolution.

Introduction

Since the single-crystal X-ray diffraction (XRD) was discovered by von Laue in 1912 (Ballon et al., 1983), various X-ray diffractometers have been developed, for example the, Debye-Scherrer-Hull diffractometer, Seemann-Bohlin diffractometer and the most well-known Bragg-Brentano diffractometer (Brentano, 1946; de Wolff, 1948; Guinebreiere et al., 2005). When a substance (crystalline or amorphous) is subjected to diffraction analysis, the sample is irradiated by X-rays to produce different degrees of diffraction (Arndt, 1986). Based on the unique diffraction pattern, the composition, crystal form, intramolecular bonds and molecular configuration can be determined. Due to the convenience, efficiency, accuracy and non-destructive testing, X-ray diffraction has been widely utilized as a definitive technique for the identification, characterization and even quantification of soil components (Barre et al., 2007; Bunaciu et al., 2015; de Araújo et al., 2021).

Soil minerals, including soil primary and secondary mineral, is one of the most important soil components (Lin et al., 2017; Zhang et al., 2006). Soil primary minerals are referred to as substances formed directly from parent rocks, which have undergone different degrees of physical weathering, but with no changes in their chemical components and crystal structures (Liu et al., 2019). Secondary soil minerals are new minerals transformed or generated from primary minerals through chemical or biological processes, which include various simple salts, secondary oxides and aluminosilicate minerals (Liu et al., 2019). Notably,

aluminosilicate minerals, commonly known as clay minerals, are important soil secondary minerals with complex crystal structures. Clay minerals are often used as indicators of diagenesis in soil. For most soil environments, several types of clay (e.g., kaolinite, chlorite, smectite, illite and mica) commonly coexist either within discrete phases or as mixed-layer minerals in which different types of clay layers coexist in the same crystal (Viennet et al., 2015; Zhang et al., 2017). Over the past decades, XRD techniques have been widely used to identify and quantify clay minerals (Han et al., 2014; Ndzana et al., 2019).

Basic principles, methodology and instrumentation

X-ray

An X-ray is a kind of electromagnetic wave with short wavelength (about 0.06–20 Å), which can penetrate a certain thickness of materials, and can make fluorescent materials emit light and induce gas ionization (Gabriel, 1977). When bombarding a metal target with an electron beam, X-rays with specific wavelengths are produced and correspond to various elements in the target, which are termed as characteristic X-rays. The essence of the X-ray is a kind of electromagnetic wave that possesses wave-particle duality. X-rays are characterized by their waviness as they travel in space at certain wavelength and frequencies. Generally, X-rays have wavelengths in the range of 0.01–100 Å, which are between γ and ultraviolet rays.

By stopping fast-moving electrons or ions suddenly, the kinetic energy of the electrons can be partially converted into X-ray energy, thus emitting X-rays. When a beam of high-speed electrons bombards a metal target, the X-ray tube produces two different kinds of X-rays, corresponding to characteristic and continuous X-rays, respectively. Moreover, characteristic and continuous X-rays have different properties, mechanisms and uses.

X-ray diffraction

An important application of X-rays is for X-ray diffraction (XRD) analysis, in which characteristic X-rays are used. The XRD is a method of research in which the diffraction pattern of a material is analyzed to obtain information about its structure, composition or atomic/molecular morphology inside it. It is used to determine the crystal structure, which causes the incident X-ray beam to diffract in many specific directions. By measuring the angles and intensities of these diffracted beams, crystallographers can produce a three-dimensional image of the electron density within the crystal. From the information, the average positions of the atoms in the crystal, as well as the chemical bonds can be determined.

Different atoms lead to various outer electron configurations, so different target materials have different X-ray characteristic wavelengths. The X-rays generated by the target material with a short wavelength are termed short-wavelength target X-rays, while long-wavelength target X-rays refer to those produced by the target material with a long wavelength. The peak of the diffraction pattern is regularly expressed along the 2θ axis when using the short-wavelength target X-ray. While, stretched peaks along the 2θ axis are obtained by using the long-wavelength target X-ray. Notably, the sample surface spacing d value obtained from the diffraction pattern is the same and independent of the target no matter what kind of target X-ray tube is used.

The diffraction peak strength depends strongly on the crystal structure, while the mass absorption coefficient (MAC) of the sample is mainly related to the wavelength of the incident ray. Therefore, the diffraction peak intensity of the same sample obtained with different X-ray sources would be slightly different. In particular, for mixed samples, the MACs between the items vary with the selected wavelength. Therefore, improper selection of X-ray wavelength may lead to inaccurate quantification of the components by XRD.

Since MAC mutations for different elements have varied wavelengths, the wavelength range is termed the absorption limit, and strong scattering fluorescence would occur when exceeding this range. As a result, the atomic numbers of the elements in the sample are generally 1–4 times smaller than that of the target element. For example, it is appropriate to use an iron (Fe) target for the analysis of a substance mainly composed of iron, cobalt (Co) and nickel (Ni), but this is not suitable for a sample containing elements of manganese (Mn), chromium (Cr), vanadium (V) and titanium (Ti).

The Bragg equation

The Bragg equation, $2d\sin\theta = n\lambda$, is the basic rule for X-ray diffraction in crystal analysis. It reflects the relationship between the crystal structure and the direction of the diffraction line. In the formula, θ , d , n , λ and 2θ are the incident angle, crystal plane spacing, diffraction order, wavelength of the incident ray and the diffraction angle, respectively (Fig. 1). The crystal spacing, d , depends on the interference index and cell type, indicating cell shape and size, which can be deduced by the Bragg diffraction angle determined to establish the relationship between the diffraction pattern and the crystal structure.

According to the above Bragg equation, if the X-ray satisfies the crystal plane, X-ray diffraction intensity will be strengthened. Therefore, when the reflection rays of different crystal faces are strengthened, it means that the difference in distance between the reflection rays of adjacent crystal faces is an integral multiple of the wavelength.

The Bragg equation is applied as below. When the d value of the crystal is known, the λ value of the characteristic X-ray can be calculated by measuring the θ value. Also, the element that produces the characteristic X-ray can be deduced based on the

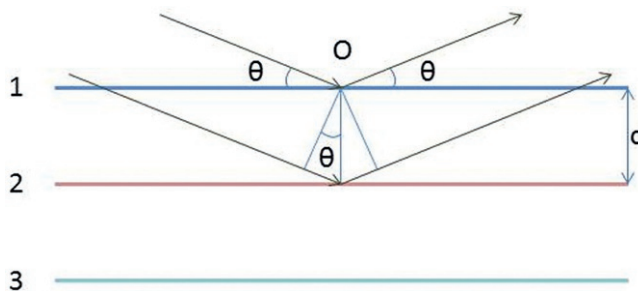


Fig. 1 Schematic diagram of the Bragg equation. Adapted with permission from Bunaciu AA, Udristioiu EG and Aboul-Enein HY (2015) X-ray diffraction: Instrumentation and applications. *Critical Reviews in Analytical Chemistry* **45**: 289–299.

λ value determined. Moreover, given the wavelength λ of the incident X-ray, the spacing d between the surface and the network can be determined by measuring θ . As a result, the crystal structure is determined or the phase is analyzed by the distance between the surface and the net.

Instrumentation

Fig. 2 shows that an X-ray diffractometer contains three basic elements: an X-ray tube, a sample holder and an X-ray detector. The X-rays are generated in a cathode ray tube, in which electrons are initially produced by heating a filament, and accelerated by the applied voltage to bombard the target material. When the generated electrons have enough energy to move the electrons out of the inner shell of the target material, characteristic X-ray spectra are produced. These spectra consist of several components, including K_α and K_β , where, K_α is composed of $K_{\alpha 1}$ and $K_{\alpha 2}$ to some extent. The wavelength of $K_{\alpha 1}$ is slightly shorter and the intensity is twice that of $K_{\alpha 2}$. The specific wavelength is a feature of the target material, e.g., copper (Cu), iron, molybdenum (Mo), chromium. By filtering the X-rays through a foil or crystal monochromator, monochromatic X-rays required for diffraction analysis can be generated. The $K_{\alpha 1}$ and $K_{\alpha 2}$ have similar wavelengths, so the weighted average of the two parameters (K_α) is used. The wavelength of the X-ray can be adjusted by changing the material of the anode target of the X-ray tube, and the intensity of the X-ray source can be controlled by varying the anode voltage and the current of the X-ray tube. Copper is the most commonly used target material for single-crystal diffraction, whose K_α radiation equals to 1.5418 Å.

The sample holder is located at the center of the goniometer circle and can be rotated around the central axis. The goniometer is an instrument used to maintain the angle and rotate the sample, while, the X-ray detector is installed on the circumference of the goniometer and can also be rotated around the central axis. To ensure that the detector is always in the reflection direction, the detector and the sample stage rotate at the same time. For instance, the sample is rotated at an angle of θ on the path of the collimated X-ray beam, while the X-ray detector is mounted on an arm and is rotated at an angle of 2θ . Then the diffracted X-ray signals are collected by recording the reflected X-ray intensity. An intensity peak appears when the geometry of the incident X-ray satisfies the Bragg equation and constructive interference occurs. Finally, the detector will record and process the X-ray signals, and the results are converted into the count rate of photons, which is subsequently output to devices such as a printer or computer monitor.

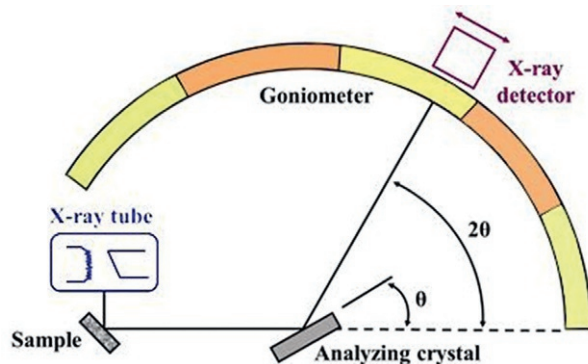


Fig. 2 Schematic diagram of a diffractometer system. Adapted with permission from Bunaciu AA, Udristioiu EG and Aboul-Enein HY (2015) X-ray diffraction: Instrumentation and applications. *Critical Reviews in Analytical Chemistry* **45**: 289–299.

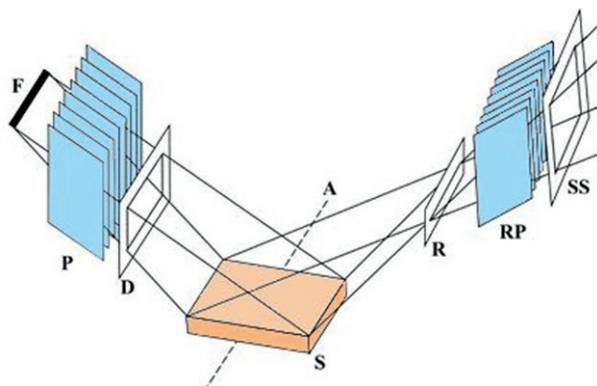


Fig. 3 Geometry of the Bragg-Brentano diffractometer. Adapted with permission from Bunaciu AA, Udristoiu EG and Aboul-Enein HY (2015) X-ray diffraction: Instrumentation and applications. *Critical Reviews in Analytical Chemistry* **45**: 289–299.

Fig. 3 illustrates the geometric arrangement known as the Bragg-Brentano para-focusing system, which includes a system source (F), Soller slits (P and RP), sample (S), divergence slit (D), and receiving slit (R). Line A is the axis of the goniometer and distances FA and AR are equal. In the system, the line source F emits a diverging beam, falling on sample S. After being diffracted, the light beam reaches the detector through the receiving slit R. The effective focal length of the light source F and the aperture of the divergence slit D determine the amount of divergence. The axial divergence is controlled by two Soller slits (sets of parallel plate collimators, P and RP) between F and A, and A and SS, respectively. The use of a narrower divergence slit will give a smaller sample coverage at a specific diffraction angle, thereby allowing a lower diffraction angle to be obtained when the sample has a larger apparent surface, which is indicative of a larger θ value. However, this can only be achieved at the expense of lower intensity. Therefore, the choice of divergence slit and its matching scatter slit is determined by the range of angles covered. It depends strongly on the intensity available to decide whether to increase the size of the slit at a given angle. The diffracted X-ray photons will be converted into voltage pulses by a photon detector, commonly a scintillation detector, which is placed behind the scatter slit. These pulses can be integrated in a rate meter, giving an analog signal on the x/t recorder, with which the scanning speed of the goniometer is synchronized, and a plot of the relationship between 2θ degrees and intensity (i.e., diffractogram) can be obtained.

Most of the basic geometric issues of diffractometers were pointed out around 2000. The width of the diffraction line is reduced by the introduction of the focusing geometry, which corresponds to the improvement in instrument resolution function (IRF). The IRF is an important driving force in the competition between X-ray diffractometer manufacturers. Besides IRF, the pattern acquisition time is another important aspect. In addition to the detector and X-ray source, the geometry of the diffractometer also affects the reduction in pattern acquisition time. In general, reducing the pattern acquisition time and improving IRF are conflicting choices. Therefore, the design of the diffractometer aims to achieve a good compromise between the two factors. Based on this, X-ray diffractometers have been developed in different directions, satisfying different purposes. However, no diffractometer is suitable for all measurements. Therefore, it is necessary to select a convenient instrument when solving scientific or technical problems.

Intensity of the diffraction line

The Bragg equation determines the direction of the diffraction line, relating it to the fundamental period of crystal structures. By measuring the diffraction direction, the symmetry type and cell properties, the crystal structure can be determined theoretically, but not the type and position of the atom arrangement.

Furthermore, the element and the atomic position in the crystal determine the diffraction intensity of the X-ray. While, the diffraction pattern depends on the intensity of the diffraction line. There are three basic elements in the diffraction pattern, i.e., the shape, peak position and intensity of the diffraction line. The diffraction intensity can be divided into integral intensity and peak height intensity. The latter corresponds to the intensity related to the peak height of the diffraction line, while the integral strength is the area below the curve and above the base. The expression for the diffraction line intensity of a substance is complex, but it can be written concisely in the following equation (Eq. 1):

$$I = I_0 \times K \times |F|^2, \quad (1)$$

where, I_0 is the monochromatic X-ray incident intensity on a unit cross-section area, F is the structure factor that depends on the crystal structure and the properties of the atoms contained in the crystal, K is a comprehensive factor, including diffraction geometry conditions, the shape and absorption properties of the samples, as well as physical constants during the experiment (e.g., temperature).

Factors affecting the XRD signal strength

Regardless of the mutual interference between incoming rays, scattered rays and the re-scattering phenomenon of the scattered waves, if the research object is an ideal polycrystal, with neither coarse grains nor preferential orientation and with uniform and continuous diffraction rings, under this condition the cumulative intensity I_{hkl} of the diffraction ring per unit length is calculated by the following formula (Eq. 2):

$$I_{hkl} = \frac{I_0}{32\pi R} \left(\frac{\mu_0}{4\pi} \right)^2 \left(\frac{e^2}{m} \right)^2 \lambda^3 (N^2 |F_{hkl}|^2 P_{hkl}) (L_p) (e^{-2M}) [A(\theta)] V, \quad (2)$$

where, e is electronic charge, m is electronic quality, N is the number of unit cells per unit volume, I_0 is the intensity of incident X-ray, λ is the X-ray wavelength, R is the distance between the diffraction line, μ_0 is the line absorption coefficient, hkl is indice of crystallographic plane, F_{hkl} represents the structural factor, p_{hkl} is the multiplicity factor, L_p is the angle factor, e^{-2M} is the temperature factor, $A(\theta)$ is the absorption factor, and V is the volume fraction of the diffracted object.

According to the above formula, the direction and intensity of the diffraction depend on the internal structure of the crystal (e.g., the type, number and arrangement of atoms in the unit cell, etc.) and its periodicity. In the same phase, the factors that affect the diffraction intensity include the degree of crystallinity of the sample (influencing the shape of the diffraction peak), purity (influencing the resolution), the fineness of the powder sample (when the θ value is the same, the absorption is different), and the radiation wavelength (in the d value phase). Meanwhile, the angle factor $\varphi(\theta)$ is different, and the preparation conditions of the powder sample (preferred orientation should be considered) also affect the diffraction pattern. Different test methods will also cause different diffraction intensities.

The XRD analysis for soil mineral qualification

The XRD techniques have been widely applied to analyze soil samples in many fields. As the most fundamental function, the technique is used to identify the components of soil samples. The components can be identified by the 2θ values and intensity of peaks according to the PDF (powder diffraction file) cards of standard samples or previous research. In most cases, for natural soil samples, the signals of quartz and clay dominate the XRD patterns because of their good crystal structures and large contents in soil. However, by analyzing the different positions and intensities of the peaks, interesting results can be obtained, such as the transformation process of minerals and identification of metal or metalloid elements in the samples. By comparing the different XRD patterns of rock samples with increasing weathering, different behaviors of chemical elements can be provided by the dissolution of silicate minerals into the surrounding groundwater system. For example, arsenic geochemistry and mineralogy as a function of particle-size in natural arsenic-enriched soil can be determined via XRD analysis and some other characterization methods. The XRD of soil samples can also assist researchers to discover the soil or ocean condition in paleogeography. On the other hand, minerals present within the soil environment exhibit characteristic crystal structures, chemical compositions and properties that determine the total reserves of essential plant nutrients. Thus, the mineral–nutrient relationships in soils are extensively assessed using cluster analysis of X-ray powder diffraction patterns and compositional methods.

The XRD-based quantification methods for soil minerals

With the complexity of soil components, various strategies have been developed, amongst which three general approaches based on the XRD technique are commonly used for the quantification of soil clay minerals, i.e., methods of known addition, the absorption-diffraction method and mineral intensity factor method (Fig. 4) (Kahle et al., 2002).

Method of known addition

The method of known addition means that known doses of the minerals are added into the sample. By extrapolating to zero, the original content of the phase can be obtained. Therefore, there is a need to consider that how well the added standard minerals match those in the soil samples. The peak intensity is strongly influenced by the structure factors, which are controlled by the atomic positions in the cell unit and the chemical compositions of the minerals. For example, isomorphous replacement of iron in clay minerals has a dominant control on the structure factor. However, the method of known addition is very time-consuming for a soil sample that is composed of several clay mineral phases, and cannot be recommended for researchers with large quantities of samples.

Absorption-diffraction method

The absorption-diffraction method has been used to quantify soil clay minerals, in which the intensities of the diffraction peak i can be determined for phase α in the mixture of unknown composition and in the pure phase, respectively. If both the mass absorption

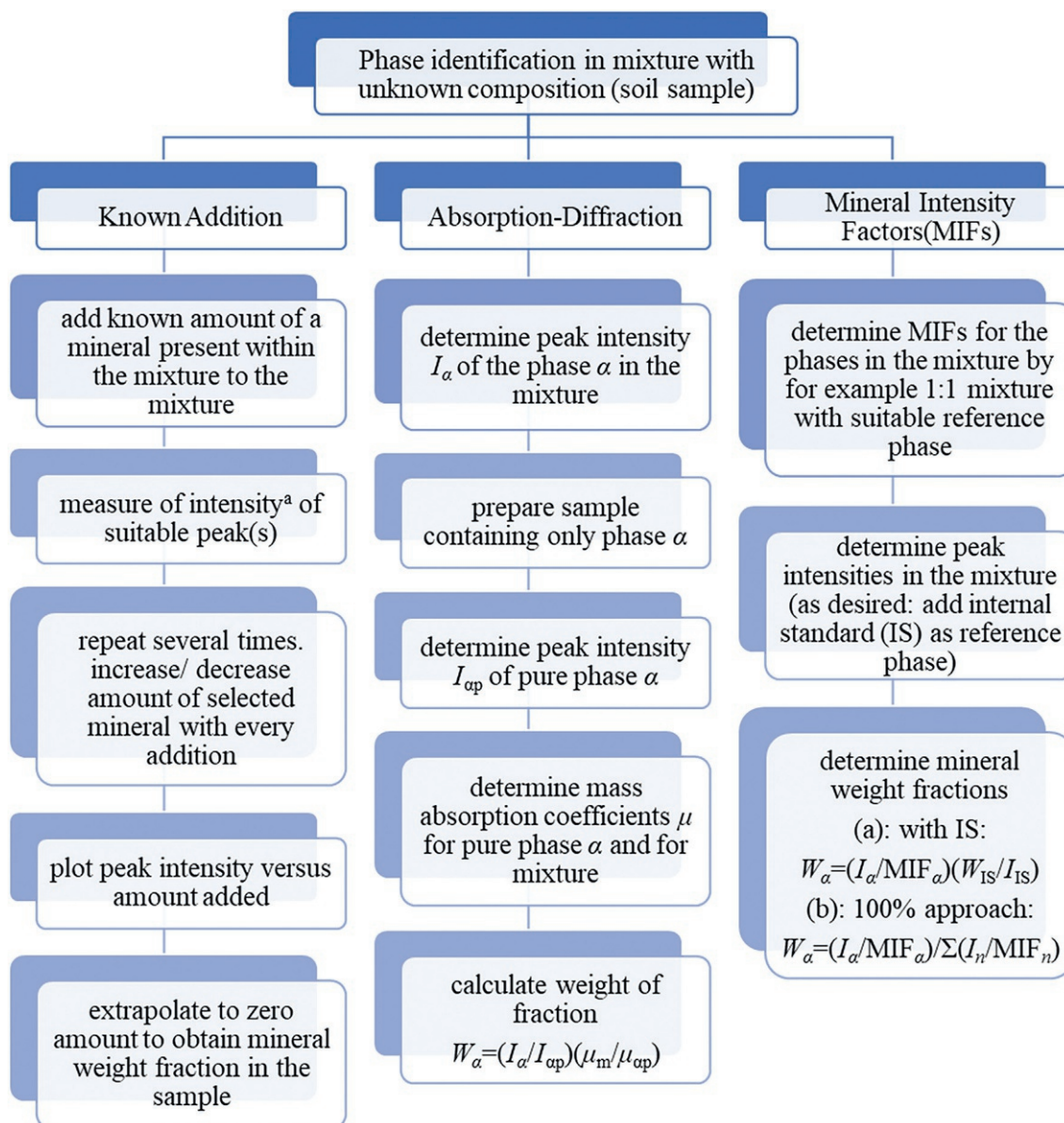


Fig. 4 Protocols for XRD-based quantification of clay minerals in soil samples. Reproduced with permissions from Kahle M, Kleber M and Jahn R (2002) Review of XRD-based quantitative analyses of clay minerals in soils: The suitability of mineral intensity factors. *Geoderma* **109**: 191–205.

coefficients of the pure phase α and the mixture are available, and the weight fraction of phase α can be determined by comparing the peak intensity ratio. However, this method is probably not suitable for the quantification of clay minerals in oriented aggregates, which could be attributed to inaccurate determination of the variability of preferred orientation between the mixture of unknown composition and the pure phase. It is a matter of fortune to select a standard mineral with diffraction characteristics identical to those of the mineral in the sample, which is similar to the method of known addition.

So far, the absorption-diffraction method and mineral intensity factors have been the most commonly used approaches for XRD-based quantification of soil minerals. The general relationship between the weight fraction of a mineral and the integrated intensity of a diffraction peak is simplified in the following equation (Eq. 3):

$$I = W(1/V^2)(1/\rho)IFl^2PL\psi(1/\mu)K \quad (3)$$

where I , W , V^2 , ρ , IFl^2 , P , L , ψ , μ and K refer to the integrated intensity of a diffraction peak, the weight fraction of the mineral, volume of the unit cell, density of the mineral, structure factor, polarization factor, Lorentz factor, powder ring distribution factor,

mean mass absorption coefficient of the mineral mixture and a constant, respectively. For a specific diffraction peak i of a known mineral, V^2 , ρ , IFI^2 , P , L and ψ are supposed to be fixed and the product of these terms can be set as a constant U . As a result, the integrated intensity I_i of the peak i can be simplified as in Eq. (4). Consequently, the peak intensity ratio of two components can be calculated as in Eq. (5) in a given mixture. The ratio of U_α/U_β refers to mineral intensity factor and is also a constant. The proportions of the minerals in a mixture can be calculated if all mineral factors are known in relation to a specific mineral standard.

$$I_i = WU(1/\mu)K \quad (4)$$

$$I_\alpha/I_\beta = (W_\alpha/W_\beta)(U_\alpha/U_\beta) \quad (5)$$

Considering the occurrence of soil clay minerals as a mixture, the signal intensity of each mineral phase in the mixture depends on various physical properties. However, the physical properties are not constants amongst different clay mineral phases, consequently leading to the well-known phenomenon that the relative 001 peak areas of clay minerals in the mixture of equal parts are commonly not equal. Therefore, it is bound to fail for the quantification of clay minerals by referencing individual 001 peak areas to the sum peak areas in a diffractogram.

Mineral intensity factor method

The mineral intensity factor method can follow two different procedures, i.e., the addition of an internal standard and the 100% approach. In the internal standard method, extra internal standards that do not exist in the sample initially are added into it before XRD measurement. The minerals selected as reference internal standards always have the following characteristics: (a) uniform particle size to minimize particle-size segregation when preparing samples; (b) the same crystalline morphology as the unknown minerals in the soil sample; (c) negligible interference with useful diffraction peaks of unknown components; (d) at least one conveniently located strong reflection. For example, pyrophyllite and silicon are commonly recommended as internal standards for the analysis of clay minerals and iron oxide in soil samples, respectively.

The internal standard method can be described theoretically as in the formula below (Eq. 6):

$$x_p = kI_p/I_s \quad (6)$$

where k is a coefficient related to the nature of component P and internal standard s, the properties of the X-ray diffraction test (i.e., the geometry of the apparatus and the wavelength of the X-ray) and the amount added from the standard. x_p is the mass % of component P. The I_p and I_s are X-ray diffraction intensities of component P and the standards, respectively.

Quantification of clay mineral by means of mineral intensity factors consists of two steps. First, the values of mineral intensity factors for phases in a soil sample must be determined using standard mineral mixtures composed of known amounts of the internal standard mineral and one other component. Second, the integrated intensity of the same diffraction peak is measured after adding a known dose of the internal standard with a specific diffraction peak. Quantification of the soil mineral is performed through dividing the integrated intensity of the mineral by the corresponding mineral intensity factors, which is subsequently multiplied with the quotient of the weight fraction.

Considering the presence of short-range-order minerals and non-crystalline materials or misidentified minerals, the sum of the weight fractions of the reference internal standard and soil clay mineral components is commonly not equal to 100%, thus leading to the preference of internal standards from the literature. However, the disadvantages of the internal standard method are related to the difficulty in the selection of standard minerals and the time taken, so another so-called 100% approach has wide application. Suppose that all phases existed in the sample, have been thoroughly identified and the sum of different phase quantities equals to 100%, the weight fraction of each phase (W_α) can then be calculated according to Eq. (7):

$$W_\alpha = (I_\alpha/MIF_\alpha)/\Sigma(I_n/MIF_n) \quad (7)$$

Where I_α and I_n represent the intensities of the selected peaks from phase α and n in the mixture, respectively, MIF_α and MIF_n refer to the mineral intensity factors for phase α and n , respectively.

The application of XRD for soil evolution

Vast areas in the Mediterranean are characterized by evaporite deposits. Large deposits have been built up in a shallow lagoon-like system, which are now widely distributed in Albania, Cyprus, Turkey and southern Italy. These deposits have been extensively investigated for soil formation on evaporites in subarid to arid environments. Although the formation of soil has achieved new significance, little is known about the evolution of the Mediterranean evaporites. Therefore, soil formation is investigated for evaporites collected from the Caltanissetta basin (Sicily, Italy), and the soil profiles with soil horizon designation is shown in Fig. 5. The lithology includes marine clay deposits, laminated marl (diatomite; Tripoli sediments), limestone, detritus deposits and gypsum (selenite). During pedogenesis, element leaching from silicates was the most dominant process in the soil, which was active before the development of clays and the Tripoli sediments. Vertisols and Mollisols evolved together with a large amount of

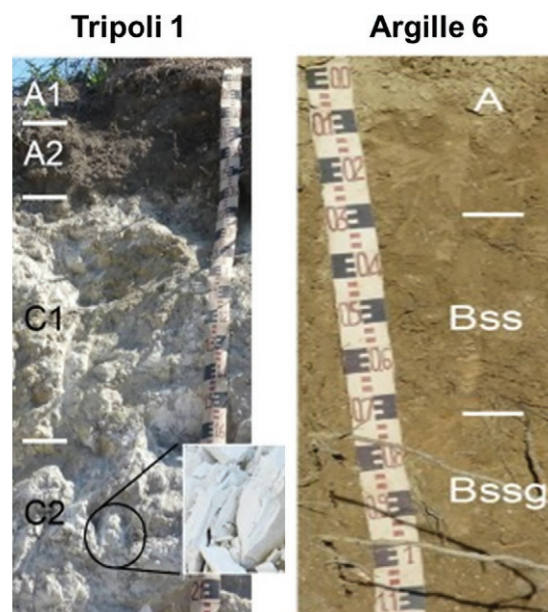


Fig. 5 The soil profiles with soil horizon designation. The soil profiles of Tripoli 1 from top to bottom refer to A1, A2, C1 and C2; The soil profiles of Argille from top to bottom refer to A, Bss and Bssg. Adapted with permission from Egli M, Plötze M, Tikhomirov D et al. (2019) Soil development on sediments and evaporites of the Messinian crisis. *CATENA* **187**: 104368.

kaolinites and oxyhydroxides (non-crystalline and crystalline forms). Weak soil formation was only recognized at sites containing gypsiferous parent material. The presence of selenite can hinder more advanced evolution. The soils developed on detritus and gypsiferous material that contain some palygorskite, which is inherited from the parent material. Moreover, shallow Aridisols or Mollisols are observed at sites that have limestone or selenite. Weathering is pronounced as indicated by a large proportion of trioctahedral minerals. In addition to inheritance from the parent material, kaolinite and smectite also undergo active formation in the soil as weathering proceeds. Plagioclase and mica are the main sources of smectite, and K-feldspar is the main source of kaolinite neoformation (Egli et al., 2019).

The X-ray diffractograms of typical samples after saturation with glycerol, Mg ions and K ions as well as calcination treatments are shown in Fig. 6. After glycerol saturation, the Tripoli samples exhibit basal reflections at 0.72, ~0.90, 1.00, 1.10 and 1.71 nm. The peak around 1.71 nm suggests the presence of smectite (Fig. 6), while, the peak at 0.91 nm is probably due to a reflection ($d(002)$) of the 1.71 nm peak. The peak at 1.00 nm refers to mica, while the diffraction peaks at 1.08, 1.12 and 1.32 nm are assigned to irregularly interstratified minerals (e.g., mica/smectite) (Fig. 6). However, the peak near 1.08 nm may be assigned to palygorskite. The quantitative analysis of minerals in the clay fraction suggests that a minor content of palygorskite is possibly present in the Tripoli, Argille 7, Calcare and particularly the Gessi samples. The peak at ~1.41 nm (Mg saturation) of the A1 horizon from Tripoli 1 (Fig. 6) remains after K saturation. This peak collapses after heat treatment, indicating the existence of some hydroxy-interlayered-smectite (HIS) or hydroxy-interlayered-vermiculite (HIV). Neither HIS nor HIV can be detected in the C2 horizon of the Tripoli soil, demonstrating that HIV is pedogenically formed. In general, there is no distinct difference in the XRD pattern between the subsoil and topsoil. In the Tripoli soil, some HIV and some additional interstratified clay minerals undergo active formation during pedogenesis, whereas changes observed in the Argille soil were barely recognizable. The Argille soil is affected by vertic processes, leading to a homogenization of the top- and sub-soil. Consequently, the mineralogical differences are minor. Compared to the Tripoli soil, the Argille soil contains a distinctly higher content of kaolinite (Fig. 6), while the Tripoli soil has a larger content of interstratified minerals (Egli et al., 2019).

Evaluation of the XRD patterns of phyllosilicates (58–64 degrees 2θ region) confirms the presence of both trioctahedral and dioctahedral phases of mica in the subsoil and topsoil (Fig. 7). Trioctahedral species are represented by peaks ranging from 59 to 61 degrees, and dioctahedral species are found, correspondingly, in the range of 61–63 degrees 2θ . A significant part of the peaks around 0.1537 and 0.1533 nm can be assigned to quartz. The peaks near 0.1520 nm are, thus, more indicative of changes in trioctahedral species. The proportion of trioctahedral species in the Argille soils seems to decrease during the pedogenesis process (the intensity of peaks around 0.1520 decrease from Argille 6A to Argille 6Bss) (Fig. 7), indicating that the trioctahedral mica undergoes active weathering. The contribution of the peak ranging from 0.1495 to 0.1498 nm (kaolinite and smectite) increases with progressive soil evolution, as indicated by the decrease in peak intensity (0.1495–0.1498 nm) from Tripoli 1C2 to Tripoli 1A1 (Fig. 7). This mica is either transformed into smectite or is an interstratified mica/smectite (Egli et al., 2019).

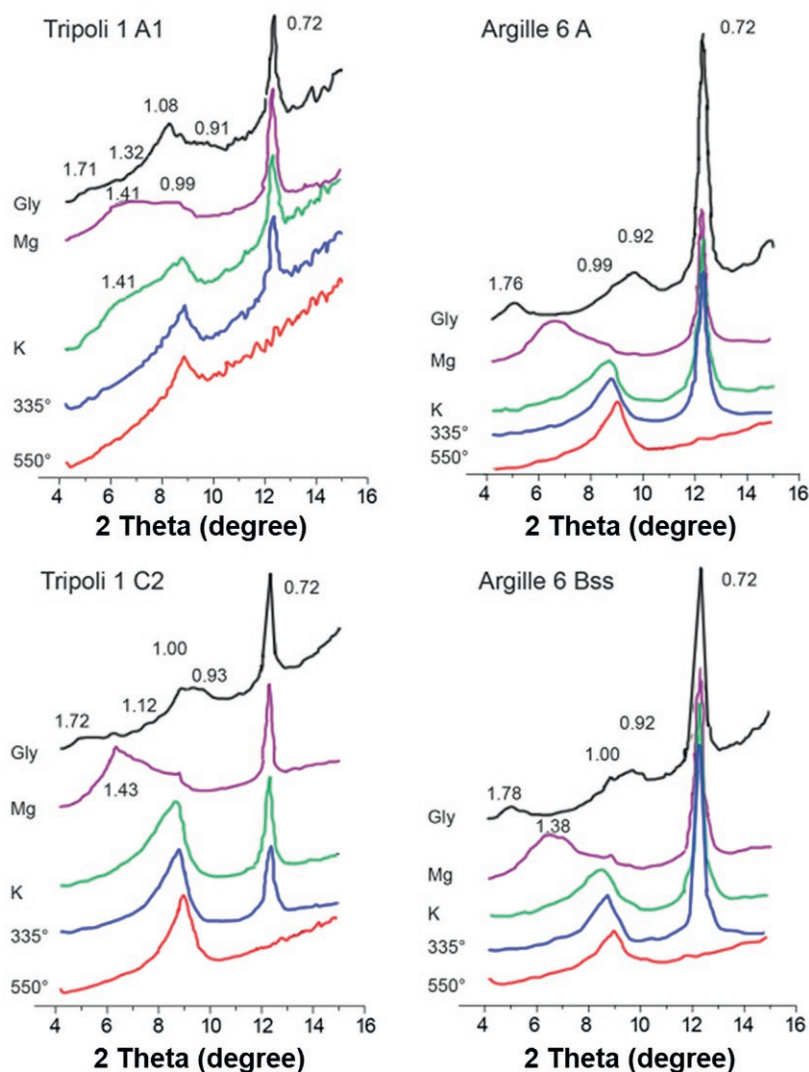


Fig. 6 X-ray diffractograms of the clay fraction of the topsoil and subsoil of Tripoli 1 and Argille 6. In the figure, Gly, Mg, K, 335 degrees and 550 degrees represent glycerol saturation of the Mg-saturated sample, K-saturated sample, heated K-saturated sample at 335 °C, heated K-saturated sample at 550 °C. The *d* spacing is given in nm. Adapted with permission from Egli M, Plötze M, Tikhomirov D et al. (2019) Soil development on sediments and evaporites of the Messinian crisis. *CATENA* **187**: 104368.

The XRD has been applied to analyze the relationship between the formation and evolution of soil clay minerals and environment in a terrestrial ecosystem of Northern China (Zhang et al., 2021). Fig. 8 shows the diffraction peaks located at 3.34, 3.54, 4.25, 4.71, 5.00, 7.10, 10.00 and 14.20 Å, which indicate the presence of quartz, illite, kaolinite and vermiculite. Moreover, consistent results of soil minerals were obtained for farmland, wetland and forest. The reflections for quartz are located at 3.34 Å and 4.25 Å, showing a negligible degree of weathering and proving its strong resistance to weathering. The peaks of K-saturated clay mineral remained unchanged after heat treatments at 300 °C and 550 °C, confirming the existence of illite, it's a non-swelling 2:1 type of clay. Fig. 9 shows that the Mg-air (air-dried) samples have a strong diffraction peak at 14.20 Å, but after treatment with glycerin, this reflection disappeared after high temperature calcination. While, the peak at 10.00 Å was strengthened, further confirming the coexistence of vermiculite (Vanderaverroet et al., 2000). Similarly, the 7.10 Å peak of K-air vanished with heat treatment, suggesting that kaolinite also existed.

The strong microbial activity and high degree of weathering in forest soil may make it a better alternative to farmland and wetland soils to represent the soil mineral evolution of a terrestrial ecosystem. By comparison with the conventional XRD method, synchrotron radiation XRD (SR-XRD) possesses various advantages such as small incident spot diameter, high angular resolution,

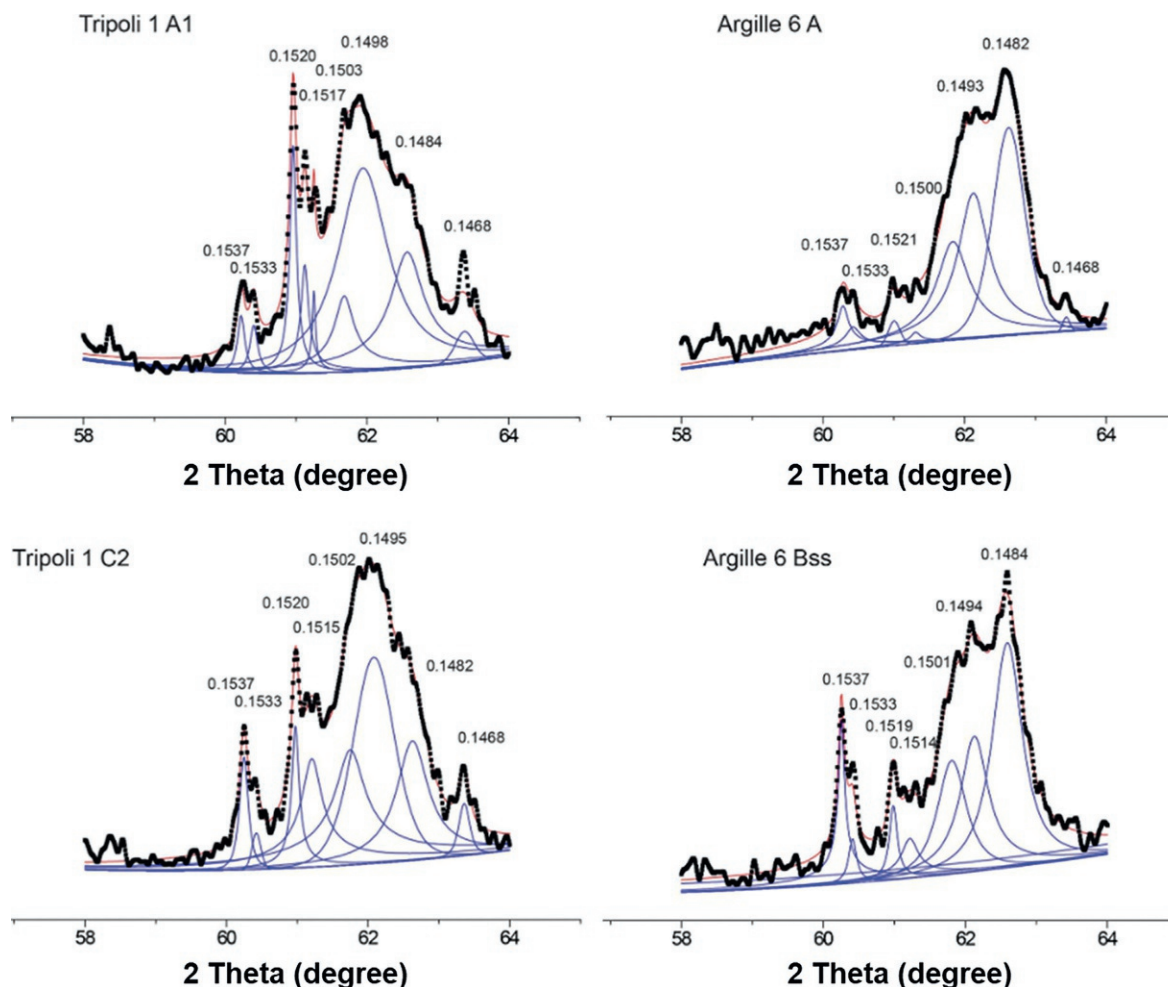


Fig. 7 Peak separation of X-ray patterns in the $d(060)$ region of some typical soil samples (clay fraction). The peak region ranging from 59 to 61 degrees 2θ can be attributed to trioctahedral minerals and the region between 61 degrees and 63 degrees 2θ to dioctahedral species. Adapted with permission from Egli M, Plötze M, Tikhomirov D et al. (2019) Soil development on sediments and evaporites of the Messinian crisis. *CATENA* **187**: 104368.

strong brightness and penetration, thus exhibiting great improvements in the accuracy of soil morphology characterization in both depth and breadth (Christensen et al., 1996; Gualtieri et al., 1996; Cheng et al., 2017). Fig. 9 shows that the diffraction peaks are more concentrated and their intensity is greater than those of conventional XRD. Moreover, SR-XRD produced higher resolution and narrower diffraction peaks for the same scanning range. For instance, the peak at 3.34 Å (referring to illite) is compact in the SR-XRD spectrum, indicating the crystallinity better. Moreover, the diffraction peak was negligible in the conventional XRD spectra, while the reflection signal obtained in the SR-XRD spectra is clear (Figs. 8 and 9). By comparison with conventional XRD spectra, there is a slight turning point at 12.20 Å for the SR-XRD spectra, which subsequently rises rapidly to 14.10 Å. This suggests that a mixed layered mineral of illite and vermiculite might exist and the diffraction peak at 14.10 Å was probably derived from HIV. In addition, the high diffraction peak at 3.19 Å confirms the presence of plagioclase in Fig. 9. Therefore, SR-XRD could serve as an alternative to achieve a more accurate and reliable analysis of clay minerals compared with the conventional XRD.

Conclusion

In this chapter we have provided a detailed introduction of XRD, including the basic principles, methodology, instrumentation, influencing factors and applications. Amongst the wide applications, the XRD-based technique was specifically summarized for soil analysis, i.e., XRD analysis for soil mineral qualification, XRD-based methods of quantification for soil minerals and application of XRD for soil evolution. Furthermore, based on the XRD results, the activity of weathering processes at different soil depths can be

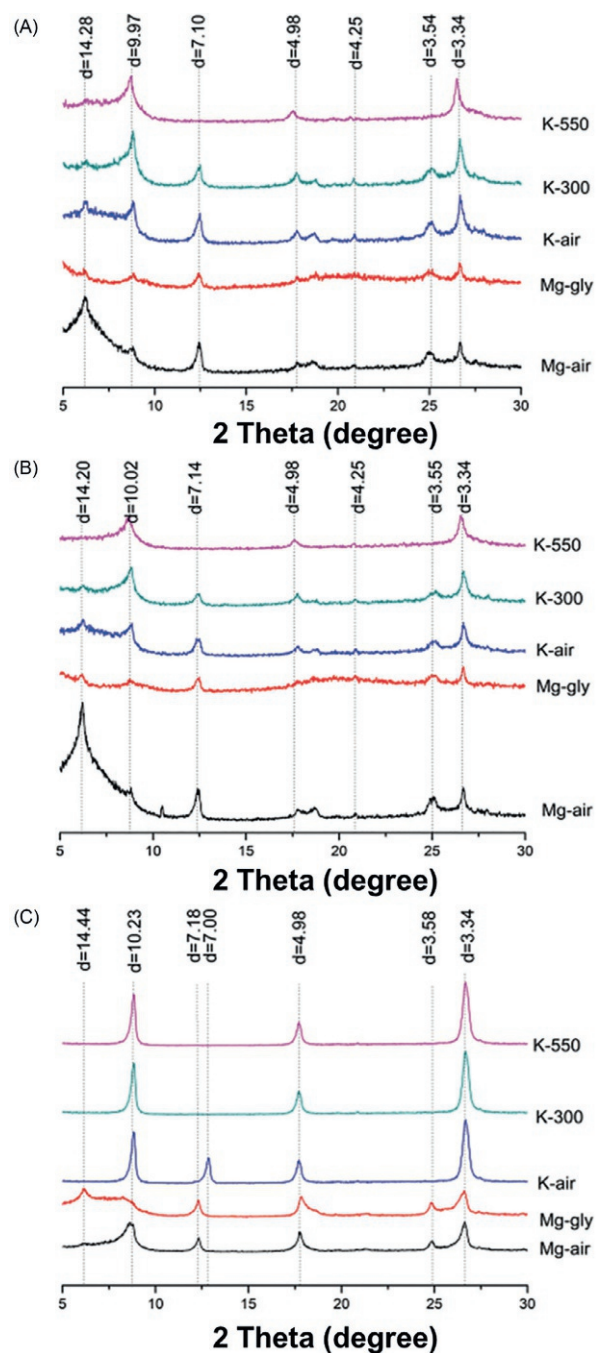


Fig. 8 The X-ray diffraction spectra of clay minerals in farmland (A), wetland (B) and forest (C). K-air and Mg-air refer to air-dried samples saturated with K and Mg ions, respectively. K-300 and K-550 correspond to K-saturated samples oven-dried at 300 °C and 550 °C, respectively. Mg-gly refers to Mg-saturated and glycerin-treated samples. Adapted with permission from Zhang Z, Sheng Q, Zhao M, Zhong J, He N, Li R, Zhang L, Guo D and Zhang J (2021) Analysis of soil clay mineral in terrestrial ecosystem using X-ray diffraction spectroscopy. *Spectroscopy Letters* 54: 65–71.

deduced by comparing the content of minerals. Generally, XRD is an efficient method for qualitative and quantitative analysis of soil components, thus providing further information about mineral transformation processes. Therefore, XRD processing will make greater contributions to promote the development of soil and environmental sciences in the future. By comparison with conventional XRD, SR-XRD can act as an alternative to achieve a more accurate and reliable analysis of clay minerals given its advantages such as small incident spot diameter, high angular resolution, strong brightness and penetration.

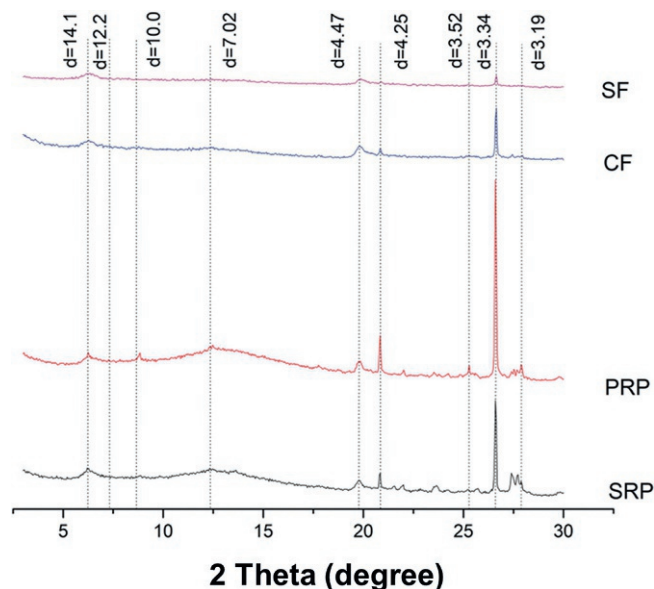


Fig. 9 Synchrotron radiation X-ray diffraction spectra of clay minerals in the forest soil. SF, CF, PRP and SRP are spruce-fir mixed forest soil, coniferous forest soil, primary and secondary red pine broadleaf mixed forest soil, respectively. Adapted with permission from Zhang Z, Sheng Q, Zhao M, Zhong J, He N, Li R, Zhang L, Guo D, Zhang J (2021) Analysis of soil clay mineral in terrestrial ecosystem using X-ray diffraction spectroscopy. *Spectroscopy Letters* 54: 65–71.

References

- Arndt UW (1986) X-ray position-sensitive detectors. *Journal of Applied Crystallography* 19: 145–163.
- Ballon J, Comparat V, and Pouxe J (1983) The blade chamber—A solution for curved gaseous detectors. *Nuclear Instruments and Methods in Physics Research* 217: 213–216.
- Barre P, Velde B, Catel N, and Abbadie L (2007) Soil-plant potassium transfer: Impact of plant activity on clay minerals as seen from X-ray diffraction. *Plant and Soil* 292: 137–146.
- Brentano JCM (1946) Parafofocusing properties of microcrystalline powder layers in X-ray diffraction applied to the design of X-ray goniometers. *Journal of Applied Physics* 17: 420–434.
- Bunaciu AA, Udristoiu EG, and Aboul-Enein HY (2015) X-ray diffraction: Instrumentation and applications. *Critical Reviews in Analytical Chemistry* 45: 289–299.
- Cheng H, Lu C, Liu J, Yan Y, Han X, Jin H, Wang Y, Liu Y, and Wu C (2017) Synchrotron radiation X-ray powder diffraction techniques applied in hydrogen storage materials: A review. *Progress in Natural Science: Materials International* 27: 66–73.
- Christensen AN, Norby P, Hanson JC, and Shimada S (1996) Phase transition of KNO_3 monitored by synchrotron X-ray powder diffraction. *Journal of Applied Crystallography* 29: 265–269.
- de Araújo W, Kloss J, and Volpato N (2021) Evaluating mechanical properties of a photosensitive acrylic resin for additive manufacturing with the addition of organophilic clays. *Journal of the Brazilian Society of Mechanical Sciences and Engineering* 43: 1–11.
- de Wolff PM (1948) Multiple guinier cameras. *Acta Crystallographica* 1: 207–211.
- Egli M, Plötze M, Tikhomirov D, et al. (2019) Soil development on sediments and evaporites of the Messinian crisis. *Catena* 187: 104368.
- Gabriel A (1977) Position-sensitive X-ray detector. *Review of Scientific Instruments* 48: 1303–1305.
- Gualtieri A, Norby P, Hanson J, and Hriljac J (1996) Rietveld refinement using synchrotron X-ray powder diffraction data collected in transmission geometry using an imaging-plate detector: Application to standard m-ZrO_2 . *Journal of Applied Crystallography* 29: 707–713.
- Guinebreiere R, Bouille A, Masson O, and Dauger A (2005) Instrumental aspects in X-ray diffraction on polycrystalline materials. *Powder Diffraction* 20: 294–305.
- Han W, Hong H, Yin K, Churchman GJ, Li Z, and Chen T (2014) Pedogenic alteration of illite in subtropical China. *Clay Minerals* 49: 379–390.
- Kahle M, Kleber M, and Jahn R (2002) Review of XRD-based quantitative analyses of clay minerals in soils: The suitability of mineral intensity factors. *Geoderma* 109: 191–205.
- Lin S, Chen Y, Liu S, and Yuan M (2017) Effect of long term fertilization on soil clay minerals in soil. *Acta Agriculturae Boreali-Sinica* 32: 142–147.
- Liu Y, Zhang M, Guan D, Song X, Yuan F, Wu J, and Wang A (2019) A review of natural science research in Changbai Mountain area: 1956. *Chinese Journal of Applied Ecology* 30: 1783–1796.
- Ndzana GM, Huang L, Zhang Z, Zhu J, Liu F, and Bhattacharyya RJ (2019) The transformation of clay minerals in the particle size fractions of two soils from different latitude in China. *Catena* 175: 317–328.
- Vanderaverroet P, Bout-Roumazilles V, and Fagel N (2000) Significance of random illite-vermiculite mixed layers in pleistocene sediments of Northwestern Atlantic Ocean. *Clay Minerals* 35: 679–691.
- Viennet JC, Hubert F, Ferrage E, et al. (2015) Investigation of clay mineralogy in a temperate acidic soil of a forest using X-ray diffraction profile modeling: Beyond the HIS and HIV description. *Geoderma* 241: 75–86.
- Zhang J, Cui D, Sui F, and Zhang Y (2006) Effect of long-term fertilization on clay mineral composition and main soil physical and chemical properties of noncalcareous fluvo-aquic soil. *Acta Pedologica Sinica* 43: 697–702.
- Zhang S, Liu Q, Yang Y, Wang D, He J, and Sun L (2017) Preparation, morphology, and structure of kaolinites with various aspect ratios. *Applied Clay Science* 147: 117–122.
- Zhang Z, Sheng Q, Zhao M, Zhong J, He N, Li R, Zhang L, Guo D, and Zhang J (2021) Analysis of soil clay mineral in terrestrial ecosystem using X-ray diffraction spectroscopy. *Spectroscopy Letters* 54: 65–71.

Soil analysis using laser-induced breakdown spectroscopy

Débora Marcondes Bastos Pereira Milori^a, Ladislau Martin-Neto^a, Hero T Gollany^b, Diego Victor de Babos^a, Luis Carlos Leva Borduchi^{a,c}, and Paulino Ribeiro Villas-Boas^a, ^aEmbrapa Instrumentação, São Carlos, SP, Brazil; ^bUSDA-Agricultural Research Service (USDA-ARS), Columbia Plateau Conservation Research Center, Adams, OR, United States; ^cInstituto de Física de São Carlos, São Carlos, SP, Brazil

© 2023 Elsevier Ltd. All rights reserved.

Introduction	654
Plant nutrients	656
Toxic elements	657
Soil carbon	658
Soil physicochemical properties	659
Conclusion	660
Acknowledgments	661
References	661

Abstract

Analyses of soil physicochemical properties, including texture, carbon (C) content, and plant nutrients, are essential for effective farm management. However, because they require several time-consuming and expensive methods, such analyses are not viable on a large scale, as in precision agriculture. An attractive alternative to these methods is laser-induced breakdown spectroscopy (LIBS), which provides quick, affordable, and simultaneous assessments of several soil characteristics. This chapter covers the latest advances in LIBS for assessing toxic elements, plant nutrients, C content, and physicochemical properties of soils. The benefits and limitations of LIBS, compared to conventional methods, will be discussed for each assessment.

Glossary

ICP-MS Inductively coupled plasma mass spectrometry
ICP-OES Inductively coupled plasma optical emission spectrometry
LIBS Laser-induced breakdown spectroscopy
LOD Limit of detection
Nd:YAG laser Neodymium-doped yttrium aluminum garnet laser

Key points

- LIBS is an alternative, attractive method for soil analysis;
- LIBS has advantages over conventional methods for soil analyses involving a large number of samples;
- LIBS provides rapid soil analyses with minimal sample preparation, no generation of residues, and simultaneous assessment of several physicochemical properties;
- The main limitations of LIBS in soil analysis are (1) matrix effects and (2) the high LOD values for elements such as N, P, S, Cl, B, Ni, and Zn;
- LIBS can assess, with decent accuracy, plant nutrients, toxic elements, and physicochemical properties;
- Verra has qualified LIBS as a valid method to support soil carbon certification;
- Further studies in LIBS are required to improve the quantification of certain plant nutrients (e.g., N, P, S, Cl, B) and toxic elements (Ni and Zn) in soil samples;

Introduction

Understanding the physicochemical characteristics of soils is crucial for farmers as it enables them to make informed and strategic decisions about farm management. By optimizing the use of inputs and minimizing environmental impacts, farmers can enhance productivity and meet the ever-increasing global demand for food. However, analyzing soil characteristics requires several techniques that are typically time-consuming, expensive, and environmentally unfriendly, as they generate toxic residues. So,

how can soil samples be reliably and promptly analyzed? The solution is the development of new technologies and instruments that give farmers access to soil characteristics rapidly, affordably, and with adequate accuracy and precision. Laser-induced breakdown spectroscopy (LIBS), which requires a single measurement, is an attractive technique that meets these criteria. Recent studies and reviews (Tavares et al., 2022; Villas-Boas et al., 2020a,b) suggest that LIBS can rapidly assess various soil properties such as texture, pH, carbon (C) content, and cation exchange capacity, in addition to the concentration of plant nutrients and toxic elements.

LIBS is a spectroanalytical technique capable of simultaneously identifying and quantifying any element on the periodic table (Cremers and Radziemski, 2013). LIBS analysis takes less than 1 min per sample and can be performed under ambient air atmospheric conditions. This technique requires a simple experimental setup with, at least, five components: a pulsed laser, a focusing lens, an emission collection optical system, a spectrometer, and a delay generator (Fig. 1). By focusing the laser pulse on the sample surface, molecular bonds are broken, and atoms and ions are heated, thus generating a plasma plume with temperatures that can exceed 50,000 K. After a few microseconds, the background emission decreases, at which time the spectrometer gate is opened to acquire the LIBS emission spectrum. By examining the resulting spectrum, the analyst can recognize the elements contained in the sample through their atomic and ionic emissions. The intensity of these emissions, which is usually obtained from the height or area of the lines, reflects the total concentration of the elements and is used in LIBS quantitative models. The bioavailable nutrient content and physicochemical properties of soils can be inferred from the entire spectrum or from the intensity of selected emission lines by multivariate machine learning models.

LIBS offers several advantages compared to traditional methods of soil property analyses. These advantages are: (i) minimal sample preparation since it allows for the direct analysis of solid samples in the form of pellets; (ii) small amount of material required for the samples undergoing analysis (around 0.5 g per sample measurement, depending on the sample and laser characteristics); (iii) simultaneous multielement determinations, including plant nutrients and toxic elements; (iv) high analytical frequency (spectral acquisition within less than 1 ms); (v) no need for chemical reagents; (vi) no generation of post-analysis chemical residues; and (vii) possibility of direct in situ soil analysis with a portable system, without the need to collect and transport the samples to the laboratory. In comparison to other analytical techniques, soil sample preparation for LIBS analysis is relatively simple, and includes sample homogenization using pestle and mortar or an automatic mechanical grinder, sample pelleting, and LIBS analysis (Fig. 2).

LIBS has not yet become a routine technique for laboratory soil analysis due to two key disadvantages (i) matrix effects and (ii) poor limit of detection (LOD) for some plant nutrients and toxic elements. The first disadvantage results from the fact that the LIBS signal depends not only on element concentration, but also on the sample's physicochemical properties, such as chemical composition, light absorption coefficient, and thermal conductivity (Borduchi et al., 2022). Because of matrix effects, developing quantitative models for samples of various composition, especially soils, is challenging. To mitigate matrix effects in soil analyses, several strategies can be used, such as the use of femtosecond or picosecond lasers, ultraviolet laser pulses, spectral pre-processing

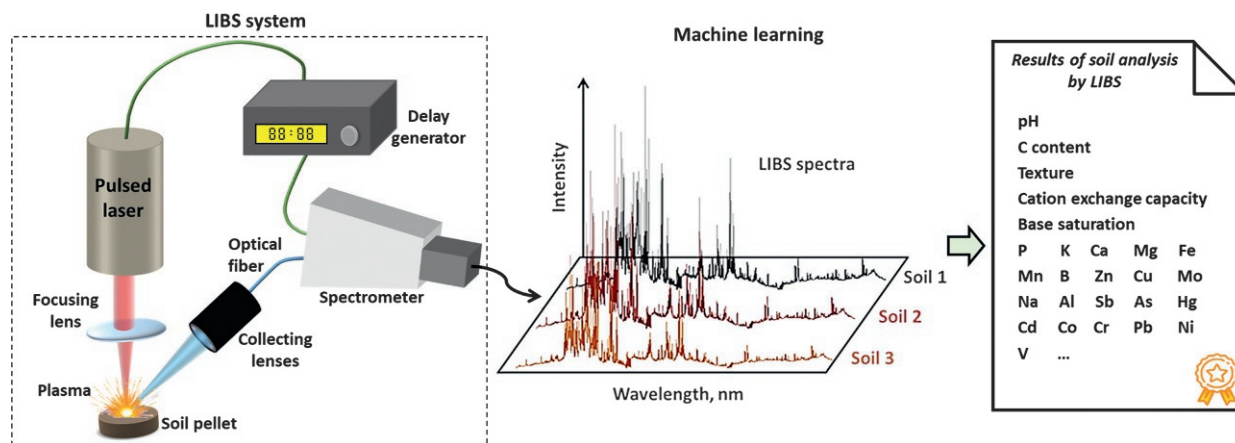


Fig. 1 Overview of soil analysis and the possible determinations using the LIBS technique.

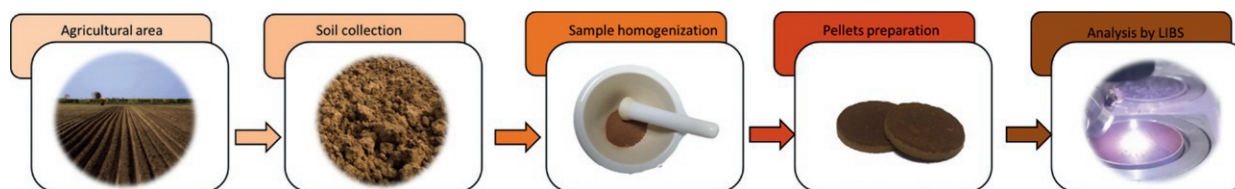


Fig. 2 Steps required for soil analysis using LIBS.

methods, and calibration-free models (Takahashi and Thornton, 2017). The second disadvantage is related to the intensity of the lines that can be detected within the spectral range (usually from 200 to 1000 nm) of a conventional LIBS system. Some elements, such as sulfur (S), potassium (P), chlorine (Cl), arsenic (As), lead (Pb), and mercury (Hg), have weak emission lines in this range, and are typically found in soil at trace concentrations. The LIBS LOD can be improved by using (i) appropriate instrument configurations, such as a double-pulse LIBS system, physical plasma confinement, and microwave antenna (Zorov et al., 2015); or (ii) specific sample preparations, such as the addition of nanoparticles (de Giacomo et al., 2016).

Despite these limitations, LIBS has the potential to be used in precision and digital agriculture. In particular, precision agriculture requires the analysis of numerous soil samples, which is time-consuming and costly, especially if conventional methods are used. However, precision agriculture may be made viable with LIBS with its advantages over traditional methods. LIBS can outperform traditional methods even if, individually, the results of the analyses are less precise and accurate than those obtained through conventional methods. This is because the large number of samples in LIBS compensates for possible inaccuracies in the assessment of a soil property. In addition, LIBS can increase the temporal resolution of the sampled area, assisting farmers in managing agricultural lands properly over time. Based on the fertility maps generated, the farmer can add inputs to the soil—such as limestone to correct soil acidity, or fertilizers—at variable rates (Bernardi et al., 2017). Thus, the synergy between precision agriculture and LIBS enables the optimization of agricultural production, lowering costs, and mitigating environmental impacts.

This chapter presents the advantages and limitations of the application of LIBS in soil analysis to quantify plant nutrients, toxic elements, and soil C, in addition to indirectly assessing soil organic matter, pH, base saturation, cation exchange capacity, and texture.

Plant nutrients

The quantification of plant nutrients in soils is crucial for the effective management of a farm because it indicates the optimum amounts of fertilizers to be applied on each field for a crop to maximize productivity, minimize costs, reduce waste, and mitigate environmental impacts. Each plant uses intrinsic mechanisms to extract essential nutrients from the soil, therefore, several methods for plant nutrient analysis have been proposed to measure the bioavailability of these elements. However, methods of plant nutrient extraction vary from country to country, are all time-consuming, require reagents (extractants that mimic the extraction of nutrients from the soil by plant roots), and produce waste, which makes them impractical to implement on a large-scale (including precision agriculture). Therefore, the LIBS technique may be a desirable alternative for quantifying macro- and micro-nutrients in soil samples (Table 1).

Table 1 LIBS limit of detections (LODs) of soil nutrients (as reviewed by Villas-Boas et al., 2020b), compared to CHNS or ICP LODs (according to the manufacturers), and typical content found in soils (Larcher, 2003).

Nutrient (Symbol)	Typical content (mg kg ⁻¹)	CHNS or ICP LOD (mg kg ⁻¹)	LIBS LOD (mg kg ⁻¹)	Most used emission lines (nm)
Macronutrients (primary)				
Nitrogen (N)	2000	100 ^a	8000 ^b	742.36, 744.23, 746.83
Potassium (K)	14,000	50	31	404.72, 518.36, 766.49, 769.90
Phosphorus (P)	800	50	251	178.3, 213.61, 255.32, 255.5
Macronutrients (secondary)				
Calcium (Ca)	15,000	2.5	180	317.9, 318.1, 393.37, 396.8, 428.3
Magnesium (Mg)	5000	2.5	4	280.27, 285.2, 383.23, 383.81, 516.73, 517.26
Sulfur (S)	700	100 ^a	16000 ^c	180.7, 182.1, 182.6, 921.3, 922.8, 923.8
Micronutrients				
Iron (Fe)	40,000	2.5	10	274.6, 274.9, 404.6, 406.4, 407.2, 489.15
Chlorine (Cl)	<100	5 ^d	75 ^e	133.6
Manganese (Mn)	1000	2.5	7	380.67, 403.07, 403.31, 403.45, 475.4, 472.7, 478.3
Boron (B)	20	5	1000	255.14 (BO band)
Zinc (Zn)	90	2.5	7	213.86, 334.50, 334.59, 472.22, 481.05
Copper (Cu)	30	2.5	0.6	282.43, 324.75, 327.4, 510.6, 515.3, 521.8
Molybdenum (Mo)	3	5	0.3	313.26, 550.65

Metallic and metalloid nutrients were measured under ambient air atmospheric conditions, while non-metallic nutrients were measured in a chamber with specific conditions, as indicated in the footnotes.

^aCHNS (CHNS elemental analyzers).

^bChamber with air at 0.04 torr; Ar at 0.3 Mpa; or Ar at 1 atm.

^cChamber with air at 0.02 mbar for VUV; or air at 1 atm for UV-VIS-NIR.

^dICP (inductively coupled plasma—mass spectrometry, ICP-MS), otherwise ICP optical emission spectroscopy (ICP-OES).

^eCO₂ at 7 torr.

Based on Villas-Boas PR, Franco MA, Martin-Neto L, Gollany HT, Milori DMBP (2020b). Applications of laser-induced breakdown spectroscopy for soil characterization, part II: Review of elemental analysis and soil classification. *European Journal of Soil Science* 71: 805–818.

The quantification of metalloid (B, boron) and non-metallic (N(nitrogen), P, S, and Cl) nutrients is difficult with the LIBS technique because the emission lines of these elements are weak in the spectral range (200–1000 nm) commonly used in LIBS systems. For this reason, the LOD for these nutrients is high, and may compromise the analysis of soil samples because the concentrations of some elements, which occur at trace levels in soils, may not be detectable (Table 1). In addition, the LIBS measurement of N, S, and Cl requires a chamber filled with a noble gas (e.g., Ar, argon) or thin (near vacuum) atmospheric air. For the measurement of soil N content, N₂ from the air can contribute to plasma formation, making it impossible to distinguish between N from the air and that from the sample. Sulfur and chloride have strong lines only in the deep ultraviolet range, which requires measurements in a vacuum, and with spectrometers designed for this range (less than 200 nm). Boron also has weak lines in the 200–1000 nm range and can be measured with a boron monoxide (BO) molecular fragment emission band at 255.14 nm (emission of boron oxide (BO) molecular fragment); however, the LOD is much higher than the typical concentration in soils (Table 1).

In contrast to the non-metallic elements, the nutrients potassium (K), calcium (Ca), magnesium (Mg), iron (Fe), manganese (Mn), zinc (Zn), copper (Cu), and molybdenum (Mo) are not difficult to detect with the LIBS technique and can be measured under ambient air atmospheric conditions because they have strong lines within the spectral range commonly used in LIBS systems. Thus, the LIBS LOD for the metallic nutrients is far lower than their typical concentrations in soils (Table 1), which makes the LIBS technique advantageous for these elements. However, the main recommendation for assessing these nutrients is to refrain from using the strongest lines, as they are often self-absorbed and may not bear a linear relation with the concentration of nutrients.

The intensity of a nutrient emission line is correlated with the total concentration of the element in question, therefore, the content of bioavailable nutrients cannot be determined directly with LIBS. To overcome this limitation, multivariate machine learning models can be developed with the LIBS spectra (Tavares et al., 2022), which contain information concerning other elements in the soil sample. Thus, the content of bioavailable nutrients can be obtained through the stoichiometric relation between the elements present in the soil samples, which can be revealed with these models.

In summary, LIBS shows great potential for large-scale analyses of plant nutrients in soils because it is fast, inexpensive, and acceptably accurate. Certain challenges, however, are yet to be overcome, such as matrix effects and the high LOD for non-metallic and metalloid nutrients. Matrix effects can be mitigated with (i) additional sample preparation (e.g., use of a binder, or matrix removal through element extraction methods), (ii) spectral treatment methods (e.g., spectral normalization and wavelet transform), and (iii) physical (calibration-free) or multivariate models (e.g., artificial neural networks). The LOD for non-metallic and metalloid nutrients can be improved with the use of molecular bands, a binder (to enhance emission), double-pulse LIBS systems, or a microwave antenna. For these elements, using a vacuum chamber may still be necessary, especially to ensure the absence of atmospheric N₂ while measuring N content in soil samples.

Toxic elements

For a farm to be managed sustainably and to keep food from becoming contaminated before it reaches the shelves, a quantification of the toxic elements in the soil is critical. Depending on their concentration in the soil, certain elements such as copper (Cu) and zinc (Zn) can be either plant nutrients or toxic. A soil is considered contaminated if the concentrations of certain elements exceed the legal threshold (Table 2), which may vary from country to country. The issue with confirming contamination by these elements in soils is that the method most often used (ICP-OES) is time-consuming, requires reagents for sample digestion, and produces waste, which makes it unviable on a large scale.

The benefits of using LIBS, including its rapid analysis time, affordable cost, and zero waste production, have attracted interest in the technique for the measurement of toxic elements in soil samples. However, traditional LIBS systems are not sensitive enough (i.e., their LODs are high compared to ICP-OES) to detect these elements at the concentrations usually found in soils, i.e., from 1 µg kg⁻¹ to 1 mg kg⁻¹ (Table 2). To reduce the LOD, several strategies were suggested (Zorov et al., 2015), such as (i) using the strong lines (high Einstein coefficient for spontaneous emission) for these elements; (ii) optimizing acquisition of the plasma emission; (iii) increasing the plasma lifetime (e.g., with physical or magnetic confinement, using a microwave antenna); (iv) increasing the plasma's energy (e.g., using double-pulse LIBS systems, electrical discharge coupled LIBS system, laser-induced fluorescence coupled LIBS system); and (v) extracting elements with an acid (phase transformation method). With these strategies LIBS can be performed with far smaller LODs than those of the reference method, except for nickel (Ni) and Zn (Table 2). In principle, the LODs for toxic elements could be further improved by combining various strategies, but that would increase the complexity of the procedure and lengthen the time required for the analysis.

Most studies on toxic elements have focused on improving the LODs, therefore, the issue of matrix effects has not been extensively addressed. In addition, the majority of these studies have used the same type of soil samples, varying only the concentrations of the toxic element of interest, resulting in minor matrix variations among the samples. In real-world situations where agricultural or industrial soil samples are evaluated, matrix effects can compromise quantitative models for these elements. Thus, several methods have been developed to address this problem, including:

Table 2 LIBS limit of detections (LODs) for toxic elements (as reviewed by Villas-Boas et al., 2020b), compared to ICP-OES LODs (according to the manufacturers), threshold values for contamination (according to Finnish Ministry of the Environment, 2007, and based on ecological, groundwater (p), and human health risks), and natural occurrence found in soils (Finnish Ministry of the Environment, 2007).

Element (symbol)	Natural occurrence (mg kg ⁻¹)	Threshold value (mg kg ⁻¹)	ICP-OES LOD (mg kg ⁻¹)	LIBS LOD (mg kg ⁻¹)	Most used emission lines (nm)
Antimony (Sb)	0.01–0.2	2 (p)	25	0.221 ^a	252.85
Arsenic (As)	0.1–25	5 (p)	25	3.3 ^b	188.98, 200.34, 228.81, 234.98
Mercury (Hg)	0.005–0.05	0.5	25	0.7 ^c	194.15, 435.83, 546.07
Cadmium (Cd)	0.01–0.15	1	2.5	0.067 ^d	214.44, 226.50, 228.80, 340.37, 346.62, 479.99, 508.58
Cobalt (Co)	01–30	20 (p)	2.5	1 ^e	399.53
Chrome (Cr)	6–170	100	2.5	0.8 ^c	336.80, 337.92, 340.33, 340.87, 342.27, 343.36, 357.87, 359.35, 360.53, 425.43, 427.48, 428.97, 433.94, 434.45, 437.42, 475.61, 478.93, 479.3, 480.10, 520.60
Copper (Cu)	5–110	100	2.5	0.6 ^f	282.43, 324.75, 327.4, 510.6, 515.3, 521.8
Lead (Pb)	0.1–5	60	25	0.94 ^d	405.78
Nickel (Ni)	3–100	50	2.5	6.2 ^g	229.00, 231.10, 231.60, 231.72, 232.00, 333.06, 336.96, 341.48, 344.62, 345.84, 346.17, 346.75, 348.59, 349.30, 352.45, 357.19, 359.77, 361.94, 508.05, 510.00, 511.54
Zinc (Zn)	8–110	200	2.5	7 ^d	213.86, 334.50, 334.59, 472.22, 481.05
Vanadium (V)	10–115	100	2.5	1 ^h	411.52, 437.92

Zinc and copper can be either nutrients or toxic, depending on their concentration in the soil. All LIBS measurements were performed under ambient air atmospheric conditions.

Particular measurement conditions are indicated in the footnotes.

^aLIBS assisted with laser-induced fluorescence for Sb emission line.

^bLIBS operating with Nd:YAG laser at 1064 nm, 450 mJ and 5–7 ns per pulse.

^cCoarse metal powder-assisted pulse CO₂ LIBS.

^dLIBS assisted with phase transformation method (solid-liquid-solid).

^eLIBS assisted with laser-induced fluorescence for Co emission lines.

^fLIBS operating with Nd:YAG laser at 355 nm, 21 mJ per pulse.

^gLIBS using hemispherical plasma spatial confinement.

^hLIBS and plasma spatial confinement.

Based on Villas-Boas PR, Franco MA, Martin-Neto L, Gollany HT, Milori DMBP (2020b). Applications of laser-induced breakdown spectroscopy for soil characterization, part II: Review of elemental analysis and soil classification. *European Journal of Soil Science* 71: 805–818.

- A phase transformation method, by which the toxic elements are extracted, which removes the matrix and then the extracts are dried and analyzed with the LIBS technique;
- Spectral normalization, which aims to reduce the variation in intensity of emission lines caused by both laser fluctuation and matrix effects;
- Multivariate models, which employ multiple linear and non-linear relationships between the intensities of the emission lines from various elements (including those in the matrix) and concentrations of the toxic elements of interest;
- Soil classification followed by regression modeling, where a regression model is developed for each soil class;
- Combination of various methods such as normalization, normal data transformation, outlier removal, noise removal, and multivariate modeling.

In recent years, significant advancements have been made in LIBS concerning the quantification of toxic elements in soil samples. However, certain challenges remain to be overcome, such as creating models for several (more than 100) soil samples of differing composition. Furthermore, even for qualitative or semi-quantitative LIBS analyses, in-the-field use of portable systems or systems embedded in field vehicles is still emerging.

Soil carbon

Soil organic matter is of great importance in the biogeochemical carbon cycle because it is the second-largest carbon pool on Earth (Lal, 2004). When properly managed, carbon that is stored as organic matter can improve physicochemical properties of the soil (soil fertility) and, consequently, crop yields. From an environmental perspective, the sequestration of carbon dioxide (CO₂) by plants, and its storage in soil organic matter (preferably with a high degree of humification), can help to mitigate the greenhouse effect, and contribute to an increase in the farmer's profitability through the payment of carbon credits.

Before they can be traded in the recent and thriving carbon market, carbon credits (1 carbon credit = 1 ton of CO₂ equivalent) must be measured by methods that quantify soil organic carbon stock. With that in mind, LIBS has

advantages compared to conventional methods, such as dry combustion and wet acid oxidation (Walkley–Black), when measuring C content in soils. The simplicity, speed, and possibility of in situ analyses are some of the advantages of this technique, in contrast to conventional methods of C quantification. LIBS analysis enables the direct quantification of total C and indirect quantification of organic and inorganic C (using stoichiometric relationships with other elements in the sample). In LIBS, soil samples are prepared as pellets, which eliminates the requirement of (i) tin capsules (used in the dry combustion method) and (ii) oxidizing acids and toxic reagents such as concentrated sulfuric acid (H_2SO_4) and potassium dichromate ($\text{K}_2\text{Cr}_2\text{O}_7$) (used in the Walkley–Black method). Thus, LIBS enables the measurement of soil C to be done more rapidly, at a more affordable cost, and in a way that aligns with the principles of “green chemistry.” The use of LIBS was recently approved by an international certifier (Verra, 2022) for soil carbon credit verification, bringing the technique one step closer to becoming a standard method for routine soil C analyses.

Despite the above benefits, challenges related to spectral interference and matrix effects still need to be overcome in the quantification of soil C in LIBS. The main C emission lines used in LIBS measurements are the atomic lines at 193.03 and 247.86 nm, which are associated with large emission probabilities (Einstein coefficients for spontaneous emission) and high relative intensities. These two lines are observed in conventional LIBS systems and under ambient air atmospheric conditions, although the first is absorbed partially by atmospheric oxygen (O_2). However, aluminum (Al) and Fe lines can interfere with the atomic C lines at 193.03 and 247.86 nm, respectively, which could make the quantification of C in soil samples difficult, especially in tropical soils, e.g., Oxisols. Super high-resolution spectrometers (i.e., resolution of less than 0.01 nm) could be used to overcome the spectral interference observed in the atomic C lines or methods that subtract the Al or Fe contribution from atomic C lines (e.g., Nicolodelli et al., 2014).

A viable strategy to overcome matrix effects is to develop a calibration model for each soil type, e.g., considering textural classes (Segnini et al., 2014). Two calibration models, one for medium-textured soil samples, and another for fine-textured soil samples, were successfully created for total soil C. However, different LODs were obtained for each model because of the texture-related matrix effects. The LOD for the fine-textured soils (0.13% C) was 59% less than that obtained for medium-textured soils (0.32% C).

After these challenges are overcome, LIBS has a great potential for enabling the measurement of C stocks in soils and, consequently, the C credit market. However, the calculation of C stocks in soils also requires soil bulk density, a physical soil property that is labor-intensive and time-consuming to measure. To enable large-scale measurement of soil C stocks, a fast technique other than LIBS that can estimate soil bulk density is required. The next section presents applications of the LIBS technique to assess several soil physicochemical properties, such as texture and pH.

Soil physicochemical properties

In addition to measuring elements, LIBS may also be used to evaluate soil characteristics such as pH, texture (percentages of sand, clay, and silt), cation exchange capacity, base saturation, and the degree of humification of soil organic matter. These properties have been effectively assessed using multielement data from the LIBS emission spectrum, stoichiometric ratios, and multivariate regression models (Table 3).

The soil physical or chemical properties can only be inferred by LIBS from the stoichiometric relationships among the constituent elements of the soil sample, except for total measurements of the elements content (Villas-Boas et al., 2020a). An example is the measurement of soil pH by LIBS (Table 2). This property, which denotes the acid–base balance in soil, can be estimated on the basis of the concentration of elements related to soil acidity or alkalinity (e.g., Al and Ca, respectively). While the former predominates in acidic soils, Ca is associated with soil alkalinity: soil liming is a regular practice to correct soil acidity. The pH of a variety of Brazilian soils was estimated, with an error of 0.4 pH units, from the emission lines of

Table 3 Models (partial least square regression—PLSR, multiple linear regression—MLR, and K-nearest neighbors—KNN), elements or spectral ranges used, and predictive errors—either root mean squared errors (RMSE) or root average errors (RAE)—for assessing soil properties using LIBS.

Soil property	Model	Elements or spectral ranges (nm)	Predictive errors
pH	PLSR ^a	Al, Ca, H, and O ^a	RMSE = 0.40 ^a
	PLSR ^b	220.84–222.21, 237.82–239.27, and 391.78–394.22 ^b	RMSE = 0.33 ^b
Cation exchange capacity	MLR ^b	C, Na, Mg, Al, Si, P, Ca, Ti, Mn, Fe, Cu, and Zn ^b	RMSE = 10.55 mmol _c dm ⁻³ ^b
	PLSR ^c	183–908 ^c	RMSE = 5.4 cmol _c kg ⁻¹ ^c
Texture	PLSR ^a	188–908 or Si, Na, Fe, Ti, Ca, K, Al, Co, Mg, V, Ba, and Be ^a	RMSE = 7% (clay), 8% (sand), 4% (silt) ^a
	PLSR ^b	220.84–222.21, 246.77–248.29, and 251.37–252.94 ^b	RMSE = 27.68 g dm ⁻³ (clay) ^b
Base saturation	PLSR ^b	201.31–272.37, 279.18–300.60 and 396.59–399.10 ^b	RMSE = 5.49% ^b
Humification degree of soil organic matter	K-NN ^a	C, Al, Mg, and Ca ^a	RAE = 33.5% ^a

^aVillas-Boas PR et al. 2020, *European Journal of Soil Science* 71:789–804.

^bTavares TR et al. 2022 *Soil and Tillage Research* 216: 105250.

^cPelagio et al., 2020, *Chemistry Select* 5:3798–3804.

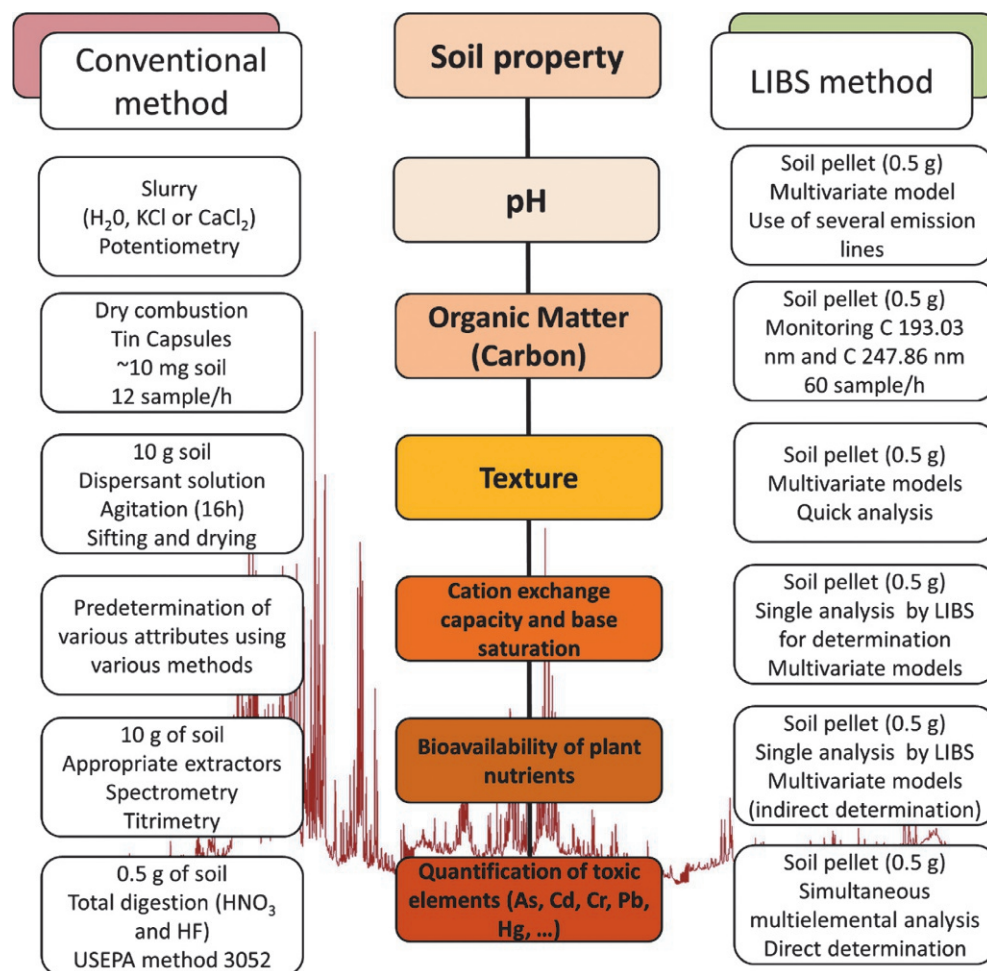


Fig. 3 Comparison of conventional methods and LIBS with respect to their advantages and limitations for the analysis of elements and main soil properties.

these two elements and those of hydrogen (H) and O (Ferreira et al., 2015). Thus, to assess any soil property requires multivariate regression models based on the linear and non-linear stoichiometric relationships among the constituent elements in the soil sample.

The simplicity of measurement, speed of analysis, and minimal sample preparation are the key LIBS advantages for these applications compared to conventional methods (Fig. 3). For instance, while traditional methods take at least 2 days to determine soil texture, LIBS analysis requires only a few minutes per sample (Villas-Boas et al., 2016). However, matrix effects can compromise the accuracy of LIBS estimates, in particular when the samples undergoing analysis have a composition that differs significantly from that of the calibration set. To ensure that LIBS is accurate and reliable, it is recommended that a spectral library should be maintained, with the frequent addition of new sample information.

Conclusion

Large-scale agricultural production is necessary to meet the rising global demand for food, which calls for the effective use of agricultural inputs to increase productivity and to mitigate environmental impacts. However, conventional methods of soil analysis are unable to meet these demands because they are costly, time-consuming, require chemical inputs, and produce residues. LIBS presents unique features (rapid analyses, minimal sample preparation, and no generation of residues) that can allow for large-scale soil analyses, as required in precision agriculture. This chapter discussed the key advantages and limitations of LIBS as a method to assess, in soil samples, (i) the total and bioavailable contents of plant nutrients; (ii) the total content of toxic elements; (iii) the total, organic and inorganic C content; and (iv) the physicochemical properties. Despite the limitations, LIBS has great potential to become a standard method for routine soil analyses and to meet the demands of large-scale agricultural production. The use of LIBS

has already been approved by Verra, an international certifier, for soil carbon credit verification, allowing companies to quantify soil C for the carbon market.

Acknowledgments

This work was supported by CNPq (grants 312775/2021-0, 40226/2021-0, 1133163/2021-0, 434380/2018-0, and 312647/2018-2) and FAPESP (grant 2013/07276-1, 2022/05451-0 and 2017/19248-3).

The mention of commercial products and organizations in this manuscript is solely to provide specific information. It does not constitute an endorsement by USDA-ARS over other products and organizations not mentioned.

References

- Bernardi ACC, Bettiol GM, Mazzuco GG, Esteves SN, Oliveira PPA, and Pezzopane JRM (2017) Spatial variability of soil fertility in an integrated crop livestock forest system. *Advances in Animal Biosciences* 8: 590–593. <https://doi.org/10.1017/S2040470017001145>.
- Borduchi LCL, Milori DMBP, Meyer MC, and Villas-Boas PR (2022) Reducing matrix effects on the quantification of Ca, Mg, and Fe in soybean leaf samples using calibration-free LIBS and one-point calibration. *Spectrochimica Acta Part B: Atomic Spectroscopy* 198: 106561. <https://doi.org/10.1016/j.sab.2022.106561>.
- Cremers DA and Radziemski LJ (2013) *Handbook of Laser-Induced Breakdown Spectroscopy*. Oxford, UK: John Wiley & Sons Ltd. <https://doi.org/10.1002/9781118567371>.
- de Giacomo A, Dell'Aglio M, Gaudioso R, Koral C, and Valenza G (2016) Perspective on the use of nanoparticles to improve LIBS analytical performance: Nanoparticle enhanced laser induced breakdown spectroscopy (NELIBS). *Journal of Analytical Atomic Spectrometry* 31: 1566–1573. <https://doi.org/10.1039/C6JA00189K>.
- Ferreira EC, Gomes Neto JA, Milori DMBP, Ferreira EJ, and Anzano JM (2015) Laser-induced breakdown spectroscopy: Extending its application to soil pH measurements. *Spectrochimica Acta Part B: Atomic Spectroscopy* 110: 96–99. <https://doi.org/10.1016/j.sab.2015.06.002>.
- Finnish Ministry of the Environment (2007) *Finnish Government Decree on the Assessment of Soil Contamination and Remediation Needs*.
- Lal R (2004) Soil carbon sequestration impacts on global climate change and food security. *Science* 304: 1623–1627. <https://doi.org/10.1126/science.1097396>.
- Larcher W (2003) *Physiological Plant Ecology: Ecophysiology and Stress Physiology of Functional Groups*, 4th Edn. Berlin: Springer.
- Nicolodelli G, Marangoni BS, Cabral JS, Villas-Boas PR, Senesi GS, dos Santos CH, Romano RA, Segnini A, Lucas Y, Montes CR, and Milori DMBP (2014) Quantification of total carbon in soil using laser-induced breakdown spectroscopy: A method to correct interference lines. *Applied Optics* 53: 2170–2176. <https://doi.org/10.1364/AO.53.002170>.
- Segnini A, Xavier AAP, Otaviani-junior PL, Ferreira EC, Watanabe AM, Sperança MA, Nicolodelli G, Villas-Boas PR, Oliveira PPA, and Milori DMBP (2014) Physical and chemical matrix effects in soil carbon quantification using laser-induced breakdown spectroscopy. *American Journal of Analytical Chemistry* 05: 722–729. <https://doi.org/10.4236/ajac.2014.511080>.
- Takahashi T and Thornton B (2017) Quantitative methods for compensation of matrix effects and self-absorption in laser induced breakdown spectroscopy signals of solids. *Spectrochimica Acta Part B: Atomic Spectroscopy* 138: 31–42. <https://doi.org/10.1016/J.SAB.2017.09.010>.
- Tavares TR, Mouazen AM, Nunes LC, dos Santos FR, Melquiades FL, da Silva TR, Krug FJ, and Molin JP (2022) Laser-Induced Breakdown Spectroscopy (LIBS) for tropical soil fertility analysis. *Soil and Tillage Research* 216: 105250. <https://doi.org/10.1016/j.still.2021.105250>.
- Verra (2022) Verified Carbon Standard Program [WWW Document]. <https://verra.org/project/vcs-program/>.
- Villas-Boas PR, Romano RA, de Menezes Franco MA, Ferreira EC, Ferreira EJ, Crestana S, and Milori DMBP (2016) Laser-induced breakdown spectroscopy to determine soil texture: A fast analytical technique. *Geoderma* 263: 195–202. <https://doi.org/10.1016/j.geoderma.2015.09.018>.
- Villas-Boas PR, Franco MA, Martin-Neto L, Gollany HT, and Milori DMBP (2020a) Applications of laser-induced breakdown spectroscopy for soil analysis, part I: Review of fundamentals and chemical and physical properties. *European Journal of Soil Science* 71: 789–804. <https://doi.org/10.1111/ejss.12888>.
- Villas-Boas PR, Franco MA, Martin-Neto L, Gollany HT, and Milori DMBP (2020b) Applications of laser-induced breakdown spectroscopy for soil characterization, part II: Review of elemental analysis and soil classification. *European Journal of Soil Science* 71: 805–818. <https://doi.org/10.1111/ejss.12889>.
- Zorov NB, Popov AM, Zaytsev SM, and Labutin TA (2015) Qualitative and quantitative analysis of environmental samples by laser-induced breakdown spectrometry. *Russian Chemical Reviews* 84: 1021–1050. <https://doi.org/10.1070/RCR4538>.

Relevant websites

- <https://bsssjournals.onlinelibrary.wiley.com/doi/abs/10.1111/ejss.12888>—Applications of laser-induced breakdown spectroscopy for soil analysis, part I: Review of fundamentals and chemical and physical properties
- <https://bsssjournals.onlinelibrary.wiley.com/doi/10.1111/ejss.12889>—Applications of laser-induced breakdown spectroscopy for soil characterization, part II: Review of elemental analysis and soil classification

Calorimetric analysis

Xingmin Rong and Qiaoyun Huang, College of Resources and Environment, State Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University, Wuhan, China

© 2023 Elsevier Ltd. All rights reserved.

Introduction	662
Isothermal microcalorimetry (IMC)	663
Determination of soil microbial activity and biomass	663
Monitoring the toxicity and decomposition of soil pollutants	665
Isothermal titration calorimetry (ITC)	665
Differential scanning calorimetry	667
Conclusions	668
References	668

Abstract

Calorimetry is the measurement of heat. It has been used since 1780 in Lavoisier and Laplace's ice calorimeter. In environmental soil science, isothermal microcalorimetry, isothermal titration calorimetry, and differential scanning calorimetry provide thermodynamic and kinetic information on various reactions and processes; they are powerful tools to elucidate almost all important topics. There has been an increase in applications of the three typical techniques in determining microbial activity, monitoring the toxicity and biodegradation of organic pollutants, evaluating the risk of metals and metalloids, and investigating the interfacial processes in soils and associated environments. Coupled with other analytical techniques, calorimetry could be promising for research in soil and environmental sciences.

Key points

- Introduction of calorimetric techniques.
- Application of calorimetry in soil science and related fields.
- Recent research advances in calorimetry.

Introduction

In nature, all the physical, chemical, and biological processes are associated with the effect of heat, which is the basis for thermal analysis and calorimetry. According to the International Confederation for Thermal Analysis and Calorimetry, thermal analysis studies the relationship between a sample property and its temperature as the sample is heated or cooled in a controlled manner. There is an overlap in the meaning of the term 'thermal analysis' and 'calorimetry'. In some books or journal papers, they are used indiscriminately or combined. In practice, thermal analysis refers to a group of techniques to study the property of a sample in a controlled temperature program. It is a technique that exists for properties or physical quantities that are measured against temperature. Here the property or physical quantity could be heat, temperature, heat flow, mass, dimension, mechanical, electrical, magnetic, acoustic, and structural characteristics of the sample concerned. Calorimetry measures heat released into or absorbed from the environment during a chemical, physical or biological reaction. Calorimetry focuses on measuring heat changes during a process under particular controlled conditions, such as constant volume or constant temperature. A typical experiment for calorimetric analysis may be designed to measure the heat capacity, enthalpy, entropy, and free energy with high precision and accuracy over a wide range of temperatures controlled by a programmed heating and cooling procedure. Precision and accuracy are the keys for analytical techniques, and also for calorimetric analysis. With the invention of the thermocouple and development of modern electronics, in particular information technology, the variation in temperature and detectable heat effect of a calorimeter could be very small, i.e., the temperature probe can measure to less than one ten-thousandth change of a degree Celsius and maintain the baseline stability at tens of micro-degree Celsius over hours.

Calorimetry is one of the oldest areas of physical chemistry. Lavoisier and Laplace were the pioneers for heat measurement and made the first calorimeter equipment in the world. The application of calorimetry in the environmental area can be traced back to Le Chatelier who determined and compared the dehydration point of several clay minerals (Plante et al., 2009). Since then, thermal analysis and calorimetry have been applied intensively in mineralogy and chemistry. Calorimetry encompasses various processes, including titration, flow, reaction, and sorption (Gaisford et al., 2016). Calorimeters can be classified based on a variety of criteria. It may not be easy to give a consistent classification and provide a clear definition of a group of calorimeters. Sarge et al. (2014) classified calorimeters into three groups: (1) isothermal calorimeters; (2) calorimeters with controlled heat exchange between the

sample and surroundings; and (3) adiabatic calorimeters. Each group can be further classified into subgroups according to practical criteria.

The first commercial calorimeter was produced in the 1950s. There are now tens of calorimetric instruments available. The most recent calorimeter is equipped with a computer to control experimental conditions and automatically collect the experimental data. The technique for measuring tiny quantities of heat and heat power is called microcalorimetry, using a microcalorimeter instrument. Beware that 'micro-' means that the detection limit of the device for heat or heat power can be microjoule or microwatt. Calorimetry has been applied in soil and environmental sciences for a long time. It focuses on the characterization of soil clay minerals, soil organic matter, soil bacterial growth, degradation of organic contaminants, and the interfacial reactions between the solid environmental components and pollutants. The applications of several most commonly used calorimetric techniques, including isothermal microcalorimetry (IMC), isothermal titration calorimetry (ITC), differential scanning calorimetry (DSC) in soil and environmental sciences (Fig. 1), will be introduced in the following sections.

Isothermal microcalorimetry (IMC)

The term 'isothermal microcalorimeter' refers to susceptible calorimetric techniques used in experiments where the temperature remains constant. The sensitivity for such instruments is typically in the range of one microwatt or even better (Wadsö, 2001). The isothermal microcalorimeter is usually designed to work in a static incubation mode. The heat released from the sample ampoules can be real-time monitoring and dynamic analysis of chemical, physical and biological processes for hours, days, even months. Isothermal microcalorimetry is a nonspecific technology. In principle, it can be applied to all types of reactions or processes accompanied by thermal effects. This technique has been employed over a wide range of studies involving life science, material science, and pharmaceutical development over the past years (Yang et al., 2005).

Soil is a complex ecosystem in which many physical, chemical, and biochemical reactions occur. All these reactions are associated with more or less heat release, which can be monitored conveniently by microcalorimetry. Isothermal microcalorimetry is a powerful tool for investigating these processes because their thermal effects depend only on the system's initial and final energy states regardless of the types of reactions and stages of the processes.

Determination of soil microbial activity and biomass

Microorganisms are the most active organic components in soils. The interactions between microorganisms and other soil constituents profoundly affect the physicochemical and biological properties of the soil ecosystem. Microorganisms are also responsible for various biochemical reactions during the evolution and formation of soils. These reactions and processes that occur as a result of microbial activity generate a flow of heat which is easily monitored with a microcalorimeter by putting a small amount of soil into the incubation ampoules. The bacterial heat release of the soil versus growth time collected by microcalorimeter interestingly shows a pattern similar to the bacterial growth curve, including a latency phase, a logarithmic phase, a steady phase, and a dead phase (Barros and Feijóo, 2003). The growth rate constant and the metabolic enthalpy can also be determined by calculating the heat evolved due to microbial metabolism in the soil (Núñez-Regueira et al., 2006). By changing the incubation condition in the ampoules, such as moisture, organic matter content, plant nutrient, and the incubation temperature, much information related to soil microbial activity and soil health can be acquired. It has been demonstrated that both total soil heat

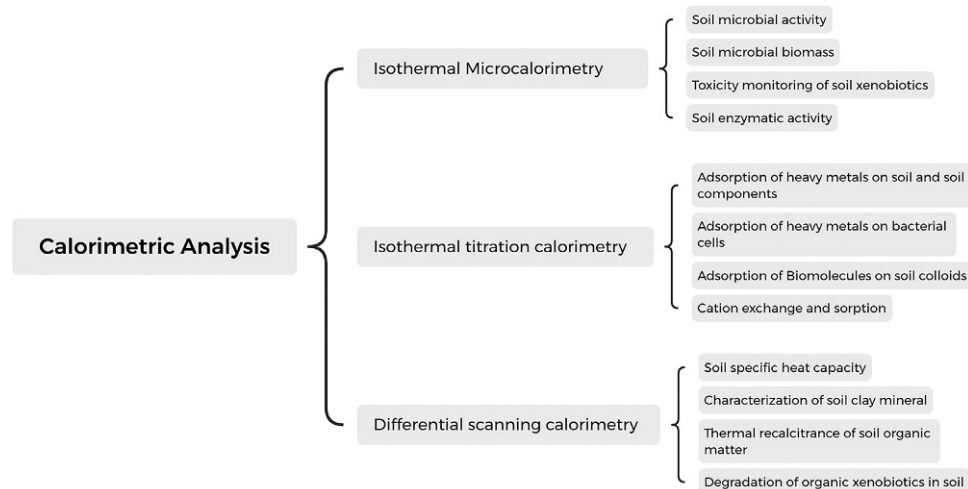


Fig. 1 Applications of calorimetric analysis in soil and environmental science.

production and the microbial growth rate constant are strongly correlated with soil moisture, and soils with a larger organic matter content generated more heat (Barros et al. 1997).

The metabolic enthalpy change (ΔH_{met}) of soil microorganisms can be calculated using the microcalorimetric method. The ΔH_{met} reflected the competition among the microorganisms, providing valuable information on the interactions among microbial populations. Microcalorimetry is a highly sensitive method to assess the overall soil microbial biomass. The soil microbial biomass C determined by microcalorimetry was strongly correlated with that measured by O_2 consumption (Raubuch and Beese, 1999). The data on microbial biomass C obtained using the chloroform fumigation method did not reflect the interactive effect on active soil microbiota. However, the data from microcalorimetry provided information not only on soil microbial activity but also on the competition among microbial populations and the inhibitory effect of metabolites (Barros et al., 2000). Traditional techniques, such as fumigation determine the biomass, corroborating the validity of the microcalorimetric method in assessing soil microbial activity and biomass. Ahamadou et al. (2009) proposed that the peak height (P_{max}) and peak time (t_{max}) obtained from the thermograms could be used as good indices of soil microbial activity. In addition, these calorimetric parameters, P_{max} , t_{max} , total heat evolution (Q), quantitatively reflect the influence of organic matter and water gradients on microbial activity (Hassan et al., 2014).

Bacteria and minerals are the essential components of soil. However, the interactions between bacteria and minerals have received scant attention. By monitoring the thermograms of bacterial growth by IMC, Rong et al. (2007) reported that kaolinite, montmorillonite, and goethite stimulated the exponential growth of *Bacillus thuringiensis*, whereas the sporulation growth of *B. thuringiensis* was somewhat inhibited by kaolinite and montmorillonite (Fig. 2). Goethite significantly depressed the sporulation and the total metabolic activity of the bacterium. Wu et al. (2014) also examined the effects of soil colloids and minerals on the metabolic activity of *Pseudomonas putida* by microcalorimetry. They found that montmorillonite markedly improved the metabolic activity of *P. putida*, whereas kaolinite, goethite, and soil colloids significantly inhibited the activity. In calorimetric monitoring of

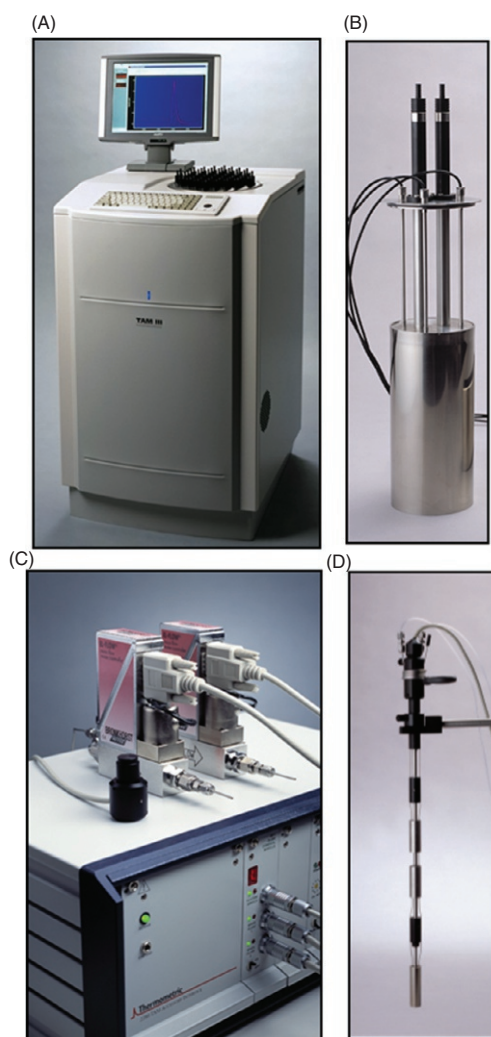


Fig. 2 Isothermal titration calorimeter (A: TAMIII instrument, B: titration unit; C: electrical system; D: single titration channel).

the biodegradation of methyl parathion in the presence of minerals, Zhao et al. (2014) found that kaolinite enhanced the biodegradation of parathion by stimulating bacterial activity. The calorimetric method provides different and complementary information to that from other techniques in characterizing microbial activities.

Monitoring the toxicity and decomposition of soil pollutants

Understanding the behavior and fate of anthropogenic xenobiotics released into the environment always attracts the attention of environmental scientists. The toxicity and rate of biodegradation of xenobiotics can be readily assessed using microcalorimetry. The microbial activity may be disturbed by xenobiotics and shows a unique heat release signature. For example, the growth of *Tetrahymena pyriformis* was progressively inhibited as the concentration of diuron increased, revealing that the state of the living system was severely altered at a concentration of $56.0 \mu\text{g mL}^{-1}$ (Bricheux et al., 2013). The IC_{50} (half-maximal inhibition concentration) was also estimated at $13.8 \mu\text{g mL}^{-1}$ by microcalorimetry. Calorimetry is a noninvasive technique. The toxicity estimates can be derived directly from the physical response of a living system. Sachs et al. (2017) demonstrated the concentration-dependent influence of U(VI) on the cell metabolism of *Brassica napus*, a plant known to be able to accumulate heavy metals. These studies showed that microcalorimetry is an effective tool for determining the growth rate constant and for rapidly detecting metabolic perturbations.

In a study on the decomposition of pesticide 2,4-dichlorophenoxyacetic acid (2,4-D) using the microcalorimetry method, microbial activity was inhibited by heavy loads of 2,4-D in soils (Prado and Airolidi, 2002). The influence of diuron on soil microbial activity indicated that the original thermal effect gradually decreased with increasing amounts of diuron in soil from 0 to $333.3 \mu\text{g g}^{-1}$. Soil microbial activity decreased by 87.4% in the presence of $10.00 \mu\text{g g}^{-1}$ free picloram. In contrast, a decrease of 54.6% was detected for the same amount of immobilized picloram, suggesting that free picloram was more toxic than immobilized picloram. Likewise, Critter and Airolidi (2001) observed significant influences of different pesticides on soil microbial activities by microcalorimetry. They found that increasing amounts of these pesticides in the soil decreased the exothermal effect from 2234 to 1987 kJ mol^{-1} for paraquat, 1670 to 1306 kJ mol^{-1} for diquat, and to a greater extent from 2239 to 589 kJ mol^{-1} for phosphamidon. Huang et al. (2009) found that the catalytic enthalpy of free phosphatase was greater than that of immobilized phosphatase on montmorillonite and soil colloids, indicating that enzymatic activities were inhibited by immobilization.

Heavy metals and metalloids are hazardous to the environment and human health. Based on heat evolution from intact cells, McGuinness and Barisas (1991) developed a bioassay for the acute toxicity of heavy metals including Hg (mercury), Cd (cadmium), Cu (copper), Pb (lead), Cr (chromium), V (vanadium), and As (arsenic). The EC_{50} of heavy metals obtained from heat production of *Photobacterium phosphoreum* agreed with the results obtained using chemiluminescent assay. The reduction in heat output from cells reflects the inhibition of cell metabolic processes by toxicants, which is a better predictor of biotoxicity than chemiluminescence because that heat production is directly related to the overall metabolism. The influences of metals or metalloids on the heat effect of microbial metabolism depend on the types and concentrations of metals or metalloids. Low concentrations of heavy metals normally stimulate the metabolic heat production of *Rhizopus nigricans*, whereas large concentrations of heavy metals depress heat production. On the basis of the microcalorimetric study, the toxicity of four metals on *R. nigricans* was of the order of $\text{Cd}^{2+} > \text{Hg}^{2+} > \text{Pb}^{2+} > \text{Cu}^{2+}$, and the IC_{50} of these heavy metals was 0.8, 1.7, 48.0, and $110 \mu\text{g mL}^{-1}$, respectively. The heat generation of *T. thermophila* was stimulated by $0.2\text{--}0.4 \text{ mg L}^{-1} \text{ Cu}^{2+}$ or $5\text{--}10 \text{ mg L}^{-1} \text{ Cd}^{2+}$. Toxicity of the two-metal system was greater than that of the one-metal system. For example, the growth of *Bacillus thuringiensis* was enhanced by low concentrations of Cu^{2+} ($5\text{--}30 \text{ mg mL}^{-1}$), possibly because Cu^{2+} ions ensured that the cells had an appropriate flow quality and enabled the synthesis of DNA and RNA. The heat output was suppressed entirely by 20 nmol mg^{-1} of Hg^{2+} due to its inhibitory effect on the phosphorylation of the mitochondria (Zhu et al., 2006). The lag phase in the power-time curve of the sulfate-reducing bacteria, *Desulfomicrobium norvegicum* and *Desulfovibrio vulgaris*, was prolonged in the presence of chromium (VI) (Chardin et al., 2002). The greater the concentration of chromium, the longer was the lag phase. Microcalorimetric study on bacterial metabolism affected by heavy metals has been regarded an appropriate method for screening chromium-resistant bacterial strains, which is of help in the bioremediation of chromium-polluted soils. Airolidi and Critter (1996) showed that the thermal effect of Latosol decreased after the addition of copper sulfate. Heat production reached zero when the amount of copper sulfate increased to 6.19 mg in 1.5 g soil, suggesting that CuSO_4 significantly inhibited the activity of soil microorganisms. A kinetic equation describing the biodegradation mechanism of glucose influenced by CuSO_4 was then fitted based on the calorimetric data. The quantitative thermal information of soils could be used to understand better the influence of heavy metals on various environmental ecosystems.

Isothermal titration calorimetry (ITC)

The isothermal microcalorimeter designed to work in titration mode is called the isothermal titration calorimeter (Fig. 2), in which one of the reaction components (titrant) is injected/titrated into a vessel filled with the other components (titrand) maintained at a constant temperature. In an isothermal titration calorimeter, the addition of titrant into the reaction vessel can be done in one single titration or be added in stepwise increments. A stirrer controlled by a computer ensures thorough mixing. The heat generated or consumed in the reaction vessel is compensated for by the Peltier pile. Both reaction kinetics and thermodynamic parameters, such as enthalpy change (ΔH), Gibbs free energy change (ΔG), entropy change (ΔE), and reaction stoichiometry, can be easily obtained from the power-time curves of the ITC experiment. Because titration microcalorimetry provides information on binding enthalpy

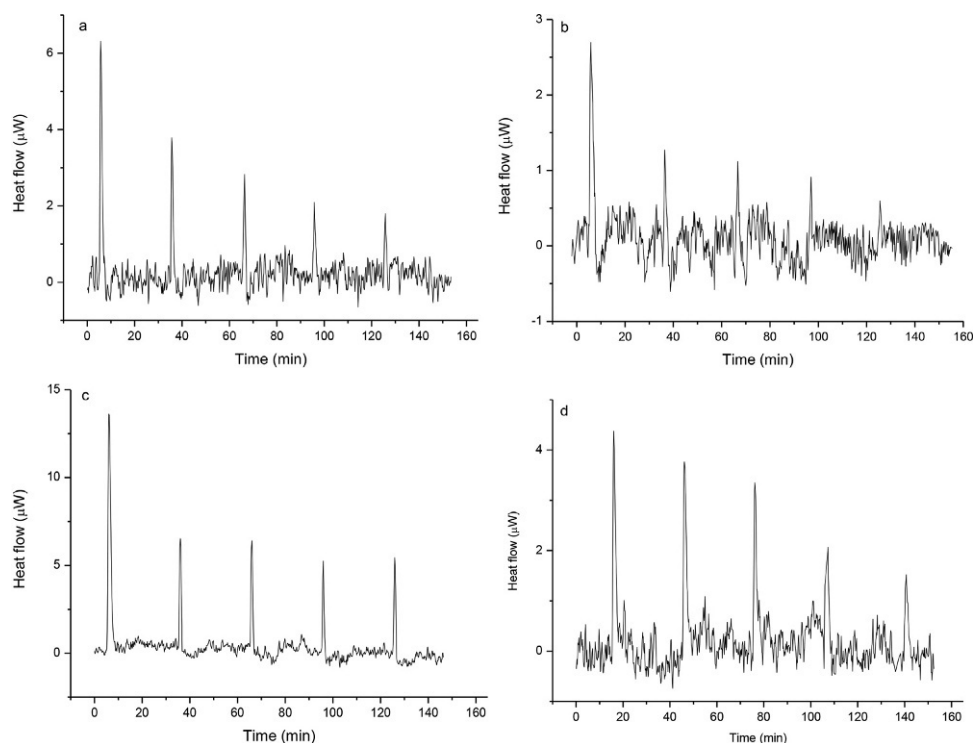


Fig. 3 The power–time curves for the titration of bacterial suspension into clay mineral systems (A, kaolinite, 25 °C; B, montmorillonite, 25 °C; C, kaolinite, 40 °C; D, montmorillonite, 40 °C) (Rong et al. 2007).

change and binding constant, it has been used in studies on specific fast reactions such as enzyme kinetics and the ligand-binding process, which requires minutes or hours to proceed to completion (Hong et al., 2021) (Fig. 3).

The ITC has been used to measure surface reactions in the soil, such as adsorption of heavy metals, cation exchange, phosphate sorption, biomolecules adsorption on soil particles, and mineral–microorganism interfacial reactions. For instance, Cd^{2+} adsorption on goethite was an exothermic process, and the heat release in adsorption was correlated with the equilibrium concentration of Cd^{2+} . The specific adsorption of Pb was predominantly driven by a change in entropy since there was no detectable heat associated with this reaction in a Ultisol and an Oxisol (Appel et al. 2002). The adsorption of Ca and Pb, when replacing exchangeable K, was endothermic. The heat exchanged was less for Pb than for Ca in both soils, indicating that more energy was required for Ca^{2+} than for Pb^{2+} to replace K^+ from surfaces of soil particles. In the Oxisol, there was no change in the amount of heat needed to replace the remaining K^+ as divalent cation saturation increased, suggesting that the exchange sites were energetically similar in this soil. However, a more significant amount of energy was required to remove K^+ as the amount of exchangeable K decreased in the Ultisol. Rhue et al. (2002) found that Na^+ ions experienced considerable difficulty in replacing exchangeable K^+ ions from the Ultisol containing an expanding 2:1 layer silicate clay. The reaction of phosphate with the Ultisol and amorphous $\text{Al}(\text{OH})_3$ was exothermic. Precipitation of Al-phosphate was shown to be endothermic, confirming that precipitation was not the primary mechanism for phosphate sorption. Cai et al. (2006) reported that the adsorption enthalpies of DNA on permanent-charge soil colloids were higher than those on variable-charge soil colloids, and DNA adsorption on organic clays was endothermic, whereas that on inorganic clays was exothermic. Similar results also showed that DNA adsorption on pure montmorillonite was an endothermic reaction and that on hydroxyl aluminum-coated montmorillonite it was exothermic (Cai et al., 2008). The ITC is a convenient method to study adsorption in a complex system, such as the bacteria–mineral composites in soil. The titration of Cu^{2+} solution into bacterial, mineral, and bacteria–mineral composite suspensions and found that Cu^{2+} adsorption on bacteria releases more heat than those of mineral and bacteria–mineral composites. In the montmorillonite- and goethite–bacteria composite systems, ITC verified that the adsorption of Cd, Cu and Pb did not conform to component additivity, and site masking models were proposed to predict the binding and distribution of heavy metals on multi-component soil interfaces, especially at large concentrations of metal ions and large contents of microbial biomass (Qu et al. 2018, Fig. 4).

The adhesion between bacterial cells and soil minerals can also be monitored by ITC. Hong et al. (2011) presented some earlier research about the adsorption of *Bacillus subtilis* on three clay minerals which was an exothermic processes, with enthalpy changes ranging from -52 to 141 kJ kg^{-1} . By further analysis of the dependence of adsorption enthalpy on temperature, pH, and salt concentration, they concluded that electrostatic attraction might play a critical role in bacterial adsorption on clay minerals. Chen et al. (2016) revealed that the entropy increases of excluded water rearrangement and dehydration play key roles in the interaction between DNA and *P. putida*-montmorillonite/goethite composites. Recently, Hong et al. (2021) reported that the negative change in enthalpy was mainly assigned to the ligand exchange of phosphate adsorption on four soil colloids in the pH range from 5.0 to 8.5.

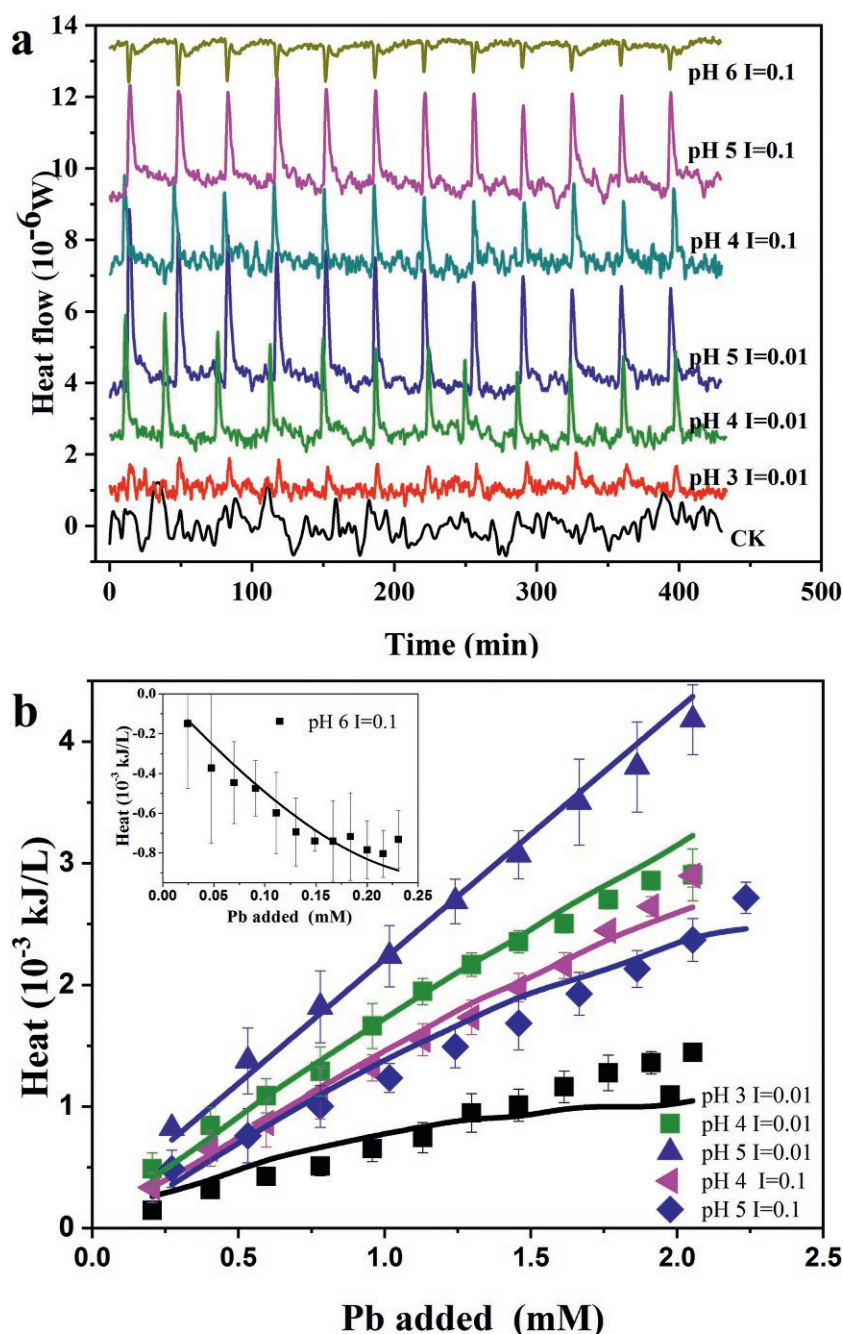


Fig. 4 The power-time curves (A) and the cumulative heat (B) for Pb titration to montmorillonite in the presence of 0.01 or 0.1M NaNO_3 at initial pH 4, 5 and 6. Solid curves represent model prediction (Qu et al. 2018).

The ITC could discriminate between the interactions of proteins and silver nanowires. The negative entropy and negative enthalpy changes of lysozyme adsorption on the silver nanowires might arise from conformational restriction and electrostatic interaction, while the bovine serum albumin and ovalbumin have very weak interaction with the nanowire with ITC detecting minimal adsorption energy.

Differential scanning calorimetry

Differential scanning calorimetry (DSC) measures the energy introduced to a system as a function of temperature. The instrument may work in two modes, heat-flux or power compensation, depending on the method of measurement used. In the heat-flux DSC,

the instrument directly determines the difference in temperature and rate of heat flow between the sample and reference. In a power compensation DSC, the power supplied is varied to compensate for deviations in temperature to establish a near-zero temperature difference between the sample and reference. The DSC is a powerful tool that is used extensively in determining the stability of biopolymers and biological membrane structures in the life sciences, molecular recognition and stability of drugs, protein denaturation, and the order–disorder transition of protein and starch in food science, and the phase transition in material science.

Differential scanning calorimetry has become increasingly popular in the characterization of soil organic matter, clay minerals, and pollutants in soil and the environmental sciences. Wang et al. (2019) systematically evaluated the effects of clay and organic matter content on soil specific heat capacity. They found that organic matter and tightly bound water are the keys for specific heat capacity, and that fine-textured soils possess a larger specific heat capacity than coarse-textured soils. Microplastics are manufactured as microbeads, capsules, fibers, or pellets with a particle size of less than 5 mm. They are widely distributed in all environmental compartments and have attracted increasing attention worldwide. The branching, impurity, additives, and sample size of microplastic polymer affected its phase transition temperatures obtained by DSC. With DSC, Tang et al. (2020) proved that aged plastic lost 12% of its crystallinity compared to that in a pristine state. Differential scanning calorimetry is usually linked with other thermal techniques, for example, thermal gravity analysis (TGA), differential thermal analysis (DTA), and microcalorimetry, or coupled with spectroscopy, microscopy, and chromatography. Dell'Abate et al., (2003) investigated the thermal behavior of humic substances extracted from temperate and tropical soils by DSC/TG coupled with mass spectrometry. The thermal recalcitrance and decomposability of the organic materials were valuable indices for the stability of humic substances.

Conclusions

Calorimetric analysis enables the continuous monitoring of samples for a prolonged period. The samples collected from the different environmental compartments can be used for calorimetric analysis as received with no need for pretreatment before loading into a calorimeter, which maintains the sample in its original form rather than destroying its structure. This is a significant advantage of thermal analysis and calorimetry over chromatography, mass spectrometry, and other techniques that need thorough and laborious sample preparation. Soil and environmental scientists have been applying the three typical calorimetric analysis techniques, IMC, ITC, and DSC in their research. These techniques offer not only new insights into the reactions and processes in soils and associated environments, such as the thermodynamic mechanisms involved in the adsorption and desorption of organic molecules, inorganic ions, and pollutants on soil particles, but also facilitate further understanding of the thermodynamics and kinetics of enzymatic reactions in the soil system.

It is expected that thermal analysis and calorimetry will be used more extensively in the field of soil and environmental science in the future. In particular, the combination of calorimetry with other techniques would help to provide an integrated understanding of the various processes related to soil health, soil pollution and remediation, soil ecology, and global change.

References

- Ahamadou B, Huang Q, Chen W, Wen S, Zhang J, Mohamed I, et al. (2009) Microcalorimetric assessment of microbial activity in long-term fertilization experimental soils of Southern China. *FEMS Microbiology Ecology* 70(2): 186–195.
- Airolidi C and Critter SAM (1996) The inhibitor effect of copper sulphate on microbial glucose degradation in red latosol soil. *Thermochimica Acta* 288(1): 73–82.
- Appel C, Rhue D, Ma L, and Reve B (2002) Heats of K/Ca and K/Pb exchange in two tropical soils as measured by flow calorimetry 1. *Soil Science* 167(12): 773–781.
- Barros N and Feijóo S (2003) A combined mass and energy balance to provide bioindicators of soil microbiological quality. *Biophysical Chemistry* 104(3): 561–572.
- Barros N, Feijóo S, and Balsa R (1997) Comparative study of the microbial activity in different soils by the microcalorimetric method. *Thermochimica Acta* 296(1): 53–58.
- Barros N, Feijóo S, Simoni A, Critter SAM, and Airolidi C (2000) Interpretation of the metabolic enthalpy change, ΔH_{met} , calculated for microbial growth reactions in soils. *Journal of Thermal Analysis and Calorimetry* 63(2): 577–588.
- Bricheux G, Bonnet J-L, Bohatier J, Morel J-P, and Morel-Desrosiers N (2013) Microcalorimetry: A powerful and original tool for tracking the toxicity of a xenobiotic on *Tetrahymena pyriformis*. *Ecotoxicology and Environmental Safety* 98: 88–94.
- Cai P, Huang Q, Jiang D, Rong X, and Liang W (2006) Microcalorimetric studies on the adsorption of DNA by soil colloidal particles. *Colloids and Surfaces. B, Biointerfaces* 49(1): 49–54.
- Cai P, Huang Q, Li M, and Liang W (2008) Binding and degradation of DNA on montmorillonite coated by hydroxyl aluminum species. *Colloids and Surfaces. B, Biointerfaces* 62(2): 299–306.
- Chardin B, Gallice P, and Bruschi M (2002) Thermodynamics of cleaning-up of waste waters polluted by chromium. *Journal of Thermal Analysis and Calorimetry* 70(2): 475–482. <https://doi.org/10.1023/a:1021628608379>.
- Chen X, Huang Q, and Chen W (2016) Adsorption of DNA by bacteria and their composites with minerals. *Geomicrobiology Journal* 33(9): 822–831.
- Critter SAM and Airolidi C (2001) Application of calorimetry to microbial biodegradation studies of agrochemicals in Oxisols. *Journal of Environmental Quality* 30(3): 954–959.
- Dell'Abate MT, Benedetti A, and Brookes PC (2003) Hyphenated techniques of thermal analysis for characterisation of soil humic substances. *Journal of Separation Science* 26(5): 433–440.
- Gaisford S, Kett V, and Haines P (2016) *Principles of Thermal Analysis and Calorimetry*. Cambridge, UK: Royal Society of Chemistry.
- Hassan W, Chen W, Cai P, and Huang Q (2014) Estimation of enzymatic, microbial, and chemical properties in Brown soil by microcalorimetry. *Journal of Thermal Analysis and Calorimetry* 116(2): 969–988.
- Hong Z, Rong X, Cai P, Liang W, and Huang Q (2011) Effects of temperature, pH and salt concentrations on the adsorption of *Bacillus subtilis* on soil clay minerals investigated by microcalorimetry. *Geomicrobiology Journal* 28(8): 686–691.
- Hong ZN, Yan J, Jiang J, Li JY, and Xu RK (2021) Direct quantification of sorption thermodynamics of phosphate on four soil colloids through isothermal titration calorimetry. *ACS Earth and Space Chemistry* 5(2): 295–304.

- Huang Q, Zhu J, Qiao X, Cai P, Rong X, Liang W, et al. (2009) Conformation, activity and proteolytic stability of acid phosphatase on clay minerals and soil colloids from an Alfisol. *Colloids and Surfaces. B, Biointerfaces* 74(1): 279–283.
- McGuinness SM and Barisas BG (1991) Acute toxicity measurements on aquatic pollutants using microcalorimetry on tissue-cultured cells. *Environmental Science and Technology* 25(6): 1092–1098.
- Núñez-Regueira L, Rodríguez-Añón JA, Proupín-Castiñeiras J, and Núñez-Fernández O (2006) Microcalorimetric study of changes in the microbial activity in a humic Cambisol after reforestation with eucalyptus in Galicia (NW Spain). *Soil Biology and Biochemistry* 38(1): 115–124.
- Plante AF, Fernández JM, and Leifeld J (2009) Application of thermal analysis techniques in soil science. *Geoderma* 153(1): 1–10.
- Prado AGS and Airoldi C (2002) The toxic effect on soil microbial activity caused by the free or immobilized pesticide diuron. *Thermochimica Acta* 394(1): 155–162.
- Qu C, Huihui D, Ma M, Chen W, Cai P, and Huang Q (2018) Pb sorption on montmorillonite-bacteria composites: A combination study by XAFS, ITC and SCM. *Chemosphere* 200: 427–436.
- Raubuch M and Beese F (1999) Comparison of microbial properties measured by O₂ consumption and microcalorimetry as bioindicators in forest soils. *Soil Biology and Biochemistry* 31(7): 949–956.
- Rhue RD, Appel C, and Kabengi N (2002) Measuring surface chemical properties of soil using flow calorimetry 1. *Soil Science* 167(12): 782–790.
- Rong X, Huang Q, and Chen W (2007) Microcalorimetric investigation on the metabolic activity of *Bacillus thuringiensis* as influenced by kaolinite, montmorillonite and goethite. *Applied Clay Science* 38(1–2): 97–103.
- Sachs S, Geipel G, Bok F, Oertel J, and Fahmy K (2017) Calorimetrically determined U(VI) toxicity in *Brassica napus* correlates with oxidoreductase activity and U(VI) speciation. *Environmental Science & Technology* 51(18): 10843–10849.
- Sarge SM, Höhne GWH, and Hemminger W (2014) *Calorimetry: Fundamentals, Instrumentation and Applications*. Weinheim, Germany: Wiley-VCH.
- Tang S, Lin L, Wang X, Feng A, and Yu A (2020) Pb (II) uptake onto nylon microplastics: Interaction mechanism and adsorption performance. *Journal of Hazardous Materials* 386: 121960.
- Wadsö L (2001) Isothermal Microcalorimetry. Current problems and prospects. *Journal of Thermal Analysis and Calorimetry* 64(1): 75–84.
- Wang Y, Lu Y, Horton R, and Ren T (2019) Specific heat capacity of soil solids: Influences of clay content, organic matter, and tightly bound water. *Soil Science Society of America Journal* 83(4): 1062–1066.
- Wu H, Chen W, Rong X, Cai P, Dai K, and Huang Q (2014) Soil colloids and minerals modulate metabolic activity of *Pseudomonas putida* measured using microcalorimetry. *Geomicrobiology Journal* 31(7): 590–596.
- Yang Y, Zhu J, Liu Y, Shen P, and Qu S (2005) Microcalorimetry is a sensitive method for studying the effect of nucleotide mutation on promoter activity. *Journal of Biochemical and Biophysical Methods* 62(3): 183–189.
- Zhao G, Huang Q, Rong X, Cai P, Liang W, and Dai K (2014) Interfacial interaction between methyl parathion-degrading bacteria and minerals is important in biodegradation. *Biodegradation* 25(1): 1–9.
- Zhu JC, Li CH, Liu Y, Zhang ZH, Hou AX, and Qu SS (2006) A microcalorimetric study of the action of mercuric chloride on the metabolism of mitochondria isolated from *Cyprinus carpio* liver tissue. *Journal of Thermal Analysis and Calorimetry* 83(1): 181–186.

Understanding of soil processes at the microscale—Use of NanoSIMS in soil science

Carsten W Mueller^a, Carmen Hoeschen^b, and Ingrid Koegel-Knabner^{b,c}, ^aDepartment of Geosciences and Natural Resource Management, University of Copenhagen, Copenhagen, Denmark; ^bTechnical University of Munich, TUM School of Life Sciences, Soil Science, Freising-Weihenstephan, Germany; ^cTechnical University of Munich, Institute of Advanced Study, Garching, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	670
The need for techniques that allow the study of soils at the process scale	670
The technical principle of NanoSIMS	671
SIMS	671
NanoSIMS	671
Requirements for samples before and during measurement	672
Understanding of soil systems from single particles to intact aggregates and soil cores	674
Mineral particles and aggregates	674
Interfaces of soils and biota	676
Detection of microbiota in the soil environment	678
Conclusion	679
References	679

Abstract

Nano-scale secondary ion mass spectrometry (NanoSIMS) offers an unprecedented tool to complement classical quantitative measurements (e.g., isotopic ratio, elemental composition) by high resolution imaging of isotope and elemental distributions. It represents a unique technique to follow isotopically labelled compounds at soil microscale hot spots. Intact soil structures can be studied and their individual associated components like microaggregates, iron (hydr)oxides, bacteria, particulate organic matter surfaces or macroaggregate and soil core sections at the relevant process scale with a resolution to ~ 100 nm. The chapter includes the technical basics, a short introduction to sample preparation, and several exemplary studies to illustrate NanoSIMS.

Key points

- Introduction to the principle of secondary ion mass spectrometry.
- Requirements for properties and preparation of soil samples for successful NanoSIMS measurements.
- Need to accompany NanoSIMS measurements by multiple imaging techniques to yield a more complete understanding of microscale soil processes.
- Approaches for the analysis of soil mineral surfaces with a special emphasize on organo-mineral associated organic matter.
- Approaches to study soil processes in intact soil structures, ranging from aggregates to larger intact soil structures like rhizosphere or soil core sections.
- Applications for using NanoSIMS in soil microbiology.

Introduction

The need for techniques that allow the study of soils at the process scale

Soils comprise a highly complex three-dimensional system in which biological, chemical and physical processes are intricately connected. Most of the mechanisms and processes that determine soil functions like carbon storage, fate of pollutants or nutrient cycling take place at nm (nanometer) to μm (micrometer) scales, but have implications at much larger scales. Thus, allocation of plant-derived organic matter (OM) takes place at the interfaces between plant litter, roots, microorganisms and soil, but has implications for large scale organic carbon (OC) storage and greenhouse gas emissions (Hallett et al., 2013). Mineral soil structures and their reactive surfaces range from nm and μm scales such as clay minerals and Fe/Al (hydr)oxides (Patzner et al., 2020) to larger structures of several μm for micro- and macro-aggregates which are hot spots for the occlusion of particulate OM and regulate the persistence of OC in soils (Witzgall et al., 2021). The exploration of very fine structures in soils and the functional association of its microscale entities dates back to the early 20th century and was refined using electron microscopy later in the same century (Mueller et al., 2013). The study of soils at the microscale for long remained rather descriptive and based on deriving functions and processes from specific soil structures. Over the last two decades, the rapid increase in the technical capability and accessibility of spectro-microscopic techniques, now allow structural information to be combined with chemical composition as well as elemental and

isotopic information (Mueller et al., 2013). An interesting tool to unravel soil microscale processes is nano-scale secondary ion mass spectrometer (NanoSIMS) due to its ability to image stable isotopes at an unprecedented spatial resolution. Starting in the 1990s with the development of the technique and the first attempts to use NanoSIMS in soil research in the early 2000s (Herrmann et al., 2007), it is now an approach that is easier to access with the increase in available laboratories and the further development of suitable experimental settings, sample preparation, and modern image classification tools. This chapter summarizes the technical principle of secondary ion mass spectrometry as a prerequisite to understand the specific requirements for sample preparation and the measurement itself. The application of NanoSIMS in soil science will be illustrated starting with the specific use to unravel the microspatial organization of organo-mineral interfaces and microscale soil structures (e.g., microaggregates), followed by plant-soil interactions, and its application in soil microbiology.

The technical principle of NanoSIMS

SIMS

Secondary ion mass spectrometry (SIMS), successfully used in environmental research, is becoming a central tool to resolve spatially the elemental composition of solid soil samples (Lamberti, 2005). An energetic ion beam (hundreds of eV to tenth of keV, e.g., O^- (oxygen), Cs^+ (cesium), Ga^+ (gallium)) is accelerated onto a sample under high vacuum conditions (10^{-8} – 10^{-10} Torr in the analysis chamber) inducing a collision cascade by kinetic energy transfer. This leads to the ejection of secondary species (ions, neutrals, electrons, atoms, clusters) from the sample surface, a process called sputtering. According to the ion dose impacting the sample, SIMS can be static (below the static limit of 10^{13} ions cm^{-2}), with no chemical damage of the sample, or dynamic, when the sample is actively eroded after the primary dose has overcome the static limit (Walker, 2010). The particles escaping the surface in an ionized state, the so-called secondary ions, are accelerated into a mass analyzer where they are differentially recorded according to their mass/charge ratio (Fig. 1). The power of the SIMS technique is the detection of elements as well as stable isotopes with a very high sensitivity in the ppm range. There are three types of mass analyzers commonly used in SIMS: time-of-flight (TOF), quadrupole, and magnetic sector analyzers, each having advantages and disadvantages (Walker, 2010). The TOF instruments are preferred for static SIMS and molecular depth profiling, quadrupole and magnetic sectors are used for dynamic depth profiling, while magnetic sector instruments are widely used in studies that require a high mass resolving power. Although, SIMS is a qualitative technique due to the different ionization potential of elements, semi-quantitative analyzes are possible by comparison to standards with known concentrations or isotopic signatures.

NanoSIMS

In the late 1990s, a new design of a dynamic SIMS instrument revolutionized the research community (Hillion et al., 1994). Using a unique co-axial configuration (see Fig. 1), Cameca Gennevilliers, France developed the so called NanoSIMS 50/50 L instrument capable of performing SIMS analyzes at an unprecedented submicron spatial resolution and high transmission (Mueller et al., 2013 and references within). The instrument is equipped with two exchangeable sources of reactive ions (Cs^+ and O^-), delivering a beam that can be focused to the nanometer scale. Lately, a new RF (radio frequency) O^- plasma source with an improved spatial resolution (~ 50 nm) was developed to replace the previous duoplasmatron (~ 200 nm lateral resolution). This enables the same spatial resolution to be achieved for positively (using O^- source) and negatively (using Cs^+ source) charged secondary ions (see Fig. 1). Presently, up to 7 masses (50 L configuration, 5 for the NanoSIMS 50) can be measured simultaneously at a lateral resolution down to 50 nm in both polarities.

The secondary ions escaping the sample surface are directed into the detection system consisting of a double focusing magnetic sector. This enables the accurate separation of isobars at the same mass number, resulting in a high mass resolving power (MRP) of $>10,000$ provided by the NanoSIMS instrument. A high mass resolution is crucial to measure ion species accurately, especially when experimenting with stable isotope labeling (^{13}C (carbon-13), ^{15}N (nitrogen-15), ^{18}O (oxygen-18), ^{29}Si (silica-29), ^{44}Ca (calcium-44)). In 'imaging mode' secondary ion species can be registered at high lateral resolution using electron multipliers (EMs, 6 movable and 1 fixed) or in the so called 'spot mode' counted using Faraday cups (FCs), with a better precision in the per mill range (no imaging in this case). By scanning the focused primary ion beam with a defined number of steps (pixels, e.g., 256×256), a fixed dwell time for each pixel (e.g., 1 ms $pixel^{-1}$), and an appropriate field of view (e.g., $25 \mu m \times 25 \mu m$) over the sample, the elemental distribution is visualized by the detected secondary ions. In general, these parameters are chosen so that the size of a pixel is smaller than the size of the primary beam to avoid over-measuring. Multiple cycles can be registered to reach better ion statistics by accumulating all counts on every single pixel, for example stacking single image layers to form an accumulated image. There are several software packages for analyzing the acquired data. Some of them are under license, such as L'Image developed by L. R. Nittler, Carnegie Institution of Washington and WinImage developed by Cameca accompanying the NanoSIMS instrument. Other software packages have been developed by users and are open source (Look@NanoSIMS developed by L. Polerecky and the OpenMIMS plugin of the Fiji-ImageJ interface) (Polerecky et al., 2012). According to the needs, the user can choose the most suitable one. In general, all of these software packages enable the first processing steps, consisting of dead time correction of the ion signal measured on the electron multipliers and drift correction of the multiple planes and their accumulation. Furthermore, isotopic/elemental ratios can be calculated.

The unique capabilities of NanoSIMS for image isotopic enrichment at a submicron lateral resolution have resulted in many applications in multidisciplinary research fields. Moreover, apart from applications in fields such as materials science, geosciences,

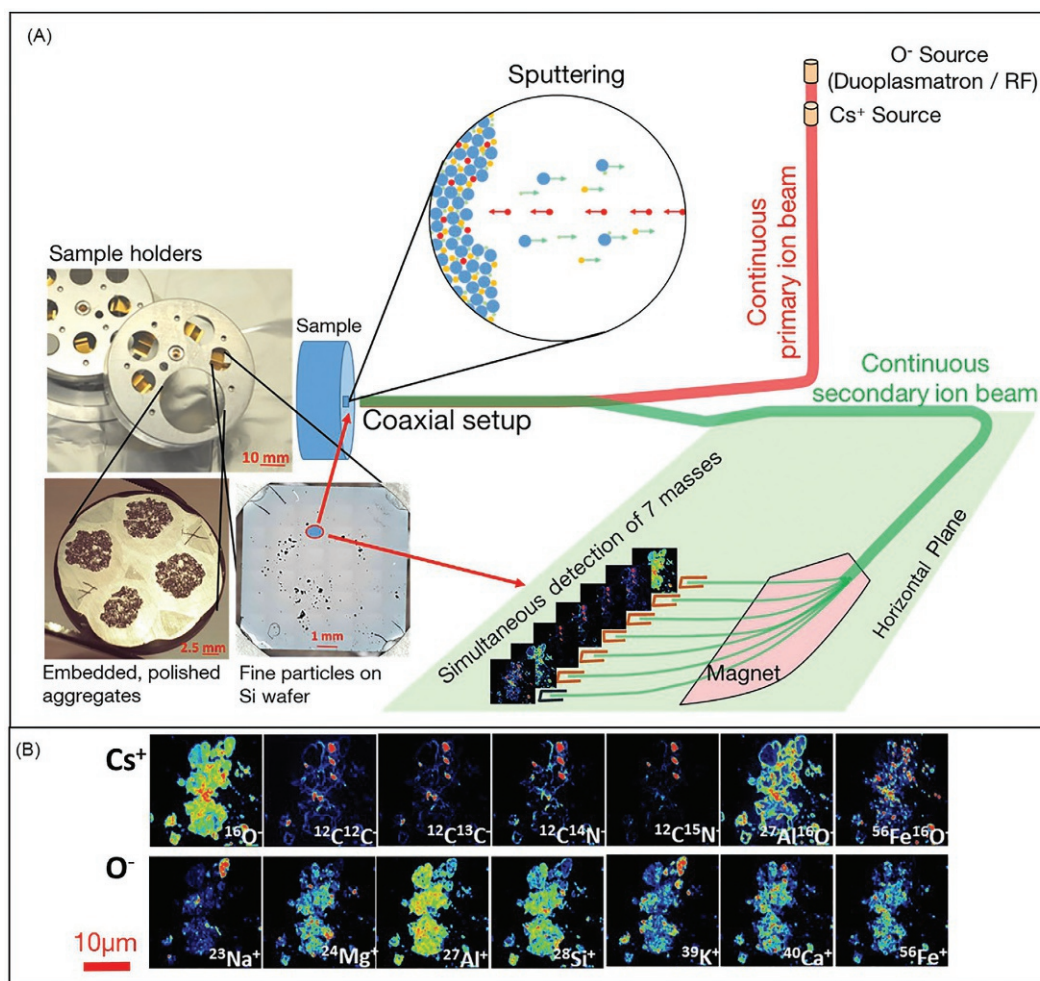


Fig. 1 Schematic diagram illustrating the technical background of the NanoSIMS technique from the sample holder, sample surface and coaxial setup in (A) to the final images in (B). (A) On the left a sample holder and two ways of mounting samples (embedded cross section and fine particles on Si-wafer) are displayed. The close up shows the sputtering process with primary ions (red arrows), and extracted secondary ions (green arrows). The coaxial setup of the NanoSIMS with primary ion beam (in red) and secondary ion beam (in green) in one plane enables the high spatial resolution of the instrument. This setup requires that secondary and primary ions have the opposite charge to enable the extraction of the ions to the spectrometer. In (B) final images are displayed from a fine grain soil sample measured with Cs⁺ and with O⁻ as primary ion, which yields negative and positive secondary ions, respectively.

cosmochemistry, microbiology, or medical science (Hatton et al., 2012; Hoppe et al., 2013; Musat et al., 2016; Pett-Ridge and Weber, 2021), NanoSIMS has revolutionized the field of soil science (Herrmann et al., 2007; Keiluweit et al., 2012; Remusat et al., 2012). It is an important advance in technology that enables us to identify the arrangement of soil constituents and gain insight into their interactions at the sub-micrometer scale, matters which have hitherto been impossible.

Requirements for samples before and during measurement

The successful utilization of SIMS for surface analyzes is determined by a variety of factors. There are universal factors which apply to all SIMS instruments and concern the sample characteristics: only samples that are stable under high vacuum conditions can be measured. Furthermore, as little as possible topographical unevenness is recommended as it strongly influences the quality of the results. It is preferable to have conductive samples, but samples can be measured after coating with a thin conductive layer (e.g., carbon, gold), which does not interfere with the species of interest. Additionally, some systems allow the measurement of isolated samples via charge compensation, e.g., by flooding the sample with electrons. It is worth noting that adjustment of the neutralization system is always challenging.

The application of isotopic labels is a widely used approach to unravel allocation and transformation of elements and organic compounds in soil microenvironments. NanoSIMS offers a wide range of detectable isotopes, therefore, one is not restricted to the most commonly used ¹³C and ¹⁵N to trace organic matter transformation or microbial activity, but may also use ¹⁸O for instance to

trace soil phosphorus as $^{31}\text{P}^{18}\text{O}$ or ^{57}Fe to study iron oxide formation. Thus, NanoSIMS can support classical quantitative isotopic ratio mass spectrometer data with high resolution imaging of the reactive sites and interfaces involved in the targeted processes.

It is important that the investigator carefully plans the experiment before starting the sample preparation and measurement campaign. When the aim is to reveal processes by identifying the location of isotopic enrichment by stable isotope labeling, which is used in many environmental sciences and intensively so in soil research, decisions must be taken on the pre-treatments: which substances, labeling isotopes, and enrichment are involved, and in what way and time scale is the experiment performed.

Once these parameters are fixed, a procedure for sample preparation for the final NanoSIMS measurements should be specified. It has to be considered that the sample holder has an upright position in the measuring chamber, so the sample must be well fixed. For soil science, from experience two ways have been established for working with fine sized particles (e.g., microaggregates, clay minerals) either on flat wafers (e.g., Si wafer) or filters or with embedded cross sections of intact soil samples. Both of these techniques have advantages and disadvantages, and the decision is mostly made based on the research question and the size of sample.

Small particles (fine silt- and or clay-sized particles not exceeding 20 μm in overall sample topography) are commonly deposited on a flat substrate that fits the SIMS sample holder (e.g., Si or GaAs wafers). This method has the advantage that the sample is not in contact with any embedding material (e.g., epoxy resin). A drawback, however, is that the topography of the particles induces possible edge artifacts and they might move in water solution. To reduce the latter, we currently use cold-substrate fixation making use of condensation rather than exposing the whole sample to water in a suspension. A wafer is cooled well before depositing particles on it by sieving, thus taking advantage of the condensation of air humidity to fix small soil particles onto the sample holder's substrate (e.g., Si wafer). In rare cases it is also possible to work directly on sample surfaces if they resemble larger flat areas such as on particulate organic matter (Witzgall et al., 2021).

The topographical drawback of small particles can be overcome by embedding them. Applying a sufficiently thick coating onto the particles' surface before embedding allows the particles' carbon species to be separated from the embedding resin, as reported by Hoeschen et al. (2015).

For structures $>20\ \mu\text{m}$ (soil aggregates, intact plant-soil systems like the rhizosphere), embedding with epoxy or polyester resins is a well-established method. However, care has to be taken that a high vacuum stable resin is used. Araldite 502 has been successfully used due to its low N content, resulting in a good contrast of the C/N ratio of this resin and SOM (Herrmann et al., 2007). It was possible to separate the Araldite epoxy NanoSIMS signal from the SOM ion signal by appropriate supervised and unsupervised image classification (Steffens et al., 2017). The embedded samples are polished as flat thin sections; they are favored over thick blocks due to a rapid outgassing before loading them into the analyzes chamber. Applications in which samples need to be embedded are gradient investigations at biogeochemical interfaces of intact larger samples (e.g., intact soil macroaggregates as in Fig. 2). This is of particular interest, for example, at the root-soil interface where exchange processes of nutrients are very intensive. As an alternative to resins, embedding in molten sulfur and cryofixation were also applied for preparing cross sections of structures $<100\ \mu\text{m}$ with the advantage of avoiding carbon contamination (Lehmann et al., 2005), but due to the rather complex preparation this approach is not widely used. Sectioning of samples with focused ion beam (FIB), usually used to produce thin sections for subsequent analysis by transmission electron microscopy, looks promising, but requires specialized equipment and only a limited volume of a sample can be processed (Vidal et al., 2018).

A single NanoSIMS measurement depicts only a very small field of view and the measurement time is costly, therefore, an important preliminary step before NanoSIMS analysis is a detailed imaging cascade of the whole sample to identify representative regions of interest (ROI). We recommend an intense investigation starting with optical microscopy, with areas of greater interest being imaged at the highest possible magnification. This step can also include staining of, for example, microorganisms to locate interesting 'hot spots' of microbial activity and or combine the imaging with phylogenetic information using fluorescent in situ hybridization (FISH) (Musat et al., 2016; Schlueter et al., 2019). Alternatively or additionally, a secondary electron microscopy (SEM) screening can be beneficial to obtain highly resolved images of ROIs. For embedded soil sections with low topographic contrast use of the backscattered electrons (BSE) detector of the SEM allows details based on elemental contrasts to be revealed. In addition, correlative imaging (e.g., helium ion microscopy, fluorescence microscopy, confocal microscopy, atomic force microscopy) are advised to target research-specific parameters and add value to the NanoSIMS analyzes (Mueller et al., 2013 and references within). To co-locate positions between all imaging methods, a coordinate transformation can be applied. Therefore, at least two distinct reference spots (unique pattern, sample edges, scratches, indications marked onto sample using laser) are referenced in each method's coordinate system.

The revolutionary coaxial configuration of NanoSIMS enables very high lateral resolution and high transmission, therefore, there is a constraint on the polarity of the primary ions sputtering the sample and the secondary ions leaving the sample surface: their charge has to be opposite. As chemical elements present different electron affinities and ionization potentials, this requires an early decision on which ion source needs to be used. For soil samples, there are two phases of interest, organic matter (OM) and mineral constituents. To depict the spatial distribution of OM, the key elements carbon and nitrogen are targeted as the species ^{12}C and ^{14}N . As carbon ionizes predominantly negatively and only poorly positively, the $^{12}\text{C}^-$ secondary ions leaving the sample need to be detected. This requires use of the Cs source to provide positive primary ions. In this configuration, nitrogen is detected as $^{12}\text{C}^{14}\text{N}^-$ (the elemental ionization of nitrogen is very poor, neither negative nor positive), while the ^{13}C (as $^{13}\text{C}^-$) and ^{15}N (as $^{12}\text{C}^{15}\text{N}^-$) stable isotopes can be measured simultaneously. The measurement of further OM-specific elements, such as phosphorus (^{31}P) and sulfur (^{32}S), and specific mineral elements (^{28}Si , $^{27}\text{Al}^{16}\text{O}^-$, $^{56}\text{Fe}^{16}\text{O}^-$) is possible. A typical configuration to cover OM and mineral compartments in soil samples treated with isotopic labels would be: $^{12}\text{C}^-$, $^{13}\text{C}^-$, $^{16}\text{O}^-$, $^{12}\text{C}^{14}\text{N}^-$, $^{12}\text{C}^{15}\text{N}^-$, $^{27}\text{Al}^{16}\text{O}^-$,

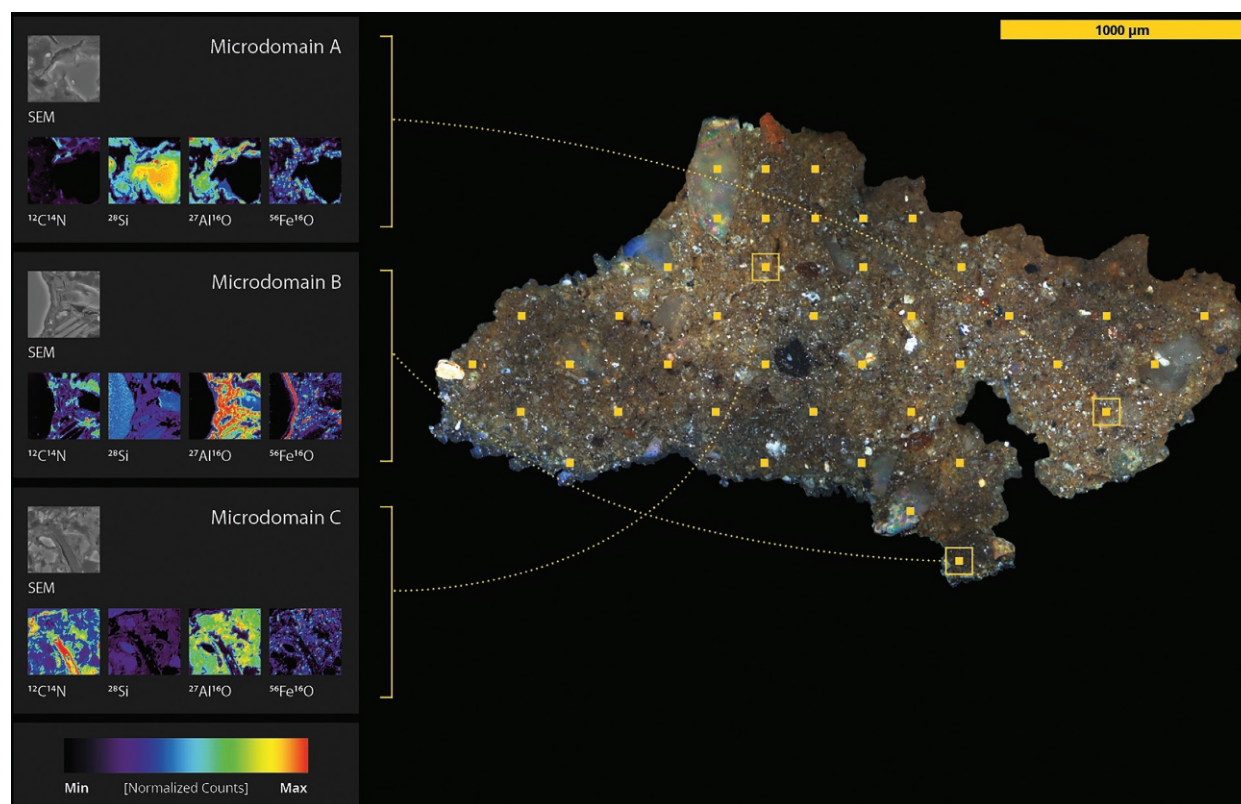


Fig. 2 A polished cross section of an embedded soil macroaggregate analyzed by Steffens et al. (2017) along a regular grid using NanoSIMS. The reflected light microscope image indicates the 40 NanoSIMS measurement locations (yellow blocks). On the left, images from a scanning electron microscope (SEM) and the NanoSIMS data for three locations are shown that represent distinct functional microdomains. All SEM and NanoSIMS images are $30\ \mu\text{m} \times 30\ \mu\text{m}$ in size. Reprinted with permission from Steffens M, Rogge DM, Mueller CW, Höschen C, Lugmeier J, Kölbl A and Kögel-Knabner I (2017) Identification of distinct functional microstructural domains controlling C storage in soil. *Environmental Science & Technology* **51**(21): 12182–12189, Copyright (Steffens et al., 2017) American Chemical Society.

$^{56}\text{Fe}^{16}\text{O}^-$. Other species such as magnesium ($^{24}\text{Mg}^+$), calcium ($^{40}\text{Ca}^+$), or potassium ($^{39}\text{K}^+$) need to be measured as positive ions ($^{24}\text{Mg}^+$, $^{40}\text{Ca}^+$, $^{39}\text{K}^+$), using the O^- source of the instrument. This configuration also allows the successful measurement of silicon, aluminum, and iron ($^{28}\text{Si}^+$, $^{27}\text{Al}^+$, $^{56}\text{Fe}^+$).

Once the sample has been well documented and the decision made of which primary ion to use, the sample is ready for loading into the NanoSIMS load lock. Soil samples tend to have charging issues on the isolating minerals (e.g., quartz grains), therefore, a thin coating (10–30 nm) with a gold-based alloy is applied independently of preparation and sample size.

According to the size of the sample, different standard sample holders can be chosen that fit into the NanoSIMS stage system with different configurations. A standard sample holder can store multiple samples with a diameter of 10 mm. For investigations of isotopic enrichment, we recommend analyzing control samples loaded on the same holder of natural isotopic abundance in the same session. Other typical sample sizes have diameters of 5 mm, 12.5 mm, or 25.4 mm.

The machine has to be carefully tuned (Mueller et al., 2013 and references within) at the outset, setting an appropriate mass resolution to detect masses affected by isobars accurately (e.g., at mass number 13 when measuring ^{13}C , which interferes with $^{12}\text{C}^1\text{H}$). The first tuning for a new sample series can take up to 1 day.

Understanding of soil systems from single particles to intact aggregates and soil cores

Mineral particles and aggregates

NanoSIMS provides elemental and or isotopic information on organo-mineral assemblages or aggregates in a spatial context from intact soil isolates. The combination of high-resolution microscopy and isotopic probing enables spatially-referenced information to be obtained on the isotopic composition of materials. Therefore, it is used to produce high-resolution 2-D maps (lateral resolution of ca. 100 nm) of the elemental/isotopic abundance in thin surface layers (penetration of the NanoSIMS primary ion beam $\sim 10\ \text{nm}$) of a soil isolate. These can either be organo-mineral associations obtained from the soil after a disintegration step, such as physical soil fractionation or from aggregates conventionally isolated from the soil by sieving. Frequently the so-called

organo-mineral associations are small micro-aggregates comprising organic and mineral particles. A major focus of such investigations has been to identify preferential associations of OM with certain mineral phases. This approach provides a visual demonstration and quantification of organic material in interaction with metal oxides and can directly prove previous indirect demonstrations by correlation. In such materials, soil minerals and organic matter can be distinguished and classified in NanoSIMS images using the distinction between OM-derived secondary ions ($^{12}\text{C}^-$, and $^{12}\text{C}^{14}\text{N}^-$) and mineral-derived secondary ions ($^{16}\text{O}^-$, $^{28}\text{Si}^-$, $^{27}\text{Al}^{16}\text{O}^-$, and $^{56}\text{Fe}^{16}\text{O}^-$). Several different mineral phases have been identified by NanoSIMS from the combined occurrence of these ions. Examples include iron and aluminum (hydr)oxides and allophane (Keiluweit et al., 2012; Patzner et al., 2020), phyllosilicates, such as smectite or illite (Steffens et al., 2017). With respect to the composition of organic matter it is to some extent possible to differentiate materials that contain only carbon based molecules, such as lipids, polysaccharides, or lignin from those that contain carbon and nitrogen based materials together, i.e., materials most probably derived from amino acids (proteins) or amino sugars (murein) (Witzgall et al., 2021). To evaluate the data obtained, suitable image analysis and statistical procedures are necessary to provide quantitative information. To facilitate interpretation of the results from NanoSIMS analyzes correlative imaging (e.g., SEM) and additional imaging data (e.g., fluorescence imaging) can be valuable.

It is possible with image analysis tools to quantify the proportion of mineral surfaces covered with organic matter and also the size and shape of the organic matter patches associated with the minerals from isolated soil organo-mineral particles (Vogel et al., 2014). At the same time information is provided on their elemental/isotopic composition. This is important and useful additional information because both the nature and strength with which OM binds to mineral surfaces are known to depend upon both the mineral type and mineral surface area.

NanoSIMS imaging may also be profitable for investigating intact three-dimensional micro-structures in soils, such as aggregates. To obtain adequately flat samples for analysis, larger soil aggregates and intact soil cores typically require embedding and subsequent cutting or thin sectioning (Mueller et al., 2017). Mineral particles pose a challenge to elemental mapping and isotope tracing experiments in intact soil matrices because they make embedding and thin-sectioning more difficult, and can cause electrical charging effects, all of which need to be considered in such analyzes (see detailed discussion in the section on sample requirements). For example, the resin embedding approach has been used to prepare a cross section of an intact soil macroaggregate for elemental in situ mapping of organic and mineral components from an arable Luvisol topsoil (Steffens et al., 2017) which is displayed in Fig. 2.

The main strength of NanoSIMS is the highly accurate isotope detection that makes it an unprecedented tool for stable isotope labeling experiments to reveal the spatial distribution and dilution of an isotopic tracer as it is incorporated into different soil compartments (e.g., microorganisms, organo-mineral associations). The co-detection of organic and mineral-derived elements enables us to study the association between these components in soils, whereas use of isotopic labels enables study of specific processes by following the label during its way into and through different soil components. It has been mostly used for tracking stable isotopes during organic matter decomposition, such as ^{13}C and ^{15}N onto distinct minerals in an intact micro-environment (e.g., Herrmann et al., 2007; Keiluweit et al., 2012; Kopittke et al., 2020; Remusat et al., 2012; Vogel et al., 2014). There are several advantages over the analysis of label(s) in bulk soil or soil fractions. One is of course that it is possible to follow the association of the labelled material directly or its transformation products specifically on mineral surfaces after isolation or within aggregates. It also provides the wide range of concentrations observed in the isolated soil fraction, and information on the processes beyond bulk data (mean values) measured with conventional isotope analysis. If both ^{13}C and ^{15}N labels are applied, NanoSIMS analyzes of specific soil fractions allow spatial differentiation of unlabeled OM (^{12}C and $^{12}\text{C}^{14}\text{N}$) and OM either double labeled (^{13}C and $^{13}\text{C}^{15}\text{N}$) or OM carrying one single label in association with minerals. Based on the NanoSIMS measurements $^{13}\text{C}^-/^{12}\text{C}^-$ and $^{12}\text{C}^{15}\text{N}^-/^{12}\text{C}^{14}\text{N}^-$ ratios can be calculated and compared to the natural abundance of an unlabeled control sample to examine how the labelled OM is incorporated and distributed in the soil material or specific soil fraction. Rumpel et al., (2015) took advantage of a field experiment with ^{13}C and ^{15}N labelled root material. They found that stabilization of root material by interaction with metal oxides occurs at different soil depths within a time frame of years. Kopittke et al. (2018) showed that N-rich microbial products preferentially attached to distinct areas of mineral surfaces compared to C-dominated moieties (Fig. 3), demonstrating the ability of soils to store additional OM in newly formed organo-mineral associations on previously OM-free mineral surfaces. They also found differences in ^{15}N enrichment between a smectite-dominated Vertisol and a kaolinite-dominated Alfisol, showing the importance of mineral composition in regulating the binding of organic compounds of most probably microbial origin in soils.

A recommended approach is also to combine the submicron characterization of isotopic label distribution by NanoSIMS with structural and chemical information from other methodologies such as TEM, scanning transmission X-ray microscopy (STXM) or near edge X-ray absorption fine structure spectroscopy (NEXAFS) (Keiluweit et al., 2012). Remusat et al. (2012) have shown how the combination of NanoSIMS, STXM and NEXAFS enabled relating the distribution of ^{15}N with the presence of microbial metabolites in intact soil aggregates, indicating the potential of this technique for in situ SOM analysis. Extracellular polymeric substances (EPS) are expected to be an important source for the formation of mineral-organic associations in soil. Liu et al. (2013) investigated the composition of EPS before and after adsorption to goethite, in combination with Raman spectroscopy and scanning transmission X-ray microscopy. They could show that organic coverage on goethite minerals was heterogeneous, consisting of $\sim 100 \times 200$ nm large patches of either lipid-rich or protein-rich material. NanoSIMS analyzes in combination with XANES of S at the K-edge provided information on the distribution and forms of As, Fe, S and OM in soils under redox level oscillations. With this approach Al-Sid-Cheikh et al. (2015) could identify ternary interactions between OM, FeO and As, which is often considered to play a major role in As-OM interactions under oxic conditions, but also a second ternary interaction between OM, S and As. Du et al. (2018) investigated the Cd sorption on ferrihydrite-OM composites formed by coprecipitation versus adsorption of OM and applied NanoSIMS in combination with Cd K-edge EXAFS. They demonstrated that the binding of Cd is a function of

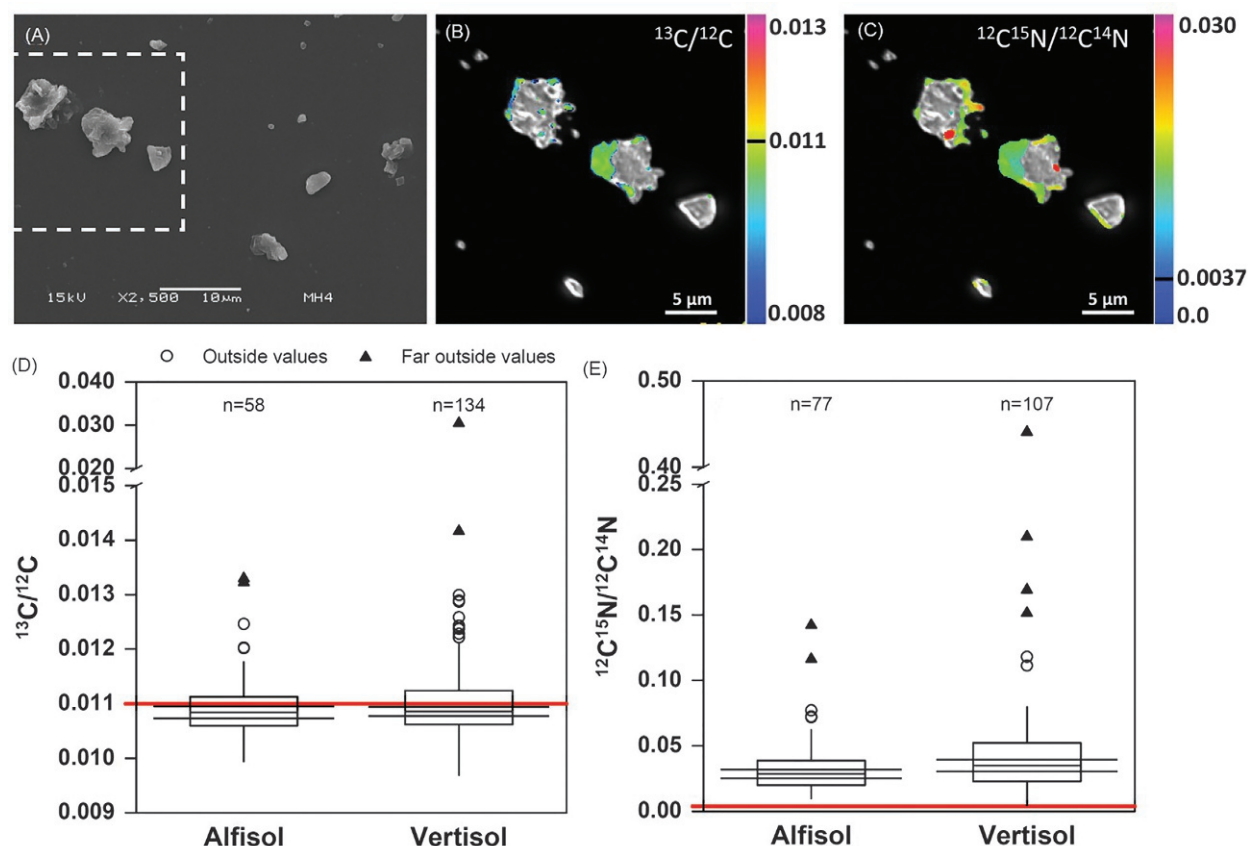


Fig. 3 Two-dimensional NanoSIMS images for microaggregates derived from an Alfisol after incubation with labelled plant residues (Kopittke et al., 2018). The images demonstrate one region of interest (ROI) that is assessed by scanning electron microscopy (SEM) prior to NanoSIMS measurement. (A) In the scanning electron microscope (SEM) image microaggregates are displayed that were deposited on a silica wafer. The dashed line indicates the region of the sample that was examined after SEM analysis using NanoSIMS. (B) The $^{13}\text{C}/^{12}\text{C}$ isotopic ratios are displayed with the ^{16}O signal in grey scale as background to depict the mineral particles. The natural abundance for ^{13}C (0.011, in panel (D)) is shown on the color scale, indicating brighter colors to correspond with ^{13}C enrichment. (C) The $^{12}\text{C}^{15}\text{N}/^{12}\text{C}^{14}\text{N}$ isotopic ratios are displayed in the same way as $^{13}\text{C}/^{12}\text{C}$, with the corresponding natural abundance marked in the color scale (see also 0.0037 in panel (F)). The sum of the C and N distribution in (B) and (C) demonstrate the total area of the microaggregates that is covered with OM. Based on a number of ROIs (n in the box plots) from multiple NanoSIMS measurements, box plots of the $^{13}\text{C}/^{12}\text{C}$ and $^{12}\text{C}^{15}\text{N}/^{12}\text{C}^{14}\text{N}$ ratios can be obtained as shown in (D) and (E). Natural isotope abundance is highlighted by the red line indicates. Hot spots of ^{13}C and ^{15}N enrichment are demonstrated by 'outside values' (>1.5 times the distance between the 25th and 75th percentiles) and 'far outliers' (>3 times the distance). The figure is adapted from Figs. 2 and 3 of Kopittke PM, Hernandez-Soriano MC, Dalal RC, Finn D, Menzies NW, Hoeschen C and Mueller CW (2018) Nitrogen-rich microbial products provide new organo-mineral associations for the stabilization of soil organic matter. *Global Change Biology* 24(4): 1762–1770.

the C loading for both types of ferrihydrite-OM compounds. In a study by Vidal et al. (2019) the TEM examination of earthworm cast aggregate ultrathin sections provided visual identification of the added labelled plant residue features. The same ultrathin section used for TEM was used for NanoSIMS tracing of elemental and isotopic composition. This combined approach allowed the identification of different types of organic matter (plant residues, microbial remains) in the cast aggregates, and at the same time the freshly incorporated residues could be differentiated from the indigenous organic matter by mapping ^{13}C using NanoSIMS. This kind of combined approach, NanoSIMS image analysis joined by complementary microscopy (SEM-EDX, STXM/NEXAFS), has been used several times now and seems to be a particularly profitable way to infer the molecular and spatial fate of labeled organic materials in a mineral matrix. It has the potential to contribute further to a mechanistic understanding of sorption, occlusion, and decomposition processes that operate at fine spatial scales. As outlined above, such analyzes should be combined with classical SEM and micromorphological studies for an accurate assessment of OM distribution in the soil and its relationship to ecological processes at larger scales.

Interfaces of soils and biota

The main hot spots for biogeochemical cycles in soils are the interfaces between biota and mineral soil constituents, with plants and their residues being the most prominent biota that drive soil processes due to their large OM input. Most organic matter enters soils either from decaying above- and below-ground plant litter or the input of living roots, so called rhizodeposits. Although large volumes of soil are affected by these processes, the transformation of plant-derived OM into SOM takes place at distinct microscale

biogeochemical interfaces between plants and their residues, microorganisms and soil minerals, either in the rhizosphere (volume of soil directly affected by root activity) or the detritosphere (volume of soil affected by decaying plant residues). NanoSIMS can also provide an approach to follow directly the fate of plant-derived OM into microorganisms and soil particles at these μm scale soil interfaces. However, these process sites are within a highly complex 3-D soil matrix that complicates sample preparation for imaging approaches. Elucidation of the physical structure of intact rhizosphere or detritosphere soil volumes is relatively easy using X-ray micro computed tomography (CT) (Schlueter et al., 2019), whereas the preparation of meaningful intact flat samples for electron microscopy or NanoSIMS requires specific sample treatments described above. The preparation of embedded cross sections ensures the integrity of the 3-D structural arrangement of the biota-soil interface for meaningful micrometer analyzes. The preservation of biological cell morphologies and spatial arrangement of analytes is mostly done by chemical fixation (Herrmann et al., 2007; Mueller et al., 2013). One classical procedure for sample preparation is to preserve the sensitive plant and microbial cell structures using chemical fixation followed by dehydration with a series of water to solvent mixtures and final embedding in a polyester or epoxy resin. The embedded sample can then be sectioned and polished if a flat sample surface with low topography is needed. This approach gave good results with NanoSIMS for studying intact rhizosphere and detritosphere samples. Other imaging techniques are also possible in combination with NanoSIMS, for example fluorescence microscopy to localize microbial cells (Schlueter et al., 2019; Vidal et al., 2018). For very sensitive samples and the need to preserve highly mobile elements e.g., potassium, it is also possible to use flash freezing and cryo-substitution (Kilburn and Clode, 2014). If the sample itself is vacuum stable and has large flat areas with low topography it can be examined directly without embedding. Fine structures like thin roots and fungal hyphae might also be analyzed directly in the NanoSIMS, and also more complex systems. For example, Witzgall et al. (2021) studied the microbial transformation of particulate OM in the detritosphere directly at the interface of ^{13}C enriched plant residues, microorganisms, and soil minerals without any embedding. If the sample surface is irregular in terms of topography there are concerns about artifacts due to differences in the ion yield of different secondary ions, but this effect is relatively small for isotopologues (e.g., $^{13}\text{C}_2$, $^{12}\text{C}_2$), the most often used molecular cluster secondary ions to trace isotopic enrichment.

As described above for mineral-associated OM, the capability of NanoSIMS to image ^{13}C and ^{15}N allows plant-derived OM at distinct plant-soil interface systems to be followed directly from the plant into microorganisms and soil particles (Gorka et al., 2019). Bulk analysis can only yield the overall enrichment for instance of a whole root, extracted microbial biomass or a sum of picked fungal hyphae, whereas NanoSIMS enables quantification of the plant-derived ^{13}C and or ^{15}N enrichment to the single cell level and to follow its transformation directly into microbial C and N as well as the buildup of SOM. This offers the opportunity to study gradients starting at the plant root, into microorganisms at the rhizoplane and mineral surfaces at certain distances from the root (Fig. 4). A fundamental property of the rhizosphere and detritosphere, the emergent development of gradients of elements and processes, can also be studied.

By separating single soil constituents (e.g., bacteria, fungal hyphae) it is possible to analyze them separately, and for example bacterial cells on fungal hyphae can be imaged (Gorka et al., 2019). However, intact biota-soil interfaces comprise tremendously complex systems for microspatial analyzes as they combine soft plant, animal or microbial cell structures and rigid soil mineral particles. In a first attempt to study the interaction and competition between plants and soil microorganisms for nitrogen in the rhizosphere, Clode et al. (2009) used a microcosm setup in which the authors grew wheat exposed to $^{15}\text{NH}_4^+$ in a loamy soil. Intact rhizosphere samples were chemically fixed, resin embedded and finally polished cross sections were produced. These samples were analyzed by transmission electron microscopy and subsequently NanoSIMS, which allowed the authors to demonstrate that microbial cells in the direct vicinity of the plant root took up most of the added ammonia. Thus, this work demonstrated for the first time the great advantage of the high spatial resolution of isotopic imaging for soil science as there is no other way to differentiate microscale elemental cycling in rhizosphere constituents (roots, bacteria, fungi, etc.) down to the single cell level. Herrmann et al. (2007) demonstrated this setup for larger soil cores and opened the way to analyze biogeochemical interfaces in more complex natural soils.

The rhizosphere, in particular, with its intricate connection between plant roots, soil microorganisms and soil minerals is at the focal point of applying NanoSIMS to follow the fate of plant-derived OM into soil microorganisms. In controlled mesocosm setups it is possible to follow the fate of photosynthetically fixed $^{13}\text{CO}_2$ from the plant root into the rhizosphere (Gorka et al., 2019). Kaiser et al. (2015) demonstrated with NanoSIMS the imaging of root cells, fungal hyphae and soil microorganisms that photosynthetically fixed OC is rapidly transferred into the roots and from there, independently from passive exudation, into arbuscular mycorrhizal and via their fungal hyphae into soil microorganisms. These authors showed that the mycorrhizal network is important for the allocation of plant-derived OC into soil microorganisms. With a split pot design in which plant roots were separated from extraradical fungal hypha, Gorka et al. (2019) demonstrated the use of two isotopic tracers, $^{13}\text{CO}_2$ for the plant-derived C and ^{15}N for litter-derived N. By NanoSIMS imaging of the $^{13}\text{C}/^{12}\text{C}$ distribution, the authors showed the uptake of hyphae-derived OC by bacterial cells residing on the hyphae itself. But this approach can also be used for samples derived from field trials. Vidal et al. (2018) used in-growth cores to sample the intact rhizosphere soil of wheat from an in situ labeling experiment, and used NanoSIMS to trace plant-derived OC into roots, rhizosphere microorganisms and soil particles. This study highlights the applicability of the methodology and analytical approach to support field experimental data by mechanistic underpinnings at the relevant process scale. In addition to the interfaces of plants and soil, those of soil biota and soil can also be assessed using NanoSIMS. Embedded cross sections of earthworm casts were used to follow the different fate of ^{13}C labelled above- and below-ground plant residues into mineral-associated organic matter using a cascade of TEM and NanoSIMS analyzes (Vidal et al., 2019). The high resolution of TEM combined with the isotopic imaging capability of NanoSIMS enabled demonstration of the intricate connection of soil microorganisms and the formation of mineral-associated OM regulated by earthworms.

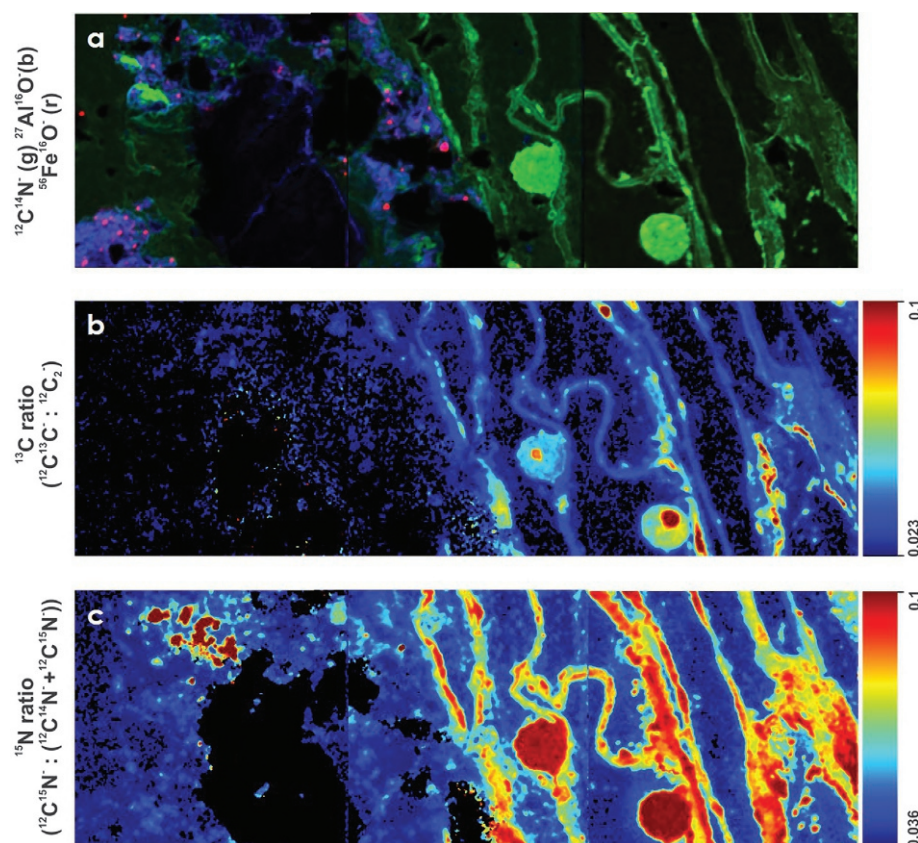


Fig. 4 Images based on NanoSIMS measurements of an epoxy resin embedded cross-section of a root-soil interface of a primary root of a 21-days old maize (*Zea mays* L.) plant grown in loam labelled by both $^{13}\text{CO}_2$ and $^{15}\text{NH}_4$ enrichment. Panel (a) displays a composite image to illustrate the intricate connection of the root with the surrounding minerals, depicting plant and microbial OM based on the distribution of $^{12}\text{C}^{14}\text{N}$ (green), and mineral soil constituents based on the distribution of $^{27}\text{Al}^{16}\text{O}$ (blue) and $^{56}\text{Fe}^{16}\text{O}$ (red). Panel (b) shows the distribution of the ^{13}C ratio, and panel (c) shows the distribution of the ^{15}N ratio. The intense high color areas highlight microscale spots enriched in root derived ^{13}C and fertilizer derived ^{15}N at the surface of a soil aggregate in the vicinity of the highly enriched root (E. Lippold, C. Hoeschen, C.W. Mueller unpublished data).

Detection of microbiota in the soil environment

Soil microbiology is a vibrant field within soil science with the application of microscopic techniques. In microbiology light microscopy and fluorescence staining are used routinely, and provide an excellent basis to build on with spectromicroscopic techniques such as NanoSIMS. Sample preparation protocols that aim to preserve very sensitive biological structures like bacterial cells are important for developing imaging cascades from light microscopy to electron microscopy and NanoSIMS. The high spatial resolution and the possibility of using isotopic tracers make NanoSIMS the perfect tool to analyze individual microbial cells. This allows us to differentiate the activity and functional importance of distinct microbial phyla and/or consortia. In environmental microbiology the application of NanoSIMS for aquatic systems is relatively straightforward by analyzing cells directly fixed on filter discs. However, the study of soil microbiota at the microscale in natural soils is hampered by their sparse distribution; most microbiota accumulate at hot spots like the rhizosphere and it is difficult to assess a sufficient number of cells against a large soil background. Either the microscale analyzes focus on the functioning of such hot spots with a sufficient presence of microbiota as discussed above for the rhizosphere and detritosphere, or other methods have to be used. A relatively simple approach to yield sufficient soil microbiota from the rhizosphere together with soil particles is by shaking root samples with adhering soil in deionized water and desiccation of the suspension on a silica wafer. Although this approach does not ensure the preservation of intact microbial cells because a fixation step is missing, it can provide insight into the isotopic enrichment of fungal and bacterial cells derived from a certain soil sample. If the shaking is done in 50% ethanol cell structure preservation will be better and ensured, and thus allow the analysis of intact microbial cells (Hestrin et al., 2021). A more comprehensive method to separate soil microorganisms from the soil and increase the number of analyzable cells was developed by Eichorst et al. (2015). After chemical fixation of the soil samples, microbial cells were extracted from the soil slurries by filtering. This approach leads to a high cell density and reduced numbers of soil particles on the filters. This approach can be used to analyze more individual cells while also allowing the combination of NanoSIMS with previous staining of the cells to obtain phylogenetic information by fluorescence in situ hybridization (FISH).

An approach independent from the single cell level was developed to be able to analyze isotopic enrichment in a larger number of different phyla (Pett-Ridge and Weber, 2021). The so called NanoSIP is based on a high-density microarray approach that measures isotopic enrichments (^{13}C , ^{15}N , ^{18}O) of taxon-specific 16S (or 18S) rRNA genes by NanoSIMS. The combination allows the taxon-specific quantification of the incorporation of substrates in complex microbial communities. This is a nice example of how the high spatial resolution and high sensitivity of the NanoSIMS can be used for analyzes with higher throughput of sample numbers that cannot be measured by classical isotopic ratio mass spectrometry.

Conclusion

The conceptual understanding of the functioning of soil systems ranging from ecosystems down to microscales has developed rapidly. We have now started to understand microscale processes better with the development and increasing access to imaging techniques capable of resolving very fine soil structures down to soil aggregates, single minerals and microbial cells. With NanoSIMS the soil science community has an unprecedented tool at hand to analyze directly the elemental composition of soil interfaces and to follow the fate of elements in complex soil systems using enrichment with stable isotopes. NanoSIMS provides the important step from a single data point of, for example, a $^{13}\text{C}/^{12}\text{C}$ ratio to the detailed analysis of microbial utilization of plant-derived organic carbon at a single cell level. NanoSIMS is a strong supporting technique to underpin conceptual ideas by direct imaging of the relevant process scale and the processes and players involved. It is up to the soil science community to integrate this tool wisely into experiments and to use the large analytical tool box in conjunction to unravel soil processes at the microscale. The future might bring more research that goes beyond the use of the classical stable isotopes ^{13}C and ^{15}N , but aims at using the full mass range that can be resolved by NanoSIMS.

References

- Al-Sid-Cheikh M, Pédrot M, Dia A, Guenet H, Vantelon D, Davranche M, Gruau G, and Delhaye T (2015) Interactions between natural organic matter, sulfur, arsenic and iron oxides in re-oxidation compounds within riparian wetlands: NanoSIMS and X-ray adsorption spectroscopy evidences. *Science of the Total Environment* 515–516: 118–128.
- Clode PL, Kilburn MR, Jones DL, Stockdale EA, Cliff JB III, Herrmann AM, and Murphy DV (2009) In situ mapping of nutrient uptake in the rhizosphere using nanoscale secondary ion mass spectrometry. *Plant Physiology* 151: 1751–1757.
- Du HH, Peacock CL, Chen WL, and Huang QY (2018) Binding of Cd by ferrihydrite organo-mineral composites: Implications for Cd mobility and fate in natural and contaminated environments. *Chemosphere* 207: 404–412.
- Eichorst SA, Strasser F, Woyke T, Schintlmeister A, Wagner M, and Woebken D (2015) Advancements in the application of NanoSIMS and Raman microspectroscopy to investigate the activity of microbial cells in soils. *FEMS Microbiology Ecology* 91: 14.
- Gorka S, Dietrich M, Mayerhofer W, Gabriel R, Wiesenbauer J, Martin V, Zheng Q, Imai B, Prommer J, Weidinger M, Schweiger P, Eichorst SA, Wagner M, Richter A, Schintlmeister A, Woebken D, and Kaiser C (2019) Rapid transfer of plant photosynthates to soil bacteria via ectomycorrhizal hyphae and its interaction with nitrogen availability. *Frontiers in Microbiology* 10: 20.
- Hallett PD, Karim KH, Bengough AG, and Otten W (2013) Biophysics of the Vadose Zone: From reality to model systems and back again. *Vadose Zone Journal* 12: 17.
- Hatton P-J, Remusat L, Zeller B, and Derrien D (2012) A multi-scale approach to determine accurate elemental and isotopic ratios by nano-scale secondary ion mass spectrometry imaging. *Rapid Communications in Mass Spectrometry* 26: 1363–1371.
- Herrmann AM, Ritz K, Nunan N, Clode PL, Pett-Ridge J, Kilburn MR, Murphy DV, O'Donnell AG, and Stockdale EA (2007) Nano-scale secondary ion mass spectrometry—A new analytical tool in biogeochemistry and soil ecology: A review article. *Soil Biology and Biochemistry* 39: 1835–1850.
- Hestrin R, Weber PK, Pett-Ridge J, and Lehmann J (2021) Plants and mycorrhizal symbionts acquire substantial soil nitrogen from gaseous ammonia transport. *New Phytologist* 231: 1746–1757.
- Hillion F, Daigne B, Girard F, and Slodzian G (1994) *A new high performance instrument: The Cameca 'NanoSIMS 50'.* *Secondary Ion Mass Spectrometry SIMS IX*. New York: John Wiley and Sons 254–257.
- Hoeschen C, Hoeschen T, Mueller CW, Lugmeier J, Elgeti S, Rennert T, and Kogel-Knabner I (2015) Novel sample preparation technique to improve spectromicroscopic analyses of micrometer-sized particles. *Environmental Science & Technology* 49: 9874–9880.
- Hoppe P, Cohen S, and Meibom A (2013) NanoSIMS: Technical aspects and applications in cosmochemistry and biological geochemistry. *Geostandards and Geoanalytical Research* 37: 111–154.
- Kaiser C, Kilburn MR, Clode PL, Fuchsluger L, Koranda M, Cliff JB, Solaiman ZM, and Murphy DV (2015) Exploring the transfer of recent plant photosynthates to soil microbes: Mycorrhizal pathway vs direct root exudation. *New Phytologist* 205: 1537–1551.
- Keilueit M, Bougoure JJ, Zeglin LH, Myrold DD, Weber PK, Pett-Ridge J, Kleber M, and Nico PS (2012) Nano-scale investigation of the association of microbial nitrogen residues with iron (hydr)oxides in a forest soil O-horizon. *Geochimica et Cosmochimica Acta* 95: 213–226.
- Kilburn MR and Clode PL (2014) Elemental and isotopic imaging of biological samples using NanoSIMS. In: Kuo J (ed.) *Electron Microscopy: Methods and Protocols*, 3rd edn, vol. 1117, pp. 733–755.
- Kopittke PM, Hernandez-Soriano MC, Dalal RC, Finn D, Menzies NW, Hoeschen C, and Mueller CW (2018) Nitrogen-rich microbial products provide new organo-mineral associations for the stabilization of soil organic matter. *Global Change Biology* 24: 1762–1770.
- Kopittke PM, Dalal RC, Hoeschen C, Li C, Menzies NW, and Mueller CW (2020) Soil organic matter is stabilized by organo-mineral associations through two key processes: The role of the carbon to nitrogen ratio. *Geoderma* 357: 9.
- Lamberti WA (2005) Imaging secondary ion mass spectrometry. In: Yao N and Wang ZL (eds.) *Handbook of Microscopy for Nanotechnology*, pp. 207–225. Boston, MA: Springer US.
- Lehmann J, Liang B, Solomon D, Lerotic M, Luizao F, Schafer T, and Wirick S (2005) Near-edge x-ray absorption fine structure (NEXAFS) spectroscopy for mapping nano-scale distribution of organic carbon forms in soils: Application to black carbon particles. *Global Biogeochemical Cycles* 19: 12.
- Liu X, Eusterhues K, Thieme J, Ciobota V, Hoeschen C, Mueller CW, Kuesel K, Kogel-Knabner I, Roesch P, Popp J, and Totsche KU (2013) STXM and NanoSIMS Investigations on EPS Fractions before and after Adsorption to Goethite. *Environmental Science & Technology* 47: 3158–3166.
- Mueller CW, Weber PK, Kilburn MR, Hoeschen C, Kleber M, and Pett-Ridge J (2013) Advances in the analysis of biogeochemical interfaces: NanoSIMS to investigate soil microenvironments. In: Sparks DL (ed.) *Advances in Agronomy*. vol. 121, pp. 1–46. Academic Press.

- Mueller CW, Hoeschen C, Steffens M, Buddenbaum H, Hinkel K, Bockheim JG, and Kao-Kniffin J (2017) Microscale soil structures foster organic matter stabilization in permafrost soils. *Geoderma* 293: 44–53.
- Musat N, Musat F, Weber PK, and Pett-Ridge J (2016) Tracking microbial interactions with NanoSIMS. *Current Opinion in Biotechnology* 41: 114–121.
- Patzner MS, Mueller CW, Malusova M, Baur M, Nikeleit V, Scholten T, Hoeschen C, Byrne JM, Borch T, Kappler A, and Bryce C (2020) Iron mineral dissolution releases iron and associated organic carbon during permafrost thaw. *Nature Communications* 11: 6329.
- Pett-Ridge J and Weber PK (2021) NanoSIP: NanoSIMS applications for microbial biology. *Microbial Systems Biology. Methods in Molecular Biology*, vol. 2349. New York: Humana.
- Polerecky L, Adam B, Milucka J, Musat N, Vagner T, and Kuypers MMM (2012) Look@NanoSIMS—A tool for the analysis of nanoSIMS data in environmental microbiology. *Environmental Microbiology* 14: 1009–1023.
- Remusat L, Hatton P-J, Nico PS, Zeller B, Kleber M, and Derrien D (2012) NanoSIMS study of organic matter associated with soil aggregates: Advantages, limitations, and combination with STXM. *Environmental Science & Technology* 46: 3943–3949.
- Rumpel C, Baumann K, Remusat L, Dignac M-F, Barré P, Deldicque D, Glasser G, Lieberwirth I, and Chabbi A (2015) Nanoscale evidence of contrasted processes for root-derived organic matter stabilization by mineral interactions depending on soil depth. *Soil Biology and Biochemistry* 85: 82–88.
- Schlueter S, Eickhorst T, and Mueller CW (2019) Correlative imaging reveals holistic view of soil microenvironments. *Environmental Science & Technology* 53: 829–837.
- Steffens M, Rogge DM, Mueller CW, Hoeschen C, Lugmeier J, Kölbl A, and Kögel-Knabner I (2017) Identification of distinct functional microstructural domains controlling C storage in soil. *Environmental Science & Technology* 51: 12182–12189.
- Vidal A, Hirte J, Bender SF, Mayer J, Gattlinger A, Hoeschen C, Schädler S, Iqbal TM, and Mueller CW (2018) Linking 3D soil structure and plant-microbe-soil carbon transfer in the rhizosphere. *Frontiers in Environmental Science* 6: 14.
- Vidal A, Watteau F, Remusat L, Mueller CW, Nguyen TT, Buegger F, Derenne S, and Quenea K (2019) Earthworm cast formation and development: A shift from plant litter to mineral associated organic matter. *Frontiers in Environmental Science* 7: 15.
- Vogel C, Mueller CW, Hoeschen C, Buegger F, Heister K, Schulz S, Schlöter M, and Koegel-Knabner I (2014) Submicron structures provide preferential spots for carbon and nitrogen sequestration in soils. *Nature Communications* 5: 2947.
- Walker AV (2010) Secondary ion mass spectrometry. In: Lindon JC (ed.) *Encyclopedia of Spectroscopy and Spectrometry*, 2nd edn, pp. 2504–2510. Oxford: Academic Press.
- Witzgall K, Vidal A, Schubert DI, Hoeschen C, Schweizer SA, Buegger F, Pouteau V, Chenu C, and Mueller CW (2021) Particulate organic matter as a functional soil component for persistent soil organic carbon. *Nature Communications* 12: 10.

Nanoparticles in soil

Di Lang, Quan Chen, Yang Liu, and Bo Pan, Yunnan Provincial Key Laboratory of Soil Carbon Sequestration and Pollution Control, Faculty of Environmental Science & Engineering, Kunming University of Science & Technology, Kunming, P. R. China

© 2023 Elsevier Ltd. All rights reserved.

Introduction	681
Natural and engineered nanoparticles in soil	682
Naturally occurring nanoparticles	682
The origin of engineered nanoparticles in soil	682
Identification of nanoparticles in soil	683
Separation of nanoparticles from other particles	683
Techniques in nanoparticle characterizations	683
Techniques for distinguishing natural and engineered nanoparticles	684
The fate of nanoparticles in soil	685
The dissolution of nanoparticles in the soil system	686
The transport of nanoparticles in the soil system	686
The aging of nanoparticles in the soil system	687
The interactions between nanoparticles and soil components	687
The assembling of nanoparticles with soil particles	687
The surface coating of organic matter on nanoparticles	687
The embedding of nanoparticles in soil aggregates	688
The impacts of nanoparticles on soil systems	688
Beneficial application of nanoparticles in agricultural activities	688
The uptake of nanoparticles by soil organisms	689
Nanoparticles as a two-blade sword to the soil system	690
Summary	690
Acknowledgments	691
References	691

Abstract

Nanoparticles exist widely in soil. They could be generated by natural processes, or could be made artificially and released into the environment. Both natural and engineered nanoparticles are active components, and greatly affect the chemical processes in the soil system. This chapter describes nanoparticles in soil taking account of the following aspects: separation and identification of nanoparticles in soil, the fate and interactions of nanoparticles in soil, and the effects of nanoparticles on soil systems. The research progress and demands are discussed regarding the characterization, environmental behavior, and toxicity assessment of nanoparticles in soil.

Key points

- Both natural and engineered nanoparticles are reactive components in soil.
- They can be separated and quantified using sophisticated techniques.
- Nanoparticles are aged in soil with continuous changes in their properties.
- The negative and positive impacts should be balanced for their application.

Introduction

The word nano is from the Greek “Nanos” meaning dwarf. It is a prefix used to describe one billionth (10^{-9}). A nanometer (nm) is a billionth of a meter, or a millionth of a millimeter. Nanoparticles (NPs) usually have sizes in the range of 1–100 nm for at least one dimension. They may be formed through natural processes (NNPs) or well-controlled engineered processes (ENPs) (Pan and Xing, 2010). Unlike traditional materials, NPs show many special physical and chemical properties, such as quantum size, surface, and quantum tunneling effects. They are widely used in magnetic, microelectronic and catalytic materials, fuel cells, and biosensors.

Although nanotechnology has a promising potential in many fields, some researchers indicated the potential of biological and environmental safety issues of nanomaterials and nanotechnology several decades ago (Colvin, 2003). Because of the nanoscale effect, NPs have been proved to have strong toxicological impacts. Thus, many researchers believe that ENPs should be applied with

great care, and should be prohibited for in situ soil remediation (Tratnyek and Johnson, 2006). However, nanomaterials may be released into the environment through different pathways. The identification and differentiation of NNPs and ENPs are therefore important for studies on NPs and their environmental risks.

Natural and engineered nanoparticles in soil

Naturally occurring nanoparticles

Natural occurring nanoparticles (NPs) occur widely in soils, sediments, volcanic dust and ash, ocean surface microlayers, and other natural waters. Soil is a reservoir of inorganic and organic NNPs such as clay minerals, humic substances, and metal oxides. The NNPs in soil are mainly the combined result of long-term physical, chemical and biological processes. For example, they could be the products of minerals weathering or microbial activities. Chemical weathering tends to produce metal oxide NNPs like magnetite and hematite. The NNPs produced by biological processes also include humic NPs as well as iron and manganese oxides.

The NNPs in soil are mainly nano-clay minerals, which are generally composed of nano-layered silicate, amorphous nano-silicate, and metal oxides such as iron and aluminum oxides. Their large surface areas make their surfaces ideal for immobilizing enzymes, retaining harmful substances, and adsorbing nutrients needed by plants. For example, natural nano-ball allophane is a hollow spherical amorphous silicate with a diameter of 3.5–5.0 nm and a surface area of 700–900 m² g⁻¹. The loaded enzymes have greater mobility and activity on the nano-ball allophane because of the hollow sphere structure. Thus, the interface reaction activity of soil particles is enhanced by the allophane nanoclay. In soil, the particle size of natural iron and aluminum oxide ranges from 5 to 100 nm. For instance, hematite (α -Fe₂O₃) NPs can quickly adsorb heavy metal ions such as Cu²⁺, and can also catalyze Mn²⁺ to manganese oxide minerals. It should be noted that NNPs in soil usually exist in the form of organic-inorganic complexes, namely organic coatings on mineral particles. For allophane, the coated organic matter may significantly increase the adsorption of heavy metals. The organic-inorganic complexes may also facilitate the electron transfer between NNPs and microorganisms or pollutants. However, the active sites could be blocked after organic matter coating, which can consequently decrease the reactivity of NNPs.

Thus, NNPs are important constituents of soils because they are involved in numerous fundamental biogeochemical processes, such as water regulation, elemental cycling and bioavailability through the sorption, retention and transport of nutrients, natural organic matter and organic-inorganic contaminants (Nowack and Bucheli, 2007). These processes are strongly influenced by the physicochemical properties of NNPs as well as their interaction with soil components.

The origin of engineered nanoparticles in soil

In the past two decades, nanotechnology has attracted great attention and has been widely used in various fields. Engineered nanoparticles (ENPs) are NPs designed and produced for a specific purpose, also with a size of 1–100 nm at least for one dimension. The ENPs can be divided simply into two categories according to whether they contain carbon or not: carbon-based material NPs and inorganic engineered NPs, respectively. The most common carbon-based material NPs are fullerenes and carbon nanotubes. They have a large specific surface area, electron transfer properties and thermal properties, and so they have been widely used in the fields of medical sciences, pollutant adsorption and degradation. The inorganic engineered nanomaterials include nano-metal, nano-metal oxides and metal salts. Silver (Ag) NPs have highly effective antibacterial properties which can inhibit a variety of bacteria, fungi and viruses, and thus are used as supplements in antibacterial fibers, fabrics, plastics, ceramics, and antibacterial coatings. Nano-Ag antibacterial products are also favored by the majority of consumers. However, with their strongly antibacterial properties, the release of Ag NPs into the environment inevitably increases concerns about the risks to the ecosystem. For example, Ag NPs may be released from commercially available Ag-containing socks during washing. In the water after washing socks, 3.0–1300 and 1.5–650 µg Ag were detected in both granular and ionic forms, respectively. Similar releases of Ag NPs can be expected for toothbrushes, clothes, and cooking wares (Nowack and Bucheli, 2007). These ENPs may be released through washing, sweating or rainwater scouring, and then enter the urban sewage network. Conventional municipal sewage treatment does not deal with NPs, and consequently they enter into the environment after wastewater treatment.

With the development of nano science and technology, more and more nanomaterials are applied to our daily life. “The Nanodatabase” (Danish Consumer Council) lists a total of 5040 ENPs-containing products for 2020, four times more than for 2012 (the first report). These ENPs are widely used in textiles, cosmetics, daily necessities, household appliances, medical, electronics, coatings, catalysts, sensors, semiconductor materials and machinery manufacturing. It is easy to envisage that during the synthesis of ENPs, the manufacturing, utilization, abrasion and disposal of nanoproducts, ENPs will be constantly released into the environment. Some of them enter the urban sewage treatment system, and may end up in sewage sludge and wastewater treatment facilities after water treatment processes. Even the land application of treated sludge and treated wastewater may introduce ENPs to the environment. A small portion of ENPs may enter the soil and water environments directly through atmospheric deposition, wet deposition, and surface runoff, and finally accumulate in the sediments and soil. With their small size and special morphology, ENPs often have unique physical and chemical properties. It should be noted that these anthropogenic particles may induce ecological toxicity, or become carriers of other pollutants, posing potential environmental risks.

Identification of nanoparticles in soil

Separation of nanoparticles from other particles

Separation or extraction of NPs is a prerequisite for studying the fate of NPs and the implications. However, the separation of NPs from the soil matrix is extremely challenging because of their small particle size and easy aggregation. The aggregation of NPs may occur among the same NPs (homo-aggregation) or among different NPs or even NPs and other substances (hetero-aggregation). In soil systems, NPs tend to associate with organic matter and other particles forming hetero-aggregates, especially in the presence of multivalent cations. Extraction of NPs from soils usually comprises a combination of different processes, such as sonication or dispersion of the particles in a liquid followed by size separation through centrifugation or filtration.

Ultrasonication is an effective method to disperse NPs from soil and has been widely used. When the ultrasonic wave is applied in the suspension, a large number of small bubbles will be produced in the liquid due to escape of the supersaturated gas. As the bubbles grow and burst continuously, they disperse the particles in the suspension. To ensure the effectiveness of the dispersion, ultrasonic temperature, energy, and time are adjusted depending on the concentration of the suspension, soil type, organic matter and even mineral compositions.

It should be noted that a one-step dispersion may not be able to fully separate individual NPs. A complete disaggregation of NPs requires that the interaction forces such as van der Waals attractions and organo-mineral interactions are overcome. Therefore, the application of chemical extractants always helps to disperse NPs more extensively. For example, alkaline conditions enhance SOM solubilization or desorption and may favor the disaggregation process of organo-mineral hetero-aggregates. Thus, a pretreatment of soil particles using NaOH, Na₂CO₃, or Na₄P₂O₇ is always of value. In addition, chemical reagents such as Na₂C₂O₄, EDTA, and Na₄P₂O₇, could be applied to complex multivalent cations released from clays. Such complexation prevents specific adsorption of multivalent cations onto particles, which is also beneficial for dispersion of the aggregates. Phosphate-based extractants are also shown to adsorb specifically on to colloids, modify their surface charge, and facilitate disaggregation processes.

After an effective dispersion of the particles, NPs may be obtained through particle separation. At present, the technology for the separation of NPs includes membrane separation (such as filtration, microfiltration, and ultrafiltration dialysis), size exclusion chromatography (SEC), hydrodynamic chromatography (HDC), and field flow fractionation (FFF). The dispersion and separation will eventually ensure a quantitative or qualitative analysis of NPs using various 'back-end' detection equipment, such as the UV-Visible spectrum (UV-Vis), inductively coupled plasma mass spectrometry (ICP-MS), multiangle laser light scattering (MALS) and dynamic light scattering (DLS). Size distribution of various types of NPs and some common analytical techniques are summarized in Fig. 1 (Klaine et al., 2008).

Techniques in nanoparticle characterizations

The particle size of NPs is one of the most important indicators to characterize their physical and chemical properties. Various methods and instruments are applied for particle-size analysis of NPs, but each technique has its own advantages and disadvantages. Currently, DLS is the most widely used instrument to determine the particle-size distribution of NPs in the aqueous phase. This technique is based on Brownian motion, and measures particles down to 2 nm. However, because the intensity of scattered light is proportional to the sixth power of the particle diameter, the analysis is heavily weighted towards larger particles. Considering the highly heterogeneous size distributions in soil, the detection of small particles may be greatly shielded by the larger particles. A preliminary separation of soil particles into different particles-size ranges may greatly enhance the reliability of DLS determination.

Nano particle tracking analysis (NTA) is also a light scattering technique for analyzing the size distribution of NPs with high resolution, which tracks the Brownian motion of NPs in suspension in real time. The particle-size distribution and concentrations are obtained using Stokes-Einstein equation. During NTA measurement, individual particles are identified and tracked, and the distance travelled by the particle in each time period is related to their hydrodynamic diameter. Thus, the NTA technique is a combination of DLS and image analysis. Because NTA tracks individual particles, it is not limited by polydisperse samples as for DLS. This technique also shows high sensitivity to particle numbers and size distributions, and requires less time than microscopic techniques. The NTA method has been successfully used in detecting NPs in environmental samples, such as fullerenes and TiO₂ NPs in river water, as well as municipal solid waste leachate. A drawback of NTA is its detection limitations, and reliable data cannot be obtained for particles below 20 nm. In addition, we cannot directly observe the morphology, structure and chemical composition of NPs by NTA, and thus a combination of image analysis is always advantageous.

Microscopy-based image analysis, such as transmission electron microscopy (TEM), scanning electron microscopy (SEM), scanning transmission electron microscopy (STEM), and atomic force microscopy (AFM), have been successfully used in the characterization of NP morphology, dispersion and size distribution. These microscopic techniques break through the limitation of 0.2 µm for traditional optical microscopes and can observe particles at a much higher resolution. For example, during TEM analysis, a short-wavelength electron beam penetrates the sample and then magnifies the image. The resolution of TEM is related to the electron beam acceleration voltage. Considering the 80–300 kV electron beam acceleration voltage applied in TEM measurement, the resolution could reach 0.1–0.2 nm. High-end TEM can even achieve atomic resolution.

Scanning electron microscopy is a technique that combines TEM and the optical microscope. It uses a very narrow focused high-energy electron beam to scan the sample. The instrument collects optical signals excited during the interaction between the beam and the substance, and magnifies the information and images to characterize the microscopic morphology of the substance.

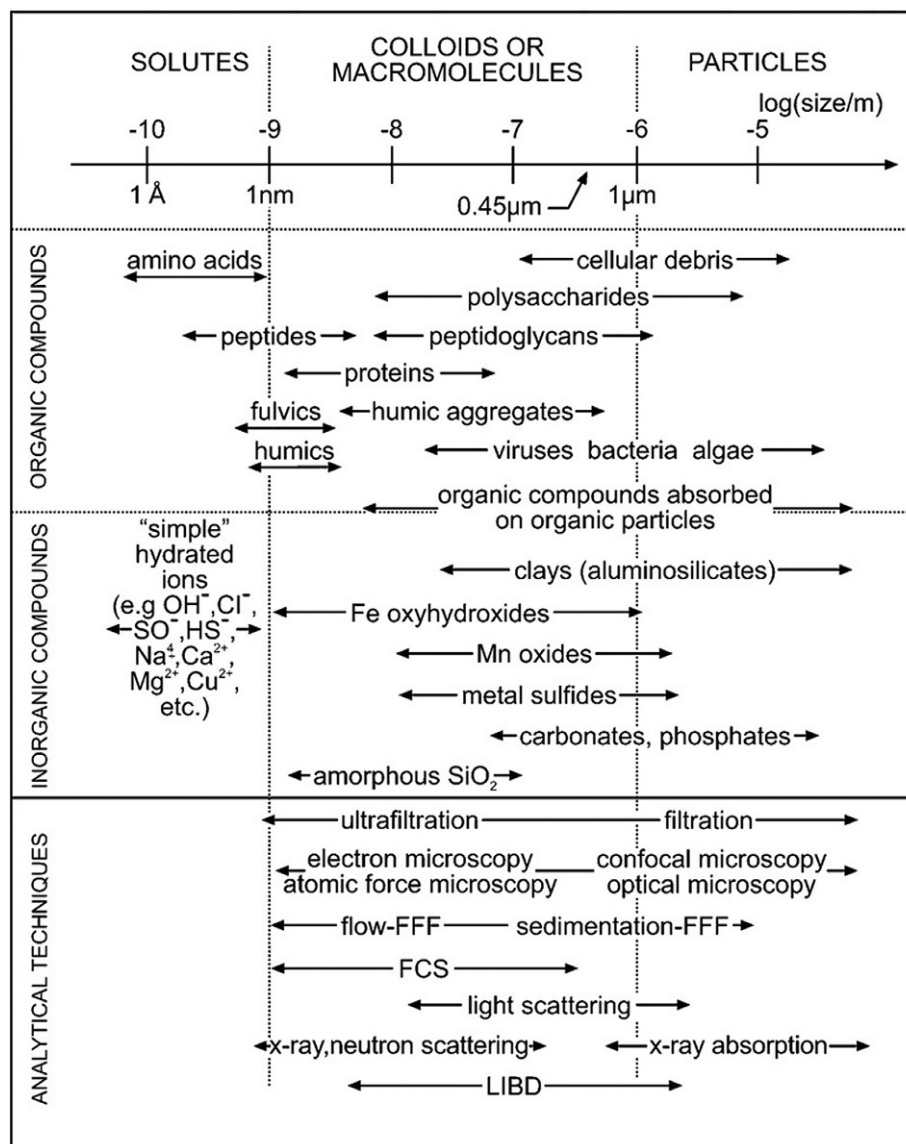


Fig. 1 Size distribution of various types of particles and their characterization techniques (FFF: field-flow fractionation; FCS: fluorescence correlation spectroscopy; LIBD: laser induced breakdown detection) (Klaine et al., 2008). From Klaine SJ, Alvarez PJ, Batley GE, Fernandes TF, Handy RD, Lyon DY, Mahendra S, McLaughlin MJ and Lead JR (2008) Nanomaterials in the environment: behavior, fate, bioavailability, and effects. *Environmental Toxicology and Chemistry* 27(9): 1825–1851. Reprinted with permission from John Wiley & Sons.

The scanning electron microscope's resolution can reach 1 nm, and the magnification can reach 300,000 times or more. Atomic force microscopy is also a powerful technique to study the surface structure of solid materials including insulators. It can detect the very weak interatomic interaction between the sample surface and a micro force sensor. When scanning the sample, the sensor detects the deformation or motion changes of the microcantilever. The surface morphology, roughness, and structural information are obtained with nanoscale resolution based on the information obtained on force distribution.

Techniques for distinguishing natural and engineered nanoparticles

Engineered nanoparticles (ENPs) possess some technically intensified properties not involved in biotic evolution. Therefore, their ecological risks are much stronger than NNPs. It is thus essential to develop analytical approaches to distinguish NNPs and ENPs in complex environmental matrices, especially in soils. The challenges in identifying ENPs in complex environmental matrices include: (1) the low concentration of ENPs (several order of magnitude lower) relative to NNPs; (2) similarity in the chemical compositions between ENPs and NNPs, and (3) the tendency of ENPs and NNPs to hetero-aggregate in the natural environment. A general process of separation and identification of NNPs and ENPs from soil is shown in Fig. 2.

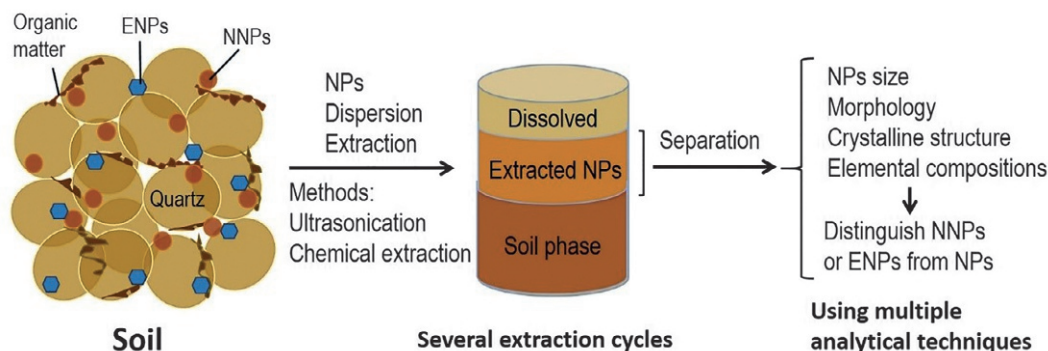


Fig. 2 Flow diagram of separation and identification of natural and engineered NPs from soil.

In soils, the majority of inorganic NNPs form during the weathering of parent materials. Thus, natural nanomaterials are likely to have similar elemental compositions to those of the rocks from which they form. For instance, natural TiO_2 minerals such as rutile and ilmenite are the main carriers (>90–95% of the whole-rock content) for titanium (Ti), Niobium (Nb), antimony (Sb), and tantalum (Ta), and important carriers (5–45% of the whole-rock content) for chromium (Cr), vanadium (V), tin (Sn), and molybdenum (Mo). Therefore, microelement analysis of the co-occurrence of elements is also one of the effective methods to distinguish ENPs and NNPs.

The elemental compositions of NPs may be easily quantified through spectral analysis, such as atomic absorption spectroscopy (AAS), plasma inductively coupled plasma emission spectroscopy (ICP-AES) and X-ray fluorescence spectroscopy (XRF), or mass spectrometry, such as ICP-MS and secondary ion mass spectrometry (SIMS), or energy spectrum analysis represented by X-ray photoelectron spectroscopy (XPS). The ICP-MS is one of the most widely used techniques to quantify inorganic NPs, because of its low concentration detection limits (down to ng L^{-1}). Inductively coupled plasma optical emission spectrometer (ICP-OES) provides detection limits in the $\mu\text{g L}^{-1}$ range, and AAS offers a detection limit in between this and ICP-MS. If the particles are individually separated, the following quantification techniques will be much more powerful for distinguishing between NNPs and ENPs. For example, asymmetric flow field-flow fractionation (AFFF) separates NPs according to their diffusion, which can be converted to equivalent hydrodynamic diameter according to the Stokes-Einstein equation. When combined with ICP-MS, this technique enables size fractionation and online analysis of the elemental composition of NPs. A critical step in the AFFF analyses of soil NPs is their dispersion into a colloidally stable suspension. Single particle ICP-MS (Sp-ICP-MS) is another continuous separation technique assisted by on-line ICP-MS. This joint system provides important information about the number, concentration and element mass content of NPs. The detection limits for Sp-ICP-MS are within the range of 1000 particles mL^{-1} for the number concentration, and 10–20 nm for the size of the mono-element NPs or >100 nm for oxides. The main goal of these techniques is to determine the total element concentration in the sample, but not physicochemical form (dissolved or particulate) or the size, charge or aggregation state.

The previously discussed electron microscopies are based on the analysis of separated particles. The size, shape, aggregation status and surface coating of particles can be directly observed. If other valid techniques could be combined with the microscopic methods, the NPs could also be effectively quantified. For example, the elemental composition (energy dispersive spectrometer, EDS) and phase structure (selected area electron diffraction, SAED) analysis of the particles are valuable complementary measures to the microscopies. Electron microscopy is also an effective method to distinguish ENPs. The HR-TEM with EDS and SEAD functions have been successfully applied to observe the polycrystalline structure of ZnS NPs (Kim et al., 2010). Considering the polycrystalline form of abiogenic ZnS NPs and the single crystals of biogenic ZnS NPs, the mechanisms of formation by abiotic processes could be extensively discussed.

The fate of nanoparticles in soil

Nanoparticles may undergo a series of morphological transformations through biological and abiotic processes in soil, including dissolution, transport, aggregation, dispersion, and aging (Tourinho et al., 2012). These environmental processes will ultimately change the persistence, bioavailability/bio-absorption, reactivity, and toxicity of NPs. Therefore, understanding the behavior of NPs is fundamental for recognizing their environmental risks and applications (Nowack and Bucheli, 2007). It should be noted that processes associated with NPs in water are different from those in soil where NPs may easily contact and interact with other soil components; this will be discussed in detail below. Thus, the dispersion and aggregation of NPs will not be included in this section, and readers are recommended to refer to previous review papers (for example, Zhang et al., 2020).

The dissolution of nanoparticles in the soil system

The dissolution of NPs refers to the process in which chemical components separate from the solid particles and diffuse into the liquid medium to form a homogeneous mixture. The extent of dissolution (solubility) is determined by the solubility of the chemical constituent in a given environment, and the rate of dissolution (kinetics) is determined by the concentration gradient and contact area between the bulk solution phase and particle surface. The kinetics of dissolution for solid NPs can be described by the Noyes-Whitney equation. The concentration of solute components can be calculated from the solubility of the solute in bulk solution, dissolution time, and the dissolution rate constant. The rate constant is positively related to the diffusion coefficient of solute components and the surface area of NPs (Borm et al., 2006), but inversely proportional to the thickness of the diffusion layer.

The effect of pH conditions on the dissolution of NPs is important. The NPs could react with H^+ (hydrogen ion) and promote the release of metal ions, or react with OH^- (hydroxyl group) to generate metallic hydroxyl complexes. Increasing the concentration of H^+ or OH^- in the soil medium will accelerate the dissolution. In general, environmental pH alters the dissolution equilibrium of a complexing agent, the protonation or hydroxylation of the ionic groups released by NP dissolution, and the NP interfacial free energy. Thus, the pH value plays a complex role in NP dissolution, depending on the structures of the complexing agents and the NPs together with their interaction strengths with the solvent.

Dissolved organic matter (DOM) is ubiquitous in the soil solution and may be actively involved in the dissolution of NPs. They contain a large number of oxygen-containing functional groups such as phenolic hydroxyl groups and carboxyl groups, which complex with the released metal cations. According to the Le Chatelier principle of chemical reaction equilibrium, coordination complexation leads to a decrease of the free ion concentration, which consequently increases the concentration gradient and promotes the dissolution of NPs. At the same time, H^+ will be released when DOM is complexed with metal cations, which is also beneficial to the dissolution of NPs as discussed above.

It should be noted that some researchers also found that DOM can inhibit the dissolution of NPs. After the adsorption of DOM on to the NPs, the thickness of the surface diffusion layer of the entire particle increases accordingly, and the dissolution of NPs is inhibited. The apparent effect of DOM on dissolution of NPs therefore depends on the balance between these two opposite effects. However, it is easy to imagine that the interactions between DOM and NPs will depend strongly on the DOM-NP pairs, and it may not be practical to propose a universal parameter, or a DOM concentration range to predict the apparent dissolution of NPs. If we have to choose between the accuracy and feasibility, a direct measurement of the exposed surface area of NPs after DOM sorption may be a useful value to predict NP dissolution.

The dissolution of NPs is also related to biological processes. For example, microorganisms and plants in soil systems may excrete different types of chemicals with abundant H^+ (such as organic acids) or strong redox potential. Furthermore, the chemicals released during the decomposition of microorganisms and plants may also interact with NPs. All the soil organism-related chemicals may act in a similar way to DOM in controlling NP dissolution.

The dissolution of NPs may be controlled by the above properties and processes, individually or simultaneously. A comprehensive model to predict NP dissolution considering the contribution of and even the balance among different processes will be valuable for the management of NPs. A modeling concept of the combination of processes and metal speciation, similar to MINTEQA (Bonito et al., 2018), may be of value for the mathematical simulation of NP dissolution.

The transport of nanoparticles in the soil system

The interaction between NPs and soil particles is the key to determine their transport, which can be described by colloidal filtration theory and is recognized as three main interactions, namely interception, sedimentation, and diffusion (Abbas et al., 2020). Interception occurs mainly in the direction of water flow when the distance between their transport trajectory and the surface of soil particles is shorter than the radius of the NPs. The NPs are attached to the surface of soil particles which settle under gravity. Diffusion is by Brownian motion during which NPs interact with soil particles.

The Derjaguin-Landau-Verwey-Overbeek (DLVO) theory is the most widely used theory to describe the NP-soil interaction. The classic DLVO theory describes Van der Waals force and the electric double layer force between NPs and soil particles. Their interaction energy can be calculated semi-quantitatively. When the distance between NPs and soil particles is smaller than 5 nm, short-range DLVO forces should be considered, including the hydration force, hydrophobic force, and steric resistance force. However, mismatches are usually observed between the theoretical predictions and experimental measurements. For example, DLVO theory considers the particle size and surface charge well, but not the shape. Namely, this model assumes that the particles are spheres, which is not true in most circumstances. The aggregates of spherical NPs may be relatively more compact and generally spherical compared to other shapes, which again is not considered in DLVO theory. The extended DLVO model considers particles of different morphologies well, which is a great step forward in describing the environmental behavior of NPs.

The mobility of ENPs can be quite different in different porous media. For example, nC_{60} was retained by sand (77%) more effectively than by glass beads (8–49%) (Wang et al., 2008), probably because of their different surface charges and roughness. Previous studies have demonstrated that the heavy metal compounds (mainly metal oxyhydroxides) can strongly retain ENPs in their porous structure (Tian et al., 2010). Under saturated conditions when soil pores are filled with water, the subsurface environment is a two-phase space where solid-water interfacial interactions and pore straining mainly control the transport processes of ENPs.

It is interesting to note that NP transport may also change soil properties such as electron conductivity and the microbial community, which are essential to the soil system. In addition, the sorption of various chemicals on NPs is generally over two orders of magnitude greater than on the soil particles (Fang et al., 2009). The transport and subsequent release of these chemicals by NPs in the soil column may result in the depletion of these chemicals in the current soil layer, but may be enriched in the soil layers below. The effect of this co-transport therefore depends on the type of the chemicals. For example, if the transported chemicals are nutrients, the deeper soil layers will in essence be fertilized. However, if a pollutant was carried by the NPs, the risks may be spread more widely.

The aging of nanoparticles in the soil system

The NPs in soil will inevitably be involved in geochemical processes, leading to changes in their physiochemical properties. This process is called the ‘aging’ of NPs, which includes physical, chemical, and biological processes (Mia et al., 2017). Apparently, NP aging may include the previously discussed individual processes, such as dissolution, transport, and their interactions with environmental matrices.

The aging process requires us to consider NPs as dynamic particles with continuously varying physicochemical properties. Since the individual processes will be discussed separately in the chapter, we use this section to emphasize three viewpoints. (1) The change of NP properties may not be monotonic. For example, although dissolution and decreased particle size are involved in the initial process of Au (gold) NP aging, a subsequent increase in particle size may occur because of Ostwald ripening (Gubicza et al., 2013). This ripening process occurs because large particles are favored more strongly than the small particles. This is different from aggregation and demands a special theoretical concept in mathematical modeling. (2) The shape of NPs may be changed following matrix or crystalline reformation. For a macroparticle, the edges may be gradually worn and the particle becomes spherical. However, for NPs, the spherical particles may be changed to bipyramid, decahedron, deca-tetrahedron, triangular plate, and rod shapes after aging (Gubicza et al., 2013). Proper consideration and description of the change in shape of NPs will be essential for modeling NP fate during their aging. (3) The surface redox reaction may accelerate or inhibit the following reaction of NPs. A typical example is nZVI (nonzerovalent iron) NPs. Some of the zerovalent iron atoms on the surface may be oxidized to Fe (III), which greatly weakened or enhanced the reactivity of nZVI NPs (Reinsch et al., 2010). Thus, the environmental roles and toxicity of nZVI NPs both decreased after aging.

According to the above discussion, it is thus easy to understand that the aging of NPs warrants a complete re-evaluation of the behavior and risks of NPs after they are released into the environment. The NPs should be considered extremely dynamic anthropogenic particles, especially when considering their interactions with soil components, as discussed in the following section.

The interactions between nanoparticles and soil components

The assembling of nanoparticles with soil particles

The assembly of NPs specifically refers to a process whereby the particles are organized with a repeating apparent morphology. The assembly of NPs has been studied for inorganic NPs, organic polymers, DNA, or proteins and employed in drug delivery, disease diagnosis and treatment, vaccines, luminescent materials, and flexible electronic devices. The most significant advantage of the assembling of NPs is to tailor their structure and thus their function. The assembling process of NPs also occurs in soil system. For example, natural organic matter (NOM) with non-amphiphilic, amphiphilic, and lipid fractions can form a hierarchical self-assembling system (Chilom et al., 2013). The assembling of NPs may contribute greatly to the generation of stable soil aggregates, in a process known as oriented attachment. The reason is that the assembling of NPs may remove the boundary between oriented NPs, decrease the total surface area and the surface free energy, and inhibit dissolution (Waychunas et al., 2005). Specifically, the heterogeneous assembling of NPs refers to the assembling among different types of NPs or among NPs and macroparticles, which could promote the formation of a more complex or ordered structure because of the typical anisotropic characteristics of NPs. These compactly formed soil aggregates are beneficial for the ecological stability of soil particles. Engineered simulation of this process may help in the development of slow-release fertilizers.

The surface coating of organic matter on nanoparticles

Soil is rich in natural organic matter (NOM), and NPs in the soil ecosystem will inevitably interact with NOM and form an ‘eco-corona.’ Because the sorption of NOM on NPs will significantly change their physical and chemical properties (Grillo et al., 2015), the sorption mechanisms have been widely discussed in previous studies. However, the interactions between NOM and NPs are considered very complicated and the reasons can be understood from the following (Fig. 3):

- (1) The NOM molecules contain various functional groups, such as carboxyl, hydroxyl, quinone, aldehyde, ketone, and sulfhydryl groups. These functional groups can act individually or simultaneously in NOM sorption through various mechanisms of electrostatic interaction, ligand exchange, hydrophobic interaction, hydrogen bonding, and cation bridging. Therefore, the types of NPs should be carefully considered in studying the sorption mechanism.

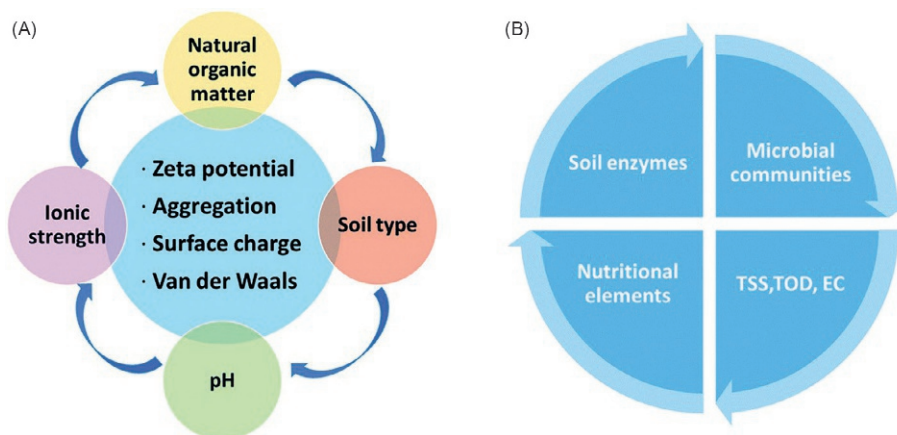


Fig. 3 Schematic summary of the interactions between soil and nanoparticles. (A) Properties in small circles are the soil factors that affect the behavior of nanoparticles (in the large circle); (B) effects of nanoparticles on soil properties. Abbreviations: EC, electricity conductivity; TOD, total oxygen demand; TSS, total suspended solids (Tan et al., 2018). From Tan W, Peralta-Videa JR and Gardea-Torresdey JL (2018) Interaction of titanium dioxide nanoparticles with soil components and plants: current knowledge and future research needs—A critical review. *Environmental Science: Nano* 5: 257–278. Reprinted with permission from Royal Society of Chemistry.

- (2) The sorption of NOM chemical components on NPs is selective, with some specific NOM molecules preferentially adsorbed. Investigations on the sorption selectivity of NOM molecules on common particles have not yet reached a common conclusion (Cai et al., 2015).
- (3) Studies on sorption are generally accompanied by a mathematical fitting of the apparent sorption curves. However, for NPs, the sorption curves do not follow the commonly applied models. Investigators have suggested that NPs may be dispersed after NOM sorption, and thus the sorption is not always thermally stable.
- (4) Considering the significant environmental roles of NPs, it may be more important to understand how NOM sorption affects their dispersion. It should also be noted that the mechanisms of NP dispersion are diverse, including surface coating, formation of micelles, and ‘unzipping’ (Singh et al., 2020). If the dispersion mechanisms by different NOM fractions are revealed, NP dispersion by NOM may be predicted better, or a strategy may be developed to obtain a highly effective dispersion reagent for NPs.

The embedding of nanoparticles in soil aggregates

The embedding of NPs in soil aggregates is different from the heterogeneous assembling of NPs and soil components. The assembling emphasizes the ordered and repeated morphology or structural units, while the embedding process is more like a random physical deposition of NPs in soil aggregates. Soil aggregates are the basic unit of soil structure and enhance aeration and drainage, reduce erosion and runoff, and promote rooting, soil fertility and productivity. The formation of soil aggregates is a comprehensive process that includes physical, chemical, and biological processes. Nanoparticles may be involved in all of these processes through agglomeration, sedimentation, co-precipitation, adsorption, and even biological digestion.

After being embedded in soil aggregates, the reactivity and mobility of NPs will greatly decrease. Because of the organic matter coating and particle blocking, accessibility of the surface of NPs, their dissolution, and surface oxidation are very low. Thus, the aging processes of NPs are greatly slowed. The embedded NPs cannot induce the generation of reactive oxygen species (ROS) and their toxicity is also significantly inhibited. However, it should be noted that the stability of NPs in the soil environment is increased. The release of NPs after the break-down of aggregates should be carefully evaluated for their consequent behavior and potential risks.

The impacts of nanoparticles on soil systems

Nanotechnology has shown its great potential in agricultural applications. However, investigators continue to identify the potential negative impacts of NPs. The debate on the double-edged effects of NPs will accompany the development of nanotechnology.

Beneficial application of nanoparticles in agricultural activities

By 2050, the human population is expected to be 9.8 billion. Food production will thus need to increase by 70% to feed this growing global population, while giving special consideration to the requirement for sustainable development of agricultural

production (Kopittke et al., 2019). Nanomaterials may serve as an efficient fertilizer or pesticide to achieve the goals of increased growth in crop yield and crop quality by improving plant mineral nutrition utilization compared with conventional fertilizer application. This may be possible by balancing the contents of essential macronutrients, mesonutrients, and micronutrients in crops, and by increasing the plant-beneficial soil microorganisms. Nutrients are generally applied to the soil (i.e., for root uptake) or to the plant leaves (i.e., foliar spray) to supplement nutrients in the soil (e.g., N (nitrogen), P (phosphorus), Mg (magnesium), Zn (zinc), Fe (iron), Mn (manganese), Si (silicon), Cu (copper)) or to control pests and diseases for crops. According to the current technology, the nano-fertilizers or pesticides mostly include the following four types.

- (1) Slow release of nano-fertilizers/pesticides. The NPs seem to be effective as a coating for reducing the rate of release and extending the retention time of macronutrient fertilizers or pesticides in the soil. For example, urea-modified hydroxyapatite NPs encapsulated into cavities of the soft wood of *Gliricidia sepium* (Jacq.) Steud showed an initial rapid release of N followed by a slower release in the next 60 days, and improved the yield of rice (*Oryza sativa* L.) more efficiently than conventional fertilizers (Kopittke et al., 2019). Due to the slow and continuous release of the embedded nutrients, NP fertilizers added to the roots can substantially increase both biomass and nutrient uptake compared with traditional fertilizers. Meanwhile, many other nutrients in the soil can be adsorbed on the surface of an NP coating because of the strong force of adsorption. Similarly, many NPs, such as Cu-, Zn- and Ag-based NPs have shown strong performances in controlling bacteria, fungi and insects, which implies the potential of NPs in protecting crops during growth or even storage.
- (2) Stabilized nano-fertilizers/pesticides. Here the stabilization is aimed at preventing fertilizers or pesticides from degradation. For instance, neem (*Azadirachtin indica* A. Juss., 1830) seed oil is known as a powerful natural insect antifeedant and a growth regulator. The effective component of azadirachtin can be easily degraded under the exposure of light, acid or alkaline media or by microorganisms. Using the nano-emulsion/solvent displacement and interfacial deposition method, Forim et al. (2013) synthesized a type of poly (ϵ -caprolactone) NP containing azadirachtin, with a solubility (260 mg L^{-1}), 13 times greater than natural azadirachtin. After 7 days of UV exposure, the photodegradation ratio of these neem NPs decreased from 100% to 10–55%. Similarly, rotenone, a botanical insecticide derived from the roots or rhizomes of tropical legumes, is limited in agricultural use due to its rapid degradation under ultraviolet light, low water solubility and strong toxicity to organisms. Using the N-(octadecanol-1-glycidyl ether)-O-sulfate chitosan nano micelles, Lao et al. (2010) successfully increased the water solubility of rotenone by 13,000 times and prolonged the release of the rotenone by 17 times.
- (3) Nanotechnology-facilitated transgene integration for crops. Recently, nanotechnologies have been developed to improve the stability of DNA during transgene integration for crops. Conjugates between positively charged NPs (e.g., gold NPs) and nucleic acid (e.g. poly-lysine) can overcome the cell wall barrier and be directed into the plant cell. The DNA is protected by NPs from enzymatic attack in the cytosol, but could be released into the nucleus (Moulick et al., 2020). The NP–DNA conjugate in combination with a magnetic field is entitled pollen magnetofection, and can be used to produce a stable transgenic plant even without tissue culture or regeneration (Moulick et al., 2020).
- (4) Fertilizers/pesticides hybrid. During growth and development, crops need a variety of nutrients rather than a single one. Thus, the demand for slow and sustainable release of various nutrients promotes the generation of hybrid nanofertilizers. At present, many researchers have formulated urea-hydroxyapatite nanohybrids for the slow release of nitrogen and phosphorus to plants. Tarafder et al. (2020) further incorporated copper, iron and zinc NPs into the urea-modified hydroxyapatite to enhance plant nutrition and disease resistance of the ladies' finger (*Abelmoschus esculentus* (L.) Moench).

Notably, various environmental factors such as temperature, soil pH and alkalinity may affect the solubility or chemical speciation (e.g., chelated, soluble, and bulk forms) of nano-fertilizers and consequently their availability to crops. In addition, the presence of organic matter may decrease the performance of inorganic components possibly through the formation of complexes or precipitates. Proper consideration of these factors will be beneficial for improving the performance of nano-fertilizers or pesticides in different scenarios.

The uptake of nanoparticles by soil organisms

The reactivity of NPs has brought a general concern that the application of NPs to soils might be involved with negative effects. For example, investigators have reported risks of NPs to soil enzymes, soil microorganisms and even soil animals. The main processes of NPs in the soil–organism system may include the adsorption of NPs to biological surfaces, the uptake and translocation of NPs inside organisms, and their storage in various tissues.

The NPs can be biologically available to soil organisms in two forms, the dissolved ions and the dispersed NPs. The uptake of dissolved ions is well-understood as passive or active transport. For NPs, in particular, their penetration through the bio-abiotic interface has attracted a great deal of research attention. Since the direct uptake of NPs by organisms has been confirmed, their risks should be carefully evaluated, considering their interactions with cell components, the generation of ROS, as well as the pollutants carried. Different techniques have been adopted to observe the penetration of NPs through cell walls and membranes. For example, SEM and TEM are the most commonly used instruments for observing the morphology of NPs outside or inside organisms. Using synchrotron-radiation micro-X-ray fluorescence, the crystalline phase and size of NPs can be identified after the internalization. Single particle-inductively coupled plasma-mass spectrometry and ^{14}C labeled methods were also successful to quantify the contents of metals or ^{14}C in plant tissues for metal-based NPs and carbon nanotubes, respectively.

It is easy to understand that NPs need to be dispersed before uptake. The NPs may be dispersed by engineered reagents, or natural chemicals such as DOM as discussed above. The dispersed NPs with sizes between 10 and 50 nm may enter biological cells via pinocytosis, while those larger than 100 nm may be absorbed by phagocytosis.

Nanoparticles as a two-blade sword to the soil system

As shown above, NPs have a great potential as fertilizers, pesticides or their additives for increasing crop production and to reduce adverse environmental problems caused by excess application of traditional fertilizers or pesticides. However, studies concerning the negative effects of NPs on soil organisms have increased year by year. Improper application of NPs may decrease the length of root, shrink root tips, result in considerable vacuolation or collapse of root epidermal and cortical cells (Nair and Chung, 2014), and reduce the abundance and diversity of soil bacterial and fungal communities (Liu et al., 2021). It is of great importance to establish strategies to overcome the negative impacts and preserve the beneficial effects of NPs to the soil system (Fig. 4). The strategies may include:

- (1) Control NPs at a safe concentration range. Almost all chemicals will show negative effects on organisms after exceeding a certain concentration. In addition, the varying environmental factors (e.g., pH values, electrolytes, temperature and organic matters) may also alter the threshold values through affecting the physicochemical properties of NPs. Full understanding of the underlying mechanisms of the biological effects of NPs should be the first step in proposing a threshold for a specific group of organisms under certain conditions. Continuous monitoring of the metabolites (e.g., siderophores) from plant roots in response to environmental stresses may also help to identify the potential risks of the system. Establishment of a database on detailed toxicity or ecotoxicity for different individuals (e.g., researchers, policy makers or product R&D people) may be the next step in the safe application of NPs.
- (2) Match positive NP-organism pairs. Because of differences in the sensitivity, tolerance and detoxification mechanisms of different organisms to even the same type of NPs, the threshold varies across different species of organisms. Positive effects of NPs may be achieved because of increased utilization efficiencies, or even through the 'hormesis' effect of NPs at low concentrations. Identifying the positive NP-organism pairs requires an extensive investigation of NP and organism combinations, which is time-consuming, costly, and labor-intensive. Modeling (such as hormetic dose-response, biotic ligand model, SAR or QSAR models) may be useful for calculating some critical values, such as concentrations producing no or small observed effects, and those that have half the maximal effect.
- (3) Use degradable NPs as carriers. Some matrixes or compositions of the nano-sized carriers may have antibacterial or anti-insect properties, and may change the microbial community in the soil with adverse effects on plant-microbe interactions. The investigation of biodegradable or photodegradable substances (e.g., chitosan or polymeric nanocarriers) can be one of the trends to develop nano-fertilizers or -pesticides in the near future.

Summary

The whole story regarding NPs in soil is about the balancing between their negative and positive effects. It cannot be over-emphasized that both NNPs and ENPs are dynamic particles, with varying reactivities during their aging, transport and interaction with soil components. Well-balanced NP properties are not only essential for nanotechnology management, but are also useful for nanotechnology applications in sustainable agriculture.

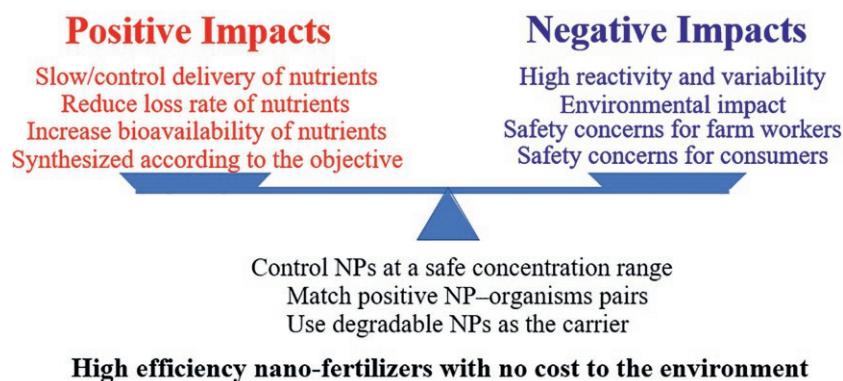


Fig. 4 The balancing between the positive and negative effects of nanoparticles.

Acknowledgments

This research was supported by the National Natural Scientific Foundation of China (41725016 and 41977334), a joint fund between NSFC-NCN (41961134002), Yunnan Provincial Scientific Innovation Team of Soil Environment and Ecological Safety (2019HC008), and the Scientific Research Fund of Yunnan Provincial Education Department (2016CYH06).

References

- Abbas Q, Yousaf B, Amina AMU, Munir MAM, El-Naggar A, Rinklebe J, and Naushad M (2020) Transformation pathways and fate of engineered nanoparticles (ENPs) in distinct interactive environmental compartments: A review. *Environment International* 138: 105646.
- Bonito MD, Lofts S, and Groenenberg JE (2018) Models of geochemical speciation: Structure and applications. In: Vivo BD, Belkin HE, and Lima A (eds.) *Environmental Geochemistry*, 2nd edn, pp. 237–305. Elsevier. ch. 11.
- Borm P, Klaessig FC, Landry TD, Moudgil B, Pauluhn JR, Thomas K, Trotter R, and Wood S (2006) Research strategies for safety evaluation of nanomaterials, part v: Role of dissolution in biological fate and effects of nanoscale particles. *Toxicological Sciences* 90: 23–32.
- Cai N, Peak D, and Larese-Casanova P (2015) Factors influencing natural organic matter sorption onto commercial graphene oxides. *Chemical Engineering Journal* 273: 568–579.
- Chilom G, Baglieri A, Johnson-Edler CA, and Rice JA (2013) Hierarchical self-assembling properties of natural organic matter's components. *Organic Geochemistry* 57: 119–126.
- Colvin VL (2003) The potential environmental impact of engineered nanomaterials. *Nature Biotechnology* 21: 1166–1170.
- Fang J, Shan XQ, Wen B, Lin JM, and Owens G (2009) Stability of titania nanoparticles in soil suspensions and transport in saturated homogeneous soil columns. *Environmental Pollution* 157: 1101–1109.
- Forim MR, Costa ES, Das Gracas Fernandes Da Silva MF, Fernandes JB, Mondego JM, and Boica Junior AL (2013) Development of a new method to prepare nano-/microparticles loaded with extracts of *azadirachta indica*, their characterization and use in controlling *plutella xylostella*. *Journal of Agricultural and Food Chemistry* 61: 9131–9139.
- Grillo R, Rosa AH, and Fraceto LF (2015) Engineered nanoparticles and organic matter: A review of the state-of-the-art. *Chemosphere* 119: 608–619.
- Gubicza J, Lábar JL, Quynh LM, Nam NH, and Luong NH (2013) Evolution of size and shape of gold nanoparticles during long-time aging. *Materials Chemistry and Physics* 138: 449–453.
- Kim B, Park CS, Murayama M, and Hochella MF (2010) Discovery and characterization of silver sulfide nanoparticles in final sewage sludge products. *Environmental Science & Technology* 44: 7509–7514.
- Klaine SJ, Alvarez PJ, Batley GE, Fernandes TF, Handy RD, Lyon DY, Mahendra S, McLaughlin MJ, and Lead JR (2008) Nanomaterials in the environment: Behavior, fate, bioavailability, and effects. *Environmental Toxicology and Chemistry* 27(9): 1825–1851.
- Kopittke PM, Lombi E, Wang P, Schjoerring JK, and Husted S (2019) Nanomaterials as fertilizers for improving plant mineral nutrition and environmental outcomes. *Environmental Science: Nano* 6: 3513–3524.
- Lao SB, Zhang ZX, Xu HH, and Jiang GB (2010) Novel amphiphilic chitosan derivatives: Synthesis, characterization and micellar solubilization of rotenone. *Carbohydrate Polymers* 82: 1136–1142.
- Liu Y, Li Y, Pan B, Zhang XY, Zhang H, Steinberg CEW, Qiu H, Vijver MG, and Peijnenburg W (2021) Application of low dosage of copper oxide and zinc oxide nanoparticles boosts bacterial and fungal communities in soil. *Science of the Total Environment* 757: 143807.
- Mia S, Dijkstra FA, and Singh B (2017) Long-term aging of biochar: A molecular understanding with agricultural and environmental implications. In: Sparks DL (ed.) *Advances in Agronomy*. Academic Press. ch. 1.
- Moullick RG, Das S, Debnath N, and Bandopadhyay K (2020) Potential use of nanotechnology in sustainable and 'smart' agriculture: Advancements made in the last decade. *Plant Biotechnology Reports* 14: 505–513.
- Nair PMG and Chung IM (2014) Impact of copper oxide nanoparticles exposure on *Arabidopsis thaliana* growth, root system development, root lignification, and molecular level changes. *Environmental Science and Pollution Research* 21: 12709–12722.
- Nowack B and Bucheli TD (2007) Occurrence, behavior and effects of nanoparticles in the environment. *Environmental Pollution* 150: 5–22.
- Pan B and Xing B (2010) Chapter three—Manufactured nanoparticles and their sorption of organic chemicals. *Advances in Agronomy* 108: 137–181.
- Reinsch BC, Forsberg B, Penn RL, Kim CS, and Lowry GV (2010) Chemical transformations during aging of zerovalent iron nanoparticles in the presence of common groundwater dissolved constituents. *Environmental Science & Technology* 44: 3455–3461.
- Singh RP, Sharma K, and Mausam K (2020) Dispersion and stability of metal oxide nanoparticles in aqueous suspension: A review. *Materials Today: Proceedings* 26: 2021–2025.
- Tan W, Peralta-Video JR, and Gardea-Torresdey JL (2018) Interaction of titanium dioxide nanoparticles with soil components and plants: Current knowledge and future research needs—A critical review. *Environmental Science: Nano* 5: 257–278.
- Tarafder C, Daizy M, Alam MM, Ali MR, Islam MJ, Islam R, Ahommed MS, Aly MAS, and Khan MZH (2020) Formulation of a hybrid nanofertilizer for slow and sustainable release of micronutrients. *ACS Omega* 5: 23960–23966.
- Tian Y, Gao B, Silvera-Batista C, and Ziegler KJ (2010) Transport of engineered nanoparticles in saturated porous media. *Journal of Nanoparticle Research* 12: 2371–2380.
- Tourinho PS, Van Gestel CAM, Lofts S, Svendsen C, Soares AMVM, and Loureiro S (2012) Metal-based nanoparticles in soil: Fate, behavior, and effects on soil invertebrates. *Environmental Toxicology and Chemistry* 31: 1679–1692.
- Tratnyek PG and Johnson RL (2006) Nanotechnologies for environmental cleanup. *Nano Today* 1: 44–48.
- Wang Y, Li Y, Fortner JD, Hughes JB, Abriola LM, and Pennell KD (2008) Transport and retention of nanoscale C60 aggregates in water-saturated porous media. *Environmental Science & Technology* 42: 3588–3594.
- Waychunas GA, Kim CS, and Banfield JF (2005) Nanoparticulate iron oxide minerals in soils and sediments: Unique properties and contaminant scavenging mechanisms. *Journal of Nanoparticle Research* 7: 409–433.
- Zhang D, Qiu J, Shi L, Liu Y, Pan B, and Xing B (2020) The mechanisms and environmental implications of engineered nanoparticles dispersion. *Science of the Total Environment* 722: 137781.

Soil chemistry and climate change

Jianming Xu, Yong Li, and Zhongmin Dai, Institute of Soil & Water Resources and Environmental Science, College of Environmental and Resource Sciences, Zhejiang Provincial Key Laboratory of Agricultural Resources and Environment, Zhejiang University, Hangzhou, China

© 2023 Elsevier Ltd. All rights reserved.

Effects of soil chemistry on greenhouse gas emissions	692
Effects of soil chemistry on carbon dioxide emissions	692
Effects of soil chemistry on methane emissions	693
Effects of soil chemistry on nitrous oxide emissions	694
Impact of climate change on soil chemistry	695
Effects of warming on soil chemistry	695
Effects of drying on soil chemistry	696
Effects of precipitation on soil chemistry	697
Effects of wildfire on soil chemistry	697
Concluding remarks and future research	698
References	698

Abstract

Greenhouse gases (GHGs) alter the Earth's climate by absorbing energy in the lower atmosphere and re-emitting it. Anthropogenic emissions of carbon dioxide (CO₂), methane (CH₄) and nitrous oxide (N₂O) contribute most to global warming. Soil is a major source and sink of GHGs. Here, the impact of soil chemistry, including soil organic matter, pH and nutrients on CO₂, CH₄ and N₂O emissions and the feedbacks of climate change, including warming, drying, precipitation and wildfires on soil chemistry are discussed.

Key points

- Soil chemistry significantly affects GHGs emissions.
- Climate changes also influence soil chemical properties.

Effects of soil chemistry on greenhouse gas emissions

Effects of soil chemistry on carbon dioxide emissions

Carbon dioxide (CO₂) is an important heat-trapping (greenhouse) gas, which is released through human activities such as deforestation and burning of fossil fuels, as well as natural processes such as respiration by organisms and volcanic eruptions. Since the beginning of the industrial era (1750), fossil fuel use, expansion of cultivated agriculture and forest clearance have led to an increase in atmospheric CO₂ from ~280 ppm to the current level of ~420 ppm (μmol mol⁻¹) in 2022. This change in atmospheric CO₂ concentrations is much larger than historic natural fluctuations over the past 20,000-year period (from the Last Glacial Maximum to 1750, from approximately 185 to 280 ppm). These elevated levels of CO₂ absorb and reradiate long-wave radiation, thereby retaining heat in the lower atmosphere and causing global warming. Carbon dioxide is the most important of the Earth's [long-lived greenhouse gases](#) contributing about two-thirds of the current [total energy imbalance](#). The average 2020 surface temperature was 0.98 °C warmer than that of the twentieth-century, and with temperature increases nearly double that in polar regions.

The present increase in atmospheric CO₂ is caused by anthropogenic emissions of CO₂. Land-use and land-use change are responsible for approximately 11% of all anthropogenic CO₂ emissions. Soils are a major source and sink of CO₂, hence approximately 10% of the atmosphere's CO₂ passes through terrestrial soils each year. The source and sink dynamics of CO₂ from soils depend on several environmental factors, such as ecosystem net primary productivity (carbon inputs), available soil water, soil temperature, soil organic matter, soil mineralogy and chemistry, and nutrient availability. Herein, we discuss the effects of soil chemistry, including soil organic matter, pH, and plant nutrients, on CO₂ emissions.

Soils are of particular importance in the atmospheric CO₂ budget as they store more organic carbon (1500 Gt) than the atmosphere (750 Gt) and plants (560 Gt) combined. Changes in the size of the soil C storage pool therefore can significantly affect atmospheric CO₂ concentrations. Soil organic carbon has decadal to centennial turnover times. Relatively inert (stable or recalcitrant) soil organic carbon is composed of molecules that are more or less resistant to further decomposition. A very small fraction of soil organic matter, and a small fraction of burnt biomass, are converted into inert forms. Natural processes and soil management practices may reduce or increase the amount of carbon stored in soil pools with turnover times on the order of tens to hundreds of years (e.g., black carbon—charcoal and biochar) and thus influence the evolution of atmospheric CO₂ over time. Conversion of natural vegetation to agriculture is a major source of CO₂, not only due to losses of plant biomass, but also increased decomposition

of soil organic matter caused by soil disturbance, and energy use associated with various agricultural practices. Conversely, the use of high yielding plant varieties, fertilizers, irrigation, residue management and reduced tillage can alleviate losses and enhance C uptake within managed areas. These processes have led to an increase of soil carbon and decrease of CO₂ emissions in some agricultural soils, especially those in arid and semiarid regions where irrigation greatly increases net primary productivity. Organic amendments may also contribute to increased CO₂ emissions. Despite all the benefits that organic amendments contribute to soil restoration, there is tremendous variability among the types of amendment. There is also considerable variability in organic matter composition and chemistry which can result in highly variable rates of organic matter decomposition, thereby altering the rates of CO₂ emission. The impact on CO₂ emissions was greater when applying organic amendments from pig manure compared with residue amendments from maize (*Zea mays* L.) cultivation.

Soil microbial activity has considerable effects on soil CO₂ emissions. Microbial activity is strongly influenced by soil pH. Therefore, management practices such as liming influence soil emissions by altering biological activity. In addition, CO₂ is released by the liming of acidic soils:

Alkaline soils have significantly higher CO₂-to-soil organic carbon ratios, and consequently soil organic carbon in alkaline soils appears less chemically stable with respect to microbial mineralization. Soil respiration decreased markedly from acid soil samples leading to lower CO₂ emissions.

In addition to soil pH, soil nutrient availability is an important factor regulating ecosystem net primary productivity (NPP), soil microbial activities and soil CO₂ emissions. Nitrogen (N) and phosphorous (P) are key nutrients that limit crop growth and microbial activities. Nitrogen availability is an important constraint on NPP at the global scale, although phosphorus and calcium may be more important limiting nutrients in many tropical and subtropical regions. Reactive nitrogen from burning of fossil fuels and biomass combustion (NO_x) and ammonia emitted by industry, animal husbandry and fertilizer use can act as a fertilizer for terrestrial plants thereby stimulating NPP. Nitrogen fertilization may enhance the formation of humified soil organic matter, which increases the residence time of carbon and decreases CO₂ emissions from soils. Conversely, nitrogen additions were shown to increase soil CO₂ efflux in a field experiment in Costa Rica, whereas N addition had no net effect or decreased soil CO₂ effluxes in China. It is generally acknowledged that P rather than N is the growth-limiting nutrient in many lowland tropical forests. Phosphorus additions to several tropical forests resulted in a strong increase in microbial biomass leading to an increase in the soil CO₂ efflux. The incorporation of crop residues (as opposed to burning) and application of inorganic fertilizers, green manures and animal wastes can be of great importance in maintaining soil fertility and hence ecosystem NPP. These practices can provide essential nutrients for agricultural crops and coupled with reductions in the burning of crop residues can reduce CO₂ emissions from soils to the atmosphere.

Effects of soil chemistry on methane emissions

Methane (CH₄), the second-most important greenhouse gas after carbon dioxide (CO₂), has contributed almost one quarter of the increase in cumulative radiative forcing for the major greenhouse gases (CO₂, CH₄, and N₂O) since 1750. Methane is also a precursor for tropospheric ozone (O₃)—both an air pollutant and greenhouse gas—influencing ozone background levels. Although methane is far less abundant in the atmosphere than CO₂, it absorbs infrared radiation more efficiently and, as a consequence, has a global warming potential (GWP) ~86 times stronger per unit mass than CO₂ on a 20-year timescale and 28-times more powerful on a 100-year basis. The atmospheric methane (CH₄) concentration reached 1895 nmol mol⁻¹ (ppb) in 2021, approximately 2.6 times greater than its estimated pre-industrial value of 722 ppb in 1750. The global network of surface observations over the past 3–4 decades indicates that methane went through a period of rapid growth from the 1980s to 1990s, almost stabilized from 1999 to 2006, and then renewed its rapid growth to the present. The overall increase in CH₄ concentration is attributable in large part to increased anthropogenic emissions arising primarily from agriculture (e.g., livestock production, rice (*Oryza sativa* L.) cultivation, biomass burning), fossil fuel production and use, waste disposal in landfills, and changes in natural methane fluxes due to climate change (e.g., melting of permafrost).

Global emissions from agriculture and waste disposal are estimated at 195 Tg CH₄ yr⁻¹, representing ~57% of total anthropogenic CH₄ emissions. Atmospheric methane concentrations are a balance between production and removal processes. Methane is removed from the atmosphere by reaction with the hydroxyl radical (·OH) in the troposphere, destruction by reactions with excited atomic oxygen and atomic chlorine (Cl) in the stratosphere and uptake and utilization in soils by methanotrophic bacteria (CH₄ oxidizers). Biogenic methane is the final product of the decomposition of organic matter by Archaea in anaerobic environments, such as water-saturated soils, wetlands, rice paddies, marine sediments, landfills, waste-water facilities, or inside animal intestines. In contrast, unsaturated oxic soils maybe a sink for atmospheric methane due to the presence of methanotrophic bacteria, which consume methane as a source of energy in which they oxidize the CH₄ to CO₂. Wetlands with temporally variable saturation can act as either a methane source or sink. Several soil factors influence methane emissions from soils, such as redox potential, alternative electron acceptors, substrate availability, temperature, diffusion, water availability, soil pH and salinity, fertilizer and manure additions. Below we discuss the impact of several soil chemistry parameters on methane emissions.

Soil organic carbon is an important component of the global carbon cycle and soil organic carbon dynamics strongly affect methane emissions. Readily mineralizable organic matter derived from primary production or organic amendments is the main electron donor resulting in methane formation in soils. Anaerobic conditions are necessary to facilitate the biochemical pathways of CH₄ production owing to the anaerobic nature of the methanogens responsible for CH₄ production by the process known as methanogenesis. Hydrogenotrophic (CH₄ production from H₂ and CO₂) and acetoclastic (CH₄ production from acetate as the

electron acceptor) methanogenesis are the most common pathways for CH_4 production in soils. During flooding of rice fields, the soil redox potential declines progressively as electron acceptors are consumed along the descending electron tower (e.g., O_2 , NO_3^- , Mn^{4+} , Fe^{3+} , SO_4^{2-} , CO_2). As soil organic matter is decomposed during periods of inundation, it releases electrons that facilitate the methanogenic process. Rice plants supply the microbes in soil and rhizosphere with organic substrates: about 30–60% of net photosynthate carbon is allocated to roots, and as much as 40–90% of this fraction is translocated into the soil as root exudates. Increasing C-substrate availability by the input of organic substances to the soil through root exudation and litter production stimulate methanogenesis and methane production. In contrast, a large portion of the CH_4 produced under anaerobic conditions may be oxidized in aerobic environments (e.g., water column of paddy fields or wetlands) before escaping to the atmosphere by the activity of methane-oxidizing bacteria, which utilize methane as a carbon and energy source. In this case, the organic matter does not directly affect methane-oxidizing bacteria, but does affect methane formation (i.e., methanogens) and other microbes in the environment, which have implications for the functioning of methane-oxidizing bacteria. The addition of organic materials, such as crop residues, to soils affects CH_4 oxidation in different ways, often depending on the C:N ratio. When the C:N ratio of the residues is low (i.e., high N content), CH_4 oxidation generally decreases. The increased release of ammonium increases ammonia oxidation, which inhibits the oxidation of CH_4 . Conversely, when the organic residues have a high C:N ratio, the CH_4 oxidation rate is not appreciably affected.

Soil pH is a crucial factor for regulating methane production in wetland rice soils in addition to the carbon supply and water regime. The optimal pH range for methanotrophic activity in soils is wide; there are reports of methanotrophic activity in acid (< 3.5) and alkaline (pH 9.5) soils. Soils generally have low methane production under acidic conditions. Moreover, the highest rates of CH_4 oxidation have been found in soils with near neutral pH conditions (~ 7.0). In general, nitrogen (N) and phosphorus (P) limitations are not frequently observed in methanogenesis. Methane production can be greatly attenuated when urea and manure are used as fertilizers. In flooded soils, CH_4 emissions were lower when ammonium sulfate was used as a fertilizer source compared to urea because of competition between sulfate-reducing bacteria and methanogens as electron acceptors. Fertilization with nitrate also decreases CH_4 emissions as the NO_3^- preferentially serves as an electron acceptor during the denitrification process. Notably, ammonium sulfate significantly reduced CH_4 emissions compared to the application of potassium nitrate in paddy fields.

The activity of methane-oxidizing bacteria largely depends on the presence of methane, therefore any factor that affects methane formation will consequently have implications for methane-oxidizing bacteria. Populations of methanotrophs in soils are affected by several environmental factors, most notably temperature, pH, type and concentration of N sources, and the concentrations of CH_4 and O_2 . The extent of the nutrient effects depends on whether Type I or Type II methane oxidizers predominate. Type I strains predominate in nutrient-rich environments, whereas the Type II strain can survive in environments where N is scarce. Methane monooxygenase requires copper (Cu) as a cofactor, and the amount of copper necessary for optimal activity varies within the community of methanotrophs. Overall, Type II methanotrophs are more active than Type I when Cu concentration is low and vice versa. While there are no conclusive results concerning potassium availability on methanotrophic activity, one report demonstrated a stimulating effect from the addition of potassium.

Effects of soil chemistry on nitrous oxide emissions

Nitrous oxide (N_2O) is the third most important greenhouse gas, after carbon dioxide (CO_2) and methane (CH_4), with a global warming potential ~ 298 times that of carbon dioxide on a 100-year time scale. It is also a potent ozone-depleting substance in the stratosphere, catalyzing the destruction of ozone in the stratosphere and contributing to the annual ozone hole in the polar regions. Ice-core measurements indicate that N_2O concentration in the atmosphere was nearly constant at about 270 ppbv in the 11,000 years prior to the beginning of the industrial era. Since the start of the industrial revolution, N_2O has increased from ~ 280 to 334 ppbv in 2021. Globally, soil ecosystems constitute the largest source of N_2O emissions, comprising approximately 65% of total N_2O emissions to the atmosphere. Specifically, 62% is derived from nitrogen fertilization and indirect emissions, 31% arises from manure management and 7% results from biomass burning. Almost all microbes and pathways involved in the biogeochemical cycling of nitrogen have the potential to produce N_2O as an intermediary product, and these processes are interrelated and interact to share intermediate products. These N_2O -generating processes include ammonia (hydroxylamine) oxidation, nitrifier denitrification, nitrite oxidation, heterotrophic denitrification, anaerobic ammonium oxidation (anammox) and dissimilatory nitrate reduction to ammonium (DNRA, or nitrate ammonification).

The key pathways involving N_2O production and consumption include nitrification and denitrification, with each of these processes mediated by specialized microbial assemblages. In the nitrification process, ammonium is first oxidized to nitrite, which is subsequently oxidized to nitrate. The N_2O is produced as a byproduct of the nitrification process. The first and rate-limiting step of the nitrification process, ammonia oxidation is carried out by ammonia oxidizing bacteria (AOB), archaea (AOA), or the recently discovered complete ammonia oxidation (comammox) bacteria. Nitrous oxide is also produced from denitrification, a process by which nitrate is progressively reduced to nitrite, nitric oxide, N_2O and N_2 under anaerobic conditions. The major abiotic determinants regulating N_2O production are soil oxygen levels, water content, pH, temperature, nitrate and nitrite concentrations, organic carbon content and micronutrient contents. Below, we discuss the impact of soil organic matter, pH and plant nutrients on nitrous oxide emissions.

Soil organic matter comprises the largest carbon stock in agroecosystems and plays a very important role in the global carbon and nitrogen cycles. The C:N ratio represents an important criterion for evaluating the quality of humus, and it also directly affects

N₂O emissions from soils. High soil N concentrations, especially those following fertilization, may enhance microbially-mediated N transformation processes leading to high N₂O emissions. The optimal C:N ratio of soil organic matter is considered to be about 10:1, whereas the ideal C:N ratio for compost is ~30 as the elevated carbon level provides sufficient food availability for microorganisms. Viewed globally, soils are the central link in the organic matter biotransformation chain with significant gaseous C and N end-products. Several studies report that N₂O fluxes increased after the application of residues, with low C:N residues releasing more N₂O than residues with high C:N ratios. The addition of organic substances has a positive effect on soil nitrification and the autotrophic nitrifier population. Similarly, organic amendments significantly increased the abundance of AOA and stimulated the growth of AOB in agricultural and riparian zone soils. The amount of organic C, as a substrate for bacteria metabolism, is a primary determinant regulating the N₂:N₂O ratio of denitrification. Large concentrations of labile soil organic carbon increase microbial activity resulting in the consumption of N₂O to produce N₂, thereby increasing the N₂:N₂O ratio.

Soil pH is another important factor influencing N₂O emissions. Several studies have demonstrated that the abundance and structure of nitrogen-cycling genes, and the rates of nitrification and denitrification, are strongly regulated by soil pH. Soil pH is a vital factor influencing the niche differentiation of the autotrophic nitrifiers AOA, AOB and comammox bacteria. The AOA generally dominate ammonia oxidation in acidic soils and either AOA or AOB may dominate ammonia oxidation in slightly acid to neutral pH soils. In contrast, the role of pH on the distribution of the recently discovered comammox bacteria are unknown. While the contribution of ammonia oxidation to N₂O emissions remains ambiguous, a meta-analysis of 100+ isotope dilution studies determined a slightly negative correlation between gross nitrification rates and soil pH. Autotrophic nitrification and the resulting N₂O flux are limited at soil pH values less than 4.5. Soil pH also affects the N₂:N₂O ratio, with N₂O/(N₂ + N₂O) having a significant negative relationship with soil pH within the normal range (pH = 5–8) of agricultural soils. Notably, a global meta-analysis of field experiments revealed that the amount of N₂O substantially increased in soils with lower pH values. A possible mechanism for the positive effect of acidity on N₂O emissions was attributed to the inhibition of nitrous oxide reductase through the disruption of enzyme assembly in the periplasm.

All forms of N inputs to agricultural soils, such as inorganic N fertilizer and organic N sources in the form of manures, slurry, legumes or post-harvest crop residues, represent a potential contribution of N substrates for subsequent N₂O emissions. The influence of fertilizer on N₂O emissions is a function of the fertilizer type (e.g., organic, NH₄⁺ or NO₃⁻), the amount, and timing of application. A meta-analysis revealed a greater loss of N₂O from urea than from (NH₄)₂SO₄ (ammonium sulfate) fertilizer treatments. The greater is the amount of ammonium in the fertilizer, the greater is the potential for nitrification, which is a source of N₂O emissions. Furthermore, the loss of N₂O can increase because the nitrite formed during the nitrification process can be used as an electron acceptor during denitrification, if O₂ is limited. Thus, both nitrification and denitrification are sources of N₂O under aerobic vs anaerobic conditions, respectively. Emissions of N₂O are greater when nitrate concentration in the soil is high. However, several studies indicate that large soil nitrate concentrations inhibit the reduction of N₂O to N₂, thereby resulting in a higher yield of N₂O. Soil nitrate concentrations as high as 277 µg N g⁻¹ soil inhibited N₂O reductase activity, which reduced the conversion of N₂O to N₂ and resulted in lower N₂:N₂O ratios in sandy and silt loam soils. In addition, denitrification enzymes require a variety of metal cofactors, including Mo (molybdenum), Fe (iron), Cu and Zn (zinc). The absolute requirement of nitrous oxide reductase (the enzyme that reduces N₂O to N₂) for Cu cofactors accounts for the critical role of this micro-nutrient in regulating the denitrification process. Thus, several soil chemical properties regulate N₂O emissions from soils, with important implications for agricultural and environmental management.

Impact of climate change on soil chemistry

Effects of warming on soil chemistry

Global warming is expected to lead to an increased global surface temperature of between 2.1 °C and 3.5 °C over the 21st century, due to greenhouse gas emissions from human activities. Warming significantly affects plant growth, organism activity and rates of chemical reactions making soil chemistry highly responsive to warming. Soil organic C, arguably the most important component in soils, comprises 1500 Gt of carbon stored in soils, which is twice that contained in the atmosphere. Warming increases plant photosynthesis and water-use efficiency, which has the potential to enhance CO₂-fixation from the atmosphere and increase C storage in soils. On the other hand, increased plant C inputs to soil may create a soil priming effect, thereby enhancing soil organic C decomposition and CO₂ emission back to the atmosphere. Therefore, several interacting feedbacks may offset the increased soil organic C from plant C inputs making conclusions somewhat tenuous.

Several global-scale studies contend that the size of the overall soil organic C pool will not change appreciably, even though soil C fluxes and biogeochemical C cycling are expected to accelerate (Kuzakov et al., 2019). Soil organic C consists of diverse compounds that have simple (e.g., glucose) to complex (e.g., lignin) chemical structures, which generally correspond to a gradient of labile to recalcitrant C compounds. The temperature sensitivity of organic carbon decomposition and feedbacks to warming depend on the degree of warming and the chemical composition of the organic compounds. In general, higher temperatures are expected to cause greater decomposition of soil organic C by a combination of biotic and abiotic processes. In addition to greatly affecting the rate of decomposition of soil labile C, warming is reported to alter the rate of decomposition of more recalcitrant C compounds, thereby leading to a positive feedback in response to climate change (Frey et al., 2013). Subsoil horizons contain more than half of globally stored soil organic C, which is considered to be relatively stable under warming due to physical and chemical stabilization mechanisms arising from interactions with the soil colloidal fraction. However, in some specific ecosystems, such as

alpine grasslands, subsoil organic C is also sensitive to warming. Root exudate inputs to subsoil horizons are associated with deepening root systems that could potentially increase soil organic C fractions, such as lipids and sugars, and decrease other C fractions, such as phenols and lignin (Jia et al., 2019). To assess ecosystem and soil carbon storage rigorously in response to climate change, it is necessary to critically examine plant carbon allocation (above-ground vs below-ground) and the dynamics of subsoil organic carbon pools.

Soil N biogeochemical cycling is intimately linked to soil C cycling, as the soil C:N stoichiometry is well constrained to a narrow range across various terrestrial ecosystems. Hence, the response of soil organic C to warming has a similar pattern to that of organic N; however, warming also affects soil inorganic N cycling and N availability. Effects of warming on soil N transformations are important because forest and agricultural ecosystems often suffer from N limitation. As mentioned previously, increased plant growth with warming will increase plant demand for available N, which may accelerate N limitation in terrestrial ecosystems, termed progressive nitrogen limitation. A recent evaluation of global warming effects on N-cycling revealed a shift in microbially-regulated processes from N immobilization to enhanced N mineralization, nitrification and denitrification (Dai et al., 2020). The increased potential for nitrification and denitrification pose a further risk for the loss of inorganic N (e.g., leaching, runoff and N_2/N_2O emissions) from terrestrial ecosystems, again accelerating N limitation in terrestrial ecosystems.

Warming also affects soil chemical properties including pH and mineral nutrient availability, such as K^+ (potassium), Ca^{2+} (calcium) and Mg^{2+} (magnesium). Organic acids and H^+ secreted by plant roots under warmer conditions can decrease soil pH, despite an increase in net primary production. Warming also affects rates of mineral weathering and nutrient mobilization by affecting plant root and microbial pathways (e.g., through respiration and root extension). Moreover, increased CO_2 concentrations accelerate mineral weathering and the release nutrient cations, such as K^{2+} , Ca^{2+} and Mg^{2+} . Furthermore, root extension may contribute to mineral weathering by physical and biochemical pathways. However, the impact of higher temperature on mineral weathering is relatively weak compared to other climate change events, such as precipitation, drying and fire. The nutrient elements released from primary minerals can also be quickly taken up by plants, thereby offsetting the increase in soil mineral nutrient pools due to enhanced chemical weathering.

Effects of drying on soil chemistry

Climate change may also induce significant drying effects in some geographical regions. Drying is a serious threat to ecosystem and agricultural productivity, and ecosystem stability, and pronounced drying may lead to widespread desertification. The percentage of land worldwide subjected to drying will rapidly increase as atmospheric temperatures rise. Under warming, the intensity, duration and frequency of droughts are rapidly increasing. The drylands are mainly distributed in the western United States, north and east Africa, Australia, and central Asia, and cover over ~45% of Earth's land surface. In humid regions, drying may cause a shift in ecosystems, such as transformation of forests to woodlands to grasslands to deserts, resulting in a large loss of biodiversity.

Drying has several prominent effects on soil chemical properties related to soil moisture, such as redox potential, salinity, alkalinity and pH. Drying causes greater oxygen concentrations and would be expected to reduce CH_4 production, denitrification and sulfate reduction in soils. Soil salinization and alkalization often occur in drylands and poorly managed irrigated lands because salts are inadequately leached or concentrate in the rooting zone due to capillary rise of water from subsoil or groundwater sources. Evaporation and transpiration further concentrate soluble salts (e.g., Na^+) in surface soils inducing soil salinization and alkalization. The high ionic strength of saline soils hinders plant availability of water due to the low osmotic potential of saline waters. Soils experiencing salinization or alkalization generally have high soil pH (pH > 8.5), which reduces the availability of several trace nutrient metals (e.g., Fe, Mn (manganese), Cu, Zn).

Ecosystem and soil drying may decouple the C, N and P biogeochemical cycles and change soil organic C, N and P pools and fluxes. Drying usually reduces primary ecosystem production as moisture becomes the limiting plant growth factor in many arid and semi-arid climates. Decreased plant cover and associated soil biological processes result in decreased returns of plant litter and root exudates to soils, causing a decrease in soil organic matter. Drying also reduces the activity of soil microorganisms and small fauna that control soil nutrient cycling. Inhibition of soil microbial activity by moisture limitation will decrease microbially-regulated organic C decomposition and nutrient mineralization. While the contribution of microorganisms to soil organic C is less than plant C inputs, the low C:N ratio of microbial biomass has a disproportional effect on nutrient cycling. Similarly, soil drying may have a negative effect on total soil N concentration as biological processes, such as photosynthesis, decomposition and N fixation, are very low compared to humid regions. The decrease in soil organic C and total N pools with climate drying becomes more prominent with increasing intensity, duration and frequency of drought, increasing the potential for a collapse of ecosystems into a state of desertification.

Drying favors the dominance of physical processes, e.g., rock weathering, leading to the release of rock-derived nutrients, especially P, K, Ca and Mg (Delgado-Baquerizo et al., 2013). With water-limited plant growth, plant uptake of nutrients is very low, which also leads to the accumulation of inorganic P and mineral nutrients in arid-region soils. Thus, drying may have a positive effect on the availability of inorganic P and base cations, whereas it has a negative effect on soil organic C and N concentrations. Soil micronutrients such as iron (Fe), manganese (Mn), copper (Cu) and zinc (Zn) play an important role in organism growth and ecosystem function. A recent study found that globally, except for Antarctica, the availability of soil Fe, Mn, Cu and Zn in drylands is less than that in humid-region soils (Moreno-Jiménez et al., 2019). Drying decreases the availability of many micronutrients by increasing soil pH. The precipitation of micronutrients with -OH are more prominent in soils with high pH rendering them less

available to plants. The change in micronutrient availability in soil due to global warming induced drying may further exacerbate food production and security, and the diversity of plants and soil organisms in arid and semi-arid ecosystems worldwide.

Effects of precipitation on soil chemistry

Heavy precipitation has appreciable effects on soil chemistry that depend on the frequency and magnitude of extreme rainfall events. A direct effect of heavy precipitation on soil chemistry is the alteration of soil moisture related properties, such as plant nutrient availability, soil pH and redox potential. Soil base cations (e.g., K^+ , Na^+ , Ca^{2+} , Mg^{2+}) constitute a source of soil alkalinity and pH buffering capacity. Heavy precipitation increases water infiltration causing enhanced leaching of base cations that may result in soil acidification (decreasing soil pH). A decrease in soil pH will have a corresponding effect on nutrient availability, such as decreased base cations, potentially increased micronutrients and inorganic P, which has optimum availability in the pH range of 6–7. In addition, higher precipitation decreases soil N availability by increased nitrate leaching, whereas phosphate that is predominantly bound to soil particles is lost by increased soil runoff and erosion. Heavier rainfall events can cause serious soil erosion, especially in arid regions where lack of soil cover cannot protect the topsoil from raindrop impact which destroys surface soil structure and reduces infiltration. Soil erosion removes the surface soil, which has the highest organic matter, nutrient and biological diversity and activity, thereby causing considerable degradation in soil quality and productivity. Further, the nutrients and organic matter eroded from soils accumulate in surface waters where they contribute to eutrophication.

If the magnitude of precipitation exceeds the hydraulic conductivity of various soil horizons, the soils may become water saturated leading to a lack of oxygen and low redox potential. Reduced redox potential under water-saturation affects the speciation of several soil elements, influencing nutrient cycling, nutrient availability and elemental toxicity. The reduction of nitrate, sulfate, Fe, Mn, metal(loids) (e.g., As (arsenic) and Cr (chromium)), methane production and anaerobic decomposition of soil organic matter tend to increase after heavy precipitation. Chemical weathering is generally increased by “excess” precipitation and leaching, thereby leading to increased clay formation. Soil desiccation favors the formation of crystalline clay minerals at the expense of noncrystalline or nanocrystalline colloids, which has profound effects on soil mineral reactivity with respect to sorption reactions and soil organic matter stabilization.

In several regions of the world, increased precipitation will result in increased acid deposition to ecosystems and soils. Acid deposition contains large concentrations of sulfuric and nitric acids, and ammonium that can oxidize to nitric acid, resulting in pH values often in the 4.0–4.5 range, but as low as 2.0. Acid deposition enhances soil acidification through H^+ inputs into soils, which at pH values less than 5 solubilizes aluminum that can become toxic to sensitive plants. The increased amount of acidic cations (e.g., H^+ and Al^{3+}) facilitates the exchange of base cations (e.g., Ca^{2+} and Mg^{2+}) from soil colloids, and with the charge balancing anions (e.g., SO_4^{2-} and NO_3^-) accelerates the leaching of the exchangeable base cations. The increased concentrations of H^+ and Al^{3+} and decreased concentrations of base cations limit soil productivity. In general, soil has a buffering capacity (e.g., exchangeable bases) that offsets the acidity from acidic deposition by neutralizing such inputs. The buffering capacity varies with soil type, with a strong dependence on the amount and type of clay minerals and organic matter content. For example, soils dominated by permanently charged minerals (2:1 clay minerals) and large organic matter content are more resistant to acid deposition (i.e., more strongly buffered) than soils dominated by variable charge minerals (e.g., 1:1 clay minerals and iron–aluminum (hydr)oxides) and small organic matter concentrations. In addition, acid deposition has less effect on alkaline than acid soils due to their inherent alkalinity (acid-neutralizing capacity). Finally, the acid inputs from deposition can increase mineral weathering compared to non-polluted precipitation resulting in increased clay formation and the release of rock-derived nutrients.

Effects of wildfire on soil chemistry

Climate change increases the risk of hot and dry weather that can result in prolonged droughts. This dramatically increases the intensity and frequency of wildfire events in forest ecosystems, which are of prime importance for water and soil protection, species conservation and carbon sequestration. Currently, about 50–85 million hectares of global forestlands burn each year due to wildfires (Alonso-Canas and Chuvieco, 2015). Plant biomass, leaf litter and a portion of organic matter at the soil surface are consumed by high-temperature burning and converted into CO_2 , H_2O and other gases, and it also generate a layer of ash composed largely of mineral materials at the soil surface. The ash layer typically has a very high pH (>8–9) and a high content of soluble cations and anions that includes several plant nutrients (e.g., P, Ca, K, etc.).

As fire intensity and temperature increase, soil and plant organic carbon undergoes transformations from slight distillation to carbonization and pyrolysis to complete oxidation. A significant loss in organic carbon content starts at 200–250 °C, with progressively larger losses up to a temperature of ~460 °C (Giovannini et al., 1988). Thus, it is typical to find a remarkable loss in ecosystem and soil carbon stocks because of high-intensity fires. Incomplete combustion generates recalcitrant products, such as black carbon, which is relatively stable compared to soil organic C compounds. In fact, low intensity fires may actually increase the soil organic C content by introducing incompletely burnt plant materials (e.g., charcoal) into the soil.

Similarly, wildfire leads to a significant loss of soil organic N by conversion into gaseous N_2 , NH_3 , N_2O and NO_x as a result of volatilization, and in some cases N limitation may be found in post-fire forest ecosystems. A global meta-analysis indicated that wildfire reduces soil C and N in surface mineral soil layers with the losses depending on forest ecosystem types and fire characteristics (Pellegrini et al., 2018). Wildfire affects soil organic C and N when soil temperatures exceed ~200 °C during combustion, but normally the temperatures only become appreciably elevated in the upper 2–3 cm of the soil (Alexis et al., 2007). Elevated soil

Table 1 Wildfire effects on soil pH, organic C, total N, total P and micronutrients.

Soil chemical property	Fire effects and mechanisms
Soil organic C	Wildfire volatilizes plant biomass, litter residues and soil organic C leading to a decrease in soil organic C; Incomplete combustion generates recalcitrant products, such as black carbon and charcoal; Increased C and N pools after lower intensity fires are observed because of the incorporation of unburned or partially unburned slash fragments into the soil layer
Soil pH, base cations and electrical conductivity	Wildfire burns much of organic biomass and leaves soluble minerals in the ash layer thereby increasing soil pH, base cations (K, Na, Ca, Mg) and electrical conductivity
Soil total N	Similar to organic C, soil organic nitrogen is converted to N_2 , NH_3 , N_2O and NO_x as a result of volatilization. This may lead to N limitation in forest ecosystems following wildfires
Soil total P	Mineral P is not easily volatilized, but rather accumulates in the ash layer leading to a significant increase of inorganic total and available P concentrations. The increase of pH in highly acidic soils further increases PO_4^{3-} availability
Soil total micronutrients	Soil micronutrients, such as Fe, Mn, Cu, Zn, are not easily volatilized and therefore accumulate in ash layer. Changes in soil pH may either increase or decrease micronutrient availability depending the optimum pH for each micronutrient's availability

temperature mainly affects the soil organic layer with temperatures reaching only 50 °C in the mineral layer, even when temperatures in the organic layer exceed 500 °C (Trammell et al., 2004). As a result, only small changes in mineral soil carbon and nitrogen pools generally occur, even following high intensity fires.

Forest fires have several effects on total and available P pools, converting the organic pool of soil P to orthophosphate, a process termed pyromineralization. Mineral P is not easily volatilized and therefore accumulates in the ash layer at the soil surface, leading to a significant increase of inorganic P concentrations (Butler et al., 2018). Since the peak in P availability occurs around pH 6.5, fire-induced changes in soil pH, especially in acid soils, has a positive effect on soil P availability. Although wildfire results in an enrichment of available P, leaching and runoff may cause a loss of P from forest ecosystems after fire events. In tropical and subtropical regions, the fire-released P is immobilized by soil Fe and Al (hydr)oxides leading to a decrease in P availability to plants. Overall, wildfires have a destructive effect on forest ecosystems and several soil chemical properties, and the negative effects increase as fire intensity and frequency increase (Table 1). The short-term effects of wildfire are significantly larger than other climate change effects, such as warming, drying and extreme precipitation events. The recovery of vegetation and soil properties may take several decades, especially if high temperatures and drought conditions further hinder forest ecosystem recovery.

Concluding remarks and future research

Soil chemistry such as soil pH, nutrient availability and organic matter affect the emissions of greenhouse gases such as CO_2 , CH_4 and N_2O . These changes cause a variety of climate and weather-related events: warming, drying, precipitation and fire. In turn, the events give considerable or slight feedback to soil chemical properties in different ways. Here, we point out some future research that will help us to understand global warming effects and climate change management. Climate change usually occurs concurrently and therefore soils suffer from multiple pressures. However, we know little about the reaction of soil chemistry to multiple climate change factors which reveal the real-world condition. We lack a unified perspective of the variation in soil chemistry at global, continental or supracontinental scale, covering diverse climates, vegetation and soil types, as every soil on the earth is affected by global warming. Correspondingly, the emissions of CO_2 , CH_4 and N_2O under climate change and their deterministic factors at the global scale require systematic investigation. Global warming may last for decades or centuries, while most findings are based on short-term experiments, e.g., microcosms and field trials (<10 years). Long-term effects of climate changes on soil chemistry based on longer-term field observations or prediction models will have important implications for the maintenance of ecosystem function, food production and sustainable management.

References

- Alexis M, Rasse R, Rumpel C, Bardoux G, Pechot N, Schmalzer P, Drake B, and Mariotti A (2007) Fire impact on C and N losses and charcoal production in a shrub oak ecosystem. *Biogeochemistry* 82: 201–216.
- Alonso-Canas I and Chuvieco E (2015) Remote sensing of environment global burned area mapping from ENVISAT-MERIS and MODIS active fire data. *Remote Sensing of Environment* 163: 140–152.
- Butler OM, Elser JJ, Lewis T, Mackey B, and Chen C (2018) The phosphorus-rich signature of fire in the soil-plant system: A global meta-analysis. *Ecology Letters* 21: 335–344.
- Dai Z, Yu M, Chen H, et al. (2020) Elevated temperature shifts soil N cycling from microbial immobilization to enhanced mineralization, nitrification and denitrification across global terrestrial ecosystems. *Global Change Biology* 26: 5267–5276.
- Delgado-Baquerizo M, Maestre FT, Gallardo A, et al. (2013) Decoupling of soil nutrient cycles as a function of aridity in global drylands. *Nature* 502: 672–676.
- Frey SD, Lee J, Melillo JM, and Six J (2013) The temperature response of soil microbial efficiency and its feedback to climate. *Nature Climate Change* 3: 395–398.
- Giovannini G, Lucchesi S, and Giachetti M (1988) Effect of heating on some physical and chemical-parameters related to soil aggregation and erodibility. *Soil Science* 146: 255–261.
- Jia J, Cao Z, Liu C, et al. (2019) Climate warming alters subsoil but not topsoil carbon dynamics in alpine grassland. *Global Change Biology* 25: 4383–4393.

- Kuzyakov Y, Horwath WR, Dorodnikov M, and Blagodatskaya E (2019) Review and synthesis of the effects of elevated atmospheric CO₂ on soil processes: No changes in pools, but increased fluxes and accelerated cycles. *Soil Biology and Biochemistry* 128: 66–78.
- Moreno-Jiménez E, Plaza C, Saiz H, Manzano R, Flagmeier M, and Maestre FT (2019) Aridity and reduced soil micronutrient availability in global drylands. *Nature Sustainability* 2: 371–377.
- Pellegrini AFA, Ahlström A, Hobbie SE, et al. (2018) Fire frequency drives decadal changes in soil carbon and nitrogen and ecosystem productivity. *Nature* 553: 194–198.
- Trammell TLE, Rhoades CC, and Bukaveckas PA (2004) Effects of prescribed fire on nutrient pools and losses from glades occurring within oak-hickory forests of central Kentucky. *Restoration Ecology* 12: 597–604.

Encyclopedia of SOILS IN THE ENVIRONMENT

Volume Two: Chemistry

EDITORS-IN-CHIEF

Michael J Goss

Professor Emeritus, School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G 2W1, CANADA

Margaret Oliver

Visiting professor, Soil Research Centre, Department of Geography and Environmental Science, Russell Building, University of Reading, Whiteknights, Reading, RG6 6DR, United Kingdom.

SECTION EDITORS

Dr Evgenia Blagodatskaya

Dept. of Soil Ecology, Helmholtz-Zentrum für Umweltforschung GmbH – UFZ, Theodor-Lieser Str. 4, D-06120 Halle (Saale), GERMANY

Professor Adrian Unc

Boreal Ecosystems and Agricultural Sciences, Grenfell Campus, Memorial University of Newfoundland, 20 University Drive, Corner Brook, Newfoundland A2H 5G4, CANADA

Dr. Siobhán Staunton

INRA, UMR Eco&Sols, Montpellier SupAgro, bâtiment 12, 2 Place Viala, 34060 Montpellier Cedex 2, Montpellier, FRANCE

Professor Baoshan Xing

University of Massachusetts – Amherst, College of Natural Sciences, Stockbridge School of Agriculture, Environmental & Soil Chemistry, Program, 410 Paige Laboratory, Amherst, MA 01003, USA

Dr Karin Mueller

New Zealand Institute for Plant & Food Research Ltd., Ruakura Research Centre, Private Bag 3123, Hamilton, New Zealand

Professor Christina Siebe

Instituto de Geología, Circuito de la Investigación Científica s/n, Ciudad Universitaria, Ciudad de México, CP 04510, México

Professor Rupert Baeumler

Institute of Geography, Friedrich-Alexander-University of Erlangen-Nuremberg, Wetterkreuz 15, 91058 Erlangen, Germany

Dr Uta Stockmann

Soil Processes and Function Organisational Unit, CSIRO Agriculture and Food, Black Mountain, GPO BOX 1700, Canberra ACT 2601, AUSTRALIA

Professor Paul Hallett

Room 2.23, Cruickshank Building, School of Biological Sciences, University of Aberdeen, Aberdeen, AB24 3UU, UNITED KINGDOM

Professor Dani Or

Department of Environmental Systems Science (D-USYS), Institute of Biogeochemistry and Pollutant Dynamics (IBP), Soil and Terrestrial Environmental Physics (STEP), CHN F 29.1 Universitaetstrasse 16, 8092 Zurich, SWITZERLAND



ELSEVIER

elsevier.com/books-and-journals

Set ISBN

ISBN 978-0-12-822974-3



9 780128 229743

Volume ISBN

ISBN 978-0-443-15209-2

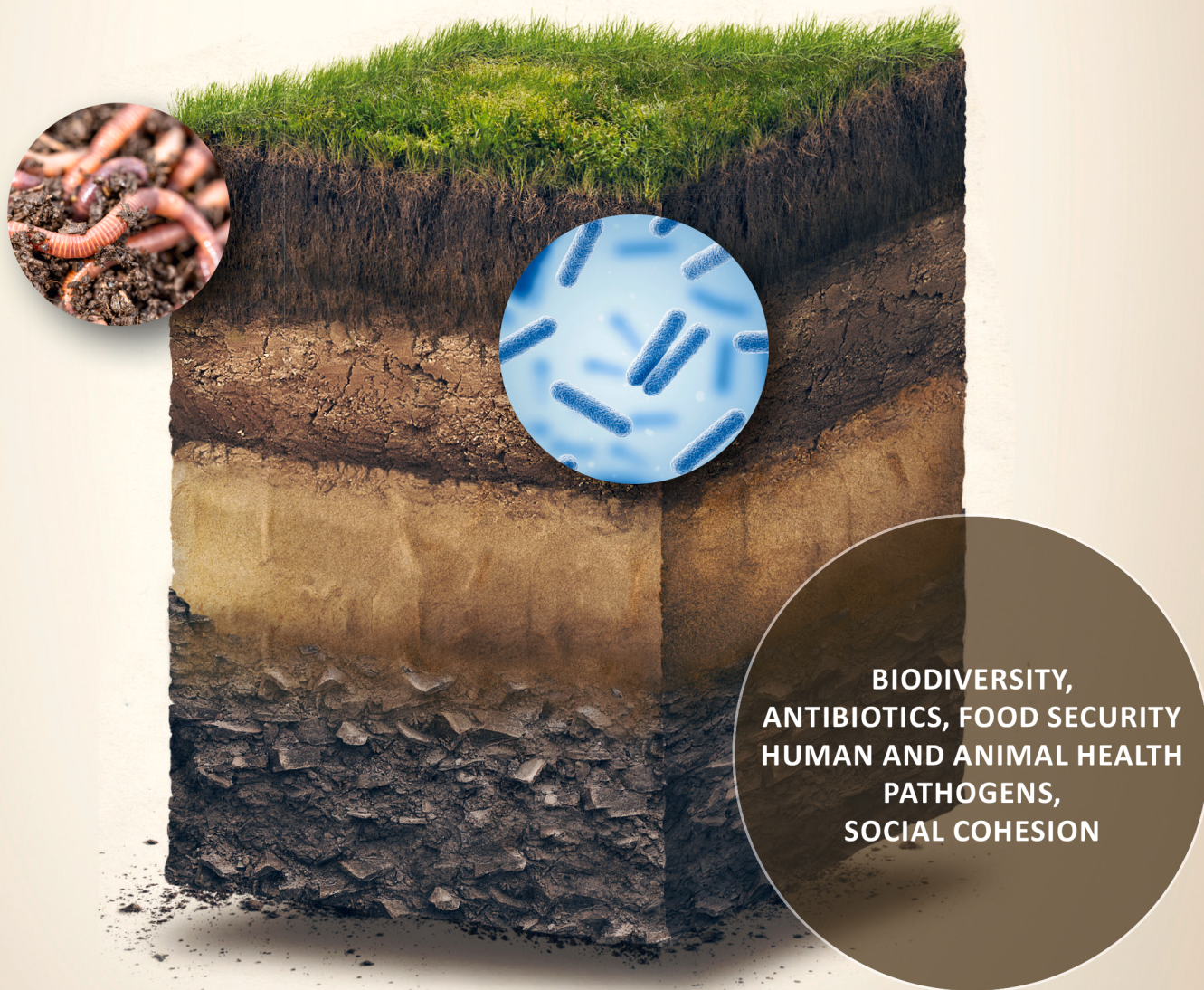


9 780443 152092

Second Edition

Encyclopedia of SOILS IN THE ENVIRONMENT

Volume Three: Environment and Management



Editors-in-Chief

Michael Goss and Margaret Oliver

Section Editors

Karin Müller & Christina Siebe

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

SECOND EDITION

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

SECOND EDITION

EDITORS-IN-CHIEF

Professor Michael J Goss

*Professor Emeritus, School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G
2W1, CANADA*

Professor Margaret Oliver

*Visiting Professor, Soil Research Centre, Department of Geography and Environmental Science,
Russell Building, University of Reading, Whiteknights, Reading, RG6 6DR, United Kingdom*

VOLUME 3

ENVIRONMENT AND MANAGEMENT

SECTION EDITORS

Karin Mueller

*New Zealand Institute for Plant & Food Research Ltd., Ruakura Research Centre, Private Bag 3123,
Hamilton, New Zealand*

Christina Siebe

*Instituto de Geología, Circuito de la Investigación Científica s/n, Ciudad Universitaria, Ciudad de
México, CP 04510, México*



ELSEVIER

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge MA 02139, United States

Copyright © 2023 Elsevier Ltd. All rights reserved

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

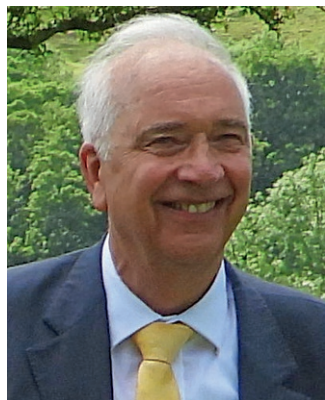
ISBN 978-0-12-822974-3

For information on all publications
visit our website at <http://store.elsevier.com>



Publisher: Oliver Walter
Acquisition Editor: Priscilla Braglia
Content Project Manager: Sona Chahal
Associate Content Project Manager: Nandhini Mahendran
Designer: Christian Bilbow

EDITORS BIOGRAPHY



Michael J Goss, PhD, MSc, BSc, is a professor emeritus in the School of Environmental Sciences at the University of Guelph. Before taking on the joint editor-in-chief responsibilities with Margaret Oliver, he held a similar position for *Soil Use and Management*, a journal of the British Society of Soil Science of which he is an honorary fellow. He was an inaugural member of the European Society for Agronomy. Michael started his career at the Letcombe Laboratory before moving to Rothamsted Experimental Station and the Macaulay Land Use Research Institute (now the James Hutton Research Institute). In 1990 he was appointed to the Research Chair in Land Stewardship at the University of Guelph and in 2005 became the director of its Regional Campus near Ottawa.

Michael's research has covered soil plant relations, soil management, and agriculture's footprint on the environment. His focus has been on the physical properties of soil and their impacts on root growth, which he showed to be influenced by the role of the root cap. He broadened his approach to include the effects of tillage and land drainage on root exploration. This work stimulated his interest in the transport of materials, such as plant nutrients, pathogenic microbes, antibiotics, and biologically active compounds through the soil and their contamination of water resources. Michael's knowledge in this field resulted in his involvement in the public inquiry into the death of seven people, including children, from a microbially contaminated drinking water supply. He contributed, both to the investigation into how the contamination occurred and if any fault could be attributed to local agricultural practices, and to how to prevent future events.

Michael also developed an interest in the interrelationships between soil microbes and root systems, particularly the tripartite symbiosis between roots of legumes, mycorrhizal fungi, and rhizobia. He has been an energetic member of collaborative research programs in France and Germany, and in Portugal, where he continues his interest in the role of mycorrhiza in sustainable land management.

Michael has authored more than 180 publications including two books, one in its 2nd edition, and edited two others.



Margaret Ann Oliver (BSc, PhD)

Margaret's interest in soil began in her penultimate year at school when learning about different types of soil and their formation. She chose the University of Bristol, United Kingdom for her first degree because the geography course offered soil science and she was also able to do a degree in geology alongside. Her love for both subjects remains undiminished. She did her PhD in soil spatial analysis at the University of Birmingham. Her early research focused on the multivariate and geostatistical analysis of soil data, which she continued to use in various research projects. Her research interests expanded to include the application of a wide range of numerical and geostatistical methods to soil and other data including pollen counts, forestry, radon emission, remotely sensed imagery, precision agriculture, and the incidence of childhood cancer. Her specific interests have been in sample design for spatial analysis and eventual mapping, risk analysis associated with prescribed thresholds in relation to soil pollutants or deficiencies of crop nutrients, and the relations between different scales of spatial variation. Her most

recent research was in the field of precision agriculture. At both the Universities of Birmingham and Reading, she taught topics in soil science, statistical methods, and spatial analysis. At Reading she had a research group working on various projects such as imagery, sampling and mapping, precision agriculture, and human health.

She has published more than 100 academic papers and has made several contributions to books. She is the co-author of three books: *Statistical Methods in Soil and Land Resource Survey*, Oxford University Press (Webster, R and Oliver, M. A. 1990); *Geostatistics for Environmental Scientists*, Second Edition, John Wiley & Sons (Webster, R and Oliver, M. A. 2007), and *How to do the Basic Steps in Geostatistics: The Variogram and Kriging*, Springer, Dordrecht (Oliver, M.A. & Webster, R.2015). She has edited two books: *Geostatistical Applications for Precision Agriculture* (Springer, Dordrecht, 2010) and *Precision Agriculture for Sustainability and Environmental Protection* (Earthscan from Routledge, 2013).

Margaret retired officially from Reading in 2004, but as yet has not managed to escape from academia. She has retained a link and a base at the University as a visiting professor since January 2005. After retirement, she developed and taught geostatistics courses to outside groups for several years. She then started on the long path of editorial work, as an associate editor of the *European Journal of Soil Science* and eventually the editor-in-chief, assistant editor of *Mathematical Geosciences*, co-editor of *Precision Agriculture* with John Stafford, and finally the editor-in-chief of the *Encyclopedia of Soils in the Environment* with Michael Goss.

Margaret was made an honorary member of the British Society of Soil Science in 2021 to reflect her commitment to soil science through fellow research, teaching, and editorial work. She was honored by the Pedometrics Commission of the International Union of Soil Science by having an award named after her in 2015. It is the Margaret Oliver award for young pedometricians.

EDITORIAL BOARD

Editors in Chief

Michael J Goss

Professor Emeritus, School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G 2W1, CANADA

Margaret Oliver

Visiting Professor, Soil Research Centre, Department of Geography and Environmental Science, Russell Building, University of Reading, Whiteknights, Reading, RG6 6DR, United Kingdom

Section Editors

Dr Evgenia Blagodatskaya

Dept. of Soil Ecology, Helmholtz-Zentrum für Umweltforschung GmbH – UFZ, Theodor-Lieser Str. 4, D-06120 Halle (Saale), GERMANY

Professor Adrian Unc

Boreal Ecosystems and Agricultural Sciences, Grenfell Campus, Memorial University of Newfoundland, 20 University Drive, Corner Brook, Newfoundland A2H 5G4, CANADA

Dr. Siobhán Staunton

INRA, UMR Eco&Sols, Montpellier SupAgro, bâtiment 12, 2 Place Viala, 34060 Montpellier Cedex 2, Montpellier, FRANCE

Professor Baoshan Xing

University of Massachusetts – Amherst, College of Natural Sciences, Stockbridge School of Agriculture, Environmental & Soil Chemistry, Program, 410 Paige Laboratory, Amherst, MA 01003, USA

Dr Karin Mueller

New Zealand Institute for Plant & Food Research Ltd., Ruakura Research Centre, Private Bag 3123, Hamilton, New Zealand

Professor Christina Siebe

Instituto de Geología, Circuito de la Investigación Científica s/n, Ciudad Universitaria, Ciudad de México, CP 04510, México

Professor Rupert Baeumler

Institute of Geography, Friedrich-Alexander-University of Erlangen-Nuremberg, Wetterkreuz 15, 91058 Erlangen, Germany

Dr Uta Stockmann

Soil Processes and Function Organisational Unit, CSIRO Agriculture and Food, Black Mountain, GPO BOX 1700 , Canberra ACT 2601, AUSTRALIA

Professor Paul Hallett

*Room 2.23, Cruickshank Building, School of Biological Sciences, University of Aberdeen, Aberdeen, AB24 3UU,
UNITED KINGDOM*

Professor Dani Or

*Department of Environmental Systems Science (D-USYS), Institute of Biogeochemistry and Pollutant Dynamics (IBP),
Soil and Terrestrial Environmental Physics (STEP), CHN F 29.1 Universitaetstrasse 16, 8092 Zurich, SWITZERLAND*

FOREWORD

Soil is the porous “skin” on the surface of the Earth that has formed over thousands of years of weathering, comprising minerals, organisms (both dead and alive), water, and air. Soil is more than just the sum of its parts. It serves several crucial ecological functions, from food and fiber production to water storage and purification to waste decomposition.

Increasingly we understand that everything is connected, and that soil is a “Critical Zone” that intersects with the hydrosphere (water), lithosphere (minerals), atmosphere (air), and biosphere (all living organisms). To quote Daniel Hillel, the 2012 World Food Prize laureate and editor of the First Edition of *The Encyclopedia of Soils in the Environment*, “[Soil] is the crucible of terrestrial life, within which biological productivity is generated and sustained...Our civilization depends on the soil...” He also promoted the importance of understanding soil to manage it effectively. This revised and reorganized second edition of the Encyclopedia aims to reflect the changing status of soil in the world. Within these five volumes, over 724 experts from many countries have provided updated, easy-to-understand chapters on soil, its properties, and its role in the environment to affect climate change, plant productivity, human health, and biodiversity.

Many technological advances have been made since the first edition of this encyclopedia that has allowed us to investigate soil physical, chemical, and especially biological properties more precisely, enabling us to identify and understand their roles in ecosystem services. The study of soils, especially in the environmental context, is critical to meeting the challenges of feeding, clothing, and cleaning up the world. The information presented in these volumes will be useful to many, including soil and environmental scientists, policymakers, land managers, educators, students, and anyone interested in learning more about soil.

Soil is under considerable threat from many sources—building and infrastructure, climate change, population pressure, and pollution, yet it is vital to agronomic productivity and continues to provide critical ecosystem services. Society needs to be engaged in how soil is managed and conserved. Most natural scientists are very dedicated and generous, including all of the authors in this Encyclopedia who have freely given their time to make these volumes so comprehensive. The study of soil, while often challenging, is fascinating—soil is a beautiful earthy material that can delight our senses with its variegated colors, gritty to smooth textures, and especially its characteristic aroma when freshly turned. Perhaps that is why gardeners and even some farmers tend to overwork their soil to its detriment—it *just feels good to get your hands in the dirt!*

Soil science is one of the most exciting, interesting, and rewarding disciplines, largely because it is multi-dimensional and interdisciplinary, basic and applied. It affects all of our lives. Thus, I encourage you to peruse these volumes and discover all kinds of interesting aspects of soil in the environment.

April Ulery

Former President, Soil Science Society of America
Professor, Soil & Environmental Sciences
New Mexico State University
Las Cruces, NM
The United States

PREFACE

This second edition of the *Encyclopedia of Soils in the Environment* follows in the footsteps of a highly rated first edition overseen by Professor Daniel Hillel as editor in chief, ably assisted by six editors. That edition provided a very valuable resource on the “state-of-the-art” in soil science in the context of the environment in the closing years of the twentieth century and the first three years of the current century. It set a standard that the editorial team for the second edition has aimed to match. The soil and environmental sciences have made huge strides in the almost 20 years since the first edition was published, and the new team has had to embrace these developments while retaining the basic components of these sciences. The project has taken 4 years to complete. It began with the appointment of the two editors-in-chief, professors Michael J. Goss and Margaret A. Oliver, followed by that of the nine subject editors and a pedology adviser.

The project was in its infancy (a few months only) when COVID-19 struck. The editors-in-chief discussed the format and structure of the second edition around the time that the first “lockdowns” around the world were being introduced. The first edition was alphabetically organized, whereas we decided to have a section-oriented approach in this edition, which we hope gives a more structured feel for the various topics embraced. The principal sections are soil biology, soil chemistry, the soil environment and its management, pedology (the origins and development of soil), pedometrics (the application of mathematics and statistics to soil), and soil physics. We were able to give our full attention to this development and to the selection of the scientists to join us on the road ahead. Furthermore, we were fortunate to choose a team that has maintained its enthusiasm despite the many tribulations we have faced along this path. The team comprises Jane Blagodatskaya and Adrian Unc (biology and microbiology), Siobhan Staunton and Baoshan Xing (chemistry), Karin Müller and Christina Siebe (environment and management), Rupert Bäumler (pedology), Uta Stockmann and Margaret Oliver (pedometrics), and Paul Hallett and Dani Or (physics). The editorial board has been mindful in their choice of subjects of the need for basic soil science and for state-of-the-art developments, with the aim that the audience for these volumes can be as large as possible.

During the gestation period of the Encyclopedia, methods of working for those in academia were turned on their heads. The change from “face-to-face” teaching to formulating and establishing online courses, initially for lockdown periods but then for much of the next 2 years, had huge consequences for academic staff worldwide. This meant that they had less time to devote to outside activities, such as writing chapters for an encyclopedia. As a result, some authors were unable to complete their chapters in time, leaving a few unfortunate gaps. The whole team has done as much as they can to limit any adverse effects, but we ask readers to appreciate the struggle that authors have had with illness (personal, within their families and work colleagues) and a great increase in their workload. Despite this, we are presenting a comprehensive document that should stand the test of the next 20 years of soils in the environment.

This century has seen an almost unprecedented interest in the soil and environmental sciences. It has been triggered not only by issues that the Earth is facing from global climate change, increasing population pressure, soil and land degradation, and food insecurity but also by the many new and sophisticated methods available, such as the DNA determination of microbial populations and their functional capabilities, proximal sensing, machine learning methods, and an ever-increasing array of technological tools. These developments have greatly increased our knowledge and appreciation of the complexity of soils, which has changed our understanding of some well-established aspects of the science. The soil and its environment are at the critical interface of the atmosphere, biosphere, hydrosphere, and lithosphere. They deserve to be taken seriously by the international community in the same way that air and water have been. Governments are now paying attention

to the soil and environment, and a legal and governance framework is emerging to look after these parts of our life support resources.

We have aimed, through our editing of chapters, to make the language of the chapters accessible to all, regardless of educational background. This does not mean that we have diminished the scientific content, but that readers should be able to gain the foothold they desire in these subjects to progress to greater knowledge and understanding of them. We have retained the overview of soil written by Daniel Hillel for the First Edition, first to honor his great contribution to soil science and second to demonstrate the significance of soil. It follows this Preface.

*Michael J. Goss
Margaret A. Oliver*

SOIL

The term “soil” refers to the weathered and fragmented outer layer of our planet’s land surfaces. Formed initially through the physical disintegration and chemical alteration of rocks and minerals by physical and biogeochemical processes, soil is influenced by the activity and accumulated residues of a myriad of diverse forms of life. As it occurs in different geologic and climatic domains, soil is an exceedingly varied body with a wide range of attributes.

Considering the height of the atmosphere, the thickness of the earth’s rock mantle, and the depth of the ocean, one observes that soil is an amazingly thin body—typically not much more than one meter thick and often less than that. Yet it is the fount of terrestrial life, within which biological productivity is generated and sustained. It acts like a composite living entity, a home to a community of innumerable microscopic and macroscopic plants and animals. A mere fistful of soil typically contains billions of microorganisms, which perform vital interactive biochemical functions. Another intrinsic attribute of the soil is its sponge-like porosity and its enormous internal surface area. That same fistful of soil may actually consist of several hectares of active surface, upon which physicochemical processes take place continuously.

Realizing humanity’s utter dependence on the soil, ancient peoples, who lived in greater closeness to nature than many of us today, actually revered the soil. It was not only their source of livelihood but also the material from which they built their homes and that they learned to shape, heat, and fuse into household vessels and writing tablets (ceramics, made of clayey soil, being the first synthetic material in the history of technology). In the Bible, the name assigned to the first human was Adam, derived from “adama,” meaning soil. The name given to the first earthling’s mate was Hava (Eve, in transliteration), meaning “living” or “life-giving.” Together, therefore, Adam and Eve signified quite literally “Soil and Life.”

The same powerful metaphor is echoed in the Latin name for human species—Homo, derived from humus, the material of the soil. Hence, the adjective “human” also implies “of the soil.” Other ancient cultures evoked equally powerful associations. To the Greeks, the earth was a manifestation of Gaea, the maternal goddess who, impregnated by Uranus (god of the sky), gave birth to all the gods of the Greek pantheon.

Our civilization depends on the soil more crucially than ever because our numbers have grown while available soil resources have diminished and deteriorated. Paradoxically, however, even as our dependence on the soil has increased, most of us have become physically and emotionally detached from it. Many of the people in the so-called developed countries spend their lives in the artificial environment of a city, insulated from direct exposure to nature, and some children may now assume as a matter of course that food originates in supermarkets.

Detachment has bred ignorance, and out of ignorance has come the delusion that our civilization has risen above nature and has set itself free of its constraints. Agriculture and food security, erosion and salination, degradation of natural ecosystems, depletion and pollution of surface waters and aquifers, and decimation of biodiversity—all these processes, which involve the soil directly or indirectly, have become abstractions to many people. The very language we use betrays disdain for that common material underfoot, often referred to as “dirt.” Some fastidious parents prohibit their children from playing in the mud and hurry to wash their “soiled” hands when the children nonetheless obey an innate instinct to do so. Thus, soil is devalued and treated as unclean though it is the terrestrial realm’s principal medium of purification, wherein wastes are decomposed and nature’s productivity is continually rejuvenated.

Scientists who observe soil closely see it in effect as a seething foundry in which matter and energy are in constant flux. Radiant energy from the sun streams onto the field and cascades through the soil and the plants growing in it. Heat is exchanged, water percolates through the soil’s intricate passages, and plant roots extract

water and transmit it to their leaves, which transpire it back to the atmosphere. Leaves absorb carbon dioxide from the air and synthesize it with soil-derived water to form the primary compounds of life. Oxygen emitted by the leaves makes the air breathable for animals, which consume and in turn fertilize plants.

Soil is thus a self-regulating bio-physio-chemical factory, processing its own materials, water, and solar energy. It also determines the fate of rainfall and snowfall reaching the ground surface—whether the water thus received will flow over the land as runoff, or seep downward to the subterranean reservoir called groundwater, which in turn maintains the steady flow of springs and streams. With its finite capacity to absorb and store moisture, and to release it gradually, the soil regulates all these phenomena. Without the soil as a buffer, rain falling over the continents would run off entirely, producing violent floods rather than sustained river flow.

Soil naturally acts as a living filter, in which pathogens and toxins that might otherwise accumulate to foul the terrestrial environment are rendered harmless. Since time immemorial, humans and other animals have been dying of all manner of diseases and have then been buried in the soil, yet no major human disease is transmitted by it. The term antibiotic was coined by soil microbiologists who, as a consequence of their studies of soil bacteria, specifically actinomycetes, discovered streptomycin (an important cure for tuberculosis and other infections). Ion exchange, a useful process of water purification, also was discovered by soil scientists studying the passage of solutes through beds of clay.

However unique in form and function, soil is not an isolated body. It is, rather, a central link in the larger chain of interconnected domains and processes comprising the terrestrial environment. The soil interacts both with the overlying atmosphere and the underlying strata, as well as with surface and underground bodies of water. Especially important is the interrelation between the soil and climate. In addition to its function of regulating the cycle of water, it also regulates energy exchange and surface temperature.

When land is cleared of vegetation and turned into a cultivated field, the native biomass above the ground is often burned and the organic matter within the soil tends to decompose. These processes release carbon dioxide into the atmosphere, thus contributing to the Earth's greenhouse effect and to global warming. On the other hand, the opposite act of afforestation and soil enrichment with organic matter, such as can be achieved through conservation management, may serve to absorb carbon dioxide from the atmosphere. To an extent, the soil's capacity to store carbon thus helps to mitigate the greenhouse effect.

Thousands of years are required for nature to create life-giving soil out of sterile bedrock. In only a few decades, however, unknowing or uncaring humans can destroy that wondrous work of nature. In various circumstances, mismanaged soils may be subject to erosion (the sediments of which tend to clog streambeds, estuaries, lakes, and coastal waters), to leaching of nutrients with attendant loss of fertility and eutrophication of water bodies, to waterlogging and impaired aeration, or to an excessive accumulation of salts that may cause a once-productive soil to become entirely sterile. Such processes of soil degradation, sometimes called "desertification," already affect large areas of land.

We cannot manage effectively and sustainably that which we do not know and thoroughly understand. That is why the tasks of developing and disseminating sound knowledge of the soil and its complex processes have assumed growing urgency and importance. The global environmental crisis has created a compelling need for a concentrated, concise, and definitive source of information—accessible to students, scientists, practitioners, and the general public—about the soil in all its manifestations in nature and in relation to the life of humans.

Daniel Hillel
May 2004

LIST OF CONTRIBUTORS FOR VOLUME 3

Stefan Abel
Independent Researcher

CL Acharya
Former Director, ICAR-Indian Institute of Soil Science, Bhopal, India

Md Mezbaul Bahar
CRC for Contamination Assessment and Remediation of the Environment (crc CARE), ATC Building, University of Newcastle, Callaghan, NSW, Australia; Global Centre for Environmental Remediation (GCER), University of Newcastle, University Drive, Callaghan, NSW, Australia

KK Bandyopadhyay
Division of Agricultural Physics, ICAR-Indian Agricultural Research Institute, New Delhi, India

Arlete Simões Barneze
Soil Biology Group, Wageningen University & Research, Wageningen, The Netherlands

Bartosz Bartkowski
Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany

Nicolas Beriot
Soil Physics and Land Management Group, Wageningen University & Research, Wageningen, The Netherlands

Bijay-Singh
Department of Soil Science, Punjab Agricultural University, Ludhiana, Punjab, India

Bhabananda Biswas
CRC for Contamination Assessment and Remediation of the Environment (crc CARE), ATC Building, University of Newcastle, Callaghan, NSW, Australia; Global Centre for Environmental Remediation (GCER), University of Newcastle, University Drive, Callaghan, NSW, Australia

Martin SA Blackwell
Net Zero and Resilient Farming, Rothamsted Research North Wyke, Okehampton, United Kingdom

Antra Boca
Latvia University of Life Sciences and Technologies, Jelgava, Latvia

Gerardo Bocco
Centro de Investigaciones en Geografía Ambiental, Universidad Nacional Autónoma de México, Mexico City, Mexico

Nanthi S Bolan
Global Centre for Environmental Remediation and International Centre for Balanced Land Use, University of Newcastle, Callaghan, NSW, Australia

Trond Børresen
Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway

Eric C Brevik
College of Agricultural, Life, and Physical Sciences, Southern Illinois University, Carbondale, IL, United States

Christopher Brock
Forschungsring, Darmstadt, Germany

Laura Cardenas
Net Zero and Resilient Farming, Rothamsted Research North Wyke, Okehampton, United Kingdom

Elizabeth Chávez-García
Department of Geography, Faculty of Philosophy and Letters, National Autonomous University of Mexico, Mexico City, Mexico

Peter Clinton
New Zealand Forest Research Insititute Ltd (Trading as Scion Research), Rotorua, New Zealand

Joanna M Cloy
Carbon, Crop and Soils, Scotland's Rural College, Edinburgh, United Kingdom

Carlos E Cordova
Department of Geography, Oklahoma State University, Stillwater, OK, United States

Sophie Cornu

Aix Marseille University, CNRS, IRD, INRA, Coll France, CEREGE, Aix-en-Provence, France

Edoardo AC Costantini

CNR-IBE—Department of Biology, Agriculture and Food Sciences, Sesto Fiorentino, Florence, Italy

Lorna Anne Dawson

Centre for Forensic Soil Science, James Hutton Institute, Aberdeen, United Kingdom; Forensic Science, School of Pharmacy and Life Sciences, Robert Gordon University, Aberdeen, United Kingdom

Igor Costa de Freitas

Universidade Federal de Minas Gerais, Instituto de Ciências Agrárias, Montes Claros, Minas Gerais, Brazil

Seth Denizen

High Meadows Environmental Institute, Princeton, NJ, United States

Thomas Eichert

University of Applied Sciences Erfurt, Erfurt, Germany

Pierre Ellssel

Division of Organic Farming, University of Natural Resources and Life Sciences, Vienna, Austria

Norm Fausey

USDA Agricultural Research Service, Soil Drainage Research Unit, Columbus, OH, United States

Luiz Arnaldo Fernandes

Universidade Federal de Minas Gerais, Instituto de Ciências Agrárias, Montes Claros, Minas Gerais, Brazil

Victoria Fernández

Systems and Natural Resources Department, School of Forest Engineering, Universidad Politécnica de Madrid, Madrid, Spain

Eleonora Flores-Ramírez

Independent Researcher

Charles Andrew Francis

Kiesselbach Crops Research Laboratory, University of Nebraska–Lincoln, Lincoln, NE, United States

Uwe Franko

Halle, Germany

Leidivan Almeida Frazão

Universidade Federal de Minas Gerais, Instituto de Ciências Agrárias, Montes Claros, Minas Gerais, Brazil

Bernhard Freyer

Division of Organic Farming, University of Natural Resources and Life Sciences, Vienna, Austria

Jürgen K Friedel

Division of Organic Farming, University of Natural Resources and Life Sciences, Vienna, Austria

Harald Ginzky

German Environment Agency, Dessau-Roßlau, Germany

AM Gordon

School of Environmental Sciences, University of Guelph, Guelph, ON, Canada

Andrew S Gregory

Rothamsted Research, Harpenden, United Kingdom

KM Hati

Division of Soil Physics, ICAR-Indian Institute of Soil Science, Bhopal, India

Jeffrey A Homburg

Catena Affiliates, Tucson, AZ, United States; School of Anthropology, University of Arizona, Tucson, AZ, United States

Esperanza Huerta-Lwanga

Soil Physics and Land Management Group, Wageningen University & Research, Wageningen, The Netherlands; Agroecología, El Colegio de la Frontera Sur, Campeche, México

Trevor Kenneth James

AgResearch, Hamilton, New Zealand

Arón Jazcilevich

Instituto de Ciencias de la Atmósfera y Cambio Climático, UNAM, Ciudad de México, Mexico

Rai S Kookana

Commonwealth Scientific and Industrial Research Organisation (CSIRO), Glen Osmond, SA, Australia

Sandeep Kumar

ICAR-Indian Agricultural Research Institute, Karnal, India

Sumanta Kundu

ICAR-Central Research Institute for Dryland Agriculture, Hyderabad, India

Rattan Lal

CFAES Rattan Lal Center for Carbon Management and Sequestration, The Ohio State University, Columbus, OH, United States

Mathieu Lamandé

Department of Agroecology, AU Foulum, Aarhus University, Tjele, Denmark

Maxine J Levin

University of Maryland, Silver Spring, MD, United States

Guy J Levy

Institute for Soil, Water and Environmental Sciences, ARO, Volcani Research Center, Rishon-Lezion, Israel

S Malleswari

Acharya N.G. Ranga Agricultural University, Agricultural Research Station, Srikakulam, Andhra Pradesh, India

Ravula Manasa
ICAR-National Academy of Agricultural Research
Management, Hyderabad, India

Ben Marchant
British Geological Survey, Keyworth, United Kingdom

Julian R Massenberg
Helmholtz Centre for Environmental Research – UFZ,
Leipzig, Germany

Lynda H McCarthy
Toronto Metropolitan University (formerly Ryerson
University), Toronto, ON, Canada

Richard McDowell
AgResearch, Lincoln Science Centre, Christchurch, New
Zealand

Steve P McGrath
Rothamsted Research, Hertfordshire, United Kingdom

Ram Swaroop Meena
Department of Agronomy, Institute of Agricultural
Sciences, Banaras Hindu University, Varanasi, India

Charles Norman Merfield
The BHU Future Farming Centre, Canterbury, New
Zealand

Kurt Möller
Institute of Crop Science, Fertilisation and Soil
Matter Dynamics, Universität Hohenheim, Stuttgart,
Germany

David Montagne
Université Paris-Saclay, INRAE, AgroParisTech, UMR
ECOSYS, Palaiseau, France

Jean Louis Morel
Laboratoire Sols et Environnement, Université de Lorraine,
INRAE, Nancy, France

Sarah Morgan
Agriculture and Environment Department, Harper Adams
University, Newport, Shropshire, United Kingdom

K Müller
Sustainable Production, The New Zealand Institute for
Plant and Food Research Limited, Auckland,
New Zealand

Lars Juhl Munkholm
Department of Agroecology, AU Foulum, Aarhus
University, Tjele, Denmark

Uri Nachshon
Institute for Soil, Water and Environmental Sciences, ARO,
Volcani Research Center, Rishon-Lezion, Israel

PKR Nair
University of Florida, Gainesville, FL, United States

K C Nataraj
Agricultural College, Naira, Acharya N.G. Ranga
Agricultural University (ANGRAU), Guntur, Andhra
Pradesh, India

Divina A Navarro
Commonwealth Scientific and Industrial Research
Organisation (CSIRO), Glen Osmond, SA, Australia

Thomas Nehls
Ecohydrology and Landscape Evaluation, Technische
Universität Berlin, Berlin, Germany

Adriana Marcela Silva Olaya
Universidad de la Amazonia, Florencia, Caquetá,
Colombia

Margaret A Oliver
Soil Research Centre, Department of Geography &
Environmental Science, School of Human & Environmental
Sciences, The University of Reading, Reading, United
Kingdom

Jen Owens
New Zealand Forest Research Institute Ltd (Trading as
Scion Research), Rotorua, New Zealand

Anna Paltseva
Environmental Science, SL Mid-Winter Fair Endowed
Professorship in Renewable Resources, University of
Louisiana at Lafayette, Lafayette, LA, United States

Kiran Pilli
ICAR-National Academy of Agricultural Research
Management, Hyderabad, India

JVNS Prasad
ICAR-Central Research Institute for Dryland Agriculture,
Hyderabad, India

Simone Priori
Department of Agriculture and Forest Sciences (DAFNE),
University of Tuscia, Viterbo, Italy

Manzoor Qadir
United Nations University Institute for Water,
Environment and Health (UNU-INWEH), Hamilton,
ON, Canada

S Rakesh
ICAR-National Academy of Agricultural Research
Management, Hyderabad, India

G Ranjith Kumar
ICAR-National Academy of Agricultural Research
Management, Hyderabad, India

KV Rao
ICAR-Central Research Institute for Dryland Agriculture,
Hyderabad, India

- Ravi Naidu
CRC for Contamination Assessment and Remediation of the Environment (crc CARE), ATC Building, University of Newcastle, Callaghan, NSW, Australia; Global Centre for Environmental Remediation (GCER), University of Newcastle, University Drive, Callaghan, NSW, Australia
- Laura Bertha Reyes-Sánchez
Agricultural Engineering Department, FES-C, National Autonomous University of Mexico, Mexico City, Mexico
- R Jane Rickson
School of Water, Energy and Environment, Cranfield University, Cranfield, United Kingdom
- Nicolas Saby
INRAE, Orléans, France
- Samaresh Sahoo
ICAR-National Academy of Agricultural Research Management, Hyderabad, India
- Peter Sanderson
CRC for Contamination Assessment and Remediation of the Environment (crc CARE), ATC Building, University of Newcastle, Callaghan, NSW, Australia; Global Centre for Environmental Remediation (GCER), University of Newcastle, University Drive, Callaghan, NSW, Australia
- Alexandra Sandhage-Hofmann
Institute of Crop Science and Resource Conservation, Division Soil Science, University of Bonn, Bonn, Germany
- Jonathan A Sandor
Agronomy Department, Iowa State University, Ames, IA, United States
- Tek B Sapkota
International Maize and Wheat Improvement Center (CIMMYT), El Batán, Mexico
- Christopher A Scott
Department of Ecosystem Science and Management, 124 Forest Resources Building, Pennsylvania State University, University Park, PA, United States
- Carmen Segura
Net Zero and Resilient Farming, Rothamsted Research North Wyke, Okehampton, United Kingdom
- Geoffroy Séré
Laboratoire Sols et Environnement, Université de Lorraine, INRAE, Nancy, France
- Bryant M Serre
McGill University, Montreal, QC, Canada
- C Siebe
Instituto de Geología, Universidad Nacional Autónoma de México, Ciudad de México, México
- Keith A Smith
School of GeoSciences, University of Edinburgh, Edinburgh, United Kingdom
- Rogério Peres Soratto
College of Agricultural Science, São Paulo State University (UNESP), São Paulo, Brazil
- Ch Srinivasarao
ICAR-National Academy of Agricultural Research Management, Hyderabad, India
- Karl Stahr
Department of Soil Science and Land Evaluation (310a), Universität Hohenheim, Stuttgart, Germany
- Jeffrey Strock
University of Minnesota, Southwest Research and Outreach Center, Lamberton, MN, United States
- John Tatarko
Rangeland Resources and Systems Research Unit, USDA Agricultural Research Service, Fort Collins, CO, United States
- NV Thevathasan
School of Environmental Sciences, University of Guelph, Guelph, ON, Canada
- Alexandra R Toland
Bauhaus-Universität Weimar, Faculty of Art and Design, Chair of Arts and Research, Geschwister-Scholl-Straße 7, Weimar, Germany
- Sevilay Topcu
University of Cukurova, Faculty of Agriculture, Department of Agricultural Structures and Irrigation, Adana, Turkey
- Viacheslav Vasenev
Soil Geography and Landscape Group, Wageningen University, Wageningen, The Netherlands
- Hai-Long Wang
Biochar Engineering Technology Research Center of Guangdong Province, School of Environmental and Chemical Engineering, Foshan University, Foshan, Guangdong, China
- Peng Wang
College of Resources and Environmental Sciences, Nanjing Agricultural University, Nanjing, China
- Yu-Jun Wang
Key Laboratory of Soil Environment and Pollution Remediation, Institute of Soil Science, Chinese Academy of Sciences, Nanjing, Jiangsu, China

Ann G Wellens

*Instituto de Ciencias de la Atmósfera y Cambio Climático,
UNAM, Ciudad de México, Mexico; Facultad de
Ingeniería, UNAM, Ciudad de México, Mexico*

Joann K Whalen

*Department of Natural Resource Sciences, Macdonald
Campus of McGill University, Sainte-Anne-de-Bellevue,
QC, Canada*

Martin Wiesmeier

*Bavarian State Research Center for Agriculture, Freising,
Germany*

Daniel Wolter

*Bauhaus-Universität Weimar, Faculty of Art and Design,
Chair of Arts and Research, Geschwister-Scholl-Straße 7,
Weimar, Germany*

Ping Wu

*Center for Agricultural Resources Research, Institute of
Genetics and Developmental Biology, Chinese Academy of
Sciences, Shijiazhuang, China*

Caio Fernandes Zani

*School of Natural and Environmental Sciences, Newcastle
University, Newcastle upon Tyne, United Kingdom*

Fang-Jie Zhao

*College of Resources and Environmental Sciences, Nanjing
Agricultural University, Nanjing, China*

CONTENTS OF ALL VOLUMES

VOLUME 1: SOIL BIOLOGY

Overview

Edaphology	1
<i>RM Goss and AL Ulery</i>	

Biodiversity	10
<i>Monica A Farfan and Diana H Wall</i>	

Soil organisms

Bacteriophage (Viruses)	16
<i>Li Bi, Ji-Zheng He, Dan-Ting Yu, Li-Li Han, and Hang-Wei Hu</i>	

Archaea	25
<i>Mikhail V Semenov</i>	

Bacteria	31
<i>Mikhail V Semenov</i>	

Soil protists	39
<i>Stefan Geisen, Nathalie Amacker, and Alexandre Jousset</i>	

Fungi including mycorrhizal fungi	49
<i>F Buscot</i>	

Soil algae	61
<i>Karen Baumann and Patrick Jung</i>	

Soil arthropods: Underfoot and all around	70
<i>Ernest C Bernard</i>	

Nematodes	105
<i>Deborah A Neher and Thomas O Powers</i>	

Earthworms: Essential ecosystem engineers providing vital ecosystem services	112
<i>Maria JI Briones</i>	

Complex structures

Soil biofilms: Evolving concepts and ecological functions	120
<i>Yichao Wu, Marc Redmile-Gordon, Peng Cai, Chunhui Gao, and Qiaoyun Huang</i>	

Biocrusts	127
<i>Matthew A Bowker, Jasmine Anenberg, Hilina Yohannes Kebede, Anita Antoninka, and Lydia Bailey</i>	

Functioning

Studying soil microbial communities: Culture-dependent approaches	139
<i>Eduardo K Mitter, Dasiel Obregon, Micaela Tosi, Heather Slinn, and Kari Dunfield</i>	

Studying soil microbial communities: Culture-independent approaches	146
<i>Tolulope Mafa-Attoye, Heather Slinn, Micaela Tosi, Evan Mayer, Dasiel Obregon, and Kari Dunfield</i>	

Studying microbial function in soil: Culture-independent techniques <i>Micaela Tosi, Dasiel Obregon, Heather Slinn, and Kari Dunfield</i>	155
Kinetics of microbial processes: General principles <i>NS Panikov</i>	168
Kinetics of microbial processes: Applications to soil biology <i>NS Panikov</i>	186
Biogeochemical cycles	
Carbon cycle in soils: Dynamics and management <i>Charles W Rice, Carlos B Pires, and Marcos VM Sarto</i>	219
Carbon storage in soils <i>Isabelle Basile-Doelsch, Tiphaine Chevallier, Marie-France Dignac, and Amandine Erktan</i>	228
Soil organic nitrogen <i>EJ Daly and KA Congreves</i>	243
Nitrification: Process, organisms and management <i>Jeanette M Norton and Yang Ouyang</i>	253
Microbial denitrification <i>Michael J Goss, Sergey Blagodatsky, and Adrian Unc</i>	261
Symbiotic nitrogen fixation <i>Florence Mus and Hsin-Hua Wu</i>	270
Nitrogen: Non-Symbiotic Nitrogen fixation in soils <i>Vadakattu VSR Gupta and Sarah S Roley</i>	283
Phosphorus in soils—Biological interactions <i>Nataliya Bilyera and Iryna Loginova</i>	293
Biological transformations of sulfur in soil <i>Chaoqun Wang and Maxim Dorodnikov</i>	303
Plant uptake of mineral nutrients <i>Angela Hodge</i>	310
Higher plants and their interactions with soil organisms	
Soil Biology: Root form and function <i>Magda Garbowski, Grégoire Freschet, Louise Jackson, Cynthia Brown, and Louise Comas</i>	321
Root mucilage: Chemistry and functions in soil <i>Yutong Jiang, Lijun Hou, Fang Ding, and Joann K Whalen</i>	332
Root exudates and microorganisms <i>Aidan De Sena, Krisztina Mosdossy, Joann K Whalen, and Chandra A Madramootoo</i>	343
Root interactions with the microbiome from the rhizoplane to the bulk soil: An overview <i>Joann K Whalen and Shamim Gul</i>	357
Microorganisms & biological processes	
Soil respiration <i>Jeremiah D Vallotton, Sergey Blagodatsky, and Adrian Unc</i>	369
Anaerobic soil processes <i>PW Inglett, D Said-Pullicino, and K Ramesh Reddy</i>	379
Mineral-organic-microbial interactions <i>Balwant Singh, Sabina Yeasmin, and Donald L Sparks</i>	387

Microbial ecology & management

Habitat for soil organisms and their functions <i>Alexandra Kravchenko and Andrey Guber</i>	407
Plant growth promoters <i>Elisa Gamalero, Gustavo Santoyo, and Bernard R Glick</i>	418
Inoculation with microorganisms <i>Maureen O'Callaghan and David Wright</i>	428
Biocontrol of soil-borne plant diseases <i>Mehrdad Mohammadiani and Matthew Bakker</i>	439
Management of soil biology for crop protection from biotic and abiotic stresses <i>Michael J Goss, Isabel Brito, and Mário Carvalho</i>	447
Trees, hedges, agroforestry and microbial diversity <i>Josiah B Judson, Joseph Holden, Pippa Chapman, and Marcelo Valadares Galdos</i>	469
Urban soil microbiome: Activity, diversity and functioning <i>Maria Korneykova, Ekaterina Kozlova, Kristina Ivashchenko, Maria Vasilieva, and Viacheslav Vasenev</i>	480
Microbiology of extreme soil environments <i>Patrice Dion</i>	494
Microbial mobility and transport in soils <i>Mohammad Reza Mosaddeghi, Adrian Unc, Ali Akbar Mahboubi, and Azadeh Safadoust</i>	512

Biology of soil amendments & contaminant degradation

Biology of compost <i>Heribert Insam, Thomas Klammersteiner, and María Gómez-Brandón</i>	522
Landfarming and its microbiology <i>Naga Raju Maddela, Balasubramanian Ramakrishnan, Kadiyala Venkateswarlu, and Mallavarapu Megharaj</i>	533
Biodegradation of organic contaminants <i>PB Hatzinger and JW Kelsey</i>	547
Biological aspects of manure management <i>Natasha Rayne and Bee Khim Chim</i>	558
Septic systems and municipal waste disposal on land <i>Charles P Gerba, Michael J Goss, and Adrian Unc</i>	567

Climate change

Climate change impacts on soil biology <i>Alberto Canarini, Lucia Fuchslueger, François-Xavier Joly, and Andreas Richter</i>	578
---	-----

VOLUME 2: CHEMISTRY**Basic Soil Chemistry**

Acidity and acidification <i>Siobhan Staunton</i>	1
pH <i>Siobhan Staunton</i>	8
Buffering capacity <i>Siobhan Staunton and Ravi Naidu</i>	15
Soil specific surface area <i>Emmanuel Arthur, Markus Tuller, and Kurt D Pennell</i>	22

Soil Analytical Chemistry

Cation exchange capacity and reactions <i>Donald S Ross and Ulf Skyllberg</i>	32
Testing for nutrients in the soil and the environment <i>Antonio P Mallarino</i>	43
Metal extraction and analysis <i>Guo Wang and Honghong Li</i>	50
Soil organic matter: Composition <i>Markus Kleber</i>	58
Physical fractionation techniques <i>Sam J Leuthold, Michelle L Haddix, Jocelyn Lavallee, and M Francesca Cotrufo</i>	68
The use of stable isotopes in soil science: Low atomic number elements <i>Chiara Pistocchi, Isabelle Bertrand, Kristýna Kantnerová, Astrid Oberson, Federica Tamburini, and Roland A Werner</i>	81
The use of radioactive isotopes to study soil pools and processes <i>Timothy I McLaren</i>	96
The use of stable isotopes in soil science: Metals <i>Cécile Quantin and Damien Guinoiseau</i>	107

Periodic Table

Alkali metals in soils <i>Guy J Levy and Asher Bar-Tal</i>	116
Alkaline earth metals in soil <i>Justin B Richardson</i>	126
Transition metals in soil <i>Siobhan Staunton</i>	135
Nitrogen and ammonia in soils <i>Nicholas T Girkin and Hannah V Cooper</i>	142
Phosphorus in soil <i>Luisella Celi, Maria Martin, and Elisabetta Barberis</i>	152
Aluminum in soils <i>Syuntaro Hiradate</i>	167
Iron in soils <i>Trolard Fabienne and Guilhem Bourrié</i>	176
Sulfur and selenium in soil <i>Dongli Liang, Fei Zhou, and Mingxing Qi</i>	192
Arsenic in soil <i>Md Imran Ullah Sarkar, Afsana Jahan, Md Tofail Hossain, Mohammad Mahmudur Rahman, and Ravi Naidu</i>	201
Non-transition elements: Metals and metalloids in soils <i>Natasha, Sana Khalid, Muhammad Shahid, Faiz Rabbani, and Camille Dumat</i>	214
Lanthanides, actinides and radon <i>Franco Ajmone Marsan and Mery Malandrino</i>	225
Halogens in soils <i>Svensson Teresia, Thiry Yves, Bueno Maïté, and Oelmann Yvonne</i>	234
Transformation of metals and metalloids by microorganisms <i>Susan M Glasauer, Euan P Burford, and Geoffrey Michael Gadd</i>	243
Dynamics of radionuclides <i>Siobhan Staunton</i>	256
Silicon and other non-metal elements from Group 14 <i>Bruno Delvaux and Zimin Li</i>	263

Soil Physical Chemistry

Chemical equilibria	274
<i>Arthur Paul Schwab</i>	
Dissolution and precipitation processes	284
<i>Zhenqing Shi, Shiwen Hu, Rong Li, and Yang Lu</i>	
Reactions, redox potential and kinetics	297
<i>Trolard Fabienne and Guilhem Bourrié</i>	
Sorption of organic chemicals in soil	315
<i>Joseph J Pignatello</i>	
Sorption—Oxyanions	327
<i>CP Schulthess, H Wijnja, W Yang, Anahita Khosravi, and Baoshan Xing</i>	
Sorption—Metals	336
<i>Daniel G Strawn and Donald L Sparks</i>	
Sorption–desorption kinetics	343
<i>Gautier Landrot and Donald L Sparks</i>	
Sorption-desorption hysteresis	355
<i>Wenhao Wu, Hongxia Zhu, and Kun Yang</i>	
Surface charge and zero-charge points	368
<i>J Chorover</i>	
Flocculation and dispersion	376
<i>Guy J Levy and Isaac Shainberg</i>	

Soil Carbon Chemistry

Soil organic matter dynamics	387
<i>Yu Luo and Jianming Xu</i>	
Soil proteins	396
<i>Siobhan Staunton</i>	
Biochar: Properties, mechanisms, and interactions in the soil	406
<i>Balwant Singh, Shamim Mia, and Marta Camps-Arbestain</i>	
Dissolved organic carbon	421
<i>Jason James and Rob Harrison</i>	
Black carbon, pyrogenic carbon	432
<i>Heike Knicker</i>	
Fire effects on soil	448
<i>Stefan H Doerr, Cristina Santín, and Jorge Mataix-Solera</i>	
Interfacial reactions of microorganisms with minerals and organic matter	458
<i>Rota Wagai, Puu-Tai Yang, and Klaus Kaiser</i>	
Pesticide fate in soils	470
<i>Pierre Benoit, Laure Mamy, Carole Bedos, and Enrique Barriuso</i>	
Soil organic matter in molecular simulations	483
<i>Martin H Gerzabek, Daniel Tunega, Edgar Galicia-Andrés, and Chris Oostenbrink</i>	

Soil Physical Analysis

Raman spectroscopy	497
<i>Anupam Das and Huiyuan Guo</i>	
Fourier transform infrared spectroscopic methods of soil analysis	510
<i>Derek Peak</i>	
Electron paramagnetic resonance spectroscopy: Part I Historic Perspectives	519
<i>Giorgio S Senesi and Nicola Senesi</i>	

Electron paramagnetic resonance spectroscopy: Part II the view forward <i>Robert L Cook, Emmanuel Ampiah, Phillip T Sprunger, and Randall W Hall</i>	533
Fluorescence Spectroscopy: Part I Principles <i>Yong Zhang, Shuai Guo, and Chaoxian Wei</i>	544
Fluorescence spectroscopy: Part II practical applications <i>Yong Zhang, Chaoxian Wei, and Shuai Guo</i>	552
Near and mid infrared soil spectroscopy <i>Wartini Ng, Brendan Malone, Budiman Minasny, and Sangho Jeon</i>	567
Electron microscopic methods (TEM, SEM and energy dispersal spectroscopy) <i>Jian Zhao and Xia Liu</i>	575
Application of atomic force microscopy (AFM) on environmental interfaces <i>Saikat Ghosh, Zhang Di, Bo Pan, and Baoshan Xing</i>	589
Applications of nuclear magnetic resonance for the study of soils <i>Daniel H Lysak, Myrna J Simpson, and Andre J Simpson</i>	605
Neutron-gamma analysis of soil elements <i>Aleksandr Kavetskiy, Galina Yakubova, Nikolay Sargsyan, Stephen A Prior, and H Allen Torbert</i>	625
X-Ray Diffraction <i>Chao Wang and Cheng Gu</i>	642
Soil analysis using laser-induced breakdown spectroscopy <i>Débora Marcondes Bastos Pereira Milori, Ladislau Martin-Neto, Hero T Gollany, Diego Victor de Babos, Luis Carlos Leva Borduchi, and Paulino Ribeiro Villas-Boas</i>	654
Calorimetric analysis <i>Xingmin Rong and Qiaoyun Huang</i>	662
Understanding of soil processes at the microscale—Use of NanoSIMS in soil science <i>Carsten W Mueller, Carmen Hoeschen, and Ingrid Koegel-Knabner</i>	670
Nanoparticles in soil <i>Di Lang, Quan Chen, Yang Liu, and Bo Pan</i>	681
Overview	
Soil chemistry and climate change <i>Jianming Xu, Yong Li, and Zhongmin Dai</i>	692
VOLUME 3: ENVIRONMENT AND MANAGEMENT	
Soil functions, potential uses of soil and ecosystem services	
Soil functions, soil potential uses and ecosystem services <i>Karl Stahr</i>	1
Farming systems to conserve land for nature <i>Rattan Lal</i>	12
Geoarchaeology and soils <i>Carlos E Cordova</i>	20
Civilizations and the role of soil, agriculture, and climate change <i>Jeffrey A Homburg, Eric C Brevik, and Jonathan A Sandor</i>	29
Ecosystem services provided by soils in highly anthropized areas (SUITMAs) <i>Jean Louis Morel, Geoffroy Séré, Viatcheslav Vasenev, and Thomas Nehls</i>	37
Structure and function of forested soils <i>Peter Clinton and Jen Owens</i>	56
Agroforestry <i>NV Thevathasan, AM Gordon, and PKR Nair</i>	68

Rangeland management <i>Alexandra Sandhage-Hofmann</i>	88
Effects on soil of grassland management for pasture, hay and silage <i>Carmen Segura, Laura Cardenas, Richard McDowell, Sarah Morgan, and Martin SA Blackwell</i>	102
Dryland farming: Technological and management options for sustainable agriculture and food systems <i>Ch Srinivasarao, G Ranjith Kumar, Ravula Manasa, Kiran Pilli, Samaresh Sahoo, S Rakesh, Sumanta Kundu, K C Nataraj, KV Rao, JVNS Prasad, and S Malleswari</i>	113
The effect of crop rotations on soil <i>Caio Fernandes Zani, Arlete Simões Barneze, Rogério Peres Soratto, and Charles Andrew Francis</i>	125
The contribution of organic farming systems to soil fertility—A systems perspective <i>Bernhard Freyer, Pierre Ellssel, Jürgen K Friedel, and Kurt Möller</i>	135
Soil functions to support natural plant communities and crops: Soil multifunctionality <i>Leidivan Almeida Frazão, Luiz Arnaldo Fernandes, Adriana Marcela Silva Olaya, and Igor Costa de Freitas</i>	146
Methane and nitrous oxide emissions from land <i>Joanna M Cloy and Keith A Smith</i>	153
Soil and land degradation	
Overview chapter on soil degradation <i>K Müller, Margaret A Oliver, and C Siebe</i>	165
Soil monitoring <i>Ben Marchant and Nicolas Saby</i>	172
Soil quality and health key indicators <i>Edoardo AC Costantini and Simone Priori</i>	181
Water induced soil erosion <i>R Jane Rickson</i>	193
Erosion by wind <i>John Tatarko</i>	208
Human exposure to aeroparticles emitted from unpaved roads due to traffic resuspension and wind erosion <i>Arón Jazcilevich and Ann G Wellens</i>	223
Soil salination processes and management <i>Uri Nachshon and Guy J Levy</i>	236
Degradation of soil structure and soil organic matter <i>Andrew S Gregory</i>	246
Drainage, soil water storage, buffering and filtering <i>Jeffrey Strock and Norm Fausey</i>	257
Environmental effects of irrigation <i>Sevilay Topcu</i>	266
Trade-offs of wastewater irrigation <i>Manzoor Qadir and Christopher A Scott</i>	277
Pollution and risks of trace elements in the soil environment <i>Peng Wang, Fang-Jie Zhao, and Steve P McGrath</i>	288
Pollutants—Persistent organic <i>Rai S Kookana and Divina A Navarro</i>	296
Pollution by microplastic in agricultural areas <i>Nicolas Beriot and Esperanza Huerta-Lwanga</i>	309

Municipal solid waste management: Production, management, and environmental effects <i>Bryant M Serre and Lynda H McCarthy</i>	321
Remediation of metals and organic contaminants in soil <i>Peter Sanderson, Md Mezbaul Bahar, Bhabananda Biswas, and Ravi Naidu</i>	333
Soil change under different scenarios	
Soil change under different scenarios <i>Joann K Whalen</i>	344
The effect of global change on the soil body <i>Sophie Cornu, David Montagne, and Antra Boca</i>	360
The effects of adequate and excessive application of mineral fertilizers on the soil <i>Bijay-Singh and Tek B Sapkota</i>	369
Foliar Application of Nutrients <i>Thomas Eichert and Victoria Fernández</i>	382
Mulches and cover crops part I: Types <i>KK Bandyopadhyay, CL Acharya, and KM Hati</i>	392
Mulches and cover crops part II: Role in soil health and climate resilient agriculture <i>KK Bandyopadhyay, CL Acharya, and KM Hati</i>	401
The potential agronomic and environmental applications of biochar: Prospects and challenges <i>Yu-Jun Wang, Ping Wu, Nanthi S Bolan, and Hai-Long Wang</i>	414
Soil management for carbon sequestration <i>Christopher Brock, Uwe Franko, and Martin Wiesmeier</i>	424
Soil tillage <i>Mathieu Lamandé, Lars Juhl Munkholm, and Trond Børresen</i>	433
Weed and soil management: A balancing act <i>Trevor Kenneth James and Charles Norman Merfield</i>	439
Terraces as traditional agricultural landscapes. Stability amidst change <i>Gerardo Bocco</i>	450
Soil quality protection policies and plans to ensure sustainability <i>Ram Swaroop Meena and Sandeep Kumar</i>	457
Construction of Technosols for green roofs and vertical green system <i>Eleonora Flores-Ramírez, Stefan Abel, Elizabeth Chávez-García, and Geoffroy Séré</i>	473
Management of park areas, sport fields, and school yards, including golf courses and public right of ways <i>Maxine J Levin and Anna Paltseva</i>	488
Soil-society relations	
Overview chapter on soil-society relations <i>C Siebe, Margaret A Oliver, and K Müller</i>	505
Soil art: Sensory and symbolic engagement with soils <i>Alexandra R Toland and Daniel Wolter</i>	509
Soil science in the criminal justice system <i>Lorna Anne Dawson</i>	521
Soil governance at the international, regional and national level <i>Harald Ginzky</i>	532
The economics of soils' contribution to human well-being <i>Bartosz Bartkowski and Julian R Massenber</i>	547

Soil and human health	555
<i>Margaret A Oliver and Eric C Brevik</i>	
Soil Science Education, Awareness and Outreach in the Modern World	572
<i>Laura Bertha Reyes-Sánchez and Eric C Brevik</i>	
Soils in landscape planning	580
<i>Seth Denizen</i>	
VOLUME 4: PEDOLOGY	
Pedology	
Pedology—Basic principles	1
<i>Rupert Bäumler, Uta Stockmann, and Margaret A Oliver</i>	
Factors of soil formation, pedogenic processes and mineralogy	
Overview chapter on factors of soil formation	7
<i>Rupert Bäumler, Margaret A Oliver, and Uta Stockmann</i>	
Factors of soil formation: Climate	14
<i>Stefaan Dondeyne, Stephan Mantel, and Seppe Deckers</i>	
Factors of soil formation: Soil biota and organic matter as a factor for soil formation and weathering processes	25
<i>Scott Werts</i>	
Soil forming factors: Parent material	34
<i>Stephen Nortcliff</i>	
Parent materials: Loess	42
<i>Tobias Sprafke</i>	
Time	48
<i>Susan E Melzer, Caroline M Yonker, and Eugene F Kelly</i>	
Factors of soil formation—Human impacts	54
<i>Jonathan A Sandor, C Lee Burras, Michael Thompson, and Skye A Wills</i>	
Dynamic soils	69
<i>Markus Egli and Dennis Dahms</i>	
Interactions between soils and climate change	82
<i>Cornelia Rumpel</i>	
Primary minerals	91
<i>Balwant Singh</i>	
Mineral components of the soil clay fraction	109
<i>Darrell G Schulze</i>	
Short-range ordered aluminosilicates	121
<i>Tetsuhiro Watanabe, James B Harsh, and Rota Wagai</i>	
Metal oxides	135
<i>Andreas C Scheinost and Balwant Singh</i>	
Soil morphology	149
<i>Phillip R Owens and Zamir Libohova</i>	
Soil micromorphology	171
<i>Rosa M Poch and Georges Stoops</i>	
Soil Classification and distribution	
History and principles of soil classification	185
<i>Maria Gerasimova and Dmitry Konyushkov</i>	

Soil classification systems: World reference base and soil taxonomy	197
<i>David Badía-Villas</i>	
World reference base for soil resources (WRB)	206
<i>Stephan Mantel, Stefaan Dondeyne, and Seppe Deckers</i>	
World soil map based on soil taxonomy	218
<i>Eswaran Padmanabhan and Paul F Reich</i>	
Soils as archives of city history	232
<i>Hermine Huot and Jean Louis Morel</i>	
Classification of soils by ecozones	
Ecozones and soils – An introduction	246
<i>Rupert Bäumler</i>	
Cold Region Soils: Part I—Distribution, Properties and Processes	250
<i>Nicolas A Jelinski, Chien-Lu Ping, and John CF Tedrow</i>	
Cold region soils: Part II – Genesis and classification	263
<i>Nicolas A Jelinski, Chien-Lu Ping, and John CF Tedrow</i>	
Boreal forest soils	280
<i>Gabriele Broll</i>	
Soils of humid cool temperate regions	289
<i>Nicolas A Jelinski, Justin B Richardson, and Edward A Nater</i>	
Grassland soils in the cool–arid–temperate ecozone (Steppe, grassland)	299
<i>Mengyang You, Xiaobing Liu, Rupert Bäumler, and Margaret A Oliver</i>	
Soils from a warm-temperate semi-tropical ecozone (Mediterranean) with humid winter months	319
<i>J Torrent, Marco Pfeiffer, and Juan José Ibañez</i>	
Arid and semi-arid subtropical and tropical ecozone (II)	332
<i>Bernhard Lucke</i>	
Humid subtropical ecozone with emphasis on Acrisols and Alisols	343
<i>Nilton Curi, Sérgio Henrique Godinho Silva, and Alberto Vasconcellos Inda</i>	
Soils of the humid and sub-humid tropics	353
<i>Karl Stahr and Ludger Herrmann</i>	
Classification of soils by properties and environmental settings	
Sodic soils	368
<i>Guy J Levy and Uri Nachshon</i>	
Organic soils	383
<i>Heike Schimmel and Wulf Amelung</i>	
The pedology of anaerobic soils	398
<i>John Hollis and Allan Lilly</i>	
Alluvium and alluvial soils	417
<i>Janis L Boettinger</i>	
Hydric Soils	424
<i>MJ Vepraskas and JF Berkowitz</i>	
Wetland soils	436
<i>Cornelius Wilhelm van Huyssteen</i>	
Paddy soils: Characteristics, productivity and management	447
<i>Stephan M Haefele</i>	
Volcanic soils and non-volcanic andic soils	458
<i>Rupert Bäumler and G Uehara</i>	

PEDOMETRICS**Pedometrics**

- Overview of Pedometrics 471
Alexandre MJ-C Wadoux, Inakwu OA Odeh, and Alex B McBratney

Statistics and soil data analysis

- Statistics in soil science 486
Richard Webster

- Multivariate analysis for soil science 499
Si Yang Han, Patrick Filippi, Mercedes Román Dobarco, Januar Harianto, Mathew S Crowther, and Thomas FA Bishop

- Machine learning and processing of large data 509
Madlene Nussbaum

- The issue of scale and change of support in the spatial analysis of environmental data 521
A Castrignanò, A Belmonte, and N Romano

Soil measurements in space and time

- Sampling for soil survey 535
DJ Brus

- Handling of soil samples in the field, in transport to the laboratory and in the laboratory 551
Claudy Jolivet, Céline Ratié, Déborah Chavrit, Nicolas Soler-Dominguez, and Antonio Bispo

- Digital Soil Morphometrics 568
Yakun Zhang, Alfred E Hartemink, Jingyi Huang, and Budiman Minasny

- Proximal soil sensing in the field 579
Mikaela Tilse, Uta Stockmann, and Patrick Filippi

- Proximal sensing of soil moisture 591
Markus Tuller, Ebrahim Babaeian, and Scott B Jones

- Proximal sensing of land surface temperature 600
Ebrahim Babaeian and Markus Tuller

- Proximal sensing of evapotranspiration 610
Ebrahim Babaeian and Markus Tuller

- Remote sensing of soil moisture 618
Alexander Gruber and Jian Peng

- Remote Sensing of Soil Organic Carbon 631
Bas van Wesemael and Sabine Chabrilat

- Pedotransfer functions and their application to soil water dynamics 642
Marcel G Schaap, Yonggen Zhang, and Attila Nemes

Spatial and temporal soil variation

- Geostatistics: Modelling spatial variation 655
Richard Webster

- Uncertainty assessment of spatial soil information 671
Gerard BM Heuvelink and Richard Webster

- Digital soil mapping: Evolution, current state and future directions of the science 684
Brendan Malone, Dominique Arrouays, Laura Poggio, Budiman Minasny, and Alex B McBratney

- Geographical information systems (GIS) and soils 696
S Grunwald and J Böhner

- Precision agriculture 710
James A Taylor

Foundations, measurements and trends in pedodiversity <i>Juan José Ibañez and Marco Pfeiffer</i>	726
Soil landscape modelling	
Modeling soil development in a landscape context <i>Peter A Finke, Budiman Minasny, and Arnaud JAM Temme</i>	739
Climate and land surface models: Role of soil <i>Toby Richard Marthews, Holger Lange, Alberto Martínez-de la Torre, Richard J Ellis, Sarah E Chadburn, and Martin G De Kauwe</i>	751
VOLUME 5: SOIL PHYSICS	
Soil structure & mechanics	
Soil structure <i>Steffen Schlüter and John Koestel</i>	1
Aggregation <i>Teamrat A Ghezzehei</i>	8
Porosity and pore-size distribution <i>John R Nimmo, Sheela Katuwal, and Maik Lucas</i>	16
Pore-scale modelling of flow and transport phenomena in soils <i>Kirill M Gerke and Marina V Karsanina</i>	25
Soil texture: Its relevance, measurement and representation internationally <i>Attila Nemes</i>	35
Soil tilth <i>Thomas Keller, Douglas L Karlen, and Paul D Hallett</i>	48
Crusts and seals: Structural <i>Shmuel Assouline and Sally Thompson</i>	57
Soil mechanics <i>Markus Berli and Paul D Hallett</i>	69
Compaction <i>JJH van den Akker, L ten Damme, M Lamandé, and Thomas Keller</i>	85
Bioturbation—Physical processes <i>Siul Ruiz, Paul D Hallett, and Dani Or</i>	100
The reinforcement of soil by plant roots <i>Alexia Stokes and Anil Yildiz</i>	115
Physical effects of soil salinity <i>David Russo</i>	123
Shrink-Swell processes in soils <i>Pascal Boivin</i>	139
Water retention	
Water properties <i>N Lu, C Zhang, and D Hillel</i>	146
Water potential <i>Dani Or, Markus Tuller, and Jon M Wraith</i>	159
Water content and water potential measurement <i>Wolfgang Durner</i>	168
Soil water retention and characteristic curve <i>Markus Tuller and Dani Or</i>	187
Soil water repellency <i>Jörg Bachmann and Marc-Oliver Goebel</i>	203

Flow & transport processes

Aeration	216
<i>Rodney T Venterea and Dennis E Rolston</i>	
Capillarity	225
<i>Dani Or and Markus Tuller</i>	
Darcy's law	240
<i>Conrad Jackisch and Eva Kröner</i>	
Diffusion	247
<i>TKK Chamindu Deepagoda and Shoichiro Hamamoto</i>	
Hydrodynamics in soil	256
<i>Ty Ferre and Art Warrick</i>	
Thermodynamics of soil water	262
<i>Pieter H Groenevelt</i>	
Immiscible Fluids	271
<i>Yukie Tanino and Dubravka Pokrajac</i>	
Overland flow	287
<i>Tammo S Steenhuis, Rajith Mukundan, and Seifu A Tilahun</i>	
Direction-dependency of soil hydraulic and mechanical properties	294
<i>Steffen Beck-Broichsitter, Heiner Fleige, Jose Dörner, and Rainer Horn</i>	
Macropores and macropore flow	306
<i>Nicholas Jarvis and Mats Larsbo</i>	
Unstable preferential flow in porous media	314
<i>Naaran Brindt, J-Yves Parlange, and Tammo S Steenhuis</i>	
Leaching processes	321
<i>Brent Clothier and Steve Green</i>	
Colloid transport	336
<i>Yan Jin, Chongyang Shen, and Volha Lazouskaya</i>	
Scaling of transport properties	353
<i>Robert Ewing and Behzad Ghanbarian</i>	
Soil influence on groundwater	364
<i>David Rudolph and Edwin Cey</i>	

Thermal properties and behaviour

Temperature effects of soil hydraulic properties	379
<i>Jan W Hopmans and Robert Horton</i>	
Heat and vapor transfer	385
<i>Sunny Eng Giap Goh, Joshua L Heitman, and Robert Horton</i>	
Radiation Balance	393
<i>Jerry L Hatfield, Kaitlin Togliatti, Thomas J Sauer, and John H Prueger</i>	
Freezing and thawing cycles	400
<i>Alan W Rempel</i>	
Arctic permafrost	410
<i>Julia Boike, Andrey Abramov, Katrina E Bennett, and Lars Kutzbach</i>	
Permafrost and cold region hydrology	419
<i>Masaki Hayashi and Ylva Sjöberg</i>	

Sensing Techniques

Soil water sensing by neutron scattering	427
<i>Steven R Evett</i>	

Time-domain reflectometry	436
<i>GC Topp and Ty Ferre</i>	
Geophysical methods for soil applications	444
<i>Sarah Garré, Guillaume Blanchy, David Caterina, Philippe De Smedt, Alejandro Romero-Ruiz, and Nataline Simon</i>	
Non-invasive imaging of soil processes	459
<i>Craig J Sturrock</i>	
Hydrology & water balance	
Soil processes in the hydrologic cycle	469
<i>Cosimo Brogi, Harry Vereecken, Heye Reemt Bogena, and Luca Brocca</i>	
Global water cycle from a soil perspective	482
<i>Michael H Young and Dani Or</i>	
Water evaporation from bare soil	490
<i>Peter Lehmann</i>	
Evapotranspiration	501
<i>Yuting Yang</i>	
Plant available water	509
<i>Quirijn de Jong van Lier, Sally D Logsdon, Everton Alves Rodrigues Pinheiro, and Paulo Ivonir Gubiani</i>	
Plant-water relations	516
<i>Ian C Dodd, Andrew D Hiron, and Jaime Puértolas</i>	
Water use efficiency: A review of spatial and temporal variability	527
<i>Simone Fatichi, Athanasios Paschalis, Sara Bonetti, Gabriele Manoli, and Christoforos Pappas</i>	
Root water uptake and myco-rhizosphere hydraulic properties	543
<i>Andrea Carminati, Maria Marin, Valentin Couvreur, Mohsen Zarebanadkouki, Stéphane Declerck, and Mathieu Javaux</i>	
Plant-soil-water relations	552
<i>Jan Vanderborght, Quirijn de Jong van Lier, Andrea Schnepf, and Mathieu Javaux</i>	
Crop and landscape water requirements	571
<i>Luis S Pereira, Isabel Alves, and Paula Paredes</i>	
Water harvesting	593
<i>Wolde Mekuria and Desalegn Tegegne</i>	
Irrigation methods	608
<i>Shmulik P Friedman</i>	
Soil–plant–atmosphere continuum	624
<i>John M Norman, Martha C Anderson, and William P Kustas</i>	
The Penman-Monteith equation	634
<i>Richard G Allen, Ayse Kilic, Guido D'Urso, and Simona Consoli</i>	
Infiltration	646
<i>Ty Ferre and Art Warrick</i>	
Solute transport	652
<i>MC Sukop and M Dentz</i>	
Lysimetry	667
<i>Thomas Pütz and Jannis Groh</i>	
Recognition of notable past soil scientists	680
<i>Michael J Goss, Evgenia Blagodatskaya, Ariena HC van Bruggen, Monika Joschko, Stefan Schrader, and Christoph Tebbe</i>	
<i>Index</i>	691

ENVIRONMENT AND MANAGEMENT

Soil functions, soil potential uses and ecosystem services

Karl Stahr, Department of Soil Science and Land Evaluation (310a), Universität Hohenheim, Stuttgart, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction: The importance of soils in the environment and landscape	2
Soil is a four-dimensional body	2
Why soils matter?	3
Soils are natural bodies	3
Soils as archives of natural and cultural history	3
Soils are natural aesthetic objects'	4
Potentials offered by soils	4
Soil functions must be protected	5
Soils are habitats for organisms	5
Soils produce food and fodder	5
Soils produce raw materials	6
Soils support energy production	6
Soils are a genetic resource	7
Soils have a transformation potential	7
Soils are air filters and regulators	7
Soils are a filter and buffer, and regulate the water cycle	7
Soils are a source of raw materials for technical processes	8
Soils cover a certain area	8
Soils have a load capacity	8
Soils can be used for transport	8
Soils can take residues	8
Soils are recreation areas	8
The role of soils in ecosystems and landscapes	9
Soils are environmental filters and buffers	9
Soils control the water regime	9
Soils support ecosystem services	9
Water, air and soil: The basic environmental components	10
History of soil ecology	11
Conclusions and outlook	11
References	11

Abstract

In soils, the lithosphere, atmosphere, hydrosphere and biosphere overlap to form the pedosphere. Soils have the potential to benefit humans and the environment. The goods and services potentially rendered by soil are called soil potentials; the active services provided by a particular soil are soil functions. Soils with organisms living in them are considered ecosystems. Soils together with ecosystems play a regulating role in larger units like landscapes. Ecosystem services include people. The studies of soil ecology and ecosystem services were initiated in the 1970s. Among all types of degradation the biggest problem for soil protection worldwide seems to be soil erosion and the increase of settlements in fertile areas.

Glossary

Abiotic potentials Represent potential soil services, where the soils regulate the water regime, clean the air and provide technical raw materials.

Areal potentials Offer soil services for any type of use.

Biotic potentials Reflect potential soil services, where the soils are used as living space.

Ecosystem services Are the benefits people obtain from ecosystems.

Soil functions Are services, which soils under a specific land use carry out for humans and the environment.

Soil potentials Relate to services, which have the potential to benefit humans and the environment.

Introduction: The importance of soils in the environment and landscape

The soils on our planet are all useful. They provide the ecological balance, and primarily they allow plants and animals to thrive (Blume et al., 2016). Civilizations have used them for various purposes, which leads to the questions: How can a particular soil be used, or is this specific soil suitable for the required use?

These questions are as old as human interactions with soils. It started in the Younger Stone Age (Neolithic, began ~10,000 years ago), when people cleared forests to convert them into agricultural lands (Stahr et al., 2020). The earliest Greek and Roman approaches to soil classification were based on agricultural suitability.

In contrast, the beginning of modern Soil Science started with approaches to understand soil development. One of the founders of modern soil science, Dokuchaev (1883), observed soils across the Russian plateau and elaborated the dependence of soils (independent natural bodies) on soil forming factors, especially climate. Building on these ideas the most important soil philosopher, the Swiss Hans Jenny (Jenny, 1941), developed the central formula of soil formation:

$$\text{Soil} = f(t) (P, C, O, H, R)$$

P = parent material, rock or sediments, C = climate, O = organisms, plants and animals, H = humans, R = relief, t = time.

He later tried to quantify the efficiency of the individual factors over time (Jenny, 1980). Schlichting (1986) conceived rock and litter as the two factors that are transformed into soil, while the other factors (climate, relief, flora, fauna, humans) catalyze or regulate the course of transformation.

$$(\text{Rock} + \text{Litter}) \times \text{Time} \xrightarrow{\text{climate} \text{---} \text{relief} \text{---} \text{flora} \text{---} \text{fauna} \text{---} \text{humans}} \text{Soil}$$

These ideas increased the understanding of soil development, but diverted from the original idea to explain soil productivity and allocate land-use capabilities, which are delivered by our soils.

Meanwhile, the widely accepted approach is to describe and classify the soils according to their properties and genesis, where each particular site has its own influence, and therefore offers specific potentials and functions.

On the other hand, we have to find measures to evaluate the abilities of the same types of soil to provide goods and services, and to understand how to protect these from degradation. Soil classification and a soil suitability rating are two different approaches. The first tries to characterize a soil itself, while the other estimates the response of a soil to the demand of a certain land use. There is no soil suitable for all possible land uses; i.e. there is a specific soil suitability for growing rice (*Oryza sativa* L.), for grazing cattle, growing asparagus (*Asparagus officinalis* L.), or playing football.

Soils can be conceived as the skin of the Earth. The utmost importance of soil for the environment becomes clearly visible from this statement. The skin always protects the body within from extreme external influences. That means, it has a stabilizing influence on the water, air and temperature regimes. It buffers the extreme external changes in the atmosphere. It also covers most of the ground surface of the earth, including mountains, deserts and oceans (there are soils, which are covered with water all the year). Soil controls the vertical fluxes of air, water and energy. But soil properties at any one place differ widely from those at others according to the different combinations of soil forming factors. Therefore, the influence on the magnitude and direction of these processes will change from place to place. In landscapes we often find a regular catenary pattern of soils based on different landforms. These soils (soil landscapes) will affect the processes that occur in the landscape and provide support for infrastructure and cultural services.

Soil is a four-dimensional body

Soils have a carrying capacity. We use soils daily by standing, walking or driving on them. By doing so we see a two-dimensional section, i.e. a plane.

As soon as we acknowledge that plants penetrate the soil with their roots, or that water is infiltrating and being stored, we perceive the soil as a body. All types of soil potentials rely on the fact that solids, liquids and gases make up the soil body. The use of soils for plant production, energy exchange, water storage, biological and ecosystem processes requires a volume of soil to be effective. Therefore, a soil must be conceived of as a three-dimensional body at least.

Pushing with a spade into the garden soil at different times of the year allows us to realize that some properties change with time. The soil may be wet or dry, sticky, plastic or hard, light or dark colored, warm, cold or even frozen. All of these properties change reversibly with time. However, water percolating through the soil may dissolve substances and leach them irreversibly out of the soil and so change its properties over time. The same happens, when silty dust containing 'heavy' metals is eventually deposited and accumulates on top of the soil. Dead plant or animal residues will be transformed to soil organic matter (humus), and primary

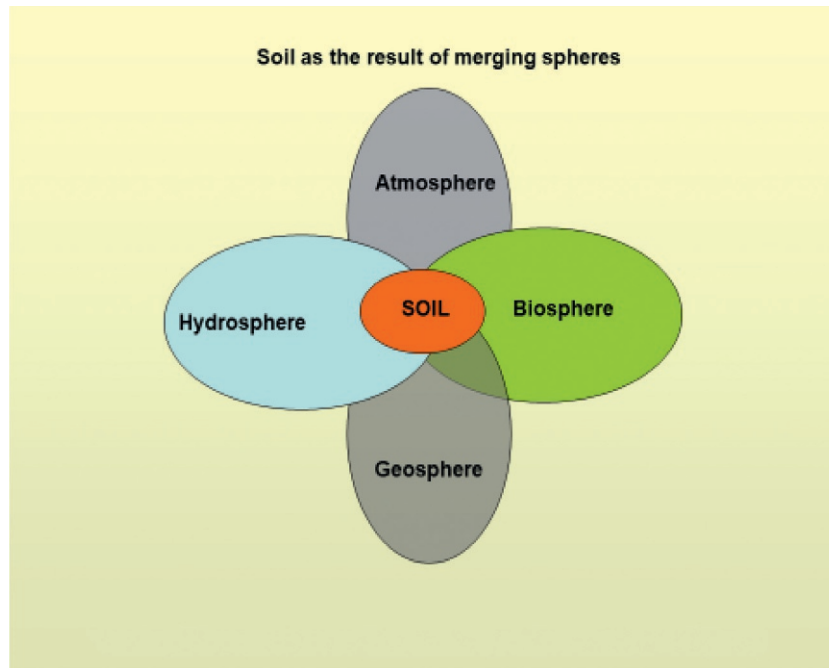


Fig. 1 Relation between the pedosphere and the other spheres. Soil is in the overlap of the other spheres.

minerals into secondary minerals. Soils in general undergo irreversible changes. Extreme changes may occur, for example, if carbonate is leached in karstic areas over millions of years it can leave a deep acid clay rich soil. We can now realize that soils are four-dimensional objects.

“Soils are natural bodies and as such four-dimensional parts of the upper Earth crust. Herein rock (lithosphere), water (hydrosphere), air (atmosphere), and living organisms (biosphere) are interlinked” (Fig. 1).

Why soils matter?

Soils are a part of our environment and their protection should be mandatory as is the protection of all other environmental components (air, water and biota).

Soils are natural bodies

Soils as four dimensional bodies occur in the environment, which means in nature. In principal all soils can be formed under natural conditions and they are part of landscapes. Even anthropogenic soils find their place in the natural environment.

Like rare birds, precious orchids or plants we also have rare soils, which may occupy very little space in the pedosphere. They often have very specific functions in the soil cover such as allowing endemic plant species to grow or influencing different fluxes in a particular way. An example may be that they provide the precipitation of particular secondary minerals. The disappearance of such soils may change the water and energy balance significantly as well as the structure of the landscape. These soils need to be protected just as the protection of ecosystems is foreseen.

Soil processes often take a very long time to change the properties of the parent material or pre-existing soils significantly. Therefore, features that developed long ago are often still visible today in paleosols. For example, in Luvisols derived on loess, we find relicts of wooden poles from houses of the Younger Stone Age or flint knives in the ancient rubbish hollows or middens of the same people. Another example is the paleosols, which occur in the Sahara and that act as documents of earlier humid climates. Soils can be important archives of the Earth's history. Destruction of those soils is an irreversible loss of natural bodies and records of natural and cultural history.

Soils as archives of natural and cultural history

Soils are often laboratories of nature with a high productivity for example of reduction, oxidation or emission. This has to be protected. Likewise with precious crystals, we might observe soils and realize that they have admirable colors and structures. Therefore, it is important to preserve the natural beauty of such soils. It will be only a small proportion of soils, that we should protect because of their aesthetic and historical information. However, we should foresee these needs.

Soils are natural aesthetic objects' (Kubiena, 1946)

Furthermore, we can take into account that our soils as a whole deliver services in the regime of nature and specifically for the desires of human society. This is still the main reason why we have to protect the soil cover in order to be able to use its functions at any time. Yes, soils produce food and fodder and control environmental processes.

Potentials offered by soils

All soils are part of an environment. Our knowledge gathered during the last 200 years allows us to estimate the soil's potential much better as long as the soils are not heavily disturbed. A specific soil always has a function, e.g. it is used for growing potatoes.

At any time, the same soil has several types of potential. In accordance to its properties it may be used as a sports field, a footpath or a raw material for bricks. It may also have the potential to become a nature reserve, if different types of vegetation are present to provide habitats and food for birds and other animals. The potential goods and services a particular soil is able to provide should be visualized before making decisions on which function the soil should be used for.

If one asks, which principal possibilities soils have to serve for humans and in the environment, three groups of potential can be identified: These are the biotic potentials, abiotic potentials and areal potentials (Table 1).

The biotic potentials are related to the existence and growth of plants, animals and microorganisms, which need the soil as a living space to deliver food and energy. We may separate this potential in different groups, for example the production of food (wheat (*Triticum aestivum* L.), sugar cane (*Saccharum officinarum* L.), bananas (*Musa acuminata*), potatoes (*Solanum tuberosum* L.), rice (*O. sativa* L.), ...), the production of raw materials (timber, sisal fibers (*Agave sisalana* Perrine), cotton (*Gossypium hirsutum* L.), ...), the production of crops for energy (firewood, oil, ...), and the possibility to guarantee species survival. Many species have their seed bank in the soil, which means their reproduction is closely connected with the soil. Many microorganisms, which may be still unknown for example, can only be protected, provided that we give them the opportunity for reproduction in soils. The most important biotic potential is the one possibility for transformation. Here organic material as well as mineral materials existing in the soil or introduced into it can be altered and have a different appearance thereafter. Soils are exciting environmental laboratories. On the other hand, this potential is the transition between biotic and abiotic potentials because organic and mineral components can be transformed.

The abiotic potentials are those of gathering and filtering water, cleaning air and providing technical raw materials. Soils are a reservoir of water, and during periods, when the water balance is positive, they generate percolating water, which supplies groundwater and springs. In general, groundwater and spring water have drinking water quality. In Germany, the annual mean amount of water delivered is about 100 l per m² with a range from zero to more than a thousand. This range is also valid worldwide. Input of dust, gases or other elements that enter the soils can be filtered and fixed. Therefore, soils are also air and water filters. Finally, many soil components like stones, gravel, sand, clay, infusorial (diatomaceous) earth or lime may be useful raw materials for construction, to produce ceramics or fertilizers. Here soils act as sources of raw materials.

The areal potentials offer the place and space covered by each soil. Society generally honors these types of potential most. A parking place for a car is generally more expensive than the equivalent area for growing wheat, which could be produced on it. Such 'places to stand' may be offered to a person, a house, a factory or others. Important and increasing areas of land are used for transport. Furthermore, soils can be used as deposits for rubbish or construction materials and poisonous or toxic substances. Often places, where raw materials were first mined or extracted, become areas of waste deposits later that are then probably contaminated. The use of soils is also important for recreation. Recreation includes the admiration of nature, garden space, sports and playgrounds for children.

Table 1 Overview of all potentials that can serve ecology and society.

(1) Biotic potentials
a. Habitat for biota
b. Growth of food and feed crops
c. Growth of materials (e.g., cotton)
d. Growth of energy crops
e. Genetic resource
f. Transformation potential (the soil lab, also abiotic)
(2) Abiotic potentials
a. Air filter and regulator
b. Filter, buffer and supplier in the water cycle
c. Source of raw materials
(3) Areal potentials
a. Load and construction capacity
b. Traffic route
c. Deposit (landfill)
d. Recreation area



Fig. 2 Copulation of earthworms (*Lumbricus terrestris*) on the soil surface of a Vertic Cambisol.

If we want to differentiate between the quality of all types of potential, we realize that all biotic types can in principle be used reversibly. That means the use will maintain the potential. This accords with the transformation potential and also the abiotic potential of storing, infiltrating and delivering water and cleaning air, and it also applies to the recreation potential. On the other hand, extraction of raw materials and some uses of the areal potential will destroy other types of soil potential irreversibly. Utilization of types of reversible potential might also affect soil features that usually guarantee the potential to provide a particular service. However, the soils might have been used in such a way that the potential is destroyed or reduced over time.

Soil functions must be protected

Soils are habitats for organisms

The biosphere intersects with the pedosphere, therefore many organisms live in the soil. The dominant families here may be termites (Isoptera) and earthworms (Lumbricidae). In addition to these two groups manifold other macro-organisms need the soil as a habitat such as beetles, mice, moles, squirrels, gophers and so on (Dunger, 2008). They are responsible for forming and maintaining the soil's structure, which in turn is needed by plants and meso- and micro-fauna to grow successfully. The living conditions of these organisms need to be preserved in order to avoid their migration or even extinction.

Soils must be preserved as natural bodies. That means a certain minimum area is required to provide sustainable populations of soil organisms (Coleman and Crosley Jr, 1996). The aim is to preserve not only rare soils, but also the more regular, widespread soils under nature oriented conditions. Increasing the intensity of use may result in the loss of natural properties (Stahr et al., 2020) (Fig. 2).

Soils produce food and fodder

The production of food and fodder is extremely important for the maintenance of human life. Huge areas of soil are predominantly covered by pastures to feed a range of livestock such as sheep, goats, camels, horses and cattle. These natural pastures are in danger from overgrazing, and wind and water erosion. However, in intensive agriculture maize (*Zea mays* L.) or soybeans (*Glycine max* (L.) Merrill) are important food and fodder products. Crops need the soil as a rooting zone, for water supply, proper aeration, sustaining a particular temperature regime, providing nutrients and stabilizing the site. Food crops like maize, wheat (*T. aestivum* L.), barley (*Hordeum vulgare* L.), potatoes, sugar beet (*Beta vulgaris* L.), sugar cane and others make specific demands on the soil to succeed, which should be provided naturally or might need to be maintained by various agricultural practices. Worldwide, rice is a crop derived from floodplains that is cultivated in paddies, which need specific site characteristics and particular irrigation procedures to succeed. Many food crops that require specific conditions such as banana (*M. acuminata*), olives (*Olea europaea* L.), cacao (*Theobroma cacao* L. subsp. cacao), vine (*Vitis vinifera* L.), citrus (*Citrus* sp.), apples (*Malus domestica* Borkh.) and others are successfully cropped under a combination of soil and climate conditions. Sustainable agricultural practices must maintain soil structure, water and nutrient supply, and avoid toxic elements and soil erosion. When we want to use soils to reduce hunger from food insecurity, we should investigate its cause and potential health effects. The potential for food production often exists but there is often a lack of awareness of it. The increase in yields during the 20th century in many countries and crops demonstrates this; the 'green revolution' (Beek, 1978; McRae and Burnham, 1981).

However, a major problem worldwide is the loss of agricultural land to other uses like transport or settlement. The best soils are often in the neighborhood of growing big cities (consider Cairo, Egypt) (Fig. 3).



Fig. 3 Cows and Horses find their fodder in an alpine valley on Histosols and Histic Gleysols.

Soils produce raw materials

Natural raw materials can be produced in forests and agricultural fields, and have played an important role since humans settled on Earth. During the recent centuries of industrialization, it seemed that natural raw materials could be replaced increasingly by technical products. This tendency is now in reverse. Timber, cottonwood for paper, sisal fibers and other such products have gained greater importance again. Some of these crops can adapt to extreme conditions e.g. saline soils, very acid soils and steep slopes. This reduces the competition on more favorable sites to allow the production of food crops. Another benefit of the crops that provide raw materials is that their resistance against toxic substances is generally higher. Finally, the products produced from such a range of soils can be technically or naturally recycled. Therefore, this soil function will remain important, but must be brought into equilibrium with the demand for food by humankind.

The above situation could be solved by the co-cultivation of food and timber or other raw materials. On an agricultural field the production of cotton or wheat excludes the other, but there are other possibilities. There is the question: Does the production of mushrooms reduce timber production. In some situations the cultivation of one plant depends on another. For example, cocoa production needs shadow from other trees, but its production clearly needs space, which reduces timber growth. That means co-cultivation may be beneficial, but generally reduces the growth of one or both cultivars.

To maintain soil function, its natural fertility must be maintained, which may benefit from co-cultivation in some places. Fertility may be in danger through loss of nutrients, strong acidification or strong alkalinization. Under specific conditions monocultures, which are at environmental (soil chemical) limits can undergo continuous changes. This will result in unfavorable conditions for the continuation of the specific culture. The biggest problem seems to be the danger of soil erosion during or following harvest, when the crops are changed and the land may be left bare.

Soils support energy production

Rapeseed (*Brassica napus* L.) or firewood are crops grown for energy. However, assimilation by plants in general is a process which collects energy. Therefore, in principle, all soils on which plants grow are sites that store energy. However, in most cases the purpose of the growth is not the latter. In the case of plant material for firewood the use as fuel is clear (Kölling et al., 2007). To gain the energy back the plants may be burned, transformed into gasoline or into biogas.

In many cases the question as to whether the soil can support energy production is only answered after harvest. For example, maize can be grown as human food, as animal fodder and for biogas production. The order may even go in a sequence of the first being fodder with the following product, i.e. animal manure, used for biogas production. To maintain the above function, the stability of nutrient cycles and avoidance of soil erosion are again important considerations (Fig. 4).



Fig. 4 Rapeseed (*Brassica napus* L.) collects energy from soil (Rendzic Leptosol).

Soils are a genetic resource

All organisms need a specific environment in which to live and survive. Many organisms from microorganisms to vertebrates use soils as their habitat. Earthworms (Lumbricidae) and termites (Isoptera) are characteristic soil specific organisms.

The influence of the overall environment may induce genetic changes in soil biota. This is especially true for microorganisms. Therefore, we can expect that new microorganisms are created by the influence of the soil or that others are genetically altered through environmental conditions provided by internal soil properties. Many microorganisms remain to be identified in soils. Also new genetic varieties of macroorganisms may find opportune conditions for survival and multiplication.

Soils have a transformation potential

In soils nothing is really stable. Changes in conditions and properties can transform most substances that occur in soils. This may be the greatest potential and function that soils have. It can be a change in organic as well as mineral materials, and there can even be a transformation of the one form into the other. In general, these transformations generate organic substances from organic materials and minerals from minerals. However, mineral carbonates or nitrates may be formed from organic materials or mineral phosphates may be included in organic molecules.

There are many examples of the above: beech (*Fagus sylvatica* L.), maple (*Acer* L.) and oak (*Quercus* L.) leaves are transformed to humic substances; crystalline silicates are transformed to clay minerals, gibbsite or opal; iron will be released from rock minerals and precipitated in the form of different oxides; Sulfides can transform into sulfates under oxidizing conditions, but may be reduced again to other types of sulfides under reducing conditions. Even technical organic or inorganic products, which do not occur in nature, can be changed into natural products. This transformation function is at the border between biotic and abiotic functions.

Soils are air filters and regulators

External and internal soil surfaces can bind particles, which are dispersed in the air. Volatile gaseous molecules and ions can bind to soil surfaces. In general, the air quality over natural soils is better than that over sealed areas. In fact, vegetated soils in subhumid and humid areas are perfect air filters. On the contrary, deserted areas are sources of dust and locally even sand. Gases may also be emitted from volcanic areas that may acidify the soils around. Anthropogenic and specific agricultural desertification will often turn the filter into an emitter. Therefore, human activities should maintain the soil in conditions that can fix particles and gases. For example, aggregates on the soil surface should be stable, or surfaces should be kept vegetated and moist.

Soils accumulate particles to a large extent, for example loess or dune sands. Such additions may totally change the properties of the underlying soils, and they may be transformed into completely different soils.

Soils are a filter and buffer, and regulate the water cycle

All soils contain a liquid phase; this is an aqueous solution. Each soil has a pore volume, which can be filled with water or solution. Soils have a certain infiltration capacity. The water infiltrated may be stored or leave the soil and move into deeper rock pore systems. On the other hand, the water may evaporate or be taken up by plants. In humid periods water infiltrates the soil and is stored until the water retention capacity is reached. Amounts of water greater than the retention capacity may become seepage water.

Water flow is controlled by the gradient of the hydraulic or matric potential and the hydraulic conductivity. Therefore, depending on these regulating values water or the soil solution may move in any direction. Because gravity influences the hydraulic potential, vertical movement has priority, but the low hydraulic conductivity of certain subsoils or parent rocks causes an often dominant lateral water movement.

The important point is that the physical processes do not take place independently. The solution always reacts with the internal surface of the soil. It may exchange cations and anions with the internal surfaces; it may dissolve or precipitate salts and it may transport particles for some distance. Water, which has passed through soil and rock is often of drinking quality, which rainwater before entering the soil never was.

The water regulating filter and buffer capacity of soils must in general be seen as an advantage for the environment. This function has very different values for different types of soil. A soil may store between 1 and 400 L m⁻³ of water, and it may receive annually between 10 and more than 2000 L m⁻² of water. The hydraulic conductivity may change by up to a factor of a million times.

The different values that determine the importance of a soil in ecosystems are strongly related to soil fertility. Regulation of the water cycle is directly related to the natural soil condition. However, users must take care to maintain the benefits. Compaction, disturbance of pore connections, erosion and deposition can markedly alter the function of soils in the water cycle. However, there are also possibilities to improve structure and water exchange by appropriate activities. Land use must take care to maintain and improve the water regime.

Soils are a source of raw materials for technical processes

Soil materials can often be used as raw materials. Most widespread is the use of loams, which can be burned to make bricks. Other more specific and precious clays are used for ceramics such as porcelain or the widely used siliceous earth. Furthermore, lime as a basis for mortar or cement and also sand and gravel are used as construction materials. To extract these materials specific soils and sites must be found.

The fundamental difference between this function and all others mentioned above is that here the use destroys the potential as well as the soils. The materials used are altered in such a way that recycling is almost impossible. An exception are the clay bricks made by sun drying loam in dry and hot areas.

Soils cover a certain area

Soils have a load capacity

On most soils you may stand. By doing so, you use the areal function. You may also put other things like your car or your house on top of a soil. The two ways that you could use this potential are: 1. You may place something on a soil within the load bearing capacity and later take it away to put it elsewhere. 2. You construct something in or on top of a soil like a gas station or a sky scraper. The first is sustainable, the second destroys further use, especially other potentials for millennia.

Soils can be used for transport

You can ride your horse across grassland. This seems anarchic, but it is a principal function a soil may offer. Animals, humans and all types of vehicles may use the soil to move across. Nowadays, this function is regularly combined with the load function. First, a road or a highway is constructed, which seals the soil, and then different forms of transport may take place. We observe that worldwide roads cover an increasing part of the land. Therefore, it is of importance to construct them in such a way that allows later withdrawal. Most roads are repaired or reconstructed over a time scale of 20 years, but they are made as if it is for ever. Here engineers must learn to protect soils (Fig. 5).

Soils can take residues

In all agricultural, horticultural and forestry functions by-products or wastes occur. In this case the products, in principal, can be brought back to the fields directly as mulch or after stabilization by, for example, composting, in which a complete recycling should occur.

On the other hand, industrial or mining by-products cannot be returned to the environment from which they were released. In these case soils must take the residue load by deposition, which leads to complete alteration of the soil properties and their potential. In such cases toxicity often has to be taken into account. Also the change is long lasting, if not permanent. To control the situation all such effects should be minimized during operations. If the 'clean-up' of a site is required the surface soil should first be removed. The deposit should produce no leachates and at the end of the 'clean-up' process about three meters of uncontaminated soil material should be put on top to allow reintegration of sites into the landscape ecosystems (Fig. 6).

Soils are recreation areas

You may lay down on a soil just for fun and recreation. Visiting a mountain landscape, climbing a range of hills and enjoying the feeling or just the landscape view are also soil services for recreation. Walking or cycling along trails through an environmentally attractive reserve is again a good soil service. Sports are also taken for recreation. In all these cases the simple existence of soils in attractive landscapes is sufficient. However, this function is often combined with the use of the load capacity, where sportsgrounds are constructed or with the traffic function, where trails must be used. Therefore, in all those cases the environmental load must be minimized and reconstruction must be considered.



Fig. 5 Soils (Vertic Cambisol) are used for traffic while maize (*Zea mays* L.) is harvested.

The role of soils in ecosystems and landscapes

Soils are environmental filters and buffers

The pedosphere is an overlap between the lithosphere, hydrosphere, atmosphere and biosphere (Fig. 1). Soil pore systems allow contact between the different phases. The pore system also provides the opportunity for transport in gaseous, liquid and solid phases. Living organisms can also use the pore system or change the pores and soil structure to facilitate their existence. They may even change the pore system in favor of their demands.

During all these transport processes physical, chemical and biological contacts and reactions occur, where pollutants, nutrients and other elements can be adsorbed and desorbed until they reach equilibria. This is how soils function as buffers in the environment. The simplest case is the transport of particles of minerals or organic compounds, which reach soils through transport from outside the soil system (pedosphere). When they enter soils their movement velocity slows down or is stopped. They are thereby enriched in the soil and the soil itself is the filter.

Soils control the water regime

Soils have an eminent importance in the water regime of any landscape. The average water retention is permanently 1000 m³ per ha against gravity. The same amount of water can be stored in addition in the short term, but it will later become stagnant or leach out to groundwater or springs. During rain events water can infiltrate into soils provided the rate of infiltration is large enough. Otherwise the danger of erosion is considerable. Water retention provides the guarantee for plant water supply and its regulation.

Water has a regulation function in soils because it is mobile in all directions. The vertical downward transport starts with infiltration, leaching and drainage. The upward movement starts with capillary rise, plant uptake, transpiration and evaporation. Lateral water movement along a hydraulic gradient is also often important. All these transport processes not only carry water but also dissolved elements, ions and suspended solid particles.

One of the most exciting transport- and soil-forming processes is the illuviation of clay particles during strong rains in mildly acid soils. The transport of humic substances and aluminum is restricted to strongly acid and very humid conditions. Under the same conditions, and especially under reducing conditions, iron and manganese can be transported in any direction. The same can happen with lime (calcium carbonate) under neutral and mildly alkaline conditions. Finally, gypsum and soluble salts can be transported under neutral and alkaline pH mainly under semiarid conditions. Therefore, we can conclude that the water regime is a major controlling factor in the chemical environment of the soil together with pH and Eh (redox potential).

Soils support ecosystem services

Ecosystem services are the benefits people obtain from ecosystems (Fisher and Turner, 2008). That means there is a focus on relations to society and therefore they also have an economic perspective (Fisher et al., 2009). The services are separated in four ways. The 'regulating services' cover climate relevant facts like carbon sequestration, biological pest control, purification of water and air, and the enabling of natural pollination. The 'provisioning services' take food, feed and raw materials into account. Furthermore, they support genetic resources, medicinal resources, production of bio-minerals as well as energy supply. 'Supporting



Fig. 6 Waste deposit on the way to being reclaimed.

services' account for nutrient cycling and the provision of habitats. Finally, 'cultural services' cover landscape aesthetics, cultural heritage, outdoor recreation and more (Naidoo and Fisher, 2020).

All the services mentioned above are possible only through the influence of soils. As a matter of urgency, we have the duty to manage our soils in such a way that guarantees the maintenance of these services for present and future generations.

Water, air and soil: The basic environmental components

In the pedosphere we observe water, air and solids. Solids can occur as minerals and organic materials. However, the pedosphere is also interconnected with the atmosphere and hydrosphere. The interrelation between soil and air is related with the exchange of dust as well as of gases, and is of utmost importance (Blume et al., 2016; Amelung et al., 2018). These fluxes between the spheres may have significant influences. Emission of methane from peat (Histosols) will change the habitat significantly. Also iron, which is released in a spring, changes the water quality significantly and may produce ochreous precipitations.

In national and international legislation and laws we find that these basic environmental components are taken in the same sequence. Water is considered to be more important than air. Air is considered more important than soil. The reason is, that water and air are generally considered as common goods, but soils are generally a private good. The potentials of soils need to be socialized adequately to be considered in the legislation that is meant to protect them.

History of soil ecology

The knowledge of soils dates back to the first agricultural societies. In Mesopotamia, Egypt, India and China such knowledge was already recorded 10,000 years ago (Blume, 2003). In Central Europe it started 7000 years ago during the Neolithic Age. Greek literature started to document such knowledge (Plato) in about 400 BCE. Hildegard von Bingen (1150–1160) (2000) in the 12th century started to record the existing knowledge, which was mainly on topsoil. All this knowledge was focused on the agricultural use of soils. Soil science as a natural science is about 200 years old only. Famous people such as J.W. von Goethe, J. von Liebig, and A. Thaer were involved in this development. The first text book on soil science dates back to Sprengel (1837). In the 19th century the development of knowledge and understanding of soils occurred in North America (Hilgard), Germany (Fallou) and Russia (Dokuchaev). All this progress was almost entirely related to agriculture, forestry and geology (Blume, 2003; Blume, 2012).

The observations of Charles Darwin on the nature and behavior of earthworms (Lumbricidae) from 1840 to 1880 was the first document of biological processes in soils (Darwin, 1881).

In the first half of the 20th century systematic research on soil genesis and soil fertility resulted in the progress of soil classification, site evaluation and soil mapping (e.g. Schlichting, 1986). The development of land evaluation methods marks the peak of the practical outcome of traditional soil science (Storie, 1978; Sys et al., 1991). The modern soil classifications Soil Taxonomy (2014) and World Reference Base, WRB (2015) still order soils mainly for agricultural and silvicultural purposes.

Soil ecology became evident only one generation after the Second World War (Ehrlich and Ehrlich, 1970). Now soil regulations on climate, ecosystem functions and relations to humans play an important role (Fisher et al., 2009). However, the traditional functions like production of food and energy crops are increasingly dependent on the progress of soil ecology (Millennium Ecosystem Assessment (MEA), 2005; Naidoo and Fisher, 2020).

Conclusions and outlook

Soils can provide the potentials to serve people and landscapes in the future, but only if care is taken to maintain them. All soils accomplish certain functions and thereby they serve their owners and users in ecosystems and landscapes.

In addition to their primary functions (producing wheat and timber, regulating the water cycle and load capacity) soils provide ecosystem services (regulating, provisioning, supporting and cultural services).

Soils will continue to be important for humans, all wildlife and the environment in the future. Therefore, soil protection must be focused upon to ensure adequate maintenance of our soils and the ecosystem services that they provide for life now and for future generations.

References

- Ameling W, Blume H-P, Fleige H, Horn R, Kandeler E, Kögel-Knabner I, Kretschmar R, Stahr K, and Wilke B-M (2018) *Scheffer/Schachtschabel Lehrbuch der Bodenkunde*, 17th edn. Heidelberg: Springer Spektrum.
- Beek KJ (1978) *Land Evaluation for Agricultural Development*. Wageningen, Netherlands: International Institute for Land Reclamation and Improvement.
- Blume HP (2003) *Die Wurzeln der Bodenkunde*. Ch. 1.3.1. In: Blume HP et al. (1996 ff) lit. cited.
- Blume HP (2012) *Die geobiowissenschaftliche Tradition*. Ch. 1.3.2.1. In: Blume HP et al. (1996 ff) lit. cited.
- Blume H-P, Brümmer GW, Fleige H, Horn R, Kandeler E, Kögel-Knabner I, Kretschmar R, Stahr K, and Wilke B-M (2016) *Scheffer/Schachtschabel Soil Science*, 1st edn. Berlin, Heidelberg: Springer.
- Coleman DC and Crosley DA Jr. (1996) *Fundamentals in Soil Ecology*. San Diego: Academic.
- Darwin C (1881) *The Formation of Vegetable Mould Through the Action of Worms With Observation of Their Habits*. London: Murray.
- Dunger W (2008) *Tiere im Boden*, 4th edn. Hohenwarsleben: Westarp.
- Ehrlich PR and Ehrlich A (1970) *Population, Resources, Environmental Issues in Human Ecology*. 383 pp San Francisco: W.H. Freeman.
- Fisher B and Turner RK (2008) Ecosystem services: Classification for valuation. *Biological Conservation* 141(5): 1167–1169.
- Fisher B, Turner RK, and Morling P (2009) Defining and classifying ecosystem services for decision making. *Ecological Economics* 61(2–3): 267–276.
- Hildegard von Bingen (1150–1160) (2000) *Liber subtilitatum diversarum naturam creaturam*. Handwritten manuscripts. Comp. Riethe and Vollmann.
- Jenny H (1941) *Factors of Soil Formation*. New York: McGraw-Hill.
- Jenny H (1980) *The Soil Resource—Origin and Behaviour*. *Ecological Studies*, vol. 37. New York: Springer.
- Kölling C, Göttlein A and Rothe A (2007) *Energieholz Nachhaltig Nutzen. Biomasseentzug und Nährstoffentzug*. LWT-Aktuell 61, pp. 32–36.
- Kubiena WL (1946) *Bestimmungsbuch und Systematik der Böden Europas*. Stuttgart: Enke.
- McRae SG and Burnham CP (1981) *Land Evaluation*. Oxford: Clarendon Press.
- Millennium Ecosystem Assessment (MEA) (2005) *Ecosystems and Human Well-Being: Synthesis*. 155 pp Washington: Island Press.
- Naidoo R and Fisher B (2020) Reset sustainable development goals for a pandemic world. *Nature* 583(7815): 198–201.
- Schlichting E (1986) *Einführung in die Bodenkunde*, 2nd edn. Hamburg: Parey.
- Sprengel C (1837) *Die Bodenkunde oder die Lehre vom Boden*. Leipzig: J. Müller.
- Stahr K, Kandeler E, Herrmann L, and Streck T (2020) *Bodenkunde und Standortlehre*, 4th edn. Stuttgart: UTB Eugen Ulmer 327.
- Storie RE (1978) *Storey Index Rating*. Special Publication, Division of Agricultural Sciences, University of California No 3203.
- Sys C, Van Ranst E, and Debaveye J (1991) *Land Evaluation. Part II*. Brussels, Belgium: General Administration for Development Cooperation.

Farming systems to conserve land for nature

Rattan Lal, CFAES Rattan Lal Center for Carbon Management and Sequestration, The Ohio State University, Columbus, OH, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of R. Lal, SHIFTING CULTIVATION, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 488-497, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00485-9>.

Introduction	12
Feeding the world by 2050	13
Soil restorative farming system	13
Strategies for sparing land	15
Nature-positive farming systems	15
Agro-ecology	16
Rights-of-soil	16
Food system transformation	17
Current and future aspirational land use	17
Farming systems for returning land to nature	17
References	18
Further reading	19

Abstract

To feed 8 billion (B) people in 2022, agriculture occupies 37.1% of Earth's land. Yet, 15% (1.2 B) of population is undernourished and 25% (2 B) is malnourished. However, 30–40% of food produced globally is wasted. To feed 9.8 B by 2050, some argue for additional land, water, chemicals, and other inputs. In contrast, the soil-centric approach of producing more from less can spare 50% of the land and 66% of the water consumed in agriculture for conservation and rewilding of nature. Translating science into action to spare land for nature requires cooperation between land managers, researchers, policy makers and the private sector.

Key points

- Enough food is produced to feed the current and future population.
- A large amount of food produced is wasted.
- Hunger and malnutrition are man-made tragedies and are not due to lack of production.
- The approach of having additional land and inputs (water, fertilizers, energy) to feed the future population is inherently flawed.
- Of the 37% of land area and 70% of total water used for agriculture, adoption of a soil-centric approach of producing more from less can conserve land and water for nature.
- Rather than producing more calories, agriculture must be nutrition sensitive.
- Prudent governance must involve implementation of policies which are pro-nature, pro-soil, pro-farmer and pro-agriculture.
- Transformation of the world food systems by adoption of innovative farming systems can reconcile the need to provide adequate, safe, and nutritious food with the necessity of restoring the environment.
- Restoration of degraded soils (2 B ha) and deserted ecosystems can also sequester carbon in the terrestrial biosphere and limit global warming to below 2 °C.
- Choice of a healthy diet, and adoption of innovative farming systems can spare land for nature.

Introduction

World population has increased from about 1 billion (B) at the time of Malthus in circa 1800 to 8 B in 2022 and is projected to be 9.7 B by 2050 and 10.4 B by 2100 (U.N., 2022). The Green Revolution, the greatest success story in agriculture, increased global cereal production from 0.88 billion ton (B t) in 1960 to 3 B t in 2020 (by a factor of 3.3) when the population increased from 3.1 B to 7.7 B (by a factor of 2.5). Contrary to Malthus' prediction, the per capita cereal production increased by 32% from 284 kg to 375 kg. Indeed, the rate of food production being more than that of population growth averted global mass starvation.

The Green Revolution miracle was created by the intensification of agriculture on 5.2 B ha of land (1.5 B ha crop land and 3.7 B ha grazing or range land) through increased use of water, fertilizers, pesticides, and other energy-based materials applied to input-responsive and dwarf varieties.

Between 1961 and 2020 the increase in inputs was 11.6 times for N fertilizer, 4.4 times for P fertilizer, 3.2 times for K fertilizer, 5.2 times for pesticides, and 2.4 times in cropland area equipped for irrigation which increased from 144 M ha to 350 M ha.

The Green Revolution, however, was cereal centric (rice (*Oryza sativa* L.), wheat (*Triticum aestivum* L.)), based on heavy inputs of chemicals and irrigation, and focused on the intensification of monoculture. Thus, it caused environmental degradation, land desertification, decline in soil quality, eutrophication of water (i.e., algal bloom, hypoxia of coastal ecosystems), air pollution, and the emission of greenhouse gases (GHGs) into the atmosphere. One-third of the Earth's soils are degraded, depleted, and polluted (FAO/ITPS, 2015) and 30–35% of all anthropogenic emissions are contributed by global food systems (IPCC, 2018). In other words, agro-ecosystems are perceived to be responsible for the widespread problem of degraded soils, polluted waters, contaminated air, dwindling biodiversity, and denuded landscapes.

The present food systems are consuming a vast quantity of natural resources. For example, 38% of the Earth's surface is used for agriculture, 75% of agricultural land (3.7 B ha) is used for raising animals, 70% of the global fresh water withdrawn is used for irrigation, and 30–35% of global GHGs are emitted by agriculture. Yet, 1 in 7 persons (1.2 B out of 8 B people in 2022) are undernourished and 1 in 4 (2 B out of 8 B) are malnourished (FAO et al., 2022). The prevalence of undernutrition was estimated at 678 million (M) in the pre-COVID era of 2019, 782 M in 2020 and 828 M in 2021 because of the pandemic-induced disruptions in food production and supply chains. The number increased to 1200 M in 2022 because of the war in Ukraine and civil strife in other regions of the world.

Thus, business as usual cannot continue. There is a strong need for a paradigm shift in global food systems through the adoption of regenerative agricultural systems. Innovative agricultural systems must reconcile the need to meet the demands of a growing and increasingly affluent human population with the absolute necessity of improving the environment. Therefore, the objective of this chapter is to deliberate innovative farming systems, which make agriculture a solution for restoring the environment by producing more from less.

Feeding the world by 2050

Global food grain production of ~3B t in 2022 is enough to feed the present human population (8 B) and future population (9.7 B) by 2050. Thus, there is no need to expand agriculture, bring new land under agro-ecosystems by deforestation, and increase the use of fertilizers and pesticides to produce more food. This approach is counter-productive and is inherently flawed; both under-nutrition and malnutrition are man-made tragedies. The prevalence of under-nutrition is caused by social, economic, and political factors rather than by any shortfall in agricultural production. Similarly, prevalence of malnutrition or hidden hunger is caused by poor diet and food grown on degraded or polluted soils (Lal, 2009; Ewing-Chow, 2020), irrigation with contaminated water (Linderhof et al., 2021), and grown on soil mulched by polyethylene plastic (Madrid et al., 2022). Irrigation water can be a source of food-borne pathogens (Steele and Odumeru, 2004) and polluted irrigation water is making people unwell. Food grown in soil and water polluted by arsenic (As) is a major health hazard globally (Kayode et al., 2021; Sandhi et al., 2022). In addition, the effects on under nutrition of deficiency in micronutrients in soil (e.g., Zn, Mn, Fe), such as the role of selenium content in wheat on human health, are widely recognized (Tamás et al., 2010).

There is the problem of food waste which may be as much as 30–50% globally (van den Bos Verma et al., 2020), which must be effectively addressed. Improving access to food by addressing poverty, inequality, civil strife, and war is critically important as has been demonstrated by the war in Ukraine in 2022 and civil strife in Horn of Africa since 1990s. Food insecurity and hunger in rich countries is closely associated with inequalities (Pollard and Booth, 2019). Increasing the use of pulses (legumes (Fabaceae)) and promoting a plant-based diet can have a positive effect on human health and that of the planet (Lal, 2017).

Therefore, the Green Revolution 2.0 must be soil centric and based on soil resilience, eco-centric and based on eco-efficiency, and knowledge centric based on scientific innovations.

The overall strategy is to increase agronomic productivity from existing land, restore deserted land and degraded soils, enhance biological N fixation by legumes, protect prime soils against urbanization and non-agricultural uses, and spare land for nature. The aim is to produce more food from less land, less water, fewer inputs of fertilizer and pesticides, less use of energy, and less emission of GHGs.

Soil restorative farming system

One-third of Earth's surface is already degraded (FAO/ITPS, 2015). Predominant processes of soil degradation are erosion by water and wind, salinization, physical degradation of soil structure and attendant soil compaction or densification, acidification, and elemental imbalance (Oldeman, 1994). Severe depletion of soil organic carbon (SOC) stocks, through widespread use of extractive farming systems by resource-poor and small land holders of developing countries and indiscriminate and excessive use of inputs by largescale commercial farmers in developed economies, is a major factor aggravating anthropogenic emissions (Lal, 2004). Thus, there is a critical and urgent need for the adoption of soil-restorative farming systems.

Farming system refers to a mix of farming enterprises (i.e., crop, livestock, aquaculture, agroforestry, and fruit crops) to which a farm family allocates its resources (soil, water, landscape, biodiversity) to manage the existing environment effectively to attain the family goals (Lal and Miller, 1990). In addition to the biophysical approach, farming systems have also been defined with the focus on social and economic considerations, and emerging environmental consensus such as climate change (Lal, 2023). Because

of the strong interconnectivity of the numerous factors involved, there is also an increasing focus on the nexus approach to identifying appropriate farming systems for the site-specific situation (Lal, 2013), and for addressing the Sustainable Development Goals (SDGs) of Agenda 2030 of the United Nations (U.N., 2015). The rationale for the identification of innovative farming systems (Fig. 1) includes environmental, social, economic, cultural, and others. Because it is highly interactive, the nexus approach is appropriate.

Both the severity and extent of soil degradation and its adverse effects on numerous ecosystem services must be addressed at local, national, and global scales through the choice of site-specific farming systems. Further, goals for achieving high agronomic production to meet the demands of growing population must be reconciled with the even more urgent and critical goal of restoring the environment while maintaining an increasing trend in achieving per capita productivity (e.g., 32% gain in per capita cereal production between 1960 and 2020). Lal (1994) outlined several indices of sustainability, some of which are briefly outlined in Table 1.

Indices outlined in Table 1 can be assessed at different spatial and temporal scales. Spatial scale can be farm household, state, national or global level. Temporal scale is important to recognize in relation to the climate factor (C_t) which appears in indices 4, 5, and 6; it is relevant both now and in the future more than ever before.

In the context of saving land for nature, the following land-based factors are relevant (Lal, 1994):

$$L = \frac{C + F}{C}, \quad (1)$$

where L is the land-use factor or land-use intensity, C is the cropping period, and F is fallow period. In the modern concept of cover cropping, F is synonymous with the duration of cover cropping.

$$LER = \sum_{i=1}^n \left(\frac{Y_i}{Y_m} \right), \quad (2)$$

where LER is the land equivalent ratio where two crops (cereal and legume, or annual and perennial as in agroforestry) are grown on the same land at the same time, Y_i and Y_m are yield of component crops and n is the number of crop species grown on the same land. In scenarios where crop species may vary widely in their maturity period, an index called area time equivalent ratio (ATER) is used.

$$ATER = \frac{1}{t} \left[\sum_{i=1}^n \left(\frac{d * Y_i}{Y_m} \right) \right], \quad (3)$$

where d is the growing period of crop m in days and t is the time period in days for which the land remains occupied, or it is the growth period of the longest duration crop in the complex farming system.

The overall goal is to use these and other emerging indices to optimize resource use, produce more from less, and save land for nature.

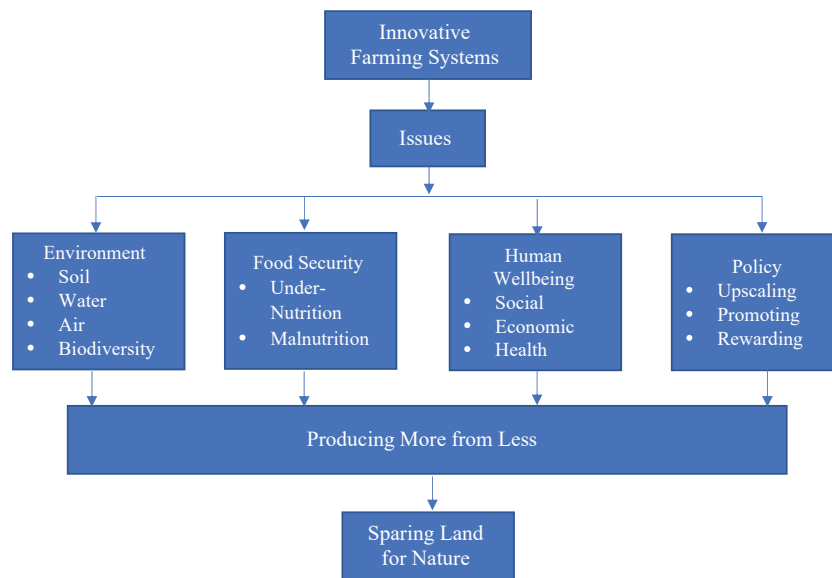


Fig. 1 Identification of innovative farming systems to make agriculture a solution to environmental issues and return our spare land for nature.

Table 1 Some indices of assessing sustainability of farming systems.

Index	Formula	Variables
1. Productivity (P)	$P = \frac{p}{R}$	P = productivity p = total production R = resource used (N, water, energy, SOC, etc.)
2. Total factor productivity (TFP)	$TFP = \frac{p}{\sum_{i=1}^n (R_i \times C_i)}$	p = total production R = resource used C = cost of the resource n = number of resources used in achieving total production
3. Coefficient of sustainability (C_s)	$C_s = f(O_i, O_d, O_m)_t$	C_s = coefficient of sustainability O_i = output per unit input that maximizes per capita productivity or profit O_d = output per unit decline in the most limiting or non-renewable resource O_m = minimum assured output t = time
4. Index of sustainability (I_s)	$I_s = f(P_i * S_i * W_i * C_i)_t$	I_s = index of sustainability S_i = alteration in soil properties W_i = change in water resources and quality C_i = modification in climactic factor t = time
5. Agricultural sustainability (A_s)	$A_s = d(P_i * S_p * W_t * C_t)dt$	A_s = agricultural sustainability P_t = productivity per unit input of the limited or non-renewable resource S_p = critical soil property (SOM stock, rooting depth) W_t = available water capacity C_t = climactic factor t = time
6. Sustainability coefficient (S_c)	$S_c = f(P_i * P_d * P_m)_t$ $S_c = d(P_i * W_t * C_t)dt$	P_t = productivity per unit input of the limited or non-renewable resource P_d = productivity per unit decline in soil property P_m = minimum assured productivity S_c = critical level of soil property W_t = soil water regime and quality C_t = climactic factor t = time

Adapted from Lal R and Miller FP (1990) Sustainable farming systems for the tropics. First Int. Symp/on Sustainable Natural Resource Management for Sustainable Agriculture: Issues, Perspectives and Prospects in Semi-Arid Tropics, Vols 1 and 2, Indian Soil Agrom. New Delhi, India. A69–A89; Lal R (1994) *Methods and Guidelines for Assessing Sustainable Use of Soil and Water Resources in the Tropics*. Washington, D.C.: USDA/SMSS Bull. 21, 78.

Strategies for sparing land

Rather than deplete and degrade, innovative farming systems must be those which enhance and restore soil functions. Among the critical soil properties are SOC content, plant available water capacity (PAWC), rooting depth, soil pH, cation exchange capacity (CEC), water infiltration capacity, water stable aggregation and other indices (Kladivko et al., 2014). Choice of soil properties may vary and depend upon ecoregion, land use, farming system and other biophysical, socio-economic, and cultural conditions. Similarly, knowledge of the critical limits of these properties for optimal soil functions (i.e., SOC content of 2% by weight in the root zone) must be known for site-specific conditions. Thus, farming systems that degrade soil, which expand agriculture and take land from nature, are characterized by 5 Ds: deplete, degrade, destroy, discard, and dominate. These farming systems must be replaced by those which return land to nature by restoring and enhancing soil functions, and are characterized by 5 Rs: reduce, reuse, recycle, regenerate, and restore. Such a transformational change, based on a soil-centric strategy to return land to nature by translation of science into action, must be supported by prudent governance and appropriate policies (Lal, 2018a,b). Because agriculture must be a solution to environmental issues and farmers or land managers are the biggest stakeholders, policies must be pro-nature, pro-farmer, and pro-agriculture.

Nature-positive farming systems

Since the World Food System Summit (U.N., 2021), there has been a growing interest in regenerative agriculture (RA) and agro-ecology (AE) (Lal, 2020b, 2021a,b, 2022a,b). Regenerative agriculture is an approach to farming that uses soil conservation and restoration as the entry point to regenerate or restore natural resources, to strengthen essential ecosystem services and to reduce and minimize the harm done. Put simply, RA is nature-positive agriculture, and site-specific farming systems are adopted to

achieve these goals (Lal, 2020b). With soil health as an entry point it suggests the use of a farming system that replaces what is removed, wise or prudent response to what is changed, predict what may happen from anthropogenic and natural perturbations, and restore SOC content and stock by creating a positive soil–ecosystem organic carbon budget. Adoption of this strategy implies creating a positive (to replace the existing negative) SOC budget. Farming systems and RA practices that can create a positive SOC budget include conservation agriculture (CA) based on mulch farming together with complex rotations and integration of crops with trees and livestock (Lal, 2015, 2020a). These and other RA practices are inspired by eco-innovation, powered by non-carbon energy, driven by a circular economy and green infrastructure, and supported by the recarbonization of the terrestrial biosphere (Lal et al., 2012) as the bedrock of sustainable farming systems.

Despite its appeal and benefits, RA is a relatively new concept and a debatable issue. Because there is no one-size-fits-all and a range of practices can be listed as RA, there is a need for its precise definition, the criteria and methods for evaluation of its impact, and specific benefits to advancing SDGs of the U.N. Furthermore, standardization of RA is needed for site-specific situations. Under its overall umbrella, what is needed is the identification of key practices which include: nature positive multifunctional landscapes, prudent governance of these landscapes through relevant policies, plans to promote land system trends toward afforestation, intensification and upscaling of climate-resilient practices, urban agriculture based on soil-less culture practiced in dense cities, and education of the general public and policy makers in making a cultural value shift in favor of nature restoration (soil, vegetation, wetlands, biodiversity) and rewilding of land spared. In other words, RA must support nature conservation, restoration of deserted ecosystems and degraded soils, eco-intensification of high-productivity agricultural lands, and rewilding of marginal agricultural landscapes.

Agro-ecology

Similar to RA, the U.N. Food System Summit (U.N., 2021) also promoted the idea of ‘agro ecology’ (AE), as another innovation in food systems for restoration of the environment (Lal, 2022a). Rooted in the basic principles of ecology, AE is a multi-disciplinary and an integrative approach aimed at enhancing environmental friendliness of food systems, while also improving human health by implementing the “One Health” concept (Howard, 1940). The modern version of the “One Health” concept states that “health of soil, plants, animals, people, ecoregions and planetary processes is one and indivisible” (Lal, 2020d). It is often argued that AE is not based on science because it does not promote biotechnology and related innovations in addressing global issues (Lal, 2022a). That being the case, AE is not widely supported by the private sector. Nonetheless, basic principles of AE comprise: (i) enhancing use efficiency and reducing inputs, (ii) using cost-effective inputs through substitutions under site-specific conditions, (iii) integrating biodiversity in agroecosystems, (iv) improving connectivity of the consumer with producers, (v) identifying and using equitable farming systems, (vi) increasing community involvement in agroecosystems, and (vii) improving food security in urban centers (Lal, 2020e).

Enhancing acceptability of the AE concept implies the increased role of science so that it is relevant to global issues such as climate change, and other foci of SDGs. Acceptability of AE can be improved by embracing multiple dimensions of sustainability including environmental, social, economic, cultural, and institutional (Lal et al., 2016). It is therefore prudent to integrate the basic principles of AE with scientific innovations to address effectively poverty, hunger and malnutrition, and food safety while restoring degraded soils and the environment.

Rights-of-soil

The finite but ecologically fragile soil resource of the world, source of numerous ecosystem services essential to human wellbeing and nature conservancy, must be protected and managed prudently. The precious soil resource is shrinking both through degradation of its quality and functionality but also with conversion to non-agricultural uses (e.g., urbanization, industrialization, infrastructure, recreation, mining). An estimated 170,000 species of soil organisms have been identified (Wall and Moore, 1999; Wall et al., 2015) and is considered a living entity. Thus, similar to any other living entity, soil also must have the right to be protected against misuse and mismanagement so that it can also flourish and thrive (Lal, 2019, 2022b). Ethically, rights-of-soil are not based solely on its economic benefits but for protection of the planetary processes against pervasive and persistent impact of the Anthropocene and drastic perturbances of Earth’s life support systems. Similar, to other ecosystems (e.g., lakes, rivers, mountains, rain forest and nature on the whole), soil must have rights against deliberate activities that pollute, erode, deplete, degrade or decertify the precious resource and jeopardize its capacity to generate ecosystem services essential to all terrestrial life. Just and prudent use of soil through innovative land use and farming systems is neither immoral nor unethical, and also not illegal. It is in this context that emerging concepts of RA, AE, and other innovative ideas which enhance soil formations are not only pertinent but also essential to both human and all terrestrial life.

The concept of rights-of-soil is also pertinent to the idea of “Planetary Boundaries” (Rockström et al., 2009). The goal is to define planetary boundaries in which humanity can operate safely. Transgressing these boundaries can have disastrous consequences for humanity and the planet. Nine planetary boundaries identified by Rockström and colleagues are: climate change (CO₂ concentration of 350 ppm in the atmosphere), ocean acidification (mean surface sea water saturation state with respect to aragonite $\geq 80\%$ of pre-industrial level), biogeochemical N cycle (limit agricultural and industrial fixation of N₂ to 35 Tg N yr⁻¹),

and P cycle (annual P inflow to oceans not to exceed 10 times the natural background weathering of P), global freshwater use, and <15% ice-free land under cropland, among others. Thus, the concept of “rights-of-soil” can be used to limit the expansion of water for irrigation. It is a pertinent tool to keep humanity within the “Planetary Boundaries.”

Food system transformation

Food systems comprise inter-linked and value-added activities including growing, transporting, processing, distributing, using, preparing, consuming, and disposing of waste products from agriculture, fisheries, and forestry. Because of the widespread problem of under-nutrition and malnutrition, together with severe issues of soil and environmental degradation, and rapid global warming, the current food systems are not sustainable. The [FAO \(2018\)](#) defined a sustainable food system “That delivers food security and nutrition for all in such a way that the economic, social, or environmental bases to generate food security and nutrition for future generations are not compromised.” In accord with this definition, present food systems have several limitations that include: (i) inability to provide adequate nutrition, and a healthy and safe diet to the entire global population, (ii) susceptibility of food producing ecosystems to degrade soils, pollute waters, aggravate global warming, reduce biodiversity, and denude landscapes, (iii) vulnerability of food systems to emission of GHGs and aggravation of anthropogenic global warming, and (iv) failure to bridge the inequality gap among ‘haves and have nots’.

Thus, there is a strong need to develop and implement nature-positive food systems that eliminate hunger and malnutrition, and restore the environment. Therefore, nature-positive farming systems must: strengthen ecosystem services, minimize disservices, promote emission-negative farming, reduce input of chemicals, use bio-stimulants and organic amendments, promote a bio circular economy, support the “One-health” concept, use digital innovations, and re-carbonize the terrestrial biosphere.

Therefore, transformation of present food systems must be through the adoption of innovations which: (i) restore soil and biophysical environments, (ii) improve social equity and address human dimensions, (iii) promote prudent governance, (iv) identify and implement site-specific game changing solutions, (v) create local and national science-policy interfaces for sustainable food and farming systems. An important and much needed transformation would be that which leads to the adoption of nutrition sensitive agriculture (NSA). Malnutrition or hidden hunger is caused by lack of protein and deficiency of essential micronutrients in the diet. Therefore, NSA must focus on improving protein and micro-nutrients in the diet and decreasing consumption of fat. Apparently, healthy soil is the basic pre-requisite of NSA, because a healthy diet must come from food grown on healthy soil.

Therefore, strategies to combat hidden hunger are the restoration of soil health, complex crop rotation and farming systems based on the integration of crops with trees and livestock, dietary diversification with focus on pulses (legumes) and plant-based diet, and behavioral change through education and communication. An important part of the education is to promote the thesis that (i) healthy diet comes only from plants or animals grown on a healthy soil, (ii) when diet is poor medicine is of little use, and (iii) when diet is correct, medicine is needed less (ayurveda). Thus, there is a close inter-connectivity between innovative farming systems and a safe nutritious healthy diet.

Current and future aspirational land use

Most estimates of agricultural land use between 2020 and 2100 indicate expansion of cropland and of the area equipped for irrigation ([Table 2](#)). However, humanity cannot afford to take more land from nature because it is neither in the interest of the people nor of the planet. The prudent strategy is to reduce the land area under agriculture, spare land and rewild it to strengthen ecosystem services and minimize disservices.

An aspirational timetable of returning land to nature by adopting innovative farming systems is shown in [Table 3](#). By adopting the concept of producing more from less combined with the choice of a healthy diet and reduced food waste, it is possible to reduce the area under cropland to about 750 Mha or less (500 Mha) and pastureland to 1500 Mha or less (1000 Mha). However, with the adoption of drip sub-irrigation, the land area equipped for irrigation can be expanded to 500 Mha by 2100, and yet the water consumed may be reduced to one third of the present use for irrigation.

Farming systems for returning land to nature

Global food systems are in need of radical transformation through the choice of a healthy diet, and eco-intensification of agro-ecosystems designed to produce more from less, and education of the general public to reduce food waste which is estimated at 30 to 40% for grains and more for fresh vegetables and fruits. Innovative farming systems, based on the integration of crops with trees and livestock, and which involve regenerative agriculture and options to restore the soil and ecosystems, must be promoted so that half of the land currently under agriculture can be spared for nature by 2100. Sparing land for nature, rewilding of the terrestrial biosphere that has been drastically disturbed by anthropogenic activities, necessitates the restoration of deserted land and degraded soils through prudent governance and policies which are pro-nature. Growth of vegetation on returned land will sequester atmospheric CO₂ and mitigate anthropogenic global warming, enhance quality and renewability of natural waters and strengthen biodiversity. This would be a ‘win-win’ option for humans and nature.

Table 2 Past, present and future agricultural land use for agriculture (Mha).

Year (AD)	Population (M)	Cropland (M ha)	Pastureland (M ha)	Irrigated Land (M ha)
1	188	130	110	7
1800	989	420	510	8
2000	6896	1381	3357	260
2014	7200	1417	3516	331
2017	7550	1396	3369	335
2020	7600	1530	3406	399
2030	8600	1660		
2050	9800	1680		431
2100	11,200	1640		

Adapted and modified from Lal R (2018a) Digging deeper: A wholistic perspective of factors affecting SOC sequestration. *Global Change Biology* 24(8) doi:<https://doi.org/10.1111/gcb.14054>; Cao B, Yu L, Li X, Chen M, Li X, Hao P, and Gong P (2021) A 1 km global cropland dataset from 10,000 BCE to 2100 CE. *Earth System Science Data* 13: 5403–5421; Schneider JM, Zabel F, and Mauser W (2022) Global inventory of suitable cultivable and available cropland under different scenarios and policies. *Scientific Data* 9: 537. doi:<https://doi.org/10.1038/s41597-022-01632-8>; FAO (2009) Agriculture towards 2050. High Level Expert Forum, 12–13 October 2009, Rome, Italy; Hurtt GC, Chini L, Sahajpal R, Froking S, Bodirsky BL, Calvin K et al. (2020) Harmonization of global land use change and management for the period 850–2100 (LUH2) for CMIP6. *Geoscientific Model Development* 13(11): 5425–5464; Ellis EC (2021) Land use and ecological change: A 12,000-year history. *Annual Review of Environment and Resources* 46: 1–33; Molotoks A, Stehfest E, Doelman J, Albanito F, Filton N, Dawson TP, and Smith P (2018) Global projections of future cropland expansion to 2050 and direct impacts on biodiversity and carbon storage. *Global Change Biology* 24(12): 5895–5908; Vale MM, Lima-Ribeiro MS, and Rocha TC (2021) Global land-use and land-cover data: Historical, current and future scenarios. *bioRxiv*, 2021-05.

Table 3 A proposed timetable for humanity to return land and water to nature.

Resource Use	Units	2020	2030	2050	2100
Cropland Area	M ha	1500	1400	1000	750
Grazing or Rangeland	M ha	3700	3500	2500	1500
Irrigated Land (Drip or Subsoil)	Km ³	390	400	450	500

Source: Adapted and modified from Lal R (2020b) Regenerative agriculture for food and climate. *Journal of Soil and Water Conservation*. doi:10.2489/jswc.2020.0620A; Lal R (2016) Feeding 11 billion on 0.5 billion hectare of cropland. *Food and Energy Security Journal* 5(4):239–251; Lal R (2023) Farming systems to return land for nature: It's all about soil health and re-carbonization of the terrestrial biosphere. *Farming Systems* 1(1): Open Access, Elsevier (In Press).

References

- Ewing-Chow D (2020) Earth's rapidly degrading soil is bad news for human health. *Food, Agriculture, Sustainability*.
- FAO (2018) *Sustainable Food Systems: Concept and Framework*. Rome, Italy: FAO. <https://www.fao.org/3/ca.2079en/CA2079EN.pdf>.
- FAO, IFAD, UNICEF, and WMO (2022) *The State of Food Security and Nutrition in the World*. Rome, Italy: FAO. <https://doi.org/10.4060/cc0639en>.
- FAO/ITPS (2015) Status of the world's soil resources. In: *Main Report. Food and Agriculture Organization of the U.N., and Intergovernmental Technical Panel on Soils, Rome, Italy*.
- IPCC (2018) Climate change and land. In: *Special Report. Summary for Policy Makers*. Geneva, Switzerland: WMO/UNEP.
- Kladivko EJ, Helmers MJ, Abendroth LJ, Herzmann D, Lal R, Castellano MJ, and Villamil MB (2014) Standardized research protocols enable transdisciplinary research of climate variation impacts in corn production systems. *Journal of Soil and Water Conservation* 69(6): 532–542.
- Lal R (1994) *Methods and Guidelines for Assessing Sustainable Use of Soil and Water Resources in the Tropics*. Washington, D.C.: USDA/SMSS Bull. 2178.
- Lal R (2004) Soil carbon sequestration impacts on global climate change and food security. *Science* 304: 1623–1627.
- Lal R (2009) Soil degradation as a reason for inadequate human nutrition. *Food Security* 1: 45–57.
- Lal R (2013) Climate-strategic agriculture and the water-soil-waste nexus. *Journal of Plant Nutrition and Soil Science* 174(4): 479–493.
- Lal R (2015) A system approach to conservation agriculture. *Journal of Soil and Water Conservation* 70(4): 82A–88A.
- Lal R (2017) Improving soil health and human protein nutrition by pulses-based cropping systems. *Advances in Agronomy* 145: 167–204.
- Lal R (2018a) Digging deeper: A wholistic perspective of factors affecting SOC sequestration. *Global Change Biology* 24(8). <https://doi.org/10.1111/gcb.14054>.
- Lal R (2018b) Saving global land resources by enhancing eco-efficiency of agroecosystems. *Journal of Soil and Water Conservation* 73(4): 100A–106A.
- Lal R (2019) Rights-of-soil. *Journal of Soil and Water Conservation* 74(4): 81A–86A.
- Lal R (2020a) Integrating animal husbandry with crops and trees. *Frontiers in Sustainable Food Systems* 4: 113. <https://www.frontiersin.org/article/10.3389/fsufs.2020.00113>.
- Lal R (2020b) Regenerative agriculture for food and climate. *Journal of Soil and Water Conservation*. <https://doi.org/10.2489/jswc.2020.0620A>.
- Lal R (ed.) (2020d) *The Soil-Human Health Nexus*, p. 335. Boca Raton, FL: CRC/Taylor and Francis.
- Lal R (2020e) Home gardening and urban agriculture for advancing food and nutritional security in response to the COVID-19 pandemic. *Food Security*. <https://doi.org/10.1007/s12571-020-01058-3>.
- Lal R (2021a) Feeding the world and returning half of the agricultural land back to nature. *Journal of Soil and Water Conservation* 76(4): 75A–78A.
- Lal R (2021b) Negative emission farming. *Journal of Soil and Water Conservation* 76(3): 61A–64A. (May 1) <http://www.jswnonline.org/content/76/3/61A.abstract>.
- Lal R (2022a) Agroecology. *Rural* 21. 18 Aug. 2022 <https://www.rural21.com/english/search/detail/article/many-ways-one-goal-achieving-sustainable-agroecosystems.html>.
- Lal R (2022b) Nature-based solutions of soil management and agriculture. *Journal of Soil and Water Conservation*. <https://doi.org/10.2489/jswc.2022.0204A>. Soil and Water Conservation Society, 1 Mar. 2022.

- Lal R (2023) Farming systems to return land for nature: It's all about soil health and re-carbonization of the terrestrial biosphere. *Farming Systems* 1(1). Open Access, Elsevier (In Press).
- Lal R and Miller FP (1990) Sustainable farming systems for the tropics. In: *First Int. Symp/on Sustainable Natural Resource Management for Sustainable Agriculture: Issues, Perspectives and Prospects in Semi-Arid Tropics, Vols 1 and 2, Indian Soil Agrom. New Delhi, India* A69–A89.
- Lal R, Lorenz K, Huttel RRJ, Schneider BU, and von Braun J (eds.) (2012) *Recarbonization of the Biosphere*, p. 545. Holland: Springer Verlaag.
- Lal R, Kraybill D, Hansen DO, Singh BR, Mosogoya T, and Eik LO (2016) *Climate Change and Multi-Dimensional Sustainability in African Agriculture*. Columbus: Springer.
- Oldeman LR (1994) The global extent of soil degradation. In: Greenland DJ and Szabolcs I (eds.) *Soil Resilience and Sustainable Land Use*, pp. 99–118. Wallingford, U.K.: CAB International.
- Rockström J, Steffen W, Noone K, Persson Å, Chapin FS III, Lambin E, and Foley J (2009) Planetary boundaries: Exploring the safe operating space for humanity. *Ecology and Society* 14(2).
- Sandhi A, Yu C, Rahman MM, and Amin MN (2022) Arsenic in the water and agricultural crop production system: Bangladesh perspectives. *Environmental Science and Pollution Research* 29(34): 51354–51366.
- U.N (2015) *Sustainable Development Goals*. New York: Department of Economic and Social Affairs, Sustainable Development.
- U.N (2021) *Food Systems Summit*. . 23rd September 2021, United Nations, New York, USA.
- U.N (2022) World population prospects 2022. In: *Summary of Results, United Nations Department of Economic Affairs, UN DESA/POP 2022/TR No 3*.
- van den Bos Verma M, de Vreede L, Achterbosch T, and Rutten MM (2020) Consumers discard a lot more food than widely believed: Estimates of global food waste using an energy gap approach and affluence elasticity of food waste. *PLoS One* 15(2), e0228369.
- Wall DH, Nielsen UN, and Six J (2015) Soil biodiversity and human health. *Nature* 528(7580): 69–76.

Further reading

- Cao B, Yu L, Li X, Chen M, Li X, Hao P, and Gong P (2021) A 1 km global cropland dataset from 10,000 BCE to 2100 CE. *Earth System Science Data* 13: 5403–5421.
- Ellis EC (2021) Land use and ecological change: A 12,000-year history. *Annual Review of Environment and Resources* 46: 1–33.
- Howard SA (1940) *An Agricultural Testament*. London, U.K.: Oxford University Press, Inc.
- Hurt GC, Chini L, Sahajpal R, Frolking S, Bodirosky BL, Calvin K, and Zhang X (2020) Harmonization of global land use change and management for the period 850–2100 (LUH2) for CMIP6. *Geoscientific Model Development* 13(11): 5425–5464.
- Kayode OT, Aizebeokhai AP, and Odukoya AM (2021) Arsenic in agricultural soils and implications for sustainable agriculture. *IOP Conference Series: Earth and Environmental Science* 655(1), 012081.
- Linderhof V, De Lange T, and Reinhard S (2021) The dilemmas of water quality and food security interactions in low-and middle-income countries. *Frontiers in Water* 3: 736760.
- Madrid B, Wortman S, Hayes DG, DeBruyn JM, Miles C, Flury M, and DeVetter LW (2022) End-of-life management options for agricultural mulch films in the United States—A review. *Frontiers in Sustainable Food Systems* 282.
- Pollard CM and Booth S (2019) Food insecurity and hunger in rich countries—It is time for action against inequality. *International Journal of Environmental Research and Public Health* 16(10): 1804.
- Steele M and Odumeru J (2004) Irrigation water as source of foodborne pathogens on fruit and vegetables. *Journal of Food Protection* 67(12): 2839–2849.
- Tamás M, Mandoki ZS, and Csapó J (2010) The role of selenium content of wheat in the human nutrition. A literature review. *Acta Universitatis Sapientiae, Alimentaria* 3: 5–34.
- Wall DH and Moore JC (1999) Interactions underground: Soil biodiversity, mutualism, and ecosystem processes. *BioScience* 49(2): 109–117.

Geoarchaeology and soils

Carlos E Cordova, Department of Geography, Oklahoma State University, Stillwater, OK, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	20
Archaeology and soil science through geoarchaeology	20
How to read soil profiles in geoarchaeology?	21
Methodological approaches	22
Soil formation processes in site taphonomy	22
Soil as chronological control: Soil genesis, pedo-stratigraphic sequences and paleosols	23
Soils in landscape evolution and paleoenvironmental reconstructions	24
Soils as an agricultural resource: Farmed soils and anthrosols	24
Scales of research and techniques	26
Final remarks	27
References	27

Abstract

Soil studies in archaeology are an essential component of archaeological research during survey and at excavation of sites, and for the interpretation of data. The application of methods and techniques developed in soil science have been adapted to solve archaeological problems. Geoarchaeology incorporates the methods and techniques of all geosciences in the field of archaeological research. Soil studies target several foci, including site formation and abandonment (site taphonomy), the reconstruction of ancient landscapes and environments, soils transformed by humans, including those transformed by ancient agricultural societies. This diversity of research encompasses studies at several scales, from those incorporating archaeological territories and landscapes to those at the microscopic level.

Key points

- Soils are an important part of archaeological research because of their connection to ancient environments and societies.
- Geoarchaeology acts as the connector between soil science and other earth science disciplines to archaeology.
- The effective application of soil science to archaeological research requires understanding of archaeological methods and the dynamics of ancient societies.
- Modern technology allows soils to be studied in archaeological contexts from microscopic to regional scales.

Introduction

Archaeology and soil science through geoarchaeology

Archaeology is the study of past societies through the analysis of physical remains left by humans. Such remains are often embedded in soils and sediments, for which soil science is an essential part in their history. In other cases, soils are considered important in archaeological research because they are essential resources for human societies, especially for those dedicated to farming. Furthermore, soils provide information about the environment and landscapes that humans inhabited in the past, which adds relevant information to the study of past societies.

Although the relation between soils and archaeology has existed for a long time, nowadays most soil research in archaeological projects develops under the umbrella of 'geoarchaeology'. Although geoarchaeology was originally defined as the application of earth science methods to archaeological research (Butzer, 1982), it has now developed into a multidisciplinary and interdisciplinary field that brings together specialists from the geosciences (e.g., geology, geography, geomorphology, and soil science) and archaeologists (Cordova, 2018). Nonetheless, the focus of geoarchaeology as a field linked to archaeological research implies that its specialists should be conversant in archaeological methods and theory.

In other words, a pedologist with no idea of archaeological matters might not be familiar with the problems archaeologists are trying to solve. It is one consideration in soil science to identify and interpret soil horizons, but another to establish a connection between the artifacts in the soil and the environmental context of the society that made them. Thus, to exemplify the proper involvement of soil research in archaeology, the following sections discuss methodological and practical cases of the involvement of soil specialists in geoarchaeological research.

How to read soil profiles in geoarchaeology?

Geoscience specialists often view stratigraphic profiles from the perspective of their own discipline. In other words, a geologist, soil scientist, and an archaeologist interpret the same stratigraphic section differently (Goldberg and MacPhail, 2006). However, under the lens of geoarchaeology those different views of the section can be brought together in an integrated view that is useful for reconstructing the evolution of a landscape associated with a particular or various ancient societies. To illustrate this point, the diagrams in Fig. 1 provide an example of a soil sequence at an archaeological site in central Mexico.

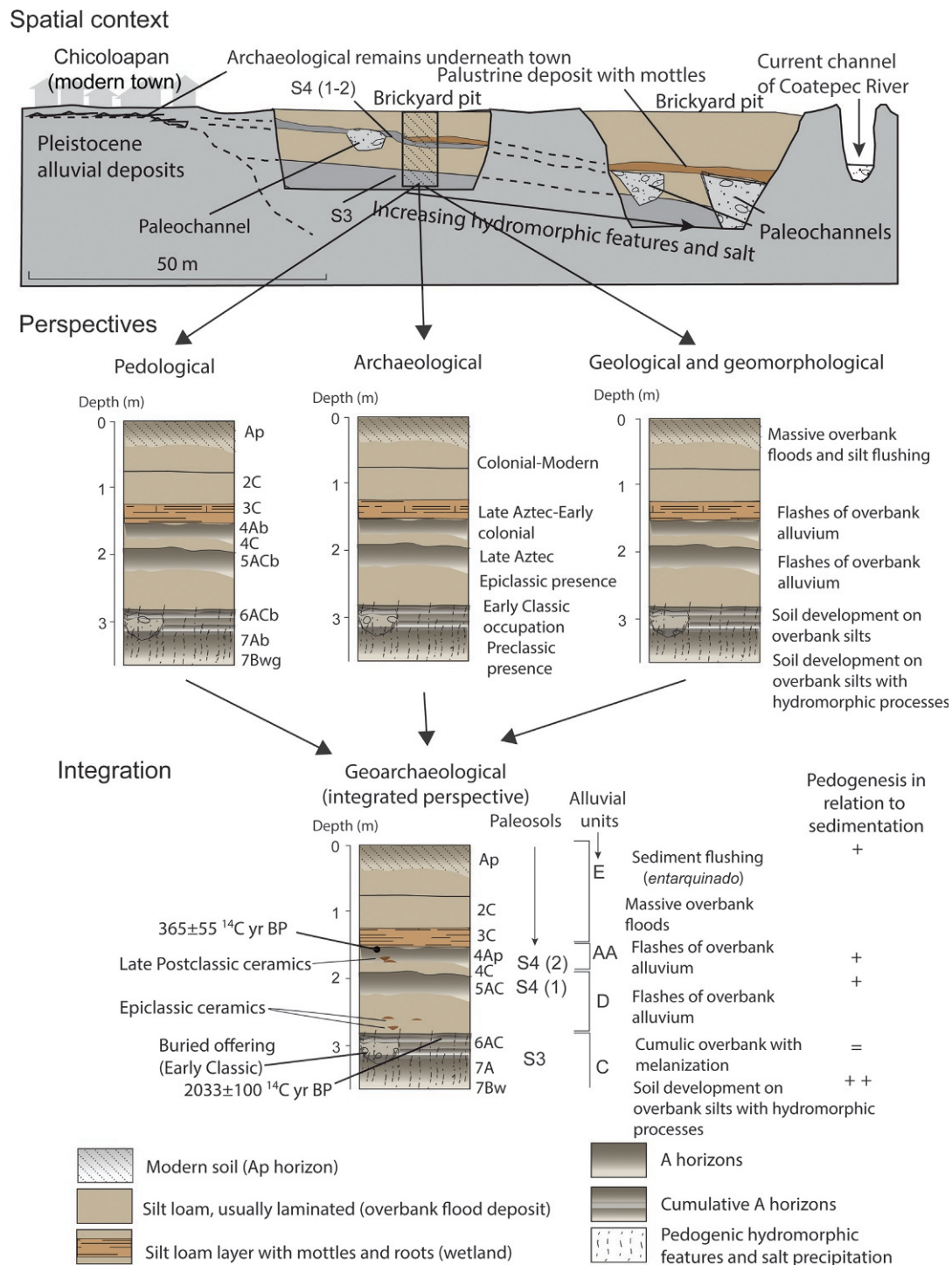


Fig. 1 View of a soil by an archaeologist, a pedologist, a geologist, and a ge archaeologist. Section extracted and modified from Cordova (2017). The locality represents a section exposed in a brickyard pit next to various archaeological sites of different ages.

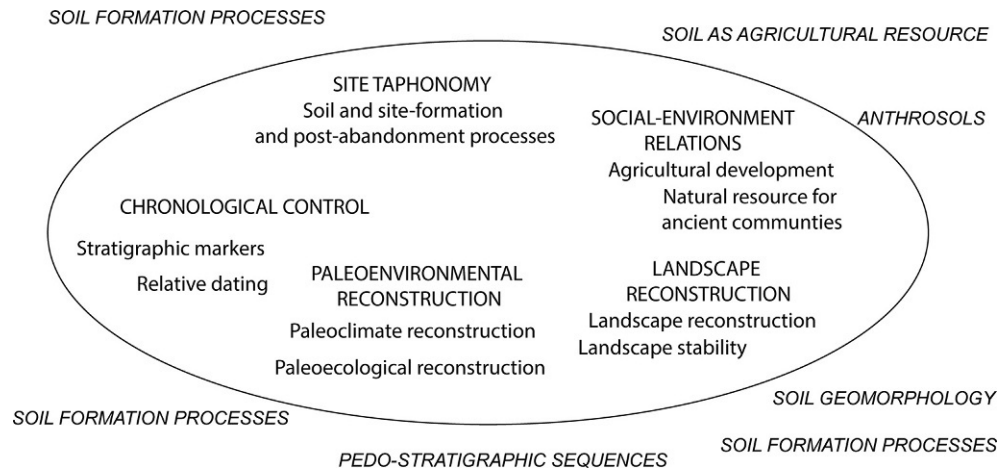


Fig. 2 Different soil science methodological approaches and geoarchaeological research areas in the application of soil science used for archaeological research.

The first aspect to notice is the spatial context of the profile (top illustration in Fig. 1), which shows an exposed wall inside a brickyard pit near the town of Chicoloapan, Mexico. Its geomorphic context indicates that it is in a former floodplain of the Coatepec River; one can see the modern channel to the left of the diagram. Second, comes the archaeological context. To the left there is a promontory where the modern town of Chicoloapan is located. However, the town sits on a Late Aztec site (c. 1325–1521 CE), and in an area with substantial occupation in the Classic (c. 200–550 CE) and Epiclassic (550–900 CE) periods.

With the spatial and archaeological context in mind, one can proceed to interpret the profile under study. However, the interpretation varies depending on which specialist looks at the profile. Thus, a soil scientist, an archaeologist, and a geologist or geomorphologist would interpret it based on their specialties (middle of Fig. 2). Therefore, it is necessary to combine these three views (bottom of Fig. 1). Thus, the lesson learned here is that the interpretation of soil development is crucial for reconstructing ancient landscapes, but also it must be combined with the sedimentological and archaeological layers, while the geomorphic context (floodplain environment) and local spatial and archaeological contexts are considered. The resulting combined interpretation will help in the understanding of landscape change and use of the soil by the ancient inhabitants.

Methodological approaches

Soil research in geoarchaeology combines methods and approaches developed in pedology as well as those of other geosciences such as geomorphology, sedimentology, stratigraphy, geophysics, geochemistry, and petrology (Goldberg and MacPhail, 2006; Rapp Jr. and Hill, 2006). The methodological approaches involving soils, however, are diverse and depend on the focus of the archaeological research. To represent this relationship better, the diagram in Fig. 2 shows the foci of research inside an ellipse and the methodological approaches for each focus around it. In the following subsections the discussion centers on the archaeological research component and how soil science approaches participate in this using examples.

Soil formation processes in site taphonomy

In archaeology a ‘site’ (i.e., archaeological site) is “any place where there are physical remains of past human activities” (SAA, 2022). There are, however, various types of archaeological sites, each of which has its own association with the environment, and consequently with soils. Prehistoric sites often lack written records and range from merely camps where people settled temporarily to villages and even cities. In some instances, even a rock wall with rock paintings or petroglyphs is considered an archaeological site. In contrast, historic archaeological sites often have written records associated with them, and they can range from small units like a house or wall to complex settlements and built structures. This makes, for example, an abandoned factory anywhere in the world as much an archaeological site as an ancient Greek temple.

The association with soils varies depending on the site. For example, a prehistoric site would have more association with soil genesis than a site with structures of more recent age. A prehistoric site may comprise many thousands of years of accumulation such that it may be embedded in the soil formation itself. From a different perspective, the locality of an enclosure where cattle were kept underwent changes due to the incidence of animal feces and trampling, thus transforming the soil from the original state which may be found outside the enclosure. It is here where soil taphonomy enters with a soil methodological approach.

Site taphonomy refers to the study of transformation of a locality from its original condition through the many additions and destructions that occurred during its occupation, to the transformation after the site was abandoned. In the case of its transformation by humans, it includes materials added to the soil intentionally, including soil brought in from elsewhere, as well as the addition of inorganic and organic materials (e.g., sand, lime, and natural fertilizers) that change the physical and chemical properties of the soil.

Linked to soil taphonomy is the study of preservation of archaeological materials under different soil conditions (e.g., pH, carbonates, sand, expansive clays, organic content, etc.) (Goldberg and MacPhail, 2006). The techniques used in soil science that are applied to site taphonomy include soil micromorphology and many others such as X-ray diffraction used to determine the chemical and physical properties of soils (see Goldberg and MacPhail, 2006).

Soil as chronological control: Soil genesis, pedo-stratigraphic sequences and paleosols

Many times, inside and outside sites soil genesis and development becomes important as means of estimating time. In archaeology there are two main methods to determine age, the absolute and relative methods. The absolute method uses techniques to establish a numerical time frame, as in the case of radiocarbon dating. Relative methods use techniques to place soils within a reference time frame. However, by combining the two methodologies it is possible to provide a better time frame for the evolution of a landscape with more resolution (see next subsection).

To understand the role of soils as a means of chronological control one must consider that some soil orders or types have more developed structure (horizons), indicating that they have formed during long periods of time, while others are poorly developed, suggesting that they are young soils. This difference, however, may also be related to the prevalent climatic conditions (see next subsection).

Because climatic conditions are strongly linked to soil development, sequences of soils (i.e., pedo-stratigraphic sequences) provide a good way to correlate environmental change with known climatic conditions through time. This approach is more common at prehistoric archaeological sites belonging to the Lithic periods (e.g., Paleolithic, Mesolithic, and Neolithic), especially in open environments (e.g., outside caves and rock shelters).

The interpretive basis of a pedo-stratigraphic sequence is the alternation of sediments and soil horizons. The latter are often referred to as paleosols, which are soil horizons that are no longer soils because they are buried by younger sediments, such as alluvium (deposited by water), eolian (deposited by wind), or volcanic (buried by ashfall, pyroclastic flows) or simply buried under a lava flow. The importance of paleosols in archaeology is that in some cases people lived on these soils or farmed them. Thus, the artifacts, if diagnostic, provide relative ages to the paleosols. Conversely, radiocarbon dating from humic acids of the organic horizons of such paleosols can provide a numerical date for the artifacts.

Fig. 3 shows a sequence of soil horizons developed on redeposited loess and alluvium exposed on a terrace at the confluence of the Kama and Volga Rivers in Russia. In this case, the different soil horizons in the sequence attest to relatively warmer periods

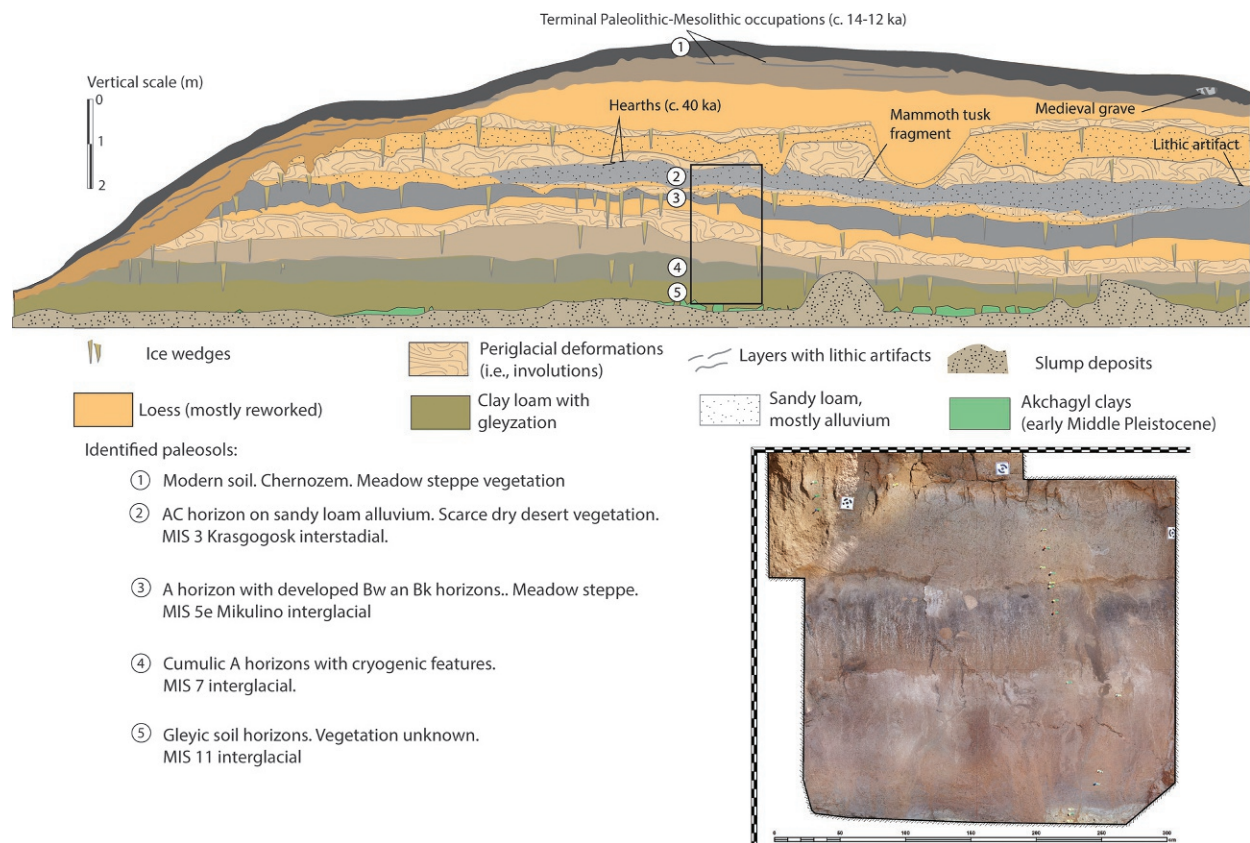


Fig. 3 Pedo-stratigraphic sequence at Beganchik on a terrace at the confluence of the Kama and Volga rivers, Russia (Cordova et al., 2021). The picture shows the details of soil horizons indicated by the square on the main diagram.

(interstadials and interglacials), while loess accumulation and deformation by ice correspond colder periods (stadials and glacials). In the sequence one can see the modern soil (1) which is the Chernozem at the top of the sequence, which caps a sediment with Mesolithic occupations during a period about 14,000–12,000 years before the present. Farther below the sequence, an alluvial soil (2) includes hearths, artifacts, and fragments of bones, clearly indicating a prehistoric site. The idea of soil formation at this latitude (approximately 55° N) suggests that soil development and human occupation needed relatively warm temperatures. The relative position of the soil and fauna associated with it and its position below a layer of loess suggested that it belonged to the MIS 3 interstadial (approximately 60,000 to 27,000 years before the present). Radiocarbon dating of one of the hearths provided two ages around 48,000 years before the present, which places the soil exactly in the Krasnogorsk interstadial (or the Mooershoofd interstadial of Western Europe), a relatively warm period across most of Europe at the time (Cordova et al., 2021). In this part of Russia, the warm conditions permitted humans to pursue animals farther north than normal. This provides a clue that similar soil horizons in the region may be associated with this warm phase in the history of the last ice age.

The sequence has another paleosol that stands out by its thickness and darkness (3), which by correlation with other dated horizons across the Russian Plain corresponds to the so-called Mikulino paleosol, which relates to the penultimate interglacial, or MIS 5e (128,000–116,000 years before present). This interglacial had warm temperatures similar to the current ones, leading to the formation of soil like that of the modern Chernozems in the Russian Plains. Although there are remains of large fauna associated with this layer, there is no evidence of human occupation. Farther down towards the bottom of the sequence, there is another paleosol with gleyic characteristics (4), which may also correspond to an earlier warm period, but only absolute dating would be able to reveal its approximate age.

Soils in landscape evolution and paleoenvironmental reconstructions

The reconstruction of past landscapes and environments involves the use of soils as a measure of age and evolution of the landscape (the surface) and the environment (surrounding conditions) through time. These are aspects of geoarchaeological research that are crucial for understanding the survival and development of human societies in the past. This approach relates to geomorphology, the study of earth surface processes and landforms. The role of soil science in geomorphology is in terms of it both as a chronological method (previous subsection) and as an indicator of landscape stability (Holliday, 2004).

Landscape stability means no erosion and no accumulation, which allows plants to develop, eventually leading to pedogenesis. In contrast, periods of erosion and accumulation hinder soil development. Furthermore, stability of the landscape results under relatively wet, mild, and stable climatic conditions. Often these stable periods foster the presence of fauna and flora, thus becoming attractive to humans (Mandel, 2008). Accordingly, if conditions are unstable, sedimentation and erosion may impede the formation of soils. Put in the context of a floodplain environment, landscape instability results in frequent flooding, at times catastrophic, that may bury a soil or impede soil formation. In contrast, paucity of floods due to stable conditions, may allow soils to develop. The relation with archaeology is that stable landscapes are more likely to sustain human populations. However, if conditions change and sedimentation (flooding) occurs, the surface is abandoned, leaving behind an archaeological site.

The best example of a stable landscape appears in the paleosol S3 (see Fig. 1) in the sequence, which has archaeological material and even burials, suggesting permanent occupation. However, the site was abandoned as sedimentation buried the soil. Such an unstable period coincided with climatic instability and consequently social instability that occurred in the last three centuries of the first millennium AD in central Mexico (Cordova, 2017). The unstable conditions drove settlers elsewhere. Nonetheless, stable conditions returned a few centuries later.

Landscape instability is not unique to floodplain environments. Upland areas can also experience unstable periods due to surface erosion. However, like the floodplain counterpart, lack of erosion may lead to soil formation and the attraction of settlers, particularly farmers. The example shown in Fig. 4 provides an idea of widespread erosion of an upland landscape in the Texcoco region of central Mexico, where the remains of soil are found only on top of pedestal rock features (numbers 1–4 on Fig. 4a–b), indicating that at some point this was a stable landscape. Interestingly, the soils developed on the pedestals (Fig. 4c) suggest that soil formation was followed by farming, as the top horizon (Ap) contains evidence of the addition of materials for improving fertility (Cordova and Parsons, 1997). Interestingly, the stable period is not so far back in time; the archaeological materials associated with the farmed horizon correspond mainly to the Late Aztec period (AD 1325–1521), or even later. Further research has revealed that the widespread erosion occurred during a climatic and social crisis that took place in the 16th and 17th centuries (Cordova, 2017; Cordova and Parsons, 1997).

Studies of soil and landscape stability may encompass periods in the order of millennia where stability does not necessarily relate to farming or demographic pressures, but periods of climatic instability. This is the case for numerous examples across the Great Plains of North America in prehistoric times (Mandel, 2008; Holliday, 2004).

Soils as an agricultural resource: Farmed soils and anthrosols

Farmed soils, and particularly those that are completely artificial are related to soil taphonomy, and landscape and environmental stability. A farmed soil is often a natural soil horizon that has been plowed, and which often has the designation Ap. The farmed horizon may have physical and chemical characteristics that suggest the addition of manure and perhaps soil from elsewhere, but in general the altered horizon would have continuity with the original subjacent soil (similar to the example in Fig. 4c). There are also cases of farmed soils with the so called 'agric' horizon, or the horizon below the plow horizon that has accumulated clay, silt and



Fig. 4 Eroded landscapes with pedestal rocks capped by a relict soil which here is used to reconstruct the pre-erosional topography of the site. The soil on the pedestal has signs of being farmed and contains archaeological material bearing a date of c. 16th century as the maximum age of the erosion. (A) Simplified topographic profile of the study area; numbers are used as reference points. (B) Photograph of the study area with reference points; (C) Soil profile from the pedestal marked as point 2; (D) Archaeological material recovered from the bottom of the Ap horizon.

humus leached from above, often suggesting the implementation of irrigation (Bockheim, 2014). In this, and many other cases, soil horizons present evidence of agricultural management in the past, an aspect that is of great interest in archaeological research (Homburg and Sandor, 2011).

In some cases, distinguishing a farmed soil from non-farmed soils may not be easy. There are cases, for example, in which soil horizons bear signatures of farming. One case is the use of stable carbon isotopes in soils under the dense tropical vegetation of the Mayan region (Fig. 5). During the Classic Maya period (roughly AD 250–900) population in the lowlands of northern Guatemala and Belize, and Mexico’s Yucatan Peninsula grew to the point that broad areas of tropical rainforest were removed, and their soils farmed by the slash-and-burn technique. In some cases, the evidence of this process is clear either because of the ash layers or artificial soils (e.g., anthrosols), whereas in others it is more difficult to observe in the field. This is the case of the soil profiles in Fig. 5 in which the Ab horizons present a shift in $\delta^{13}\text{C}$ towards the C_4 plants, often grasses and maize, *Zea mays* L. (C_4 plant) (Beach et al., 2018). Non-farmed soils retain the low signature of C_3 plants. This is the case for those horizons above and below Ab, which sustained trees (typical C_3 plants). Thus, the Ab horizon with the C_4 plant signature is an isotopic marker of the Classic period.

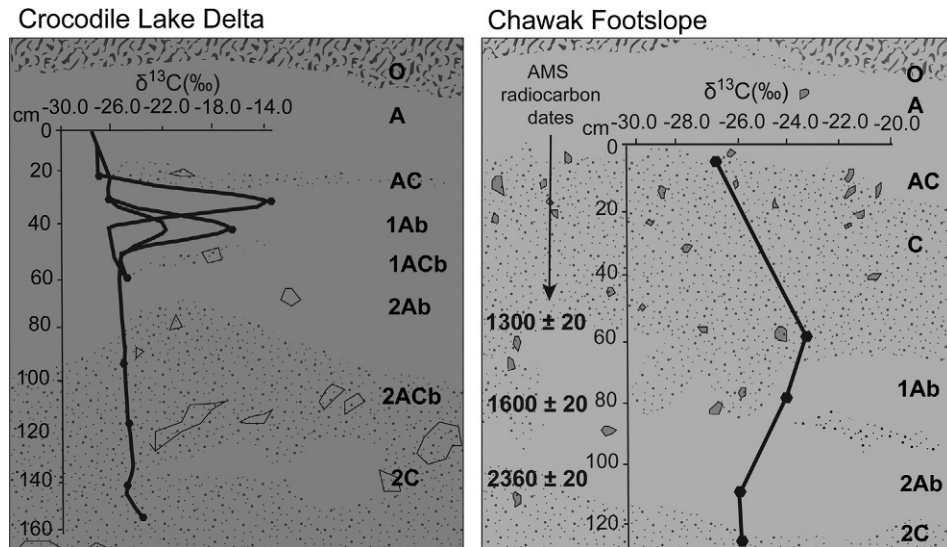


Fig. 5 Selection of soil profiles in the Maya lowlands, displaying $\delta^{13}\text{C}$ through their horizons. Adapted from Beach T, Ulmer A, Cook D, Brennan ML, Luzzadder-Beach S, Doyle C, Eshleman S, Krause S, Cortes-Rincon M and Terry R (2018) Geoarchaeology and tropical forest soil catenas of northwestern Belize. *Quaternary International* 463: 198–217.

Farmed soils can also be entirely artificial soils, in which case they are referred to as ‘anthrosols’ because their formation depends entirely on human activities. In this case an anthrosol is a result of the deliberate accumulation of materials, or in some cases of the indirect accumulation of materials, that create a surface where plants can grow. The study of anthrosols in archaeology varies around the world, depending on the geomorphic, biogeographic, and climatic context. One of the most studied soils in recent times is the so called ‘dark’ or ‘black earths’. The term may be used indiscriminately for examples of dark soils or those of anthropic origin in different parts of the world. Dark earths (also known as ‘terra preta’) are soils developed with the accumulation of burned and slashed materials by ancient farmers in the Amazon region (Kern et al., 2004). They are visually distinctive from the typical Oxisols developed under a wet tropical climate. Similar soils have also been reported in the Mayan region, often in contexts where terracing involved the creation of soils, and where the dark soils are distinctive from the pale or light color of limestone derived soils (Beach et al., 2018). The European black earths, on the other hand are soils that are often typical of archaeological contexts in urban areas. Their origin is often debated, although in many cases they have resulted from the ruralization of urban or peri-urban areas in the Middle Ages (Nicosia and Devos, 2014). Their formation is often associated with the accumulation of manure or with artificially induced pastures, although this varies in many places suggesting that these soils may also have had multiple origins and formed under different socioeconomic circumstances (Nicosia and Devos, 2014).

Scales of research and techniques

Soils in geoarchaeology can be studied at different spatial scales. Thus, one can study soils across large areas especially when soil characteristics vary across the landscape, while at the same time looking at the properties of those soils inside and outside sites in more detail (Table 1). The scale used in soil research in geoarchaeology also depends on the stage of the archaeological process under study (e.g., survey, excavation, and analysis). In general, soil studies in survey will require involvement at macro- and meso-scales, whereas those involved in excavation are likely to focus on micro spatial scales. The methods and techniques developed in soil science at each scale are used in archaeological research at the different phases, namely survey, excavation, and laboratory research (right column in Table 1).

Table 1 Different scales of research and their associated methodological approaches and application in the archaeological process.

Scale	Methodological approach	Stage in the archaeological research process
Macroscale	Mapping, remote sensing, GIS, stratigraphic correlation, spatial analysis	Survey, analysis
Mesoscale	Pedostratigraphic profiles, geochemistry, macrofossils.	Survey, excavation, analysis
Microscale	Micromorphology, trace elements, microfossils, geochemistry	Excavation, analysis

At the macroscale, archaeological research involves mainly survey, though in geoarchaeology there might be research associated with the search for evidence of environmental change (including climatic change) and landscape change. Thus, at this stage soil mapping becomes an important part of the research, which is often aided by geographic information systems (GIS). Additionally, archaeological research uses remote sensing techniques especially to look for irregularities on the ground represented by differences in vegetation and soils that may indicate previous anthropogenic developments. Such large scale investigation then often leads to mesoscale research associated with excavation at certain sites. It is in this part of the process where soil science may contribute to studying landscape stability/instability and environmental change.

At the mesoscale, soil methodologies and techniques are often applied to smaller areas. In this case, it may involve survey of a small area of a site or a full site. This could still be part of the survey phase, but also of excavation. It may also involve the use of geophysical tools to provide data that are sometimes used to search for areas of occupation below the subsoil. Test pits (1 m × 1 m) are often dug to expose sediments and soils, which may be followed by field sampling of the profile and laboratory analyses. Samples may be taken preserving the natural soil aggregation to make slides for micromorphological investigations. The sampling phase might also involve examining exposures of soils outside settlement sites, especially in areas of accumulation like floodplains. Techniques used here may involve quick analyses of physical and chemical properties of the soil, especially to study soil development or indicators of human transformation of the soil in the past.

At the microscale the main tasks are detailed excavations of sites, and it is here where soil science contributes more to the aspect of soil taphonomy. Techniques used here mainly involve soil micromorphology and certain chemical analyses. Outside archaeological sites, microscale studies may target anthrosols and farmed soils, for which soil micromorphology would provide valuable information about the materials they are made of, or as in the case of the profiles in Fig. 5, would provide information on the processes of farming. Micromorphology would also provide information on soil processes induced by humans such as clay translocation, salinization, and so on.

Final remarks

Soil research is an integral part of archaeological research, which is now developing in tandem with geology, geomorphology and other geosciences under the broad field of geoarchaeology. Soil studies are an essential part of the archaeological research during survey, excavation, and laboratory analysis.

Among the many contributions of soil research to archaeology, one can cite the following:

- The process of abandonment and degradation of archaeological sites and their artifacts
- Source of absolute and relative dates for stratigraphic sequences
- Survey of archaeological sites
- Physical and chemical properties of soils inside and outside archaeological site
- Climatic and environmental changes as well as stability of the landscape
- Reconstruction of past ecosystems
- Soils as an agricultural resource
- Details on past agricultural practices

The study of soils in geoarchaeological research encompass various spatial scales. The macroscale involves regional survey of soils and their association with ancient settlements for which a series of mapping techniques are implemented. This also involves the study of soils as a measure of landscape stability and environmental change. At the mesoscale soil research includes the inspection of sites and the transformation of soils by human activities, which normally comprises studies of the physical and chemical properties of soils. Finally, at the microscale, soil properties of certain layers or horizons are examined, often associated with site taphonomy, but also with farmed and anthropogenic soils.

In conclusion, it is possible to see that methods and techniques used in soil science have been adapted to archaeological research via the broad research approach known as geoarchaeology. This means, however, that soil scientists dedicated to this field should be conversant with the theoretical and methodological approaches in other geosciences and archaeology.

References

- Beach T, Ulmer A, Cook D, Brennan ML, Luzzadder-Beach S, Doyle C, Eshleman S, Krause S, Cortes-Rincon M, and Terry R (2018) Geoarchaeology and tropical forest soil catenas of northwestern Belize. *Quaternary International* 463: 198–217.
- Bockheim JG (2014) *Soil Geography of the USA: A Diagnostic-Horizon Approach*. Heidelberg: Springer.
- Butzer KW (1982) *Archaeology as Human Ecology*. Cambridge: Cambridge University Press.
- Cordova CE (2017) Pre-Hispanic and Colonial Flood Plain Destabilization in the Texcoco Region and Lower Teotihuacan Valley, Mexico. *Geoarchaeology* 32: 64–89.
- Cordova CE (2018) *Geoarchaeology: The Human-Environment Approach*. London: IB Tauris-Bloomsbury.
- Cordova CE and Parsons J (1997) Geoarchaeology of an Aztec dispersed village.
- Cordova CE, Vyazov LA, Blinnikov MS, Ponomarenko EV, Ponomarenko DS, Sitdikov AG, and Salova YA (2021) Stratigraphy and Paleolithic landscapes of the Beganchik site at the Kama-Volga confluence. *Поволжская археология* 37: 8–21. <https://doi.org/10.24852/pa2021.3.37.8.21>.
- Goldberg P and MacPhail RI (2006) *Practical and Theoretical Geoarchaeology*. Oxford: Blackwell Publishing.

- Holliday VT (2004) *Soils in Archaeological Research*. New York: Oxford University Press.
- Homburg JA and Sandor JA (2011) Anthropogenic effects on soil quality of ancient agricultural systems of the American Southwest. *Catena* 85: 144–154.
- Kern DC, Costa MLD, and Frazão FJL (2004) Evolution of the scientific knowledge regarding archaeological black earths of Amazonia. In: Glaser B and Woods WI (eds.) *Amazonian Dark Earths: Explorations in Space and Time*, pp. 19–28. Berlin, Heidelberg: Springer.
- Mandel RD (2008) Buried paleoindian age landscapes in stream valleys of the central plains, USA. *Geomorphology* 101: 342–361.
- Nicosia C and Devos Y (2014) Urban dark earth. In: Smith C (ed.) *Encyclopedia of Global Archaeology*, pp. 7532–7540. New York: Springer.
- Rapp G Jr. and Hill CL (2006) *Geoarchaeology*, 2nd edn New Haven, CO: Yale University Press.
- Society for American Archaeology (SAA) (2022) *What is Archaeology*. Available at: <https://www.saa.org/about-archaeology/what-is-archaeology> (Last viewed June 16, 2022).

Civilizations and the role of soil, agriculture, and climate change

Jeffrey A Homburg^{a,b}, Eric C Brevik^c, and Jonathan A Sandor^d, ^aCatena Affiliates, Tucson, AZ, United States; ^bSchool of Anthropology, University of Arizona, Tucson, AZ, United States; ^cCollege of Agricultural, Life, and Physical Sciences, Southern Illinois University, Carbondale, IL, United States; ^dAgronomy Department, Iowa State University, Ames, IA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	29
Soil management in early civilizations	30
Asia and Oceania	30
Europe	30
Africa	31
Americas	31
Effects of ancient agriculture on soils and societies	31
Sustainability and degradation	32
Climate change and ancient cultures	33
Conclusions	35
References	35

Abstract

Civilization is tied to soil through our reliance on agriculture and building materials. As early civilizations developed new agricultural strategies, their knowledge of soil also expanded. Major innovations included irrigation, terracing, use of plows, contour tillage, soil classification, and maintaining soil fertility through intercropping, crop rotations, fallowing, and amendments like manure and ash to develop sustainable agricultural systems. In several noteworthy cases, soil degradation contributed to civilization collapse. In addition, soils are influenced by and influence the global climate system. Changing climates have contributed to past civilization collapses, and societies that rely on degraded soils are less resilient and adaptable to climate change.

Key points

- Describe major agricultural soils and systems for ancient civilizations of the world
- Review examples of agricultural soil management, sustainability, resilience, and degradation in the context of climate change in the past
- Demonstrate the relevance of soil–agriculture–climate change interactions for past civilizations in modeling and adapting to climate change in the future

Introduction

The success of modern and ancient societies is inextricably linked to the availability and distribution of productive, well-watered soils and landforms suitable for agriculture. Ancient agricultural systems permitted cultures to obtain more food and store it for longer periods, which in turn supported permanent settlements and growing populations beyond that possible by hunting and gathering. Agriculture is tied to soil that functions as the medium for anchoring plants and supplying water and nutrients to crops. Human knowledge of soil must have evolved since the earliest days of agriculture, but it is noteworthy Neolithic settlements were located near rich soils that were well-suited to support agriculture. This indicates that these ancient farmers already had relatively sophisticated knowledge about soil fertility and productivity (Montgomery, 2007). As early civilizations rapidly developed new agricultural strategies, their knowledge and management of soil also expanded. Major innovations in the first few millennia of agriculture included irrigation, terracing, plows, contour tillage, and soil classification. Soil fertility was maintained by intercropping, crop rotations that included legumes, fallowing, manuring, and the addition of ash to develop sustainable agricultural systems that persisted for millennia. By contrast, poor agricultural management has caused soil degradation that contributed to civilization collapse. Changing climates have also contributed to such collapse in the past, with societies coping with degraded soils that were more susceptible to problems caused by climate change. It is important to understand soil–climate–human interactions. Gaining a holistic understanding transcends any single academic field. Interdisciplinary and transdisciplinary studies that integrate the expertise of soil scientists, agronomists, climatologists, archeologists, and various physical, biological, and social scientists are necessary to plan better for climate change and how to manage it. Climate change and its impact on both soils and civilizations is a major topic of interest today as a way to adapt to changes in climate in the future.



Fig. 1 Approximate locations of countries, civilizations, and culture areas referenced in this chapter, plus the 1930s Dust Bowl. The red oval encircling the Mediterranean Sea approximates the expanse of the Roman Empire. Photographs: Colca Valley (courtesy of Jonathan Sandor), terra preta and nearby non-terra preta soil (courtesy of David Laird), and farmer working in a terraced paddy in the Philippines (courtesy of International Rice Research Institute and Gene Hettel).

Soil management in early civilizations

Many early civilizations developed ways to improve agricultural production, and much of our understanding of early soil knowledge is from archeological studies (Brevik et al., 2018) (Fig. 1). The earliest evidence of soil manipulation for agricultural production dates to approximately 9000 BCE; between 7500 and 500 BCE major innovations included irrigation terracing, early plows, and contour tillage (Troeh et al., 2004). Observations of natural plant growth, soil colors and textures, well-watered landscape positions, and trial and error were ways that early people selected the best land for farming. Soil spatial patterns were recognized by different civilizations to select cropping (Krupenikov, 1992) and settlement sites by 3000–2000 BCE. Early Chinese (Gong et al., 2003) and Greek civilizations (Krupenikov, 1992) were classifying soils by 300 BCE. The study of soils had not yet advanced to a science in early civilizations, but ancient cultures were well aware of the importance of soils and how to manipulate them to their advantage. Krupenikov (1992) argues that soil science appeared as an empirical field about 500–0 BCE. Examples of early knowledge from different parts of the world are reviewed below.

Asia and Oceania

The oldest indications of human management of soil found to date come from Asia, near Jarmo in Iraq, and date to about 9000 BCE (Troeh et al., 2004) and shortly afterwards in Abu Hureyra in modern Syria (Montgomery, 2007). Irrigation was developed in modern-day Iraq by 7500 BCE (Troeh et al., 2004), and early plows were developed in the Middle East between 6000 and 4000 BCE, which revolutionized soil preparation for planting (Hillel, 1991). The first cultivation and plow appeared about the same time in fertile land in Mesopotamia, but it was necessary to increase crop production by means other than just adding new agricultural fields (Montgomery, 2007). Sheep and goats were domesticated by about 8000 BCE in the Zagros Mountains area and manure was used to fertilize fields (Montgomery, 2007). Furthermore, fields were created from 'artificial' soils that formed small islands in marshes as early as 3000 BCE (Fitzpatrick, 2004). Terraces were first used in the Near East around 4000–3000 BCE. Agricultural production began in China not long after it began in the Middle East (Montgomery, 2007), with rice (*Oryza sativa* L.) cultivated by 7000 BCE, paddy soils used by 4000 BCE (Cao et al., 2006), and terraces by 1000 BCE. Plowing was a common practice in India by 3000 BCE and farmers amended soils in Uzbekistan to improve both fertility and texture as early as 2000 BCE (Krupenikov, 1992). Early drainage canals found in Papua New Guinea date to 7000 BCE with more extensive drainage systems and the construction of raised soil beds by 4000 BCE (Vasey, 2002).

Europe

Agriculture spread from the Middle East west into Europe, starting with Turkey and Greece about 6300 BCE (Montgomery, 2007). Ancient Greek philosopher-scientists recognized differences between soils by the second millennium BCE; they are credited with the first recorded works that show knowledge of soil properties and with developing the concept of a soil profile (Krupenikov, 1992). They understood that plant nutrients were supplied by the soil and wrote of water storage in soils (Hillel, 1991).

Their attention to soils allowed the Greeks to match crops to the best locations for production (Krupenikov, 1992). Despite this, soil erosion and degradation became major problems in ancient Greece.

Romans utilized manure and green manure to enhance soil fertility, built terraces to reduce erosion (Troeh et al., 2004), classified soils, and developed soil fertility tests. Because the Roman Empire encircled the Mediterranean and reached far into western and northern Europe (see Fig. 1), Roman ideas on soils and agriculture were widespread. In other parts of Europe, contour tillage was developed in the British Isles prior to the Roman conquest, and some cultivated land on steep slopes was reforested for conservation purposes by 900 CE (Krupenikov, 1992).

Africa

North Africa had some very successful ancient agricultural societies. The ancient Egyptian civilization lasted for 3000 years, from 3300 BCE to 332 BCE, based on agriculture sustained by frequent flooding of the Nile River that restored fertility by depositing organic-rich sediments. Farther west, Carthage became a political power around 600 BCE based on the cultivation of cereals, fruits, vegetables, and nuts. The Roman conquest of Carthage in 146 BCE secured land for crops needed to feed Rome, and the Romans were very interested to learn of Mago's handbook of Carthaginian agriculture (Montgomery, 2007). The Carthaginians had advanced cultivation and irrigation systems, but wind and water erosion eventually removed enough topsoil around Carthage to make agriculture unsustainable.

Americas

The earliest evidence of agriculture in the Americas dates to between 8500 and 7000 BCE in Peru and Mexico (Vasey, 2002). Maize (*Zea mays* L.) was domesticated by 7000 BCE, which led small dispersed villages in Mesoamerica to grow into larger towns that served as cultural and economic centers (Montgomery, 2007). Mayan civilization thrived in Central America from about 600 BCE until 800–900 CE as population increased from around 200,000 to 3–6 million, supported by dynamic and ever evolving agricultural systems that varied spatially and temporally (Dunning et al., 2012). Population grew as agriculture was intensified. Mayan agriculture evolved from slash and burn under low population density to raised fields, drainage by canals in lowlands, and extensive terraces on slopes, but they lacked domesticated animals and manure for renewing soil fertility.

Crops were first cultivated by Native Americans in the Midwest and Eastern USA by 5000 BCE. Intercropping, ash amendments, and fallowing rejuvenated soil fertility. Agricultural management in the American Southwest included runoff terraces, rock mulch, and irrigation systems matched to particular soils and landforms (Homburg and Sandor, 2011). Domesticated maize spread from Mexico's Tehuacan Valley north to the American Southwest by 2000 BCE.

Agricultural terraces date back at least 4000 years in Peru, and many ancient terraces are still cultivated today (Denevan, 2001; Sandor and Homburg, 2017). Intercropping, crop rotations that included legumes, fallowing, and ash were used to maintain soil fertility. Animal domestication in South America provided manure that was absent elsewhere in the Americas. Native Americans in the Bolivian Amazon established raised fields and canal systems that supported complex societies with a relatively high population density (Lombardo et al., 2015).

In the Amazon Basin, human activities created patches of Dark Earth, or terra preta, between 500 BCE and 1640 CE. Terra preta have thickened A horizons formed from surficial accumulation of archeological residues at ancient settlements and associated agricultural fields and gardens. These residues include charcoal, ash, organic refuse, fragments of pottery, and stone, bone, and shell artifacts. Most tropical soils are marginal for agriculture due to intensive chemical weathering, but terra preta have high fertility resulting from elevated soil organic matter, pH, cation exchange capacity, nutrients (especially phosphorus (P), calcium (Ca), magnesium (Mg), zinc (Zn), and manganese (Mn)), and favorable physical properties, including greater workability, permeability, water-holding capacity, aggregate stability, and root penetrability. It is unknown if terra preta were initially created incidentally or intentionally, but over time these fertile soils were managed deliberately for agriculture (Lombardo et al., 2021). Crop productivity is significantly increased in terra preta relative to neighboring areas where highly weathered and leached, acidic Oxisols and Ultisols typify tropical environments. Recalcitrant, pyrogenic carbon from charcoal and stable humus account for much of their fertility. The sustainability, food security, and environmental benefits of terra preta are recognized today. They sequester substantial amounts of carbon, which has spurred research over the last two decades on the properties and benefits of biochar (charcoal produced by pyrolysis of biomass in an oxygen-limited environment for uses other than as fuel) as a soil amendment and many other purposes. Biochar has become a leading candidate for both mitigating climate change and boosting agricultural productivity sustainably.

Effects of ancient agriculture on soils and societies

Humans have manipulated soils for at least 11,000 years to improve agricultural production. A common result was soil degradation (Richter, 2007). Water and wind erosion are natural geologic processes, but agriculture is known to accelerate erosion and strip away topsoil faster than it can form in many environments. On a worldwide basis, rill, gully, and sheetwash erosion remains the dominant cause of soil degradation. Compaction, nutrient loss, and salinization are also major causes of soil degradation. Table 1 summarizes the main physical and chemical properties used for assessing soil degradation. There are also numerous examples of sustainable systems. We can learn from both sustainable and unsustainable practices in the past. This section provides examples of both sustainability and soil degradation severe enough to have caused or contributed significantly to civilization decline.

Table 1 Soil properties that may indicate soil degradation caused by cultivation.

Soil property	Criteria for recognizing degradation: Typical causes and consequences
A horizon thickness	Decreased thickness caused by water or wind erosion. Reduces important organic matter-enriched surface layer that can be exploited by plants for water, nutrients, and oxygen. Shallower depth to possible root-limiting subsurface layers such as strongly developed argillic horizons.
Soil structure	Macromorphology: lowered grade of granular or subangular blocky structure, trend toward massive state, especially in surface horizons. Commonly results from compaction and organic matter decline.
Bulk density	Compaction (increase in bulk density above that of natural condition) associated with degradation in soil structure. Compaction and structure degradation commonly retard seed germination and root growth, reduce root access to water, oxygen, and nutrients, reduce diffusion of gases, and decrease water infiltration and available water capacity.
Organic carbon	Decrease in organic C is common under conventional cultivation. Results from accelerated microbial oxidation of organic matter in disrupted, exposed soil aggregates, and other effects of agriculture. Numerous benefits of organic matter for soil physical, chemical, and biological properties important to plant growth are well documented.
Nitrogen	Decrease in total N accompanies declining organic matter in agricultural soils, though C:N ratio tends to decrease. Nitrate and ammonium are plant available forms of N, which is commonly a key limiting factor for plant growth in all regions, including arid regions.
Phosphorus	P (both total and available) is another macronutrient that has been shown to decrease due to cultivation in some cases. Phosphorus is a key ecological and soil indicator because of its low mobility, low availability to plants, and long-term stability of its forms in soils.
pH	Very high soil pH can indicate salt accumulation (which is measured by electrical conductivity). Sodic soil conditions (recognized by high exchangeable sodium) can be prevalent in agricultural soils of arid and semiarid regions. Acidification and low pH is common in humid and tropical environments due to intensive leaching and weathering. Detrimental effects on many plants, including crop species, occur both through direct chemical effects and through soil structural deterioration. pH is also an indicator of the availability of nutrients to crops.

Adapted from Brevik EC, Homburg JA, and Sandor JA (2018) Soils, climate, and ancient civilizations, in Horwath WR, and Kuzyakov Y (Eds.), *Climate Change Impacts on Soil Processes and Ecosystem Properties*, pp. 1–28. Developments in Soil Science, vol. 35, Elsevier B.V. Amsterdam.

Sustainability and degradation

Mesopotamia is a classic example of anthropogenic land degradation that contributed to civilization collapse, but along the southern reaches of the Tigris and Euphrates rivers agriculture was sustained for five millennia (3000 BCE to 1980 CE) based on the production of dates (*Phoenix dactylifera* L.), vegetables, and water buffalo. High concentrations of N and P in soils indicate a slow buildup of organic-rich materials through anthropogenic activity. This lifestyle lasted until large-scale dams were built on the Tigris and Euphrates and marshes were drained in the 1970s and 1980s (Fitzpatrick, 2004).

In his classic work “Farmers of Forty Centuries,” King (1911) noted that despite dense populations and relatively little arable land, civilizations in eastern Asia managed to sustain agriculture for 4000 years. King attributed their success to several factors: (1) abundant rain and long growing seasons that allowed multiple crops per year, (2) intercropping, (3) matching crops to local conditions, (4) good water management that combined irrigation and dryland cropping systems on the same land, depending on season, (5) using mulch for water conservation during drier periods, and (6) using multiple nutrient sources as fertilizers (e.g., green manure, compost, ash, animal and human manure, and legumes in crop rotations). King concluded that the favorable climate was essential for sustaining East Asian agriculture. The obvious corollary is that climate change could undermine the sustainability of these practices. Paddy soils in Asia have a number of unique properties because of the alternations between being saturated and drained over time, which cause anaerobic and aerobic conditions and reduction and oxidation processes in the soil. The result is elevated organic matter levels, altered mineral weathering and clay formation, acidification, puddled surfaces, albic E horizons, and redoximorphic features (that is, iron and manganese depletions and concentrations) where water levels fluctuate.

Some agricultural soils in the Andes Mountains of South America have been cultivated for at least 1500 years and still remain productive (Sandor and Homburg, 2017). Although located in steep mountainous terrain, terraces were built to create gentle slopes for arable land and provided a way to manage irrigation. Manure and ash from fire hearths were incorporated into terraced soils. These practices created thickened A horizons in terraced soils that had lower bulk density and elevated C (carbon), N (nitrogen), and P levels compared to nearby, natural uncultivated soils on steep mountain slopes. However, some researchers have questioned the sustainability of these terrace systems during prolonged droughts (Branch et al., 2007). One important lesson here is that a given system may be sustainable under some climates, but not others.

Unfortunately, there are many examples of soil degradation in human history, with one of the most cited examples being Mesopotamia. Sumerians irrigated successfully for a long period, but their civilization declined as soils became too saline and waterlogged for crop growth due to salts in irrigation water and rising groundwater levels caused by irrigation (Hillel, 1991). The Babylonians rose to power after the Sumerians until Babylonian canals filled with silt eroded from surrounding hills, triggered by forest clearance from hills around river valleys to obtain timber and create grazing areas for livestock (Troeh et al., 2004). Along the borders of southern Iraq’s marshlands, the same location where agriculture persisted for five millennia, rice was produced as early as 550 BCE. Accelerated sedimentation covered rice fields over time so that fields were abandoned and new rice fields created, leaving large expanses of unproductive land behind. This example shows how proper water management and soil conservation are critical to sustainable agriculture.

Another classic example of cultural collapse due to soil degradation and resource exploitation is found on Easter Island (Rapa Nui), where Polynesian settlers began shade gardening in forests about 1200 CE but shifted to intensive dryland farming as extensive deforestation and soil erosion occurred within a century of occupation (Sherwood et al., 2019). These early settlers arrived on an island vegetated by mesic forest including palm trees (*Paschalococos disperta* Dransfield), but deforestation was nearly complete by 1650 CE. Sweet potato (*Ipomoea batatas* (L.) Lam.) was their main staple, but other crops included banana (*Musa* sp.), taro (*Colocasia esculenta* (L.) Schott), paper mulberry (*Broussonetia papyrifera*, Syn. *Morus papyrifera* (L.)), and in all likelihood, bottle gourd (*Lagenaria siceraria* (Molina) Standl.) (Sherwood et al., 2019). Soils at many sites were degraded on the island, mainly by leaching and loss of nutrients, especially P, whereas others remained fertile in localized area, especially soils formed from the weathering of lapilli tuff, a type of volcanic pyroclastic. At one time Easter Island may have supported a population of as many as 10,000 people, but by 1650 CE the population had declined and the society collapsed. When James Cook visited Easter Island in 1774, he found about 1000 people only. Hypotheses for explaining the collapse on Easter Island include: (1) overexploitation and degradation of the island's resources, especially palm trees, but also soil and water resources, (2) contact with European explorers who spread disease vectors, and (3) environmental changes beyond human control. Of these, the first seems the most likely cause. If true, Easter Island shows what can happen when humans abuse ecosystems. If the third hypothesis is true, it serves as a warning regarding the potential impact of environmental changes such as climate change. It is also possible, if not likely, that a combination of factors is the best explanation. This is a sobering possibility as we contemplate a modern world with an environment that is degraded in many ways coupled with a changing climate.

Climate change and ancient cultures

Both climate and climate change have strongly influenced human societies throughout history. Warmer climates at the end of the Pleistocene facilitated the beginnings of agriculture and complex societies. Many early civilizations in both the Old World (Mesopotamia, Egypt, Indus, Greece, Roman Empire, and other places in Asia, Africa, and Europe) and the Americas (Peru, highland Mexico, American Southwest) arose in arid regions. Favorable climates provided conditions for past societies to take root and flourish. Climate change, especially prolonged drought, has been implicated as a critical factor in the demise and collapse of some civilizations in the past.

Some civilization failures were probably caused by a combination of climate changes and cultural factors. A prime example were the Norse people in Greenland. The climate in Greenland was unusually mild when the Norse first settled Greenland in the late 900 s CE, followed by the onset of the Little Ice Age about 1300 CE and complete abandonment of Norse settlements in Greenland in the 1400s CE. Clinging to cultural traditions has been suggested as an explanation for the failure of Norse settlements in Greenland as the Little Ice Age brought on a colder climate (Fig. 2). Debate exists over whether soil degradation was a problem for



Fig. 2 Standing architecture of the Hvalsey Church and associated masonry rubble from collapsed farm buildings in an ancient Norse settlement in southern Greenland. Courtesy of Wikipedia: https://en.wikipedia.org/wiki/History_of_Greenland#/media/File:Hvalsey_Church.jpg.

Norse settlements in Greenland, with some researchers concluding there is no evidence of significant widespread soil degradation while others have concluded erosion was a significant problem in at least some locations (Brevik et al., 2018). There are, for example, buildings at some sites that were buried by drifting sand, likely to have been initiated by overgrazing. However, there is little debate that climate played a role in the Norse demise. The abandonment of Norse settlements in Greenland in the 1400s CE after the onset of the Little Ice Age led many researchers to conclude that the Norse settlements in Greenland failed due to climate. Studies indicate a change in diet from primarily terrestrial food sources to greater dependence on marine food sources during the Norse presence in Greenland; a cooling climate is one reason that may have spurred this dietary shift. Despite adjustments such as a shift in diet, the Norse settlements in Greenland eventually failed. It is noteworthy that Inuit communities, by contrast, were able to survive the same climate change that the Norse settlements in Greenland could not. The decline of the Norse Greenland settlements was probably caused by a combination of climate change and a conservative culture that was resistant to change, leaving the Norse vulnerable to environmental alterations. The message from this for the future is that land degradation is not the only challenge to the persistence of a civilization, cultural inflexibility may also be a serious liability in a world of changing climate.

Societies may not only be negatively affected by climate change in an absolute sense, but also by increased climate variability and unpredictability. Previous climate change during the Holocene that has affected societies globally (e.g., the Medieval Warm Period and the Little Ice Age) or regionally are attributed to oceanic and atmospheric circulation cycles such as the El Niño-Southern Oscillation, cyclic and episodic changes in solar radiation, and volcanic activity. Humans must now be added to this list of drivers of climate change.

Climate change is a significant factor in social change in the past and present, but it is also important to recognize that causes for the rise and fall of societies involve complex interacting social, economic, and environmental variables, not just climate change. Views that climate change determined past social change are too simplistic because societies are not passive in the face of environmental change and variability, but actively respond and adapt (Kirch, 2005). Determining the cause(s) of past social transformation or collapse is controversial. This complexity in the interaction of humans and environment is evident in the wide range of human responses to climate change over space and time. Some environments are inherently riskier than others, and some societies have been more resilient than others. Archeologists distinguish different degrees of vulnerability of societies to instability and failure in the face of environmental change and other pressures. Some societies that lacked the flexibility to respond to environmental change were caught in a 'rigidity trap.'

It is also important to recognize the challenges in measuring and interpreting past climate change in studies of ancient societies (Ingram, 2015). A number of tools exist for reconstructing paleoclimate (Fagan, 2008) with varying degrees of accuracy and precision, and multiple independent approaches are important. Even the finest of these climate markers, such as tree-rings, which provide precipitation and temperature data on an annual scale, are still too imprecise for reconstructing conditions during crop growth, such as during the stages of maize development withing a growing season.

Cases from several world regions, climates, and cultures illustrate the broad range of climatic and environmental impacts on water resources, conditions during the crop growing season, and food production encountered by past societies. These include many combinations of environmental and social conditions, and outcomes for past societies. Some researchers have identified collapses in ancient societies in which severe drought played a significant role. These have occurred in northern China, the Indus Valley, Mesopotamia, Peru and the Andes, the Maya region, and North America (Dunning et al., 2012; Fagan, 2008; Montgomery, 2007). Societal collapse has involved political upheaval, conflict, and long-term depopulation. To illustrate the complexity of social transformation in relation to climate change, the drought of the mid-1100s AD in the American Southwest associated with the demise and depopulation of Chaco Canyon affected other Southwestern cultures in the Mesa Verde and Mimbres areas (local population shifts) differently (Ingram, 2015). In some cases, systematic demographic and land use change in response to climate change has been a consistent component of some social structures, as in the 'verticality' adjustments of Andean people and the movement of agriculture to higher elevations with more moisture during drier periods or to lower elevations during wetter periods.

There are a number of cases in which ancient and traditional societies were able to respond successfully to climate change and remain in place. Of potential relevance today are the innovations developed by past peoples in actively adapting to climate challenges and change, including new technologies, crops, and agricultural strategies. For example, varieties of maize were developed in the Americas that were adapted to aridity or high altitudes. In the American Southwest, desert plants such as agave (*Agave* sp.) were domesticated in part to offset food shortages when other crops failed. Several of these adapted crops, the landscapes on which they were grown, and the techniques by which they were managed are not used today but have potential (Sandor and Homburg, 2017). One of the most important principles in adaptive traditional agricultural systems is their diversity in terms of crops and field location. Agricultural diversification is a deliberate strategy of risk management in the face of environmental vagaries. Strategies of landscape risk management for climate variability, widely practiced in arid regions such as the American Southwest, include spatially diverse field locations, combinations of irrigated agriculture in valleys with dryland farming at valley edges and in uplands, and specially adapted crops and crop management, combined with food storage techniques. Irrigation agriculture may have been the mainstay, but upland production could protect the soil against damaging floods or freezes relative to fields in valley floors that must cope with cold air drainage risks. Innovative technologies developed in response to climate fluctuation, water scarcity, and climate change-related hazards such as soil erosion. Solutions to these problems include agricultural terraces, lithic mulch and rock configurations for concentrating and conserving soil moisture and moderating temperatures, water harvesting, wells, and qanats (gently sloping underground channels with vertical shafts used to access and transport water from groundwater to agricultural fields).

Climate change not only impacts societies directly, but also through complex interactive effects among hydrologic, geomorphic, and soil processes, and vegetation change. Positive or negative outcomes in terms of impact on agriculture and society are possible,



Fig. 3 1930s Dust Bowl on the southern Great Plains, USA. (A) massive dust storm during the Dust Bowl. (B) heavily eroded agricultural fields and windrow of redeposited topsoil along a fence line eroded from agricultural fields left in the aftermath of a dust storm. Courtesy of the United States Department of Agriculture, Natural Resources Conservation Service, USA.

but negative outcomes are more likely. For example, stream channel widening and incision caused by climate change after centuries of geomorphic stability played a role in reorganization of the prehistoric Hohokam society in the Phoenix Basin of Arizona by disrupting extensive canal irrigation systems (Waters, 2006).

Another key point is that it can be difficult to separate the causes of land degradation from either natural climate change or anthropogenic impacts because they often operate together through feedback processes. Degradational effects of climate change may be triggered by anthropogenic landscape and ecological change or vice versa. There are instances in which climate change may not have caused landscape degradation such as accelerated erosion had it not been preceded by anthropogenic change. An added dimension to this today is human-induced climate change.

Researchers infer a link between climate change and anthropogenic environmental change that has negatively affected many past societies. Soil-geomorphic and ecological degradation associated with climate change and agriculture has been documented in several world regions. Common examples include accelerated soil erosion in sloping agricultural lands in China, Europe, the Mediterranean and Middle East, the Mayan area, and the American Southwest (Butzer, 1982; Hillel, 1991; Kennett and Beach, 2013; Kirch, 2005; Montgomery, 2007; Nabhan et al., 2020; Sandor and Homburg, 2017). A historic example is the Dust Bowl in the U.S. southern Great Plains, where extensive clearing and cultivation of prairie soil combined with a periodic drought led to severe wind erosion, land abandonment, and depopulation (Montgomery, 2007) (Fig. 3).

Climate change, now linked with human causes, presents serious challenges to societies today. The failures and successes of ancient societies in their land use and response to a wide range of climate and environmental change can provide valuable long-term insight for addressing current problems such as water supply and agricultural production under difficult and variable climate conditions.

Conclusions

Many scientists now consider that we are living in the Anthropocene, an age when human influences at a global scale have become so pronounced and pervasive that humans are the single most defining geologic force of the era (Zalasiewicz et al., 2015). This influence has made many soils human-natural bodies rather than just natural bodies (Yaalon and Yaron, 1966). Another effect of humans on the natural system has been our influence on changes to global climate. Soils are changed by climate and, in turn, influence climate. Consequently, as we seek to understand how living in a world of changing climate may affect modern civilizations, understanding how soils will respond, and how that affects humans is imperative. One approach to undertaking this task is to model what the future will look like given the changes that we expect to see. However, human civilizations have a long record of living under changing climates in the past. Through interdisciplinary and transdisciplinary studies that link soil scientists, archeologists, and other scientists, we can gain a more holistic understanding of the impact of climate change on past human societies and natural resources such as soils that they depended on and have the opportunity to observe those impacts. By paying attention to and learning from our past, we can improve the models that aim to predict our future and plan to manage and minimize potential impacts better.

References

- Branch NP, Kemp RA, Silva B, et al. (2007) Testing the sustainability and sensitivity to climatic change of terrace agricultural systems in the Peruvian Andes: a pilot study. *Journal of Archaeological Science* 34: 1–9.
- Brevik EC, Homburg JA, and Sandor JA (2018) Soils, climate, and ancient civilizations. In: Horwath WR and Kuzyakov Y (eds.) *Climate Change Impacts on Soil Processes and Ecosystem Properties. Developments in Soil Science*, vol. 35, pp. 1–28. Amsterdam: Elsevier B.V.

- Butzer KW (1982) *Archaeology as Human Ecology: Method and Theory for a Contextual Approach*. Cambridge: Cambridge University Press.
- Cao ZH, Ding JL, Hu ZY, et al. (2006) Ancient paddy soils from the Neolithic age in China's Yangtze River Delta. *Naturwissenschaften* 93: 232–236.
- Denevan WM (2001) *Cultivated Landscapes of Native Amazonia and the Andes*. New York: Oxford University Press.
- Dunning NP, Beach TP, and Luzzadder-Beach S (2012) Kax and kol: collapse and resilience in lowland Maya civilization. *Proceedings of the National Academy of Sciences* 109: 3652–3657.
- Fagan BM (2008) *The Great Warming: Climate Change and Rise and Fall of Civilizations*. New York: Bloomsbury Press.
- Fitzpatrick RW (2004) *Changes in Soil and Water Characteristics of Natural, Drained and Re-flooded Soils in the Mesopotamian Marshlands: Implications for Land Management Planning*. CSIRO Land and Water Client report.
- Gong Z, Zhang X, Chen J, and Zhang G (2003) Origin and development of soil science in ancient China. *Geoderma* 115: 3–13.
- Hillel D (1991) *Out of the Earth: Civilization and the Life of the Soil*. Berkeley: University of California Press.
- Homburg JA and Sandor JA (2011) Anthropogenic effects on soil quality of ancient agricultural systems of the American Southwest. *Catena* 85: 144–154.
- Ingram SE (2015) Human vulnerability to dry periods. In: Ingram SE and Hunt RC (eds.) *Arid Lands Agriculture: Understanding the Past for the Future*, pp. 131–160. Tucson: University of Arizona Press.
- Kennett DJ and Beach TP (2013) Archeological and environmental lessons for the Anthropocene from the Classic Maya collapse. *Anthropocene* 4: 88–100.
- King FH (1911) *Farmers of forty centuries, or permanent agriculture in China, Korea and Japan*. Privately published by Mrs. Madison, Wisconsin: F.H. King. https://soilandhealth.org/wp-content/uploads/King_Farmersof40Centuriespt1.pdf. https://soilandhealth.org/wp-content/uploads/King_Farmersof40Centuries.pt2_.pdf. Accessed 29 August, 2022.
- Kirch PV (2005) Archaeology and global change: the Holocene record. *Annual Review of Environment and Resources* 30: 409–440.
- Krupenikov IA (1992) *History of Soil Science from Its Inception to the Present*. New Delhi: Oxonian Press.
- Lombardo U, Denier S, and Veit H (2015) Soil properties and pre-Columbian settlement patterns in the Monumental Mounds Region of the Llanos de Moxos, Bolivian Amazon. *The Soil* 1: 65–81. <https://doi.org/10.5194/soil-1-65-2015>.
- Lombardo U, Arroyo-Kalin M, Huisman H, et al. (2021) Evidence confirms an anthropic origin of Amazonian Dark Earths (Comment to Silva, L. C. R. et al., 2021). "A new hypothesis for the origin of Amazonian Dark Earths" *Nature Communications* 12(1): 127. *Nature Communications* 13:1–7.
- Montgomery DR (2007) *Dirt: The Erosion of Civilizations*. Berkeley: University of California Press.
- Nabhan GP, Riordan EC, Monti L, et al. (2020) An Aridamerican Model for Agriculture in a Hotter, Water Scarce World. *Plants, People, Planet* 2: 627–639.
- Richter D (2007) Humanity's transformation of Earth's soil: Pedology's new frontier. *Soil Science* 172(12): 957–967.
- Sandor JA and Homburg JA (2017) Anthropogenic soil change in ancient and traditional agricultural fields in arid to semiarid regions of the Americas. *Ethnobiology* 37(2): 196–217.
- Sherwood SC, Tilburg JA, Barrier CR, et al. (2019) New excavations in Easter Island's statue quarry: Soil fertility, site formation and chronology. *Journal of Archaeological Science* 111: 1–22.
- Troeh FR, Hobbs JA, and Donahue RL (2004) *Soil and Water Conservation for Productivity and Environmental Protection*, 4th edn Upper Saddle River: Prentice Hall.
- Vasey DE (2002) *An Ecological History of Agriculture 10,000 BC to AD 10,000*. West Lafayette: Purdue University Press.
- Waters MR (2006) Prehistoric human response to landscape change in the American Southwest. In: Doyel DE and Dean JS (eds.) *Environmental Change and Human Adaptation in the Ancient American Southwest*, pp. 26–45. Salt Lake City: University of Utah Press.
- Yaalon DH and Yaron B (1966) Framework for man-made soil changes-an outline of metapedogenesis. *Soil Science* 102(4): 272–277.
- Zalasiewicz J, Waters CN, Williams M, et al. (2015) When did the Anthropocene begin? A mid-twentieth century boundary level is stratigraphically optimal. *Quaternary International* 383: 196–203.

Ecosystem services provided by soils in highly anthropized areas (SUITMAs)

Jean Louis Morel^a, Geoffroy Séré^a, Viacheslav Vasenev^b, and Thomas Nehls^c, ^aLaboratoire Sols et Environnement, Université de Lorraine, INRAE, Nancy, France; ^bSoil Geography and Landscape Group, Wageningen University, Wageningen, The Netherlands; ^cEcohydrology and Landscape Evaluation, Technische Universität Berlin, Berlin, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	37
Definitions	39
Specific features of soils of highly anthropized areas	39
Soils of highly anthropized areas	39
Soil formation in highly anthropized areas	40
Common characteristics of soils in anthropized areas	41
Composition	41
The pH	42
Organic matter	42
Contaminants	43
Diversity of soils in anthropized areas	44
Anthropization produces a large diversity of soils	44
The urban soil mosaic	44
Classification	44
Functions of SUITMAs	44
Support for roots	44
Reservoir of biodiversity	48
Carbon storage	49
Ecosystem services provided by SUITMAs	49
Restoration of soil functions to improve ecosystem services	51
Conclusion	53
References	53
Further reading	55

Abstract

Land under human influence, such as urban, industrial, traffic, mining, and military areas are characterized by a great diversity of soils with specific properties. Some are close to natural soils, but most are greatly modified or even deliberately constructed. These soils of highly anthropized areas render a wide range of ecosystem services, provided that their functions (e.g. carbon storage, habitat for biodiversity) are not altered or that restoration operations are carried out.

Introduction

Soils in highly anthropized areas, such as in cities, industrial complexes, along streets, railways, at airfields, in open pits or underground mining areas and in areas used for military actions can perform a wide range of functions of primary interest (Levin et al., 2017). In cities for instance, they support human activities, e.g. habitat, stable ground for housing, industrial activities, transport and greening. These functions are primarily considered and widely accepted by landowners, policy makers, and city dwellers as they offer monetary income and a minimum of wellbeing.

With emergence of the concept of ecosystem services (MEA, 2005), efforts to identify the greatest number of contributions to the mitigation of global change (e.g. climate, biodiversity), ensuring well-being and prosperity of citizens have led to questions about the role of soils in anthropized areas including those greatly modified by human activities. These soils are primarily found in urban areas, but also in industrial, traffic, mining, and military areas (Burghardt et al., 2015). In view of the great diversity of locations where soils are profoundly affected by human activities, in 1998, Burghardt and Morel proposed SUITMA to designate heavily modified Soils in Urban, Industrial, Traffic, Mining, and Military areas (Burghardt et al., 2015). This definition emphasizes the location of soils, i.e. very heavily anthropized areas, rather than their specific characteristics. The acronym was launched at the 18th World Congress of Soil Sciences held in Montpellier, and scientists who dedicate their research to these soils have been meeting every 2 years since at the SUITMA conferences. Despite the impacts of human actions on the properties of such soils that can lead to some degradation, SUITMAs can provide a wide range of services, namely habitat, provisioning, regulating and cultural services (to humans).

Among anthropized areas, cities are of particular interest because most of the world's population now lives in them. It was only in the last two decades that urban soils have attracted increasing interest from policy makers, municipalities, and soil scientists.

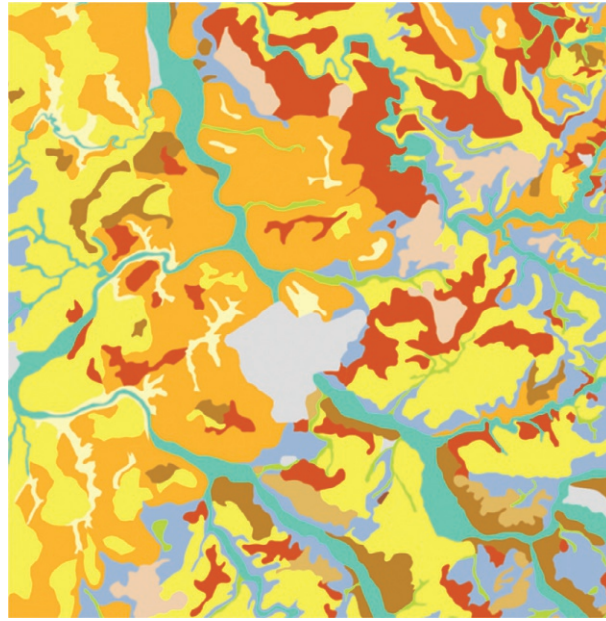


Fig. 1 Cities as they appear on soil maps (here city of Nancy, France) – from Soil map of France, GIS Sol – RMT *Sols et Territoires*, 1/250000 – legend: gray: city; yellow & orange: soils on limestone material; green: soils on alluvial material; dark brown: soils on loam.

Of course, they were not ignored in the past (e.g. Bastie, 1965; Blume and Runge, 1978; Burghardt, 1994; de Kimpe and Morel, 2000), but often considered marginal and, therefore, set aside or even referred to as ‘non-soils.’ Cities, due to the challenges that their complexity induces for soil science and soil mapping, are however still depicted with gray or white spots on soil maps (Fig. 1).

Global soil map resources such as S-world or SoilGrids ignore urban soil information by masking urban areas (Poggio et al., 2021) (Fig. 2).

Also, emblematic, one of the earliest scientific publications on urban soils was not about the ecosystem services they provide, but about their contamination status (Purves, 1966). But now they are recognized as soils with specific characteristics of primary importance, especially in cities (Riddle et al., 2022).

In this chapter, the existing and potential role of SUITMAs and especially of urban soils as providers of ecosystem services in areas dominated by humans as the main controlling factor of the environment are discussed. After giving definitions, special emphasis will be placed on the main characteristics of soils, and the state of their functions will be described. Finally, the ecosystem services provided by highly anthropized areas will be presented, including the services linked to the mitigation of global climate change.



Fig. 2 Urban areas on the SoilGrids layer of soil organic stocks (case of Amsterdam, Netherlands) (<https://soilgrids.org/>).

Definitions

Specific features of soils of highly anthropized areas

Three main features characterize the soils of highly anthropized areas: (i) the presence of technogenic material, (ii) sealing, and (iii) soil construction.

Technogenic material: in addition to soil material of natural origin, other compounds are intentionally introduced into soils, e.g. for road construction, waste deposition, dumping of dredged sediments, soil remediation etc. – or accidentally accrued during the city's history which can be as long as 6000 years. The so-called 'cultural layers,' usually containing construction and demolition debris, can have a thickness of several meters and store considerable amount of carbon, nitrogen and phosphorous, also pollutants and artifacts, i.e. materials created, substantially modified or brought to the surface by humans (Mazurek et al., 2016). Cities often grew on their own 'ashes.'

Technogenic material in soil science is material that is of (i) natural origin but not in its autochthonous location; it may show natural layering, evidence of transport and dumping by vehicles, pumps or bulldozers. This can be dredged material from rivers pumped to a new site, bedrock in mining areas dumped in heaps or tailings or topsoil material from other sites (e.g. transported material); (ii) material that is itself the product or byproduct of technological processes, i.e. an artifact, which is not found in nature like concrete, steel, bricks, or even limestone from chemical processes having a purity which cannot be found in nature; (iii) mixtures of both (i) and (ii).

Soil sealing is defined as the destruction or covering of the ground by an impermeable material.¹ Sealing is the strategy that paves the way for the beginning of colonization and civilization. Paved roads date back to 6000 BCE in the city of Ur in Mesopotamia. Paving is now well established in most cities around the world. In the more recent past, it made it possible to go from unsanitary streets, with mud, waste, smells, to clean sidewalks with sewers, where moving from one place to another in the city became easier and more comfortable, and where life in general became healthier (pandemics like diarrhea, typhus and cholera were reduced). These soils are thus designed to fulfill transport functions for the city at the expense of ecological functions. Pavements require the implementation of a series of technologies, which contribute to stabilize the soil, allow the introduction and concealing of wires, tubes, pipes, etc., and finally allow the circulation of humans and goods either on foot or with vehicles ranging from donkey carts to heavy trucks. However, soil sealing is considered one of the main causes of soil degradation and flooding in the European Union (EU), where more than 2.3% of the EU territory is now covered with an impermeable substrate (Corine Land Cover, EU soil map). Soil sealing is a primary issue regarding the destruction of important soil functions, e.g. water infiltration and runoff, biodiversity, local climate. The functioning of soils as a component of the urban ecosystem, including water flows and root paths under the impervious surfaces, is not well established.

Soil construction: constructed soils are defined as deliberate mixtures of technogenic materials that are associated and implemented to meet specific requirements (Deeb et al., 2020). Constructed wetlands, green roofs and green walls are examples of deliberately designed and constructed soils in urban areas. They are developing rapidly and promise to meet citizens' needs for greener and cooler streets and buildings. For example, urban green infrastructure contributes to a comfortable urban microclimate by cooling air temperature and increasing air humidity (Deeb et al., 2020). Another example is the construction of soils on degraded sites to restore their functions and to seek the conditions for the development of a new ecosystem (Séré et al., 2010; Slukovskaya et al., 2019). These examples will be detailed later to outline the consequences on ecosystem services.

Soils of highly anthropized areas

Anthropogenic soils are created or heavily modified by humans. Highly transformed soils are not only found inside cities, i.e. urban soils, but in every place where human activity is carried out and is converted into an artificial environment. The soil is one of the first components to undergo profound modifications. To highlight the importance of the human factor in the formation and evolution of soils in highly anthropized areas, the soils are identified and grouped according to their location, i.e. SUITMA.

Urban soils: are anthropogenic soils, located inside urban areas (recent or old) (Morel et al., 2005). These soil types are strongly heterogeneous, and are developed from natural, anthropogenic, or naturally occurring but anthropogenically displaced materials or mixtures of both. They show a change in soil properties due to anthropogenic impacts such as sealing and intensive use (Blume, 1992). They are developed from a mixture of materials that are different from the surrounding non-urban soils. Thus, urban soils are named to categorize and group them by their most obvious common characteristic, their location in urban areas. The National Resources Conservation Service (USDA, USA) defined urban soils as follows: "Urban soils are found in watersheds that provide drinking water, food, waste utilization, and natural resources to communities. Urban soils also are located within cities in park areas, recreation areas, community gardens, green belts, lawns, septic absorption fields, sediment basins and other uses" (USDA, online). Here, implicitly the interrelation of soils and their use are highlighted. In urban areas, most soils are simply managed to contribute to the urban landscape, and they are shaped without considering their functional roles in urban ecosystems. This is the case of soils from construction areas, which support lawns, trees, playgrounds, roadsides, etc.

Unlike natural soils, forest soils or agricultural soils, the anthropogenic impact can - if any - only be reconstructed with the help of history. The effects are understandable in retrospect but not logically or predictably so, and in summary urban soils could often be

¹<https://esdac.jrc.ec.europa.eu/themes/soil-sealing>

described as chaotic. However, urban soils have developed from 'natural soils' (including arable land, forests, wasteland). Unlike agricultural soils, urban soils are characterized by multiple different disturbances for different purposes. They have been and are subjected to extractions, additions, mixing, compaction, leveling, contamination, drainage, and sealing. Some of their original horizons have been completely removed by the excavations. Natural soil horizons may also be buried by excavated material or technogenic material transported from elsewhere. As a result, spatial patterns of urban soil properties are determined by a complex combination of conventional (e.g. elevation, vegetation, geomorphology) and city-specific (e.g. functional zoning and urban history), which determine an uneven spatial structure and a unique spatial variability of urban soils even at short distances.

(Pseudo)natural soils: in addition to highly anthropized soils, there are soils in the urban environment that have been only slightly altered by human activity. They are still very close to the original natural soils from which they developed, but they show changes due to urbanization. These might include compaction, the presence of artifacts, a different water regime and, in general, exposure to higher average and maximum temperatures than in their original environment. Thus, in cities, markedly transformed soils coexist with (pseudo)natural soils, the latter being found mainly at sites unsuitable for construction or in protected areas such as urban forests, parks, home gardens and community area (e.g. Tiergarten Park in Berlin, Germany, a former hunting reserve of the Dukes and Kings of Prussia which has become a public park)

Industrial soils: they are soils like those in urban areas, but with very specific characteristics due to the various operations that have taken place during industrial activity. Industries are often built near natural or man-made bodies of water. Consequently, their installation requires not only waterproofing but also the contribution of considerable quantities of materials to stabilize and level the surface, but also to avoid any threat linked to flooding by neighboring waters and high groundwater levels. In general, wastes from industrial activity are (minimally) managed directly on site. Lack of knowledge or awareness results in tailings, heaps and ponds of solid and liquid materials, intermediate by-products and wastes for storage or disposal. Some of the waste contains contaminants which, after industry closure and future use of the site, must be removed or inactivated to ensure safe redevelopment. Remediation is based on physical (e.g. thermal desorption, soil venting), chemical (e.g. chemical oxidation or reduction, soil washing) and biological (e.g. biodegradation, phytoremediation) treatments. Each of the technologies has a specific effect on the composition of the soil and its further operation. For example, thermal desorption removes most of the organic matter and leaves a biologically dead, mostly structureless mineral residue, while phytoremediation in general improves soil fertility and accelerates ecosystem restoration at the cost of a long remediation time.

Soils under traffic: these are typically drained, levelled, and compacted soils and their paving. Most of them are sealed soils, as defined above. Along the roads, water is re-directed and infiltrated or drained into sewers. Traffic-specific contaminants are dusts and soot from exhausts of trains, lorries and cars, particulates from rails, wheels, tires, brakes, pavements, and mechanical abrasions of metals due to friction in all parts of the vehicles. Furthermore, railways and roads are managed to be free of vegetation and ice by application of pesticides and deicing salts.

Mining soils: open pit mining greatly affects soils over very large surface areas. Historically, mining activities have led to the foundations of cities, some of them with a short peak of activity (e.g. Dawson in the Yukon, Canada) others where mining began a long and prosperous history (Freiberg, Clausthal-Zellerfeld, Germany; Zhelesnogorsk and Norilsk in Russia). After closure, mining sites must often be remediated and restored, the original soils can seldom be recovered. Another trajectory is then followed toward the development of a new ecosystem. Remediation of mining sites is based on soil engineering, i.e. on all the procedures and processes developed from the knowledge of soil science and environmental engineering to create new soil profiles whose functions have been improved and, if possible, could get as close as possible to those of the natural soils. With their small organic matter content, strategies are generally based on the addition of organic amendments, e.g. compost, manure, or materials excavated from natural soils. Nowadays, to preserve natural soils, the remediation of mining sites uses secondary materials which are recycled such as manure and compost (Guo et al., 2022).

Military soils: military operation is another human activity with a strong effect on soils. Either effective war or military maneuvers destroy original soils and introduce materials and contaminants (e.g. trace elements, radionuclides, toxic organic compounds). An example is the Verdun area (NE France on banks of Meuse River) which was strongly impacted during World War I. After one century, the soil has evolved toward the original soil type but still contains many artifacts, archives of the dramatic period. Unfortunately, today and in the past wars cause considerable damage to soil properties and functions, relocating soil and introducing technogenic materials (Rawtani et al., 2022).

Soil formation in highly anthropized areas

In anthropized areas, after the original natural soil genesis, soil formation after urbanization is mainly controlled by the human factor (Figs. 3A and B). Human activity controls the parent material (e.g. composition, compaction), transforms the landscape (e.g. relief, water dynamics), introduces and controls plants and animals and, in cities, generates specific climatic conditions (e.g. heat island effect). Therefore, pedogenic processes in SUITMAs are mainly under human influence. However, as discussed in another chapter of this book (Huot and Morel, 2023), processes observed in SUITMAs are like those described for natural soils (Huot et al., 2015). In short, the dominant processes are accumulation (e.g. organic matter from vegetation), transformation with alteration of the primary phases and neoformation of minerals, oxidation–reduction reactions, decarbonization, and transport, such as leaching (e.g. sulfates, carbonates), and precipitation into deeper layers (e.g. secondary calcite). Many other processes may also occur depending on the composition of the parent material. The specificity of the pedogenic evolution of heavily anthropized soils is the coexistence of processes that never or rarely occur simultaneously in the natural environment, such as, for example, the

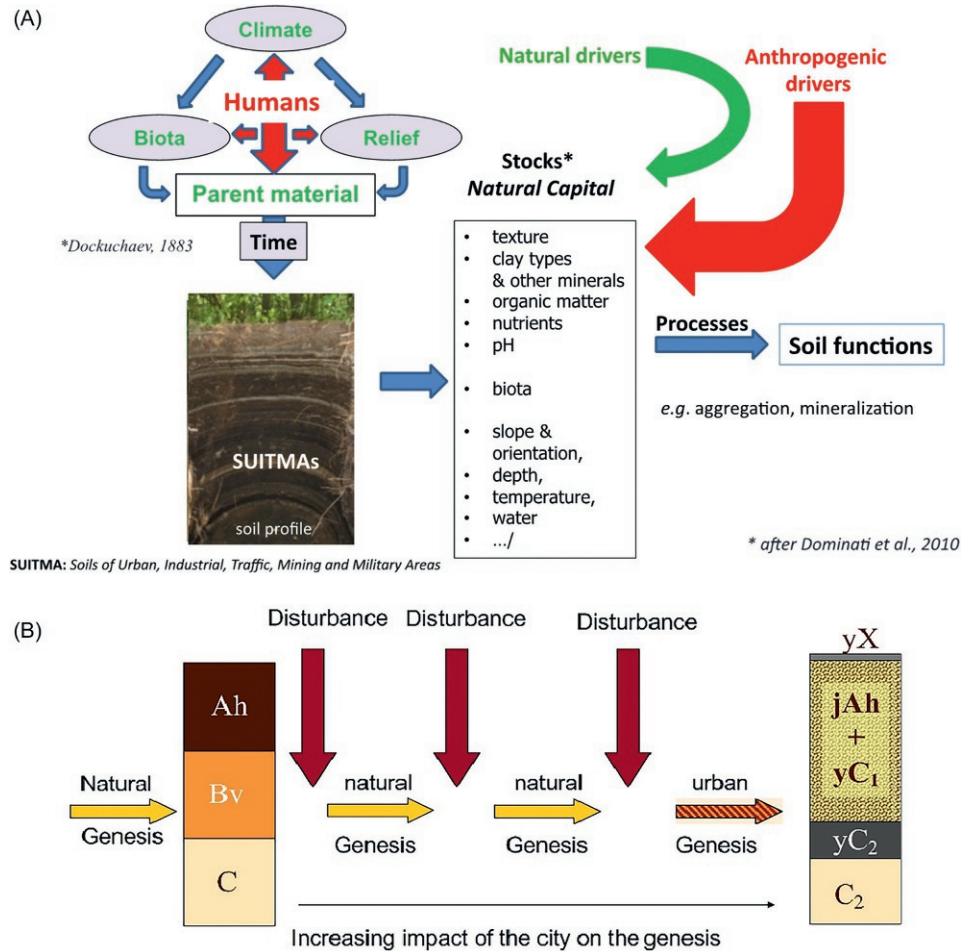


Fig. 3 (A) Soil formation in highly anthropized areas (drawing J.L. Morel). (B) Concept of urban soil genesis depicting the importance of anthropogenic disturbances and their highly dynamic conditions for the existing soil (drawing T. Nehls).

decarbonation or carbonation observed together with andosolization, particularly in the presence of industrial sludge (Huot et al., 2015). Unlike natural soils, SUITMAs are young, poorly evolved soils because the parent material subjected to pedogenesis is mostly recent. As a result, very anthropized soils are characterized by an A/C type profile, which can evolve toward A/B/C soil horizons. In addition, with the presence of finely divided material, i.e. highly reactive material, a rapid differentiation of horizons during the early stages of pedogenesis can be observed.

In the following sections, we summarize the state of knowledge on urban soils, their formation, general characteristics, functions, and role in the urban ecosystem as a provider of ecosystem services, and the means to improve and to restore services from degraded land.

Common characteristics of soils in anthropized areas

Non-sealed soils are characterized by specific characteristics compared to natural soils (Riddle et al., 2022).

Composition

Despite the great diversity of soil types, SUITMAs have common characteristics. They are mainly composed of coarse materials. They can contain many anthropogenic materials, also called 'artifacts,' which are derived from rubble and various wastes (e.g. concrete, asphalt, bricks, glass, ceramics, oxides, plaster, plastic, coal, wood, waste, ballast) and industrial materials (e.g. slag, mine waste, dredged sediments, rock waste). There are also natural coarse materials, such as aggregates, sand and gravel, widely used in construction operations. Among the artifacts, some have properties likely to interfere with the fate of contaminants, metals, organic molecules, pathogens including viruses. An example is brick, which is mainly present in the coarse fraction and has a very dense internal pore system, giving a high water-holding capacity and the possibility for organisms to enter the pores (roots, fungi, bacteria, virus) (Fig. 4). Bricks are widely present in the soils of anthropized areas around the world.

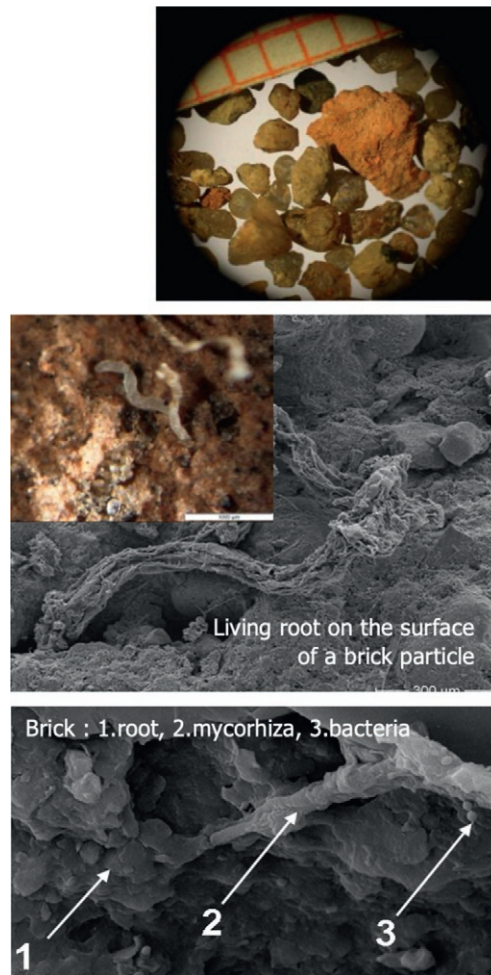


Fig. 4 Biological functions of bricks (Nehls et al., 2013).

The pH

The pH values of soils are, in general, higher in urban areas than in rural areas (Fig. 5). This is due to the presence of large amounts of carbonates in anthropized areas such as cities and traffic areas but also, to a lesser extent, industrial areas. An example is the use of lime for road construction which increases soil pH significantly in acidic areas (e.g. Andosols, Kida and Kawahigashi, 2015). Materials for housing construction are also very rich in carbonates (e.g. concrete). With their high pH values, the physical and chemical properties of urban soils are therefore different from those of natural soils in the same region. Consequently, their functioning, both physically (structure) and chemically (reactivity, availability of elements, fate of charged particles), and biologically (e.g. microbial activity), is very specific. They could be similar to soils developed on natural calcareous materials and function like these soils (e.g. dynamics of nutrients, P, trace elements; organic matter).

Organic matter

The organic matter composition of SUITMAs is highly variable. Some have a large organic matter content (e.g. disposal of combustion residues and organic waste, soils used for horticulture). Black carbon, a very stable organic material produced by incomplete combustion (e.g. heating, fires, vehicle emissions) also accumulates in urban soils, thus contributing to a high percentage of soil organic carbon (e.g. 28%; Nehls and Shaw, 2010). Other soils contain small concentrations of organic matter, for example after the removal of topsoil by civil engineering (e.g. road and building construction; mining). Others may contain a large amount of very stable C (e.g. petroleum hydrocarbon, polycyclic aromatic hydrocarbons, PAHs) which can be found very deep in the profile of industrial soils. Global reviews showed: (i) higher C stocks in urban soils compared to the natural counterparts with an especially high variability in high latitudes (Vasenev and Kuzyakov, 2018); (ii) that Technosols, despite their high variability, are among the soils with the largest organic carbon stocks on Earth (Allory et al., 2021).

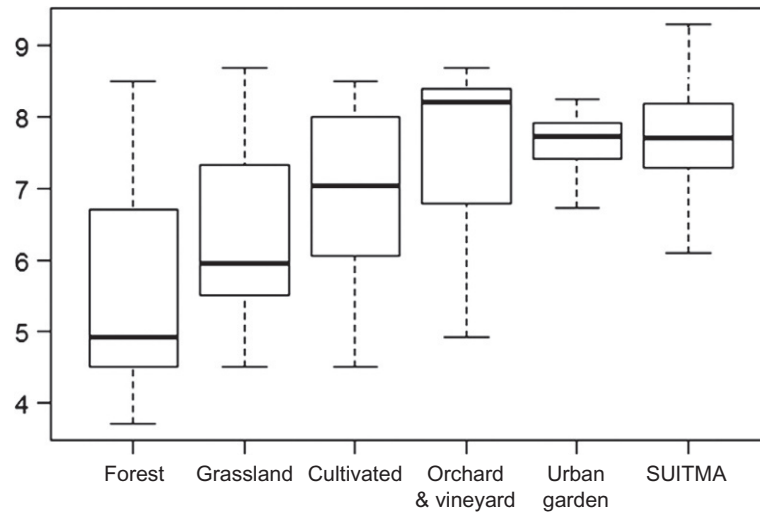


Fig. 5 The pH in French topsoil according to land use ($n = 2450$) (Joimel et al., 2016).

Contaminants

Most SUITMAs contain contaminants, either organic (e.g. hydrocarbons, polychlorobiphenyls PCBs) or inorganic (e.g. metals), or both. They come from all urban and industrial activities (e.g. traffic, wastes). In general, soils of anthropized areas exhibit higher concentrations of contaminants than in other areas (Fig. 6). Gardens are particularly exposed to contaminations, and, on average, there is twice the concentration of metals in garden soils than in agricultural soils (Schwartz et al., 1995). The contaminants react with the solid organo-mineral phase of the soil. This results in a partition between the two solid and liquid phases, which determines their fate. They can be degraded (organic contaminants), immobilized (precipitation, sorption, absorption, complexation), or introduced into the soil solution where they can be transferred to different targets (e.g. organisms, groundwater). Soil composition and properties (e.g. clays, organic matter, pH, redox potential) are key factors that determine the fate of pollutants in soils. In urban soils, because of their high pH, the risk of transfer of 'heavy' metals to targets such as plants is low. On the other hand,

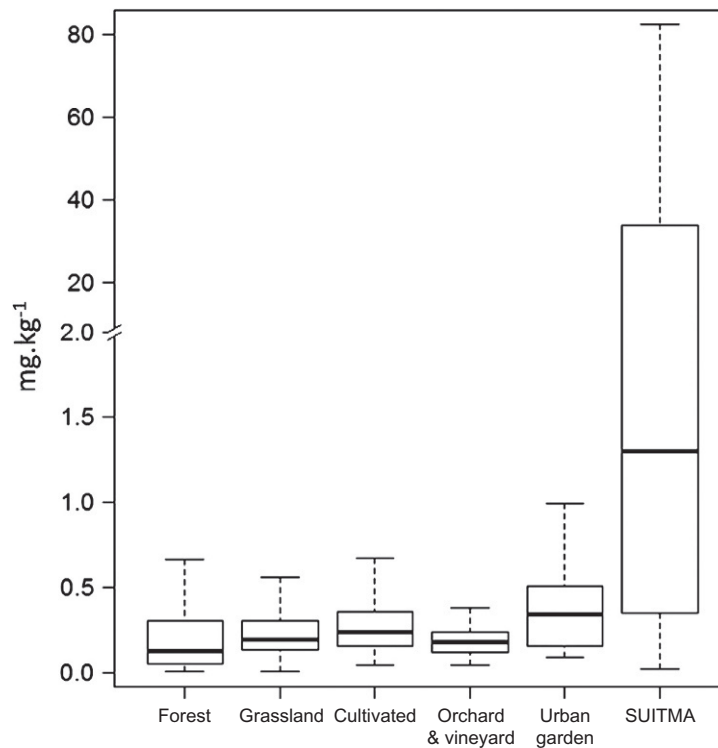


Fig. 6 Cadmium in French topsoil according to the land use (Joimel et al., 2016).

soil particles (e.g. clays, silts) carrying contaminants can be transported in the form of dust and contribute to the exposure of nearby population to toxic compounds or even pathogens that can affect their health. These factors also play similar roles with respect to charged viral particles. Salinization by de-icing reagents is another threat, which urban soils in cold and temperate climates are exposed to. Large concentrations of soluble salts suppress soil microbiota and damage root systems (Gavrichkova et al., 2020).

Diversity of soils in anthropized areas

Anthropization produces a large diversity of soils

SUITMAs cover a wide range of soil types. The main reason is the variety of parent materials from which they derive. These parent materials come from the numerous mixing, compacting, leveling and sealing operations of the initial materials, and from the export of earthy materials and import of exogenous materials (e.g. soils, waste). As a result, soil diversity in heavily anthropized areas is greater than in agricultural areas where operations tend to standardize soil properties (Fig. 7A). Soil formation and evolution in highly anthropized areas remain, however, linked to the climatic conditions as observed along a latitude gradient (Fig. 7B).

The urban soil mosaic

Cities are characterized by the 'urban soil mosaic' (Pouyat et al., 2010), which is clearly represented on soil maps (e.g. New York City; Fig. 8). City soil maps are an important link between soil science and city management, as they provide useful information for decision makers to optimize land use, and provide a safe and good quality environment for city dwellers. For example, they can be used to provide an inventory of soil properties in open spaces to manage municipal water and sewers, control soil pollution, identify wetlands and wildlife habitats, assist in park management, and understand the drainage, runoff and flow of urban streams (Shaw et al., 2007).

Classification

SUITMAs are considered in soil classifications, especially in the World Reference Base for Soil Resources (2022) (WRB). They are classified into three groups of WRB soils: Regosols, Anthrosols and Technosols. The Regosols of SUITMAs are young soils with an A/C type profile and are characterized by urban classifiers (e.g. urbic, spolic, technical, toxic, transportic). Anthrosols are marked by the presence of a large content of organic matter (e.g. horticultural soils) at great depth (e.g. hortic, irrigric, plaggic or terric horizon ≥ 50 cm thick). Organic matter has accumulated over the years by the application of manure, household waste, and charcoal. Old gardens with long and intensive horticultural use are examples of Anthrosols (Fig. 9).

Technosols are soils that contain a high proportion of anthropogenic materials, or artifacts, resulting from human technologies (at least 20% in the first meter of the profile) and/or a continuous impervious layer (e.g. geomembrane). This group was suggested by Lehman (2006) and Rossiter (2007). Constructed soils are classified as Technosols because they contain a large amount of technogenic materials, including transported soil material, wastes, sludges, etc. (Fig. 10). SUITMAs are also recognized by national classifications, German, French, Polish, USA, and Russian classifications, to integrate anthropogenic processes. The International Committee for Anthropogenic soils has defined Human-Altered and Human-Transported soils in Soil Taxonomy (Galbraith, 2018).

Functions of SUITMAs

Like natural soils, SUITMAs fulfill a wide range of functions (reservoir of nutrients, support for roots, reservoir of biodiversity, filtration and transformation of substances, storage, source of raw materials, physical support for human activities, archives). Pseudo-natural SUITMAs (e.g. parks, urban forests) that show little change (e.g. compaction) from the corresponding natural soils, perform most of these functions. But when anthropogenic factors tend to dominate the formation and evolution of soils, these functions can be seriously degraded. Soil sealing is an example of a degraded function that can cause dramatic damage to people and their property (Fig. 11). The case of floods illustrates the dependence of city dwellers on natural capital (stocks). The figure also shows the links between stocks, processes, functions and services and shows that it is possible to give an economic value to stocks in the event of flooding.

Examples of the status of three main soil functions in SUITMAs is presented below.

Support for roots

Root establishment and development is highly dependent on the physical properties of the soil. The presence of coarse materials, rubble, impermeable layers, large voids, contamination, excess water, etc. affect root development in SUITMAs. Again, sealing is a major obstacle for roots and a major disturbance for plants, such as trees. In the city, trees encounter very contrasting physical soil conditions whether they are in parks or have been planted along streets (Fig. 12).

Soil compaction can induce not only plant growth inhibition but also water stagnation and subsequent hydromorphic conditions. Soil porosity is generally significantly affected by the anthropization gradient (Fig. 13).

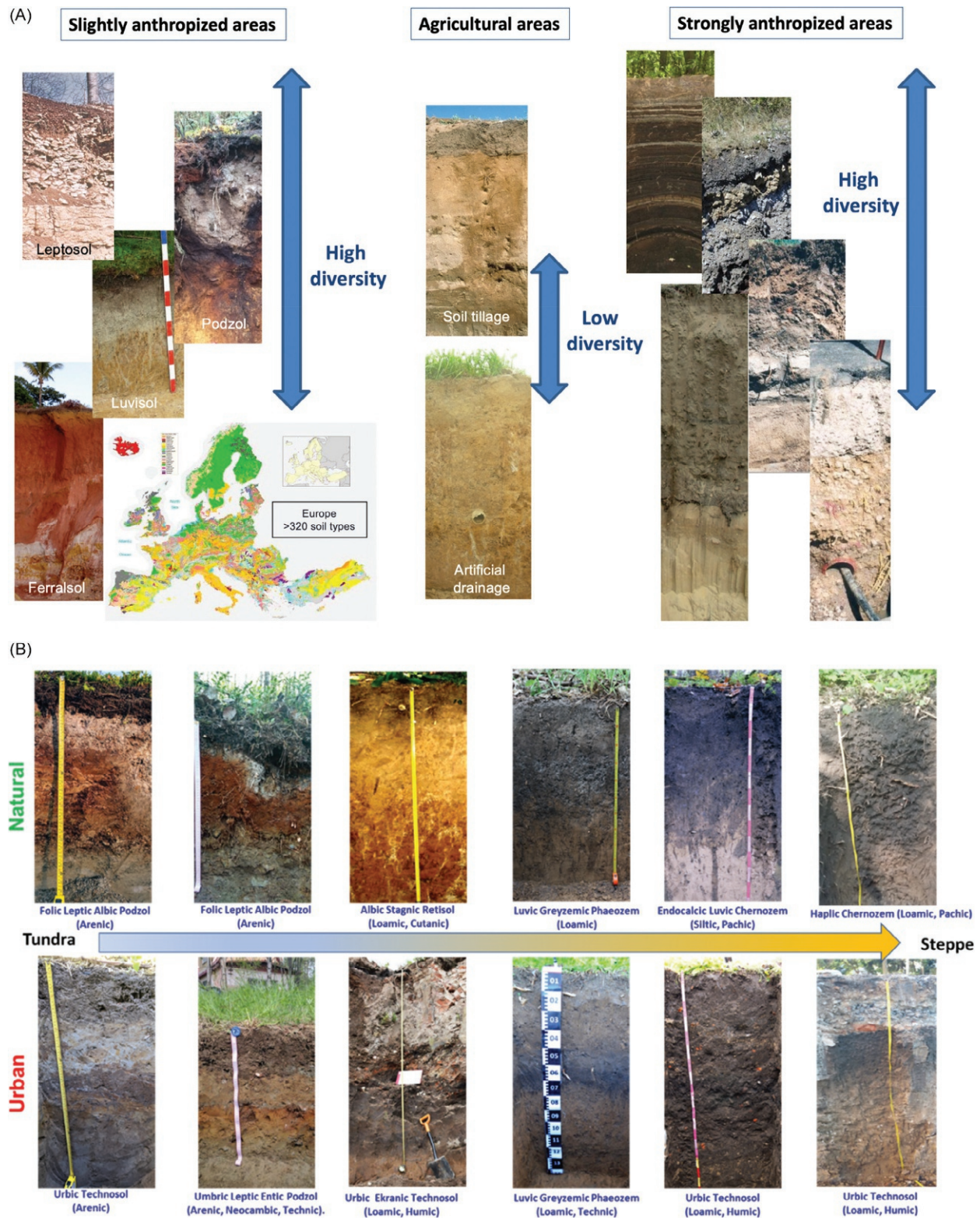


Fig. 7 Anthropization produces a large diversity of soils – (A) comparison between natural, agricultural and anthropized areas; (B) sequence of natural and urban soils according to the latitudinal gradient in Russia.

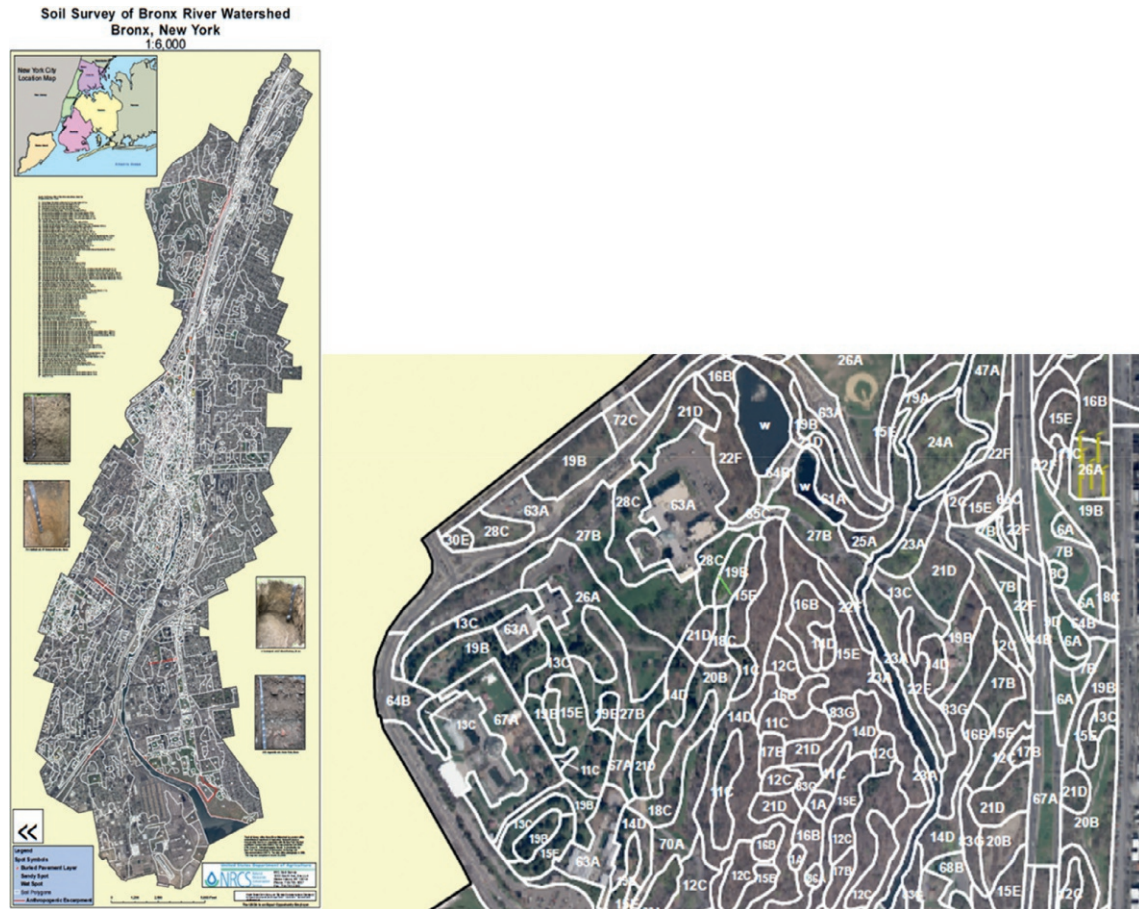


Fig. 8 The 'urban soil mosaic' as defined by Pouyat et al. (2010) (Soil map of New York City; New York City Soil Survey Staff; Shaw et al., 2007).



Fig. 9 Example of Anthrosol in an old garden (Tolstoi House, SUITMA Conference tour, 2017, Moscow, Russia).

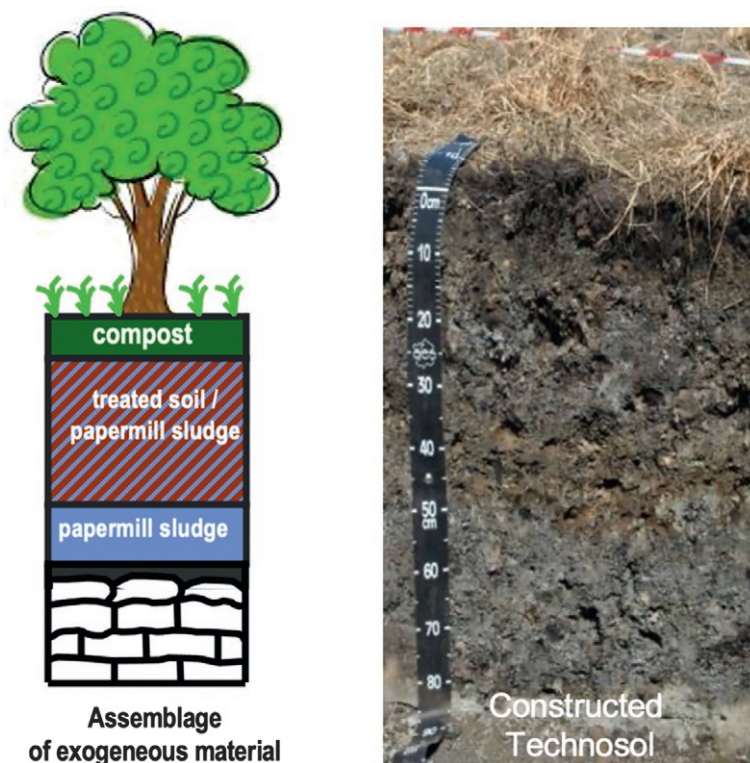


Fig. 10 Constructed Technosol to restore soil functions after industrial activities (Séré et al., 2010).

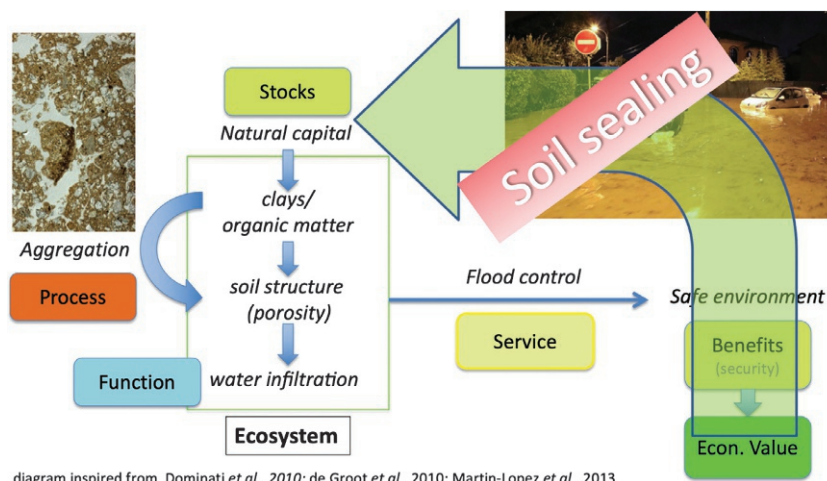


diagram inspired from Dominati et al., 2010; de Groot et al., 2010; Martin-Lopez et al., 2013

Fig. 11 Stocks, processes, functions, services, benefits, and value of the soil natural capital as affected by soil sealing (drawing J.L. Morel).



Fig. 12 Contrasting conditions for trees in cities.

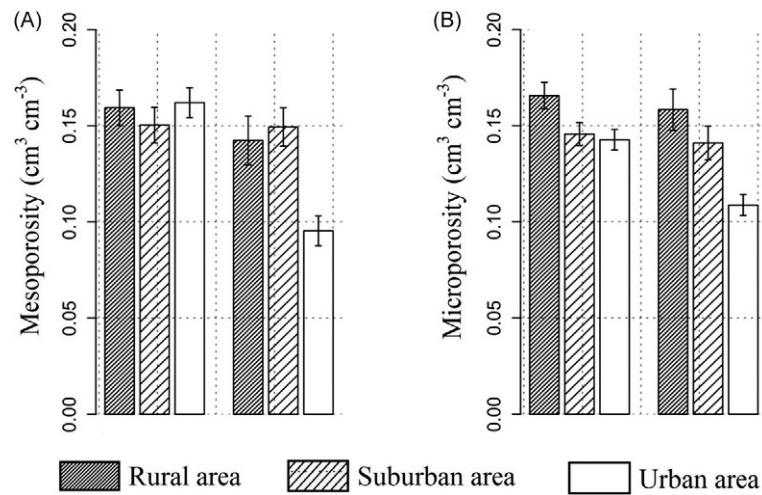


Fig. 13 Porosity along an anthropization gradient (Paris) (A) lawns, (B) woods (Foti et al., 2020).

Reservoir of biodiversity

SUITMAs offer a wide variety of habitats for biodiversity. All groups of soil organisms are present in urban soils, i.e. earthworms, arthropods, nematodes, microorganisms, living under the specific environmental conditions of the urban environment (e.g. temperature, available water, compaction, contamination) (Guilland et al., 2018). However, they are notably different in composition, morphology, and functioning according to the gradient of anthropization from forest areas to urban areas (Fig. 14).

Soil degradation induced by former industrial activities includes contamination, presence of artifacts, small organic matter and nutrient content, low available water, etc. Such conditions are in general not favorable to organisms. From a survey of a set of brownfields, it was observed that the taxonomic microbial diversity was not significantly affected by the gradient of metal pollution (Lemmel et al., 2019). However, the metabolic functions showed a decrease along the gradient. Therefore, the biodiversity in soils of heavily anthropized areas exhibits specific features which open important perspectives regarding the role of soil organisms in the functioning of urban soils and hence of the urban ecosystem.

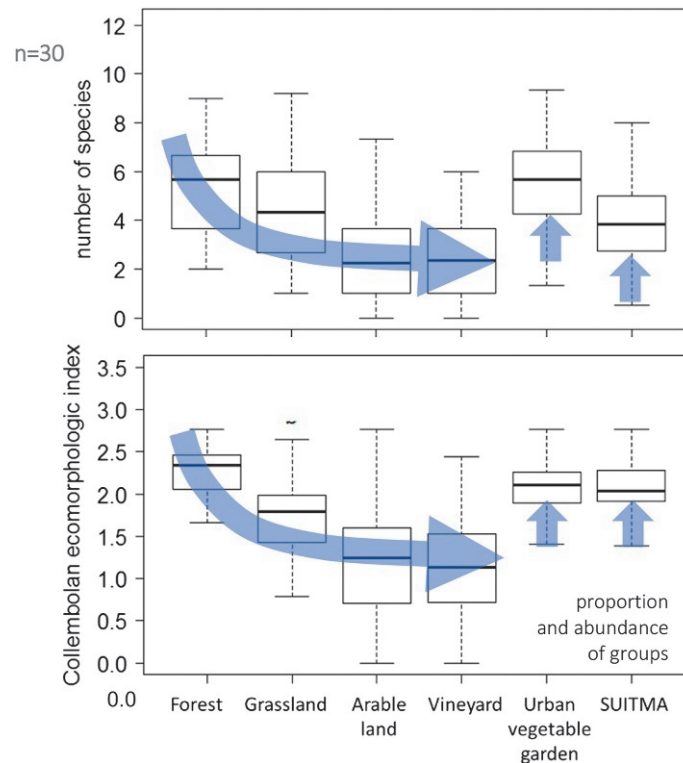


Fig. 14 Soil macrofauna along a gradient of anthropization (from forest to urban areas) (Joimel et al., 2016).

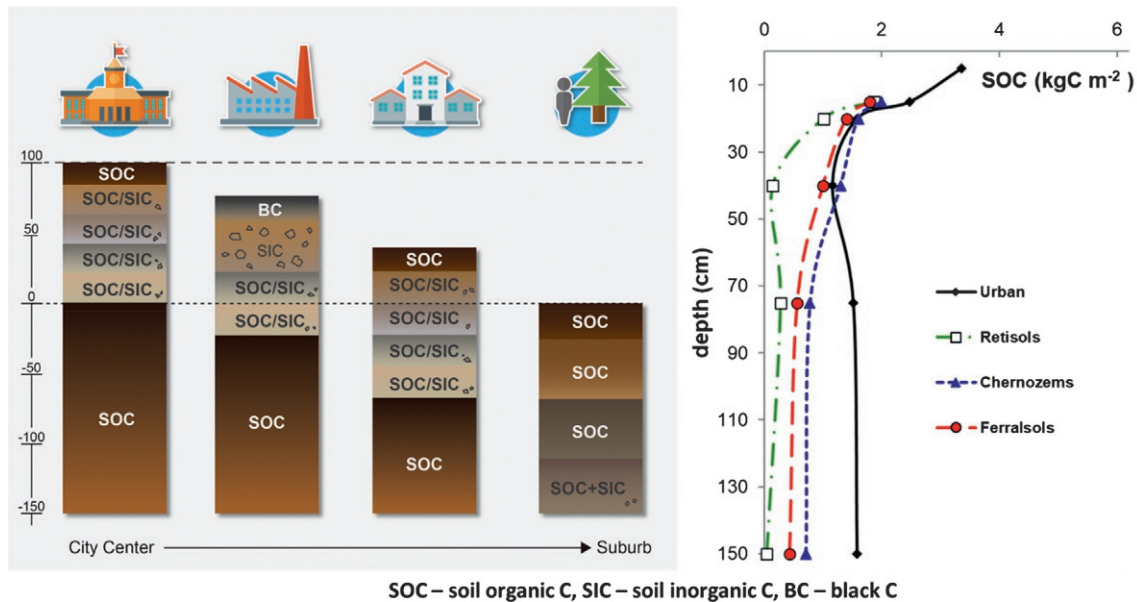


Fig. 15 Accumulation of carbon in soils of urban areas (Vasenev and Kuzyakov, 2018).

Carbon storage

Carbon accumulates in urban soils; the inputs originate from suburban areas (food, wood, raw materials, peat material) and bring large amounts of organic compounds into the city. Also, other forms of C other than organic matter contribute to the accumulation of C in urban soils from waste (plastic, rubber, soot, charcoal) and from construction materials (e.g. concrete). Over the years, this process leads to a significant increase of C in soil which was estimated to reach 20–30 kg m² per century (Fig. 15; Vasenev and Kuzyakov, 2018). Therefore, SUITMAs represent a potential carbon sink which could be relevant to contribute to C storage to meet the objectives of C neutrality by 2050, considering the huge surfaces of anthropized areas (e.g. parks and gardens, urban agriculture, green roofs and walls, roads, railways, airports, commercial areas, brownfields – mines and industry, military areas; <https://www.allianceenergie.fr/>).

Estimates of C in urban soils have been made showing a range of 6–10 kg C m⁻² in the first 30 cm of the soil profile (Pouyat et al., 2006). Forest land contains about 15 kg m⁻², and crop land less than 8 kg m⁻². The unsealed areas reached more than 18 kg C m⁻² while the sealed soils contained only 6 kg C m⁻². Similar figures were obtained by Cambou et al. (2018) for the city of Paris. At the same time, the resistance of these stocks to biodegradation can be rather low, especially under the urban heat island effect. For example, urban lawns on Technosols constructed from peat–sand mixtures in Moscow city were C sources, especially in the central areas where the heat island effect was the strongest (Vasenev et al., 2021). In addition, due to the presence of carbonated material and incorporation of charcoal, inorganic C in anthropized areas represents a significant contribution to C storage (Vasenev and Kuzyakov, 2018).

Based on a meta-analysis of carbon stocks in Technosols, two major origins of these anthropized soils have been highlighted: (i) the organic carbon stocks are very different from one Technosol to another, with an unparalleled variability compared to other natural soils; (ii) the distribution of organic carbon stocks over depth does not respond to the usual decrease with depth because many anthropized soils have a large organic carbon content in the deeper horizons (Allory et al., 2021) (Fig. 16).

Ecosystem services provided by SUITMAs

Like natural soils, SUITMAs provide a wide range of ecosystem services. In densely populated areas, SUITMAs are important for urban agriculture (e.g. horticulture, gardening, green roofs and green walls, parks and street trees), essential in the water cycle (e.g. urban hydrology, infiltration, water transport), sequestration of pollutants, in air quality regulation (e.g. PM, NO_x), climate regulation (emission of greenhouse gases, heat-island effect). They play a key role in supporting biodiversity in anthropized areas because they host plants and ecosystems (fauna, flora, microflora).

The potential to provide a wide range of ecosystem services, each at a high value, decreases with human pressure (Fig. 17). There has been a proposal to sort the different types of SUITMAs according to their ability to provide ecosystem services (Fig. 18; Morel et al., 2015). Four groups have been proposed, including pseudo-natural SUITMAs, represented by only slightly altered soils found in urban forests, constructed SUITMAs, including garden soils, constructed Technosols, green roofs, and SUITMAs in landfills, either for the disposal of municipal waste or industrial waste, finally the sealed SUITMAs, including paved soils and completely sealed soils.

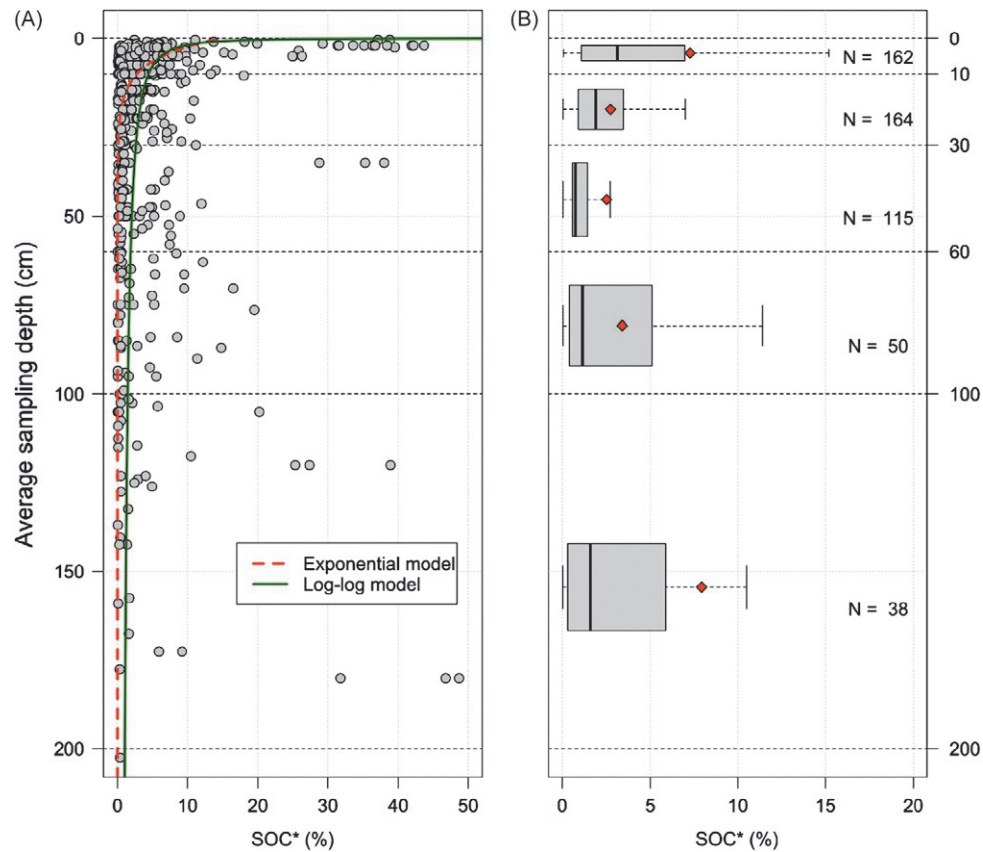


Fig. 16 Distribution of soil organic carbon content with depth in Technosols and fitted models and boxplots of SOC* content for 0–10 cm, 10–30 cm, 30–60 cm, 60–100 cm and 100–200 cm of average sampling depth (Allory et al., 2021).

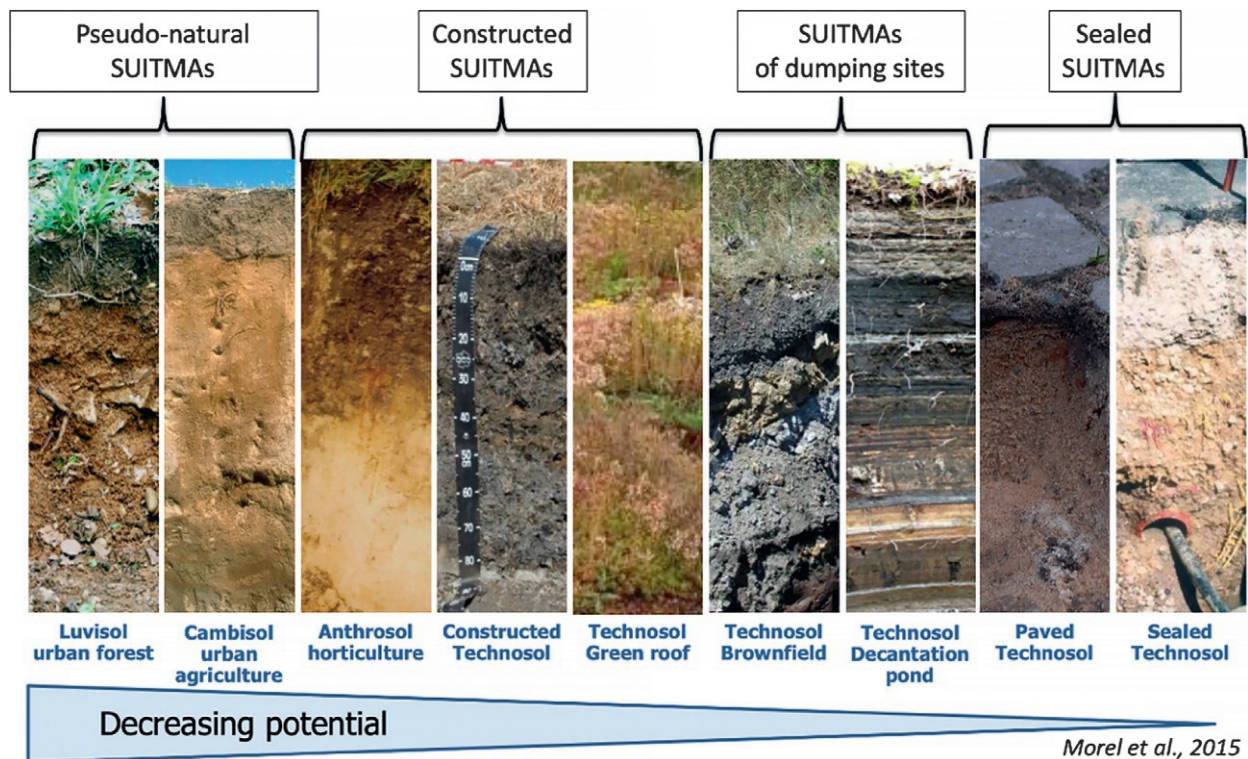


Fig. 17 Groups of SUITMAs according to their potential to provide ecosystem services (Morel et al., 2015).

Ecosystem services O, +, ++, +++		groups of SUITMAs			
		Pseudo-natural	Constructed	Dumping sites	Sealed
Provisioning	Food production	++	++	O	O
	Non-food biomass (energy, fiber, wood)	+++	++	++	O
	Reservoir of minerals	+	+	+++	O
	Fresh water supply	O	+	O	+++
Regulating	Run-off and flood control	+++	+++	++	+
	Control of erosion	++	+	++	O
	Pollution attenuation	++	+++	++	O
	Waste recycling	+	+++	+++	O
	Global climate regulation	+++	++	++	+
	Local climate regulation	+++	++	+	O
	Air purification	+++	++	+	O
	Noise attenuation	++	+++	++	+
Cultural	Recreation (parks, gardening, tourism)	+++	++	O	O
	Archives of human history	+	+	+++	++
	Landscape	++	+++	+	+
	Education	+++	+++	++	+
Ecosystem functioning	Reservoir of biodiversity	+++	+++	++	O
	Support for vegetation	+++	+++	++	O

Fig. 18 Estimated value of ecosystem services provided by groups of SUITMAs (Morel et al., 2015).

Examples of two services provided by contrasting groups of SUITMAs, each resulting from deliberate construction are: (i) green roofs and (ii) sealed soils.

Green roofs and green walls are completely artificial soils. They are designed to fulfill several functions, and as such they are providers of a large range of services. Regulating services include mitigation of water flow, notably flood control, filtration of water, decrease of urban temperature through evapotranspiration, and thermal insulation of buildings, improvement of air quality through the capture of pollutants such as PM and NO₂. They also provide habitats for biodiversity, not only for plants but also for arthropods, birds, etc. Green roofs can also contribute to urban agriculture and provide provisioning services such as food (roof gardens). Green roofs contribute also to cultural (gardening and related social activities) and esthetic services. Attention should be paid to the quality of the materials that are used for the construction of green roofs, to avoid any leaching and dissemination of trace elements.

Sealed soils are designed to provide several ecosystem services to support the well-being of city dwellers through the supply of water and energy (hosts of pipes and wires) and support safe and clean transportation and habitat activities. They are also designed to manage rainwater and ensure a rapid disposal of runoff water. Soils in cities are also shaped to provide a smooth and attractive landscape, either sealed or vegetated soils, highlighting architectural heritage and pseudo-natural spaces. Overall, in all countries, soil sealing has been and remains mainly seen as a step toward a better life in the city. However, as mentioned earlier, sealed soils can also be detrimental. Among the issues are increasing risks of flooding resulting from loss of the infiltration function in the event of heavy rain, the loss of agricultural and forest land due to the conversion of natural soils into buildings, roads, parking areas, commercial centers, airports, etc. Also, the heat storage function is exacerbated and reinforces the heat island effect. Regarding plants, sealed soils offer a poor habitat for roots, and biodiversity is rather low. Sealed soil covered by dead plant leaves represents a threat for pedestrians during the rainy days of autumn. There are many questions regarding the functioning of soil in the deeper horizons, including water infiltration, development of roots, aeration, dynamics of elements and carbon, etc.

Restoration of soil functions to improve ecosystem services

Soil science knowledge combined with technical equipment has made it possible to develop new ways to restore soil functions of degraded areas. This dynamic was triggered by the considerable shift of the economy of many industrialized countries during the period of the 1970s–1990s, leading to huge areas of industrial wasteland resulting from industrial, mining, and military activities. Together with the conversion to new economic and residential activities, large areas have been left to evolve spontaneously. The strategy is based on nature-based solutions, in particular soil construction and agroforestry ecosystems (Fig. 19).

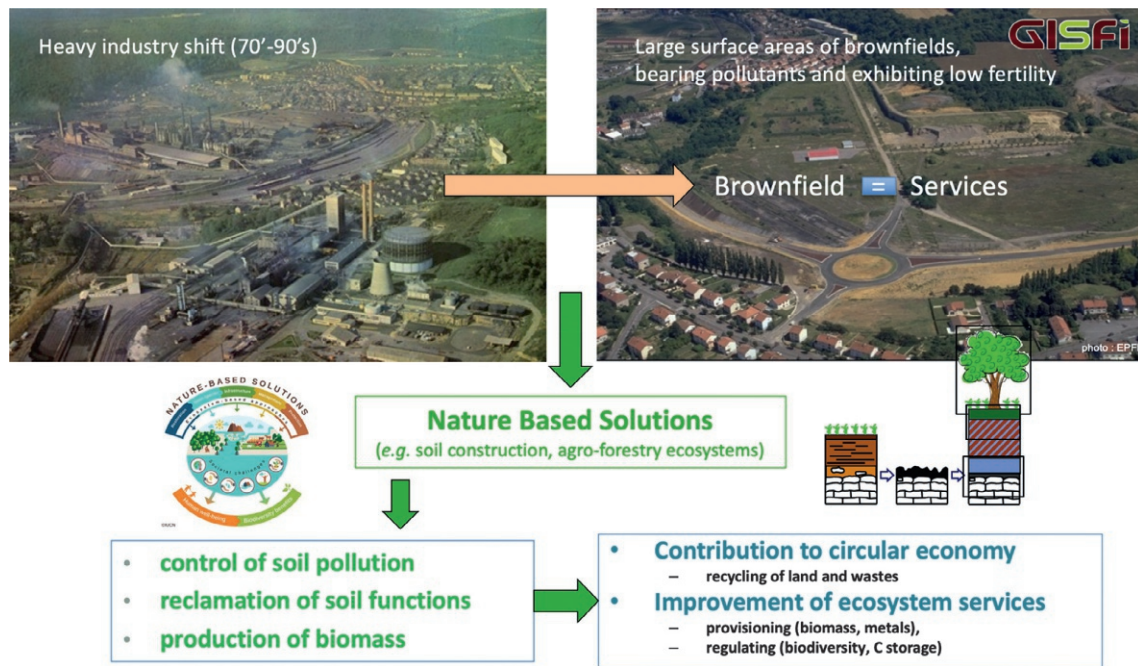


Fig. 19 From an unused degraded site to a valuable resource (drawing J.L. Morel).

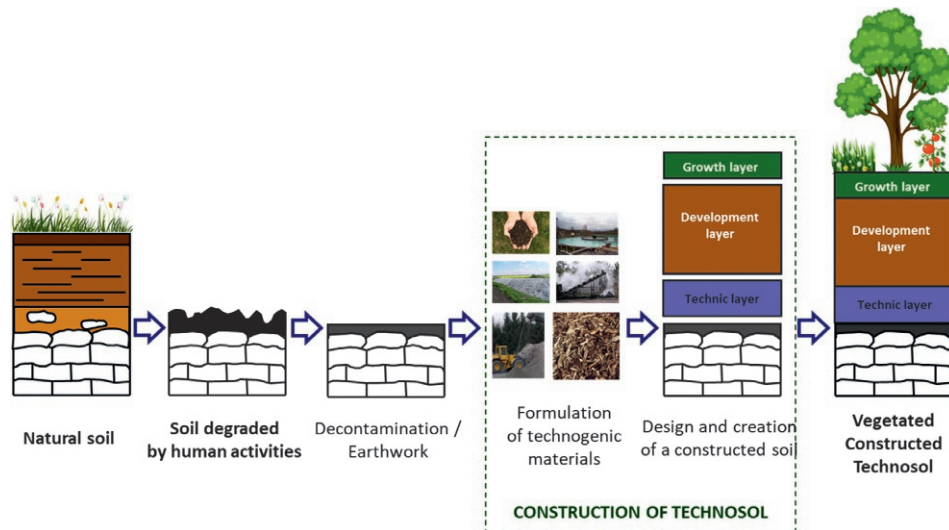


Fig. 20 Soil construction principle (Séré et al., 2010).

Constructed Technosols are fully artificial soils that have been created to reclaim degraded sites (e.g. brownfields, mining sites). Such an operation creates new soil profiles whose properties and functioning are close or even analogous to those of natural soils (Séré et al., 2010; Fig. 20).

Soil functions are restored and enable the development of new ecosystems. Beforehand, such soil profiles were specifically designed to fulfill a set of functions that provided many ecosystem services from industrial and mining sites, initially poor in organic matter and nutrients, salinized, containing pollutants and artifacts, compacted, therefore unfavorable to root development, and more generally to life. Soil construction should be based on the use of secondary materials (e.g. urban and industrial wastes, treated earth), for which the content of contaminants (e.g. trace elements, organic compounds, pathogens) should be assessed before construction. This strategy makes it possible to reduce the excavation and transport of clean soil material collected from natural soils. Using secondary materials thus preserves agricultural and forest soils.

Constructed soil provides several ecosystem services, including provisioning (biomass), regulation (biodiversity, filtering and exchange, C storage, use of anthropogenic material to protect soil natural capital, local climate). A constructed Technosol implemented on a former coking plant (Gisfi, Homécourt, NE France) has demonstrated the capacity to provide several ecosystem

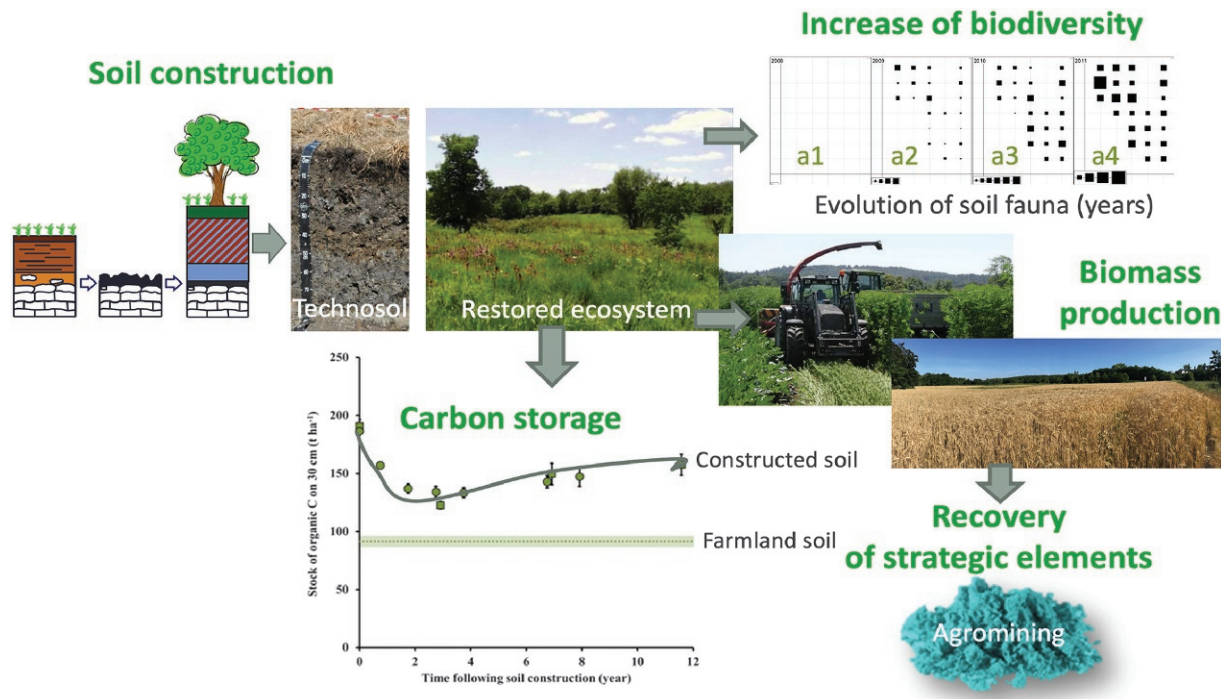


Fig. 21 Ecosystem services provided by a former brownfield site after soil construction (Lorver program, Gisfi, 2018; <http://lorver.org/accueil/index.html> (drawing J.L. Morel).

services in a similar or even better way than natural soils (Fig. 21) (Hedde et al., 2019; Rees et al., 2019). For the food supply service, fodder yields were like those rendered by neighboring agricultural soils. Woody species have been installed, contributing to the supply of biomass for industrial use (e.g. energy, fiber). The recovery of metals (e.g. nickel) can also be considered by agromining if the soil contains the element of interest. Construction of the soil, then development of the vegetation contributed to a significant carbon supplement, which reached 520 t C ha⁻¹ after 3 years, four times more than the analogous natural surrounding soils. The restored ecosystem has allowed the installation and development of many species (e.g. collembolas, nematodes, earthworms), with different kinetics depending on the soil organisms considered. Diversity and abundance were like those of referenced natural grasslands.

Conclusion

Anthropization generates a great diversity of soils with specific characteristics, pH, coarse texture, compaction, sealing, pollutants, etc. The multiplicity of human actions over time results in a mosaic of soils with differing characteristics at short distances. Until recently, the soils of highly anthropized areas (SUITMA) have been dominantly perceived as degraded and contaminated soils if not non-soils. However, they can fulfill a large array of functions of great importance in the urban ecosystem. These functions are the basis of several ecosystem services, among them local and global climate regulation (mitigation of the heat island effect, C storage), biodiversity habitat induced by the variety of soil types in the anthropized areas, and regulation of the water cycle.

Soil functions in anthropized areas, however, are frequently affected by the physical and chemical stresses that occur in densely populated areas (e.g. compaction, sealing, contamination). They have been disregarded until very recently for their role as a three-dimensional reactor in the urban ecosystem; these soils now represent an asset for city managers to improve the quality of life of city dwellers. With the growing interest in 'urban soils' shown by pedology, new knowledge is being acquired. Soil science devoted to urban areas is quite different from that of natural and agricultural areas. They differ in at least two ways, the first is a soil science oriented toward the prediction of soil evolution, i.e. the future of soils instead of the past, and the second is a practical aspect, i.e., soil engineering specifically designed to perform given functions to meet the demand of cities for ecosystem services.

References

- Allory V, Séré G, and Ouvrard S (2021) A meta-analysis of carbon content and stocks in Technosols and identification of the main governing factors. *European Journal of Soil Science*. <https://doi.org/10.1111/ejss.13141>.
- Bastie J (1965) The soil, a primordial component of the urban. *Annales de Géographie* 74: 708–713.

- Blume H-P (1992) *Handbuch des Bodenschutzes: Bodenökologie und -belastung – vorbeugende und abwehrende Schutzmaßnahmen*. 2. Aufl., 794 S. Landsberg/Lech: ecomed Verlagsgesellschaft.
- Blume H-P and Runge M (1978) Genese und Ökologie innerstädtischer Böden aus Bauschutt. *Zeitschrift für Pflanzenernährung und Bodenkunde* 141: 727–740.
- Burghardt W (1994) Soils in urban and industrial environments. *Zeitschrift für Pflanzenernährung und Bodenkunde* 157: 205–214.
- Burghardt W, Morel JL, and Zhang GL (2015) Development of the soil research about urban, industrial, traffic, mining, and military areas (SUITMA). *Soil Science and Plant Nutrition* 61(suppl 1): 3–21.
- Cambou A, Shaw RK, Huot H, Vidal-Beaudet L, Hunault G, Cannavo P, Nold F, and Schwartz C (2018) Estimation of soil organic carbon stocks of two cities, New York City and Paris. *Science of the Total Environment* 644: 452–464.
- de Kimpe CR and Morel JL (2000) Urban soil management: A growing concern. *Soil Science* 165: 31–40.
- Deeb M, Groffman PM, Blouin M, Egendorf SP, Vergnes A, Vasenev V, Cao DL, Morin T, and Séré G (2020) Using constructed soils for green infrastructure –challenges and limitations. *Soil, European Geoscience Union*. <https://doi.org/10.5194/soil-6-413-2020>.
- Foti L, Barot S, Gignoux J, Grimaldi M, Lata J-C, Lerch TZ, Nold F, Nunan N, Raynaud X, Abbadié L, and Dubs F (2020) Topsoil characteristics of forests and lawns along an urban–rural gradient in the Paris region (France). *Soil Use and Management* 37: 749–761. <https://doi.org/10.1111/sum.12640>.
- Galbraith JM (2018) Human-altered and human-transported (HAHT) soils in the U.S. soil classification system. *Soil Science and Plant Nutrition* 64: 190–199. <https://doi.org/10.1080/00380768.2018.1442682>.
- Gavrichkova O, Brykova RA, Brugnoli E, Calfapietra C, Cheng Z, Kuzyakov Y, Liberati D, Moscatelli MC, Pallozzi E, and Vasenev VI (2020) Secondary soil salinization in urban lawns: Microbial functioning, vegetation state, and implications for carbon balance. *Land Degradation and Development* 31: 2591–2604. <https://doi.org/10.1002/ldr.3627>.
- Gisfi (2018) *Groupeement d'Intérêt Scientifique sur les Friches Industrielles*. <http://gisfi.univ-lorraine.fr>.
- Guilland C, Maron PA, Damas O, and Ranjard L (2018) Biodiversity of urban soils for sustainable cities. Review. *Environmental Chemistry Letters*. <https://doi.org/10.1007/s10311-018-0751-6>.
- Guo M, Zhong X, Liu W, Wang G, Chao Y-Q, Huot H, Qiu R, Morel JL, Watteau F, Séré G, and Tang Y (2022) Biogeochemical dynamics of nutrients and rare earth elements (REEs) during natural succession from biocrusts to pioneer plants in REE mine tailings in Southern China. *Science of the Total Environment* 828: 154361. <https://doi.org/10.1016/j.scitotenv.2022.154361>.
- Hedde M, Nahmani J, Séré G, Auclerc A, and Cortet J (2019) Early colonization of constructed technosols by macro-invertebrates. *Journal of Soils and Sediments* 19: 3193–3203. <https://doi.org/10.1007/s11368-018-2142-9>.
- Huot H and Morel JL (2023) *Encyclopedia of Soil Science*. This issue- to be later completed.
- Huot H, Simonnot M-O, and Morel JL (2015) Pedogenetic trends in soils formed in technogenic parent materials. *Soil Science* 180: 182–192.
- Joimel S, Cortet J, Jolivet CC, Saby NPA, Chenot ED, Branchu P, Consalès JN, Lefort C, Morel JL, and Schwartz C (2016) Physico-chemical characteristics of topsoil for contrasted forest, agricultural, urban, and industrial land uses in France. *Science of the Total Environment* 545–546: 40–47.
- Kida K and Kawahigashi M (2015) Influence of asphalt pavement construction processes on urban soil formation in Tokyo. *Soil Science and Plant Nutrition* 61: 135–146.
- Lehman A (2006) Technosols and other proposals on urban soils for the WRB (World Reference Base for Soil Resources). *International Agrophysics* 20: 129–134.
- Lemmel F, Maunoury-Danger FF, Fanesi A, Leyval C, and Cebon A (2019) Soil properties and multi-pollution affect taxonomic and functional bacterial diversity in a range of French soils displaying an anthropisation gradient. *Microbial Ecology* 7: 993–1013. <https://doi.org/10.1007/s00248-018-1297-7>. Springer Verlag.
- Levin MJ, Kim K-HJ, Morel JL, Burghardt W, Charzynski P, and Shaw RK (2017) *Edited on behalf of IUSS, Working Group SUITMA: Soils within Cities: Global approaches to their sustainable management*. Stuttgart: Suitma group, Catena-Schweizerbart. 253 p.
- Mazurek R, Kowalska J, Gasiorok M, and Setlak M (2016) Micromorphological and physico-chemical analyses of cultural layers in the urban soil of a medieval city - A case study from Krakow. *Poland Catena* 141: 73–84. <https://doi.org/10.1016/j.catena.2016.02.026>.
- MEA (2005) *Millennium Ecosystem Assessment*. <https://www.millenniumassessment.org/en/index.html>.
- Morel JL, Schwartz C, Florentin L, and de Kimpe C (2005) Urban soils. In: *Encyclopedia of Soils in the Environment*, 202–208.
- Morel JL, Chenu C, and Lorenz K (2015) Ecosystem services provided by soils of urban, industrial, traffic, mining, and military areas (SUITMAs). *Journal of Soils and Sediments* 15: 1659–1666. <https://doi.org/10.1016/B0-12-348530-4/00305-2>.
- Nehls T and Shaw RK (2010) Black carbon in soils: Relevance, analysis, distribution. *Soil Survey Horizons, Soil Science Society of America* 51: 79–84. <https://doi.org/10.2136/sh2010.3.0079>.
- Nehls T, Rokia S, Mekiffer B, Schwartz C, and Wessolek G (2013) Contribution of bricks to urban soil properties. *Journal of Soils and Sediments* 13: 575–584. <https://doi.org/10.1007/s11368-012-0559-0>.
- Poggio L, De Sousa LM, Batjes NH, Heuvelink GBM, Kempen B, Ribeiro E, and Rossiter D (2021) SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *The Soil* 7: 217–240. <https://doi.org/10.5194/soil-7-217-2021>.
- Pouyat RV, Yesilonis ID, and Nowak DJ (2006) Carbon storage by urban soils in the United States. *Journal of Environmental Quality* 35: 1566–1575.
- Pouyat RV, Szlavetz K, Yesilonis ID, Groffman PM, and Schwarz K (2010) In: Aitkenhead-Peterson J and Volder A (eds.) *Chemical, Physical, and Biological Characteristics of Urban Soils, Agronomy Monograph* 55. *Urban Ecosystem Ecology* 119–152. <https://doi.org/10.2134/agronmonogr55.c7>.
- Purves D (1966) Contamination of urban garden soils with copper and boron. *Nature* 210(5040): 1077.
- Rawtani D, Gupta A, Khatri N, Rao PK, and Hussain CM (2022) Environmental damages due to war in Ukraine: A perspective. *Science of the Total Environment* 850: 157932. <https://doi.org/10.1016/j.scitotenv.2022.157932>.
- Rees F, Dagois R, Derrien D, Fiorelli JL, Watteau F, Morel JL, Schwartz C, Simonnot MO, and Séré G (2019) Carbon storage in constructed technosols: In situ monitoring over a decade. *Geoderma* 337: 641–648. <https://doi.org/10.1016/j.geoderma.2018.10.009>.
- Riddle RL, Siebecke MG, Weindorf DC, Shaw RK, and Scharenbroch BC (2022) Chapter Four - Soils in urban and built environments: Pedogenic processes, characteristics, mapping, and classification. *Advances in Agronomy* 173: 227–255. <https://doi.org/10.1016/bs.agron.2022.02.004>.
- Rossiter DG (2007) Classification of urban and industrial soils in the world reference base for soils resources. *Journal of Soils and Sediments* 7: 96–100.
- Schwartz C, Fetzner KD, and Morel JL (1995) Factors of contamination of garden soils by heavy metals. In: Prost R (ed.) *CD-Rom, Contaminated Soils, Third International Conference on the Biogeochemistry of Trace Elements, Paris*.
- Séré G, Schwartz C, Ouvrard S, Renat JC, Watteau F, Villemin G, and Morel JL (2010) Early pedogenic evolution of constructed Technosols. *Journal of Soils and Sediments* 10: 1246–1254.
- Shaw RK, Reinhardt L, and Isleib J (2007) *Soil survey of Bronx River watershed, Bronx, NY*. Natural Resources Conservation Service: US. – Department of Agriculture. <http://www.soilandwater.nyc/urban-soils.html>.
- Slukovskaya MV, Vasenev VI, Ivashchenko KV, Morev DV, Drogobuzhskaya SV, Ivanova LA, and Kremenetskaya IP (2019) Technosols on mining wastes in the subarctic: Efficiency of remediation under Cu-Ni atmospheric pollution. *International Soil and Water Conservation Research* 7: 297–307. <https://doi.org/10.1016/j.iswcr.2019.04.002>.
- Vasenev V and Kuzyakov Y (2018) Urban soils as hot spots of anthropogenic carbon accumulation: Review of stocks, mechanisms and driving factors. *Land Degradation and Development* 29: 1607–1622. <https://doi.org/10.1002/ldr.2944>.
- Vasenev V, Varentsov M, Konstantinov P, Romzaykina O, Kanareykina I, Dvornikov Y, and Manukyan V (2021) Projecting urban heat island effect on the spatial-temporal variation of microbial respiration in urban soils of Moscow megalopolis. *Science of the Total Environment* 786. <https://doi.org/10.1016/j.scitotenv.2021.147457>. paper № 147457.

Further reading

Introduction

- IUSS Working Group WRB (2014) *World reference base for soil resources 2014. International soil classification system for naming soils and creating legends for soil maps. (No. 106), World Soil Resources Reports*. Rome: FAO.
- Stoorvogel JJ and Mulder VL (2021) *A comparison, validation, and evaluation of the s-world global soil property database Land*. 10, paper № 544 <https://doi.org/10.3390/land10050544>.
- World Population Prospects (2019) *Highlights United Nations*. Department of Economic and Social Affairs, Population Division. https://population.un.org/wpp/Publications/Files/wpp2019_10KeyFindings.pdf.

Definitions

- Galagoda RU, Jayasinghe GY, Halwatura RU, and Rupasinghe HT (2018) The impact of urban green infrastructure as a sustainable approach towards tropical micro-climatic changes and human thermal comfort. *Urban Forestry & Urban Greening* 34: 1–9. <https://doi.org/10.1016/j.ufug.2018.05.008>.
- Libessart G, Franck-Néel C, Branchu P, and Schwartz C (2022) The human factor of pedogenesis described by historical trajectories of land use: The case of Paris. *Landscape and Urban Planning* 222: 104393. <https://doi.org/10.1016/j.landurbplan.2022.104393>.
- Slukovskaya MV, Vasenev VI, Ivashchenko KV, Dolgikh AV, Novikov AI, Kremenetskaya IP, Ivanova LA, and Gubin SV (2021) Organic matter accumulation by alkaline-constructed soils in heavily metal-polluted area of Subarctic zone. *Journal of Soils and Sediments* 21: 2071–2088. <https://doi.org/10.1007/s11368-020-02666-4>.
- Smagin AV, Dolgikh AV, and Karelin DV (2016) Experimental studies and physically substantiated model of carbon dioxide emission from the exposed cultural layer of Velikii Novgorod Eurasian. *Soil Science* 49: 450–456. <https://doi.org/10.1134/S1064229316040116>.
- USDA online https://www.nrcs.usda.gov/wps/portal/nrcs/detail?cid=nrcs142p2_053986

Soil formation in highly anthropized areas

- Huot H, Cortet J, Watteau F, Milano V, Nahmani J, Sirguez C, Schwartz C, and Morel JL (2019) Diversity and activity of soil fauna in an industrial settling pond managed by natural attenuation. *Applied Soil Ecology* 132: 34–44.
- Huot H, Simonnot M-O, Marion P, Yvon J, de Donato D, and Morel JL (2013) Characteristics and potential pedogenetic processes of a Technosol developing on iron industry deposits. *Journal of Soils and Sediments* 13: 555–568.

Functions of anthropized soils

- O'Riordan R, Davies J, Stevens C, and Quinton JN (2021) The effects of sealing on urban soil carbon and nutrients. *Soil, European Geoscience Union* 7: 661–675. <https://doi.org/10.5194/soil-7-661-2021>.
- Shaw RK, Hernandez LA, Levin MJ, and Muniz E (2017) Promoting soil science in the urban environment – partnership in New York City, NY, USA. *Journal of Soils and Sediments* 18: 352–357. <https://doi.org/10.1007/s11368-016-1456-8>.
- Shchepeleva AS, Vasenev VI, Mazirov IM, Vasenev II, Prokhorov IS, and Gosse DD (2017) Changes of soil organic carbon stocks and CO₂ emissions at the early stages of urban turf grasses' development. *Urban Ecosystem* 20: 309–321. <https://doi.org/10.1007/s11252-016-0594-5>.

Ecosystem services provided by SUITMAS

- Bouzouidja R, Séré G, Clavier R, Ouvrard S, Nuttens L, and Lacroix D (2018) Green roof aging: Quantifying the impact of substrate evolution on hydraulic performances at the lab-scale. *Journal of Hydrology* 564: 416–423. <https://doi.org/10.1016/j.jhydrol.2018.07.032>.
- Pugh TAM, MacKenzie AR, Whyatt JD, and Hewitt CN (2012) The effectiveness of green infrastructure for improvement of air quality in urban street canyons. *Environmental Science & Technology* 46: 7692–7699. <https://doi.org/10.1021/es300826w>.
- Schwager J, Schaal L, Simonnot MO, Clavier R, Ruban V, and Morel JL (2015) Emission of trace elements and retention of Cu and Zn by mineral and organic materials used in green roofs. *Journal of Soils and Sediments* 15: 1789–1801.

Restoration of soil functions to improve ecosystem services

- Séré G, Schwartz C, Ouvrard S, Sauvage C, Renat JC, and Morel JL (2008) Soil construction: A step for ecological reclamation of derelict lands. *Journal of Soils and Sediments* 8: 130–136.
- Villeneuve C, Séré G, Schwartz C, Watteau F, Jimenez A, and Cortet J (2018) Rapid changes in soil nematofauna in the first years after Technosol construction for the reclamation of an industrial brownfield. *Eurasian Journal of Soil Science* 51: 1266–1273. <https://doi.org/10.1134/S1064229318100149>.

Structure and function of forested soils

Peter Clinton and Jen Owens, New Zealand Forest Research Institute Ltd (Trading as Scion Research), Rotorua, New Zealand

© 2023 Elsevier Ltd. All rights reserved.

Introduction	56
Human use of forests and forest soils	58
Intensive management of forest and forest soils	59
Microbiome: Roots and mycorrhizae	60
Organic matter recycling in forests: Litterfall and the forest floor	60
Global carbon dynamics and greenhouse gases	61
Watershed nutrient dynamics and hydrology	62
Soil and forest resilience	63
Sustainable forest management: Criteria for indicators	64
Soils in urban forests	65
Future pressures and requirements for forest soils	65
Summary and conclusions	65
Acknowledgments	65
References	66

Abstract

Forest soils provide vital ecosystem services such as filtering and allocating water, storing carbon, and providing habitat. However, changes to climate and increased requirements for the spaces that forests occupy for other land uses such as agriculture and urban development increases pressure on forest soils. This questions their ability to deliver ecosystem services in the future. This chapter reviews global forest soils, their microbiome, nutrient dynamics, indicators of sustainable forest management, how climate change will impact forests' ability to provide ecosystem services, and finally discusses impending pressures on forest soils.

Glossary

Microbiome The microorganisms in a particular environment.

Mineralization The process by which chemicals present in organic matter are decomposed or oxidized into easily available forms to plants.

Nitrogen fixation The chemical processes by which atmospheric nitrogen is assimilated into organic compounds, especially by certain microorganisms as part of the nitrogen cycle.

Phytobiome Plants, their environment, and their associated communities of organisms.

Weathering The process of wearing or being worn by long exposure to the atmosphere.

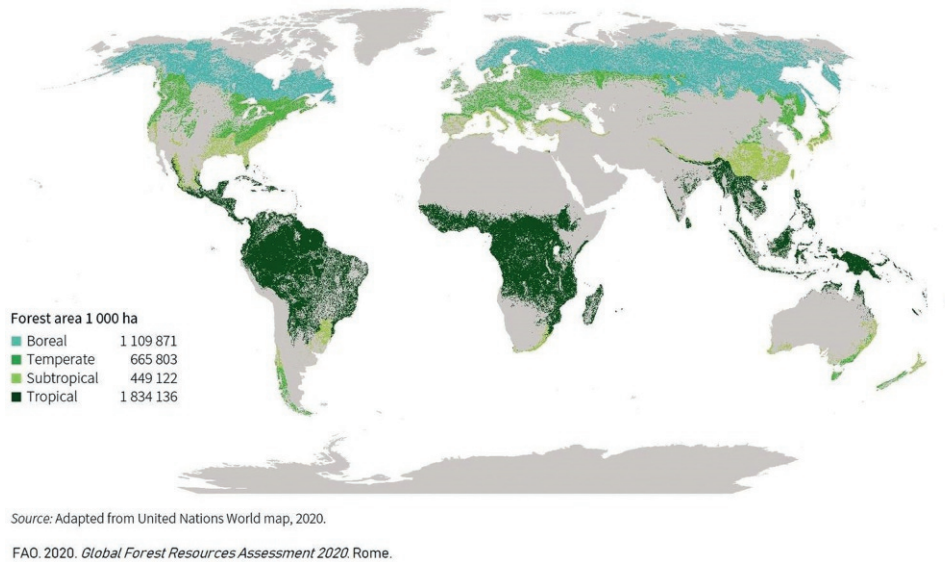
Key points

- Forests and forest soils provide ecosystem services including water filtration and carbon storage.
- Forests are under increased pressure from climate change and changes in land use which will affect soils and their capacity to supply ecosystem services.
- Soil properties are critical to forest resilience which is influenced by forest management activities.

Introduction

Why are forest soils and the forests that they support so important? Forests, and therefore forest soils, represent key ecosystems that are essential to life on Earth, support many aspects of nature, and provide society with a wide range of ecosystem services and benefits. The protection and maintenance of these services and benefits is fundamental to the natural world and to human well-being. The topic of forest soils is as broad as it is deep and includes, but is not limited to, the many aspects of soil biology, soil chemistry, and soil physical processes, effects of management, and understanding of natural processes that contribute to the wide variation in the world's forest soils.

The global distribution of forests, by climatic domain

**Fig. 1** The global distribution of forests by climatic zones.

Forests cover an area of 4.06 billion ha, representing approximately one third of the world's land surface area and are widely distributed, only excluding ice, snow, and rock covered areas. They are concentrated in five countries (Russian federation, Brazil, Canada, the United States, and China) that have 55% of the global forest cover. Tropical forests represent 45% of forest cover with 27, 10, and 11% occurring in Boreal, Temperate, and subtropical zones, respectively (Fig. 1). Of the total global area of forests, primary forests cover about 1.01 billion ha. These primary forests are characterized by low levels of human disturbance; they are dominated by natural processes, and consist of native species. The remaining forests are either regenerating or highly modified natural forests or are forests that have been planted and managed with a range of goals. Because of the wide and variable areas covered by forests, forest soils occur in many geographical locations around the world, creating an extraordinarily large potential for variation in the situations (forest types, geologies, and climate) in which the soils were formed. This has created a multitude of properties reflecting a correspondingly large diversity of topography, time scales, vegetation, and natural disturbance regimes together with the effects of human activity during the Holocene/Anthropocene. With increasing pressures from climate change impacts such as fire or invasive forest species, and increasing intensification of management, forest soils will continue to develop to express new properties with time.

Forest soils can be characterized as supporting potentially large, long-lived individual trees with deep root systems, substantial litter layers, and coarse woody debris. Forests represent a store of approximately 662 gigatons of carbon, of which 55% is stored in deadwood, litter, and soil organic matter (Food and Agriculture Organization, 2020). The deadwood and litter are important components of forest soils and provide important habitats for a wide range of forest animals, insects, and microorganisms. Forests also contain communities of understorey species, a wide diversity of soil dwelling organisms, including what is now widely termed as the soil microbiome (fungi, bacteria, amoeba, and Oomycetes), which through their activities, also cycle carbon and nutrients into and through forest soils. Key forest ecosystem processes of litterfall, carbon, and nutrient cycling are essential to maintaining productivity and these processes depend on this wide range in biodiversity to drive forest growth.

Forests and the forest soils that support them are widely seen as key to global sustainability now and into the future (World Business Council for Sustainable Development, 2015). The WBCSD suggested that the world needs more forests to meet current and future expectations for wood products, and for environmental services and benefits from forests. The WBCSD highlighted that up to 300% more fiber is required to meet the expected global shortage, which anticipates that the annual demand for wood will triple by 2050 to more than 10 billion m³ year⁻¹. This increased demand is attributed to population growth, rise of the middle class, and changing demands, for example, bioenergy and biofuels, and the rise of the bioeconomy. In addition to these increased demands, forests currently directly affect the livelihoods of 20% of the world's population, provide approximately 75% of clean freshwater, and support 9% of the world's primary energy demands while removing and storing vast quantities of carbon dioxide. Forests also support habitat for many important species with approximately 80% of terrestrial biodiversity found in forests (World Business Council for Sustainable Development, 2015). This importance of forests highlights just how fundamental forest soils are to so much of what forests provide. In terms of global efforts to stem climate change and the associated impacts, forests are seen as one of a few mechanisms that can have an effective impact in the relevant time frames, again further highlighting why forest soils are so important.

Human use of forests and forest soils

Forests have been a key source of timber and non-timber products for many generations. As civilization has developed, forests have been cleared to provide land for buildings, towns, cities, and agriculture. Out of this period of liquidation of old growth forests, forest management practices have emerged that have allowed for the protection and conservation of old growth forests and the emergence of intensively managed planted forests that now supply much of the world's supply of timber products. As a result, a wide range of forest types have been recognized with the key distinction between natural and planted forests (Table 1).

In the absence of harvesting for timber, trees in natural forests can be long-lived, and there are many examples of forest ecosystems around the world that contain populations of trees that are many hundreds or thousands of years old. Natural forests can, and in many cases, have sometimes undergone numerous natural disturbances such as fire, wind, earthquakes, disease, or insect attack. These disturbances can result in tree death, toppling of stems and tipping over of root plates, creating large amounts of coarse woody material. Large scale disturbance of forest can result in significant soil disturbance which will have implications for the ongoing development of the forest soil. The soil horizons can be mixed through the action of root plates tipping over or with potentially large amounts of organic matter buried beneath landslides. In particular, wind throws often create what is known as pit and mound topography, and freshly exposed mineral soils can become rapidly weathered. The mounds also act as elevated sites for seedling establishment and regeneration. In areas of the world where it occurs, volcanic disturbance, such as eruptions, can have a profound effect on forest soils. Volcanic eruptions can generate large amounts of heat, which can bake the soil and make it unsuitable for plant growth. Lava flows from an eruption can destroy the rooting systems of trees, making it difficult for them to recover. The loss of trees can also cause changes in soil hydrology, leading to increased erosion and runoff. In the long term, forests typically recover from these disturbances, but the process can take many years. After an eruption, the ash and other debris can blanket the ground, smothering vegetation and altering the soil structure and restricting infiltration. Ash deposits can lead to changes in the soil composition and nutrient availability, which in turn can affect the types of forest vegetation that can grow in the area. Since the last glaciation, there have been several generations of forests and cycles of natural disturbances with ensuing cycles of soil development and renewal as forests regenerate and stand dynamics alter the structure of the forests with time.

Planted forests managed for production of timber can also undergo these same disturbances, but the predominant disturbance occurs at the time of harvesting where there is potential for soil disturbance and modification during harvesting and stand replacement activities. Through time, a range of harvesting systems has been used to extract timber. These systems have become more intensive, as old growth forests were cleared for timber and to make way for expanding agricultural and urban land uses. Early on it was recognized that timber supply from old growth forests would not last forever. Therefore, in many places around the world,

Table 1 Example of FAO classification criteria for different types of natural and planted forests.

Natural forest		Planted forest				Non-forest
Primary	Modified natural forests	Semi-natural forests		Plantations		Trees outside forest (TOF)
		Assisted natural regeneration	Planted component	Protective	Productive	
						
Forest of native species, with no clearly visible indications of human activities and the ecological processes are not significantly disturbed.	Forest of naturally regenerated native species with clearly visible indications of human activities.	Silvicultural practices for intensive management (weeding, fertilizing, thinning, selective logging).	Forest of native species, established through planting, seeding or coppice of planted trees.	Forest of native or introduced species, established through planting or seeding mainly for provision of services.	Forest of introduced or native species established through planting or seeding mainly for production of wood or non-wood goods.	Stands smaller than 0.5 ha; trees in agricultural land (agroforestry systems, shelterbelts, orchards); trees in urban environments; and scattered along roads and in landscapes.

Adapted from Food and Agriculture Organization (2020) *Global Forest Resources Assessment 2020 Term and Definitions*. Roma: FAO.

plantations of timber producing species were established on previously forested lands, sometimes with introduced exotic species or with local indigenous species. There are 292 million ha of planted forests. Of these planted forests 131 million ha are considered to be plantations with more than 50 million ha of these managed intensively. Globally, 44% of plantation forests are planted with introduced species, primarily with members of two key genera, *Pinus* and *Eucalyptus*. These species are fast growing and provide timber that is suitable for many industrial purposes. The implications of these intensively managed forests for soil development beneath them is unknown, especially when the effects of management activities such as harvesting, site preparation, and stand management are considered.

Intensive management includes the use of herbicides, fertilizers, and thinning and pruning of stems with the goal of increasing timber production. Practices such as thinning and pruning do return organic matter to the soil surface which then enters the carbon cycle, providing habitat and energy for a range of forest dwelling organisms. The use of chemicals such as herbicides and fertilizers can have a variety of effects on plant nutrient and carbon cycling, although their use is confined to certain times during the life of the forests. These practices can result in shorter rotations, increasing the consequences of forest harvesting for soil disturbance. If done poorly, the use of heavy harvesting machinery can result in compaction and or mixing of soil horizons. However, most harvesting operations are regulated to avoid harvesting when conditions could result in soil damage, and to ensure practices such as brush mats consisting of non-merchantable stems, branches, and tree-tops, can be used to mitigate machine induced disturbance on harvesting trails. Other harvesting systems, such as continuous cover (the practice of maintaining a continuous canopy while removing selected trees), are also used to minimize disturbance to the forest and soils, particularly on slopes that are prone to erosion. Other harvesting practices include single tree or group selection or small coupe harvesting, which aim to minimize harvesting impacts on the forest and soils. Clear cutting of large coupes is still widely used in commercial plantations and can be highly mechanized, allowing for whole tree harvesting which can result in removal of not just the stems, but tree-tops, branches, and needles and leaves from the forest to log processing sites located on road edges. The processing of logs at the stump with suitable harvesting heads minimizes the loss of nutrients and returns the nutrient rich needles and leaves to the site together with the branches and treetops. The harvesting of trees for timber also presents the opportunity to use harvesting residues for energy purposes.

Natural disturbances to forest soils are usually more intense and occur over a shorter duration than those caused by forest management activities. In addition, natural disturbances typically affect a larger area of the forest and are less selective in terms of the plants and animals that are affected. As a result, natural disturbances can have a more profound impact on the structure and function of forest soils.

Intensive management of forest and forest soils

Wood has been widely used as a source of energy for cooking and heating since early civilizations. The use of wood for these purposes is still very common today and there has been widespread interest for many years in using wood to produce heat or steam, or steam for generating electricity, or as a feed stock for liquid biofuels or other biochemicals. Interest in using woody biomass for bioenergy has continued to grow since the first oil shocks in the 1970s. These shocks spiked an interest in alternative sources of energy, and now in modern times when concerns are fueled by efforts to reduce emissions from fossil fuels, the use of woody biomass for producing wood pellets for heating or biofuels for the road, marine, and aviation transport sectors is receiving more attention than ever. Biomass harvesting has increased, through the use of harvest residues from conventional harvested stands. More recently, purpose grown short rotation stands of fast-growing species such as willows have been established and managed for biomass crops. Another practice that has emerged with the harvesting of forest residues is the practice of stump harvesting. This is where the whole stump and as much of the root systems as possible is extracted and ultimately processed into bioenergy. This is not widely practiced as it is seen as resulting in considerable soil disturbance and greater loss of site nutrients.

The removal of harvest residues and the nutrients and carbon they contain was identified as an area of concern e.g., [Weetman and Webber \(1972\)](#) early on. In response, forest and soil scientists in many countries set up experiments in the late 1980s and early 1990s to examine the long-term impacts of intensive harvesting practices that involved the removal of harvest residues from conventionally harvested stands (stem only harvesting vs. whole tree harvesting). There has been much debate and scientific research about the sustainability of this practice and more recently about the climate effects of bioenergy ([Cowie et al., 2021](#)).

Most of the trials in planted forests have only examined the impact of the intensity of removal of harvesting residue at the end of one rotation; therefore, the question remains as to what might happen to plant nutrient resources at the end of subsequent rotations. Will nutrient removals be offset through atmospheric inputs or weathering of parent materials? This issue will gain more attention in the coming years as an increasing proportion of the current planted forest estate undergoes its second or third rotation. As the world tackles the goal of zero emissions and reductions in the availability and use of fossil fuels, the benefits of bioenergy production from forest residues are likely to increase, with uncertain consequences for forest soils. The findings generated by earlier research can be used to address issues arising around the sustainability of biomass harvesting (e.g., [Vance et al. \(2014\)](#)) and could prove useful in underpinning any development of local guidelines (e.g., [Titus et al. \(2021\)](#)) to support the use of forest residues for bioenergy or feedstocks for tree based biorefineries.

In some cases, the results provide confidence that not only is the productive capacity of most sites unlikely to be impacted after one rotation, but they also suggest that the future security of forest carbon pools and sequestration capacity will also be maintained in many situations. Care will have to be taken in situations where natural soil fertility is poor or where soil physical properties will not tolerate frequent disturbance during harvesting. Regardless of what the early results will have shown, it will be important to

continue this line of research because there will be an increasing area of planted forests entering their third or fourth rotations, where rotations are short and demand for harvest residues is likely to increase.

In addition to fertilizer application to replace nutrients removed during harvesting, strategies to ameliorate nutrient removals, particularly nitrogen, may include the use of nitrogen-fixing plants during periods of increased nutrient demand until canopy closure. There are numerous examples of suitable nitrogen-fixing species being sown into forest stands, and the occurrence of naturally occurring nitrogen-fixing species which can be considered weeds because of their competitive interactions with trees.

Other practices have also evolved to replace nutrients removed during harvesting and involve the use of common waste streams, for example, municipal wastewater treatment plant sludges commonly known as biosolids. Biosolids contain both carbon and nutrients, and their application to forest soils is seen as minimizing their negative consequences because forests are separate from the human food chain. The application of biosolids can benefit tree growth, especially where the forests are intensively managed, for example, where harvest residues are removed. However, concerns are still raised about possible organic and inorganic contaminants in biosolids. Therefore, it is essential that application programs are closely monitored to reduce the potential for any unintended consequences to arise.

Effluents from municipal wastewater treatment plants can also be applied to forests. One waste stream arising from the use of biomass is wood ash. Ash is rich in cations and is seen as a good way to replace the nutrients removed in the wood and bark of trees.

The planting and growth of forests together with the development of carbon and nutrient cycles have also been widely used to rehabilitate soils at the landscape level, particularly in the case of open cast mining. Once established on the reclaimed soils, trees provide protection to the soil surface and the production of litter returns carbon and nutrients to the soil. The growing trees provide habitats to a wide range of forest animals and insects, and the accumulating litter layer provides resources and habitat for more insects and animals.

Microbiome: Roots and mycorrhizae

Mycorrhizae have long been recognized as fundamentally important to the success of plant life. Mycorrhizal fungi can be found in association with the fine roots typical of tree root systems throughout the soil, wherever roots extend to. There are two key types of mycorrhiza common to most trees. These are endo or ecto mycorrhizae. The initial understanding of the importance of fungi in forest soils has now been extended to include the bacteria in forest soils found close to roots. Overall, the microbial communities consisting of fungi and bacteria found close to roots are known as the root microbiome and is part of the plant phytobiome. Through modern molecular tools and ecological principles, new understanding of the functions of the root microbiome has emerged. Previously, the understanding of the role of bacteria in forest soils was limited to functions, for example, nitrogen fixation and nitrogen cycling. Of far greater importance now is the idea of a forest wide web and the interconnectedness of trees through the microbiome, building on the concept of Darwin's 'root-brain hypothesis' and plant responses to environmental signals. This is a fast-moving area of research that is more than likely to impact greatly on our understanding of forest soils and the interactions between trees and microbes. This is similar to our understanding of the communication between the human gut microbiome and human brain that is redefining our understanding of human health and well-being. Managing the tree root microbiome or engineering the rhizosphere is considered to offer ways to increase forest health, productivity, and resilience under climate change.

Organic matter recycling in forests: Litterfall and the forest floor

Forests cycle carbon, nutrients, and water between the soil, atmosphere, and trees. Trees create sugars to support their growth by absorbing carbon dioxide and water during photosynthesis. Organic matter is deposited on the forest floor as leaves, twigs, and branches, and forests also add organic matter to soil through fine root turnover and decomposition of microbial biomass. This organic matter is decomposed by bacteria, fungi, and other soil organisms that releases nutrients that were locked in organic matter via mineralization, making them available for uptake by trees and other plants (Krishna and Mohan, 2017). This cycle of nutrient uptake and release helps to maintain fertile soil, preserve a healthy forest ecosystem, and allows forest soils to supply trees in forests with up to 60% of their nutritional requirements.

The development of a litter layer on the forest floor differentiates forest soils from those in other ecosystems such as grasslands. However, not all forests develop a distinct litter layer. Litter layers may not develop in forests where rates of litterfall are low, decomposition rates outpace deposition of litter which occurs in some tropical systems, or where accumulated litter is regularly removed or burned. Surface soils in forests are characterized by high organic matter content resulting from litter decomposition. However, it is not always possible to delineate a distinct organic layer from the underlying mineral soil, if, for example, management practices mix topsoil into the subsurface during harvesting or if the soil profile has been disturbed by animals.

Soil organic matter is a critical component of soil quality; it improves forest productivity by enhancing nutrient and water availability. It also plays an important role in soil structure by contributing to aggregate formation, increasing soil particle surface area which enhances water absorption, and by leading to the production of humic substances (humins, humic acids, and fulvic acids). These compounds are unlikely to be further decomposed due to their chemical structure (Berg and McClaugherty, 2020). Humus found on the forest floor shows distinct development and properties in different types of forests (Fig. 2).

Trees, particularly in boreal forests, have been shown to access free amino acids through their mycorrhizal associations, which has led to the development of organic nitrogen fertilizer that can be used instead of traditional mineral nitrogen fertilizer.

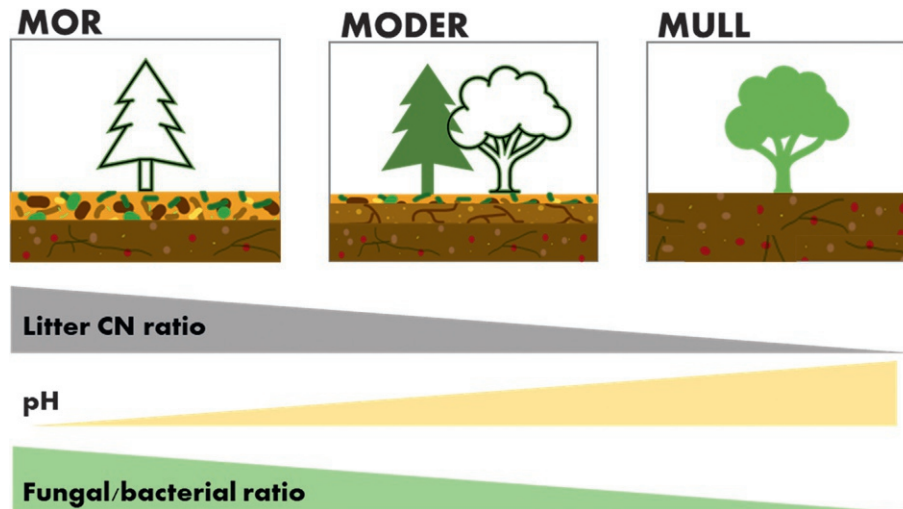


Fig. 2 Litter on the forest floor decomposes into simpler organic components. Humification breaks down plant debris leading to the formation of different types of humus under different forest types. Mor humus develops under coniferous forests under cool, wet climate conditions, where the breakdown of organic material is slow due to the low abundance of soil biota. This type of humus is characterized by a matted litter layer and an abrupt boundary between mineral soil and organic soil. Mull humus develops under deciduous forests where base-rich plant litter is actively broken down by abundant soil biota. It is characterized by a thin or absent litter layer, and a horizon of organic-enriched mineral soil. Moder is the intermediate between mull and mor types. This has a loosely structured litter layer and gradual boundary between mineral and organic layers.

Forest management practices that influence soil organic carbon		Afforestation and conversion of primary forests	Harvesting and harvest residue management
		Biomass removal and fire management	Nitrogen fertilizer addition and N-fixing species
Stand density and thinning	Tree species selection and diversity		
		Soil disruption and compaction from site preparation	Herbivory regulation

Fig. 3 Many forest management practices can increase or decrease forest soil organic stock.

Forest management activities can increase or decrease a soil's organic matter content, depending on the nature of the activities (Fig. 3). Examples of forest management practices that can increase soil organic matter include afforestation and adding nitrogen as fertilizer. Other management practices, including harvesting and removing harvest residue, generally have a negative impact on forest soil organic matter content because they remove the litter and wood that fuels decomposition and therefore nutrient cycling in forest soils (Mayer et al., 2020).

Global carbon dynamics and greenhouse gases

Forest soils are important for regulating atmospheric carbon dioxide levels and play a significant role in the global carbon cycle. Trees and other vegetation absorb carbon dioxide from the atmosphere through photosynthesis and store it in their leaves, stems, and roots. When leaves and other organic matter fall to the forest floor and decompose, they release carbon dioxide back into the atmosphere through respiration. Carbon is also transferred from trees to soil by rhizodeposition, which is the process whereby plants secrete organic matter and nutrients into the soil through their roots. Rhizodeposition can stimulate microbial activity in the

soil, which can lead to increased decomposition of organic matter and greater nutrient release, and has been shown to increase soil organic matter, improve soil structure, and increase soil water holding capacity. The net balance of the amount of carbon that is stored in forests, and the rate at which it is released as carbon dioxide, can have a significant impact on global climate change. Because forests cover a substantial area of the world's land surface, small changes to forest coverage can have substantial impacts on global carbon dynamics.

Forests mitigate greenhouse gas emissions by sequestering carbon and storing it for long periods of time. Carbon dioxide absorbed by trees is stored long-term in the trees themselves and the wood-based products they are used to produce, and as organic matter in forest soil (Pan et al., 2011). Globally, forest soils store 163 t ha^{-1} of carbon as organic matter, or 45% of all carbon in forests, suggesting soils have the potential to play a major role in mitigating climate change (Food and Agriculture Organization, 2020b).

There are many factors that contribute to the role of forest soils in global carbon dynamics. For example, the type of trees in a forest can affect the amount of carbon that can be stored in soils. Trees with deep roots tend to store more carbon in their soils than trees with shallow roots. Likewise, older forests tend to have more carbon stored in their soils than younger forests. Forest management, climate change-induced disturbances, and human activities can also impact the carbon storage capacity of forest soils. For example, by reducing or eliminating organic matter inputs from trees to soils, deforestation can lead to a decrease in the amount of carbon that is stored in soils, and forest fires can release large amounts of carbon into the atmosphere (Mayer et al., 2020).

Carbon stored in forest soils may help to offset other greenhouse gases such as methane and nitrous oxide. Forest soils can function as net sources or sinks of methane and nitrous oxide which are potent greenhouse gases that contribute to climate change. Methane is produced in soils by methanogenic microorganisms under oxygen depleted conditions which leads to methane emissions from soils. However, methane is also consumed by methanotrophic microorganisms in the presence of oxygen, which leads to the soil taking up more methane than they produce. Most forests situated on unsaturated soils are net sinks of methane. There is some evidence, however, that suggests net uptake of methane by forest soils is declining, especially at latitudes between 0 and 60°. This is due to changes in precipitation patterns and soil hydrological fluxes that reduce methane diffusion from the atmosphere into the soil which inhibits methane uptake by methanotrophic soil microorganisms (Ni and Groffman, 2018). Nitrous oxide is produced in soils by biological activity or chemical reactions. The majority of nitrous oxide emitted from soil is produced during biological denitrification in low oxygen conditions when carbon and nitrate are available (Zhu et al., 2013). Nitrous oxide is an intermediate product of denitrification, whereas complete denitrification produces nitrogen gas, a natural and abundant constituent of the atmosphere. In oxygen depleted soils, complete denitrification outpaces incomplete denitrification. Soils can become net sinks for nitrous oxide as uptake occurs. In forest soils with low levels of nitrate and forest floors with high C:N ratios, we can generally expect forest soils to be net sinks for nitrous oxide. However, it is unclear whether global forest soils are net sources or sinks of nitrous oxide because conditions are highly variable and other biological processes such as nitrifier denitrification can produce and take up nitrous oxide (Chapuis-Lardy et al., 2007).

Forest management practices have a significant impact on emissions of carbon dioxide and other greenhouse gases from soils. Soil carbon dioxide emissions can be up to 10 times higher in managed forests than in unmanaged forests. This can be partly attributed to soil disturbance that occurs during forest management. Clearcutting and replanting disturbs the soil and can release carbon dioxide and other greenhouse gases that were previously trapped in the soil. It also alters the soil conditions, thereby affecting nitrous oxide and methane fluxes. For example, removing large swaths of trees in boreal regions suddenly eliminates tree water uptake, and this could elevate local water tables leading to wetter soils that are more enabling to nitrous oxide and methane production. In contrast, by removing only certain trees, selective logging can decrease nitrous oxide and methane emissions from soils. This is because the remaining trees create a microclimate that is more conducive to aerobic soil conditions, thereby preventing conditions that enhance production of nitrous oxide and methane.

Watershed nutrient dynamics and hydrology

Forests regulate water quantity and quality in many regions around the world. Changes in forest cover alter the supply of water resources by influencing precipitation patterns, drinking water supply and quality, flood control, stream flows, and recreational activities (Holmes et al., 2017). Forested watersheds supply a large proportion of the world's accessible fresh water for domestic, agricultural, industrial, and ecological needs in both upstream and downstream areas (Food and Agriculture Organization, 2013).

Forest soil conditions and properties strongly influence how much water is stored in soil, how quickly water moves through the soil matrix, and how quickly water moves from soil to surrounding surface or groundwater sources, to plants, and back to the atmosphere (Black, 1996). Forested soils have unique characteristics that affect their ability to hold and transport water. Unlike most other ecosystems, forest soils receive large inputs of leaf litter and woody debris, and decomposition of this litter contributes to the development of soils with a large content of organic matter, and in many non-tropical forests, a litter layer. Large organic matter content in soils and forest litter layers are both associated with large pore spaces, which contribute to infiltration rates that often exceed rainfall intensity. As a result of these characteristics, forest soils have high water storage capacity, and the presence of a litter layer both holds water and creates a physical barrier that reduces water loss from soil as evaporation (Edwards et al., 2015).

Soil hydrology and biogeochemistry are intimately linked in forested watersheds (Fig. 4). Forested watersheds improve water quality and enhance water storage; however, forest management affects nutrient cycling, water transport, and surface and groundwater quality and quantity. Management activities such as road construction, harvesting, site preparation for regeneration, and

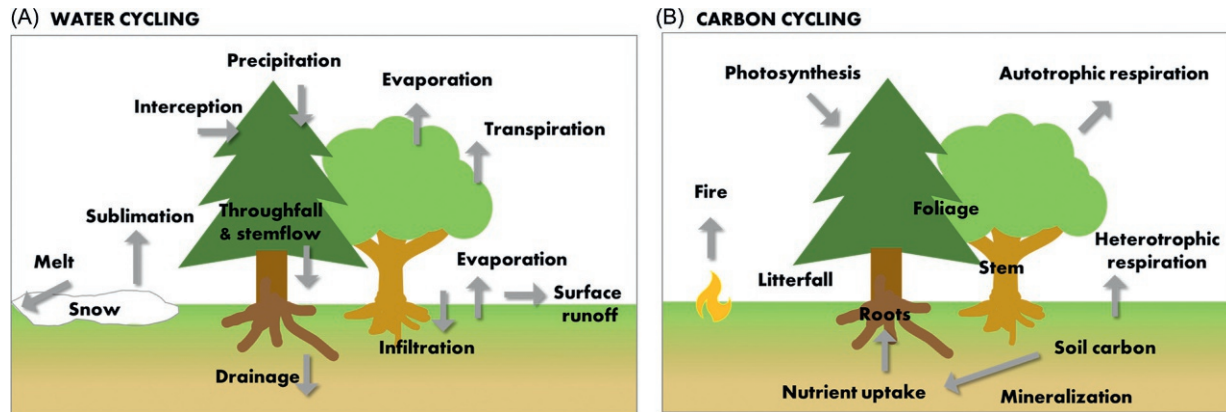


Fig. 4 Water and biogeochemical cycling in forests are tightly linked. Left (A) shows a simplified representation of water cycling in forests while right (B) shows a simplified representation of carbon cycling.

fertilization are all linked to reduced water quality in forested watersheds. These activities cause changes to sediment loads, stream temperature, dissolved oxygen, and dissolved nutrients.

Nitrogen is the primary nutrient that limits tree growth, but boron and other micronutrients can also reduce forest productivity. Many studies linking soil hydrology and biogeochemistry in watersheds have focused on nitrogen, partly because of its effect on productivity, but also because nitrogen in the form of nitrate is a pollutant in surface and groundwater. Nitrate leaching from forest soils to water sources can be a problem in some circumstances, for example, following harvesting when nitrogen uptake by trees is reduced, when planting new forests on poor quality soils that require the addition of nitrogen fertilizer, or when planting forests on land with a history of animal grazing and agricultural fertilizer use because nitrogen concentrations may already exceed tree requirements. There is also a risk of phosphorus loss in sediments following harvesting disturbance in the absence of best practices such as lack of riparian areas.

Soil and forest resilience

Resilience refers to a forest's capacity to withstand and recover from change and disturbance. The concept is usually associated with plant recovery, which is determined by easily measured metrics such as tree growth, reproduction, abundance, canopy cover, species diversity, and biomass. Forests are under increased pressure from climate change and disturbance from wildfires, insect attacks, and storms. Biotic and abiotic soil characteristics are critical to the survival and recovery of trees following such events (Ibanez et al., 2019).

Forest resilience relies on reciprocal relationships between trees and soils. Reductions in forest canopy cover or stand density will reduce organic matter inputs from litter and woody detritus, which can, in turn, reduce forest soil organic matter content, thereby decreasing soil water holding capacity, altering nutrient availability and transport, and modifying soil ecology. Thus, these changes in tree canopy or stand density will influence how soil conditions respond to changes in temperature, solar radiation, and water inputs, thereby altering a soil's capacity to support forest growth.

Climate change can affect soils in several ways, and these changes can impact forest resilience. Increased intensity and frequency of droughts will lead to more soil erosion and degradation. By reducing soil water content, droughts are also associated with increased soil temperatures; wet soils have a greater specific heat capacity requiring more energy to change their temperature, and therefore, wet soils have a greater ability to regulate temperature changes. Warmer soil temperatures can stress trees and increase tree mortality. Changes in precipitation patterns can affect the amount and timing of available water in soils, which can affect their ability to support forest growth. Soils that have experienced wildfire are often more susceptible to erosion, which can lead to the loss of soil and plant nutrients. This makes them susceptible to the development of hydrophobic conditions, reduced infiltration, and increased runoff as a result of wildfire burning off soil organic matter. They also have increased susceptibility to flooding and encroachment by invasive species. Overall, the effects of reduced tree growth, and more frequent droughts and fires, can lead to a cycle of ecosystem and soil degradation, which will reduce forest soil health and the resilience of forests to these disturbances.

By disrupting and altering the flow of energy and nutrients between soils and trees, climate change-induced disturbances can alter a soil's capacity to support tree growth thereby affecting organic matter inputs from trees to soil. Soils with a large organic matter content and good structure provide habitat to support healthy populations of soil organisms that cycle nutrients and organic matter and help to protect soils from erosion. Fertile soils that are protected from erosion are better able to recover from disturbances such as insect attacks. While disturbances such as insects can cause significant damage to trees, if the soils are healthy, the trees are more likely to survive and recover.

Forest management activities can influence the capacity of soils to support forest regeneration following harvest. Harvesting can interrupt energy, water, and nutrient flows between the soil and trees. However, whether such interruptions result in changes to long-term forest ecosystem functions will depend on many factors; steep slopes, shallow soils, limited nutrient supplies, and

absence of vegetation reduce the ability of a forest to recover following disturbance (Ibanez et al., 2019). The type of harvest activities will also affect forest soil conditions. Selective harvesting of natural forests only removes one or a small number of trees leaving the rest of the stand intact, thereby minimizing the effects on soil and the rest of the forest. However, clearcut harvesting of natural forests removes all trees, and removing the canopy creates soil conditions similar to those that occur following natural disasters which can severely reduce forest resilience. Harvesting machinery can influence physical soil properties which can hinder seedling establishment after harvest. Soils that have been damaged or compacted by heavy equipment are less able to support forest regeneration and tree growth.

Sustainable forest management: Criteria for indicators

Since 1990, an estimated 420 million ha of forests have been cleared for other land uses, including the production of timber. Loss of forests also means that the soils beneath the forests are subject to the influence of different land uses. Soils under either agricultural or urban land uses are no longer considered forest soils although their properties reflect this previous land use.

Since 1990, afforestation has reduced the net loss of forests; planting trees on non-forested land has reduced net losses of forests by 178 million hectares. Afforestation of agricultural land or degraded lands is widely seen around the world and in some cases, large gains in productivity have been observed on previous agricultural land (e.g., Beets et al. (2019)) due to the legacy of fertilizer inputs.

The trends of increasing deforestation and the potential negative impacts of poor management practices have seen the evolution of local, national, and international systems for monitoring and reporting on sustainable forest management practices, including the protection of forest soils and their productive capacity. Today, in response to these concerns, certification schemes have emerged for forests managed for timber products, and now 13% of global forest cover is covered by the Programme for the Endorsement of Forest Certification (300 million ha (Programme for the Endorsement of Forest Certification, 2021)), while 228 million ha of forests are covered by the Forest Stewardship Council Certification Scheme (Forest Stewardship Council, 2021). At the national level, many countries report the state of their forests and forest management through the Montreal Process for Sustainable Forest Management. This process has established a set of key criteria and indicators of sustainable forest management which are reported on regularly at the national level. This has allowed for trends in the criteria and indicators to be assessed at a global level and includes the state of forest soils. More recently, the United Nations Sustainable Development Goals have emerged and are relevant to the management of forest soils and the contribution they make to nature and human well-being (Fig. 5).

Soil quality underpins the productive capacity of forests which provide important ecosystem services and benefits. The quality of forest soils has been of interest for some time (Schoenholtz et al., 2000) because of the relationship of many soil quality indicators with forest productivity, which has been the focus of many studies. Amongst important soil quality indicators are nitrogen, phosphorus, C:N ratio, soil bulk density as well as total soil porosity. These have generally been applied to surface mineral soils and do not consider the organic horizons, which are an important feature of forest soils. They also do not consider deeper soils or processes not seen as directly important to plant growth. For example, the weathering of rock fragments within the root zone of trees and parent materials deeper in soils are not considered, but are areas where research could help to improve our understanding on the sustainability of managed forest soils. Remote sensing, geospatial techniques, and mapping technologies are providing useful insight into the nature of forests, and fusion of this new knowledge about forests with new sensing technologies applied to forest soils will increase our understanding about the evolving nature of forest soils.



Fig. 5 Goal 15 of the United Nations Sustainable Development Goals (SDGs) is Life on Land. The SDGs are related to one another, with the thicker lines and proximity to goal 15 representing other SDGs that are more closely related to goal 15.

Modern molecular tools have been applied to forest soils for several years, but more recently cheaper, high throughput analytical systems have provided an explosion of new understanding of forest soil microbial ecology. Further advances in the study of eDNA (Ruppert et al., 2019) will allow for an even greater understanding of the diversity of life in forest soils and how this varies between major forest types and with management. There is a current focus on what is now known as the forest microbiome and what this consists of, and the role it plays in the health and productivity of forests.

Soils in urban forests

The term urban soil describes soils located in urban areas that have been affected by human activities such as construction, pollution, and changes to the natural landscape, including the planting of trees. Forests and trees found in urban areas provide benefits to the environment and human health including reducing air pollution, providing shade, and improving mental health. Some soil classification systems already recognize urban soils as a distinct soil class with identifiable characteristics, and urban soils are becoming increasingly important as more than half of the world's human population live in urban areas. Some characteristics of urban soils may impede tree growth. For example, urban soils are often compacted by foot traffic and construction activity, which can limit oxygen availability to tree roots and inhibit water and nutrient infiltration and transport to tree roots. Urban soils often have a higher pH than other forest soils due to the presence of concrete and other building materials. Urban soils also tend to have higher concentrations of 'heavy' metals and other pollutants due to runoff from roads and buildings (Scharenbroch et al., 2005).

Like other forest soils, urban soils that are forested can sequester carbon and there is some evidence that this can be related to tree species. Relative to natural and production forests, less is known about nutrient cycling and carbon content of urban forest soils. Tree species vary in their litter quality, and also in their rates of litter decomposition (Binkley and Giardina, 1998). Research suggests that these differences contribute to differences in soil organic carbon content in urban soils, with greater soil organic carbon in urban forests planted with coniferous species than in urban forests planted with broadleaved species (Xu et al., 2021).

Future pressures and requirements for forest soils

There are several interrelated challenges ahead for forest soils. First, world population is likely to be close to 9.8 billion by 2050, creating future global demand for the provision of wood and fiber products, in the order of 10 billion m³ year⁻¹. This will require either larger areas of forests to be harvested, or greater productivity from existing forests. At the same time, the results of global efforts to keep global temperature rise below 2 degrees Celsius will be apparent, therefore, the challenge for forests and forest soils is how they can continue to be key carbon sinks and reservoirs that contribute to meeting the Paris Agreement targets while simultaneously providing the wood and fiber resources for traditional uses and meet new demands for biomass resources to replace fossil carbon derived products. At the same time, it is also expected that there will be greater disturbance from climate related events, for example, wildfires, extreme winds and floods, disease, and pest outbreaks that will result in loss of forests. This will jeopardize their ability to contribute to efforts to meet the Paris Agreement targets to provide timber resources, and at the same time put more pressure on remaining forests (Fig. 6). Whether or not there will be increasing pressure to clear more forests to meet demand for land from the growing population remains to be seen, given current trends to intensify agricultural production together with new trends in addressing demands for protein. Increasing population density in existing urban areas may offset the need to clear forested land, but on the other hand the need to move climate threatened communities might result in more loss of forests. To meet these interrelated challenges, it will be essential for the management and conservation of forests and forest soils to be a priority for future governments along with more collaboration internationally to secure these critical resources.

Summary and conclusions

Forest soils are as diverse as the forests they support but increasingly they will reflect the outcomes of many more years of human influences and homogenization of their properties. It will become even more important to maintain and enhance the productivity of our managed forests, given the reliance on forestry to reduce atmospheric carbon levels and substitute fossil fuels at the same time while protecting our remaining primary forests. This will require careful management of the soil resources ensuring that they can support increases in forest productivity while contributing to carbon sequestration and storage. Increasing global demand for timber products is creating more pressure on forests, and new management practices will need to be developed to buffer the soil resources against any pressures resulting from intensification (e.g., shorter rotations, higher stocked stands, more disturbance, and residue removal). However, new technologies promise the potential to avoid some negative impacts, such as robotic harvesting systems that protect the soil surface from disturbance while improving worker safety (Parker et al., 2016).

Acknowledgments

The authors wish to thank Serajis Salekin and Priscilla Lad for their comments and suggestions, and Loretta Garrett for the concept behind Fig. 5.

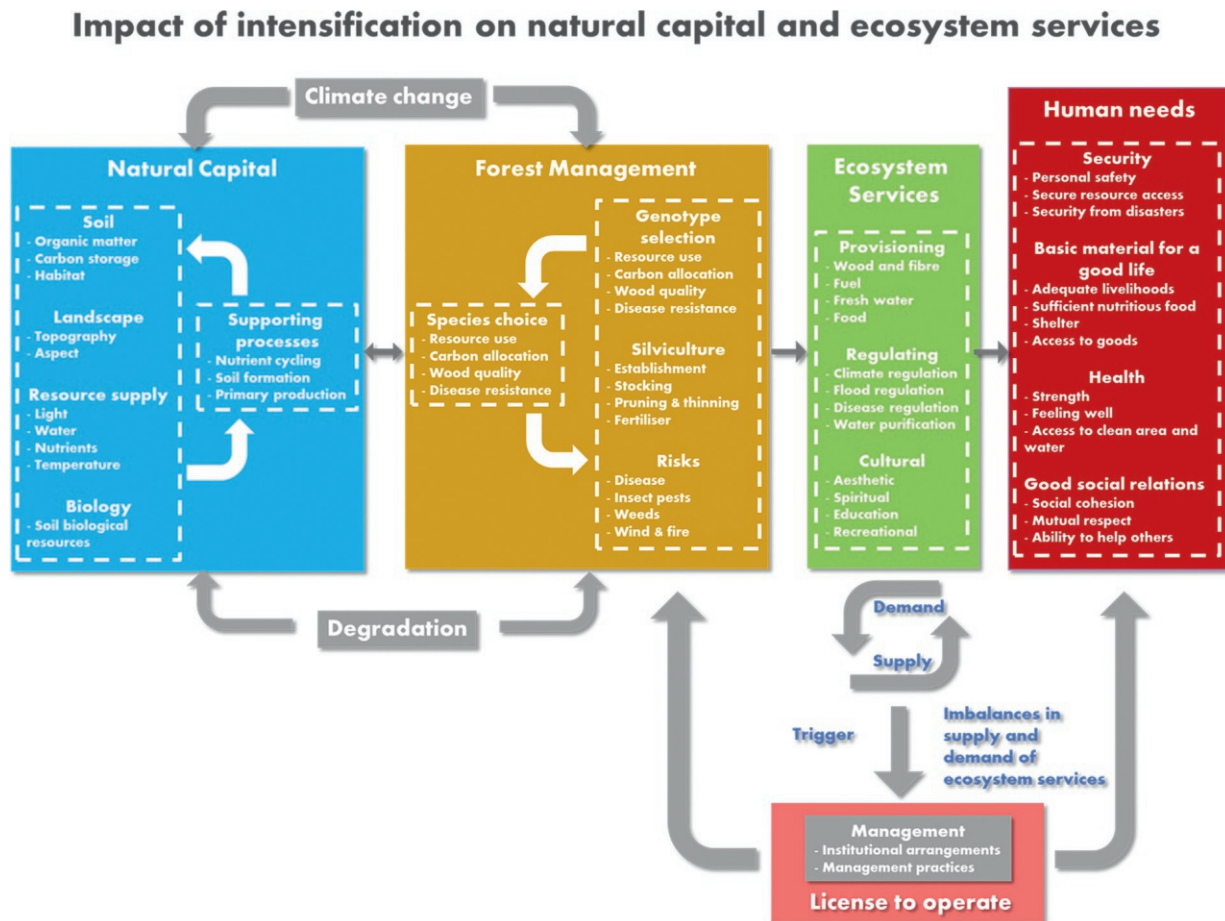


Fig. 6 This diagram shows how soil and other natural capital are important for management decisions in forestry. When trees are planted, they will interact with the natural capital including soils and the supporting processes such as nutrient cycling. Climate change will affect the natural capital and forest management decisions. Managing and protecting soil will impact ecosystem services under climate change and will future proof the planted forests through realizing the maximum benefits from planted forests. Moreover, it will give information to underpin licenses to operate.

References

- Beets P, Kimberley M, Garrett L, Paul T, and Matson A (2019) Soil productivity drivers in New Zealand planted forests. *Forest Ecology and Management* 449: 117480.
- Berg B and McLaugherty C (2020) *Plant Litter: Decomposition, Humus Formation, Carbon Sequestration*. Springer Nature.
- Binkley D and Giardina C (1998) Why do tree species affect soils? The warp and woof of tree-soil interactions. In: *Plant-Induced Soil Changes: Processes and Feedbacks*. Springer.
- Black PE (1996) *Watershed Hydrology*. CRC Press.
- Chapuis-Lardy L, Wrage N, Metay A, Chotte JL, and Bernoux M (2007) Soils, a sink for N₂O? A review. *Global Change Biology* 13: 1–17.
- Cowie AL, Bermdes G, Bentsen NS, Brandão M, Cherubini F, Egnell G, George B, Gustavsson L, Hanewinkel M, and Harris ZM (2021) Applying a science-based systems perspective to dispel misconceptions about climate effects of forest bioenergy. *GCB Bioenergy* 13: 1210–1231.
- Edwards PJ, Williard KW, and Schoonover JE (2015) Fundamentals of watershed hydrology. *Journal of Contemporary Water Research & Education* 154: 3–20.
- Food and Agriculture Organization (2013) *Forest and Water International Momentum and Action*. Roma: FAO.
- Food and Agriculture Organization (2020) *Global Forest Resources Assessment 2020: Main Report*. Rome: Italy.
- Forest Stewardship Council (2021) Facts & Figures. Available at: <https://fsc.org/en/facts-figures> (Accessed 3 March, 2022).
- Holmes TP, Vose J, Warziniack T, and Holman B (2017) Forest ecosystem services: water resources. In: Sills EO, Moore SE, Cabbage FW, McCarter KD, Holmes TP, and Mercer DE (eds.) *Trees at Work: Economic Accounting for Forest Ecosystem Services in the US South*. USDA.
- Ibanez I, Acharya K, Juno E, Karounos C, Lee BR, McCollum C, Schaffer-Morrison S, and Tourville J (2019) Forest resilience under global environmental change: Do we have the information we need? A systematic review. *PLoS One* 14: e0222207.
- Krishna M and Mohan M (2017) Litter decomposition in forest ecosystems: A review. *Energy, Ecology and Environment* 2: 236–249.
- Mayer M, Prescott CE, Abaker WE, Augusto L, Cécillon L, Ferreira GW, James J, Jandl R, Katzensteiner K, and Laclau J-P (2020) Tamm Review: Influence of forest management activities on soil organic carbon stocks: A knowledge synthesis. *Forest Ecology and Management* 466: 118127.
- Ni X and Groffman PM (2018) Declines in methane uptake in forest soils. *Proceedings of the National Academy of Sciences* 115: 8587–8590.
- Pan Y, Birdsey RA, Fang J, Houghton R, Kauppi PE, Kurz WA, Phillips OL, Shvidenko A, Lewis SL, and Canadell JG (2011) A large and persistent carbon sink in the world's forests. *Science* 333: 988–993.
- Parker R, Bayne K, and Clinton PW (2016) Robotics in forestry. *New Zealand Journal of Forestry* 60: 8–14.
- Programme for the Endorsement of Forest Certification (2021) *PEFC Global Statistics—June 2021*. Geneva. Available: <https://cdn.pefc.org/pefc.org/media/2021-08/725619c9-2460-4061-866d-95e160251648/22bd782f-bd0d-5840-a4bb-843400f15bea.pdf> (Accessed July 27, 2022).

- Ruppert KM, Kline RJ, and Rahman MS (2019) Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation* 17: e00547.
- Scharenbroch BC, Lloyd JE, and Johnson-Maynard JL (2005) Distinguishing urban soils with physical, chemical, and biological properties. *Pedobiologia* 49: 283–296.
- Schoenholtz SH, van Miegroet H, and Burger J (2000) A review of chemical and physical properties as indicators of forest soil quality: Challenges and opportunities. *Forest Ecology and Management* 138: 335–356.
- Titus BD, Brown K, Helmisari H-S, Vanguelova E, Stupak I, Evans A, Clarke N, Guidi C, Bruckman VJ, and Varnagiryte-Kabasinskiene I (2021) Sustainable forest biomass: A review of current residue harvesting guidelines. *Energy, Sustainability and Society* 11: 1–32.
- Vance ED, Aust WM, Strahm BD, Froese RE, Harrison RB, and Morris LA (2014) Biomass harvesting and soil productivity: Is the science meeting our policy needs? *Soil Science Society of America Journal* 78: S95–S104.
- Weetman G and Webber B (1972) The influence of wood harvesting on the nutrient status of two spruce stands. *Canadian Journal of Forest Research* 2: 351–369.
- World Business Council for Sustainable Development (2015) *Reporting matters: Redefining performance and disclosure, WBCSD 2015 Report*. Available: http://docs.wbcsd.org/2015/11/WBCSD_Reporting_Matters_2015_Interactive.pdf (Accessed July 27, 2022).
- Xu X, Sun Z, Hao Z, Bian Q, Wei K, and Wang C (2021) Effects of urban forest types and traits on soil organic carbon stock in Beijing. *Forests* 12: 394.
- Zhu X, Burger M, Doane TA, and Horwath WR (2013) Ammonia oxidation pathways and nitrifier denitrification are significant sources of N₂O and NO under low oxygen availability. *Proceedings of the National Academy of Sciences* 110: 6328–6333.

Relevant websites

- <https://www.scionresearch.com/about-us/about-scion/corporate-publications/scion-connections/past-issues-list/scion-connections-issue-37-september-2020/forest-microbiomes-small-actors-play-a-big-part>.
- <https://iba.org/eng/iba-publications/infographics>.
- <https://www.fao.org/forestry/99325/en/>.

Agroforestry

NV Thevathasan^a, AM Gordon^a, and PKR Nair^b, ^aSchool of Environmental Sciences, University of Guelph, Guelph, ON, Canada; ^bUniversity of Florida, Gainesville, FL, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of P.K. Ramachandran Nair, Agroforestry: Trees in Support of Sustainable Agriculture, Reference Module in Earth Systems and Environmental Sciences, Elsevier, 2013, ISBN 9780124095489, <https://doi.org/10.1016/B978-0-12-409548-9.05088-0>.

Introduction	68
Emergence of agroforestry as a land-use approach	69
Agroforestry: Concepts and principles	69
Agroforestry: Practices and systems	70
Distribution and area of agroforestry systems	71
Brief description of major agroforestry systems	72
Intercropping (alley cropping and other forms of tree intercropping)	72
Multistrata systems	74
Protective agroforestry systems	76
Silvopasture	76
Tree woodlots and specialty crops	78
Agroforestry and ecosystem services	78
Soil improvement	78
Water quality enhancement and nutrient interception	79
Biodiversity conservation	79
Carbon sequestration and climate change mitigation	79
System-level carbon sequestration in temperate agroforestry land-use systems (Canada)	81
Tree-based intercropping (alley cropping) system – A case study	81
Carbon sequestration in riparian buffer systems (RBS) in southern Ontario, Canada	83
Concluding remarks: The role of agroforestry in land management in the 21st century	84
Acknowledgment	85
References	86
Further reading	87

Abstract

Agroforestry incorporates trees/shrubs into farming systems, and allows for the production of trees and crops or livestock from the same piece of land in order to obtain economic, ecological, environmental and social benefits. Climate-resilient land-use systems that promote biodiversity, nutrient cycling and sustainability as an adaptation strategy to enhance people's life globally are being promoted by several international organizations. In this context, agroforestry land-use systems create multifunctional landscapes contributing to climate-resilient processes and social well-being and therefore considered as the future of global land-use. This book chapter highlights such land-use systems supported by recent research outcomes.

Introduction

Agroforestry involves the deliberate growing of trees and shrubs with crops, sometimes with animals, in interacting combinations for a variety of objectives. Agroforestry farming practices have been used throughout the world for millennia but have only attracted scientific attention and attained prominence as relevant land-use practices since the late 1970s. Today, acting as an interface between agriculture and forestry, agroforestry is considered to be a promising and sustainable approach to land use. Improvement and utilization of multipurpose trees that produce food, fodder, timber, and fuelwood in integrated farming systems offer sustainable solutions to several serious land management issues such as food security, environmental protection, and climate change, in both developing countries and the industrialized world. In the former, agroforestry is largely aimed at food security, whereas in industrialized countries, agroforestry land-use systems are often recommended for environmental protection (biodiversity enhancement, non-point source pollution and soil erosion control, and the development of long-term and efficient nutrient cycling mechanisms). In addition, agroforestry systems (AFS) can provide micro-climatic modification for enhanced water-use efficiency, and climate-change mitigation (carbon sequestration, enhancement of climate resilience) as well as product diversification to promote on-farm economic viability. This chapter presents an account of the major forms of agroforestry systems in both tropical and temperate situations and their role and potential for addressing current and future land-management issues.

Emergence of agroforestry as a land-use approach

The practice of growing trees and crops has been prevalent for many centuries in different parts of the world, especially under subsistence farming conditions. For example, homegardens are reported to have been associated with fishing communities living in moist tropical regions around 10,000 BCE, and domestic animals were introduced into forests for foraging in Europe around 4000 BCE. However, it has only been since the late 1970s that these indigenous forms of growing trees and crops/animals together have come to be thought of as a truly scientific land-use scenario.

Several factors contributed to this development. During the mid-to-late 1900s, agriculture and forestry in the industrialized nations had been successfully transformed into commercial, single-commodity-production enterprises. When the newly independent nations of the developing world were faced with the problem of feeding their large populations during the 1950s and 1960s, policy makers thought that the best solution was to embrace the successful agricultural models common in the industrialized world. Several food-production technologies were subsequently developed with an emphasis on the production of preferred crop species in monocultural (sole-crop) stands with a heavy reliance on agrochemical inputs (fertilizers, insecticides, herbicides, etc.), mechanization, and irrigation. In the 1970s, these efforts, collectively called the Green Revolution, resulted in substantial gains in food production and helped avert large-scale hunger in some regions of the world. At that time, most of the international agricultural research centers (IARC's) of the Consultative Group on International Agricultural Research (CGIAR) system and national country programs focused on a few individual food crops and their production technologies for sole-crop production systems. Similar patterns of single-species stands of trees were also adopted in plantation forestry and fruit orchards in many parts of the tropics and that resulted in significant gains in wood and fruit production. However, many farmers, especially the poorer ones in the tropics and subtropics, were often cultivating their crops in mixed stands for more than one season, and were commonly nurturing trees in their cropping systems for a myriad of reasons. These traditional, mixed production systems of raising food crops, trees, and animals together, as well as exploiting a multiple range of products from natural woodlots did not fit into this "modern" single-commodity paradigm, and were ultimately discouraged.

Serious doubts began to be expressed after the advent of the Green Revolution, about the relevance of single-commodity strategies and the policies promoting them. In particular, there was concern that the basic needs of the poorest farmers were neither being considered nor adequately addressed. It soon became clear that many of the technology packages that contributed to the Green Revolution, such as irrigation systems, fertilizers, and pesticides, were not affordable by the poor farmer on the one hand, and in addition, their large-scale application had serious environmental consequences on the other. It was also recognized that most tropical soils, which are poorer and more easily degraded than temperate-zone soils, were unable to withstand the impact of high-input technologies. In addition, the disastrous consequences of increasing rates of deforestation in the world's tropical regions also became a matter of serious concern.

Faced with these problems, land-use experts and institutions intensified their search for appropriate strategies that would be socially acceptable, enhance the sustainability of the production base, and meet the need to produce multiple outputs. These collective efforts led to a critical investigation of the shunned and historical farming practices that utilized combinations of trees, crops, and livestock concurrently on the same land unit. The inherent advantages of traditional land-use practices involving trees – sustained yield, environmental conservation, and multiple outputs – were quickly recognized. Agroforestry thus began to come of age in the late 1970s and the establishment in Kenya in 1977 of the International Council for Research in Agroforestry (ICRAF) [www.icraf.cgiar.org], now known as the World Agroforestry Centre, signified the institutionalization of agroforestry and heralded the beginning of the era of agroforestry as a multidisciplinary science.

In the more industrialized regions, many of which are found in the temperate zone, the evolution of "modern" agroforestry was slower than in the tropics. J. Russell Smith's classical work "Tree Crops: A Permanent Agriculture" (Smith, 1950) is considered to have created a new wave of interest in agroforestry; he argued that "an agricultural economy based almost entirely upon annual crops such as corn and wheat is wasteful, destructive of soil fertility and illogical." In the temperate zone, adoptions of highly-technical agricultural practices and aspects of the green revolution have resulted in many environmental and social problems. This ultimately led to much increased research and practical activity in the development of sustainable farming systems and among the prominent disciplines and practices advocated, agroforestry land-use systems emerged as a viable alternative.

The value of temperate agroforestry systems has now been formally acknowledged, and many have been described for specific regions (e.g. Coleman et al., 2018). Many of these systems are common-sense adaptations of historical knowledge that exists on the benefits of incorporating trees into farming systems (see Smith, 1950), but many are new applications of systems that have been successful in other situations. Most modern textbooks on agroforestry confine discussion of temperate systems to individual chapters (Nair et al., 2021; Coleman et al., 2018; Nair and Garrity, 2012), although both research and descriptive information on temperate agroforestry is available for localized regions such as Canada, United Kingdom, Europe, China, Chile, Argentina and New Zealand (Coleman et al., 2018).

Agroforestry: Concepts and principles

Agroforestry is a loosely defined term. Several definitions – long and short – and associated descriptions were proposed for agroforestry during the 1980s and 1990s. However, a consensus has since emerged that agroforestry is one of the terms (including

agriculture, forestry, and several subsets within them) that cannot – and perhaps need not be – defined exactly, and that the concept is more important than the definition. Today, there is a general understanding that agroforestry refers to the purposeful growing of trees and crops in interacting combinations for a range of objectives including both a variety of products or commodities and a vast array of environmental and other ecological services. Basically, it involves the deliberate mixing of trees and shrubs – collectively called woody perennials or trees – on the same unit of land as agricultural crops or animals, either in some form of spatial arrangement or temporal sequence. In the resulting systems, there is significant ecological and economical interaction between the woody and the non-woody components. An agroforestry system will have more than one output, and its ecology and economics will be more complex than in a monocultural agricultural or forest system.

Four key “I” words are sometimes used to denote the essence of agroforestry: intentional, intensive, interactive (biophysical interactions between the perennial and annual components), and integrated (system components are integrated both temporally and spatially) (Garrett et al., 2022). The term “intentional” implies that systems are intentionally designed and managed as a whole unit, and “intensive” means that the systems are intensively managed for productive and protective benefits, although agroforestry based on unimproved (wild) trees is generally not considered intensive. It is often emphasized that all agroforestry systems are characterized by three basic sets of attributes: productivity (production of preferred commodities as well as productivity of land resources), sustainability (conservation of the production potential of the resource base), and adoptability (acceptance of the practice by the farming community or other targeted clientele). These different concepts are captured in ICRAF’s characterization of agroforestry as “a dynamic, ecologically sound system of natural resource management; by integrating trees on farms and in the agricultural landscape, it helps diversify and sustain production for enhanced economic, environmental and social benefits.”

As noted earlier, the motivation for agroforestry research differed between the tropical and the temperate regions. This is to be expected, given that the structure and management of land-use systems are determined not only by biological components, but also by site-specific ecological features and local social and cultural characteristics. This is true for all forms of land use, be it agriculture (including animal production) or forestry. These inescapable differences manifest themselves in the nature, characteristics, objectives, and expectations of agroforestry practices in the two broad geographical regions; nonetheless, agroforestry practices in both regions embrace the same concepts and principles.

In addition to agroforestry, several other terms ending with “forestry” became prominent in the 1970s and 1980s as a consequence of increasing global interest in tree planting activities. These would include terms such as community forestry, social forestry, and farm forestry. Although not defined precisely, these terms emphasize the participation of people in tree planting activities, not necessarily in association with agricultural crops and/or animals as in agroforestry, but with social objectives that rank equally in importance with production objectives. Social forestry refers to using trees and/or tree planting specifically to pursue social objectives, usually the betterment of the poor, through the delivery of the benefits of tree products and/or tree planting to the local people. Community forestry, a form of social forestry, refers to tree planting activities undertaken by a community on communal or so-called common lands; it is based on the direct participation of local inhabitants in the process, either by growing trees themselves, or by processing the tree products locally. Farm forestry, a term used mainly in Asia, refers to tree planting on farms. The major distinction between agroforestry and these other terms is that while agroforestry emphasizes the interactive association between woody perennials and agricultural crops and/or animals, the other terms refer solely to tree planting. Several other terms and initiatives such as permaculture, agroecology, and farming systems that encompass the concept of integrated land use (with or without trees, such as regenerative agriculture) have also emerged in the last several decades. In practice, however, all of these activities and practices that include trees directly or indirectly refer to growing and using trees to provide food, fuelwood, fodder, medicines, building materials, and cash income. Only blurred lines, if any at all, separate them, and they all embrace agroforestry concepts and technologies. Therefore, in common land-use parlance, these different terms are often used as synonyms, and sometimes out of context.

Agroforestry: Practices and systems

A wide range of traditional land-use systems that fit the concept of agroforestry have been identified from different parts of the world and many others being practiced may remain to be identified. The most organized effort to understand the complexity, diversity, and extent of indigenous agroforestry systems in the tropics was a global inventory undertaken by ICRAF between 1982 and 1987. That activity, aimed at systematically collecting, collating, and evaluating data pertaining to such systems around the tropics, assembled for the first time a substantial body of information on many AFS including their structure and function, and their merits and weaknesses. Based on this comprehensive database, a classification scheme for AFS was proposed based on several criteria:

1. The structure of the system, based on the nature and arrangement of the three major groups of components: a. agrisilviculture (trees/crops), b. silvopasture (trees/animals) and c. agrosilvopasture (trees/crops/animals).
2. The primary function of the system: a. production of basic needs (such as food, fodder, fuelwood and cash), and b. protection or service functions (such as soil fertility build-up, environmental amelioration, wind-protection, and erosion control)
3. Spatial arrangement of components, i.e. the time sequence of component arrangement: a. Simultaneous (trees intercropped with crops or livestock), or b. Sequential (trees and crops rotated in sequences).
4. Level of management: a. subsistence to commercial, or b. household consumption to market-oriented

Table 1 Major agroforestry practices in the tropics and the temperate regions.

Agroforestry practice	Brief description
Tropical agroforestry	Fast-growing, preferably leguminous, woody species grown in crop fields; the woody species pruned periodically at low height (<1.0 m) to reduce shading of crops; the prunings applied as mulch into the alleys as a source of organic matter and nutrients, or used as animal fodder.
Alley cropping (hedgerow intercropping)	
Homegardens	Intimate multistory combinations of a large number of various trees and crops in homesteads; livestock may or may not be present.
Improved fallow	Fast-growing, preferably leguminous, woody species planted and left to grow during fallow phases between cropping years for site improvement; woody species may yield economic products.
Multipurpose trees (MPTs) on farms and rangelands	Fruit trees and other MPTs scattered haphazardly or planted in some systematic arrangements in crop or animal production fields; trees provide fruits, fuelwood, fodder, timber, etc.
Silvopasture:	Integrating trees in animal production systems:
• Grazing systems	– Cattle grazing on pasture under widely spaced or scattered trees.
• Cut and carry system (Protein banks)	– Stall-feeding of animals with high-quality fodder from trees grown in blocks on farms.
Shaded perennial-crop systems	Growing shade-tolerant species such as cacao and coffee under or in between overstory shade-, timber-, or other commercial tree crops.
Shelterbelts and windbreaks	Use of trees to protect fields from wind damage, sea encroachment, floods, etc.
Taungya	Growing agricultural crops during the early stages of establishment of forestry (timber) plantations.
Temperate-zone agroforestry	Trees planted in single or grouped rows in herbaceous (agricultural or horticultural) crops in the wide alleys between the tree rows.
Alley cropping	
Forest farming	Utilizing forested areas for producing specialty crops that are sold for medicinal, ornamental, or culinary uses.
Riparian buffer strips	Strips of perennial vegetation (tree/shrub/grass) planted between croplands/pastures and streams, lakes, wetlands, ponds, etc.
Silvopasture	Combining trees with forage (pasture or hay) and livestock production.
Windbreaks	Row trees around farms and fields, planted and managed as part of crop or livestock operation to protect crops, animals, and soil from wind hazards.
Bioenergy	Short-rotation woody crops such as willow (<i>Salix</i> spp.), miscanthus (<i>Miscanthus</i> spp.) or switchgrass (<i>Panicum virgatum</i>), grown in conjunction with other on-farm agroforestry practices such as intercropping, or as a sole-crop

Source: Modified from Nair PKR, Gordon AM, and Mosquera-Losada MR (2008) Agroforestry. In: Jorgensen SE and Fath BD (eds.) Ecological Engineering. pp. 101–110. Oxford, U.K: Elsevier.

Other classification schemes of AFS have also been proposed, but essentially, they all are based upon the above criteria and concepts. Various characterizations and descriptions of AFS such as agrisilvicultural systems for fuelwood production in semiarid lands, silvopastoral systems for animal production on sloping lands, multistrata homegardens in humid tropics, etc., are common in the agroforestry literature. Moreover, descriptions of existing systems as well as recommendations concerning potential agroforestry technologies for specific agroecological zones include a mixture of various forms of agroforestry in terms of the nature and arrangement of components, and often, several AFS can be found within the same ecological regions. Thus, in general, any specific agroforestry practice cannot be identified exclusively with a specific ecological region, and vice versa.

A list of major global agroforestry practices is presented in Table 1. Local adaptations and manifestations of these (and other) practices exist as innumerable agroforestry systems with site-specific characteristics, especially in the tropics and subtropics. Compared with the tropics, agroforestry practices and systems in the temperate zone are less diverse and complex. The Association for Temperate Agroforestry (AFTA: www.aftaweb.org) has recognized five major agroforestry practices in North America: Alley cropping, forest farming, riparian buffer strips, silvopasture, and windbreaks. Other temperate-zone AFS include ancient tree-based agricultural systems involving a large number of multipurpose trees such as chestnuts (*Castanea* spp.), oaks (*Quercus* spp.), carob (*Ceratonia siliqua* L.), olive (*Olea europaea* L.), and figs (*Ficus* spp.) in the Mediterranean region. The dehesa system, grazing under oak trees with strong linkages to recurrent cereal cropping in rangelands, is also a very old European practice. In addition, bioenergy systems, where fast-growing woody perennials are grown in short-rotation within an agricultural context, would also qualify as an agroforestry system.

Distribution and area of agroforestry systems

The nature and intensity of the type of agroforestry practiced in a locality or region is determined primarily by its ecological potential and socioeconomic conditions. Humid lowlands and other tropical regions of the world that are not constrained by climatic limitations such as extreme temperatures and long dry seasons are home to a large diversity of plant species and therefore numerous and diverse agroforestry systems. Moreover, socioeconomic factors such as human population pressure, greater availability of labor, smaller land-holding size, complex land tenure, and lack of proximity to markets often support more and diverse agroforestry systems in such regions. In general, small family-farms, subsistence food crops, and emphasis on the role of trees in improving soil

quality of agricultural lands are characteristic of tropical agroforestry systems. On the other hand, environmental protection is the main motivation for agroforestry in temperate regions, where intensive, mechanized monocultural production systems of agriculture and forestry have contributed to reduced biodiversity and loss of forest resources and wildlife habitat, and increased environmental hazards such as erosion, non-point source pollution of ground water and rivers, and greenhouse-gas emissions.

As noted above, multistrata systems (homegardens, shaded perennial systems, and multilayer tree gardens) are common in tropical regions with high human population, while less intensive systems are common in areas with lower population density. In the moist tropical and dry regions, the nature of agroforestry systems is influenced by climate and population pressure: homegardens and multilayer tree gardens are found in the relatively wetter areas; tree intercropping systems such as multipurpose trees on croplands, silvopastoral (grazing) systems, and protective systems of windbreaks and shelterbelts are common in these regions. In the tropical montane regions that have favorable rainfall regimes, sloping lands and rolling topography make soil erosion an issue of major concern; consequently, soil conservation is one of the main objectives of agroforestry in these regions. Shaded perennial systems, the use of woody perennials in soil conservation, and silvopastoral systems are the major forms of agroforestry in such tropical highlands. Several other specific systems also exist in the tropics such as apiculture with trees, and aquaculture involving trees and shrubs; however, information on these systems was too meager to warrant their inclusion here.

Estimating the area under AFS on a farm itself is challenging because of the lack of clarity and proper procedures for delineating the area influenced by trees in a mixed stand of trees and crops (Nair et al., 2009). In simultaneous systems, the entire area occupied by multistrata systems such as homegardens and shaded perennial systems and intensive tree-intercropping situations can be listed as agroforestry. However, most AFS are rather extensive, where the components, especially trees, are not planted at regular spacing or density such as in parkland systems and extensive silvopastures. Such situations exist also in Europe and other temperate regions where the mosaics of intensive field systems, with hedgerows and patches of woodland are common features of agricultural landscapes. The problem of estimating the area under agroforestry is more difficult in the case of practices such as windbreaks and boundary planting where the trees are planted at wide distances between rows (windbreaks) or around agricultural or pastoral parcels (boundary planting), because the influence of trees both above and below ground extends beyond the obviously visible area of trees. For windbreaks, the rule of thumb is that the area protected from wind erosion extends laterally to 10 times the height of trees in the central core of windbreaks. The problem has a different dimension of difficulty when it comes to sequential tropical systems such as improved fallows and shifting cultivation. In such situations, the beneficial effect of trees and other woody vegetation (in the fallow phase) on the crops that follow them (in the cropping phase) is believed to last for a variable length of time (years). Despite these temporal and spatial issues, a recent ICRAF survey (www.worldagroforestry.org/advanced-search?search_api_fulltext=high-resolution+remote-sensing+) using high-resolution remote-sensing quantified the areas of agricultural landscapes with at least 10% tree cover worldwide as nearly a billion hectares. Considering the prevalence of agroforestry practices in non-agricultural lands, the area currently under agroforestry is estimated to be 1.6 billion ha globally (Table 1).

Brief description of major agroforestry systems

Intercropping (alley cropping and other forms of tree intercropping)

In the tropics, alley cropping was developed by researchers in the 1970s and 1980s in an attempt to find alternatives to the long-term fallows of shifting cultivation. The main motivation and rationale for alley cropping was the perceived soil-improving potential of fast-growing, leguminous woody species and their ability to withstand repeated cutting. It was hypothesized that by integrating these woody species into food crop production systems in simultaneous combinations, some of the benefits of natural long-term fallows such as recycling nutrients, suppressing weeds, and controlling erosion on sloping land, could be realized without having to keep the land out of agricultural production. The practice consists of growing arable crops between hedgerows of planted shrubs and trees, preferably leguminous species that are periodically pruned to prevent shading to crops (Fig. 1). The biomass obtained by pruning the trees and shrubs during the crop-growing season is put back to the alleys as a source of mulch and green manure. Besides, leguminous woody species add nitrogen to the system through biological nitrogen fixation (Table 2).

Many studies on alley cropping, of both biophysical and socioeconomic nature, were undertaken in various parts of the tropics during the 1980s and 1990s. Indeed, alley cropping was the most widely researched topic in tropical agroforestry during that period. In interpreting the results of these studies, some experts have used the data to defend alley cropping, others to denigrate it. But, the merits or demerits of alley cropping, or for that matter, any agroforestry technology, cannot be judged on the basis of any single criterion or short-term result. Benefits of yield and economic returns from multiple products such as fuelwood and fodder, and the possibility for avoiding the risk associated with monocultural production systems must be carefully weighed against drawbacks.

Relatively limited ecological adaptability and comparatively higher labor demand are the two key issues concerning alley cropping. The provision of nutrients through decomposing mulch, a basic feature of alley cropping, depends on the quantity of the mulch as well as on its quality and time of application. If the ecological conditions do not favor the production of sufficient quantities of mulch (as is the case in dry tropics), there is no perceptible advantage in practicing alley cropping. The issue of high labor demand is associated with the pruning of the shrubs and the incorporation of the biomass in the field. There are also some issues of pest management problems of mixed species stands as well as newly introduced species, and tree-crop competition for growth resources such as nutrients, water, and light. All these issues have contributed to poor rates of farmer adoption of alley-cropping technology.



Fig. 1 Tropical alley cropping. Fast-growing, coppicing, nitrogen-fixing shrubs and trees (“Fertilizer Trees”) are grown in rows 4–8 m apart in crop fields and pruned periodically (4–6 week intervals); the succulent and easily decomposable tree biomass so obtained is returned to the cropped alleys as a source of nutrient for crops. The photo shows *Gliricidia sepium* grown with maize (*Zea mays*), a practice followed by thousands of farmers in eastern and southern Africa. Photo credit: ICRAF, Nairobi, Kenya.

Table 2 Global distribution and area of different agroforestry system sub-groups^a.

AFS sub-group	Major AF practices	Distribution (major agro-ecol./geographical regions) ^b		Approx. area (million ha) ^b	
		Tropical	Temperate	Tropical	Temperate
Multistrata systems	Homegardens	Humid: wet, moist, and montane (rainfall >1000 mm year ¹)		100	
Tree intercropping	Shaded perennials		Forest farming		
	Alley cropping	Rainfall >800 mm year ¹	??	50	50
Silvopasture	Trees on farmlands	Throughout tropics	??	550	50
	Cut-and –carry and browsing	Wet and moist; rainfall >1000 mm year ¹		300	150
	Grazing under trees	Semi-arid to arid	N. America, Europe, subtropical highlands		
Protective systems	Windbreaks, shelterbelts	Semi-arid and arid lands; coastal areas	N. America, Europe, China	200	100
	Soil conservation hedges	Sloping lands in higher rainfall areas	N. America (Riparian buffer strips)		
	Boundary planting	Throughout	Windbreaks		
Agroforestry woodlots	Firewood and fodder	Drylands		50	
	Land reclamation	Degraded lands (eroded, salt-affected)			
TOTAL				1250	350

^aEstimates based on reported values in the literature.

^bIncludes potential areas for adoption.

Source: Nair PKR and Garrity D (2012) *Agroforestry – The Future of Global Land Use, Advances in Agroforestry*. Springer Dordrecht Heidelberg New York London.

Improved fallow with “fertilizer trees” is a related agroforestry technology that has been promoted in nutrient-depleted African soils. Basically, one or a few (mixed) tree species, usually biological nitrogen fixers, are planted at high density as a substitute to natural fallow, to achieve the benefits of the latter in a shorter time of 2–3 years versus 15–20 years of natural fallow. In this case also, the production of biomass in planted fallows as well as the potential of planted fallows to ameliorate soil fertility is controlled by several factors: environmental conditions, soil type, land degradation, length of the fallow period, density of tree planting, tree management, and soil and climatic conditions.

In temperate regions, alley cropping refers to growing herbaceous (agricultural) crops in the alleys between rows of widely-planted trees that are planted in single or grouped rows (Fig. 2). This arrangement allows the use of farm machinery and equipment and reduces the need for manual labor. Black walnut (*Juglans nigra*), a valuable timber- and nut-producing tree is the species most widely used in alley cropping in North America. The common spacing adopted is 12.5 m between tree rows and 3 m between trees within a row (270 trees ha⁻¹). The driving force behind most temperate-zone alley cropping are the economic benefits from the intercrop. Soil



Fig. 2 Alley cropping system in Guelph, Ontario, Canada, where field crops [e.g. Corn (*Zea mays*) are grown between widely spaced (15 m) walnut (*Juglans nigra*) tree rows]. Photo credit: Naresh Thevathasan.

conservation in gently sloping lands could be an additional advantage. Various aspects of this practice have received considerable research attention in the eastern and Midwestern USA and Canada. Other similar systems in the temperate regions include traditional intercropping practices used in the establishment of fruit and nut trees including olive (*Olea* spp.) and grapes (*Vitis* spp.) in Europe, with pecan (*Carya illinoensis*) trees in southeastern U. S., and with paulownia (*Paulownia* spp.) trees in China.

The most widespread form of agroforestry, in terms of the area involved, is perhaps the extensive intercropping under scattered stands of trees, particularly in the semiarid regions of sub-Saharan Africa and Asia. The so-called parkland system of sub-Saharan Africa (Fig. 3) extends across the entire semiarid belt from Senegal in the west to Ethiopia and Kenya in the east. It originated from the practice of the traditional shifting cultivation in which farmers conserved in their fields the trees that are desirable for shade, fodder, fruits, and medicines. The trees that are most common include *Faidherbia albida*, *Parkia biglobosa*, and *Vitellaria paradoxa*. The understory species include both crops such as millet (*Pennisetum glaucum*), cassava (*Manihot esculenta*), cotton (*Gossypium* sp.), maize and others depending on climatic and soil conditions, as well as grasses for animal grazing, where animals are let loose for grazing over crops residues and grasses. The practice is therefore considered as extensive intercropping or silvopasture.

Multistrata systems

Intensive, multispecies, tree-based farming systems such as homegardens, shaded perennial stands, and other complex agroforests are common agroforestry practices in the humid and subhumid parts of the tropics, especially in the lowlands (Fig. 4). The structural and functional diversity and various other characteristics of these systems have been well described and reviewed. Homegardens have a long tradition of providing food and nutritional security as well as environmental sustainability in smallholder production systems, often in heavily populated regions of lowland humid tropics in South- and Southeast Asia, and to a small extent in other tropical and subtropical regions. Although a few systems that have some similarities to tropical homegardens can be found in parts of the temperate regions as well (e.g., the satoyama system in Japan), intensive multispecies homegardens are a unique agroecosystem of the tropics.

Shaded-perennial systems that involve growing tree crops such as coffee (*Coffea* sp.) and cacao (*Theobroma cacao*) under the shade of overstorey tree species (Fig. 5) represent another traditional example of high-intensity crop combination that has some unique ecological features and commercial value. In addition to coffee and cacao, several of the tropical fruit-and-nut producing tree species that are harvested annually or at shorter intervals are often grown in association with understory- or overstorey species. In some situations, the species combinations and management features of these systems are very similar to those of homegardens such that, at the landscape- and village level, there is a continuum of plant associations from homegardens nearer homes to multistrata tree gardens away from the homes.

Characterized as the epitome of sustainability, these multispecies tree-crop combinations are excellent examples of efficient land-management systems by smallholder farmers. With heavy reliance on human (often family) labor for the farm operations, conventional tillage operations involving machinery are minimal or absent in such systems. Some of the distinguishing ecosystem sustainability features of these systems include efficient and “closed” nutrient cycling facilitated by continuous litterfall and decomposition and very little export of nutrients from the system by way of harvested products, the reliance on organic manure



Fig. 3 The Parkland system of sub-Saharan Africa. Deliberate integration of various trees and shrubs on crop- and grazing lands is a dominant traditional practice in extensive semiarid to sub-humid regions in sub-Saharan Africa. The trees provide food, fuel, fodder, medicinal products, building materials, and saleable commodities. The picture shows the shea tree (*Vitellaria paradoxa*) along with sorghum (*Sorghum bicolor*) crop. Inset: Women carrying head-loads of harvested shea fruits for processing. Photo credit: Peter Lovett.



Fig. 4 Homegardens of Jamaica. Homegardens that consist of intimate, multistory combinations of various trees and crops around homesteads provide food and cash security as well as social and esthetic benefits to millions of farm households around the tropics. The picture shows a homegarden with various fruit- and nut-producing trees such as breadfruit tree (*Artocarpus*), mango (*Mangifera indica*), and papaya (*Carica papaya*); and other food- and cash crops such as plantains (*Musa* spp.) and coffee (*Coffea* sp.) in a homegarden in Jamaica. Photo credit: P.K.R Nair.



Fig. 5 *Coffea* sp. with *Erythrina* in Costa Rica. Shaded perennial system. Shaded perennial-crop systems involve growing shade-tolerant commercial species such as coffee (*Coffea* sp.) and cacao (*Theobroma cacao*) under overstory shade trees. The photo shows coffee (dark-colored bush) under *Erythrina poeppigiana*, a leguminous tree (light-colored overstory) in Costa Rica. The tree is pollarded (pruned) two to three times a year to regulate shade received by the coffee bush.

and plant materials with consequent avoidance of chemical-fertilizer use, and predominance of deep-rooted trees, which collectively contribute to the high levels of C sequestration and climate-change mitigation in these systems, as evidenced by the recent studies in the homegardens of Kerala, India, and the shaded cacao systems of Bahia, Brazil.

Protective agroforestry systems

These systems encompass the use of trees and shrubs for exploiting their ecosystem-protection benefits by planting them as windbreaks, riparian buffers, soil conservation hedges, and similar other configurations. Windbreak practices include shelterbelts, timber belts, and hedgerows, and are planted and managed as part of a crop or livestock operation. Field windbreaks are used to protect a variety of wind-sensitive row, forage, tree, and vine crops, to control wind erosion, and to provide other benefits such as improved bee pollination of crops and wildlife habitat (Brandle et al., 2009). For livestock, windbreaks help reduce animal stress and mortality, feed and water consumption, and odor, while timber belts are managed windbreaks designed to increase the value of the forestry component. Shelterbelts are also planted along coastal areas to reduce the impact of sea encroachment and protect crops from salt-water damage. Riparian and “upland” buffers are strips of permanent vegetation, consisting of trees, shrubs, and grasses that are planted and managed together. Riparian buffers are placed between agricultural land (usually crop land or pastureland) and water bodies (rivers, streams, creeks, lakes, wetlands, etc.) to reduce runoff and non-point source pollution, stabilize stream banks, improve aquatic and terrestrial habitats, and provide harvestable products (Fig. 6). Upland buffers are placed along the contour within agricultural crop lands to reduce runoff and non-point source pollution, improve internal drainage, enhance infiltration, create wildlife habitat and connective travel corridors, and provide harvestable products. As discussed under alley cropping, the frequently pruned rows of trees and shrubs planted across the contour in crop production fields help reduce soil erosion and the pruned biomass serves as a source of nutrient to crops or can be used as animal fodder. Riparian buffers are a common feature of the landscape in the U.S. North Central region as well as in Canada. However, it is not a common land-use system in the European countries, and therefore the need to enhance riparian plantings in Europe has been discussed by Englund et al. (2021).

Silvopasture

Silvopasture that combines trees, forages, and shrubs/trees with livestock operations is another type of agroforestry practice that is popular in both tropical and temperate regions. Broadly, there are two major forms of silvopasture: grazing- and tree-fodder systems. In grazing systems, cattle are allowed to graze on pasture under widely spaced or scattered trees (Fig. 7), whereas in the tree-fodder systems, the animals are stall-fed with fodder from trees or shrubs grown in blocks (fodder banks) on farms. Most silvopastoral systems in Africa and other developing regions of the world involve extensive open grazing by free-roaming animals under scattered natural stands of trees and shrubs mostly in semiarid to arid areas, as in the parklands of sub-Saharan Africa. More intensive grazing systems of silvopasture are practiced in Latin America where animals are penned in parcels of land with barbed wire living-fence and grazing is regulated. Such organized “silvopastoral systems are also becoming popular in the extensive Cerrado



Fig. 6 Riparian buffer. Riparian Buffer: Bear Creek National Restoration Watershed (designated under the Clean Water Action Plan, 1999) and the USDA designated Bear Creek National Research and Demonstration Site. The pictures are of various kinds of riparian buffers along Bear Creek. Photo credit: The Iowa State University NREM Riparian Buffer Team, ISU, USA.



Fig. 7 Silvopasture in Brazil: Beef cattle grazing on *Brachiaria brizantha* (grass) grown under *Eucalyptus* hybrid trees on a farm in Paracatu, Minas Gerais, Brazil.

region of Brazil. The most intensive silvopastoral system is the stall feeding of animals by fodder from trees grown elsewhere, which is a very common practice in smallholder farming systems of Africa and South America. The grazing system of silvopasture has recently gained prominence as an environmentally desirable approach to managing degraded pasture lands in industrialized countries. Silvopasture embodies the goals of desirable management practices for climate-change mitigation (especially in terms of grazing and manure addition) and C sequestration. Additional benefits of silvopasture include water quality improvement, soil conservation, esthetics, and providing shade to cattle (animal welfare). Alternative land uses including sustainable forest management, outdoor recreation, and ecotourism, and most encouragingly silvopasture, are considered highly compatible with traditional ranching and includes several elements of best management practices for ranchers.

Tree woodlots and specialty crops

These terms are used to denote agroforestry practices that are undertaken for special situations and needs. Examples include growing tree woodlots as fodder banks (for production of cut-and-carry tree-fodder), boundary planting of trees for production of firewood, small timber, poles and fence posts; tree planting for reclamation of degraded lands such as saline soils and mined land; establishing tree woodlots for biomass and bioenergy production such as in the Canadian prairies, growing specialty products such as ornamentals, honey, and high value crops (e.g., ginseng, shiitake mushrooms) for niche markets; and a whole host of such “non-conventional” land-use systems. The extent of people and areas involved and the economic benefits derived from such activities are seldom documented. Among these activities only large-scale tree woodlots would count as major activities in terms of carbon sequestration and climate change mitigation, although all would be important for their economic, social, and cultural benefits.

One of the best ways of reclaiming areas degraded by severe soil erosion, acidity, salinity or alkalinity, and other forms of land degradation is by planting tree species that are adapted to such conditions and leaving the land under the tree woodlot for several years to 20 or more, depending upon local conditions, before planting crops or fodder grasses is considered. Several such programs, under the banner “wasteland development,” supported by various government agencies and with the involvement of nongovernmental organizations and local communities are popular in India. Another example of large-scale implementation of protective agroforestry is the “fallow regeneration” that is gaining momentum in different parts of arid and semiarid sub-Saharan Africa. For example, the recent rejuvenation of the centuries-old parkland agroforestry systems of the Sahel, through the spread of farmer-managed natural regeneration has been documented to have occurred on nearly 5 million hectares in Niger alone.

Agroforestry and ecosystem services

The concept of agroforestry is based on the expected role of on-farm and off-farm tree production in supporting sustainable land-use and natural-resource management. While the aboveground and belowground diversity provides more stability and resilience for the system at the site level, the system provides connectivity with forests and other landscape features at the landscape and watershed levels. These ecological foundations manifest themselves in providing environmental quality and ecosystem-service functions, those commonly associated with AFS being soil improvement, water quality enhancement, biodiversity conservation, and C sequestration/climate change mitigation and adaptation.

Soil improvement

One of the tree-mediated benefits of considerable advantage in the tropics is that trees and other vegetation improve the productivity of the soil beneath them. Research results during the past three decades show that three main tree-mediated processes determine the extent and rate of soil improvement in agroforestry systems. These are: (i) increased nitrogen input by N_2 -fixing trees, (ii) enhanced availability of nutrients resulting from production and decomposition of tree biomass, and (iii) greater uptake and utilization of nutrients from deeper layers of soils by trees, the roots of which extend much deeper into the soil than roots of common crops; however, the research base on role of trees in the uptake of nutrients from different soil layers in agroforestry systems is comparatively limited.

Nitrogen-fixing trees (NFTs) and other “fertilizer trees” are a valuable resource in agroforestry systems. Farmlands in many parts of the developing world generally suffer from the continuous depletion of nutrients as farmers harvest without adequately fertilizing or fallowing the land. One promising way for overcoming the acute problem of the low nutrient status of soils, such as African soils in general, is to enable smallholders to use fertilizer-tree systems (Fig. 1) that increase on-farm food production. Some of the widely held assumptions about their benefits could, however, be wrong or incomplete. Because of methodological difficulties in quantifying N_2 fixation, especially in older trees, our understanding of the extent of N_2 fixation, and, therefore, the benefit that is actually realized by using NFTs in agroforestry systems is unsatisfactory. Furthermore, it is not clearly understood how much of the N_2 that is fixed by a NFT is actually utilized or potentially made available to an associated crop during its growth cycle, and how much goes into the soil’s N store for eventual use by subsequent crops.

Biomass decomposition patterns vary greatly among agroforestry tree species. Several biomass (litter)-quality parameters, based on the chemical composition of plant tissues, have been developed to interpret these patterns: ratios of C to N, polyphenols to N, lignin to N, and (polyphenols/lignin) to N. Using this information, we can develop management strategies to manipulate the decomposition of plant biomass in agroforestry systems, thereby regulating the rates of nutrient release in the short term, and improving soil fertility, via improved soil organic matter status, in the long term. Compared to our knowledge of aboveground biomass additions in agroforestry systems, much less is known about the dynamics of belowground biomass.

The other major avenue of soil improvement in agroforestry is through soil conservation. When properly designed and managed, agroforestry techniques can contribute to ecosystem protection and restoration functions by reducing water- and wind erosion and enhancing soil productivity.

The presence of deep-rooted trees in the system can contribute to improved soil physical conditions and higher soil microbiological activities under agroforestry. About 2 billion ha of land, a third of total farmland, in developing nations are estimated to be degraded through erosion, salinity, and fertility depletion. The potential of agroforestry to reduce the hazards of erosion and desertification as well as to rehabilitate such degraded land and to conserve soil and water has been well recognized. The soil ameliorative potential of agroforestry systems has recently been demonstrated in the temperate zone as well.

Water quality enhancement and nutrient interception

The so-called “safety-net” effect of tree roots – the ability of deep-rooted trees to absorb nutrients that have leached below the rooting zone of agronomic crops and recycling them via leaf litter and fine root turnover and thus improving the nutrient-use efficiency in the system as a whole – could have an important application in the heavily fertilized sandy soils that have low nutrient retention capacities. The capacity of tree roots to capture nutrients from the deeper soil horizons can enhance nutrient storage in the plant-soil system and thereby reduce the amount of nutrients that might otherwise be transported to ground and surface water through runoff and leaching. Research during the past decade has shown that riparian forest buffers can remove significant amounts of sediment, nutrients, and pesticides from both surface and subsurface waters and thus reduce non-point-source pollution of water bodies in the industrialized regions.

However, if riparian buffers are not established on all agricultural streams within a watershed, the nutrient levels in the stream can remain high and encourage eutrophic conditions. For example, in a study in Washington Creek Ontario, Canada (Plascencia-Escalante, 2008), upstream nutrient loading resulted in high nitrate and phosphorus levels in water adjacent to riparian plantings. However, the shade created by the riparian trees reduced the solar radiation incidence on the water and thereby prevented periphyton growth or algae bloom. The study suggested that algae blooms can be prevented even under high nutrient levels in the water if adequate shade is provided by the riparian trees. In the same study, algae blooms were observed in the same stream where there was no shade or riparian tree planting. Fig. 8 depicts the gap fractions associated with two riparian buffer types; natural forest and rehabilitated riparian buffer.

Biodiversity conservation

When landscapes are increasingly being fragmented and remaining patches of natural vegetation are reduced to isolated habitat islands consequent to human activities, mixed species AFS could play a significant role in helping to maintain a higher level of biodiversity and provide greater landscape connectivity. This can occur in at least three ways: (i) intensification of AFS leading to reduced exploitation of protected areas, (ii) increasing biodiversity in working landscapes through the expansion of AFS into traditional farmlands, and (iii) increasing the species diversity of trees in farming systems. Where croplands occupy most of the landscape, riparian forest buffers and field shelterbelts can be essential for maintaining plant and animal biodiversity, especially under a changing climate scenario. The trade-offs between ecosystem conservation and agricultural production can convincingly be addressed by shifting the view from the plot to the landscape scale and integrating biodiversity-friendly land-use systems such as agroforestry into development strategies.

Carbon sequestration and climate change mitigation

Agroforestry systems are often perceived to have higher potential to sequester carbon (C) than comparable single-species crop or pasture systems. The underlying premise is the niche complementarity hypothesis, which states that a larger array of species in a system leads to a broader spectrum of resource utilization, making the system more resilient, and implies that plant species in a mixed system use resources in a complementary way. The estimates of global annual C storage in AFS range from 0.29 to 15.21 Mg ha⁻¹ year⁻¹ above-ground, resulting in pools of 30–300 Mg C ha⁻¹ up to 1 m depth in the soil (Table 3). Recent studies under various AFS in diverse ecological conditions have shown that tree-based agricultural systems, compared to treeless systems, stored more C in deeper soil layers near the tree than away from the tree, and higher soil organic C content was associated with higher species richness and tree density. Furthermore, C3 plants (trees) have been found to contribute to more stored C in the silt clay fractions (<53 mm diameter) that constitute more stable C than C4 plants, in deeper soil profiles.

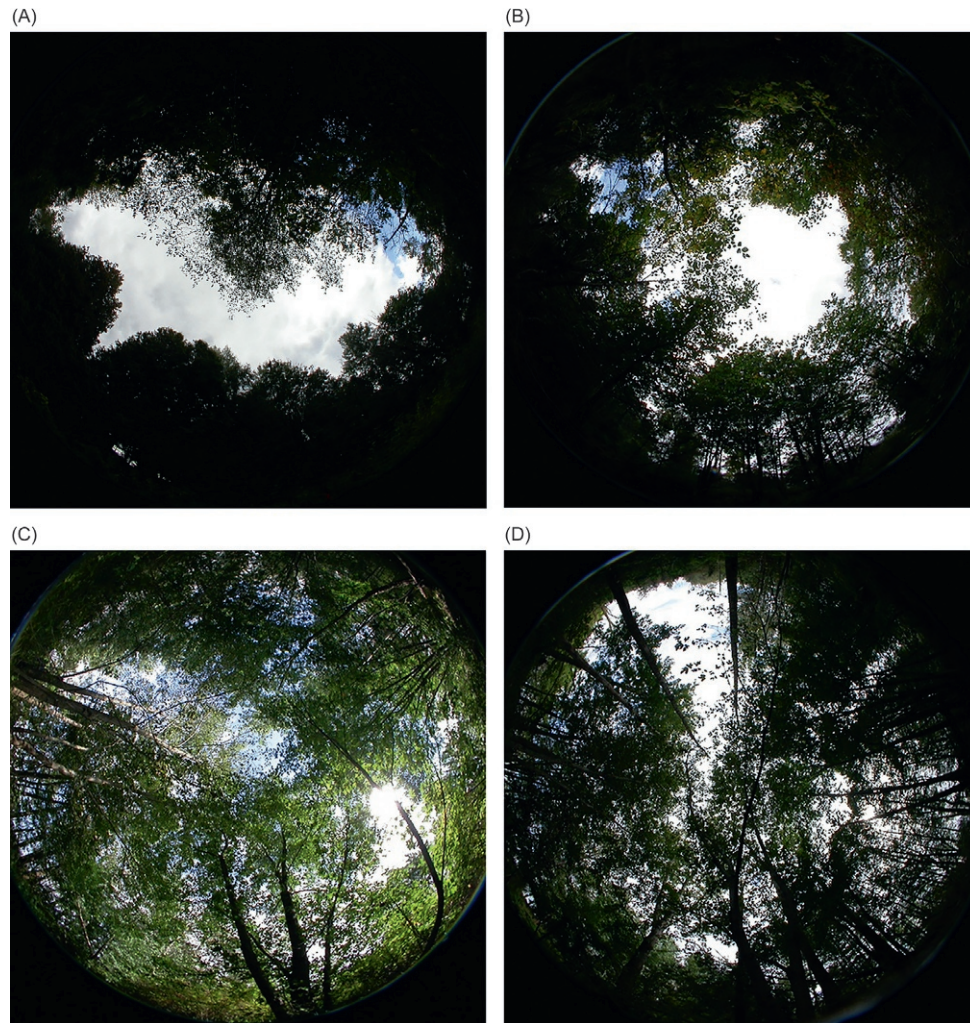


Fig. 8 Hemispherical photos taken under the water surface under two Natural Forest (NF) and two Rehabilitated Area (RA) sites at Washington Creek. Both land sites had over-stream riparian vegetation. (A) NF 1 Total gap fraction: 19.73% and Canopy cover: 81.7% (B) NF 2 Total gap fraction: 14.30% and Canopy cover: 86.7% (C) RA 1 Total gap fraction: 12.21% and Canopy cover: 87.79% (D) RA 2 Total gap fraction: 12.49% and Canopy cover: 89.51%. From: Plascencia-Escalante FO (2008) An analysis of some components of the nitrogen cycle as affected by land use adjacent to the riparian zone of a southern Ontario stream. PhD Thesis. Guelph, ON: University of Guelph.

Table 3 Estimates of carbon stocks and carbon sequestration potential (CSP) in agroforestry systems in different ecological regions of the tropics.

Agroforestry system	Major AF practices	Estimated C stock range (Mg ha^{-1})		C sequestration potential ($\text{Mg ha}^{-1} \text{ yr}^{-1}$)	
		Above ground	Belowground (esp. soils)	Above ground	Below ground (soils)
Multistrata systems	Homegardens	2–18	Up to 200	0.5–3	1.5–3.5
	Shaded perennials	5–15	Up to 300	1–4	1.0–5
Tree Intercropping	Alley cropping	Up to 15	Very low to 150	0.5–4	1.5–3.5
	Multipurpose trees on farmlands	Up to 12	Very low to 150	0.2–2.5	1.5–3.5
Silvopasture	Tree fodder (Browsing, cut-and-carry)	1.8–3	Very low to 80	0.3–4	1.0–2.5
	Grazing under trees	1.5–8	Very low to 60	0.3–2	0.4–1.0
Protective systems	Windbreaks, Shelterbelts, Soil conservation hedges, Boundary planting	1.5–7	Very low to 60	0.7–2	0.4–1.0
Agroforestry Tree Woodlots	Woodlots for firewood, fodder, land reclamation	1.5–7	Very low to 60	1–5	1.0–6.0

Source: Nair PKR and Garrity D (2012) *Agroforestry – The Future of Global Land Use, Advances in Agroforestry*. Springer Dordrecht Heidelberg New York London.

The amount of C sequestered in AFS depends to a great extent on environmental conditions and system management. Based on these outcomes, the potential role that AFS can contribute to climate-change mitigation and adaptation strategies is now being recognized and pursued in many development programs around the world.

The expectations of AFS in terms of environmental benefits are different between tropical and temperate regions. In the tropics where small family farms, subsistence food crops, and declining soil productivity are common features, the emphasis on the role of trees in numerous and diverse AFS is on improving the soil quality of agricultural lands and providing food and nutritional security. On the other hand, environmental protection is the cornerstone of agroforestry motivations and applications in temperate (industrialized) regions; it needs to be noted however, that although there is a tendency to equate “temperate” with “industrialized” regions, some temperate countries (or temperate regions of countries) such as China, Russia, Chile, and Argentina may be less motivated by environmental concerns than countries such as Canada, the US, Europe, New Zealand and Australia. In the tropics, soil-related issues such as soil fertility improvement, soil and water conservation, and reclamation of saline-, acid-, and other degraded soils have been major areas of agroforestry research and application during the 1980s and 1990s. Although these are environmental quality issues, they were pursued mainly as a means of augmenting the soil’s capacity to produce crops sustainably. In temperate regions, on the other hand, agroforestry was seen more as an environmental protection opportunity through the exploitation of the beneficial role of trees for water-quality enhancement in agricultural watersheds, reduced wind and water erosion, etc. In both situations, however, the focus was at the local or regional scale. During the past several decades, environmental issues of a global scale such as carbon sequestration/climate change and biodiversity conservation have become increasingly important.

System-level carbon sequestration in temperate agroforestry land-use systems (Canada)

Tree-based intercropping (alley cropping) system – A case study

The case study results presented here are from a long-term tree-based intercropping research project that was initiated in 1987 on a 30-ha site at the University of Guelph Agroforestry Research Station, Ontario, Canada (43° 32' 28" N latitude, 80° 12' 32" W longitude). Ten tree species namely, silver maple (*Acer saccharinum*), Hazelnut (*Corylus avellana*), white ash (*Fraxinus americana*), black walnut (*Juglans nigra*), Norway spruce (*Picea abies*), Hybrid poplar (*Populus* spp.), red oak (*Quercus rubra*), black locust (*Robinia pseudoacacia*), willow (*Salix discolor*), and eastern white cedar (*Thuja occidentalis*) were planted and annually intercropped with either corn (*Zea mays*), soybean (*Glycine max*), or winter wheat (*Triticum aestivum*). Tree rows were spaced at 12.5 or 15 m apart with a within-row tree spacing of 3 m or 6 m. The soil type is sandy loam (Typic Hapludalf). Crops were planted between the tree rows every year according to local ‘standard’ cultural practices.

Soil organic carbon (SOC) gains and carbon sequestration in trees were monitored from the early 1990s to 2018, a period of over 25 years. The high rate of poplar leaf biomass addition ($1 \text{ Mg C ha}^{-1} \text{ year}^{-1}$; Thevathasan and Gordon, 1997) over a total period of 8 years resulted in an absolute increase of SOC of approximately 1% close to the tree row at 15 cm soil depth, and this effect extended into the alley for up to approximately 4 m. This is about a 35% relative increase (% difference between 3.25 and the mean of 2.32 and 2.50) in SOC close to the tree rows over the given period of time and reflects leaf biomass inputs in the early 1990s, when trees were small and the major portion of litterfall was distributed close to the tree row (2–3 m) (Table 4). By 2002, the poplar trees were 14 m tall and leaf biomass was evenly distributed across the crop alley up to distances of 15 m. This resulted in a slow but inexorable increase in SOC in the middle of the crop alley, as illustrated in Fig. 9.

At the age of 25-years, the same system was investigated for carbon sequestration in soil and in trees by Wotherspoon et al. (2014). The study investigated hybrid poplar, Norway spruce, red oak, black walnut and white cedar when they were intercropped with soybean. Total C stocks for poplar, cedar, oak, walnut, spruce and a soybean sole-cropping system were 113.4, 99.4, 99.2, 91.5, 91.3, and 71.1 t C ha⁻¹, respectively at a tree density of 111 trees ha⁻¹, including mean tree C content and soil organic C stocks (Fig. 10). Net C flux for poplar, spruce, oak, walnut, cedar, and soybean sole-crop were +2.1, +1.6, +0.8, +1.8, +1.4 and – 1.2 t C ha⁻¹, y⁻¹, respectively. Results suggest that integrating these five tree species into a soybean agricultural system contributed to

Table 4 Poplar (*Populus* spp.) litterfall distribution in a poplar–barley intercropping system during the 1993 and 1994 growing seasons when trees were 6 and 7 years old respectively, Guelph, Ontario, Canada.

Distance from the poplar tree row (m)	Litterfall biomass ($\text{Mg ha}^{-1} \text{ y}^{-1}$)	
	1993	1994
0–2.5	2.67 ± 0.04	2.76 ± 0.14
2.5–6.0	0.52 ± 0.05	0.61 ± 0.06

Source: Thevathasan NV and Gordon AM (1997) Poplar leaf biomass distribution and N dynamics in a poplar–barley intercropping system in southern Ontario. *Agroforestry Systems* 37: 79–90.

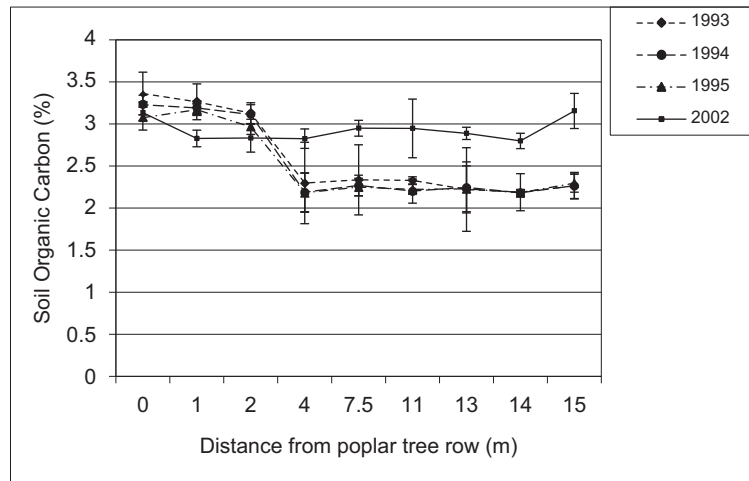


Fig. 9 Soil organic carbon content at various distances from poplar tree row in 1993, 1994, 1995 and 2002, when the trees were 6, 7, 8 and 15 years old, in the tree-intercropping experiment in southern Ontario, Canada. Error bars that overlap indicate that associated values are not significant at $P < 0.05$. Source: Thevathasan NV and Gordon AM (1997) Poplar leaf biomass distribution and N dynamics in a poplar-barley intercropping system in southern Ontario. *Agroforestry Systems* 37: 79–90.

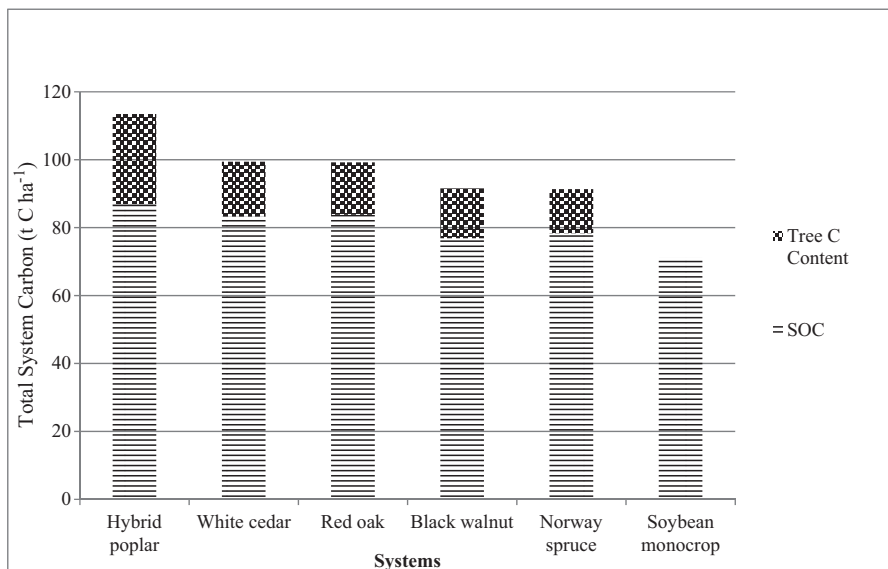


Fig. 10 System level greenhouse gas emission reduction potential of five tree species commonly found in tree-based intercropping systems over the past 25 years in comparison with a conventional agricultural system in southern Ontario, Canada. Source: Wotherspoon A, Thevathasan NV, Gordon AM, and Voroney RP (2014) Carbon sequestration potential of five tree species in a 25-year-old temperate tree-based intercropping system in southern Ontario, Canada. *Agroforestry Systems* 88: 631–643.

greater atmospheric CO₂ sequestration potential when the intercropped tree systems were compared to a conventional agricultural system.

Even though the net flux for the soybean sole-crop was negative ($-1.2 \text{ t C ha}^{-1} \text{ y}^{-1}$), when a crop rotation such as corn-soybean-wheat is considered, agricultural SOC can be maintained, if not increased with proper Best Management Practices (BMP) (Ijzerman et al., 2022; Wotherspoon et al., 2014). However, the presence of perennial trees provides an additional pathway to increase C capture in agricultural systems, as indicated in this study.

At the same study site, it was interesting to note the change in carbon sequestration as the system matured. Ijzerman et al. (2022) reported system-level C sequestration (above and belowground tree C + SOC) for four tree species grown on the same site, although

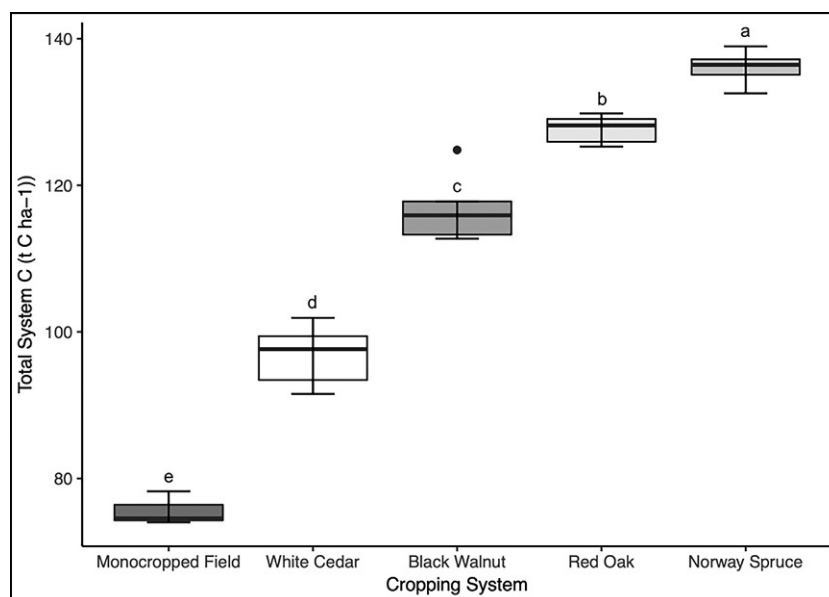


Fig. 11 Mean total system C (t C ha^{-1}) + SE for four tree species in a 31-year-old tree-based intercropping system in southern Ontario, Canada and a monocropped soybean field ($n = 5$), to a depth of 30 cm. Total system C includes SOC stock and tree/soybean C content. a, b, c, d, e means with different superscripts between cropping systems are significantly different at $P < 0.05$. Source: Ijzerman MM, Bazrgar AB, Gordon AM, and Thevathasan NV (2022) Quantification of the carbon sequestration potential of a 31-year-old tree-based intercropping system in southern Ontario, Canada. *Advances in Environmental and Engineering Research* 3: 1–14.

the trees were 31-years old when data were collected in 2018. When comparing Fig. 9 (system age = 25 years) and Fig. 10 (system age = 31), the increase in system-level C sequestration associated with respective trees species and in the soil is obvious.

Results presented in Fig. 11 align with studies done on the same site in previous years. Peichl et al. (2006) reported that the poplar Tree-based Intercropping (TBI) system had 16.0 t C ha^{-1} more than a barley monocropped system, 13 years after establishment. Wotherspoon et al. (2014) reported that, at a tree density of $111 \text{ trees ha}^{-1}$, hybrid poplar, white cedar, red oak, black walnut sequestered, respectively, 42.3, 28.3, 28.1, 20.4 and 20.2 t C ha^{-1} more than a soybean monocrop, 25 years after establishment. Although total C was highest in the poplar TBI system, which is in agreement with previous studies (Thevathasan and Gordon, 2004; Peichl et al., 2006), an abatement in net assimilation rates was observed, possibly due to tree age. However, slow-growing species such as Norway spruce, red oak, black walnut and white cedar were found to add continuously larger amounts of carbon as they continued to mature (Ijzerman et al., 2022).

While earlier studies provide evidence that high initial C sequestration was attributed to fast-growing species such as hybrid poplar, Ijzerman et al. (2022) have highlighted the importance of incorporating slow-growing spruce and oak species for long-term C storage. This may suggest that for fast-growing tree species management strategies such as relay planting, where there is a secondary planting around 10 years after the initial TBI establishment, may provide continuously high C sequestration rates as the TBI system matures. This would allow for each species to contribute maximum system C sequestration benefits to the TBI system.

Carbon sequestration in riparian buffer systems (RBS) in southern Ontario, Canada

Ofosu et al. (2021) compared the SOC (0 to 60 cm soil depth) between different RBS that were established on previous agricultural lands and associated adjacent agricultural fields. SOC quantities within the different RBS types were significantly higher than values reported for adjacent agricultural fields (Fig. 12).

The above results show that within watershed areas in the province of Ontario, if a 15 m wide RBS was established on either side of the stream, significant gains in SOC can be realized in the top 60 cm of the soil layer. This accumulation will be in addition to the C sequestration in trees in (above and belowground biomass) the riparian buffer area. The same authors (Ofosu et al., 2022) reported SOC and biomass carbon associated with different riparian buffer types. The results are presented in Fig. 13.

The results from these studies reveal a great opportunity for climate change mitigation in the province of Ontario, Canada, in other jurisdictions in Canada, and other temperate nations. In Ontario, there are 150,000 ha of watersheds associated with 36 Conservation Authorities (<https://www.ontarioconservationareas.ca/>). The province of Ontario therefore has the potential to sequester carbon in the long-term along many water bodies if riparian buffers are prioritized in the land holdings of the Conservation Authorities.

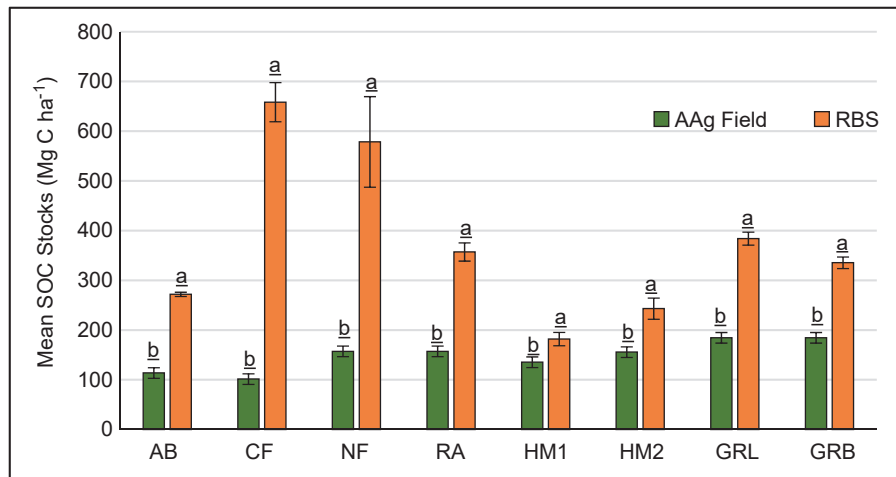


Fig. 12 Comparison of mean soil organic Carbon stocks (Mg C ha^{-1}) in diverse riparian buffer systems (based on equivalent soil mass basis) and their adjacent conventional agricultural fields (0–60 cm depth) in Southern Ontario, Canada, 2017–2018. Means followed by same letter are not significantly different, Tukey's multiple comparisons of means ($P \geq 0.05$). [Coniferous (AB = Ausable buffer, CF = cedar forest), Deciduous (NF = natural forest, RA = rehabilitated buffer), Grass (HM1 = Hullet Marsh 1, HM2 = Hullet Marsh 2, GRL = Grassland, GRB = Grass buffer)]. Source: Ofosu E, Bazrgar AB, Coleman BRW, Deen B, Gordon AM, Voroney RP, and Thevathasan NV (2021) Soil organic carbon enhancement in diverse temperate riparian buffer systems in comparison with adjacent agricultural soils. *Agroforestry Systems* 96: 623–636.

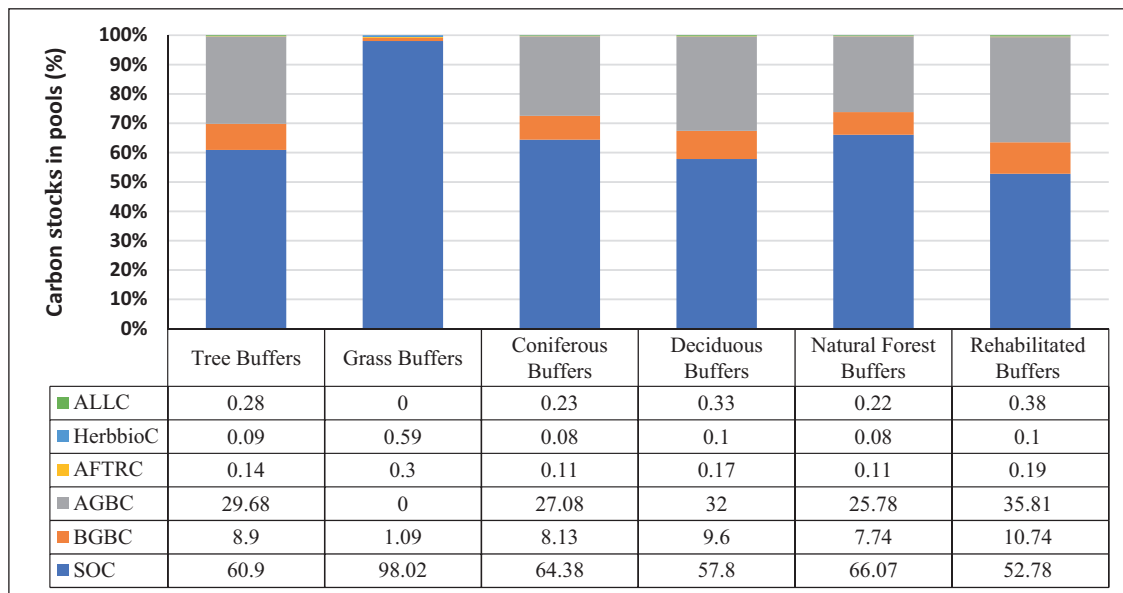


Fig. 13 Percentage of carbon stocks in various carbon pools [soil organic C (SOC, 0–60 cm), belowground woody biomass C (BGBC), aboveground woody biomass C (AGBC), annual fine root turnover C (AFTRC), annual leaf litter C (ALLC), herbaceous biomass C (HerbbioC)], of diverse riparian buffer systems [tree buffers (coniferous, deciduous), grass buffers, rehabilitated buffers, and natural forest buffers] in Southern Ontario, Canada sampled in 2017–2019. NOTE: the 0 value shown for ALLC and AGBC in Grass Buffers is as a result of the problematic measuring of leaf litter in grass buffers and the absence of trees and shrubs (aboveground woody biomass) in grass buffers. Source: Ofosu E, Bazrgar AB, Coleman BRW, Deen B, Gordon AM, Voroney RP, and Thevathasan NV (2022) Diverse temperate riparian buffer types promote carbon sequestration in southern Ontario, Canada. *The Forestry Chronicle* 98: 103–118.

Concluding remarks: The role of agroforestry in land management in the 21st century

Agroforestry began to attract the attention of the international development and scientific community in the 1970s as a means to increasing and sustaining agricultural production in marginal lands and remote areas of the tropics that did not directly benefit from the Green Revolution. Thanks to the research and development efforts at various local, regional, and global levels, agroforestry is today recognized as having the potential to offer much toward sustainable land management and environmental integrity in poor

and rich nations alike. Recent reports on the role of agroforestry in land management in the 21st century have identified food security, climate change mitigation and adaptation, and other ecosystem services as the key areas where the impact of agroforestry systems could be conspicuous in the immediate future.

In sub-Saharan Africa, which is home to three-quarters of the world's "ultra-poor" (incomes of less than US\$ 0.50 per day), approximately 218 million people, roughly 30% of the continent's total population, struggle with hunger daily. Soil-fertility depletion, exacerbated by continuous cropping without nutrient inputs in the inherently poor soils is a major challenge in the dryland ecosystems that dominate these regions. Reversing this trend and restoring soil health in the smallholder African farming systems has become a major development policy issue on the continent.

In the temperate regions of the world, agroforestry practices and principles contributing to economic and environmental services are now being more recognized by landowners. However, high establishment and labor costs associated with agroforestry land-use systems is a challenge for wider adoption in this region. Adoption of practices to strictly meet environmental, ecosystem and conservation goals must be supported by well-planned government policies that provide incentives for tree-based farming systems. These incentives can be in the form of full subsidies to integrate trees into agricultural landscapes, cost-shared programs, and reduction in property taxes for contributing to pro-environmental or ecosystem services. Mandatory environmental regulations such as the preservation of water quality in agricultural streams by establishing riparian buffers and windbreak systems to reduce soil erosion can also enhance adoption. Development of carbon credit programs (sale of C credits = enhanced economic return) is also a novel approach given the current emphasis on climate mitigation efforts promoted in this region.

Agroforestry is a key approach to the integration of climate change adaptation and mitigation objectives, often generating significant co-benefits for local ecosystems and biodiversity. Synergies between climate change adaptation and mitigation actions are particularly likely in situations involving income diversification with tree and forest products. These options also reduce the susceptibility of land use systems to extreme weather events, enhance soil fertility, and favor the conservation and restoration of forest and riparian corridors. While existing multistrata and tree intercropping systems will continue to provide substantial climate change mitigation benefits, large-scale initiatives in grazing land management, working trees in drylands, and the establishment of vegetative riparian buffer and tree woodlots are the most promising agroforestry pathways for both adaptation and mitigation. Sequestration of carbon in soils and in the biota, along with payments to resource-poor farmers for the ecosystem services rendered, would be a timely win-win strategy in the fight against food-insecurity and global warming.

The role of agroforestry in carbon sequestration is now well articulated for many agroforestry systems. Consequently, there is considerable interest in the creation of bio-carbon investment funds to channel carbon offset payments from developed countries to stimulate more carbon sequestration while simultaneously enhancing the livelihoods of small landholders and the environment. These investments will encourage development pathways resulting in higher landscape-level carbon stocks. Major public investment in the farming community can reward farmers for ecosystem services that their enterprises provide to society, in both the developed world and the emerging economies such as China and India. This suggests that the integration of trees into agricultural systems will be a major opportunity in the coming decades.

The gains and developments of more than five decades of agroforestry research and development are certainly impressive. In particular, the second half of the 1990s witnessed a large number of top-quality research endeavors and specialized publications. Undoubtedly, agroforestry is now on a firm scientific footing and is today no longer a mysterious enigma that defies science and scientific principles, as it was perceived two decades ago. Agroforestry is well on its way to becoming a specialized science at a level similar to those of crop and forest science. These hard-earned scientific gains need now to be put into practice in terms of solving the problems alluded to in this chapter. Process-oriented research needs to be extended to hitherto under-researched aspects of agroforestry, and a rethinking is also needed on overall research and development efforts at all levels (local, regional, and global).

Relatively under-researched aspects include the "agro" (i.e., crop) component of agroforestry. All crops and their cultivars/varieties that are used in agroforestry research today are those that have been developed for sole-crop conditions, and they may not be the best for situations such as those that occur in agroforestry systems. Domestication and improvement of indigenous and underexploited trees is also an area that can yield immediate dividends. Furthermore, research endeavors have so far been mostly limited to plot- and field-levels and now need to be scaled up to landscape, watershed, and ecosystem levels. Another important area of research priority is policy and market issues pertaining to the valuation of non-timber forest products. Enhancing the adoption potential of agroforestry technologies through improved extension techniques and methodologies for impact assessment are two additional areas that need immediate research attention. In general, all agroforestry land-use systems discussed in this chapter have their strengths and challenges (weaknesses). Even though these practices promote environmental, ecological, economic and social aspects, the outcomes are often influenced by 'internal and external' elements. Nair et al. (2016) have summarized these elements using SWOT analysis (Fig. 14). Agroforestry systems, be they practiced in the tropics or temperate regions, provide many basic needs and ecosystem services, and thus contribute to many regional developmental goals. These under-exploited systems have the potential to develop into a set of major land-use options in the 21st century.

Acknowledgment

This chapter is an updated version of a previous chapter in this volume by P.K. Nair (Nair, 2004), and draws heavily on that publication as well as current and previous work by all 3 authors, as cited in this chapter. Certain passages and texts have also been taken from Nair (2013) *Agroforestry: Trees in Support of Sustainable Agriculture* (<https://doi.org/10.1016/B978-0-12-409548-9>).

	POSITIVE	NEGATIVE
	Strengths	Weaknesses
INTERNAL 	Food- and nutritional security	Long-term investment
	Monetary benefits: higher farm incomes and income diversification	Relatively unimproved farming practices
	Risk reduction; Satisfaction	Difficulty in commercialization
	Multifunctionality	Lack of research and technical skills
	Niche complementarity in maximum utilization of natural resources	Lack of institutional recognition and policy support
	- Ecological diversity - Enhancement of ecosystem and environmental services	- Difficulty in extrapolating to other areas - Labor intensive in relation to managing both components (trees and the annual crops)
	Landscape aesthetics	
EXTERNAL 	Opportunities	Threats
	Potential government support	Potential future tax on "subsidiary income"
	Increase in land value	Uncertainty about future government regulations
	Environmental and recreational benefits	Excessive lopping and exploitation of preferred species
	Climate-smart agriculture and climate-change mitigation	Changes in Wood import policies
	Enhancement of socio-cultural values and land ethics	- Policy and governance issues related to input subsidies - Local land tenure regulations may inhibit adoption - Change in ownership of the land can threaten the longevity of the practice/s

Fig. 14 Internal and external influence on opportunities and challenges associated with agroforestry systems. Adapted from Nair PKR, Viswanath S, and Lubina PA (2016) Cinderella agroforestry systems. *Agroforestry Systems* 91: 901–917.

05088-0), a chapter appearing in the book: Reference Module in Earth Systems and Environmental Sciences, published by Elsevier. In this revised chapter, additional and current information has been included on aspects of temperate agroforestry that contribute to the sequestration of soil carbon.

References

- Brandle JR, Hodges L, Tyndall J, and Sudmeyer RA (2009) Windbreak practices. In: Garrett HE (ed.) *North American Agroforestry: An Integrated Science and Practice*, 2nd edn, pp. 75–104. Madison, WI: American Society of Agronomy.
- Coleman BRW, Gordon AM, and Newman SM (2018) *Temperate Agroforestry Systems*, 2nd edn UK: CABI International.
- Englund O, Börjesson P, Mola-Yudego B, Berndes G, Dimitriou I, Cederberg C, and Nicolae S (2021) Strategic deployment of riparian buffers and windbreaks in Europe can co-deliver biomass and environmental benefits. *Communications Earth and Environment* 2: 176. <https://doi.org/10.1038/s43247-021-00247-y>.
- Garrett HE, Jose S, and Gold MA (2022) *North American Agroforestry*, 3rd edn Madison, WI: American Society of Agronomy.
- Ijzerman MM, Bazrgar AB, Gordon AM, and Thevathasan NV (2022) Quantification of the carbon sequestration potential of a 31-year-old tree-based intercropping system in southern Ontario, Canada. *Advances in Environmental and Engineering Research* 3: 1–14.
- Nair PKR (2004) Agroforestry. In: Hillel D and Hatfield J (eds.) *Encyclopedia of Soils in the Environment*, 1st edn, pp. 35–44. Amsterdam, NL: Elsevier/Academic Press.
- Nair PKR and Garrity D (2012) *Agroforestry – The Future of Global Land Use, Advances in Agroforestry*. Springer Dordrecht Heidelberg New York London.
- Nair PKR, Kumar BM, and Nair VD (2009) Agroforestry as a strategy for carbon sequestration. *Journal of Plant Nutrition and Soil Science* 172: 10–23.
- Nair PKR, Viswanath S, and Lubina PA (2016) Cinderella agroforestry systems. *Agroforestry Systems* 91: 901–917.
- Nair PKR, Kumar BM, and Nair VD (2021) *An introduction to agroforestry: Four decades of scientific developments*, 2nd edn Switzerland: Springer Nature.

- Oforu E, Bazrgar AB, Coleman BRW, Deen B, Gordon AM, Voroney RP, and Thevathasan NV (2021) Soil organic carbon enhancement in diverse temperate riparian buffer systems in comparison with adjacent agricultural soils. *Agroforestry Systems* 96: 623–636.
- Oforu E, Bazrgar AB, Coleman BRW, Deen B, Gordon AM, Voroney RP, and Thevathasan NV (2022) Diverse temperate riparian buffer types promote carbon sequestration in southern Ontario, Canada. *The Forestry Chronicle* 98: 103–118.
- Peichl M, Thevathasan NV, Gordon AM, Huss J, and Abohassan R (2006) Carbon sequestration potentials in temperate tree-based intercropping systems, Ontario, Canada. *Agroforestry Systems* 66: 243–257.
- Plascencia-Escalante FO (2008) *An analysis of some components of the nitrogen cycle as affected by land use adjacent to the riparian zone of a southern Ontario stream*. PhD Thesis Guelph, ON: University of Guelph.
- Smith JR (1950) *Tree crops – A permanent agriculture*. NY: The Devin-Adair Co.
- Thevathasan NV and Gordon AM (1997) Poplar leaf biomass distribution and N dynamics in a poplar-barley intercropping system in southern Ontario. *Agroforestry Systems* 37: 79–90.
- Thevathasan NV and Gordon AM (2004) Ecology of tree intercropping systems North temperate region: Experiences from southern Ontario, Canada. *Agroforestry Systems* 61: 257–268.
- Wotherspoon A, Thevathasan NV, Gordon AM, and Voroney RP (2014) Carbon sequestration potential of five tree species in a 25-year-old temperate tree-based intercropping system in southern Ontario, Canada. *Agroforestry Systems* 88: 631–643.

Further reading

The following are some of the major publications in agroforestry the publication of the previous version of this encyclopedia.

- Alavalapati JRR and Mercer DE (eds.) (2004) *Valuing agroforestry systems: Methods and applications*. The Netherlands: Kluwer Academic Pub.
- Coleman BRW, Gordon AM, and Newman SM (eds.) (2018) *Temperate Agroforestry Systems*, 2nd edn UK: CABI International.
- Dagar JC, Singh AK, and Arunachalam A (eds.) (2014) *Agroforestry systems in India: Livelihood security and ecosystem services*. The Netherlands: Springer.
- Dawson IK, Torquebiau E, Malézieux E, Iiyama M, Sileshi GW, Kehlenbeck K, Masters E, McMullin S, and Jamnadass R (2013) Agroforestry, food and nutritional security. In: *ICRAF Background Paper for the International Conference on Forests for Food Security and Nutrition, FAO, Rome, 13–15 May, 2013* accessible at: <http://bit.ly/10oEWkC>.
- Garrett HE, Jose S, and Gold MA (2022) *North American Agroforestry*, 3rd edn Madison, WI: American Society of Agronomy.
- Jose S and Gordon AM (eds.) (2008) *Toward Agroforestry Design: An Ecological Approach*. NY: Springer.
- Jose S (ed.) (2010) Agroforestry for ecosystem services and environmental benefits. *Reprinted from Agroforestry Systems* 76(1). Springer, Netherlands.
- Kumar BM and Nair PKR (eds.) (2006) *Tropical Homegardens: A Time-Tested Example of Sustainable Agroforestry*. The Netherlands: Springer.
- Kumar BM and Nair PKR (eds.) (2011) *Carbon Sequestration in Agroforestry Systems: Opportunities and Challenges*. The Netherlands: Springer.
- Leakey RRB (2012) *Living with the Trees of Life: Towards the Transformation of Tropical Agriculture*. UK: CABI International.
- Nair PKR and Garrity DP (eds.) (2012) *Agroforestry – The Future of Global Land Use*. The Netherlands: Springer.
- Nair PKR, Kumar BM, and Nair VD (2021) *An Introduction to Agroforestry - Four Decades of Scientific Developments*, 2nd edn Switzerland: Springer.
- Nair PKR, Nair VD, Kumar BM, and Showalter JM (2010) Carbon sequestration in agroforestry systems. *Advances in Agronomy* 108: 237–307.
- Nair PKR, Rao MR, and Buck LE (eds.) (2004) *New Vistas in Agroforestry: A Compendium for the 1st World Congress of Agroforestry*. The Netherlands: Springer.
- Rigueiro-Rodríguez A, JH MA, and Mosquera-Losada MR (eds.) (2008) *Agroforestry in Europe*. The Netherlands: Springer.
- Snelder DJ and Lasco RD (eds.) (2008) *Smallholder tree growing for rural development and environmental service: Lessons from Asia*. The Netherlands: Springer.
- Zomer RJ, Trabucco A, Coe R, and Place F (2009) Trees on farm: An analysis of global extent and geographical patterns of agroforestry. In: *ICRAF Working Paper no. 89. World Agroforestry Centre, Nairobi, Kenya*.

Some Additional Resources

- Nair PR (2011) Agroforestry systems and environmental quality: Introduction. *Journal of Environmental Quality* 40(3): 784–866.
- Agroforestry Systems, published by Springer, The Netherlands, is the leading scientific journal exclusively for agroforestry; accessible on-line: springer.com/journals.
- The Overstory is an electronic newsletter exclusively for agroforestry with many subscribers in 159 countries; web-site: www.agroforester.com/overstory/ovs.html; e-mail: overstory@agroforester.com.
- ICRAF (icraf.cgiar.org), the World Agroforestry Centre, Nairobi, Kenya, is an international agricultural research institution under the Consultative Group on International Agricultural Research, and maintains an extensive database on various subjects related to agroforestry.

Rangeland management

Alexandra Sandhage-Hofmann, Institute of Crop Science and Resource Conservation, Division Soil Science, University of Bonn, Bonn, Germany

© 2023 Elsevier Ltd. All rights reserved.

This is an update of G.L. Anderson, RANGE MANAGEMENT, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 360-365, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00479-3>.

Introduction	88
Rangeland management	89
Animals	90
Domestic livestock	90
Wildlife	92
Multiple species grazing	92
Stocking rate	92
Grazing system	94
Water	95
Fire	95
Monitoring	96
Rangeland degradation	96
Vegetation	96
Invasive species	97
Woody plant encroachment	97
Soil	98
Physical degradation	98
Chemical degradation	99
Biological degradation	99
Outlook	100
References	100

Abstract

Rangeland management is essential to the sustainable use of rangeland by protecting, improving and restoring the rangeland's basic resources—soil, water, plants, and wildlife. Important management principles include livestock species, stocking rates, grazing systems, fire suppression or prescribed burning, and water supply. Management must be flexible and adaptive in response to uncertainties and variation. The effects of management actions need to be monitored on the ground and/or with new methods such as remote sensing or modeling. Many rangelands are affected by constant vegetation changes, soil degradation, and social and environmental changes (climate change). Future rangeland management must address these challenges through adaptive management strategies.

Introduction

Rangelands are the dominant land cover type on Earth. There are different definitions of rangeland, but most are a combination of land cover type, referring to vegetation or even a biome, and land use; only the emphasis between these aspects varies. There are general features that are characteristic of rangeland:

- (1) (semi-) natural vegetation
- (2) suitable for extensive grazing or browsing by wildlife and/or domestic animals
- (3) non-cultivated land

Rangelands can be defined as “lands that support predominantly native or naturalized vegetation capable of sustaining native or domestic grazing and/or browsing ungulates.” These characteristics distinguish rangelands from cropland and pastures, where the latter is used for the production of adapted, forage plants for domesticated livestock. Here, the land is seeded, usually with introduced species, sometimes with native plants and combined with intensive land use.

Rangelands have a high spatial and temporal heterogeneity, which is inherent (e.g., climatic variability, topographic features, soils, plant communities) as well as generated by disturbances (grazing, fire). They provide a suite of provisioning, regulating, supporting and cultural ecosystem services, meeting the growing demand for livestock products (Erb et al., 2016) with carbon sequestration, biodiversity conservation (Kremen and Merenlender, 2018), and recreation certainly being among the most important.

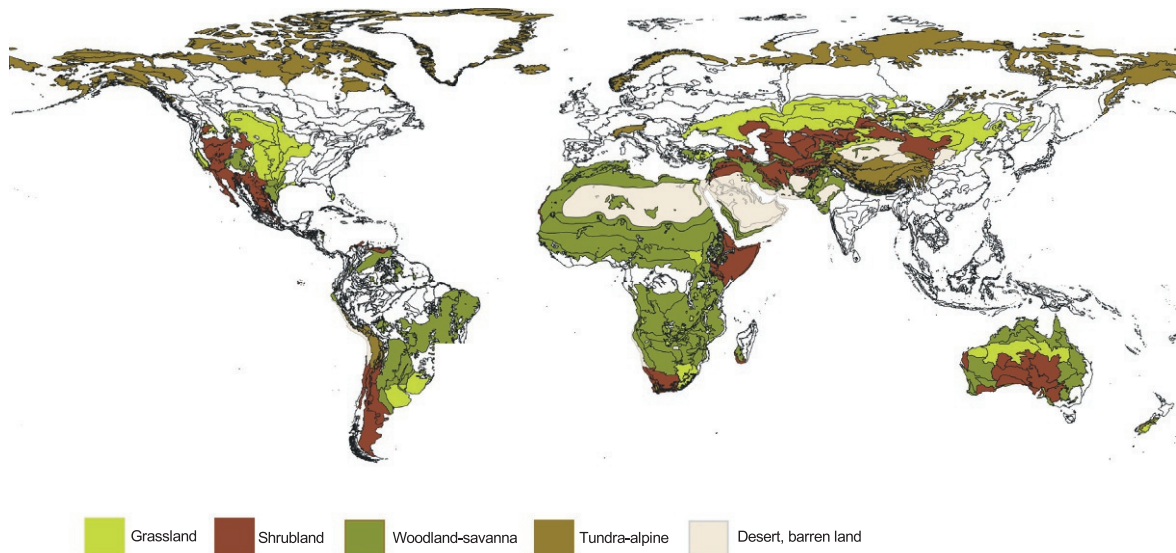


Fig. 1 Global map of rangelands as derived from World Wildlife Fund terrestrial ecoregions map. Data download: <https://www.worldwildlife.org/publications/terrestrial-ecoregions-of-the-world>.

Rangelands are diverse and not restricted to specific ecosystems. Mostly they are found in arid areas characterized by limited rainfall, often sparse vegetation and climatic extremes. However, they also occur in all bioclimatic zones of the world, with conditions ranging from arid to sub-humid, and average annual precipitation spanning 150–1500 mm (Fig. 1). Rangelands exist on all continents, with the highest percentages in sub-Saharan Africa and Asia (except the Middle East). Estimates of the proportion of the world covered by rangeland vary depending on the definition and the regions included, but rangeland is generally thought to account for about 47–51% of the Earth's terrestrial surface, or 68.5 million km² (IPCC, 2014). Higher values of 54% are also documented when deserts are included in the calculations (ILRI et al., 2021). The extent of rangeland is constantly changing as portions of it are lost by conversion to cultivated land (in continental temperate zones) or, conversely, rangeland area is increased by conversion of forested land to human-made grassland (in equatorial and subtropical zones) (ILRI et al., 2021).

Due to different climates and soils, rangelands vary in their biota, which has led to classification into rangeland types. Basic rangeland types by climate and vegetation are grassland, shrubland, woodland, savanna, and tundra (Fig. 1). Grasslands are treeless and dominated by plants of the *Gramineae* family. Savannas, forests, and shrublands are distinguished primarily by their varying tree cover, while tundra areas are characterized by a treeless plain with tufted perennial plants and lichens in arctic regions. In addition to these types, there are numerous subtypes, some of which are defined only for distinct regions and specific criteria in certain parts of the world (e.g., Eurasian steppe, South African Highveld, American prairies, and South American pampas).

Rangelands are used by a large number of people in pastoral systems or livestock-dependent societies. The number of pastoralists is currently estimated to be 200 million, worldwide (FAO, 2022). Pastoralists are commonly defined as people who derive more than 50% of their income from extensive livestock production and livestock products. However, there is a huge variety of pastoral systems around the world, categorized by economy (subsistence or commercial pastoralists), the contribution of agriculture to income (pure pastoralists and agro-pastoralists), management strategies or proportion of movement (mobile and sedentary) (for references see Sandhage-Hofmann (2016)).

Rangeland management

Livestock and grazing are still the main land uses in rangelands, even though increasing types of use are being added by, for example, tourism, agro-pastoralism, conservation, mining, and energy production. There are several, mostly linked, strategies for sustainable management for livestock, where the overall goal is to maintain rangelands in good condition, sustaining their basic ecosystem functions and services. Rangeland management developed as a discipline that aims to ensure sustained yields of rangeland products while protecting, improving and restoring the basic range resources: soil, water, plants, and animal life. The management has evolved from a pure exploitation of rangelands (end of 19th and beginning of 20th century), to steady-state management (early last century), followed by ecosystem management (late 1970s), to today's resilience-based approach of rangeland management (Briske, 2017). This means people changed from being simply users of the rangeland resources to being stewards of the resources. Ecological and social change, uncertainty, and incomplete knowledge are part of resilience-based rangeland management. Resilience can be defined as the capacity of a system to absorb disturbance and reorganize to maintain function, structure, feedbacks and identity. Exceeding the resilience, ecosystems may form alternative stable ecosystems. The resilience-based approach also recognizes different spatial scales. Rangelands are viewed as social-ecological systems resulting in stakeholder engagement. Consequently, rangeland

management must be flexible and adaptive enough not only to tailor grazing and management activities to suit unpredictable environmental conditions, but also to respond to changing policy, economics, demands for ecosystem services, and management capacity (for references see [Briske, 2017](#)). Such flexibility and adaptability constitute a serious challenge, especially given climate change and decreasing profitability of range-fed livestock. Adaptive management concepts have the ability to cope with heterogeneity and disturbance and respond to changes, and have replaced static management prescriptions in rangeland management ([Allen et al., 2017](#)).

Even though rangeland management has developed from exploitation to an approach that is sustainable for the rangeland and its users and responsive to unexpected events, the basic grazing principles—animal species, stocking rate, grazing systems (duration and season of grazing), and fire—are still the mainstays).

Animals

The management of livestock and their grazing behavior is an important topic with respect to sustainability and the worldwide problem of rangeland degradation, which is assumed to be caused also by overgrazing. When selecting livestock for a particular rangeland, the rangeland manager must decide about livestock species and the composition of classes in the herd, as well as stocking rate and stocking strategy.

Domestic livestock

Approximately 12 species of domesticated animals dominate global livestock production ([Table 1](#)). The most important of these are cattle ([Fig. 2](#)), sheep, and goats, with global populations of approximately 1.5, 1.3, and 1.1 billion (B), respectively (including indoor housing, which represents only a small portion mainly in Europe). Livestock numbers continue to grow to meet increasing demand for meat, particularly in Asia, Africa and South America. During 2000–20 the increase in cattle, sheep, and goats numbers was 33%, 11%, and 50%, respectively ([FAOSTAT, 2022](#)).

Livestock species are adapted differently to various types of rangeland, and five key factors are involved in the selection of appropriate livestock species: climate—including animal responses to climate change ([Mitchell et al., 2018](#)), water requirement, vegetation, topography, and socio-economic conditions.

- (1) Climatic conditions include air temperature, where livestock species differ in withstanding climatic extremes. It is important to recognize that not only the air temperature, but also the thermoneutral zone (TNZ) of the animals and their thermal safety are important for the selection of species ([Mitchell et al., 2018](#)). The TNZ is defined as the range of effective ambient temperatures (EAT) at which an animal does not have to regulate its body temperature actively. For healthy cattle, a range between 5 °C and 25 °C is recommended, whereas sheep and goats can cope with higher and lower temperatures. The TNZ is influenced by many factors, including breed, age, body condition, type of coat and pigmentation of the hide and hair, wind speed, and humidity ([Table 1](#)).
- (2) Adequate water supply is critical for livestock health and production. Water requirements vary significantly depending on the livestock species, and consumption of water is influenced by numerous factors like age, gender, pregnancy, lactation, activity, type of diet, and feed intake ([Table 1](#), [Fig. 3](#)). In addition, environmental properties control water intake. For example, water requirement of beef cattle rises as temperatures rise from 4.4 °C to 32.2 °C ([Fig. 3](#)).
- (3) Depending on food preference, herbivores are divided into three major groups: (i) Grazers such as cattle, horses, and mules eat primarily grass and have the digestive capabilities to process large quantities of relatively low-quality forage. They have a limited



Fig. 2 Cattle in South African Highveld, Thaba Nchu, Free State South Africa. Photo: Sandhage-Hofmann.

Table 1 Number of different livestock species worldwide and their main features.

Species	Number of animals worldwide [1]	Water requirement L/day	Water requirement average L day ⁻¹	Thermal-comfort zone	Vegetation	Physiological adaption	Topographical adaption
Cattle	1.5 bio			5–25 °C [5]	Tall, coarse grass	Grazers	Flat terrain
Feedlot cattle: backgrounder cattle		15–40	25				
Lactating cows with calves		40–80	55				
Sheep	1.3 bio	4–5		0–30 °C	Shrubs, forbs, grass	Intermediate feeders	Rough, steep terrain
Feeder lamb		3.6–5.2	4.40				
Gestating meat ewe		4.0–6.5	5.25				
Lactating dairy ewe		9.4–11.4	10.4				
Goat	1.1 bio	4–5		0–30 °C	Shrubs, bushes	Browsers	Rough, steep terrain
Milking			3.7				
Dry cows, bred heifers and bulls			2.6				
Horses	60 mio			5–25 °C [4]	Coarse grass	Grazers	
Small		13–20	16.5				
Large		39–59	49				
Camels	22 mio			10–40 °C [1]		Browsers	Flat terrain
Donkeys	44 mio.		20	23–32 °C [2]		Grazers and Browsers	Rough, steep terrain
Mules	45 mio		20			Grazers and Browsers	
Yak	13,8 mio		10	8–14 °C [3]	Grass, coarse plants, sedges	Grazers	

[1] FAO (2022), [2] Ake et al. (2013); [3] Wiener et al. (2003), [4] Morgan (1998), [5] Washington University (2015), for references [2–5] see Sandhage-Hofmann (2016).

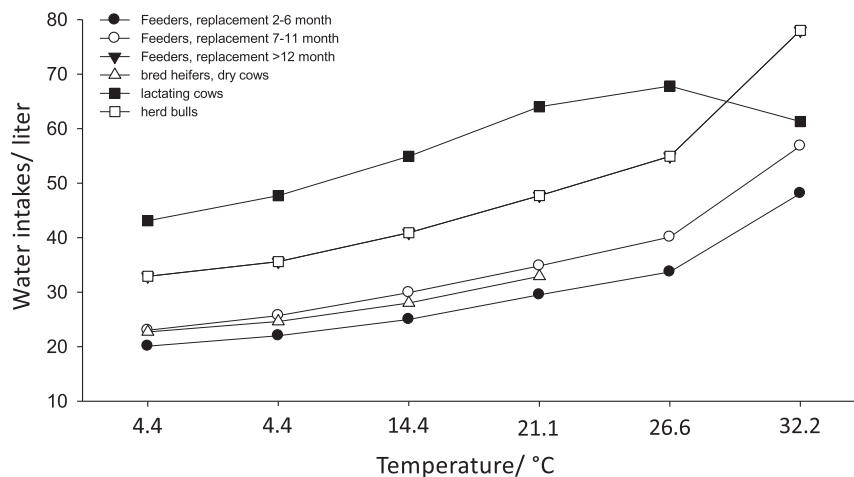


Fig. 3 Daily water intake of animals under different temperature regimes. According to N.C.R. (2000) *National Research Council. Nutrient Requirements of Beef Cattle: Seventh Revised Edition: Update 2000*. Translated from English by. Washington, DC: The National Academies Press.

- ability to choose between plants and plant parts. (ii) Intermediate feeders, which include sheep, can graze selectively and eat certain plant parts such as small leaves or buds. Sheep tend to eat more green forage than grass. (iii) Browsers such as goats are able to chew twigs and strip individual leaves from woody stems, resulting in a higher quality diet than cattle. In extreme situations such as forage shortages, most livestock adapt to different plant species.
- The topography of the rangeland can be a decisive factor in the selection of animals. Goats and sheep are versatile animals and can live in diverse environments (Table 1). This includes grazing and or browsing in varied, steep, and rocky terrain. In contrast, cattle are not well adapted to rocky and steep grounds and prefer flat areas.
 - Socio-economic conditions can influence the selection of livestock species. This is especially true in traditional rangeland systems, where certain species and their numbers are associated with wealth and power.

Wildlife

Many rangelands are also subject to grazing or browsing by wildlife (Holechek et al., 2010). In the past decades, game ranching has become an increasingly important option to benefit from rangeland. This includes bison production, which is becoming increasingly popular in the United States and Canada. In sub-Saharan Africa, numerous wildlife ranches have been established since the 1960s because they provide increased or supplemental income to livestock production (Bothma et al., 2004). The growing demand to see wild animals in nature is being met by a growing supply of (guided tours) over farm areas. Another reason for the increase of game ranching, apart from economic considerations, is the fact that native animals are better adapted to local conditions (foraging, water, climate), particularly in areas with extreme climate conditions (Holechek et al., 2010). Climate change is likely to intensify this trend.

Multiple species grazing

A promising tool for balancing the impacts on rangeland caused by one livestock species is applying multi-species grazing systems, i.e., keeping two or more animal species on the same rangeland simultaneously (Martin et al., 2020). Benefits of diversity in multi-species grazing systems include resource use efficiency, biodiversity conservation, productivity, profitability, and animal health and welfare. For example, different feeding habits among livestock species can reduce competition for feed if the dietary overlap between the species is low and forage availability is sufficient.

Stocking rate

The use of appropriate stocking rates is one of the most critical decisions for successful range management. The stocking rate is defined as the number of livestock per unit land area for a given period and is the basic relationship between livestock and forage resource (SRM, 2015). Stocking rate affects livestock production, rangeland condition and economic outcome. The objective is to balance the forage demand of grazing animals with that of forage production over an annual forage production cycle. The amount of forage an animal requires depends on its size, age, species, and whether it is growing, pregnant, lactating, or dry. There are several schemes to relate these various conditions to an “animal unit” (AU); most of them use a calculation based on the weight of the animal. Others use animals of different sizes, which are directly proportional to their weight. Table 2 gives an overview of the most common calculations for the animal unit.

Based on the animal unit or livestock unit, an expected amount of forage is used to calculate the stocking rate. The Society for Range Management calculates that one animal unit would consume 2–2.6% of its live weight per day, which represents 9.1 kg of dry

Table 2 Common livestock units/animal units of different regions worldwide.

Region	Europe	North America	South America	Tropics	Australia	New Zealand	Asia
Unit equivalent	Dairy cow	Beef cow	Cattle mature	Tropical cow	2-year-old Merino sheep	Ewe with one lamb	Cattle mature
Weight equivalent	600 kg	455 kg	455 kg	250 kg	45 kg	55 kg	455 kg
Livestock equivalents							
Cattle mature	1.00	1.00	0.70	0.70	8.0	6.3	0.65
Cattle (<2 years)	0.40	0.80			7.0	4.5	
Horse	0.70	0.80	0.65	0.80	10.0		0.65
Sheep	0.17	0.15	0.10	0.10	1.1	1.0	0.1
Goat	0.17	0.10	0.10	0.10	1.1	0.8	0.1
Mules	0.40		0.60	0.70			
Llama	0.17			0.7			
Alpacas				0.3			
Camel		1.10		1.00			
Buffalo				1.5			0.7
Wildlife equivalents							
Bison	1.00	0.90					
Red deer	0.20					2.0	
References	[5, 6]	[2, 4] [Bison management]	[3]	[1, 11]	[7, 10]	[8, 9]	[3]

[1] FAO: Tropical Livestock Units (TLU); [2] Womach (2005); [3] Chilonda and Otte (2006); [4] FAO (2003); [5] Schweizer Bundesrat (2016); [6] European Commission for Agriculture and Environment; [7] McLaren (1997); [8] Cornforth and Sinclair, A.G. (1984); [9] Woodford and Nicol (2004); [10] Government of Western Australia; [11] Jahnke et al. (1988), for references: see Sandhage-Hofmann (2016).

Adapted from references [1] FAO: Tropical Livestock Units (TLU); [2] Womach (2005); [3] Chilonda and Otte (2006); [4] FAO (2003); [5] Schweizer Bundesrat (2016); [6] European Commission for Agriculture and Environment; [7] McLaren (1997); [8] Cornforth and Sinclair, A.G. (1984); [9] Woodford and Nicol (2004); [10] Government of Western Australia; [11] Jahnke et al. (1988), for references: see Sandhage-Hofmann A (2016) Rangeland management. *Reference Module in Earth Systems and Environmental Sciences*.

matter (DM) per day⁻¹ or 3318 kg DM year⁻¹ (Holechek et al., 2010). Although fewer studies are available for the tropics and subtropics, estimates are comparable and approximately 2.5% of live weight would be consumed for one tropical livestock unit (TLU), which represents 6.25 kg DM day⁻¹ and 2281 kg DM year⁻¹.

The following equation can be used to calculate the stocking rate:

$$\frac{(\text{Total land area (acre)}) \times (\text{end of season standing forage (lbs. acre}^{-1}) \times (\text{forage utilization (lbs. acre}^{-1}))}{(\text{AU forage demand per AU per day (lbs. acre}^{-1}) \times (\text{number of days grazed}) *}$$

*Bidwell et al. (2017).

The high variability of forage production and the variability in annual and interannual precipitation are the main challenges for setting appropriate stocking rates. There is a common understanding that stocking rates should be flexible and adaptable, and include the consideration of social parameters such as current perceptions, prevalent norms, and historical contexts, especially in communal land (Rasch et al., 2016). The use of flexible stocking rates, which are adapted to fluctuating annual forage production or stocking rates, that are adjusted based on precipitation during the current growing season (adaptive management) are preferable to fixed stocking rates, which are based on the carrying capacity of the land. Carrying capacity (CC) is the maximum number of animals that a rangeland can support over the long term without degrading the natural resources of the area. It can be seen as a basic guideline for animal numbers in the management unit, because on many rangelands the carrying capacity is less than forage production.

Stocking rates also depend on the amount of forage available on the rangeland, measured at the end of the season (Redfearn and Bidwell, 2017). Forage production is highly variable and not only affected by climate, but also by soil fertility, texture, and vegetation management. Traditional methods for measuring the end of season standing biomass rely on time-consuming hand clipping in long-term exclosures at the end of season. This is increasingly being replaced by remote sensing imagery from satellites or, more recently, drone estimates that can support adaptive rangeland management during the grazing season by providing real-time estimates of forage production (Fig. 4) (Liu et al., 2019). Additional factors affecting stocking rates include the grazing system practiced, distance to water, slope, fire, and climatic variation (see Sandhage-Hofmann, 2016).

Calculation of appropriate stocking rates must include wildlife densities, because livestock and wildlife may compete for the same forage. For nature reserves or game farms, which have become increasingly prevalent, especially in Southern Africa, knowledge of "wildlife units" is necessary. The different feeding behaviors of wildlife (browsing, grazing, intermediate) and the widely varying diets of wildlife species make it necessary to consider both grazing and browsing units (GU, BU) for calculating wildlife stocking rates (Table 3), taking into account the grazing and browsing capacity of rangelands. This allows for a more accurate determination of the overall stocking rate.

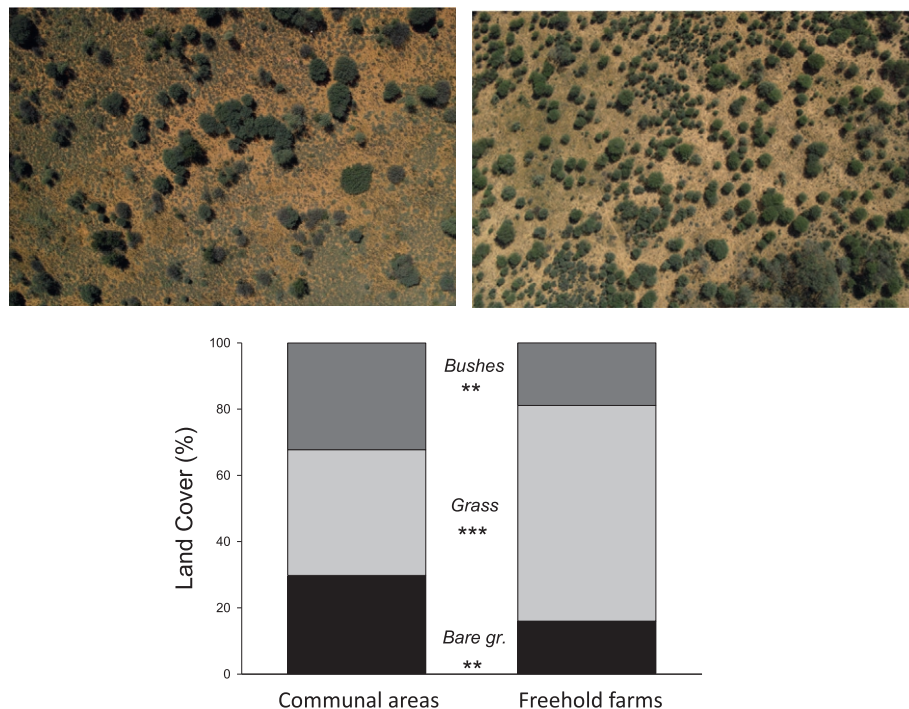


Fig. 4 Rangeland in South African Highveld, Thaba Nchu, Free State South Africa. Photo: Sandhage-Hofmann.

Table 3 Wildlife units of main wildlife species.

<i>Unit</i>	<i>Grazer unit</i>	<i>Browser unit</i>	<i>Animal unit</i>
Abbreviation	GU	BU	AU
Unit equivalent	Blue wildebeest	Kudu	Beef cow
Weight equivalent	180 kg	140 kg	455 kg
Wildlife equivalents			
Bison			0.9
Blue wildebeest	1.0	1.21	
Kudu	0.8	1.0	0.61
Waterbuck	1.1	1.33	0.58
Gemsbok	1.1	1.36	0.54
Eland	2.0	2.44	0.96
Impala	0.3	0.4	0.18
	[1]	[1]	[2, 3]

[1] Bothma et al. (2004); [2] Miller (2002); [3] Agricultural Special Valuation Guidelines (2014), for references see Sandhage-Hofmann, 2016.

Adapted from Miller et al. (2002); Bothma et al. (2004), for references see Sandhage-Hofmann A (2016) Rangeland management. *Reference Module in Earth Systems and Environmental Sciences*.

Grazing system

The introduction of grazing systems as a management option was implemented quite recently in the use of rangelands; in the 20th century. Previously, rangeland had been characterized by free-ranging herds of wild herbivores, and the traditional transhumant movement of cattle farmers (Briske, 2017). Animals moved in response to forage supply, fire, predators, and other factors (Teague et al., 2011). Early in the 20th century, settled farmers put continuous grazing pressure on localized areas. In view of rangeland degradation, various kinds of grazing systems have emerged to stop deterioration and to handle the spatio-temporally heterogeneous distribution of biomass, which is a common feature of most rangelands.

Continuous and rotational grazing are the two dominant grazing systems in the world. Continuous grazing is the most common grazing system in rangelands (Teague et al., 2011); it allows livestock to graze throughout the year (year-round continuous grazing) or during a specific period of the year (continuous grazing during the growing season). In this continuous grazed system, no resting periods for vegetation and soil are introduced. Livestock graze forage selectively and tend to use preferred local areas. Consequently, the stocking rate in these frequently grazed patches may be much higher than the intended stocking rate. Therefore, spatial patterns may develop in which some areas and plants are heavily grazed while others are not or only lightly used. In addition, livestock are concentrated around water points where permanently high grazing pressure can lead to the development of grazing gradients and consequently soil and vegetation degradation (Sandhage-Hofmann et al., 2015).

In rotational grazing systems, the rangeland is divided by fencing into several paddocks or camps among which the animals are moved regularly, with a resting period for recovery of soil and vegetation between each grazing interval. Multiple specializations of rotational grazing systems exist, which differ mainly in the length of grazing or resting time (Table 4). Longer grazing periods of several weeks and months are contrasted to short grazing periods between 2 and 10 days. In this context, the term Holistic Planned Grazing™ describes an intensive, rotational, time-controlled approach.

Rotational grazing or multi-paddock grazing systems are often applied rigidly rather than adaptively, and advantages compared to continuous grazing are low. But recently, adaptive multi-paddock grazing includes short grazing periods, moderate grazing

Table 4 Basic rotational grazing systems with a single herd.

<i>Grazing system</i>	<i>Stocking rate</i>	<i>Length of grazing</i>	<i>Length of rest</i>	<i>management</i>	<i>strategy</i>
Deferred rotation	Moderate	Long	Moderate	Fixed	HPG ^a
Rest rotation	High	Short	Long	Fixed	HPG
High intensity-short duration	High	Short	12 month ^a		
			Long		HUG ^b
			> 2 month ^a		
Time-controlled grazing	High	Short			
Short duration	High	Short	Moderate		HPG
		5 days	4–6 weeks ^a		

^aHigh-performance grazing strategy for selective grazing of preferred plants.

^bHigh-utilization grazing for grazing on preferred and unpreferred plants.

After Briske et al. (2008); Thorne et al. (2007), references in Sandhage-Hofmann A (2016) Rangeland management. *Reference Module in Earth Systems and Environmental Sciences*.

during the growing season, leaving adequate plant cover followed by adequate planned recovery, and matching livestock to forage biomass.

The debate over which is the best grazing system continues. Several publications have pointed out that, despite many prejudices, research could not support the argument that the anticipated disadvantages of continuous grazing systems lead to an overall degradation of rangeland. A recent review of grazing systems around the world showed that assessments often take into account only single indicators (e.g., only soil) but not the entire rangeland system, meaning vegetation, soil, livestock, and social interactions. In addition, a positive or negative effect of a particular grazing system sometimes applies only to a particular type of rangeland, while the effect may be different in other climates (di Virgilio et al., 2019). In addition, temporal and spatial scales of the studies vary. In view of this, the benefits of rotational systems for forage production, ubiquitous use of crops, restoration of vitality and improvement of soil properties, which have been demonstrated in numerous studies, should be generalized with caution (e.g., Hughes et al., 2011; Teague et al., 2011; Kotzé et al., 2014; Sandhage-Hofmann, 2015; Roche et al., 2015, for references see Sandhage-Hofmann (2016)).

At the very least, there seems to be a consensus that adaptive, goal-oriented management systems are required, which account for spatial heterogeneity, inherent uncertainty, risk, and incomplete knowledge (Briske et al., 2020). Management is then often “learning by doing,” and managers act despite risk and uncertainty of how measures will play out. Risk and uncertainty become greater when medium- or long-term decisions are being taken.

A basic knowledge of plants and animals is essential, together with an understanding of abiotic factors like soil and climatic circumstances of the rangeland. Grazing systems and stocking rates should be variables that are used adaptively as tools in rangeland management systems. Social circumstances should also be included in considerations of the management systems. This includes the fundamental role of livestock’s safety net function in communal areas, but also prevalent norms such as the role of livestock as a status symbol (Rasch et al., 2016).

Water

The number and location of water points are important factors in controlling the movement and concentration of livestock, as well as in controlling the distribution of grazing in rangelands (Holechek et al., 2010). Water can be supplied by wells and boreholes (artesian or pumped), dams and ponds and by pumping and piping from water bodies, or sometimes by trucks (Syria and Jordan). Water pumping techniques include wind turbines, gravity distribution, solar energy, and solar panels to prevent freezing. In traditional systems, particularly in semi-arid areas, wells, water harvesting, and storage techniques such as *birka* (cisterns) in Somalia and *hafir* (dug tanks) in several Arabic countries are common. Sometimes, water is transported by truck to the rangelands (Syria and Jordan).

Ideal distances between water points depend on the livestock species and age, topography, climate, season, type of vegetation and grazing system. Livestock are averse to walking long distances to water. In general, goats and herded sheep can be without water for a longer time than cattle as they will probably derive water from food. They were reported to depend on a 3–5-day watering, while cattle rely on water supply within a period of 3 days. Of cattle’s activity, 80–90% takes place within a range of 5 km from water, meaning that 5–6 km between water points could be optimal in terms of achieving a moderate grazing distribution and also preventing overgrazing. However, several studies in various countries have suggested different recommendations, from 1.6–3.2 km (North America), 6–10 km in dry seasons for cattle and 4–5 km for sheep (South Africa), 3 km for sheep, 6 km for cattle, and 7 km for red kangaroos (Australia) with variations according to wet or dry seasons (for all references see Sandhage-Hofmann (2016)).

A balanced supply of water is recommended, because a widespread availability of water, allowing herbivorous mammals to continue grazing, can lead to feed shortages when there is a lack of rainfall (Holechek et al., 2010). In contrast, there may be negative effects on various ecological properties; especially in large paddocks with few water points, the areas near water tend to be overgrazed.

Fire

Fire is critical to maintaining biodiversity and ecological processes in rangelands and is a natural component of most of them. However, the beneficial effect of fire has been underestimated for a long time and prescribed burning has been reduced, resulting in a recent increase in number and extent of wildfires (Starns et al., 2019). But planned burning and/or fire suppression are both essential parts of rangeland management. Prescribed burning is the use of fire under specific conditions to increase forage productivity, reduce litter, improve growth of favorable plants, combat undesirable plants (e.g., brush encroachment), and control pests and diseases. Fire can also have a positive effect on soil by increasing pH, plant nutrients and exchangeable cations, and NO_3^- immediately after burning. Negative effects may also result, such as water repellent surfaces which hinder water infiltration and enhance runoff and erosion. It is important to adapt the fire regime, intensity, frequency and season, to local conditions to avoid negative effects on plant performance and soil. In addition, improper use can be a threat to human life, property, community assets, air quality, and other rangeland aspects. Prescribed burning differs from the reaction-driven treatment that is generally applied following wildfires (Holechek et al., 2010).

Monitoring

The effectiveness of management actions can only be assessed through monitoring. Monitoring strategies include seasonal observations with short-term adjustments of management as well as long-term monitoring with, consequently, long-term management strategies to cope with interactions between livestock pressure, droughts, invasive species, bush encroachment, wildfires and climate change (McCord and Pilliod, 2022). In relation to the evolution of adaptive rangeland management, adaptive monitoring is required which responds to new management questions, changing environmental and socioeconomic conditions, improved monitoring methods, models, tools, and experience in implementing the monitoring program. Monitoring requires (i) setting objectives and questions, (ii) selecting methods and sampling locations, (iii) collecting data, and (iv) analyzing and interpreting the data. Evaluation of these results should then be used to adjust management, if necessary. Monitoring tools developed from field-based quantitative and qualitative measurements may evolve to new techniques with remote sensing tools plus machine learning and model development. Promising tools for monitoring rangelands are drones and increasingly smart-phone cameras (sometimes even with an integrated LiDAR sensor). Historical and present trends of land cover data can also be used to evaluate previous management actions, and to guide future management decisions.

Rangeland degradation

Rangeland degradation is defined as a reduction in the ability to provide ecosystem functions and services such as forage production and quality for livestock, water quality and quantity, biodiversity conservation, or carbon sequestration. Rangelands are highly dynamic ecosystems that include changes in ecological and social properties at different temporal and spatial scales. However, interactions of social and ecological factors can cause transitions to undesirable alternative states where ecosystem services are negatively affected. In arid and semiarid ecosystems, these transitions are often observed not as gradual, predictable succession, but as abrupt regime shifts to alternative states ("state changes") (Archer et al., 2022). Currently, many rangelands are altered, as evidenced by persistent changes in vegetation (plant species composition), soil degradation, or invasive species (Nandintsetseg et al., 2007; Stafford Smith et al., 2007; Williams and Jackson, 2007; Dai, 2011, for references see Sandhage-Hofmann (2016)). Drivers of ecological and social changes in rangelands can be summarized into (i) human population growth and the associated unsustainable use of natural resources; (ii) climate-induced warming and increasing climate variability (iii) accelerated land conversion and fragmentation, including exurban development and (iv) reduced effectiveness of traditional common governance, where developing countries are facing additional challenges like aging, a shift in societal preferences away from rangelands and low profit margins in livestock production (Briske et al., 2020).

Rangeland degradation is a major problem worldwide, but the assessment of degradation varies significantly due to differences in the methods, benchmarks, and scales (temporal and spatial) under consideration, which lead to different estimates about the extent (Lund, 2007; Reed et al., 2015). Some authors have suggested that more than 50% of the world's rangeland area is degraded, while others estimate that it is up to 75% (Seymour et al., 2010; for references see Sandhage-Hofmann (2016)). The rangeland atlas estimated degradation based on the productivity status of rangelands, which resulted in an area of 20% (excluding deserts), where developing countries are more affected (ILRI et al., 2021). Rangeland degradation is usually first perceived in the condition of the vegetation, but other ecological properties are also affected by degradation phenomena.

Vegetation

Degradation of rangeland vegetation has been attributed to a combination of overstocking with livestock, deforestation, woody-plant encroachment, and invasion by non-native plant species. Historical-cultural and social conditions also contribute, as does global climate change. Livestock overgrazing is considered to be one of the main causes of vegetation degradation by decreasing both the productivity and resilience of host species, reducing vegetation cover, increasing unpalatable species, and decreasing species diversity.

Grazing (and browsing) is a natural ecological process in all rangelands. Herbivores consume plant material including its compounds, such as proteins and carbohydrates, which are digested and provide them with energy and nutrients (Teague et al., 2011). Dung and urine release close the loop. Grazing has many direct or indirect effects on the vegetation. Whether they improve or degrade the vegetation depends on the timing, duration, kind of grazing, rangeland type, but also aridity. For example, plant species richness improves in wet regions when grazing is compared to non-grazing, while plant richness is negatively affected in dry regions, and stocking rates explain variation in the data (Gao and Carmel, 2020). In general, studies have reported that light and moderate stocking rates have no or positive effects on plant productivity and composition, especially in fertile rangelands. Causes are attributed to increasing spatial heterogeneity or vegetation patchiness of plant communities caused by selective grazing patterns of the herbivores. In contrast, high long-term stocking rates lead to overgrazing and negative effects on the forage value of rangelands due to depletion of some preferred species and their replacement by non-preferred species. Overgrazing is defined here as repeated continuous grazing which exceeds the recovery capacity of the plant community and creates degraded rangeland (SRM, 2015). Effects on vegetation can be manifold, e.g., the herbaceous biomass changed from perennial grasses to annual species, the portion of bare soil increased; species changed from palatable to unpalatable species, or, shifts in grass species dominance

occurred with (over-) grazing (for references see [Sandhage-Hofmann \(2016\)](#)). In this category, invasive species and encroachment of woody plants deserve special attention.

Invasive species

Invasion of plants in rangelands is a major threat worldwide. In general, the term invasive plants refers to the intentional or unintentional introduction and spread of exotic plant species (alien) into an area where they did not previously exist. These plants have the ability to thrive and spread aggressively outside their native range, eventually resulting in the replacement of indigenous plant species. Invasion processes include introduction, establishment, spread and impact, where invaders' survival and abundant establishment is ensured by early reproduction, large numbers of seeds and the ability to become established easily in degraded environments. Disturbances such as human agro-pastoral activities or alterations in patterns of fire regimes could be the causes of invasion. Or, management-induced changes in native plant communities might provide opportunities for invasion or a further increase of alien and native species ([Poland et al., 2021](#)).

The challenges posed by invasive plants in rangelands are of serious concern and are expected to increase in the next several decades. Invasive terrestrial plants can change key system processes such as productivity and nutrient cycling. Depending on their plant functional traits, invaders can accelerate carbon cycling and can lead to carbon storage decline in rangelands. They impact plant community structure and diversity, alter forage production for livestock and wildlife, and threaten livestock production. Depending on their life cycle they can increase soil erosion and deplete nutrients and reduce soil moisture ([Poland et al., 2021](#)). Despite negative effects, sometimes invasive plants can help to restore degraded rangelands and provide some benefits to the system. Invasive species can lead to the development of so-called "novel ecosystems." In future, the control of these novel ecosystems should not be based on the restoration of the original ecosystem alone, but should also consider the potentially higher biodiversity and ecosystem services these new ecosystems may have.

Woody plant encroachment

Most rangelands consist of a mixture of herbaceous and woody vegetation. In the last century, this balance has been disturbed and shifted in favor of shrubs and trees at the expense of the herbaceous (grassy) layer in many areas of the world. This so-called woody plant encroachment (WPE) is expressed in a global vegetation greening by native or non-native woody plants ([Fig. 5](#)). The distinction from "invasive species" (see above) is not always clear because alien woody plants (e.g., *Prosopis* in Kenya) also contribute to the phenomenon of WPE. Worldwide, 10–20% of drylands have already undergone encroachment by woody plants, but the real rates of encroachment vary depending on the region and methods of assessment. It is estimated that absolute encroachment rates range from zero to 3.3% cover year⁻¹, with an average of 0.85% cover year⁻¹. Rates are higher in the early stages of bush encroachment, and decrease with time. The WPE occurs in a wide range of climates, with one of the most striking shifts toward woody plants occurring in the Arctic, known as the "greening" of the Arctic ([Myers-Smith et al., 2020](#)).

Drivers and causes of WPE are inter-related and context specific. These factors include alterations in fire and grazing regimes, exclusion of browsing megaherbivores, landscape fragmentation, and land abandonment as internal factors as well as elevated atmospheric CO₂, rising N deposition and climate change as external drivers ([Archer et al., 2022](#)). Environmental variables such as soil, topography, and precipitation regimes modulate the intensity and extent of WPE. The effect of grazing, in particular, has long been criticized. Extended periods of heavy grazing suppress the production of herbaceous forage, creating space for the proliferation of woody plants. The lower grass fuel loads reduce the frequency and severity of wildfire, which seems to control the spread of bushes. Biological feedbacks like soil-fauna activities, or enhanced nutrient cycling beneath the bushes, may promote WPE. Nowadays, it is well accepted that causes of WPE are an interaction of several variables where the main factor can vary in different regions. Effects of WPE are manifold, including effects on carbon pools in biomass and soils ([Sandhage-Hofmann et al., 2020](#)), hydrology, biodiversity, nutrient cycling, composition of plants, invertebrates and vertebrates, and increased risks of vector-borne diseases.



Fig. 5 Bush encroachment, North Western Province, South Africa. Photo: Sandhage-Hofmann.

Several strategies have developed to combat WPE. “Brush management” tries to eliminate woody plants by cutting, burning, and chemical clearing to foster the recovery of herbaceous biomass, but results are variable, probably also because of the depletion of soil seed banks. Adapted stocking rates are also necessary to avoid overgrazing with its above-mentioned effects. Livestock systems with multiple species, grazers and browsers, can not only stabilize livelihoods through diversification and grazing, but browsing of megaherbivores could counteract WPE. As WPE is not necessarily associated with functional decline, it can have positive effects on ecosystem services (e.g., carbon sequestration) and management should carefully balance livestock production with other benefits of WPE.

Soil

Soil degradation is widespread in rangelands and characterized by the decline and reduction of ecosystem goods, and services provided by soils. It includes physical, chemical, and biological processes.

Physical degradation

Compaction

Trampling by animals is the main cause of alterations in physical properties of the soil. The force that a grazing animal imposes on the soil is a function of the weight of the animal and the area of contact between the animal’s hooves and the soil surface. The pressure varies between different animal species, with cattle applying the highest pressure (Table 5). One single livestock or animal unit (LSU, AU) can produce pressure on the soil beneath a single hoof of about 123 kPa, which may significantly increase when the animal moves (Bilotta et al., 2007).

The direct negative effect of trampling on soils is compaction, which affects bulk density, water infiltration rates, aggregate stability, penetration resistance and porosity (Roesch et al., 2019). Compaction itself is dependent on soil properties like texture or water content, with for example sandy soils are less prone to it. Volume and connectivity of large pores are the main properties related to infiltration and transport of water and air in soils, and both decrease with soil compaction. Macropores are particularly sensitive to deformation, and their compaction leads to higher bulk density and increases the risk of water runoff. Higher soil bulk density because of increased animal trampling has been observed for different grazing animals and rangeland systems. In a meta-analysis of grazing impacts on soils, Byrnes et al. (2018) confirmed findings from other authors, that all livestock strategies and intensities increased soil compaction, where the effect was greatest under heavy grazing, and if comparing rotational and continuous grazing, continuously grazed soils were more compacted. Trampling by animals may also lead to aggregate disruption, which is most common in arid and semiarid rangelands. Aggregation is an integrative indicator for the soil’s overall quality and closely related to several ecosystem services, and functions, including physical and hydrological functioning of soil, and sequestration of organic carbon. Trampling by animals results in a loss of stable aggregates, which decreases infiltration rates (for references see e.g., Roesch et al., 2019). Aggregate stability decreases with increasing grazing intensity and in places where animals congregate such as water-points, shaded areas, and fences (Teague et al., 2011). Soil compaction by trampling of animals depends on soil texture, moisture, and type (Sandhage-Hofmann et al., 2015), grazing intensity and strategy (Fig. 6).

Soil loss

The loss of soil from land surfaces by erosion is widespread globally and adversely affects the productivity of natural ecosystems like rangelands. In arid and semiarid rangelands, soil erosion has been widely recognized as an important soil degradation process and is considered one of the main factors responsible for declining soil fertility and desertification (Okin et al., 2001; Michaelides et al., 2009; Bestelmeyer et al., 2006; for references see Sandhage-Hofmann (2016)). Fine-size soil particles are lost, as is organic matter, which plays an essential role in the formation of stable aggregates. Plant nutrients are lost with water and or wind erosion.

Livestock grazing with high stocking rates and inflexible stocking strategies are thought to increase the susceptibility of rangelands to wind erosion by reducing vegetation cover and modifying the stability of the soil surface, but responses to management are complex and can differ between rangeland types, soil types, vegetation, and livestock species. For example, taller vegetation can inhibit erosion by reducing wind speed, for example (Gao et al., 2014). Soil erosion is influenced by vegetation cover, plant community composition, distance between grasses, and size of and connectivity between bare soil patches. Furthermore, the

Table 5 Live weight, foot area and static pressure of different livestock species.

	Mass kg	Total foot area cm ²	Static pressure kPa
Cattle	306–612	264–460	98–192
Sheep	40–54	55–84	48–83
Goats	40	55	60–73
Horses	400–700	736	54–95
Camels	450–650	1644	27–40

Adapted from Greenwood and McKenzie (2001) and Evaluation and Review Report, New Zealand, 2010, for references see Sandhage-Hofmann A (2016) Rangeland management. *Reference Module in Earth Systems and Environmental Sciences*.



Fig. 6 Degradation at a water point, North Western Province, South Africa. Photo: Sandhage-Hofmann.

occurrence of physical or biological soil crusts affects rates of erosion, where grazing destroys soil crusts directly through trampling, and indirectly by altering vascular plant communities that interact with the crust communities. Quantifying the impact of rangeland management on rates of wind and water erosion remains a challenging task and therefore many soil erosion models and applications have been developed. But, the discrepancy between modeled systems and reality is great, and often model calibrations are only valid for a specific region (Batista et al., 2019). Therefore, generalizations should be made with care.

Chemical degradation

Nutrients

Livestock grazing affects the contents and dynamics of nutrients and soil organic matter (SOM) through the soil–vegetation–animal system in many ways, including trampling, consumption, excreta deposition, redistribution, and export (Sandhage-Hofmann, 2016). Two contrasting pathways influence nutrient concentrations in rangeland soils: (i) enrichment and (ii) loss. Enrichment of nutrients occurs at locations of highest grazing pressure, where herbaceous biomass consumed by livestock is returned to the soil in form of dung and urine, as well as through a centripetal nutrient flow to the water holes at natural water points. While urine contains high concentrations of urea-nitrogen, potassium, sodium and chloride, dung contains relatively high concentrations of phosphorus, calcium, magnesium and organic nitrogen. The patches of excreta cover only a small proportion of the grazed area, and are especially large around water points, beneath trees, at kraals, or along fences (Sandhage-Hofmann et al., 2015) resulting in a spatially heterogeneous distribution of high nutrient concentrations. High stocking rates are associated with such nutrient accumulation, whereas under rotational grazing, rangeland resources are used more evenly, and manure and urine are more randomly distributed.

Nutrient losses may be caused by high grazing pressure and is accompanied by depletion of plant cover and litter input. In addition, defoliation can affect plant rates of photosynthesis, root/shoot ratios, C allocation, fine root mass, and plant root exudates (for references see Byrnes et al., 2018). Reported effects of grazing or different grazing strategies on plant nutrients range from negative to no effects (for references see Sandhage-Hofmann, 2016).

Nowadays, woody plant encroachment in many rangelands does counteract nutrient losses and improve the soils. Various processes contribute to these so-called “islands of fertility.” These include litter from leaf fall, stem flow, and through-fall; droppings from birds and mammals; accumulation of nutrients from areas beyond the crowns by means of lateral tree roots; and increased termite activity. Greater organic matter and nutrient input in encroached rangelands restores rangeland soils (e.g., Sandhage-Hofmann et al., 2020), but at the cost of palatable grass species, which threaten the livelihood of many people.

Biological degradation

Soil biological degradation involves the decline in activity and species diversity of soil biota, especially of microorganisms, which play critical roles in the cycling of nutrients, and leads to the decline of soil organic carbon, the loss of soil biodiversity and biological activity (Soto et al., 2021).

Soil organic carbon

Worldwide, a huge number of studies have evaluated the effect of grazing on soil organic carbon stocks, contents, or pools. Soil organic carbon is positively associated with many soil functions that are critical in rangeland production, including nutrient and pH buffering, water retention, and soil structural stability. It is difficult to generalize the impact that grazing has on soil carbon dynamics because the balance between soil carbon inputs and outputs in grazed rangelands varies greatly among systems. Research

has shown mixed results of grazing effects on soil organic carbon (SOC) and soil organic nitrogen (SON). In general, higher grazing intensities and overgrazing are thought to decrease soil C and N contents by directly removing above-ground herbaceous biomass, and by reducing below-ground C inputs through lower root production and higher root turnover. Continuous grazing compared with no grazing showed that continuous grazing had negative effects on C content at low, medium, and high grazing intensities. In contrast, carbon values appeared to be higher when rotational grazing was compared with no grazing (Byrnes et al., 2018). The direction and magnitude of change is influenced by many factors like climate (aridity), soil type, vegetation type, grazing management, and history of grazing. High long-term grazing intensities are leading to more bare soil patches susceptible to wind erosion and subsequent soil organic carbon loss (see above). In addition, if plant composition changes, the chemistry of plant litter, patterns of plant biomass allocation, litter production, and the spatial distribution of nutrients change. Depending on the intensity, grazing pressure may slow rates of decomposition by lowering the carbon-to-nitrogen (C:N) ratio of litter; alternatively, decreased standing biomass may accelerate the decomposition by increasing soil temperature, but decreasing soil water content may neutralize this effect. Furthermore, grazing may convert different carbon pools from a relatively recalcitrant pool into more mineralizable carbon fractions.

A recent meta-analysis (287 studies) summarized the effects of different grazing intensities on a variety of soil properties (Lai and Kumar, 2020). Results showed that of the 15 variables at 0–10 cm soil depth under heavy grazing, nine were negatively affected. This number reduced to six under medium grazing intensity. The soil properties significantly affected were bulk density, penetration resistance, soil organic carbon, and microbial community under both grazing intensities. The authors concluded that heavy grazing is more detrimental for soil quality than moderate and light grazing. But the mechanisms of the interactions between grazing and environmental factors on soil properties are still unclear and require more research.

Outlook

Rangelands and pastoral societies face social and environmental challenges, such as changes in land tenure, land-use, intensification, fragmentation, settlement, population growth, wildlife conflicts, conservation policies, and, most importantly, climate change. These changes are no longer occurring gradually, with time to adapt, but more often abruptly and unpredictably. This so-called tipping point behavior occurs when a system or parts of it cross a threshold that leads to a regime change to another, often undesirable state. Not every such regime shift is necessarily detrimental, and potentially positive changes in rangelands in the future also have to be recognized that could lead to new adapted management strategies. Degradation of rangelands continues to be a major problem, and restoration efforts are required. These include various management principles discussed above, but they must also be feasible and reasonable. The issue of benchmarking, i.e., what condition a rangeland should be restored to, should be given more consideration in the future. In addition, the culture and history of societies are important considerations. For example, planned measures may be much more difficult to implement in communal areas of developing countries with higher financial costs than in industrialized societies with commercial farmers who have the appropriate economic background. This means, culture and history of societies can be important for successful rangeland management for maintaining, improving or restoring rangeland condition.

References

- Allen CR, Angeler DG, Fontaine JF, Garmestani AS, Hart NM, Pope KL, and Twidwell D (2017) Adaptive management of rangeland systems. In: Briske DD (ed.) *Rangeland Systems. Springer Series on Environmental Management* pp. 373–394. Springer.
- Archer SR, Peters DPC, Burruss ND, and Yao J (2022) Mechanisms and drivers of alternative shrubland states. *Ecosphere* 13(4): 13:e3987.
- Batista PV, Davies J, Silva ML, and Quinton JN (2019) On the evaluation of soil erosion models: Are we doing enough? *Earth-Science Reviews* 197: 102898.
- Bidwell T, Elmore D, and Hickman K (2017) *Stocking Rate Determination on Native Rangeland*. Oklahoma Cooperative Extension Service. 2017. Available online <http://osufacts.okstate.edu>.
- Bilotta GS, Brazier RE, and Haygarth PM (2007) The impacts of grazing animals on the quality of soils, vegetation, and surface waters in intensively managed grasslands. *Advances in Agronomy* 94: 237–280.
- Bothma JDP, Van Rooyen N, and Van Rooyen M (2004) Using diet and plant resources to set wildlife stocking densities in African savannas. *Wildlife Society Bulletin* 32(3): 840–851.
- Briske DD (2017) *Rangeland Systems. Processes, Management and Challenges*. Springer Series of Environmental Management.
- Briske DD, Coppock DL, Illius AW, Fuhlendorf SD, and Niu K (2020) Strategies for global rangeland stewardship: Assessment through the lens of the equilibrium–non-equilibrium debate. *Journal of Applied Ecology* 57(6): 1056–1067.
- Byrnes RC, Eastburn DJ, Tate KW, and Roche LM (2018) A global meta-analysis of grazing impacts on soil health indicators. *Journal of Environmental Quality* 47(4): 758–765.
- di Virgilio A, Lambertucci SA, and Morales JM (2019) *Sustainable grazing management in rangelands: Over a century searching for a silver bullet*. 283: Agriculture, Ecosystems & Environment 106561.
- Erb KH, Lauk C, Kastner T, Mayer A, Theurl MC, and Haberl H (2016) Exploring the biophysical option space for feeding the world without deforestation. *Nature Communications* 7: 11382.
- FAO (2022) <https://www.fao.org/pastoralist-knowledge-hub/en/>. 2022. Available online: [Accessed].
- FAOSTAT (2022) FAOSTAT, Crops and Livestock, 2022. Available online <https://www.fao.org/faostat/en/#data/QCL>. Accessed.
- Gao J and Carmel Y (2020) A global meta-analysis of grazing effects on plant richness. *Agriculture, Ecosystems & Environment* 302: 107072.
- Gao TM, Zhang RQ, and Guo JY (2014) Research on wind erosion of Xilamuren Grassland, Inner Mongolia. In: *Advanced Materials Research*. Trans Tech Publications.
- Holchek JL, Pieper RD, and Herbel CH (2010) *Range Management Principles and Practices*, 6th edn. Upper Saddle River, NJ: Prentice-Hall Inc.
- ILRI, IUCN, FAO, WWF, UNEP, and ILC (2021) *Rangeland Atlas*. Nairobi, Kenya: ILRI.
- IPCC (2014) *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

- Kremen C and Merenlender AM (2018) Landscapes that work for biodiversity and people. *Science* 362(6412): 1101–1104.
- Lai L and Kumar S (2020) A global meta-analysis of livestock grazing impacts on soil properties. *PLoS One* 15(8): e0236638.
- Liu H, Dahlgren R, Larsen R, Devine S, Roche L, O'Geen A, Wong A, Covello S, and Jin Y (2019) Estimating rangeland forage production using remote sensing data from a small unmanned aerial system (sUAS) and planetscope satellite. *Remote Sensing* 11(5): 595.
- Lund HG (2007) Accounting for the world's rangelands. *Rangelands* 29(1): 3–10. [https://doi.org/10.2111/1551-501X\(2007\)29\[3:AFTWRJ\]2.0.CO;2](https://doi.org/10.2111/1551-501X(2007)29[3:AFTWRJ]2.0.CO;2).
- Martin G, Barth K, Benoit M, Brock C, Destruel M, Dumont B, Grillot M, Hübner S, Magne M-A, Moerman M, Mosnier C, Parsons D, Ronchi B, Schanz L, Steinmetz L, Werne S, Winckler C, and Primi R (2020) Potential of multi-species livestock farming to improve the sustainability of livestock farms: A review. *Agricultural Systems* 181: 102821.
- McCord SE and Pilliod DS (2022) Adaptive monitoring in support of adaptive management in rangelands. *Rangelands* 44(1): 1–7.
- Mitchell D, Snelling EP, Hetem RS, Maloney SK, Strauss WM, and Fuller A (2018) Revisiting concepts of thermal physiology: Predicting responses of mammals to climate change. *The Journal of Animal Ecology* 87(4): 956–973.
- Myers-Smith IH, Kerby JT, Phoenix GK, Bjerke JW, Epstein HE, Assmann JJ, John C, Andreu-Hayles L, Angers-Blondin S, and Beck PS (2020) Complexity revealed in the greening of the Arctic. *Nature Climate Change* 10(2): 106–117.
- Poland TM, Patel-Weyand T, Finch DM, Miniati CF, Hayes DC, and Lopez VM (2021) Invasive species in forests and rangelands of the United States. In: *A Comprehensive Science Synthesis for the United States Forest Sector*. Springer Nature.
- Rasch S, Heckelet T, and Johannes Oomen R (2016) Reorganizing resource use in a communal livestock production socio-ecological system in South Africa. *Land Use Policy* 52: 221–231.
- Redfearn DD and Bidwell T (2017) *Stocking Rate: The Key to Successful Livestock Production*. Oklahoma Cooperative Extension Service, PSS-2871. 2017. Available online <http://osufacts.okstate.edu>.
- Reed MS, Stringer LC, Dougill AJ, Perkins JS, Athlapheng JR, Mulale K, and Favretto N (2015) Reorienting land degradation towards sustainable land management: Linking sustainable livelihoods with ecosystem services in rangeland systems. *Journal of Environmental Management* 151: 472e485.
- Roesch A, Weisskopf P, Oberholzer H, Valsangiacomo A, and Nemecek T (2019) An approach for describing the effects of grazing on soil quality in life-cycle assessment. *Sustainability* 11(18): 4870.
- Sandhage-Hofmann A (2016) *Rangeland Management, Reference Module in Earth Systems and Environmental Sciences*. Elsevier.
- Sandhage-Hofmann A, Kotzé E, van Delden L, Dominiak M, Fouché HJ, van der Westhuizen HC, Oomen RJ, du Preez CC, and Amelung W (2015) Rangeland management effects on soil properties in the savanna biome, South Africa: A case study along grazing gradients in communal and commercial farms. *Journal of Arid Environments* 120: 14–25.
- Sandhage-Hofmann A, Löffler J, Kotzé E, Weijers S, Wingate V, Wundram D, Weihermüller L, Pape R, du Preez CC, and Amelung W (2020) Woody encroachment and related soil properties in different tenure-based management systems of semiarid rangelands. *Geoderma* 372: 114399.
- Soto RL, Martínez-Mena M, Padilla MC, and de Vente J (2021) Restoring soil quality of woody agroecosystems in Mediterranean drylands through regenerative agriculture. *Agriculture, Ecosystems & Environment* 306: 107191.
- SRM (2015) Glossary of Terms, Society for Range Management. Available online <http://globalrange-lands.org/rangelandswest/glossary>.
- Starns HD, Fuhlendorf SD, Elmore RD, Twidwell D, Thacker ET, Hovick TJ, and Luttbeg B (2019) Recoupling fire and grazing reduces wildland fuel loads on rangelands. *Ecosphere* 10(1): e02578.
- Teague WR, Dowhower SL, Baker SA, Haile N, DeLaune PB, and Conover DM (2011) Grazing management impacts on vegetation, soil biota and soil chemical, physical and hydrological properties in tall grass prairie. *Agriculture, Ecosystems & Environment* 141(3–4): 310–322.

Effects on soil of grassland management for pasture, hay and silage

Carmen Segura^a, Laura Cardenas^a, Richard McDowell^b, Sarah Morgan^c, and Martin SA Blackwell^a, ^aNet Zero and Resilient Farming, Rothamsted Research North Wyke, Okehampton, United Kingdom; ^bAgResearch, Lincoln Science Centre, Christchurch, New Zealand; ^cAgriculture and Environment Department, Harper Adams University, Newport, Shropshire, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction	102
What are grasslands?	102
Grassland soils	103
Soil carbon sequestration in grasslands	104
Grassland production systems	104
Impact of grassland production systems on soil properties	105
Effect of grazing on soil properties	106
Impacts of grazing under continuous-stocking management vs. no grazing (exclosure) on soils	106
Impacts of rotational or time-controlled grazing vs. continuous grazing on soil properties	107
Effect of silage and hay production and mixed management (grazing and mowing) on soil properties	107
Management strategies for pasture, hay and silage	108
Application of inorganic fertilizers	108
Application of organic amendments on grassland soils	108
Application of lime on grassland soils	109
Irrigation of grassland soils	109
Re-seeding	109
Impact of different vegetation/species	110
Silvopasture	110
Conclusion	111
References	111

Abstract

In this chapter we address how grassland management affects soils, recognizing that this relationship is not unidirectional, and it is driven by interacting factors such as soil texture, climate, botanical composition, production systems (grazing or mowing for hay and silage) and management practices. Grasslands that are managed for pasture, hay and silage would not naturally be grasslands, but have been established on areas that historically would have been forest but have been cleared and developed for agricultural purposes. Consequently, grasslands have diverse soil textures that influence other physico-chemical properties and processes in soil.

Key points

- The focus of this chapter is on soils under temperate agricultural grasslands, specifically those that are managed for pasture, hay and silage production.
- Soil processes in grasslands are mainly driven by soil texture, climate, botanical composition, production systems (grazing or mowing for hay and silage) and management practices (e.g., grazing intensity, fertilizer application, irrigation, re-seeding).
- Impacts of grazing on soils are affected by the degree of defoliation, urine and dung deposition and trampling, all of them influenced by stocking rates and the type of grazing system. Impacts of mowing on soils are affected by cutting frequency and soil compaction from vehicle traffic.

Introduction

What are grasslands?

Grassland is one of the largest of five major categories of ecosystem defined by the Pilot Analysis of Global Ecosystems (White et al., 2000), the others being forest ecosystems, agroecosystems, coastal ecosystems and freshwater systems. Broadly, the term grassland refers to “terrestrial ecosystems dominated by herbaceous and shrub vegetation and maintained by fire, grazing, drought or freezing temperatures.” However, the definition of grassland varies considerably across studies and authors, and their distinction from rangelands (a term often used for grasslands) can be confusing (see Rangelands chapter in this Encyclopedia). In general, we could

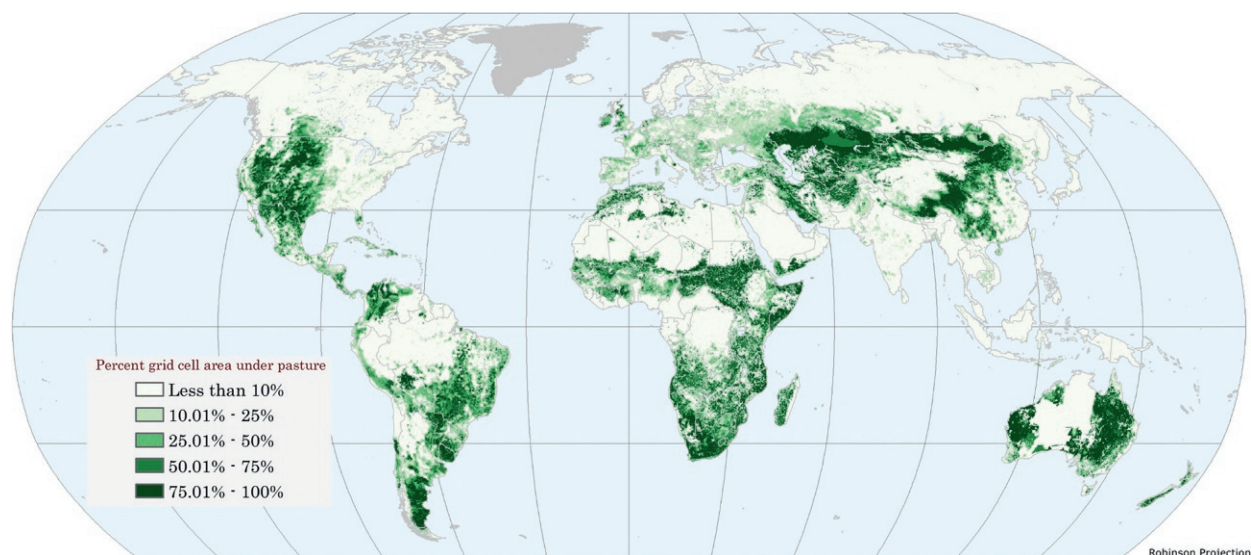


Fig. 1 Pasture global distribution in 2000. The Global Pastures data set represents the proportion of land areas used as pasture land (land used to support grazing animals) in the year 2000. Source: Center for International Earth Science Information Network (CIESIN), Columbia University. Global Agricultural Lands: Pastures, 2000: Global [Map]. 2012. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). <https://doi.org/10.7927/H47H1GGR>. Ramankutty N, Evan AT, Monfreda C, and Foley JA (2010) *Global Agricultural Lands: Pastures, 2000*. Palisades, New York: NASA Socioeconomic Data and Applications Center (SEDAC). <https://doi.org/10.7927/H47H1GGR>

say that rangeland supports predominantly native or naturalized vegetation capable of sustaining native or domestic grazing or browsing ungulates, whereas pastures (i.e., agricultural grasslands) are used for livestock production (Fig. 1).

The large extent of grasslands, which cover approximately 70% of the World's agricultural area, means that they are globally important for maintaining high levels of biodiversity and food production, regulating nutrient cycling and water quantity and quality, and supporting cultural services, such as recreation, tourism and cultural heritage. Globally, grasslands contain ca. 20% of global soil carbon stores and therefore play a major role in climate change (Stockman et al., 2013). However, around 49% of the area covered by grassland currently experiences some level of degradation due to overgrazing, eutrophication, land use changes, invasive species, climate extremes and fire (Bardgett et al., 2021). Agricultural grassland can be considered under three categories based on the intensity of their management:

- (1) Unimproved grassland—typically used for hay or extensive grazing and not receiving fertilizers
- (2) Semi-improved grassland—typically subject to additions of fertilizer to improve productivity and used for more intensive grazing or hay production
- (3) Improved grassland—typically plowed and re-seeded with grass species in monoculture, and heavily fertilized for intensive grazing or silage production.

Climatically, grasslands can be divided into tropical and temperate grasslands. Tropical grasslands are characterized by existing on highly weathered soils, which are more susceptible to nutrient and organic matter losses, and the dominance of C_4 grasses. In general, the typical rainfall patterns strongly affect biomass productivity and decomposition processes in tropical soils. The main threat in temperate grasslands is intensification, arising due to grazing, mowing, reseeding and inorganic and organic amendment applications, whereas tropical grasslands are threatened by overgrazing, fire and climate change. Different soil conditions and plant inputs to the soils reflect the differences in physico-chemical and biological processes in soils, such as the soil organic carbon turn-over rate (the time it takes for a carbon atom to pass from its entrance in an ecosystem to its exit), which generally is longer in tropical soils.

The focus of this chapter is on soils under temperate grasslands, which occupy 8% of the world's land area (White et al., 2000), specifically those that are managed for pasture, hay and silage production.

Grassland soils

In grasslands, soil processes are mainly driven by soil type, climate, botanical composition of the sward and management practices (Byrnes et al., 2018; Russell and Bisinger, 2015), including land use history and duration of management. Natural grasslands typically occur on Chernozems (Mollisols in US Soil Taxonomy), characterized by a thick, dark organic and nutrient rich surface horizon, but managed grasslands can occur on a wide range of soil types. Many of the grasslands that are managed for pasture, hay and silage would not naturally be grasslands, but have been established on areas that historically would have been forest but have been cleared and developed for agricultural purposes. Consequently, grasslands have diverse soil textures that influence other

physico-chemical properties and processes in soil. For example, fine-textured soils tend to have a greater water holding capacity than coarser-textured soils; in addition, fine textured particles also play a key role in soil aggregation. Precipitation and temperature directly affect primary production in grasslands as well as decomposition rates in soils. Unfavorable local weather conditions (too dry or too wet) can promote and aggravate soil physical damage, soil erosion, water quality degradation and greenhouse gas emissions—all negative effects that are associated with grazed pastures under high intensity (Soussana and Lemaire, 2014). Other climate effects are mediated via plant communities or soil faunal activity, both of which vary with climate (Ingram and Fernandes, 2001). An increase in plant biodiversity (richness and abundance) impacts positively on productivity and microbial activity (Lange et al., 2015), and can be promoted through grazing management, leading to a higher functionality of grassland ecosystems.

Soil condition can be determined through the measurement of key soil indicators, and these are crucial to develop appropriate and sustainable management strategies (Bardgett et al., 2021; White et al., 2000). For example, indicators such as soil nutrient and carbon content, soil depth, litter characteristics (mass, distribution and incorporation), soil respiration and enzymatic activity are linked to soil fertility and climate change mitigation, while soil stability and water regulation can be assessed by measuring soil organic matter (SOM), soil loss, soil aggregate stability, soil texture, bulk density, water holding capacity and electrical conductivity. Measurements of other chemical soil properties (pH, cation exchange capacity, salinity) are useful to evaluate the capacity of a soil to filter and retain pollutants.

Most grasslands have a large inherent SOM content, especially near the soil surface, that supplies plant nutrients, helps to reduce soil erosion and increases cation exchange and water holding capacities. Although the effect on water holding capacity is relatively small, organic matter creates macropores that can significantly increase water infiltration and gas transport. Therefore, in grassland systems it is vital to maintain the concentrations of SOM (Conant et al., 2017). Major organic matter inputs to grassland soils are provided by animal manures and roots (fine roots and root exudates), and relatively small amounts from plant litter as the majority of biomass is removed. All of these modify soil structure and soil biochemistry.

Soil carbon sequestration in grasslands

Temperate grasslands have an estimated soil organic carbon (SOC) sink capacity between 105 and 295 Gt C (0–1 m) according to Stockman et al. (2013). The potential of soil carbon sequestration is determined by several defining factors related to soil type and physical properties such as mineralogy, texture, stoniness, bulk density and aeration (Ingram and Fernandes, 2001). In general, under the same carbon input and climate, soils with fine texture can store more carbon than soils with coarse texture due to greater soil aggregation and the carbon protection role performed by clay and fine silt.

Increasing grassland biomass productivity, for example, by increasing nutrient application, and therefore carbon inputs to the soil, will enhance the capacity of grasslands to store carbon. However, the ability to maintain and stabilize soil carbon depends on climate, the degree of protection within soil aggregates, the soil carbon saturation level (highly dependent on soil texture), its quality and turnover rates, all factors that are strongly influenced by the types of grassland production system and management practices (Gilmullina et al., 2020; Necpalova et al., 2014).

In this chapter we address the question of how grassland management affects soil quality, recognizing that this relationship is not unidirectional, and it is driven by the abovementioned interacting factors. We also raise the question of suitability of soils for a particular type of grassland use, for example, is the land suitable for regular trafficking by heavy farm vehicles required for mowing and fertilizer spreading, or is better suited to light grazing? Can this be better managed taking account of seasonal changes (rain, snow, drought)?

Grassland production systems

Over the past 30 years, grasslands have been used increasingly for bioenergy production and there are attempts to include grass protein as part of the human diet. Nevertheless, the main purpose for producing grass within the food chain is to feed livestock that typically provide dairy or meat products for human consumption. In this sense, the main grassland production systems, classified according to the fate of the grass, are pasture, hay and silage.

Pasture is defined as land used for the production of grass together with associated species, such as legumes and forbs, which are directly consumed by grazing livestock (Fig. 2).

In many regions grazing is only possible during warmer, drier months of the year because the cessation of grass growth over winter means there is insufficient forage to feed the livestock. For this reason, the vegetation is cut to produce hay and silage that can be stored and fed to the animals. More specifically, hay is produced by cutting grasslands and allowing the vegetation to dry in the field before being pressed into bales (Fig. 3).

Silage can be produced from similar grasslands, but instead of drying before storage, it is either packed in silage clamps or in wrapped bales (Fig. 4), both of which must be airtight to enable fermentation and acidification, which preserves the forage. This results in a more easily digestible fodder that is more nutritious than hay (e.g., with higher protein content).

On many livestock farms, fields will be managed for both pasture and hay or silage (mowed grassland) in a single season, depending on the capacity for storage. However, on farms with permanently housed animal units, fields will typically be used solely for either silage or hay production with no grazing taking place.



Fig. 2 Cattle grazing on improved pasture in southwest UK. Courtesy Rothamsted Research.



Fig. 3 Haymaking on an unimproved pasture in southwest UK. Courtesy Rothamsted Research.

The management of grasslands, for either pasture or hay and silage production, impact soils in two distinctive ways: (i) the need for cutting and harvesting of grass, typically using heavy machinery in hay or silage production, and (ii) the lack of natural returns of organic matter (by excretion) to grasslands used for hay and silage production due to the lack of animals. These distinct production systems (grazed grasslands, mowed grasslands and the combination of both) and their impacts on soils form the focus of the next section.

Impact of grassland production systems on soil properties

Grassland production systems can vary in terms of their physical impact on the soil, biomass productivity, plant community structure and biogeochemical cycling, especially carbon and nitrogen cycles. Even though both grazed and mowed grasslands imply removing biomass from the system, specific factors associated with each production system can cause different impacts on soils.



Fig. 4 Collecting and wrapping silage from improved pasture in southwest UK. Courtesy Rothamsted Research.

Effect of grazing on soil properties

Impacts of grazing on soils are affected by the degree of defoliation, urine and dung deposition and trampling. There are many different ways in which pastures can be grazed, and the manner in which grazing is managed can have significant impacts on soils. Perhaps the most common management system is continuous grazing, whereby livestock are left permanently on the pasture throughout the grazing season, although stocking numbers may be adjusted according to the availability of grass. There is growing popularity for the practice of rotational grazing, sometimes referred to as “mob” grazing or adaptive multi-paddock (AMP) grazing. In these systems livestock are contained in small paddocks within larger fields. The length of time grazing is allowed within each paddock, and how long the rest period is, can vary considerably depending on the specific rotational stocking approach that is used, which can (or should) ultimately be determined by soil properties. The grazing time allowed within each paddock can vary anywhere from about 1 week to a few hours, and equally the rest period can range anywhere from 21 days up to 50+ days.

Impacts of grazing under continuous-stocking management vs. no grazing (exclosure) on soils

Continuous stocking describes a method of grazing management where animals have unrestricted and uninterrupted access to large units of land or pasture for a prolonged period of time when grazing is allowed (Allen et al., 2011). The main management variable in a continuous-stocking grazing system is the stocking density, which can be broadly categorized into light-, moderate- and heavy-stocking or grazing. There is a wide range of soil responses to different grazing intensities, and the effects are frequently inconsistent due to an array of site-specific factors which play important moderating roles (Byrnes et al., 2018). Russell and Bisinger (2015) reported that SOM concentrations are higher in grazed pastures compared with non-grazed pastures, which in turn increases soil carbon and nutrient sequestration and water-holding capacity. The positive impact of grazing on SOC has also been reported by others, even when productivity is reduced. In such cases organic matter inputs as urine and dung deposited by grazing livestock counterbalance the reduced plant inputs, and belowground inputs can be boosted as a result of a greater sward diversity (Bai and Cotrufo, 2022).

Conversely, Wang et al. (2011) reported that overgrazing resulted in decreases in soil carbon stocks in China, but this effect was reversed by exclusion of animals over 30 years. Similarly, Byrnes et al. (2018) concluded from their meta-analysis that continuous grazing, regardless of grazing intensity, significantly reduced SOC, carbon: nitrogen ratio and total nitrogen content compared to no grazing (exclosure); but reduced grazing intensity tended to lower soil bulk density and increase SOC. A recent meta-analysis by Lai and Kumar (2020) found that heavy-grazing significantly increased the soil bulk density and penetration resistance, and reduced SOC and water content, along with nitrate, microbial biomass carbon and total nitrogen concentrations compared with soils exposed to no-grazing. Moderate-grazing had a less negative impact on soils, but still resulted in significantly increased bulk density, penetration resistance, pH, and phosphorus concentration (0–10 cm) but decreased phosphorus (10–30 cm), SOC (10–30 cm) and total nitrogen (10–30 cm) concentrations. Light-grazing had minimal negative impact on soils, resulting in increased SOC (0–10 cm) and soil ammonium concentrations compared to no grazing (Lai and Kumar, 2020). Banerjee et al. (2000) reported light, continuous-grazing systems to have the largest soil microbial biomass and nutrient mineralizing activity compared to heavy, continuous or rotational grazing. Furthermore, Zhao et al. (2017) found that high-intensity, continuous grazing

results in losses of both soil bacterial and fungal communities, and an increase in the ratio of soil fungal to bacterial communities, with associated decreases in nutrient turnover rates.

The species of grazing animals can also influence the impacts of grazing on the soil. Although changes in SOC can be linked to stocking density of livestock, for sheep the stocking density generally has little impact on soil carbon stocks. [Lai and Kumar \(2020\)](#) stated that cattle grazing had a greater impact on SOC, total nitrogen and available potassium compared with sheep grazing, but less impact on soil penetration resistance.

Impacts of rotational or time-controlled grazing vs. continuous grazing on soil properties

Rotational grazing is becoming an increasingly popular form of grazing management, which involves recurring periods of grazing and rest among three or more paddocks within an area of grazing land ([Allen et al., 2011](#)). In the United States, a study by [Conant et al. \(2003\)](#) found total soil carbon was 22% higher under a rotational system compared with a continuously grazed system while [Oates and Jackson \(2014\)](#) reported that soils under management-intensive, rotational grazing had a favorable carbon content compared with soils under continuous grazing or used for making hay. In contrast, no differences in soil carbon stocks were found when comparing soils under rotational grazing with those under continuous grazing in a Canadian study.

Another grazing strategy is silvopasture, which is the practice of integrating trees, forage, and the grazing of domesticated animals in a mutually beneficial way. It utilizes the principles of managed grazing, is one of several distinct forms of agroforestry and has been identified as one of the most beneficial climate change solutions ([Feliciano et al., 2018](#)).

Effect of silage and hay production and mixed management (grazing and mowing) on soil properties

Mowing frequency, timing and cutting height impact soil properties via their influence on plant inputs to soils. In general, soil temperature, soil compaction and bulk density tend to increase as number of cutting events increase, whereas soil moisture and soil porosity decrease as a result of above-ground biomass removal. Macro and micronutrient availability can vary depending on the frequency of cutting, plant uptake and the antagonistic relationships between many elements in soils that affect their availability to plants.

One of the main impacts on soils under mowing is compaction resulting from vehicle traffic. The degree of compaction can vary, depending on machinery pressure and mowing frequency but also on soil texture (with sandy soils less prone to severe compaction than clay soils), soil structure and soil moisture. [Hargreaves et al. \(2019\)](#) reported that soil bulk density was greater in soils exposed to trafficking by vehicles rather than trampling by cattle. Therefore, silage fields, which are exposed to multiple passes by vehicles during fertilizer application, cutting, windrowing and harvesting, up to four times a year, are more prone to compaction, with susceptibility to compaction greatest when soils are wet. This has knock-on effects, reducing yield and also nutrient use efficiency due to the smaller uptake of applied nutrients. Inefficient use of nutrients can result in losses via runoff and subsequent eutrophication of surface waters, or in the case of nitrogen, increases in rates of denitrification and greenhouse gas emissions from soils. The implementation of double cropping for silage production (e.g., conventional tillage winter rye harvested in May or June followed by no-till corn silage) can improve soil structure, and increase saturated hydraulic conductivity and the resilience of the soil to traffic.

In managed grasslands in Southwest France, [Gilmullina et al. \(2020\)](#) found that mowed systems had smaller SOC and total nitrogen concentrations than grazed systems, although both systems had relatively large carbon stocks. These authors suggested that grazing was preferable to mowing because it leads to better substrate quality and more efficient microbial functioning under the same soil and climatic conditions. However, contrasting results have been observed under different climate conditions. Under an aridity gradient, [Liu et al. \(2022\)](#) reported that high-intensity grazing decreased plant species richness and SOC whereas hay production had less detrimental effects on ecosystem functions and enhanced plant species richness. In semiarid grasslands with sandy soils, alternating mowing and grazing promotes the coexistence of diverse plant species, balancing the nutrient cycling and ensuring good levels of biomass production ([Koncz et al., 2020](#)). Regarding plant diversity, there is no clear evidence that increasing plant species richness influences SOC and TN stocks under hay cut management in England. However, the presence of legumes in grasslands plays a key role in C and N sequestration ([Sollenberger et al., 2019](#)).

Vegetation recovery time after mowing is crucial to avoid soil degradation. In areas with a high frequency of hay or silage production, fertilizer application is required to compensate the depletion of soil nutrients over time ([Mayel et al., 2021](#)). However, the effects on soils of long-term applications of inorganic fertilizers and practices such as liming need to be monitored to prevent soil acidification and a decrease in soil nutrient availability.

There is some evidence that the combined management of grazing and cutting in a single season can have positive effects on soil properties ([Gavrichkova et al., 2008](#)) resulting from enhanced rhizodeposition and associated increased microbial activity. These were reported to be accompanied by decreased rates of SOM mineralization, meaning the microbial community that developed became more energy efficient. In these mixed management grasslands, soil compaction by animal trampling and vehicles is also affected by the combined effects of soil texture and climate, which can affect not only changes in soil structure but also grassland yield and N₂O fluxes. For example, damage to soil structure and reductions in silage production (as dry matter yield) have been observed in UK soils with a large clay content compared to sandier soils, mainly during cold wet weather ([Hargreaves et al., 2019](#)). In humid temperate grazed grasslands, increases in soil temperature, soil moisture and cutting events contributed to N₂O pulses

from soils in spring and summer, highlighting the fact that changes to the timing of grazing and mowing can reduce greenhouse (GHG) emissions.

In grasslands that are grazed over winter, supplementary feeding sites can result in localized increases in soil compaction from increased treading but can also have positive effects on SOM and soil nutrient concentrations, such as soil P and K (Sollenberger et al., 2019). In this case, the regular moving of feeding sites is recommended to favor homogeneity of excreta and nutrient distribution.

Management strategies for pasture, hay and silage

Grassland management strategies, such as the application of inorganic and organic fertilizers, liming and irrigation can increase carbon storage in mineral soils due to greater nutrient availability, biomass production and an improvement in soil aggregation (Conant et al., 2017), although these practices can also alter nutrient cycling and result in leaching and groundwater pollution from nutrients (Soussana and Lemaire, 2014). Many studies show the importance of optimized, sustainable grassland management that considers the local context. Conant et al. (2017) carried out a meta-analysis of 192 studies and found that fertilizer addition, increased species diversity (including legumes), irrigation, and reduced cultivation resulted in increases in soil carbon stocks of between 0.1 and 1.0 t C ha⁻¹ yr⁻¹. However, effects were specific to soil type, climate and sward mix. Grazing effects on soil and carbon sequestration also depend on grazing intensity as described above. Moderate and light grazing can improve SOC and bulk density, whereas heavy grazing can impact soil structure, reducing soil porosity and infiltration as a consequence of compaction (Mayel et al., 2021). In contrast, poorly managed grasslands can deplete SOM and alter the botanical composition, decreasing productivity and plant diversity, hence, reducing resilience of grasslands to climate change (Russell and Bisinger, 2015).

Application of inorganic fertilizers

All three grassland production systems require nutrient inputs to maintain productivity because each experiences a net removal of nutrients from the soil. Based on intensity and quantity of fertilizers applied, the order typically declines from silage to pasture and to hay (Table 1). However, due to the many interacting factors involved, it does not necessarily follow that the degree of negative impact on soils follows the same order. The objective of fertilizer application is to maintain productivity, and therefore theoretically the nutrient status of soils under each should, on average, be the same. The reason there are fewer inputs to hay soils is that typically only one cut per year will be made, whereas for silage, three cuts per year are typically made, and grazed pasture has continuous removal throughout the grazing season. However, pasture has the benefit of receiving excreted returns from livestock, thus replenishing the nutrient status of the soils, meaning they require fewer manufactured fertilizer inputs, even though the distribution of the “natural” fertilizer can be highly heterogeneous. Currently the inclusion of legumes such as clover into swards is increasing, meaning that the need to apply inorganic N fertilizers is greatly reduced due to its N fixation role, and in some cases eliminated. However, other inorganic fertilizers such as phosphate and muriate of potash are still required to provide macronutrient to the soil (P and K).

Overall, the effects on carbon stocks of nutrient applications to grassland soils are unclear because there aren't many field measurements of the effects of nutrient availability on production and carbon stocks. Schipper et al. (2017) concluded that the application of phosphorus to four flat and hill country sites in New Zealand over 24–60 years resulted in no changes in carbon stocks.

Application of organic amendments on grassland soils

Organic fertilizer refers to those organic amendments applied to soils, other than direct deposition of excreta by grazing animals, and includes animal manure, sewage sludge and compost, along with rendering waste, guano, brewery waste, digestate and other bio-wastes. They can contribute large amounts of nutrients to soils, for example, estimates of the potential amount of cattle manure produced in different countries show a range between 8,304,000.00 (UK) and 77,094,000.00 Mg y⁻¹ (US) (Thangarajan et al., 2013). They are favored from the farmer's point of view due to their potential to increase soil fertility and productivity but can be vulnerable to nutrient losses to water and the atmosphere due to priming effects and their susceptibility to promote nitrification and denitrification. Moreover, there is a concern about the potential risk of toxic elements inputs (e.g., heavy metals) mainly under

Table 1 Cutting regimes and typical fertilizer application rates for different grassland management systems.

Management	Cutting regime per annum	Typical N fertilizer rate per annum, N ha ⁻¹	Typical P fertilizer rate per annum, P ₂ O ₅ ha ⁻¹
Silage	3+ cuts	335 kg split over three applications	75 kg split over three applications
Pasture	Continuously grazed	240 kg split over three applications	45 kg split over three applications
Hay	1 cut	100 kg, single application	40 kg, single application

From <https://www.yara.co.uk/crop-nutrition/grassland/> Consulted on September 2022.

intense rainfall events when organic amendments are applied. A rigorous regulation should consider the field conditions to ensure a safe application and provide guidance. For instance, grazing is not allowed for 3 weeks after spreading sewage sludge in the UK, and there is a soil testing requirement. The method of application of organic amendments influences the form of nitrogen lost, for example, trailing shoe, splash plate.

Application of lime on grassland soils

Many grasslands are in areas of high rainfall, meaning that one of the main causes of grassland soil acidification is the leaching of cations down the soil profile. It can also be enhanced by the addition of ammoniacal fertilizer, predominantly in pasture and silage systems, and less so in hay systems. This can have important consequences for grass productivity, with the optimal soil pH typically being around 6.5 for most grassland species. To raise the pH of soils, lime is commonly applied. This can have significant impacts on soils, including increasing the availability of nutrients, especially phosphorus, as well as impacts on microbial communities. [Lochon et al. \(2018\)](#) reported that an increase in pH associated with liming resulted in both changes in microbial communities and increased microbial activity leading to increased SOM decomposition. Similar results have been seen in the Park Grass Long Term Experiment in the UK, where measurements have been taken over more than 100 years and have shown that adding lime increases SOM decomposition rates. However, effects on total SOM depend on how far the impacts on decomposition are offset by increased incorporation of carbon inputs into stabilized carbon pools.

In a study by [Sochorová et al. \(2016\)](#) hay yield was significantly increased by application of lime. Initially, the carbon content in the top 10 cm of soil also increased with liming, but decreased if inorganic nitrogen and phosphorus fertilizers were added simultaneously. Therefore, long-term hay production can lead to a reduction in soil carbon stocks. To protect soil quality, interactions between management strategies must be considered when making decisions aimed at increasing yield.

Irrigation of grassland soils

Not all grassland soils are irrigated, but climate change means that in many regions application of water is having to be considered for the maintenance of productivity, especially for silage and pasture systems. These high-production, irrigated systems differ from their dryland counterparts in nutrient requirements and losses, especially for carbon, nitrogen and phosphorus. For example, [Carmona et al. \(2020\)](#) reported that irrigated swards decreased the proportion of carbon apportioned to roots by 35% when irrigated, and in combination with an increase in root-derived carbon and changes in the forms of carbon upon decomposition, resulted in a net reduction in soil carbon. [Schipper et al. \(2013\)](#) also found that under the same phosphorus fertilizer regime, irrigated and sheep-grazed pasture treatments had up to 4.2% less soil carbon than a dryland treatment.

Under irrigation, an increase in pasture production requires more inputs of nitrogen and phosphorus to soils. If grazed, excess nitrogen from protein in the pasture is excreted in urine which can deliver nitrogen into soils at rates too high for the sward to utilize (e.g., dairy cattle urine patches contain 600–1000 kg nitrogen ha⁻¹), resulting in the surplus being available for loss as nitrate by leaching and runoff following nitrification. Because irrigation can produce large yields of pasture (e.g., 18 t of dry matter ha⁻¹ yr⁻¹), stocking rates of 4–5 cows per hectare are achievable, and if located on shallow stony soils can result in large leaching losses of nitrate-nitrogen (>60 kg nitrogen ha⁻¹ yr⁻¹) ([Chapman et al., 2021](#)).

For phosphorus losses from grassland systems, stocking rate is less important than available soil phosphorus concentrations. This is because pasture production protects the soil surface from treading damage and the loss of particulate phosphorus in runoff. Hence in contrast to dryland grasslands, much phosphorus loss from soils in irrigated grasslands occurs due to leaching in dissolved form rather than as the result of soil erosion. The magnitude of loss, however, depends on the quotient of available soil phosphorus concentrations and soil phosphorus sorption capacity. In most agriculturally productive soils, moderate or high sorption capacity can minimize losses if the soil is maintained at an agronomically optimum level (e.g., 20–30 mg phosphorus kg⁻¹ measured by Olson test, which was originally developed for neutral and alkaline soils, but is often applied across soils of all pH values), but can be insufficient if the soil is shallow or is supplied with too much or inappropriate forms of phosphorus. Indeed, over a period of 14 years, phosphorus losses in leachate increased by an order of magnitude (from 0.14 to 1.4 kg phosphorus ha⁻¹ yr⁻¹) where phosphorus was applied in effluents from housed dairy cows compared with the same amount of phosphorus being applied as superphosphate fertilizer ([McDowell et al., 2019](#)). Increased losses were caused by organic matter blocking phosphorus sorption sites and the presence of more mobile phosphorus forms in the effluent.

The type and timing of irrigation can affect nitrogen and phosphorus losses. Large losses of phosphorus have been noted (e.g., up to 42% of that applied) if irrigation was applied within 3 weeks of the fertilizer ([Hart et al., 2004](#)). Although some of these losses can be abated by using less water-soluble forms of phosphorus and reductions in phosphorus losses have been facilitated by moves toward spray irrigation.

Re-seeding

The re-seeding of pasture and silage systems is a common practice as new cultivars are developed, or mixed swards are replaced due to decreasing diversity, especially loss of legumes over time. Re-seeding without plowing the soil is particularly common in mixed swards to maintain diversity over time. However, the impact of re-seeding on soils can be significant after plowing because this practice promotes SOC and nitrogen losses via mineralization processes ([Necpalova et al., 2014](#)). The periods of time over which

reseeding occurs can vary, but commonly swards are replaced every 25 years, and Liáng et al. (2020) calculated that plowing soils with this periodicity could result in annual carbon losses of $0.16 \text{ t C ha}^{-1} \text{ yr}^{-1}$.

Impact of different vegetation/species

Many pastures and silage systems have moved from mixed swards to monocultures of perennial ryegrass (*Lolium perenne* L.), sometimes with nitrogen-fixing white clover (*Trifolium repens* L.). Grass breeding has tended to favor the allocation of resources to above-ground biomass rather than root systems, meaning carbon inputs to many pasture and silage systems soils have been reduced. However, mixed swards with large species diversity have often been shown to be more productive due to their varied plant traits, indicating there is little competition between component species for the same resources under periods of stress. Furthermore, such swards can also increase soil carbon stocks (Lange et al., 2015). The introduction of legumes into pastures has been shown to result in increases in soil carbon stocks, due to the stimulation of both plant and microbial growth by the increase in available nitrogen from fixation by the legumes. More generally, increasing sward diversity has been shown to maintain and even increase soil carbon stocks (Lange et al., 2015). Diverse swards generally require re-sowing every 5 years, and as mentioned previously, this can impact on soil carbon, but it may also alleviate topsoil compaction, even though plow pans can develop deeper in the profile. Another consequence of plowing is that it regularly mixes the topsoil, potentially reducing heterogeneity in nutrient distribution established during grazing, whereas under permanent or long-term pasture systems this does not occur. Despite the fact that plowing can benefit sward establishment when reseeding, soil disturbance can result in changes in soil structure and promote mineralization processes (Necpalova et al., 2014), contributing to SOC depletion in the topsoil. The absence of vertical mixing in long term pastures is well illustrated in Fig. 5, in which a pasture soil, not plowed for approximately 80 years, has a stone free upper horizon of approximately 5 cm depth, resulting from bioturbation activity depositing soil that has passed through organisms such as earthworms and deposited a stone free layer at the surface due to the inability for earthworms to pass stones.

Silvopasture

The adoption of silvopastoral systems that integrate trees and livestock, ideally at low to medium density, can enhance biodiversity and provide a wide range of soil-based ecosystem services, such as the regulation of nutrients and water quality (Mayer et al., 2022). Moreover, silvopasture can be a sustainable option to reduce the footprint of livestock production by increasing carbon sequestration and reducing gaseous emissions. Feliciano et al. (2018) found that the greatest increase in soil carbon sequestration worldwide ($4.38 \text{ t C ha}^{-1} \text{ yr}^{-1}$) resulted from the transition of a grassland system to a silvopastoral system. However, in temperate climates, there is considerable variation in SOC sequestration rates in silvopastoral systems. Some studies have not found significant increases in SOC stocks after grassland conversion into silvopasture systems, even after several decades (Fornara et al., 2018). This is likely due to soil legacy factors, or limited potential for SOC sequestration because the soil was already at its maximum possible level. However, planting trees has been seen to increase more stable soil carbon fractions and micro-aggregates, improving ecosystem resilience (Baah-Acheamfour et al., 2014). Douglas et al. (2020) suggested that changes in SOC could be associated with planted species rather than silvopasture implementation. Therefore, a careful selection of tree species and areas suitable for silvopasture establishment need to be considered before any land management change.



Fig. 5 Permanent pasture soil profile showing stone-free upper layer created by bioturbation over ca. 80 years in southwest UK. Image from Martin Blackwell.

Conclusion

Pasture, hay and silage can be produced on a wide range of soils and their productivity is affected by a combination of management, climate and soil type. In fact, the soil type often drives the type of management. Overall, the intensity of management tends to decrease in the order silage, pasture and hay, but it does not necessarily follow that soils under the less intensively managed hay systems are in better condition. This is largely due to the complexity of factors that interact in different soils. Despite there being some consistent differences in the way soils are managed under either silage, pasture or hay systems, the impacts on soil properties can be diverse under different conditions, demonstrating that management of soils is complex, as are soil systems in general.

References

- Allen VG, Batello C, Berretta EJ, Hodgson J, Kothman M, Li X, Mclvor J, Milne J, Morris C, Peeters A, and Sanderson M (2011) An international terminology for grazing lands and grazing animals. *Grass and Forage Science* 66(1): 2–28.
- Baah-Acheamfour M, Carlyle CN, Bork EW, and Chang SX (2014) Trees increase soil carbon and its stability in three agroforestry systems in central Alberta, Canada. *Forest Ecology and Management* 328: 131–139.
- Bai Y and Cotrufo MF (2022) Grassland soil carbon sequestration: Current understanding, challenges, and solutions. *Science* 377(6606): 603–608.
- Banerjee MR, Burton DL, McCaughey WP, and Grant CA (2000) Influence of pasture management on soil biological quality. *Journal of Range Management* 53(1): 127–133.
- Bardgett RD, Bullock JM, Lavorel S, Manning P, Schaffner U, Ostle N, Chomel M, Durigan G, Fry LE, Johnson D, and Lavelle JM (2021) Combatting global grassland degradation. *Nature Reviews Earth & Environment* 2(10): 720–735.
- Byrnes RC, Eastburn DJ, Tate KW, and Roche LM (2018) A global meta-analysis of grazing impacts on soil health indicators. *Journal of Environmental Quality* 47(4): 758–765.
- Carmona CR, Clough TJ, McNally SR, Beare MH, Tregurtha CS, and Hunt JE (2020) Seasonal irrigation affects the partitioning of new photosynthate carbon in soil. *Soil Biology and Biochemistry* 143: 107751.
- Chapman DF, Dalley DE, Edwards GR, Cameron KC, Malcolm BJ, Clement A, Romera AJ, Pinxterhuis IB, Beukes PC, Di HJ, Bryant RH, and Curtis J (2021) Production, profit and nitrogen flows in irrigated dairy systems representing different industry development pathways: The Pastoral 21 experience in Canterbury. *New Zealand Journal of Agricultural Research* 64: 3–35.
- Conant RT, Six J, and Paustian K (2003) Land use effects on soil carbon fractions in the southeastern United States: I. Management intensive versus extensive grazing. *Biology and Fertility of Soils* 36: 386–392.
- Conant RT, Cerri CE, Osborne BB, and Paustian K (2017) Grassland management impacts on soil carbon stocks: a new synthesis. *Ecological Applications* 27(2): 662–668.
- Douglas G, Mackay A, Vibart R, Dodd M, Mclvor I, and McKenzie C (2020) Soil carbon stocks under grazed pasture and pasture-tree systems. *Science of the Total Environment* 715: 136910.
- Feliciano D, Ledo A, Hillier J, and Nayak DR (2018) Which agroforestry options give the greatest soil and above ground carbon benefits in different world regions? *Agriculture, Ecosystems & Environment* 254: 117–129.
- Fornara DA, Olave R, Burgess P, Delmer A, Upson M, and McAdam J (2018) Land use change and soil carbon pools: Evidence from a long-term silvopastoral experiment. *Agroforestry Systems* 92(4): 1035–1046.
- Gavrichkova O, Moscatelli MC, Grego S, and Valentini R (2008) Soil carbon mineralization in a mediterranean pasture: Effect of grazing and mowing management practices. *Agrochimica* 52: 285–296.
- Gilmullina A, Rumpel C, Blagodatskaya E, and Chabbi A (2020) Management of grasslands by mowing versus grazing—impacts on soil organic matter quality and microbial functioning. *Applied Soil Ecology* 156: 103701.
- Hargreaves PR, Baker KL, Graceson A, Bonnett S, Ball BC, and Cloy JM (2019) Soil compaction effects on grassland silage yields and soil structure under different levels of compaction over three years. *European Journal of Agronomy* 109: 125916.
- Hart MR, Quin BF, and Nguyen ML (2004) Phosphorus runoff from agricultural land and direct fertilizer effects: A review. *Journal of Environmental Quality* 33: 1954–1972.
- Ingram JSI and Fernandes ECM (2001) Managing carbon sequestration in soils: Concepts and terminology. *Agriculture, Ecosystems & Environment* 87(1): 111–117.
- Koncz P, Vadász-Besnyői V, Csathó AI, Nagy J, Szerdahelyi T, Tóth Z, Pintér K, Fóti S, Papp M, Balogh J, and Gecse B (2020) Carbon uptake changed but vegetation composition remained stable during transition from grazing to mowing grassland management. *Agriculture, Ecosystems & Environment* 304: 107161.
- Lai L and Kumar S (2020) A global meta-analysis of livestock grazing impacts on soil properties. *PLoS One* 15(8): e0236638.
- Lange M, Eisenhauer N, Sierra CA, Bessler H, Engels C, and Griffiths RI (2015) Plant diversity increases soil microbial activity and soil carbon storage. *Nature Communications* 6: 6707. <https://doi.org/10.1038/ncomms7707>.
- Liáng LL, Kirschbaum MU, Giltrap DL, Wall AM, and Campbell DI (2020) Modelling the effects of pasture renewal on the carbon balance of grazed pastures. *Science of the Total Environment* 715: 136917.
- Liu J, Isbell F, Ma Q, Chen Y, Xing F, Sun W, Wang L, Li J, Wang Y, Hou F, and Xin X (2022) Overgrazing, not haying, decreases grassland topsoil organic carbon by decreasing plant species richness along an aridity gradient in Northern China. *Agriculture, Ecosystems & Environment* 332: 107935.
- Lochon I, Carrère P, Revalliot S, and Bloor MG (2018) Interactive effects of liming and nitrogen management on carbon mineralization in grassland soils. *Applied Soil Ecology* 130: 143–148.
- Mayel S, Jarrah M, and Kuka K (2021) How does grassland management affect physical and biochemical properties of temperate grassland soils? A review study. *Grass and Forage Science* 76(2): 215–244.
- Mayer S, Wiesmeier M, Sakamoto E, Hübner R, Cardinael R, Kühnel A, and Kögel-Knabner I (2022) Soil organic carbon sequestration in temperate agroforestry systems—A meta-analysis. *Agriculture, Ecosystems & Environment* 323: 107689.
- McDowell RW, Gray CW, Cameron KC, Di HJ, and Pellow R (2019) The efficacy of good practice to prevent long-term leaching losses of phosphorus from an irrigated dairy farm. *Agriculture, Ecosystems & Environment* 273: 86–94.
- Nečpalova M, Li D, Lanigan G, Casey IA, Burchill W, and Humphreys J (2014) Changes in soil organic carbon in a clay loam soil following ploughing and reseeded of permanent grassland under temperate moist climatic conditions. *Grass and Forage Science* 69: 611–624.
- Oates LG and Jackson RD (2014) Livestock management strategy affects net ecosystem carbon balance of subhumid pasture. *Rangeland Ecology & Management* 67: 19–29. <https://doi.org/10.2111/REM-D-12-00151.1>.
- Russell JR and Bisinger JJ (2015) Forages and Pastures Symposium: Improving soil health and productivity on grasslands using managed grazing of livestock. *Journal of Animal Science* 93(6): 2626–2640.
- Schipper LA, Dodd MB, Pronger J, Mudge PL, Upsdell M, and Moss RA (2013) Decadal changes in soil carbon and nitrogen under a range of irrigation and phosphorus fertilizer treatments. *Soil Science Society of America Journal* 77: 246–256.
- Schipper LA, Mudge PL, Kirschbaum MUF, Hedley CB, Golubiewski NE, Smail SJ, and Kelliher FM (2017) A review of soil carbon change in New Zealand's grazed grasslands. *New Zealand Journal of Agricultural Research* 60: 93–118.

- Sochorová L, Verbruggen E, Hejman M, Schellberg J, Kiers T, and Collins-Johnson N (2016) Long-term agricultural management maximizing hay production can significantly reduce belowground C storage. *Agriculture Ecosystems & Environment* 220: 104–114.
- Sollenberger LE, Kohmann MM, Dubeux JC Jr., and Silveira ML (2019) Grassland management affects delivery of regulating and supporting ecosystem services. *Crop Science* 59(2): 441–459.
- Soussana J-F and Lemaire G (2014) Coupling carbon and nitrogen cycles for environmentally sustainable intensification of grasslands and crop-livestock systems. *Agriculture, Ecosystems and Environment* 190: 9–17. <https://doi.org/10.1016/j.agee.2013.10.012>.
- Stockman U, Adams M, Crawford J, Field D, Henakaarchchi N, Jenkins M, Minasny B, McBratney A, De Remy de Courcelles R, Singh K, Wheeler I, Abbott L, Angers D, Baldock J, Bird M, Brookes P, Chenu C, Jastrow J, Lal R, Lehmann J, O'Donnell A, Parton W, Whitehead D, and Zimmermann M (2013) The knowns, known unknowns and unknowns of sequestration of soil organic carbon. *Agriculture, Ecosystems and Environment* 164: 80–99.
- Thangarajan R, Bolan NS, Tian G, Naidu R, and Kunhikrishnan A (2013) Role of organic amendment application on greenhouse gas emission from soil. *Science of the Total Environment* 465: 72–96.
- Wang S, Wilkes A, Zhang Z, Chang X, Lang R, Wang Y, and Niu H (2011) Management and land use change effects on soil carbon in northern China's grasslands: A synthesis. *Agriculture Ecosystems and Environment* 142: 329–340.
- White R, Murray S, and Rohweder M (2000) *Pilot Analysis of Global Ecosystems (PAGE): Grassland Ecosystems*. World Resources Institute. 70. http://pdf.wri.org/page_grasslands.pdf.
- Zhao F, Ren C, Shelton S, Wang Z, Pang G, Chen J, and Wang J (2017) Grazing intensity influences soil microbial communities and their implications for soil respiration. *Agriculture Ecosystems and Environment* 249: 50–56.

Dryland farming: Technological and management options for sustainable agriculture and food systems

Ch Srinivasarao^a, G Ranjith Kumar^a, Ravula Manasa^a, Kiran Pilli^a, Samaresh Sahoo^a, S Rakesh^a, Sumanta Kundu^b, K C Nataraj^c, KV Rao^b, JVNS Prasad^b, and S Malleswari^d, ^aICAR-National Academy of Agricultural Research Management, Hyderabad, India; ^bICAR-Central Research Institute for Dryland Agriculture, Hyderabad, India; ^cAgricultural College, Naira, Acharya N.G. Ranga Agricultural University (ANGRAU), Guntur, Andhra Pradesh, India; ^dAcharya N.G. Ranga Agricultural University, Agricultural Research Station, Srikakulam, Andhra Pradesh, India

© 2023 Elsevier Ltd. All rights reserved.

This is an update of G.A. Peterson, DRYLAND FARMING, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 414–417, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00260-5>.

Introduction	114
Problems in dryland farming	114
Climate change	114
Soils under dryland farming	115
Soil degradation of rainfed dryland ecosystems	115
Climatic variation in drylands	115
Dryland crops and cropping systems	116
Livelihood challenges in dryland ecosystems	116
Technological solutions for sustainable dryland farming	116
Water technology solutions	117
Micro irrigation	118
Hydrogel applications in water deficient areas	118
Soil management solutions	118
Land leveling	118
Conservation agriculture	118
Integrated nutrient management	119
Crop management solutions	120
Intercrops	120
Foliar sprays	120
Tolerant crops	120
Agroforestry	121
Integrated farming system	121
Dryland farming in India-case study	121
Conclusion and way forward	123
References	123
Further reading	124

Abstract

In dryland areas, production in agricultural and allied sectors is small and uncertain because of the scarcity of precipitation (<500 mm). Optimum utilization of resources (water, nutrients, forest, land etc.) is crucial to increase the agricultural output and improve the livelihoods of people living in dryland areas. Nutrition rich cereals, many traditional crops and large livestock populations are unique features of dryland ecosystems. For sustainable agricultural systems, efficient water management (in situ water harvesting, farm ponds, micro irrigation, hydrogels etc.), soil and land management (improving soil organic carbon (SOC), mulching, cover crops and biochar etc.), cropping practices (intercropping, conservation agriculture, tolerant crops and agro-forestry etc.) are critical.

Key points

- Climate variability in drylands areas
- Sustainable management options of dryland farming systems
- Soil degradation in dryland ecosystems
- Soil and crop management solutions in dryland areas
- Potential of dryland farming in different agro-ecological regions

Introduction

Drylands cover about 41% of the Earth's land surface with 38% of the world's population (Osman, 2018), which is distributed over approximately 6.1 billion hectares, and home to estimated 2 billion people, about 90% of them in developing countries (FAO, 2019). Most of the drylands are distributed in Africa (2000 M ha) Asia (2000 M ha), North America (760 M ha), Oceania (680 M ha), Europe (300 M ha) and South America (56 M ha). The majority of drylands are in continents like Asia (11 million km²) and Africa (nearly 13 million km²) (De Pauw, 2009). About 99% of the surface area of countries like Botswana, Burkina Faso, Iraq, Kazakhstan, Maldives and Turkmenistan is covered with drylands (Hori et al., 2011). The United Nations Convention to Combat Desertification (UNCCD) has indicated that drylands mainly occur in arid, semi-arid and dry sub-humid regions with an aridity index (AI) between 0.05 and 0.65. Drylands are classified into four types based on the AI as i) hyper-arid – AI <0.05, ii) arid – AI 0.05 to <0.20, iii) semi-arid – AI 0.20 to <0.50 and iv) dry sub-humid – AI 0.50 to <0.65. On the basis of the Aridity Index, 47.2% of the global land areas are drylands including: hyper-arid types ~7.5%, arid types ~12.1%, semi-arid types ~17.7% and sub-humid types ~9.9% (Hori et al., 2011). Different global land use systems in drylands are given in Fig. 1.

Drylands around the world encounter several constraints that limit crop productivity to a great extent. Most farmers in rainfed areas are small holders who have limited access to land, water resource, finance, market facilities, etc. Rainfed soils are usually coarse-textured and have small soil organic matter (SOM) content, low nutrient and water retention capacity, and multi nutrient deficiencies. As a result, crops grown in these soils result in inadequate yields. These lands are highly vulnerable to climate change because the agriculture sector is directly linked with climate that controls the rates of production. Crop production under moisture stress conditions presents a serious challenge to dryland farmers to achieve sustainability in crop production.

Problems in dryland farming

Climate change

Climate change in dryland regions is worsening soil conditions with prolonged periods of drought, accelerated desertification, and is reducing soil fertility by causing negative effects on biodiversity and vegetation cover – all of which undermine food and nutrition security. Climate change is gradually increasing the global area of drylands, estimates suggest that the area has already increased by 4% since 1948–1962, and by 2100 dryland areas have possibly increased by 11–23%, constituting up to 56% of the global land surface (Huang et al., 2017). By 2030, population growth linked with the expansion of drylands due to climate change might have increased the number of people living under challenging conditions by up to 70% (World Bank, 2017). The gap between farm yields and yield potential is small in developed countries like the European Union (EU) and USA (Global Yield Gap and Water Productivity Atlas), whereas in developing countries productivity is lower and the yield gap larger. Therefore, it is necessary to increase crop yields and stability over time via resistance and resilience to climate change. Increasing the temporal stability of crop yield is a major goal for smallholder dryland agriculture in the context of climate change.

Most of the countries in Africa already face semi-arid conditions that make agriculture challenging and climate change will be likely to reduce the length of the growing seasons and force large regions of marginal agriculture out of production. As per the research prediction model, it is suggested that climate change will reduce crop yields (specifically winter crops) up to 11% and 32%

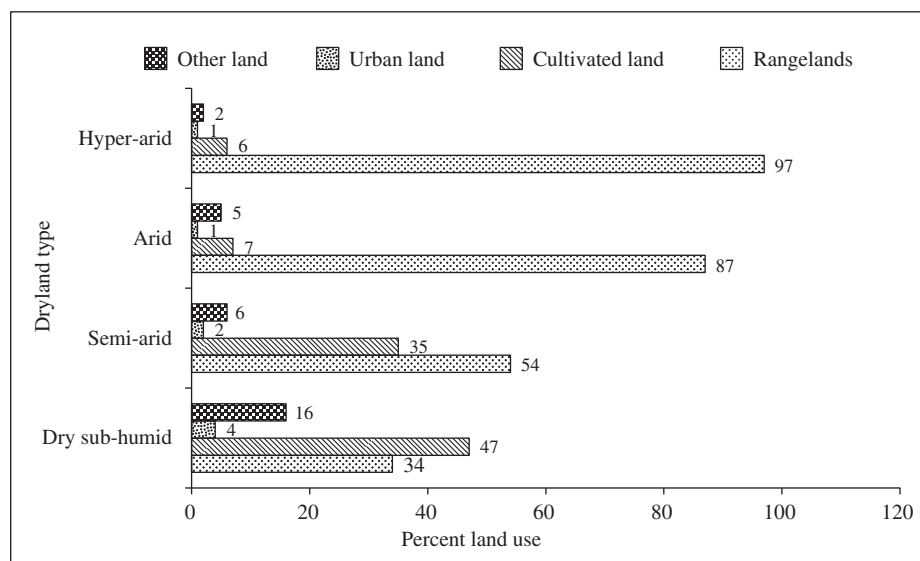


Fig. 1 Different global land use systems in drylands (Osman, 2018).

by 2050 and 2080, respectively in India (Kumar and Sharma, 2022). and this may increase by up to 90% by 2100 (Kumar et al., 2019). Without increased resilience, up to 3.8 billion people across the climate-vulnerable global drylands will be deprived of their livelihoods if the effects of intensifying climate change continue as predicted (ICARDA, 2020). Even if climate change follows only the present estimated trajectory, 10% of the total agricultural economic value could be lost worldwide by 2050 (Swiss Re Institute, 2021). However, the impacts of climate change will be much more extreme in dryland ecosystems, where the economy, and food and nutrition security are inherently linked with agriculture.

Soils under dryland farming

Global dryland soils belong to several orders of Soil Taxonomy such as Entisols, Alfisols, Aridisols, Vertisols and Mollisols (Soil Survey Staff, 2022). Dryland soils are classified as Calcisols, Gypsisols, Steppe soils and Leptosols in the World Reference Base for Soil Resources (WRB, IUSS, 2022). Climate, vegetation and parent materials are the primary factors of soil formation in drylands. Precipitation, temperature and evapotranspiration are major components of the climate. Under dryland environments rocks and minerals can undergo physical weathering, whereas because of the lack of moisture bio-chemical weathering is limited. Dryland soils have some common features all over the world such as coarse to medium texture, low moisture retention, accumulation of salts, lime (calcium carbonate) or gypsum, small organic matter contents, low biological activity and weak soil profile development (Srinivasarao et al., 2009).

Soil degradation of rainfed dryland ecosystems

Globally, the biophysical status of 5670 million ha of land is declining, of which 1660 million ha (29%) have human induced land degradation. Global land degradation has increased by 5–10 million hectares per year (Stavi and Lal, 2015). An increase in land degradation due to the increasing frequency of droughts was recorded in the Sub-Saharan region of Africa from 1968 to 1990 by Ibrahim et al. (2015). The impending land degradation causes greater threats to crop production and livelihoods of the people in developing countries. There is a gradual increase in soil degradation in India, with estimates of about 147 million hectares (Mha) of land area that has degraded (Bhattacharyya et al., 2015). However, various organizations reported different values of degraded land area in India viz., 107.4 Mha in 1994 reported by the Ministry of Agriculture (MoA) and 146.8 Mha in 2004 estimated by the National Bureau of Soil Survey and Land Use Planning (NBSS&LUP). The process of land degradation has mainly been caused by the removal of vegetation in dry, sub-humid and semi-arid regions enhancing soil erosion by wind, which is dominant in arid regions (Srinivasarao et al., 2021). More than 70% of land degradation was due to water and wind erosion, and 15% due to salinity.

Climatic variation in drylands

Weather is always a major concern in dryland farming, climatic factors such as precipitation, temperature, humidity and wind determine the climatic conditions of an area. Variation in climatic factors is a major concern in all agricultural regions, but it is the dominant factor in dryland farming regions. Annual precipitation in semi-arid region ranges from less than 50% to more than 200% of the annual average, and crop yields vary from zero to three or more times the average (Stewart, 2016). There is clear evidence that global surface temperatures are rising, and this is a serious concern for dryland regions because there is little evidence of increasing precipitation. Therefore, these already dry regions may become even drier. Global surface temperature has risen by 0.08 °C per decade since 1880, and the rate of warming over the past 40 years has been more than twice that at 0.18 °C per decade since 1981. The April 2022 global average surface temperature was 0.85 °C warmer than the 20th century average and tied with the average surface temperature in April 2010 as the fifth highest of the 143-year record (<https://www.ncei.noaa.gov/access/monitoring/monthly-report/global/202204>). All three locations mentioned in Table 1 are semi-arid (Aridity Index 0.2–0.5), but the climates are very diverse and require completely different cropping systems and management. Amman, Jordan has a humid period during the winter months. Rajkot, India, also has a humid period during the summer when precipitation exceeds potential evapotranspiration. In contrast, in Bushland, Texas, potential evapotranspiration exceeds average rainfall amounts every month of the year. With regard to the growing days Amman (119 days) and Bushland (0 days) are classified as sub humid and dry, but Rajkot (96 days) is considered as semi-arid.

Table 1 Agro climatic characteristics of different dryland farming locations (Stewart, 2016).

Location	Precipitation (P) (mm)	Potential Evapotranspiration (ETP) (mm)	P/ETP
Rajkot, India	674	2430	0.28
Texas, USA	470	1880	0.25
Ammon, Jordan	417	1608	0.26

Table 2 Major dryland crops and cropping systems in various regions across the world.

Region/Country	Dryland crops and Cropping systems
India	Mung bean (<i>Vigna radiata</i> L.), Moth bean (<i>Vigna aconitifolia</i> Jacq) Marechal), Cowpea (<i>Vigna unguiculata</i> L. Walp), Cluster bean (<i>Cyamopsis tetragonoloba</i> L. Taub), Pearl millet (<i>Pennisetum glaucum</i> L.) and Sesame (<i>Sesamum indicum</i> L.) Crops: Pearl millet (<i>Pennisetum glaucum</i>), Sorghum (<i>Sorghum bicolor</i> L. Moench), Cluster bean (<i>Cyamopsis tetragonoloba</i> L.), Pigeon pea (<i>Cajanus cajan</i> L.), Cowpea (<i>Vigna unguiculata</i> L.), Sesame (<i>Sesamum indicum</i> L.) and Groundnut (<i>Arachis hypogaea</i> L.).
Nepal	Elephant foot yam (<i>Amorphophallus paeoniifolius</i> (Dennst.) Nicolson), taro (<i>Colocasia spp</i> (L.) Schott), cassava (<i>Manihot esculenta</i> Crantz), winter bean (<i>Vicia faba</i> L.), air potato (<i>Dioscorea bulbifera</i> L.), cush-cush yam (<i>Dioscorea trifida</i> L.f.) and brinjal (<i>Solanum melongena</i> L.), sugarcane (<i>Saccharum officinarum</i> L.), turmeric (<i>Curcuma longa</i> L.), ivy gourd (<i>Coccinia grandis</i> (L.) Voigt) and legumes
Pakistan	Wheat, maize, sugarcane, pulses and vegetables etc.
Australia	Wheat, grain legumes such as field peas and lupins, oilseed crops such as canola or sunflower (<i>Helianthus annuus</i> L.), pasture legumes such as subterranean clover (<i>Trifolium subterraneum</i> L.) and medics (<i>Medicago spp.</i>)
The Canadian Prairies and the U.S. Northern Great Plains	The traditional rotation of spring wheat-summer fallow has largely been replaced with rotations based on apulses, such as lentil (<i>Lens culinaris</i> Medik.), field pea, chickpea (<i>Cicer arietinum</i> L.), and faba-bean (<i>Vicia faba</i> L.), and oilseed crops, such as canola (<i>Brassica napus</i> L.) and sunflower (<i>Helianthus annuus</i> L.), which are commonly grown in rotation with spring wheat.
U.S. Central Great Plains	Crop rotation of winter wheat with summer fallow. Other crops rotated with wheat in the Central Great Plains include sorghum, sunflower, and proso millet (<i>Panicum miliaceum</i> L.).
U.S. Southern Great Plains	Winter wheat is commonly grown in rotation with grain sorghum and other dryland crops include maize, cotton (<i>Gossypium hirsutum</i> L.), and proso millet.
Eastern and Southern Africa (ESA)	Maize (<i>Zea mays</i> L.), sorghum (<i>Sorghum bicolor</i> (L.) Moench), wheat (<i>Triticum aestivum</i> L.), barley (<i>Hordeum vulgare</i> L.), teff, chickpea (<i>Cicer arietinum</i> L.), faba bean (<i>Vicia faba</i> L.), enset bean (<i>Ensete ventricosum</i> (Welw.) Cheesman) and coffee (<i>Coffea Arabica</i> L., <i>Coffea Canefora</i> L.), cassava (<i>Manihot esculenta</i> Crantz), millet (<i>Panicum miliaceum</i> var. <i>miliaceum</i> L.), beans (<i>Phaseolus vulgaris</i>), sweet potato (<i>Ipomoea batatas</i> (L.) Lam.), Cotton (<i>Gossypium hirsutum</i> L.), tea (<i>Camellia sinensis</i> (L.) Kuntze)

Dryland crops and cropping systems

There are many constraints that limit crop yields in dryland areas, and dryland farming will become increasingly important to produce adequate food and fiber for the ever-increasing world population. Water supplies for expanding dryland farming areas are becoming increasingly limited and more expensive to develop. Wheat (*Triticum aestivum* L.), barley (*Hordeum vulgare* L.), millets and pulses (Pigeon pea (*Cajanus cajan* (L.) Millsp.), Chick pea (*Cicer arietinum* L.) and Cowpea (*Vigna unguiculata* L.) etc.) are the major global dryland crops (Table 2). Wheat and barley are the most important cereal crops in the Mediterranean, where high temperatures and water scarcity restrict yield during grain filling (Jacobsen et al., 2012). In South Asia, cereals are cultivated in all countries except the Maldives. In Pakistan, 90% of the country's food grain production comes from rice (*Oryza sativa* L.) and wheat. Potato (*Solanum tuberosum* L.) is one of dominant crops in most dryland agricultural regions in China because the demand for water by potato during the growing season largely matches the seasonal distribution of rainfall and temperature. Dryland cropping zones in Australia mainly occur between the 300 and 600 mm rainfall isohyets. The Great Plains region of the United States and Canada is an area of widespread dryland crop production, with wheat being the dominant crop. In the dryland systems in Eastern and Southern Africa (ESA), maize (*Zea mays* L.) is the dominant crop and accounts for 60–70% of the total cropped area in Zambia and Zimbabwe and >90% of the total cereal production in Zimbabwe, Kenya, Malawi, Zambia and Mozambique.

Livelihood challenges in dryland ecosystems

In addition to crop production and productivity issues, there are several livelihood challenges in dryland areas which the farmers are having to face including small and fragmented land holdings, lack of resources, poor infrastructure combined with severe unemployment for most of the year.

Technological solutions for sustainable dryland farming

Most of the countries with malnutrition and famine around the globe are in dryland areas, and crop production in many of these countries is expected to become more challenging; these regions already have huge gaps in potential crop yield. Technology is the major driving force for agricultural growth. Improved technologies should result in increased gains in crop production which can be significant over time. Considering the costs and constraints of resources such as water, nutrients and energy, the genetic enhancements in crop characteristics and their productivity levels should be coupled with improved efficiency of inputs. This will be possible only by using the existing improved technologies and by developing new ones. With advanced management and

greater inputs of water and fertilizers in poorly resourced dryland areas their efficiency of use could increase together with enhanced yields by 2–5-fold. Specialized dryland management practices such as water harvesting and reducing soil moisture loss can increase yields by 5–15% on average across the semi-arid tropics (SAT) regions and reduce year-to-year yield variability (Fischer et al., 2009).

Water technology solutions

In arid and semiarid regions, potential evaporation is generally higher than precipitation and non-uniform in distribution, resulting in frequent drought periods during the crop-growing season. Furthermore, surface runoff usually results from the high intensity rains and dry soil surface that resists water infiltration. As the water shortage in dry areas is a recurring crisis, people need information on how to capture and use every available drop of water efficiently. Water harvesting is a method of water collection that has been applied historically in arid and semiarid regions where rainfall is either not sufficient to sustain good crop growth or where, due to the erratic nature of precipitation, the risk of crop failure is very high. Water harvesting is now being applied all over the world. As new water harvesting techniques have been developed, more regions are employing water harvesting to help offset pressures on existing water resources. In Asia, many communities have emerged and are thriving in spite of the harsh arid conditions, and their social life has evolved around water scarcity and indigenous water harvesting techniques. Water harvesting works by collecting rainwater and runoff from a large catchment area into a small target area. In Sri Lanka, policymakers have focused on alleviating seasonal water scarcity in the dry zone with large-scale storage tanks and inter-basin transfers. In Nepal, community management of water systems has been popular (Ariyabandu, 2008). Supplemental irrigation during mid-season droughts potentially increases and stabilizes dryland land crops by 30–40%. Among the many types of intervention, establishing farm ponds is the most important and promising technology in the integrated watershed management program. It also provides other environmental benefits. Farm ponds can have multiple uses such as supplemental irrigation through efficient methods of water application (drip and sprinkler irrigation), fish culture or duck farming integrated with poultry, and also protect from scarcity of precipitation. Farm ponds and check dams minimize the water stress at critical stages of crop growth, which can lead to an increase in crop yield and water productivity in dry regions (Fig. 2). These structures provide local water and ensure food security by enhancing crop productivity and climate resilience.



Farm ponds



Check dam

Fig. 2 Rain water harvesting structures (Rangareddy District, Telangana State, India).

Micro irrigation

In dryland regions, fertilizer-use efficiency is lower than that in irrigated agriculture mainly because of poor fertilizer and crop management practices, and water stress at critical growth stages. Efficient use of harvested water through pressurized micro irrigation systems is crucial for enhancing water-use efficiency (WUE) thereby achieving greater crop productivity in drylands. Of a total area of about 11.1 M ha under micro-irrigation in the world, the greatest coverage has been recorded in Asia and Oceania (5.6 M ha), followed by Europe and North America (2.5 M ha each) and Africa (0.4 M ha). Israel is a good example, it is a desert nation with water scarcity that has become a water surplus nation because growers have adopted micro-irrigation practices, especially drip irrigation that has reduced the water used for irrigation by about 75%. The area under micro irrigation in India is about 60.9 M ha which represents 8.1% of the total irrigated area in the International Commission on Irrigation and Drainage (ICID) countries; while Israel has 99.63% of its total irrigated area (0.23 M ha) under micro irrigation (Saxena and Ramana Rao, 2019). Dryland soils are severely deficient in bioavailable nutrients, and fertilizer application with irrigation is an important technological option for efficient water and nutrient management. Fertilizer-use efficiency through fertigation ranges between 80 and 90%, which helps to save a minimum of 25% of nutrients. Among different fertigation methods, drip irrigation systems provide 90% efficacy, sprinkler irrigation about 70–80% and conventional surface irrigation about 60–70% (<https://www.niti.gov.in/>). Nutrient losses through erosion can be effectively minimized by the fertigation technique, which significantly improves the WUE, water productivity, crop yields and livelihoods.

Hydrogel applications in water deficient areas

Hydrogels are synthetic polymers, able to absorb and hold water 80–180 times of their volume because their hydrophilic functional groups swell upon contact with water while retaining their structure. Application of hydrogels at 2 g kg⁻¹ improved the water holding capacity of sand from 171% to 402%, which reduced the irrigation requirements. Hydrogels increase water availability in topsoils and enhance field crop germination by 2–4 days, improve crop survival by 7.2–28.6% and also have positive long-term effects on water retention throughout the entire cropping season (Tibirkov et al., 2021).

Soil management solutions

Soils hold the key to productivity and resilience to climate vagaries in dryland agriculture. The potential limiting factors for productivity enhancement in drylands are the loss of fertile soil through erosion, depletion of soil organic matter, emerging secondary and micronutrient deficiencies, soil compaction, surface crusting, and loss of soil biodiversity. Many common agricultural practices, especially plowing, disk-tillage and vegetation burning, accelerate the decomposition of SOM which results in soils being susceptible to wind and water erosion (FAO, 2005). Sequestration and rates of decomposition of SOC influence soil structure and porosity; rate of water infiltration and water holding capacity; diversity and biological activity of soil organisms; and plant nutrient availability. However, recommended management practices (RMPs) like mulch farming, conservation agriculture, cover cropping, agroforestry enhance soil health and allow sustained agricultural productivity. These practices need to be promoted as contributing to soil, land and water conservation.

Land leveling

Land leveling mitigates climate change by reducing the demand for irrigation water and improving water-use efficiency and it also reduces farm machinery operating time by 10–15%. The latter reduces the time of tractor working hours by 3¼ hours ha⁻¹ compared with traditional plowing, harrowing and leveling. For every hour of use, tractors of 37,300–41,030 W power use 4.25 L of diesel and emit 2.6 kg CO₂ L⁻¹ of diesel. Laser land leveling technology increased the land-use efficiency in cropping by 3–6% when several smaller fields were combined into one larger field. It also increased WUE by 12–40%, fertilizer-use efficiency by 10–13%, and reduced methane gas emissions by at least 20% (Sander et al., 2014). Laser land leveling includes benefits such as saving land, water, and agronomic inputs, increasing yield, and decreasing postharvest losses. It also results in the saving of energy of 3.0–6.9 GJ ha⁻¹ and decreases emissions by 1151–1486 kg CO₂-eq ha⁻¹ in countries like Cambodia, Philippines, Thailand, Vietnam and India (Nguyen-Van-Hung et al., 2022).

Conservation agriculture

Conservation agriculture (CA) practices such as no-tillage or reduced tillage, mulching, cover crops and crop residue recycling can effectively maintain soil moisture, increase yield and income, and improve farmland ecosystems in dryland ecosystems. Intensive soil cultivation breaks down soil organic matter (SOM) resulting in the loss of soil organic carbon in the form of CO₂; thus contributing to the decline in soil health and climate change. The CA practices consist of reduced tillage practices, legume crops in rotations and retention of crop residue cover which contribute to holistic and efficient soil and land management systems. Recycling of crop residues (by avoiding residue burning) can act as soil cover and can reduce the loss of water and soil while adding many essential nutrients into the soil (Fig. 3) (Srinivasarao et al., 2015). In drylands of the USA and Canada, where crop production is limited due to lack of N use and soil water availability, conservation tillage with long term diversified crop rotations involving legumes and cover crops has helped to increase water storage, decrease soil degradation and enhance dryland crop yields. In dry



Fig. 3 Conservation agriculture with a maize (*Zea mays* L.)–Horse gram (*Macrotyloma uniflorum* (Lam.) Verdc.) system.

Mediterranean climates, CA improved yields by 20–120% compared to traditional tillage systems because of improved plant nutrient availability and soil moisture (Piggin et al., 2011). In Ethiopia, CA enhanced crop yields by reducing soil erosion and runoff with substantially enhanced soil organic matter contents (Oicha et al., 2010). Mulching insulates the soil and buffers it from temperature fluctuations in dryland ecosystems. Straw mulching is widely used for many crops, not only to reduce evaporation from the soil, but also to increase rainwater infiltration. It increased soil water storage capacity in the top 30 cm of the soil by 17% and grain yield by 14–41.8% compared to the conventional method in Nigeria (Adeboye et al., 2017).

China, Japan, and South Korea are the greatest users of plastic film mulch; they account for 80% of its worldwide use. In China, plastic mulching has enhanced wheat and maize production by 33.2 and 33.7%, respectively (Liu et al., 2014). Cover crops in dryland agriculture improve the storage of nutrients and water, which enhances soil organic matter contents and they are also used as a surface mulch. Incorporation of legumes as cover crops in dryland agriculture increases the yield, nutrient availability and available water content. Further, long-term legume incorporation results in improvement of SOC, microbial biomass, and increases the soil fertility particularly in light-textured soils of drylands. Leguminous green manure crops grown prior to wheat in China enhanced the wheat yield (13%) and water-use efficiency (28%), and maintained the soil water balance better than fallow–wheat crop rotations (Zhang et al., 2016). In some Australian drylands, N fertilizer is not applied for wheat crops, but is substituted by N-fixing legumes which provide sufficient N to achieve good wheat yields. In countries like Zimbabwe (SE Africa), grain yield was greater for millet grown following legume cropping compared with continuous pearl millet cropping. Crop residue recycling is one of the key components of CA. Crop residues improve the soil health status and supply the necessary elements to the plants. In dryland areas, crop residue burning in rice, cotton, pigeon pea, castor, chilies, turmeric and other crop systems is still common, and large amounts of vegetable and fruit biomass are not recycled. About 25% of nitrogen (N) and phosphorus (P), 50% of sulfur (S), and 75% of potassium (K) uptake by cereal crops are retained in crop residues, making them valuable nutrient sources. Crop residues of rainfed crops such as cotton, castor, pigeon pea and maize are chopped and left on the soil surface to act as a mulch-cum-manure (Srinivasarao et al., 2020b). In India residue burning contributes 0.05% to the total GHG emissions, and it is also detrimental for soil microorganisms, causing a huge loss in biomass, organic C, and surface water and also leads to the formation of hydrophobic soils. Converting farm residues into biochar through pyrolysis instead of surface burning and the addition of biochar with fertilizer nutrients improved yields of several rainfed crops. Biochar added to the soil also mitigated the conditions from mid-season droughts through enhanced water retention (Srinivasarao et al., 2020b).

Integrated nutrient management

The potential of dryland soil to produce good yields is limited by the widespread deficiency of micro and secondary nutrients such as Zn (zinc), B (boron), and sulfur, in addition to deficiency of macronutrients and organic carbon. The productivity of various crops in dryland soils of the semi-arid tropics (India, East Africa, Australia, South America) is limited because of the nutrient poor soils. The yield of various crops in dryland regions can be profitability enhanced with better management of natural resources, the use of chemical fertilizers, and organic amendments. Incorporation of organic matter either in the form of crop residues, farm yard manure (FYM), compost or vermicompost together with chemical fertilizers has been shown to improve soil structure, water retention capacity, and water storage capacity of the soil profile (Ranjith et al., 2018). Combined application of inorganic and organic fertilizers reduced total N losses (41.2%) and P losses (33.3%), prevented soil erosion, and helped to increase the macronutrient pool in soil for sustaining crop productivity on a long term basis (Bouraima et al., 2015). Balanced fertilization showed an increase of 3 to 31% in the productivity of water in cotton and groundnut crops. There are ample opportunities for farmers and stakeholders to reduce GHG emissions that are resulting from agricultural production by reducing external chemical

inputs and increasing the usage of natural manures that ultimately help in improving crop productivity and the quality of agricultural produce in dryland farming systems (Srinivasarao et al., 2012).

Crop management solutions

Crop management solutions are an important strategy to achieve higher yields and to avoid production losses in dryland farming areas. Farmers need to adopt a range of innovative and frontier technologies to maintain the demand for food and other agricultural products from the world's growing population. Some crop management solutions are given below.

Intercrops

Intercrops ensure better weed control, reduce disease and pest incidence, and increase and stabilize overall productivity. It can be regarded as a biological insurance against risks under irregular or aberrant rainfall conditions. Intercrops are climate friendly, improve efficient use of the land and help to improve food security across the world. Productivity of cereal–legume intercroops was often reported to be higher and more sustainable in rainfed dryland systems than cereal monocropping (Fig. 4). Intercropping systems are a common feature of small-scale farming throughout the tropics and subtropics of Asia, Africa and Latin America, and are believed to contribute up to 15–20% of global food production (Lithourgidis et al., 2011). However, they are not widely used in large-scale agriculture in developed countries. Inter cropping systems in Southern Australian wheat belt regions indicated that among the intercrop groups studied, 'Peaola' (canola–field pea intercroops in which 70% are intercroops) showed a 50% increase in productivity, and 'cereal–grain legume intercroops' showed a 64% increase in productivity over monoculture (Fletcher et al., 2016).

Foliar sprays

Under dryland conditions, lack of soil moisture hampers the absorption of fertilizers that are applied to the soil; foliar application of plant nutrients, however, overcomes these limitations. Foliar spray has a negligible effect on various soil factors such as soil pH, fixation of fertilizers to clay, availability of soil moisture, and on the rhizosphere (soil which attached around the roots). Furthermore, very small quantities of fertilizers are required for foliar sprays which reduces inputs. Foliar application of concentrated urea solution enhanced dryland wheat yield by 31% and cotton by 13–51%. In Pakistan, foliar applied nitrogen also resulted in greater nitrogen use efficiency, net return value: cost ratio and relative increase in income compared to fertilizers applied to the soil. In drylands, foliar application of nutrients may increase the yield and yield components by decreasing water and nutrient stress, and by increasing water and nutrients use efficiencies. Some precautions are required when applying foliar sprays because a high dose fertilizer may cause burning of plant leaves, foliar fertilizers have to be compatible with other agrochemicals like herbicides, and water use has to be optimized.

Tolerant crops

Tolerance to dry conditions is the most important and essential quality of dryland crops. Most of the plants vary in their ability to tolerate aridity, and yields under arid conditions are based on their degree of tolerance. Developing varieties that are tolerant to



Fig. 4 Maize (*Zea mays* L.) + Cowpea (*Vigna unguiculata* L. Walp) intercropping system. (Cereal+legume).

Table 3 Degree of Tolerance of different crops in Global drought conditions (Creswell and Martin, 1998).

<i>Crop</i>	<i>Degree Of Tolerance^a</i>
Corn (<i>Zea mays</i> L.)	1.0
Sorghum (<i>Sorghum bicolor</i> (L.) Moench)	1.5
Pearl Millet (<i>Pennisetum americanum</i> (L.) Leeke)	2.5
Common Bean (<i>Phaseolus vulgaris</i> L.)	1.0
Cowpea (<i>Vigna unguiculata</i> L.)	1.5
Pigeon Pea (<i>Cajanus cajan</i> (L.) R. Wilczek)	2.0
Mung Bean (<i>Vigna radiate</i> L.)	2.0
Cassava (<i>Manihot esculenta</i> Crantz)	2.0
White Yam (<i>Dioscorea rotundata</i> Poir.)	1.0
Mixta Squash (<i>Cucurbita argyrosperma</i>) K. Koch	1.5
Okra (<i>Abelmoschus esculentus</i> (L.) Moench)	1.5
Pomegranate (<i>Punica granatum</i> L.)	2.0
Cashew (<i>Anacardium occidentale</i> L.)	2.5
Date (<i>Phoenix dactylifera</i> L.)	3

^aRated from 0 (no tolerance) to 3 (high tolerance).

multiple abiotic and biotic stresses with stress-tolerant Quantitative Trait Loci (QTLs), genes and alleles in elite cultivars is an efficient way of achieving climate resilience and with easy access to farmers (Pathak et al., 2018). In rainfed areas, farmers should focus on crop diversification such as climate-smart, hardy crops (millets); and include more high-value fruit crops such as dragon fruit (*Selenicereus undatus* (Haworth) D. R. Hunt) and pomegranate (*Punica granatum* L.) in drought prone areas (Table 3).

Agroforestry

In dryland areas, a change from cultivation with no trees to agroforestry can improve the soil organic carbon stocks, which can help to mitigate climate change. Soil organic matter content increases when trees are included in agroforestry through the addition of litter both above and below ground, which increases soil microbial dynamics compared to monocropping systems in rainfed regions. Agroforestry systems are widely practiced in drylands worldwide. These practices are predominant in the drylands of Eastern and Central African countries like Tanzania, Uganda, Ethiopia, Kenya, northern Africa, sub-Saharan Africa and Asia. Agroforestry practices such as alley cropping of barley and forage crops under pistachio (*Pistacia vera* L.), olive (*Olea europaea* L.) and vineyards are practiced in Turkey (Gundogan et al., 2010). Agroforestry technologies in Israel are based on an improved version of an age-old tradition of redirecting floodwaters to plots surrounded by dikes or hand-dug pits in which trees or shrubs are planted. Improvement includes the incorporation of nitrogen fixing trees and food crops which contribute to both food security and ecosystem services (Israel 21C, 2016). This multifunctional land-use strategy needs increased attention from agricultural communities and policy makers because agroforestry is one viable way to sustain soil health.

Integrated farming system

In view of the decline in per capita availability of land, it is imperative to develop strategies and agricultural technologies that enable adequate employment and income generation, especially for small-holders (farmers with <2.0 ha of land) who constitute the vast majority of the farming community in dryland farming systems. Adoption of integrated farming systems that involve different components such as crops, livestock, poultry, fisheries, etc., is a powerful tool and holds the key to ensuring income, employment, livelihood and nutritional security for small and marginal farmers in dryland regions. Dryland farmers need to adopt resource conserving technologies such as crop diversification and recycling of farm waste. Improvement in productivity, fixed income and net growth in profit, balanced diet, recycling of farm residues and sustainable growth in agriculture are the main benefits of an integrated farming system. In a well-managed integrated farming system, nutrients in manure are returned to the soil where they are used to produce feed crops for livestock production. More than 90% of the total population of large and small ruminants are kept on mixed farms in South Asia. About 55% of China's population farm using integrated crop–livestock systems, which utilizes 83% of the total cropland, produces 90% of beef and mutton, and 50% of pork and poultry in China (Hou et al., 2008).

Dryland farming in India-case study

In India, around 51% of the total cultivated area is under dryland farming. It contributes about 42% of the total food production and plays a crucial role in safeguarding the country's food security. Dryland farming systems are complexed with a variety of cropping systems, agroforestry and livestock production. Almost 80% of cereals, 50% of maize, 88% of soybean, and 65% of pigeon peas and chickpea, among other crops, are produced by dryland farming in India. Overall, dryland farming in India produces 40%

of food and supports two thirds of the livestock (Srinivasarao et al., 2015). Soils of dryland ecosystems are generally poor in soil organic carbon (SOC) stocks because of high temperatures and low rainfall. The critical yield gaps for the major crops of dryland farming in India ranged from 0.1 to 1.2 Mg ha⁻¹ (Srinivasarao et al., 2013). The main soil types (Soil Taxonomy and World Reference Base classifications) are Inceptisols, Cambisols (95.8 M ha), Entisols, Umbrisols (80.1 M ha), Alfisols, Luvisols (79.7 M ha) Vertisols (26.3 M ha), Aridisols, Calcisols, Solonchaks (14.6 M ha), Oxisols, Ferrasols (0.3 M ha) and miscellaneous soil types (23.1 M ha). Alluvial soils occupy the largest area and the problems associated with dryland agriculture are not so acute in these soils. Problems are, however, encountered frequently in Vertisols and Alfisols, which are largely distributed in the north western, central and southern parts of India. Rajasthan, Maharashtra, Karnataka, Madhya Pradesh, Gujarat, Uttar Pradesh and Tamil Nadu are the principal states that produce most of the dryland crops. The average yields are quite small for cereals (coarse grains) at around 1954 kg ha⁻¹, and oilseeds and pulses ~1265 kg ha⁻¹ and ~ 806 kg ha⁻¹, respectively. The national annual average rainfall in India is 1200 mm, yet regional differences can be as high as 10,000 mm per year in the northeast and as low as 150 mm per year in the desert regions. In India the total micro-irrigation potential is around 69 M ha; drip irrigation systems provide 90% efficacy, sprinkler 70–80% and conventional surface irrigation about 60–70% (<https://www.niti.gov.in/>). The Government of India is supporting this technology at the large scale with several incentives (Srinivasarao et al., 2014) (Table 4).

In Indian agriculture, the role of institutions becomes more crucial as there are structural- and enterprise-specific constraints like high transaction costs, lack of market integration, and interlocking of output markets which only institutions and organizations can

Table 4 Different technologies used to improve water use efficiency of rainfed farming in India.

Recommended domain	Existing practice	Improved technology	Performance
Royalaseema Region, A.P. India Telangana	Flood irrigation in groundnut solo cropping Excess rainfall runs off and farmers are unable to collect it for use during dry spells.	Supplemental irrigation from farm ponds through sprinklers Farm ponds are constructed at the lower side of the fields with a capacity to store 100–200 m ³ of runoff per hectare.	Yield of groundnut increased by 25% Tomato as a test crop showed a payback period of 10 years with benefit:cost (B:C) ratio 1.5 and internal rate of return (IRR) 18.9%. Method is more viable and economical in <i>Vertisols</i>
Northern Dry zone of Karnataka	No in situ moisture conservation practices during <i>kharif</i> (monsoon) season resulting in soil erosion and low yield of <i>rabi</i> (post monsoon) crops	Making square compartments (0.15 m height) on the field to retain rainwater and stop soil erosion.	Increase in yield of <i>rabi</i> (post monsoon) sorghum, sunflower, safflower and chickpea with 40, 35, 38 and 50%, respectively. It also significantly controls runoff.
Northern Dry zone of Karnataka	No in situ moisture conservation practices in sodic medium and deep black soils resulting in soil erosion and poor yields	Gravel and sand mulching in sodic <i>vertisols</i> during summer	Practice gives yield advantage of 125% in groundnut, 214% in <i>rabi</i> (post monsoon) sorghum, 300% in sunflower, 300% in chickpea and 366% in greengram. Cost of sand application can be recovered within 2 years of cropping
Northern Dry zone of Karnataka	Fallow crop during <i>kharif</i> (monsoon) and cultivation of sorghum, sunflower and chickpea during <i>rabi</i> (post monsoon) resulted in splash erosion and high runoff in <i>kharif</i> (monsoon)	Cover cropping with quick growing species like sunhemp, greengram, cucumber and ridge gourd in <i>kharif</i> (monsoon) to reduce runoff and splash erosion.	With cover cropping in <i>kharif</i> (monsoon), yield advantage in various <i>rabi</i> (post monsoon) crops varies from 43% to 300% with a B:C ratio up to 5.47.
Northern Dry zone of Karnataka	Farmers make bunds up to a height of 0.45 m without any provision for a runoff outlet causing soil erosion due to upstream pressure.	Created inlet spillway arrangement for draining water to the next plot downstream.	Allowing rainfall infiltration into the soil profile making it possible for two crops even in drought years.
Northern Dry zone of Karnataka	Farmers depend on check dam or nala bunds for recharging the dug wells when they go dry.	Surface water harvesting and diverting into a defunct well through filter bed having different layers of sand, boulders, charcoal etc.	Normal size wells (5*5*10 m) can store about 0.4–0.5 million liters of water per year.
Malwa region of Madhya Pradesh	Soybean cultivated on flat bed system as a rainfed <i>kharif</i> (monsoon) crop resulting in either moisture stress or they are prone to water logging in black soils eventually decreasing yields.	Broad bed furrow system (BBF) of 105 cm width are made, soybean seed rate reduced by 25%.	BBF enhanced seed yield by 36% with B:C ratio of 2.53. BBF system helps draining of excess water and improves soil physical condition in black soil.
Scarce rainfall zone of Maharashtra	Fallow land during <i>kharif</i> (monsoon) and practice monocropping of sorghum during <i>rabi</i> (post monsoon) with one or two harrowings for weed control.	Ridges (20-cm height) and furrows (45-cm width) are made at the onset of monsoon to avoid soil erosion and increase soil water storage potential.	Conserves 45% more moisture, retains it for a longer period and increases <i>rabi</i> (post monsoon) sorghum yield by 53% with B:C ratio of 1.76.

tackle effectively. Institutions help small farmers by reducing transaction costs, managing or reducing risk, building social capital, enabling collective action, and or readdressing missing markets. Many institutional innovations are emerging that are successfully bringing farmers together to aggregate their small marketable surpluses. Village Climate Risk Management Committee (VCRMC), Agro-Advisory Services (AAS), Farmers' Producers Organization (FPO), Custom Hiring Centers for Farm Implements and Agricultural Markets etc. are major innovations which help farmers to collaborate. The Indian Council of Agriculture Research (ICAR) has developed more than 650 contingency plans for all rural districts of the country towards drought preparedness, and real time responses towards overall stability of rainfed dryland systems in the country. It covers field crops, horticulture, livestock, fishery and poultry resilient technology options to implement with location specificity and with the Union and State Governments together (Srinivasarao et al., 2020a). These successful experiences are shared with other SAARC countries and have been presented at the United Nations Framework Convention on Climate Change platforms towards climate change adaptation and resilient food systems.

Conclusion and way forward

Developing and implementing moisture conservation technologies are imperative for drylands to deal with national food security and climate change. There is a range of innovative solutions available to help farmers cope with climate variabilities. Some of these technologies proficiently increase the WUE and crop productivity; such technologies need more promotion through appropriate policy options. In general, farmers of dryland area depended on the narrower range of solutions for their livelihoods, thus they are more vulnerable to climate effects. If the most feasible and cost-effective technologies can be brought into their fields, it would contribute significantly to the overall development of dryland agriculture. Efficient rainwater management and its utilization, reducing soil erosion and managing soil health with sustainable soil management practices, legume intercrops and resource conservation practices play pivotal roles in dryland crop production. Governing bodies should focus on the vulnerable sectors of drylands to find out the most appropriate technologies and develop effective policies in order to bring sustainability in the dryland ecosystem. In addition, the promotion of farmer producing organizations (FPOs), village markets, value addition and processing, storage structures, market linkages, agri startup are potential systems for the sustainability of dryland ecosystems and food systems across the world.

References

- Adeboye OB, Schultz B, Adekalu KO, and Prasad K (2017) Soil water storage, yield, water productivity and transpiration efficiency of soybeans (*Glycine max* L. Merr) as affected by soil surface management in Ile-Ife, Nigeria. *International Soil and Water Conservation Research* 5(2): 141–150. <https://doi.org/10.1016/j.iswcr.2017.04.006>.
- Ariyabandu R (2008) *Swings and Roundabouts: A Narrative on Water Policy Development in Sri Lanka. Working Paper 296*. Overseas Development Institute. 111 Westminster Bridge Road, London SE1 7JD.
- Bhattacharyya R, Ghosh BN, Mishra PK, et al. (2015) Soil degradation in India: Challenges and potential solutions. *Sustainability* 7: 3528–3570.
- Bouraima AK, He B, and Tian T (2015) Runoff, nitrogen (N) and phosphorus (P) losses from purple slope cropland soil under rating fertilization in Three Gorges Region. *Environmental Science and Pollution Research* 23: 4541–4550.
- Creswell R and Martin FW (1998) *Dryland Farming: Crops and Techniques for Arid Regions*. ECHO technical note, Fort Myers, FL, USA.
- De Pauw EF (2009) Management of dryland and desert areas. In: Verheye WH (ed.) *Land Use, Land Cover and Soil Sciences*. Belgium: National Science Foundation Flanders-Belgium and Geography Department, University of Gent.
- Fischer G, Van Velthuizen H, Hizsnyik E, and Wiberg D (2009) *Potentially Obtainable Yields in the Semi-Arid Tropics. Global Theme on Agroecosystems Report no. 54 Patancheru 502 324, Andhra Pradesh, India*. International Crops Research Institute for the Semi-Arid Tropics 68.
- Fletcher AL, Kirkegaard JA, Peoples MB, Robertson MJ, Whish J, and Swan AD (2016) Prospects to utilise intercrops and crop variety mixtures in mechanised, rain-fed, temperate cropping systems. *Crop & Pasture Science* 67(12): 1252. <https://doi.org/10.1071/cp16211>.
- Gundogan R, Merdun H, Demirkiran A, Hall N, Agturk R, Erol A, and Ersahin S (2010) Temporal variation of soil moisture under alley cropping system in Pistachio in semi-arid region of Turkey. *International Journal of Agriculture and Biology* 12(4): 601–605.
- Hori Y, Stuhlberger C, and Simonett O (2011) Desertification – A visual synthesis. In: *United Nations Convention to Combat Desertification (UNCCD)*, Zoi Environment Network (Zoi) <http://www.unccd.int/Lists/SiteDocumentLibrary/Publications/Desertification-EN.pdf>.
- Hou F, Nan Z, Xie Y, Li X, Lin H, and Ren J (2008) Integrated crop-livestock production systems in China. *The Rangeland Journal* 30(2): 221–231.
- Huang J, Li Y, Fu C, Chen F, Fu Q, Dai A, Shinoda M, Ma Z, Guo W, Li Z, Zhang L, Liu Y, Yu H, He Y, Xie Y, Guan X, Ji M, Lin L, Wang S, Yan H, and Wang G (2017) Dryland climate change: Recent progress and challenges. *Reviews of Geophysics* 55(3): 719–778.
- Ibrahim YZ, Balzter H, Kaduk J, and Tucker CJ (2015) Land degradation assessment using residual trend analysis of GIMMS NDVI3g, soil moisture and rainfall in Sub-Saharan West Africa from 1982 to 2012. *Remote Sensing* 7(5): 5471–5494.
- Jacobsen SE, Jensen CR, and Liu F (2012) Improving crop production in the arid Mediterranean climate. *Field Crops Research* 128: 34–47.
- Kumar A and Sharma P (2022) *Impact of climate variation on agricultural productivity and food security in rural India. Economics Discussion Papers, No 2013-43*. Kiel Institute for the World Economy. Available at SSRN 4144089 <http://www.economics-ejournal.org/economics/discussionpapers/2013-43>.
- Kumar KM, Sridhara CJ, Hanumanthappa M, and Marimuthu S (2019) A review of impacts and mitigation strategies of climate change on dryland agriculture. *Current Journal of Applied Science and Technology* 33: 1–12.
- Lithourgidis AS, Dordas CA, Damalas CA, and Vlachostergios DN (2011) Annual intercrops: An alternative pathway for sustainable agriculture. *Australian Journal of Crop Science* 5: 396–410.
- Liu E, He W, and Yan C (2014) 'White revolution' to 'white pollution'—agricultural plastic film mulch in China. *Environmental Research Letters* 9: 091001.
- Nguyen-Van-Hung B, C., Sandro, J., et al. (2022) Precision land leveling for sustainable rice production: Case studies in Cambodia, Thailand, Philippines, Vietnam, and India. In: *Precision Agriculture*, 1–20. <https://doi.org/10.1007/s11119-022-09900-8>.
- Oicha T, Cornelisc WM, Verplancke H, Nyssen J, Govaerts B, Behailu M, Haile M, and Deckers J (2010) Short-term effects of conservation agriculture on Vertisols under tef (*Eragrostis tef* (Zucc.) Trotter) in the northern Ethiopian highlands. *Soil and Tillage Research* 106: 294–302.

- Osman KT (2018) Dryland soils. In: *Management of Soil Problems*. Cham: Springer https://doi.org/10.1007/978-3-319-75527-4_2.
- Pathak H, Nayak AK, Jena M, et al. (2018) *Rice Research for Enhancing Productivity, Profitability and Climate Resilience*. Cuttack, Odisha: ICAR-National Rice Research Institute 527.
- Piggin C, Haddad A, and Khalil Y (2011) Development and promotion of zero-tillage in Iraq and Syria. In: *Proceedings of the 5th world congress on conservation agriculture, Brisbane, 26–29 September* 304–305.
- Ranjith KG, Konde NM, Tamgadge DB, et al. (2018) Effect of STCR equation on soil characters and nutrient availability in cotton under Rainfed conditions in Vertisols. *International Journal of Chemical Studies* 6(3): 723–726.
- Sander BO, Samson M, and Buresh R (2014) Methane and nitrous oxide emissions from flooded rice fields as affected by water and straw management between rice crops. *Geoderma* 235–236. <https://doi.org/10.1016/j.geoderma.2014.07.020>.
- Saxena C and Ramana Rao KV (2019) *Micro-Irrigation for Higher Water Productivity in Horticultural Crops*. Peter, KV 179–199.
- Srinivasarao Ch, Ravindra CG, Venkateswarlu B, et al. (2009) *Carbon Sequestration Strategies under rainfed production Systems of India*. Hyderabad, Andhra Pradesh, India: Central Research Institute for Dryland Agriculture.
- Srinivasarao Ch, Venkateswarlu B, Dixit S, et al. (2012) Water Productivity improvement through participatory soil fertility management in rainfed districts of Andhra Pradesh. *Indian Journal of Dryland Agricultural Research and Development* 27: 23–32.
- Srinivasarao Ch, Venkateswarlu B, Lal R, et al. (2013) Sustainable management of soils of dryland ecosystems of India for enhancing agronomic productivity and sequestering carbon. *Advances in Agronomy* 121: 254–329.
- Srinivasarao Ch, Ravindra Chary G, Mishra PK, et al. (2014) *Rainfed Farming – A Compendium of Doable technologies*. Hyderabad: All India Coordinated research Project for Dryland Agriculture, Indian Council of Agricultural Research 194.
- Srinivasarao Ch, Lal R, Prasad JVNS, et al. (2015) Potential and challenges of rainfed farming in India. *Advances in Agronomy* 133: 113–181.
- Srinivasarao Ch, Rao KV, Gopinath KA, et al. (2020a) Agriculture contingency plans for managing weather aberrations and extreme climatic events: development, implementation and impacts in India. *Advances in Agronomy* 159: 35–91.
- Srinivasarao Ch, Subha Lakshmi C, Kundu S, et al. (2020b) Integrated nutrient management strategies for rainfed agro-ecosystems of India. *Indian Journal of Fertilizers* 16(4): 344–361.
- Srinivasarao Ch, Rakesh S, Ranjith Kumar G, et al. (2021) Soil degradation Challenges for sustainable agriculture in tropical India. *Current Science* 120(3): 492–500.
- Stavi I and Lal R (2015) Achieving zero net land degradation: challenges and opportunities. *Journal of Arid Environments* 112: 44–51.
- Stewart BA (2016) Dryland Farming, Reference Module in Food. *Science*. <https://doi.org/10.1016/B978-0-08-100596-5.02937-1>.
- Swiss Re Institute (2021) *Swiss Re Sustainability Report*. Zurich, Switzerland: Swiss Re Ltd. 125.
- Tibirkov AP, Tibirkova NN, Filin VI, and Kuzin AG (2021) The efficiency of applying soil hydrogel to the light-chestnut soils of the Volga and Don inter stream area to stimulate seed sprouting for agricultural crops. *IOP Conference Series: Earth and Environmental Science* 659(012018): 2020.
- Zhang D, Yao P, Na Z, Cao W, Zhang S, Li Y, and Gao Y (2016) Soil water balance and water use efficiency of dryland wheat in different precipitation years in response to green manure approach. *Scientific Reports* 6(1): 1–12.

Further reading

- Biradar CM, Thenkabail PS, Noojipady P, Li Y, Dheeravath V, Turral H, Velpuri M, Gumma MK, Gangalakunta ORP, Cai XL, and Xiao X (2009) A global map of rainfed cropland areas (GMRCA) at the end of last millennium using remote sensing. *International Journal of Applied Earth Observation and Geoinformation* 11(2): 114–129.
- FAO (2005) *The Importance of Soil Organic Matter: Key to Drought-Resistant Soil and Sustained Food and Production*. Rome: Food and Agriculture Organization of the United Nations 2005.
- FAO (2019) *Trees, forests and land use in drylands: The first global assessment – Full report*. FAO Forestry Paper No. 184. Rome.
- ICARDA (2020) Enhancing Dryland Livelihoods Helps Tackle Climate-Change Induced Migration. Available at: <http://www.icarda.org/media/blog/enhancing-dryland-livelihoods-helps-tackle-climate-change-induced-migration>.
- Israel 21C (2016) 10 Top Ways Israel Fights Desertification. Available at: <http://www.israel21c.org/top-10-ways-israel-fights-desertification/>.
- IUSS Working Group WRB (2022) *World Reference Base for Soil Resources. International soil classification system for naming soils and creating legends for soil maps*, 4th edn. Vienna, Austria: International Union of Soil Sciences (IUSS).
- Shirley L, Lingling L, Xie J, and Jeffrey A (2020) Coulter, Tillage system affects soil water and photosynthesis of plastic-mulched maize on the semiarid Loess Plateau of China. *Soil and Tillage Research* 196. <https://doi.org/10.1016/j.still.2019.104479>.
- Soil Survey Staff (2022) *Keys to Soil Taxonomy*, 13th edn USDA-Natural Resources Conservation Service.
- World Bank (2017) Confronting Drought in Africa's Drylands: Opportunities for Enhancing Resilience. Available at: <https://blogs.worldbank.org/voices/africa-s-drylands-opportunities-cut-vulnerability-drought-and-famine-are-within-reach>.

The effect of crop rotations on soil

Caio Fernandes Zani^a, Arlete Simões Barneze^b, Rogério Peres Soratto^c, and Charles Andrew Francis^d, ^aSchool of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom; ^bSoil Biology Group, Wageningen University & Research, Wageningen, The Netherlands; ^cCollege of Agricultural Science, São Paulo State University (UNESP), São Paulo, Brazil; ^dKiesselbach Crops Research Laboratory, University of Nebraska–Lincoln, Lincoln, NE, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of C.A. Francis, CROP ROTATIONS, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 318-322, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00253-8>.

Historic overview of crop rotations	125
Development and current crop rotation systems worldwide	126
Potential benefits and drawbacks of crop rotations	128
Soil carbon sequestration	128
Fertility and yield	129
Pest control	131
Biodiversity	132
Social and economic aspects	132
Final remarks	133
References	133

Abstract

Crop rotation has been implemented since the early days of agriculture's history; however, it has changed over the years with systems being neglected in some regions, but also improved in others. In this chapter, the focus is on the history and current characteristics of crop rotation, what has been done to improve it worldwide, and how different climate conditions can affect it. In addition, the many benefits of crop rotation for different aspects of production systems such as yields, soil carbon sequestration, and pest control, are considered including some social-economic effects in various countries and regions.

Key points

- Crop rotation can underpin a sustainable production system.
- It can increase soil fertility and also carbon sequestration.
- Microclimate can affect the success of crop rotation.
- Pests and diseases can be suppressed by crop rotation.

Historic overview of crop rotations

Conceptually, crop rotation consists of the temporal alternation of different crop species cultivated in the same field (or group of fields) in a recurrent succession. This contrasts with the prolonged planting of the same crop types across years/seasons, i.e. the so-called monocropping or continuous monoculture cultivation practice. Crop rotation may also involve leaving the land fallow for a given season (or seasons), which might be used as pasture or just left bare.

Crop rotation has been used since the early days of agriculture's history, dating back to the predynastic Egyptian period (about 8500 BCE) and coming into use either by accident or by design (Britannia, 2022). In the past, farmers implemented crop rotation primarily based on their seasonal calendar, which gave them the traditional basis for their planting pattern. Over the years, the idea of crop rotation was developed further and consolidated primarily based on observed positive results in crop yields. In short, this occurs mainly because each crop releases or absorbs different nutrients, which helps to balance out eventual deficits and or surplus thereof in the soil.

Other historical reports, including biblical writings, also suggest that crop rotation was successfully used by the ancient Jews, Egyptians, Romans and Greeks (6000 BCE). There is written evidence of the recognition of the importance of crop rotations and also of putting specific crops on unique soil types during Roman times. Although Cato the Censor (second century B.C.) dedicated most of his writing on farming to the cultivation of grapes (*Vitis vinifera* L.) and olives (*Olea europaea* L.), he also discussed the importance of applying manure and of fallowing land before planting annual or perennial crops. His writings describe the importance of vetch (*Vicia sativa* L.), beans (several species), and lupins (*Lupinus* spp.) in helping to improve the soil or fertilize the land, and that cereal crops following these legumes benefited and produced higher yields. Varro and Virgil (first century BC) recognized the importance

of alfalfa (*Medicago sativa* L.), a deep-rooted perennial legume, in providing improved fertility for the crops that followed, and they recommended rotations, fallow periods, and specifically a bean (*Phaseolus vulgaris* L.)—spelt wheat (*Triticum spelta* L.) rotation. Pliny (first century AD) wrote extensively on farming, including crop rotations in the context of Italy and the Mediterranean region (Francis, 2005). Likewise, there are written records of the early effects of crop rotations and advice against the continuous planting of cereals in the literature from the Islamic era (8th–11th centuries AD) in the Iberian Peninsula (Bolens, 1994).

The benefit of crop rotation on crop yields has been reported by many researchers and farmers over the years. Chinese farmers have had success for millennia with the implementation of the well-known three-field system. Crop rotation was also a popular practice during the medieval and early-modern European periods. In England, the implementation by Sir Charles Townshend of the so-called 'Norfolk Rotation', with four different crops, usually barley (*Hordeum vulgare* L.)-clover (*Trifolium* spp.)-ryegrass (*Lolium* spp.)-wheat (*Triticum aestivum* L.)-turnips (*Brassica rapa* L.), helped the country's agricultural production to almost triple in the 18th century. Even before such periods, there is some evidence suggesting that in prehistoric Europe (probably during the Old Stone Age) a sequence of different annual crops and pastures was grown in rotation after land-clearing, but often followed by regrowth of the natural vegetation. This practice was also similar to the more familiar slash-and-burn or swidden agriculture, which is still practiced in some parts of the world, particularly in the tropics.

In the Middle Ages (5th to the late 15th centuries) before the Renaissance, there was apparently a reversion to crop-fallow sequences. There are limited reports of perennial pasture-grass plantings in rotation with cultivated cereal crops. The use of animal manure was prevalent. Perhaps the effects of manure applications were more apparent in a given year, if there were adjacent areas with and without manure, than the more subtle differences that were the result of rotation, the effects of which were not easily observed. In the seventeenth and eighteenth centuries, more organized legume-cereal rotations began to appear in reports from England. They often consisted of 2–4 years of vegetable or fodder crops, 1 year in legumes, and 1 or 2 years in cereals. Mixed farming with crops and livestock was prevalent, and the pastures and animal manure undoubtedly contributed to cereal yields. These rotation systems were taken to North America by early settlers from northern Europe.

Cropping diversity in many early systems included multiple species as well as crop rotations. In Europe, the small grains were often mixed cultures of cereal species that were used for grain and for fodder. The well-known maize (*Zea mays* L.)-beans (*Phaseolus* spp.)-squash (*Cucurbita* spp.) systems of Central America and Mexico as well as the intercrop and relay crop systems of maize-beans-potato (*Solanum tuberosum* L.) in the Andean Zone are examples of diverse mixtures that evolved over centuries as people selected these New World crop species (Francis, 2005).

With an emerging understanding of soil chemistry and the responses of plants to applied manures and other soil amendments, crop rotations in the first half of the twentieth century focused again on cereal-legume sequences (Francis, 2005). Such a simplistic rotation was almost entirely phased out, in particular in developed countries, after World War II when large supplies of synthetic nitrogen (N) and other multiple-nutrient fertilizers became the new 'normal'. The specialization in field crops, especially cereals, and disappearance of livestock from most farms further pushed the systems into continuous cropping or a very simplified 2-year cereal-legume rotation. Such changes confirmed that even though farmers were implementing crop rotation for many years, they did not know about its impact on the soils and in the environment. Only in the past two decades, with the growing interest in organic agriculture and more integrated systems, followed by the emerging concerns around the sustainability of agriculture in the mid-1980s, longer and more complex rotations have attracted public attention around the world again. This is particularly so in wealthy countries such as the United States and most of the developed European countries but often forgotten by developing countries. This is, however, particularly important for the latter group, especially those located in the tropical regions of the globe, where soil and climate conditions allow a more intensify farming system, and therefore the implementation of conservationist practices, such as crop rotation, is essential.

Development and current crop rotation systems worldwide

The current crop rotation systems vary considerably depending on climate conditions (as rainfall has an effect on crop production), production systems, and social aspects of the region. Cereal-legume rotations are among the most frequently used systems, dating back to before Roman times (Martinez, 2005), and they still provide the foundation for the design of most crop sequences today. Rotation among crops from different families is one of the most frequently used approaches today in the temperate zone (Francis, 2005). In areas of adequate rainfall, above 600 mm year⁻¹, a maize-soybean [*Glycine max* (L.) Merr.] rotation is prevalent in the Corn Belt of the United States Midwest (Sahajpal et al., 2014). In areas with lower rainfall, irrigation is used to assure a crop, and rotations are less common in these systems. A summary of the general types of crop rotations is presented in Fig. 1, showing the array of different levels of genetic diversity that have been exploited in farmers' systems across the centuries.

Rotating different varieties of the same cereal, rotating different cereals, and rotating cereals with legumes are all strategies to provide some of the soil fertility and plant protection benefits that come from diverse cropping sequences. More diverse are the sequences of summer-winter crops and rotations of pastures with crops, but most diverse of all are the annual-perennial sequences and combinations such as alley cropping. Alley cropping is characterized by planting rows of trees or shrubs to create alleys within which agricultural or horticultural crops are produced. The trees or shrubs provide additional valuable products as fruits, nuts, wood, or fiber (Soratto et al., 2022). Integrated crop-livestock systems are also a form of sustainable intensification of agriculture that relies on synergistic relationships between elements of plant and animal systems to bolster critical agroecosystem processes, with potential impacts on nutrient cycling, soil health, and resilience to weather anomalies (Reddy, 2016; Peterson et al., 2020). In these systems, annual crops are rotated with perennial pastures for animal grazing, and they are especially relevant for tropical

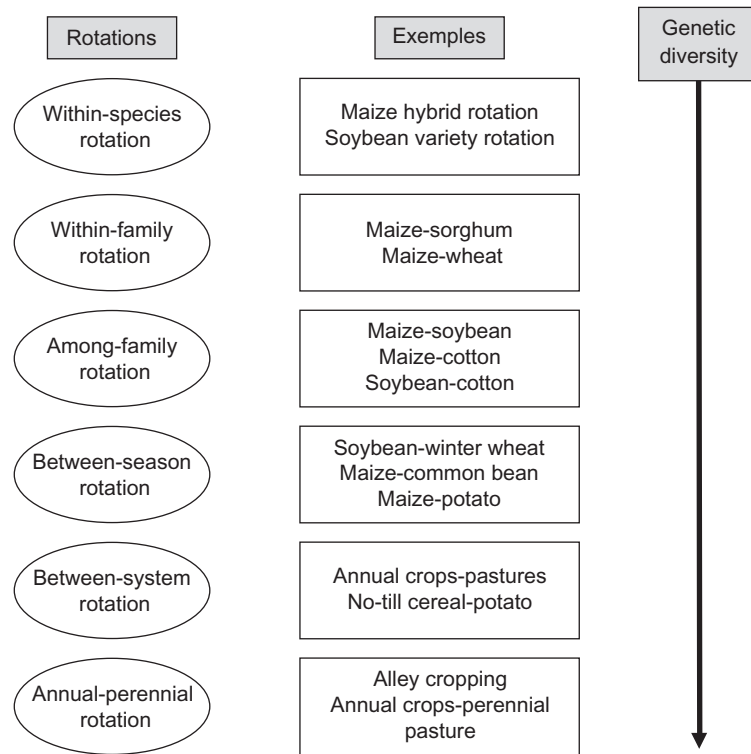


Fig. 1 Different types of crop rotation, crop-pasture rotation, and long-term perennial alternatives, showing the increasing genetic diversity of specific systems (from top to bottom). Adapted from Francis CA (2005) Crop rotations. In: *Encyclopedia of Soils in the Environment*, pp. 318–322. Elsevier. doi:10.1016/b0-12-348530-4/00253-8.

regions (Salton et al., 2014). While those crop rotation models are gaining popularity once again, current agricultural systems are still predominantly monoculture-driven, which often depend on aspects such as synthetic inputs, mechanization, irrigation, and fossil fuels. Such a model, nonetheless, has proved increasingly unsustainable, with several studies indicating rates of soil loss that are greater than its rate of formation, water pollution and its scarcity, also environmental damage, loss of biodiversity, and effects on human health (Francis, 2005; Salton et al., 2014).

The use of grassland in crop rotation has been shown to have a positive impact on production. However, it might depend on the type of grass used. Grasses have different resource uptake strategies known as acquisitive and conservative strategies. Acquisitive grasses have specific traits related to the more fast-growing species with quicker nutrient uptake and growth, while conservative grasses have a slower growth rate and lower rate of respiration (Reich, 2014). The implementation of these contrasting species in the rotation might also affect the productivity and sustainability of production because it affects the rate of nutrient uptake.

There are huge differences between sustainable production systems in the semi-arid tropics and semi-arid temperate regions, not only influenced by rainfall events, but also by the production system per se. For example, crop production in countries in Africa and Asia are in village units. This means that if the production system is not sustainable at some point, people make necessary changes in land use or even migrate to another region to start the system over again. On the other hand, production systems in Australia for example are characterized by being family organized so that if sustainability or profits are not achieved, the cropping practices change and possibly also ownership of the land.

In some parts of the world, crop rotation is often used in association with direct drilling. This is when crops are sown directly into the stubble without cultivation or preparing the soil after the previous crop and all crop residue is left on the soil surface. In Brazil, the direct drilling system has been increasing rapidly, especially in recent years. It is estimated that today the system with rotation covers more than 33 million hectares, that is, about 61% of the area with annual crops in the country (Fuentes-Llanillo et al., 2021). Farmers are advised to leave crop residues in the field to cover the soil properly and maintain or increase soil organic matter: a minimum of 6–7 Mg ha⁻¹ year⁻¹ for the subtropical high-altitude climate, 7.5–8.0 Mg ha⁻¹ year⁻¹ for the subtropical low altitude climate and transition between subtropical and tropical climates, and 10–12 Mg ha⁻¹ year⁻¹ for tropical climates (Canalli et al., 2019). One of many benefits of direct drill associated with crop rotation is the effect on soil temperature, because the crop residues help to maintain the ambient temperature which helps to avoid crop damage. For example, the crop rotation system in Brazil varies considerably because of the vast range of climate conditions over the different regions (Canalli et al., 2019). The use of crop rotation systems in the South Brazil is more favored because it is possible to produce summer and winter crops due to the ample and even rainfall distribution over the year, whereas in Central Brazil (with an almost total absence of rain between May-August) it is difficult to grow winter crops (as it requires a suitable microclimate or irrigated agriculture). The main crops grown in Central Brazil are those that increase the input of crop residues to the soil surface (Canalli et al., 2019). This requires that they have certain characteristics,

such as rapid development and increased drought tolerance (Salton et al., 2014; Possamai et al., 2022). Although crop rotation is an important system for many aspects including sustainability, the microclimate and climate conditions are a barrier in many regions to the success in implementing this system worldwide.

It is difficult to suggest the best system for the design a crop rotation system because as mentioned above several factors might have a great impact on the production. It is important to choose species and cultivars with high productivity and large C/N ratio which delays decomposition and provides a proper soil cover and/or initial straw supply. With technological advances and greater knowledge nowadays, the use of crop rotation could be improved considerably regardless of the type of soil, region and climate. The success of these production systems also depends on knowledge of the particular features of each region and environment, including climate and socioeconomic variables. Such aspects will influence the establishment of the most appropriate rotation and type of management to be used, which will ensure its benefits and minimize the risks for production (Francis, 2005; Peterson et al., 2020).

Potential benefits and drawbacks of crop rotations

Crop rotation can be considered one of the key factors on the path towards sustainable agriculture. In general, crop rotation strategies can contribute to a range of important benefits (Francis, 2005; Ouda et al., 2018; Renard and Tilman, 2019), including but not limited to:

- Soil organic matter increments
- Amelioration in soil physical, chemical, and biological conditions and consequently longevity of the agricultural systems, fertility, water availability to plants, and crop yields
- Reductions in costs related to synthetic inputs (including fertilizers, pesticides, and other agrochemicals, etc.)
- Enhancement to overall farming activities and timing
- Reduction in rates of erosion by wind and water
- Boosting of overall biodiversity
- New market opportunities
- Resilience of the agricultural systems against climate as a general risk, as well as economic risks
- Reducing risks from climate-change-related threats

Crop rotation can favor many of the above aspects, but it is important to emphasize that soil management practices are a crucial part of the success and or failure of the chosen crop rotation system. Choices of crop species, harvesting techniques, tillage practices, among others, need to be carefully considered because they can have synergies and/or antagonistic effects that might either add to or counteract the benefits. For example, reduced tillage often maintains more residue on the soil surface, reduces erosion, and conserves soil moisture compared with plowing and other intensive land preparation and cultivation methods (Fuentes-Llanillo et al., 2021; Possamai et al., 2022). The additional stored moisture opens up a farmer's options for crop rotation and increasing cropping intensity through the growing season. Catch crops of short-cycle legumes may be planted following a principal cereal grain crop, or a green manure crop planted contemporaneously with or following a cereal, in combinations often not possible in many locations without this additional moisture (Salton et al., 2014; Peterson et al., 2020). In addition, reduced tillage increases earthworm populations, slows oxidation and the breakdown of organic matter to increase the long-term nutrient supply to crops (Jordan et al., 1997; Possamai et al., 2022). Rotations of tillage practices that reach different soil depths can reduce the potential for soil compaction that leads to the formation of a plow or disk pan. Conversely, depending on the crops grown in the rotation, soil type, climate, among others, reducing tillage intensity may increase the risk of plant diseases and weed infestation, meaning that the success of one is highly dependent on the other (Ouda et al., 2018). Accordingly, rotating dissimilar crops may pose a challenge for the implementation of reduced tillage, because it might require the use of different pieces of equipment at different times of the year, potentially also increasing the costs. On the other hand, in some countries/regions, the rotation of some specific crops might only be possible after adjustments in tillage intensity; due to improvements in soil moisture as mentioned above (Francis, 2005).

Crop rotation might require a large initial investment, for example, in seeds and specific types of machinery, and in addition the technical knowledge, skills, and time required from the farmer. Therefore, despite the many benefits claimed for crop rotations, they cannot be regarded as a 'silver bullet' solution. Crop rotation must, however, be regarded as one of the many actions that together with others will help on the path towards a more sustainable agricultural future, where products are not the only goal, but other ecosystem services are also valued (Ouda et al., 2018; Peterson et al., 2020). With this in mind, below are some of the main benefits that crop rotations could deliver in terms of profit, soil condition, and sustainability aspects, which altogether may represent the general basis for the future of agriculture.

Soil carbon sequestration

A crop rotation and its duration are defined by a fixed schedule of different crops planted annually or over several years, which might also include pasture or grasses and or fallow periods, with a given frequency with which one of the above uses returns to the

same agricultural field, i.e. piece of land. The alternation of crops associated with the implementation of a wider diversity of main crops and pasture or grasses and or fallow periods, can lead to changes in quantity, as well as the quality, of soil organic matter inputs (above- and below-ground) (Canalli et al., 2019). Accordingly, many recent pieces of research have indicated that crop rotation could help to enhance the delivery of many ecosystem services including improved soil organic carbon (SOC) storage when compared to monoculture systems (Salton et al., 2014; Ouda et al., 2018; Zani et al., 2021, 2022).

Increases in the quantity and quality of soil organic matter inputs are also known to benefit other soil features that are critical to the success of SOC accumulation and eventually to carbon sequestration in agroecosystems (Salton et al., 2014; Ouda et al., 2018). It is important to highlight that in this text the terms SOC accumulation and sequestration have distinctive meanings and must not be used interchangeably. The former simply indicates that SOC stocks, i.e. the balance between carbon inputs and carbon losses coming from the soil are positive for a given piece of land. Carbon sequestration, however, must consider a much broader view, which will determine whether the balance between carbon inputs and all losses, including non-carbon sources (e.g. N₂O emissions) and other carbon losses not coming from the soil (e.g. through fossil fuels uses), is positive or not. For example, the consumption of fossil fuels used to produce fertilizers and to till the soil and manage crops emits carbon into the atmosphere and these emissions are discounted from the carbon accumulated in the soil. Continuous increments in soil organic matter inputs can have a positive effect on physical (e.g. soil structure, bulk density, aggregate stability, water capacity), chemical (e.g. nutrient cycling, electrical conductivity, soil fertility), and biological (e.g. macro-, meso-, and micro-fauna) aspects, all of which are services associated with soil quality (Jordan et al., 1997; Ouda et al., 2018). Improvements in overall soil quality (i.e. physical, chemical, and biological aspects) help roots to develop and grow better, which in turn can have a direct positive effect on SOC accumulation at different depths, i.e. shallow and deeper horizons (Salton et al., 2014; Zani et al., 2021, 2022). The soil organic pool, in turn, has a direct effect on properties such as soil fertility and biodiversity and nutrient cycling that are advantageous for ecosystem services (Albizua et al., 2015; Tiemann et al., 2015).

Soil organic carbon enrichments depend on the effects of the individual crops in the rotation, but as an overall rule the more crops in the rotation the greater is the potential for SOC increments. In crop rotation systems, the diversification of root functionality with an extensive, more fibrous, and deep-rooted system usually results in larger SOC stocks compared with monoculture (Crusciol et al., 2015; Triberti et al., 2016; Zani et al., 2022). Many studies have indicated a direct relationship between root biomass, rhizo-deposits, and microbes as sources of below-ground carbon (Kell, 2012; Hirte et al., 2018). The development of a dense root system may also lead to improved soil aggregate stability (i.e. improved soil structure), and favor the protection and stabilization of soil organic matter, which is key for carbon sequestration (Zani et al., 2021). In addition, different crops will add a distinctive amount and composition of crop residues to the soil either through above-ground as well as below-ground inputs, resulting in a wide range of carbon and nutrient supply. A promising crop rotation design for the purpose of SOC accumulation is the old well-known integration of arable crops with pasture systems. While the main aim of temporary grass-clover swards is to increase productivity, nutrient supply, fertility and to break potential disease infestations, research has shown that grass-clover and or alfalfa ley periods in crop rotations can also increase SOC, with a maximum potential close to 0.4 t C ha⁻¹ year⁻¹ (Chen et al., 2015; Assmann et al., 2017; Zani et al., 2022).

Whether a ley period is used for cutting for hay or livestock grazing can also have a direct effect on the potential SOC stock. In grazing systems, extra inputs might occur from forage residues and animal manure, together with potential stimulation or change in root dynamics, turnover, and exudates from the pattern of plant removals or defoliation by herbivores (Salton et al., 2014; Crusciol et al., 2015; Triberti et al., 2016; King and Blesh, 2018; Peterson et al., 2020). Nonetheless, from a carbon sequestration perspective, grazing ruminants might also increase greenhouse gas emissions and/or reduce land available for cereal crop production, which emphasizes the complexity of decision-making about crop rotation systems.

Ultimately, it is only possible to conclude that crop rotation systems will regulate the composition of carbon pools, their stability and/or decomposability. However, as mentioned before, while crop rotation could in many cases enhance SOC accumulation, agroecosystems consist of several other management practices (and soil features or characteristics) that will together with the crop rotation, regulate inputs and outputs and therefore overall carbon sequestration potential. To summarize, an integrated approach, considering direct and indirect effects (e.g. crop growth in the rotation, market prices, management practices, soil type and characteristics, and its effect on carbon inputs and outputs, i.e. through greenhouse gas emissions) is needed when it comes to the real potential of carbon sequestration in agroecosystems that use crop rotation systems (Francis, 2005; Ouda et al., 2018).

Fertility and yield

Crop rotation has been suggested as a strategy for improvements in soil fertility and consequently crop yields. In North America, regardless of the growing conditions, a more diverse rotation system resulted in an increase in maize yield of 28% (on average) over time. It was also observed that under unfavorable conditions, yield losses can be reduced by approximately 14–90%, which also depends on other factors such as weather and soil condition in the specific array of years of the rotation per se (Bowles et al., 2020). Even under a simplistic 2-year succession system, e.g. maize-soybean, there are indications of increased yields of the former of between 5% and 10%. Similar yield increases have also been observed in other cereals grown in rotation with legumes, for example, grain sorghum-soybean, where the latter showed an approximate 10% increase in yields (Francis, 2005; Bowles et al., 2020).

Crop rotation's positive effects on yields are often attributed to improved soil fertility that accrues from the combination of a cereal-legume sequence and the increased soil organic matter that improves soil quality. The benefits of crop rotation on yields are also reinforced with current advances in hybrids and varieties of major field crops, which can lead to increases in yields of ~10%

Table 1 Examples and mechanisms for crop rotation contributions to soil fertility and nutrient cycling.

Examples or mechanisms	Contribution to soil fertility
Cereal-legume rotation	Improves soil structure, texture, tilth, quality Increases water percolation, aeration, nitrification Increase N availability in the system Promotes more efficient nutrient and water use
Different root structures	Shallow-deep species exploit more water and recycle more nutrients Improve soil aggregation and structure
Legume-grass sequence	Increase soil organic matter, N availability, and reduces pests
Biennial and perennial legumes	Increase soil organic matter and N availability, manages weeds
Longer rotations and diversity	Improves soil quality, increases soil organic matter and biodiversity of agroecosystem
Pasture-crops sequences	Increases soil organic matter, improves nutrient cycling and use-efficiency, and reduces pests

Adapted from Francis CA (2005) Crop rotations. In: *Encyclopedia of Soils in the Environment*, pp. 318–322. Elsevier. doi:10.1016/b0-12-348530-4/00253-8.

(or more), particularly when part of a crop rotation system (Francis, 2005). However, other numerous improvements to the soil system can be related to improved yields in crop rotation systems. Several specific effects of crop rotations are summarized in Table 1.

Crop rotation's positive effect on yields is a consistent observation, even when all apparent nutrient needs of crops are satisfied and there are no obvious limiting pest and disease problems, which reinforces the benefits it might provide to soil. If there are serious limiting factors that cannot be remedied by other means, for example, a severe insect infestation or plant pathogen infection, crop rotations may increase yields even more compared to monoculture systems. The main reason for this might be because different plant species have different susceptibilities and strengths and as such, they might suppress the pathogen population and/or decrease the disease severity when used interchangeably (i.e. responding differently to stresses) (Ouda et al., 2018).

Crop rotation that combines cereal-legume sequences (e.g. maize-alfalfa) not only has higher yields than continuous maize, but the need for additional N is minimal in the first year. Equally, rotation of summer and winter crops, annual and perennial crops, and crops with pasture for grazing all reduce the need to apply fertilizers. In a rotation with soybean, this is quite common because it fixes much of its N requirements, and fertilizer is often not applied to this crop (Francis, 2005; Bowles et al., 2020). In general, the reduced need for fertilizers in crop rotation results from the rotation of crop root systems, patterns of nutrient uptake, timing and different nutrient needs, and diversity in crop residues.

The importance of growing N-fixing legumes in crop rotations is their ability to fix atmospheric N with their root nodules, leaving available N for the subsequent crops. In addition to its potential benefit to productivity, this has also been linked to a potential reduction of N₂O emissions if these N-fixing legumes are associated with other grass species compared to monocultures in grass-legume mixtures (Barneze et al., 2020). There is a well-known relationship between productivity and N₂O emissions, which has been proved in many studies throughout Europe (Moreau et al., 2015). The N available in the system can be taken up by plants which increases above-ground biomass, while decreasing the emissions. The use of crop rotation has normally been associated with a reduction of fertilizer application, which can also be linked to the reduction of N emissions from the system. On average, a farmer who implemented crop rotation with legumes can increase their yield by 20% (particularly in maize) while having a cumulative reduction of N₂O of 35% (Behnke et al., 2018). This is especially important because agricultural systems are now required to be increasingly sustainable to accord with current climate change mitigation.

Improvements to the soil system as a result of crop rotation can also explain the positive effects this system can offer in areas with less rainfall. Under such conditions, if wheat, for instance, is grown in rotation with a fallow year, enough soil moisture accumulates during fallow to make the crop profitable. An average of 25% of the fallow-year moisture is stored for a succeeding crop, with the rest lost to evaporation. However, this is sufficient in traditional systems to make the wheat-fallow rotation consistently profitable. Recently such systems as Eco fallow, with minimal tillage and a sequence of two crops in 3 years (e.g. wheat-proso millet (*Panicum miliaceum* L.), wheat-sunflower (*Helianthus annuus* L.), or wheat-maize) with short fallow periods, have intensified cropping and made better use of available rainfall. In dry climates, many reports have shown consistent improvements in yields by applying longer crop rotation systems (Tamburini et al., 2020). It is possible that rotating deep- and abundant-rooted crops could favor not only water use efficiency but also other soil system aspects, as for example, soil structure, water transport, which can be used for shallow-rooted cash crops, and the allocation or recycling of other nutrients such as potassium (K) and phosphorus (P).

Despite the potential benefits that crop rotation may offer to soil fertility and yields, there are also inconsistent observations in longer crop rotations. Beyond the two-crop succession system, such as maize-soybean, the inclusion of other annual crops is thought to be related to a potential suppression from the likely use of higher inputs and consequently adverse effects on the beneficial functions that crop rotation may offer in low-input systems. Such uncertainties make the implementation of more diversified crop rotation systems challenging in the current high-demanding intensive systems.

In summary, it is important to highlight that while crop rotation can improve soil fertility and consequently increase crop yields, it is difficult to eliminate completely issues such as water and nutrient deficits, weeds, pests, and/or diseases. This is particularly

questionable for more diversified crop rotation systems, where the benefits of crop rotations might be surpassed by adverse effects from unavoidable high inputs. Nevertheless, a crop rotation system that puts such aspects as the focus for its design can certainly lead to a significant decrease in dependence on synthetic inputs, which can sometimes even eliminate the continuous use of N and synthetic pesticides.

Pest control

The current changes in climate (temperature, rainfall, etc.) together with the occurrence of extreme weather events around the globe has indicated a potential increase in pest outbreaks and the need to implement strategies for preventing and/or improving crop resilience. Planting different crops in successive seasons or years can effectively disrupt the reproductive cycles of many weeds, insects, plant pathogens, or other pest species. As such, crop rotation has been recommended as a biological method for promising effective pest control (EU, 2009; Skellern and Cook, 2018).

The rationale behind crop rotation's effectiveness against pests is that there is no host available for crop-specific parasites because of the additional time within the rotation between the growth of the potential host crops (i.e. their life-cycle is disrupted). A well-planned crop rotation can decrease the potential for weeds, pests, and diseases to reduce crop yields and their quality, and can also diminish the costs of their control. One well-known and successful example of the use of crop rotation for such purposes is the organic crop production systems. Organic systems rely almost entirely on rotations plus genetic resistance to manage crop pests without synthetic pesticide applications. Problems with soil-borne pathogens such as bacteria, fungi, and nematodes can be reduced by crop rotations that include different plant species and/or various crop families (cereals, oilseeds, and legumes). If that is accompanied by lengthening the rotation cycle, a wider spectrum of the shoot and root characteristics, biological and chemical traits, and crop management, can be achieved, which adds ecological complexity and benefits pest control (Marja et al., 2021). Nonetheless, as highlighted before, crop rotation alone cannot be regarded as a 'silver bullet' solution and the organic systems are there to 'teach' us that this must be accompanied by other management practices, including promising new ideas such as priming the soil and the use of natural materials (e.g. pheromones) among others. Nevertheless, for some soil biotic pests, rotation is the most readily and perhaps the only economic way to control them (especially in developing countries and tropical regions). Several mechanisms or methods for pest control (including a few examples) are listed in Table 2.

Longer rotations that include annual and perennial plants, or crops and pasture, can provide benefits for both soil fertility and pest protection. Annual crops allow frequent cultivation that will often suppress perennial weeds. Perennial crops may provide year-long competition that shades out annual weed species. Crop and animal mixed farming and a sequence of several years of annual crops followed by perennial grass-legume pasture mixes and grazing help to diversify income sources, and in addition control many weeds and other pests that can become problematic with the continuous cultivation of annual crops. Rotations of different animal species in the pastures such as ruminants, sheep, pigs, poultry (although unlikely at the large scales), have different preferences for grass length, for example, and also make more complete use of available forage. This also reduces animal diseases that can become problematic (they can possibly become endemic) in the continuous grazing of animals in a small space. For example, goats will eat many weed species that are overlooked by other grazing ruminants, while chickens eat insect larvae from the manure deposits of larger grazers.

While crop rotations might help to control many pests and diseases, it is important to stress once again that its success will depend on other soil management practices such as seed treatments, seeding time, and tillage (as aforementioned), among others. Furthermore, there are still contradictions in the current literature regarding its effectiveness on weed control (Marja et al., 2021),

Table 2 Mechanisms of crop rotation contributions to plant protection against weeds, insects, pathogens, nematodes, and other biotic problems.

<i>Mechanisms or process</i>	<i>Contribution to plant protection</i>
Cereal-legume sequences	Favors different weed spp., facilitates chemical and mechanical control
Unlike crop sequences	Reduces soil-borne pathogens
Maize-soybean sequence	Controls maize stem borer (<i>Busseola fusca</i>), maize rootworm (<i>Diabrotica</i> spp.)
Vegetable-cereal sequence	Controls nematodes
Potato-legume sequence	Controls wireworms (several species)
Continuous plant cover	Provides competition with weeds
Legume cover crops	Competition with undesirable weeds
Alfalfa in crop sequence	Control of annual weeds, increased predators and parasites
Strip-cutting alfalfa	Increases insect movement, increases natural enemies
Diverse crops in sequence	Soil biology diversity, suppression of soil pathogens, facilitates weed control
Cropping diversity over time	Promotes soil microorganism biodiversity

Adapted from Francis CA (2005) Crop rotations. In: *Encyclopedia of Soils in the Environment*, pp. 318–322. Elsevier. doi:10.1016/b0-12-348530-4/00253-8.

which seems to be due to the lack of other soil management control methods rather than crop rotation per se. This further emphasizes the importance of delivering multifunctional agroecosystems where crop rotation together with other soil management practices are key factors that must not be neglected.

Biodiversity

It is already known that having a greater diversity of plants improves soil microbial biodiversity and consequently soil and water quality, but the impact of rotational crop diversity on soil biodiversity is not well understood. Studies that compared soils under monoculture with those under different crop rotations showed an increase of 33% in soil carbon in mega-aggregates that formed with the diversity of crops in rotations (Tiemann et al., 2015). This increase in soil C is associated with an increase in soil structural stability because the diverse crops in rotation improve the quality and quantity of residues going into the soil, favoring the diversity of microbial communities and their activities. With the increase of microbial activities in the soil, soil organic matter can be trapped in soil mega-aggregates. Furthermore, increasing the microbial activity per se under rotations can shift the microbial community by having greater abundance of fungi compared to bacteria (Tiemann et al., 2015). Fungi are known to have an important role in the development of stable soil structure by binding the soil particles. The formation of these mega- and micro-aggregates help to trap C and N in the soil. The implementation of crop rotation has been shown to have a great impact on the composition of the microbial community, which is associated with improving soil structure and boosting soil fertility.

The use of different crops in rotation can also improve or change soil biodiversity because of the diversity in residual root exudates and plant litter (also by altering symbiotic associations) from preceding crops; these provide C substrates for a range of soil microorganisms. Crops with different root systems improve soil structure by providing optimum conditions for the growth of both more diverse and abundant biological communities. This enhanced biological diversity might favor aspects such as resource use efficiency and temporal stability of agricultural systems (Tamburini et al., 2020).

Research has also shown that crop rotation increases the richness and abundance of soil macrofauna. An experiment in Zambia showed that a maize-legume rotation increased the soil populations of earthworms, millipedes and termites (Sileshi et al., 2008). The use of green manure or fallow after rice (*Oryza sativa* L.) also enhanced soil macrofauna in the rice cropping system in China. There are also indications that monoculture cropland areas in Australia, and North and South America could benefit from the use of crop rotation in terms of pollinator abundance for pollination as well as weed control (Tamburini et al., 2020). A recent global database showed that diversified farming systems, where crop rotation is regarded as a cornerstone practice, effectively contribute to biodiversity including positive effects on species from 33 taxonomic orders (spanning insects, plants, birds, mammals, eukaryotes, annelids, fungi, and bacteria) (Jones et al., 2021). This is crucial for the diversity and multifunctionality of agricultural soils and the delivery of many ecosystem services that it provides to society.

Social and economic aspects

It is important to note that the choice of which crop rotation to put into practice is influenced by a wide range of factors, including biological interactions and contributions to yield on the farm, relative prices of differently adapted crops, and potential for production and marketing from crop and animal systems. Farmers' potential to use rotations may also depend on the agreements in place with landowners. Further, political decisions in the European Union and United States on crop commodity support payments and promotion for global markets for some selected crops narrow the farmers' options. Farmers around the EU, in particular, have chosen (sometimes even forced) to change their crop rotation accordingly.

Political decisions in the Baltic countries to position themselves for entry into the European Union caused major changes in cropping systems during the 1990s, as farmers began to compete in a regional and global market and abandoned the guaranteed markets that were common under the former Soviet system. While this situation holds true for agriculture in most parts of the United States and European Union, for some countries or regions, especially in developing countries, market price is often the key factor affecting farmers' choice of crops in a specific rotation and/or even implementing a monoculture system. For example, if prices are attractive for cereals, farmers might choose a cereal rotation that comprises mainly wheat, maize, sorghum and barley. Conversely, if legumes and/or forage seed prices are more attractive, farmers might establish a rotation that includes different bean species and peas, for example, or extend the forage seed crop for more than a year.

While financial factors are expected to be the main factor affecting crop rotation choice in developing countries, some events prove, as has happened throughout history, once again to disrupt more stable markets in developed countries. A recent example is the current conflict between Russia and Ukraine, where the latter's capacity to export crops has been greatly affected. Russia, on the other hand, is a leading exporter of energy and fertilizers, as well as crops such as wheat, barley, and sunflower seeds, but its capacity to export has been reduced because of many international sanctions. All of these have a direct effect on farmers (and of course consumers) and the price of certain crops has increased markedly.

In addition to financial factors, there are also other elements, including the availability of machinery, personal needs, family farming traditions (including knowledge and education of the farmers), and/or certain site-specific peculiarities, that influence farmers' choices when implementing a crop rotation system. It is within this very complex economic, biological, personal, and

Table 3 Mechanisms of crop rotation contributions to economic and social consequences for farming systems.

Mechanism or process	Contribution to economic or social viability
Rotation of fields	Places uniform area in all crops each year
Diverse crops in rotations	Provides forages, cereals, cash crops, soil building
Sequences of pesticides	Avoids build-up of resistance of insects, pathogens, weeds
Different tillage depths	Avoids build-up of plow-pan or compaction layer
Range in crop-planting dates	Spreads labor needs, reduces equipment costs
Range in harvest dates	Reduces peak labor demand, reduces equipment costs, increase high prices window
Diverse array of crops/products	Buffers price changes in marketplace gives direct sales
Different skills, labor needed	Increases interaction with local community for labor

Adapted from Francis CA (2005) Crop rotations. In: *Encyclopedia of Soils in the Environment*, pp. 318–322. Elsevier. doi:10.1016/b0-12-348530-4/00253-8.

political framework that crop rotations must be discussed and where many farmers (depending on the country and region) encounter the main hurdles that may discourage its implementation.

Despite these potentially adverse aspects, it is important to emphasize that there are positive economic and social aspects that result from diverse plant combinations on the farm. Table 3 lists several direct economic benefits or those that will develop as a result of biological and other soil changes in fields over time.

The rotation of large fields or within-field diversity by annual strip-cropping systems will diversify farm income and protect against a weak market for a single commodity. Planting summer and winter crops, or annual and perennial crops, will spread the demand for labor and equipment throughout the year and optimize the use of the area (Soratto et al., 2022). Such rotations provide income at different times of the year. The diversification into more intensive crops and products provides the opportunity to use labor from other farms or nearby communities, perhaps in addition and opens new markets for direct sale.

Economic diversity can be enhanced by both crop rotations and the diverse products that come from the farm. If value-added activities on the farm or in the local community can make effective use of available labor when field operations are less intense, this strategy can increase farm income and return to investment. Crop rotations and diversity of products open new economic opportunities for on-farm sales or cooperation with neighbors in other direct sales options. A diverse farm is also attractive to visitors for educational programs, farm stays, hunting, and other creative ways to add value to the rural landscape and its natural resources (e.g. increase biodiversity of the agroecosystem and consequently its resilience).

Final remarks

Although economic rewards, consolidation, and specialization drove many farmers away from rotations in the latter years of the twentieth century, it is likely that cropping diversity and rotations will become more important in the future. Government subsidy programs that link payments to the conservation of soil, water, and other natural resources often stipulate field and farm diversity as conditions for farmers to continue to qualify for payments (although this does not seem to be a global trend yet). Organic farming regulations and subsidies in some countries specify the types of rotations that must be used for farmers to qualify. There is often a rule against continuous cropping of the same crop species and a requirement to include legumes in at least half of the years in a crop sequence (BIO Intelligence Service, 2010).

With increased costs of energy, the nutrients from green manure and animal wastes become more attractive and cost-effective. When animal manure from concentrated confinement operations becomes an excessive liability because of water-quality issues, there will be greater pressure from society and from regulators to use this resource and cycle it back into the crop production process. Diverse crop rotations provide multiple opportunities throughout the year to apply manure and compost back into production fields. There is a growing appreciation to design agroecosystems that can mimic the diversity of natural ecosystems so that they can maximize productivity and minimize negative impacts on the environment. Crop rotations provide multiple benefits to farmers and society and will become increasingly important in tomorrow's agriculture.

References

- Albizua A, Williams A, Hedlund K, and Pascual U (2015) Crop rotations including ley and manure can promote ecosystem services in conventional farming systems. *Applied Soil Ecology* 95: 54–61. <https://doi.org/10.1016/j.apsoil.2015.06.003>.
- Assmann JM, Martins AP, Anghinoni I, de Oliveira Denardin LG, de Holanda Nichel G, de Andrade Costa SEVG, et al. (2017) Phosphorus and potassium cycling in a long-term no-till integrated soybean-beef cattle production system under different grazing intensities in subtropics. *Nutrient Cycling in Agroecosystems* 108: 21–33. <https://doi.org/10.1007/s10705-016-9818-6>.
- Barneze AS, Whitaker J, McNamara NP, and Ostle NJ (2020) Legumes increase grassland productivity with no effect on nitrous oxide emissions. *Plant and Soil* 446: 163–177. <https://doi.org/10.1007/s11104-019-04338-w>.

- Behnke GD, Zuber SM, Pittelkow CM, Nafziger ED, and Villamil MB (2018) Long-term crop rotation and tillage effects on soil greenhouse gas emissions and crop production in Illinois, USA. *Agriculture, Ecosystems and Environment* 261: 62–70. <https://doi.org/10.1016/j.agee.2018.03.007>.
- BIO Intelligence Service (2010) *Environmental Impacts of Different Crop Rotations in the European Union. Final report*, European Commission DG ENV Available at: https://ec.europa.eu/environment/agriculture/pdf/BIO_crop_rotations%20final%20report_rev%20executive%20summary_.pdf (Accessed 22/06/2022)
- Bolens L (1994) *Agronomos andaluces de la Edad Media*. Granada: Uni versidad de Granada, Institute de Estudios Almerienses.
- Bowles TM, Mooshammer M, Socolar Y, Calderón F, Cavigelli MA, Culman SW, Deen W, Drury CF, García AG, Gaudin ACM, Harkcom WS, Lehman RM, Osborne SL, Robertson GP, Salerno J, Schmer MR, Strock J, and Grandy AS (2020) Long-term evidence shows that crop-rotation diversification increases agricultural resilience to adverse growing conditions in North America. *One Earth* 2-3: 284–293. <https://doi.org/10.1016/j.oneear.2020.02.007>.
- Britannica (2022) Origins of Agriculture. Available at: <https://www.britannica.com/topic/agriculture> (Accessed: 30/5/2022).
- Canall LBS, Conceição PC, and Cassol C (2019) Produção de Biomassa. In: Bertol OJ, Colozzi Filho A, Barbosa GMC, Santos JB, and Guimarães MF (eds.) *Manual de manejo e conservação do solo e da água para o estado do Paraná*, pp. 133–137. Curitiba: NEPAR-SBCS.
- Chen W, Huang D, Liu N, Zhang Y, Badgery WB, Wang X, and Shen Y (2015) Improved grazing management may increase soil carbon sequestration in temperate steppe. *Scientific Reports* 5: 10892. <https://doi.org/10.1038/srep10892>.
- Crusciol CAC, Nascente AS, Borghi E, Soratto RP, and Martins PO (2015) Improving soil fertility and crop yield in a tropical region with palisadegrass cover crops. *Agronomy Journal* 107(6): 2271–2280.
- EU (2009) *Establishing a Framework for Community Action to Achieve the Sustainable Use of Pesticides*. Available at: <https://eur-lex.europa.eu/legalcontent/EN/ALL/?uri=celex%3A32009L0128> (Accessed November 01.09.2022)
- Francis CA (2005) Crop rotations. In: *Encyclopedia of Soils in the Environment*, pp. 318–322. Elsevier. <https://doi.org/10.1016/b0-12-348530-4/00253-8>.
- Fuentes-Llanillo R, Telles TS, Soares Junior D, Melo TR, Friedrich T, and Kassam A (2021) Expansion of no-tillage practice in conservation agriculture in Brazil. *Soil & Tillage Research* 208: 104877. <https://doi.org/10.1016/j.still.2020.104877>.
- Hirte J, Leifeld J, Abiven S, Oberholzer H-R, and Mayer J (2018) Below ground carbon inputs to soil via root biomass and rhizodeposition of field-grown maize and wheat at harvest are independent of net primary productivity. *Agriculture, Ecosystems & Environment* 265: 556–566. <https://doi.org/10.1016/j.agee.2018.07.010>.
- Jones SK, Sánchez AC, Juventia SD, and Estrada-Carmona N (2021) A global database of diversified farming effects on biodiversity and yield. *Scientific Data* 8: 212. <https://doi.org/10.1038/s41597-021-01000-y>.
- Jordan D, Stecker JA, Cacio-Hubbard VN, Li F, Gantzer CJ, and Brown JR (1997) Earthworm activity in no-tillage and conventional tillage systems in Missouri soils: A preliminary study. *Soil Biology and Biochemistry* 29(3–4): 489–491. [https://doi.org/10.1016/s0038-0717\(96\)00038-7](https://doi.org/10.1016/s0038-0717(96)00038-7).
- Kell DB (2012) Large-scale sequestration of atmospheric carbon via plant roots in natural and agricultural ecosystems: Why and how. *Philosophical Transactions of the Royal Society, B: Biological Sciences* 367(1595): 1589–1597. <https://doi.org/10.1098/rstb.2011.0244>.
- King AE and Blesh J (2018) Crop rotations for increased soil carbon: Perenniality as a guiding principle. *Ecological Applications* 28: 249–261.
- Marja J, Erja H, Heikki J, Katja K, Mari N, Sari H, and Lauri J (2021) Effects of crop rotation on spring wheat yield and pest occurrence in different tillage systems: A multi-year experiment in finish growing conditions. *Frontiers in Sustainable Food Systems* 5. <https://doi.org/10.3389/fsufs.2021.647335>. 2571–581X.
- Martínez NA (2005) Agriculture and food from the Roman to the Islamic Period in the North-East of the Iberian peninsula: Archaeobotanical studies in the city of Lleida (Catalonia, Spain). *Vegetation History and Archaeobotany* 14(4): 341–361.
- Moreau D, Pivato B, Bru D, Busset H, Deau F, Faivre C, Matejcek A, Strbik F, Philippot L, and Mougé C (2015) Plant traits related to nitrogen uptake influence plant-microbe competition. *Ecology* 96(8): 2300–2310. <https://doi.org/10.1890/14-1761.1.26405754>.
- Ouda S, Zohry AE, and Noreldin T (2018) *Crop Rotation: An Approach to Secure Future Food*. Switzerland: Springer.
- Peterson CA, Bell LW, Carvalho PCF, and Gaudin ACM (2020) Resilience of an integrated crop–livestock system to climate change: a simulation analysis of cover crop grazing in southern Brazil. *Frontiers in Sustainable Food Systems* 25: 604099. <https://doi.org/10.3389/fsufs.2020.604099>.
- Possamai EJ, Conceição PC, Amadori C, Bartz MLC, Ralisch R, Vicensi M, and Marx EF (2022) Adoption of the no-tillage system in Paraná State: A (re)view. *Revista Brasileira de Ciência do Solo* 46: e0210104. <https://doi.org/10.36783/18069657rbscs20210104>.
- Reddy PP (2016) Integrated crop–livestock farming systems. In: *Sustainable Intensification of Crop Production*, pp. 357–370. Springer. https://doi.org/10.1007/978-981-10-2702-4_23.
- Reich PB (2014) The world-wide ‘fast–slow’ plant economics spectrum: A traits manifesto. *Journal of Ecology* 102: 275–301. <https://doi.org/10.1111/1365-2745.12211>.
- Renard D and Tilman D (2019) National food production stabilized by crop diversity. *Nature* 571: 257–260. <https://doi.org/10.1038/s41586-019-1316-y>.
- Sahajpal R, Zhang X, Izaurrealde RC, Gelfand I, and Hurr GC (2014) Identifying representative crop rotation patterns and grassland loss in the US Western Corn Belt. *Computers and Electronics in Agriculture* 108: 173–182. <https://doi.org/10.1016/j.compag.2014.08.005>.
- Salton JC, Mercante FM, Tomazi M, Zanatta JA, Concencço G, Silva WM, and Retore M (2014) Integrated crop–livestock system in tropical Brazil: Toward a sustainable production system. *Agriculture, Ecosystems & Environment* 190: 70–79. <https://doi.org/10.1016/j.agee.2013.09.023>.
- Sleshi G, Mafongoya PL, Chintu R, and Akinnifesi FK (2008) Mixed-species legume fallows affect faunal abundance and richness and N cycling compared to single species in maize/fallow rotations. *Soil Biology and Biochemistry* 40(12): 3065–3075. <https://doi.org/10.1016/j.soilbio.2008.09.007>.
- Skellern MP and Cook SM (2018) The potential of crop management practices to reduce pollen beetle damage in oilseed rape. *Arthropod-Plant Interactions* 12: 867–879. <https://doi.org/10.1007/s11829-017-9571-z>.
- Soratto RP, Perdoná MJ, Parecido RJ, Pinotti RN, and Gitari HI (2022) Turning biennial into biannual harvest: Long-term assessment of Arabica coffee–macadamia intercropping and irrigation synergism by biological and economic indices. *Food and Energy Security* 11: e365. <https://doi.org/10.1002/fes3.365>.
- Tamburini G, Bommarco R, Wanger TC, Kremen C, van der Heijden MGA, Liebman M, and Hallin S (2020) Agricultural diversification promotes multiple ecosystem services without compromising yield. *Science Advances* 6: eaba1715. <https://doi.org/10.1126/sciadv.aba1715>.
- Tiemann LK, Grandy AS, Atkinson EE, Marin-Spiotta E, and McDaniel MD (2015) Crop rotational diversity enhances belowground communities and functions in an agroecosystem. *Ecology Letters* 18(8): 761–771. <https://doi.org/10.1111/ele.12453>.
- Triberti L, Nistri A, and Baldon G (2016) Long-term effects of crop rotation, manure and mineral fertilisation on carbon sequestration and soil fertility. *European Journal of Agronomy* 74: 47–55. <https://doi.org/10.1016/j.eja.2015.11.024>.
- Zani CF, Gowing J, Abbott GD, Taylor JA, Lopez-Capel E, and Cooper J (2021) Grazed temporary grass-clover leys in crop rotations can have a positive impact on soil quality under both conventional and organic agricultural systems. *European Journal of Soil Science* 72: 1513–1529. <https://doi.org/10.1111/ejss.13002>.
- Zani CF, Lopez-Capel E, Abbott GD, Taylor JA, and Cooper JM (2022) Effects of integrating grass-clover leys with livestock into arable crop rotations on soil carbon stocks and particulate and mineral-associated soil organic matter fractions in conventional and organic systems. *Soil Use and Management* 38: 448–465. <https://doi.org/10.1111/sum.12754>.

The contribution of organic farming systems to soil fertility—A systems perspective

Bernhard Freyer^a, Pierre Ellssel^a, Jürgen K Friedel^a, and Kurt Möller^b, ^aDivision of Organic Farming, University of Natural Resources and Life Sciences, Vienna, Austria; ^bInstitute of Crop Science, Fertilisation and Soil Matter Dynamics, Universität Hohenheim, Stuttgart, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	136
Thematic focus of the article	136
Background and definition of organic farming	136
Elements of organic farming systems	137
Soil health and soil fertility	137
Interaction between organic farming practices and soil properties	137
Impact of farming systems	137
Impact of cropping systems	138
Crop rotation systems	138
Tree based systems	140
Impact of agronomic measures	140
Soil tillage systems	140
Nutrient and organic matter systems	141
Conclusions	142
Main findings	142
Ways forward	144
References	144

Abstract

Soils in organic farming develop differently from those in mainstream farming due to different farming practices including crop rotation and agronomic measures. For health, environmental and energy reasons, organic farming does not use a number of inputs that are common in mainstream agriculture or conventional farming. As a consequence, cropping systems need to be adapted to take over or compensate for the functions of the excluded inputs. This article summarizes the effects of organic farming practices on selected properties of organically managed soils. As farming practices are decisive for the development of soil fertility, we use them as an organizational framework.

Glossary

Conventional or mainstream farming Agricultural practice, defined by the production of a small number (three to four) of crops, together with one catch crop (*Sinapis alba* or *Phacelia tanacetifolia*), using only mineral fertilizer together with pesticide application and all at increased intensity, mainly inversion tillage.

Key points

- Organic farming excludes herbicides for weed control and synthetic pesticides against pests and diseases, industrial nitrogen fertilizers and some mineral fertilizers, and limits the nutrient and carbon input via animal feed or any fertilizer according to the carrying capacity and productivity of the site (see organic guidelines)
- High diversity of crop rotations and specifically the share of forage legumes positively impacts soil fertility indicators
- Intercropping, integrated as often as possible in organic systems between two main crops, and organic amendments also contribute positively to several soil fertility indicators
- While solid organic manure primarily strengthens diverse physical and biological soil characteristics, and provides low amounts of ammonium and nitrate nitrogen, liquid manures supply the crop with highly available nutrients, specifically nitrogen as ammonium ions.
- Controversially discussed are the impacts of different soil tillage combinations on soil fertility. The organic approach, combining shallow turning and mixing the soil with loosening deeper soil layers, seems to be beneficial for soil fertility, but the final assessment also requires an integrative analysis of soil type, climatic situation, combined crop rotation, share of forage legumes and quantity and quality of intercropping

- Weeds might compete with the biomass production of the main crops, but also make their specific contribution to soil fertility, e.g., root biomass and distribution, or nutrient uptake
- The role of root exudates of the diverse crop species and varieties is not yet sufficiently researched with regard to their contribution to soil fertility (e.g., SOM)
- There is insufficient knowledge of organic farming systems for (semi-)arid and (sub-) tropical areas
- More systemized long-term trials including diverse crop rotations, forage legumes, tillage systems and organic amendments under different pedoclimatic conditions are necessary to clarify open questions

Nomenclature

AMF	Arbuscular mycorrhizal fungi
BMP	Best management practices
C	Carbon
CEC	Cation exchange capacity
DM	Dry matter
FM	Fresh matter
MBC	Microbial biomass carbon
N	Nitrogen
OF	Organic farming
SOC	Soil organic carbon
SOM	Soil organic matter

Introduction

Thematic focus of the article

This article mainly refers to a selection of key literature informing about the effects of organic farming practices on biological, physical and chemical aspects of soil fertility. The majority of research findings are based on comparative trials with conventional farming.

Background and definition of organic farming

At the beginning of the last century, Organic Farming (OF) entered the agricultural debate. Since these early days, the main approaches to OF have focused on agroecological practices that placed more emphasis on agronomic measures (e.g., crop rotation, cover cropping, organic manure, bio- and mineral fertilizers) to manage soil fertility and less on external inputs. Pioneers in this context include Lady Balfour, Hermann Brauner, Sir Albert Howard, Ewald Könemanns, Hans and Maria Müller, Hans Peter Rusch, Rudolf Steiner and others. All of them highlighted a specific aspect of fertility management in OF, e.g., R. Steiner on the combination of arable and livestock farming and the role of bio-dynamic preparations, Sir A. Howard on the use of compost and H.P. Rusch on soil microorganisms. Notwithstanding this, soil fertility is at the heart of the thinking of all the pioneers, which is a key element in understanding OF approaches and practices.

Nowadays, OF builds on the basic principles of the International Federation of Organic Agriculture Movements (IFOAM), including health, ecology, fairness and care with several obligatory, recommended or facultative guidelines (Table 1).

Table 1 Obligatory and facultative practices in organic farming systems.

<i>Farming practices</i>	<i>Obligatory</i>	<i>Facultative</i>
Soil tillage—shallow inversion and deep loosening		x
Crop rotation rules		x
Forage legume integration		x
Preference for organic amendments		x
Limitations for nutrient import via fertilizers and feed quantity and type	x	
Exclusion of synthetic nitrogen fertilizer	x	
Exclusion of superphosphate, potassium chloride, and some other industrially produced fertilizers	x	
Exclusion of a large group of synthetic pesticides incl. herbicides	x	
Exclusion of quick lime; inclusion of carbonic and dolomitic acidic lime	x	
Exclusion of Genetically modified organisms (GMO)	x	

Elements of organic farming systems

As a result of the complex interactions between different system components, the soil fertility management in organic farming relies on a long-term integrated approach rather than the more short-term targeted solutions common in conventional agriculture. In this paper we refer to the most discussed practices for assessing the impact of organic farming on soil fertility: (1) crop rotation; (2) cover cropping; (3) tillage; (4) fertilizers; (5) soil pH regulation; (6) weeds.

Soil health and soil fertility

Soil fertility is defined as “the function of the soil to act as a mediator of nutrients, water and air for plants and edaphon (soil life)” (Ottow, 2011: 421). It is the results of physical, chemical and biological processes that lead to the release of nutrients, water, aeration and stability to the plant and edaphon (Ottow, 2011), as well as the absence of any substances that may inhibit growth. Biological processes drive the transformation of carbon and nutrient cycling, improve soil structure, and regulate pests and diseases. Soil characteristics of farming systems are: diversity, abundance and activity of microorganisms and the proportion of earthworms, and other soil organisms such as arthropods; quantity and quality of soil organic matter and its characteristics, including ratios and contents of C:N, lignin, polyphenols and tannins; a high potential to prevent nutrient leaching, the proportion of bio-pores, water infiltration rates and water holding capacity, and soil aggregate stability. As the analytical products of fulvic acids and humic substances in soils are in question, we will not refer to previous research in this field.

The effects of farming practices occur in different layers of soils. We distinguish plant production-based (cropping systems) and agronomic practices (Fig. 1), which differ according to farm specific practices but also between farm types (e.g., arable cropping system (stockless or mixed), or horticultural systems), which may have different implications for soil fertility.

Interaction between organic farming practices and soil properties

Impact of farming systems

While many reviews and results of field studies on soils refer specifically to single factors influencing soil fertility parameters, the investigations cited in this section are a selection mainly comparing the two farming systems as a whole—organic and conventional—with their effects on soil properties.

Soil organic carbon levels are indicative of soil health, carbon mitigation, and provide evidence of soil quality and functions. When comparing multiple conventional and organic long-term studies, several authors have found increased aggregate stability, decreased soil bulk density, reduced runoff and soil losses, and higher SOC content in organic farming due to greater organic matter inputs from diverse sources (Gattinger et al., 2012), indicating that soils in OF are less prone to soil crusting. Jacobs et al. (2018) mainly found more SOC closely related to crop yield, a result that contradicts the previous observations, as yields in organic systems are approx. 30% lower than in conventional systems. Meta-regressions could not precisely determine the drivers, but differences in external C inputs and crop rotations (e.g., clover-grass mixtures as integral part of organic rotations) are important factors.

Organic Guidelines	Agronomic interventions		Cropping systems interventions		IFOAM Principles		
	Tillage systems		Crop rotation systems				
	Intensive (medium to deep ploughing; several tools)		Diversity (in minimum seven main crops and in addition intercrops)				
	Extensive (conservation tillage; reduced tillage; no tillage)		Forage legumes (annual and multiannual legumes mainly with grasses)				
	Systemic tillage (intensive and extensive elements combined)		Intercrop (cover-, green manure-, forage-, mulch crops etc.)				
	Nutrient and organic matter systems		Field herbs / weeds				
	Organic (solid / liquid animal manure; plant-based compost; industrial)		Tree systems				
	Minerals (macronutrients, trace elements)		Hedges (Multifunctional, species rich)				
	Lime (carbonic or dolomitic acid lime (slaked lime / natural lime))		Alleys (Leguminous species rich with multiple functions)				
	Soil fertility properties						
	Physical properties		Chemical properties			Biological properties	
	Aggregate stability		Nitrogen			Soil organic carbon	
	Bulk density		Other macronutrients			Microbial biomass	
	Biopore system		Trace elements			Microbial diversity	
	Water infiltration		C:N ratio			Enzyme activities	
	Waterholding capacity		Cation exchange capacity			Root exudates	
	Runoff / Soil loss / Siltation		pH			Antiphytopathogenic potential	

Fig. 1 Agricultural interventions of organic farming systems and their impact on selected soil characteristics.

Regardless of the farming system, a greater investment in SOC may store more nutrients, but also increases mineralization (steady state). Organic systems often show increased microbial biomass carbon, microbial biomass nitrogen, total nitrogen, greater total phospholipid fatty-acids and dehydrogenase, urease and protease activities than conventional systems, indicating higher biological soil fertility in OF systems. With greater SOC, the application of chemical fertilizers as well as supplemental irrigation could be at least partially reduced without having a negative impact on crop yields (Lal, 2020). It can be expected that better microbial performance in the organic system could lead to increased nutrient availability.

In comparative studies, soils in organic systems show the greatest ability to store and provide water to plant roots, quickly drain excess water and facilitate root proliferation (Di Prima et al., 2018). It is argued that the most relevant effect on the soil water balance is likely to be an increased water infiltration capacity depending on soil aggregate stability and bio-pores, as SOC content itself has little effect on the water holding capacity of soils (Minasny and McBratney, 2018). It remains an open question why a serious increase in water content in soils is rarely observed with a larger SOC content. One interpretation could be the immediate uptake of water by plants as well as the relocation of water to deeper horizons and to groundwater. A prerequisite for this assumption is that a more SOC increases aggregate stability and the proportion of bio-pores, thus optimizing the infiltration of water. If humus supply and soil cover via crops is high, as well as the majority of fertilizers are of organic material, and the application of mineral fertilizers is comparatively less than in conventional systems, an increased risk of nutrient leaching is relatively unlikely.

However, Williams et al. (2017) found a tenfold cumulative increase in water infiltration in a long-term experiment over 40 years in a temperate zone when they compared organic farming practices with green manure (OGM) and animal manure (OAM) to a conventional control treatment. They detected larger and more stable aggregates in the organic treatments, with the weight diameter of water-stable soil aggregates in the upper 15 cm depth increased by 50%.

There is ongoing debate as to why soil fertility and biodiversity indicators are greater on organic farms, but productivity is generally less (Ponisio et al., 2015). One example of factors explaining smaller yields is the reduced use of fertilizers, especially nitrogen, energy and pesticides. There is evidence that increased physical and biological fertility of soils in organic systems could lead to an approximation of crop yields of both systems under extreme weather conditions or that the organic system even outperforms the conventional approach (Lotter et al., 2003). Therefore, it is debated whether the yield gap between these systems could narrow with future climate changes.

Based on a number of reviews and reports on long-term trials with different soil indicators, it can be concluded that organic systems have positive effects on soil fertility in multiple ways. However, it is not always possible to determine which factor(s) of the system are the driving factors that explain the observed differences in crop yields.

A common explanation is that at the beginning of the growing season and in the following crucial growth phases, nitrogen availability does not follow the nitrogen demand of the crop in a timely manner due to soil biological processes. Other positive soil fertility properties do not compensate for this.

Impact of cropping systems

Crop rotation systems

Systems perspective

Crop rotations in organic farming differ from those under conventional farming mainly due to a greater diversity of crop species, the integration of forage legumes with grass mixtures, a larger proportion of cover crops, and to some extent, intercropping (Freyer, 2003, 2019). The approach to the design of crop rotations—i.e., the sequence of crops over time—includes a change in the family of the selected crop species from year to year, the succession of spring and winter crops (under tropical conditions also the change of long- and short-rainy season crops from different crop families), of humus reproducing and humus demanding crops, weed suppressing and weed sensitive crops, of large (e.g., alfalfa) and small water-consuming crops, a shift between deep (e.g., alfalfa) and shallow rooting crops (taproot vs. fibrous), and nitrogen-fixing with nitrogen-demanding crops (Freyer, 2003, 2019). There should be also more emphasis on the root exudates of the diverse crops used for intercropping. Long-term studies are the main source of evidence for the multiple effects of crop rotations on soil fertility in organic farming.

Crop rotations in OF take over many functions that are externalized in the conventional system and substituted by inputs such as nitrogen fertilizers, herbicides for weed control and other pesticides to control pests and diseases. It is recommended that, in OF, each crop rotation includes approx. 20–25% forage legumes, the share of root crops should not exceed 30% and cereals approx. 60%, and whenever possible cover crops should be included.

Effects of diversified organic crop rotation systems

Field experiments document a positive influence of diverse crop rotations on microbial biomass nitrogen, total carbon and total nitrogen (Lori et al., 2017) and appear to correlate linearly with the volume of 0.2–3.0 µm diameter pores (“protective” pore space), the multifunctional pore system and a decrease in soil bulk density. The improvement in soil quality is particularly evident in labile soil C and N pools, both of which are important for maintaining N fertility in organic systems and the ability to supply nutrients to crops (Delate et al., 2017). These diverse crop rotations combine species-specific root effects (e.g., type and extent of root structural impacts) to influence the individual mechanisms underlying a stable, crumbly aggregate structure and multi-functional pore system, as well as changes in physicochemical properties. Versatile crop rotations also influence microbial communities of the rhizosphere and the incidence of soil-borne pathogens, and lead to reduced incidence of pests and diseases in susceptible crops.

Forage legumes

In organic systems, the overall nitrogen budget depends primarily on forage legumes (including annual and perennial forage legumes), with biological nitrogen fixation between 80 and 500 kg N ha⁻¹ y⁻¹.

As SOC at soil depths from 30 to 100 cm accounts for 36% of the total, as established by a study in Germany, crops with large root biomass and exudate formation play a key role in determining the ability of organic farming systems to sequester carbon (Jacobs et al., 2020). They thus can compensate, to some extent, for the reduced aboveground net primary production in comparison with systems without forage legumes and driven by mineral fertilizers (Jacobs et al., 2020).

Forage legumes such as *Trifolium pratense*, *Medicago sativa*, *Desmodium* spp. or *Trifolium alexandrinum*, often grown as a mixture with diverse grass species (such as *Lolium perenne*, *Phleum pratense*, *Festuca pratensis*), are known to contribute to multiple ecosystem services that can influence overall system sustainability. Perennial legume crops, such as alfalfa, consistently increase soil organic carbon (SOC) content, showing higher levels of labile organic matter fractions, biochemical activity and aggregate stability (Tully and McAskill, 2020), and have more soil biota (earthworms, nematodes, bacteria and fungi) than conventional crop rotations.

Thus, a greater proportion of root biomass in the subsoil compared to annual crops (Fan et al., 2016), estimated at 3–6 t DM ha⁻¹ y⁻¹ compared to 1–3 t DM ha⁻¹ y⁻¹ for grass species, has a positive effect on soil SOC content, soil aggregate stability, soil water infiltration capacity, prevention of soil compaction and erosion through continuous soil cover and weed suppression. They alleviate structural damages in the subsoil through “biological drilling” (Cresswell and Kirkegaard, 1995), by greatly increasing the number of large pores (>2 mm) in the subsoil compared to continuous cultivation of annual crops and pastures. In another study, soils under organic management with an alfalfa-based crop rotation showed increased rootability, thinner roots and greater root density of succeeding crops compared to conventional rotations (Monaci et al., 2017).

Under tropical soil, climate and agricultural conditions—in addition to the two-year grass-clover or alfalfa-, forage legumes such as pigeon pea (*Cajanus cajan*), lablab (*Lablab purpureus*), desmodium (e.g., *Desmodium intortum*) and others improve the total N content in the uppermost soil strata in the long-term and increase soil organic carbon in the upper soil layer (Kumar et al., 2020). Grain legumes with large biomass production, such as mung bean (*Vigna radiata* L.) and lablab (*Lablab purpureus*), are most effective in improving the structure of a compacted heavy soil.

Intercropping

In our case, we define intercropping (cover crops; catch crops) as plants that grow between two main crops with different uses. Theoretically, these crops can serve as fodder for animals, as food crops, as feedstock for anaerobic digestion, as green manure for erosion control and from a biodiversity perspective, or as a pre-crop for cereals and root crops, which has a positive effect on soil fertility. Two types of sowing are of relevance:

- *Undersown crops*: Primarily white and red clover, mixed with grasses, are used as undersown crops, especially in cereals. The transfer of fixed nitrogen during growth in the shade of cereals can be expected to be small (Høgh-Jensen and Schjoerring, 1997). After harvesting the main crop, the clover grass mixture develops into a main crop or continues as an intercrop. When cultivated as an intercrop, it can be assumed that after a growth period starting in march under a winter cereal, and harvested in October or the following year before start of a summer crop, approximately 40–80 kg N ha⁻¹ of fixed nitrogen is available for the following crop and the biomass is between 1 and 4 Mg DM ha⁻¹, the latter with positive impacts on soil structure, biology and weed suppression. Relay cropping, defined as an undersown crop in a main crop at the end of the growing season to take advantage of water surpluses, allows multiple uses. The expected yields vary between 0.5 and 3 t DM ha⁻¹.
- *Blank seeding*: These crops are sown after a main crop and harvested before the subsequent main crop. They need a growing period of at least 6 weeks, up to approx. 4 months if they are not frost-resistant, or they continue to grow until the following spring. Clover mixtures can fix up to about 60–80 kg N ha⁻¹ if cultivated until spring of the next year.

The nutrients in the biomass can serve as an early and easily available nutrient source in spring, specifically in case of leguminous crop mixtures (Marcillo and Miguez, 2017). Other effects include a reduced soil erosion risk, improved soil aggregate stability through soil cover, a dense rooting system providing the development of bio-pores, as well as the uptake of nutrients that prevents their leaching, always according to the plant material characteristics, as well as soil and climate conditions.

The biomass production of intercrops depends on species characteristics, climate, soils as well as sowing date and duration in the field. The substantially greater importance of the roots of intercrops compared to the aboveground biomass for soil C processes is discussed by various authors (Ladoni et al., 2016). The effect of intercropping on SOC is controversially discussed. Some authors argue that the effectiveness of these crops for carbon sequestration, per se, is overestimated, while others found significant increases in SOC concentrations after several years, although intercropping did not show an overall impact on SOC concentrations over short time periods. Factors influencing these processes are the plant species, the total plant biomass, the age of the material, its C:N ratio, the proportion of polyphenols, lignin, and root exudates.

Intercrop diversity increases carbon input to the rhizosphere and soil microbial biomass in the soil and may also have allelopathic effects on pests, diseases and weeds (Bahadur et al., 2015). Mixtures usually produce a larger yield of root biomass than the pure stands and therefore have a higher potential for efficient soil exploration by roots.

Tree based systems

The literature on agroforestry and alley cropping is comprehensive, but only few studies are available that specifically address the effects in organic farming systems. However, from a systemic point of view, agroforestry and especially alley cropping, is a fundamental element in organic farming as it contributes to many ecosystem services that are necessary for the maintenance of the organic system.

The concept of agroforestry is based on the observation that higher crop yields are obtained at a certain distance from trees or in places where trees were previously grown, and hence a higher fertility of soils in such areas can be expected. Various field experiments and reviews on alley cropping systems in different climate-soil environments have shown the ability of prunings from several tree species to deliver sufficient nutrients to meet crop demands, except for phosphorus. Shrubs and trees in diverse combinations have multiple functions in such a system, including mitigation and adaptation to climate change (Reyes et al., 2021). Under organic farming conditions, the agroforestry species of importance are those that contribute to soil fertility and do not compete with the main crop for agronomic inputs, i.e., with tap roots, rich root biomass, limited water demand (drought resistance), and the ability to fix nitrogen. The often-deep-rooted systems enable the uptake of nutrients from deeper soil layers, creating bio-pores that are relevant as habitats for microorganisms.

Patterns of nutrient release from organic materials are partly determined by their chemical composition as well as by soil type and weather conditions. Leguminous biomass releases nitrogen immediately from leaf and thin young root material, unless they contain high levels of lignin or polyphenols. Litter and roots of both legumes and non-legumes immobilize or mineralize N, depending on the size of the C:N ratio, with mineralization being associated with a small ratio. Competition between trees and plants for resources is often argued over, but several studies report no significant competitive disadvantages for crop yields when yield is summed over the entire plot size (Yamoah et al., 1986).

Impact of agronomic measures

Tillage and nutrient management determine detailed aspects of cropping systems. Their fine-tuning is necessary for soil fertility. The interplay between cropping and tillage systems as well as fertilizer strategies are decisive for crop yield, while avoiding negative effects on soil fertility.

Soil tillage systems

Systems approach

The overall directive for soil tillage in organic farming is the shallow mixing but only partial inversion of the upper fertile soil layer from 10 to 20 cm depth, together with the loosening of deeper layers with a cultivator (Peigné et al., 2018). Using this approach, the aims are: (i) to avoid or limit mixing upper fertile soil layers with deeper less fertile soil layers; (ii) to prevent a compacted layer developing between upper and lower soil horizons; (iii) to loosen the subsoil and fertile upper soil layers to create optimal conditions for root and plant growth, but also to avoid soil compaction and erosion. Tillage modifies soil structure and crop residue distribution, which in turn affects the substrate availability for soil micro-organisms and their ability to break down soil organic matter and release nutrients for crop growth. Soil tillage is also a key management tool in organic farming to suppress weeds, pests and diseases.

In addition to this general directive, multiple combinations of tillage technologies are used—such as shallow inversion tillage and non-inversion minimum tillage, no-tillage and cover crop mulch-based systems, making a general assessment of the impact of “organic” soil tillage technologies on soil fertility challenging. Likewise, “deep” plowing encompasses a variety of tillage systems and techniques with different impacts on soil quality and soil functions, but most publications on field studies in OF lack a detailed description of the plowing technique, thus the interpretation of their impact on soil fertility indicators is limited.

Diversified soil tillage systems

The main positive effects of reduced tillage systems are the reduction in the risk of soil erosion, owing to more stable soil aggregates and a mulch layer over the soil surface. In addition, reduced tillage systems result in a more pronounced stratification of soil properties (e.g., SOC, microbial activity, macronutrients, such as phosphorus and potassium), compared with “deep” inversion plowing (Jacobs et al., 2020). However, SOC enrichment at the soil surface due to reduced tillage is offset, for example, by smaller contents deeper in the soil profile. The net effects are hotly debated, the relevance in relation to annual anthropogenic CO₂ emissions is small (Hunt et al., 2020). Reasons are that deep plowing might reduce the SOC in the upper layer due to less incorporation of organic matter, while in the lower layers organic matter is also enriched, which is not the case with reduced tillage. If trials only focus on the first 15 cm soil depth they ignore the development in the lower layers, which might lead to these controversial observations.

Reduced tillage is associated with an increase in earthworm species and activity (Peigné et al., 2018), shifts in microbial communities, increased survival of arbuscular mycorrhizal fungi (AMF) and less siltation due to higher aggregate stability and humus accumulation in the upper soil layer, in addition to a reduction in fuel consumption and required workload. The effects on soil are mainly observed in the upper layer. If both soil layers are analyzed, more SOC is observed in deeper layers when deep plowing is applied. In those cases, the positive effects on the soil biota are more evenly distributed between upper and lower soil

layers. Depending on the specific conditions prevailing, in some cases reduced tillage systems, in other cases more intensive tillage systems perform better in terms of their effects on the associated soil characteristics.

Most experiments on organic farming indicate yields are smaller in response to reduced tillage, but larger yields have also been reported (Peigné et al., 2018). The most important factors determining yield were reduced nitrogen supply and greater weed infestation under reduced tillage (Cooper et al., 2016). In cases where reduced tillage led to increased weed pressure and thus reduced yields, a solution could be found through the use of clover grass, a crop rotation of spring and winter crops and thus the possibility of integrating cover crops.

In the case of no-tillage, soil warming and mineralization is slower, while biodiversity and biomass of soil organisms are enhanced. The risk of weed infestation is also very high. Mulching to reduce weed pressure additionally delays soil warming. After years of no-till, consolidation due to natural soil settlement is to be expected. Thus, no-tillage can be an option, but is accompanied by several challenges. Compared with seed beds without residues (conventional no-till), improved drill designs, e.g., with mulch cover, might reduce weed pressure, increase soil moisture, aggregate stability, soil organism activity and, specifically in case of maize or soybean, also reduce the risk of siltation and erosion, and thus might become a feasible approach for the organic system. Yields are even greater in conventional no-till systems and therefore further technological innovations are necessary for the mulching system to contribute both to soil fertility and crop yield.

Each tillage system has advantages and disadvantages, therefore shallow inversion plowing and the two-layer plow, or shallow loosening and temporary deeper loosening of soils could be a compromise to maintain or enhance the positive effects on soil fertility without negatively affecting crop yield.

Nutrient and organic matter systems

Systems approach

In organic farming the idea of nutrient and organic matter application is to optimize soil conditions via the “feeding of soil organisms,” increasing carbon content and optimizing soil chemical, physical and biological conditions. In practice, this is the case when biomass-rich solid organic amendments are applied. A large proportion of these fertilizer-derived nutrients first undergo a mineralization or decomposition process by soil organisms before they are available to plants. In the case of liquid animal fertilizers (slurry and biogas-slurry) or organic commercial liquid fertilizers (organic industrial fertilizers; biofertilizers) based on food and brewery residues, urban and communal waste and others, the nutrients are offered directly to the plant roots in larger proportions.

Farmgate nutrient budgets show that results can vary greatly between farms and farming systems (arable with and without livestock, vegetable production, etc.) depending on the share of imported nutrients (Reimer et al., 2020). Although nutrient export is comparatively smaller from OF than most conventional farms, soil stocks of P, K and other macro and trace elements are reduced over time (Reimer et al., 2020). Only off-farm fertilizers (and feeds) that are qualitatively and quantitatively accepted by organic standards can be used to contribute to balancing the farm nutrient budget, i.e., offsetting nutrient deficits in the soil, whereas on-farm organic matter can be redistributed from field to field but does not compensate for nutrient exports at the farm level.

The nutrient export of nitrogen is balanced through N-fixation by legumes as well as atmospheric deposition, which in Europe can contribute annually in excess of 20 kg ha⁻¹. All other nutrients have to be supplied mainly from off-farm, depending on the results of the soil analysis, the nutrient balance, the site-specific ability of tap-root crops like alfalfa to mobilize nutrients in the subsoil, and further information on the state of soil fertility, e.g., via spade diagnosis.

Therefore, to maintain soil fertility over the long term, a strict strategy to ensure efficient use of on-farm available nutrients from soil nutrient stocks is required, including optimal on-farm nutrient cycling via organic matter, constant supply of organic matter to managed fields via forage legumes and cover crops, and limiting nutrient losses through erosion and leaching.

Organic fertilizers

Organic matter sources are diverse—crop residue, straw, green manure, farmyard manure, compost, digestates, or residues from food processing—and contribute differently to soil fertility. Organic manures positively influence the functional diversity of microorganisms in the soil. Organic manure also improves soil physical properties with a significant increase in soil porosity and hydraulic conductivity, and a decrease in bulk density. Farmyard manures of different qualities (fresh or in different composted stages) show positive effects on soil parameters like SOC, earthworms and aggregate stability.

In organic systems with animal husbandry, solid manure application can improve soil fertility indicators, although their direct contribution to crop yield is relatively small, due to the small proportion of ammonium and nitrate nitrogen. In the case of liquid manures, nutrient availability for crops is large, specifically that of ammonium nitrogen, and directly enhance crop yield. Relatively small quantities can be applied using rates of approximately 10–20 m³ ha⁻¹ according to the crop demand during the vegetation period, limit the risk of nutrient leaching. The contribution of liquid manure to crop yield greatly depends on the application technique and its capacity to avoid losses by ammonia volatilization.

Mineral fertilizers

In organic farming, most readily soluble synthetic mineral fertilizers are prohibited with the objectives of reducing energy inputs and of feeding soil organisms rather than plants directly. There is also the working hypothesis that product quality and product shelf life are negatively affected, if highly soluble fertilizers are applied. All forms of mineral nitrogen fertilizers are excluded by the organic farming guidelines, instead rhizobia bacteria from legumes take over the supply of nitrogen, accompanied by numerous other functions with positive effects on soil fertility.

After an analysis of the nutrient status of the soils, the nutrient balance of the farm and approval by an accredited organic farming certifier, the addition of mineral fertilizers is permitted. However, the philosophy of only applying mineral fertilizers that need to be converted by microorganisms is only partially fulfilled, e.g., highly soluble potassium sulfate is allowed.

Under organic management, soils with a low level of available P content (e.g., analyzed via P calcium acetate lactate extract) were often found (Kolbe, 2015), whereas a larger organic P (P_{org}) pool was observed than under conventional management. This pool is particularly important in grassland soil, less so under arable farming.

The main challenges for soil fertility in terms of providing adequate nutrients in quality and quantity for plant growth, is the non-adapted nutrient stoichiometry of the most commonly used base fertilizers—solid animal manures or composts and commercial industrial fertilizers (including mineral fertilizers)—as they do not match the crop specific demand. For example, cattle or poultry manure contain twice as much or even more phosphorus per unit compared to plant N and P offtakes. Also, the composition of liquid manures is closer to the nutrient stoichiometry of vegetative products. As a result, the use of soil nutrients is inefficient, leading to an inappropriate supply of nutrients in terms of timing and quantity, with the risk of nitrogen losses through gases and leaching.

Liming

The pH of agricultural soils must be within an optimal range to ensure an efficient use of fertilizers, i.e., nutrient storage and availability, biological nitrogen fixation, aggregate stability, and a pronounced soil life. pH values between 5.8 and 7.2 offer optimal living conditions for microorganisms. In the case of pH values that are below 5.5 the effective CEC is reduced due to the reduced cation buffering capacity. A further consequence of decreasing pH is the limitation of available molybdenum, which negatively affects the nitrogenase activity and biological nitrogen fixation.

The organic guidelines allow the use of natural lime, such as limestone ($CaCO_3$) or its transitional lithotypes like lime marl, crayon or calcareous marine algae, while quicklime (CaO) or slaked lime ($Ca(OH)_2$) (structural liming) are excluded (Table 1). The composition and abundance of most of the soil fauna is affected by natural liming, including fungi, bacteria, archaea, nematodes, earthworms, and microarthropods. The effects of natural liming on soil biota and biological processes vary significantly, but with evidence of positive effects (Holland et al., 2018). Structural liming (quick lime and carbonic acid lime (slaked lime or natural liming)) may have negative short-term effects on microbial communities, earthworm and coleopteran populations. Liming of acidic soils stabilizes the organic matter content and improves mineralization (Fageria, 2012). To what extent the continuous application of different types of carbonic acid lime can lead to partially similar effects of quicklime is not known. The extent to which the pH value in organic farming differs from that in conventional farming is also not yet known, although some results from long-term experiments do not indicate any significant differences (Rodas-Gaitan et al., 2022). Although in organic farming forage legumes contribute to a pH reduction, in conventional farming mineral fertilizer composition varies and therefore also their influence on soil acidity. In all systems, a continuous application of animal manure and compost could contribute to an increase in soil pH, provided the initial pH is less than 7.4 (Whalen et al., 2000).

With regard to soil physical characteristics, natural lime application can aid in fostering a favorable soil structure and aggregates by developing clay-humus-complexes, increasing the stabilization of organic matter in aggregates and thus reducing the risk of soil erosion and silting, while promoting soil aeration, water infiltration and a favorable thermal balance (Holland et al., 2018).

The type of liming and the composition with other minerals, elements and as well as the particle size play an outstanding role in low input systems, especially their effects on soil biota and crop yield, as confirmed by several authors (Li et al., 2019). Currently, there is a lack of research focusing on organic systems, e.g., in relation to the type and composition of liming. However, such research would be of great importance, as a better adaptation of liming strategies could contribute to a reduction of yield differences between organic and conventional agriculture.

Conclusions

Main findings

There is evidence in the scientific literature that the contribution organic farming systems to soil fertility in terms of biological (e.g., soil microbial activity, earthworm population, etc.) and some soil physical (e.g., aggregate stability) properties is relatively large compared to many conventional systems (Table 2). At the same time, at least in organic farming systems, there is a long-term threat to soil (chemical) fertility in terms of soil phosphorus and potassium status due to long-term depletion resulting from negative structural farm balances of these nutrients. Field experiments also indicate that under climate change conditions, the relative yield difference between organic and conventional agriculture could decrease.

Often it is not a single practice but the organic system as a whole that contributes to the physical and biological fertility of soils, while the exclusion or incomplete application of relevant practices (e.g., adequate recycling of nutrients lost in exported products or by leaching) can have negative consequences for overall biomass productivity and thus soil fertility. Consequently, if some key organic practices, such as consistent nutrient recycling, are not applied, the conventional system provides similar or better results on some soil fertility indicators. Recent reviews and meta-analyses have shown that organic farming systems have more soil organic carbon. However, it can be assumed that when comparing organic farms without forage legumes with heavier yielding conventional farms, the latter show greater SOC content, whereas integrating forage legumes into organic farms results in larger SOC values compared to conventional farms of medium productivity.

Table 2 Effects of organic farming practices on soil fertility in comparison to conventional systems.

Sector/practice	Comparison of organic with conventional systems
<i>Cropping systems</i>	
Crop rotation and organic fertilizers	• Similar to higher levels of SOC depending on yields and straw management in conventional systems and forage crops in the organic crop rotations • Less soil-borne disease • Higher microbial biomass C, N and activity, higher diversity of soil microbial community • Smaller bulk density, greater aggregate stability or stable, crumbly aggregate structure • Reduced soil crusting, greater share of multi-functional pore system • Less runoff, less soil loss • Greater enzyme activities (phospholipid fatty-acids, and dehydrogenase, urease and protease activities) • Similar or higher soil N mineralization potential • Larger share of labile organic N
Forage legumes	• Large root biomass and rhizodeposition; nitrogen input • Overall positive effects on physical, chemical and biological soil indicators
Cover crops	• Short term improvement of soil physical conditions, small long-term increase in SOC • Increased availability of nutrients for the following crops (mainly with legumes)
Herbs / weeds	• Diversification of insect habitats, root exudates, microorganisms, and nutrient decomposition
<i>Tree- and shrub-based systems</i>	
Hedges	• Important element of the organic farming system (erosion control, biomass production, biodiversity, habitat for beneficials, etc.) • Systemized data of different hedge types functionalities are missing
Alley crops	• Fundamental part of the organic farming system, but integration in practice is still limited • Positive effects on soil erosion reduction, SOC, and nitrogen if leguminous species are used; transfer of nutrients from lower to upper soil layers
<i>Tillage systems</i>	
Tillage	• Several trade-offs, no-tillage is partly critical due to the risk of increased weed pressure in the organic system and as a consequence reduced crop biomass/yield
<i>Nutrient and organic matter systems</i>	
Organic fertilizer	• Increase in SOC and positive effects on diverse soil biological and physical indicators • Risk of long-term imbalances in the soil due to an imbalanced nutrient stoichiometry of organic fertilizers compared to the demand of crops
Mineral fertilizer	• No mineral nitrogen fertilizer allowed • Biological N-fixation via legumes with multiple positive effects on soil fertility • Lower use of phosphorus and potassium mineral fertilizer reduces risk of nutrient losses
External inputs	• Lower use of external inputs leads to lower nutrient concentrations in animal manure • Nutrient balances of phosphorus, potassium and other nutrients tend to be negative
Liming and pH	• Not in general a specific characteristic of one or the other system • Besides liming, animal manure and compost could be of relevance for pH regulation

These findings underline that the combination of crop rotations, based on forage legumes and cover crops, together with application of organic amendments and a moderate-intensity of soil tillage, represent the most promising practices with positive effects on the different soil fertility indicators, but always require a detailed site-specific adaptation of the whole crop-technical system. Forage legume-based cropping systems play a key role as they provide large inputs of carbon below- and aboveground and nitrogen, and have multiple positive effects on soil biota and soil physical and chemical characteristics. Forage legumes are the strongest biological competitor against weeds, which, at least in arable cropping systems, are one of the most important biological regulating factors limiting organic crop yield. Cover cropping as a key strategy in organic farming systems has multiple positive effects on soil fertility, especially with legumes mitigating the effects of nitrogen as a yield-limiting factor. Depending on the length of the vegetation period of cover crops, seed mixtures and climatic conditions, they have different effects on soil indicators. Minor positive effects on SOC can be expected after several years of application.

The various soil tillage systems sometimes have contradictory effects on soil properties, weed and disease incidence, which affects the overall productivity of the cropping system. Systems with reduced or no-tillage (the latter is rarely established) applied on organic farms have led to a stratification of many soil properties with positive effects on soil organic matter and several biological, chemical and physical soil parameters, primarily in the upper soil layers, while the values of these indicators are reduced in deeper soil layers.

Differences between the tillage system impact on total SOC in deeper soil layers are not convincingly established and are probably of only minor relevance. Due to observed reductions in crop yield as a consequence of increased weed pressure, especially under more humid site conditions, reduced mineralization of nutrients and impairment of soil structure (compaction), reduced tillage has not yet become established on most organic farms. The regulation of soil pH to create an optimal pH range for

microorganism diversity and activity is fundamental to nutrient availability in organic systems. Less important for soils but more for crop yield is the relationship between nutrient availability and the mismatched nutrient stoichiometry of fertilizers according to crop demand.

Ways forward

There are some factors that influence crop yield levels with direct consequences for soil fertility. For future improvement of organic farming systems, a better understanding of the main yield-limiting factors in different regions, cropping systems and single crops is needed to identify approaches to improve the overall system productivity and performance. Possible yield-limiting factors are either weeds, pests or diseases, or the supply of nitrogen and sulfur, which tend to have short-term effect, while phosphorus, potassium as well as sulfur are likely to have a longer-term yield limiting effect, with negative feedback effects on the nitrogen balance by impairing the potential of legumes to fix nitrogen.

A higher yield level will also result in more crop residues, potentially increasing positive effects on soil physical and biological properties. A better understanding of how weed control can be improved using modern technologies to provide further options to reduce the overall intensity of tillage could also be useful to improve approaches to protect soil from erosion.

Improving the overall recycling rate of nutrients between urban centers and agriculture is also a major challenge to achieve more sustainable overall nutrient cycles in our societies and improve soil chemical fertility, not only in OF systems. Organic industrial fertilizers as well as biofertilizers (biostimulants) is a relatively new group of fertilizers, where knowledge on soil fertility and crop yield is limited and needs further research. The multiple effects on soil biota, which vary according to lime type and properties, may not be well understood and require further research to optimize soil quality and effects on crop yields.

By their very nature, changes in soils and productivity may occur shortly after the implementation of a measure or with a certain time lag, sometimes of an additive nature. They may be of long duration or affect soils only during the current growing season, or become relevant only if repeated over time. The manifold effects of agricultural practices on soil fertility only become visible after several years of their application or are subject to changes over time, which are also due to weather influences. Therefore, long-term experiments are needed to better understand the effects of individual practices on soil fertility.

A further challenge is the identification of multiple effects, for example: (1) interactions between cover crops, tillage systems and weed pressure; (2) additive effects of multiple technical interventions and fertilizer amendments on soil fertility; (3) effects of diverse organic farming practices on aggregate stability and water dynamics; (4) research on root exudates and their impact on the inhibition of seed germination and crop growth rate. It is therefore useful to distinguish between the components of soil fertility that change relatively slowly, perhaps over the course of a crop rotation or in some cases over decades, and the direct contribution of materials such as fertilizers and manures.

As more extensive conventional systems overlap with the organic system, a differentiation of both systems has its limitations. There is a great deal of variety and nuance in both systems, which requires a classification of farming systems to better understand the distinction between the organic systems themselves and in comparison to the various conventional systems.

Further research is also needed regarding a multipurpose use of our landscape for biodiversity, fiber, energy and tree or wood production. This includes the impact of photovoltaic modules as integral part of a landscape (e.g., instead or in combination with hedges), or the effects of production of crops under photovoltaic modules on crop growth, water balance, temperature, etc. There is also a need for further research of tree-based systems in relation to: (1) soil fertility; (2) competition of resources between trees or shrubs and crops and landscape water balance.

Currently, most of the literature on organic farming systems refers to temperate regions, and less to arid or (sub-) tropical conditions. Here, the need for research is very great and urgent due to the fact that climate change increases the risk to food security, while a better understanding of soil fertility under organic farming conditions could contribute to an increase in production, especially in regions with limited access to external sources such as fertilizers or pesticides.

References

- Bahadur S, Verma S, Prasad S, Madane A, Maurya S, Gaurav VV, and Sihag S (2015) Eco-friendly weed management for sustainable crop production—A review. *Journal of Crop and Weed* 11(1): 181–189.
- Cooper J, Baranski M, Stewart G, Nobel-de Lange M, Bärberi P, Fließbach A, and Casagrande M (2016) Shallow non-inversion tillage in organic farming maintains crop yields and increases soil C stocks: A meta-analysis. *Agronomy for Sustainable Development* 36(1): 22.
- Cresswell H and Kirkegaard J (1995) Subsoil amelioration by plant-roots—the process and the evidence. *Soil Research* 33(2): 221–239.
- Delate K, Cambardella C, Chase C, and Turnbull R (2017) A review of long-term organic comparison trials in the US. *Sustainable Development of Organic Agriculture* 101–118.
- Di Prima S, Rodrigo-Comino J, Novara A, Iovino M, Pirastru M, Keesstra S, and Cerdà A (2018) Soil physical quality of citrus orchards under tillage, herbicide, and organic managements. *Pedosphere* 28(3): 463–477.
- Fageria NK (2012) Role of soil organic matter in maintaining sustainability of cropping systems. *Communications in Soil Science and Plant Analysis* 43(16): 2063–2113.
- Fan J, McConkey B, Wang H, and Janzen H (2016) Root distribution by depth for temperate agricultural crops. *Field Crops Research* 189: 68–74.
- Freyer B (2003) *Crop Rotation*. Stuttgart: Eugen Ulmer GmbH & Co.
- Freyer B (2019) The role of the crop rotation in organic farming. In: Köpke U (ed.) *Improving Organic Crop Cultivation*, pp. 547–568. Burleigh Dodds Science Publishing Limited.
- Gattinger A, Müller A, Haeni M, Skinner C, Fließbach A, Buchmann N, and Scialabba NE-H (2012) Enhanced top soil carbon stocks under organic farming. *Proceedings of the National Academy of Sciences* 109(44): 18226–18231.

- Høgh-Jensen H and Schjoerring J (1997) Interactions between white clover and ryegrass under contrasting nitrogen availability: N₂ fixation, N fertilizer recovery, N transfer and water use efficiency. *Plant and Soil* 197(2): 187–199.
- Holland J, Bennett A, Newton A, White P, McKenzie B, George T, and Hayes R (2018) Liming impacts on soils, crops and biodiversity in the UK: A review. *Science of the Total Environment* 610: 316–332.
- Hunt JR, Celestina C, and Kirkegaard JA (2020) The realities of climate change, conservation agriculture and soil carbon sequestration. *Global Change Biology* 26(6): 3188–3189.
- Jacobs A, Flessa H, Don A, Heidkamp A, Prietz R, Dechow R, and Schneider F (2018) *Landwirtschaftlich genutzte Böden in Deutschland: Ergebnisse der Bodenzustandserhebung*. Thünen Report.
- Jacobs A, Poeplau C, Weiser C, Fahrion-Nitschke A, and Don A (2020) Exports and inputs of organic carbon on agricultural soils in Germany. *Nutrient Cycling in Agroecosystems* 118(3): 249–271.
- Kolbe H (2015) Wie ist es um die Bodenfruchtbarkeit im Ökolandbau bestellt: Nährstoffversorgung und Humusstatus? *Bodenfruchtbarkeit-Grundlage erfolgreicher Landwirtschaft (BAD Tagungsband 2015)* 89–123.
- Kumar S, Meena RS, Datta R, Verma SK, Yadav GS, Pradhan G, and Mashuk H (2020) Legumes for carbon and nitrogen cycling: An organic approach. In: *Carbon and Nitrogen Cycling in Soil*, pp. 337–375. Springer.
- Ladoni M, Basir A, Robertson PG, and Kravchenko AN (2016) Scaling-up: Cover crops differentially influence soil carbon in agricultural fields with diverse topography. *Agriculture, Ecosystems & Environment* 225: 93–103.
- Lal R (2020) Soil organic matter content and crop yield. *Journal of Soil and Water Conservation* 75(2): 27A–32A.
- Li Y, Cui S, Chang SX, and Zhang Q (2019) Liming effects on soil pH and crop yield depend on lime material type, application method and rate, and crop species: A global meta-analysis. *Journal of Soils and Sediments* 19(3): 1393–1406.
- Lori M, Symnackiz S, Mäder P, De Deyn G, and Gattinger A (2017) Organic farming enhances soil microbial abundance and activity—A meta-analysis and meta-regression. *PLoS One* 12(7), e0180442.
- Lotter DW, Seidel R, and Liebhardt W (2003) The performance of organic and conventional cropping systems in an extreme climate year. *American Journal of Alternative Agriculture* 18(3): 146–154.
- Marcillo G and Miguez F (2017) Corn yield response to winter cover crops: An updated meta-analysis. *Journal of Soil and Water Conservation* 72(3): 226–239.
- Minasny B and McBratney A (2018) Limited effect of organic matter on soil available water capacity. *European Journal of Soil Science* 69(1): 39–47.
- Monaci E, Polverigiani S, Neri D, Bianchelli M, Santilocchi R, Toderi M, and Vischetti C (2017) Effect of contrasting crop rotation systems on soil chemical and biochemical properties and plant root growth in organic farming: First results. *Italian Journal of Agronomy* 12(4).
- Ottow JC (2011) *Mikrobiologie von Böden: Biodiversität, Ökophysiologie und Metagenomik*. Springer-Verlag.
- Peigné J, Vian J-F, Payet V, and Saby NP (2018) Soil fertility after 10 years of conservation tillage in organic farming. *Soil and Tillage Research* 175: 194–204.
- Ponisio LC, M'Gonigle LK, Mace KC, Palomino J, De Valpine P, and Kremen C (2015) Diversification practices reduce organic to conventional yield gap. *Proceedings of the Royal Society B: Biological Sciences* 282(1799): 20141396.
- Reimer M, Möller K, and Hartmann TE (2020) Meta-analysis of nutrient budgets in organic farms across Europe. *Organic Agriculture* 10(1): 65–77.
- Reyes F, Gosme M, Wolz KJ, Lecomte I, and Dupraz C (2021) Alley cropping mitigates the impacts of climate change on a wheat crop in a mediterranean environment: A biophysical model-based assessment. *Agriculture* 11(4): 356.
- Rodas-Gaitan H, Fritz J, Dahn C, Köpke U, and Joergensen RG (2022) Biodynamic compost effects on soil parameters in a 27-year long-term field experiment. *Chemical and Biological Technologies in Agriculture* 9(1): 1–11.
- Tully KL and McAskill C (2020) Promoting soil health in organically managed systems: A review. *Organic Agriculture* 10(3): 339–358.
- Whalen JK, Chang C, Clayton GW, and Carefoot JP (2000) Cattle manure amendments can increase the pH of acid soils. *Soil Science Society of America Journal* 64(3): 962–966.
- Williams DM, Blanco-Canqui H, Francis CA, and Galusha TD (2017) Organic farming and soil physical properties: An assessment after 40 years. *Agronomy Journal* 109(2): 600–609.
- Yamoah C, Agboola A, and Wilson G (1986) Nutrient contribution and maize performance in alley cropping systems. *Agroforestry Systems* 4(3): 247–254.

Soil functions to support natural plant communities and crops: Soil multifunctionality

Leidivan Almeida Frazão^a, Luiz Arnaldo Fernandes^a, Adriana Marcela Silva Olaya^b, and Igor Costa de Freitas^a, ^aUniversidade Federal de Minas Gerais, Instituto de Ciências Agrárias, Montes Claros, Minas Gerais, Brazil; ^bUniversidad de la Amazonia, Florencia, Caquetá, Colombia

© 2023 Elsevier Ltd. All rights reserved.

Introduction	146
Biomass production	146
Nutrient cycling	147
Climate regulation and carbon sequestration	148
Water purification and regulation	149
Functional and intrinsic biodiversity provision	149
Transformation of potentially harmful elements and compounds	150
Conclusion	150
References	150

Abstract

The main ecological functions of soil include nutrient cycling, carbon storage and turnover, water regulation, soil habitat, above-ground diversity regulation, biotic regulation, buffering, and transformation of potentially harmful elements and compounds. The optimization of productivity concomitant with maintaining beneficial soil conditions is determined by various factors, including the physical, chemical, and biological qualities essential for the sustainable production of plant biomass and food security. As favorable soil conditions are essential for delivering multiple ecosystem services, in this chapter, we discuss the various soil functions and correlate them with several effects of human activities, agricultural practices and land use systems for different soil types and climate zones.

Key points

- Soil conservation management practices improve soil conditions and productivity.
- Soil fertility and biodiversity maintenance reduce the impact of climate change on agroecosystems.
- Sustainable management is critical for soil structure maintenance and water purification and regulation.

Introduction

As a central interface between the atmosphere, hydrosphere, lithosphere and biosphere, the soil fulfills various functions (Brevik et al., 2015), including sustaining plant biomass and animal production nutrient and water cycling, carbon storage, buffering, and transformation of potentially toxic elements and compounds (Haygarth and Ritz, 2009), which turn the soil into an essential component of food security and, consequently, life on the planet (Rojas et al., 2016). Additionally, sustainable soil management promotes other essential environmental services, such as water quality, balance of atmospheric gases, and biodiversity (Fig. 1).

Biomass production

The soil has the capacity to supply nutrients and water to produce plant biomass for human use, providing food, feed, fiber, and fuel within natural or managed ecosystem boundaries (Sandén et al., 2019). Productive soils have sufficient and balanced nutrient levels and support plant growth through adequate physical, chemical and biological quality, humidity and temperature, and contain no toxic elements (Lopes and Guilherme, 2007).

In agriculture, the soil plays a critical role. Besides providing physical support for plant growth, the soil must have an adequate porosity for root respiration, water infiltration, storage, and availability, a balanced pH and essential nutrients to support biomass production. Soil particle size (sand, silt, and clay) distribution also influences the productivity by affecting the soil porosity, nutrient retention and availability (e.g., cation exchange capacity [CEC]). Additionally, topography (slope) and climate (precipitation and temperature) influence soil productivity through influencing weathering and erosion, and also affect plant growth and development. Vegetation type, microorganisms, enzyme activity, other soil organisms, and the amount of above- and belowground organic

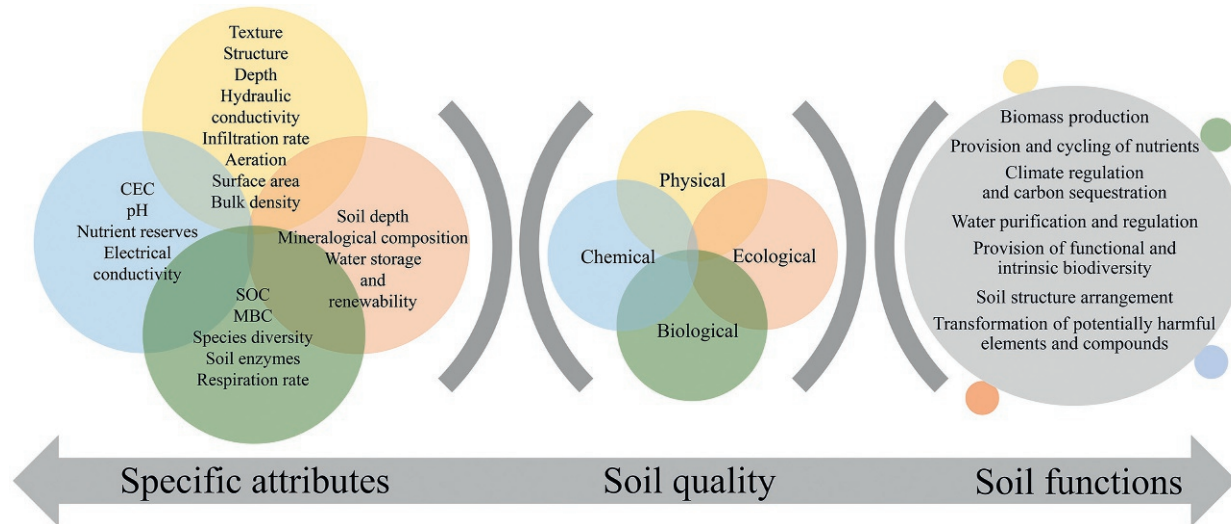


Fig. 1 Soil functions related to soil quality. SOC: soil organic carbon; MBC: microbial biomass carbon; CEC: cation exchange capacity. Adapted from Lal R (2016) Soil health and carbon management. *Food and Energy Security* 5(4): 212–222, <https://doi.org/10.1002/fes3.96>.

residues are affected by climate. Increased production of organic residues results in high carbon inputs in humid climates, which promote nutrient cycling, soil aggregation, and structure resulting from, for example, increased microbial activities, which increase the soil's CEC (Karlen, 2005).

Even when nutrients and water are limited, soil fertility management can improve soil health and increase the primary productivity by enhancing nutrient cycling and the capability of water capture and storage (Foley et al., 2011). Soil productivity is the capacity of the soil to perform productive functions and to optimize yield through adapting different management practices. Soil productivity can be determined by measuring the difference between inputs (such as water, nutrients, genetic resources, and pest control) and outputs (production or crop yield). Soil erosion, nutrient extraction exceeding nutrient inputs, organic matter loss, acidification, salinization, flooding, and soil compaction are the primary causes of soil productivity loss. Erosion removes the soil surface layer, which is generally more fertile because of its higher biological activity and organic matter content (Karlen, 2005).

Minimizing soil disturbance, increasing plant diversity, and maximizing soil through management practices such as zero tillage, crop residue management, and the use of cover crops can improve soil health in agroecosystems by influencing biotic and abiotic productive soil components, that favor soil productivity (Miner et al., 2020), and thus, gradually increase soil organic matter content. Techniques reducing surface runoff, such as terracing and strip cultivation, contribute to soil quality maintenance (Karlen, 2005).

Since soil microorganisms are fundamental for nutrient cycling and have essential interactions with plants, management practices that favor the abundance and diversity of microbial communities can increase crop productivity, reduce production costs by reducing needs of mineral fertilizer and pesticide use, and promote sustainability in agroecosystems (Cardoso et al., 2013).

Nutrient cycling

Essential substances for plant life, i.e., necessary for the plant to complete its life cycle, are called nutrients. According to the elemental analysis, plants are composed of mineral and non-mineral elements, essential (nutrients) or not for the life cycle. Non-mineral nutrients are carbon (C), hydrogen (H), oxygen (O) and represent approximately 90% of the total elemental composition of plants and are obtained from water and air. Mineral nutrients are nitrogen (N), phosphorus (P), potassium (K), calcium (Ca), magnesium (Mg), sulfur (S), boron (B), chlorine (Cl), copper (Cu), iron (Fe), manganese (Mn), molybdenum (Mo), zinc (Zn), cobalt (Co) and nickel (Ni). There are other elements that are beneficial to plants but are not essential, such as silicon (Si), sodium (Na), cobalt (Co) and selenium (Se). Mineral elements represent about 10% of the elemental composition of plants and are obtained from the soil (Taiz et al., 2015).

Nutrient cycling is one of the ecosystem services provided by soils, and it is influenced by soil biological and physicochemical properties, agroecosystems management, and climate. Most nutrients needed for plants are provided through soils and need to be available when required. The factors responsible for soil formation and, consequently, for the availability of nutrients are: parent material, climate, living organisms, time and relief (topography). The factors climate and organisms (active factors) act together with time and relief on the parent material. The availability of water (rainfall) and temperature variations are the main climate elements responsible for the chemical and physical changes of rock minerals (parent material) (Weil and Brady, 2017).

Plant-absorbed nutrients return to the soil and are recycled through the mineralization of organic compounds present in litter or crop residues. However, since in agricultural systems, nutrients are removed from the system through harvests and the crop residues

remaining in the system might not have sufficient nutrients to meet the nutrient demand of the subsequent crop, nutrient replacement practices via application of mineral fertilizer or organic amendments and other corrective measures are necessary to supplement natural nutrient input (for example, nitrogen via biological fixation) (Schröder et al., 2016) to ensure that nutrient inputs and outputs are balanced in the long term.

The ability of soil to recycle nutrients is affected by the capacity of soil to receive and retain nutrients, to make nutrients bioavailable, and to facilitate the extraction and export of nutrients by crops (Schröder et al., 2016), processes that are influenced by soil properties including content of Fe^{2+} , Al^{3+} , and Ca^{2+} compounds, CEC, pH, micronutrients content (important for the symbiosis of rhizobia and legumes), parent material, organic matter content, texture, porosity, soil bulk density, drainage, water retention, microbial activity (important for organic matter mineralization), climatic conditions (temperature and humidity), and management practices in agricultural systems (Van Breemen, 1995; Delgado and Gomez, 2016; Schröder et al., 2016).

Plants take up nutrients (typically in the form of dissolved ions) from the soil solution. The ions adsorbed, precipitated as soluble salts, or available as mineralized organic compounds constitute the “labile nutrient pool,” which is in equilibrium with the soil solution. Labile nutrients added to the nutrients in the soil solution constitute the “bioavailable nutrient pool.” The balance of nutrients and exchange rates between the labile fraction and the soil solution is influenced by chemical (adsorption/desorption or precipitation/solubilization) and biological (immobilization/mineralization) interactions controlling the ionic activities of the solution and transport of nutrients to plant roots (Delgado and Gomez, 2016). Following absorption, nutrients are distributed among plant leaves, stems, and roots, and nutrients are returned to the soil through senescence, exudation, and leaching. Weathering of primary minerals, atmospheric deposition, groundwater supply, and other sources increase the soil nutrient content, whereas leaching reduces it (Van Breemen, 1995).

In tropical ecosystems, where temperature and humidity are higher, rock weathering is more intense (Weil and Brady, 2017). Under these conditions, the soils are deeper, more acidic and comparatively poor in nutrients and non-recalcitrant organic matter. On the other hand, in arid and cold climate conditions, weathering is less intense and the soils are shallower, with a greater amount of nutrients and organic matter.

In highly weathered soils, soil organic matter plays an important role for the cation exchange capacity (CEC), nutrient availability and contaminant filtering through sorption in organo-clay complexes. Soil microorganisms are responsible for the decomposition of organic material (including xenobiotic organic pollutants) and, consequently, nutrient cycling. As for the source of carbon, microorganisms can be grouped into autotrophs and heterotrophs. Autotrophic microorganisms produce their own food (reduced carbon) through photosynthesis and chemosynthesis. Heterotrophic microorganisms, on the other hand, need reduced carbon (organic compounds) for their metabolism (Bollag and Stotzky, 2000). This way, microbial heterotrophic metabolism (decomposition of organic materials) regulates essential processes for maintaining life on planet, such as nutrient cycling and the ability to retain and exchange cations and anions in the soil.

Maintaining adequate contents of soil organic matter is essential for nutrient cycling. Thus, in cultivated areas, conservation management practices (e.g., no-till) should be encouraged. On the other hand, at least in the short term, the addition of organic materials to the soil can stimulate and accelerate the mineralization of the soil organic matter. When adding organic residues, there are intense short-term changes in the mineralization of soil organic matter - positive priming effect. In contrast, the delay in the mineralization of soil organic matter due to the addition of organic materials is known as the negative priming effect (Kuzakov, 2010).

Degradation of organic substrates results from several sequential chemical reactions mediated or catalyzed by enzymes (e.g., ligninases, cellulases, proteases, glucosidases, galactosidases, phosphatases, amidases, urease, sulfatase) that play a fundamental role in ecosystems, acting as catalysts for several reactions that result in the decomposition of organic materials, nutrient cycling and formation of soil organic matter (Tabatabai, 1994). This way, all biochemical transformations on the planet are also dependent on or are related to the presence of enzymes.

Climate regulation and carbon sequestration

In general, the soil is considered the largest reservoir of terrestrial organic carbon (C), storing two times the mass of C present in the atmosphere and three times the mass of C in the biota that live on the land surface (Lal, 2018). By holding this huge carbon stock, the soil plays an important role in climatic change mitigation.

The soil ability to regulate the climate is determined based on their potential to take up nitrogen from the atmosphere, promote C sequestration, and the reduction of greenhouse gas (e.g., CH_4 and N_2O) emissions (Van de Broek et al., 2019). A positive balance between soil carbon inputs and outputs allows the soil to function as a carbon sink, enabling improved climate regulation and sustainability (Shovik et al., 2015).

The soil C sequestration - the process of transferring carbon dioxide (CO_2) from the atmosphere into stable terrestrial C pools that are not re-emitted back into the atmosphere - is influenced by climatic factors (rainfall and temperature), soil texture, mineralogy, amount of crop residues, and/or addition of organic inputs, as well as by the management practices adopted by farmers (e.g., no-till).

Clayey soils generally have a greater potential for C sequestration than sandy soils, due to the influence the clay on organic matter stabilization processes impacting the longevity and persistence of SOC (Schmidt et al., 2011). In clay dominant soils the chemical adsorption of C onto the surfaces of clay minerals and the physical occlusion within microaggregates reduce the access of soil microbes to organic C (Sissoko and Kpombrekou-A, 2010).

Diverse mechanisms including electrostatic bonding, van der Waals force, hydrogen bond and coordination with oxyhydroxides and silicate clays are associated to the C absorption in clay minerals. The properties of the organic compounds, as well as the type of clay, the soil acidity and moisture can result in different mechanisms occurring at the same time.

Conservation-oriented soil management can contribute to the reduction of greenhouse gas emissions and the increase of CO₂ sequestration, which results in improvements in soil fertility and biodiversity, reduction of erosion processes and water pollution. These positive effects can help to reduce the impacts of climate change on agroecosystems (Smith, 2012; Paustian et al., 2016) in the long-term, even though the organic carbon present in the soil is highly dynamic. For example, the formation of stable microaggregates and organomineral complexes can promote the protection of organic carbon from microbial processes for millennia (Lal, 2016).

Several management strategies can be used to increase carbon inputs in agroecosystems, such as the use of species with deeper and larger root systems, crop rotations, use of cover crops, irrigation in places with water deficit, fertilization in soils with nutritional deficits to increase crop productivity, and use of perennial crops to minimize soil disturbance. Less intensive soil tillage, such as zero tillage, has been shown to promote carbon sequestration (Bai et al., 2019; Lee et al., 2019) by favoring aggregation and aggregate stability which is considered as the main mechanism promoting increased C (Lal, 2016; Van de Broek et al., 2019).

Water purification and regulation

Water purification and regulation is the ability of a soil to remove harmful compounds from water and the capability of the soil to receive, store, and conduct water, thereby contributing to the maintenance of high surface and groundwater quality, prevention of droughts, floods, and erosion, and food security maintenance and climate change mitigation (Wall et al., 2020).

In general, excessive use of agrochemicals can contaminate the soil and lead to the pollution of water resources. Additionally, land use change and degradation of agricultural lands can alter water quality through soil erosion, dissolution, and leaching of nutrients and other pollutants. However, through processes, such as precipitation, adsorption, desorption, ion exchange, oxidation-reduction, and metabolic decomposition, the soil can remove many soil pollutants favoring the remediation of natural ecosystems and agricultural lands (Cheng et al., 2021).

The ability of the soil to regulate water storage is influenced by environmental characteristics (e.g., precipitation volumes, frequency and intensity), soil chemical and physical characteristics (e.g., quality and quantity of organic matter, pH, type of clay minerals, and bulk density), and agronomic practices. Organic soil amendments and the adoption of practices that result in soil surface cover affect positively the water regulation function of soils by enhancing physical and hydraulic properties in soil, including soil porosity, bulk density, saturated hydraulic conductivity and aggregate stability (Jian et al., 2020). The surface cover protects the soil structure against degradation resulting in surface sealing, can increase the soil organic matter contents (Poeplau and Don, 2015), microbial biomass, community composition and biological activity, processes that promote stable soil aggregation, reduction in soil bulk density, increases in soil porosity, which in turn improves the flow of water and solutes in soil (Blanchy et al., 2023).

The conservation and improvement of soil structure are essential for the performance and maintenance of soil quality and water retention (Rabot et al., 2018). Soil structure is the spatial arrangement between mineral particles and soil organic matter and the arrangement and size of the pores between these particles, considering its stability against external forces (e.g., machine traffic and raindrop impact). Soil structure is considerably altered by soil management practices and cultivation techniques, which can modify the soil water holding capacity as well as water flow (Delgado and Gomez, 2016).

Total porosity, macroporosity, and connectivity between soil pores are also important indicators of soil structural quality, and can be used to assess soil functions and the several processes that occur after agricultural operations. Desirable values for the aforementioned indicators are generally positively correlated with large biomass yields, adequate water storage and filtration, nutrient storage, recycling, carbon storage, favorable habitat for a greater diversity of soil organisms to develop and perform their activities, and provide adequate soil physical support (Rabot et al., 2018).

Functional and intrinsic biodiversity provision

Soil biodiversity is critical for the maintenance of ecosystem services, and is controlled by a combination of environmental, biological, chemical, and physical processes (Thiele-Bruhn et al., 2012). Soil biodiversity comprises a multitude of organisms and processes interacting with an ecosystem that supports life above ground (Van Leeuwen et al., 2019).

Soil fauna (bacteria, fungi, protists, and invertebrates) is directly related to several ecosystem functions, such as nutrient cycling and storage, formation and turnover of soil organic matter, decomposition, crop production, reduction of pathogenicity potential, pest control, and increased resistance and resilience to the effects of extreme weather conditions. Therefore, the increase in soil biodiversity increases plant diversity in biomes and the self-regulatory capacity of agricultural and natural ecosystems (Thiele-Bruhn et al., 2012; Delgado-Baquerizo et al., 2020; Wall et al., 2020, 2015). Soil biota can promote soil aggregation, thereby improving water infiltration owing to the effect of fungi and bacteria (Wall et al., 2015). The performance of these soil biodiversity functions is substantially influenced by the physicochemical properties of the soil, particularly pH and organic carbon content (Van Leeuwen et al., 2019), but also by the climate and management decision in agricultural systems.

Soil functions to support natural plant communities and crops are strongly linked to the metabolic activity of microorganisms and soil organic matter. Thus, the soil is a biochemical system highly regulated by the processes of decomposition of organic materials and inorganic transformations, carbon stock and greenhouse gas emissions (Bollag and Stotzky, 2000; Tabatabai, 1994). In the context of soil system functionality, determining enzyme activity is a tool used to verify whether the soil is fulfilling its functions to support natural plant communities and crops. In general, the enzymes linked to the carbon cycle (beta-glucosidase and dehydrogenase), nitrogen (urease), phosphorus (phosphatases) and sulfur (arylsulfatase) are more sensitive to changes that occur in the soil and have been proposed for use as quality indicators (Mendes et al., 2019).

Agricultural activities and management systems can alter the structure and richness of biological communities (Thiele-Bruhn et al., 2012). Soil management practices that alter organic matter content affect biological activity. Organic matter is a source of carbon for various organisms, such as soil microbiota (Delgado and Gomez, 2016). Conservation management techniques, such as no-till, crop residue maintenance, cover crops, crop rotation, integrated pest management, integrated soil fertility management (combination of organic and mineral fertilizers), and agroforestry practices increase soil biodiversity, thereby improving human, animal, and plant health owing to the suppression of disease-causing organisms and the supply of clean air, water, and food (Wall et al., 2015).

Transformation of potentially harmful elements and compounds

Soil acts as a geochemical sink for contaminants and a natural buffer controlling the transport of chemical elements and substances to the atmosphere, hydrosphere, and biota (Kabata-Pendias, 2011) and consequently plays a global ecological role. Therefore, processes of soil pollution can affect quickly and significantly the physical, chemical and biological properties of soil impacting the essential ecological functions of soil. According to Simon (2014), natural and human impacts can cause point-like or diffuse soil contamination and pollution. Industrial, agricultural, and urban activities release potentially harmful trace elements, which can cause soil and water contamination. In agricultural soils, Zheng et al. (2007) revealed that potentially toxic elements such as Cd, Cu, Hg, Pb, and Zn may influence plant nutrient uptake, and their concentrations in the tissues of food and fodder crops can affect the quality of food and drinking water with potential implications to human health.

The availability of potentially harmful elements and compounds to plants and microorganisms depends on the form they are present in the soil solution. This presence is governed by soil composition, soil reaction, oxidation-reduction conditions, and reaction kinetics, which in turn depends on soil attributes and their tendency to form insoluble precipitates and co-precipitates with other minerals, complexes with organic matter, and promote mineral absorption (Camargo et al., 2001).

Although the selectivity of each element for adsorption sites in soil particles differ considerably due to the chemical, physical, and mineralogical characteristics of the soils, the amount of metals retained is positively correlated with organic matter, CEC, specific surface, and clay and oxide content. pH is one of the most important soil solution parameters because it is positively correlated with the adsorption of metals (Ross, 1994).

Considered as a constituent of the soil matrix in the form of organic colloids, humus is made up of recalcitrant molecules from vegetable and microbial, combined through polymerization and resynthesis reactions, with phenolic compounds left over from lignin (Sutton and Sposito, 2005). The main effects of humus are the improvement in plant nutrition and in the physical and chemical properties of the soil, including the immobilization of potentially toxic elements and organic pollutants (Ameri and Tehranifar, 2013).

The remediation of potentially toxic elements also can be achieved through immobilization using different soil amendments, and details can be found in the chapter "Remediation of soils polluted with heavy metals & organic pollutants." However, soil properties such as pH, clay, sesquioxides, and organic matter content, and processes such as sorption/desorption and redox processes, are the key factors governing the effectiveness of amendments through the immobilization of potentially toxic elements in soils.

Conclusion

Soil functions can be affected by many biological, chemical, and physical factors, as well as the complex interactions among these factors. However, the ability of a soil to provide these functions depends on its properties. In this context, anthropogenic activities directly influence the sustainable productivity of soils, and conservation management systems are vital in maintaining soil health and quality and ensuring sustainable food production.

References

- Ameri A and Tehranifar A (2013) Effect of humic acid on nutrient uptake and physiological characteristic *Fragaria ananassa* var: camarosa. *Acta Horticulturae* 6: 77–79. <https://doi.org/10.17660/ActaHortic.2014.1049.54>.
- Bai X, Huang Y, Ren W, Coyne M, Jacinthe PA, Tao B, Hui D, Yang J, and Matocha C (2019) Responses of soil carbon sequestration to climate-smart agriculture practices: A meta-analysis. *Global Change Biology* 25(8): 2591–2606.

- Blanchy G, Bragato G, Di Bene C, Jarvis N, Larsbo M, Meurer K, and Garré S (2023) Soil and crop management practices and the water regulation functions of soils: A qualitative synthesis of meta-analyses relevant to European agriculture. *The Soil* 9: 1–20. <https://doi.org/10.5194/soil-9-1-2023>.
- Bollag JM and Stotzky G (2000) *Soil Biochemistry*. New York: Marcel Dekker.
- Brevik EC, Cerdà A, Mataix-Solera J, Pereg L, Quinton JN, Six J, and Van Oost K (2015) The interdisciplinary nature of soil. *The Soil* 1: 117–129. <https://doi.org/10.5194/soil-1-117-2015>.
- Cardoso EJB, Vasconcellos RLF, Bini D, Miyauchi MYH, Santos CA, Alves PRL, Paula AM, Nakatani AS, de Moraes Pereira J, and Nogueira MA (2013) Soil health: Looking for suitable indicators. What should be considered to assess the effects of use and management on soil health? *Scientia Agricola* 70: 274–289. <https://doi.org/10.1590/S0103-90162013000400009>.
- Camargo OA, Alleoni LRF, and Casagrande JC (2001) Reações dos micronutrientes e elementos tóxicos. In: Ferreira ME, Cruz MCP, Van Raij B, and Abreu CA (eds.) *Micronutrientes e elementos tóxicos na agricultura*, pp. 89–124. Jaboticabal, São Paulo: CNPq/FAPESP/POTAFOS.
- Cheng K, Xu X, Cui L, Li Y, Zheng J, Wu W, Sun J, and Pan G (2021) The role of soils in regulation of freshwater and coastal water quality. *Philosophical Transactions of the Royal Society B* 376(1834): 20200176.
- Delgado A and Gomez JA (2016) The soil. Physical, chemical and biological properties. In: Villalobos FJ and Fereres E (eds.) *Principles of Agronomy for Sustainable Agriculture*, pp. 15–26. Cham, Switzerland: Springer International Publishing AG. https://doi.org/10.1007/978-3-319-46116-8_2.
- Delgado-Baquerizo M, Reich PB, Trivedi C, Eldridge DJ, Abades S, Alfaro FD, Bastida F, Berhe AA, Cutler NA, Gallardo A, García-Velázquez L, Hart SC, Hayes PE, He JZ, Hseu ZY, Hu HW, Kirchmair M, Neuhauser S, Pérez CA, Reed SC, Santos F, Sullivan BW, Trivedi P, Wang JT, Weber-Grellon L, Williams MA, and Singh BK (2020) Multiple elements of soil biodiversity drive ecosystem functions across biomes. *Nature Ecology & Evolution* 4(2): 210–220. <https://doi.org/10.1038/s41559-019-1084-y>.
- Foley JA, Ramankutty N, Brauman KA, Cassidy ES, Gerber JS, Johnston M, Mueller ND, O'Connell C, Ray DK, West PC, Balzer C, Bennett EM, Carpenter SR, Hill J, Monfreda C, Polasky S, Rockström J, Sheehan J, Siebert S, Tilman D, and Zaks DPM (2011) Solutions for a cultivated planet. *Nature* 478: 337–342. <https://doi.org/10.1038/nature10452>.
- Haygarth PM and Ritz K (2009) The future of soils and land use in the UK: Soil systems for the provision of land-based ecosystem services. *Land Use Policy* 26(S1): S187–S197. <https://doi.org/10.1016/j.landusepol.2009.09.016>.
- Kabata-Pendias A (2011) *Trace Elements in Soils and Plants*, 4th edn Boca Raton: CRC Press/Taylor & Francis.
- Karlen DL (2005) Productivity. In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*, pp. 330–336. Elsevier. <https://doi.org/10.1016/B0-12-348530-4/00241-1>.
- Kuz'yakov Y (2010) Priming effects: Interactions between living and dead organic matter. *Soil Biology and Biochemistry* 42: 1363–1371. <https://doi.org/10.1016/j.soilbio.2010.04.003>.
- Jian J, Lester B, Du X, Reiter M, and Stewart R (2020) A calculator to quantify cover crop effects on soil health and productivity. *Soil & Tillage Research* 199: 104575. <https://doi.org/10.1016/j.still.2020.104575>.
- Lal R (2016) Soil health and carbon management. *Food and Energy Security* 5(4): 212–222. <https://doi.org/10.1002/fes3.96>.
- Lal R (2018) Digging deeper: A holistic perspective of factors affecting soil organic carbon sequestration in agroecosystems. *Global Change Biology* 24(8): 3285–3301.
- Lee H, Lautenbach S, Nieto APG, Bondeau A, Cramer W, and Geijzenborgh IR (2019) The impact of conservation farming practices on Mediterranean agro-ecosystem services provisioning—A meta-analysis. *Regional Environmental Change* 19: 2187–2202. <https://doi.org/10.1007/s10113-018-1447-y>.
- Lopes AS and Guilherme LRG (2007) Fertilidade do solo e produtividade agrícola. In: Novais RF, Alvarez VH, Barros NF, Fontes RLF, Cantarutti RB, and Neves JCL (eds.) *Fertilidade do Solo*. Sociedade Brasileira de Ciência do Solo: Viçosa.
- Mendes IC, Souza LM, Sousa DMG, Lopes AAC, Reis-Junior FB, Lacerda MPC, and Malaquias JV (2019) Critical limits for microbial indicators in tropical oxisols at post-harvest: The FERTBIO soil sample concept. *Applied Soil Ecology* 139: 85–93. <https://doi.org/10.1016/j.apsoil.2019.02.025>.
- Miner GL, Delgado JA, Ippolito JA, and Stewart CE (2020) Soil health management practices and crop productivity. *Agricultural & Environmental Letters* 5: e20023. <https://doi.org/10.1002/ael2.20023>.
- Paustian K, Lehmann J, Ogle S, Reay D, Robertson GP, and Smith P (2016) Climate-smart soils. *Nature* 532(7597): 49–57. <https://doi.org/10.1038/nature17174>.
- Poeplau C and Don A (2015) Carbon sequestration in agricultural soils via cultivation of cover crops – A meta-analysis. *Agriculture, Ecosystems & Environment* 200: 33–41. <https://doi.org/10.1016/j.agee.2014.10.024>.
- Rabot E, Wiesmeier M, Schlüter S, and Vogel HJ (2018) Soil structure as an indicator of soil functions: A review. *Geoderma* 314: 122–137. <https://doi.org/10.1016/j.geoderma.2017.11.009>.
- Rojas RV, Achouri M, Marouli J, and Caon L (2016) Healthy soils: A prerequisite for sustainable food security. *Environmental Earth Sciences* 75(3): 1–10. <https://doi.org/10.1007/s12665-015-5099-7>.
- Ross SM (1994) Retention, transformation and mobility of toxic metals in soils. In: Ross SM (ed.) *Toxic Metals in Soil-Plant Systems*, pp. 63–152. New York: Wiley.
- Sandén T, Trajanov A, Spiegel H, Kuzmanovski V, Saby NPA, Picaud C, Henriksen CB, and Debeljak M (2019) Development of an agricultural primary productivity decision support model: A case study in France. *Frontiers in Environmental Science* 7: 58. <https://doi.org/10.3389/fenvs.2019.00058>.
- Schmidt M, Torn M, Abiven S, Dittmar T, Guggenberger G, Janssens IA, Kleber M, Kögel-Knabner I, Lehmann J, Manning DAC, Nannipieri P, Rasse DP, Weiner S, and Trumbore SE (2011) Persistence of soil organic matter as an ecosystem property. *Nature* 478: 49–56. <https://doi.org/10.1038/nature10386>.
- Schröder JJ, Schulte RPO, Creamer RE, Delgado A, Van Leeuwen J, Lehtinen T, Rutgers M, Spiegel H, Staes J, Toth G, and Wall DP (2016) The elusive role of soil quality in nutrient cycling: A review. *Soil Use and Management* 32(4): 476–486. <https://doi.org/10.1111/sum.12288>.
- Shovik D, Singh BPB, Biswapati M, Amitava R, and Bahadur SH (2015) Soil organic carbon: Towards better soil health, productivity and climate change mitigation. *Climate Change and Environmental Sustainability* 3: 26–34. <https://doi.org/10.5958/2320-642X.2015.00003.4>.
- Simon L (2014) Potentially harmful elements in agricultural soils. In: Bini C and Bech J (eds.) *PHEs, Environment and Human Health*, pp. 85–150. Dordrecht: Springer. https://doi.org/10.1007/978-94-017-8965-3_3.
- Sissoko A and Kpomblekou-A K (2010) Carbon decomposition in broiler litter-amended soils. *Soil Biology and Biochemistry* 42(4): 543–550. <https://doi.org/10.1016/j.soilbio.2009.10.007>.
- Smith P (2012) Soils and climate change. *Current Opinion in Environmental Sustainability* 4(5): 539–544. <https://doi.org/10.1016/j.cosust.2012.06.005>.
- Sutton R and Sposito G (2005) Molecular structure in soil humic substances: The new view. *Environmental Science & Technology* 39: 9009–9015. <https://doi.org/10.1021/es050778q>.
- Tabatabai MA (1994) Soil enzymes. In: Weaver RW, et al. (ed.) *Methods of Soil Analysis. Part 2. Microbiological and Biochemical Properties*. Madison: SSSA. Book Series 5.
- Taiz L, Zeiger E, Moller IM, and Murphy A (2015) *Plant Physiology and Development*, 6th edn Sunderland: Sinauer Associates.
- Thiele-Bruhn S, Bloem J, Vries FT, Kalbitz K, and Wagg C (2012) Linking soil biodiversity and agricultural soil management. *Current Opinion in Environmental Sustainability* 4(5): 523–528. <https://doi.org/10.1016/j.cosust.2012.06.004>.
- Van Breemen N (1995) Nutrient cycling strategies. *Plant and Soil* 168(1): 321–326. <https://doi.org/10.1007/BF00029344>.
- Van de Broek M, Henriksen CB, Ghaley BB, Lugato E, Kuzmanovski V, Trajanov A, Debeljak M, Sandén T, Spiegel H, Decock C, Creamer R, and Six J (2019) Assessing the climate regulation potential of agricultural soils using a decision support tool adapted to stakeholders' needs and possibilities. *Frontiers in Environmental Science* 7: 131. <https://doi.org/10.3389/fenvs.2019.00131>.
- Van Leeuwen JP, Creamer RE, Cluzeau D, Debeljak M, Gatti F, Henriksen CB, Kuzmanovski V, Menta C, Pérès G, Picaud C, Saby NPA, Trajanov A, Trinsoutrot-Gattin I, Visioli G, and Rutgers M (2019) Modeling of soil functions for assessing soil quality: Soil biodiversity and habitat provisioning. *Frontiers in Environmental Science* 7: 113. <https://doi.org/10.3389/fenvs.2019.00113>.
- Wall DH, Nielsen UN, and Six J (2015) Soil biodiversity and human health. *Nature* 528(7580): 69–76. <https://doi.org/10.1038/nature15744>.

- Wall DP, Delgado A, O'Sullivan L, Creamer RE, Trajanov A, Kuzmanovski V, Henriksen CB, and Debeljak M (2020) A decision support model for assessing the water regulation and purification potential of agricultural soils across Europe. *Frontiers in Sustainable Food Systems* 4: 115. <https://doi.org/10.3389/fsufs.2020.00115>.
- Weil R and Brady N (2017) *The Nature and Properties of Soils*, 15th edn Columbus: Pearson.
- Zheng N, Wang Q, and Zheng D (2007) Health risk of Hg, Pb, Cd, Zn, and Cu to the inhabitants around Huludao Zinc Plant in China via consumption of vegetables. *Science of the Total Environment* 383: 81–89. <https://doi.org/10.1016/j.scitotenv.2007.05.002>.

Methane and nitrous oxide emissions from land

Joanna M Cloy^a and Keith A Smith^b, ^aCarbon, Crop and Soils, Scotland's Rural College, Edinburgh, United Kingdom; ^bSchool of GeoSciences, University of Edinburgh, Edinburgh, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction	153
Methane	154
Sources and sinks	154
Soil sources	154
Soil sink for methane	155
Efforts to reduce methane emissions	158
Nitrous oxide	158
Soil sources	158
Nitrous oxide emissions from agricultural soils	159
Soil emissions determinants	160
Potential for reducing nitrous oxide emissions	162
Emission factors	162
Crop-based biofuels and nitrous oxide emission factors	162
Summary	163
References	163

Abstract

Accumulation of the greenhouse gases carbon dioxide (CO₂), methane (CH₄) and nitrous oxide (N₂O) in the Earth's atmosphere, resulting from human activities, is responsible for recent global heating and associated climate change. Natural soil emissions of CH₄ and N₂O are substantial, and increase after land conversion to agriculture, e.g., flooded rice fields release CH₄ and manures and synthetic nitrogen fertilizers promote the release of N₂O. Some mitigation of N₂O emission is achievable by improved nitrogen use efficiency. Replacement of fossil fuels by crop-based biofuels is controversial, because reduced CO₂ emissions can be largely offset by increases in those of N₂O.

Key points

- Emissions of the greenhouse gases methane (CH₄) and nitrous oxide (N₂O) from land contribute to global heating.
- The current concentration of CH₄ in the atmosphere is now over 2.5 times that in pre-industrial times.
- The global warming potential of CH₄ is c. 30 times that of CO₂ over a 100-year period, and >80 times greater over a 20-year period.
- Natural CH₄ sources include wetlands and termites; anthropogenic ones are principally rice fields and landfills.
- The main sinks for CH₄ are atmospheric processes.
- N₂O in the atmosphere has increased sharply over recent decades.
- Natural soils are the biggest emitter of N₂O, followed by agricultural soils.
- Increased use of N fertilizers and land conversion to agriculture are mainly responsible for increased N₂O emissions.
- N₂O emissions from crops grown for biofuels can offset much of the benefit from reduced fossil fuel use.

Introduction

Absorption of infrared radiation by natural concentrations of water vapor, carbon dioxide (CO₂), and trace gases in the atmosphere is responsible for the mean surface temperature of the Earth being approximately +15 °C rather than −18 °C. Additional heating caused by emissions of greenhouse gases (GHGs), resulting from human activities, has been superimposed on this very large, natural heating process. Although this extra heating is comparatively small (estimated to have currently raised the global mean temperature by 1.1 °C above the average for 1850–1900 (IPCC, 2021)), it is now accepted as being responsible for undesirable climate change, already manifested by extreme heatwaves and storms. Consequently, international efforts are being mobilized to limit emissions, and to restrict the maximum global temperature rise to 1.5 °C above pre-industrial levels, as embodied in the Paris Agreement signed in 2015 (UNFCCC, 2016).

The most important individual GHG is CO₂, but substantial contributions to global heating are made by trace gases, in particular methane (CH₄) and nitrous oxide (N₂O), both of which come in part from emissions from land, and which have much greater global warming potentials (GWP) than CO₂. The GWP of a GHG is a measure of its contribution to atmospheric global heating over a fixed period of time relative to that of CO₂, which is assigned a GWP of one. One kilogram of CH₄ has a GWP 28–36 times greater than 1 kg of CO₂ over a 100-year period, and 84–87 times greater over 20 years, while the GWP of 1 kg of N₂O is 265–298 times greater. A further contribution to global heating comes indirectly from tropospheric ozone, which is produced by light-catalyzed reactions involving nitric oxide (NO) and organic compounds emitted by vegetation. Nitric oxide is a natural product emitted from soils as well as from combustion processes. As well as being a source of CH₄ and N₂O, soils also act as a minor sink for CH₄ and N₂O already in the atmosphere (Smith, 2010; Reay et al., 2010).

Methane

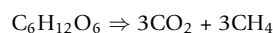
Sources and sinks

The concentration of CH₄ in the atmosphere is now well over double that in the preindustrial era, up from approximately 0.7 ppm (0.7×10^{-6} mol mol⁻¹) to almost 1.9 ppm. The lifetime of CH₄ in the atmosphere is quite short, c.10 years; the main removal pathway or sink is through oxidation by UV-created hydroxyl radicals (OH) in the atmosphere, while 5% is oxidized by soil microorganisms. No single source is responsible for the atmospheric increase, but soils (including soil fauna) contribute approximately 50% of total emissions (Table 1). Global emissions of CH₄ are estimated at 559 (range 540–568) Tg CH₄ year⁻¹ for the 2003 to 2012 decade (Saunois et al., 2016).

Soil sources

Methane is formed in soils by a process called methanogenesis, which is the microbial breakdown of organic compounds in strictly anaerobic conditions (Conrad, 1996). A very low redox potential is required for this process, and CH₄ production does not begin until reduction of molecular oxygen, nitrate, iron(III), manganese(IV) and sulfate (all of which maintain the potential at higher levels) is complete. Such low-redox conditions are predominantly found in soils where prolonged waterlogging is a normal feature, e.g., natural wetlands and flooded rice fields, and beneath the soil cover of closed landfills, as well as in the sediments at the bottoms of lakes. The original identification of CH₄ as a main component of the gas released as bubbles from natural marshes was made by Alessandro Volta in Italy in 1776. The gas is still known by the name ‘marsh gas’ given to it at that time.

The organic compounds that provide the carbon for CH₄ formation in marshes, bogs, and flooded soils are released from living roots or come from the microbial degradation of plant remains in a bogland plant community, or rice straw ploughed into the soil of a paddy field prior to the establishment of the next crop. The final step in the decomposition is performed by methanogens, i.e., methane-producing archaea (Stams and Plugge, 2010), giving rise to CH₄ and CO₂ in equimolar quantities, according to the equation:



The CH₄ formed in this way can migrate to the surface and be emitted into the atmosphere by one of three different pathways (Fig. 1). Diffusion can take place in solution from the point of formation in an anaerobic layer, upward to water layers containing

Table 1 Estimated soil sources of CH₄ emissions and soil and atmospheric sinks (taken from IPCC, 2013).

Sources	Global annual emission/sink size (Tg CH ₄ year ⁻¹)
<i>Natural</i>	
Wetlands	217
Termites	11
Permafrost	1
<i>Anthropogenic</i>	
Rice paddies	36
Landfills	75
Total from soils/soil fauna	340
Total from all sources	678
<i>Sinks</i>	
Soils	–28
Tropospheric OH	–528
Stratospheric OH	–51
Tropospheric Cl	–25

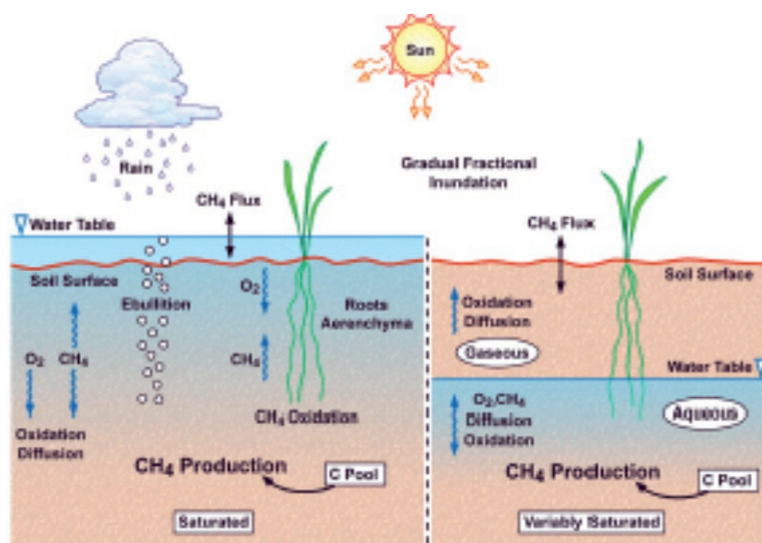


Fig. 1 Schematic representation of biological and physical processes involved in methane cycling between terrestrial ecosystems and the atmosphere. Left: fully inundated conditions; right: variably saturated conditions. Reproduced, with permission, from Riley WJ, et al. (2011) Barriers to predicting changes in global terrestrial methane fluxes: analyses using CLM4Me, a methane biogeochemistry model integrated in CESM. *Biogeosciences* 8: 1925–1953.

oxygen, where much of the CH₄ is oxidized and only a fraction outgasses to the atmosphere. When sufficient gas is produced to exceed the solubility product, bubbles form in the water layer, and force their way to the surface—the process of ebullition that prompted Volta's original investigation. The speed of this process prevents any significant oxidation. The third route is via the internal structures of vascular plants. Many species that are adapted to life in flooded environments contain continuous air spaces that connect the leaves and stems to the roots. These structures, known as aerenchyma (Fig. 2), have evolved to transport oxygen needed for cell division to the root meristems, but serve equally well as channels for the transport of CH₄ from the root environment to the atmosphere.

In natural wetlands, CH₄ emissions vary with temperature, depth to the water table and the type and amount of vegetation present (Turetsky et al., 2014). Fig. 3 shows the steep increase in CH₄ emissions from boreal bogs and fens with increasing temperature and decreasing depth to the water table. Although many investigations have been made into emissions from temperate and boreal wetlands, the degree of uncertainty about total emissions, including the additional methane emitted as Arctic permafrost thaws, is still large. The uncertainty is greater still over emissions from tropical wetlands, because fewer studies have been made, and there is thought to be a much bigger contribution from seasonally flooded environments (e.g., Ringeval et al., 2014). Tropical river basins such as those of the Amazon and Paraná in South America, and the Congo in Africa, are prime examples of such seasonal flood plains, in which there is typically a rapid development of lush vegetation with the advent of the water, followed by decomposition under the floodwater and consequent CH₄ production. However, the flooded areas vary considerably in extent and mean flooding depth from year to year.

In the absence of adequate experimental data, mathematical modeling of regional emissions has made an important contribution to the assessment of wetland emissions. Observed seasonal variations in the concentrations of CH₄ in the atmosphere at different latitudes can now be simulated reasonably well by such models. Generally these models predict that more than 50% of global wetland emissions are of tropical origin.

Rice cultivation is also a major CH₄ source. Estimates compiled by the Intergovernmental Panel on Climate Change (2013) put the annual global emission from rice fields at approximately 36 Tg CH₄ year⁻¹ (Table 1). The area devoted to the rice crop has expanded considerably in recent decades, and global production is predicted to rise still further in the future, to meet the needs of rising populations. This trend has undoubtedly contributed to the increase in overall CH₄ emissions observed in the last few decades. An international program of research has been carried out to investigate the principal factors that determine the magnitude of emissions from rice. This has shown that soil type, crop fertilization, the amount of straw or organic manure incorporated into the soil before planting, and the continuity of the flooding regime, as well as climate, are all important.

In warm climates, CH₄ production also occurs in some aerobic soils, as well as in wetlands, but here the source is the microorganisms that live symbiotically in the guts of some species of burrowing and mound-forming termites, and not the free-living soil microflora (Bignell, 2010).

Soil sink for methane

The phenomenon of CH₄ oxidation by soil bacteria is general in aerated soils (Fig. 1). In well-drained and therefore well-aerated soils, in the absence of termites, any CH₄ present is largely that which has diffused in from the atmosphere, where the concentration

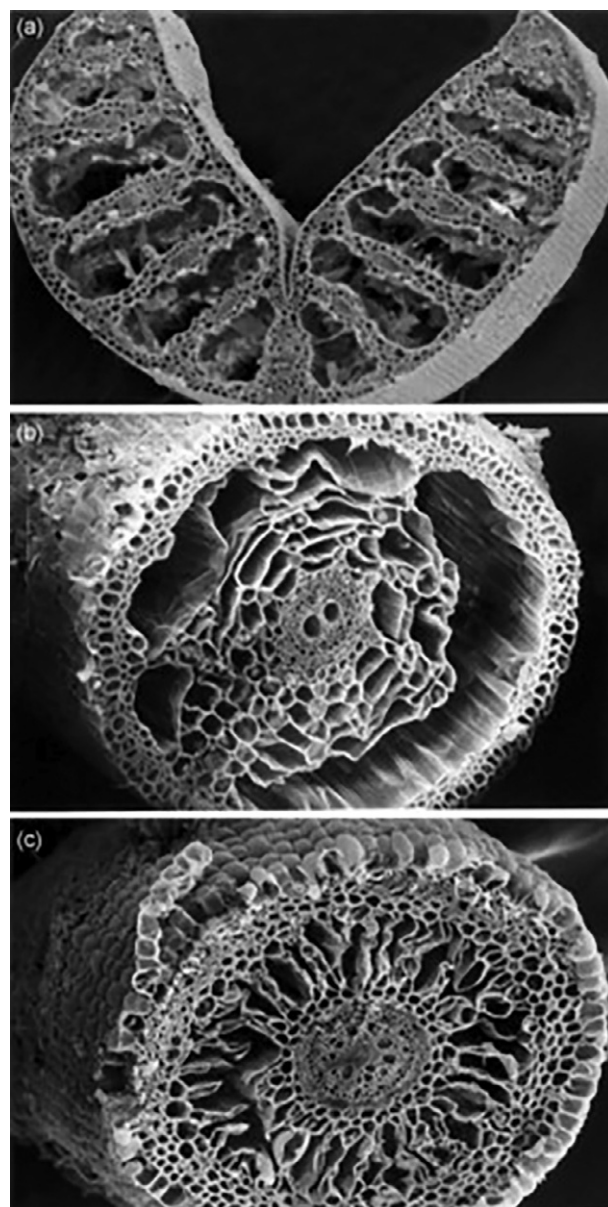


Fig. 2 Scanning electron micrographs of transverse sections through the vegetative organs of some vascular wetland plants, showing the aerenchyma through which oxygen and methane can readily diffuse: (A) *Eriophorum angustifolium* leaf; (B) *Carex echinata* root (at a depth of 100–110 mm); and (C) *Molinia caerulea* root (at a depth of c. 150 mm). Reproduced with permission from Thomas KL, et al. (1996) Role of wetland plants in the diurnal control of CH₄ and CO₂ fluxes in peat. *Soil Biology and Biochemistry* 28: 17–23.

is approximately 1.9 ppm. This CH₄ is oxidized to CO₂ by methanotrophic bacteria that are adapted to living on this very low concentration of substrate. The overall global sink for atmospheric CH₄ in soils is estimated at approximately 28 Tg year⁻¹; this is about 4% of total annual emissions from all sources (Table 1).

Soil bulk density and water content, and their consequent effects on gas movement and penetration in the soil profile, have a major impact on the rate of oxidation of atmospheric CH₄ in soils. Rates are highest in coarse-textured forest soils with well-developed soil structure and a permeable surface organic layer. Increasing soil acidity reduces oxidation; part of the explanation is that in acid conditions plant litter decomposes less readily and accumulates at the soil surface. This litter, particularly the leaves of broadleaved deciduous trees, acts as a diffusion barrier to gas exchange between the soil and the atmosphere. The effect of soil temperature on oxidation rate is small, and this is attributed to limitations imposed by the supply of available CH₄ substrate, due to the combined effects of diffusion resistance and low atmospheric concentration.

In addition to the oxidation of atmospheric CH₄ in soils, there is also oxidation of a fraction (perhaps 20–30%) of the CH₄ produced in soils before it escapes to the atmosphere. The methanotrophic bacteria responsible live in, e.g., the thin, aerated

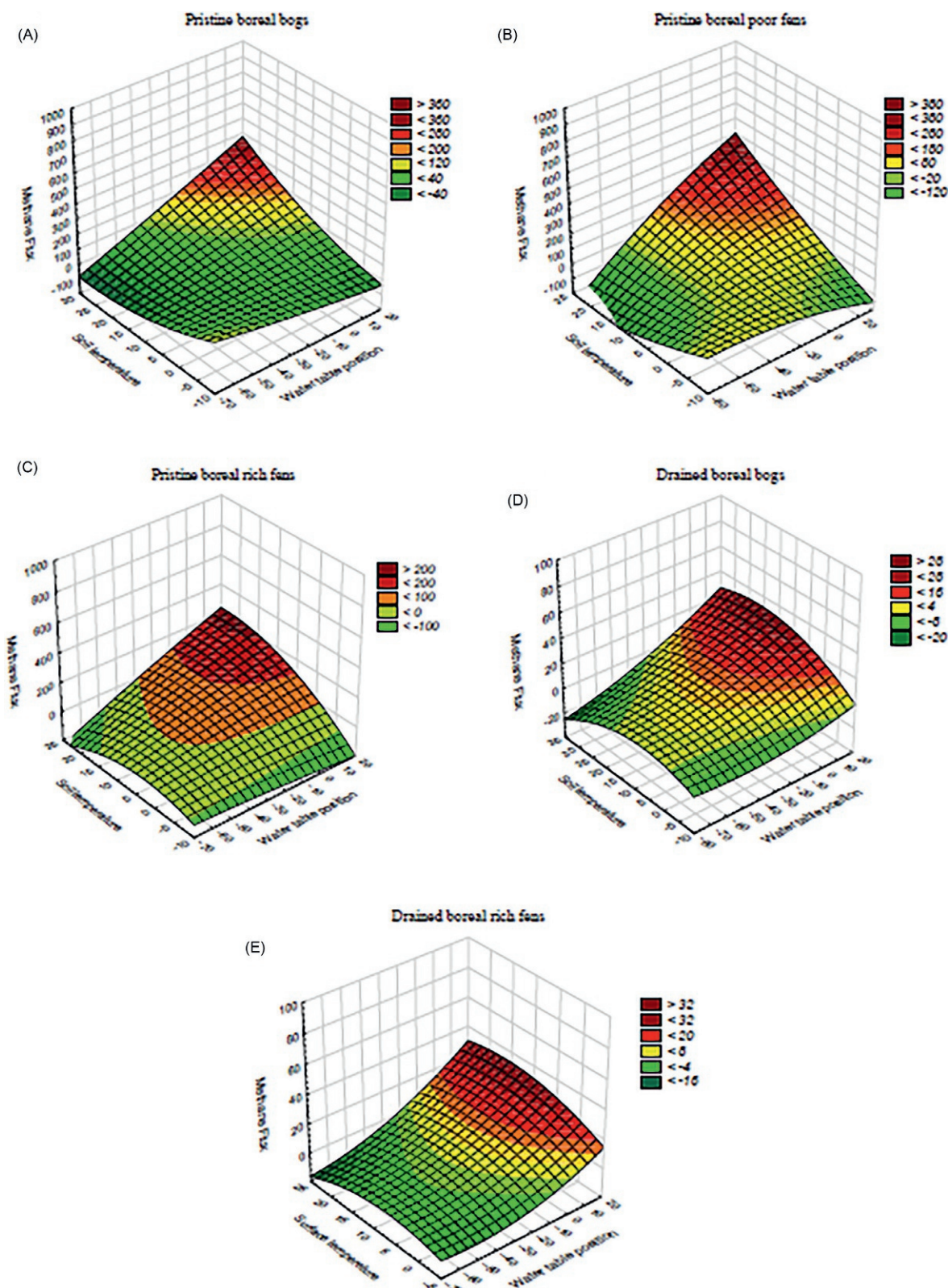


Fig. 3 Interactions between soil temperature and water table govern methane flux. Surface plots of relationships between observed instantaneous methane flux, soil temperature and water table position across (A) pristine boreal bogs, (B) pristine boreal poor fens, (C) pristine boreal rich fens, (D) drained bogs, (E) drained rich fens. Reproduced, with permission, from Turetsky MR, et al. (2014) A synthesis of methane emissions from 71 northern, temperate, and subtropical wetlands, *Global Change Biology* 20: 2183–2197.

sheath of soil that exists around the roots of vascular plants growing in flooded conditions, and in intercellular spaces within plant root systems. They also occur in aerobic hummocks on the tops of bogs, as well as in aerated cover soils above anaerobic refuse layers in landfills. In such environments, average CH₄ concentrations may be very much higher than those resulting from gas diffusion from the atmosphere, and different mixed populations of bacteria, capable of assimilating CH₄ at a faster rate than those adapted to atmospheric concentrations, become established.

As a result of deleterious effects of soil disturbance and additions of nitrogenous fertilizers on methanotrophic organisms, conversion of natural soils to agricultural use reduces CH₄ oxidation rates by approximately two-thirds. Recovery of oxidation rate from land-use change or fertilization is slow; it may take 100 years or more to return to pre-disturbance rates. The reasons for this unusually long period of adjustment may be related to slow-changing soil properties such as soil organic matter composition or microbial community structure (Suwanwaree and Robertson, 2005).

Efforts to reduce methane emissions

Methane is a good target for mitigation because it has a short atmospheric lifetime (c. 10 years) and the benefits of reduced emissions would be apparent over a relatively short time (Cloy et al., 2012). Attempts are being made to reduce net CH₄ emissions from two major sources: sanitary landfills and flooded rice fields. Paradoxically, the creation of closed landfills for municipal wastes to prevent other problems associated with unconfined piles of refuse has brought about a major increase in the amounts of CH₄ produced. This is because compacted clay layers in the soil caps exclude atmospheric oxygen, thus allowing the landfills to become anaerobic. However, the general recognition of the undesirable contribution of these sites to global heating has led to the development of effective engineering solutions for application to large modern landfills. These range from the venting and burning on-site of the CH₄ (converting it to CO₂, which on a 20-year timescale is less than one-eightieth as potent a GHG), to its collection and use in energy production, thus gaining an additional benefit through reducing the use of fossil fuel for that purpose. In both the USA and the EU, these measures have brought about a reduction in landfill CH₄ emissions (though in developing countries increased landfilling will likely result in increased emissions) (Bogner and Spokas, 2010). For many older and smaller landfills, however, there is little scope for economic installation of CH₄ collection and recovery equipment. For such environments the more effective option is the promotion of CH₄ oxidation in the cover soil. The oxidation is optimal in cover soils of intermediate textures and moderate water contents, which permit a reasonable rate of diffusion through the soil matrix and thus allow oxidation to take place, but avoid the development of cracks through which gas may be vented rapidly to the surface.

There is potential for reducing substantially the average CH₄ emissions from traditionally flooded paddy fields. In recent years there has been a move away from permanently flooded production systems, to more productive cultivation systems involving periodic flooding or dryland production (Conen et al., 2010). Water management that involves multiple cycles of flooding and draining of the soil has been shown to reduce CH₄ emissions by 40%. Reduced inputs of crop residues prior to flooding can also reduce emissions. Such changes to traditional methods can be made without detriment to crop yields and sometimes bring with them other agronomic or cost benefits—at worst they can usually be introduced without increasing the cost or difficulty of farming operations. However, persuading farmers to use such new methods takes much advisory effort, and the rate of adoption varies greatly among different societies, unless incentives are introduced.

Nitrous oxide

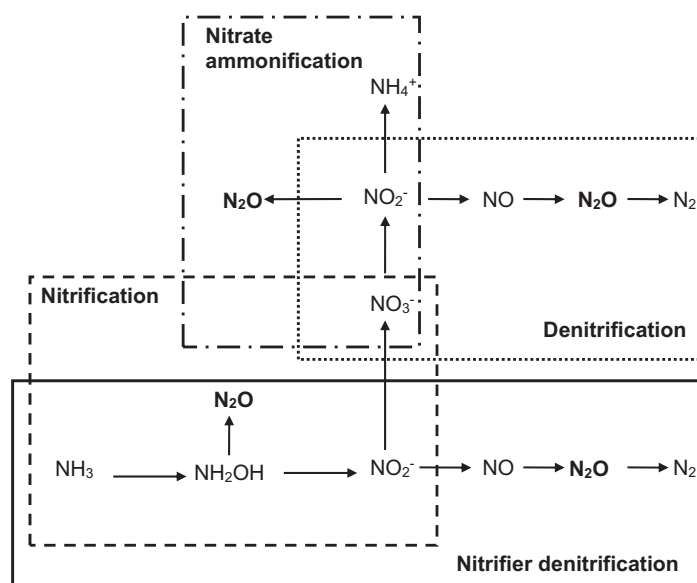
Soil sources

The natural cycle of N₂O in the environment involves microbial and chemical production in soils and waters, and ultimate decomposition in the stratosphere. Approximately 65% of all atmospheric emissions of N₂O are from soils. Ice-core data show that the concentration of N₂O in the atmosphere in the preindustrial era was approximately 275 ppb ($275 \times 10^{-9} \text{ mol mol}^{-1}$), but this had increased to about 331 ppb in 2018 (Tian et al., 2020). This increase in atmospheric N₂O is attributed to increased anthropogenic emissions, through increased production and use of N fertilizers, tropical land conversion from forest to agriculture, increased biomass burning, etc. Recent modeling also suggests an increased emission from natural soils, as a result of global N deposition and climate change (Tian et al., 2020). The extra emissions have not been accompanied by any corresponding increases in the sinks for N₂O. The dominant sink is the reaction with ozone in the stratosphere to produce NO, which was the process that first generated the environmental concern about N₂O because of its potential for increasing UV radiation at the Earth's surface (Table 2). In addition, a very small soil sink (uptake of N₂O in soil) has been suggested, and attributed to a reduction of overlying atmospheric N₂O to N₂ during denitrification, the only proven terrestrial biological N₂O sink. However, there is no direct supporting experimental evidence (Smith, 2010).

Nitrous oxide production in soil results predominantly from a range of microbial processes: nitrification, denitrification, nitrifier denitrification and nitrate ammonification (Conrad, 1996). Their pathways are shown in Fig. 4; those generally regarded as most important for N₂O production are autotrophic nitrification, involving ammonia-oxidizing *Nitrosomonas* and *Nitrosospira* bacteria, and heterotrophic denitrification, brought about by a wide range of bacteria, together with some archaea and fungi (Baggs and Phillipot, 2010). Non-biological formation of N₂O, chemodenitrification, can also occur in soil, either through the reaction of

Table 2 Estimated soil sources of N₂O emissions and stratospheric sink (Tg N₂O-N year⁻¹) (taken from IPCC, 2013).

Sources	Global annual emission	Range
<i>Natural</i>		
Soils under natural vegetation	6.6	3.3–9.0
<i>Anthropogenic</i>		
Agriculture (direct emissions from cultivated/fertilized soils and animal production)	4.1	1.7–4.8
Atmospheric N deposition on land	0.4	0.3–0.9
Total from soils/agriculture	11.1	5.3–14.7
Total from all sources	17.9	
Observed atmospheric increase	3.6	3.5–3.8
Stratospheric sink	14.3	4.3–27.2

**Fig. 4** Soil microbial processes leading to N₂O production. In: Smith KA (2017) Changing views of nitrous oxide emissions from agricultural soils: Key controlling processes and assessment at different spatial scales. *European Journal of Soil Science* 68: 137–155.

nitrite with the constituents of organic matter, or by the reduction of nitrate by iron (II). The available evidence suggests that the former pathway constitutes a very small source compared with the microbial ones, but the latter could be more significant.

There are strong relationships between the state of aeration of soils and the balance between nitrification and denitrification as the main sources of N₂O emissions. As soil wetness increases, the availability of oxygen from the atmosphere decreases, and nitrification is inhibited. Conversely, denitrification, which is very variable in time and space, is generally promoted. Both processes can occur in the soil in close proximity to each other (Conrad, 1996).

Natural soil N₂O emissions per unit area in the tropics are about 50% higher than the global average, because many lowland, highly weathered tropical soils have excess nitrogen relative to phosphorus (Davidson et al., 2007; Tian et al., 2020). These tropical emissions have been estimated to total 3 Tg N₂O-N year⁻¹ while those from temperate forests, temperate grasslands and tropical dry savannas are each thought to contribute in the order of 1 Tg year⁻¹ (Kroeze et al., 1999). A global total of 6.6 Tg N₂O-N year⁻¹ from all natural soils has been estimated by the IPCC (Table 2). The uncertainty associated with this estimate is in the order of ±50%. Global long-term average soil N₂O emissions from croplands were estimated at 3.6 Tg N₂O-N year⁻¹ (ranging from 2.96–4.35 Tg N₂O-N year⁻¹) from 1996 to 2013 (Yang et al., 2021).

Nitrous oxide emissions from agricultural soils

Reactive N compounds are introduced into the natural terrestrial environment primarily by biological nitrogen fixation (BNF) in forests and grasslands, particularly in the tropics. The N input by this pathway in the pre-industrial era (up to 1860) was estimated by Galloway et al. (2004) at 120 Tg N year⁻¹ in 1860, but human activity has now roughly doubled the N supply to the biosphere

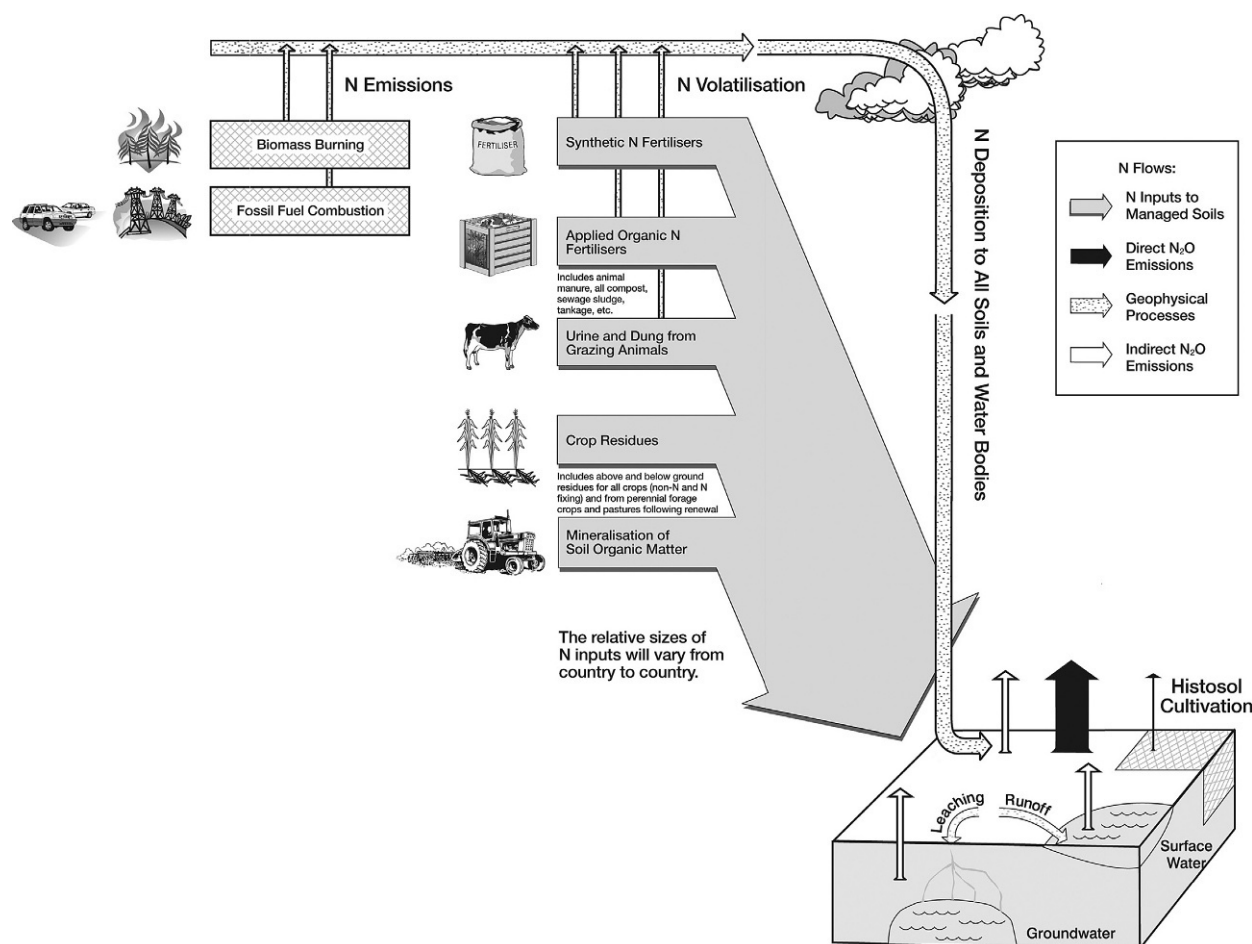


Fig. 5 Sources and pathways of reactive nitrogen compounds that result in N_2O emissions from soils and waters. From de Klein C, et al. (2006) N_2O emissions from managed soils and CO_2 emissions from lime and urea application. In Eggleston HS, Buendia L, Miwa K, Ngara Y, and Tanabe K. (eds.), *2006 IPCC Guidelines for National Greenhouse Gas Inventories*, Vol. 4, Chapter 11, Hayama, Japan: IGES.

through synthetic N fertilizers—the worldwide demand for 2020 was 109 Tg N (FAO, 2019)—together with a smaller amount through BNF by leguminous crops, and also inadvertently as NO_x from fossil fuel combustion. Consequently, human-induced emissions of N_2O are dominated by N additions to land, and have increased by 30% over the past four decades to 7.3 (4.2–11.4) Tg N year⁻¹ (Tian et al., 2020).

Fig. 5 shows schematically the sources and pathways of reactive N compounds that result in N_2O emissions from soil and water. As the Figure shows, agricultural soil emissions are not only from land receiving mineral N fertilizers, but also from land receiving animal manures, either applied by farmers or directly deposited as the excreta of grazing livestock. The steady increase in the size of livestock populations over time has contributed substantially to the present-day emissions level. Increases in food production are regarded as essential to meet the needs of a growing population, over the next few decades, and are likely to require increased use of N, both for crops intended for direct human consumption, and for animal feed crops; this trend may be expected to increase overall N_2O emissions.

Soil emissions determinants

Fluctuations in soil physical conditions and mineral N content in agricultural soil cause large variations in N_2O emissions. Consequently, rainfall events and applications of fertilizer that cause step-changes in soil wetness and mineral N content, respectively, are often associated with emission pulses, as shown in Fig. 6. The largest fluxes generally occur when soil water-filled pore space (WFPS) values are high, indicating that denitrification is the major process responsible.

On grazed land, the deposition of excreted N is spatially very variable. A small patch of soil surface may receive N as urine at a rate equivalent to some hundreds of kilograms per hectare, whereas land between patches receives none. Also, treading of the soil surface in wet conditions by animal hooves can create localized compacted zones in which water collects and the soil becomes

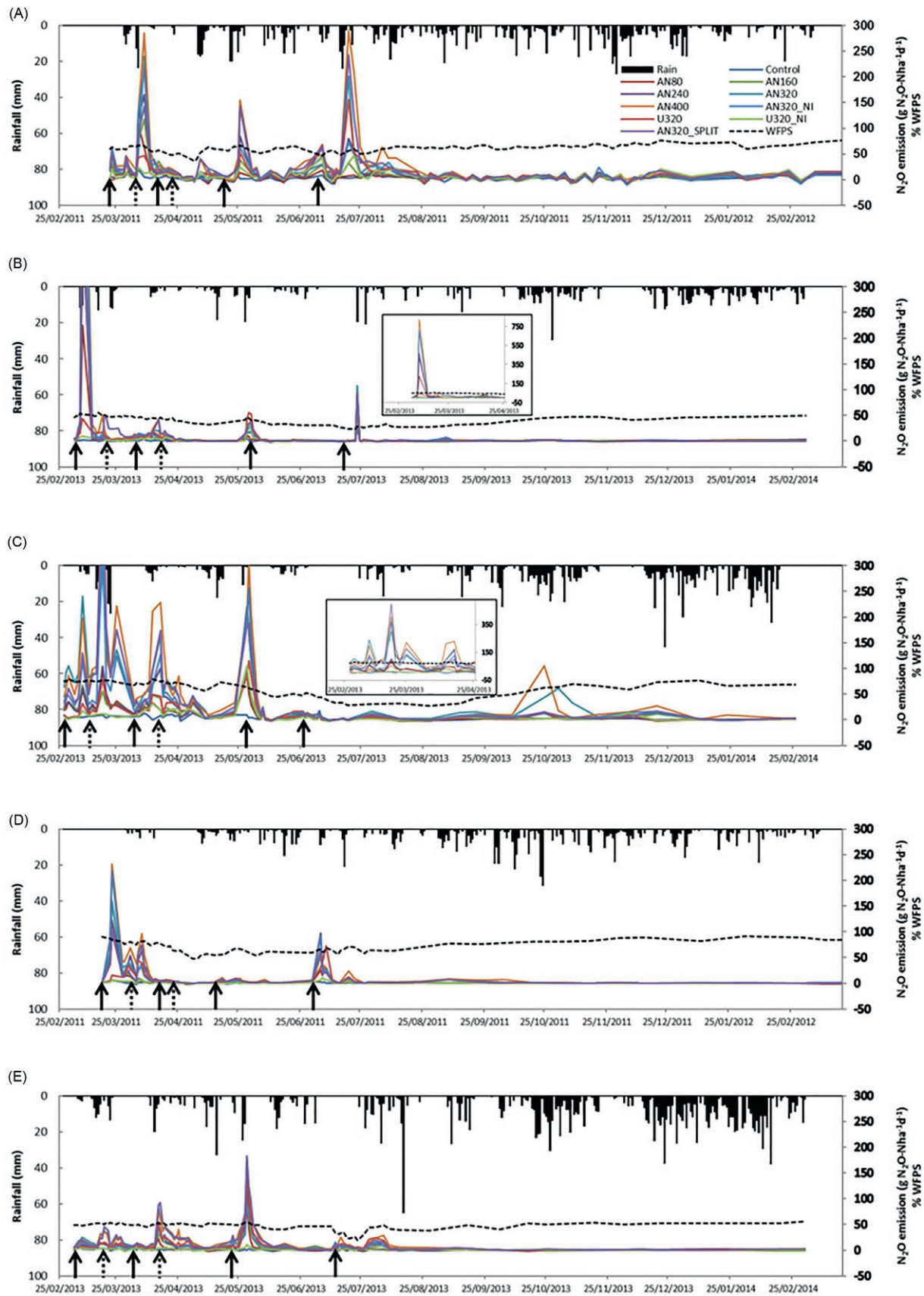


Fig. 6 Nitrous oxide emissions from grassland at 5 UK sites, fertilized with different forms of nitrogen fertilizer: ammonium nitrate (or calcium ammonium nitrate) and urea, with and without the nitrification inhibitor dicyandiamide. The figure shows variation in daily precipitation and soil water-filled pore space (WFPS) throughout the experimental period, and mean N_2O emissions from all N treatments. Continuous arrows show dates of main N fertilizer applications, and dashed arrows show dates of extra fertilizer additions in split application treatments. From Cardenas LM, et al. (2019) Nitrogen use efficiency and nitrous oxide emissions from five UK fertilised grasslands. *Science of the Total Environment* 661: 696–710.

anaerobic. Such factors lead to the occurrence of ‘hotspots’ of microbial activity and N_2O emission; overall the average emission rate from grazed grassland may be two to three times that from N-fertilized grass grown as a crop to be cut for winter feed. In Western Europe at least, the emissions from cut grassland are, in turn, substantially higher than emission rates from arable land cropped with wheat and barley. This is attributable to a combination of soil physical conditions and the timing of N applications. However, the emissions are very variable between sites and between successive seasons, and for any crop type several years’ data are required to obtain a robust estimate of annual N_2O flux (Dobbie and Smith, 2003).

Potential for reducing nitrous oxide emissions

With food security an issue (see Chapter on Soil and Human Health), reducing production is not a viable option. Instead, ‘sustainable intensification’ practices that aim to produce ‘more with less,’ e.g., improving N-use efficiency (NUE), offer the best opportunity of reducing emissions from agricultural soils. In general, the procedures recognized as helping in this direction include:

- Avoiding excessive inputs of N and optimizing plant N through soil/plant testing to determine fertilizer N needs;
- Minimizing fallow periods to limit mineral N accumulation;
- Improving the timing of and or splitting of N applications;
- Using ‘controlled release’ N fertilizers, which make the N available for crop uptake over an extended period;
- Adding urease inhibitors and nitrification inhibitors to urea and ammonium fertilizers;
- Matching the fertilizer type to the soil water regime, e.g., by using an ammonium form in circumstances where wet soil conditions are likely to cause excessive denitrification losses from nitrate forms.

The widespread adoption of these measures could reduce total emissions by 20–25%.

Emission factors

National emissions of GHGs are generally reported using Intergovernmental Panel on Climate Change (IPCC) methodology. For each significant source, an emission factor (EF) is multiplied by the size of the source: the so-called ‘activity data’. Thus, e.g., direct soil emissions of N_2O are calculated by multiplying the fraction of the applied N that is emitted as N_2O , by the total mass of N applied. ‘Default values’ for EF are used where there are no measurement data for the country involved that meet the necessary quality criteria: 1% for direct soil emissions from applied fertilizer and manure N, 2% for N deposited by grazing livestock, 0.3–0.4% for indirect emissions following N loss by leaching, run-off or volatilization (IPCC, 2006; Cloy et al., 2012). Using an ensemble of process-based models, Tian et al. (2020) estimated a global agricultural soil EF of 1.8% (1.3–2.3%), which is considerably larger than the IPCC Tier 1 default for direct emission of 1%, and suggest that this is caused by a strong interactive effect between N additions and other global environmental changes. They calculated that anthropogenic N_2O emissions increased from 5.6 (3.6–8.7) Tg N year^{-1} in the 1980s to 7.3 (4.2–11.4) Tg N year^{-1} in 2007–16, at a rate of 0.6 ± 0.2 Tg N year^{-1} per decade ($P < 0.05$). Up to 87% of this increase (71% directly and 16% indirectly) resulted from N additions to soils. The bottom-up land emissions estimate during 1998–2016 was on average 1.8 Tg N year^{-1} higher than the top-down estimate, but showed a slightly slower rate of increase of 0.8 ± 0.2 Tg N year^{-1} per decade (95% confidence interval; $P < 0.05$) compared with 1.1 ± 0.6 Tg N year^{-1} per decade ($P < 0.05$) from the top-down approach.

An alternative ‘top-down’ approach took the reactive N (Nr) entering agricultural systems as the source of all agriculture-related N_2O emissions. On the basis of the observed rate of increase in the atmospheric concentration of N_2O , this gave a global EF of 3–5% for the year 2000 (Crutzen et al., 2008). Furthermore, including two other Nr sources—the N mineralized as a result of land-use change and N deposited as NO_x —and applying an overall EF of 4% shows a very close match with the observed upward trend in the atmospheric concentration of N_2O over the 140 years between 1860 and 2000 (Smith et al., 2012), as shown in Fig. 7.

Crop-based biofuels and nitrous oxide emission factors

The agricultural sector is faced with a growing demand for food and the consequent need to increase food supply. However, the surge in the production of crops for biofuels has also increased the pressure on land resources. Life cycle analysis (LCA) models generally indicate that crop-based biofuels generate GHG savings compared with fossil fuels. However, applying a 4% global EF for N_2O , and a global mean value for nitrogen use efficiency (NUE) of 40–50% can cancel out these GHG savings (Crutzen et al., 2008). Use of these global figures is inevitably controversial. There are circumstances in which sophisticated farm management results in a much higher NUE than the global average, which would correspondingly diminish the N_2O emission per unit of crop-based biofuel. However, given the absence of any controls on production methods for biofuel feedstocks, there is a case for the precautionary use of the global mean values, when assessing the potential environmental benefits from switching from fossil fuels to biofuels. Overall, it appears that deriving biofuels from biomass, from waste feedstocks, or high-yielding bioenergy crops with low N demand grown on currently unproductive land, rather than from N-fertilized crops, would be the best option for GHG mitigation (Smith and Searchinger, 2012).

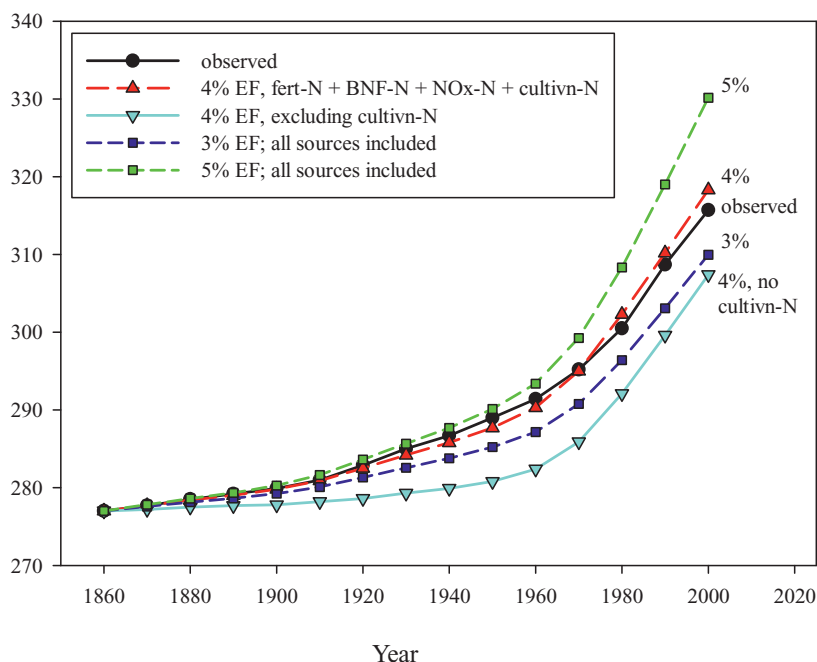


Fig. 7 Comparison between observed increase in atmospheric concentration of N_2O over the period 1860–2000 (black circles) and calculated increase based on a 4% emission factor (EF) for new reactive N, including N from synthetic fertilizers, from additional N from biological N fixation (BNF) above the rate of fixation in 1860, from NO_x -N, and from N mineralized from soil organic matter as a result of land-use change to agriculture (red triangles). Bottom curve (inverted triangles) includes all these sources except mineralized N. Blue and green squares indicate the increase for EFs of 3 and 5%, respectively. Based on data in Davidson EA (2009) The contribution of manure and fertilizer nitrogen to atmospheric nitrous oxide since 1860. *Nature Geosciences* 2: 659–662; and Butterbach-Bahl K, et al. (2011) Nitrogen as a threat to the European greenhouse balance. In Sutton MA, et al. (eds.) *The European Nitrogen Assessment*, pp. 434–462. Cambridge, UK: Cambridge University Press. Adapted from Smith KA, et al. (2012) The role of N_2O derived from biofuels, and from agriculture in general, in Earth's climate. *Philosophical Transactions of the Royal Society, Series B* 367: 1169–1174.

Summary

Emissions of the greenhouse gases carbon dioxide (CO_2), methane (CH_4) and nitrous oxide (N_2O) are responsible for global heating, resulting from human activities. Methane and N_2O are more potent greenhouse gases than CO_2 , and their emissions from land are significant. Natural CH_4 sources include wetlands and termites; anthropogenic ones are principally rice fields and landfills. The main sinks for CH_4 are atmospheric processes, with only a small soil sink. Natural soils are the biggest terrestrial emitter of N_2O , followed by agricultural soils. Increased use of N fertilizers, land conversion to agriculture, and climate change, are mainly responsible for increased N_2O emissions over recent decades. The N_2O from crops grown for biofuels can offset much of the benefit from reduced fossil fuel use.

References

- Baggs E and Phillipot L (2010) Microbial terrestrial pathways to nitrous oxide. In: Smith KA (ed.) *Nitrous Oxide and Climate Change*, pp. 4–35. London: Earthscan.
- Bignell DE (2010) Termites. In: Reay D, Smith P, and van Amstel A (eds.) *Methane and Climate Change*, pp. 62–73. London: Earthscan.
- Bogner JE and Spokas K (2010) Landfills. In: Reay D, Smith P, and van Amstel A (eds.) *Methane and Climate Change*, pp. 175–200. London: Earthscan.
- Cloy JM, Rees RM, Smith KA, Goulding KWT, Smith P, Waterhouse A, and Chadwick D (2012) Impacts of agriculture upon greenhouse gas budgets. In: Hester RE and Harrison RM (eds.) *Environmental Impacts of Modern Agriculture*, pp. 57–82. Cambridge, UK: Royal Society of Chemistry. Issues in environmental science and technology. Issue 34.
- Conen F, Smith KA, and Yagi K (2010) Rice cultivation. In: Reay D, Smith P, and van Amstel A (eds.) *Methane and Climate Change*, pp. 115–135. London: Earthscan.
- Conrad R (1996) Soil microorganisms as controllers of atmospheric trace gases (H_2 , CO, CH_4 , OCS, N_2O , and NO). *Microbiological Reviews* 60: 609–640.
- Crutzen PJ, Mosier AR, Smith KA, and Winiwarter W (2008) N_2O release from agro-biofuel production negates global warming reduction by replacing fossil fuels. *Atmospheric Chemistry and Physics* 8: 389–395.
- Davidson EA, et al. (2007) Recuperation of nitrogen cycling in Amazonian forests following agricultural abandonment. *Nature* 447: 995–998.
- Dobbie KE and Smith KA (2003) Nitrous oxide emission factors for agricultural soils in Great Britain: The impact of soil water-filled pore space and other controlling variables. *Global Change Biology* 9: 204–218.
- Food and Agriculture Organization of the United Nations (FAO) (2019) *World Fertilizer Trends and Outlook to 2022*. Rome: FAO.
- Galloway JN, Dentener FJ, Capone DG, et al. (2004) Nitrogen cycles: Past, present, and future. *Biogeochemistry* 70: 153–226.
- IPCC (Intergovernmental Panel on Climate Change) (2006) In: Eggleston S, Buendia L, Miwa K, Ngara T, and Tanabe K (eds.) *IPCC Guidelines for National Greenhouse Gas Inventories*. Hayama, Japan: IGES.

- IPCC (Intergovernmental Panel on Climate Change) (2013) *Climate change 2013: The physical science basis*. Fifth Assessment Report. Working Group I Cambridge, UK: Cambridge University Press.
- IPCC (Intergovernmental Panel on Climate Change) (2021) *Climate change 2021: The physical science basis*. Sixth Assessment Report. Working Group I Cambridge, UK: Cambridge University Press.
- Kroeze C, Mosier A, and Bouwman L (1999) Closing the global N_2O budget: A retrospective analysis 1500–1994. *Global Biogeochemical Cycles* 13: 1–8.
- Reay D, Smith P, and van Amstel A (eds.) (2010) *Methane and Climate Change*. London: Earthscan.
- Ringeval B, Houweling S, van Bodegom PM, Spahni R, van Beek R, Joos F, and Röckmann T (2014) Methane emissions from floodplains in the Amazon Basin: challenges in developing a process-based model for global applications. *Biogeosciences* 11: 1519–1558.
- Saunio M, Bousquet P, Poulter B, et al. (2016) The global methane budget: 2000–2012. *Earth System Science Data* 8: 697–751.
- Smith KA (ed.) (2010) *Nitrous Oxide and Climate Change*. London: Earthscan.
- Smith KA and Searchinger TD (2012) Crop-based biofuels and environmental concerns. *Global Change Biology. Bioenergy* 4: 479–484.
- Smith KA, Mosier AR, Crutzen PJ, and Winiwarter W (2012) The role of N_2O derived from crop-based biofuels, and from agriculture in general, in Earth's climate. *Philosophical Transactions of the Royal Society, B: Biological Sciences* 367: 1169–1174.
- Stams AJM and Plugge CM (2010) The microbiology of methanogenesis. In: Reay D, Smith P, and van Amstel A (eds.) *Methane and Climate Change*, pp. 14–26. London: Earthscan.
- Suwanwaree P and Robertson GP (2005) Methane oxidation in forest, successional, and no-till agricultural ecosystems: Effects of nitrogen and soil disturbance. *Soil Science Society of America Journal* 69: 1722–1729.
- Tian H, Xu R, Canadell JG, et al. (2020) A comprehensive quantification of global nitrous oxide sources and sinks. *Nature* 856: 250–256.
- United Nations Framework Convention on Climate Change (UNFCCC) (2016) *The Paris Agreement*. New York: United Nations.
- Turetsky MR, Kotowska A, Bubier J, et al. (2014) A synthesis of methane emissions from 71 northern, temperate, and subtropical wetlands. *Global Change Biology* 20: 2183–2197.
- Yang Y, Liu L, Zhang F, Zu W, Liu X, Wang Z, and Xie Y (2021) Soil nitrous oxide emissions by atmospheric nitrogen deposition over global agricultural systems. *Environmental Science & Technology* 55: 4420–4429.

Relevant websites

www.ipcc.ch—The Intergovernmental Panel on Climate Change.

<https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions>—US Environmental Protection Agency

Overview chapter on soil degradation

K Müller^a, Margaret A Oliver^b, and C Siebe^c, ^aSustainable Production, The New Zealand Institute for Plant and Food Research Limited, Auckland, New Zealand; ^bSoil Research Centre, Department of Geography & Environmental Science, School of Human & Environmental Sciences, The University of Reading, Reading, United Kingdom; ^cInstituto de Geología, Universidad Nacional Autónoma de México, Ciudad de México, México

© 2023 Elsevier Ltd. All rights reserved.

Introduction	165
Factors controlling soil degradation	166
Types of soil degradation	167
Physical soil degradation	167
Chemical soil degradation	168
Biological soil degradation	169
Soil remediation	170
References	170

Abstract

Globally, soils are being degraded at an unprecedented rate. This chapter provides an overview on the extent of soil degradation and its main anthropogenic controlling factors. Most controlling factors are related to land mismanagement or overexploitation to satisfy the increasing demands of the growing global population. Positive feedback loops between climate change, land-use changes and socio-political pressures exacerbate soil degradation. This chapter also briefly introduces the chapters that focus on soil quality and monitoring, physical and chemical soil degradation processes, for example, erosion, pollution and salination, and biological soil degradation processes. Novel remediation approaches for contaminated soils are summarized, which aim to restore degraded lands to safe and beneficial reuse.

Key points

- Soil degradation is a major global problem
- It is caused by anthropogenic and natural factors
- Physical soil degradation processes include wind and water erosion, and the loss of soil structure
- Chemical soil degradation processes are mostly related to anthropogenic activities and include salination, overfertilization, and pollution with trace elements, microplastics and organic chemicals
- Major biological soil degradation processes are losses of soil organic matter and biodiversity. The functioning of soil biodiversity and community composition are still poorly understood. The relevance of biological soil degradation, however, is more and more recognized

Introduction

Soils are degrading worldwide at an unprecedented rate mainly resulting from unsustainable human activities (FAO and ITPS, 2015). This chapter provides an overview of chapters in the Encyclopedia that deal with soil quality and monitoring, soil degradation, and the remediation of contaminated soils.

Soil, the skin of the earth, plays a fundamental role for humanity. In concepts integrating the economic worldview with ecological principles, soil is considered as natural capital stock. Soils deliver many essential ecosystem services such as provisioning services, regulating and supporting services (Costanza et al., 1997). Soils provide humans with 98.8% of our food (Kopittke et al., 2019) and are essential for achieving many of the Sustainable Development Goals (SDGs) including zero hunger, clean water and flood regulation (Keesstra et al., 2016). The total value of ecosystem services provided by soils is estimated at US\$ 11.4 trillion (McBratney et al., 2017). Soil is considered a non-renewable resource in the human lifetime because its formation takes centuries to millennia. It is a living ecosystem, and its functioning including, for example, nutrient cycling, water infiltration and retention, and contaminant filtering, is affected by natural forces and anthropogenic activities.

Soil functioning and the delivery of ecosystem services depend on soil quality. There is a growing body of literature on defining soil quality, identifying its key indicators, measuring changes in soil quality, and defining soil degradation (Gomiero, 2016; Bünemann et al., 2018). Constantini and Piori (2023) discuss the importance of informed selection and analysis of soil quality indicators and highlight that interpretation and use of the data require close cooperation of professionals with different and specific expertise. Definitions and common determinators are needed for developing soil protection policies. Monitoring of soil quality

and policy implementation remain challenging because changes in soil properties are often only apparent on decadal timescales and soil monitoring is a long-term commitment (Marchant and Saby, 2023). The authors of this chapter highlight essential features such that soil monitoring networks need to be cost-effective and robust to changing land uses, technological advances, and evolving soil quality priorities.

To date, however, the geographical distribution and total areas affected by soil degradation are still only roughly known. Estimates of global soil degradation range between less than 1 Gha to more than 6 Gha (Gibbs and Salmon, 2015; Fig. 1) reflecting the challenges of defining 'soil degradation', unreliable quantitative and qualitative data sources, differences in models and their assumptions, and missing information in general. By 2050, it is estimated that soil degradation will affect 90% of the soils globally (FAO and ITPS, 2015). Once a soil is degraded, its restoration in general is not straightforward. It can be costly and take a long time depending on the type and degree of the degradation, inherent land and soil properties, land use and climate. Better understanding of the drivers, threats and consequences of soil degradation, plus assessments of the costs of non-action, will support the development of a global soil protection policy and effective soil protection strategies.

Factors controlling soil degradation

Global population growth and changing diets in many parts of the world have increased the demand for food, feed and fuel, and have necessitated agricultural intensification to ensure food security for the growing population (Kraamwinkel et al., 2021). Significant yield increases achieved during the 'Green Revolution' in the middle of the last century were possible through massive increases in agricultural resource inputs, including water, energy, pesticides and fertilizers, and other technological advances such as the double-cropping and development of new cultivars. One of the harmful unintended environmental consequences of the 'Green Revolution' was widespread soil degradation (John and Babu, 2021).

Land-use change has threatened soil for millennia. Deforestation is the first step to convert natural vegetation into agricultural land and it often induces soil organic carbon loss and soil erosion processes, particularly in the tropics (Don et al., 2011; Pimentel

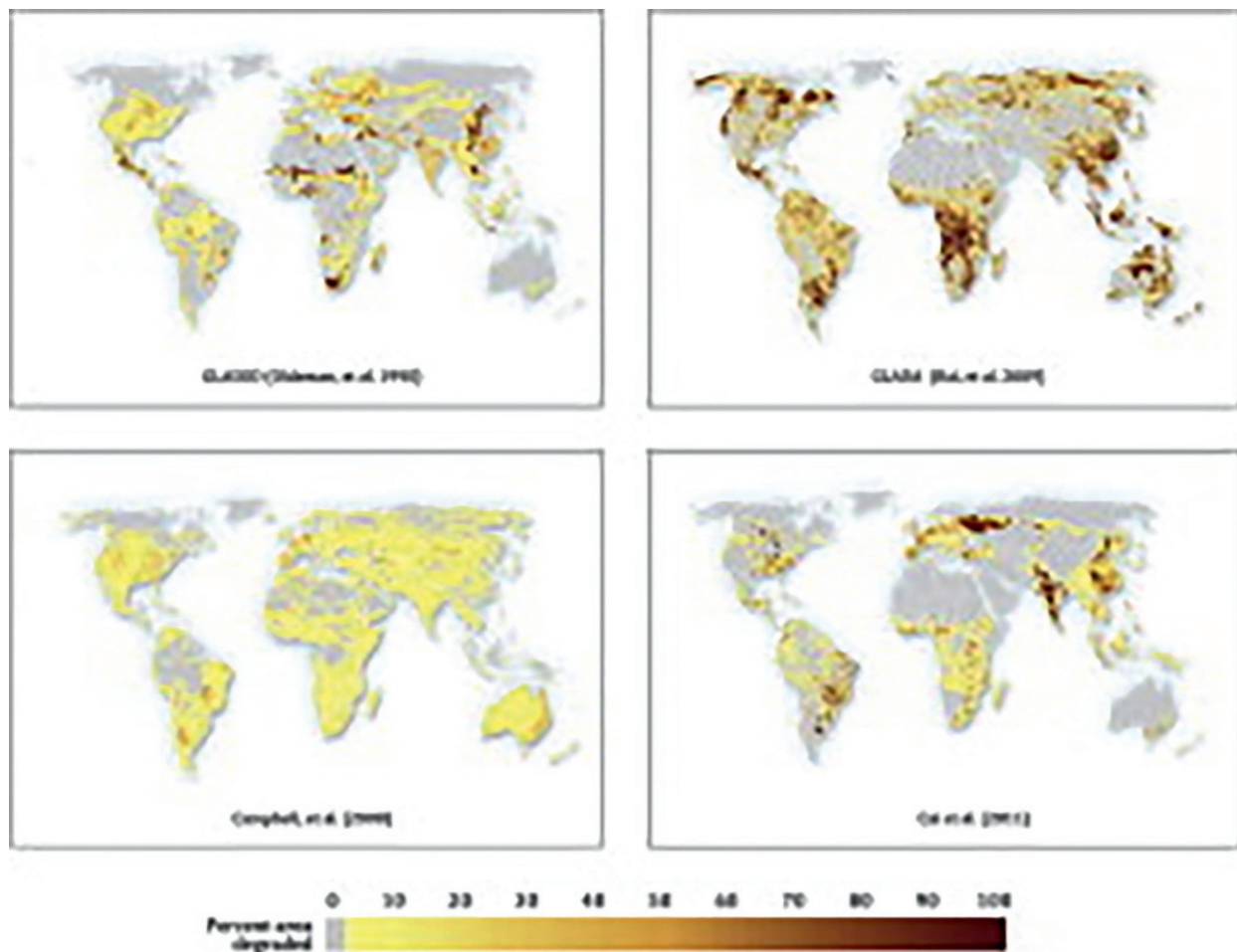


Fig. 1 Maps of land areas (percent of cell area; 20 k grid) affected by degradation; each panel represents a different method of modelling method described by Gibbs and Salmon (2015). Taken from Gibbs HK and Salmon JM (2015) Mapping the world's degraded lands. *Applied Geography* 57: 12–21.

and Burgess, 2013). Land-use change to extensive grasslands for cattle has reduced soil biodiversity and often leads to soil compaction (Centeri, 2022). Irrigation with poor quality water has increased soil salinity in many subtropical regions (Topcu, 2023; Qadir and Scott, 2023). Implementation of high value cash crops such as avocado (*Persea americana* Mill.), oil palm (*Elaeis guineensis* Jacq.), etc. might reverse rates of erosion because these crops keep the soil covered. However, they introduce agrochemicals in large amounts, which have been shown to reduce soil biological activity and can pollute soil and groundwater in the mid to long term (Meena et al., 2020). Recently, urbanization is sealing fertile land in some regions of Europe and North America and is now recognized as the most important soil degradation factor in populated areas (Scalenghe and Marsan, 2009).

Global population expansion further fosters the production of large volumes of waste, and augments pressure from other economic sectors such as industry, transport, and tourism on soil. The demands of agriculture and other economic sectors on soil are often competing and conflicting and contribute to soil degradation (Fig. 2).

Climate change is presenting a further threat to soil. Temperature increases, increased frequency of extreme rainfall events, and prolonged droughts are affecting biomass production and vegetation cover of natural ecosystems and its soils, and these effects are also closely interrelated with human land-use decisions. Resulting land-use changes go hand in hand with changing soil management practices that can further enhance soil degradation. Positive feedback loops between climate change, land-use changes and social and political pressures can exacerbate soil degradation.

Types of soil degradation

Lal (2015) classified soil degradation into physical, chemical, biological, and ecological degradation (Fig. 3). In the following chapters, a selection of types of physical and chemical soil degradation and their controlling factors are outlined. While types of biological soil degradation are not explicitly included in this section, soil organic matter loss is discussed jointly with the loss of soil structure in the chapter by Gregory (2023), and we provide a brief overview on biological degradation here.

Physical soil degradation

Physical soil degradation can, for example, occur as soil erosion, loss of soil structure, desertification, sealing and crust formation (Fig. 3). The chapters by Rickson (2023) and Tartako (2023) describe soil erosion processes, which are defined as the detachment and transport of the usually fertile topsoil material by water or wind. Climate change, land-use change, removal of vegetation and soil disturbance by human activities can accelerate topsoil loss, often exceeding the rates of soil renewal. As soil is lost to erosion, vital ecosystem services delivered by soil are also lost, and eventually the remaining soil is rendered infertile. Rickson (2023) presents the fundamental principles of water erosion, both measurement and monitoring approaches at different scales, and discusses modelling methodology as mitigation strategies. Tartako (2023) highlights that wind erosion, which occurs when strong winds impact loose, dry, or bare soils, is a global concern with significant on-site and off-site effects. The author uses cropland as an example to explore causes, effects, and controls of wind erosion. The general principles, however, are applicable to all landscapes.

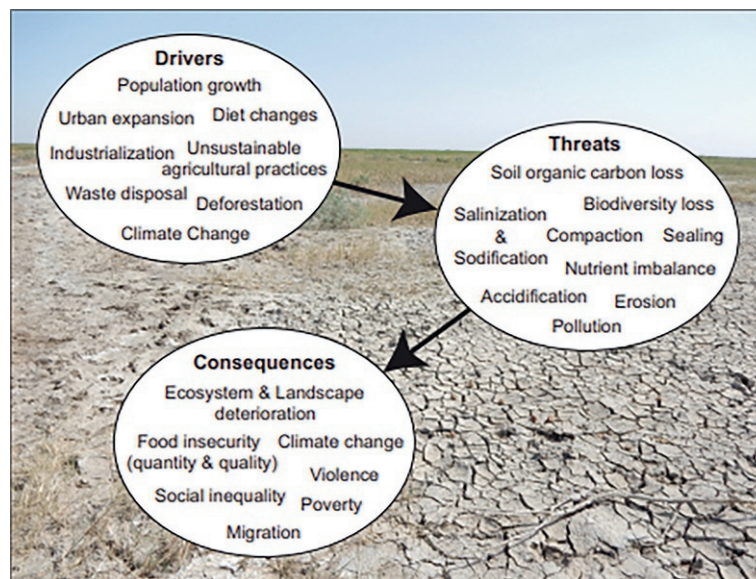


Fig. 2 Drivers, threats and consequences of soil degradation (Kraamwinkel et al., 2021).

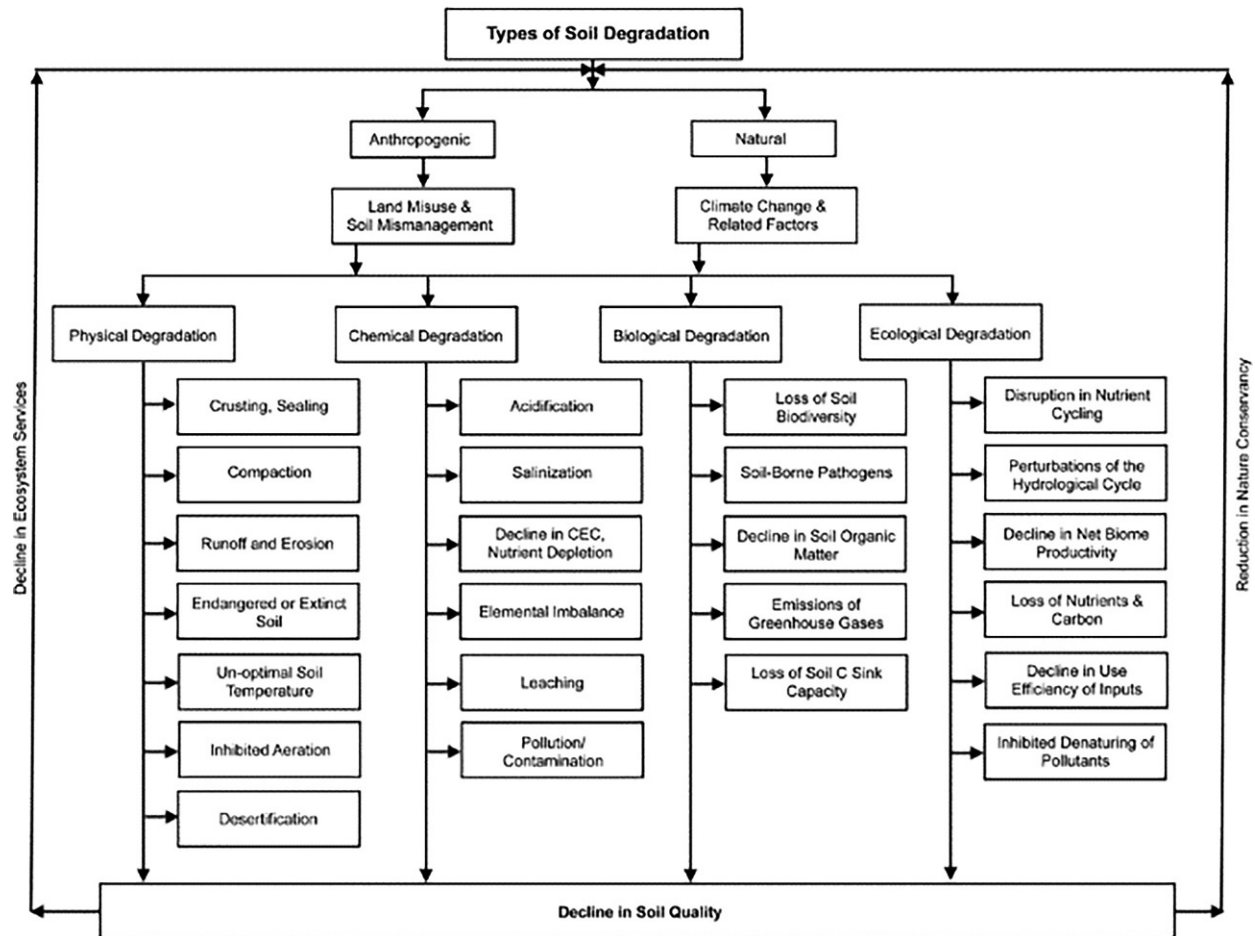


Fig. 3 Types of soil degradation (Lal, 2015).

A loss of soil structure, which is closely related to the loss of soil organic matter, is another physical process of soil degradation. Soil structure and organic matter are two dynamic soil properties that facilitate most soil processes at all spatial and temporal scales. The largest component of organic matter is carbon, and hence degradation of soil structure can have implications for the global carbon cycle and climate change. Gregory (2023) summarizes the symbiotic degradation of soil structure and organic matter including the main causes and effects of degradation on soil structure and soil organic matter, and the subsequent effects on soil functions.

Chemical soil degradation

Soil degradation can also occur as a result of chemical processes. Acidification, salination, elemental imbalances, decline in cation exchange capacity and soil contamination are examples of chemical soil degradation (Lal, 2015).

Agricultural productivity has been increased through increasing resource inputs. For example, during the last 70 years mineral fertilizer applications fundamentally changed soil and crop management. Singh and Sapkota (2023) outline how fertilizers increase the availability of essential plant nutrients in soils and enhance the natural productivity of soils. However, excessive fertilizer applications may also degrade soils by altering their natural functioning, which can lead to major and sometimes irrevocable changes in soil characteristics, e.g., nitrogen fertilizer may cause soil acidification, and soils may be contaminated with trace elements contained in rock phosphates, or in fungicides such as those used preventively in vineyards or other permanent high value cash crops.

Similarly, Topcu (2023) describes how irrigation of agricultural land has contributed to yield increases. Although approximately 25% only of the world's arable land is under irrigation, it accounts for more than 60% of the cereal production worldwide and 50% of the sales value of all crops harvested come from irrigated agriculture (Angelakis et al., 2020). Freshwater resources, however, are scarce and water scarcity is a major global issue (Mancosu et al., 2015). Current irrigation practices affect the availability and quality of soil and water resources. Agriculture is the largest water user and accounts for almost 80% of the global water use (Famiglietti, 2014). Topcu (2023) outlines options for improving the environmental performance of irrigation while increasing water

productivity in irrigated agriculture. Achieving these complementary goals is important for the long-term sustainability of irrigated agriculture.

The use of treated wastewater for irrigation offers a significant opportunity to increase productivity, particularly in water scarce areas. Most non-potable options for the use of wastewater require quality lower than that of drinking water, and wastewater can be treated for irrigation. Qadir and Scott (2023) provide an overview of wastewater irrigation and assess the trade-offs between the positive aspects of wastewater irrigation (year-round water supply, savings on fertilizer, less cost for pumping, increase in soil organic matter, and greater income from high-value crops) in relation to the negative risks (unbalanced plant nutrient application, increased soil salinity, metal ion pollution, and exposure to emerging contaminants and pathogens). Balancing these trade-offs requires positive outcomes to be ensured while minimizing risks through wastewater treatment and use compatible with its quality. A serious problem in developing countries is the use of large volumes of untreated or inadequately treated wastewater in agriculture, where governments do not spend resources to build wastewater treatment plants (Jones et al., 2021).

It is estimated that, worldwide, more than 50% of croplands will be salinized by 2050 (Jamil et al., 2011). Major causes of anthropogenic soil salination in arable lands are irrigation with saline water and improper soil drainage. Nachson and Levy (2023) differentiate between anthropogenic and natural salination processes. The authors discuss the effects of soil salination and practices that can be employed to overcome this challenge, including soil reclamation.

The expansion of irrigated areas is unsustainable, however, without proper drainage and funding for the appropriate implementation of technical and socioeconomic aspects. Drainage, the topic addressed by Strock (2023), is defined as the natural or artificial removal of water from the land either through surface drainage, subsurface drainage or a combination of both. The practice of land drainage may be beneficial if poorly drained conditions result from shallow groundwater or soils have low permeability. In agricultural settings, many soils require artificial drainage to improve field trafficability for field operations and to protect growing crops from excess soil water. Drainage has direct and indirect effects on the provision of ecosystem services including impacts on soil organic carbon stocks, organisms living in the soil, nutrient cycling, water retention, greenhouse gas emissions, and nutrient losses to surface and groundwater.

Soil contamination is defined as the presence of substances in soil, at concentrations higher than would either occur naturally, or through sustainable soil management practices. Soil contaminants include trace elements, radionuclides, pesticides, plant nutrients, fuels, solvents, crude oil, other organic substances, pharmaceuticals, and personal care products. Sources of contamination can be divided into point and nonpoint (diffuse) and can occur through applications of agricultural inputs, mining activities, combustion of fossil fuels, spills, wet and dry deposition, land application of by-products, and flooding (FAO and ITPS, 2015). In the Encyclopedia, three soil pollutants are discussed that result in soil degradation, namely trace elements, organic chemicals and microplastics.

Soils are the major sink for trace elements released from agricultural (e.g., mineral and organic fertilizers, pesticides) and industrial activities (e.g., mining, disposal of wastes). While trace elements do not undergo microbial or chemical degradation their chemical speciation can change, which affects their bioavailability, fate, and subsequent risk in soils. Wang et al. (2023) emphasize that trace elements of anthropogenic origin are generally present in more mobile and bioavailable fractions.

Kookana and Navarro (2023) highlight the problem of the long persistence time of some organic pollutants in the soil environment; the persistent organic pollutants (POPs). Their chapter provides an overview of how the chemical properties and fundamental processes of conventional POPs (e.g., polyaromatic hydrocarbons, polybrominated diphenyl ethers) as well as new POPs (e.g., *per*- and poly-fluoro alkyl substances - PFAS) control their fate and behavior in the soil environment and determine their potential impact on biota.

The chapter by Huerta Lwanga and Beriot (2023) introduces the issue of plastic pollution in agricultural soil. Plastic pollution occurs when plastic materials are fragmented into macroplastics (>5 mm) and microplastics (<5 mm), or when they are applied inadvertently with amendments such as composts or sewage sludge. The impacts of plastic particles on soil physical properties, plants, invertebrates and vertebrates, the biotic and abiotic transport of micro and nanoplastics through soils, and their transfer into the food chain are discussed.

Another aspect of the global population expansion is the generation of large volumes of diverse waste, which can cause soil degradation if not properly managed. As a subset of global waste, municipal solid waste includes household, business, institutional, and construction waste that may be processed, composted, recycled, and or disposed of at managed sites. Serre and McCarthy (2023) describe various municipal waste management strategies in Western societies, including recycling, composting, energy recovery, treatment and disposal. The purpose of such waste management regulations is to minimize the potential contamination of terrestrial ecosystems (by physical littering; landfill leachate) and aquatic resources (e.g., potable water, fisheries), and to preserve human health.

Biological soil degradation

Soil biological degradation refers to the depletion of the soil organic carbon pool, loss of soil biodiversity and soil biological functions, a reduction in soil carbon sink capacity, and increased greenhouse gas emissions from soil to atmosphere (Lal, 2015).

Losses of soil organic carbon in both, intensively and extensively managed agricultural production systems are contributing to increasing greenhouse gas emissions to the atmosphere (Cloy and Smith, 2023; Lal, 2004). Gregory (2023) summarizes the symbiotic degradation of soil structure and organic matter including the main causes and effects of these processes.

Soil is considered a living organism with more than 170,000 species of soil organisms identified (Lal, 2022). These organisms play a vital role in the functioning of the soil ecosystem, its stability and sustainability. Key ecosystem services provided by soil organisms are the formation and turnover of soil organic matter including mineralization and sequestration of carbon, fixation of atmospheric nitrogen, disease transmission and prevention, and nutrient cycling. The soil biological community also modifies soil physical structure through bioturbation and microaggregate formation, thereby, contributing to water regulation in soils. Soil organisms influence plant health through their effects on pathogens, nutrient cycling and uptake.

Studies have shown that anthropogenic activities such as land-use change and agricultural intensification (irrigation, cultivation, fertilization, etc.) and climate change can reduce abundance and diversity of soil biological communities (Tsiafouli et al., 2015). However, to date, it is not well understood whether such changes in biological communities have effects on the functioning of ecosystems. High functional redundancy has commonly been proposed as an explanation for the lack of clear relationships between soil microbial diversity and specific soil functions, e.g., organic carbon mineralization (Nannipieri et al., 2003). The shifting focus from short- to long-term effects of redundancy on ecosystem functioning reveals that redundant species can buffer against environmental changes (Jurburg and Salles, 2015). More recent research highlights the role of soil biodiversity for the provisioning of multiple ecosystem functions at various time points (Mori et al., 2016; Wagg et al., 2019, 2021).

Proposed indicators of biological degradation of soil include measurements of soil biodiversity and community composition, soil enzyme activities as indicators of nutrient cycling processes, accumulation of toxic molecules, and changes in the soil redox status (Sims, 1990). Qualification of soil biota according to traits, such as cellulose degraders, has been proposed to monitor biological soil health (van der Putten et al., 2023).

Soil remediation

The final chapter by Sanderson et al. (2023) is dedicated to soil remediation following soil contamination. In the recent decades, an emphasis on sustainability is increasingly placed on remediation technologies and new remediation approaches and techniques have been developed. Broadly, remediation technologies fall into the categories ‘containment and isolation’, ‘immobilization’, ‘separation and removal’ and ‘detoxification’. Remediation of contaminated soils is a critical part of the restoration of degraded lands to safe and beneficial reuse. Remediation will not only help to restore the ecosystem environments but will also provide additional land back into the pool required to sustain human life and food security.

References

- Angelakis AN, Zaccaria D, Krasilnikoff J, Salgot M, Bazza M, Roccaro P, Jimenez B, Kumar A, Yinghua W, Baba A, Harrison JA, Garduno-Jimenez A, and Fereres E (2020) Irrigation of world agricultural lands: Evolution through the millennia. *Water* 12: 1285.
- Bünemann EK, Bongiorno G, Bai Z, Creamer RE, De Deyn G, de Goede R, Flesskens L, Geissen V, Kuyper TW, Mäder P, Pulleman M, Sukkel W, van Groenigen JW, and Brussaard L (2018) Soil quality – A critical review. *Soil Biology and Biochemistry* 120: 105–125.
- Centeri C (2022) Effects of Grazing on Water Erosion, Compaction and Infiltration on Grasslands. *Hydrology* 9: 34.
- Cloy JM and Smith KA (2023) Methane and nitrous oxide emissions from land. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.
- Constantini EAC and Priori S (2023) Soil quality and health key indicators. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.
- Costanza R, d'Arge R, de Groot R, Farber S, Grasso M, Hannon B, Limburg K, Naeem S, O'Neill RV, Pareulo J, Raskin RG, Sutton P, and van den Belt M (1997) The value of the world's ecosystem services and natural capital. *Nature* 387: 253–260.
- Don A, Schuhmacher J, and Freibauer A (2011) Impact of tropical land-use change on soil organic carbon stocks – A meta-analysis. *Global Change Biology* 17: 1658–1670.
- Famiglietti JS (2014) The global groundwater crisis. *Nature Climate Change* 4: 945–948.
- FAO and ITPS (2015) *Status of the World's Soil Resources (SWSR) – Main Report*. Rome, Italy: Food and Agriculture Organization of the United Nations and Intergovernmental Technical Panel on Soils.
- Gibbs HK and Salmon JM (2015) Mapping the world's degraded lands. *Applied Geography* 57: 12–21.
- Gomiero T (2016) Soil degradation, land scarcity and food security: Reviewing a complex challenge. *Sustainability* 8: 281.
- Gregory AS (2023) Symbiotic degradation of soil structure and soil organic matter. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.
- Huerta Lwanga E and Beriot N (2023) Pollution by microplastics in agricultural areas. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.
- Jamil A, Riaz S, Ashraf M, and Foolad MR (2011) Gene expression profiling of plants under salt stress. *Critical Reviews in Plant Sciences* 30: 435–458.
- John DA and Babu GR (2021) Lessons from the aftermaths of green revolution on food system and health. *Frontiers in Sustainable Food Systems* 5.
- Jones ER, van Vliet MTH, Qadir M, and Bierkens MFP (2021) Country-level and gridded estimates of wastewater production, collection, treatment and reuse. *Earth System Science Data* 13: 237–254.
- Jurburg SD and Salles JF (2015) Functional redundancy and ecosystem function—the soil microbiota as a case study. *Biodiversity in Ecosystems-Linking Structure and Function* 29–49.
- Keesstra SD, Bouma J, Wallinga J, Tittonell P, Smith P, Cerdà A, Montanarella L, Quinton JN, Pachepsky Y, van der Putten WH, Bardgett RD, Moolenaar S, Mol G, Jansen B, and Fresco LO (2016) The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals. *The Soil* 2: 111–128.
- Kookana RS and Navarro DA (2023) Soil pollution, organic pollutants. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.
- Kopittke PM, Menzies NW, Wang P, McKenna BA, and Lombi E (2019) Soil and the intensification of agriculture for global food security. *Environment International* 132: 105078.
- Kraamwinkel CT, Beaulieu A, Dias T, and Howison RA (2021) Planetary limits to soil degradation. *Communications Earth & Environment* 2: 249.
- Lal R (2004) Soil carbon sequestration impacts on global climate change and food security. *Science* 304: 1623–1627.
- Lal R (2015) Restoring soil quality to mitigate soil degradation. *Sustainability* 7: 5875–5895.
- Lal R (2022) Nature-based solutions of soil management and agriculture. *Journal of Soil and Water Conservation* 77: 23A–29A.
- Mancosu N, Snyder RL, Kyriakakis G, and Spano D (2015) Water scarcity and future challenges for food production. *Water* 7: 975–992.
- Marchant BP and Saby N (2023) Soil monitoring. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.

- McBratney AB, Morgan CLS, and Jarrett LE (2017) The value of soil's contributions to ecosystem services. In: Field DJ, Morgan C, and McBratney AB (eds.) *Global Soil Security*. Cham: Springer International Publishing.
- Meena RS, Kumar S, Datta R, Lal R, Vijayakumar V, Brtnicky M, Sharma MP, Yadav GS, Jhariya MK, Jangir CK, Pathan SI, Dokulilova T, Pecina V, and Marfo TD (2020) Impact of agrochemicals on soil microbiota and management: A review. *Land* 9: 34.
- Mori AS, Isbell F, Fujii S, Makoto K, Matsuoka S, and Osono T (2016) Low multifunctional redundancy of soil fungal diversity at multiple scales. *Ecology Letters* 19: 249–259.
- Nachson U and Levy GJ (2023) Soil salination processes and management. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.
- Nannipieri P, Ascher J, Ceccherini MT, Landi L, Pietramellara G, and Renella G (2003) Microbial diversity and soil functions. *European Journal of Soil Science* 54: 655–670.
- Pimentel D and Burgess M (2013) Soil Erosion threatens food production. *Agriculture* 3: 443–463.
- Qadir M and Scott C (2023) Trade offs of wastewater irrigation. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.
- Rickson J (2023) Soil erosion by water. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.
- Sanderson P, Biswas B, Mezbaul Bahar M, and Naidu R (2023) Remediation of soils polluted with heavy metals and organic pollutants. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.
- Scalenghe R and Marsan FA (2009) The anthropogenic sealing of soils in urban areas. *Landscape and Urban Planning* 90: 1–10.
- Serre BM and McCarthy LH (2023) Municipal waste disposal on land. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.
- Sims GK (1990) Biological degradation of soil. In: Lal R and Stewart BA (eds.) *Advances in Soil Science: Soil Degradation*. vol. 11, pp. 289–330. New York, NY: Springer New York.
- Singh B and Sapkota TB (2023) The effects of adequate and excessive application of mineral fertilizers on the soil. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.
- Strock J (2023) Drainage, soil water storage, buffering and filtering. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.
- Tartako J (2023) Erosion by wind. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.
- Topcu S (2023) Environmental effects of irrigation. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.
- Tsiafouli MA, Thébault E, Sgardelis SP, de Ruiter PC, van der Putten WH, Birkhofer K, Hemerik L, de Vries FT, Bardgett RD, Brady MV, Bjornlund L, Jørgensen HB, Christensen S, Hertefeldt TD, Hotes S, Gera Hol WH, Frouz J, Liiri M, Mortimer SR, Setälä H, Tzanopoulos J, Uteseny K, Piži V, Stary J, Wolters V, and Hedlund K (2015) Intensive agriculture reduces soil biodiversity across Europe. *Global Change Biology* 21: 973–985.
- van der Putten WH, Bardgett RD, Farfan M, Montanarella L, Six J, and Wall DH (2023) Soil biodiversity needs policy without borders. *Science* 379: 32–34.
- Wagg C, Hautier Y, Pellkofer S, Banerjee S, Schmid B, and van der Heijden MG (2021) Diversity and asynchrony in soil microbial communities stabilizes ecosystem functioning. *eLife* 10.
- Wagg C, Schlaeppi K, Banerjee S, Kuramae EE, and van der Heijden MGA (2019) Fungal-bacterial diversity and microbiome complexity predict ecosystem functioning. *Nature Communications* 10: 4841.
- Wang P, Zhao F-J, and McGrath SP (2023) Pollution and risks of trace elements in the soil environment. In: *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.

Soil monitoring

Ben Marchant^a and Nicolas Saby^b, ^aBritish Geological Survey, Keyworth, United Kingdom; ^bINRAE, Orléans, France

© 2023 Elsevier Ltd. All rights reserved.

Introduction	172
Underlying principles	173
Why monitor soils?	173
The starting point	174
Soil monitoring networks	175
Protocols	176
Monitoring sites	176
Soil sampling	177
Quality control	178
Recent developments in soil monitoring practices	178
Summary	179
References	179

Abstract

Soil monitoring is required to determine where threats to soil function and health are appearing, to assess the effectiveness and sustainability of land management practices and to understand soil processes better. Many changes in soil properties are only apparent on decadal time-scales and therefore soil monitoring is a long-term commitment. Soil monitoring networks must be established with thorough and carefully recorded protocols that are cost-effective and robust to changing land uses and technological advances, and that can adapt to address evolving soil health priorities and questions. This chapter considers the many elements of such a protocol.

Key points

- Soil monitoring requires careful and long-term planning.
- The motivation for establishing a soil monitoring network must inform the protocols and design of the network.
- The network protocols must be robust to changing policy and science priorities, landscapes, organizational structures and technology over decadal time-scales.
- New technologies and measurement techniques have the potential to improve the accuracy and resolution of soil monitoring networks and the breadth of soil processes and functions that can be monitored.

Introduction

In the first edition of this encyclopedia, Loveland and Bellamy (2005) described the great commitment and considerable planning that is required to monitor the status of soils and to provide reliable information about changes that are occurring. They highlighted the need for robust protocols so that comparable soil measurements and information can be gathered many years apart and for decadal scale funding to ensure that the monitoring efforts are fully exploited. More than 15 years on, many technological advances have occurred. As standard, soil measurement sites are now reliably located using Global Positioning Systems (GPS). Advances in disciplines such as metagenomics and techniques for measuring more types of soil pollutants mean that more properties of the soil can be measured. Remote sensing methods have increased the spatial extent of soil and landscape information, whereas proximal soil sensing methods permit spatially intensive measurements. Both types of sensors have also led to an increase in the achievable temporal frequency of measurements. Furthermore, adoption of advanced statistical and machine learning techniques have aided the interpretation of the sometimes huge soil datasets that result from monitoring. However, the underlying principles from 2005 remain true. Changes in the soil can only be reliably quantified if a monitoring survey is designed according to sound statistical theory, if consistent measurement methods are used throughout the survey, if appropriate data analysis methods are applied and if all elements of the soil monitoring protocol are reliably documented. This still requires considerable and long-term financial commitment. Therefore, in this revised chapter we make only small modifications to Loveland and Bellamy's (2005) description of the methodology and principles that underpin surveys for soil monitoring. We do, however, briefly discuss how recent technological advances might improve soil monitoring and the extent to which they can compensate when the optimal monitoring procedures cannot be followed.

Underlying principles

Soil monitoring (SM) is often driven by the desire to understand soil quality and changes in it. This usually requires the establishment of soil quality indicators or properties. This can, and does, have a profound influence on what type of SM is carried out, how a program is designed and executed, and what it costs (e.g., [Black et al., 2008](#)). Effective SM implies long-term commitment and very tight control of potential sources of variability. Decisions taken at the planning stage are of extreme importance, and careful attention to them can save much money and trouble and lead to greater reliability. Two essential points are:

- There is no universal system of soil monitoring; the means must fit the purpose and vice versa.
- There is a clear distinction between a soil inventory program and a soil monitoring program.

The first point might seem obvious, but is often overlooked. What is appropriate for an industrial site and its surroundings, where potential hazards to human health might be the motivation, can be wholly inappropriate in an area where maintenance of a particular soil ecosystem is the target or at the national-scale where multiple concerns must be considered.

The second point is that an inventory is a collection of information about a place. For soils, it might be the features observed at the location of a soil profile and the properties of the profile horizons. 'Monitoring' implies temporal change, i.e., observation, sampling, and measurement repeated over time. There is thus a different intent and commitment in soil monitoring compared with an inventory, although the latter might be the starting point (baseline) of a soil monitoring program. It is stressed that a single set of measurements in time does not constitute soil monitoring.

Most commonly, soil monitoring targets change in a specific property or properties, e.g., pH, metal content, biomass. However, it can be equally important to study a process, e.g., soil erosion. Increasingly, there is a demand to understand potential soil functions ([van Leeuwen et al., 2017](#)). Linking soil monitoring to other environmental monitoring programs can maximize long-term benefits from the data collected. Planning soil monitoring needs to take these different objectives into account.

Finally, although many of the points discussed below might also apply to small parcels of land, this chapter is mostly concerned with the concept of soil monitoring applied at the regional or national scale.

Why monitor soils?

Since the turn of the century, there has been growing acceptance that not enough is known about soils and the changes that they are, or might be, undergoing, especially under increased pressure from human activities ([European Commission, 2002](#)). This contrasts with the effort that has gone into the assessment of air and water quality. There are concerns that pressures on soils, usually driven by economic needs, will lead to irreversible decline in their ability to maintain ecosystem diversity, food security, carbon storage, fiber and timber production and this will lead, indirectly, to serious problems with water and air quality and climate regulation. Thus, the whole framework of sustainable development would be compromised. There have been many statements over the last few decades pointing to the need to consider soils in terms of the:

- establishment of a set of principles for the rational use of soils and protection of them against irreversible degradation;
- pursuit of programs of soil conservation and reclamation and appropriate use of soils;
- recognition of soils as a fragile and essentially irreplaceable resource;
- importance of specific soil properties such as those that determine best use and land management practices
- need for inventories and monitoring programs to establish baseline soil properties and changes in these over time.

Against this background, soil monitoring can be instigated for a number of reasons:

- A. Basic science: are soil properties changing and, if so, where and at what rate and in what direction, and what does this tell us about our understanding of the world? For example, establishment of the magnitude of change in soil carbon content could improve inputs into models of the global carbon cycle;
- B. Political: the need for information from which to formulate, test or improve policy, e.g., conventions on transboundary pollution require reductions in the acid load to soils; this could be demonstrated by showing that soil pH is remaining above a given value over a given time;
- C. Legal: whether change has occurred to a soil that is in conflict with legislative obligations, e.g., demonstrate that loadings of so-called 'heavy metals' are within statutory limits;
- D. Financial: are the properties of a soil changing (or likely to change) over a given time-frame such that the soil is, or will become, more or less valuable in a given context, e.g., is there a potential problem from pollution that could adversely affect land that would otherwise be valuable for development.

Much of this thinking is increasingly considered in terms of the ability of soils to perform several functions, often simultaneously ([van Leeuwen et al., 2017](#)), e.g., food and fiber production, filtering, buffering, acting as a pool of genetic resource and new antibiotics, support for diversity of above- and below-ground ecosystems, keeper of the archaeological record, support for the built environment. Thus, soil monitoring is under pressure to provide information on the performance (or potential performance) of these functions and, often, to guide choices as to which set of functions should be 'allocated' to a parcel of land in preference to

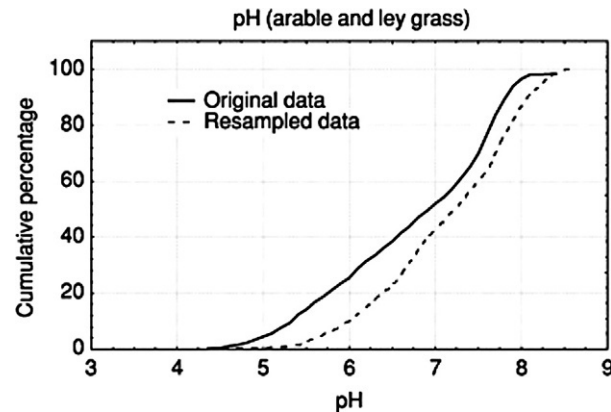


Fig. 1 Change in topsoil pH between 1980 and 1995 under an arable/ley grassland rotation in England and Wales ($n = 900$) Loveland PJ (1990) The National Soil Inventory: survey design and sampling strategies. In: Lieth H and Markert B (eds) *Element Concentration Cadasters in Ecosystems*, pp. 73–80. Weinheim, Germany: VCH Verlagsgesellschaft. Data from the National Soil Inventory of England and Wales and figure previously used in Loveland and Bellamy (2005).

another, e.g., build here but not there. The measurement of properties is often no longer the absolute *raison d'être* of soil monitoring, although it is often the baseline. Fig. 1 shows that over 15 years the pH of topsoils (0–15 cm) increased under arable/ley grass rotation in England and Wales, i.e., the soils became less acid. This finding has been interpreted (Kirk et al., 2009) as a mixture of the reduction in acid deposition (policy response), better targeting of the use of agricultural lime (improved agronomic practice), and deep plowing into less acid subsoil because of larger, more powerful farm machinery (economic driver). All these reasons can be viewed as useful outcomes of soil monitoring and could lead to more focused targeting of future environmental research.

The starting point

Whatever the driver for soil monitoring, some fundamental questions need to be addressed (e.g. Black et al., 2008):

- A. What information is already available about soil type(s), properties, scale of observations, etc.?
- B. Is this information internally consistent, e.g., are, or were, all the data obtained in the same way, or in ways comparable enough for the purpose intended?
- C. What information is not available and does this matter?
- D. At what spatial and temporal scale is further information required?
- E. Is information required simply as values of a property, or in ways that describe soil functions; if so, which function is of greatest interest?
- F. What is the target, e.g., all soils in an area, the common soils, rare soils, a specific ecosystem, a specific soil process?
- G. What magnitude of change is of interest, over what time scale, and is this a realistic target?
- H. Is the property measurable or is further research needed?
 - I. Is it necessary to collect more information, or can the need be met from surrogate measurements, pedotransfer functions or from extrapolation, interpolation and classification of existing data?

Organizational and practical questions arise:

1. How can soil monitoring yield the greatest amount of information for the greatest number of people for the least cost?
2. At what time interval is the monitoring to take place?
3. Is this interval suitable for all soil properties of interest?
4. Can the needs of the users be satisfied with one soil monitoring design, i.e., number and layout of sites; if not, what modifications are needed?
5. Will one organization do everything, or is it to be a consortium (if so, who leads?)
6. Does the lead organization have the ability, infrastructure and will to keep that role for a sufficient period? It is no use setting up a 10-year program if those involved drop out after 5 years;
7. Are there agreed protocols for the design of the monitoring network, for the location, layout, and description of sites, for the soil sampling, for the preparation and storage of samples, for laboratory work, for quality control; if not, who will orchestrate this?
8. Who will keep the samples, the raw data, the processed data, and be responsible for them into the future?
9. Who 'owns' the data, i.e., are there issues of copyright, intellectual property rights, confidentiality, and controls on the release of data?

10. Is there an agreed publication policy for the output from the soil monitoring, given that many years might pass before sufficient data are available to allow robust statements to be made about change?
11. What will the various options cost, bearing in mind that the costs of very long-term site maintenance, quality control, and storage can be substantial?
12. Will the monitoring sites remain accessible in the future (urbanization, site access)?

Many of these points might seem trivial. However, the time scale over which soil monitoring must run before large amounts of meaningful data are acquired has to be seen in decades. Thus, the objectives, responsibilities, targets, and costs need to be thought through with great care and appropriate commitments obtained.

Soil monitoring networks

Many countries have conventional soil survey maps showing the spatial distribution of soil types. Although the scale of the map controls the smallest area of any soil type that can be shown, even relatively coarse-scale maps, e.g., 1:250000 scale, can indicate the relative abundance of soil types. At a very early stage, it has to be decided whether the soil monitoring network is to be capable of sampling rare soils, i.e., those occupying only a small proportion of the landscape. As a guide, a network that is expected to 'capture' a soil occupying 5% of the landscape with 95% confidence would need several thousand sampling locations (Table 1). An example is the National Soil Inventory (NSI) for England and Wales (UK; approx. 150,000 km²), which sampled almost 5700 points on a regular grid, but only captured about 400 of the 700 soil series that are known to occur in the two countries (Loveland, 1990). If no soil maps exist, then the problem becomes even more pressing.

Numerous designs (see examples in Fig. 2) have been proposed for national or regional soil monitoring networks (Arrouays et al., 2012). The NSI network is based on sampling at the intersections of a 5 km × 5 km-square grid. The Forest Soil Monitoring Network in Europe is based on a regular 16 km × 16 km grid, and there are many other examples of the use of square grids. The most efficient regular geometric grid is triangular, but the square grid has traditionally been favored because it can be co-located easily with a national geographic grid, thus making it easy to find points on the ground using conventional maps. This provision has become less important with the common use of GPS. One caution often expressed about square grids is that the periodicity of the grid might match that of some periodicity of the soil itself. However, whilst this might be true at the field scale, e.g., rows of crops in orchards and olive groves, 'tramlines' in arable crops, etc., no evidence has yet been presented that such periodicity exists for very

Table 1 Size of random sample required (*n*) to estimate the proportion (*P*) of soils of a specified type within limits $P \pm 0.1P$ for three levels of confidence, calculated from the normal approximation.

<i>P</i> or $1-P$	<i>Confidence Limits</i>	<i>n</i>		
		80%	90%	95%
0.5	±0.05	164	269	384
0.4	±0.04	245	331	576
0.3	±0.03	382	627	896
0.2	±0.02	655	1075	1536
0.1	±0.01	1474	2420	3456
0.05	±0.005	3112	5109	7296

Adapted from Webster R and Oliver MA (1990) *Statistical Methods in Soil and Land Resource Survey*. New York: Oxford University Press, with permission and previously used in Loveland and Bellamy (2005).

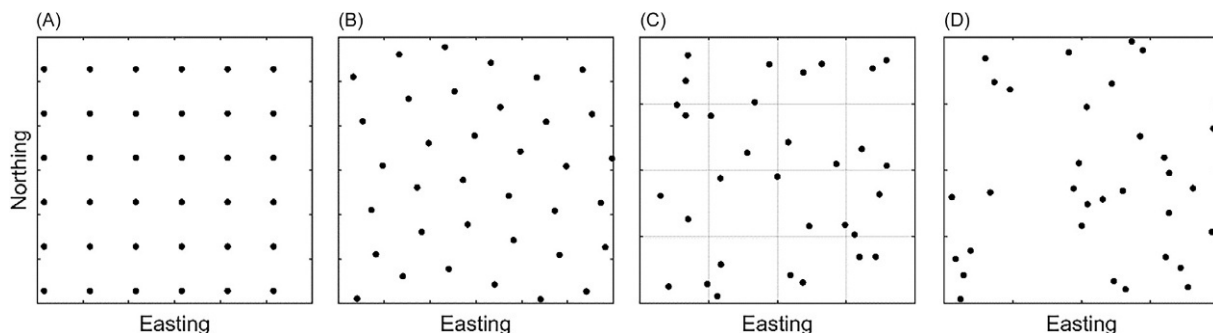


Fig. 2 Realizations of different spatial sampling designs: (a) a 36-point regular square grid, (b) a 36-point unaligned design, (c) a 32-point stratified random sample (strata are the marked squares) (d) a 32-point simple random sample.

large areas of land. In situations where periodicity could be a concern, an unaligned sampling might be used. According to this design, each cell of a regular square grid is sampled once but the position of the measurement within the cell is varied. The y-coordinate of any one column and the x-coordinate of any one row are selected at random and fixed for that column or row. The design ensures that measurements are reasonably evenly dispersed across the study area but they are not aligned to any periodic features.

Alternatively, networks can be designed based on a probabilistic selection of points. Simple random sampling can be inefficient as the points might not be evenly dispersed across the study region. This issue can be partly addressed by stratifying the sampling according to spatial sub-regions or according to covariate information such as soil classes. The Soil Sampling Programme (SSP) for the Netherlands implemented a stratified random sampling approach. The primary stratification was based on water table depth and pedologically homogeneous classes (Knotters et al., 2022).

The pros and cons of systematic and probabilistic designs have been widely discussed (e.g. de Gruijter et al., 2006; Black et al., 2008; Arrouays et al., 2012). In summary, data collected from probabilistic designs are amenable to classical statistical analyses leading to unbiased and valid estimates of the mean across reporting units and quantification of the uncertainty of these estimates. In contrast, systematic designs cover the study region more efficiently and are well suited to geostatistical analyses and mapping of the resultant data. However, model-based techniques must be used to analyze data from systematic designs and these techniques require the practitioner to make assumptions about the statistical distribution from which the measured values are realized and the correlation between measurements from different sites.

An alternative design approach is weighted or purposive sampling, i.e., the sampling points are located on the basis of, for example, an existing network of sites such as experimental farms, sites for other forms of monitoring, selected soil-landscape combinations or, possibly, perceived soil functions. The principal problems with this approach are:

- It might not yield enough points to allow the spatial structure of the data to be estimated reliably by geostatistical methods—too much of the landscape is omitted;
- There is a high risk of introducing bias into the estimation of the statistical parameters.

However, such schemes can yield reliable baseline data for particular soil properties, e.g., lead content, provided that the limitations are appreciated. Well-conducted monitoring networks of this kind are being run, for example, in Switzerland, Austria, The Netherlands and France (Saby et al., 2017) among others.

Protocols

Whatever is done and by whom, it is absolutely essential to have protocols in place for all aspects of the work: site selection, accuracy of location, sampling strategy at each site, sampling, sample treatment in field and laboratory, analytical procedures, data storage and handling, and data access. Uncontrolled sources of potential variability are the enemy of long-term monitoring. Peoples' memories are not reliable over long timescales, and there is no certainty that those people will be available some years later to answer any questions. Information including meta-data, sampling protocols and perhaps photographs must be stored securely in digital repositories and proper recording and training undertaken. Above all, it is vital that information is shared between investigators to ensure safe long-term storage.

Technological developments might lead to improvements in protocols over time. For example, more accurate methods of measuring soil properties could be developed. Careful thought must be given to adopting these advances in later rounds of a SM network. The improved accuracy must be weighed against difficulties in comparing measurements made according to different protocols. If protocols are changed then archived samples from previous sampling rounds should be analyzed according to the new protocol, or new samples should be analyzed according to the previous protocol, to determine whether adjustments should be made to the new measurements.

Monitoring sites

Whatever design of survey is chosen, it is important to avoid bias in the exact location of measurements within a sampled site i.e., the person on the ground should not be allowed to pick and choose where to locate a soil core and how to sample.

The number of sites (n) should be sufficient to estimate the status of each soil property for each round of sampling and its rate of change between sampling rounds. However, the precision of an estimate of this status or rate of change depends upon the standard deviation of these quantities which are unknown until the data have been collected and analyzed. Hence, it is not possible to guarantee the precision of a monitoring survey prior to sampling. Furthermore, the standard deviation of change cannot be determined easily from a single sampling round (Lark, 2009). This means, that it is not possible to guarantee that resampling a specified proportion of an existing baseline survey will lead to an estimate of the rate of change with sufficient precision. A limited reconnaissance survey could be conducted to determine the relevant standard deviations and the required number of sites but, more usually, plausible modelling assumptions are used to simulate soil property data upon which different sampling designs are tested. For instance, Saby et al. (2008) used simulated data which indicated that for almost all national-scale SM networks in Europe, the

minimum detectable change in soil organic carbon stocks is greater than the annual greenhouse gas emissions. Therefore, a length of time of about 10 years between sampling rounds would be required to detect change in most countries.

In general, the rate of change in soil properties is established efficiently if the same sites are revisited in each round of the survey. However, it is important to ensure that non-representative land management practices are not applied at these sites because they are known to be part of a monitoring network and to influence the results of the survey deliberately (e.g., if a land manager is aware that rewards for environmental stewardship will be determined based on soil health metrics at specified sites, [de Gruijter et al., 2016](#)).

Although it might be possible to control activities at a small number of key sites, it is uncommon to be able to tie land owners or managers to particular land-use agreements for decades. Thus, sites can be built upon, become dumps for waste, become polluted and thus unacceptably dangerous to work on, or the land use might change so drastically that the site is no longer fit for monitoring purposes or cannot be accessed. The latter is always difficult to assess. Some would argue that if a soil is converted from arable to woodland, monitoring the progress of that change is a matter of considerable interest. Others might argue differently because they want to know how cultivated land behaves in the long term. Whatever is decided, there is a need to avoid the situation whereby the loss of a relatively small number of sites compromises the whole network. To use the example from England and Wales (UK), approximately 3% of the c. 900 sites under arable/ley grassland in 1980 had been lost to agriculture by 1995. This change was interesting in itself, but the 'loss' of these sites did not seriously alter the conclusions drawn from the data for the remaining sites (approximately 870). It is also strongly recommended that the network should have sufficient points to be robust to these losses of sites.

The simplest site is a sampling point, a spot on the ground located to a given accuracy. Points need to be located geographically according to a protocol that fits the purpose of the network. Experience has shown that people with some background of field work, using good topographic maps (1:10000 or 1:25000 scale) and/or aerial photographs, can easily relocate a point to within 10 m of a target point. With GPS, relocation to within less than a meter is possible, although nearby buildings, trees, and hilly ground can seriously distort signals. It can be valuable to mark these points to aid future relocation. A vandal-proof approach is the burial of either a metal plate (a sheet of aluminum alloy approximately 25-cm square and 5-mm thick) or of a small but powerful permanent magnet (wrapped in a plastic bag) at approximately 50-cm depth, i.e., well below plow depth. These can easily be found again after many years by appropriate hand-held, electromagnetic detectors.

There are many alternative arrangements of kinds of sites. Two of the most common are:

- Area-based sites, in which it is sufficient for the purposes of monitoring to analyze composite samples of soil that are representative of an entire parcel of land (e.g. [Oliver et al., 2006](#)). This approach is quite common in what might be termed 'agricultural monitoring,' where the argument is that fields or other parcels of land are subject to such intense management that the whole parcel becomes the site;
- Key sites or control sites that are specifically managed to be available for revisiting in the long term. An area of relatively uniform soil is identified (often of a hectare or more) and sampling plots are laid out across it in sufficient number to allow resampling according to a strict protocol that avoids resampling of previous plots for many years. These kinds of sites can be extended to allow investigation of large-scale processes such as soil erosion or the effects of afforestation, etc., but it is axiomatic that large sites, with their (partial) dedication to the long-term study of specific soil processes are expensive to run. In most countries, therefore, that have followed this thinking, the number of such sites is small (10–15 is not uncommon).

There is a belief that a small number of such key sites should be a component of any soil monitoring network as they can form a framework by which the quality control of a much larger network can be achieved.

Soil sampling

There are two key questions to be asked when considering soil sampling:

- Which soil layers are to be sampled, e.g., soil horizons (the people doing the sampling need to be able to recognize these) or fixed-depth sampling, and are litter layers to be included?
- Is the sample to be taken at a point, e.g., from the face of a soil pit or with an auger, or is it to be a bulked sample (of how many cores?) taken using the geographic point as a centroid?

In practical terms, often a combination of these approaches is used. Surface layers are readily sampled around a centroid, e.g., 25 subsamples at 4-m intervals on a 20 m × 20 m grid, a 'random' walk across a parcel of stated extent, subsampling along the radii of a circle, or a numbered layout of sufficient sampling plots to allow a considerable number of resamplings without the risk of revisiting disturbed soil. There is much to be said for this 'bulkied' approach as at least a partial answer to dealing with the short-range spatial variation of soils. Others argue for a formal, geostatistical approach to each site, so that the structure of its spatial variation is known and an appropriate sampling scheme drawn up. This can, however, have large cost implications if it is to be applied to hundreds of sites. Whatever is done should not compromise future sampling—the temptation to remove the whole of the topsoil layer at the centroid should be resisted; similarly, soil pits should be small.

Finally, the sample must be big enough, because there are usually several people who want some of it. At least 2 kg on an air-dried, less than 2 mm basis is a useful target for present and future analyses. However, the French national soil monitoring survey

has sometimes found that 5-kg samples are insufficient to support all potential research activities. Even a mass of 2 kg has implications for the volume of organic (peat) soils sampled. Sample storage needs considerable prior discussion with all who might have an interest, e.g., the chemists might be happy with dried material, but what do the biologists want? Long-term 'fresh' sample storage implies access to specialized facilities—Do they exist? Can they handle the number of samples (this year, next year, every X years, etc.)? Storage of frozen samples can pose similar problems. Air-dried material can be stored in screw-cap glass bottles with Teflon lid liners (but these can be costly). Special consideration should be given to the storage of samples for microbial analysis (e.g., [Schroeder et al., 2021](#)). The amount of shelf space required for the next 10, 20, 50, etc. years needs to be estimated, as well as the cumulative weight of the material. An intelligible sample-identification system is required so that samples can be retrieved, and a mechanism is necessary that ensures no one arbitrarily moves the samples or throws them away.

If the physical properties of soils are of interest, e.g., bulk density, aggregate stability, water-retention characteristics, and hydraulic conductivity, then other sampling techniques are required (e.g., undisturbed cores or samples of known volume) that are best discussed with the appropriate specialists. It might be necessary to designate a separate area of the site for this kind of work, as considerable disturbance can be involved. Above all, it is necessary to allow for sufficient sampling space to accommodate all potential measurements without compromising any of the others.

Quality control

This is not just a laboratory matter, but applies equally well to field operations. Whichever is under consideration, protocols, testing, and training are paramount; slipshod procedures at any stage can undermine the whole enterprise.

In the field, there is much to be said for having a formal recording system—the site location, local land characteristics, aspect, slope, etc. being recorded in standard ways, and (permitted) deviations from these should be recorded also. For example, it is common to find that the specific geographic point is not accessible (e.g., because it is covered by a road or building). Rather than abandon the site, limited, structured deviation is permitted, e.g., move 100 m north, then east, then south, etc., the details of the deviation being recorded. Such information can be recorded on proformas on portable computers.

Long-term quality control requires that a proportion of the samples are analyzed in duplicate or triplicate during the current program and during every subsequent resampling. This is to detect long-term drift. The proportion is a matter for discussion, but, as the monitoring program continues, the amount of 'repeat' analysis becomes the largest component of the laboratory work, with implications for costs. If samples are to be sent to distant laboratories, a system has to be established for this. For example: Are field refrigerators or 'cool boxes' required? How quickly should the samples be sent (overnight for biological monitoring)? Does the laboratory know what to do when the samples arrive? Protocols have to be set up for all these matters and tested before the monitoring commences, e.g., can the protocol deal with samples arriving at the start of a public holiday?

Ideally, all the determinations should be made in one laboratory but, even where this is not possible, thought should be given to dividing the work into blocks, e.g., all the metals are determined in one laboratory, all the organics in another, all the biology at a third. Are the chosen laboratories up to standard and are they familiar with soils? Do they have formal QA procedures in place and are they accredited and audited? Do they use widely recognized standard methods (International (ISO) standards should be used wherever possible)? Are they viable in the long term? Cheapest is not necessarily best.

Similar approaches are necessary with regard to output from the soil monitoring program. The data need to be held securely in a system that has a good prospect of long-term stability in terms of software and staffing. Data ownership, intellectual property, personal data regulations and related issues need to be clarified at an early stage, as does a policy over the publication of results.

Recent developments in soil monitoring practices

More than 15 years after Loveland and Bellamy's original encyclopedia article, the need for soil monitoring is undiminished. A harmonized framework for soil monitoring in Europe is one of the four pillars of the "Soil Deal for Europe" European Union Mission ([Panos Panagos et al., 2022](#)). Emphasis continues to be placed on monitoring the capacity of soils to (often simultaneously) perform vital functions such as primary production, water purification and regulation, carbon sequestration and climate regulation, soil biodiversity and habitat provisioning and recycling of nutrients. Soil pollution by metals remains a major concern and many more chemical contaminants are being monitored (e.g. [Froger et al., 2021](#); [Yu et al., 2020](#)). However, within Europe, the data from broad-scale soil monitoring networks are unbalanced both geographically and in terms of the functions that can be assessed ([van Leeuwen et al., 2017](#)).

In recent years, considerable emphasis has been placed on monitoring the amount of carbon stored in soils. Land managers and policy makers need to understand better the circumstances in which soil carbon sequestration can be achieved in order to limit the amount of carbon dioxide in the atmosphere and the rate of increase of global temperatures. However, [Saby et al. \(2008\)](#) note that the minimum detectable change in carbon stocks for almost all European SM networks is greater than annual greenhouse gas emissions and therefore it is unlikely that these networks can be used for annual carbon accounting. It would take at least a decade to address these shortcomings of current SM networks and to estimate the rate of change of soil properties based on two rounds of sampling. Therefore, alternative approaches are being sought.

Recent reviews (e.g. Arrouays et al., 2012; Filippi et al., 2016) of soil monitoring activities emphasize the use of legacy data that was not collected for SM purposes. Information about the rate of change of soil carbon stocks, for example, can be obtained by resampling sites where measurements had been made previously or by examining trends in databases of soil measurements such as those collected to inform agricultural management practices. However, as described in relation to purposive sampling above, the absence of a statistical design for these data sets means that biases can occur and novel statistical methodologies (Saby et al., 2017) and accompanying model assumptions are required to make appropriate adjustments.

Information on soil carbon accounting is required across spatial scales to understand both the impacts of activities within a single field and the accumulative effect across a landscape or nation. National-scale soil monitoring networks are not suitable for determining farm-scale trends, but advances have been made in the design of sampling schemes to enable efficient measurement of soil carbon trends at this scale (de Gruijter et al., 2016) and in the use of proximal sensors to predict soil carbon stocks. For example, Viscarra Rossel et al. (2017) developed a soil condition analysis system that analyses soil cores within the field and includes a γ -ray attenuation densitometer to measure bulk density and a visible-near-infrared sensor to determine soil carbon concentrations through the use of machine learning algorithms.

The potential to use remote sensing for large-scale assessments has also been explored. In contrast to the multi-year intervals between phases of a soil monitoring network, optical satellite data can be obtained approximately weekly (unless prevented by cloud cover) and used to monitor quantities such as soil moisture or soil carbon (Angelopoulou et al., 2019; Babaeian et al., 2019). The quantities of interest are often not directly measured by the sensors but inferred from other quantities in conjunction with statistical or machine learning models. The quality of the calibration of these models, which can possibly vary in both time and space, can lead to an additional component of uncertainty in the inferred trends. In addition, although there is some evidence that the soil carbon content of bare soils can be inferred from remote sensing data, vegetation, crop residues and variable soil moisture conditions can all confound the models that underlie these predictions (Environmental Defense Fund and Woodwell Climate Research Center, 2021). Smith et al. (2020) advocate a soil carbon monitoring, reporting and verification system that integrates information from short- and long-term experiments, process models, spatial data, activity data (e.g. details of farm management practices), remote sensing and soil monitoring networks.

These approaches potentially increase the temporal and spatial resolution at which the soil can be measured, but it is vital that the principles of reliable soil monitoring described in this chapter are not forgotten. Practitioners must consider whether the sensor measurements are truly representative of the relevant soil property at the appropriate spatial-scale. Care must be taken to ensure that measurements made at different times are truly comparable and the approaches used to analyze the data must be thoroughly documented.

Summary

Soil monitoring programs need to be 'fit for purpose'. At the start, a clear answer to the question 'Why are we doing this?' is needed. Considerable planning is required to achieve success. A large number of issues need to be considered and decided at an early stage and written into robust protocols. Above all, soil monitoring has considerable spatial, temporal, and cost implications, and there has to be a commitment at the decadal scale. Novel remote and proximal sensing technologies offer opportunities to observe the state of soils more frequently, but only if these basic principles underpinning reliable soil monitoring are followed.

References

- Angelopoulou T, Tziolas N, Balafoutis A, Zalidis G, and Bochtis D (2019) Remote sensing techniques for soil organic carbon estimation: A review. *Remote Sensing* 11: 676.
- Arrouays D, Marchant BP, Saby NPA, Meersmans J, Orton TG, Martin MP, Bellamy P, Lark RM, and Kibblewhite M (2012) Generic issues on broad-scale soil monitoring schemes: A review. *Pedosphere* 22: 456–469.
- Babaeian E, Sadeghi M, Jones SB, Montzka C, Vereecken H, and Tuller M (2019) Ground, proximal, and satellite remote sensing of soil moisture. *Reviews of Geophysics* 57: 530–616.
- Black H, Bellamy P, Creamer R, Elston D, Emmett B, Frogbrook Z, Hudson G, Jordan C, Lark M, Lilly A, Marchant B, Plum S, Potts J, Reynolds B, Thompson R, and Booth P (2008) *Design and Operation of a UK Soil Monitoring Network*. United Kingdom: Environment Agency.
- de Gruijter JJ, Bierkens MFP, Brus DJ, and Kottler M (2006) *Sampling for Natural Resource Monitoring*. Springer.
- de Gruijter JJ, McBratney B, Minasny B, Wheeler I, Malone BP, and Stockmann U (2016) Farm-scale soil carbon auditing. *Geoderma* 265: 120–130.
- Environmental Defense Fund & Woodwell Climate Research Center (2021) *Agricultural Soil Carbon Credits*. <https://www.edf.org/sites/default/files/content/agricultural-soil-carbon-credits-protocol-synthesis.pdf> (accessed on November 2021).
- European Commission (2002) Towards a thematic strategy for soil protection. *Official Journal COM* 179: 35.
- Filippi P, Minasny B, Cattle SR, and Bishop TFA (2016) Monitoring and modeling soil change: The influence of human activity and climatic shifts on aspects of soil spatiotemporally. *Advances in Agronomy* 139: 153–214.
- Froger C, Saby N, Jolivet CC, Boulonne L, Caria G, Freulon X, de Fouquet C, Roussel H, Marot F, and Bispo A (2021) Spatial variations, origins, and risk assessments of polycyclic aromatic hydrocarbons in French soils. *Soil* 7: 161–178.
- Kirk GJD, Bellamy PH, and Lark RM (2009) Changes in soil pH across England and Wales in response to decreased acid deposition. *Global Change Biology* 16: 3111–3119.
- Knotters M, Teuling K, Reijneveld A, Lesschen JP, and Kuikman P (2022) Changes in organic matter contents and carbon stocks in Dutch soils, 1998–2018. *Geoderma* 414: 115751.
- Lark RM (2009) Estimating the regional mean status and change of soil properties: two distinct objectives for soil survey. *European Journal of Soil Science* 60: 748–756.
- Loveland PJ (1990) The National Soil Inventory: Survey design and sampling strategies. In: Lieth H and Markert B (eds.) *Element Concentration Cadasters in Ecosystems*, pp. 73–80. Weinheim, Germany: VCH Verlagsgesellschaft.
- Loveland PJ and Bellamy PH (2005) Environmental monitoring. In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*. Elsevier.

- Oliver MA, Archer JR, Baxter SJ, Todd AD, and Skinner RJ (2006) The Representative Soil Sampling Scheme of England and Wales: A statistical analysis of topsoil pH and nutrient status between 1971 and 2001. *Soil Use and Management* 22: 372–382.
- Panos Panagos P, Montanarella L, Mirco Barbero M, Annette Schneegans A, Laura Aguglia L, and Jones A (2022) Soil priorities in the European Union. *Geoderma Regional* 29: e00510.
- Saby NPA, Bellamy PH, Morvan X, Arrouays D, Jones RJA, Verheijen FGA, Kibblewhite MG, Verdoodt A, Berényi Úveges J, Freudenschub A, and Simota C (2008) Will European soil-monitoring networks be able to detect changes in topsoil organic carbon content? *Global Change Biology* 14: 2432–2442.
- Saby NPA, Swiderski C, Lemerrier B, Walter C, Louis BP, Eveillard P, and Arrouays D (2017) Is pH increasing in the noncalcareous topsoils of France under agricultural management? A statistical framework to overcome the limitations of a soil test database. *Soil Use and Management* 33: 460–470.
- Schroeder J, Lisa Kammann L, Mirjam Helfrich M, Christoph C, Tebbe CC, and Christopher Poeplau C (2021) Impact of common sample pre-treatments on key soil microbial properties. *Soil Biology and Biochemistry* 160: 108321.
- Smith P, Soussana JF, Angers D, Schipper L, Chenu C, Rasse DP, Batjes NH, van Egmond F, McNeill S, Kuhnert M, Arias-Navarro C, Olesen JE, Chirinda N, Fornara D, Wollenberg E, Álvaro-Fuentes J, Sanz-Cobena A, and Klumpp K (2020) How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. *Global Change Biology* 26: 219–241.
- van Leeuwen JP, Saby NPA, Jones A, Louwagie G, Micheli E, Rutgers M, Schulte RPO, and Spiegel H (2017) Gap assessment in current soil monitoring networks across Europe for measuring soil functions. *Environmental Research Letters* 12: 124007.
- Viscarra Rossel RA, Craig R, Lobsey CR, Sharman C, Flick P, and McLachlan G (2017) Novel Proximal Sensing for Monitoring Soil Organic C Stocks and Condition. *Environmental Science and Technology* 51: 5630–5641.
- Yu H, Liu Y, Shu X, Ma L, and Pan Y (2020) Assessment of the spatial distribution of organochlorine pesticides (OCPs) and polychlorinated biphenyls (PCBs) in urban soil of China. *Chemosphere* 243: 125392.

Soil quality and health key indicators

Edoardo AC Costantini^a and Simone Priori^b, ^aCNR-IBF—Department of Biology, Agriculture and Food Sciences, Sesto Fiorentino, Florence, Italy; ^bDepartment of Agriculture and Forest Sciences (DAFNE), University of Tuscia, Viterbo, Italy

© 2023 Elsevier Ltd. All rights reserved.

Introduction	181
From soil quality to soil health	182
Different approaches on soil indicators: Quantitative and qualitative, field and laboratory analyzes, visual assessments	183
Selecting indicators	184
Soil indicators across spatial and temporal scales	185
Most common soil quality and health indicators	186
Water supply	186
Oxygen supply	188
Nutrient supply	188
Regulation of runoff	189
Sediment production	189
Groundwater recharge	189
Purifying capacity and regulation of contaminants	189
Carbon sequestration	189
Biological activity and cycling of organic matter and nutrients	190
Biodiversity pool	190
Conclusions	191
References	191

Abstract

The concepts of soil quality and soil health concern the ability of soil to function and provide ecosystem services. This chapter reports a historical and critical overview of the two terms and a review of the methods and indicators currently used to assess and monitor soil quality and health. Different approaches are described and guidelines on the selection and use of indicators across spatial and temporal scales are provided. Key indicators are introduced, grouped according to their relevance to provide 10 specific soil functions, with their main references, sensitivity to changes, and relative cost of sampling and analysis.

Key points

- The passage from the concepts of soil fertility to soil quality and soil health.
- Main principles in the use and selection of soil indicators.
- Considering spatial and temporal variations.
- Most commonly used soil quality and health indicators.

Introduction

Water and air quality are usually assessed considering their degree of pollution, as it impacts directly the health of humans and animals, and natural ecosystems. The soil quality concept is more complex, not only because soil encompasses solid, liquid, and gaseous phases, but also because soils can be used for a larger variety of purposes. Soil quality is not limited to the degree of soil pollution but is commonly defined much more broadly as “the capacity of a soil to function within ecosystem and land-use boundaries to sustain biological productivity, maintain environmental quality, and promote plant and animal health” (Doran and Parkin, 1994 in Bünemann et al., 2018). The quality of soils determines agricultural and forestry sustainability, environmental quality and, as a consequence, plant, animal, and human health. Changes in soil quality over time can be due to natural events or, more often, to human use. Assessment of soil quality by the use of indicators is then mandatory to inform policymakers, farmers, and citizens about the impact of human activities on soil and its functions.

From soil quality to soil health

Historically, the term “soil quality” has been strictly related to the concept of land evaluation to support primary production. Indeed, the soil ecosystem service that was first considered by humanity was the production of food, feed and fibers, therefore assessments of the suitability of a soil for crop growth may have been taken into account since the origin of agriculture. Before the introduction of the term soil quality, the concept of “soil fertility” was used to describe the capacity of the soil to provide water and nutrients for plants. The FAO (Food and Agriculture Organization) described soil fertility as “the ability of the soil to supply essential plant nutrients and soil water in adequate amounts and proportions for plant growth and reproduction in the absence of toxic substances which may inhibit plant growth” (<http://www.fao.org>).

Although soil fertility is a basic concept that refers to the function of biomass production, the soil provides several other functions, important for ecosystems and human well-being. Another term has been then introduced, namely “soil functionality,” which means the ability of a soil to provide functions and ensure the provision of the various ecosystem services. The European Commission published the communication “Towards a Thematic Strategy for Soil Protection” (COM 2006.231) where seven soil functions have been defined: (i) production of food and biomass, (ii) storage, filtering and transformation of compounds, (iii) habitats for living creatures and gene pools, (iv) physical and cultural environment, (v) source of raw materials, (vi) carbon pool, and (vii) archive of geological and archeological heritage.

To summarize, soil properties determine the soil functionality, that is the ability to produce functions that grant ecosystem services. The concept of “soil quality” is closely related to functionality, but it emphasizes the multi-functionality behavior of the soil, and it is related to land use. The better is the soil quality, the greater are the functions that the soil can perform optimally.

The need to assess the quality of soils at a certain scale, from continental scale to single farm, is mainly linked to the monitoring of the effects of human activities on soil resources. In the history of humankind, humans have often neglected the soil more than other natural resources, and this lack of awareness resulted in failures and the end of civilizations, namely the Maya in Central America, Vikings in Iceland, Rapa Nui people in Easter Island, and Mesopotamian civilizations. More recently, the catastrophic failures in land management of the 1930s in the United States began with ignorance of the Great Plains’ soil resource, which was described as “indestructible and immutable” in the 1909 Bureau of Soils Bulletin.

Efforts to define and assess soil quality can be traced to publications from the 1960s to the 1970s. One of the oldest mentions in the scientific literature is by Klingebiel and Montgomery (1961), who defined soil quality as “an attribute of a soil that cannot be measured directly, but which is deduced from soil properties and soil behavior under defined conditions. Fertility, productivity, and erodibility are examples of soil qualities.” Later, Mausel (1971, in Bünemann et al., 2018) when soil mapping in Illinois (United States), defined soil quality as “the ability of soils to yield corn, soybeans, and wheat under conditions of high-level management.” It is clear that, historically, the concept of soil quality matched with “land capability,” defined as the intrinsic capacity of a soil to contribute to agricultural production. The high-quality soils or “prime farmlands” were the soils of the first two classes of land capability, while the last classes were those only suitable for more or less productive forestry and pasture or just for ecological purposes (Klingebiel and Montgomery, 1961).

Doran and Parkin (1994 in Bünemann et al., 2018) considered the focus on agricultural productivity to be too restrictive and proposed that soil quality should be disconnected from productivity. Besides agricultural productivity, the soil quality concept should include the ability of soils to contribute to environmental quality and to promote plant, animal, and human health. In 1993, the U.S. National Research Council (NRC) published an agenda for U.S. policymakers, titled “Soil and water quality: an agenda for agriculture”, which stressed the importance of soil quality and its application in land management. The agenda reported that national policies should not only focus on controlling erosion and conserving agricultural productivity, but they should also consider other threats to soil quality, namely salinization, compaction, acidification, and loss of biological activity. In the same year, the U.S. NRC also established the Soil Quality Institute (SQI), with the mission to develop tools for soil quality assessment. During the 1990s, one of the first methods used to assess soil quality was through the development and use of soil quality scorecards. In 2001, the U.S. Department of Agriculture published the first guidelines for “Soil Quality Assessment in Conservation Planning,” where a series of indicators and a field record were proposed.

Sojka et al. (2003) (in Bünemann et al., 2018) criticized the “soil quality” paradigm, expressing dissatisfaction with the idea of a universal soil quality index and with the formal adoption of soil quality evaluation. They also affirmed that soil quality lacks standards and oversimplifies the complex interactions occurring in soil, therefore the concept of “quality of soil management” should be preferred. Although their proposal did not find much support in the scientific society, their criticisms have influenced the development of soil quality evaluation, in which soil management has become central. Soil quality assessment should provide scientific tools to evaluate the management of soil resources, considering not only biomass production but also other ecosystem services provided by the soils (Bünemann et al., 2018).

The original concept of soil quality is continuously evolving and recently has often been replaced by “soil health,” to emphasize the behavior of the soil to be a dynamic living resource. Some authors consider the terms “soil health” and “soil quality” to be equivalent and interchangeable, while others differentiate the meaning of the two words. Soil quality is related to soil functions or what it does, while soil health presents a holistic view of the soil as a dynamic living resource (Lehmann et al., 2020). Soil health has also been illustrated as an analogy to the health of an organism or a community. More specifically, soil health has been defined as the “capacity of soil to function as a vital living ecosystem that sustains biological productivity, maintains environmental quality and promotes plant, animal, and human health” (Doran and Zeiss, 2000). Recently, the FAO Technical Group on Soils (ITPS)

defined soil health as “the ability to sustain productivity, diversity and environmental services of terrestrial ecosystems” (<http://www.fao.org>).

Soil health should be maintained, promoted, or recovered through the implementation of sustainable soil management practices. In recent years, several foundations with the mission to promote, safeguard, and enhance soil health through scientific research and advancement have developed in several countries of the world, often supported by their governments. Policymakers have increased awareness of soil health and its importance for the environment and human health. In addition, some major companies are starting soil health programmes to manage their supply chains more sustainably.

In 2012, member countries of the FAO established the Global Soil Partnership (GSP) as a collaborative mechanism to ensure soil governance and promote sustainable soil management (SSM). The GSP is developing technical and policy tools to adapt the principles of sustainable soil management to local needs and stakeholders, as well as to encourage investment in technical cooperation, education, and awareness (<http://www.fao.org>). After the International Year of Soils 2015 and the revision of the World Soil Charter, the GSP published the Voluntary Guidelines for Sustainable Soil Management (VGSSM). The VGSSM has been aimed at being a reference that provides general technical and policy recommendations on SSM for a wide range of stakeholders, including governments, forest and land managers, agricultural advisors, farmers, civil society and academia. In 2020, FAO-ITPS (<http://www.fao.org>) published the “Protocol for the Assessment of Sustainable Soil Management,” where they reported the recommended indicators and the approaches to evaluate SSM. A farmer-to-farmer training system, called the “Global Soil Doctors programme” was also developed by the GSP in 2020. The programme aims to educate farmers on sustainable soil management and the preservation of soil health through a set of tools including educational material, a soil testing method (STM), and a soil testing kit (STK) for preliminary soil analyzes (<http://www.fao.org>).

The European Union has also adopted the term soil health. In a recently published document entitled “Taking Care of the Soil is Taking Care of Life” soil health is defined as “the continuous capacity of soils to provide ecological functions for all forms of life.” The document puts soil at the center of the actions to be taken to achieve the sustainable development goals set by the European Green Deal (https://ec.europa.eu/info/publications/caring-soil-caring-life_en).

Different approaches on soil indicators: Quantitative and qualitative, field and laboratory analyzes, visual assessments

Assessment of soil quality and health indicators (SQI, SHI), and monitoring the direction of change with time, are fundamental to determine the sustainability of land management and to verify the effectiveness of the strategies for conservation management practices. An important aspect to take into account for soil quality assessment is the identification of a set of indicators or attributes that are sensitive to changes in land management, which influence the capacity of a soil to fulfill its functions. Such indicators can be assessed using qualitative or quantitative techniques and should be easy to measure as well as sensitive to variations in climate and management (Bünemann et al., 2018). Approaches that target farmers and stress the educational aspect usually involve qualitative or semi-quantitative indicators, which can be easily assessed in the field, deliver rapid results, and facilitate communication between farmers and scientists (Fig. 1).

Standardization of the methods to collect and group indicators is also fundamental, although monitoring soil quality is usually planned with specific goals and areas of the study. For example, specific studies on soil quality have been made to evaluate the effects of tillage types in croplands, different types of agricultural management of extensive agricultural areas, land-use change, and

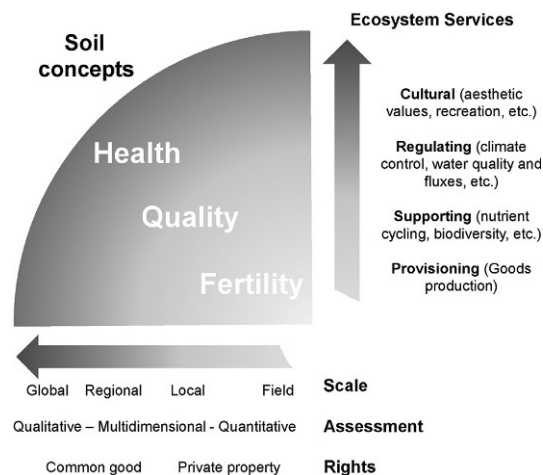


Fig. 1 Graphical sketch to summarize how the different soil concepts, namely soil health, quality, and fertility, are affected by the scale and methods of assessment. On the right, how the different concepts are relevant for soil ecosystem services. The figure was inspired by the sketch of Lehmann et al. (2020).

urban soils. The reliability of SQIs mainly depends on their ability to represent the spatial and temporal variations within a certain study area. Therefore, it is difficult to find a universal approach to evaluate soil quality worldwide and in all situations. Increasing awareness of soil resources and national regulations have catalyzed a proliferation of various SQI and SHI, guidelines, and scorecards for field surveys. Among the different approaches, qualitative methods based on direct visual assessment of field indicators are the most rapid, easy, and cost-effective. One of the most widely used worldwide is the Visual Soil Assessment (VSA), proposed by the FAO to assess soil quality worldwide, which is based on the visual appraisal of key soil “state” and plant performance indicators. The VSA guidelines have been developed for general land cover categories (annual crops, pastures, orchards) and some specific crops (maize, olive orchards, vineyards, wheat) (<http://www.fao.org>).

Another widely used visual method is based only on soil physical characteristics, and it is called Visual Evaluation of Soil Structure (VESS) (Franco et al., 2019). This method is included in the “spade methods” as it involves the sampling of an undisturbed soil slice 0.25-m deep, 0.1-m thick and 0.2-m wide using a straight spade. The VESS scorecard includes the evaluation of structure quality, consistency, aggregate-size distribution, porosity, and roots frequency.

Visual methods are important because they can be included among the techniques of citizen sciences, as they can be executed by non-experts of the soil. On the other hand, visual soil assessment alone cannot fully evaluate the state of ecosystem services driven by biological, chemical, and hydrological soil processes. A combination of visual assessment and low-cost laboratory analysis is needed to improve soil quality determination.

In the United States, the Cornell Soil Health Test offers various soil health testing packages for farmers and policymakers. Furthermore, the U.S. Department of Agriculture developed the “Cropland In-Field Soil Health Assessment Guide” as a qualitative diagnostic tool for field use to help U.S. conservation planners. The guide includes 11 indicators that can be easily monitored in the field to assess the impact of soil management on croplands.

In addition to the qualitative methods described above, quantitative methods to assess soil quality have also been proposed. To capture the overall state of the soil, quantitative methods usually have complex and multivariate approaches, such as “least square regression” models. In some cases, the measurement of a specific soil feature can be relatively time-consuming and becomes impractical to analyze in the laboratory. In that case, the literature has reported several pedotransfer functions (PTFs) to estimate some soil parameters through the measurement of basic soil properties, which are easier to analyze and are often available in the soil databases. The PTFs are frequently used to estimate the hydrological properties of soils, in particular, saturated permeability (K_{sat}) and available water capacity (AWC), but others have been developed to estimate many other properties, like oxygen diffusion coefficient and cation exchange capacity (CEC).

Importantly, the soil classification and taxonomic systems, such as the World Reference Base (IUSS Working Group WRB, 2014), and Soil Taxonomy (Soil Survey Staff, 1999) are not only based on soil genetic processes but also characteristics related to soil fertility and quality. Soil evaluation based on soil classification has been conducted for many purposes, such as assessment of soil productivity (Eswaran et al., 1997, see Bünemann et al., 2018), cadastral quality and value of soils (Sapozhnikov and Granina, 2021), as well as soil quality in urban environments (Makki et al., 2021). Therefore, the type of soil can be used as an indicator of soil quality, with different possible interpretations according to the detail of the taxonomic level. Soil classification systems usually focus on the subsurface horizons in addition to the topsoil, therefore, soil types are better indicators of slowly changing environmental conditions.

Selecting indicators

Many frameworks exist for soil quality and health assessments, and they are characterized by various indicators and their aggregation. Such indicator frameworks can be divided into three groups concerning their main focus: (i) assessing soil quality of a certain territory, based on detailed field measurements or by soil database and associated models; (ii) comparing the impact of different soil management systems; (iii) elaborating on the status of specific soil threats. In addition, the selection of indicators should be made based on the type of land of interest: croplands, forests, polluted soil, etc. A review by Bünemann et al. (2018) reported the frequency of the indicators used in the literature. The most commonly used to assess soil quality and health are soil organic carbon/matter, pH, available phosphorus and potassium, total nitrogen, water storage, and bulk density. Other indicators are less commonly used because of the difficulty and/or costs of the measurement, for example, labile organic carbon, soil respiration, microbial biomass, earthworms, and micronutrient availability. Further indicators are used to assess the state of specific threats, for example, sodicity–salinity for salinization processes, or presence of contaminants. The technical literature reports a long list of possible soil indicators, and there is no general agreement about the ones to use in the different national legislations and policies. In some cases, indicators are grouped in a comprehensive soil quality index equation, after the application of weight factors provided by the experts.

A possible strategy is the setting of a minimum dataset to overcome problems like costs of measurements, collinearity, as well as excessive complexity of the relationships between indicators and management options (Bünemann et al., 2018). Selection can be carried out through a statistical data reduction by multivariate techniques such as principal component analysis (PCA), redundancy analysis (RDA), and discriminant analysis. A selection of a minimum dataset of soil indicators is also recommended by the FAO, the Soil Quality Institute, and the EU expert panel, while larger datasets are advised for specific purposes.

Interestingly, some methodological approaches include features of soil health indicators that are not directly related to soil characteristics. For example, the FAO protocol includes the evaluation of soil productivity through the weight of the total biomass or

an estimate of the dry biomass per unit area. The EU expert panel includes vegetation cover and soil sealing in the minimum dataset, because of their direct impact on soil health in farmland, forestry, and urban settings. The Implementation Plan of the Soil Deal for Europe also considers landscape heterogeneity and forest cover (<https://ec.europa.eu/>). The inclusion of landscape features within soil indicators shows the attempt to use indirect traits of soil health that can be monitored through remote sensing, such as satellite images. A recent approach in assessing the effectiveness of management strategies is by combining analysis of the spatial pattern of vegetation with qualitative soil surface indicators. This simplified but effective methodology allows the monitoring of variation in landscape functioning in space and time, and it is particularly suitable for the assessments carried out at intermediate territorial scales (Costantini et al., 2016).

Soil indicators across spatial and temporal scales

The soil information needed to assess and monitor soil quality and health depends on the spatial and temporal scales. The area of consideration can be punctual like a site on a landscape, a small research plot, etc., or as large as a nation or a continent. The reliability of broader-scale assessment will depend strongly on the quality of the database in terms of the spatial distribution of observations and quality of the data, as well as on the efficiency of models to interpolate the data in the area of interest. Two approaches that deal with spatial scale have usually been used by researchers (Fig. 2). The first is based on the selection of the most sensitive set of indicators, the method of their aggregation, and the spatial distribution, for the geographic scale for which soil quality assessment is performed. In general, broader scales will include easier, more economic, and less precise indicators than detailed scales. In this approach, the quantity of data is more important than the quality, because the interpolation is data-driven. The second approach is based on precise and detailed punctiform measurements that can be expanded to a larger area of investigation by combining interpolation and extrapolation models. This approach is more costly but provides more detail on

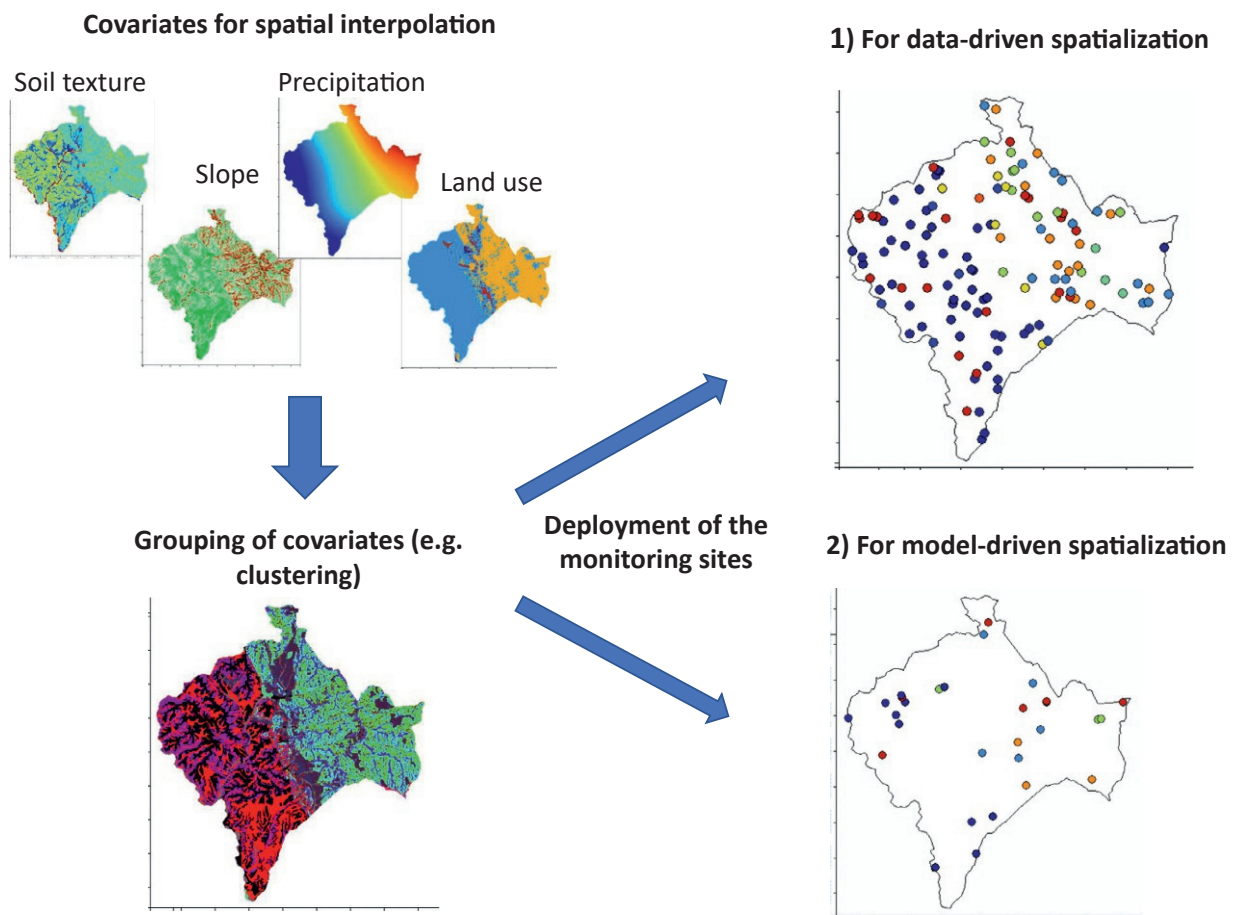


Fig. 2 Deployment of the monitoring sites. After the selection of the covariates for spatial interpolation, they are grouped and used for the deployment of the monitoring sites. The spatialization can then follow two alternative ways: data or model-driven. The first one is more suitable for a large number of easy to measure and cheap indicators, representative of the territory. The second way is for precise and sensitive (e.g., biological) indicators, when they are strongly related to one or more covariates of interpolation (e.g., land use).

the investigation at the monitoring sites. In this case, the extrapolation is model-driven, therefore the selection and calibration of the model to spatialize the information are fundamental to obtain reliable soil quality assessment of a certain area.

The selection of monitoring time frames is also fundamental for soil quality assessment because the effect of climate, soil moisture conditions, stage of plant growth, and human factors can cause great temporal variability in indicator status. Biological and biochemical indicators, in particular, show strong temporal variability, mainly due to fluctuations of soil temperature and moisture. The response time for an individual indicator to change, as a result of management, determines the appropriate time interval for monitoring indicator changes.

Indicators with a strong sensitivity to rapid changes (Table 1) may be used as either an “early warning system” of degradation, or to assess the effectiveness of soil management changes at the field scale. Different management practices affect different aspects of soil functionality, especially indicators associated with labile organic carbon, enzymatic activity, and biodiversity. For instance, changes in tillage intensity appear to have greater effects than organic vs. conventional practices (van Es and Karlen, 2019).

It must be also considered that for the soil system many natural or anthropic interventions can be seen as “catastrophic events,” in the sense that they largely disturb and disrupt soil characteristics and functionalities. Exceptional events, like flooding and consequent accumulation of sediment, aeolian erosion and deposition, slumps, earth slides, or gully erosion, can result in truncation of the soil profile or its burial. Similarly, anthropic activities such as land leveling, slope reshaping, and deep ploughing can profoundly modify the soil profile and its quality (Costantini et al., 2018). In these cases, which are increasingly frequent in both space and time, there is little value in considering the sensitivity of soil quality indicators because they might all change abruptly. On the other hand, at the landscape spatial scale, it is critical to consider indicators that focus on long-term changes, to overcome the short-term variation in land use and management. Soil indicators that change slowly are more related to inherent soil qualities and can reflect the impacts of climate and long-term management changes (Pellegrini et al., 2018).

The placement of monitoring sites should reflect the soil spatial and temporal variability of the indicators used. This is perhaps the most delicate and professional part of the work. The areas selected must be representative of both the soil and management practice to be assessed. A more accurate delimitation of the study area within a management type can take place based on the characterization of soil through proximal sensing by, for example, geoelectric, spectrophotometric and radiometric sensors. It is crucial to make comparisons within the same type of soil. The great variety of soil properties, even within a limited territory, means that the values of soil indicators cannot be compared with those at a different soil site. The risk is that of arriving at wrong or approximate evaluations, as a consequence of inadequate sampling. Correct use of the indicators passes from a comparison with the measurements carried out on the same area before starting the testing of the management practice, or on similar and nearby areas that have not received the treatment studied but have the same soil type.

Most common soil quality and health indicators

The main soil indicators can be grouped according to their relevance to provide specific functions. In this chapter, we focus on the soil indicators that are related to soil quality and health, are affected by management, and sensitive to variations. The indicators related to soil functions that are not directly related to soil management, namely the ability of the soil cover to support human activities and infrastructures (bearing capacity) or to be used as a source of raw materials, the role of an archive of geological and archeological heritage, its cultural and esthetic values, are omitted. Among the indicators described below, those that are most relevant and commonly used in monitoring programmes are reported in Table 1.

Water supply

One of the most important soil functions is to provide and regulate the water uptake by plants and other organisms that live in the soil. Soil degradation very often goes together with the impairment of this paramount function. The potential water supply can be assessed and monitored in different ways, but the most common soil indicators are available water capacity, rooting depth and volume, bulk density, and soil salinity.

The most common indicator used to estimate the amount of potentially available soil water for plants is the so-called “available water capacity” (AWC), which refers to the water held in the soil between “field capacity” (matrix potential of fine earth (<2 mm) at -33, -10 kPa for sandy soils) and “permanent wilting point” (matrix potential of -1500 kPa). The easily available water is the fraction of AWC held between -10 and -200 kPa, which is more affected by temporal variation and soil physical conditions, in particular macroporosity and structure. In addition, the study of the relationship between water tension and volume (water retention curve) can provide useful indicators of the soil structural quality. Many times, AWC is estimated through particle size distribution.

The amount of water that is potentially available for plants depends on the soil volume that can be explored by roots. Rooting volume is based on multiple factors. First, rooting depth, the distance between the soil surface and a horizon or layer preventing root penetration, such as a consolidated substrate, a cemented pedogenetic horizon, a layer very rich in salts, or a water table. Second, there is the need to subtract the volume of soil that cannot be penetrated by roots, namely the quantity of unaltered skeleton, very firm clods, and the parts of the soil profile volume that show very marked compaction (Piori et al., 2021).

Soil penetration resistance can be useful for estimating the compaction of soil horizons. It indicates root-impeding layers, such as hardpans or dense soil layers, and can be used to compare relative strengths among similar soil types. Moreover, field observation of soil consistence and root growth pattern can give useful hints on where to place the measurements.

Table 1 Key soil functionality indicators.

<i>Soil indicator</i>	<i>Sensitivity to changes^a</i>	<i>Method of evaluation and analysis</i>	<i>References</i>	<i>Cost of sampling and analysis^b</i>
Available water capacity	Low	Sandbox and pressure-plate extractors	Cassel and Nielsen (1986)	Medium
Rooting depth and volume	Low	Visual assessment, bulk density, penetration resistance	Priori et al. (2021)	Low
Bulk density	Medium	The core method	Arshad et al. (1997)	Low
Penetration resistance	Medium	Penetrometer	Herrick and Jones (2002)	Low
Wet aggregate stability	Medium	Wet sieve procedure	Le Bissonnais (1996)	Medium
Hydraulic conductivity	Medium	Two-ponding head method; disc permeameter; pedotransfer functions	Reynolds et al. (2000)	Medium
Infiltration capacity	Medium	Double ring infiltrometer; disc permeameter	Lowery et al. (1997)	Medium
Crusting susceptibility	Medium	Cone penetrometer; pedotransfer functions	ASAE (1994)	Low
Erodibility	Low	Pedotransfer function	Torri et al. (1997)	Low
Total nitrogen	High	Micro-Kjeldahl determination; dry combustion	FAO-GSP-GLOSOLAN (2021)	Medium
Potentially mineralizable nitrogen	High	Anaerobic incubation	Drinkwater et al. (1997)	Medium
Available phosphorus	Medium	Olsen, Bray, and Mehlich methods using sodium bicarbonate or other extractants	FAO-GSP-GLOSOLAN (2021)	Medium
pH	Low	Electrode system	FAO-GSP-GLOSOLAN (2021)	Low
Cation exchange capacity (CEC)	Low	Sum of cations on the exchange complex	Sikora and Moore (2014)	Medium
Base saturation of the CEC complex	Low	Calculation—percentage of bases on the total exchangeable cations	Sikora and Moore (2014)	Low
Sodicity—Exchangeable Sodium Percentage (ESP) of the CEC	Low	Calculation—percentage of sodium on the total exchangeable cations	Sikora and Moore (2014)	Low
Salinity—Soil Electrical Conductivity (EC)	Medium	Standard electrical conductivity meter system on saturated paste of different soil: water ratios	FAO-GSP-GLOSOLAN (2021)	Low
Total carbonate content	Low	Gas-volumetric method using Dietrich-Fruhling Calcimeter	Sparks (1996)	Medium
Thickness of the Ah horizon	Low	Visual assessment	Jahn et al. (2006)	Low
Organic carbon	Medium	Walkley-Black titration and colorimetric method; dry combustion	FAO-GSP-GLOSOLAN (2021)	Medium
Labile organic carbon	High	Hot-water extractable carbon; particulate organic matter; active carbon	Weil et al. (2003 in Bünemann et al., 2018)	Medium
C:N ratio	Medium	The ratio between organic carbon and total nitrogen		Medium
Microbial biomass carbon	High	Substrate induced respiration (SIR); chloroform fumigation extraction (CFE)	Alef and Nannipieri (1995)	High
Microbial respiration rate	Medium	Laboratory-based soil respiration measurement (static or dynamic)	Bastida et al. (2008 in Bünemann et al., 2018)	High
Enzymatic activity: <i>N</i> -acetyl- β -glucosaminidase, β -glucosidase, butyrate esterase, acid phosphatase, arylsulphatase, β -xylosidase, cellulose and acetate esterase	High	Incubation and fluorescence	Marx et al. (2001)	High
DNA and RNA profiling	High	RNA genes amplified via Polymerase Chain Reaction (PCR), analyzed by Denaturing Gradient Gel Electrophoresis (DGGE)	Kowalchuk (2004)	High
Phospholipid fatty acids	High	Solid-phase extraction and chromatography	Quideau et al. (2016)	High
QBS-ar index	Medium	Berlese–Tullgren extractor	Parisi et al. (2005 in Bünemann et al., 2018)	Medium

^aExcluding intense soil perturbation, such as anthropic earthworks, strong manuring, liming and additions of other materials. High = day and single management variations; medium = seasonal and main year management variations; low = long term variations, land-use changes.

^bHigh = specialistic analysis with skilled personnel and/or high costs; medium = routine wet laboratory analysis; low = low-cost laboratory analysis, field assessment.

Soil bulk density is an indicator of the porosity of the soil mass, therefore of the maximum volume that can be occupied by water and air. It is also called apparent volumetric mass, to differentiate it from the real volumetric mass, or the real density, which depends on the densities of the soil minerals and organic components. It is calculated as the dry weight of the soil divided by its volume, including both the volume of soil particles and the pores. Particular care must be taken to exclude gravels from the computation. Bulk density not only reports the quantity of voids in the soil mass but also reflects the degree of aggregation of the soil particles (the soil structure). Therefore, it is strictly related to the ability of plant roots to penetrate into the soil mass, and also of the soil to function for structural support, water and solute movement, and gas exchanges.

Soil salinity is the accumulation of water-soluble salts in the soil profile, mainly chloride, sulfate, carbonate and bicarbonate of sodium, magnesium, calcium, and potassium. Salinity increases the osmotic potential of the soil water solution, which limits the plant's capacity to absorb water. Salinity can be estimated using electrical conductivity (EC).

Oxygen supply

Most of the organisms living in soil and plant roots need good aeration and the presence of oxygen in the soil atmosphere. Good soil health and functionality go together with a large presence of interconnected macropores, letting the soil atmosphere, which is richer in carbon dioxide, to be oxygenated with atmospheric air. The kind of macropores that ensure an effective diffusion of oxygen within the soil mass are those produced by the biological activity of soil fauna, flora, and roots. The soil indicators used to assess potential oxygen supply are mainly related to soil structure. Apart from wet aggregate stability, particle size, bulk density, and air capacity are among the most utilized. Soil porosity can also be estimated by visual assessment or by the micromorphometric method.

Oxygen supply is also related to sodicity because excess sodium can lead to the destruction of soil structure through the swelling and dispersion of clay particles and the formation of low permeability layers that restrict root growth. Clay dispersion also reduces water infiltration throughout the soil profile and causes waterlogging, clogs drainage pipes, reduces the bearing capacity of the soil (restricting the use of machinery to work the soil), and enhances surface and subsurface soil erosion.

There are also good macro- and micro-morphological methods to assess the state of soil oxygenation. In particular, the field assessment of soil structure and consistence, as well as the study in the field and under the microscope of porosity and redoximorphic features, are considered in the most important soil classification schemes. In the WRB classification system (IUSS Working Group WRB, 2014), in particular, the presence of more or less expressed stagnic properties indicates that the infiltrated water cannot be drained at depth and saturates the soil profile. Gleyic properties, instead, refer to the presence of shallow groundwater for most of the year when soil temperatures permit biological activity.

Nutrient supply

Plants need nutrients in different supply ranges, which are usually higher for productive crops. Among basic macronutrients (nitrogen (N), phosphorus (P), potassium (K), sulfur (S), magnesium (Mg) and calcium (Ca)), total nitrogen and available phosphorus are the most frequently suggested in soil quality and health monitoring schemes (EJP-SOIL, 2021). Plant available phosphorus, in particular, is often the most limiting nutrient for crop and forage production. It is used to monitor chemical soil fertility as it is a stable element, and its mobility in the soil is limited. In addition, potentially mineralizable nitrogen, which represents the fraction of nitrogen that can be easily decomposed by soil microorganisms, is considered in monitoring the effects of soil management on nitrogen availability during the growing season and as an indirect measure of biological activity.

Minor nutrients (molybdenum (Mo), copper (Cu), zinc (Zn), manganese (Mn), iron (Fe), nickel (Ni), boron (Bo), selenium (Se), and chlorine (Cl)), although essential for plant growth, are only considered when symptoms of deficiency in plants have been recognized.

Other main soil indicators of nutrient supply are pH, cation exchange capacity, base saturation, and soil mineralogy. Soil pH gives an important indication of the availability of plant nutrients, and different crops thrive at different pH values. Soil pH may change in response to management activities such as liming, fertilizer addition, irrigation that leads to the accumulation of salts, and some forms of soil pollution. Soil pH controls the solubility and mobility of heavy metals, such as Al (aluminum), Fe, Mn, Cu, and Zn, and nutrients, such as phosphorus. It also controls the toxicity of many heavy metals.

The CEC of soil represents its capacity to hold and exchange positively charged cations, in particular, H, Al, Ca, Mg, K, and Na. The CEC is a characteristic of the soil that is influenced by both very stable (soil mineralogy) and dynamic (organic carbon) properties and affects the soil's retention of fertilizers and other ameliorants. The base saturation of the CEC depends on the presence of Ca, Mg, K, and Na, in contrast to the acidic saturation of H and Al, and it is the basis for cation availability and the buffering capacity of soil pH.

Some indicators highlight the proportion of different cations on the exchange complex (e.g., K/Mg, Ca/Mg). This is particularly important for Na^+ , as an excess of it can negatively alter cell morphology, plant photosynthesis, and chlorophyll production. The exchangeable sodium percentage (ESP) is the amount of Na^+ adsorbed on the surfaces of the soil particle as a proportion of the CEC, while the sodium absorption ratio (SAR) is the relative concentration of Na^+ to Ca^{2+} and Mg^{2+} .

The contents of total and active calcium carbonate are important in regulating the supply of nutrients for plants and microorganisms, and their abundance may decrease the iron uptake, but can also interfere with the supply of water and oxygen. The quantity and activity of carbonates can vary according to soil management, in particular the quality and quantity of irrigation water. Degraded soils may show higher values of carbonates (Costantini et al., 2018).

Regulation of runoff

Soil surface characteristics regulate water runoff and soil in good health can limit runoff by increasing infiltration, while degraded soils on slopes show increased runoff. Soil indicators related to runoff are particle size, surface stoniness, infiltration capacity, and crusting susceptibility. The presence on the surface of gravel and stones acts as a mulch, protecting the surface aggregates from raindrop impact, favoring water infiltration, and reducing runoff. Soil infiltration rate is a measure of how fast water enters the soil under saturated or non-saturated conditions. Water entering too slowly can produce ponding conditions and reflect soil compaction, thus enhancing surface runoff and water erosion. The susceptibility of soil crusting depends on the soil characteristics, such as fine and coarse silt, clay, and organic matter content. Both soil infiltration and crusting susceptibility are strongly influenced by porosity and aggregate stability.

Sediment production

Soil characteristics regulate the production of sediments through erosion. Soil erosion can be caused by wind, water, or anthropogenic activities such as soil tillage. The volume of sediments detached from the soil can be measured in the field with various tools and methods, or by observing visible evidence of soil losses, or can be estimated through laboratory experiments or models that consider soil, climate, morphology, land use and management.

Soils regulate the production of sediments delivered to water bodies through the susceptibility to erosion. The main indicator is soil erodibility. Soil erodibility is the degree of resistance to the impact of raindrops on the soil surface and the shearing action of runoff water. Similar to crusting susceptibility, it can be considered a soil property that depends on soil texture, organic matter content, and aggregation. Other soil indicators related to erosion risk are surface roughness and stoniness, hydrophobicity, but also the content of soluble salts and sodicity in the soil profile, which can cause the formation of erosion tunnels under the soil surface.

Groundwater recharge

An important soil function is the ability to allow the infiltrating water to replenish the groundwater. The process is regulated by the capacity of the soil to infiltrate and hold water and let it percolate throughout the entire profile. Key factors are the depth of groundwater table and the presence of soil horizons that impede vertical movement (e.g., massive clayey layers, thick petrocalcic and petroferic horizons), or create preferential subvertical (e.g., horizons with vertic properties) or subhorizontal flows (e.g., fragic horizons).

The main soil indicators to be considered are infiltration capacity, AWC and hydraulic conductivity of the most limiting horizon of the soil profile.

Purifying capacity and regulation of contaminants

Soils in good condition are an essential filter for clean water and are low in contaminants. Water that moves through the soil is cleaned by physical, chemical, and biological processes. Small pores mechanically filter pollutants and, within pores, the negatively charged soil components, mainly organic matter and clays, capture chemicals with a positive charge, such as many toxic chemicals and ammonium. The bases held on the exchange complex contribute to neutralizing the soil solution, while pH regulates the chemical behavior of toxic elements. Soil organic matter can also have positive charges able to capture some negatively charged chemicals, such as nitrate. Soil bacteria and fungi may transform the pollutants into non-toxic substances such as carbon dioxide and water.

Although the purifying capacity of soil is a complex process, the main indicators used are cation exchange capacity, base saturation, pH, organic matter content, and rooting depth and volume. Indicators of the biodiversity pool are also relevant, but since it is difficult to relate them directly to purifying capacity, they are only used indirectly, or in specific cases of bioremediation.

Any substance that exceeds naturally occurring levels in the soil and poses health risks to life forms, in particular humans, is a contaminant. Soil indicators refer to the presence of contaminants whose nature, location, or quantity produces undesirable effects on the environment or human health. They can be very variable, such as heavy metals, different types of pesticides, excess nutrients, hydrocarbons, microplastics, etc. In Europe, the pollutants most frequently monitored are Cd (cadmium), Co (cobalt), Cr (chromium), Cu, Ni, Pb (lead), Zn, among the potentially toxic elements, while there are organochlorine pesticides, polycyclic aromatic hydrocarbons, and polychlorinated biphenyls, among organic pollutants (EJP-SOIL, 2021).

Carbon sequestration

Soils support vegetation that photosynthetically captures carbon dioxide and host soil organisms responsible for its storage and recycling. Soil organic matter has long been the most widely used quality indicator.

The main source of soil organic matter (SOM) is the above- and below-ground residues of vegetation. The humification and decomposition of these organic materials sustain the soil food chain, as SOM is used as a source of energy for the soil micro- and meso-fauna and fungi. At the same time, mineralization of the plant residues releases nutrients into the soil solution, where they

become accessible for uptake by the vegetation's root system. Soil organic matter is subjected to microbial degradation and its persistence can vary depending on both chemical recalcitrance and physical protection.

The main soil indicator is the organic carbon content, also called the organic carbon density. Derived from it is the carbon stock, which refers to the organic mass of the soil, calculated at the reference depth, considering bulk density and stoniness. Other derived indicators are the temporal changes in organic carbon and/or stock and the stratification throughout the soil profile. The stratification ratio is calculated by the SOC concentration at a shallow depth divided by that at a deeper depth. The stratification ratio is higher in soils that have stored more carbon in the topsoil, whereas it is lower when for example, water or wind erosion has occurred, there is compaction, fertility is poor, carbon input was limited or could not accumulate.

The thickness of the Ah horizon (topsoil mineral horizon with humus accumulation) is another useful indicator to monitor carbon sequestration and the variation in soil health.

Soil organic matter is composed of different functional fractions, which are defined according to their persistence capacity vs. decomposability. There are soil indicators used for monitoring the labile organic carbon fraction or "active" pool, which encompasses only very few per cent of the overall carbon pool and represent the most readily decomposable fraction. Among the most used, there are the hot water extractable carbon (Ghani et al., 2003 in [Bünemann et al., 2018](#)), the particulate organic matter (POM) (Cambardella and Elliott, 1992 in [Bünemann et al., 2018](#)), and the active carbon (Weil et al., 2003 in [Bünemann et al., 2018](#)). The POM, in particular, is the fraction of total organic matter which does not pass through a filter and that typically ranges in size from 0.053 and 2 mm, while active carbon is extracted with dilute, instead of concentrated, potassium permanganate to react with only the most readily oxidizable (active) forms of soil C. The labile organic carbon is responsive to land-use change and management practices, therefore, it has often been considered in monitoring soil health.

Biological activity and cycling of organic matter and nutrients

It is the activity of soil organisms that recycle dead organic matter and mineral inputs, producing mineral forms that plants and microflora can use for vegetative growth, and is at the basis of the soil food web. Several biological indicators have been widely used to monitor soil life in space and time. The most commonly used are the C:N ratio, microbial biomass carbon, microbial respiration rate, and enzymatic activity.

Biological properties are very sensitive to external conditions, as microorganisms require only a brief time to reproduce (from hours to days), allowing them to react to pressure and to transfer genetic modifications swiftly at population levels. This makes these indicators very sensitive to changes, but also increases the relevance of proper sampling. The cycle of organic matter and nutrients is usually evaluated in the laboratory as potential activity, because routine methods for open field measurements have not been standardized (e.g., Tea Bag Index, earthworm test). Potential activity means metabolic activity, including enzymatic activities that soil microbes are capable of developing under optimal laboratory conditions. The decomposition of SOM is carried out by microorganisms through the enzymatic attack of SOM and microbial respiration: extracellular enzymes degrade SOM through hydrolytic or oxidative processes, producing assimilable dissolved organic matter that can be rapidly incorporated by microbes.

Biologically active forms of SOM can function as short-term indicators of longer-term changes in SOM. Autoclaved-citrate extractable protein is an indicator of organically bound nitrogen that is easily mineralized by microbes ([Hurisso et al., 2018](#)).

Biodiversity pool

Healthy soils are capable of hosting great biodiversity. The term "soil biodiversity" is used in a genetic sense and denotes the number of distinct species (richness) in an ecosystem, and their proportional abundance (evenness), but can be extended to encompass phenotypic, functional, structural, or trophic diversity ([Benedetti and Mocali, 2010](#)).

Soil biodiversity plays an important role in supporting the sustainable productivity of ecosystems and regulating multiple other ecosystem services, including nutrient cycling, organic matter decomposition, pollutant degradation, and pathogen control in terrestrial ecosystems. However, the linkages between diverse soil organisms and ecosystem functions remain unclear, and the roles played by bacteria, fungi, protists, and invertebrates for multiple types of ecosystem functions remain largely unresolved. Soil biological diversity reflects the variability among living organisms including microorganisms (e.g., bacteria, fungi, protozoa) and mesofauna (e.g., nematodes, acari, and springtails), as well as the macro-fauna (e.g., earthworms, ants and termites).

Indicators for soil biodiversity include the collection of macro- and meso-organisms in the field by pitfall trapping, their identification and counting by a trained person, following extractions in the laboratory. Regular genomic analysis can also assess biodiversity more accurately at the microbial level. Currently, many methods are available to assess soil microbial diversity. The use of molecular techniques to investigate the microbial diversity of soil communities continues to provide a new understanding of soil properties and quality. Analysis of the soil-extracted nucleic acid sequences (DNA and RNA profiling) provides a powerful tool for characterization of the entire microbial community. The metagenomic approach is another method to assess simultaneously both soil microbial diversity and function ([Benedetti and Mocali, 2010](#)).

The analysis of types and amounts of different phospholipid fatty acids (PLFA) is a biochemical approach that offers an alternative to molecular techniques because it reflects both microbial taxonomic and functional diversity. The amount of total

PLFA can be used as an indicator of viable microbial biomass; further characterization can be done based on the specific signature of biomarker fatty acids (Quideau et al., 2016).

Soil mesofauna composition (microarthropods <2 mm) has been proposed for the evaluation of soil ecosystem services, in particular, biodiversity pools. Soil-dwelling animals play a significant role in the colonization and restoration of degraded biological habitats. Their role includes litter fragmentation, soil aggregation and porosity formation, water infiltration, and the distribution of organic matter in soil horizons. The greater the number of different mesofauna groups adapted to the soil habitat, the better is the soil functionality. Healthy soil systems show a set of ecosystem niches and related organisms, while stressed soils are poorer, both in species and individuals. The QBS-ar index has been developed to overcome difficulties related to identification at the species level, by focusing on the evaluation of adaptability to hypogeal life. Higher values correspond to more complex and soil-adapted communities.

Conclusions

The selection of appropriate indicators for the scope of soil quality and health assessment and monitoring, the methods of analysis, the elaboration and interpretation of results, their reporting and the guidelines given to farmers, policymakers and the general public, are still the object of great debate and a wealth of studies. In general, the list of suggested indicators is never exclusive, leaving room for the use of further indicators. Similarly, the choice of a minimum dataset should have sufficient degrees of freedom to meet the aims of the assessment and the availability of resources. However, the inclusion of at least one or two indicators for each of the physical, chemical, and biological properties of the soil is always recommended to arrive at a proper evaluation of soil functions.

When databases are incomplete for some important soil characteristics, measured data may be substituted by information derived from pedofunctions or rules, based on local, benchmark soils. On the other hand, if the datasets are limited or we make use of many derived characteristics, the benchmark sites are particularly important for validating the results given by the indicators.

The relevance of trained and specific competencies in the use of soil indicators cannot be overemphasized. The selection of indicators and methods of analysis, knowledge of the local soil types and their variability, interpretation and use of the results all require close cooperation of professionals with different and specific expertise.

Future developments should go towards increasing integration between field observations and measurements, use of proximal and remote sensors, and laboratory analysis. It should always be kept in mind that soil is a complex system where processes act throughout the whole profile, not only in the topsoil, and that simplification (use of few indicators or synthetic indicators) and approximations (derivative indicators, pedofunctions) can easily lead to incorrect, incomplete or biased results.

References

- Alef K and Nannipieri P (1995) *Methods in Applied Soil Microbiology and Biochemistry*. Academic Press.
- Arshady MA, Lowery B, and Grossman B (1997) Physical tests for monitoring soil quality. *Methods for Assessing Soil Quality*, vol. 49, pp. 123–141. Madison, Wisconsin, USA: Soil Science Society of America, Inc.
- ASAE (1994) Soil cone penetrometers. In: *S313.2 in Standards Engineering Practices Data*. St. Joseph, MI, USA: ASAE Standards.
- Benedetti A and Mocali S (2010) Exploring research frontiers in microbiology: The challenge of metagenomics in soil microbiology. *Research in Microbiology* 161: 497–505.
- Bünemann EK, Bongiorno G, Bai Z, et al. (2018) Soil quality—A critical review. *Soil Biology and Biochemistry* 120: 105–125.
- Cassel DK and Nielsen DR (1986) Field capacity and available water capacity. *Methods of Soil Analysis: Part 1 Physical and Mineralogical Methods*, vol. 5, 901–926.
- Costantini EAC, Branquinho C, Nunes A, et al. (2016) Soil indicators to assess the effectiveness of restoration strategies in dryland ecosystems. *Solid Earth* 7: 397–414.
- Costantini EA, Castaldini M, Diago MP, et al. (2018) Effects of soil erosion on agro-ecosystem services and soil functions: A multidisciplinary study in nineteen organically farmed European and Turkish vineyards. *Journal of Environmental Management* 223: 614–624.
- Doran JW and Zeiss JR (2000) Soil health and sustainability: Managing the biotic component of soil quality. *Applied Soil Ecology* 15: 3–11.
- Drinkwater LE, Cambardella CA, Reeder JD, and Rice CW (1997) Potentially mineralizable nitrogen as an indicator of biologically active soil nitrogen. *Methods for Assessing Soil Quality*, vol. 49, pp. 217–229. Madison, WI, USA: Soil Science Society of America.
- EJP-SOIL (2021) Deliverable 2.2 Stocktaking on Soil Quality Indicators and Associated Decision Support Tools, Including ICT Tools. Online https://ejpsol.eu/fileadmin/projects/ejpsol/WP2/Deliverable_2.2_Stocktaking_on_soil_quality_indicators_and_associated_decision_support_tools_including_ict_tools.pdf
- FAO-GSP-GLOSOLAN (2021). Standard Operating Procedures (SOPs) [online] <http://www.fao.org/global-soil-partnership/glosolan/soil-analysis/standard-operating-procedures/en/#c763834>
- Franco HHS, Guimarães RML, Tormena CA, Cherubin MR, and Favilla HS (2019) Global applications of the Visual Evaluation of Soil Structure method: A systematic review and meta-analysis. *Soil and Tillage Research* 190: 61–69.
- Herrick JE and Jones TL (2002) A dynamic cone penetrometer for measuring soil penetration resistance. *Soil Science Society of American Journal* 66: 1320–1324.
- Hurisso TT, Moebius-Clune DJ, Culman SW, Moebius-Clune BN, Thies JE, et al. (2018) Soil protein as a rapid soil health indicator of potentially available organic nitrogen. *Agricultural and Environmental Letters* 3(1), 180006.
- IUSS Working Group WRB (2014) World reference base for soil resources. In: *World Soil Resources Report, 103*. Rome (Italy): FAO.
- Jahn R, Blume HP, Asio VB, Spaargaren O, and Schad P (2006) *Guidelines for Soil Description*. FAO.
- Klingebiel AA and Montgomery PH (1961) *Land-Capability Classification (No. 210)*. Soil Conservation Service, US Department of Agriculture.
- Kowalchuk GA (2004) *Molecular Microbial Ecology Manual*. Springer Science & Business Media.
- Le Bissonnais YL (1996) Aggregate stability and assessment of soil crustability and erodibility: I. Theory and methodology. *European Journal of Soil Science* 47: 425–437.
- Lehmann J, Bossio DA, Kögel-Knabner I, and Rillig MC (2020) The concept and future prospects of soil health. *Nature Reviews Earth & Environment* 1(10): 544–553.
- Lowery B, Hickey WJ, Arshady MA, and Lal R (1997) Soil water parameters and soil quality. *Methods for Assessing Soil Quality*, vol. 49, pp. 143–155. Madison: SSSA.
- Makki M, Thestorf K, Hilbert S, Thelemann M, and Makowsky L (2021) Guideline for the description of soils in the Berlin metropolitan area: An extension for surveying and mapping anthropogenic and natural soils in urban environments within the German soil classification system. *Journal of Soils and Sediments* 21(5): 1998–2012.

- Marx MC, Wood M, and Jarvis SC (2001) A microplate fluorimetric assay for the study of enzyme diversity in soils. *Soil Biology and Biochemistry* 33(12–13): 1633–1640.
- Pellegrini S, Agnelli AE, Andrenelli MC, et al. (2018) Using present and past climosequences to estimate soil organic carbon and related physical quality indicators under future climatic conditions. *Agriculture, Ecosystems & Environment* 266: 17–30.
- Priori S, Pellegrini S, Vignozzi N, and Costantini EAC (2021) Soil Physical-Hydrological Degradation in the Root-Zone of Tree Crops: Problems and Solutions. *Agronomy* 11: 68.
- Quideau SA, McIntosh AC, Norris CE, et al. (2016) Extraction and analysis of microbial phospholipid fatty acids in soils. *Journal of Visualized Experiments* (114): 54360.
- Reynolds WD, Bowman BT, Brunke RR, et al. (2000) Comparison of tension infiltrometer, pressure infiltrometer, and soil core estimates of saturated hydraulic conductivity. *Soil Science Society of America Journal* 64(2): 478–484.
- Sapozhnikov PM and Granina NI (2021) Assessment of the soil cover in the Irkutsk Region by cadastral value. *IOP Conference Series: Earth and Environmental Science*, vol. 629, p. 012024. IOP Publishing. No. 1.
- Sikora FJ and Moore KP (2014) *Soil Test Methods From the Southeastern United States*. *Southern Cooperative Series Bulletin* 419: 54–58. Fayetteville, AR: University of Arkansas.
- Soil Survey Staff (1999) Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys. *Agriculture Handbook*, vol. 436. Natural Resources Conservation Service.
- Sparks DL (ed.) (1996) *Methods of Soil Analysis. Part 3. Chemical Methods*, In: *Soil Science Society of America Book Series*, No. 5.
- Torri D, Poesen J, and Borselli L (1997) Predictability and uncertainty of the soil erodibility factor using a global dataset. *Catena* 31(1–2): 1–22.
- van Es HM and Karlen DL (2019) Reanalysis validates soil health indicator sensitivity and correlation with long-term crop yields. *Soil Science Society of America Journal* 83(3): 721–732.

Water induced soil erosion

R Jane Rickson, School of Water, Energy and Environment, Cranfield University, Cranfield, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction	194
Processes of water induced soil erosion	194
Consequences of water induced erosion	196
On-site consequences	196
Off-site consequences	196
The total costs of soil erosion	197
Types of water induced soil erosion	197
Factors affecting water induced erosion	197
Rainfall erosivity	197
Soil erodibility	198
Slope steepness and slope length	199
Land use and land cover	199
Land management practices	200
Soil erosion control measures	201
Agronomic methods of soil erosion control	201
Mechanical methods of soil erosion control	202
Terraces	202
Grass waterways	203
Soil management methods of soil erosion control	203
Gully erosion control	205
Effectiveness of land management measures in controlling soil erosion	205
Conclusion/summary/outlook: A short concluding paragraph	205
References	206

Abstract

Soil erosion is the detachment and transport of slope forming materials by erosive agents. Rates of water-induced soil erosion are determined by interactions between rainfall and runoff, soil type, slope steepness and length, vegetation and land management. As soil is lost through the processes of sheetwash, rill and gully erosion, vital ecosystem goods and services delivered by soils are compromised. The impacts of this justify the use of erosion control measures, which include agronomic methods (the use of vegetation or simulated vegetation); mechanical methods (the use of field engineering structures); and soil management methods (the improvement of soil conditions).

Glossary

Bench terrace A step like landform, designed and built on steep slopes to reduce slope length and erosion risk.

Channel terrace An excavated channel (comprising a cut and fill section) running across slope to intercept any potentially erosive overland flow.

Ecosystem goods and services Ecosystems underpin all human life and activities. The goods and services they provide are vital to sustaining well-being, and to future economic and social development.

Rainfall erosivity The ability of rainfall to cause soil erosion.

Soil erodibility The susceptibility of soil to soil erosion processes.

Soil erosion The detachment and transport of slope forming materials by erosive agents.

Soil loss ratio The relative soil loss under a certain vegetation or land management practice, compared to the soil loss from bare soil that is cultivated up/down slope.

Soil loss tolerance The actual rate of soil erosion at which a deterioration or loss of one or more soil functions does not occur.

Key points

- This chapter defines soil erosion as the detachment and transport of slope forming materials by erosive agents.
- The factors affecting erosion are discussed, namely interactions between rainfall and runoff, soil type, slope steepness and length, vegetation and land management.

- The on-site and off-site consequences (and costs) of erosion are presented.
- The chapter concludes with descriptions of soil erosion control measures, namely agronomic methods (the use of vegetation or simulated vegetation); mechanical methods (the use of field engineering structures); and soil management methods (the improvement of soil conditions).

Nomenclature

t ha⁻¹ yr⁻¹ Metric tonnes per hectare per year

Introduction

Soil erosion is the detachment and transport of slope forming materials by erosive agents, including rainfall, runoff, wind, tillage, and co-extraction on root crops and land-based machinery. This chapter focuses on water-induced erosion of soil on hillslopes. Erosion of river beds, banks and channels, and the movement of slope forming materials solely under gravity (i.e., mass movements).

In natural systems, soil erosion is in balance with soil formation, as an example of ‘dynamic equilibrium’ in the environment. However, land use change, removal of vegetation and soil disturbance by human activities can lead to accelerated rates of soil loss, often exceeding those of soil renewal, which will lead to an unsustainable system. As soil is lost to erosion, the vital ecosystem goods and services delivered by soil resources are also lost. These include production of food, fiber, fodder and biofuels; storage and supply of water resources (helping to mitigate droughts and floods); sequestration and storage of carbon (to regulate our climate); and protection of terrestrial habitats (to support biodiversity) (Adhikari and Hartemink, 2016; Haygarth and Ritz, 2009). Soils have been linked to most of the United Nations Sustainable Development Goals (UN SDGs; (Keesstra et al., 2016)) which drive international governments’ programs and policies in delivering outcomes such as No Poverty (SDG1), Zero Hunger (SDG2), Good Health and Well-being (SDG3), Clean Water (SDG6), Clean Energy (SDG8), Climate Action (SDG13) and Life On Land (SDG15).

Local, regional and global rates of water induced erosion vary over both space and time, and depend on how they were estimated. Measurement and monitoring of soil erosion can be undertaken by field surveys of erosion features such as rills and gullies (Types of water induced erosion), that can be mapped to show the spatial extent of erosion and how this evolves over time (known as ‘sequential erosion surveys’). Field erosion plots, usually covering an area of several m² can be set up to observe erosion processes under natural or simulated rainfall. Catchment-scale studies often involve measurement of sediment loads and concentrations (i.e., eroded soil) in streams and rivers. Remote sensing imagery can measure and monitor soil erosion over large areas, although the resolution of the imagery will tend to capture only large-scale (i.e., >1 m²) erosion features. Finally, soil erosion prediction models have been developed, based on the wealth of observed data on rates of erosion the factors affecting erosion, and also a better understanding of erosion processes. These models are classified as empirical, conceptual, or process-oriented, and predict erosion processes at different spatial and temporal scales. Several of these models are reviewed in Borrelli et al. (2021), including the Universal Soil Loss Equation (USLE) (Wischmeier and Smith, 1978) and the EUROSEM model (Morgan et al., 1998).

The FAO (2019) reports a range of soil erosion rates from 2 to almost 50 t ha⁻¹ year⁻¹. This is compared to rates of soil formation which are estimated to be roughly equivalent to 0.7 t ha⁻¹ year⁻¹ on a global scale (Wakatsuki and Rasyidin, 1992). Tolerable rates of soil erosion, defined as the “actual soil erosion rate at which a deterioration or loss of one or more soil functions does not occur”, range between 0.2 and 2.2 t ha⁻¹ year⁻¹, (Verheijen et al., 2009). The soil loss rates quoted are likely to increase due to (a) expansion of cropland, especially in Sub-Saharan Africa, South America and Southeast Asia (Borrelli et al., 2017) and (b) climate change, where higher intensity, more frequent and longer duration rainfall events are also likely to increase rates of soil erosion (O’Hara et al., 1993).

Processes of water induced soil erosion

The first process of soil erosion occurs when individual or aggregated soil particles are detached from the soil surface by raindrop impacts and or surface overland flow. Detachment occurs when the shear forces of these eroding agents exceed the soil’s resistance to detachment. Thus, the rate of soil detachment is determined by the kinetic energy of the erosive agent (i.e., the mass and velocity of the raindrops and or overland flow) and detachability of the soil (Eq. 1). In the case of rainsplash erosion, the depth of any surface water layer can help protect the soil from the force of the impacting drops.

$$DET = k.KE^a.e^{-b.h} \quad (1)$$

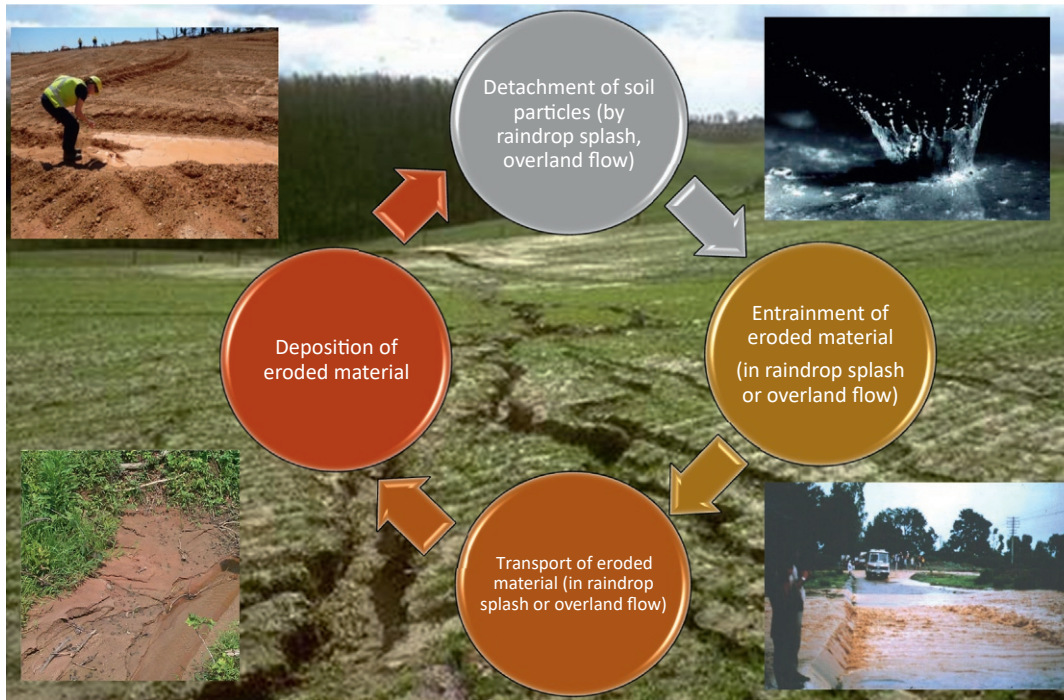


Fig. 1 The cyclical nature of soil erosion processes.

where, DET = detachment of material ($\text{g m}^{-2} \text{s}^{-1}$), k = index of detachability (g l^{-1}), KE = kinetic energy (based on the mass and velocity of the erosive agent) ($\text{J m}^{-2} \text{s}^{-1}$), a, b = experiment coefficients related to soil texture, h = depth (mm) of surface water layer (if any) $KE = 1/2 m v^2$, where m = mass of an object and v = the velocity by which it moves.

The detached material is then transported away, either by raindrop splash (in the jets of water spreading out from the point of raindrop impact) or by overland flow. For rainsplash erosion, the resulting degree of erosion is limited by the small distances covered by the rainsplash jet trajectories. This is usually less than 1 m. On the other hand, overland flow can transport detached particles over much longer distances, depending on the particle size of the material being carried and the kinetic energy, i.e., volume and velocity of the flow. Deposition of transported material occurs when the energy of the erosive agent is insufficient to carry the eroded material. This might be caused by a reduction in the volume of the overland flow (e.g., cessation of rainfall) and or in the velocity of the flow (e.g., friction imparted to the flow). The deposited material may be redetached in the following rainfall event and the whole detachment-transport-deposition process restarts (Fig. 1).

The relationships between soil particle size, flow velocity and energy, and erosion processes are shown in the Hjulström curve (Fig. 2), where soil particle size (x axis) and flow velocity (y axis) will determine whether the particle is detached, transported or deposited.

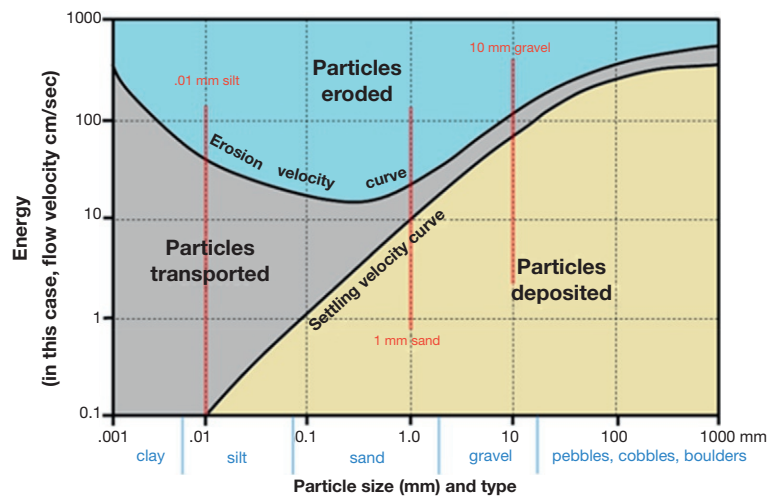


Fig. 2 The Hjulström Curve, showing the relationship between particle size, flow velocity and erosion process (detachment, transport and deposition).

Consequences of water induced erosion

On-site consequences

Loss of soil depth from erosion has major impacts on soil functions including food production, water regulation, carbon sequestration and storage, biodiversity, and protection of cultural features. Soil erosion reduces the depth and volume of rooting media for vegetation growth. Nutrients, carbon and organic matter held in the soil are preferentially eroded as they are associated with the more erosion-susceptible, finer textured materials (especially silts and very fine sands), making erosion a selective process. Seeds and seedlings can also be washed away with erosion, together with soil biota (flora and fauna), making the remaining 'uneroded' soil less biologically active. This in situ soil also has smaller amounts of nutrients, lower infiltration rates and less water holding capacity, making it more susceptible to further erosion. As a result, a soil profile that has been degraded by erosion is less able to support vegetation cover that may (a) protect it against future erosion events and (b) produce crops to ensure food security (Zhang et al., 2021) (Fig. 3). Soil erosion features can also disrupt field operations and vehicular access to land.

Off-site consequences

Eroded material can be transported by overland flow to the wider hydrological network (Fig. 4). High sediment concentrations in water (turbidity) can affect the aquatic environment in terms of temperature (sediment can absorb more heat than water), oxygen (reduced dissolved O₂ levels) and reduced depth of light penetration. These changes can damage aquatic flora and fauna. For example, sediment can have a significant negative effect on fish populations, as suspended soil particles may damage fish gills and make channel beds unfavorable for spawning (Lazar et al., 2010). Aquatic plant communities can be scoured by soil particles in the flow and smothered by deposited sediments. Water bodies with high sediment concentrations are also unsuitable as sources of irrigation water or sites for hydroelectricity (due to abrasive damage to pipework and turbines). Expensive water treatment to remove sediment can be necessary before it is suitable for domestic or industrial supplies. Recreational activities such as angling, boating and water sports are also impeded by large sediment concentrations in water bodies. Deposited sediment in waterbodies reduces water storage capacities, increasing the risk of flooding during and after high rainfall and runoff events. Deposited sediment also impedes navigation, requiring disruptive and expensive dredging activities to remove the eroded materials.

In addition to these physical and biological effects, eroded soil can be an effective carrier of pollutants and contaminants. Sediments that are easily transported also have the highest chemical adsorption capacity due to their relatively large specific surface area. Agrochemicals such as nitrates, phosphates, herbicides and pesticides can be adsorbed and absorbed onto eroded sediment, and transported throughout the catchment. With the selective nature of erosion, the concentrations of these chemicals can be higher in the eroded sediment than in the in situ soil where they were originally applied. Phosphate is especially associated with eroded sediment, and its concentration in and subsequent release from deposited sediments can lead to eutrophication of waterbodies as excessive nutrient loading encourages algal blooms which deplete oxygen levels for other aquatic species.



Fig. 3 Severe gully erosion on agricultural land in the United Kingdom.



Fig. 4 Poor water quality due to high concentrations of eroded soil from agricultural land, United Kingdom.

The total costs of soil erosion

Putting a financial value on the on-site and off-site consequences of soil erosion is challenging, as many of these impacts cannot be easily monetarized. [Pimentel et al. \(1995\)](#) estimated that the total annual costs of soil erosion by water in the US was \$7.4 billion, which is the equivalent of \$14.4 billion at 2022 prices. Accounting for inflation, the annual costs of soil erosion in England, Scotland and Wales alone were estimated at approximately £285 million at 2022 prices ([Graves et al., 2015](#)). These costs justify the use of soil erosion control measures.

Types of water induced soil erosion

For soil detachment to occur, rainsplash and overland flow must reach a critical shear force to overcome the soil's shear resistance. This can occur at a range of spatial scales, from a few centimeters to several meters. At a very localized scale, 'pedestal erosion' occurs where no overland flow is present, but rainsplash erosion alone preferentially removes soil if there is no surface protection from the raindrop impacts (e.g., bare ground). Rates of soil loss in this case are fairly small ([Morgan, 2005](#)). However, where overland flow is generated (and with sufficient volume, velocity and kinetic energy), then localized flow concentrations (due to surface irregularities caused by roughness elements such as vegetation or stones) can exceed the shear resistance of the soil and detachment by flow will take place. Where there is no obvious eroded channel, this is termed 'wash', 'sheet' or 'interill' erosion. Where flow concentrations do occur, the velocity and kinetic energy of the flow may be sufficient to create small channels or 'rills'. By definition, these erosion features are small enough that they can be removed by the next rainfall event or by conventional agricultural operations, such as plowing ([Fig. 5](#)). However, as these small channels develop, they receive more overland flow, setting up a positive feedback loop which may result in the formation of a 'gully'. These deeply incised, permanent erosion features are associated with very high rates of soil loss and are notoriously difficult to remove. A comparison of different forms of soil erosion and their associated soil losses is given in [Table 1](#).

Factors affecting water induced erosion

There are many factors affecting water induced rates of soil erosion. Many are ubiquitous in time and space because the processes of erosion are universal, even if absolute rates vary. It is important to understand the factors affecting erosion as this can inform the design, use and effectiveness of soil erosion control practices. Many of these factors are inter-related, but here they are discussed independently.

Rainfall erosivity

The potential ability for rainfall to cause erosion is termed 'erosivity'. The raindrops from 1000 mm of rainfall, hitting a hectare of land ($100\text{ m} \times 100\text{ m}$) provide the energy equivalent of 60,000 kcal (250×106 joules) per year. This 60,000 kcal represents about the same energy of 8 L of petrol ([Pimentel and Burgess, 2013](#)). The kinetic energy (KE) available for soil erosion depends on the raindrop mass and drop velocity, which are related: larger drops have higher velocities and hence higher kinetic energies. Raindrop



Fig. 5 Severe rill erosion on an arable field, Herefordshire, United Kingdom.

Table 1 Comparison of different forms of soil erosion.

Form	Mass ^a	Typical velocity ($m\ s^{-1}$)	Kinetic energy ^b	Energy for erosion ^c	Observed sediment transport ^d ($g\ cm^{-1}$)
Raindrops	R	6.0	$18R$	$0.036R$	20
Overland flow	$0.5R$	0.01	$2.5 \times 10^{-5}R$	$7.5 \times 10^{-7}R$	400
Rill flow	$0.5R$	4^e	$4R$	$0.12R$	19,000

^aAssumes rainfall mass of R of which 50 per cent contributes to runoff.

^bBased on $\frac{1}{2}mv^2$.

^cAssumes that 0.2 per cent of the kinetic energy of raindrops and 3 per cent of the kinetic energy of runoff is utilized in erosion.

^dTotals observed in mid-Bedfordshire, England, on an 11 degree slope, on sandy soil, over 900 days. Most of the energy of raindrops contributes to soil particle detachment rather than transport.

^eEstimated using the Manning equation of flow velocity for a rill, 0.3 m wide and 0.2 m deep, on a slope of 11 degree, at bankfull, assuming a roughness coefficient of 0.02. From Morgan RPC (2005) *Soil Erosion and Conservation*, 3rd edn., Blackwells.

size depends on a range of meteorological factors including cloud ambient temperature, air turbulence, particulate concentrations and the weather system (e.g., cold or warm fronts; frontal, convective or relief rainfall). As rainfall intensity increases ($mm\ h^{-1}$), the median drop size (D_{50}) also increases (Laws and Parsons, 1943; Eq. 2).

$$D_{50}\ (mm) = 2.23 * (I/25.4)^{0.182} \quad (2)$$

where I = Intensity ($mm\ h^{-1}$).

Each raindrop falls at its 'terminal velocity' (where it experiences zero acceleration). For example, a raindrop of 1.5 mm diameter falls at a terminal velocity of $5.5\ m\ sec^{-1}$. For a larger 3.0 mm diameter drop, the terminal velocity is $8.0\ m\ sec^{-1}$. Raindrop velocity can be affected by wind, with a reported 3.5-fold increase in rainfall kinetic energy with 'wind driven rainfall' (Helming, 2001; Marzen et al., 2017). The total mass of drops (and their associated terminal velocities) for a rainstorm will indicate that event's erosivity (potential ability to cause erosion).

With the many factors affecting erosivity, simple indices of rainfall erosivity have been devised to quantify the ability of rainfall to cause water-induced erosion. Ideally these indices should be (a) strongly correlated with soil loss and (b) determined from easily available rainfall data. Examples include the 'EI₃₀ index', which is based on the total KE of a storm and the maximum 30-min intensity (Wischmeier and Smith, 1978). It was developed in the US, using data from experimental soil erosion plots. The $KE > 25\ mm\ h^{-1}$ index was devised because the EI₃₀ index was weakly correlated with soil loss in tropical countries (Hudson, 1981). The strongest correlations with rates of soil erosion were given by the sum of the storm KE for all rainfall events with an intensity $> 25\ mm\ h^{-1}$.

Soil erodibility

Soils' susceptibility to erosive agents is termed 'soil erodibility'. It is affected by soil properties such as soil texture, where soils with a large proportion of silt and very fine sand particles are especially susceptible to erosion. This is because larger particle sizes (sands) require more energy to be transported (see Fig. 2), whereas finer textured soils (e.g., clays) tend to adhere to each other, again requiring more energy to overcome these cohesive bonds. Silts and very fine sands lack either of these characteristics. The organic

matter content of the soil is also important as this affects the binding of soil particles and their ability to resist detachment by rainfall and runoff. This resistance to erosive forces is also due to soil chemistry and the attraction of positive and negative ions held on the mineral and organic components of the soil. The chemical interaction of soil particles, minerals and organic matter leads to the formation of polycationic and organo-mineral bridges between particles.

Soil organic matter also improves the soil's water holding capacity, reducing the risk of generating potentially erosive overland flow during or after a storm event. Organic matter also supports soil biology (such as earthworms, fungi, bacteria and plant root exudates) which helps improve soil aggregation and soil structure, stabilizing soil pore networks and connectivity. This facilitates infiltration of rainfall and reduces the risk of overland flow generation. Stable aggregates are also strongly correlated with reduced soil erodibility (Amezket, 1999; Nciizah and Wakindiki, 2015).

External factors such as freeze-thaw, rapid wetting (slaking) and extreme heat (leading to hydrophobic soils) can reduce the aggregation of soil particles, making them more susceptible (erodible) to erosive forces. The way the soil is managed can also affect its erodibility. Field practices such as intensive and repeated cultivations and tillage operations break up stable soil aggregates, making the soil more susceptible to detachment and transport by rainfall and runoff. Removing crop residues after harvest reduces soil organic matter which in turn increases soil erodibility. Large rates of livestock stocking can also remove vegetation, making the soil beneath more susceptible to erosion.

Natural and synthetic soil amendments and conditioners have also been developed to reduce soil erodibility, by making the soil either hydrophilic (reducing overland flow) or hydrophobic (increasing soil resistance to detachment). Products used for runoff and soil erosion control include seaweed extracts, molasses, compost teas and bitumen.

Soil erodibility indices have been derived to generalize and simplify all the relevant factors affecting a soil's susceptibility to erosion. Ideally, these must be (a) correlated with rates of soil erosion and (b) derived from simple and accessible data sets. Examples include the 'K' erodibility index, based on soil texture (silt and very fine sand), organic matter %, permeability and soil structure (Wischmeier and Mannering, 1969).

Slope steepness and slope length

"Obviously, the steeper the slope, the greater the erosion. Not surprisingly, the amount of erosion is not just proportional to the steepness of the slope, but rises rapidly as the slope increases. The length of slope has a similar effect" (Hudson, 1981). As slope length increases, the volume of overland flow accumulates and erosion risk increases because the flow has a greater transporting capacity. The acceleration of that flow also increases on longer slopes, increasing flow velocity and kinetic energy. However, there will be a point at which the flow is carrying too much sediment for its transport capacity and deposition may occur on very long slopes.

An increase in slope gradient (steepness) is associated with higher velocities and kinetic energy of overland flow, which can be expended in soil detachment and transport. Slope gradient has even more effect on flow erosivity than slope length, as expressed in the exponents used in the following equation (Zingg, 1940):

$$E = f(S^{1.4} L^{0.6}) \quad (3)$$

where, E = soil loss ($\text{t ha}^{-1} \text{yr}^{-1}$), S = slope gradient (%), L = slope length (m).

In addition to slope length and steepness, slope profile (cross section) and plan (cross slope) are also important because they determine the concentration or divergence of overland flow. Where landform concentrates overland flow, rates of soil erosion will be higher.

Land use and land cover

The effectiveness of vegetation in controlling soil erosion is best demonstrated when that protective cover is removed, and bare soil is exposed to erosive rainfall and overland flow. Plant traits determine the erosion-control effectiveness of individual plants and vegetation stands. The canopy of the plant will intercept raindrops, so shattering them into smaller drop sizes and reducing their velocity. In turn, this reduces the erosivity of those drops. Plant canopies also reduce the volume of rainfall reaching the ground (where it may generate overland flow), by either evaporation of the rainfall that is intercepted and held on the canopy or by extending the time it takes for the rainfall to reach the ground surface. This reduces the effective intensity of the rainfall and thus its erosivity. Plant stems impart a roughness to surface flow, leading to a reduction in the velocity, kinetic energy and erosivity of the flow, due to frictional effects, turbulence and eddying. Plant stems also absorb some of the flow shear forces that would otherwise apply to the soil surface. The reduction in flow velocity due to the frictional effects of plant stems reduces the transport capacity of the flow, leading to deposition of previously eroded material. Plant stems can also have a physical filtering effect on sediment in the flow, again leading to sediment deposition upslope of the vegetation stand. Plant roots encourage infiltration of rainfall, so reducing the generation of potentially erosive overland flow. Fine roots in the surface soil layer increase soil shear strength (through adhesion and cohesion between the root and soil particles), so reducing soil susceptibility to detachment and soil erodibility. Finally, roots add organic matter to the soil, which is strongly correlated to soil erodibility. More indirectly, vegetation debris provides a protective litter layer on the soil surface, as well as encouraging soil biological activity, and associated aggregate stability and low erodibility (Fig. 6).

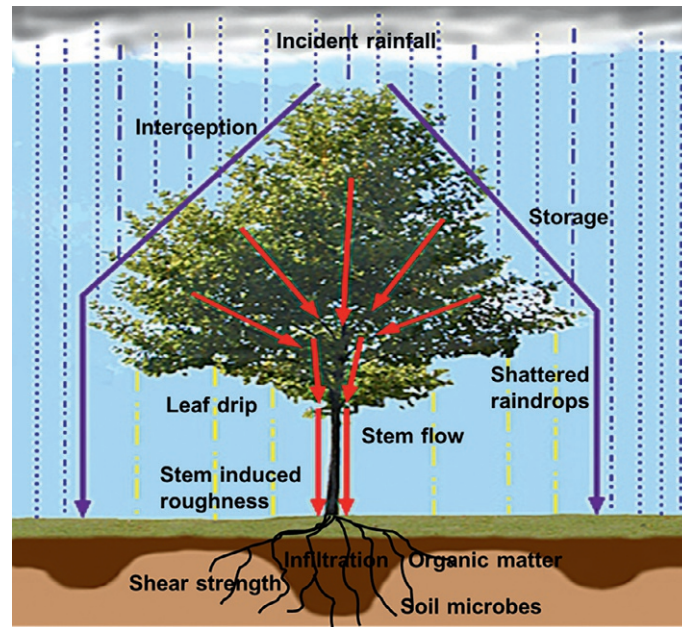


Fig. 6 The role of vegetation on soil erosion processes.

The overall effectiveness of vegetation in controlling soil erosion is simply expressed in the 'C' factor of the Universal Soil Loss Equation (Wischmeier and Smith, 1978), where

$$C = C_i + C_{ii} + C_{iii} \quad (4)$$

where, C = effect of vegetation on soil loss and erosion processes, C_i = above ground canopy effects, C_{ii} = at or on ground canopy effects, C_{iii} = residual effects of land use (effects of roots, subsurface stems, effects of soil biological activity, soil structure, organic matter content, soil bulk density, effects of tillage).

In the C factor, the rate of soil erosion under a given vegetation cover is expressed as a 'soil loss ratio' of that under bare soil, all other factors being equal (Table 2). This ratio is determined for the whole cropping cycle, from seedbed preparation to harvest (and management of any crop residues, post-harvest). The original C factors were calculated from observations of soil losses on erosion plots in the US under many different vegetation covers (including unvegetated plots). The C factor for a bare soil is 1, whereas it is 0.001 for a deciduous forest, all other factors being the same (i.e. 0.1% of the soil loss from the bare soil equivalent). While the C factor is a useful device to show the relative effectiveness of vegetation to control erosion, it does not explain the mechanisms by which the different vegetation types (or their components) control soil losses. Soil erosion prediction models such as the modified Morgan, Morgan and Finney model give better expressions of the way vegetation affects the processes of soil detachment and transport (Morgan and Duzant, 2008).

Land management practices

As shown by the 'C' factor above, the choice of land use or crop type will affect the degree of soil surface protection from erosive forces. Arable and horticultural crops (especially row crops) are particularly vulnerable to erosion due to (a) the high degree of soil disturbance during cultivations, especially during seed bed preparation and (b) long periods of bare soil, both post cultivation and post-harvest. Multiple tillage operations will destroy soil structure and soil aggregation, so reducing infiltration and increasing the generation of runoff and soil erosion. Creation of a loose, fine tilth seedbed can lead to surface sealing, impeded infiltration, greater runoff generation and higher rates of erosion. The small soil aggregates created will be highly susceptible to erosion. Soil erodibility is further increased by the loss of soil carbon through oxidation as tillage exposes deeper soil layers to the atmosphere. Soil faunal biomass, including fungal mycelia and earthworms, is important for maintaining cohesive bonds within and between soil aggregates, but soil biology can be adversely effected during soil disturbance.

Reducing the number and or intensity of cultivations preserves soil structural stability and aggregation. Less disturbed aggregates are able to resist erosive forces, as cohesive bonds within the aggregates are maintained. Fewer, less invasive cultivations mean more of the vegetative biomass remains on or within the soil, which can absorb the energy of erosive agents, so reducing erosion rates. For example, a rough, 'trashy' soil surface imparts friction to overland flow, causing turbulence and eddying, resulting in a reduced flow velocity and kinetic energy to erode. Carrying out field operations up and down slope (compared to across slope or on the contour) can concentrate overland flow, so much so that it becomes erosive. Land management practices to control soil erosion are discussed in more detail in the following section below.

Table 2 The cropping ('C') factor for selected land uses/crops (expressed as a ratio of soil erosion under bare soil).

Crop/Land Use	C factor	Crop/Land Use	C factor
Bare soil	1	Millet/sorghum (<i>Sorghum bicolor</i> (L.) Moench)	0.4–0.9
Barley (<i>Hordeum vulgare</i> L.)	0.28	Mulch (e.g., crop residues)	0.025 (100% cover)
Beans (<i>Phaseolus vulgaris</i> L.)	0.2–0.4		0.3 (50% cover)
Cabbage (<i>Brassica oleracea</i> L. <i>Capitata</i> group)	0.6		0.58 (20% cover)
Cashew (<i>Anacardium occidentale</i> L.)	0.4	Oil palm (<i>Elaeis guineensis</i> Jacq.)	0.1–0.3
Cassava (<i>Manihot esculenta</i> Crantz)	0.6	Orange (<i>Citrus × sinensis</i> (L.) Osb.)	0.3
Chilli (<i>Capsicum annuum</i> L.)	0.6	Pasture	0.015
Cocoa (<i>Theobroma cacao</i> L.)	0.1–0.3	Pineapple (<i>Ananas comosus</i> (L.) Merr.)	0.38
Coconut (<i>Cocos nucifera</i> L.)	0.4	Potato (<i>Solanum tuberosum</i> L.)	0.2–0.3
Coffee (<i>Coffea</i> L.)	0.1–0.3	Prairie/savannah grass	0.01–0.10
Corn / maize (<i>Zea mays</i> L.)	0.25	Sorghum (<i>Sorghum bicolor</i> (L.) Moench)	0.65
Cotton (<i>Gossypium arboreum</i> L.)	0.3–0.7	Soybean (<i>Glycine max</i> (L.) Merrill)	0.421
Cultivated grass	0.004–0.01	Sugar beet (<i>Beta vulgaris</i> L.)	0.2–0.3
Deciduous forest	0.001	Sugarcane (<i>Saccharum officinarum</i> L.)	0.4
Disturbed forest	0.04	Sunflower (<i>Helianthus annuus</i> L.)	0.7
Dry evergreen forest	0.019	Sweet potato (<i>Ipomoea batatas</i> (L.) Lam.)	0.6
Floriculture	0.386	Tea (<i>Camellia sinensis</i> (L.) O. Kuntze)	0.1–0.3
Forest/woodland	0.001–0.002 (with undergrowth)	Tomato (<i>Solanum lycopersicum</i> Mill.)	0.6
	0.001–0.004 (no undergrowth)	Vine (<i>Vitis vinifera</i> L.)	0.6
Golf course	0.015	Watermelon (<i>Citrullus lanatus</i> Shrade.)	0.6
Jute (<i>Corchorus olitorius</i> L.)	0.6	Wheat (<i>Triticum aestivum</i> L.)	0.1–0.2 (winter sown)
Mango (<i>Mangifera indica</i> Linn.)	0.15		0.2–0.4 (spring sown)

Soil erosion control measures

Just as human interference can aggravate rates of soil erosion (e.g., by removing vegetation cover), so protective land management practices can be used to reduce soil losses. The control of soil erosion by water (sometimes referred to as 'soil conservation') can be classified into three approaches:

- Agronomic methods: the use of vegetation or simulated vegetation.
- Mechanical methods: the use of field engineering structures.
- Soil management methods: the manipulation of soil conditions.

These methods are inter-related, and in many situations, a single approach is insufficient to control soil erosion on its own and several approaches need to be combined to achieve tolerable erosion rates. Most practices have been developed for agricultural land, but can be applied on other land uses such as construction sites and on engineered slopes (e.g. road and railway embankments and cuttings).

Agronomic methods of soil erosion control

Vegetation is highly effective at controlling soil erosion (Coppin and Richards, 2007). As described above, the canopy, stem, root and litter components of plants interact with soil erosion processes. Agronomic methods of soil erosion control apply these principles in the field. Maximum vegetation cover on the soil surface will protect the soil from erosive raindrops and overland flow, and absorb their kinetic energies. High stem and root densities reduce overland flow volumes, velocities and energy to erode. These effects are achieved through practices such as nurse cropping, high density planting, cover and companion cropping, and crop rotations. Vegetation can be used to reduce slope length and gradient when grown in strips parallel to the contour. The strips may be of different crops (e.g. strip cropping) or as a single vegetative strip across the slope (e.g. field buffer strips). The strips restrict the generation of overland flow and encourage deposition of eroded sediment, so that local slope gradients can be reduced too.

Agroforestry systems grow perennial trees or shrubs together with annual crops. The trees form across-slope barriers to overland flow (also known as 'alley cropping'), so reducing erosion (Nair et al., 1995). The deep tree roots encourage infiltration and reduce overland flow, and leaf fall from the trees adds organic matter to the soil, so reducing its erodibility. Mulches and geotextiles or erosion control blankets (Bhattacharyya et al., 2010) simulate the components of live vegetation that control soil erosion (Fig. 7). By covering the soil, these materials protect it from incoming rainfall and overland flow. When incorporated or buried into the soil, these materials simulate root effects on erosion. Natural-based mulches and geotextiles degrade over time, adding organic matter to the soil, so reducing erodibility. The C factor for soil erosion control geotextiles is estimated at 0.144.



Fig. 7 Nurse crop with a straw mulch (foreground) and coir geotextile (background) for soil erosion control on a new road embankment.

Mechanical methods of soil erosion control

Field engineering structures are used to control erosion on steep slopes, where the erosion risk exceeds the soil loss tolerance.

Terraces

Terraces control soil erosion by reducing slope length and to a lesser extent, slope steepness, so reducing the volume, velocity, and kinetic energy of overland flow to cause soil erosion. There are many different types of erosion control terrace, often developed for local environmental conditions (e.g., the ‘fanya juu’ terraces of eastern Africa; Fig. 8). Other types of terrace include channel and bench terraces. Channel terraces are recommended on slopes no steeper than 7° (12%) (Sheng, 1972). The distance between consecutive terraces down slope can be determined by equations based on site conditions, such as rainfall, soil type and slope gradient. Distances between terrace channels vary as the slope gradient changes, with shorter terrace intervals on steeper slopes. This may cause practical problems for mechanized farming. Channel terraces consist of a dug channel and a filled ridge (constructed from the cut material), excavated across the slope, usually at a gentle gradient of $1:250 = 0.4\%$ slope to carry any overland flow, but at a non-erosive velocity.

Bench terraces are step-like structures, constructed across the slope. They prevent soil erosion by breaking up natural slope lengths into smaller, shorter components, each with a horizontal (or near horizontal) bench and a vertical ‘riser’ (Fig. 9). The dimensions of the benches and risers depend on slope steepness, the desirable width of cultivable land (i.e. bench width) and the acceptable height of the steep wall of the terrace, or ‘riser’ where it remains structurally stable and will not collapse. The construction



Fig. 8 Construction of ‘fanya juu’ terraces for soil and water conservation, Machakos, Kenya.



Fig. 9 Bench terrace, showing height of riser and width of bench.

of the level bench requires the over-steepening of the terrace riser. To prevent collapse of the riser, it must be supported with stones, rocks, masonry, timber or simply a dense network of vegetation roots.

Grass waterways

Field waterways are designed to carry overland flow downslope, at a safe velocity, without causing erosion. These waterways are ideally located in natural depressions to minimize the need to excavate a channel. They are usually lined with dense, short vegetation, stones or geotextiles (Fig. 10). These linings permit high velocity flows without causing erosion. The width, depth and cross sectional area of the waterway are determined by the volume of overland flow they have to carry, the slope gradient and the properties of the channel lining.

Soil management methods of soil erosion control

Farming operations carried out up and down the slope will increase the risk of overland flow generation and soil erosion. Changing to contour cultivations across the slope will reduce these risks significantly. But this is not always practical, especially on complex or steep slopes, where the arrangement of fields on the contour makes field operations such as harvesting difficult and incurs high energy consumption. Contour ridging creates raised ridges across slope, which will reduce localized slope length and effective catchment area for overland flow generation. The ridges can also intercept any overland flow that is generated. Ridges can be joined together at regular intervals by 'ties', so creating small basins, which trap any overland flow and eroded soil, conserving both soil and water within the 'tied ridges' (Fig. 11).

'Conservation tillage' involves the relative reduction of soil disturbance and retention of vegetative material on and within the soil to control soil erosion (Holland, 2004). The term incorporates a continuum of soil management methods including zero/no till, minimum till, mulch tillage, direct drilling, non-inversion tillage, sub-surface mulch tillage and cultivations that combine plowing and planting operations simultaneously (Fig. 12). Conservation tillage aims to maintain soil fertility and control erosion (Cogo et al., 1984). It can also reduce input costs to land managers, as fewer, less invasive cultivations expend less energy and fuel, resulting in considerable rates of adoption in the 1970s in the US during the oil crisis. It should be noted that the success of conservation tillage is related to soil type and timeliness of operations. For example, there is evidence that reducing the number of cultivations can increase bulk density on structureless soils, leading to overland flow generation, which can only be alleviated by deep sub-soiling periodically. Supporters of conservation tillage claim that these are short term effects, which are off-set in the medium to long term by the improvement of soil structure resulting from increased levels of biotic activity and organic matter contents.

Soil conditioners are applied to soils susceptible to erosion. Hydrophilic soil conditioners including tree bark, compost teas, palm oil effluent and seaweed extract often add organic matter to the soil that increases cohesive bonds and the stability of aggregates. The improved soil structure and infiltration capacity reduce the risk of runoff and soil erosion. Hydrophobic soil conditioners including latex, emulsions and molasses, form an impermeable film on the soil surface, which increases the soil's resistance to erosion. The impermeable layer also inhibits infiltration and thus increases the volume of overland flow, but this is unable to detach the soil particles that are bound together by the conditioner.



Fig. 10 Stages of installing a grass waterway for soil erosion control (clockwise from top left: rill erosion in concentrated flow path; marking the width of the waterway; installing a geotextile lining; established grass waterway).

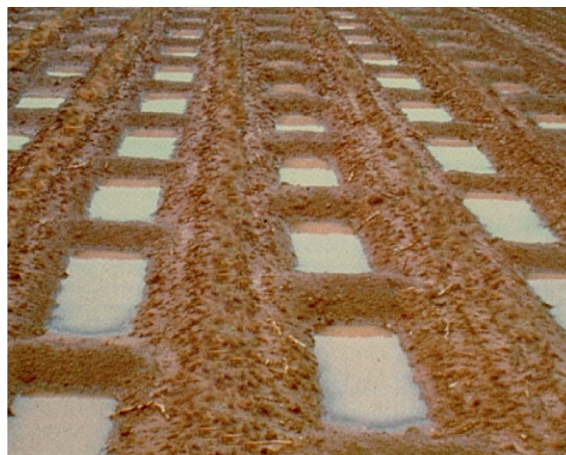


Fig. 11 Tied ridging for soil and water conservation.



Fig. 12 Seedbed conditions: (a) conventional tillage (multiple operations creating a very fine tilth) and (b) conservation tillage (direct drill into previous seasons crop residues).

Gully erosion control

Gullies are erosion features with permanent channels, but intermittent flow. They tend to be deeper than wide. They are notoriously difficult to control, due to their erratic hydrological behavior, high rates of runoff and soil loss, and also they cannot be removed by normal field operations such as plowing. Soil loss is generated at many sites: erosion of the adjacent hillside, and the gully headwall, banks and bed. Prioritizing where soil erosion control measures are needed most can be a challenge, requiring multidisciplinary knowledge of hydraulic and civil engineering, biology, ecology, geology, geomorphology and soil science. The costs of gully reclamation are extremely high, yet difficult to justify in such an eroding and degrading landscape, demonstrating the importance of the principle that “prevention is better than cure”.

Controlling runoff in the gully catchment can reduce the volume of water reaching the gully. Stabilizing the gully banks with wattling, brush layering or stone filled baskets (gabions) can reduce bank collapse and deflect flow away from the gully sides. Check dams (or ‘grade stabilization structures’) are installed in the gully floor to act as an obstruction to flow. This reduces flow velocity and encourages sedimentation upstream of the dam. In time, this deposited material changes the channel gradient and vegetation may establish, retarding flow velocities further. Also, the trapped sediments can store storm runoff. Over time, the pattern of sedimentation behind the dams will create a ‘stepped’ long profile. Check dams can be made of live or dead brush material, rocks (with or without gabions), sandbags, hay bales, timber, sheet steel, masonry, concrete or earth embankments. The spacing of the check dams is given as:

$$S = \frac{He}{K \cdot \tan G \cdot \cos G} \quad (5)$$

where S = Spacing between dams (m), He = Effective dam height (spillway crest to gully floor; m), G = Gully gradient (degrees), K = Constant related to volume of sediment deposited, = 0.3 (if $\tan G < 0.2$), = 0.5 (if $\tan G > 0.2$) (Heede and Mufich, 1974).

Effectiveness of land management measures in controlling soil erosion

The relative effectiveness of some land management practices on rates of soil erosion can be expressed as a ‘soil loss ratio’ (i.e., the ratio of soil loss under a certain land management practice compared to that where no soil control measures are used (Table 3)). Here, the baseline of up and downslope cultivation is given a relative ‘P’ (protection) value of 1. The relative reduction in soil erosion from using different land management practices is then expressed as a ratio of this value.

Conclusion/summary/outlook: A short concluding paragraph

Soil erosion is a natural process, operating in dynamic equilibrium with the processes of soil formation. However, rates of soil erosion can be accelerated by human activities such as removal of protective vegetation, soil disturbance by cultivations and oversteepening of engineered slopes. Soil erosion is exacerbated by climate change (in particular more intense, frequent and erosive rainfall events). Soil erosion and all forms of soil degradation will increase, given the growing demands and pressures put on our finite and limited land resources to supply the world with sufficient food, water and energy, while protecting terrestrial biodiversity. Thus the impacts of soil erosion are felt not only by land managers, but by the whole of society as soil erosion curtails the delivery of many goods and services that originate from soils. Controlling soil losses is vital to maintain these goods and services, but many farmers are often unconvinced about whether erosion control measures will bring benefits (either economic or environmental) to their businesses, either in the short or long term. This suggests that more evidence is needed of the effectiveness of erosion control measures. Also, payments or subsidies are required to encourage uptake of erosion control measures because financial rewards may only accrue in the long term, often beyond the financial planning horizons of many land managers.

Table 3 Relative reduction in soil loss from different land management practices (The 'P' Factor).

Land management practice	P value
Up & down slope cultivation	1.0
Cross slope cultivation	0.75
Contour farming	0.50
Strip cropping, across slope	0.37
Strip cropping, on the contour	0.25
Contouring (on slopes 1–25%)	0.6–0.9 (depending on the slope gradient)
Strip cropping	0.45–0.68 (ditto)
Grass strip 10% cover	0.55–0.65 (ditto)
Grass strip 20% cover	0.50–0.55 (ditto)
Grass strip 30% cover	0.4–0.50
Grass strip 40% cover	0.35–0.40
Grass strip 50% cover	0.30–0.35
Level bench terrace	0.14
Reverse slope bench terrace	0.05
Outward sloping bench terrace	0.35
Level retention bench terrace	0.01

References

- Adhikari K and Hartemink AE (2016) Linking soils to ecosystem services—A global review. *Geoderma*. <https://doi.org/10.1016/j.geoderma.2015.08.009>.
- Amezketta E (1999) Soil aggregate stability: A review. *Journal of Sustainable Agriculture* 14(2–3): 83–151.
- Bhattacharyya R, Smets T, Fullen MA, Poesen J, and Booth CA (2010) Effectiveness of geotextiles in reducing runoff and soil loss: A synthesis. *Catena* 81(3): 184–195. <https://doi.org/10.1016/j.catena.2010.03.003>.
- Borrelli P, Robinson DA, Fleischer LR, Lugato E, Ballabio C, Alewell C, Meusburger K, Modugno S, Schütt B, Ferro V, Bagarello V, Van Oost K, Montanarella L, and Panagos P (2017) An assessment of the global impact of 21st century land use change on soil erosion. *Nature Communications* 8(1): 2013. <https://doi.org/10.1038/s41467-017-02142-7>.
- Borrelli P, Alewell C, Alvarez P, Anache JAA, Baartman J, Ballabio C, Bezak N, Biddocccu M, Cerdà A, and Chalise D (2021) Soil erosion modelling: A global review and statistical analysis. *Science of the Total Environment* 780: 146494.
- Cogo NP, Moldenhauer WC, and Foster GR (1984) Soil loss reductions from conservation tillage practices. *Soil Science Society of America Journal* 48(2): 368–373. <https://doi.org/10.2136/sssaj1984.03615995004800020029x>.
- Coppin NJ and Richards IG (2007) *Use of Vegetation in Civil Engineering*. London: Ciria.
- FAO (2019) *Soil Erosion: The Greatest Challenge to Sustainable Soil Management*. FAO.
- Graves AR, Morris J, Deeks LK, Rickson RJ, Kibblewhite MG, Harris JA, Farewell TS, and Truckle I (2015) The total costs of soil degradation in England and Wales. *Ecological Economics* 119: 399–413. <https://doi.org/10.1016/j.ecolecon.2015.07.026>.
- Haygarth PM and Ritz K (2009) The future of soils and land use in the UK: Soil systems for the provision of land-based ecosystem services. *Land Use Policy* 26(supplement 1): 187–197. <https://doi.org/10.1016/j.landusepol.2009.09.016>.
- Heede BH and Mufich JG (1974) Field and computer procedures for gully control by check dams. *Journal of Environmental Management*.
- Helming K (2001) Wind speed effects on rain erosivity. In: Stott DE, Mohtar RH, and Steinhardt GC (eds.) *10th International Soil Conservation Organization Meeting*, pp. 771–776. USDA-ARS.
- Holland JM (2004) The environmental consequences of adopting conservation tillage in Europe: Reviewing the evidence. *Agriculture, Ecosystems & Environment* 103(1): 1–25. <https://doi.org/10.1016/j.agee.2003.12.018>.
- Hudson NW (1981) *Soil Conservation*. London: Batsford.
- Keesstra SD, Bouma J, Wallinga J, Tittone P, Smith P, Cerdà A, Montanarella L, Quinton JN, Pachepsky Y, van der Putten WH, Bardgett RD, Moolenaar S, Mol G, Jansen B, and Fresco LO (2016) The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals. *The Soil* 2(2): 111–128. <https://doi.org/10.5194/soil-2-111-2016>.
- Laws JO and Parsons DA (1943) The relation of raindrop-size to intensity. *Eos, Transactions American Geophysical Union* 24(2): 452–460.
- Lazar AN, Butterfield D, Futter MN, Rankinen K, Thouvenot-Korppoo M, Jarritt N, Lawrence DSL, Wade AJ, and Whitehead PG (2010) An assessment of the fine sediment dynamics in an upland river system: INCA-Sed modifications and implications for fisheries. *Science of the Total Environment* 408(12): 2555–2566. <https://doi.org/10.1016/j.scitotenv.2010.02.030>.
- Marzen M, Iserloh T, de Lima JLMP, Fister W, and Ries JB (2017) Impact of severe rain storms on soil erosion: Experimental evaluation of wind-driven rain and its implications for natural hazard management. *Science of the Total Environment* 590–591: 502–513. <https://doi.org/10.1016/j.scitotenv.2017.02.190>.
- Morgan RPC (2005) *Soil Erosion and Conservation*, 3rd edn. Blackwells.
- Morgan RPC and Duzant JH (2008) Modified MMF (Morgan-Morgan-Finney) model for evaluating effects of crops and vegetation cover on soil erosion. *Earth Surface Processes and Landforms* 33(1): 90–106. <https://doi.org/10.1002/esp.1530>.
- Morgan RPC, Quinton JN, Smith RE, Govers G, Poesen JWA, Auerswald K, Chisci G, Torri D, and Styczen ME (1998) The European soil erosion model (EUROSEM): A dynamic approach for predicting sediment transport from fields and small catchments. *Earth Surface Processes and Landforms* 23(6): 527–544. [https://doi.org/10.1002/\(SICI\)1096-9837\(199806\)23:6<527::AID-ESP868>3.0.CO;2-5](https://doi.org/10.1002/(SICI)1096-9837(199806)23:6<527::AID-ESP868>3.0.CO;2-5).
- Nair PKR, Kang BT, and Kass DCL (1995) Nutrient cycling and soil-erosion control in agroforestry systems. *Agriculture and the Environment: Bridging Food Production and Environmental Protection in Developing Countries, ASA Special Publications*, vol. 60, 117–138.
- Ncizah AD and Wakindiki IC (2015) Physical indicators of soil erosion, aggregate stability and erodibility. *Archives of Agronomy and Soil Science* 61(6): 827–842. <https://doi.org/10.1080/03650340.2014.956660>.
- O'Hara SL, Street-Perrott FA, and Burt TP (1993) Climate change and soil erosion. *Nature* 364(6434): 197. <https://doi.org/10.1038/364197c0>.
- Pimentel D and Burgess M (2013) Soil erosion threatens food production. *Agriculture* 3(3): 443–463.
- Pimentel D, Harvey C, Resosudarmo P, Sinclair K, Kurz D, McNair M, Crist S, Shpritz L, Fitton L, Saffouri R, and Blair R (1995) Environmental and economic costs of soil erosion and conservation benefits. *Science (New York, N.Y.)* 267(5201): 1117–1123. <https://doi.org/10.1126/science.267.5201.1117>.

- Sheng TC (1972) Bench terracing. *Journal of the Scientific Research Council of Jamaica* 3: 113–127.
- Verheijen FGA, Jones RJA, Rickson RJ, and Smith CJ (2009) Tolerable versus actual soil erosion rates in Europe. *Earth-Science Reviews* 94(1–4): 23–38. <https://doi.org/10.1016/j.earscirev.2009.02.003>.
- Wakatsuki T and Rasyidin A (1992) Rates of weathering and soil formation. *Geoderma* 52(3): 251–263. [https://doi.org/10.1016/0016-7061\(92\)90040-E](https://doi.org/10.1016/0016-7061(92)90040-E).
- Wischmeier WH and Mannering JV (1969) Relation of soil properties to its erodibility. *Soil Science Society of America Proceedings* 33(1): 131–137.
- Wischmeier WH and Smith DD (1978) *Predicting Rainfall Losse, a Guide to Conservation Planning*. Agriculture Handbook, vol. 357, U.S. Department of Agriculture 67.
- Zhang L, Huang Y, Rong L, Duan X, Zhang R, Li Y, and Guan J (2021) Effect of soil erosion depth on crop yield based on topsoil removal method: A meta-analysis. *Agronomy for Sustainable Development* 41(5). <https://doi.org/10.1007/s13593-021-00718-8>.
- Zingg AW (1940) Degree and length of land slope as it affects soil loss in run-off. *Agricultural Engineering* 21: 59–64.

Erosion by wind

John Tatarko, Rangeland Resources and Systems Research Unit, USDA Agricultural Research Service, Fort Collins, CO, United States

© 2023, Published by Elsevier Ltd.

Disclaimer: Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

EEO Statement: USDA is an equal opportunity provider and employer.

Introduction	209
Effects of wind erosion	210
Wind erosion processes	210
The wind	212
Particle movement	212
Soil surface erodibility	213
Vegetation	215
Wind erosion modeling	215
Wind erosion measurement	216
Wind erosion control	218
Vegetation and vegetative residues	218
Barriers	219
Strip cropping	220
Soil aggregation and roughness	220
Chemical treatments	221
Emergency control	221
Summary	221
Appendix: Supplementary material	222
References	222

Abstract

Soil erosion by wind is a dynamic physical process, leading to significant soil, air, water, and human health concerns. It occurs as a result of strong winds impacting on loose, dry, bare soils. Although wind erosion occurs on natural landscapes throughout the world, this chapter focuses primarily on wind erosion on cropland and discusses the various impacts of erosion by wind. However, the same principles also apply to natural landscapes. It also examines the processes of wind erosion including the wind profile, modes of particle transport, soil surface conditions, and vegetation as well as modeling, measurements, and control of wind erosion.

Key points

- Wind erosion is the result of the physical interaction of wind forces with the soil.
- Wind erosion impacts not only the soil but also air and water quality as well as human health.
- Wind erosion models help improve understanding of wind erosion processes and control.
- Measurement of wind erosion is needed to design controls, validate models, and assess soil–wind interactions.
- A wide variety of wind erosion control practices have been developed.

Symbols

u_*	Friction velocity (m s^{-1})
$u(z)$	Wind speed at height z (m s^{-1})
z	Height (m)
z_0	Aerodynamic roughness height (m)
u_{*t}	Threshold friction velocity (m s^{-1})
u_t	Minimum wind velocity initiating particle movement (m s^{-1})

Introduction

Wind erosion is a dynamic physical process that occurs when strong winds impact loose, dry, or bare soils. If allowed to continue over time, it can lead to soil degradation. Fine, fertile soil particles are removed during wind erosion (wind-induced particle movement), reducing soil productivity and causing significant on-site and off-site problems. The widespread social and economic hardships that occurred in the United States (U.S.) during the disastrous Dust Bowl days of the 1930s were caused primarily by wind erosion on cropland (Fig. 1). Much progress has been made in reducing the effects of wind erosion in the U.S. through soil conservation efforts made by individual landowners with the technical assistance of the United States Department of Agriculture, Natural Resource Conservation Service (USDA, NRCS). However, wind erosion continues to be an international problem. A recent review of estimates of present-day global dust emissions varied from 500 to about 330 Tg year⁻¹ (Shao et al., 2011). Dust clouds may still be seen in many parts of the U.S. (Fig. 2). Studies by the USDA-NRCS show that wind erosion causes about 7210 million metric tons per year of soil to be lost from U.S. cropland (USDA, 2020). Although significant wind erosion occurs in arid desert



Fig. 1 A large 'roller' type dust storm moves across the land during the 1930s Dust Bowl in Colorado U.S. Photo credit: USDA-NRCS.



Fig. 2 A dust storm engulfing a farm in Iowa, U.S. Photo credit: USDA-NRCS.

regions and semiarid grasslands, in this chapter we explore the causes, effects, and control of wind erosion on cropland. However, most of the principles of erosion discussed here also apply to natural landscapes.

Global wind erosion was estimated by the United Nations Food and Agriculture Organization (FAO, 2019); they found that North Africa, the border between Iran and Afghanistan, and the Gobi Desert are areas with the highest rates of wind erosion of $>1.0 \text{ t ha}^{-1} \text{ year}^{-1}$. Other areas with less but still severe wind erosion (i.e., $<1.0 \text{ t ha}^{-1} \text{ year}^{-1}$) include the drylands of Iran and Afghanistan, the Arabian Peninsula, as well as dry areas of Argentina, Australia, Chile, Mexico, and the western United States (FAO, 2019). Fig. 3 shows global areas of wind erosion vulnerability based on global climate and soil properties.

Estimates of annual crop loss due to all erosion range from 0.1% to 0.4% that projects a total reduction in productivity of 10.25% by 2050 (FAO and ITPS, 2015). Wind erosion affects crop growth and yield by: (1) removal of the most fertile components of the soil surface (i.e., clays and organic matter), (2) incorporation of subsoil into the surface layer (i.e., degrades soil structure), and (3) a decrease in the rooting depth of the soil including reduced water holding capacity.

Effects of wind erosion

The on-site effects of wind erosion are generally considered to be those in or very near the eroding field, which may be further divided into long-term cumulative effects and immediate effects. The effects of wind erosion on the soil are cumulative and where it is a problem, soil losses often exceed the rates of soil formation dramatically. In extreme situations, wind may erode the entire topsoil horizon and reveal the underlying, less fertile horizons. Wind erosion is a winnowing process that, while it moves all the surface soil in a field, tends to remove the finer, more fertile parts of the soil and transport these materials away from the field. The result is a soil with a lower percentage of silt and clay, lower cation exchange capacity, lower organic matter, and reduced water-holding capacity. All of these factors serve to reduce the productivity of the soil and to diminish crop yields.

Immediate on-site effects of wind erosion also result in economic losses to agriculture. During windstorms, sandblast injury from wind eroding sand grains striking and abrading plants can damage crops, especially seedling stands. Depending on the intensity of the event, seedling stands may be retarded or stunted later in growth and development, damaged to the point that replanting is warranted, or completely destroyed. As the seedlings start to produce true leaves, and especially during exponential growth phases, tissue conversion to woody matter reduces the susceptibility of the plant to damage and the increase in vegetative ground cover tends to reduce wind erosion frequency and intensity. In general, small grain cereal crops are less susceptible to sandblast injury than maize (i.e., corn), soybeans, and sunflowers, which are more tolerant than cotton, cabbage-related crops, and legumes. The least sandblast-injury-tolerant crops are garden vegetables and flowers.

Studies suggest that the off-site impacts from wind erosion may be much larger than the on-site impacts (Pimentel and Burgess, 2013). The estimated off-site costs of wind erosion in the western U.S. range from \$4 to \$12 billion per year (Huszar and Piper, 1986). Off-site impacts from wind erosion are caused primarily by the release of fugitive dust, which may travel long distances and which constitutes most of the airborne dust that settles in homes and on automobiles, imposing costs for cleaning, reduced recreational opportunities, and impaired human and animal health. Estimates of the annual flux of airborne dust deposition over land areas range from less than 10 to about $200 \text{ t km}^{-2} \text{ yr}^{-1}$. As expected, these estimates vary with respect to distance and direction from the major source regions. The source areas of suspended dust are primarily deserts, agricultural operations in arid and semiarid environments, unpaved roads, dry lake beds, braided streams, deltas, and glacial outwash. These areas can be located on any mid-latitude continent, and the literature comprises studies of dust source regions on all continents.

Some suspended dust has been traced across continents, oceans, and around the world (Fig. 4). Significant amounts of suspended dust from Africa are transported across the Atlantic each year, which is deposited in the Caribbean and the U.S. with potential implications for ecological, environmental, and global climate change (Prospero and Lamb, 2003). Fugitive dust obscures visibility and negatively affects air quality. Reduced visibility contributes to the cancelation of airline flights as well as numerous automotive accidents and resulting deaths annually (Van Pelt et al., 2020). Reduced visibility also depresses tourism and recreation revenues. Fugitive dust fouls machinery and adversely impacts environmental health. Pathogens, both plant and animal, may be transported thousands of kilometers on dust particles, even across oceans. Dust from wind erosion has been shown to contribute to global climate change (Prospero and Lamb, 2003) and reductions in snow pack and water supply (Clow et al., 2016). A small fraction of suspended particles may cause health problems when inhaled. These particles are known as PM_{2.5} and PM₁₀, which are particulate matter less than 2.5 and 10 μm in size, respectively.

Wind erosion processes

Soils are susceptible to wind erosion if they are bare of vegetation, exist on extensive or flat areas, lack roughness or aggregation, and are subjected to strong wind forces. Soil erosion by wind occurs when the wind velocity at the soil surface exceeds a critical threshold value, and soil particles detach and move in response to wind shear forces (Tatarko et al., 2019). A soil surface that is rough or protected with non-erodible material, such as vegetation or large stable clods or aggregates, requires a stronger wind to initiate particle movement than a bare, smooth surface.

Wind Erosion Vulnerability

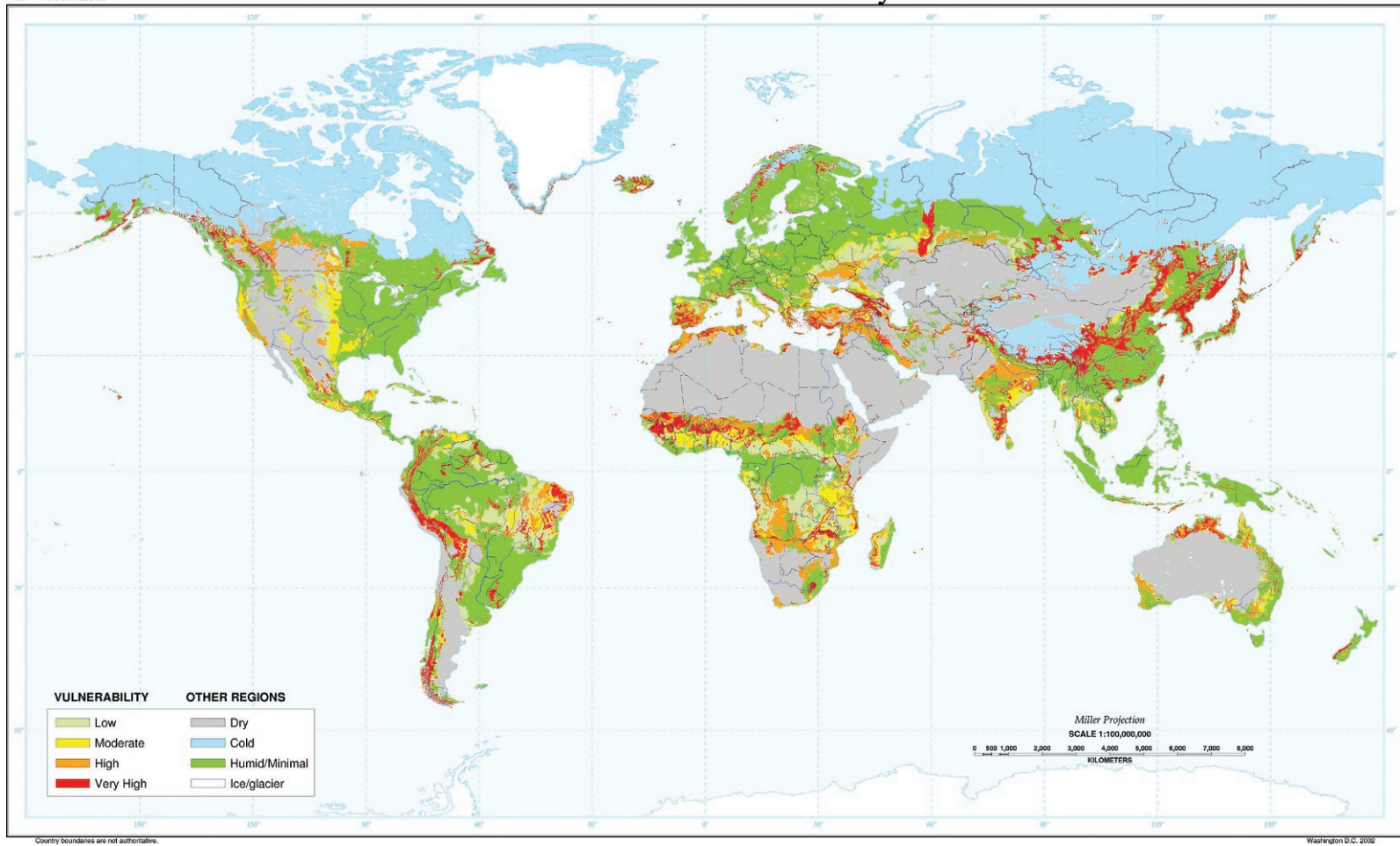


Fig. 3 The extent of global vulnerability to wind erosion. Map credit: USDA-NRCS. Available at: https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/use/worldsoils/?cid=nrcs142p2_054007.

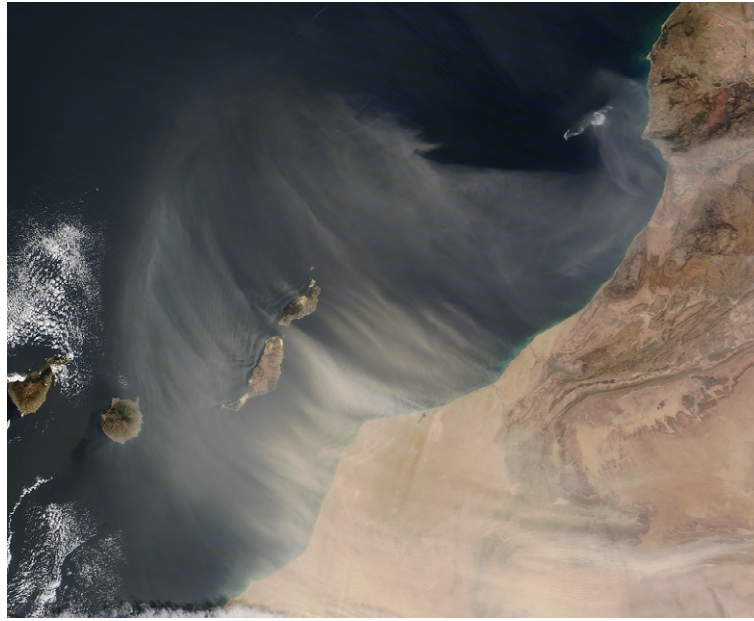


Fig. 4 Dust blowing off the coasts of Morocco and Western Sahara in early March 2012. The image is from NASA's Moderate Resolution Imaging Spectrometer (MODIS).

The wind

Wind is simply air in motion, caused by differences in air pressure in the Earth's atmosphere. Since air has mass, when it is moving, it has energy. That energy is what moves soil during wind erosion. Wind erosion is the result of the physical interaction of the wind force on the soil surface. The Earth's surface exerts a drag on the wind, resulting in a vertical profile of wind speed that is described by a semi-logarithmic equation:

$$u(z) = \frac{u^*}{k} \ln \left(\frac{z}{z_0} \right), \quad (1)$$

where $u(z)$ (m s^{-1}) is the wind speed at height z (m); u^* (m s^{-1}) is the friction velocity, which indicates the amount of atmospheric turbulence (its value is independent of height for a given wind profile); k is the von Karman constant (0.4), a dimensionless number; and z_0 (m) is the aerodynamic roughness height, which is a measure of the roughness of the ground surface. From Eq. (1), it is evident that there is a sharp decrease of mean horizontal wind speed as the surface is approached. This gradient in wind speed produces a shearing force caused by surface roughness resulting in a drag on the air flow. The shearing force produces a tangential stress on the surface, called the shear stress, which is equal to the product of the square of the friction velocity and fluid density. The shear stress provides the interchange of momentum necessary for erosion to occur. The susceptibility of a soil to transport due to the effects of wind or water is called soil erodibility. Soil wind erodibility is often assessed using the friction velocity at which particle movement begins, called the threshold friction velocity (u_{*t}). The minimum wind velocity initiating particle movement is known as the threshold velocity (u_t). There is no single threshold velocity but rather a range of velocities, depending on the condition of the soil surface, including plants present, roughness, aggregation, and soil surface moisture.

Particle movement

A soil particle's susceptibility to wind erosion is affected by its size, density, shape, and composition. The quartz sand particle diameter most susceptible to wind erosion is approximately $100 \mu\text{m}$ (Bagnold, 1941). This size of particle begins to move with a wind velocity of approximately 6 m s^{-1} , measured at a height of 2 m. Particles larger than this size require greater wind velocities to initiate movement due to the greater volume and mass of the larger particles. Similarly, particles smaller than $100 \mu\text{m}$ require greater wind velocities to move because of the increased surface area to volume ratios and greater inter-particle bonding forces for the smaller particles. Wind-blown materials erode in three modes of transport known as creep, saltation, and suspension (Fig. 5). Each mode has unique characteristics and effects.

In creep mode (often also called surface creep), soil particles 0.5 to 1 mm in diameter move by rolling or sliding along the soil surface because they are too massive to be lifted into the air. They continue to move until the wind slows, they are stopped by other particles, or they are trapped in a sheltered location, such as a crop furrow or a vegetated area. Creep contributes to loss and deposition within a localized area, filling fence rows and road ditches or burying plants with eroded soil.

Particles between 0.1 and 0.5 mm in size tend to move in saltation (bouncing) mode where they are lifted into the air and accelerated by the wind, returning to surface with greater energy and horizontal momentum, typically within 8 m downwind.

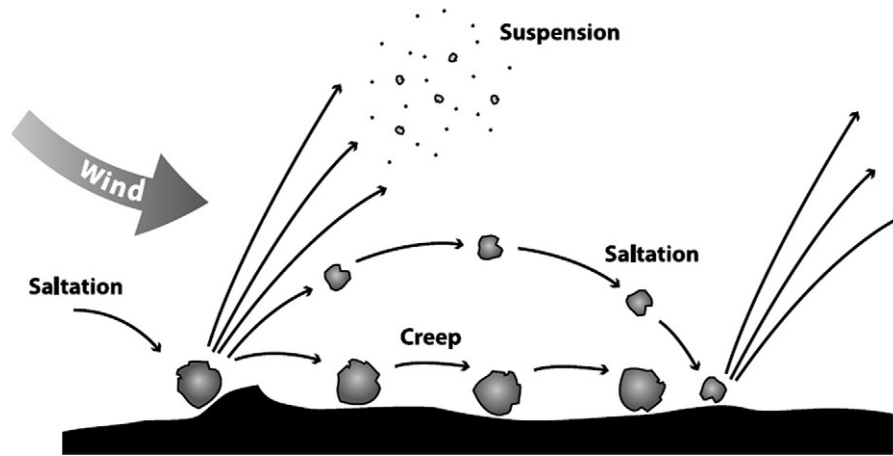


Fig. 5 Soil particles under wind erosion can move in creep, saltation, and suspension mode of transport. Photo credit: USDA-ARS.

Saltating particles return to the soil surface with a force that is a function of their mass and speed, resulting in the disruption of soil aggregates and surface crusts and the release of other particles, thus accelerating the erosion process. Saltation is what causes damage to young plants, threatening their survival. In mature plants, saltation can damage flowering parts and their fruit, reducing their marketability. Like particles under creep, saltating particles continue to move until the wind slows or they are trapped in sheltered areas such as road ditches or streams.

The third mode of particulate transport is suspension, which involves soil particles less than 0.1 mm in diameter. As these fine soil particulates are lifted from the surface, turbulent eddies carry them high into the air. These fine materials may travel great distances in the atmosphere before returning to Earth. Suspension sized particles degrade air quality and cause visibility and health problems. It is this size that includes PM_{2.5} and PM₁₀, both of which are regulated as health hazards by many countries.

For a particle to be lifted to any height, the upward velocity of the eddy must be greater than the terminal velocity of the particle. Sand-sized quartz grains have significantly higher terminal velocities than silt- and clay-sized particles as well as soil particulate organic matter and are rarely lifted to heights greater than a few meters. While in most cases sand-size particles are redeposited in the source field or nearby, in extreme events fine sands may enter true suspension mode and be transported great distances.

As saltating particles are ejected from the surface, they initially rise at an angle of about 25–50° and the force of the wind accelerates them constantly until gravity returns them to the surface. The force these returning particles exerts on the soil surface often results in the ejection of several sand grains from the surface, so the number of grains in saltation often increases exponentially over a short distance of tens of meters in sandy soils. This phenomenon has been called the avalanche effect. There is a limit to the number of grains that can be entrained for any given wind speed, however, the accelerating sand grains exert a drag on the passing wind that results in the transfer of wind energy into the energy of the sand particles and heat of friction. There is thus an equilibrium known as the transport capacity where as much saltating material is being deposited as is being removed from the surface. This is generally not true for particles moving in the suspension mode, as the quantity of suspended particles continues to increase upwards into the atmosphere with increasing fetch across the eroding surface.

Soil surface erodibility

Although wind drives the wind-erosion process, the condition of the soil surface often controls whether or not a soil is transported by the wind. The susceptibility of the soil surface to erosion is called soil erodibility. Any factor that reduces the impact of the wind on the soil surface reduces wind erosion. In addition, conditions that increase the cohesion of particles to produce larger non-erodible aggregates or increase the resistance of particles to detachment from the surface also will reduce wind erosion. For example, intergranular cohesion associated with surface soil moisture increases the threshold velocity needed for particle movement. All of the factors contributing to soil erodibility are the basis for developing control measures, which will be discussed later.

Surface soil texture (the proportions of sand, silt, and clay) is another important factor affecting the erodibility of a site. Sandy soils tend to have weak stability and are easily moved by wind. The USDA-NRCS has related surface soil texture to soil erodibility, classifying the soil into wind erodibility groups according to soil texture and carbonate content (Table 1; USDA-NRCS, 2002). Calcareous soils (soils that contain enough calcium carbonate to effervesce in the presence of dilute hydrochloric acid) tend to be more erodible than similar noncalcareous soils. The surface soil texture and carbonate content are difficult to change and are generally considered intrinsic or static soil properties. However, they may change spatially, over very short distances. Other surface soil properties are temporal in nature and may change annually, seasonally, or even daily. Thus, the effects of these properties on wind erosion may be subject to management practices.

Soil surface roughness, also called microrelief or microtopography, can modify the effect of the wind on the surface, physically protect part of the surface from abrasion caused by blowing particles, and trap particles that may be moving. In tilled soils, tillage tools often create an oriented roughness, or ridges, in the direction of tillage and random roughness due to the random orientation

Table 1 Relation of soil texture to soil erodibility.

Predominant soil texture class of surface layer	Wind erodibility group (WEG)	Soil erodibility index (I) (mg ha ⁻¹ year ⁻¹) ^a
Very fine sand, fine sand, sand, or coarse sand	1	694 ^b 560 493 403 358
Loamy very fine sand, loamy fine sand, loamy sand, loamy coarse sand, or sapric organic soil materials	2	300
Very fine sandy loam, fine sandy loam, sandy loam, or coarse sandy loam	3	193
Clay, silty clay, non-calcareous clay loam, or silty clay loam with more than 35% clay	4	193
Calcareous loam and silt loam or calcareous clay loam and silty clay loam	4 L	193
Non-calcareous loam and silt loam with less than 20% clay, or sandy clay loam, sandy clay, and hemic organic soil materials	5	125
Non-calcareous loam and silt loam with more than 20% clay, or non-calcareous clay loam with less than 35% clay	6	108
Silt, non-calcareous silty clay loam with less than 35% clay, and fibric organic soil material	7	85
Soils not susceptible to wind erosion due to coarse surface fragments or wetness	8	—

^aThe soil erodibility index is based on the relationship of dry soil aggregates greater than 0.84 mm to potential soil erosion.

^bThe "I" factors for WEG 1 vary from 160 for coarse sands to 310 for very fine sands. For coarse sand with gravel, use a low figure. For very fine sand without gravel, use a higher value. When unsure, use an "I" value of 220 as an average figure.

Adapted from the USDA, NRCS (2002) *National Agronomy Manual*. 3rd edn. Title 190 Part 502. Wind Erosion, USDA Natural Resources Conservation Service. Available online at: <https://efotg.sc.egov.usda.gov/>.

of soil aggregates and clods. The effectiveness of tillage-induced roughness on wind erosion depends upon wind direction relative to the orientation of the ridges. Winds in the direction perpendicular to tillage are affected by both ridges and clods while wind in the direction parallel to tillage is only affected by the random roughness.

In agricultural fields, tillage not only creates surface roughness but also disrupts the soil and creates an unconsolidated surface layer of soil aggregates, or clods of various sizes. Some of the aggregates are small enough to be eroded by wind. Aggregates of mineral soils <0.84 mm diameter are generally considered erodible. Larger particles of organic soils are blown because they have much lower density than mineral soils due to the relatively large content of organic material present. The estimated potential wind erosion is related to the amount of nonerodible aggregates, as indicated in Table 2 (USDA-NRCS, 2002). Practices that encourage the development of nonerodible sized aggregates reduce the amount of wind erosion in tilled fields.

The ability of soil aggregates to resist breakdown due to abrasion caused by blowing sand grains or other applied forces is called aggregate stability. Fragile soils have low aggregate stability and are easily disrupted during wind erosion events, causing additional fine material to be added to the wind stream. Aggregate stability is quantified by measuring the degree of aggregate disruption caused by repeated sieving in a rotary sieve or by determining the amount of energy per unit mass needed to crush aggregates of a given size using an aggregate crushing device (Hagen et al., 1995).

Soils often form consolidated surfaces, called crusts, that are more resistant to erosion. Crusts form through physical, chemical, and biological processes. Rainfall often impacts the soil surface with enough force to disrupt aggregates and rearrange the particles

Table 2 Soil erodibility index (I) in units of mg ha⁻¹ as determined by percentage of nonerodible soil, >0.84 mm in size.

Tens ^a	Units (mg ha ⁻¹)									
	0	1	2	3	4	5	6	7	8	9
0	—	694	560	493	437	403	381	358	336	314
10	300	293	287	280	271	262	253	244	237	228
20	220	213	206	202	197	193	186	181	177	170
30	166	161	159	155	150	146	141	139	134	130
40	125	121	116	114	112	108	105	101	96	92
50	85	81	74	69	65	60	56	54	52	49
60	47	45	43	40	38	36	36	34	31	29
70	27	25	22	18	16	13	9	7	7	4
80	4	—	—	—	—	—	—	—	—	—

^aThe rows and columns represent the amount of nonerodible soil. For example, if nonerodible soil is 25%, first find the 20 in the Tens row and then the 5 in the Units column to find the erodibility index of 193 mg ha⁻¹.

Adapted from the USDA, NRCS (2002) *National Agronomy Manual*. 3rd edn. Title 190 Part 502. Wind Erosion, USDA Natural Resources Conservation Service. Available online at: <https://efotg.sc.egov.usda.gov/>.

on the surface of unconsolidated tilled soils to form a relatively thin physical crust. The crust is denser and more resistant to abrading particles than the unconsolidated soil immediately below it. Very sandy soils form very weak crusts that are easily destroyed by blowing sand particles. However, soils with even a small amount of clay and silt will bind together to form a layer that is 40–70 times as resistant to erosion as entirely erodible soil of the same texture. In rangeland conditions, algae, bacteria, fungi, and other organisms may form biological crusts that are very resistant to wind erosion until disturbed.

In sandy soils, a thin layer of loose, highly erodible sand may form on the surface of a physical crust. This loose sand is very susceptible to the wind and starts to move at relatively low wind speeds. Such loose material will abrade crusts until they fail, typically resulting in a more erodible surface. The amount and distribution of this loose, erodible sand on the soil surface are related to the potential wind erosion of a site. If little or no loose erodible sand is present, erosion generally does not occur.

Vegetation

Vegetation, whether growing, dead, standing, or flat, reduces the susceptibility of the soil surface to erosion. Standing vegetative cover effectively raises the roughness height (z_0) and reduces the shear stress acting on the surface behind the barrier. Standing biomass stems absorb more wind energy than flat vegetative material and are more than twice as effective at reducing wind energy than flat residues. The effectiveness of standing biomass at reducing erosion increases with the height, density, and stiffness of the plants). Modern harvesting equipment such as finger headers on grain combines has resulted in taller grain residue and better wind erosion control. Flat plant residue, while not greatly affecting the value of z_0 , also serves to protect the soil surface from erosion. Dense flat residue and low-growing vegetation effectively shelter the surface against the force of the wind, and thus protect the underlying soil surface. Where the flat residue is less dense, it acts much like random soil roughness elements. If saltation is initiated in a localized area of the field, the saltating grains may strike relatively nonerodible plant material and settle to the ground, effectively stopping avalanching saltation. Where adequate water from rainfall or irrigation is available, winter-cover crops are often planted as a wind-erosion control technique. The adoption of no-till cropping systems throughout the world, most notably in the North American Great Plains, South America, and Australia (Derpsch et al., 2010), has resulted in improved residue management and better control of wind erosion. However, the advent of the biofuel industry and harvesting of crop residues to make ethanol has the potential to reverse this healthy trend. In cropped fields, the crops and resulting crop residues are relatively uniformly distributed. In native plant communities however, the protection afforded by vegetation may be spatially variable. Recent trends in vegetation on some grazing lands are moving away from uniform grasslands toward communities composed of shrubs with bare soil surfaces in between. In such communities, the bare areas tend to form areas of highly erodible soils and the shrubs capture the sediment resulting in nutrient islands and elevated soil surfaces. These isolated elements may funnel wind over the bare areas, resulting in more soil movement. Although the redistribution of soil in such situations is obvious, little is known about net soil erosion in native plant communities.

Wind erosion modeling

Models provide a means to improve knowledge of the many factors that contribute to the processes of wind erosion. In addition, models not only allow users to forecast soil loss given the land condition, they can also be used to evaluate the processes and interactions of climate, land condition, and management on soil loss, without the need for time-consuming and costly field measurements. Several wind erosion models for agricultural as well as natural conditions have been developed over the past decades that vary in degree of complexity and capabilities. A review of many established wind erosion models was provided by Jarrah et al. (2020). These models vary in degree of complexity, temporal and spatial scales simulated, and details of inputs and outputs. A few examples of the more widely used models are discussed below.

The first wind erosion prediction model that was widely adopted in the U.S. was the Wind Erosion Equation (WEQ), developed by the USDA (Woodruff and Siddoway, 1965). The WEQ was developed to predict long-term average annual soil loss from farm fields. The WEQ has been used by the USDA, NRCS to determine soil loss due to wind erosion and to aid producers in designing control practices as well as for various farm programs and the U.S. Natural Resource Inventory (USDA, 2020). The general form of the equation, $E = f(I, K, C, L, V)$, expresses soil loss (E) as functionally related to soil erodibility (I), ridge roughness (K), a climatic index (C), field length (L), and a vegetative factor (V). In practice, the I value is determined from a table similar to Tables 1 or 2 and a series of charts and nomographs are used to apply the other factors in the equation.

The Wind Erosion Prediction System (WEPS) model was developed by the USDA beginning in 1985 and was released to the public in 2010. The overall objective of the WPES project was to take advantage of state-of-the-art wind erosion research and computer technology for improved predictions of wind erosion for conservation planning and wind erosion control (Tatarko et al., 2019). The WEPS is designed to make daily or shorter-time-interval estimates of wind erosion. It is a modular computer-based model that simulates soil erodibility, plant growth, and residue decomposition as affected by weather and field management to predict cropland soil loss at the field scale. The model uses a series of complex submodels that account for the effects of weather, crop growth and decomposition, soil surface properties, soil hydrology, land management, and erosion. In general, the model simulates the weather, grows a specified crop, and determines when erosion occurs based on the soil condition at the time that winds exceed the threshold velocity. Inputs can be as simple as providing the field location (i.e., for weather), the soil present, and a list of management practices applied to the field. The output includes estimates of saltation plus creep and suspended sediment, including PM10 and PM2.5. As a process-based model, WEPS can be used globally with inputs described above without the need for

calibration to a specific location. The WEPS has been applied in studies for Argentina, Burkina Faso, Canada, China, Germany, Kazakhstan, Mexico, and Sweden (Tatarko et al., 2019).

Webb et al. (2006) developed the Australian Land Erodibility Model (AUSLEM). This daily time-step model resulted from a systems analysis of the factors controlling wind erosion. The AUSLEM was developed as a geographic information system tool to assess land susceptibility to wind erosion across western Queensland, Australia. Required inputs include plant cover, soil moisture, soil texture and soil surface roughness. The model was found to perform well in the arid southern and western regions of Queensland.

The Aeolian EROsion (AERO) model was recently developed in the U.S. to assess the impacts of land management and environmental change on rangelands (Edwards et al., 2022). The AERO model uses standardized soil and vegetation monitoring data to develop indicators for wind erosion effects on land health and air quality assessments. The model approach applies a Generalized Likelihood Uncertainty Estimation (GLUE) framework for calibration using vegetation transect, sediment transport (i.e., flux), and meteorological data collected from standardized wind erosion monitoring sites. The model results suggest AERO effectively represents temporal variability in eolian transport.

Concerns about the effects of dust on global climate change has created interest in using models to estimate atmospheric dusts, produced primarily by desert or semiarid ecosystems. Advances in the understanding of atmospheric transport of soil-derived dusts have been made by laboratories in France and Australia. Scientists at the Laboratoire Interuniversitaire des Systèmes Atmosphériques, Creteil, France have developed a dust production model (DPM) considering the size distribution of erodible particles and surface roughness as controls on dust emissions (Alfaro and Gomes, 2001). Scientists with the Commonwealth Scientific and Industrial Research Organization (CSIRO) in Australia have developed a model accounting for the effects of soil, climate, and vegetation to estimate sand drift intensity and dust emission for use at scales ranging from paddocks to regions (Shao et al., 1996). The model known as the Wind Erosion Assessment Model (WEAM), has been coupled with a geographic information system to estimate dust emission over a large area of Australia. Modeling wind erosion in rangelands is a particular challenge due to the complex interaction of vegetation and the wind on the soil surface.

Wind erosion measurement

Measurement of wind erosion is necessary to understand the underlying processes as well as determine the effects of control measures, validate models, and assess the intensity of erosive forces. A wide variety of devices has been developed for this purpose (Webb et al., 2019). Wind erosion assessments typically examine two factors that affect the magnitude of soil loss: (i) erosivity of the climate and (ii) erodibility of the soil surface. Climate erosivity entails standard meteorological measurements for wind, relative humidity, temperature, and precipitation (Webb et al., 2019). Wind velocity is measured as a profile with typically 3 to 6 anemometers positioned at various heights above the surface (Fig. 6). This arrangement is also needed to calculate the aerodynamic roughness height (z_0) of the surface and the wind shear velocity (u^*).

It is important at this point, to distinguish the difference between horizontal soil flux and soil erosion, both of which are reported in erosion research literature. Horizontal soil flux is the mass of sediment passing a vertical plane at a given point on the landscape (e.g., kg m^{-2}). There is usually little knowledge of where on the surrounding landscape the flux began or the surface conditions for which the emissions occurred. Therefore, horizontal flux is typically further used to derive soil erosion, which is the mass loss of soil per unit area of ground surface (e.g., also kg m^{-2}). Horizontal flux can also be used to calculate sediment deposition. Estimates of soil redistribution by wind are often based on the vertical distribution of sediment in transport at multiple spots within a field. If the total sediment in transport at one location is greater than that at an upwind location, net erosion is indicated for the surface interval between the samplers. If the inverse is true, then deposition is indicated for the interval between the samplers. Many types of samplers have been developed to measure the transported sediment at multiple heights above the soil surface and the relative merits of these samplers have been compared (Goossens et al., 2000). The ideal sampler has a high sampling efficiency of collecting sediment without disturbing the wind stream or developing suction or back-pressure. Sampling efficiency for most samplers usually decreases with particle size in the wind stream. Since about 50% of the blowing sediment is transported within 2 cm of the soil's surface, sampling close to the surface is desirable to make accurate estimates of transported material.

A combination creep and saltation sampler was developed by the USDA to sample very near to the soil surface, including creep size, at heights of 0–3, 3–9, and 9–20 mm (Zobeck, 2002). The sampler rotates into the wind and each inlet has a separate collection pan. This sampler is used in combination with other samplers such as Big Spring Number Eight (BSNE) or Modified Wilson and Cooke (MWAC) samplers (discussed below) to analyze the entire saltation zone.

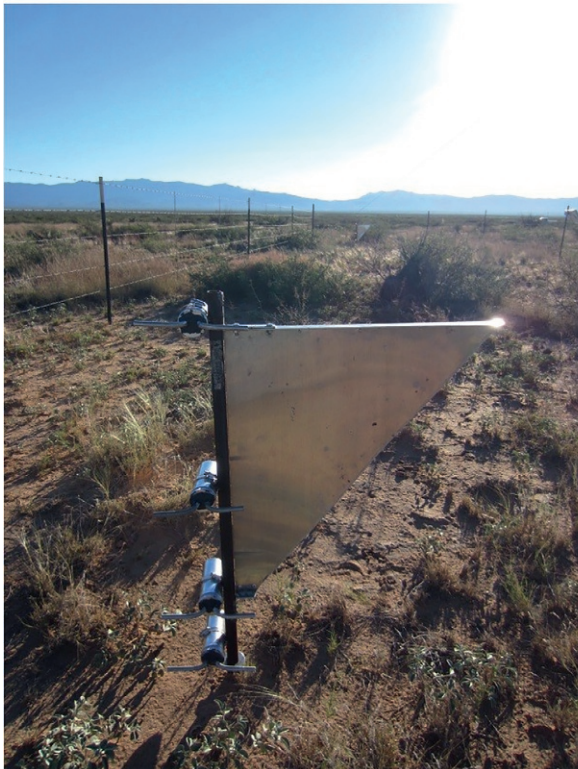
Several types of samplers have been developed to sample the saltation zone, which is about 1–2 m above the soil surface. The BSNE samplers are wedge-shaped samplers with 60-mesh screens on the top or sides to relieve back-pressure and 2-cm-wide and 5-cm-high inlets (Fryrear, 1986). Several BSNE samplers are mounted on a pole to sample at various heights. The samplers have wind vanes to ensure they are oriented into the wind. A similar sampling scheme, the MWAC catchers use bottles mounted on a pole with glass or metal tubes as inlets and outlets (Wilson and Cook, 1980). The larger volume within the bottle allows wind velocities to diminish and eolian sediment to settle within the bottle. This highly versatile and inexpensive sampler may also be modified and mounted to assess the directions from which eolian transport occurs (Fig. 7).

The samplers discussed above are termed passive samplers because they require no external source of power to function. Although these samplers may be used to integrate the sediment transport over periods of time between field visits and sediment



Fig. 6 A Meteorological tower with six anemometers for measuring the wind profile. Photo credit: USDA-ARS.

(A)



(B)



Fig. 7 Two aspects (A) and (B) of the Modified Wilson and Cook (MWAC) wind erosion sediment sampling system. Photo credit: USDA-ARS.



Fig. 8 A Sensit device used to measure saltation impacts of moving soil grains. Photo credit: USDA-ARS.

collection, they reveal little about the temporal nature of the sediment transport. For this purpose, saltation impact sensors with piezo-electric elements such as the Sensit or microphone capsules such as the Saltiphone are commonly used (Fig. 8). These sensors allow researchers to determine at what wind speed and z_0 wind erosion begins.

Sampling of the suspended dust load requires more sophisticated instrumentation. Sharratt and Pi (2018) compared the performance of five instruments for measuring suspended PM10 above eroding soils. Optical-based instruments such as the DustTrak are often used in the field to record real-time dust fluxes at a specific height above a point on the landscape. More expensive and larger instruments have also been developed for long-term monitoring of atmospheric dust. The model used by the U.S. Environmental Protection Agency is the Tapered Element Oscillating Microbalance (TEOM). It draws a constant volume of air through a small filter and weighs the filter at predetermined intervals.

While wind erosion measurement devices and samplers give good estimates for a given event or short-term loss, they are rarely employed for periods exceeding weeks or months. Until the last 50 years, there has been no reliable way to estimate long-term rates of wind erosion. However, atmospheric testing of nuclear weapons in the 1950s and 1960s released radioactive isotopes that were distributed worldwide and deposited, primarily by rainfall, on the soil surface. Many of these isotopes, including ^{90}Sr strontium, ^{137}Cs cesium, and $^{239/240}\text{Pu}$ plutonium, are strongly adsorbed to the surfaces of soil clays and organic matter. Strong isotope pulses in 1959 and 1963–64 created a very sharp rise in soil borne radioisotope tracers. Transport of these soil borne tracers is almost exclusively by actual physical movement of the particles to which they are adsorbed. Radioisotope tracers have been used for the last three decades to study the long-term redistribution of wind eroded soil since the radioisotopes were deposited ending in 1964. Radioisotope activity of soil cores sampled at an undisturbed reference site such as native grassland is compared to the activity of soil cores sampled in a field or site for which the long-term estimate of soil redistribution is desired. The loss or gain of radioisotope activity relative to the reference site has been shown to be proportional to the erosion or deposition of soil at the site of interest. Computer models that determine soil redistribution rates from radioisotope activities in different landscapes were developed and validated against, carefully measured wind erosion studies on the U.S. Southern High Plains by Van Pelt et al. (2007).

Wind erosion control

Any control treatment that reduces the wind at the soil surface to below that of the threshold of particle movement, will control wind erosion. A wide variety of practices have been developed, which have been summarized below. The more common control practices for wind erosion are briefly described below.

Vegetation and vegetative residues

The easiest and most cost effective control of wind erosion for most situations is to establish and maintain adequate vegetative ground cover. The value of plant materials to reduce wind erosion has been recognized for several decades. Any practice that keeps live crops or preserves crop residues on the surface after harvest will reduce potential wind erosion (Fig. 9). Live plants and crop residues protect the soil from wind by absorbing a portion of the direct force of the wind, trapping moving soil particles, and promoting aggregation of soil particles. Planting crop rows perpendicular to the prevailing winds is more effective in the control of



Fig. 9 (A) Vegetative wind erosion control includes keeping live plants on the surface and (B) maintaining sufficient residues after harvest. Photo credits: USDA-NRCS.

wind erosion than rows parallel to the wind. In addition, practices that leave vegetation in a standing orientation are more effective in controlling wind erosion than residues that are flat. Reducing excessive tillage (i.e., reduced or no-tillage management) will maintain surface residues and generally reduce wind erosion. Maintaining a soil surface area cover of 30% has been shown to reduce soil erosion to tolerable levels, with only marginal improvements in soil loss prevention achieved with higher percentages of cover (Fryrear, 1985).

Barriers

Wind erosion tends to increase downwind with field distance due to the avalanche effect, therefore, an effective reduction in field length is often accomplished by windbreaks or barriers. Barriers may consist of single rows of trees or shrubs (Fig. 10A) or elaborate windbreaks called shelter belts that have multiple rows of vegetation, including trees and shrubs (Fig. 10B). The use of trees planted in windbreaks or shelter belts to control wind erosion dates back to the U.S. Dust Bowl of the 1930s. The effectiveness of windbreaks depends on their height, porosity, and orientation to prevailing winds. Barriers planted upwind and perpendicular to prevailing winds are more effective than those parallel or downwind of the field. Barriers have been shown to reduce wind speeds for as much as 30 times the height of the barrier (Hagen and Skidmore, 1971). As a general rule however, wind breaks reduce wind erosion for a distance of up to 10 times their height in the downwind direction and three times their height in the upwind direction. Barriers also capture snow, reduce evapotranspiration, improve crop yields, shelter homes and livestock from the wind, and provide fuelwood and wildlife habitat (Cable, 1999). Unfortunately, the number of effective shelter belts has greatly diminished and many have been intentionally removed (Sorenson and Marotz, 1977). More recently, a large-scale tree planting in China was begun with the Three-North Shelterbelt Programme, an effort to establish eventually 4.06 million km² of trees along the border of northern China's encroaching desert by 2050. Similar projects have been attempted across the Sahara Desert in northern Africa in Burkina Faso, Nigeria, and Ethiopia with only limited success (Goffner et al., 2019).

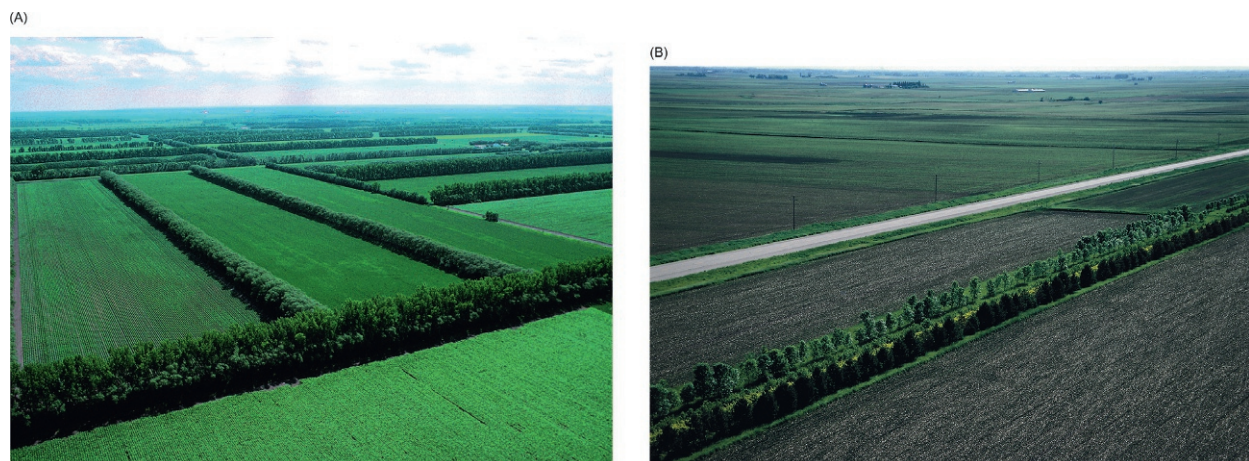


Fig. 10 (A) Field windbreaks in northwest Iowa and (B) multiple row windbreak or shelter belt. Photo credits: USDA-NRCS.

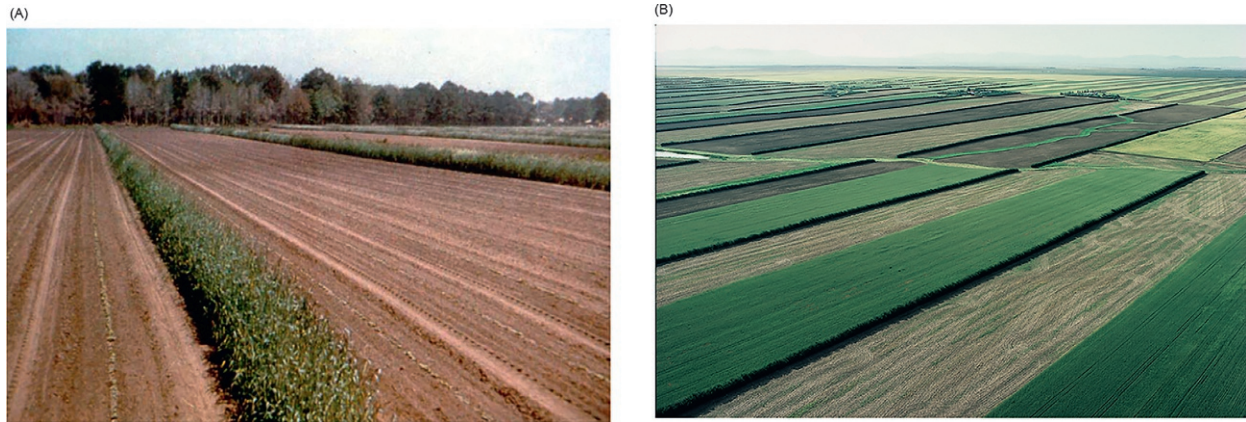


Fig. 11 (A) Perennial grass barriers protect cucumbers in South Carolina and (B) an example of strip cropping in Montana consisting of live wheat alternating with fallow areas. Photo credits: USDA-NRCS.

More recently, annual crops have been used as more closely spaced shelter belts within a field. Spacing of these ‘wind strips’ varies from two rows of wheat planted in the furrows between rows up to 5–10 m-wide strips of vegetation planted at 100–300-m intervals throughout the field (Fig. 11A). Widths of greater than 4.5 m of such herbaceous barriers also trap saltating particles to provide additional protection from the effects of wind erosion. Other advantages of these types of barriers are their ease of establishment and low cost. In emergency situations, other physical barriers such as a snow fence may be used. The recommended porosity of barriers is about 40% for best results (Hagen and Skidmore, 1971; Papesch, 1992).

Strip cropping

Strip cropping involves planting crops in alternating strips of crops susceptible to wind erosion with strips having a cover resistant to wind erosion (Fig. 11B). This practice shortens the field length in the direction of the prevailing wind, thus reducing the avalanche effect and trapping moving particles in strips with cover. Like many practices, the strips should be oriented perpendicular to the prevailing erosive winds. Wheat-fallow systems sometimes use this practice where the fallow strips alternate with growing wheat one year and are interchanged the following year.

Soil aggregation and roughness

In some situations, it is not possible to maintain a vegetative cover on the soil surface due to crop type (e.g., cotton or beans) or management practices (e.g., tillage for weed control). In such circumstances, tillage may be used to maintain suitable soils in a cloddy condition (Fig. 12). The goal is to create soil aggregates large enough to resist the wind force and trap smaller moving



Fig. 12 Tillage in this field produces a rough cloddy condition that is resistant to wind erosion. Photo credit: USDA-ARS.



Fig. 13 Ridges can provide wind erosion control when oriented perpendicular to the prevailing erosive winds. Photo credit: USDA-ARS.

particles. Soil texture is an important factor in creating wind resistant clods. Coarse textured soils such as sands do not have enough clay and silt to bind particles together to form aggregates. Soils with higher clay and silt contents, (e.g., loams, silt loams, clay loams) consolidate to form strong aggregates that are somewhat more resistant to erosive winds and saltating grains. Care must be used to ensure tillage is done when the soil is moist enough to create clods. Tilling dry soils may exacerbate the problem by further pulverizing the soil and creating even more erodible particles.

Ridges and furrows perpendicular to the prevailing erosive wind direction provide improved protection compared to simple random roughness from soil clods. Ridge and furrow systems also trap drifting soils in the lower position of furrows, thus reducing the advance of erosion across a field (Fig. 13). While a cloddy soil surface will absorb more wind energy than a flat, smooth surface, a soil surface that is both ridged and cloddy will absorb even more wind energy. However, roughness can degrade over time as the result of rainfall, freezing and thawing, as well as the wind itself and should not be considered a permanent control.

Chemical treatments

Chemical treatments, such as applying polyacrylamide or plant processing byproducts have also been used to control soil blowing and suppress dust (Movahedan et al., 2012). These chemical treatments bind the surface soil particles together. Any disturbance to the treated areas greatly reduces the effectiveness of the treatment, however. For example, precipitation or saltating soil from adjacent untreated areas that may abrade the treated areas will cause the chemical treatment to fail. Chemical treatments also tend to be expensive and are usually restricted to smaller, high-value crops or areas.

Emergency control

The controls presented so far are designed to be preventative measures. However, once a field actively begins to erode, emergency controls may need to be applied as soon as possible to avoid damage to the soil, crop, and the environment. The various methods of emergency control focus on stopping the process of erosion that is active. Addition of residues, either plant or animal (e.g., manure) have been successfully used to control active wind erosion. Typically plant residues require anchoring with a stubble puncher or disk to prevent the residue from blowing away in strong winds. Another method is emergency tillage, which is done perpendicular to the wind to create a rough, cloddy, or even ridged surface to reduce the wind force and trap moving particles. Implements such as listers, chisels, shovels, and sandfighters are often used for emergency tillage to roughen the soil surface and create clods. This is typically used only when other measures have failed and is considered only a temporary control.

Summary

Scientists at various universities and government agencies throughout the world continue to research the processes, causes, and control of soil erosion by wind, which continues to threaten a sustainable food supply worldwide. This research has the ultimate goal of developing better control strategies for improved management for a sustainable agriculture that protects the soil, air, and water resources. A comprehensive understanding of the principles of wind erosion is needed for developing control strategies. However, extreme weather resulting from climate change can cause drought and shortage of protective vegetation that will exacerbate the problem.

Appendix: Supplementary material

Supplementary material related to this chapter can be found on the accompanying CD or online at <https://doi.org/10.1016/B978-0-12-822974-3.00006-9>.

References

- Alfaro SC and Gomes L (2001) Modeling mineral aerosol production by wind erosion: Emission, intensities and aerosol size distributions in source area. *Journal of Geophysical Research* 106: 18075–18084.
- Bagnold RA (1941) *The Physics of Blown Sand and Desert Dunes*. New York: William Morrow.
- Cable TT (1999) Nonagricultural benefits of windbreaks in Kansas. *Great Plains Research: A Journal of Natural and Social Sciences* 417: 41–53.
- Clow DW, Williams MW, and Schuster PF (2016) Increasing aeolian dust deposition to snowpacks in the Rocky Mountains inferred from snowpack, wet deposition, and aerosol chemistry. *Atmospheric Environment* 146: 183–194.
- Derpsch R, Friedrich T, Kassam A, and Hongwen L (2010) Current status of adoption of no-till farming in the world and some of its main benefits. *International Journal of Agricultural and Biological Engineering* 3: 1–25.
- Edwards BL, Webb NP, Galloza MS, et al. (2022) Parameterizing an aeolian erosion model for rangelands. *Aeolian Research*. (In Review).
- FAO (2019) *Soil Erosion: the Greatest Challenge to Sustainable Soil Management*. Rome: Food and Agriculture Organization of the United Nations. 104 pp. License: CC BY-NC-SA 3.0 IGO.
- FAO and ITPS (2015) *Status of the World's Soil Resources – main report*. In: *Food and Agriculture Organization of the United Nations and Intergovernmental Technical Panel on Soils*. Rome, 649 pp. Available online at: <http://www.fao.org/3/a-i5199e.pdf>.
- Fryrear DW (1985) Soil cover and wind erosion. *Transactions of the American Society of Agricultural Engineers* 28: 781–784.
- Fryrear DW (1986) A field dust sampler. *Journal of Soil and Water Conservation* 41: 117–120.
- Goffner D, Sinare H, and Gordon L (2019) The Great Green Wall for the Sahara and the Sahel Initiative as an opportunity to enhance resilience in Sahelian landscapes and livelihoods. *Regional Environmental Change* 19: 1417–1428.
- Goossens D, Offer Y, and London G (2000) Wind tunnel and field calibration of five aeolian sand traps. *Geomorphology* 35: 233–252.
- Hagen LJ and Skidmore EL (1971) Turbulent velocity fluctuations and vertical flow as affected by windbreak porosity. *Transactions of the American Society of Agricultural Engineers* 14: 634–637.
- Hagen LJ, Schroeder B, and Skidmore EL (1995) A vertical soil crushing-energy meter. *Transactions of the American Society of Agricultural Engineers* 38: 711–715.
- Huszar PC and Piper SL (1986) Estimating the off-site costs of wind erosion in New Mexico. *Journal of Soil and Water Conservation* 41: 414–416.
- Jarrah M, Mayel S, Tatarko J, Funk R, and Kuka K (2020) A review of wind erosion models: data requirements, processes, and validity. *Catena* 187: 104388.
- Movahedan M, Abbasi N, and Keramati M (2012) Wind erosion control of soils using polymeric materials. *Eurasian Journal of Soil Science* 2: 81–86.
- Papesch AJG (1992) Wind tunnel test to optimize barrier spacing and porosity to reduce wind damage in horticultural shelter systems. *Journal of Wind Engineering and Industrial Aerodynamics* 44: 2631–2642.
- Pimentel D and Burgess M (2013) Soil erosion threatens food production. *Agriculture (Basel, Switz.)* 3: 443–463.
- Prospero JM and Lamb PJ (2003) African droughts and dust transport to the Caribbean: climate change implications. *Science* 302: 1024–1027. <https://doi.org/10.1126/science.1089915>.
- Shao Y, Raupach RM, and Leys JF (1996) A model for predicting aeolian sand drift and dust entrainment on scales from paddock to region. *Australian Journal of Soil Research* 34: 309–342.
- Shao Y, Wyrwoll KH, Chappell A, et al. (2011) Dust cycle: An emerging core theme in Earth system science. *Aeolian Research* 2: 181–204.
- Sharratt B and Pi H (2018) Field and laboratory comparison of PM10 instruments in high winds. *Aeolian Research* 3: 42–52.
- Sorenson CJ and Marotz GA (1977) Changes in shelterbelt mileage statistics over four decades in Kansas. *Journal of Soil and Water Conservation* 32: 276–281.
- Tatarko J, Wagner L, and Fox F (2019) The wind erosion prediction system and its use in conservation planning. In: Wendroth O, Lascano RJ, and Ma L (eds.) *Bridging Among Disciplines by Synthesizing Soil and Plant Processes, Advances in Agricultural Systems Modeling*. vol. 8, pp. 71–101. Madison, WI: ASA, CSSA, and SSSA. <https://doi.org/10.2134/advagricsystmodel8.2017.0021>.
- USDA (2020) Summary Report: 2017 National Resources Inventory. In: *Natural Resources Conservation Service, Washington, DC, and Center for Survey Statistics and Methodology*. Ames, Iowa: Iowa State University Available online at: <https://www.nrcs.usda.gov/>.
- USDA-NRCS (2002) National Agronomy Manual. In: *Title 190 Part 502. Wind Erosion*, 3rd edn USDA Natural Resources Conservation Service Available online at: <https://efotg.sc.egov.usda.gov/>.
- Van Pelt RS, Zobeck TM, Ritchie JC, and Gill TE (2007) Validating the use of 137Cs measurements to estimate rates of soil redistribution by wind. *Catena* 70: 455–464.
- Van Pelt RS, Tatarko J, Gill T, et al. (2020) Dust source characterization on Lordsburg Playa in Southwestern New Mexico. *Geoenvironmental Disasters* 7: 34.
- Webb NP, McGowan HA, Phinn SR, and McTainsh GH (2006) AUSLEM (Australian Land Erodibility Model): a tool for identifying wind erosion hazard in Australia. *Geomorphology* 78: 179–200.
- Webb NP, Herrick JE, Van Zee JW, et al. (2019) Standard Methods for Wind Erosion Research and Model Development: Protocol for the National Wind Erosion Research Network. USDA-ARS Jornada Experimental Range, P.O. Box 30003, MSC 3JER, NMSU Las Cruces, New Mexico 88003. 75 pp. Available online at: <https://winderosionnetwork.org/>.
- Wilson SJ and Cook RU (1980) Wind erosion. In: Kirkby MJ and Morgan RPC (eds.) *Soil Erosion*, pp. 217–251. Chichester: Wiley.
- Woodruff NP and Siddoway FH (1965) A wind erosion equation. *Soil Science Society of America Proceedings* 29: 602–608.
- Zobeck TM (2002) Field measurement of erosion by wind. In: *Encyclopedia of Soil Science*, pp. 503–507. New York: Marcel Dekker, Inc.

Human exposure to aeroparticles emitted from unpaved roads due to traffic resuspension and wind erosion

Arón Jazcilevich^a and Ann G Wellens^b, ^aInstituto de Ciencias de la Atmósfera y Cambio Climático, UNAM, Ciudad de México, Mexico;
^bFacultad de Ingeniería, UNAM, Ciudad de México, Mexico

© 2023 Elsevier Ltd. All rights reserved.

Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the Universidad Nacional Autónoma de México.

Introduction	223
Exposure to air pollution	224
Road dust health effects	226
Road dust emissions	227
Road traffic resuspension emissions	227
Road wind erosion	227
Roadside sampling methodologies	228
Manual sweeping and vacuum cleaning	228
Direct sampling	228
Wet dust sampler (WDS)	228
Testing re-entrained aerosol kinetic emissions from roads (TRAKER) and EMMA	228
SNIFFER	228
System for the continuous aerosol measurement of particulate emissions from roads (SCAMPER)	229
Online resuspension chamber	229
Portable in situ wind erosion laboratory	229
Vertical profile	229
Active concentration profilers	230
Emission factors for road wind erosion	231
From dust to dose	231
Final considerations	232
References	234
Further reading	235

Abstract

Resuspended dust emitted by unpaved roads contains particles that penetrate the respiratory system. When these particles come from contaminated soils, toxins reach the brain or blood circulatory system, generating health disorders. Scientific studies assessing the effect on human health, and the methodologies to determine their roadside transport, together with their physical and chemical characteristics, are presented. Techniques to collect road dust samples due to vehicular and wind erosion are described. Transport models relating vehicular and road wind erosion resuspension and corresponding human exposure are also discussed. Some of the literature results indicate that children and the elderly are at special risk.

Key points

- Road dust, generated by traffic resuspension and wind erosion, affects human health.
- Health effects depend on the chemical composition of the road surface.
- The smaller the particle size of road dust, the more dangerous to human health.
- Children are at special risk from road dust emissions.
- Vehicular resuspension exposure decreases rapidly with increasing distance from the road.

Introduction

Road dust consists of solid particles that are generated by mechanical processing of materials, including crushing, grinding, rapid impact, handling, detonation, and decrepitation of organic and inorganic materials such as rock, ore, and metal (Khan and Strand, 2018). This road dust, or aeroparticles, becomes airborne when local windspeed is above a certain threshold, or by traffic resuspension caused by vehicular aerodynamic wake and shear stresses on the road surface set off by vehicular traction.

While road dust from paved roads contains mainly minerals, brake wear, tire wear and carbonaceous emissions, unpaved road emissions contain the same chemical compounds as those found on the road surface. In general, road aeroparticles emissions are associated with vehicle traffic and road moisture content. Dust emissions from unpaved roads have been found to vary directly with the fraction of silt (particles smaller than 75 μm in diameter) in the road surface materials. Typical silt contents vary according to the type of unpaved road. Industrial unpaved roads contain from 0.2% to 29%, while rural dirt roads with compacted, bladed, and crowned local material have still higher silt contents.

Many communities and regions around the world with predominantly unpaved roads are affected by road-dust. The latest shale oil boom, starting in 2007 in the North Dakota Bakken region, has resulted in a renewed interest by the scientific community in road dust, since many of the oil fields are communicated by dirt and gravel roads. Besides oil, other industrial activities such as mining are major sources of trace elements contained in road-dust, mostly due to waste and emissions during the processing of the extracted materials such as tailings, waste rock deposits and smelting operations (Rodríguez Martín et al., 2014).

Table 1 shows some examples of the chemical compounds found near unpaved roads, depending on the characteristics of the surroundings. Typically, the largest metal concentrations are found in the finest size fractions of the road dust. The partial list of examples in Table 1 point out that the chemical compounds found in road dust depend on the soil composition and industrial activity of the surroundings.

The exploitation of coal, gold, uranium, wolfram/tungsten, tin, platinoids, and polymetallic sulfides cause the most severe cases of soil pollution whenever mining is carried out. Pollution of soil from trace elements at sites surrounding coal mines were reported to be the highest in Barapukuria in Bangladesh, Ledo in India, Ptolemais-Amynteon in Greece, and the Tibagi River in Brazil (Sahoo et al., 2016). Gold mining operations can cause emissions of mercury (Hg), cyanide, arsenic (As), zinc (Zn), lead (Pb), nickel (Ni), and other toxic elements.

During the transfer and transport of ore concentrates to the smelter, spillages are a potential source of road soil pollution. Particulate and gaseous emissions from a smelter that lacks appropriate abatement technologies is a local and distant source of trace elements pollution of soils (FAO and UNEP, 2021).

Many rocks are naturally rich in trace elements, which are mobilized when extracting them from the ore reservoirs. Mining wastes still contain trace elements, but in concentrations at which it is not cost-effective to extract them, but which may still pose a risk to the environment and human health. This is the case of Manganese (Mn) in the mining district of Molango, Hidalgo, Mexico, where the mineral waste contains 18% of Mn (Riojas et al., 2009).

Exposure to air pollution

Exposure refers to a contact between an airborne contaminant with the human skin or the respiratory system. Therefore, exposure requires the simultaneous occurrence of two events: a pollutant concentration at a particular place and time, and the presence of a person at that place and time. Exposure is expressed quantitatively by a description of the duration of the contact and the relevant pollutant concentration. It can be acute if the exposure lasts seconds or minutes, or it can be chronic if it lasts days or months. The accumulation of repeated acute exposures leads to chronic exposure.

Concentration and exposure are different concepts. Concentration is the amount of a substance per volume or mass of air at a certain time and place, whereas exposure describes an interaction between the environment and a subject. Also, there is a distinction between exposure and dose. Dose is the amount of the pollutant that crosses one of the body's boundaries and reaches a target tissue. For example, it can be obtained by dividing the mass or volume of a pollutant by the blood volume.

The size and chemical composition of road dust are the main components to assess its health impact. The size of the particles contained in dust determines where these are allocated in the respiratory system. Particle size is described using the equivalent aerodynamic diameter, D , defined by:

$$D = D_g k \sqrt{\frac{\rho}{\rho_0}},$$

where D_g is the geometric diameter, ρ is the aerosol density, ρ_0 is a reference density (1 g cm^{-3}) and k is the shape factor, equal to 1 in the case of a perfect sphere. Essentially, the aerodynamic diameter of an arbitrary particle is equal to the diameter of a spherical particle with a density of 1 g cm^{-3} , which has the same inertial properties in the aerosol as the particle of interest.

As shown in Fig. 1, a particle whose equivalent aerodynamic diameter is 5–10 μm , or PM_{5-10} , is respirable or inhaled, i.e., it enters the respiratory system. In this overall-used notation, PM stands for *particulate matter*. PM_{1-5} material is allocated in the thoracic region of the lung, whereas PM_1 reaches the alveolar level. Here is where interchange between oxygen and CO_2 takes place, and the pollutant may reach and enter the blood stream.

Besides size, the chemicals contained in the particles in solid or liquid phase determine toxicity. This is measured by the hazard index, HI. The hazard index is defined as “the sum of hazard quotients for substances that affect the same target organ or organ system” (US-EPA, 2021). The hazard quotient is “the ratio of the potential exposure to the substance and the level at which no adverse effects are expected” (US-EPA, 2021). The HI is used to indicate the approximate effects of an agent on an organ system.

Table 1 Chemical compounds reported in road dust near unpaved roads.

<i>Road type</i>	<i>Soil composition</i>	<i>Location</i>	<i>Chemical elements/compounds determined in the road dust</i>	<i>Autors</i>
Petrochemical complex construction site	Not specified	Central Taiwan	• Si (10.3%), Al (4.2%), Ca, Fe, K and organic carbon (13%) in the fine fraction. • In the coarse fraction, Si and Al concentrations increment about 2.5 times.	Wei et al. (1996)
Unpaved roads and trails in 29 military and civilian sites	Different soil types, composition not specified	Western United States	• Si, Ca, Al, Fe and K, concentrations depending on the site	Labban et al. (2004)
Public unpaved road; predominantly clay loam	Sand (19–39%), silt (39–56%), clay (22–26%)	New Mexico State University, United States	• C, O, Al, and Si. To a lesser extent Ca, K, Mg and Fe.	Williams et al. (2008)
Surroundings of power plants, impact from unpaved roads and neighboring cities	Not specified	Upper Silesia, Southern Poland	• Cr, Pb, Mn and Se as main trace elements in PM₁₀. • Around 66–70% of the trace elements were found in the PM_{2.5} fraction, unless for Pb, where it increased to 80%.	Zajusz-Zubek et al. (2018)
Unpaved roads in an abandoned mining site in a desertic area	Not specified	Sonora, México	• Fe (18000–20,000 mg kg⁻¹), Ti (2100–2800 mg kg⁻¹), Mn (850–1100 mg kg⁻¹), Zn (435–470 mg kg⁻¹), Pb (227 mg kg⁻¹), As (33–146 mg kg⁻¹), Sb (21–214 mg kg⁻¹), Cu (22–41 mg kg⁻¹) in the soil of unpaved roads. • Concentrations were approximately twice as much in road dust • Pb, As, Zn concentrations were above allowed limits for active mining waste	Del Rio-Salas et al. (2019)
Unpaved roads in farm fields, sandy loam	Sand (62–90%), silt (5–30%), clay (4–8%)	La Pampa, Argentina	• Glyphosate in PM₁₀ (inside farm fields 382–454 µg kg⁻¹; outside farm fields 39–639 µg kg⁻¹) • Aminomethylphosphonic acid (AMPA) in PM₁₀ (inside farm fields 900–4138 µg kg⁻¹; outside farm fields 98–500 µg kg⁻¹) • Organic matter <1%	Ramirez et al. (2021)

PM_{2.5} and PM₁₀ correspond to particulate matter with less than 2.5 and 10 µm equivalent aerodynamic diameter, respectively.

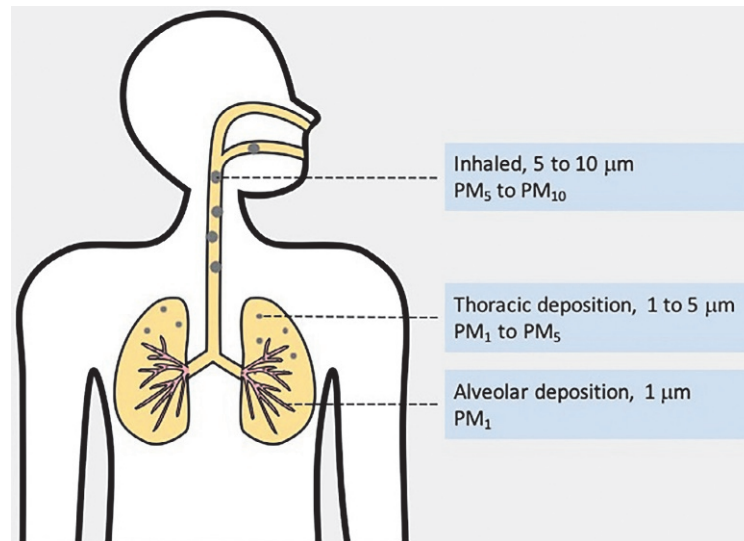


Fig. 1 Schematic showing particle deposition in the respiratory system, according to their size.

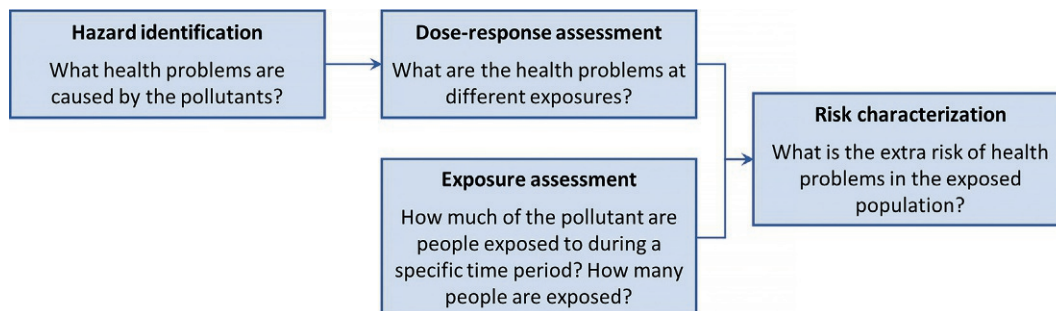


Fig. 2 The USA Environmental Protection Agency 4 step risk assessment process. Based on Khan RK and Strand MA (2018) Road dust and its effect on human health: A literature review. *Epidemiol Health* 40: e2018013. doi: 10.4178/epih.e2018013.

As shown in Fig. 2, exposure and dose response are part of the USA Environmental Protection Agency (EPA) risk assessment process. As such, most studies of wind erosion and traffic resuspension of unpaved roads with contaminated soils deal with the issues described in the assessment process boxes.

Road dust health effects

Although several studies have investigated the geochemical composition of road dust, relatively few scientific studies have assessed the effects of road dust on human health. As of 2018, a summary of the current knowledge of the scientific community on the effects of road dust on human health is presented in the review article by Khan and Strand (2018). Their main findings are summarized below.

The main organs affected by road dust are the respiratory and cardiovascular systems. Diseases related to the respiratory system are chronic obstructive pulmonary disease, asthma, fungal infections, allergies, and carcinomas. The health effects in the cardiovascular system are emergency cardiovascular diseases and increased mortality. Other diseases are low birth weight and non-specific carcinomas.

Among the chemical elements with higher HI often encountered in road dust are Pb, As, platinum group elements, aluminum, vanadium, palladium, and polycyclic aromatic hydrocarbons or PAH. Natural elements such as erionite and offretite were also found in some the studies.

Inhaled and ingested Pb compounds are associated with respiratory tract inflammation and may lead to respiratory tract cancer. Potgieter-Vermaak et al. (2012) found that Pb and chromium (Cr) compounds in road dust were present in human body fluids, indicating exposure risks to road dust. Pb is known to be responsible for deficits in neurobehavioral and cognitive development in childhood. Reports have also found that Pb exposure results in reproductive system dysfunction, and that microcytic anemia results in conditions such as hypertension and chronic renal failure.

Xu et al. (2013) found greater concentrations of As than other components in road dust near industrial and commercial areas. They found a higher HI for children near industrial plants, due to road dust. Additionally, the risk of cancer associated with As exposure via road dust was high in children. As has both carcinogenic and non-carcinogenic effects on humans. In a meta-analysis, As was found to be associated with cardiovascular disease, birth defects, neurologic and cognitive disorders, diabetes, ototoxicity, peripheral vascular diseases, and disorders of blood cells such as anemia and leukopenia (Tchounwou et al., 2003). This review also found an association between higher levels of As exposure and carcinomas of the skin, liver, lungs, kidney, colon, and urinary bladder.

Colombo et al. (2008) established an association between respiratory tract diseases and platinum-group elements in road dust. They reported that the bioavailability of the platinum-group elements depended on their concentration in road dust. Platinum (Pt) was found to have higher bioavailability in human gastric fluid than palladium (Pd) or rhodium. Zereini et al. (2021) found a higher concentration of platinum-group elements in PM₁₀ road dust and found that Pd had a greater concentration than Pt or rhodium. They reported that Pd was more soluble than other platinum-group elements, thus causing a potential health risk to humans. Transport activities due to manganese mining that are harmful to health are reported in Riojas et al. (2009), Jazcilevich et al. (2012) and Cortez-Lugo et al. (2018).

Long-term exposure to aluminum (Al) was found to be associated with Alzheimer's disease. Aluminum was found to be associated with respiratory allergies, such as asthma, in aluminum industry workers. The accumulation of Al can cause cardiac hypertrophy leading to cardiac failure. Al deposition in the body was found to cause inflammation in the hepatobiliary system, and to be associated with anemia and a low reproductive rate in rats.

Jiang et al. (2014) found high levels of PAHs in road dust. They estimated that the risk of cancer from PAH in road dust was higher through the dermal and ingestion routes than through inhalation. They found that children were more susceptible to the hazard than adults, due to their greater engagement in hand-to-mouth activities outdoors and smaller body weight.

Although inhalation is cited as a main route of exposure, ingestion and skin exposure to pollutants are also important. A vulnerable population often appearing in dust road pollution literature are children. For example, Mar et al. (2008), found a strong association between respiratory symptoms in children and PM_{2.5}.

Road dust emissions

The resuspension and wind erosion phenomena affect health because they entail high particulate matter concentrations reaching values in the mg m⁻³ scale, and since the emissions take place in near proximity of humans and their property. It should be clear that the phenomena described here involve scales of less than 150 m. These are appropriate scales to study contaminated soils affecting contiguous communities. The field of long distance transcontinental or regional erosion transport phenomena is not discussed here.

Road traffic resuspension emissions

A central factor of unpaved road aeroparticles is that particle resuspension concentrations are orders of magnitude higher than PM tailpipe emissions: if exhaust PM concentrations are measured in µg m⁻³, atmospheric concentrations emanating from dirt roads can reach the order of mg m⁻³ for short distances and periods of time. If the resuspended dust contains toxic material, its health effects can become acute, and their accumulation may lead to severe health effects.

Since PM tailpipe emissions have been substantially reduced by new internal combustion and aftertreatment technologies, resuspension emissions measurement has become more relevant. Moreover, with the advent of electric cars, resuspension is becoming a main emission source, together with brake and tire wear. Therefore, the main effort in the scientific community has been to develop techniques to evaluate resuspension emission factors for the purpose of building emission inventories for regulatory purposes. The aim has been to obtain how much mass per km is resuspended by a vehicle, and not necessarily to measure corresponding roadside concentrations pertinent to exposure studies. Nevertheless, some of the techniques used in the field of resuspension emissions measurement can be used for the purposes of roadside exposure, since they provide chemical and size distributions.

Mitigation methods for dust resuspension on unpaved roads include paving, speed control and chemical stabilization. Cortez-Lugo et al. (2018) present a case where road paving was used to decrease the amount of resuspended Mn in a mining community. Another frequently used and cheap stabilization measure is the use of dust suppressants, which can reduce PM₁₀ emissions by a factor of 5–60 (Fitz and Bumiller, 2021).

Road wind erosion

Soil erosion corresponds to the removal of detached material from the surface soil and subsequent short transport of particles. It is primarily originated by water and wind. The most important transport mechanism in wind erosion of soils is saltation, where particles (0.1–0.5 mm) fall to the ground and transfer their momentum to other soil particles in a series of short bounces along the surface of the ground, detaching and mobilizing them into the air (Avecilla et al., 2015).

The abrasion of a surface is related to the diameter of the saltating particles, which determines the kinetic energy of each impact. Larger particles manage to suspend particles with smaller sizes in the air. This is especially important in polluted soils, as toxic

compounds can be transported to the atmosphere by this mechanism. As unpaved roads generally contain locally selected and stabilized soils, they can be a major source of airborne pollutants if they are covered with toxic materials, for example, in localities near ore extraction sites, as mentioned in the introduction.

Roadside sampling methodologies

A compendium of roadside resuspension emission techniques can be found in [Casotti and Alves \(2021\)](#). The main findings are presented below. There is no clear agreement on which methodology is more appropriate to perform a road dust sampling campaign. Choices vary from basic sweeping tools to sophisticated devices, as will be shown.

Manual sweeping and vacuum cleaning

In most cases, a procedure to collect road dust accumulated on the surface is done with a dustpan and a brush. Also, vacuum cleaners are used, although the possibility of an incomplete collection exists if no wet brushing is deployed afterwards. These techniques imply a degree of uncertainty since they are not operator independent.

Care must be taken when the samples arrive to the laboratory. For example, the removal of large debris and other undesired objects must be sieved. Then, desired size fractioning is performed in 3–6 size intervals as, for example: <38 µm, <50 µm, <63 µm, <100 µm and even <250 µm. Also, after the sieving process, the finest fraction can be resuspended in a mixing chamber to simulate real world conditions.

Direct sampling

Different research teams hone their procedures according to the local conditions. For example, in Scandinavian countries it is necessary to identify and attribute sources of particulate matter which operate in late winter and spring, such as winter tire equipment and antiskid road treatments ([Gustafsson et al., 2019](#)). This is when accumulated residues from tire and road wear attain higher values and the resuspension process make them reach the respirable fraction of particulate matter. The direct techniques described in [Casotti and Alves \(2021\)](#) are:

Wet dust sampler (WDS)

This device of the Swedish National Road and Transport Research Institute (VTI) is equipped with a spray nozzle that thoroughly cleans the surface ([Gustafsson et al., 2019](#)). It employs high-pressure water to clean small circular road surface areas, and compressed air to move the sample from the washing unit to a sample bottle. The WDS has been used in several Swedish national research projects, mainly for investigating the influence of road operation measures on dust emissions. Models of the WDS can estimate the contribution to PM₁₀ from road dust and other non-exhaust particles. The content of the bottles can then be used to obtain size distribution by laser granulometry.

Testing re-entrained aerosol kinetic emissions from roads (TRAKER) and EMMA

The TRAKER consists of laser photometer instruments mounted on a vehicle, measuring PM concentrations in real time. One set of photometers is installed behind the front tires of the vehicle and measures the resuspended particles. Another set is deployed usually ahead of the front bumper, to obtain background concentrations. Road dust concentrations of PM₁₀ and PM_{2.5} are calculated as the difference between the values obtained close to the tire and the respective background ([Kuhns et al., 2003](#)) via:

$$T = T_T - T_B$$

where T is the signal with background correction, expressed in mg m⁻³, T_T is the PM concentration behind the tire, and T_B is the background concentration in front of the bumper.

A modified version of TRAKER, called EMMA, was proposed by [Hussein et al. \(2008\)](#). This device was used to carry out an extensive sampling campaign over 300 km of the Stockholm area in Sweden, from late spring until autumn 2006.

SNIFFER

Similar to the previous methodology, the mobile laboratory SNIFFER was developed and tested on 20 km of road in Helsinki, Finland, to measure traffic exhaust emissions under real driving conditions from a moving vehicle. It was further adapted for non-exhaust particles ([Pirjola et al., 2004](#)). It consists of two different inlets, one located above the windshield at 2.4 m above the ground, and the other above the bumper, located behind the left rear tire, at 0.7 m from the road surface; see [Fig. 3](#).

EMMA and SNIFFER differ in the sampling inlet location (front tires versus rear tires) and inlet air flow, and the instrument deployed for particle measurements: the optical properties of a DustTrak (a light-scattering laser photometer to measure aerosol mass) versus the gravimetric principles of the TEOM (tapered element oscillating microbalance). Measurements show, in both cases, a positive correlation between concentrations and vehicle speed, but PM₁₀ levels measured by SNIFFER are higher than those collected by EMMA.

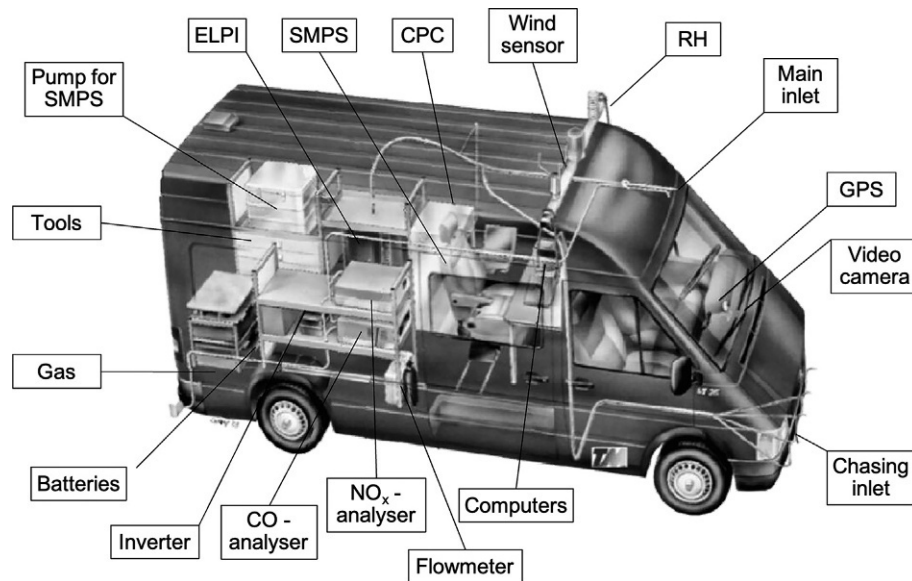


Fig. 3 SNIFFER mobile laboratory van. Reproduced with permission from Pirjola L, Parviainen H, Hussein T, Valli A, Hämeri K, Aalto P, Virtanen A, Keskinen J, Pakkanen TA, Mäkelä T and Hillamo RE (2004). "Sniffer"—A novel tool for chasing vehicles and measuring traffic pollutants. *Atmospheric Environment* **38**(22): 3625–3635. doi: 10.1016/j.atmosenv.2004.03.047.

System for the continuous aerosol measurement of particulate emissions from roads (SCAMPER)

TRAKER and SNIFFER focus on tire-road interactions, and they might be disregarding the resuspension of particles provoked by the vehicular aerodynamic wake turbulence. Therefore, [Fitz and Bufalino \(2002\)](#) proposed SCAMPER, a method with the aim of incorporating resuspension mechanisms and directly estimating emission factors. As in the other mobile laboratories, SCAMPER relies on a vehicle on which special equipment is installed and a trailer. PM₁₀ concentrations are measured through three laser-optical instruments: one, on the vehicle roof (background), and two on the trailer. To sample in the trailer only the resuspended particles of the wake, the exhaust gasses are by-passed, and contamination is avoided. Experiments included measurements in a motorway around Aachen (Germany) at a speed up to 100 km h⁻¹.

Online resuspension chamber

[Jancsek-Turóczy et al. \(2013\)](#) designed a mobile sampler for resuspended dust, consisting of a stainless-steel hood placed on a mobile platform to collect road dust. For this purpose, the hood is connected to a leaf blower which simulates the effects of turbulence generated by traffic or wind on road surfaces. The resuspended particles are collected in a cyclone separator that measures the PM₁–PM₁₀ resuspended dust fractions, while PM₁ samples are collected on a filter. The design of the collection method was not intended to be quantitative but can provide the proportion of the PM₁ and PM_{1–10} fractions in resuspended dust.

Different authors have presented variations to the online resuspension chamber to increase accuracy, reduce material loss and to be able to adjust the sampling area. This type of device is mainly used to compare dust loadings over different pavement materials in the dry season, and in different geo-morphological and wind conditions.

Portable in situ wind erosion laboratory

A direct sampling method with the added advantage of determining the potential for wind erosion and dust emission from soil surfaces is the Portable In-Situ Wind Erosion Lab (PI-SWERL; [Etyemezian et al., 2007](#)). This device uses an annular ring that rotates directly above the soil test surface. Dust and sand are mobilized by the shear created by the rotating ring. Dust concentrations within the chamber that encloses the annular ring are measured by light scattering. The PI-SWERL provides a robust index of wind erosion or dust emission potential. Compared to traditional field wind tunnels used for the same purpose, the PI-SWERL offers significant economy in size, portability, and ease of use.

Vertical profile

Most methodologies presented so far involve the collection of road dust directly from the surface, either in dry or wet conditions. To obtain vertical profiles of PM concentrations, canisters are placed on a roadside location. As shown in [Fig. 4](#), the system consists of 6-cm high, and 7-cm wide cylindrical canisters placed vertically. The theoretical assumptions are essentially that particles descend towards the ground with a rate influenced by their deposition velocity and particles rise upwards induced by turbulent eddies ([Amato et al., 2012](#)).

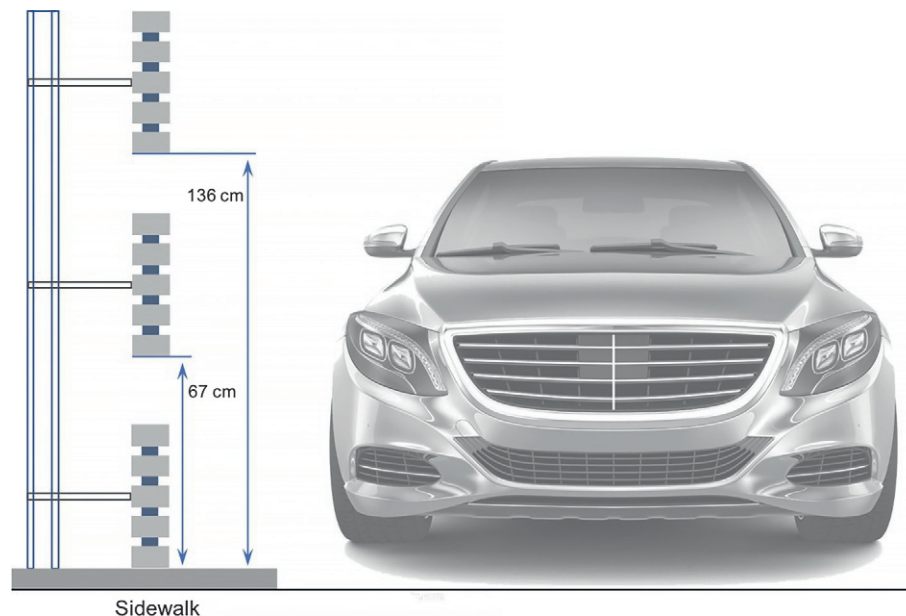


Fig. 4 Vertical profile method. Schematic of sampler array displayed on the roadside. Based on Casotti RI and Alves CA (2021) Road dust resuspension: A review. *Atmospheric Research* **261**: 105740. doi: 10.1016/j.atmosres.2021.105740.

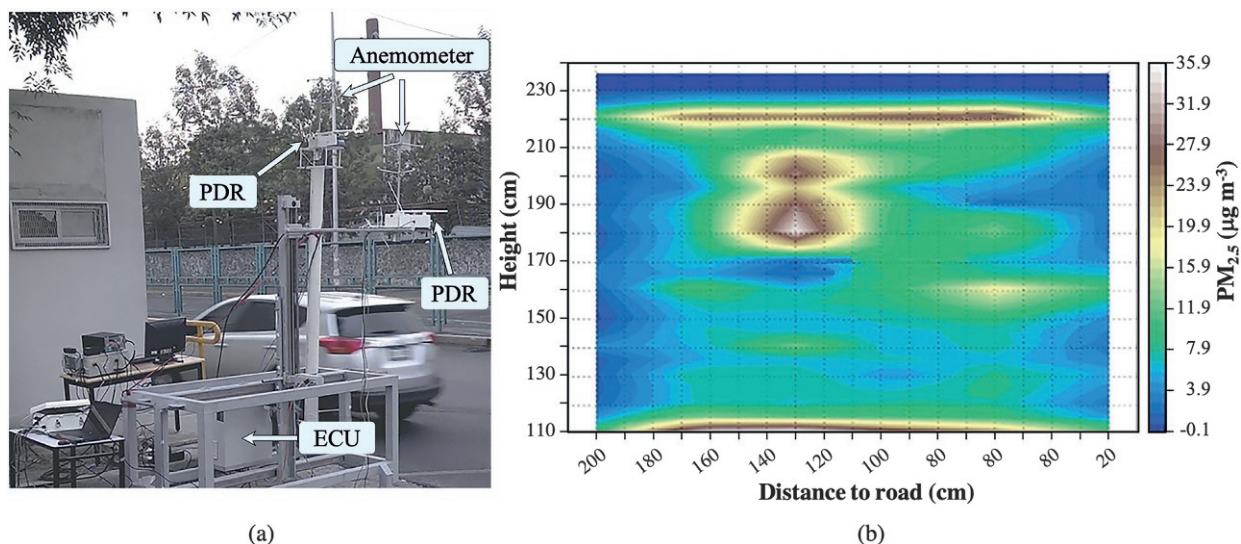


Fig. 5 In (a), electro mechanically controlled device instrumented with $PM_{2.5}$ laser-optical sensors (PDR1500) deployed on a roadside. In (b), corresponding integrated $PM_{2.5}$ concentrations on a 2×2 m orthogonal plane.

Active concentration profilers

The methodologies presented by Casotti and Alves (2021) were not developed with the aim of estimating exposure. As stated in the introduction, exposure requires pollutant concentration and human presence at the same time. Recently, devices have been proposed to obtain near-road concentration fields where pedestrians interact with the pollution. In this way, potential exposure levels are possible to obtain. An electromechanical device continuously samples and localizes roadside atmospheric concentrations on a $2 \text{ m} \times 2 \text{ m}$ plane orthogonal to the road (Jazcilevich et al., 2019). The technique discriminates background from immediate emissions but does not distinguish between resuspension and tailpipe emissions. Nevertheless, if the device were applied in a road-dust setting, where PM resuspension emissions are orders of magnitude greater than tailpipe emissions, it could yield resuspension exposure information. In the limited number of experiments using this device on paved roads, emitted $PM_{2.5}$ concentrations were located at preferred heights on the roadside. These heights are dictated in part by vehicular aerodynamics, and some coincide with children's height. Results also showed that exposure decreases rapidly with increasing distance from the road. The device and an example result are shown in Fig. 5.

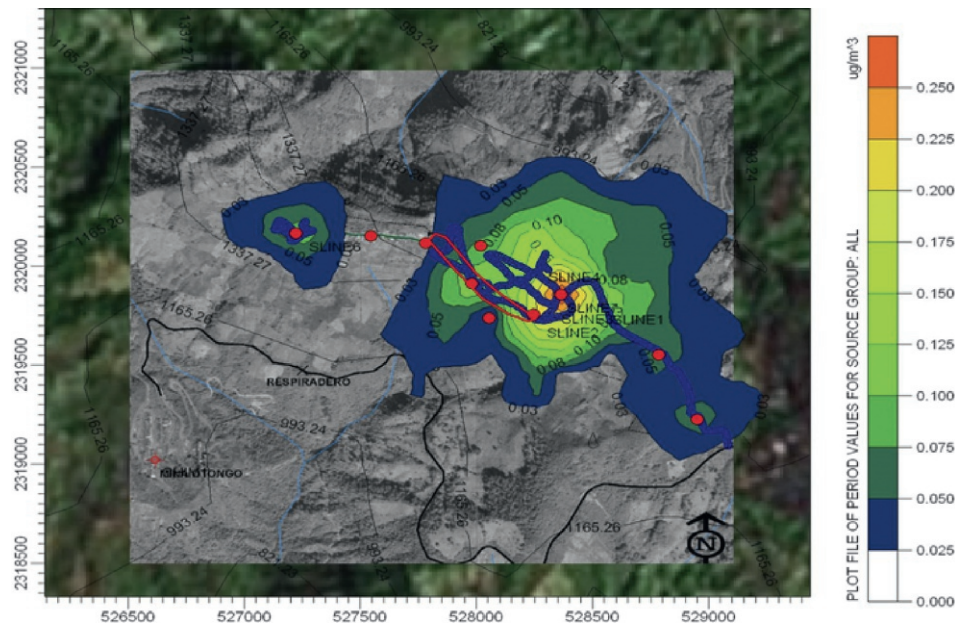


Fig. 6 Mn concentrations emanating from wind erosion of dust roads in a mining community in Molango, Hidalgo, Mexico, for a typical day in January. Based on Riojas H, Rodríguez S, Hernández D, Cortez M, Shilmann A, Santibañez R, Sabido E, Rodríguez Y, Solís R, Montes S, Ríos C, Mondragón A, Rosas I, Siebe C, Shimada K, Jazcilevich A, Wellens A, Paz F and Catalán M (2009) *Exposición a manganese de la población residente en un distrito minero (México) – fase II. Informe Final. Donación 103052-001. Proyecto conjunto INSP, INNN, PUMA, IG, CCA.*

Emission factors for road wind erosion

Wind erosion PM emissions factors of unpaved roads can be obtained by wind tunnel studies, or the PI-SWRL device. Results vary since emission factors depend heavily on physical soil characteristics. Jia et al. (2013) mention 8 m/s as a threshold wind speed for soil erosion to take place, while Díaz-Nigenda et al. (2018) observed wind erosion emissions with wind speeds over 6 and 9 m/s respectively in two cities in southern Mexico. However, other wind tunnel experiments show that for a windspeed of 3 ms⁻¹ a wind erosion emission factor of 0.0276 mg m⁻² s⁻¹ can be determined over an unpaved road (Riojas et al., 2009).

The emission factors as a function of wind speed obtained in wind tunnel experiments can be used in Gaussian air dispersion models, to simulate a point or line emission. An example of the output of such modeling technique is shown in Fig. 6 for a Mn mining community in central Mexico (Riojas et al., 2009). This study obtained PM_{2.5} and Mn concentrations using AERMOD, considering the unpaved roads as a sequence of sources where emissions were proportional to the road area. AERMOD is a steady-state Gaussian plume dispersion model aimed at short-range (<50 km) air pollution dispersion from point, line, area and volume sources (Cimorelli et al., 2003). Since gaussian dispersion models are based on the solution of the advection-diffusion equation, mass conservation is assured. Therefore, they give a better estimate of how dust concentrations decrease with distance than when using average, linear or exponential attenuation factors.

In Fig. 6, the area most affected by Mn due to wind erosion from dirt roads is shown. The PM_{2.5} emission factors for unpaved roads were obtained in a wind tunnel as a function of surface wind speed from road surface samples. Mn emission factors were adjusted from PM_{2.5} emission factors by determining the Mn content in PM_{2.5} by means of chemical analysis of the soil samples. Modeling results indicated that the effect of wind erosion is local: Mn and PM_{2.5} concentrations increase in areas where several unpaved roads converge. Simulated average Mn concentrations were two to three times above the reference concentration for inhalation exposure, corresponding to 0.05 µg/m³ (US-EPA, 1993). The extension of the area where the reference concentration is exceeded under typical meteorological conditions for the studied community was found to be around 1.2 × 1.3 km².

Wind erosion control can be achieved by two means: reducing wind speed at ground level and increasing soil cohesion, thus improving soil resistance to wind (Duniway et al., 2019). To reduce surface wind speed, green fences can be planted as windbreaks between homes and unpaved roads. To increase soil cohesion, it is highly recommended to pave the roads in rural areas where minerals or other polluting compounds are exploited, especially when these roads are frequently used and cross communities.

From dust to dose

Since exposure relates the pollutant concentrations that come in contact with the human body, we consider it an appropriate name for the material discussed below.

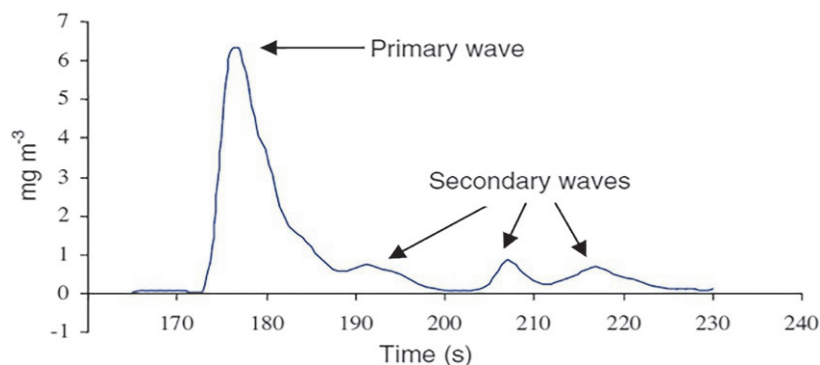


Fig. 7 Typical resuspension $PM_{2.5}$ wave form showing a primary followed by secondary waves. From Jazcilevich A, Wellens A, Siebe C, Rosas I, Bornstein RD and Riojas-Rodríguez H (2012) Application of a stochastic vehicular wake erosion model to determine $PM_{2.5}$ exposure. *Aeolian Research* 4: 31–37. doi: 10.1016/j.aeolia.2011.11.002.

Soil erosion, caused by either traffic resuspension or wind, results in accelerated transport of hazardous materials. This is of particular concern because of potential increased inhalation exposure over a short time. Nevertheless, studies regarding the ensuing mechanisms responsible for this phenomenon are lacking.

The study of exposure to roadside pollution in rural unpaved roads provides a chance to observe isolated events with large concentrations, thus acting as a research magnifying glass providing information about the fluid mechanics in which exposure takes place. For example, in Jazcilevich et al. (2012), it was possible to observe typical shapes of the resuspension of a $PM_{2.5}$ concentration signal generated by a truck traveling on an unpaved road, as shown in Fig. 7. Note that the exposure scale shown is in the range of $mg\ m^{-3}$, several orders of magnitude above the prescribed 1, 8 or 24 h averaged PM health norms, which are given in $\mu g\ m^{-3}$.

To obtain potential inhaled mass, the constraint in short time scales requires appropriate analytical methods for those scales. In the case of Jazcilevich et al. (2012), no time averaging techniques were used since they smear maxima exposure levels. This can reduce maxima by an order of magnitude. Accumulative acute effects can thus be severely underestimated. Also, by using a stochastic model, the turbulent character of the phenomena was preserved. It allowed the estimation of real-life inhaled mass by a pedestrian.

The stochastic model is shown in Fig. 8. The lognormal probability distribution (pdf) was obtained by observing the repeated resuspension signal of a truck passing in front of the instruments under calm wind conditions. The lognormal pdf was tested using Kolmogorov-Smirnov and Anderson-Darling test statistical techniques (Stephens, 1974). To project this information to other near road distances and heights, a function, whose independent variables are height, z , and distance to the road, x , was built. It was obtained by collecting and measuring the weight of resuspended dust mass in vertically and horizontally placed sampling devices at varying distances. The scheme used for this task is reminiscent of the vertical scheme described in Fig. 4, but a set of recipients were added in the horizontal, orthogonal to the road. A function whose independent variables were height and distance to the road was thus obtained. By appropriately normalizing this function, an attenuation factor function $f(x, z)$ was applied to the pointwise $PM_{2.5}$ measurements to project them to the desired distance and height. The stochastic model was tested using a “witness” sensor. A version of the model for the dry-warm and dry-cold season was obtained to account for the corresponding sensible soil heat fluxes variations. Dry-warm conditions resulted in much higher exposure levels.

By repeating the stochastic experiment at desired heights and distances to the road, potential inhaled $PM_{2.5}$ with 15% Mn was obtained for different populations. Results for the dry-warm season are shown in Fig. 9. The number of daily trucks passing in front of an individual was 8, and distance to the road 50 cm. The inhalation rate was calculated as a function of activity, using data from Gratt (1996).

To obtain dosage, the inhaled mass shown in Fig. 9 must be divided by blood or target organ volume. This volume is a function of gender and age. It can be seen that in light to moderate activity, children receive a higher dosage of Mn than adults.

Final considerations

Exposure to road dust resuspension from unpaved roads demands information on composition and spatial concentration fields that come in contact with the human body. The scientific community is advancing technological schemes to acquire resuspension data mainly to obtain emissions factors, i.e., vehicular emitted resuspension mass per km. Measurement techniques to obtain spatial information on road dust, where pedestrian activity takes place, is not being pursued with the same energy. The vertical scheme described Fig. 4, in conjunction with the attenuation function $f(x, z)$ described in Fig. 8, can be implemented to obtain distance-to-the-road concentrations and obtain spatial pollutant fields. This scheme obtains potential acute exposure. Active profilers placed on the roadside as shown in Fig. 5, obtain spatial data to estimate pedestrian exposure.

Since roadside exposure intake is directly related to inhalation rate, analysis must be done in time scales of seconds, thus including information on turbulent phenomena. Otherwise, average smearing leads to severe underestimation of inhaled

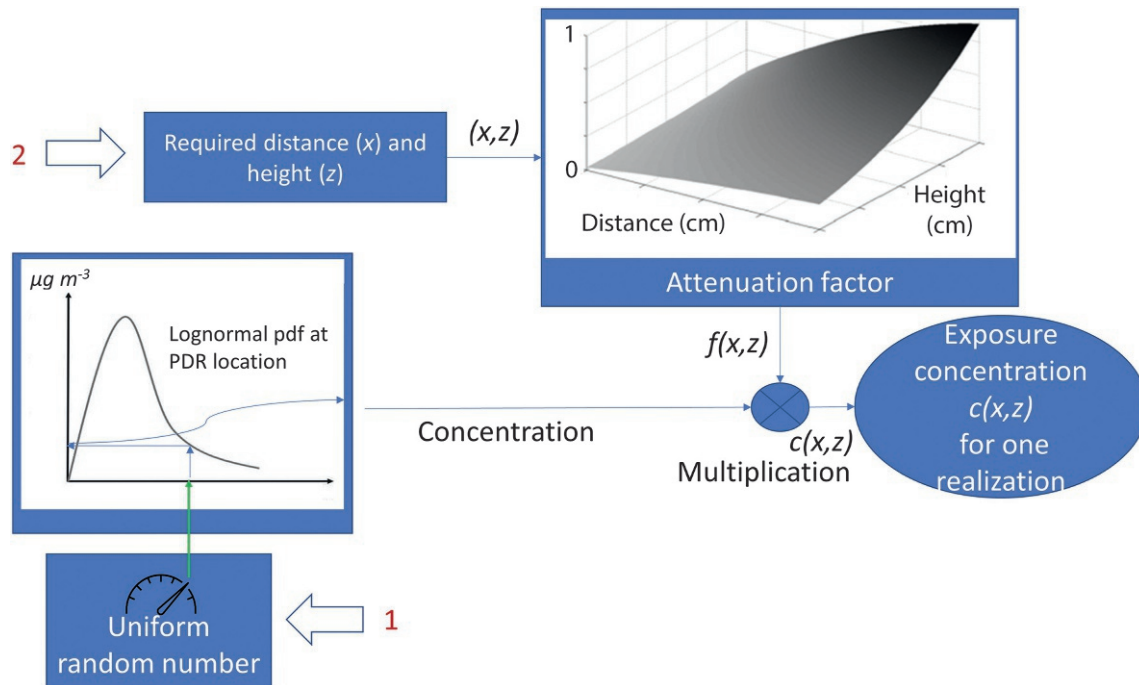


Fig. 8 Schematics of the stochastic exposure model. A realization is initiated by generating a uniform random number in the ordinate of the lognormal pdf, as indicated by arrow 1. This selects a concentration, which is multiplied by an attenuation factor $f(x, z)$ to appropriately localize the concentration at a selected height, z , and distance x to the road, as indicated by arrow 2. Based on Jazcilevich A, Wellens A, Siebe C, Rosas I, Bornstein RD and Riojas-Rodríguez H (2012) Application of a stochastic vehicular wake erosion model to determine PM_{2.5} exposure. *Aeolian Research* 4: 31–37. doi: 10.1016/j.aeolia.2011.11.002.

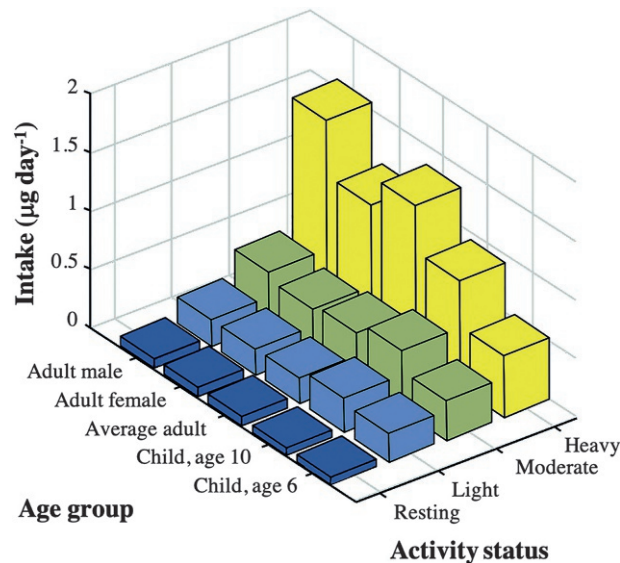


Fig. 9 Respiratory intake of Mn ($\mu\text{g day}^{-1}$) calculated by the stochastic model for eight truck passes at a distance of 50 cm to the road, as a function of age and activity status. Based on Jazcilevich A, Wellens A, Siebe C, Rosas I, Bornstein RD and Riojas-Rodríguez H (2012) Application of a stochastic vehicular wake erosion model to determine PM_{2.5} exposure. *Aeolian Research* 4: 31–37. doi: 10.1016/j.aeolia.2011.11.002.

pollutants. When working at time scales of seconds, features such as increased inhaled dosage in children and specific population, plus relevant spatial and turbulent flow characteristics containing pollutants are obtained. For isolated exposure events, a stochastic modeling scheme was presented that preserved the random character of the phenomena.

Using wind tunnel or in situ experiments with the PI-SWERL device, road soil emissions factors can be obtained as a function of wind speed. Line sources in Gaussian transport models can thus be implemented. The larger the road area, the larger the dust emission. This is clearly shown in road intersections, where emission areas are additive. Gaussian models identify affected areas, where abatement measures can be allocated.

Regarding road-dust effects in human health, further research on the interaction of pollutants and its potentiation could be an important research area. Computational Fluid Dynamics (CFD) to obtain pollution fields at time and space scales in the order of seconds and centimeters is an interesting area of further development in human exposure. Also, abatement measures such as vegetation barriers and optimal routes could be designed and tested using CFD techniques.

References

- Amato F, Karanasiou A, Moreno T, Alastuey A, Orza JAG, Lumbrales J, Borge R, Boldo E, Linares C, and Querol X (2012) Emission factors from road dust resuspension in a Mediterranean freeway. *Atmospheric Environment* 61: 580–587. <https://doi.org/10.1016/j.atmosenv.2012.07.065>.
- Avecilla F, Panebianco JE, and Buschiazio DE (2015) Variable effects of saltation and soil properties on wind erosion of different textured soils. *Aeolian Research* 18: 145–153. <https://doi.org/10.1016/j.aeolia.2015.07.005>.
- Casotti RI and Alves CA (2021) Road dust resuspension: A review. *Atmospheric Research* 261: 105740. <https://doi.org/10.1016/j.atmosres.2021.105740>.
- Cimorelli AJ, Perry SG, Venkatram A, Weil JC, Paine RJ, Wilson RB, Lee RF, Peters WD and Paumier JO (2003) *AERMOD: Description of Model Formulation*. U.S. Environmental Protection Agency Report, EPA 454/R-03-002d, North Carolina.
- Colombo C, Monhemius AJ, and Plant JA (2008) Platinum, palladium and rhodium release from vehicle exhaust catalysts and road dust exposed to simulated lung fluids. *Ecotoxicology and Environmental Safety* 71: 722–730. <https://doi.org/10.1016/j.ecoenv.2007.11.011>.
- Cortez-Lugo M, Riojas-Rodríguez H, Moreno-Macías H, Montes S, Rodríguez-Agudelo Y, Hernández-Bonilla D, Catalán-Vázquez M, Díaz-Godoy R, and Rodríguez-Dozal S (2018) Evaluation of the effect of an environmental management program on exposure to manganese in a mining zone in Mexico. *NeuroToxicology* 64: 142–151. <https://doi.org/10.1016/j.neuro.2017.08.014>.
- Del Río-Salas R, Ayala-Ramírez Y, Loredó-Portales R, Romero F, Molina-Freaner F, Minjarez-Osorio C, Pi-Puig T, Ochoa Landín L, and Moreno-Rodríguez V (2019) Mineralogy and geochemistry of rural road dust and nearby mine tailings: A case of ignored pollution hazard from an abandoned mining site in semi-arid zone. *Natural Resources Research* 28: 1485–1503. <https://doi.org/10.1007/s11053-019-09472-x>.
- Díaz-Nigenda E, Tatarko J, Morales-Iglesias H, Hernández MZ, and Vázquez MW (2018) Measurement and modeling air quality impacts of dust emissions from unpaved roads in Tuxtla Gutiérrez, Chiapas. *Geosciences* 8(8): 284. <https://doi.org/10.3390/geosciences8080284>.
- Duniway MC, Pfennigwerth AA, Fick SE, Nauman TW, Belnap J, and Barger NN (2019) Wind erosion and dust from US drylands: A review of causes, consequences, and solutions in a changing world. *Ecosphere* 10(3): e02650. <https://doi.org/10.1002/ecs2.2650>.
- Etyemezian V, Nikolich G, Ahonen S, Pitchford M, Sweeney M, Purcell R, Gillies J, and Kuhns H (2007) The Portable In Situ Wind Erosion Laboratory (PI-SWREL): A new method to measure PM10 Windblown dust properties and potential for emissions. *Atmospheric Environment* 41(18): 3789–3796. <https://doi.org/10.1016/j.atmosenv.2007.01.018>.
- FAO and UNEP (2021) *Global Assessment of Soil Pollution*. Report. Rome, Italy. ISBN 978-92-5-134469-9. <https://www.fao.org/3/cb4894en/online/src/html/copyright.html>.
- Fitz DR and Bufalino C (2002) Measurement of PM10 emission factors from paved roads using on-board particle sensors. In: *USEPA 11th International Emission Inventory Conference*, Atlanta, GA, April 15–18. <https://www3.epa.gov/ttnchie1/conference/ei11/dust/fitz.pdf>.
- Fitz DR and Bumiller K (2021) SCAMPER monitoring platform to measure PM10 emission rates from unpaved roads in real-time. *Atmosphere* 12: 1301. <https://www.mdpi.com/2073-4433/12/10/1301>.
- Gratt L (1996) *Air Toxic Assessment and Management*. New York, NY: Van Nostrand Reinhold, pp. 388.
- Gustafsson M, Blomqvist G, Järskog I, Lundberg J, Janhall S, Elmgren M, Johansson C, Norman M, and Silvergren S (2019) Road dust load dynamics and influencing factors for six winter seasons in Stockholm, Sweden. *Atmospheric Environment* 2: 100014. <https://doi.org/10.1016/j.aeoa.2019.100014>.
- Hussein T, Johansson C, Karlsson H, and Hansson HC (2008) Factors affecting non-tailpipe aerosol particle emissions from paved roads: On-road measurements in Stockholm, Sweden. *Atmospheric Environment* 42(4): 688–702. <https://doi.org/10.1016/j.atmosenv.2007.09.064>.
- Jancsek-Turóczi B, Hoffer A, Nyíró-Kósa I, and Gelencsér A (2013) Sampling and characterization of resuspended and respirable road dust. *Journal of Aerosol Science* 65: 69–76. <https://doi.org/10.1016/j.jaerosci.2013.07.006>.
- Jazcilevich A, Wellens A, Siebe C, Rosas I, Bornstein RD, and Riojas-Rodríguez H (2012) Application of a stochastic vehicular wake erosion model to determine PM_{2.5} exposure. *Aeolian Research* 4: 31–37. <https://doi.org/10.1016/j.aeolia.2011.11.002>.
- Jazcilevich A, de la Cruz Zavala J, Erazo Arcos AM, Kanda I, and Rosas I (2019) Sidewalk pollution flows caused by vehicular traffic place children at a higher acute exposure risk. *Journal of Exposure Science & Environmental Epidemiology* 29. <https://doi.org/10.1038/s41370-018-0083-4>.
- Jia Q, Huang Y, Al-Ansari N, and Knutsson S (2013) Dust emission from unpaved roads in Luleå, Sweden. *Journal of Earth Sciences and Geotechnical Engineering* 1792-9660 3(1): 1–13.
- Jiang Y, Hu X, Yves UJ, Zhan H, and Wu Y (2014) Status, source and health risk assessment of polycyclic aromatic hydrocarbons in street dust of an industrial city, NW China. *Ecotoxicology and Environmental Safety* 106: 11–18. <https://doi.org/10.1016/j.ecoenv.2014.04.031>.
- Khan RK and Strand MA (2018) Road dust and its effect on human health: A literature review. *Epidemiol Health* 40: e2018013. <https://doi.org/10.4178/epih.e2018013>.
- Kuhns H, Etyemezian V, Green M, Hendrickson K, McGown M, Barton K, Pitchford M (2003) Vehicle-based road dust emissions measurements. Part II: Effect of precipitation, wintertime road sanding, and street sweepers on inferred PM₁₀ emission potentials from paved and unpaved roads. *Atmospheric Environment* 37(32): 4573–4582. [https://doi.org/10.1016/S1352-2310\(03\)00529-6](https://doi.org/10.1016/S1352-2310(03)00529-6).
- Labban R, Veranth JM, Chow JC, Engelbrecht JLP, and Watson JG (2004) Size and geographical variation in PM₁, PM_{2.5} and PM₁₀: Source profiles from soils in the Western United States. *Water, Air, & Soil Pollution* 157: 13–31. <https://doi.org/10.1023/B:WATE.0000038855.96607.c6>.
- Mar TF, Larson TV, Stier RA, Claiborn C, and Koenig JQ (2008) An analysis of the association between respiratory symptoms in subjects with asthma and daily air pollution in Spokane, Washington. *Inhalation Toxicology* 16: 809–815. <https://doi.org/10.1080/08958370490506646>.
- Pirjola L, Parviainen H, Hussein T, Valli A, Hämeri K, Aalto P, Virtanen A, Keskinen J, Pakkanen TA, Mäkelä T, and Hillamo RE (2004) “Sniffer”—A novel tool for chasing vehicles and measuring traffic pollutants. *Atmospheric Environment* 38(22): 3625–3635. <https://doi.org/10.1016/j.atmosenv.2004.03.047>.
- Potgieter-Vermaak S, Rotondo G, Novakovic V, Rollins S, and Van Grieken R (2012) Component-specific toxic concerns of the inhalable fraction of urban road dust. *Environmental Geochemistry and Health* 34: 689–696. <https://doi.org/10.1007/s10653-012-9488-5>.
- Ramírez HNB, Aparicio VC, and Méndez MJ (2021) First evidence of glyphosate and aminomethylphosphonic acid (AMPA) in the respirable dust (PM10) emitted from unpaved rural roads of Argentina. *Science of the Total Environment* 773: 145055. <https://doi.org/10.1016/j.scitotenv.2021.145055>.
- Riojas H, Rodríguez S, Hernández D, Cortez M, Shilmann A, Santibañez R, Sabido E, Rodríguez Y, Solís R, Montes S, Ríos C, Mondragón A, Rosas I, Siebe C, Shimada K, Jazcilevich A, Wellens A, Paz F and Catalán M (2009) *Exposición a manganeso de la población residente en un distrito minero (México) – fase II*. Informe Final. Donación 103052-001. Proyecto conjunto INSP, INNN, PUMA, IG, CCA.
- Rodríguez Martín JA, Gutiérrez C, Escuer M, García-González MT, Campos-Herrera R, and Aguila N (2014) Effect of mine tailing on the spatial variability of soil nematodes from lead pollution in La Unión (Spain). *The Science of the Total Environment* 473–474: 518–529. <https://doi.org/10.1016/j.scitotenv.2013.12.075>.
- Sahoo PK, Md. Equeenuddin SK, and Powell MA (2016) Trace elements in soils around coal mines: Current scenario, impact and available techniques for management. *Current Pollution Reports* 2(1): 1–14. <https://doi.org/10.1007/s40726-016-0025-5>.
- Stephens MA (1974) EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association* 69: 730–737.

- Tchounwou PB, Patlolla AK, and Centeno JA (2003) Carcinogenic and systemic health effects associated with arsenic exposure—A critical review. *Toxicologic Pathology* 31: 575–588. <https://doi.org/10.1080/01926230390242007>.
- US-EPA (1993) *Integrated Risk Information System (IRIS) —Manganese*. U.S. Environmental Protection Agency. https://iris.epa.gov/ChemicalLanding/&substance_nmbr=373.
- US-EPA (2021) *National Air Toxics Assessment: Glossary of Terms*. US Environmental Protection Agency. Consulted on December 15, 2021. <https://www.epa.gov/national-air-toxics-assessment/nata-glossary-terms>.
- Wei CC, Chen WC, and Wang CS (1996) Development of source profiles of agricultural waste combustion and road dust in the Meliao area in Taiwan. *Chemical Engineering Communications* 151(1): 41–52. <https://doi.org/10.1080/00986449608936540>.
- Williams DS, Shukla MK, and Ross J (2008) 400. *Atmospheric Environment* 42(16): 3899–3905. <https://doi.org/10.1016/j.atmosenv.2008.02>.
- Xu S, Zheng N, Liu J, Wang Y, and Chang S (2013) Geochemistry and health risk assessment of arsenic exposure to street dust in the zinc smelting district, Northeast China. *Environmental Geochemistry and Health* 35: 89–99. <https://doi.org/10.1007/s10653-012-9463-1>.
- Zajusz-Zubek E, Mainka A, and Kaczmarek K (2018) Determination of water-soluble elements in PM_{2.5}, PM₁₀, and PM_{2.5-10} collected in the surroundings of power plants. *E3S Web of Conferences* 28: 01042. <https://doi.org/10.1051/e3sconf/20182801042>.
- Zereini F, Alsenz H, Wiseman CL, Püttmann W, Reimer E, Schleyer R, Bieber E, and Wallash M (2021) Platinum group elements (Pt, Pd, Rh) in airborne particulate matter in rural vs. urban areas of Germany: Concentrations and spatial patterns of distribution. *Science of the Total Environment* 416: 261–268. <https://doi.org/10.1016/j.scitotenv.2011.11.070>.

Further reading

- Amato F (2018) In: Amato F (ed.) *Non-Exhaust Emissions: An Urban Air Quality Problem for Public Health—Impact and Mitigation Measures*, 1st edn Academic Press. ISBN: 978-0128117705.
- Amato F, Karanasiou A, Moreno T, Alastuey A, Orza JAG, Lumbreras J, Borge R, Boldo E, Linares C, and Querol X (2012) Emission factors from road dust resuspension in a Mediterranean freeway. *Atmospheric Environment* 61: 580–587. <https://doi.org/10.1016/j.atmosenv.2012.07.065>.
- Armstrong B, Hutchinson E, Unwin J, and Fletcher T (2004) Lung cancer risk after exposure to polycyclic aromatic hydrocarbons: A review and meta-analysis. *Environmental Health Perspectives* 112: 970–978. <https://doi.org/10.1289/ehp.6895>.
- Avecilla F, Panebianco JE, and Buschiazio DE (2015) Variable effects of saltation and soil properties on wind erosion of different textured soils. *Aeolian Research* 18: 145–153. <https://doi.org/10.1016/j.aeolia.2015.07.005>.
- Fashola MO, Ngole-Jeme VM, and Babalola OO (2016) Heavy metal pollution from gold mines: Environmental effects and bacterial strategies for resistance. *International Journal of Environmental Research and Public Health* 13(11): 1047. <https://doi.org/10.3390/ijerph13111047>.
- Goossens D and Buck B (2014) Dynamics of dust clouds produced by off-road vehicle driving. *Journal of Earth Sciences and Geotechnical Engineering* 4(2): 1–21. http://www.sciencedirect.com/Upload/GEO/Vol%204_2_1.pdf.
- Holguin F, Flores S, Ross Z, Cortez M, Molina M, Molina L, Rincon C, Jerrett M, Berhane K, Granados A, and Romieu I (2007) Traffic-related exposures, airway function, inflammation, and respiratory symptoms in children. *American Journal of Respiratory and Critical Care Medicine* 176(12): 1236–1242. <https://doi.org/10.1164/rccm.200611-1616OC>.
- Karanasiou A, Amato F, Moreno T, Lumbreras J, Borge R, Linares C, Boldo E, Alastuey A, and Querol X (2014) Road dust emission sources and assessment of street washing effect. *Aerosol and Air Quality Research* 14: 734–743. <https://doi.org/10.4209/aaqr.2013.03.0074>.
- Lanzerstorfer C (2021) Toward more intercomparable road dust studies. *Critical Reviews in Environmental Science and Technology* 51(8): 826–855. <https://doi.org/10.1080/10643389.2020.1737472>.
- Understanding soil risks and hazards. Muckel GB (ed.) (2004) *Using Soil Survey to Identify Areas With Risks and Hazards to Human Life and Property*. Lincoln, Nebraska: United States Department of Agriculture, Natural Resources Conservation Service, National Soil Survey Center.
- Ngezahayo E, Ghataora GS, and Burrow MPN (2019) Factors affecting erosion in unpaved roads. In: *Proceedings of the 4th World Congress on Civil, Structural, and Environmental Engineering (CSEE'19) Rome, Italy*. <https://doi.org/10.11159/icgre19.108>.
- Parvej S, Naik DL, Sajid HU, Kiran R, Huang Y, and Thanki N (2021) Fugitive dust suppression in unpaved roads: State of the art research review. *Sustainability* 13(4): 2399. <https://doi.org/10.3390/su13042399>.
- Perez L, Lurmann F, Wilson J, Pastor M, Brandt SJ, Künzli N, and McConnell R (2012) Near-roadway pollution and childhood asthma: Implications for developing “win-win” compact urban development and clean vehicle strategies. *Environmental Health Perspectives* 120: 1619–1626. <https://doi.org/10.1289/ehp.1104785>.
- Querol X, Karanasiou A, Amato F, Vasconcelos C, Alastuey A, Viana M, Moreno T, Plana F, Pérez N, Cabañas M, Bartoli R, Martínez S and Sosa M et al. (2016) *Technical guide to reduce road dust emissions in Southern Europe*. AIRUSE LIFE 11/ENV/ES/584. http://airuse.eu/wp-content/uploads/2013/11/R28_AIRUSE-TechGuide-road-dust-emission-reduction.pdf.
- Rodríguez EN, McLaughlin MJ, and Pennock D (2018) *Soil Pollution: A Hidden Reality*. Rome, Italy: Food and Agriculture Organization of the United Nations. ISBN: 978-92-5-130505-8156.
- Sexton K and Ryan PB (1988) Assessment of human exposure to air pollution: Methods, measurements, and models. In: Watson AY, Bates RR, and Kennedy D (eds.) *Air Pollution, the Automobile, and Public Health*. Washington: National Academies Press. ISBN: 0-309-56826-9 704 pages. https://www.ncbi.nlm.nih.gov/books/NBK218150/pdf/Bookshelf_NBK218150.pdf.
- US-EPA (2006) *Unpaved Roads. AP-42: Compilation of Air Emissions Factors, Chapter 13, Section 2.2*. U.S. Environmental Protection Agency. <https://www3.epa.gov/ttnchie1/ap42/ch13/final/c13s0202.pdf>.
- Yulevitch G, Danon M, Krasovtsov B, Fominykh A, Svet N, Tsesarsky M, and Katra I (2020) Evaluation of wind-induced dust-PM emission from unpaved roads varying in silt content by experimental results. *Atmospheric Pollution Research* 11(2): 261–268. <https://doi.org/10.1016/j.apr.2019.10.010>.

Soil salination processes and management

Uri Nachshon and Guy J Levy, Institute for Soil, Water and Environmental Sciences, ARO, Volcani Research Center, Rishon-Lezion, Israel

© 2023 Elsevier Ltd. All rights reserved.

Global extent and economic damage to agriculture	236
Origin of salinity in soils	237
Salination of cultivated soils	237
Definition	237
Classification of water quality by salinity	237
Salt dissolution, transport, and precipitation	239
Effect of salination on soil	239
Effects on the soil ecosystem	239
Effects on soil properties	240
Impact of salination on plants	240
Crop yield and soil salinity	240
Salinization management	242
Monitoring salt affected soils	242
Coping with soil salinization	243
Reclamation of salt affected soils	243
Summary	244
References	244

Abstract

Soil salination is a major worldwide concern, which affects mainly arable lands, due to irrigation with low quality water and or improper drainage practices that may result in rising of the groundwater level. Many studies have focused on the major negative impacts that soil salination has on soil productivity and associated harmful effects on the agricultural and natural environments. Here, basic soil salination processes and terms are presented, and their impact on soil and crop production is discussed. The chapter also presents key insights regarding common practices used to cope with soil salination and the reclamation of saline soils.

Key points

- Soil salination is a major process of soil degradation.
- Soil salination results from natural (primary) and anthropogenic (secondary) salination processes.
- Major causes of anthropogenic soil salination in arable lands are irrigation with saline water and improper soil drainage.
- Crop development and yield are harmed by soil salination and various practices can be employed to overcome this challenge, including soil reclamation.

Global extent and economic damage to agriculture

Soil is one of the essential resources on the globe, affecting the supply of food, fiber, and clean fresh water, maintaining biodiversity and overall protection of the ecosystem. Soil also plays a central role in the carbon cycle by acting as a major sink. However, globally soil is being continuously depleted due to various degradation processes, with soil salination being a major cause (Ezlit et al., 2010).

Excessive levels of salts occur in large areas of soil around the world and profoundly affect land use. Six percent of terrestrial land is naturally salinized in a process termed primary soil salination, in which salt stored in the soil or groundwater is mobilized to the land surface in the development process of a landscape (Rengasamy, 2006). However, much greater and faster soil salination processes are observed in cultivated lands where 20% of total cropland and 33% of irrigated croplands worldwide are salinized as a result of agricultural practices (Singh, 2015). This type of soil salination, a result of anthropogenic activity, is termed secondary soil salination.

Salinized areas are increasing steadily at a rate of 10% per year and it is estimated that, worldwide, more than 50% of croplands will be salinized by 2050 (Jamil et al., 2011). Naturally, the economic burden of soil salination is considerable, with the global cost of irrigation-induced salinity being in the order of US\$27 billion per year (Qadir et al., 2014).

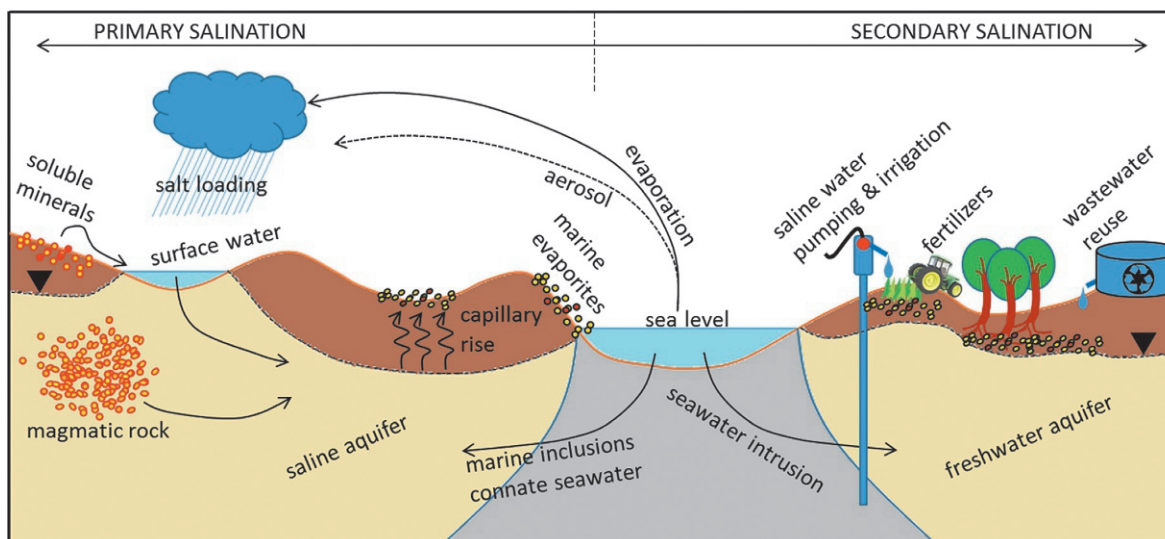


Fig. 1 Major processes associated with primary and secondary salination (adapted from Maniçoba da Rosa Ferraz Jardim et al., 2020). Light and dark brown colors indicate water saturated and un-saturated terrestrial areas, respectively.

Origin of salinity in soils

As aforementioned, there are two types of controlling factors for soil salination: primary and secondary (Fig. 1). Primary salination results from natural processes including: (i) physical or chemical weathering and transport from parent material, geological deposits or groundwater. This is often observed in closed basins in arid areas, such as the Canadian prairies, where soil leaching is minor and evaporation is high; (ii) geological events, such as volcanic activity or specific formations with high solubility, such as halite (NaCl) that can increase salt concentration in groundwater; (iii) salt water intrusion and tidal fluctuations leading to submergence of low-lying lands by sea water, as observed in many coastal areas worldwide, including the south-east of the United States; (iv) hydrological processes in the presence of shallow and saline groundwater that cause water flow upward by capillarity from the water table to replenish evapotranspiration, subsequently leading to solute accumulation close to the soil surface; (v) deposition of salts by rainfall, mainly along coastal areas, which may account for 10–25% of the total yearly contribution from weathering and is often overlooked.

Unlike primary salination, secondary salination is introduced by human interventions; mainly irrigation with saline water or ill-suited irrigation practices in soils with poor drainage conditions (Ezlit et al., 2010). Modern agriculture in dry areas consumes about 500–1000 mm of water per year while under extreme cases may even consume up to 2000 mm. Globally, and in particular in dry areas, high quality water is in short supply and the use of irrigation water rich in solutes is becoming ever more common. Moreover, overuse of fertilizers through irrigation systems (fertigation) may also increase solute concentration in the irrigation water as farmers frequently add fertilizers in quantities greater than plant demand (Nachshon, 2018). Consequently, the risk of soil salination due to unsustainable irrigation practices with water rich in salt is rising (Rengasamy, 2006).

Salination of cultivated soils

Definition

Salt is an ionic compound formed by the reaction of an acid and a base. Once salt is formed, it may go through cycles of dissolution, transport and crystallization. The salt content of soils above which plant growth is affected depends upon several factors, among which are soil texture, composition of the salt, and plant species. The electrical conductivity (EC) of the saturation extract has been adopted by the US Salinity Laboratory Staff (1954) as the preferred scale for estimating soil salinity, with saline soils being defined as soils with EC exceeding 4 dS m^{-1} .

Classification of water quality by salinity

The criterion for soil salinity ($\text{EC} > 4 \text{ dS m}^{-1}$), as proposed by the US Salinity Laboratory Staff (1954), was determined in saturated water extracts of soils because under this practice water content of the saturated extract can be closely associated with the naturally-occurring state of the soil solution, and can thus be related to plant-growth response. However, the practice of soil saturated water extract is relatively time consuming, and requires high levels of skill, precision, and expertise to reach saturation point, especially in clay soils. Therefore, over recent years it has become more common to determine soil salinity by various soil to

Table 1 Interpretation of 1:1 soluble salts test, for various soil textures (Moebius-Clune et al., 2017).

Degree of salinity	Crop residue	EC (dS m ⁻¹) by soil texture			
		Coarse sand to loamy sand	Loamy fine sand to loam	Silt loam to clay loam	Silty clay loam to clay
Non-saline	Almost negligible effect	0–1.1	0–1.2	0–1.3	0–1.4
Slightly-saline	Yield of the most sensitive crops reduced	1.2–2.4	1.3–2.4	1.4–2.5	1.5–2.8
Moderately saline	Yield of most crops reduced	2.5–4.4	2.5–4.7	2.6–5.0	2.9–5.7
Strongly saline	Only tolerant crops yield well	4.5–9.0	4.8–9.5	5.1–10.1	5.8–11.5
Very strongly saline	Only very tolerant crops yield well	>9.0	>9.5	>10.1	>11.5

water ratio (soil:water) extracts (Kargas et al., 2018). In this practice, the EC of the supernatant of a given soil:water extract (1:1, 1:2, 1:5) is measured by a calibrated electric conductivity meter. Table 1 presents interpretation of EC readings of a 1:1 soil:water extract for different soil textures and related degrees of soil salinity.

The EC readings, correlated to salt concentration expressed in terms of total dissolved solids (TDS, mg L⁻¹), and ionic concentration (mmol_c L⁻¹), for various salt solutions are presented in Fig. 2. Based on these correlations, it is commonly accepted for a mixture of salts to refer to the relationship between TDS and EC for various salt solutions as follows (Rhoades, 1996):

$$\text{TDS} = 640 \times \text{EC} \quad (1)$$

where EC is in units of dS m⁻¹.

As can be seen from the data in Fig. 2, the EC of soil is not only determined by the concentration of ions but also by the characteristics of the ions or salinity chemistry (US Salinity Laboratory Staff, 1954). Consequently, the ionic composition of the soil solution should also be taken into account for proper interpretation of salinity stress and soil-plant interactions (Rengasamy, 2010). In general, in the Commonwealth of Independent States (CIS, formerly the USSR) and many other countries, classification of soil salinity is based on total soluble salts (TSS, %; g salt 100 g⁻¹ dry soil), and salinity chemistry which is determined by the concentration ratios of anions (HCO₃⁻, Cl⁻, SO₄²⁻) and additionally by cations (Na⁺, Ca²⁺, Mg²⁺) in 1:5 soil–water extraction (meq L⁻¹). This type of classification suggests that together with the TSS, the threshold level or toxicity of salts (highly soluble salts in toxic quantities to plants) is associated with the relation between anionic and cationic ions, i.e., type and degree of salinity. Recently, Ismayilov et al. (2021) stated that once the salinity type of the soil is established, EC (1:5) values can be safely used for evaluation of the degree of soil salinity in irrigated land in the context of sustainable soil and crop management.

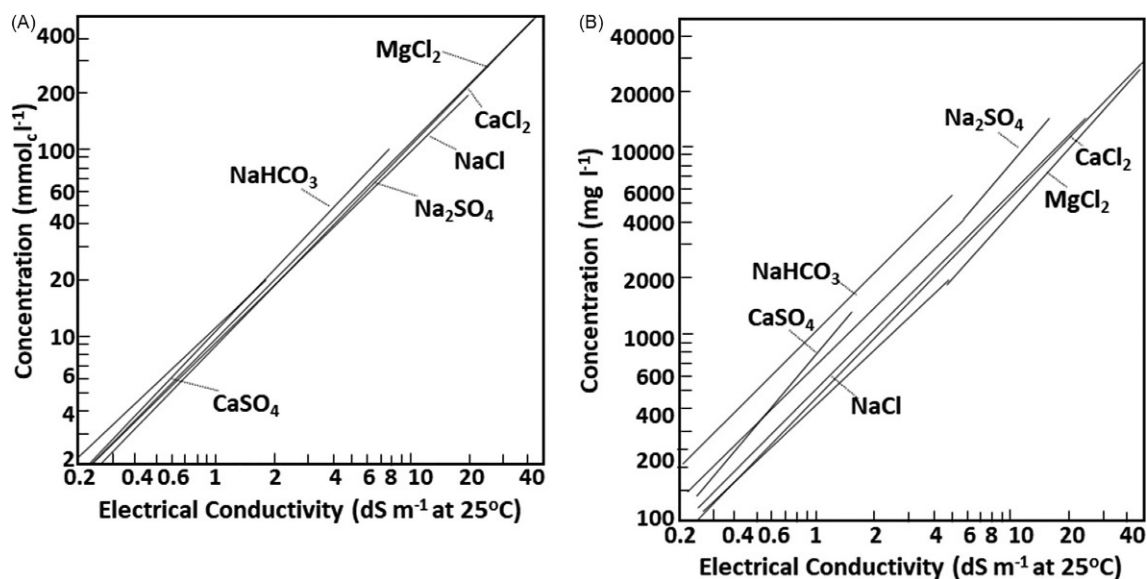


Fig. 2 Relationship between salt concentration and electrical conductivity of various salt solutions: salt concentration in (A) millimoles of charge per liter and (B) milligrams per liter. Adapted from the US Salinity Laboratory Staff (1954) Diagnosis and improvement of saline and alkali soils. *Agriculture Handbook*, vol. 60, pp. 83–100, US Salinity Laboratory Staff.

Salt dissolution, transport, and precipitation

Many salts are relatively soluble in water, as the electrostatic attraction forces between water dipoles and the anions and cations forming the salt exceeds the attraction forces between the positive and negative ions that compose the salt (Hillel, 2000). Consequently, pore water and surface water may dissolve crystalline salts that are found in the soil, and transport the solutes in the water by diffusion and advection (Hillel, 2000; Nachshon, 2018).

Salt accumulation and precipitation in cultivated soils is usually a result of evaporation and transpiration processes. These processes take the pore water out of the soil, while leaving behind the solutes that are dissolved within. Consequently, pore water solute concentration increases until saturation is reached and salt crystals begin to precipitate. Salt precipitation usually occurs as narrow layers at the upper boundary of the evaporation front or near the roots, where water uptake by the roots takes place (Nachshon, 2018). When the evaporation front is at the soil-atmosphere interface, salt accumulates at the soil surface either as an efflorescent (above soil surface) or as a subflorescent crust (below soil surface), depending mainly on salt type. A conceptual scheme of root zone soil salination due to upward capillary flow of water from a shallow-saline aquifer and transport of solutes upwards, towards the soil surface and the root system, where evaporation and transportation are taking place is presented in Fig. 3. These processes result in the accumulation and precipitation of salt in the soil. At locations where the capillary flow reaches the soil surface, the salt precipitates as an efflorescent salt crust.

Effect of salination on soil

Effects on the soil ecosystem

The recent review of Daliakopoulos et al. (2016), and references cited therein, highlights the role of soil salination in determining ecological soil functions. Salination may negatively affect soil respiration, residue decomposition, nitrification, and denitrification through a decrease in soil biodiversity and microorganism activity that often lead to loss of soil organic matter. Loss of the latter enhances aggregate destabilization and soil sensitivity to wind and water erosion. Moreover, extreme salination could result in eliminating all vegetation, leaving the soil bare and thus more susceptible for further physical degradation. Yet, naturally saline soils may comprise unique gene reservoirs that deserve protection.

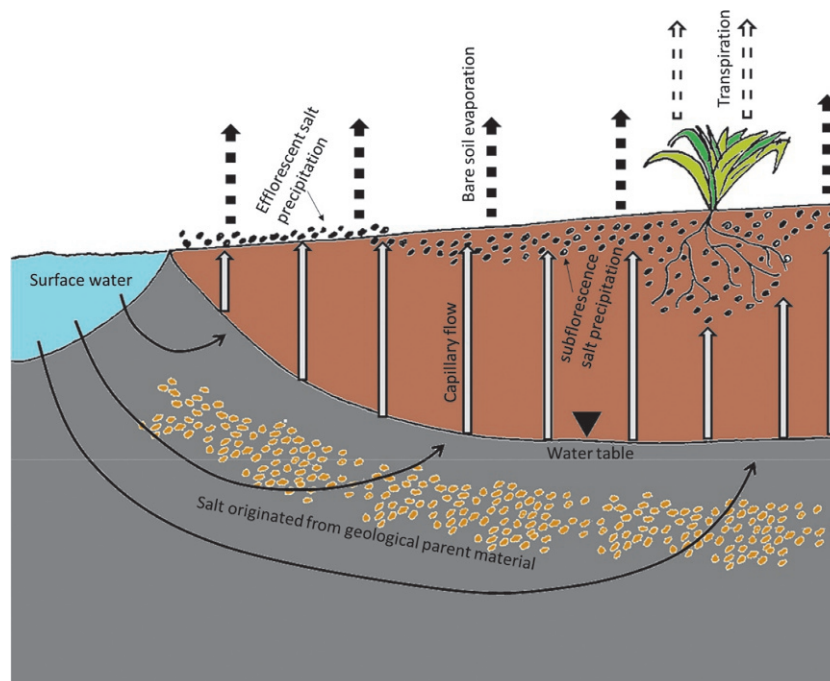


Fig. 3 Conceptual scheme of soil salination due to upward transport of solutes towards the soil surface with capillary flow. At the soil surface and the root system, roots and evaporation are consuming water and salts accumulate and precipitate in the upper parts of the soil profile, as an efflorescent salt crust where capillary flow reaches the soil surface, and as subflorescence where the water table is too deep and the capillary flow does not reach the soil surface.

Effects on soil properties

Salt precipitation, especially subflorescence precipitation, greatly affects soil mechanical properties as salt crystallization may lead to soil cementation that holds sand and silt particles together and increases the shear strength and stiffness of the soil during the drying process. On the other hand, in cohesive (rich in clay and/or organic matter) soils, subflorescence salt crystallization may enhance weathering processes and reduce soil aggregate stability as the crystallized salt at the pores applies pressure against the pore walls that leads to disaggregation (Nachshon, 2016). Furthermore, subflorescence salt precipitation in the soil pores changes the hydrological properties of the soil and reduces percolation and evaporation processes as it reduces matrix porosity, permeability and vapor diffusivity. Salt precipitation as efflorescent salt crust has less effect on physical properties of the soil-matrix itself, as crystallization occurs above the soil surface. Nevertheless, efflorescent salt precipitation strongly affects evaporation because it alters conditions at the boundary layer between the soil and the atmosphere.

In salt affected soils that are also sodic, i.e., soils in which the sodium adsorption ratio (SAR) is greater than 5, the soil is considered a sodic soil and the conditions may adversely affect soil structure and stability, and subsequently the growth of most crop plants. However, the interrelationship between the SAR of an equilibrium solution and EC of the soil solution is central in dictating whether sodic or non-sodic behavior (i.e., swelling and dispersion of soil clay) will be observed (Levy and Nachshon, in press).

Impact of salination on plants

The negative effects of salinity on plant growth are related to: (i) lowering of soil solution osmotic potential; (ii) specific ions toxicity, and (iii) nutritional imbalance.

Reduction of the osmotic potential between soil pore water and plant root cells, because of an increase of soil pore water solute concentration, leads to restrained water uptake by plants and reduces their ability to survive and produce, especially in arid areas (Rengasamy, 2006; Nachshon, 2018). The osmotic pressure, P_o [kPa], of a solution is a colligative property and is related to the total salt concentration rather than the concentration of individual ionic species. Therefore, P_o can be estimated from the solution electrical conductivity expressed in units of dS m^{-1} (Rhoades, 1996):

$$P_o = 0.36 \times \text{EC} \quad (2)$$

Certain solutes which accumulate in plants over time can negatively affect various biological functions of the plant. For example, high concentrations of boron, chloride, sodium and carbonate are toxic for many plants and result in the drying of leaves, damage to fruit, and can affect root growth (Munns and Tester, 2008; Nachshon, 2018). Excess salinity may also cause nutrient deficiencies or imbalances due to the competition between sodium and/or chloride and nutrients such as calcium, potassium and nitrate, which are essential for plant development (Hu and Schmidhalter, 2005).

Crop yield and soil salinity

For agricultural purposes, salt tolerance of crops is described by various response functions that correlate soil salinity to yield reduction. The first correlations that were developed described a linear reduction of yield as salinity increases (Magistad et al., 1943; Nachshon, 2018). Later on, it has been recognized that crop yield is not reduced until a threshold salinity level is exceeded, beyond which the yield decreases linearly with rising salinity. Further research has presented more complicated functions that correlate yield with soil salinity by exponential ratios (Steppuhn et al., 1998) and a combination of two and three linear segments that describe different stages of crop response to salinity (Shani and Dudley, 2001), as demonstrated in Fig. 4.

Today, the function of van Genuchten and Hoffman (1984) is still widely used to describe percentile reduction of yield (f_{EC}) as a function of the soil saturated paste electrical conductivity (EC_e):

$$f_{\text{EC}} = \frac{1}{1 + \left(\frac{\text{EC}_e}{\text{EC}_{50}}\right)^P} \quad (3)$$

where EC_{50} is the electrical conductivity of a soil saturated paste that leads to a 50% reduction in yield, P is a plant parameter, that in many cases is equal to 3 (Shani et al., 2007). Typical values of EC_{50} , for various crops are described in the literature, and selected samples are presented in Table 2 (Shani et al., 2007; Nachshon, 2018).

Plants can be divided into four groups (sensitive to tolerant) according to their response to salinity. Of the field crops, *Hordeum vulgare* L. (barley), *Gossypium herbaceum* L. (cotton), and *Beta vulgaris* L. subsp. *vulgaris* (sugar beet) belong to the tolerant group, *Triticum aestivum* L. (wheat) and *Glycine max* (L.) Merrill (soya bean) belong to the moderately tolerant group, *Arachis hypogaea* L. (peanut) and *Oryza sativa* L. (rice) belong to the moderately sensitive group, and *Phaseolus vulgaris* L. (beans) and *Vigna unguiculata* L. (cowpeas) belong to the salt-sensitive group.

Moreover, plant response to salinity depends also on soil, climate, and plant factors. Soil-water content and frequency of irrigation affect the tolerance of crops to soil salinity. Shortening the irrigation interval minimizes the deleterious effects of salinity.

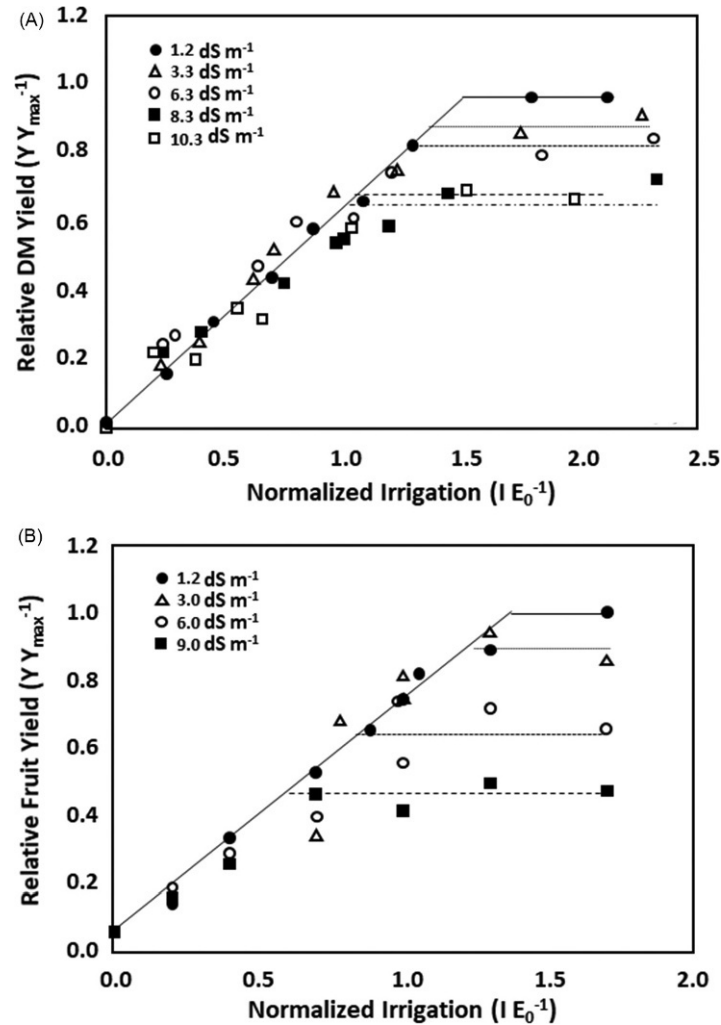


Fig. 4 (A) Relative dry matter (DM), and (B) yield ($Y Y_{max}^{-1}$), for corn (A) and melon (B) as affected by normalized irrigation (the ratio of irrigation [I] to potential evaporation [E_0]) and salinity treatments. Reproduced with permission from Shani U and Dudley LM (2001) Field studies of crop response to water and salt stress. *Soil Science Society of America Journal* **65**: 1522–1528.

Table 2 The EC_{50} values of selected crops.

Crop	EC_{50} [dS m ⁻¹]
Sunflower	10
Melon	7.5
Tomato	5.5
Corn	5
Pepper	2.5
Grape	2.5
Onion	3

The salt tolerance of many crops is enhanced when daily drip irrigation is used. Climate can also modify plant response to salinity with tolerance to salt often being reduced under hot, dry conditions.

Plant factors such as stage of growth, variety, and rootstock also affect plant response to salinity. Rice, barley, wheat, and corn are most sensitive to salinity during the early seedling stage. Sugar beet and sunflower (*Helianthus annuus* L.) are more sensitive to salinity during germination than in later stages of growth. Salinity effects on the vegetative growth of many plants (e.g., cotton, wheat) are greater than on seed or fiber production. In general, salinity suppresses top growth more than root growth. The type of rootstock of woody plants may also affect the specific tolerance of fruit crops to salinity.

Salinization management

Monitoring salt affected soils

Managing of salt affected soils requires knowledge of their location. Hence, monitoring the development of and mapping of soil salinity are central for effectively handling such soils. Mapping of salt affected soils at a global scale by remote observations (satellites, aerial photograph) may illustrate the scope of the soil salinization problem. Yet, it should be born in mind that, typically, remote observations of soil salinity can identify salt features on the soil surface in the visible spectrum, or through multispectral and hyperspectral remote sensing indices that depict only surface soil properties or vegetation health that can serve as a proxy for salinity.

However, mapping at a regional, farm or field scale is necessary for the design of management practices needed for alleviating the salinity problem. Salinity monitoring at farm or field scale have traditionally been carried out using local salinity sensors and soil sampling. In recent years, use of non-invasive geophysical techniques such as electromagnetic induction (EMI) (Hendrickx et al., 1992), vertical electrical sounding (VES) (Masoud et al., 2019) and electrical resistivity tomography (ERT) (Rasul et al., 2018) are gradually being adopted. However, for optimal use of data collected by these means, especially for mapping subsurface soil salinity, the measurements should be combined with actual ground data.

Finally, multi-scale integrated assessments that imply a combination of remote sensing, field data and various modeling approaches can improve the development of soil salination risk maps, useful to land managers and users. Fig. 5 presents an example of a soil salinity and sodicity risk map produced by Daliakopoulos et al. (2016).

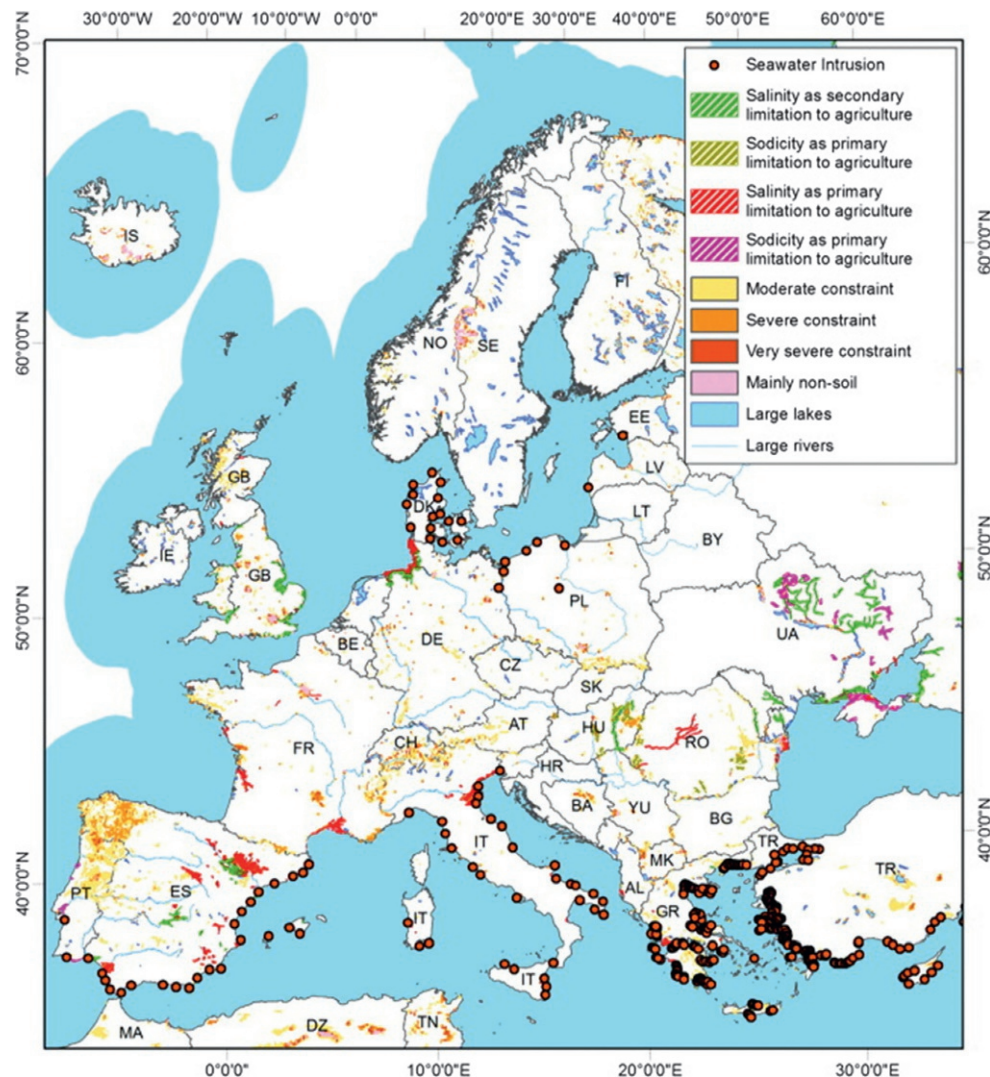


Fig. 5 Distribution of saline and sodic soils in terms of agricultural constraint and as primary and secondary limitations to agricultural use, and areas of seawater intrusion in Europe. From Daliakopoulos IN, Tsanis IK, Koutroulis A, Kourgialas NN and Varouchakis AE (2016) The threat of soil salinity: A European scale review. *Science of the Total Environment* 573: 727–739.

Coping with soil salinization

Like most other environmental concerns, the best way to cope with soil salination is to prevent it by using good quality water for irrigation, avoid land-use changes and agricultural practices that reduce infiltration and soil leaching processes. However, good quality water is in short supply in many areas, especially dry areas, where the need for irrigation is high, precipitation is low and is exceeded by evapotranspiration (Nachshon, 2018). Therefore, it is essential to improve agricultural and irrigation practices in order to maintain flourishing and sustainable agriculture in saline environments.

To avoid salt accumulation in the root zone during repeated cycles of irrigation and evapotranspiration, the common remedy is to apply water in volumes greater than evapotranspiration to deliberately cause a fraction of the applied water to flow downwards through the root zone and flush away the excess salts.

The required volume of excess water to leach the salt downwards is termed the 'leaching requirement' (LR). The LR is a function of irrigation EC (EC_i) and required EC of the saturation extract water from the root zone soil (EC_t) to ensure a yield reduction less than 10%:

$$LR = \frac{EC_i}{EC_t - EC_i} \quad (4)$$

The EC_t values for different crops can be found in the literature (Shani et al., 2007).

Although the practice of leaching is effective with respect to salt leaching from the root zone, it has three negative aspects: (i) it reduces water-use efficiency because more water is needed for irrigation; (ii) in the case of irrigation with saline water, the elevated volume of water used for irrigation results in an increase of net mass of salt that is added to the soil environment; and (iii) under conditions of poor drainage, it may cause groundwater levels to rise and water logging, which further increases the risk of soil salination and reduction of yield.

To avoid the rising of groundwater levels, artificial drainage systems of the soil profile may be implemented. The artificial drainage of groundwater is generally carried out by means of drains that could consist of ditches, pipes, or 'mole channels' into which groundwater flows as a result of the hydraulic gradients existing in the soil. The drains themselves are made to direct the water, by gravity or by pumping, to a drainage outlet, which may be a stream, a lake, an evaporation pond or the sea. In some places, drainage water may be recycled or reused for agricultural, industrial or residential purposes, as well as for agroforestry or the irrigation of ornamental plants. Because drainage water may contain potentially harmful concentrations of salts, fertilizer nutrients, pesticide residues, and various other toxic chemicals, as well as biological pathogens, it is not enough to provide a means for 'getting rid' of it; attention must be paid to the quality of the water to be disposed of and to the long-term consequences of its disposal.

Reclamation of salt affected soils

In ideal soils under piston flow conditions, the EC of soil water and the applied water would be equal after the passage of one pore volume of rainfall or irrigation water. However, water flow in soils is not ideal for two reasons: (i) water flows faster through soil cracks and large pores between soil aggregates than through smaller pores within aggregates. This difference in localized water flow velocities is commonly referred to as 'bypass flow,' or 'preferential flow'; (ii) salt diffusion and chemical reactions are not sufficiently rapid to allow salt concentration to be the same for all pores of differing sizes, particularly during periods of rapid water movement.

Salt transport models have provided considerable insight into these processes. However, suitable mathematical models are not available to describe the multiple concurrent processes of water flow, including spatially variable hydraulic properties and all the chemical reactions—exchange, dissolution, precipitation, and CO_2 liquid-gas equilibria—involved in salt transport at the field scale. Consequently, reclamation guidelines are largely based on experimental relations obtained from field and lysimeter reclamation experiments.

The method of water application and soil type are the primary variables affecting the amount of water required to reclaim saline soils. For saline soils, 60% or more of the salts initially present is leached by a depth of water equal to the depth of soil (Fig. 6A) under continuously ponded conditions. Sandy loam soils with low saturated water contents have greater leaching efficiency than clay loam or clay soils (Fig. 6A). Soil pores in sandy soils are more uniform in diameter than in clay loam or clay soils. The finer-textured soils can have large cracks with large pores when dry, and fine pores within aggregates when wet. Such bypass channels reduce leaching efficiency. The water requirement for leaching can be reduced by the intermittent application of ponded water, particularly for fine-textured soils (Fig. 6B). Under intermittent ponding, unsaturated conditions prevail in the soil profile and the inefficient bypass flow in large cracks and pores is prevented. Similarly, with sprinkler irrigation, reclamation occurs under unsaturated conditions and the efficiency of leaching increases. Sodic clay soils tend to suffer from swelling and dispersion processes that reduce their hydraulic conductivity and affect their soil structure and stability. This can be moderated by the addition of gypsum (calcium sulfate, $CaSO_4$) or other substances that contribute divalent cations to the soil (see Levy and Nachshon, in press).

In the case of extreme soil salination conditions, where soil leaching and drainage practices are impractical and conventional agriculture is impossible, one may use halophyte species for agriculture. Halophyte crops include domesticated wild species that are resilient to salinity and produce edible yield and useful biomass. The development of halophytic species for agriculture has great potential as it will generate new croplands in areas that currently are not suitable for agriculture such as the coastal areas and very saline inland areas. Furthermore, halophyte plants (e.g., *Salicornia bigelovii* Torr) can be used to reclaim saline soil, as some plants may extract salt from the soil with the water consumed, and accumulate it in their tissues (e.g., *Suaeda maritima* (L.) Dumort and

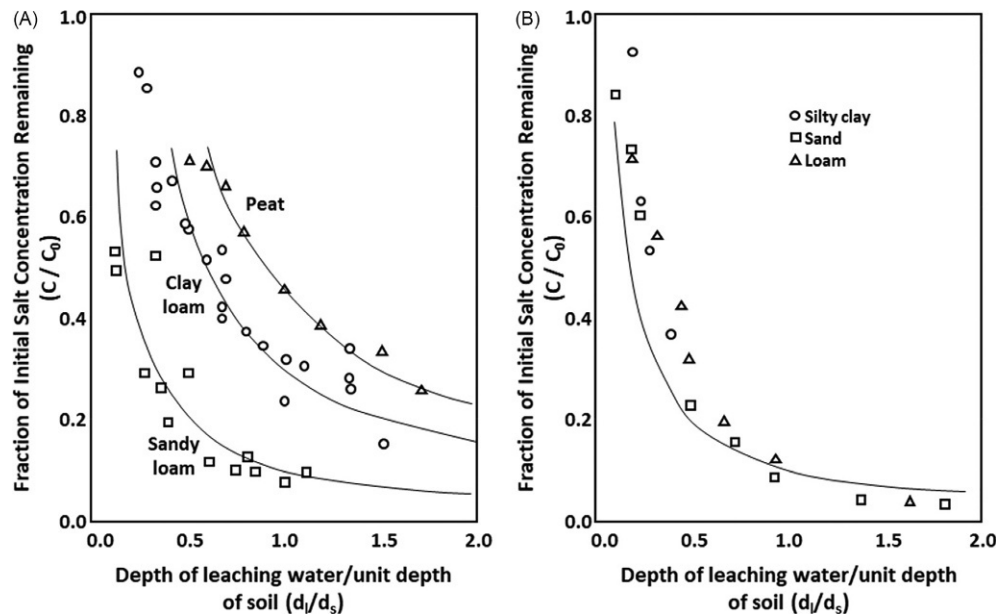


Fig. 6 Depth of leaching water per unit depth of soil required to reclaim a saline soil by continuous ponding (A) and depth of leaching water per unit depth of soil required to reclaim a saline soil by ponding water intermittently (B). Adapted from Oster JD, Shainberg I and Abrol IP (1996) Reclamation of salt affected soils. In Agassi M (ed.) *Soil Erosion Conservation and Rehabilitation*, pp. 315–342. New York: Marcel Dekker.

Sesuvium portulacastrum (L.) L). In addition, control of soil erosion in saline environments may be achieved by the use of halophyte plants that stabilize soil structure with their roots (e.g., *Atriplex lentiformis* (Torr) S. Wats). This type of soil reclamation, termed phytoremediation, has the potential to be a cost-effective and an environmentally sound technology for remediation of salt impacted soils (Hasanuzzaman et al., 2014; Nachshon, 2018).

Summary

Soil has always been a valuable resource for mankind, but the ever growing world population applies an significant increasing pressure on cultivated lands that has turned them into a limited resource because soil has a very long formation period (in the order of thousands of years). Therefore, if not handled in a sustainable manner, soils, particularly those suited for agricultural activities, might be in short supply in a few decades. Under arid and semiarid conditions, and in regions of poor natural drainage, there is a real hazard of salt accumulation in soils. The processes by which soluble salts accumulate in soils include irrigation with water containing salts, upward movement of moderate-to-saline water from high water tables, and weathering of primary and secondary minerals in the soils. Accumulation of soluble salts in the soil solution imposes stress on crop development and growth that can lead to decreased yields and, in severe cases, complete crop failure. Use of poor-quality water is increasing due to growing municipal demand for good-quality water and the need to dispose of municipal wastewater and agricultural drainage water. As a result, a better understanding of the reclamation of salt affected soils is becoming an increasingly important component of water and soil management to ensure the long-term sustainability of irrigated agriculture.

References

- Daliakopoulos IN, Tsanis IK, Koutroulis A, Kourgialas NN, and Varouchakis AE (2016) The threat of soil salinity: A European scale review. *Science of the Total Environment* 573: 727–739.
- Ezlit YD, Smith RJ and Raine SR (2010) *A Review of Salinity and Sodicty in Irrigation*. Cooperative Research Centre for Irrigation Futures. Irrigation Matters Series No. 01/10.
- Hasanuzzaman M, Nahar K, Alam M, Bhowmik PC, Hossain M, Rahman MM, and Fujita M (2014) Potential use of halophytes to remediate saline soils. *BioMed Research International* 2014. <https://doi.org/10.1155/2014/589341>.
- Hendrickx JMH, Baerends B, Raza ZI, Sadig M, and Chaudhry MA (1992) Soil salinity assessment by electromagnetic induction of irrigated land. *Soil Science Society of America Journal* 56(6): 1933–1941.
- Hillel D (2000) *Salinity Management for Sustainable Irrigation: Integrating Science, Environment, and Economics*. Washington, DC: World Bank Publications.
- Hu Y and Schmidhalter U (2005) Drought and salinity: A comparison of their effects on mineral nutrition of plants. *Journal of Plant Nutrition and Soil Science* 168: 541–549.
- Ismayilov AI, Mamedov AI, Fujimaki H, Tsunekawa A, and Levy GJ (2021) Soil salinity type effects on the relationship between the electrical conductivity and salt content for 1: 5 soil-to-water extract. *Sustainability* 13: 3395. <https://doi.org/10.3390/su13063395>.
- Jamil AS, Riaz MA, and Foolad MR (2011) Gene expression profiling of plants under salt stress. *Critical Reviews in Plant Sciences* 30: 435–458.
- Kargas G, Chatzigiakoumis I, Kollias A, Spiliotis D, Massas I, and Kerkides P (2018) Soil salinity assessment using saturated paste and mass soil:water 1:1 and 1:5 ratios extracts. *Watermark* 10–11: 1589.

- Levy GJ and Nachshon Y (in press) Sodic soils. In Goss MJ and Oliver MA (eds), *Encyclopedia of Soils in the Environment* 2nd edn, Amsterdam: Elsevier.
- Magistad OC, Ayers AD, Wadleigh CH, and Gauch HG (1943) Effect of salt concentration, kind of salt, and climate on plant growth in sand cultures. *Plant Physiology* 18: 151–166.
- Maniçoba da Rosa Ferraz Jardim A, Silva T, Souza L, and de Sá Souza M (2020) Interaction of agroecosystem intercropped with forage cactus-sorghum in the semi-arid environment: A review. *Journal of Environmental Analysis and Progress* 5: 69–87. <https://doi.org/10.24221/jeap.5.1.2020.2743.069-087>.
- Masoud AA, Koike K, Atwia MG, El-Horiny MM, and Gemail KS (2019) Mapping soil salinity using spectral mixture analysis of landsat 8 OLI images to identify factors influencing salinization in an arid region. *International Journal of Applied Earth Observation and Geoinformation* 83: 101944.
- Moebius-Clune BN, Moebius-Clune DJ, Gugino BK, Idowu OJ, Schindelbeck RR, Ristow AJ, van Es HM, Thies JE, Shayler HA, McBride MB, Kurtz KSM, Wolfe DW, and Abawi GS (2017) *Comprehensive Assessment of Soil Health. The Cornell Framework*, 3rd edn. Ithaca, New York: Cornell University.
- Munns R and Tester M (2008) Mechanisms of salinity tolerance. *Annual Review of Plant Biology* 59: 651–681.
- Nachshon U (2016) Seepage weathering impacts on erosivity of arid stream banks: A new conceptual model. *Geomorphology* 261: 212–221.
- Nachshon U (2018) Cropland soil salinization and associated hydrology: Trends, processes and examples. *Watermark* 10: 1030. <https://doi.org/10.3390/w10081030>.
- Qadir M, Quillérrou E, Nangia V, Murtaza G, Singh M, Thomas R, Drechsel P, and Noble A (2014) Economics of salt-induced land degradation and restoration. *Natural Resources Forum* 38(4): 282–295.
- Rasul H, Zou L, and Olofsson B (2018) Monitoring of moisture and salinity content in an operational road structure by electrical resistivity tomography. *Near Surface Geophysics* 16(4): 423–444.
- Rengasamy P (2006) World salinization with emphasis on Australia. *Journal of Experimental Botany* 57: 1017–1023.
- Rengasamy P (2010) Soil processes affecting crop production in salt affected soils. *Functional Plant Biology Journal* 37: 613–620.
- Rhoades JD (1996) Salinity: Electrical conductivity and total dissolved solids. In: Sparks D, Page A, Helmke P, Loeppert R, Soltanpour PN, Tabatabai MA, Johnston CT, and Sumner ME (eds.) *Methods of Soil Analysis*. Soil Science Society of America. <https://doi.org/10.2136/sssabookser5.3.c14>.
- Shani U and Dudley LM (2001) Field studies of crop response to water and salt stress. *Soil Science Society of America Journal* 65: 1522–1528.
- Shani U, Ben-Gal A, Tripler E, and Dudley LM (2007) Plant response to the soil environment: An analytical model integrating yield, water, soil type, and salinity. *Water Resources Research* 43: 8.
- Singh A (2015) Soil salinization and waterlogging: A threat to environment and agricultural sustainability. *Ecological Indicators* 57: 128–130.
- Steppuhn H, Wang H, and Gan Y (1998) Evaluating Russian wild ryegrass emergence from saline seedbeds using the Gompertz function. *Canadian Agricultural Engineering* 40: 241–247.
- US Salinity Laboratory Staff (1954) Diagnosis and improvement of saline and alkali soils. *Agriculture Handbook*. vol. 60, pp. 83–100. US Salinity Laboratory Staff.
- van Genuchten MT and Hoffman GJ (1984) Analysis of crop production. In: Shainberg I and Shalhevet J (eds.) *Soil Salinity Under Irrigation*, pp. 258–271. New York: Springer.

Degradation of soil structure and soil organic matter

Andrew S Gregory, Rothamsted Research, Harpenden, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction	247
Causes of soil structure and SOM degradation	247
Land use	247
Tillage	249
Compaction	251
Climate change	252
Effects of soil structure and SOM degradation on soil functions	253
Net primary productivity	253
Regulation of water and air	254
Biodiversity support	254
Climate change mitigation	254
Other soil functions and ecosystem services	255
Summary and outlook	255
References	255

Abstract

Soil structure and soil organic matter (SOM) have an interactive relationship and mediate all soil processes at all spatial and temporal scales. Soil structure and SOM can be degraded through land use, tillage, compaction, and climate change mechanisms. This impacts on the ability of soil to execute its soil functions and ecosystem services such as net primary productivity, regulation of water and air, biodiversity support, and climate change mitigation. Management interventions, particularly in agriculture, may slow or reverse degradation of soil structure and SOM, although there is much that future research can focus on to increase our knowledge.

Glossary

Alkyl A functional group of an organic chemical containing only carbon and hydrogen atoms arranged in a chain, considered to be more difficult for soil microorganisms to decompose.

Arbuscular mycorrhiza fungi Fungi which form a symbiotic association with plant roots for mutual benefit in exchanging nutrients and carbon.

O-alkyl A functional group mainly representing carbohydrates, including cellulose and hemicellulose, in fresh plant material, considered to be labile for soil microorganisms.

Plastic limit Occasionally referred to as the lower plastic limit, this is the water content above which the soil passes from a non-plastic to a plastic consistency state, indicating the 'workability' of the soil.

Suberin A complex plant macromolecule, common in roots, formed of long-chain fatty acids which form a protective barrier against water and solute movement.

Key points

- Soil structure and soil organic matter (SOM) have an interactive relationship, mediating all soil processes at all spatial and temporal scales, and a degradation of one will impact the other.
- Soil structure and SOM are degraded by a variety of mechanisms and processes, including land use, tillage, compaction, and climate change.
- Soil structure and SOM degradation affect the ability of soil to carry out its functions in support of wider ecosystem services, including net primary productivity, regulation of water and air, biodiversity support, and climate change mitigation.
- Management interventions can slow or reverse degradation of soil structure and SOM, such as modifying agricultural practices to retain or increase SOM, minimizing soil mechanical stresses to reduce soil structure degradation, and proactively remediating previous degradation.
- Future research should exploit recent advances in the non-destructive characterization of soil structure, the physical, biological, and chemical characterization of SOM, and modeling, to increase our understanding of the interactive relationship between soil structure and SOM.

Nomenclature

C	Carbon
C ₂ H ₄	Ethylene
Ca	Calcium
CH ₄	Methane
CO ₂	Carbon dioxide
K	Potassium
Mg	Magnesium
Mn	Manganese
N	Nitrogen
N ₂ O	Nitrous oxide
NH ₃	Ammonia
NO ₂	Nitrate
P	Phosphorus
S	Sulfur
SOM	Soil organic matter

Introduction

Soil structure and soil organic matter (SOM) have an interactive relationship. Soil structure is the spatial architectural arrangement of solid primary particles, either separately or more commonly in secondary agglomerated units (aggregates), with the associated non-solid pore network. Amongst other mechanisms, aggregates can be created by SOM components, such as the initiation of aggregation through clay mineral adsorption and alignment on particulate SOM. Aggregates can then be stabilized by other SOM components, notably microbial products or the living microorganisms themselves. In turn, SOM components can be sequestered and protected physically in stabilized aggregates or on mineral surfaces. Other SOM components will turn over in a functioning soil. In short, soil structure and SOM are interdependent.

In natural systems soil structure and SOM are controlled by soil forming factors. In classic pedology theory these factors are climate, organisms, relief, parent material and time. The same factors are also important in non-natural systems, but with the additional factor of management. All agricultural and forestry systems have a strong management factor which impacts on many soil properties and functions. In agricultural and forestry systems, it can be argued that soil structure and SOM are controlled more by management than any of the other pedological factors.

In many non-natural systems, management practices and interventions can often lead to a degradation of soil structure and SOM where the main objective is not based directly on enhancing these. In agricultural systems, for instance, soil structure and SOM can be degraded in the course of interventions imposed to maximize crop or livestock productivity. This is not always the case however as it is recognized increasingly that soil structure and SOM are key to achieving any main objective, and sometimes positive interventions are made. It is also the case that soil structure and SOM can degrade in natural systems too.

Structure and SOM are dynamic properties of the soil and are key to the ability of the soil to carry out its functions in support of wider ecosystem services. Soil structure and SOM mediate all soil processes at all spatial and temporal scales. Degradation of soil structure and SOM are therefore undesired in a healthy soil. Degradation of SOM may relate to either a measurable loss of the total quantity of SOM, or a decline in the quality of SOM particularly with respect to soil functioning, or both. The greater part of SOM is comprised of carbon (C), and hence this can have implications for the global C cycle and climate change. This article gives a very generalized summary of the interactive degradation of soil structure and SOM. It describes the main causes and effects of degradation on soil structure and SOM, the subsequent effects on soil functions, and an outlook to the future in this area.

Causes of soil structure and SOM degradation

There are many natural or management-imposed mechanisms which can lead to the degradation of soil structure and SOM. This article is not exhaustive, but four principal themes known to affect soil structure and SOM interactively are discussed below: land use; tillage; compaction; and climate change (Fig. 1).

Land use

The way in which land is used has a great control on both soil structure and SOM, and this can lead to their degradation (Young et al., 1998; Ferreira et al., 2022) (Fig. 2). In soil under long-term land use, SOM attains an equilibrium based on the balance between inputs and outputs (Johnston et al., 2009). Soil under natural and semi-natural systems, and agricultural and forestry soils

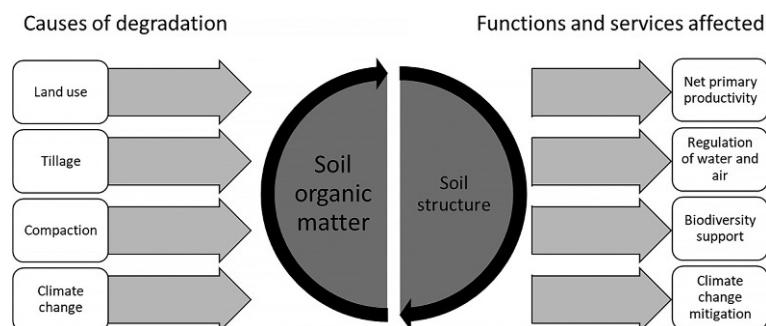


Fig. 1 The main causes of interactive degradation of soil structure and soil organic matter (left) and the effect on soil functions and ecosystem services (right).

Land use	Tillage	Compaction	Climate change
• natural to agricultural systems • perennial to annual systems	• inversion tillage • soil disturbance	• soil densification • soil strength	• changes to soil temperature • changes to soil hydrology

Fig. 2 A summary of the causes of interactive degradation of soil structure and soil organic matter.

under long-term perennial vegetation, have greater SOM contents compared to agricultural soils under annual crops (Sanderman et al., 2017). Perennial systems have a continuous vegetative covering with a constant source of organic inputs through leaf litter and root exudates. They also support a healthy soil biology and microbiology which process such organic inputs. Contrastingly, soils under annual systems generally have smaller organic inputs to the soil because much of the above-ground biomass is harvested and removed. Such systems also tend to be disturbed by cultivations or tillage, which encourages the loss of SOM. The intensity of management of annual systems is an important factor. Systems where management intervenes to apply organic amendments, or to include phases of grass leys in crop rotations, will suffer less degradation of SOM (Johnston et al., 2009). However, this lesser degradation of SOM through use of organic amendments may only be felt in the topsoil, unless applied at very high rates, with subsoils still degraded in SOM beyond what they would be in comparable perennial systems.

It follows that any change in land use affects SOM. Degradation of SOM usually follows a change from natural to semi-natural to perennial to annual systems (The Royal Society, 2020). Typical degradative land use changes are from woodland or grassland to arable systems (Fig. 3). This has been a land use change that has been present over the course of human history as increasingly more land has been put into agricultural production. Restorative land use changes such as taking land out of annual arable systems and back into grassland can restore SOM, but experience shows that it is much harder and slower to regain SOM compared to the rapidity with which SOM can be lost (Or et al., 2021). It can take several decades for soil to adjust to a new land use (Johnston et al., 2009).

A change in land use can also lead to changes in the biochemical characteristics of SOM (Shepherd et al., 2001). Different plant types will have different compositions of compounds and functional groups. Compounds such as amino acids, organic acids and celluloses tend to be labile, or available to microorganisms, whereas fats, waxes and lignins tend to decompose at a much slower rate. There is also a range of functional biochemical groups, such as O-alkyl C and alkyl C groups which are, relatively, available and stable, respectively, though that is a generalization and overall decomposition rates are also controlled by accessibility. Biochemistry is also controlled by the plant part it derives from. For instance, the macromolecule suberin is particularly abundant in root tissue and is believed to be highly resistant to decomposition. The biochemical composition determines bioavailability by controlling the enzymatic rates required for microorganisms to process. Another factor affecting bioavailability to microorganisms is the C-to-nitrogen (N) ratio (C:N). Application of some organic amendments in agriculture can reduce C:N ratios, which favors faster decomposition. In general, systems with a greater diversity of above-ground vegetation will support a greater diversity of SOM in the underlying soil than systems with reduced diversity.

Shifts from perennial grassland or woodland systems to annual arable systems that degrade SOM often cause degradation of soil structure. Rapid reversal of decades of structure formation can occur following land use change. Loss of SOM leads to a loss of structural stability, particularly the stability of soil aggregates (Tisdall and Oades, 1982; Shepherd et al., 2001). This is manifest in a decrease in the mean size of stable aggregates. Land use change has been shown to have a particular effect in reducing particulate SOM content which is thought to contain many of the structural stabilizing components. Loss of soil structure can be evident in a decrease in overall porosity. Larger water-transmitting pores (macropores) are affected more so than smaller water-retaining pores (micropores). Interventions such as the application of organic amendments in arable systems can arrest these declines to some



Fig. 3 Soil cores taken from the Highfield Ley-Arable long-term experiment at Rothamsted Research (UK), approximately 60 years after land use change from existing perennial grassland (left) to annual arable (center) or permanent bare fallow (right). Note the loss of the darker organic layer in the soils under arable and fallow treatments compared to grass.

extent, though the effects are generally only felt in the topsoil. Such amendments tend to increase the particulate SOM content which is linked to soil structure improvement.

Soil structure degradation increases susceptibility of a soil to raindrop impact in annual arable systems where there is a discontinuous vegetative cover of the soil surface. Soil structure decline can be associated with an increase in fine soil particles that are readily-dispersed. Such dispersed soil particles can form a hard seal of aligned particles upon drying to create a surface crust. Surface crusts can be slowly- or completely impermeable compared with an open stable structure often associated with perennial systems (Dexter, 1988). This has implications for the amount of water that can enter the soil, and the availability of water to growing plants and crops.

Loss of SOM can also degrade soil structure by increasing the susceptible to forces of rapid wetting. SOM can impart a degree of hydrophobicity on soils, particularly below a critical water content threshold where there is an inverse relationship between SOM and soil wetting rate (Kay, 1998). Rapid wetting rates of dry soil under annual cropping systems can cause slaking where air entrapped within aggregated structures rapidly increases in pressure to cause an implosion (Tisdall and Oades, 1982; Dexter, 1988; Hallett and Bengough, 2013). In contrast, organic compounds supplied in perennial system increase bond strengths of aggregates and slow the wetting rate through increased hydrophobicity or water repellency. Extracellular polymeric substances, common in non-degraded soil, have been found to reduce the rate of wetting and drying (Hallett and Bengough, 2013).

An extreme example of land use change is soil sealing, where the soil is permanently covered by artificial surfaces of urban areas or infrastructure in the built environment (Huber et al., 2008). This can degrade the structure of the remaining underlying soil due to increased overburden pressures (Ferreira et al., 2022). Although existing SOM can be protected against further mineralization through being sealed, there is often a net loss of SOM from removal of SOM-rich topsoil, and there are few practical opportunities to increase SOM stocks of sealed soil.

Tillage

Tillage or cultivation is the most significant form of mechanical disturbance to the soil and has been described as the largest land engineering operation on the planet (Or et al., 2021). Tillage is conducted to prepare the soil for a crop. 'Conventional' tillage often involves use of a moldboard plow to invert the upper soil layer (typically 20 cm or so), and this continues to be common. Secondary

tillage treatments may then follow, such as use of disks or power harrows to break up soil. Much interest now is on reducing the impact of tillage on the soil, with variations that might be umbrella-termed 'reduced' tillage which includes minimum or shallow tillage, right through to completely zero tillage (also synonymous with direct drilling). Reduced or zero tillage often form a basis of wider agricultural strategies including conservation agriculture and regenerative farming. The full effects of any kind of tillage are site-specific, but for some cropping systems and regions, particularly aridic zones, this is a feasible option.

The rationale and benefits of conventional tillage for soil structure are to modify the structure to break up problematic structures such as surface crusts and large clods to create a seedbed with a range of aggregate and pore sizes (Dexter, 1988; Hallett and Bengough, 2013; Or et al., 2021). Aggregate size variation ensures good seed-soil contact. Aggregates formed through tillage are more likely to be spherical shaped rather than irregular, which generate a porous structure. The variation in pore sizes is important to maintain or create both larger pores to keep the seedbed aerobic and smaller pores that retain plant-available water. Conventional tillage reduces surface roughness which is important for uniformity of seed placement depth and germination. When conducted at the correct time when the soil is friable, typically when just drier than the soil plastic limit (Dexter, 1988), it can improve structure by creating a seedbed and incorporating SOM. This can be a great benefit of conventional tillage to SOM in that it is incorporated below the surface where it is more likely to be protected. Other benefits of tillage include controlling weeds, pests and pathogens arising from the previous crop (Hallett and Bengough, 2013).

Despite these benefits, conventional tillage is often associated with degradation of soil structure and SOM (Fig. 2). Tillage affects structure-forming and stabilizing agents negatively (Shepherd et al., 2001). Plant root exudates help to bind soil particles into aggregates, but tillage can disrupt these and reverse the process. Fungal hyphae are a key aggregating agent as they physically enmesh soil particles together (Kay, 1998). Hyphae are easily destroyed by tillage causing loss of structure (The Royal Society, 2020). Arbuscular mycorrhiza, believed to play a key role in aggregate stability, are particularly affected by tillage (Tisdall and Oades, 1982). Glomalin is a protein produced by arbuscular mycorrhiza fungi which has been shown to respond rapidly to changes in tillage to impart stability (Redmile-Gordon et al., 2020), although the full effects are not fully established and are almost certainly context-specific (Goss and Kay, 2005). Soil structure degradation can sometimes be observed following land use change from undisturbed to tilled systems before noticeable differences in SOM are measured. This has been ascribed to the changes in fine roots and fungal hyphae. At the soil surface, tillage exposes soil to raindrop impact which can increase erosion rates. Aggregates created by tillage are not as stable as those formed in natural systems. Tillage also negatively affects soil macrofauna, most notably earthworm species (*Lumbricus*). Often biopores (those created by macrofauna) are the key component of the soil macroporosity (Or et al., 2021).

Conducted at the wrong time, tillage can degrade soil structure (Fig. 4). Tillage is best suited to periods when the soil is in a brittle condition close to the lower plastic limit for the preparation of seedbeds. As finer-textured soils have a greater range of plasticity than do coarser-textured soils, they are particularly susceptible to plastic deformation when tilled under wet conditions (Hallett and Bengough, 2013). The pore network can be adversely affected—particularly the water-transmitting macropores. Inappropriate or ill-timed tillage can create an impermeable pan at the maximum depth of the plough inversion—termed a plough pan—particularly where there is repeated smearing at the same depth of the plough moving over wet soil (The Royal Society, 2020). Increases in porosity from tillage can be short-lived if the loosened soil slumps when wetted due to aggregate instability. Some finer-textured soils can recover from tillage-induced plastic deformation through wet-dry cycles (and freeze-thaw cycles) that cause shrinking and swelling of the soil—a process that has been termed 'self-mulching'.



Fig. 4 Conventional tillage by moldboard plough (seen here on the Broadbalk Winter Wheat classical experiment at Rothamsted Research, UK) is designed to create a seedbed when the soil is in a brittle state close to the plastic limit, but when conducted under wet conditions tillage can lead to smearing, plastic deformation and degradation of soil structure.

Conventional tillage increases the oxidation of SOM causing its loss. Tillage negatively affects macrofauna populations with implications for SOM. Deep-burrowing earthworms and bioturbation are a key process by which SOM is delivered to the subsoil where it can be stabilized under localized conditions that lessen or prevent oxidation and mineralization. The soil plastic limit—an index important for predicting soil mechanical behavior—is linked to SOM. Soils depleted of SOM have been found to have lower plastic limits meaning that the risk of degrading soil through plastic deformation is at a smaller water content when soils of greater SOM would be in a more brittle state suitable for seedbed preparation (Obour et al., 2018). To minimize the potential for soil structure degradation, such soils may have to be tilled when drier, reducing the window for soil operations. This is important for hard-setting soils (Dexter, 1988).

Reduced or zero tillage in arable agriculture can lessen SOM and soil structure degradation through reducing soil disturbance and maintaining crop residues. Comparison of different tillage practices on SOM have to be made with care however and with reference to soil depth. Apparent improvements in soil structure and SOM in the topsoil (e.g., an increase in topsoil SOM from reduced tillage) might actually be more reflective of a change in the vertical distribution of SOM rather than an overall change (e.g., if SOM is buried below the topsoil under conventional tillage).

Compaction

Compaction is the process whereby soil loses pore volume and hence increases in bulk density. Compaction can be due to natural causes, such as compression in the subsoil arising from overburden stress, and densification in non-rigid fine-textured soils as they shrink upon drying. Localized compaction occurs naturally as plant roots elongate into the soil and necessarily push soil particles together to create the volume for expansion (Dexter, 1988). More often though, compaction is caused by the way in which soils are managed. Very occasionally, compaction is a desired objective in specific circumstances such as the use of soil for foundations and engineering works. Some degree of compaction can be beneficial to achieve good seed-soil contact and to reduce the risk of crop lodging in soils under arable systems. Commonly however, compaction is a degradation process than has clear deleterious effects on soil structure and SOM (Dexter, 1988; Ferreira et al., 2022) (Fig. 2).

Compaction in agricultural systems is largely caused by the use of traffic and machinery to conduct operations (Ferreira et al., 2022). Routine operations such as tillage, drilling, fertilizer application, and harvest are mostly conducted by tractor-powered machinery. In intensive agricultural systems the machinery and loads have tended to increase over time. The extent of compaction is dependent on a number of factors, including ground contact pressure, the number of passes and the porosity and stability of the soil. Wheeling by traffic under wet conditions can deform the soil through shear stresses. Compaction can also occur in agricultural soils under grazing livestock systems (Hu et al., 2021). The compaction pressures under the hooves of larger livestock, such as cattle, can be similar to those of larger tractors, and equally degrading under wet conditions. Some regions have seen significant increases in both the stocking density and the average mass of livestock under some intensive systems. When combined with over-grazing, grassland compaction can be severe, although it is less-well characterized compared to traffic-induced compaction in arable systems.

Compaction increases bulk density. There is often a non-uniform propagation of stress down the soil profile, though this can vary with the equipment used. Compaction-induced porosity loss is not evenly distributed amongst the pore sizes. The water-transmitting macropores are affected most, and there is a corresponding loss of air capacity and pore connectivity (Kay, 1998). Some SOM-stabilized biopores are able to resist compaction more so than pedological macropores however. There is often a change in the alignment of pores from predominantly vertical towards predominantly horizontal, with an associated platy soil structure. This anisotropy of the soil pore network has implications for the hydraulic functioning of soil and gaseous exchange processes (Ferreira et al., 2022), root growth and burrowing by soil fauna.

Soil strength is increased through compaction. Strength can increase to such an extent that root penetration is impacted or prevented completely (Fig. 5). Compaction can cause the creation of an impenetrable pan. This has deleterious effects on the ability of plants to grow. Roots respond to soil structure change due to the effects on cell division and expansion rates. This is manifest in retarded root elongation and increases in root diameter as they expand radially, rather than longitudinally, and thicken in strong soil (Hallett and Bengough, 2013). Root systems are often shallow and restricted to the upper soil layers as a result. Signaling between roots and shoots can then affect above-ground plant growth and crop yield. Studies have shown reactive oxygen signaling and ethylene and auxin responses which control root growth characteristics upon encountering mechanically-hard layers in soil, including 'bending' of roots and proliferation of root hairs to try and explore a way around the obstacle (Jacobsen et al., 2021), with repercussions for leaf growth, morphology, and stomatal closure.

Compaction-induced soil structure degradation can affect SOM. Plant-derived SOM is reduced where plant growth is retarded by compaction. Leaf litter inputs are reduced consequently. Of particular importance is the distribution of roots in the soil profile, which are typically restricted to the upper soil layers where compaction is present (Hallett and Bengough, 2013). The loss of SOM can reduce the resilience of the soil to withstand further physical stresses. SOM affects the ability of soil to compact at smaller stress compared to greater stresses. Under specific conditions however, compaction has been linked to a reduction of SOM loss likely due to reduced loss of dissolved SOM through reduced infiltration (Deurer et al., 2012).

Tillage can reduce bulk density in the upper plough layer however, though the longevity of such changes varies. Soils that ordinarily are under reduced or zero tillage management sometimes require conventional tillage to counter natural increases in soil bulk density in such systems over time (Dexter, 1988; Ball et al., 1997). Traffic-derived soil loading can be reduced further by increasing the contact area of tyres or using tracked vehicles to reduce contact pressure. Reversing compaction effects through tillage is easier done for light-textured soils, though sandy soils are vulnerable to compaction, compared to heavier-textured soils where the

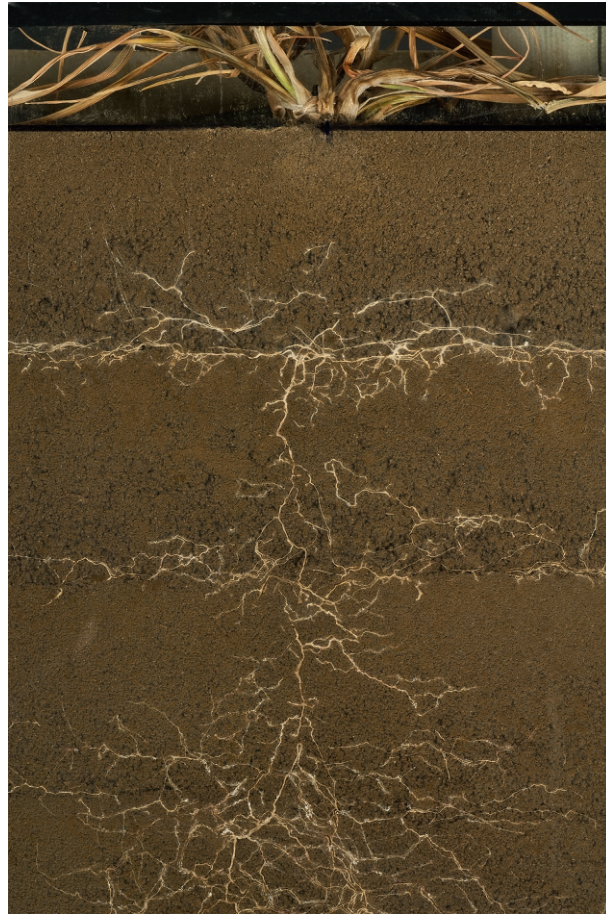


Fig. 5 Compaction affects the ability of roots to penetrate strong soil layers and can lead to lateral rather than vertical growth, as can be demonstrated in experimentation with rhizotrons.

compaction will have tended to result in plastic deformation of soil. Subsoil compaction is particularly difficult because it is rarely detected and the negative effects can be cumulative, less-easily rectified by agricultural intervention, and costly when rectification is possible (Hu et al., 2021).

Some soils are able to recover soil structure to a greater or lesser extent from compaction. Clay soils, especially those containing expanding clay minerals (e.g., smectite, vermiculite) can recover structure through their ability to shrink and swell in response to drying-wetting or freezing-thawing cycles (Dexter, 1988; Hallett and Bengough, 2013). This has been termed 'self-mulching' and SOM is thought to be involved also (Kay, 1998).

Climate change

Global climates are changing. In the northern latitudes, for instance, the trajectory is for hotter, drier summers and warmer, wetter winters than currently, with an increased occurrence of low frequency-high magnitude events such as heatwaves and intense storms (Young et al., 1998; The Royal Society, 2020). Climate change is sure to impact all soil properties directly or indirectly, and soil structure and SOM are no exception (Fig. 2). Soil structure and SOM are explicitly linked to soil biology and microbiology and their processes will undoubtedly respond to changes in climate (Young et al., 1998). The direction of response however is unknown as our understanding of the effects is still in its infancy. Fuller understanding will depend on further controlled experiments (e.g., Free-Air Carbon dioxide Enrichment (FACE) techniques) and advances in modeling. Below is a summary of some current broad ideas.

Greater warming and atmospheric carbon dioxide (CO_2) concentration may increase microbial activity, resulting in either a net loss of SOM through oxidation, and may increase net primary productivity, resulting in a net gain of SOM from greater inputs of shoot- and root-derived residues to the soil. It has been suggested that plant material growing under elevated atmospheric CO_2 conditions may shift in its chemical composition (Young et al., 1998). This could have implications for the turnover of SOM if the lability of the plant-derived residues changes. In addition, there could be changes in the allocation of photosynthate, with evidence suggesting a greater proportion of C being allocated to the roots, with greater turnover of roots as a result (Young et al., 1998). Some models predict significant loss of SOM while others suggest seasonal patterns of decreased respiration during the

growing season and increased respiration in the non-growing season, based largely on the effects of predicted environmental changes in temperature, moisture and CO₂ on plant growth and microbiological processes.

Structure stability will likely degrade if those instance where SOM declines. Organic peat soils can suffer irreversible shrinkage upon drying. Prolonged drying can increase soil hydrophobicity and soil strength. If climate change affects patterns of wetting and drying, then these will impact on soil structure and SOM. In general, rewetting events tend to encourage bursts of microbial activity, which can alter soil structure and expose SOM to oxidation (Young et al., 1998). During periods of intense rainfall, leaching of polyvalent cations such as calcium (Ca) and magnesium (Mg), which play a role in soil structure development and stabilization, may create a more acidic soil susceptible to dispersion (Young et al., 1998). Sea level rise could degrade soil structure in coastal soils where inundation of sea water would be an increased threat due to increased salinity which commonly reduces aggregate stability. Excessive accumulation of soluble salts can affect soil microorganisms with implications for SOM. Increased incidence of wildfires can have degrading effects on SOM as it is thermally oxidized, with effects on soil structure as a result (Ferreira et al., 2022), and likely increases in soil hydrophobicity with implications for soil water dynamics.

Effects of soil structure and SOM degradation on soil functions

Soil structure and SOM are key controls on the ability of soil to carry out its functions in support of wider ecosystem services. The effect of degradation of soil structure and SOM on four principal types of function or service are discussed below: net primary productivity; regulation of water and air; biodiversity support; and climate change mitigation (Fig. 1).

Net primary productivity

In many cases the primary function that a soil carries out is to support vegetation, ranging from climax vegetation in natural ecosystems, to extensively-managed vegetation in semi-natural systems, to intensively-managed crops in agricultural systems. The key parameter of this function is the net primary productivity of the system, including the yield of food and fiber. These are severely affected by soil structure and SOM degradation (Ferreira et al., 2022; The Royal Society, 2020) (Fig. 6).

Degradation of soil structure affects plant growth through increased soil strength, retarded root growth, reduced aeration, increased waterlogging, and a decrease in soil friability (Ball et al., 1997). Crop growth is significantly challenged where soil strength exceeds 2 MPa (Hallett and Bengough, 2013). Degradation arising from tillage can destroy biopores created by earthworms and similar soil engineers, which are often the pores that plant roots use preferentially (Dexter, 1988; Or et al., 2021). Plant productivity is decreased due to effects on decreased soil exploration and soil resource capture by the roots. Excessive water retention arising from degraded soil structure can create anaerobic conditions known to retard root elongation. Crop growth is affected where the macroporosity, or air-filled porosity is less than 10% by volume (Dexter, 1988). Ponding at the soil surface from a compacted soil structure reduces infiltration and vertical hydraulic conductivity and can affect the ability to carry out farm operations in spring on agricultural soils. Erosion on grasslands can have severe effects locally on growth and silage quality. Yields are affected more by soil structure degradation in wetter years, perhaps through waterlogging at key growth stages, compared to drier years (Ball et al., 1997; Hu et al., 2021), although the reverse is more likely to be the case when the strength of soil increases significantly as it dries.

Soil structure degradation can reduce microbial functions. Compaction-induced structural degradation has been associated with a reduction in N mineralization. Denitrification and ammonia (NH₃) volatilization can increase with compaction-induced structural degradation (Ball et al., 1997). Similarly, phosphorus (P) and potassium (K) nutrient uptake rates are reduced in structurally-degraded soils that become waterlogged and C can be lost in dissolved forms, whereas concentrations of other elements and compounds, such as ethylene (C₂H₄) and manganese (Mn), can increase to toxic levels. Soil structural degradation can also increase the susceptibility of a crop to soil-borne diseases and fungi.

Degradation of SOM affects food, fiber and productivity, and any such reduction in these functions has a feedback effect in reducing the amount of plant residues returned to the soil which form SOM. Soil erosion leads to the loss of topsoil and its components of nutrients and SOM, which reduces productivity and yield (The Royal Society, 2020). Degradation of SOM has concomitant effects on supporting plant growth through its effects on plant nutrients. Loss of SOM can reduce the exchange of important nutrients such as N, P, and sulfur (S) (Johnston et al., 2009).

Net primary productivity	Regulation of water and air	Biodiversity support	Climate change mitigation
• reduced growth and yield • reduced plant residue return	• waterlogging and anaerobism • surface runoff and flooding	• reduced habitat space and food • reduced biodiversity	• loss of soil carbon • emission of greenhouse gases

Fig. 6 A summary of the main effects of interactive degradation of soil structure and soil organic matter on soil functions and ecosystem services.

Regulation of water and air

Soil structure degradation has important repercussions for the regulation of water and air (Fig. 6). The propensity for compaction to reduce macroporosity decreases infiltration rates. Compaction preferentially affects macropores and can change the alignment of pores from vertical to parallel with respect to the soil surface. Vertical hydraulic conductivity is severely compromised as a result (The Royal Society, 2020). This increases the risk of surface runoff, through both infiltration-excess and saturation-excess means, and subsequent erosion, loss of topsoil, flooding and pollution of water courses (Ball et al., 1997; The Royal Society, 2020). This is particularly important for surface-applied agrichemicals which can lead to eutrophication of freshwaters through the loss of N and P, and movement of sorbed herbicides and pesticides in eroded soil. Some biopores can be resistant to compaction-induced soil structural degradation however. Biopores created by growing roots and burrowing earthworms tend to have walls strengthened by organic residues and clay particles aligned to the direction of root growth or earthworm movement, making them particularly stabilized. Wider environmental problems can arise from flooding, including the silting of water courses from eroded soil, damage to buildings and infrastructure, and the loss of life (Hu et al., 2021; Ferreira et al., 2022). Tillage can remove surface residues leading to loss of soil water through evaporation.

In the soil, soil structural degradation can increase the potential for waterlogging and reduced aeration (Ball et al., 1997). In the most extreme cases this can cause anaerobic conditions to develop in the soil. Small changes in air-filled porosity can have large effects on diffusivity and air permeability. Storage and transport of heat in soils is also affected. Conversely there have been instances of increases in preferential flow with compaction. Drying of the topsoil can occur where root penetration is impaired, and roots are concentrated in the upper layers. The loss of SOM reduces the amount of water retained by the soil, although the magnitude of this may be small and is inevitably soil-specific (Minasny and McBratney, 2018). A corresponding reduction in the sorption capacity of soils with SOM loss would be expected, which could impact on the ability of soils to retain nutrients and agrichemicals.

Biodiversity support

Soil-dependent biodiversity is dependent on habitat space and a supply of organic material. It follows therefore that degradation of soil structure and SOM, respectively, will adversely affect biodiversity (Huber et al., 2008) (Fig. 6), although the full soil structure-SOM-biology interaction remains little-understood.

The soil pore network component of soil structure provides the habitat for soil biodiversity. It is important that soil has a diversity of soil structural components, providing a variety of habitats and niches, as this tends to lead to increased biodiversity (The Royal Society, 2020). Processes that degrade soil structure by removing larger pores, such as compaction, tends to adversely affect macrofauna such as springtail (Collembola), arthropods (Arthropoda), potworms (Enchytraeidae), and both surface-dwelling and deep-burrowing earthworm species (Gregory et al., 2015; The Royal Society, 2020). Earthworms are linked to productivity and yield. Termites and ants may be affected by soil structure and SOM degradation, but their influence tends to be highly localized around social nests compared to the dispersed populations of non-social earthworms. Soil structural degradation can reduce soil-pupating larvae. Effects on butterflies and moths (Lepidoptera) and some bird species have been noted, presumably due to the effect on soil-dwelling prey (Gregory et al., 2015).

The quantity and quality of SOM, as the primary food source, largely controls the soil microbial biomass and can also control biodiversity. Studies following land-use change between annual arable cropping and perennial grass systems have shown rapid changes in the soil microbiology, including increases in fungal populations and the mesofauna in the latter (Hirsch et al., 2017), although bacterial diversity is unaffected (Hirsch et al., 2009). Studies have reported that increased quantity of SOM can increase populations of larger detritivores (anecic earthworms) and decrease smaller groups (springtail), and that low-quality recalcitrant SOM is linked to increases in the relative proportion of fungi and fungal-feeding nematodes (Nematoda) (Sauvadet et al., 2016). SOM has been positively correlated with the abundance of flies (Diptera) and aerially-active beetles (Coleoptera) in arable soils (Gregory et al., 2015). Lower fungal biomass and fungal-to-bacterial biomass ratios have been found in soils of smaller SOM content compared with undisturbed SOM-rich soils. Fungi are particularly affected by soil structure degradation (e.g., especially tillage operations) and a change in land use can decrease SOM (Tisdall and Oades, 1982).

Climate change mitigation

Soils have a crucial role to play to help mitigate climate change through its part in the global C cycle. Soils contains more C globally than the atmosphere and vegetation combined. Small changes in C fluxes can have a great effect on the global C cycle therefore. Degradation of SOM causes a release of C from the soil to the wider environment—most commonly to the atmosphere. Soils could flip between being a sink and a source of C if degradation occurs (Fig. 6). Principle examples of where a soil may change from C sink to C source include losses of SOM from cultivation of semi-natural land, urbanization, drainage of naturally humid soils (e.g., peatlands and fenlands), and peat extraction (Cannell et al., 1999).

Soil structure degradation can exacerbate problems. Emissions of methane (CH_4) can occur under low-oxygen conditions when SOM is degraded by methanogenic bacterial microorganisms that excrete it as a by-product (The Royal Society, 2020). Similarly, emissions of nitrous oxide (N_2O) occur under anaerobic conditions, in the presence of nitrate (NO_3^-) and available C, when denitrification occurs (Ball et al., 1997; Hu et al., 2021). As CH_4 and N_2O are significantly more potent greenhouse gases than CO_2 (approximately 30 and 300 times the global warming potential of CO_2 , respectively), this can seriously compromise the ability

of soils to mitigate climate change. Erosion of SOM-containing topsoil can release previously-protected C to the atmosphere (Ferreira et al., 2022).

Other soil functions and ecosystem services

In addition to the four principal areas discussed above, degradation of soil structure and SOM can impact on other soil functions and ecosystem services.

SOM plays a key role in the ability of soils to buffer the effects of potentially-toxic substances, in part by chelation and adsorption (Ferreira et al., 2022). Loss of SOM can reverse this and release toxic elements. Waterlogging following structural degradation causes reducing conditions in soils which can impact on buffering and contaminant filtering. This can affect the quality of drinking water supplies and can lead to eutrophication problems in freshwaters.

Soils are the foundation for many amenity uses of land, such as for sport and recreation. Many amenity uses are reliant on a healthy resilient vegetative covering, such as turf. It is apparent therefore that degradation of soil structure and SOM has a negative impact on the function of soils to support most amenity uses through effects on quality of turf and the ability of the ground to resist stresses. There are some minor exceptions to this, such as the deliberate degradation of soil structure by compaction for the sport of cricket. Soils are also the repository of cultural heritage. Soil structural degradation can negatively affect the preservation of cultural artifacts and archeology in the soil. However, whether waterlogging following structural degradation preserves or degrades archeological materials in soils is not clear.

Soils have an important function as a support for the built environment and infrastructure. This is an example of a function that is often dependent on deliberately degrading soil structure and SOM to create a compacted inert foundation for building or engineering works. Nevertheless, changes in soil structure such as shrinking can adversely affect soils as building foundations (Young et al., 1998).

Summary and outlook

Soil structure and SOM have an interactive relationship and hence it is demonstrably the case that a degradation on one will impact the other. This article has sought to summarize the key processes and mechanisms by which soil structure and SOM can be degraded. Some of those processes and mechanisms are natural, such as those associated with natural pedological soil-forming factors, including the stresses imparted by the ordinary climatic conditions under which a soil is subjected. However, it is the case that many of the ways in which soil structure and SOM are degraded are as a direct or indirect result of management interventions. Management, particularly in agriculture, controls surface vegetation, inputs of organic residues from vegetation, livestock and amendments, and soil physical conditions through tillage and traffic-induced compaction. All of these impact on soil structure and SOM. Subsequently these have implications for soil functions and ecosystem services that are reliant on soil structure and SOM, including net primary productivity, regulation of water and air, and biodiversity support. Potentially mitigating, but most likely aggravating, the degradation of soil structure and SOM is the changing climate, which itself is affected by the dynamics of soil structure and SOM.

The positive message is that we have as a foundation a basic theoretical fundamental understanding of what causes the degradation of soil structure and SOM, and how it affects the functions and services we ask of soils. There is the caveat that soils are arguably the most complex systems on the planet with huge natural variation in characteristics, which makes their study and understanding equally challenging but compelling. Generalizations often have to be adapted to local conditions. Nevertheless, there are broad management interventions that ought to slow or reverse degradation of soil structure and SOM. We can modify agricultural practices to prevent or minimize SOM degradation, including use of crop rotations, grass leys, perennials, cover cropping, agroforestry, and organic amendments. We can reduce soil structure degradation by minimizing tillage, reducing traffic tyre or track pressures, using controlled traffic systems, reducing livestock density, and reducing soil stresses at particularly vulnerable times, to reduce soil compaction. Reversals of previous soil structure and SOM degradation may be achieved through addition of organic amendments and careful soil engineering. Seemingly small changes in SOM may be manifest in disproportionate great effects on soil structure-controlled properties and functions, because certain SOM components, such as particulates and polysaccharides, are likely more important than total SOM per se.

We can still improve our knowledge of the fundamental soil structure-SOM interaction, how this is controlled by management, and the subsequent effects on soil functions and ecosystem services. Future research will be able to exploit developing technologies and methodologies. These include the great advances that have been made in the non-destructive characterization of soil structure (see Non-invasive imaging of soil processes) and physical, biological, and chemical characterization of SOM (see Spectroscopic Methods (FTIR, Diffuse Reflectance Spectroscopies)) together with the greater power of mechanistic modeling to predict future trajectories.

References

- Ball BC, Campbell DJ, Douglas JT, Henshall JK, and O'Sullivan MF (1997) Soil structural quality, compaction and land management. *European Journal of Soil Science* 48: 593–601. <https://doi.org/10.1046/j.1365-2389.1997.00120.x>.

- Cannell MGR, Milne R, Hargreaves KJ, et al. (1999) National inventories of terrestrial carbon sources and sinks: The U.K. experience. *Climatic Change* 42: 505–530. <https://doi.org/10.1023/A:1005425807434>.
- Deurer M, Müller K, Kim I, et al. (2012) Can minor compaction increase soil carbon sequestration? A case study in a soil under a wheel-track in an orchard. *Geoderma* 183–184: 74–79. <https://doi.org/10.1016/j.geoderma.2012.02.013>.
- Dexter AR (1988) Advances in characterization of soil structure. *Soil & Tillage Research* 11: 199–238. [https://doi.org/10.1016/0167-1987\(88\)90002-5](https://doi.org/10.1016/0167-1987(88)90002-5).
- Ferreira CSS, Seifollahi-Aghmiuni S, Destouni G, Ghajarnia N, and Kalantari Z (2022) Soil degradation in the European Mediterranean region: Processes, status and consequences. *Science of the Total Environment* 805: 150106. <https://doi.org/10.1016/j.scitotenv.2021.150106>.
- Goss MJ and Kay BD (2005) Soil Aggregation. In: Zobel RW and Wright SF (eds.) *Roots and Soil Management: Interactions between Roots and the Soil*. Agronomy Monograph 48pp. 163–180. Madison, WI: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America. <https://doi.org/10.2134/agronmonogr48.c9>.
- Gregory AS, Ritz K, McGrath SP, et al. (2015) A review of the impacts of degradation threats on soil properties in the UK. *Soil Use and Management* 31(supplement 1): 1–15. <https://doi.org/10.1111/sum.12212>.
- Hallett PD and Bengough AG (2013) Managing the soil physical environment for plants. In: Gregory PJ and Nortcliff S (eds.) *Soil Conditions and Plant Growth*, 12th edn., pp. 238–268. Chichester: Wiley-Blackwell. <https://doi.org/10.1002/9781118337295.ch8>.
- Hirsch PR, Gilliam LM, Sohi SP, et al. (2009) Starving the soil of plant inputs for 50 years reduces abundance but not diversity of soil bacterial communities. *Soil Biology & Biochemistry* 41: 2021–2024. <https://doi.org/10.1016/j.soilbio.2009.07.011>.
- Hirsch PR, Jhureea D, Williams JK, et al. (2017) Soil resilience and recovery: Rapid community responses to management changes. *Plant and Soil* 412: 283–297. <https://doi.org/10.1007/s11040-016-3068-x>.
- Hu W, Drewry J, Beare M, Eger A, and Müller K (2021) Compaction induced soil structural degradation affects productivity and environmental outcomes: A review and New Zealand case study. *Geoderma* 395: 115035. <https://doi.org/10.1016/j.geoderma.2021.115035>.
- Huber S, Prokop G, and Arrouays D, et al. (eds.) (2008) *Environmental Assessment of Soil for Monitoring*. In: Vol. I, *Indicators & Criteria*, EUR 23490 EN/Luxembourg: Office for the Official Publications of the European Communities.
- Jacobsen AGR, Jervis G, Xu J, Topping JF, and Lindsey K (2021) Root growth responses to mechanical impedance are regulated by a network of ROS, ethylene and auxin signalling in Arabidopsis. *New Phytologist* 231: 225–242. <https://doi.org/10.1111/nph.17180>.
- Johnston AE, Poulton PR, and Coleman K (2009) Soil organic matter: Its importance in sustainable agriculture and carbon dioxide fluxes. *Advances in Agronomy* 101: 1–57. [https://doi.org/10.1016/S0065-2113\(08\)00801-8](https://doi.org/10.1016/S0065-2113(08)00801-8).
- Kay BD (1998) Soil structure and organic carbon: A review. In: Lal R, Kimble JM, Follett RF, and Stewart BA (eds.) *Soil Processes and the Carbon Cycle*. *Advances in Soil Science* 169–197. Boca Raton: CRC Press LLC.
- Minasny B and McBratney AB (2018) Limited effect of organic matter on soil available water capacity. *European Journal of Soil Science* 69: 39–47. <https://doi.org/10.1111/ejss.12475>.
- Obour PB, Jensen JL, Lamandé M, Watts CW, and Munkholm LJ (2018) Soil organic matter widens the range of water contents for tillage. *Soil & Tillage Research* 182: 57–65. <https://doi.org/10.1016/j.still.2018.05.001>.
- Or D, Keller T, and Schlesinger WH (2021) Natural and managed soil structure: On the fragile scaffolding for soil functioning. *Soil & Tillage Research* 208: 104912. <https://doi.org/10.1016/j.still.2020.104912>.
- Redmile-Gordon M, Gregory AS, White RP, and Watts CW (2020) Soil organic carbon, extracellular polymeric substances (EPS), and soil structural stability as affected by previous and current land-use. *Geoderma* 363: 114143. <https://doi.org/10.1016/j.geoderma.2019.114143>.
- Sanderman J, Hengl T, and Fiske GJ (2017) Soil carbon debt of 12,000 years of human land use. *Proceedings of the National Academy of Sciences of the United States of America* 114: 9575–9580. <https://doi.org/10.1073/pnas.1706103114>.
- Sauvadet M, Chauvat M, Fanin N, Coulibaly S, and Bertrand I (2016) Comparing the effects of litter quantity and quality on soil biota structure and functioning: Application to a cultivated soil in Northern France. *Applied Soil Ecology* 107: 261–271. <https://doi.org/10.1016/j.apsoil.2016.06.010>.
- Shepherd TG, Saggar S, Newman RH, Ross CW, and Dando JL (2001) Tillage-induced changes to soil structure and organic carbon fractions in New Zealand soils. *Australian Journal of Soil Research* 39: 465–489. <https://doi.org/10.1071/sr00018>.
- The Royal Society (2020) *Soil Structure and Its Benefits: An Evidence Synthesis*. <https://royalsociety.org/-/media/policy/projects/soil-structures/soil-structure-evidence-synthesis-report.pdf>
- Tisdall JM and Oades JM (1982) Organic matter and water-stable aggregates in soils. *Journal of Soil Science* 33: 141–163. <https://doi.org/10.1111/j.1365-2389.1982.tb01755.x>.
- Young IM, Blanchart E, Chenu C, et al. (1998) The interaction of soil biota and soil structure under global change. *Global Change Biology* 4: 703–712. <https://doi.org/10.1046/j.1365-2486.1998.00194.x>.

Drainage, soil water storage, buffering and filtering

Jeffrey Strock^a and **Norm Fausey^{b,*}**, ^aUniversity of Minnesota, Southwest Research and Outreach Center, Lamberton, MN, United States;
^bUSDA Agricultural Research Service, Soil Drainage Research Unit, Columbus, OH, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of N.R. Fausey, DRAINAGE, SURFACE AND SUBSURFACE, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 409–413, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00352-0>.

Introduction	257
Climate change	258
Temperature	258
Precipitation	258
Soil hydrology	259
Drainage	259
Surface drainage	259
Subsurface drainage	259
Conventional free drainage	260
Controlled drainage	261
Environmental effects of drainage	261
Practices to improve water quality and peak flows	262
Buffers and filters	262
Controlled drainage	262
Cover crops	262
Wetlands	262
Bioreactors	263
Saturated buffers	263
Drainage water recycling	263
Water storage	263
Conclusion and outlook	263
References	264

Abstract

Drainage is the natural or artificial removal of water from the land either through surface drainage, subsurface drainage or a combination of both. The practice of land drainage may be beneficial if soil water conditions create a need for drainage due to shallow groundwater or soils with low permeability. In agricultural settings, the natural, internal drainage of most soils is adequate to prevent waterlogging, but many soils require artificial drainage to improve field trafficability for field operations and to protect growing crops from excess soil water conditions. Drainage helps to provide an optimal balance of water and air in the soil necessary for plant growth, as well as the diverse function and services of soil-dwelling organisms.

Key points

- Artificial drainage mitigates risk and damage to crop production from excess soil water conditions.
- Drainage provides optimal conditions for growing plants and soil organisms.
- Drainage system planning, design, installation, operation, and maintenance are critical for success.

Introduction

Drainage is a natural process and is a component of the global hydrologic cycle whereby water moves across, in, through, and out of the soil as a result of various forces. Historically, land drainage might be undertaken for one or more of the following reasons: improvement of the agricultural potential of soils; the reclamation of wet soils to improve the stability of roads and building sites; or for health reasons, to reduce the potential for malaria or other infectious diseases. According to the International Commission on Irrigation and Drainage (ICID, 2017), the global land area estimated to be artificially drained is about 194 Mha. Drainage of the soil

* Retired.

can be accelerated by the use of surface and subsurface drainage practices. Soils that require accelerated drainage typically have some impermeable or slowly permeable feature at or below the soil surface that either limits water from entering the soil or after the water has entered the soil and restricts its movement deeper into the soil and underlying materials at a rate that allows efficient and economical use of the water resource.

Drainage systems permit or improve crop growth and production through removal of non-capillary water from the soil profile. The essential functions of these drainage systems are to prevent flooding through the rapid removal of surface and non-capillary soil water, lower shallow water tables to prevent crop stress, and create suitable conditions for field operations such as tillage, fertilization, application of crop protection products, and harvest. Wet soils can also delay planting, which has a direct negative effect on yield potential due to a reduction in growing degree days. In addition, plant and root diseases are more common in wet, poorly drained soils and plants may be more susceptible to diseases in these stressful growing conditions (Strock, 2018).

The importance of proper drainage and drainage system management are critical for crop production on naturally poorly drained soils in humid and sub-humid regions affecting millions of hectares. In addition, drainage has direct and indirect effects on the provision of ecosystem services including impacts on soil organic carbon (SOC) stocks, organisms living in the soil, nutrient cycling, water retention on the landscape, groundwater recharge, greenhouse gas (GHG) emissions, nutrient losses to surface and groundwater, riparian vegetation, wildlife habitat, and the resilience of cropping systems to climate change.

Climate change

There is an increasing body of evidence that supports the proposition that climate change will impact ecosystems, communities, and human livelihood (Meerhoff et al., 2022; IPCC, 2013; Kumar et al., 2014; Parmesan and Yohe, 2003; Patz et al., 2005; Wheeler and von Braun, 2013). Two commonly ascribed effects of climate change are increasing temperature and extreme weather variability. Anticipated changes in the frequency and intensity of extreme events (e.g. droughts, heatwaves, severe storms, extreme rains) may also affect ecosystem stability and resilience (García-Palacios et al., 2018; Pimm et al., 2019). Hydrologic and thermal effects of projected climate change will have direct and indirect consequences for freshwater ecosystems, for example, human water use, increased incidence of flooding and low flows, the rate of growth and distribution of freshwater organisms, and impacts on water quality (van Vliet et al., 2013).

Temperature

Temperature is an important environmental factor that effects physical, chemical and biological properties and processes in aquatic ecosystems. Water temperature can affect chemical (e.g., dissolved oxygen concentration) and biological (e.g., decomposition, growth rate of aquatic organisms) processes occurring in a water body. Water temperature fluctuations can occur naturally or as a result of human disturbances such as thermal pollution, riparian vegetation removal (e.g. deforestation), and climate change.

One of the earliest responses to climate change is expected to be alterations in river and lake water temperature. Factors influencing the thermal regime of surface waters can generally be classified into four groups: (1) atmospheric conditions (e.g. solar radiation, air temperature); (2) topography (e.g. latitude/altitude, shading); (3) streambed (e.g. geothermal conduction, groundwater input); and (4) stream discharge (e.g. inflow/outflow) (Caissie, 2006). Stream discharge, mostly a function of water hydraulics (e.g. inflows and outflows), mainly influences the heating capacity (volume of water) and or cooling through mixing of water from different sources including influx of precipitation, surface runoff, subsurface drainage, and groundwater.

Precipitation

The need for drainage can vary within a growing season due to seasonal climate patterns and differences in crop tolerance to excess water throughout the growing season. Climate projection models predict that, in the future, areas with high precipitation will become even wetter, especially in the Northern Hemisphere (Holopainen et al., 2016; IPCC, 2019). Changing precipitation patterns will complicate the effects of agricultural subsurface drainage, particularly in regions where the frequency and magnitude of heavy rain events are increasing relative to average precipitation (IPCC, 2013). Projected increases in precipitation patterns under future climate scenarios will probably increase surface runoff and subsurface drainage, making water management more challenging. It is likely that more frequent, intense and higher amounts of precipitation will increase nutrient loss from agricultural land and affect non-point source pollution thereby increasing nutrient concentrations in receiving water bodies. Nutrient exports from agricultural watersheds are mainly attributed to surface and subsurface drainage networks that promote rapid drainage, but also to the relatively large inputs these watersheds receive from fertilizer and livestock manures. Subsequently, as water temperature and nutrient concentrations increase, algal growth is expected to be stimulated, leading to water eutrophication and algal blooms. Consequently, eutrophication will be accelerated due to the increased nutrient supply, higher temperatures, and greater net primary production.

Soil hydrology

Precipitation, as well as irrigation, delivers water to the land surface. The frequency, intensity, duration, rate, and amount of precipitation and the properties of the soil determine the relative amounts of water that will infiltrate, runoff, or remain on the soil surface temporarily as surface water storage. Water that runs off does so by the natural process of surface drainage, sometimes called runoff or overland flow. The water that remains temporarily stored on the surface will eventually either evaporate or infiltrate into the soil matrix. Water that infiltrates into the soil increases the soil water content and may ultimately be evaporated back to the atmosphere directly, be taken up by plants, or move deeper into the soil where it could then undergo lateral flow and move beyond the area of infiltration.

The forces of cohesion, adhesion, and gravity control the redistribution of water that infiltrates into the soil. Water infiltrates into a network of pores within and between soil particles and structural aggregates and adheres to these components. Additional soil water accumulates as water coheres to water already in contact with the soil particles. Within small pores, enough water can be held against the force of gravity to fill the pores completely. Pores that hold water against the force of gravity are called capillary-size pores. In larger pores, all of the water that enters cannot be held against the force of gravity and some of the water moves downward through the soil in these pores. This process is called percolation; percolation is a natural drainage process. Water that percolates deeper into the soil may: enter capillary-size pores at greater depths and increase the soil-water content; encounter restrictive features that block its descent and cause accumulation of water in the form of a perched water table; reach and add to an existing water table; and or reach and replenish an aquifer.

Drainage

Land drainage can be accomplished by surface drainage, subsurface drainage or a combination of both systems.

Surface drainage

Surface drainage diverts excess water from the soil surface directly to streams, thereby reducing the amount of water that will move into and possibly through the soil. Surface drainage is accomplished by land forming, grading, or smoothing to remove small depressions where water collects and to provide a uniform slope for the removal of surface water. The orientation of the graded land and thus direction of surface drainage is normally toward an adjacent open drain (ditch). If the ditch is sufficiently deep it will also receive water transported laterally through the soil profile. A field is considered to have good surface drainage if it has a small amount of depressional storage that must be satisfied before surface drainage occurs. In contrast, poor surface drainage is characterized by a large amount of depressional storage that must be satisfied before surface drainage can occur. Water management by means of surface drainage enhances overland runoff and thus requires less intensive subsurface drainage than would be needed if surface drainage were poor. The main concerns regarding surface drainage are increases in rates of peak flow in receiving waters, loss of wetlands, increased loss of sediment and sediment-bound nutrients, and the loss of other pollutants found in surface runoff, such as pesticides and bacteria.

Subsurface drainage

Subsurface drainage describes the process of removal of that water which has infiltrated into the soil in excess of the amount that can be held by capillary forces against the force of gravity. The goal of subsurface drainage systems is to remove non-capillary water from the upper layers of the soil profile as quickly as possible to ensure an adequately aerated root zone and trafficability for critical field operations such as planting and harvesting. Subsurface drainage systems to manage shallow water tables are designed to remove excess pore water from the soil profile to a specific design depth, typically between 60 and 120 cm. Excess pore water is defined as moisture contents greater than field capacity, defined as pressures less than 10-kPa tension for coarse textured soils and 33-kPa tension for fine textured soils. When soil moisture exceeds field capacity, or when the water table rises above the drain depth, soil water from larger pores can move to subsurface drainpipes or ditches. Subsurface drainage, provided by ditches and drainpipes (aka 'tile'), collects and diverts non-capillary water from within the soil directly to ditches, or streams and rivers. Buried drainpipes may be constructed of clay, concrete, or perforated, corrugated polyethylene pipe. Drainpipes act as conduits that transmit water through their walls and can be placed at a range of depths and spacing's in the soil. By connecting ditches and buried drainpipes, subsurface drainage systems can be created to remove non-capillary water from broad areas of soil.

Design, installation, and operation of subsurface drainage systems require thorough planning and consideration of factors including: climate, crop requirements, soil properties, and site characteristics (Strock et al., 2011). It is widely accepted that some crops are more tolerant of poorly drained, wet conditions than others. Water management for most crop production is important during the growing season. As drainage pertains to climate, it is conceivable that there are times during the growing season when an entire rainfall will infiltrate into the soil. At other times, when rainfall is frequent and intense, the soil may become saturated with water and it may be necessary to remove excess water within 24 h in order to minimize crop damage. There may also be times when

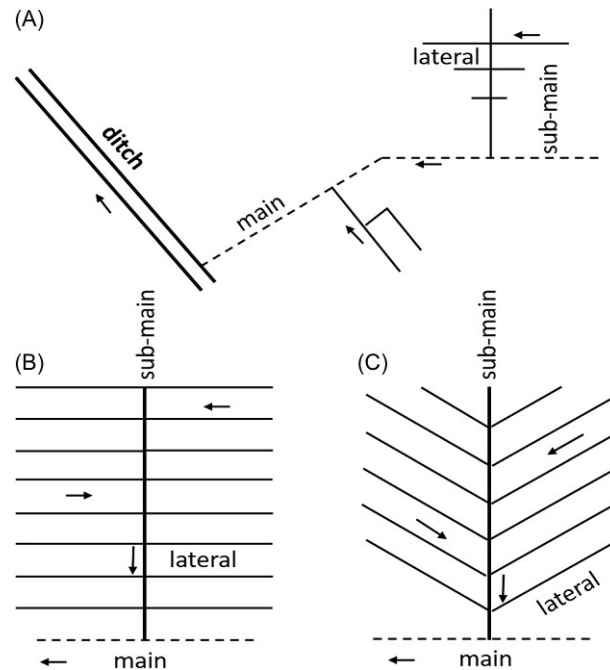


Fig. 1 Schematic diagram of the conceptual drainage system design patterns (A) Random, (B) Parallel, and (C) Herringbone.

the intensity of rainfall exceeds the infiltration capacity of the soil and water is forced to run off the soil surface. Subsurface drainage systems to manage shallow water tables consists of two distinctly different types: conventional free drainage and controlled drainage.

Conventional free drainage

Subsurface drainage systems are commonly installed in either a random, parallel, or herringbone pattern (Fig. 1). A random design may be used where field topography is rolling in order to remove excess water from wet areas in a field. In the case of the random pattern, a single tile drain may be installed through several isolated depressions. A parallel drainage system consists of lateral tiles that are oriented perpendicular to a main field tile drain to drain an entire field. The herringbone configuration consists of lateral drain tiles, oriented in a diagonal manner, entering a main drain from both sides.

Design, installation, and operation of a subsurface drainage system requires thorough planning and consideration of many factors including: soil properties, climate, crop, economic, and environmental factors. The risks of excess water on crop production and profitability depend on these factors. To a great extent drainage design, installation and operation are site-specific activities. Subsurface drainage design is a process of selecting drainage depth, spacing, and layout to mitigate the negative consequences associated with excess water in the soil profile, while balancing the costs associated with the drainage system (Sands, 2003). The ultimate choices of drain depth, spacing, and layout have to be considered to optimize drainage system design where net returns on investment (benefits minus the costs of the system) are maximized over the life of the system.

The proper depth and spacing of subsurface drains depend on multiple factors including site conditions, management factors, and soil characteristics. It is not within the scope of this chapter to describe fully drainage system planning, design and installation, however, there are resources readily available to aid in the process (e.g. Smedema et al., 2004). Briefly, drain depth and spacing should be determined such that the drainage system can reduce the depth of the water table between the drain lines to an elevation that is most beneficial to plant growth within 24–48 h after a rain. Subsurface drain pipes come in various diameters ranging from 5 to 60 cm in diameter. These drains are usually installed at depths of 0.6 to 1.5 m deep. Drain spacing for parallel drains can be determined empirically for a given soil using the Hooghoudt equation (Eq. 1) as described by van der Ploeg et al. (1999):

$$L = \left[\frac{8K_2md_o + 4K_1m^2}{R} \right]^{1/2}, \quad (1)$$

where L is the distance between drain lines, R is the design drainage rate, K_1 and K_2 are the hydraulic conductivities of the soil layers above and below the drain, respectively, m is the midpoint water table elevation above the drain tile, and d_o is the effective depth to the impermeable layer. The use of Eq. (1) begins with selecting a drainage rate, R , for a specific site based on soil, crop, climate, and other factors. Design drainage rates can be found in Framji et al. (1987). Typically, drainage rates range from 6 to 40 mm^d. Once the design drainage rate and drainage depth are selected, drain spacing can be determined solving Eq. (1) for L .

Although all of the potential combinations of site conditions, management factors, and soil characteristics cannot be adequately described in the context of this chapter, in general, as soil texture becomes finer (i.e. clay content increases) it may be necessary to

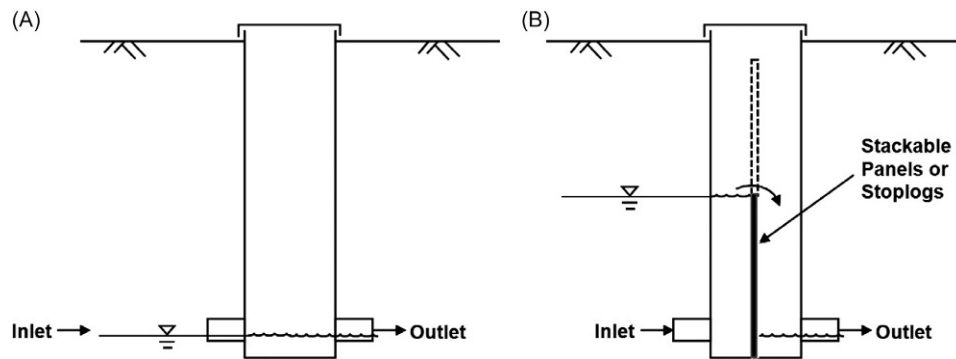


Fig. 2 Schematic diagram of the conventional free drainage and controlled drainage concepts, using an in-field water control structure.

install drain pipes closer together. In contrast, as the soil texture becomes coarser (i.e. clay content decreases) drain spacing may become wider.

Controlled drainage

The previous section focused on conventional free drainage where the drainage system is managed so that the water table is effectively lowered to the depth of the drainage system outlet. This level of management permits farming under conditions where soils have poor natural drainage and is sufficient to satisfy agricultural production goals. However, in some cases, this level of management may lead to excessive drainage which is undesirable during the growing season since it reduces the amount of plant available soil water and could lead to increased water stress and yield reductions during unusually dry periods. Excessive drainage may also lead to leaching losses of important fertilizer nutrients and other constituents transporting them to surface waters where they may lead to impaired water quality.

Controlled drainage is a practice that has developed for temporarily storing water in the soil profile and reducing nitrogen loss from agricultural fields with subsurface drainage (Skaggs et al., 2010; Helmers et al., 2022). Controlled drainage influences soil hydrology since it can be used to raise a shallow water table, which increases soil wetness in the upper part of the soil profile. Controlled drainage is accomplished through management of the drainage outlet which can be set at any level between the ground surface (undrained) and the drain depth (conventional free drainage) depending on farming needs, precipitation–evapotranspiration conditions, soil water content, and water table elevation (Fig. 2). Raising the outlet elevation of a controlled drainage system effectively reduces overall drainage volume and consequently nutrient loss (Skaggs et al., 2010; Youssef et al., 2018; Helmers et al., 2022).

The suitability of a field for controlled drainage must be considered during the planning and design stages of the drainage project. Landscapes which support the implementation of controlled drainage are normally flat, with slopes less than 0.5%. As slopes approach 1%, the number of control structures required to maintain a shallow water table within the desired management depth increases as does cost. At slopes greater than 1%, drainpipes can be installed following natural topographic contours. Controlled drainage can be used with any drain spacing; however, narrower drain spacing is common in order to maintain a relatively 'flat' shallow water table. Narrow drain spacing also reduces the risk of yield loss due to excess wetness during the growing season because water is removed from the soil at a faster rate. Narrower drain spacing also increase drainage system cost.

To maximize the benefits of controlled drainage, farmers/managers must actively manage the control structures throughout the year. Controlled drainage systems can be managed during the growing season to alleviate plant stresses caused by dry conditions, potentially resulting in higher crop yields. In contrast, if poorly managed, controlled drainage may also have an adverse effect on crop growth and yield during wet growing seasons since it may delay and or reduce the removal of non-capillary soil water in the root zone, contributing to plant stresses due to wet conditions (Youssef et al., 2022). The potentially positive and negative effects of controlled drainage on crop performance depend on the frequency, duration, timing and severity of dry and or wet conditions and how intensively a controlled drainage system is managed in response to these conditions.

The optimization of crop performance by controlled drainage is typically achieved by managing water table depths between 45 and 60 cm from the soil surface, halfway between drainpipes. Management of the drainage system outlet depth will vary depending on factors including drainage system design, soil texture, crop type, and growth stage. Ideal water table depths for coarse textured (e.g. sandy) soils, which are susceptible to dry conditions, should be shallower and closer to 45 cm. In contrast, for fine textured (e.g. clay loam) soils, which would be susceptible to wet conditions, should be managed at 60 cm or deeper.

Environmental effects of drainage

Although many sources, including wastewater treatment plants, domestic septic systems, and urban nonpoint sources contribute water and nutrients to surface waters, agriculture is the foremost source in many areas of the world. Numerous adverse effects of drainage on water quality have been identified; specifically, artificial drainage increases nitrogen and phosphorus loads to

downstream waters (Dinnes et al., 2002; Gentry et al., 2007; Wang and Li, 2019). Some of the adverse effects of drainage are avoidable and some can be minimized by the adoption and implementation of various management practices.

Recognition that artificial drainage contributes to impaired surface waters has led to the development of methods and technology for water table management (Strock et al., 2010), and also practices aimed at addressing the reduction in both nitrogen and phosphorus loads (Baulch et al., 2019; Xia et al., 2020). Loss of nitrate-nitrogen from artificially drained agricultural land has been the focus of numerous research studies. Research has concentrated on factors including precipitation; soil organic matter mineralization; cropping system; timing, source, and rate of nitrogen application; use of nitrification inhibitors and slow-release nitrogen fertilizers; tillage; and drainage system depth and spacing (Randall and Goss, 2008). Nitrate-nitrogen loading has been associated with hypoxia in coastal waters, but recent evidence indicates that phosphorus also plays a critical role (Dodds, 2006; Rabalais and Turner, 2019). Phosphorus loss has been an important water quality issue for decades because of the critical role P plays in eutrophication. Historically, the majority of research on phosphorus losses has focused on surface runoff and erosion because such losses were considered negligible from subsurface drainage. More recently, research has revealed a better understanding of the magnitude and importance of subsurface drainage on dissolved phosphorus transport and loading to surface water (King et al., 2015). Subsurface drainage has been shown to reduce total phosphorus losses because drainage reduces surface erosion and surface runoff which are the main drivers for sediment associated phosphorus export (Wagena and Easton, 2018).

Practices to improve water quality and peak flows

Drainage coupled with climate change has the potential to affect hydrology and diffuse nutrient export from agricultural landscapes. Climate change is causing an increase in the magnitude and frequency of large precipitations events, and longer, more frequent droughts. Increasing flood and drought risk contribute to the need for a robust suite of hydrology and water management practices to support adaptation and mitigation options in the near future. The increase in amount and intensity of precipitation will contribute to greater diffuse nutrient and sediment export from landscapes.

Buffers and filters

Buffers and filter strips have been found to be effective in trapping contaminants, especially particulate pollutants (Dabney et al., 2006). There are numerous terms and definitions for filter strips and riparian buffers including vegetative filter strips, conservation filters, conservation buffers, riparian buffer zone, buffer strips, riparian buffer. Filter strips primarily slow runoff velocities and filter out sediment and other pollutants from surface runoff and erosion. Filter strips can be located anywhere on the landscape and do not need to be adjacent to a surface waterbody. In contrast, riparian buffers must be located adjacent to a surface waterbody such as a lake or river. While there has been extensive use of filter strips and riparian buffers to improve water quality, it is generally not possible to suggest a size for them to meet specific water quality improvement goals due to difficulties with predicting their performance effectiveness (Helmets et al., 2008). The performance of a buffer or riparian system depends on the field, topographic, and climatic conditions at the site. The main constraints to establishing buffer or filter systems are associated with the cost of establishing them and in taking land out of production.

Controlled drainage

Controlled drainage was described in detail in the section ‘Subsurface drainage.’

Cover crops

Row crop agriculture is under scrutiny with the aim to reduce runoff, erosion, and nutrient concentrations and loads in drained agricultural landscapes. The use of cover crops and applying appropriate rates of fertilizer, especially N, for maize (*Zea mays* L.) are potential management strategies to reduce nutrient losses in tile drainage water (Dinnes et al., 2002). Cover crops have the potential to decrease soil erosion and particulate nutrient loss in surface runoff and leaching, however their effectiveness is reduced when germination and stand cover are limited by poor germination, by lack of fall precipitation, or short growing seasons (Strock et al., 2004; Noland et al., 2018).

Wetlands

Natural, restored, and constructed wetlands represent a potentially promising practice for treating agricultural drainage whether the influent consists of surface water runoff, subsurface drainage, or a combination of both. Research has demonstrated that wetlands can provide removal of suspended sediments, biological treatment of nitrogen, and storage of carbon and phosphorus (Kadlec and Knight, 2008; McCarty and Ritchie, 2002). Wetlands are very efficient at removing nitrate. They are also considered effective in removing phosphorus associated with suspended sediment; however, they are less effective in retaining phosphorus than nitrogen. Wetlands are generally considered to be phosphorus sinks due to sediment accumulation of bound inorganic and organic phosphorus.

The effectiveness of wetlands depends on several factors including: climate; temperature; hydraulic loading and residence time; wetland to watershed area ratio; type and density of vegetation; the characteristics of the microbial communities; and the distribution and characteristics (concentration) of the influent water. Optimal reduction in nutrients occurs when hydraulic residence time is long and the rate of hydraulic loading is low. One disadvantage of wetland restoration is that the wetlands often disrupt the continuity of farming practices because of their location, thus the targeted positioning of constructed wetlands to treat agricultural drainage is promising.

Bioreactors

Recently, researchers have evaluated the use of organic fibers, in particular corn cobs and woodchips, etc., for the remediation of nitrate from subsurface drainage (Christianson et al., 2012; Jaynes et al., 2008). Bioreactors are excavated areas (beds) filled with organic fibers through which subsurface drainage waters are directed for treatment. There have been several different designs for denitrifying bioreactors including, in-field denitrification walls (Jaynes et al., 2008), edge-of-field bioreactors (Woli et al., 2010) and stream bed bioreactors (Robertson and Merkley, 2009). With sufficient hydraulic retention time in the bioreactor, denitrification is enhanced and nitrate loads are reduced in bioreactor effluent. The organic fibers act as a carbon source for microorganisms to convert NO_3^- to N_2O and N_2 gases through denitrification. The state of bioreactor science has been well summarized by Christianson and Schipper (2016).

Saturated buffers

Saturated buffers are a conservation practice that diverts subsurface tile drainage into riparian soils through a network of distribution pipes that effectively slow the discharge of agricultural drainage water to waterways while simultaneously extracting nutrients. Subsurface drainage is diverted by a water level control structure to a series of perforated drainpipes that run parallel to a ditch or stream. Drainage water seeps through the soil beneath the riparian buffer, artificially raising the water table between the drainpipes and the ditch. Recent research concluded that when proper site conditions and design considerations are met, saturated buffers can be an effective method for reducing nitrate transport from subsurface drainage systems (Utt et al., 2015).

Drainage water recycling

Drainage water recycling comprises a water storage reservoir linked to an in-field subsurface drainage system. Excess water is drained from the field and is subsequently stored in the reservoir to be recycled as irrigation water, either through surface or subsurface irrigation systems. This practice is most effective in fields with relatively dense drainage (i.e., narrow drain spacing). Benefits of drainage water recycling include the diversion of nutrients (e.g., nitrogen and phosphorus), and other contaminants (e.g., sediment) away from downstream waters (Frankenberger et al., 2017). In addition, reservoir storage can potentially buffer peak flows. Drainage water recycling has been shown to increase crop yields (12–19%), and with even greater increases (25–29%) under moderate drought conditions (Frankenberger et al., 2017; Baker et al., 2012). Reservoir construction, however, results in a loss of farm revenue because of taking land out of production but improved yields on the remaining acreage may mitigate those losses.

Water storage

Increasing or restoring water storage in the landscape can help reduce nutrient export from watersheds (Wilson et al., 2019). In addition, water storage practices can be implemented to achieve reductions in flow and associated channel erosion. Water storage practices are those that have the ability to mitigate water flow in terrestrial and aquatic environments (Rhees et al., 2022). Examples of potential water storage practices used in the USA are listed in Table 1. Water storage practices can impact water transport through retention or detention, or by affecting components of the hydrologic cycle (e.g., evapotranspiration). Before adoption and establishment of water storage practice, the following key factors should be considered including site suitability, water storage volume, co-benefits (e.g. multiple ecosystem services), limitations (e.g., legal, physical, regulatory, financial, or social constraints), and cost.

Conclusion and outlook

Agricultural drainage has enabled vast areas of marginally productive soils to become highly productive. Adequate surface drainage together with good subsurface drainage can effectively protect crops from excessive soil water conditions. Improving drainage of poorly drained soils improves trafficability and eliminates periods of excess water during the growing season that may result in delayed planting and harvest, and impair crop growth and limit yield potential.

We highlighted management practices and common conservation practices that farmers, land managers, and conservation professionals currently use to decrease agriculture's impact on water quality.

Controlled drainage is a practice that has long been recognized as an effective best management practice for mitigating the effects of agricultural drainage on water quality, especially related to nitrogen. In addition, benefits of water conservation and of higher

Table 1 Examples of water storage practices for reducing flow and nutrient export.

Practice	Description
Pond (NRCS Practice Code 378)	Used to reduce peak flows, prevent erosion, and treat runoff from agricultural and urban lands.
Controlled Drainage (NRCS Practice Code 554)	Uses a control structure to temporarily detain and store soil water in fields with tile drainage systems.
Wetlands (NRCS Practice Codes 656 and 657)	Used to store water through retention and evapotranspiration.
Off-channel impoundments (NRCS Practice Code 362)	Used to divert water from a water course to a storage area using levees or embankments.
Drainage Water Recycling (NRCS Practice Codes 436 and 443)	Used to retain surface and subsurface drainage water in a reservoir to be recycled as irrigation water, either through a surface or subsurface irrigation system.
Ditch and Culvert Improvements	Used to make structural modifications that allow management (i.e., slow) the flow of water.
Riparian Restoration (NRCS Practice Codes 390 and 391)	Used to restore ecological and hydrological conditions to a natural regime adjacent to a stream channel.
Floodplain Connection	Used to reconnect a modified or disconnected channel with its existing floodplain to allow for inundation during higher flow conditions.

NRCS: Natural Resources Conservation Service.

Adapted from Rhees S, Van Offelen H, Wall D (2022). *Water Storage: A Planning and Decision Support Framework*. https://bwsr.state.mn.us/sites/default/files/2022-02/WaterStorage%20Feb2022FinalDraft_1.pdf. Accessed 14 February, 2023.

crop yields may be attainable when controlled drainage systems are properly managed. Active management of controlled drainage structures is critical for achieving higher yields.

There are emerging issues related to the effects of climate change on precipitation and the emphasis on soil health, sustainability, and resilience. Future attention needs to be directed toward the adoption and implementation of technology (e.g. wireless communication, sensors, artificial intelligence, machine learning, and informatics) for managing, modeling, and monitoring drainage systems focused on productivity, economics, and environmental quality.

References

- Baker JM, Griffis TJ, and Oshner TE (2012) Coupling landscape water storage and supplemental irrigation to increase productivity and improve environmental stewardship in the U.S. Midwest. *Water Resources Research* 48. <https://doi.org/10.1029/2011WR011780>.
- Baulch HM, Elliott JA, Cordeiro MR, Flaten DN, Lobb DA, and Wilson HF (2019) Soil and water management: Opportunities to mitigate nutrient losses to surface waters in the northern Great Plains. *Environmental Reviews* 27: 447–477.
- Caissie D (2006) The thermal regime of rivers: A review. *Freshwater Biology* 51: 1389–1406.
- Christianson LE and Schipper LA (2016) Moving denitrifying bioreactors beyond proof of concept: Introduction to the special section. *Journal of Environmental Quality* 45(3): 757–761.
- Christianson L, Helmers M, and Bhandari A (2012) A practice-oriented review of woodchip bioreactors for subsurface agricultural drainage. *Applied Engineering in Agriculture* 28(6): 861–874.
- Dabney SM, Moore MT, and Locke MA (2006) Integrated management of in-field, edge-of-field, and after-field buffers. *Journal of the American Water Resources Association* 42: 15–24.
- Dinnes DL, Karlen DL, Jaynes DB, Kaspar TC, Hatfield JL, Colvin TS, and Cambardella CA (2002) Nitrogen management strategies to reduce nitrate leaching in tile-drained Midwestern soils. *Agronomy Journal* 94(1): 153–171.
- Dodds WK (2006) Nutrients in the “dead zone”: The link between nutrient ratios and dissolved oxygen in the northern Gulf of Mexico. *Frontiers in Ecology and the Environment* 4: 211–217.
- Framji KK, Garg BC, and Kaushish SP (1987) *Design Practices for Covered Drains in an Agricultural Land Drainage System—A World-Wide Survey*. New Delhi, India: International Commission on Irrigation and Drainage.
- Frankenberger JB, Reinhart L, Bowling C, Hay M, Youssef J, Strock X, Jia M, and Helmers BA (2017) *Questions and Answers about Drainage Water Recycling for the Midwest*. Transforming Drainage Project.
- García-Palacios P, Gross N, Gaitán J, and Maestre FT (2018) Climate mediates the biodiversity–ecosystem stability relationship globally. *Proceedings of the National Academy of Sciences* 115: 8400–8405.
- Gentry LE, David MB, Royer TV, Mitchel CA, and Starks KM (2007) Phosphorus transport pathways to stream in tile-drained agricultural watersheds. *Journal of Environmental Quality* 36: 408–415.
- Helmers MJ, Isenhardt TM, Dosskey M, Dabney SM, and Strock JS (2008) Chapter 4. Buffers and vegetative filter strips. In: *Final Report: Gulf Hypoxia and Local Water Quality Concerns Workshop*. ASABE, St. Joseph, MI 43–58.
- Helmers MJ, Abendroth L, Reinhart B, Chighladze G, Pease L, Bowling L, Youssef M, Ghane E, Ahiablame L, Brown L, Fausey N, Frankenberger J, Jaynes D, King K, Kladvivko E, Nelson K, and Strock J (2022) Impact of controlled drainage on subsurface drain flow and nitrate load: A synthesis of studies across the U.S. Midwest and Southeast. *Agricultural Water Management* 259: 107265.
- Holopainen R, Lehtiniemi M, Meier HE, Albertsson J, Gorokhova E, Kotta J, and Viitasalo M (2016) Impacts of changing climate on the non-indigenous invertebrates in the northern Baltic Sea by end of the twenty-first century. *Biological Invasions* 18: 3015–3032. <https://doi.org/10.1007/s10530-016-1197-z>.
- IPCC (2013) Climate change. 2013: The physical basis. In: Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, ... Midgley PM (eds.) *Contribution of Working Group 1 to the Fifth Assessment Report of the IPCC*. Cambridge, UK: Cambridge University Press.
- IPCC (2019) *Special Report on the Ocean and Cryosphere in a Changing Climate*. 1–1203UN.
- Jaynes DB, Kaspar TC, Moorman TB, and Parkin TB (2008) In situ bioreactors and deep drain-pipe installation to reduce nitrate losses in artificially drained fields. *Journal of Environmental Quality* 37: 429–436.
- Kadlec RH and Knight RL (2008) *Treatment Wetlands*. CRC Press.
- King KW, Williams MR, Macrae ML, Fausey NR, Frankenberger J, Smith DR, Kleinman PJA, and Brown LC (2015) Phosphorus transport in agricultural subsurface drainage: A review. *Journal of Environmental Quality* 44: 467.

- Kumar S, Lawrence DM, Dirmeyer PA, and Sheffield J (2014) Less reliable water availability in the 21st century climate projections. *Earth's Future* 2: 152–160. <https://doi.org/10.1002/2013EF000159>.
- McCarty GW and Ritchie JC (2002) Impact of soil movement on carbon sequestration in agricultural ecosystems. *Environmental Pollution* 116: 423–430.
- Meerhoff M, Audet J, Davidson TA, De Meester L, Hilt S, Kosten S, Liu Z, Mazzeo N, Paerl H, Scheffer M, and Jeppesen E (2022) Feedbacks between climate change and eutrophication: revisiting the allied attack concept and how to strike back. *Inland Waters* 1–42. <https://doi.org/10.1080/20442041.2022.2029317>.
- Noland R, Wells MS, and Peterson H (2018) Enhancing cover crop nitrogen uptake with improved establishment. *Better Crops* 102: 31–34.
- Parnesan C and Yohe G (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature* 421: 37–42.
- Patz JA, Campbell-Lendrum D, Holloway T, and Foley JA (2005) Impact of regional climate change on human health. *Nature* 438: 310–317.
- Pimm SL, Donohue I, Montoya JM, and Loreau M (2019) Measuring resilience is essential to understand it. *Nature Sustainability* 2: 895–897.
- Rabalais NN and Turner RE (2019) Gulf of Mexico hypoxia: Past, present, and future. *Limnology and Oceanography Bulletin* 28: 117–124.
- Randall GW and Goss MJ (2008) Nitrate losses to surface water through subsurface, tile drainage. In: *Nitrogen in the Environment*, pp. 145–175. Academic Press.
- Rhees S, Van Offelen H, and Wall D (2022) Water Storage: A Planning and Decision Support Framework. https://bwsr.state.mn.us/sites/default/files/2022-02/WaterStorage%20Feb2022FinalDraft_1.pdf. Accessed 14 February, 2023.
- Robertson WD and Merkley LC (2009) In-stream bioreactor for agricultural nitrate treatment. *Journal of Environmental Quality* 38: 230–237.
- Sands GR (2003) Drainage coefficient. In: Trimble SW, Stewart BA, and Howell TA (eds.) *Encyclopedia of Water Science*, pp. 118–120. NY, NY: Marcel Dekker.
- Skaggs RW, Youssef MA, Gilliam JW, and Evans RO (2010) Effect of controlled drainage on water and nitrogen balances in drained land. *Transactions of the ASABE* 53: 1843–1850. <https://doi.org/10.13031/2013.35810>.
- Smedema SK, Smedema LK, and Vlotman WF (2004) *Modern Land Drainage: Planning, Design and Management of Agricultural Drainage Systems*. London, U.K.: Taylor and Francis Group.
- Strock JS (2018) Drainage requirements to maintain soil health. In: Reicosky D (ed.) *Managing Soil Health for Sustainable Agriculture*. vol. 2, pp. 147–168. Burleigh Dodds Science Publishing.
- Strock JS, Porter PM, and Russell MP (2004) Cover cropping to reduce nitrate loss through subsurface drainage in the northern U.S. corn belt. *Journal of Environmental Quality* 33: 1010–1016.
- Strock JS, Kleinman PJA, King KW, and Delgado JA (2010) Drainage water management for water quality protection. *Journal of Soil and Water Conservation* 65: 131A–136A.
- Strock JS, Sands GR, and Helmers MJ (2011) Subsurface drainage design and management to meet agronomic and environmental goals. In: Hatfield JL and Sauer TJ (eds.) *Soil Management: Building a Stable Base for Agriculture*, pp. 199–208. Madison, WI: ASA-SSSA.
- Utt N, Jaynes D, and Albertson J (2015) *Demonstrate and Evaluate Saturated Buffers at Field Scale to Reduce Nitrates and Phosphorus from Subsurface Field Drainage Systems*. Agricultural Drainage Management Coalition.
- van der Ploeg RR, Kirkham MB, and Marquardt M (1999) The Colding equation for soil drainage: Its origin, evolution, and use. *Soil Science Society of America Journal* 63: 33–39.
- van Vliet MT, Franssen WH, Yearsley JR, Ludwig F, Haddeland I, Lettenmaier DP, and Kabat P (2013) Global river discharge and water temperature under climate change. *Global Environmental Change* 23: 450–464.
- Wagena MB and Easton ZM (2018) Agricultural conservation practices can help mitigate the impact of climate change. *Science of the Total Environment* 635: 132–143.
- Wang ZH and Li SX (2019) Nitrate N loss by leaching and surface runoff in agricultural land: A global issue (a review). *Advances in Agronomy* 156: 159–217.
- Wheeler T and Von Braun J (2013) Climate change impacts on global food security. *Science* 341: 508–513.
- Wilson HF, Casson NJ, Glenn AJ, Badiou P, and Boychuk L (2019) Landscape controls on nutrient export during snowmelt and an extreme rainfall runoff event in northern agricultural watersheds. *Journal of Environmental Quality* 48: 841–849. <https://doi.org/10.2134/jeq2018.07.0278>.
- Woli KP, David MB, Cooke RA, McIsaac GF, and Mitchell CA (2010) Nitrogen balance in and export from agricultural fields associated with controlled drainage systems and denitrifying bioreactors. *Ecological Engineering* 36(11): 1558–1566.
- Xia Y, Zhang M, Tsang DC, Geng N, Lu D, Zhu L, et al. (2020) Recent advances in control technologies for non-point source pollution with nitrogen and phosphorous from agricultural runoff: Current practices and future prospects. *Applied Biological Chemistry* 63: 1–13.
- Youssef MA, Abdelbaki AM, Negm LM, Skaggs RW, Thorp KR, and Jaynes DB (2018) DRAINMOD-simulated performance of controlled drainage across the U.S. Midwest. *Agricultural Water Management* 197: 54–66.
- Youssef M, Strock J, Bagheri E, Reinhardt BD, Abendroth L, Chighladze G, et al. (2022) Impacts of controlled drainage on corn yield under varying precipitation patterns: A synthesis of studies across the U.S. Midwest and Southeast. *Agricultural Water Management* 275: 107993. <https://doi.org/10.1016/j.agwat.2022.107993>.

Relevant Website

International Commission on Irrigation and Drainage (2017) World drained area. Available at http://www.icid.org/database_icid.html. Accessed 9 December 2022.

Environmental effects of irrigation

Sevilay Topcu, University of Cukurova, Faculty of Agriculture, Department of Agricultural Structures and Irrigation, Adana, Turkey

© 2023 Elsevier Ltd. All rights reserved.

This is an update of S. Topcu, C. Kirda, Irrigation: Environmental Effects[☆], Reference Module in Earth Systems and Environmental Sciences, Elsevier, 2013, ISBN 9780124095489, <https://doi.org/10.1016/B978-0-12-409548-9.05194-0>.

Introduction	266
Environmental effects of irrigated agriculture	268
Water shortage	268
Waterlogging and salinity	270
Pollutants attributable to irrigation	271
Plant nutrients (fertilizers)	271
Pesticides	272
Metals	273
Public health concerns	273
Sustainability	274
Conclusions and outlook	274
References	274
Further readings	275

Abstract

The chapter describes the importance of irrigated agriculture, the major water consumer, and reviews the effects of irrigation on the availability as well as the quality of land and water resources. Moreover, it underlines the problems related to current irrigation practices, such as pollutants attributable to irrigation, public health concerns as well as socio-economic and policy issues. Finally, it outlines options for improving the environmental performance of irrigation while increasing water productivity in irrigated agriculture. These are complementary goals which are also important for the long-term sustainability of irrigated agriculture.

Key points

- The earliest archeological evidence of irrigation in agriculture dates back to about 6000 BCE.
- Crop yields under irrigation can be increased by 100 to 500% compared with those under rainfed conditions.
- Globally, irrigated agriculture uses the highest portion (72%) of the world's water resources, with a low water use efficiency and productivity.
- Overexploitation of water resources and excessive water usage in irrigation cause water shortages, changes in flow regimes of rivers, rising water tables, triggers drainage and salinity problems.
- Excess water application and particularly surface irrigation methods promote runoff and deep percolation which carry pesticides and fertilizers together with microorganisms and pathogens into rivers, lakes and aquifers.
- Inappropriate wastewater and irrigation practices can have harmful health effects including cholera, typhoid, ascariasis, amebiasis, giardiasis, and entero-invasive *Escherichia coli*.

Introduction

The delta plains of the Euphrates and the Tigris rivers in Mesopotamia, and the Nile in Egypt have witnessed numerous civilizations. Archeological remains of irrigation structures such as barrages, channels, and deep wells date back to as early as 4000–6000 BCE. The real advancement in irrigated agriculture, however, started in the twentieth century with the development of large irrigation projects in Egypt, Pakistan, India, and the USA. Irrigation contributed significantly to the large crop yields achieved following the Green Revolution in the 1960s, which is commonly attributed to the breeding of high-yielding crop cultivars, and use of fertilizers and pesticides.

Irrigation development projects facilitate the improvement of infrastructure such as: roads, telecommunications and energy networks; improvement of public services in health care and education; promotion of investments in agribusiness and related areas; and creation of new employment opportunities. While country-wide prices of agricultural produce tend to decrease under irrigated agriculture, the living standards of the whole population tend to improve; hence the above improvements may lead to substantial

socio-economic development. Irrigated agriculture was also key to alleviating poverty in Asian countries, and its continuing expansion has been considered essential in meeting the needs of increasing populations in the developing world.

The replenishment of water in the plant root zone during periods when natural rainfall is insufficient for good plant growth is the classic definition of irrigation. Irrigation is essential to ensure high crop yields in arid and semiarid regions. Supplementary irrigation in humid and sub-humid regions not only guarantees high crop yields, but also improves the quality of those crops. It is well documented that crop yields under irrigation increase by 100 to 500% compared with yields obtained under rainfed conditions in arid regions as well as globally (e.g. Jaramillo et al., 2020). Irrigated areas increased from a mere 63 Mha in 1900 to 111 Mha in 1950 and to 339 Mha in 2018. Fig. 1 shows the expansion of irrigated areas since 1960 and compares it with the global increases in population and cereal production. Prior to 1950, the areas equipped for irrigation gradually increased (Siebert et al., 2015) and the rate of expansion was 2.2% annually until the late 1970s. The rate of expansion then leveled off to around 1% per year in the 2000s. Although irrigated agriculture is practiced on nearly all continents, 70% of the globally irrigated areas are in Asia, particularly in Pakistan, China and India. For example, an area of about 16 Mha produces more than 90% of the entire agricultural and livestock production in Pakistan (see Sobia et al., 2018 in Qureshi, 2020).

Production of major broad acre crops (wheat (*Triticum aestivum* L.), maize (*Zea mays* L.), rice (*Oryza sativa* L.) and soybean (*Glycine max* (L.) Merrill) shows a strong association with the increasing acreage of irrigated areas (Fig. 1). Although only approximately 25% of the world's arable land is under irrigation, it accounts for more than 60% of the cereal production worldwide and 50% of the sales value of all crops harvested come from irrigated agriculture (Angelakis et al., 2020). Although around 30% only of the world's cereal-planted area is under irrigation, 42% of global cereal production originates from these areas. Eliminating irrigation would reduce cereal production by 20% (see Siebert and Döll, 2010 in Siebert et al., 2015). Based on a 'business-as-usual' scenario, the Food and Agriculture Organization of the United Nations (FAO) estimates that the world will need about 60% more food by 2050, and that irrigated food production will increase by more than 50% over the same period (FAO, 2017). The necessary amount of water, however, is not available: the FAO recognizes that the amount of water withdrawn by agriculture can only realistically increase by 10% (UN, 2021).

In the future, the largest expansion of irrigated areas will take place in Eastern and Near-East countries and in North Africa where farmland is limited and irrigation is the only means of increasing agricultural production. However, for Sub-Saharan Africa and particularly for the developed countries only a marginal expansion of the irrigated area is projected until 2050. An estimated three-quarters of the farmland in East and Near-East countries and North Africa will have irrigation projects implemented in the near future. Similarly, irrigated agriculture will extend to more than 90% of cultivable land in South Asia. However the projected net increase, with an annual growth rate of 0.24% in irrigated land, is much lower than that achieved during the last 50 years (FAO, 2011). The major constraints to the expansion of irrigated areas are the scarcity of land and water resources, increased demand for municipal and industrial water allocations that are in competition with agriculture as well as the rising costs of irrigation project investments.

The expansion of irrigated areas is unsustainable, however, without proper drainage and funding for the appropriate implementation of technical and socioeconomic aspects. History has witnessed the disappearance of numerous civilizations not only because of wars and long spells of drought or floods, but also as a result of the unintended consequences of irrigated agriculture. The Babylonian and Sumerian empires in Mesopotamia are examples of past civilizations whose collapse, in part, was caused by poorly practiced irrigation and drainage.

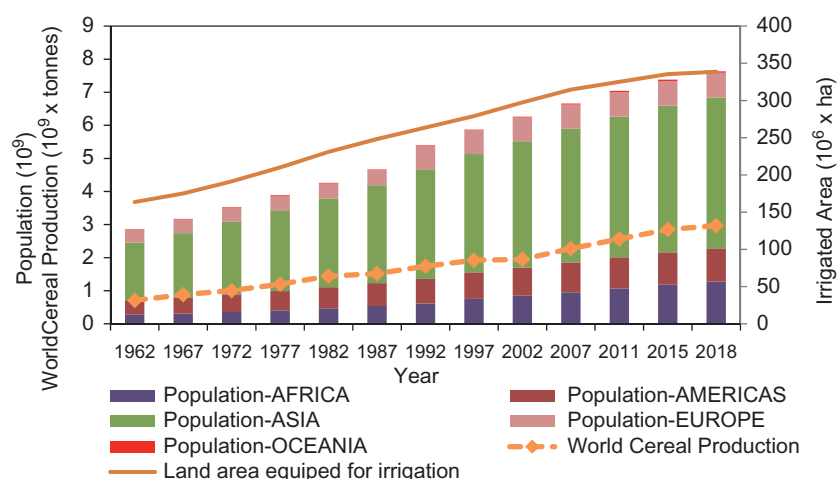


Fig. 1 Population, irrigated areas and major broad acre crops (maize, wheat, rice and soybean) production (1960–2018). Compiled from on-line databases of the U.N. Food and Agriculture Organization (FAO-FAOSTAT, <http://www.fao.org/faostat/en/#data>).

Environmental effects of irrigated agriculture

Irrigation development has several positive socio-economic as well as ecological effects such as the creation of employment opportunities, rural development and poverty alleviation, transformation of drylands into green areas that produce large amounts of oxygen and fix carbon dioxide and accommodate, through fields, ditches, canals and reservoirs, the habitats where a wide range of wildlife thrives, as opposed to deserts and undeveloped natural arid lands (Angelakis et al., 2020). A comparative study shows that irrigation generally has a positive effect on soil fertility under arid conditions and a negative effect in semi-arid, semi-humid and humid tropical zones (Nikolskii et al., 2019).

Some of these effects can also be considered negative from the ecological point of view, since they alter or even destroy natural ecosystems in arid and semiarid regions. For example irrigation projects, that transform large areas of dry cereal agro-ecosystems into irrigated crop systems dominated by maize, erode plant diversity in Spain (Fagúndez et al., 2016). Irrigation causes far-reaching land transformation, which is often permanent and irreversible and takes place mainly in two ways: (a) through direct modifications of the land surface, which occur when water conveyance and distribution networks are constructed, and lands are cleared, shaped and leveled for farming and irrigation; and (b) through in-depth transformations that take place when the water and salt balance in the region are changed as a consequence of importing large quantities of water and salts into the area (see Squatriti, 2002 in Angelakis et al., 2020).

Dam construction and, more importantly, water withdrawals from groundwater and surface water have altered not only freshwater flow dynamics (see Döll et al., 2009 in Siebert et al., 2015) but also variations in water storage in surface water bodies and aquifers (see Döll et al., 2012 in Siebert et al., 2015). Overexploitation of groundwater, changing flow regimes of rivers, and rising water tables are some of the environmental issues to be addressed. Excessive water usage triggers drainage and salinity problems, which may cause irreversible loss of soil fertility. Erosion and sedimentation, as well as salinization, have gone 'hand-in-hand' with the deterioration of farmland and the eventual collapse of some social structures. At the same time, salts, organic matter, solids of several sizes, fertilizers and pesticides may leach from the soil and run off into rivers, streams or aquifers, together with microorganisms and several pathogens. Furthermore, large bodies of water stored in dams and diversion channels can cause significant changes in local climate, while the extent of wetlands that have important ecological functions of biodiversity, nutrient retention and flood control can be reduced.

Water shortage

Global water use has increased six fold over the past 100 years and continues to grow steadily at a rate of about 1% per year as a result of increasing population, economic development and shifting consumption patterns (UNESCO, 2020). The demand for fresh water increased fourfold during the period 1950–1960 only, compared to the previous decades, due to the implementation of new irrigation projects and the increase in demand for municipal and industrial water. Global water demand is projected to increase by 20 to 30% by 2050 (see UNESCO, 2018 in UNESCO, 2020). Approximately 3.2 billion people live in water-scarce agricultural areas, of whom, 1.2 billion live in severely water-constrained agricultural areas. Population growth is a key driver of water scarcity as it implies rising demand for this precious natural resource. Consequently, the annual amount of available freshwater resources per person has declined by more than 20% in the past two decades. More than 60% of irrigated cropland (171 Mha) is highly water-stressed (FAO, 2020).

Globally, irrigated agriculture uses the highest portion (72%) of the world's water resources (Fig. 2) (UN-Water, 2021), and accounts for over 90% in arid countries (FAO, 2018). Groundwater irrigation plays a vital role in boosting agricultural production and livelihoods of rural communities in many parts of the world. For example, the area in South Asia equipped for irrigation has tripled since 1950 (Siebert et al., 2015). India, Pakistan, and Bangladesh are the largest groundwater users in South Asia, and more

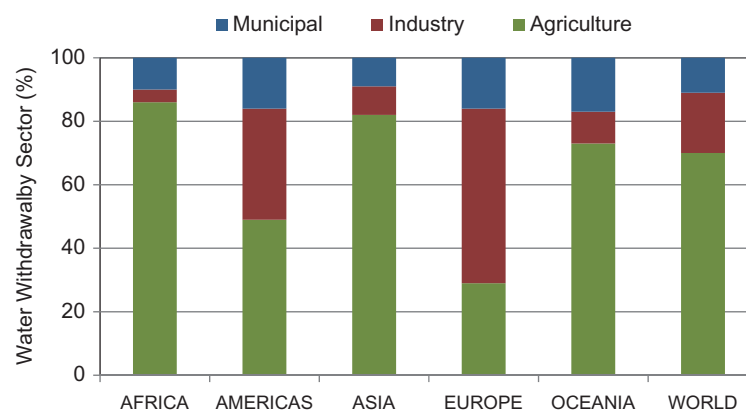


Fig. 2 Percentage of water withdrawal by sector and continent. FAO-AQUASTAT database (http://www.fao.org/nr/water/aquastat/water_use/index.stm).

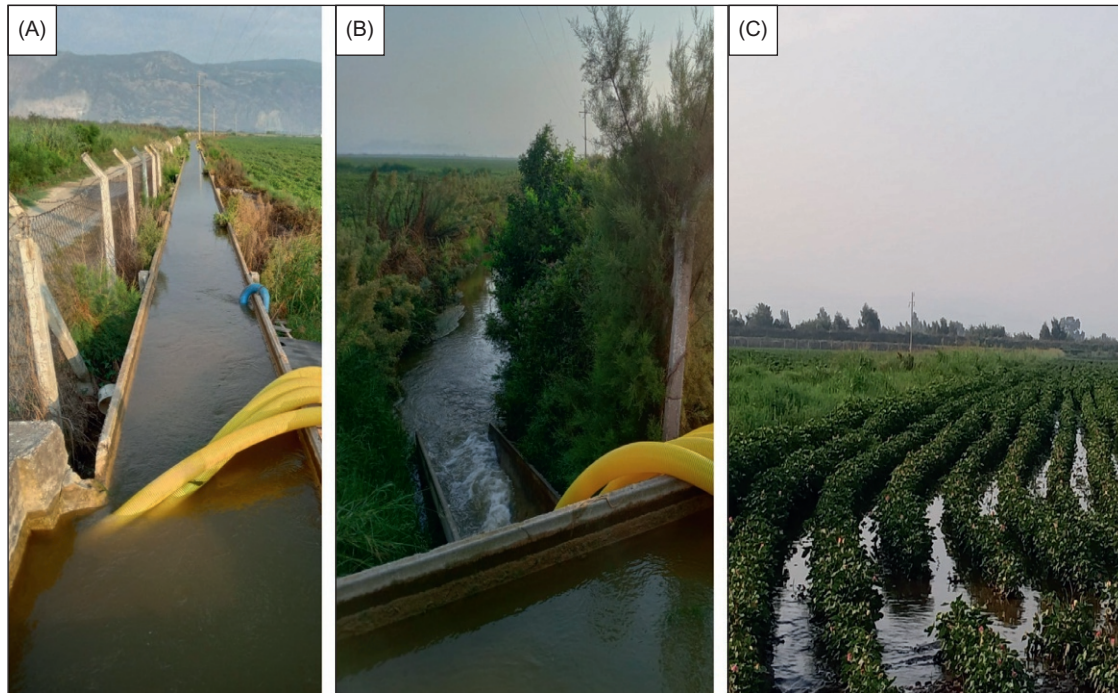


Fig. 3 Water distribution by open concrete canals throughout the fields (A), water withdrawal from those canals to the field-head ditches (B) and water application by flooding irrigation method (C).

than 85% is used for agricultural purposes, compared to 40% in the rest of the World (Siebert et al., 2015; Chindarkar and Grafton, 2019). These three countries irrigate 48 Mha using groundwater, approximately 42% of the global groundwater-fed cropland (see Mukherjee et al., 2014 in Qureshi, 2020). An estimated 35.9 Mha of agricultural lands are subject to the indirect use of urban wastewater (Thebo et al., 2017).

Water consumption by crops is only a small portion of the total water diverted to irrigated agriculture because of losses in conveyance networks, poor water application efficiency, excess water application due to inappropriate water management, policies and unrealistically low water pricing. The delivery efficiency of the Indus irrigation system (consisting of perennial rivers, a network of unlined canals, and distribution channels) is only 36%. Similar surveys from Europe indicate that in open canal distribution networks, water losses are estimated at up to 40% in unlined ditches and up to 25% in lined canals. The efficiency of irrigation application ranged from 45% to a maximum of 60% in areas where water was distributed via open canals. Greater irrigation efficiencies have been achieved in pipe systems (e.g. Phocaides, 2007). The water use efficiency (WUE) in irrigated agriculture increased up to 44% globally, varying between the regions and countries: low in Southeast Asia (19%) and sub-Saharan Africa (30%) and high in Northern Africa (69%) and South Asia (55%) in the early 2000s compared to late 1990s (FAO, 2011). Despite technical advances, WUE in Pakistan ranges between 35% and 40% (see Khan et al., 2008 in Qureshi, 2020) which is a result of the application of excess irrigation water by the flood irrigation method (Qureshi, 2020). The Fig. 3 shows examples of water distribution by open concrete canals throughout the fields, water withdrawal from those canals to the field-head ditches and water application by the flood irrigation method.

The global average 'economic water productivity', which is defined as the value added per volume of water withdrawn in all water-using sectors, has increased 4% globally since 2015 and is around USD 15 m⁻³. Significant differences exist among countries and regions, the lowest regional economic water productivities are USD 2 m⁻³ in Central and Southern Asia, around USD 7 m⁻³ in sub-Saharan Africa where the irrigated areas have more than doubled since 1960. The highest values are USD 50 m⁻³ in Oceania and USD 38 m⁻³ in Europe and North America (UN-Water, 2021; FAO, 2018).

The source of the water withdrawn (groundwater or surface water) affects water storage in surface water bodies and aquifers in different ways. For example, in regions with extensive surface water withdrawals for irrigation, such as Southern China, net abstractions from groundwater are negative, i.e. groundwater is recharged by irrigation. The opposite occurs in areas dominated by groundwater withdrawals for irrigation, such as in the High Plains aquifer of the central USA, where net abstraction of surface water is negative because return flow of withdrawn groundwater recharges the surface water compartments (see Döll et al., 2012 in Siebert et al., 2015).

With overexploitation of groundwater resources for irrigation in many countries, aquifer depletion has been observed for many semi-arid and arid regions. More than one million deep wells are opened annually in the Indian peninsula and Pakistan. In the Baluchistan province of Pakistan, groundwater tables are declining at a rate of 2 to 3 m annually, and without effective management measures about 20% of the agricultural areas may lose access to groundwater by 2025 (see Tariq et al., 2020 in Qureshi, 2020). The

water table depth in Northern China increased from 8 to 50 m between 1967 and 2000. Similarly, it was reported that water levels of 10 aquifers in Mexico decreased by 1.8–3.3 m annually. In the USA and Mexico, around 20% of irrigation water comes from non-renewable groundwater. In Iran and Saudi Arabia, where rainfall and surface freshwater are extremely scarce, non-renewable groundwater provides the largest contribution to irrigation water, 40 and 77% respectively (Wada, 2012). In India, cropping intensity is expected to decrease by 20% nationwide and by 68% in groundwater depleted regions in 2025 (Jain et al., 2021).

Overexploitation of aquifers and the use of groundwater from deeper aquifers for irrigation can cause secondary salination in agricultural lands (see Greene et al., 2016 in Mastrocicco and Colombani, 2021) since groundwater salinity and some natural (geogenic) hazardous elements tend to increase with increasing depth (van Weert et al., 2009). The contamination due to geogenic sources such as arsenic, fluoride, and iron as well as salinity problems in coastal and inland areas are the major land and water quality issues in many parts of the world e.g. in India (Garduño et al., 2011), the Indus Basin in Pakistan and the Murray-Darling Basin in Australia (see Greene et al., 2016 in Mastrocicco and Colombani, 2021) and Mediterranean countries (Mastrocicco and Colombani, 2021).

Furthermore, overexploitation of deep wells close to coastal areas triggers seawater intrusion and thereby causes degradation of water quality, which in turn endangers the sustainability of irrigated agriculture. As a result of groundwater overexploitation, land subsidence triggered by aquifer compaction can be observed, resulting in high socio-economic impacts for the affected communities such as in the Konya closed Basin in the Central Anatolian region of Turkey (Caló et al., 2017). Mexico City has experienced land subsidence up to 30 cm year⁻¹ due to massive GW extraction for municipal and agricultural water use, resulting in decreasing water resources, damaging infrastructure, and increasing flood risk and water contamination (see Chaussard et al., 2014 in Caló et al., 2017).

Irrigation withdrawals have also contributed to the shrinkage of vast lakes: Lake Chapala in Mexico lost 80% of its volume between 1979 and 2001, and the Aral Sea almost disappeared by the end of the 20th century as irrigation withdrawals for cotton (*Gossypium hirsutum* L.) reduced inflow (FAO, 2011). Climate change may also exacerbate the situation in water-scarce regions in the world. Thus, on the one hand climate change could increase water demand for agriculture, primarily for irrigation, due to prolonged dry periods and severe drought (FAO, 2020), but on the other hand the share of water available for agriculture is expected to decline to 40% by 2050 (OECD, 2012).

Waterlogging and salinity

Salinity problems associated with irrigated agriculture have been recognized for a long time. Salinity causes several problems such as poor soil structures particularly due to sodium, poor plant growth and less yield, ecological and socio-economic problems and also food insecurity. Soil salinity that develops because of mismanaged farm irrigation systems is called secondary salinity. Excess use of irrigation water (Fig. 3) can cause a rise in the water table, and the resulting problem from waterlogging can become a major constraint to high crop yields. Soil salinity and waterlogging are interrelated and both occur due to lack of drainage and excess irrigation-water application. These problems, however, are often ignored at the planning stage of irrigation projects. This is not because of the lack of technical understanding, but mainly because of populist political considerations and commitments to extend irrigated acreage to larger areas without drainage. Drainage systems are largely underground, and therefore have the least political appeal. The value of the drainage systems is difficult to appreciate in the short term, since they function as preventive measures for soil salinization that may occur in the future.

In areas lacking properly designed drainage systems, the water table soon rises to the soil surface, continuing evaporation takes place, and salt accumulation speeds up (Fig. 4). Waterlogged soils can also collapse quickly through dispersion of clay particles, which is especially the case in dispersive or sodic soils. Furthermore, roads and building foundations can be weakened by an accumulation of salts within the natural soil structure. Waterlogging within the rootzone leads to depletion of soil oxygen and carbon dioxide accumulation, chemical conversion of non-toxic chemicals into their reduced form (which can be toxic, e.g., sulfate reduced to sulfide), denitrification and emission of nitrous oxide (N₂O), one of the major greenhouse gases, and reduction in nitrogen fixation by the nodules of legume crops (see IPBES, 2018 in UNESCO, 2020).

The global area of all salt-affected soils, most of which are naturally salty, is about 1000 Mha; they occur in about 100 countries (see IPBES, 2018 in UNESCO, 2020). Irrigated land damaged by salinization is estimated globally to be 60 Mha and the top eight contributing countries are Pakistan (3.2 Mha), China (7 Mha), USA (5.2 Mha), India (20 Mha) (Squires and Glenn, 2011), Uzbekistan, Iran, Iraq and Turkey. An additional 60–80 Mha are affected to some extent by waterlogging and related salinity (FAO, 2011). Some well-known regions where salinization occurs include the Aral Sea Basin in Central Asia, the Indo-Gangetic Basin in India, the Indus Basin in Pakistan, the Yellow River Basin in China, the Euphrates Basin in Syria and Iraq, the Murray-Darling Basin in Australia, and the San Joaquin Valley in USA (Qadir et al., 2014). In the Indus Basin 21% of the total irrigated area, which produces more than 90% of Pakistan's grain, is currently affected by salinity. In the populous Sindh province, 43% of the irrigated area is affected (Qureshi, 2020). About 10% of all river stretches in Africa and Asia have high salinity levels primarily associated with agricultural irrigation (see UNEP, 2016 in UNESCO, 2020).

The cost of salt-induced land degradation in 2013 was USD 441 per hectare; a simple benefit transfer suggests the annual economic losses could be 27 billion USD (Qadir et al., 2014). It is estimated that the areal extent of naturally occurring and human-induced salt-affected soils are increasing due to enlargement of irrigated areas and the impact of climate change.



Fig. 4 Soil salinity in Harran Plain in Turkey. Photo by M. Ali Çullu.

Pollutants attributable to irrigation

Excessive and inefficient irrigation can affect soil and water quality by causing erosion, transporting nutrients, pesticides, and 'heavy metals.'

Sediment load carried in returning waters from large-scale irrigation projects and erosion occurring in watersheds may be a source of major pollutants and result in serious environmental consequences. There are two different sources of soil erosion: (1) erosion occurring during a flood that silts up water reservoirs, (2) erosion in areas with poorly planned irrigation projects where land capability classifications are disregarded, and surface irrigation methods are practiced with no grading of the land. Sediments originating from the erosion of fertile surface soil in agricultural areas are carried to downstream areas in returning irrigation waters.

Water reservoirs, man-made small lakes, and large dams may be filled with silt from erosion before they complete their economic life, if forestation is not planned and implemented. Silt and sediments found in water reservoirs may also carry plant nutrients (that may cause algal blooms) and pesticide residues (with the risk of serious health effects). Heavy metals are also transported into water reservoirs by soil erosion.

Plant nutrients (fertilizers)

Use of fertilizers is a successful low-cost means of achieving high crop yields under irrigation. As a result fertilizer demand has increased globally, supported by global economic growth, leading to an excessive application of fertilizers: "the more, the better". According to the FAO statistics, world NPK (nitrogen, phosphorus, potassium) fertilizer consumption reached 188 million tons (MT) in 2018 (Fig. 5) and is expected to increase to 231 and 263 MT by 2030 and 2050, respectively (Alexandratos and Bruinsma, 2012) unless sustainable strategies such as improved water and fertilizer management, use of cover crops and fertilizer technology are widely practiced to increase nutrient-use efficiency and reduce nitrate leaching.

Only 30–50% of applied nitrogen fertilizer (e.g. Cassman et al., 2002) and ~45% of phosphorus fertilizer (Smil, 2000) is taken up by crops, and significant amounts of the applied nitrogen and phosphorus remain in the soil after the crops have been harvested. Excessive application of fertilizers and manure or inefficient use of the main nutrients (N and P) in fertilizers are the main contributors to environmental issues linked to agriculture; these failures also increase the production costs for the farmers. Such non-point nutrient losses pollute soil resources, harm off-site ecosystems, degrade water quality and aquatic ecosystems, and contribute to greenhouse gas emissions.

In irrigated agriculture, excessive water application increases nitrate leaching, leading to a vicious circle where low nitrogen (N) availability is compensated for by increasing fertilizer applications which result in low nitrogen use efficiency (NUE) by the

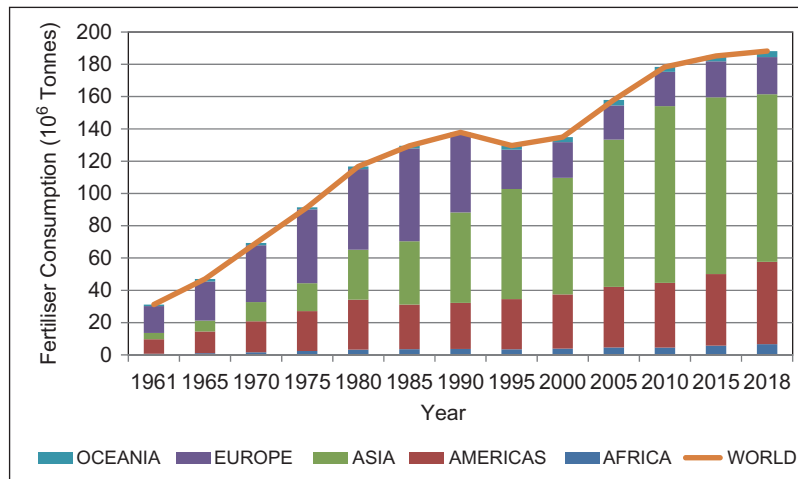


Fig. 5 Annual fertilizer consumption by continent (1961–2018). (Compiled from FAOSTAT–Agricultural Data, FAO on-line database.)

crops and a deleterious impact on groundwater. Total global leaching and run-off of nitrogen is estimated at 32.6 million tons per year, the large majority from poor agricultural practices (Mekonnen and Hoekstra, 2015).

Nitrate fertilizers have high solubility, and the nitrate ion is easily leached into groundwaters with the application of excess irrigation water because of its strong mobility. Return flows from irrigated areas with high sediment loads may also contain large nitrate concentrations. The increase in nitrate and phosphorus in lakes is the main cause of eutrophication, algal blooms, and depletion of oxygen, which in turn can lead to the mass die-off of fish and it can also reduce fattening in cattle. Various parts of the worldwide have nitrate values higher than the WHO standards (50 ppm), and nitrate pollution from both rainfed and irrigated agriculture and, specifically, from the use of fertilizers and pesticides, is a common denominator on all continents (Abascal et al., 2022).

Water and nitrogen availability remain, globally, the most limiting factors for crop growth (Mueller et al., 2012). Therefore, irrigated agriculture requires specific farm water management practices to increase water- and nitrogen-use efficiency that may differ from rainfed systems. Results of a meta-analysis showed that improved water management, for example adjusting irrigation application to crop water requirements, using deficit irrigation techniques, improved irrigation scheduling, and water saving irrigation technologies such as drip irrigation, can help to optimize this. These approaches provide the most effective strategy and treatments for reducing nitrate leaching, followed by practicing improved fertilizer management, use of cover crops and use of improved fertilizer technologies (Quemada et al., 2013). The application of fertilizers and chemicals such as pesticides to crops together with water through the drip or sprinkler system is known as fertigation or chemigation. Fertilizers and plant protection chemicals that are water soluble can be applied efficiently and directly to the plants and or plant rootzone at the required time and in the right amount through the drip irrigation system. Compared to conventional methods of fertilizer application under flood irrigation, fertigation and chemigation prevent pollution by increasing the efficiency of fertilizer and pesticide use, in addition to increasing the efficiency of irrigation application under drip irrigation (Fig. 6).

Pesticides

Evaporation from large bodies of water stored in dams, networks of channels, and large irrigated fields contributes to increasing humidity, which in turn creates a favorable environment for the spread of plant diseases and pests. This has led to the use of pesticides, one of the major inputs of irrigated agriculture, which dates back to as early as the fifteenth century when arsenic salts were used as herbicides and insecticides. Growers prefer to grow crops of high cash value under irrigated agriculture resulting in the widespread practice of continuous monoculture, which can lead to pest and disease problems.

Only 1% of total pesticides are used effectively to control insect pests on target plants (see Bernardes et al., 2015 in Tudi et al., 2021), the remaining pesticides penetrate or reach non-target plants and environmental media. Irrigation practices play a major role in the transport and then accumulation of pesticide residues encountered in water resources. Irrigating at a rate that exceeds the rate of infiltration of soil promotes runoff that can carry pesticides with it. Irrigation that promotes deep percolation of water beyond the root zone of plants also causes the leaching of agro-chemicals such as pesticides to ground water. Both types of water movement, runoff and deep percolation, are of particular concern in areas where surface irrigation methods such as flooding are practiced. In addition, the frequent irrigation of coarse-textured soils and use of surface irrigation methods on sloping lands can also increase the risk of pesticide transport and accumulation in surface and ground water resources (SDWF, 2022).

Climate change affects pest and disease outbreaks and also accelerates the pesticide behavior of volatilization, runoff, leaching processes, and pesticide degradation (Tudi et al., 2021). There are many reports about pesticide contamination of both surface water and groundwater worldwide (e.g. see Brauns et al., 2018 and Aravinna et al., 2018 in Tudi et al., 2021). For example, the United States Geological Survey (USGS) found several pesticides in more than 90% of water and fish samples collected from streams in the USA (see Rose et al., 2018 in Tudi et al., 2021).



Fig. 6 An example of a simple drip irrigation system: (A) control unit including filter and fertilizer tanks, control valves, (B) main and lateral water distribution pipes and drippers near the plant rows.

A study of over 2000 agricultural workers in Africa, Asia, and Latin America found that many workers suffered from acute pesticide poisoning and reported the following symptoms: chronic headaches, dizziness, convulsions, and nausea (PAN, 2010). About 3000 children were poisoned by pesticides in eastern China's Zhejiang province, where irrigated agriculture is dominant, between 2006 and 2015, with most cases occurring during the farming season (Fan, 2017).

Metals

Fertilizers and pesticides that are used in irrigated agriculture contain a wide array of metals such as arsenic (As), boron (B), mercury (Hg), copper (Cu), cadmium (Cd), lead (Pb), nickel (Ni), and zinc (Zn). While boron is often added to the soil with irrigation water, arsenic (see Varol and Şekerçi, 2018 in Abascal et al., 2022) is included in many insecticides and herbicides, and phosphorus fertilizers are the major source of zinc and cadmium pollution (see in Mazhar and Ahmad, 2020 in Abascal et al., 2022) in soils and groundwater. In addition, the presence of nitrates from the use of chemical fertilizers implies the co-existence of other ions in groundwater, such as F^- (fluoride), SO_4^{2-} (sulfate), and K^+ (see Roy et al., 2020 in Abascal et al., 2022). Runoff and tail waters carrying sediments from irrigated areas often contain metals that are discharged into drainage channels and then mixed with other surface water resources. These may cause serious environmental damage such as the selenium (Se) pollution reported in the San Francisco Delta area in the American West (Presser and Luoma, 2006). Prevailing intensive irrigation has caused selenium enrichment in groundwater and also its accumulation in soils in Punjab. Similarly, irrigation has been found to be a critical source of Hg (38.63%) and Ni (22.69%) in Zhejiang province in China (Ma et al., 2021).

Public health concerns

In large-scale irrigation schemes, the large bodies of water stored by dams and water-conveying open-channel systems and related local climate change may cause an increase in pests and parasites that carry diseases. Among the diseases spread by parasites and insects feeding in water sources are malaria, dengue fever and the Zika virus, schistosomiasis (bilharzia), yellow fever, encephalitis, and river blindness. Encephalitis is largely found in the rice fields of South, Southeast, and West Asia; lymphatic filariasis, across all continents; river blindness, largely in Africa and Central America; and malaria and schistosomiasis, across all continents but largely in Africa. A study on the incidence of malaria found that the large dams in major river basins of sub-Saharan Africa increased cases from 0.9 to 1.7 million annually depending on the year and the malarial burden (Kibret et al., 2021). Of the approximately 800 million people at risk of schistosomiasis worldwide, 100 million live in close proximity to dams and irrigation schemes (see Steinmann et al., 2006 in Lund et al., 2021).

Large nitrate concentrations in drinking water as a result of fertilizer use in irrigated agriculture is also a public health threat, causing the deadly health disorder methemoglobinemia, particularly in human infants.

In water-scarce regions, municipal and industrial wastewaters are often recycled and rerouted into irrigation schemes. Although treatment of such water sources is common practice in the developed world, it is rare in developing countries. Wastewater from

poultry farms, cattle ranges, and household sewage may pose a serious health hazard if used for the irrigation of freshly consumed vegetables. Among the diseases spread by the use of untreated wastewater for irrigation are cholera, typhoid, ascariasis, amebiasis, giardiasis, and entero-invasive *Escherichia coli* (e.g. [Blumenthal and Peasey, 2002](#)). In addition to regulatory mechanisms and controls of water quality used for irrigating crops that are eaten without cooking (e.g. leafy salad crops), it is suggested that controlling the water source, regime, and timing of irrigation may help to reduce harmful contamination ([Banach and van der Fels-Klerx, 2020](#)).

Use of untreated wastewater for irrigation causes the transfer of Pb, Cd, and Ni from the roots to the edible parts of vegetables, posing a long-term health risk for example in India (Sharma et al., 2007 in [Ahmed et al., 2018](#)) and in Bangladesh (see Tasrina et al., 2015 in [Ahmed et al., 2018](#)). Metals related to wastewater irrigation, and intensive fertilizer and pesticide applications, both in irrigated and rainfed agriculture, affect human health. The health problems depend on the ingested dose of such metals; they range from excessive shyness to simple stomach ailments, nausea, vomiting and diarrhea, to blood, lung, kidney complications or diseases, gastrointestinal cancer and even death. In particular young children and unborn fetuses are highly sensitive (e.g. [Andrews and Henrique, 2006](#)).

Sustainability

Sustainable irrigation development has to be capable of delivering a constant and reliable supply of water of sufficient quality for irrigation to support a constant or increasing level of food production, while keeping its impact on the environment within sustainable levels. Implementation of integrated approaches for land and water resources management comprising integrated plant nutrient and pest management and crop–livestock systems can promote the sustainability of irrigated agriculture.

The history of agriculture has shown that sustainable irrigated agriculture requires adequate salt balance and drainage, in addition to water balance for sustainable soil and water resources management. The technology to accomplish this exists, but its use is limited by: lack of economic incentives for growers to improve their existing irrigation structures; need for investment in pressurized irrigation systems and drainage works; lack of training in and experience with effective management practices; public opposition to the possible effects of irrigation return flows on the environment; and administrative constraints such as low budget allocations for timely maintenance and management of irrigation schemes.

Conclusions and outlook

Irrigation enables agricultural production in arid regions where precipitation is insufficient to meet plant water requirements. In humid and sub-humid regions, applied irrigation water supplements available soil moisture and facilitates high crop yields and quality by providing a critical buffer against periodic droughts. Crop yields under irrigation can be increased by up to 500% compared to rainfed agriculture. This makes irrigation a vital part of the agricultural economy in many countries.

However, increasing competition between water using sectors, overextraction of groundwater reserves, decreases in quantity and quality of surface water resources, shifts in climatic conditions are constraining the ability of hydrological systems to meet irrigation needs in many agricultural regions of the world. In particular, projected increases in the frequency and severity of droughts will probably increase future agricultural water demand in arid and semiarid regions. In addition, excess use of water in many irrigated lands causes deterioration of the soil and water resources, which not only harms the environment but also diminishes the productivity of irrigated agriculture. Hence, the resilience of irrigated agriculture will depend on the successful implementation of sustainable irrigation policies and management strategies to adapt to water scarcity and to reduce related environmental effects.

Effective strategies and practices for integrated land and water resources management may increase resource productivity while reducing the negative effects. Inefficient use of pesticides can be replaced with integrated pest management practices. On-farm water conservation strategies, such as improved land leveling, zero tillage, cover crops and advanced irrigation technologies, such as drip irrigation, have the potential to save a significant amount of water, reduce nutrient losses and save energy. Smart and precision agriculture strategies, which enable the applications of inputs (seed, water, fertilizer, pesticides, etc.) to be only as much as, when and where needed, can help to increase resource use efficiency and protect the environment. However, the negative environmental effects of irrigation can only be significantly minimized once sustainable living strategies such as reducing food loss and waste are practiced together with sustainable irrigated agriculture.

References

- Abascal E, Gómez-Coma L, Ortiz I, and Ortiz A (2022) Global diagnosis of nitrate pollution in groundwater and review of removal technologies. *Science of the Total Environment* 810. <https://doi.org/10.1016/j.scitotenv.2021.152233>.
- Ahmed M, Matsumoto M, and Kurosawa K (2018) Heavy metal contamination of irrigation water, soil, and vegetables in a multi-industry district of Bangladesh. *International Journal of Environmental Research* 12: 531–542. <https://doi.org/10.1007/s41742-018-0113-z>.
- Alexandratos N and Bruinsma J (2012) *World agriculture towards 2030/2050: The 2012 revision ESA working paper 12-03*. Rome: FAO.
- Andrews A and Henrique G (2006) *Heavy metal pollution and agriculture. An attempt to assess to existence and effects of heavy metal contamination in our natural environment*. TMS 424 Public information Bulletin.

- Angelakis A, Zaccaria D, Krasinikoff J, Salgot M, Bazza M, Roccaro P, Jimenez B, Kumar A, Yinghua W, Baba A, Harrison JA, Garduno-Jimenez A, and Fereres E (2020) Irrigation of world agricultural lands: Evolution through the Millennia. *Watermark* 12: 1285.
- Banach JL and van der Fels-Klerx HJ (2020) Microbiological reduction strategies of irrigation water for fresh produce. *Journal of Food Protection* 83(6): 1072–1087. <https://doi.org/10.4315/JFP-19-466>.
- Blumenthal UJ and Peasey A (2002) Critical review of epidemiological evidence of the health effects of wastewater and excreta use in agriculture. Unpublished document prepared for World Health Organization, Geneva. www.who.int/water_sanitation_health/wastewater/whocriticalrev.pdf.
- Caló F, Notti D, Galve JP, Abdikan S, Görüm T, Pepe A, and Balık Şanlı F (2017) DInSAR-based detection of land subsidence and correlation with groundwater depletion in Konya Plain, Turkey. *Remote Sensing* 2017(9): 83.
- Cassman KG, Dobermann A, and Walters D (2002) Agroecosystems, nitrogen-use efficiency, and nitrogen management. *Ambio* 31: 132–140.
- Chindarkar N and Grafton RQ (2019) India's depleting groundwater: When science meets policy. *Asia & the Pacific Policy Studies* 1–17.
- Fagúndez J, Olea PP, Tejedo P, Mateo-Tomás P, and Gómez D (2016) Irrigation and maize cultivation erode plant diversity within crops in mediterranean dry cereal agro-ecosystems. *Environmental Management* 58(1): 164–174. <https://doi.org/10.1007/s00267-016-0691-5> Epub 2016 Mar 19 PMID: 26994604.
- Fan LY (2017) China finds pesticide office to combat pollution, Overuse. <http://www.sixthtone.com/news/1000987/china-finds-pesticide-office-to-combat-pollution%2C-overuse#>. Accessed May 18, 2021.
- FAO (2011) *The State of the World's Land and Water Resources for Food And Agriculture (SOLAW)—Managing Systems at Risk*. Rome/London: Food and Agriculture Organization of the United Nations/Earthscan.
- FAO (2017) *Water for Sustainable Food and Agriculture: A Report Produced for the G20 Residency of Germany*. Rome: FAO.
- FAO (2018) *Progress on Water Use Efficiency—Global Baseline for SDG 6 Indicator 6.4.1 2018*. Rome: FAO/UN-Water56.
- FAO (2020) *The State of Food and Agriculture 2020. Overcoming Water Challenges in Agriculture*. Rome: FAO.
- Garduño H, Romani S, Sengupta B, Tuinhof A, and Davis R (2011) *India, Groundwater Governance Case Study*. Water Papers 71727 Washington DC: World Bank.
- Jain M, Fishman R, Mondal P, Gillian, Galford L, Bhattarai N, Naeem S, Lall U, Balwinder-Singh, and DeFries RS (2021) Groundwater depletion will reduce cropping intensity in India. *Science Advances* 7(9): eabd2849.
- Jaramillo S, Graterol E, and Pulver E (2020) Sustainable transformation of rainfed to irrigated agriculture through water harvesting and smart crop management practices. *Frontiers in Sustainable Food Systems* 4: 437086.
- Kibret S, McCartney M, Lautze J, Nhamo L, and Yan G (2021) The impact of large and small dams on malaria transmission in four basins in Africa. Springer Nature *Scientific Reports* 11: 13355.
- Lund AJ, Rehkopf DH, Sokolow SH, et al. (2021) Land use impacts on parasitic infection: A cross-sectional epidemiological study on the role of irrigated agriculture in schistosome infection in a dammed landscape. *Infectious Diseases of Poverty* 10: 35.
- Ma J, Chen Y, Weng L, Peng H, Liao Z, and Li Y (2021) Source identification of heavy metals in surface paddy soils using accumulated elemental ratios coupled with MLR. *International Journal of Environmental Research and Public Health* 18: 2295.
- Mastrocicco M and Colombani N (2021) The issue of groundwater salinization in coastal areas of the mediterranean region: A review. *Watermark* 13: 90. <https://doi.org/10.3390/w13010090>.
- Mekonnen MM and Hoekstra AJ (2015) Global gray water footprint and water pollution levels related to anthropogenic nitrogen loads to fresh water. *Environmental Science and Technology* 49: 12860–12868.
- Mueller ND, Gerber JS, Johnston M, Ray DK, Ramankutty N, and Foley JA (2012) Closing yield gaps through nutrient and water management. *Nature* 490: 254–257. <https://doi.org/10.1038/nature11420>.
- Nikolskii YN, Aidarov IP, Landeros-Sanchez C, and Pchyolkina VV (2019) Impact of long-term freshwater irrigation on soil fertility. *Irrigation and Drainage* 68(10): 993–1001.
- OECD (2012) *Environmental Outlook to 2050*. Paris: OECD.
- PAN International Pesticide Action Network International (2010) *Communities in Peril: Global Report on Health Impacts of Pesticide Use in Agriculture*. PAN International Pesticide Action Network International.
- Phocaides A (2007) *Handbook on Pressurized Irrigation Techniques*. Rome: FAO.
- Presser TS and Luoma SN (2006) Forecasting selenium discharges to the San Francisco Bay-Delta Estuary; ecological effects of a proposed San Luis Drain extension. In: *U.S. Geological Survey Professional Paper 1646* 196. <https://pubs.usgs.gov/pp/p1646/>.
- Qadir M, Quillerou E, Nangia V, Murtaza G, Singh M, Thomas RJ, Drechsel P, and Noble AD (2014) Economics of salt-induced land degradation and restoration. *Natural Resources Forum* 38(4): 282–295.
- Quemada M, Baranski M, Nobel-de Lange MNJ, Vallejo A, and Cooper JM (2013) Meta-analysis of strategies to control nitrate leaching in irrigated agricultural systems and their effects on crop yield. *Agriculture, Ecosystems and Environment* 174: 1–10.
- Qureshi AS (2020) Groundwater governance in Pakistan: From colossal development to neglected management. *Watermark* 12: 3017.
- SDWF (2022) *Safe Drinking Water Foundation*. <https://www.safewater.org/fact-sheets-1/2017/1/23/pesticides>.
- Siebert S, Kummu M, Porkka M, Döll P, Ramankutty N, and Scanlon BR (2015) A global data set of the extent of irrigated land from 1900 to 2005. *Hydrology and Earth System Sciences* 19: 1521–1545.
- Smil V (2000) Phosphorus in the environment: Natural flows and human interferences. *Annual Review of Energy and the Environment* 25: 53–88.
- Squires VR and Glenn EP (2011) Salination, desertification and soil. Erosion. (Encyclopedia of Life Support Systems (EOLSS), Ed.) In: *The Role of Food, Agriculture, Forestry and Fisheries in Human Nutrition (Vol. III)*. Australia.
- Thebo AL, Drechsel P, Lambin EF, and Nelson KL (2017) A global, spatially explicit assessment of irrigated croplands influenced by urban wastewater flows. *Environmental Research Letters* 12: 074008.
- Tudi M, Daniel Ruan H, Wang L, Lyu J, Sadler R, Connell D, Chu C, and Phung DT (2021) Agriculture development, pesticide application and its impact on the environment. *International Journal of Environmental Research and Public Health* 18: 1112.
- UNESCO (2020) *United Nations World Water Development Report 2020: Water and Climate Change*. Paris: UNESCO.
- UN-Water (2021) *United Nations Summary Progress 2021: SDG 6 Indicators*. https://www.unwater.org/app/uploads/2021/03/Infographics_Progress-on-water-and-sanitation.png.
- Van Weert F, Van Gun J, and Reckman J (2009) *Global Overview of Saline Groundwater Occurrence and Genesis*. Utrecht: IGRAC.
- Wada Y (2012) *Non-sustainable Groundwater Sustaining Irrigation*. GWF Discussion Paper 1205 Canberra, Australia: Global Water Forum.

Further readings

- Cui G, Lu Y, Zheng C, Liu Z, and Sai J (2019) Relationship between Soil Salinization and Groundwater Hydration in Yaoba Oasis, Northwest China. *Water* 11: 175. <https://doi.org/10.3390/w11010175>.
- Diamond J (2011) *Collapse: How Societies Choose to Fail or Succeed*. Penguin.
- Kanter DR (2018) Nitrogen pollution: A key building block for addressing climate change. *Climatic Change* 147(1–2): 11–21.
- Krishan G (2019) *Groundwater Salinity*. Current World Environment. 14/2.

- Monsef HA, Smith SE, and Darwish K (2015) Impacts of the Aswan High Dam after 50 years. *Water Resources Management* 29(6): 1873–1885. An International Journal –Published for the European Water Resources Association (EWRA).
- Pérez-Lucas G, Vela N, Aatik AE, and Navarro S (2018) Environmental risk of groundwater pollution by pesticide leaching through the soil profile. In: Larramendy M and Soloneski S (eds.) *Pesticides—Use and Misuse and their Impact in the Environment*. IntechOpen.
- Shahid SA (2013) Developments in salinity assessment, modeling, mapping, and monitoring from regional to submicroscopic scales. In: Shahid SA, Abdelfattah MA, and Taha FK (eds.) *Developments in Soil Salinity Assessment and Reclamation—Innovative Thinking and Use of Marginal Soil and Water Resources in Irrigated Agriculture*, pp. 3–43. Dordrecht/Heidelberg/New York/London: Springer.
- Quemada M, Baranski M, Nobel-de Lange MNJ, Vallejo A, and Cooper JM (2013) Meta-analysis of strategies to control nitrate leaching in irrigated agricultural systems and their effects on crop yield. *Agriculture, Ecosystems and Environment* 174: 1–10.

Trade-offs of wastewater irrigation

Manzoor Qadir^a and Christopher A Scott^b, ^aUnited Nations University Institute for Water, Environment and Health (UNU-INWEH), Hamilton, ON, Canada; ^bDepartment of Ecosystem Science and Management, 124 Forest Resources Building, Pennsylvania State University, University Park, PA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	277
Wastewater use for irrigation	278
Tradeoffs of wastewater irrigation	280
Nutrients	280
Organic matter and organic carbon	281
Salts	282
Metal and metalloids	283
Emerging pollutants	284
Pathogens	285
Conclusions	285
References	286

Abstract

Most governments in developing countries do not invest enough to treat wastewater, therefore, the farmers use untreated wastewater for irrigation. The positive aspects of such irrigation are a reliable and year-round water source, little or no spending on fertilizer, less pumping cost than with groundwater, an increase in soil organic matter, and greater income from market-ready crops. The negative risks include imbalanced nutrient addition, salinity, metal ion contamination, emerging contaminants, and potential exposure of farmers and consumers to pathogens and contaminants. Balancing these tradeoffs requires ensuring positive outcomes while minimizing risks through wastewater treatment and use compatible with its quality.

Key points

- Wastewater irrigation has increased over time because of rising volumes of wastewater from population growth, urbanization, and higher living standards coupled with growing urban demand for food, forage, and other urban and peri-urban agricultural produce.
- There is a large variation between developed and developing countries as well as among countries within different economic groups regarding safe management of wastewater.
- Positive aspects of wastewater irrigation are a year-round water supply, savings on fertilizer, less pumping cost, increase in soil organic matter, and greater income from high-value crops.
- Negative risks include imbalanced plant nutrient application, increased soil salinity, metal ion pollution, and exposure to emerging contaminants and pathogens.
- Safe and productive use of wastewater in agriculture helps in improving livelihoods of farming communities whilst protecting the environment and human and animal health.

Introduction

Freshwater ecosystems provide a range of crucial services such as drinking water, water for agriculture and energy, natural solutions for water purification and climate resilience, and habitats for fish and other aquatic species as well as recreation and tourism. Whilst the overall balance of freshwater at the global level suggests enough freshwater availability, freshwater scarcity has increased over time in specific regions, countries, and parts of countries due to uneven distribution of human population densities and water resources. Climate change and shifting weather and rainfall patterns have triggered freshwater scarcity (World Bank, 2018), which is expected to persist and deepen in arid and semi-arid areas, including those with dense population.

Recognized as a key challenge to sustainable development, freshwater scarcity is considered as a prospective cause of tensions within and across countries (UN-Water, 2020). Water-scarce areas characteristically fall under arid and semi-arid climates and depend on some combination of seasonal rainfall, mountain headwater runoff, and groundwater (with aquifers being non-renewable in most situations). Such areas have very few rivers, often originating in other countries. In addressing the water scarcity challenge, water-scarce countries and communities need to consider alternatives to conventional water resources to narrow

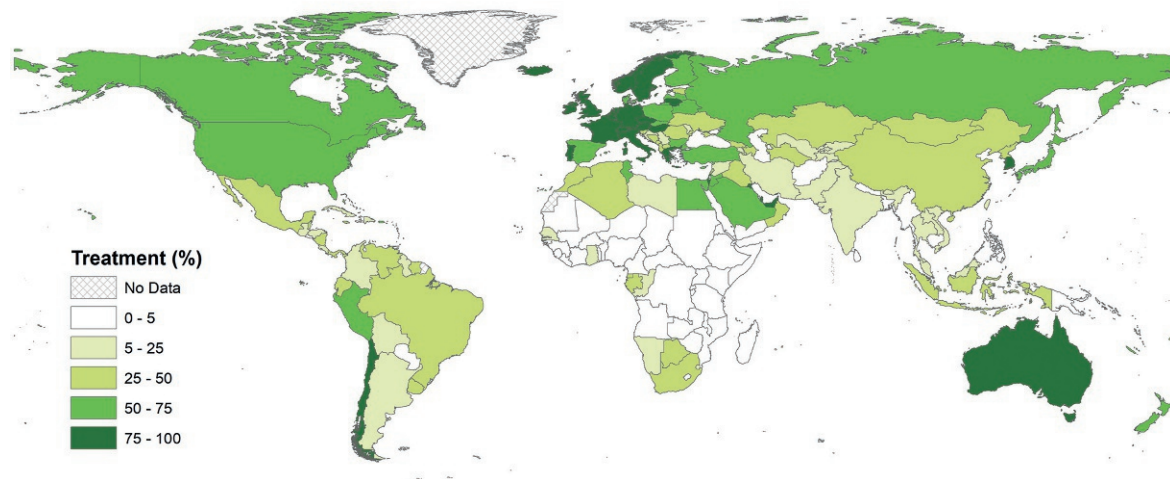


Fig. 1 Wastewater treatment (%) at the country scale. From Jones ER, van Vliet MTH, Qadir M and Bierkens MFP (2021) Country-level and gridded estimates of global wastewater production, collection, treatment, and re-use. *Earth System Science Data* 13: 237–254; reproduced under the Creative Commons Attribution 4.0 License.

the water demand–supply gap. In this regard, safe collection and treatment of municipal wastewater and its productive use offer a significant opportunity (Jones et al., 2021; Faruqui et al., 2004).

Although agriculture is the dominant user of freshwater in water-scarce areas, the sector has been conceding its share gradually to household, municipal, and industrial activities where demand is increasing rapidly. Since the use of freshwater in domestic, municipal, and industrial sectors generates wastewater, the volumes of wastewater are on the increase and available for fit-for-purpose use in agriculture after pertinent treatment (Qadir et al., 2020). Fit-for-purpose use is defined as treating wastewater to a water quality standard acceptable by a specific user. Thus, fit-for-purpose treatment can optimize the water quality for its next role. For example, most non-potable use options for wastewater require a quality lower than that of drinking water, so that secondary treatment is often adequate for irrigation of non-food crops (WWAP, 2017).

Given water scarcity, the use of treated wastewater needs to be a common practice in water-scarce areas. In contrast, the use of large volumes of untreated or inadequately treated wastewater in agriculture is common in developing countries (Jones et al., 2021). Despite multiple demonstrated benefits of safely managed wastewater for the farmers and associated communities, the progress regarding wastewater treatment and fit-for-purpose use of treated wastewater remains very slow in most developing countries. Countries with different levels of economic development demonstrate large variation in wastewater treatment (Fig. 1). On average, the countries falling under the high-income category treat 70–74% of the wastewater generated, followed by upper-middle-income countries (38–43%), lower-middle-income countries (26–28%), and finally low-income countries (4–8%) (Sato et al., 2013; Jones et al., 2021). The drivers behind wastewater reuse are a complex mixture of social, economic, geographic, and hydrological factors, and data are very limited globally. An estimated 41 billion km³ of treated wastewater is intentionally used (Jones et al., 2021), particularly in the Middle East and North Africa, Australia, western North America, and southern Europe.

By acknowledging the importance of water quality protection and wastewater treatment, the 2030 Agenda for Sustainable Development has a specific target in the Sustainable Development Goal (SDG) 6, which addresses safe water and sanitation for all. Addressing water quality and wastewater treatment explicitly, SDG Target 6.3 states: “By 2030, improve water quality by reducing pollution, eliminating dumping and minimizing release of hazardous chemicals and materials, halving the proportion of untreated wastewater and substantially increasing recycling and safe reuse globally.”

This chapter provides an overview of wastewater irrigation and assesses the tradeoffs between the positive aspects of wastewater irrigation (year-round water supply, savings on fertilizer, less pumping cost, increase in soil organic matter, and greater income from high-value crops) in relation to the negative risks (imbalanced soil nutrient application, increased soil salinity, metal ion pollution, and exposure to emerging contaminants and pathogens).

Wastewater use for irrigation

While wastewater has multiple uses—agricultural and landscape irrigation, groundwater recharge, agroforestry systems, aquaculture, cooling of thermoelectric power stations, dust control, and instream ecological conservation—its most prevalent use remains for agricultural and landscape irrigation (WWAP, 2017). In most developing countries, large volumes of untreated wastewater



Fig. 2 Municipal wastewater is a valuable source of water, nutrients, and energy, but large volumes are released to the environment in untreated form. Farmers in dry areas pump untreated wastewater to irrigate nearby fields. Credit: Manzoor Qadir.

are used for irrigation in urban and peri-urban settings (Fig. 2) where farmers' land holdings are small and these farmers are not able to optimize the volume or quality of the wastewater they receive (Qadir et al., 2010; Jones et al., 2021).

A two decades-old estimate of wastewater irrigation suggested an area of up to 20 million ha worldwide (Scott et al., 2004 citing 2001 work); this estimate was also cited in Jiménez and Asano (2008). Later, a global, spatially explicit assessment of wastewater-irrigated areas reveals the area under direct and indirect wastewater irrigation at 36 million ha, of which 29 million ha are likely to be exposed to untreated wastewater flows (Thebo et al., 2017). These estimates consider a cropping intensity of 1.48 and wastewater irrigation in both diluted and undiluted forms. Using the same cropping intensity (1.48) considered by Thebo et al. (2017), recent estimates of wastewater production by Qadir et al. (2020) suggest an irrigation potential with undiluted wastewater at 42 million ha.

Despite increasing water scarcity and emerging problems of the deterioration in water quality coupled with climate change events, the developing countries in arid and semi-arid areas are far from achieving the desired levels of wastewater treatment. Wastewater treatment plants either do not exist or existing plants are overwhelmed by poor operation and maintenance. In the case of operational wastewater treatment plants, they are often run well beyond their design capacity as they receive more wastewater than the volumes for which they were designed to treat in a specific timeframe and a given efficiency (Sato et al., 2013). Such realities of wastewater treatment status on the ground question the technical feasibility and robustness of wastewater treatment processes (if they exist at all), and as a result, the quality of treated wastewater is highly variable, from no treatment to full treatment (Grangier et al., 2012). Although the policies and regulations prohibiting the agricultural use of untreated or inadequately treated wastewater are in place, they remain far from their desired implementation (Qadir et al., 2010).

Despite the benefits demonstrated, the use of wastewater for irrigation falls short of its potential due to a range of constraints, such as: (1) inadequate public budget allocation for adequate wastewater management—collection of domestic and industrial wastewater separately, their pertinent treatment, and safe use or disposal; (2) limited comprehensive assessments on the impacts of long-term irrigation with untreated wastewater on the environment and human health; (3) incomplete economic analyzes without undertaking sound assessments and feasibility studies on the anticipated benefits from resource recovery—water, nutrients, precious metals, and energy—and comparative evaluation of the costs of other alternate sources of water; (4) lack of cost-recovery mechanisms for safely managed wastewater to fund pertinent wastewater managed systems; (5) disparity between water pricing and regional or local water scarcity without valuing treated wastewater, leading to the shutting down of wastewater treatment plants or supplementing costs with substantial subsidies to keep the wastewater management systems functional; (6) lack of clarity on ownership or use rights over wastewater once it is treated and released from treatment plants to local water bodies; (7) lack of political will and effective public policies to prioritize and encourage implementation of water recycling and reuse systems; (8) lack of clarity among wastewater-related national agencies and local institutions for effective management of wastewater; (9) little understanding of soil properties such as their filter and buffering capacity while considering wastewater irrigation systems; and (10) lack of training and capacity development to produce skilled human resources capable of handling the complexity resulting from wastewater collection, treatment, and reuse systems.

In general, the lack of commitment on the part of several developing-country governments to advocate and support comprehensive wastewater treatment programs has led to the lack of understanding among farmers and households about the anticipated environmental and health benefits from wastewater treatment and irrigation with treated water (Sato et al., 2013). This contrasts with formal promotion of desalination as another water-scarcity adaptation mechanism, despite inherent tradeoffs in its use (Qadir et al., 2007; Wilder et al., 2016). Because the wastewater treatment plants in low- and lower-middle-income countries, if any, are away from urban areas, urban households do not greatly appreciate the benefits of wastewater treatment and reuse in water-scarce

environments. Although the governments in such situations collect fees for water connections and wastewater services, they do not invest enough on the treatment of wastewater. On the ground, the water resources management systems are inefficient in most developing countries and do not consider appropriately the potential gains from treated water as a precious resource, which can be used for a range of opportunities such as agricultural and landscape irrigation.

Tradeoffs of wastewater irrigation

Despite issues with the quality of untreated and inadequately treated wastewater and its reuse in agriculture, there are certain positive trade-offs of wastewater irrigation. Most farmers in water-scarce low-income and lower-middle-income countries and communities with access to wastewater do not hesitate to use it even in untreated or inadequately treated forms for irrigation (Qadir et al., 2010; WWAP, 2017). The reasons are that (1) wastewater is the only water source available for irrigation throughout the year, (2) its use for irrigation reduces the need for purchasing certain farm inputs like fertilizer as wastewater is rich in essential nutrients, (3) irrigation with wastewater involves less energy cost if the alternative clean water source is deep groundwater, (4) wastewater irrigation provides greater farm work opportunities due to high cropping intensities and yields, and (5) wastewater enables farmers in peri-urban areas to produce high-value crops for sale in nearby urban markets (Minhas and Samra, 2004; Murtaza et al., 2010).

A reliable supply of irrigation water and nutrients for crop growth are crucial inputs to agricultural production systems, therefore, wastewater irrigation delivers both inputs to a significant extent (Hamilton et al., 2005). This is crucial in situations where wastewater is the only source of irrigation water available throughout the year or where other sources of irrigation water are extremely limited to support viable irrigation systems (WWAP, 2017). As the farmers in wastewater-irrigated areas often have small land holdings, wastewater irrigation provides them with the opportunity to operate year-round farming systems, which help these farmers to increase their income per unit of land area while alleviating poverty (Fig. 3).

Together with positive benefits and controlling factors, there is a range of tradeoffs, which may be negative or positive depending on the treatment of wastewater or lack of it as well as the quality and availability of wastewater (Lazarova and Bahri, 2005; Vulliet et al., 2011; Montes-Grajales et al., 2017).

Nutrients

Although municipal wastewater is rich in nutrients, the concentrations of these nutrients vary widely as demonstrated in wastewater samples collected from different locations and sources (Qadir et al., 2020). Such nutrient concentrations also depend on wastewater treatment and nutrient removal processes during specific treatment processes. Based on data on wastewater quality from different locations in South Asia, the estimates reveal that concentrations of the nutrient elements are highly variable, for example nitrogen is found in the range of 11 and 98 mg L⁻¹ and phosphorous of 1–30 mg L⁻¹ (Qadir and Scott, 2010). Based on 53 wastewater quality datasets from across the world, Qadir et al. (2020) estimate the average concentrations of nitrogen as 43.7 mg L⁻¹ and phosphorus as 7.8 mg L⁻¹.

For the wastewater-irrigating farmers, the nutrients in wastewater are termed as a major driving force for irrigation with untreated wastewater (Table 1). As most farmers believe wastewater treatment removes most nutrients, they prefer untreated wastewater over



Fig. 3 Wastewater is an important asset for resource-poor farmers, supporting local livelihoods. Year-round supply often means fodder for animal feed is cultivated using wastewater. Credit: Jeroen Ensink.

Table 1 Potential contribution of irrigation with wastewater in terms of nutrient addition to the soil. Derived from the data on nutrient concentrations in 53 wastewater samples.

Nutrient	Nutrient concentration (mg L ⁻¹)	Fertilizer contribution (kg ha ⁻¹)		
		Irrigation rate (5000 m ³ ha ⁻¹)	Irrigation rate (8000 m ³ ha ⁻¹)	Irrigation rate (10,000 m ³ ha ⁻¹)
Nitrogen	43.7	219	350	437
Phosphorus	7.8	39	62	78
Potassium	16.5	83	132	165

From Qadir M, Drechsel P, Cisneros BJ, Kim Y, Pramanik A, Mehta P and Olaniyan O (2020) Global and regional potential of wastewater as a water, nutrient, and energy source. *Natural Resources Forum* 44: 40–51.

treated wastewater for irrigation. In contrast to farmers' perception, the evaluation of nutrient concentrations in untreated and treated wastewater suggests that treated wastewater contains sufficient nutrient concentrations to meet crop nutrient requirements (Yadav et al., 2002). In addition to macronutrients, a range of micronutrients such as iron, zinc, manganese, and copper are also added to the soil through wastewater irrigation.

While most wastewater contains enough nutrients to meet crop requirements, managing required concentrations of nutrients in wastewater is a challenge due to possible negative impacts stemming from their excessive addition to wastewater-irrigated soils. In the case of macronutrients such as nitrogen and phosphorus, there are three likely pathways: (1) discharge of nutrient-rich untreated and inadequately treated wastewater to freshwater bodies may pose a significant challenge to the receiving waters over the long run, causing eutrophication of natural water bodies, undesirable growth of algae and weeds in wastewater channels and soils; (2) surplus and uncontrolled nitrogen application through wastewater irrigation may lead to excessive vegetation and green biomass resulting in a delay in crop maturity and a consequent decrease in economic yield; and (3) leaching of nitrogen may cause groundwater pollution, and methaemoglobinemia in infants drinking such nitrate-rich groundwater. Methaemoglobinemia reduces the ability of blood to carry vital oxygen around the body. Nitrates are highly soluble and can easily be moved to groundwater through the soil profiles in wastewater-irrigated areas.

The fertilizer value of wastewater has great importance because nutrients in wastewater contribute to crop requirements, therefore, periodic monitoring is essential to estimate the nutrient loads in wastewater and adjust fertilizer applications accordingly (Lazarova and Bahri, 2005). In addition to the negative effects from excessive addition of nitrogen to wastewater-irrigated soils, such applications of nitrogen also lead to increased emissions of nitrous oxides into the atmosphere.

Pertinent management of municipal wastewater and recovery of excess nutrients such as nitrogen and phosphorus by the wastewater treatment process offers environmental protection from minimizing eutrophication, health benefits with the elimination of methaemoglobinemia, and financial benefits from the recovery of nutrients (UN-Water, 2020). A recent assessment of nutrient recovery from wastewater suggests that nutrient concentrations in wastewater can account for 14.4% of global demand for nitrogen and 6.8% for phosphorus as fertilizers. In terms of potential economic gains, the recovery of nitrogen and phosphorus from wastewater may generate annual revenue of \$11.3 billion (Qadir et al., 2020).

Organic matter and organic carbon

The occurrence of organic matter in wastewater may have a range of positive impacts on wastewater-irrigated soils depending on the nature of the organic material added through wastewater irrigation. The addition of organic matter to the soil through wastewater improves soil structure, acts as a store of nutrients essential to meet crop nutritional needs, and enhances soil chemical properties, such as cation exchange capacity. The latter helps to adsorb undesirable metal ions on the soil cation exchange complex and to decrease their availability in soil solution for plant uptake.

Beyond the beneficial effects on some soil chemical properties, wastewater irrigation improves certain soil physical characteristics, such as aggregate stability, water holding capacity, and hydraulic conductivity (Minhas and Samra, 2004). Periodic monitoring of these properties revealed a 24% increase in soil hydraulic conductivity after 15 years of wastewater irrigation, which was increased further to 39% after 25 years of wastewater irrigation when compared with freshwater-irrigated soils (Fig. 4). There was also an improvement in other soil physical properties investigated in the study (aggregate stability and water holding capacity). Such properties contribute to enhancing water storage in the soil, thereby increasing water use efficiency and crop productivity per unit of water and land. In another study on irrigation with wastewater for 3 years, Biswas and Mojid (2018) observed consistent but small increases in soil porosity, hydraulic conductivity, and water holding capacity in a soil irrigated with wastewater when compared with a freshwater-irrigated soil.

Field studies have shown that organic carbon in wastewater-irrigated soils increases regardless of soil characteristics. Minhas and Samra (2004) reported that long-term wastewater irrigation resulted in higher organic carbon levels than those irrigated with freshwater pumped from groundwater. For example, there was an 80% increase in soil organic carbon from 1.42% in

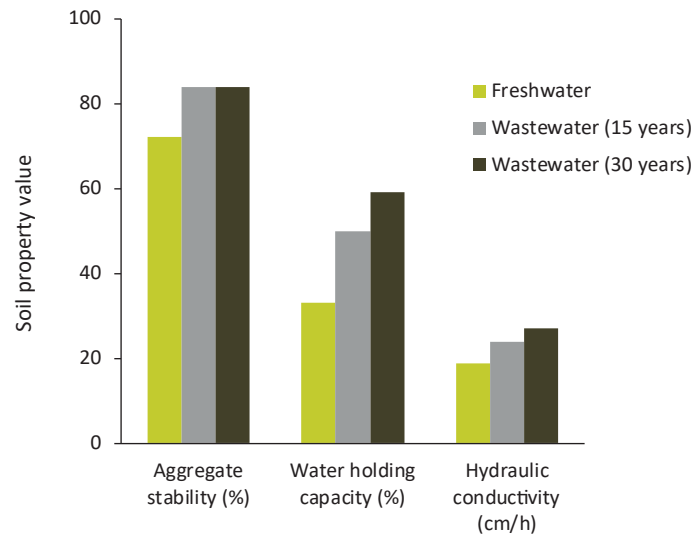


Fig. 4 Effects of 15 and 25 years of wastewater irrigation on some selected soil physical properties—aggregate stability (%), water holding capacity (%), and hydraulic conductivity (cm h^{-1})—when compared with freshwater irrigation. Modified from Minhas PS and Samra JS (2004) *Wastewater Use in Peri-Urban Agriculture: Impacts and Opportunities*. Karnal, India: Central Soil Salinity Research Institute. 75p.

freshwater-irrigated soils to 2.56% in wastewater-irrigated soil over a period of 15 years. Extending wastewater irrigation to 25 years resulted in an increase in soil organic carbon to 4.63% in wastewater-irrigated soil, suggesting a 226% increase over the freshwater-irrigated soil (Minhas and Samra, 2004). Organic and inorganic fractions of soil carbon are not only important for the soil to perform its agricultural productivity and environmental conservation functions, but also provide a pool of soil carbon that plays an important role in the global carbon cycle (Lal, 2004).

In contrast to a range of positive effects of organic matter added to wastewater-irrigated soils, there may be negative impacts resulting from (1) plugging of micro irrigation systems such as drippers and sprinklers; (2) hypoxic conditions in water holding bodies (tanks and ponds) due to depletion of dissolved oxygen; and (3) possibility of increased mortality in fish and other aquatic species (Qadir and Scott, 2010).

Salts

Wastewater contains a greater diversity of soluble salts than freshwater because it receives salts from a range of sources. The amount and type of salts used in a sector (domestic, municipal, industrial) affect the quality of wastewater it generates. For example, in the tannery industry, animal skins are usually salted with 50–100% salt and hides with 40–50% salt on a mass basis (Money et al., 2009). Such a large input of salt in the tannery industry generates wastewater containing salt concentrations in the range of $10\text{--}50\text{ g L}^{-1}$ ($16\text{--}80\text{ dS m}^{-1}$). For comparison, salt concentrations in domestic wastewater are low and range from $0.3\text{ to }0.5\text{ g L}^{-1}$ ($0.5\text{--}0.8\text{ dS m}^{-1}$), although both tannery and domestic wastewaters are broadly categorized as wastewater (Lazarova and Bahri, 2005).

Wastewater quality gets more complex when industrial waste streams are not discharged into separate sewers, rather into main municipal wastewater channels that carry wastewater to the treatment plants or to the channels carrying wastewater to farmers' fields. The challenge with the mixed wastewater is the removal of salt as the technologies (cation exchange resins or reverse osmosis membranes) are costly to ensure good-quality water (Toze, 2006; Qadir, 2018).

The amount and type of salts in wastewater and the treatment of wastewater or lack of it affect the quality of wastewater it generates. The effects of excess concentrations of salts in wastewater (saline wastewater) are different from those dominated by excess levels of sodium (sodic wastewater) (Oster et al., 2016; Qadir et al., 2021). Long-term irrigation with sodium-rich wastewater may result in the development of sodic soils, which exhibit structural problems driven by certain physical processes. Such soil degradation processes and conditions alter the geometry of soil pores, thereby affecting soil hydraulic properties such as infiltration rate and hydraulic conductivity. Such soil structural degradation may affect root penetration, seedling emergence, surface runoff, erosion, and tillage operations. After field studies for 8 years while irrigating olives with wastewater, Erel et al. (2019) found a steady increase in soil sodicity (higher exchangeable sodium percentage) in the wastewater-irrigated site reaching nearly double that of the freshwater site. They recommended close monitoring of soil sodicity while irrigating with wastewater to avoid soil degradation and allow consideration of practices leading to soil remediation.

Some industries release wastewaters with elevated levels of magnesium and potassium together with sodium. This is especially the case for wastewaters generated by dairies, piggeries, wineries, and olive or tomato processing plants. Research and practice in

recent years have demonstrated adverse effects of potassium and magnesium, in addition to sodium, on soil structure with degradation occurring through reduction in water movement into and through irrigated soils (Rengasamy and Marchuk, 2011; Smith et al., 2015). With increasing concentrations of potassium and magnesium in wastewaters (Oster et al., 2016), there is a need for a reassessment of the guidelines on irrigation water quality by considering the scale of the potential negative impacts of these cations to ensure the long-term sustainability of irrigated agriculture (Qadir et al., 2021).

Irrigation with saline and sodic wastewater may also affect groundwater quality. Wastewater generated from some industries, such as effluent from tanneries, is a major cause of groundwater contamination in surrounding areas and consequently may become a major socio-political issue, in addition to its environmental implications (Money et al., 2009). In the case of well-drained soils, salts and other contaminants may move through the soil profile into unconfined aquifers. The quality of wastewater in terms of salt concentration, soil characteristics such as texture and drainage, and the initial quality of the receiving groundwater are important factors that determine the extent to which salts in wastewater can affect groundwater quality. If the groundwater has high inherent salt concentration, then additional salt entering the groundwater from the soil will have a limited effect. But where groundwater has lower salt concentration, the impact of some salt addition may not be too adverse if the groundwater is not used for any other nearby purposes, e.g., drinking water supply.

Metal and metalloids

A range of metal ions and metalloids are essential for adequate plant growth, but they become toxic at elevated concentrations beyond permissible limits. Certain potentially toxic metals and metalloids are naturally present in the environment in trace amounts and are ingested with food, water, and air. Several of these metals and metalloids are of particular concern due to their adverse effects on agricultural productivity as well as on environment health (Gupta et al., 2012). For example, metal ions such as cadmium, mercury, and lead do not have essential functions, but they are detrimental, even in small quantities, to plants, animals and humans, and accumulate because of their long biological half-life (Hamilton et al., 2005). Other metals and metalloids, such as manganese, zinc, and copper are essential micro-nutrients in small concentrations, but harmful to crops when they surpass maximum allowable concentrations (Table 2).

Most of the industries in developing countries discharge untreated wastewater containing variable concentrations of metals and metalloids into municipal wastewater conveyance systems. Without separation of industrial and domestic wastewater, the wastewater channels carry a blend of industrial and domestic wastewater. Estimates reveal that most untreated industrial-domestic-mix wastewaters contain higher concentrations of chromium, lead, and cadmium than their permissible limits in irrigation water (Murtaza et al., 2010).

Several studies have been carried out to evaluate the implications of wastewater irrigation on the concentrations of metals and metalloids in soils and crops (Qadir et al., 2000; Minhas and Samra, 2004; Lazarova and Bahri, 2005; Hamilton et al., 2007). In a comprehensive sampling program undertaken in two peri-urban areas of Faisalabad, Pakistan, Simmons et al. (2010) quantified the effects of long-term untreated wastewater irrigation on soil quality, wheat yield, and chemical composition of wheat grain and straw. The metal ion concentrations in soils remained below the EU Maximum Permissible Levels for cadmium (Cd), lead (Pb), and zinc (Zn) in sludge-amended soils. In the case of wastewater-irrigated wheat, cadmium and lead concentrations remained below the Permissible Levels for these metals in feed materials. In addition to the studies addressing the impacts of wastewater irrigation on metal ion concentrations in soils and crops, there is a need to regulate heavy metal loads over time as metals and metalloids tend to accumulate in wastewater-irrigated soils. In long-term wastewater irrigation systems, high loads of metal ions may limit crop production. Consequently, the farmers may need to grow plant species that exclude metals, or to use the plants for

Table 2 Distinct groups of selected metal ions based on their bioavailability, phytotoxicity, and risks.

Group	Metal	Soil adsorption	Phytotoxicity and risks
1	Chromium, silver, and titanium	Low solubility and strong retention in soil	Low phytotoxicity and little associated risks as these metals are taken up by plants in negligible concentrations
2	Arsenic, mercury, and lead	Strongly adsorbed by soil colloids	Plant roots may take up these metals but do not translocate to shoots, though elevated arsenic from mixed geogenic-anthropogenic sources has been found in grains. These metals are generally not phytotoxic except at very high concentrations
3	Boron, copper, manganese, nickel, and zinc	Less strongly adsorbed by soil colloids than Groups 1 and 2	Readily taken up by plants and phytotoxic at concentrations that pose little risk to human health
4	Cadmium, cobalt, molybdenum, and selenium	Least adsorbed of all metals	Pose human and animal health risks at plant tissue concentrations that are not generally phytotoxic. Bioaccumulation of these metals occurs through the soil-plant-animal food chain

Modified from Hamilton AJ, Boland A-M, Stevens D, Kelly J, Radcliffe J, Ziehl A, Dillon PJ and Paulin R (2005) Position of the Australian horticultural industry with respect to the use of reclaimed water. *Agricultural Water Management* 71: 181–209.



Fig. 5 Long-term irrigation with untreated or inadequately treated wastewater may lead to absorption of contaminants by crops to elevated levels harmful to the health of humans and animals consuming the crops. Credit: Manzoor Qadir.

other purposes than human and animal food, such as biofuels, ornamental plants, etc. Soil properties such as pH, organic matter and clay content are important to reduce metal mobility and bioavailability.

Based on a survey carried out along the Musi River in India, [Minhas and Samra \(2004\)](#) revealed the transfer of metal ions from wastewater to cows' milk from para grass grown on wastewater-irrigated soil and fed to the animals. The proportion of samples showing excessive amounts of pollutants in para grass samples ranged from 4% for cadmium to 100% for lead. Milk samples were highly contaminated with both metal ions ranging from 1.2 to 40 times higher than the permissible limits. [Qadir et al. \(2000\)](#) found that leafy vegetables irrigated with untreated wastewater accumulated cadmium in greater amounts than non-leafy vegetables on an equivalent mass basis. [Sharma et al. \(2007\)](#) concluded that wastewater irrigation increased the contamination of edible parts of vegetables with cadmium, lead, and nickel resulting in potential long-term health risks. Generally, metal ion concentrations in plant tissue increase with their respective concentration in irrigation water. Concentrations in the roots were usually found to be higher than in the leaves of a range of plant species ([Qadir and Scott, 2010](#)).

Unregulated use of wastewater for irrigation leading to the excessive accumulation of metals and metalloids in soils is detrimental as it is extremely difficult to remove such constituents from soils once they reach the loading limits of certain metals and metalloids ([Fig. 5](#)). Excessive concentrations of metals and metalloids in wastewater-irrigated soils may subsequently cause (1) metal and metalloid toxicity to plants grown on contaminated soils; (2) absorption by crops, resulting in metal and metalloid levels in plant tissues which may be harmful to the health of humans and animals consuming these crops; and (3) leaching of metals and metalloids from soils to groundwater or their accumulation in surface water triggered by surface runoff, thereby rendering the water contaminated and unsafe for other uses ([Murtaza et al., 2010](#)). In relation to groundwater, the use of untreated or inadequately wastewater for surface irrigation may result in leaching of metals and metalloids and other undesirable constituents to groundwater with severe impacts on its quality.

Emerging pollutants

Emerging pollutants refer to any synthetic or naturally occurring chemical or any microorganism that is not commonly monitored in the environment but has the potential to cause known or suspected adverse ecological and/or human health effects ([WWAP, 2017](#)). The changes in living styles and living standards have led to the introduction of a range of such pollutants to wastewater, which consist of but are not limited to endocrine disruptor compounds, hormones, residual pharmaceuticals, drugs, and active residues of personal care products. The presence of such emerging contaminants in wastewater is a matter of growing concern. Such contaminants may have adverse biological effects in animals and humans even at low-level exposures. A range of chemicals have been identified as potential endocrine-disrupting contaminants ([Lu et al., 2020](#)), including industrial contaminants.

Some residual pharmaceuticals and personal care products tend to survive wastewater treatment. Soils irrigated with wastewater containing such contaminants may not hold them resulting in their percolation through the soil to the groundwater. Although many residual pharmaceuticals and personal care products may not be toxic at low concentrations, but they may result in health implications through their effects on the immune and hormonal systems of animals and humans even at low concentrations. There

may be some unspecified toxic effects in the form of rising antibiotic-resistance of bacteria through repeated exposure to antibiotics in wastewater and contaminated streams (Bouwter, 2005).

While the presence of endocrine disrupting chemicals in the environment and the potential ecological consequences are alarming, their concentrations found in the water bodies are usually very low. When they are present in the soil, rapid microbial degradation or adsorption by the soil organic matter occurs. As a result, they may not enter the plant tissue through plant roots in significant concentrations.

The occurrence of many contaminants of emerging concern is often related to discharges from the wastewater treatment plants because of the widespread use of many of these compounds and the lack of technologies with sufficient removal efficiency (Ferreiro et al., 2020). In fact, the legislation related to wastewater treatment (Directive 2000/60/EC, Directive 2008/56/EC, Directive 2013/39/EU) does not yet include many of these compounds. Thus, the wastewater treatment plants remove a part only of such contaminants as the treatment plants are not specifically designed to remove them from wastewater (Prasse et al., 2015; Montes-Grajales et al., 2017). Contaminants of emerging concern have even been found in potable water (Vulliet et al., 2011). Consequently, recent research has focused on avoiding the accumulation of certain contaminants of emerging concern in drinking water.

Chemical stability and slow natural mitigation of certain contaminants makes their remediation a difficult environmental challenge. The degree to which persistent pollutants in wastewater need to be treated depends on (1) pollutant loads, i.e., concentration in wastewater and wastewater volume over time; (2) behavior of these compounds in the soil, which could be assessed through specific bioavailability tests before implementing costly remediation strategies; (3) soil properties, such as soils with large buffering capacities (adequate pH, high soil organic matter content, high cation exchange capacity, medium to deep profile) have greater capacity to filter larger pollutant loads than soils with smaller buffering capacities. For the sites already contaminated with these compounds, the approach usually taken is to isolate the affected sites, and either remove the contaminated soil or rely on phytoremediation (Qadir and Scott, 2010).

In developing countries, the exposure of crops and farmers to organic contaminants is probably higher through pesticide application than organic contaminants in irrigation water. The challenge of any related risk mitigation starts with its assessment, which is costly if based on actual analysis. Thus, such assessments are mostly based on observations (WHO, 2006). An alternative between both extremes may be risk assessment based on easy to monitor environmental factors and application practices (Simmons et al., 2010). More difficult and costly would be the analysis of emerging contaminants, like residual pharmaceuticals or endocrine disruptor compounds (WHO, 2006). A crucial step minimizing the exposure of crops and farmers as well as accumulation of emerging contaminants in soils refers to ensuring separate collection and pertinent treatment of liquid and solid wastes from related industries.

Pathogens

The occurrence of certain viruses, bacteria, protozoa, helminth eggs, and fecal coliforms in wastewater can cause a range of communicable diseases for wastewater-irrigating farmers and traders, and consumers of wastewater-grown food. Such diseases may consist of, but are not limited to, diarrhea, typhoid, dysentery, cholera, gastroenteritis, ascariasis, hepatitis, ulcer, and food poisoning (Grangier et al., 2012; Kesari et al., 2021).

The use of antibiotics for human beings to treat specific diseases has increased over time and it continues to intensify. The use of antibiotics in other sectors such as in poultry and dairy farms and animal production facilities, biochemical industries, and agriculture is common practice, and such use has also increased. Extensive use and/or misuse of antibiotics has resulted in an increase in multi-resistant bacteria, which carry multiple resistance genes (Xu et al., 2017). These multidrug-resistant bacteria collected through the sewerage and wastewater conveyance network end up in the wastewater treatment plants. Thus, it can be inferred that the wastewater treatment plants may serve as a hotspot of antibiotic-resistant bacteria and antibiotic resistance genes. The U.S. Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO) have already declared that antibiotic-resistant bacteria are an imminent hazard to human health (Kesari et al., 2021).

Among the viruses, coronavirus disease 2019 (COVID-19) in individuals who do not die can damage the lungs, heart, and brain, which increases the risk of long-term health problems. In severe cases, COVID-19 can cause loss of life. One diagnostic near-real-time option that has been highly visible after the pandemic started is detecting the virus causing COVID-19, namely SARS-COV2, in wastewater generated by communities, municipalities, and cities. A big advantage of SARS-COV2 monitoring in wastewater is that it picks up the virus associated with a vast number of people while providing early insights into the transmission trajectory of the virus, e.g., if it is increasing or declining (Larsen and Wigginton, 2020).

Conclusions

While water scarcity is recognized as a key challenge to sustainable development, water-scarce countries and communities need to consider increasingly alternate water resources, such as treated wastewater, to narrow the water demand–supply gap. While the use of treated wastewater needs to be enhanced in water-scarce areas for multiple benefits, large volumes of untreated or inadequately treated wastewater continue to be used for agricultural irrigation in developing countries. Thus, the progress on treatment of

wastewater and fit-for-purpose use of treated wastewater remains slow in most developing countries. Wastewater use for irrigation presents a complex series of tradeoffs, which offer crucial considerations for managing wastewater collection, treatment, and discharge.

The constituents of major concern for environmental risks from untreated or inadequately treated wastewater include excessive concentrations of metals and metalloids, nutrients, salts and specific ionic species, emerging contaminants, and certain viruses, bacteria, protozoa, helminth eggs, and fecal coliforms. The strategies for contaminant risk management have been focused mainly on wastewater treatment options, which have demonstrated positive impacts of safely managed wastewater in agriculture. These should be complemented with farm-based measures that can also play an important role in further minimizing contaminant risk for the environment and human health under situations where untreated or partially treated wastewater is used for irrigation.

There is a need to develop tools and models to evaluate risks from wastewater and implement pertinent risk reduction approaches as a part of the process to assist policy makers in deciding on available treatment options under specific conditions considering environmental, social, and economic dimensions. The right combination of flexible policies, effective institutions, and financial planning and support can help in executing risk management approaches. This is particularly important in areas generating large volumes of untreated wastewater where such interim measures are needed to gradually reach a point when most wastewater would be available in treated form for its safe and productive use.

References

- Biswas SK and Mojid MA (2018) Changes in soil properties in response to irrigation of potato by urban wastewater. *Communications in Soil Science and Plant Analysis* 49: 828–839.
- Bouwer H (2005) Adverse effects of sewage irrigation on plants, crops, soils, and groundwater. In: Lazarova V and Bahri A (eds.) *Water Reuse for Irrigation: Agriculture, Landscapes, and Turf Grass*, pp. 235–263. Boca Raton, USA: CRC Press.
- Erel R, Eppel A, Yermiyahu U, Ben-Gal A, Levy G, Zipori I, Schaumann GE, Mayer O, and Dag A (2019) Long-term irrigation with reclaimed wastewater: Implications on nutrient management, soil chemistry and olive (*Olea europaea* L.) performance. *Agricultural Water Management* 213: 324–335.
- Faruqui NI, Scott CA, and Raschid-Sally L (2004) Confronting the realities of wastewater use in irrigated agriculture: Lessons learned and recommendations. In: Scott CA, Faruqui NI, and Raschid-Sally L (eds.) *Wastewater Use in Irrigated Agriculture: Confronting the Livelihood and Environmental Realities*, pp. 173–185. Wallingford: CABI Publishing.
- Ferreiro C, Gómez-Motos I, Lombraña JI, de Luis A, Villota N, Ros O, and Etxebarria N (2020) Contaminants of emerging concern removal in an effluent of wastewater treatment plant under biological and continuous mode ultrafiltration treatment. *Sustainability* 12: 725.
- Grangier C, Qadir M, and Singh M (2012) Health implications for children in wastewater-irrigated peri-urban Aleppo, Syria. *Water Quality Exposure and Health* 4: 187–195.
- Gupta N, Khan DK, and Santra SC (2012) Heavy metal accumulation in vegetables grown in a long-term wastewater-irrigated agricultural land of tropical India. *Environmental Monitoring and Assessment* 184: 6673–6682.
- Hamilton AJ, Boland A-M, Stevens D, Kelly J, Radcliffe J, Ziehl A, Dillon PJ, and Paulin R (2005) Position of the Australian horticultural industry with respect to the use of reclaimed water. *Agricultural Water Management* 71: 181–209.
- Hamilton AJ, Stagnitti F, Xiong X, Kreidl SL, Benke KK, and Maher P (2007) Wastewater irrigation the state of play. *Vadose Zone Journal* 6: 823–840.
- Jiménez B and Asano T (2008) *Water Reuse: An International Survey of Current Practice, Issues and Needs*. IWA Publishing.
- Jones ER, van Vliet MTH, Qadir M, and Bierkens MFP (2021) Country-level and gridded estimates of global wastewater production, collection, treatment, and re-use. *Earth System Science Data* 13: 237–254.
- Kesari KK, Soni R, Jamal QMS, et al. (2021) Wastewater treatment and reuse: A review of its applications and health implications. *Water, Air, and Soil Pollution* 232: 208.
- Lal R (2004) Soil carbon sequestration impacts on global climate change and food security. *Science* 304: 1623–1627.
- Larsen DA and Wigginton KR (2020) Tracking COVID-19 with wastewater. *Nature Biotechnology* 38: 1151–1153.
- Lazarova V and Bahri A (2005) *Water Reuse for Irrigation: Agriculture, Landscapes, and Turf Grass*. Boca Raton, USA: CRC Press.
- Lu J, Zhang C, Wu J, Zhang Y, and Lin Y (2020) Seasonal distribution, risks, and sources of endocrine disrupting chemicals in coastal waters: Will these emerging contaminants pose potential risks in marine environment at continental-scale? *Chemosphere* 247: 125907.
- Minhas PS and Samra JS (2004) *Wastewater Use in Peri-Urban Agriculture: Impacts and Opportunities*. Karnal, India: Central Soil Salinity Research Institute. 75p.
- Money CA, Ramasami T, Chandra Babu NK, et al. (2009) Salinity reduction in tannery effluents in India and Australia. In: *ACIAR Project (AH/2001/005) Final Report*. Canberra: Australian Centre for International Agricultural Research. 51 pp.
- Montes-Grajales D, Fennix-Agudelo M, and Miranda-Castro W (2017) Occurrence of personal care products as emerging chemicals of concern in water resources: A review. *Science of the Total Environment* 595: 601–614.
- Murtaza G, Ghafoor A, Qadir M, Owens G, Aziz MA, Zia MH, and Saifullah. (2010) Disposal and use of sewage on agricultural lands in Pakistan: A review. *Pedosphere* 20: 23–34.
- Oster JD, Sposito G, and Smith CJ (2016) Accounting for potassium and magnesium in irrigation water quality assessment. *California Agriculture* 70: 71–76.
- Prasse C, Stalter D, Schulte-Oehlmann U, Oehlmann J, and Ternes TA (2015) Spoilt for choice: A critical review on the chemical and biological assessment of current wastewater treatment technologies. *Water Research* 87: 237–270.
- Qadir M (2018) Addressing trade-offs to promote safely managed wastewater in developing countries. *Water Economics and Policy* 4: 1871002.
- Qadir M and Scott CA (2010) Non-pathogenic tradeoffs of wastewater irrigation. In: Drechsel P, Scott CA, Raschid-Sally L, Redwood M, and Bahri A (eds.) *Wastewater Irrigation and Health: Assessing and Mitigating Risks in Low-income Countries*, pp. 101–126. Earthscan-International Development Research Centre (IDRC)-International Water Management Institute (IWM).
- Qadir M, Ghafoor A, and Murtaza G (2000) Cadmium concentration in vegetables grown on urban soils irrigated with untreated municipal sewage. *Environment, Development and Sustainability* 2: 11–19.
- Qadir M, Sharma BR, Bruggeman A, Choukr-Allah R, and Karajeh F (2007) Non-conventional water resources and opportunities for water augmentation to achieve food security in water scarce countries. *Agricultural Water Management* 87: 2–22.
- Qadir M, Wichelns D, Raschid-Sally L, McCornick PG, Drechsel P, Bahri A, and Minhas PS (2010) The challenges of wastewater irrigation in developing countries. *Agricultural Water Management* 97: 561–568.
- Qadir M, Drechsel P, Cisneros BJ, Kim Y, Pramanik A, Mehta P, and Olaniyan O (2020) Global and regional potential of wastewater as a water, nutrient, and energy source. *Natural Resources Forum* 44: 40–51.
- Qadir M, Sposito G, Smith CJ, and Oster JD (2021) Reassessing irrigation water quality guidelines for sodicity hazard. *Agricultural Water Management* 255: 107054.
- Rengasamy P and Marchuk A (2011) Cation ratio of soil structural stability (CROSS). *Soil Research* 49: 280–285.
- Sato T, Qadir M, Yamamoto S, Endo T, and Zahoor A (2013) Global, regional, and country level need for data on wastewater generation, treatment, and reuse. *Agricultural Water Management* 130: 1–13.

- Scott CA, Faruqui NI, and Raschid-Sally L (2004) Wastewater use in irrigated agriculture: Management challenges in developing countries. In: Scott CA, Faruqui NI, and Raschid-Sally L (eds.) *Wastewater Use in Irrigated Agriculture: Confronting the Livelihood and Environmental Realities*, pp. 1–10. Wallingford: CABI Publishing.
- Sharma RK, Agrawal M, and Marshall F (2007) Heavy metal contamination of soil and vegetables in suburban areas of Varanasi, India. *Ecotoxicology and Environmental Safety* 66: 258–266.
- Simmons RW, Ahmad W, Noble AD, Blummel M, Evans A, and Weckenbrock P (2010) Effect of long-term un-treated domestic wastewater re-use on soil quality, wheat grain and straw yields and attributes of fodder quality. *Irrigation and Drainage Systems* 24: 95–112.
- Smith C, Oster JD, and Sposito G (2015) Potassium and magnesium in irrigation water quality assessment. *Agricultural Water Management* 157: 59–64.
- Thebo AL, Drechsel P, Lambin EF, and Nelson KL (2017) A global, spatially-explicit assessment of irrigated croplands influenced by urban wastewater flows. *Environmental Research Letters* 12: 074008.
- Toze S (2006) Reuse of effluent water—Benefits and risks. *Agricultural Water Management* 80: 147–159.
- UN-Water. (2020) *UN-Water Analytical Brief on Unconventional Water Resources*. Geneva: Switzerland.
- Vulliet E, Cren-Olive C, and Grenier-Loustalot MF (2011) Occurrence of pharmaceuticals and hormones in drinking water treated from surface waters. *Environmental Chemistry Letters* 9: 103–114.
- WHO (World Health Organization) (2006) *Guidelines for the Safe Use of Wastewater, Excreta and Grey Water: Volume 2. Wastewater Use in Agriculture*. Geneva: WHO.
- Wilder MO, Aguilar-Barajas I, Pineda-Pablos N, Varady RG, Megdal SB, McEvoy J, Merideth R, Zuniga-Teran AA, and Scott CA (2016) Desalination and water security in the US–Mexico border region: Assessing the social, environmental and political impacts. *Water International* 41: 756–775.
- World Bank (2018) *Beyond Scarcity: Water Security in the Middle East and North Africa. MENA Development Report*. Washington, DC: The World Bank.
- WWAP (United Nations World Water Assessment Programme) (2017) *The United Nations World Water Development Report 2017. Wastewater: The Untapped Resource*. Paris: UNESCO.
- Xu YB, Hou M, Li YF, Huang L, Ruan JJ, Zheng L, et al. (2017) Distribution of tetracycline resistance genes and AmpC β -lactamase genes in representative non-urban sewage plants and correlations with treatment processes and heavy metals. *Chemosphere* 170: 274–281.
- Yadav RK, Goyal B, Sharma RK, Dubey SK, and Minhas PS (2002) Post-irrigation impact of domestic sewage effluent on composition of soils, crops and ground water—A case study. *Environment International* 28: 481–486.

Pollution and risks of trace elements in the soil environment

Peng Wang^a, Fang-Jie Zhao^a, and Steve P McGrath^b, ^aCollege of Resources and Environmental Sciences, Nanjing Agricultural University, Nanjing, China; ^bRothamsted Research, Hertfordshire, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction	288
Sources of trace elements in polluted soils	290
Pesticides	290
Fertilizers	291
Organic manures (animal manures, compost, and sewage sludge)	291
Wastewater for irrigation	292
Mining processes	292
Manufacturing and recycling industries	293
Energy production industries	293
Characteristics of trace elements in soils and potential risk pathways	293
Summary	294
References	294

Abstract

Trace elements may be either essential or toxic to organisms. The most common trace elements in soils are Cd, As, Cr, Hg, Pb, Ni, Zn, and Cu. Pollution of trace elements in soils may be natural or anthropogenic, with the latter including agricultural and industrial sources. Different trace elements pose variable potential risks to ecosystems, food safety and human health. Cadmium and As have low phytotoxicity, and a high risk of soil–food chain transfer posing a threat to food safety and human health, whereas Zn, Ni, Cu, and Cr are ecotoxic at elevated concentrations, thereby posing risks to productivity and functions of ecosystems including C and N cycles.

Key points

- Trace elements enter soils from natural and anthropogenic sources.
- Different trace elements pose different risk pathways to organisms.
- Cadmium and arsenic pose a high risk to food safety and human health.
- Zinc, Ni and Cu are highly toxic at elevated concentrations to ecosystems, threatening ecological functions.

Introduction

A trace element is a chemical element whose concentration (or other measure of amount) is very small (a ‘trace amount’). Some of them are known to be essential micronutrients to soil microorganisms, plants, and or animals, such as iron (Fe), cobalt (Co), copper (Cu), manganese (Mn), nickel (Ni), zinc (Zn), molybdenum (Mo), and selenium (Se). They are essential for metabolic activities at low concentrations, but when they are present at large concentrations exceeding a certain threshold, they are highly toxic to humans and other species. Some are considered to be nonessential (or toxic), such as arsenic (As), cadmium (Cd), lead (Pb), mercury (Hg), and chromium (Cr) (e.g., Cr(VI)) (Table 1; (Rahman and Singh, 2019)). These toxic elements do not have a specific threshold of risk (meaning that it is unclear whether there is any level of exposure that poses no risk to at least some organisms) and are therefore described as the most problematic elements. Even at certain concentrations, these toxic elements can activate resistance of soil bacteria, posing a potential threat to ecological functions.

Both essential and nonessential trace elements occur naturally in soils from geological and pedological processes. Where there is no anthropogenic influence, trace elements are normally present at small concentrations in total or in bioavailable form. They are often immobile and therefore do not pose problems for environmental or human health. In this case, the total concentration of a trace element in soil is called the natural (or geochemical) background value. This value often varies among soils, depending on the mineralogical composition of the soil parent material and on pedogenetic processes (e.g., soil formation). Soils can be polluted due to anthropogenic activities because they are the major sink for trace elements released from agricultural activities (e.g., mineral fertilizers, organic fertilizers including manure and sewage sludge, pesticides, and wastewater irrigation), and rapidly expanding industrial activities (e.g., mining, disposal of wastes, coal combustion residues, and atmospheric deposition) (see references in Wuana and Okieimen (2011)). Unlike organic contaminants, once trace elements have been introduced into soils, they do not undergo microbial or chemical degradation, and thus their total concentration persists for a long time after their introduction.

Table 1 Characteristics of toxic trace elements.

	<i>Arsenic (As)</i>	<i>Cadmium (Cd)</i>	<i>Chromium (Cr)</i>	<i>Mercury (Hg)</i>	<i>Lead (Pb)</i>
Global production (1000 kg year ⁻¹)	36,000–44,000	20,000–24,000	18,000–30,000 × 10 ³	2300	4800 × 10 ³
Abundance in Earth's crust (mg kg ⁻¹)	1.8–2.1	0.15	102–140	0.067–0.085	10.0–14.0
Main valency states in soils	As(III), As(V)	Cd(II)	Cr(III), Cr(VI)	Hg(0), Hg(I), Hg(II)	Pb(II), Pb(IV)
In lithosphere	300+ As-bearing minerals as arsenopyrite (FeAsS), cobaltite (CoAsS), niccolite (NiAs), realgar (AsS), orpiment (As ₂ S ₃), and arsenolite (As ₂ O ₃)	Greenockite and hawleyite (CdS), otavite (CdCO ₃), niedermayrite (Cu ₄ Cd(SO ₄) ₂ (OH) ₆ ·4H ₂ O), and cadmoselite (CdSe)	Ferrochromite (FeCr ₂ O ₄) and crocoite (PbCrO ₄)	Metallic (Hg)(0) or as cinnabar (HgS), corderoite (Hg ₃ S ₂ Cl ₂), and livingstonite (HgSb ₄ S ₈)	100+ Pb-containing minerals as galena (PbS), crocoite (PbCrO ₄), anglesite (PbSO ₄), pyromorphite (Pb ₅ (PO ₄) ₃ Cl), coronadite (PbMn ₈ O ₁₆), and cerussite (PbCO ₃)
Anthropogenic sources	Mining, smelting, combustion of fossil fuels, and production of glass and semiconductors, fertilizers, and chemotherapeutic drug of cancer	Ni–Cd batteries, coatings and platings, stabilizers for plastics, fossil fuel combustion, phosphate fertilizer, waste incineration, mining, ferrous and non-ferrous metal production, and cement production	Tanning, mining, textile dyes manufacturing, wood preservation, metal corrosion inhibition, electroplating, aircraft industry, and cleaning of glassware	Artisanal and small-scale gold mining (ASGM), coal combustion, production of non-ferrous and ferrous metals, cement production, pesticides, and fertilizers	Mining, ore processing, Pb-acid battery recycling (PABC), Pb-containing gasoline in petrol, pipes, ammunition, pesticides, pigment in paints, electronic wastes, dyes, and ceramic glazes
Natural sources	Weathering of rock, microbial colonization, and volcanic eruption	Dust storm, sea salt spray, volcanic activities, weathering, erosion, and wildfire	Tectonic and hydrothermal events	Degassing, weathering of rock, volcanic eruption, and wild fire, and reemission	Natural deposits, natural fires, sea salt spray, and volcanic eruption
Residence time of in atmosphere	7–10 days	Days to weeks	<10 days	~1 year	<10 days
Effect at cellular level	Carcinogenic and neurotoxic	Carcinogenic	Carcinogenic and mutagenic	Neurotoxic	Neurotoxic
ATSDR rank list in priority (2015)	1st	7th	17th	3rd	2th
Pure Earth rank list in priority (2015)	–	6th	5th	3rd	1st

Adapted from Rahman Z and Singh VP (2019) The relative impact of toxic heavy metals (THMs) (arsenic (As), cadmium (Cd), chromium (Cr)(VI), mercury (Hg), and lead (Pb)) on the total environment: An overview. *Environmental Monitoring and Assessment* 191: 419.

Table 2 Average total metal concentrations (mg kg⁻¹) in different parent materials.

	Arsenic (As)	Cadmium (Cd)	Chromium (Cr)	Copper (Cu)	Nickel (Ni)	Lead (Pb)	Antimony (Sb)	Uranium (U)	Zinc (Zn)
Upper crust	2	0.1	35	14	19	17	0.3	2.5	52
Granite, granodiorite	3	0.1	10	12	5	20	0.3	4	50
Gabbro, basalt	0.7	0.2	250	90	130	4	0.2	0.5	100
Ultramafic rocks	0.7	0.05	2300	40	2000	0.05	0.1	0.02	60
Sandstone	0.5	<0.04	35	2	2	10	0.05	1.3	20
Shales	13	0.25	100	45	70	22	1	3.2	100
Limestones	1.5	0.1	5	6	5	5	0.15	1	40
Coal	10	1	20	20	20	20	2	2	50

From Alloway BJ (2013) Sources of Heavy Metals and Metalloids in Soils. Springer Netherlands with permission.

However, it is often the case that some elements undergo changes in chemical forms (i.e., their chemical speciation), which importantly affects their bioavailability, fate, and subsequent risk in soils. Typically, trace elements of anthropogenic origin are generally present in more mobile and bioavailable fractions.

Sources of trace elements in polluted soils

Trace element contents in soils may be influenced strongly by geogenic or anthropogenic inputs and therefore may be long-standing or recently elevated. Pedogenetic processes of weathering of parent materials naturally result in the presence of trace elements at the levels that are regarded as trace or natural background. However, certain types of rocks have particularly large concentrations of trace elements (Table 2; Alloway, 2013) and weathering and pedogenetic processes may cause an increased concentration of some elements in soils. For example, shales generally have larger concentrations of many elements. Trace elements of geogenic origin might be detrimental to human and environmental health. Some endemic diseases are often found in an area where soils have abnormal concentrations of trace elements. For instance, elevated arsenic in drinking water and foodstuffs in some areas have caused endemic chronic poisoning by As (Chen, 2014).

Anthropogenic activities disturb or accelerate nature's slow geochemical cycles of trace elements. As a result, most soils under anthropogenic influences may accumulate one or more trace elements at significantly larger concentrations than the natural background values. This poses risks to ecosystems including plants, microorganisms and soil animals, and threatens the health of animals, humans, or other receptors. Trace elements are defined as 'a pollutant' in the soil environment where (i) the rates of introduction of trace elements into soil from anthropogenic activities are more than natural ones, (ii) via anthropogenic activities, their concentrations are larger than their natural background values, and (iii) trace elements cause adverse effects to soil quality or become transferred to other environmental receptors where the potential for direct exposure or risk are greater. Sometimes, certain trace elements are associated with specific activities, such as irrigation of crops with As-laden groundwater, while many others often result from a wider range of sources through various anthropogenic activities. The main sources of soil pollution can be divided into agricultural sources (i.e., pesticides, mineral and organic fertilizers, wastewater for irrigation, biosolids and manures) and industrial sources (i.e., mining, manufacturing and recycling industries, energy production industries, and air-borne deposition).

Pesticides

Many pesticides used extensively in agriculture and horticulture in the past and currently are sources of trace elements in soils. Of the 1800 active ingredients held in the Compendium of Pesticide Common Names, 219 contain one or more trace elements in their composition (Wood, 2021). Examples of such pesticides are Cu-based fungicidal sprays (e.g., Bordeaux mixture, CuSO₄; copper oxychloride, etc.) that are applied in vineyards and orchards. Fungicides and pesticides containing As, Cu, Mn, and Zn were used historically to protect citrus (*Citrus* spp), apples (*Malus domestica* Borkh.), peaches (*Prunus persica*), strawberries (*Fragaria x ananassa* Duchesne), and other fruit crops. Lead arsenate (AsHO₄Pb) was used in fruit orchards for many years to control parasitic insects (Alva, 1992). Organomercury compounds such as phenylmercury acetate were used as seed dressings (Byford, 1985). Chromated copper arsenate was used as a wood preservative (Hingston et al., 2001). Arsenic-containing compounds (e.g., monosodium methyl arsenate, MSMA) were also used extensively to control pests and control weeds (Matteson et al., 2014). Such applications can potentially cause soil pollution. This problem is of particular concern when contaminated soils are used for food production. With the growing awareness of risk to environmental and human health posed by pesticides, many that contain hazardous chemicals including trace elements have been banned for use. Soil pollution associated with pesticides is often local and restricted to particular sites or crops.

Fertilizers

Plants require not only macronutrients (e.g., N (nitrogen), P (phosphorus), K (potassium), S (sulfur), Ca (calcium), and Mg (magnesium)), but also essential micronutrients. These nutrients can come naturally from the soils themselves through the dissolution of minerals, desorption from minerals, decomposition of organic matter, and other processes. Some soils are deficient in essential micronutrients for plant growth and crops must be supplied with them to increase crop production. An example of such is if cereal crops are grown on Zn-deficient soils they are occasionally treated with Zn as an addition to the soil or as a foliar spray. In addition, large quantities of N, P, K fertilizers and organic fertilizers are regularly applied to soils in intensive farming systems to increase crop production. Phosphorus fertilizers are a major source of Cd in agricultural soils because they often contain relatively large concentrations of Cd that is naturally present in the raw material used (i.e., phosphate rock) for P fertilizer production (McLaughlin et al., 2021). Approximately 85% of global P fertilizer production is from sedimentary deposits in which average Cd content is about 69-fold higher than that of non-phosphate containing rocks (Van Kauwenbergh, 2010). Numerous long-term field studies have observed that continuous use of phosphatic fertilizers inadvertently increased Cd concentrations in soils in the U.S., Canada, Australia, New Zealand, and EU (European Union) countries (McLaughlin et al., 2021). Inevitably, repeated applications of high rates of P fertilizers containing large concentrations of Cd increased those of phytoavailable Cd in soils and crop Cd concentrations. Soil Cd concentrations in relation to P fertilization depend on the balance between Cd input and Cd removal over time (Fig. 1; Salmanzadeh et al., 2017). A reduction in Cd concentration in P fertilizer can reduce the increase in soil Cd concentration.

Organic manures (animal manures, compost, and sewage sludge)

Solid residues after treatment of municipal sewage sludge (biosolids), livestock manures and composts, are a valuable source of macronutrients, micronutrients as well as organic matter. The application of such organic manures can improve fertility and physicochemical properties in agricultural soils through organic carbon enrichment and the slow release of nutrients. Their application is a common practice in many countries and is also the cheapest way of recycling or disposal. In the United States, more than half of approximately 5.6×10^8 kg dry sewage sludge is applied to land annually across the country (Silveira et al., 2003). Similarly, in the European community over 30% of the sewage sludge is used as fertilizer in agriculture, and in Australia, most of the dry biosolids produced each year (over 17.5×10^6 kg) are applied to agricultural land (Silveira et al., 2003).

Organic manures also contain trace elements, leading to the accumulation of these metals in soils, e.g. As, Cd, Cu, Pb, Hg, Ni, Zn, etc. In intensive livestock production, for example, the feeds may contain essential trace elements such as Cu and Ni, and toxic trace elements such as Cr, Pb, and As. The incorporation of metals such as Cu and Zn into animal feeds is a common practice, because they can improve production efficiency. It has been reported that concentrations of some trace elements in manures increased greatly from 1990 to 2020 (Wang et al., 2014; Zhang et al., 2012). Copper sulfate, as a footbath, is often used to treat lameness in dairy cattle and also used as a widespread growth promoter in pig and poultry systems (Xiong et al., 2010). Zinc oxide is added to pig feed as a cure for diarrhea (Poulsen, 1998). Roxarsone is an arsenic-containing compound widely used as a feed additive in poultry industries to control coccidiosis and promote growth (Chapman and Johnson, 2002). These applications in the animal feed result in larger concentrations of trace elements in manures. This may potentially cause soil pollution if the manures produced from

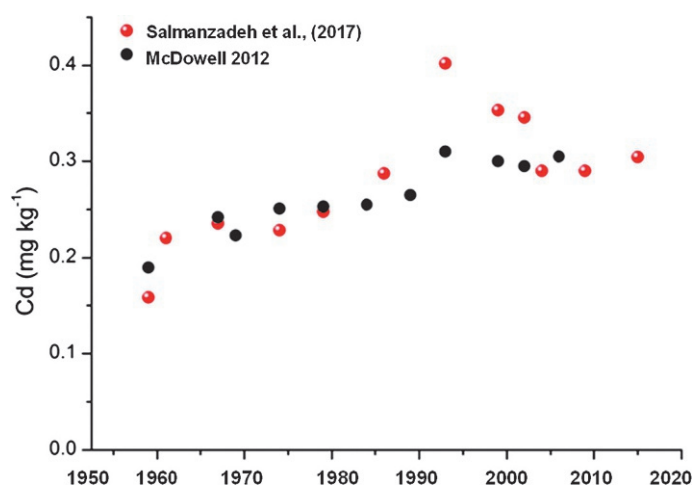


Fig. 1 The mean concentrations of Cd in topsoil samples from Winchmore research farm on the Canterbury Plains of the South Island, New Zealand, over the period 1959–2015 related to Cd-containing P fertilizers. The concentration of Cd in fertilizer has decreased since the 1980s. From Salmanzadeh M, Hartland A, Stirling CH, Balks MR, Schipper LA, Joshi C, and George E (2017) Isotope tracing of long-term cadmium fluxes in an agricultural soil. *Environmental Science & Technology* with permission.

animals on such diets are continuously applied to agricultural land. Under certain conditions, trace elements introduced to soils through applications of biosolids can be leached downwards through the soil profile and can potentially pollute groundwater (Gianico et al., 2021).

Wastewater for irrigation

Worldwide, approximately 20 million hectares of arable land are irrigated with wastewater in developing countries, especially in or close to cities, where it is plentiful (Scott et al., 2004). The types and concentrations of trace elements in the wastewater depend on the sources of effluents and the technologies used to treat them (Scott et al., 2004). Wastewater has been widely used for irrigation in agriculture, especially in developing countries. A few countries have reported the direct use of untreated wastewater for irrigation, and for example the volume reaches 4 million cubic meters per year in Mexico. Although wastewater provides low-cost resources for irrigation, irrigation with untreated wastewater can contribute to the accumulation of trace elements, and organic contaminants such as persistent organic pollutants (POPs) and perfluoroalkyl and polyfluoroalkyl substances (PFAS), as well as a source of dangerous pathogens, and pharmaceuticals and personal care products in soils (Islam et al., 2018). Crops irrigated with untreated wastewater can potentially accumulate toxic trace elements in their edible parts at an elevated concentration which pose risks to human health.

Mining processes

Mining is a potential major source of trace elements in soils, not just from the mining operation itself, but mostly from wastes and emissions during the processing of the extracted materials such as in tailings, waste rock deposit, and from smelting operations (Table 3) (FAO and UNEP, 2021). Mining waste often contains one or more trace elements at concentrations at which extraction is not cost effective, but they may still pose a risk to the environment and human health. Mining to exploit coal, gold (Au), uranium (U), tungsten (W), tin (Sn), platinoids and polymetallic sulfides, have caused the most severe cases of soil pollution, according to the reports by the FAO and ITPS (FAO and ITPS, 2015). Mercury is often used for the extraction of gold and consequently gold mining operations often result in the emission of mercury. It is estimated that large-scale gold mining is responsible for about 2.7×10^6 kg of mercury entering the soil each year (UNEP, 2019). Tailings are made of water and fine mineral particulates and are often constructed on site by means of a dam. They may cause soil pollution when the tailings dam fails or from wind erosion that disperses the fine mineral particulates from the surface. This often leads to soil pollution of areas at long distances from the sources. There have been many accidents involving ruptures of tailings dams historically, releasing toxic waste rich in As, Cd, Cu, Pb, and Zn that then accumulated in soils or reached surface waters. In addition, extensive Pb and Zn ore mining and smelting have often resulted in soil pollution from water or wind erosion of deposits of ore and rock waste (Hu et al., 2019). Where mining wastes that contain reduced sulfur minerals (e.g., pyrites) are exposed to oxygen and water, acid mine drainage (AMD) tends to form. This can lead to the formation of large volumes of highly acidic wastewaters containing large concentrations of sulfates, iron, and trace elements. In the event of a leakage, AMD could lead to the pollution of water bodies and subsequent transfer to soil. Spillage during the transfer and transport of ore concentrates to the smelter is another potential source of soil pollution. Particulate and gaseous emissions from a smelter that lacks appropriate control technologies are a local and long-distance source of trace elements to soils. For instance, worldwide 85–90% of total atmospheric Cd emissions arise from anthropogenic sources and most of the Cd in the atmosphere is associated with suspended particulate matter (aerosols) in the form of oxides, sulfates, and chlorides (McLaughlin et al., 2021). In Europe, Cd emissions decreased markedly from 2500 Mg year⁻¹ in the 1960s to 68–191 Mg year⁻¹ between 1990 and 2017 with the installation of more efficient flue gas desulfurization equipment in smelters and power plants (McLaughlin et al., 2021). This downward trend in atmospheric Cd concentrations have caused a marked decrease in deposition of Cd into soils.

Table 3 Trace elements likely to occur in soils in contaminated environments.

Trace elements	Industrial sources
Antimony (Sb)	Electronics, metal smelting
Arsenic (As)	Insecticide manufacture and use, wood preservation
Beryllium (Be)	Nuclear, munitions, and electronics
Cadmium (Cd)	Pigments, alloys, electroplating, batteries, mining, and smelting industries
Chromium (Cr)	Plating, tanning, inks, wood treatment, pigments, and dyes
Copper (Cu)	Mining and smelting, wood treatment, electronics, plumbing, copper and brass manufacture
Lead (Pb)	Batteries, tetraethyl lead production, mining, smelting, plumbing
Mercury (Hg)	Chemical plant, mining and smelting, refining gold and silver, electrolysis industry, chlor-alkali plant, paper mills
Nickel (Ni)	Plating and metal works, batteries, iron and steel industries, electronics
Selenium (Se)	Smelting, paint, pigment, electronics
Silver (Ag)	Mining, photographic, electroplating, silver manufacture
Thallium (Tl)	Cement industry, electronics
Zinc (Zn)	Mining and smelting, plating, galvanizing, runoff from galvanized surfaces, print and paper

Manufacturing and recycling industries

The textile industry may also release toxic trace elements in effluents. Large amounts of wastewater, produced in the dyeing and finishing process often contain trace elements including As, Cd, Co, Cu, Pb, Hg, and Ni (Table 3). Leather manufacture and tanneries produce large amounts of solid and liquid by-products including Cr(III) compounds, thus tannery effluents are highly enriched in Cr (Alvarez-Bernal et al., 2006). During the process, Cr(III) can be oxidized to its much more toxic hexavalent state (chromate, Cr(VI)). The tanning process has caused many cases of Cr pollution in soils. Soil pollution by chromate-containing tanning effluents may contaminate water bodies or underground waters because of the greater mobility of chromate than trivalent chromium.

The manufacturing and recycling of Pb-acid and Ni-Cd batteries are important sources of trace element pollution in soils (McLaughlin et al., 2021). About 25,000 Mg of Cd were produced in 2017 of which 80% of the Cd produced was used for the manufacture of batteries (U.S. Geological Survey, 2019). The manufacturing of Pb-acid batteries consumes about 85% of Pb worldwide and is growing, especially in the Asian market with the rapid development of industries that use Pb-acid batteries such as electric bikes, electric power vehicles, and photovoltaic devices. Pollution occurs during the production and recycling of batteries when factories lack proper pollution control and regulatory oversight. For example, the production of Pb-acid batteries can generate and release fine Pb dust through atmospheric dispersion, leading to its deposition on soils (FAO and UNEP, 2021). For example, a survey from lead battery manufacturing and recycling in seven African countries has demonstrated that lead concentrations were 26,000 mg kg⁻¹ in soils surrounding the plants, posing significant health risks to the public (Gottesfeld et al., 2018).

Energy production industries

Coal-based energy production generates a large volume of wastes, including fly ash, bottom ash, boiler slag, and flue gas. Coal wastes in landfills and surface impoundment ponds often result in frequent leaching and exposure of materials to weathering and erosion due to the lack of adequate pollution control measures. There are many examples of polluted soils around coal waste deposits. Coal fly ash contains silicon, aluminum and trace elements such as Cd, As, Cu, Hg, and other organic contaminants such as PAHs (polycyclic aromatic hydrocarbons). The concentrations of trace elements vary with the type of coal and are often found to be positively correlated with sulfur concentrations (FAO and UNEP, 2021). Atmospheric deposition from coal combustion is an important pathway by which trace elements accumulate in soils (Bian et al., 2009). Coal combustion largely contributes mercury emissions to the soil, mainly through airborne particulates. Some trace metals such as As, Cd, Pb can also volatilize during high-temperature processing and then convert to oxides and condense as fine particulates (Wuana and Okieimen, 2011). The airborne particulates are eventually deposited on land or water bodies and thus most forms of fossil fuels are an important form of soil pollution. For example, 85–90% of total atmospheric Cd emissions globally arise mainly from fossil fuel combustion, nonferrous metal manufacturing, iron and steel production, cement production, and waste incineration; they are an important source of soil pollution (WHO, 2011).

Characteristics of trace elements in soils and potential risk pathways

Trace elements are commonly found in polluted soils, with most reports being for Cd, As, Cr, Hg, Pb, Ni, Zn, and Cu. Soil pollution with potentially toxic trace elements may pose risks and hazards to ecosystems, food safety, and humans through many pathways, depending on the basic chemistry of the trace elements (Table 4 and Fig. 2). The ecotoxicological effects and risks of trace elements not only depend upon their total concentrations in soils, but also upon their bioavailability. The concept of a 'soil–plant barrier' (Chaney, 1980; Zhao and Wang, 2020), proposed to describe the ability of trace elements to enter the food chain, can be used to divide trace elements into four categories. The first category has very low solubility in the soil so that their bioavailability to plants is also very low; these elements include Pb, Hg and Cr(III). The second category is those elements that can be strongly sorbed on root surfaces or strongly chelated in the root cells so that their translocation to other parts of the plant is very limited; elements such as Hg and Pb. The third category can cause phytotoxicity and decrease crop production, and so the damaged crops are less likely

Table 4 Critical risk pathways of trace elements.

<i>Risk Pathways</i>	<i>Potential trace elements</i>	<i>Risk</i>
Soil → Plant	Zn, Cu, Ni, etc.	Ecosystem
Soil → Micro- or Macroorganism	Zn, Cu, Ni, etc.	Ecosystem
Soil → Plant → Human	Cd, As, MeHg	Human health
Soil → Plant → Animal → Human	Cd, As, MeHg	Human health
Soil → Underground water → Human	Cr(VI), As	Underground water
Soil → Oral, Skin → Human	Pb	Human health

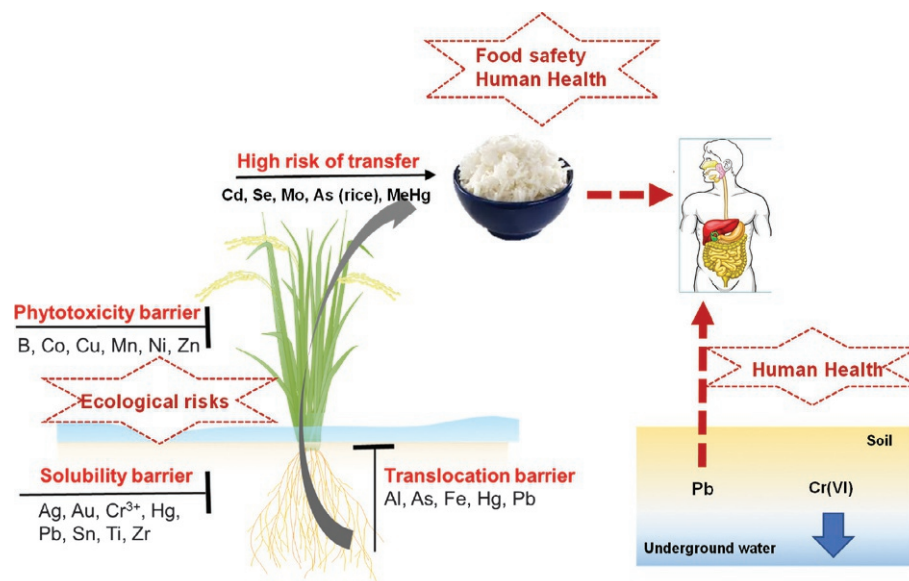


Fig. 2 Soil-plant barriers and the risk of toxic element transfer to the food chain in the paddy rice system.

to be consumed by humans or domestic animals; elements include Cu, Zn, Ni and As. The fourth category has relatively strong solubility in soils and high translocation in plants and causes toxicity to humans at levels that do not cause phytotoxicity; the elements include Cd, selenium (Se), molybdenum (Mo) and As. However, Se and Mo rarely cause toxicity and in contrast, their deficiencies in animals and humans are the more common. Cadmium is the most important toxic trace metal threatening global food safety through the soil to the food chain (McLaughlin et al., 2021). Arsenic also has a high risk of soil-food chain transfer in paddy rice systems, a worldwide concern threatening rice production and food safety (Meharg et al., 2009; Meharg and Zhao, 2012). Rice is the most important contributor to dietary Cd and As intake and sometimes methylmercury (MeHg, CH_3Hg^+) (see references in Zhao et al. (2022)).

In view of the risk to human health associated from soil pollution, the critical pathways include direct ingestion or contact with contaminated soil (e.g. Pb; soil-oral, skin), the food chain (e.g. Cd, As and MeHg; soil-plant-human or soil-plant-animal-human), drinking of contaminated groundwater (e.g. Cr(VI), As), and reduction in soil functions or quality through ecotoxicity (e.g. Zn, Cu, Ni, etc.) (Table 4).

Summary

Trace elements, including essential and toxic elements, occur naturally in soils related to geological and pedological processes. Anthropogenic activities disturb or accelerate the accumulation of trace elements in soils, resulting in soil pollution. Sources of trace elements that lead to soil pollution are very varied and range from various agricultural sources to industrial activities. To prevent and reduce soil pollution with trace elements, great effort should be made to reduce the emission of trace elements into soils by reducing the production and use of toxic trace elements in agricultural activities, and in minimizing emissions from industries or by shifting industrial production toward more sustainable systems in which the emissions are reduced. Trace elements differ largely in their potential to pose risks to ecosystems, food safety, and humans. To reduce soil pollution by trace elements, prevention is the first and foremost action against pollution. Scientifically, further research is needed to fill knowledge gaps in the pollution cycle. Legislative actions and monitoring networks are also required to be strengthened. Finally, public awareness and communication should be improved.

References

- Alloway BJ (2013) *Sources of Heavy Metals and Metalloids in Soils*. Netherlands: Springer.
- Alva AK (1992) Micronutrients status of Florida soils under citrus production. *Communications in Soil Science and Plant Analysis* 23: 2493–2510.
- Alvarez-Bernal D, Contreras-Ramos SM, Trujillo-Tapia N, Olalde-Portugal V, Frías-Hernández JT, and Dendooven L (2006) Effects of tanneries wastewater on chemical and biological soil characteristics. *Applied Soil Ecology* 33: 269–277.
- Bian Z, Dong J, Lei S, Leng H, Shouguo M, and Wang H (2009) The impact of disposal and treatment of coal mining wastes on environment and farmland. *Environmental Geology* 58: 625–634.
- Byford WJ (1985) A comparison of fungicide seed treatments to improve sugar beet seedling establishment. *Plant Pathology* 34: 463–466.

- Chaney RL (1980) Health risks associated with toxic metals in municipal sludges. In: Bitton G, Damron BL, Edds GT, and Davidson JM (eds.) *Sludge: Health Risks of Land Application*, Ann Arbor, MI.
- Chapman HD and Johnson ZB (2002) Use of Antibiotics and Roxarsone in Broiler Chickens in the USA: Analysis for The Years 1995 to 2000. *Poultry Science* 81: 356–364.
- Chen CJ (2014) Health hazards and mitigation of chronic poisoning from arsenic in drinking water: Taiwan experiences. *Reviews on Environmental Health* 29: 13–19.
- FAO and ITPS (2015) *Status of the World's Soil Resources (SWSR): Main Report*. Rome.
- FAO and UNEP (2021) *Global Assessment of Soil Pollution: Report*. Rome.
- Gianico A, Braguglia CM, Gallipoli A, Montecchio D, and Mininni G (2021) Land Application of Biosolids in Europe: Possibilities, Con-Strains and Future Perspectives. *Water* 13: 103.
- Gottesfeld P, Were FH, Adogame L, Gharbi S, San D, Nota MM, and Kuepouo G (2018) Soil contamination from lead battery manufacturing and recycling in seven African countries. *Environmental Research* 161: 609–614.
- Hingston JA, Collins CD, Murphy RJ, and Lester JN (2001) Leaching of chromated copper arsenate wood preservatives: A review. *Environmental Pollution* 111: 53–66.
- Hu J, Lin B, Yuan M, Lao Z, Kangming W, Zeng Y, Liang Z, Li H, Li Y, Zhu D, Liu J, and Fan H (2019) Trace metal pollution and ecological risk assessment in agricultural soil in Dexing Pb/Zn mining area, China. *Environmental Geochemistry and Health* 41: 967–980.
- Islam MA, Davor Romić M, Akber A, and Romić M (2018) Trace metals accumulation in soil irrigated with polluted water and assessment of human health risk from vegetable consumption in Bangladesh. *Environmental Geochemistry and Health* 40: 59–85.
- Matteson AR, Gannon TW, Jeffries MD, Haines S, Lewis DF, and Polizzotto ML (2014) Arsenic retention in foliage and soil after monosodium methyl arsenate (MSMA) application to turfgrass. *Journal of Environmental Quality* 43: 379–388.
- McLaughlin MJ, Smolders E, Zhao FJ, Grant C, and Montalvo D (2021) Chapter one - managing cadmium in agricultural systems. In: Sparks DL (ed.) *Advances in Agronomy*. Academic Press.
- Meharg AA and Zhao F-J (2012) *Arsenic & Rice*. Dordrecht: Springer Netherlands.
- Meharg AA, Williams PN, Adomako E, Lawgali YY, Deacon C, Villada A, Cambell RCJ, Sun G, Zhu Y-G, Feldmann J, Raab A, Zhao F-J, Islam R, Hossain S, and Yanai J (2009) Geographical variation in total and inorganic arsenic content of polished (white) rice. *Environmental Science & Technology* 43: 1612–1617.
- Poulsen HD (1998) Zinc and copper as feed additives, growth factors or unwanted environmental factors. *Journal of Animal and Feed Sciences* 7: 135–142.
- Rahman Z and Singh VP (2019) The relative impact of toxic heavy metals (THMs) (arsenic (As), cadmium (Cd), chromium (Cr)(VI), mercury (Hg), and lead (Pb)) on the total environment: An overview. *Environmental Monitoring and Assessment* 191: 419.
- Salmanzadeh M, Hartland A, Stirling CH, Balks MR, Schipper LA, Joshi C, and George E (2017) Isotope tracing of long-term cadmium fluxes in an agricultural soil. *Environmental Science & Technology* 51: 7369–7377.
- Scott CA, Faruqi NI, and Raschid-Sally L (2004) *Wastewater Use in Irrigated Agriculture: Confronting the Livelihood and Environmental Realities*. Wallingford: CABI.
- Silveira M, Alleoni L, and Guilherme L (2003) Biosolids and heavy metals in soils. *Scientia Agricola* 60.
- U.S. Geological Survey (2019) Mineral Commodity Summaries. <https://www.usgs.gov/centers/nmic/cadmium-statistics-and-information>.
- UNEP (2019) *Global Mercury Assessment 2018*. Switzerland.
- Van Kauenbergh SJ (2010) *World Phosphate Rock Reserves and Resources*. Muscle Shoals: International Fertilizer Development Centre (IFDC).
- Wang H, Dong Y, and Wang H (2014) Hazardous metals in animal manure and their changes from 1990 to 2010 in China. *Toxicological & Environmental Chemistry* 96: 1346–1355.
- Cadmium. WHO (ed.) (2011) *WHO Air Quality Guidelines for Europe*, edited by WHO Regional Office for Europe. Copenhagen.
- Wood AM (2021) Compendium of Pesticide Common Names. <https://pesticidecompendium.bcp.org/index.html>.
- Wuana RA and Okieimen FE (2011) Heavy Metals in Contaminated Soils: A Review of Sources, Chemistry, Risks and Best Available Strategies for Remediation. *ISRN Ecology* 2011: 402647.
- Xiong X, Yanxia L, Wei L, Chunye L, Wei H, and Ming Y (2010) Copper content in animal manures and potential risk of soil copper pollution with animal manure use in agriculture. *Resources, Conservation and Recycling* 54: 985–990.
- Zhang F, Li Y, Yang M, and Li W (2012) Content of heavy metals in animal feeds and manures from farms of different scales in northeast china. *International Journal of Environmental Research and Public Health* 9: 2658–2668.
- Zhao F-J and Wang P (2020) Arsenic and cadmium accumulation in rice and mitigation strategies. *Plant and Soil* 446: 1–21.
- Zhao F-J, Tang Z, Song J-J, Huang X-Y, and Wang P (2022) Toxic metals and metalloids: Uptake, transport, detoxification, phytoremediation, and crop improvement for safer food. *Molecular Plant* 15: 27–44.

Pollutants—Persistent organic

Rai S Kookana and Divina A Navarro, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Glen Osmond, SA, Australia

© 2023 Elsevier Ltd. All rights reserved.

Introduction	296
Persistent organic pollutants (the conventional and new POPs)	297
Conventional POPs	298
New POPs (PFAS)	298
Physicochemical properties of POPs in soils	298
Conventional POPs	298
New POPs (PFAS)	299
Fate and behavior of POPs in the soil environment	300
Persistence of organic chemicals in soils	300
Conventional POPs	300
New POPs (PFAS)	300
The 'ageing' of POPs within soils	301
Sorption-related mechanisms	301
Sorption retarded intraparticle diffusion	302
Intraorganic matter diffusion	303
Additional mechanisms involved in sorption of new POPs (PFAS)	304
The (Bio)availability and bioaccumulation of POPs in soils	304
Availability to plants	306
Availability to soil microflora	306
Availability to grazing animals	306
Availability to earthworms	307
Concluding remarks	307
References	307

Abstract

Pollution can lead to disease and premature death in the world today. In 2015, about half a million deaths were estimated to be associated with soil pollution. Many organic pollutants can persist in the soil environment for a long time and are referred to as persistent organic pollutants (POPs). This chapter provides an overview of how the chemical properties and fundamental processes of conventional POPs (e.g., polyaromatic hydrocarbons, polybrominated diphenyl ethers) as well as new POPs (e.g., Per- and Poly-fluoro Alkyl Substances—PFAS) control their fate and behavior in the soil environment and determine their potential impact on biota.

Key points

- The environmental burden of persistent organic pollutants (POPs) is growing.
- Per- and Poly-fluoro Alkyl Substances (PFAS) are new POPs that are now ubiquitous in the environment.
- PFAS have markedly different properties from conventional POPs (e.g., PAHs, PBDEs) and many PFAS being hydrophilic are more mobile than the conventional POPs in the environment.
- The conventional approaches to assess fate and behavior of POPs (e.g., models based on octanol: water partition coefficient $-K_{ow}$) do not apply to PFAS and need to be refined.
- Currently, the fate and behavior of PFAS in the soil environment is poorly understood.

Introduction

"Pollution is the largest environmental cause of disease and premature death in the world today", according to The Lancet Commission on Pollution and Health (Lancet, 2018). In 2015, about half a million deaths were estimated to be associated with soil pollution. Deaths caused by pollution tend to be disproportionately higher in low- to middle-income countries, as well as among minorities and marginalized sections of society.

Organic pollutants are ubiquitous in the environment and their environmental burden is continuously growing, largely because of anthropogenic activities and the synthesis of myriad chemicals over last two centuries. A wide range of organic chemicals in industry, such as polyaromatic hydrocarbons, flame retardants, pesticides can find their way to the soil environment after their use.

Soil can be a major sink for many of the organic pollutants released into the environment. These compounds enter the environment through a wide variety of processes including inefficient combustion, dry and wet deposition from the atmosphere, widespread pesticide use, and wastewater drainage from industrial and domestic sources into sewers or directly into the soil environment. Once in a soil, high-molecular-weight non-ionic organic compounds may persist for decades.

Many organic pollutants can persist in the soil environment for a long time (e.g., the pesticide DDT is still detected in soil environment despite its widespread application was stopped some 40 years ago) due either to their inherent chemical properties not favoring degradation (e.g., strength of the chemical bonds) or to their binding to the soil solid phase (organic and mineral matter). Strong binding and sequestration can protect them from breakdown by soil microorganisms. These persistent chemicals are generally referred to as persistent organic pollutants (POPs).

Soil POPs are potentially bioavailable to many organisms, including soil microflora, macrofauna (e.g., earthworms), plants, animals and humans (Jones, 2021). Soil POPs may be taken up by plants and other organisms, enter the food chain and thus impact human and ecosystem health. Therefore, the bioavailability of POPs to organisms is important in determining the potential hazard presented by a soil contaminated by POPs. Their mobility and bioavailability may also influence the effectiveness of soil remediation measures.

A range of properties of chemicals (e.g., hydrophobicity, electrostatic charge, volatility) control their distribution among, and impact on, various environmental compartments. Several fundamental processes (e.g., sorption and desorption) governing the distribution of POPs among the various phases of the soil (i.e., solid-, air- and aqueous-phase) determine their mobility, bioavailability and potential environmental impact.

This chapter provides an overview of how the chemical properties of conventional POPs (e.g., polyaromatic hydrocarbons, organochlorine pesticides) and new POPs (e.g., Per- and Poly-fluoro Alkyl Substances—PFAS) control their fate and behavior in the soil environment. Being so unique in their properties, PFAS are being discussed separately, where appropriate, as new POPs.

The chapter draws heavily from the previous version by D. Johnson (Rothamsted Research, Harpenden, United Kingdom) published in the first edition of this Encyclopedia (Johnson, 2005). The original article has now been updated with a new class of POPs, namely the Per- and Poly-fluoro Alkyl Substances (PFAS) due to their extreme persistence and bioaccumulation. Some PFAS (PFOS and PFOA) have already been added to the Stockholm Convention List of POPs.

Persistent organic pollutants (the conventional and new POPs)

Since the 19th century, humans have produced and released vast quantities of organic compounds into the biosphere. These compounds are widely distributed throughout the environment and many are considered carcinogenic and mutagenic. This has led to the prioritization of organic pollutants by the United States Environmental Protection Agency (USEPA) and the European Community (EC). These lists include polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs), polychlorinated dibenzo-*p*-dioxins and furans (PCDD/Fs), chlorobenzenes (CBs), phthalate acid esters, pesticides and other miscellaneous compounds. Many members of these classes of chemicals are extremely persistent. According to the Stockholm Convention on Persistent Organic Pollutants (POPs), the “POPs are organic compounds that are resistant to environmental degradation through chemical, biological, and photolytic processes”. The Stockholm Convention is an international treaty to protect human health and the environment from POPs.

In 2001, a group of 12 chemicals were identified as POPs (the initial POPs). However, in 2009 it was decided to add 16 new chemicals to this list. These included some chemicals used as pesticides (e.g., Lindane, Endosulfan), flame retardants (some polybrominated diphenyl ethers—PBDEs), and as water repellents and fire-fighting foams (e.g., perfluorooctane sulfonic acid—PFOS). The current list of all POPs plus information on their use, production and replacement options is available at the Stockholm Convention’s website (the link is provided at the end of this chapter). Among these, the per- and poly-fluoroalkyl substances (PFAS) are a large group of chemicals which are quite unique in their chemistry and are very different from the original POPs identified in 2001.

The PFAS have emerged during the last decade as one of the major soil pollutants around the globe (Wang et al., 2017). According to OECD, nearly 5000 chemicals belong to PFAS family; however, only 256 of these are currently commercially relevant (Buck et al., 2021). Many PFAS are considered both extremely persistent and extremely mobile in the soil environment (e.g., anionic or ionizable PFAS). Their high mobility and extreme persistence together lead to their ubiquitous presence in soil and water and potential exposure to humans and wildlife. Some PFAS, such as PFOS, PFOA, PFHxS (and their salts and associated substances) have either been placed on the Stockholm Convention list of priority pollutants (A, B or C) or are under consideration for such classification. Agreement on phasing out of PFOS and PFOA (where possible) has already been reached under the Stockholm Convention in 2009 and 2019, respectively. However, alternatives are currently not available for some of the products containing PFOS and PFOA.

Conventional POPs

A majority of POPs (e.g., PAHs, PCBs, PCDD/Fs, CBs and organochlorine pesticides such as DDT, dieldrin, heptachlor, hexachlorobenzene) are hydrophobic in nature. They are characterized by their relatively high octanol/water partition coefficient ($\log K_{ow}$) and low water solubility. Thus, once in a soil they tend to be resistant to microbial degradation and prone to bioaccumulation. Atmospheric transport processes such as 'global distillation' result in an almost ubiquitous contamination of soils, often thousands of kilometers from major POP sources. The highest soil concentrations of POPs are usually found near industrial point sources, urban areas, and in organic-rich soils, as POPs tend to accumulate in organic matter. The disused sites in the manufacturing industry in many western economies have left large areas of soil contaminated by such POPs and requiring remediation before they can be re-developed.

New POPs (PFAS)

Since commencement of their production in the 1950s, PFAS have been extensively used in a range of industrial (e.g., production of Gore-Tex™, Teflon™, electronic goods, metal plating) and commercial applications (e.g., stain- and water-resistant textiles, food packaging) as well as in military training activities (e.g., in fire-fighting). The PFAS class also includes polymers (PTFE or Teflon™) which are also used in stain- and water-resistant textiles and food packaging. The non-polymer surfactant PFAS are used as aqueous film-forming foams (AFFFs) to extinguish fires and are more abundant in the environment, due to their regular use at airports and defence facilities, for example. Consequently, PFAS are now ubiquitous in various environmental compartments, such as soil and water, including in oceans (Krafft and Riess, 2015a; Krafft and Riess, 2015b), and have been found to bioaccumulate and biomagnify (Giesy and Kannan, 2001). The accumulation of PFAS in soils reported from various countries is apparent from the data compiled in Table 1 (Brusseau et al., 2020).

The PFASs generally have a hydrophobic 'tail' that contains a high proportion of fluorine (C—F chain), a hydrophilic functional group (e.g., carboxylate, sulfonate) and in many cases a 'spacer' organic group linking these two portions of the compound together. Replacing the hydrogen atoms (in a hydrocarbon chain in organic molecules) by fluorine atoms allowed a range of unique chemical properties that enabled this group of chemicals to be used for a diverse range of functions and applications mentioned above (Krafft and Riess, 2015a).

Physicochemical properties of POPs in soils

Conventional POPs

The principal conventional POPs found in soils are PAHs, PCBs, CBs, and PCDD/Fs. Table 2a displays comparative physicochemical data for several selected POPs that occur in contaminated soils. Clearly, these compounds are highly hydrophobic in nature (high values of $\log K_{ow}$) and consequently they readily partition onto the soil organic carbon (OC) phase. Therefore, their sorption or partition coefficient (K_d) in soils is often normalized against the fraction of OC in soils (i.e., $K_{oc} = K_d/\text{fraction of OC}$). The $\log K_{oc}$ values in Table 2a show a strong correlation with their K_{ow} values.

Table 1 Per- and poly-fluoro alkyl substances (PFAS) concentrations reported in soil at primary-source contaminated sites in various countries.

Locations	Type of site (number of sites)	Maximum PFOS ^a Concentration $\mu\text{g kg}^{-1}$	Maximum PFOA ^a Concentration $\mu\text{g kg}^{-1}$
Australia	Airports (6)	84,200	6400
	Fire training area (unspecified)	460,000	3200
Belgium	PFAS manufacturing (1)	7800	114
Canada	Crash site (1)	9.3	29
China	PFAS manufacturing (1)	2583	50
	PFAS Industrial Park (1)	0.4	5.3
Norway	Fire training area (4)	8924	141
	Fire training area (6 airports)	17,400	75
Sweden	Fire training area (4)	8520	219
United States	Fire training area (1)	36,534	11,484
	AFFF ^b Source Zones (10 military installations)	9700	58
	AFFF Source Zones (Many military installations)	373,000	50,000
Global	Median	8722	83
Background level	Median	2.7	2.7

^aPFOS: Perfluorooctane sulfonic acid; PFOA: Perfluorooctanoic acids.

^bAqueous film forming foam (AFFF).

Based on Brusseau ML, Anderson RH and Guo B (2020) PFAS concentrations in soils: Background levels versus contaminated sites. *Science of the Total Environment* **740**: 140017.

Table 2a Physicochemical properties of conventional POPs (e.g., selected chlorobenzenes, polycyclic aromatic hydrocarbons, polychlorinated biphenyls, and polychlorinated dibenzo-*p*-dioxins and furans).

Compound/Acronym	Chemical formula	Log K_{ow}	Log K_{oc}	$K' H$	Log K_{oa}
1,4-DCM	CH_2Cl_2	3.5	3.43	6.46×10^{-2}	4.69
HCB	C_6Cl_6	5.5	5.25	2.72×10^{-2}	7.07
Naphthalene	$C_{10}H_8$	3.37	3.14	1.74×10^{-2}	5.13
Phenanthrene	$C_{14}H_{10}$	4.57	4.4	1.31×10^{-3}	7.45
Benzo(a)pyrene	$C_{20}H_{12}$	6.04	6.24	1.86×10^{-5}	10.77
Coronene	$C_{24}H_{12}$	6.75	6.4	1.72×10^{-7}	13.51
2,3,7,8-TCDD	$C_{12}H_4Cl_4O_2$	6.8	6.2	1.35×10^{-3}	9.67
OCDD	$C_{12}Cl_8O_2$	8.2	6.97	2.76×10^{-4}	11.76
PCB-18	$C_{12}H_7Cl_3$	5.72	5.15	1.95×10^{-2}	7.43
PCB-52	$C_{12}H_6Cl_4$	5.83	4.99	1.75×10^{-2}	7.59
PCB-101	$C_{12}H_5Cl_5$	6.3	6.3	8.74×10^{-3}	8.36
PCB-138	$C_{12}H_4Cl_6$	6.69	6.24	1.00×10^{-3}	8.69
PCB-206	$C_{12}HCl_9$	7.94	6.66	5.59×10^{-3}	10.19
PentaBDE (BDE-47; BDE-99)	$C_{12}H_6Br_4O$; $C_{12}H_5Br_5O$	6.64–6.97	4.89–5.10	1.20×10^{-5}	—
OctaBDE (BDE-203)	$C_{12}H_2Br_8O$	6.29	5.92–6.22	7.50×10^{-8}	—
DecaBDE (BDE-209)	$C_{12}Br_{10}O$	6.26	6.80	1.62×10^{-6}	—

log K_{ow} , octanol/water partition coefficient; log K_{oc} , octanol/carbon coefficient; $K' H$, Henry's constant; log K_{oa} , octanol/air coefficient.

Taken from Johnson D (2005) Pollutants/persistent organic (POPs). In: Hillel D (ed.), *Encyclopedia of Soils in the Environment*, pp. 264–271. Elsevier.

Table 2b Physicochemical properties of selected PFAS.

Acronym	Chemical formula	CAS number	pK_a (range) ^a	Log K_{oc} ^b	K_{aw} ^c
PFBA	$C_3F_7CO_2^-$	45048-62-2	0.169–0.444	1.88	1.72E-06
PFPeA	$C_4F_9CO_2^-$	45167-47-3	0.432–0.638	1.37	5.84E-06
PFHxA	$C_6F_{11}CO_2^-$	92612-52-7	0.745–0.921	1.31	2.20E-05
PFHpA	$C_6F_{13}CO_2^-$	120885-29-2	1.7–2.1	1.63	5.76E-05
PFOA	$C_7F_{15}CO_2^-$	45285-51-6	–0.5–3.8	2.02	2.32E-04
PFNA	$C_8F_{17}CO_2^-$	72007-68-2	<1.6 ^d	2.43	9.34E-04
PFDA	$C_9F_{19}CO_2^-$	73829-36-4	2.575	2.93	3.72E-03
PFUnA	$C_{10}F_{21}CO_2^-$	196859-54-8	2.606	3.48	1.90E-02
PFBS	$C_4F_9SO_3^-$	45187-15-3	<1.6 ^d	1.41	1.78E-05
PFHxS	$C_6F_{13}SO_3^-$	108427-53-8	<1.6 ^d	1.96	9.72E-05
PFOS	$C_8F_{17}SO_3^-$	45298-90-6	–3.03 ^d	2.79	2.30E-03
6:2 FtS	$C_6F_{13}C_2H_4SO_3^-$	27619-97-2	~–2.00 ^d	N.A.	N.A.
6:2 FtSaB	$C_{15}H_{19}F_{13}N_2^+O_4S^-$	34455-29-3	~1.8; ~10.1 ^d	N.A.	N.A.
6:2 FtSaAm	$C_6F_{13}C_2H_4S(O)_2N^+(H)C_3H_6N(CH_3)_2$	1383438-86-5	~9.2; ~10.7 ^d	N.A.	N.A.

N.A.: not applicable due to complex chemistry and mechanisms involved or not available.

Note: Due to their surface activities, PFAS tend to concentrate at the octanol/water interface and thus accurate values of K_{ow} of PFAS are not available. Furthermore, due to their interfacial interaction their adsorption at the air/water interface (K_{aw}) can be a significant phenomenon for their retention in soils.

^aExperimental values of the acid dissociation constants (pK_a) are taken from Burns et al., 2008

^bThe literature K_{oc} values compiled by Brusseau, 2019a.

^cThe K_{aw} (air/water adsorption coefficient) values reported by Brusseau, 2019b.

^dValues estimated through predictive models (Burns et al., 2008).

New POPs (PFAS)

The PFAS, a large group of substances (polymers and non-polymers) including solids, liquids and gases, have a broad range of physical, chemical and biological properties (Buck et al., 2021). The chemical properties of PFAS are very different from those of conventional POPs (Table 2b). The PFAS are aliphatic substances characterized by a partially or fully fluorinated alkyl chain and a functional head-group (such as carboxylates, sulfonates, sulfonamides, phosphonates and alcohols) (Buck et al., 2021). Properties of these chemicals include their exceptional thermal and chemical stability—attributable to the strength of the C—F bond in the fluoroalkyl tail, stronger polarization and greater hydrophobicity than alkyl chains of comparable length. The addition of functional head-groups increases their water solubility. Consequently, the functionalized fluorochemical has surfactant properties because of both hydrophobic and hydrophilic moieties (Krafft and Riess, 2015a). The key features of fluorinated surfactants include their

surface activity in both aqueous and organic solvent systems, reduced surface tension, which results in superior wetting, spreading, and levelling properties on all types of surfaces, effective emulsification in specialty applications, and extreme stability both chemically and thermally (Krafft and Riess, 2015a).

The PFAS are unique due to the properties of the C—F bond, with high electronegativity and polarity (Krafft and Riess, 2015a). Therefore, many of the PFAS compounds (e.g., carboxylates and sulfonates) are present in the anionic form over the environmentally relevant pH range (4–9). These PFAS are highly water-soluble and consequently have a high mobility in the environment. The most common detections in wastewater treatment, fresh water and ground water systems have been of anionic PFAS such as perfluorooctane sulfonic acid (PFOS) and perfluorooctanoic acid (PFOA). The acidic head-group makes them hydrophilic, whereas the tail made of the C—F chain makes them hydrophobic and consequently their partitioning properties also depend on the length of the C—F chain (Table 2b). On the other hand, several PFAS contain amide and sulfonamido head-groups and are cationic in nature. Furthermore, some PFAS (e.g., fluorotelomer sulfonamido alkylbetaines, FTABs) are zwitterionic. For example, perfluorooctane amidoalkyl betaine (PFOAB) contains other ionizable functional groups in addition to a positive charge in its structure ($-N^+(CH_3)^2-$); this theoretically results in four species with different charges. Clearly, PFAS have more complex chemistries than conventional POPs.

Fate and behavior of POPs in the soil environment

Legislation relating to organic compounds in soil is largely based on toxicological criteria with the aim of protecting wildlife and human health. Thus, it is important that the ‘availability’ of compounds to transfer from the soil to organisms is understood in a quantitative manner. Recent research on the fate and behavior of organic chemicals in soils suggests that the availability of an organic compound from the soil into the food chain may depend markedly on the form in which a compound is present in a soil and how it has interacted with the soil solid-phase. This is influenced by several factors, including its residence time within the soil, physicochemical properties of the compound and soil, and how it entered the soil. Desorption of sorbed compounds into the soil solution enhances biodegradation, leaching, volatilization, plant uptake and other chemical or photochemical transformations. Whilst these processes are generally enhanced by the chemical being present in the bulk soil solution, chemical degradation, photochemical transformations and biodegradation of soil-sorbed compounds are also possible. Furthermore, soil-sorbed compounds may also be lost from areas of soil by wind and water erosion, uptake by invertebrates or direct ingestion of soil by grazing livestock.

Persistence of organic chemicals in soils

The persistence and relative significance of each loss pathway in a soil depends on the physicochemical properties of the organic chemical and environmental conditions. However, regardless of the main loss pathway, the overall loss of organic chemicals from soils is often biphasic. This means that a short period of rapid initial loss is most often followed by a longer period of slow dissipation. Loss processes in soils exhibit similar patterns. The primary rate-limiting factors governing this biphasic behavior are thought to be fundamental sorption or desorption mechanisms that control the distribution of a chemical between the solid and aqueous or gaseous phases of soil and hence, the supply of chemicals available to the various loss pathways. The availability of organic compounds in a soil will affect both the toxicological threat and the potential to remediate a contaminated site. Further details relating to specific technologies available for contaminated soils can be found in the chapter on the remediation of polluted soils.

Conventional POPs

The persistence and overall loss of organic chemicals from soils following their entry into a soil depend on the physicochemical properties of the soil and the chemicals themselves, environmental factors, including temperature and precipitation, and anthropogenic factors such as cultivation and drainage.

Typically, organic contaminants with short half-lives exhibit relatively rapid dissipation with most of the residues ultimately being removed from the soil. Such decay curves are typical of compounds that are volatile, water-soluble, or easily degraded (or a combination of these). Examples of this type of chemical are the lighter chlorinated benzenes and volatile aromatic hydrocarbons. At the other extreme, chemicals that are nonvolatile, relatively water-insoluble and recalcitrant, dissipate slowly. Such chemicals are strongly sorbed and persist in soils over often extremely long periods. Examples include PCDDs and PCDFs. Most organic compounds fall within these two extremes, with an initial period of rapid loss followed by a much slower dissipation of the chemical. This leads to a ‘persistent residual fraction’ of the compound forming in the soil, the size of which will be determined largely by the physicochemical properties of the compound and soil. As mentioned earlier, a loss profile of this shape is often referred to as ‘biphasic’ and is typical of organic chemicals with intermediate sorption potentials.

New POPs (PFAS)

The chemical structure of PFAS contains the C—F bond, the strongest covalent bond known in organic chemistry (Krafft and Riess, 2015a). The small size and unique properties of fluorine make the PFAS thermally and chemically stable. Therefore, per-fluorinated compounds, PFAS, which are fully fluorinated (e.g., PFOS, PFOA) are extremely persistent. On the other hand, the poly-fluorinated

compounds, which also contain hydrocarbon (C—H) or other non-fluorinated bonds (e.g., C—H or C—O) in the structure, are susceptible to biotic or abiotic breakdown and transformations (similar to conventional POPs). However, the breakdown of the polyfluorinated compounds often leads to a transformation into shorter chain perfluorinated compounds as terminal degradation products, and therefore, are considered as potential precursor compounds. The degradation process also produces transient degradation intermediates that may persist in the soil environment. Examples of such transformation include degradation of fluorotelomer alcohols (FTOH) or fluorotelomer sulfonamido alkylbetaines (FTABs—see Table 2b) into intermediate PFAS as well as terminal perfluoroalkyl carboxylates of various chain lengths (e.g., PFOA). Several studies on soils, sediments and sludges (reviewed by Butt et al., 2013) have shown that breakdown or transformation of 6:2 FTOH and 8:2 FTOH through biotic and abiotic pathways could result in the formation of terminal perfluoroalkyl substances such as PFOA, PFHxA (typical yields of 1–2%) and several (5–12) intermediate metabolites (both transient and terminal) some of which may be covalently bound to biological macromolecules and thus become non-extractable (making up to 10–35% of the original mass). In terms of the processes of ‘ageing’ and formation of bound residues, discussed below, there may be commonality between these intermediate products and the conventional POPs. However, more work is needed to understand this phenomenon fully for PFAS.

The ‘ageing’ of POPs within soils

As discussed above and can be seen in Fig. 1, when an organic compound enters a soil, it undergoes a number of loss processes including leaching, volatilization, chemical and biological degradation, and plant uptake. In addition, sorption- and transport-related processes mean that an increasing component of the soil-borne compound becomes ‘irreversibly bound’ within the soil matrix, i.e., it cannot be extracted by conventional analytical techniques. The result is that, as the residence times of organic compounds increase within a soil, they become increasingly difficult to extract. This phenomenon has become known as the ‘ageing effect.’ There is also an assumption that there is an analogous decline in a compound’s bioavailability as soil residence time increases and extraction efficiency declines, but this may be erroneous (see later).

Fig. 1 demonstrates that this process of declining extractability of soil-bound compounds is, in part, due to the formation of irreversibly bound residues within the soil matrix. The development of irreversibly bound residues is a major contributing factor in the formation of persistent residual fractions of organic chemicals in soils. Addressing this issue is complicated by the fact that experimental data are highly dependent on the extraction method used and knowledge that the efficiency of extraction of an organic chemical from a soil will decline with increasing contact time between a chemical and the soil, i.e., the ‘ageing effect.’

Several groups of mechanisms are thought to be involved in the ‘ageing effect’ of organic chemicals in general, including POPs. The processes involved can be split into transport-related and sorption-related mechanisms. Transport-related mechanisms are those where sorption or desorption of organic chemicals is controlled by water flow in soils, especially in structured soils, where, for example, a network of cracks may develop due to swelling and shrinkage of soils, leading to mobile and immobile domains for chemical transport.

Sorption-related mechanisms

Sorption mechanisms of conventional POPs can be divided into two distinct classes: diffusive mass transfer and chemisorption. Diffusive mass transfer mechanisms include film diffusion, pore diffusion, retarded intraparticle diffusion, and intraorganic matter diffusion. Film diffusion relates to the resistance encountered by an organic chemical when crossing a thin film of water surrounding soil solids or aggregates, or when moving from the solid phase into the bulk solution or gaseous phases. Depending

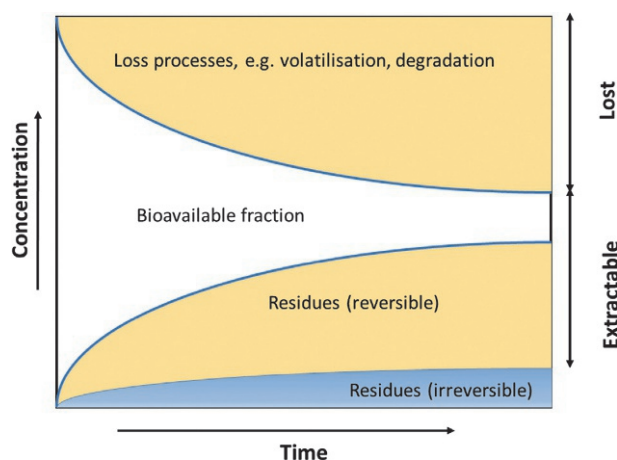


Fig. 1 Dynamics of organic chemical persistence, bioavailability and loss in soils. After Johnson D (2005) Pollutants/persistent organic (POPs). In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*, pp. 264–271. Elsevier.

on the nature of sorbents and sorbate (within soil, i.e., organic matter, inorganic minerals), the sorption process may involve various stages: (i) film diffusion across the liquid film surrounding the external surfaces of soil particles, (ii) intraparticle diffusion into the liquid in the pores and along the pore walls and (iii) slower phase of sorption.

Chemisorption occurs when a functional group of an organic chemical interacts with reactive moieties on soil solids through the formation of strong chemical bonds, principally covalent bonds. Such interactions give rise to the formation of a fraction that is irreversibly sorbed, despite exhaustive extraction procedures, and has become known as 'bound residues' in the area of pesticide science. However, pesticides are often ionic, in contrast to nonionic compounds such as nonionic POPs, which are less likely to form chemical bonds with soil solids. Thus, chemisorption is unlikely to be a major rate-limiting factor governing the persistence and loss of hydrophobic organic chemicals in soils. Where a contaminant tends to ionize or contains polar functional groups (such as in the case of PFAS), then chemisorption may be important. In addition, biologically mediated degradation processes may alter the structure of nonionic organic chemicals so that they are more susceptible to chemisorption. In biologically-mediated degradation, the irreversible chemisorption of metabolites can also occur, and it may compete with biodegradation. It is important to understand fully the formation of the persistent residual fraction in soil, as it may determine whether the compounds remain chemically and, more importantly, biologically unavailable.

Sorption retarded intraparticle diffusion

Of the mass transfer mechanisms involved in the formation of persistent residual fractions it is generally agreed that sorption retarded intraparticle diffusion (SRPD) and sorption retarded intraorganic matter diffusion (SROMD) are the two primary rate-limiting factors. The SRPD has been defined as 'aqueous-phase diffusion of a solute within pores of microporous particles (e.g., sand grains) mediated by retardation resulting from instantaneous sorption to pore walls.' However, SRPD is often used as an umbrella term for several similar processes that occur by diffusion of organic compounds through pores. The above definition of SRPD refers to the molecular diffusion of chemicals in pore water that is retarded, chromatographic-like, by local sorption on pore walls. This process is more marked in larger mesopores (diameter $>2\ \mu\text{m}$), and the process may or may not be aided by sorption to and within organic matter lining these pores (see later for description of SROMD). As the diameter of a pore decreases, its tortuosity tends to increase, further hindering diffusion processes.

Fig. 2 illustrates graphically the various stages of diffusion that occur in exceptionally narrow pores (a few nanometers in diameter) traversed by a model PAH, benzo(a)pyrene. Compounds tend to enter these nanopores as they diffuse along gradually

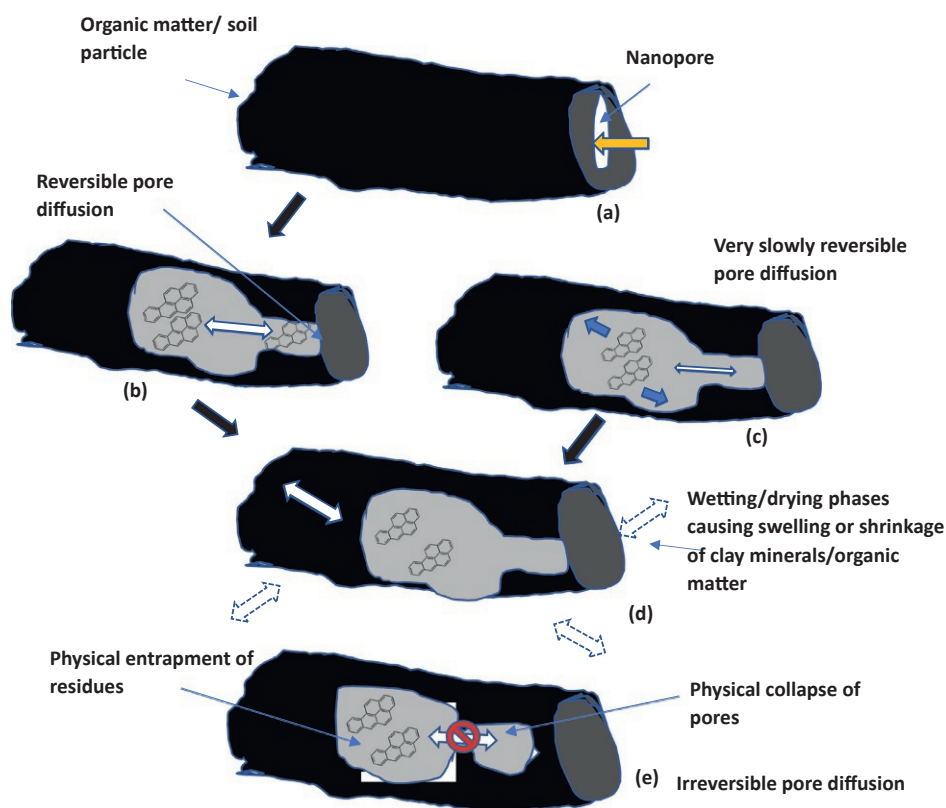


Fig. 2 A conceptual diagram showing the mass-transfer limited processes leading to sorption-retarded pore diffusion. After Johnson D (2005) Pollutants/persistent organic (POPs). In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*, pp. 264–271. Elsevier.

narrowing mesopores ($>2\ \mu\text{m}$) followed by micropores ($<2\ \mu\text{m}$) and eventually end up in nanopores whose diameter may be only a few nanometers in diameter (Fig. 2a). In extremely narrow pores, diffusion is often exceptionally slow due to the high viscosity of pore water. Polar minerals have one or more layers of water strongly sorbed on their surfaces. This ordered layering of molecules increases the viscosity of the water and greatly slows molecular diffusion within nanopores (Fig. 2b). The water contained in nanopores may be 'ice-like' and thereby greatly restrict solute diffusion. The tortuosity of pores becomes an increasingly important factor in the retardation of pore diffusion. In extremely narrow and tortuous pores, steric hindrance may further retard desorption (Fig. 2c). When this occurs, sorption is only very slowly reversible, contributing to the persistent residual fractions discussed earlier. Natural wetting-drying phases of a field soil will lead to expansion and contraction of clay mineral lattices (Fig. 2d), which can induce physical collapse of pores and physical entrapment of solute molecules within 'blind alleys' (Fig. 2e).

Fig. 2 therefore represents a series of steps that progressively restrict solute diffusion and retard desorption leading to persistent residual fractions of organic compounds in soils. These compounds, although chemically unaltered, are usually unavailable to traditional chemical extraction techniques and the assumption is that they are also not bioavailable. However, it is important to remember that, although such compounds are often referred to as irreversibly bound, this may not always be the case. Physical destruction of the soil aggregates they are associated with may release the compounds or at least increase their biological and chemical availability. Acidification dissolves the inorganic oxide cements that hold soil aggregates together and hence may similarly increase rates of pore diffusion and sorption.

Intraorganic matter diffusion

Intraorganic matter diffusion (IOMD) is the other important rate-limiting process involved in the desorption of persistent organic compounds from soils. Soil organic matter can exist as surface coatings, lining pores, or as discrete particles. Organic matter diffusion may be the primary rate-limiting process in soils with large organic matter contents but, since slow sorption also occurs in soils with little to no organic matter, it is clearly not the only rate-limiting sorption process. The IOMD is considered to be diffusion solely within the matrix of organic matter (OM), which is thought to resemble a polyelectrolyte polymer mesh, with retardation resulting from reversible sorption on the internal surfaces of organic matter. Organic matter polymers are said to have glassy (condensed, rigid) or rubbery (expanded, flexible) structures (Fig. 3). Diffusion in polymers occurs either by a place change mechanism, in which movement is accomplished by cooperative interchange of position in polymer segments and the penetrating molecule, or by a defect mechanism where the penetrant may jump between lattice defects, voids, and pores. The OM polymers consist of condensed, rigid 'glassy' phases through which diffusion is slow (Fig. 3b) or more expanded and flexible 'rubbery' structures (Xing and Pignatello, 1997) through which diffusion is faster (Fig. 3a).

In addition, some organic compounds may undergo 'Langmuir-like' adsorption (Fig. 3c) in internal voids in the OM matrix, which has been likened to a hole-filling mechanism (Xing and Pignatello, 1997). These slow diffusion processes lead to the formation of physically protected compounds which, as in the case of SRPD, are often mistakenly labelled as irreversibly bound residues. The OM diffusion and binding processes are in fact transitory since environmental variables may swell or soften

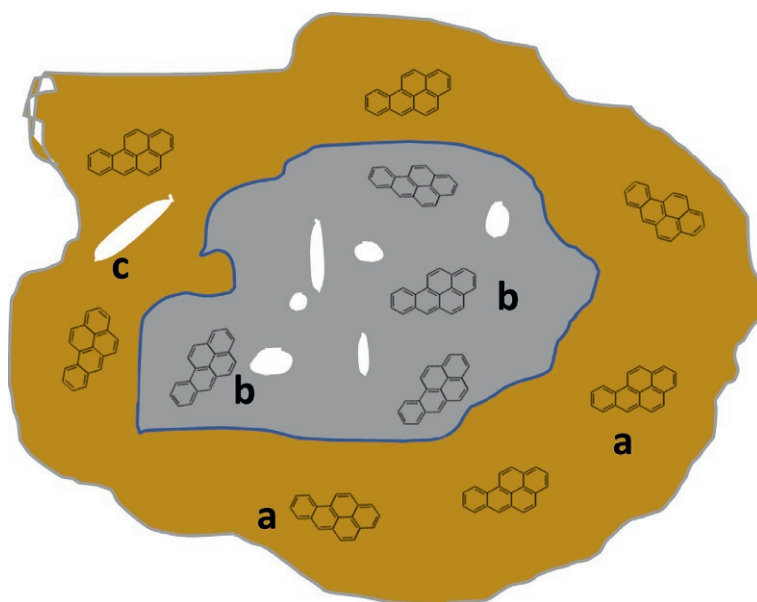


Fig. 3 A conceptual diagram describing the processes involved in sorption of organic chemicals in soil organic matter consisting of sorption on a rubbery phase (a), on a glassy phase (b), and through hole filling (c) mechanisms. Not to scale; Conceptualization based on Xing B and Pignatello JJ (1997) Dual-mode sorption of low-polarity compounds in glassy poly(vinyl chloride) and soil organic matter. *Environmental Science & Technology* **31**: 792–799.; Johnson D (2005) Pollutants/persistent organic (POPs). In: Hillel D (ed.), *Encyclopedia of Soils in the Environment*, pp. 264–271. Elsevier.

the OM matrix, converting it from a glassy to a more rubbery state and accelerating diffusion processes. Increasing temperatures, change in pH, and the presence of co-solvents can all act to increase the diffusivity of tightly bound compounds.

Recent work on new POPs (PFAS) has also shown that diffusion-related processes can make a significant contribution to their sorption on soil organic and mineral matter. For example, Xiang et al. (2018) in their studies on sorption of PFOA on various size fractions of soil particles observed that the pore sizes in their soils ranged from 8 nm to 16.9 nm in size (i.e., in the mesopore size range), thus allowing intraparticle diffusion of the linear structure of PFOA. They found that both the film and the intraparticle diffusion processes made a significant contribution to PFOA sorption on various size fractions of the soil. The role of intraparticle diffusion diminished as pore sizes became smaller. Intraparticle diffusion in the inner matrix of organic and mineral matter is expected to slow down their release and thus can potentially sequester the residues of organic compounds in the soil matrix.

In summary, it is probable that several mass transfer mechanisms are responsible for retarded diffusion leading to the formation of persistent residual fractions within most soils. Of these, the two most important are usually SROMD and SRPD. Although much work is needed to determine the exact contribution of each mechanism, it is likely that in most soils a synergy of both SROMD and SRPD is primarily responsible for the formation of persistent residual fractions. Residual fractions are often considered both biologically and chemically unavailable. However, as the preceding passages have shown, the mechanisms that lead to their formation are often transitory. This is particularly important when considering biological availability since the conditions that are known to enhance rates of diffusion often occur in the intestines of biological organisms (low pH, mechanical manipulation, and elevated temperatures). Hence, soil ingestion by an organism may have an impact upon a compound's availability and traditional extraction techniques used to determine the risk presented by a persistent residual fraction may be inaccurate.

Additional mechanisms involved in sorption of new POPs (PFAS)

The mechanisms elucidated in Figs. 2 and 3 above, are equally applicable to the sorption of several new POPs (pesticides and PBDEs flame retardants). However, for PFAS in soils several additional mechanisms are involved in their sorption and retention in soils, as discussed below.

The mechanisms of sorption or retention of PFAS in soils are more complex (Li et al., 2018) than conventional POPs (which are non-ionic and highly hydrophobic) because of their complex chemistry (e.g., amphiphilic and surfactant properties as well as surface activity) as shown in Fig. 4. The PFAS are composed of a hydrophobic C—F chain (tail) and a polar functional group (head). Therefore, their hydrophobicity increases with the length of the tail and consequently their sorption affinity to soil organic carbon (represented by the K_{oc} parameter,) also increases with increasing length of the C—F chain (Table 2b). While sorption or partitioning of some hydrophobic PFAS (long-chain compounds with >C8) on soil organic carbon may be similar to conventional POPs and may be the key controlling factor of their sorption behavior in soils. The sorption of the short-chain compounds is primarily driven by electrostatic interaction with the soil solid phase (Nguyen et al., 2020). For example, in the case of carboxylates or sulphonates the acidic head-group makes them anionic, which does not favor their sorption in net negatively charged soils. Only certain soils, such as those rich in Fe and Al oxides, could carry a net positive charge at an environmentally relevant pH range (4–9) and have the ability to sorb anionic POPs. It is noteworthy that several PFAS may exist as a cation or a zwitterion depending on the soil pH. While those containing amide and sulfonamido headgroups are cationic in nature, fluorotelomer sulfonamido alkylbetaines (FTABs) can form zwitterions in soils. In contrast to conventional POPs, the electrostatic interactions with the soil solid phase (both organic as well as mineral matter) play a major role in the sorption of PFAS in soils (Du et al., 2014; Li et al., 2018; Mejia-Avendaño et al., 2020).

Many PFAS, being surfactants, are surface-active compounds and tend to accumulate at the fluid-fluid interface (Fig. 4b). This has implications for the adsorption of PFAS in unsaturated soils with significant air:water interfacial area or those containing immiscible organic liquids i.e., non-aqueous phase liquid (NAPL) as co-contaminants (e.g., trichloroethylene). In such soils their adsorption at the fluid (air, NAPL)-water interface can be enhanced several fold compared to that in saturated soils (commonly used in the measurement of sorption coefficients on the soil solid phase). It has been noted that where sorption by the soil solid phase is small (e.g., sandy soils with low organic carbon) the contribution of interfacial adsorption towards PFAS retention in soils is higher. Brusseau (2018) observed that in unsaturated soils, air:water interfacial adsorption could contribute up to >80% of total retention in sandy soils (with low solid phase sorption) containing significant air-water interfacial areas, depending on the degree of saturation of soil pores with water. Several factors including solid phase sorption, soil texture, moisture content, PFAS concentration, co-contaminants and soil salinity influence the extent of interfacial adsorption of PFAS in soils.

Currently, the sorption behavior in soils of polyfluorinated PFAS with complex chemistry (e.g., FTAB zwitterions, fluorotelomer amines such as 6:2 FtSaAm, and other polyfluorinated precursor) is poorly understood, is more complex than expected and is difficult to predict by bulk soil properties (Barzen-Hanson et al., 2017).

The (Bio)availability and bioaccumulation of POPs in soils

The term 'bioavailability' is often discussed in the literature but rarely defined. In relation to the focus of this chapter, bioavailability refers to the extent to which humans and ecological receptors are exposed to POPs in soil. This is a broad definition and encompasses bioavailability to plants and microorganisms by a variety of potential routes. The amount of compound present in an organism will be a function of the rate of absorption versus the rate of excretion.

The POPs in soils are potentially bioavailable to many different organisms. In this chapter, a brief overview is provided of the bioavailability of POPs to those organisms that are thought to be most important. These are plants, bacteria and fungi, and grazing

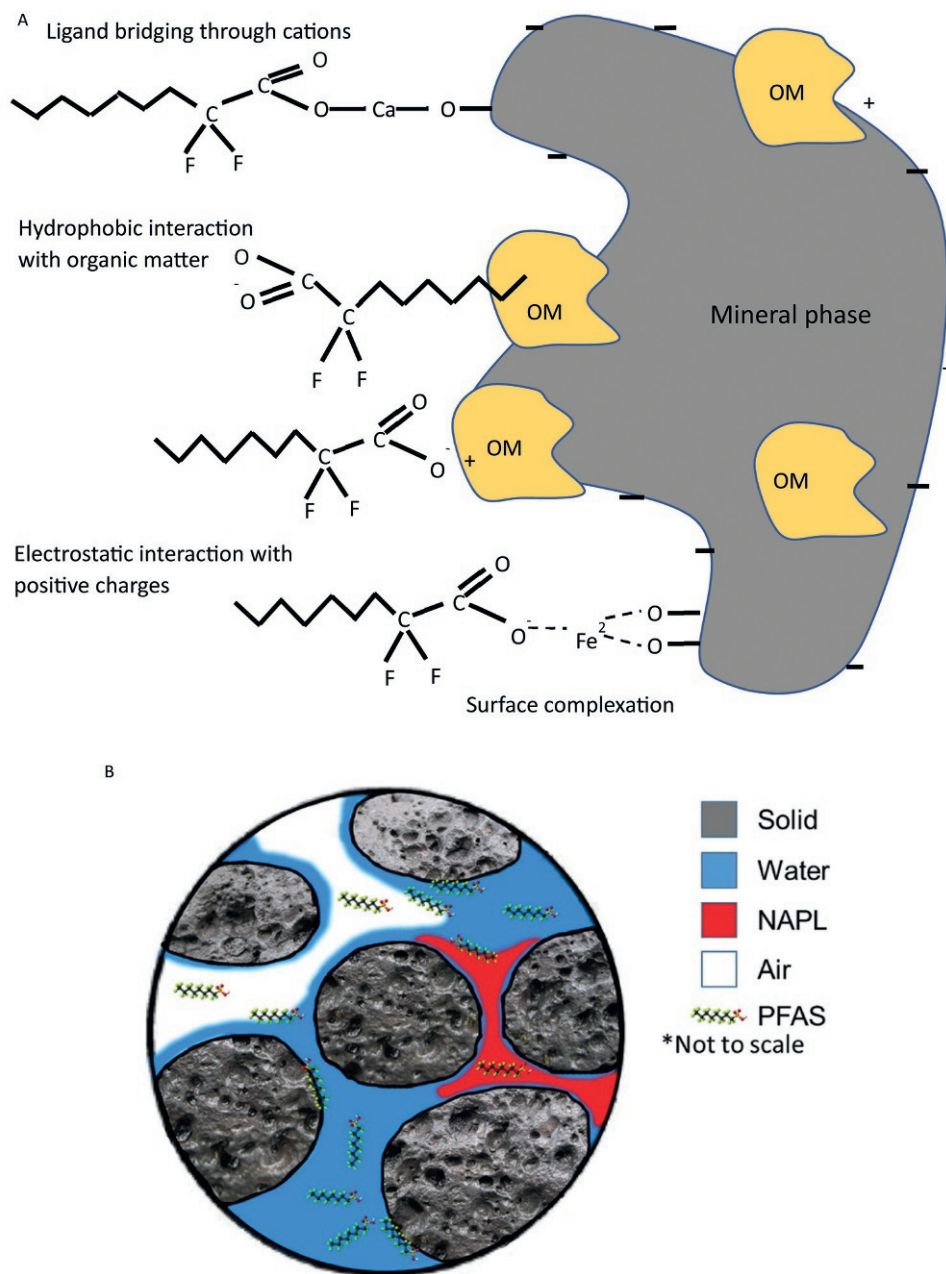


Fig. 4 Conceptual diagrams (a and b) describing a range of sorption processes involved in sorption of PFAS in soils. (a) Electrostatic and hydrophobic interactions with soil organic and mineral matter.; (b) Fluid-fluid interfacial adsorption. (a) Reproduced with permission from Li Y, Oliver D and Kookana RS (2018) A critical analysis of published data to discern the role of soil and sediment properties in determining sorption of per and polyfluoroalkyl substances (PFAS). *Science of the Total Environment* **628–629**: 110–120.; (b) Reproduced with permission from Brusseau ML (2018) Assessing the potential contributions of additional retention processes to PFAS retardation in the subsurface. *Science of the Total Environment* **613–614**: 176–185.

animals. Macroinvertebrates, in particular earthworms, are dealt with in more depth as they may represent important routes for POPs to enter the food chain. Organisms such as earthworms that are in constant physical contact with the soil may accumulate considerably greater burdens of POPs in their bodies than the surrounding soil. These concentrations may not be great enough to cause either acute or chronic toxicity in the earthworms. However, many species feed on earthworms and will concentrate the lipophilic compounds in their body fat. The result is an increase in POP body burdens at a higher trophic level of the food chain. This phenomenon has become known as biomagnification and may lead to symptoms of toxicity in organisms that have little contact with soil POPs.

Despite being very different from hydrophobic POPs in terms of chemical properties, the bioaccumulation and biomagnification (through dietary uptake) of PFAS has been demonstrated in a wide range of organisms such as plants, invertebrates, fish and

humans (Houde et al., 2011; ITRC, 2020). While hydrophobic POPs have the tendency to migrate from water to the lipids of organisms, PFAS are proteinophilic (e.g., PFOS accumulate in serum albumin protein) and thus are somewhat unique among POPs. The PFAS (especially carboxylates and sulphonates) are predominantly found in protein-rich compartments (such as blood and liver) rather than in the fatty tissues. The conventional models, based on equilibrium octanol-water partition coefficients, cannot be applied to PFAS. In fact, the surface activity of ionic PFAS surfactants (being both hydrophobic and hydrophilic) prevents the empirical determination of K_{ow} . Nevertheless, the hydrophobic part of PFAS interacts with lipids and therefore the bioaccumulation of such compounds has been found to be related to their C—F chain length. The chemistry of PFAS head groups also affect their bioavailability and generally the PFAS with sulfonic acid headgroup are more bioaccumulative than carboxylic acid (Conder et al., 2008). The short chain PFAS (<7C-F chain) have low bioaccumulation and biomagnification potential.

Availability to plants

Uptake of POPs and other organic chemicals can occur through a number of pathways, such as passive and active uptake of the chemical by plant roots from soil solution, transport from the root to the shoot, transfer of the chemical from soil to soil air (depending on its volatility) and in turn uptake by plants through soil-air-plant pathway and particulate deposition to plant surfaces (Collins et al., 2006). In general, translocation of nonionic organic compounds (conventional POPs) from the roots to shoots is negligible because of low water solubility, high K_{ow} and strong associations of many compounds to soil particles and organic matter. Plant uptake of nonionic organic chemicals from soils is usually dominated by vegetative uptake of contaminated vapor from the surrounding air (McLachlan, 1996). Furthermore, many POPs are volatile in nature and may transfer into the gaseous phase from soil depending on the ambient conditions (temperature, humidity, etc.). Although large concentrations of soil-borne nonionic contaminants are rarely available for translocation by plant roots (due to their strong binding to soil particles), heavily contaminated soils can influence the concentrations of organic compounds in above-ground vegetation by the soil-air-plant route.

Perfluoroalkyl acids (PFCAs) being hydrophilic in nature are taken up by plants passively through water and therefore, their accumulation in plants is mainly controlled by rates of plant transpiration (influenced by plant species and environmental conditions). The bioavailability of PFAS in soil porewater will be regulated by their sorption affinity to soils, depending on PFAS chemistry (e.g., chain length, charge characteristics and hydrophobicity) as well as soil properties (organic carbon, clay content, pH and salinity). However, in a study on PFOS and PFOA uptake by seven different plant species it was observed that the PFAS concentrations in plant biomass were positively correlated with root protein content and negatively correlated with the root lipid content (Wen et al., 2016). This shows that because PFAS are proteinophilic, their uptake and translocation in plants could also be associated with protein regulation.

Availability to soil microflora

The 'availability' of organic compounds in soils to microorganisms is critical for their degradation and removal from the environment by either natural attenuation or bioremediation. Microbial biodegradation is a major process affecting the persistence of POPs in soils. The extent of POP bioavailability to bacteria in different soils may differ markedly, even under the same optimum growth conditions. Sorption has an important influence on the bioavailability of hydrophobic organic compounds in soils. A reduction in the rates of degradation with increasing soil sorption capacity has been noted for POPs due to their reduced bioavailability. Furthermore, a persistent residual fraction of POPs may remain unavailable to soil bacteria after remediation. This can be attributed to the slow migration of an increasingly large fraction of the POPs into the soil organic matter and mineral matrix with time.

As discussed earlier, POP chemical availability decreases with increasing residence time in soil (Alexander, 2000). This reduced chemical availability is often mirrored by an analogous decline in bioavailability. This 'ageing effect' may reduce the efficiency of bioremediation of a contaminated soil as a function of time. However, the reduced bioavailability may also reduce the potential toxicological threat of organic compounds in a soil. While a great deal of research on the bioavailability of POPs has been carried out for developing an accurate assessment of their bioavailability in soils, little is currently known about the bioavailability of PFAS. However, given PFOS and PFOA are anionic in nature, these are expected to be readily bioavailable to soil microflora. Nevertheless, these compounds and many other PFAS are extremely resistant to microbial degradation.

Availability to grazing animals

Soil ingestion by grazing farm animals can form a significant fraction of the total dry-matter intake, depending on the grazing habit of animal and forage conditions (e.g., plant species, sward height, etc.). While in some circumstances, soil formed up to 26% of the total diet ingested by sheep, generally the ingested soil forms <5% of the total diet (Duarte-Davidson and Jones, 1996; Weber et al., 2018; Collas et al., 2020). Free range chickens are more prone to POP contamination, as they take up more soil (per body weight) than other farm animals (Weber et al., 2018). For hens ingesting approximately 30 g soil per day, soils levels of 2–4 ng PCB-TEQ kg⁻¹ can exceed European Union standards (Weber et al., 2018). Thus, ingested soil may represent a potentially major route of uptake for soil-borne organic contaminants by farm animals. It follows that consumption of meat, eggs and or dairy products may be an important exposure pathway to humans (Weber et al., 2018). In the case of PFOS and PFOA, these compounds are more readily taken up by plants from soil solution and therefore the contamination of forage is likely to form a significant pathway in addition to soil ingestion.

Availability to earthworms

Earthworms can be extremely numerous, especially in pastureland where their biomass can be up to 272 g m⁻². Add to that the fact that an individual earthworm can consume up to 400 times its own weight in soil over a 12-month period, their importance in the dynamics and persistence of organic compounds in soils becomes evident. This high density and activity mean that up to 25% of the surface layers of soil may be turned over annually. A relatively large body lipid content means that earthworms accumulate POPs. Earthworms may facilitate faster release of certain sorbed residues, since a large quantity of soil passes through their guts and is consequently exposed to digestive enzymes that act on soil organic matter. Hence, the assumption that there is an analogous decline in chemical and biological availability of organic compounds may break down when an active (in this case gastrointestinal) uptake mechanism is involved as opposed to a more passive mechanism, such as contact between skin and soil.

Bioaccumulation of a wide range of PFAS (mostly PFCAs) in earthworms has been reported in the literature (e.g., Rich et al., 2015), generally increasing with the increasing chain length of PFCAs (7 to 12C—F). For equivalent chain length the bioaccumulation factors were found to be typically higher for sulphonic acids than for carboxylic acid analogues. Bioaccumulation of polyfluoroalkyl substances (e.g., 6:2 FTAB present in AFFFs), that upon transformation can form PFOA and PFOS, has also been found to related to their chain lengths (Munoz et al., 2020).

Concluding remarks

The family of POPs (under the Stockholm Convention) is growing. In addition to the original group of 12 chemicals, 16 new chemicals were added to the POP family in 2009. These included some with chemical properties similar to those of old POPs (e.g., flame retardants), but others such as PFAS had a unique and very different chemistry from that of the older POPs. Owing to their unique chemistry some of their environmental fate and transport processes and mechanisms resulting in their persistence, sorption, bioavailability and transport are very different from those of conventional POPs. Therefore, the conventional models, such as those based on their hydrophobicity and equilibrium octanol-water partition coefficients (K_{ow}), cannot be applied to PFAS. While new models and approaches are needed for PFAS, currently their fate and behavior in the soil environment remain poorly understood. A massive research effort is currently being directed internationally on developing better understanding of PFAS in the environment.

References

- Alexander M (2000) Aging, bioavailability, and overestimation of risk from environmental pollutants. *Environmental Science & Technology* 34: 4259–4265.
- Barzen-Hanson KA, Davis SE, Kleber M, and Field JA (2017) Sorption of fluorotelomer sulfonates, fluorotelomer sulfonamido betaines, and a fluorotelomer sulfonamido amine in national foam aqueous film-forming foam to soil. *Environmental Science & Technology* 51: 12394–12404.
- Brusseau ML (2018) Assessing the potential contributions of additional retention processes to PFAS retardation in the subsurface. *Science of the Total Environment* 613–614: 176–185.
- Brusseau ML (2019a) Estimating the relative magnitudes of adsorption to solid-water and air/oil-water interfaces for per- and poly-fluoroalkyl substances. *Environmental Pollution* 254B: 113102.
- Brusseau ML (2019b) The influence of molecular structure on the adsorption of PFAS to fluid-fluid interfaces: Using QSPR to predict interfacial adsorption coefficients. *Water Research* 152: 148–158.
- Brusseau ML, Anderson RH, and Guo B (2020) PFAS concentrations in soils: Background levels versus contaminated sites. *Science of the Total Environment* 740: 140017.
- Buck RC, Korzeniowski SH, Leganis E, and Adamsky F (2021) Identification and classification of commercially relevant per- and poly-fluoroalkyl substances (PFAS). *Integrated Environmental Assessment and Management* 17: 1045–1055.
- Burns DC, Ellis DA, Li H, McMurdo CJ, and Webster E (2008) Experimental pKa determination for perfluorooctanoic acid (PFOA) and the potential impact of pKa concentration dependence on laboratory-measured partitioning phenomena and environmental modeling. *Environmental Science & Technology* 42: 9283–9288.
- Butt CM, Muir DCG, and Mabury SA (2013) Biotransformation pathways of fluorotelomer-based polyfluoroalkyl substances: A review. *Environmental Toxicology and Chemistry* 33: 243–267.
- Collas C, Mahieu M, Badot PM, et al. (2020) Dynamics of soil ingestion by growing bulls during grazing on a high sward height in the French West Indies. *Scientific Reports* 10: 17231.
- Collins C, Fryer M, and Grosso A (2006) Plant uptake of non-ionic organic chemicals. *Environmental Science & Technology* 40: 45–52.
- Conder JM, Hoke RA, de Wolf W, Russell MH, and Buck RC (2008) Are PFCAs bioaccumulative? A critical review and comparison with regulatory criteria and persistent lipophilic compounds. *Environmental Science & Technology* 42: 995–1003.
- Du Z, Deng S, Bei Y, et al. (2014) Adsorption behaviour and mechanism of perfluorinated compounds on various adsorbents—A review. *Journal of Hazardous Materials* 274: 443–454.
- Duarte-Davidson R and Jones KC (1996) Screening the environmental fate of organic contaminant in sewage sludge applied to agricultural soils: II. The potential for transfer to plants and grazing animals. *Science of the Total Environment* 185: 59–70.
- Giesy JP and Kannan K (2001) Global distribution of perfluorooctane sulfonate in wildlife. *Environmental Science & Technology* 35: 1339–1342.
- Houde M, De Silva AO, Muir DCG, and Letcher RJ (2011) Monitoring of perfluorinated compounds in aquatic biota: An updated review. *Environmental Science & Technology* 45: 7962–7973.
- Interstate Technology and Regulatory Council (ITRC) (2020) *PFAS Technical and Regulatory Guidance Documents and Fact Sheet PFAS-1*. Washington, DC.
- Johnson D (2005) Pollutants/persistent organic (POPs). In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*, pp. 264–271. Elsevier.
- Jones KC (2021) Persistent organic pollutants (POPs) and related chemicals in the global environment: Some personal reflections. *Environmental Science & Technology* 55: 9400–9412.
- Krafft MP and Riess JG (2015a) Selected physicochemical aspects of poly- and perfluoroalkylated substances relevant to performance, environment and sustainability-part one. *Chemosphere* 129: 4–19.
- Krafft MP and Riess JG (2015b) Per- and polyfluorinated substances (PFASs): Environmental challenges. *Current Opinion Colloid Interface* 20: 192–212.
- Lancet (2018) The lancet commission on pollution and health. *Lancet* 391: 462–512.

- Li Y, Oliver D, and Kookana RS (2018) A critical analysis of published data to discern the role of soil and sediment properties in determining sorption of per and polyfluoroalkyl substances (PFAS). *Science of the Total Environment* 628–629: 110–120.
- McLachlan MS (1996) Bioaccumulation of hydrophobic chemicals in agricultural food chains. *Environmental Science & Technology* 30: 252–259.
- Mejia-Avendaño S, Zhi Y, Yan B, and Liu J (2020) Sorption of polyfluoroalkyl surfactants on surface soils: Effect of molecular structures, soil properties, and solution chemistry. *Environmental Science & Technology* 54: 1513–1521.
- Munoz G, Desrosiers M, Vetter L, et al. (2020) Bioaccumulation of zwitterionic polyfluoroalkyl substances in earthworms exposed to aqueous film-forming foam impacted soils. *Environmental Science & Technology* 54: 1687–1697.
- Nguyen TMH, Bräunig J, Thompson K, et al. (2020) Influences of chemical properties, soil properties, and solution pH on soil–water partitioning coefficients of per- and polyfluoroalkyl substances (PFASs). *Environmental Science & Technology* 54: 15883–15892.
- Rich CD, Blaine AC, Hundal L, and Higgins CP (2015) Bioaccumulation of perfluoroalkyl acids by earthworms (*Eisenia fetida*) exposed to contaminated soils. *Environmental Science & Technology* 49: 881–888.
- Wang Z, DeWitt JC, Higgins CP, and Cousins IT (2017) A never-ending story of per- and polyfluoroalkyl substances (PFASs)? *Environmental Science & Technology* 51: 2508–2518.
- Weber R, Herold C, Hollert H, et al. (2018) Reviewing the relevance of dioxin and PCB sources for food from animal origin and the need for their inventory, control and management. *Environmental Sciences Europe* 30: 42.
- Wen B, Wu Y, Zhang H, et al. (2016) The roles of protein and lipid in the accumulation and distribution of perfluorooctane sulfonate (PFOS) and perfluorooctanoate (PFOA) in plants grown in biosolids-amended soils. *Environmental Pollution* 216: 682–688.
- Xiang L, Xiao T, Yu PF, et al. (2018) Mechanism and implication of the sorption of perfluorooctanoic acid by varying soil size fractions. *Journal of Agricultural and Food Chemistry* 66: 11569–11579.
- Xing B and Pignatello JJ (1997) Dual-mode sorption of low-polarity compounds in glassy poly(vinyl chloride) and soil organic matter. *Environmental Science & Technology* 31: 792–799.

Relevant websites

- <http://chm.pops.int/Home/tabid/2121/Default.aspx>—The Stockholm Convention on POPs.
- <http://chm.pops.int/TheConvention/ThePOPs/tabid/673/Default.aspx>—List of POPs at the Stockholm Convention.
- <https://www.oecd.org/chemicalsafety/portal-perfluorinated-chemicals/>—OECD portal on PFAS.
- <https://pfas-1.itrcweb.org/fact-sheets/>—Interstate Technology Regulatory Council, USA. PFAS Fact sheets.

Pollution by microplastic in agricultural areas

Nicolas Beriot^a and Esperanza Huerta-Lwanga^{a,b}, ^aSoil Physics and Land Management Group, Wageningen University & Research, Wageningen, The Netherlands; ^bAgroecología, El Colegio de la Frontera Sur, Campeche, México

© 2023 Elsevier Ltd. All rights reserved.

Microplastics in agriculture	309
What do we call plastic?	309
Plastic use in agriculture	309
Plastic mulches in agriculture	310
Plastic, from ubiquitous use to ubiquitous contaminants in agricultural soil	311
Uncertainties shrouding plastic content assessment in soil	314
Effects of plastic residues on the environment	315
Effect on soil physicochemical properties	315
Effect on the soil microbiome	315
Effect on plants	315
Effect on soil invertebrates and vertebrates	316
Effect on soil ecosystem services	317
Transport of microplastics	317
Abiotic transport	317
Biotic transport	317
Microplastics and humans	318
Concluding comments	319
Acknowledgments	319
References	319
Further reading	320

Abstract

This chapter introduces the issue of plastic pollution in agricultural soil. Plastic pollution occurs when diverse plastic materials such as pipes, plastic breeding tunnels or plastic mulch are fragmented into macroplastics (>5 mm) and microplastics (<5 mm), or when they are applied inadvertently with amendments such as composts or sewage sludge. The impacts of plastic particles on soil physical properties, plants, invertebrates and vertebrates are also discussed. The biotic and abiotic transport of micro and nanoplastics is explained briefly, and also their transfer into the food chain where their potential effects on various body systems of animals and humans, for example, those on human blood, have been documented.

Microplastics in agriculture

In recent years we have become more aware of several factors that promote soil contamination; we know that soil is a natural entity or body and it requires centuries to millennia to be formed. Therefore, as a non-renewable resource its functioning (i.e., nutrient cycles, water retention, infiltration) is affected by various agricultural practices, and strong measures need to be taken if we wish to restore soils or promote their re-establishment. We know that soil can be further eroded, thus making it infertile. In this chapter we will discuss the contamination by microplastics at agricultural sites.

What do we call plastic?

A common definition of plastic is a material that is at least partly made of an organic polymer and can be molded into solid, non-soluble, objects. For example, polyethylene is composed of a chain of carbon atoms whereas polylactic acid, a polymer used in biodegradable plastics, is composed of a chain of lactic acid (Fig. 1).

Most conventional plastics are petroleum-based, meaning that they are made from fossil resources such as natural gas, oil or coal. In Europe plastic production accounts for 4–6% of all the oil and gas used, and agricultural plastics represent about 4% of the plastic production in Europe. Plastics can also be produced from crops in which case they are called bio-based. For example, sugar cane (*Saccharum officinarum* L.) can be processed to produce ethylene, which can then be used to manufacture polyethylene. Plastics can be made of a single polymer or a blend of several types, assembled in different ways. The two main polymers used in agriculture are polyethylene (PE) and polypropylene (PP). Additives are used to adjust the elasticity, color, mechanical strength and degradability of the plastic. The chemical composition and manufacturing are tailored to fit the plastic's intended function.

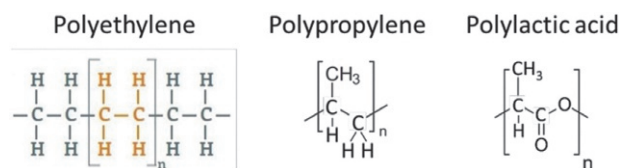


Fig. 1 Chemical structure of common polymers used in agriculture: polyethylene (PE), polypropylene (PP) and polylactic acid (PLA).

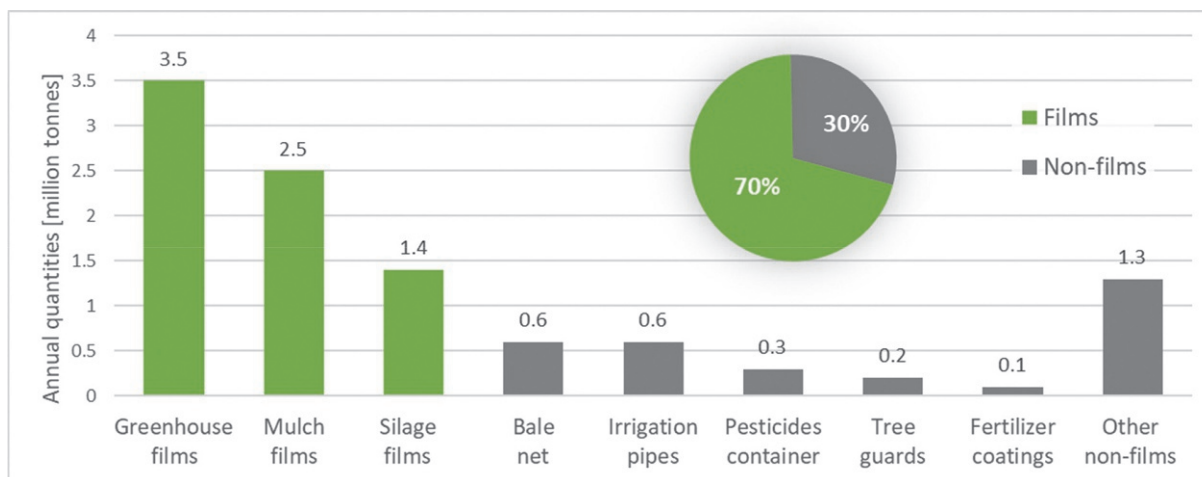


Fig. 2 Estimated global annual quantities of agricultural plastics for different uses add up to 10.5 million tonnes. Plastic food packaging is not included and represents an additional 37 million tonnes. Data from Moine 2018, APE-Europe 2021, FAO (2021) Assessment of Agricultural Plastics and Their Sustainability—A Call for Action. Rome, Italy.

Plastic use in agriculture

In agriculture, plastics are mostly used in the form of film as it constitutes a light, resistant, elastic, cheap and waterproof barrier (Fig. 2, FAO, 2021). Plastic mulch, greenhouse cover and bale wrapping (Fig. 3) are the three main agricultural practices that use plastic films. Plastics also provide other services: crop protective nets; irrigation pipes; ropes and twines; crates and containers for farming products; controlled-release coatings. Some of these plastics are placed directly in contact with the soil. This is the case for mulches, irrigation pipes and controlled-release coatings. We will focus on plastic mulches as they are an example of a wide application of plastic that is in direct contact with the soil.

Plastic mulches in agriculture

An estimated 2.5 million tonnes of plastic mulch are used annually worldwide, covering about 140,000 km² (twice the size of the Netherlands and Belgium together, FAO, 2021). Plastic mulch is used for one or all of the three following reasons:

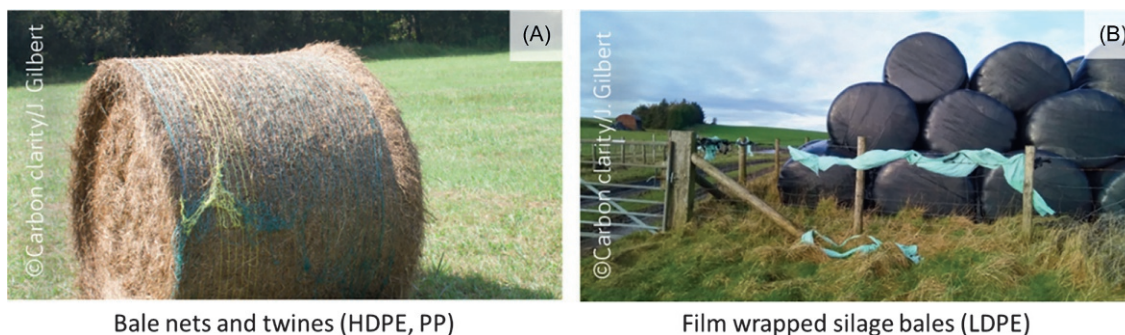


Fig. 3 Bale wrapping with nets and twines (A.), with silage film (B.) and the main polymers used, HDPE, PP and LDPE. Adapted from FAO (2021) Assessment of Agricultural Plastics and Their Sustainability—A Call for Action. Rome, Italy.

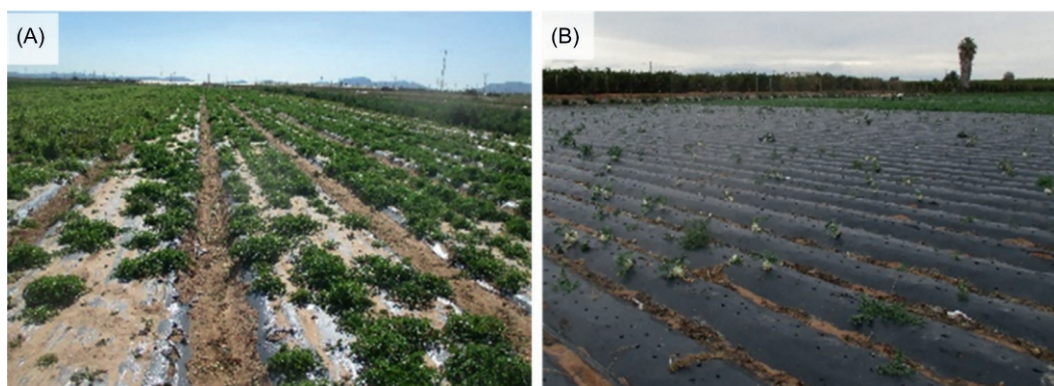


Fig. 4 Fields covered with low density polyethylene (LDPE) plastic mulch before the harvest of parsnip (Panel A) and after the harvest of Kohlrabi (*Brassica oleracea*, var. *gongylodes*, Panel B) in Southeast Spain.

- *Increasing soil temperature*: Plastic mulch was first noted for its ability to increase soil temperature in the 1950s. Higher soil temperatures increase nutrient availability, enhance nutrient uptake by roots, increase the number and activity of soil microorganisms, speed up plant germination and growth, and can help to control pathogens, e.g., *Tuta absoluta* in tomato (*Lycopersicon esculentum* Mill.) plants leading to higher and earlier yields. Therefore, plastic mulch may reduce the need for fertilization (for appropriate references for this section please see Further reading).
- *Increasing water use efficiency*: Water use efficiency is estimated by dividing the yield per hectare by the total amount of water applied. Plastic mulch is a barrier that prevents water evaporation from the soil and therefore increases water availability for plants. Plastic mulch can also increase rainwater harvesting when associated with ridge–furrow tillage; the ridges being mulched by plastic and plants growing in the furrows. An analysis of 266 studies in China, showed that plastic mulching significantly increased crop yield by 24% and water use efficiency by 28% on average (Gao et al., 2019).
- *Decreasing weed growth*: Opaque (often black) plastic mulch limits weed growth by preventing the light from reaching the soil. Plastic mulch can reduce weed emergence by 64–98% during the growing season, depending on the surface covered with plastic. In this sense plastic mulch can contribute to reducing the use of herbicides. Nevertheless, with a clear or transparent plastic mulch, the application of a pre-emergent herbicide is needed.

Because of these benefits, plastic mulch is used under various climatic conditions, in open fields and in greenhouses for the production of different crops, and under both organic and conventional farming practices. Plastic mulch is often partially buried in the soil to prevent it from being blown away by the wind (Fig. 4).

In summary, we showed that plastics are very diverse and provide many benefits to farmers. Intensive agriculture has great potential to feed the world population, however, this success is not without drawbacks.

Plastic, from ubiquitous use to ubiquitous contaminants in agricultural soil

The general public and the scientific community first became aware of the issue of plastic debris by observing their accumulation in the oceans after Captain Moore in 1997, since then, land has been identified as a major source and reservoir of plastic debris. Plastic debris is formed through the fragmentation of plastics in the environment. Plastic degradation relies on two main processes: weathering and biodegradation.

Weathering refers to abiotic reactions such as thermal degradation, photo-degradation, oxidation, hydrolysis and to mechanical degradation (e.g., wind or plowing). Weathering plays an important role in the degradation processes, as weathered plastic will undergo faster biodegradation. For example, photo-degradation can change the chemical structure of plastic polymers making them easier to degrade for microorganisms (Napper and Thompson, 2019).

Biodegradation of a polymer is a biological process leading to its complete or partial conversion to water, CO₂ (carbon, dioxide), CH₄ (methane) and new biomass by microorganisms (mostly bacteria and fungi). The biodegradation process can be divided in three different steps (Sander, 2019):

1. The organisms colonize the polymer and grow on its surface (Fig. 5).
2. The organisms degrade the polymer. They do this mostly by secreting enzymes (e.g., hydrolases) that can depolymerize the polymer. Depolymerization is the breaking of chemical bonds in the polymer that leads to smaller molecules. The main process of depolymerization is catalyzed hydrolysis with enzymes.
3. Finally, the hydrolysis products released from the polymer are used as an energy source or a carbon source for the microorganisms leading, for example, to the emission of CO₂ or the increase of biomass.

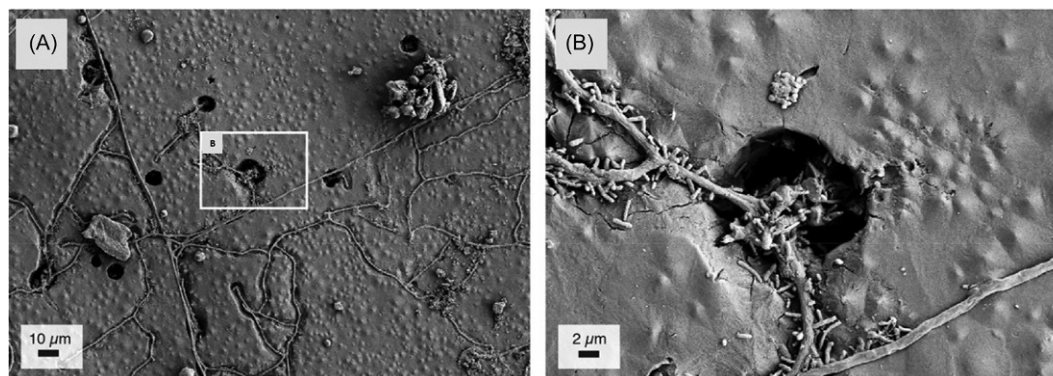


Fig. 5 Scanning electron microscopy images of the surfaces of polybutylene adipate-co-terephthalate (PBAT) films that were incubated with an agricultural soil in the laboratory (6 weeks; temperature of 25 °C). The images illustrate colonization of PBAT films by both fungi and unicellular organisms. Panel B shows a magnification of the area highlighted in panel A. Adapted from Sander M (2019) Biodegradation of polymeric mulch films in agricultural soils: Concepts, knowledge gaps, and future research directions, *Environmental Science and Technology*, 53: 2304–2315.

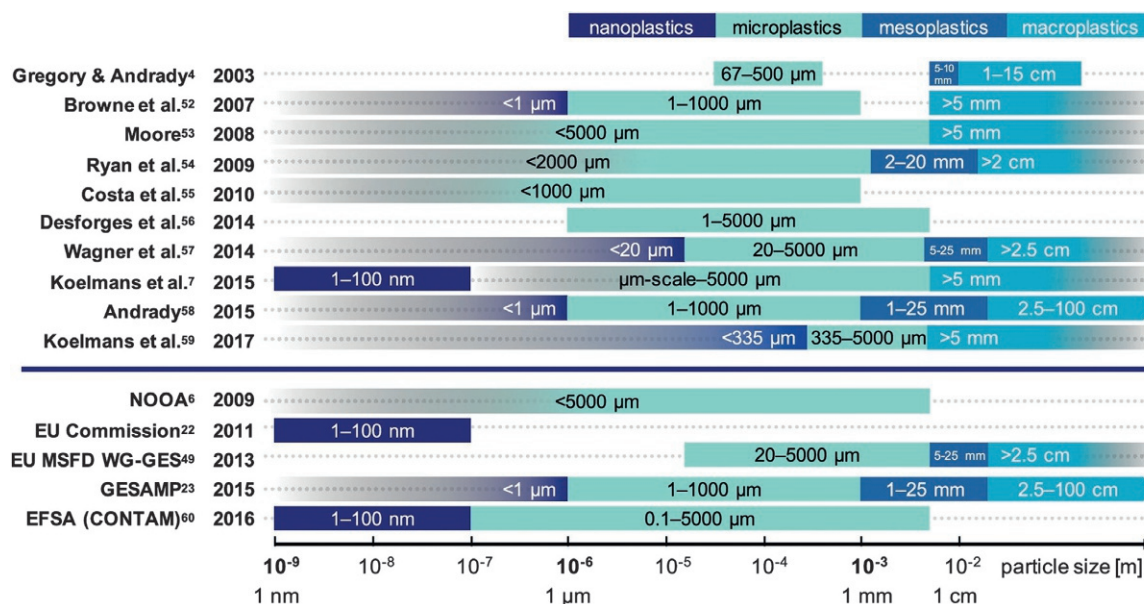


Fig. 6 Examples of the categorization of plastic debris according to size as defined in scientific literature. Adapted from Hartmann NB, T Hüffer, RC Thompson, M Hassellöv, A Verschoor, AE Daugaard, S Rist, T Karlsson, N Brennholt, M Cole, MP Herrling, MC Hess, NP Ivleva, AL Lusher and M Wagner (2019) Are we speaking the same language? Recommendations for a definition and categorization framework for plastic debris. *Environmental Science & Technology*, 53(3): 1039–1047.

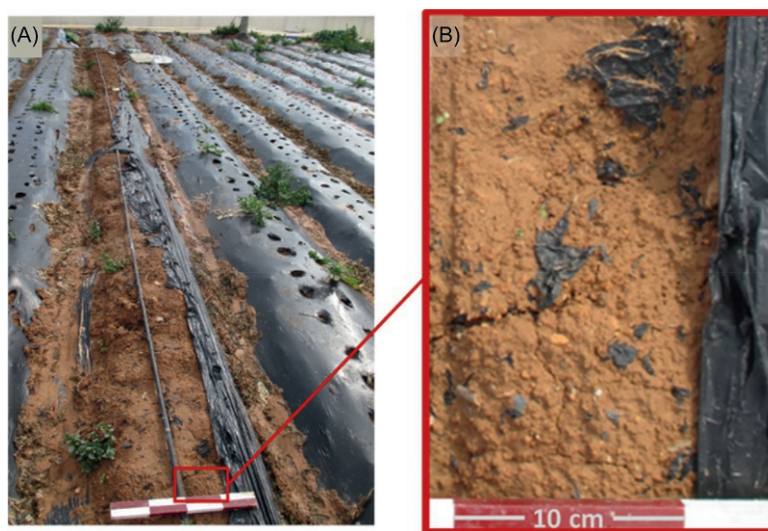
The degradation of plastic produces smaller and smaller debris. Debris size is frequently classified into three categories: nanoplastics < microplastics < macroplastics (Hartmann et al., 2019). However, the size of plastic debris is, in fact, a continuous variable and no defined boundaries exist for these categories (Fig. 6). Most often the boundary between microplastics and macroplastics will be based on the method of identification used, and so varies from one study to the other from 500 µm to 5 mm.

Plastic debris can also be classified on their origin. As plastics have many functions, plastic debris have many different origins. For agricultural sources we can distinguish direct sources: the plastic is used for the beneficial action it is expected to offer; and indirect sources: the plastic is transported to the field without expecting a benefit from it. In addition, when debris is small, it is interesting to know if it was manufactured like this or has been degraded. Therefore, we distinguish primary microplastic from secondary microplastic which are particles derived from the fragmentation of bigger debris. Secondary microplastics have gone through degradation, which has modified their size, shape, physical and or chemical properties.

Examples of direct sources of plastic in agriculture are plastic mulch, irrigation pipes and plastic coatings (Table 1). The degradation of plastic mulch and irrigation pipes into smaller debris is a direct source of secondary microplastics (Fig. 7), whereas fertilizers coated in plastic are a direct source of primary microplastics. Organic fertilizers such as sewage sludge (Corradini et al., 2019) and compost are examples of indirect sources of microplastics. However, sewage sludge can contain primary microplastics from cosmetics and compost can contain secondary microplastics from the fragmentation of a wrongly sorted plastic item. When

Table 1 Examples of direct and indirect sources of primary and secondary microplastics in agriculture.

		Degradation stage of the plastic	
		Primary microplastics	Secondary microplastics
Origin of the plastic	Direct source	Microplastic coating for slow release fertilizers	Degradation of agricultural plastics in the soil, such as twines, plastic mulch, irrigation pipes
	Indirect source	Microplastic from personal care products in sewage sludge	Tire fragment in sewage sludge, wrongly sorted plastic packaging in compost, oversight contamination


Fig. 7 Light density polyethylene plastic mulch after harvest of Kohlrabi in Southeast Spain. The plastic mulch covers an irrigation pipe to improve the water use efficiency (Panel A). The sides of the mulch film are buried into the soil making complete removal impossible and leading to debris accumulation (Panel B).

they are applied to the field for fertilization, contaminated sewage sludge and compost bring with them microplastics that are of no use to the farmers.

The occurrence of plastic in agricultural soils has been described in many studies, with a range from 0 particles kg^{-1} to 50,000 particles kg^{-1} (e.g., van den Berg et al., 2020). For example, it has been found in soils after plastic mulching (Meng, 2021), and compost (van Schothorst et al., 2021) and sewage sludge applications (van den Berg et al., 2020). In China, the extensive use of low density polyethylene (LDPE) plastic mulch is leading to worrying levels of residues (Liu et al., 2014). Detrimental effects of plastic residues have been observed when the contents in soil exceed 270 kg ha^{-1} (Gao et al., 2019). Estimated a 3% decrease in crop yield for every additional 100 kg ha^{-1} of film residue in soil.

The LDPE is the most used polymer for plastic mulches; it has to be removed from the fields after harvest and incomplete removal leads to plastic debris accumulation (for appropriate reference please see Further reading). Some plastic producers have tried to improve the degradation processes of plastic mulch by adding pro-oxidant additives to LDPE to avoid the need for plastic mulch removal and thus, the accumulation of plastic debris. With these aims in mind, pro-oxidant additive containing (PAC) plastic, also called "oxo-degradable" or "oxo-biodegradable," was developed but it has proved to degrade poorly in soils (Hogg et al., 2016) leading to it being banned in June 2019 by the European Parliament. Another attempt to solve the problem of plastic debris accumulation from mulch has resulted in the production of biodegradable (BIO) plastic polymers. Biodegradable mulch can be made from a diversity of biobased or petroleum-based polymers or a blend of both (Sintim et al., 2020). Biodegradable mulches are expected to degrade into water, CO_2 , methane, energy and new biomass in soil with the help of microorganisms. These mulches were designed so that they would degrade by 90% after 2 years in the soil (Standard ISO 16929). However, research has found that biodegradable plastic mulch does not degrade as fast as expected under some field conditions, and leaves residues in the soil after more than 2 years (Sintim et al., 2020). In addition to its potential incomplete degradation, BIO plastic mulch also presents a threat to plant growth. Two studies have found a reduction in the production of wheat (Qi et al., 2018) and common beans (Meng, 2021) in the presence of 2% (w/w) of BIO residues added to the soil. Processes leading to the decrease in crop production are not well understood yet and need more investigation.

One explanation for the decrease in crop production could be the reduction of beneficial microorganisms or the increase of plant pathogens in the presence of BIO plastics. Plastic debris can affect the soil microbiome, and BIO plastics have a stronger potential to do so because they are degraded by microorganisms (Sander, 2019). Plastic debris could be habitats for microorganisms, with the

formation of a biofilm, for example, (Amaral-Zettler et al., 2020). In addition, plastics could be a source of nutrients, mostly carbon. Several organisms have been reported to degrade plastic, even LDPE. Finally, plastic could release chemicals that are toxic for some microorganisms. As explained previously, we have barely any information about the chemicals present in commercial plastic and they could be released due to plastic degradation in the soil. We know that some additives can directly affect the soil microbiome (for appropriate reference please see Further reading).

Plastics and their additives can affect other organisms. Study of the exposure of soil animals to microplastics showed that many organisms do ingest plastics. However, ecotoxicological effects at environmental concentrations are still uncertain (Ng et al., 2018). Microplastics ingested by organisms can travel into the food chain. For example, microplastic concentrations increased from soil (0.87 ± 1.9 particles g^{-1}), to earthworm casts (14.8 ± 28.8 particles g^{-1}), to chicken feces (129.8 ± 82.3 particles g^{-1}) in home gardens in southeast Mexico (Huerta Lwanga et al., 2017). An increase in earthworm mortality has been recorded when exposed to 28%, 45%, and 60% of microplastics in litter (Huerta Lwanga et al., 2016), but these plastic contents are higher than average environmental contents. Moreover, nanoplastics decreased the growth, locomotor activity, and intestinal microbiota viability of snails that were feeding on plants grown in soil with 10 mg kg^{-1} and 100 mg kg^{-1} nanoplastics added (Chae and An, 2020). The effects of plastics on terrestrial organisms can also be caused by additives and other contaminants that may leak from the plastic debris (for appropriate reference please see Further reading). Many plastic additives are suspected of being endocrine disruptors. This means that they may interfere with animal hormones and therefore impact the entire organism (for appropriate reference please see Further reading). Plastic debris is also ingested by larger organisms resulting in various negative effects. These effects include blockage of the intestinal tract, inhibition of gastric enzyme secretion, reduced feeding stimuli, decreased steroid hormone levels, delays in ovulation and even failure to reproduce. Though it is still difficult to draw definitive conclusions from the literature, early studies about plastics in soil concur with the wider base of aquatic plastic toxicology in the sense that plastics are a threat to soil biota (Helmberger et al., 2020). Toxicity of aquatic plastic debris has been better described and marine animals like sea turtles illustrate well the problem of plastic contamination in the environment. Water and wind transport link land and aquatic plastic contamination, making plastic debris a ubiquitous contaminant.

Uncertainties shrouding plastic content assessment in soil

Microplastic residues are more or less everywhere, but they are not easy to spot. However, it is not necessarily their size but their diversity which makes them difficult to detect. We cannot over emphasize that plastics are diverse and complex contaminants. Commercial plastics can be composed of a blend of different polymers and contain several additives. Plastic additives have been proved to be toxic, possibly even more toxic than the main polymer (for appropriate reference please see Further reading). The detection of plastic additives would rely on targeted screening of likely chemicals, which occurs for pesticides for example. However, additives are the secret ingredients used by industries making targeted screening difficult. If we leave out additives for simplicity, we can apply various techniques to assess plastic contamination. For plastic polymers we can measure the number of particles, contact surface, mass of debris, type of polymer and their degradation over time.

The extraction method of Zhang et al. (2018) was developed for LDPE and consisted of a stereomicroscope visual selection (SMVS). The microplastics are identified visually based on their shape, color, brightness, and response to heat. The SMVS method has the advantage of being relatively fast and cheap, but relies on observer judgment and attention which can vary with different users. Spectral methods (e.g., Raman or Fourier transform infrared) are used to identify several types of plastic by comparing the spectra of the plastics analyzed with a spectral library. Spectral methods provide a 2-D image of the sample with number, surface area and best polymer type matched in the library for each particle (Corradini et al., 2021). Other methods do not directly report the size or number of particles but analyze the mass of all plastic present in the sample. This is the case for thermal extraction desorption gas chromatography mass spectrometry (TED GC-MS) (for appropriate reference please see Further reading) and pyrolysis–gas chromatography–mass spectrometry (Py-GC-MS). Essentially, plastic particles are burned and the molecules emitted are analyzed by gas chromatography. It is a sensitive and well-established method for the characterization and mass-quantification of many polymer types and their organic additives (for appropriate references of this section see Further reading).

Each method relies on a different property of the plastic and therefore is likely to underestimate the total plastic content. In addition, each method is suitable for a certain particle size. For example, with SMVS, smaller particles are more difficult to identify than larger ones simply because they are more difficult to detect. It is another factor suggesting underestimation of the total plastic number. In general, the smaller the plastic becomes the more difficult it is to analyze. The identification of nanoplastics remains a challenge and new methods have to be validated in environmental samples. Raman spectroscopy and pyrolysis–gas chromatography/mass spectrometry (py-GC-MS) are the most promising methods. In addition, sample pre-processing and the method of plastic extraction are also major factors for microplastic analysis. To summarize, the diversity of plastics, of matrices (e.g., soils, composts, feces) and of variables (number, area, mass, polymer type) mean that there is no single method able to answer all questions (for appropriate references of this section see Further reading).

A combination of different methods could provide a better description of environmental microplastic samples. For example, to compare the quantity of plastic left after spiking known amounts of plastic in soil, the best option could be to use a thermal desorption method to estimate the mass balance and to combine it with the SMVS method to investigate evolution of the plastic size distribution. Spectral methods will be less relevant in this case since the type of polymer applied would already be known. On the other hand, a spectral method is recommended to identify the type of plastic in field assessments. In summary, we

recommend adapting the method of assessment to the specific interest of the study. For example, to estimate the probability of particle ingestion the amount of debris will be most important (Helmberger et al., 2020); for the formation of biofilms, the total area will be the dominant factor (Sander, 2019); and for input/output balances the mass is generally favored. Moreover, many processes will be influenced by the size of the debris. For example, a study on plastic ingestion may focus on millimeter-size debris for mammals, micrometer-size for earthworms (Helmberger et al., 2020) and nanometer-size for plant roots (Chae and An, 2020). Now that we have a better idea of what we are measuring, what are the expected consequences for environmentally relevant plastic contents?

Effects of plastic residues on the environment

Effect on soil physicochemical properties

Various studies have tested the effects of plastic debris on soil physical properties. Qi et al., 2020 tested macro- and micro-plastic particles from the addition of LDPE and BIO mulches (0.5%, 1% and 2% *w/w*) to a sandy soil. They reported a decrease in field capacity with LDPE debris and an increase with BIO debris suggesting a strong effect of the type of plastic on water retention. No effects were observed for the soil pH, electrical conductivity and aggregate stability. Under various conditions de Souza Machado et al. (2019) observed a reduction of aggregate stability in the presence of polyamide beads (diameter of 15–20 μm) and polyester fibers (~5000 μm in length and ~8 μm in diameter). In another study they also showed that polyester and polyacrylic fibers reduced aggregate stability. They suggested that the fiber shape, as an elastic and linear particle creates tension inside the aggregates (Fig. 8). Apart from confirming the effect of plastic on soil properties, these findings highlight the fact that different results can be expected from different plastic particle polymers, shapes and sizes. In addition, most studies emphasize that aging of the plastic in the soil will possibly modify the interaction with soil particles; and soil properties (i.e., soil density, infiltration, water holding capacity De Souza Machado et al., 2019), are closely related with soil type and texture (Yang et al., 2020). Therefore, further research is needed to focus on different soil types, plastic shapes and types.

Effect on the soil microbiome

Meng et al. (2023) and Qi et al. (2020) observed a modified soil bacterial community after the introduction of plastic mulch debris. There are different hypotheses that could explain the effect. First, the plastic may alter the soil physical properties, which could then modify the soil microbiome. For example, if biodegradable plastic could increase the soil water retention as reported in Qi et al. (2020), organisms adapted to higher water content would thrive. Another hypothesis is that plastic could serve as a habitat for certain microorganisms, with the formation of a biofilm (Amaral-Zettler et al., 2020). In addition, plastics could be a source of nutrients, mostly carbon. Several organisms have been reported to degrade plastic even LDPE. Finally, plastic could release chemicals that are toxic for some microorganisms. As mentioned above, we have little information about the chemicals present in the commercial plastics that could be released with plastic degradation in the soil. It is known that some plastic additives can directly affect the soil microbiome (for appropriate reference see Further reading).

Effect on plants

Qi et al. (2018) and Meng (2021) also observed a significant effect of BIO plastic but not of LDPE on wheat (*Triticum aestivum* L.) and common bean (*Phaseolus vulgaris* L.), respectively. The impact of BIO could come from effects discussed earlier such as change in

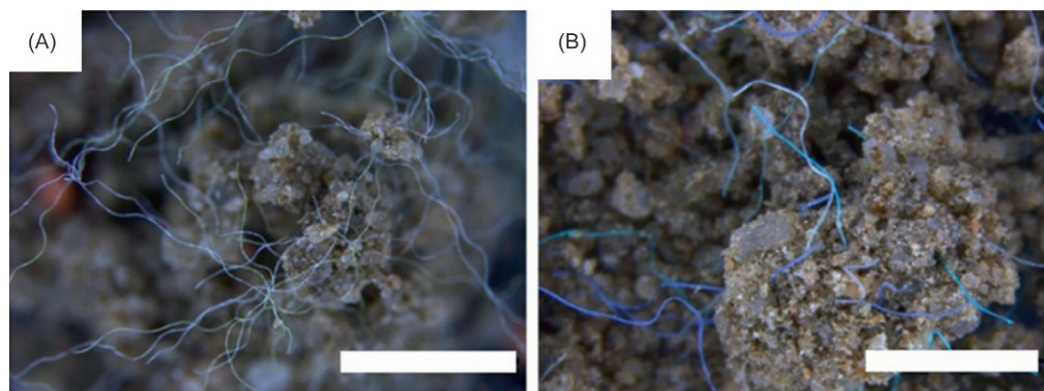


Fig. 8 Polyester fibers (A) and polyacrylic fibers (B) incorporated into soil aggregates. The white bar in each panel represents 1 mm size. Adapted from de Souza Machado AA, CW Lau, J Till, W Kloas, A Lehmann, R Becker and MC Rillig (2018) Impacts of microplastics on the soil biophysical environment. *Environmental Science & Technology* 52(17): 9656–9665.

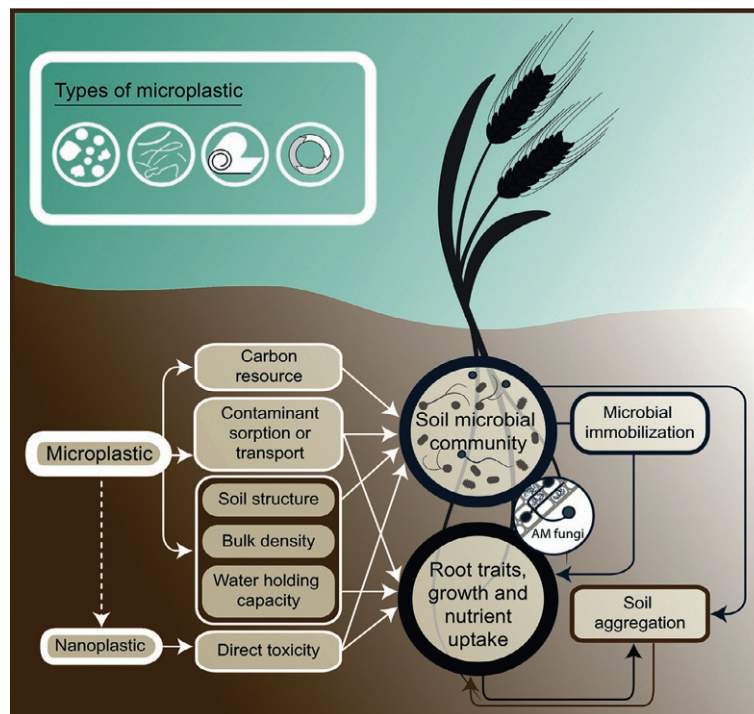


Fig. 9 Effects of microplastics on plants. Taken from Rillig MC, Lehmann A, de Souza Machado AA, & Yang G (2019) Microplastic effects on plants. *New Phytologist*, 223(3): 1066–1070.

the soil physical properties or in the microbiome (Meng et al., 2023). There is uncertainty about whether beneficial or pathogenic organisms might be suppressed or enhanced by the BIO plastic. Another theory could be nutrient competition between the organisms degrading the BIO plastic and the plant. For instance, Meng (2021) reported less nitrate but more organic nitrogen in soils with 2% BIO plastic. This effect would depend on the fertilization applied and degradation of the plastic. More research is required for greater understanding. Moreover, a direct effect of plastics on the plant should be considered. Rillig et al., 2019 (Fig. 9) describe effects on soil conditions that consequently affect plants (i.e., soil microbial community, root traits). Bosker et al. (2019) reported that nanoplastics blocked the pores of cress (*Lepidium sativum* L.) seed capsules and reduced the rate of germination. Nanoplastics could also potentially enter the roots leading to oxidative stress in the plant (Giorgetti et al., 2020). As they are more easily degraded it is likely that BIO plastics generate more nanoplastics, which explains why they have more impact on plant growth. Finally, chemicals leaked from the BIO plastic also have unknown consequences for the plant. All these hypotheses remain unverified and could be combined to explain detrimental effects of plastic residues on the plant. If plastic ends up in plants we need to consider the consequences of plastic further up the food chain.

Effect on soil invertebrates and vertebrates

Recent studies have clearly indicated how microplastics produce changes in soil macroinvertebrates. Earthworms, for example, ingest microplastics (Huerta Lwanga et al., 2016), and carry them in and on their bodies (Huerta Lwanga et al., 2016; Rillig et al., 2017). The effects observed are weight loss, oxidative stress (Rodríguez-Seijo et al., 2018), histological damage (Rodríguez-Seijo et al., 2018) and mortality (Huerta Lwanga et al., 2016).

Larger concentrations of microplastics were found inside earthworm biogenic structures compared to the bulk soil. They also transport microplastics into the deeper soil layers (Rillig et al., 2017; Huerta Lwanga et al., 2022). In addition, other invertebrates such as snails, springtails, and mites are also affected by the presence of microplastics; they carry the microplastics inside them (Song et al., 2019) and displace them, and the predator–prey relationship in mites is affected. When this contaminant is present, macro and meso invertebrates are also affected, and there is a response or adaptation that we need to understand better about which processes are involved in these adaptations.

Vertebrates such as sheep, vultures and chicken are vulnerable to microplastics through their ingestion (Beriot et al., 2021; Borges-Ramírez et al., 2021; Huerta Lwanga et al., 2017). There are few studies that have reported how such animals, which also interact with agricultural systems (sheep), ingest the plastic mulch products if they are not removed, and then carry them in their bodies to nearby sites.

Effect on soil ecosystem services

When earthworms or macroinvertebrates lose weight, or are less active in decomposing organic matter, promoting infiltration or aggregation (Rillig et al., 2019, de Souza Machado et al., 2019, fig. 9) there are adverse effects on ecosystem services in terms of change and degradation. The incorporation of organic matter from the soil surface to the deeper soil layers is also altered, as the earthworms transport microplastics from the soil surface to the soil matrix, which contaminate the tunnels and burrows they build. The nutrient cycle and water retention are also affected when microplastics are introduced by earthworms (Huerta Lwanga et al., 2022). Moreover, the number and sizes of burrows are also modified when plastic particles are on the soil surface (for appropriate reference see Further reading). More research is necessary to understand this process better at the level of the earthworm burrows or galleries, together with the microplastics and the microorganisms that inhabit them.

Less germination of *Lolium perenne* L. has been noticed together with weak soil aggregate stability when LDPE microplastics were present in soils, PLA also decreased earthworm biomass in combined systems with this plant (for appropriate references see Further reading).

Transport of microplastics

The transport of microplastics is not well described in many of the existing studies, nevertheless it is very probable that microplastics are everywhere; there are not pristine areas without microplastics. They have been found in remote areas such as in the Swiss Alps or the Arctic (for appropriate reference please see Further reading), even in those places where there are no clear traces of human activities.

The transport of microplastics in terrestrial environments is carried out by abiotic and biotic factors (Huerta Lwanga et al., 2022); among the most important abiotic factors are air and water. Once microplastics are present on the soil surface they are moved by the wind (soil erosion), or run off by the water, leached or moved beyond.

Abiotic transport

The transport of microplastics is size dependent; it is estimated that small particles such as micro or nanoplastics are highly transportable compared to macroplastics. In agricultural areas, once the soil is polluted with microplastics they will move from the soil surface into deeper layers according to the type of soil matrix and follow preferential flow paths (Yu et al., 2019; Huerta Lwanga et al., 2022). There is a need to understand how the processes are taking place, and which types of plastics, soils, and sizes are involved in microplastic transport (Huerta Lwanga et al., 2022). Transport of microplastics to the ground water has already been reported. Polyethylene, polyester, and polyvinylchloride microplastics (50–150 µm) have been found in ground water, and in drinking water, representing a potential health risk (for appropriate reference please see Further reading). Wind transport of microplastics has been reported (Rezaei et al., 2019); this involves the fibers which are more susceptible to this type of transport.

Biotic transport

Among biotic factors, microplastics are being transported through the biogenic structures of invertebrates and vertebrates in the soil (feces), through the tunnels or burrows in the case of earthworms and through the bodies of invertebrates or vertebrates. Microplastics are even taken as “food” in the case of springtails (Maaß et al., 2017), or snails (Song et al., 2019).

Once the invertebrates in soil are in contact with microplastics, depending on the concentrations and time of exposure, they become impregnated with the particles and electrostatically transport them horizontally or vertically (Fig. 10). Earthworms ingest microplastics together with leaf litter and transport them from the soil surface to the lower soil layers, in particular those earthworms that form vertical tunnels, (i.e., the microplastics in hotspots, Huerta Lwanga et al., 2022, Rillig et al., 2017).

The situation is similar for terrestrial vertebrates such as chickens (Huerta Lwanga et al., 2017), sheep (Beriot et al., 2021), and vultures (Borges-Ramírez et al., 2021), which, once they have ingested the plastics, carry them in their bodies until they are defecated or regurgitated.

It becomes more complicated when the plastic particles are combined with other substances such as pesticides, metals or the same additives that are added to plastics to make them more flexible or resistant, then the microplastics carry other pollutants (Huerta Lwanga et al., 2022). Nevertheless, this brief chapter cannot discuss the interactions that take place between plastic particles and other pollutants. During transport, the interaction between plastics and pollutants such as pesticides play a significant role. For example, pesticides can be adsorbed or not according to the type of pesticide and the type of plastic, a situation that might occur at industrial agricultural sites due to the use of plastics, agrochemicals and pesticides. More research is required to understand these phenomena.

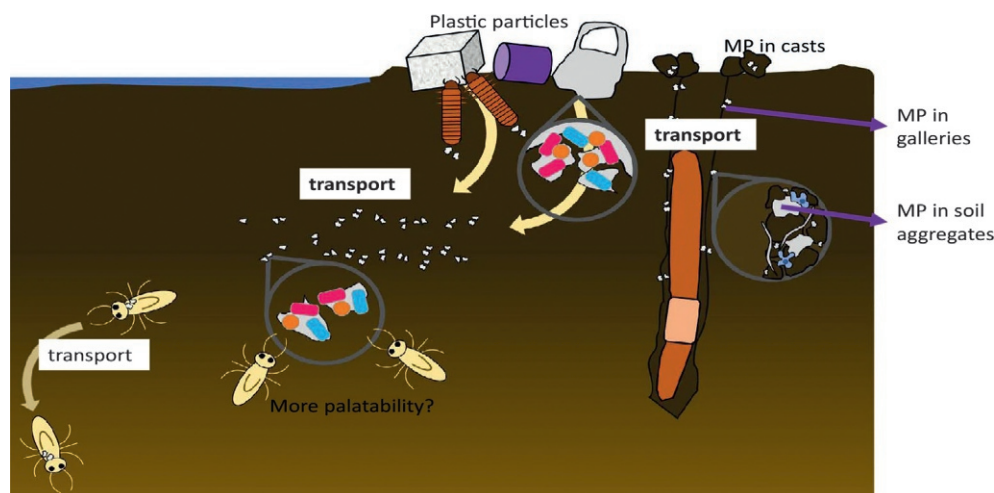


Fig. 10 Microplastic (MP) transport by soil invertebrates. Taken from Helmberger MS, Tiemann LK, & Grieshop MJ (2020) Towards an ecology of soil microplastics. *Functional Ecology*, 34(3): 550–560 and modified by Huerta Lwanga EH, Beriot N, Corradini F, Silva V, Yang X, Baartman J, . . . & Geissen V (2022) Review of microplastic sources, transport pathways and correlations with other soil stressors: A journey from agricultural sites into the environment. *Chemical and Biological Technologies in Agriculture*, 9(1): 1–20.

Microplastics and humans

It has recently been reported that humans contain microplastics in their feces, and Leslie et al., 2022 reported $1.6 \mu\text{g mL}^{-1}$ of plastic particles in human blood. The main plastic residues were polyethylene terephthalate, polyethylene and polymers of styrene. The implications for humans being exposed to different types of plastics, and also to different pesticides is yet to be explored. The toxicity of plastic particles to soil invertebrates (weight loss, and mortality depending on concentrations and time of exposure) is known, but not in humans. Prata et al., 2020, described how exposure of humans to microplastics by ingestion, dermal contact, and respiration (Fig. 11) can affect various human systems (parasympathetic, and sympathetic). Genotoxic and cytotoxic effects of plastic particles in blood cells were explored by Çobanoğlu et al. (2021); the authors concluded that microplastics can generate genetic instability in blood cells, but they did not observe a clear cytotoxic effect on the studied lymphocytes, and recommend more



Fig. 11 Microplastics effects on humans. After Prata JC, da Costa JP, Lopes I, Duarte AC, & Rocha-Santos T (2020) Environmental exposure to microplastics: An overview on possible human health effects. *Science of the Total Environment*, 702: 134455.

research. It is already reported that microplastic fibers have been detected already in lung biopsies, and the authors consider that injury or death is expected if people are chronically exposed to airborne microplastics (for appropriate reference please see Further reading). The sources of airborne microplastics are different: among them are fibers from synthetic clothes, erosion and synthetic rubber tires. Further research is required to understand the effects of microplastics on humans, under different conditions, and when other stressors are also present as pesticides or heavy metals.

Concluding comments

Microplastics are particles smaller than 5 mm, and are emergent pollutants. They are present in most agricultural soils tested. They have mainly entered agricultural systems due to the fragmentation of plastic materials used in agriculture (e.g., plastic mulches, greenhouses, twines), the application of inputs to the soil (e.g., plastic coated fertilizers, sludge, compost) or through external contamination (e.g., atmospheric deposition, dumping). Microplastics are transported by biotic and abiotic factors, disseminated along the food chain, and result in strong potential environmental and human risks.

To summarize, much still remains to be studied and understood with respect to microplastics. They are not static or isolated, and also carry other pollutants, and consequently various interactions take place in the soil, with or without the organisms that inhabit it. We invite the reader to go deeper into the subject by consulting some sites of interest like those included in European projects such as the H2020 projects MINAGRIS (www.minagris.eu) and PAPILLONS (www.papillons-h2020.eu).

Acknowledgments

Research is a team work, therefore we thank all those who have done research with us and NWO women in STEM (19342), and also we thank all investigations related with microplastics in this area, because every day we understand more how microplastics interact with the environment, and so we expect in the near future to find solutions for mitigating this pollution.

We also thank you “the readers of this chapter,” because we firmly believe that more heads think better than one (Mexican proverb), and much better when those heads are informed, please share this information with your students and colleagues.

References

- Amaral-Zettler LA, Zettler ER, and Mincer TJ (2020) Ecology of the plastisphere. *Nature Reviews Microbiology* 18(3): 139–151.
- Beriot N, Peek J, Zornoza R, Geissen V, and Lwanga EH (2021) Low density-microplastics detected in sheep faeces and soil: A case study from the intensive vegetable farming in Southeast Spain. *Science of the Total Environment* 755: 142653.
- Borges-Ramírez MM, Escalona-Segura G, Huerta-Lwanga E, Iñigo-Elias E, and Rendón-von Osten J (2021) Organochlorine pesticides, polycyclic aromatic hydrocarbons, metals and metalloids in microplastics found in regurgitated pellets of black vulture from Campeche, Mexico. *Science of the Total Environment* 801: 149674.
- Bosker T, Bouwman LJ, Brun NR, Behrens P, and Vijver MG (2019) Microplastics accumulate on pores in seed capsule and delay germination and root growth of the terrestrial vascular plant *Lepidium sativum*. *Chemosphere* 226: 774–781.
- Chae Y and An Y-J (2020) Nanoplastic ingestion induces behavioral disorders in terrestrial snails: Trophic transfer effects via vascular plants. *Environmental Science: Nano* 7(3): 975–983.
- Çobanoğlu H, Belivermiş M, Sıkdokur E, Kılıç Ö, and Çayır A (2021) Genotoxic and cytotoxic effects of polyethylene microplastics on human peripheral blood lymphocytes. *Chemosphere* 272: 129805.
- Corradini F, Meza P, Eguiluz R, Casado F, Huerta-Lwanga E, and Geissen V (2019) Evidence of microplastic accumulation in agricultural soils from sewage sludge disposal. *Science of the Total Environment* 671: 411–420.
- Corradini F, Beriot N, Huerta-Lwanga E, and Geissen V (2021) uFTIR: An R package to process hyperspectral images of environmental samples captured with μ FTIR microscopes. *SoftwareX* 16: 100857.
- de Souza Machado AA, Lau CW, Kloas W, Bergmann J, Bachelier JB, Faltin E, and ... Rillig MC (2019) Microplastics can change soil properties and affect plant performance. *Environmental Science & Technology* 53(10): 6044–6052.
- FAO (2021) *Assessment of Agricultural Plastics and Their Sustainability—A Call for Action*. Rome, Italy: FAO.
- Gao H, Yan C, Liu Q, Ding W, Chen B, and Li Z (2019) Effects of plastic mulching and plastic residue on agricultural production: A meta-analysis. *Science of the Total Environment* 651: 484–492.
- Giorgetti L, Spanò C, Muccifora S, Bottega S, Barbieri F, Bellani L, and Ruffini Castiglione M (2020) Exploring the interaction between polystyrene nanoplastics and *Allium cepa* during germination: Internalization in root cells, induction of toxicity and oxidative stress. *Plant Physiology and Biochemistry* 149: 170–177.
- Hartmann NB, Hüffer T, Thompson RC, Hasselöv M, Verschoor A, Daugaard AE, Rist S, Karlsson T, Brennholt N, Cole M, Herrling MP, Hess MC, Ivleva NP, Lusher AL, and Wagner M (2019) Are we speaking the same language? Recommendations for a definition and categorization framework for plastic debris. *Environmental Science & Technology* 53(3): 1039–1047.
- Helmlinger MS, Tiemann LK, and Grieshop MJ (2020) Towards an ecology of soil microplastics. *Functional Ecology* 34(3): 550–560.
- Hogg DG, Adrian, Hann S, and Ettlinger S (2016) *The Impact of the Use of “Oxo-Degradable” Plastic on the Environment*. E. C. D.-G. f. Environment.
- Huerta Lwanga E, Gertsen H, Gooren H, Peters P, Salánki T, Van Der Ploeg M, and ... Geissen V (2016) Microplastics in the terrestrial ecosystem: Implications for *Lumbricus terrestris* (Oligochaeta, Lumbricidae). *Environmental Science & Technology* 50(5): 2685–2691.
- Huerta Lwanga E, Mendoza Vega J, Ku Quej V, Chi JDLA, Sanchez del Cid L, Chi C, and ... Geissen V (2017) Field evidence for transfer of plastic debris along a terrestrial food chain. *Scientific Reports* 7(1): 1–7.
- Huerta Lwanga EH, Beriot N, Corradini F, Silva V, Yang X, Baartman J, and ... Geissen V (2022) Review of microplastic sources, transport pathways and correlations with other soil stressors: A journey from agricultural sites into the environment. *Chemical and Biological Technologies in Agriculture* 9(1): 1–20.
- Leslie HA, Van Velzen MJ, Brandsma SH, Vethaak D, Garcia-Vallejo JJ, and Lamoree MH (2022) Discovery and quantification of plastic particle pollution in human blood. *Environment International*, 107199.

- Liu EK, He WQ, and Yan CR (2014) 'White revolution' to 'white pollution'—Agricultural plastic film mulch in China. *Environmental Research Letters* 9(9): 091001.
- Maaß S, Daphi D, Lehmann A, and Rillig MC (2017) Transport of microplastics by two collembolan species. *Environmental Pollution* 225: 456–459.
- Meng F (2021) *From Plastic Mulching to Microplastic Pollution: An Effect Assessment of Microplastics in the Soil-Plant System*. Doctoral dissertation Wageningen University and Research.
- Meng F, Harkes P, van Steenbrugge JJ, and Geissen V (2023) Effects of microplastics on common bean rhizosphere bacterial communities. *Applied Soil Ecology* 181: 104649.
- Napper IE and Thompson RC (2019) Environmental deterioration of biodegradable, oxo-biodegradable, compostable, and conventional plastic carrier bags in the sea, soil, and open-air over a 3-year period. *Environmental Science & Technology* 53(9): 4775–4783.
- Ng E-L, Huerta Lwanga E, Eldridge SM, Johnston P, Hu H-W, Geissen V, and Chen D (2018) An overview of microplastic and nanoplastic pollution in agroecosystems. *Science of the Total Environment* 627: 1377–1388.
- Prata JC, da Costa JP, Lopes I, Duarte AC, and Rocha-Santos T (2020) Environmental exposure to microplastics: An overview on possible human health effects. *Science of the Total Environment* 702: 134455.
- Qi Y, Yang X, Pelaez AM, Huerta Lwanga E, Beriot N, Gertsen H, Garbeva P, and Geissen V (2018) Macro- and micro- plastics in soil-plant system: Effects of plastic mulch film residues on wheat (*Triticum aestivum*) growth. *Science of the Total Environment* 645: 1048–1056.
- Qi Y, Ossowicki A, Yang X, Huerta Lwanga E, Dini-Andreote F, Geissen V, and Garbeva P (2020) Effects of plastic mulch film residues on wheat rhizosphere and soil properties. *Journal of Hazardous Materials* 387: 121711.
- Rezaei M, Riksen MJ, Sirjani E, Sameni A, and Geissen V (2019) Wind erosion as a driver for transport of light density microplastics. *Science of the Total Environment* 669: 273–281.
- Rillig MC, Ingrassia R, and de Souza Machado AA (2017) Microplastic incorporation into soil in agroecosystems. *Frontiers in Plant Science* 8: 1805.
- Rillig MC, Lehmann A, de Souza Machado AA, and Yang G (2019) Microplastic effects on plants. *New Phytologist* 223(3): 1066–1070.
- Rodríguez-Seijo A, da Costa JP, Rocha-Santos T, Duarte AC, and Pereira R (2018) Oxidative stress, energy metabolism and molecular responses of earthworms (*Eisenia fetida*) exposed to low-density polyethylene microplastics. *Environmental Science and Pollution Research* 25(33): 33599–33610.
- Sander M (2019) Biodegradation of polymeric mulch films in agricultural soils: Concepts, knowledge gaps, and future research directions. *Environmental Science and Technology* 53: 2304–2315.
- Sintim HY, Bary AI, Hayes DG, Wadsworth LC, Anunciado MB, English ME, Bandopadhyay S, Schaeffer SM, DeBruyn JM, Miles CA, Reganold JP, and Flury M (2020) In situ degradation of biodegradable plastic mulch films in compost and agricultural soils. *Science of the Total Environment* 727: 138668.
- Song Y, et al. (2019) Uptake and adverse effects of polyethylene terephthalate microplastics fibers on terrestrial snails (*Achatina fulica*) after soil exposure. *Environmental Pollution* 250: 447–455.
- van den Berg P, Huerta-Lwanga E, Corradini F, and Geissen V (2020) Sewage sludge application as a vehicle for microplastics in eastern Spanish agricultural soils. *Environmental Pollution* 261: 114198.
- van Schothorst B, Beriot N, Lwanga EH, and Geissen V (2021) Sources of light density microplastic related to two agricultural practices: The use of compost and plastic mulch. *Environments* 9.
- Yang X, Guo X, Huang S, Xue S, Meng F, Qi Y, Cheng W, Fan T, Lwanga EH, and Geissen V (2020) Microplastics in soil ecosystem: Insight on its fate and impacts on soil quality. In: He D and Luo Y (eds.) *Microplastics in Terrestrial Environments: Emerging Contaminants and Major Challenges*, pp. 245–258. Cham: Springer International Publishing.
- Yu M, Van Der Ploeg M, Lwanga EH, Yang X, Zhang S, Ma X, and ... Geissen V (2019) Leaching of microplastics by preferential flow in earthworm (*Lumbricus terrestris*) burrows. *Environmental Chemistry* 16(1): 31–40.
- Zhang S, Yang X, Gertsen H, Peters P, Salánki T, and Geissen V (2018) A simple method for the extraction and identification of light density microplastics from soil. *Science of the Total Environment* 616: 1056–1065.

Further reading

In this section you will find references of complementary readings to the text, and also references that reinforce the arguments presented in this chapter, we invite you to dive deeper into the topic of microplastics contamination in agricultural systems.

- Asma C and Kaouthar L-G (2017) Population dynamics of the tomato leaf miner *Tuta absoluta* (Meyrick) (Lepidoptera: Gelechiidae) in Tunisia natural conditions. *Journal of Entomology and Zoology Studies* 5(4): 427–432.
- Bergmann M, Mützel S, Primpke S, Tekman MB, Trachsel J, and Gerdtts G (2019) White and wonderful? Microplastics prevail in snow from the Alps to the Arctic. *Science Advances* 5(8): eaax1157.
- Boots B, Russell CW, and Green DS (2019) Effects of microplastics in soil ecosystems: Above and below ground. *Environmental Science & Technology* 53(19): 11496–11506.
- Bullard JE, Ockelford A, O'Brien P, and Neuman CM (2021) Preferential transport of microplastics by wind. *Atmospheric Environment* 245: 118038.
- Cai H, Xu EG, Du F, Li R, Liu J, and Shi H (2021) Analysis of environmental nanoplastics: Progress and challenges. *Chemical Engineering Journal* 410: 128208.
- de Souza Machado AA, Lau CW, Till J, Kloas W, Lehmann A, Becker R, and Rillig MC (2018) Impacts of microplastics on the soil biophysical environment. *Environmental Science & Technology* 52(17): 9656–9665.
- Dümichen E, Eisenraut P, Bannick CG, Barthel A-K, Senz R, and Braun U (2017) Fast identification of microplastics in complex environmental samples by a thermal degradation method. *Chemosphere* 174: 572–584.
- Hermabessiere L, Dehaut A, Paul-Pont I, Lacroix C, Jezequel R, Soudant P, and Duflos G (2017) Occurrence and effects of plastic additives on marine environments and organisms: A review. *Chemosphere* 182: 781–793.
- Huerta Lwanga E, Gertsen H, Gooren H, Peters P, Salánki T, van der Ploeg M, and ... Geissen V (2017) Incorporation of microplastics from litter into burrows of *Lumbricus terrestris*. *Environmental Pollution* 220: 523–531.
- Jabran K (2019) Use of mulches in agriculture: Introduction and concepts. In: Jabran K (ed.) *Role of Mulching in Pest Management and Agricultural Sustainability*, pp. 1–14. Cham: Springer International Publishing.
- Kong X, Jin D, Jin S, Wang Z, Yin H, Xu M, and Deng Y (2018) Responses of bacterial community to dibutyl phthalate pollution in a soil-vegetable ecosystem. *Journal of Hazardous Materials* 353: 142–150.
- Lear G, Kingsbury JM, Franchini S, Gambarini V, Maday SDM, Wallbank JA, Weaver L, and Pantos O (2021) Plastics and the microbiome: Impacts and solutions. *Environmental Microbiome* 16(1): 2.
- Mai L, Bao L-J, Wong CS, Zeng EY, and Zeng EY (2018) Chapter 12—Microplastics in the terrestrial environment. In: *Microplastic Contamination in Aquatic Environments*, pp. 365–378. Elsevier.
- Moine BL (2018) Worldwide plasticulture—A focus on films. In: *21st CIPA Congress. Arcachon, France*.
- Prata JC (2018) Airborne microplastics: Consequences to human health? *Environmental Pollution* 234: 115–126.
- Steinmetz Z, Wollmann C, Schaefer M, Buchmann C, David J, Tröger J, Muñoz K, Frör O, and Schaumann GE (2016) Plastic mulching in agriculture. Trading short-term agronomic benefits for long-term soil degradation? *Science of the Total Environment* 550: 690–705.

Municipal solid waste management: Production, management, and environmental effects

Bryant M Serre^a and Lynda H McCarthy^b, ^aMcGill University, Montreal, QC, Canada; ^bToronto Metropolitan University (formerly Ryerson University), Toronto, ON, Canada

© 2023 Elsevier Ltd. All rights reserved.

Author Contribution: This encyclopedic entry was conceptualized by BMS and LHM, and the synthesis of literature was conducted by BMS and LHM. The preparation of the original draft, in addition to review and edits, were conducted by BMS and LHM.

Introduction to municipal solid waste (MSW)	321
Definition of municipal solid waste (MSW)	321
History of municipal solid wastes (MSW)	322
MSW production, composition, and elimination	323
Current MSW production in the western world	323
Ongoing MSW strategies	323
Recyclable and reclamation products	325
Organic waste and compost strategies	325
Energy recovery	326
The MSW challenges	331
Concluding remarks	331
References	331

Abstract

Global population expansion generates large volumes of diverse waste. As a subset of global waste, municipal solid waste (MSW) includes household, business, institutional, and construction waste that may be processed, composted, recycled, and or disposed of at managed sites. Modern sustainable development is predicated on transitioning from linear (take–make–dispose) to circular (take–make–reuse) waste economies; however, underlying all waste management, is a concern for human, and aquatic-, and terrestrial- ecosystem health. In response, this chapter describes various waste management strategies in Western societies, including recycling, composting, energy recovery, treatment and disposal. Moreover, the potential impact of each strategy on soils is emphasized.

Key points

- Municipal Solid Waste (MSW) involves diverse waste products, including non-hazardous recyclables, compostables, in addition to household, business, institutional, and construction waste.
- MSW can sorted into streams as compost, recyclable material, biosolids (as a product of wastewater treatment), energy recovery streams (thermal and non-thermal waste-to-energy, pyrolysis, gasification, landfill gas utilization), and should no further resource be able to be extracted, all other wastes are landfilled.
- When successful, engineered solutions can reduce the volume of waste, improve resource recovery, and limit the potential toxicity to soils, and the broader terrestrial and aquatic environment.
- Despite decades of waste reduction education globally, the waste generated per capita continues to increase unsustainably.
- Potential pathways of soils and water contamination include, but are not limited to: (i) landfill leachate (precipitation over landfill products, which may mobilize contaminants into soil, groundwater, and surface waters); (ii) gas production from landfills and waste incineration; (iii) non-compostable wastes and environmentally-deleterious compounds, which may be present in biosolids or composts, being incorporated into soils through land-application.

Introduction to municipal solid waste (MSW)

Definition of municipal solid waste (MSW)

In North America, municipal solid waste (MSW) refers to any non-hazardous or hazardous items that consumers have no further use for and throw away. Municipal solid waste is diverse, including recyclables and compostable materials, as well as waste from households, businesses, institutions, and construction and demolition sites. Once collected, MSW is recycled, processed, or disposed of at managed sites. Similarly, the European Commission defines MSW as “waste from households, as well as other waste which, because of its nature or composition, is similar to waste from households” (European Commission (EUC), 1999).

Following a linear consumption model (i.e., take–make–dispose), Western societies have, for centuries, metabolized raw materials into the food, water, air, and shelter needed to sustain human life, repetitiously creating municipal solid waste (and waste generally) as a by-product.

Municipal solid waste (MSW) is waste predominantly generated by urban populations and collected for disposal by municipal authorities (Tchobanoglous and Kreith, 2002). In Canada, MSW includes waste generated from residential and industrial, commercial, and institutional (ICI) sources. The ICI sources can include wastes from office buildings, shopping malls, schools or hospitals. The MSW can also include (i) organic (putrescible) wastes, which are biodegradable and compostable, such as food scraps, lawn and garden waste, used paper products and cardboard; (ii) recyclable waste such as plastics, metals, paper, and cardboard; (iii) construction, renovation, and demolition (CRD) waste such as wood (clean, engineered, treated, and painted), asphalt roofing and drywall; in addition to (iv) residual materials that comprise wastes that cannot be recycled or composted.

Municipal solid waste differs from liquid waste that is collected through sewage and sanitary networks to be processed subsequently in wastewater treatment plants (WWTP). However, MSW and liquid waste overlap in the creation of processed sewage sludge (i.e., biosolids) that is then managed as solid waste and either (i) land-applied, (ii) disposed of in landfills, or (iii) incinerated (Puddephatt et al., 2022).

Waste production and management are regulated in many countries, albeit differently by jurisdiction. Nevertheless, waste management regulations globally were enacted to preserve human health, and minimize the potential contamination of aquatic resources (e.g., potable water, fisheries) and terrestrial ecosystems (by physical littering; landfill leachate). In addition, regulations govern the generation of gases during waste decomposition (including methane, carbon dioxide, carbon monoxide, hydrogen sulfides, etc.) (see Manning, 2005, for review).

History of municipal solid wastes (MSW)

In North America, historically, waste from residential, industrial, commercial, and institutional sources was collected and dumped into rivers, oceans, and unlined landfills, as land and water were viewed as an inexhaustible ‘sink’ for waste. However, as the unsustainability of such practices became very clear, and by the 1960s, citizens and lawmakers mobilized, and Congress passed the Solid Waste Disposal Act (SWDA) of 1965. This framework was the beginning for better waste control from all sources. Concomitantly, with the formation of the United States Protection Agency (USEPA) in 1970, federal lawmakers worked with the states to pass the Resource Conservation and Recovery Act (RCRA) of 1976. A transformation of the waste industry has been steadily making progress over the past half century, with RCRA national waste disposal goals that include (i) protecting human health and the environment from MSW hazards, (ii) conserving energy and natural resources, (iii) reducing amounts of waste generated, and iv) managing waste in an environmentally-sound manner.

With the formation of the European Environmental Agency (EEA) in 1994, the European Commission subsequently set forth incremental reductions in municipal solid waste through the Landfill Directive of 1999; among other waste reduction indicators, by 2015, they set out to reduce the landfill of biodegradable products by 65%; and in a 2018 referendum, they set the goal of limiting the share of landfilled MSW to 10% by 2035. For example, Ireland has since exceeded the recommended 50% of recycled waste at 53% and has surpassed the broader 70% of material recovery (recycling and compostables) at 84%.

The coupled reduction and effective management of wastes are essential to sustainable development, and thus, climate change mitigation. It has been postulated that if all wastes could be recycled and reused for new materials and products, forming a ‘circular economy’ as opposed to a linear consumption model (i.e., take–make–dispose), this would typify sustainable development.

In part, the rote belief that waste disposal is linear (take–make–dispose), has been propelled by past systematic and cultural resistance in Western societies to reuse large quantities of waste and prevent potential sources of contamination. Thus, disposal to landfill has been the dominant disposal procedure for residues in North America and parts of Europe (excluding Switzerland, Germany, Finland, Sweden, and Belgium).

However, it is noteworthy that since 1995, the amount of waste disposed to landfill has decreased (62% in 1995 to 42% by 2017) for OECD (Organization for Economic Co-operation and Development) countries and in the U.S., as the reduction of waste at source, before it enters the waste stream, has become a high priority for state and local governments in recent years. Yunis and Aliakbari (2021) suggested that since Canada was generating less waste over time, the country has “partially decoupled waste generation from economic growth.”

Concomitantly, alternative methods for solid waste disposal are being adopted in countries with the luxury of sustainable development (i.e. Western societies), including: (i) biogas recovery from landfills, (ii) reuse and repurposing of post-consumer plastic (9% plastic recycling rate in Canada, 8.7% in the United States), (iii) quarry rehabilitation with leachate liners and caps for landfilled solid waste, and (iv) the land application of sewage biosolids for soil and crop nutrient amendment (McCarthy and Loyo, 2015). Nevertheless, current Western waste generation and management processes, operating among individual citizens through to entire corporations, remains largely linear and consumptive (Bowen et al., 2020).

Once municipal solid wastes are landfilled, these materials are subjected to the natural processes of decomposition (both anaerobic and aerobic), and can affect root systems, influence pore water and gas compositions in the unsaturated and saturated zones, and can have potentially deleterious effects on the aquatic and terrestrial environment. However, the effects are not limited to when wastes are disposed at a landfill. In the genesis of landfills, where landfills are placed is important as it renders the land unusable for the duration of landfill activities. They also degrade soil fertility and plant growth through the removal of organic and plant matter, destroy the soil profile, and disturb the ecological community at a site and beyond. Indeed, the broader effects of landfill include reduced ecosystem services (e.g., carbon sequestration in soil and biomass, bioremediation of



Fig. 1 Earl Bales Ski and Snow Board Centre in urban Toronto, Ontario. Courtesy City of Toronto, Parks Forestry and Recreation (2019) Retrieved from <https://www.toronto.ca/data/parks/prd/facilities/complex/2766/index.html>.

compounds); modification of watershed drainage routes, and change in vegetation patterns; reduced biodiversity, habitat fragmentation; and in select instances, mechanical (e.g., wind-blown litter into terrestrial and aquatic environments), and chemical (e.g., long-range transport of landfill gases through the atmospheric compartment, which are deposited in distant terrestrial and aquatic compartments) dispersion of pollutants. Even when landfills are capped, prior soil needs to be remediated and or allochthonous soil and matter may be introduced to the site (which involves degrading another site).

Through the advent and use of leachate liners in both Europe and North America, landfill sites have limited their potentially deleterious environmental effects and are often restored to new uses through capping (preventing atmospheric deposition of compounds) and revegetation. For example, in Toronto, Canada, the *Earl Bales Landfill* has been capped and converted into a city park and downhill ski hill (see Fig. 1), while another Toronto landfill has been capped and contoured to form a golf course (*BraeBen Golf Course*; formerly Britannia Landfill).

To understand the complex consequences of MSW application on soil systems (i.e., compost, land-application, landfilled), it is necessary to understand the production of municipal solid wastes, their composition and subsequent degradation, and ecotoxicological processes. Moreover, this chapter reviews how the potential for environmental impact of MSW is evaluated and remediated.

MSW production, composition, and elimination

Current MSW production in the western world

As world populations undergo development and demographic transition, the World Bank estimates that the amount of global waste is likely to grow from 2.01 Bt year⁻¹, to 2.59 Bt year⁻¹ in 2030, and 3.40 Bt year⁻¹ by 2050 (The World Bank (TWB), 2018). While there are current improvements in per capita efficiency of waste generation (as waste minimization), globally there is increasing waste creation overall (TWB, 2018), as can be seen in the U.S.: see Fig. 2 below. For example, in 1990, the U.S. generated 188.96 million tonnes, while in 2018, that figure had risen to 265.2 million tonnes.

The focus on increasing efficiency suggests that there already exists waste infrastructure to manage MSW effectively. However, today, nearly 46% of the global population is without basic waste management systems and that building systematic waste management capacity is needed first (TWB, 2018). Thus, most of the following information focusses on North American and Western European MSW strategies and their ongoing efforts to reduce the impact of MSW pollution to water, land, and air. Table 1 gives the proportion of recyclables, compostables, incinerated, and land-applied to landfill solid wastes by country for 2000 and 2018. The selected countries follow Manning's review (Manning, 2005) (Japan, United Kingdom, United States) with a current addition of Canada.

Ongoing MSW strategies

In the Western world, a waste management hierarchy has been strongly encouraged for the past few decades (Fig. 3), with a preferred strategy for managing and minimizing the impacts of MSW. Source reduction and reuse, or the prevention of waste in the first place,

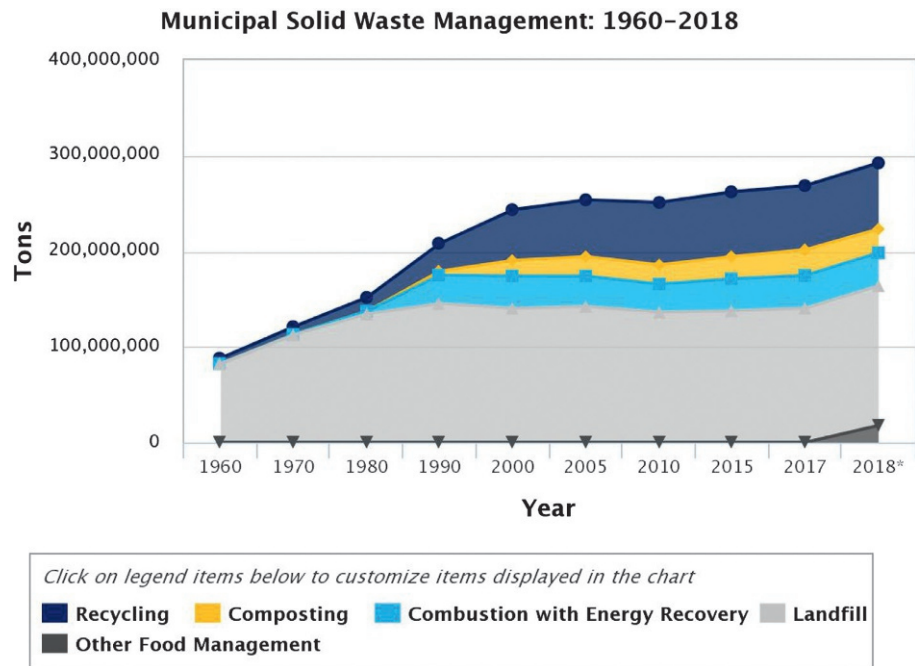


Fig. 2 Municipal solid waste management in the United States from 1960 to 2018. Note, units are provided in U.S tons (1 US ton = 0.907185 metric tonne). Courtesy United States Environmental Protection Agency (2018) *National Recycling Strategy: Part One of a Series on Building a Circular Economy for All*. Retrieved from: <https://www.epa.gov/system/files/documents/2021-11/final-national-recycling-strategy.pdf>.

Table 1 Amounts of MSW generated by Canada, Japan, U.K., U.S., broken down by end disposal. For 2000 (Top) and 2018 (Bottom). Data courtesy OECD (2022).

Countries	Recyclables	Compostables	Incinerated	Landfilled	Other
<i>tonnes, thousands year⁻¹ (2000)</i>					
Canada	5159	980	1158	22,010	0
Japan	7860	0	40,304	3084	3607
United Kingdom	2836	937	2456	27,563	162
United States	48,090	14,923	30,599	127,242	0
<i>tonnes, thousands year⁻¹ (2018)</i>					
Canada	6945	2873	1278	24,446	0
Japan	8354	181	33,721	439	0
United Kingdom	8386	5191	12,046	4463	832
United States	62,614	22,580	31,352	132,612	0



Fig. 3 Four-tiered waste management hierarchy. Courtesy United States Environmental Protection Agency (2018) *National Recycling Strategy: Part One of a Series on Building a Circular Economy for All*. Retrieved from: <https://www.epa.gov/system/files/documents/2021-11/final-national-recycling-strategy.pdf>.

is being implemented at the manufacturing, consumer, and institutional level. While recycling has been practiced, it has had varying degrees of success in waste reduction for the past half century. As of 2018, the USEPA reported 62.5 million tonnes of MSW was recycled, or less than 25% of the total MSW generated (265.2 million tonnes).

Recyclable and reclamation products

Recyclable municipal solid waste is that diverted from traditional solid waste disposal streams (e.g., land-application, landfill, incineration) for meaningful reuse and repurposing. Troschinetz and Mihelcic (2009) identified some of the barriers to adoption of MSW recycling, including: (i) waste collection and segregation, (ii) household education, (iii) technological and human abilities, and (iv) government policy. These concerns are ratified in the USEPA *National Recycling Strategy* which prioritizes (i) the market for recycled materials, (ii) collection and materials sorting infrastructure, and (iii) building education capacity and policies which reduce contamination in recyclable material streams, while promoting circularity.

Recyclable wastes are post-consumer products, such as plastics, that are transformed and remanufactured into new products with productive use. The United States reports that of the 62.6 million tonnes recycled, 66% is paper and paperboard, while plastic represents 4.4% of all redirected material (Fig. 4). In Canada, 3 million tonnes of plastic wastes are produced annually; however, only ~290 thousand tonnes are recycled (roughly 9%). What is necessary are systems to deal with the diversity of recyclable products. These include paper products (e.g., packaging, cardboard, newsprint, and office paper), containers (ferrous metal cans and vessels, aluminum, glass, plastics, and composites of the aforementioned sources), household hazardous wastes (e.g., paints, pesticides, and unregulated hazardous wastes), construction and demolition debris, batteries, computers and other electronics ('e-wastes'), used tires, oil and oil filters, and discarded appliances that are difficult-to-process and require niche reclamation technologies.

Recyclable products are collected from households and industrial and commercial sources in a variety of ways, including: (i) dedicated recyclable curbside collection (commingled or source-separated), (ii) collected with MSW and sorted at facilities, (iii) user-delivered to facilities through purchase (e.g., as buy-back) or donation programs, and (iv) through individual disposal agreements from industrial and commercial waste generators.

Once collected, wastes are recycled differentially. Primary mechanical recycling involves material that is shredded and melted into new pellets to make new products of similar quality. Secondary recovery involves processing reused material into products with lower qualities (i.e., lower grade plastics (i.e., downgrading)). Tertiary recovery involves extracting chemical constituents (i.e., chemical or feedstock recycling to depolymerize polymers). Quaternary recovery involves extracting energy from the product (i.e., energy-from-waste processes such as valorization). The multiple types of recycling above are ranked in decreasing order of societal value, given the reliance on raw materials (Hopewell et al., 2009). Newer biodegradable plastics are decomposed similarly to putrescible wastes, as the products are composed of large starch and cellulose concentrations, requiring decomposition from lignocellulosic bacteria and fungi in either aerobic and or anaerobic digestion systems).

Organic waste and compost strategies

Organic (putrescible) waste from food processing and consumption, in addition to collected leaf litter and garden waste, are considered. Compostable MSW from households and businesses is collected and managed in localized waste management units (e.g., municipalities in Canada, United States and Europe) and brought to industrial composting facilities. Selectively, compostables are remediated on-site by homeowners or local groups when municipality-led disposal is not available (e.g., backyard aerobic composters).

At industrial facilities, organic waste is decomposed either by aerobic (i.e., oxygen-rich putrefaction) or anaerobic mechanisms (i.e., anoxic or oxygen-poor decomposition). Typically, households compost aerobically. When MSW is delivered to composting

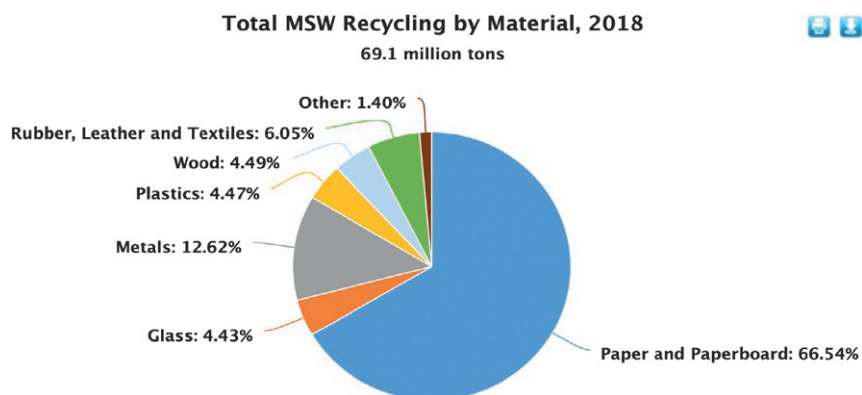


Fig. 4 Total MSW Recycling in 2018, by Material, Courtesy USEPA (2018). Note, units are provided in U.S tons (i.e., 69.1 million U.S. tons = 62.7 million tonnes). United States Environmental Protection Agency (2018) *National Recycling Strategy: Part One of a Series on Building a Circular Economy for All*. Retrieved from: <https://www.epa.gov/system/files/documents/2021-11/final-national-recycling-strategy.pdf>.

facilities, the substrate is sorted to remove non-compostable objects (e.g., plastics, non-putrescible wastes generally), and is prepared for primary composting through mechanical processes (i.e., grinding or shredding). Organic wastes may be primed for subsequent chemical and biological processes through adding leachate, water, and or inoculants such as bacteria (i.e., actinomycetes) and fungi (organisms specifically evolved to degrade more recalcitrant compounds and fixing nitrogen). While inoculants vary depending on geography and by the composition of organic matter, they predominantly include *Trichoderma* sp. (fungi producing cellulase and chitinase, degrading carbohydrates, and controlling pathogenicity, respectively) and *Pleurotus* sp. (i.e., spent mushroom substrate (SMS), or 'white rot,' decomposing lignin, cellulose, and hemicellulose materials from paper products and lawn wastes). Moreover, biological inoculation with microbiota is a conservative process, as wastes from digesters are used as bacterial and fungal community cultivars in subsequent composting. Collectively, mechanical, and chemical processes, in addition to microbial inoculation, improve the efficiency of decomposition.

During aerobic decomposition, the heat released from biological activity typically regulates pathogenic abundance to acceptable limits. Composts are typically placed into oxygen-rich environments (i.e., windrows) to be cured, where microorganisms (i.e., either atmospherically-deposited or from initial inoculation) slowly metabolize organic carbon into carbon dioxide and soil humus, and fix atmospheric nitrogen into soil nitrates to produce collectively a nutrient-rich organic matter. During curing, microorganisms such as thermophilic fungi (i.e., white-rot'), continue to break down more complex lignocellulosic structures. The compost is screened to reduce material size and remove relict physical contaminants (e.g., glass, metal) and stored until used. Nevertheless, there is potential for soil and water contamination through land application of compostables for soil and crop nutrient amendment. There are benefits to soil through increasing organic matter (i.e., source of carbon for plant development, improved water retention capacity), and the addition of soil macronutrients (i.e., N (nitrogen), P (phosphorus), K (potassium), and S (sulfur)). However, there is always potential, albeit mostly negligible, for the introduction of contaminants and physical litter that was not sufficiently screened and pre-processed (e.g., microplastics).

Under anoxic (i.e., oxygen-deprived) decomposition conditions, putrescible wastes are typically processed within high-solids systems (moisture content less than 80%) or low-solids (or wet systems) (moisture content above 80%). Furthermore, within each category, anaerobic digestion can occur within single or multiple stage processes. Within a multi-stage process, primary decomposition occurs in an acidic (pH 5–6) environment to allow for lignocellulose degradation (e.g., using *Cellulomonas* sp. and *Clostridium* sp. bacteria) and the creation of acetates and hydrogen (e.g., *Syntrophomonas* and *Syntrophobacter* genera). Subsequently, the feedstock is transferred to another vessel, and the pH is raised to neutral (~6.2–7) to optimize the transformation of methane gas by methanogens (e.g., heterotrophic methane bacteria convert acetic acid (CH_3COOH) into methane gas (CH_4)) (Verstraete et al., 2002). While methane gas production can occur in a single stage, the sole decomposition environment means that neither step (i.e., acidogenesis for acetate production, or methanogenesis) is optimized. The resulting methane gas may be used for heat and to power engines. The digestives (the settled products of anaerobic digestion) may be reused within inoculation of the digester (introducing decomposition bacteria from the end of one batch into the beginning of another).

Similar to aerobic digestives, anaerobic digestives may be stabilized and land-applied, either for soil and crop nutrient amendment, or brownfield remediation. With regard to potential soil contamination, anaerobic digestion typically further degrades more recalcitrant and persistent compounds, thus posing less ecotoxicological risk to soil ecosystems.

Energy recovery

In 2015, the European Commission presented a Circular Economy Package to the European Parliament to promote a shift from a linear economy, where resources have traditionally been extracted, used, and subsequently disposed of as waste. In this document, it was strongly recommended that the use of materials, and waste generation be minimized, which could be achieved through source minimization, reclamation and recycling of discarded products, composting of organic biomass, and energy recovery from MSW to become a renewable energy source (EUC, 2015). Energy can be recovered from waste by landfill gas recovery, combustion, pyrolyzation and gasification, and anaerobic digestion in processes that convert non-recyclable waste into useable heat, electricity, or fuel. The European Commission further suggested that this renewable energy could contribute 27% by 2030 to the final energy production budget. Energy recovery from MSW can provide a local source of renewable energy while the overall volume of waste being disposed of in landfills is reduced. In addition, the waste-to-energy (WtE) strategy, as a source of renewable energy, is expected to contribute to a reduction in the dependency on fossil fuel use, leading to a reduction in GHG (greenhouse gas) emissions. The number of WtE facilities for MSW in 2016 was 1618 plants globally, with 512 plants in Europe, 822 plants in Japan, and 88 in the United States (Scarlat et al., 2019).

In the U.S., combustion of MSW was initiated in the late 1880s and as a very primitive form of energy-from-waste, the heat generated from the incinerators was used to boil water that subsequently powered steam generators to generate electricity to be used domestically and in industrial processes. However, by the 1960s, the impact on air and water quality from these incinerators led to the Clean Air Act of 1970, with regulations to control emissions. While the combustion of MSW increased through the 1980s with cleaner emissions, and with 34 million tons of MSW diverted to WtE processes by 2017, landfill is still considered a more viable, albeit short-term, option in America as space limitations are not as important as in smaller countries such as Japan and some European nations. In addition, public opposition and the significant initial funding needed to build combustion plants for energy recovery, is slowing the implementation of WtE facilities further in North America.

The critical need for the increased development of WtE facilities globally cannot be overstated when dependence on traditional fossil fuels for energy is no longer sustainable. Thus, while prior sections described the essential need for recycling and composting

as the second tier of the MSW hierarchy (see Fig. 3), energy recovery from waste is an important third component that includes landfill gas utilization, incineration with energy recovery, pyrolysis and gasification, and anaerobic digestion with biogas production.

Landfill gas (LFG) utilization

In regions of the world where land is perceived as abundant (Canada, United States, etc.), MSW is often disposed in landfills. When biogenic components of that waste (paper fiber, food waste, wood, etc.) are buried, a complex series of reactions occurs and anaerobic microorganisms degrade the organic matter to produce landfill gas (LFG) composed primarily of methane (40–60%) and carbon dioxide, with trace amounts of volatile organic compounds (VOCs) providing less than 1% of the remaining gas components (Sullivan, 2010). After fossil fuel combustion and agriculture, landfills produce the third largest anthropogenic source of methane in the U.S. and as a potent GHG, this compound must be captured and changed to a less destructive form. While controlling its emission is crucial, the ability to use captured methane to generate energy is a sustainable strategy that is being implemented in many regions, especially in North America.

The capture, collection, and processing of methane from landfills for multiple purposes is shown in Fig. 5. Vertical and horizontal wells are installed into the MSW where 67% (open landfills) to 84% (closed landfills) of the gas is extracted and piped to a main collection system where it is flared (burned off) or treated for beneficial use, including industrial and institutional uses, arts and crafts, pipeline gas, and vehicle fuel. When landfill gas cannot be used for beneficial use, it is flared off to become relatively harmless carbon dioxide. However, for beneficial use, including pipeline gas, the methane must be treated to remove impurities, with primary treatment removing moisture and particulates and secondary treatment implemented to remove siloxanes and sulfur compounds that can damage equipment. The landfill gas can be sent in pipelines to boilers and kilns, to be used similarly to natural gas, or to provide heat and power to local greenhouses. In addition, artists' studios engaged in glass-blowing, metal-working, or pottery that are energy-intensive can use this beneficial energy source and gas turbines can produce electricity to sell commercially or to be used on-site.

Thermal waste-to-energy (WtE)

In regions of the world where land constraints dictate against significant landfill strategies, MSW is often diverted to WtE facilities, where cogeneration incineration, or incineration with energy recovery, is often the prevalent strategy, particularly in Japan (80% MSW was combusted in 2015) and the European Union (Kaza and Bhada-Tata, 2018). Advanced thermal treatment, including pyrolysis and gasification, and the anaerobic digestion of biodegradable waste, are also significant WtE processes (Scarlat et al., 2019). In a 2019 report, the International Environmental Technology Centre (IETC) of the United Nations Environmental

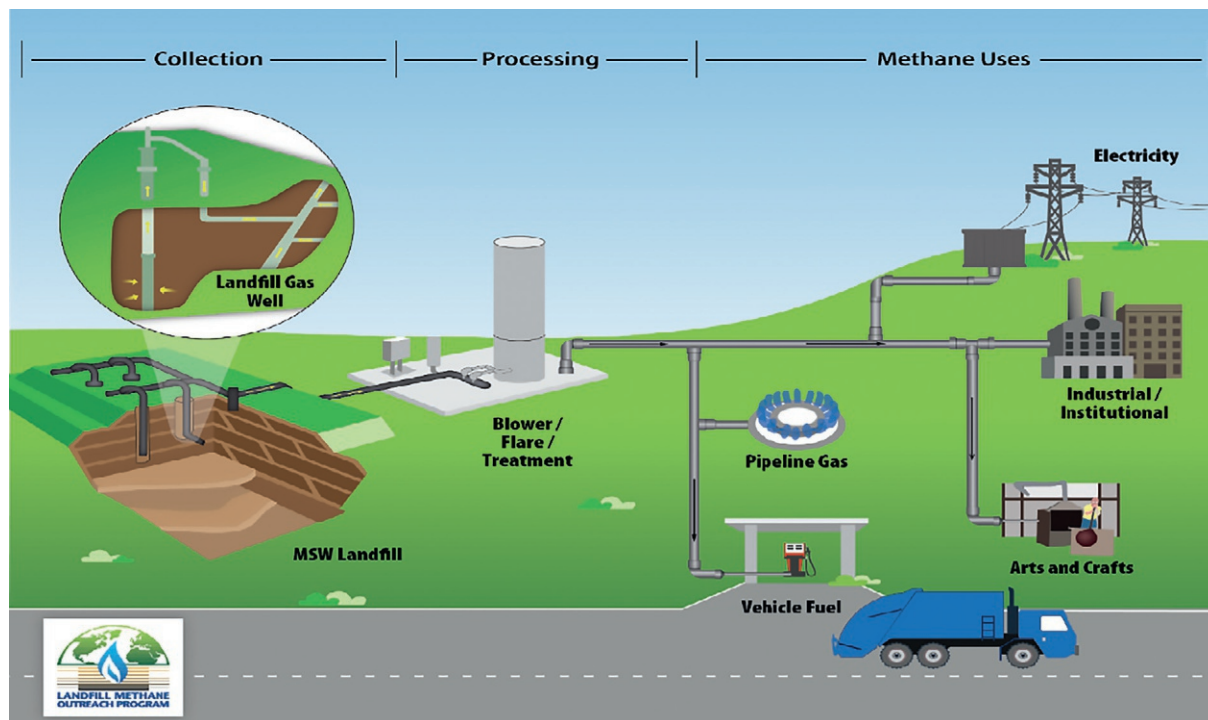


Fig. 5 Landfill gas collection, processing, and beneficial uses of methane. Courtesy United States Environmental Protection Agency (2021) *Basic Information about Landfill Gas*. Retrieved from: <https://www.epa.gov/lmop/basic-information-about-landfill-gas>.

Programme (UNEP) stated that over 1700 thermal WtE facilities are in operation worldwide, with 80% of the plants located in developed countries, and that 200 thermal WtE plants are under construction in developing countries (UNEP, 2019).

Combustion (incineration) with energy recovery

The MSW is the waste feedstock during the incineration process where the volume of waste can be reduced by 90% and where GHG emissions are reduced by diverting waste from landfills and open burning. Concerns about toxic emissions from modern incinerators, including dioxins, furans, and mercury, have been addressed with advanced air pollution control (APC) strategies that are implemented in countries with WtE facilities. The waste is combusted under controlled conditions and electricity and/or heat are subsequently generated, together with non-toxic bottom ash and toxic fly ash. The bottom ash, which usually is composed of silica, quartz, calcium, iron oxide, and aluminum oxide, can be recycled as construction material or used in the production of cement, while fly ash often contains a high content of 'heavy' metals and is disposed of in a hazardous waste landfill (UNEP, 2019). In the European Union, cogeneration data suggest that both electricity (500 kWh t^{-1}) and heat (1000 kWh t^{-1}) can be generated for industrial and district heating. Pollution from toxic emissions is reported to be low due to a combination of high temperature combustion that destroys most organic compounds, advanced air pollution control systems, and typically strict government emission regulations (Kaza and Bhada-Tata, 2018). Although incineration technology has advanced in the past decades, it is still a relatively expensive strategy and is used primarily in high-income countries where high technical capacity exists with strong environmental regulations. European Union countries with the largest capacities for MSW incineration are led by Germany (24 million tonnes) and France (16 million tonnes), followed by Sweden, Netherlands, Denmark, Norway, and Austria (Scarlat et al., 2019). In 2018, 30.8 million tonnes of MSW were combusted in the U.S. for energy recovery.

Pyrolysis and gasification

Advanced thermal treatment (ATT) technologies such as pyrolysis and gasification convert MSW not into heat and electricity (incineration energy recovery) but into synthetic gas (syngas) or liquid fuel such as tar or a solid material (char) that can be used as a soil amendment (Kaza and Bhada-Tata, 2018). The resulting syngas is largely composed of hydrogen (H_2) and carbon monoxide (CO) and is itself a fuel due to the flammability of H_2 and CO . It can be used as the hydrogen source in fuel cells, because high-temperature solid oxide fuel cells can use the gas mixtures, including carbon dioxide and methane, produced by gasification. Pyrolysis involves high temperature heating of MSW in oxygen-free environments; higher temperatures ($>760^\circ\text{C}$) produce more syngas and lower temperatures ($450\text{--}730^\circ\text{C}$) generate both liquid fuel and gas. Gasification also involves high temperatures in an environment containing some oxygen (O_2), but not as much O_2 as that required for incineration (Kaza and Bhada-Tata, 2018). Syngas can be used to produce electricity, or further processed to remove contaminants such as chloride and potassium to produce vehicle fuel.

While the use of MSW or biosolids in gasification processes to produce energy from waste is, in principle, sound, there are few instances globally of successful MSW facilities that use this strategy. Gasification is not new; heating organic matter or fossil fuels such as coal in a high temperature environment with a controlled amount of oxygen has been used to supply energy since the 1800s. However, a reliable feedstock that is of a consistent quality is difficult to obtain from MSW as it is a mixed composite of various products, and upstream separation of waste to secure a constant waste composition is necessary. Most pyrolysis and gasification plants initially pretreat incoming MSW feedstock by drying, sterilizing, and separation of recyclables. The remaining waste is then heated to produce the syngas, tar, and char, with the former being cleaned to remove particles and soluble substances to produce a 'clean gas' for electricity generation and/or heat (Kaza and Bhada-Tata, 2018).

The benefits of advanced thermal treatment technologies are substantial because the process diverts MSW from landfills, reduces methane emissions, and the char produced is an important soil fertility amendment. However, the technologies are relatively new; for instance, only a few full-scale operations exist in Japan and Europe where advanced technical knowledge and financial resources are available (Kaza and Bhada-Tata, 2018).

Non-thermal waste-to-energy (WtE) (anaerobic digestion)

Processes to convert food and outdoor green waste into fertilizers and biogas have been discussed under composting procedures that use oxygen, and anaerobic systems that exclude oxygen. More information on energy recovery during anaerobic digestion is provided here. In the absence of oxygen, anaerobic microorganisms degrade the organic components and produce a semi-solid fertilizer and biogas (Fig. 6). The process is in a closed vessel, or reactor, which allows a feedstock containing many materials such as manure, biosolids, and MSW biomass to co-digest. This anaerobic digestion process typically reduces the waste feedstock by 50% while conserving nutrients and killing up to 95% of pathogenic microbes. The biogas can be used to generate electricity or be further refined to supply natural gas (Kaza and Bhada-Tata, 2018).

The biogas produced is primarily methane (50–75%), a natural gas, and can be used to provide heat, generate electricity, and power cooling systems. In addition, biogas can be further purified to remove carbon dioxide (CO_2), hydrogen sulfide (H_2S), and water to generate renewable natural gas (RNG). After appropriate treatment (including drying), the digestate can be used as a nutrient-rich fertilizer, animal bedding, or as a foundation for bioplastics, among other uses. Anaerobic digestion of MSW is used extensively in Europe and is becoming economically feasible as source separation practices improve. With increasingly high landfill disposal costs, and favorable energy prices, non-thermal Waste-to-Energy (WtE) practices are expanding globally (Kaza and Bhada-Tata, 2018).

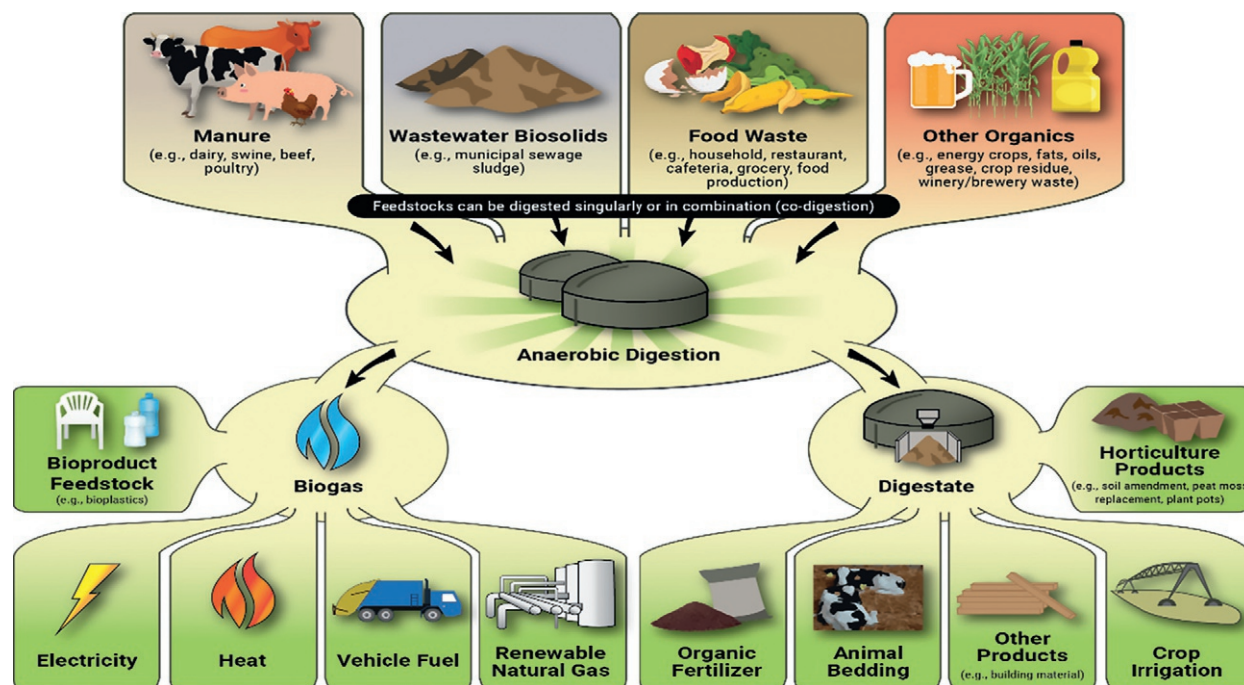


Fig. 6 Feedstock from agriculture, wastewater treatment plants, MSW, other industries undergo anaerobic digestion to produce biogas and digestate that can be used for fertilizers and other beneficial uses. Courtesy United States Environmental Protection Agency (2019) *Methane Emissions in the United States: Sources, Solutions & Opportunities for Reductions*. https://www.epa.gov/sites/default/files/2019-06/documents/methane_emissions_overview_may2019.pdf.

Treatment and disposal: Landfill

The MSW that is unable to be composted or recycled, either because of its composition or the lack of capacity to do so (i.e., insufficient waste management infrastructure), or not utilized in WtE facilities requires final disposal. In North America, these MSWs are disposed of to landfill or are incinerated. In the United States, the USEPA estimates 67.9% of MSW is either sent to landfill (56.1%) or incinerated (11.8%). These solid wastes are collected similarly to recycled wastes through curbside household collection and individual generator disposal agreements.

When solid wastes are landfilled, they are presumed to be localized and geographically constrained (Sivakumar-Babu et al., 2017). Landfill typically involves daily compaction of the surface with new wastes until capping (i.e., sealing a completed landfill), in turn, anaerobic conditions are established for waste degradation, similar to composting. However, there is potential for widespread environmental impact if leachate (i.e., contaminated precipitation through contact with wastes) is not contained and or unprocessed and enters watercourses, surface water, permeates soil (through infiltration and percolation), and subsequently groundwater. Leachates characteristically contain a large amount of organic matter (OM), have low pH (acidic with carbonic acid and basic with ammonium), chemical oxygen demand (COD), ammonia (NH_4), nitrogen gas (N_2), and metals (e.g., mercury (Hg)), in addition to other emerging contaminants of concern (ESOCs). Prior to discharge, OM is removed from leachate, as determined by the degree of chemical and biological oxygen demand (COD, BOD respectively) and ammonium content, to avoid potential toxicity to aquatic and terrestrial environments.

The composition of landfill leachate is typically characterized by age; for instance, younger landfill leachate has products of acidogenic anaerobic decomposition (where acidifying bacteria convert hydrolysis products and soluble chemicals into short-chain acids, alcohols, hydrogen, aldehydes, and carbon dioxide). Moreover, younger leachate may contain volatile fatty acids (VFAs) because of large amounts of biodegradable matter, and the high moisture content in landfilled waste (Welander et al., 1997). By comparison, mature landfill materials are further along the decomposition process where methanogenic microorganisms will transform acetate ($\text{C}_2\text{H}_3\text{O}_2$) and hydrogen (H_2) into CH_4 (methane) and carbon dioxide (CO_2), which are then emitted to the atmosphere. The remaining leachate is dominated by non-biodegradable humic substances. As gas creation from leachate is a potential source of atmospheric contamination, engineered solutions capture end-member gases (e.g., methane, carbon dioxide) for energy production. Apart from leachate-derived gas production, the other landfill gases are reported in low proportions and vary because they depend upon what wastes are discarded (atmospheric nitrogen (N_2), hydrogen (H_2), carbon monoxide (CO), and hydrogen sulfide (H_2S) (Manning, 2005). While gas production directly affects atmospheric compartments, it results in the indirect effects of the long-range transport of pollutants, which are subsequently precipitated onto terrestrial (including soil) and aquatic environments.

Newer landfills typically collect and treat leachate, either on-site or at wastewater treatment facilities, to mitigate potential deleterious environmental effects. Treated singly and concomitantly, leachate is processed similarly to sewage treatment through

(i) aerobic biological systems, such as activated sludge or aerated lagoons, (ii) anaerobic biological systems, such as bioreactors or anoxic lagoons, (iii) physicochemical treatment, where environmental properties are modified for optimal degradation (e.g., pH adjustments, deoxygenation, chemical precipitation, reduction and oxidation); and or (iv) coagulation using binders (e.g., lime, ferric chloride, electrical systems).

Treatment and disposal: Incineration

When incinerated, the volume of waste is significantly reduced. Historically, air pollution from incineration was of potential concern. Modern incinerators typically include scrubbers and other strategies to prevent atmospheric contamination. For instance, scrubbers exist for many persistent organic pollutants (POPs) such as carcinogenic dioxins (i.e., polychlorinated dibenzodioxins) and furans, in addition to metals (e.g., Hg, Fe, Pb) (Zhang et al., 2020).

Incineration typically occurs in waste-to-energy facilities (WtE) and wastewater sludge incinerators (e.g., incineration of sludge); however, the exact energy potential depends on the composition of the waste or dried sludge (e.g., 7,594,390 J kg⁻¹ for food waste to 101,874,148 J kg⁻¹ for polypropylene products (i.e., plastics) (Tillman, 2012). As a result of combustion, predominant compounds include silicon (Si), aluminum (Al), iron (Fe), calcium (Ca), magnesium (Mg), potassium (K), sodium (Na), and chlorine (Cl), in addition to the metals cadmium (Cd), chromium (Cr), copper (Cu), mercury (Hg), nickel (Ni), lead (Pb), and zinc (Zn). Furthermore, common volatile organic compounds (VOCs) of methylene (CH₂), CO, CO₂, CH₄, ethane (C₂H₆), and amino radicals (i.e., amide ion, NH₂), in addition to water vapor (H₂O(g)) are common (Tillman, 2012). Similarly, the exact gases released are contingent on the source material, necessitating understanding of the chemistry of different combustibles (i.e. source compounds, interactions, end members) (Tillman, 2012). Finally, the ash from incineration may be recovered for iron (i.e. as 'ferrous recovery') and other metals (as a form of tertiary recovery) or sent for disposal.

Incineration gases can be cleaned by (i) separation processes (i.e., washing, leaching, electrochemical processes, thermal treatment, flocculation and filtration, ion exchange, crystallization, and evaporation to transform gas products to non-environmentally-deleterious compounds before they are re-emitted into the atmosphere), (ii) solidification and stabilization (i.e. stabilization with and without additives, binders, or macro-encapsulation), and (iii) thermal methods (vitrification, sintering, and fusion) (see Van der Sloot et al., 2001, for review). For example, limestone (calcium carbonate and dolomite) is popularly used for the removal of hydrochloric acid (HCl) and sulfur dioxide (SO₂) as a form of stabilization with additives.

Biosolids (i.e., processed sewage sludge)

Biosolids (or sewage sludge) are a major by-product of primary and secondary treatment of residential and industrial effluent in wastewater treatment plants (Puddephatt et al., 2022). Biosolids are created in similar ways to the processing of landfill leachate, and disposed of similarly to other solid wastes (including incineration and landfilling). The treated product is also used similarly to decomposed putrescible matter in their land-application for soil and crop nutrient amendment and remediating brownfields (similar to compostables). Biosolids are characteristically high in organic matter (OM) and soil macronutrients such as phosphorus (P), nitrogen (N), potassium (K), and sulfur (S), and are applied as liquids or solid composites to land. In 2012, Canada produced more than 660,000 metric tons (dry weight, dw) of biosolids, with almost half of them being produced in southern Ontario. In the United States, 51% of biosolids were disposed through land-application, 22% were landfilled, 16% incinerated, and 11% were subject to other management practices (e.g., brownfield remediation, sold to the public for lawn and garden soil amendments, etc.).

As a by-product of wastewater treatment, there are often engineered strategies to remove organic contaminants (e.g., flame retardants, personal pharmaceuticals and care products (PPCPs), potentially toxic metals (e.g. mercury, lead), and pathogens (e.g. enteroviruses, *Salmonella* sp., *Listeria* sp., and protists such as *Giardia*) from sewage sludge (Flemming et al., 2017). See McCarthy and Loyo (2015), for a review of the production and treatments of biosolids. Many of the compounds found in trace amounts, and predominantly studied in biosolids, are emerging substances of concern (ESOCs). While ESOCs are ambiguously defined, there is consensus on the common characteristics including i) that they have been detected in at least one environmental compartment (i.e., terrestrial, aquatic, atmospheric) and are present in relatively small (trace) amounts and ii) that they are generally not being regulated, or at least not for a recent concern, such as possible endocrine disruptor effects of poly-chlorinated bisphenols (PCBs), which were banned because of their persistent and toxic nature (San Francisco Estuary Institute (SFEI), 2013).

At wastewater treatment plants, incoming wastewater from sanitary networks is treated through physical, mechanical, chemical, and biological means in two stages of primary and secondary treatment. Primary treatment is a physical process where heavier solids (e.g., tree branches, rocks, and feces) are removed; secondary treatment is a biological process where organic pollutants in effluent are oxidized. Moreover, during secondary treatment, the addition of chemicals may be used to precipitate and coagulate macronutrients from the secondary wastewater (aluminum sulfate (Al₂(SO₄)₃) or calcium hydroxide (Ca(OH)₂)). Once settled, the solids (or sludge) in both primary and secondary treatment processes are sent to a digester where organic matter, ammonia, and amino species are converted to carbon dioxide, water, and nitrate.

Biosolids are also produced using anaerobic digestion and dewatering, where anoxic microorganisms continue to break down the organic material present in the original sludge and create a product much less offensive in terms of odor and organic contaminants. In the anaerobic digestion process, the four microorganism groups (hydrolyzers, acidogens, acetogens, and subsequently, methanogens) convert lignocellulosic compounds and suspended solids to organic acids, then into acetate and hydrogen (H₂), which are subsequently evolved into the more stable compounds of carbon dioxide (CO₂), ammonia (NH₃), and methane (CH₄). In both aerobic and anaerobic digestion, the volume of sludge, and abundance of pathogens (approximately 95% removal) is reduced from the source sludge (Toronto Water, 2019). At a subsequent stage, dewatering reduces water content and pathogens in

the resulting biosolids. Dewatering is a mechanical process where the addition of chemicals releases water bound to organic matter, which is then removed by centrifugation, pressing, or belts. The newly formed biosolids, with a reduced pathogen and volatile organic compounds (VOCs) load, are then applied to land, incinerated, or disposed of in landfills.

The MSW challenges

Globally, an estimated 2 billion tonnes of MSW were generated in 2016 and projected to increase to 3.4 billion tonnes by 2050, if current practices of MSW production continue (Kaza et al., 2018). Approximately 37% of waste is landfilled, 33% is openly dumped, 19% is recycled and composted, and 11% undergoes modern incineration with energy recovery. Adequate waste disposal or treatment is almost exclusively the domain of high- and upper-middle-income countries while lower-income countries rely on open-dumping (93% compared to only 2% in high-income countries) (Kaza et al., 2018). In addition, high-income countries, despite education and government incentives, continue to generate large amounts of MSW. In the United States, the generation of MSW has increased from 79.92 million tonnes in 1960, to 264.9 million tonnes in 2018, translating into a generation rate of 1.2 kgs/person/day in 1960, to 2.3 kgs/person/day, respectively. This increase has occurred despite comprehensive efforts on source reduction and use, recycling and composting, and energy recovery from combustion (zero in 1960; 12% in 2018). In Europe, while increasingly sophisticated engineered technologies have seen the rise of WtE facilities, Scarlat et al. (2019) report that MSW as an energy source is underexploited in most European countries, with only a very small percentage of the existing waste being converted to energy. Lastly, low-income countries are expected to increase their waste generation by more than three times by 2050, with open-dumping speculated to remain the major strategy of disposal (Kaza et al., 2018).

Effective waste management and disposal practices will, and has been, affected by natural and anthropogenic disasters (i.e., hurricanes, earthquakes, floods, tornadoes, forest fires, pandemics and war). While the immediate effect of such calamities will be obvious, infrastructure to support MSW strategies will be influenced negatively, including stewardship of long-term landfills and the continuation of WtE facilities. Municipal policies focused on MSW source reduction, reuse and recycle plans, and composting options will also, out of necessity for preserving human life, be reduced. In response, the United States Environmental Protection Agency continues to develop plans for MSW preparedness in the face of disastrous weather events, including the need to deal with huge amounts of storm debris and flood residues properly. Other catastrophic events such as chemical accidents and domestic security attacks similarly pose additional challenges for municipalities globally.

Concluding remarks

The World Business Council for Sustainable Development (WBCSD), in their *Vision 2050 – The New Agenda for Business Report*, suggest that the projected 9 billion people by 2050 must learn to live within the limits of the resources of Planet Earth. The report further proposes that material demand, consumption, and production become transformed strategies to address the truth – that non-renewable resources are truly limited. Thus, practices that mandate closed-loop recycling make the concept of the obsolescence of waste a normalized business practice (WBCSD, 2021).

In support, the 2015 European Commission's Report to the European Parliament outlined the necessity for a circular economy in the waste management field, similarly stipulating a multinational plan for waste and resource management. Strategies include source minimization, reclamation and recycling of discarded products to be reengineered for multiple purposes including raw materials for manufacturing, composting of organic matter, and energy recovery to become a renewable energy source, have all been described. The reduction in GHG emissions and a decrease in energy and water use are mandatory in this circular economy plan (UNEP, 2019). Lastly, in the United States, new attitudes and technological advances in life-cycle product development, use, and retirement are ushering in a new era of 'waste avoidance,' rather than current strategies of 'waste management,' through improved design and reuse and recycling mandates.

The effects of MSW disposal on their environment are broadly similar to the solutions we need to reduce and close the loop on otherwise unsustainable and unprincipled resource consumption: these are to consider the broader socio-political and environmental system on which they depend. Thus, solutions need be transboundary and they need to invoke principles of 'strong sustainability,' which principally involves changing Western consumption patterns through education and incentivization, ultimately bringing about large-scale behavior changes. Finally, these solutions need to support research and innovation to improve the efficiency of existing recycling and composting processes.

References

- Bowan PA, Kayaga SM, Cotton AP, and Fisher J (2020) Municipal solid waste management performance. *Journal of Studies in Social Sciences* 19.
- European Commission (EUC) (1999) Council Directive 1999/31/EC of 26 April 1999 on the landfill of waste. Retrieved from: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31999L0031>.
- European Commission (EUC) (2015) *Towards a circular economy: a zero waste programme for Europe*. COM (2014). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, Brussels.
- Flemming CA, Simhon A, and Odumeru JA (2017) Pathogen characterization of fresh and stored mesophilic anaerobically digested biosolids. *Water Environment Research* 89(11): 2031–2042.

- Hopewell J, Dvorak R, and Kosior E (2009) Plastics recycling: Challenges and opportunities. *Philosophical Transactions of the Royal Society, B: Biological Sciences* 364(1526): 2115–2126.
- Kaza S and Bhada-Tata P (2018) *Decision Maker's Guides for Solid Waste Management Technologies*. Urban Development Series, World Bank Group. Retrieved from: <https://openknowledge.worldbank.org/handle/10986/31694>.
- Kaza S, Yao L, Bhada-Tata P, and van Woerden F (2018) *What a Waste 2.0. A Global Snapshot of Solid Waste Management to 2050*. Urban Development Series, World Bank Group. Retrieved from: <https://openknowledge.worldbank.org/handle/10986/30317>.
- Manning DAC (2005) Waste Disposal on Land. Municipal. In: Hillel (ed.) *Encyclopedia of Soils in the Environment*. Elsevier.
- McCarthy LH and Loyo J (2015) Risks associated with municipal biosolids application to agricultural lands in Canada – Literature review. *Canadian Water Network*. Retrieved from: <https://cwn-rce.ca/wp-content/uploads/2015/08/McCarthy-Risks-Biosolids-2015.pdf>.
- Organization for Economic Cooperation and Development (OECD) (2022) *Municipal waste (indicator)*. <https://doi.org/10.1787/89d5679a-en>.
- Puddephatt KJ, McCarthy LH, and Serre BM (2022) Assessing the Potential Chronic, Sublethal and Lethal Ecotoxicity of Land-Applied Biosolids on *Folsomia candida* and *Lumbricus terrestris*. *Ecotoxicology*. <https://doi.org/10.21203/rs.3.rs-1350257/v1>.
- San Francisco Estuary Institute (SFEI) (2013) Emerging Contaminants Workgroup. Retrieved from <http://www.sfei.org/rmp/ecwg>.
- Scarlat N, Fahl F, and Dallemard JF (2019) Status and opportunities for energy recovery from municipal solid waste in Europe. *Waste and Biomass Valorization* 10: 2425–2444. <https://doi.org/10.1007/s 12649-018-0297-7>.
- Sivakumar-Babu GL, Lakshmikanthan P, and Santhosh LG (2017) Assessment of landfill sustainability. In: *Sustainability Issues in Civil Engineering*, pp. 257–269. Singapore: Springer.
- Sullivan P (2010) The Importance of Landfill Gas Capture and Utilization in the U.S. Retrieved from: https://www.scsengineers.com/wp-content/uploads/2015/03/Sullivan_Importance_of_LFG_Capture_and_Utilization_in_the_US.pdf.
- Tchobanoglous G and Kreith F (2002) *Handbook of Solid Waste Management*, 2nd edn. New York: McGraw-Hill.
- The World Bank (TWB) (2018) What a Waste: An Updated Look into the Future of Solid Waste Management. Retrieved from: <https://www.worldbank.org/en/news/immersive-story/2018/09/20/what-a-waste-an-updated-look-into-the-future-of-solid-waste-management>.
- Tillman DA (2012) *Incineration of Municipal and Hazardous Solid Wastes*. Elsevier.
- Toronto Water (2019) City of Toronto Wastewater Treatment Plant Annual Reports - Ashbridges Bay Wastewater Treatment Plant 2019 Annual Report. City of Toronto. Retrieved from: <https://www.toronto.ca/wp-content/uploads/2019/05/8f0f-2018-TAB-Annual-Report-FINAL-ecopy.pdf>.
- Troschinetz AM and Mihelcic JR (2009) Sustainable recycling of municipal solid waste in developing countries. *Waste Management* 29(2): 915–923.
- United Nations Environmental Programme (UNEP) (2019) International Environmental Technology Centre (IETC). Waste to Energy. Considerations for Informed Decision-Making. Retrieved from: <https://www.unep.org/ietc/resources/publication/waste-energy-considerations-informed-decision-making>.
- Van der Sloot HA, Kosson DS, and Hjelmar O (2001) Characteristics, treatment and utilization of residues from municipal waste incineration. *Waste Management* 21(8): 753–765.
- Verstraete W, Doulami F, Volcke E, Tavernier M, Nollet H, and Roels J (2002) The importance of anaerobic digestion for global environmental development. *Doboku Gakkai Ronbunshu* 2002(706): 97–102.
- Welander U, Henrysson T, and Welander T (1997) Nitrification of landfill leachate using suspended-carrier biofilm technology. *Water Research* 31(9): 2351–2355.
- World Business Council for Sustainable Development (WBCSD) (2021) Time to Transform. Retrieved from: <https://www.wbcd.org/Overview/About-us/Vision-2050-Time-to-Transform/Resources/Time-to-Transform>.
- Yunis J and Aliakbari E (2021) *Generation and Management of Municipal Solid Waste: How's Canada Doing?* Fraser Institute.
- Zhang J, Zhang S, and Liu B (2020) Degradation technologies and mechanisms of dioxins in municipal solid waste incineration fly ash: A review. *Journal of Cleaner Production* 250: 119507.

Relevant websites

- https://c2f5e172-e73c-41cc-bca8-22c93ecbc51c.filesusr.com/ugd/c3fa71_d835cf33c9b040c0b72e93799efc1298.pdf—Please see a link to the webpage with all references used to conduct this review.
- https://www.ec.gc.ca/gdd-mw/3E8CF6C7-F214-4BA2-A1A3-163978EE9D6E/13-047-ID-458-PDF_accessible_ANG_R2-reduced%20size.pdf—Technical Document on Municipal Solid Waste Organics Processing.
- https://publications.gc.ca/collections/collection_2020/eccc/en14/En14-405-2020-eng.pdf—National Waste Characterization Report: The Composition of Canadian Residual Municipal Solid Waste
- <https://www.canada.ca/en/environment-climate-change/services/managing-reducing-waste/reduce-plastic-waste.html>—Plastic waste and pollution reduction.
- <https://www.fao.org/3/y5104e/y5104e05.html>—Composting process and techniques.
- https://www.epa.gov/sites/default/files/2019-08/documents/multi_stage_anaerobic_digestion_fact_sheet_p10053f0.pdf—Biosolids Technology Fact Sheet: Multi-Stage Anaerobic Digestion.
- <https://www.epa.gov/system/files/documents/2021-11/final-national-recycling-strategy.pdf>—National Recycling Strategy: Part One of a Series on Building a Circular Economy for All.
- <https://www.epa.gov/smm/energy-recovery-combustion-municipal-solid-waste-msw#>—Energy Recovery from the Combustion of Municipal Solid Waste (MSW).
- <https://www.epa.gov/mop/basic-information-about-landfill-gas>—Basic Information about Landfill Gas.
- <https://www.epa.gov/agstar/how-does-anaerobic-digestion-work>—How does Anaerobic Digestion Work?

Remediation of metals and organic contaminants in soil

Peter Sanderson^{a,b}, Md Mezbaul Bahar^{a,b}, Bhabananda Biswas^{a,b}, and Ravi Naidu^{a,b}, ^aCRC for Contamination Assessment and Remediation of the Environment (crc CARE), ATC Building, University of Newcastle, Callaghan, NSW, Australia; ^bGlobal Centre for Environmental Remediation (GCER), University of Newcastle, University Drive, Callaghan, NSW, Australia

© 2023 Elsevier Ltd. All rights reserved.

This is an update of E. Lombi, R.E. Hamon, Remediation of Polluted Soils, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 379-385, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00087-4>.

Introduction	333
Remediation approaches	334
Capping	334
Engineered containment cells	335
Soil stabilization/solidification	335
Chemical stabilization	336
Bioremediation	336
Phytoremediation	337
Soil washing	338
Soil flushing	339
Chemical reduction	339
In situ oxidation	339
Electrokinetic remediation	340
Soil vapor extraction	341
Thermal treatment	341
Conclusions	342
References	342

Abstract

Various remediation approaches have been demonstrated for the management of metal and organic contaminants. Remediation approaches may be broadly categorized as “Isolation,” “Immobilization” “Degradation/toxicity reduction.” The feasibility (effectiveness, cost) of any given approach depends on the nature of contamination and site-specific conditions and constraints. The chosen remediation technology must render the site suitable for the future proposed land use. Increasingly, consideration of sustainability has led to emphasis on risk-based approaches which seek to reduce risks to human health and ecosystems to ensure contaminated land is made suitable for a proposed future land use.

Key points

- A range of technologies is available for the remediation of metals and organic contaminants
- Remediation techniques may be categorized as “Isolation,” “Immobilization” “Degradation/toxicity reduction”
- Applicability of remediation approaches depends on the contaminants of concern and site-specific considerations
- Some technologies are applied in situ and some require ex situ treatment of soil

Introduction

Widespread contamination of land has resulted from human activities such as mining, industrial and energy production, agriculture, transport, and waste generation and disposal. This contamination is often attributed to the emission and exposure of “heavy” metals, metalloids and a wide spectrum of organic and other inorganic compounds. Among metals and metalloids, lead (Pb), arsenic (As), chromium (Cr), cadmium (Cd), mercury (Hg), copper (Cu), nickel (Ni), zinc (Zn) and antimony (Sb) are priority elements of potential concern for ecological and human health. Similarly, in the case of organic contaminants, petroleum hydrocarbons, polycyclic aromatic hydrocarbons (PAHs), trichloroethylene (TCE), polychlorinated biphenyl (PCBs), dioxins, pentachlorophenol (PCP), perfluoroalkyl and polyfluoroalkyl substances (PFAS), and dichlorodiphenyltrichloroethane (DDT) are commonly found soil contaminants.

Remediation of contaminated soils is a critical part of the restoration of degraded lands to safe and beneficial reuse. This is becoming increasingly important as urban populations are expanding into areas that were previously used for industries and agriculture.

It is recognized that a sustainable risk-based approach is needed to tackle the scale of this issue. A risk-based approach seeks to reduce risks to human health and ecosystems associated with contaminated land so that they are suitable for a proposed future land use.

Recent decades have seen extensive research and development on remediation approaches and technologies. Many conventional remediation technologies, such as “pump and treat,” capping and containment, solidification and stabilization, soil washing, thermal treatment, in situ chemical oxidation, and in situ biodegradation are being applied commercially, while improvements to existing technologies and development of new approaches continue. An emphasis on sustainability is increasingly placed on remediation technologies considering the site and contaminants characteristics. Approaches, such as excavation and disposal of contaminated soil to landfill are becoming more undesirable.

Broadly, remediation technologies fall into the categories “containment and isolation,” “immobilization,” “separation and degradation” and “detoxification.”

Remediation approaches

Table 1 provides an overview of the main remediation techniques and their application.

Capping

Capping is the process of covering contaminated soil or landfill waste to create a barrier to the hazardous materials. It is a non-removal soil remediation approach that does not destroy or remove the contaminants, but keeps them in place. Caps isolate the contaminants to prevent them from spreading further, though there remains risk of exposure and mobilization where the integrity of the cap and liners is not preserved and in the future beyond the lifetime of the containment layers. Capping techniques can be utilized for a variety of reasons, including restricting access of contaminant to people and wildlife, reducing water percolation from site of concern, preventing the escape of gases containing harmful volatile chemicals, and stopping hazardous material from being blown by the wind.

Selection of an appropriate capping system depends on the contaminant’s characteristics, the size of the site, rainfall, and future use. A cap can be made simply by placing a single layer of material over mildly contaminated soil. However, highly contaminated soil or hazardous waste landfill may require several layers of different materials, including a vegetative layer, drainage layer, geomembrane, and clay layer (Fig. 1). The clay layer acts as the barrier because of its low permeability, which is sometimes coupled with a geo-synthetic membrane to make it more effective (US EPA, 2012a).

Capping of contaminated areas is one of the most simple, rapid, and cost-effective solutions for reducing risks from soil contamination. However, it is not a permanent remediation solution because the contamination remains, and caps and geomembranes can deteriorate over time. In addition, the seasonal fluctuation of groundwater at the capping site, as well as surrounding hydrogeological features must be considered. These may influence the cap stability. Furthermore, the risk of slippage must be assessed if a cap is installed on sloping land (Liu et al., 2018).

Table 1 Overview of soil remediation approaches.

Category	Techniques	Processes	Contaminants susceptible to the technique	Advantages	Limitations
Isolation	Capping, Containment Cells and Sub-surface barriers	Physical	Metals and organics	Simple, rapid and cost-effective	Contamination remains on site and needs to be monitored
Immobilization	Solidification and Chemical Stabilization	Physical and Chemical	Metal(loid)s and organics	Low cost; can be both in situ and ex situ	Requires binding materials; Long-term stability requires consideration of other factors (e.g., dosage and regeneration of binding material)
Degradation/toxicity reduction	Bioremediation, Phytoremediation, Soil Washing, Chemical Reduction/Oxidation, Electrokinetic, Thermal Treatment and Soil Vapor Extraction	Biological, Chemical and Physical	Metal(loid)s and organics	Applicability and feasibility depend upon contaminant and site-specific conditions	

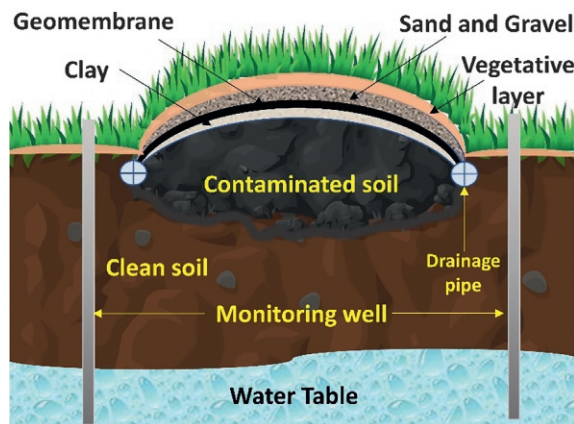


Fig. 1 Capping of contaminated soil with several layers. Scheme adapted after US-EPA (2012a) A Citizen's Guide to Capping, EPA 542-F-12-004.

Engineered containment cells

Engineered containment cells (ECCs) are designed to contain wastes and contaminated soil completely on site. They provide a more fully engineered approach than capping alone by limiting infiltration and providing a full contaminant barrier between the contaminated substances and the clean soil at the site (Fig. 2). This method of remediation depends on a feasibility assessment whether the contaminants identified require temporary or permanent containment rather than their destruction or removal (CRC CARE, 2019a).

An ECC needs a layer of poorly permeable or impermeable material as a barrier on the sides and bottom surface of the containment cell. Clay, such as bentonite or a mixture of clay and soil is often used as the barrier layer or "geo membrane." The ECCs also use an engineered cap for limiting infiltration and exposure of contaminants within the cell. An engineered cap may include a layer of vegetation to stabilize the capping layer, along with providing phytoremediation (US-EPA, 2012b). An ECC differs from other engineered barriers such as vertical engineered barriers (VEBs) which are installed vertically in the subsurface in trenches for containing the contaminants at their origin (in situ). On the other hand, ECCs are typically used to consolidate and contain excavated material from sites.

This remediation method is a choice of technology where the clean-up of contaminants would be expensive and technically difficult to execute. Landfill or active metal smelter sites may be examples of such target land sites.

Soil stabilization/solidification

Stabilization/solidification (S/S) refers to as a remediation technology whereby the mobility of contaminants is contained or restricted using binding materials. These binding materials can be cementitious and or might include additional stabilizing materials such as biochar, modified clays, iron nanoparticles, slag.

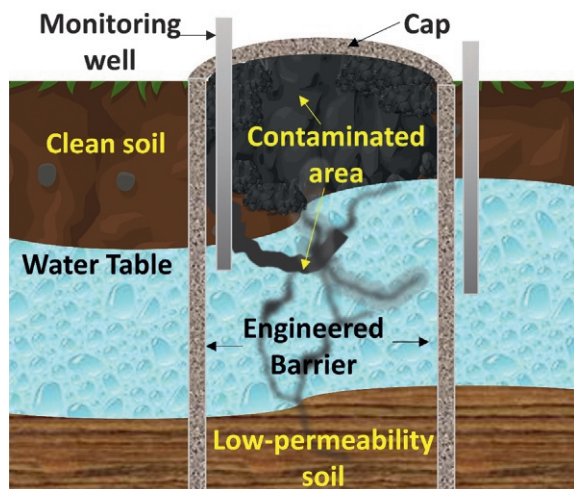


Fig. 2 Engineered containment barrier of contaminated soil. Scheme adapted after US-EPA (2012b) A Citizen's Guide to Engineered Vertical Barriers. In: Agency, E. P. (ed.). EPA.

Stabilization is the process where a chemical reaction is expected between the modifier and contaminant(s) of concern, aiming to cease or minimize the leachability and volatilization of those contaminants. On the other hand, solidification is defined as the process where the remediating material entraps or encapsulates the contaminants into its physical matrix. However, these two approaches are often combined in the process, and are thus historically known as S/S in the remediation technologies (Bates and Hills, 2015).

The stabilization/solidification approach can be applied both in situ and ex situ, and for remediating inorganic and organic contaminants in soil. The in situ approach is applied where the affected soil is not removed or excavated from its original place. In the case of ex situ, the soil is removed and taken to a dedicated stockpile area or facilities where the binding material is mixed with the pile.

Stabilization/solidification is suitable mostly for remediating soils that are contaminated with metallic compounds, high molecular weight polycyclic aromatic hydrocarbons, long-chain petroleum hydrocarbons or larger chlorinated compounds. This technology was developed in the 1950s and became popular in the remediation industry, probably due to its ability to achieve remediation goals with a minimum investment. However, its limitations, including the inability to remove or destroy contaminants from the soil have resulted in a loss of interest in this remediation process in some areas such as the USA and Europe, despite growing practice in some other countries like China (Shen et al., 2019).

Chemical stabilization

Chemical stabilization represents a cost effective and sustainable risk-based approach to the remediation of contaminated soil. The technique involves the incorporation of amendments into the soil to render the contaminant less available (less mobile and less bioavailable), and therefore it represents less risk. It is considered a sustainable technique as it does not involve offsite disposal of contaminated material and often makes use of relatively low-cost amendments. Industrial waste by-products such as red mud, steel slag, and waste materials high in phosphate are few of those amendment materials.

Chemical stabilization is applied largely for metal contamination, which cannot be treated by degradation, and is most suited to diffuse contamination of the surface soil. It is adaptable to site-specific conditions and can be applied in place on site. The application can be formulated based on treatability trials to achieve specific remediation targets, such as reducing the leachable concentration or reducing the bioavailable concentration below acceptable limit. It is most typically applied to diffuse surface contamination by incorporation of the amendment into the surface layer of the soil.

Stabilization of the contaminant is achieved by chemical reaction. The mechanism of stabilization depends on the nature of the amendment applied. Broadly, soil amendments for chemical stabilization may be classified into three categories: those that stabilize by increasing soil pH, those that result in the precipitation of specific less soluble minerals, and those that provide a surface for sorption of the contaminant in the soil.

Soil amendments that increase soil pH include lime and alkaline by products such as red mud and magnesium oxide. These stabilize metals in the soil by increased sorption to soil and precipitation which occurs at higher pH (pH 7–9 is typically optimal).

Amendments that target the precipitation of particular minerals include phosphates and sulfates, which aim to stabilize metals by the precipitation of low solubility metal phosphates and metal sulfates.

Materials that provide increased sorption include those with iron and aluminum oxides such as steel slag, red mud, fly ash and biochar. Chemical stabilization is achieved by specific and non-specific adsorption of metals to these surfaces.

Bioremediation

Bioremediation techniques involve the use of microorganisms to clean up contaminated soils, sediments, water, or air. This is not a new concept, as people have been utilizing this technology since the beginning of civilization, for example through composting. However, the term “bioremediation” is relatively recent. Bioremediation takes advantage of natural processes and relies on microorganisms such as bacteria, fungi, actinomycetes, and cyanobacteria, which can be found in nature or cultured in the laboratory.

The basic principle of bioremediation is that microorganisms can be harnessed to degrade or transform contaminants into nonhazardous or less hazardous chemicals (Fig. 3). These organisms obtain energy and nutrients for their growth and reproduction by oxidizing organic matter. To accomplish this, they require the presence of other chemicals that can be reduced, which are generally the oxidized inorganic compounds, such as oxygen, nitrate, iron(III), manganese(IV), sulfate, and carbon-dioxide (Palmisano and Hazen, 2003).

Under favorable conditions, microorganisms are able to gain energy or carbon from the breakdown of organic contaminants. Although most organic contaminants are biodegradable, metals and radionuclides are not. Despite this, microorganisms can interact with these non-degradable contaminants, altering their oxidation state through electron transfer, and thereby converting them from one chemical form to another.

The success of bioremediation depends not only on the types of contaminants but also on the characteristics of the site. When the remediation depends solely on indigenous microorganisms without any human intervention, it is referred to as intrinsic bioremediation or natural attenuation. On the other hand, when microbial activities are anthropogenically accelerated using microbial growth stimulating agents, such as nutrients or oxygen, this is known as engineered bioremediation or enhanced

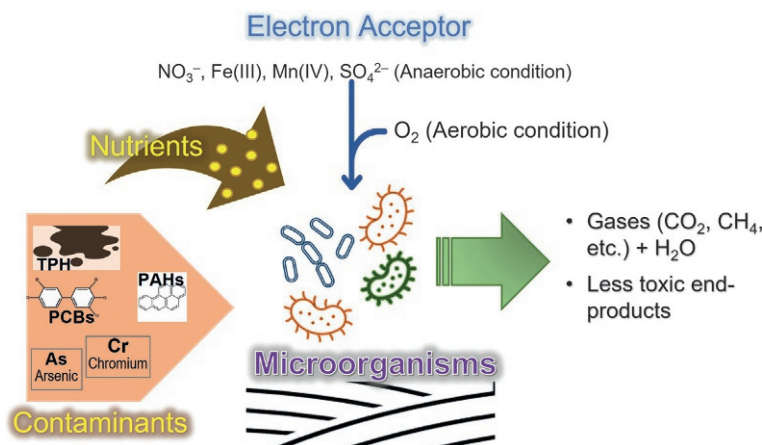


Fig. 3 Microorganisms degrade organic contaminants and transform metals to forms that are less toxic in bioremediation.

bioremediation. Bioremediation can be conducted at the site where the contamination occurs, termed as in situ processes whereas off-site remediation (ex situ processes) involves removal of contaminated media to a treatment area (NRC, 1993).

Bioremediation is a viable alternative to traditional clean-up procedures, as it can potentially be more efficient, safer, and less expensive. There are several successful reports of bioremediation being applied to contaminated medium, however those exclusively in soil are limited. Among soil bioremediation examples, the remediation of fuel oil-contaminated soil at the Penrice Soda premises in Osborne, South Australia, stands out. It was carried out using a combination of biopile and enhanced bioremediation techniques, treating affected area of 3000 m³ with approximately 5000 KL of spill (ZWSA, 2013). Likewise, hydrocarbon remediation from Antarctic soil (Martínez Álvarez et al., 2017) or other examples given in the review by Stepanova et al. (2022) are worth noting.

Phytoremediation

Phytoremediation broadly refers to the use of plants for the remediation of contaminated soil. It is a green and low-cost remediation technology which has particular applicability to stabilize and remediate degraded environments such as those affected by mining and other industrial activities. It is best suited to shallow and widely dispersed contamination because its effectiveness relies on the root zone of the plants. Phytoremediation has been applied to the remediation of metals and organic contaminants, including PCBs, PAHs, BTEX compounds (benzene, toluene, ethylbenzene, and xylene), TCE, explosive residues, and pesticides.

There are two main applications of phytoremediation for metals: phytostabilization and phytoextraction. Phytostabilization refers to the stabilization of contaminants in the soil, while phytoextraction involves the extraction of contaminants from the soil. For organic contaminants, the phytoremediation process may involve immobilization, volatilization and mineralization of these compounds. Immobilization of contaminants occurs in the root environments, while certain compounds release to the air through plant roots, stems and leaves (volatilization), and mineralization is the conversion of toxic compound to non-toxic constituents (Tripathi et al., 2020).

The nature of the application defines how phytoremediation is designed and implemented. Plants differ in their ability to take up contaminants in soil, therefore for phytoextraction, plants known as “hyperaccumulators” are used to maximize uptake of the contaminant in soil into the above ground biomass of the plant. The plants may then be harvested and disposed of after sufficient uptake of the contaminant into the biomass of the plant. In the case of phytostabilization, where the plants will accumulate and stabilize the contaminant in the root zone of the plant, the effectiveness of different plant species will also vary.

The effectiveness of different species of plants depends on their physiology and the contaminant of concern because some contaminants are readily taken up by plants due to their chemical properties. The plant species used must also be resistant to toxicity from the contaminant, which would restrain its growth. Generation of sufficient biomass is also required for effective phytoremediation. Researchers have investigated which plants are most effective for different applications (D’Orazio et al., 2013; Jiménez et al., 2011; Gobelius et al., 2017). For example D’Orazio et al. (2013) reported that the greatest efficiency of pyrene removal from soil of 32% was achieved with *Medicago sativa* L. (Alfalfa). Jiménez et al. (2011) examined the accumulation of zinc, lead and copper from mine tailings by three plant species. They found *Diurichia viscosa* (L.) Greuter, 1973 to be a suitable “phytoextractor,” whereas *Euphorbia pithyusa* subsp. *cupanii* was considered a good “phytostabilizer” (Jiménez et al., 2011). Gobelius et al. (2017) assessed the capability of local plant species for the removal of PFAS from soil. Mixed stands of silver birch (*Betula pendula* Roth) and spruce (*Picea abies* (L.) H. Karst), and an understory of ground elder (*Aegopodium podogragaria* L.) had the potential to remove 1.4 g of PFAS per year per hectare.

The efficiency of phytoremediation may be restrained for several reasons due to soil properties, type of contaminants, plant physiology, and climatic conditions (temperature, rainfall and water availability). The phytoavailability of a contaminant relates to

its availability for plant uptake, which for metals relates to pH, solubility of the mineral form, and association with soil. For organic contaminants, it is related to sorption, degree of hydrophobicity, and a log octanol–water partition coefficient (Alkorta and Garbisu, 2001).

To meet the desired remediation goal, it may be necessary to use a chemical amendment to enhance the process. This can involve the use of a chelate or acid for phytoremediation of metals, or a surfactant in the case of organic contaminants. These substances promote the phytoavailability of the contaminant to increase their uptake by the plant.

In addition, recent research has focused on the role of endophytes, which are microorganisms found inside the internal tissues of plants, particularly in the context of organic contaminant degradation, promotion of plant growth, and increase of tolerance to contaminant toxicity (Feng et al., 2017). For example, a strain of *Enterobacter* sp. living in poplar (*Populus* sp.) plants could degrade TCE. This plant was harvested from a TCE contaminated soil site (Kang et al., 2012). Furthermore, Yao et al. (2022) reported that an endophytic bacterial community helped the black locust tree (*Robinia pseudoacacia* L.) in coping with metal stress while growing in the metal tailings of the Manao Mountain area in China.

Soil washing

Soil washing is a waste volume minimization treatment technology based on separation by physical and or chemical processes. It is a widely established treatment for the remediation of contamination in soils.

The technology enables the separation of contaminant from bulk soil, reduction of waste volume, easier handling of waste and reuse of the bulk soil fraction on the site as backfill. Contaminants sorbed onto fine soil particles are separated from bulk soil in an aqueous system, usually on the basis of particle size or density. The means of physical separation include:

- Mechanical (vibrating screens),
- Hydrodynamic (settling velocity, i.e., sedimentation),
- Gravity (density),
- Froth flotation (uses air or mechanical agitation to suspend particles or oil based on its hydrophobic properties),
- Magnetic separation (magnetic properties of materials and contaminants present in soil),
- Attrition scrubbing (agitation to remove contaminant from coarse particles).

The wash solution may use a basic leaching agent, surfactant, acid (pH adjustment), or chelating agent for enhanced removal of organics and metals (Fig. 4).

The target contaminant groups for soil washing are: metals, semivolatile organic compounds (SVOCs), selected volatile organic compounds (VOCs), and pesticides.

There are several considerations and limitations to the applicability and effectiveness of soil washing including:

- Additives must be of low toxicity and ideally be biodegradable,
- Contaminant removal efficiency for complex waste mixtures, high clay soils or very large concentrations,
- Large humic content in soil may require pre-treatment,
- Treatment of the washing solution is needed to recover contaminants from the waste stream,
- Recycling of the extraction solution may be required for cost effectiveness,
- The effect of the chemical washing solution may render soils unsuitable for some future use due to alteration of the soils' physical and chemical properties by the chemical washing process.

The remaining bulk soil is not totally contaminant-free. Therefore, for soil washing to be successful, the level of contamination must be below a site-specific action limit (e.g., based on risk assessment).

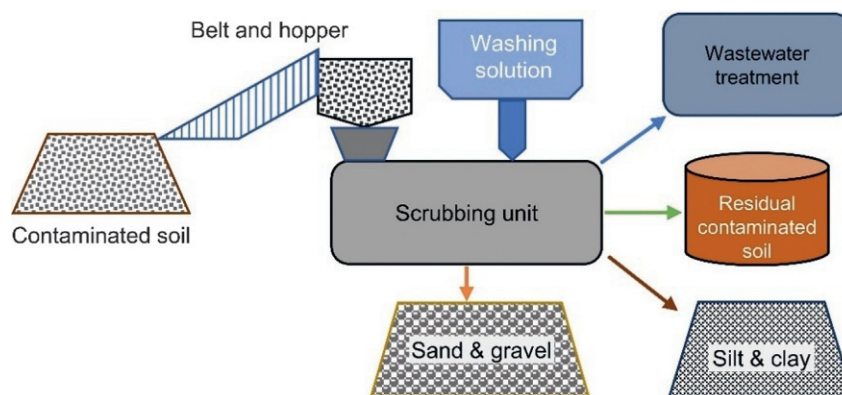


Fig. 4 Schematic diagram of soil washing treatment.

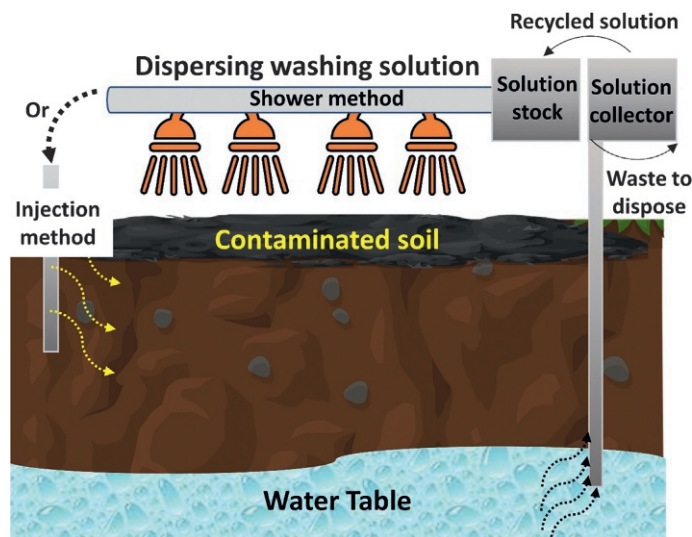


Fig. 5 Schematic diagram of soil flushing treatment.

Other considerations relate to the volume of contaminated soil to be treated (volume or coarse to fine material) and the capacity of the washing plant (tons hr^{-1}). Economies of scale should be applicable because of the large upfront costs of establishing the treatment succession on site; therefore the cost per cubic meter will be higher for a smaller site.

Soil flushing

Soil flushing is an in situ remediation technique that operates on a similar principle to chemical soil washing. However, in soil flushing the aqueous washing solution is introduced to the soil in place by injection or flooding. Contaminants in the soil are desorbed into the solution and extracted for treatment by extraction wells (Fig. 5).

Chemical reduction

Chemical reduction or in most cases the in situ chemical reduction (ISCR) technique is applied to convert a more toxic form of the contaminant into a less toxic form. Transformation of hexavalent chromium (Cr(VI)) or a dense non-aqueous phase liquid (DNAPL) compound trichloroethene (TCE) is one such case. In the case of Cr the tri-valent form is preferable due to its much lower toxicity and stability over the severely toxic Cr(VI) (Fan et al., 2016).

For this method of remediation, a chemical or material is used that has the electron transfer capacity to reduce the target compound or ions. Zero valent iron (ZVI) is a dominant type of material used. The particle size of ZVI varies, and so does its effectiveness. It varies in size from macro to extremely small particles, which is often referred to as nano zero valent iron (nZVI).

The mode of application of this reducing material depends on the site of the soil and the type of contaminant. It can be (i) applied directly or (ii) injected via a permeable reactive barrier (PRB) (Fig. 6). In the case of direct injection, the reactive material is presented as a slurry using water or nontoxic oil (e.g., vegetable oil) before being poured into several holes at the contaminated soil site. On the other hand, PRB is applied to treat the leachate flow of the contaminated site by creating a trench for the barrier where the reactive agent is placed. The outlet of the leachate is deemed to be less toxic once passed through this permeable trench (Fig. 6) (US-EPA, 2012c).

With the progress of this technology, ISCR material can be applied to achieve both reductive transformation of the target contaminant and also to enhance the bioremediation. The latter is achieved when the ISCR material like ZVI is incorporated with an organic additive, which is a microbial growth enhancer.

In situ oxidation

Like in situ chemical reduction (ISCR), the principle of in situ chemical oxidation (ISCO) is the conversion of the toxic form of a contaminant to it a less toxic form. For example, the reduction of TCE occurs in presence of permanganate following the chemical reaction of:



Oxidative breakdown of solvents, fuels or types of pesticides is the commonly targeted application. To that end, a range of reactive agents known as oxidants, including hydrogen peroxide, ozone, persulfate, and permanganate are applied. Depending on the target contaminant and source area, the type of oxidant is chosen and applied underground. The dissipation of these injections encounters

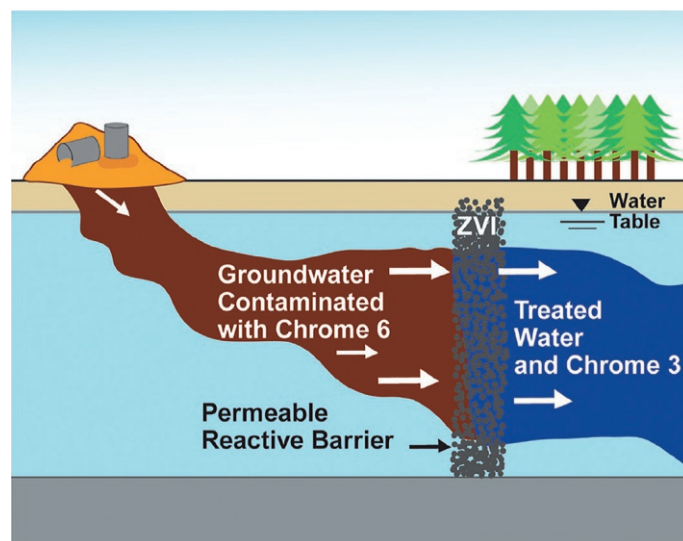


Fig. 6 The illustration of a permeable reactive barrier made of zero valent iron (ZVI) for the transformation of hexavalent chromium (US-EPA, 2012c).

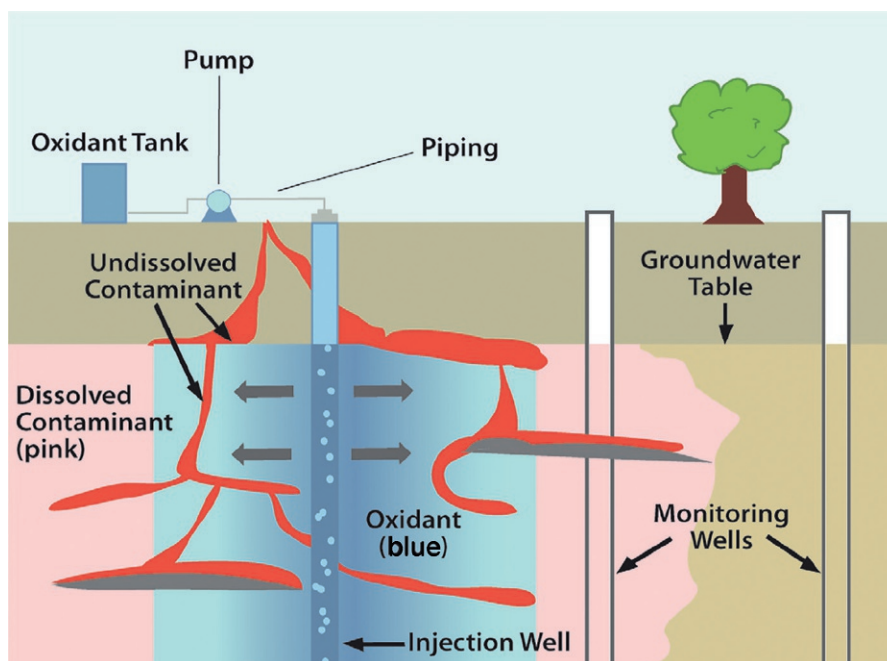


Fig. 7 The illustration of in situ oxidation treatment system where the oxidants are typically injected underground by pumping them into wells (US-EPA, 2012c).

the target contaminant and acts as an electron receptor. This dissipation occurs most effectively when the agents are supplied below the surface through an installation with several holes at various depths from the surface to depth (Fig. 7) (US-EPA, 2012c).

Improvement of oxidants and the mode of application have been tested. The surfactant mediated oxidation process of polycyclic aromatic hydrocarbons (PAHs) and total petroleum hydrocarbons (TPHs) was deemed rapid and enhanced (Wang et al., 2013). Surfactants could increase the bioavailability of these petroleum compounds to be treated by oxidants.

The ISCR has a long history of remediation practice because of its fast performance and low-cost; also, it is performed underground leaving the surface soil arrangement intact. However, the application of ISCR is more appropriate for the remediation of groundwater than for the soil matrix.

Electrokinetic remediation

Electrokinetic remediation is an environmental technology that was designed specifically for removing pollutants from soil, sediments, and sludge, although it may also be used on any solid porous material. This technology uses a low-level direct current between electrodes placed across a section of contaminated media (Acar and Alshawabkeh, 1993).

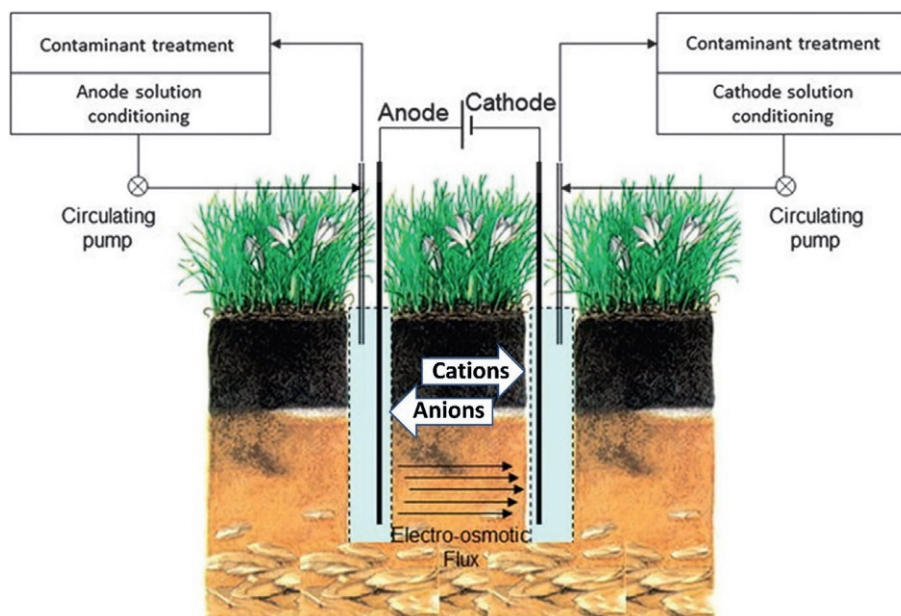


Fig. 8 Ion transport mechanisms in electrokinetic remediation (Cameselle, 2014).

The electrokinetic remediation principle is similar to that of a battery. When the electrodes, cathodes and anodes are introduced into the soil matrix and a low-intensity current is applied, electroosmosis and ion migration take place (Fig. 8). The direction of aqueous phase contaminant movements depends on their charge. Anions and negatively charged organic molecules move toward the anode, whereas metals and positively charged organic compounds move toward the cathode. The contaminants can then be recovered or deposited at the electrode.

Electrokinetic remediation can be used to clean up soils contaminated with radionuclides, metals, and organic compounds. The application of this remediation technique to non-polar organic contaminants can be enhanced by the use of different surfactants.

The advantage of the electrokinetic remediation technique is its potential to be used both in situ and ex situ at a low cost. However, this is a time-consuming process and issues such as excessive heat generation in the vicinities of electrodes, and changes in soil pH could be a matter of concern. The timeframe required for electrokinetic remediation may be up to one or more years depending on the type of contaminants, size of treatment area, electrode spacing, the nature of treatment (biodegradation or chemical oxidation) and site-specific conditions such as aquifer conductivity and lithology (FRTR, 2020a).

Soil vapor extraction

Soil vapor extraction (SVE) is a conventional in situ remediation technique that is applied for the treatment of volatile and some semi-volatile organic contaminants (most commonly petroleum hydrocarbons and chlorinated VOCs) in the vadose zone of soil. The vapor extraction works by producing advective flow within the vadose zone by a pressure gradient created by the application of a vacuum to the extraction wells.

The technology comprises three processes, namely extraction, treatment, and venting.

There is a range of approaches available for each of the treatment processes. The treatment system typically comprises (CRC CARE, 2019b): extraction wells, a vacuum pump or blower, a low permeability surface cap, a vapor or liquid separator, and soil vapor treatment system. The soil vapor treatment system may take the form of granular activated carbon (filtration of contaminants from vapor), combustion and condensation. Venting is by exhaust to the atmosphere with a final scrubbing by the filtration system before release to the atmosphere (Fig. 9).

The effectiveness of the treatment may be affected by site conditions including lithology and heterogeneity of the subsurface, temperature, soil moisture, natural organic carbon content, and depth to groundwater (FRTR, 2020b). These factors affect the permeability of the air flow and therefore the efficiency of the process.

Contaminant properties also need to be considered when applying the treatment because SVE can be used only to remediate VOCs with a Henry's Law constant >0.01 or vapor pressure >0.5 mmHg (FRTR, 2020b). It can also be applied to non-aqueous phase liquids, dissolved organics, the sorbed phase and free soil vapor.

The costs of SVE tend to increase with the size and depth of contamination.

Thermal treatment

Thermal soil remediation removes organic contaminants from soil and other solid media by subjecting the material to temperatures high enough to volatilize the target contaminant. It is widely applied to organic contaminants, including persistent organic contaminants such as polychlorinated biphenyls, organochlorines and pesticides. A combined approach of injecting hot air stream

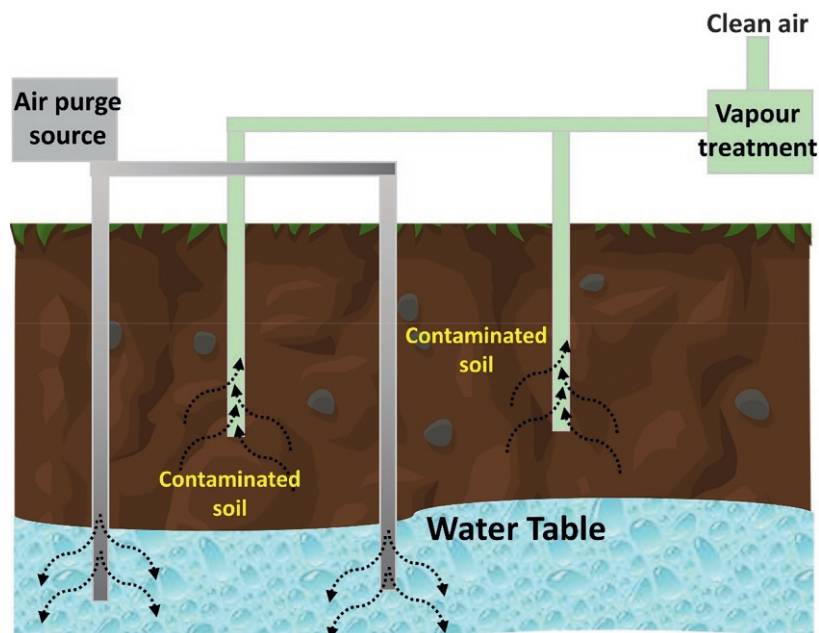


Fig. 9 Schematic diagram of soil vapor extraction in combination with the purging of air to the sub-surface.

and vapor extraction is often referred to as “Thermally enhanced SVE” (FRTR, 2020c). Remediation may be achieved over a relatively short timescale (weeks to months). It works by separating the contaminant from the soil by heating it to boiling point and transforming the contaminant to the gas phase. The gas then passes through an off-gas treatment system to destroy or absorb/adsorb off-gases generated by the thermal treatment.

The thermal treatment may have a range of configurations (Vidonish et al., 2016):

- Fired thermal desorption (soil fed through a rotary kiln with a flame),
- Enhanced thermal conduction (heating rods pipe hot air through the soil mound),
- In situ thermal (several methods, including hot air injection, steam injection, electromagnetic heating, electrical resistance heating, vitrification and electrothermal dynamic stripping), and
- Indirectly heated thermal desorption (similar to fired thermal, but the heat source is external to the desorption chamber).

For thermal treatment, soils of lower permeability (silty and clay) are less expensive to remediate as they require less gas flow. Similar to other treatment technologies, the depth and thickness of the contamination zone is associated with higher costs of application of the technology.

Conclusions

Several technologies are available for the remediation of contaminated soil. Choosing the appropriate remediation technology depends on the type of contaminants and the site conditions. In addition, sustainability needs to be considered, such as cost effectiveness and potential environmental impacts and benefits. The configuration of treatment technologies should be tailored to site-specific application to optimize the remediation of a given site.

References

- Acar YB and Alshawabkeh AN (1993) Principles of electrokinetic remediation. *Environmental Science & Technology* 27(13): 2638–2647. <https://doi.org/10.1021/es00049a002>.
- Alkorta I and Garbisa C (2001) Phytoremediation of organic contaminants in soils. *Bioresource Technology* 79(3): 273–276.
- Bates E and Hills C (2015) *Stabilization and Solidification of Contaminated Soil and Waste: A Manual of Practice*. Édit Media H.
- Cameselle C (2014) Electrokinetic transport in soil remediation. In: Kreysa G, Ota K, and Savinell RF (eds.) *Encyclopedia of Applied Electrochemistry*. New York, NY: Springer. https://doi.org/10.1007/978-1-4419-6996-5_120.
- CRC CARE (2019a) *Technology Guide: Soil-Contaminant. National Remediation Framework*. CRC for Contamination Assessment and Remediation of the Environment.
- CRC CARE (2019b) *Technical Guide: Soil Vapour Remediation. National Remediation Framework*. CRC for Contamination Assessment and Remediation of the Environment.
- D’Orazio V, Ghanem A, and Senesi N (2013) Phytoremediation of pyrene contaminated soils by different plant species. *CLEAN–Soil, Air, Water* 41(4): 377–382.
- Fan D, O’Brien Johnson G, Tratnyek PG, and Johnson RL (2016) Sulfidation of nano zerovalent iron (nZVI) for improved selectivity during in-situ chemical reduction (ISCR). *Environmental Science & Technology* 50: 9558–9565.

- Feng NX, Yu J, Zhao HM, Cheng YT, Mo CH, Cai QY, and ... Wong MH (2017) Efficient phytoremediation of organic contaminants in soils using plant–endophyte partnerships. *Science of the Total Environment* 583: 352–368.
- FRTR (2020a) Technology Screening Matrix—Electrokinetic-Enhanced Remediation. <https://frtr.gov/matrix/Electrokinetic-Enhanced-Remediation/>. Accessed 24/01/2023.
- FRTR (2020b) Technology Screening Matrix – Soil Vapour Extraction. <https://frtr.gov/matrix/Soil-Vapor-Extraction/>. Accessed 02/11/2021.
- FRTR (2020c) Technology Screening Matrix – Soil Thermal Treatment. <https://frtr.gov/matrix2/section4/4-9.html>. Accessed 02/11/2021.
- Gobelius L, Lewis J, and Ahrens L (2017) Plant uptake of per-and polyfluoroalkyl substances at a contaminated fire training facility to evaluate the phytoremediation potential of various plant species. *Environmental Science & Technology* 51(21): 12602–12610.
- Jiménez MN, Bacchetta G, Casti M, Navarro FB, Lallena AM, and Fernández-Ondoño E (2011) Potential use in phytoremediation of three plant species growing on contaminated mine-tailing soils in Sardinia. *Ecological Engineering* 37(2): 392–398.
- Kang JW, Khan Z, and Doty SL (2012) Biodegradation of trichloroethylene by an endophyte of hybrid poplar. *Applied and Environmental Microbiology* 78(9): 3504–3507.
- Liu L, Li W, Song W, and Guo M (2018) Remediation techniques for heavy metal-contaminated soils: Principles and applicability. *Science of the Total Environment* 633: 206–219.
- Martínez Álvarez LM, Ruberto LAM, Lo Balbo A, and Mac Cormack WP (2017) Bioremediation of hydrocarbon-contaminated soils in cold regions: Development of a pre-optimized biostimulation biopile-scale field assay in Antarctica. *Science of the Total Environment* 590–591: 194–203.
- NRC (1993) *In Situ Bioremediation—Evaluation. I*. National Research Council (U.S.), National Academy of Sciences.
- Palmisano A and Hazen T (2003) *Bioremediation of Metals and Radionuclides: What It Is and How It Works*, 2nd edn. <https://doi.org/10.2172/820771>.
- Shen Z, Jin F, O’connor D, and Hou D (2019) Solidification/stabilization for soil remediation: An old technology with new vitality. *Environmental Science & Technology* 53: 11615–11617.
- Stepanova AY, Gladkov EA, Osipova ES, Gladkova OV, and Tereshonok DV (2022) Bioremediation of soil from petroleum contamination. *Processes* 10: 1224.
- Tripathi S, Singh VK, Srivastava P, Singh R, Devi RS, Kumar A, and Bhadouria R (2020) Phytoremediation of organic pollutants: Current status and future directions. In: *Abatement of Environmental Pollutants*, pp. 81–105. Elsevier.
- US-EPA (2012a) *A Citizen’s Guide to Capping*. . EPA 542-F-12-004.
- US-EPA (2012b) In: AGENCY, E. P (ed.) *A Citizen’s Guide to Engineered Vertical Barriers*. EPA.
- US-EPA (2012c) In: AGENCY, E. P (ed.) *A Citizen’s Guide to In Situ Chemical Reduction*. EPA.
- Vidonish JE, Zygourakis K, Masiello CA, Sabadell G, and Alvarez PJ (2016) Thermal treatment of hydrocarbon-impacted soils: A review of technology innovation for sustainable remediation. *Engineering* 2(4): 426–437.
- Wang WH, Hoag GE, Collins JB, and Naidu R (2013) Evaluation of surfactant-enhanced in situ chemical oxidation (S-ISCO) in contaminated soil. *Water, Air, & Soil Pollution* 224: 1713.
- Yao Y, Zhang X, Huang Z, Li H, Huang J, Corti G, Wu Z, Qin X, Zhang Y, Ye X, Fan H, and Jiang L (2022) A field study on the composition, structure, and function of endophytic bacterial community of *Robinia pseudoacacia* at a composite heavy metals tailing. *Science of the Total Environment* 850: 157874.
- ZWSA (2013) Zero Waste SA. <https://www.greenindustries.sa.gov.au/publications-management-of-contaminated-soils>.

Soil change under different scenarios

Joann K Whalen, Department of Natural Resource Sciences, Macdonald Campus of McGill University, Sainte-Anne-de-Bellevue, QC, Canada

© 2023 Elsevier Ltd. All rights reserved.

Introduction	344
Soil change due to erosion	345
Soil change due to erosion: The case of the Chinese Loess Plateau	345
Soil change due to loss of organic matter	349
Organic matter for soil stability	349
Soil change from organic matter loss after deforestation: The case of New Zealand pastures	350
Organic matter for soil fertility	352
Soil change due to organic matter loss following cultivation: The Great Plains prairie grasslands	352
Soil change due to element imbalance	356
Conclusions	358
References	359

Abstract

Soils change naturally as the parent material and minerals are progressively weathered, and in response to the growth of plants, animals and microorganisms across landscapes. In the past century, human-induced changes to the global climate and land use have put soils at risk of degradation. Erosion of fertile topsoil, loss of soil organic matter, element imbalances resulting from salinization, trace metal pollution and excessive fertilization, are some of the most damaging soil changes that we must resolve to sustain our soils, a vital natural resource. This chapter explains the reasons for these changes in soil, and the human activities that are likely to accelerate or slow this rate of change. Possible scenarios to improve the soil's resilience to change and prevent its depletion are presented.

Key points

- Soil change is associated with geological processes and land use conversion for agriculture.
- Soil erosion and loss of soil organic matter from fragile landscapes are associated with food insecurity, water shortages and ecological damage.
- Loss of sediments and soil organic matter can be prevented by revegetating and stabilizing vulnerable areas with built control structures, and adopting conservation agricultural practices.
- Element balances in soil are affected by improper fertilizer use and irrigation, as well as contaminants deposited from atmospheric pollution.
- Judicious fertilizer and water use are needed to control the amounts of essential and non-essential elements that affect biological processes and abiotic reactions in soil.
- Local, regional and national programs are needed to measure and monitor soil change.

Introduction

The United Nations declared 2021–2030 as the Decade on Ecological Restoration, which is a call for action to prevent, halt and reverse ecosystem degradation worldwide. We must rebuild ecosystems to achieve the sustainable development goals of food security, climate change mitigation, water and biodiversity conservation and eradication of poverty. This requires us to address issues related to land degradation, which affects an estimated 25% of the land surface and is estimated to reach 95% of the global land area by 2050. Land degradation is the change in the biophysical environment that causes deterioration or loss of productive soils. Such a change in the soil condition can limit the growth of desirable vegetation, which we use as food, fuel, fiber and depend upon for oxygen, regulation of hydrologic processes and other essential ecological services. Soil change can also diminish its biodiversity, capacity for biogeochemical cycling and ability to attenuate contaminants in the environment.

Soil change is largely a human-induced phenomenon that has dire consequences for the world's population. The erosion of about 24 billion tons of fertile soil each year impacts 3.2 billion people, approximately 40% of the world's population. About 33% of soils worldwide have lost soil organic matter due to historical expansion of agriculture and pastoralism, which converted many native ecosystems (e.g., peatlands, forests, grasslands) for arable land use. Element imbalance due to soil salinization affects more than 830 million hectares of cultivated land worldwide and is expected to impact about 50% of arable land by 2050. These alarming trends must be stopped and reversed to ensure that soils continue to support vital ecosystem functions, creating a stable or improved land resource that can provide the necessary habitat and products that are needed to fulfil the Earth's biological needs.

This chapter explains the root causes of soil change due to human activities related to the conversion of natural ecosystems for agricultural use in the context of global climate change. Case studies are presented to explain the consequences of soil change,

as well as the time needed for regional-scale restoration processes to be effective. These examples offer some concrete solutions to the problems that result from soil change, while explaining the necessity of stakeholder involvement, policies and economic investments to control soil change.

Soil change due to erosion

Erosion is part of the natural weathering process that leads to the physical and chemical degradation of rocks, sedimentary materials and minerals, moving smaller particles and dissolved substances from one location to another in the landscape. Soil erosion refers to the physical movement of soil particles and associated organic matter, generally from the soil surface, following abrasion by mechanical forces. Wind, water, ice and machinery exert sufficient force to dislodge particles and move them to a new location, with the help of entropy and gravitational forces. Natural soil erosion, also referred to as geological erosion, occurs at a relatively slow rate of <0.2 mm per year, which is equivalent to the annual removal of approximately $1 \text{ Mg topsoil ha}^{-1}$ (Montgomery, 2007). In the absence of human activities, geological erosion is responsible for gradually moving soil from weathered landscapes and depositing particles in new locations, over hundreds to thousands of years. Geological erosion adds dust and water-borne sediments to the soil profile developed through weathering, at a rate of $0.04\text{--}0.08$ mm per year; this rate of soil formation is close to the global average rate of erosion of 0.05 mm per year under native vegetation (Montgomery, 2007). Thus, there is an equilibrium between soil formation and erosion in ecosystems that are free from human disturbance.

Soil change due to erosion: The case of the Chinese Loess Plateau

Geological erosion creates landforms of major importance, such as the Chinese Loess Plateau, which covers $640,000 \text{ km}^2$ in Gansu, Shaanxi and Shanxi provinces. Home to more than 100 million people, the Loess Plateau is the world's largest formation of yellow wind-blown sediment, which develops into a unique soil type called loess. Soil on the Loess Plateau originates from mountains (the Gobi Altai Mountains, the Hangayn Mountains and the Qilian Mountains) in south-central Mongolia and northwestern China. Physical weathering of the rocks in these mountains, aided by freeze-thaw and flowing water, generates small-sized grains that gradually move into the Gobi Desert and surrounding sand deserts. The prevailing winds, together with monsoons and dust storms, transport silt-sized particles from the Gobi and other deserts for hundreds of kilometers. The gradual decline of wind energy upon arrival at the Loess Plateau means that larger, heavier particles are deposited in the northwest region and finer clay particles are found in the southeast of the area. Aeolian sediments are deposited at a rate of about 0.1 mm per year, so it takes around 10,000 years to produce 1 m of loess soil. The wind-borne dust is gradually transformed into a grey-yellow palaeosol through three soil-forming processes: (1) primary calcite particles within the dust are dissolved in rainwater and leached, leading to secondary CaCO_3 deposition on silt and clay particles, (2) organic matter accumulates due to the growth and turnover of organic substances from plants, microorganisms and animals, and (3) some minerals in the loess are gradually weathered into finer clay particles and form complexes with organic substances and iron oxides (Zhao et al., 2017).

Agriculture speeds up soil erosion, particularly on hilly, sloping landscapes in the Chinese Loess Plateau. Rates of soil erosion are nearly two orders of magnitude greater in conventional agriculture, with a global average of 3.9 mm soil eroded per year, compared to under native vegetation (Montgomery, 2007). Reports of soil erosion on the Loess Plateau date back more than 2000 years, when farming began in the region. Ongoing soil loss changed the geomorphology, such that 70% of the Loess Plateau became dominated by hills and gullies as the loess gradually washed into tributaries of the Yellow River. Excessive sediment loading in the Yellow River is linked to more than 1500 floods between 602 BC and AD 1947, with at least 26 major and 9 severe shifts in the channel of the Yellow River during this time (Douglas, 1989). Each flood inundated a large area of the South China Plain, depositing huge quantities of silt and sand across the delta, displacing villagers and negatively affecting river commerce. By the 1980s, the overuse and overgrazing of highly-erodible loess produced some of the highest erosion rates ever recorded. Maximum annual suspended sediment loads of 5000 Mg km^{-2} to more than $35,000 \text{ Mg km}^{-2}$ in tributaries of the Yellow River (Table 1) led some observers to remark that the loess highlands below Hekouzhou were the most severely eroded land on Earth (Douglas, 1989). Around this time, the Yellow River reached its capacity for sediment loading, meaning that the river was carrying the maximum sediment it could

Table 1 Maximum annual suspended sediment loads recorded in tributaries draining the Chinese Loess Plateau from 1956 to 1980.

River tributary	Station	Drainage area (km^2)	Maximum annual suspended sediment load (Mg km^{-2})
Yellow	Longmen	497,559	4944
Weihe	Huaxian	106,498	9953
Huangfu	Huangfu	3199	53,454
Beiluo	Zhuangtuo	25,154	8587
Kuye	Wenchuan	8645	35,049
Wuding	Chuankou	30,217	14,561

Adapted from Douglas I (1989) Land degradation, soil conservation and the sediment load of the Yellow River, China: Review and assessment. *Land Degradation and Development* 1: 141–151. doi: 10.1002/ldr.3400010206.

transport. Centuries of sediment deposition have created an elevated river that rises as much as 10 m above the level of the surrounding land. Millions of inhabitants in the Yellow River Delta live with the threat of disastrous flooding, which endangers human life and destroys property, because the volume of fast-flowing, sediment-laden water in the Yellow River has the capacity to breach dams, dikes and levee banks. Government officials and their advisors agree that perpetual control of soil erosion on the Loess Plateau is the only way to ensure public safety and prevent further degradation of the Yellow River.

Stabilizing the loess is the most direct way to reduce sediment loading in the Yellow River. This begins with terracing, a land-forming practice that cuts a sloped land surface into a series of flat platforms that resemble steps. Terracing was a traditional practice in this area that was improved by better design and construction after the 1950s. Less than 1% of the Loess Plateau was terraced in the mid-1950s, and this increased to about 15% of the land area by 2018 (Ji et al., 2021). Terracing prevents soil loss because the outer edge of the terrace has an elevated earth embankment, at least 20-cm high, that prevents slope erosion by retaining sediments, water and dissolved substances. Terraced fields retain more water, as much as 100–200 mm after heavy rainfall events, without appreciable erosion, and always have higher soil moisture than sloped land (Liu et al., 2011). Greater soil moisture supports the growth of deep, extensive plant roots, resulting in more above-ground biomass and larger crop yields. In Zhuanglang County, the average grain yield was 27% greater in 3-year old terraces and 53% more in 7-year old terraces than on hilly fields with $\geq 10^\circ$ slope (Table 2; Liu et al., 2011). Since roots are generally not harvested, an intact root system with long, dense root hairs and lateral branching roots can effectively adhere to soil in the root:soil contact surface area, thereby preventing erosion.

Although terraces can retain about 10 times more sediment than cultivated slopes, landowners must manage their terraced fields carefully to avoid soil loss through wind and water erosion. Low-disturbance tillage methods such as minimum-tillage, no-tillage and contour plowing, which turns the soil across the slope rather than up-down the slope, are preferred to conventional plowing methods. These low-disturbance techniques can reduce the volume of surface runoff by more than 50% and prevent up to 75% of sediment loss from terraced fields (Wei et al., 2014). Including perennial crops in the rotation, intercropping to increase the vegetative cover in row-cropped systems, applying green manure, retaining straw as a mulch after crop harvest, and planting drought-tolerant varieties that produce high yields with less water are also promoted on the Loess Plateau (Wei et al., 2014). These agricultural practices are expected to prevent erosion while simultaneously increasing water retention and improving soil fertility, which will sustain crop yields and contribute to the economic prosperity of landowners in this region.

Erosion control in the hilly and gully regions of the Loess Plateau requires construction of water-retaining structures such as water caves, water banks and check dams that can trap surface runoff and sediments. Check dams built in the gullies of sloping lands are particularly effective for retaining sediments (Fig. 1). They consist of an earth dam, which may be built on top of stone or wood,

Table 2 Winter wheat production in dryland terraced fields, relative to unimproved hillslopes, during the 1990s in Zhuanglang County, eastern Gansu Province in the Chinese Loess Plateau.

Age of terraced field	Average yield (Mg ha^{-1})	Yield gain relative to unimproved hillslope (Mg ha^{-1})	Yield gain relative to unimproved hillslope (%)
First year of establishment	2.9	0.1	4
After 3 years of construction	3.6	0.8	27
After 5 years of construction	3.8	1.0	35
After 7 years of construction	4.3	1.5	53
Unimproved land with $\geq 10^\circ$ hillslope	2.8		

Adapted from Liu XH, He BL, Li ZX, Zhang JL, Wang L and Wang Z (2011) Influence of land terracing on agricultural and ecological environment in the loess plateau regions of China. *Environmental Earth Sciences* **62**: 797–807. doi: 10.1007/s12665-010-0567-6.

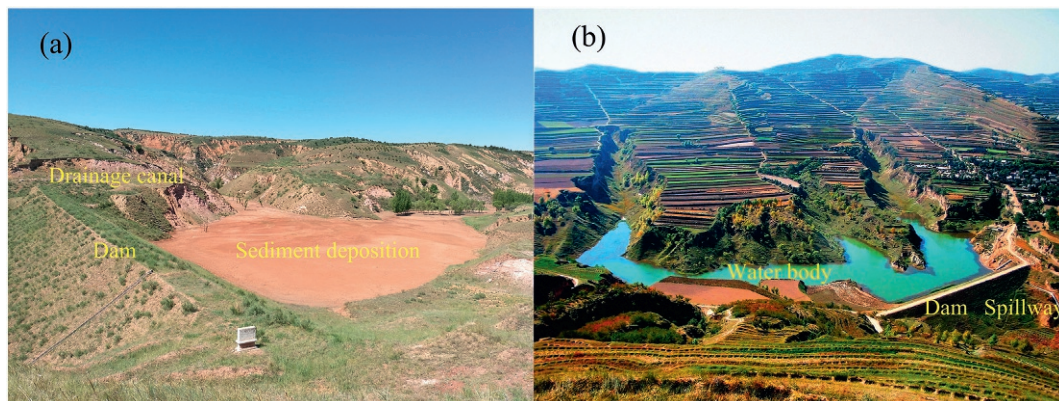


Fig. 1 Check dam in the Wuding River watershed of the Chinese Loess Plateau, showing the extent of sediment deposition (a), and water retention (b) behind the check dam. From Shi P, Zhang Y, Ren ZP, Yu Y, Li P and Gong JF (2019) Land-use changes and check dams reducing runoff and sediment yield on the Loess Plateau of China. *Science of the Total Environment* **664**: 984–994. doi: 10.1016/j.scitotenv.2019.01.430.

and a culvert that discharges flood water, generally without a spillway. The primary function of a check dam is to prevent gully incision by the network of headward-growing tributaries (first order streams of the Yellow River). Check dams can reduce the erosion energy to nearly zero in areas where gully and gravity erosion were the main cause of stream incision and bank failure that loaded sediments into the Yellow River (Shi et al., 2020; Wang et al., 2021). Sediments accumulate behind the earth dam, where they remain even if the dam fails, and effectively reduce the runoff velocity and erosion energy for many years. An estimated 110,000 check dams were constructed in thousands of gullies on the Loess Plateau since the 1960s. Each one has capacity for 0.01–5 million m³ of sediments, depending on the geomorphology of the gully, which could be up to 50-m deep, leading to a total retention of about 4 billion m³ of sediments by 2020 (Wang et al., 2021).

Check dams in gullies will prevent sediments from entering the Yellow River. A study in the Loess Mesa Ravine Region demonstrated that more 90% of the surface runoff and sediments from hillslopes were retained in a single check dam, although it only occupied 2.3% of the total conserved land (Fig. 2; Xu et al., 2012). The check dam was far more effective at controlling soil

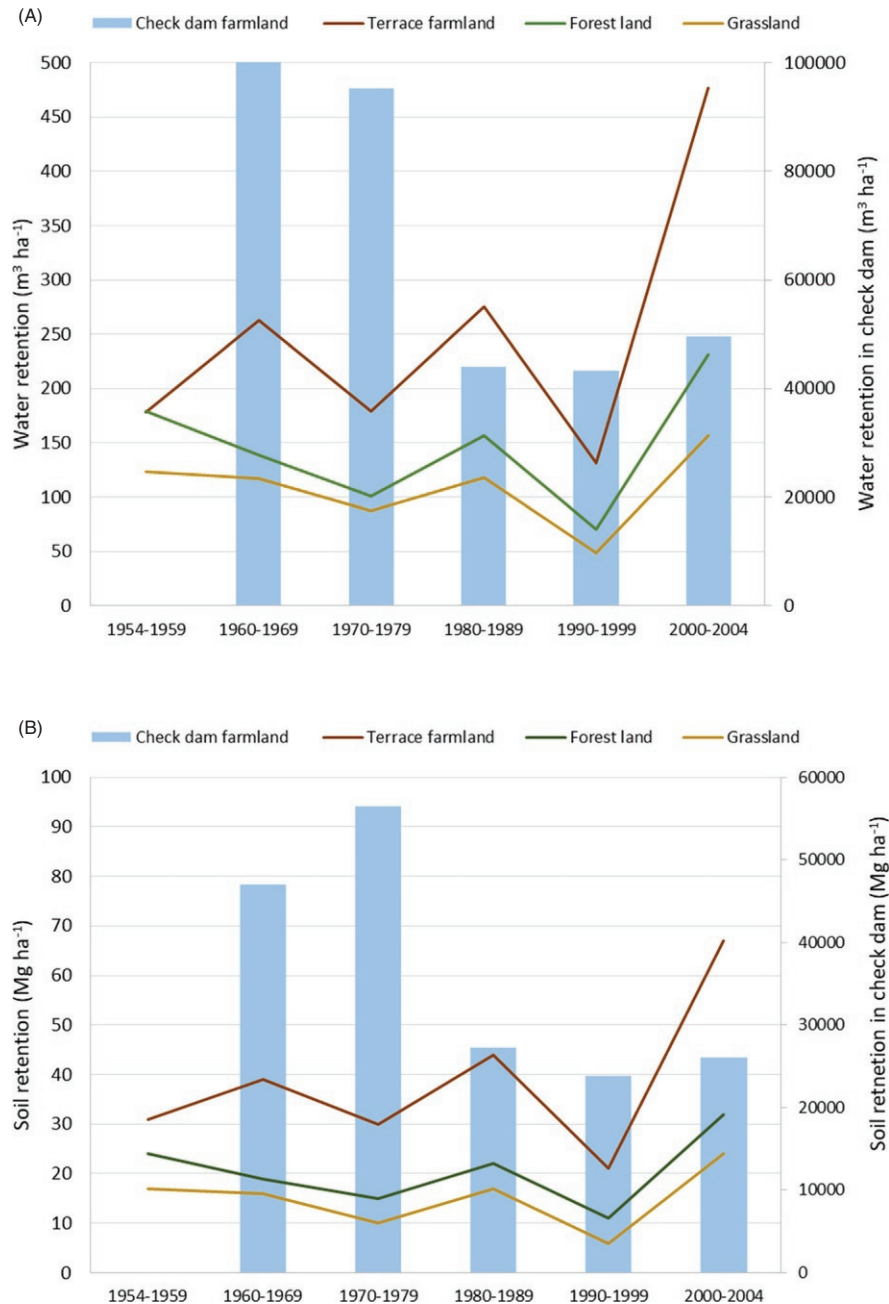


Fig. 2 Check dam farmland retained more than 90% of the (a) water and (b) eroded soil, compared to other conservation practice (terracing, tree-planting and grassland restoration) in the Nanxiaohogou Catchment, a 36.3 km² watershed in the Loess Mesa Ravine Region near Qingyang, eastern Gansu Province in the Chinese Loess Plateau. Adapted from Xu XZ, Li MJ, Liu B, Kuang SF and Xu SG (2012) Quantifying the effects of conservation practices on soil, water, and nutrients in the Loess Mesa Ravine Region of the Loess Plateau, China. *Environmental Management* 49: 1092–1101. doi: 10.1007/s00267-012-9835-4.

erosion than terracing and forest or grassland revegetation in this region. Sediments collected in check dams created more than 18,100 ha of new farmland by 1999, and this increased to 69,000 ha by 2007 (Shi et al., 2020). Furthermore, the farmland with check dams is 8–10 times more productive than the surrounding hillslopes, and it can produce about twice as much grain yield per unit area as terraced fields (Shi et al., 2020). Despite its small land area, check dam farmland could generate about 15% of the annual grain production on the Loess Plateau, which contributes to food security and economic prosperity in this region.

Traditionally, large tracts of the Chinese Loess Plateau were covered with perennial vegetation including forests, shrublands and grassland. As the human population of the Loess Plateau increased, from 8 million about 2000 years ago to 36 million in 1949 and more than 81 million in 1985, much of the native vegetation was removed as land was converted for agricultural use and steeper slopes were cultivated. Forests covered 50% of the Loess Plateau around 2000 years ago, but were only 6% of the land use by 1949 (Ji et al., 2021). The rapid expansion of cultivated lands, coupled with overgrazing by goats and excessive gathering of fuelwood, severely depleted vegetation in the uncultivated areas of the loess highlands, leaving the steep slopes vulnerable to severe gully and slope erosion by the 1970s and 1980s. Tree planting programs began in the mid-1950s with a more intensive, large-scale afforestation scheme starting in the late 1970s to revegetate the bare hills and minimize soil erosion. However, only about 25% of trees planted during 1950–1980 survived because the chronic water shortage in this area could not support species like the locust tree (*Robinia pseudoacacia* L.), Chinese pine (*Pinus tabulaeformis* Carr.) and sea buckthorn (*Hippophae rhamnoides* L.). Furthermore, trees grow relatively slowly in the semi-arid climate of the north and north-central Loess Plateau. It can take many years before trees produce sufficient leaf surface area to intercept rainfall, reducing its erosive force on the soil surface. Tree roots also protect soil from erosion, since the soil mass attached to tree roots is 13 times greater than the weight of the tree root system (Tanikawa et al., 2021), but it requires time for the tree roots to grow and stabilize the soil mass. Better tree growth occurs on slopes when a check dam is constructed, since the earth dam will prevent trees and shrubs from being destroyed by gully erosion. After installing a check dam and then planting trees, the vegetation cover on two hilly watersheds of the Loess Plateau increased from 1% to 5% to between 40% and 50% from 1955 to 1990 (Wang et al., 2021). The government provides financial compensation for tree planting, which resulted in the reestablishment of forest on about 11–15% of the land in the Loess Plateau from 1986 to 2018 (Fig. 3; Ji et al., 2021).

It is easier to establish perennial grasses and legumes than woody plants in the semi-arid regions of the Loess Plateau. Restored grasslands increased at a rate of about 6700 km² per year from 1986 to 2018, reaching an estimated 43% of the land area by 2018 (Fig. 3; Ji et al., 2021). Grasslands are well-adapted to grow in this area and more effective than other types of vegetation for retaining water and reducing sediment loss. Grasslands restored through natural revegetation are dominated by native plants from the genus *Artemisia*, *Agriophyllum squarrosum* and *Lespedeza davurica*, as well as the grasses *Agropyron cristatum*, *Bothriochloa ischaemum*, *Cleistogenes songorica* and *Stipa* species. Passive grassland restoration occurs when sloping farmland is taken out of cultivation, which improves the soil's resistance to erosion within 15–20 years. For example, soil resistance to slope erosion peaked after 18 years, and this was associated with grassland species that produced more root mass, especially roots that were 0.5–1 mm long, near the soil surface (Guo et al., 2020). At other locations, vulnerable slopes were reseeded with perennial legumes like alfalfa (*Medicago sativa* L.), after excluding grazing animals. However, the large water demand of alfalfa and other introduced plants means that they cannot be sustained in the longer term. The current strategy is to reduce the populations of water-demanding plants by broadcasting seeds of native vegetation and water-saving plants. This should maintain the vegetative cover whilst conserving the scarce soil water resources in the region.

Restoration of the Loess Plateau requires systematic, government-sponsored programs to coordinate land use activities in the area. For example, the 'Grain-for-Green' program that began in 1999 aimed to stop the cultivation of unstable loess highlands, the hilly and plateau-gully lands. The goal of this program was to replant or allow natural revegetation on hilly lands with >25° slope, as well as on degraded cropland and barren land. By the end of 2012, the Grain-for-Green program was credited with converting 9.1 million ha of cropland to forest, and an additional 0.64 million ha of cropland was revegetated as grassland (Fig. 4; Wu et al., 2020). This program is considered a model to improve land use in other parts of China, and worldwide. However, the success of this program required a major shift in government policy. One of the most important policy changes was declaration of the 'household responsibility system', which allowed decades-long leases of the land. Secure land rights gave farmers and herders more incentive to care for the land because a family could expect future benefits from their investment in land improvement. Another key action was to implement a payment scheme to compensate farmers for converting sloping cropland into woodland or grassland. Initially, the Grain-for-Green program promoted afforestation of ecological trees and shrubs, but by 2012 the policy changed to give greater financial compensation to landowners that planted edible tree crops. Consequently, orchards of fruit and nut trees have gradually replaced annual crops and perennial species on the steeper slopes. Today, orchards cover about 15% of the Loess Plateau (Ji et al., 2021), and farmers on the Loess Plateau are the leading producers of apples in China because fruit production is more profitable than grain production in this region.

Deforestation, overgrazing and intensive agriculture were the leading causes of the severe soil degradation and erosion across the Chinese Loess Plateau. The restoration efforts have greatly reduced soil loss from this area, but they have cost millions of dollars. For example, the World Bank and its partners in China invested more than US\$500 million in the Grain-for-Green program. Financial compensation remains an important way to stabilize rural incomes and encourage millions of people to maintain the erosion control structures across this fragile landscape. This example, which is the world's largest erosion control program, demonstrates that humans can reverse negative soil changes due to erosion. While the Grain-for-Green program appears to have been successful for restoring soil productivity, there remain challenges ahead. More water is needed to maintain the current land use. Temperature is rising and rainfall is declining due to global climate change. These climatic changes led to an unexpected ecohydrological outcome, namely a reduction of –0.46 mm per year in streamflow from watersheds across the Loess Plateau (Chen et al., 2020). Therefore, we must continually consider the multiple human-ecological dimensions of soil restoration efforts in our quest to rebuild sustainable soils on the Chinese Loess Plateau.

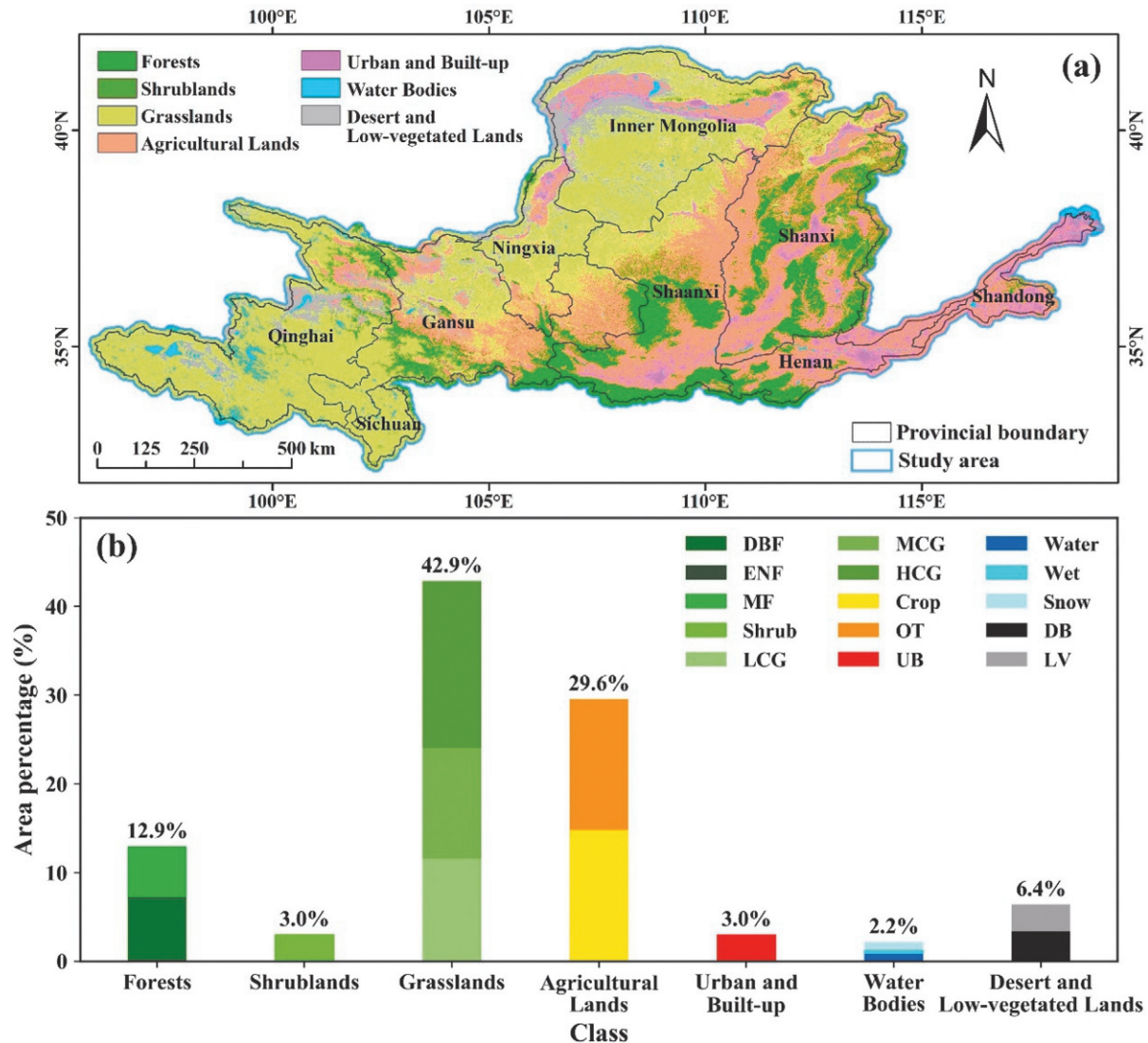


Fig. 3 Spatial distribution of (a) land use/cover of the Yellow River basin in 2018 and (b) the average area percentage of each land use/cover class in 2018, based on Google Earth Engine. In (b), the abbreviations are DBF: deciduous broadleaf forest, ENF: evergreen forest, MF: mixed forest, Shrub: shrublands, LCG: low coverage grasslands, MCG: medium coverage grasslands, HCG: high coverage grasslands, Crop: cultivated cereal cropland, OT: orchards and terraces, UB: urban land, Water: surface water, Wet: wetlands, Snow: snow/ice covered land, DB: bare soil, LV: low-vegetated lands. From Ji Q, Liang W, Fu B, Zhang W, Yan J, Lü Y, Yue C, Jin Z, Lan Z, Li S and Yang P (2021) Mapping land use/cover dynamics of the Yellow River Basin from 1986 to 2018 supported by Google Earth Engine. *Remote Sensing* **13**: 1299. doi: 10.3390/rs13071299.

Soil change due to loss of organic matter

Soil organic matter (SOM) is the litter and debris in soil that comes from plants, animals and microorganisms (Fig. 5). It consists of larger materials that are readily identified as having a biological origin, such as plant residues (leaves, stems, branches and roots) and animal dung, as well as microscopic organic substances that associate with soil minerals in organo-mineral complexes. Organic compounds in organo-mineral complexes are either the decomposition products of plant and animal residues, or secretions and cellular debris produced by plant roots, soil animals and soil microorganisms.

Organic matter for soil stability

The SOM content in grassland and forest ecosystems is maintained in a steady-state by the regular inputs of carbon derived from plant photosynthates, and supplemented by microbially-mediated nitrogen and sulfur fixation, as well as phosphorus and sulfur weathering from parent material. The steady-state is affected somewhat by geological erosion, which moves SOM across the landscape and transfers sediments to downstream environments. However, in most uncultivated ecosystems, the accrual of new organic carbon is equivalent to or greater than the carbon lost during decomposition (e.g., as carbon dioxide respired by soil

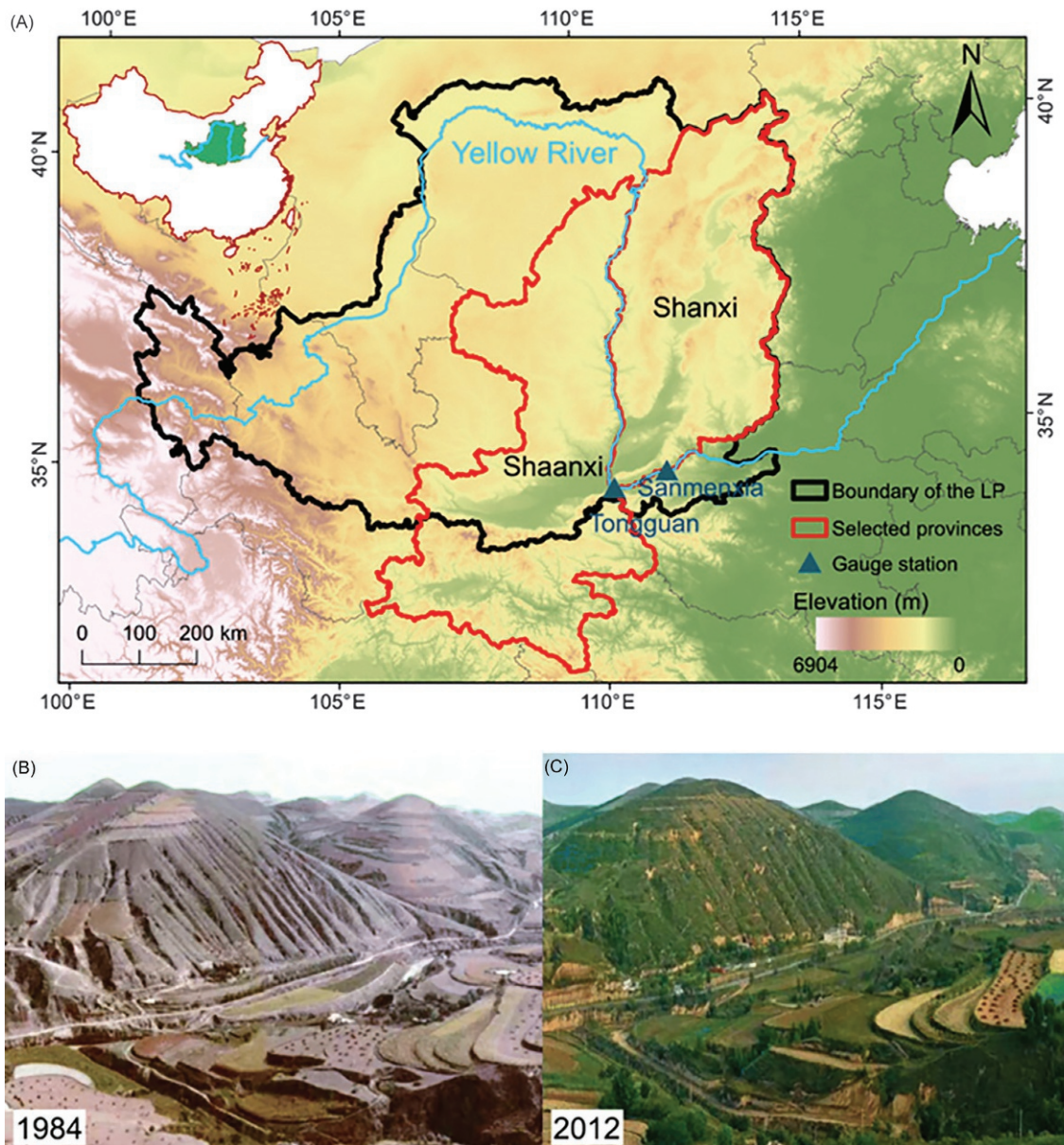


Fig. 4 (a) The boundary of the Chinese Loess Plateau and provinces in this region, with (b) a typical landscape in 1984 and (c) after 12 years of land management under the Grain to Green Program. From Wu XT, Wei YP, Fu BJ, Wang S, Zhao Y and Moran EF (2020) Evolution and effects of the social-ecological system over a millennium in China's Loess Plateau. *Science Advances* 6: eabc0276. doi: 10.1126/sciadv.abc0276.

organisms and dissolved organic carbon that leaches through the soil profile). Organic matter agglomerates with clay and silt minerals to form an aggregate structure. The connected network of macropores between aggregates and micropores within aggregates allow water and dissolved solutes to percolate through the soil matrix. The pore network facilitates air exchange with the atmosphere, creates channels for plant root growth, provides a microhabitat for microbial life and allows soil organisms to move through soil layers, including the surface litter. When SOM is lost following the conversion of forests and grasslands for agriculture, this can drastically affect the soil stability.

Soil change from organic matter loss after deforestation: The case of New Zealand pastures

The islands of New Zealand in the southwest Pacific Ocean were the last large habitable landmass to be settled by humans—first by Māori people around 1300 and later by European settlers in the 1800s. Geographically, the country is dominated by mountains (>1000 m, about 20% of the land surface) and hilly landscapes with steep non-arable hills below 1000 m representing ~40% of the land surface. At least 75% of New Zealand's land surface was forested until about 1200 years ago, but this changed after



Fig. 5 Decomposing litter and organic debris becomes soil organic matter (SOM) in the dark brown topsoil above the mineral subsoil of a Brazilian Oxisol. The SOM contains an abundance of nutrients and stores water for growing plants. The SOM is a component of soil aggregates, which stabilize the soil structure, reducing compaction, erosion and surface crusting, while increasing water infiltration into the soil. Photograph from Washington State Department of Commerce ©2016.

colonization. Hand-felling of forest trees has occurred for centuries, but the rate of deforestation reached a peak around the beginning of the 20th century. After trees were removed, the area was burned and then planted with grass for pastoral use or cropped. Currently, cultivated and pastoral lands represent more than 50% of the total land area (about 14 million hectares) in New Zealand.

In the Waipaoa Basin (North Island, New Zealand), deforestation has significantly modified the landscape in the past 200 years. The European settlers to this area converted ~90% of the native podocarp forest to pasture using repeated slash-and-burn techniques starting in the 1870s. Recently burned land was seeded with ryegrass (*Lolium perenne* L.), clover (*Trifolium repens* L.) and cocksfoot (*Dactylis glomerata* L.) to feed livestock, and systematically overgrazed with cattle and sheep to prevent native vegetation from regrowing. Initially, the land was quite prosperous, supporting 10–12 sheep per hectare. By 1910 there was an increase in soil erosion and declining soil fertility, which required farmers to reduce the sheep stocking density by about 50%. By 1920, the immense soil erosion problems were acknowledged in a New Zealand Geological Survey Bulletin, which warned of the impending danger from landslides in the area (Cervovski-Darriau and Roering, 2016). It took 40–50 years before soil degradation was evident because deforestation practices in Waipaoa Basin removed timber and above-ground components (stems, branches, leaves) but left roots undisturbed. Eventually, the roots of felled trees decayed and SOM levels plummeted. Grazing animals compacted the soil surface, reducing water infiltration into the soil profile. Furthermore, the soil parent material in this region is weakly lithologic sandstone and mudstone that has a natural vulnerability to landslides when rapidly rewetted after dry periods (Fig. 6). Episodic heavy rainfall delivers large volumes of water that flow across the denuded, geologically unstable hilly landscape, resulting in severe soil erosion, gullying and landslides. Consequently, the annual suspended sediment load in the Waipaoa River increased more than six times after European settlement of the region, and was one of the highest sediment yields in the world that is directly attributed to human modification of the landscape (Kettner et al., 2007) (Fig. 7).

In response, watershed managers in the Poverty Bay Catchment Board developed a major flood control and soil conservation scheme in the 1950s that led to dedicated reforestation programs in the upstream and headwater regions of the Waipaoa Basin. This reduced catchment erosion rates of nearly 30 mm per year in the 1950s and 1960s to less than 10 mm per year by 2010 (Cervovski-Darriau and Roering, 2016). Reforestation is an effective strategy for land restoration, but it takes time for trees to establish and stabilize the soil mass, particularly in the landslide-prone areas of the hill slopes where ongoing soil movement often topples or buries saplings before they have a chance to grow. The success of reforestation programs also requires agencies to suitable incentives that encourage local communities to plant trees on their land-holdings. One effective policy has been to exclude grazing animals, especially cattle, from the unstable slopes, which will allow native trees to regrow. For example, the natural regeneration of 10-year-old shrubland of endemic manuka (*Leptospermum scoparium* J.R. Forst & G. Forst) and kanuka (*Kunzea ericoides* (A. Rich.) Joy Thomps. Var. *ericoides*) sustained 65% less damage from landslides than pasture, and there was 90% less soil loss under a 20-year-old stand than under pasture (Bergin et al., 1995).

Reforestation, whether by deliberate tree-planting or natural revegetation, can effectively control soil erosion. The amount of soil stabilization depends upon the stand density of planted or natural regeneration forests, as well as root traits of woody plants and the magnitude of the root–soil adhesion (Fig. 8). An extra benefit of reforestation is the ability of trees to sequester carbon.

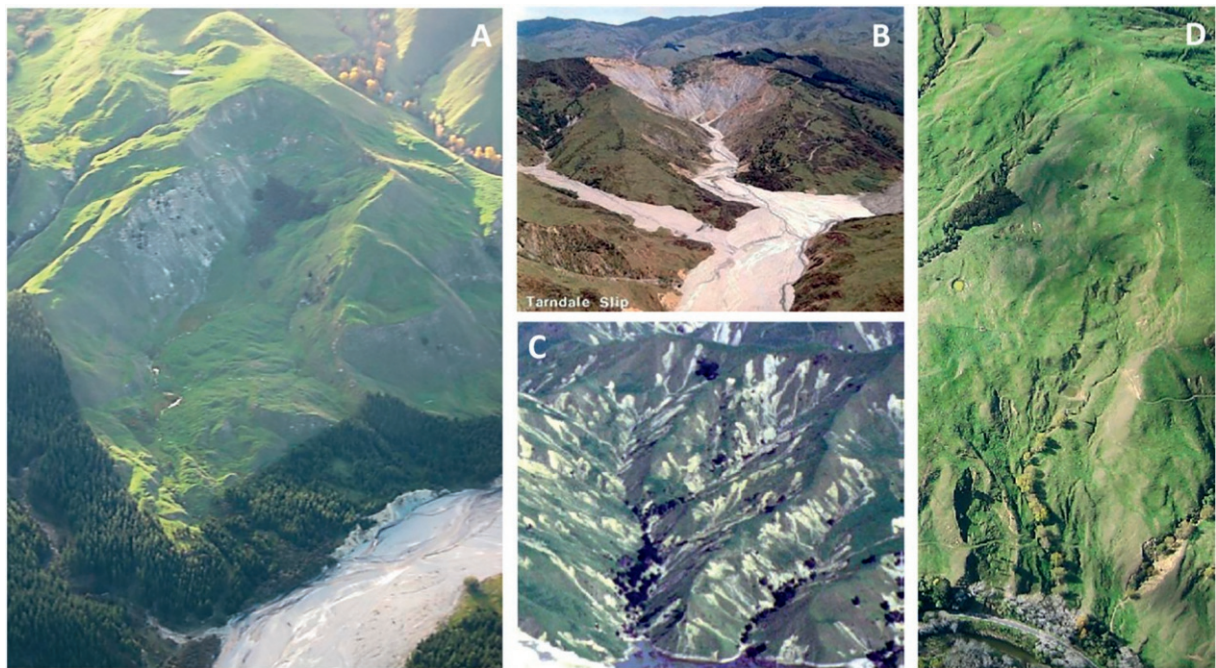


Fig. 6 Photographs of common sediment production processes in the Waipaoa River watershed, New Zealand. (a) Deep-seated landslides often deliver sediment and other terrigenous materials to channels. (b) Many gullies, like the Tarndale Slip, developed after European settlement, and expanded in size as water dislodged sediments from the land surface and carried them downstream. (c) Shallow landslides like these triggered by the 1988 Cyclone Bola mobilize the soil mantle. (d) Slow-moving earthflows impart a fluid-like appearance to hillslopes and generate rapid slope relaxation in response to surface water movement. From Kuehl SA, Alexander CR, Blair NE, Harris CK, Marsaglia KM, Ogston AS, Orpin AR, Roering JJ, Bever AJ, Bilderback EL, Carter L, Cerovski-Darriau C, Childress LB, Corbett DR, Hale RP, Leithold EL, Litchfield N, Moriarty JM, Page MJ, Pierce LER, Upton P and Walsh JP (2016) A source-to-sink perspective of the Waipaoa River margin. *Earth-Science Reviews* 153: 301–334. doi: 10.1016/j.earscirev.2015.10.001.

At least 1.45 million ha of marginal pastoral land has the potential for natural regeneration, which represents a possible carbon gain of about 2.9 million Mg C yr⁻¹ or nearly one-third of the annual energy-related carbon emissions in New Zealand (Trotter et al., 2005). In 2018, the New Zealand government announced an ambitious strategy to plant one billion trees by 2028 to capture carbon dioxide and reach its Paris Agreement targets for 2030 and 2050. Forested lands will be eligible for carbon sequestration payments from New Zealand's Emission Trading Scheme. Conversion of permanent pastures to forests should be a positive step for soil conservation and water protection, and is expected to rebuild the SOM content across hilly landscapes.

Organic matter for soil fertility

Soil organic matter is a well-known source of soil fertility because it contains essential plant nutrients in organic forms—nitrogen (N), phosphorus (P) and sulfur (S) that are hydrolyzed into soluble ions—and it has charged surfaces that retain cations, such as potassium (K⁺), magnesium (Mg²⁺), calcium (Ca²⁺), and anionic forms of boron (B⁻) and molybdate (Mo⁻) that are required for plant nutrition. Ions released or exchanged from SOM are essential for the nutrition of soil microorganisms, plants and animals. Loss of SOM is often associated with low soil fertility, for several reasons. Soils with low organic matter content have smaller reserves of the essential macro- and micro-nutrients. Soil organic matter acts like a sponge for water, and water is often the most limiting factor for plant growth and other soil biological processes. Furthermore, soil aggregate structure is diminished following the loss of SOM, which interferes with biological processes and impedes water and gas exchange between the soil and its surroundings (e.g., the atmosphere and the vadose zone). Intensive cultivation is well-known to deplete the SOM, leading to an overall loss of soil fertility.

Soil change due to organic matter loss following cultivation: The Great Plains prairie grasslands

Prairie grasslands covered more than 1.7 million km² in the Great Plains of central North America before the arrival of European settlers. The underlying black, dark brown and brown soils (Mollisols) contained appreciable quantities of organic matter that accumulated from thousands of years of biomass production in the grassland. A native tallgrass prairie can produce from 5 Mg to 10 Mg shoot biomass ha⁻¹ per year, with an additional 8–15 Mg root biomass ha⁻¹, adding an estimated 4–5 Mg C ha⁻¹ in plant residues to grassland soils each year (Huggins et al., 1998). This is supplemented by decay-resistant charcoal left after grasslands burned through natural fires (e.g., started by lightning strikes) and cultural burning by Native Americans as part of their hunting, animal husbandry and plant cultivation practices. Consequently, soils in the Great Plains have an estimated soil C storage of about 70–130 Mg C ha⁻¹ in the topsoil to 30-cm depth (DeLuca and Zabinski, 2011).

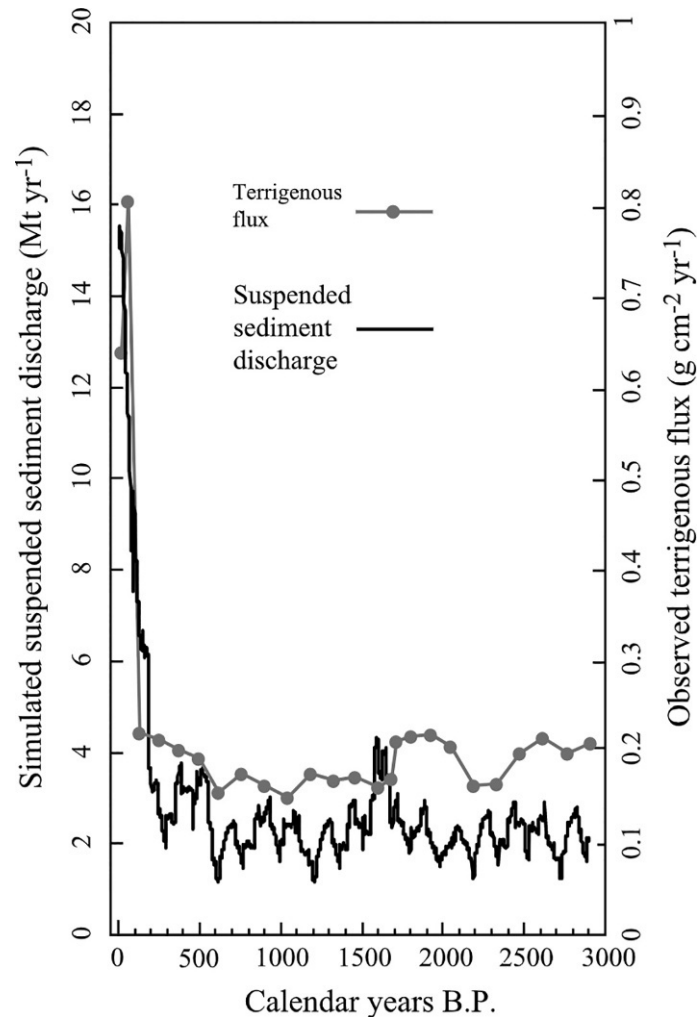


Fig. 7 Widespread land use changes following European deforestation of the Waipaoa Basin (North Island, New Zealand) in the 1870s, for establishing pasture grasses for sheep and other grazing animals, have caused a profound change in the Waipaoa River system. Prior to the 1870s, the annual mass of eroded soil and other terrestrial material (expressed as the terrigenous flux) was around 0.2 g cm^{-2} . Initial land disturbance increased the suspended sediment discharge in the Waipaoa River system by 350%, and this reached 660% once the headwaters were deforested in the 1920s. From Kettner AJ, Gomez B and Syvitski JPM (2007) Modeling suspended sediment discharge from the Waipaoa River system, New Zealand: The last 3000 years. *Water Resources Research* **43**: W07411. doi: 10.1029/2006WR005570.

Government policies in the late 1800s and early 1900s allowed European settlers to establish farms across the region, and was supported by U.S. and Canadian legislation such as the Land Act, the Homestead Act, the Dominion Lands Act and the Swamp Land Act. Settlers cultivated about 1.4 million ha per year of tallgrass prairies in the mid-1800s, with an additional 500,000 ha per year of mixed and short grass prairies in the early 1900s. By the mid-1930s, more than 90% of the original prairie had been converted to cultivated crops, primarily maize (*Zea mays* L.), wheat (*Triticum aestivum* L.), and oats (*Avena sativa* L. 1753). By 2020, less than 5% of the original, undisturbed tallgrass and mixed grass prairies remained in the northeastern Great Plains.

The initial conversion of prairie grassland to agriculture was accompanied by a significant loss of SOM. An estimated 45–70% of the soil organic carbon in the native prairie was lost in the first 25 years of cultivation in Illinois and Missouri (Huggins et al., 1998). The main reason for this is that annual cereal crops produced relatively little root biomass compared to native prairie, and most of the above-ground biomass is harvested. Grain accounts for about 50% of the above-ground biomass of annual cereal crops and is removed from the field at harvest. Often the straw is also harvested for animal bedding, fuel and other purposes. Roots remain in the field, but add a modest amount of organic matter to the soil because annual grain crops produce only about 20% of the plant biomass. This contrasts with native grasses whose roots form 50–70% of the plant biomass and introduced grasses with a root system weighing 35–50% of the plant biomass (Table 3; Wilsey and Polley, 2006). This means that a grain crop producing 10 Mg above-ground biomass ha^{-1} grows 2 Mg root biomass ha^{-1} each year, whereas a tallgrass prairie with 10 Mg above-ground biomass ha^{-1} has $5\text{--}7 \text{ Mg}$ root biomass ha^{-1} . Furthermore, annual cereal crops are photosynthetically active and fix carbon dioxide from the atmosphere for a relatively short growth period, around 100–150 days, whereas a perennial grassland photosynthesizes throughout



Fig. 8 Reforestation of steep slopes in New Zealand with interplanting of native and exotic trees. Photograph by Tree Time Corporation Limited ©2020.

Table 3 Root mass fraction, expressed as the fraction of root mass (shallow and deep roots) in the total plant biomass (roots and shoots) of selected crop and grass species grown in the Great Plains.

<i>Plant species</i>	<i>Root mass fraction</i>
Crops	
Corn (<i>Zea mays</i>)	0.18
Soybean (<i>Glycine max</i>)	0.15
Oats (<i>Avena sativa</i>)	0.40
Barley (<i>Hordeum vulgare</i>)	0.50
Wheat (<i>Triticum</i> spp.)	0.20
Sunflower (<i>Helianthus</i> spp.)	0.06
Permanent hayfield	0.87
Introduced C ₄ grasses	
Yellow bluestem (<i>Bothriochloa ischaemum</i>)	0.20
Kleingrass (<i>Panicum coloratum</i>)	0.23
Dallisgrass (<i>Paspalum dilatatum</i>)	0.26
Native C ₄ grasses	
<i>Schizachyrium scoparium</i>	0.50
<i>Sorghastrum nutans</i>	0.39
<i>Bouteloua curtipendula</i>	0.34
<i>Bothriochloa laguroides</i>	0.25
<i>Sporobolus asper</i>	0.57
Native forbs	
<i>Echinacea purpurea</i>	0.48
<i>Ratibida columnifera</i>	0.19
<i>Salvia azurea</i>	0.49

Adapted from Wilsey BJ and Polley HW (2006) Aboveground productivity and root-shoot allocation differ between native and introduced grass species. *Oecologia* **150**: 300–309. doi: 10.1007/s00442-006-0515-z.

the frost-free period. Consequently, the annual carbon input from plant residues decreased from about 5 Mg C ha⁻¹ under native prairie to less than 1 Mg C ha⁻¹ with conversion to agricultural production (Huggins et al., 1998). This resulted in a loss of nearly 50 Mg C ha⁻¹, representing about 27% of the total carbon in the soil profile to 1.5-m depth (DeLuca and Zabinski, 2011).

The SOM content in cultivated prairies may continue to decline or it may be stabilized, depending on how the land is managed. Declining SOM contents are associated with intensive tillage and frequent fallow periods in the Great Plains region. Intensive tillage refers to a combination of primary (plowing) and secondary (harrowing) operations, which may occur in the spring or fall. Tillage has several advantages because it aerates the soil, prepares the seedbed for planting, disrupts weeds and speeds up soil warming. It also loosens compacted layers and improves water infiltration. The main disadvantage of tillage is that it increases carbon loss due to soil biological respiration by accelerating decomposition of plant residues and disrupting aggregates that contain occluded organic matter. Consequently, the SOM contents tend to be lower in intensively tilled soils than in those with conservation tillage in semi-arid and arid regions. Plowing reduced the soil organic carbon content by 40% in the native prairie and by 27% in a restored prairie (cultivated for 37 years, then replanted with crested wheatgrass, *A. cristatum* (L.) Gaertn., for 11 years). When conservation tillage (no-tillage) occurred on the same sites, the soil organic carbon content declined by 20% in the native prairie and by 4% in the restored prairie (Doran et al., 1998). The benefit of conservation tillage methods such as no-tillage and zone tillage, which disturbs the soil in only a narrow band directly below the crop row, is that these methods preserve the soil structure and leave surface litter as a slowly-decomposing mulch that also retains water.

A fallow period occurs when arable land is not planted with a cash crop during the growing season. Traditionally, the aim is to improve soil fertility and increase soil moisture, because vegetation growing on fallow land is tilled into the soil to retain nutrients and water, rather than exported from the field. However, fallow periods are associated with loss of SOM, particularly when coupled with intensive tillage. Doran et al. (1998) observed that 32% of soil organic carbon was lost from topsoil (0–30-cm depth) after 27 years of winter wheat–fallow cropping in western Nebraska. Modern conservation tillage systems, coupled with greater availability of irrigation in the Great Plains region and the development of drought-tolerant crop varieties, has greatly reduced the need for fallow. From an economic perspective, it is more profitable to grow oilseeds and pulses in rotation with cereal crops in the semi-arid regions of former prairie grasslands. Diversifying the crop sequences in the Great Plains region is expected to improve soil fertility and biological functions, and may replenish the SOM, depending on the crop rotation. Annual crops like lentils (*Lens culinaris* Medik.), field pea (*Pisum sativum* subsp. *Arvense* (L.) Asch.), flax (*Linum usitatissimum* L.) and canola, rapeseed (*Brassica napus* L.) have a limited effect on SOM because of their relatively low residue input. Appreciable gains in soil organic carbon are recorded when perennial crops (+6.2% soil organic carbon) and cover crops (+12.5% soil organic carbon) are included in the crop sequence because these crop species produce more roots and other residues than annual cereals and legumes (King and Blesh, 2018).

Although SOM content can be regenerated by adopting conservation tillage and diverse crop rotations in the former prairie grasslands of the Great Plains, drier hillslopes that erode when cultivated are still vulnerable to SOM loss. Restoring the SOM in erodible areas of the landscape is possible, but the land must be taken out of cultivation and a permanent vegetative cover must be established. The Conservation Reserve Program is a government-subsidized program in the United States that pays landowners to re-vegetate erodible cropland for 10–15 years. One way to achieve this is to restore the native prairie by applying a ‘seed-hay’, the aboveground biomass containing dried seeds from a natural prairie remnant, on the surface of a cultivated field (Fig. 9).



Fig. 9 Prairie strips, the light-green strips in cropland with dark green foliage, are strategically revegetated in perimeter areas of crop fields, eroded hillsides, terrace channels, on pivot corners, along waterways, or in low-lying, waterlogged areas. This is an affordable and environmentally beneficial agricultural conservation practice, supported by the Conservation Reserve Program, that supports biodiversity, regeneration of the soil organic matter, prevents soil erosion and protects water quality. Photograph by Omar de Kok-Mercado, Iowa State University ©2020.



Fig. 10 Riparian buffers are vegetated areas near streams where trees, shrubs and other perennial plants are grown to shade and protect the stream from the impact of adjacent land uses. This is an affordable and environmentally beneficial agricultural conservation practice, supported by the Conservation Reserve Program, that supports biodiversity and prevents soil and streambank erosion, thereby improving water quality in the aquatic environment. Photograph by Seneca Conservation District ©2022.

Prairie restoration effectively increased the soil organic carbon level to 68–74% of the carbon content in the native prairie, whereas the continuously cultivated soil had 50% or less soil organic carbon than the native prairie in Texas (Potter and Derner, 2006). The Conservation Reserve Program also encourages the planting of filter strips, riparian buffers, grassed waterways, field windbreaks, shelterbelts and salt-tolerant vegetation, while the complementary Wetland Reserve Program promotes restoration of wetlands and flood-prone areas (Fig. 10). Replacing cropland with woody species promotes carbon sequestration because carbon accounts for about 45% of the structural components of trees and shrubs. In undisturbed environments, woody species will increase the soil organic carbon, but the carbon gain depends upon the area covered by woody plants, the species planted and disturbances such as fire and grazing that remove part of the woody biomass. Since their inception, the Conservation Reserve Program and Wetland Reserve Program have resulted in the voluntary conversion of approximately 11 million hectares of farmland into conserved lands. This has had the effect of increasing SOM, preventing soil erosion, and protecting vital waterways and wildlife habitat.

Soil change due to element imbalance

Element imbalances are observed in nature from heterogeneities in the geological processes that are responsible for soil formation, coupled with climatic conditions. For example, soils may have large inherent salt concentrations because the parent materials contained salts: calcium, magnesium, sodium, iron (Fe), manganese (Mn) and aluminum (Al). These elements associate with carbonates, bicarbonates, chlorides, sulfates and phosphates, which are dissolved through weathering to release salt ions into the soil pore water. Salinity problems in cultivated soils generally arise because the water source used for irrigation has an elevated salt concentration with an electrical conductivity $>1 \text{ dS m}^{-1}$. Poor drainage of naturally saline soil is another factor contributing to salinity. This occurs when water containing salts moves towards the soil surface, rather than leaching through the soil profile. As the water evaporates, salts are left behind as precipitates and form salt layers and crusts near the soil surface.

Plant growth is suboptimal in saline soils because of osmotic stress. This causes water to move from the weakly saline cytoplasm of plant root cells into the saline soil pore water surrounding the roots. This inhibits the transport of substrates and cofactors into plant cells and leads to the accumulation of reactive oxygen species that cause oxidative stress in the plant. The osmotic effect depends on the dominant cations present, with $\text{Na}^+ > \text{Ca}^{2+} > \text{Mg}^{2+}$ because large Na^+ concentration in cell walls can induce Ca^{2+} and K^+ deficiencies. Osmotic effects are also associated with salt-associated anions and $\text{Cl}^- > \text{SO}_4^{2-}$ because of their potential toxicity and interference with NO_3^- uptake in salt-sensitive plant species. Salt imbalance can be corrected by irrigating with desalinized water and adding fertilizers to balance the plant nutrient requirements, but this increases the crop production costs. Clearly, global change leading to soil salinization is a threat to the sustainability of the world's food supply.

Soil changes due to the accumulation of trace metals that are not essential for plant growth, through natural or anthropogenic processes. Element imbalance is often indicated by the presence of trace metals in edible parts of crops, as well as their biomagnification in food webs, at levels that threaten human or animal health. Some trace metals become elevated in natural

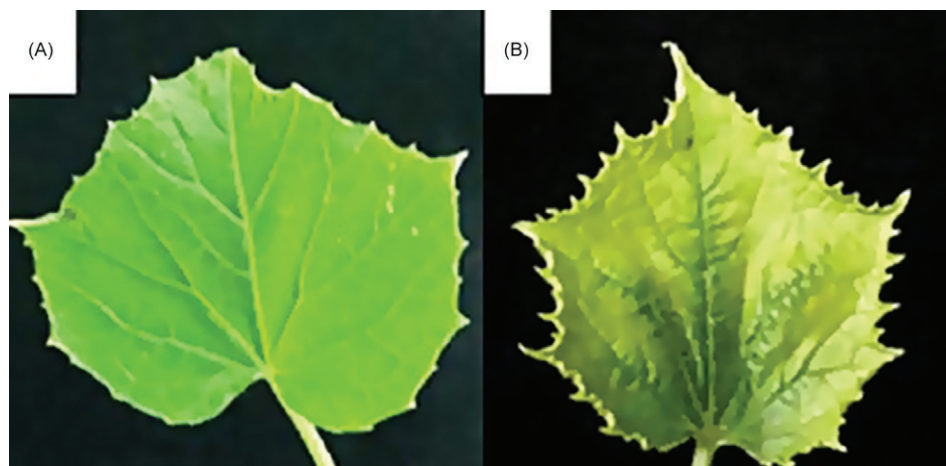


Fig. 11 Element imbalance in cucumber leaves (*Cucumis sativus* L.) exposed to cadmium (Cd) for 20 days. (a) normal leaf with no Cd exposure. (b) chlorotic leaf showing signs of reactive oxygen stress after exposure to 500 μM Cd. From Sun S, Li M, Zuo JH, Jiang WS and Liu DH (2015) Cadmium effects on mineral accumulation, antioxidant defence system and gas exchange in cucumber. *Zemdirbyste-Agriculture* **102**: 193–200. doi: 10.13080/z-a.2015.102.025.

soils due to weathering of enriched parent material. For example, cadmium (Cd) associates with sulfides, carbonates and phosphorites resulting in elevated cadmium concentrations in mafic and ultramafic rocks that contain 0.001–0.60 mg Cd kg⁻¹, as well as black shales with 10–500 mg Cd kg⁻¹ (Page et al., 1987). Geogenic weathering releases a significant amount of cadmium into soil, which readily leaches into groundwater and surface water. Cadmium is a carcinogen for humans and the WHO Guidelines for Drinking-Water Quality recommend a guideline value of 3 μg Cd L⁻¹ in potable water. Anthropogenic sources of cadmium, which is a non-essential element, enter soil through mining, atmospheric deposition of combustion products and application of cadmium-containing fertilizers. This has led to cadmium pollution in soil and groundwater worldwide, exacerbated by the tendency for cadmium to form water-soluble complexes with anions, such as CdCl⁺ and Cd(SO₄)₂²⁻, and associate with dissolved organic matter. Plants accumulate Cd from soil and should be consumed sparingly to avoid negative health outcomes (Fig. 11).

A global assessment of trace metals, also known as toxic elements, reveals more than 5 million sites with contaminated soil. The situation is particularly dire in China where approximately 4 million ha of arable land is moderately or highly contaminated from pollution and more than 25% of the farmland has excessive amounts of cadmium, chromium (Cr), lead (Pb), tin (Sn) and zinc (Zn), leading to a crop production loss of about 10 million Mg per year (Shao et al., 2011). In Europe and the Americas, the cities and industrialized areas are more likely to have soil contaminated with toxic elements from power plants, industries, and heavy traffic. Soil contamination is related to the geographical and socio-economic development, industrial activities, land use, population density and local climate in each area. Several technologies have the potential to decontaminate local soils, depending on the extent of pollution and regulatory limits for each toxic element. Soil washing, electrokinetic remediation, phytoremediation, bioremediation and site stabilization are effective methods of removing toxic elements or retaining them in the soil, to prevent off-site contamination. Control of the anthropogenic sources of metals in the environment, which can be achieved by adopting green chemistry technologies and preventing metal emissions to the atmosphere and waterways, are additional strategies that should reduce soil changes due to pollution in the future.

Element imbalance in cultivated lands is linked to a change in the soil fertility level. It is common to apply agricultural lime, organic amendments and fertilizers to improve soil fertility, since this produces economically profitable yields. However, some amendments are inherently unbalanced in their element composition, making it difficult to choose an application rate that provides the right quantities of macro- and micro-nutrients to match the agronomic crop requirements. This is the case for recycled wastes such as animal manure, municipal wastewater biosolids and incineration ash, which generally have less nitrogen but more phosphorus, nutritive substances and trace elements than required by crops. Annual applications of beef cattle feedlot manure according to local farmer practices provided 1.6–9.4 Mg P ha⁻¹ in a 16-year period, but the barley (*Hordeum vulgare* L.) crop required no more than a tenth of the applied phosphorus (barley harvest removed 0.2–0.6 Mg P ha⁻¹ during the same period; Whalen and Chang, 2001). This resulted in up to 10-fold more plant-available phosphorus concentration in the topsoil of manured soil than unmanured soil, and substantial phosphorus leaching through the soil profile, below the crop root zone. After 25 years of continuous application of beef cattle manure to the same field, there was a linear increase in the soluble salts (Na, K, and Mg) with increasing rates of manure application and more extractable copper and zinc.

Element imbalance in soil is linked to the deterioration of water quality in watersheds, worldwide. Nitrogen that is not used by the crop is lost by leaching, surface runoff and gaseous emissions. Phosphorus leaves the soil through leaching and surface processes (sediment loss and runoff) in quantities that greatly exceed water quality standards. For example, the sediment-associated particulates from agricultural fields in a cold humid region contained as much as 3181 μg P L⁻¹ in surface runoff and 1346 μg P L⁻¹ in surface tile drainage water (Poirier et al., 2012); more than two orders of magnitude above the 30 μg total P L⁻¹ level in eutrophic waters. Eutrophication of aquatic systems that receive excess nitrogen and phosphorus from agricultural runoff is



Fig. 12 Dairy slurry manure applied at a rate that supplies essential plant elements in balance with the needs of the grass-based hayfield, based on the farm nutrient management plan. Judicious use of fertilizer will support crop yield targets while protecting air and water quality. Photograph from Irish Farmers Journal @2018.

associated with cyanobacteria blooms in freshwater systems and coastal estuaries. Blooms of cyanobacteria and other noxious algae are responsible for the closure of recreational fisheries, deteriorating water quality and the loss of biodiversity. Furthermore, hazardous algal blooms produce toxic substances like cyanotoxins that cause acute and chronic health problems in plants, animals and humans, constituting a major risk to public health.

Soil changes due to element imbalance in the essential plant nutrients are detected by regular soil testing, every 3–5 years. To avoid further soil change, which could accelerate nutrient losses into downstream environments and the atmosphere, farmers must follow fertilizer plans that constrain the nutrient inputs to agricultural fields (Fig. 12). Fertilizers must be applied when they are needed by the crop and in close proximity to the roots, for maximum nutrient uptake. Field margins must be stabilized to prevent erosion, trap sediments and retain runoff, often by deliberately planting a buffer with permanent vegetation such as perennial grasses, shrubs and trees. Constructed wetlands and other environmental filters can be built to intercept nutrient-rich water and prevent it from entering downstream waterways. Depending on the jurisdiction, compliance may be achieved through voluntary programs or through regulations and structured policies that limit livestock production and applications of fertilizing materials to vulnerable catchments. The joint efforts of farmers, agricultural planners and watershed managers is essential to prevent soil element imbalances from having a long-term impact on water quality.

Conclusions

Soil change is the result of geological and anthropogenic processes. Soil erosion and SOM loss increase dramatically when undisturbed land is converted for agricultural use. It is possible to halt and reverse soil change by choosing agricultural practices that reduce the frequency and magnitude of disturbances that can destabilize soil structure and decompose SOM. Land forming activities such as terracing and check-dams, coupled with revegetation programs and conservation agriculture practices, have proven an effective strategy to regenerate the soil productivity of fragile landscapes within years to decades. Element imbalances are a soil change associated with human activities such as irrigation, industrial emissions of trace metals and over-application of fertilizing materials. Restoration of soil with an element imbalance will take years, but is possible with improved irrigation practices, nutrient management plans, and application of remediation technologies to contaminated soils. The response of restorative soil interventions to fluctuations in soil temperature and moisture associated with global climate change is difficult to predict and remains a challenge in designing effective strategies to combat soil change. Although soil change is observed at a global scale, solutions must involve local stakeholders who are familiar with and invested in restoring the distinct soil type in their area, considering the topography, climate and goals of the restoration project. Some lands are readily improved after conservation agriculture practices

are implemented, and others require regenerative agriculture solutions. Still other areas should be taken out of cultivation altogether because the land is too fragile to be disturbed, or because it is more valuable as a wildlife or water conservation area. Sustainable land use decisions will determine the future trajectory of soil change, and necessarily involve conversations between landowners, lease-holders, soil specialists, watershed managers and government agencies.

References

- Bergin DO, Kimberley MO, and Marden M (1995) Protective value of regenerating tea tree stands on erosion-prone hill country, east coast, North Island, New Zealand. *New Zealand Journal of Forestry Science* 25: 3–19.
- Cerovski-Darriau C and Roering JJ (2016) Influence of anthropogenic land-use change on hillslope erosion in the Waipaoa River Basin, New Zealand. *Earth Surface Processes and Landforms* 41: 2167–2176. <https://doi.org/10.1002/esp.3969>.
- Chen H, Fleskens L, Baartman J, Wang F, Moolenaar S, and Ritsema C (2020) Impacts of land use change and climatic effects on streamflow in the Chinese Loess Plateau: A meta-analysis. *Science of the Total Environment* 703: 134989. <https://doi.org/10.1016/j.scitotenv.2019.134989>.
- DeLuca TH and Zabinski CA (2011) Prairie ecosystems and the carbon problem. *Frontiers in Ecology and the Environment* 9: 407–413. <https://doi.org/10.1890/100063>.
- Doran JW, Elliott ET, and Paustian K (1998) Soil microbial activity, nitrogen cycling, and long-term changes in organic carbon pools as related to fallow tillage management. *Soil & Tillage Research* 49: 3–18. [https://doi.org/10.1016/S0167-1987\(98\)00150-0](https://doi.org/10.1016/S0167-1987(98)00150-0).
- Douglas I (1989) Land degradation, soil conservation and the sediment load of the Yellow River, China: Review and assessment. *Land Degradation and Development* 1: 141–151. <https://doi.org/10.1002/ldr.3400010206>.
- Guo MM, Wang WL, Kang HL, Yang B, and Li JM (2020) Changes in soil properties and resistance to concentrated flow across a 25-year passive restoration chronosequence of grasslands on the Chinese Loess Plateau. *Restoration Ecology* 28: 104–114. <https://doi.org/10.1111/rec.13057>.
- Huggins DR, Buyanovsky GA, Wagner GH, Brown JR, Darmody RG, Peck TR, Lesoing GW, Vanotti MB, and Bundy LG (1998) Soil organic C in the tallgrass prairie-derived region of the corn belt: Effects of long-term crop management. *Soil & Tillage Research* 47: 219–234. [https://doi.org/10.1016/S0167-1987\(98\)00108-1](https://doi.org/10.1016/S0167-1987(98)00108-1).
- Ji Q, Liang W, Fu B, Zhang W, Yan J, Lü Y, Yue C, Jin Z, Lan Z, Li S, and Yang P (2021) Mapping land use/cover dynamics of the Yellow River Basin from 1986 to 2018 supported by Google Earth Engine. *Remote Sensing* 13: 1299. <https://doi.org/10.3390/rs13071299>.
- Kettner AJ, Gomez B, and Syvitski JPM (2007) Modeling suspended sediment discharge from the Waipaoa River system, New Zealand: The last 3000 years. *Water Resources Research* 43: W07411. <https://doi.org/10.1029/2006WR005570>.
- King AE and Blesh J (2018) Crop rotations for increased soil carbon: Perenniality as a guiding principle. *Ecological Applications* 28: 249–261. <https://doi.org/10.1002/eap.1648>.
- Liu XH, He BL, Li ZX, Zhang JL, Wang L, and Wang Z (2011) Influence of land terracing on agricultural and ecological environment in the loess plateau regions of China. *Environmental Earth Sciences* 62: 797–807. <https://doi.org/10.1007/s12665-010-0567-6>.
- Montgomery DR (2007) Soil erosion and agricultural sustainability. *Proceedings of the National Academy of Sciences of the United States of America* 104: 13268–13272. <https://doi.org/10.1073/pnas.0611508104>.
- Page AL, Chang AC, and El-Amamy M (1987) Cadmium levels in soils and crops in the United States. In: Hutchinson TC and Meema KM (eds.) *Lead, Mercury, Cadmium and Arsenic in the Environment*, pp. 119–146. Chichester: John Wiley & Sons Ltd.
- Poirier S-C, Michaud AR, and Whalen JK (2012) Bioavailable phosphorus in fine-sized sediments transported from agricultural fields. *Soil Science Society of America Journal* 76: 258–267. <https://doi.org/10.2136/sssaj2010.0441>.
- Potter KN and Derner JD (2006) Soil carbon pools in central Texas: Prairies, restored grasslands, and croplands. *Journal of Soil and Water Conservation* 61: 124–128.
- Shao H, Chu L-Y, Xu G, Yan K, Zhang L-H, and Sun J (2011) Progress in phytoremediating heavy-metal contaminated soils. In: Sherameti I and Varma A (eds.) *Detoxification of Heavy Metals. Soil Biology*, vol. 30, pp. 73–90. Berlin: Springer. https://doi.org/10.1007/978-3-642-21408-0_4.
- Shi P, Feng ZH, Gao HD, Li P, Zhang XM, Zhu TT, Li ZB, Xu GC, Ren ZP, and Xiao L (2020) Has “Grain for Green” threaten food security on the Loess Plateau of China? *Ecosystem Health and Sustainability* 6: 1709560. <https://doi.org/10.1080/20964129.2019.1709560>.
- Tanikawa T, Ikeno H, Todo C, Yamase K, Ohashi M, Okamoto T, Mizoguchi T, Nakao K, Kaneko S, Torii A, Inagaki Y, Nakanishi A, and Hirano Y (2021) A quantitative evaluation of soil mass held by tree roots. *Trees - Structure and Function* 35: 527–541. <https://doi.org/10.1007/s00468-020-02054-y>.
- Trotter C, Tate K, Scott N, Townsend J, Wilde H, Lambie S, Marden M, and Pinkney T (2005) Afforestation/reforestation of New Zealand marginal pasture lands by indigenous shrublands: The potential for Kyoto forest sinks. *Annals of Forest Science* 62: 865–871. <https://doi.org/10.1051/forest:2005077>.
- Wang ZY, Chen ZY, Yu S, Zhang Q, Wang Y, and Hao JW (2021) Erosion-control mechanism of sediment check dams on the Loess Plateau. *International Journal of Sediment Research* 36: 668–677. <https://doi.org/10.1016/j.ijsrc.2021.02.002>.
- Wei W, Chen LD, Zhang HD, Yang L, Yu Y, and Chen J (2014) Effects of crop rotation and rainfall on water erosion on a gentle slope in the hilly loess area, China. *Catena* 123: 205–214. <https://doi.org/10.1016/j.catena.2014.08.002>.
- Whalen JK and Chang C (2001) Phosphorus accumulation in cultivated soils from long-term annual applications of cattle feedlot manure. *Journal of Environmental Quality* 30: 229–237. <https://doi.org/10.2134/jeq2001.301229x>.
- Wilsey BJ and Polley HW (2006) Aboveground productivity and root-shoot allocation differ between native and introduced grass species. *Oecologia* 150: 300–309. <https://doi.org/10.1007/s00442-006-0515-z>.
- Wu XT, Wei YP, Fu BJ, Wang S, Zhao Y, and Moran EF (2020) Evolution and effects of the social-ecological system over a millennium in China's Loess Plateau. *Science Advances* 6: eabc0276. <https://doi.org/10.1126/sciadv.abc0276>.
- Xu XZ, Li MJ, Liu B, Kuang SF, and Xu SG (2012) Quantifying the effects of conservation practices on soil, water, and nutrients in the Loess Mesa Ravine Region of the Loess Plateau, China. *Environmental Management* 49: 1092–1101. <https://doi.org/10.1007/s00267-012-9835-4>.
- Zhao JB, Luo XQ, Ma YD, Shao TJ, and Yue YL (2017) Soil characteristics and new formation model of loess on the Chinese Loess Plateau. *Geosciences Journal* 21: 607–616. <https://doi.org/10.1007/s12303-016-0069-y>.

The effect of global change on the soil body

Sophie Cornu^a, David Montagne^b, and Antra Boca^c, ^aAix Marseille University, CNRS, IRD, INRA, Coll France, CEREGE, Aix-en-Provence, France; ^bUniversité Paris-Saclay, INRAE, AgroParisTech, UMR ECOSYS, Palaiseau, France; ^cLatvia University of Life Sciences and Technologies, Jelgava, Latvia

© 2023 Elsevier Ltd. All rights reserved.

Introduction	360
Global change, the dynamics and impact on the fluxes of matter and energy responsible for soil evolution	361
Global change and its dynamics	361
Impact of global change on soil stocks, structure and fluxes	363
Anticipated soil changes due to global change	365
Soil resilience and vulnerability to global change	365
Impact of global change on soil: Synergy or antagonism?	366
Assessing soil change under global change, a scientific challenge	366
Concluding remarks	367
References	368

Abstract

Without soil, no life on land is possible because soil provides the means for plant growth and plays a crucial role in water and carbon cycling among others. Human-induced global changes related to climate, land use and management are altering soils. The concepts of ecosystem services and soil capital value them, their functions and prevent their degradation. A major challenge for soil science is to quantify the relationships better between global changes, and soil processes and properties. The aim is to predict soil changes and understand whether soils will continue to deliver the services required in the long-term and to what extent soil management can mitigate climate change.

Key points

- Soil is a dynamic system characterized by feedback loops, linking external forcing variables (or factors of pedogenesis), pedological processes, soil properties and ultimately soil functions and services.
- As a result of the complex interactions between climate, land-use and or management changes acting on various spatial and time scales, the direction, intensity and spatio-temporal pattern of future soil changes remain largely unknown.
- First indications of the effect of global change on a range of soil types, processes and properties have been observed already.
- In the future, effects of global change on soils will largely exceed well identified situations such as the reduction of Cryosols, soil degradation processes (e.g. erosion or desertification) or manageable soil properties (e.g. soil organic carbon stocks).
- With the irreversibility of numerous soil processes, a changed climate that can increasingly remove primary minerals, clay particles and base cations from soil threatens soil fertility, quality and health.
- If land-use or land management changes could be used theoretically to counteract climate change, their practical implementation at the pace and spatial scales of climate change is doubtful and could have counterproductive effects.
- Soil management must be planned and implemented to limit the risk of mutual reinforcement of climate and land-use changes that are already inducing cataclysmic soil degradation under extreme weather events.
- To face these arising and mostly inevitable challenges, there is an urgent need to develop tools and methodologies able to capture the new evolution of soil under global changes.
- The feedback loops between external forcing variables and soil properties are currently insufficiently considered in global modelling approaches, in which the soil compartment is generally poorly described.

Introduction

Soils are open systems produced by a complex combination of processes of input, transformation, transfer and export of materials and energy. Their occurrence and intensity are a function of external environmental conditions, otherwise called factors of pedogenesis. These factors include the parent material, climate, relief, living beings and time.

Under the control of these external environmental conditions (factors of pedogenesis), the physical, chemical or biological processes transform and structure the upper layers of the Earth's crust, i.e. the parent material of the soils, whose initial properties are gradually replaced by more stable constituents and organization. The emergence of these new properties is the basis for the ecosystem functions and services that soils provide to humanity. They also affect how the soil responds to external environmental conditions. For example, the physical fragmentation of a hard rock will gradually favor the infiltration of water into the soil and

modify the pedoclimatic conditions; the redistribution of soil particles by erosion will modify the relief; the formation and deepening of the soil will allow its colonization by new plant or animal species which will, in turn, participate in the processes of alteration of the geological substrate. The soil is therefore a dynamic system characterized by feedback loops, linking external forcing variables (or factors of pedogenesis), pedological processes and soil properties. Thus, any modification of the external forcing variables is likely to modify some pedological processes and properties, which in turn will change how the external variables affect the soil depending on thresholds, positive or negative feedbacks or even hysteresis.

From the pedological factors, two – climate and vegetation – have constantly changed through time, notably over the most recent period, in which humans have strongly affected both and will continue to do so in the near future. These changes, also known as global change, have been shown to affect a range of soil properties with some changing rapidly and others gradually (Rounsevell et al., 1999). Agriculture, for example, affects soil rapidly, at a decadal to century time scale (Cornu et al., 2012; Guo et al., 2010; Montagne et al., 2016) and, in many cases, in an irreversible way (Sombroek, 1990; Rounsevell et al., 1999; Montagne et al., 2016). The effects of climate change on soils are more difficult to identify and quantify. However, its impacts may have serious consequences for soil productivity and sustainability of agriculture and forestry as well as the stability of our infrastructure. The changes induced in the soil may in turn act on land use through changes in fertility or on climate via the release of greenhouse gasses. These feedback loops are currently not considered sufficiently in global modelling approaches, in which the soil compartment is generally poorly described.

Therefore, quantifying the impact of global change on soils is a challenge that must be addressed. A major difficulty in addressing this is the large uncertainty about the kinetics of natural soil processes. Consequently, the direction, intensity and spatial distribution of change in the soil under global changes, and their consequences on the delivery of soil services and human well-being still remain largely unknown.

Global change, the dynamics and impact on the fluxes of matter and energy responsible for soil evolution

Global change and its dynamics

The concept of global change, defined in the 1970s, considers human activities as part of the planetary system together with atmosphere, biosphere, hydrosphere and the geosphere. It can be divided in two main types of change: climate change and land use changes.

Climate change caused by anthropogenic emissions of greenhouse gasses is widely recognized, and has resulted in a noticeable increase of the Earth's surface temperature of 0.85 °C since 1880 (Fig. 1). This increase is responsible for the increasing temperature of permafrost areas since the 1980s and a decrease in ice mass and snow cover in most regions, with noticeable consequences for the hydrological regime. In addition, an increase in rainfall has been observed since the beginning of the 20th century in the Northern hemisphere, but with no clear change recorded elsewhere. A change in the frequency of extreme events has been recorded with a decrease in cold extreme events, and an increase in hot extreme events and strong rainfall events (Jia et al., 2019). Regardless of the

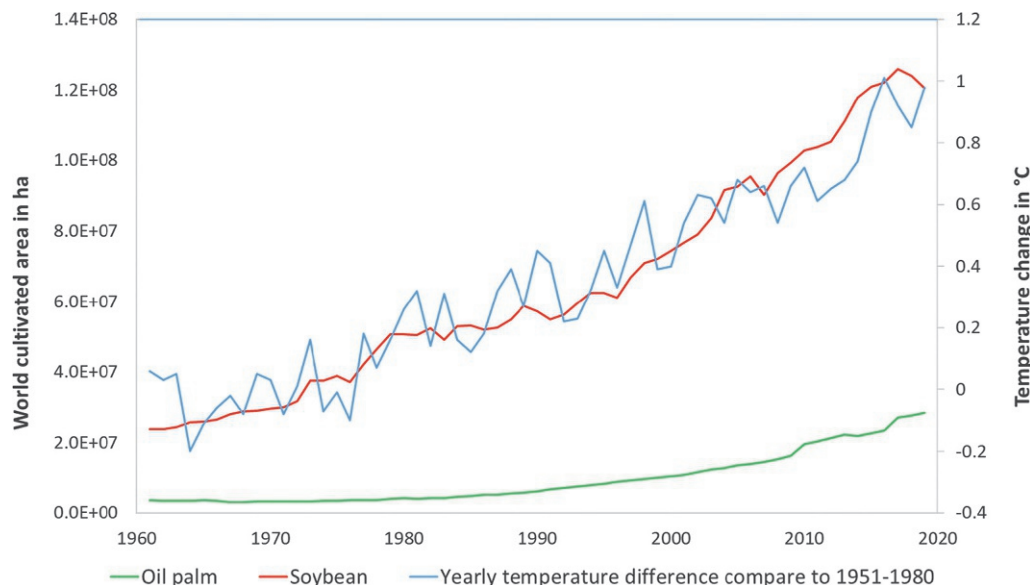


Fig. 1 Climate change represented by the relative temperature increase 1951–1980 (NASA data: https://data.giss.nasa.gov/gistemp/graphs_v4/) and global land-use change represented by the surface area converted to soybean and palm oil plantations (FAO data: <https://www.fao.org/faostat/en/#home>) between 1961 and 2020.

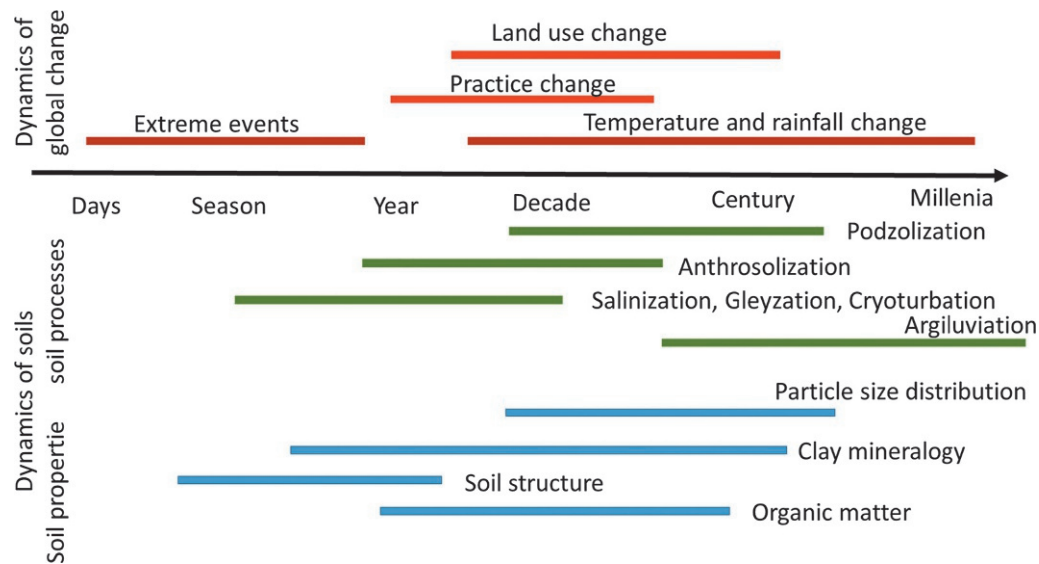


Fig. 2 Global change and examples of time steps in soil change (days to millenia).

climate scenario considered, these trends are predicted to continue over the 21st century together with a decrease of rainfall in the arid medium latitudes (Hartmann et al., 2013). An increase in the speed of such progressive changes or even abrupt changes in temperature or precipitation pattern, however, cannot be completely excluded, at least locally, as a result of the atmospheric CO₂ (carbon dioxide) accumulation or climatic break points such as the melting of polar ice caps or changes in ocean global circulations. Gradual changes in temperature and precipitation will induce slow, low amplitude and continuous modifications in soils, whereas extreme events will induce rapid, large amplitude and irregular changes, all of which will accumulate over time (Fig. 2).

Land cover is the biophysical state of the Earth's surface and immediate subsurface. *Land use* involves both the manner in which the biophysical attributes of the land are manipulated and the purpose for which it is used. In other words, land use can be described by the terms 'forestry, parks, farmlands, etc.' as well as by biophysical manipulations such as the use of fertilizers, logging, change of forest species composition, etc. Purely natural disturbances of the land cover, e.g., wildfires have caused large-scale changes like the loss of tree canopy and gain in short vegetation in large patches of boreal Canada, eastern Alaska and central Siberia (Song et al., 2018). However, a much stronger force for change is from human impact. Land-use change is traditionally a local-scale human practice with an annual time step (e.g., deforestation of a plot), its evolution at the global scale has however a decadal to century time step (Fig. 2).

Today, as much as 50% of the Earth's ice-free land surface has been transformed, and virtually all land has been affected in some way by such processes as co-adapted landscapes, climate change, and tropospheric pollution (Turner et al., 2007). The *Millennium Ecosystem Assessment* (2005) states that, while these changes in land uses are not recent, they have been more rapid and extensive since the 1950s as illustrated by the exponential increase in surface area cultivated with soybean (*Glycine max* (L.) Merrill) and palm trees (Arecaceae) during that period (Fig. 1). Nevertheless, a global increase in tree cover has also been observed from 1982 to 2016 together with a reduction in bare soil. The net gain in tree cover was mostly due to increases in the continental and polar regions, and across all climate domains for montane systems (Song et al., 2018). This increase counteracted losses of tree cover in the tropics and in many arid and semi-arid ecosystems where increasing conversion to intensive agriculture is occurring. In northern temperate countries, the surface area dedicated to agriculture has tended to decline while the intensification of agriculture continues to produce more food on smaller areas of land for an increasing human population, and a new challenge of biomass production to substitute fossil fuels emerges. On a positive note, new agricultural diversification practices are emerging worldwide (crop and non-crop rotation, and diversification, various forms of conservation agriculture, organic farming, etc.), many of which show a win-win situation in the relationship between crop yields and multiple ecosystem services (Tamburini et al., 2020). The area under alternative farming systems is still small (~72.3 million hectares of the ~5 billion hectares of agricultural land worldwide are under organic farming) but steadily increasing (FAOSTAT data). In Europe, for example, the area under organic farming currently constitutes around 9% of all agricultural land (Eurostat data).

Overall, land use and land cover changes are considered to be the major causes and consequences of global change on land. Song et al. (2018) estimated that from 1982 to 2016, of all land changes, 60% were associated with direct human activities and 40% with indirect factors such as climate change (e.g., woody vegetation growth in the Arctic). Land use and land cover changes are estimated to have contributed 25% of cumulative carbon emissions to the atmosphere since industrialization (Friedlingstein et al., 2020). Conversely, climate change affects cultivation in many regions of the world (Jia et al., 2019). Indeed, the continuing increase in

global temperature, and the shifts in precipitation patterns in turn affect the net primary production and biome distribution. These complex interactions between climate and land-use changes lead to a complex spatial and temporal distribution of their impacts on the environment at the global scale. Both changes acting synergistically or antagonistically on net primary production, decomposition, leaching and other soil forming factors and processes, will undoubtedly alter the development of soils in ways that are difficult to predict.

Impact of global change on soil stocks, structure and fluxes

Soil is made up of organic and mineral solid phases (or soil stocks) that are spatially organized (soil structure) and form pores, within which water, gas and heat transfers take place. It is also the site of significant biodiversity, which interacts with the soil's solid phases and their structure. Both, climate and land use changes can act directly or indirectly on these soil stocks, structure and fluxes.

By increasing air temperature and modifying precipitation, climate change strongly affects the soil water regime and plant growth, and thus acts indirectly on the biologically mediated soil processes. An increase in annual precipitation leads directly to increased runoff and infiltration, while increases in temperature affect evaporation and transpiration. Temperature and water supply influence solute chemistry and rates of weathering through thermodynamic and kinetic interactions between minerals and solutions, with warmer and wetter environments experiencing higher rates of weathering (Rounsevell et al., 1999). Both air temperature and precipitation have a direct effect on organic matter turnover. Climate, through its control on plant community distribution, has an indirect effect on the quantity and quality of organic matter inputs into soil (Fig. 3).

As mentioned above, land-use change includes a change in vegetation (land cover) as well as a large variety of management practices from tillage, irrigation, fertilization and amendments to more intrusive modifications such as artificial drainage or soil

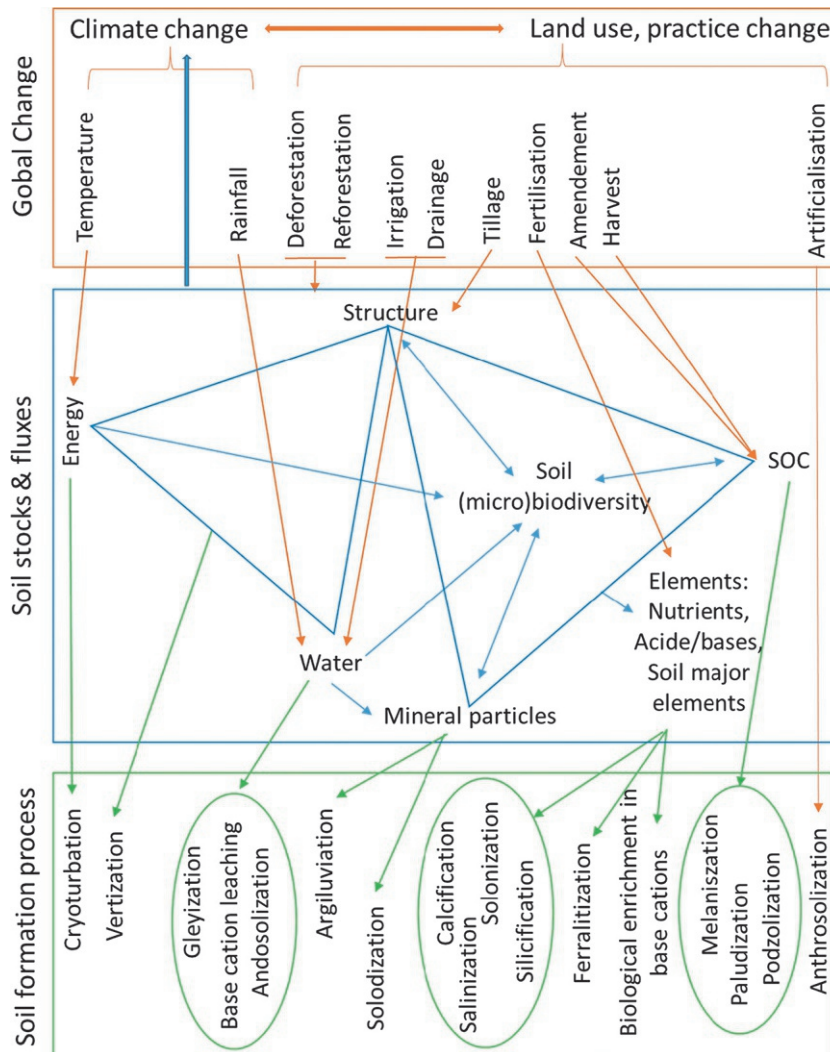


Fig. 3 Impact of global change on soil stocks (including soil biodiversity), structure and fluxes of matter and energy in soils, and their consequences for the evolution of soil processes.

artificialization (mainly urbanization, mining, industrialization and traffic), and soil construction. Large gaps in knowledge related to the effects of land management persist, in particular for human-induced changes in terrestrial ecosystems that do not result in conversions of land cover. The impact of land use on soil stocks, organization and fluxes is therefore much more diverse than that of climate change and cannot be described in detail here mainly because of the current lack of understanding on this topic. Nevertheless, three main categories can be identified: i) practices that act mainly on soil stocks like fertilization, amendments and harvest; ii) practices that act mainly on soil organization and fluxes like irrigation, drainage and tillage, and; iii) land cover change (forestation, deforestation, artificialization, etc.) that act on soil stocks, organization and fluxes.

Changes in land cover induce direct changes in biotic diversity, actual and potential primary productivity (and therefore biogeochemical cycling), runoff and rates of erosion. The increase in tree cover and reduction of bare soil described above (Song et al., 2018) suggest that globally we could expect an increased accumulation of soil organic carbon (SOC) and potentially an increase in weathering with greater root activity. However, this net gain results from a net loss in the tropics that is outweighed by a net gain in the non-tropics. Thus, in the tropics, where deforestation and conversion of forests to farmland are prominent, a loss of SOC and soil by erosion (a destructive process) can be expected (Fig. 4). Tree species selection can also affect various chemical and physical soil properties such as the interception of precipitation (less for evergreen species), water repellency of the forest floor (higher for coniferous species especially under dry conditions), mineralization and accumulation of organic matter and nutrients, soil pH (generally lower for conifers) and finally mineral weathering (Fig. 3). Overall, the effect of tree species selection on soil properties depends on other environmental variables, e.g., climate and soil mineralogy (Augusto et al., 2014).

Soil artificialization brings elements to the soil by atmospheric deposition (airborne deposits) or by allochthonous material of natural or artificial (rubbles notably) origin. This results most often in changes in soil pH, increases in stoniness and decreases in SOC compared to natural soils, and in some cases – in the formation of a totally new soil type – Anthrosol (Cornu et al., 2021) (Fig. 3).

Fertilization, amendments and harvest mainly act on soil stocks. In forests, harvest reduces standing biomass and soil C stocks, especially when wood residues are removed. The effect of crop harvest is mixed, and often depends on the presence and intensity of pre- and post-harvest management practices (Singh et al., 2018). Fertilization is designed to increase plant productivity, which in turn increases C-inputs to the soil and ultimately the exchange of C between land and the atmosphere. In addition, N fertilization of croplands can lead to soil acidification, when not countered by liming, as demonstrated in Chinese soils, which experienced a 0.2- to 0.3-unit pH decrease in 20 years (Guo et al., 2010). Potassium and N fertilization have been shown to influence exchanges between the soil solution and the soil mineral phase with significant effects on clay mineralogy (Matocha et al., 2016).

Tillage, irrigation and drainage act mainly on the soil structure and fluxes that in turn modify soil stocks, including soil biodiversity. For example, tillage, designed to maintain soil porosity and control weeds, directly affects water fluxes in soils but it also has numerous indirect effects. It reduces soil biodiversity and modifies its composition, for example, the bacteria to fungi ratio or the relative abundances of epigeic, endogeic and anecic earthworms through various mechanisms among which the burying of soil litter is of primary importance. Plowing disrupts aggregate structure, aerating the soil, activating microbial decomposition, releasing nutrients from organic matter and thus depleting SOC. Some evidence indicates that for most soil types adoption of

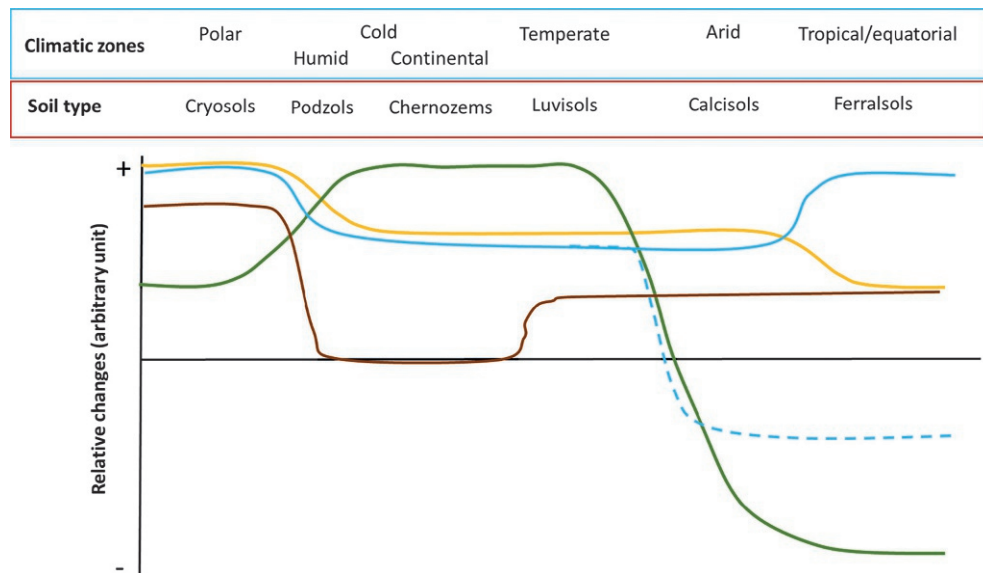


Fig. 4 Climate (temperature and rainfall) and trends in land-use change for the main Köppen-Geiger climatic zones and resulting vulnerability to global change of the dominant soils. Blue represents rainfall, yellow temperature, green forest area, brown soil vulnerability. In arid and tropical zones the solid blue line describes precipitation in Africa and the dashed line South America. Increasing soil vulnerability is due to thawing in polar areas, weathering and leaching of base cations in the temperate zone, salinization or erosion in the arid zone and erosion after deforestation in tropical areas. In the cold zone, the limited soil vulnerability is due to either an advanced stage of weathering (Podzols, World Reference Base; Spodosols, Soil Taxonomy) or a strong buffering capacity (Chernozems, WRB; Mollisols, Soil Taxonomy).

no-tillage practices after tillage-based management does not significantly increase SOC stocks (Powlson et al., 2014), but changes its depth distribution. No-tillage does, however, generally reduce soil erosion.

Irrigation significantly enhances plant productivity where water limits plant growth with variable secondary consequences on SOC inputs and storage. By increasing water inputs, irrigation increases soil moisture, water fluxes through soil, and can generate anoxic soil conditions when water is in excess. Dissolved elements or particulate matter are brought into soil with irrigation water, some of them being quite relevant in soil development such as salts or contaminants, which can lead to soil salinization.

Similar to irrigation, soil drainage primarily affects water fluxes in soils because it is intended to eliminate excess water to improve soil trafficability and productivity. However, by decreasing soil moisture and promoting better aeration than under conditions that formerly experienced anoxic periods, drainage tends to favor SOC depletion rather than SOC storage, especially in organic soils. By removing excess water from an initially largely closed soil system, drainage increases flows of water and dissolved and particulate matter through and out of soils.

Anticipated soil changes due to global change

Soil resilience and vulnerability to global change

At the global scale, climate is the main factor for soil formation. There is a close dependence between soil processes and properties with soil temperature and soil moisture. Five main zones are commonly differentiated from polar to equatorial climates on the basis of the current global soil distribution (Fig. 4). Soils from very cold climates are characterized by water that occurs primarily as ice and by cryoturbation processes. Cold and humid environments are dominated by the podsolization process, in which the degradation of soil organic material is hindered by cold temperatures, and the percolation of organic acids completely leaches base cations from the soil and dissolves primary weatherable minerals. Toward the mid latitudes, with higher soil temperatures and precipitation, moderate leaching of base cations and transport of particulate matter dominate soil formation (argiluviation, solodization, Fig. 3). The most humid and hot climates of the tropical zones are characterized by ferrallitization involving fast mineralization of soil organic matter, almost complete chemical weathering of primary minerals, strong leaching of solubilized base cations and silicon (Si), and the accumulation of secondary kaolinite, iron (Fe) or aluminium (Al) oxides. In contrast to the humid climates of boreal, temperate and tropical zones, in which the soil formation is dominated by export of materials, that under drier climates involves the accumulation of organic matter (melanization), secondary carbonates and gypsum (calcification), secondary Si (silicification) or various kind of soluble salts (salinization). Despite early recognition of the importance of climate on soil formation, it has in general been ignored in climate change research with the exception of some rare, mostly qualitative, and descriptive studies (Sombroek, 1990; Rounsevell et al., 1999).

Nevertheless, several consequences of climate change may be inferred from the zonal distribution of soil such as: (i) a marked reduction of cryoturbation processes in the current polar climatic zone (Meredith et al., 2019; Fig. 4); (ii) an increase in organic matter turnover and leaching of base cations in either particulate or dissolved forms from the polar to humid cold and temperate zones where temperatures and precipitation are predicted to increase (Fig. 4); and (iii) decreasing organic matter inputs into soils and increasing accumulation of soluble salts of Na (sodium), Ca (calcium), Mg (magnesium) and K (potassium) in arid mid-latitude and subtropical zones where temperatures and precipitation are predicted to increase and decrease, respectively (Fig. 4). The accumulation and leaching of base cations are both theoretically possible under climate change. However, losses in base cations are almost irreversible, therefore, climate change is more likely to induce losses of basic cations and reduce soil fertility than the reverse. Enrichment in basic cations and subsequent improvement in soil fertility may be expected only in the case of soil rejuvenation, if unweathered materials are newly exposed at the soil surface. Such massive rejuvenation has occurred in the past, for example, in northern temperate areas after the last deglaciation, or in tropical areas after aridification.

At the extremes of the climate gradients, where climate change is most severe, thawing of Cryosols (World Reference Base; Gelisols in Soil Taxonomy) in polar areas and desertification in semi-arid to sub-tropical areas have already been observed. In the mid-latitudes, vegetation shifts to the north have been detected, as well as increases in surface and subsurface soil temperature in continental Europe (Khudiyakov and Reshotkin, 2017). Such changes in the most dynamic soil properties are probably the first recordable signals of deeper changes affecting not only soils at the extremes of the climate gradient but also soils under all types of climates. The speed of these changes is, however, poorly understood.

In soils that will experience increased weathering of primary minerals or leaching of base cations, three main situations can be distinguished. Soils already completely depleted in primary minerals and base cations (e.g. Podzol and Ferralsol, World Reference Base; Spodosol and Oxisol, Soil Taxonomy) are likely to be affected little by climate change when not subjected to erosion induced by land-use change (Fig. 4). Soils rich in carbonates, clay minerals, or soil organic matter (e.g. Chernozem, World Reference Base; Mollisol, Soil taxonomy), which act as buffers for soil processes such as argiluviation, podzolization or chemical weathering, should be resilient to climate change (Fig. 4) provided that their stocks of these elements are not totally depleted. Exhaustion of these stocks would represent a clear threshold for the future evolution of these soils. The kinetics of dissolution of carbonates or clay minerals, or turnover of soil organic matter are key properties to consider in their evolution. For soils already poor in these constituents, the buffering capacity of the soil will be reduced quickly and the threshold reached rapidly. These soils are expected to have a limited resilience to climate change (Fig. 4). This is especially the case for large areas of non-calcaric, clay-poor Cambisols (Inceptisols) and Luvisols (Alfisols) (World Reference Base; Soil Taxonomy) in the cool, northern regions, which are among the most productive soils

on Earth. Currently their development is limited by their young age and cool climates, but it could be significantly and very rapidly affected by increasing temperatures and precipitation.

Human effects on pedogenesis are mainly discussed from a theoretical point of view, and quantification of the consequences on soil evolution remain fragmented. In the body of research studying human impact on soils, most deal with: (i) soil degradation processes such as erosion, salinization and desertification, or soil contamination that are directly and strongly linked to human practices; (ii) a single soil property taken individually such as soil porosity, soil water retention, soil organic carbon or earthworm populations to name just a few; and finally, (iii) time scales rarely exceeding a few years (Richter et al., 2007). With its pivotal role on climate change and soil functions, changes in soil organic carbon contents have been widely studied, including time spans ranging from decades to centuries. It has recently been demonstrated, however, that human activities do not only affect soil organic carbon contents and related soil degradation processes, but also most other soil properties. Among these properties, carbonate or clay contents are of particular interest because of their capacity to buffer soil changes (Montagne et al., 2016). Furthermore, a large range of soil processes from podzolization to argiluviation in humid temperate climates is affected (Cornu et al., 2008; Montagne et al., 2016).

One of the main difficulties in identifying possible trajectories for human-induced soil evolution relies on the multidirectional impacts that land use and management changes can have on the buffering of soil properties and consequently on soil resilience. For example, afforestation is known to protect soils physically from rainwater, to increase soil organic carbon content and to contribute to the biological enrichment of base cations and consequently to increase soil resilience. At the same time, afforestation also favors weathering and soil acidification, particularly under coniferous forests, which can reduce soil resilience (Cornu et al., 2008). Similarly, N-fertilization or nutrient export through harvests hastens soil acidification in conventional cropping systems (Guo et al., 2010) and reduces soil resilience, but liming practices increase soil pH and improve soil resilience (Montagne et al., 2016). Finally, irrigation can lead to either leaching or the addition of base cations or clay particles depending on the quality and management of irrigation water. Human occupation and activities will have very contrasting, sometimes opposite, effects on soil evolution. They will depend on the precise combination of chosen practices and of their practical implementation as evidenced by the almost unlimited diversity of Anthrosols (World Reference base).

Impact of global change on soil: Synergy or antagonism?

Climate and human induced changes in land use and management may mutually reinforce each other. This is, for example, the case in semi-arid areas where the decreasing amounts but increasing intensities of rainfall lead to enhanced erosion and where deforestation for cultivation or (over)grazing exerts pressure in the same direction. This is also the situation in areas where rainfall is increasing and where extensive soil sealing prevents the infiltration of rainwater into soils. These two effects combine and cause flooding and landslides. When combined with events of extreme intensity or with soils of low resilience, synergies between climate and land use or management changes can give rise to extreme phenomena as exemplified by major wildfires in Australia, Canada, Greece, Russia or the United States, and by massive landslides and floods in the South of France or more recently in Germany.

Land use and management changes can, however, also counteract the impact of climate change. One example from temperate regions is the increase in tree cover, which could partially or totally offset the increased turnover of organic matter caused by the increase in temperature. In cultivated fields, practices that bring organic matter and other amendments to the soil surface may be of interest to regulate the stock of SOC, carbonates, clay minerals or base cations. For SOC, the “4 per 1000” initiative promotes land use and management changes that could offset the impact of the agricultural sector on climate change or even mitigate it. Although theoretically possible with soils, the practical mitigation of the effects of climate change suffers from several limitations. First, the effects of climate change are already being observed and affect large areas over the world, but the adoption of practices to counteract them depends on political, economic and sociological conditions that will be unlikely to evolve at the pace and spatial scales of climate change. The practice of N fertilization has, for example, induced the acidification of Chinese soils in the past 20 years (Guo et al., 2010) despite the fact that liming practices have been well known and used for millennia to maintain soil pH. Furthermore, when practices to counteract change rely on finite resources (e.g., phosphate fertilization) or consume energy (e.g., for extracting, transporting or spreading limestone amendments), their large-scale use will rapidly become unsustainable or have counterproductive effects on climate change. If changes in soil use and management are insufficient to mitigate climate change and its impacts significantly and globally, they must be designed and implemented to avoid further negative consequences.

Assessing soil change under global change, a scientific challenge

As discussed above, some soil properties are already changing whereas others will change as a result of the complex feedbacks between the factors of soil formation, the direction and intensity of soil processes and soil properties. As far as the dynamics of soil properties are concerned, two types of soil properties are defined: those that can be managed by human activities (SOC, pH, nutrients, topsoil structure, i.e., the macroscopic organization of solid grains and pores) and the so-called inherent soil properties (texture, mineralogy, soil type, soil depth, horizon sequence, structure at depth). When addressing global change, time steps from a few decades to a century are important (Fig. 2). Despite early recognition of the dynamic nature of all soil properties on a time scale ranging from several decades to a century, inherent soil properties are still considered stable in time. This is probably due to the lack of systematic research on the multi-decadal impact of land use and soil management on changes in soil properties. Disregarding

this dependence of the soil's inherent properties on time, which strongly affects the manageable ones as demonstrated for SOC (Mathieu et al., 2015), induces errors in predicting the dynamics of the manageable properties.

The quantification of medium to long-term changes in soil properties, in particular of inherent properties, is challenging. Long-term field experiments are the tools of choice for studying evolution of soil inherent properties (Filippi et al., 2016), however, they are too rare to provide an exhaustive characterization of all the potential combinations of soil resilience with changes in climate, land use or management. Another classical approach in studying change in soil is the use of soil sequences, in which only one pedological factor changes while others remain as constant as possible. This assumption is, however, sometimes difficult to verify (Filippi et al., 2016). Studies of soil sequences that vary mainly with climate (so called climosequences) were performed along latitudinal or altitudinal gradients. They demonstrated a significant impact of climate on soil evolution. However, in this approach both temperature and rainfall evolve jointly in a way that may differ from that induced by climate change and the change of atmospheric CO₂ concentration is not considered. Some studies have also shown by an analogical approach with artificial drainage as a proxy for the projected increase in soil drainage due to climate change in temperate Luvisols (Montagne and Cornu, 2010), that the impact of climate on soil evolution could be significant over a few tens of years. Conversely, anthroposequences have been used to quantify the effect of different cultivation practices on soil, notably tillage reduction, artificial drainage, organic amendments among others. However, these studies were mainly performed on a local scale and for a few types of soil only. When considering global change, other approaches have to be considered, for example, monitoring networks that have been implemented increasingly over the last few years at the national scale. Their implementation is time- and resource-consuming, however, and their results do not elucidate the effects of global change unless at least two campaigns separated by a few decades are available at the global scale. A last, but challenging, option is the reuse of legacy data that are available worldwide (Filippi et al., 2016).

All these approaches involve studying past changes. They may be used to provide future projections, but because future combinations of climate, land use, and soil type are unknown, they may fail to deliver a good result. Modelling approaches are thus needed.

Two main types of modelling approaches exist in the literature: (i) empirical or statistical, and (ii) more deterministic models. Purely empirical or statistical predictions, despite being highly appropriate for applied purposes and decision-making, may not be useful for predicting future states of soils over the long term (century or more). Alternatively, deterministic modelling enables us to describe the interaction of soil processes with the pedological factors. Most of the existing models, however, depict the effect of various agricultural practices on the evolution of a single soil property (e.g. SOC content, sensitivity to erosion). Although they do not consider important soil internal feedbacks, they often integrate other soil properties in the form of input data (e.g. clay content in SOC dynamics models). A few attempts at pedogenesis modelling developed over the last decade explicitly consider the fluxes of solid constituents, solute and water transfers, agricultural practices and the feedback loops between soil functioning and soil structure at the profile (e.g. SoilGen, Finke, 2012) or the landscape scale (e.g. MILESD, Vanwalleghe et al., 2013; HydroLorica, van der Meij et al., 2019). They could, therefore, simulate the effect that changes in climate or human practices could have on soils because they (i) consider interactions between physical and chemical processes; (ii) take into account climate factors, land-use and agricultural practices as boundary conditions; and (iii) provide the depth distribution of a large range of soil properties. They do not, however, generally contain a dynamic description of the other compartments of the ecosystem, notably plant development. On the other hand, dynamic global vegetation models (DGVM) have been developed to predict the impacts and feedback of climate and land use on vegetation and carbon budgets. The DGVMs notably predict primary production, water, carbon and nitrogen fluxes to the atmosphere and surface water, as well as their storage in the soil (Le Quére et al., 2018). However, in these models, for which soil data availability is a key issue, soils are the least well described compartment (Dai et al., 2019). This situation leads to significant biases in DGVM predictions (Folberth et al., 2016), that can exceed the effect of climate variability (Koven et al., 2013). Therefore, while these models represent the best tool we currently have for simulating the long-term effect of human-induced global change on soil, they require a better description of the interactions and temporal evolution of soil properties.

Concluding remarks

In soil science, climate and human activities were recognized very early as significant factors for soil formation, but the understanding of current climate and land use or management changes on soil is limited to: (i) a subset of soil degradation processes; (ii) manageable soil properties, in particular organic carbon concentration and stocks; and (iii) relatively short time scales in general. In addition, the precise dynamics of how soil responds to climate or human impact is still poorly understood. However, global change has already had an impact, and will continue to affect all soil fluxes and stocks with consequences not only on the manageable soil properties but also on a large range of inherent soil processes and properties. Moreover, the examples given above show that soil change may be faster than previously thought, especially for the inherent properties.

Predicting the impact of global changes on soil processes and properties remains a challenging task given (i) the interlinking of different and complex changes in patterns of climate, land use and management as well as of soil resilience, (ii) the shift from a pedogenesis driven by slow, low amplitude and continuous changes in soils to a pedogenesis driven by more rapid, large amplitude and irregular changes due to the increasing frequency of extreme climatic events and the globalization of land use and land management abrupt changes, and (iii) the current lack of adequate tools to assess unprecedented combinations of climate, land use and management on a set of interacting soil processes and properties. Understanding these complexities is, however, becoming urgent as climate or human-induced changes in soil properties will inevitably have either positive or negative effects on the future

delivery of soil-based ecosystem services such as biomass production, water storage and purification, climate regulation, erosion and flood regulation, etc. Beyond this complexity, there is no doubt that climate change is more likely to induce additional removal of primary or clay minerals than the reverse. Similarly, while examples of soil deterioration due to human activities are frequent, examples of soil recovery are less common. For such reasons, climate and land use changes are considered as two major factors for soil degradation. Halting human-induced soil degradation is unlikely to be sufficient to mitigate climate change significantly. It should, however, be at least the minimum objective to avoid or reduce cataclysmic changes and maintain sufficient resilience in the soil to climate change. This would give scientists and practitioners enough time to understand better the impact of global changes on soils, and to develop better approaches and tools that could help us maintain soil ecosystem services to at least the same level as they are now.

References

- Augusto L, De Schrijver A, Vesterdal L, Smolander A, Prescott C, and Ranger J (2014) Influences of evergreen gymnosperm and deciduous angiosperm tree species on the functioning of temperate and boreal forests. *Biological Reviews of the Cambridge Philosophical Society* 33. <https://doi.org/10.1111/bvr.12119>.
- Cornu S, Besnault A, and Bermond A (2008) Soil podzolization induced by reforestation as shown by sequential and kinetic extractions of Fe and Al. *European Journal of Soil Science* 59: 222–232.
- Cornu S, Montagne D, Hubert F, Barré P, and Caner L (2012) Evidence of short-term clay evolution in soils under human impact. *Comptes Rendus Geoscience* 344: 747–757.
- Cornu S, Keller C, Delolme C, Béchet B, Schwartz C, and Vidal-Beaudet L (2021) Pedological characteristics of SUITMAS: a snapshot. *Geoderma*. <https://doi.org/10.1016/j.geoderma.2021.115321>.
- Dai Y, Shangquan W, Wei N, et al. (2019) A review of the global soil property maps for Earth system models. *The Soil* 5: 137–158. <https://doi.org/10.5194/soil-5-137-2019>.
- Filippi P, Minasny B, Cattle SR, and Bishop TFA (2016) Chapter Four - Monitoring and modeling soil change: The influence of human activity and climatic shifts on aspects of soil spatiotemporally. *Advances in Agronomy* 139: 153–214. <https://doi.org/10.1016/bs.agron.2016.06.001>. eds Sparks D.L.
- Finke PA (2012) Modeling the genesis of Luvisols as a function of topographic position in loess parent material. *Quaternary International* 265: 3–17. <https://doi.org/10.1016/j.quaint.2011.10.016>.
- Folberth C, Skalský R, Moltchanova E, et al. (2016) Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations. *Nature Communications* 7.
- Friedlingstein P, O'Sullivan M, Jones MW, Andrew RM, Hauck J, Olsen A, Peters GP, Peters W, Pongratz J, Sitch S, Le Quéré C, Canadell JG, Ciais P, Jackson RB, Alin S, Aragão LEOC, Arneeth A, Arora V, Bates NR, et al. (2020) Global carbon budget 2020. *Earth System Science Data* 12(4): 3269–3340.
- Guo JH, Liu XJ, Zhang Y, Shen JL, Han WX, Zhang WF, Christie P, Goulding KW, Vitousek PM, and Zhang FS (2010) Significant acidification in major Chinese croplands. *Science* 327(5968): 1008–1010. <https://doi.org/10.1126/science.1182570>.
- Hartmann DL, Klein Tank AMG, Rusticucci M, Alexander LV, Brönnimann S, Charabi Y, Dentener FJ, Dlugokencky EJ, Easterling DR, Kaplan A, Soden BJ, Thorne PW, Wild M, and Zhai PM (2013) Observations: Atmosphere and surface. In: Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, ... Midgley PM (eds.) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, p. 96. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.
- Jia G, Shevliakova E, Artaxo P, De Noblet-Ducoudré N, Houghton R, House J, Kitajima K, Lennard C, Popp A, Sirin A, Sukumar R, and Verchot L (2019) Chapter 2: Land-Climate Interactions. In: Shukla PR, Skea J, Calvo Buendia E, Masson-Delmotte V, Pörtner H-O, Roberts DC, ... Malley J (eds.) *Climate Change and Land: An IPCC Special Report on Climate Change, Desertification, Land Degradation, Sustainable Land Management, Food Security, and Greenhouse Gas Fluxes in Terrestrial Ecosystems*, 186.
- Khudyakov OI and Reshotkin OV (2017) Soil evolution in relation to modern climate warming. *Theoretical and Applied Ecology* 2: 38–43.
- Koven CD, Riley WJ, Subin ZM, et al. (2013) The effect of vertically resolved soil biogeochemistry and alternate soil C and N models on C dynamics of CLM4. *Biogeosciences* 10: 7109–7131. <https://doi.org/10.5194/bg-10-7109-2013>.
- Le Quéré C, Andrew RM, Friedlingstein P, et al. (2018) Global Carbon Budget 2018. *Earth System Science Data* 10: 2141–2194. <https://doi.org/10.5194/essd-10-2141-2018>.
- Mathieu JA, Hatté C, Balesdent J, et al. (2015) Deep soil carbon dynamics are driven more by soil type than by climate: A worldwide meta-analysis of radiocarbon profiles. *Global Change Biology* 21: 4278–4292. <https://doi.org/10.1111/gcb.13012>.
- Matocha CJ, Grove JH, Karathanasis TD, and Vandiviere M (2016) Changes in soil mineralogy due to nitrogen fertilization in an agroecosystem. *Geoderma* 263: 176–184. <https://doi.org/10.1016/j.geoderma.2015.09.002>.
- Meredith M, Sommerkorn M, Cassotta S, Derksen C, Ekaykin A, Hollowed A, Kofinas G, Mackintosh A, Melbourne-Thomas J, Muelbert MMC, Ottersen G, Pritchard H, and Schuur EAG (2019) Polar regions. In: Pörtner H-O, Roberts DC, Masson-Delmotte V, Zhai P, Tignor M, Poloczanska E, ... Weyer NM (eds.) *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*.
- Millennium Ecosystem Assessment (2005) *Ecosystems and Human Well-Being: Wetlands and Water*. Washington, DC: World Resources Institute 86.
- Montagne D and Cornu S (2010) Do we need to integrate a soil module to model climate change? *Climatic Change* 98(1–2): 75–86.
- Montagne D, Cousin I, and Cornu S (2016) Changes in the pathway and the intensity of albic material genesis: Role of agricultural practices. *Geoderma* 268: 156–164.
- Powlson DS, Stirling CM, Jat ML, Gerard BG, Palm CA, Sanchez PA, and Cassman KG (2014) Limited potential of no-till agriculture for climate change mitigation. *Nature Climate Change* 4(8): 678–683. <https://doi.org/10.1038/nclimate2292>.
- Richter DD, Hofmockel M, Callahan MA Jr., Powlson DS, and Smith P (2007) Long-term soil Experiments: Keys to managing Earth's rapidly changing ecosystems. *Soil Science Society of America Journal* 71: 266–279.
- Rounsevell MDA, Evans SP, and Bullock P (1999) Climate change and agricultural soils: Impacts and adaptation. *Climatic Change* 43: 683–709.
- Singh BP, Setia R, Wiesmeier M, and Kunhikrishnan A (2018) Agricultural management practices and soil organic carbon storage. In: *Soil Carbon Storage: Modulators, Mechanisms and Modeling*. Elsevier Inc <https://doi.org/10.1016/B978-0-12-812766-7.00007-X>.
- Sombroek WG (1990) Soils on a warmer earth: Tropical and subtropical regions. In: Scharpenseel HW, Schomaker M, and Ayoub A (eds.) *Soils on a Warmer Earth*, pp. 157–174. Amsterdam: Elsevier.
- Song XP, Hansen MC, Stehman SV, Potapov PV, Tyukavina A, Vermote EF, and Townshend JR (2018) Global land change from 1982 to 2016. *Nature*. <https://doi.org/10.1038/s41586-018-0411-9>.
- Tamburini G, Bommarco R, Wanger TC, Kremen C, van der Heijden MGA, Liebman M, and Hallin S (2020) Agricultural diversification promotes multiple ecosystem services without compromising yield. *Science Advances* 6(45). <https://doi.org/10.1126/SCIADV.ABA1715>.
- Turner BL, Lambin EF, and Reenberg A (2007) The emergence of land change science for global environmental change and sustainability. *Proceedings of the National Academy of Sciences* 104(52): 20666–20671. <https://doi.org/10.1073/pnas.0704119104>.
- Van der Meij WM, Temme AJAM, Wallinga J, et al. (2019) Modelling soil and landscape evolution – The effect of rainfall and land use change on soil and landscape patterns. *Soil Discussions*. <https://doi.org/10.5194/soil-2019-82>.
- Vanwallegheem T, Stockmann U, Minasny B, et al. (2013) A quantitative model for integrating landscape evolution and soil formation. *Journal of Geophysical Research - Earth Surface* 118: 331–347. <https://doi.org/10.1029/2011JF002296>, 2013.

The effects of adequate and excessive application of mineral fertilizers on the soil

Bijay-Singh^a and Tek B Sapkota^b, ^aDepartment of Soil Science, Punjab Agricultural University, Ludhiana, Punjab, India; ^bInternational Maize and Wheat Improvement Center (CIMMYT), El Batán, Mexico

© 2023 Elsevier Ltd. All rights reserved.

Introduction	370
Mineral fertilizers as sources of plant nutrients	370
Effects of fertilizers on the nutrient supplying capacity of soils	372
Nitrogen	372
Phosphorus	372
Potassium	373
Micronutrients	374
Effect of fertilizer application on soil organic matter	374
Fertilizer use and soil acidification	375
Effect of fertilizers on soil microbial biomass	376
Effect of fertilizer use on enzyme activity in soil	377
Fertilizer effects on soil erodibility	378
Soil physical properties as influenced by fertilizer use	378
Fertilizers and heavy metal contamination of soils	378
Effects of fertilizer use on pollution of surface and ground waters, and climate change	379
Balanced application of fertilizer nutrients and soil properties	379
Soil properties under integrated nutrient management based on mineral fertilizers and organic manures	379
Summary	380
References	380

Abstract

Fertilizers increase availability of essential plant nutrients in the soil and alter many chemical, biological and physical soil properties. Fertilizers lead to the accumulation of soil organic matter (SOM), but the excessive application of nitrogen fertilizers is deleterious to SOM, may contribute to nitrate pollution of freshwaters and emission of nitrous oxide – a greenhouse gas. Nitrogen fertilizers can also cause soil acidification. Phosphatic fertilizers may contribute to eutrophication of surface waters, and heavy metal impurities in these may contaminate soils. Balanced fertilization positively affects soil organic matter and soil biota; and in general soil microbial life is positively influenced.

Glossary

Ammonification and urea hydrolysis Ammonification is the primary microbial process that converts reduced organic nitrogen to ammonium. In the case of urea hydrolysis, the enzyme urease catalyzes the conversion of urea to ammonia and carbonic acid.

Nitrification Biological oxidation of ammonium to nitrite and then to nitrate. Both steps are carried out by two different types of prokaryotes.

Prophylactic fertilizer use Application of fertilizers as a preventive measure to avoid development of nutrient deficiency in the crop to be grown on the soil.

Recommended fertilizer rate Crop and region specific fertilizer application rates, which result in profits for the farmers based on the cost of fertilizer and income from the crop, together with minimal losses of nutrients to the environment.

Shannon diversity index A measure of diversity that combines species richness (the number of species in a given area) and their relative abundances.

Soil acidification Reduction of soil pH and build-up of hydrogen cations due to production of various acids in the soil, active uptake of some nutrient ions by biota through proton pumps, transformation of nitrogen applied to the soil, and acid deposition on the soil.

Soil Erodibility Measure of the resistance of soil to erosion. Particles in highly erodible soils are readily displaced and transported by water or wind.

Key points

- Mineral fertilizers supply essential plant nutrients but also affect many soil functions both positively and negatively
- A large portion of nitrogen applied to soil as fertilizer converts into organic forms and becomes part of the soil nitrogen pool
- Fertilizers applied to crops at recommended rates promote the production of both above- and below-ground plant biomass leading to the accumulation of soil organic matter
- Fertilizer nitrogen applied at above the recommended rates can accelerate microbial decay of organic matter already present in the soil and also contribute to nitrate pollution of fresh water bodies and emission of nitrous oxide – a greenhouse gas
- Nitrogen fertilizers lead to soil acidification through microbial nitrification of ammonium ions and leaching of nitrate ions from the soil.
- Fertilizers in general positively influence the microbial life in soil
- Fertilizer use reduces soil erodibility through enhanced plant growth leading to soil stabilization and increased organic matter contents
- Heavy metals and radionuclides contained as impurities in phosphatic fertilizers may contaminate soils
- Balanced application of N, P, and K and integrated management of mineral fertilizers with organic manures enhance the positive effects of fertilizers in the accumulation of soil organic matter, biota, and physical properties
- Long-term use of phosphatic fertilizers in intensive agriculture leads to a build-up of surplus phosphorus in the soil and its transfer via runoff results in eutrophication of lakes and rivers

Introduction

Soil is a complex and variable natural resource, critical for life on earth, and natural living medium for the growth of plants. It provides mankind with most of its food, nutrients, and energy needs. It is a habitat for many organisms and constitutes a living system characterized by many complex biological functions that interact with the abiotic, physical, and chemical environment. From the perspective of agro-ecosystems, a healthy soil supports, nurtures and sustains food production with minimum adverse effects on the environment. Soil health is reflected in several management-sensitive soil properties and the ability of soil to carry out different ecosystem functions, respond to agricultural interventions, and thwart processes that degrade it.

In the last century, humans have altered agricultural soils more than in any comparable period in human history. While tillage and irrigation constitute the two major anthropogenic interventions in the functioning of agricultural soils, during the last 70 years the application of mineral fertilizers has resulted in a fundamental change in soil and crop management to improve global food security. Fertilizers were introduced in the late 19th and early 20th centuries, but global nitrogen (N) and phosphorus (P) fertilizer use was negligible in 1900 and increased very slowly to about 7 M nutrient t in 1950. It was only after World War II when large scale production of N fertilizers became possible by the Haber–Bosch process that fertilizer use increased exponentially to support the rapidly growing human population. Thus, massive additions of mineral fertilizers to agricultural soils started only after 1950 (Fig. 1) and it was an integral component of the Green Revolution witnessed from the 1950s onwards in different parts of the world. According to an estimate, increased food production achieved by the application of N fertilizers now supports about half of the global population (<https://ourworldindata.org/fertilizers>, accessed 24 January 2022). However, mineral fertilizers are chemicals that can potentially alter the natural functioning of soils and lead to major and sometimes irrevocable changes in their characteristics.

High yields are obtained by applying recommended fertilizer application rates for different crops. But many farmers tend to apply them at excessive rates to avoid the risk of yield loss from nutrient deficiencies. For example, after the 1980s, in many countries in East and South Asia, fertilizer N consumption increased several fold more than in the North America and West Europe, where levels have either stabilized or decreased since 1990 (Fig. 1). In 2019, China and India, with about 36% of the global population, used 42.45% of the 107.7 Mt. total fertilizer N consumed globally; North America and Western Europe consumed only 16.95% (<http://www.fao.org/faostat/en/#data/RFN>; Accessed 24 January 2022). Fertilizers, particularly when applied in excess, can lead to losses of nutrients from the soil–plant system to the environment and degrade its quality. The challenge ahead is to understand the changes in soil properties from fertilizer use and to manage them to keep soils healthy to support sustainable food production with minimal environmental impact. Efficient fertilizer management and sustainable soil health will ensure production of sufficient food from the land already under cultivation and allow land to be spared for purposes other than agricultural production.

Mineral fertilizers as sources of plant nutrients

Mineral fertilizers are chemicals that contain elements essential for the growth of plants. The three primary plant nutrients supplied in large amounts by fertilizers are N, P, and potassium (K). Chemicals supplying zinc (Zn), boron (B), copper (Cu), iron (Fe),

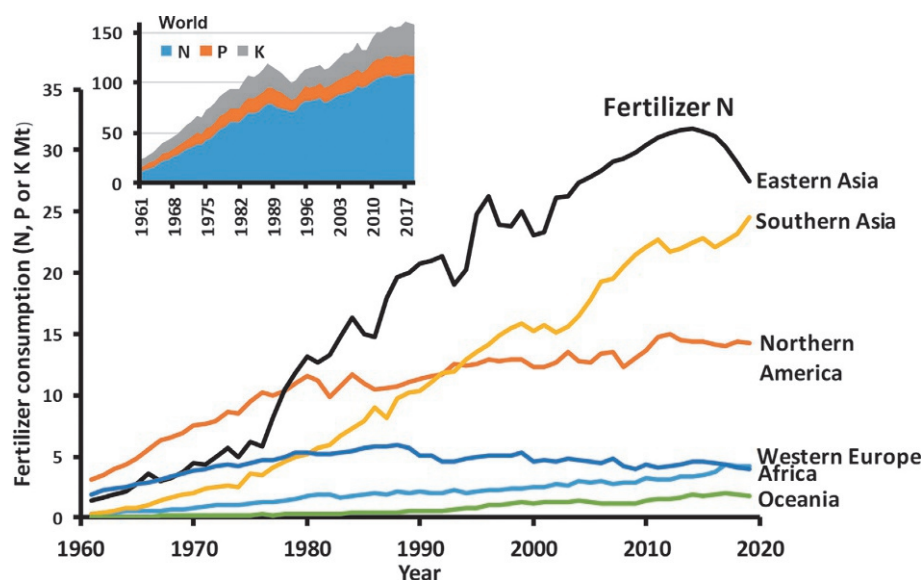


Fig. 1 Global fertilizer nitrogen, phosphorus, and potassium use from 1961 to 2019. Data source: <http://www.fao.org/faostat/en/#data/RFN>, accessed 24 January 2022.

manganese (Mn), and molybdenum (Mo) to plants are required in very small quantities and are classified as micronutrient fertilizers.

Fertilizers are used to achieve sustainable agricultural production. Global consumption of mineral fertilizers has increased from 11.39 Mt. N, 4.74 Mt. P, and 7.15 Mt. K in 1961 to 107.74 Mt. N, 18.95 Mt. P, and 31.03 Mt. K in 2019 (Fig. 1). In particular, N is vital for high yields because it is required in large quantities for building plant proteins that are essential for converting solar radiation into carbohydrates that control plant growth. The effect of applying fertilizer N to various crops is generally conspicuous in terms of plant growth and crop yields. Therefore, its use has increased from 1961 to 2019 by almost 9.5 times compared to four times for both P and K. Furthermore, to ensure food security, fertilizer N has been heavily subsidized by governments in countries such as China and India resulting in higher than necessary applications to reduce the risk of yield losses from N deficiency. Effects of fertilizers on soils differ depending on rates of application. Nutrient application can be as single-nutrient fertilizers or compound multi-nutrient fertilizers. There are two ways to produce compound fertilizers: (1) by combining two or more ingredients through chemical reaction (complex fertilizers), and (2) by blending two or more granular fertilizers. Their effects on the soil are governed by the amount of nutrients applied, nature of the chemicals in the fertilizer, method of application (soil surface application, band placed or mixed into the soil), general soil conditions (such as temperature, soil water content, soil texture and structure), and environmental conditions at the time of application.

Ammonia produced by the energy-intensive Haber–Bosch process is the most used source for producing different N fertilizers. Urea remains the most commonly used N fertilizer all over the world. In 2018, farmers applied 48.1% N as urea, 20.5% in the form of ammonia and various ammonium salts, and 20.3% as NPK, NP, and NK compound fertilizers (<https://www.ifastat.org/databases/plant-nutrition>, accessed 3 August 2021). When applied to the soil, urea is readily hydrolyzed to NH_4^+ and carbon dioxide by the urease enzyme present in soil. This reaction raises soil pH which may lead to ammonia volatilization of some of the applied urea N. The NH_4^+ -N produced from urea hydrolysis or through applying NH_4^+ fertilizers may be immobilized into soil organic matter (SOM), taken up by plant roots, or converted to NO_3^- through nitrification – a bacterial process dependent on soil temperature, moisture, aeration, and pH. In well-aerated soils, NH_4^+ -N is converted to NO_3^- -N within a few weeks. Plant roots can readily take up and transport both NH_4^+ and NO_3^- ions.

Most phosphatic fertilizers are produced from rocks containing fluorapatite and hydroxyapatite. By treating phosphate-rich rock with strong acids, the P-containing minerals are converted to water-soluble P salts. When the rock is treated with sulfuric acid, normal or single superphosphate fertilizer containing a mixture of monocalcium phosphate and gypsum in about equal proportion is produced. Triple superphosphate fertilizer is produced by treating phosphate rock with phosphoric acid. Nitric phosphate fertilizers containing both N and P are produced when nitric acid is used to solubilize P-containing minerals in the phosphate rock. Superphosphate application to soils results in an acidic reaction. Phosphorus-containing ammonium phosphate fertilizers are produced by reacting ammonia with phosphoric acid. When applied to the soil, diammonium phosphate is basic in reaction while monoammonium phosphate is acidic. However, nitrification of NH_4^+ reduces soil pH a few weeks after the application of either of the two ammonium phosphates.

The two common K fertilizers are potassium chloride (muriate of potash) and potassium sulfate, although potassium nitrate is also used sometimes to supply K to crops. Potassium chloride, the most widely used source of K on farms is mined from rock

deposits left by evaporites in different regions of the world (e.g. Canada). Potassium sulfate is produced by reacting potassium chloride with sulfuric acid. All K fertilizers supply K^+ ions to the soil–plant system.

Effects of fertilizers on the nutrient supplying capacity of soils

Nitrogen

In all soils, the large amount of N retained in SOM is referred to as the soil N pool. Mineral-N (NH_4^+ + NO_3^-) is released from this pool to a smaller pool through mineralization–immobilization turnover, a microbial process controlled by soil temperature and moisture conditions (Fig. 2). Nitrogen applied to soil systems through various input sources such as mineral fertilizer, organic matter and crop residues plus N-fixation and atmospheric deposition, adds to these soil N pools. Nitrogen contained in the soil mineral-N pool can be taken up by plant roots and microbial organisms but is also susceptible to losses by leaching, ammonia volatilization, and denitrification. Data generated using ^{15}N -labeled fertilizers in 93 studies from all over the world show that the average recovery of fertilizer N by harvested grains and above-ground plant residues of rice (*Oryza sativa* L.), wheat (*Triticum aestivum* L.), and maize (*Zea mays* L.) crops is only 44% (Ladha et al., 2005). Remaining N taken up by crop plants is supplied through mineralization of N from the organic N pool in the soil that has been enriched by fertilizer N applications in previous years. Research on 217 fields with grain crops to which ^{15}N -labeled fertilizers were applied, revealed that even with large rates of fertilizer N average plant N uptake from the soil N pool was about 60% (Gardner and Drinkwater, 2009). A portion only of fertilizer N applied to agricultural soils (about 50%) contributes directly to the mineral N pool of which some may be lost from the soil–plant system (Fig. 2). Remaining fertilizer N converts into organic forms to become a part of SOM. Sebilo et al. (2013) showed that 61 to 65% of the one-time application of ^{15}N -labeled fertilizer to a sugar beet (*Beta vulgaris* L.)–winter wheat annual crop rotation on a chalk soil was used by the crops adequately fertilized by unlabeled N during the next three decades. Nevertheless, 12 to 15% of the ^{15}N was recovered in the soil N pool three decades after its application. An important implication of the interaction between fertilizer N and soil N is that a portion of the annual fertilizer N inputs that exceeds the N either removed by the harvested crops or becomes a part of the soil N pool, is vulnerable to losses to the environment. Because some fertilizer N is always incorporated into the soil N pool and becomes available for uptake by crop plants subsequently suggests that unless adequate fertilizer N is applied to crops, the soil N pool will gradually be depleted and soil quality adversely affected. Thus the soil N pool, if not maintained by sufficient fertilizer N and or through organic manure application and recycling of legume residues, will not enable full yield potential and soil biodiversity to be achieved.

Phosphorus

Phosphorus is held by different soil components with a continuum of bonding energies (Syers et al., 2008). The various conceptual pools of soil P and movement of P between pools are shown in Fig. 3. In the soil solution, P is present as orthophosphate ions, principally as $H_2PO_4^-$ and to a lesser extent as HPO_4^{2-} , which are bioavailable for plant uptake. While P concentration in the soil solution at 10^{-4} M is considered high, in P deficient soils it is 10^{-6} M, and in some tropical soils with very poor fertility it can be as low as 10^{-8} M (Syers et al., 2008). In the readily available P pool, which is in equilibrium with P in the soil solution, P is weakly bonded to the surfaces of soil components. Phosphorus is readily transferred from this pool to the soil solution when P is taken up

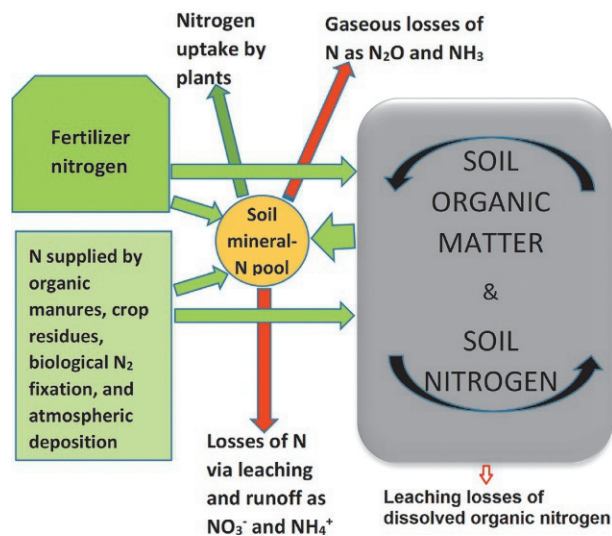


Fig. 2 Schematic diagram showing the interactions between fertilizer N and soil N, and its effects on crop N uptake and loss.

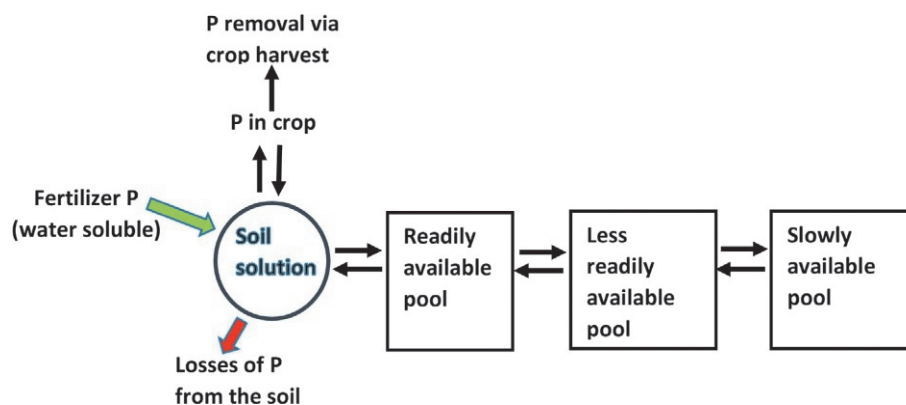


Fig. 3 Schematic diagram showing the fate of fertilizer P in relation to different P pools in the soil. Adapted from Syers JK, Johnston AE, and Curtis D (2008) *Efficiency of Soil and Fertilizer Phosphorus Use: Reconciling Changing Concepts of Soil Phosphorus Behaviour with Agronomic information*, FAO Fertilizer and Plant Nutrition Bulletin 18. Rome: Food and Agriculture Organization of the United Nations.

by plant roots. In the less readily available soil P pool, it is adsorbed onto the soil matrix or internal surfaces of soil components. Over time P in this pool can revert to the readily available pool. There is also a slowly available P pool, in which P is very strongly bound to soil components through (1) complexation with soil minerals, (2) precipitation as sparingly soluble compounds, or (3) location in the soil matrix and inaccessible to plant roots. Equilibria between different P pools in the soil primarily involve adsorption and diffusive penetration reactions that may be largely reversible, therefore, plant availability of fertilizer P may decrease with time. Reversible transfer of P between the three pools is important for its behavior in soil; only a fraction of the total P applied through water-soluble P-fertilizers remains in soil solution while the remaining P is distributed between the readily- and less readily-available P pools.

Soils rarely contain sufficient P in the soil solution to meet the demand from crops. Phosphorus in soil solution is replenished several times a day from readily and less readily available P pools. Globally, about two thirds of the total agricultural land contains too little plant-available P. To avoid sub-optimum yields of crops associated with P deficiency, farmers often adopt a prophylactic approach and over-apply P fertilizers. The largest rates of fertilizer P application have been by farmers in East Asia, Western Europe, the mid-western United States, and southern Brazil. Depending on the soil type, its physical condition and chemical status, crop, and environmental conditions, 10–20% only of the fertilizer P applied in the first year is utilized by the crops; the remaining P adds to the soil reserves leading to its accumulation in slowly available P soil pools (Fig. 3). For example, from 1965 to 2007 cumulative P input of 1115 kg ha^{-1} in cropland in Europe from fertilizers and manures was more than three times the cumulative P removal by crops (Sattari et al., 2012). This 'residual' or 'legacy' P represents not only a significant environmental risk in terms of P-related eutrophication of rivers, lakes, and estuaries but is also a P reserve that can potentially be exploited to reduce fertilizer P use if appropriate technological and management interventions are developed and used.

Potassium

Potassium fertilizers are applied to most agricultural soils because most soil K (10 to 20 g kg^{-1} soil) is only potentially available for plant uptake. In terms of plant availability and complex dynamics in the soil, K can be differentiated into four soil pools: water-soluble or soil solution K, exchangeable K, non-exchangeable K, and structural K. There are dynamic equilibrium reactions between these pools. Plants and microbes take up K from the water-soluble pool (0.1–0.2% of total soil K), but the exchangeable K (1–2% of total soil K), which is electrostatically bound on the surfaces of clay minerals and humic substances, can be released quickly into the soil solution. Non-exchangeable and structural K (incorporated in the crystal lattice structure of minerals) account for 96–99% of total soil K, and are considered to be slowly- and potentially-available K sources for plants, respectively.

Removal of K by crops and losses by leaching in coarse-textured soils and soil erosion are not being adequately replaced. Less than 10% of K removed from the soil is counterbalanced by fertilizer K and other inputs. Application of less than the required amount of fertilizer K can lead to unbalanced fertilization and may also result in significant depletion of soil K reserves and reduced soil fertility. Application of K immediately leads to an increase in concentrations of soil solution K and to adsorption of K onto clay minerals as exchangeable K. Therefore, one application of fertilizer K to crops is generally sufficient to meet their requirement; only in coarse-textured soils with low cation exchange capacity (CEC) does it need to be applied in two or three split doses. However, irrespective of the cropping system, less fertilizer K applied than the crop requirement will compromise soil fertility and crop production. Nevertheless, rates of fertilizer K application are often based on optimal rates of fertilizer N rather than the crop's actual K requirements.

Micronutrients

Micronutrient deficiencies have increased markedly since the middle of the 20th century largely due to the intensification of arable farming, increased use of N, P and K fertilizers, cultivation of marginal and reclaimed land, loss of topsoil by erosion, liming of acid soils, and decreasing use of organic manures. While deficiency of Zn has emerged as the most ubiquitous micronutrient problem, B deficiency is the second most widespread. Deficiencies of Fe, Cu, Mn and Mo vary in importance in different regions of the world depending upon crop and soil properties. Regular use of micronutrient fertilizers is increasing both as soil applications (Zn, B, Cu and Mo), foliar sprays (Fe and Mn) or seed treatment (Mo). Simple compounds such as zinc sulphate, sodium borate, and copper sulphate are generally used for soil application.

Application of micronutrients to soils increases production of both above- and below-ground crop biomass, which leads to the accumulation of SOM. Additional SOM not only contributes to soil health, but also determines the distribution of micronutrients between soil colloids and solution by controlling sorption of micronutrients through active functional groups and by forming soluble complexes. The formation of stable micronutrient complexes facilitates uptake of micronutrients by plants (Dhaliwal et al., 2019). Since Zn, Cu, B and Mo are relatively less sensitive than Fe and Mn to redox changes in the soil, their availability to plants is more governed by SOM.

Effect of fertilizer application on soil organic matter

Soil organic matter regulates a host of biological, chemical, and physical soil processes and contributes to soil functions. Although the relative importance of many functions performed by SOM differs with soil type, climate, and land use, one of the most important is as a nutrient reservoir for plants and ultimately for animals and humans. Organic matter content of agricultural soils is a key indicator of soil health. If its content declines because of microbial mineralization and topsoil erosion, soil quality deteriorates, the benefits of SOM decline and there are adverse effects on N nutrition of crop plants. Because of the fundamental coupling of microbial carbon (C) and N cycling and the close correlation between the mineralization of C and N, SOM acts not only as the largest source of N for plants but also the largest sink for applied fertilizer N in agroecosystems.

Application of fertilizers, particularly N, to crops can result in the accumulation or decline of SOM. When applied as the crop demands, fertilizer N promotes the production of both above- and below-ground plant biomass which increases SOM contents. Excessive applications of fertilizer N may, however, accelerate SOM loss through microbial decay of plant litter and organic C already present in the soil. The accumulation of SOM from fertilizer use in cropping systems is widely accepted, but its loss because of excessive fertilizer N application has been demonstrated only recently by Poffenbarger et al. (2017). In continuous maize cropping for 16 years, SOM storage in the surface soil layer increased with increasing application of N, reaching a maximum at the recommended amount for maize. Application of less than the recommended N stimulated crop growth resulting in increasing crop residue inputs into the soil, which in turn increased the rate of SOM accumulation. On the other hand, application of fertilizer N to maize above the recommended rate did not increase crop biomass production significantly above that of the recommended level. Increased amounts of residual inorganic N in the soil from excessive fertilizer N stimulated the mineralization of SOM resulting in its net loss. Fig. 4 provides a conceptual depiction of the response of SOM as a function of fertilizer N application up to and beyond the recommended level. Considerable residual inorganic N resulting from excessive fertilizer N applications accelerates the mineralization of SOM by supplying N for the growth of microbes involved in this process. Application of fertilizer N at large rates also reduces the C:N ratio of crop residues, which accelerates their decay. Thus the response of SOM to fertilizer N is controlled by multiple processes but the extent of C immobilization or SOM mineralization is determined by the rate of fertilizer N application.

Increases in SOM and associated soil quality properties through the application of recommended levels of N, P, and K fertilizers in intensively managed multiple cropping systems have been recorded by several researchers cited by Bijay-Singh (2018). Table 1 gives the average increase in soil organic C (SOC) with fertilizer use estimated by four meta-analyses based on different crops and cropping systems in various regions of the world. Although Ladha et al. (2011) observed an average increase of 8% in SOC with fertilizer N use compared with that in no-N plots, the extent of increase in tropical, humid subtropical, and temperate soils ranged from 3 to 16%. The positive effect of fertilizer use on SOM in tropical soils was greater than that in temperate regions probably because tropical soils inherently contain small amounts of SOM. It is interesting to note that increases in SOM with fertilizer use have been observed in tilled soils under crop rotations. Tillage enhances the decomposition of organic matter in most soils. Therefore, the positive effect of fertilizer use on increasing SOM can be assumed to be higher than that shown in Table 1, and interpreting the changes in SOM with fertilizer use in long-term experiments needs to consider the interactions between tillage and fertilizer use.

In a meta-analysis of experiments in which soil C was isotopically labeled to distinguish between old and new SOM, it was found that N application substantially increased new soil C stocks (Huang et al., 2020). The N-induced increases were greatest in ecosystems receiving small amounts of N; however, the positive effects are likely to diminish with continuous N enrichment. In soils with poor mineral N supply, soil microbes fulfil their N demands by accelerating the decomposition of organic matter. Recently, Li et al. (2017) reported that application of recommended amounts of fertilizer N in soils with low N content reduced the decomposition of SOM already in the soil. Cultivation invariably reduces SOM, but the application of fertilizers reduces the extent of this decline.

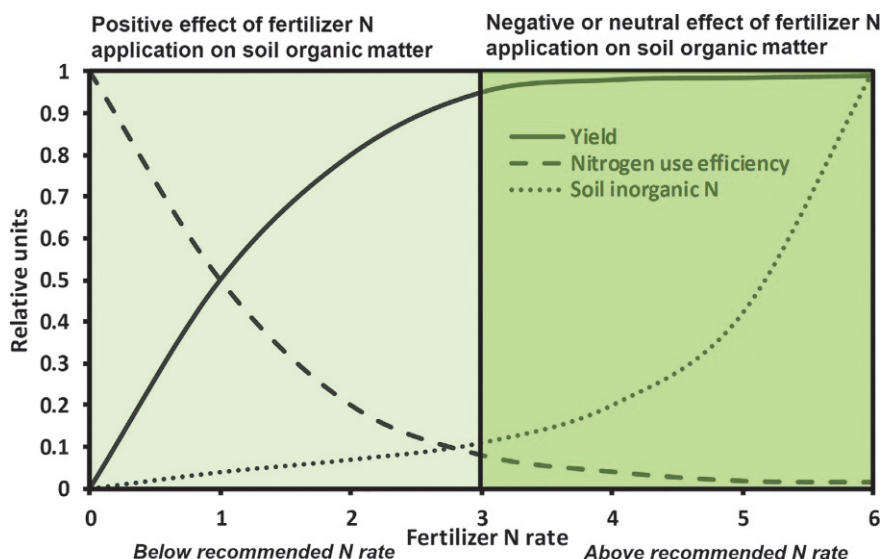


Fig. 4 Conceptual diagram depicting effects of fertilizer N application to crops at below- and above-recommended rates on soil organic matter through relationships between increasing fertilizer N levels and yield and crop residue production, N use efficiency (N output per unit of N input), and residual soil inorganic N. Expected effects of fertilizer N application at below- and above-optimum levels on soil organic matter responses are shown above the light and dark green areas of the plots. Modified from Poffenbarger HJ, Barker DW, Helmers MJ et al. (2017) Maximum soil organic carbon storage in Midwest US cropping systems when crops are optimally nitrogen-fertilized. *PLoS One* 12: e0172293. doi: 10.1371/journal.pone.0172293.

Fertilizer use and soil acidification

Microbial nitrification of NH_4^+ ions and leaching of NO_3^- ions constitute the two processes that produce H^+ ions when N fertilizers are applied to the soil (Fig. 5). Uptake of NO_3^- ions by crop plants releases OH^- ions, which neutralize a portion of H^+ ions released during nitrification and nitrate leaching. If plants take up all the NO_3^- ions produced from urea and ammonium fertilizers, net H^+ ion production is zero. It implies that N fertilizers mostly lead to soil acidification if application rates exceed crop demands. Nevertheless, when NO_3^- ions are leached, they carry with them positively charged basic cations Ca^{+2} , K^+ , Mg^{+2} , and Na^+ , which accelerates soil acidification (Fig. 5). If the N fertilizer applied is ammonium sulfate, leaching of both NO_3^- and SO_4^{+2} ions deplete basic cations resulting in increased acidification. Urea is less acidifying than NH_4^+ -based fertilizers because OH^- ions produced during urea hydrolysis and ammonification neutralize a portion of H^+ ions produced from the nitrification of NH_4^+ ions.

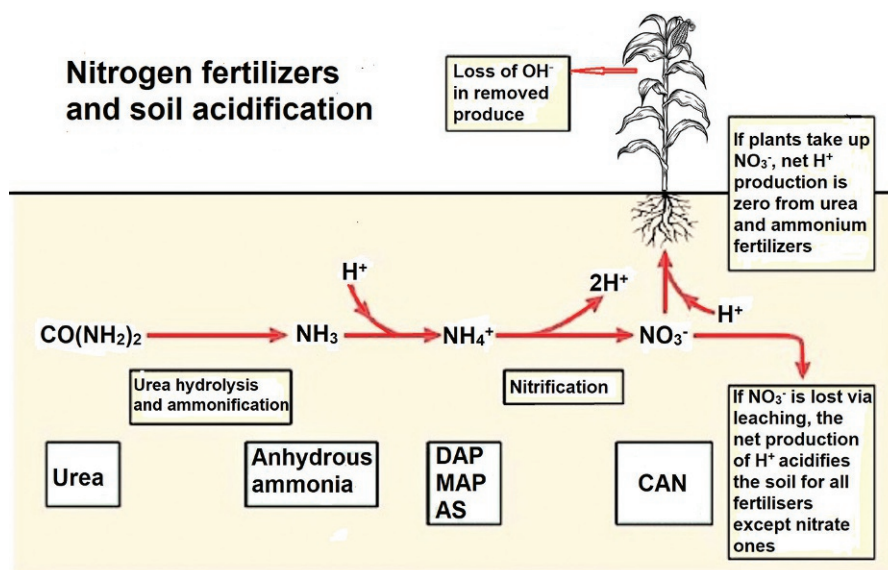


Fig. 5 Schematic diagram showing soil acidification from the application of different nitrogen fertilizers for crop production. DAP: Diammonium phosphate, MAP: Monoammonium phosphate, AS: Ammonium sulfate, CAN: Calcium ammonium nitrate.

Dissolution of monocalcium phosphate, the P carrying compound in superphosphate fertilizer, in the soil results in the formation of dicalcium phosphate and phosphoric acid around the fertilizer granules. Subsequently, phosphoric acid is dissociated into H_2PO_4^- and H^+ ions. However, acidification of the soil with the application of superphosphate is mild because a portion of H^+ ions is neutralized by OH^- ions released during adsorption of H_2PO_4^- onto soil particles. Because H_2PO_4^- ions are strongly adsorbed onto soil particles, H_2PO_4^- -induced leaching of basic cations hardly occurs.

Buffering capacity of the soil to resist changes in pH constitutes one of the most important factors that determine the extent of soil acidification. Soil constituents such as calcium carbonate (in calcareous soils), organic matter, and Fe and Al (aluminium) oxides contribute to the pH buffering capacity of soils. Calcareous soils in semi-arid and arid regions of the world are highly buffered. Soils in temperate regions are generally neutral or slightly acidic. However, highly weathered soils in the tropics are mostly acidic and possess little or no pH buffering capacity. With the continuous use of fertilizer N in such poorly buffered soils, when base cations such as Ca^{+2} and Mg^{+2} are depleted during the acidification process, Al^{+3} is released from soil minerals, often reaching toxic levels. Toxic levels of Al in the soil solution affect root cell division and inhibit root elongation. Aluminium also reacts with P, reduces its solubility in the soil and creates P deficiency in plants.

Guo et al. (2010) noted that large rates of fertilizer N application in a major crop-production area in China between the 1980s and 2000s resulted in severe soil acidification. Based on paired data from 154 field locations, the average pH declined by 0.50 units in the topsoil. Transformations of fertilizer N in the soil released 20 to 221 kilomoles H^+ ha^{-1} year^{-1} in four different cropping systems. Uptake of base cations contributed a further 15 to 20 kilomoles H^+ ha^{-1} year^{-1} to soil acidification. A meta-analysis of 106 studies from all over the world revealed that on average soil pH declined by 0.26 units from N additions and the rate of decline was linearly related to rates of N application (Tian and Niu, 2015). When the duration of the experiment was longer than 20 years, the extent of soil acidification due to N application was less. In the early stages of significant N additions, base cations Ca^{+2} , Mg^{+2} , and K^+ play a crucial role in buffering against N-induced soil acidification, but in many regions of the world, N additions are now transitioning to the Al^{+3} -based buffering phase.

Effect of fertilizers on soil microbial biomass

Microorganisms play an important role in maintaining soil fertility and productivity. The impact of fertilizers on different assemblages of soil microorganisms can have important implications for sustainable agriculture. Mineral fertilizers interact with microbial communities in the soil and affect biological activity, soil bacterial biomass, and community composition through enhanced cropping system productivity, crop residue return, SOM accumulation, and increased nutrient availability. Fertilizers also affect microbial life in the soil indirectly by altering pH. Soil acidification caused by the application of N fertilizers can have a considerable negative effect on soil organisms. Large amounts of nutrients from excessive applications of fertilizers may also decrease soil microbial activities.

Positive or negative effects of fertilizer use in crop production on the soil biota are determined by the type and amount of fertilizer used, the mode of application and duration. For example, the potential negative effect of a large concentration of ammonia fertilizer applied in bands on soil microbial life is usually short-term and confined to the application zones. Fertilizer application to crops usually increases soil microbial biomass. Several studies have reported a significant increase of soil microbial biomass, rates of microbial respiration, and community functional diversity with the long-term use of mineral fertilizers in cropping systems (Bijay-Singh, 2018). In some studies, the application of P together with N further stimulated soil microbial biomass. A significant correlation between microbial properties and SOM content suggests that the application of mineral fertilizers positively influences microbial life in the soil indirectly through the accumulation of SOM. In general, microbial activity in the soil is primarily limited by the availability of C substrate. But in agroecosystems when demand for N exceeds its supply, fertilizer N may exert a positive effect on soil microbial activity although there might be a negative impact from associated reduced pH. In two global meta-analyses from 55 studies on lowland rice (Geisseler et al., 2017) and 64 studies on upland crops (Geisseler and Scow, 2014), the application of mineral fertilizers (N or NPK) resulted in an overall increase of 15–29% in microbial biomass C compared to the no-fertilizer control treatments (Table 2). In upland crops, the increase in microbial biomass C averaged 48% in soils with pH 7 or higher but a decrease was observed with a pH below 5. When the duration of the experiment was less than 10 years, an average decrease of 22.7%

Table 1 The extent of increases in average soil organic C content with the application of fertilizer N compared to unfertilized controls in meta-analyses of long-term fertilizer experiments under different cropping systems in different regions of the world.

Crops	Region	Duration of long-term experiments	Number of experiments/ observations	Percent increase in soil organic C	References
Non-lowland rice crops	World	5–130 years	64	12.8	Geisseler and Scow (2014)
Cereal crops	World	6–158 years	104	8	Ladha et al. (2011)
Wheat, barley, oats, sugar beets, potato, maize, sorghum, rye	Europe	16–108 years	20	10	Körshens et al. (2013)
Irrigated or rainfed lowland rice	Asia	10 weeks– 58 years	46	13.4	Geisseler et al. (2017)

Table 2 The average response of soil microbial biomass carbon (MBC) to the application of mineral fertilizers (N or NPK) in meta-analyses conducted on experiments on lowland and non-lowland rice.

Experiments and treatments	Number of data sets	Percent change in MBC in +N treatments over no-N control
Non-lowland rice crops experiments (Geisseler and Scow, 2014)		
All data sets	107	+15.1
pH in +N treatment: <5	17	-3.1
pH in +N treatment: 5-7	39	+9.3
pH in +N treatment: >7	17	+48.2
Duration of experiment: 5-10 years	18	-22.7
Duration of experiment: 10-20 years	34	+17.2
Duration of experiment: longer than 20 years	55	+30.5
Lowland rice experiments (Geisseler et al., 2017) ^a		
Mineral N	19	+21.0
Mineral NPK	48	+29.0

^aNeither the duration of the experiment nor the annual rate of fertilizer N application influenced the response of MBC to mineral fertilizer N application.

in microbial biomass C was observed because of fertilizer N application. However, in experiments longer than 10 years, a large positive effect of fertilizer use on microbial biomass C was recorded (Table 2). In soils under lowland rice, the duration of the experiment or fertilizer N rate (40–830 kg N ha⁻¹) did not influence the effect of fertilizer use on microbial biomass C.

Long-term use of mineral fertilizers in agro-ecosystems affects the soil microbial biomass, but there are conflicting reports about its influence on the composition of the bacterial community. Guo et al. (2019) did not observe a significant effect of NPK fertilizer application on microbial communities. Some studies also reported that functional diversity, prokaryote and fungal taxonomic diversity were not significantly different between NPK treated plots and no-fertilizer plots. Zhang et al. (2017) studied the effects of fertilizer NPK applications on bacterial diversity and community structure in acidic (pH 5.8), near-neutral (pH 6.8), and alkaline (pH 8.4) soils and found that they did not affect operational taxonomic unit richness or the Shannon diversity index in the alkaline soil. In contrast, both microbiological diversity indices were significantly lower under fertilizer NPK application in acidic and near-neutral soils. In general, greater abundance and diversity of soil micro- and macro-organisms in optimally fertilized fields can be attributed to the accumulation of crop biomass/organic matter in the topsoil, regulation of soil moisture and temperature, and enhanced growth of roots and consequent distribution of root litter and rhizosphere exudates.

Over-use of mineral fertilizers in cropping systems can potentially damage habitats of biological communities and disrupt their functions. Treseder (2008) conducted a meta-analysis of 82 published studies and found that the average decline of microbial biomass under conventional rates of fertilizer N application was 15.4%. In studies that examined different groups of microbes separately, fungi and bacteria were not significantly altered. However, declines in the abundance of microbes as well as fungi were observed in studies with longer duration of fertilizer N over-use.

Effect of fertilizer use on enzyme activity in soil

Microorganisms in the soil synthesize and secrete extracellular enzymes, which contribute to forming and decomposing organic materials. Soil enzymes are sensitive indicators of soil fertility and productivity. Hydrolases are important soil enzymes because of their role in acquiring C, N, and P during primary metabolism. Oxidoreductases facilitate SOM decomposition. The release of biologically available nutrients from organic compounds is controlled by β -1,4-glucosidase (β G), β -1,4-N-acetylglucosaminidase (NAG), leucine aminopeptidase, and acid phosphatase (AP). The intensity and direction of different biochemical processes in the soil can be gauged from enzyme activities involved in the various processes. Microbial demand for nutrients and the response of agroecosystems to environmental changes are also reflected in extracellular enzyme activities.

A meta-analysis based on 107 datasets from 64 non-lowland rice long-term experiments from around the world (Geisseler and Scow, 2014) revealed that activities of protease, urease, and phosphatase were not significantly affected by N fertilization. Phosphatase activity was strongly inhibited by P fertilization. Another meta-analysis based on 64 studies conducted by Chen et al. (2018) showed that N fertilizer application increased the activities of NAG and urease enzymes by 5.5 and 11.6%, respectively. The activity of glycine aminopeptidase was reduced by 33% but no response was observed for the non-specific protease and leucine aminopeptidase. Jian et al. (2016) conducted a meta-analysis of 65 published studies and reported that N fertilizer application in agricultural soils significantly increased β -D-cellobiosidase, hydrolytic C acquisition enzymes, AP, β -1,4-xylosidase, β G, α -1,4-glucosidase, and urease activities by 6.4, 9.1, 10.6, 11.0, 11.2, 12.0, and 18.6%, but decreased peroxidase, oxidative decomposition and phenol oxidase by 6.1, 7.9 and 11.1%, respectively. The effect of fertilizer N on enzyme activities does not show a simple increasing or decreasing trend because the effect varies with soil type, composition of the soil microbial community, and the kind of enzyme reaction. Mineral fertilizers, particularly N, may also indirectly affect activities of soil enzymes through their effect on soil properties like pH.

Fertilizer effects on soil erodibility

Soil erodibility is defined as the susceptibility of soil particles to be detached and become available for erosion by wind, water, or ice. Soil erosion tends to increase with decreased ground cover. Ground cover protects soils from the impact of rainfall and wind, and can also buffer these impacts on soils with weak aggregate stability resulting from low SOM contents. Fertilizer use on farms does not directly influence soil erodibility, but it improves crop cover and aggregate stability through the accumulation of SOM. Under a well-developed crop canopy rainwater is held in the canopy reducing the risk of runoff and soil erosion. In well-fertilized fields, soil also tends to remain in place because of a more extensive root system. For example, in fields with a 2% slope, soil erosion was reduced by more than 9% in the fertilized fields compared to the controls (Bhattacharyya et al., 2015). In studies conducted in seven countries in Asia by the International Board for Soil Research and Management, soil erosion was reduced from 50 to 15 t ha⁻¹ year⁻¹ in the cropped fields with fertilizer application.

Soil physical properties as influenced by fertilizer use

Fertilizers affect soil physical properties primarily by increasing the production of the above-ground and root biomass. The incorporation of organic matter into the soil through crop residues and root biomass leads to improvement in soil structure and water retention capacity, increased infiltration rate, and reduced bulk density. Increased stability of aggregates and soil porosity results in an improved root environment and an increased percentage of macro-aggregates adds to the protection of C and N in the soil.

The use of fertilizers can also change soil physical properties independently of the SOM. Pernes-Debuyser and Tessier (2004) observed that in uncultivated plots to which various fertilizer treatments were applied each year for 70 years, the same small SOM content was recorded but the soil physico-chemical environment and soil physical properties differed strongly among plots. By reducing soil pH and CEC, long-term use of NH₄⁺ fertilizers resulted in degradation of hydraulic properties of the soil. On the other hand, amendments supplying bases such as CO₃²⁻, OH⁻, O²⁻ increased soil pH and CEC as well as its saturation with calcium, which improved soil structural cohesion and water flow properties.

Pore geometry of the soil is modified by fertilizer-induced increased rooting depth and root proliferation in field crops. It affects water transmission characteristics of the soil profile as well as the volume of soil exploited by the roots for water and nutrients. Schjonning et al. (2005) observed that while long-term application of fertilizers and organic manures did not affect soil hydraulic conductivity in the plough layer of a sandy loam soil, improved root growth conditions resulted in increased macro-porosity in the subsoil and near-saturated conductivity for water.

Fertilizers and heavy metal contamination of soils

As P fertilizers are produced from phosphate rock, these can contain substantial amounts of impurities such as cadmium (Cd), lead (Pb), chromium (Cr) and some radionuclides such as uranium (U), thorium (Th), and polonium (Po). There is a wide variation in the heavy metal content of rock phosphates mined in different regions of the world and this is reflected in the P fertilizers produced. Accumulation of heavy metals in the soil adversely affects not only its biological characteristics, but a decrease of soil pH with repeated applications of N and P fertilizers may also facilitate the uptake of heavy metals by plants leading to their entry into the animal and human food chain. The elements applied to soil from P fertilizers mainly accumulate in the surface layer of the soil in bioavailable forms. In acidic soils with coarser textures, such elements are more available than in alkaline soils and those containing large amounts of clay.

Among the heavy metal contaminants in P fertilizers, Cd has received the most attention because of its adverse effects on human health. Depending on the source, single superphosphate can contain between 2 and 40 mg kg⁻¹ Cd; the range in triple superphosphate can be from <10 to >100 mg kg⁻¹ Cd. In most arable soils, P fertilizers constitute the main source of Cd. Several studies have reported a gradual accumulation of heavy metals and Cd in agricultural soils from repeated applications of P fertilizers (Khan et al., 2018). Surveys show that Cd concentrations in topsoils follow the pattern of distribution of fertilizer P applications. Khan et al. (2018) also summarized research reporting increases in radionuclides in soils from long-term P applications. However, there is insufficient evidence for potential long-term effects of P fertilizer use on radiobiological exposure.

If P is applied through ammonium-containing fertilizers such as diammonium phosphate and monoammonium phosphate to neutral or alkaline soils, in which substantial leaching of nitrate-N occurs, the resulting temporary soil acidification of the rhizosphere will mobilize both phosphate and Cd. In contrast, if P is repeatedly supplied to crops through single or triple superphosphate, the induced changes in soil pH will transform Cd into less bioavailable compounds such as cadmium phosphate.

Microbial biomass and enzyme activity in the soil are important indicators for the extent of soil contamination with heavy metals such as Cd and Pb. Significant inhibition of microbial activity in soils contaminated with heavy metals has been reported in several studies. In a meta-analysis based on 671 observations, Aponte et al. (2020) found the activities of arylsulfatase, dehydrogenase, β -glucosidase, urease, acid phosphatase, alkaline phosphatase and catalase were linearly reduced in response to soil contamination with Pb, Zn, Cd and Cu.

Effects of fertilizer use on pollution of surface and ground waters, and climate change

With fertilizer N consumption in East and South Asia surpassing that in North America and Europe (Fig. 1), nitrate pollution of freshwaters and emission of nitrous oxide (N_2O), a long-lived greenhouse gas that accumulates in the atmosphere and contributes to climate change and stratospheric ozone depletion, are now becoming pervasive global problems. The present trends of nitrate-N enrichment of groundwater reflect the legacies of current and past applications of N fertilizers and manures. In addition to the rates of fertilizer N use, factors such as the extent of irrigated area, and annual mean temperature determine the spatial and temporal distribution of nitrate-N in groundwater beneath cropland. Depending upon fertilizer rate, soil nitrate leaching varies from 10 to 30%. One of the most comprehensive assessments of N_2O emissions by Tian et al. (2020) revealed that 43% of the global emissions during 2007–2016 ($7.3 \text{ Tg N year}^{-1}$) were from anthropogenic sources. Direct emissions from N application in agriculture contributed to 52% of emissions linked with anthropogenic activities. Furthermore, up to 71% of the increase in anthropogenic N_2O emissions from the 1980s to 2007–2016 was due to direct emissions from fertilizer applications in agriculture. Important strategies to control nitrate leaching and N_2O emissions from cropland include application of fertilizer N as needed by the crop and synchronizing the time of its application with the pattern of N uptake by crops and irrigation management.

Both N and P contribute to eutrophication, but the main focus has been on P because a single atom of P is required to produce as much phytomass as 16 atoms of N and 106 atoms of C. Even relatively small losses of P from agriculture to surface water bodies may contribute to eutrophication. With the continued application of P in fertilizers and manures to soils under intensive agriculture, there is surplus P in the soil. Although applied P is strongly bonded with soil components and the release of P to the soil solution or runoff water depends on soil type, surplus P in the soil remains the principal cause of its transfer to surface water bodies. Phosphorus in streams and lakes is also delivered through erosion of P fertilized soil. Application of fertilizer P based on soil tests is the best way to reduce the contribution of fertilizers to eutrophication. Omitting P application for some years on soils with a large P content without affecting yield can also be effective. Other measures are maintenance of proper soil pH, incorporation of fertilizer P rather than broadcasting, and proper timing of fertilizer application (Smil, 2000).

Balanced application of fertilizer nutrients and soil properties

Nutrient imbalances in the soil resulting from the application of N, P, and K through fertilizer rates lower or higher than those recommended for a crop can prove deleterious for many soil functions. The adverse effect of imbalanced use of fertilizer N, P, and K on the performance of different crops and cropping systems is well proven. Long-term experiments, in which N, P and K were applied in combination according to soil test results and crop demands led to a greater accumulation of SOM compared to controls where only one or two of the three plant nutrients were applied (Bijay-Singh, 2018). Increases in both SOM and biomass C with balanced nutrient application result from larger root and above-ground crop biomass production. Increased soil microbial functional diversity has also been observed in long-term balanced fertilization experiments. Balanced application of fertilizer N, P, and K can sustain bacterial and archaeal abundances, and bacterial community structure, which contribute to soil fertility. Based on measurements of soil organic C, total N, available P, pH, basic cations, microbial biomass C, microbial biomass size, and crop yield, Belay et al. (2002) inferred that judicious and balanced application of mineral fertilizers can maintain soil productivity and soil quality in the long-term.

Soil properties under integrated nutrient management based on mineral fertilizers and organic manures

Integrated management of mineral fertilizers and organic manures has evolved as an approach to achieve sustainable crop production while ensuring maintenance/improvement of soil fertility and soil health. It is an ecologically sound and economically optimized nutrient management system based on using all possible organic and inorganic nutrient sources in an efficient and integrated manner. Application of organic nutrient sources together with fertilizers, affects soil properties differently from fertilization based only on mineral fertilizers. In general, optimal use of organic and inorganic sources of plant nutrients always leads to improvement in the soil environment for plant growth. Application of organic sources results in the accumulation of SOM which improves physical, chemical and biological properties of the soil. It is not only the nutrients supplied by organic sources but also the other benefits of SOM enrichment that increase crop yields under integrated nutrient management.

Several studies with different crops, particularly from China and India, have reported increased SOM with the application of organic manures together with fertilizers compared to only fertilizer use. Organic sources of nutrients supply C-rich organic compounds to generally C-limited microbial communities in arable soils. Therefore, integrated nutrient management leads to increased soil microbial biomass and enzyme activities in the soil. A meta-analysis of studies conducted on rice and wheat from 1989 to 2016 in South Asia revealed that integrated nutrient management increased average soil organic C by 23.2% in rice and 16.2% in wheat compared with mineral fertilizers (Sharma et al., 2019). The increases in soil organic C were 26.5% in loamy soils and 12.3% in clayey soils. Furthermore, total N, available P, available K, and soil microbial biomass C were positively influenced by integrated nutrient management. In a meta-analysis of 58 long-term experiments in China, Waqas et al. (2020) inferred that the accumulation of soil organic C was greater when organic manures were integrated with balanced and recommended applications of

fertilizer N, P, and K rather than with their unbalanced use. In a study based on 80 long-term experiments in Europe, [Zavattaro et al. \(2017\)](#) reported that application of bovine farmyard manure and bovine liquid slurry in combination with mineral fertilizers increased SOC, respectively by 33.4% ($n = 54$) and 17.3% ($n = 10$) over mineral fertilizer treatment at a similar N rate. Corresponding increases in soil organic N content were 22.4% ($n = 17$) and 15.5% ($n = 11$).

Summary

In the imminent future, mankind will continue to depend upon agriculture and agriculture will continue to depend upon mineral fertilizers. However, mineral fertilizers are chemicals that can potentially alter and sometimes cause irrevocable changes in soil characteristics and may affect normal functioning of the soil. To reduce the risk of reduced yields because of nutrient deficiencies, farmers sometimes apply fertilizers in amounts that are greater than necessary, and effects on soils may be adverse. A major portion of applied fertilizer N converts into organic forms to become a part of SOM and adds to the soil N pool. Depending upon the crop, climate and soil conditions, a substantial amount of N taken up by crop plants is met from the N mineralized from the soil N pool that has been enriched by fertilizer N applications in previous years. The major portion of P applied as water-soluble fertilizers is distributed between the readily- and less readily-available P pools in the soil. Fertilizer K when applied to the soil immediately leads to an increase in K in the soil solution, but most is adsorbed on clay minerals as exchangeable K.

Application of recommended rates of fertilizers to crops promotes increases in both above- and below-ground plant biomass leading to increased organic matter content in the soil. However, excessive use of fertilizer N may accelerate microbial decay of plant litter and organic matter already present in the soil. Application of N fertilizers leads to soil acidification through microbial nitrification of ammonium ions and leaching of nitrate ions from the soil. Although urea is less acidifying than NH_4^+ -based fertilizers because OH^- ions are produced during urea hydrolysis, buffering capacity of the soil to resist changes in pH determines the extent of soil acidification. The application of mineral fertilizers in general positively influences microbial life in the soil, although soil acidification from the use of N fertilizers can have a considerable negative effect on soil organisms.

Through improved crop cover, development of an extensive root system, and increased aggregate stability with SOM accumulation, fertilizer use on farms can reduce soil erodibility. Fertilizer use also improves various soil physical properties and water transmission characteristics of the soil profile. Heavy metals and radionuclides contained as impurities in P fertilizers may contaminate soils. Pollution of groundwater due to nitrate leaching beneath agricultural lands, emission of N_2O , and loss of P from fertilized soils by runoff leading to eutrophication may happen when fertilizers are not managed efficiently. Balanced application of N, P, and K and integrated management of mineral fertilizers with organic manures generally enhance the accumulation of soil organic matter and biological properties of the soil.

Enhancing fertilizer use efficiency is going to be one of the biggest challenges in future agriculture. With increasing fertilizer prices, improved nutrient management will ensure that sustainable and economic food production is achieved without excessive use of fertilizers. Reduced fertilizer application will result in healthy soils with minimal risks of environmental pollution and climate change associated with fertilizer use. The continuous increase in the price of fertilizer nutrients from increasing costs of energy, mining, storage and transport, and declining accessibility of phosphate rock reserves also require efficient and effective fertilizer management. The sustainable and efficient management of fertilizers will be beneficial for the economy, for human health and for the environment.

References

- Aponte H, Meli P, Butler B, et al. (2020) Meta-analysis of heavy metal effects on soil enzyme activities. *Science of the Total Environment* 737: 139744. <https://doi.org/10.1016/j.scitotenv.2020.139744>.
- Belay A, Claassens A, and Wehner FC (2002) Effect of direct nitrogen and potassium and residual phosphorus fertilizers on soil chemical properties, microbial components and maize yield under long-term crop rotation. *Biology and Fertility of Soils* 35: 420–427. <https://doi.org/10.1007/s00374-002-0489-x>.
- Bhattacharyya R, Ghosh BN, Mishra PK, et al. (2015) Soil degradation in India: Challenges and potential solutions. *Sustainability* 7: 3528–3570. <https://doi.org/10.3390/su7043528>.
- Bijay-Singh (2018) Are nitrogen fertilizers deleterious to soil health? *Agronomy* 8: 48. <https://doi.org/10.3390/agronomy8040048>.
- Chen H, Li D, Zhao J, Xiao K, and Wang K (2018) Effects of nitrogen addition on activities of soil nitrogen acquisition enzymes: A meta-analysis. *Agriculture, Ecosystems & Environment* 252: 126–131. <https://doi.org/10.1016/j.agee.2017.09.032>.
- Dhaliwal SS, Naresh RK, Mandal A, Singh R, and Dhaliwal MK (2019) Dynamics and transformations of micronutrients in agricultural soils as influenced by organic matter build-up: A review. *Environmental and Sustainability Indicators* 1–2: 100007. <https://doi.org/10.1016/j.indic.2019.100007>.
- Gardner JB and Drinkwater LE (2009) The fate of nitrogen in grain cropping systems: a meta-analysis of ^{15}N field experiments. *Ecological Applications* 19: 2167–2184. <https://doi.org/10.1890/08-1122.1>.
- Geisseler D and Scow KM (2014) Long-term effects of mineral fertilizers on soil microorganisms – A review. *Soil Biology and Biochemistry* 75: 54–63. <https://doi.org/10.1016/j.soilbio.2014.03.023>.
- Geisseler D, Linquist BA, and Lazicki PA (2017) Effect of fertilization on soil microorganisms in paddy rice systems – A meta-analysis. *Soil Biology and Biochemistry* 115: 452–460. <https://doi.org/10.1016/j.soilbio.2017.09.018>.
- Guo JH, Liu XJ, Zhang Y, et al. (2010) Significant acidification in major Chinese croplands. *Science* 327: 1008–1010. <https://doi.org/10.1126/science.1182570>.
- Guo Z, Han J, Li J, Xu Y, and Wang X (2019) Effects of long-term fertilization on soil organic carbon mineralization and microbial community structure. *PLoS One* 14: e0211163. <https://doi.org/10.1371/journal.pone.0211163>.
- Huang X, Terrer C, Dijkstra FA, Hungate BA, Zhang W, and van Groenigen KJ (2020) New soil carbon sequestration with nitrogen enrichment: a meta-analysis. *Plant and Soil* 454: 299–310. <https://doi.org/10.1007/s11104-020-04617-x>.

- Jian S, Li J, Chen JI, et al. (2016) Soil extracellular enzyme activities, soil carbon and nitrogen storage under nitrogen fertilization: A meta-analysis. *Soil Biology and Biochemistry* 101: 32–43. <https://doi.org/10.1016/j.soilbio.2016.07.003>.
- Khan MN, Mobin M, Abbas ZK, and Alamri SA (2018) Fertilizers and their contaminants in soils, surface and groundwater. *Encyclopedia of the Anthropocene* 5: 225–240. <https://doi.org/10.1016/B978-0-12-809665-9.09888-8>.
- Körschens M, Albert E, Armbruster M, et al. (2013) Effect of mineral and organic fertilization on crop yield, nitrogen uptake, carbon and nitrogen balances, as well as soil organic carbon content and dynamics: Results from 20 European long-term field experiments of the twenty-first century. *Archives of Agronomy and Soil Science* 59: 1017–1040. <https://doi.org/10.1080/03650340.2012.704548>.
- Ladha JK, Pathak H, Krupnik TJ, Six J, and van Kessel C (2005) Efficiency of fertilizer nitrogen in cereal production: Retrospects and prospects. *Advances in Agronomy* 87: 85–156. [https://doi.org/10.1016/S0065-2113\(05\)87003-8](https://doi.org/10.1016/S0065-2113(05)87003-8).
- Ladha JK, Kesava Reddy C, Padre AT, and van Kessel C (2011) Role of nitrogen fertilization in sustaining organic matter in cultivated soils. *Journal of Environmental Quality* 40: 1756–1766. <https://doi.org/10.2134/jeq2011.0064>.
- Li XG, Jia B, Lv J, Ma Q, Kuzyakov Y, and Li FM (2017) Nitrogen fertilization decreases the decomposition of soil organic matter and plant residues in planted soils. *Soil Biology and Biochemistry* 112: 47–55. <https://doi.org/10.1016/j.soilbio.2017.04.018>.
- Pernes-Debuysse A and Tessier D (2004) Soil physical properties affected by long-term fertilization. *European Journal of Soil Science* 55: 505–512. <https://doi.org/10.1111/j.1365-2389.2004.00614.x>.
- Poffenbarger HJ, Barker DW, Helmers MJ, et al. (2017) Maximum soil organic carbon storage in Midwest US cropping systems when crops are optimally nitrogen-fertilized. *PLoS One* 12: e0172293. <https://doi.org/10.1371/journal.pone.0172293>.
- Sattari SZ, Bouwman AF, Giller KE, and van Ittersum MK (2012) Residual soil phosphorus as the missing piece in the global phosphorus crisis puzzle. *Proceedings of the National Academy of Sciences of the United States of America* 109: 6348–6353. <https://doi.org/10.1073/pnas.1113675109>.
- Schjønning P, Iversen BV, Munkholm LJ, Labouriau R, and Jacobsen OH (2005) Pore characteristics and hydraulic properties of a sandy loam supplied for a century with either animal manure or mineral fertilizers. *Soil Use and Management* 21: 265–275. <https://doi.org/10.1111/j.1475-2743.2005.tb00398.x>.
- Sebilo M, Mayer B, Nicolardot B, Pinay G, and Mariotti A (2013) Long-term fate of nitrate fertilizer in agricultural soils. *Proceedings of the National Academy of Sciences of the United States of America* 110: 18185–18189. <https://doi.org/10.1073/pnas.1305372110>.
- Sharma S, Padbhushan R, and Kumar U (2019) Integrated nutrient management in rice–wheat cropping system: An evidence on sustainability in the Indian subcontinent through meta-analysis. *Agronomy* 9: 71. <https://doi.org/10.3390/agronomy9020071>.
- Smil V (2000) Phosphorus in the environment: Natural flows and human interferences. *Annual Review of Energy and the Environment* 25: 53–88. <https://doi.org/10.1146/annurev.energy.25.1.53>.
- Syers JK, Johnston AE, and Curtis D (2008) *Efficiency of Soil and Fertilizer Phosphorus Use: Reconciling Changing Concepts of Soil Phosphorus Behaviour with Agronomic information, FAO Fertilizer and Plant Nutrition Bulletin 18*. Rome: Food and Agriculture Organization of the United Nations.
- Tian D and Niu S (2015) A global analysis of soil acidification caused by nitrogen addition. *Environmental Research Letters* 10: 024019. <https://doi.org/10.1088/1748-9326/10/2/024019>.
- Tian H, Xu R, Canadell JG, et al. (2020) A comprehensive quantification of global nitrous oxide sources and sinks. *Nature* 586: 248–256. <https://doi.org/10.1038/s41586-020-2780-0>.
- Treseder KK (2008) Nitrogen additions and microbial biomass: A meta-analysis of ecosystem studies. *Ecology Letters* 11: 1111–1120. <https://doi.org/10.1111/j.1461-0248.2008.01230.x>.
- Waqas MA, Smith P, Wang X, et al. (2020) The influence of nutrient management on soil organic carbon storage, crop production, and yield stability varies under different climates. *Journal of Cleaner Production* 268: 121922. <https://doi.org/10.1016/j.jclepro.2020.121922>.
- Zavattaro L, Bechini L, Grignani C, et al. (2017) Agronomic effects of bovine manure: A review of long-term European field experiments. *European Journal of Agronomy* 90: 127–138. <https://doi.org/10.1016/j.eja.2017.07.010>.
- Zhang Y, Shen H, He X, et al. (2017) Fertilization shapes bacterial community structure by alteration of soil pH. *Frontiers in Microbiology* 8: 1325. <https://doi.org/10.3389/fmicb.2017.01325>.

Foliar Application of Nutrients

Thomas Eichert^a and Victoria Fernández^b, ^aUniversity of Applied Sciences Erfurt, Erfurt, Germany; ^bSystems and Natural Resources Department, School of Forest Engineering, Universidad Politécnica de Madrid, Madrid, Spain

© 2023 Elsevier Ltd. All rights reserved.

Introduction	383
Barrier properties of leaf surfaces	383
Nutrient uptake through leaves	383
Nutrient uptake through the cuticle	384
Cuticular penetration of lipophilic solutes	385
Cuticular penetration of hydrophilic solutes	385
Role of specific leaf structures	386
Nutrient uptake through stomata	386
Role of external factors	387
Humidity effects	387
Humidity effects on solute concentration on the leaf surface	387
Humidity effects on the permeability of leaf surfaces	388
Overall humidity effects on uptake rates	388
Environmental effects on the barrier properties of leaf surfaces	389
Effects of adjuvants	389
Practical importance of foliar application of nutrients	389
Poor nutrient availability in soils	389
Dry topsoil	389
Periods of large nutrient demand	390
Improving quality of horticultural commodities	390
Bio-fortification	390
Efficiency of foliar nutrition	390
Summary	391
References	391

Abstract

Leaves are protected by the cuticle, stomata and other epidermal structures against the uncontrolled exchange of matter with the environment. Nevertheless, solutes such as mineral fertilizers, biostimulants or pesticides may penetrate the external plant surface. Meanwhile, foliar application of plant nutrients is a common agricultural practice, especially for micronutrients. This chapter summarizes the current knowledge about the barrier properties of both the cuticle and stomata against the penetration of solutes. Practical aspects of foliar fertilization are outlined.

Glossary

Adjuvant Multifunctional and heterogeneous chemicals which may improve spray formulation performance.
Cell wall Hydrated extracellular material made of polysaccharide polymers which surrounds plant cells.
Cuticle Lipidized part of the epidermal cell wall of at least aerial plant organs.
Deliquescence For a certain temperature, relative humidity at or above which solid salt particles begin to become a liquid.
Efflorescence For a certain temperature, relative humidity at or below which a highly concentrated solution begins to crystallize.
Partition coefficient Estimate of the degree of lipophilicity or hydrophilicity of a substance.
Stomata Micrometric size pores in the leaf surface controlling gas exchange.
Trichomes Epidermal cells which grow outwards and may be glandular or not, single celled or multicellular.

Key points

- Plant leaves can absorb solutes such as mineral nutrients and agrochemicals.
- The penetration of leaf surfaces is a passive process.
- Nutrients enter leaves by diffusion.
- Nutrients can be taken up both through the cuticle and through stomata.
- Air humidity is decisive for the rate of foliar absorption.

- Foliar uptake can be fast and efficient.
- Foliar fertilization is of particular importance for micronutrients and calcium.
- Uptake can be improved by adding adjuvants to the foliar formulations.
- Leaf burn can be the result of inadequate application of foliar fertilizers.

Introduction

Under natural conditions plants absorb water and mineral nutrients with their roots. In agriculture, however, foliar nutrient application is a common practice to complement root fertilization, particularly when nutrient availability in soil is limited and nutrients have to be supplied quickly (Eichert and Fernández, 2012). Since most foliar fertilizers can be mixed with pesticides in the sprayer tank, foliar fertilization is often part of routine cultural practices in commercial agriculture.

Foliar application is often preferred to the application of fertilizer to soil when constraints such as soil pH or calcium carbonate limit the availability of nutrients such as zinc (Zn), iron (Fe), and manganese (Mn) or when deep-rooted perennial plants are grown in soils which tend to immobilize nutrients such as potassium (K). The addition of Fe salts to an alkaline or calcareous soil is usually ineffective because soil mineral Fe (often Fe oxides) is poorly available at high pH.

There are several situations in which foliar nutrition can effectively replace or complement (Table 1) soil fertilization. The amount of nutrients that can be supplied through the leaves is limited because above a certain threshold concentration foliar sprays can be toxic. Foliar fertilizer application can meet plant nutrient demand if the requirement for a given nutrient is low, which is the case for microelements (boron (B), copper (Cu), Fe, Mn, molybdenum (Mo), and Zn) or when soils already have adequate fertility, so that only small additional amounts are needed. Multiple foliar applications may be needed to ensure adequate nutrient delivery.

Because of the usually high efficiency of foliar fertilization, this practice is regarded as a promising technique in reducing the amount of chemicals in the environment, thereby promoting sustainable agriculture. While roots are organs specialized in nutrient absorption and root uptake is therefore more or less under physiological control, nutrient uptake by leaves is not. The penetration of leaf surfaces is a passive physical process governed by the laws of diffusion. Therefore, both properties of the applied nutrient solution (e.g., concentration and solubility of the nutrient) and of the leaf surface (e.g., permeability and wettability) affect the uptake process. Several variables influence the effectiveness of foliar-applied fertilizers and may explain the often-reported inconsistency in plant response to foliar nutrition.

Barrier properties of leaf surfaces

To minimize uncontrolled exchange of matter with the environment, leaf surfaces of terrestrial plants are covered with a cuticle. In general, the cuticle can be considered a non-living, hydrophobic skin with low permeability for water, gases and solutes such as mineral fertilizers. To enable CO₂ (carbon dioxide) uptake required for photosynthesis, leaves are equipped with stomata as adjustable apertures occurring in the leaf surface, which optimize the trade-off between CO₂ uptake and water loss. The evolutionary development of the cuticle and stomata as barriers against the uncontrolled exchange of matter was the prerequisite for colonization of the land surface by higher plants, but these barriers do not fully impede the penetration of dissolved nutrients into the leaves. This enables the supply of mineral nutrients to plants by foliar sprays as an alternative or supplement to soil application.

Table 1 Major agronomic and physiology-related conditions promoting the effectiveness of foliar nutrient supply.

Objective	Example
Prevent or cure transient nutrient deficiency	When nutrient demand exceeds root uptake rate
Bypass soil conditions limiting nutrient availability or uptake and accelerate plant response	Unfavorable pH and conditions for root growth, including soil temperature, moisture, and aeration
Increase the buildup of nutrient reserves for storage and successive remobilization the following year	It applies to perennial plants and to nutrients that are phloem-mobile such as nitrogen
Cure or prevent nutrient deficiencies of specific organs that occur despite optimal root uptake due to problems in within-plant distribution of a given nutrient	It applies to nutrients such as Ca which move mainly through the transpiration stream toward the leaves, leading to calcium-related disorders in fruits, e.g., in apple and in tomato
Supply nutrients to plants with deep root systems when soil surface application of fertilizers is almost without effect	It applies to trees, especially for nutrients such as K, Ca, or Mg, whose penetration to root zone layer is sometimes difficult and in general to conditions of absence of irrigation or rain

Nutrient uptake through leaves

Foliar-applied nutrients may penetrate the leaf surfaces through the cuticle and stomata, and it is assumed that both pathways play an important role for the uptake of foliar-applied nutrients. Moreover, cuticle injuries caused by, for example mechanical stress caused by wind, may enable solute uptake. In contrast to root uptake, where physiological processes are involved, the penetration of leaf surfaces by solutes is a purely passive process driven by the concentration gradient between the surface and leaf interior. After penetration of the leaf surface, mineral nutrients will reach the interior of the plant and are thus protected from washing off by rain, for example, and may be utilized by the plant. However, to fulfill their physiological functions most elements have to be taken up into the leaf cells. This is the first point where the uptake process is under physiological control.

The cuticular and stomatal pathways differ fundamentally in their physico-chemical characteristics: The cuticle is solid and rather lipophilic, constituting an effective barrier for the diffusion of large and/or polar, hydrophilic molecules. Ions, in particular, which are the most common constituents of mineral fertilizers, are highly polar. Therefore, the cuticle represents a substantial barrier against the uptake of foliar fertilizers. By contrast, penetration of stomata is assumed to take place in thin water films covering the walls of the pores. Therefore, this absorption pathway can be more efficient for water soluble fertilizers. However, it was shown that this pathway is not present in all stomata, which may contribute to variable rates of uptake.

Fig. 1 summarizes the current view on uptake pathways for foliar applied substances with a special emphasis on water soluble solutes. At low external relative air humidity (RH) the cuticle is an effective barrier against the penetration of applied hydrophilic solutes. The hydrophobic properties of the cuticle prevail because only few water molecules are present (Fig. 1A). This changes at high RH or underneath a drop of aqueous solution. More water molecules are preferentially absorbed in the cuticle by hydrophilic cell wall constituents. Water clusters may form bridges between the leaf surface and leaf interior in which applied solutes may penetrate the cuticle (Fig. 1B). The surfaces of some stomatal pores may be coated by thin water films enabling the diffusion of solutes into the leaves (Fig. 1C).

Nutrient uptake through the cuticle

The surface of plants is covered with the cuticle, a non-living 'skin' protecting plants from excessive water loss by transpiration and uncontrolled exchange of solutes with the environment. The structure and composition of the cuticle vary greatly among plant species, varieties, organs and developmental stages (Fig. 2). The cuticle consists of cell wall material, cutin, a polymer forming a three-dimensional matrix in which waxes are embedded (intra-cuticular waxes), and waxes on the surface of the cuticle (epicuticular waxes). Waxes are predominantly apolar and thus make the cuticle and leaf surfaces rather hydrophobic. Variable amounts of polysaccharide fibrils and pectin lamellae may extend from the epidermal cell wall, binding the cuticle to the underlying tissue. In many plant species cell wall polysaccharides may extend to the outermost leaf surface. In contrast to waxes, polysaccharides are

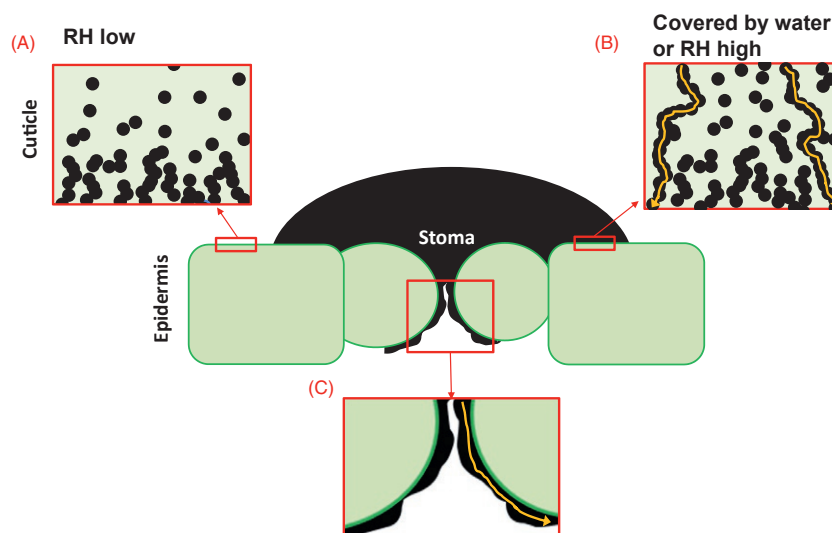


Fig. 1 Pathways for the penetration of hydrophilic solutes across leaf surfaces. Water (black) may be present in the cuticle (A and B) or in some stomata (C), enabling the diffusion of hydrophilic solutes present on the leaf surface into the leaves (orange arrows). At low relative humidity (RH), water in the cuticle is present mostly in the inner regions adjacent to the epidermal cells, whereas in the outermost regions only little water is absorbed (A). Only if RH is high or underneath a drop of liquid water, the cuticle absorbs enough water from the outer side to create continuous aqueous connections crossing the cuticle (B). In some stomata, water clusters may be present and form thin water films creating diffusion pathways for the penetration of solutes by diffusion (C). Since RH within the stomatal pores is generally increased by transpiration, these stomatal water films may exist at lower external RH and in the absence of liquid water drops on the leaves.

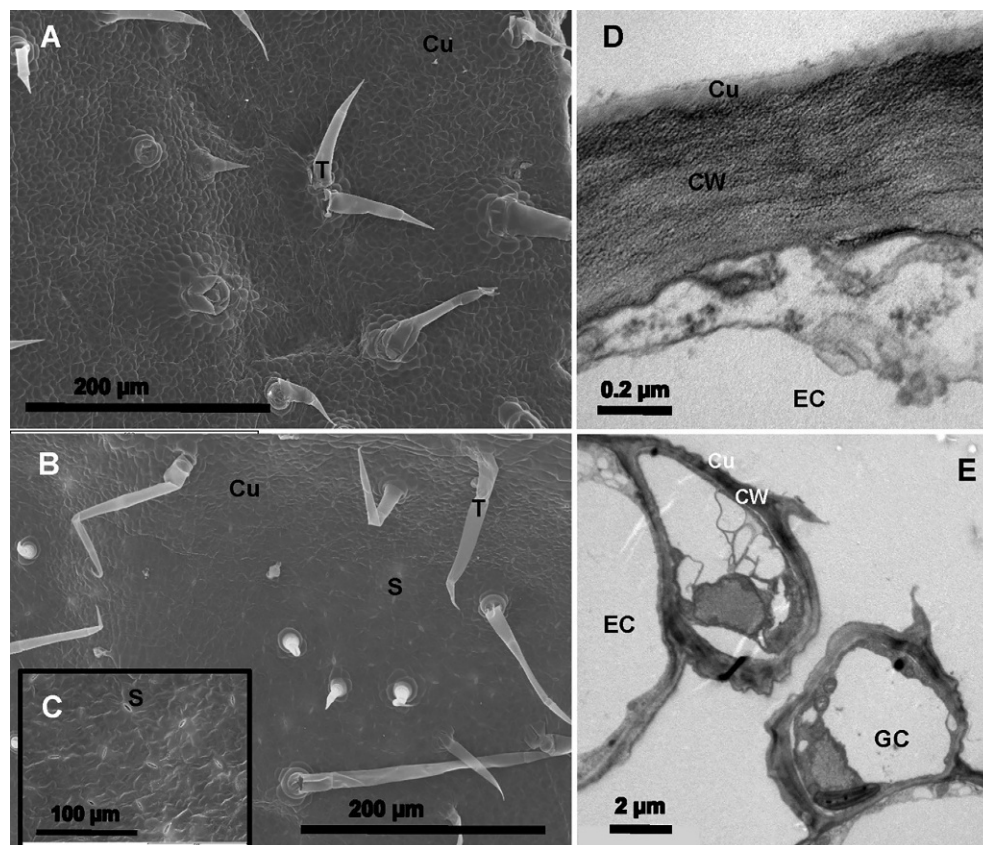


Fig. 2 Topography (A–C) and cross sections of a cucumber leaf observed by scanning (A–C) and transmission (D and E) electron microscopy. Both the upper (A) and lower (B and C) sides are covered with a cuticle (Cu) and have trichomes (T). Many stomata (S) are present in the lower side (B and C). All epidermal cells including trichomes are surrounded by a cell wall (CW) and have a cuticle (Cu) at the surface (D and E). Stomata are formed by two guard cells (GC) surrounded by a cuticle and a cell wall (E).

hydrophilic and may therefore adsorb water molecules, a process which is thought to promote the formation of aqueous pores in the cuticle enabling the penetration of the cuticle by polar solutes (see below).

Cuticular penetration of lipophilic solutes

The more or less hydrophobic nature of the cuticle makes it an effective barrier against penetration by hydrophilic, polar solutes, such as mineral fertilizers. However, lipophilic molecules, such as certain pesticides, may penetrate cuticles at high rates. The penetration of cuticles by lipophilic molecules is described by the solution-diffusion model. According to this model, molecules have to dissolve in the cuticle before they cross the cuticle by diffusion. The rate of penetration of a molecule is a function of its solubility and mobility in the cuticle according to:

$$P = D \times K / \Delta x, \quad (1)$$

where P is the permeance (m s^{-1}), D ($\text{m}^2 \text{s}^{-1}$) is the diffusion coefficient in the cuticle, K is the partition coefficient of the molecule between the external solution and the cuticle as a measurement of solubility, and Δx (m) is the path length of diffusion. The solubility parameter K takes into account the chemical affinity between the permeating molecule and the cuticle, whereas the diffusion coefficient D is a function of the size of the molecule. Because diffusion takes place in the narrow voids in the cuticle, which have diameters in the order of 1–5 nm, the cuticle is highly size selective and larger molecules may be practically excluded from diffusion across it.

Cuticular penetration of hydrophilic solutes

With a few exceptions, such as boric acid or urea, foliar fertilizers are applied as ions, which are very polar and have a very low solubility in the cuticle. Applying the solution-diffusion model to these polar, hydrophilic substances results in extreme underestimation of the observed rates of penetration across the cuticle. Therefore, many researchers assume that hydrophilic solutes penetrate the cuticle by an alternative pathway called ‘polar pores’ or ‘aqueous pores’. It was suggested that these ‘pores’ are formed by clusters of water molecules adsorbed by the hydrophilic moieties of the cuticle, mainly by intra-cuticular cell wall material (Fernández et al., 2017, 2021). It has been observed that the rates of penetration of hydrophilic solutes increase with increasing air

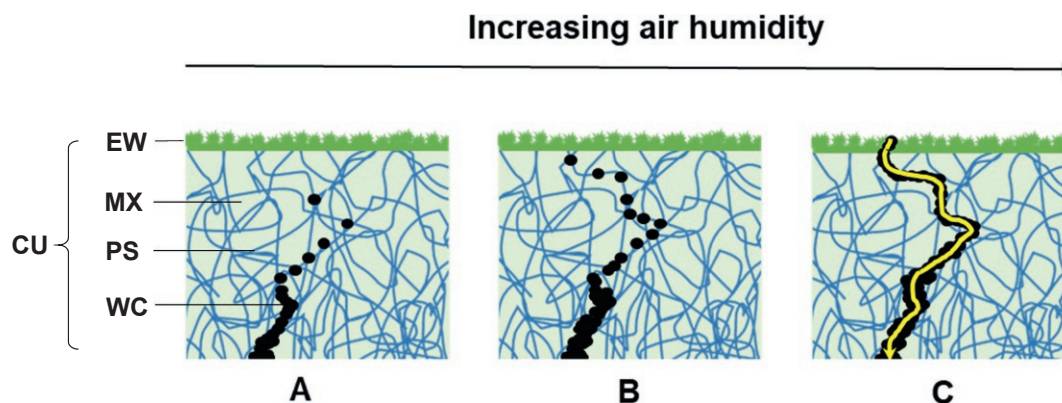


Fig. 3 Model of the formation of aqueous connections ('pores') traversing the cuticle. In this simplified model, the cuticle (CU) consists of a matrix of cutin and embedded waxes (MX) interspersed with hydrophilic domains, mostly provided by polysaccharides (PS) and other hydrophilic constituents. The overlying layer of epicuticular waxes (EW) facing the outer side lacks polysaccharides. Water clusters (WC, black dots) are formed by adsorption of water by the hydrophilic domains. If air humidity is low, water clusters mainly originate from the epidermal cells underneath the cuticle (A). With increasing external air humidity, more water is absorbed by the cuticle from the outer surface (B). At high humidity tortuous connections between the leaf surface and the leaf interior emerge (C). Externally applied solutes may diffuse in these connections through the cuticle (orange arrow in C). For clarity, other water clusters in the cuticle adjacent to the depicted emerging connection are not shown.

humidity, and uptake increases markedly when the air approached saturation with water vapor. These observations gave rise to the 'dynamic pore model' of cuticular penetration of hydrophilic solutes depicted in Fig. 3.

According to this model, water is adsorbed by hydrophilic constituents of the cuticle and forms water clusters in the cuticle. In dry air, water clusters are practically restricted to the lower part of the cuticle where the epidermis provides a constant water source. Small amounts of water only will be absorbed by the outer cuticle, and hence less functional 'pores' traversing the cuticle will exist. With increasing air humidity, more water is absorbed by the cuticle eventually creating continuous connections between the leaf surface and the epidermis. The permeability of cuticles is thus not constant but depends on environmental conditions.

From an ecophysiological perspective, the variable permeability of the cuticle makes perfect sense: under dry conditions, the small water content of the cuticle minimizes the permeability of the cuticle, not only for solutes but also for water. Therefore, water loss by the plant is reduced. If air is saturated with water, transpiration ceases and the flow of water and solutes into leaves is interrupted. Under these conditions, many plants actively pump water from the roots to the leaves where drops of xylem sap are exuded on the tips or edges of leaves, a phenomenon called guttation. This occurs mainly when transpiration is suppressed and shows, that the cycling of water and nutrients in the plants through the xylem and phloem has to be sustained. The considerable water permeability of cuticles at high air humidity will facilitate transpiration, thereby sustaining water and nutrient transport in the xylem.

From a practical perspective, the variable permeability of cuticles has to be taken into account when applying foliar fertilizers or pesticides. The water absorption ('swelling') of the cuticle after longer periods of high air humidity or even rainfall increases rates of penetration. On the one hand, this may be beneficial when large rates of uptake have to be achieved. On the other hand, uptake rates higher than usual may pose the risk of leaf damage. In this case, application doses should be reduced accordingly.

Role of specific leaf structures

A range of studies showed that the cuticle covering leaf veins and trichomes may represent preferential uptake routes for nutrients applied to the leaves. It was shown that calcium (Ca) was preferentially absorbed by veins of beech leaves and that the cuticle above veins differed from the surrounding cuticle in structure and composition (Bahamonde et al., 2018). In sunflower leaves non-glandular trichomes are important entry sites for foliar-applied Zn. To date the reason for the higher cuticular permeability above these leaf structures and its ecological significance is not well understood.

Nutrient uptake through stomata

In terrestrial plants, the stomata are the sites of exchange of gases (mainly CO₂, O₂ (oxygen)) with the atmosphere. Their number per mm² of leaf surface varies between about 20 in succulents, 100–200 in most annual species, and more than 800 in certain tree species. In terrestrial plants, stomata are usually more abundant (most annual species) or confined (many tree species) to the lower (abaxial) leaf surface. By contrast, in leaves of aquatic plants like sacred lotus (*Nelumbo nucifera* Gaertn.) they may only occur on the upper side.

Stomata are required to enable photosynthesis. The cuticle itself can hardly be penetrated by gases like CO₂ which is required to convert solar energy into chemical energy (sugars). Therefore, stomata as 'holes' in the leaf surface are required. The aperture of stomata can be adjusted. When open they enable the uptake of CO₂ into leaves by diffusion. However, open stomata also inevitably enable the escape of H₂O (water) from the leaf interior. Thus, photosynthesis is always combined with water losses, which have to

be compensated for by water uptake from the soil. The rate of water loss by transpiration depends on environmental conditions, mainly on air humidity, temperature, and wind. When rates of water losses are too high, plants react by closing stomata.

At first sight, stomata may appear as the designated entry points for foliar applied substances because the sprayed solution should simply infiltrate the leaves through the stomatal pores. However, both practical experience and theoretical considerations make clear that this is not the case. If stomata could be infiltrated spontaneously by solutions on the leaf surface, this would happen during every precipitation event, at least when stomata are present on the upper leaf surface. Leaves infiltrated by water can be easily identified with the naked eye because of the darker color of the infiltrated tissue. Moreover, if water could easily infiltrate the leaf interior and occupy all the normally air-filled intercellular spaces, diffusion of CO_2 to the sites of photosynthesis in the mesophyll cells would be strongly restricted. After rainfall, the rates of photosynthesis would decrease dramatically until the infiltrated water had transpired, which would take at least several hours. Apparently, stomata are protected against the infiltration of aqueous solutions due to their specific architecture (Schönherr and Bukovac, 1972).

Although both common sense and scientific considerations exclude the possibility of spontaneous infiltration of stomata by water and solutions, many scientific studies have shown clear evidence of a promoting effect of stomata on foliar uptake. In many studies rates of foliar uptake into hypostomatous leaves, i.e., leaves with stomata only on the lower leaf sides, were higher through the lower (abaxial) leaf sides than through the upper (adaxial) side. Other studies reported positive correlations between rates of uptake and stomatal density or stomatal aperture.

The reason for this apparent contradiction is that stomatal uptake of foliar applied fertilizers was assumed to be possible only when the complete solution, i.e., the dissolved solute (= the fertilizer salt) plus the solvent (= water) entered the leaf. This uptake process is also known as 'mass-flow'. Infiltration of stomata by mass-flow can be induced when the surface tension of solutions is lowered by the addition of effective surfactants, such as certain organosilicones.

Meanwhile, it is clear that the assumption of infiltration of stomata by mass-flow as a prerequisite of stomatal uptake of normal solutions was wrong. It has been shown that uptake of solutes such as fertilizers occurs by diffusion, not by mass-flow. The final evidence for this mechanism was the observation that hydrophilic particles (43 nm diameter) suspended in water entered leaves through stomata by diffusion along the walls of the pore (Eichert et al., 2008). However, it was also shown that only a small portion, usually less than 10% of stomata covered by a foliar-applied droplet of solution, are penetrated (Eichert and Goldbach, 2008). It is therefore likely that the penetrability of stomata is not a native, *a priori* property of stomata but that it is acquired *a posteriori*. This possibly occurs by modifications of the, initially rather hydrophobic, pore wall cuticle, due to the effect of, for example, deposited hygroscopic particles, microbes growing in the stomata chamber or salts ascending the pore, rendering it more wettable (Eichert and Burkhardt, 2001). Although not all stomata present on the leaf surface are involved in foliar solute uptake, the contribution of this penetration pathway may be very important and, depending on the type of solute, be as great or even greater than the cuticular route (Eichert and Goldbach, 2008).

Role of external factors

Humidity effects

Humidity effects on solute concentration on the leaf surface

Foliar-applied solutions are usually rather diluted and thus not in equilibrium with the water potential of the atmosphere. They will therefore start to evaporate as drops on the leaves or even before they reach the leaf surfaces. This will result in a reduction of droplet volume accompanied by an increase in solute concentration. During equilibration with the atmosphere, solutions may dry out completely or remain liquid. The final state mainly depends on external air relative humidity (RH) and the type of solute.

When humidity increases, previously dried residues can absorb water from the atmosphere and re-dissolve in this water. For each pure solute, there is a certain threshold RH above which this process starts. This threshold is called deliquescence relative humidity (DRH) or the deliquescence point. The value of DRH is specific for a given solute and is affected by temperature.

If the RH is below the DRH, solutions can completely dry out during equilibration with the atmosphere. This process is called efflorescence. Efflorescence often occurs at values of RH well below the DRH (Fernández et al., 2020). In contrast to DRH, the efflorescence humidity (ERH) depends on the actual conditions and therefore no defined ERH values exist. Efflorescence of solutions below DRH is the result of kinetic inhibition. This can occur in the absence of crystallization nuclei, which normally trigger the crystallization process. These nuclei can for instance be, other salt crystals, small dust particles or specific surface structures.

Both DRH and ERH are important parameters controlling foliar uptake processes: ERH controls the initial drying of fertilizer spray drops while DRH affects the re-hydration of droplet deposits due to e.g., dew, fog or high environmental RH (Fernández et al., 2020). If RH is at, or just above, the DRH or ERH, solute concentrations are maximal. With increasing RH above the DRH, concentrations decrease and theoretically approach zero when RH increases to saturation.

The concentration gradient between the solution on the leaf surface and the leaf interior is the driving force of foliar uptake, therefore, RH will have a strong effect on foliar penetration rates. Moreover, at RH below ERH uptake will cease due to immobilization on the leaf surface. Uptake will resume, however, once RH increases again above DRH, e.g., during the night. The weather conditions and the diurnal rhythm of RH and temperature will thus result in fluctuating concentrations on the leaf surface, making the prediction of rates of uptake difficult under field conditions. Values of DRH and ERH differ considerably between fertilizer salts of the same element (Table 2). The appropriate choice of fertilizer salt may thus help to avoid excessive uptake rates and thus leaf burn.

Table 2 Nutrient carriers normally used in foliar spray formulations.

<i>Element</i>	<i>Compounds</i>	<i>DRH</i>
N	Urea	
	(NH ₄) ₂ SO ₄	63
	NH ₄ NO ₃	
P	K ₂ HPO ₄	92
	KH ₂ PO ₄	
	H ₃ PO ₄	95
	Ca(H ₂ PO ₄) ₂	
K	KCl	86
	KNO ₃	95
	K ₂ CO ₃ · 2 H ₂ O	44
	KH ₂ PO ₄	95
	K ₂ HPO ₄	92
	K ₂ SO ₄	98
Ca	CaCl ₂ · 6 H ₂ O	33
	Ca(NO ₃) ₂ · 4 H ₂ O	53
	Ca-acetate	100
	Ca-propionate	95
Mg	MgCl ₂ · 6 H ₂ O	33
	Mg(NO ₃) ₂ · 6 H ₂ O	56
	MgSO ₄ · 7 H ₂ O	90
	MgSO ₄ · 7 H ₂ O	90
S	B(OH) ₃ , Borax (Na ₂ B ₄ O ₇), Na-octoborate (Na ₂ B ₈ O ₁₃)	
B	FeSO ₄ , Fe(III)-chelates	
Fe	MnCl ₂ , Mn(NO ₃) ₂ , MnSO ₄	
Mn	ZnCl ₂ , ZnSO ₄ , Zn(II)-chelates, ZnO	
Zn	CuCl ₂ , CuSO ₄ , Cu-Chelates	
Cu		

For selected compounds values of deliquescence relative humidity (DRH) are given.

Based on data from Fernández V, Sotiropoulos T and Brown P (2013) *Foliar Fertilization: Scientific Principles and Field Practices*. First edn. IFA: Paris, France.

Humidity effects on the permeability of leaf surfaces

Increasing RH leads to cuticular water sorption as explained above, and thus facilitates the permeability for polar solutes. It has been shown that an increase of RH from 50% to saturation may increase the permeability of isolated cuticles for ions by two orders of magnitude (Fernández and Eichert, 2009). Stomatal aperture may also react to air humidity. In dry air, plants may close stomata to avoid excessive water loss. Control and adjustment of stomatal aperture, however, is a complex process governed by a range of additional interacting factors, such as soil water availability, plant water status or irradiation. Therefore, no strict relationship between RH and aperture exists and the stomatal penetration pathway will thus probably be less RH dependent than the cuticular pathway.

Overall humidity effects on uptake rates

The RH affects both the solute concentration of fertilizer spray drops deposited onto leaf surfaces and permeability of the leaf surface, at least of the cuticle. Below ERH (drying of drops) or DRH (re-wetting of dry residues) foliar fertilizers are dry and practically no uptake can be expected. Immediately above ERH or DRH solutions are at saturation, solute concentrations are maximal and so is the driving force of uptake. When RH increases to saturation, solute concentrations will decrease due to dilution by the water absorbed from the atmosphere. However, with increasing RH leaf permeability will increase. Therefore, RH has a counter effect on the driving force of uptake and the permeability of the leaf. The resulting rates of uptake can be predicted, in a similar way to Ohm's law, from the product of both parameters, driving force and permeability, according to:

$$F = \Delta c \times P, \quad (2)$$

where F is flux (mol m² s⁻¹), a measure of rates of uptake, Δc (mol m⁻³) is the concentration difference across the leaf surface responsible for the driving force, and P is the permeance (m s⁻¹, see Eq. (1)).

This simple equation supports the conclusion that with high RH (close to or at saturation), when leaf permeability is very high, rates of uptake will nevertheless be low because of the small solute concentrations on the leaf. Therefore, the widespread assumption that high RH will result in the highest rates of foliar uptake is not necessarily true. Rather moderately high RH, at best just above DRH or ERH of the fertilizer salt may result in the largest uptake rates.

Environmental effects on the barrier properties of leaf surfaces

The prevailing environmental conditions during plant growth have a direct influence on the structural and chemical composition of the leaf surface. They may for example, affect cuticle thickness or amount and composition of epicuticular waxes. This will influence wettability and permeability, and thus rates of uptake from foliar sprays. Shading during plant development may decrease the amount of wax per leaf area, whereas high temperatures can modify the morphology and composition of epicuticular waxes. Relative humidity affects the amount of wax per leaf area and wax crystal morphology.

In general, young, partially expanded leaves and fruits are regarded as more penetrable than mature plant parts. This may be linked to, for example, the presence of trichomes, increased stomatal densities and stomatal functionality. However, the barrier function of the surfaces of older plant surfaces can be modified by the erosion of waxes, mechanical injuries and pathogens.

Nutrient deficiencies may also affect leaf structure and function, also including the leaf surface. Nutrient deficiencies may cause changes in trichome or stomatal density and morphology. For example, compared to Fe-sufficient peach and pear leaves, Fe-chlorotic leaves were found to have decreased stomatal pore size and conductance, in addition to lower leaf cuticle weight per unit surface and concentration of soluble cuticular lipids (Fernández et al., 2008). Effects on leaf surface properties and permeability were also reported in relation to deficiencies of other nutrients such as K, Mn, B, phosphorus (P) and Zn. In most cases, nutrient deficiencies decreased the permeability of leaf surfaces and thus rates of uptake. Hence, existing nutrient deficiencies may hamper their correction by foliar sprays. This supports one of the basic principles of plant nutrition: it is better to prevent than to cure nutrient deficiencies!

Effects of adjuvants

Adjuvants (literal translation: 'helpers') are chemicals added to foliar-applied solutions to facilitate the penetration process. They may have different modes of action: 'surfactants' reduce the surface tension of the spray solution with the aim of increasing the contact area between liquid drops and leaf surface. 'Stickers' aim at avoiding the run-off and drip-off of sprays. 'Humectants' are added to keep the active ingredients moist and thus available for uptake. 'Plasticizers' increase the plasticity of the cuticle, thereby enhancing the penetration rates of solutes.

Selection of an appropriate adjuvant is often crucial for improving penetration rates, and this will depend on a range of factors such as properties of the target surface, the compound to be applied and its concentration. However, surfactants may also have an effect on leaf physiology for example by increasing leaf transpiration or altering the structural or chemical composition of the cuticle.

Practical importance of foliar application of nutrients

Foliar fertilization is of increasing importance in agricultural production worldwide. The application of foliar sprays, either as a supplement for root applications or under conditions of limited nutrient availability in the soil, can help to preserve crop yield and quality with low environmental impact. While foliar fertilization has been traditionally used to correct nutrient deficiencies, there is an increasing trend to using nutrient sprays to prevent deficiency symptoms, particularly for elements with limited phloem mobility such as Ca or micronutrients. In commercial agricultural production, foliar fertilization is of particular practical relevance under the following conditions:

Poor nutrient availability in soils

Many soil properties can limit element solubility and the uptake of nutrients by plant roots. For example, due to the poor availability of P in many soils, P is a widespread limiting factor for crop production worldwide. On the other hand, the limited solubility of several micronutrients such as Fe, Mn, Cu and Zn is also a common problem in calcareous soils with high pH, which is exacerbated by drought. The application of micronutrient sprays to control or avoid the occurrence of micronutrient deficiencies under conditions of limited nutrient availability in the soil is one of the most important practical uses of foliar fertilization worldwide.

Foliar Fe sprays may induce the re-greening of chlorotic leaves, and restore some leaf physiological properties such as the rate of photosynthesis or transpiration. Manganese and Zn foliar sprays can also be very effective for overcoming Mn and Zn deficiency in crop plants.

However, micronutrient sprays may have a low efficacy in terms of limited distribution to different plant parts and metabolic functionality.

Dry topsoil

In semi-arid regions, the lack of available water in the topsoil often reduces nutrient availability during the growing season, even though water may still be available in the subsoil. Under conditions of water and nutrient shortage, crop yield and quality can be ensured by complementing standard root fertilization with foliar sprays. Foliar sprays may also contribute to plant drought

tolerance, but physiological effects caused by drought (e.g., stomatal closure) may decrease the rate of foliar absorption of foliar fertilizers.

Periods of large nutrient demand

During the reproductive stage, root nutrient uptake may be compromised by the transport of carbohydrates to reproductive organs (sink competition). Foliar fertilizer application during such periods can compensate for reduced root nutrient uptake during these periods. Foliar nitrogen (N) sprays have been found to be beneficial for legumes when there is sink competition for carbohydrates between developing seeds and root. In cereals, grain quality and yield can be increased by late-season foliar N applications.

Boron plays a key role in plant reproduction, thus it is important to ensure an adequate plant B status, and foliar sprays can be useful for this purpose. In annual crops, B sprays before flowering have proven effective. In fruit-tree species with high B phloem mobility, foliar B sprays supplied in autumn can successfully increase the B concentration of reproductive and vegetative organs, and improve fruit setting in the next growing season.

Foliar sprays of nutrients such as N, K, B and micronutrients are often supplied to fruit crops during the growing season to improve flowering, fruit set and plant development, for example.

Improving quality of horticultural commodities

The occurrence of Ca-related disorders in fruits and vegetables is a major problem affecting the marketability, storability and shelf-life of horticultural commodities. Such damage develops during fruit growth and is caused by localized Ca deficiency in plant organs with generally low transpiration. Calcium has poor phloem mobility and its supply to the fruits depends on Ca delivery via xylem flow. Because of the low transpiration rate of fruits which have low stomatal densities or non-functional stomata, little Ca will reach these organs even under good Ca availability conditions for root absorption. Subsequently, Ca sprays applied to the aerial part of fruit trees are often supplied during the growing season to prevent the occurrence of Ca-related disorders and improve fruit quality and storability.

Bio-fortification

Bio-fortification is the enhancement of the nutritional value of food crops for human consumption. Foliar sprays applied alone or in combination with root fertilization can increase the concentration of micronutrients in foods and alleviate micronutrient deficiencies in humans, as reported after micronutrient enrichment of grain crops or rice (*Oryza sativa* L.). The bio-fortification of staple food crops or horticultural commodities with foliar sprays of beneficial elements such as selenium (Se) or iodine (I) can also be achieved.

Bio-fortification has been intensively studied in the case of Zn. Almost 50% of soils in cereal growing areas in the world have been found to be deficient in Zn, either because Zn concentrations are inherently low or Zn availability is reduced by high soil pH. In many cases, foliar Zn application has been found to be superior in increasing grain Zn concentration in cereals compared to soil application (Cakmak and Kutman, 2018).

Efficiency of foliar nutrition

Theoretically, foliar fertilization is more target-oriented and environmentally-friendly than soil treatments because nutrients are delivered directly to sink plant organs and there is less risk of environmental contamination. Plants often have a rapid response to nutrients supplied as foliar sprays compared to root fertilization. However, the efficiency of foliar sprays can be variable because of the complexity surrounding the foliar uptake process. Consequently, complete reliance on foliar sprays for meeting plant nutrient demand in full cannot be achieved in commercial agriculture, because this may compromise crop yield and quality, particularly in the case of macronutrients which are taken up in large amounts by plants.

The efficacy of foliar nutrient sprays is determined by many environmental, physico-chemical and physiological factors associated with the plant, the environment and the spray formulation. An array of problems limiting the effectiveness of foliar sprays may arise, such as:

1. Limited leaf wetting and retention of sprays when treating unwettable, even water-repellent leaves, or if applying spray solutions without appropriate surfactants to unwettable plant surfaces.
2. Washing off by rain following the application of nutrient sprays.
3. Low rates of nutrient uptake due to low foliar permeability as affected by, for example, plant physiological factors or environmental conditions.
4. Rapid drying of spray solutions, particularly under low RH and high temperature conditions, resulting in poor nutrient uptake.
5. Low mobility of the nutrients sprayed within the plant following foliar application
6. Limitations related to the concentration range of foliar-applied macronutrients for meeting plant demand without inducing phytotoxicity (leaf burn).

7. Foliar fertilization may support spore germination and colony growth of pathogens
8. Potential transient nutrient imbalances when single nutrient sprays are supplied to the plants.

Leaf burn and defoliation are major risks when applying foliar sprays. Different degrees of damage may occur, ranging from the appearance of necrotic spots to complete defoliation. This can be the result of excessive nutrient concentrations, the use of hygroscopic active ingredients or the presence of surfactants, which significantly increase the rate of penetration of the applied nutrients and are often phytotoxic when applied at large concentrations. Environmental conditions during application are also important. Leaf burn is more likely to occur when solutions are sprayed during periods of high irradiation and high temperature. Therefore, application in the morning or evening is usually recommended.

Summary

Although leaves are not designed to be natural sites of nutrient uptake, fertilizers can be taken up through the leaf surfaces. Foliar applied substances penetrate the external surface either through the cuticle or through stomata. In contrast to root uptake, foliar uptake is a passive process solely ruled by the laws of diffusion. Therefore, care has to be taken to avoid the negative effects of foliar application such as leaf burn.

In most cases leaf damage results from excessive uptake. This risk can be reduced by using adequate nutrient concentrations and selecting appropriate compounds. Salts differ in their ability to absorb and bind water from the atmosphere. This has a strong effect on the feasibility of nutrient absorption by the leaf surface because uptake requires that the substances are present in dissolved form. Efflorescence, the drying of foliar sprays, occurs at different levels of relative air humidity (RH). The level of RH at which solutions dry out is named efflorescence relative humidity (ERH). Likewise, deliquescence, the re-wetting of salt residues, occurs after RH has exceeded a defined threshold; the deliquescence relative humidity (DRH). Values of DRH and ERH differ considerably between salts, and the appropriate choice of fertilizer salt may help to avoid excessive rates of uptake and thus leaf burn.

Foliar nutrient application is of great importance whenever nutrient availability in the soil is the limiting factor. This is a typical problem in alkaline soils where micronutrients such as Fe or Mn are immobilized. Application of calcium sprays in fruit production is another important field of application.

Foliar nutrient application is a target-oriented and environmentally-friendly fertilization strategy that reduces the risk of environmental contamination. It allows faster nutrient delivery to plants compared to root fertilization. In the case of micronutrients, but also for calcium supply to fruits, foliar fertilization is already an established fertilization strategy with increasing practical importance.

References

- Bahamonde HA, Gil L, and Fernández V (2018) Surface properties and permeability to calcium chloride of *Fagus sylvatica* and *Quercus petraea* leaves of different canopy heights. *Frontiers in Plant Science* 9: 494.
- Cakmak I and Kutman UB (2018) Agronomic biofortification of cereals with zinc: A review. *European Journal of Soil Science* 69: 172–180.
- Eichert T and Burkhardt J (2001) Quantification of stomatal uptake of ionic solutes using a new model system. *Journal of Experimental Botany* 52: 771–781.
- Eichert T and Fernández V (2012) Uptake and release of elements by leaves and other aerial plant parts. In: Marschner P (ed.) *Marschner's Mineral Nutrition of Higher Plants*, pp. 71–84. Oxford: Academic Press.
- Eichert T and Goldbach HE (2008) Equivalent pore radii of hydrophilic foliar uptake routes in stomatous and astomatous leaf surfaces—further evidence for a stomatal pathway. *Physiologia Plantarum* 132: 491–502.
- Eichert T, Kurtz A, Steiner U, and Goldbach HE (2008) Size exclusion limits and lateral heterogeneity of the stomatal foliar uptake pathway for aqueous solutes and water-suspended nanoparticles. *Physiologia Plantarum* 134: 151–160.
- Fernández V and Eichert T (2009) Uptake of hydrophilic solutes through plant leaves: Current state of knowledge and perspectives of foliar fertilization. *Critical Reviews in Plant Sciences* 28: 36–68.
- Fernández V, Eichert T, Del Río V, López-Casado G, Heredia-Guerrero JA, Abadía A, Heredia A, and Abadía J (2008) Leaf structural changes associated with iron deficiency chlorosis in field-grown pear and peach: Physiological implications. *Plant and Soil* 311: 161–172.
- Fernández V, Bahamonde HA, Peguero-Pina JJ, Gil-Pelegrín E, Sancho-Knapik D, Gil L, Goldbach HE, and Eichert T (2017) Physico-chemical properties of plant cuticles and their functional and ecological significance. *Journal of Experimental Botany* 68: 5293–5306.
- Fernández V, Pimentel C, and Bahamonde HA (2020) Salt hydration and drop drying of two model calcium salts: Implications for foliar nutrient absorption and deposition. *Journal of Plant Nutrition and Soil Science* 183: 592–601.
- Fernández V, Gil-Pelegrín E, and Eichert T (2021) Foliar water and solute absorption: An update. *The Plant Journal* 105: 870–883.
- Schönherr J and Bukovac MJ (1972) Penetration of stomata by liquids. Dependence on surface tension, wettability, and stomatal morphology. *Plant Physiology* 49: 813–819.

Mulches and cover crops part I: Types

KK Bandyopadhyay^a, CL Acharya^b, and KM Hati^c, ^aDivision of Agricultural Physics, ICAR-Indian Agricultural Research Institute, New Delhi, India; ^bFormer Director, ICAR-Indian Institute of Soil Science, Bhopal, India; ^cDivision of Soil Physics, ICAR-Indian Institute of Soil Science, Bhopal, India

© 2023 Elsevier Ltd. All rights reserved.

This is an update of C.L. Acharya, K.M. Hati, K.K. Bandyopadhyay, MULCHES, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 521–532, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00250-2>.

Introduction	392
Types of mulches	393
Organic mulches	393
Organic residue mulch	393
Wastes from cattle sheds as mulch	394
Spreading of farm yard manure (FYM) as mulch	394
Cover crops	394
Other organic mulches	395
Inorganic mulches	396
Plastic mulch	396
Gravel mulch	397
Crushed rock	398
Other materials used as mulch	398
Methods of mulch application	398
Limitations of mulching and cover crops	398
Conclusions	399
References	399

Abstract

Mulches, both organic and inorganic favorably ameliorate the hydrological regime of soil, and are gaining significance in the context of climate change. In-situ mulching practiced under conservation agricultural systems can replace crop residue burning. Similar to mulches, growing cover crops can protect soils from erosion and enhance soil fertility. However, mulching has some limitations such as the disposal of non-biodegradable plastic mulch, competing demand for crop residues by other enterprises, lack of machinery and knowhow among the farmers for field operations, particularly in the case of developing countries. Appropriate policy decisions may be taken for addressing these limitations.

Introduction

Mulch is a layer of dissimilar material separating the soil surface from the atmosphere, and its application aims to modify the soil environment beneficially (Acharya et al., 2005, 2018). Mulch may be organic (crop residues, stubble mulch etc.) or inorganic (plastic sheets, gravels etc.). It may be generated in-situ by crop residues left on the soil surface or it may be living such as a perennial legume mulch, or it may be grown or produced ex-situ for example, straw, sawdust, plastic products etc. The practice of mulching has been widely used as a management tool for centuries in many ancient civilizations. The Great Plains region of USA, which was once a dust bowl, has become a major grain producer with the adoption of conservation tillage whereby crop residues or flattened straw mulch are left on the surface. The idea of using plastics as a mulch originated in Hawaii, where it was successfully used in pineapple cultivation.

Cover crops provide ground cover, and thus growing cover crops is a principal agronomic approach to erosion control and enhancing soil fertility. The merits of cover crops are well established. There is, for example, a wide choice of species and planting techniques that allow adequate plant establishment in most temperate or tropical settings, thus making cover cropping feasible under most soil, climatic, and cultural conditions. The goals of cover cropping have varied historically between single- and multi-purpose, and may range from single crop stands grown principally for ground protection, to mixed stands of food or feed species. The latter are, in large part, established to produce a crop (e.g., grain or hay in north-temperate climates and pulses in tropical climates), but eventually all systems result in soil quality improvement that may take the form of: (1) physical conditioning; (2) nitrogen accumulation; and (3) organic matter enhancements to the benefit of any arable terrain, be it sloping or flat land.

For example, the Ministry of New and Renewable Energy (MNRE, 2009), Government of India has estimated that about 500 Mt. of crop residues are generated every year and of these 91–141 Mt. of residues are burnt. The rice (*Oryza sativa* L)–wheat (*Triticum*

aestivum L.) cropping systems in the north western (NW) states of India produce about 34 million tonnes of rice residues of which Punjab alone contributes about 65%. The mechanized harvesting and threshing of rice using combine harvesters is a common practice in NW India. In the process, residues are left behind the harvesters in a narrow strip (windrow) in the field. Disposal or utilization of the leftover residue in the short window of 10–20 days for the timely planting of a wheat crop is a difficult task. Therefore, the farmers commonly opt to burn the rice residues because of lack of access to user-friendly, cost- and time-effective options. It is estimated that in the NW states of India about 23 million tonnes of rice residues are burned annually (NAAS, 2017). This adversely affects air quality in adjacent areas resulting in various health hazards in addition to degrading soil health. It is estimated that burning 1 tonne of rice residue releases 13 kg particulate matter, 60 kg CO (carbon monoxide), 1460 kg CO₂ (carbon dioxide), 3.5 kg NO_x (nitrogen oxides), 0.2 kg SO₂ (sulfur dioxide). The black carbon emitted during residue burning warms the lower atmosphere and it is the second most important contributor to global warming after CO₂. The emission of high levels of PM_{2.5} and PM₁₀ in the air causes chronic diseases like cardiopulmonary disorders, irrecoverable lung capacity or asthma in human populations. If part of this residue was utilized as mulch, it would contribute to improving the soil, environment and human health.

This chapter reviews the types of mulches and cover crops and methods of their application to soil and limitations to the adoption of mulches and cover crops by the farming community. A companion chapter deals with the role of mulches and cover crops on soil health and climate resilient agriculture.

Types of mulches

Soil itself forms a mulch called “dust or soil mulch” owing to the formation of a dry layer at the soil surface, the effectiveness of which depends on its texture (Prihar et al., 1968), thickness and porosity (Acharya and Prihar, 1969), nature of cations in the exchange complex and prevailing climatic conditions (Acharya and Abrol, 1978; Acharya et al., 1979). Apart from soil mulch there are different types of materials ranging from natural plant residues, to gravel to various industrial products such as paper or plastic sheets, which are used as mulches.

Organic mulches

Organic mulches include live mulch, cover crops, in situ mulching under conservation agriculture and ex situ mulches using the materials derived from natural sources, such as field crops, trees, or natural bio-waste products.

Organic residue mulch

These include hay, straw, wood chips, bark, sawdust, wood shavings, peat moss, compost, wastes from cattle sheds, farmyard manure (FYM), grass clippings, paper, and materials such as cocoa shells (*Erythroxylum coca* Lam.) and buckwheat (*Fagopyrum esculentum* Moench) hulls. Organic materials for mulching should be stable and remain in place at least through the growing season of an annual crop and longer for perennial crops. More durable organic materials having high C:N ratios (>80:1) and provide longer-lasting benefits compared with readily decomposable plant residues. Durability of organic mulches depends upon its composition. Factors influencing the rapidity of residue decomposition include the C:N ratio, cellulose and lignin contents, climatic conditions, and community composition of soil microorganisms. High lignin and cellulose containing cereal straws produce more weather-resistant mulches than legumes (Fig. 1).

The effectiveness of organic or vegetative mulches also depends upon their thickness. For adequate protection of the soil the mulch should not be too thick suppressing plant growth. Vegetative mulches, with their low thermal conductivities, decrease the conduction of heat into and out of the soil. Therefore, the mulched soil is cooler during the day and warmer at night than bare soil; it



Fig. 1 Straw mulching in a vegetable crop. Source: <https://horticulture.ucdavis.edu/>.

insulates the soil surface. Vegetative mulches also affect soil temperature by changing the radiant energy balance of the system. The radiation balance is influenced by the reflection of incoming radiation by surface residues, differential heating of soil and air, and reduction in evaporation of soil water from the surface. Soil temperature generally decreases as reflectance increases. Age of the residue, color, geometry (standing, flattened or matted), distribution and amount are the main characteristics that affect radiation reflection. Reflection is usually greatest from bright residues such as wheat straw and decreases with residue aging (discoloration) and decomposition.

Organic mulches have the advantage of being biodegradable and will eventually be broken down by microorganisms, increasing soil organic matter (SOM) content and improving soil structure or tilth. Organic mulches reduce the rate of evaporation from the soil surface, depending on their thickness, type, orientation and color. The material density of mulch affects the thickness and surface coverage obtained for a given weight, therefore, a low density material such as wheat straw reduces evaporation more effectively than higher density materials such as sorghum (*Sorghum bicolor* (L.) Moench) stubble or cotton (*Gossypium hirsutum* L.) stalks.

Crop residue mulch may be of two types, i.e., surface mulching with crop residues or rooted crop residues. One of the principles of conservation agriculture is that at least 30% of the area should be covered with crop residue mulch. After crop harvest with a combine harvester, the field is left with crop residues with their roots anchored to soil, which serve as mulch and the next crop is sown into this mulched condition using "Happy seeder or Turbo seeder" techniques.

Wastes from cattle sheds as mulch

Many indigenous technical knowledge (ITK) has been developed by the farmers in countries of the developing world to prepare nutrient rich mulch material, which can be followed under similar agroclimatic regions elsewhere. For example, in sub-mountainous hill regions of north-west India, farmers spread wastes from cattle-sheds in standing maize (*Zea mays* L.). Leafy materials like Basooti (*Adhatoda vasica*), Bhang (*Canavisa sativa*) and other locally available wild shrubs with prolific growth rates during the monsoon season and have no fodder value, therefore, are spread beneath animals to provide bedding. This material is collected each day together with dung and urine and spread in the standing maize crop. This makes a good mulching material to conserve and carry over the moisture from the time when the monsoon starts receding (Acharya et al., 2001). The material is left over after the maize harvest and is incorporated into the soil at the time of field preparation for sowing wheat. The practice not only conserves moisture but also enriches the soil to provide nutrients to the subsequent wheat crop because some of the materials are nutrient rich, for example, Basooti leave residues that contain on average 3.4% N, 0.4% P and 3.8% K.

Spreading of farm yard manure (FYM) as mulch

Farmyard manure has an important role worldwide in supplying humus forming materials which improves soil structure, water holding capacity, microbial populations and their activity, cation exchange capacity and resistance to soil erosion. For example, in the mid-hill region of north-west Indian Himalayan region, instead of incorporation FYM is spread over the fields after the sowing of wheat. This is done primarily to moderate the thermal regime as the minimum soil temperature during winter is very low. Depending upon the quantity applied, the coverage provided by FYM also helps in moisture conservation of seed-zone depth in addition to enriching soil fertility (Acharya et al., 2001). This technology can be adopted in similar agroclimatic regions for improving soil health and crop productivity in other regions of the world.

In the cold desert Himalayan regions the winter temperature goes as low as -30°C making life difficult for both humans and livestock. Here, cattle sheds are attached to the farmers' houses. In general, the ground floor in each house is used as a cowshed so that animals can be conveniently looked after during winter. The green leaves of pine trees (*Pinus sylvestris* L.), dry pine leaves, dry grasses or other bushes are put into the cattle shed as bedding. The dung and urine are collected from these beds every 3–4 days and stored in a heap within the cattle shed during winter months, so that decomposition is enhanced by the relatively high temperature conditions indoors. It is then placed in small heaps in the fields. These heaps are covered with a thin layer of soil to avoid dispersion of the manure by wind. This material is converted in due course to farmyard manure (FYM). This technology can be adopted in similar agroclimatic regions of the developing world countries.

As the winter season sets in, and before snowfall, the fields are prepared and crops such as coarse grains, wheat, barley (*Hordeum vulgare* L.) and some more remunerative vegetable crops like green pea (*Pisum arvense* L.) are sown. These fields are then covered with a layer of FYM as mulch. During summer, as temperature rises and the snow starts to melt, the FYM provides not only good nutrition but also a good cover to moderate the hydrologic and thermal regime for better seed germination. Beneath the mulch, chemical and biological activity are enhanced that encourage further decomposition and release of nutrients. The early crop of green pea with higher productivity fetches a good price in the market not only for its taste but also for its other quality attributes.

Cover crops

Live or cover crop mulch is also an important organic mulch; it is a covering of living plant material that is planted or allowed to remain after another crop has been planted. The live mulch system is based on the principles of mixed cropping. For example, a perennial legume might be established to smother weeds and then a seasonal grain crop is grown in strips through the cover crop. Cover crops succeed if they are low-growing and do not compete for light, moisture or nutrients with the main crop. The live mulch system is more successful in humid and sub-humid regions with little or no water deficit than in semiarid or arid regions. It has similar benefits to surface-applied organic mulches such as increased soil carbon content, increased rainwater infiltration and reduced soil erosion. Cover cropping is seen as green culture; given its role in SOM accumulation, soil structure improvement, soil

fertility increase, and weed and disease suppression with minimal or no chemical inputs. Thus, it should gain in popularity and political importance with sustainable agricultural practices now considered vital worldwide for soil and human health.

Types of cover crops

The three basic principles of conservation agriculture are (i) reduced tillage, (ii) residue retention (at least 30% of the soil surface should be covered) and (iii) crop diversification (FAO, 2019). The purpose of crop diversification is to select crops that will cover the soil year round; legumes are mostly preferred for this purpose. In general, the choice of cover crops depends on many factors viz., crop, soil, climate conditions, availability of resources and socio-economic factors. Within the limits of adaptability of any cover crop to prevailing climate and soil conditions, foremost among the desired characteristics is establishment vigor, persistence to the end of the required period, and ease of eradication to make way for the succeeding crop, usually the main crop (Edwards and Burney, 2005). The termination of cover crops may be achieved by their incorporation into the soil by tillage or by the application of broad spectrum herbicides, which is called brown manuring. The age-old practice of green manuring or the later developed practice of brown manuring are examples of cover crops.

Suitable cover crops include both monocotyledonous and dicotyledonous species. The former botanical group, which comprises mostly cereals and grasses, are doubly useful in soil erosion control because, in addition to the direct role of their vegetation shielding the soil surface against aggregate breakdown by raindrop impact, their fibrous root systems bind the soil, thus increasing its resistance to entrainment by overland flow. On the other hand, the soil-protection attributes of a dicotyledonous cover are due to their broad-leaf nature, specifically their characteristically overlapping system of leaves geared to intercepting rainfall and minimizing splash (Fig. 2).

Success in cover-crop establishment and function depends on species choice. But species suitability is hardly ever limiting considering the wide range of potential cover crops.

Cover crop mulches suppress weed growth through modification of the microclimate and release of allelochemicals. However, yield suppression of the food crops has been reported due to allelopathic effects of the cover crop on the main crop and competition for moisture during periods of drought stress. Cover crops also help to increase biodiversity in the system and help in disease and pest control. Some crops like the cereal rye (*Secale cereale* L.) and hairy vetch (*Vicia villosa* Roth) can be used safely as cover crop mulches. Cereal rye is one of the best winter-hardy cover crops grown in the mid-western U.S. Perennial ryegrass (*Lolium perenne* L.) is a potential cover crop that is commonly used as a living mulch for some vegetables such as cabbage (*Brassica oleracea* L.).

Living mulch, on the other hand, is a type of low-growing ground cover crop that blankets soil throughout the growing season of the main crop. It can be incorporated with the main crop by inter-seeding or over-seeding, co-seeding or otherwise companion planting. Some of the potential living mulches are: hairy or crown vetch, flatpea (*Lathyrus sylvestris* L.), clover (*Trifolium* L.), ryegrass etc. Living mulches grow for a long time with the main crop, whereas cover crops are incorporated into the soil or killed with herbicides.

Other organic mulches

Recycled paper is an important organic mulch, and can be considered as a biodegradable alternative to plastic mulch. Paper does not require disposal at the end of the season and its use can reduce solid waste and energy costs. Paper might decompose too rapidly, however, to provide ground coverage throughout a growing season. Treating the paper with waste cooking oil delays its breakdown and enhances soil warming by increasing the paper's translucency to sunlight.



Fig. 2 *Crotalaria* cover crop (<https://www.istockphoto.com/photo/crotalaria-cover-crop-keeps-soil-moisture-gm487086108-72723713>).

Inorganic mulches

Various inorganic materials such as plastic film, gravels, crushed rocks etc. are used as mulches. These do not add nutrients or humus to soil and do not decompose, except after long exposure to weathering.

Plastic mulch

Plastics are the most widely used inorganic mulch materials. Plastic film has played a major role in plant cultivation by creating some mechanical protection at the soil surface and a suitable microclimate in terms of temperature distribution, retention of humidity and the supply of carbon dioxide to the stomata of lower leaves of small plants. Plastic mulches have been used commercially for vegetable production since the early 1960s. It can be used very effectively with trickle or drip irrigation particularly for the cultivation of vines (*Vitis vinifera* L.) and strawberries (*Fragaria × ananassa* Duchesne) where it encourages lateral development of the root system. In many of the developed countries plastic film, such as low-density polyethylene (LDPE) film, is rapidly replacing conventional vegetative mulching because of its effectiveness and commercial viability, particularly for high value crops. Around 30–60% of the cultivated area is covered with a thick polythene film (30–150 µm). The plants can grow through holes in the LDPE film. It is often used under gravel or stone mulches for ornamental plants. The effectiveness of the plastic mulch depends upon its color, which determines the energy-radiating behavior and affects the temperature of the surface of the mulch and of the underlying soil. Recently, new plastic products for farmers such as photodegradable, co-extruded (black/white), and different colored mulches. They can maintain cool soil temperatures during summer in tropical countries, increase warming of soil in spring in temperate regions and repel aphids with the silver colored mulch. By raising the soil temperature, plant growth is accelerated producing earlier yields in temperate regions. Harvest can be 7–14 days earlier under black plastic while clear plastic can accelerate harvest by 21 days. For enhancing the earliness of vegetable crops in temperate regions, the mulch maximizes the soil-warming near infra-red radiation and minimizes the amount of visible light transmitted, which increases weed growth under the mulch.

Plastic mulches are mainly used for high value vegetable and cash crops because of the high cost and difficulty of keeping it in place (Fig. 3). Disposal of the plastic is an environmental concern (Steinmetz et al., 2016), and it is always completely removed from the field and can remain in soil for a long duration (Briassoulis et al., 2015). Liu et al. (2014) reported that long-term use of plastic film mulches in Chi/na has resulted in an estimated accumulation of 50–260 kg hm⁻² of residual plastics at 0–20 cm soil depth, which inhibited the growth of subsequent crops. Also other environmental problems are, for example, carbon footprint of plastic mulches (life cycle assessment), microplastics in soils being taken up by crops and animals, and ending up in the food chain, and potentially faster removal of pesticides from the plastic cover. There is also debate around potential degradation of soil quality under plastic mulches through shifting C/N ratio, reduction of carbon stocks, shifting microbial community composition to more fungi-dominated etc. Biodegradable plastic mulches have now been developed to resolve some of the environmental concerns (Bandopadhyay et al., 2018). Muskmelons (*Cucumis melo* L.), tomatoes (*Solanum lycopersicum* L.), pepper (*Piper nigrum* L.), cucumbers (*Cucumis sativus* L.), squash (*Cucurbita maxima* Duchesne), eggplants (*Solanum melongena* L.), watermelons (*Citrullus lanatus*) and okra (*Abelmoschus esculentus* (L.) Moench) are some of the vegetable crops that have shown a significant increase in early yield, and fruit quality when grown on plastic mulch. Some less valuable crops like sweet corn (*Zea mays* L.), snap beans (*Phaseolus vulgaris* L.), southern peas (*Vigna unguiculata* (L.) Walp.) and pumpkins (*Cucurbita pepo* L.) have also shown similar responses in temperate climates. Based on the ability to transmit, absorb or reflect specific wavelengths, plastic mulches are divided into the following categories:

- a. *Black plastic mulch*: Black is the most commonly used color of polyethylene mulch. This is chosen more often than clear polythene as a mulch because of its ability to control weeds by excluding photosynthetically active radiation. Black polythene placed in direct contact with the soil can warm the soil to a depth of about 8–10 cm. Under black plastic film the temperature of



Fig. 3 Plastic mulching in a tomato field. Source: <http://www.tomatonews.com/en/>.

- the soil during the day is not significantly higher and in certain cases it is even lower than of bare ground, but during the night the temperature is always 0.5–4 °C higher than that of uncovered ground. This is desirable for warm-season vegetables such as peppers, tomato, eggplants etc. It is also widely used for the cultivation of pineapples in Hawaii, USA for reaping higher yields.
- b. *Transparent or clear plastic mulch*: Clear plastics have the greatest warming potential. They are transparent to incoming radiation and trap the longer wavelengths re-radiating from the soil, thereby producing a greenhouse effect that warms the soil underneath. Under transparent plastic film during daytime soil temperature remains 2–10 °C higher than bare ground while at night differences of 2–4 °C are observed depending upon season, soil type and its moisture content. For warm season crops, such as vines, earlier and higher yields can be achieved with clear plastic mulches. However, as clear mulch is transparent to visible light, weed infestation under it may negate the benefits of soil warming.
 - c. *White plastic mulch*: Under white plastic mulch the temperature always remains lower than the uncovered ground due to its high reflectance and low transmittance of solar radiation. Since energy exchange is very poor, this film is used either in regions with a high level of sunshine where it is required to reduce the energy and lower the soil temperature or in regions of low luminosity where the amount of reflected light can be increased on the lower and middle leaves of crop plants.
 - d. *Colored plastic mulch*: These are opaque, like black plastics, and are made with many different pigments. They do not allow light to pass through, but rather, reflect specific wavelengths of light up into the crop canopy. These mulches may be useful for enhancing crop production. For example, red mulches are effective in promoting growth of tomatoes. The other response observed with different mulch colors is the increase or decrease in insect populations depending on reflected light and possibly temperature. Yellow appears to attract insects, especially cucumber beetle, and silver appears to repel aphids. In addition, pesticide applications could be reduced if instead of spraying it on the entire field, some rows with yellow mulch (e.g., one row of yellow for every five or six rows covered by differently colored mulches) were sprayed with insecticides to control high pest populations. Thus, yellow plastic mulches can be used to attract insects, much like a trap crop (Fig. 4).
 - e. *Infra-Red transmitting mulches (IRT)*: These have been developed recently and are wavelength selective mulches. The effect of IRT mulches on soil temperature is intermediate between clear and black polythene mulches. These mulches are pigmented to reduce the amount of visible light transmitted and, thereby, reduce weed growth underneath. Thus, these blend the soil warming characteristics of clear mulch with the weed control ability of black mulches. The IRTs contain very specific pigments that give them the unique ability to transmit a maximum of near infrared radiation and a minimum (14–16%) of visible light. Some visible light is required to be transmitted to maximize soil warming. The benefits of IRTs are greatest for early season plantings in temperate regions. Though these mulches are more expensive than black or clear plastics, some growers find them useful for high-value, heat-loving crops like melons.

Gravel mulch

Gravel mulching is an age-old practice and can be very effective for water conservation (both in enhancing infiltration and in suppressing evaporation even in layers as thin as 5–10 mm). The problem with this type of mulch is that once applied, gravels cannot be removed, do not decompose and might hinder future farm operations or land uses, and also the high cost of mining, transport and application in the field.



Fig. 4 Colored mulch film. Source: <http://agritech.tnau.ac.in/>.

Crushed rock

Crushed volcanic rock or stones are useful around plants subject to crown rot. Their color influences the temperature of the soil underneath. Black rock absorbs heat and keeps the soil temperature warmer than normal, while white rock reflects sunlight and maintains a cooler temperature below but can generate too much heat for the above ground portion of some plants to survive.

Other materials used as mulch

Sprayable preparations of latex, asphalt, oil, fatty alcohols etc. are also used as mulches for certain high-income crops such as strawberries. They inhibit evaporation by forming monomolecular films on the soil surface. Some water proofing chemicals like silicone application on clods on the soil surface forms a layer of water repellent aggregates. These aggregates then constitute a dry mulch that serves as a barrier to evaporation. This layer also enhances infiltration by allowing the treated surface zone to act as a retention reservoir to prevent runoff.

Methods of mulch application

The methods of mulch application vary with soil type, climate, crop and the type of mulch material used. Paper and plastic mulches are generally spread and fastened down either between the rows or over the rows of the crop. The plant seedlings in the latter case can grow through suitable openings in the material. Plastic mulches used in commercial cultivation are generally applied mechanically. A 15–23 cm border of plastic is usually buried leaving 76–91 cm exposed. Nowadays tractor operated mulch spreaders have been developed for laying plastic mulches in large fields.

Organic mulches are usually spread a few centimeters thick on the soil surface. Their depth depends upon the mulch material and types of crop on which it is applied. Thicker layers are normally placed around trees and shrubs while finer and smaller materials are generally used for small flowers and vegetables. The finer and smaller is the particle size, the thinner the layer needs to be. Thick layers of very fine material block air to the roots of plants. For trees, mulch cover should reach the tree's drip line or the outer perimeter of the branches. It should preferably be kept at least 15 cm from the tree trunk and two centimeters from plant stems.

Under conservation agriculture, sufficient residues remain on the surface after harvest of the previous crop. Sometimes, the left over residues in patches create problems for sowing of the next crop. Some engineering interventions can help in addressing this problem. Recently, a super straw management system (Super-SMS) has been developed by Punjab Agricultural University, Ludhiana (Punjab), India, which comes as an attachment to the Combine Harvester (NAAS, 2017). This straw management system can spread rice residues uniformly during the rice harvest. Then the next crop is sown in the standing crop residue using the Turbo Happy seeder. Harvesting of rice by super-SMS fitted combine harvesters allows concurrent sowing of the subsequent wheat crop, which saves time, energy and requires only one irrigation by utilizing the residual moisture of rice fields. This practice also reduces traffic and compaction. Further, this technology can replace rice residue burning, a problem for air quality.

Mulches are normally applied and spread uniformly on the soil surface except in the case of vertical mulching. Here crop residues, stubble from the previous crop or other agricultural waste materials are placed in vertical trenches at intervals of every few meters to recharge the profile with water and conserve moisture in low rainfall areas of the semi-arid tropics. The few centimeters of stubble above the trenches can also reduce runoff and sediment loss from the field. In vertical mulching, holes are dug around the base of a tree or shrub that is stressed. The holes are filled with a mixture of organic material. The back fill and aeration can dramatically improve root growth and reduce or eliminate stress. The other advantages of vertical mulching include: reduced compaction, increased drainage, oxygen to the root system, soil organic matter, soil organisms, and water retention. The aeration and soil treatment create an environment that mimics nature and allows for greater nutrient uptake, increased root growth and the potential regeneration of a damaged root system.

Limitations of mulching and cover crops

Despite the positive aspects of mulching, it has some limitations, which must be taken into account while following the practice of mulching.

- (i) If mulch is applied before sowing it hinders sowing and many post sowing operations like fertilizer applications, irrigation etc. Mulches pose particular problems for the arable farmer because tillage tools become clogged with the residue, thus planting under the residue is not always successful. Some specific equipment is required for post sowing operations in mulched plots.
- (ii) In shallow potato (*Solanum tuberosum* L.) planting if mulch is managed improperly it may lead to a greater percentage of stolon becoming aerial shoots and more tubers becoming sunburned. The mulch layer in potato fields should be removed once a full canopy is achieved to avoid disease and pest infestation.
- (iii) In areas of high rainfall, thick mulching may lead to waterlogging conditions and provide an environment for disease infestation. If the soil is too wet under mulch, denitrifying bacteria will increase and may cause a loss of N.

- (iv) The effectiveness of mulch depends upon the depth and extent of its coverage. However, if the mulch is applied too thickly it may hinder seedling emergence through the deep impenetrable barrier. Therefore, it is suggested that a thin layer of mulch is used at planting to avoid impeding seedling emergence, then followed with a thicker application once the seedlings are established.
- (v) Plastic mulches usually require pick up and disposal at the end of the season, and their manufacture and disposal entail significant environmental costs because these are not biodegradable, and microplastics especially pose a potential environmental threat. Plastic mulch can interfere with recharge of the soil profile by rain or overhead irrigation, which can lead to water deficits in vegetables if drip irrigation is not used.
- (vi) Numerous phytotoxic substances associated with various crop residues have been shown to limit growth of several crop species reportedly through exudation, leaching or microbial production of allelopathic chemicals.
- (vii) Organic mulches such as grass and straw often contain seeds that allow weeds to grow and hamper crop growth.
- (viii) There are competing demands for crop residues for other uses in the semi-arid tropics of the developing countries. Most of the plant residues are used as cattle fodder, therefore, farmers might not adopt this practice where fodder is scarce. In the tropics, crop residues are also used for fencing, roofing and as a source of household fuel. Mulching is thus practiced under conditions where the marginal return from this practice is higher than its other competing demands.
- (ix) Crop residue mulches reduce the temperature of the surface soil layers. In cool climates, this reduction in soil temperature shortens the growing season of arable crops while in wet areas higher soil moisture may induce gleying and anaerobic conditions. The limiting problem of using residue mulch on the soils in temperate regions is its inability to remove excess water to allow the soil to warm up in time for planting the next crop.
- (x) The mulch competes with the main crop for nitrogen during decomposition if the C:N ratio of the material is high due to immobilization of the soil nitrogen by microorganisms.
- (xi) The surface residue mulch may generate a more favorable habitat for soil and surface-dwelling insects and pathogens. The residue left at the surface reduces water loss, moderates temperature extremes and provides food resources for pests and increases their infestation. Furthermore, for many plant pathogens, residues provide a source of food, a place to live and a place to reproduce. A significant proportion of root infecting organisms, for example, depend upon crop residues for their survival in soil. Root diseases caused by *Pythium* species are favored by wet soils and cool temperature, especially in maize production.
- (xii) In areas of water scarcity, cover crops compete with the main crops for water and nutrients. If the cover crops do not yield any economic output then farmers are reluctant to grow them even though they are beneficial for soil health. Rather than green manure crops, growing short duration legumes may be promoted in between cereal crops, which can serve the dual purpose as cover crop and also yield some economic benefit to the farmers.
- (xiii) Some of the limitations of living mulch include: (a) It competes for nutrients and water with the main crop which can reduce the yield; (b) It slows down the maturity of the main crop; (c) Some mulch produces allelopathic effects on the crop; (d) Specialized management practices are required; (e) The temperature below the mulch is lowered.

Conclusions

Mulching can be considered as a win-win technology because it can solve the problem of straw burning, improve the environmental quality and crop productivity. Mulches can be of organic and inorganic nature. Site-specific mulching practices should be developed. Because mulching improves the soil hydrological regime, this practice can mitigate the adverse effect of climate change and climate variability. Therefore, mulching can be considered as a potential climate resilient technology. However, this technology has some limitations such as the disposal of non-biodegradable plastic mulch, competing demand of crop residues by other enterprises, lack of machinery and knowhow among the farmers for field operations in the presence of mulches, particularly in the case of developing countries. Appropriate policy decisions may be taken to address these limitations. Together with mulching, short duration pulses may be promoted as cover crops in cereal based rotations to improve soil health, reduce soil erosion, and improve crop productivity and farmers' income. Our motto should be "Grains for human consumption and residues for improving soil health."

References

- Acharya CL and Abrol IP (1978) Exchangeable sodium and soil water behaviour under field conditions. *Soil Science* 125: 310–319.
- Acharya CL and Prihar SS (1969) Vapour losses through soil mulch at different wind velocities. *Agronomy Journal* 61: 666–668.
- Acharya CL, Sandhu SS, and Abrol IP (1979) Effect of exchangeable sodium on the field drying pattern of soil at two evaporative demands. *Soil Science* 127: 56–62.
- Acharya CL, Sood MC, and Sharma RC (2001) In: Acharya CL, Ghosh PK, and Subba Rao A (eds.) *Indigenous Nutrient Management Practices—Wisdom Alive in India*, p. 395. Bhopal: Indian Institute of Soil Science.
- Acharya CL, Hati KM, and Bandyopadhyay KK (2005) Mulches. In: Hillel D, Rosenzweig C, Powlson DS, Scow KM, Singer MJ, Sparks DL, and Hatfield J (eds.) *Encyclopedia of Soils in the Environment*, pp. 521–532. Elsevier Publication.
- Acharya CL, Bandyopadhyay KK, and Hati KM (2018) Mulches: Role in climate resilient agriculture. In: *Earth Systems and Environmental Sciences*, <https://doi.org/10.1016/B978-0-12-409548-9.11654-9>.
- Bandopadhyay S, Martin-Closas L, Pelacho AM, and DeBruyn JM (2018) Biodegradable plastic mulch films: Impacts on soil microbial communities and ecosystem functions. *Frontiers in Microbiology* 9: 819.

- Briassoulis D, Babou E, Hiskakis M, and Kyrikou I (2015) Analysis of longterm degradation behaviour of polyethylene mulching films with pro-oxidants under real cultivation and soil burial conditions. *Environmental Science and Pollution Research* 22: 2584–2598.
- Edwards L and Burney J (2005) Cover crops. In: *Earth Systems and Environmental Sciences*, 311–318. <https://doi.org/10.1016/B978-0-12-409548-9.11654-9>.
- FAO (2019) Conservation Agriculture. Available online at: <http://www.fao.org/conservation-agriculture/overview/what-is-conservation-agriculture/en/>. (accessed August 2019).
- Liu EK, He WQ, and Yan CR (2014) 'White revolution' to 'white pollution'- agricultural plastic film mulch in China. *Environmental Research Letters* 9: 3.
- MNRE (Ministry of New and Renewable Energy Resources) (2009) Govt. of India, New Delhi. www.mnre.gov.in/biomassresources.
- NAAS (2017) *Innovative Viable Solution to Rice Residue Burning in Rice-Wheat Cropping System Through Concurrent Use of Super Straw Management System Fitted Combines and Turbo Happy Seeder*. Policy Brief No. 2 New Delhi: National Academy of Agricultural Sciences. 16p.
- Prihar SS, Singh B, and Sandhu BS (1968) Influence of soil and climatic environment on evaporation losses from mulched and unmulched pots. *Journal of Research* 5: 320–328.
- Steinmetz Z, Wollmann C, Schaefer M, Buchmann C, David J, Troger J, et al. (2016) Plastic mulching in agriculture. Trading short-term agronomic benefits for long-term soil degradation? *Science of the Total Environment* 550: 690–705.

Mulches and cover crops part II: Role in soil health and climate resilient agriculture

KK Bandyopadhyay^a, CL Acharya^b, and KM Hati^c, ^aDivision of Agricultural Physics, ICAR-Indian Agricultural Research Institute, New Delhi, India; ^bFormer Director, ICAR-Indian Institute of Soil Science, Bhopal, India; ^cDivision of Soil Physics, ICAR-Indian Institute of Soil Science, Bhopal, India

© 2023 Elsevier Ltd. All rights reserved.

Introduction	401
Effect of mulching on soil health	402
Soil moisture regime	402
Controlling evaporation from the soil surface	402
Improving infiltration	402
Soil moisture retention	403
Water condensation at night	404
Controlling runoff and soil erosion	404
Effect of mulching on the environment	407
Effects on crop growth and input-use efficiency	408
Modification of crop growth	408
Weed competition	408
Seedling emergence	408
Root growth	408
Improvement of input-use efficiencies	409
Role of cover crops on soil quality improvement	411
Role of cover crops on runoff and soil erosion control	411
Recommendations and policy decisions on mulching and cover crops	411
Conclusions	412
References	412

Abstract

Mulches may improve soil health, moderate the soil moisture regime, buffer temperature fluctuations, facilitate aeration, and crop growth, seedling emergence, root growth, and foster efficient water and nutrient use. Mulching also improves the environment by facilitating carbon sequestration and reducing greenhouse gas emissions from the soil. Similar to mulches, growing cover crops protects the soil, improves biodiversity and is one agronomic approach to erosion control and enhancing soil fertility. Policy decisions to overcome limitations in adopting conservation agriculture, mulching and cover cropping are required to provide technical and financial support to farmers.

Introduction

Mulching has a great potential for improving soil health by favorable modification of the soil hydrological regime, carbon sequestration, minimizing soil erosion, improving agricultural productivity and mitigation of climate change. Developing adaptation and mitigation strategies for the adverse impacts of climate change are amongst the most serious challenges of the present century. With increasing awareness of the effects of climate change by both national and international bodies, ensuring climate resilience has become a major goal. The aim of such efforts is to address the vulnerability that communities, states, and countries have to the environmental consequences of climate change. Currently, efforts for climate resilience encompass social, economic, technological, and political strategies that are being implemented at all scales of society. Various climate resilient technologies have been developed in agriculture to mitigate the impact of climate change. For example, the National Initiative for Climate Resilient Agriculture by the Government of India has outlined some technologies in this context: resource conservation practices and technologies for natural resource management, and efficient use of resources and inputs for improved crop, livestock and fisheries production etc. In situ retention of biomass and crop residues as mulches has been identified as a climate resilient technology for improving soil health, soil moisture conservation and enhancing water-use efficiency. Mulch has a buffering effect and it dampens the influence of environmental factors on soil. The magnitude of this effect depends upon the quality, quantity and durability of the mulch material, soil type and climatic conditions. Mulching influences the soil hydrological and thermal regime through the radiation balance, rate of heat and water vapor transfer, and heat capacity of soil. Thus it can protect the crop from the adverse impact of climate change and climate variability. Organic mulches add nutrients to soil through microbial decomposition and contribute to carbon sequestration, which can be used as a mitigation strategy for global warming.

Mulching also improves the physical soil conditions by enhancing soil aggregation, which helps in carbon sequestration and conservation of water in soil by reducing evaporation, increasing infiltration and retarding runoff and erosion, favorably modifies

the soil's thermal regime, improves the soil chemical environment and biological activity in soil. Modification of the soil microclimate by mulching favors seedling emergence, root proliferation, and suppresses weeds. Thus the soil edaphic environment under mulches improves crop productivity, enhances input-use efficiency and reduces environmental pollution. It is useful as an alternative to cover crops in dry areas where insufficient rain prevents their establishment before the onset of heavy rain or strong winds, or where a cover crop competes for moisture with the main crop. However, in areas with assured water supply cover crops serve as a useful technology to improve soil health and carbon sequestration. Improper application of mulches, i.e., too thick a layer of mulch can lead to an anaerobic environment under heavy rainfall leading to loss of nitrogen through denitrification, disease and pest infestation, some allelopathic materials produced in some crop residues retard crop growth, and mulches with high C:N ratios can lead to immobilization of N. A proper site-specific methodology needs to be followed for mulching and cover cropping under different soil, crop and climatic conditions to achieve improved soil health, sustainable crop production and mitigate the adverse impact of climate change.

This chapter examines the role of mulches and cover crops on soil health, environmental protection, crop growth and input-use efficiency. Part I of this topic deals with the types of mulches and cover crops and methods of their application.

Effect of mulching on soil health

Soil health can be defined as "the continued capacity of soil to function as a vital living system, within ecosystem and land use boundaries, to sustain biological productivity, maintain the quality of air and water environments, and promote plant, animal and human health" (Doran et al., 1996). Soil health or quality includes three groups of mutually interactive attributes, i.e., soil physical, chemical and biological quality, which must be restored to its optimum to sustain or even improve productivity in the long run.

Soil moisture regime

Mulching favorably influences the soil moisture regime by controlling evaporation from soil surface, improving infiltration and soil moisture retention and facilitating condensation of water at night due to temperature reversals.

Controlling evaporation from the soil surface

Mulching reduces evaporation from the soil surface by reducing the intensity of radiation and wind velocity. It retards evaporation mainly during the initial energy-limited stage of drying, but is less effective in controlling evaporation during the falling rate or supply-controlled stage. During this stage, the soil surface dries under the mulch and the vapor transfer of water to the atmosphere is inhibited more by the dry soil than the overlying vegetative mulch. Furthermore, the effectiveness of porous mulches, including surface cultivated soil mulch, decreases with time. Over long periods the cumulative evaporation losses from cultivated and no till soils may be similar even on soils that responds to the mulch application. When the loss of evaporation is decreased in the soil during the terminal phase of evaporation and water loss occurs only through movement in soil as vapor, the rates of evaporation are governed by the porosity and thickness of the dry layer (Acharya and Prihar, 1969). In addition to porosity and thickness of the dry layer, climatic conditions, especially wind velocity, also govern evaporation loss at this stage (Prihar et al., 1968; Acharya and Prihar, 1969). A more porous soil and lower wind velocity might lose less moisture through vapor transfer than a less porous soil and high wind velocity.

Higher rates of mulching retard the energy reaching the soil surface and hence limit evaporation losses at the constant rate stage (Bond and Willis, 1969). The retardation of initial evaporation can also enhance the process of internal drainage and thus can allow more water to migrate into the deeper parts of the profile, where it is less likely to be lost by evaporation and can be stored.

Mulching also affects the pattern of evaporation by offering resistance to water vapor flow from the soil surface to the atmosphere and by increasing the thickness of the relatively non-turbulent air zone above the soil, thus changing the boundary layer conditions of the soil-air interface. Relative magnitude of the reduction in evaporation for different mulch amounts is much greater under high evaporative demand. The cumulative reduction in evaporation due to mulching follows a quadratic relationship with time. The magnitude and time required to reach maximum reduction in evaporation depend on the method and rates of mulch application, soil type and magnitude of evaporative demand (Fig. 1) (Prihar et al., 1996). Orientation of the residue, i.e., flat or matted vs. standing affects the porosity and thickness of the layer and thus the rate of evaporation from the soil surface. With increasing amounts of standing residue higher wind speeds are needed to induce evaporation, and rates of water loss are reduced. Residue position also affects soil temperature (Table 1), which in turn affects the rate of evaporation through its influence on the vapor pressure of the soil water (Unger, 1990).

Water saving under mulching is prominent if rains are frequent. However, under extended dry periods mulches may keep the soil surface moist for longer and hence the first stage of evaporation is extended, but the net saving of water is negligible.

Improving infiltration

Mulching with organic materials serves as a barrier for runoff, which enhances water infiltration. Second, mulch intercepts rainwater and protects the soil. Furthermore, decomposition of organic mulches adds organic matter, improves the macro porosity and stability of soil and thereby improves water transmission properties (Lal, 1987), which enhance infiltration and recharge of the soil profile. Straw mulch application increases soil water storage and storage efficiency. The amount of water storage in the soil profile,

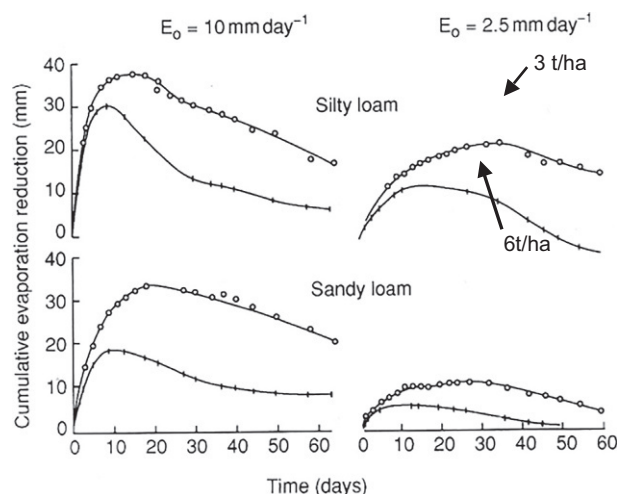


Fig. 1 Rate of cumulative evaporation reduction (CER) in two different soils under two wheat (*Triticum aestivum* L.) straw mulch application rates (3 t ha⁻¹ and 6 t ha⁻¹) at two evaporative demands (E₀). Reproduced from Prihar, SS, Jalota, SK and Steiner, JL (1996) *Soil Use and Management* 12: 152.

Table 1 Average daily soil surface temperature with wheat (*Triticum aestivum* L.) straw in different positions during August to September, Colorado, America.

Straw position	Soil surface temperature (°C) ^a
Bare soil	47.8 c
Flat straw	41.7 b
¾ flat and ¼ standing	39.6 b
½ flat and ½ standing	32.2 a

^aColumn values accompanied by different letters are significantly different at $P = 0.01$ (Duncan's multiple range test).
Reproduced from Unger PW (1990) *Advances in Soil Science*, 13, 27–68.

Table 2 Effect of wheat (*Triticum aestivum* L.) straw mulch rate on soil water storage during fallow,^a water storage efficiency and dryland grain sorghum (*sorghum bicolor* L.) yield in Bushland, Texas.

Mulch rate (Mg ha ⁻¹)	Water storage (mm)	Storage efficiency (%) ^b	Grain yield (Mg ha ⁻¹)	Total water use (mm)	Water use efficiency (kg m ⁻³)
0	72 c	22.6 c	1.78 c	320	0.56
1	99 b	31.1 b	2.41 b	330	0.73
2	100 b	31.4 b	2.60 b	353	0.74
4	116 b	36.5 b	2.98 b	357	0.84
8	139 a	43.7 a	3.68 a	365	1.01
12	147 a	46.2 a	3.99 a	347	1.15

Column values followed by the same letter are not significantly different at 5% level (Duncan's multiple range test).

^aFallow duration 10–11 months.

^bPrecipitation averaged 318 mm.

Reproduced from Unger PW (1990) *Advances in Soil Science*, 13, 27–68.

storage efficiency, total water use and water-use efficiencies of dryland crops increase with increased mulch rates (Table 2), and may outweigh the benefits of reduced evaporation in terms of water conservation (Unger, 1990).

Soil moisture retention

Mulching improves moisture retention properties of soil through its effect on pore-size distribution and soil structure. Higher rates of mulch application increase soil-water retention more at lower suctions (Lal, 1987) with the increase in macropores and inter-aggregate pores resulting from enhanced soil organic matter content and greater activity of soil fauna, e.g., earthworms and termites in mulched plots.

Water condensation at night

Stone and gravel mulches induce lateral movement of heat and vapor, which could in turn collect water under the stones due to condensation at night in amounts sufficient to serve as the sole water source for some species of desert plants.

Controlling runoff and soil erosion

Mulching invariably decreases soil erosion and often reduces runoff rates and volumes. Mulch cover protects the soil from raindrop impact and surface sealing, increases the infiltration of water and decreases run-off velocity through physical resistance to water flow. In general, water loss through runoff decreases exponentially with increasing mulch amounts (Table 3) (Erenstein, 2002). Runoff water is filtered through the mulch, which reduces sediment loads and erosion. Mulching also decreases sediment concentration in runoff water through the protective effect of crop residues against raindrops.

Ranaivoson et al. (2017) reported that soil erosion decreased markedly with increasing amounts of surface crop residues (Fig. 2). Residue amounts of 2–4 t dry matter ha⁻¹ reduced soil erosion by about 80% compared to bare soil. With 8 t dry matter ha⁻¹ of residues and above, there was virtually no more soil loss. It was observed that when around 60% of the soil was covered soil loss was reduced by an average of 80%. The boundary line showed that the maximum effects of surface residue amounts on soil erosion were reached with 2 t ha⁻¹ of residues (Fig. 2a). This corresponds to 80% soil cover or more (Fig. 2b). Residue cover dissipates raindrop energy and reduces the velocity of runoff flows by the interception of precipitation, thereby decreasing soil detachment (Gilley et al., 1986). It protects the physical structure of topsoil, improves aggregate stability and minimizes surface crusting (Jordán et al., 2010), which facilitates infiltration and hence reduces runoff and soil erosion.

Acharya et al. (1993), in a 5-year study with maize (*Zea mays* L.)–wheat cropping sequence on a silty clay loam under rainfed conditions, showed that under conservation tillage, involving the concept of agro-forestry, the minimum tillage with a mulch of robinia (*Robinia pseudocasia* L.) planted on either side of the terraces to generate mulching material and crop residues (1:1), the runoff losses were 7–19% of the total rainfall compared to 19–34% under conventional tillage. Therefore, more rainwater entered the soil profile under the former than the later treatment.

Table 3 Effects of rice (*Oryza sativa* L.) straw mulch rates on runoff and soil loss in Nigeria.

Mulch rate (Mg ha ⁻¹)	Run off (%)	Soil loss (Mg ha ⁻¹)	Marginal conservation effect (Mg soil) (Mg mulch) ⁻¹
0	75.4	9.6	–
2	43.4	2.3	3.7
3	15.2	0.5	1.8
6	5.4	0.1	0.13
12	0	0	0.017

Reproduced from Erenstein O (2002) *Soil and Tillage Research*, 67, 115–133.

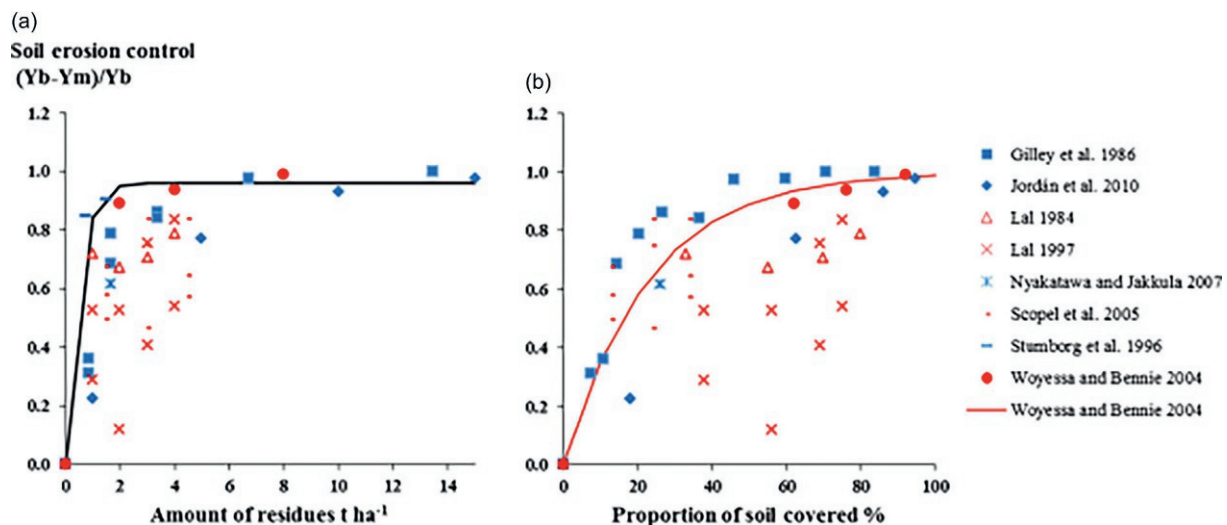


Fig. 2 Effect of crop residue mulching on soil erosion control; Y_m = soil erosion under mulching and Y_b = soil erosion under bare soil. Source: Ranaivoson et al. (2017) *Agronomy for Sustainable Development* 37: 26.

The rate of soil loss decreases exponentially with the increase in percentage area covered by mulch. The mulch factor (MF) defined as the ratio of soil loss with mulch to that without mulch, is related to the percentage residue cover (RC) or mulch cover by the following expression (Morgan, 2005):

$$MF = e^{-a RC}, \quad (1)$$

where, a ranges between 0.03 and 0.07; generally a is set to 0.05. Eq. (1) is valid for rill erosion, whereas inter-rill erosion decreases linearly with mulch cover. For most soils, percentage mulch residue cover and slopes longer than 25 m mean the contribution of inter-rill erosion to total soil loss is quite small. Thus, Eq. (1) is generally used to compute MF. This mulch factor is multiplied by RC values to account for the effect of mulching in the universal soil loss equation (USLE). A minimum 50% soil cover is required to achieve soil loss reductions. Ideally, a mulch should cover 70–75% of the soil surface. The optimum mulch rate can be computed for a pre-selected set of conditions using the Manning equation for flow velocity and the relationship between Manning's n and the mulch rate is given below:

$$n = (r^{0.67} S^{0.5})/v, \quad (2)$$

where, r is the hydraulic radius, (m) S is the slope and v is the desired maximum flow velocity (m s^{-1}).

According to Foster et al. (1982), n_m , Manning's n as affected by a mulch cover, can be calculated as.

$$\left(n^{3/2} - n_s^{3/2} \right)^{2/3}, \quad (3)$$

where, n_s is Manning's n for a particular soil, s.

The mulch rate (M , kg m^{-2}) is related to the Manning's n for mulch (n_m) by the following relationship.

$$n_m = 0.071 M^{1.12} \quad \text{for inter-rill erosion} \quad (4)$$

$$n_m = 0.105 M^{0.84} \quad \text{for rill erosion} \quad (5)$$

Experimental evidence shows that an adequate quantity of crop residue mulch can mitigate the effect of the degree of slope in reducing soil loss.

Mulching is also effective in reducing wind erosion by reducing its velocity on the soil surface, provided that the mulch is not blown up by wind. Standing crop residues are more effective than flattened surface mulches in reducing wind erosion. The effectiveness of standing crop residues depends upon size and number of stalks per unit area.

Modifying the soil thermal regime

Mulches have a moderating influence on the soil thermal regime and the effect varies among soils, climates, mulch materials and their application rates. They increase soil temperature during cooler weather and decrease it during hot spells. In general, mulches have a damping effect on the amplitude of the diurnal fluctuations in soil temperature. Organic mulching enhances the soil temperature at night and in the early morning, but it decreases the daytime temperature compared to unmulched plots (Bhagat and Acharya, 1987).

Mulching is commonly practiced for some fruit crops like strawberries to keep the fruit clean by avoiding direct contact with soil and moderating the soil hydrological and thermal regimes for better productivity and quality of the fruit. Transparent polythene mulch produces the maximum soil temperature, whereas organic mulches like pine needles (*Pinus*) and grass (*Poaceae*) mulches lower it. Black polythene mulch, however, does not alter the maximum soil temperature appreciably (Fig. 3). The application of straw mulch can lower the maximum soil temperature due to interception of the incoming solar radiation, high reflectivity and low heat conductivity, the magnitude of which depends upon soil wetness, incidental radiation, rate of mulch application, and time of the year (Prihar, 1986). The decrease in maximum soil temperature by straw mulching during the early stage of growth of sugarcane (*Saccharum officinarum* L.) can significantly improve the yield.

Blanco-Canqui et al. (2006) observed that soil temperature decreased with the increase in rates of maize stover cover, except during winter (Fig. 4). The bare soil without stover consistently had the highest soil temperature for all stover rate treatments. The temperature from June to October decreased with the decrease in rate of stover removal (Blanco-Canqui and Lal, 2009). Stover removal at rates >25% strongly altered the soil temperature regime. On average, there was a drop of 5 °C when stover removal rate increased from 0 to 75% regardless of soil type. No significant differences in soil temperature were observed for stover removal rates of 75% and 100%. Lower soil temperatures in soils with residue retention are attributed to the fact that mulch cover significantly alters the radiation balance. Light colored stover cover has a high albedo and reflects incoming solar radiation at the soil surface (Chaudhary and Acharya, 1993; Horton et al., 1994).

Changing soil aeration

Crop residue mulches improve soil aeration by promoting free exchange of gases between soil and atmosphere. This is facilitated by increased soil structural stability, total porosity and macroporosity, decreased surface crusting, which all contribute to improved overall soil drainage. The rate of oxygen diffusion is higher under mulches. The gaseous composition of soil air under a mulch depends on the nature of its material (C:N ratio), its rate of decomposition, the soil moisture regime and climatic conditions. Plastic

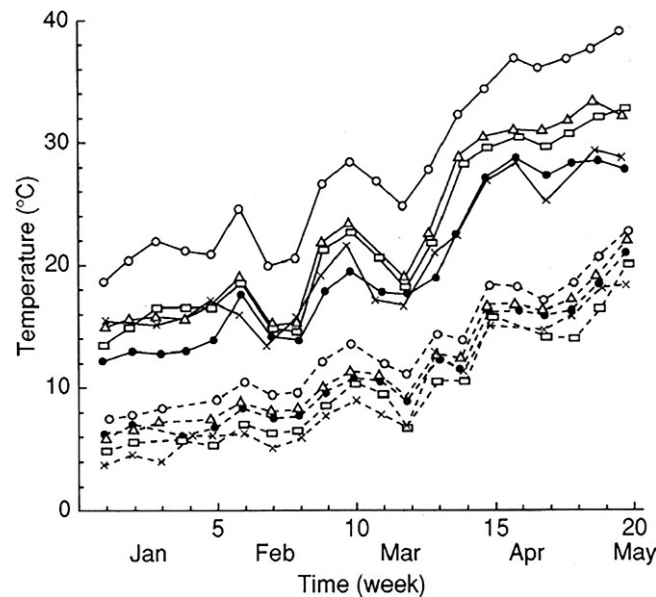


Fig. 3 Mean weekly soil temperature measured daily at —14.30 h and—06.30 h at 5-cm depth under different mulch treatments and unmulched control in strawberry (*Fragaria vesca*) on a Typic Hapludalf. ○ transparent polythene; ● black polythene; △ unmulched control; □ pine needle mulch. Reproduced from Gupta R and Acharya CL (1993) *Journal of the Indian Society of Soil Science* 41: 17–25.

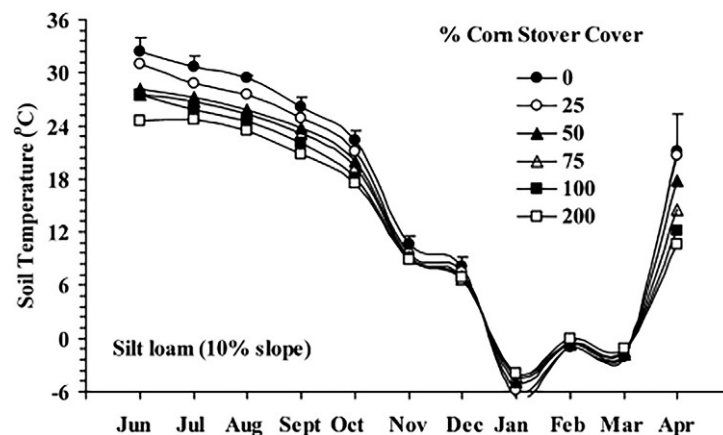


Fig. 4 Response of soil temperature to varying rates of stover cover in a no-till corn in Ohio. The error bars represent the LSD values of the mean. Source: Blanco-Canqui and Lal (2009). *Critical Reviews in Plant Science*, 28:139–163, 2009.

mulch is practically impervious to carbon dioxide (CO_2), a gas that is of prime importance for photosynthesis. The CO_2 that accumulates under plastic escapes through the planting holes causing a 'chimney effect', which results in localized large CO_2 concentrations that accelerates crop growth.

Improving soil structure

Mulching improves soil structural properties directly by preventing raindrop impact and indirectly by promoting biological activity. Organic mulching improves the mean weight diameter of water stable aggregates, because of the addition of organic matter with decomposition of the mulch by soil microorganisms. The mean weight diameter of water stable aggregates increases with increased mulching rates (Table 4) (Lal, 1987).

In general, under mulched conditions, the bulk density of the soil is lower than without mulching. Bulk density decreases with increased mulch residue. This can be attributed to higher organic matter contents (Hati et al., 2006), improved aggregation and consequently increased total porosity, decreased compaction (Leroy et al., 2008) and increased root growth, earthworm activity and biopore formation (Bandyopadhyay et al., 2010). Chatterjee et al. (2016) reported a significant increase in the mean weight diameter, percentage of water stable aggregates, soil physical health indices like the 'S' index and least limiting water range (LLWR) from the application of wheat residue mulch in maize under a sandy loam soil.

Table 4 Effect of maize (*Zea mays* L.) straw mulch rate on pore size distribution of 0–5 cm layer of an Alfisol after 18 months of initiation of the experiment.

Mulch rate (Mg ha^{-1})	Porosity (%)			Mean weight diameter (mm)
	Macro ($>75 \mu\text{m}$)	Meso ($30\text{--}75 \mu\text{m}$)	Micro ($5\text{--}30 \mu\text{m}$)	
0	17.9	7.1	23.1	1.2
2	20.6	7.2	22.4	1.4
4	27.7	8.4	18.7	2.1
6	28.9	8.3	17.9	2.0
12	37.0	8.2	13.2	2.4
LSD (0.05)	5.3	NS	4.2	–

Reproduced from Lal R (1987) *Tropical Ecology and Physical Edaphology*, pp. 618–685. Wiley-Interscience Publication.

Improving the soil chemical environment

Organic mulches add organic matter and plant nutrients to soil with decomposition which increases carbon sequestration. The volatilization and leaching losses of nitrogen are reduced under mulched conditions. During mulch decomposition soil mineral nitrogen is immobilized by microbes. Cation exchange capacity and thus soil fertility are substantially influenced by increased organic matter contents in soils containing predominantly low activity clays. Furthermore, decomposition of organic mulches adds organic acids to the soil lowering the soil pH, which influences the bioavailability of many plant nutrients such iron, zinc and copper. Mulches also influence the availability of nutrients through their influence on physical and biological soil conditions described above.

Changing soil biological activity

Soil biological activity is either directly influenced by the supply of food substrates by organic mulches or indirectly influenced by both organic and inorganic mulches through altering soil hydrological and thermal regimes. The activity and diversity of soil organisms are influenced by mulch quality and application rates. Soil microorganisms (bacteria, fungi and actinomycetes) and fauna (earthworm, millipede, centipede etc.) help in decomposing soil organic matter and hence enhance nutrient availability. However, mulching with materials of high C:N ratio results in nitrogen immobilization. Changes in soil physicochemical properties during mulching can promote changes in microbial community structure and its metabolic function. Zhang et al. (2020) reported that bacterial communities were significantly higher (3.2%) in peanut (*Arachis hypogaea* L.) mulched soil than that under polyethylene film (1.2%) because of the relative abundance of *Nitrospirae* in the peanut mulched soil. Similarly, the fungal communities were significantly higher (33.7%, 21.9%) in peanut hull mulched soil than that in polyethylene film mulched soil (14.9%, 6.5%), due to the relative abundance of *Mortierellomycota* and *Basidiomycota*. Peanut hull mulching increased the diversity of fungal communities at 0–20-cm soil depth and the diversity of bacterial communities at 20–40-cm soil depth. At the microbial functional level, there was an enrichment of bacterial functional features, including amino acid transport, and metabolism and energy production and conversion, and there was an enrichment of fungal functional features, including undefined saprotrophs, plant pathogens and soils apotrophs. They concluded that organic mulching had a positive regulatory effect on the soil bacterial and fungal communities and ecosystem functions, and so, is more suitable for tea (*Camellia sinensis* (L.) Kuntze) plantations. Rabary et al. (2008) reported that soil β -glucosidase and acid phosphatase activities were significantly higher for the direct seeding mulch-based systems using crop residues than for those using living mulches in Madagascar. Application of *Trichoderma viride* together with waste mulch in ratoon sugarcane increased contents of organic C, available P and microbial biomass of soil. The residue decomposition and growth promotion abilities of *Trichoderma* produced a synergistic effect with mulch from waste compared with its burning and removal. With degradation, crop residue mulch served as a source of energy for enhanced multiplication of soil bacteria and fungi and provided suitable niches for plant–microbe interaction. Trash mulching together with microbial consortia containing *Gluconacetobacter diazotrophicus*, *Trichoderma viride* and *Pseudomonas fluorescens* effectively improved soil quality properties, and sugarcane and sugar yields in a multi ratooning system. (Yadav et al., 2009). Mulching with organic residues of weeds like wild sage (*Lantana camara* L.) or eupatorium (*Eupatorium adenophorum* Spreng.) under conservation tillage practices enhances the earthworm population with increased food substrate and better hydrological and thermal regimes under the mulch (Acharya et al., 1998).

Effect of mulching on the environment

Intensive plough tillage (PT), use of agricultural chemicals and the burning of crop residues are major farm activities that emitting greenhouse gases (GHGs). On the other hand, no till (NT) and crop residue retention improve soil structure and enhance C sequestration and reduce GHG emissions. With an estimated annual global production of crop residues at 3.76 Pg (Lal, 2005),

conversion of 15% of C in the residues into stable SOC fractions implies C sequestration of 0.15–0.175 Pg year⁻¹. Chatterjee et al. (2018) reported that application of wheat residue mulch in a maize crop significantly increased the total organic carbon (TOC) concentration by 14.9% at 0–5-cm soil depth compared to the no mulch treatment in a sandy loam soil. Water stable aggregate-associated carbon concentration in large macro- and micro-aggregates increased significantly by 16.7% and 11.8%, respectively with crop residue mulching. Application of crop residue mulch also resulted in a significant increase in labile and non-labile pools of carbon at 0–5-cm soil depth compared to the no mulch treatment. This increase in TOC and the non-labile pool of OC will facilitate sequestration of SOC and will contribute to reducing GHG emissions to the atmosphere and thus reduce global warming. Nawaz et al. (2017) reported that on a silt loam soil mulching reduced bulk density and improved total soil porosity. More TC (16.16 g kg⁻¹), SOC (8.36 mg L⁻¹) and soil microbial biomass carbon (152 µg g⁻¹) were recorded in soil under NT than plough tillage (PT). Mulch application also decreased soil temperature (0–5 cm) and penetration resistance (0–60 cm). Adoption of long-term NT reduced the GHG emission. Average fluxes of GHGs under NT were 1.84 g CO₂ –C m⁻² day⁻¹, 0.07 mg CH₄–Cm⁻² day⁻¹ and 0.73 mg N₂O–Nm⁻² day⁻¹ compared with 2.05 g CO₂–Cm⁻² day⁻¹, 0.74 mg CH₄–Cm⁻² day⁻¹ and 1.41 mg N₂O–Nm⁻² day⁻¹, respectively, for PT. Mulching had no impact on CO₂ and CH₄ (methane) emissions but enhanced N₂O (nitrous oxide) emissions.

Effects on crop growth and input-use efficiency

Modification of crop growth

The effect of mulches on crop growth depends upon the combination of climate, type of crop and soil. Mulching influences three principal factors affecting crop growth, i.e., weed competition, seedling emergence and root growth, which influence crop productivity.

Weed competition

Straw mulching suppresses weed growth through delayed emergence, reduced weed count and a smothering effect. Weed control efficacy depends on the type of mulch material used. Black plastic mulches are more effective for weed control than organic mulches in thin layers as light does not penetrate through them.

Seedling emergence

Moisture deficit or unfavorable soil temperature in the seed zone limit seedling emergence even though there may be sufficient water lower in the soil profile. Mulch keeps the surface layers wetter for longer, which enhances seed germination and seedling emergence. The time required for emergence initiation and maximum emergence are shorter under mulches. Straw mulch significantly increases percentage of seedling emergence both under dry and rainfall situations (Prihar, 1986). Under dry conditions, seedling emergence increases due to lower maximum temperature and reduced evaporation losses under mulches, whereas under rainfall conditions seedling emergence increases due to retardation of crust formation, which otherwise hampers seedling emergence.

Root growth

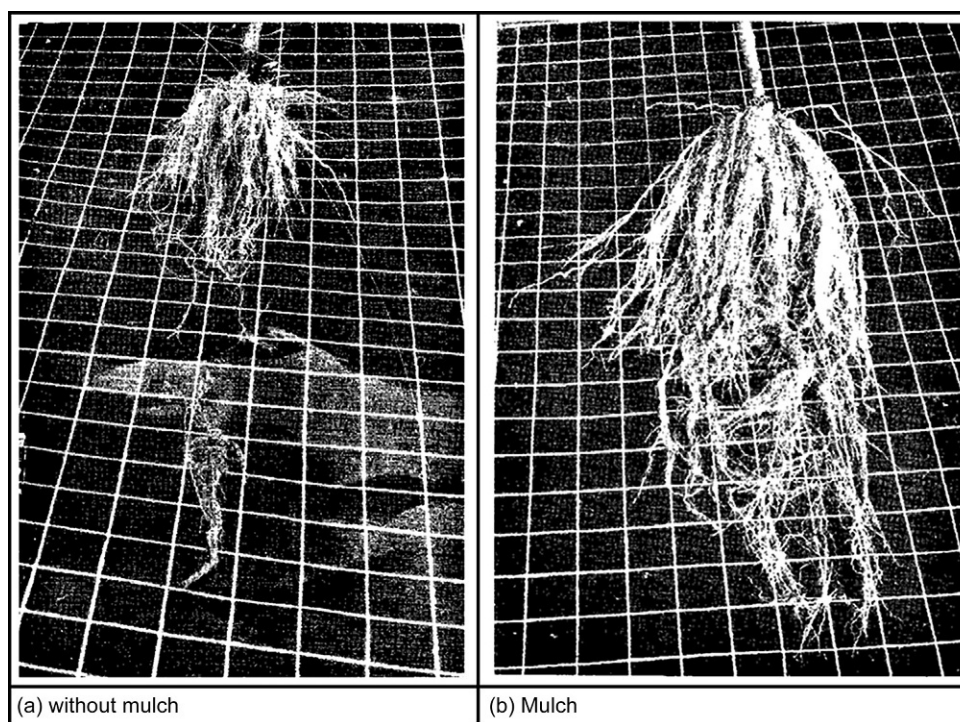
Mulching favorably influences root growth and distribution in the upper soil layer due to favorable soil temperature and higher soil water content under mulches. Higher soil water content in the surface layer reduces the mechanical resistance to growing roots. Application of organic mulches to the previous standing crop helps in conserving and carrying over residual soil moisture at the time of sowing of the subsequent rainfed crop, leading to better root proliferation and higher crop yield, in regions where weather starts cooling in the post rainy season (Table 5).

In Alfisols of the north western Himalayas favorable effects of no-tillage with mulch on root growth was recorded compared to no-tillage (NT) without mulch in a maize–wheat sequence (Acharya and Sharma, 1994). The no-tillage without mulch showed a significant increase in the bulk density at 0–0.15-m depth. Mulching with pine needle at 10 t ha⁻¹ with NT, however, prevented compaction of the untilled layer at the surface and showed a significantly lower bulk density at the surface (0.15 m) depth. It was interesting to note that, even after heavy rain, no surface stagnation of rain water at any time was observed in plots under NT with mulch compared to NT without mulch. Furthermore, mulching under both no-tillage and conventional tillage in a silty clay loam soil significantly increased root length and rooting density of crops owing to the evident improvements in soil water, temperature and bulk density. Acharya and Bhagat (1984) reported that pine needle mulch applied on conventionally cultivated plots resulted in much higher rooting density of maize at depths below 10 cm compared to that under no mulch (Plate 1). This was ascribed to moderation of the soil physical environment that also resulted in better drainage of water during heavy rains. In a sandy loam soil, application of wheat residue mulch resulted in a significant increase in root length density and root mass density of maize at 0–15 and 15–30-cm soil depth due to better availability of soil moisture and favorable soil temperatures under mulching (Chatterjee et al., 2017). Acharya et al. (1998) reported that there was a significant increase in the root mass density of rainfed wheat under minimum tillage and lantana mulch that moderated the hydrological and thermal regimes compared to repeated conventional tillage in a maize–wheat cropping system. Thus, during prolonged dry spells the deeper roots may extract water stored in the lower layers of the soil profile and would help in mitigating the adverse effect of prolonged drought on crop growth.

Table 5 Effect of tillage and nitrogen on root growth and grain yield of rainfed wheat (*Triticum aestivum* L.) in a silty-clay loam soil (Typic Hapludalf).

Mulch Tillage (T)	Nitrogen	Root mass density (kg m ⁻³)		Grain yield (Mg ha ⁻¹)
		Tillering	Flowering	
T ₁ : Mulching previous standing maize (<i>Zea mays</i> L.) with wild sage (<i>Lantana camara</i> L.) followed by its incorporation with conventional tillage immediately prior to sowing of wheat	N ₆₀	0.19	0.61	2271
	N ₁₂₀	0.21	0.67	2711
T ₂ : As T ₁ but wheat sowing following conservation tillage and retaining the residues of wild sage as mulch	N ₆₀	0.26	0.68	2532
	N ₁₂₀	0.29	0.78	2942
T ₃ : As T ₁ but mulching by eupatorium (<i>Eupatorium perfoliatum</i> L.) instead of wild sage	N ₆₀	0.20	0.58	1850
	N ₁₂₀	0.24	0.65	2330
T ₄ : As T ₂ but mulching with eupatorium instead of wild sage	N ₆₀	0.25	0.60	2252
	N ₁₂₀	0.29	0.72	2623
T ₅ : Farmer's practice of conventional tillage after maize harvesting	N ₆₀	0.13	0.41	1231
	N ₁₂₀	0.17	0.53	1282
LSD (<i>P</i> = 0.05)	Mulch—Tillage (T)	0.05	0.06	433
		0.03	0.04	274
	Nitrogen (N)	0.07	0.08	610
	Interaction (T X N)			

Adapted from Acharya CL, Kapur OC and Dixit SP (1998) *Soil and Tillage Research*, 46, 153–163.

**Plate 1** Rooting pattern of 60-days old maize (*Zea mays* L.) plants under conventional tillage (a) without mulch and (b) with wheat (*Triticum aestivum* L.) straw mulch. Reproduced from Acharya CL and Bhagat RM (1984) *Proceedings of the Indian National Science Academy Part B* 50 (4): 441–448.

Improvement of input-use efficiencies

The favorable effects of mulching on crop growth and yields through an improved soil environment enhance both water- and nutrient-use efficiency. Increases in water availability in the upper layers because of a reduction in evaporation and runoff losses under mulches is one of the reasons for increased input-use efficiency. Furthermore, modification of the soil hydrological and thermal regimes under mulches favor root proliferation, which can tap water and nutrients from a larger soil volume, thus improving water- and nutrient-use efficiency. Mulching results in better coordination of demand and supply of water resulting in complete extraction of profile-stored water, which improves water-use efficiency. As the grain yield–water use relation is linear

within the domain of limited water supplies, any increase in water supply beyond a minimum threshold would be directly reflected in increased grain yield and water-use efficiency (WUE).

Higher soil water content, particularly in the initial stages of crop growth, results in greater total water use by the crop, leading to more plant growth and higher yields. Another factor contributing significantly to higher yield is more efficient use of growing season precipitation when the surface is covered with mulches. The final consequence of higher yields with relatively small increases in total water use is higher WUE with mulching compared to without mulching.

Application of pine needles from pine forests as mulch at the time of sowing of some tuber crops like potato (*Solanum tuberosum* L.) improves the yield and WUE and results in the saving of nitrogen fertilizer and irrigation water (Table 6) (Acharya and Kapur, 2001). Hence, within certain limits, mulches can act as a substitute for both irrigation water and nitrogen fertilizer.

With increased water availability under mulching, crop response to fertilizer increases, thereby, improving the fertilizer-use efficiency. Enrichment of soil with plant nutrients through organic mulching materials boosts crop production. Furthermore, mulching helps in moderating the soil's thermal regime, which has a profound influence on nutrient transformations in soil, root growth and nutrient uptake by crop plants, and enhances nutrient-use efficiency by crops. Chatterjee et al. (2017) reported that application of wheat residue mulch to a maize crop on a sandy loam soil resulted in significant increase in root length density, grain yield, and water- and nitrogen-use efficiency by maize.

In rainfed areas, sowing of crops, especially the winter crops in the post monsoon season, is delayed by late arrival of winter rains and hence the efficiency of water- and nutrient-use is low. Under such situations, the presence of soil surface cover in the form of an organic mulch compensates for the crop growth and grain yield compared to unmulched bare soil. Sharma and Acharya (2000) reported that the presence of *Lantana camara* L. mulch (retained over the surface after harvest of maize in the post monsoon season) and delayed sowing of rainfed wheat under no-tillage together with *Lantana camara* mulch after the late arrival of winter rain compensated more in terms of crop growth and grain yield than sowing rainfed wheat under conventional cultivation, but without mulch (as farmers practice).

Das et al. (2015) reported similar results of higher grain yield and thereby increased input-use efficiency from the north-east Indian Himalayan region with in-situ soil moisture conservation in a maize–toria (*Brassica campestris* L.) system using residues of the preceding rainy season maize crop and mulching with locally available weed biomass of common ragweed (*Ambrosia artemisiifolia* L.), applied before the sowing of toria. The productivity of toria was enhanced by about 99%, due only to the retention of maize stalk cover (MSC) as a mulch. Mulching with MSC + *Ambrosia* sp. (5 t ha^{-1}) + poultry manure (5 t ha^{-1}) recorded the highest seed yield for toria (four-year average: 641 kg ha^{-1}), which was 228 and 64% higher than no mulching (control) and MSC alone, respectively. MSC + *Ambrosia* mulch 10 t ha^{-1} gave the highest net returns and benefit: cost (B:C) ratio of the maize–toria system. The overall B:C ratios were better under zero tillage (ZT) than conventional tillage (CT). Similarly, in the same region, double mulching with in-situ maize stover mulch (MSM) + fresh biomass of white hoary pea (WHP-*Tephrosia candida* DC.) and MSM + fresh biomass of ragweed (RW0) showed improved soil moisture content, relative leaf water content of crops during the dry season, higher soil organic carbon (SOC) content and stocks at 0–15-cm depth, maximum soil microbial biomass carbon and dehydrogenase activity, and the highest system and water productivity than for no mulch (Ngangom et al., 2020).

Furthermore, the presence of mulch on the surface in conservation agriculture increased the average yield of wheat by 2–4% compared to conventional till wheat and saved in the cost of labor, fuel, chemicals, etc. It saved about 20 L of fuel per hectare due to sowing wheat in a single operation, increased nutrient-use efficiency by continuous recycling of residues over 3–4 years, and produced the same yield with 30–40 kg ha^{-1} less nitrogen use and significantly higher (10–15%) nutrient-use efficiency, produced more crop per drop of water through reduction in evaporation loss equivalent to about 45 mm (0.45 million liter) during the wheat season. It also reduced the risk of biotic and abiotic stresses, improved the environment by reduction in GHGs, reduced air pollution caused by rice crop residue burning, reduced terminal heat effects, as straw mulch lowered the canopy temperature in wheat which

Table 6 Tuber yield and water use efficiency (WUE) of potato (*Solanum tuberosum* L.) as influenced by mulching and N application (pooled over 4 years).

Treatment ^a	Autumn potato		Spring potato	
	Yield (Mg ha^{-1})	WUE ($\text{kg ha}^{-1} \text{ cm}^{-1}$)	Yield (Mg ha^{-1})	WUE ($\text{kg ha}^{-1} \text{ cm}^{-1}$)
60 kg N ha^{-1}	8.46	479.9	13.30	257.0
120 kg N ha^{-1}	9.86	528.2	15.15	271.2
Pine needle (<i>Pinus sylvestris</i> L.) mulch @ 10 Mg ha^{-1} + 60 kg N ha^{-1}	12.69	711.0	16.33	340.9
Pine needle mulch @10 Mg ha^{-1} + 120 kg N ha^{-1}	14.52	735.5	18.34	395.4
LSD (P = 0.05)	1.19	—	1.46	—

^aAll treatments received a uniform dose of farmyard manure (FYM) @ 20 Mg ha^{-1} .

Reproduced from Acharya CL and Kapur OC (2001) *Indian Journal of Agricultural Sciences*, 71, 306–309.

helps in adapting to terminal heat, and resulted in a reduction in air pollution by PM_{2.5} and PM₁₀ particles, black carbon and noxious gases, by preventing burning of crop residues (NAAS, 2017).

Role of cover crops on soil quality improvement

Cover crops contribute to the improvement of soil quality through their decomposition by soil microbes. Their decomposition depends upon their nature, contact between soil and cover crop residue, soil moisture content and temperature, and microbial population. Decomposition of cover crops adds to the soil organic matter, which improves the soil physical condition and also contributes to the fertility status of soil. The degree of enrichment depends on the quantity and quality of cover-crop biomass.

Cellulose-rich plants or plant parts degrade far more rapidly than lignin-rich plants such as mature grasses. Leaf portions of the shoot system degrade far more rapidly than those of the stem. With decomposition, the resulting increases in SOM benefit the soil by improving its structure, thus enhancing gaseous exchange, internal drainage, nutrient exchange and water retention. Increased SOM leads to greater soil biodiversity, which is at the center of all SOM dynamics and increased aggregate stability which, in turn, means a reduced tendency to erode (Edwards and Burney, 2005).

From the standpoint of soil nutrient enrichment, cover crops contribute to the nutrient pool through the mineralization of decaying biomass under the influence of soil biota, which release the full range of plant nutrients in an available form. Where the decaying material is leguminous, nitrogen enrichment is predominant because of the low C/N ratio of this family of plants and, particularly, the capacity of their roots to fix atmospheric nitrogen in symbiotic association with bacteria. Thus, by root rupture or necrosis, the soil eventually benefits from this nitrogen-fixing activity. *Pueraria phaseoloides* L., a leguminous cover crop popular in the tropics, accumulated between 150 and 250 kg N ha⁻¹ within 4–18 months in parts of West Africa. The resulting nutrient enrichment of the soil following cover-crop fallows generally means an increase in crop production and savings on fertilizer costs, particularly where fertility is poor in areas such as Latin America and Africa. Tropical leguminous cover crops are also known for their foliar abundance. Some species contribute over 95% to total litter fall and are particularly favored in orchards and tree plantations if they are shade-tolerant and non-competitive (Edwards and Burney, 2005).

Role of cover crops on runoff and soil erosion control

Increased foliar mass of the cover crop reduces splash erosion with increasing ground coverage because of interception of the rainfall impact, which can be incorporated as a crop management factor in calculating soil loss. Cover crops work by reducing splash detachment and by intercepting the splashed soil, thus limiting its trajectory movement (Edwards and Burney, 2005). The effectiveness of cover crops in reducing soil erosion depends on its ground coverage, i.e., leaf area index and root structure. Parsons et al. (1994) reported that a comparison of cover-management effectiveness under varying degrees of canopy cover with a field rainfall simulator, showed that soil loss was not affected by a canopy of potatoes of up to 5%. A canopy of 70% was required before the crop cover could keep the soil reasonably in place, and a 90% canopy was needed before it was as effective as a mature stand of red clover (*Trifolium aestivum* L.) in the same study. Over longer distances, the progressive build-up of runoff generates an increasing shear stress on the soil surface. As the effects of runoff increase relative to the constant effects of splash downslope, the root structure of the cover crop becomes increasingly important for keeping the soil in place.

Recommendations and policy decisions on mulching and cover crops

Appropriate policy decisions may be taken for the wide adaptability of mulch and cover crop technology as suggested below.

- Proper research and development activities for site-specific conservation agriculture practices and mulching should be promoted.
- Awareness should be developed among farming communities about the importance of mulching, cover crops and conservation agriculture using field demonstration and extension services
- Legislation should be enacted to stop burning of crop residues by farmers. Some incentives to the farmers may be given to adopt conservation agriculture with minimum tillage and spreading of crop residues as a mulch rather than their burning.
- Practicing all three principles of conservation agriculture yields the desired benefits. Promoting conservation agricultural practice involving reduced tillage, residue retention and improved crop rotation including legumes, is a step toward sustainable agriculture that not only conserves water and favorably moderates physical, chemical and biological soil health and facilitates soil organic carbon sequestration, but also helps with greater nutrient-, water- and energy-use efficiency and minimizing soil erosion.
- Production of machinery and assured access by farmers to perform conservation operations including minimum till, seed and fertilizer placement etc. in one go. For small and marginal farmers, this machinery should be made available for hire.
- Growing short duration legumes in cereal based rotations should be promoted by giving some incentives to the farmers because it can act as a cover crop and improve soil health, and could provide pulses to meet the nutritional requirement of the world's growing population.

Conclusions

Mulching has a great potential for improving the soil's physical, chemical and biological health. This practice also helps in improving root and shoot growth, productivity and input-use efficiency. Further, this practice can improve environmental quality by enhancing soil organic carbon sequestration and reducing greenhouse gas emissions. However, this technology has some limitations. Appropriate policy decisions should be taken for addressing these limitations with the provision of technical and financial support to farmers. Together with mulching, short duration pulses could be promoted as cover crops in cereal based rotations to improve soil health, reduce soil erosion, and improve crop productivity and farmers' income. All three principles of conservation agriculture need to be followed to achieve the desired benefits. Site-specific mulching, cover cropping and conservation agricultural practices should be developed for sustainable climate-smart agriculture.

References

- Acharya CL and Bhagat RM (1984) Infiltration behaviour, root development and yield of rainfed maize (*Zea mays* L.) under different soil management. *Proceedings of the Indian National Science Academy B* 50(4): 441–448.
- Acharya CL and Kapur OC (2001) Using organic wastes as compost and mulch for potato (*Solanum tuberosum*) in low water-retaining hill soils of north-west India. *Indian Journal of Agricultural Sciences* 71: 306–309.
- Acharya CL and Prihar SS (1969) Vapour losses through soil mulch at different wind velocities. *Agronomy Journal* 61: 666–668.
- Acharya CL and Sharma PD (1994) Tillage and mulch effects on soil physical environment, root growth, nutrient uptake and yield of maize and wheat on an Alfisol in north-west India. *Soil and Tillage Research* 32: 291–302.
- Acharya CL, Masand SS, Angiras NN, Kapur OC, and Sharma PK (1993) Conservation tillage with alley farming in relation to soil erosion and crop production in rainfed conditions of western Himalayas. *Indian Journal of Hill Farming* 6: 69–77.
- Acharya CL, Kapur OC, and Dixit SP (1998) Moisture conservation for rainfed wheat production with alternative mulches and conservation tillage in the hills of north-west India. *Soil and Tillage Research* 46: 153–163.
- Bandyopadhyay PK, Saha S, Mani PK, and Manda B (2010) Effects of organic inputs on aggregate associated organic carbon concentration under long-term rice–wheat cropping system. *Geoderma* 154: 379–386.
- Bhagat RM and Acharya CL (1987) Effect of soil management on rainfed wheat in northern India. 1. Hydrothermal regime and root growth. *Soil and Tillage Research* 9: 65–77.
- Blanco-Canqui H and Lal R (2009) Crop residue removal impacts on soil productivity and environmental quality. *Critical Reviews in Plant Sciences* 28(3): 139–163.
- Blanco-Canqui H, Lal R, Post WM, and Owens LB (2006) Changes in long-term no-till corn growth and yield under different rates of stover mulch. *Agronomy Journal* 98: 1128–1136.
- Bond JJ and Willis WD (1969) Soil water evaporation: Surface residue rate and placement effects. *Soil Science Society of America Proceedings* 33: 445–448.
- Chatterjee S, Bandyopadhyay KK, Pradhan S, Singh R, and Datta SP (2016) Irrigation, crop residue mulch and nitrogen management practices influence on soil physical quality. *Journal of the Indian Society of Soil Science* 64: 351–367.
- Chatterjee S, Bandyopadhyay KK, Pradhan S, Singh R, and Datta SP (2017) Yield and input use efficiency of maize (*Zea mays* L.) as influenced by crop residue mulch, irrigation and nitrogen management. *Journal of the Indian Society of Soil Science* 65: 199–209.
- Chatterjee S, Bandyopadhyay KK, Pradhan S, Singh R, and Datta SP (2018) Effects of irrigation, crop residue mulch and nitrogen management in maize (*Zea mays* L.) on soil carbon pools in a sandy loam soil of Indo-gangetic plain region. *Catena* 165: 207–216.
- Chaudhary RS and Acharya CL (1993) A comparison of evaporative losses from soil of different tilths and under mulch after harvest of rice. *Soil and Tillage Research* 28: 191–199.
- Das A, Ghosh PK, Verma MR, Munda GC, Ngachan SV, and Mandal D (2015) Tillage and residue mulching effect on productivity of maize (*Zea mays*)—toria (*Brassica campestris*) cropping system in fragile ecosystem of Northeast India Himalayas. *Experimental Agriculture* 51(1): 107–125.
- Doran JW, Sarrantonio M, and Liebig M (1996) Soil health and sustainability. In: Sparks DL (ed.) *Advances in Agronomy*. vol. 56, pp. 1–54. San Diego: Academic Press.
- Edwards L and Burney J (2005) Cover crops. In: *Earth Systems and Environmental Sciences*, pp 311–318. <https://doi.org/10.1016/B978-0-12-409548-9.11654-9>.
- Erenstein O (2002) Crop residue mulching in tropical and semi-tropical countries: An evaluation of residue availability and other technological implications. *Soil and Tillage Research* 67: 115–133.
- Foster GR, Johnson CB, and Moldenhauer WC (1982) Hydraulics of failure of unanchored corn stalk and wheat straw mulches for erosion control. *Transactions of the American Society of Agricultural Engineers* 25: 940–947.
- Gilley JE, Finkner SC, and Varvel GE (1986) Runoff and erosion as affected by sorghum and soybean residue. *Transactions of the American Society of Agricultural Engineers* 29: 1605–1610.
- Hati KM, Swarup A, Singh D, Misra AK, and Ghosh PK (2006) Long-term continuous cropping, fertilization and manuring effects on physical properties and organic carbon content of a sandy loam soil. *Australian Journal of Soil Research* 44: 487–495.
- Horton R, Kluitenberg GJ, and Bristow KL (1994) Surface crop residue effects on the soil surface energy balance. In: Unger PW (ed.) *Managing Agricultural Residues*, pp. 143–162. Boca Raton, FL: CRC Press.
- Jordán A, Zavala LM, and Gil J (2010) Effects of mulching on soil physical properties and runoff under semi-arid conditions in southern Spain. *Catena* 81: 77–85. <https://doi.org/10.1016/j.catena.2010.01.007>.
- Lal R (1987) Farming systems. In: Lal R (ed.) *Tropical Ecology and Physical Edaphology*, pp. 618–685. Wiley-Interscience Publication.
- Lal R (2005) World crop residues production and implications of its use as a biofuel. *Environment International* 31: 575–584.
- Leroy BLM, Herath HSMK, Sleutel S, De Neve S, Gabriels D, Reheul D, and Moens M (2008) The quality of exogenous organic matter: Short-term effects on soil physical properties and soil organic matter fractions. *Soil Use and Management* 24: 139–147.
- Morgan RPC (2005) *Soil erosion and conservation*, 3rd edn. UK: Blackwell Publishing Ltd. p. 304.
- NAAS (2017) *Innovative Viable Solution to Rice Residue Burning in Rice-Wheat Cropping System Through Concurrent Use of Super Straw Management System Fitted Combines and Turbo Happy Seeder*. Policy Brief No. 2. New Delhi: National Academy of Agricultural Sciences. p. 16.
- Nawaz A, Lal R, Shrestha RK, and Farooq M (2017) Mulching affects soil properties and greenhouse gas emissions under long term No-Till and Plough-Till systems in Alfisols of Central Ohio. *Land Degradation and Development* 28: 673–681.
- Ngangom B, Das A, Lal R, Idapuganti RG, Layek J, Basavaraj S, Babu S, Yadav GS, and Ghosh PK (2020) Double mulching improves soil properties and productivity of maize-based cropping system in eastern Indian Himalayas. *International Soil and Water Conservation Research* 8(3): 308–320.
- Parsons TS, Burney JR, and Edwards LM (1994) *Canadian Agricultural Engineering* 36: 127–133.
- Prihar SS (1986) Fertilizer and water use efficiency through the aid of mulching. In: *Proceedings of Seminar Organized by Fertilizer Association of India (Northern Region) at Dehradun, India, May 10–11, 1986*, pp. 31–44.
- Prihar SS, Singh B, and Sandhu BS (1968) Influence of soil and climatic environment on evaporation losses from mulched and unmulched pots. *Journal of Research Ludhiana* 5: 320–328.

- Prihar SS, Jalota SK, and Steiner JL (1996) Residue management for reducing evaporation in relation to soil type and evaporativity. *Soil Use and Management* 12: 150–157.
- Rabary B, Sall S, Letourmy P, Husson O, Ralambofetra E, Moussa N, and Chottel JL (2008) Effects of living mulches or residue amendments on soil microbial properties in direct seeded cropping systems of Madagascar. *Applied Soil Ecology* 39: 236–243.
- Ranaivoson L, Naudin K, Rapoche A, Affholder F, Rabcharisoa L, and Corbeels M (2017) Agroecological functions of crop residues under conservation agriculture. A Review. *Agronomy for Sustainable Development* 37(26): 1–17.
- Sharma PK and Acharya CL (2000) Carry-over of residual soil moisture with mulching and conservation tillage practices for sowing of rainfed wheat (*Triticum aestivum*L.). *Soil and Tillage Research* 57: 43–52.
- Unger PW (1990) Conservation tillage systems. *Advances in Soil Science* 13: 27–68.
- Yadav RL, Shukla SK, Suman A, and Singh PN (2009) Trichoderma inoculation and trash management effects on soil microbial biomass, soil respiration, nutrient uptake and yield of ratoons sugarcane under subtropical conditions. *Biology and Fertility of Soils* 45: 461–468.
- Zhang S, Wang Y, Sun L, Qiu C, Ding Y, Gu H, Wang L, Wang Z, and Ding Z (2020) Organic mulching positively regulates the soil microbial communities and ecosystem functions in tea plantation. *BMC Microbiology* 20: 103–116.

The potential agronomic and environmental applications of biochar: Prospects and challenges

Yu-Jun Wang^a, Ping Wu^b, Nanthi S Bolan^c, and Hai-Long Wang^d, ^aKey Laboratory of Soil Environment and Pollution Remediation, Institute of Soil Science, Chinese Academy of Sciences, Nanjing, Jiangsu, China; ^bCenter for Agricultural Resources Research, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Shijiazhuang, China; ^cGlobal Centre for Environmental Remediation and International Centre for Balanced Land Use, University of Newcastle, Callaghan, NSW, Australia; ^dBiochar Engineering Technology Research Center of Guangdong Province, School of Environmental and Chemical Engineering, Foshan University, Foshan, Guangdong, China

© 2023 Elsevier Ltd. All rights reserved.

Introduction	414
Typical properties of biochar	415
Heat treatment temperature (HTT)	415
Feedstock composition	416
Residence time	416
Rate of heating	416
Biochar applications in agriculture and other environments	416
Climate change mitigation	416
Soil improvement	417
Salinity and drought stress amelioration	417
Immobilization of toxic metals	418
Removal of organic pollutants	419
Unintended consequences	420
Conclusions and perspectives	422
References	423

Abstract

Biochar is a pyrogenic carbonaceous material with well-developed porous structure and tunable functionality, and has triggered great interests in its agronomic and environmental benefits. However, unintended consequences related to biochar application are poorly understood. This chapter provides a systematic review on the structural development and characterization of biochar, which impact its function as a soil amendment including carbon sequestration, soil improvement, salinity and drought stress amelioration, and remediation of polluted soils. The relationships between biochar properties and its applications as well as the potential risks on plants and humans are also analyzed critically. Future research foci are suggested for green and sustainable applications of biochar.

Key points

- Feedstock and strategies for biochar production and modification are described.
- Effects of feedstock used and pyrolysis conditions on biochar's physicochemical properties are discussed.
- Agricultural and environmental benefits of biochar application to soils are reviewed.
- Potential ecological and human health risks associated with biochar application to soils are discussed.
- Future research outlooks are proposed.

Introduction

Biochar is a carbon-rich porous material produced from the thermal decomposition of raw biomass under an oxygen-limited atmosphere. Various biomass materials have been used for biochar production including agricultural, forestry, food and garden wastes, livestock carcasses and manure, sewage sludge, biosolids, and aquatic organisms such as micro- and macro-algae. The strategies for biochar production include slow and fast pyrolysis, hydrothermal carbonization, flash carbonization, gasification, and torrefaction (Fig. 1). Slow pyrolysis and hydrothermal carbonization are the most predominantly used methods for biochar production. Biochar is a heterogeneous mixture of an amorphous less-carbonized fraction and a microcrystalline graphitic structure. It is characterized by well-developed porosity, large specific surface area (SSA), and abundant oxygen-containing functional groups (OFGs) and minerals. Specific properties of biochar are mainly regulated by feedstock composition and heat treatment temperature (HTT) (Lehmann, 2007; Yang et al., 2021).

Biochar has gained increasing attention because of its potential agronomic and environmental benefits through carbon sequestration, enhancing soil productivity, alleviating salinity and drought stress, and immobilizing contaminants in soils. The recalcitrant nature of biochar's carbon structure and its strong adsorption capacity to CO₂, together with sustainable biochar

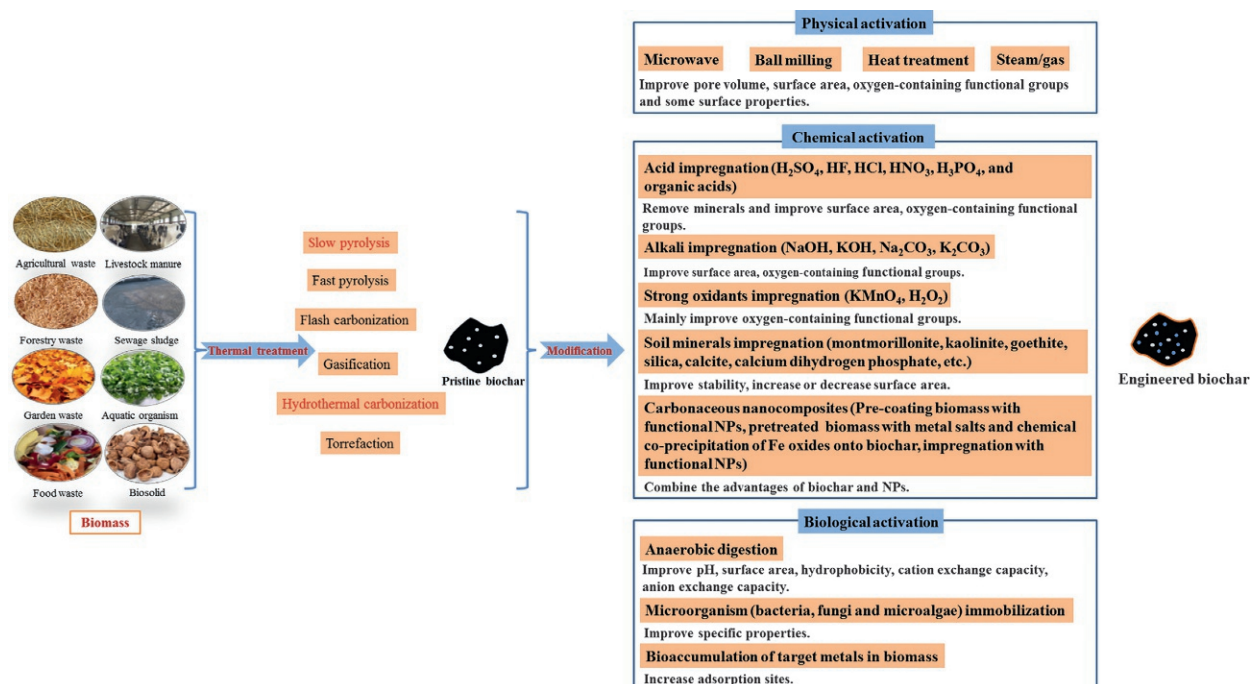


Fig. 1 Biochar production and modification.

production and soil application have the potential to enhance long-term carbon storage and remove carbon from the atmosphere, thereby mitigating climate change. Biochar also provides a potential route to increasing plant productivity and crop yields by improving soil quality and soil fertility levels. The improvement of soil quality and fertility after biochar incorporation into soil can also contribute to the alleviation of salinity and drought stress to plants. Finally, biochar's capability to immobilize harmful contaminants in soil makes it an effective material to improve food safety in areas of contaminated soils. However, biochar's functions as a soil amendment are greatly regulated by its physicochemical properties. Additionally, biochar aging in soils may cause considerable changes in its physicochemical properties, thereby affecting biochar's applications in agriculture and environment (Saifullah et al., 2018; Wang et al., 2020).

Although there are various agronomic and environmental benefits of biochar application to soil, potential risks related to ecological and human health are of great concern and need to be addressed before large-scale application in agricultural practices. It has been reported that some toxic substances inherent in biochar such as polycyclic aromatic hydrocarbons (PAHs) and heavy metals may be released into soils, having negative effects on soil biota. Persistent free radicals (PFRs) present in biochar that depend on feedstock and production conditions can induce the generation of reactive oxygen species (ROS), thus causing agro-environmental risks (Godlewska et al., 2021). Compared to the numerous studies outlining the multiple functions of biochar in agriculture and the environment, the potential risks of biochar application in soils have seldom been addressed.

The main objective of this chapter is to provide a balanced perspective on the beneficial effects and critical issues of biochar application to soil considering both the agricultural ecosystem and also wider environmental aspects. The aims of this chapter are to (i) analyze structural development and characterization of biochar in the context of heat treatment conditions; (ii) highlight potential agronomic and environmental benefits of biochar application to soil; and (iii) discuss potential unintended consequences. Future research outlooks are also put forward.

Typical properties of biochar

Heat treatment temperature (HTT)

Heat treatment temperature (HTT) is considered to be a dominant parameter influencing biochar properties. Broadly speaking, carbon and mineral contents, pH, alkalinity, aromaticity, surface area, porosity, and zeta potential increase with increasing HTT. While yield, albedo, cation exchange capacity (CEC), oxygen-containing functional groups (OFGs), nitrogen and sulfur contents, and aliphaticity/polarity decrease with increasing HTT. The hydrophobicity of biochar first increases, then decreases with increasing HTT (Fig. 2) (Xiao et al., 2018). The effect of HTT on biochar characteristics is attributed to the loss of volatile organic compounds (VOCs), nitrogen (N) and sulfur (S), and redistributions of its pore structure.

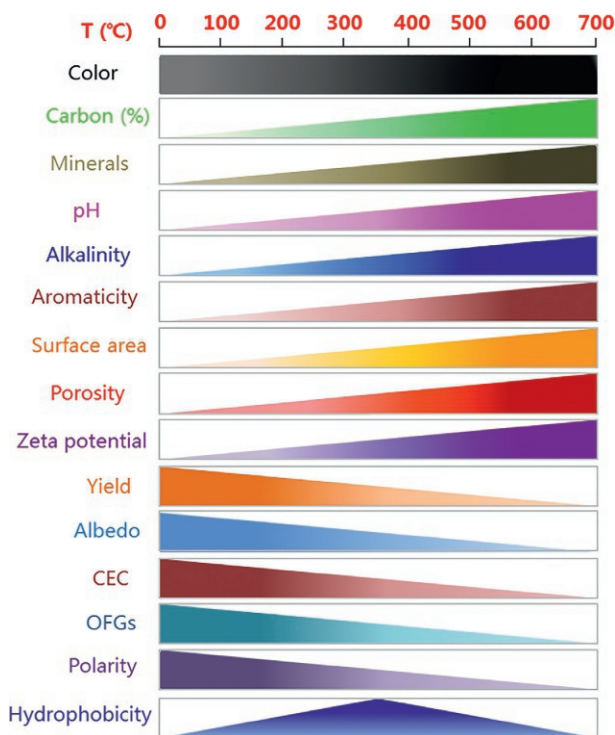


Fig. 2 Biochar property variations with increasing heat treatment temperatures.

Feedstock composition

The composition of biomass also plays an important role in biochar properties. Biomass (e.g., wood, pine, bamboo) with higher contents of lignin results in larger yields, carbon contents, fixed carbon content, greater surface area and porosity, and lower ash contents. Wet feedstock including livestock manure and sewage sludge derived biochar has higher ash and nutrient contents, CEC, pH and alkalinity, but a lower carbon content, surface area and porosity compared to biochar derived from dry feedstocks like agricultural, forestry and garden wastes. Moreover, wet feedstock-based biochar contains less fixed carbon compared with biochar produced from dry biomass (Lian and Xing, 2017; Nidheesh et al., 2021). Therefore, it is reasonable to speculate that dry biomass-based biochar with a large fixed carbon content are of high quality, especially in relation to long-term carbon storage which suggests it can contribute to climate change mitigation.

Residence time

Residence time is also a factor affecting biochar properties. Research indicates that the surface area and porosity initially increase, then decrease with increasing residence time. Sufficient reaction time accelerates the release of volatile matter and the development of pore structure. However, prolonged exposure to high temperatures may obliterate the pore structures of biochar. Increasing residence time may lead to an increase in fixed carbon content (Leng et al., 2021).

Rate of heating

The rate of heating also affects biochar's surface area and porosity. At first, an increase followed by a decrease in surface area and pore volume of biochar with an increase in rate of heating was proposed. Increasing the rate of heating accelerates heat and mass transfer, thus increasing surface area and total pore volume of biochar. However, an excessively high rate of heating may lead to the rapid diffusion of volatile matter, thus blocking the entrance of pore channels (Leng et al., 2021).

Biochar applications in agriculture and other environments

Climate change mitigation

Biochar has potential to mitigate climate change through carbon sequestration by storing carbon within the biochar itself and by capturing CO₂ from the atmosphere as well as by reducing CO₂ emissions. The abundant OFGs on the surface of biochar contribute to the adsorption of CO₂. Additionally, biochar application to soils reduces the content of labile C in soils, thus reducing CO₂ emissions from soils (Lehmann, 2007). Greenhouse gas (e.g., CH₄ and N₂O) emissions from soils may also be affected by

the application of biochar to soil. Inhibition of methanogenic archaeal activity, increase in the *pmoA* gene abundance and activity of methanotrophs, and decrease in the ratio of methanogenic to methanotrophs abundances have been found to be responsible for the reduction of CH₄ emissions from soil after biochar application. Of these processes the increased activity of methanotrophic proteobacteria and decreased ratio of methanogens to methanotrophs abundances are mostly responsible for the reduction of CH₄ emissions. The mechanisms responsible for the reduction in N₂O emissions mainly include: (i) biochar liming effect increases *nosZ* gene abundances, leading to an accelerated reduction of N₂O to N₂; (ii) biochar increases the adsorption of NO₃⁻ and/or NH₄⁺, decreasing the availability of inorganic nitrogen substrate for nitrification and denitrification; (iii) improves soil aeration and O₂ content, restraining the denitrification process; (iv) inhibits nitrification or denitrification processes by releasing toxic compounds such as PAHs into soils; and (v) increases the availability of dissolved organic carbon (DOC), leading to complete denitrification (Xiao et al., 2018; Yang et al., 2021).

Soil improvement

In agriculture, biochar can improve soil fertility by providing direct nutritional value and improving nutrient-use efficiency. Biochar can be used as a slow-release fertilizer by releasing inherent mineral fractions (e.g., N, phosphorus (P), potassium (K), calcium (Ca), magnesium (Mg)) into soils. The slow-releasing property of biochar is due to its unique structure. The porous network within biochar and the unique association of nutrient elements with the carbon matrix contribute to slow nutrient release. These properties also endow it with strong sorption capacity for nutrients, enabling their concentration followed by slow desorption into soil (Hagemann et al., 2017; Xiao et al., 2018). The loading of metallic (oxyhydr)oxide nanoparticles (NPs) on biochar increased the sorption performance for phosphate and other nutrient elements. Therefore, the metallic (oxyhydr)oxide-biochar nanocomposites have great potential in recycling nutrients from wastewater and holding them in the soil as a fertilizer for a long time (Yao et al., 2013). The feedstock composition plays a crucial role in determining biochar's contents and species of elemental fertilizers. For example, straw-derived biochar contains abundant silicon; soybean (*Glycine max* (L.) Merrill) root-derived biochar is characterized by a rich N fraction; eggshell yields biochar containing a large amount of calcium; sewage sludge-based biochar has abundant phosphate and other minerals. These various attributes of biochar result in its potential use as slow-release fertilizers. Moreover, HTT, soil properties and nutrient element species have a significant influence on the characteristics of biochar fertilizer release. It was reported that N became less available with increasing HTT by transforming N from NH₄⁺ to heterocyclic-N. Low soil pH contributed to the release of N, P and K, while it suppressed the release of dissolved silicon (Si) from biochar. Silicic acid was more available to plants than crystal silicon within biochar (Wang et al., 2019; Xiao et al., 2018).

Biochar also has positive effects on improving soil quality by improving microaggregation and soil structure as well as by increasing soil water holding capacity. The appropriate application of biochar with high ash contents is recommended for improving the quality of acidic soil by increasing the soil pH and ameliorating Al toxicity. The mechanisms for alleviating soil Al toxicity by biochar are delivered through its strong alkalinity and its high sorption capacity to Al (Qian et al., 2013). Furthermore, application of biochar to soil can stimulate microbial biomass and improve microbial community structure by providing bioavailable carbon sources, which indirectly could also improve soil quality. Also, the beneficial rhizosphere environment for microbiota and earthworm populations induced by biochar amendment may result in improved plant growth. On the other hand, biochar application can suppress soil pathogens by triggering beneficial microbial activities, thereby enhancing plant growth.

Salinity and drought stress amelioration

Salinity and drought are major abiotic stresses that limit crop productivity worldwide. Application of biochar has promise for improving soil properties and plant growth in salt-affected soils. Acidic biochar application to salt-affected soils has been shown to be effective in improving soil physical (e.g., porosity, water holding capacity, bulk density, and soil aggregation), chemical (e.g., electrical conductivity, pH, and exchangeable sodium percentage), and biological (e.g., diversity and growth of soil organisms) properties. Laboratory and field studies have shown that acidic biochar application to salt-affected soils alleviated salinity stress on plants and improved plant growth through the direct release of essential macronutrients (e.g., N, K, P) and micronutrients (e.g., copper (Cu), zinc (Zn), iron (Fe), and manganese (Mn)), which offset nutrient deficiency. Indirect benefits from biochar application to salt-affected soils on the improvement in crop productivity include: (i) improved soil physical, chemical, and biological properties; (ii) reduced bioavailability of Na due to its adsorption on biochar; (iii) reduced oxidation stress by decreasing malonaldehyde (MDA), O₂•⁻, and hydrogen peroxide (H₂O₂) contents; (iv) reduced osmotic stress by increasing water-use efficiency; (v) improved stomatal conductance, and photosynthesis efficiency; (vi) enhanced seed germination; (vii) lowered levels of phytohormone production; and (viii) the stimulation of microbial activities (Saifullah et al., 2018). Biochar in conjunction with other materials such as chemical soil amendments, nutrients, and microbes proved to be superior in promoting plant growth and productivity under salt stress compared to biochar alone (Saifullah et al., 2018).

Ammonia (NH₃) volatilization is a major concern in salt-affected soils. The pore structure, large surface area, and abundant OFGs of biochar play an important role in ammonium (NH₄⁺) retention through surface adsorption and ion exchange to NH₄⁺. Reduction of NH₃ volatilization could increase nitrogen-use efficiency in salt-affected soils amended with biochar. It should be noted that a large application of alkaline biochar may facilitate NH₃ volatilization in salt-affected soils resulting from the competitive adsorption between NH₃/NH₄⁺ and salt ions and aggravated inhibition of nitrification (Zhu et al., 2020).

Biochar application was also proposed to mitigate drought stress. It improves water-use efficiency through enhancing the rate of photosynthesis and plant growth. Xylem water potential, stomatal conductance and chlorophyll content are also enhanced.

Additionally, biochar plays an active role in osmoregulation, resulting in reduced sodium (Na) uptake and increased K uptake. Combining biochar with other materials (e.g., silicon (Si) nutrition and plant growth promoting rhizobacteria) had greater positive effects on plant growth under drought stress than the sole application of biochar (Mansoor et al., 2021).

Immobilization of toxic metals

Biochar has diverse influences on the immobilization of toxic metals in soils. Numerous laboratory or field studies on biochar application for the immobilization of cationic heavy metals (e.g., lead Pb(II), cadmium Cd(II), Cu(II), Zn(II), and mercury Hg(II)) in soils have shown promising results. Biochar can immobilize and reduce the bioavailability of metal cations through several mechanisms including cation exchange, electrostatic attraction, surface and inner-sphere complexation, cation- π interaction, and precipitation (Fig. 3) (Ahmad et al., 2014). The abundant OFGs (e.g., hydroxyl, carbonyl, and phenolic groups) on biochar occur mainly in dissociated forms and can be deprotonated at a certain range of pH values, favoring the cation exchange, electrostatic attraction and surface complexation between biochar and metal cations. The abundant alkali and alkaline earth metallic species in biochar also participate in cation exchange with cationic heavy metals. The mineral oxides (e.g., ferric oxide (Fe_2O_3) particles, and manganese oxide (MnO_x) particles) and organic carbon in biochar are responsible for the inner-sphere complexation with cationic metals. The aromatic structure of biochar promotes the cation- π interaction. In addition, co(precipitation) of cationic metals with biochar's inorganic phases containing carbonate, silicate, and phosphate was also widely reported. Biochar-induced precipitates such as hydrocerussite ($\text{Pb}_3(\text{CO}_3)_2(\text{OH})_2$), lead hydroxypyromorphite ($\text{Pb}_5(\text{PO}_4)_3\text{OH}$), azurite ($\text{Cu}_3(\text{CO}_3)_2(\text{OH})_2$), zinc silicate (Zn_2SiO_4), zinc hydroxide ($\text{Zn}(\text{OH})_2$), cadmium carbonate (CdCO_3), and cadmium phosphate ($\text{Cd}_3(\text{PO}_4)_2$) are relatively stable and have a low bioavailability in soils (Wu et al., 2020). The relative contribution of these mechanisms in the immobilization of metals by biochar varies among metal cations. Several studies have reported that cation exchange and surface complexation with OFGs were the predominant mechanisms for the immobilization of Cd(II) and Pb(II). However, other studies have reported that surface complexation and precipitation dominated the sorption of Pb(II). Notably, HTT has been shown to influence the mechanism through which biochar immobilizes metal cations in soils. Despite different contributions of various immobilization mechanisms, some general trends can be identified: (i) cation exchange and surface complexation with OFGs are the dominant mechanisms for cationic metals sorption by biochar produced under low HTTs; whereas, with increasing HTTs the contributions of these mechanisms are superseded by other processes including cation- π interaction and precipitation due to the decrease of OFGs and formation of graphitic structures; (ii) the inorganic phases of biochar, especially if derived from livestock manure and sewage sludge rich in minerals, play an important role in the immobilization of cationic metals through precipitation (Wu et al., 2020). An increase in soil pH induced by the application of biochar (i.e., liming effect) can also accelerate the precipitation and adsorption of cationic metals on soil particles, indirectly enhancing the immobilization of cationic metals. However, an increase in Cu(II) mobility in soils amended with biochar was reported in several studies, which was associated with increased dissolved organic C (DOC) forming soluble Cu-DOC complexes. The discrepancy between Cu(II) activation and immobilization was mainly due to the different contents of biochar DOC. Generally, the DOC contents are relatively high for relatively low-temperature biochar (<500 °C), which has negative effects on the retention of Cu(II) in soils by the formation of soluble Cu-DOC complexes. On the

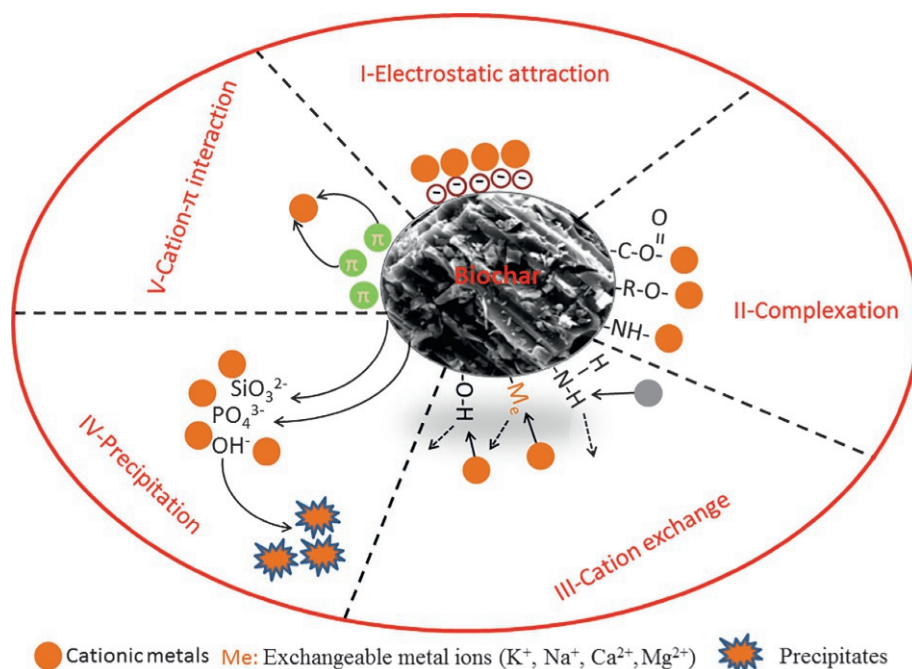


Fig. 3 Schematic diagram showing the various mechanisms of biochar for the immobilization of cationic metals.

other hand, the blocking of biochar pores by DOC prevents physical adsorption of Cu(II). For high-temperature biochar ($>600^{\circ}\text{C}$), DOC is generally deficient, and processes such as electrostatic attraction, cation- π interaction and precipitation can offset the role of DOC in Cu(II) mobilization (Ahmad et al., 2014).

Purposive tailoring of the physicochemical properties of biochar can greatly enhance its effectiveness in the immobilization of heavy metal(loid)s in soils. Biochar tailoring can be achieved by physical, chemical, and biological modifications. Chemical modification, especially involving metal salts or metal oxides has been widely used to reduce the mobility, availability, and toxicity of cationic metals in soils. The attachment of metal oxide NPs on the surface of biochar can not only increase the number of effective active sites but also the electrostatic potential, favoring the immobilization of cationic metals in soils (Wu et al., 2020). In addition to the modification of metal salts or metal oxides mentioned above, heteroatom modification was also found to be effective in improving the sorption capacity for cationic metals. For example, phosphorus-modified biochar effectively reduced Cd(II) and Pb(II) bioavailability by forming precipitates with low solubility such as cadmium hydroxylapatite ($\text{Cd}_5(\text{PO}_4)_3\text{OH}$), $\text{Cd}_3(\text{PO}_4)_2$, pyromorphite ($\text{Pb}_5(\text{PO}_4)_3\text{Cl}$), lead phosphate ($\text{Pb}_3(\text{PO}_4)_2$), and $\text{Pb}_5(\text{PO}_4)_3\text{OH}$. Sulfur modified biochar was also found to reduce Cd, Cu and Hg(II) bio-accessibility remarkably by forming stable sulfide precipitates (Wang and Wang, 2019).

Biological modification associated with microbial immobilization technology has created broad interest in toxic metal immobilization. Metal-resistant microorganism-biochar immobilized complexes can directly or indirectly remediate soils contaminated with cationic metals. The addition of the cationic metal-tolerant strain *Pseudomonas* sp. NT-2 immobilized on biochar enhanced the stability of Cd(II) and Cu(II) by increasing the fraction of residual Cd(II) and Cu(II) as well as by reducing the fraction of exchangeable and carbonate bound Cd(II) and Cu(II) in soil. The synergistic effects were attributed to surface complexation with OFGs and precipitation on biochar as well as the surface complexation with abundant functional groups (e.g., carboxyl, hydroxyl, amine, phosphate) on microbial cells. The application of phosphate solubilizing bacteria immobilized on biochar significantly enhanced the stabilization of Pb(II) in soil by forming stable precipitates such as $\text{Pb}_5(\text{PO}_4)_3\text{OH}$ and hydroxylpyromorphite ($\text{Pb}_{10}(\text{PO}_4)_6(\text{OH})_2$). Biochar-based microbial immobilization technology as a bio-augmentation method has also shown promising results for indirect remediation of soils contaminated with cationic metals. Inoculation of cationic metal-resistant bacteria or plant growth promoting bacteria immobilized on biochar improved the phytoextraction efficiency of cationic metals by promoting plant growth and metal accumulation in plants (Wu et al., 2021).

Biochar aging in soil may have diverse effects on metal immobilization performance. Biochar contains more OFGs after oxidation by chemical and biological processes in soils, and tends to adsorb greater amounts of metals due to enhanced surface complexation with OFGs. Nevertheless, the loss of inorganic minerals and the blockage of biochar pores are thought to diminish biochar sorption capacities for cationic metals. Overall, the long-term effectiveness of biochar in the stabilization of cationic metals gradually weakened in long-term field trials (Wang et al., 2020).

For metal anions (mainly arsenic (As) and chromium (Cr)), biochar also presents great potential in the remediation of As and Cr contaminated soils. The species of As and Cr are sensitive to soil conditions such as redox potential and soil pH. It has been demonstrated that biochar produced from sewage sludge had a high As(V) immobilization capacity in soil, mainly due to the inner-sphere complexation of As(V) with large contents of Fe_2O_3 in biochar. Immobilization of As(III) in paddy soils can be achieved by the adsorption of As(III) on biochar through electrostatic attraction and surface complexation which oxidizes As(III) to the less mobile As(V). However, pristine biochar also showed conflicting behavior in the immobilization of As in paddy soils. It was found that As mobility increased in soils treated with biochar. This may be attributed to the increase in soil pH and the competition of As with soluble P in biochar (Ahmad et al., 2014). Biochar addition also increased the abundance of genes such as *arrA* and *arsC*, which were associated with the reduction of As(V) to mobile As(III) species. These contrasting outcomes may be related to different properties of biochar and soil (Kumarathilaka et al., 2019). The use of modified biochar is, therefore, effective in immobilizing As in paddy soils and reducing As accumulation in rice grains. It has been reported that biochar loaded with Fe or Mn oxide NPs greatly improved As(III) and As(V) removal kinetics and capacity. Both the redox transformation of As(III) to less mobile As(V) species and the adsorption of As(III) and As(V) on biochar increased after loading with Fe or Mn oxide NPs (Kumarathilaka et al., 2019). Studies have clearly demonstrated that Si application to soil could reduce As accumulation in rice tissues, which is attributed to reduced As uptake and increased Si uptake by Si uptake transporters. Therefore, biochar rich in Si derived from wheat and rice straw can be used to reduce As uptake to rice tissues. The main mechanisms responsible for Cr(VI) sorption on biochar are assumed to be the adsorption of Cr(VI) on the surface of biochar by electrostatic interactions between the Cr(VI) anion and the positively charged functional groups and the subsequent reduction to Cr(III). The Cr(III) was more tightly immobilized on biochar through surface complexation and precipitation. Sugarcane bagasse-derived biochar showed greater immobilization capacity of Cr(VI) in soils due to its lower pH compared with straw-derived biochar. Low-temperature ($<500^{\circ}\text{C}$) biochar was beneficial for the retention of Cr(VI) in soils because of the large contents of OFGs in biochar, which were responsible for the reduction of Cr(VI) to Cr(III). Biochar modification with nZVI effectively facilitated the immobilization of Cr(VI) by enhancing the reduction of Cr(VI) to Cr(III). An immobilization efficiency of Cr(VI) of 100% could be achieved by the addition of 0.8% biochar-nZVI nanocomposites to Cr contaminated soil (Zheng et al., 2021).

Removal of organic pollutants

Biochar has shown strong affinity to many organic pollutants in contaminated soils, thereby reducing their detrimental effects on plants. Adsorption and partitioning are the common mechanisms controlling the bioavailability of organic pollutants in soils. Adsorption is a non-linear and competitive process; whereas partitioning represents the opposite process. Adsorption mechanisms include H-bond, π - π electron donor-acceptor interaction (π - π EDA), hydrophobic effect, pore-filling, and electrostatic attraction.

Sorption mechanisms of hydrophobic organic compounds (HOCs) evolved from partition-dominant at low HTTs to adsorption-dominant at high HTTs due to an increase in hydrophobicity and porosity of biochar. The contributions of H-bond and partitioning decreased; whereas the contributions of hydrophobic effect, π - π EDA, and pore filling increased with increasing HTTs because of the thermal structural transition of biochar. The nature of HOCs also affected the sorption capacities of biochar. It was found that the sorption of polar organic compounds such as phthalates (PAEs) was dominated by the adsorption process (π - π EDA, hydrophobic interactions, pore-filling); whereas the sorption of nonpolar compounds such as PAHs was mainly controlled by partitioning.

The sorption behavior of ionizable organic compounds (IOCs) is different from that of HOCs. The mechanisms responsible for the sorption of IOCs mainly included electrostatic attraction, π - π EDA, and H-bond. The sorption of IOCs depended strongly on soil pH and zero point of charge of biochar (pH_{zpc}). When soil pH < pH_{zpc} of biochar, biochar with net positive charges on its surface is suitable for the sorption of anionic pesticides. On the contrary, biochar is negatively charged at soil pH > pH_{zpc} of biochar and this promotes the sorption of cationic pesticides (Lian and Xing, 2017).

Biochar aging in soils may have diverse influences on the sorption capacities of organic pollutant. Natural aging of biochar produced under low HTTs can introduce more OFGs on the surface of biochar, contributing to π - π EDA and thus a substantial increase in the sorption capacities of organic pollutants. On the other hand, the introduction of OFGs during biochar aging may lead to the formation of water clusters, inhibiting the approach of organic pollutants to the biochar surface. Moreover, the blockage of biochar pores can cause a decrease in physical adsorption of organic contaminants (Wang et al., 2020).

Degradation is another important removal pathway for organic pollutants from soils. In-situ remediation of soils through advanced oxidation processes (AOPs) were highly-effective in the complete removal of organic pollutants from soils. A recent study found that degradation efficiency of bisphenol A in soil of 99% could be achieved by the biochar-activated persulfate (PS) oxidation process. The persistent free radicals (PFRs) such as oxygen-centered and semiquinone-type free radicals in biochar could activate PS to generate the hydroxyl radical ($HO\bullet$) and sulfate radical ($SO_4^{\bullet-}$), which were responsible for the degradation of bisphenol A (Liu et al., 2020). Biochar loaded with spinel ferrite, Fe oxide NPs, and nZVI showed superior performance in activating PS and H_2O_2 to generate $HO\bullet$ and $SO_4^{\bullet-}$ to degrade organic pollutants in soils than pristine biochar. The Fe^{2+} in spinel ferrite- and Fe oxide NPs-supported biochar could serve as an effective activator for PS or H_2O_2 to produce $SO_4^{\bullet-}$ and $HO\bullet$. The nZVI on biochar could activate PS directly to generate $SO_4^{\bullet-}$. The Fe^{2+} or amorphous Fe(II) generated during the activation process could also induce PS to form $SO_4^{\bullet-}$. The contributions of $HO\bullet$ and $SO_4^{\bullet-}$ in the degradation of organic contaminants were different. For example, several studies found that $SO_4^{\bullet-}$ was the predominant reactive oxygen species (ROS) in the degradation of decabromodiphenyl ether in acid and neutral soil conditions. Other studies reported that both $HO\bullet$ and $SO_4^{\bullet-}$ were involved in the degradation of *o*-nitrochlorobenzene, while $SO_4^{\bullet-}$ played a more important role. Understanding of the heterogeneous activation processes and degradation mechanisms contributes to the development of AOPs.

Biodegradation is considered to be an eco-friendly and sustainable method for removing organic contaminants from soils. Biochar addition into soils stimulates the microbial degradation of organic contaminants by increasing the abundance of organic pollutant-degraders. It has been reported that the relative abundance of main PAH-degraders (e.g., *Azospirillum*, *Pseudomonas*, *Magnetospirillum* and *Bacillus*) increased in soils amended with biochar pyrolyzed at high temperatures (500–700 °C) (Ding et al., 2021). It is important to note that biochar aging in soils may favor the biodegradation of organic pollutants. An increase in OFGs after aging would promote biochar as an electron shuttle between colonized organic pollutant-degraders on the external and internal surfaces of biochar and organic pollutants (Wang et al., 2020). More importantly, the immobilization of organic pollutant-degraders with biochar as a carrier has greater potential in the biodegradation of organic pollutants than free microbes. Biochar exhibits well-developed porosity, large surface area, considerable mechanical strength, and tunable surface functionality, which are suitable for the attachment and growth of organic pollutant-degraders. Moreover, immobilized cells have greater resistance and adaptability to disturbance in the soil environment than free cells (Wu et al., 2021). Several studies suggested that petroleum hydrocarbons (PHCs)-degrading bacteria immobilized on biochar were effective in the bioremediation of soils contaminated with PHCs. The immobilized cells played a significant role in shorting the half-lives of PHCs and enhancing their biodegradation in soils. The immobilized cells were also efficient in the long-term degradation of PHCs in soils. The addition of other organic pollutant-degraders such as polyacrylamide-degrading bacteria (*Klebsiella* sp.), PAH-degrading bacteria (*Pseudomonas* sp. SDR4 and *M. alpina* JDR7), tebuconazole-degrading bacteria (*Alcaligenes faecalis* WZ-2), and methylene blue-degrading bacterial consortium immobilized on biochar is an attractive and efficient approach for the remediation of soils contaminated by organic pollutants by combining the advantages of highly efficient sorption of biochar and microbial degradation (Fig. 4) (Wu et al., 2021).

It is noteworthy that microbial degradation of organic pollutants is time-consuming and the reaction time for complete degradation is much longer than AOPs. Selecting suitable fast-growing organic pollutant-degrading bacteria is feasible to enhance the biological activity and efficiency of microbial degradation of immobilized cells (Wu et al., 2021). Additionally, insufficient mineralization of organic contaminants usually generated more recalcitrant or toxic intermediate metabolites, which could be toxic to soil biota (Yu et al., 2020). However, chemical degradation also has its limitations such as high energy input, large operational costs, and secondary pollution through the chemicals added. A high degradation efficiency and economic promise may be achieved by coupling microbial degradation with advanced oxidation processes, which should be further developed and evaluated.

Unintended consequences

Although biochar application may have agronomic and environmental benefits, its potential ecological and human health risks have to be considered. It has been reported that some biochars have several harmful substances including PAHs, polychlorinated

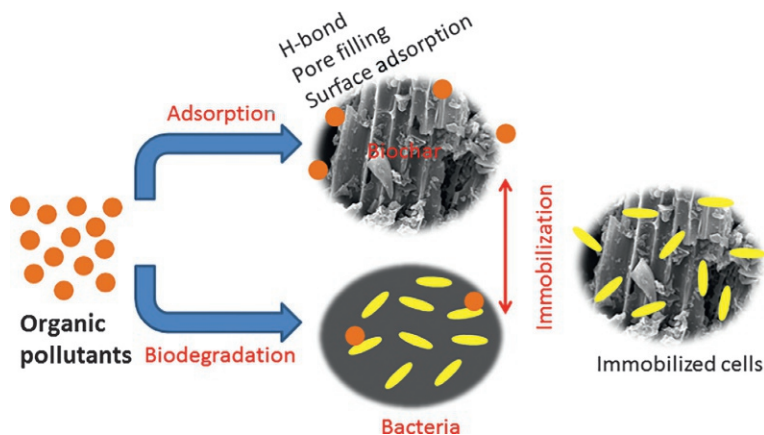


Fig. 4 The removal mechanisms of organic pollutants by cells immobilized on biochar.

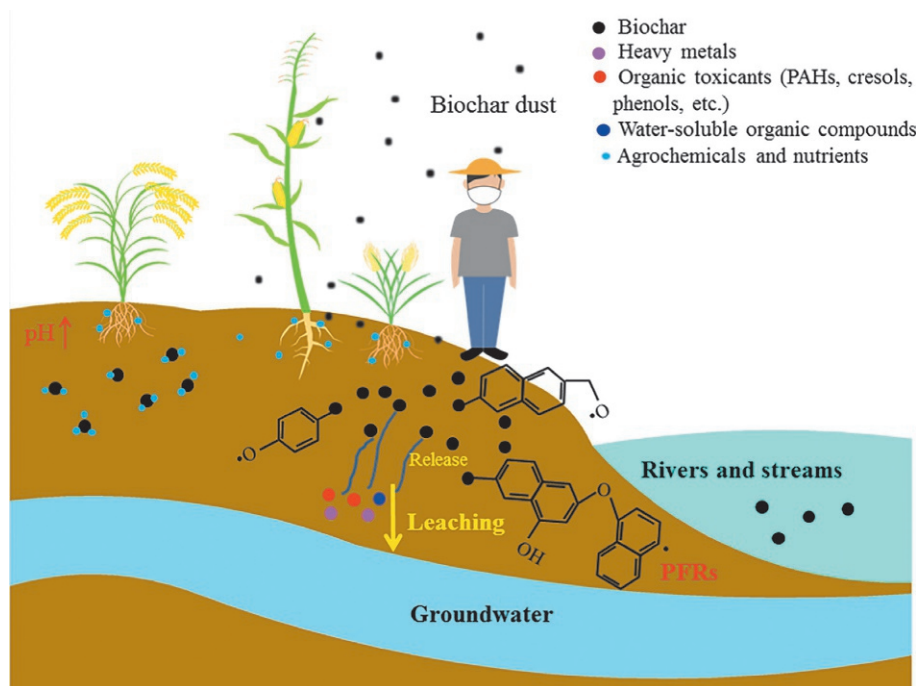


Fig. 5 Potential risks associated with biochar application to soils.

dibenzo-p-dioxins and -furans (PCDD/Fs), volatile organic compounds (VOCs), PFRs, heavy metals, and water-soluble organic compounds (WSOCs) (Fig. 5) (Godlewska et al., 2021). The toxic effects of these harmful substances are related to their bioavailable fraction, which is mainly determined by the feedstock used, HTT and method of biochar production. For example, PAHs are formed during incomplete combustion of biomass. Research suggested that biochar produced by slow pyrolysis contained less PAHs than biochar produced by fast pyrolysis. The bioavailable PAH contents in biochar also vary with feedstocks used. It is suggested that *Miscanthus*-derived biochar has higher contents of PAHs than other biomass-derived biochar. Larger amounts of bioavailable PAHs were often found in medium- temperature biochar (400–600 °C). In general, the bioavailable fraction of PAHs is relatively small for the majority of biochar. The PCDD/Fs form mainly on the biochar surface during thermal decomposition of feedstock containing abundant chlorine, such as food waste. Low HTTs (200–400 °C) and short duration time (<120 min) favor the formation of PCDD/Fs on the biochar surface. It is estimated that PCDD/Fs always exist in insignificant quantities in biochar and pose a rather marginal risk to organisms. However, the repeated application of biochar may increase the ecological risk of PCDD/Fs. The VOCs are formed on the biochar surface or inside biochar pores by thermochemical transformations of biomass. The contents of VOCs in biochar decrease with increasing HTTs. A number of studies have reported that stable and large molecular PFRs are formed in biochar during the carbonization process. It was found that the electron paramagnetic resonance (EPR) signals in lignin-derived biochar were higher than that in cellulose-derived biochar. The EPR signals increased with increasing HTT for the

majority of biochar (Liao et al., 2014). Besides organic compounds, heavy metals (e.g., Cd(II), Cu(II), and Pb(II)) in biochar are of great concerns because of their toxicity. Heavy metal contaminants mainly originate from feedstocks rich in heavy metals, especially sewage sludge, animal manure, and plants grown on heavy metal-contaminated soils. The conversion of biomass with abundant heavy metals to biochar is considered a promising method for safe disposal of the biomass as well as significantly reducing the bioavailability and leaching of heavy metals. The contents of certain heavy metals increased with increasing HTTs, which may be attributed mainly to a 'concentration effect' resulting from the decrease in biochar yield. The bioavailability of heavy metals depends on heavy metal contents in raw biomass as well as the transformation and dissolution of heavy metals in biochar (Godlewska et al., 2021). The release of water-soluble organic compounds (WSOCs) from hydrochar was greater than that from biochar. With increasing HTT, the contents of phenols and organic acids in WSOCs increased (Hao et al., 2018).

Despite ecologically acceptable concentrations of these harmful substances in most biochar, it may pose an ecological risk to soil biota. For example, research reported that some contaminants (e.g., PAHs, cresols, methylated phenols) accumulated in biochar could be toxic to soil microorganisms. The PFRs in biochar might inhibit germination, damage plasma membrane, and retard plant growth. Other studies confirmed that PAHs in biochar can increase mortality and hamper the reproduction of soil invertebrates. Biochar-related dust emissions may have potentially adverse impacts on human health when applied to agricultural soils. The potential ecotoxicity of nano-sized biochar after ball milling was higher than that of pristine biochar due to its nanotoxicity. Except for the ecological risk directly induced by harmful substances in biochar, the changes in soil physicochemical properties after biochar addition may indirectly cause ecotoxicity to soil organisms. Increased soil pH after biochar incorporation may have a toxic effect on the reproduction of earthworms. Biochar's strong adsorption capacity for nutrients and thus their reduced bioavailability may have negative impacts on plant growth. Similarly, the strong water retention capacity of biochar may decrease the accessibility of water to plants (Fig. 5). It should be noted that low rates of biochar application may not have negative effects on soil biota, but application of biochar at excessively large rates is likely to generate unintended consequences.

Conclusions and perspectives

In this chapter, the agricultural and environmental implications of biochar application to soils are discussed. Numerous studies have demonstrated the great potential of biochar in various agricultural and environmental applications including climate change mitigation, soil improvement, the amelioration of salinity and drought stress, and remediation of polluted soils. However, potentially negative implications, and ecological and human health risks related to biochar application should not be overlooked before its large-scale application in agricultural and environmental practices. Further research is needed to enhance the potential agricultural and environmental benefits of biochar may include:

- (i) Considering the variety of feedstock used and pyrolysis conditions, biochar properties are highly variable. Differences in biochar properties affect its functions as a soil amendment as well as its environmental risks. Further research is needed to combine several methods to understand the charring mechanisms more precisely and to predict the resulting biochar properties. This would be useful to produce biochar with properties that would be suitable as a soil amendment and with little risk of releasing harmful substances.
- (ii) Owing to the excellent performance of tailored engineered/modified biochar, it has been widely used to enhance remediation efficiency of contaminants in soils. However, its potential and the unintended ecotoxicological consequences are seldom addressed. The potential toxicities of metal oxide NPs (e.g., ZnO, Fe₃O₄, TiO₂ NPs) to soil organisms have been elucidated in previous studies. The loading of metal oxide NPs on biochar may be ecotoxic to soil biota. Moreover, the leaching of metal ions from biochar-metal oxide nanocomposites may also lead to ecotoxicity of soil biota. It has been reported that the application of biochar-embedded nZVI could inhibit the growth of soil microorganisms. Therefore, it is necessary to monitor and assess the potential risk related to the application of engineered biochar.
- (iii) The long-term effects of biochar on the stabilization and bioavailability of contaminants, herbicides, pesticides, and nutrients in soils should be explored further. The changes in biochar physicochemical properties and soil conditions over time are likely to influence sorption and desorption kinetics of contaminants, thereby affecting the fate of contaminants in soils. For example, the increase in dissolved organic matter (DOM) induced by biochar aging may accelerate the activation of As (III, V) and Cu(II) in soils. Decrease in the long-term stabilization performance of biochar in soil may lead to the release of heavy metals sorbed to biochar. The long-term aging of biochar in soils can pose positive or negative effects on nutrient-use efficiency. Thus, long-term field experiments should be conducted to examine the aging effect on biochar's efficacy as a soil amendment.
- (iv) Although biochar can reduce bioavailability of contaminants and thus reduce their uptake by crops, its application would increase production costs, thus, little public acceptance is expected. To maximize the remediation benefits of biochar, more work is needed to identify the rates of application of biochar that achieve the remediation purpose as well as minimize input costs. Additionally, combining applications of biochar with other remediation technologies such as microbial remediation and cultivating low-metal accumulation crop cultivars could reduce input costs. Furthermore, biochar as a carrier for nutrients, moisture, pesticides and beneficial microorganisms such as N-fixing rhizobium bacteria could be attractive options and promote field application of biochar products.

References

- Ahmad M, Rajapaksha AU, Lim JE, Zhang M, Bolan N, et al. (2014) Biochar as a sorbent for contaminant management in soil and water: A review. *Chemosphere* 99: 19–33.
- Ding Z, Zhang F, Gong H, Sun N, Huang J, et al. (2021) Responses of phenanthrene degradation to the changes in bioavailability and microbial community structure in soils amended with biochars pyrolyzed at low and high temperatures. *Journal of Hazardous Materials* 410: 124584. <https://doi.org/10.1016/j.jhazmat.2020.124584>.
- Godlewska P, Ok YS, and Oleszczuk P (2021) The dark side of black gold: Ecotoxicological aspects of biochar and biochar-amended soils. *Journal of Hazardous Materials* 403: 123833. <https://doi.org/10.1016/j.jhazmat.2020.123833>.
- Hagemann N, Joseph S, Schmidt HP, Kammann CI, Harter J, et al. (2017) Organic coating on biochar explains its nutrient retention and stimulation of soil fertility. *Nature Communications* 8: 1089. <https://doi.org/10.1038/s41467-017-01123-0>.
- Hao S, Zhu X, Liu Y, Qian F, Fang Z, et al. (2018) Production temperature effects on the structure of hydrochar-derived dissolved organic matter and associated toxicity. *Environmental Science & Technology* 52: 7486–7495.
- Kumarathilaka P, Seneweera S, Ok YS, Meharg AA, and Bundschuh J (2019) Mitigation of arsenic accumulation in rice: An agronomical, physico-chemical, and biological approach – A critical review. *Critical Reviews in Environmental Science and Technology* 50: 31–71.
- Lehmann J (2007) A handful of carbon. *Nature* 447: 143–144.
- Leng L, Xiong Q, Yang L, Li H, Zhou Y, et al. (2021) An overview on engineering the surface area and porosity of biochar. *Science of the Total Environment* 763: 144204. <https://doi.org/10.1016/j.scitotenv.2020.144204>.
- Lian F and Xing B (2017) Black Carbon (Biochar) in water/soil environments: Molecular structure, sorption, stability, and potential risk. *Environmental Science & Technology* 51: 13517–13532.
- Liao S, Pan B, Li H, Zhang D, and Xing BS (2014) Detecting free radicals in biochars and determining their ability to inhibit the germination and growth of corn, wheat and rice seedlings. *Environmental Science & Technology* 48: 8581–8587.
- Liu J, Jiang S, Chen D, Dai G, Wei D, et al. (2020) Activation of persulfate with biochar for degradation of bisphenol A in soil. *Chemical Engineering Journal* 381: 122637. <https://doi.org/10.1016/j.cej.2019.122637>.
- Mansoor S, Kour N, Manhas S, Zahid S, Wani OA, et al. (2021) Biochar as a tool for effective management of drought and heavy metal toxicity. *Chemosphere* 271: 129458. <https://doi.org/10.1016/j.chemosphere.2020.129458>.
- Nidheesh PV, Gopinath A, Ranjith N, Praveen Akre A, Sreedharan V, et al. (2021) Potential role of biochar in advanced oxidation processes: A sustainable approach. *Chemical Engineering Journal* 405: 126582. <https://doi.org/10.1016/j.cej.2020.126582>.
- Qian L, Chen B, and Hu D (2013) Effective alleviation of aluminum phytotoxicity by manure-derived biochar. *Environmental Science & Technology* 47: 2737–2745.
- Saifullah, Dahlawi S, Naeem A, Rengel Z, and Naidu R (2018) Biochar application for the remediation of salt-affected soils: Challenges and opportunities. *Science of The Total Environment* 625: 320–335.
- Wang J and Wang S (2019) Preparation, modification and environmental application of biochar: A review. *Journal of Cleaner Production* 227: 1002–1022.
- Wang Y, Xiao X, Xu Y, and Chen B (2019) Environmental effects of silicon within biochar (sichar) and carbon–silicon coupling mechanisms: A critical review. *Environmental Science & Technology* 53: 13570–13582.
- Wang L, O'Connor D, Rinklebe J, Ok YS, Tsang DCW, et al. (2020) Biochar Aging: mechanisms, physicochemical changes, assessment, and implications for field applications. *Environmental Science & Technology* 54: 14797–14814.
- Wu P, Wang Z, Wang H, Bolan NS, and Wang Y (2020) Visualizing the emerging trends of biochar research and applications in 2019: A scientometric analysis and review. *Biochar* 2: 135–150.
- Wu P, Wang Z, Bhatnagar A, Jeyakumar P, Wang H, et al. (2021) Microorganisms-carbonaceous materials immobilized complexes: Synthesis, adaptability and environmental applications. *Journal of Hazardous Materials* 416: 125915. <https://doi.org/10.1016/j.jhazmat.2021.125915>.
- Xiao X, Chen B, Chen Z, Zhu L, and Schnoor JL (2018) Insight into multiple and multilevel structures of biochars and their potential environmental applications: A critical review. *Environmental Science & Technology* 52: 5027–5047.
- Yang Q, Zhou H, Bartocci P, Fantozzi F, Masek O, et al. (2021) Prospective contributions of biomass pyrolysis to China's 2050 carbon reduction and renewable energy goals. *Nature Communications* 12: 1698. <https://doi.org/10.1038/s41467-021-21868-z>.
- Yao Y, Gao B, Chen J, and Yang L (2013) Engineered biochar reclaiming phosphate from aqueous solutions: mechanisms and potential application as a slow-release fertilizer. *Environmental Science & Technology* 47: 8700–8708.
- Yu T, Wang L, Ma F, Wang Y, and Bai S (2020) A bio-functions integration microcosm: Self-immobilized biochar-pellets combined with two strains of bacteria to remove atrazine in water and mechanisms. *Journal of Hazardous Materials* 384: 121326.
- Zheng C, Yang Z, Si M, Zhu F, Yang W, et al. (2021) Application of biochars in the remediation of chromium contamination: Fabrication, mechanisms, and interfering species. *Journal of Hazardous Materials* 407: 124376.
- Zhu H, Yang J, Yao R, Wang X, Xie W, et al. (2020) Interactive effects of soil amendments (biochar and gypsum) and salinity on ammonia volatilization in coastal saline soil. *Catena* 190: 104527.

Soil management for carbon sequestration

Christopher Brock^a, Uwe Franko^b, and Martin Wiesmeier^c, ^aForschungsring, Darmstadt, Germany; ^bHalle, Germany; ^cBavarian State Research Center for Agriculture, Freising, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	425
Carbon sequestration—Potential and constraints	425
Dynamics of organic carbon in soils	425
Turnover of organic matter in soils	425
Management impact on organic matter turnover	426
Carbon inputs	426
Carbon outputs	426
Carbon saturation	427
Carbon sequestration strategies	427
Soil management practices for carbon sequestration	428
Land-use change	428
Cropland soil management	428
Improved crop rotations and crop residue management	428
Organic fertilizers or amendments	428
Biochar	428
Agroforestry	429
Reduced tillage and deep inversion tillage	429
Mineral amendments	429
Grassland soil management	429
Carbon sequestration for the mitigation of climate change	429
Constraints of carbon sequestration for climate change mitigation	429
Carbon supply in a systems perspective	430
Carbon sequestration in soils compared with fossil C emissions	430
Conclusion	430
References	431

Abstract

Carbon management in soils has a great potential both for climate change mitigation and adaptation. Mitigation can be achieved directly by the sequestration of carbon in the soil. Further carbon is usually added in the form of organic matter. As organic matter is a key factor of soil fertility and health, carbon management indirectly contributes to climate change mitigation by fostering ecosystem services that have the potential to reduce the need for external inputs and technical interventions in land management. Moreover, improving ecosystem services enhances the resilience and adaption of land-use in a changing climate.

Key points

- Carbon management in soils has a large potential for both climate change mitigation and adaptation.
- Carbon management in soils is based on the supply of soils with organic matter (and biochar), and or on the control of microbial turnover.
- Carbon management contributes directly to climate change mitigation by stabilizing carbon on mineral surfaces and within soil aggregates and micropores.
- Carbon management indirectly contributes to climate change mitigation by fostering ecosystem services that have the potential to reduce the need for external inputs and technical interventions, which in turn reduces industrial emissions and fuel consumption.
- Ecosystem services related to carbon management further improve the resilience and adaptiveness of land-use systems in a changing climate.
- Carbon sequestration in soils is a dynamic process, where different carbon fractions have different turnover times. Carbon may be stabilized in soils for a long time, but there is no permanent sequestration.
- Carbon sequestration in soils does not discharge other sectors from the responsibility to phase out fossil carbon emissions in the near future.

Introduction

Throughout the history of arable farming it has been acknowledged that organic matter supply to soil is a prerequisite for the development of fertile soils. In Europe, the 1st century Roman agronomist Columella stated that it is necessary to supply arable soils with organic matter, as the export of crop products creates an imbalance between inputs and outputs of organic matter on the field scale (Columella: *De re rustica libri duodecim*). Later, Albrecht Daniel Thaer developed a table to calculate the demand of different crops for organic manure (Thaer: *Grundsätze der rationellen Landwirthschaft*, vol.2). And Justus von Liebig, better known for the mineral nutrient theory and law of the minimum, stressed that organic matter supply is invaluable to maintain soil fertility (Liebig: *Chemische Briefe*). Awareness for the need to supply soils with organic matter was not only a European perspective, but is a characteristic of traditional arable societies throughout the world. A famous example is the formation of Terra Preta soils in the Amazonas by the build-up of organic matter to improve the fertility of pedogenetically poor tropical soils.

The aim of supplying soils with organic matter has always been to improve soil fertility, or -to be concise- soil functions in arable crop production. Carbon sequestration has only recently become an aim of soil organic matter (SOM) management to mitigate climate change.

Carbon sequestration—Potential and constraints

The idea of carbon sequestration is to capture atmospheric carbon and store it in soils. Soils comprise the second-largest carbon pool in the world (after oceans). Carbon in soils is either carbonate in soil minerals, or carbon in organic matter. Of these two, the organic matter pool is dynamic and can be increased (or decreased) by soil management. Carbon sequestration strategies therefore aim to increase organic carbon in soils.

According to Lal et al. (2018), carbon (C) sequestration in soils could achieve a C gain of 333 Pg by the end of this century, corresponding to a decrease of the atmospheric C concentration by 156 ppm. The “4 per 1000” initiative launched in 2015 at the United Nations Climate Change Conference in Paris claims that an annual increase of C levels in soils by 0.4% would offset approximately 30% of actual annual global C emissions for a yet undefined timespan until C saturation in soils has been reached (Minasny et al., 2017). Thus, the potential of soils in mitigating climate change is, in principle, considerable.

It must be considered, though, that both the calculation of Lal et al. (2018) and the “4 per 1000” concept refer to best-case scenarios that would require huge efforts in land management on a global level. Even though the annual rate of C sequestration proposed by the “4 per 1000” initiative may be feasible for many soils from a biophysical point of view, several barriers must be taken into account with regard to the implementation. Effective carbon sequestration in soils depends not only on technical and physical factors, but even more so on economic, social and political factors in human societies. These findings do not counteract the potential of soils to sequester carbon in principle, but they need to be taken into account in the public discussion, and in the development of cross-sector strategies for climate change mitigation.

Soil management for carbon sequestration not only contributes to climate change mitigation by capturing atmospheric carbon, but also supplies soils with organic matter which has an important effect on soil fertility and health. Recommended measures to increase soil carbon by management aim at increasing organic matter inputs, in particular by adapting crop rotations and cropping systems, and also the stability of added carbon in soils. The supply of soils with organic matter is a key factor for soil fertility and soil health, therefore, measures for carbon sequestration are synergistically connected to soil fertility management. Fertile soils, in turn, need fewer external inputs in the production systems, which contributes to decreasing greenhouse gas (GHG) emissions in the industrial sector. Therefore, production and climate change mitigation can be linked more easily in agriculture than in most other sectors.

Dynamics of organic carbon in soils

To utilize carbon management for climate change mitigation, it is necessary to recall the dynamics and functions of organic matter in soils.

The accumulation and depletion of organic matter in soils are basically biological processes. Apart from carbonate in mineral fertilizers and calcareous soils, the carbon in soils is of organic origin in living and dead organisms, and their turnover products. Organic matter in soils is continuously processed, and it is likely that there are no stable products at any stage of the turnover process, even though different materials may have different turnover times (Lehmann and Kleber, 2015). Carbon can be stored only temporarily in soils by incorporation into the living microbial biomass (short-term), or through physical protection from microbial breakdown by the soil structure or on mineral surfaces. Further, organic matter turnover is considerably decelerated under suboptimal turnover conditions, e.g. in wet soils, or, more generally, in subsoils.

Turnover of organic matter in soils

In the early soil organic matter models it was assumed that organic matter can be treated as one homogenous fraction. But over time, the limitations of this concept became more and more clear. Today, it is generally accepted that soil organic matter has to be subdivided into different pools with individual turnover times (Paul, 2016), where the mechanisms for stabilization were identified as (i) protection by incorporation in soil aggregates and (ii) protection on mineral surfaces. According to Marschner et al. (2008),

chemical recalcitrance is important mainly in the initial phase of organic matter turnover, while physical protection is the main mechanism in the long-term stabilization of organic matter in soils. Under suboptimal living conditions for microbes, biological activity and, correspondingly, the turnover of organic matter are strongly reduced in micropores and within aggregates. Therefore, clay-rich soils with tight aggregation and a large proportion of micropores have a greater potential to stabilize and sequester carbon than light soils with poorer aggregation.

Biological activity in soils depends on temperature, moisture and (with aerobic organisms) aeration. Extreme conditions (waterlogging, drought, low and high temperature) reduce microbial activity. Potential carbon sequestration, therefore, is greater under suboptimal conditions due to decreased turnover of organic matter. As a consequence, organic matter in the subsoil has longer turnover times than the same material in the topsoil, even if the material is not chemically recalcitrant and, in principal, accessible. Carbon inputs in subsoil, e.g., via plant roots, therefore play an important role with regard to carbon sequestration.

Essentially, we have to distinguish between the turnover of organic material that has been added to the soil by natural or technological sources, and the decomposition of SOM itself. While it is commonly assumed that soil microbes play the crucial role in this process, it has become increasingly obvious that the meso- and macro-fauna in soils also contribute considerably to organic matter turnover. This is partly due to the fact that faunal activities disintegrate and relocate litter thus creating new surfaces for improved decomposition. On the other hand, there are also changes in turnover intensity induced by bioturbation, especially by the activity of earthworms that improve aggregation with increased protection of organic matter.

Management impact on organic matter turnover

Soil texture, structure, temperature, moisture and aeration not only interact with each other, but also with management measures, in particular organic matter supply and tillage. The supply of soils with organic matter improves soil structure by fostering aggregation in light soils and by increasing the volume of meso- and macro-pores in heavy soils. Tillage disrupts soil aggregates, making occluded organic matter available for soil microbes. Therefore, both organic matter supply and tillage basically improve living conditions for soil microbes and thus increase biological activity. Provided that organic matter is supplied, this observation does not imply a reduction in carbon sequestration. Furthermore, the increase in microbial turnover will usually be overcompensated for by the addition of organic material itself and by growth of the microbial population. In addition, organic matter supply improves growth conditions for plants, leading to more biomass production and related residue inputs.

Tillage may also improve plant growth, which in turn increases biomass input with residues to the soil. Meta studies have shown that reducing tillage intensity has only a small positive effect on the amount of soil organic matter over the whole soil profile (e.g. Luo et al., 2010). Therefore, the proliferation of microbial activity by aggregate disruption should not be overrated.

Another important issue that has gained attention only recently are the stoichiometric implications of fertilizer additions for organic matter dynamics. As organic matter in soils has a comparatively narrow C:N ratio, it is clear that nitrogen is needed to retain organic carbon in the turnover process (Craine et al., 2007). Carbon-rich substrates with a wide C:N ratio have only a limited potential to increase soil organic matter, because a predominant share of the carbon will be lost in microbial turnover through respiration. In many cases, potential C retention during turnover is higher for organic material with a narrow C:N ratio than for organic material with a wide C:N ratio. This is true despite a faster turnover of N-rich material, if N is prevented from leaching by incorporation into the microbial biomass and organic turnover products. On the other hand, the application of N-rich fertilizers without significant coupled C input (synthetic N fertilizer, biogas liquid after slurry separation, dung water, etc.) may cause a priming effect in the turnover of native soil organic matter.

Carbon inputs

The conversion of atmospheric carbon into plant biomass is the only true C input in the soil-atmosphere carbon balance. All other organic matter amendments, such as compost and animal manure, re-distribute plant-derived carbon. They do not add new carbon to soils, but transfer it in space and or time. It is therefore important to calculate the build-up of plant biomass to estimate carbon inputs. A major challenge in this context is to quantify plant residues on the surface (litter, stubble) and below ground (roots, exudates). The amount of plant residues depends on both environmental conditions at a site and physiological properties of the respective plant. Several studies report data on plant residues, but because of the large effort required for data collection, the available data are still not sufficient for reliable site-specific calculations of plant biomass inputs.

Biochemical properties of different materials affect the turnover times of carbon in soils. It has been shown that roots contribute more to soil organic matter build-up than shoots (Kätterer et al., 2017). Further, carbon supply with microbially pre-digested substrates such as compost, is more effective in stabilizing carbon in soils than green plant biomass (Spaccini and Piccolo, 2019). With regard to the carbon balance, however, it has to be considered that original substrate carbon is lost in the composting process.

Carbon outputs

Carbon leaves the soil via microbial respiration as methane or as dissolved organic carbon (DOC). The most important pathway is microbial respiration, and this process occurs in all soils without waterlogging. Respiration depends on the microbial community, the environmental conditions and the properties of organic substrates. Suboptimal conditions (stress) and a wide C:N ratio of substrates result in decreased efficiency of the microorganisms to incorporate carbon, while carbon losses per unit input increase (Manzoni et al., 2012).

Methane is produced in soils with waterlogging when organic material is decomposed under anaerobic conditions. Methane is 38 times more effective as a greenhouse gas than carbon dioxide. The production of methane therefore is an important process with

regard to climate change, but with limited relevance in arable farming systems on non-flooded soils. Methane is also consumed by certain bacteria in soils under aerobic conditions, which counteracts greenhouse gas emissions.

Dissolved organic carbon is a side product of microbial organic matter turnover. It may be leached into water bodies and even reduce freshwater quality. Even though the leaching of DOC is not a C loss to the atmosphere, this process may require more attention in the management of organic matter.

Carbon saturation

Soil organic matter dynamics follow an exponential time course, approaching a theoretical steady state if an input of organic material with a given quality is applied continuously. The process of carbon saturation depends on soil properties that are related to the physical protection of organic matter, and on environmental conditions that regulate biological activity and turnover (Craig et al., 2021). Carbon saturation is usually calculated as the difference between actual and maximum mineral-associated organic carbon stocks, where the fine fraction of soil texture (clay, or clay + fine silt) is used as a basis. Carbon saturation limits the potential for further sequestration of carbon at a site. Soil with a low carbon saturation may sequester more additional carbon than soil that has already reached a high stage of saturation.

The approach has proven applicability, even though it has been criticized because this calculation does not capture less stable soil carbon pools that may contribute significantly to total carbon storage (Barré et al., 2017). In fact, soils with a low texture-related potential to stabilize carbon (sandy soils) may hold considerable amounts of carbon under appropriate management (Roß et al., 2022). In this situation, carbon is not stored by protection from turnover. Rather, it is 'stored' in turnover by balancing supply and decomposition at a high level. The unprotected carbon pool must be taken into account in carbon sequestration strategies, calling for a more dynamic perspective (Fig. 1). It is necessary to consider that a large supply of organic carbon is usually associated with a considerable input of nitrogen, which can cause environmental damage. On the other hand, the supply of soils with organic matter improves the growing conditions for the plants, which should facilitate N uptake and export. Increasing the unprotected carbon pool, therefore, is a very valuable strategy for the ecological intensification of crop production, but requires thorough monitoring and management of nutrient flows.

There are no reliable global estimates of carbon saturation in managed soils. However, regional calculations (e.g. Wiesmeier et al., 2014) showed that arable soils had a lower carbon saturation than grassland and forest soils. Correspondingly, arable soils had a large calculated potential to store additional carbon, while forest soils were almost saturated. It should be noted, however, that the calculations were performed according to the texture-related approach described above, where potential 'dynamic' carbon storage in less stable fractions is not accounted for.

Carbon sequestration strategies

Carbon sequestration in soils can be achieved either by reducing the turnover of organic matter or by increasing organic inputs. Reducing turnover requires techniques that decrease the accessibility of organic matter to soil microorganisms, or the application of material with reduced degradability (i.e. biochar). The advantage of this strategy is that C storage is quite permanent if conditions remain unchanged. On the other hand, the strategy neglects the importance of organic matter supply to foster biological processes and soil health.

Soil health, a term referring to the ability of soils to function as a living system with a high degree of self-regulation, is largely dependent on soil biology (Lehman et al., 2015). As organic matter plays a crucial role in the soil food web, its decomposition and turnover are of vital importance for maintaining soil health. In particular, turnover provides microorganisms with energy and nutrients, contributes to the biological regulation of pathogens as well as to the decomposition of pollutants, and releases

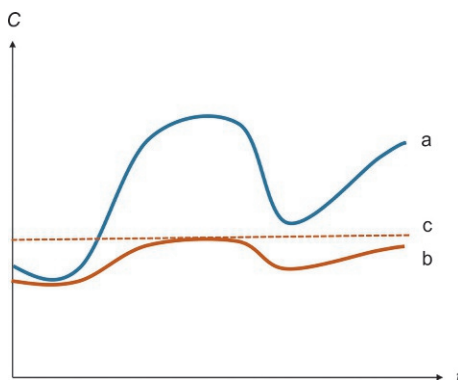


Fig. 1 Dynamics of the unprotected (a) and protected (b) carbon pools in the soil. Carbon changes in both pools over time, but the dynamic is much stronger in the unprotected pool. Carbon sequestration strategies need to consider both pools, calling for a dynamic perception of carbon sequestration. The dotted line (c) denotes the physical maximum of carbon protection in the specific soil.

plant-available nutrients. Moreover, soil health has the potential to reduce the need for external inputs, such as synthetic fertilizers and pesticides (Lehman et al., 2015), which in turn reduce greenhouse gas emissions in the industrial sector.

In contrast to decelerating turnover, increasing organic matter inputs is a synergistic solution for carbon sequestration and the management of soil health. Techniques in this strategy include diverse crop rotations, agroforestry, cover cropping, and organic fertilization. It must be considered, however, that this strategy builds up a very dynamic C pool whose maintenance requires continual effort. Further, increasing organic inputs may even stimulate a priming effect on organic matter turnover, or induce emissions of N₂O (nitrous oxide) and CH₄ (methane), which severely impair the positive effect in climate change mitigation.

Soil management practices for carbon sequestration

Land management has a large impact on carbon dynamics. In this context, the form of land use (cropland, grassland, forest, etc.) is more important than the management within land-use types. The conservation or establishment of land-use types with large carbon storage is therefore of crucial interest for maintaining or improving C storage in soils. With cropland soils, there are various options to improve C levels. Improved crop rotations, residue management and agroforestry mainly increase C inputs, while organic fertilization and biochar application supply carbon in more stable forms that in addition improve soil fertility and health. Further, reduced tillage can affect turnover conditions, and mineral amendments may be used to stabilize carbon better in soils. Carbon sequestration in grassland soils can be improved by adapted grazing, cutting and fertilization strategies.

Land-use change

The conversion of intensively managed cropland to more extensively managed land-use types such as grasslands or forests results in a long-term increase in SOC in most cases (Poeplau et al., 2011). In particular, the establishment of permanent grassland on former cropland is a promising strategy to sequester C, although potential GHG emissions related to grassland use have to be accounted for. A change in land use from cropland to forest normally also results in C sequestration, but the intensity of the SOC increase is strongly related to the tree species. Afforestation with coniferous species often results in a shift from stable to labile SOC that is prone to faster turnover and degradation. For the conversion of grassland to forest there is no consistent effect on SOC stocks, although in temperate regions SOC tended to decrease (Poeplau et al., 2011). The build-up of SOC induced by the conversion of cropland to grassland or forest has often been related to increased C inputs into the soil. However, in temperate regions, recent research has shown that the C input into cropland and grassland soils is similar, but that the quality of organic matter (more root-derived C inputs in grassland soils) rather than quantity controls the differences in SOC stocks between land-use types (Jacobs et al., 2020).

Cropland soil management

Improved crop rotations and crop residue management

Increasing the diversity of cropping systems has been shown to be a valuable tool for the build-up of SOC (McDaniel et al., 2014). In particular, the integration of legumes, ley grassland, cover crops, undersown crops, deep-rooting crops and perennial crops into cropping systems often leads to sequestration of C, mainly by increasing root-derived C inputs. An increased retention of crop residues is another strategy to enhance the C input into agricultural soils and their SOC stocks. In particular, the retention of aboveground harvest residues such as stubble, beet leaves and straw or their return via organic fertilizers (slurry, manure or digestate) is an efficient way to increase C inputs and thus SOC stocks (Bolinder et al., 2020). However, it should be mentioned that the use of harvest residues, e.g. for biogas production or utilization of the straw, might be more advantageous in some cases in terms of climate protection than their retention on agricultural land to increase SOC stocks.

Organic fertilizers or amendments

The return of exported harvest residues to agricultural land in the form of organic fertilizers such as manure, slurry, digestate and compost is a central component of a balanced OM supply to soils and decisive for sustainable cultivation with regard to nutrient recycling (Gross and Glaser, 2021). However, an increase of SOC based on external organic fertilizers does not represent C sequestration, as organic matter is merely relocated and locally concentrated. Such a relocation reduces the SOM reproduction potential of the areas from which the external organic matter originates. Nevertheless, organic fertilizers do contribute to C sequestration indirectly when crop yields and thus C inputs increase due to improved soil fertility.

Biochar

Amending agricultural soils with thermally converted organic matter (biochar) contributes to C sequestration due to its high stability against microbial degradation (Lehmann et al., 2006). In general, a pyrolysis temperature >450 °C is needed to produce biochar that has high overall stability in order to decrease its turnover time in soils and to contribute to C sequestration. However, the effects of the spatial redistribution of organic matter mentioned above in the context of organic fertilizers have to be taken into account. Benefits in terms of C sequestration, improved soil fertility and yields by biochar have so far mainly been demonstrated in tropical and subtropical soils and to a much smaller extent in temperate soils (Jeffery et al., 2017).

Agroforestry

Agroforestry systems integrate woody species in agricultural land and contribute to C sequestration, mainly due to increased above- and below-ground C inputs. Evidence for C sequestration in soils of agroforestry systems was mainly found in tropical and subtropical regions, but also in temperate environments (Chatterjee et al., 2018; De Stefano and Jacobson, 2018). The incorporation of trees in agricultural land may result in additional C inputs by tree roots and rhizodeposition, particularly in subsoils, but also by litterfall, the return of pruning residues, understory vegetation and increased yields on adjacent agricultural land (De Stefano and Jacobson, 2018). Moreover, reduced soil disturbance and erosion, and retarded decomposition of recalcitrant tree litter may contribute to observed SOC increases in agroforestry systems.

Reduced tillage and deep inversion tillage

It is often assumed that reduced tillage or no-till systems may contribute to C sequestration due to improved soil aggregation and thus physical protection of organic matter, while conventional tillage leads to a disruption of (macro)aggregates and increased mineralization of organic matter. However, meta studies and systematic reviews that also considered subsoil layers have provided evidence that reduced tillage and no-till systems merely induce a vertical redistribution of C and do not contribute to a net C sequestration (Luo et al., 2010). In reduced tillage/no-till systems SOC stocks are often increased in the upmost topsoil (0–10 cm), but are lower in the lower topsoil compared to conventional tillage, probably because of reduced incorporation of organic matter. In contrast, deep inversion tillage in soils with compacted subsoil layers (so-called hardpans caused by natural compaction processes, or indurated layers caused by ploughing) or for pasture renewal may contribute to C sequestration as the burial of C-rich topsoil material results in reduced decomposition of OM, while fresh OC inputs in exposed subsoils depleted in C may be effectively stabilized (Alcántara et al., 2016).

Mineral amendments

The application of mineral amendments such as clay-rich materials to agricultural soils is a promising strategy to increase their SOC storage potential. In addition to the incorporation of external clay-rich substrate in topsoils, subsoil material is often mixed into sandy topsoils to improve their chemical and physical properties, reduce soil erosion and nutrient leaching and increase SOC stocks (Schapel et al., 2018). Moreover, the incorporation of crushed silicate rocks such as basalt is discussed as an option to sequester inorganic C in the form of stable carbonates. Silicate by-products from industrial processes such as steel slag could also be used. However, evidence for organic or inorganic C sequestration by mineral amendments is still poor and additional GHG emissions related to mining or excavation, grinding, transport and the application of minerals have to be taken into account (Chenu et al., 2019).

Grassland soil management

Several improved grassland management practices are discussed that increase SOC stocks in managed grassland soils. In grazed grassland ecosystems, improved grazing management such as optimized grazing intensity or holistically planned grazing may contribute to C sequestration (McSherry and Ritchie, 2013). In general, reduced grazing intensity or spatial and temporal grazing exclusion normally results in increased SOC stocks because of increased above- and below-ground C inputs, particularly in semi-arid environments. In temperate regions, optimized grassland management intensity with regard to the frequency of cutting and fertilization may increase SOC stocks (Conant et al., 2017). Managing plant diversity in temperate grasslands could also be a strategy to foster C sequestration due to increased rhizosphere C input (Yang et al., 2019). Moreover, sward renewal and integration of deep rooting and/or leguminous species could result in increased belowground C inputs and increased SOC storage.

Carbon sequestration for the mitigation of climate change

There is increasing interest in carbon sequestration in soils as a strategy for the offset of GHG emissions to mitigate climate change. Until now, carbon sequestration in soils has not been included in the United Nations Clean Development Mechanism (<https://cdm.unfccc.int/>), or in the European Emission Trading Mechanism (Directive (EU) 2018/410). However, more private sector initiatives for voluntary climate impact compensation have discovered carbon sequestration in soils as a promising business model to counteract climate change and generate income for farmers by selling carbon credits to companies and other clients (Wiesmeier et al., 2020). In principal, this development has to be appreciated, as it raises awareness about the great importance of soils and soil services. Further, measures for carbon sequestration in managed land can have multiple benefits in addition to carbon storage. At the same time, constraints and limitations of carbon sequestration have to be taken into account to make use of the potential benefit and avoid trade-offs.

Constraints of carbon sequestration for climate change mitigation

Concerning potential carbon sequestration in soils, approaches such as the “4 per 1000” initiative have been criticized because of their exaggerated estimates and lack of consideration of the feasibility of the measures. The “4 per 1000” initiative accounts for the

physical limits of carbon accumulation in soils, but an increase in carbon stocks of 0.4% per annum is hardly achievable and realistic as a global average. Rumpel et al. (2020) summarized the barriers and constraints of the initiative. The authors not only identify biophysical, but also socio-economic and political issues that need to be addressed. With regard to biophysical barriers the authors recall that the potential increase of carbon contents in soils depends on the stage of carbon saturation, and on environmental conditions. At the interface of environmental and management factors, the availability of organic matter to build up SOM must be taken into account. Further, cultural traditions and or socio-economic conditions and political issues can make it difficult for farmers to adopt new techniques and revise their farming systems. In a regional case study from Germany, Wiesmeier et al. (2020) calculated a potential mean annual carbon stock increase of 0.1‰ as a realistic scenario. This is in accord with Schlesinger and Amundson (2018), who assume that carbon sequestration in soils may decrease net carbon emissions to the atmosphere by roughly 5%. It must be noted, though, that this number does not comprise indirect beneficial effects of organic matter build-up.

Carbon supply in a systems perspective

Carbon sequestration must be contemplated in a systems perspective. In fact, increasing soil carbon has many more advantages apart from storing a certain amount of C that benefit climate change mitigation: As stated in the introduction to this chapter, carbon supply is usually related to the build-up of soil fertility, which in turn allows for a reduction of external inputs, such as synthetic fertilizers and pesticides. Currently, the indirect climate change mitigation potential of carbon sequestration has not been quantified on a global scale because the effects usually result from complex system changes, and not from carbon build-up alone. However, there is considerable evidence for ecological services provided by the organic matter supply to soils. According to Bommarco et al. (2013), increasing SOM amounts and diversifying crop rotations are the most effective measures to promote the support of ecological services in agroecosystems. Based on a large European survey, Garrat et al. (2018) showed that increased SOC in soil can partially replace nitrogen fertilizer and reduce pest pressure in cereals under conventional management, even without further agroecological optimization of farming systems. Luo et al. (2018) found that organic amendments significantly improved crop yields and soil functions in a global meta-analysis of field experiments. Against this background, Baveye et al. (2020) argue that the improvement of soil functions by organic matter supply is of much greater relevance for climate change mitigation than carbon sequestration itself.

It is definitely necessary to look at management options for carbon sequestration in a different way. Wiesmeier et al. (2020) pointed out that measures for carbon sequestration may have so-called leakage effects on other sites, meaning that improving carbon sequestration by increasing carbon inputs may withdraw carbon from other sites. This is particularly relevant for all measures that reduce productivity (such as set-aside), as they may induce land-use changes or other forms of intensification at another site to compensate for the decrease in product output. Further, it has been suggested that organic matter supply may lead to reduced N efficiency use and increased losses because of problems in synchronizing N release and uptake in agricultural soils. In addition, nitrogen release from organic amendments might even increase GHG emissions, and N₂O in particular. Undoubtedly, there is also a need for improvement and optimization in the development and implementation of agroecological management strategies and techniques. The potential benefit of organic matter supply to soils for improved ecological and economic performance of agricultural systems is much greater than the potential disadvantage.

Carbon sequestration in soils compared with fossil C emissions

The most important threat to climate change mitigation by carbon sequestration in soils, however, is that this option may distract from the need to phase out C emissions from fossil sources. In the scientific literature, it is well-recognized that carbon sequestration directly and indirectly contributes to climate change mitigation, but that there is definitely no alternative to the phasing out of fossil C emissions in the near future (Seddon et al., 2020). As described in this chapter, carbon is dynamic in soils, and the potential of soils to store carbon further is not infinite. Therefore, the attempt to offset fossil C emissions by carbon sequestration in soils is a risky strategy without a long-term perspective. Permanent storage of C, which would be the prerequisite for the true offset of fossil C emissions, cannot be guaranteed. The balancing of fossil C emissions by carbon sequestration in soils may even have a detrimental impact on climate change, especially if carbon is lost again from soils over time (Thamo and Panell, 2016). Furthermore, the finite potential of carbon accumulation in soils makes carbon sequestration in soils a transitional solution on the way to a post-(fossil) carbon economy.

Conclusion

Building up soil carbon contents through agricultural management is much more than merely removing carbon from the atmosphere. On the one hand, agricultural soils worldwide have a great potential to sequester atmospheric carbon and compensate for a, yet limited, amount of GHG emissions. On the other hand, organic matter in soils is a key factor of soil fertility and health, and therefore plays a crucial role in the ecological intensification of agricultural production. Building up soil organic matter is thus both a mitigation and adaptation strategy in the context of climate change.

There is no permanent storage of organic matter in soils. Soil carbon is stored reversibly in fractions with different stability and turnover times due to protection from microbial breakdown and chemical recalcitrance. Long-term stabilization of organic carbon

in managed soils depends on the sorptive capacity of the mineral fraction and on compactness of the soil structure. In agricultural soils, the largest proportion of the carbon is incorporated in organic matter that is involved in biological turnover. Unlike other carbon capture and storage technologies, carbon sequestration in agricultural soils must be understood as a dynamic process, where carbon is not stored in a permanent pool, but in the cycling of organic matter.

Organic matter supply to soils is both a key driver of carbon build-up and of soil fertility, therefore, farming *with* carbon should replace the idea of farming *for* carbon. Carbon sequestration in managed soils is both a prerequisite and an effect of ecological intensification in agricultural production. It therefore has a great potential to counteract climate change, but it does not discharge other sectors from their responsibility to phase out fossil carbon emissions in the near future.

References

- Alcántara V, Don A, Well R, and Nieder R (2016) Deep ploughing increases agricultural soil organic matter stocks. *Global Change Biology* 22(8): 2939–2956.
- Barré P, Angers DA, Basile-Doelsch I, Bispo A, Cécillon L, Chenu C, Chevallier T, Derrien D, Eglin TK, and Pellerin S (2017) Ideas and perspectives: Can we use the soil carbon saturation deficit to quantitatively assess the soil carbon storage potential, or should we explore other strategies? *Biogeosciences*. <https://doi.org/10.5194/bg-2017-395>. Preprint.
- Baveye PC, Schnee LS, Boivin P, Laba M, and Radulovich R (2020) Soil organic matter research and climate change: Merely restoring carbon versus restoring soil functions. *Frontiers in Environmental Science* 8: 579904.
- Bolinder MA, Crotty F, Elsen A, Frac M, Kismányoky T, Lipiec J, Tits M, Tóth Z, and Kätterer T (2020) The effect of crop residues, cover crops, manures and nitrogen fertilization on soil organic carbon changes in agroecosystems: A synthesis of reviews. *Mitigation and Adaptation Strategies for Global Change* 25: 929–952.
- Bommarco R, Kleijn D, and Potts SG (2013) Ecological intensification: Harnessing ecosystem services for food security. *Trends in Ecology & Evolution* 28: 4.
- Chatterjee N, Nair PKR, Chakraborty S, and Nair VD (2018) Changes in soil carbon stocks across the Forest-Agroforest-Agriculture/Pasture continuum in various agroecological regions: A meta-analysis. *Agriculture, Ecosystems and Environment* 266: 55–67.
- Chenu C, Angers DA, Barré P, Derrien D, Arrouays D, and Balesdent J (2019) Increasing organic stocks in agricultural soils: Knowledge gaps and potential innovations. *Soil and Tillage Research* 188: 41–52.
- Conant RT, Cerri CEP, Osborne BB, and Paustian K (2017) Grassland management impacts on soil carbon stocks: A new synthesis. *Ecological Applications* 27(2): 662–668.
- Craig ME, Mayes MA, Sulman BN, and Walker AP (2021) Biological mechanisms may contribute to soil carbon saturation patterns. *Global Change Biology* 27(12): 2633–2644.
- Craine JM, Morrow C, and Fierer N (2007) Microbial nitrogen limitation increases decomposition. *Ecology* 88(8): 2105–2113.
- De Stefano A and Jacobson MG (2018) Soil carbon sequestration in agroforestry systems: A meta-analysis. *Agroforestry Systems* 92(2): 285–299.
- Garrat MPD, Bommarco R, Kleijn D, Martin E, Mortimer SR, Redlich S, Senapati D, Steffan-Dewenter I, Świątek S, Takács V, van Gils S, van der Putten WH, and Potts SG (2018) Enhancing soil organic matter as a route to the ecological intensification of European Arable Systems. *Ecosystems* 21: 1404–1415.
- Gross A and Glaser B (2021) Meta-analysis on how manure application changes soil organic carbon storage. *Scientific Reports* 11: 5516.
- Jacobs A, Poeplau C, Weiser C, Fahrion-Nitschke A, and Don A (2020) Exports and inputs of organic carbon on agricultural soils in Germany. *Nutrient Cycling in Agroecosystems* 118(3): 249–271.
- Jeffery S, Abalos D, Prodana M, Bastos AC, van Groenigen JW, Hungate BA, and Verheijen F (2017) Biochar boosts tropical but not temperate crop yields. *Environmental Research Letters* 12: 6.
- Kätterer T, Bolinder MA, Andrén O, Kirchmann H, and Menichetti L (2017) Roots contribute more to refractory soil organic matter than above-ground crop residues, as revealed by a long-term field experiment. *Agriculture, Ecosystems and Environment* 141: 184–192.
- Lal R, Smith P, Jungkunst HF, Mitsch WJ, Lehmann J, Nair PKR, McBratney AB, de Moraes Sá JC, Schneider J, Zinn YL, Skorupa ALA, Zhang H-L, Minasny B, Srinivasrao C, and Ravindranath NH (2018) The carbon sequestration potential of terrestrial ecosystems. *Journal of Soil and Water Conservation* 73: 6.
- Lehman RM, Cambardella CA, Stott DE, Acosta-Martinez V, Manter DK, Buyer JS, Maul JE, Smith JL, and Collins HP (2015) Understanding and enhancing soil biological health: The solution for reversing soil degradation. *Sustainability* 7: 988–1027.
- Lehmann J and Kleber M (2015) The contentious nature of soil organic matter. *Nature* 528: 60–68.
- Lehmann J, Gaunt J, and Rondon M (2006) Bio-char sequestration in terrestrial ecosystems—A review. *Mitigation and Adaptation Strategies for Global Change* 11(2): 403–427.
- Luo Z, Wang E, and Sun OJ (2010) Can no-tillage stimulate carbon sequestration in agricultural soils? A meta-analysis of paired experiments. *Agriculture, Ecosystems and Environment* 139(1): 224–231.
- Luo G, Lia L, Friman V-P, Guo J, Guo S, Shena Q, and Ling N (2018) Organic amendments increase crop yields by improving microbe-mediated soil functioning of agroecosystems: A meta-analysis. *Soil Biology and Biochemistry* 124: 105–115.
- Manzoni S, Taylor P, Richter A, Porporato A, and Ågren G (2012) Environmental and stoichiometric controls on microbial carbon-use efficiency in soils. *New Phytologist* 196: 79–91.
- Marschner B, Brodowski S, Dreves A, Gleixner G, Gude A, Grootes PM, Hamer U, Heim A, Jandl G, Ji R, Kaiser K, Kalbitz K, Kramer C, Leinweber P, Rethemeyer J, Schäffer A, Schmidt MWI, Schwark L, and Wiesenberger GLB (2008) How relevant is recalcitrance for the stabilization of organic matter in soils? *Journal of Plant Nutrition and Soil Science* 171: 91–110.
- McDaniel MD, Grandy AS, Tiemann LK, and Weintraub MN (2014) Crop rotation complexity regulates the decomposition of high and low quality residues. *Soil Biology and Biochemistry* 78: 243–254.
- McSherry ME and Ritchie ME (2013) Effects of grazing on grassland soil carbon: A global review. *Global Change Biology* 19(5): 1347–1357.
- Minasny B, Malone BP, McBratney AB, Angers DA, Arrouays D, Chambers A, Chaplot V, Chen ZS, Cheng K, Das BS, Field DJ, Gimona A, Hedley CB, Hong SY, Mandal B, Marchant BP, Martin M, McConkey BG, Mulder VL, O'Rourke S, Richer-de-Forges AC, Odeh I, Padarian J, Paustian K, Pan G, Poggio L, Savin I, Stolbovoy V, Stockmann U, Sulaeman Y, Tsui C-C, Vågen T-G, van Wesemael B, and Winowiecki L (2017) Soil carbon 4 per mille. *Geoderma* 292: 59–86.
- Paul EA (2016) The nature and dynamics of soil organic matter: Plant inputs, microbial transformations, and organic matter stabilization. *Soil Biology and Biochemistry* 98: 109–126.
- Poeplau C, Don A, Vesterdal L, Leifeld J, Van Wesemael BAS, Schumacher J, and Gensior A (2011) Temporal dynamics of soil organic carbon after land-use change in the temperate zone—Carbon response functions as a model approach. *Global Change Biology* 17(7): 2415–2427.
- Roß C-L, Baumecker M, Ellmer F, and Kautz T (2022) Organic manure increases carbon sequestration far beyond the “4 per 1000 Initiative” goal on a sandy soil in the throw long-term field experiment DIV.2. *Agriculture* 12: 170.
- Rumpel C, Amiraslani F, Chenu C, García Cardenas M, Kaonga M, Koutika L-S, Ladha J, Madari B, Shirato Y, Smith P, Soudi B, Soussana J-F, Whitehead D, and Wollenberg E (2020) The 4p1000 initiative: Opportunities, limitations and challenges for implementing soil organic carbon sequestration as a sustainable development strategy. *Ambio* 49: 350–360.
- Schapel A, Marschner P, and Churchman J (2018) Clay amount and distribution influence organic carbon content in sand with subsoil clay addition. *Soil and Tillage Research* 184: 253–260.
- Schlesinger WH and Amundson R (2018) Managing for soil carbon sequestration: Let's get realistic. *Global Change Biology* 25: 386–389.
- Seddon N, Smith A, Smith P, Key I, Chausson A, Girardin C, House J, Srivastava S, and Turner B (2020) Getting the message right on nature-based solutions to climate change. *Global Change Biology* 27: 1518–1546.
- Spaccini R and Piccolo A (2019) Amendments with humified compost effectively sequester organic carbon in agricultural soils. *Land Degradation & Development* 31(10): 1206–1216.

- Thamo T and Panell DJ (2016) Challenges in developing effective policy for soil carbon sequestration: Perspectives on additionality, leakage, and permanence. *Climate Policy* 16: 973–992.
- Wiesmeier M, Hübner R, Spörlein P, Geuß U, Hangen E, Reischl A, Schilling B, von Lützow M, and Kögel-Knabner I (2014) Carbon sequestration potential of soils in southeast Germany derived from stable soil organic carbon saturation. *Global Change Biology* 20(2): 653–665.
- Wiesmeier M, Mayer S, Burmeister J, Hübner R, and Kögel-Knabner I (2020) Feasibility of the 4 per 1000 initiative in Bavaria: A reality check of agricultural soil management and carbon sequestration scenarios. *Geoderma* 369: 114333.
- Yang Y, Tilman D, Furey G, and Lehman C (2019) Soil carbon sequestration accelerated by restoration of grassland biodiversity. *Nature Communications* 10(1): 718.

Soil tillage

Mathieu Lamandé^{a,b}, Lars Juhl Munkholm^a, and Trond Børresen^b, ^aDepartment of Agroecology, AU Foulum, Aarhus University, Tjele, Denmark; ^bFaculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway

© 2023 Elsevier Ltd. All rights reserved.

This is an update of MR Carter, E McKyes, Cultivation and Tillage, Editor(s): D Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 356–361, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00514-2>; RL Raper, Subsoiling, Editor(s): D Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 69–76, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00267-8>

Introduction	433
Tillage systems	433
Why tillage?	436
Why NOT tillage?	436
Soil fragmentation in tillage	437
Water range for tillage	438
Outlook	438
References	438

Abstract

Tillage is the mechanical manipulation of the soil to modify or ameliorate its physical condition with regard to plant growth for human and animal food production. Tillage operations include incorporation of fertilizers or crop amendments, seedbed preparation, seeding, and post-emergent cultivation for weed control. The requirements for tillage depend on the crop, the soil type and the climate conditions. Inadequacy of the tillage intensity with these conditions locally might lead to severe damages of the environment and the crop production systems.

Key points

- Tillage is the mechanical manipulation of the soil needed for crop production.
- Tillage operations are combined into various tillage systems.
- Requirement for tillage must be carefully investigated to ensure a sustainable crop production.

Introduction

Tillage is an activity of preparing the land for growing crops by mechanical agitation of soil. Soil movement can also occur due to natural (e.g., freeze-thaw processes) and biotic (e.g., earthworm movement) processes; however, tillage generally refers to the direct intervention of a soil manager, who uses tillage tools to modify, manipulate, or ameliorate the soil physical condition. This activity can be human-powered, draft animal powered or carried out using powered machinery. In a crop-production system, tillage is employed to facilitate crop establishment, modify soil structure, incorporate fertilizers and soil amendments (e.g., lime and manure), control plant residues and weeds, and alleviate both climatic and soil constraints to agricultural production. Over time, various tillage tools have been developed to perform specific soil-moving operations. A combination of tillage operations and their timing results in the development of a tillage system to provide specific functions in given situations. Tillage is a fundamental part of cropland farming. It has an instant and strong effect on soil physical properties and soil biology, and soil physical properties affect the outcome of soil tillage (Fig. 1).

Tillage systems

Traditionally, most tillage operations consist of primary tillage followed by secondary tillage (Fig. 2). This traditional combination is referred to as conventional tillage, although this term is also applied to any tillage system that has been adopted for a period. Primary tillage usually consists of relatively deep (15–25 cm) tillage, using a moldboard or chisel plow. In contrast, secondary tillage is relatively shallow (5–10 cm) and used to pulverize and consolidate the soil to form a suitable seedbed. Both primary and secondary tillage are usually conducted uniformly across a field. The last few decades of the 20th century have seen concerted efforts to reduce the level of energy used for plant production, and therefore the amount and degree of tillage by adoption of reduced, minimum, or shallow tillage. Reduced tillage allows a reduction in depth, degree, and frequency of tillage, while minimum tillage refers to the minimum soil manipulation necessary for crop production or other soil tillage requirement. Shallow tillage involves restricting tillage to a shallow soil depth (less than 15 cm). In no-tillage systems, crop seeds are planted directly into the stubble and

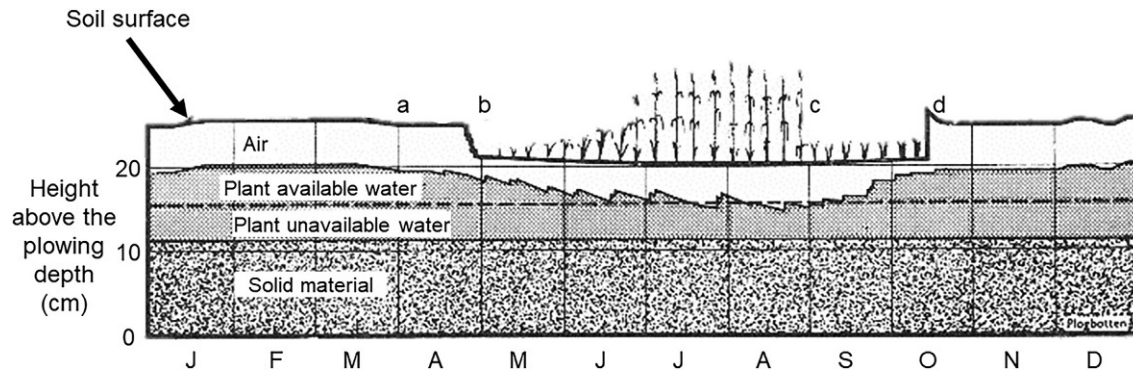


Fig. 1 Tillage has a strong and instant effect on soil physical properties. This diagram represents the dynamics of the air-filled and water-filled volumes (both available and unavailable water for plants) in a cultivated topsoil during a year: (a) autumn plowed topsoil before the start of cultivation in spring, (b) after the seedbed preparation and the seeding in spring, (c) the lowest height of soil surface above the plowing depth after the harvest, and (d) immediately after the autumn plowing. After Andersson S and Håkansson I (1966) Structure dynamics in the top soil. A field study. In: *Swedish with English summary. Grundförbättring 1966: 3*. Swedish University of Agricultural Sciences: Uppsala, Sweden.

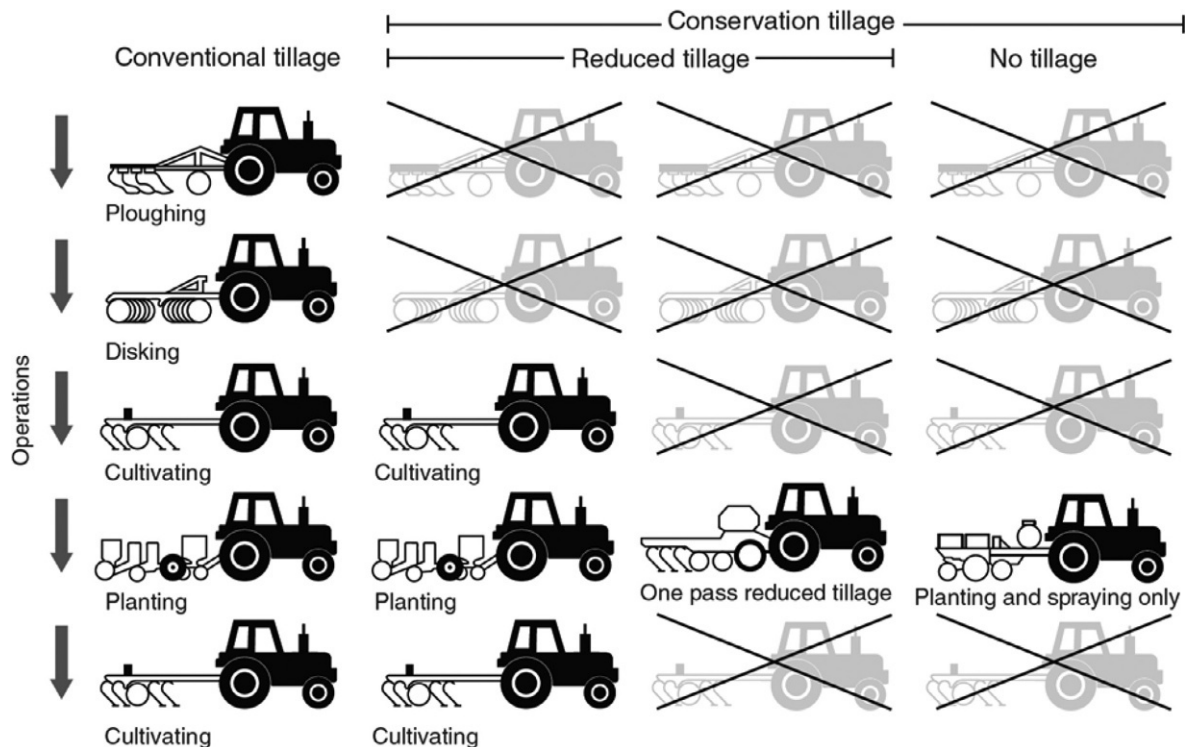


Fig. 2 Tillage systems with various tillage intensity. Ploughing and diskking are primary tillage operations, while cultivating and planting are secondary tillage operations. After Hallett PD and Bengough AG (2013) Managing the soil physical environment for plants. In Gregory PJ and Nortcliff S (eds.), *Soil Conditions and Plant Growth*, 1st edn, pp. 238–268. Blackwell Publishing Ltd.

residue of the previous year's crop (Figs. 2 and 3). A special planter is generally required to penetrate residual vegetation and a compacted soil surface. However, occasional tillage may be required to control weeds that are resistant to herbicides.

Overall, the main impetus behind the reduction in tillage intensity is to reduce the degree of soil degradation often associated with excess tillage, and to reduce tillage costs in commercial agriculture. In many cases, uniform tillage across a field may not be required or desirable. For row crops (e.g., maize (*Zea mays* L.), sugar beet (*Beta vulgaris* L.), oilseed rape (*Brassica napus* L.)) tillage can be restricted to specific areas across a field (strip or zone tillage). Seedbed preparation can be confined to the planting row, while the remaining area receives less or a different type of tillage treatment. In these not planted areas, a coarse soil structure is maintained to facilitate water infiltration. Further to this, crop residues can be kept on the soil surface outside the planting rows to reduce evaporation and protect soil surface from rain drops and wind.

Basic soil tillage operations consist of soil inversion or fragmentation (Table 1). Specific tillage tools operating at a range of soil depths are used to achieve these operations. Soil inversion is best accomplished using a moldboard plow, while soil-loosening can



Fig. 3 Examples of tillage operations in a long-term tillage and rotation experiment: (a) moldboard plowing to 20–25 cm depth; (b) rotavator harrowing to 8–10 cm depth; (c) harrowing to 3–4 cm depth; (d) no tillage using a single disk planter. Photos: Lars Andreassen.

Table 1 The characteristics, aims and some essential effects of common tillage operations.

Tillage operation	Main machines involved	Principal aims and effects achieved on soil structure
Inversion tillage	Moldboard plow, offset disks	Inverts, fragments and loosens the soil; buries weeds and residues from the previous crop, incorporates manure and organic amendments.
Ridge till	Ridge tillage cultivator and modified planter	Soil is pushed by pairs of winged tines to form ridges up to 20 cm deep. The harvest residues on the surface are removed and the top of the ridge flattened by a horizontal blade so that the seed is planted in the ridges without residue cover in the row. Contrasting soil conditions are created in the ridge and between the ridges. Residues and organic amendments are mixed with soil in the sides of the ridges each year. The ridges warm up quicker than flat land in spring, in part because the ridges drain more readily.
Reduced tillage	Rigid or sprung tines, disks, powered tines, rotary implements	Fragments and loosens the soil. Harvest residues can be left on the soil surface. Can be performed to a depth greater than 100–150 mm (deep reduced tillage) or to a depth of less than 100 mm (shallow reduced tillage).
Stubble-mulch tillage	Tines or disks	Fragments and loosens the soil. Residues are retained on the soil surface.
Zero-tillage/no-till/direct drilling	Drill with tines or disks	Seed is drilled into the stubble of the previous crop with minor soil disturbance confined to the row and does not extend below seeding depth.
Zone or Strip-tillage	Rigid or sprung tines	Tillage with coulters or rigid tines (down to 15 cm below the soil surface) in a narrow strip of soil. Seeding takes place in the tilled zones, with the inter-row area remaining undisturbed and where crop residues are retained.
Strategic tillage	Rigid or sprung tines, disks, powered tines, moldboard plow	A system predominantly using reduced tillage equipment but includes the use of moldboard plowing as a flexible responsive to seedbed concerns (typically weeds) or the need to incorporate organic amendments, manure or crop residues.
Rotational plowing	Rigid or sprung tines, disks, powered tines, moldboard plow	Reduced tillage or no-till is used routinely for soil preparation but a plow is used at key points in the rotation.
Secondary tillage	Powered or unpowered tines, disks, rotary implements, rollers	Tillage practices used after the principal tool to create a seedbed and provide additional weed control
Remedial tillage	Subsoiler tines, rotary spades or mole plow	Deep tillage techniques used to remediate soil compaction.

be achieved using several tillage tools. Rotary cultivators are well suited for mixing soil and often used to mix soil fertilizer and organic amendments into the soil. Plant residues can interfere with some tillage tools. Coulters or disks are often used to cut or divert crop residue away from the immediate tillage zone. Repeated or heavy traffic in the field might lead to soil compaction in the subsoil, i.e., the soil below the tillage depth. Subsoiling is the attempt to mechanically loosen the compacted subsoil (remedial tillage operations, Table 1). An efficient subsoiling operation should be performed at a water content where the soil is friable to avoid smearing. A mechanically loosened subsoil has a low strength and can be easily and severely recompacted.

In many cropping situations and climates, innovations in tillage practices have led to the development of conservation tillage or conservation agriculture systems. Conservation tillage is a generic term used to describe tillage systems that have the potential to protect soil by reducing soil loss (by wind and water erosion) relative to some form of conventional tillage (Fig. 2 and Table 1). A well-accepted operational definition of conservation tillage is a cover of 30% or greater field area of crop residue on the soil surface after planting. Thus, leaving crop residues on the soil surface is fundamental for conservation tillage systems. Conservation agriculture can be viewed as an extension of the conservation tillage concept as it includes: (1) minimal soil disturbance, (2) Permanent soil organic cover (residues and living plants) and (3) Species diversification.

Why tillage?

The traditional aims of tillage are to improve soil structure for crop growth, incorporate organic amendments into the soil, and to control weeds. The last goal can often be met by use of herbicides, and this has led to the development of no-tillage systems, where tillage is confined to soil disturbance associated with crop seeding or planting (Figs. 2 and 3). Soil tillage is also involved in soil-water conservation and regulation, and in soil-erosion control. Table 2 outlines some of the various pros and cons of tillage in agricultural systems. The role of tillage in soil structure improvement has various facets related to crop growth and productivity. Tillage is conducted to improve soil functions such as water and air regulation and flow, increasing infiltration, improving drainage, and enhancing the water-storage capacity of the soil. It is used to create a desirable aggregation size distribution conducive for crop seed-soil contact. Soil loosening to alleviate soil compaction can also be a key objective of tillage. Such an operation is labeled subsoil loosening or subsoiling when it is carried out below the normal depth of tillage (20–30 cm). For subsoiling, straight, curved or parabolic shanks are typically used. Tillage is also used to manipulate crop residues, either by burying and mixing them into the soil, or by retaining them at the soil surface to serve as mulch for soil protection or as a barrier to reduce runoff and soil erosion, improve water infiltration and limit evaporation. The depth and intensity of tillage needed are not the same for all soil types and cropping situations. Some soils, due to climate (e.g., cool, wet soils) or type (e.g., high clay or sand content, or poor permeability), have a relatively high need for tillage. In contrast, there are soils with a relatively low tillage requirement, as they are able to regenerate soil structure under natural processes associated with soil wetting and drying, and freeze-thaw processes.

Why NOT tillage?

Improved and optimal growth condition for the crops has been the aim of tillage for thousands of years (Fig. 4). However, the need for soil tillage should be carefully investigated, as operations performed in non-optimal conditions might be detrimental for plant growth and the environment (Table 2). For almost 200 years ago, the moldboard plow opened new possibilities in plant production allowing for crop rotations that include grass, and for incorporation of manure, both of which resulted in higher yields. However, the extensive use of moldboard plowing to convert natural arid grassland, the Great Plains, to arable land in North America resulted in severe dust storms in the 1930s, damaging significantly both the ecology and the agriculture. These events, referred as the Dust Bowl in the United States, sparked interest in conservation tillage (Faulkner, 1943). Plowing is both energy and time consuming. The lack of horses to pull the plow during the 1st World War encouraged farmers to drop the plowing operations (Holladay, 1918). A similar trend was observed in connection to the energy crisis in the 1970s, and during the last decade owing to the high prices of fossil fuels. Nowadays, the need for plowing is again questioned in order to reduce CO₂ emissions to the atmosphere.

Table 2 Pros and cons of soil tillage.

<i>Pros</i>	<i>Cons</i>
Ideal seedbed	Breakdown soil structure
Weed control	Promote runoff
Reduce compaction	Promote erosion
Incorporate organic matter	Destroy macropores
Disease control	Promote drying
Soil warming	Disturb soil biota
Break up crust	Increase temporally SOM turnover
Improve infiltration, drainage, water storage	Increase daily temperature variation in soil
Decrease pesticide use	Increase energy consumption and working time



Fig. 4 Representation of a tillage operation at Aspeberget, Sweden, Bronze Age (c. 1700 BC–500 BC). These rock carvings are UNESCO's World Heritage property. Photo: Mathieu Lamandé.

Mixing, loosening and inverting the topsoil layer affect the soil quality over time (Karlen et al., 1994). High erodibility after plowing has also been a strong driver in the development of plowless systems. The 1930s Dust Bowl in the United States initiated and accelerated the development of conservation tillage technology in North America. Erosion affects soil quantity and quality, and has negative effects on the environment (eutrophication of surface water bodies, reduced water quality, sediments with large amount of adsorbed phosphorus or pesticides). When tillage operations are performed in wet conditions, the soil may be compacted, and the soil functions may be impacted. Soil tillage operations, especially moldboard plowing, increase evaporation and water loss from the soil. The inversion of the top layer by plowing produces a soil surface without any plant residues that usually protect it from sealing and crusting, from erosion and evaporation. In dry areas, conservation tillage (see the “Tillage system” section for a definition) helps reducing evaporation and keeping the soil water for plant growth (Duley and Matthew, 1947). Plant residues on the soil surface help also to reduce the daily soil temperature variation, mainly by increasing the reflection of the solar energy (i.e. by increasing the albedo of the soil surface) (Hay et al., 1978). The content of organic matter is very important for quality and health of mineral soils. In soils used for annual arable cropping, the content of soil organic matter (SOM) is often decreasing over time, due to export of harvested products and residues. Soil tillage, and especially moldboard plowing, affects the SOM content mainly by mixing into the tilled layer, but not by reduction of the total amount in the soil profile (Gómez-Muñoz et al., 2021). After some years of conservation tillage, the content of SOM in the surface layer will increase, due to concentration of organic matter inputs in this layer. However, the content of SOM will decrease deeper in the soil profile due to lack of tillage induced soil mixing. Nevertheless, an increase of SOM content in the surface layer may be very important for the soil properties connected to plant growth.

Soil fragmentation in tillage

Soil fragmentation is the process of crumbling of soil fragments under applied stress. The easier it is to fragment the soil the easier it is to perform tillage (Utomo and Dexter, 1981). The ease of soil fragmentation depends on the internal soil pore structure, the soil water content or more precisely the soil water potential, the soil mechanical and hydraulic properties, and the mechanical action of the tillage implement. A soil fragment is a conglomeration of organic and inorganic particles that cohere to each other more than to the neighboring particles. The aim of tillage is to obtain an overall loosening of soil without disturbance of the microstructure of the soil matrix. In addition, the satisfactorily soil loosening should be achieved efficiently with regard to energy consumption. Therefore, brittle failure of soil fragments is desirable. Brittle fracture is the soil failure by crack propagation along the weakest planes of soil fragments. Cracks propagate rapidly through the soil fragments without a need to increase the mechanical stress applied. There is little plastic deformation associated with brittle failure, which means a low energy absorption during fragmentation. The ability of soil to display brittle failure is often quantified by measuring the tensile strength, i.e., the mechanical stress or force per unit area that is required to break it along the weakest failure planes, for differently sized soil fragments. Ideally, the soil should easily fail into a particular size range of smaller fragments—making up an ideal seedbed. This ability is labeled soil friability. A friable soil is characterized by a high ease of fragmentation of large aggregates or clods and a low ease of fragmentation of minor aggregates into undesirable small elements.

Table 3 Soil water range for tillage reported in the literature.

Soil workability limits			Soils studied	Climate	Source
Wet limit	Optimum water content	Dry limit			
$\omega = 0.24 \text{ kg kg}^{-1}$	—	$\omega = 0.18 \text{ kg kg}^{-1}$			Sitkei (1967) cited in Dexter and Birkas (2004)
—	θ at maximum Proctor density	—	Silty clay loam and silt loam	Humid continental climate	Wagner et al. (1992)
θ at plastic limit (θ_{PL})	$0.9 \theta_{PL}$	θ at which the strength of soil is twice the strength at $0.9 \theta_{PL}$	Clay loam to silty clay loam	Temperate	Dexter and Bird (2001)
θ at $\varphi = -100 \text{ hPa}$	—	θ at $\varphi = -1250 \text{ hPa}$	Loam	Tropical	Hoogmoed et al. (2003)
θ at $\varphi = -125 \text{ hPa}$	—	θ at $\varphi = -3162 \text{ hPa}$	Clay	Tropical	Hoogmoed et al. (2003)
—	0.7θ at $\varphi = -50 \text{ hPa}$	—	Sand to clay	Continental temperate humid to continental dry	Mueller et al. (2003)
$\omega = 0.40 \text{ kg kg}^{-1}$	—	$\omega = 0.33 \text{ kg kg}^{-1}$	Clay	Temperate	Gülser et al. (2009)

θ : volumetric water content; ω : gravimetric water content; φ : matric potential.

After Obour PB, Lamandé M, Edwards G, Sørensen CG, Munkholm LJ (2017) Predicting soil workability and fragmentation in tillage: A review. *Soil Use and Management* **33**: 288–298.

Water range for tillage

The soil normally displays optimal friability at a specific range in water content—the so-called water range for tillage. At wetter state, the soil will tend to deform plastically and at a drier state the soil will need strong force to fragment and tend to pulverize rather than fragment it into a desirable size range of smaller fragments. The water range for tillage will typically increase with increasing soil organic carbon content and decrease with degree of compaction and clay content (Table 3).

Outlook

Soil tillage has been essential to improve and sustain crop growths since the very beginning of agriculture 12,000 years ago. Various tillage operations have been developed since, human or animal powered, then leading to the modern mechanized tillage systems. The type of tillage operations and the intensity of tillage should be carefully investigated for given pedo-climatic regions to ensure a sustainable crop production without detrimental consequences on the environment. The timing, with regard to soil mechanical conditions, and the intensity of a tillage operation will determine the quality of the results. Appropriate tillage systems will prevent soil erosion, soil compaction, flooding, and pollution of water bodies.

References

- Dexter AR and Bird NRA (2001) Methods for predicting the optimum and the range of soil water contents for tillage based on the water retention curve. *Soil and Tillage Research* 57: 203–212.
- Dexter AR and Birkas M (2004) Prediction of the soil structures produced by tillage. *Soil and Tillage Research* 79: 233–238.
- Duley FL and Matthew OR (1947) Ways to till the soil. In: *Yearbook of Agriculture 1943–1947*. Washington: U. S. Government Printing Office.
- Faulkner EH (1943) *Plowman's Folly*. Norman: University of Oklahoma Press, 174.
- Gómez-Muñoz B, Jensen LS, Munkholm L, Olesen JE, Møller Hansen E, et al. (2021) Long-term effect of tillage and straw retention in conservation agriculture systems on soil carbon storage. *Soil Science Society of America Journal* 85: 1465–1478.
- Gülser C, Selvi KÇ, and Iç S (2009) Some mechanical properties and workability of soils in Karadeniz Agricultural Research Institute field. *Journal of Agricultural Machinery Science* 5: 423–428.
- Hay RKM, Holmes JC, and Hunter A (1978) The effect of tillage, direct drilling and nitrogen fertiliser on soil temperature under a barley crop. *The Journal of Soil Sciences* 29: 174–183.
- Holldack L (1918) Die Kulturmethode Jean. *Mitteilungen der Deutschen Landwirtschafts-Gesellschaft* 19: 21–22.
- Hoogmoed WB, Cadena-Zapata M, and Perdok UD (2003) Laboratory assessment of the workable range of soils in the tropical zone of Veracruz, Mexico. *Soil and Tillage Research* 74: 169–178.
- Karlen D, Wollenhaupt NC, Erbach DC, Berry EC, Swan JB, et al. (1994) Long-term tillage effects on soil quality. *Soil and Tillage Research* 32: 313–327.
- Mueller L, Schindler U, Fausey NR, and Lal R (2003) Comparison of methods for estimating maximum soil water content for optimum workability. *Soil and Tillage Research* 72: 9–20.
- Sitkei G (1967) *Soil mechanical problems of agricultural machines (Amezőgazdasági gépek talajmechanikai problémái)* (in Hungarian). Akadémiai Kiadó, Budapest.
- Utomo WH and Dexter AR (1981) Soil friability. *Journal of Soil Science* 32: 203–213.
- Wagner LE, Ambe NM, and Barnes P (1992) Tillage-induced soil aggregate status as influenced by water-content. *Transactions of the ASAE* 35: 499–504.

Weed and soil management: A balancing act

Trevor Kenneth James^a and Charles Norman Merfield^b, ^aAgResearch, Hamilton, New Zealand; ^bThe BHU Future Farming Centre, Canterbury, New Zealand

© 2023 Elsevier Ltd. All rights reserved.

Introduction	440
Weeds	441
What is a weed?	441
Why manage weeds?	442
Plants and soil health	442
Plant diversity	442
Bare soil	443
A balancing act	443
Weed and soil management interactions	443
Tillage	444
Vertical distribution of the weed seedbank	444
Vegetation clearance, minimum and no-tillage	444
Soil applied herbicides	445
In-crop weed management	445
Integrated weed management	445
Herbicides and mechanical weeding	446
Cover crops	446
Green manures and catch crops for nutrient management	446
Smother crops	447
Living mulch	447
Relay cropping and undersowing	447
System level techniques	447
Rotations	448
Crop and pasture mixtures - intercropping	448
Grazing management	448
Conclusion	448
References	449

Abstract

Weed management interacts with soil in many ways. Its primary aim of reducing the number and biomass of non-crop plants means that it simplifies agroecosystems, although it produces negative ecological and environmental outcomes including reducing soil health. At the same time weed management is utterly essential, as without it weeds would overrun crops within a few years, dramatically reducing yield. Integrated weed management (IWM) is where, physical (e.g. mechanical hoeing), chemical (i.e., herbicides), biological (e.g. biological control) and ecological (e.g. rotations) techniques are used in an integrated and symbiotic manner. Integrated Weed Management is the approach recommended by the international weed science community.

Glossary

Allelopathic Suppression of plant growth and seed germination by a toxin released from a nearby plant.

Broad spectrum herbicide Herbicide that kills a wide range of plants, both monocotyledons (grasses) and dicotyledons (broad leaves).

Dedifferentiate The process by which differentiated plant meristems (buds) i.e., root, stem and cambium, can change into another form of meristem (e.g., root meristem changes into shoot meristem).

Fallow To leave an agricultural field bare of plants, either using herbicides, or more typically tillage.

Herbicide resistance Herbicide resistance is the inherited ability of a plant to survive and reproduce following exposure to a dose of herbicide normally lethal to the wild type. In a plant, resistance may occur naturally or be induced by techniques such as genetic engineering or selection of variants produced by tissue culture or mutagenesis.

Herbicide tolerance Herbicide tolerance is the inherent ability of a species to survive and reproduce after herbicide treatment. This implies that there was no selection or genetic manipulation to make the plant tolerant; it is naturally tolerant.

Organ of perennation The structure within a plant that allows it to persist over multiple years, e.g., a bulb or tap root.

Raunkiaer system A plant categorizing system based on life-form, devised by Danish botanist Christen C. Raunkiaer.

Residual herbicide A herbicide that is applied to the soil and kills weed seedlings as they germinate and emerge.

Selective herbicide A herbicide that will kill some plant species (the weeds) but not others (the crop).

Systemic herbicide A herbicide that translocates (moves) within a plant's vascular system.

Therophyte Annual plants that complete their lifecycle rapidly in favorable conditions and survive unfavorable seasons as seeds.

Translocate To move within a plant's vascular system (phloem and xylem).

Weed seed rain The deposition of viable weed seeds from their parent plant into the weed seedbank.

Weed seedbank All the weed seeds that reside in the soil.

Key points

- Weed management aims to reduce plant biodiversity to benefit the crop, but reduced biodiversity has multiple negative environmental outcomes including on soil health.
- Weed management is essential, without it weeds will overrun crops, and yield will be drastically reduced.
- Integrated weed management (IWM) integrates many different weed management techniques achieving synergies among them, is the approach recommended by weed scientists.
- Techniques such as mechanical weeding and herbicides tend to have the greatest negative impacts, particularly when non-crop plant numbers are reduced to very low levels and soil is left bare.
- Techniques that use biological and ecological approaches, e.g. living mulch, have lower negative impacts and can also have positive impacts.
- Weed management is therefore a balancing act between achieving sufficient weed management to achieve good yields, while, minimizing the negative impacts it can have on soil and all aspects of the farm ecosystem.

Introduction

Weed management interacts with soils in multiple ways. The primary aim of weed management is to limit weed populations below the levels where they cause economic losses, both in current and future crops. Weed management therefore deliberately aims to reduce plant biodiversity, both in the amount and range of non-crop species, which produces a cascade of negative ecological effects, including reduced soil health. Where weed management achieves a very large reduction in weed populations bare soil can result which is at risk of erosion and other damage. Tillage (e.g. mouldboard ploughing/inversion tillage) is an integral part of weed management in many farming systems, particularly cropping systems (e.g., cereals and vegetables), and tillage's negative impacts on soil are well documented, including in this encyclopedia. In response to the harmful effects of tillage, especially inversion tillage, on soil health, reduced/minimum tillage and no-tillage systems were developed and have been widely adopted. Changing tillage systems can have large impacts on both weed flora as well as soil health.

The soil is where the weed seedbank is situated, and particularly for therophyte weeds, managing the weed seedbank is critical for effective weed management, especially under non-chemical and integrated weed management systems (Fig. 1). Soil health and



Fig. 1 Weed seed banks can contain millions of seeds giving rise to large numbers of weeds, *Digitaria sanguinalis* (L.) Scop. uncontrolled in maize (*Zea mays* L.).

function also have a large impact on the weed seedbank, as there are many organisms in soil, from microorganisms through invertebrates to vertebrates that predate on seeds, so soil management and soil health can affect how much of the weed seedbank is lost to predation. In comparison, mechanical techniques to reduce the weed seedbank, such as fallowing are highly damaging to soil due to both repeated tillage and absence of living plants, resulting in compaction, loss of structure, reduction in soil biological diversity, etc.

Integrated weed management (IWM) is the approach promoted by weed scientists as the most effective long-term and damage limiting approach to weed management. It systematically integrates physical, chemical, biological and ecological approaches using the technique that achieves optimal weed control with the fewest negative outcomes for the wider environment for any given issue. However, physical and chemical techniques are often harmful to soil, and greater use of biological and ecological techniques is required. Additionally, over reliance on chemical control can lead to herbicide resistant weeds. One such technique is cover cropping, which has many permutations and can achieve good weed management while at the same time avoiding the damage and consequences of physical and chemical techniques and even improving soil health at the same time.

System level farm tools for weed management, for example, diversified rotations, can have positive effects on soil health, e.g. the inclusion of a pasture phase in a cropping rotation, improves all aspects of soil quality and health, and reduce the soil seedbank and therefore the need to use harmful agrichemicals.

While weed management has multiple effects on soil, many negative, effective weed management is critical in farming, particularly cropping. If not controlled, weed populations can rapidly reach exceptionally high levels to the point that crop yield is dramatically reduced, even to the level of complete crop loss. Further, for any given crop or pasture there will always be many different species of weeds present and any given weed species can infest a very wide range of crops and pastures, which means weeds are the ubiquitous pests of crops and pasture. This contrasts with invertebrate pests and diseases (e.g. fungi and bacteria) of plants which are typically highly host specific, i.e., any given pest can attack only a narrow range of host species, and vice versa any given crop species is only attacked by a small range of pests and diseases. Weeds can also host crop plant pests and diseases. Effective weed management is therefore vital for successful agriculture and horticulture.

Weed management therefore interacts with soil in many different ways, and while often having negative effects, there are many techniques available to help ameliorate negative effects and increase positive effects. Weed management is therefore a balancing act, between achieving sufficient weed control for good crop yield and quality, while minimizing negative impacts on soil and the rest of the farm environment.

Weeds

What is a weed?

While those working with weeds in the field have a clear intuitive understanding of what a weed is, there is no agreement on a precise definition. For example, the Weed Science Society of America defines a weed as “A weed is a plant that causes economic losses or ecological damage, created health problems for humans or animals, or is undesirable where it is growing,” while the European Weed Research Society defines weeds as “any plant or vegetation, interfering with the objective or requirements of people.” However, both of these highlight that defining a plant as a weed is a subjective not objective decision, particularly when it comes to defining a specific plant or population of plants at a given time, as weeds. For example, a crop species such as *Triticum aestivum* L. (common wheat) can become a weed in the following year in a crop of *Brassica napus* L. (oil seed rape).

The advent of the herbicides around the 1940s, provided farmers with new tools that made the control of weeds much easier. Many plants that had not previously been considered weeds, because the effort to control them was much greater than the benefits of removing them, were reclassified as weeds. This has tended to result in all non-crop plants now being considered weeds. However, other approaches exist; for example, traditional farmers in the tropical agroecosystems of south-eastern Mexico divide plants in the field into three groups, 1 the crop, 2 buen monte (good plants) and 3 mal monte (bad plants) (Chacón and Gliessman, 1982). Buen monte are plants that establish themselves in the field (i.e., they are not planted) but are recognized as still being of value, often as food, but also for other benefits, e.g. medicinal. Therefore, these plants are not removed, and may even be cultivated if they are of sufficient value and or scarcity. These plants can be defined as non-crop plants. Mal monte are plants that establish themselves in the field, but have no value or worse, have negative values (they are bad). For example, they can smother the crop and reduce yield, or are toxic/harmful, so they are therefore removed, i.e., they are weeds. Therefore, for these farmers not all non-crop plants are weeds, rather, at the individual plant and species level, a decision is made as to whether they are good albeit non-crop plants or bad and therefore weeds (Fig. 2).

Within the developed world with intensive farming systems a similar reappraisal is occurring. For example, Storkey and Neve (2018) argue that weeds have value in agroecosystems not only for environmental values such as biodiversity, but also within an ecological framework if they are managed. The increased weed diversity can be beneficial for yield, system stability, sustainability and resilience, e.g. in the face of weather extremes due to climate change. As weed management can have a number of negative impacts on soil, and other environmental outcomes, e.g. insect populations and biodiversity, moving from the view that any non-crop plant is a weed and must be controlled to a more nuanced view that non-crop plants can have multiple benefits, including agronomic, is going to be vital to reduce the negative outcomes of weed management, particularly the evolution of herbicide resistant weeds.



Fig. 2 Mal monte (bad plants), in this example the mal monte was the previous crop of radish (*Raphanus sativus* L.) but now a mal monte in the present crop of barley (*Hordeum vulgare* L.).

Why manage weeds?

The primary reason for managing weeds is that they compete with crop and pasture plants, for nutrients, water and or light, thereby reducing yield/production and therefore decreasing the amount of food and fiber produced, and where crops are sold, profit. Weeds can also be harmful in other ways: for example, they can harbor crop pests and diseases, they can reduce the quality of produce (e.g. *Avena fatua* L. (wild oat) seed contaminating wheat grain for flour, weed seeds contaminating planting seed, or burs in sheep's wool), they can be toxic to livestock, and increase production costs from in-field weed plant control, to post harvest contamination and cleaning.

Weeds are also a ubiquitous pest, in that in any given crop or pasture, many tens, if not hundreds, of weed species will be present, over a very large range of populations, and, any given weed species is able to infest a very wide range of crops or pasture (cropping weeds and pasture weeds tend mostly to be different species; cropping weeds mostly annuals, pasture weeds mostly perennials). Weeds therefore have a 'many-to-many' relationship, meaning that all crops and all pastures will have weeds present (there is no such thing as a weed-free crop or pasture), and any one weed species can occur in many different crops or pasture. In comparison, invertebrate pests and diseases typically have a 'one-to-one' or 'few-to-few' relationship with their hosts, in that any given species of pest or disease can only attack one or just a few host species, and any one host is susceptible to a tiny fraction of the total invertebrate pests and diseases that attack plants. If environmental and ecological conditions are favorable, any crop species may be unaffected by most of its potential pests and diseases at a given time and in ideal conditions it may be completely free from all of them. Weeds are therefore a far greater threat to crops and pasture than invertebrate pests and diseases.

Many weeds, particularly annuals and biennials, can produce prodigious amounts of seed, for example tens even hundreds of thousands of seeds per plant. If uncontrolled, the populations of these weeds will grow exponentially, such that within a few years, their sheer numbers will make it impossible to grow crops. Successful weed management is therefore not just focused on controlling weed plants growing within a crop or pasture, it is even more focused on preventing weeds going to seed and increasing the size of the weed problem in the future.

The many harmful attributes of weeds, their ubiquitous nature and their massive reproduction rates mean that effective weed management is always required in every crop and pasture if the long-term productivity of farming is to be maintained. Effective weed management is therefore essential.

Plants and soil health

All plants: crops, non-crop plants and weeds perform the key function of the primary suppliers of energy and nutrients to soil biology which in turn controls soil health (Bender et al., 2016). This is at the heart of the interaction of weed management and soil. Since their development in the mid 20th century, herbicides have taken over weed control in most developed countries, which is now leading to the problems of herbicide resistant weeds and to depleted soil biology.

Plant diversity

Living plants are the primary controlling factor of soil biology and therefore soil health. This is because, as autotrophs/primary producers, it is plants that harvest energy from the sun and use that to extract carbon dioxide from the atmosphere to create complex

organic compounds (i.e., carbohydrates, fats, and proteins), which are then transferred to the soil ecosystem directly by living plant roots and also to a lesser extent from dead plant material, both roots and foliage (Sokol et al., 2019). Where plants are eliminated from a soil, by herbicide application or tillage for example, the soil's primary source of energy and complex organic chemistry is lost, and its health starts to decline.

Further, ecological science has now clearly demonstrated that plant diversity is also critical for maximizing soil health attributes, such as soil organic matter/soil carbon, and ecosystem diversity, particularly microbial diversity. A diversity of plant species delivers a much greater diversity of complex organic compounds, and due to different root morphologies they explore different soil depths and aggregate spaces which in turn affects nutrient and water dynamics all of which creates even greater soil biodiversity (Weisser et al., 2017). Plant biomass is critical to soil health but is optimized with greater plant diversity.

Bare soil

A lack of plants is highly detrimental to soil health. In tropical and temperate climatic zones of most ecosystems, soil is generally covered by plants, both living and their dead residues, e.g. the detritus layer in forests. This plant cover protects the soil from harmful factors such as becoming hot under direct sunlight, damage to soil structure from rain-drop impact, flooding from high rainfall events that can erode soil, high winds that can blow soil away and compaction from livestock. Intensive weed management often results in bare soil among the crop plant stems, which is therefore at risk of being eroded by wind and water (sheet, rill and gully erosion) and an overall reduction in soil health.

Where plants are removed from soil for extended periods, e.g. herbicide strips under perennial crops such as apples and grapes, soil health can be dramatically reduced, resulting in severe compaction, loss of structure, organic matter and soil biology.

A balancing act

The fundamental aim of weed management is to reduce the number and populations of weed plants in crops and pastures, and therefore weed management inherently works against soil health. It is therefore critical to understand the details of the interaction of weed management and soil health, and how best to minimize harm and maximize positive outcomes, i.e., a balance is required. Weed control needs to be good enough so that, in both the short and long terms, weeds do not have negative impacts on crops and pasture. However, weeds must be controlled only to the point that yield loss and other negative production outcomes of weeds are avoided, but no further, or replace *mal monte* with *buen monte*, such as cover crops (discussed below), so that the negative environmental outcomes of weed management are minimized. That balancing act is the focus of the rest of this chapter.

Weed and soil management interactions

The Weed Seedbank

In arable and vegetable systems most crop species are short-term annual plants, therefore the weed species that have a good evolutionarily fit in these systems are also short-term annuals that can complete their lifecycle within the production time of the crop. Longer lived plants, i.e., biennials and perennials, do not survive as living plants but are typically killed between crops by tillage or herbicides. Under the Raunkiaer system, these short-term annuals are classified as therophytes, as the resting buds are in the form of seeds. From this perspective it is not the weed plants growing in the crop that are the fundamental part of annual weeds' lifecycle, rather it is the seeds in the soil, i.e., the weed seedbank, that is the permanent life stage of annual weeds. Effective annual weed management therefore needs to have a strong focus on the weed seedbank in the soil; so once again, weed management and soil are intimately interlinked.

The size of the weed seedbank can vary by orders of magnitude, for example, Rahman et al., (1996) found the number of weed seeds per kilogram of soil collected from six maize (*Zea mays* L.) fields in Waikato, New Zealand, varied from 1 to 695 seeds kg⁻¹ dry soil. The number of weeds germinating from that soil was strongly correlated ($r = 0.96$) to the number of seeds in the seedbank (Rahman et al., 1996). While there is little or no direct interaction between the weed seedbank and crop plants, they tend to emerge simultaneously (competition) and a large weed seedbank will result in a large number of in-crop weeds which will have a large negative impact on the crop. Therefore, minimizing the weed seedbank is critical for reducing in-crop weed plants.

The primary means of managing, i.e., reducing, the weed seedbank is to prevent in-crop weed plants seeding and refilling the weed seedbank, a process called the 'weed seed rain.' Managing in-crop weed plants is the dominant focus of herbicide-based weed management, which is so effective in crop weed management and will also minimize weed seed rain. Here, practices such as cover crops that reduce, but do not eliminate, seed rain are better than herbicides for managing the weed seedbank. There can also be a large loss of seeds in the weeks immediately following seed shed through predation by vertebrates such as birds and rodents, soil dwelling invertebrates such as carabid beetles, and over the longer term by microorganisms such as fungi (Gómez et al., 2013; Kennedy, 1999). Both soil and weed management can have a significant impact on seed predating species and their abundance, both positive (e.g. cover crops), and negative (e.g. herbicides), so in turn it can have a considerable impact on the weed seedbank in both the short and longer term.

For non-herbicide weed management systems, management of the weed seedbank is particularly important. One technique for diminishing the weed seedbank is tilled fallow, where the soil is repeatedly tilled to shallow (<10 cm) depths, during the growing

season, to encourage weed seeds to germinate and then kill the emerged weed seedlings (Rahman et al., 2001). While this can be highly effective, regular tillage and keeping fields free from vegetation over extended periods, it is highly damaging to soil structure, reduces organic matter and overall soil health, so this technique has fallen out of favor. A more targeted approach involves false and stale seedbeds. These are implemented by setting up the land for planting, either with tillage or herbicides, but, then delaying planting by one to 3 weeks to allow weed seeds to germinate, particularly where tillage has stimulated germination. Then in a false seedbed the weedlings are killed through very shallow re-tillage (<5 cm), or in a stale seedbed by herbicides or thermal weeding, e.g., a flame weeder (Merfield, 2015). These can be very effective at reducing in-crop weeds, while having much less impact on soil health than fallow, because of the short duration when the soil is free from plants and the low intensity of the re-tillage used.

The weed seedbank occupies the soil which means that its management and soil management are intertwined. What is good for seedbank management may be damaging to soil and vice versa, whereas actions that benefit soil, e.g. leaving more weed plants, may increase the size of the weed seedbank. Again, balance is required.

Tillage

The purpose of tillage is often considered to be to improve soil structure, enhance water infiltration and aeration, incorporate fertilizers, increase crop seed germination and boost crop and pasture growth, however, a key function of tillage is also weed management. There are two main approaches: first, to eliminate living weed plants and also plant residues from the field prior to establishing a new crop or pasture. Second, tillage changes the vertical distribution of the weed seedbank and different forms of tillage or no-tillage produce different distributions that can have important implications for weed management.

Vertical distribution of the weed seedbank

In reduced tillage systems the seedbank is stratified closer to the soil surface while inversion tillage creates a more even distribution through the soil profile (Fig. 3).

Weed species differ in their response to different tillage intensities with some increasing under reduced tillage and others decreasing, with often opposite outcomes for inversion tillage (Table 1).

Tillage approaches can therefore be varied to address particular weed problems. For example, where grass weeds build up in no- and min-tillage systems, occasional use of inversion tillage, e.g., once in 5 years, will bury the grass seeds at depth. Grass seeds lack the hard seed coat of dicotyledons, therefore they die after a few years of burial and are eliminated from the seedbank.

Vegetation clearance, minimum and no-tillage

Traditionally, clearing vegetation was done with the mouldboard plough which inverts the soil, burying the weeds and residues on the soil surface and bringing up soil free from plants and residues. The scientific evidence is very clear that tillage has multiple negative impacts on soil, with inversion tillage having the biggest effect. A classic example of this was the dust bowls in North America in the 1930s and 1940s. They resulted in the introduction of reduced and minimum tillage approaches such as disk harrows that only till 5 to 10-cm deep, which are used instead of the plough to kill off weeds and incorporate residues. The logical end point of reducing tillage intensity is no-tillage systems, where only the absolute minimum amount of soil and residue is disturbed to drill and germinate crop seeds successfully. No-tillage systems are particularly important in regions with low rainfall

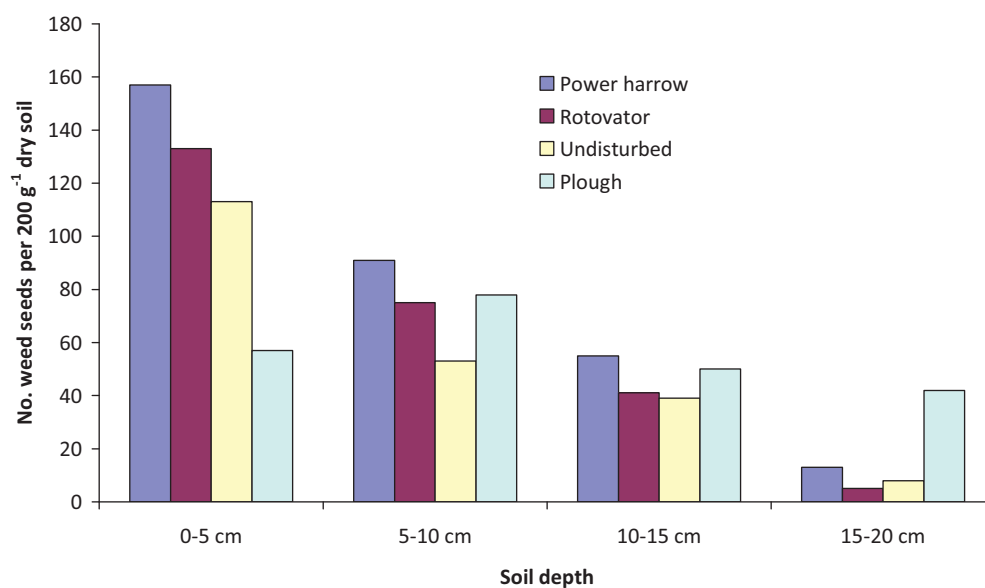


Fig. 3 Distribution of weed seeds (total no 200 g⁻¹ dry soil) in the soil profile after cultivation treatments (Rahman et al., 2000).

Table 1 Influence of tillage systems on densities of selected summer annual weed species in *Zea mays* L. (maize) and *Glycine max* (L.) Merr. (soybean) production systems.

Weed species	Tillage response ^a
<i>Abutilon theophrasti</i>	—
<i>Amaranthus retroflexus</i>	+, 0, —
<i>Cassia obtusifolia</i>	—
<i>Cenchrus incertus</i>	+
<i>Chenopodium album</i>	+, 0, —
<i>Echinochloa crus-galli</i>	+
<i>Panicum dichotomiflorum</i>	+
<i>Setaria complex</i>	+
<i>Xanthium strumarium</i>	—

^a+ = Increased density with reduced tillage, 0 = no change with reduced tillage, — = decreased density with reduced tillage.

and erosion prone soils, such as in central South and North America. No-tillage systems, however, rely on broad-spectrum, systemic herbicides, principally glyphosate, to replace the role of tillage in eliminating weeds and other plants prior to sowing new crops, such that the use of herbicides for this purpose is referred to as chemical ploughing (Baker and Saxton, 2007). However, the frequent and continuous use of the same type of herbicide is a risk because of selecting herbicide resistant weeds. The main aim of minimum and no-tillage is protection of the soil. To quote Baker and Saxton (2007; p. 1.) “No farming technique yet devised by humankind has been anywhere near as effective as no-tillage at halting soil erosion and making food production truly sustainable.”

In addition to achieving the minimum level of soil disturbance necessary to accomplish successful crop emergence, no-tillage also aims to maintain as high a level of soil cover as possible, either by living plants or plant residues. This is also a key aim of conservation agriculture. Soil cover is vital to protect the soil from wind and water erosion and from becoming excessively hot due to solar radiation; therefore, it helps to maximize soil health by regulating soil temperature and water dynamics. No-tillage therefore reduces farming’s impact on the soil by minimizing soil disturbance and maintaining soil cover, but it is only currently possible by using herbicides.

This shows the close relationship between tillage, including no-tillage, and weed/vegetation management, and how different tillage/no-tillage approaches affect soil health.

Soil applied herbicides

Herbicides are divided into two groups, foliar and soil applied, based on how they are absorbed by plants. To be effective, soil applied herbicides face multiple challenges in that factors such as soil texture (especially clay content), organic matter content, biological activity, pH, temperature and moisture content can all affect the herbicide’s effectiveness. Most of these interactions are well understood and for any given herbicide there are detailed guidelines for successful use (Zimdahl, 2018a).

Soil applied herbicides can also directly negatively affect soil biology including microorganisms, though it is generally accepted that the effects are temporary rather than permanent (Zimdahl, 2018b). Leaching of herbicides, into ground and surface water is also a risk. Leaching is regulated by the same soil properties that determine the herbicide’s effectiveness, e.g. clay and organic matter particles that adsorb a herbicide and reduce its efficacy will also reduce its leaching. A herbicide’s properties, e.g. solubility, also have a large impact with rainfall and irrigation being additional factors (Zimdahl, 2018b).

In-crop weed management

Integrated weed management

In-crop weed management has been dominated by herbicides since the 1950s, however, the international weed science community is clear that herbicide centric weed management systems face multiple challenges, such as evolved resistance to herbicides, and that changing to an integrated weed management (IWM) approach is essential for ongoing effective weed management (MacLaren et al., 2020). Like all integrated pest management systems, IWM monitors weed problems in crops and pasture, and only when action or economic thresholds have been reached, is a weed control activity undertaken. Those weed control activities can be broken down into four main approaches: physical, such as mechanical hoeing; chemical, i.e., herbicides; biological, e.g. biological control agents (BCAs); and ecological, e.g. rotations, and intercropping. When all four approaches are combined in a symbiotic and integrated fashion IWM is achieved (Fig. 4).

However, unlike insect and disease integrated management, where past, current and future pest occurrences are only weakly linked or not linked at all, weeds are strongly linked over time because of the weed seedbank for annual and biennial weeds, and because perennial weeds live for many years. Therefore, IWM needs to manage this temporal nature of weeds actively by instigating

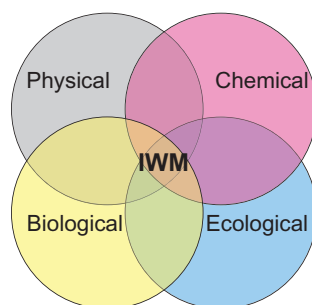


Fig. 4 Combining physical, chemical, biological and ecological weed management techniques to create integrated weed management (IWM).

proactive measures, for example, diversified and long rotational sequences, rather than relying on monitoring and then controlling living weed plants in crops and pasture. A further problem of threshold-based weed management is that some weeds are difficult to control once they have emerged and thus require the prophylactic use of soil-acting residual herbicides.

Herbicides and mechanical weeding

While the weed science community is clear on the need to move to IWM, change on the farm is being achieved more slowly. Mainstream farmers often still rely heavily on herbicides, and alternative farming systems, such as organic agriculture, which prohibit the use of synthetic herbicides, often relies heavily on mechanical approaches such as interrow hoeing. The latter prohibits the use of synthetic pesticides and herbicides and often relies heavily on mechanical approaches such as interrow hoeing. Both of these can achieve good weed control, such that the vast majority of the plant species and biomass in the field is the crop alone. As discussed at the start of this chapter, monoculture and bare ground are both deleterious for soil health and many other environmental outcomes, so, alternatives are required. This is particularly pertinent to herbicide resistance, as a move by mainstream farmers from herbicides to mechanical techniques is not necessarily beneficial for the soil or wider environment, resulting in reduced soil structure and organic matter, increased compaction, and greater use of fossil fuels. Therefore, biological and ecological approaches should be favored. Cover crops are one such example that also have the potential to improve soil health at the same time.

Cover crops

Cover crop is the general term to describe plants that are grown for purposes other than for immediate sale to generate profit. The terms cash crop and non-cash crop are also used to describe the two purposes. The reasons for growing cover crops are many; the key ones include building soil organic matter, improving soil health, fixing nitrogen, and reducing leaching, soil erosion, agrichemical runoff, and weed management.

Cover crops/non-cash crops are sub-divided into a diverse range of approaches. The key message regarding all the cover crop approaches is that they use biology/ecology rather than physical and chemical means to manage weeds. By increasing the number of plant species and biomass in crops and pasture, positive gains can be made for soil health; the opposite of what physics and chemistry typically achieve.

It should be noted that there are no agreed definitions of the various cover crop approaches. There is variation in use among the English speaking countries and even within them, with some terms being used interchangeably and the same term being used to describe different approaches.

Cover crops can be initially divided into two main approaches: first, those grown in the period between cash crops and second, those grown with the cash crops. In the first approach, there is no issue of competition between the cash and cover crops, though there will be lag effects due to changes in plant nutrients, moisture and residues. For the second approach there is the potential for significant competition between cash and cover crops which has to be carefully managed, but it is also possible to have cover crops that have a synergistic effect on the crop and therefore increase yield.

Green manures and catch crops for nutrient management

The term green manure is mostly used for cover crops that can fix atmospheric nitrogen into reactive nitrogen that can be used by living organisms. These are mostly legumes, which have a symbiotic relationship with diazotrophs that live in nodules on the roots. These are of particular value in organic agriculture which prohibits the use of synthetic and soluble nitrogen fertilizers, so green manures are an important route by which nitrogen is replaced.

A catch crop aims to catch or take up nutrients that could potentially leach out of a soil, principally nitrogen, and convert inorganic nutrients into organic forms which are more resistant to leaching and also provide nutrition for the soil biome, augment soil organic matter, improve soil structure, etc.

Green manures and catch crops are both grown in-between cash crops. In tropical regions this might be at the same time, whereas typically in temperate regions they are grown over winter when no cash crops are growing. Their benefit for weed management is that they can compete strongly with weeds, preventing them from proliferating and seeding. They can also provide habitats for seed predators such as beetles and rodents. Some of the plants are allelopathic, so can further inhibit weed plant growth and also the germination of weed seeds, and in addition their residues/ground cover can prevent the emergence of weeds.

Smother crops

Smother crops are similar to green manures and catch crops in that they are grown in-between cash crops, but, their main purpose is to smother weeds through strong competition from both roots and foliage. These are mainly used to manage particularly challenging weeds such as *Cirsium arvense* (L.) Scop. (Californian/Canadian thistle) and *Elymus repens* (L.) Gould (couch grass). These types of weeds are hard to kill as they have creeping roots or stems that can spread rapidly through the soil. They often have substantial organs of perennation which provide them with stores of energy and nutrients to regrow after damage, and most can also dedifferentiate and regrow even after having been damaged or broken up by tillage. Control of these weeds can be very challenging, even with herbicides, and where they fail the alternative is often extended fallow, sometimes with deep cultivation that is highly damaging to soil health.

The smother crop species need to be matched to specific weeds to ensure that they can compete effectively with the weed. For example, *C. arvense* has a creeping root system at a depth between 20 cm and 1 m, and when broken up by mechanical tillage all the pieces can dedifferentiate to produce new roots and shoots resulting in more plants than before. Few systemic herbicides can translocate sufficiently throughout the root system to kill it entirely. One smother crop option is an autumn sown mixture of *Secale cereale* L. (ryecorn/cereal rye) and *Vicia* species e.g., *Vicia sativa* L. (common vetch). This achieves sufficient biomass and height over winter when *C. arvense* is dormant, such that when the thistle emerges in spring it immediately experiences considerable competition for light, which increases further as the ryecorn can grow to a height of two meters. This is twice the typical height of *C. arvense*, with the vetch climbing up the ryecorn thereby blocking nearly all light to the thistle. In addition, ryecorn has a highly competitive, deep, fibrous root system which competes strongly with the thistle roots. *Elymus repens* has shallow creeping stems which can dedifferentiate at the nodes into both shoots and roots, such that it easily regrows after being broken up. *Trifolium pratense* L. (red clover) is an effective smother crop because it grows higher than *E. repens* and has thick dense foliage that blocks the light to the *E. repens* leaves. Red clover is a legume that can fix its own nitrogen giving it a competitive advantage if the soil is deficient in nitrogen over couch grass, which will be nitrogen deficient. Smother crops are therefore a particularly clear illustration between physical and chemical approaches that can harm the soil, compared with biological/ecological approaches which can have positive benefits for soil health.

Living mulch

Living mulches, also known as companion cropping, are where a non-crop plant is grown under the crop plants to cover the soil completely. A classic example is *Trifolium repens* L. (white clover) growing under *Zea mays* L. (maize). The clover's green leaves intercept and change the light spectrum reaching the soil which has a strong inhibitory effect on weed seed germination, and if weed seeds do germinate, they face strong competition from the white clover. In addition to being a highly effective means of controlling weeds, living mulches can have many other benefits such as supplying nitrogen to the crop, providing shelter, nectar, alternative prey and pollen (SNAP) to beneficial insects that can help control crop pests (Gurr et al., 2017). This is in addition to the benefits to soil compared to the bare ground created by herbicides and mechanical weeding. However, in dryer climates living mulches can compete for water with the crop, so the technique is not universal.

Relay cropping and undersowing

Relay cropping and undersowing are where the following crop (cash and non-cash) is sown into an established crop. Relay cropping, as in relay race, is where the following crop is established close to harvest of the preceding crop, often being drilled between current crop rows. Undersowing is where the following crop (cash and non-cash) is established soon after the preceding crop is sown and is often broadcast, for example, a pasture or a green manure. In terms of weed management, there is no tillage between the changes in crops so the germination impetus from tillage is lacking, nor is there a period where the soil is plant free; these both inhibit weed seed germination and establishment. For soil health there is a greater diversity of plant species during the overlap period, no tillage, and a continuous cover of plant foliage and living plant roots, i.e., there is no plant free gap between crops, all of which is beneficial for soil health.

All of these cover crop examples illustrate the potential of using biological and ecological techniques to manage weeds between and within cash crops, achieving excellent weed control while avoiding the damage that can be done to soil health by physical and chemical techniques.

System level techniques

As discussed in the IWM section, due to the temporal nature of weeds, system level techniques are also vital for effective weed control. Rotations are the prime example of a system level technique, together with plant mixtures/intercropping, and grazing management techniques in pastoral systems.

Rotations

Rotations are where different crop species, both annual crops and pasture, are grown in sequence over time. For example, two to three years of pasture, then *Triticum aestivum* L. (wheat), *Solanum tuberosum* L. (potatoes), *Allium cepa* L. (onions), *Avena sativa* L. (oats), for 1 year each and then returning to pasture. Rotations have multiple benefits for both production and the farm ecosystem, including weed management (Liebman and Dyck, 1993) and soil health (Merfield, 2019). By increasing the diversity of crops grown, no weed species has continual ideal conditions for growth and reproduction, particularly when there are both permanent crops such as pasture and short-term annuals, such as potatoes and onions. Most annual weeds cannot survive in pasture as they are preferentially eaten by stock and likewise the perennial weeds that prosper in pasture are killed by tillage and the crop renewal process involving the use of herbicides. Also varying planting between spring and autumn crops, e.g. winter and spring wheat and spring planted oats, changes the time of crop establishment and therefore which weed species emerge with the crop. For soil health, rotations increase the diversity of plant species through time, and in the pasture phase it is possible to have a highly diverse range of species. Pasture, being perennial, and with the fibrous roots of grasses, achieves far greater gains in soil health across all measures than any cropping system, even no-tillage, so its inclusion in a cropping rotation is exceptionally beneficial.

Crop and pasture mixtures - intercropping

As discussed at the start of this chapter, increasing the diversity of plants in space, i.e., growing a mixture of crop or pasture species together (intercropping) is beneficial for soil health. It can also improve crop production factors such as weed, insect pest and disease suppression, an overall increase in yield compared to the same area of land sown in a monoculture system i.e., the land equivalency ratio (LER) (Nyfeler et al., 2009; Wendling et al., 2017), as well as providing wider environmental and ecosystem benefits (Tamburini et al., 2020). The improved weed management in mixed species is principally due to the different species taking up different ecological spaces or niches leaving less ecological space for weeds.

As the number of plant species in a mixture increases, however, particularly if they are not closely related, e.g. a mixture of monocotyledons and dicotyledons, then the number of herbicides that can be used that will not kill one or more of the crop species decreases considerably. Selective herbicides are easiest to use in monocultures, which is partly why organic farmers, who are prohibited from using any synthetic herbicides, are more likely to use crop mixtures. However, the increased level of weed control that can be achieved with mixtures, potentially with a small amount of mechanical weeding around crop establishment, may give sufficient weed control so that herbicides are not required.

Grazing management

The traditional approach to grazing is 'set stocking' where livestock are spread out across all the available pasture. This results in stock preferentially eating the most palatable plants and leaving unpalatable, toxic and other harmful plants, which by definition in a pastoral farming system are weeds. The continuous grazing of palatable plants and non-grazing of weeds results in the weeds having a competitive advantage ecologically and they proliferate. Rotational grazing, 'rotates' /moves stock around the available pasture, with only small areas being made available to them at any one time. Once they have grazed the defined pasture area, they are moved to the next area, and prevented from returning to the area they have just grazed. This allows the pasture to recover from grazing and achieve a considerable percentage of its maximum biomass. This means the pasture is more competitive with the weeds because it is not being continually grazed and the weeds ungrazed. Also, many pasture weeds are simply unpalatable rather than being harmful, and in a rotational grazing system these are often inadvertently eaten by the stock together with the palatable plants; therefore, they also suffer a setback then with grazing. The intensity of rotational grazing varies considerably, with stock being left on a larger area of pasture for a period of up to a week, through to very small areas of pasture with the stock being moved several times a day. Well implemented rotational grazing is therefore a key weed management tool for pasture. Pasture can grow considerably more biomass and achieve greater rooting depth under rotational grazing and the stock are not continually trampling the soil as in set stocking system, therefore, long-term rotational grazing management can also improve soil health by increasing soil organic matter and structure and reducing compaction.

Conclusion

Weed management can have significant impacts on soil: both positive and negative. Many of the negative outcomes are associated with physical and chemical management techniques, while positive outcomes are often the result of biological and ecological techniques. Like much of farming, weed management and weed science are in a state of considerable change, even a paradigm shift. Just as chemical herbicides largely replaced mechanical weed control in the 1950s, with increasing herbicide resistance, a lack of new herbicide chemistry since the 1980s and withdrawal of older chemistry due to new regulatory requirements, many farmers are returning to mechanical techniques. The latter are often updated with sophisticated technology to the point that individual weeds, even at the cotyledon stage, can be identified by computer vision systems and each weed then eliminated individually, e.g. by high voltage electricity. However, replacing chemicals with mechanical weeding is not necessarily positive for soil health, and may be more harmful. What is required is an increased focus on and use of biological and ecological techniques such as cover cropping, intercropping, diversified rotations and rotational grazing, which can achieve excellent weed management while also being beneficial for soil.

References

- Baker CJ and Saxton KE (eds.) (2007) *No-tillage Seeding in Conservation Agriculture*, 2nd edn. Wallingford, UK: Food and Agriculture Organization of the United Nations.
- Bender SF, Wagg C, and van der Heijden MGA (2016) An underground revolution: Biodiversity and soil ecological engineering for agricultural sustainability. *Trends in Ecology & Evolution* 31: 440–452.
- Chacón JC and Gliessman SR (1982) Use of the “non-weed” concept in traditional tropical agroecosystems of south-eastern Mexico. *Agro-Ecosystems* 8: 1–11.
- Gómez R, Liebman M, and Munkvold G (2013) Weed seed decay in conventional and diversified cropping systems. *Weed Research* 54: 13–25.
- Gurr GM, Wratten SD, Landis DA, and You M (2017) Habitat management to suppress pest populations: Progress and prospects. *Annual Review of Entomology* 62: 91–109.
- Kennedy AC (1999) Soil microorganisms for weed management. In: Buhler DD (ed.) *Expanding the Context of Weed Management*, pp. 123–138. New York: Food Products Press.
- Liebman M and Dyck E (1993) Crop rotation and intercropping strategies for weed management. *Ecological Applications* 3: 92–122.
- MacLaren C, Storkey J, Menegat A, Metcalfe H, and Dehnen-Schmutz K (2020) An ecological future for weed science to sustain crop production and the environment. A review. *Agronomy for Sustainable Development* 40: 24.
- Merfield CN (2015) False and Stale Seedbeds: The most effective non-chemical weed management tools for cropping and pasture establishment. *The FFC Bulletin* 25.
- Merfield CN (2019) *Rotations and their Impact on Soil Health*. Lincoln: The BHU Future Farming Centre.
- Nyfelor D, Huguenin-Elie O, Suter M, et al. (2009) Strong mixture effects among four species in fertilized agricultural grassland led to persistent and consistent transgressive overyielding. *Journal of Applied Ecology* 46: 683–691.
- Rahman A, James TK, Grbavac N, and Mellsop J (1996) Spatial distribution of weed seedbank in maize cropping fields. In: O’Callaghan M (ed.) The 49th New Zealand Plant Protection Conference, pp. 291–295. Auckland, New Zealand: The New Zealand Plant Protection Society Inc.
- Rahman A, James TK, Mellsop J, and Grbavac N (2000) Effect of cultivation methods on weed seed distribution and seedling emergence. In: Zydenbos SM (ed.) 53rd New Zealand Plant Protection Conference, pp. 28–33. Auckland, New Zealand: New Zealand Plant Protection Society Inc.
- Rahman A, James TK, and Grbavac N (2001) Potential of weed seedbanks for managing weeds: A review of recent New Zealand research. *Weed Biology and Management* 1: 89–95.
- Sokol NW, Kuebbing SE, Karlsen-Ayala E, and Bradford MA (2019) Evidence for the primacy of living root inputs, not root or shoot litter, in forming soil organic carbon. *New Phytologist* 221: 233–246.
- Storkey J and Neve P (2018) What good is weed diversity? *Weed Research* 58: 239–243.
- Tamburini G, Bommarco R, Wanger TC, et al. (2020) Agricultural diversification promotes multiple ecosystem services without compromising yield. *Science Advances* 6: eaba1715.
- Weisser WW, Roscher C, Meyer ST, et al. (2017) Biodiversity effects on ecosystem functioning in a 15-year grassland experiment: Patterns, mechanisms, and open questions. *Basic and Applied Ecology* 23: 1–73.
- Wendling M, Büchi L, Amossé C, et al. (2017) Specific interactions leading to transgressive overyielding in cover crop mixtures. *Agriculture, Ecosystems & Environment* 241: 88–99.
- Zimdahl RL (2018a) Introduction to chemical weed control. In: *Fundamentals of Weed Science*, 5th edn, pp. 392–416. London, UK: Academic Press.
- Zimdahl RL (2018b) Herbicides and soil. In: *Fundamentals of Weed Science*, 5th edn, pp. 447–462. London, UK: Academic Press.

Relevant Websites

- <http://www.bhu.org.nz/future-farming-centre/>—The BHU Future Farming Centre.
- <http://the-jena-experiment.de/>—The Jena Experiment.
- <http://www.plant-teams.eu/>—DIVERSify: Designing Innovative plant teams for Ecosystem Resilience and agricultural Sustainability.
- <http://www.weedscience.org/>—International Herbicide-Resistant Weed Database.
- <http://wssa.net/>—Weed Science Society of America.
- <http://www.ewrs.org/>—European Weed Research Society.

Terraces as traditional agricultural landscapes. Stability amidst change

Gerardo Bocco, Centro de Investigaciones en Geografía Ambiental, Universidad Nacional Autónoma de México, Mexico City, Mexico

© 2023 Elsevier Ltd. All rights reserved.

Introduction	450
Basic definitions and typology	451
Traditional slope management and geographic distribution of agricultural terracing	452
Case studies	453
Terracing in Asia	453
Honghe Hani Rice Terraces (China)	453
Rice Terraces of the Philippine Cordilleras (Philippines)	453
Terracing in Latin America. The Andes and Central Mexico	454
Peruvian Andes	454
The Valley of the Colca River	454
Terraces in the Mexican Mixteca Alta UNESCO Geopark	454
Concluding remarks	454
References	455

Abstract

Agricultural landscapes, particularly in hilly to mountainous tropical terrains, have been constructed through millennia by local, peasant or indigenous societies on the basis of their traditional environmental knowledge. Agricultural terracing is an excellent example of an agricultural landscape. As ancient as agriculture itself, terracing has been practiced by rural communities in fragile or marginal environments in all continents, from North Atlantic subarctic islands to dry and wet tropical regions. The ancient techniques of terrace building, still used in many regions by small-holders, peasants managing subsistence agriculture, have been complemented or replaced by contemporary techniques employing mechanical equipment particularly for excavation in the frame of modern agriculture. In this chapter, emphasis will be on the first type of terracing, i.e., as a traditional agricultural landscape. Terraces have been largely devoted to subsistence agriculture, and were able to feed substantial numbers of people under various political and organizational systems. Regardless of the environmental characteristics and the level of social peasant organization, the purposes of any type of terrace are soil and water conservation and rain-fed or irrigated agriculture in marginal or fragile sloping terrains. In every instance, local, traditional slope management plays a key role. Abandonment has been also occurring in various places through history, probably accounting for the decline of terracing in many sites. Despite differences in the type of human-environment relationship, abandonment seems to follow economic crises in agriculture that in turn trigger migration.

... where flat lands suitable for agriculture do not naturally exist
man has been forced to create such surfaces artificially.

Spencer and Hale, 1961, p. 1.

Key points

- Agricultural terraces are discussed as cultural landscapes and heritage sites.
- Traditional environmental and slope knowledge is highlighted as a basis for terrace development.
- Focus is on terracing in tropical mountainous developing countries.
- Case studies are put forward to depict a balance between current farming in terraces and terrace abandonment.

Introduction

Agricultural landscapes, particularly in hilly to mountainous tropical terrains, have been constructed through millennia by local, peasant or indigenous societies on the basis of their traditional environmental knowledge. Agricultural landscapes are cultural landscapes involving either temporary or permanent interventions on slopes or valleys aimed at managing and improving local environmental conditions to develop usually complex agricultural systems. Sometimes referred to as rural landscapes, they have been highly regarded as efficient means to use and conserve natural resources particularly plants, soils and water. It has even been argued that cultural landscapes may provide more efficient access to resources than natural landscapes.

In many instances terraces have become part of the cultural heritage of human societies at the national or regional level. The United Nations Educational, Scientific and Cultural Organization (UNESCO) institutionalized the concept of world heritage, and in 1972 established the convention of the world's cultural and natural heritage. In this frame more than 1000 sites have been recognized since as natural, cultural or mixed heritage sites around the world. One site category is that of cultural landscapes, among which agricultural sites are prominent. Likewise, the UNESCO geopark network includes significant agricultural landscapes. In addition, since 2005 the FAO has designated 62 Globally Important Agricultural Heritage Systems (GIAHS) in 22 countries, most of them tropical, developing ones (see UNESCO and FAO websites).

Agricultural terracing is an excellent example of an agricultural landscape. As ancient as agriculture itself, terracing has been practiced by rural communities for millennia in fragile or marginal environments on all continents, from North Atlantic subarctic islands to dry and wet tropical regions (Fig. 1).

Terracing is considered a paramount human modification of sloping land for social adjustment to changing environmental conditions and to develop sustainable agriculture (Varotto et al., 2019). The practice has deserved scientific attention since the 1950s for both archeologic and environmental academic fields (for example, Spencer and Hale, 1961). The interest expanded during the last decades. In 2010, the International Terraced Landscapes Alliance (ITLA) was founded to revisit terraces as a cultural resource to meet current social demands: agriculture, food, tourism, among others (see ITLA web site, below).

While it is not clear whether there was one single center where terracing originated, several reports indicate that in China terracing is as old as 4500 years, in the Peruvian Andes 4400, in central Mexico 3500 and in Yemen, Jordan and Israel at least 3000 years old. Terracing traditions coupled to soil management seem to have been initiated separately from, but later on became integrated with, traditions of hydraulic management in some regions. The ancient techniques of terrace building, still used in many regions by small-holders and peasants managing soils and water for subsistence agriculture, have been complemented or replaced by contemporary techniques employing mechanical equipment particularly for excavation in the frame of modern agriculture. In this chapter, emphasis will be on the first type of terracing, i.e., as a traditional agricultural landscape.

Basic definitions and typology

An agricultural terrace is a portion of humanly created flat terrain, disposed step-like downslope or across a water channel. In this way both the gradient and the length of the slope or the channel are reduced. Overland flow, interrill and rill erosion of the soil are thus controlled, while soil sedimentation, infiltration capacity and residual humidity are stimulated. The purpose of terracing is to create a relatively flat surface for cultivation in otherwise poorly suitable strongly sloping terrain with shallow soils. Locally, terraces are referred to by different names in different languages. For instance, *bancas* and *andenes* are widespread terms for terraces in Spanish speaking countries.

A terrace comprises two components. On the one hand a stone, soil or shrub (or a combination of those) riser perpendicular to the main slope gradient or to the water flow direction in a channel. The base of the riser upslope is at the same level as the upper portion of the riser downslope. On the other hand, a second element is the flat surface immediately above the riser (Fig. 2). In this design, traditional builders expect that following storms sediment will be delivered downslope and the terrace will be filled up with fresh upslope surface materials. With time, the riser height must be increased to trap sediments efficiently downslope. In addition, a terrace may be accompanied by a canal to distribute incoming runoff and to allow for irrigation. The length and width of a terrace

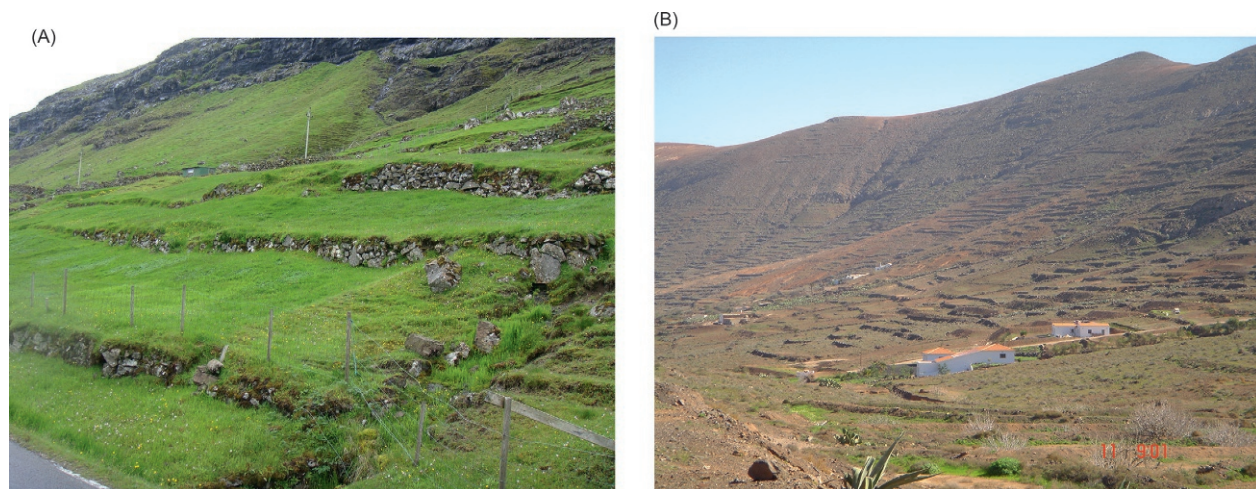


Fig. 1 Modern terracing in contrasting environments. (A and B) Grassy terraced land nearby Tórshavn, Faroe Islands; (C) Sloping terraces in Fuerteventura, Canary Islands. Photographs by the author.

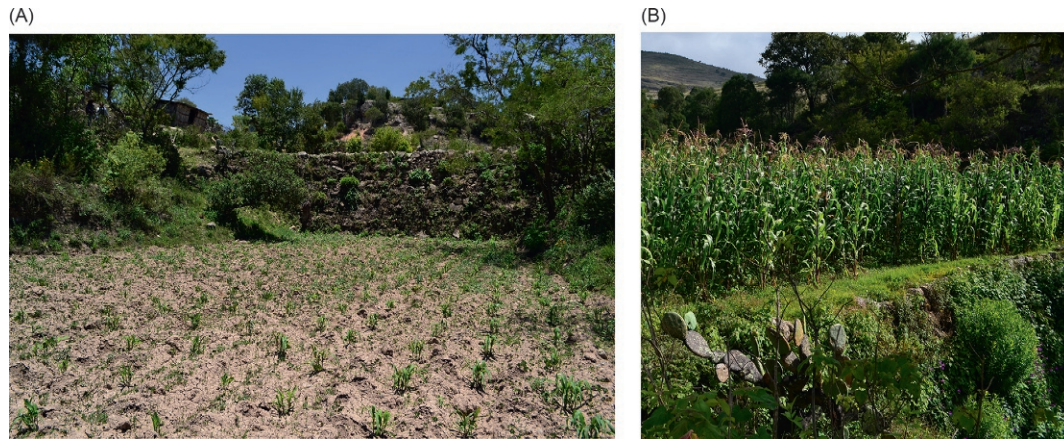


Fig. 2 Cross-channel terrace with stone riser, under rain-fed cajete maize (*Zea mays* spp.) at the onset of the rainy season. Mixteca Alta geopark. (A) at the onset of the rainy season. (B) at the end of the rainy season. Photographs by Quetzalcóatl Ramírez-Orozco.

will depend on the slope or channel gradient. Slope conditions will in turn control the process of digging and filling whenever feasible.

Terraces have been classified in several typologies according to form, dimensions, building materials and landscape position (among others, see [Spencer and Hale, 1961](#)). One simple, albeit comprehensive classification, encompasses three major types: cross-channel terrace to control ephemeral drainage and trap sediment delivered along water courses; contour or sloping terrace, ranging from a barely modified slope to a stone-built, excavated bench terrace, usually irrigated; valley-floor terrace on relatively flat terrain sometimes used for irrigation purposes.

Terrace types are based on designs that may be as elementary as simple cross-channel weirs, trapping incoming soil and water downslope. They may evolve through transitional forms up to more complex structures involving backslope digging and earth filling of the foreslope against the embankment if regolith is deep enough. In areas of shallow surface material filling may have been accomplished by hand, in baskets, or even removing top soil to allow the washing downslope of the subsoil material. Irrigation may also require complex canal design either to manage excess water in wet tropical areas, or to conserve limited water supply in dry regions. More complex designs require stronger social organization and engineering capabilities.

Regardless of the environmental characteristics and the level of social peasant organization, the purposes of any type of terrace are soil and water conservation and rain-fed or irrigated agriculture in marginal or fragile sloping terrains. In every instance, local, traditional slope management plays a key role.

Traditional slope management and geographic distribution of agricultural terracing

Traditional slope management is strongly based on traditional soils and landscape knowledge by farmers worldwide, but with a focus on hilly to mountainous tropical countries. This type of knowledge involves knowledge on landform, runoff, soil and farming practice; it is derived from long-term observations, coupled to religious beliefs and transmitted orally from generation to generation.

Slope management by farmers focuses on slope dynamics as governed by gravity; in fact it deals with the management of runoff and sedimentation. It is well documented that this involves the recognition of overland flow, soil detachment, transport and sedimentation by water, as well as soil quality for cropping along the *catena*, from summit to base level through slope, foot-slope and toe-slope positions. Terrace farms practice traditional cropping systems, either rain-fed or efficiently irrigated, involving limited use of external inputs and minimizing environmental impacts.

Recognition of the need to build terraces on sloping terrain and the benefits that could be obtained seem to underpin the decision-making process leading peasants to invest time, energy and resources to intervene on slopes and channels. Traditional knowledge is crucial in such processes, involving human-environment relationships at a local level but in disparate sites. Various local practices, which produced similar outcomes elsewhere, share common features and may be based on general principles behind local slope management.

This may explain why terracing has originated and developed in different parts of the world without any contact between social groups involved in the practice. The number of cases and the size of areas cropped either currently or in the past amount to very large figures. An inventory would be a titanic effort outside the scope of this chapter. Here, a short list of countries where terracing is or has been important is summarized by (sub)continent. Rice terracing (paddy fields) has been reported in scientific publications for mountains in Yemen, China, Japan, Nepal, Bhutan, India, Indonesia and the Philippines. In Africa, examples in Ethiopia, Tanzania and Rwanda have been documented. For Europe, cases are reported especially but not only for the Mediterranean region (particularly Spain and Italy). In Latin America, most examples are reported in the Andes (especially Ecuador, Peru and Bolivia) and Central America (especially Mexico and Guatemala). Other cases have been mentioned for Southwest United States.

Case studies

To depict the scope of terraced landscapes, a few case studies are briefly summarized below for Asia and Latin America. In both regions, terracing has been and still is a prominent agricultural landscape feature. The cases reviewed here emphasize prehistoric origin and the challenges to continuous use by local communities.

Terracing in Asia

Honghe Hani Rice Terraces (China)

The Honghe Hani Rice Terraces in the Ailao Mountains in Yunnan Province, Southwest China are a 1300 year old UNESCO World Heritage Cultural Landscape site and an FAO Globally Important Agricultural Heritage System. The Hani people have built and maintained an impressive paddy system encompassing 3000 terraces occupying 4700 ha, over 16,600 ha of total land, and transformed the valley of the Hong River in a living cultural landscape providing food and services to the rural population in three towns and 18 villages (Bai et al., 2013).

Traditional landscape knowledge coupled to strong religious beliefs and a solid social organization allowed the Hani to develop an integrated soil and water management system on very steep slopes, from the forests at the top of the valley side at nearly 2800 masl down to the river banks at approximately 250 masl (Li, 2017). The forests upslope operate as a sponge for the roughly 1350 mm yearly average seasonal rainfall. A complex canal system distributes the water through the rice terraces downslope to reach the river channel. Downslope water movement also involves organic fertilization (manure) for terrace cultivation of several varieties of rice and other staple crops (Yuan et al., 2014).

The Hani rice system is well known for the harmonious spatial integration of its four landscape components: forest, village, terrace and river. The system has been successful in achieving a long-term combination of agroecological and livestock food production, accounting for nearly 50% of the peasant income, together with soil and water conservation, and disaster mitigation, particularly drought and hazards from mass movement (Yuan et al., 2014).

Ritual landscape components, deities and beliefs, and other immaterial features are at the core of annual religious festivals celebrating water, terrace and forest. Beliefs and local expert knowledge contribute to traditional water management. The village water manager, elected by the villagers on the basis of expertise and social recognition, controls the allocation of water to the terraces. The distribution of water is carefully calculated by means of a traditional allocation system as a function of the size of the irrigation area.

The prominent landscape aesthetic value of forest, terraces and traditional dwellings has triggered increasing tourist activity. To be sustainable, for locals and visitors, development of careful land use planning of the terrace system is desirable and that follows appropriate guidelines to face various current challenges. Beliefs and traditions as well as local knowledge in land and water management tend to diminish, especially among younger generations following access to modern education and rural to urban migration. The red rice (*Oryza punctata*, Kotschy ex Steud.) system, the major crop, is sufficient to feed local villagers, but modern cropping techniques increasingly modify the traditional system. Coordination among national cultural heritage authorities, local leaders and villagers could help to provide a proper answer to challenges (Zhan and Zhang, 2015). Research on human-environment links from both traditional and modern perspectives could also serve this purpose.

Rice Terraces of the Philippine Cordilleras (Philippines)

This steep terraced cultural landscape is on Luzon Island, in the remote Philippine Cordillera mountains, covering a range of soil and climatic conditions able to support high population densities. Of great aesthetic beauty, the wet rice contour terrace system has been built and maintained by the Ifugao ethnic group, a minority indigenous community that has occupied the region for ca. 2000 years.¹ The Ifugao terrace system also is a UNESCO World Heritage Cultural Landscape site and an FAO Globally Important Agricultural Heritage System. It is considered an example of a large-scale irrigated system outside a centralized polity (Acabado et al., 2019).

As well as the Hani rice example, major attributes of the Philippine system are the spatial distribution of landscape components where traditional cropping knowledge, beliefs and rituals are combined: the upland forest, the paddy terraces, the traditional villages and the water management system allow for irrigation through a complex network of canals (Camacho et al., 2016). Although colonial and Christian domination affected the balance in the human-environment interaction, partly guided by the local nature-inspired deities, the Ifugao managed to maintain the integrity of the system.

The Ifugao system is considered an example of adaptation of local communities to the conditions of remote environments, however, it is vulnerable to social, economic and technological changes, and it was included temporarily in the list of UNESCO endangered sites. Out migration following economic constraints, limits the availability of human power to maintain the terrace system and comply with the yearly farming calendar. A Community-Based Land Use and Zoning Plan (CBLUZP) has been recognized as critical to ensure that the conditions of integrity are maintained.

¹There exists some controversy in the literature with respect to the age of the terracing.

Terracing in Latin America. The Andes and Central Mexico

Peruvian Andes

Terracing in the Andes is emblematic. Ancient and current terrace agriculture has taken place and is still ongoing on many thousands of terraces occupying extensive areas in several countries originally occupied by the Inca (Borsdorf and Stadel, 2015). Origins of Andean terracing can be dated to pre-Inca times, but major development occurred during the climax of the Inca empire. Many rich cultural corridors have been recognized as heritage sites where sloping and bench terracing, locally called *andenes*, coupled to complex canal irrigation systems are widespread.

The Sacred Valley of the Incas, considered the heart of the empire, is one such example. Located at a short distance from Cusco, the ancient capital of the Incas, the Valley embraces many archeological sites, including the famous Machu Picchu. The Valley was a core area for maize and other staple crops because of the development of extensive agricultural terracing and irrigation systems on steep slopes along altitudinal gradients. This allowed for a large production of diverse crops to satisfy the needs of such a highly cultured civilization. Similar terracing systems occur in many other Andean valleys, connected through a large road system and also recognized as a multinational heritage site by UNESCO (Borsdorf and Stadel, 2015).

The Valley of the Colca River

A large proportion of Peruvian terraced landscape on deep Andean valleys has been abandoned following environmental and social economic change through the centuries. A thoroughly studied example is that of the Colca River Valley in the Department of Arequipa, Southwestern Andes. This semi-arid valley encompassing a relief amplitude of 3000 m has been terraced for at least 4000 years (Denevan, 2001). Original terracing was developed by the Collagua pre-Inca aboriginal group, but the peak of agricultural activity followed Inca occupation of the valley during the 15th century, a century before the Spanish conquest. Seasonal rainfall and canal irrigation allows current terrace agriculture of maize, quinoa, broad beans, alfalfa, and barley. In addition, farmers graze cattle, sheep, llamas, and alpacas on the high plateau above the valley and on fallow sloping fields.

Historic terrace abandonment has been explained on demographic grounds, particularly the decline in indigenous population during the 16th century following Spanish domination. Lack of human power to maintain terraces and canals is at the core of the process. Coupled to climate change that has affected the amount and seasonality of rainfall as well as changes in economic and political agricultural policies, abandonment has proceeded through to the present time due to out migration. The major concern however relates to conditions for restoration of an originally wise agricultural system based on terraces in an otherwise fragile and marginal environment.

Terraces in the Mexican Mixteca Alta UNESCO Geopark

The UNESCO Global Geoparks are sites and landscapes of international geological significance combining conservation with sustainable development while involving local communities. There are 169 UNESCO Global Geoparks in 44 countries. Mixteca Alta, Oaxaca UNESCO Global Geopark is in Central-Southern Mexico, a subhumid mountain region with a long cultural tradition, and home of the Mixtec indigenous group. Many of the 35 geosites of this geopark are related to the intensive use of land for rain-fed farming purposes through terracing over millennia, a system that could feed a population as large as 50,000 inhabitants. Human presence in Mixteca Alta (circa 3400–3500 years BP) has been established by radiocarbon dating of soil organic carbon present in agriculture terraces known locally as *lamabordos*. Traditional agriculture based on terraces has been ongoing ever since.

Nowadays, population density and demographic growth are low. Because of lack of economic opportunities, emigration to urban areas is significant. Following migration, terrace maintenance has become increasingly difficult and abandonment is widespread in the area. However, traditional agriculture in the geopark can still feed the local population. Terracing has proved efficient in adjusting cropping to changing climatic conditions, particularly a decrease in rainfall and varying onset of the rainy season (Orozco-Ramírez et al., 2019). A variety of maize has been adapted to terrace conditions by the *cajete* farming system, a traditional practice that uses residual soil moisture (Figs. 2–4).

Concluding remarks

Traditional agricultural terracing is a prominent cultural landscape worldwide. Developments in disparate places suggest multiple origins in different environments and cultures, indicating a complex human-environment relationship through history. These systems have proved efficient in terms of sustainable agricultural practices, especially small scale rain fed farming developments in mountainous tropical developing countries.

Devoted to subsistence agriculture, terraces can feed large populations under various political and organizational systems. Abandonment has also been occurring in different places throughout history, probably accounting for the decline of terracing at most sites. Despite differences in the type of human-environment relationship, abandonment seems to follow economic crises in agriculture which in turn triggers migration. Thus, maintenance of terrace structures as well as the human power required to perform farming activities becomes impractical.

Interest in terracing and the chance to overcome the challenges posed by modernization is increasing. In 2010, the International Terraced Landscapes Alliance (ITLA) was founded to recover the memory of the terraces as landscapes of the culture of their inhabitants. Joint interest by academic groups and social organizations is expressed in the biannual celebration of international



Fig. 3 Farmers reinforcing a riser to reach proper height at Mixteca Alta geopark. Photograph by Quetzalcóatl Orozco-Ramírez.



Fig. 4 Traditional maize agriculture in the Mixteca Alta geopark terrace. Photograph by the author.

congresses that discuss the past, present and future of terrace landscapes. These efforts recognize the problems but also the potential of terrace farming when humans at large face crucial problems in the rural domain. Maintaining terraces is an important aspect of food security and human health.

References

- Acabado SB, Koller JM, Chin-Hsin L, Lauer AJ, Farahani A, Barretto-Tesoro G, Reyes MC, Martin JA, and Peterson JA (2019) The short history of the Ifugao rice terraces: A local response to the Spanish Conquest. *Journal of Field Archaeology* 44: 195–214. <https://doi.org/10.1080/00934690.2019.1574159>.
- Bai Y, Min Q, Liu M, Yuan Z, Xu Y, Cao Z, and Li J (2013) Resilience of the Hani Rice Terraces System to extreme drought. *Journal of Food, Agriculture and Environment* 11: 2376–2382.
- Borsdorf A and Stadel C (2015) The cultural development of the Andes. In: In Borsdorf A and Stadel C (eds.) *The Andes. Springer Geography* pp. 99–118. Switzerland: Springer International Publishing.
- Camacho LD, Gevaña DT, Carandang AP, and Camacho SC (2016) Indigenous knowledge and practices for the sustainable management of Ifugao forests in Cordillera, Philippines. *International Journal of Biodiversity Science, Ecosystem Services & Management* 12(1–2): 5–13. <https://doi.org/10.1080/21513732.2015.1124453>.
- Denevan WM (2001) Terrace and irrigation origins and abandonment in the Colca Valley. In: Denevan WM (ed.) *Cultivated Landscapes of Native Amazonia and the Andes*, pp. 185–211. Oxford: Oxford University Press.
- Li Y (2017) A brief analysis of the conservation and management of the Honghe Hani Rice Terraces in China. *Journal of World Heritage Studies* 6–10. Special Issue.

- Orozco-Ramírez Q, Bocco G, and Solís-Castillo B (2019) Cajete maize in the Mixteca Alta region of Oaxaca, Mexico: Adaptation, transformation, and permanence. *Agroecology and Sustainable Food Systems* 44: 1162–1184. <https://doi.org/10.1080/21683565.2019.1646374>.
- Spencer JE and Hale GA (1961) The Origin, nature, and Distribution of Agricultural terracing. *Pacific Viewpoint* 2: 1–40. <https://doi.org/10.1111/apv.21001>.
- Varotto M, Bonardi L, and Tarolli P (eds.) (2019) *World Terraced Landscapes: History, Environment, Quality of Life*. In: *Environmental History*, vol. 9. Switzerland: Springer. https://doi.org/10.1007/978-3-319-96815-5_8.
- Yuan Z, Lun F, He L, Cao Z, Min Q, Bai Y, Liu M, Cheng S, Li W, and Fuller AM (2014) Exploring the State of Retention of Traditional Ecological Knowledge (TEK) in a Hani Rice Terrace Village, Southwest China. *Sustainability* 6: 4497–4513.
- Zhan G and Zhang J (2015) Hani Rice Terraces of Honghe—The harmonious landscape of nature and humans. *Landscape Research* 40: 655–667. <https://doi.org/10.1080/01426397.2015.1060299>.

Relevant websites

UNESCO and FAO sites on Honghe Hani Rice and Ifugao Rice terraces, including field photographs

<https://whc.unesco.org/en/list/722/>.
<http://www.fao.org/giahs/giahsaroundtheworld/designated-sites/asia-and-the-pacific/ifugao-rice-terraces/en/>.
<http://www.fao.org/fao-stories/article/en/c/1168911/>.
<https://whc.unesco.org/en/list/1111/>.
<http://www.fao.org/giahs/giahsaroundtheworld/designated-sites/asia-and-the-pacific/hani-rice-terraces/detailed-information/es/>.
<https://medium.com/thoughts-on-world-heritage/living-cultural-landscapes-the-honghe-hani-rice-terraces-59156e406b7e>—UNESCO.

UNESCO sites on Peruvian terraces, Cusco and Main Andean Road—Qhapaq Ñan

<https://whc.unesco.org/en/list/273/>.
<https://whc.unesco.org/en/qhapaqnan/><https://whc.unesco.org/en/list/1459>.

UNESCO web site on Geoparks and Mexican Mixteca Alta Geopark

<https://en.unesco.org/global-geoparks>.

Web site of Terraced Landscape Organization

<http://terracedlandscapes2019.es/en/background/>.

Soil quality protection policies and plans to ensure sustainability

Ram Swaroop Meena^a and Sandeep Kumar^b, ^aDepartment of Agronomy, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi, India; ^bICAR-Indian Agricultural Research Institute, Karnal, India

© 2023 Elsevier Ltd. All rights reserved.

Introduction	457
Soil protection legislation and regulations: A global perspective	458
Programs and acts for agricultural waste management	462
Action plans for policymakers and planners	463
Establish links for policy implementation	466
Financial incentives-A need for supportive interventions	467
Private organizations for investment in soil quality and environmental protection	468
Case study: India	469
New policy initiative recommendations for soil quality protection in India	470
Conclusion	471
Acknowledgments	472
References	472

Abstract

Soil quality and environmental protection necessitate ground-level policy and legal frameworks, followed by strict and appropriate planning, execution, and regulatory reforms on a time-to-time basis, in accordance with the requirements and situations for soil management. Soil quality and protection of the environment have received international recognition in the policy arena. There is an urgent need to strengthen our knowledge and initiate action to protect soil quality as the soil is considered a living body. This chapter helps to provide a roadmap to feasible plans, and policies to protect soil quality by including them as part of land-use planning. Therefore, governments should encourage and incentivize farmers to enhance eco-friendly land management technologies and soil ecosystem services.

Key points

- Contributes to the protection, restoration, and enhancement of soil organic carbon for landowners.
- Land management technologies and soil ecosystem services should be encouraged and incentivized by governments.
- Enhancing soil quality through existing value chains, policy frameworks, and financial mechanisms.
- Environmental protection and soil quality must receive an international recognition.

Introduction

Soil is a living body that offers several ecosystem services for humans (Wall and Nielsen, 2012). It provides underlying support for the economy and society. Soil has been exploited continuously for hundreds of years. Since the green revolution in the 1960s, the exploitation of soil for food production has intensified and this is likely to increase further in the future with increasing population and food requirements. A global ecosystem loss of US\$ 6.3–10.6 trillion was estimated in 2017 due to soil degradation, which was equivalent to 10–17% of the *gross domestic product* (GDP) (Land and Soil Policy, 2019). Every year, erosion leads to the loss of 25–40 billion tonnes of fertile topsoil, which significantly reduces the soil's ability to store and maintain carbon, water, and nutrients. This ultimately causes loss of crop yield and food quality. An estimated 766 million tonnes of cereal production is lost every year due to soil erosion. Further, if action is not taken in time to improve soil quality, over 253 million tonnes of cereals are projected to be lost globally due to erosion and poor soil quality by 2050. As a result, an estimated 1.5 million km² of cropland would be removed from production, which is roughly equivalent to the intensively cultivated arable land of South Asia (Land and Soil Policy, 2019).

Presently, the United Nations (UN) defines the “Sustainable Development Goals (SDGs)” with well-defined indicators and targets leading to the overall sustainability of the Earth ecosystem in all respects or spheres of human society. Soil quality and environmental sustainability can be accomplished through the adoption of eco-friendly technologies and measures by the farming community and stakeholders that lead to harmony in resource use and ecosystem services.

Soil quality should be managed by adopting sustainable agricultural practices such as crop diversification, maintaining constant vegetative ground cover, adding crop waste, reducing the use of pesticides, avoiding tillage, etc. (Sherwood and Uphoff, 2000). The conservation and protection of soil may contribute directly to climate change resilience, and thus are critical for providing food,

water, and energy for humanity. Globally, several countries have initiated action plans to prevent the deterioration of soil quality and to enhance the sustainability of natural resource use.

On an international scale, various programs and schemes for preventing soil erosion and protecting soil quality have been implemented. Therefore, it requires appropriate land and soil policies with increased attention to the soil as a major sink for carbon accumulation and thus its direct contribution to climate change mitigation. The formulation and implementation of effective policies are required at the national level for restoring and maintaining soil quality by farmers who can be incentivized for positive environmental outcomes on the basis of soil tests. Simultaneously, there is a need to strengthen further the research, development and extension network which needs serious attention by the policymakers.

This chapter will help in implementing and promoting the policies, plans and thrust areas to restore and protect soil quality by including them in land-use planning. In this planning, many sectors will have to come together and contribute to meet common goals, such as food supply, water supply, and conservation of biodiversity. Therefore, the government should encourage and incentivize farmers, who are the main link in this context, to adopt eco-friendly land management technologies for healthy ecosystem services. Likewise, small and medium agribusinesses engaged in sustainable land management should also be supported by governments.

Soil protection legislation and regulations: A global perspective

The existing policies or legislation are specific to a particular aspect of soil and its functions. Therefore, they do not constitute a coherent soil protection policy because they do not apply to all soils and do not deal with all threats to soil quality. However, current policies reflect a wide dimension and address various issues and areas related to the sustainable management of soil resources.

Recently, there has been increasing awareness of the importance of sustainable soil management with the implementation of legislation and regulations for adopting improved soil management practices and limiting soil losses, globally. Following the International Year of Soils in 2015, the International Union of Soil Sciences (IUSS) declared the International Decade of Soils (2015–2024) to continue raising awareness about the significance of soil for life (IUSS, 2016). During this time, the IUSS Executive Committee and Council are playing a critical role in determining overarching goals and directions (Fig. 1) (IUSS, 2016).

The United Nations has included sustainable soil management (SSM) in the SDGs to set new global benchmarks for soil management. The SDGs such as SDG1 (no poverty), SDG2 (zero hunger), SDG3 (good health and well-being), SDG6 (access to clean water), SDG7 (affordable and clean energy), SDG11 (sustainable cities), SDG12 (responsible consumption and production), SDG13 (combating climate change), and SDG15 (life on land) draw attention to the need to protect soil quality and improve degraded soils (Fig. 2) (Policy Brief, 2019). In 2012, the UN General Assembly initiated the Global Soil Partnership (GSP) under the Food and Agriculture Organization (FAO) to promote soil protection, conservation, and productivity (Policy Brief, 2019). In 2016, the GSP and FAO published the Voluntary Guidelines for Soil Sustainable Management (VGSSM), an attempt to define Sustainable

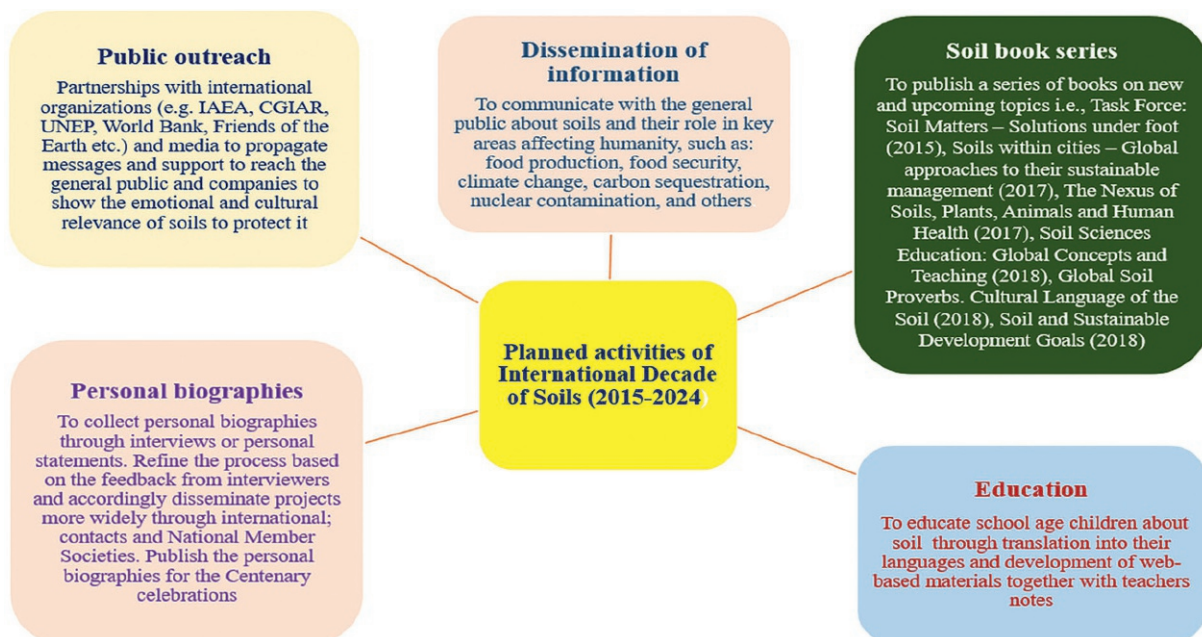


Fig. 1 Planned activities of the International Decade of Soils (2015–2024); CGIAR: Consultative Group on International Agricultural Research; IAEA: Consultative Group on International Agricultural Research; UNEP: United Nations Environment Programme.

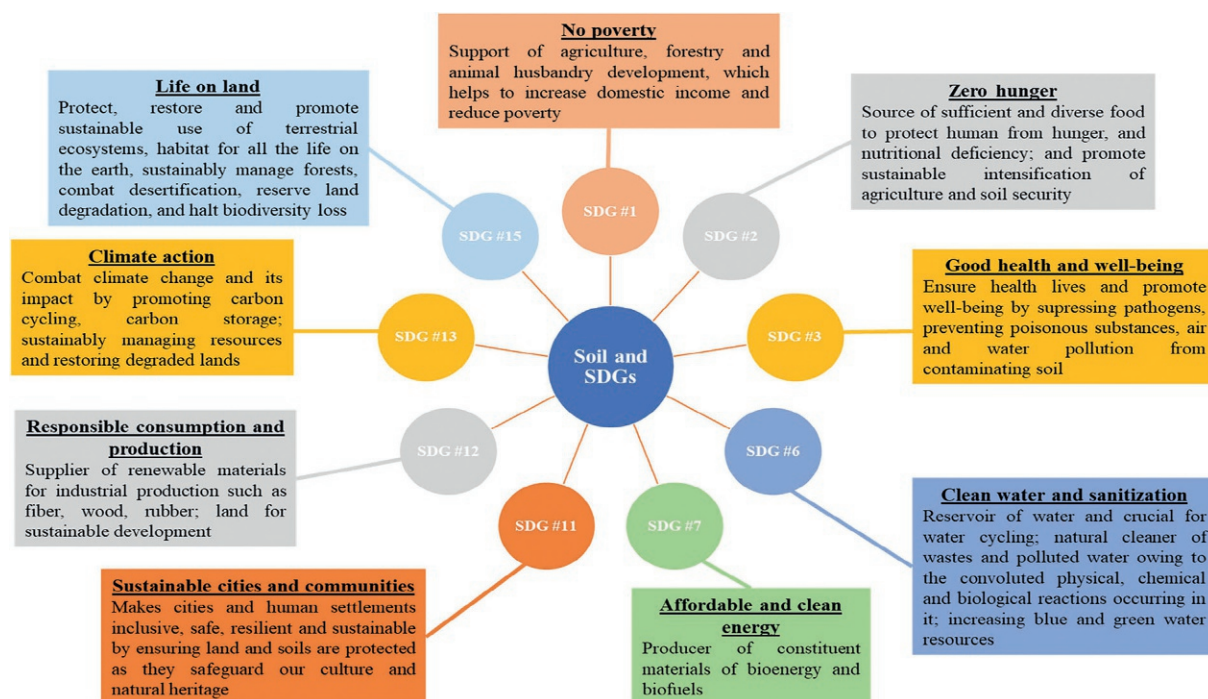


Fig. 2 Soil and Sustainable Development Goals (SDGs).

Soil Management (SSM) practices at the global level (Policy Brief, 2019). The FAO and its member countries are supporting the policies and legislation for improving global soil governance for the last decade (FAO and ITPS, 2015).

Soil quality is a global concern that has been mentioned in three international treaties, which contain relevant provisions on soil protection: UN Conventions, i.e., the Convention to Combat Desertification (UNCCD) of 1994, the Framework Convention on Climate Change (UNFCCC) of 1992 and the Convention on Biological Diversity (UNCBD) of 1992 (IUCN, 2018). The main purposes of the UNCCD are to combat desertification and mitigate the effects of drought. The UNCBD includes the preservation and sustainable use of biological diversity for soil and terrestrial ecosystems. The UNFCCC focuses on adaptation and mitigation actions, which include sinks and reservoirs of greenhouse gas emissions (GHGs).

As a result of the great efforts of international agencies and unions, in the last few years several laws, acts, regulations and legislations have been enforced and others are under consideration across several regions and countries. For example, the United Nation's soil Principles, EU Soil Policy for 2030 (Table 1), soil protection legislation and regulations of Greece (Table 2) and the United States (Tables 3–5).

Table 1 Soil protection laws of European Unions (European Commission Published a Policy Report, 2006)

Law/framework	Directive/regulation	Ways in which it helps to protect soil
Environmental Liability Directive	Directive 2004/35/EC	Sets most appropriate measures to remediate land damage
Industrial Emissions Directive	Directive 2010/75/EU	Ensures the operations to ensure soil quality
Environmental Impact Assessment Directive	Directive (85/337/EEC)	Identify, describe and assess the direct and indirect environmental impacts including land
Sewage Sludge Directive	Directive (86/278/EEC)	Encourage and regulate the use of sewage sludge in agriculture to prevent their harmful effects on soil quality
Regulation on fertilizers	Regulation (EU) 2019/1009	Regulates thresholds for contaminants, and presence in fertilizer products, particularly cadmium, to minimize soil pollution
Land use, land-use change and forestry Regulation	Regulation (EU) 2018/841	Sets a commitment for each Member State to ensure that GHG (greenhouse gas) emissions from land use are accounted for
Common Agriculture Policy (CAP) Implementation of soil management measures and associated obligations to advance soil protection in both agricultural and forestry systems	Regulation 1307/2013	and regulation 1306/2013
EU Biodiversity Strategy for 2030	–	Support the recovery of nature, limits soil sealing; restore degraded ecosystems, in particular, soil; protect soil fertility, and reduce soil erosion and the over-use of nutrients

Table 2 Regulations and legislation for soil protection in Greece.

<i>Laws/Action Plans</i>	<i>Theme</i>
Environment Protection Law (1986)	Soil erosion, conservation and salinization/sodification
Ministerial Decision No. 80568/4225/91 on methods, conditions and restrictions for the use of sludge from domestic and municipal waste-water treatment in agriculture (1991)	Soil conservation and pollution
Law No. 2508 on sustainable urban planning (1997)	Soil conservation and sealing
National action plan for combating desertification (2001)	Soil erosion, soil organic carbon loss, salinization/sodification
Law on the protection and management of water resources (2003)	Soil conservation
Ministerial Joint Decree No. 107017 on assessment of the effects of certain plans and programs on the environment in compliance with Directive 2001/42/EC of the European Parliament and the Council (2006)	Soil conservation
Presidential Decree No. 148 on environmental liability for prevention and remedying of environmental damage—Compliance with Directive 2004/35/CE of the European Parliament and the Council (2009)	Soil restoration and pollution
Joint ministerial decision on criteria for agricultural land quality (2010)	Soil conservation, quality and nutrient imbalance
Biodiversity conservation law (2011)	Soil erosion and conservation
Law No. 4014 on the environmental licensing of works and activities, and regulation of illegal constructions in connection with environmental stability (2011)	Soil conservation
Law No. 4042 on the protection of the environment through criminal law, on waste management and other provisions, in compliance with EU Directives 2008/99/EC and 2008/98/EC (2012)	Soil pollution

Data source: Schismenos et al., 2022.

Table 3 Soil health protection acts and programs enforced in the United States (USDA, 2022).

<i>Act/program</i>	<i>Action for soil protection</i>
Soil Conservation and Domestic Allotment Act (1936)	The government pays farmers to reduce production to conserve soil and prevent erosion
Soil and Water Resources Conservation Act (1977)	Providing broad strategic assessment and planning authority for soil conservation and natural resources management
Environmental Quality Incentives Program (EQIP)	Providing financial and technical assistance to agricultural producers to address natural resources and soil management
Conservation Stewardship Program (CSP)	Helps agricultural producers maintain and improve their existing conservation systems and adopt additional conservation activities to address priority resource concerns, particularly soil
Agricultural Conservation Easement Program (ACEP)	Helps landowners, land trusts, and other entities to protect, restore, and enhance wetlands, grasslands, and working farms and ranches through conservation easements
Healthy Forests Reserve Program (HFRP)	Helps landowners to restore, enhance and protect forestland resources through easements and financial assistance to enhance carbon sequestration and others

The United Nations Global Compact facilitated the development of “Soil Principles” for sustainable soil management. Soil principles aim to increase the understanding of key actors to protect soils through long-term programs, and to strengthen existing policies for soil conservation:

1. Protect soil from physical, chemical and biological degradation, limit erosion and avoid deforestation.
2. Restore soils on degraded, stranded and marginal lands.
3. Maintain soil-based ecosystem services, water availability and quality.
4. Enhance soil productivity according to its natural capacity.
5. Develop extension services, and knowledge systems, and promote innovation.
6. Communicate the importance of soil.

These principles are the framework for collaboration between the United Nations, governments, companies, civil society and other stakeholders. They offer a framework of outcomes and actions for companies to align with the SDGs and Agenda 2030. To support these principles, the UN Global Compact encourages companies to commit to global soil quality, share and promote better soil management practices, including soil quality in strategic planning, and work in partnership with stakeholders to ensure healthy

Table 4 State-wise legislation in the United States (Soil Health Institute, 2022).

Bill	Bill No.	State
California's Healthy Soils Initiative (HSI) (Agricultural Lands: Greenhouse Gases: Healthy Soils Program)	SB.1350	California
Agriculture Climate Adaptation Tools Program	AB.409	California
Environmental Farming Incentive Program	SB.253	California
An Act Establishing A Regenerative Agriculture Program	HB.06647	Connecticut
HB 1578 Relating to Climate Change	HB.1578	Hawaii
Farmer Apprentice Mentoring Program; Whole Farm System; Appropriation	HB.1505, SB.1544	Hawaii
Property Tax Exemption for Agricultural Land planted in Cover Crops	HSB.78, SSB.1123	Iowa
An Act for a State-wide Soil resource health and monitoring system	HF.102	Iowa
A bill for an act creating an agricultural land, soils, water quality, and farm tenure initiative committee	HF.103	Iowa
An Act requiring water quality plans for persons receiving financial assistance for establishing soil and water conservation plans	HB.406	Iowa
Amends the Soil and Water Conservation Districts Act	HB.2737, SB.1980	Illinois
Amends the Department of Natural Resources Act	HB.2819	Illinois
For legislation relative to storm water management and climate change adaptation incentives. Revenue.	SB.1663	Massachusetts
An Act to promote healthy soils and agricultural innovation within the Commonwealth	H.873, S.438	Massachusetts
An Act promoting healthy soils for reducing greenhouse gases and the effects of climate change in the Commonwealth	S.2402	Massachusetts
Maryland Healthy Soils Program	HB.1063	Maryland
Agriculture—Cost-Sharing Program—Fixed Natural Filter Practices	HB.687	Maryland
Agriculture—Maryland Healthy Soils Grant Program	HB.1176	Maryland
A resolution to declare April 28-May 5, 2019, as Soil and Water Stewardship Week in the state of Michigan.	HR.0084	Michigan
Working lands drinking water protection pilot program and appropriation	SB.1637	Minnesota
Drinking water protection pilot program established, and money appropriated.	HF.1569	Minnesota
Board of Water and Soil Resources Soil Health state-wide action plan funding provided and money appropriated	SF.1418, HF.1609	Minnesota
Create the Healthy Soils Task Force and add a Use for a Fund	LB.243	Nebraska
Adopt the Soil Health and Productivity Incentive Act	LB. 729	Nebraska
Provide for a climate change study	LB.283	Nebraska
Healthy Soil Act	HB. 204	New Mexico
Soil & Water Conservation District Day	SM.75	New Mexico
Carbon Farming Act	SB.4875	New Mexico
Green New Deal	A05334-A	New York
Oklahoma Carbon Sequestration Enhancement Act	HB.1192	Oklahoma
Relating to greenhouse gas emissions; declaring an emergency.	HB.2020	Oregon
Relating to greenhouse gas emissions; prescribing an effective date.	SB1507 A	Oregon

Table 5 Soil protection legislation introduced recently in the Senate and House of the United State (<https://www.congress.gov/bill>).

Legislation	Actions for soil protections
Healthy Soils Healthy Climate Act (2021) (amendment to the Food Security Act of 1985)	Create permanent payments within the environmental quality incentives program for soil health practices and carbon sequestration
Healthy Soil, Resilient Farmers Act (2020)	Facilitating soil health transition loan and soil health transition plan
Study on Improving Lands Act (2021)	Direct a study on the state of soil health on federal lands that determine impacts such as grazing, wildfire, recreation, and invasive species
RESTORE Act (2021)	selection of forest landscape to conduct a forest landscape project on specified management activities, including restoring soil health
Farmers Fighting Climate Change Act (2021)	This bill expands the Conservation Stewardship Program of the USDA to provide for the establishment of climate change mitigation and reduce GHGs emissions
Future of Agricultural Resiliency And Modernization Act (2021)	Award grants to carry out projects that reduce GHG emissions, improve soil health, and increase carbon sequestration
Naturally Offsetting Emissions by Managing and Implementing Tillage Strategies Act (2021) (amendment to the Food Security Act of 1985)	To optimize the sequestration of carbon and the reduction of net GHGs emissions in agriculture systems
Agriculture Resilience Act (2021)	To address the impact of climate change on agriculture, and for other purposes including soil quality
Climate Stewardship Act (2021)	To provide incentives for agricultural producers to carry out climate stewardship practices, to provide for increased reforestation across the US
BIOCHAR Act (2021)	To establish a biochar demonstration project and biochar grant program to promote soil health

soils. When viewed in the wider context of Agenda 2030, companies that support the Soil Principles will catalyze investment and action in support of the world's soils and contribute to the implementation of SDGs.

The EU's soil policy for 2030 establishes a framework for protecting soils and ensuring their sustainable use. To accomplish this vision, three major objectives were set: (i) all EU soil ecosystems should be healthy and resilient and continue to provide core services; (ii) there is no net land taken and soil pollution is reduced so that it is no longer harmful to people's health or ecosystems; and (iii) soil protection, restoration, and sustainable management are common practices. To attain these targets, the strategy emphasizes some of the key targets:

- Submitting a separate legislative proposal on soil health by 2023 to support the goals of the EU soil strategy to achieve healthy soil by 2050.
- Making sustainable soil management the new standard by providing a plan for farmers to have free soil testing, and encouraging sustainable soil management through best management practices and common agricultural policy (CAP).
- Reclaiming eroded soils and cleaning up contaminated areas.
- Preventing desertification through the development of a uniform approach for assessing it and land degradation.
- Consider establishing legally enforceable targets to prevent the draining of organic soils and wetlands and restore drained and managed peat lands to adapt and militate against climate change.

At the EU level, there was no binding overarching framework that strategically defined policy priorities or parameters for soil protection. Soil protection outcomes are mostly derived from other laws delivering environmental objectives and are not explicitly soil-focused, as follows.

Programs and acts for agricultural waste management

The continuous growth in the world's population has raised the demand for food and increased agricultural production, which results in the generation of tons of agricultural waste at a global level (Sabiiti, 2011). Agricultural waste includes materials generated by various agricultural activities including crop residues (leaves, residual stalks, husks straw, roots, shells, etc.), manures, wastes from slaughterhouses and poultry houses, runoff, and leaching of fertilizers, pesticides and silt drained from fields, and processing residues including husks and domestic waste (e.g., skins of fruits) (Fig. 3).

The improper disposal of agricultural waste has adverse effects on the environment, therefore, its recycling and reuse can help in protecting the environment (He, 2019). The proper management of crop residue can generate bioresources such as organic fertilizers for farming communities which helps to boost soil quality. These organic wastes are rich in organic carbon, nitrogen,



Fig. 3 Agricultural waste.

phosphorus, potassium, and other beneficial micronutrients that improve soil structure and fertility to a certain extent. In addition, it encourages efficient energy use of agricultural wastes. These waste products can be used for biocomposting, which can provide organic manure to protect soil from deterioration. With proper planning, public awareness, and implementation of government policies and laws, agricultural waste can be managed and used properly to generate beneficial bioresources (Sabiiti, 2011). Policies are needed to raise awareness of the public and farmers to these benefits. In addition, this also calls for the setting up organizations or institutions to harness agricultural wastes as a potential resource in farming and energy production (Sabiiti, 2011). In recent years, the management of agricultural wastes has become an issue of concern for policymakers to ensure ecological agriculture and sustainable development. There is a need to focus on policies and regulations to recycle agricultural solid waste efficiently into useful products to generate new income and employment opportunities (Wang et al., 2016). This will require better use of technology and incentives, a change in philosophy and attitudes, and better approaches to agricultural waste management. Government initiatives such as subsidies, encourage agriculture-linked enterprises and cultivate service entities, etc. that can save resources, develop the economy, and protect the environment by increasing and improving the recycling of agricultural wastes (Wang et al., 2016).

The benefits of crop residue management and the adverse effect of residue burning on soil quality need to be understood by farmers through training to ensure soil quality. In South Asian countries, governments should implement effective strategies for in-situ crop management at the time of rice harvest to use the waste (Meena et al., 2021). This may include initiatives to create awareness, support for machinery required, scientific guidance, and the strict enforcement of rules and regulations by local administrations (Fig. 4). In addition, nexus thinking in environmental management is an emerging concept that can be beneficial to governments and other sectors of society. The concept of nexus thinking promotes stakeholders' involvement and integration at a higher level, going beyond disciplinary boundaries, which could provide a benchmark for resolving issues like residue burning. Efforts are required at all levels, including basic and strategic research, technology dissemination, strengthening stakeholder links, financial support, policy formulation, and implementation.

Action plans for policymakers and planners

The task for framing effective soil policies toward sustainability is difficult to formulate and implement. It requires assessment of soil quality indicators and properties, soil functions, land capability classification, land use, etc. to develop policies for managing the soil ecosystem (Fig. 5). Further, technologies such as low input agriculture, organic farming, and precision farming would benefit soil quality by optimizing the use of fertilizer and nutrient cycling within the agroecosystem to maintain soil quality. From a policy perspective, optimum fertilizer use should be given major emphasis to promote soil quality. Management schemes and plans should be implemented at the local, regional, and national levels, depending on soil and local conditions. Above all, the main

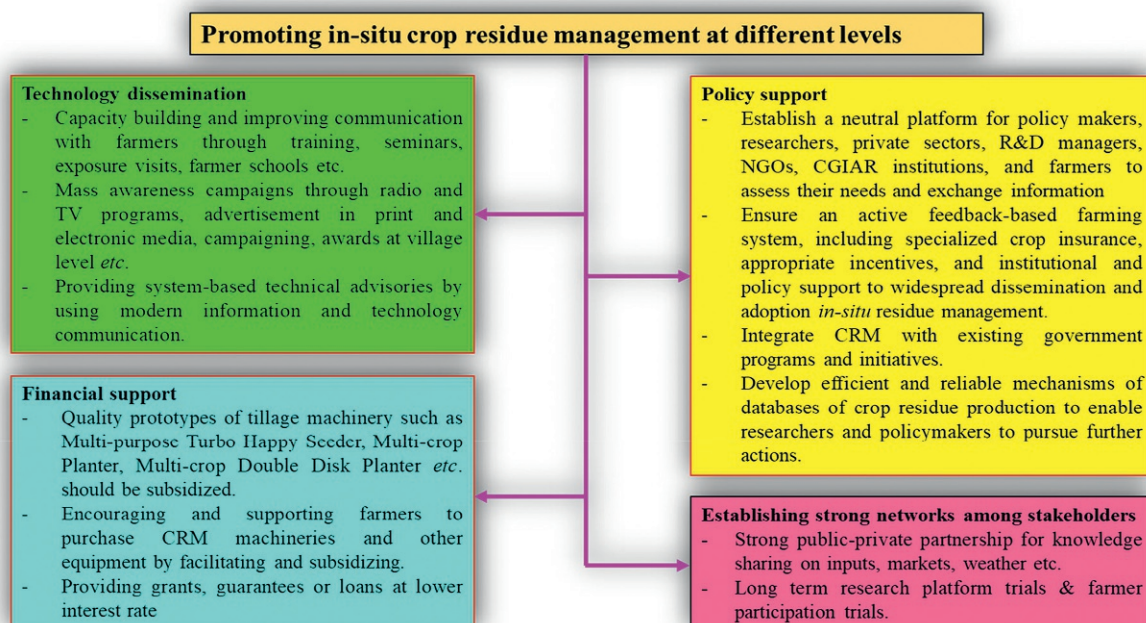


Fig. 4 Plans and policies for in-situ crop residue management (CRM). CGIAR: Consultative Group on International Agricultural Research; NGOs: Non-governmental organizations.

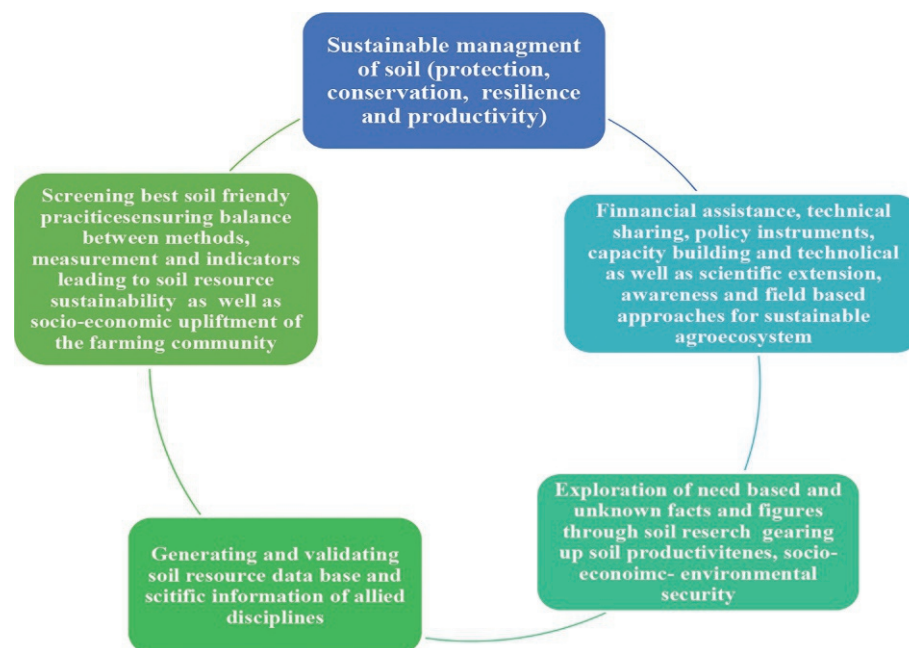


Fig. 5 Action plans for policymakers.

action plans should include the reduction of land loss and restoration of degraded land, should promote reasonable access to land and secure land tenure, and enhance capacity-building, technology transfer and financial support by various methods (United Nations, 2009).

i) Reduce land losses and restore degraded land

- Find interrelationships between climate change, land degradation, and desertification, including the need to promote integrated land, and water management to address various issues;
- Support sustainable land management by increasing soil and water productivity, and improved security of tenure for people living in vulnerable areas;
- Protect land resources sustainably using scientific and indigenous knowledge;
- Encourage the improvement and restoration of degraded lands by integrating pastoral and agricultural land use, as well as best management practices;
- Address the risk of erosion and land losses caused by sea-level rise, particularly in low-lying coastal states and areas, by planning land-use access and adaptation programs for climate change;
- Control soil erosion and enhance soil water holding capacities through sustainable forest and agroforestry management;
- Promote national action plans and international cooperation to prevent the spread of sand dunes, and reduce the severity and frequency of sandstorms by strengthening early warning systems, restoring vegetation, and supporting rural community initiatives in affected areas, primarily in developing countries;
- Encourage appropriate well established practices and indigenous knowledge concerning land use, water management, and farming activities;
- Converge sectorial policies and programs at the national level to stop and reverse land degradation and create coordination through convergence.
- Improve cooperation and coordination between the authorities that manage land and water resources;
- Strong support for policies aimed at reducing land degradation while also creating jobs and contributing to the eradication or elimination of poverty in developing countries;
- Protect and conserve soil and land resources through land use, spatial planning, and sustainable agricultural practices such as agroforestry, eco-farming, diversification, reduced tillage, perennial vegetation-based land cover, and enhancing capacity-building for sustainable development of peoples;
- Develop land management policies that address the direct and indirect causes of land degradation, such as erosion, salinization, alkalization, desertification, pollution, and loss of SOM.
- Adopt policies that will restore the soil's physical integrity, improve the nutrient status, and increase SOM content;
- Strengthening capacity, especially in developing countries, for monitoring soil quality and land degradation, including looking for and evaluating the information on biophysical and socio-economic perspectives of land degradation.

- Ensure a proper balance between social, environmental, and economic aspects of sustainability, livelihoods, and food productivity in land management and policies;
 - Encourage inventory, assessing, monitoring, and understanding of the potential of land to support ecosystem functions, including creating a global soil map with the help of local and indigenous knowledge systems, and accordingly, include long-term strategies in land-use planning for soil management;
 - Empower local authorities to execute sustainable land policies consistent with local conditions and national priorities;
 - Build partnerships and networks among the stakeholders in land-use planning and management to gain a better understanding of land-related issues by considering sustainable goals and indigenous knowledge to enable land development and identify potential areas for conservation;
 - Provide financial assistance to encourage private and public investments in research and development for sustainable land management, combat land degradation and ensure land managers' access to credit;
 - Encourage initial start-up at the small scale for resource-constrained smallholders to acquire and adopt sustainable land management practices and technologies;
 - Establish a new center of excellence and renovate the existing centers as part of land policy, tenure, and land management;
 - Establish the set of norms and standards for a proper review, monitoring, and evaluation of land policies and their indicators.
- ii) Promote reasonable access to land and secure land tenure
- Promote soil and water conservation by ensuring land tenure security for rural land users, particularly women, indigenous people, and other vulnerable groups;
 - Establish cost-effective and efficient land administration systems, including transparent tenure and registration systems, that may contribute to investment and best land management practices, using the latest information technologies;
 - Recognize different laws and or systems that govern land access and tenure among states, improve institutional and legal frameworks for the administration of property rights and tenure systems considering smallholder farmers, established land tenure, rural poor, customary tenure, and local practices; and adopt laws and policies that promote equitable and secure access to land and tenure for the poor, especially women, indigenous people, and other marginalized groups;
 - Establish a mechanism for the dissemination of additional knowledge and information on land rights and implement civic education campaigns to educate the public;
 - Recognize rights for other uses such as grazing and wood gathering, which often contribute to women's livelihoods;
 - Allow and promote a more equitable role for women in land-use planning and management decision-making;
 - Eliminate the obstacles to the full understanding of the rights of the people living under colonial and foreign occupations, which are incompatible with the dignity and honor of a human being and therefore need to be removed.
- iii) Enhance capacity-building, technology transfer and financial support

Capacity-building, training, and development of skills for soil quality management should be integrated with the advancement of new techniques and tools. Proper research, development, and extension activities enhance the various governmental schemes, plans, and policies among farming communities. Training and extension activities need to be carried out at the grassroots level among farmers in terms of generating awareness about the soil resources together with positive outcomes of adopting various eco-friendly technologies. Government policies must be aligned with farmers' varying priorities and needs (Katyal et al., 2016). Further, the improvement and establishment of a knowledge base on soil technology and sustainable agroecosystems are linked with innovation, research, and development by involving interdisciplinary approaches. This would lead to the identification and screening of key environmental datasets that underpin the ecological and environmental perspectives and their sustainable management for human civilization (Kumar et al., 2015). The key action plans for capacity-building and enhancing technology adoption among farmers are given below:

- Strengthening the scientific base and advancing research on drought and desertification for making informed decisions on land management, and measuring and monitoring the impact on programs to control the deterioration of soil quality;
- Develop appropriate technologies for assessing, analyzing, and quantifying the consequences, impacts, and remedial options of deteriorated soil quality and desertification.
- Support for developing nations based on mutual terms through sharing and scaling up of advanced science-based practices and technologies, including best management practices and practical experience in combating land degradation and desertification. For example, sustainable agriculture practices and vegetation conservation or rehabilitation can serve advanced science-based practice together with management.
- Strengthening investment in land management to encompass sustainable forest management and land-use planning to improve soil quality and restore degraded land in arid, semi-arid, and semi-humid environments.
- Encourage and invite developed nations and the Global Environment Facility Council to provide adequate, consistent, and timely monetary support to areas affected by land degradation;
- Provide financial assistance to implement reform of land tenure in developing nations and countries in transition to promote sustainable land use, efficient governance, comprehensive planning, and equitable access to land;
- Encourage countries, particularly developing countries, to improve their scientific understanding of land resource systems by strengthening their technological capacities, including, as appropriate, supporting pilot projects to test their research findings;
- Enhancing human resource capacity through educating and training people for effective land management, particularly in developing nations.

Even after policy formulation, coordination among stakeholders needs to be improved to achieve efficient policy implementation and outcomes. Further, uptake could be achieved through need-based continued advice, information support, monitoring, and unpaid partnership. So, policies should encompass effective regulation and partnerships since land is usually privately owned and cannot be protected by the government alone. For that, farmers, land managers, local bodies, bureaucrats, policymakers, developers, researchers, and construction companies must work together under an umbrella to address diverse issues to promote healthy soils. Technology demonstration for awareness and popularization of soil quality management practices through public-public, public-private, and public-private-producer and private-private partnerships assures the successful implementation of processes at their level (Fig. 6).

A soil framework must be established for future policy development as well as to assist in monitoring and evaluating soil plans in order to strengthen the partnership with government, private agencies, and stakeholders. In the process of policy development and evaluation of its effectiveness, effective monitoring provides critical evidence on the state and changes of our soils (DEFRA, 2009). This will facilitate a more effective exchange of information, in particular for the advancement of soil research. A Soil Framework has to be developed in conjunction with advice from key stakeholders, keeping in mind that soil is a vital part of our economy, environment, and heritage, which needs to be protected for current and future generations. It should aim to promote the sustainable management and protection of soils through effective monitoring and evaluation of soil plans that are consistent with the economic, social, and environmental needs of the country. The following 13 activities are identified as being vital parts of this framework (The Scottish Government, 2009):

[illegible]

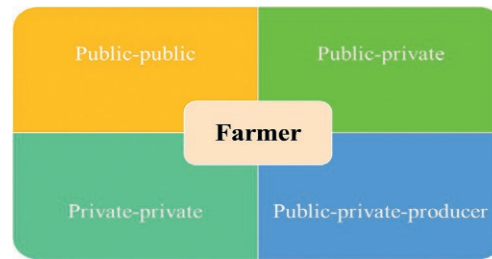


Fig. 6 Farmer-centric policy implementation.

Financial incentives-A need for supportive interventions

All benefits must be monetized to determine whether or not sustainable farming practices will deliver an appropriate return on investment at the farm level. A higher net income in farmers' fields can be achieved by improved marketing of farm products, reduction in the cost of cultivation, direct payments, subsidies, and insurance (Klauser and Negra, 2020). A combination of all these incentives can reduce barriers to the implementation of improved land management practices and facilitate farmers' adoption of them. Also, to remove barriers to the adoption of soil health-promoting practices, agricultural developmental activities need to be carried out together with policy and financial interventions adapted to specific conditions. While designing agricultural developmental activities, it should also be considered that there are still a significant number of farmers who do not have adequate resources, inputs, risk tolerance capacity, knowledge, and assets needed for the adoption of soil health-promoting activities. To increase the rate of adoption of these soil-health promoting activities among smallholders, these practices should be designed considering the local and regional conditions, agronomic variables, and policy framework context.

When soil health-promoting activities enable yield improvement or reduce production costs, there is a strong incentive for their adoption, and only help to facilitate them will be required. In the case of delayed agronomic benefits, smallholder farmers will not be sufficiently encouraged to adopt new farming practices until additional support is provided to offset new costs and the risks involved. A temporary or permanent subsidy (i.e., carbon offset payments) may be required for smallholders to adopt soil health-promoting practices when their environmental benefits extend beyond the farm. A farmer's income can be increased through increased yield owing to the adoption of improved production practices and the accrual of environmental benefits in the form of finance, subsidies, premium markets for agricultural goods, or payments for ecosystem services. Current policies and financial systems do not provide enough financial incentives to smallholders for implementing better soil management practices, therefore, fiscal support is needed in many cases.

Well-designed subsidies, land registration, and effective crop varieties are some of the fundamental elements of an enabling policy environment (WBCSD, 2018). For example, government subsidies or philanthropic funding can create multiple types of support for smallholders, such as soil amendments, customized fertilizers, and crop varieties in rotations that improve soil health. Governmental subsidies can include financial support for technical support and setting up marketing infrastructures such as storehouses, credit registrations, and price information (McCann et al., 2018).

There are several well established and new financial plans and strategies for smallholder production systems to incentivize farm-level contributions to climate change mitigation. Smallholder production has not yet demonstrated the feasibility of some newer approaches for achieving environmental security. An appropriately designed credit, loan, or insurance scheme that bridges the gap to an immediate return on investment could enable smallholders to follow soil-health promoting practices. Value chain finance could be provided to smallholders such as insurance, receivables financing, and pre-financing (Havemann, 2016). Farm credit targeted to smallholders often requires novel risk screening and collateral requirements allowing appropriate interest rates without credit histories or assets.

Soil health-related interventions are crucial for developing well-conceived fiscal structures/strategies and other specific arrangements because their success relies on agronomic yield and environmental benefits. Debt-based financing policies could increase the farm-level risk if they are not harmonized with the farmer's capability to repay loans. It is therefore important to follow debt instruments with approaches that enhance market value (such as higher market prices, Payment for Ecosystem Services—PES) or reduce their net costs (such as access to affordable, high yielding seeds). Several schemes for carbon credit and PES have been designed and tested to enhance the return on investment of improved adoption of management practice (Henderson et al., 2018). A program on PES can help land managers earn additional income generated by the adoption of sustainable land management practices and technologies. To ensure the success of the program, a strong network of local and national institutions is needed to ensure land tenure and an effective market-based payment mechanism. For instance, the Forestry Act (1996) of Costa Rica allows land users to be compensated for ecosystem services provided off-farm through forest conservation certification. As a result, deforested lands have been successfully restored through the effective implementation of natural resource management practices.

For PES to work, there must be buyers willing to pay for environmental services (for example, development finance institutions, businesses emitting GHGs, and water supply activities), and it depends on the establishment of low-cost monitoring and transaction instruments. Farmers can receive direct monetary support (usually through an aggregating entity) to adopt specific practices for a set period of time (i.e., payments on a unit area basis) or payments for environmental performance (i.e., carbon

sequestration). In addition, the differences in methods used as well as the high cost of assessing environmental performance at the farm level (such as carbon stocks and water quality) have hindered PES approaches. Prices for agricultural carbon credits have been low and volatile as a result of inconsistencies in soil carbon sequestration and benefits accruing from new practices, and poor agreements about the eligibility and permanence for compliance schemes (Meena et al., 2022).

The majority of smallholders in developing and underdeveloped countries do not have access to financial incentives that deliver improved environmental outcomes, such as lower costs for agricultural loans and PES schemes. Smallholder farming systems are fragmented, and the sustainability and scalability of PES initiatives depend on reliable metrics, and transaction costs (Salzman et al., 2018). Philanthropic, public and private sources will probably combine to supply finance for testing interventions and creating new arrangements for smallholders in the short term. The funding for such initiatives may be provided in the form of blended finance, whereby structured vehicles combine sources of funding for specific goals under particular objectives, or simply mobilize funds from a variety of sources to support certain interventions in a coordinated way. In the long run, the financing needed to continue new value chain activities is likely to be financed mainly by the private sector (such as value chain companies or commercial banks) or from public loan programs.

Investment to improve soil quality can only produce value at the farm level if the agronomic practices promoted, either individually or collectively, bring about augmented profitability for the short term. The benefits of soil health-promoting practices (i.e., improved crop yield and resilience, abridged production costs) can be converted into near-term return on investment, if required inputs, equipment, and knowledge are utilized effectively.

Financial assistance can be provided to farmers for:

Managing land to maintain, restore or promote natural heritage and environmental sustainability.

Managing land to adapt to and mitigate climate change

Managing land to avert, reduce or protect from environmental risks

Improving or protecting the quality and health of soil and plants

Conserving natural resources, habitats, and biodiversity

Starting or improving the productivity of agricultural, horticultural, or the forestry system in an environmentally sustainable manner

Encouraging food production in an environmentally sustainable manner

Financial assistance for promoting these kinds of activities among farmers may be given:

In the form of, among others, grants, loans, or guarantees

Subject to conditions the government considers appropriate, which may include the provisions under which financial assistance is to be repaid or otherwise made good (with or without interest)

To a person connected to a third-party scheme made or operated by that person

In the form of subsidies for farm inputs and machinery that promote the agenda of sustainable land management.

Overall, we can progress toward the improvement of soil quality through existing value chains, policy frameworks, and financing mechanisms. Demand-driven, network-based research and development and technology transfer can cater to cross-value chain interventions together with co-investment partnerships (e.g., with off-taker companies, input suppliers, insurance companies, financial institutions, and traders). Over the past few decades, the CGIAR (Consultative Group on International Agricultural Research) and its partner countries have built a significant body of research on soil aspects. A continuing and prolonged investment is required to advance quantification and prognostic capacity concerning the relationship between investment in strengthening smallholder capacity, changes in soil quality, and improvements in soil management as well as changes in risk and resilience at the farm and community levels. Farm management options can be tested and validated through research and development and scaling efforts in partnership with organizations committed to building farmer capacity.

Private organizations for investment in soil quality and environmental protection

Increasing carbon sequestration would help to maintain or improve soil quality while also reducing emissions and mitigating some of the negative effects of climate change. Investment in soil quality and environmental protection varies as per the interests of private organizations, either to maintain or increase their revenue, to reduce or avoid the costs of their products, to gain reputation, or to open up new financial opportunities (IUCN, 2018). Hence, to achieve the desired outcomes like increased productivity, preserving biodiversity, and environmental conservation, science-based solutions should be piloted and implemented. Recently, with increasing consumer awareness, several global companies have realized the importance of investing in soil quality and environmental protection for the future of their business. Healthy soil and better environmental conditions increase crop productivity, with more secure supply chains and better prospects to meet the global food needs of the growing population (IUCN, 2018).

Several large companies are committed to reducing carbon emissions significantly and promoting actions that will lead to augmented sustainability in agricultural systems. For example, DANONE Inc. has begun a program to become carbon neutral by 2050, which will influence the farm management programs in Belgium (DANONE, 2021). The Nature Conservancy (TNC), a global organization with headquarters in Arlington, Virginia, USA, has funded five new private enterprises that have potential in key areas

of soil quality improvement (The Nature Conservancy, 2021). While Kula Bio is committed to helping farmers to enhance crop production and reduce environmental impact by providing sustainable N solutions through developing fungal biological N fertilizers that reduce runoff losses of N. Pattern Ag offers soil microbiota analysis, and based on that, makes recommendations for the management of farm inputs. Stony Creek Colors aims to improve market-based alternatives for growers to diversify crop rotations by promoting sustainable, plant-based dyes like indigo for the cosmetics, textile, and food markets. Swarm Farm's automated, robotic farm technologies enable the precise application of fertilizers, pesticides, herbicides, and other crop-protection chemicals. The collaborative efforts by the firms promote a regenerative and nature-positive production paradigm.

Case study: India

The Government of India (GoI) has taken many initiatives to check soil degradation and improve/maintain soil fertility status in the country through launching several plans/programs/schemes/projects. These programs encourage the agricultural community to adopt best practices in agricultural waste management and avoid crop residue burning (Fig. 7). For instance, the National Mission for Sustainable Agriculture (NMSA), Soil Health Card (SHC) Scheme, Soil Conservation in the Catchments of River Valley Project and Flood Prone River (RVP&FPR), Reclamation and Development of Alkali and Acid Soils (RADAS), Nutrient-based Subsidy (NBS), production of neem-coated urea, organic farming, Paramparagat Krishi Vikas Yojana (PKVY), National Action Plan for Climate Change (NAPCC), National Mission for a Green India (NMGI) under NAPCC, National Water Mission (NWM), National Mission on Oilseeds & Oil Palm (NMOOP), and Mission for Integrated Development of Horticulture (MIDH) (NAAS, 2018; Chaudhari, 2018) are involved in soil health management, directly or indirectly. The NMSA was launched in April 2014 to restore and reclaim problematic soils (Chaudhari, 2018). The GoI is promoting NMSA for agricultural sustainability in India, followed by the Soil Health Management (SHM) sub-scheme under the NMSA with the prime goal of establishing soil testing laboratories, biofertilizers, and vermicomposting units coupled with training and demonstration to farmers to promote fertilizer application based on plant demands (NAAS, 2018). To support this program, the National Mission on Soil Health Card began in 2015, which provides farmers with information regarding the nutrient status of soil through testing and appropriate fertilizer recommendations (Chaudhari, 2018). The MIDH, Rashtriya Krishi Vikas Yojana (RKVY), and the National Programme on Organic Production (NPOP) promote organic farming for enhancing soil quality. While RADAS promotes the development of land resources through Macro Management of Agriculture schemes. The NBS policy has been in force since April 2010, under which a fixed rate of subsidy is given for different macro- and micro-nutrients every year. Furthermore, urea, which is the most widely used fertilizer, is available to farmers only in the neem-coated form, which led to a 10% decrease in its consumption and improved crop yield by 6–17% (NAAS, 2018). It helped to reduce its use, nitrate leaching, greenhouse gas (GHG) emissions and improved crop yields.

The GoI started to pay attention to this concern and enforced Section 144 of the Civil Procedure Code (CPC) to ban the burning of paddy stover. The National Green Tribunal (NGT), a government enterprise, established under the National Green Tribunal Act, aims to restrain crop burning through recycling initiatives and increasing public awareness.

In addition to the government's strict action on residue burning, the GoI implemented the National Biogas and Manure Management Program (NBMMP) under the "Waste to Energy Mission" to establish biogas plants to use surplus residues effectively and prevent pollution (Shukla, 2007). Under this program, about 5 million families have installed biogas plants. The use of surplus residues in bio-composting is also being promoted under RKVY, i.e., a state Plan Scheme launched in 2007 with Central Assistance;

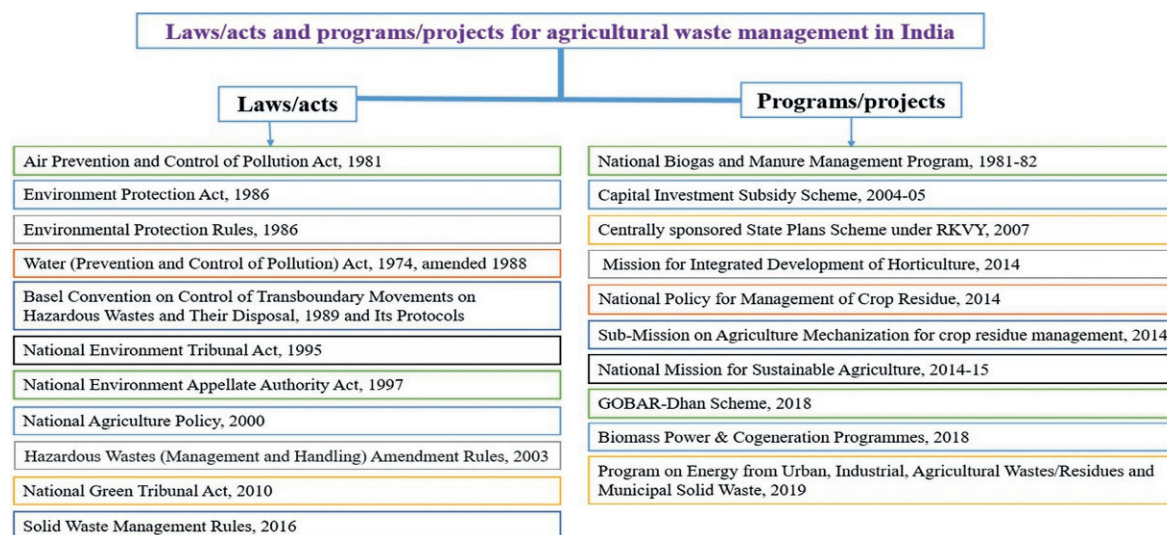


Fig. 7 Laws and programs for waste management.

farmers are trained in agro-waste bio-conversion and bio-compost production through demonstrations and training. National Policy for the Management of Crop Residue (NPMCR) initiated in 2014 by the Ministry of Agriculture & Farmers' Welfare (MoA&FW) promotes technologies for in-situ management and monitor of the monitoring of crop residues (Datta and Meena, 2021). In support of this, the Department of Agriculture Cooperation & Farmer's Welfare (DAC&FW) was launched under the scheme "Sub-Mission on Agriculture Mechanization for Crop Residue Management" in April 2014 to boost farm mechanization. The mission aims to demonstrate CRM machinery in farmers' fields at a rate of ~US\$ 55 ha⁻¹. The scheme mainly comprises three components, (i) Establish Farm Machinery Banks for Custom Hiring of in-situ CRM machinery, (ii) Financial Assistance to farmers for the Procurement of Agriculture Machinery and Equipment for in-situ CEM, and (iii) Information, Education and Communication for awareness on in-situ CRM. Mass awareness campaigns through radio and television programs, advertisements in print and electronic media documentation, demonstration camps, capacity-building programs, star campaigning, awards at the village level, etc. are being implemented across the nation. As a result of these collective efforts, residue burning events have been reduced by 30% compared to the base year of 2016 (MoA and FW, 2021). To stop residue burning and associated nutrient burning in soil, national guidelines have also been issued (NAAS, 2018).

New policy initiative recommendations for soil quality protection in India

Over the last three decades, the GoI has made a significant investment in halting soil degradation and declining soil fertility (Chaudhari, 2018), and several development plans have already been undertaken. At the national level, soil quality management is difficult to achieve without robust collaboration and networks between public and private organizations and institutions. The use of information and communication technologies may improve the relevance and effectiveness of these programs. A strong hands-on training network would be a key component of successful soil quality management programs. Major government policies must be re-examined in the light of the country's shifting priorities and the full potential of large programs like the soil health mission. The following are some of the key policies that must be implemented to restore, improve, and preserve soil quality and ensure agricultural sustainability in the country (Fig. 8) (NAAS, 2018; Rotz et al., 2021):

Solid collaboration and networks are critical for the successful execution of soil quality management initiatives at the national level. With the introduction of new technologies, the following issues must be considered:

1. In the country, land-use policies need to be revised to prevent productive lands from being transferred to other uses so that farmers have options to select crops and cropping systems according to the availability of water and other resources, land capability, socio-economic conditions, and the prevailing market price of the product so that sudden fluctuations in the market price can be avoided.
2. An urgent need to review and strengthen technologies developed by State Agricultural Universities and ICAR institutions to combat wind and water erosion. It is critical to pay more attention to the Himalayan region where the loss of soil, water, and human and animal wealth is rapid and alarming.

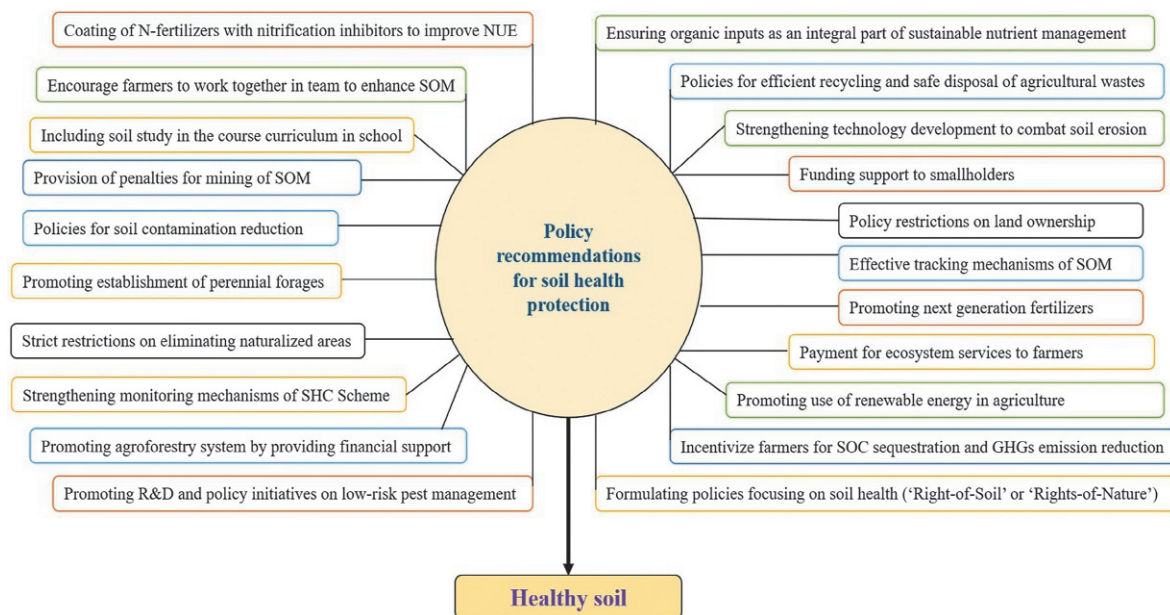


Fig. 8 Policy initiative recommendations for soil quality protection in India.

3. The SHC scheme must have a robust monitoring mechanism through identifying and strengthening key laboratories to ensure the quality and authenticity of the soil test data to make the scheme worthwhile. The scheme should also include plant tissue testing together with soil testing by appointing trained staff to provide advisory services to farmers to ensure practical implementation of recommendations.
4. Biofertilizers, microbial consortia, crop rotation, and organic manures should be an integral part of sustainable nutrient management. Local administrators (e.g., agricultural supervisor, research officers) must ensure their promotion and adoption among the farmers at the micro-level. To promote efficient utilization of organic manures and crop-livestock production systems, training and fiscal incentives should be given to farmers for better manure management and preparation of nutrient-rich composting.
5. Coating of N-fertilizers with nitrification inhibitors needs to be promoted to improve nutrient use efficiency (NUE) and reduce nitrate leaching. For this, incentives may be provided to the fertilizer producers opting to manufacture value-added fertilizers. Farmers should also be incentivized to adopt these products rather than imposing rules and regulations on them.
6. Farmers are compensated for agroecological practices and enhancing SOM through a wide range of programs comprising intended stewardship with due support from the administration.
7. Introduction of a long-term low-risk pesticide and contamination management policy that includes research and development as well as rules, regulations, knowledge transfers, and monetary enticements to make global agriculture a spearhead in low-risk pest management.
8. The SOM stock could be enhanced by expanding farmer-to-farmer support programs by encouraging farmers to work together in a team.
9. Supporting farmers in bringing more areas under a sustainable agroforestry system through implementing programs in which farmers get financial help.
10. Restrictions on the removal of buffer zones, windbreaks, and naturalized areas unless and until other habitats or areas are created.
11. Provide concrete financial and extension assistance to enrich the soil with OM, specifically through the establishment of perennial forages.
12. A policy of cross-compliance that involves penalties for mining SOM from their field by adopting exhaustive farm practices and technologies.
13. Tracking and measurement of soil quality and SOM on all farms in every province.
14. Introducing and implementing policies and strategies for the reduction of soil contamination as for water contamination policies.
15. Increase funding support for non-industrial farms, and small-scale and community-based farming at the state and national levels.
16. Implement policy restrictions on land ownership based on factors such as income, access to capital, properties, farm holdings, and, interest in provisioning food.
17. Promotion of food provisioning, consumption, and soil quality as a part of the primary and secondary school curriculum.
18. New methods of nutrient management, such as fertigation, nano-urea fertilizers, etc. should be promoted.
19. Efficient recycling and safe disposal of agricultural wastes should be strictly followed by introducing the policy in governmental campaigns.
20. The use of solar and wind energy in agricultural operations should be promoted by subsidizing the required equipment/instruments.
21. The introduction of carbon credits or incentives to farmers practicing conservation agriculture to sequester carbon and reduce GHG emissions.

Need to be supported by targeted plans and policies for soil management:

Support for soil developmental activities to monitor and forecast change in soil;

Enhance awareness and education about soil-related issues through integrating such topics across the curriculum, i.e., biology, geography, geology, economics, etc.

Investment in the development, testing, and dissemination of sustainable soil management practices and technologies.

Introduce appropriate and effective regulation and incentives, for instance, removal of subsidies on harmful practices including excessive use of synthetic fertilizer and pesticides. Subsidies could also be given to farmers to encourage the purchase of tools, machinery, and other inputs that are not harmful to soils.

Enhance capacity to manage soil resources sustainably to achieve regional, national, and international food security.

Conclusion

A sustainable soil ecosystem is needed to promote the global economy of the farming community. Key policy initiatives in terms of soil quality protection, maintenance of soil fertility, management, and recycling of agro-residues are vital components for the successful implementation of various scientific approaches toward sustainable soil quality. This chapter helps to promote policies,

plans, and strategies to enhance soil quality by including them as part of land-use planning. A variety of sectors must contribute to this planning to meet common goals such as food supply, water supply, and conservation of biodiversity. Therefore, governments should encourage and incentivize farmers to adopt eco-friendly land management technologies and other ecosystem services. Various eco-friendly practices and technological interventions can help with soil quality protection and boost agricultural productivity. Likewise, the formation of small and medium-sized agribusinesses engaged in sustainable land management should also be facilitated with financial support by the government.

Acknowledgments

The leading author is thankful to the SERB (DST, Govt. of India), New Delhi (Ref. File No. CRG/2020/001774) and Institute of Agricultural Sciences, B.H.U., Varanasi for providing the necessary facilities to conduct the experiment on soil carbon stabilization.

References

- Chaudhari SK (2018) Research and development initiatives for managing soil health in India. *Harit Dhara* 1(1): 3–5. <http://iiss.nic.in/eMagazine/v1i1/5.pdf>.
- DANONE (2021) Towards Carbon Neutrality. <https://www.danone.com/impact/planet/towards-carbon-neutrality.html>. Accessed on September, 2021.
- Datta R and Meena RS (2021) *Soil Carbon Stabilization to Mitigate Climate Change*. Springer. https://doi.org/10.1007/978-981-33-6765-4_1.
- DEFRA (2009) *Safeguarding Our Soils: A Strategy for England*. Nobel House, London: Department for Environment, Food and Rural Affairs.
- European Commission Published a Policy Report (2006); https://ec.europa.eu/environment/soil/soil_policy_en.htm
- FAO and ITPS (2015) *Status of the World's Soil Resources (SWSR)—Main Report*. Rome, Italy: Food and Agriculture Organization of the United Nations and Intergovernmental Technical Panel on Soils.
- Havemann T (2016) *Value Chain Finance for Agricultural Climate Change Resilience*. Wageningen: Technical Centre for Agricultural and Rural Cooperation.
- He K (2019) *Ecological Compensation for Agricultural Wastes Recycling*. Beijing, China: People's Publishing House.
- Henderson B, Cacho O, Thornton P, van Wijk M, and Herrero M (2018) The economic potential of residue management and fertilizer use to address climate change impacts on mixed smallholder farmers in Burkina Faso. *Agricultural Systems* 167: 195–205. <https://doi.org/10.1016/j.agry.2018.09.012>.
- IUCN (2018) New Report Highlights the Business Case for Investing in Soil Health. <https://www.iucn.org/news/water/201812/new-report-highlights-business-case-investing-soil-health>.
- IUSS (2016) Soils on the global agenda. In: Hurni H, Giger M, and Meyer K (eds.) *Developing International Mechanisms for Sustainable Land Management. Prepared with the Support of an International Group of Specialists of the IASUS Working Group of the International Union of Soil Sciences (IUSS)*, p. 64. Bern: Centre for Development and Environment.
- Katyal JC, Chaudhari SK, Dwivedi BS, Biswas DR, Rattan RK, and Majumdar K (2016) Soil health: Concept, status and monitoring. *Bulletin Indian Society of Soil Science* 30: 1–98.
- Klauser D and Negra C (2020) Getting down to earth (and business): Focus on African Smallholders' incentives for improved soil management. *Frontiers in Sustainable Food Systems* 4: 576606. <https://doi.org/10.3389/fsufs.2020.576606>.
- Kumar P, Kumar S, and Joshi L (2015) Policies for restricting the agriculture residue burning in Punjab. *Socioeconomic and Environmental Implications of Agricultural Residue Burning. Springer Briefs in Environmental Science* New Delhi: Springer. https://doi.org/10.1007/978-81-322-2014-5_6.
- Land and Soil Policy (2019) *Global Environment Outlook—GEO-6: Healthy Planet, Healthy People*. Cambridge: Cambridge University Press, pp. 372–397. <https://doi.org/10.1017/9781108627146.021>. UN Environment (ed.).
- McCann M, Mehta R, Shakhovskoy M, Zook D, and Carroll T (2018) *Getting Smarter on Subsidy: The Role of Grant Funding in Smallholder Finance*. Washington DC: ISF Advisors.
- Meena RS, Kumar S, Sheoran S, Jhariya MK, Bhatt R, Yadav GS, Gopinath KA, Srinivasa Rao C, and Lal R (2021) Soil organic carbon restoration in India: Programs, policies, and thrust areas. Chapter 13. In: Rattan Lal (ed.) *Soil Organic Matter and Feeding the Future*, p. 34. Abingdon: CRC Press/Taylor & Francis Groups. <https://doi.org/10.1201/9781003102762-13>.
- Meena RS, Yadav A, Kumar S, Jhariya MK, and Jatav SS (2022) Agriculture ecosystem models for CO₂ sequestration, improving soil physicochemical properties, and restoring degraded land. *Ecological Engineering* 176: 106546. <https://doi.org/10.1016/j.ecoleng.2022.106546>.
- MoA & FW (2021) Central Sector Scheme on 'Promotion of Agricultural Mechanization for In-Situ Management of Crop Residue in the States of Punjab, Haryana, Uttar Pradesh and NCT of Delhi. <https://www.pib.gov.in/PressReleaseDetail.aspx?PRID=1696263>.
- NAAS (2018) *Soil Health: New Policy Initiatives for Farmers Welfare. Policy Brief No. 3*. New Delhi: National Academy of Agricultural Sciences19.
- Policy Brief (2019) Sustainable Soil Management. https://www.deltares.nl/app/uploads/2019/04/Soils4EU_D1.3_PolicyBrief_DEF.pdf.
- Rotz S, MacRae R, and Martin RC (2021) *A Canadian Soil Health Vision and Strategies*. Food Policy for Canada. <http://foodpolicyforcanada.info.yorku.ca>. (Assessed in August, 2021).
- Sabitri EN (2011) Utilising agricultural waste to enhance food security and conserve the environment. *African Journal of Food, Agriculture, Nutrition and Development* 11(6): 1–9.
- Salzman J, Bennett G, Carroll N, Goldstein A, and Jenkins M (2018) The global status and trends of payments for ecosystem services. *Nature Sustainability* 1: 136–144. <https://doi.org/10.1038/s41893-018-0033-0>.
- Schismenos S, Emmanouloudis D, Stevens GJ, Katopodes ND, and Melesse AM (2022) Soil governance in Greece: A snapshot. *Soil Security* 100035.
- Sherwood S and Uphoff N (2000) Soil health: Research, practice and policy for a more regenerative agriculture. *Applied Soil Ecology* 15(1): 85–97.
- Shukla PR (2007) *Biomass Energy Strategies for Aligning Development and Climate Goals in India*. The Hague, The Netherlands: Environmental Assessment Agency.
- Soil Health Institute (2022). <https://soilhealthinstitute.org/resources/catalog/>
- The Nature Conservancy (2021) Accelerating the Regenerative Revolution: The Nature Conservancy Invests in Agri-Tech Firms to Speed Progress Against Ambitious Soil Health Goals. Published on March 08, 2021 | Arlington, VA <https://www.nature.org/en-us/newsroom/tnc-invests-emerging-agri-tech-firms-to-speed-progress-against-ambitious-soil-health-goals/>.
- The Scottish Government (2009) The Scottish Soil Framework. <https://www.gov.scot/publications/scottish-soil-framework/>. ISBN: 9780755.
- USDA (2022) United States Department of Agriculture. <https://www.nrcs.usda.gov/wps/portal/nrcs/main/national/programs/financial/>.
- Wall DH and Nielsen UN (2012) Biodiversity and ecosystem services: Is it the same below ground? *Nature Education Knowledge* 3(12): 8.
- Wang B, Dong F, Chen M, Zhu J, Tan J, Fu X, Wang Y, and Chen S (2016) Advances in recycling and utilization of agricultural wastes in China: Based on environmental risk, crucial pathways, influencing factors, policy mechanism. *Procedia Environmental Sciences* 31: 12–17.
- WBCSD (2018) *The Business Case for Investing in Soil Health*. Geneva: WBCSD.

Construction of Technosols for green roofs and vertical green system

Eleonora Flores-Ramírez^a, Stefan Abel^a, Elizabeth Chávez-García^b, and Geoffroy Séré^c, ^aIndependent Researcher; ^bDepartment of Geography, Faculty of Philosophy and Letters, National Autonomous University of Mexico, Mexico City, Mexico; ^cLaboratoire Sols et Environnement, Université de Lorraine, INRAE, Nancy, France

© 2023 Elsevier Ltd. All rights reserved.

Introduction	474
Green roofs and vertical green systems	474
Technical aspects of green roofs	475
Technical aspects of vertical green systems	476
Constructed Technosols for green roofs and vertical green systems	478
Physical requirements of GR and VGS Technosols	478
Water holding capacity and drainage	478
Particle structure, bulk density (ρ_b), and air capacity	481
Chemical requirements of GR and VGS Technosols	481
pH	481
Buffer capacity	481
Cation exchange capacity	481
Chemical stability	481
Biological activity	482
Parent materials and constructed Technosol composition	482
Natural soils and natural mineral materials	482
Processed natural materials	483
Recycled mineral materials	483
Natural and recycled organic materials	484
Substrate composition – defining the optimal ratios	484
Aging and early pedogenesis	484
Benefits provided by GR and VGS in the urban landscape	485
Energy consumption and energy balance of buildings	485
Flood risk reduction	486
Air quality	486
Biodiversity	486
Conclusion	486
References	487

Abstract

Owing to the scarcity of green spaces in cities, the installation of green infrastructure on façades and roofs have gained importance in providing a range of ecosystem services. Compared with natural soils, common growing media for green roofs and vertical green systems on façades consist of predetermined mixtures of organic and inorganic, natural, modified and synthetic materials, arranged in a soil-like horizon structure to provide anchorage, store and provide nutrients and water but allow fast drainage. They are referred to as Technosols, soils having large amounts of manufactured and modified materials (artifacts), and can be adapted for each greenery system.

Key points

- Green roofs and vertical green systems are engineered environmental design features installed for economic, environmental, or ecological reasons.
- Green roofs and vertical green systems must have appropriate growth media to sustain vegetation in harsh environmental conditions, such as strong winds, long periods of direct sunlight, drought, or strong rainfall.
- Growth media of green roofs and vertical green systems can be classified as Technosols and consist of deliberate mixtures of organic and inorganic, natural, processed and/or recycled materials, arranged in a soil-like horizon structure.
- Preferred and commonly used materials in the mixtures are coarse, lightweight and porous to retain water and simultaneously allow a fast drainage, while avoiding high weight loads on the structures.
- Green roofs and vertical green systems mitigate the urban heat island effect, reduce energy consumption for cooling buildings, reduce air pollution, create habitats for biodiversity, and regulate flooding through the runoff reduction.

Introduction

Since the beginning of civilization urban areas have been cultural hotspots, centralizing government activities, health care, transportation, industrial and commercial activity. They thus play an important role for humans. But urbanization always implies drastic human-made modification of the local environment along with a huge consumption of resources and disposal of the residues.

The world population of humans is continually increasing, being expected to reach more than 9 billion by 2050 with an increasing proportion living in urban areas. As a consequence, cities are continually expanding. In 2009, urban areas covered 3.5 million km² (i.e., 2.6% of the total land areas excluding Antarctica and Greenland), in which 0.6% of soils are covered by impervious surface (Liu et al., 2014).

Urbanization impacts negatively on the environment by disrupting water and nutrient cycles and emitting pollutants. The land use change, and especially surface sealing with concrete or asphalt for roads and housing is one of the main drivers of such disruption: it implies biodiversity loss, decreased groundwater recharge, water uptake and evaporation by the vegetation, and raised surface runoff. Thus, the urban surfaces heat rapidly causing the urban heat island effect (Scalenghe and Marsan, 2009).

A conjunction of strategies to mitigate the negative impacts of urbanization and increase the resilience of urban areas consists of the conservation and refurbishment of existing vegetated areas, as well as the generation of new green infrastructure. Green infrastructure (GI) is defined as natural, semi-natural areas, and engineered environmental design features that are strategically interconnected also with other natural areas and provide multi-ecological functions (Maes et al., 2015). Crucial ecological functions provided by GI are the improvement of air and water quality, reduction of water runoff into sewer systems, filtering of noise, microclimate stabilization, food production, improvement of psychological and physiological human health, and strengthening social relationships by attracting people to outdoor spaces (Ampim et al., 2010; Haaland and van den Bosch, 2015; Mentens et al., 2006; Jaffal et al., 2012). Economic advantages include increased tourist activities, reduced energy use for indoor temperature regulation, lower wastewater treatment costs, and increased property value, which generates more tax revenue (Tzoulas et al., 2007; Abass et al., 2020).

GI ranges from completely human-made constructions (e.g., green roofs) to natural ecosystems. Some examples of GI are:

- public and private parks and gardens
- tree pits
- orchards, urban agricultural and horticultural lots
- green façades
- extensive, intensive and semi-intensive green roofs
- infrastructure for rainwater harvesting such as bioswales, infiltration basins and trenches, filter strips, and ponds

Despite their desirability, the installation of new large green areas in densely populated cities entails many challenges. The number of available areas is limited, many of them have inappropriate soil quality to sustain vegetation. Also, the installation of green areas competes for priority with land designated for residential, commercial, or other service uses. Constructing a new green area requires a large amount of financial, human and natural resources. In particular, high quantity natural soil material from periurban areas has to be introduced to offset the poor quality of urban soils. Moreover, municipal administrations hesitate to implement green infrastructure projects due to the difficulty of evaluating their economic benefits (Haaland and van den Bosch, 2015).

In this chapter we focus on the green infrastructures of green roofs and vertical green systems, as they are versatile design features compatible with residential and commercial land use, thus reducing conflicts of interest and providing ecological and social benefits associated with vegetated areas. We present an overview of the technical aspects of green roofs and vertical green systems and how their installation design and the local conditions constrain the criteria for establishing the growing media and the plant selection. In particular, the pedogenic characteristics of the growing media and the commonly and alternative materials used are addressed.

Green roofs and vertical green systems

Due to the scarcity of available areas, the use of alternative available spaces that were not necessarily intended to be vegetated (secondary urban greening; Flores-Ramírez et al., 2018) have gained importance, such as façades and roofs. Green roofs (GR) and vertical green systems (VGS) have the potential to provide a wide range of ecosystem services. The installation can be intentionally designed from the beginning of a projected building or can be adapted to cover existing structures by retrofitting. The selection of the vegetation can have esthetic (greenery), protective (against wind erosion and storm water impacts), or productive (agriculture) purposes. Importantly, their creation requires almost no horizontal space on the ground, and are thus a versatile and promising option for growing cities.

Together with green balconies, GR and VGS conform what is called greenery systems, and are considered part of the architectural style of buildings (Fig. 1). In recent decades, greenery systems have spread widely as architectural and constructive elements in the design of façades and roofs (Abass et al., 2020). A commercial boom with new substrates, structures, and plant selection has thrived in many countries of the global north, where regulations for their installation have been established (e.g. the German Regulation for GR "Dachbegrünungsrichtlinien" FLL, 2018). But they are also spreading to other latitudes (Asia, Latin America). However, the esthetic and technical aspects have been prioritized over the environmental ones (Besir and Cuce, 2018).



Fig. 1 From left to right: Examples of an extensive green roof (first two left images) a semi-intensive green roof (third image), a green façade (fourth image) and a modular living wall (right). Photos Eleonora Flores, Stefan Abel and Geoffroy Séré.

Technical aspects of green roofs

Green roofs (eco-roofs, roof gardens, living roofs) can be defined as roofs coated with a layer of growing medium, which supports vegetation (Abass et al., 2020). Depending on the depth of the growing medium and the technical and biotic possibilities and challenges, the GR are classified into three groups: extensive, semi-intensive, and intensive (Table 1). The extensive GR implies a light construction design, it is shallow and easy to maintain as the commonly used vegetation with shallow root systems is quite stress resistant within local climate conditions. The selected plants should endure long dry periods without irrigation (e.g., species from the family Crasulaceae, and the *Sedum* genus in particular) or be adapted to areas with frequent precipitation (e.g., combinations with moss and lichens). The semi-intensive and intensive systems have more volume for root growth and commonly include an irrigation system. These systems can support a wider selection of plants, ranging from herbs, woody bushes and even small trees (<10 m), but deeper substrates also require more maintenance. For all types of GR but particularly for extensive GR systems, it is highly recommended that native plant species are selected, as they are well-adapted to local climatic conditions. The hours of intensive solar radiation as well as shading have also to be considered in vegetation selection and the inclination angle of the GR to be installed.

In general, despite all the current variations and commercial possibilities that the market offers, the basic structure of GR from top to bottom (Fig. 2), apart from the vegetation and the substrate, consists of:

- Filter sheet: made of a polymer non-woven fabric, e.g., fleece, which avoids finer particles of the substrate being transported downwards with the infiltrating water or the penetration of roots, both of which can clog the drainage system.

Table 1 GR characteristics.

Characteristic	Extensive GR	Semi-intensive GR	Intensive GR
Structural requirements	Suitable for retrofitting buildings, sloped roofs, large sized areas	Suitable for retrofitting buildings, medium sized areas	Requires planning in advance of construction of whole building Not suitable for sloped roofs
Thickness of substrate layer	Low weight load capacity 6–20 cm	Moderate weight load capacity 10–25 cm	High weight load capacity >25 cm
Particle size distribution of drainage layer (when made of aggregates)	2–8 mm to 2–12 mm	4–8 mm to 2–16 mm	4–8 mm to 16–32 mm
Weight (under water saturated conditions)	50–150 kg m ⁻²	120–350 kg m ⁻²	> 300 kg m ⁻²
Common vegetation	Small: succulents (suitable for arid and semi-arid zones), grasses, mosses (in humid climates)	Medium size: Shrubs, grasses, seasonal herbaceous plants	Large size: from grasses, succulents, shrubs, and small trees.
Irrigation	Seldom. In case of extreme drought only	Periodically/Seasonally	Regularly
Roof slope [%] and inclination [°]	Up to 100% (45.0°)	< 20% (< 11.3°)	< 5% (2.9°)
Costs	Low - medium	Medium	High
Construction and maintenance investment	Low	Medium	High

Modified from González-Méndez B and Chávez-García E (2020) Re-thinking the Technosol design for greenery systems: Challenges for the provision of ecosystem services in semiarid and arid cities. *Journal of Arid Environments* 179: 104191.

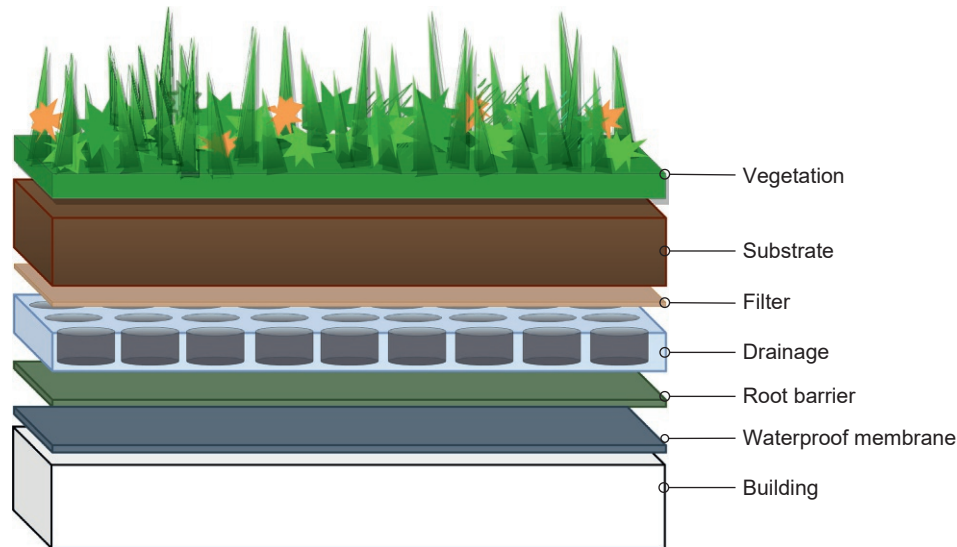


Fig. 2 Schematic profile of a simple extensive GR.

- Drainage layer: either made of mineral materials with large grain sizes that allow the water to percolate rapidly (such as gravel, expanded clay or pozzolana) or geosynthetic drainage layers with a special core system. This layer is necessary to reduce the risk of water stagnating in the root zone.
- Root barrier: often a geomembrane that prevents root penetration and consequent damage of lower waterproofing layer or the building structure. It can also be a sealed concrete or asphalt layer or consist of metal trays, which act as a waterproofing sheet. A properly installed geomembrane with welded junctions can work as root barrier and waterproofing sheet simultaneously.

Some variations include a single-layer construction, consisting of a vegetation supporting layer with drainage and filtering functions.

In a multilayer construction it is common to install an insulation layer between the root-barrier geomembrane and the waterproofing layer or the building structure.

The installation of an irrigation system on semi-intensive and intensive GR helps to sustain the vegetation, especially in areas with hot and dry climate conditions during the vegetation period.

Further technical considerations:

- I. The weight load capacity to install a GR must bear the total weight of substrate under water-saturated conditions, further installed facilities, plus the biomass weight. The structures and the substrates should therefore preferably be light.
- II. Low flammability: resistance against flying sparks and radiating heat. A maximum of 20% (*w/w*) organic matter (OM) is therefore recommended (FLL, 2018).
- III. It is desirable that the GR substrates have light hues, or high albedo, therefore reducing the adsorption of solar radiation and the consequent heating.

Technical aspects of vertical green systems

Vertical green systems (also called vertical gardens, green walls, vertical greens, vertical landscaping and bio-walls) are a diverse arrangement of installation designed to create a vegetation envelope on vertical building walls.

VGS consists of two systems relying on the facility for the plant root system and whether implemented outdoors or indoors (Fig. 3):

- a) green façades: the root system grows directly in the soil or in substrate-filled containers and the aerial part of the plants cover the building envelope. The vegetation cover develops directly on the wall (either with adventitious roots or adhesive pads) or on supporting structures attached on the wall but leaving a gap in between. Different performances in terms of temperature regulation will depend on this characteristic (Fig. 4A–C).
- b) living walls or green walls: the root system of the plants grows in large or small modules, panels, bags or similar structures attached to a supporting structure along the wall. Living walls therefore require more maintenance and a more elaborate infrastructure to place the growing substrates and the irrigation system than green façades (Fig. 4D). However, they allow a variety of plants to contribute to the uniform covering of the wall simultaneously, with modules that can be replaced easily if needed. This kind of VGS is preferred for indoor spaces. Among the diverse forms of living wall, the continuous system has no

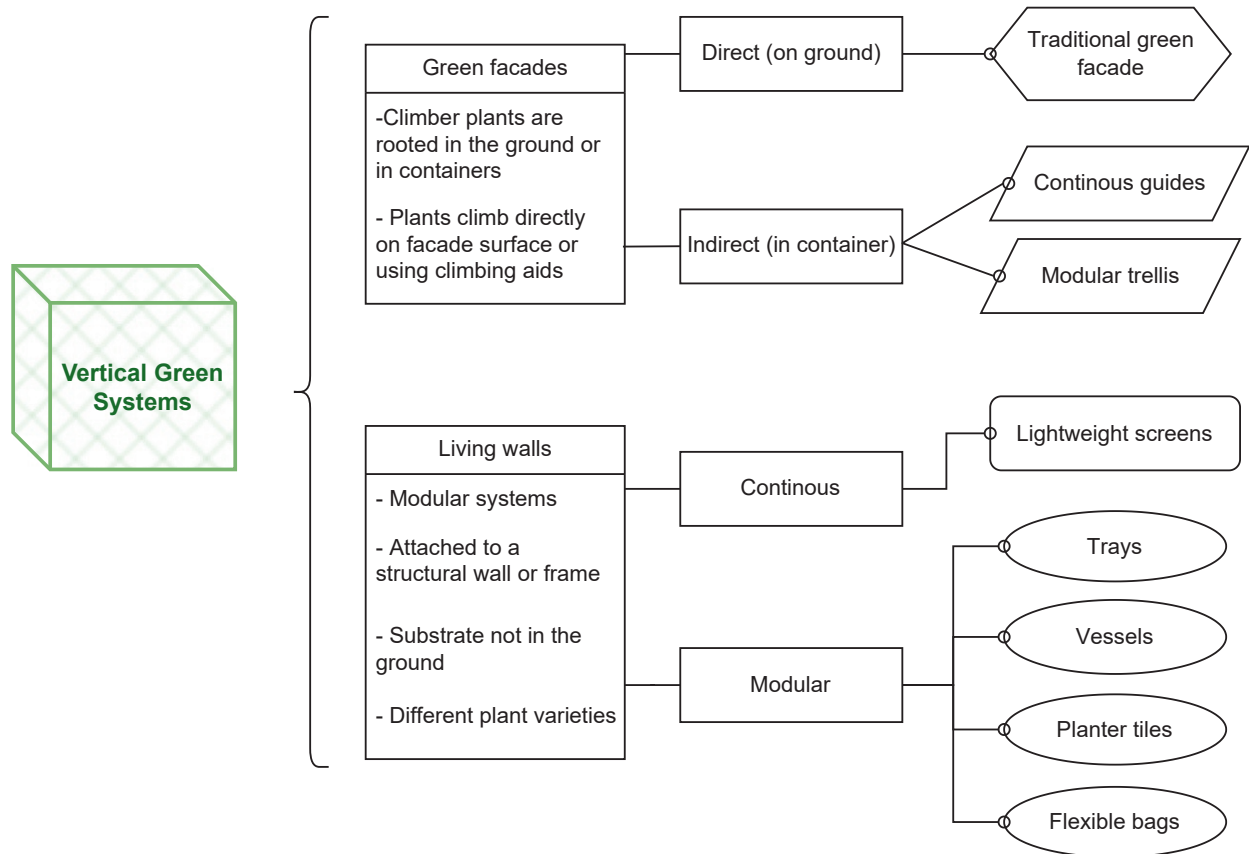


Fig. 3 Differences between vertical green systems. Modified from Manso M and Castro-Gomes J (2015) Green wall systems: A review of their characteristics. *Renewable and Sustainable Energy Reviews* 41: 863–871.

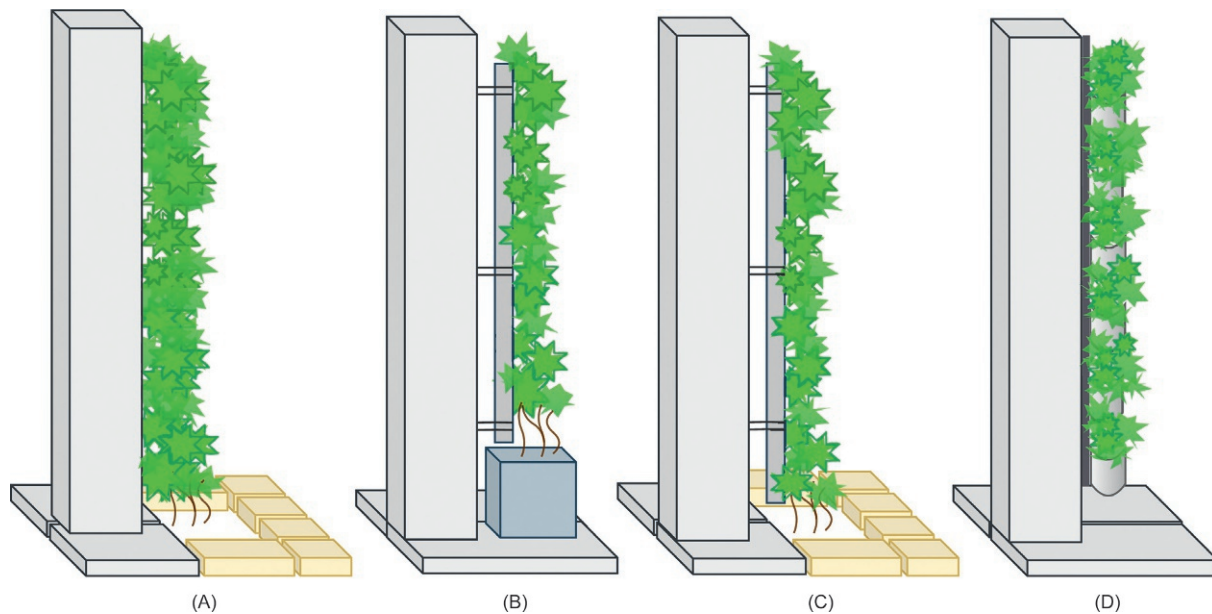


Fig. 4 Examples of vertical greenery: (A) direct green facade, (B) indirect green facade with supporting structures for continuous guides, (C) direct green facade with supporting structures for continuous guides, (D) modular living wall with flexible bags.

substrate. It has a hydroponic system with lightweight absorbent screens where the plants are inserted and are continuously irrigated with a nutrient solution. The modular living walls do have lightweight substrates where the plants can easily root (Besir and Cuce, 2018; Manso and Castro-Gomes, 2015).

Constructed Technosols for green roofs and vertical green systems

The GR and VGS are totally human-made. Their construction challenge is to provide proper growth conditions for vegetation on surfaces that typically are not planned for that. Therefore, apart from the formal structural and architectural conditions, a proper growth medium and appropriate plant selection is crucial.

Unlike natural soils which develop in-situ over time, common growing media for GR and VGS consist of deliberate mixtures of organic and inorganic, natural, modified and synthetic materials, arranged in a soil-like horizon structure to provide anchorage for roots, store plant available water, drain water excess, as well as retain and provide nutrients for plants (Ampim et al., 2010; Rokia et al., 2014; Damas and Coulon, 2016). The growing media lies on an impervious layer, either concrete, a geomembrane or it is retained in a pot-like container.

In the World Reference Base of Soil Resources (WRB) such artificially constructed soils are classified as Technosols (IUSS Working Group WRB, 2022). They notably contain large amounts of artifacts (solid or liquid substances that are created or substantially modified by humans) ranging from ca. 20% (v/v) to a complete anthropogenic composition. They can be created circumstantially or with specific functional purposes, as in this case, to fulfill plant-supporting functions (Rokia et al., 2014; Damas and Coulon, 2016; Flores-Ramírez et al., 2018) for a diversity of GI substrates.

These Technosols can be mainly described as isolatic (from the Italian *isola*, island) due to the implementation design with soil depth of ≤ 100 cm separated from the native soil surface by an impervious layer (IUSS Working Group WRB, 2022). The supplementary qualifiers can be 'folic' for the installed organic layer, not constantly water saturated and well aerated, 'drainic' for systems with an artificial drainage, and 'transportic' with substrates formed from allochthonous natural soil by transportation, which can be used to elaborately describe the Technosols of GR and VGS. Further qualifiers depend on the technical origin and characteristics of parent materials.

Commonly, the GR and VGS growing media is a mixture of extracted and relocated materials from natural origin (e.g., sand, peat, light volcanic rocks as pumice and scoria) and diverse commercial and recycled materials (e.g. expanded clay, compost, bricks, tiles, debris). The construction of Technosols as plant growing media implies the intentional optimization of the physical, chemical and biological characteristics of the mixtures adapted to the local climate conditions inside the boundaries of the infrastructure requirements the GR and VGS. Therefore, there is no all-purpose growing medium, however, it has to meet the general physical and chemical requirements discussed next.

Physical requirements of GR and VGS Technosols

Main physical characteristics to be considered for GR and VGS substrates are: large water holding capacity, large hydraulic conductivity to avoid waterlogging and facilitate drainage, small bulk density, sufficient air filled and total porosity.

Water holding capacity and drainage

An adequate growing media should be able, simultaneously, to balance the fast drainage needs with the capacity to store water that is available to plants, i.e., hydraulic conductivity and water holding capacity.

The capacity of a growing media to store, as well as to drain water, depends on its pore system. The available water capacity (AWC) is the water retained against gravity by the pores ranging from 0.2 to 50 μm . Smaller pores retain water that the roots cannot extract, and larger pores drain the water faster than the root can take it up.

In natural soils a well-balanced mixture of organic matter (OM), clay and silt provide an appropriate matrix for water retention. However, fixed weight and load capacities of the supporting structures limits the weight of GR and VGS (except green façades), thus, in general the amount of natural topsoil (with a large particle density and heavy when saturated with water) must be reduced and replaced with lighter materials able to store water. Also, the content of finest particles, particularly clay, must be reduced as they could migrate downwards, and clog the drainage systems. Some guidelines recommend reducing the silt and clay particles with sizes ≤ 0.063 mm at $\leq 20\%$ (v/v) for intensive GR and $\leq 15\%$ (v/v) for extensive GR. This has led to a general use of growing media, which includes a large content of porous materials and up to 20% (w/w) of OM (González-Méndez and Chávez-García, 2020). Examples of light porous materials are expanded clay, expanded shale, scoria, and perlite.

Commonly, the intraporous structure within the porous material retains water while allowing the drainage of water excess through the larger intra- and interparticle pores (Fig. 5). They have a low bulk density and large grain sizes, which provide good aeration and reduce the clogging of the drainage systems. The AWC depends further on the connectivity of the intra- and interparticle systems. A disruption in capillary continuity can hinder water exchange and thus reduce AWC (e.g., the ceramic cover of expanded clay). Also, some porous materials partly exhibit closed pores which are not accessible and do not contribute to AWC, e.g., some scoria (Flores-Ramírez et al., 2018). Their performance can be improved by crushing.

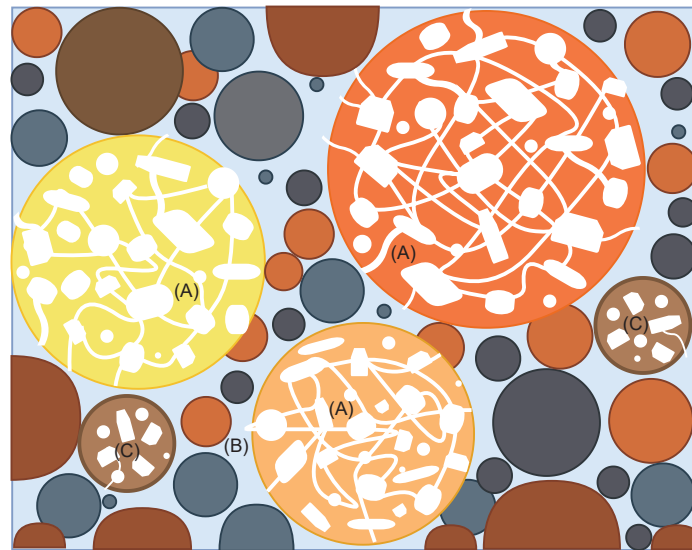


Fig. 5 Schematic representation of (A) intrapore system within porous materials, (B) interparticle pores among porous and solid particles, and (C) porous materials which partly exhibit closed pores.

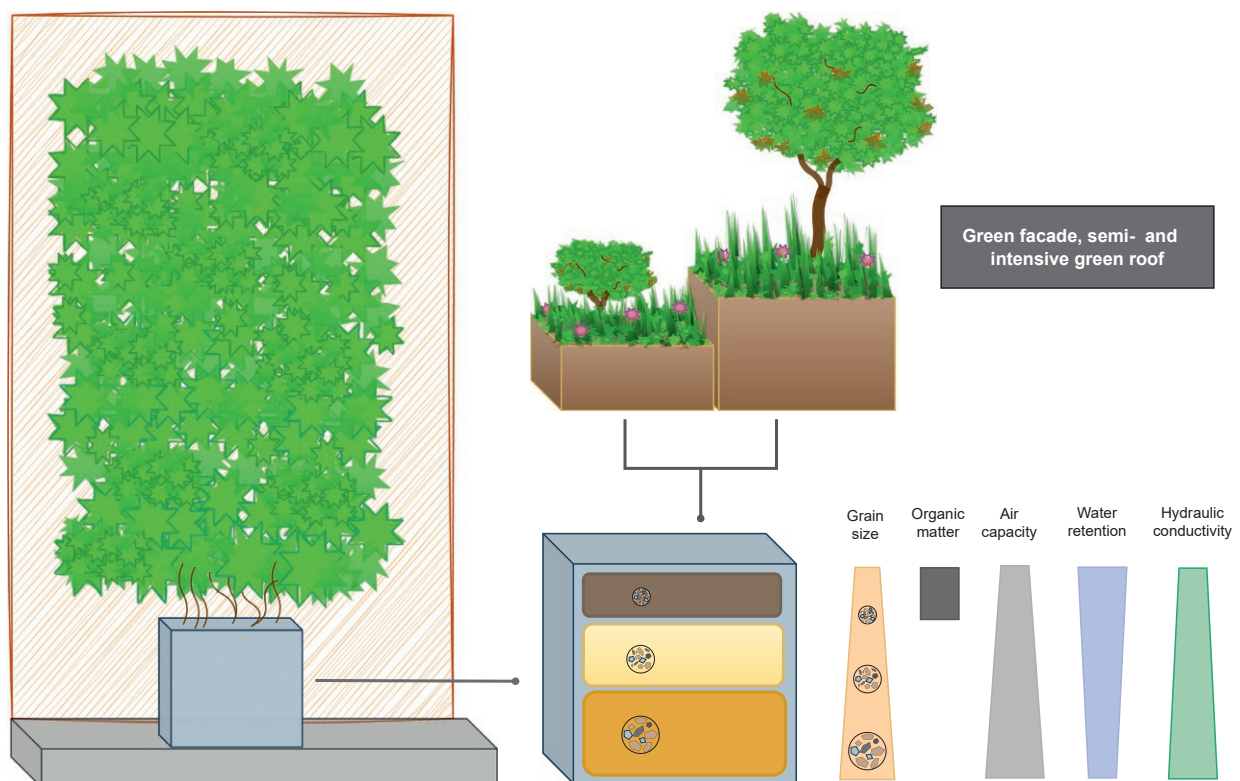


Fig. 6 General physical characteristics of the substrates in the GI where the substrate depth allows a multilayer system. Particle sizes, hydraulic conductivity and air capacity increase from top to the bottom, and the water retention concentrates in the rooting zone.

As previously mentioned, having a drainage layer at the bottom of the structure is crucial for all GR and VGS to avoid waterlogging and plant 'drowning' after large rainfall events. In addition, a high hydrostatic potential may also harm the waterproofing layer and may lead to water leaking.

A multilayer system (Fig. 6), each with various functional substrate layers and with particle sizes increasing from the top to the bottom, can maximize the AWC without decreasing adversely the hydraulic conductivity of the overall installation. Thus, a balanced relationship with water storage at the top, capillary rise at the intermediate layers and drainage at the bottom is recommended.

Organic matter beneath the growth substrate has to be avoided, as it decomposes, forming a biofilm that can clog the drainage, affecting plant growth, and increasing the weight of the GR (Ampim et al., 2010).

Extensive GR and modular systems of living walls

As this design has the shallowest substrate layer and root zone, and lacks an irrigation system, the AWC of the growth substrate is most crucial for plant growth and has to be maximized. An AWC of at least 15% (v/v) should be achieved. However, due to limits in weight load capacity the use of common substrates of a multilayer system is restricted and mostly coarse grained and porous substrates are used, which exhibit reduced AWC (Fig. 7). For drainage, a hydraulic conductivity of $0.6\text{--}70\text{ mm min}^{-1}$ is recommended for a multilayered GR construction and $60\text{--}400\text{ mm min}^{-1}$ for a monolayered construction (FLL, 2018). In arid or very humid regions the requirements may be even higher.

Semi-intensive and intensive GR

These GI are more similar to a conventional terrestrial garden; the AWC may not be the main concern as they have a deeper rooting area and commonly an irrigation system, which can nevertheless be optimized through the use of a proper substrate. For such systems the weight of the substrates used is often the crucial factor and light substrates are preferred. For drainage, a hydraulic conductivity of $0.3\text{--}30\text{ mm min}^{-1}$ for a multilayered GR construction and $60\text{--}400\text{ mm min}^{-1}$ for a monolayered construction is recommended (FLL, 2018). In arid or very humid regions the overall requirements for the substrate may be even higher.

Green façades

This kind of GI have a special challenge, because they have a relatively small volume for rooting and water supply, given the large aerial biomass they sustain. This implies a maximized leaf area for transpiration (and plant species-specific transpiration rates). It thus translates into greater watering needs.

For a central Europe climate (Berlin), Hoelscher et al. (2016) found an average daily transpiration rate (measured with daily sap flow rates) of $0.7\text{ L d}^{-1}\text{ m}^{-2}$ of covered wall area and a variation in the maximum daily transpiration rates (hot summer days) of the three plants studied from 1.3 to up to $2.3\text{ L d}^{-1}\text{ m}^{-2}$ of covered wall area: *Parthenocissus tricuspidata* (adhesive pads) < *Hedera helix* (adventitious roots) < *Fallopia baldschuanica* (needs climbing aids), which can roughly be translated into irrigation needs. For example, a leaf area of 36 m^2 of *P. tricuspidata* would need between 18 L d^{-1} to up to 46.8 L d^{-1} in summer. This illustrates the importance of irrigation or a sufficient AWC of the substrates.

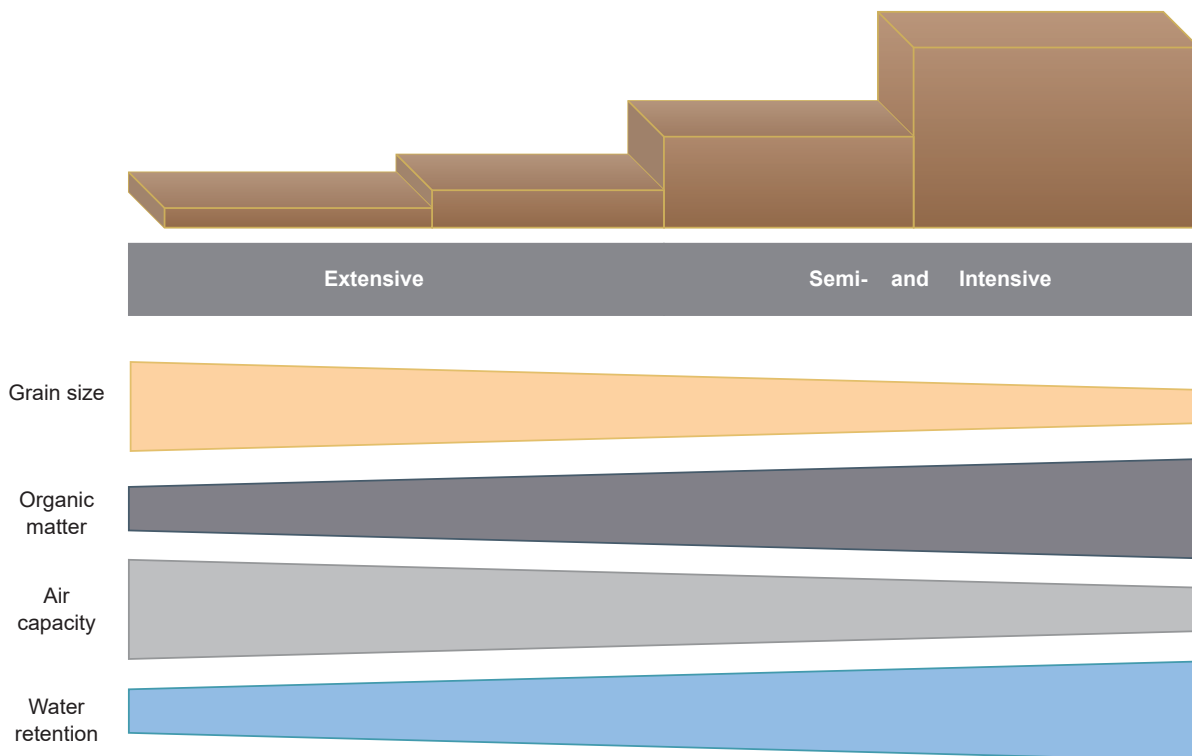


Fig. 7 General physical characteristics of the substrates in the three types of GR, from the shallow substrate layer of the extensive GR to the thicker substrate layers of semi-intensive and intensive GR.

Particle structure, bulk density (ρ_b), and air capacity

The constructed Technosol must have sufficient structural stability to avoid wind removal of the substrate, compaction, and to provide overall stability for the plant-substrate system against mechanical stress due to wind. This is mainly determined by the particle size distribution and the particle shape. Coarse crushed angular solid or porous particles provide a better interlinking of the substrate lattice and are recommended as structure-forming materials for extensive GR. For extensive GR a poorly-graded particle size distribution is recommended to minimize weight load, whereas well-graded particle size distribution is recommended for intensive GR and VGS.

Nevertheless, when the weight load capacity enables large loads, an installation of well-graded substrates on extensive GR maximizes the AWC.

When substrates with poorly-graded particle size distribution are used, the filter-stability of the substrate itself is critical leading to suffusion of the small particles and therefore a filter fleece should be applied to avoid clogging.

The largest particle size recommended are:

- For layers up to 10 cm: 12.5 mm
- For layers >10 cm: 16 mm

and a maximum content of 15–20% (v/v) for particles <0.063 mm (FLL, 2018).

The bulk density (ρ_b) of the substrate installed on supporting structures must be small to avoid exceeding the structural weight load capacities. Furthermore, the ρ_b corresponds with the root penetration resistance, thus a small ρ_b enables proper root growth. A dense root network is crucial for structural stability of the system, increasing the resistance against wind removal of the substrate as well as windthrow.

According to FLL (2018), appropriate bulk densities range from 0.6 to 1.2 g cm⁻³ or 1.0 to 1.8 g cm⁻³ at field capacity for all GR types. To achieve such a ρ_b unintended compaction during installation has to be avoided.

Substrates should have an air capacity of at least 20% (v/v) to guarantee adequate drainage (Ampim et al., 2010; FLL, 2018). At an air capacity less than 5% (v/v), which implies simultaneous high level of compaction, poor aeration as well as impedance of root growth occurs. Some researchers claim that an optimal distribution in terms of solid, water, and air phases should be 40, 40, and 20% (v/v), respectively (Ampim et al., 2010).

Chemical requirements of GR and VGS Technosols

Important chemical properties include pH, buffering capacity, cation exchange capacity (CEC) and chemical stability, which determine the nutrient availability and the leachate quality. It is important to consider that one of the problems presented by the use of chemical properties as indicators of Technosol quality is its high seasonal variability and Technosol management (e.g., nutrient supply, irrigation, climate.).

pH

The commonly accepted range of pH for optimal plant growth ranges between 5.5 and 8.5 (FLL, 2018), as plant nutrients are usually most available between pH 5.5 and 7.0. However, it will depend on the plant preferences (e.g., wetland plants prefer acid pH) and also on the environmental conditions, such as acid rain, which needs to be buffered by alkaline substrate to counteract the acidification.

Buffer capacity

The soil buffering is governed by the protonation and deprotonation of minerals (e.g., aluminosilicates or CaCO₃) and organic matter, which act as sinks for H⁺ and OH⁻ ions to maintain a relatively stable pH despite acidifying or alkalizing factors. It maintains appropriate pH conditions for a wide range of plants and microorganisms. The composition of the constructed Technosols determines the buffer capacity depending on its minerals. The application of fertilizers or other soil conditioners can alter the buffer capacity (Huot et al., 2015).

Cation exchange capacity

The components able to absorb and desorb plant nutrients (e.g., Ca, Mg and K) in GR and VGS substrates are mainly the OM and some clays of colloidal size (< 1 µm), which have a surface charge and thus provide a high CEC. However, common inorganic substrates lack clay and most of their components have a low or neglectable CEC (e.g., perlite, sand etc.). In these cases, plants growing on these substrates may need additional nutrient supplies (Ampim et al., 2010). Nutrient leaching has to be avoided adjusting fertilizer application to the nutrient demand of plants.

Chemical stability

Weathering of inorganic components and degradation of organic components might, over time, modify the CEC and pH of a Technosol depending on the rate of both processes, just as in natural soils. Consequently, the C:N of the organic matter and climatic factors, such as precipitation and temperature, determine the mineralization rate and thus the release of nutrients for

microorganisms and plant uptake. The chemical stability can be a decisive factor for the long-term stability of a Technosol (Séré et al., 2010; Huot et al., 2015).

Biological activity

The biological activity refers to the organisms living in soils and their interaction with the soil matrix. As in natural soils, organic substrates as well as dead tissues of the plants growing on the GI are transformed by organisms colonizing the constructed Technosols, especially bacteria, algae, protozoa and fungi but also nematodes and plants. This transformation occurs through changes in pH driven by acids released by plants and bacteria, the processes of mineralization (conversion of organic components into soluble mineral compounds) and humification (decomposition of fresh organic matter, which results in stable organic substances). Mineralization products are much simpler molecules (e.g., NH_4 , NO_3^- , PO_4^{3-}) that can be absorbed by plants. Humic substances can form strong bonds with clay minerals and oxides (organomineral associations). Pedoturbation lead by nematode activity modify the Technosol matrix by making bores and burying materials at greater depth. This can lead to the formation and stabilization of aggregates and modify the distribution and sizes of the pores, implying a change in the water retention and aeration capacities (Huot et al., 2015).

The biological activity as a sensitive parameter for organic residues transformation is very dynamic and a quality criterion for the soil fertility and aggregate stability. Indicators of the biological activity include microfauna and mesofauna biomass, soil respiration, metabolic quotient ($q\text{CO}_2$), and diverse enzymatic activities. Microbial biomass and activity correlate with carbon and nitrogen contents of the soil, potentially mineralizable nitrogen, soil respiration and enzymatic activity. Soil respiration is most widely used as an indicator of microbial activity and decomposition rate of organic substrates. Estimation of the enzymes involved in the C, N and P cycles allows the calculation of the functional diversity of the microbial communities. The microbial biomass expresses the number of organisms, mainly fungi, bacteria, and protozoon that have a relationship with the flow of energy and nutrient transfers.

Biological and biochemical properties respond quickly to changes in soil management and are sensitive to environmental stress.

In constructed Technosols the biological activity depends on the physical and chemical properties of the substrate's parent material, affecting water retention, water flow and nutrient supply. In GR and VGS a large number of substrates lack well established microbial communities, as most of their components have been sterilized through their thermal production process (commercial materials, e.g., expanded clay) or exhibit poor colonization (e.g., natural perlite, recycled construction debris; Huot et al., 2015; González-Méndez and Chávez-García, 2020). Therefore, to improve plant support capability, deliberate addition of OM, for example with an already active microbial community, microbial inoculums or earthworms, is recommended (John et al., 2017; Jusselme et al., 2019). Autonomous colonization of the substrates by micro and mesofauna is also possible, but it is a slow process.

Parent materials and constructed Technosol composition

A constructed Technosol for GR and VGS must address mainly the suitability for the kind of GI, the target vegetation, the climatic conditions, as well as the availability and cost of the components. There are already a large number of commercial mixtures available in the market that may meet these requirements. However, a non-commercial environmentally-focused Technosol design can address a number of crucial social and environmental issues by prioritizing the use of locally-sourced materials and the incorporation of locally-available wastes. This will limit the negative impacts of natural soil extraction and transportation, and help to close the global nutrient cycle, which is disrupted by the urbanization processes; in addition, it reduces the installation costs.

Among the common components (commercial and non-commercial) used for Technosol construction are natural soils, as well as artifacts (both may provide OM input). Among the artifacts are technogenic materials, particular recycled materials and wastes, as well as semi-natural materials including excavated soil materials (from C horizons) and sediments. These materials are described in more detail in the following paragraphs.

Natural soils and natural mineral materials

The inclusion of natural soil for extensive GR is not frequent, owing to its disadvantages related to weight and drainage clogging, but is used in other GR and VGS. Preferred natural soils are sandy-loam, turfy-soil, sandy-silt, and sandy soils (González-Méndez and Chávez-García, 2020). The advantages of including natural topsoils rely on their content of micro and macro biota, OM and nutrients. The main disadvantage is the environmental impact of the extraction and transportation of large quantities of topsoil from surrounding agricultural or forest areas to vegetate others. For example, in the surroundings of Mexico City, between 45 and $532 \times 10^3 \text{ Mg a}^{-1}$ of organic forest soil (locally called *tierra de monte*) were extracted from 1997 to 2014 (Abbruzzini et al., 2022) and in France, $3 \times 10^6 \text{ m}^3$ of topsoil are used each year (Rokia et al., 2014) to produce ornamental plants and the installation of other GI.

In contrast to natural soils, natural mineral materials obtained in mines can be purchased with defined characteristics and later mixed and optimized according to purpose. Commonly used natural mineral materials for GR and VGS are sand, gravel, clay, and volcanic rocks. Among the later, particularly lava (scoria), pumice, perlite, or any pozzolan material are widely used (Fig. 8A and B). Sand and gravel provide good drainage, but exhibit a large ρ_b when installed. Clay provides in general good nutrient retention, but may clog the drainage. Therefore, clay should be used in small amounts, and sand and gravel preferably in VGS on ground. The

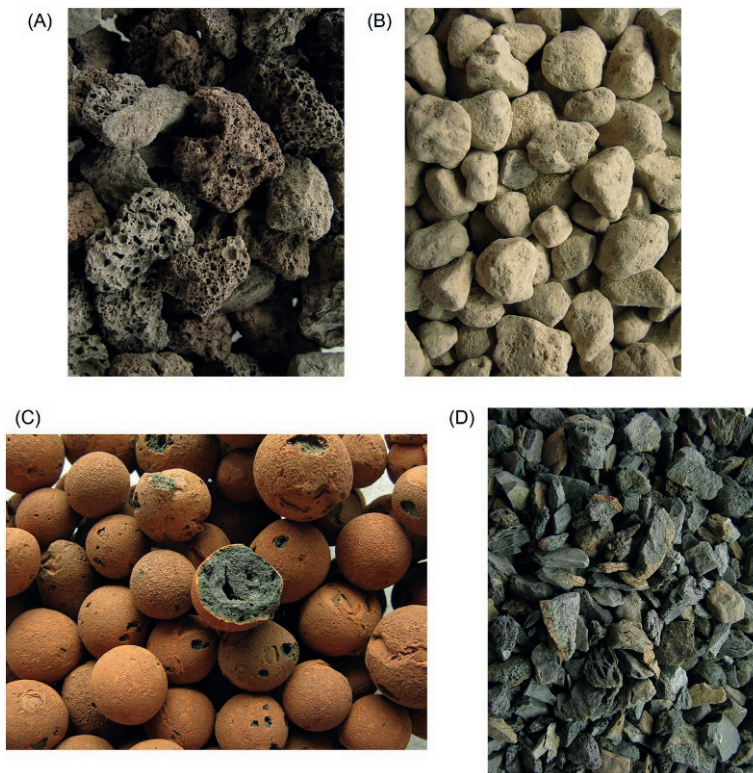


Fig. 8 (A) Lava (scoria), and (B) pumice as examples of porous natural mineral materials. Thermally expanded clay (C) and (D) thermally expanded slate as examples of porous processed natural materials.

porous volcanic rocks commonly increase the AWC of the constructed Technosols. Although some may have closed, occluded, or dead-end pores not available for water storage, like some pumices (Flores-Ramírez et al., 2018), they can be an ingredient of a mixture to improve the drainage, aeration, and lightness of the mixture.

Processed natural materials

Clay, slate, perlite, vermiculite, and glass commonly have large porosity when thermally expanded (Fig. 8B and C). Although for some materials the AWC may be poor, they can be used for drainage and to reduce the mixtures weight (González-Méndez and Chávez-García, 2020). Their production implies a very high energy input as the natural minerals are heated to around 1200 °C, increasing their environmental footprint when produced.

Recycled mineral materials

Mineral wastes are the main part of overall accrued wastes worldwide. They include wastes from construction works, soil excavation and mining operations. Efforts to minimize solid waste disposal in landfills and to encourage recycling are increasingly common (Hendriks and Pietersen, 2000).

Using locally available wastes for the construction of Technosols is a viable strategy to diminish the demand of natural soil for the installation of GI (Abbruzzini et al., 2022). Characteristics of urban wastes differ, and the possibilities and suitability of their use as components for constructed Technosols must be assessed previously.

From the huge diversity of wastes and by-products that can be found in urban areas only some of them are suitable for use as constructed Technosol components, such as construction and demolition waste (C&D) that constitutes around 35% of the total solid waste worldwide (Hendriks and Pietersen, 2000). To use C&D as substrate it needs generally to be processed and crushed. In many countries only the minor share of the overall waste is recycled (Gálvez-Martos et al., 2018), while in others the recycling of C&D is quite common (e.g., in Germany 59.7 t out of 73.9 Mt. of wastes were recycled in the overall construction industry; UBA, 2021). Thus, there is still a big potential in many countries to reuse C&D, such as in GI.

Main criteria to select waste materials for Technosol construction are the following (Damas and Coulon, 2016):

1. Non-toxic for human health or living organisms.
2. Easy to handle: granular solid materials should be preferred (no liquid and pasty wastes) and.

3. Appearance: the wastes have to generate minimal disturbances to the local population (i.e., organoleptic criteria such as smell or dust emission).

Countries may have legal requirements for the incorporation of recycled wastes in GI to be taken in account.

If the wastes fulfill those basic criteria, at least one additional criterion is required. Either they allow vegetation development, have geotechnical properties (such as stability, plasticity and cohesion), or both. Bricks may be a good example of both characteristics, as they are easily penetrated by roots to extract the water stored in their pores (Nehls et al., 2013).

Following the use of C&D and excavated soils, a constructed Technosol can be produced along with the installation of new buildings. Consequently, the general costs of the GI will be reduced and their inclusion in low-income neighborhoods may mitigate some environmental justice issues. Among the most used recycled mineral materials used as components of constructed Technosols for GR and VGS are crushed clay bricks, crushed roof tiles, crushed concrete, building rubble, and soils removed from construction sites.

Natural and recycled organic materials

Common materials providing OM to GR and VGS are compost, biochar, coconut coir, bark and processed sawdust. Compost is produced by the biological decomposition of organic wastes and acts as fertilizer. However, its quality depends on its maturity and the feedstock used. The usage of large volumes should be avoided due to high water retention and increase of the total load during rain events, as well as the possible contribution to eutrophication (Ampim et al., 2010; González-Méndez and Chávez-García, 2020).

Biochar is produced by pyrolyzation of biomass. Biochars generally have high C-content, porosity and CEC. These properties are suitable to increase overall substrate AWC, while decreasing its bulk density. However, their properties are highly variable and mainly depend on the production conditions and feedstock. Additionally, there can be negative effects due to the addition of pollutants (salts, PAHs, dioxins, furans, etc.) or due to nutrient immobilization (González-Méndez and Chávez-García, 2020).

Compressed or composted coconut coir, can be a suitable substrate due to its lightweight. Additionally, its water content when saturated can amount up to the quintuple of its own weight, which has to be considered in the load capacity of the infrastructures (Vijayaraghavan and Raja, 2014).

Sawdust is a low-cost and lightweight material, and may improve AWC, whereas bark may improve the air capacity. However, depending on the tree species, some of their components can be toxic to some seedlings, making a pretreatment such as composting necessary (Ampim et al., 2010).

Rubber crumb, made by grinding of e.g., disposed tires, may improve the properties of coarse-grained growing media by decreasing their bulk density and increasing the water retention (Zhao et al., 2009). However, this kind of material may contain PAH and may emit microplastics (Armada et al., 2022).

Soil amendments such as superabsorbent hydrogels, made out of polyacrylamide or modified cellulose, which are three-dimensional matrixes of highly hydrophilic crosslinked polymers, can be used to improve water retention (Chang et al., 2010).

Substrate composition – defining the optimal ratios

Many of the materials listed above, depending on their availability, costs, properties and durability could be selected to be mixed at different ratios. But how to know which mixture is adequate as a substrate? Guidance on the mixture ratios can be deduced from the bulk density and organic matter requirements established in the regulations indicated above.

The necessary and acceptable ratio of each mineral components of the mixed substrate has to be estimated regarding the weight load capacity (particularly for GR) and the ρ_b of the substrate. Thereby the individual ρ_b and total ρ_b of the mixture have to be considered. The more well-graded the mixture, the greater the expected ρ_b . The same applies to the AWC; however, to a certain point the air capacity decreases. Large proportions of fine particle should therefore be avoided. In a further attempt to optimize substrate mixtures, Willaredt et al. (2022) proposed determining mixing ratios by modeling and evaluating the water retention characteristics of the mixtures based on the water retention characteristics of individual components.

The addition of OM, such as compost, which contains relatively large amounts of nutrients, should be estimated by considering the nutrient demand of the vegetation. For other organic substrates, such as bark and peat, nutrient release occurs more slowly; the added proportion of these should therefore be determined by their effects on physical properties.

The proportion of OM should not exceed 20% (w/w).

Aging and early pedogenesis

As Technosols, GR and green facade substrates are susceptible media prone to early and fast pedogenesis (especially after their installation), which will lead to a change in their physical and chemical properties (Séré et al., 2010; Huot et al., 2015). The mixtures of each created substrate will act as parent materials, and they will determine the presence of specific pedogenetic processes under the influence of other soil-forming factors, such as climate and biota, similar to that in natural soils (Huot et al., 2015). Evidence for

these fast changes have been generally assessed through physical and hydraulic properties, although with contrasting results due to the heterogeneity of the Technosols.

For example, some studies have shown that water retention has increased relative to new substrates because of changes to grain size (e.g., increase of finer particles), number of micro-pores and macro-pores, and greater bulk density over time (Bouzouidja et al., 2018). However, other authors report the lack of influence of the age of GR and green façades on their hydraulic behavior (Mentens et al., 2006). Finally, evidences of an early pedogenesis were highlighted on GR substrates considering fine particles eluviation, organic matter mineralization and modification of soil matrix (Fig. 9; Bouzouidja et al., 2018).

The assembling of knowledge about pedogenesis in GR and VGS based on the relation between physicochemical properties and substrate performances over time is ongoing and has been accelerating for some time, with more information expected in the near future.

Benefits provided by GR and VGS in the urban landscape

The promotion of biodiversity and ecosystem services is an emerging major issue for the renewal of urban areas. GR and VGS contribute to some extent to the improvement in the environmental quality of urban areas, particularly of areas considered as 'dead spaces', which in the case of roofs represent 20–30% total surface after urbanization (Besir and Cuce, 2018). The provision of ecosystem services identified at the beginning of this chapter provided by the GR and VGS mainly result from their capacity to:

- regulate energy consumption and balances of buildings, through shading, evapotranspiration, and thermal insulation
- regulate water fluctuations in urban areas by increasing local water retention and evapotranspiration
- improve air quality
- improve local biodiversity

Energy consumption and energy balance of buildings

Diverse reductions in levels of energy demand of buildings have been reported for different climate conditions and GIs when compared with bare roofs or where walls were poorly insulated. The evapotranspiration rates and thermal properties of GR and VGS are strongly connected to the vegetation species used and to the substrate depth. These GI act as insulators decreasing external heat flow as the growing medium stores energy and the foliage provides shade and evaporates water (Jaffal et al., 2012; Hoelscher et al., 2016). In a mediterranean climate, Silva et al. (2016) monitored the evapotranspiration and assessed the energy balance in

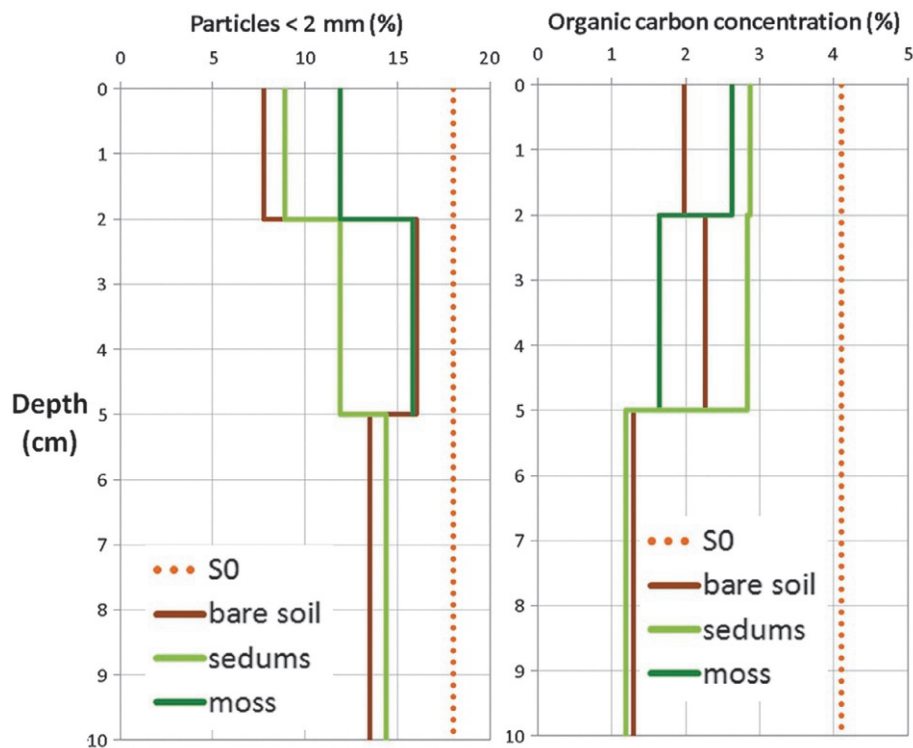


Fig. 9 Evidences of an early pedogenesis on GR substrates considering fine particles eluviation and organic carbon mineralization after 4 years. S0 means original substrate. Adapted from Bouzouidja R, Rousseau G, Galzin V, Clavierie R, Lacroix D, and Séré G (2018) Green roof ageing or isolatic technosol's pedogenesis? *Journal of Soils and Sediments*, <https://doi.org/10.1007/s11368-016-1513-3>.

buildings having an extensive, a semi-intensive or an intensive GR. Buildings with any of the three GR had similar heating needs in winter, but those having semi-intensive and intensive GR needed less cooling energy (36% and 17% of the energy used in the building having an extensive GR, respectively). The three GR performed better than conventional roofs (white and black) in terms of annual energy use in the following order: conventional black roof > conventional white roof = extensive GR > semi-intensive GR > intensive GR (up to 70% less energy than conventional black roofs).

Hoelscher et al. (2016) showed that green façades have a cooling effect on the surface wall temperature, with differences up to 15.5 °C lower on vegetated walls than on bare walls, even performing better than bare white walls with high solar reflectance. The cooling effect is mainly provided by shading more than by plant transpiration.

Flood risk reduction

Green roofs contribute to the urban water management as their substrates store rainwater, thus reducing water runoff and flood risks (Mentens et al., 2006). Kolb (2004) mention a reduction of annual runoff by up to 70%. Moran et al. (2004) reported up to 85% reduction in peak flow rate. The frequency of precipitation and soil water content before rain event play a crucial role in the quantity of runoff delay (Bouzouidja et al., 2018).

Air quality

The macro- and micromorphological characteristics of the vegetation can increase the deposition surface to catch particulate matter, NO₂, and other pollutants improving thus the air quality (Tomson et al., 2021). Therefore, the installation of GR and VGS in adjacent buildings to the street canyons of highly transited roads is an appropriate strategy to improve air quality in urban areas.

Biodiversity

Numerous works show that GR and VGS can generate niches and supplementary habitat for a number of plant species, micro, *meso* and some macrofauna. A wider selection of (preferably native) plant species is recommended, with a further selection of plants that can provide forage and habitat for diverse biota, such as insects and birds (Williams et al., 2014). Also, these GI can be designed with a focus on creating specific microhabitats for target species, such as birds, by including a special topography of the substrate or small water ponds (Fernández Cañero and González Redondo, 2010).

Further increase of the biodiversity in GR and VGS is driven by colonization of new species over time, correlating positively with the proximity of larger patches of greenery and the mobility of the biota (Mayrand and Clergeau, 2018; Williams et al., 2014). However, some authors (Wang et al., 2022; Williams et al., 2014) emphasize that there is a lack of studies, which cover a larger number of taxa, including mollusk, fungi and arthropods, systematized methods to assess the biodiversity in GR and VGS, and a particularly poor understanding of effects on biodiversity in developing countries.

Furthermore, despite the fact that local benefits of individual GI have been confirmed, significantly larger areas covered by GI would be needed to improve the connectivity for species between green area fragments within a city, to act as stepping stones, implying a broad urban design and focused public policies (Mayrand and Clergeau, 2018; Williams et al., 2014). Nevertheless, GR and VGS can, to some extent, function as a mitigation tool for biodiversity loss from urbanization, even if their effectiveness is not comparable with natural habitats or ground-level urban complex artificial greening environments (Wang et al., 2022).

Conclusion

GR and VGS are purposely designed nature-based solutions. Installation of such green infrastructures has substantially increased within the past decade as awareness of stakeholders was raised by the need to reduce energy costs and the environmental impacts of construction design. The substitution of traditional roofs and building envelopes with GR and VGS can contribute to decrease the vulnerability of cities against impacts of climate change because of their contribution to mitigate the urban heat island, the reduction of energy consumption for cooling buildings, the creation of habitats for biodiversity, as well as the regulation of flooding through the runoff reduction.

Purposes for installation of GR and VGS are diverse and can rely on economic (e.g., reduced energy cost, additional esthetic value of real estate), environmental (e.g., regulating building thermal temperature and urban hydrological conditions) or ecological reasons (e.g., creating habitats, or pollutant removal).

The constructional requirements of targeted building determine the design of the installations and the Technosol selection. Substrate depth, its composition and plant species must therefore be appropriately selected to build a viable greenery in urban areas. The environmental footprint of its installation can be markedly reduced by using local and recycled material.

Substrates of GR and VGS require demanding physicochemical properties (e.g., small bulk density, large water retention, adequate nutrient provision) to create proper Technosols with a wide range of specific functionalities. The substrates properties and performances may evolve over time because of their early pedogenesis under the effect of such factors as climate and biodiversity. Such evolution has to be taken into account to assess their middle and long-term sustainability, although the studies of such processes are currently limited, but ongoing.

The GR and VGS can function as a mitigation tool for the biodiversity loss due to urbanization to some extent. A broad urban design and focused public policies aimed to increase the green infrastructure-covered areas would improve the connectivity for species between larger spots, to act as wildlife corridors, leveraging the overall functioning of the urban green infrastructures.

References

- Abass F, Ismail LH, Wahab IA, and Elgadi AA (2020) A review of green roof: definition, history, evolution and functions. *IOP Conference Series: Materials Science and Engineering* 713(1): 012048. IOP Publishing.
- Abbruzzini TF, Mora L, and Prado B (2022) Evaluation of Technosols constructed with construction and excavation debris for greenhouse production of ornamental plants. *Journal of Soils and Sediments* 1–12.
- Amim PA, Sloan JJ, Cabrera RI, Harp DA, and Jaber FH (2010) Green roof growing substrates: Types, ingredients, composition and properties. *Journal of Environmental Horticulture* 28(4): 244–252.
- Armada D, Liompart M, Celeiro M, Garcia-Castro P, Ratola N, Dagnac T, and de Boer J (2022) Global evaluation of the chemical hazard of recycled tire crumb rubber employed on worldwide synthetic turf football pitches. *Science of the Total Environment* 812: 152542.
- Besir AB and Cuce E (2018) Green roofs and facades: A comprehensive review. *Renewable and Sustainable Energy Reviews* 82: 915–939.
- Bouzouidja R, Rousseau G, Galzin V, Clavier R, Lacroix D, and Séré G (2018) Green roof ageing or isolatic technosol's pedogenesis? *Journal of Soils and Sediments*. <https://doi.org/10.1007/s11368-016-1513-3>.
- Chang C, Duan B, Cai J, and Zhang L (2010) Superabsorbent hydrogels based on cellulose for smart swelling and controllable delivery. *European Polymer Journal* 46(1): 92–100.
- Damas O and Coulon A (2016) *Créer des sols fertiles: du déchet à la végétalisation urbaine, Create Fertile Soils: From Waste to Urban Greening*, Paris, Editions Le Moniteur 335.
- Fernández Cañero R and González Redondo P (2010) Green roofs as a habitat for birds: A review. *Journal of Animal and Veterinary Advances* 9(15): 2041–2052.
- FLL – Landscape Development and Landscaping Research Society e.V (2018) *Guidelines for the Planning, Construction and Maintenance of Green Roofs*. 0–120.
- Flores-Ramírez E, Abel S, and Nehls T (2018) Water retention characteristics of coarse porous materials to construct purpose-designed plant growing media. *Soil Science & Plant Nutrition* 64: 181–189. <https://doi.org/10.1080/00380768.2018.1447293>.
- Gálvez-Martos JL, Styles D, Schoenberger H, and Zeschmar-Lahl B (2018) Construction and demolition waste best management practice in Europe. *Resources, Conservation and Recycling* 136: 166–178.
- González-Méndez B and Chávez-García E (2020) Re-thinking the Technosol design for greenery systems: Challenges for the provision of ecosystem services in semiarid and arid cities. *Journal of Arid Environments* 179: 104191.
- Haaland C and van den Bosch CK (2015) Challenges and strategies for urban green-space planning in cities undergoing densification: A review. *Urban Forestry & Urban Greening* 14: 760–771. <https://doi.org/10.1016/j.ufug.2015.07.009>.
- Hendriks CF and Pietersen HS (eds.) (2000) *Report 22: SUSTAINABLE raw materials: construction and demolition waste—state-of-the-art report of RILEM technical committee 165-SRM*. In: vol. 22. RILEM Publications.
- Hoelscher M-T, Nehls T, Jänicke B, and Wessolek G (2016) Quantifying cooling effects of facade greening: Shading, transpiration and insulation. *Energy and Buildings* 114: 283–290.
- Huot H, Simonnot MO, and Morel JL (2015) Pedogenetic trends in soils formed in technogenic parent materials. *Soil Science* 180(4/5): 182–192.
- IUSS Working Group WRB (2022) *World Reference Base for Soil Resources. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*, 4th edn. Vienna, Austria: International Union of Soil Sciences (IUSS).
- Jaffal I, Ouldoukhite S, and Belarbi R (2012) A comprehensive study of the impact of green roofs on building energy performance. *Renewable Energy* 43: 157–164.
- John J, Kernaghan G, and Lundholm J (2017) The potential for mycorrhizae to improve green roof function. *Urban Ecosystem* 20: 113–127.
- Jusselme MD, Pruvost C, Motard E, Giusti-Miller S, Frechault S, Alphonse V, et al. (2019) Increasing the ability of a green roof to provide ecosystem services by adding organic matter and earthworms. *Applied Soil Ecology* 143: 61–69.
- Kolb W (2004) Good reasons for roof planting – Green roofs and rainwater. *International Conference on Urban Horticulture, Acta Horticulturae* 643: 295–300.
- Liu Z, He C, Zhou Y, and Wu J (2014) How much of the world's land has been urbanized, really? A hierarchical framework for avoiding confusion. *Landscape Ecology* 29: 763–771. <https://doi.org/10.1007/s10980-014-0034-y>.
- Maes J, Barbosa A, Baranzelli C, Zulian G, Batista e Silva F, Vandecasteele I, Hiederer R, Lique C, Paracchini ML, Mubareka S, Crisiani CJ, Castillo CP, and Laval C (2015) More green infrastructure is required to maintain ecosystem services under current trends in land-use change in Europe. *Landscape Ecology* 30: 517–534. <https://doi.org/10.1007/s10980-014-0083-2>.
- Manso M and Castro-Gomes J (2015) Green wall systems: A review of their characteristics. *Renewable and Sustainable Energy Reviews* 41: 863–871.
- Mayrand F and Clergeau P (2018) Green roofs and green walls for biodiversity conservation: a contribution to urban connectivity? *Sustainability* 10(4): 985.
- Mentens J, Raes D, and Hermans M (2006) Green roofs as a tool for solving the rainwater runoff problem in the urbanized 21st century? *Landscape and Urban Planning* 77: 217–226.
- Moran A, Hunt B, and Jennings G (2004) A North Carolina Field Study to Evaluate Green roof Quantity, Runoff Quality, and Plant Growth. In: *2nd Greening Rooftops for Sustainable Communities Conf.*, Portland, 2–4 June 446–460.
- Nehls T, Rokia S, Mekiffer B, Schwartz C, and Wessolek G (2013) Contribution of bricks to urban soil properties. *Journal of Soils and Sediments* 13: 575–584.
- Rokia S, Séré G, Schwartz C, Deeb M, Fournier F, Nehls T, Damas O, and Vidal-Beaudet L (2014) Modelling agronomic properties of Technosols constructed with urban wastes. *Waste Management* 34: 2155–2162. <https://doi.org/10.1016/j.wasman.2013.12.016>.
- Scalenghe R and Marsan FA (2009) The anthropogenic sealing of soils in urban areas. *Landscape and Urban Planning* 90(1–2): 1–10.
- Séré G, Schwartz C, Ouvrard S, Renat JC, Watteau F, and Morel JL (2010) Early pedogenic evolution of constructed Technosol. *Journal of Soils and Sediments* 10: 1246–1254. <https://doi.org/10.1007/s11368-010-0206-6>.
- Silva CM, Gomes MG, and Silva M (2016) Green roofs energy performance in Mediterranean climate. *Energy and Buildings* 116: 318–325.
- Tomson M, Kumar P, Barwise Y, Perez P, Forehead H, French K, et al. (2021) Green infrastructure for air quality improvement in street canyons. *Environment International* 146: 106288.
- Tzoulas K, Korpela K, Venn S, Yli-Pelkonen V, Kazmierczak A, Niemela J, and James P (2007) Promoting ecosystem and human health in urban areas using Green Infrastructure: A literature review. *Landscape and Urban Planning* 81(3): 167–178.
- UBA (2021) *Umweltbundesamt Deutschland, 2021; 12. Monitoring-Bericht Kreislaufwirtschaft Bau*.
- Vijayaraghavan K and Raja FD (2014) Design and development of green roof substrate to improve runoff water quality: Plant growth experiments and adsorption. *Water Research* 63: 94–101.
- Wang L, Wang H, Wang Y, Che Y, Ge Z, and Mao L (2022) The relationship between green roofs and urban biodiversity: A systematic review. *Biodiversity and Conservation* 31(7): 1771–1796.
- Willaredt M, Peters A, and Nehls T (2022) Predicting water retention curves for binary mixtures—concept and application for constructed technosols. *Hydrology and Earth System Sciences Discussions* 1–23.
- Williams NS, Lundholm J, and Scott MacIvor J (2014) Do green roofs help urban biodiversity conservation? *Journal of Applied Ecology* 51(6): 1643–1649.
- Zhao S, Wang L, and Duo L (2009) Effects of waste crumb rubber on medium characters and growth of *Lolium perenne* L. *Pakistan Journal of Botany* 41(6): 2893–2900.

Management of park areas, sport fields, and school yards, including golf courses and public right of ways

Maxine J Levin^{a,*} and Anna Paltseva^b, ^aUniversity of Maryland, Silver Spring, MD, United States; ^bEnvironmental Science, SL Mid-Winter Fair Endowed Professorship in Renewable Resources, University of Louisiana at Lafayette, Lafayette, LA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	489
Statement of problems	489
Case studies	489
High Line Park, New York City	490
Camden Yards Sports Complex in the City of Baltimore, Maryland	491
Imitated natural marshland in Southeastern Louisiana	492
Soil considerations for sports fields, public access and open spaces	493
Sport fields—Pitches	493
Vegetation cover on golf courses	497
Soil quality and potential problems	497
Understanding types of drainage	497
The importance of soil quality and organic matter content	497
Soil aeration and compaction-soil management and cultivation	498
Compaction and turfgrass cultural intensity	498
Turfgrass drought survival	498
Management tips to improve turfgrass drought survival	499
Brown or straw colored (dormant) lawn-turf strategies	499
Additional strategies for dormant turf	499
Nutrient management	500
Considerations for nitrogen (N)	500
Plant selection based on soil pH to minimize need for fertilizer and pH adjustment	501
The science behind turf management	502
Park soil contamination and remediation	503
Conclusion	503
References	503
Further reading	504

Abstract

This article synthesizes management considerations for environmentally sustainable stormwater control, maintenance, and construction of park areas, including golf courses and athletic fields. Recommendations cover landscape analysis, grading techniques, soil evaluation for plant selection, understanding soil types and water movement, and using the right tools and techniques to solve drainage issues. Examples include stormwater facility solutions using natural sustainable landscape design, as well as the construction principles and best infield mix for sports fields. Additionally, the article highlights challenges in managing vegetation along highways and rights-of-way. As synthetic fertilizers and pesticides are being phased out, park facilities and sports field managers must prioritize safety measures while maintaining acceptable playing surface quality and function.

Key points

- Management considerations for environmentally sustainable stormwater control, maintenance, and construction of park areas and sports fields.
- Recommendations include analyzing soil conditions, landscape surface grading techniques for water control in public areas and sports fields, evaluating soil conditions for plant selection, understanding how water moves through different types of soil, and applying the right tools and techniques to solve drainage issues.
- Examples of sustainable stormwater facility solutions, essential construction principles for the best infield mix for sports fields, and managing vegetation along highways and rights-of-way.
- Safety measures are important when reducing traditional inputs and maintaining acceptable playing surface quality on sports fields, golf courses, and public park thoroughfares.

*Retired USDA NRCS Soil Interpretation Leader.

Introduction

Soil management of public green spaces is a complex issue that involves creating functional and visually appealing environments while maintaining sustainable and non-toxic conditions. This is a balancing act for managers, who must consider soil type, quality, and conditions to determine stormwater management plans, plant selection, pesticide and herbicide use, and potential building sites.

Nature-based solutions have been recognized as more effective and cost-efficient than artificial alternatives. According to reports by ARUP, London, IUCN, and others in the World Economic Forums (2020, 2022, 2023), these solutions can be 50% more cost-effective on average and bring additional benefits such as improving resilience, boosting the health of citizens, and contributing to a safer and cleaner environment (Cavendish, 2022).

Managers must maintain a critical balance between aesthetics, function, and safety for the environment and human health. The public expects clean and green spaces that are free from toxicity and contribute to surface water pollution or human health safety concerns. Public entities that build and manage these properties are striving to show success in constructing and maintaining sustainable recreational facilities.

Many countries and local governments have developed regulations, ordinances, and guidelines for non-agricultural land management to protect environmental concerns, water quality, and human health. These regulations often require nutrient management plans or integrated pest management plans. Materials and engineering specifications for design and maintenance are also encouraged or regulated for physical human safety. ASTM International is an international standards organization that develops and publishes voluntary consensus technical standards for surface quality and safety design of school yards, playgrounds, sports fields, and golf courses.

To meet demands for safer alternatives, managers are rapidly developing and testing new techniques, such as using organic fertilizers and cutting back on pesticide and herbicide usage, to create well-kept athletic and leisure spaces. This is a challenging task, but with the help of professionals in various fields, it can be achieved.

Statement of problems

Prior to beginning sustainable maintenance planning for all public spaces, such as a park, sports field, golf course, or highway structure, environmental concerns have to be taken into account during the construction phase. The planning process for park and recreation facilities involves three levels: (1) system-wide master planning; (2) site planning for individual facilities; and (3) operational or maintenance planning.

While this chapter focuses on design and maintenance, success is heavily dependent on master planning, which should prioritize sustainability and environmental safety solutions. Master plans for parks should include a developmental history of the park system and changes it has undergone over time. Background information should define the planning area, planning process, and public engagement strategies. Site inventories using GIS mapping, site surveys, visits, and review of historical information can gather data on current environmental conditions, including historical features, environmental issues, soil conditions, vegetation, endangered species, topography, floodplains, wetlands, riparian buffers, and potential contamination from past agricultural or industrial use.

Permitted land uses should also be evaluated to determine if the intended use is allowed within the zoning code of the jurisdiction, together with all state and local ordinances for environmental and water quality protection. If the local jurisdiction does not own the land, deed restrictions or easements should be researched to determine any restrictions on land use, regardless of the landlord.

An inventory and analysis of recreation programs and park facilities may occur during site visits, focus group sessions, and interviews with municipal staff members, parks and recreation board members, and other recreation providers and stakeholders. This analysis identifies maintenance issues, risk management issues, and preventive maintenance needs.

Map data can be collected to determine current and potential future locations for parks, open spaces, and natural areas, as well as passive and active recreation sites. Any barriers associated with architecture, communication, transportation, or service that may limit use by people of various abilities should be identified and addressed through state or other acts, such as the Americans with Disabilities Act (ADA), with self-evaluation and the development of an action plan to overcome them.

When a strategic plan is in place and maintenance issues, preventive maintenance needs, and risk management issues are identified, soil-related aspects become a top priority. These include stormwater management, infiltration, and drainage; drainage and surface “skin” soil quality; grading issues; nutrient runoff/leaching control; compaction and aeration for turfgrass open areas; and maintenance issues for fertility, drainage, plant selection, and runoff.

To address human health and safety issues, ASTM Standards should be considered and incorporated into initial designs and maintenance plans for public spaces.

Case studies

The following case studies are examples of how stormwater management issues, drainage, runoff, and water quality in public areas have been addressed following identifying problems during strategic planning. The active consideration of ASTM standards was central in planning and design, but public opinion and function guided these innovative solutions.

High Line Park, New York City

Numerous case studies illustrate how particular cities have worked with both hardscape (the non-living, man-made elements of a landscape, e.g., pavement, buildings, walls, and other structures made of concrete, brick, or stone) and greenscape (the living, natural elements of a landscape, such as grass, trees, shrubs, and other plants) infrastructure solutions to create a system of dry and safe public areas. Soil analysis and management are the core of these systems. Parks, gardens, and green spaces are the first contacts for environmentally sustainable practices that city dwellers have within urban environments and the opportunity to provide for themselves and surrounding landscapes. Control of runoff and water quality from sealed lands (i.e., parking lots, roads, housing) and use of the space that was once an industrial or contaminated landfill have been taken on by cities through green space planning and sustainable maintenance to protect and conserve human health and welfare. The New York City High Line Park and transport corridor are prime examples of this ethic. The park is a green structure that repurposes formerly industrial infrastructure into a public space. Environmentally sustainable soil management is part of this highly successful renovation design and maintenance program, which took place between 2004 and 2009. Its fascinating history and development with NYC government and community support can be found with pictures at History | The High Line <https://www.thehighline.org/history/>.

Soil management on the High Line is integrated with plant selection, watering, pest management, and cleaning to produce a safe yet pleasing environment for thousands of tourists and locals who walk the Line daily. Planting beds contain formulated soil material that provides structure and consistency for adequate infiltration and a non-toxic stable planting medium. Native, drought-tolerant, and low-maintenance species sourced from a 100-mile radius make up the plant selection of the park, reducing the resources that go into the landscape and providing food and shelter for wildlife species (e.g., native pollinators). About 50% of the High Line's plants are native species, with many of them produced by local growers, making them more adaptable to the local climate, reducing the amount of plant failure and replacement costs. The gardens are diverse and ever-changing, with more than 15 distinct planting zones and 110,000 plants. The High Line's green roof system with drip irrigation allows the planting beds to hold as much water as possible. Combined with the drought tolerance of the High Line's plantings, the gardens need little supplemental watering. Garden waste is composted on-site, reducing the inputs to the waste stream and the need for energy-intensive transport to a separate compost facility. This, in turn, allows the recycling of essential nutrients in the soil to avoid supplemental fertilizer (Figs. 1 and 2).

The High Line, a public park in New York City, has implemented an integrated pest management program that relies on plant species well-suited to the climate to resist pests and disease. To avoid the use of pesticides in such a densely populated area, pests and infested plant materials are removed by hand. The park also prioritizes providing shelter for pollinators, such as wild bees, by



Fig. 1 Historical view of Highline railroad 2011 by Joel Sternfeld. <https://www.thehighline.org/history/>.



Fig. 2 The High Line park, NYC in winter. Photo by Anna Paltseva.

preserving fallen plant debris and dried stalks and stems, and incorporating as many native plants as possible. Additionally, the park mandates the use of eco-friendly cleaning solutions and post-consumer paper products for maintenance needs. Due to its elevated and narrow structure, the park faces challenges in snow removal following winter storms. Therefore, snow is removed by hand using shovels, power equipment, and eco-friendly ice melting products that are safe for the environment.

Camden Yards Sports Complex in the City of Baltimore, Maryland

Similarly, the Camden Yards Sports Complex in Baltimore, Maryland, underwent a \$5-million renovation of its pedestrian spine, known as Ravens Walk, to improve stormwater management while preserving the visitor experience. The project includes enhanced security features, new hardscaping, impervious surfaces, new landscaping, and stormwater improvements. The renovation was completed in 2019 and involved managing runoff from large parking lots to prevent parking lot contaminants from entering nearby waterways. This project is part of a larger trend of stormwater-centered investments by managers of professional sports venues across the United States, including those in Atlanta, Georgia; San Francisco, California; and Minneapolis, Minnesota (Figs. 3 and 4).

To improve stormwater management, several measures were implemented around the walkway at Camden Yards Sports Complex. The first step was to plant trees and small gardens to create a natural buffer against the flow from the parking lots. Additionally, an underground biofiltration system was installed, designed by Bio Clean Environmental Services (Carlsbad, California). Large drains were strategically placed to collect runoff, which then flows through a specialized filtration medium that simulates a natural wetland's absorption power. This treatment process helps remove pollutants, and the treated stormwater flows at a controlled rate into the surrounding area, meeting pollutant-removal benchmarks without requiring any input from surface-level plants. To provide comprehensive stormwater management with minimal above-ground infrastructure, five of these biofiltration units were installed around the Pedestrian Spine.



An innovative, underground biofiltration system helps the walkway discourage flooding and prevent water pollution. Image courtesy of Bio Clean Environmental Services.



Fig. 3 Trees and Gardens planted as natural buffer from parking lots at Camden Yards. Image courtesy of Anna Paltseva.

Imitated natural marshland in Southeastern Louisiana

Southeastern Louisiana University (Mandeville, LA) provides an excellent example of a relatively low-tech solution for imitating natural marshlands to treat freshwater ponds and water features contaminated by stormwater runoff. By using softscape techniques derived from soil conservation and reclamation strategies for shoreline erosion and buffer strips, marshland technology aims to improve water quality throughout St. Tammany parish in Louisiana.

Within the Del Sol subdivision of Covington, LA, a star-shaped, eight-armed frame of polyvinyl chloride piping with 1.2-m (4-ft) wide, vinyl-coated crab wire between the piping acts as a platform and supporting structure for marsh plants such as maidencane, arrowhead, and spider lily. The plants are attached to the netting, forming a natural filtering mechanism for the pond water. The artificial floating marsh removes excess nutrients and turbidity caused by fertilizers, pesticides, and herbicides. This project demonstrated an improvement in water quality in the first year, becoming much clearer with a considerable reduction in turbidity. Additionally, the marshland frames create a habitat for minnows and juvenile fish, significantly improving the ecosystem functioning of the pond.



Fig. 4 Underground drainage system at Camden Yards. Image courtesy of Anna Paltseva.

Similar applications have great potential for sustainably creating a more natural ecosystem associated with water features in golf courses or stormwater retention ponds in parks and open spaces. Similar structures have been experimentally installed in harbor public land viewscapes such as in the harbor of the City of Baltimore to improve water quality, fish nursery habitat, and public awareness (Fig. 5).

Soil considerations for sports fields, public access and open spaces

Sport fields—Pitches

A pitch's surface can be made of sod (grass), artificial turf, sand, clay, gravel, concrete, or other materials. Pitch care or turf management refers to the maintenance work required to keep a sporting pitch ready for use. Different types of sporting pitches present unique challenges, and the protocols and specifications for maintenance vary considerably depending on the sport and the use of artificial surfaces. Unique skill sets and approaches based on the soil are also needed to care for sand-based athletic fields or native soil fields. In most cases, regardless of the sport, there are four main attributes of a good grass pitch: good drainage, adequate grass coverage, a flat surface, and a low level of weeds.

Historically, soccer and American football pitches have had natural grass covers. However, the stresses on a pitch and winter weather lead to the pitch being returfed regularly. Typically, the existing turf is removed to a depth of 40 mm turf and 110 mm of soil, and the replacement turf is grown to ensure consistency and weed-free ground. A pitch can be returfed within four days, which involves removing and re-laying 400 cubic meters of turf and soil.

Football pitch technology has advanced considerably since the 1970s and 1980s, when almost all pitches turned into mud baths by December. Nowadays, technology includes drainage pipes at 5 m spacing, a gravel raft, a sand (suspended water table) rooting zone, undersoil heating, and additional lighting on the surface stimulating growth. Management techniques have also advanced, with more emphasis on soil and plant biology, morphology, zoology, and physiology being introduced into turf grass and soil management. There are lots of materials and methods available to reinforce a natural grass playing surface. Artificial grass offers an alternative to natural grass for football stadiums. The performance of this surface has often been questioned as not being truly natural. However, artificial grass has definite advantages, especially when a stadium has heavy or multi-use requirements. Artificial grass also has a benefit in settings hostile to natural grass, such as low sunlight intensity or water scarcity (Fig. 6).

The key specifications for surface “skin” soil management of recreational areas developed for the United States are summarized in Table 1. The specifics of these guidelines are beyond the scope of this chapter.



Fig. 5 Baltimore Harbor creates habitat for minnows and stabilization, water quality features similar to that described in pilot project in Southeastern Louisiana. Photo by Maxine Levin.



Fig. 6 Multi-use natural grass and sport field system for public access in Baltimore Harborplace area using ASTM standards for surface “skin” and engineered drainage on reclaimed industrial areas and right of ways. Photo by Maxine Levin.

Table 1 Major ASTM specifications and guidelines for surface soil management considerations on sports fields and playgrounds^a.

Guideline	Content
ASTM F 2107-08 (2020) Standard Guide for Construction and Maintenance of Skinned Areas on Baseball and Softball Fields	This guide covers techniques that are appropriate for the construction and maintenance of skinned areas on baseball and softball fields. This guide provides guidance for the selection of materials, such as soil, sand, gravel, crushed stone, crushed brick, calcined clay, diatomaceous earth, vitrified clay, etc., for use in constructing or reconditioning skinned areas and for the selection of management practices that will maintain a safe and playable skinned surface. Although this guide is specific to baseball or softball, it has application to other sports where ball bounce, ball roll, or player footing, or a combination thereof, are of importance
ASTM F 2270-12 (2018) Standard Guide for Construction and Maintenance of Warning Track Areas on Sports Fields	This guide covers techniques that are appropriate for the construction and maintenance of warning track areas on sports fields. This guide provides guidance for the selection of materials, such as soil and sand for use in constructing or reconditioning warning track areas and for selection of management practices that will maintain a safe and functioning warning track. Although this guide has applications to all sports where a warning track surface may be required or desired, it has specific applications to baseball/softball
ASTM F 2000-19 Standard Guide for Fencing for Baseball and Softball Fields	This guide provides recommended minimum requirements for various types of fences used in softball and baseball ballfields and other sports facilities, and practices for installation
ASTM F 1938-98 (2017) Guide for Safer Use of Movable Soccer Goals	This guide presents directions for the installation, use, and storage of full-size or nearly full-size movable soccer goals. It is expected that these guidelines can help prevent deaths and serious injuries resulting from soccer goal tipover. These guidelines are intended for use by parks and recreation personnel, school officials, sports equipment purchasers, parents, coaches, and any other members of the general public concerned with soccer goal safety
FIELDS—NATURAL GRASS/ROOTZONE	This guide covers techniques that are appropriate for the construction of high performance sand-based rootzones for sports fields. This guide provides guidance for the selection of materials, including soil, sand, gravel, peat, and so forth, for use in designing and constructing sand-based sports turf rootzones
ASTM F 2396-11 (2019) Standard Guide for Construction of High Performance Sand-Based Rootzones for Sports Fields	
ASTM F 1632-03 (2018) Standard Test Method for Particle Size Analysis and Sand Shape Grading of Golf Course Putting Green and Sports Field Rootzone Mixes	This test method covers the determination of particle size distribution of putting green and other sand-based root- zone mixes. Particles larger than 0.05 mm (retained on a No. 270 sieve) are determined by sieving. The silt and clay percentages are determined by a sedimentation process, using the pipet method. This procedure was developed for putting green rootzone mixes, those assumed to have sand contents of 80% by weight or greater. Particle size analysis of soils may be performed by this test method or Test Method D 422. This test method also describes a qualitative evaluation of sand particle shape
ASTM F 1647-11 (2018) Standard Test Methods for Organic Matter Content of Athletic Field Root Zone Mixes	These test methods cover the determination of the percent organic matter of a putting green root zone mixture using a loss on ignition method or the Walkley Black method. These test methods are useful for quantifying the organic matter content of volume ratio mixed root zone mixes. Test Methods D 2974 is recommended for peat and other organic soils
ASTM F 1815-11 (2018) Standard Test Methods for Saturated Hydraulic Conductivity, Water Retention, Porosity, and Bulk Density of Athletic Field Root Zones	These test methods cover the measurements of saturated hydraulic conductivity, water retention, porosity (including distribution of capillary and air-filled porosity at a known soil suction), and bulk density on sand-based root zone mixes to be used for construction and topdressing of golf course putting greens including United States Golf Association (USGA) recommended greens, golf course tees, sand-based sports fields, or other highly trafficked turfgrass areas. These test methods are designed for sand-based mixes and are not intended for use with fine or medium textured soils, for example, sandy loams and loams
ASTM F 2060-00 (2018) Standard Guide for Maintaining Cool Season Turfgrasses on Athletic Fields	This guide covers the minimum requirements for maintaining cool season turfgrasses used for natural surface athletic fields. Practices covered include mowing, fertilizer application, irrigation, core cultivation, overseeding, and pest management
ASTM F 2269-11 (2018) Standard Guide for Maintaining Warm Season Turfgrasses on Athletic Fields	This guide covers the minimum requirements for maintaining warm-season turfgrasses used for natural surface athletic fields. Practices covered include mowing, fertilizer application, irrigation, core cultivation, winter overseeding, pest management, and requirements for management of dormant turf winter overseeded with cool-season turf (see also Guide F 2060)
FIELDS—SYNTHETIC TURF	These test methods establish a recommended list of test methods to be used for the identification of physical property characteristics and comparison of the performance properties of synthetic turf systems or components for athletic and recreational uses, or both
ASTM F 1551-09 (2017) Standard Test Methods for Comprehensive Characterization of Synthetic Turf Playing Surfaces and Materials	
ASTM F 2765-014 Standard Specification for Total Lead Content in Synthetic Turf Fibers	This specification applies to the maximum content of lead in fibers used in synthetic turf. This specification outlines a test method for sample preparation and a test method for analyzing the total lead content in synthetic turf fibers. This specification outlines guidelines for reporting total lead content in synthetic turf fibers. This specification applies only to synthetic turf fibers manufactured after September 1, 2009. Note 1—It is the goal of the industry to reduce lead content to 100 mg/kg (ppm) by September 1, 2011

(Continued)

Table 1 (Continued)

<i>Guideline</i>	<i>Content</i>
ASTM F 1015-03 (2017) Standard Test Method for Relative Abrasiveness of Synthetic Turf Playing Surfaces	This test method is applicable to both laboratory and field measurement of synthetic turf surfaces used for sports. Data obtained from the procedure of this test method are indicative of the relative abrasiveness of fabric or carpet type synthetic playing surfaces
ASTM F 2898-11 (2019) Standard Test Method for Permeability of Synthetic Turf Sports Field Base Stone and Surface System by Non-confined Area Flood Test Method	This test method may be used to determine the permeability rate of synthetic turf playing field systems, playing field systems with pad or premolded drainage boards, or both, playing field systems with premolded panel base systems, porous and non-porous pavement systems, or base stone systems in the field, or a combination thereof, by non-confined area flood test method. This system is suitable for use on the finish synthetic turf playing surface and on the stone base system below the playing system
FIELDS—IMPACT ATTENUATION	This test method covers the measurement of certain shock-absorbing characteristics, the impact force-time relationships, and the rebound properties of playing surface systems. This test method is applicable to natural and artificial playing surface systems and to components thereof. Typical playing surfaces are wrestling mats, football fields, soccer fields, playgrounds, and so forth
ASTM F 355-16e1 Standard Test Method for Impact Attenuation of Playing Surface Systems, Other Protective Sport Systems, and Materials Used for Athletic, Recreation and Play	
ASTM F 1936-19 Standard Specification for Impact Attenuation of Turf Playing Systems as Measured in the Field	This specification establishes an in situ test method and maximum impact attenuation value for all types of turf playing systems and for a number of sport-specific field layouts. It also includes a protocol for determining test point locations on fields that are lined for multiple sports
ASTM F 1702-10 (2018) Standard Test Method for Measuring Impact-Attenuation Characteristics of Natural Playing Surface Systems Using a Lightweight Portable Apparatus	This test method is used to determine the impact-attenuation characteristics of natural turfgrass and soil playing surface systems with a lightweight portable apparatus. This test method can be used to compare the impact attenuation characteristics of natural playing surface systems, as well as assessing the effects of management practices on the impact attenuation characteristics. This test method also can be used to assess the compactibility of natural playing surfaces by recording g-max values or penetration of successive impacts, or both. This test method provides a procedure for assessing impact attenuation characteristics in the field, on both actual playing surfaces and research plots. Numerical data will not be comparable to data obtained using a different missile mass or geometry, different drop height, or different standard method, for example, Test Method
ASTM F 2650-17e1 Standard Terminology Relating to Impact Testing of Sports Surfaces and Equipment	This terminology covers terms related to impact test methods and impact attenuation specifications of sports equipment and surfaces

^aThe complete list and guidance documents are available electronically through the ASTM International website or American Sports Builders Association. Relevant ASTM Standards—Fields (sportsbuilders.org) <https://sportsbuilders.org/page/RelevantASTMStandards-fields>).

Vegetation cover on golf courses

The increasing popularity of golf has led to the development of turf management, which refers to the skills needed to maintain a golf course. The green and tee box are the principal areas of concern, and many golf courses are now built in environments that are hostile to natural grass cover. Essentially, the grass grows in a hydroponic or sterile environment with speedy drainage, requiring regular feeding and watering. The key characteristics of a good green are speed and consistency. Faster greens are preferred, and for tournament play, the greens should be as fast as possible. The main factors influencing green speed are:

- Mowing height, which today is typically between 3 and 6 mm. Top courses are striving for even lower cutting heights due to the demand for faster greens. In the 1960s, green mowers were not capable of cutting below 6 mm, but nowadays, they can cut below 3 mm.
- Frequency of mowing. During the growing season, daily mowing is required. For faster speeds and professional tournaments, greens may be cut twice a day or even double cut (two cuts in two directions, one immediately following the other).
- Rolling can also improve green speed, but excessive rolling can compact the soil.
- Topdressing helps with consistency and reduces thatch buildup.
- Grooming reduces the thickness of clumps.
- Verticutting removes excessive thatch, forces the grass blades to stand upright, thins out excessive growth, and speeds up greens.
- Scarification removes moss and prevents moisture collection on the green.
- Beneficial bacteria, fungi, and nematodes are used as stand-alone applications or in integrated programs to control key turf pests and diseases.
- Aerating removes excessive organic matter, modifies the rootzone composition, and improves rooting and drainage.

The rough and fairway present the best opportunity for sustainable, clean, and green maintenance in golf course design. Functionality is more about the practical attributes of the turf, such as playability, slope stabilization, or runoff reduction. Performance equals aesthetics and function. Expectations vary widely, and each management scenario may prioritize different aspects of turf performance:

- Example #1, a high-end private golf course: aesthetics and function are likely to be top priorities. Such sites are managed for maximum playability and are simultaneously expected to be green and manicured in appearance.
- Example #2, an earthen dam: appearance takes a back seat in favor of efforts to ensure that a hardy, perennial cover is in place to minimize the potential for adverse runoff or soil erosion.

Golf course planning can use the fairway and sightline, as well as the rough, water features, and sand traps, to take on aspects of the natural landscape and add to the sustainability outcomes with low maintenance and water needs, water and sediment retention, and no inputs for fertilizer and pesticides. These features allow for the propagation of native species and pollinators without sacrificing function on the high-maintenance tee box and greens.

Soil quality and potential problems

Understanding types of drainage

Evaluating drainage on a local scale, such as in gardens and sports fields, requires an assessment of the ability of the surface soil to drain water. Natural drainage from the soil profile drives any design alterations needed to create dry surfaces for play and maintenance of green space. There are two types of drainage to consider: sub-surface and surface.

Sub-surface drainage is important to prevent water from accumulating on the surface of unsealed fields during heavy rain. This is typically achieved by adding a layer of mixed gravel or cinders about 45–60 cm below the surface of the seedbed. This layer has high porosity, allowing moisture to be swiftly moved away from the playing area. However, it also prevents moisture from moving up from below the surface. Proper irrigation facilities can minimize the need for upward movement of moisture, and the adverse effects of sub-surface drainage are negligible.

Surface drainage is achieved by maintaining an even surface level, free of compaction and depressions that can cause standing water to pool. The standard is to create a slightly sloping surface of about 1–1.5 cm per meter to allow for surface runoff and maximum infiltration. This technique has been used throughout the United States as a standardized manner of preparing surfaces for well-drained yet green sports fields.

The importance of soil quality and organic matter content

In addition to considering the drainage and surface grading, it's also important to pay attention to the quality of the soil "skin" of a playfield or pitch (field of play), especially if there is green turf. The soil quality and organic matter content affect infiltration and seedbed viability, which can impact player safety. The composition of the seedbed is critical because it will be exposed to the elements and heavy impacts. Ideally, the seedbed should consist of around 10%–15% organic matter, 10%–20% coarse sand, 10% clay, and the remaining sandy loam with clay content not exceeding 15%.

Soil aeration and compaction-soil management and cultivation

Soil management and cultivation is crucial for maintaining healthy soil for plants to grow. Aeration, which involves replacing soil air with air from the atmosphere, is essential in sustaining adequate soil oxygen for root growth and other growth-related activities. However, soil compaction caused by vehicular and foot traffic in parks, sports fields, golf courses, public thoroughfares, and right of ways can be a major contributor to poor aeration, leading to turf management problems and alterations in the physical properties of the soil.

Soil compaction can decrease the total pore space, reducing macropore (large) spaces that are essential for internal drainage, channels for root growth, and air exchange, while increasing micropore (small) spaces that retain water. The latter can also reduce the soil oxygen content, which is necessary for root respiration and growth, and limit nutrient uptake by roots. Furthermore, compacted soils can reduce water infiltration and percolation rates, leading to changes in irrigation practices and scheduling.

Compaction increases soil strength and density, which can reduce rooting density and depth. Additionally, it can increase water retention, prolonging wet soil conditions and delaying spring green-up by delaying soil warming, and intensify compaction in trafficked areas. In contrast, dry compacted soils in summer warm up faster and promote high soil temperatures, dramatically slowing root growth. Therefore, it is important to manage and cultivate soil to prevent soil compaction and promote adequate aeration for plant growth.

Compaction and turfgrass cultural intensity

Soil-related stresses caused by soil compaction can negatively affect the growth of turfgrass, leading to the need for changes in management practices. Different cultivation methods are available to turfgrass managers, including coring techniques such as hollow tine, solid tine, or shatter coring, slicing, and spiking. However, these methods differ in their effectiveness in achieving important cultivation objectives, such as thatch removal, alleviation of soil compaction, and preparations to enhance seed-soil contact for overseeding purposes. Power raking and vertical cutting may reduce thatch but do not reduce soil compaction, which is the primary issue. Hollow tine and solid tine coring methods minimize soil compaction and alleviate thatch by promoting aeration and new growth. Slicing and spiking, commonly used on golf courses, do not improve soil aeration to the same extent as core cultivation practices but are less disruptive to putting surfaces. Coring, especially hollow tine, is more disruptive to surface uniformity and requires a more extended period for recovery. However, it is the most effective method for improving aeration and alleviating compaction. A relatively new technology that uses short bursts of water injected under high pressure has been shown to reduce compaction and improve aeration, equal to or better than hollow tine core cultivation, with less surface disruption.

Soil compaction often necessitates increased fertilizer and pesticide or herbicide use. Low shoot density encourages encroachment of weedy species tolerant of compacted sites, such as annual bluegrass, goosegrass, and knotweed, resulting in greater herbicide use and cost. Reduced turfgrass vigor and favorable moisture conditions can lead to disease activity with brown patch and Pythium blight. Turfgrass nutrient uptake from compacted soils may be reduced by 10%–30%, leading to chlorotic, nitrogen-deficient turf. Additional nitrogen further reduces rooting potential because the symptoms may not necessarily be due to nutrient deficiency but somewhat limited root activity and viability.

Irrigation efficiency drops due to low infiltration rates associated with compacted sites that may promote pooling (ponding) and surface water runoff, reducing the amount of water available to recharge the root zone. Typically, multiple cycling of irrigation water (small amounts of water applied over an extended period using several irrigation events) must be practiced, complicating irrigation scheduling. Due to poor drainage and restricted rooting, small doses of water must be applied more frequently, creating wetter conditions that intensify compaction.

In addition, soil compaction encourages plant tissues to become more succulent and have lower carbohydrate levels (a condition associated with poor drainage and frequent irrigation), reducing environmental stress tolerance (heat, cold, drought, disease, wear). Any factor that contributes to increased tissue succulence (e.g., excess nitrogen, water, shade, and close mowing) combined with compaction will reduce root and shoot response more than any of these factors when considered alone. Additional maintenance is required to reduce or eliminate the symptoms.

With intensive maintenance on groomed tee boxes and greens in golf courses, alleviating soil compaction and improving aeration can have various positive effects, such as the release of gases such as CO₂ (which begins immediately after coring), increased soil infiltration rate, enhanced rooting within core aerifier holes (within 2–3 weeks after coring), decreased thatch accumulation following core cultivation, and increased fertilizer uptake and use due to aeration. Aeration also promotes incorporation of immobile materials such as lime and phosphorus into the root zone.

In summary, soil compaction can have significant negative effects on turfgrass health, and proper cultivation methods such as coring can help alleviate these issues. Improved aeration can have numerous positive effects on turfgrass maintenance, including improved nutrient uptake and irrigation efficiency.

Turfgrass drought survival

When rainfall is insufficient, and water resources become limited, supplemental irrigation is often required to sustain turf growth. Under rainfall deficit conditions, irrigation of landscape plantings such as turf is often among the first activities to be placed on

water use restrictions. In this case, maintaining functional and high-quality turf with less water is required. Guidance from drought-prone areas can inform any area of the country or world to conserve water and yet maintain visual green sightlines and targeted sport function. Conservative irrigation management through drip and selective watering, controlling drainage and runoff, and managed native buffer space are critical for water quality protection. Water conservation strategies are routinely practiced in semi-arid and arid regions such as the southwest desert of the United States. These strategies include (1) incorporating water-use-efficient plant material into the landscape, (2) implementing water-conserving management practices, and (3) maximizing irrigation efficiency by controlling leaching, pooling, or ponding of irrigation water and surface water runoff. The following information from University of Massachusetts Turfgrass extension is a good example of taking specific recommendations from the desert and applying them to drought conditions.

Management tips to improve turfgrass drought survival

1. Mowing and nutrition

To promote healthy turf growth and drought survival, mow at the higher end of the recommended mowing height range for your species and cultivars (around 5–7.5 cm). This will create shade on the soil surface, encourage deeper root growth, and better water retention. Use nitrogen fertilizer in moderation, with at least 30% slow-release nitrogen. During drought conditions, avoid fertilizing non-irrigated turf. Excessive nitrogen can promote fast shoot growth rates and more drought-sensitive tissues, and reduce root depth and number. A moderate green color is healthier than a dark green color. In the fall, make sure to have adequate phosphorus and potassium levels (based on a soil test) to encourage drought recovery. If there is no potassium deficiency, apply potassium at about 50%–75% of the nitrogen applied.

2. Water or irrigation efficiency

To prevent surface runoff and puddling, apply water at a rate that the soil can absorb it. Apply only the amount of water that can be stored in the root zone without leaching. Check the rooting depth of the soil using a garden trowel, shovel or soil probe. In Massachusetts, most turfgrass species require 2.5–3.8 cm of water per week. For low traffic areas where wear tolerance is not a major concern, schedule irrigation to allow for mild to moderate stress (indicated by wilting and leaf rolling or folding) between watering events to promote deeper rooting and drought survival. Avoid over-irrigating as it can cause succulent growth, depletion of soil oxygen, and soggy conditions leading to soil compaction. Address issues that reduce soil infiltration rates like soil compaction, thatch, and hydrophobic soils to improve irrigation efficiency in the fall.

3. Control root-related stresses in the spring and fall

- Soil pH—maintain pH between 6.0 and 7.0 because root growth and activity are limited below pH 5.5 and above pH 7.5.
- Compaction—manage compaction to maximize rooting potential by maintaining optimum aeration (soil oxygen content) and gas exchange (to limit toxic CO₂ levels), minimizing water logging, and prolonging soil wetness.
- Apply herbicides intelligently—preemergence herbicides are more toxic to roots than to shoots. Hence, turfgrass shoots may be unaffected while root activity is severely reduced. An herbicide should be applied only as needed and according to label directions. Spot treat for broadleaf weeds when possible, to limit stress and avoid blanket applications when air temperatures exceed 29.4 °C.
- Thatch—manage thatch to keep levels below 4 cm to promote deeper rooting into the soil and limit rooting confined to the thatch. Thatch generally has poor nutrient and water-holding characteristics.

Brown or straw colored (dormant) lawn-turf strategies

During prolonged periods of little or no water (from precipitation or irrigation), grass growth will cease, resulting in a straw-colored turf. However, the perennial parts of the plant, including crown tissues near or at the soil surface, and nodes on lateral stems (rhizomes and stolons), remain active and capable of regenerating new shoots and roots when there is significant rainfall. With normal summer dormancy, recovery and green-up may take 2 weeks or more, and 100% green-up may not be achieved following extended periods of dormancy (45 to 60+ days) with little or no rainfall. Timely watering and proper management, as previously outlined, can help avoid drought-induced dormancy if high-quality and functional turf is desired.

Additional strategies for dormant turf

- Mowing—don't mow unless necessary to reduce stress; mow high and infrequently; avoid mowing mid-day, late afternoon, and early morning are preferred; use a sharp mower!
- Nutrient application—consider a drought fertilizer application program in the fall to enhance recovery; superior root and stem development (rhizome and stolon) are critical for recovery, and low nitrogen and high potassium rates are helpful (i.e., 15-0-30 type or similar analysis); avoid straight nitrogen applications.
- Irrigation—if irrigation cannot be applied properly and on a timely basis it is preferred to allow the turf to enter dormancy (buds associated with crowns and lateral stems are extremely drought hardy); unirrigated turf in a dormant state produces a

healthy turf in a short time after drought ceases; watering improperly during drought-induced dormancy accelerates the depletion of carbohydrate reserves which reduces drought-recovery potential.

- Pest problems—inspect for disease (summer patch) and insects (chinch bugs) regularly which may go undetected when the turf is dormant and not actively growing; weed pressure and numbers may increase, which compete for limited soil moisture, reduce herbicide usage (spot treat for weeds) during dormant periods.
- Renovations—consider renovations in the fall if drought recovery is poor or thatch levels are excessive (2.5 cm).

Nutrient management

First and foremost, in any evaluation and prescription for best management practices, whether “green and clean” or traditional aesthetic:

- Read the landscape and analyze site and soil conditions;
- Implement proper grading techniques;
- Evaluate soil conditions for plant selection;
- Identify different types of soil and learn how water moves through each; and
- Apply the right tools and techniques to solve drainage, compaction and fertility issues.

The development and implementation of a nutrient management plan (NMP) is critical to sustainable maintenance, with a priority on environmental protection and enhancement. An effective NMP should address sound agronomic practices that promote healthy turf, gardens, and fields, while also protecting water bodies. The following guidelines outline the components of an NMP that can be adapted to a variety of turf management systems. For instance, UMass Extension’s Best Management Practices for Lawn & Landscape Turf manual provides detailed coverage of the components of an NMP specific to lawn turf management in Section 7. This manual can be accessed online at <http://extension.umass.edu/turf/publications-resources/best-management-practices>. Although the manual focuses on lawn and landscape turf, the presented information can be applicable to the management of other types of parkland or right of ways, such as high-value playing surfaces or shallow maintenance turf areas.

A complete NMP should include:

Site analysis and mapping, such as:

- Identification and mapping of environmentally sensitive areas and areas at high risk for off-site movement of nutrients.
- Mapping specific buffer zones delineated in environmentally regulated areas (e.g., Zone I wellheads, wetlands, and certain coastal zones), and for the protection of environmental resources.
- Mapping of areas receiving fertilizer or amendments containing nutrients.
- Soil tests to determine the chemical and physical condition of the soil and ascertain recommendations for adjustments.

Determination of need for nutrient inputs based on soil testing, turf function and quality desired, and proximity to environmentally sensitive sites such as:

- the lowest input rate of nitrogen, phosphorus, and other nutrients will produce dense turf cover and promote deep rooting.
- form (source) of nutrients appropriate for the management plan and the use of the turf on the site.
- appropriate placement of fertilizer and amendments containing nutrients.
- application frequency of fertilizer and amendments containing nutrients based on the nitrogen and phosphorus characteristics of each material.
- timing of fertilizer applications so that maximum nutritional availability corresponds with periods of active turfgrass growth.
- fertilizer and amendments containing nutrients should not be applied before spring green-up or when turf is dormant.
- fertilizer and amendments containing nutrients should not be applied when the ground is frozen.
- fertilizer and amendments containing nutrients should not be used as de-icers.
- Ample available phosphorus (P) is critical to the success of turf establishment, renovations, and major repairs. P levels applied in conjunction with such activities should be adequate for rapid turfgrass germination, growth, and development. Specific considerations for phosphorus (P):
- determination of P levels by a soil test, and where and when the application of phosphorus-containing materials may and may not be allowed by state or local regulations.
- application of phosphorus-containing fertilizers and other nutrient-containing materials, regardless of the source of nutrients or the purpose of the application, should not exceed levels recommended by a soil test.
- accounting for all phosphorus inputs into the management program and turf system, such as fertilizers, turfgrass clippings retained in the system, composts, compost derivatives and other organic amendments, topdressings and other materials.

Considerations for nitrogen (N)

- determination of nitrogen (N) need based on turfgrass species, time of year, turf vigor, plant response, and frequency of turf use.
- determination of N rate based on turf needs as influenced by turf function and use as well as by presence of and proximity to environmentally sensitive areas, keeping N level to the lowest possible required to realize management objectives while protecting the environment.

- use of slow-release sources of N as often as possible based on management objectives.—Accounting of all N inputs into the management program and turf system. These may include fertilizers, retained turfgrass clippings, composts, compost derivatives and other organic amendments, topdressings, corn gluten used as an herbicide or fertilizer; and carryover N from prior seasons.

Plant selection based on soil pH to minimize need for fertilizer and pH adjustment

Plants require specific pH levels for optimal growth. For most ornamental plants, a pH range of 6.0–7.0 is ideal. However, some plants such as azaleas, gardenias, camellias, and related species can tolerate lower pH levels between 4.5 and 5.5. As stated above, turfgrass species, on the other hand, require a pH range of 5.5–7.0 for optimal nutrient uptake and performance. For more information on the optimal pH ranges for ornamental flower trees and bushes, you can refer to Fact Sheet 372 by Mississippi State University Soil pH and Fertilizers | Mississippi State University Extension Service (msstate.edu) Turf species range of pH are listed in Tables 2 and 3 (Bill's Will, 2017).

Adjusting soil pH through the use of lime, sulfur, or fertilizers is feasible to some extent, depending on soil texture, organic matter content, and buffering capacity. However, this method is not always a long-term or effective solution for maintenance, and it may require additives that can negatively impact water quality in runoff and leachate. Furthermore, adding lime to increase soil pH levels may increase the levels of native heavy metals such as arsenic (As), cadmium (Cd), or chromium (Cr), potentially affecting the

Table 2 Soil pH for grass types.

Grass pH ranges	pH-LOW	pH-HIGH
Bahiagrass (<i>Paspalum notatum</i>)	6.0	7.5
Argentine—Pensacola—Tifton-9		
Bentgrass (<i>Agrostis</i>)	5.6	7.0
Colonial and Creeping		
Bermuda (<i>Cynodon dactylon</i>) Hybrids and Common	5.6	7.0
Bluegrass (<i>Poa pratensis</i>)	5.7	7.4
Buffalo (<i>Bouteloua dactyloides</i>)	5.6	7.0
Carpetgrass (<i>Axonopus aureus</i>)	6.0	8.0
Centipede (<i>Eremochloa ophiuroides</i>)	4.3	5.8
Fescue (<i>Festuca rubra</i>) Red and Creeping	5.6	6.8
Fescue Tall (<i>Festuca arundinacea</i>)	5.6	7.0
Rye Annual (<i>Lolium multiflorum</i>)	5.8	7.4
Perennial Rye (<i>Lolium perenne</i>)	5.8	7.4
St. Augustine (<i>Stenotaphrum secundatum</i>)	6.3	7.8
Zoysia	5.5	7.0

Grass types with the lowest and highest pH level tolerated.

Adapted from Bill's Will: Misc Musings of a Techie: Easy Care Lawns My Way (Blog April 23, 2017) (willyvon1-willyswill.blogspot.com). <http://willyvon1-willyswill.blogspot.com/2017/04/easy-care-lawns-my-way.html>.

Table 3 The power of pH

Turfgrass	Soil pH tolerance	Ideal soil pH
<i>Warm season grasses</i>		
St. Augustine (<i>Stenotaphrum secundatum</i>)	5.0–8.5 ^a	5.5–7.2
Centipede (<i>Eremochloa ophiuroides</i>)	4.5–7.2	5.0–5.5
Bermuda (<i>Cynodon dactylon</i>)	5.5–7.5	6.0–7.0
Zoysiagrass (<i>Zoysia</i>)	5.0–7.8	6.0–6.5
Buffalo grass (<i>Bouteloua dactyloides</i>)	6.0–8.0	6.0–7.0
Seashore Paspalum (<i>Paspalum vaginatum</i>)	4.5–9.0	5.5–8.0
<i>Cool season grasses & soil pH</i>		
Kentucky Bluegrass (<i>Poa pratensis</i>)	6.0–7.2	6.0–7.2
Perennial Ryegrass (<i>Lolium perenne</i>)	6.0–7.5	6.0–7.0
Creeping Bentgrass (<i>Agrostis stolonifera</i>) (golf course greens)	5.5–7.2	5.5–6.5
Tall Fescue (<i>Festuca arundinacea</i>)	4.7–8.5	5.5–6.5
Fine Fescue (mix- <i>Festuca rubra</i> , <i>Fescue brevipila</i> , <i>Fescue ovina</i> L)	5.5–7.0	5.5–6.5

^aIron Chlorosis develops over pH of 7.5.

Adapted from The power of pH (2018).

biological components of otherwise healthy soils. A more sustainable approach is to utilize plant species that are best suited to the native pH of the soil for maintaining green vegetation.

Implementation of cultural practices that maximize nutrient uptake by plants, reduce nutrient waste and reduce off-site movement of nutrients. These practices include:

- minimization of bare soil and maintenance of dense turf cover to reduce erosion, prevent runoff and increase nutrient retention within the turf/soil system.
- Implement practices that result in rapid turf establishment following new planting or repairs.
- application of fertilizer and amendments containing nutrients only to turf, avoiding designated environmentally sensitive areas.
- application of fertilizer and amendments containing nutrients in a manner that prevents entry into surface waters or conduits, for example, catch basins that lead to surface waters or other environmentally sensitive areas.
- application of fertilizer and amendments containing nutrients where intended and not on hardscapes (driveways, sidewalks, roadways, non-vegetated surfaces). Should materials inadvertently land on such surfaces, they should be removed promptly and handled properly.
- handling of turfgrass clippings in an appropriate manner:
- retention of clippings within the turf system whenever possible.
- management of turfgrass clippings so that they are not allowed to enter surface waters or conduits such as catch basins that lead to surface waters.
- clippings that are removed should be composted or handled in a manner that does not threaten the environment.
- mowing at the highest end of the proper mowing height range for the grasses and cultivars present and for the turf use to maintain turf density and maximize rooting.
- irrigation practices promote infiltration and plant uptake, do not lead to runoff and leaching, and encourage deep and extensive turfgrass rooting.
- relief of compaction through appropriate cultural practices such as core aeration.
- Implementation of proper storm water management techniques aimed at reducing off-site movement of soil and nutrients.

Detailed application records for inputs of fertilizer and nutrient-containing materials. Field management teams recommend having soil tested when constructing a new field for fertility purposes and periodically once the field is established. If minerals are lacking, fertilizers can be applied 7–10 days prior to seeding and should be raked into the upper 5–7.5 cm of the topsoil.

Certain functional aspects of different turfgrass species and varieties may aid the turf manager in reducing inputs to meet specific management objectives. For example, grasses with lower consumptive water use rates may perform better under drought conditions while at the same time conserving irrigation water. Endophytic grasses may reduce damage from surface-feeding insects, and in turn, reduce the need for insecticide applications.

The science behind turf management

Future studies on sustainable turf-based public areas should consider public opinion and aim to achieve attainable human health outcomes while maintaining clean and green environments. Several studies have investigated the effects of turf management practices, including greenhouse gas emissions and fertilizer applications. For example, a study conducted by the University of Minnesota Crookston on Kentucky bluegrass turf found that while fertilizers increased turf health and greenness, they also increased greenhouse gas emissions, particularly during urea application. Researchers recommended the use of slow-release fertilizers such as Milorganite to lower emissions and maintain healthy turf. The study also considered cultivation practices such as verticutting and hollow tine aerification. It was observed that plants that had been aerified had higher emission rates than those that were untreated, while verticutting led to lower emissions, especially of CO₂ and N₂O, but decreased overall turfgrass quality.

The University of Massachusetts Extension Turf program has produced an extensive manual of Best Management Practices for turfgrass, which focuses on environmental protection and human safety for use with lawns and public space functions. The information presented in this guide, BEST MANAGEMENT PRACTICES FOR LAWN AND LANDSCAPE TURF version 1.51, is research-based and expertly audited knowledge intended to help practitioners make informed decisions focusing on scientific principles and practices of integrated pest management (IPM). For established, healthy turf, the guide recommends conducting soil chemical analyses at least every three years and monitoring pH annually. Critical soil test results for optimum growth have been summarized for US New England states turfgrass management, as well as recommendations for proper plant selection and optimum match with soil pH.

Another study, conducted at the University of Tennessee, aimed to determine the effect of soil type and grass species on head injury criterion (HIC). Soil moisture content was also considered and divided into three categories: dry (0.06–0.16 m³/m³), acceptable (0.17–0.29 m³/m³), and wet (0.30–0.40 m³/m³). The analysis used the Clegg hammer, a 2.25 kg hammer, to measure the surface hardness and the F334-E missile. The results of the study suggest that grass species have a minimal impact on HIC threshold, but soil type makes a significant difference. Of all the types analyzed, the USGA sand root zone was proven to be the optimum root zone for minimizing head injury risk to players. These root zones also suffered the least adverse effects due to water levels.

Park soil contamination and remediation

Soil contamination is a serious threat to human health globally, with heavy metals being a common source of pollution. Heavy metal contamination can result from industrial activities, gasoline emission deposits, lead paint, traffic, mining activities, among others, which can contaminate both industrial and residential areas. The potential health risks from exposure to heavy metals in contaminated soils were explored in various studies.

A study conducted near Athens, Greece, which converted a former shooting range into a public park, found concentrations of lead, nickel, and zinc using a portable X-ray Fluorescence Spectrometer (XRF). The study also explored the relationship between heavy metal contamination and human health risks using health risk indicators from the USEPA. The study revealed that ingestion risks were higher than skin contact or inhalation, and children were more susceptible to non-carcinogenic threats than adults (Urrutia-Goyes et al., 2017).

In a study in Galway City, Ireland, the use of XRF determined Lead, Copper, Zinc, and Arsenic contamination at a public park, which was previously a garbage disposal site for residential and industrial waste. The heavy metals were found at high concentrations, much greater than the background levels. The study advised immediate soil remediation to prevent children who play there from exposure (Carr et al., 2008).

The soil in an urban residential area in China was evaluated for heavy metal concentrations. The study used XRF to determine levels of heavy metals, and of the 80 samples, 25 were found to have high-risk contamination for lead, copper, zinc, and nickel. The most contaminated samples were closer to roads, which are most exposed to pollution from traffic. The researchers called for stricter zoning regulations to keep residential and recreational areas further from high industrial activities (Liu et al., 2022).

Soil remediation methods include containment, transforming, or removing the contaminated soils, and the distribution of heavy metals in the soil is a significant factor in selecting the appropriate method. In-situ techniques are more efficient because the soil remains in place, while ex-situ remediation, such as excavation and transportation, can be messy and expensive. Encapsulating the area of contamination with concrete or clay is the best choice for small, highly contaminated areas. For a permanent fix, soil flushing involves injecting the soil with an extraction fluid, and chemical immobilization is the best choice for low to moderately contaminated soils. Phytoremediation is a developing method that relies on plants to remove or stabilize metals into harmless forms, and compost application and mulching are low-cost methods for binding and diluting heavy metals in soils (Paltseva et al., 2022).

In conclusion, the potential health risks from exposure to heavy metal contamination in public parks and residential areas is a serious concern. Soil remediation is necessary to prevent exposure to heavy metals, and various methods are available depending on the level of contamination and distribution of heavy metals in the soil. The selection of the appropriate method should consider the budget, time availability, and geographic limitations.

Conclusion

In recent years, concerns about the use of synthetic pesticides and herbicides have led to questioning traditional maintenance methods on public recreational facilities such as golf courses and sports fields. This has posed a challenge for managers to maintain aesthetically pleasing, healthy fields while focusing on human and environmental health. The primary goals of most modern turf management programs are to achieve the expected performance level with minimal input and environmental impact. These principles can be extended to most systems of sustainable management and engineering design for parks, gardens, sports fields, golf courses, and public rights of way.

Managers of park facilities and sports fields at all levels must strive to reduce or eliminate the use of pesticides and synthetic fertilizers on natural turfgrass playing surfaces. However, this poses a challenge to balance acceptable playing surface quality and visual aesthetics with environmental impacts and safety, with the latter being a priority. Unlike home lawns, sports fields, golf courses, and public park thoroughfares present unique problems for functionality, such as higher traffic, larger turf areas, and thus greater environmental impacts. These factors require a greater focus on watershed analysis, drainage, compaction, and runoff before addressing plant selection, fertility, and pest or weed management.

Nutrient management plans and integrated pest management (IPM)-based programs guide the manager to achieve the expected performance level while minimizing pesticide use or excessive use of fertilizer, whether organic or synthetic. Organic management programs aim to achieve the desired performance without adding synthetic fertilizers or pesticides into the system through methods such as organic fertilizers, natural weed suppression, and organic field aeration. Although research in this area is just beginning and sparse, there are many examples in the US and worldwide of pilot programs with eco-friendly alternatives to field maintenance that prioritize sustainability for the future.

References

- Bill's Will (2017) Misc Musings of a Techie: Easy Care Lawns My Way. (Blog April 23, 2017) (willyvon1-willyswill.blogspot.com) <http://willyvon1-willyswill.blogspot.com/2017/04/easy-care-lawns-my-way.html>.
- Carr R, Zhang C, Moles N, et al. (2008) Identification and mapping of heavy metal pollution in soils of a sports ground in Galway City, Ireland, using a portable XRF analyzer and GIS. *Environmental Geochemistry and Health* 30: 45–52. <https://doi.org/10.1007/s10653-007-9106-0>.
- Cavendish W (2022) *Opinion: Digital is key to unlocking the potential of nature based solutions.*

- Liu S, Peng B, and Li J (2022) Ecological risk evaluation and source identification of heavy metal pollution in urban village soil based on XRF technique. *Sustainability* 14(9): 5030. <https://doi.org/10.3390/su14095030>.
- Paltseva AA, Cheng Z, McBride M, Deeb M, Egendorf SP, and Groffman PM (2022) Legacy lead in urban garden soils: Communicating risk and limiting exposure. *Frontiers in Ecology and Evolution* 10: 873542. <https://doi.org/10.3389/fevo.2022.873542>.
- Urrutia-Goyes R, Argyraki A, and Ornelas-Soto N (2017) Assessing lead, nickel, and zinc pollution in topsoil from a historic shooting range rehabilitated into a public urban park. *International Journal of Environmental Research and Public Health* 14(7): 698. <https://doi.org/10.3390/ijerph14070698>.

Further reading

- Aquatrols (2018) *The Power of Soil pH*. (Blog December 7, 2018) Education, Featured, Soil and Water Quality <https://blog.aquatrols.com/power-soil-ph/>.
- ARUP and World Economic Forum (2022) *BiodiverCities by 2030: Transforming cities' relationship with nature*. <https://www.weforum.org/biodivercities-by-2030/>.
- Brettingen J, Schwab R, Lunseth S, and Qiu Y (2016) *Sustainable Turf Management for City of Watertown Athletic Fields*. Resilient Communities Project (RCP), University of Minnesota. Retrieved from the University of Minnesota Digital Conservancy <https://hdl.handle.net/11299/184937>.
- Complete Communities Toolbox: Attractive, Inclusive, Efficient, Healthy & Resilient Places*. University of Delaware. <https://www.completecommunitiesde.org/planning/inclusive-and-active/parks-rec-master-planning/#steps>.
- Dickson K, Strunk W, and Sorochan J (2018) The effect of soil type and moisture content on head impacts on natural grass athletic fields. *Proceedings* 2(6): 270. <https://doi.org/10.3390/proceedings2060270>.
- Ebdon JS (2011) *Management Tips to Improve Turfgrass Drought Survival*. University of Massachusetts, Amhurst Center for Agriculture, Food, and the Environment UMass Extension Turf Program. <https://ag.umass.edu/turf/fact-sheets/cultural-practices>.
- Folkner JS (1957) *Construction, Renovation and Management of Athletic Fields*. Tucson, AZ: Agricultural Experiment Station, University of Arizona <https://doi.org/10.1520/t3339>.
- IUCN (2020) *Guidance for Using the IUCN Global Standard for Nature-Based Solutions – A User-Friendly Framework for the Verification, Design and Scaling up of Nature-Based Solutions*. <https://doi.org/10.2305/IUCN.CH.2020.09.en>.
- Jones KD (2020) *Soil pH and Fertilizers*. Mississippi State University Extension Publication Number: IS372 Mississippi State University Extension Service. (msstate.edu) <https://extension.msstate.edu/publications/soil-ph-and-fertilizers-0>.
- Kumar P, et al. (2021) Nature-based solutions efficiency evaluation against natural hazards: Modelling methods, advantages and limitations. *Science of the Total Environment* 784: 25. <https://www.sciencedirect.com/science/article/pii/S0048969721021288>.
- Lanier JD (2016) *Turf Maintenance Levels: Stress and Expectations*. University of Massachusetts Amhurst Center for Agriculture, Food, and the Environment UMass Extension Turf Program. <https://ag.umass.edu/turf/fact-sheets/cultural-practices>.
- Liu L, Li W, Song W, and Guo M (2018) Remediation techniques for heavy metal-contaminated soils: Principles and applicability. *Science of the Total Environment* 633(15): 206–219. <https://doi.org/10.1016/j.scitotenv.2018.03.161>.
- Parks and Recreation Master Planning | Planning for Complete Communities in Delaware. Newark, DE: University of Delaware. completecommunitiesde.org.
- Rice L, Beasley J, Gaston L, Sanders K, and Munshaw G (2019) Planting rate and nitrogen fertility affect runoff losses during hybrid Bermudagrass establishment. *Agrosystems, Geosciences and Environment* 2(1–5): 190057. <https://doi.org/10.2134/age2019.07.0057>.
- Storm Water Report and Jacques J (2021) *Baltimore's Camden Yards joins growing group of stormwater-minded sports venues*. <https://stormwater.wef.org/2021/01/baltimores-camden-yards-joins-growing-group-of-stormwater-minded-sports-venues/>.
- Stormwater Report (2016) *NYC Marshlands Jamaica Bay or Louisiana artificial Marshlands to buffer and control storm water*. <https://stormwater.wef.org/2016/10/university-scientists-develop-artificial-marshland-treat-louisiana-stormwater/>.
- Sustainable Practices | The High Line. <https://www.thehighline.org/sustainable-practices/>.
- University of Massachusetts, The UMASS Extension Turf Program (2016) *BEST MANAGEMENT PRACTICES FOR LAWN AND LANDSCAPE TURF version 1.51 prepared by: Mary C. Owen, Extension Turf Specialist Jason D. Lanier, Extension Educator Select technical content contributed by: Natalia Clifton – Extension Pesticide Specialist M. Bess Dicklow – Extension Plant Pathologist Dr. Scott Ebdon – Turf Agronomist Dr. Geunhwa Jung – Turf Pathologist Jason Lanier – Extension Educator Mary Owen – Extension Turf Specialist Randall Probstak – Extension Weed Specialist Dr. John Spargo – Soil Scientist Dr. Patricia Vittum – Turf Entomologist and additional sources as designated in the text*. https://ag.umass.edu/sites/ag.umass.edu/files/pdf-doc-ppt/lawn_bmp_establishment_2016_final.pdf; <http://extension.umass.edu/turf>.
- Walker KS and Smith KE (n.d.) *The Effects of Cultivation Practices and Fertilizer Use on the Mitigation of Greenhouse Gas Emissions from Kentucky Bluegrass Athletic Fields*. The University of Minnesota Crookston.
- White O and Semmelrock C (2021) What is an "organic" golf course, and why aren't there more of them? *Golf*. Retrieved December 4, 2021, from <https://golf.com/travel/courses/organic-golf-course-pesticide-complications/>.
- Young RF (2010) Managing municipal green space for ecosystem services. *Urban Forestry & Urban Greening* 9(4): 313–321. <https://doi.org/10.1016/j.ufug.2010.06.007>.

Overview chapter on soil-society relations

C Siebe^a, Margaret A Oliver^b, and K Müller^c, ^aInstituto de Geología, Universidad Nacional Autónoma de México, Ciudad de México, México; ^bSoil Research Centre, Department of Geography & Environmental Science, School of Human & Environmental Sciences, The University of Reading, Reading, United Kingdom; ^cSustainable Production, The New Zealand Institute for Plant and Food Research Limited, Auckland, New Zealand

© 2023 Elsevier Ltd. All rights reserved.

Introduction	505
The perception of nature and cultural aspects	506
Education	506
Planning and landscaping	506
Law and policy	507
Economy and soil evaluation	507
Soil and animal and human health	507
References	508

Abstract

Traditionally the study of soil has been handled by the natural and applied sciences, and those closely linked to food production or to forestry. Yet new disciplines and sectors of society are increasingly looking towards the soil with the emergence of new concepts being brought to soil science. This chapter provides an overview on how awareness about the importance of soil for civilization has changed over time and describes briefly some emerging topics of soil-society relations beyond the ones traditionally studied. It aims to introduce briefly the chapters in this section, which describe these relationships in more detail.

Key points

- Soil science is an interdisciplinary science.
- Its study does not only incorporate natural sciences, but also increasingly social sciences, the law and art.
- In parallel to the agricultural and environmental sciences and forestry, human and animal health, urban planning and architecture are increasingly demanding knowledge on soil.
- Emerging disciplines such as forensic science need to characterize soil.
- Major societal challenges such as inequality, climate change and antimicrobial resistance, and most Sustainable Development Goals (SDGs) need to include knowledge of the soil.
- Education in interdisciplinary scientific work and in inter-institutional cooperation requires knowledge of the soil and should be fostered to achieve sustainable soil management and reverse soil degradation.

Introduction

The scientific study of soil has focused for centuries on soil formation and on understanding the patterns of soil distribution in landscapes (pedology and soil geography) and in investigating soil functioning, i.e., biological, chemical and physical soil processes. As soil supports crops, land-use systems and agricultural management practices are additional necessary components of an encyclopedia of soil. Land use and management changes in response to shifts in society's demands, require a better understanding of soil-environmental interrelationships and biotechnological advancement. Environmental sciences emerged decades ago and they investigate soil response to climate change and pollution. The respective chapters in the Encyclopedia provide updates on the knowledge acquired in all these areas of soil science. This Encyclopedia of Soils in the Environment includes additional chapters on novel topics in which knowledge has been acquired on many aspects of soil that we might not consider soil science. These chapters highlight how soil is playing an increasingly important role in the 21st century. They show an imaginative and exciting aspect of soil that many will not be familiar with.

Apart from the traditional disciplines involved in the study of soil, many other disciplines, particularly from the social sciences, humanities and arts are increasingly interested in soil. Lawyers develop policies to regulate soil use, to protect it from degradation and to establish rules to rehabilitate the soil when it has been degraded by human mismanagement. Economists value ecological functions to incorporate them into cost-benefit analyses. People working in the health sector need a better understanding of environmental conditions, including those related to the soil because it is now clearly recognized that soil, animal and human health are closely interrelated (Brevik et al., 2020; Gibbs, 2014).

Relatively young disciplines such as forensic sciences often need to include soil characteristics as evidence that links a crime scene to potential suspects. This chapter provides an overview of these emerging soil-society relations.

The chapter further acknowledges the importance of education and outreach to ensure greater understanding of the importance of soil that will help to protect the soil from degradation. Throughout history, urban societies in particular have lost awareness about the importance of soil for their daily lives, but it has never reached the perilous degree of the current century.

The perception of nature and cultural aspects

Soil supports terrestrial life, yet awareness of its importance for daily life is lacking especially in urban societies; it is generally strong within some rural societies that have depended directly on the soil resource for generations. Rural societies until the industrial revolution developed an understanding of soil-plant-water relationships through the close daily contact with soil by growing their food, observing water infiltration and runoff, and disposing their organic residues into soil to maintain its fertility. They gained empirical knowledge which has been passed down for generations. [Homburg et al. \(2023\)](#) document that some civilizations in East Asia or the Andes have developed sustainable management practices in harmony with the environment for centuries and even millennia. In many places population growth, globalization and climate change have led to soil degradation. Their chapter provides examples of the collapse of modern and ancient civilizations based on their access to good quality soil and especially on the knowledge of their members with respect to sustainable land-use practices. Special attention is given to the adaptability of ancient civilizations to climate change, which is not only a current major challenge, but has also been so in the past.

Urban societies are disconnected from the soil as the natural source of food because food is provided by markets, and soil is not considered in relation to potable water, which is brought by pipes through the water supply network, runoff is regulated by the subsurface sewerage system. The nutrient recycling function of soil is further not acknowledged because residues are taken away by the waste collection service. Currently more than half of the world's population lives in cities, however urban societies are often not aware of the importance of soil to sustain terrestrial life. Further, reduced contact with soil does not allow people to engage with soil protection or conservation ([Siebe et al., 2018](#)).

Art can help to foster awareness and communicate the importance of soil for society. The innovative chapter by [Toland and Wolter \(2023\)](#) illustrates how artists are not only considering soil in artistic research, but are also contributing to life-long learning and soil literacy. The different formats of art expression as visual and performing arts, film, literature, music, and various design disciplines address themes that are often beyond those considered by soil scientific investigation. The inclusion of ethical, political and esthetic topics into the overall discourse about soil provides important incentives for inter- and trans-disciplinary research. Art reaches much broader audiences, which in turn helps to promote soil protection in the political agenda. Emotional connections to soil allow people to develop a sense of responsibility to preserve this life sustaining resource. The chapter addresses soil sensitization through soil color, sound, smell, touch and taste; the latter is explained through the concept of 'terroir' which explains subtle taste differences in food or beverages (for example wine) determined by local soil, relief and climate conditions of particular regions.

Education

As stated above, it has become imperative to increase awareness of the importance of soils, and to introduce and enforce soil information in education programs. The chapter by [Reyes-Sánchez and Brevik \(2023\)](#) explores important concepts in soil science education, some of the more common modes of education delivery, groups that soil science educators commonly reach out to, and professional soil science groups that engage in public education. This chapter clearly points out that teaching involves more than just the delivery of discipline-specific content. Active learning approaches (i.e. learning by doing) have proven to be the most effective strategies. Education has to reach all age groups of society, from children to adults, and foster inter- and trans-disciplinary thinking. Different people are increasingly engaging in soil science education, including teachers, members of soil societies and professionals in communication science. Teaching can take place in the formal classroom, in the field, or the laboratory; at the university level, it should also promote academic-industrial partnerships that expose students to complex real-world problems to develop the students' critical thinking skills and expose them to a variety of perspectives. Educating children and teenagers is especially important because their understanding of the value of soil will commit them to its preservation.

Planning and landscaping

The growing world population demands more and more land to satisfy its needs. Land-use systems necessarily have to adapt to the prevailing environmental conditions. Ancient societies designed their land-use systems according to the landscape attributes, potential and environmental conditions, and to their needs (food, clothing, etc.). However, the green revolution of the last century exported and globally distributed technological packages to optimize food production for diets deemed adequate by global health and nutrition organizations ([Denizen, 2023](#)). Current land-use systems have prioritized soil functions related to good extraction over soil functions that regulate and balance the water and nutrient cycles. This has led to the global likelihood that more soil is lost than is formed. In his chapter, Seth Denizen invites change in planning paradigms by incorporating rates of soil formation into the design of land-use systems instead of assessing resource stocks or documenting degradation. Professionals in charge of land-use planning acknowledge just a few attributes that dominantly refer to 'land' characteristics, as dimension, inclination, ownership,

distance to infrastructure, facilities for water supply, exposure to natural hazards, etc. The chapter addresses the problems of including soil as a 'layer' in Geographic Information Systems in land planning, by emphasizing that the often taxonomic legend of soil maps does not allow local cultural, empirical and political aspects to be considered, which strongly affect sustainable land-use practices. Much work needs to be done to foster communication between land planners and local stakeholders, as well as experts in different disciplines such as soil science, economy, politics, etc., to design circular economies centered on building soil resources rather than consuming them (Denizen, 2023).

Law and policy

The jurisdictional perception of soil has changed in recent decades as pieces of land are no longer viewed just as assets owned privately by people, by institutions, or by municipal or state entities who have a right to use and exploit the soil, but is moving towards making land owners responsible to maintain soil ecological functioning. The law and policy aspects involving this new paradigm are analyzed in the chapter by Ginsky (2023). Assurance of global food security, mitigation and adaptation to climate change and the conservation of biodiversity have triggered policy actions to meet the Sustainable Development Goals established by the United Nations General Assembly (United Nations, 2015). Governments are implementing and expanding the legal framework to ensure soil protection and provide incentives for sustainable management practices. International institutions are designing organizational and institutional arrangements to help achieve sustainable soil management. These actions aim to create a system of soil governance, which Ginsky (2023) explains in his chapter. He summarizes the existing regulatory structure internationally and provides some examples of actions taken by local governments. He concludes that containment of soil degradation will be achieved only by strong and effective interinstitutional cooperation.

Forensic soil science is an emerging discipline where soil science is applied in criminal justice systems. It provides soil information and characterization methods to support law enforcement agencies in investigating criminal, environmental, and serious and organized crime. The chapter by Dawson (2023) provides an introduction to this new discipline and explains how information on soil, its minerals, and plant materials helps to determine their provenance and to understand evidence at crime scenes. The soil sampling and characterization methods are carefully adapted to meet the needs of criminal justice systems.

Economy and soil evaluation

Human wellbeing largely relies on goods and services provided by soil. However, soil is a limited resource, particularly when current rates of soil degradation are considered. Many soil ecosystem services are collective goods, which are compromised by mismanagement by the land owners. The preferential use or exploitation of one particular soil function (for example provision of food) often leads to the under-provision of other soil functions (preservation of natural vegetation biodiversity). Public preference for the different soil functions dictates the extent of under-provision of other soil functions and can be estimated by economic nonmarket valuation (Bartkowski and Massenberg, 2023). It has become increasingly necessary to provide incentives to promote sustainable land-use practices and to give economic value to soil services so that they can be considered in decision making. A particular challenge is that many soil services are not traded on international markets and thus do not have a monetary value. The chapter by Bartkowski and Massenberg (2023) provides an economic perspective on soils' contributions to human well-being and indicates potential ways to provide nonmarket valuations of soil services. Policy instruments help to align soil management with public preferences and to optimize soil management to preserve soil multifunctionality. Understanding of the economic effects of soil use and management contributes to effectively and efficiently protecting soils.

Soil and animal and human health

Soil is linked to human and animal health, not only by the nexus soil-food provision and quality, but also because many diseases are transmitted and possibly originated in soil. Polluted soil with potentially toxic elements, radionuclides and organic compounds put human and animal health at risk through the soil-plant-water-air pathways. Soil components such as clay minerals have been used since ancient times to treat skin and stomach ailments. Soil harbors microorganisms such as bacteria and fungi from which many well-known antibiotics have been produced. It is also likely to be a vital source of future ones which is important because of the increase in antibiotic resistance genes. This has been recognized as one of the most important challenges humankind is facing currently, together with global climate change. The chapter by Oliver and Brevik (2023) provides an overview of the different interconnections between soil and human and animal health and emphasizes the opportunities to foster preventative medicine. Awareness about the strong interrelations between environmental, animal and human health is provided by the One Health Initiative (<https://onehealthinitiative.com/>), which also gives incentives to increase interdisciplinary work further on this topic.

References

- Bartkowski B and Massenberg J (2023) The economics of soil's contribution to human well-being. In: *Encyclopedia of Soils in the Environment*, 2nd edn. Elsevier.
- Brevik EC, Slaughter L, Singh BR, et al. (2020) Soil and human health: Current status and future needs. *Air, Soil and Water Research* 13. <https://doi.org/10.1177/1178622120934441>.
- Dawson LA (2023) Soil science in the criminal justice system. In: *Encyclopedia of Soils in the Environment*, 2nd edn. Elsevier.
- Denizen S (2023) Soils in landscape planning. In: *Encyclopedia of Soils in the Environment*, 2nd edn. Elsevier.
- Gibbs EPJ (2014) The evolution of One Health: A decade of progress and challenges for the future. *Veterinary Record* 174(4): 85–91. <https://doi.org/10.1136/vr.g143>.
- Ginsky H (2023) Soil governance at the international, regional and national level. In: *Encyclopedia of Soils in the Environment*, 2nd edn. Elsevier.
- Homburg J, Sander J, and Brevik E (2023) Civilizations and the role of soil, agriculture and climate change. In: *Encyclopedia of Soils in the Environment*, 2nd edn. Elsevier.
- Oliver M and Brevik E (2023) Soil and human health. In: *Encyclopedia of Soils in the Environment*, 2nd edn. Elsevier.
- Reyes-Sánchez LB and Brevik EC (2023) Soil science education, awareness and outreach in the modern world. In: *Encyclopedia of Soils in the Environment*, 2nd edn. Elsevier.
- Siebe C, Cram S, and Mora-Palomino L (2018) Enhancing awareness about the importance of urban soils. In: Lal R (ed.) *Urban Soils. Advances in Soil Science*, pp. 371–374. Boca Raton/London/New York/Leiden: Taylor and Francis Group. ch. 17.
- Toland A and Wolter D (2023) Soil art: Sensory and symbolic engagement with soils. In: *Encyclopedia of Soils in the Environment*, 2nd edn. Elsevier.
- United Nations (2015) *General Assembly, Seventieth Session, No. 11688, Agenda items 15 and 116, Resolution adopted by the General Assembly on 25 September 2015*. Transforming our world: The 2030 Agenda for sustainable development, A/RES/70/1, p. 1

Soil art: Sensory and symbolic engagement with soils

Alexandra R Toland and Daniel Wolter, Bauhaus-Universität Weimar, Faculty of Art and Design, Chair of Arts and Research, Geschwister-Scholl-Straße 7, Weimar, Germany

© 2023 Elsevier Ltd. All rights reserved.

This chapter is a longer, edited and translated version of the paper, *Künstlerische Beiträge zur Bodenbewusstseinsbildung: Soil Sensing*. In *Handbuch der Bodenkunde* (Handbook of Soil Science) edited by Hans-Peter Blume, Peter Felix-Henningsen, Hans-Georg Frede, Georg Guggenberger, Rainer Horn, and Karl Stahrand. 41st Edition, published in 2021.

Introduction	509
Soil chromology	510
Sounding soil	512
Soil scenting	514
Terroir	515
Multisensory absorption	517
Conclusions and outlook	518
References	519

Abstract

This chapter presents a sample of a much larger field of artistic research on soils. Alongside advancements in proximal soil sensing technologies, artists contribute new insights on soil sensing by facilitating multisensory experiences, developing new vocabularies for what is sensed, raising awareness about soil ecosystems and their complexities, and producing innovative research on soil phenomenology that can be cited by artists and scientists alike and inspire new interdisciplinary collaborations. Artistic soil sensing, with its diverse foci and forms of presentation and mediation, can target different audiences and segments of society, playing an important role in soil communication, life-long learning and soil literacy-building efforts.

Introduction

The arts, including the visual and performing arts, film, literature, music, and various design disciplines, play a special role in raising awareness about soil protection in that they address a range of different audiences and bring attention to themes of ethical, political and esthetic nature that are often beyond the scope of soil scientific investigation (Feller et al., 2015; Toland et al., 2018). Without claiming the recognition of a new art historical genre, we will use the term “Soil art” in the following to refer to artistic activity with soil materials or in soil environments “that is produced by artists in a multitude of genres and media, to be understood, among other things, as artwork that may contribute to wider environmental and soil protection discourses and awareness-raising efforts” (cf. Feller et al., 2015). Over the past couple of decades, the recognition of artistic works in soil scientific contexts, such as publishing, academic networking, grant consortia, and scientific conferences, and also the role of soil in artistic contexts, such as exhibitions, residencies, performances, and art publications, has evolved from being a marginal interest on both sides to becoming a new frontier for inter- and trans-disciplinary research (First-Arai, 2019; Toland et al., 2018; Adams and Montag, 2016). Collaboration and knowledge transfer across scientific and artistic disciplines has increased since the 2015 UN International Year of Soils, while new forms of communication and public engagement are needed to address soil erosion, soil pollution, soil organic carbon (SOC) loss, soil sealing, and many other “wicked problems”¹ confronting soil health today.

Science communication typically represents a marginal investment in soil scientific research, at least in third party funded fundamental research. Feller et al. (2015) observe that “the last decades show, however, that the activities of the soil science community and its traditional partners have been insufficient in protecting our soils and landscapes.” Soil art is seen as an integral partner in expanding the outreach and knowledge dissemination activities of the soil science community. At least since 2015, raising awareness of soils can no longer be considered a niche interest of soil scientists, but has been “mainstreamed” thanks in part to events such as the 2015 Expo under the motto “Feeding the Planet, Energy for life,” the 2015 UN International Year of Soils and dedicated UN decade of soils 2015–24. Beyond the inclusion of smaller artistic contributions to soil science meetings (e.g., the Soil Cinema and poster exhibition organized by A. Toland at the 20th World Congress of Soil Science in Jeju in 2014 and the Our Living Soils events and exhibitions on occasion of the 22nd WCSS in Glasgow in 2022), several residencies and research groups have recently been developed with a focus on soils, including the Soil Culture residencies in the UK, the Zone2Source residencies in Amstel Park, Amsterdam, Netherlands, Cargo Collective’s Soil LAB in Domaine Départementale de Chamarande, France, Soil Futures at Arts Catalyist in Sheffield, UK, and the Anthropogenic Soils cluster at the Oslo School of Environmental Humanities (OSEH),

¹Design theorists Rittel and Webber (1973) introduced the concept of “wicked problems” to describe complex challenges to society that are symptomatic of other problems and cannot be solved with one approach. The term has been picked up in the environmental humanities and Anthropocene studies in reference to complex problems such as climate change and soil desertification.

Norway. By facilitating relationships and creating identity with specific sites and ecologies, artists have actively contributed to increasing soil awareness and improving soil literacy.

With historical roots in the materiality and visual imagery of landscape painting and the Land Art and Earthworks of the 1960s, 1970s, and 1980s, artists today conceptualize their works on soils and with soil materials, by consciously integrating the historical, cultural, and site-specific dynamics of soils into their compositional, curatorial, material and media choices. Artists have investigated the formal esthetic properties of soils, gaining knowledge on how soil materials react chemically and can be manipulated, e.g., in the kiln, on canvas, or in large scale installations.² Artists explore the embodied, sensory reactions of humans to soils and soil environments in situ or on display, bringing awareness to how knowledge of soil formation is created through sensory experience. Artists regularly engage with landowners and farmers, collecting and integrating their impressions and experiences as primary material to be exhibited to wider audiences (see, e.g., Katie Revell's Farmerama podcasts) or as secondary material in the conceptualization of installations, interventions and more (see, e.g., description of masharu's work below). Soil art, in other words, can be seen as a form of empirical research on soil materials, environments, histories, embodied experiences, and social worlds, mediated through exhibitions, performances, publications, audio, film, and social media formats, and beyond.

While soil art might not seem immediately related to soil science in a strictly disciplinary sense, compared to more established "border disciplines" (c.f. Thompson Klein, 1990, p. 65) such as soil hydrology, geophysics, or agrochemistry, the commonalities between soil science and art are many. Both use keen and systematic observation to derive new understandings of soils; both strive for novel interpretations of nature-based phenomena; both are evaluated based on their creative use of concepts and methods; both are uncertain or risky in their outcomes; both develop processes or products that can be communicated to different target audiences. Indeed these commonalities of "novelty, creativity, uncertainty, systematization, and transferability" represent the main criteria for research itself as outlined in the OECD Frascati Manual—Guidelines for Collecting and Reporting Data on Research and Experimental Development (OECD, 2015 p.28). Positioning soil art as a form of empirical research or artistic knowledge-creation thus goes beyond the recognition of the role of art in raising awareness and facilitating public outreach. Soil art can do more. It mediates soil scientific knowledge to broader audiences, but also generates artistic knowledge and understanding about soils and their central roles in local ecologies and communities. More than simply a mouthpiece for the natural sciences, soil art can be seen as a small but growing field of research-creation in the arts and humanities that expands the scope of soil literacy and enhances the knowledge-base, already well established in the fields of soil physics, soil chemistry, soil biology, and soil taxonomy, and also the applied knowledge held by land managers, foresters, farmers and other practitioners. While a comprehensive overview of soil art is given in the publication *Field to Palette—Dialogues on Soil and Art in the Anthropocene* (Toland et al., 2018), this chapter outlines several key areas of development and draws on several case studies to outline the unique potential of soil art within the soil science community.

The chapter is structured around different approaches to creating sensory experiences with soils, emphasizing how multisensory engagement with soils, can lead to an increased sense of responsibility for soil health, and as an appreciation for the entangled relationships of human societies and the soils that sustain them. Artistic methods are shown to complement the techno-scientific field of proximal soil sensing³ with the aim of making emotional connections between human bodies and soil bodies. On the one hand, they include examples of technologically-conditioned modes of perception through the use of instruments, on the other hand they include direct modes of sensing with the human body.⁴ The idea is that while soil sensing has become essential for monitoring various properties such as soil organic carbon content, erosion rates, and water and nutrient availability, good data do not necessarily lead to good land use practices. Methods of artistic soil sensing can extend the technologically driven "new paradigm for agriculture" (Rossel and Bouma, 2016) by creating sensory experiences that evoke attentiveness and meaningful identifications with place, foster curiosity and wonder, and underline the bodily and emotional connections to soil as a fundamental basis for good practice (cf. Krzywoszynska, 2021).

Soil chromology

Central to discussions on soil sensing is the role of color in the interpretation and use of soils. This has been shown by some of the earliest forms of human communication about their environments in the cave art at Lascaux, France, to contemporary advancements in soil spectroscopy. The range of different approaches to studying soil color and its myriad applications could be subsumed under the term of "soil chromology." Soil scientists Ed Landa (2004) and Fiorenzo Ugolini (2010) provide initial insights on the history, chemistry and meaning of soil color, which is important for many artists interested in soil. The differentiation of soil color has been systematized and standardized over time, most prominently by the painter Albert H. Munsell, whose system for describing soil color is still used today. The Munsell Color System provides "a grammar of color" (Munsell and Cleland, 1921), encoding hue (Hue), gray value (Value), and saturation (Chroma) into numerical values for determining, for example, humus content, oxygen availability, and iron compounds under optimal visual conditions using moist soil samples and direct sunlight at right angles to the

²See, respectively, the works of glass artist Sarah Hirneisen, painter Ulrike Arnold, and conceptual artist Lara Almarcegui in *Field to Palette: Dialogues on Soil and Art in the Anthropocene* (Toland et al., 2018).

³Proximal soil sensing (PSS) refers to electrochemical, spectroscopic, geophysical and mechanical sensing methods used to gather data on multiple scales for scientific study, soil mapping, and land management recommendations. See Adamchuk et al. (2015) for a comprehensive overview.

⁴See Schneider & Zemanek, 2020, and Gabrys, 2016, for critical discussions on environmental sensing technologies.

color chart. Contemporary artists have often turned to Munsell color charts in their public engagement with soils, but go beyond Munsell's system to create their own taxonomies of soil color using artistic means. As has been described in Toland and Wessolek (2014) and Feller et al. (2015), bringing attention to the diversity of soil color is a common thread in artistic work with soils. This is especially evident in the works of Ellie Irons' participatory cartographic work in Philadelphia and New York; Peter Ward's collection of earth pigments in former mining areas in western England (eARTh—An Exploration of Earth Pigments in North Devon, 2008–18); Elvira Wersche's elaborate geometric floor installations using different colored sands from around the world; Ulrike Arnold's abstract in situ paintings using local soil samples; environmental art pioneer Herman de Vries' collection of boxed soil samples and soil rubbings from around the world organized as a soil archive; Koichi Kurita's grid installations and glass bottles of sifted soil samples from over 30,000 communities in Japan; Margaret Boozer's installation "Five Borough Reconnaissance Soil Survey" spatial interpretation of the New York Soil Survey; the Soil Alliance of Lower Austria's outreach project with Irena Racek, and the "One Earth Altar" collection of soil books started by Marianne Greve for Expo2000 in a church in Schneverdingen, Germany. Despite different compositional, material, and presentation format choices and exhibition outcomes, all these works bring attention to the enormous diversity of soil color that is usually hidden beneath our feet. The collections of soil samples in these artworks are archives of soil diversity, that despite having been moved from places of origin, create an awareness for soil as more than a homogenous brown substrate in which to grow crops. Artists shine light on a veritable subterranean rainbow of difference usually inaccessible to human vision.

In addition to these and other artistic approaches to communicating soil diversity through the archiving and presentation of large amounts of soil samples, color also plays a role in the assessment of sites and narration of their histories, vulnerabilities and potentials. Recent observations of soil art projects at exhibitions, festivals, and residencies include the renaissance of soil chromatography as a qualitative, visual method for site assessment, as well as the development of new tools for color-coding soil properties, both with the assumption that color can communicate snapshots of soil health at particular sites. Developed in the 1950s by Erenfried Pfeiffer, an associate of Rudolf Steiner, soil chromatography has been promoted in biodynamic farming circles as an easy and inexpensive method for visually analyzing the organic content, mineral make up and microbiological activity of a given soil (Pfeiffer, 1984). To make a soil chromatogram, a soil sample is first mixed with a dilute solution of sodium hydroxide to gently break down the mineral and organic components of the soil. The mixture is then applied to a round filter paper treated with silver nitrate by pouring the sample into a petri dish and allowing it to enter the middle of the filter paper through the capillary action of a wick inserted in the center of the paper. A unique radial image consisting of channels and spikes emerges as the different components are soaked up by the paper at different rates. Over the past several years the technique has been embraced and popularized by artists and artist groups such as Debra Solomon (Fig. 1), Claire Pentecost and Brian Holmes, Julien Chollet and mikrobiomik (various workshops, ongoing), Aino Johansson and Sirja Moberg, Hannah Fletcher, Scott Hunter (ongoing), Phil Lambert and others.

The reliability of soil chromatography for soil health assessment has been questioned by scientists such as Ford et al. (2021) who carried out extensive testing of the method and concluded that only weak correlations between chromatographic readings and soil variables could be found, however, the esthetic potential of the method should not be underestimated. What is unique about soil chromatography is its reliability and replicability as an artistic means of public engagement with soils in diverse settings. It is important to note that the artists mentioned above have usually produced soil chromatograms in workshop settings, field trips, and other participatory settings, facilitating hands-on experiences and in-depth exchanges about soil health, often with people who are not professionally connected to soils. Soil chromatography offers a way of creating colorful holistic portraits of local soils that



Fig. 1 Soil Portrait #26, Debra Solomon, 2016, sampled from self-made topsoil produced in situ at Urbaniahoeve Food Forest Ecosystem in Amsterdam Noord. SP#26 depicts a young, well-aerated soil, comprising a high percentage of partially decomposed organic materials. Reprinted with permission by the artist.



Fig. 2 Soiled. Exhibition view of Soiled: iGEM International Competition (February 27 to March 12, 2016), SVA Chelsea Gallery, New York City. Photo by Raul Valverde. Reprinted with permission by the artist.

become objects of contemplation, conversation and wonder, and ideally an increased feeling of responsibility for soil health in gardens and farms, urban spaces and backyards, and beyond. For artists, it is not the accurate correlation between soil properties with the radial patterns on the filter paper that determines the reliability of soil chromatography as a method, but rather the continued interest of art spaces and their audiences with this specific means of soil portraiture and its potential for connecting people to local ecologies.

Beyond work with soil pigments and soil chromatography, another group of artists and designers are using color to code various soil properties to make field data more user-friendly. Under the direction of bioart pioneer Suzanne Anker, a mixed group of art and science students took part in the prestigious iGEM (International Genetic Engineered Machine) competition in 2015 with the project “Soiled,” which won a gold medal for Best Art and Design Project. Informed by developments in soil spectroscopy, Soiled applies color as a metric for analysis. Using a spectrophotometer, an assigned wavelength of light was beamed through a soil solution to measure how much light was absorbed. The team developed a microfluidic chip that assigned quantitative values to color dots, resulting in a color map (Fig. 2) visualizing the available nutrients of soils sampled from over 50 locations in New York City. What is notable about Soiled is that it shows how artists can develop and are interested in developing quantitative methods for depicting soil diversity, and that soil art can capture not only individual portraits of soils but can be used as a reliable (and beautiful) tool for visualizing their bio-chemical characteristics. Albert Munsell’s gift to the soil science community was a hands-on, systematic tool for visually assessing mineral composition, humus content, and soil processes such as carbonate accumulation, among other things. It is only natural to expect that artists will contribute to soil sensing methods of the future.

Sounding soil

While the above projects use visual methods for capturing impressions and information about soil, we might next ask: what does soil sound like, and more importantly how could listening to soils improve our understanding of soil ecosystems? This question has led to fascinating research and development in both the sciences and the arts. Although Liu et al. (1993) had started testing acoustic methods for determining soil texture in the early 1990s, Tekeste et al. (2002) have since used sound waves to detect compacted layers, Koparde et al. (2018) have used echo signal processing in soil base profiling and sub-bottom profiling, and O’Shaughnessy and Sui (2018) have introduced radio frequency (RF) telemetry to precision agriculture. Worthy of mention on the artistic side are Florian Dombois’ (1999)⁵ “earthquake sounds,” Shu Lea Cheang and Martin Howse’s (2013)⁶ use of sensors to sonify rotting vegetables in “Composting the City/Composting the Net” (see also Howse’s “earthboot” and other noise works), Daniel Wolter’s

⁵See Earthquake Sounds Volume 1: Kobe 16/01/1995, 20:46 UT CD (numbered, unlimited edition). Online: floriandombois.net/works/earthquake-sounds.html Accessed Oct. 10, 2022.

⁶See project description for Composting the City/Composting the Net: <https://archive.transmediale.de/content/composting-the-city-composting-the-net>. Accessed Oct. 10, 2022.

(2019-2023) audio-visual analysis of different landscape types in the Biosphere Reservation Chorin, and Olle Corneer and Martin Lübke's (2009)⁷ "Harvest Terrafon," in which the Alunda Choir of Uppsala, Sweden, pulled an oversized gramophone through grooves in a field to amplify cropland soils.

Marcus Maeder et al.'s case study at the interface of artistic research, acoustic ecology and soil science, "Sounding Soil," represents one of the world's first studies dedicated to the acoustic phenomenology of soils. From 2017 to 2020 Maeder led a 20-member team as part of a research consortium between the Zurich University of the Arts ZHdK (Institute for Computer Music and Sound Technology), the Swiss Federal Institute for Forest, Snow and Agriculture Research (WSL), the Swiss Federal Institute of Technology Zurich (ETH), the National Soil Monitoring Office (NABO), and the Biovision Foundation for Ecological Development to develop methods for acoustic analysis of biological activity in topsoils. In 2017, 20 test sites were selected based on land use type and humus types present. For the recordings, sensors from the field of physical acoustics, including piezoelectric sensors, were modified for use in soils.⁸ Using piezoelectric sensors, vibrations were converted into electrical voltage, which in turn was amplified and captured as sound signals by a recording device. The recordings were made with probes about 30-cm deep and 60 cm in diameter, representing the area of the soil profile with the most biological activity. In addition, soils samples at the recording sites were taken with a puncture cylinder (Kempson corer, 30 cm ø, 20 cm length) and analyzed in the laboratory. Using the Berlese and Winkler method, soil fauna were extracted from the sample material and identified taxonomically and counted (cf. Maeder et al., 2019).

In October 2018 an immersive installation was opened at the Paul Klee Centre in Bern, consisting of a converted shipping container with a green roof and dimly lit listening room (Fig. 3). Small tubes for air and light in the ceiling represented macropores in the soil matrix. Sounds of the soil recordings included clicking and tapping sounds, vibrations, hissing, sighing and a variety of mechanical scratching movements. The muted walls of the structure created the impression of depth and seclusion reminiscent of the feeling of diving underwater. Surges of activity and intensity of audible phenomena in the recordings drew attention to the different soil factors in different locations. Weather and above-ground atmospheric sounds were added to create holistic "soundscapes" of the soils sampled.⁹ The term soundscapes, first introduced by R. Murray Schafer in 1977, refers to, "...any acoustic field of study. We may speak of a musical composition as a soundscape, or a radio program as a soundscape or an acoustic environment as a soundscape. We can isolate an acoustic environment as a field of study just as we can study the characteristics of a



Fig. 3 Sounding Soil, Installation in the Paul Klee Centre, Bern, Switzerland, 2018, Foto: Marcus Maeder. Reprinted with permission by the artist.

⁷See *Harvest Terrafon*. Project description on cdm online: <https://cdm.link/2009/11/a-gramophone-that-plays-the-earth-instead-of-vinyl-and-a-sonic-iphone-epidemic/> Accessed October 10, 2022.

⁸See also Murad et al. (2021) on the use of piezoelectric sensors to measure soil bulk density.

⁹To provide accessibility to those unable to travel to Bern, the recordings and a map and short description notes are available on the project website <https://soundmap.soundingsoil.ch/>. Accessed Oct.10, 2022.

given landscape” Sterne (2012). Like other acoustic engagements with soils, Maeder’s project encourages visitors to think differently about the local soundscapes that surround them and to imagine how the soil in their own backyard might compare acoustically to that in a field, a forest, or a construction site.

Beyond the esthetic qualities of Maeder’s soundscapes, the Sounding Soil study has relevance for soil health research. Biological activity is a key indicator of soil ecosystem functionality, especially in the assessment of soil biodiversity (see Wagg et al. (2014)). Sounding Soil investigates the complexity of soil soundscapes, asking whether the degree of land use, as well as humus composition, influence complexity values. Acoustic complexity plays an important role in ecoacoustic research because it can be used as an indicator of biodiversity for a particular unit of the landscape (see Pieretti et al., 2011). Within a soundscape, the diversity and complexity of sound data are measurable quantities. The acoustic complexity index (ACI) was used to quantify the acoustic complexity of soil organisms directly without including anthropogenic sound sources, which are commonly found in recordings (cf. Pieretti et al., 2011). Furthermore, significant differences in acoustic complexity were found between different land use types, with the greatest complexity for extensively managed pastures and the least for cropland soils. Forest soils and intensively managed pastures were in the middle range. The analysis further showed a weak positive correlation between the values of acoustic complexity and those of biological diversity of higher groups of soil organisms analyzed in the laboratory. For cropland and pasture soils, increased diversity of soil fauna correlated with decreasing management intensity, which in turn resulted in greater acoustic complexity. Among forest soils, there was a correspondence of curves between ACI measurements and the respective humus forms as habitat for soil fauna, with a gradual increase from mor humus to mull humus (cf. Maeder et al., 2019). Following developments in proximal soil sensing, acoustic analytical approaches can be considered a cost-effective alternative to more conventional forms of field measurement and laboratory analysis, in which disturbance to the soil matrix is reduced. Sounding Soil demonstrates that artistic research can serve as both as a tool for analyzing soil ecological properties and a means of facilitating sensory esthetic experience and raising awareness about the complexity of life underground.

Soil scenting

In her essay “The Aesthetics of Smells and Tastes,” philosopher Emily Brady (2005, p. 177) emphasizes the fundamental role of the so-called lower senses in appreciating one’s surroundings: “Sniffing and savoring not only form a fundamental pathway to sensory perception of our environment, but they also help define the quality and character of people, places, and events.” For soils, one of the most common associations and direct forms of esthetic appreciation is smelling and tasting, for example, in growing, preparing, and eating food grown in soil. In an attempt to evoke olfactory connections between soil and place, designer and food systems planner Lynn Peemoeller created a series of artisanal soil scents arranged on a product stand she designed for the international stakeholder conference “Global Soil Week” in 2013 in Berlin. Between plenary sessions on soil policy and sustainable development, conference guests were able to pause their agendas to tune into their own physical experiences with soils. For 5 days, Peemoeller handed out perfume samples of Tegel Forest soils, compost from a nearby community garden, and an ambiguous cocktail of street residue collected from Potsdamer Platz (all locations in Berlin), evoking powerful associations and memories by activating olfactory experience in the most unlikely of places. Peemoeller’s soil scent stand stood out against the typical conference venue tables with pamphlets and soil society merchandise, with the provocative question: what does soil health smell like?

The extraction of soil for use in perfumes has a long history, originating in the town of Kannauj, the so-called “perfume capital” of India, in the north-central rural state of Uttar Pradesh. In the book, *Rain: A Natural and Cultural History*, Cynthia Barnett (2016) tells the story of Australian mineralogists Isabel Joy Bear and Richard Grenfell Thomas, who published an article in the journal *Nature* in 1964 identifying the smell of fresh rain on the ground as coming from plant compounds called terpenes that were absorbed by clay soils. They called the smell “petrichor,” although it already had a name in Hindi—mitti attar, or earth perfume (Barnett, 2016). Today, one of the most popular commercially available earth perfumes is also called Petrichor. Distilled by the U.S.-based Demeter Fragrance Library, Petrichor similarly evokes the scent of soil after rain. Following the work of Bear and Thomas, a group of chemists at Brown University were able to identify the enzyme responsible for the smell of Petrichor as geosmin (see Jiaoyang et al., 2006). The presence of geosmin, literally “the smell of the earth,” is also a criterion for soil quality checklists.¹⁰ Olfactory descriptions for the screening include impressions such as sweet, foul, pleasant or unpleasant, musty, acrid, or soothing smells of soils.

Perfumers and other olfactory artists like Sissel Tolaas, Kate McLean, Lynn Peemoeller and others could help soil scientists, land managers, and the general public sharpen their vocabulary of soil odors by conveying abstract associations of combinations of mineral and organic compositions. While compounds such as alcohols, ketones, nitrogen-containing compounds, resins, tannins, etc., can be measured with gas sensors,¹¹ verbal associations remain beyond the realm of numerical data collection. Smells are experienced as olfactory sensations that directly inform the body about the health or hazard of soils. They form the foundation of what Porteous (1985) termed the “Smellscape,” or the cognitive ordering of spatial units based on particular smells or combinations thereof. The development of a soil smell vocabulary was the focus of Petrichor, a multi-part, multi-disciplinary science-art research project by artist Anaïs Tondeur in collaboration with anthropologists Marine Legrand and Germain Meulemans, ecologist Alan Vergnes, designer Yesenia Thibault-Picazo, and curator Nathalie Blanc. The team came together as part of the Sustainable

¹⁰See, for example, the Illinois Soil Quality Initiative (ISQI) for agricultural soil screening tests, which include smells as indicators.

¹¹See, e.g., recent research on artificial olfactory systems (AOS) by Zhu et al., 2019.



Fig. 4 Anaïs Tondeur (2016) Installation view, *Soils fictions*, Curator: COAL, Orangerie, Chamarande, France. Reprinted with permission by the artist.

Cultures Laboratory, initiated by COAL and the Domaine départemental de Chamarande during the 2015 International Year of Soils and continued with further iterations through 2018. Inspired by the work of Bear and Thomas, described above, the Petrichor group sampled soils at urban sites in Paris, Montpellier, and Aberdeen, in an attempt to develop an olfactory soil classification as envisioned by Parisian chemists in the 19th century.¹² Using methods of speculative fiction and anthropological fieldwork, the project invited participants, including gardeners, refugees, activists, graduate students, and elementary school students, to reflect on life in the city and the body's inherent connection to soil, even in predominantly sealed places where it is less visible. The results of the fieldwork were presented as an immersive installation (Fig. 4) of soil distillates, distillation apparatuses, and photographs and videos of soil sampling. As with Peemoeller's soil perfumes, Petrichor uses distinctive smells to evoke associative responses toward the development of a new vocabulary for soil smellscape. Analogous to work being done with soil color and color instrumentation design for soil analysis, olfactory artists demonstrate ways of assessing soil materials and sites that go beyond soil health checklists and proximal soil sensing to create sensory experiences with soils that evoke emotional responses and associative language. The description of the smell of a parking lot or forest after rainfall evokes an entirely different emotional response from a graph showing geosmin levels in the gases emitted from the same parking lot or forest.

Terroir

In Laura Parker's project, *"Taste of Place,"* a table is set and the California-based artist pours clumps of soil into wine glasses. She then activates olfactory experience by adding water and vigorously mixing. Guests spend time smelling, reflecting, remembering, smelling again, then trying to put their experience into words. "Salty, silty, mineral, wet, earthy, and green were just a few of the adjectives we threw out," writes Anne Zimmerman (2009), blogger for Cullinate, "all seemed like logical descriptors for soil." After sharing tasting notes, guests are served with vegetables, cheeses and eggs grown on the different soils in the wine glasses. As the event progresses, sensory connections are made between food and the soils they come from: a childhood memory of eating potatoes or

¹²See project description online: <https://anaistondeur.com/petrichor>. Accessed Oct. 10, 2022.

carrots from a grandparent's garden, the mouthwatering feeling of sauerkraut or kimchi made from a neighbor's cabbage. Parker's work trains guests to recognize and articulate "terroir" or the subtle differences in taste influenced by local conditions, such as soils, topographical relief, and the climate of a particular region.

Making connections between soil and food is a central strategy of many artists working with soils. Toland et al. (2018) initiated dialogues between artists and scientists who focus on food production and terroir in their work, including: a discussion between eco-art pioneer Agnes Denes' and former IUSS president Rattan Lal on the role of urban agriculture in food security and world peace; an exchange between soil scientist Brent Clothier and fourth generation farmer and artist Matthew Moore about terroir as a means of communicating the productive capacity of soils; social scientist Parto Tehrani-Krönner's ideas about "Meal Culture" as an alternative concept to food security; Roxanne Swentzell's experimental use of permaculture to promote indigenous diets; and introspection on soil pollution and community health brought to light in conversations between Bonita Ely and Richard MacEwen, and Amy Franceschini and Regine Debatty. Like Parker's Taste of Place, the works discussed were often created in participatory settings that invited reflection through sensory experience.

While these and other artists highlight the connection between food and soils in their work, the concept of terroir rarely refers to the eating of soil itself. Geophagy, or the desire to eat soil, offers a unique perspective on the sensory experience of tasting soil. In this respect, the work of Amsterdam-based artist masharu is worthy of note. Masharu's extensive artistic explorations of geophagy offer inspiring insight into the historical, culinary, and medical practices that have all but disappeared from Western cultures (Fig. 5).¹³ Drawing on their own childhood experiences of eating soil, masharu became interested in geophagic practices after moving to the Netherlands from Russia. In Amsterdam, masharu encountered clays and chalks from Tanzania, Nigeria, Ghana, and Suriname, among other places, and began interviewing clay traders and tracing the origins of edible soils. Limiting the palette to mineral samples and avoiding fresh organic topsoils for health reasons, masharu uses a variety of forms and formats to share their enthusiasm and knowledge about geophagy. These include a video archive and interactive website that describe different cultural practices of eating soil, tasting workshops, and an installation of variable dimensions for the growing collection. As part of the Edible Earth Museum, masharu has built an archive and database of more than 200 soil samples from 24 countries, acquired as gifts and on field trips, from online sources, and from health and specialty stores.

Anthropologist Sera Young (2012) has documented historical, ethnographic, and biomedical findings on picacism, the consumption of non-food artifacts including soil. While eating clay is considered unhealthy and even psychologically pathological in most of Europe and North America, it is still widely practiced by traditional healers in Africa and South America. In an interdisciplinary workshop in 2018, Young joined masharu and other artists, health researchers, geologists, and food safety experts to discuss the paradox of clay consumption as a healing and even medical act, and the taboo of pica as a psychological disorder.¹⁴ For masharu, anthropological questions about how and why people eat soil are central to their participatory investigations: it is important not only to create taste experiences, but also to understand the individual histories, memories, and cultural contexts of clay consumption.

For the project Claynialism in 2015 masharu collaborated with Arie Syarifuddin from the Jawa Barat (Indonesia) based, Jatiwangi Art Factory and the ceramist Lonny van Ryswyck, from Atelier NL (Netherlands).¹⁵ Together, they explored issues of food colonialism and searched for traces of geophagy in former Dutch colonies and also in contemporary cities in the Netherlands.



Fig. 5 Masharu, Pimba. Closing the Gap. "Interview with Sha about tastes, sensations, traditions, addictions, fears and joys." Amsterdam, November 2017. Reprinted with permission by the artist.

¹³See Pimba: Bridging The Gap. <http://masharu.nl/Pimba-Closing-the-gap>. Accessed Oct. 10, 2022.

¹⁴See Let them eat clay: IPR anthropologist bridges science and art at Amsterdam Workshop. Northwestern Institute for Policy Research: <https://files.cargocollective.com/721507/Let-them-eat-clay.pdf>. Accessed Oct. 10, 2022.

¹⁵See Claynialism. Collaboration with Arie Syarifuddin, Jatiwangi art Factory (ID) and Atelier NL (NL). <http://masharu.nl/Claynialism>. Accessed Oct. 10, 2022.

In a paper for the performance symposium Unknown Grounds, Masharu (2019, p. 31) asks: “What is behind the traditions of eating the earth? (...) How do the material properties of the earth influence taste?” Following the formal language and fascination about terroir described above, masharu and their colleagues collected mineral samples from the Netherlands, developed regional tasting notes, and conducted chemical analyses, which were in turn documented in interactive installations that ask visitors to consider the western suppression of geophagy as a cultural practice. In some iterations of the project, mineral soils from masharu’s archive were pressed into edible bars reminiscent of chocolate. On other occasions, wafer-thin clay tea cups could be eaten or filled with other edibles to create flavor combinations. At the “Our Living Soils” exhibition at Zone2Source Amsterdam, the artist passed out Stroopwafel with Luvos clay inside. In all presentation formats and exhibition sites, two elements of the work are central: edible clays and people coming together to taste them. Again, the participatory nature of the work stands out as a salient feature of artistic soil sensing. It is not soil sensing alone that constitutes the art, but rather the collective experience of interacting meaningfully with soils in diverse groups that are often distanced from soils.

Multisensory absorption

The projects described thus far have been singled out for their attention to particular senses. The complexity of soils, however, demands a multisensory sensibility to fully appreciate the soil ecosystems from which individual projects are developed. The ocular experience of perceiving color or acoustic experience of listening to a soundscape is inextricably linked to the sense of touch, smell, taste, temperature, gravity and magnetism. Early soil science terminology once recognized the sensory properties in naming soil types, such as terra rossa (Italian for red soil) or rendzina (Polish word describing a sound made by a plow hitting rocks). Soil art provides esthetic means for a “sensorial turn” in the soil sciences, analogous to proximal soil sensing (cf. Stein et al., 2004). The last case study we would like to present is a work that embraces the multisensorial reality of soil, which is also inherently social, in that sensory awareness can be shared, taught, and learned.

“Absorption” is generally known as the process by which one material (the sorbent) takes up and retains another material (the sorbate).¹⁶ In soil scientific contexts this amounts to the water holding capacity of the soil, which depends on a soil’s textural qualities, organic matter content, and topographical position in the landscape. Interpreted through artistic means, absorption can be understood as the experience of stimuli being fully taken up by all the senses—to become absorbed. The room installation, “Absorption,” by the Berlin- and New York-based artist Asad Raza is an invitation to become absorbed by the smell, feel, sound and sight of soil. Although the work has been exhibited in multiple locations, including Carriageworks, Sydney (2019), the “Down to Earth,” (Fig. 6) exhibition at Gropius Bau Berlin (2020), the Ruhr Triennial in Essen (2021) and most recently at the “We are Compost/Composting the We” exhibition at the Centre of Contemporary Art Glasgow (2022), Absorption is reconstructed under similar conditions each time with reproducible reliability. Despite local variability, Raza’s (2019) methods are clearly outlined: (1) recruit a coordinator, scientific advisors, and cultivators who will identify and gather local waste products and natural



Fig. 6 Asad Raza, Absorption, 2020. Cultivator tilling the “neosoil” for the exhibition “Down to Earth” at Martin Gropius ‘Bau, Berlin. Photo: A. Toland. Reprinted with permission by the artist.

¹⁶<https://www.encyclopedia.com/science-and-technology/physics/physics/absorption>

ingredients for the creation of a “neosoil” that is to be mixed together (e.g., using cement mixers) in order to fill an entire gallery space; (2) create engagement with visitors through regular tilling and measuring of the neosoil by hired cultivators who interact with visitors when they show interest in the work, answer questions, ask questions, and offer empty cloth bags to be filled and taken home; (3) give away all leftover neosoil to local allotment gardens and members of the community at the end of the exhibition.

Raza’s work makes reference to Walter de Maria’s iconic “Earth Room,” which was similarly exhibited in multiple locations and confronted visitors with a room full of earth-like substrates to be absorbed by multiple senses. The only remaining Earth Room, consisting of a meter-thick layer of soil was installed by de Maria in 1977 in a loft at 141 Wooster Street, NYC, and has been watered and weeded for over 40 years by DIA Art Foundation staff member Bill Dillworth. Dillworth’s maintenance of the Earth Room happens during off hours to allow visitors to experience the work without interruption. In addition to referencing de Maria’s use of minimalism to create or affect, “Absorption” also references “relational art,” described by Bourriaud (1998) as “art taking as its theoretical horizon the realm of human interactions and its social context, rather than the assertion of an independent and private symbolic space.” The point of “Absorption” is not only to have a multisensory experience, but to anchor that experience in the relationships with diverse materials and to those who care and cultivate them. The cultivators in this sense have been selected from a large pool of applicants and have gone through intense training that teaches them on the one hand how to measure soil conditions and on the other, how to read and communicate with an audience.

Touching or walking on the soil of de Maria’s Earth Room remains outside the boundaries of the artwork, which can only be viewed through a small doorway shielded by glass. While the Earth Room encourages contemplation of soil as an omnipresent natural material in juxtaposition to the sealed city streets outside the gallery, Raza explicitly invites visitors to touch and literally become “absorbed” with soils directly sourced from the city outside. These include, for example, in the Glasgow iteration: washed fine-grained building sand, dredged silt from Scottish canals, red terracotta grogged marl, crushed rubble, crushed oyster shells (formerly used as building mortar), spent hops and barley from local breweries, coffee grounds and cocoa husks, hair from the local barber, vegetable waste from local restaurants, and peat-free compost from the municipality. Compared to de Maria’s ubiquitous “Earth” of unknown (presumably agricultural) origin and composition, the materials in “Absorption” reflect the unique economic and political history of the city, listed clearly on a poster leading into the room and explained by the cultivators.

Coming in from the street, one steps onto the soft and uneven ground, comparable to that of a freshly plowed field. The acoustic environment overwhelms the body in muted silence broken up only by the sound of rhythmic raking while the air is heavy with a smell reminiscent of burning chocolate and rainfall. A trained cultivator explains that she is digging freshly discarded cacao waste from a local roastery into the substrate. Its composition represents an ideal-typical soil of 30% sand, 30% silt, 30% clay and 10% organic matter and calcareous materials. A collection of tools and instructions typically used in field work and soil mapping introduces visitors to basic instruments of soil assessment. After systematically folding in new materials with spades, cultivators rake and spray the surface and explain how to take pH measurements, use Munsell charts to determine organic content via color codes, and perform finger tests to train visitors how to recognize soil textures. The artwork uses the same techniques for assessment in the forest or field to assess the health of the neosoil “as neologism”. A group of soil scientists visiting the exhibition in Glasgow on occasion of the 22nd World Congress of Soil Science debate whether “Absorption” is a room full of amendments, a technogenic substrate, or a constructed soil, generating interest not only from general audiences, but inviting reflection on the meaning and origin of technical terms sometimes taken for granted by experts.

In reference to Latour’s (2015) Gaia Lectures, published during the International 2015 Year of Soils, “Absorption” is finally a critical examination of human presence in the planet’s “critical zone” and in the pedosphere in particular, as well as the presence of soil in human zones. Before visitors leave the exhibition, they are encouraged by cultivators to put some neosoil into bags to take home and further cultivate. In this way, visitors are not only given a gift of (portable) sensory experience, but they are expected to appropriate the artwork themselves by relocating materials that have already been relocated and caring for something that has already been cared for. Challenging standard definitions of soil, which require an extended period of place-based weathering and evidence of soil formation, Raza’s neosoil implies a definition that accepts displacement as a condition of the Anthropocene but also considers care as a necessary criterion for soil being soil. Degraded, carbon depleted soil loses its status of soil and becomes “dirt” in the absence of care. As a work of social sculpture, Absorption empowers visitors to “care forward” and take greater responsibility for soils, starting with one bag. It is in their hands from that moment on to find a place for soil in their lives. The small print at the bottom of the cloth bags reads: “composed of animate stardust, congregated here, absorbed in this.”

Conclusions and outlook

This chapter presents a representative sampling of a much larger field of artistic research on soils. As an overview, it by no means claims to be exhaustive. Together with advancements in proximal soil sensing technologies, artists contribute new insights on soil sensing by facilitating multisensory experiences, developing new vocabularies for what is sensed, raising awareness about soil ecosystems and their complexities, and producing innovative research on soil phenomenology that can be cited by artists and scientists alike and inspire new interdisciplinary collaborations. Artistic soil sensing, with its diverse formats, foci, and forms of presentation and mediation, can target different audiences and segments of society, playing an important role in soil communication, life-long learning and soil literacy-building efforts. Given the long history of sense-based measurement in soil science, from the use of finger tests and Munsell color charts to body-extending sensing technologies to gather soil data at different spatial and temporal scales, soils offer fertile ground for exploring human understanding of sensory perception itself. Artistic soil sensing is thus

not only a relevant area of research and development for soil systems science and the arts, but also for the fields of sensory studies, media anthropology and phenomenology. Beyond the five Aristotelian senses, it is worthwhile to consider the ways soil organisms rely on a host of additional senses such as temperature, gravity, balance, and magnetism, among others to navigate life in the pedosphere. Extending the realm of soil sensing not only to new technologies but also to other non-human organisms provides a multispecies context for further research on the subject.

What is only briefly touched upon in this chapter and urgently needs attention is the potential of artistic soil sensing in contexts of environmental justice, food sovereignty and seed saving, rights to nature, and indigenous soil management and land-back movements. Howes and Classen (2014), among other scholars of sensory studies, have emphasized the need for understanding sensory stimuli in social contexts. For the 2022 exhibition “We are compost/Composting the We at the Center for Contemporary Art Glasgow,” Désirée Coral makes reference to gustatory experience with heirloom seeds from the Americas and copper spoons that throw shadows in the shape of South American countries in “Eating the Ancestors.” Compared to some of the extensive earth pigment collections mentioned earlier, the display of soil-filled jars in the “Memorial to Peace and Justice” exhibition in Montgomery, Alabama has a haunting effect in the realization that the colorful, neatly ordered soil samples are gathered from selected sites of over 4000 documented terror-lynchings of African Americans in the United States. Meanwhile, the multisensory esthetics of Walter de Maria’s “Earth Room” are recontextualized in the earthworks of Delcy Morelos, who pays tribute to Andean and Amerindian cosmologies through the composting of aromas from cassava flour, cacao powder, and spices from the artists native to Columbia into large scale displays of soil materials. Otobong Nkanga’s “There’s No Such Thing as Solid Ground” (2020, Martin Gropius Bau Berlin) similarly uses a multisensory approach to explore the tensions between extractive relationships and practices of care and repair. These and other artists create sensory experiences to highlight the social and political entanglements of soils as historical archives with witnessing and storytelling capacities, and to frame soils as effective media for healing and reparation. Given the violent history of colonial land take, displacement of agricultural communities, destruction of prairies, forests, and wetlands for highest economic gain, and the erasure of indigenous land use practices, knowledge and technologies, both the sciences and arts are called upon to reflect, repair and decolonize. Artists can show that raising awareness about soils is not only about facilitating sensory experience and esthetic appreciation in and of itself but can offer space for critical reflection on the consequences of extractive capitalism and colonialism, and the roles of science in the past, present and future of soil cultivation. Beyond the display of color and texture, taste, smell, and sound, Soil art can do more.

References

- Adamchuk V, Allred B, Doolittle J, Grote K, and Viscarra Rossel RA (2015) *Tools for Proximal Soil Sensing. Soil Survey Manual. Natural Resources Conservation Service. U.S. Department of Agriculture Handbook 18th Edition.*
- Adams C and Montag D (2016) *Soil Culture: Bringing the Arts Down to Earth.* Manchester, UK: Gaia Project Press.
- Barnett C (2016) *Rain: A Natural and Cultural History.* New York: Penguin Books.
- Bourriaud N (1998) *Relational Aesthetics.* Dijon: Les Presse Du Reel, 14.
- Brady E (2005) Sniffing and savoring—The aesthetics of smells and tastes. In: Light A and Smith JM (eds.) *The Aesthetics of Everyday Life*, pp. 177–193. New York: Columbia University Press.
- Feller C, Landa E, Toland A, and Wessolek G (2015) Case studies of soil in art. In: *Soil—An Interactive Open Access Journal of the European Geosciences Union.* EGU.
- First-Arai L (2019) *Artists Find Common Ground in Soil from Around the World.* Hyperallergic. <https://hyperallergic.com/487683/artists-find-common-ground-in-soil-from-around-the-world/>.
- Ford BM, Stewart BA, Tunbridge DJ, and Tilbrook P (2021) Paper chromatography: An inconsistent tool for assessing soil health. *Geoderma* 383: 114783. 2021. ISSN 0016-7061.
- Gabrys J (2016) *Program Earth: Environmental Sensing Technology and the Making of a Computational Planet.* Minneapolis/London: University of Minnesota Press.
- Howes D and Classen C (2014) *Ways of sensing. Understanding the Senses in Society.* London: Routledge.
- Jiaoyang J, Xiaofei H, and Cane DE (2006) Geosmin biosynthesis. Streptomyces coelicolor germacradienol/germacrene D synthase converts farnesyl diphosphate to geosmin. *Journal of the American Chemical Society* 128: 8128–8129.
- Koparde P, Wagh R, and Phatak D (2018) Soil base profiling/sub-bottom profiling by echo signal processing. In: *IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA)*, 41–46. <https://doi.org/10.1109/CSPA.2018.8368682>.
- Krzywoszynska A (2021) *Sensing Soils to Live With(in) Gaia. Keynote Presentation at Sols Sous-Sols Dans la Transition Socio-Ecologique.*
- Landa ER (2004) Albert H. Munsell: A sense of color at the interface of art and science. *Soil Science* 169: 83–89.
- Latour B (2015) Face à Gaïa. Huit conférences sur le nouveau régime climatique. In: *Paris: La Découverte—Les Empêcheurs. Deutsche Übersetzung “Kampf um Gaia: Acht Vorträge über das neue Klimaregime”.* Berlin: Suhrkamp Verlag. 2017.
- Liu W, Gaultney LD, and Morgan MT (1993) Soil texture detection using acoustic methods. In: *ASAE Paper 93–1015*, St. Joseph, MI: American Society of Agricultural Engineers.
- Maeder M, Gossner MM, Keller A, and Neukom M (2019) Sounding soil. An acoustic, ecological & artistic investigation of soil life. In: *World Forum for Acoustic Ecology (WFAE) (Hg.): Soundscape: The Journal of Acoustic Ecology. Sound + Environment: Sonic Explorations. Volume 18, 18 / 2019.* Fairfield, Victoria, Burnaby, BC, 5–14.
- Masharu (2019) *Museum of Edible Earth. Unknown Grounds Performative Symposium at VHDG Kunstinitiatief Voorheen de Gemeente. Reader.* 37–46. https://files.cargocollective.com/721507/03A_MashaRu_publication_UnknownGrounds.pdf.
- Munsell AH and Cleland TM (1921) *A Grammar of Color: Arrangements of Strathmore Papers in a Variety of Printed Color Combinations According to The Munsell Color System.* Mittleague, Massachusetts: The Strathmore Paper Company.
- Murad MOF, Minasny B, Malone B, and Crossing K (2021) Measuring soil bulk density from shear wave velocity using piezoelectric sensors. *Soil Research* 59: 107–117. <https://doi.org/10.1071/SR19395>.
- O’Shaughnessy SA and Sui R (2018) Advanced tools for irrigation scheduling. In: *Advances in Agricultural Machinery and Technologies*, 1st edn. Boca Raton: CRC Press.
- OECD (2015) Frascati manual 2015: Guidelines for collecting and reporting data on research and experimental development. In: *The Measurement of Scientific, Technological and Innovation Activities.* Paris: OECD Publishing. <https://doi.org/10.1787/9789264239012-en>.
- Pfeiffer E (1984) *Chromatography Applied to Quality Testing.* Wyoming and Rhode Island: Bio-Dynamic Literature.
- Pieretti N, Farina A, and Morri D (2011) A new methodology to infer the singing activity of an avian community: The acoustic complexity index (ACI). *Ecological Indicators*. vol. 11, pp. 868–873. Amsterdam: Elsevier.

- Porteous JD (1985) Smellscape. *Progress in Physical Geography: Earth and Environment* 9(3): 356–378. <https://doi.org/10.1177/030913338500900303>.
- Raza A (2019) *Absorption Technical Manual. Guidelines and Specifications for Cultural Institutions and Curators Producing New Iterations of the Work Titled Absorption by Asad Raza*.
- Rittel HW and Webber MM (1973) Dilemmas in a general theory of planning. *Policy Sciences* 4(2): 155–169.
- Rossel RAV and Bouma J (2016) Soil sensing: A new paradigm for agriculture. *Agricultural Systems* 148(2016): 71–74. ISSN 0308-521X. <https://doi.org/10.1016/j.agsy.2016.07.001>.
- Schneider B and Zemanek E (2020) *Spürtechniken. Von der Wahrnehmung der Natur zur Natur als Medium*. Sonderheft der Medienobservationen. www.medienobservationen.de.
- Stein BE, Spence C, and Calvert G (2004) *The Handbook of Multisensory Processes*. Cambridge, Mass: MIT Press.
- Sterne J (ed.) (2012) *The Sound Studies Reader*. London: Routledge. (Hg.).
- Tekeste MZ, Grift TE, and Raper RL (2002) *Acoustic Compaction Layer Detection*, ASAE Paper 02–1089. St. Joseph, MI: American Society of Agricultural Engineers.
- Thompson Klein J (1990) *Interdisciplinarity: History, Theory, and Practice*. Detroit, MI: Wayne State University Press.
- Toland A, Wessolek G, and Churchman G (2014) Picturing soil—Aesthetic approaches to raising soil awareness in contemporary art. In: Landa E (ed.) *The Soil Underfoot: Infinite Possibilities for a Finite Resource*, pp. 83–101. Taylor & Francis.
- Toland A, Noller J, and Wessolek G (2018) *Field to Palette—Dialogues on Soil and Art in the Anthropocene*. Boca Raton: CRC Press/Taylor and Francis.
- Ugolini F (2010) Soil colors, pigments and clays in paintings. In: Landa ER and Feller C (eds.) *Soil and Culture*, pp. 67–82. Dordrecht, Heidelberg, London and New York: Springer.
- Wagg C, Bender SF, Widmer F, and van der Heijden MGA (2014) Soil biodiversity and soil community composition determine ecosystem multifunctionality. *Proceedings of the National Academy of Sciences of the United States of America* 111(14): 5266–5270. <https://doi.org/10.1073/pnas.1320054111>.
- Young SL (2012) *Craving Earth: Understanding Pica—The Urge to Eat Clay, Starch, Ice, and Chalk*. New York: Columbia University Press.
- Zhu L, Jia H, Chen Y, Wang Q, Li M, Huang D, and Bai Y (2019) A Novel method for soil organic matter determination by using an artificial olfactory system. *Sensors* 19(15): 3417. <https://doi.org/10.3390/s19153417>.
- Zimmerman A (2009) Culinate.com <http://www.culinate.com/search/q.vt=top,q=dirt/235913>.

Soil science in the criminal justice system

Lorna Anne Dawson, Centre for Forensic Soil Science, James Hutton Institute, Aberdeen, United Kingdom; Forensic Science, School of Pharmacy and Life Sciences, Robert Gordon University, Aberdeen, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction	522
The history of forensic soil science	523
Theory of soil transfer	523
Soil characteristics	523
Soil sampling at a crime scene	523
Soil examination and methods of analysis	525
Soil evidence	526
Soil databases	526
Soil classification	528
Human-made soils	528
Chain-of-custody	528
Initial description and characterization of samples	528
Soil inorganic analysis	529
Soil organic characterization	529
Soil particle characterization	529
Evaluation of soil information	529
Communication	530
Concluding remarks	530
References	530

Abstract

Forensic soil science is the application of soil science in the context of the law. Soil materials can be used in forensic investigations to evaluate potential associations to support law enforcement agencies in investigating criminal, environmental, and serious and organized crime. Soil recovered from a questioned item, with an unknown history can be compared with known locations such as crime scenes and alibi locations. Forensic soil science is used to assist in search operations where soil can help narrow down areas to focus the ground search activities (provenancing) and as trace comparison and evidence in court. Established concepts, standardized approaches and professionalism have been developed in soil science and should be carefully applied regarding the legal context to allow the appropriate and safe application of soil science within the criminal justice systems worldwide.

Glossary

ENFSI APST The Animal Plant and Soil Traces expert working group (APST) supports the aims and objectives of the European Network of Forensic Science Institutes (ENFSI) in the area of casework analysis of all kinds of biological traces of non-human origin and soil traces.

FTIR Fourier transform infrared (FTIR) spectroscopy is an established analytical tool for studying soil. The sensitivity of FTIR spectroscopy to local structural environments of infrared active soil components offers a multitude of applications that complement conventional soil science analytical techniques.

GIS A geographic information system (GIS) is a type of database containing geographic data (that is, descriptions of phenomena for which location is relevant), combined with software tools for managing, analyzing, and visualizing those data.

IR Infrared spectroscopy (IR spectroscopy or vibrational spectroscopy) is the measurement of the interaction of infrared radiation with matter by absorption, emission, or reflection. It is used to study and identify chemical substances or functional groups in solid, liquid, or gaseous forms.

IUGS IFG The International Union of Geological Sciences (IUGS) Initiative on Forensic Geology (IFG) was established at UNESCO headquarters in Paris, France on 22 February 2011. The aim of the IUGS-IFG is to develop forensic geology internationally and promote its applications.

Pedology (From the Greek *pedon* = soil) Is the soil science discipline concerned primarily with understanding the variety of soils and their distribution and is most directly concerned with the key questions concerning sampling, descriptions, processes of soil formation including the quality, extent, distribution, spatial variability, and interpretation of soils from microscopic to megascopic scales. This description and interpretation of soils can be used in addressing the questions 'What is the soil like?' and 'Where does it come from?' (i.e., provenance determination) in studies relating to characterizing and locating the sources of soils to make forensic comparisons.

Provenancing Provenancing is the capability to spatially constrain the likely region of origin of a questioned sample of earth-related material.

Questioned sample A questioned sample is a sample with unknown origin, often recovered from a suspect or an item associated with a suspect.

SEM Scanning electron microscopy is a type of electron microscopy that produces images by rastering a focused electron beam across the surface of a sample. This technique produces detailed images of a specimen's surface. Its applications include surface topography and spatially resolved elemental composition using Energy Dispersive Spectroscopy (EDS) or Energy Dispersive X-ray Spectroscopy Analysis (EDXA). SEM produces images up to a magnification of 500,000 $\times$ and a typical resolution power of a few nanometers.

Wax markers The cuticular wax compounds found in plant leaves, which slowly decompose and break down in soil, leave behind a distinctive profile reflecting the history of plant inputs to the soil. The most common and widely studied wax-marker compound class is the hydrocarbons, including *n*-alkanes, and the free and esterified long-chain fatty alcohols (high molecular-weight, straight-chain primary alcohols, derived from natural fats and oils) and fatty acids (carboxylic acids with a long aliphatic chain, containing a hydrocarbon chain and a terminal carboxyl group).

XRD X-Ray Diffraction or XRD, is a technique for analyzing the atomic or molecular structure of materials. It is non-destructive, and works most effectively with materials that are, or part, crystalline.

Key points

- Forensic soil science is the application of soil science in the context of the law.
- Forensic soil science can provide data (intelligence) to assist in search operations where soil, minerals and botanical information can help narrow down geographical areas to focus the ground search activities (provenancing).
- Soil recovered from questioned items, with an unknown (questioned) history, such as vehicles, clothing, and tools, can be compared with specific known locations such as a crime scene (trace evidence).
- Soil forensic characterization involves a staged strategic approach; first with visual examination of physical characteristics of soil (color, texture and morphology) then chemical, physical and or biological analyzes.
- Established concepts, standardized approaches and professionalism have been developed in soil science and are carefully applied to allow the appropriate and safe application of soil science within the criminal justice systems.
- Effective communication is vitally important, from the initial discussion with the investigating officer to the presentation of evidence in court.

Introduction

Forensic soil science is the science of the application of soil information to help answer legal questions, problems, or test related scenarios or hypotheses and to help provide evidence in a court of law. Soil is composed of a range of different-sized mineral particles (described as sand, silt, and clay), the inorganic components, and the organic matter components (often described by its individual organic compounds). Soil has many complex biological, chemical, physical, mineralogical, organic, and hydrological properties, which change through space (horizontally and vertically down a soil profile) and through time (soil can take thousands to millions of years to form and minutes to be lost). Soil is found across most of our terrestrial world and can be both natural and human made. Natural soils contain natural minerals and natural organic materials, which in natural soils reflect the underlying geological parent material, as well as the land use, climate, and vegetation that has grown there. Human-made soils reflect human influence on their formation, mainly in urban environments, which often results in the presence of small amounts of manufactured materials, such as brick fragments, paint flecks, plastics or fly ash. Soil components and resultant characteristics are diverse, and this potential diversity and variation has enabled forensic soil examiners to distinguish soils that otherwise may appear similar to the untrained eye (e.g., Murray and Tedrow, 1992; Pye, 2007; Ruffell and McKinley, 2008; Ritz et al., 2008; Murray, 2011; Fitzpatrick, 2013; Dawson and Mayes, 2014; Donnelly et al., 2021). Traditional soil science is a different discipline from forensic soil science, the latter requiring professional knowledge of the discipline of soil science, as well as knowledge of the legal aspects. These include adherence to legal protocols, an understanding of crime scene protocols, evidential requirements of forensic practitioners, chain of custody, and the nature of legal constraints, including effective communication in court. It is critical to establish the most appropriate methods to use, both for examination, analysis and for the evaluation of any resultant data and how to communicate those findings in a court of law (Dawson et al., 2021). Soil and related materials can be used for intelligence work such as assisting police in search operations with soil recovered from a spade, for example (Aitkenhead et al., 2014; Harrison et al., 2016). In addition, comparison of trace evidence analysis between soil found adhering to a questioned item such as a suspect's shoe or vehicle with soil from a crime scene and alternative locations, enables evaluation of the likelihood that the questioned soil originated from the crime scene or alternatively, another proposed location (Dawson and Morgan, 2015).

The history of forensic soil science

Forensic soil science has only recently (since the 1990s) been accepted in courts of law. However, the principles have been used since Roman times where hooves of enemies' horses were examined to ascertain from where they had most likely traveled. The first published case of a forensic soil comparison was with sand characteristics which helped police solve a crime that took place on a Prussian railroad in April 1856. A barrel containing silver coins had been emptied and refilled with sand during transit and Prof. Christian Gottfried Ehrenberg sampled soil from all the railway stations along the line, using a light microscope. He examined features of the soil particles (color and shape) and compared the soil substituted in the barrel to the sand sampled at the rail stations, narrowing it down to the one from which the sand originated (Science and Art, 1856). Sir Arthur Conan Doyle published several stories involving his fictional character, Sherlock Holmes, where forensic soil comparisons were carried out as part of criminal investigations. Holmes's assistant Dr John Watson described him as being able to "tell at a glance different soil from each other. After walks, has shown me splashed upon his trousers, and told me by their color and consistence in what part of London he had received them." In the early 1900s George Popp, a forensic scientist in Frankfurt, successfully examined soil, minerals, and dust from items of clothing to help solve two criminal cases (Murray, 2011). In 1893, the Australian scientist Hans Gross suggested that "perhaps dirt on someone's shoes could tell you more about where a person had last been than toilsome inquiries". Soil information was starting to be used by police in the United States, United Kingdom, Russia, and Australia between the 1900s and the 1990s. However, soil analyzes had become very specialized and expensive for use in most forensic laboratories and forensic agencies worldwide. As a result, soil forensic evidence was often missed or ignored, often buried within other trace evidence. Soil science techniques were generally not deployed on complex forensic cases until the 2000s. At that time, it was mainly mineralogy and palynology that was used, often by individuals within university departments. The recent surge in research around the world into the mineralogy, geochemistry and biology of forensically important soil materials has helped bring forensic soil science to the forefront (Dawson and Hillier, 2010; Donnelly et al., 2021). Forensic soil science is now being used to deliver powerful forensic evidence with significant benefits to criminal investigations in many countries across the world (Di Maggio et al., 2017).

Theory of soil transfer

The transfer of trace evidence is governed by the Locard Exchange Principle, which states that "when two surfaces come into physical contact there is a mutual exchange of trace evidence between them" (Locard, 1930). The exchange can take the form of soil from a location transferring to the shoes of a person who walked through that area. These types of transfers are referred to as primary transfers (e.g., evidence is transferred from the soil surface to the shoe and later recovered from the shoe, such as in the tread of the sole). After the point of primary transfer, any subsequent movements of that soil material such as from the shoe to the footwell mat of a vehicle are referred to as secondary transfers. These secondary transfers of soil materials can be significant in evaluating the nature and source(s) of contact. Soils can provide information linking (or excluding) persons to or from crime scenes, control sites, or alibi sites. Higher order transfers such as tertiary transfers of soil can also occur, such as from a scene to footwear to a carpet surface and to another person walking and encountering that area. Although it is useful to be aware of such high-level transfers, they can present interpretative problems because the ultimate source of trace evidence may be difficult to identify (Murray, 2011). Most cases involve primary or secondary associations between questioned items such as footwear, vehicles, tools such as spades, and clothing to identify locations of interest. Soil trace evidence can link an object or person to the scene of a crime, as well as rule out an individual suspect or to support a proposed alibi.

Soil characteristics

Soil materials can potentially provide powerful 'contact trace' evidence that assists in criminal investigations. They can be distinctive, have a strong likelihood of transfer and persistence (especially when wet and when they contain some clay and organic matter), can be nearly invisible and be relatively easily collected and characterized. There are thousands of different soil types, with a vast range and complexity that enables forensic soil examiners to distinguish between individual soil samples submitted for examination and characterization. These characteristics vary horizontally across a landscape and vertically with depth down a soil, which can be useful for comparison at a burial site. Human created features within soils (artifacts such as paint and plastics) can make soils even more distinctive. Soil also usually has a strong capacity to persist, particularly the fine sized fractions (clay and silt) and the organic matter in soils. When the soils are wet, they transfer and persist more easily than when dry. The presence of small amounts of fine soil material is often not visible to the naked eye, contrasting with other more colorful trace materials such as blood and paint.

Soil sampling at a crime scene

The relevance of soil evidence should be accurately documented and collection and preservation of any soils and related materials must be maintained rigorously to retain the integrity of the soil evidence as part of the chain of custody (Ruffell and McKinley,

2008). The location and number of reference samples to collect from individual crime scenes or alibi sites will depend on several factors—including activity, variability, soil type, weather conditions, vegetation cover, contact points such as footwear marks, tyre treads, all of which will affect the depth of sampling, number and size of samples to be collected and type of tool/container used (see Fig. 1 and Pirrie et al., 2021a, 2021b). If the scene involves a burial site, samples will need to be collected at any potential contact points with tools such as spades, and from every soil horizon, at the base of the burial environment and from the surface layer where contact with footwear may have occurred.

Soil sampling schemes can be divided into ‘targeted sampling’ of localized areas where distinguishing features are noted in the soil (e.g., tyre marks) or on some other surface (e.g., a footprint on the floor) or ‘random sampling’ in cases where there is an absence of obvious features at or around a location of interest (i.e., a systematic sampling of the location to obtain an unbiased sample). Clean sampling tools should always be used (i.e., an alcohol wiped trowel). Artist’s palette knives are useful for sampling very thin layers of mud or surface contact soil (Fig. 2). It is recommended that samples are stored carefully for transport in ‘rigid

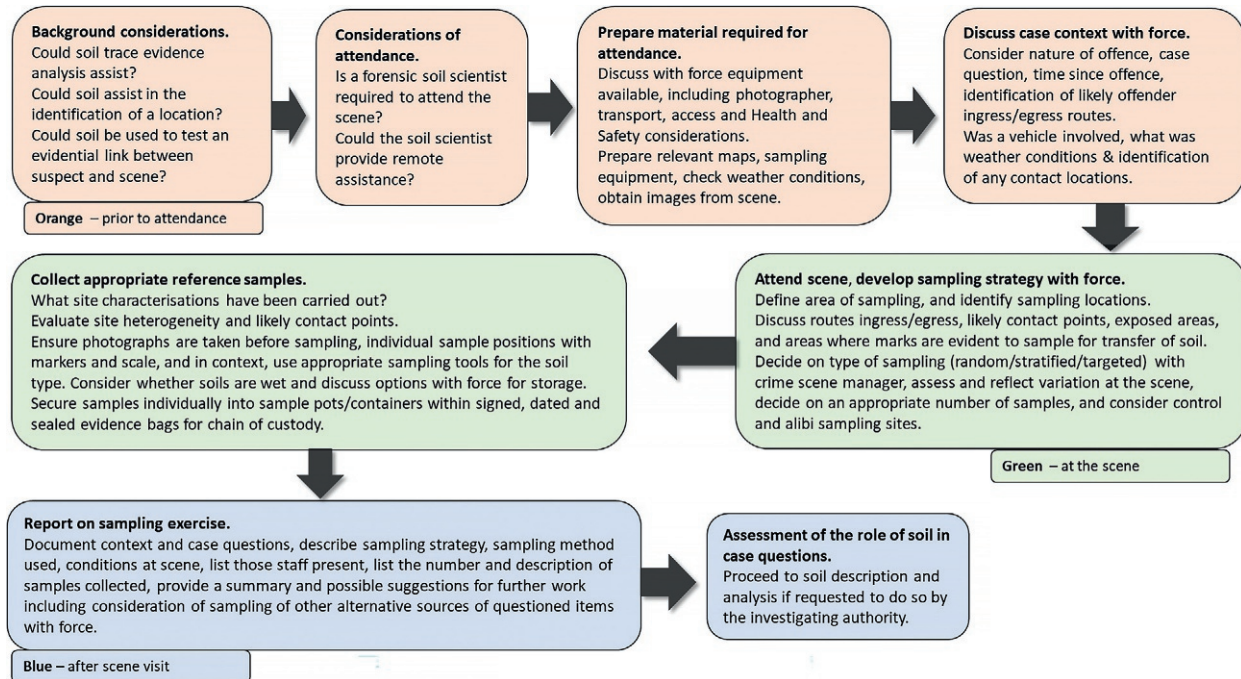


Fig. 1 Proposed considerations for soil sampling at a crime scene. Also see material in Pirrie D, Ruffell A, Dawson LA and McKinley J (2021) Crime scenes: Geoforensic assessment sampling and examination. In: Donnelly L, Pirrie D, Harrison M, Ruffell A and Dawson L (eds.), *A Guide to Forensic Geology*. Geological Society: London.



Fig. 2 Example of the method of collection of a surface soil sample at a crime scene. Lorna Dawson.

plastic containers' rather than loose within bags to help retain the soil structure to allow sub sampling of individual layers in soil stratigraphy if appropriate. Although not always practical, if the soil is in a moist condition, the sample should ideally be air dried before packing. If biological material such as plant leaves are within the soil sample, it should be packaged using clean cardboard box or paper bags and stored at $\sim 4^{\circ}\text{C}$ to prevent deterioration.

Soils can be collected by scene of crime officers at crime scenes, by forensic examiners when examining items of physical evidence (such as a questioned item of footwear) for other evidence such as blood or fibers, or by the forensic soil scientist. (See Fig. 5; Pirrie et al., 2021a).

Soil examination and methods of analysis

In the soil science laboratory, a magnifier lens and binocular light microscopes are usually the first piece of equipment to be used (after viewing the samples with the naked eye) (Fig. 3). Soil materials are first described using conventional soil pedological features (i.e., soil color, texture, structure). Soils can then be characterized using a range of analytical methods. Stones and vegetation fragments are removed with sterile tweezers and can be described separately if required. It is important that individual soil traces with similar color and texture are recovered separately to provide the most appropriate comparison with single source reference samples from any site of interest (Fig. 4).



Fig. 3 Examination of an item of footwear for subsequent analysis. Lorna Dawson.



Fig. 4 Sampling and recovery of soil and vegetation from an item of footwear for subsequent analysis. Lorna Dawson.

To characterize the inorganic fraction quantitatively, methods such as X-ray diffraction (XRD) (for the mineral profile) or inductively coupled plasma (ICP) (for the elemental profile) can be used, while chromatographic methods such as gas chromatography (GC) can be used for the organic fraction profile (Dawson and Hillier, 2010; Dawson, 2017). Where only a trace amount of sample is available, Fourier Transform Infrared Spectroscopy (FTIR), Raman spectroscopy, Scanning Electron Microscopy (SEM), Energy Dispersive X-ray Spectroscopic Analysis (EDXA) can be used. The method adopted depends on many factors, including the size and condition of the questioned samples, equipment available and the characteristics of local soils (Fig. 5). If possible, the methods chosen first should be non-destructive, followed by quantitative methods where within sample variation can be characterized and compared with related quantitative databases. The use of Standard Operating Procedures (SOPs) is recommended (e.g., Testoni et al., 2019) as are standardization and accreditation of methods where possible.

Methods for characterizing soils for a forensic comparison, broadly involves subdividing methods into four stages, each comprising several steps and involving a combination of techniques (e.g., various descriptive or morphological and analytical methods). Consequently, soil characterization requires a multidisciplinary approach, which combines descriptive and analytical methods together with information on spatial context such as that in the maps of geographical information systems (GIS). The progression of a soil examination through each of the stages (Fig. 5) will depend on several factors such as the amount of sample available, the environment and land cover at the scene, results from the initial stages of examination and the digital information available. Several forensic practitioners have suggested a similar sequence or succession approach of sediment or soil analysis for exclusionary or comparative forensic investigations. Recently a guide was published by the European Network of Forensic Science Institutes (ENFSI APTST, 2019), and a book was published by the Geological Society of London through the International Union of Geological Sciences, Initiative on Forensic Geology (IUGS, IFG) group, outlining guidelines to options for search and evidence collection, analysis, and reporting (Donnelly et al., 2021). However, there is no one single authoritative scene of crime manual or laboratory methods manual for soils; the justice system of the country in which the work is being carried out will dictate many of the most appropriate methods and approaches adopted. The approach and method of each forensic situation should be taken on its own merits (such as size and condition of questioned sample, type of soil, degree of mixing, etc.) and the approach chosen according to existing conditions. However, any approaches adopted must involve standard approaches for the recording, describing, and chain of custody of case materials.

Forensic soil comparisons of a questioned soil sample from an unknown origin with one or more soil samples of a known origin (such as a crime scene or alibi location) are preferably based on several properties. This enables the expert to evaluate the significance of observed similarities and differences, and to arrive at a conclusion regarding a possible association (See Fig. 2; Pirrie et al., 2021a, 2021b). Soil can be used to indicate or compare provenance, and therefore be used as 'intelligence' to narrow areas of search during an investigation and subsequently for trace evidence comparison if required by the investigating authority. Evaluative comparison of soil on one article of evidence compared to soil on another, or compared to a known location, can and has been used as evidence in many courts of law (e.g., Dawson and Hillier, 2010; Dawson and Mayes, 2014). Forensic evidence is often available only as very small amounts of soil material and will therefore often limit the evidential strength of such comparisons.

Most forensic cases involving soil materials are usually complex and the challenges of associating relevant information from one source with another often requires the application of sophisticated field and laboratory methods and soil science discipline-based knowledge. The methods used for comparing soil samples must be practical, inexpensive, accurate, and applicable to both small and large samples.

Soil evidence

The aim of forensic soil analysis is to test for a potential association (link) between a soil sample taken from a questioned item (e.g., shoes, clothing, tools, or vehicle) with a specific location such as a crime scene. Methods of sample collection and analyzes chosen must be able to discriminate between soil samples from separate locations. All sampling methods, sample storage security, soil characterization, analytical methods, and data handling must be subject to rigorous procedures. They should be transparent and have a secure chain of custody, so that they can be discoverable and able to be questioned if presented in court (Ruffell et al., 2021). The forensic investigation of soil usually involves site investigation and characterization; soil sampling to represent the crime scene and additional potential contact source locations; soil collection and recovery of soil from one or more questioned items; soil characterization, evaluation, and comparison, and reporting and communicating the results found, often to a court.

Soil databases

Mapping of the surface and subsurface of both natural and human-made soils provides crucial information in relation to geographical location, function, land use, and management. In many countries in the world, soil data have been encoded into computer-compatible form. In Scotland the SE web is a free source of soil data (Home | Scotland's environment web) and for the UK the UK Soil Observatory is available (UK Soil Observatory). Each nation has their own maps and digital data, with varying extent and resolution. One of the main concerns with the use of national soil databases for criminal investigations is that they were usually created for agricultural or environmental reasons, and are therefore often lacking in surface soil characteristics and the characteristics mapped (e.g., soil carbon, pH, texture) are often not those used in forensic soil science. For that reason, many forensic soil scientists

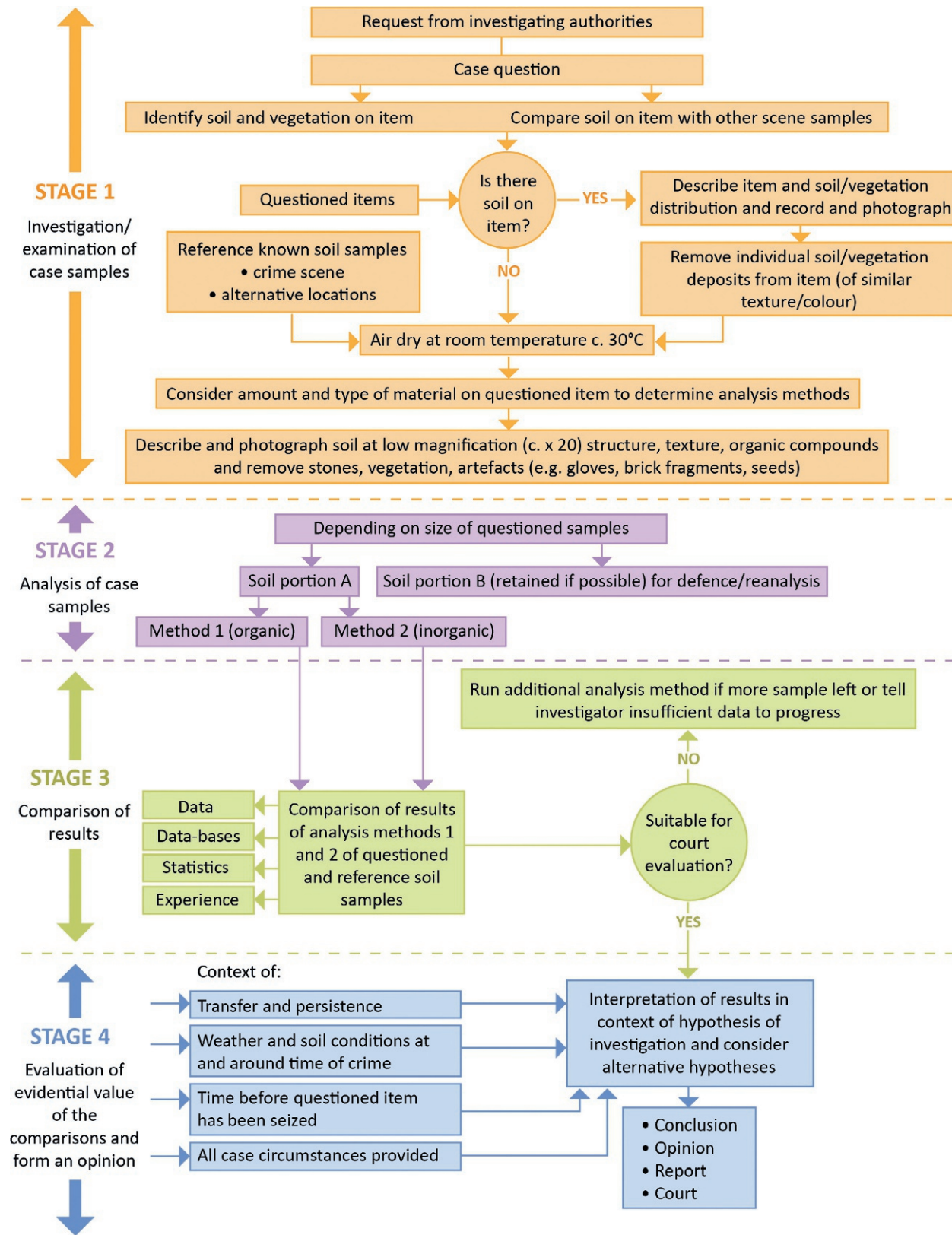


Fig. 5 A proposed sequence for soil examination, analysis, comparison, and evaluation to assist with Criminal Justice Investigations. Also see ENFSI APTST (2019) *Best Practice Manual for the Forensic Comparison of Soil Traces*. Best Practice Manual for the Forensic Comparison of Soil Traces (enfsi.eu); Donnelly LJ, Pirrie D, Harrison, Ruffell A, Dawson LA (eds.) (2021) *A Guide to Forensic Geology*. Geological Society: London, Special Publication.

have created their own databases containing representative samples across rural, urban, and peri-urban areas, including surface samples, with the methods that they most commonly adopt for forensic case work.

Soil classification

Soil classification provides a structured conceptual framework for describing and understanding soil properties and soil formation, and helps to organize soil knowledge, often through national soil survey. The two most widely used international soil classification systems are the World Reference Base (WRB) for Soil Resources (IUSS Working Group WRB, 2015) and Soil Taxonomy (Soil Survey Staff, 2014). The WRB is an international system for soil classification and developed under the direction of the Food and Agriculture Organization of the United Nations (FAO). However, many countries also have their own national classifications. Soil maps show the spatial distribution of soils across the land and were originally produced by field surveyors who walked over the landscape, looking at the soils and other features such as vegetation and topography, drawing boundaries between the different soil types. Mapping methods have evolved over time, and now techniques such as aerial photography and satellite-based global positioning systems (GPS) are used and are linked with analytical databases. Soil maps are produced at a range of scales from those covering large areas (countries and regions; 1:100000 or smaller scale) to detailed areas such as farms or parklands (1:10000 or larger scale). Each soil has its own characteristics which can be described by the soil scientist (e.g., morphology, mineralogy, and organic matter composition). These classifications allow us to narrow down areas in search operations, such as through the comparison of analytical attributes of a questioned soil with the attributes of samples held within the spatially referenced databases.

Human-made soils

Human-made soils are those that contain buried objects, or artifacts, derived from human habitation and management of urban environments. These can be particularly important in forensic case work because many crimes are committed in urban and peri-urban areas. Other terms used in soil classification systems include 'human-altered' and 'human-transported' soils (HAHT) in Soil Taxonomy, 'Technosols' (WRB), while Soil Taxonomy and the Soil Classification for England and Wales includes a 'man-made' soil category. The soil survey for Scotland does not provide for human-made soils in its taxonomic system. The Anthroisol soil group relates to the effects of long-term cultivation on soil formation, whereas the Technosol soil group is characterized by the presence of materials, which have been manufactured or relocated by humans, therefore, Technosols would more obviously represent most urban soils. Technosols are also present in non-urban environments. Anthropogenic soils are characterized by strong spatial heterogeneity, as a result of the many different potential inputs of manufactured materials (e.g., brick fragments, fecal material, fly ash, plant debris, compost, plastics, or toxic wastes) and mixing of the original (natural) soil material (e.g., gardens, parks, or cemeteries) with introduced materials such as compost and sand and can be particularly discriminating in forensic investigations (Dawson et al., 2019). Soils from mines or quarries are another class of anthropogenic soils, which are strongly influenced but are not necessarily found in or near cities. Anthropogenic soils are diverse, heterogeneous, and complex, which enables forensic soil examiners to distinguish between soils, which may otherwise appear similar. They can also often reflect a history of multiple influences, which has been very useful in understanding and characterizing soil differences in soil forensic comparisons.

Chain-of-custody

Of paramount importance in both soil collection or sampling and soil characterization or evaluation is continuity of evidence or chain-of-custody, which must be maintained and always documented throughout an entire soil forensic investigation. Preservation of soil material must be maintained to retain the integrity of the soil materials. Consequently, secure sealing of dry soil samples in sample bags or bottles with tamper-indicating serrated evidence tape and sample storage is vital. Secure sample storage and rigorous procedures ensure that information about each sample is unambiguously retrievable and can be presented in court with certainty (Fitzpatrick and Donnelly, 2021).

Initial description and characterization of samples

Initial screening (i.e., morphological comparison examination, qualitative color comparisons) of whole soil samples is to provide a visual comparison of samples (i.e., hand-held samples or specimens, soil profiles, and samples or specimens under a stereo-binocular light microscope) (ENFSI APTST, 2019). To do this, the soil must first be described systematically and characterized using standard soil morphological descriptors and tests such as color, consistency, structure, texture, segregations or coarse fragments (charcoal, ironstone, or carbonates), and abundance of roots or pores; to deduce whether a soil sample can potentially be used as evidence. These soil morphological descriptions follow strict conventions whereby a standard array of data is described in

a sequence, and each term is defined according to both International, (e.g., USDA Field book) and National standard systems for describing and sampling soils.

Soil inorganic analysis

Characterization of the soil inorganic fraction is usually carried out following sample selection and size fractionation through a standardized mesh-sized sieve (usually 1 or 2 mm). In addition, other selected methods, such as X-Ray Diffraction (XRD), Diffuse Reflectance Infrared Fourier Transform (DRIFT) or ICP and XRD are often used for the identification, characterization, and quantitative analyzes of minerals in soil (if enough sample is recovered). The advantage of XRD methods in forensic soil science is based on the unique character of the diffraction patterns of crystalline and even poorly crystalline soil minerals that they provide. Elements and their oxides, polymorphic forms (e.g., anatase and rutile polymorphic forms of TiO_2 , titanium dioxide) and mixed crystals can be distinguished by this semi-destructive examination. Part of the comparison involves identification of as many of the crystalline components by reference to a database of XRD data or to a local collection of standard reference diffraction patterns, coupled with expert interpretation. The XRD patterns can be compared for different soil samples to delineate how closely the samples relate to each other. Infrared (IR) spectroscopy is nondestructive and can be rapidly applied with the mid-infrared part of the electromagnetic spectrum, which is sensitive to organic materials, clay minerals, and quartz because of the absorption of infrared light at the vibrational frequencies of molecular functional groups constituting these materials (Dawson and Hillier, 2010) although they are not quantitative. Magnetic susceptibility methods, which have become a powerful tool to detect magnetic materials in soils (e.g., maghemite, magnetite) that are present at amounts below the detection limits of both XRD and IR, can also be applied (Murray, 2011).

Soil organic characterization

Characterization and the quantification of the organic fraction in soil samples can be carried out using several approaches. The quantification of soil wax markers, although destructive, can be carried out on samples <45 mg in size, and enables the analysis of individual aggregates or peds, which allows a better comparison between traces of soil. The *n*-alkanes and alcohols in soil reflects both current and past land use at a particular site. These, in general, provide information at a more local scale than inorganic methods on petroleum hydrocarbons and fecal material within soil samples (Dawson, 2017; McCulloch et al., 2018).

The soil pollen profile can also be used to characterize the organic component of soils, but it requires transparency of associated databases of pollen profiles to enable any evaluation to be carried out (Ezegbogu, 2021; Uitdehaag, 2021) which is not always available in all countries. In addition, more contemporary approaches such as Raman spectroscopy, stable isotope analysis and DNA metabarcoding are being developed. But there are major challenges with these methods that include a lack of optimization to forensic expectations and lack of robust data interpretation (Ezegbogu, 2021). Several research papers have been reported to show that the DNA profile of a soil bacterial community could be obtained from samples of soil recovered from potential crime scenes (e.g., shoes or clothing), and could be compared with profiles that are representative of the site of collection. There are issues, however, with DNA extraction for the comparison of soil type, and the very small scales of temporal and spatial variation (Demaneche et al., 2017). At present, the analytical procedures are not sufficiently robust for results to be acceptable in courts of law.

Soil particle characterization

In many forensic investigations of soil, the amount of soil available for analyzes may be extremely small (e.g., thin coatings or single particles weighing around 0.5–5 mg). The use of petrography is a major and precise method of studying and screening soils for discrimination in forensic science and is used in several countries across the world. Around 50 common minerals can be observed by the naked eye. A hand lens or low power stereo-binocular microscope enables the forensic soil scientist to detect the mineral properties better (particle shape and surface texture) and provide more accurate mineral identification. The petrographic microscope is also used for characterizing microfossils such as pollen grains, grass spores, opal phytoliths and diatoms. However, these detailed identifications may represent only a small part of the source sample and require specialized training, are subjective, and will be unlikely be part of national forensic laboratories.

Evaluation of soil information

Landform and soil mapping can be used by forensic soil scientists to integrate and extrapolate soil information from one scale to the next to create a coherent model of soil information from microscopic observations to the landscape scale (e.g., using soil, geological maps, terrain analysis, remote sensing, and geophysics). This combined information is used for geographic sourcing to identify the origin of a sample by placing constraints on the environment from which the sample originated.

There are many ways to evaluate soil evidence as to its evidential strength, and each country has its own approach, ranging from simple 'cannot be excluded as sharing a common origin' to Bayesian likelihood ratio determinations (ENFSI APTST, 2019). The types of methods applied and types of soil, along with available relevant validated databases, and the type of legal system will influence the system adopted (Dawson et al., 2021).

Communication

Investigators are often trying to establish if there are linkages or associations between victims, suspects, crime scenes, and individual items, which is the case with any form of physical evidence. Raymond Murray's book 'Evidence from the Earth' provides many high-profile case examples in which earth materials and related methods have contributed significantly to solving cases from around the world. Soil evidence has already proven useful in helping to solve criminal, civil, and environmental cases. Innovative ideas, methods, or applications of field and laboratory approaches in soil science have been critical in developing coherent, predictive results and the building of soil models from microscopic to landscape scales, to solve difficult soil-based investigations at a range of scales, often involving complex issues. In particular, in countries where the legal systems involve juries as the triers of fact (e.g., the United Kingdom, Australia, the United States of America), the expert witness should be able to explain their findings in an easily understood manner so that any jury member can understand what the evidence can tell us, and of equal importance, what the evidence is not able to tell us (Dawson et al., 2021; Dawson and Gunnicliffe, 2017). Soil is a complex entity, but the duty of the expert witness (including that of the forensic soil scientist) is to the court and along with that duty is the need to deliver the results and opinion in a clear and straightforward way so that it is understood.

Concluding remarks

Forensic soil science is the application of soil science within the context of the law. It is currently used in many countries, including the United Kingdom, Australia, the Netherlands, Germany, Spain, Italy, the United States of America, and Russia, and is developing rapidly across other regions of the world. Soil materials can be used in forensic investigations to evaluate potential associations to support law enforcement agencies in investigations of criminal, environmental, wildlife and serious and organized crime. Soil recovered from a questioned item with an unknown history can be compared with known locations such as a crime scene. Forensic soil science is used to assist in search operations where soil can help to narrow down areas on which to focus the ground search activities (provenancing). It can also be used for trace evidence comparison, where a questioned soil recovered from an item such as a vehicle, footwear or clothing can be compared with known samples such as from a crime scene. Soil forensic investigation involves the collection of soil samples from a crime scene location and the recovery of soil from questioned items, followed by a process of soil description, soil characterization and comparative analyzes. Soil samples used in forensic case work require careful collection, consistent with standard forensic best practice, followed by analysis and interpretation by a suitably qualified and experienced forensic soil scientist. Soil characterization should involve a staged strategic approach, which normally involves visual examination of the soil's physical characteristics, (i.e., color, texture and morphology), followed by chemical, physical and/or biological analysis, using the most appropriate technique for the soil type and soil conditions under consideration. Individual components within the soil can be identified and compared, such as leaf fragments, hydrocarbons, fecal substrates, minerals, pollen grains or recovered artifacts. Integration and extrapolation of the soil information can be scaled up to create a model of soil information, which often involves soil and land cover maps and geographical information systems (GIS). Progression of a soil forensic examination through each of the stages depends on the amount of sample available, the conditions of the sample, the case question, and the outcome of the initial stages of the examination. Established concepts, standardized approaches and professionalism have been developed in soil science and should be carefully applied regarding the legal context to allow the appropriate and safe application of soil science within the criminal justice systems worldwide.

References

- Aitkenhead M, Coull MC, and Dawson LA (2014) Predicting sample source location from soil analysis using neural networks. *Environmental Forensics* 15: 281–292.
- Dawson LA (2017) Soil organic characterisation in forensic case work. *Episodes Journal* 40(2): 157–165.
- Dawson LA and Gunnicliffe C (2017) *Managing the myths - the CSI effect*. November. <https://www.microbiologysociety.org/publication/current-issue/microbiology-in-popular-culture/article/managing-the-myths-the-csi-effect-in-forensic-science.html>.
- Dawson LA and Hillier S (2010) Measurement of soil characteristics for forensic applications. *Surface and Interface Analysis* 42: 363–377.
- Dawson LA and Mayes RW (2014) Criminal and environmental soil forensics: Soil as physical evidence in forensic investigations. In: Murphy BL and Morrison RD (eds.) *Introduction to Environmental Forensics*, 3rd edn, pp. 457–486. Oxford: Academic Press. Ch. 12.
- Dawson LA and Morgan R (2015) Forensic Geoscience. In: Walport M (ed.) *Annual Report. Forensic Science and Beyond*, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/506462/gs-15-37b-forensic-science-beyond-evidence.pdf.
- Dawson LA, Macdonald LM, and Ritz K (2019) *Plant Wax Compounds and Soil Microbial DNA Profiles to Ascertain Urban Land Use Type*. London: Geological Society. Special Publications, No. 492, 14 November 2019.
- Dawson LA, Auchie D, and Parratt D (2021) The judicial system, reporting and giving evidence in court. In: Donnelly L, Pirrie D, Harrison M, Ruffell A, and Dawson L (eds.) *A Guide to Forensic Geology*. London: Geological Society.

- Demaneche S, Schauser L, Dawson LA, Franqueville L, and Simonet P (2017) Microbial soil community analyses for forensic science: Application to a blind test. *Forensic Science International* 270: 153–158.
- Di Maggio RM, Donnelly L, Al Naimi K, Barone PM, Da Silva Salvador FA, Dawson L, Dixon R, Fitzpatrick F, Gradusova O, Nesterina E, Peleneva M, Ushacova O, Molina Gallego CM, Pirrie D, Ruffell A, McKinley J, Sagripanti G, Villalba D, Schneck B, Sugita R, Wach G, Silva R, and Forbes S (2017) Global developments in forensic geology. *Episodes* 40(2): 120–131. <https://doi.org/10.18814/epiiugs/2017/v40i2/017014>.
- Donnelly LJ, Pirrie D, Harrison, Ruffell A, and Dawson LA (eds.) (2021) *A Guide to Forensic Geology*. London: Geological Society. Special Publication 2021.
- ENFSI APTST (2019) *Best Practice Manual for the Forensic Comparison of Soil Traces*. Best Practice Manual for the Forensic Comparison of Soil Traces (enfsi.eu)
- Ezegbogu MO (2021) Identifying the scene of a crime through pollen analysis. *Science and Justice* 61(3): 205–213.
- Fitzpatrick RW (2013) Soil: Forensic analysis. In: Jamieson A and Moenssens A (eds.) *Wiley Encyclopedia of Forensic Science*. John Wiley and Sons, Ltd. The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ. <https://doi.org/10.1002/9780470061589.fsa096.pub2> (August 2013 updated articles) <http://onlinelibrary.wiley.com/book/10.1002/9780470061589>.
- Fitzpatrick RW and Donnelly LJ (2021) Introduction to forensic soil science and forensic geology: A synthesis. In: Fitzpatrick RW and Donnelly LJ (eds.) *Forensic Soil Science and Geology*. vol. 492, pp. 1–32. Geological Society of London. <https://doi.org/10.1144/SP492-2021-81>. Special Publications.
- Harrison K, Dawson LA, and Mackinnon G (2016) Interdisciplinary approaches to the search and location of buried bodies: A United Kingdom context. In: Kars H (ed.) *Soil Forensics*. Springer.
- IUSS Working Group WRB (2015) *World Reference Base for Soil Resources*. World Soil Resources Report 106 Rome: FAO. <http://www.fao.org/3/a-i3794e.pdf>.
- Locard E (1930) Analyses of dust traces parts I, II and III. *American Journal of Police Science* 1: 276–298. 401–418, 496–514.
- McCulloch G, Dawson LA, Ross MJ, and Morgan R (2018) The discrimination of geoforensic trace material from close proximity locations by organic profiling using HPLC and plant wax marker analysis by GC. *Forensic Science International*. <https://doi.org/10.1016/j.forsciint.2018.02.009>.
- Murray RC (2011) *Evidence From the Earth: Forensic Geology and Criminal Investigation*, 2nd edn. Missoula, MT: Mountain Press Publishing.
- Murray R.C., Tedrow J.C.F. (1992) *Forensic Geology*. Prentice Hall; 1992.
- Pirrie D, Ruffell A, and Dawson LA (2021a) Geological evidence recovery from exhibits. In: Donnelly L, Pirrie D, Harrison M, Ruffell A, and Dawson L (eds.) *A Guide to Forensic Geology*. London: Geological Society.
- Pirrie D, Ruffell A, Dawson LA, and McKinley J (2021b) Crime scenes: Geoforensic assessment sampling and examination. In: Donnelly L, Pirrie D, Harrison M, Ruffell A, and Dawson L (eds.) *A Guide to Forensic Geology*. London: Geological Society.
- Pye K (2007) *Geological and Soil Evidence: Forensic Applications*. Boca Raton, FL: CRC Press.
- Ritz K, Dawson L, and Miller D (eds.) (2008) *Criminal and Environmental Soil Forensics, Soil Forensics International, Edinburgh Conference Centre: Springer*.
- Ruffell A and McKinley J (2008) *Geoforensics*. Chichester: John Wiley & Sons, Ltd.
- Ruffell A, Pirrie D, and Dawson LA (2021) Geological evidence analysis. In: Donnelly L, Pirrie D, Harrison M, Ruffell A, and Dawson L (eds.) *A Guide to Forensic Geology*. London: Geological Society.
- Science and Art (1856) Curious use of the microscope. *Scientific American* 11: 240.
- Soil Survey Staff (2014) *Keys to Soil Taxonomy*, 12th edn Washington, DC: United States Department of Agriculture Natural Resources Conservation Service.
- Testoni SA, Melo VF, Dawson LA, Salvador FAS, and Kunii PA (2019) Validation of a Standard Operating Procedure (SOP) for forensic soils investigation in Brazil. *Revista Brasileira de Ciência do Solo* 43: e0190010. <https://doi.org/10.1590/18069657rbcs20190010>.
- Uitendhaag SCA (2021) Forensic soil comparison. *Utrecht Studies in Earth Sciences*. vol. 227, p. 115. Utrecht University.

Soil governance at the international, regional and national level

Harald Ginzky, German Environment Agency, Dessau-Roßlau, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	533
The background of soil governance	533
Definition of soil governance	534
Value of soils for society	534
Global status of land and soil degradation	534
Challenges of soil governance	534
Soil threats and drivers	535
Summary	535
Soil governance at the international level	535
Sustainable development goals – the objective of a “land degradation neutral world”	536
Treaty provisions	537
Global soil partnership – Food and Agriculture Organization	538
Right to food	538
Sustainable soil management as a common concern of humankind	538
Recommendation and outlook	538
National and regional level	539
Core elements of good soil governance at the national level	539
Soil as natural resource which provides several crucial social and ecological services	539
Regulatory concept: Framework law and sectoral provisions	540
Complementary provision in other pieces of legislation	540
Relevance of soil data and information	540
Specified requirements for all soil uses in a legally binding form	541
Procedural rights and courts	541
Tenure rights	541
Importance of administrative arrangements – the case of implementation and enforcement	541
Soils, culture and ethics	542
Excursus one: Soil governance in Germany	542
Excursus two: Methodological observations	543
Transboundary trade and sustainable soil management	544
Outlook	544
References	544

Abstract

The chapter first analyses what is meant by the term soil governance and what the particular challenges are. It addresses the national and regional levels of soil governance, together with an explanation of the existing regulatory structure at the international level. As sustainable management of soils is inextricably linked to and carried out based on the conditions specific to a location, national regulation is most essential. To provide more specific insights, this chapter also explores – in an excursus – soil governance in Germany and an assessment of the tools and sources of information, including databases. The interlinkage of international trade law and sustainable soil management is also explained.

Key points

- Overview of soil governance at national, regional and international level
- Challenges of soil governance
- Insights to the regulatory concept in Germany
- Importance of effective administrative arrangements
- Cultural and ethical perceptions on soils
- Methodological considerations

Introduction

The perception of soils has changed significantly over the last 10 years. Previously, soils were often regarded as the poor, in the sense of being neglected, cousin of water, air and biodiversity (Ginzky, 2020). Soils were seen as ‘dirt’; the many social and ecological functions they provided were taken for granted and therefore, in a way, ‘overlooked and not valued.’

Even worse from a conservation perspective, soil was often only perceived as land, which could be possessed and owned by private persons or institutions, or by municipal or state entities. These owners and or users were often regarded as having an exclusive and unlimited right to use and exploit the land and the soil beneath it (Ruppel and Ginzky, 2021; Ginzky et al., 2019). Thus, a responsibility of the land owner or user to maintain the soil quality was not assumed.

There were several processes which led to a new perception of soils. The food crisis at the beginning of this century, and all its consequences, triggered this new development as it became clear that fertile soils are a precondition for food security (Ginzky, 2020). States and private investors also started to try to secure access to land in other countries for future needs, referred to as ‘and grabbing,’ thereby often jeopardizing the interests and necessities of the local communities (Windfuhr, 2016; Tamasang, 2022). Moreover, scientific evidence has shown that fertile soils are the second largest biological carbon sink on the planet; thus, they are indispensable to climate mitigation and adaptation processes (IPCC, 2018, 2020). In this context it also became clear that fertile soils are needed for the production of plants for renewable energy and potentially for the uptake and sequestration of already emitted greenhouse gases (IPCC, 2018). In addition, with regard to the dramatic biodiversity crisis faced globally, studies have emphasized that soils host an almost infinite amount of biodiversity (IPBES, 2018). Given all of these factors, an overwhelming consensus developed that the ongoing land and soil degradation needs to be stopped and that the maintenance and enhancement of soils are a precondition to fighting these various crises and implementing truly sustainable development.

Starting around 2010, several scientific and or political fora started to discuss and promote the topic of soils and their social and ecological services. One important move was the Berlin Call for Action of 2011. Several relevant international bodies and entities assembled in Berlin, such as the United Nations Convention to Combat Desertification (UNCCD), the Food and Agriculture Organisation (FAO) and others, and concluded with a call for more soil awareness, joint efforts to develop a better scientific understanding on soils primarily through the selection of data and for more soil governance.¹ In the same year, the UNCCD took up the issue, working on the concept of a “Zero land degradation world” (Lal et al., 2012; Hannam, 2022). In 2012, world leaders at the Rio + 20 Conference jointly recognized the “economic and social significance of good land management” and stressed that “desertification, land degradation and drought are challenges of global dimension” as well as “the need for urgent action” (United Nations, 2012). September 2015 was then another milestone. The UN General Assembly in New York adopted the 2030 Sustainability Agenda (United Nations, 2015) with the 17 Sustainable Development Goals (SDGs), including 169 targets. In target 15.3 the world leaders agreed to strive to achieve a “land degradation neutral world” by 2030. The UNCCD subsequently established itself as the international lead organization for the implementation of this objective at its Conference of the Parties at the end of 2015 (Boer et al., 2016).

The importance of ‘dealing well with soils’ was furthermore promoted and brought to the fore by a sequence of so-called “Global Soil Weeks” – large workshops with international attendance that took place annually in Berlin and Nairobi between 2012 and 2019.² In addition, several reports by the Intergovernmental Panel on Climate Change (IPCC) and the Intergovernmental Science-Policy Panel on Biodiversity and Ecosystem Services (IPBES), in particular after 2018, underlined the interface between sustainable soil management, and both climate mitigation and adaptation as well as the biodiversity crisis (IPCC, 2018 and IPBES, 2018).

Awareness is a good thing. Nevertheless, sustainable soil management requires considerable knowledge, political will and good governance. As pioneers on the topic of soil governance, Ben Boer and Ian Hannam have to be mentioned. In the early years of this century, they dedicated a great deal of work to analyzing soil law at the international as well as the national level (inter alia Boer and Hannam, 2003 and Hannam and Boer, 2002). The *International Yearbook of Soil Law and Policy* has now published five volumes, each more than 400 pages, and SoILEX, a database of existing legal instruments on soil protection throughout the world provided by the FAO, are important mechanisms to promote discussion as to appropriate regulatory instruments and concepts, and to create a platform for exchange and debate.

This chapter is about soil governance. The intention is to provide an instructive overview on its most important aspects, primarily for non-lawyers. The chapter includes what is meant by the term soil governance and the particular challenges, the existing regulatory structure at the international level, followed by that at the national and regional levels, in an excursus, soil governance in Germany, the interlinkage of international trade law and sustainable soil management, and finally some concluding remarks on the interface of soil governance and transformative processes towards sustainability and climate neutrality.

¹<https://www.umweltbundesamt.de/en/soil-protection-at-international-level#background>.

²<https://globalsoilweek.org/>.

The background of soil governance

Before discussing regulation in itself, it is important to understand first what 'soil governance' means. Second, it is important to emphasize the potential value of soil for society. In this context, the social and ecological services that soils provide must be highlighted. Third, the particularities which render soil governance challenging should be mentioned. Finally, the potential drivers and soil threats need to be explained.

Definition of soil governance

Governance is a term which needs clarification. The FAO defines soil governance as follows: "Soil governance concerns policies and strategies and the processes of decision-making by nation states and local governments on how the soil is utilized."³ In this chapter soil governance is taken to include legal provisions, and organizational and institutional arrangements that are intended to regulate the use of soils. A full picture of soil governance would also need to entail the financial resources of relevant entities, the interplay between various actors, such as policy makers, scientific institutions and civil society, as well as cultural and religious conventions and perceptions. Soil governance can be found at all levels, municipal as well as national, regional (transboundary) (for example, the European Union and the Alpine region) and at international level.

Value of soils for society

Until recently, soils have not been adequately valued. This undervaluing of soils has been the dominant perception for a long time. There are several reasons. Soils have been primarily perceived as the dirt under our shoes. Others may have seen soils as a given and therefore 'eternal' or 'inexhaustible.' Others may have taken the view that nature is there to serve the purposes of human beings.

A more realistic view tells us that soils are indispensable for the future of human beings, which will only be possible through sustainable development as they provide the following services (Ginzky, 2020):

- Basis for food production and production of plants for renewable energy
- Carbon sequestration, therefore important to climate change mitigation
- Biodiversity host
- Vital role in world's biological cycle by inter alia storing nutrients and filtering groundwater from hazardous substances
- Climate adaptation, particularly in urban areas
- Water storage in case of draught and extreme rainfall
- Basis for nature and recreational areas for human beings
- Cultural and biological archive

Without fertile or healthy soils, inter alia, food security, climate mitigation and adaptation as well as maintenance of biodiversity could hardly be achieved. Inadequate sustainable soil management may thus lead to poverty, migration within or beyond national boundaries and even political or military conflicts about natural resources or due to societal instabilities (Ginzky, 2020).

Furthermore, and it is not to be underestimated, soils as part of nature are culturally rich and historically significant. Culture relates to the traditions, emotions, societal behaviors and even art influenced by, deeply ingrained and tied to soils. They also serve as an archive of human history, containing biological remnants from past (mis)uses and societies. Loss of soil could then also mean a loss of culture and or history (Heuser, 2017).

Soils are generally immobile and local, however, erosion by wind or water may move them to another location. Despite being generally immobile, soils provide the above mentioned social and ecological services which may have a direct or indirect transboundary effect. Hence, the potential negative consequence of soil degradation, such as food insecurity, political conflicts, climate change, are in essence transboundary (Flasbarth, 2016).

Global status of land and soil degradation

Looking at the current state of soils worldwide, figures of the steady or even accelerating rate of land degradation are alarming. About one third of the world's soils are already affected (Ginzky, 2016). An area the size of Italy has been and will be, without intervention, continuously degraded each year (Linz and Lobos, 2016).

Challenges of soil governance

Soil governance faces particular challenges. First, most human activities affect soils, such as agriculture, residential development (urban or rural), transport and manufacturing. (Ginzky (2016), p. 3).⁴ Almost any use of soil will cause degradation to some extent.

³<https://www.fao.org/global-soil-partnership/areas-of-work/soil-governance/en/>.

⁴One exemption is if people act in aquatic settings. For example, fishing or shipping do not necessarily have a direct effect on soils.

Soil governance is therefore a cross-cutting issue as most human activities need to be considered with regard to sustainable soil management.

A second aspect which poses specific challenges to soil governance is that soils are mostly owned by individual persons, private companies, local communities or state entities. For a long time, the overall perception has been that the plot of land which entails the soil belongs to the owner, who has an exclusive and somewhat untouchable right to exploit this 'resource.'

As a third challenge, insecure tenure rights are an additional challenge in many regions of the world (Ginzky and Ruppel, 2021; Munyuki-Hungwe, 2020). Insecure tenure may cause many hindrances to sustainable soil management as user rights, responsibilities and accountability remain unclear (Windfuhr, 2016).

Fourth, one must not forget that soil as a resource is often of high economic value as it can be exploited in many respects, such as, inter alia, for building houses, infrastructure or industrial facilities, growing food (or nowadays also fuel crops) or for grazing as well as receiving carbon credits for carbon uptake (Bodde et al., 2020). It immediately becomes clear that governance for sustainable soil management may conflict with the economic interests of individual users or owners of soils as they do not want to see their individual exploitation rights limited or restricted.

And last but not least, a more scientific issue: There are many different types of soil which differ significantly with regard to their components, structure and other aspects. Thus, each soil type carries its own characteristics, specific resilience and vulnerabilities with regard to different uses. These different natural realities need to be considered when drafting soil governance instruments as 'one-size-fits-all' is neither appropriate nor effective.

Soil threats and drivers

Most human activities have effects on soils. There are, however, a few main factors that control the degradation of soils. First, unsustainable agricultural and forestry practices are a major factor in most parts of the world. These unsustainable practices could have completely different motives or reasons: lack of knowledge regarding extent or severity of impacts or consequences, necessity due to poverty, inadequate technology as well as economic benefit or greed.

Second, urbanization typically causes negative effects on soils, posing threats such as sealing (i.e. paving over healthy soils with concrete foundations) and contamination. The same applies to land taken for infrastructure, such as streets or railways, and industrial uses. Mining may also provoke degrading processes on soils. Finally, effects from climate change or the biodiversity crisis, such as unpredictable seasonal weather, extreme weather events or new pests, could have detrimental effects on soils (Ginzky, 2020). Furthermore, poverty and hunger on the one hand and weak governance on the other could be drivers of soil degradation (Orubebe, 2020; Ruppel and Ginzky, 2021).

Different threats to soil arising from these various uses may be identified. The ability to categorize the different risks posed to soils may help to focus the debate better on potentially effective regulatory concepts and instruments. The EU Thematic Strategy for Soils of 2006 differentiated eight types of soil threats, including: soil erosion, decline in organic matter, local and diffuse contamination, sealing, compaction, and decline in biodiversity, salinization, floods and landslides.⁵ Water scarcity or drought, eutrophication as well as changing biodiversity were not included in that policy document, but they should be seen as additional soil threats needing to be dealt with through soil governance.

Summary

Soil governance is a complex task for several reasons: it is cross-cutting, it has been undervalued, it may potentially conflict with economic interests, it must cover diverse types of soils, etc. At the same time, soils provide numerous social and ecological services which are of great value for society. Soils are indispensable, and this renders soil governance a major challenge.

There is no blue print which fits all states. The challenges faced by each state with regard to their different main soil threats and drivers, as well as their necessities and priorities should be the starting point for their individual sustainable soil governance processes. Thus, these aspects are also decisive to determine which specific actions are appropriate and needed for the national contexts.

Soil governance at the international level

International environmental law is very different from national law as there is no single legislator, but states who decide on the topics, the level and the structures of international cooperation. Thus, international environmental law is mainly entailed in international treaties, such as the United Nations Framework Convention on Climate Change (UNFCCC).⁶ Provisions in international treaties are only obligatory for states that have signed and ratified the respective treaties. For citizens and legal entities, provisions of international treaties only become mandatory after they have been transposed into national legislation (Bodde et al., 2020).

⁵EU Commission (2012).

⁶The Text of the UNFCCC may be found at: www.unfccc.int

The preparation and negotiation of international treaties is very time- and resource-consuming as all states follow their own political priorities in these discussions. All states must accept the various provisions in such treaties, therefore, compromises have to be negotiated and reached to gain consensus.

Most international environmental treaties have established their own organizational entities: *inter alia*, a Conference of the Parties, which meets annually or biannually and is normally the political body of the treaty regime; a Secretariat, which is in charge of facilitating the negotiations amongst the State Parties and for the Convention's continuous management; and a scientific advisory body (Bodle et al., 2020).

In addition to treaty provisions, international customary law may set mandatory obligations. International customary law is usually not codified (not written) and may come into force if states have accepted and applied the respective obligation over a long period of time. For example, international customary law obliges a state in whose territory an activity is intended to take place which may have negative effects on a neighboring country to arrange for a consultation process with the potentially affected country.

Finally, there is so-called 'soft law,' which is not legally binding but may be politically highly relevant. Decisions of the organs of an international organization, such as the UN General Assembly, are considered to be soft law.

States only enter into international cooperation if there is a good reason and if international cooperation has an added value. As the ecological functions of soils, such as ensuring food supply, fighting climate change and the biodiversity crisis, are trans-boundary in nature, international cooperation is certainly reasonable and required to deal with these enormous challenges.

How international cooperation on sustainable soil management is organized will be elucidated in this subsection.

Sustainable development goals – the objective of a “land degradation neutral world”

In 2015, the UN General Assembly adopted the 2030 Agenda for Sustainable Development, which entails commitments for both developing and developed countries (United Nations, 2015). Several of the 17 SDGs and the 169 accompanying targets have implications for soil protection. All SDGs need to be considered in their complexity and jointly to achieve the ambitions of the 2030 Sustainable Development Agenda. Siloed thinking, meaning one SDG for itself, would not be sufficient, also with regard to sustainable soil management (Ehlers, 2016).

The SDG 15 and Target 15.3 are, however, the most pertinent ones as SDG 15 requests “to halt and reverse land degradation” and target 15.3 demands that states “... by 2030, combat desertification, and restore degraded land and soil, including land affected by desertification, drought and floods, and strive to achieve a land-degradation neutral world” (Boer and Hannam, 2017; Altvater et al., 2014).

It is important to mention that the objective of a land degradation neutral world (or 'land degradation neutrality' (LDN)) has been increasingly perceived as a kind of a leitmotiv of international and national soil governance (Boer et al., 2016; Bodle et al., 2020; Fowler and Hannam, 2019). Although the 2030 Agenda, including the SDGs, is not legally binding, it is politically important as it was endorsed by all states.

During COP 12 of the UNCCD, a definition of LDN was unanimously adopted:

Land degradation neutrality is a state whereby the amount and quality of land resources necessary to support ecosystem functions and services and enhance food security remain stable or increase within specified temporal and spatial scales and ecosystems.

(UNCCD, 2015)

Soil is an essential element of land, therefore the LDN objective is directly applicable to soils (Ginzky, 2017). The definition furthermore underlines that the “amount and quality” has to “remain stable or increase.” Thereby, it is stressed that the neutrality objective has to be thought of in both quantitative and qualitative terms. Moreover, as land and soil degradation cannot be completely avoided, restoration or rehabilitation must be actively undertaken to achieve the objective of 'neutrality,' expressed in the phrase “remain stable or increase” (Ehlers, 2016).

The UNCCD – with its entity aimed at the 'science-policy interface' – has developed the conceptual framework for understanding LDN and recommendations on the necessary measures to achieve it. This conceptual framework is intended to support states with the implementation of the LDN objective in their respective home countries. The conceptual framework is clear, concise and instructive and, in that sense, a very useful instrument.

The UNCCD's 'LDN response hierarchy' (UNCCD, 2017)⁷ entails three necessary categories of actions: avoid, reduce and reverse (Cowie et al., 2018). Avoid is aimed at preventing degradation; in other words, that no degradation occurs. The term reduce is aimed at actions to ensure that the negative effects of an activity on soils are minimized to the lowest level possible. Reverse includes measures to restore or to rehabilitate degradation processes. In the view of UNCCD, 'restoration' seeks to “re-establish the pre-existing biotic integrity” and 'rehabilitation' is intended to “reinstatement the ecosystem functionality with the focus on provision

⁷In addition, many other matters have to be clarified, defined and arranged, i.e. the following points: A baseline has to be defined, the status of soil/land have to be assessed, indicators for the assessment of the status, of land degradation and of restoration/rehabilitation actions as well as monitoring requirements have to be defined. For further information, see Ehlers, 2016, Minelli et al., 2016 and Cowie 2018.

of goods and services" (Cowie et al., 2018). Finally, pursuant to the conceptual framework, "a pro-active focus on planning" needs to be in place by which "anticipated losses" should be counterbalanced by "planned gains" (Cowie et al., 2018).

These four categories of required measures need to be taken into account when implementing LDN nationally. They could also serve as evaluation criteria as to whether and to what extent LDN is already implemented (Ginzky, 2020).

Treaty provisions

There is currently no international treaty specifically dealing with soils (Ginzky, 2016; Boer et al., 2016; Bodle et al., 2020). Thus, no international treaty sets a general framework for sustainable soil management or acts like an anchor for this topic. In fact, the topic is addressed by various international treaties from different angles. Most belong to international conventions which were developed and finally concluded in the context of the first Rio Conference on Environment and Development (also known as the Earth Summit) in 1992, shortly after the dissolution of the former Soviet Union. That significant change in the political landscape followed by a relaxation of international tensions worldwide provided a window of opportunity for international cooperation.

The UN Convention to Combat Desertification (UNCCD),⁸ the Convention on Biological Diversity (CBD)⁹ and the UN Framework Convention on Climate Change (UNFCCC) were the three conventions resulting from the Rio Conference, all of which contain provisions relevant to the sustainable management of soils. The three treaties apply almost universally, meaning to almost all states as signatories, and entail legally binding provisions that have to be complied with by the State Parties.¹⁰ However, each of the three treaties has a specific substantive focus and is thus per se limited with regard to sustainable soil management. Moreover, it could be said that no treaty contains soil specific requirements. Soils are therefore dealt with more as a side-effect in these three treaties. In the following, these three regimes are analyzed concerning their provisions on sustainable soil management.

The UNCCD is the most relevant regime for the management of soils as it expressively refers to 'land.' However, the UNCCD, formally taken, is only applicable to so-called 'dry-lands.' Dry-lands, pursuant to the legal definition of the UNCCD, occupy about 40% of the Earth's terrestrial surface (Boer and Hannam, 2003). Thus, the UNCCD is formally only applicable to less than half of the land globally. Moreover, the UNCCD does not contain any soil-specific actions. Countries affected by land degradation are obliged to implement National Action Programs (NAP) and report back to the UNCCD on their individual NAP (Hannam and Boer, 2002). Thus, a country's NAP could entail measures specifically addressing sustainable soil management.

As explored above, in 2015 the UNCCD established itself as the leading regime for the implementation of a 'land degradation neutral world' by various decisions taken by the Conference of the Parties (Boer et al., 2016). Furthermore, the UNCCD has launched the voluntary target setting program for Land Degradation Neutrality (LDN) (Mastrojeni, 2018). More than one hundred states have joined.¹¹

The scope of application of the CBD could, in general, encompass all soil functions (Ginzky, 2016; Hannam and Boer, 2002). It would allow for holistic regulation of sustainable soil management. However, the CBD does not currently foresee specific measures with regard to soils. There are only a few programs that might have positive benefits for soils, like that on agricultural biodiversity (Wolff and Kaphengst, 2016).

The prime focus of the UNFCCC is climate change issues. With regard to soils, the UNFCCC provisions on accounting, maintaining and restoring sinks and reservoirs provide a general legal framework for action, which is relevant to soils as a carbon sink. However, the UNFCCC does not establish specific requirements concerning the management of soils (Boer et al., 2016; Streck and Gay, 2016; Bodle et al., 2020). In 2015, the Paris Agreement was signed and has set new and more ambitious quantitative targets, which aim to hold the increase of global average temperature to "well below 2 degrees" and at best at 1.5 degrees Celsius (see Article 2.1 (a) Paris Agreement). The Paris Agreement was seen as a major step forward in international climate governance, both with regard to the more ambitious target and the introduction of a new regulatory concept (Boer et al., 2016). The latter required State Parties to submit "Nationally Determined Contributions" and for those to be evaluated 5 years later, i.e., 2021, thereby establishing a control mechanism over their actions towards their commitments.

The Paris Agreement, however, did not establish specific soil-related obligations. A major challenge for the implementation of the Paris Agreement has been the accounting of the Nationally Determined Contributions (Paris Agreement, Paragraph 31.3). To this end, a working group has been established, which was also tasked with determining a state's requirements for soil and land management (Fee, 2018).

The three most relevant international treaties, namely the UNCCD, the CBD and the UNFCCC/Paris Agreement, do not entail, either alone or jointly, a complete regulatory concept for sustainable soil management. Furthermore, the UNCCD, the CBD and the UNFCCC do not include soil-specific provisions. Given the cross-cutting nature and the particular challenges in governing soils – inter alia, the huge variety of soils, the interface with tenure rights, the potential for economically conflicting interests – the actual level or even indirect effects of international regulations on soil conservation at the local level is thus very limited. As these treaties and regimes act somewhat according to their own logic, there is no overarching soil regulatory regime that sets a general framework or coordinates the various actions.

⁸The Text of the UNCCD may be found at: www.unccd.int.

⁹The Text of the CBD may be found at: www.cbd.int.

¹⁰"Almost", because the United States of America are not a Party to the CBD.

¹¹<https://www.unccd.int/actions/ldn-target-setting-programme>.

Global soil partnership – Food and Agriculture Organization

The Food and Agriculture Organization (FAO), together with the Global Soil Partnership, is another important actor in the field of sustainable soil management. In December 2012, the FAO established the Global Soil Partnership (GSP) with regional branches on all continents to promote sustainable soil management and to enable international and regional cooperation. The GSP has been very instrumental, in particular, in strengthening regional cooperation, collection of soil-related information – inter alia, through the new Technical Panel on Soils,¹² the exchange of information and knowledge and finally on awareness-raising at all levels.

Jointly, the FAO and GSP promote scientific knowledge on the various soil threats and drivers. Several large international workshops have been held on relevant soil topics, such as soil pollution, erosion and biodiversity.¹³ Based on these workshops, informative reports have been published. In addition, myriad other information on sustainable soil management has been provided. Thus, the FAO has clearly extended its focus from agricultural soils to a more comprehensive soil-related perspective (Bodde et al., 2020).

Right to food

Food security is one of the essential services that soils provide to society. The eradication of hunger and poverty are of keen interest primarily in developing countries. Sustainable soil management and the enforcement of human rights are thus closely linked. For many years, international law has recognized the human right to food. It is understood as an element of the right to life according to International Covenant on Civil and Political Rights.¹⁴ Furthermore, the human right to food is expressly mentioned in the International Covenant on Economic, Social and Cultural Rights (Article 11).¹⁵ Both international treaties have been ratified by about 170 states and thus almost universally applicable (Elver, 2018; Ruppel, 2021). Access to land and natural resources is often a precondition to ensuring food security – particularly in developing countries (Poderatti, 2022; Tamasang, 2022). Thus, soil governance is not only an environmental challenge, but needs to be considered with regard to its human rights dimensions, with a particular focus on the human right to food.

Sustainable soil management as a common concern of humankind

There is an ongoing debate about whether sustainable soil management is to be recognized as a common concern of humankind (Boer, 2014). The concept of ‘common concern of humankind’ is a principle of international law, which, if applicable to something, commits states to more intensified attention to it and increased international cooperation (Cottier et al., 2014; Bowling et al., 2016; Ginzky, 2017). The concept of common concern of humankind primarily refers to global problems or challenges which “inevitably transcend the boundaries of single states and require collective action in response.”¹⁶ Several international treaties have explicitly stated that a specific challenge, for example to fight climate change (as stated in the UNFCCC), is a common concern of humankind.

Sustainable soil management deals with social and/or ecological services which are transboundary in nature and challenges requiring cooperation between states, e.g. poverty eradication, hunger prevention, climate mitigation and adaptation and reversing biodiversity decline, therefore, it could be seen as a common concern of humankind. Several international joint statements, such as the outcome document of the Rio + 20 Conference “The Future – we want”, emphasize this assumption. However, as there is no international soil treaty, a legally binding expression that this principle should apply is yet to come (Ginzky, 2017).

Recommendation and outlook

Although there is no standalone international treaty on sustainable soil management, the last few years have seen considerable momentum to regulate and increase sustainable soil management better through international cooperation. Most importantly, the global community officially recognized ‘land degradation’ as a joint challenge in the Rio + 20 outcome document “The future – we want” and through the objective of land degradation neutrality in target 15.3 of Sustainable Development Goal 15.

The UNCCD has been very supportive in this context by establishing itself as the international lead organization, unanimously adopting the definition of LDN and creating the voluntary target setting program. The FAO and GSP primarily support the promotion of sustainable soil management by providing scientific information.

It is debatable whether to negotiate and finally adopt a new treaty on sustainable soil management would be a clever strategy given the resources required and the time needed. Nevertheless, the urgency of appropriate, targeted measures must be taken into account, inter alia, due to the unacceptable realities of extreme hunger and poverty as well as the effects of the ongoing climate change and biodiversity crises affected by and/or linked to unsustainable soil management. A proposal has been submitted to apply the regulatory framework of the Paris Agreement to soils (Fowler and Hannam, 2019; Bodde et al., 2020), which was based on the agreement of an overarching objective: State Parties should be obliged to contribute to the extent that is nationally possible, but

¹²<https://www.fao.org/global-soil-partnership/itps/en/>.

¹³<https://www.fao.org/soils-portal/en/>.

¹⁴<https://www.ohchr.org/EN/ProfessionalInterest/Pages/CCPR.aspx>.

¹⁵<https://www.ohchr.org/en/professionalinterest/pages/cescr.aspx>.

¹⁶Shelton (2009), p. 83.

with reporting and re-assessment responsibilities embraced. It would certainly be a step forward if such a concept could be agreed upon quickly.

It seems promising that there is now better and intensified cooperation between the UNCCD and FAO/GSP, as well as perhaps the UNFCCC. A particular focus of these cooperative endeavors, however, should be the development of appropriate measures on how sustainable soil management could be achieved locally in the light of various drivers: agriculture, industry, land take and sealing due to housing and infrastructure, among others. Another essential aspect is the provision of sufficient financial resources and technical support to developing countries.

Location-specific soil governance is crucial to deal with the challenges. At the international level, general recommendations on appropriate concepts and instruments should be developed to inform and assist national policy makers.

National and regional level

The previous section has shown that soil governance internationally could mainly serve two purposes: first, to create, maintain and enhance cooperation amongst states with regard to this common challenge of soil protection, and second, to set incentives on states to engage nationally. These 'incentives' at the international level could be very different, including legal obligations, the provision of funding or development aid, capacity building or economic benefits.

Sustainable soil management needs to be initiated and implemented with specific regard to location as the challenges have to be addressed and resolved locally. Soils are bound to land, which is immobile and therefore contained within the boundaries of a sovereign nation. Thus, soil governance at national or regional level is crucial to work towards sustainable soil management. Soil governance at the regional level means the cooperation of states within a region, such as the Member States within the European Union (Raffelsiefen and Straßburger, 2016; Heuser, 2022) or African countries within the African Union (Kibugi, 2017, 2020), which could consist of shared agendas, programs, policies and legal provisions. The institutional settings and involvement of civil society are also of importance.

There has been considerable interest in soil governance at national level in recent years. For example, Germany has initiated a process to revise its soil governance completely in order to adapt, in particular, to the climate change and biodiversity crises and to other current needs.¹⁷ Its aim is to draft a regulatory regime which is robust enough to tackle the 'real' challenges. The EU Commission published its new Soil Strategy in November 2021, which sets the objective of achieving 'healthy soils' throughout the European Union by 2050. To this end, a draft 'soil health law' should be submitted by the European Commission in early 2023.¹⁸ Mexico has started a process on the 'gobernanza del suelo,' translated as governance of the soil (Ortiz Garcia et al., 2022). A similar debate was initiated by the Asian Research Institute for Environmental Law for the South-East Pacific Region, aimed at informing on soil governance of countries in the region.¹⁹ And the Pan-African Parliament is in the process of drafting model legislation on sustainable soil management, which, according to the procedures of the Pan-African Parliament, would then be recommended to all national parliaments in Africa for consideration and obviously necessary adaptation to the national context and refinement (Ruppel and Ginzky, 2021).

Soils are essential to so many important aspects of our lives, therefore it could be expected that endeavors to improve soil governance at national and regional levels will arise in many if not almost all places. Thus, an intense debate on good soil governance is important and must be moderated and extended widely.

It is not possible to provide information on the current soil governance in all states. There are other sources for this kind of information such as FAO SoILEX. The following subsections are intended to provide insight into to what is needed for 'good' soil governance at national and regional levels. To this end, it should be elucidated which core regulatory elements of good soil governance, in the sense of legal mechanisms and instruments, are needed. Second, some light should be shed on the importance of the often-neglected issue of tenure rights. Third, the importance of appropriate institutional settings should be discussed to implement and enforce effectively soil-related legal provisions. In addition, some short observations should be made with regard to the interface of soil governance and cultural and ethical aspects, followed by an excursus on current German soil governance as a concrete example.

Core elements of good soil governance at the national level

The following overview on the core elements of good soil governance are not meant to be comprehensive; the chapter provides a general compilation of the most relevant aspects.

Soil as natural resource which provides several crucial social and ecological services

Sustainable soil management faces cross-cutting challenges. Thus, soils have to be managed sustainably with regard to all uses, threats and drivers. To cope with these challenges, it is instrumental to mention three objectives that should be upheld for each national legislation. It is important that the objectives are understood as legally binding: First, each legislation should acknowledge

¹⁷See the "contract" of the new German Government: <https://www.bundesregierung.de/breg-de/aktuelles/koalitionsvertrag-2021-1990800>.

¹⁸https://ec.europa.eu/environment/publications/eu-soil-strategy-2030_en.

¹⁹<https://www.umweltbundesamt.de/en/webinar-soil-governance-southeast-asia-in-focus>.

that soils are a natural resource, which needs to be protected and used in a sustainable manner to stress that soils which are owned privately nevertheless bear value that is of interest to all. Thereby, it would be clarified that owners of land and soils do not have an unconditional right to exploit the owned resource (Ruppel and Ginzky, 2021).

Second, each legislation should stipulate as a legally binding objective that the social and ecological services of soils are to be maintained and at best enhanced. Such an objective would also oblige landowners to manage their land and soil sustainably, respecting the services provided by the soils. Moreover, such an objective would also bind the governmental entities. Third, to abide by the international agreements, it would also help and be effective to stipulate land degradation neutrality as a legally binding objective of the national soil legislation. Thereby, a requirement would be established on the implementation of this objective by, *inter alia*, the LDN response hierarchy with the four elements of avoid, reduce and reverse as well as precautionary planning as foreseen in the UNCCD LDN conceptual framework (Bodle and Stockhaus, 2019).

Regulatory concept: Framework law and sectoral provisions

A few states only have their own act on soil issues (Peake and Robb, 2022). Moreover, these specific acts often do not address all soil aspects comprehensively. For example, the German Federal Act on Soils primarily entails provision concerning the identification, assessment and rehabilitation of brownfield sites (areas of land previously developed rather than unconverted greenfield sites). In addition, several legal studies have concluded that a legislative framework on sustainable soil management is recommendable because it could determine the most essential aspects of good soil governance, such as the objective and main operative provisions. At the same time, however, it would need to leave regulation of the specificities of certain soil threats, such as the handling of pesticides, urbanization processes, building of infrastructure, to specific provisions in other pieces of legislation specific to certain sectors (with regard to the envisaged model law in Africa – Ruppel and Ginzky, 2021).

What could or should such a framework law entail? It seems to be reasonable that it should stipulate the three objectives of soils being a natural resource, maintenance of the social and ecological services of soils and implementation and enforcement of the objective of land degradation neutrality. Furthermore, it should establish a procedure that would allow for *ex ante* control of activities that may be detrimental to soils. The applicant for such activities would need to be obliged to at least minimize or at best completely prevent probable negative effects to the soil. To have the necessary information available to propose an informed decision as to whether the activity may proceed, the framework law should require an environmental impact assessment to be conducted for such activities. And moreover, an obligation should be included that the applicant and operator of such an activity that causes negative effects on soils which could not be avoided must compensate by enhancing soil conditions in other locations. Thereby, the objective of land degradation neutrality as the balance of ongoing land and soil degradation and rehabilitation or restoration projects (Ehlers, 2016) could be realized (Bodle, 2017).

Complementary provision in other pieces of legislation

As mentioned above, there are many drivers of soil degradation and various soil threats. For example, the ongoing land take by housing or transportation infrastructure development, which may result in an extreme form of land and soil degradation called sealing; unsustainable agricultural practices, e.g. inappropriate use of peatlands which sequester carbon and are therefore important for climate mitigation; contamination of soils at brownfield sites or by new ubiquitously introduced and often persistent hazardous substances or plastics; as well as the role of soils for climate adaptation particularly in urban areas. These different drivers, threats and uses of soil require specific regulatory approaches. It is not possible to outline all of these regulations in the framework law concerning sustainable soil management. Thus, specific pieces of legislation will need to supplement it and flesh out these different regulatory approaches.

Nevertheless, the framework law should set the three overarching objectives towards which all these sectoral provisions need to be directed in order to manage effectively, coherently and concisely all uses of soils and all drivers and threats.

Relevance of soil data and information

Another important aspect of good environmental governance is the availability of data and information. Without a sufficient level of information, effective governance is not possible. The selection, assessment, documentation and distribution of data and information are additional aspects that should be dealt with in the framework law with its overarching nature (Ginzky, 2022).

There are several categories of necessary soil data and information. Information is necessary on the condition of soils in a specific location, the potential services a specific soil plot could provide, its vulnerability and its current uses. In addition, information is also needed on the potential effects of specific activities, such as the impact of pesticides, of (partial) sealing or of the introduction of contaminants. Finally, information on the techniques used also would be essential, including measures to reduce the negative effects.

To have such information available, appropriate scientific entities must be established in all countries. They should gather, evaluate and assess soil data and information. Furthermore, the information needs to be made available to policy makers and other users (private sector, farmers).

To create these scientific entities, to establish a strong science–policy interface and to reach out to civil society is a long-term task, but nevertheless, it is indispensable. The legislative basis for the scientific entity and for the collection, management and security of soil data and information should also be provided by the framework law (Ginzky and Ruppel, 2021).

Specified requirements for all soil uses in a legally binding form

Many research studies have shown that the success of soil governance in practice depends considerably on legally binding standards for the various soil threats, inter alia, compaction, erosion, soil sealing, loss of carbon and biodiversity, salinization, water scarcity, flooding (Ginzky, 2022; Ginzky and Ruppel, 2021). Standards are meant to set conditions for certain activities. They should entail standards on environmental quality (results-based) or on the design of the techniques applied (practice-based). Thereby, it should be determined which and at what level of negative effects on the environment, in our case specifically on soils, are acceptable. In other words, what kind of use is permitted legally and what is prohibited. For example, what amount of fertilizer may be applied, or not only what kind of chemicals may be released from industrial facilities but also what level of residue is acceptable in the soil?

These standards need to be legally binding to bind both the land/soil user and commit the competent governmental entities to uphold that standard. Legally binding standards would then work towards two ends: First, the user would know what is legally permissible. Second, the competent authorities would be instructed as to what needs to be implemented and enforced; not vague obligations and ambitions, but very clear and specific requirements.

Experience shows that the development and adoption of such standards is very demanding and time- and resource-consuming. As legally binding standards determine what is permissible and what is not, the adoption of such standards is in most cases very controversial. Thus, it helps that the FAO has published guidelines on various soil threats; for example, on pollution, erosion and biodiversity, which could be used as a basis for the national decision-making process or at least as a kind of anchor for all competent authorities that are in charge of implementing the national provisions.²⁰

Procedural rights and courts

Finally, stakeholders and civil society need to become sufficiently involved. The commonly accepted governance instruments, e.g. access to environmental information, public participation and access to justice, need to be implemented with regard to activities that are relevant to soils. Respective provisions could be foreseen in either sectoral provisions or in the framework law mentioned above.

In a few counties, such as Kenya (Kameri-Mbote, 2021), specific courts have been established to make judgments on environmental and land issues. Such an approach could serve three needs. First, to increase the level of judges' legal knowledge about how the law relates to issues of soil use and protection. Second, to emphasize the importance of these challenges in the national context. Third, to foster access to justice for those seeking redress due to unsustainable soil management.

Tenure rights

Clarity on land tenure is crucial for sustainable soil management because management responsibility is often linked to tenure. In other words, if it is not clear who owns or has access to a plot of land and the soil below, there is no clarity as to who is in charge of ensuring sustainable handling of it. In addition, there is a social dimension of land tenure as access to land is extremely important in many states where people depend on agriculture for their livelihoods and subsistence. As land has become a commodity of interest due to the food crisis at the beginning of the century, the acquisition of land for current and future uses by often powerful national or foreign investors has placed local communities under additional pressure.

In some countries and regions, the lack of clarity over tenure is caused by the conflict of so-called modern law and traditional concepts of land ownership. African countries are relevant case studies in this context (Ruppel and Ginzky, 2021; Tamasang, 2022; Towela Sambo, 2021). Foreign or even national investors may benefit from such 'obscurity' to strengthen their interests, particularly in weak or corrupt land rights systems.

Lack of legal clarity on tenure is often a strong impediment to sustainable soil management, particularly in developing countries as many examples have shown. In cases of legal non-clarity, people are not clear about their responsibilities or duties and there is less incentive to treat the land well because that individual does not know how long they will have access to it, so short-term exploitation may be prioritized. In addition, strong investors may misuse this legal lack of clarity to circumvent the proper registration process or stake a claim in the knowledge that they will win support rather than a diffuse number of less powerful interested holders, for instance (Tamasang, 2022; Ginzky and Ruppel, 2021; Ginzky et al., 2021a).

Clarification of tenure rights would thus serve several purposes. First, it could support poverty eradication. Second, it is one important instrument to control national or foreign investors by avoiding illegitimate land acquisitions. Simply, clarity on land tenure is the basis for more environmental responsibility, which would be particularly useful to strengthen sustainable soil management.

Importance of administrative arrangements – the case of implementation and enforcement

Soil governance largely depends on effective institutional settings and arrangements for its implementation and enforcement. This aspect is often neglected because of its technical nature. Nevertheless, as experience has shown, brilliant substantive provisions are only ink on the paper without effective implementation and enforcement (Ginzky et al., 2021a).

There are mainly four administrative tasks for governmental entities to effect soil governance. First, the gathering of soil information, which also requires inter alia a lot of environmental monitoring on the ground, assessment of the data gathered

²⁰<https://www.fao.org/soils-portal/en/>

and its later dissemination to policy makers and civil society. Second, legally binding standards need to be developed and adopted. Third, potentially detrimental activities must be subject to an *ex-ante* control mechanism by governmental entities. Fourth, compliance with substantial provisions needs to be monitored and, in the case of non-compliance, enforced (Ginzky, 2022).

These tasks have to be attributed to specific governmental entities to clarify who is in charge and to avoid overlaps of responsibilities and lack of coordination. To this end, the roles and responsibilities of the various ministries and competent authorities need to be determined clearly by legal provisions.

Although implementation and enforcement seem to be primarily a technical issue, the complexity of the challenge to establish an effective institutional setting should not be underestimated. Responsibilities accompany reputation, influence and power, therefore, the decision-making process could be politically very controversial amongst the entities involved.

Soils, culture and ethics

Nature and landscape trigger emotions, perceptions of the interlinkages of nature and human beings and thus traditions, culture and ethical values (Heuser, 2017). The Catholic religion has a specific understanding of nature and soils, inter alia, regarding humans as double-being: 'earthbound' and spiritual (Voigt, 2019). Similarly, for other world religions such as Buddhism or Hinduism.

It is important to recognize that the magnitude and diversity of cultural or ethical perceptions is somewhat endless (Shikongo, 2016). Cameroon alone assembles more than 270 culturally different entities with almost the same number of languages. Thus, it is called 'l'Afrique en miniature' (Tamasang, 2021). A North American indigenous perspective on Mother Earth asks: "who is governing who?" (Poitras, 2022).

These cultural or ethical perceptions at local, regional or national levels should be taken into account when drafting instruments and concepts for soil governance. It is important to consider different cultural or ethical understandings of societal groups respectfully, such as local communities, in attempting to understand how soils are used, by whom, for what purpose, etc. Such a respectful approach is also reasonable from the point of view of law enforcement as people are more willing to comply with regulations if their perceptions have been referred to in the legislation.

Excursus one: Soil governance in Germany

This subsection gives example of national soil governance (Peake and Robb, 2022), which shows how complicated it is to achieve comprehensive and effective soil governance.

Industrialization in Germany during the middle of the 19th Century caused severe contamination of soil and the local groundwater bodies, leading to about 425,000 brownfield sites. Calculations of the costs for rehabilitation exceed more than 500 billion Euro (Ginzky, 2021). A second driver is the threat to soil linked to ongoing industrial activities. Urbanization and related necessary infrastructure, such as streets and railways, are a third factor (Schlacke and Jürschik, 2018). The land take in Germany of about 58 ha per day is still unsustainably high; about half of which is ultimately sealed. The fourth threat to soil is linked to the continuous intensification of agriculture. Climate change is the last (but probably not least) driver of soil impacts in Germany, which has not yet been thoroughly assessed scientifically.

Germany is one of the few states that has a specific Soil Protection Act (SPA).²¹ The German Soil Protection Act was adopted in 1998 and therefore 2018 marked its 20th anniversary. Paragraph 1 of the German Soil Protection Act stipulates that soil functions must be maintained sustainably and or rehabilitated, and that to this end both reactive and precautionary measures have to be taken. The overall objective of the law is broad enough to encompass all drivers of soil degradation and all soil threats. Moreover, the law addresses soil as a natural resource and stresses the need to foster its ecosystem services (functions is another term) of soils.

However, the core operative provisions of this act focus on the regulation of and how to clean up contaminated industrial sites, including landfill sites (§§ 4 and 10 SPA). The Soil Protection Act's focus is thus mainly on the rehabilitation of soils contaminated by industrial uses.

The law also outlines some provisions with regard to 'good agricultural practices' (§ 17 SPA). However, these provisions are not mandatory and are very vague. They relate to the requirement that agricultural practice should consider the site-specific conditions of a plot of soil. Almost no legally binding standards are in place with regard to these good agricultural practices, and the obligation to comply does not have a stringent behavior-altering effect.

Germany is a very industrialized country, therefore, there is also a need to avoid future contamination of the soil by ongoing or future industrial activities. Thus, the German Act on Emission Control²² requires *ex-ante* permission for industrial sites. The permission to build industrial sites may only be granted if detrimental effects on the environment, including soils, are prevented. With regard to the maintenance of soil functions, before production starts the operator is required to examine and document the status of soils on the industrial site. After closure of the site, if significant negative effects have been caused the operator is required to rehabilitate the site to the 'original' status (§ 5 BImSchG). The provision clearly creates an incentive to avoid negative effects on soils caused by new industrial uses.

²¹Gesetz zum Schutz vor schädlichen Bodenveränderungen und zur Sanierung von Altlasten (Bundes-Bodenschutzgesetz - BBodSchG), 17.3.1998.

²²Gesetz zum Schutz vor schädlichen Umwelteinwirkungen durch Luftverunreinigungen, Geräusche, Erschütterungen und ähnliche Vorgänge (BImSchG), 15.3.1974.

Negative effects on soils by urbanization are dealt with in a specific piece of legislation. The German law on town planning and buildings²³ requires that soils should be managed economically and carefully. The act also demands town planning processes be structured so that the various concerns, i.e. need for accommodation and infrastructure, space for industrial uses, protection of the environment, social security, etc., must be taken into account and potential conflicts between them be settled. The act stipulates that if a town planning decision causes negative effects on nature, including soils, these negative effects have to be compensated for. This provision is of great importance for nature conservation, although it also has some drawbacks (Bodle et al., 2020). Concerning soils, the provision cannot be thoroughly implemented because important information is lacking on the condition of soils throughout Germany and no binding standards have been adopted yet.

Agriculture may be the cause of several soil threats. Contamination could result from the input of manure, fertilizer or pesticides. For all these categories of substances, specific pieces of legislation are in place in Germany, setting limits for emissions, uses or environmental quality standards (Möckel et al., 2014). For almost all other, not substance related, soil threats caused by agriculture, such as soil erosion, decline in organic matter, decline in biodiversity or salinization, no legally binding standards have been adopted.

Land take, or infrastructure such as streets, airports or railways, are dealt with by further legislation. Infrastructure measures usually also require *ex-ante* permission. The competent authority must, before granting the permission, consider all relevant concerns including soil aspects. Similar to urbanization, it is mandatory to offset detrimental effects on nature caused by infrastructure developments. However, there are the drawbacks mentioned above.

The Soil Protection Act stipulates precautionary limits with regard to contamination which are legally binding. These limits have to be considered when other sectoral provisions are implemented and effects on soils are to be managed. With regard to soil contamination, the German Soil Protection Act could therefore be regarded as a kind of a framework regulation.

In summary, it has been shown that German soil governance addresses the various drivers of soil degradation and threats by a combination of the overarching Soil Protection Act and several sectoral provisions. Legally binding standards, however, have been adopted only with regard to soil contamination. For most other soil threats, legally binding requirements, for example, on compaction or erosion are missing. Consequently, law enforcement is weak with regard to this (Ginzky et al., 2019). The division of soil-related provisions over several acts is not a problem in itself, but what weakens the legal implementation or enforcement is if there are no legally binding standards to which users may be held, such as those of environmental quality (Ginzky et al., 2019).

Of note: Land degradation neutrality has not yet been included as an objective in German soil governance (Bodle et al., 2019).

Excursus two: Methodological observations

It is a complex and demanding task to either draft or amend soil legislation. To be able to submit well substantiated proposals as a legislator for political discussion, it is essential to apply a convincing methodology. The following subsection provides some general considerations on how such a task could be approached. This methodological approach has been used to inform existing legislation at the national level (Ruppel and Ginzky, 2021) and for the various country reports in the volumes of the *International Yearbook of Soil Law and Policy*.

The methodology encompasses four main steps. First, a description of the historical, societal, economic and political conditions and circumstances is required. This could be used primarily to inform foreigners. Nevertheless, it could also be instrumental for national discussion to establish a common 'starting point.' Second, it is appropriate for determining what are the main soil threats and drivers in the respective country. This may differ significantly. The outcome of the analysis should be used to decide on the focus of the national soil legislative process.

Third, if the focus is on amendment of an existing law, an analysis of the legal provisions is needed. This should refer mainly to the country's core threats and drivers. Technically speaking, the analysis should include constitutional provisions, environmental laws, including subsidiary environmental standards and guidelines, tenure rights, as well as international and regional treaties and agreements. Finally, the institutional, organizational and financial settings should be analyzed, including aspects such as data gathering and distribution, administration, roles, communities, local chiefs to determine the strengths and weaknesses of current arrangements and to identify room for improvement. This is of eminent importance because implementation and enforcement are key (Ginzky, 2022).

With regard to the last two steps of the investigation, it is advisable to investigate solutions that have been applied successfully in other national legislation. Considerable knowledge is available to this end. The *SoiLEX*, as described above, is a database that assembles information on most national legislation and is thus an indispensable tool in this context. In addition, each of the five volumes of the *International Yearbook of Soil Law and Policy* includes several legal analyses of national soil governance systems; in total more than 30 chapters which explain both the strengths and weaknesses. In addition, the comparative work for a model law on sustainable soil management in Africa (Ruppel and Ginzky, 2021) and the two anthologies on Indian (Desai, 2022) and African soil law (Yahyah et al., 2020) could be considered. Therefore, all this knowledge and expertise could and should be considered when drafting national legislation on sustainable soil management.

²³Baugesetzbuch (BauGB), 23.6.1960.

Transboundary trade and sustainable soil management

International trade on the one hand and food security and sustainable soil management on the other hand are closely interlinked, which is also addressed by the SDGs (Ehlers, 2016; Elver, 2018). Food security depends strongly on the sustainable management of soils because with water it is the essential natural resource to grow food, at least in most developing countries. Food security, therefore, could be regarded as a precondition to achieve sustainability in general (Ruppel, 2021).

International trade, on the one hand, could support the aim of food security by several mechanisms, such as the immediate supply of food but also with the supply of seeds, fertilizers or agricultural technical equipment (Ruppel, 2021). On the other hand, reality suggests that international trade could also have a huge destabilizing effect on food security in regions because of disruptions to supplies, growing unilateralism and competition over agricultural resources, including soil itself (Windfuhr, 2016). Moreover, international trade could and has been an instrument of geopolitics (Ruppel, 2021).

International trade in agricultural products needs to ensure that the sustainable management of soils in both exporting and importing countries is not impeded or even made impossible. To this end, two main objectives should be regarded as requirements to international trade in agricultural products. First, it should not lead to direct or indirect negative effects on soils. Second, international trade in raw agricultural products which are exported by a developing country should be combined with an obligation on the importing country (potentially or likely in the Global North) to support the development and establishment of processing facilities in the exporting country (potentially or likely in the Global South). This requirement would create more autonomy in those countries in the Global South, which is in accord with the spirit of the SDGs (Ankersmit and Partiti, 2020).

It needs to be highlighted that current international trade law is somewhat unclear with regard to the two objectives just mentioned. Basically, it is currently about enabling more trade, i.e. free trade (Ruppel, 2021). Environmental standards in relation to how they factor these into trade structures and processes need to be developed further to work towards better food security and to avoid repercussions for sustainable soil management in various locations throughout the world.

Nevertheless, to avoid negative and hindering effects on soils from international trade in agricultural products is another crucial challenge to allow for sustainable soil management locally, which needs to be further discussed.

Outlook

The twenties of this century will be the decade of transformation. This will be either by design or by disaster. Sustainable soil management will be one core element to achieve transformation by design with other crucial topics, such as energy efficiency, energy supply systems based on renewables, a circular economy and sustainable agriculture. Thus, one could state emphatically that this decade will be the decade of soils. Whether this is over-stated or not, to achieve sustainable management of soils is non-negotiable.

Sustainable soil management can only be achieved with good governance. There are other factors, such as awareness, knowledge, capacity building and economic instruments. The chapter first explained the complexity and importance of sustainable soil management and soil governance. In addition, the existing soil-related instruments at the international level have been outlined. Moreover, the most important core elements for good governance of sustainable soil management have been explained.

There is no alternative to success. Thus, there is a need to speed up, to cooperate, to exchange views, to experiment and to evaluate. Transformation is not a result, it is primarily a process (Lehndorff, 2021; Ginzky et al., 2021b). It is hoped that this chapter lays out a basis to inform and support all experts, stakeholders and interested people in their ambition to work towards a world in which sustainable management of soils is not only an aspiration, but a reality.

References

- Altwater S, Dooley E, and Roberts E (2014) Legal instruments to implement the objective of "land degradation neutral world" in international law. Umweltbundesamt, Texte 09/2015. Available at: https://www.umweltbundesamt.de/sites/default/files/medien/378/publikationen/texte_19_2015_legal_instruments_to_implement_the_objective_land_degradation_neutral_world.pdf.
- Ankersmit L and Partiti E (2020) *Alternatives for the 'Energy and Raw Materials Chapters' in EU trade agreements – An inclusive approach*. Available: *Alternatives-for-the-'Raw-materials-and-Energy-Chapters'-in-EU-trade-agreements-web.pdf* (power-shift.de).
- Bodle R (2017) Implementing Land Degradation Neutrality at National Level: Legal Instruments in Germany. *International Yearbook of Soil Law and Policy* 2: 287–308.
- Bodle R and Stockhaus (2019) Geeignete Rechtsinstrumente für die nationale Umsetzung der bodenbezogenen sustainable development goals, insbesondere des Ziels einer "land degradation neutral world" Available at: <https://www.umweltbundesamt.de/publikationen/geeignete-rechtsinstrumente-fuer-die-nationale%20>.
- Bodle R, Stockhaus H, Wolff F, Scherf CS, and Oberthuer S (2020) Improving International Soil Governance: Analysis and Recommendations. available at: <https://www.umweltbundesamt.de/publikationen/improving-international-soil-governance-analysis>.
- Boer B (2014) Land degradation as a common concern of humankind. In: Lenzerini F and Vrdoljak F (eds.) *International Law for Common Goods*, 289–307.
- Boer B and Hannam I (2003) Legal Aspects of the sustainable Soils. *Rev Eur Com Int Environmental Law* 12(2): 149–163.
- Boer B and Hannam I (2017) *Land Degradation Law. Legal Studies Research Paper No. 17/100*. The University of Sydney.
- Boer B, Ginzky H, and Heuser I (2016) International Soil Protection Law: History, Concepts and Latest Developments. *International Yearbook of Soil Law and Policy* 1: 49–72.
- Bowling C, Pierson E, and Ratté S (2016) The Common Concern of Humankind: A Potential Framework for a New International Legally Binding Instrument on the Conservation and Sustainable Use of Marine Biological Diversity in the High Seas. *White Paper* 1–15. under: http://www.un.org/depts/los/biodiversity/prepcom_files/BowlingPiersonandRatte_Common_Concern.pdf.
- Cottier T, Aerni P, Karapinar B, Matteotti S, de Sepibus J, and Shingai A (2014) The Principle of Common Concern and Climate Change. *Archiv des Völkerrechts* 52: 293–324.

- Cowie AL, Orr BJ, Castillo Sanchez VM, Chasek P, Crossman ND, Erlewein A, Louwagie G, Maron M, Metternicht GI, Minelli S, Tngberg AE, Walter S, and Welton S (2018) Land in Balance: the scientific conceptual framework for Land Degradation Neutrality. *Environmental Science and Policy* 79: 25–35.
- Desai B (2022) *Soil Law and Governance in India*, forthcoming.
- Ehlers K (2016) Chances and Challenges in Using the Sustainable Development Goals as a New Instrument for Global Action Against Soil Degradation. *International Yearbook of Soil Law and Policy* 1: 73–84.
- Elver H (2018) Human Rights Based Approach to Sustainable Agricultural Policies and Food Security. *International Yearbook of Soil Law and Policy* 3: 347–372.
- EU Commission (2012) *Report: The implementation of the Soil Thematic Strategy and ongoing activities*. COM (2012) 46 final.
- Fee E (2018) Implementing the Paris Climate Agreement: Risks and Opportunities for Sustainable Soil Management. *International Yearbook of Soil Law and Policy* 3: 249–270.
- Flasbarth J (2016) Soils need International Governance: A European Perspective for the First Volume of the International Yearbook of Soil Law and Policy. *International Yearbook of Soil Law and Policy*, 1: 15–20.
- Fowler RJ and Hannam I (2019) Critique of the Report “Improving International Soil Governance: Analysis and Recommendations” *International Yearbook of Soil Law and Policy*, 4: 177–196.
- Ginzky H (2016) *Bodenschutz weltweit – Konzeptionelle Überlegungen für ein internationales Regime*. Handbuch Boden 1–32.
- Ginzky H (2017) The sustainable management of soils as a common concern of humankind: How to implement it? *International Yearbook of Soil Law and Policy*, vol. 2, 433–450.
- Ginzky H (2020) Good Governance for “Sustainable Soil Management”? on National and International Level: How to do it? In: Yahyah H, Ginzky H, Kasimbazi E, Kibugi R, and Ruppel OC (eds.) *Legal Instruments for sustainable soil management in Africa*, 35–54.
- Ginzky H (2021) Soil protection governance in Germany. In: Ginzky H, et al. (ed.) *International Yearbook of Soil Law and Policy*, vol. 4, 295–333.
- Ginzky H (2022) The case of implementation and enforcement. In: *Soil Security – Special Issue “Soil Governance”*. Available: Soil governance: The case of implementation and enforcement - ScienceDirect.
- Ginzky H and Ruppel OC (2021) Executive Summary. In: Ruppel OC and Ginzky H (eds.) *African Soil Protection Law – Mapping out options for a model legislation for improved sustainable soil management in Africa – a comparative legal analysis form Cameroon, Kenya and Zambia*, Nomos, 31–52.
- Ginzky H, Ruppel O, Kibugi R, and Engelberg W (2019) Workshop “Implementing Land Degradation Neutrality in Afrika: Means, Instruments and Institutional Challenges ?”: Outcome Document. *International Yearbook of Soil Law and Policy*, vol. 4, 143–154.
- Ginzky H, Kameri-Mbote P, Tamasang CF, Towela Sambo P, and Ruppel OC (2021a) Mapping out options for model legislation for sustainable soil management in Africa. In: *African Soil Protection Law – Mapping out options for a model legislation for improved sustainable soil management in Africa – a comparative legal analysis form Cameroon, Kenya and Zambia*, Nomos, 379–454.
- Ginzky H, Loewe C, and Neßhöver C (2021b) Lessons from the Corona Crisis: New guiding principles required for environmental and sustainability policy? – a discussion paper. at: <https://www.umweltbundesamt.de/en/publikationen/lessons-from-the-corona-crisis-new-guiding>.
- Hannam (2022) Soil governance and land degradation neutrality. In: *Soil Security – Special Issue “Soil Governance”*. Available: <https://www.sciencedirect.com/journal/soil-security/special-issue/10VQ9HFPD6Z>.
- Hannam I and Boer B (2002) *Legal and Institutional Frameworks for sustainable use of soils*. IUCN Environmental Policy and Law Paper No. 45.
- Heuser IL (2017) Development of Soil Awareness in Europe and other Regions: Historical and Ethical Reflections About European and (International) Soil Protection Law. *International Yearbook of Soil Law and Policy*, vol. 2, 451–474.
- Heuser I (2022) Soil Governance in the current European Union Law and in the European Green Deal. In: *Soil Security – Special Issue “Soil Governance”*. Available: <https://www.sciencedirect.com/journal/soil-security/special-issue/10VQ9HFPD6Z>.
- International Panel on Biodiversity and Ecosystem Services (2018) The Assessment Report on Land Degradation and Restoration. Available: <https://www.bing.com/search?q=IPBES+land&form=QBLH&sp=1&pq=ipbes&sc=8-5&q=8&sk=cvid=C04406F7D8EB4ECE9C048F6870D009AE>.
- International Panel on Climate Change (2018) Special report: Global Warming of 1,5 degree Celsius. Available: <https://www.ipcc.ch/sr15/>.
- International Panel on Climate Change (2020) Special report: Climate Change and Land. Available: <https://www.ipcc.ch/srcc/>.
- Kameri-Mbote P (2021) Country report for Kenya. In: *African Soil Protection Law – Mapping out options for a model legislation for improved sustainable soil management in Africa – A comparative legal analysis form Cameroon, Kenya and Zambia*, Nomos, 53–172.
- Kibugi R (2017) Soil Health, Sustainable Land Mangement and Land Degradation n Africa: Legal Options on the Need for a Specific African Soil Convention or Protocol. *International Yearbook of Soil Law and Policy*, vol. 2, 387–412.
- Kibugi R (2020) In: Yahyah H, Ginzky H, Kasimbazi E, Kibugi R, and Ruppel OC (eds.) *Salvation of Soils: Exploring the Windows of Opportunity through Nationally Determined Contributions to implement Sustainable Soil Management, Legal Instruments for Sustainable Soil Management in Africa* 15–34.
- Lal R, Safriel U, and Boer B (2012) *Zero Net Land Degradation: A New Sustainable Development Goal for Rio+ 20*. Report for UNCCD.
- Lehndorff S (2021) New Deal Means Being Prepared for Conflict. Available: <https://www.vsa-verlag.de/nc/buecher/detail/artikel/new-deal-means-being-prepared-for-conflict/>.
- Linz F and Lobos I (2016) Boden und Land in der internationalen Nachhaltigkeitspolitik – von der globalen Agenda zur lokalen Umsetzung. *Zeitschrift für Umweltrecht* 195–199.
- Mastrojeni G (2018) UNCCD COP 13: From Awareness to Action in a Complex World. *International Yearbook of Soil Law and Policy*, vol. 3, 229–248.
- Minelli S, Erlewein A, and Castillo V (2016) Land Degradation Neutrality and the UNCCD: From Political Vision to Measurable Targets. *International Yearbook of Soil Law and Policy*, vol. 1, 85–104.
- Möckel S, Köck W, Rutz C, and Schramek J (2014) *Rechtliche und andere Instrumente für vermehrten Umweltschutz in der Landwirtschaft, Berlin*.
- Munyuki-Hungwe R (2020) Overview of main challenges with regards to land tenure in Africa: Factual and legal aspects. In: Yahyah H, Ginzky H, Kasimbazi E, Kibugi R, and Ruppel OC (eds.) *Legal Instruments for sustainable soil management in Africa*, 121–132.
- Ortiz Garcia S, et al. (2022) Soil governance and sustainable agriculture in Mexico. In: *Soil Security – Special Issue “Soil Governance”*. Available: <https://www.sciencedirect.com/journal/soil-security/special-issue/10VQ9HFPD6Z>.
- Orubebe BB (2020) Soil Governance and Sustainable Land Use System in Nigeria: the Paradox of Inequalities. *Natural Resource Conflict and Ecological Diversity in a Federal System* 157–180.
- Peake LR and Robb C (2022) The global standard bearers of soil governance. *Soil Security – Special Issue “Soil Governance”* 6. <https://doi.org/10.1016/j.soisec.2022.100055>.
- Poderatti G (2022) Human rights aspects and soil governance: A special focus on land grabbing. In: *Soil Security – Special Issue “Soil Governance”*. Available: <https://www.sciencedirect.com/journal/soil-security/special-issue/10VQ9HFPD6Z>.
- Poitras M (2022) Soil Governance: Wo is governing who ? A North American Indigenous perspective on mother Earth. In: *Soil Security – Special Issue “Soil Governance”*. Available: <https://www.sciencedirect.com/journal/soil-security/special-issue/10VQ9HFPD6Z>.
- Raffelsiefen M and Straßburger (2016) The Protection of Soils: Does the European Union Live up to its own Ambitions? *International Yearbook of Soil Law and Policy*, vol. 1, 389–410.
- Ruppel OC (2021) Soil Protection and the right to food: Sustainability implications for global Climate Governance and World Agricultural Trade? In: Ruppel OC and Ginzky H (eds.) *African Soil Protection Law – Mapping out options for a model legislation for improved sustainable soil management in Africa – A comparative legal analysis form Cameroon, Kenya and Zambia*, Nomos.
- Ruppel OC and Ginzky H (2021) African Soil Protection Law – Mapping out options for a model legislation for improved sustainable soil management in Africa – a comparative legal analysis form Cameroon, Kenya and Zambia, Nomos. Available: <https://www.nomos-elibrary.de/10.5771/9783748908043/african-soil-protection-law>.
- Schlacke S and Jürschik U (2018) Urbanisation and Urban Land Use: A Normative Compass for Sustainable Urban Governance. *International Yearbook of Soil Law and Policy*, vol. 3, 3–27.
- Shikongo ST (2016) Greetings to the Launch of the Yearbook from an African Perspective. *International Yearbook of Soil Law and Policy*, vol. 1, 3–8.
- Streck C and Gay A (2016) The Role of Soils in International Climate Change Policy. *International Yearbook of Soil Law and Policy*, vol. 1, 105–128.

- Tamasang CF (2021) Country report for Cameroon. In: *African Soil Protection Law – Mapping out options for a model legislation for improved sustainable soil management in Africa – A comparative legal analysis form Cameroon, Kenya and Zambia, Nomos*, 177–294.
- Tamasang CF (2022) Land tenure legislation and soil security concerns in Cameroon. In: *Soil Security – Special Issue “Soil Governance”*, Available: <https://www.sciencedirect.com/journal/soil-security/special-issue/10VQ9HFPD6Z>.
- Towela Sambo P (2021) Country report for Zambia. In: *African Soil Protection Law – Mapping out options for a model legislation for improved sustainable soil management in Africa – A comparative legal analysis form Cameroon, Kenya and Zambia, Nomos*, 295–378.
- UNCCD (2015) Report of the Conference of the Parties on its twelfth session, held in Ankara from 12 to 23 October 2015. Part two. ICCD/COP(12)/20/Add.1 <http://www.unccd.int/Lists/OfficialDocuments/cop12/20add1eng.pdf>.
- UNCCD (2017) Conceptual Framework for Land Degradation Neutrality. see at: https://www.unccd.int/sites/default/files/documents/2019-06/LDN_CF_report_web-english.pdf.
- United Nations (2012) *General Assembly, Resolution adopted by the General Assembly on 11 September 2012, 66/288. The Future We Want*.
- United Nations (2015) General Assembly, Seventieth Session, No. 11688, Agenda items 15 and 116, Resolution adopted by the General Assembly on 25 September 2015. In: *Transforming our world: the 2030 Agenda for sustainable development*, 1. A/RES/70/1.
- Voigt M (2019) Perceptions of soils in catholic theology. *International Yearbook of Soil Law and Policy*, vol. 3, 357–368.
- Windfuhr M (2016) FAO: voluntary Guidelines on responsible Governance of Tenure of Land, Forests and Fisheries – Relevance, Reception and First Experience in Reception. *International Yearbook of Soil Law and Policy*, vol. 1, 203–218.
- Wolff F and Kaphengst T (2016) The UN Convention on Biological Diversity and Soils: Status and Future Options. *International Yearbook of Soil Law and Policy*, vol. 1, 129–148.
- Yahyah H, Ginzky H, Kasimbazi E, Kibugi R, and Ruppel OC (2020) *Legal Instruments for sustainable soil management in Africa*.

The economics of soils' contribution to human well-being

Bartosz Bartkowski and Julian R Massenberg, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	547
Soil-based ecosystem services	548
Collective goods and property rights regimes	549
Economic valuation of soil-based ecosystem services	550
Incentive-based instruments for soil protection	552
Conclusions	553
References	554

Abstract

Soils influence human well-being in manifold ways. However, their multifunctionality implies trade-offs. Moreover, soils are a scarce stock of natural capital. Many soil-based ecosystem services are collective goods, whose provision is affected by externalities of soil management. This chapter provides an economic perspective on soils' contributions to human well-being and the associated challenges in terms of public preferences, property rights regimes and policy. It offers an overview of key concepts such as soil-based ecosystem services, property rights and collective goods dilemmas, economic valuation as well as incentive-based policy instruments that help align soil management with public preferences and societal well-being.

Glossary

Collective good A good providing benefits that can be enjoyed by anyone and from whose enjoyment no-one can be excluded.

Externality Side effect of an economic activity (e.g. soil management) that affects third parties without their consent.

Incentive-based policy instruments Policy instruments that rely on shifting relative prices (e.g. taxes, offsets), rather than prohibiting or commanding specific actions (e.g. emission or technology standards).

Information asymmetry A situation among at least two interacting agents, in which one of the agents possesses more information relevant to their interaction than the other(s).

Institutions Conventions, norms and formally sanctioned rules of a society (e.g. laws, social norms, property rights).

Marginal value The value attached by people to an incremental change of one unit (quantity or quality) of a good or service.

Market failure A situation in which unregulated market activity leads to societally undesirable outcomes (e.g. due to externalities).

Opportunity costs Foregone benefits of alternative uses of a resource (e.g. the profit from agricultural yields foregone in the event of conserving a piece of land).

Relative scarcity A good is scarce if its intended uses exceed the available quantity. Scarcity is relative if the good can be substituted with respect to a given use.

Social dilemma A situation in which independent rational behavior of individuals leads to an inferior overall outcome as compared with a situation in which individual actions were coordinated.

Key points

- Soils provide multiple benefits to human well-being (soil-based ecosystem services).
- Soil management affects the provision of soil-related collective goods.
- Existing legal property right regimes obscure the public-good nature of soil-based ecosystem services.
- Economic valuation can help demonstrate the contributions of soils to human well-being and the trade-offs involved.
- Incentive-based policy instruments can help align soil management with public preferences.
- The design of effective and efficient incentive-based soil policy instruments is challenging.

Introduction

Soils influence human well-being in manifold ways, e.g. by storing carbon, absorbing, storing and filtering water, providing habitat for biodiversity as well as contributing to the production of biomass that can be used as food, feed, energy or materials (Baveye et al.,

2016; Dominati et al., 2010). These numerous benefits illustrate soils' multifunctionality. However, soils cannot be managed to maximize all their contributions to human well-being (soil functions or soil-based ecosystem services) simultaneously. Their multifunctionality implies trade-offs. In this sense, they can be considered an essentially scarce resource. Their scarcity is relative, as many of their functions can be substituted or replaced by man-made alternatives, although imperfectly – examples of such substitution are synthetic fertilization, water treatment facilities, flood protection infrastructure, irrigation systems etc. Given the relative scarcity of soils in their multifunctionality and their multiple contributions to human well-being, soils can and should be considered an economic good or, more precisely, a natural capital stock (Dominati et al., 2010). In this view, soil functions and, ultimately, soil-based ecosystem services are benefit streams derived from this stock of natural capital.

Further, while some of the ecosystem services provided by soils benefit predominantly those who own and/or manage the land where they are located, many soil-based ecosystem services provide wider benefits to society. These ecosystem services can be considered public or collective goods. Against this background, a person's private activity, e.g. biomass production, may have negative or positive impacts on a third person not directly involved in the activity, or on society at large. In the case that the person undertaking the action does not consider costs or benefits imposed on others, an uncompensated damage (or spill-over benefit) exists, i.e. a so-called 'externality.' It is the primary role of societal institutions (laws, rules, norms) to internalize them, meaning that measures are undertaken to take into account uncompensated damage or benefits (formally: align private costs/benefits and social costs/benefits). This internalization of externalities can be based on three approaches, depending on context:

- State intervention in the form of corrective taxes (or subsidies in the case of positive externalities).
- Direct negotiations among affected parties, with the aim for contractual internalization (well-defined property rights are required).
- Bottom-up establishment of collective institutions (norms, rules) by those affected (Ostrom, 1990).

The fact that soils exhibit collective good characteristics suggests that there is a need for policy intervention to correct market failure and underprovision of the soil-related collective goods that arise due to soil degradation. This chapter provides an economic perspective on soils' contributions to human well-being and the associated challenges in terms of public preferences, property rights regimes and the design of policy interventions that aim at internalization.

Soil-based ecosystem services

The importance of soils for human well-being can be highlighted and analyzed by means of the ecosystem service concept (Dominati et al., 2010). This concept aims to demonstrate the contributions of ecosystems to human well-being and the associated value humans place on ecosystems (TEEB, 2010). Although scientific interest in the concept has risen substantially since the millennium, so far there has been no clearly defined terminology or a standardized classification of ecosystem services. Ecosystem services are often defined as the indirect and direct contributions of ecosystems to human well-being (TEEB, 2010). The development of the Common International Classification of Ecosystem Services (CICES) is considered an important contribution to a common understanding and standardization (Haines-Young and Potschin, 2018). The CICES differentiates ecosystem services into three overarching categories: provisioning services (e.g. food, timber), regulating and maintenance services (e.g. climate regulation, flow regulation, soil formation) and cultural services (e.g. recreation, esthetic experiences). Within these three main categories, CICES version 5.1 distinguishes between 83 so-called classes of ecosystem services (Haines-Young and Potschin, 2018).

To illustrate the complex relationship between ecosystems and human well-being in the form of a 'production chain,' Haines-Young and Potschin (2010) developed the so-called cascade model (Fig. 1). This model assumes a stepwise relationship between the biophysical context, processes and structures created by living organisms and their potential to provide ecosystem services, and the socio-cultural and economic context associated with benefits and value to humans. Final ecosystem services, i.e. all services that contribute directly to human well-being, are the link between the ecosystem functions (including soil functions) on the one hand and benefits and values on the other hand. Thus, there is a need to distinguish between an ecosystem's potential to provide ecosystem services and the actual supply and demand for ecosystem services, which by definition cannot exist in isolation of human needs (Haines-Young and Potschin, 2010). While the standard version of the cascade model is characterized by a unidirectional flow, decision-making may influence the biophysical structure through pressures, e.g. unsustainable management of resources.

As discussed above, soils fulfil multiple functions (production of biomass, water storage and filtration, nutrient cycling, carbon storage and habitat provision), which form the basis for the provision of ecosystem services. While the ecosystem service concept could make the contributions of soils to human well-being and the impacts of soil management on the provision of ecosystem services visible, CICES does not explicitly specify soil-based ecosystem services (Haines-Young and Potschin, 2018). Thus, defining ecosystem services provided by soils remains challenging because ecosystem services are characterized by complex interactions, e.g. between below-ground and (natural and human-driven) above-ground processes. Therefore, identical soils may provide varying levels of ecosystem services under different management regimes; soil-based ecosystem services are the result of an interplay between inherent as well as manageable soil properties (Dominati et al., 2010) and human needs and preferences, which determine for which soil functions there is demand (making them soil-based ecosystem services). For instance, fertilizer input affects the amount of fiber and food production. In this context, Paul et al. (2021) distinguished between (i) *soil-related* ecosystem services, i.e.

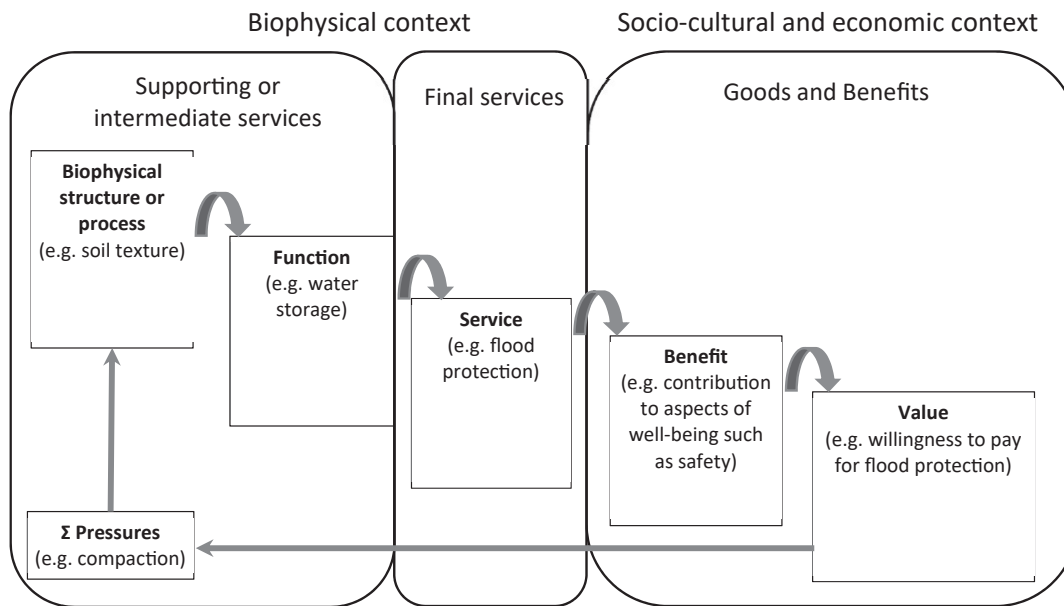


Fig. 1 The ecosystem service cascade model. Own illustration adapted from Haines-Young R and Potschin M (2018) *Common International Classification of Ecosystem Services (CICES) V5.1: Guidance on the Application of the Revised Structure*; see also, Haines-Young R and Potschin M (2010) The links between biodiversity, ecosystem services and human well-being. In: Raffaelli DG and Frid CLJ. *Ecosystem Ecology: A New Synthesis*. Cambridge: Cambridge University Press, 110–139.

all services affected by soil properties, e.g. soil quality by weathering processes (here called soil-based ecosystem services), and (ii) ecosystem services *affected by agricultural management of soil*. Of the 83 ecosystem service classes included in CICES, 29 were identified as soil-related and 40 were classified as affected by soil management, while 23 classes were part of both subsets. Furthermore, reconciling the contributions of soils to human well-being with the CICES classification is hampered by CICES's dualistic conception of biotic ecosystem services, e.g. biomass provision or carbon storage, and abiotic ecosystem services, e.g. groundwater for drinking or control of liquid flows. Soil-related services are often relevant for both categories as they are linked in the interaction zone between pedosphere, hydrosphere, atmosphere and biosphere (Paul et al., 2021). Additionally, the linear relationship between soil functions and soil-based ecosystem services suggested by the cascade model does not necessarily hold. Thus, in the literature definitions of soil-based ecosystem services correspond to some extent to definitions of soil functions (Baveye et al., 2016), but most definitions of soil functions arguably do not have a sufficiently clear link to human well-being. For example, the soil function 'water storage' is highly relevant for the soil-based ecosystem service 'flood protection,' but in the absence of further factors (proximity to human settlements, rivers or streams as source of flooding etc.), the function will not become an ecosystem service.

Collective goods and property rights regimes

Many of the soil-based ecosystem services mentioned in the previous section are public or collective goods. This means that others cannot be excluded from enjoying the benefits of the provision of these ecosystem services. For instance, while increasing the organic carbon content in soils has numerous benefits to the land users themselves (e.g. in terms of soil moisture and productivity), carbon sequestration and storage in soils also removes greenhouse gases from the atmosphere, thus contributing to the provision of a global collective good (a stable climate) enjoyed by every person alive as well as (and especially) future generations. However, while the land user can reap the private benefits of soil management, outweighing the associated costs of management changes, this is not the case for collective goods. Here, the land user as provider of a collective good bears the entire cost of its provision, in particular the opportunity costs resulting from not being able to manage differently, but they cannot reap the benefits, unless some institution has been set up to remunerate them.

The challenge posed by soils in this context is that they are multifunctional, and the private-good and collective-good aspects cannot be easily separated. Building upon Lancaster's consumer theory (Lancaster, 1966), which posits that consumers are interested in the characteristics of a good (e.g. taste, size, color in the case of an apple) rather than in the good as such, soils can be considered a bundle of characteristics (soil functions and soil-based ecosystem services), which contribute to human well-being in different ways (Bartkowski et al., 2018).

This has profound consequences for the conflict between *de facto* and *de jure* property rights toward soils. While in legal terms, soils are part of 'land,' which can be owned privately and bought and sold in the market, *de facto*, many of the ecosystem services provided by them are collective goods that defy private ownership and marketability. Moreover, private and collective goods are not a binary category, but rather a spectrum. To analyze this spectrum and locate various soil-based ecosystem services within it, it is useful to distinguish a set of actions that can be executed toward any economic good (Schlager and Ostrom, 1992). These actions are:

- Access, i.e., the physical interaction with land/soils;
- Withdrawal, i.e., enjoyment of the 'fruits' provided by land/soils;
- Management, i.e., modifying and regulating land/soils and their properties;
- Exclusion, i.e., preventing others from access, withdrawal, and/or management;
- Alienation, i.e., transferring the land to another person or entity (by selling or giving away).

Different persons or entities can possess different *de facto* and *de jure* rights toward the same good, i.e. control different actions related to this good. For instance, while the tenant may possess the exclusive right to cultivate a piece of land for a contractually defined period (access, withdrawal, management, exclusion), the land owner retains the right to sell the property in question (alienation) (Bartkowski et al., 2018). The same intuition holds for various soil-based ecosystem services, as exemplarily shown in Fig. 2. In particular, almost everyone can enjoy the benefits of collective-good ecosystem services (e.g. climate regulation or biodiversity) and no-one can be effectively excluded from enjoying them. In the absence of societal institutions that will address this problem, the land user does not have reasons to bear the cost of managing the soil to provide those soil-based ecosystem services that exhibit collective good properties. The objectives of management will thus be restricted to a subset of ecosystem services that directly affects the interests and well-being of the land user, thus leading to a social dilemma or collective-goods problem.

It is impractical to assign distinct legal property rights to each soil-based ecosystem service, and to define the actions permitted for each affected individual, entity or group. However, given how legal property rights are commonly assigned (their object being simply the land), the observation that *de facto* property rights of third parties are also affected creates a rationale for state intervention in the form of institutions (policy instruments) that provide incentives to land users to align their management with the interests of the affected parties. Thus, the social dilemma can potentially be resolved. However, before this can be done, it is first necessary to determine who is affected and how they are affected, i.e. what are the public preferences for various soil-based ecosystem services.

Economic valuation of soil-based ecosystem services

The primary motive for economic valuation of the environment, e.g. of soils, is to demonstrate the importance of ecosystem services for human well-being and the value these have for society. The focus is particularly on those ecosystem services that are not traded on markets and therefore do not have prices (nonmarket valuation). Further, economic valuation allows for the expression of trade-offs and assessment of the impact of alternative actions on human welfare (TEEB, 2010). A heuristic to assess the aggregate value of benefits associated with soil-based ecosystem services is the so-called Total Economic Value (TEV) framework. 'Total' refers here to the aggregate of different economic, i.e. preference-based value types and categories (see Fig. 3) and should not be confused with the absolute value derived by an ecosystem service. Usually, economic valuation of ecosystem services captures marginal values, i.e. the value of a small incremental change in the quantity or quality of the ecosystem service, which depends on the relative scarcity of the ecosystem service and, thus, its current levels of provision. Higher marginal value is associated with increasing scarcity in ecosystem service supply – below a minimal threshold, sometimes called 'critical natural capital,' when the ecosystem service supply becomes non-substitutable and, thus, essential to human survival, economic valuation ceases to be meaningful (Farley, 2008). Therefore, the aim is not to estimate the absolute value derived from an ecosystem service, but the marginal value associated

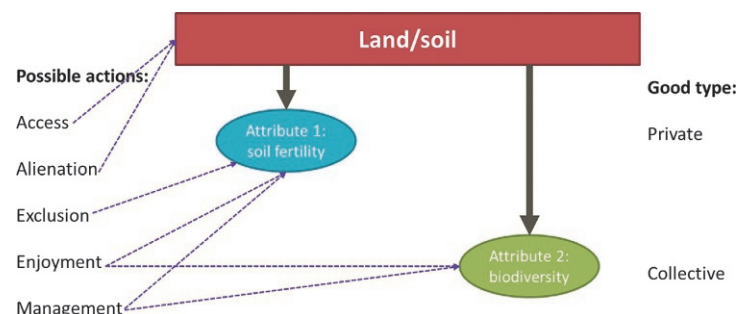


Fig. 2 Stylized relationships between different types of attributes of land/soil and various actions covered by *de facto* and *de jure* property rights. Bartkowski B and Bartke S (2018) Leverage points for governing agricultural soils: A review of empirical studies of European farmers' decision-making. *Sustainability* 10, 3179. doi: 10.3390/su10093179 reproduced by permission; licensed CC BY 4.0.

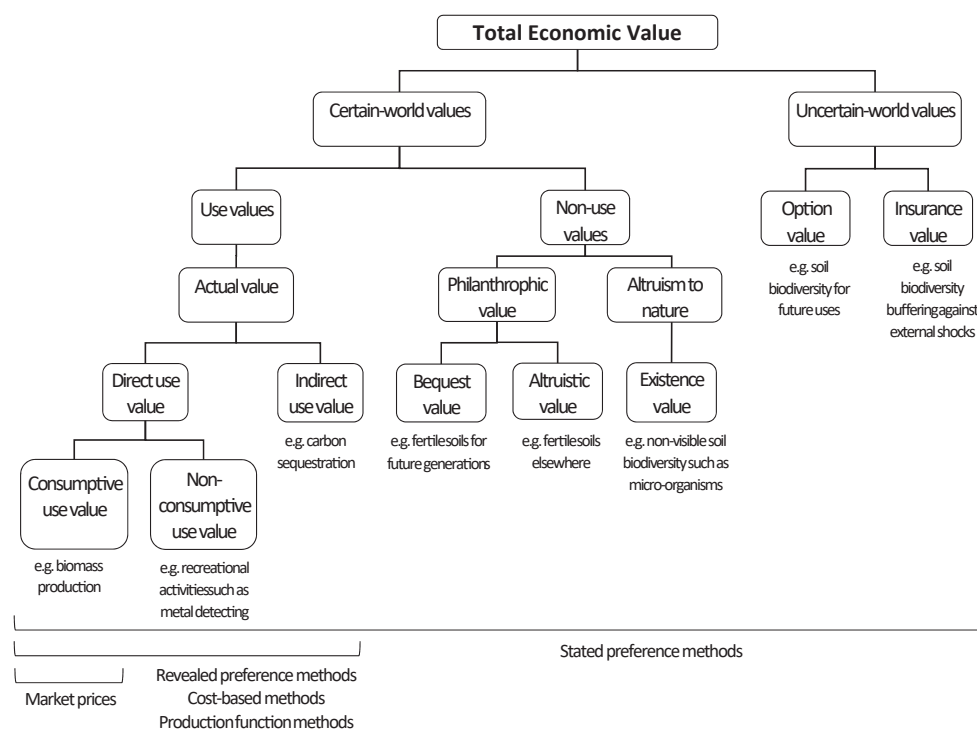


Fig. 3 The Total Economic Value framework with soil-relevant examples. Own illustration based on Modified from Bartkowski B, Bartke S, Helming K, Paul C, Techen A-K, and Hansjürgens B (2020) Potential of the economic valuation of soil-based ecosystem services to inform sustainable soil management and policy. *PeerJ* 8: e8749. doi: 10.7717/peerj.8749.

with a change in the state of the world. For example, a decision on whether to re-wet a peatland area would be informed by the marginal changes in ecosystem service provision with respect to the various value categories of the TEV, described in the following.

The various value categories comprise two overarching types of value: 'certain-world values' and 'uncertain-world values.' Certain-world values are independent of the uncertainty regarding future states of the world, whereas uncertain-world values derive from the inherent uncertainty over future demand and supply of soil-based ecosystem services (Bartkowski et al., 2020). Certain-world values can be divided into 'use values' and 'non-use values.' As suggested by the name, use values refer to values derived from the use of environmental goods which may be 'direct' and 'consumptive' (e.g. biomass production), direct and 'non-consumptive' (e.g. recreational activities such as metal detecting) or 'indirect' or 'passive' (e.g. climate regulation through carbon storage). Direct use values are often, though not always, reflected in market transactions, whereas indirect use values generally are not. The concept of non-use values captures all values of ecosystems that either relate to altruism toward other humans (philanthropic value) or toward nature (existence value). Philanthropic value may refer to the satisfaction of knowing that future generations will have access to benefits from a given (bundle of) ecosystem services or an ecosystem (bequest value) or that other people currently living have access to such benefits (altruistic value). Satisfaction derived from the knowledge that an ecosystem or species simply exists is referred to as 'existence value.' Despite the fact that non-use values are associated with other humans or non-human entities, these types of value are usually understood in terms of individual preferences. In other words, individual measures of value are of relevance with respect to individual satisfaction, e.g. of knowing that soils will provide benefits for future generations. Turning to the notion of uncertain-world values, the associated values account for the temporal dimension of soils and the provision of soil-based ecosystem services. The so-called 'option value' is associated with the continued ability to respond to future preferences and demand, i.e. it focuses on the preservation of options to use soils in the future and to enjoy the potential benefits of soil-based ecosystem services. 'Insurance value' arises from soils' potential to reduce variance in soil-based ecosystem service supply, in particular with regard to disturbances, and thus to supply uncertainty. Both types of value are strongly linked to biodiversity (Bartkowski et al., 2020; Pascual et al., 2015).

A variety of valuation methods can be used to estimate the TEV of ecosystems or ecosystem services. The choice of method is highly context specific, and there are multiple conflicting factors to consider, including the firmness of the theoretical underpinning of a method, the cost of its application, the associated data demands, flexibility and comprehensiveness. The primary goal of each nonmarket valuation method is the inference of preference information from proxies, mostly information about choice behavior. Two common options are the elicitation of preferences for hypothetical scenarios with questionnaires ('stated preference methods') and observation of individual market behavior that can be linked to ecosystem services ('revealed preference methods'). Among stated preference methods, one may distinguish contingent valuation, which is based on a direct question about respondents'

willingness to pay for a hypothetical change in the provision of an ecosystem service, and discrete choice experiments, which are based on respondents choosing among scenarios that are characterized by variable attributes (e.g. soil-based ecosystem services). Revealed preference methods include the travel cost method (inference of the value of an ecosystem, a landscape etc. from expenditures people incur to visit them) and hedonic pricing (analysis of market data – e.g. real estate or land markets – where price variations can be partly explained by spatially explicit provision of ecosystem services). Stated preference methods are the most comprehensive and flexible, but rather effort- and cost-intensive. Furthermore, they are based on the analysis of hypothetical choices, which may be biased and not reflect how people would behave in the real world. Revealed preference methods reflect actual choice behavior, but they can only be applied to a limited subset of ecosystem services, require substantial amounts of data and assume that people are aware of the benefits they derive from ecosystems. In some cases, the data demands and costs associated with preference-based valuation methods may be prohibitive. In such cases, more crude yet easier to apply methods are sometimes used, such as those based on market prices of comparable goods ('market price-based method'), the costs of replacement, preservation or restoration of the ecosystem service in question ('cost-based methods') or the estimation of production functions that link the production of market goods to ecosystem inputs ('production function method').

In the context of soils, cost-based methods or market price proxies are applied most often to estimate the economic value of soil-based ecosystem services, instead of state-of-the-art preference-based valuation methods (Bartkowski et al., 2020). For instance, in a particularly comprehensive study, Dominati et al. (2014), use replacement costs and market price proxies to calculate the economic values of a range of soil-based ecosystem services (including food provision, flood protection, nutrient cycling and climate regulation). However, cost-based methods are problematic, as they do not provide information about preferences for collective goods, but merely give insights about costs of technical solutions. These costs may be an inadequate proxy for values if public preferences differ significantly. In addition, market prices can only be considered rough proxies because, as discussed above, most environmental goods are non-marketable or rather the associated marketed goods are not equivalent to the soil-based ecosystem services. Furthermore, existing economic valuation studies consider only a minor part of the range of soil-based ecosystem services discussed above. Primarily, economic values of climate regulation (carbon storage), decomposition (nutrient cycling) and (cost of) soil erosion have been investigated in the empirical literature. Thus, the potential of economic valuation to inform decision-making has not yet been fully realized. Hence, there is a need to consider the whole range of soil-based ecosystem services including trade-offs among them and for the application of state-of-the-art methods (Bartkowski et al., 2020).

Incentive-based instruments for soil protection

As discussed above, soils provide multiple ecosystem services that are valued by the public and exhibit collective good characteristics. The latter implies that their provision is unlikely to be ascertained by markets, given conventional legal property rights regimes. Rather, a corrective intervention by the state is required. However, the spatial heterogeneity of soils suggests that the required corrections in management will vary across locations; also, their effects may manifest at different spatial and temporal scales. Furthermore, the demand for ecosystem services not only depends on the local potential of soils to provide them, but also on the availability of substitutes (e.g. flood protection through infiltration and water storage in soils can be substituted by flood protection walls), the characteristics of the local population (e.g. in terms of income and wealth or attitudes) and the opportunity costs of managing for a given ecosystem service (for instance, in highly productive areas, the opportunity costs of extensification are high). Therefore, in most cases, a mix of different policy instruments will be required to address soil multifunctionality in heterogeneous landscapes (Bartkowski et al., 2021a). Relatively rigid policy instruments such as legal standards or zoning are unlikely to be efficient on their own and should at least be complemented by incentive-based instruments. In some cases, overly rigid policies (e.g. bans of particular practices) may even backfire – e.g., a glyphosate ban has been shown to entail increasing pressure on soils due to a shift to mechanical weed control (Böcker et al., 2020).

The main challenge in crafting effective and efficient policy instrument mixes is that soil protection and the provision of soil-based ecosystem services constitute a so-called principal–agent problem. Land users are the agents who are supposed to manage the land and soil in accordance with the requirements and incentives provided by the principal, i.e. the state or, more broadly speaking, society. However, this principal–agent relationship is hampered by information asymmetry between the state (represented by regulatory agencies) and land users. This asymmetry is especially relevant in cases where the effectiveness of management changes is likely to depend on land users' local, often tacit, knowledge including that about (opportunity) costs. Further, the regulatory agency has only imperfect capability to monitor the land user's compliance with requirements and regulations. Both types of information asymmetry impose two major challenges, which are called 'adverse selection' and 'moral hazard,' respectively. In the context of soil protection, adverse selection particularly refers to where soil-friendly management is taking place: where it is most effective (i.e. according to the interest of the principal) or where it is cheapest (i.e. according to the interest of the agent). Properly designed instruments help to avoid adverse selection and to incentivize the implementation of protection measures where they are most effective, e.g. by offering payments for measured improvements in soil condition, rather than simply for the adoption of specified practices (Bartkowski, 2021). In contrast to adverse selection, moral hazard becomes relevant after the contractual agreement (or, analogously, after introduction of binding management standards). The regulatory agency is then unable to monitor perfectly the compliance by the land user, who has an incentive to manage in a way that will maximize their profit, rather than in the agreed/mandated, soil-friendly manner. In this context, digitalization and remote sensing are increasingly important technologies that reduce the costs of monitoring of compliance (Ehlers et al., 2021).

A third major challenge related to information asymmetry is called 'additionality.' Ex ante, the regulatory agency usually does not know what the land user will do in the absence of incentives for soil protection. Given that the provision of incentives is costly for the agency, it is rational to provide them only if the land user would not have changed to soil-friendly management in their absence. Otherwise, the land user receives a windfall profit for an activity they would have undertaken anyway, and the scarce funds used to fund non-additional behavior cannot be used to finance the provision of collective goods elsewhere. However, the land user's future decisions are affected by many factors (Bartkowski and Bartke, 2018), and the counterfactual case of no incentives is difficult to grasp, especially for longer-term problems. In some cases, it may therefore be rational (from the point of view of the regulating agency) to provide incentives for non-additional behavior to prevent a return to the undesirable earlier behavior when circumstances change (e.g. when prices for relevant inputs or outputs change, making opportunity costs of soil protection higher).

These three major challenges should be considered when evaluating the effectiveness and efficiency of policy instruments, including incentive-based instruments. The latter include payments, taxes and (tradable) offsets. Each of these can be found in the context of soil protection: for instance, in agriculture, agri-environmental payments are a common instrument to align soil management with societal preferences (Bartkowski et al., 2021b); taxes have been used and discussed especially as a means to reduce pesticide use (Möhring et al., 2020); tradable offsets have been discussed and implemented intensively with respect to initiatives aiming to increase the soil organic carbon content as a strategy for climate change mitigation (negative emissions) (Thamo and Pannell, 2016).

Agri-environmental payments are the most widespread incentive-based instrument and a specific form of 'payments for ecosystem services.' Here, an important distinction is between action-based and result-based variants. In the case of action-based payments, the payment is not directly linked to the environmental outcome that is supposed to be enhanced or protected. For instance, rather than paying for changes in soil organic carbon, the regulatory agency may offer payments for the adoption of practices that are assumed to be beneficial in terms of carbon sequestration and storage (e.g. non-inversion tillage). In theory, this approach would only be efficient if the regulatory agency had perfect, spatially explicit, ex-ante knowledge about the effects of the regulated activity. In practice, action-based payments are quite common due to the relatively low requirements in terms of measurement and monitoring. Result-based payments are the theoretically preferable alternative in the face of information asymmetries between the regulatory agency and the land users (Burton and Schwarz, 2013). Here, the payments are issued for ex-post measured changes in the environmental good. This approach is more efficient than action-based payments, but it poses high demands in terms of measurement and monitoring. Furthermore, it implies greater risk for land users, as they cannot be certain that their actions will lead to the envisioned results, thus securing them compensation payments (Derissen and Quaas, 2013). Within the European Union's Common Agricultural Policy, one of the largest agricultural and agri-environmental policy frameworks worldwide, none of the few result-based payment schemes is related to soils. An alternative, hybrid approach consists in offering payments on the basis of model predictions regarding the spatially explicit effects of management changes (Bartkowski et al., 2021b). A similar approach has already been tried out in the context of private soil carbon offset schemes (TerraCarbon, Indigo Ag, 2020) as a lower-cost alternative to regular, expensive measurements of changes in soil organic carbon. In the future, further developments in remote sensing technologies could also help address the challenge of prohibitive measurement/monitoring costs (Ehlers et al., 2021).

Taxes are the opposite of payments and have been used less intensively in the context of soil protection, mainly due to the difficulties of monitoring that is required for effective environmental taxes. They are primarily relevant when it comes to making problematic inputs (e.g. synthetic pesticides) more expensive (Möhring et al., 2020). In the context of soils, a slightly more widespread approach has been tradeable offsets. These have been used mainly in the context of nonpoint-source pollution such as nitrate (Stephenson and Shabman, 2017) and in the context of soil carbon (Thamo and Pannell, 2016). However, the former is associated with rather high transaction costs (i.e. costs of setting up schemes, monitoring, organizing a trade market etc.) and its effectiveness has been questioned (Stephenson and Shabman, 2017), while the latter has faced a major obstacle in the need to establish permanence (Thamo and Pannell, 2016), which is required for a real climate effect of soil carbon storage.

Overall, the design of effective and efficient soil protection policy mixes is quite challenging and requires the context-specific consideration of numerous unintended effects such as adverse selection or non-additionality. The increase in understanding of the complexity of land users' decision-making processes (Bartkowski and Bartke, 2018) adds to the challenge by demonstrating the inadequacy of existing instruments. However, effective and efficient policy mixes are essential to align the management of soils with the public preferences for soil-based ecosystem services.

Conclusions

Soils are scarce, both in terms of quantity and, especially, quality. They provide multiple benefits to humans, in the form of soil-based ecosystem services. However, this multifunctionality is characterized by trade-offs. Many of the ecosystem services provided by soils are collective goods, which implies that they will be underprovided in the absence of additional incentives. The extent of underprovision depends on the public preferences for soil-based ecosystem services, which can be estimated by means of economic nonmarket valuation. This can then inform the design of policy instruments required to align soil management with public preferences and to optimize soil management so as to enhance the multifunctionality.

To protect soils effectively and efficiently, it is necessary to understand the relevant economic effects adequately. For instance, soils are heterogeneous; therefore, their potential to provide ecosystem services, the opportunity costs of their protection as well as

the most effective management practices vary spatially. Economic valuation can help illuminate the extent of public preferences for non-private soil-based ecosystem services, but not all valuation methods are equally adequate and precise; any number may not be significantly better than no number at all. Furthermore, when designing policy instruments, challenges that arise from information asymmetry in the principal-agent setting need to be properly taken into account to avoid wasting scarce resources that could otherwise be used to finance the provision of other public goods.

References

- Bartkowski B (2021) Don't throw efficiency out with the bathwater: A reply to Jeffery and Verheijen (2020). *Environmental Science & Policy* 122: 72–74. <https://doi.org/10.1016/j.envsci.2021.04.011>.
- Bartkowski B and Bartke S (2018) Leverage points for governing agricultural soils: A review of empirical studies of European Farmers' Decision-Making. *Sustainability* 10: 3179. <https://doi.org/10.3390/su10093179>.
- Bartkowski B, Hansjürgens B, Möckel S, and Bartke S (2018) Institutional economics of agricultural soil ecosystem services. *Sustainability* 10: 2447. <https://doi.org/10.3390/su10072447>.
- Bartkowski B, Bartke S, Helming K, Paul C, Techen A-K, and Hansjürgens B (2020) Potential of the economic valuation of soil-based ecosystem services to inform sustainable soil management and policy. *PeerJ* 8: e8749. <https://doi.org/10.7717/peerj.8749>.
- Bartkowski B, Bartke S, Hagemann N, Hansjürgens B, and Schröter-Schlaack C (2021a) Application of the governance disruptions framework to German agricultural soil policy. *SOIL Discussions* 1–22. <https://doi.org/10.5194/soil-2021-16>.
- Bartkowski B, Droste N, Ließ M, Sidemo-Holm W, Weller U, and Brady MV (2021b) Payments by modelled results: A novel design for agri-environmental schemes. *Land Use Policy* 102: 105230. <https://doi.org/10.1016/j.landusepol.2020.105230>.
- Baveye PC, Baveye J, and Gowdy J (2016) Soil "ecosystem" services and natural capital: Critical appraisal of research on uncertain ground. *Frontiers in Environmental Science* 4. <https://doi.org/10.3389/fenvs.2016.00041>.
- Böcker T, Britz W, Möhring N, and Finger R (2020) An economic and environmental assessment of a glyphosate ban for the example of maize production. *European Review of Agricultural Economics* 47: 371–402. <https://doi.org/10.1093/erae/fby050>.
- Burton RJF and Schwarz G (2013) Result-oriented agri-environmental schemes in Europe and their potential for promoting behavioural change. *Land Use Policy* 30: 628–641. <https://doi.org/10.1016/j.landusepol.2012.05.002>.
- Derissen S and Quaas MF (2013) Combining performance-based and action-based payments to provide environmental goods under uncertainty. *Ecological Economics, New Climate Economics* 85: 77–84. <https://doi.org/10.1016/j.ecolecon.2012.11.001>.
- Dominati EJ, Patterson M, and Mackay A (2010) A framework for classifying and quantifying the natural capital and ecosystem services of soils. *Ecological Economics* 69: 1858–1868. <https://doi.org/10.1016/j.ecolecon.2010.05.002>.
- Dominati EJ, Mackay A, Green S, and Patterson M (2014) A soil change-based methodology for the quantification and valuation of ecosystem services from agro-ecosystems: A case study of pastoral agriculture in New Zealand. *Ecological Economics* 100: 119–129. <https://doi.org/10.1016/j.ecolecon.2014.02.008>.
- Ehlers M-H, Huber R, and Finger R (2021) Agricultural policy in the era of digitalisation. *Food Policy* 100: 102019. <https://doi.org/10.1016/j.foodpol.2020.102019>.
- Farley J (2008) The role of prices in conserving critical natural capital. *Conservation Biology* 22: 1399–1408. <https://doi.org/10.1111/j.1523-1739.2008.01090.x>.
- Haines-Young R and Potschin M (2010) The links between biodiversity, ecosystem services and human well-being. In: Raffaelli DG and Frid CLJ (eds.) *Ecosystem Ecology: A New Synthesis*, pp. 110–139. Cambridge: Cambridge University Press.
- Haines-Young R and Potschin M (2018) *Common International Classification of Ecosystem Services (CICES) V5.1: Guidance on the Application of the Revised Structure*.
- Lancaster KJ (1966) A new approach to consumer theory. *Journal of Political Economy* 74: 132–157. <https://doi.org/10.1086/259131>.
- Möhring N, Ingold K, Kudsk P, Martin-Laurent F, Niggli U, Siegrist M, Studer B, Walter A, and Finger R (2020) Pathways for advancing pesticide policies. *Nature Food* 1: 535–540. <https://doi.org/10.1038/s43016-020-00141-4>.
- Ostrom E (1990) *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.
- Pascual U, Tormansen M, Hedlund K, Brussaard L, Faber JH, Foudi S, Lemanceau P, and Jørgensen SL (2015) On the value of soil biodiversity and ecosystem services. *Ecosystem Services* 15: 11–18. <https://doi.org/10.1016/j.ecoser.2015.06.002>.
- Paul C, Kuhn K, Steinhoff-Knopp B, Weißhuhn P, and Helming K (2021) Towards a standardization of soil-related ecosystem service assessments. *European Journal of Soil Science* 72: 1543–1558. <https://doi.org/10.1111/ejss.13022>.
- Schlager E and Ostrom E (1992) Property-rights regimes and natural resources: A conceptual analysis. *Land Economics* 68: 249–262. <https://doi.org/10.2307/3146375>.
- Stephenson K and Shabman L (2017) Can water quality trading fix the agricultural nonpoint source problem? *Annual Review of Resource Economics* 9: 95–116. <https://doi.org/10.1146/annurev-resource-100516-053639>.
- TEEB (2010) *The Economics of Ecosystems and Biodiversity: Ecological and Economic Foundations*. London and Washington: Earthscan.
- TerraCarbon, Indigo Ag (2020) *Methodology for Improved Agricultural Land Management Version 1.0. Verified Carbon Standard*.
- Thamo T and Pannell DJ (2016) Challenges in developing effective policy for soil carbon sequestration: Perspectives on additionality, leakage, and permanence. *Climate Policy* 16: 973–992. <https://doi.org/10.1080/14693062.2015.1075372>.

Soil and human health

Margaret A Oliver^a and Eric C Brevik^b, ^aSoil Research Centre, Department of Geography & Environmental Science, School of Human & Environmental Sciences, The University of Reading, Reading, United Kingdom; ^bCollege of Agricultural, Life, and Physical Sciences, Southern Illinois University, Carbondale, IL, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	555
History of soil and human health	556
Soil and food security	557
Nutrition and selected trace elements	559
Iron	559
Iodine	561
Selenium	561
Zinc	562
Effects of soil entering the human body	562
Ingestion	562
Inhalation	563
Soil absorption through skin lesions	563
Soil organisms	564
Effects of soil pollutants – metals, organic chemicals, and radionuclides	566
Medicines from soil	567
Antibiotics	567
Clays	569
The future	570
References	571

Abstract

The links between soil and human health are many; they include essential nutrients, food security, exposure to pollutants and pathogens, beneficial organisms, medications, and clean water. Ancient civilizations were aware of the influence of soil on human health, but scientific study of the relations is more recent. Human health benefits from the biological, chemical, and physical properties and processes in soil. We embrace the following aspects of the relations: history, soil and food security, nutrition and four important trace elements, effects of soil entering the body, pollutant effects of metals, organic chemicals, radionuclides, and soil organisms, and medicines from soil.

Key points

- Some relations between soil and human health were known in ancient history
- Soil and food security are under increasing pressure because of growth in the world's population with potentially serious consequences for human health
- Trace elements from soil play a vital role in human health; we focus on iodine (I), iron (Fe), selenium (Se) and zinc (Zn)
- Soil can enter the human body by ingestion, inhalation or through skin lesions and can affect human health
- Some metal elements, organic chemicals, radionuclides and organisms can have a deleterious effect on human health
- Soil has been an important source of microorganisms that have provided medicines such as antibiotics.

Introduction

There is a multitude of links between soil and human health, including the supply of essential nutrients through the soil–plant–(animal)–food system, food security, exposure to pollutants and pathogens, the effects of beneficial organisms, production of medications, and purification of water. All the human health benefits supplied by soil take advantage of various aspects of the biological, chemical, and or physical properties of soil.

Recognition that soil influences human health goes back thousands of years, but the scientific study of the links between soil and human health is much more recent. There has also been a major increase in interest concerning the relations between soil and human health in recent years, with identifiable trends that aligned with major events. For example, a large amount of attention was focused on radionuclides deposited on the soil and entering the food chain after the testing of atomic bombs in the 1950s and 1960s and following accidents at the Chernobyl and Fukushima nuclear plants. Increased attention has also been given to organic

chemicals and 'heavy' metals as the environmental movement gained momentum in the 1970s and 1980s and this continues to the present (Brevik and Sauer, 2015).

The goal of this chapter is to provide a brief, relatively comprehensive overview of the connections between soil and human health, which will serve as a good primer for those who want an introduction to the topic. A number of recent review articles have discussed various aspects of the soil and human health relationship, providing broad overviews of the topic (e.g. Oliver and Gregory, 2015; Brevik et al., 2020), links to ecosystem services (e.g., Keith et al., 2016), and the influence of soil health on human health (Kemper and Lal, 2017). Reviews such as these, and the numerous references cited in them and in this chapter, can provide additional information as needed.

History of soil and human health

People have been aware for a long time that the environment affects their health either directly or indirectly through their food and drinking water, and the air they breathe. Several early civilizations referred to relations between the environment and human health; they drew their inferences from repeated observations of conditions rather than from scientific investigation. Scientific study of the role of the soil on animal and human health is little more than 100–200 years old, but the notion that soil has such a role goes back to at least 1400–1500 BCE (Brevik and Sauer, 2015). Around 1550 BCE an Egyptian physician collected medical facts about many health issues that he recorded on papyrus (now known as the Ebers Papyrus, 1500 BCE), which was later sealed in a tomb near Thebes and obtained by Georg Ebers in 1872. The Bible also depicts Moses as understanding that fertile soil was essential to the well-being of his people in about 1400 BCE as he entered Canaan after leaving Egypt (Numbers 12, 18–20) (Brevik and Burgess, 2013). Despite such early observations, most of these effects remained unexplained until the twentieth century, and some still remain so.

Archeologists and medical historians have identified evidence of poor health from tissues and mummies from prehistoric times that can be linked to environmental conditions (see Davies et al., in Selinus et al., 2013). For example, goiter was prevalent in ancient China, Greece, Egypt and later the Inca state of Peru. It was often treated with seaweed, although people at the time did not know why the treatment was effective. The practice was based on observations and on trial and error. Archeologists have also noted relations between health and the environment from the analysis of bone, teeth, diet and the nutritional status of prehistoric humans and animals (Szostek and Glab, 2001). For example, the change from a hunter–gatherer society to an agricultural-based economy led to major dietary change and to evidence of iron (Fe) deficiency, i.e. iron deficiency anemia (Roberts and Manchester, 2007). This is because iron in plants is more difficult to absorb than that from meat.

The ancient Greeks can be regarded as more or less founding the basic concepts of modern medicine which they based on ideas from Egypt. The work of Hippocrates (a Greek physician, ca 460 BCE–377 BCE) and his fellow workers resulted in the recording of many scientific facts, probably beginning the tradition of studying the causes of diseases rather than simply observing the symptoms when prescribing a cure. Hippocrates noted in his treatise *Airs, Waters, Places* (see Brevik and Sauer, 2015) that water comes from soil, which further links with the list provided for consideration in a proper medical evaluation that included the ground: "... whether it be naked and deficient in water, or wooded and well-watered, and whether it lies in a hollow, confined situation, or is elevated and cold...". In this book he also mentions the soil in relation to the environment six times.

Hippocrates and his co-workers recognized that some diseases were always present in a given population, i.e. 'endemic,' whereas others that were not always present but which occurred in greater frequency at certain times, they called 'epidemic.' The introductory paragraphs of Hippocrates's *Airs, Waters, Places* summarize key factors such as climate, quality of the soil, water, way of life and nutrition (Rosen and Rosen, 1993). For more than 2000 years, this was the basic epidemiological text that provided the theoretical underpinning for understanding endemic and epidemic diseases. No fundamental change occurred until the late nineteenth century when new sciences, such as bacteriology and immunology, emerged (Tountas, 2009). Ideas from the era of Hippocrates about the environmental determinants of health are still relevant.

From the seventeenth to the nineteenth centuries there was increasing awareness of the links between the soil and human health, but until the middle of the twentieth century much of this was anecdotal (references to many points in this paragraph are given in Brevik and Sauer, 2015). Nevertheless, in 1876 Robert Koch determined that anthrax was caused by the soil-borne pathogen *Bacillus anthracis*. There were several other developments toward the end of the nineteenth century to support the effects of soil on human health such as *Mycobacterium leprae* (leprosy; 1873), *Clostridium tetani* (tetanus; 1884) and *Escherichia coli* (1885), for example. In the early 1920s experiments were designed to investigate relationships between soil and health. *Medical Testament*, published in the United Kingdom in 1939, indicated that illness in the UK was increasing because of poor nutrition related to agricultural practices that depleted nutrients in the soil. At the same time (1938) the United States Department of Agriculture (USDA) *Yearbook of Agriculture* stated the importance of minerals from the soil in food on the well-being of animals and humans. The isolation of antibiotics from soil bacteria by Selman Waksman in the 1940s was a major breakthrough (see below). In 1946 Eve Balfour, who studied agriculture at Reading University College during WWI, established the Soil Association arising from her understanding of the relations between soil and health. In 1959 André Voisin's book showed how this topic was gaining momentum and that the medical profession should give more attention to deficiencies in the soil. Toward the end of the twentieth century research was increasing to provide greater understanding of the role of soil in several types of human illness. Nevertheless, major gaps in knowledge remain and the need for focused research continues.

The early 21st century has seen a growth in interdisciplinary research which is fundamental for progress in our knowledge of the relations between soil and human health. However, direct knowledge of these relations remains difficult to obtain by empirical means. With the development in genomic methods greater understanding of some aspects of the relations are possible. With safer and more nutritious food we should need less medical care and have a better quality of life.

Soil and food security

The concept of food security developed over several decades from the 1970s, leading to the most commonly used current definition: the state when “all people, at all times, have physical and economic access to sufficient, safe, and nutritious food to meet their dietary needs and food preferences for an active and healthy life” (FAO, 1996). Food security is complex because it involves many factors and systems. The availability of food depends on suitable land and soil for agriculture, natural conditions, and socio-economic conditions, all of which have multiple facets (Fig. 1).

To achieve food security for all people is a huge challenge with an increasing population (1950 2.5 billion; 2021 7.9 billion), changes in eating habits, social balances and land ownership, climate change and increasing degradation of the soil. Over the past 50 years food production has largely kept pace with demand, but there remain approximately 2 billion people worldwide in moderate to severe food insecurity. This has two facets, the lack of sufficient food and of certain essential elements. The former is widespread in the developing countries of Africa, Asia and parts of Latin America, but the latter can affect both developed and developing countries. In addition, the COVID-19 pandemic of 2020–2022 has increased food insecurity dramatically in almost every country.

Insufficient food leads directly to ‘protein–energy’ malnutrition caused by a lack of carbohydrates, proteins and fats in the diet; there is also a deficit in all major macronutrients and certain micronutrients. Some causes of food security are linked to under-nutrition and are related to the soil or agriculture, whereas others are social and political (Table 1). Under-nutrition leads to stunted growth, poor health, disability, wasting of muscle and fat reserves, and an increased risk of diabetes and cardiovascular disease in adulthood. It impairs the immune system and enhances the effects of toxic elements on humans. Food security is also related to poverty because people cannot afford to buy food even if it is available. There are strong social forces at play in such situations.

The soil affects food availability and its quality: it is the basis for 98.8% of the food consumed by humans. The concept of food security is well established, but that of soil security is in its infancy. Soil is the means to provide food, fiber and biofuels, but this must be done sustainably. Soil is a fragile and essentially non-renewable resource that has suffered degradation in many parts of the world. Increasing demands on the soil and degradation have sometimes led to an unstable system in which the soil becomes more vulnerable to threats and less able to perform its functions.

Inherent soil fertility embraces nutrient availability, stable porous structure, large available water content and diverse microbial and faunal communities that facilitate root and shoot growth. These can be maintained or improved by sustainable management practices, including modern technologies. The use of crop rotations, cover crops, crop residues, animal and green manures and other organic wastes can improve soil fertility, soil organic matter (SOM) content and mitigate soil degradation. Organic farming alone cannot resolve issues of food security because of its generally smaller yields, therefore, integrated management should be considered that includes the use of organic and mineral fertilizers to maintain both human and soil health.

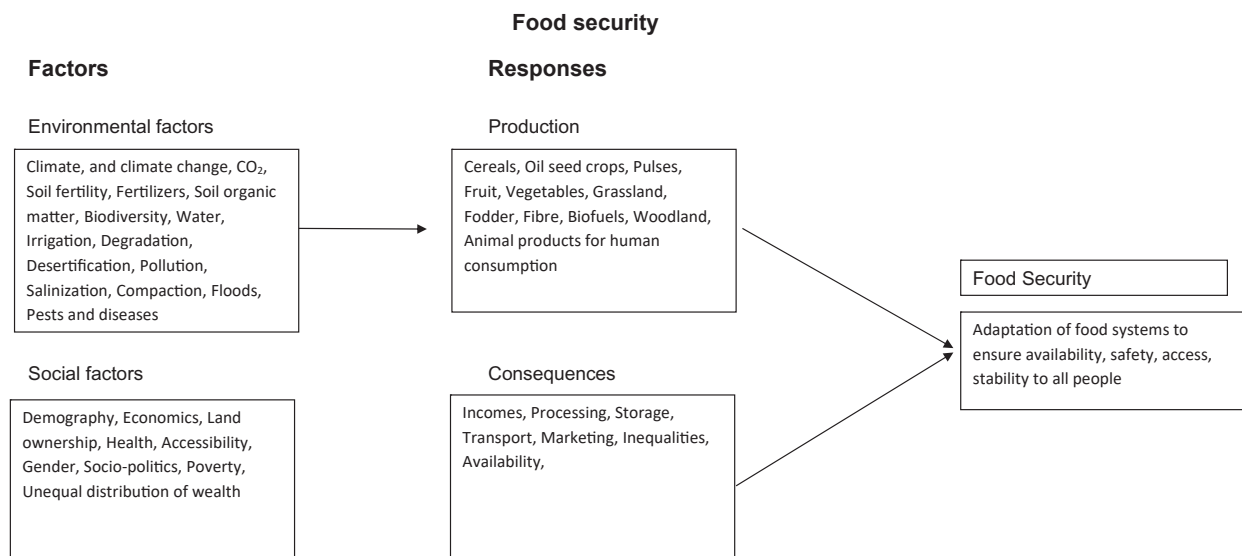


Fig. 1 The factors and systems that link to form the complex web of food security.

Table 1 Essential mineral nutrients (in addition to carbon, hydrogen and oxygen) in plants and mammals, together with their recommended daily allowance (RDA) or adequate intake (AI – where no RDA has been established) in humans, function in mammals and common sources in the human diet.

Element	Essentiality		RDA or AI/mg	Function in mammals	Common sources in the human diet
	Plants	Mammals			
Primary element					
Calcium (Ca)	Yes	Yes	1300	Essential constituent of bones and required for synthesis and function of blood cells, and function of muscle, heart and digestive system	Kale, collards, mustard greens, broccoli, dairy products
Nitrogen	Yes	Yes	105	Essential constituent of DNA, RNA and proteins	Beans, lentils, beef, lamb, cheese, eggs, milk
Phosphorus (P)	Yes	Yes	700	Essential constituent of bones, cells and energy processes plus other functions	Nuts, beans, peas, lentils, grains, meats, eggs, dairy products
Potassium (K)	Yes	Yes	4700	A systemic electrolyte and essential in regulating ATP with sodium	Fruits, cereals, vegetables, beans, peas, lentils, dairy products, meats
Secondary element					
Magnesium (Mg)	Yes	Yes	420	Required for bones and processing ATP	Seeds, nuts, beans, peas, lentils, whole grains, dark green vegetables
Sulfur (S)	Yes	Yes		Constituent of amino acids – diets with sufficient protein contain adequate sulfur so there is no RDA	Bok choy, broccoli, brussels sprouts, cabbage, cauliflower, collard greens, onions, shallots, leeks, garlic
Chlorine (Cl)	Yes	Yes	2300	Pump functions of cells and hydrochloric acid in stomach	Dairy products, meats, eggs
Trace element					
Boron (B)	Yes	Yes		Has a role in some animals in vitamin D activation, but uncertain whether this is a nutritional effect. Helps to build and maintain bones and is involved in Ca and Mg metabolism	Raisins, peanuts, fruits, beans, broccoli, carrots, onions, dark green leafy vegetables
Iron (Fe)	Yes	Yes	18	Required for many proteins and enzymes particularly hemoglobin in blood	Meats, especially red meat
Manganese (Mn)	Yes	Yes	2.3	A cofactor in enzyme reactions	Whole grains, beans, peas, lentils, nuts, tea
Copper (Cu)	Yes	Yes	0.9	Required component of many redox enzymes	Beans, peas, lentils, whole grains, nuts, peanuts, mushrooms, chocolate, organ meats (offal)
Zinc (Zn)	Yes	Yes	11	Required component of many enzymes	Nuts, whole grains, beans, peas, lentils, meats, organ meats (offal)
Nickel (Ni)	Yes	Yes		Suggested role in some animals in enzyme function and iron adsorption and metabolism in humans	Dry beans, cocoa products, some nuts, wheat, shellfish, vegetables
Molybdenum (Mo)	Yes	Yes	0.045	Component of several oxidase enzymes	Beans, peas, lentils, dark green leafy vegetables, organ meats (offal)
Sodium (Na)	Beneficial	Yes	1500	A systemic electrolyte and essential in regulating ATP with potassium	Dairy products, meats, eggs
Selenium (Se)	Beneficial	Yes	0.055	An essential cofactor of antioxidant enzymes	Grain products, nuts, especially Brazil nuts, garlic, broccoli (if grown on high-Se soils), meats from Se-fed livestock, seafood
Cobalt (Co)	Beneficial	Yes		Required in the synthesis of vitamin B ₁₂ (cobalamine) which is synthesized by bacteria	Green vegetables, cereals, mushrooms
Minimum trace element					
Aluminium (Al)	Beneficial			No known function in healthy humans but known toxic effects	
Silicon (Si)	Beneficial	Possibly		Possible beneficial role in human bone health and wound healing	Cereals, pumpkins, water
Human trace element only					
Iodine (I)		Yes	0.15	Essential for the synthesis of thyroid hormones, possibly as an antioxidant and for the functioning of mammary and salivary glands	Vegetables, cereals, fruit
Fluorine (F)				Not an essential nutritional element but can help in the prevention of dental cavities	Tea, kale, leeks, lettuce, mustard, onions, peas, spinach, tomatoes, organ meats (offal), fish
Human minimum trace element					
Lithium (Li)		Possibly		Physiological function unknown but feeding studies have suggested benefits in some animals	Vegetables, drinking water
Vanadium (Va)		Possibly		Not essential, but is a cofactor that enhances or inhibits some enzymes	Parsley, leafy vegetables, asparagus, rye, black pepper, mushrooms
Chromium (Cr)		Possibly		Not essential although some studies suggest a role in sugar metabolism	Pulses, nuts, meats, seafood
Toxic trace					
Arsenic (As)		Possibly		May play a role in DNA methylation and cancer prevention, but also toxic effects	Rice, seafood
Cadmium				No known function in healthy humans, but known toxic effects	
Lead (Pb)				No known function in healthy humans, but known toxic effects	

Adapted from Table 1, Oliver MA and Gregory PJ (2015) Soil, food security and human health: A review. *European Journal of Soil Science* 66: 257–276.

During the last 50 years the global population has tripled and production of the major cereal grains (maize (*Zea mays* L.), wheat (*Triticum aestivum* L.), rice (*Oryza sativa* L.) and barley (*Hordeum vulgare* L.)) has also tripled, furthermore diets have changed to consume more meat. Fairly recent technological innovations and intensified production have had some adverse effects on the environment. For example, excessive applications of nitrogen and phosphorus have polluted fresh and coastal waters, and intensive land use and vegetation clearance have resulted in soil and land degradation. Soil organic matter has also declined, and this maintains many key soil functions that underpin fertility and resistance to erosion. This is also associated with a loss of soil biodiversity which can affect the soil's ability to perform many ecosystem functions. Some elements occur in potentially toxic concentrations that diminish crop yields and can accumulate in the food chain with detrimental consequences for human health. Salinization and desertification reduce the availability and productivity of land for food production. Compaction from the large machinery used in intensive agriculture leads to reduced yields and increased erosion. Nutrient mining leads to decreased productivity and food security. This is a major problem in the tropics. In most of Sub-Saharan Africa and Asia the soil supports only poor yields because it is either inherently infertile, over used or it requires the addition of fertilizers. Degradation further impoverishes farmers in developing nations because they lack the resources to remedy the situation, leading to greater food insecurity, under-nutrition and poor health.

The nutritional value of many foods is markedly affected by the soil, especially where food is produced and consumed locally. In complex food chains where food derives from different sources and processed foods are often supplemented with essential minerals and vitamins the effects of soil are less evident. Crops require 14 mineral elements (listed in Table 2) and deficiency in any one reduces crop yields. In addition to these elements, humans require small amounts of several other trace elements, most of which are from plants. Deficiencies in trace elements occur worldwide (see below). The diets of more than two-thirds of the world's population lack or are deficient in one or more of these essential mineral elements, with more than 60% deficient in Fe, >30% in Zn, almost 30% in I and about 15% in Se. Deficiencies of calcium (Ca), copper (Cu) and magnesium (Mg) are also prevalent in many countries.

Nutrient malnutrition also results from crops produced on soil with poor phytoavailability of the elements essential to human nutrition. Although plant breeding has increased the calories and protein in human diets, the nutritional health of many people, in both developing and developed countries, has deteriorated. Nutrient deficiencies can have far-reaching consequences to human health, but they are often disregarded because they can be difficult to identify. They can result in anemia, diseases of the immune system, mental retardation, and cardiovascular diseases, to name but a few, and have more impact during pregnancy, lactation and periods of rapid growth (childhood). Table 3 summarizes health disorders linked to a range of trace element deficiencies. Nutrient deficiencies are most likely to be evident in developing countries. Nevertheless, they might be an underlying factor for many diseases in developed countries, particularly selenium and zinc.

Future food security will depend on a major change in attitude in agriculture toward sustainable intensification that will halt further degradation of the soil. There is an urgent need to regenerate the soil by matching inputs to the needs of the crop and soil, increase soil organic matter content, remediate contaminated land, improve the quality of irrigation water and scheduling, and the use of technology and precision agriculture.

Nutrition and selected trace elements

Crops require 17 mineral elements, 14 of which are obtained from soil (Table 1). Deficiency in any one restricts plant growth and reduces crop yields. In addition to these elements essential for plants, humans require small amounts of an additional 13 trace elements (Table 1). Humans get most of these elements from plants (Brevik et al., 2020); hydrogen (H), (oxygen) O₂, carbon (C), nitrogen (N), sodium (Na), potassium (K), Ca, Mg, phosphorus (P), S and chlorine (Cl) make up about 99.9% of atoms in the human body and all, except for H, O₂ and C, have soil as their source (Table 1). The remaining 0.1% of atoms come from trace elements. Trace elements can have both beneficial and toxic effects on humans, especially where the range for optimal intake is narrow. The trace elements required by humans cannot be synthesized; therefore, there is a direct link between soil, food and under-nutrition (McMillan, 2002). The diets of more than two-thirds of the world's population lack or are deficient in one or more of these essential mineral elements (White and Broadley, 2009).

Mineral deficiencies occur worldwide, but some elements also occur in potentially toxic concentrations that diminish crop yields. Such elements may be taken up by plants and animals and accumulate in the food chain with detrimental consequences for human health (Oliver and Gregory, 2015). We focus on four trace elements only (iodine, iron, selenium and zinc) because of space, but the deficiencies of these are known to have substantial effects on human health.

Iron

Soil affects the amount of iron in food, but its absorption by humans is complex and depends on many factors (Table 1). Availability of Fe to plants is limited by its solubility in soil which decreases from the amorphous iron oxides to ferrihydrate to goethite to hematite. In aerobic soil, the activity and availability of Fe³⁺ is enhanced by many soil bacteria and fungi.

Iron deficiency in soil, plants and animals results in too little Fe in the human diet. Iron-deficiency anemia (IDA) is one of the most severe effects of nutritional deficiencies in the world affecting some 25% of the world's population and about 47% of

Table 2 A summary of soil and other associated factors that affect human health.

	<i>Human health</i>	<i>Soil factors</i>	<i>Other factors</i>	<i>Food sources of elements</i>
Food insecurity Under-nutrition	Protein–energy malnutrition, stunted growth, disability, immunocompromised, death.	Poor soil, poor nutrient status and maintenance, organic matter and biodiversity depletion, salinization, sodification, acidification, drought, desertification, compaction, wind and water erosion	Transport and distribution, poverty, politics, economics, loss of agricultural land to other uses	
Food quality – elements				
Hidden hunger				
Iron (Fe)	Present in hemoglobin, myoglobin. Iron deficiency anemia (IDA) – cognitive development in children, pregnancy, reduced immunity and thyroid function.	Poor bioavailability – Fe^{2+} more available than Fe^{3+} . Deficiency occurs in aerobic, sandy, calcareous and alkaline soil types. Toxicity occurs in anaerobic and acid ($\text{pH} < 5.5$) soil. Availability depends on redox potential, pH and microbiology (siderophores and phytosiderophores).	Deficiency in food affects humans. Deficiency and toxicity reduce crop yields. Fertilizers can maintain Fe status but expensive. Zinc enhances Fe absorption by humans.	Red meats, liver, egg yolk, dried fruits, beans, lentils.
Iodine (I)	Hypothyroidism - goiter, irreversible brain damage, cretinism, mental retardation and perinatal mortality.	Dry and wet deposition from atmosphere, volatilization from soil and re-deposition. Soil organic matter, microbial activity, texture, Eh and pH affect I fixation. Bioavailability depends on redox and $\text{pH} - \text{IO}_3^-$ (iodate) more available than I (iodide).	Deficiency in humans – deficiency in food and presence of goitrogens. Selenium involved with I in synthesis of thyroxine.	Cod, shellfish, eggs, seaweed, milk, cranberries.
Selenium (Se)	Deficiency – reduces antioxidant and redox functions. Linked to coronary diseases, miscarriage, malignancies. Toxicity – selenosis.	Deficiency is more widespread than toxicity. Microbial processes, pH and redox potential affect Se status. Selenate generally more available than selenite, although some challenge this.	Fertilizers effective in supplementing Se (see example from Finland).	Brazil nuts, cashew nuts, shellfish, tuna, swordfish, seeds, whole grains, meats.
Zinc (Zn)	Metabolic promoter. Gene expression. Deficiency affects physical growth, immune system, reproductive health, dermatological conditions, neurobehavioral problems.	Deficient on leached tropical, sandy, calcareous, alkaline and flooded soil types. Toxic in soil with $\text{pH} < 4$ and deficiency common in soil with $\text{pH} > 7$ and large concentrations of phosphate. Redox status affects form.	Cu inhibits Zn uptake in plants.	Meats, shellfish, whole grain cereals, sesame and pumpkin seeds, pulses, cashew nuts.
Environmental factors				
Elemental contamination	Arsenic (As) (vascular disorders), cadmium (Cd) (kidney and liver disease), fluorine (F) (fluorosis), lead (Pb) (neurological damage and hematologic effects), Zn (reproductive effects).	Toxicity by metals may relate to geology, mining, industrial activity, fertilizers, pesticides, windborne dusts, sewage sludges, irrigation. May accumulate in plants and be ingested or may be ingested directly from soil by mouth or inhalation.	Soil degradation can lead to increased wind and water transport over long distances of soil pollutants.	
Organic chemical contamination	Carcinogenic, neurological and immune problems.	PCBs, PAHs, carbamates, furans accumulate in soil from industry, pesticides and applied biosolids. Some persist in soil (POPs) and enter food chain.	Sources food, windborne dust and soil. Children are most susceptible.	
Pharmaceutical contamination	Emasculation resulting from estrogen in environment. Antibiotic resistance from environmental effects. Unknown human effects from additive effects of mixtures of compounds.	Sewage water and sludges applied to land contain a range of pharmaceutical agents.	Antibiotics and other drugs given to animals enter the environment from sewage sludges with largely unknown effects.	
Soil pathogens	Bacteria (anthrax, tetanus), fungi (Valley fever), protozoa (diarrhea) helminths (anemia, ascariasis), prions (transmissible spongiform encephalopathies).	Occur naturally in soil, some related to anthropogenic activity. Transported long distances by wind. Ingestion by mouth, skin lesions and inhalation.	Desertification and salinization: poor crop cover resulting in greater soil transport of dust containing pathogens.	

Adapted from Table 2, Oliver MA and Gregory PJ (2015) Soil, food security and human health: A review. *European Journal of Soil Science* 66: 257–276.

Table 3 Examples of health problems that can be caused by nutrient deficiencies.

Nutrient	Deficiency diseases/problems
Boron	Osteoporosis, inflammation, atherosclerosis, prostate cancer, neurological problems
Calcium	Low bone mass, osteoporosis, arteriosclerosis, hypertension, degenerative joint diseases
Chromium	Glucose intolerance, diabetes, hyperlipidemia, peripheral neuropathy, encephalopathy, cardiovascular disease
Copper	Hypochromic microcytic anemia, neutropenia, Alzheimer's, ischemic heart disease, hypercholesterolemia, osteoporosis, reduced immunity, cardiovascular disease, secondary iron deficiency
Fluorine	Dental caries, osteoporosis
Iron	Iron-deficiency anemia, fatigue, cognitive impairment, reduced immunity, preterm delivery, fetal growth retardation
Iodine	Goiter, hypothyroidism, mental retardation, cretinism, perinatal mortality, congenital anomalies, cancer
Magnesium	Cardiovascular disease, arteriosclerosis, diabetes, obesity, metabolic syndrome, hypertension, osteoporosis, diabetes, cancer, asthma, muscle weakness
Manganese	Weight loss, transient dermatitis, nausea
Molybdenum	Stomach and esophagus cancer, tooth decay
Nickel	Glucose intolerance, osteoporosis
Phosphorus	Anemia, reduced immunity, muscle weakness, osteoporosis
Potassium	Hypertension, cardiovascular disease, muscle weakness, dementia, abnormal heartbeat, diabetes, stroke
Selenium	Keshan and Kashin–Beck diseases, inflammation, reduced immunity, pseudo albinism, cancer, phenylketonuria, mental retardation
Sulfur	Pain, allergic rhinitis
Vanadium	Hypertension, glucose intolerance
Zinc	Stunted growth and learning, psychomotor and neurobehavioral problems, adverse impacts on maturation of spermatozoa and fetal development, weakened immune system, acrodermatitis enteropathica, pustular dermatitis, alopecia, diarrhea, emotional disorders, weight loss, infections, immune dysfunctions, hypogonadism in males, neurosensory disorders, problems with the healing of ulcers

Sources: Gupta UC and Gupta SC (2014) Sources and deficiency diseases of mineral nutrients in human health and nutrition: A review. *Pedosphere* 24 (1) 13–38, [https://doi.org/10.1016/S1002-0160\(13\)60077-6](https://doi.org/10.1016/S1002-0160(13)60077-6), Oliver MA and Gregory PJ (2015) Soil, food security and human health: A review. *European Journal of Soil Science* 66: 257–276, and Brevik EC, Slaughter L, Singh BR, Steffan JJ, Collier D, Barnhart P, and Pereira P (2020) Soil and human health: Current status and future needs. *Air, Soil, and Water Research* 13: 1–23.

pre-school children (WHO, 2008 see website reference). Iron-deficiency anemia impairs the cognitive development of children from infancy to adolescence and immunity leading to increased susceptibility to infectious diseases and mortality.

Iodine

Iodine was the first element to be recognized as essential to human health. Goiter is the most obvious effect of iodine deficiency, but the more insidious effects are mental retardation, cretinism and perinatal mortality. Cretinism may affect 5–15% of the population in areas of endemic I deficiency.

The main source of iodine in soil is from the oceans by atmospheric deposition in either rain or dry deposits (Table 2). There is no simple correlation between I in soil and distance from the sea, although the largest concentrations are typically 0–50 km from the shore. Iodine deficiency occurs worldwide; it is most pronounced in high mountain areas, inland areas of continents and some alluvial plains. The soil's potential to fix iodine affects its concentration, which depends principally on soil organic matter content, but also on other soil properties given in Table 2. The bioavailable fraction (Table 2) important for health is iodide, which is favored by acidic and anaerobic conditions. In general <25% of soil iodine is immediately available (see Fuge, 2013). The less mobile and available iodate occurs in dry, oxidizing, alkaline conditions such as in calcareous soil. Fig. 2 summarizes the main features of the iodine cycle. Biological methylation of iodine (CH_3I) in soil by a variety of bacteria plays an important role in its transfer to the biosphere.

Insufficient I in crops, animal products and drinking water are the primary causes of iodine deficiency diseases. The effects of iodine deficiency on human health are among the most serious of the trace element deficiencies especially for pregnant women and young children. Iodine is required in only minute amounts, but it is essential in the thyroid gland to synthesize thyroxine, which is crucial for growth and development.

Selenium

Selenium occurs in soil as selenide (H_2Se), selenate (SeO_4^{2-}), selenite (SeO_3^{2-}), elemental Se and organic forms such as selenomethionine and selenocysteine. Selenite is more strongly adsorbed by soil and is less soluble and more stable than selenate. Acid conditions favor the stability of selenite (Se^{4+}).

Selenium is an essential trace element for humans with a narrow range between dietary deficiency (<40 μg Se per day) and toxicity (>400 μg Se per day). Severe human deficiency in Se is rare; the main examples are in China where Keshan and Kashin–Beck diseases are evident (Oliver and Gregory, 2015). Many people, however, have too little Se in their diet with possibly 0.5–1 billion people directly affected by Se deficiency.

Selenium deficiency is implicated in a wide range of medical conditions (Tables 1 and 3). It reduces the antioxidant and redox functions of selenoproteins making people more prone to viral infections (Fairweather-Tait et al., 2011), inflammation and environmental pollutants. Selenium has been added to fertilizers since 1984 in Finland by law.

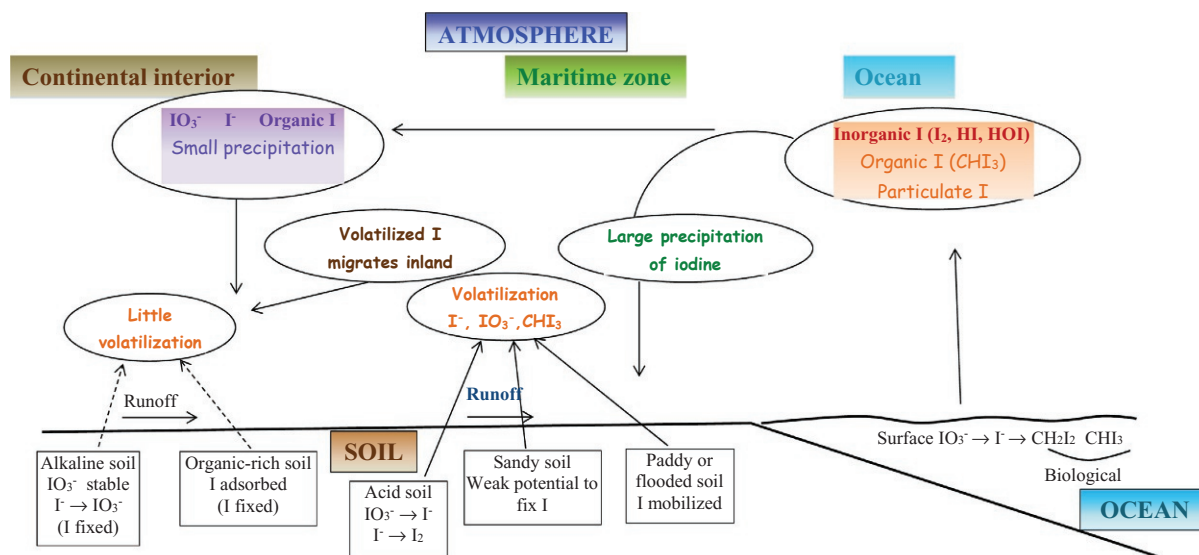


Fig. 2 Schematic diagram that summarizes the main features of the iodine cycle. Adapted from Fig. 5, Oliver MA and Gregory PJ (2015) Soil, food security and human health: a review. *European Journal of Soil Science* 66: 257–276.

Selenium toxicity in humans, which can cause selenosis, is much less common than deficiency. Toxic concentrations, $>3 \text{ mg Se kg}^{-1}$, can occur locally on seleniferous parent materials and can also result from irrigation with Se-rich water.

Zinc

Zinc deficiency affects about a third of the world's population. In children it causes stunted growth and learning, psychomotor and neurobehavioral problems. It also affects the formation and maturation of spermatozoa and fetal development and the immune system. Severe Zn deficiency gives rise to acrodermatitis enteropathica (a genetic disorder), pustular dermatitis, alopecia, diarrhea, emotional disorders, weight loss, infections, immune dysfunction hypogonadism in males, neurosensory disorders and problems with the healing of ulcers.

The solubility of Zn is generally small and pH dependent (Table 2). Factors affecting Zn availability to plants are total Zn content, pH, contents of organic matter, clay and calcium carbonate, redox conditions, microbial activity in the rhizosphere, soil moisture status, concentrations of other trace elements and macronutrients, especially P, and climate (Alloway, 2008). Copper and large phosphate concentrations in soil can cause Zn deficiency. Although Zn is not susceptible to oxidation or reduction, the soil's redox status can affect its form in soil.

Zinc is essential for animal and human health although the effects of its deficiency in humans were not recognized until the 1960s. Zinc deficiency can be overcome by its addition to fertilizers, which increase Zn bioavailability in the diet.

Effects of soil entering the human body

Humans can be exposed to soil directly by the pathways of ingestion, inhalation during respiration and absorption or penetration of the skin, i.e., dermal contact. These pathways result in exposure to: (i) toxic chemicals that can be carcinogenic, can adversely affect cognitive and respiratory function, growth and fertility; (ii) pathogens such as bacteria, fungi and helminths; and (iii) small irritant particles such as asbestos and clays.

Ingestion

Soil can be ingested involuntarily by eating contaminated food such as unwashed vegetables and fruits and by hand-to-mouth activity, which is especially prevalent in children. If the soil is contaminated with metals, organic chemicals, or pathogens this can result in adverse health effects. The nervous systems of small children are especially sensitive to Pb, more of which can accumulate in their brains than in older people because of the incomplete development of the blood–brain barrier. Lead is a cumulative element in the body and according to the World Health Organization (WHO, 2019 – see website below), “there is no level of exposure to lead that is known to be without harmful effects.”

The deliberate consumption of soil, especially clayey soil, has a long history and is amazingly widespread. The term geophagy was coined in ~1850 to describe this custom, however, it has existed for at least two millennia. Clark (1969) described an anomalous clay layer at a prehistoric site at the Kalambo Falls in Zambia indicating that clay might have been eaten by hominids

(*Homo erectus* or *Homo sapiens*) (see “History of soil and human health”). Von Humboldt reported that clay was eaten by the Otomac tribe along the Orinoco River, and he saw mothers give their children lumps of clay in Peru to keep them quiet. Geophagia is most common in developing countries where there is poverty and many people suffer from malnutrition. Clay is traded for consumption in India, Africa and the USA, among other locations. Recent research suggests that there are two main explanations: (i) to compensate for mineral deficiencies and detoxification, or (ii) that it is learned behavior, i.e. cultural. Clay soil can release essential elements that humans require by cation exchange, such as Cu, Fe, Mg, Mn (manganese), K (potassium), Si (silicon) and Na, in the acid conditions of the stomach. Geophagy is not necessarily beneficial because, for example, Ca-rich clay can affect the utilization of Zn and Fe, and inhibit their absorption by the body. Avicenna (980 to 1037 CE) suggested that geophagia led to anemia. It can also result in the ingestion of harmful bacteria, parasites and environmental pollutants.

Johns and Duquette (1991) suggested that detoxification is a more satisfactory explanation for geophagy because clay soil can absorb dietary toxins, such as alkaloids, tannin, and bacterial toxins. They indicated that clay was used to make acorns more palatable and to detoxify bitter potatoes (*Solanum* spp.). The clays used for geophagy have a large cation exchange capacity and specific surface area; they can absorb and adsorb elements and microbes. Kaolin and smectite clays are effective in medicinal use because they can bind with toxins, aflatoxins and bacteria. Because of their effect on the intestinal mucosa, clays can reduce nausea, diarrhea, and intestinal pH.

Soil is often contaminated with enteric bacteria and helminths when it is ingested. For example, *Campylobacter*, *Escherichia coli* and *Shigella* spp. are Gram-negative bacteria that do not form spores, and possibly cause up to 6 million deaths per year. They are the second most common cause of infant mortality. Most helminths are nematodes; they include hookworm, round worms, whipworms and pinworms. There are probably two billion people worldwide infected by helminths, and Bultman et al. (see Selinus et al., 2013) suggested that 130,000 deaths per annum are caused by nematode infections. The history of human health (see above) shows that they have always been part of human existence. *Ascaris lumbricoides* (round worm), which causes ascariasis, an infection of the small intestine, is the largest and most widespread nematode. Infections result from poor sanitation and infected soil ingested directly or from dirty vegetables and fruit. Helminths are transmitted by eggs present in human feces, which contaminate soil in areas where sanitation is poor. Adult worms live in the intestine where they produce thousands of eggs each day. In areas that lack adequate sanitation, these eggs contaminate the soil. They produce a wide range of symptoms including diarrhea, abdominal pain, general malaise and weakness.

Inhalation

Soil can be inhaled as dust, especially in dry regions with fine soil. There has been increasing recent realization that dust inhaled from urban soils can contain metals and other contaminants from industrial and domestic emissions, buildings, fuels, roads and tyre particles. Cadmium and lead are of particular concern for human health because of their accumulation in the body. Contaminated soil can be transported over very long distances.

Soil rich in fibrous minerals such as tremolite and erionite (asbestos-like minerals) are known to cause malignant pleural mesothelioma. There are examples of such effects in Cappadocia, central Turkey and northeast Corsica, together with concerns in western North Dakota, USA.

Pathogens can be inhaled easily because of their small size. The fungal pathogen *Coccidioides immitis* causes Valley Fever (coccidioidomycosis) in the southwestern USA. The yeast-like fungi *Cryptococcus neoformans* and *C. gattii* occur in the soil of tropical and sub-tropical regions. Inhalation of their spores leads to cryptococcosis in humans, which can affect the nervous system as cryptococcal meningitis or the lungs as pneumonia (Redshaw et al., 2013). Over the past decade or so outbreaks of disease associated with *C. gattii* have occurred in temperate regions of Canada and the USA; these might be a response to a warming climate. Infection from *Hantavirus* can result from inhalation of wind-blown particles from infected soil. This can occur with tillage of soil when loose material containing *Hantavirus* might be inhaled. Infections can cause either ‘hantavirus pulmonary syndrome’ (HPS) or ‘hemorrhagic fever with renal syndrome’ (HFRS). Hantavirus infections can occur worldwide, but in the USA most cases occur in the arid and semiarid western states. Pathogens in raw sewage used for irrigation in dry areas can be easily picked up and transported by the wind and then inhaled. This could also be an important pathway for arsenic and lead to enter the body.

Radon is a radioactive gas that humans inhale in certain areas and which causes malignancies. Radon decays to solid daughters, e.g. ^{218}Po (polonium), ^{214}Pb , ^{214}Bi (bismuth) and ^{214}Po , and epidemiologists suspect that inhaled radon and the solid decay products retained in the lungs lead to or contribute to lung cancer. However, it is also implicated in other forms of cancer, including certain childhood ones. Radon accounts for about 50% of the radiation dose received by people living in the United Kingdom.

Soil absorption through skin lesions

One of the most common diseases associated with soil is tetanus which results from infection by *Clostridium tetani* when it enters the body through cuts and abrasions. These anaerobic bacilli occur in the intestines of man and other animals and enter the soil from feces. Spores from the bacterium can remain dormant in the soil for many years and then germinate and multiply in the anaerobic conditions of human tissue. The bacilli remain localized but they produce an exotoxin that passes into the blood stream which affects the motor nerve endings and cells. It causes spasm and convulsions, and death in about 50% of cases. The disease is well controlled by immunization.

Parasites such as hookworm can also enter the body through abrasions. Hookworm infects over 500 million of the world's population. There are two main species of hookworm, *Ancylostoma duodenale* and *Necator americanus*. Their eggs hatch to rhabditi-form larvae and then become filariform larvae. The latter enter the body through abrasions in the feet and pores in skin, then move through the blood stream into the lungs, trachea and esophagus. When adult they attach themselves to the walls of the small intestine where they produce their eggs which are then defaecated to complete the cycle. They cause bleeding which causes anemia when the infestations are severe.

An interesting link between soil and human health is a non-filarial elephantiasis called Podoconiosis, a form of lymphoedema that causes swelling of the lower legs resulting from blockage of lymph nodes by soil. It appears to be associated with being bare foot and with a particular type of fine clay soil that occurs in the tropics. Podoconiosis is now the most common form of elephantiasis. It has been suggested that micrometric particles of soil containing kaolinite, amorphous silica, some quartz and iron oxides penetrate bare skin and are engulfed by macrophages in the lymphatic system of the lower limb, inducing an inflammatory response in the lymphatic vessels leading to lymphoedema and fibrosis.

Soil organisms

There are several ways that soil organisms can be sub-divided. One is into micro- and macro-organisms based on size. Both micro- and macro-organisms play important roles in the overall soil ecosystem, and thus are critical for creating a suitable environment for crop production as well as several other ecological functions. However, the influence of soil microorganisms on human health has been extensively studied while similar work regarding macroorganisms is much rarer (Brevik et al., 2020). Another subdivision is based on the time the organism spends in the soil (Table 4), and from a soil and human health perspective is useful in understanding possible modes of transmission and infection. Classifying soil organisms as pathogenic or non-pathogenic, and what they are pathogenic to, is also important in soil and human health studies. Plant and human pathogens both impact human health; the plant pathogens do so indirectly with their influence on crop production and the human pathogens do so directly. Examples of a few common human pathogens found in soil are given in Table 5 and displayed in Fig. 3.

Human pathogens are often introduced to soil through anthropogenic activities. Disposing of sewage with on-site infrastructure such as septic systems, applying raw sewage effluent or manure (particularly human effluent or manure) to crops, and land disposal of municipal and clinical wastes are all ways that pathogens may be introduced to soil. Application of organic wastes to soil is an important part of a circular economy, and these wastes contribute to soil biological, chemical, and physical properties. Pathogens can be killed through proper composting prior to application of solid wastes to soil and treatments such as chlorine, ozone, or ultraviolet light prior to applying liquid wastes. A healthy soil ecosystem with much biodiversity can also help to keep pathogens under control (Wall et al., 2015).

Exposure to soil microorganisms likely plays an important role in developing and regulating the human immune system. Much of the study in this area has focused on allergic responses, with studies showing that those who were raised in environments where they were exposed to soil were considerably less apt to develop allergies than those raised where they had little soil exposure (Brevik et al., 2020). Soil microorganisms also seem to have positive influences on physical health. In mouse studies gut microbes acquired from soil increased anti-inflammatory response, and in humans providing a killed *Mycobacterium vaccae*, which is a common soil saprophyte, improved the performance of chemotherapy. Findings such as these indicate that having contact with soil microorganisms is beneficial to humans. Blum et al. (2019) have proposed that a close link has formed between the soil microbiome and the human intestinal microbiome during human evolution, and that this link continues to evolve as human lifestyles and settings change. They are particularly concerned that as humans have moved to a more urban lifestyle, with less soil contact, the human intestinal microbiome has been depleted, which has a negative impact on human health.

The soil ecological system, including both micro- and macro-organisms, is important in crop production. This occurs because organisms like earthworms and plant roots take part in creating soil voids and structure, which is critical to achieve satisfactory water infiltration and storage and aeration in the soil. Soil organisms also control the decomposition of soil organic matter, releasing plant available nutrients into the soil solution and creating glues that help bind primary soil particles into structural units. And soil organisms themselves add to soil organic matter, lending their tissues to the system when they die, which recycles the nutrients found in them. These relationships have been well-known for many years, and they are all important to the production of adequate

Table 4 Classification of soil organisms based on the amount of time they spend in the soil.

Classification	Time
Permanent	The organism is a permanent inhabitant of the soil and can complete its entire life cycle in the soil
Periodic	The organism requires part of its life cycle to be completed in the soil
Transient	The organism may be found in the soil naturally, but the soil is not required to complete its life cycle
Incidental	Organisms that are introduced into the soil through anthropogenic activities

Table modified from Collier D and Brevik EC (2021) Soils and human health: Communication between soil scientists and health care providers. In Lal R (ed.), *The Soil-Human Health-Nexus*. 59–80. Boca Raton, FL: CRC Press.

Table 5 Examples of soil organisms that can have negative impacts on human health.

Organism	Common name	Point of infection/ gateway	Disease name	Associated problems	Soil residency	Incidence	Geographic distribution
<i>Necator americanus</i>	Hookworm	Intestine	Ancylostomiasis	Chronic anemia, diarrhea, cramps	Periodic	1.2 billion/year (includes <i>A duodenale</i>)	Central and South America, southern Asia, Australia, Pacific Islands
<i>Strongyloides stercoralis</i>	Roundworm	Small intestine	Strongyloidiasis	Abdominal pain, diarrhea, rash	Permanent	100 million/year	Tropical, subtropical, and temperate regions
<i>Enterobius vermicularis</i>	Pinworm	Colon	Enterobiasis	Anal itching	Incidental	200 million total	Temperate regions
<i>Entamoeba histolytica</i>	Giardia	Intestine	Amebiasis	Diarrhea, dysentery, liver abscesses	Incidental	40 million	Worldwide
<i>Giardia lamblia</i>		Large intestine	Giardiasis	Diarrhea, abdominal cramps, bloating, fatigue, weight loss	Transient-incidental	500,000	Worldwide
<i>Dientamoeba fragilis</i>		Gastrointestinal tract	<i>D. fragilis</i> infection	Mild abdominal pain, gas, diarrhea	Unknown	2–4% of the population in developed countries	Worldwide
<i>Aspergillus</i> sp.		Respiratory, possibly through contaminated biomedical devices	Aspergillosis	Wheezing, fever, cough, allergic sinusitis, chest pain, shortness of breath	Permanent	1–2/100,000	Worldwide
<i>Coccidioides</i> sp.		Respiratory, sometimes trauma to skin	Coccidioidomycosis (Valley fever)	Flu-like symptoms, fever, cough, rash, headache, muscle aches	Permanent	15/100,000	Southwestern United States, northern Mexico, microfoci in Central and South America
<i>Trichophyton</i> sp., <i>Microsporum</i> sp., <i>Epidermophyton</i> sp.	Ringworm	Skin contact	Tinea corporis	Itchy, red circular rash with healthy-looking skin in the middle		Common	Worldwide
<i>Bacillus anthracis</i>		Respiration, skin trauma, ingestion	Anthrax	Ulcer w/black necrotic center, severe breathing problems, shock, nausea, fever, vomiting, diarrhea, abdominal pain	Periodic	Rare	South and Central America, southern and eastern Europe, Asia, Africa, the Caribbean, the Middle East
<i>Clostridium tetani</i>	<i>E. coli</i>	Skin trauma	Tetanus (Lockjaw)	Painful muscle tightening	Permanent	500,000 annual deaths	Worldwide
<i>Escherichia coli</i>		Ingestion	Diarrhea	Bloody diarrhea, abdominal pain, vomiting, fever	Incidental	5 million/year	Worldwide
<i>Clostridium botulinum</i>		Ingestion	Botulism	Vision problems, slurred speech, dry mouth, difficulty swallowing, muscle weakness	Permanent	Rare	Worldwide
<i>Clostridium</i> sp. (other than 2 above)		Skin trauma	Gas gangrene	Air under skin, blisters, fever, drainage, pain	Permanent	Common before antibiotics	Worldwide
<i>Rickettsia</i> sp.		Tick bite	Rocky Mountain spotted fever, other tick fevers and spotted fevers	Fever, nausea, vomiting, severe headache, pain, rash	Periodic	250 to 1200 cases/year in the United States	Worldwide

Table modified from Collier D and Brevik EC (2021) Soils and human health: Communication between soil scientists and health care providers. In Lal R (ed.), *The Soil-Human Health-Nexus*. 59–80. Boca Raton, FL: CRC Press.

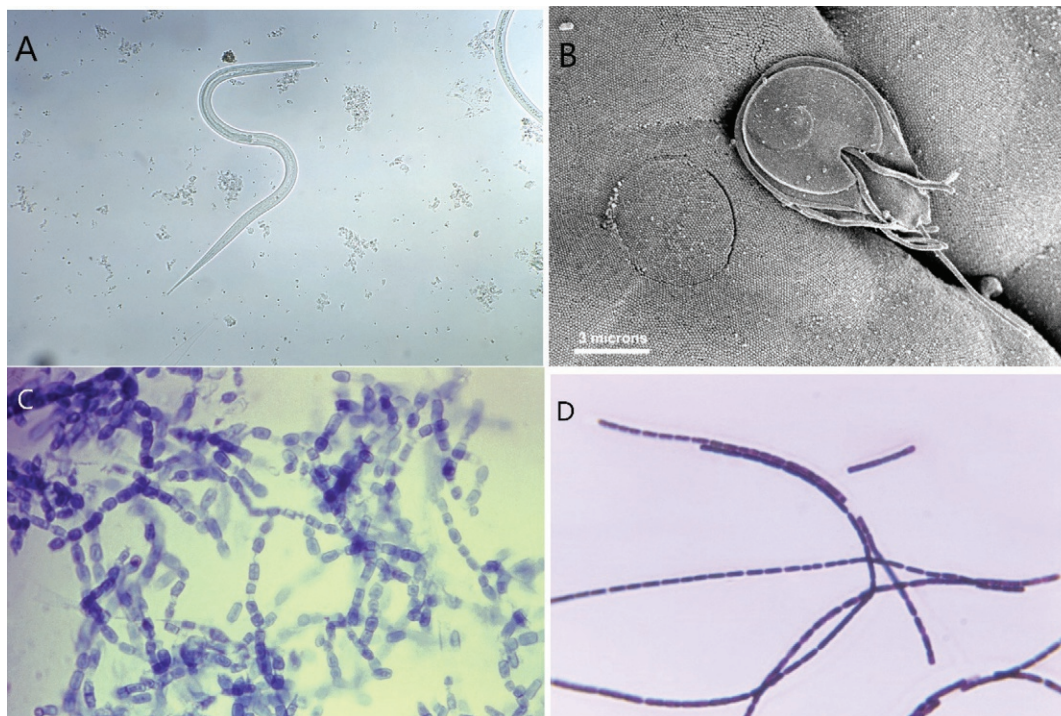


Fig. 3 Examples of some soil pathogens. (A) Infective, third-stage filariform larvae (L3; center of picture) of *Strongyloides stercoralis* at 128 × magnification. L3 larvae are found in soil and can invade a human host by penetrating the skin. (B) A *Giardia* sp. protozoan is shown in the center of this scanning electron microscopic image. The circular lesion to its left shows where a *Giardia* sp. parasite was previously attached to the intestinal surface. (C) *Coccidioides immitis* at 600 × magnification, showing the chains of barrel-shaped arthrospores that are common to this organism. (D) A Gram-stained photomicrograph showing several rod-shaped Gram-positive *Bacillus anthracis* bacteria arranged in long strands. (A) Image courtesy of the Centers for Disease Control and Prevention, image ID# 1548. (B) Image courtesy of the Centers for Disease Control and Prevention, image ID# 11647. (C) Image courtesy of the Centers for Disease Control and Prevention, image ID# 16303. (D) Image courtesy of the Centers for Disease Control and Prevention, image ID# 20496.

amounts of nutritious crops, a fundamental requirement for human health. However, they also represent indirect benefits for plant production. More recently the ways that soil microorganisms directly aid crop production have been explored in greater depth. Manipulating the soil microbial community found in the rhizosphere through inoculations with bacteria or fungi that promote plant growth has shown the ability to increase the quantity and nutritional quality of harvests (Brevik et al., 2020). Some of these inoculations are even able to improve crops when soil nutrient levels are low. These studies highlight the co-dependent nature under which plants and soil microbial communities have evolved, and the benefits hold true for a wide range of crops. Learning how to manage soil ecological systems appropriately will be an important part of achieving sustainable agricultural production and improving human health through soil-based benefits.

As we look toward the future, there are still many aspects of the soil organisms-human health nexus that need investigation. Most of the research to date has focused on soil microorganisms, investigating links between human health and soil macroorganisms has been largely overlooked. There is also a need to understand better the biological processes that underpin the improvements in yield and quality seen when the proper plant–soil microbe relationships are established, how exposure to soil microbes helps to build the human immune system, and links between the soil microbiome and human microbiome. Sequencing and metagenomics will likely assist in establishing these relationships. There is still an abundance of work to be done in this area.

Effects of soil pollutants – metals, organic chemicals, and radionuclides

Humans have released metals and metalloids (henceforth, metals for the purposes of this chapter) into the environment for thousands of years, but the rate at which metals have accumulated in soils and other environmental media at levels that lead to pollution accelerated rapidly during the Industrial Revolution. Although often associated with adverse effects on human and environmental health, metals are also quite important to human health; many of the elements essential to human life fall into these categories. Physiologically, metals take part in oxidation–reduction reactions in the human body, act as antioxidants or pro-oxidants, and play a role in relief from oxidative stress (Madeshiya et al., 2021). The challenge with metals is that the difference between deficient and toxic can be very small; in some cases this difference may be just a few tens to hundreds of grams per ha. When metals that are not essential for human health are present at toxic levels, they can react with the sulfhydryl enzyme and restrict

Table 6 Examples of pollution indices that have been used to provide information on the relationship between soil pollution levels and human or environmental health risks.

<i>Index</i>	<i>Basic measurement being made</i>
Contamination degree	The sum of the contamination factors for a given sample.
Contamination factor	The ratio between the content of a given metal and its background concentration.
Enrichment factor	Measures the impact of anthropogenic activity on soil trace element concentrations. It does this by comparing the ratio of the concentration of an element of interest to the concentration of a metal with a low variability of occurrence to the ratio of the background levels of the same two elements.
Geoaccumulation index	Evaluates soil contamination by a single trace element by calculating the logarithm of its concentration divided by 1.5 times its background concentration.
Modified degree of contamination	Degree of contamination divided by the number of analyzed elements.
Pollution load index	Determined by the number of times that the trace element concentrations in a soil exceed the background concentration
Single pollution index	Used to determine which trace element is the greatest threat in a given soil environment, calculated as the concentration of a trace element in the soil divided by its background value. Similar to contamination factor.
Sum of pollution index	The sum of all single pollution indices calculated for a given soil. Similar to degree of contamination.

cellular production by replacing essential metals on the enzyme's active sites (Ullah et al., 2020). Determining what level of metal accumulation in soil represents a threat is not straightforward. Several soil chemical properties, including pH, Eh (redox potential), organic matter content, clay content and mineralogy, and the presence of other ions, especially other metals, help determine the threat level for a given amount of metal in any given soil. To help researchers and policy makers address these differences between soils, several indices have been developed (Table 6).

There are several routes through which metals can be enriched in soils, including natural sources, industrial sites and wastes, e-wastes, traffic emissions, coal and fuel combustion residues, mining, paints, sewage, pesticides, fertilizers, and petrochemicals. Metals can affect several systems in the human body, including the blood, liver, brain, kidneys, skin, and lungs. Human health problems caused by metals include neurological degenerative processes (e.g. Parkinson's disease and Alzheimer's disease), cancers in the various body systems mentioned above, and osteoporosis (Brevik et al., 2020). One of the major concerns with metals in the environment is that they do not biodegrade over time. Once a metal is present in a soil, it has the potential to cause human and environmental health issues over very long periods of time.

A wide variety of materials fall under the heading of organic chemicals. Some are naturally occurring, such as petroleum and bitumen, but many of the organic chemicals that are major pollution concerns in the modern world are synthetic. Some of these synthetic chemicals were produced purposely (e.g. plastics, lubricants, fuels, solvents, herbicides, insecticides, and preservatives), but several are also unintended byproducts (e.g. polychlorinated dibenzo-p-dioxins (PCDDs or dioxin) and polychlorinated dibenzofurans (PCDFs)) of organic chemical synthesis. Many of these synthetic organic chemicals are highly resistant to microbial decomposition (Brevik and Burgess, 2013).

Among the synthetic organic chemicals, the particularly problematic formulations have become known as persistent organic pollutants (POPs). Chemicals classified as POPs are determined at meetings of the Stockholm Convention. The original list of 12 POPs was established in 2001 and became known as the 'dirty dozen.' Since then, nine more chemicals were added in 2009, one in 2011, one in 2013, three in 2015, two in 2017, and one in 2019 (Table 7). Some of the chemicals on the Stockholm Convention's list can still be used in limited applications, but some cause so many environmental problems that their use is now completely banned. In addition to the existing POPs, three chemicals are currently under review by the Stockholm Convention for possible future listing, Dechlorane Plus, Methoxychlor, and UV-328. Seventeen of the 29 POPs can be used as pesticides, and many of these have ended up in agricultural soils through pesticide applications. Health problems related to organic chemical use include diabetes, various cancers, cardiac failure, asthma, hepatic necrosis, dizziness, nausea, and headaches (Brevik et al., 2020).

Radionuclides in soil can also present health problems. They can be introduced to the soil through anthropogenic activities such as disposal of medical wastes, disasters such as Chernobyl (April 26, 1986) and Fukushima (March 11, 2011) or occur naturally. With many radionuclides the concern is uptake by crops, which pass the radioactive elements on to humans through the food web. A variety of cancers can be caused by exposure to this radiation. The decay process of radionuclides found naturally in soil also generates radon, which may accumulate in basements or other confined spaces with limited air movement and be inhaled, introducing radioactive elements to the lungs. Radon exposure combined with tobacco smoking is particularly problematic in the development of lung cancer.

Medicines from soil

Antibiotics

The most significant medical use of soil has been in the isolation of antibiotics from soil organisms, mainly bacteria and fungi. Antibiotics have been used for millennia, for example, ancient Egyptians and Greeks used various molds and plant extracts to treat

Table 7 Persistent organic pollutants (POPs) identified by the Stockholm Convention (SC). Some POPs can still be used for specific purposes as outlined in the SC.

Chemical	Year added	Source ^a	Annex in SC	Additional notes
Aldrin	2001	P	A	
Chlordane	2001	P	A	
Dichlorodiphenyltrichloroethane (DDT)	2001	P	B	DDT still used against mosquitoes in several countries to control malaria
Dieldrin	2001	P	A	
Endrin	2001	P	A	
Heptachlor	2001	P	A	
Hexachlorobenzene (HCB)	2001	P, IC, UP	A & C	
Mirex	2001	P	A	
Toxaphene	2001	P	A	
Polychlorinated biphenyls (PCB)	2001	IC, UP	A & C	Has specific exemptions under Annex A
Polychlorinated dibenzo-p-dioxins (PCDD)	2001	UP	C	
Polychlorinated dibenzofurans (PCDF)	2001	UP	C	
Alpha hexachlorocyclohexane	2009	P	A	No exceptions or acceptable uses
Beta hexachlorocyclohexane	2009	P	A	No exceptions or acceptable uses
Chlordecone	2009	P	A	No exceptions or acceptable uses
Hexabromobiphenyl	2009	IC	A	No exceptions or acceptable uses
Hexabromodiphenyl ether, heptabromodiphenyl ether	2009	IC	A	Can be used in accordance with the provisions of Part IV of Annex A
Lindane	2009	P	A	Human use for control of head lice and scabies as second line treatment
Pentachlorobenzene	2009	P, IC, UP	A & C	No exceptions or acceptable uses
Perfluorooctane sulfonic acid, its salts and perfluorooctane sulfonyl fluoride	2009	IC	A & B	Acceptable purposes and specific exemptions in accordance with Part III of Annex B, amended 2019
Tetrabromodiphenyl ether, pentabromodiphenyl ether	2009	IC	A	Has specific exemptions under Part V of Annex A
Technical endosulfan and its related isomers	2011	P	A	Exemptions for crop-pest complexes in accordance with the provisions of part VI of Annex A
Hexabromocyclododecane	2013	IC	A	Expanded and extruded polystyrene in buildings in accordance with the provisions of part VII of Annex A
Hexachlorobutadiene	2015	IC, UP	A & C	No exceptions or acceptable uses, added to annex C in 2017
Pentachlorophenol and its salts and esters	2015	P	A	Pentachlorophenol for utility poles and cross-arms in accordance with the provisions of part VIII of Annex A
Polychlorinated naphthalenes	2015	IC, UP	A & C	Can be used for production of polyfluorinated naphthalenes, including octafluoronaphthalene
Decabromodiphenyl ether	2017	IC	A	Exemptions for certain uses in vehicles, aircraft, textiles, additives in plastic housings etc., polyurethane foam for building insulation
Short-chain chlorinated paraffins	2017	IC	A	Allowed as additives in transmission belts, rubber conveyor belts, leather, lubricant additives, tubes for outdoor decoration bulbs, paints, adhesives, metal processing, plasticizers
Dicofol	2019	P	A	No exceptions or acceptable uses

^aP – pesticide, IC – industrial chemicals, UP – unintentional production/by-products.

Table from Brevik EC, Slaughter L, Singh BR, Steffan JJ, Collier D, Barnhart P, and Pereira P (2020) Soil and human health: Current status and future needs. *Air, Soil, and Water Research* 13: 1–23.

infections. However, it was not known until the late eighteenth century that the infections were caused by bacteria. Penicillin was the first antibiotic to be discovered from the soil-borne fungus, *Penicillium notatum*, but at the time the term ‘antibiotic’ had not been coined. Penicillin was isolated by Sir Alexander Fleming in 1929 by accident on a culture of *Staphylococcus* on which *P. notatum* landed and created bacteria free zones. The antibiotic penicillin was mass produced in the early 1940s in the USA based on a process developed in Oxford, UK, and used widely by 1944 to treat troops for infections. Sir Alexander Fleming, Ernst Boris Chain and Sir Howard Walter Florey received the Nobel Prize in Physiology or Medicine in 1945 for their work on penicillin.

Selman Waksman was the first to coin the term antibiotic in 1943. In the 1940s, Waksman and his team at Rutgers University isolated several inhibitory compounds (actinomycin, neomycin and streptomycin) from *Actinomyces* and *Streptomyces* spp. (of the phylum *Actinobacteria*). These are prokaryotic organisms that are Gram-positive, filamentous and mainly aerobic and saprophytic bacteria that produce spores. Waksman’s team went on to discover over 20 inhibitory substances from *Actinomycetes* by 1944.

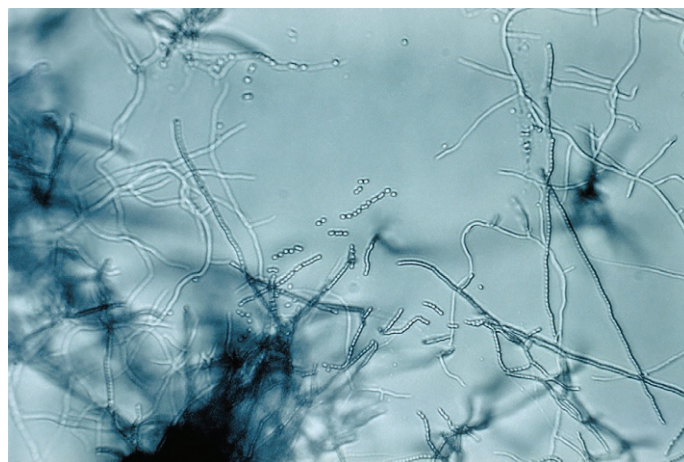


Fig. 4 Magnified image of a culture of *Streptomyces* sp. bacteria. Note the branching filamentous hyphae, abundant aerial mycelia, and long chains of small spores. These features are characteristic of all *Streptomyces* spp. Photo courtesy of the Centers for Disease Control and Prevention, image ID# 2983.

Waksman, (1952) found that some microbes compete with one another for survival by secreting secondary metabolites as a chemical defense (for example antibiotic, antifungal, immunosuppressant, anti-tumor and antihelminthic products) to gain a competitive advantage in the soil ecosystem (Ratnaddass et al., 2006). The production of antibiotics and other metabolites by *Actinobacteria* appears often to be controlled by diffusible hormones produced during the filament forming part of their life cycle.

The complex secondary metabolites produced by *Actinomycetes* account for over two thirds of all known antibiotics, of which the majority is produced by *Streptomyces*. Some *Streptomyces* (Fig. 4) species produce a single antibiotic, but others produce a range of different compounds and compound classes (Barka et al., 2016). The soil is a reservoir of potentially new antibiotics because bacteria are continuously developing new ones to resist new stresses. Until recently, however, the physiology and genetic make-up of only a few bacteria have been studied because most (>99%) cannot be cultured in the laboratory. Now advances in genomics methods can enable the physiology and genetics of microbes that cannot be cultured to be determined. A recently discovered (2014) soil bacterium, temporarily known as *Eleftheria terrae*, was found to produce a new antibiotic, teixobactin. It has been found to be effective against multidrug-resistant bacteria such as *Staphylococcus aureus* and *Mycobacterium tuberculosis*. Another new group of antibiotics known as malacidins was discovered recently (2018) from *Streptomyces albus* after examining 2000 soil samples. The great advantage of the malacidins is that they are also effective against multidrug-resistant pathogens. When tested against *S. aureus* malacidins were found to be effective. Soil fungi, for example the cephalosporin group, are also major sources of antibiotics. The successful immunosuppressant cyclosporine is derived from the fungus *Tolypocladium inflatum*.

Several antibiotics derived from soil bacteria are widely used as chemotherapeutic drugs such as the anthracyclines Daunorubicin and Dexorubicin, which are produced by *S. peucetius*. Actinomycin D, which was one of the first antibiotics discovered, is commonly used to treat many types of cancerous tumors. The National Cancer Institute (Bethesda, Maryland) launched an effort in 2015 to encourage research into natural products from the environment such as from soil as a source of new drugs.

The need to find new antibiotics is critical because of the increase in resistance of many pathogens to several currently used antibiotics. A major cause of this resistance has developed from the release of antibiotics to soil and water through applications of human and animal waste to soil and water. These exogenous antibiotics affect soil microorganisms by changing their enzyme activity and ability to metabolize various sources of carbon. Furthermore, the mixing of exogenous bacteria with indigenous antibiotic-producing bacteria (Wellington et al., 2013) can result in horizontal gene transfer that produces antibiotic-resistant genes (ARGs). The ARGs relate to agricultural and human activities in several ways, one of which is their incomplete metabolism in humans and animals.

Clays

The ancient Greeks used Lemnos earth, a mud most probably composed of clay, for skin problems and snake bites. The phyllosilicate clays have had a long history of medicinal use because of their capacity for adsorption and absorption, and surface charge. The clays used in pharmaceuticals include smectites such as bentonite, palygorskite (attapulgite), kaolinite and talc, which have large specific areas with strong adsorptive and absorptive capacity, rheological properties, are chemically inert and have little or no toxicity. They can remove secretions, contaminants and toxins from the skin and adsorb aflatoxins, which are naturally occurring carcinogenic mycotoxins, toxic secondary metabolites produced by several *Aspergillus* species (commonly known as molds) and found in all soils. Clays used for gastrointestinal therapy are palygorskite (a magnesium aluminium silicate) and kaolinite because of their large specific area and sorption capabilities. These clays are used to treat diarrhea because they absorb water and gases and can absorb toxins, bacteria and viruses.

Otto and Haydel (2013) examined the antibacterial properties of natural clay mixtures in vitro and tests on their samples showed that *Escherichia coli* and methicillin-resistant *Staphylococcus aureus* (MRSA) were killed in 24 h. They considered that the antibacterial properties are due to adsorbed ions on the clays, namely Fe^{2+} , Cu^{2+} and Zn^{2+} . Brunet de Courssou treated the Buruli ulcer caused by a necrotizing flesh-eating mycobacterium, *Mycobacterium ulcerans*, with two French green clays. One of the clays killed the bacterium, which is endemic in Central and West Africa. However, the mechanism by which the clay minerals inhibit bacterial growth remains unknown.

The future

Humankind has learned much about the links between soil and human health, starting with experiential, trial and error approaches early in our history to the complex, carefully planned scientific experiments that are run today. However, true scientific study of the relations between soil and human health did not really start until about the 20th century, and we still have much to learn.

The impacts of soil properties and processes on human health are due to complex interactions that are difficult for any single discipline to study adequately and do not translate well to the use of traditional approaches that isolate a single variable to determine how it affects the overall system. It has been conclusively demonstrated that similar levels of, for example, soil contamination by cadmium at two different locations can have two very different outcomes in terms of human health (Brevik and Burgess, 2013). This is because factors such as soil physical and chemical properties influence cadmium availability to plants, and through that just how the contamination impacts human health. Simply isolating cadmium as a variable and seeing how the system reacts is not sufficient. Several indices have been developed to evaluate the health risk associated with metals found in soil, and these are better than just using total contaminant concentration (Brevik et al., 2020). However, their use assumes that the soil itself is the only factor determining the effects on health. Human diet, for example, has also been shown to play a role. Returning to the cadmium example above, the diet in one location was deficient in iron and zinc compared to the other, and cadmium retention is increased 15 times in those with iron and zinc deficient diets compared to those with diets that are not deficient (Brevik and Burgess, 2013).

As we seek to understand complex interactions new models and statistical techniques will be required, and we will need to learn from other fields that work with similar challenges. There are several approaches utilized in modern epidemiology that are able to account for multiple variables but are not typically used in soil science, including causal diagrams, marginal structural models, and propensity score methods (Brevik et al., 2021). Epidemiologists also deal with environmental exposures, just as soil scientists and others with an interest in soil and human health links do, providing promise that techniques that work for epidemiology may also work in soil and human health research. Omics approaches in biology are another good example that could be applied to the study of soil pathogens and the broader soil microbial community. As we look to the future, it will be important to learn from related fields and investigate the applicability of their techniques to address soil and human health challenges.

Multiple disciplines engage in the knowledge necessary to address the links between soil and human health holistically. Agronomists, biologists, crop scientists, epidemiologists, geographers, geologists, lawyers, medical doctors, public health officials, sociologists, soil scientists, and others have roles to play. The disciplines in this list are diverse, do not all speak the same professional language, not all of them typically work together, and they do not communicate using the same professional outlets. Finding ways to facilitate communication and cooperation between these diverse professional groups will be a key part of moving our knowledge of soil and human health links forward, and how to implement that knowledge effectively.

Another challenge we face is how to raise awareness concerning the importance of soil to human health among the general public, health agencies and governments to initiate a preventive approach to maintaining healthy populations. Research areas that are perceived as important receive higher levels of funding than those that have less recognition (Brevik et al., 2020). Several options for developing this recognition have been suggested, including using social marketing or social media platforms to build recognition of soil and human health issues by connecting them to concepts such as soil health or soil security (Brevik et al., 2020). These are concepts that already have a level of acceptance among non-soil scientists or are connected to concepts such as energy, food, and water security that have wider acceptance. In addition, they have direct, easily understood links to soil and human health.

So, to understand soil and human health links adequately we need to find ways to account for changes in multiple variables that are all acted on by different parts of the overall environment, from air to biota to soil and continuing to water. This will involve approaching scientific study of the problems encountered in non-traditional ways. We need to engage multiple disciplines that have not traditionally worked together and facilitate communications between them; each disciplinary expert needs to have full understanding of what the others are saying and be able to work with them at a high level. Beyond this, we need to engage with the public and communicate the importance of what is being done to build support for research on soil and human health. This is truly a Grand Challenge! However, the rewards for success will also be substantial. Many of the connections between soil and human health can lead us to better nutrition, avoidance of pollutants, enhanced allergy resistance, and the like. Therefore, if we can improve our knowledge of the complex links within the soil–human health nexus that should allow us to improve health with less reliance on, for example, medications and medical procedures, in other words, to enhance our ability to practice preventive medicine.

References

- Alloway BJ (2008) *Zinc in Soils and Crop Nutrition*, 2nd edn Paris, France: International Zinc Association, Brussels; International Fertilizer Industry Association.
- Barka EA, Vatsa P, Sanchez L, Gaveau-Vaillant N, Jacquard C, Klenk H-P, Clément C, Ouhdouch Y, and van Wezel GP (2016) Taxonomy, physiology, and natural products of *Actinobacteria*. *Microbiology and Molecular Biology Reviews* 80: 1–43.
- Blum WEH, Zechmeister-Boltenstern S, and Keiblinger KM (2019) Does soil contribute to the human gut microbiome? *Microorganisms* 7: 287. <https://doi.org/10.3390/microorganisms7090287>.
- Brevik EC and Burgess LC (eds.) (2013) *Soils and Human Health*. Boca Raton, FL: CRC Press.
- Brevik EC, Khaleedian Y, and El-Ramady H (2021) Assessing the complex links between soils and human health: An area of pressing need. *Frontiers in Soil Science* 1: 731085. <https://doi.org/10.3389/fsoil.2021.731085>.
- Brevik EC and Sauer TJ (2015) The past, present, and future of soils and human health studies. *Soil* 1: 35–46.
- Brevik EC, Slaughter L, Singh BR, Steffan JJ, Collier D, Barnhart P, and Pereira P (2020) Soil and human health: current status and future needs. *Air, Soil, and Water Research* 13: 1–23.
- Clark JD (1969) *Kalambo Falls Prehistoric Site*. vol. 1. Cambridge: Cambridge University Press.
- Fairweather-Tait SJ, Bao Y, Broadley MR, Collings R, Ford D, Hesketh JE, et al. (2011) Selenium in human health and disease. *Antioxidants & Redox Signaling* 14: 1337–1383.
- FAO (1996) *Report of the World Food Summit*. Rome: Food and Agriculture Organization of the United Nations.
- Fuge R (2013) Anthropogenic sources. In: Selinus O, Alloway B, Centeno JA, Finkelman RB, Fuge R, Lindh U, and Smedley P (eds.) *Essentials of Medical Geology*, pp. 59–74. Dordrecht: Springer.
- Johns T and Duquette M (1991) Detoxification and mineral supplementation as functions of geophagy. *American Journal of Clinical Nutrition* 53: 448–456.
- Keith AM, Schmidt O, and McMahon BJ (2016) Soil stewardship as a nexus between Ecosystem Services and One Health. *Ecosystem Services* 17: 40–42.
- Kemper KJ and Lal R (2017) Pay dirt! Human health depends on soil health. *Complementary Therapies in Medicine*. <https://doi.org/10.1016/j.ctim.2017.04.005>.
- Madeshiya A, Banerjee P, Santra S, Ghosh N, Karmakar S, Bagchi D, Roy S, and Das A (2021) Role of metals and metalloids in redox biology. In: Bagchi D and Bagchi M (eds.) *Metal Toxicology Handbook*, pp. 3–12. Boca Raton: CRC Press.
- McMillian RJ (2002) Malnutrition. In: Katz SH and Weaver WW (eds.) *Encyclopedia of Food and Culture*, pp. 431–435. New York: Scribner.
- Oliver MA and Gregory PJ (2015) Soil, food security and human health: A review. *European Journal of Soil Science* 66: 257–276.
- Otto CC and Haydel SE (2013) Exchangeable ions are responsible for the in vitro antibacterial properties of natural clay mixtures. *PLoS One* 8: e64068.
- Ratnaddass A, Michellon R, Randriamanantsoa R, and Séguin L (2006) Effects of soil and plant management on crop pests and diseases. In: Uphoff N, Ball AS, Fernandes E, Herren H, Husson O, and Laing M, et al. (eds.) *Biological Approaches to Sustainable Soil Systems*, pp. 589–602. Boca Raton, FL: Taylor & Francis Group.
- Redshaw CH, Stahl-Timmons WM, Fleming LE, Davidson I, and Depledge MH (2013) Potential changes in disease patterns and pharmaceutical use in response to climate change. *Journal of Toxicology and Environmental Health, Part B* 16(5): 285–320. <https://doi.org/10.1080/10937404.2013.802265>.
- Roberts CA and Manchester K (2007) *The Archaeology of Disease*, 3rd edn Ithaca, New York: Cornell University Press.
- Rosen G and Rosen G (1993) *A History of Public Health*. Baltimore: The Johns Hopkins University Press.
- Selinus O, Alloway B, Centeno JA, Finkelman RB, Fuge R, Lindh U, and Smedley P (eds.) (2013) *Essentials of Medical Geology*, Revised Edition Dordrecht: Springer.
- Szostek K and Glab H (2001) Trace elements in human teeth from a Neolithic common grave at Nakonowo (Central Poland). *Variability and Evolution* 9: 51–59.
- Tountas Y (2009) The historical origins of the basic concepts of health promotion and education: the role of ancient Greek philosophy and medicine. *Health Promotion International* 24: 185–192.
- Ullah I, Ditta A, Imtiaz M, Mehmood S, Rizwan M, Rizwan MS, Jan AU, and Ahmad I (2020) Assessment of health and ecological risks of heavy metal contamination: a case study of agricultural soils in Thall, Dir-Kohistan. *Environmental Monitoring and Assessment* 192: 786.
- Waksman SA (1952) *The Literature on Streptomycin 1944–1952*. New Brunswick, NJ: Rutgers University Press.
- Wall DH, Nielsen UN, and Six J (2015) Soil biodiversity and human health. *Nature* 528: 69–76.
- Wellington EMH, Boxall ABA, Cross P, Feil EJ, Gaze WH, Johnson-Rollings AS, et al. (2013) The role of the natural environment in the emergence of antibiotic resistance in Gram-negative bacteria. *Lancet Infectious Diseases* 13: 155–165.
- White PJ and Broadley MR (2009) Biofortification of crops with seven mineral elements often lacking in human diets – iron, zinc, copper, calcium, magnesium, selenium and iodine. *New Phytologist* 182: 49–84.

Relevant websites

- <https://www.epa.gov/superfund/soil-bioavailability-superfund-sites-human-health>—EPA
- <http://www.fao.org/3/cb4894en/online/src/html/copyright.html>—FAO and UNEP.
- <https://rodaleinstitute.org/blog/10-ways-to-connect-soil-and-human-health/>—Rodale Institute.
- <https://soilhealthinstitute.org/wp-content/uploads/2019/03/SH-HH-post-conference-report-Final-030519.pdf>—Soil Health Institute.
- <https://apps.who.int/iris/bitstream/handle/10665/43894/97892?sequence=1>—WHO.
- <https://www.who.int/news-room/fact-sheets/detail/lead-poisoning-and-health>—WHO.

Soil Science Education, Awareness and Outreach in the Modern World

Laura Bertha Reyes-Sánchez^a and Eric C Brevik^b, ^aAgricultural Engineering Department, FES-C, National Autonomous University of Mexico, Mexico City, Mexico; ^bCollege of Agricultural, Life, and Physical Sciences, Southern Illinois University, Carbondale, IL, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	572
Important concepts in soil science education	573
Modes of educational delivery	574
Outreach by professional groups	574
International Union of Soil Sciences	574
Latin-American Soil Science Society	574
Soil Science Society of America and others	575
Museums as educational spaces	575
Groups that soil science educators interact with	576
The importance of primary and secondary education	576
To teach is not the same as to educate	576
How do we educate to build values that ensure respect for the soil?	576
What should we do?	578
Working with children	578
Opening the professional soil science curriculum to inter- and trans-disciplinarity	578
Provide adequate support and training to those who teach soil science	578
Concluding statements	578
References	579

Abstract

Soil science education takes place in the formal classroom, laboratory, and extension and outreach settings. It can be provided by professionally trained teachers in schools and universities, individual scientists reaching out to the public, or professional societies looking to expand knowledge of the discipline. Regardless of the form, setting, or identity of those delivering this education, it is important to understand that teaching involves more than just the delivery of discipline-specific content. Soils knowledge is used to address a wide range of modern problems. Inter- and trans-disciplinary thinking, communications, and interpretative skills are also needed to address global challenges.

Key points

- Why soil science?
- Why we need to change?
- To teach does not mean to educate.
- How should we educate to create soil preservation values that move us towards sustainability?
- How should we change soil science teaching to guarantee life on the planet?
- Why start at the primary education level?
- What do we propose to do?

Introduction

There are many reasons to educate people in soil science. Without fertile soil, there are no plants or animals for food, no clean water to drink, no oxygen to breathe, that is, no life or development to reach or sustain. Furthermore, soil provides fiber, fuel, and building materials, all of which make soil essential to human health and wellbeing (Brevik et al., 2019). Soil is the natural habitat of an immense number of organisms and microorganisms, and therefore, a source of biodiversity; it filters, recirculates, and defines the distribution of water; it is a waste deposit; and an excellent catalyst in green chemical reactions and therefore a real path to reduce pollutants in the environment (Reyes-Sánchez, 2012a). Soil is a natural resource, but also a social, economic, cultural, political and patrimonial good of mankind (Reyes-Sánchez, 2018). Soil loss influences poverty, displacement, inequality, violence, and injustice as a result of famine; therefore, it is essential that we study and preserve soil (Reyes-Sánchez, 2012b). Education is very important in the preservation of soil, because “that which is not known is not valued and therefore will not be protected or cared for” (Fig. 1).

Soil science education takes place in multiple ways, including formal classroom settings with a variety of delivery methods, internships, outreach to groups ranging from children to adults, and many other modes (Brevik et al., 2020; Field et al., 2017).

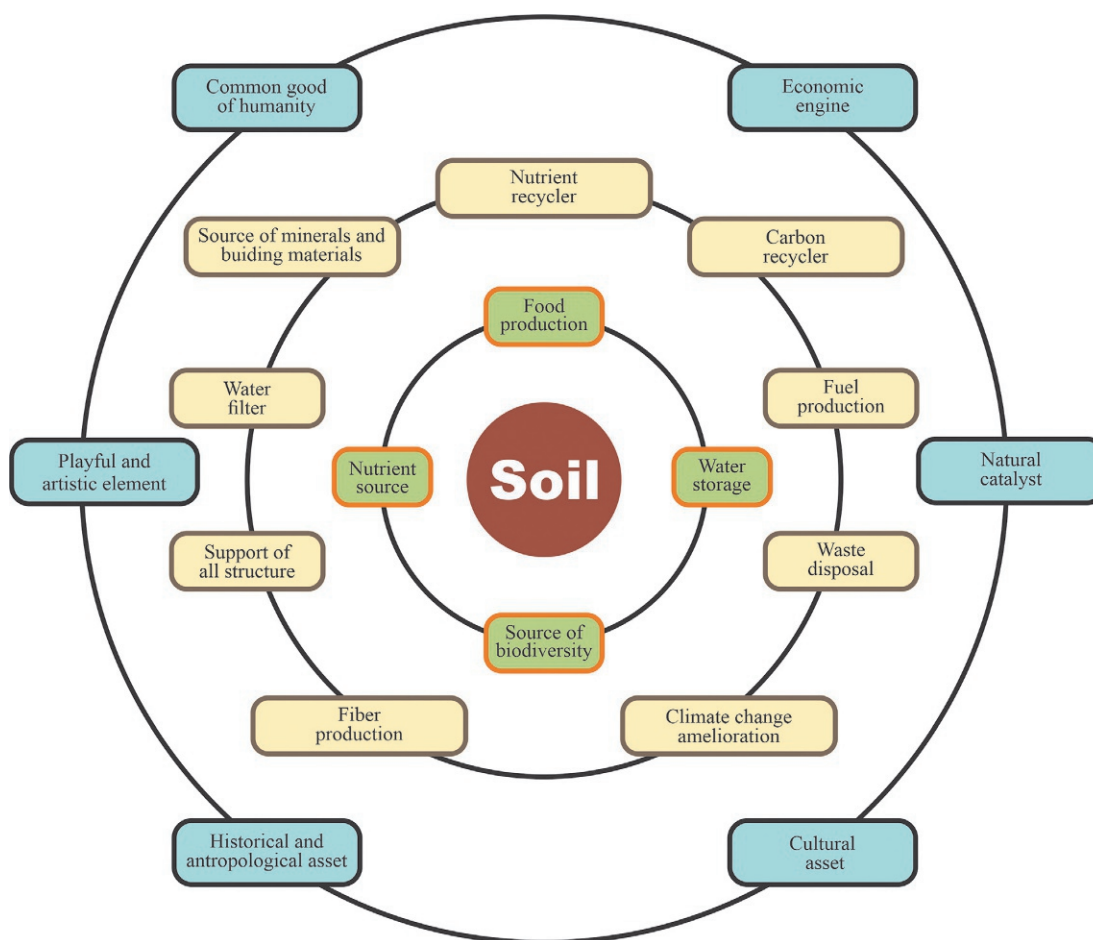


Fig. 1 The importance of soil to all.

Education can be provided by professionally trained teachers, professional societies, or others who have an interest in the topic. In addition, delivery modes change over time along with new technologies and the interests and experiences of students and teachers. The COVID-19 pandemic also created change in educational approaches to soil science education (Mahler et al., 2021).

This chapter will explore important concepts in soil science education, some of the more common modes of education delivery, groups that soil science educators commonly reach out to, and soil science professional groups that engage in public education. There will also be some discussion of ways that the COVID-19 pandemic affected soil science education and outreach, and how that may lead to future changes.

Important concepts in soil science education

All disciplines have a certain core body of knowledge that forms the unique foundation of that discipline. Churchman (2010) identified five subdisciplines that form the foundation of soil science: (1) pedology, (2) soil biology, (3) soil chemistry, (4) soil mineralogy, and (5) soil physics. If these subdisciplines are accepted as being essential to understand soil science, training in them and the supporting coursework to understand them is essential in soil science education (Brevik et al., 2020).

However, soil science education needs to go beyond this. Modern soil science has become interdisciplinary and even transdisciplinary in its nature. This makes it necessary to expand our vision for soil science education to include coursework that provides inter- and trans-disciplinary values, thoughts, procedures, and actions that will allow graduates to address social, economic, legal and environmental problems to achieve true development for all (Reyes Sánchez, 2016). Another way to address this is through academic-industrial partnerships that expose students to complex real-world problems as a way to develop the students' critical thinking skills and expose them to a variety of perspectives (Field et al., 2017).

Understanding the complexity of soil science and the issues it addresses requires understanding that soil is itself a system. Therefore, if we want to explain adequately the world in which soil scientists work, which is complex, integrated, interdependent, interdisciplinary, and transdisciplinary, our teaching should also mirror this. It needs to include the fields of applications and experiences that we seek to exert both within the school context, from preschool to doctorate, as well as in broader society.

Modes of educational delivery

As early as 1657, Juan Amos Comenius identified the need to teach using a comprehensive and unitary educational system. This system would provide a logical structure that allowed teachers to facilitate satisfactory learning, which in turn would influence thinking and transform it for the benefit of the individual and society (Comenio, 2000).

Science, including soil science, needs to be taught by guiding students to think and work as scientists, whether this involves teaching in the classroom, the field, or the laboratory. Whether addressing children, teenagers, postgraduate students, or other stakeholders, this needs to be congruent with the concepts and methods of science. Questions need to be chosen that define progressive hypotheses or research questions that address the problem that we intend to solve, with methods then proposed to solve the problem. Therefore, science teaching as a scientific activity should have 'a goal', 'a method', and 'an application field' appropriate to the school context that allows students to arrive at answers to the research problems that arise from observation and definition (Reyes-Sánchez, 2018).

A wide array of delivery modes are available for soil science education. The most common delivery methods are traditional lectures, laboratory analyses, and field trips (Krzić et al., 2018). Other approaches that are documented in the soil education literature include soil judging (Smith et al., 2020), online delivery of course materials (Ulery et al., 2020), open-source textbooks (Moorberg, 2020), and flipped classrooms (Ramírez II et al., 2022). Any of these can be formatted to provide the scientific teaching approach advocated above, but they have to be developed with that goal in mind, and it has been suggested that active learning approaches are the most effective (Smith et al., 2020).

The COVID-19 pandemic that began in 2020 represented a serious challenge to soil science education. Teachers around the world were forced to move their classes to distance delivery at very short notice, and in many cases this approach to soil science classes had never been given serious consideration (Mahler et al., 2021). Delivery modes prior to the COVID pandemic covered all options previously discussed, ranging from traditional lectures to highly active learning approaches, and each of these presented their own challenges. The journal *Natural Sciences Education* recently published a special issue that presented and analyzed the approaches taken by several soil science teachers to address the COVID-19 emergency (Mahler et al., 2021). Students generally indicated that they preferred face-to-face activities to the distance learning options (Wyatt, 2021), and they were less engaged in their learning. Synchronous approaches did show more engagement with the subject than asynchronous approaches (Walker and Koralesky, 2021). Many instructors developed online materials during this time that they plan to incorporate into future classes, albeit not always used exactly as they were during the COVID-19 emergency, because they feel these materials will enhance their future classes (Brown and Krzić, 2021). It will be important to consider carefully the lessons learned during COVID-19 and how they can help improve and expand soil science education in the future (McCauley, 2021).

Outreach by professional groups

Many professional soil science societies engage in educational outreach efforts at multiple levels. This section will not attempt to provide a comprehensive listing of these professional groups and their efforts, but will provide some examples.

International Union of Soil Sciences

The International Union of Soil Sciences (IUSS) has called for explicit recognition that fertile soil is essential for all agricultural practices that guarantee food security, and that soil supports all human, social, and ecosystem functions (IUSS, 2002). Consequently, the IUSS called on everyone to commemorate World Soil Day every December 5. On December 5th, 2015, the IUSS launched the "International Decade of Soils 2015–2024" (IUSS–IDS) with the Vienna Declaration (<https://www.iuss.org/international-decade-of-soils/>). The two main objectives of IUSS–IDS are "Stop land degradation as the most insidious and underestimated challenge of the 21st century, and direct the main focus of all activities on school age children and young adults who will be teenagers and young adults in 10 years time" (Hom, 2015). Therefore, interdisciplinary soil science education, awareness, and outreach is a fundamental goal of the IUSS, both to face the environmental challenges of today and to be able to confront those that will occur in the future.

The IUSS has addressed this with many recent educational projects, including a children's book project that was run jointly with the Food and Agriculture Organization of the United Nations (FAO) in 2020 (<http://www.fao.org/world-soil-day/bookcontest/en/>) and 2021 (<https://www.fao.org/world-soil-day/concurso-de-libros/en/>), publishing the book "Soil Science Education: Global Concepts and Teaching" (Kosaki et al., 2020), and "The IUSS Goes to the School" education project (<http://www.iuss-goes-to-school.org.mx/>). The IUSS website for this education project provides links to books, videos, experiments, and other educational resources to help children learn about soil. Some of these resources are available in multiple languages. The IUSS financially supports the education and dissemination initiatives of its national soil science societies through its stimulus fund program.

Latin-American Soil Science Society

The Latin-American Soil Science Society (SLCS; Sociedad Latinoamericana de la Ciencia del Suelo) has collaborated with its member countries since 1987 developing the education project "Thus are the Soils of my Nation" (<https://www.slcs.org.mx/index.php/educacion-y-ensenanza/asi-son-los-suelos-de-mi-nacion>), a project to introduce students to the soils found in their home country. This is done using children's books, plays, games, educational aids, and training for teachers. "Thus are the soils of my Nation" works with primary school children using the school orchard as the center of their activities to achieve the main

objective: “that children understand that the soil is an indispensable natural resource to preserve life on Earth” (<https://www.slcs.org.mx/index.php/fotos>). The School Orchard Program in educational centers for the development of environmental and sustainable competencies of teachers and students is promoted by teachers, national soil science societies, FAO, and NGOs (<https://educacio-valencia.es/es/proyectos/el-huerto-escolar-como-herramienta-educativa/>) (<https://www.bodeninfo.net/projekte/bodenmacht-schule/>) (https://www.euskadi.eus/contenidos/documentacion/inn_doc_ed_ambiental/es_def/adjuntos/800001c_huerto_escolar_c.pdf; <https://lifelab.org/> <https://edibleschoolyard.org/>).

An indispensable part of “Thus are the soils of my Nation” is the “Symposium of Educative Innovations on Teaching and Learning Soil Science” as a permanent space open to the expression and presentation of children’s research within the congresses of the Latin American Society of Soil Science (<https://www.slcs.org.mx/index.php/educacion-y-ensenanza/simposio-de-innovaciones-educativas-en-la-ensenanza-de-la-ciencia-del-suelo>). Another important part of this Latin American collaboration is the multi-language glossary of soil science terms (<https://cit.iec.cat/DMCSE/default.asp?opcion=0>).

Soil Science Society of America and others

The Soil Science Society of America (SSSA) has extensive resources for kindergarten to twelfth grade teachers (K–12 teachers) at <https://www.soils4teachers.org/>. These include the ability to ask a soil scientist a question, a soil science glossary, pre-developed lessons and activities that are aligned with USA education standards, posters, webinars, videos, and other materials. The SSSA also developed monthly educational videos as part of the Year of Soils in 2015, and these can be found at <https://www.soils.org/iys/monthly-videos/>. Many of these SSSA resources have been available for over a decade as of 2022 and new resources continue to be added.

Some other worldwide education resources available from professional societies include the British Soil Science Society materials for primary schools (http://www.soil-net.com/downloads/soilsafari_teachernotes.pdf) and education supports of the French Association pour l’Etude des Sols: <https://www.afes.fr/ressources/bibliotheque> or FNEassociation (<https://www.fne-aura.org/notre-offre/rhone/nos-kits-pedagogiques-a-decouvrir/>).

Museums as educational spaces

Important educational efforts have been made worldwide through museums with outstanding collections of soil profiles and their corresponding descriptions (Table 1).

Table 1 Museums with soil profile collections.

Country	Museums
Argentina	Universidad Nacional de Río Cuarto, Instituto Nacional de Tecnología Agropecuaria.
Australia	The University of Sydney (Australian Technology Park).
Austria	University of Vienna.
Belgium	KU Leuven.
Brazil	Universidade Federal de Lavras, Universidade Federal de Roraima, Universidade Federal de Santa Maria.
Canada	University of Alberta, University of British Columbia, Great Lakes Forestry Center (Ontario).
Colombia	Museo de Suelos Ciro Molina Garcés, Museo de Suelos Instituto Geográfico Agustín Codazzi.
China	China Soil Museum; The Modern Soil Monolith Exhibition Center.
Estonia	Soil Museum Estonian University of Life Sciences.
Germany	Museum für Bodenschätzung in Eickendorf (Börde).
India	Kerala Forest Research Institute.
Indonesia	Museum Tanah (Bogor Soil Museum).
Japan	Natural Museum of History and Science, Natural Resource Inventory Museum, Tsuchino-Yakata, Hokkaido.
Netherlands	World Soil Museum, Wageningen University, VU University, Rijksuniversiteit Groningen, HAS Hogeschool, VHL University of Applied Sciences, Museonder, Geologisch Streekmuseum ‘de IJsselvallei’.
Peru	Museo de Suelos.
Russia	Vasily Dokuchaev Museum of Soil Science, St Petersburg; Williams Museum of Soil and Agriculture, Moscow.
Spain	Universidad de Murcia, Universidad de Granada, Insititu Cartogràfic de Catalunya.
Taiwan	National Taiwan University, Research Institute.
Thailand	Soil Museum Bangkok.
United Arab Emirates	Emirates Soil Museum.
United Kingdom	Cranfield University, Natural History Museum, London
United States of America	Kansas State University, University of Idaho, Texas A&M, Virginia Tech, West Virginia University, University of Georgia, Smithsonian’s National Museum of Natural History (2008–2009), Cayuga Nature Center (NY), Diné College (AZ), American Museum of Natural History (NY), St. Louis Science Center.
Uzbekistan	State Research Institute of Soil Science and Agrochemistry.

Adapted from Stoof CR, Candel JHJ, van der Wal LAGM, Peek G (2019) Soil lacquer peel DIY: Simply capturing beauty. *The Soil* 5: 159–175.

Groups that soil science educators interact with

Soil science education can take many forms and be done with many different groups of students. These range from children in primary and secondary schools to undergraduate and post-graduate students in colleges and universities. But it also includes farmers and ranchers, in settings often referred to as extension, and outreach to the public. Citizen science projects have become popular over the last decade, and several that involve soils have been undertaken (e.g., Sandén et al., 2020). In this context, the FAO launched the Global Soil Doctors Programme in 2020. This is a farmer-to-farmer training initiative being implemented at the global level on a voluntary participation basis and centers attention on the importance of soil as a vital resource for farmers, policy makers, development planners, agricultural extension workers, NGOs, private sectors, and stakeholders (<https://www.fao.org/global-soil-partnership/pillars-action/2-awareness-raising/soil-doctor/es/>). This is being done in collaboration with soil scientists, soil science societies, and universities, with an emphasis on Bangladesh, Malawi, and Mexico (https://www.fao.org/global-soil-partnership/resources/highlights/detail/en/c/1470466/?fbclid=IwAR0qmaCfdaiWFbDlG_GSPpWj6qTqYpzGb3IY0ofhokzCYB0MlqmBqbqR8VA).

The importance of primary and secondary education

*Like the soil, no matter how rich it may be, cannot bear fruits if it is not cultivated
The mind without cultivation cannot produce either*
Seneca

According to the main goals of the IUSS “International Decade of Soils: 2015–2024”, the IUSS aspires to be a leader in the process of educating our future scientists, as well as in generating public awareness of soil sustainability (Horn, 2015). Educating children and teenagers is a priority for the IUSS because the future of soil, and therefore of life on the planet, will be in their hands. The responsibility to ensure that children and teenagers understand the value of the soil resource, and commit to its preservation, is ours. We need to teach children and teenagers why the soil is important in their daily lives.

To teach is not the same as to educate

*“It produces an immense sadness to think that nature speaks
while humankind does not listen to her”*
Victor Hugo

When UNESCO (1971) defined environmental education as “the cognitive process of recognition of values and clarification of concepts to promote the necessary attitudes to understand and appreciate the interrelationships between man, his culture, and his biophysical environment”, it made clear the intrinsic relationship between the education process and human construction of environmental awareness and culture. However, lack of knowledge about the value that natural resources in general, and soil resources in particular, have in our lives has not encouraged citizens to create a culture of environmental preservation that is manifested in their daily actions. The increased urbanization of society, and its detachment from the food production process, means that a significant proportion of the human population lacks a fundamental understanding of soil as an invaluable resource, its functions, and the benefits and services (including food production) received from soil. Therefore, it is important to build citizen awareness about the value of the soil resource (Brevik et al., 2018).

Teaching is not the same as educating. Educating entails the explicit and implicit formation of values that emotionally and cognitively implant the information transmitted, thus making possible the progressive formation of a conscience that in the medium- and long-term builds, little by little, a culture of environmental consciousness (Reyes-Sánchez, 2012a). It is about the contents of all we teach being taught with the goal of preserving the soil for sustainability.

When we teach soil science, we are transmitting basic and specialized information at various levels. We teach how to use that knowledge in theory and practice and how to use that information for interpretations, corrections, and projections. Unfortunately, that does not create an environmental consciousness or culture for preservation that supports the soil’s sustainability. Knowledge accumulates when we teach and it is not possible to transmit what you do not know. However, public awareness requires education. To reach a position where soil preservation becomes important, we need to do more than teach people about soil. We need to educate them to preserve the soil resource because it is indispensable for life itself. To do this it is necessary to explain to people why soil is essential for them and why soil loss affects us all, regardless of the social class to which we belong and regardless of the location we live in (rural or urban, continent we live on, etc.). “Soil preservation is a common good of humankind” (Declaración de Mar del Plata, 2012).

How do we educate to build values that ensure respect for the soil?

*“Educating is more difficult than teaching
To teach you only need to know, but to educate, you need to be”*
Albert Einstein

According to the pedagogical perspective of Antón Makarenko (1959), the teaching of science through education must harmonize social interests with the particular interests that students have, involving them in the search for solutions to everyday problems. For Arca et al. (1990), learning science implies learning to change the ways one sees phenomena, reasoning, speaking, and emotions; taking account of these simultaneously. Gómez and Reyes-Sánchez (2004) argue that education to preserve natural resources to advance towards sustainability is an objective that goes beyond one subject or a set of them in the curriculum. They posit that education is not about reproducing actions focused on techniques, but rather an invitation to debate and reflect on the type of technology and social organization that allows people to live in harmony with each other and the natural environment. Creating environmental awareness in society involves education to change individual and collective behaviors (something done with social marketing; Brevik et al., 2018). This implies a transformation that affects the personal lifestyle as seen in consumption, health, civic attitudes, and equality, and prepares whoever is educated in this way to exercise democracy as a way to create the same scale of values. By teaching and learning science in this way, we form values, we form a consciousness, and we build a preservation culture (Fig. 2).

All these approaches indicate that we require a change of perspective globally. Although science can help, pollution, devastation, degradation, and loss of soil resources are not problems that can be solved technologically given the prevailing values of competition, lack of solidarity, exclusion, uniformity, or hierarchy, governed by the global market (Gómez and Reyes-Sánchez, 2004; Gutiérrez, 1995; Gil and Vilches, 2005; UNESCO, 2009).

However, we can pretend to teach values as simple theoretical entities and even impose them from a decontextualized and rote learning perspective, in which the purpose is repetition, oral or written, as a synonym for learning. This is neither pedagogically nor environmentally desirable. What is needed is to inform humans with the knowledge and awareness that allows them to assume different ways of using soil and new forms of coexistence between humans and the soil resource (Reyes-Sánchez, 2006).

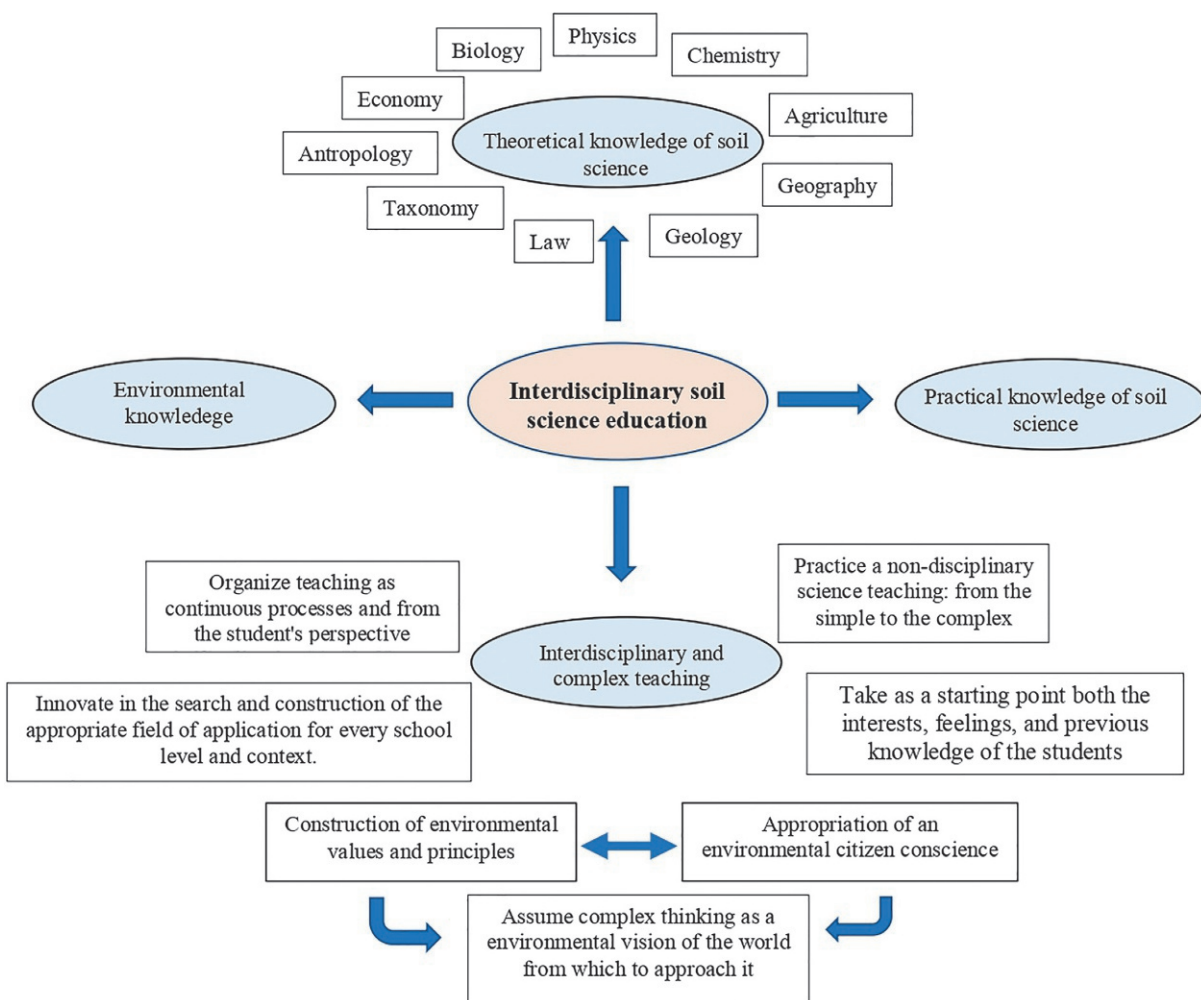


Fig. 2 Interdisciplinary soil science education has no borders and neither does its teaching.

What should we do?

*We propose to educate in soil science on a planet that is a natural system
and work together as a team to put the gaze
and concern of the world on soil*

The following points are proposed to create an enhanced global level of soil education: (a) work with children and teenagers to educate them about and create an awareness of soils and their importance, (b) open the professional soil science curriculum to inter- and trans-disciplinary research and teaching, and (c) provide adequate support and training to those who teach soil science.

Working with children

We, today's soil scientists, can decide what soil science knowledge to impart, how, and to whom. It is up to us to decide if we should continue to limit the teaching of soil science to highly specialized professionals, or if we are willing to make a personal commitment to make soil education accessible to students who are children and teenagers who are practically blank pages, to provoke their interest, enthusiasm, and love for soil (Reyes-Sánchez, 2020). This would involve researchers and university professors working with teachers at pre-university levels to create a new paradigm in the teaching of soil science. As shown in the "Outreach by professional groups" section, many soil scientists are already taking steps in this direction. This is important work that needs to continue.

Opening the professional soil science curriculum to inter- and trans-disciplinarity

The disciplinary subdivisions of science, including soil science, have their value. But we cannot fail to recognize that in reality there are no simple or isolated problems, but many complex problems that face our modern world. Therefore, it is not possible to address these problems only from the isolated view of the disciplinary divisions. We need to work to build, through inter- and trans-disciplinary integration from various fields of knowledge, specialized knowledge in soil science that allows us to address complex problems and propose viable solutions. This includes multicausal and interdependent perspectives. It is important for the modern soil scientist to be exposed to fields ranging from biology, chemistry, and physics to communications, economics, the law, and sociology (Brevik et al., 2020).

The United Nations Sustainable Development Goal (SDG) 4 is "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all". An interdisciplinary practice and teaching of soil science could help construct an education that advocates the search for solutions that are culturally and ethically acceptable, economically feasible, energetically desirable, and environmentally respectful of the balance of ecosystems as the foundation of a socially responsible science. This would environmentalize¹ all knowledge through a continuous methodological process that allows both building the knowledge that children and teenagers have from the simple to the complex, and pedagogically organizing all actions in an equally processual way. This would establish stages of knowledge construction, with increasing levels of complexity in terms of relationships and interrelations, and thus closer to a true perception of reality. Approaching soil in an inter- and trans-disciplinary way is important in understanding that soil is itself a system. When we talk about development goals, we talk about the development and sustainability of natural and social systems. These systems are open and dissipative, in a constant state of change (Reyes-Sánchez, 2012b).

Provide adequate support and training to those who teach soil science

During the Tbilisi in 1987, Rio in 2000, and Johannesburg in 2002 summits, both the UN and UNESCO emphasized the importance of training teachers whose activities and decisions significantly influence the education of future citizens and prepare them to meet the challenge of responding to present and future problems. Achieving this implies empowering and sensitizing people so they recognize there are no pre-established rules that are always valid and they know how to make decisions that will lead to soil preservation. It also implies ensuring that people can change their behaviors according to circumstances, publicly defend their opinions and recognize those of others, engage in dialogue, and be able to negotiate (Gómez and Reyes-Sánchez, 2004). Therefore, among the primary characteristics to develop in teachers so that they can transfer them to students are the development of values and attitudes. In this way future citizens will be capable of making and participating in all kinds of decisions with solid arguments based on knowledge and manifested firmly through reflective attitudes, but above all, based on analysis carried out in the light of the values formed.

Concluding statements

Education for soil sustainability requires that it is participatory, effective, multicultural, dialogic, democratic, investigative, interdisciplinary, and activist. It is important to teach children, teenagers, and adults not just to be content with knowledge, but also how to be competent workers, reflective citizens, social critics, and agents of change (Reyes-Sánchez, 2020). This will be required if we are

¹A methodological process that approaches the environmental problems in a cognitive and interdisciplinary way through education and the practice of science with the objective of building a preservation culture and conscience by providing models for the management and exploitation of natural resources towards the development for all (Reyes-Sánchez, 2012b).

to have a society committed to sustaining the soil resource. It will be important for soil scientists, educational institutions, and professional societies to commit to such an educational vision. The national soil science societies in countries around the world can join with each other and the IUSS to train future citizens of the world in the knowledge and environmental awareness necessary to achieve this.

References

- Arca M, Guidoni P, and Masón P (1990) *Enseñar ciencia*. Barcelona: Paidós/Rosa Sensat.
- Brevik EC, Steffan JJ, Rodrigo-Comino J, Neubert D, Burgess LC, and Cerdà A (2018) Connecting the public with soil to improve human health. *European Journal of Soil Science* 70: 898–910.
- Brevik EC, Peregr L, Pereira P, Steffan JJ, Burgess LC, and Gedeon CI (2019) Shelter, clothing, and fuel: Often overlooked links between soils, ecosystem services, and human health. *Science of the Total Environment* 651: 134–142.
- Brevik EC, Krzic M, Itkin D, Uchida Y, and Chau HW (2020) Guidelines for under- and post-graduate students. In: Kosaki T, Lal R, and Sánchez LBR (eds.) *Soil Sciences Education: Global Concepts and Teaching*, pp. 31–48. Stuttgart: Schweizerbart Science Publishers.
- Brown S and Krzic M (2021) Lessons learned teaching during the COVID-19 pandemic: Incorporating change for future large science courses. *Natural Sciences Education* 50: e20047.
- Churchman GJ (2010) The philosophical status of soil science. *Geoderma* 157: 214–221.
- Comenio JA (2000) *Didáctica Magna*. México City: Porrúa.
- Declaración de Mar del Plata. 2012. *Latin-American Soil Science Society*. Accessed at: <http://slcs.org.mx/index.php/es/informacion-general/declaraciones/8-mar-del-plata>
- Field DJ, Yates D, Koppi AJ, McBratney AB, and Jarrett L (2017) Framing a modern context of soil science learning and teaching. *Geoderma* 289: 117–123.
- Gil PD and Vilches A (2005) Década de la educación para el desarrollo sostenible. Algunas ideas para elaborar una estrategia global. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias* 2(1): 91–100.
- Gómez MM and Reyes-Sánchez LB (2004) La educación ambiental, imprescindible en la formación de las nuevas generaciones. *TERRA Latinoamericana* 22(4): 515–522.
- Gutiérrez J (1995) *La Educación Ambiental: Fundamentos teóricos, propuesta de transversalidad y orientaciones extracurriculares*. Madrid: La Muralla.
- Horn R (2015) *IUSS Vienna Soil Declaration*. file:///C:/Users/SIU856519303/Downloads/vienna_soil_declaration_december_7_%20(1).pdf.
- IUSS (2002) *Proceedings XVII World Congress of Soil Science*. Bangkok, Thailand: IUSS.
- Kosaki T, Lal R, and Reyes-Sánchez LB (eds.) (2020) *Soil Science Sciences Education: Global Concepts and Teaching*. Stuttgart: Schweizerbart Science Publishers.
- Krzic M, Yates TT, Basiliko N, Pare MC, Diochon A, and Swallow M (2018) Introductory soil courses: A frontier of soil science education in Canada. *Canadian Journal of Soil Science* 98: 343–356.
- Mahler RL, Krzic M, Merkle BG, Moorberg C, and Brevik EC (2021) Natural sciences education in a COVID-19 world. *Natural Sciences Education* 50: e20067.
- Makarenko A (1959) *Poema Pedagógico*, 6ª edición México City: Progreso.
- McCauley DJ (2021) COVID-19 forced rapid changes in education, but which changes should we keep? *CSA News* 66(10): 44–47.
- Moorberg CJ (2020) An open annotated bibliography for soil and water conservation: A case study. *Natural Sciences Education* 49: e20014.
- Ramirez S II, Teten S, Mamo M, Speth C, Kettler T, and Sindelar M (2022) Student perceptions and performance in a traditional, flipped classroom, and online introductory soil science course. *Journal of Geoscience Education* 70: 130–141.
- Reyes Sánchez LB (2016) *La educación como política pública de concientización para la preservación del suelo como recurso natural limitante para la existencia de la vida, en el libro: "Redescubriendo al suelo y su importancia ecológica"*. FES-Zaragoza UNAM y CONACYT.
- Reyes-Sánchez LB (2006) La enseñanza de la ciencia del suelo en el contexto del desarrollo sustentable. *TERRA Latinoamericana* 24(3): 431–439.
- Reyes-Sánchez LB (2012a) Enseñanza de la ciencia del suelo: Estrategia y garantía de futuro. *Spanish Journal of Soil Science* 2(1): 87–99.
- Reyes-Sánchez LB (2012b) Aporte de la química verde a la construcción de una ciencia socialmente responsable. (Green Chemistry's contribution in the construction of socially responsible science). *Educación Química* 23(2): 222–229.
- Reyes-Sánchez LB (2018) Edaphological approaches to advancing Sustainable Development Goals: An educational perspective to build a citizen preservation culture. In: Lal R, Horn R, and Kosaki T (eds.) *Soil and Sustainable Development Goals of the U.N.*, pp. 139–155. Stuttgart: Schweizerbart Science Publishers.
- Reyes-Sánchez LB (2020) Educating to build a citizen preservation culture. In: Kosaki T, Lal R, and Sánchez LBR (eds.) *Soil Sciences Education: Global Concepts and Teaching*, pp. 49–58. Stuttgart: Schweizerbart Science Publishers.
- Sandén T, Spiegel H, Wenng H, Schwarz M, and Sameel JM (2020) Learning science during teatime: Using a citizen science approach to collect data on litter decomposition in Sweden and Austria. *Sustainability* 12: 7745.
- Smith CMS, Chau HW, Carrick S, van Dijk JL, Balks MR, and O'Neill TA (2020) Learning by doing is more memorable: The practice of pedagogically aligned learning in university level soil science in New Zealand. In: Kosaki T, Lal R, and Reyes-Sánchez LB (eds.) *Soil Sciences Education: Global Concepts and Teaching*, pp. 183–190. Stuttgart: Schweizerbart Science Publishers.
- Ulery A, Muise AS, Carroll KC, Chamberlin B, White L, Martinez P, Spears L, and Gleason J (2020) Impact of multimedia learning tools in agricultural science classes. *Natural Sciences Education* 49: e20011.
- UNESCO (1971) *Man and Biosphere Program*. <https://en.unesco.org/mab>
- UNESCO (2009) *Conferencia Mundial sobre Educación para el Desarrollo Sostenible*.
- Walker KA and Koralesky KE (2021) Student and instructor perceptions of engagement after the rapid online transition of teaching due to COVID-19. *Natural Sciences Education* 50: e20038.
- Wyatt B (2021) Insights into student participation in a soil physics course during COVID-19 emergency online learning. *Natural Sciences Education* 50: e20036.

Soils in landscape planning

Seth Denizen*, High Meadows Environmental Institute, Princeton, NJ, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	580
Turning the soil layer on: Two challenges for soil mapping	581
Making decisions: The politics of cause and effect in socio-ecological systems	584
After the collapse of the layer cake model: Rethinking the relationship between soil and land-use	586
Soil futures: A new discipline	586
References	587

Abstract

Soil often plays a key role in landscape planning. However, there are some key limitations to conventional approaches for incorporating soil data into planning models. The separation of soil and land-use layers in conventional planning maps has led to a systemic misunderstanding of the complex socio-ecological relationships that drive processes like land degradation and climate change induced migration. Failed conceptual models for how to imagine the relationship between soils and human environments should lead us to reconsider the way soils are made an object of planning in the 21st century.

Key points

- The desire to use soil as a resource shapes the way it is understood empirically.
- Soil-based disciplines are epistemologically diverse and produce soil knowledge that is not always mutually compatible.
- Layer-based planning maps traditionally obscure the socio-ecological relationships that are key to soil health and conservation
- More effective approaches to incorporating soil knowledge into planning processes will require each discipline to rethink its methods and values

Introduction

The majority of the world's soil resources are degraded. In technical terms, they are in fair, poor, or very poor condition according to the Food and Agriculture Organization's most recent inventory (Food and Agriculture Organization (FAO) and Intergovernmental Technical Panel on Soils (ITPS), 2015). Roughly half of all people on Earth live in landscapes degraded by erosion, salinization, compaction, acidification, chemical pollution, or subsidence and by 2050 between 50 and 700 million people will be forced to migrate as climate change further erodes the material basis of life in arid and low-lying regions (Montanarella et al., 2018). 20th century modes of urbanization and agriculture have had profound consequences for the world's soil resources. The realization that these consequences are best described as "degradation" should be cause for some concern, as the presence of anthropogenic soils are now a plausible stratigraphic indicator for our planetary geologic epoch (Certini and Scalenghe, 2011).

Landscape planning is a field that cuts across many of the processes that drive soil degradation. Processes such as urbanization, deforestation, and agriculture are all key concerns of landscape planning. However, the scope and scale of these concerns makes expertise in the subject highly distributed. There are environmental planners, landscape architects, geographers, land-use planners, foresters, civil engineers, urban ecologists, and landscape ecologists, all of whom claim some expertise in what could be summarized in a word: land. Rather than something intrinsic to the terrestrial Earth, 'land' is better understood as something that emerges as a byproduct of these forms of expertise (Li, 2014). In other words, land is what happens when soil is interpreted as use or value.

The implications of this for landscape planning are important and often misunderstood. There has been a tendency to believe that the way soil is valued has nothing to do with how soil is understood as an empirical object. Values, in this model, float above the fundamental science of what soil is and how it works, as if each mode of valuing soil simply represents a deliberately partial view of what is actually well-defined in an entity that is independent and autonomous of those values. In practice, this assumption leads to a great deal of confusion about how soil knowledge could be incorporated into the social and political process of planning.

When soil is understood as property, infrastructure, agricultural resource, or ecosystem service, it has been interpreted through the lens of a specific value. What is crucial to understand, however, particularly for people doing interdisciplinary work like planning, is that different ways of valuing the soil have produced different epistemologies for knowing the soil. Over time, the tendency to continually improve the methodology through which such a value is detected in the soil produces its own tools, techniques, terminology, and protocols. For example, the value of soil stability for foundations leads to the application of concepts

*Princeton-Mellon Fellow in Architecture, Urbanism, and the Humanities.

such as 'strain rate,' as well as the invention of tools and protocols to measure it in different soils, and then eventually to the building regulations that become codified as civil law protections for local and national governments.

Over the course of the 20th century, the fundamental importance of soil for all aspects of political and economic life has resulted in a radiation (to use the evolutionary term) of soil-related disciplines, each with their own expertise in detecting different values in the soil. Crucially for the planner, not all of this knowledge is mutually compatible. For example, the soil that geotechnical engineers describe deep under a building is, for agronomists, not soil at all. In this case, what is meant by the word 'soil,' differs considerably, as do the tools and techniques used to describe it. While these differences are clear to the researchers and practitioners in each respective soil-related discipline, planners are tasked with overlapping many different forms of expertise, often quite literally. In what follows, I will outline a brief conceptual history of how soil knowledge has been 'overlapped' with other forms of knowledge in 20th century planning procedures. This history will serve as an introduction to understanding the challenges soil poses to conventional planning paradigms, and how knowledge about the soil might be utilized more effectively in planning processes in the future.

Turning the soil layer on: Two challenges for soil mapping

Historically, the way in which knowledge about soil has been incorporated into planning models is through a very particular technology, which began as the map overlay. The map overlay has been the 20th century engine of landscape planning and its related disciplines, particularly as it developed as the conceptual basis of geographic information systems (GIS). In these models soil typically appears, unsurprisingly, as a layer in order to be overlapped with other kinds of geospatial information (Fig. 1).¹ The soil layer in landscape planning maps has long provided crucial insights about geotechnical stability, agricultural potential, landslide risk, forest quality, recreational value, septic suitability, and flood risk. The ability to compare land use against soil properties has appeared early and often in the decision-making trees used by landscape planners (Marsh, 2010).

However, soil is not easily reduced to a map layer, even when planning goals are very narrowly defined (Gordon and Gordon, 1981). The inherent three-dimensionality of soil profiles and their catenas across diverse landscapes and climate regimes has challenged soil scientists since the earliest days of the profession. It is also important to understand that these early days were not very long ago. The concept of soil as something that consists of layers that grow and are distinct from rock dates only to the 1880s. The challenge since this recent beginning has been to see patterns within the astonishing diversity of soils around the world and, crucially, to name these patterns. However, this problem of how to name the soil (also known as soil taxonomy) has historically been inseparable from the problem of how to map the soil. The World Reference Base (WRB) soil taxonomy is a good example of this close relationship between mapping and naming, as it began as the map legend to the FAO–UNESCO World Soil Map (Nachtergaele, 2003). This United Nations funded mapping project was the first and most important attempt to make soil knowledge available to planners, and its ambitions were vast (Selcer, 2018). The World Soil Map was intended to provide a definitive catalog of the world's soil resources, making it possible to plan soil on a global scale. However, to accomplish this goal, vastly different kinds of soil knowledge had to be reconciled. René (Rudy) Dudal, who was the lead author of the *World Soil Map*, describes different systems of soil classification as not just describing things differently, but as describing 'different realities.'

This work was undertaken primarily in an attempt to overcome the lack of correlation that has hitherto characterized the field of soil cartography, in which various countries had established their own nomenclatures, survey methods and systems of classification. In particular, those who favoured a genetic classification of soils found themselves in opposition to those who advocated a classification based on morphological criteria. This made any comparison or exchange of information extremely difficult. The point had been reached at which the nomenclatures used could not be translated because terms which at first glance appeared to have equivalents in other languages did not, in fact, cover the same realities; this entailed a very great risk of misunderstanding and practical error (Dudal and Batisse, 1978).

For Dudal and his team, what was untranslatable between systems of soil classification around the world was not language but values. Just as geotechnical engineers and agronomists classify soil in ways that are not only different but mutually exclusive (as we saw above), so too do rice farmers in the Mekong Delta and tax-collectors in Central Mexico. For global systems of soil taxonomy to emerge at all, a consensus had to emerge among diverse forms of expertise for which soil properties to privilege in the naming (and mapping) of soil. In this sense, what Dudal is describing above is not just the rectification of maps at different scales and projections, but the rectification of the values that would determine the tools, techniques, and protocols of soil naming at a global scale.

This kind of scientific consensus is still an ongoing project, but a key benefit of this procedure has already become apparent. Establishing a common language for communicating soil knowledge has made it possible to share knowledge between regions with similar soil conditions that are separated by large geographic distances. It has also facilitated a greater integration between botanical and pedological research, revealing new ecological relationships and agricultural opportunities. While such a consensus is also particularly helpful for interdisciplinary work like planning, the system has its hazards. In data collection terms, soils are not like topography, where the distance between data collection and mapping is very small or fully automated. The labor of interpretation that takes place between observing soil properties and mapping them is still significant and often costly. This labor of interpretation

¹Fig. 1 is an ideogram originally developed as a nationally distributed poster for a 1984 seminar led by Bernard J Niemann, Jr., entitled "Seminar on the Multipurpose Cadastre: Modernizing Land Information Systems in North America." The figure was also republished in Tulloch DL (1999) Theoretical Model of Multipurpose Land Information Systems Development. *Transactions in GIS* 3, (3): 259–283.

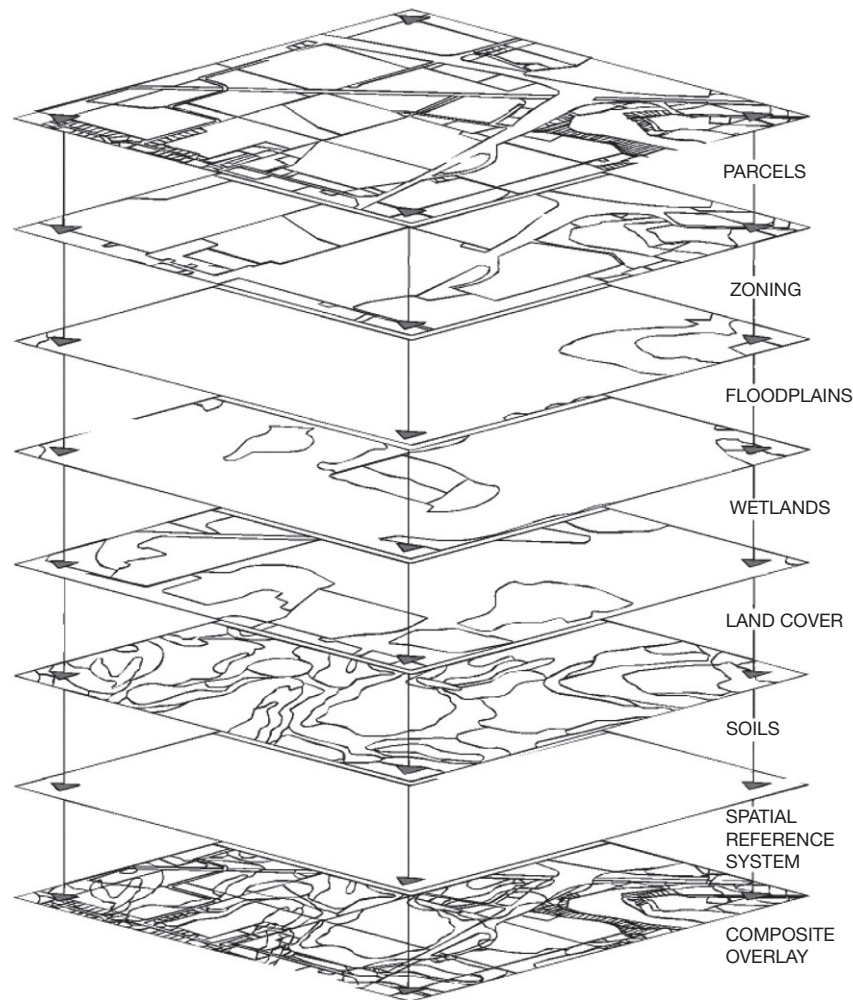


Fig. 1 The traditional map overlay method for landscape planning.

takes diverse forms around the world, and the assumption that this process results in an equivalent and continuous data set over large geographical areas has made soil mapping a deeply political exercise (Selcer, 2018).

The widespread assertion that so-called “tropical soils” have poor soil fertility turned out to be an error based in part on Eurocentric assumptions about plant nutrient cycling (Richter and Babbar, 1991). This Eurocentric bias within the structure of Western soil taxonomic systems created, and continues to create, poor policy decisions in tropical and subtropical regions where soils have formed at the intersection of vastly different socio-ecological contexts. For example, undervaluing soil fertility in the Colombian Amazon resulted in the exclusion of small agro-ecological producers from state support and financial credit, exacerbating their struggle to find alternatives to cocoa production amid the US–Colombia policy of anti-drug aerial fumigation (Lyons, 2020).

Arid landscapes have also been subject to similar errors in the evaluation of soil quality. A wide variety of landscapes outside temperate rainfall norms have been mislabeled as deserts, which in the temperate imagination serves as a synonym for low-value land. In some contexts it has also served as a metonym for soil degradation, particularly when their environmental histories have been written by European colonizers (Davis, 2007, 2016). This kind of ‘climate’ in the colonial imagination is still active and commonplace today. For example, in 1946 a doctor named Richmond Anderson decided to study the diet of indigenous Hñähñu people living in the Central Mexican Plateau for the United States Public Health Service in order to document what he believed would be a case of severe malnutrition (Anderson et al., 1946). This region of Mexico receives less than 500 mm of rain per year and has historically been considered infertile and of no agricultural value. Instead of malnutrition, Anderson found a population eating a mainly vegetarian diet of xerophytic plants with no nutritional deficit whatsoever. Many of these plants relied on crassulacean acid metabolism (CAM) photosynthesis, and produced a healthy diet from soil that could never sustain the kinds of foods recommended by global public health standards, such as wheat (*Triticum aestivum* L.), rice (*Oryza sativa* L.), or dairy products.

In this sense, the concept of soil fertility contains within it hidden landscapes, climates, and botanical values that are not neutral, and culturally specific. It would be a mistake, however, to think that these different interpretations of soil fertility are simply the result of different opinions about how to interpret the same soil data. For instance, one might imagine that soil fertility is simply a

subjective concept applied to objective soil data. The problem with this interpretation is that in each of the above cases, human beings were actively intervening in the soil in order to give it new qualities, which in an objective sense, were not already there. Understanding how to use the litter layer to farm under fumigated Amazonian canopy, or how to maintain soil moisture with CAM plants and nitrogen fixing legumes in arid environments like the Central Mexican Plateau resulted in soils with objectively different properties.

This socio-ecological dimension of soil as an empirical object is precisely what makes it so difficult to isolate as the kind of map layer we see in Fig. 1, and so far there have been two primary problems. To recapitulate, the first problem was that the socio-ecological relations between soil and human societies resulted in different kinds of soil knowledge, and therefore different soil taxonomies, which were mutually incompatible at a global scale. In other words, different ideas about how to use soil resulted in different systems for describing it, and therefore different ways of drawing the soil layer on a map. The second problem was that even once a consensus had been established for how to describe and classify soil at a global scale, socio-ecological processes intervened in the process of soil formation at a regional scale in ways that undermined the utility of the classification for planning purposes. The result is that the World Soil Map, like soil taxonomy itself, looks increasingly like an unfinishable project, as it emerges from ongoing practices that shift how the soil is interpreted as value.

For planners, the best place to observe this dual problem of soil mapping is in urban areas, where both of these problems contribute to an almost complete absence of data (See Fig. 2).² On maps of urban areas, soils are typically given generic labels like 'urban land' or 'fill,' which signify the absence of a taxonomic category more than the presence of a well-described soil. Whereas surrounding areas may have detailed information about the soil's horizon structure, pH, texture, and minerals, urban soil often appears homogenous on soil surveys despite its immense material complexity. Like the arid or Amazonian soils above, human intervention has transformed the objective properties of urban soil as an empirical object. Urban processes have effectively intervened in the soil's formation, changing its porosity, chemistry, biology, and capacity for gas exchange, making it indecipherable to the dichotomous keys of, for example, 'Soil Taxonomy,' which is the United States system of classification shown in Fig. 2. For natural historical reasons, the empirical practices that shaped the world's most comprehensive soil taxonomic systems in the 20th century were simply not developed to understand soils formed through processes of urbanization. This has resulted in poor descriptions of the urban soil environment, and an ever widening gap in the soil survey as urban areas expand around the globe.

Today, there is an active and vigorous scientific culture growing around the project of developing new empirical practices capable of interpreting urban soil. Many new initiatives exist to expand traditional soil taxonomic categories to include anthropogenic soils. Soils of Urban, Industrial, Traffic, Military, and Mining Areas (SUITMA) is an active working group of the International Union of Soil Sciences (IUSS), and its members have produced a surge of new research over the last decade. Since 2006, the World Reference Base (WRB) soil taxonomy incorporates high level categories for anthropogenic soils. This is critical research, not only for landscape planning, but for urban environmental health and justice broadly defined. The capacity of soil to filter urban toxicity, mitigate flooding, and support healthy environments is essential for planning urban environments.

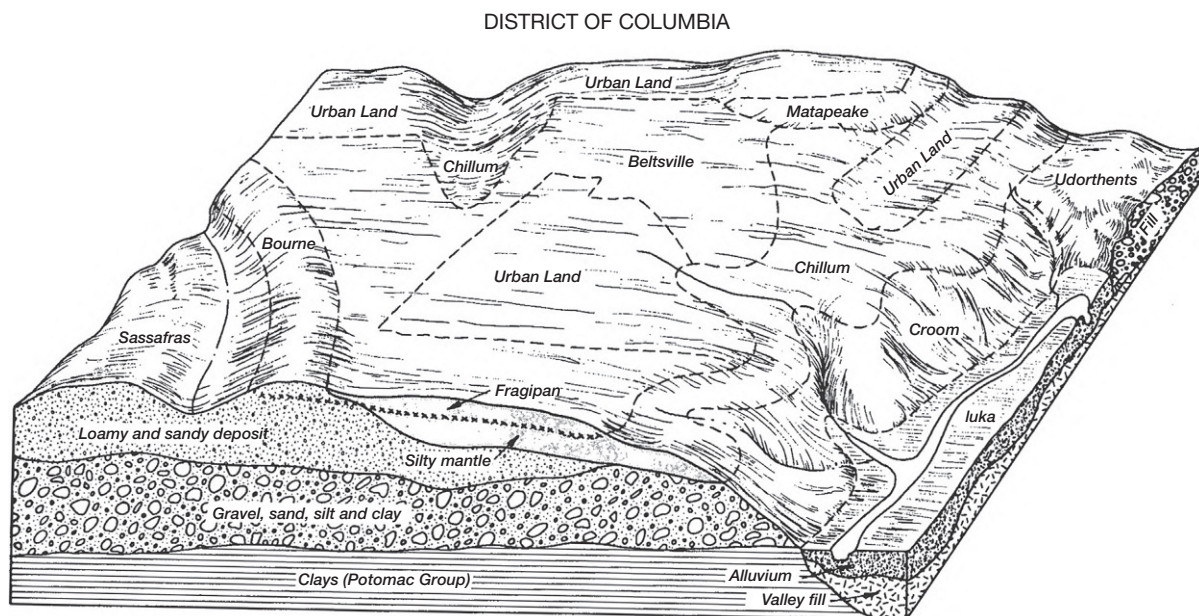


Fig. 2 'Urban land' and other classifications from the Soil Survey of the District of Columbia, Washington DC.

²Figure from Smith H (1976) *Soil Survey of District of Columbia*. Washington DC: US Department of Agriculture Soil Conservation Service (USDA).

The socio-ecological dimension of urban soil, however, continues to pose problems for a taxonomic project whose boundaries were circumscribed by the distinction between human and natural history. As such, there has been a tendency to treat anthropogenic soils as if they are simply the same natural materials as before, but simply re-organized into new kinds of shapes and mixtures. In this paradigm, soil science proceeds by analogy. Construction debris is interpreted as if it were a coarse layer of calcareous rock, and the long-term pedological effects of these new calcareous rock layers in the landscape are analyzed according to the chemistry of calcium. This approach has value, but something very specific is lost in the transformation. The process of recognizing only the chemical and physical properties of human history that are legible to existing natural historical frameworks of soil knowledge has a severe side effect, which is that it acts as an efficient filter, removing politics from the soil's history. To remove politics from a soil's history is to limit its utility for planning. For the planner it would be important, for example, to know that the construction debris came from urban renewal. Unlike the quartz or greywacke that a pedologist might identify in a soil profile, construction debris is produced through social relationships, and these relationships matter. If planners do not understand where urban soil comes from, both historically and materially, it becomes difficult to connect policy decisions to their pedological consequences and therefore to make different decisions about what kinds of soil properties to cultivate in the future. Rather than seeing the task of planning in this context as merely assessing the resource value of soil, it is important to recognize that in an urban context, planning produces soil by shaping the conditions of its future formation. The future toxicity, fertility, porosity, or stability of urban soils, among other properties, will depend on decisions that are made today.

Making decisions: The politics of cause and effect in socio-ecological systems

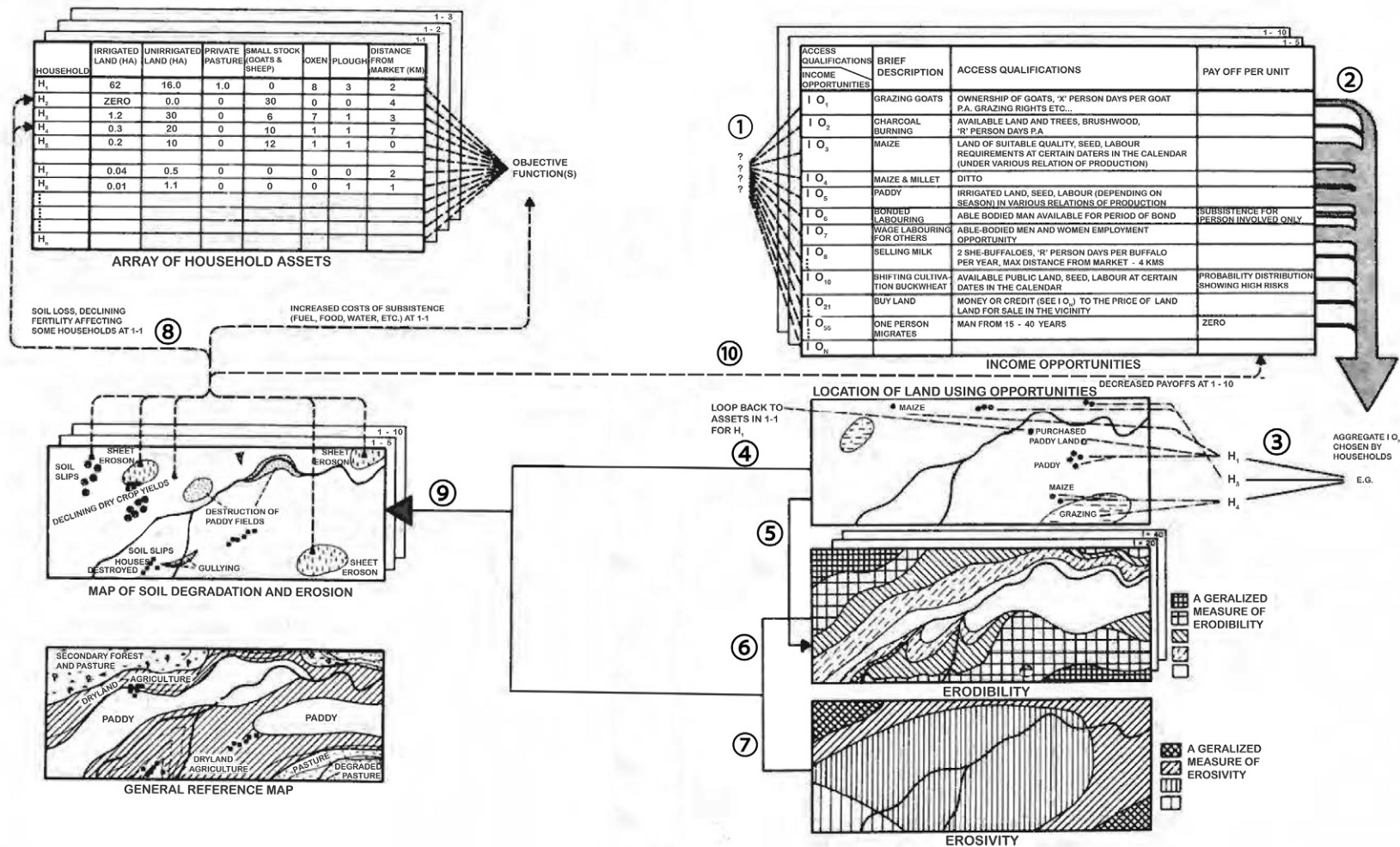
From this perspective, it should start to become clear why visualizing soil as an independent layer that one could overlay with other land-use layers on a planning map is problematic. In a naïve way, one might ask if soils determine the land-use layer, or if land use determines the soil layer? The division between these layers formalizes an assumption about the natural historical quality of soil that events in the 20th century have called into question or, at the very least, made inconvenient. Today, these layers are increasingly inextricable.

Fundamentally, thinking about the relationship between the soil and land use layers of a planning map is about connecting soil to its consequences. What are the consequences of soil and soil formation for 'land,' as a category of use and value? As we have seen, this question is not easy to answer empirically. The empirical process of analyzing soil already contains certain values that are not always independent of land use, but there is also another difficulty here. Even when soil processes are considered entirely 'natural,' isolating cause and effect relationships can be difficult, and this is a problem that soil scientists have struggled with for generations. The great soil scientist Hans Jenny tells a story about the difficulty of narrating cause and effect in the Mendocino pygmy forest, in California (Jenny et al., 1969). According to Jenny, botanists would bring their students to Mendocino and explain how the pygmy forest was the result of a very acidic soil, whereas soil scientists would bring their students to the same spot and explain, to the contrary, that it was the acidic leaves of the forest's ericaceous plants that produced the acidic soil. In this system plants and soils were co-producing one another, and for Jenny, this was a story about the need to think dialectically about the distinction between biology and geology in ecosystems in general. However, Jenny could just as easily have been speaking about the relationship between people and soil in a city or on a farm. There is a reciprocal relationship between soils and land use that is bi-directional and time-dependent. Causality in these systems is often more circular than linear. While this kind of holistic environmental thinking may produce less reductive descriptions of purely ecological systems, they tend to present a political problem for planners and politicians, however, whose policy decisions are measured by how clearly they appear to be the origin of change.

The first systematic attempt to think about this problem in relation to soil was Piers Blaikie's, 1985 book *The Political Economy of Soil Erosion in Developing Countries* (Blaikie, 1985). Here, Blaikie's central question was why policy initiatives to reduce soil erosion were so often unsuccessful, or even worsened the problems they intended to solve. His conclusion was that soil erosion tends to be perceived by development organizations as a purely technical problem to be remedied with terracing, agricultural equipment, crop rotation, population control, or the technical education of a misinformed peasantry. The inability of international development agencies and non-governmental agencies (NGOs) to solve soil erosion problems was a result of the narrowness of this technical approach, which according to Blaikie was narrow by design, given the need to demonstrate policy effectiveness through quantitative cause and effect benchmarks that could be reported to funders. Blaikie called this the "colonial approach to erosion and conservation," and argued that soil degradation was actually a distant and downstream byproduct of the interaction between soil and social relations, many of which were determined by global economic forces and local historic inequalities in land tenure, taxation, and political agency (Blaikie, 1985, p. 53). In Blaikie's research, there was no easy linear causal mechanism available to address soil erosion, only complex bi-directional models of socio-ecological change. Consider, for example, the difference between the typical layer-based model of soil shown in Fig. 1 and Blaikie's model in Fig. 3. Here, soil is not represented as a layer that can be overlapped with other social and ecological data. Instead, the soil map-layer appears as a dynamic form within the socio-ecological processes that shape it. In this model, the linear boundaries of the soil shown on the map reflect social relations as much as soil formation. In Fig. 3 one might say that the layer-cake model (Fig. 1) has collapsed under its own weight like a soufflé.

Blaikie (1985) focused only on soil erosion, but his findings are still relevant to the way we think about the socio-ecology of soil. For example, today's United Nations Sustainability Goals (SDG 12) call for reductions in carbon dioxide (CO₂) emissions in modern agricultural practices. While CO₂ emissions provide a quantifiable benchmark for progress on this goal, it is only achievable through significant social and political change. Chefs and seed savers may turn out to be as important to this transition as farmers,

Land use decisions and soil erosion



politicians and regulatory bodies, particularly as climate change affects the seasonality and water efficiency of food. Setting goals for what kinds of soils to cultivate requires an intervention in fundamental aspects of both gastronomy and economy, and from this perspective soil science looks increasingly like a planning discipline.

After the collapse of the layer cake model: Rethinking the relationship between soil and land-use

In Southern China, the discovery of soil-dependent strains of cadmium-excluding rice has made the interconnection between soil and planning particularly clear. In the southern provinces of Hunan, Guizhou, Jiangsu and Guangdong, mining and industrial corridors have left a legacy of cadmium-contaminated soils in close proximity to rice production. According to a national survey by China's Ministry for Land and Resources in 2014, almost 20% of China's arable land is now contaminated (Zhao et al., 2015). The cadmium that the rice plant absorbs from these soils is transferred to the grain, contaminating the food supply and causing "Itai-itai" disease, which translates as "it hurts-it hurts" disease, in reference to the severe spine and joint pain that cadmium poisoning causes. Rice is currently a staple crop for 3 billion people, which means that the vast legacy of post-industrial toxicity in Asia cannot easily be fixed by simply abandoning rice production in these areas. Even if this was an option, the precise boundaries of the contamination are not known. To survey agricultural soils for contamination in China required an astonishing 787,500 soil samples to cover an area of 6.3 million square kilometers. At this scale, each sample represented an area of 8 km². Moreover, the most dangerous soils in this context are the ones that are only slightly contaminated, where cadmium levels are not high enough to kill the plants but still high enough to accumulate in the rice grain. How many more samples would be needed to delineate clearly the subtle edge of this boundary for China as a whole?

Over the last several decades researchers have been trying to address this problem through a soil-dependent pathway that takes advantage of the biology of the rice plant. As a wetland species adapted to aquatic environments, the aerenchyma of the rice plant transports oxygen to the roots for respiration. Roughly 30–40% of this oxygen is lost to the environment through root pores and wherever it comes into contact with iron in the soil a metallic 'plaque' forms on the roots (Cheng et al., 2014). This rusty plaque is an effective filter against cadmium uptake, and the result is that specific strains of oxygen exuding rice in combination with iron rich soil have the potential to dramatically improve food safety in China (Chen et al., 2017; Rahimi et al., 2021).

If there was a 21st century post-industrial version of Hans Jenny's parable in the Mendocino coastal terraces, this might be it. A plant–soil interaction has re-drawn land-use boundaries in China through selective plant breeding, activism, land-use change, and public health research. No one element in the process could be called the origin of this new socio-ecological compromise. It is important to note however, that the cadmium is still in the soil and will stay there for the foreseeable future. Nothing has been cleaned up, and in this sense there is little to celebrate. However, in the 21st century, toxicity is not a binary condition of 'present' or 'absent' but rather a pervasive condition that has disproportionate effects on low income and people of color in both wealthy and developing nations. Instead of avoiding toxicity, which requires resources and is typically a privilege reserved for the environmental enclaves of the wealthy, it will be increasingly important to develop transparent and public strategies for surviving this condition. For the chemistry and biology of toxicity to be included in political forums there needs to be public trust in the way toxicity is represented, or made visible. For trust to be established, these discussions should not only occur in the context of a public health crisis. Transparent environmental monitoring and reporting of health risks needs to become a public utility, like electricity or water, to provide the time and space necessary to address environmental problems with long-term planning rather than short-term crisis measures. Soil is arguably the most critical buffer against environmental toxicity. Conserving the soil properties that support this function will increasingly require this kind of integrated approach to understand the geo-biology of land-use change.

Soil futures: A new discipline

Given the centrality of soil to global sustainability goals, what should the future of soil science and landscape planning look like? What would it mean to cultivate an expertise in this area adequate to the challenges we are now facing in an uncertain climate?

Today, the vast majority of published information on the relationship between soil and land use falls into one of two categories: resource extraction or damage control. In the resource extraction paradigm, the properties of soil are assessed for how well they might support or prohibit a particular land use. In the damage control paradigm, a particular land use has caused some degree of soil degradation and the question is what to do about it, or how things might have been done differently. In these paradigms, soil is either a pristine natural resource to be utilized or the material aftermath of some lamentable event. What these well-known genre's of applied soil research have in common is the basic conceptual division enacted by the layer-based model of soil in the human environment: in the first paradigm soil is a cause and land use is the effect, whereas in the second the order is reversed: land-use causes soil degradation. In order to think about soil and landscape planning in the 21st century, a new genre of soil narrative will need to be invented. Rather than assessing resource stocks or documenting degradation, one question we could ask is the following: how could a future soil be planned through the soil forming effects of land-use relationships? Could these relationships be environmentally just, decolonial, and climate aware? What kinds of circular economies might cultivate soil resources rather than degrade them? Designing land-use strategies around their soil forming effects would constitute the methodology for a discipline that does not yet exist. However, the political and environmental conditions which produce the need for such a discipline are already clear. If soil and landscape planning remain as distinct as the title of this entry suggests, then perhaps we need a new discipline to address these challenges. What would a degree in 'soil futures' look like?

References

- Anderson RK, Calvo J, Serrano G, and Payne GC (1946) A study of the nutritional status and food habits of Otomi Indians in the Mezquital Valley of Mexico. *American Journal of Public Health and the Nation's Health* 36: 883–903. <https://doi.org/10.2105/AJPH.36.8.883>.
- Blaikie PM (1985) *The Political Economy of Soil Erosion in Developing Countries*, Longman Development Studies. London, New York: Longman.
- Certini G and Scalenghe R (2011) Anthropogenic soils are the golden spikes for the Anthropocene. *The Holocene* 21: 1269–1274. <https://doi.org/10.1177/0959683611408454>.
- Chen Z, Tang Y-T, Yao A-J, Cao J, Wu Z-H, Peng Z-R, Wang S-Z, Xiao S, Baker AJM, and Qiu R-L (2017) Mitigation of Cd accumulation in paddy rice (*Oryza sativa* L.) By Fe fertilization. *Environmental Pollution* 231: 549–559. <https://doi.org/10.1016/j.envpol.2017.08.055>.
- Cheng H, Wang M, Wong MH, and Ye Z (2014) Does radial oxygen loss and iron plaque formation on roots alter Cd and Pb uptake and distribution in rice plant tissues? *Plant and Soil* 375: 137–148. <https://doi.org/10.1007/s11104-013-1945-0>.
- Davis DK (2007) *Resurrecting the Granary of Rome: Environmental history and French Colonial Expansion in North Africa*. Ohio University Press series in ecology and history Athens: Ohio University Press.
- Davis DK (2016) *The Arid Lands: History, Power, Knowledge*. Cambridge, Massachusetts: The MIT Press.
- Dudal R and Batisse M (1978) The soil map of the world. *Nature and Resources* 14: 2–6.
- Food and Agriculture Organization (FAO), Intergovernmental Technical Panel on Soils (ITPS) (2015) *Status of the World's Soil Resources (SWSR) – Main Report*. Rome, Italy: Food and Agriculture Organization of the United Nations.
- Gordon SI and Gordon GE (1981) The accuracy of soil survey information for urban land use planning. *Journal of the American Planning Association* 47: 301–312. <https://doi.org/10.1080/01944368108976512>.
- Jenny H, Arkley RJ, and Schultz AM (1969) The pygmy forest-podsol ecosystem and its dune associates of the Mendocino coast. *Madrone* 20: 60–74.
- Li TM (2014) What is land? Assembling a resource for global investment. *Transactions of the Institute of British Geographers* 39: 589–602. <https://doi.org/10.1111/tran.12065>.
- Lyons KM (2020) *Vital Decomposition: Soil Practitioners + Life Politics*. Durham and London: Duke University Press.
- Marsh WM (2010) *Landscape Planning: Environmental Applications*, 5th edn Hoboken, NJ: Wiley.
- Montanarella L, Scholes R, Brainich A, and Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (2018) *The IPBES Assessment Report on Land Degradation and Restoration*. Bonn, Germany: Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services.
- Nachtergaele FOF (2003) The future of the FAO legend and the FAO/UNESCO soil map of the world. In: Eswaran H, Rice T, Ahrens R, and Stewart BA (eds.) *Soil Classification: A Global Desk Reference*, p. 263. Boca Raton, FL: CRC Press.
- Rahimi M, Rahimi G, Ebrahimi E, and Moradi S (2021) Assessing the distribution of cadmium under different land-use types and its effect on human health in different gender and age groups. *Environmental Science and Pollution Research* 28: 49258–49267. <https://doi.org/10.1007/s11356-021-12881-2>.
- Richter DD and Babbar LI (1991) Soil diversity in the tropics. *Advances in Ecological Research* 21: 315–389. [https://doi.org/10.1016/S0065-2504\(08\)60100-2](https://doi.org/10.1016/S0065-2504(08)60100-2).
- Selcer P (2018) *The Postwar Origins of the Global Environment: How the United Nations Built Spaceship Earth*, *Columbia Studies in International and Global History*. New York: Columbia University Press.
- Zhao F-J, Ma Y, Zhu Y-G, Tang Z, and McGrath SP (2015) Soil Contamination in China: Current Status and Mitigation Strategies. *Environmental Science & Technology* 49: 750–759. <https://doi.org/10.1021/es5047099>.

Encyclopedia of SOILS IN THE ENVIRONMENT

Volume Three: Environment and Management

EDITORS-IN-CHIEF

Michael J Goss

Professor Emeritus, School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G 2W1, CANADA

Margaret Oliver

Visiting professor, Soil Research Centre, Department of Geography and Environmental Science, Russell Building, University of Reading, Whiteknights, Reading, RG6 6DR, United Kingdom.

SECTION EDITORS

Dr Evgenia Blagodatskaya

Dept. of Soil Ecology, Helmholtz-Zentrum für Umweltforschung GmbH – UFZ, Theodor-Lieser Str. 4, D-06120 Halle (Saale), GERMANY

Professor Adrian Unc

Boreal Ecosystems and Agricultural Sciences, Grenfell Campus, Memorial University of Newfoundland, 20 University Drive, Corner Brook, Newfoundland A2H 5G4, CANADA

Dr. Siobhán Staunton

INRA, UMR Eco&Sols, Montpellier SupAgro, bâtiment 12, 2 Place Viala, 34060 Montpellier Cedex 2, Montpellier, FRANCE

Professor Baoshan Xing

University of Massachusetts – Amherst, College of Natural Sciences, Stockbridge School of Agriculture, Environmental & Soil Chemistry, Program, 410 Paige Laboratory, Amherst, MA 01003, USA

Dr Karin Mueller

New Zealand Institute for Plant & Food Research Ltd., Ruakura Research Centre, Private Bag 3123, Hamilton, New Zealand

Professor Christina Siebe

Instituto de Geología, Circuito de la Investigación Científica s/n, Ciudad Universitaria, Ciudad de México, CP 04510, México

Professor Rupert Baeumler

Institute of Geography, Friedrich-Alexander-University of Erlangen-Nuremberg, Wetterkreuz 15, 91058 Erlangen, Germany

Dr Uta Stockmann

Soil Processes and Function Organisational Unit, CSIRO Agriculture and Food, Black Mountain, GPO BOX 1700, Canberra ACT 2601, AUSTRALIA

Professor Paul Hallett

Room 2.23, Cruickshank Building, School of Biological Sciences, University of Aberdeen, Aberdeen, AB24 3UU, UNITED KINGDOM

Professor Dani Or

Department of Environmental Systems Science (D-USYS), Institute of Biogeochemistry and Pollutant Dynamics (IBP), Soil and Terrestrial Environmental Physics (STEP), CHN F 29.1 Universitaetstrasse 16, 8092 Zurich, SWITZERLAND



ELSEVIER

elsevier.com/books-and-journals

Set ISBN

ISBN 978-0-12-822974-3



9 780128 229743

Volume ISBN

ISBN 978-0-443-15210-8

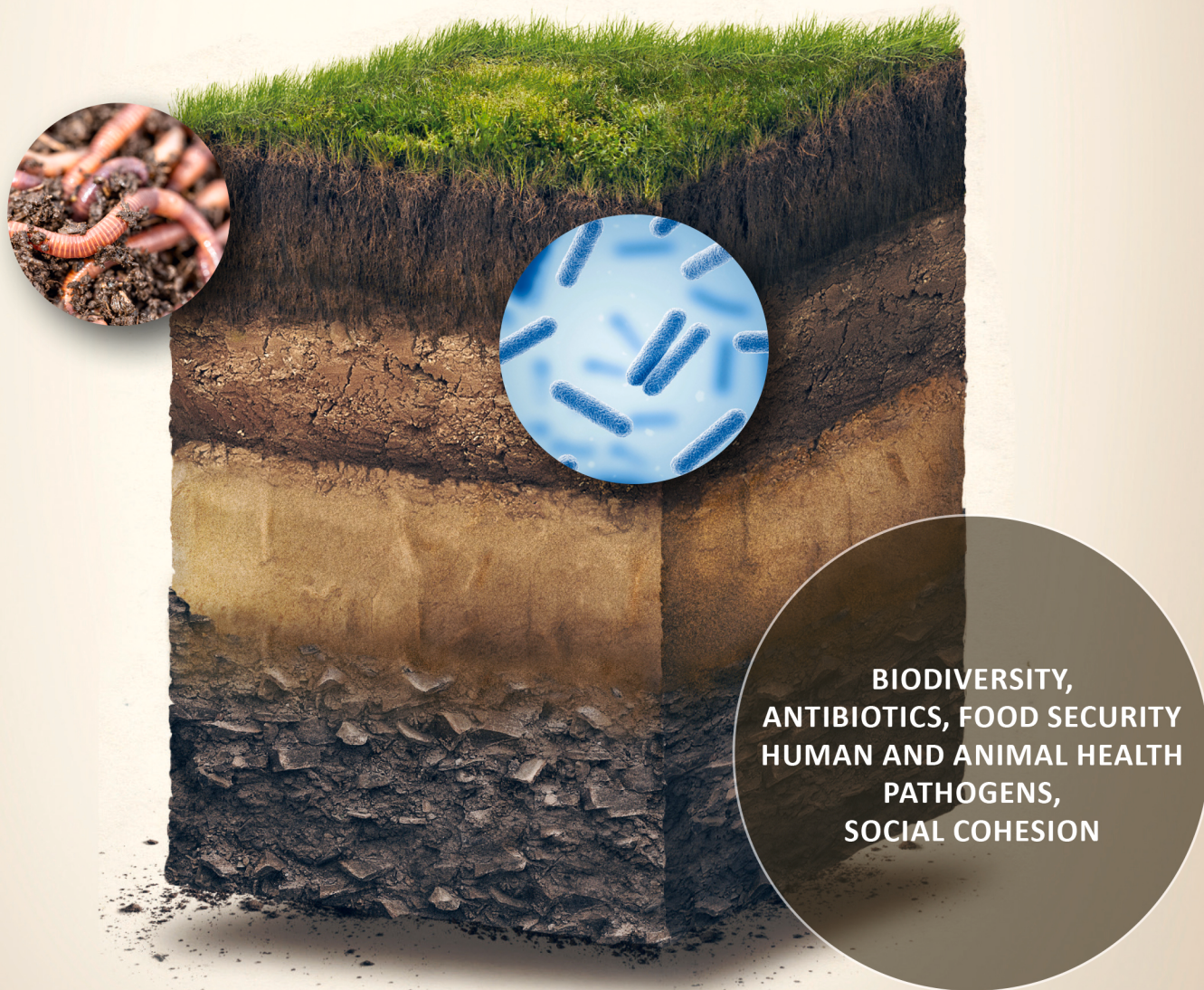


9 780443 152108

Second Edition

Encyclopedia of SOILS IN THE ENVIRONMENT

Volume Four: Pedology & Pedometric



Editors-in-Chief

Michael Goss and Margaret Oliver

Section Editors

Rupert Baeumler & Uta Stockmann

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

SECOND EDITION

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

SECOND EDITION

EDITORS-IN-CHIEF

Professor Michael J Goss

*Professor Emeritus, School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G
2W1, CANADA*

Professor Margaret Oliver

*Visiting Professor, Soil Research Centre, Department of Geography and Environmental Science,
Russell Building, University of Reading, Whiteknights, Reading, RG6 6DR, United Kingdom*

VOLUME 4

PEDOLOGY

PEDOMETRICS

SECTION EDITORS

Rupert Baeumler

*Institute of Geography, Friedrich-Alexander-University of Erlangen-Nuremberg,
Wetterkreuz 15, 91058 Erlangen, Germany*

Uta Stockmann

*Soil Processes and Function Organisational Unit, CSIRO Agriculture and Food,
Black Mountain, GPO BOX 1700, Canberra ACT 2601, AUSTRALIA*



ELSEVIER

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge MA 02139, United States

Copyright © 2023 Elsevier Ltd. All rights reserved

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

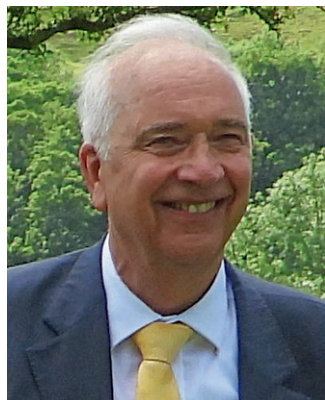
ISBN 978-0-12-822974-3

For information on all publications
visit our website at <http://store.elsevier.com>



Publisher: Oliver Walter
Acquisition Editor: Priscilla Braglia
Content Project Manager: Sona Chahal
Associate Content Project Manager: Nandhini Mahendran
Designer: Christian Bilbow

EDITORS BIOGRAPHY



Michael J Goss, PhD, MSc, BSc, is a professor emeritus in the School of Environmental Sciences at the University of Guelph. Before taking on the joint editor-in-chief responsibilities with Margaret Oliver, he held a similar position for *Soil Use and Management*, a journal of the British Society of Soil Science of which he is an honorary fellow. He was an inaugural member of the European Society for Agronomy. Michael started his career at the Letcombe Laboratory before moving to Rothamsted Experimental Station and the Macaulay Land Use Research Institute (now the James Hutton Research Institute). In 1990 he was appointed to the Research Chair in Land Stewardship at the University of Guelph and in 2005 became the director of its Regional Campus near Ottawa.

Michael's research has covered soil plant relations, soil management, and agriculture's footprint on the environment. His focus has been on the physical properties of soil and their impacts on root growth, which he showed to be influenced by the role of the root cap. He broadened his approach to include the effects of tillage and land drainage on root exploration. This work stimulated his interest in the transport of materials, such as plant nutrients, pathogenic microbes, antibiotics, and biologically active compounds through the soil and their contamination of water resources. Michael's knowledge in this field resulted in his involvement in the public inquiry into the death of seven people, including children, from a microbially contaminated drinking water supply. He contributed, both to the investigation into how the contamination occurred and if any fault could be attributed to local agricultural practices, and to how to prevent future events.

Michael also developed an interest in the interrelationships between soil microbes and root systems, particularly the tripartite symbiosis between roots of legumes, mycorrhizal fungi, and rhizobia. He has been an energetic member of collaborative research programs in France and Germany, and in Portugal, where he continues his interest in the role of mycorrhiza in sustainable land management.

Michael has authored more than 180 publications including two books, one in its 2nd edition, and edited two others.



Margaret Ann Oliver (BSc, PhD)

Margaret's interest in soil began in her penultimate year at school when learning about different types of soil and their formation. She chose the University of Bristol, United Kingdom for her first degree because the geography course offered soil science and she was also able to do a degree in geology alongside. Her love for both subjects remains undiminished. She did her PhD in soil spatial analysis at the University of Birmingham. Her early research focused on the multivariate and geostatistical analysis of soil data, which she continued to use in various research projects. Her research interests expanded to include the application of a wide range of numerical and geostatistical methods to soil and other data including pollen counts, forestry, radon emission, remotely sensed imagery, precision agriculture, and the incidence of childhood cancer. Her specific interests have been in sample design for spatial analysis and eventual mapping, risk analysis associated with prescribed thresholds in relation to soil pollutants or deficiencies of crop nutrients, and the relations between different scales of spatial variation. Her most

recent research was in the field of precision agriculture. At both the Universities of Birmingham and Reading, she taught topics in soil science, statistical methods, and spatial analysis. At Reading she had a research group working on various projects such as imagery, sampling and mapping, precision agriculture, and human health.

She has published more than 100 academic papers and has made several contributions to books. She is the co-author of three books: *Statistical Methods in Soil and Land Resource Survey*, Oxford University Press (Webster, R and Oliver, M. A. 1990); *Geostatistics for Environmental Scientists*, Second Edition, John Wiley & Sons (Webster, R and Oliver, M. A. 2007), and *How to do the Basic Steps in Geostatistics: The Variogram and Kriging*, Springer, Dordrecht (Oliver, M.A. & Webster, R.2015). She has edited two books: *Geostatistical Applications for Precision Agriculture* (Springer, Dordrecht, 2010) and *Precision Agriculture for Sustainability and Environmental Protection* (Earthscan from Routledge, 2013).

Margaret retired officially from Reading in 2004, but as yet has not managed to escape from academia. She has retained a link and a base at the University as a visiting professor since January 2005. After retirement, she developed and taught geostatistics courses to outside groups for several years. She then started on the long path of editorial work, as an associate editor of the *European Journal of Soil Science* and eventually the editor-in-chief, assistant editor of *Mathematical Geosciences*, co-editor of *Precision Agriculture* with John Stafford, and finally the editor-in-chief of the *Encyclopedia of Soils in the Environment* with Michael Goss.

Margaret was made an honorary member of the British Society of Soil Science in 2021 to reflect her commitment to soil science through fellow research, teaching, and editorial work. She was honored by the Pedometrics Commission of the International Union of Soil Science by having an award named after her in 2015. It is the Margaret Oliver award for young pedometricians.

EDITORIAL BOARD

Editors in Chief

Michael J Goss

Professor Emeritus, School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G 2W1, CANADA

Margaret Oliver

Visiting Professor, Soil Research Centre, Department of Geography and Environmental Science, Russell Building, University of Reading, Whiteknights, Reading, RG6 6DR, United Kingdom

Section Editors

Dr Evgenia Blagodatskaya

Dept. of Soil Ecology, Helmholtz-Zentrum für Umweltforschung GmbH – UFZ, Theodor-Lieser Str. 4, D-06120 Halle (Saale), GERMANY

Professor Adrian Unc

Boreal Ecosystems and Agricultural Sciences, Grenfell Campus, Memorial University of Newfoundland, 20 University Drive, Corner Brook, Newfoundland A2H 5G4, CANADA

Dr. Siobhán Staunton

INRA, UMR Eco&Sols, Montpellier SupAgro, bâtiment 12, 2 Place Viala, 34060 Montpellier Cedex 2, Montpellier, FRANCE

Professor Baoshan Xing

University of Massachusetts – Amherst, College of Natural Sciences, Stockbridge School of Agriculture, Environmental & Soil Chemistry, Program, 410 Paige Laboratory, Amherst, MA 01003, USA

Dr Karin Mueller

New Zealand Institute for Plant & Food Research Ltd., Ruakura Research Centre, Private Bag 3123, Hamilton, New Zealand

Professor Christina Siebe

Instituto de Geología, Circuito de la Investigación Científica s/n, Ciudad Universitaria, Ciudad de México, CP 04510, México

Professor Rupert Baeumler

Institute of Geography, Friedrich-Alexander-University of Erlangen-Nuremberg, Wetterkreuz 15, 91058 Erlangen, Germany

Dr Uta Stockmann

Soil Processes and Function Organisational Unit, CSIRO Agriculture and Food, Black Mountain, GPO BOX 1700, Canberra ACT 2601, AUSTRALIA

Professor Paul Hallett

*Room 2.23, Cruickshank Building, School of Biological Sciences, University of Aberdeen, Aberdeen, AB24 3UU,
UNITED KINGDOM*

Professor Dani Or

*Department of Environmental Systems Science (D-USYS), Institute of Biogeochemistry and Pollutant Dynamics (IBP),
Soil and Terrestrial Environmental Physics (STEP), CHN F 29.1 Universitaetstrasse 16, 8092 Zurich, SWITZERLAND*

FOREWORD

Soil is the porous “skin” on the surface of the Earth that has formed over thousands of years of weathering, comprising minerals, organisms (both dead and alive), water, and air. Soil is more than just the sum of its parts. It serves several crucial ecological functions, from food and fiber production to water storage and purification to waste decomposition.

Increasingly we understand that everything is connected, and that soil is a “Critical Zone” that intersects with the hydrosphere (water), lithosphere (minerals), atmosphere (air), and biosphere (all living organisms). To quote Daniel Hillel, the 2012 World Food Prize laureate and editor of the First Edition of *The Encyclopedia of Soils in the Environment*, “[Soil] is the crucible of terrestrial life, within which biological productivity is generated and sustained...Our civilization depends on the soil...” He also promoted the importance of understanding soil to manage it effectively. This revised and reorganized second edition of the Encyclopedia aims to reflect the changing status of soil in the world. Within these five volumes, over 724 experts from many countries have provided updated, easy-to-understand chapters on soil, its properties, and its role in the environment to affect climate change, plant productivity, human health, and biodiversity.

Many technological advances have been made since the first edition of this encyclopedia that has allowed us to investigate soil physical, chemical, and especially biological properties more precisely, enabling us to identify and understand their roles in ecosystem services. The study of soils, especially in the environmental context, is critical to meeting the challenges of feeding, clothing, and cleaning up the world. The information presented in these volumes will be useful to many, including soil and environmental scientists, policymakers, land managers, educators, students, and anyone interested in learning more about soil.

Soil is under considerable threat from many sources—building and infrastructure, climate change, population pressure, and pollution, yet it is vital to agronomic productivity and continues to provide critical ecosystem services. Society needs to be engaged in how soil is managed and conserved. Most natural scientists are very dedicated and generous, including all of the authors in this Encyclopedia who have freely given their time to make these volumes so comprehensive. The study of soil, while often challenging, is fascinating—soil is a beautiful earthy material that can delight our senses with its variegated colors, gritty to smooth textures, and especially its characteristic aroma when freshly turned. Perhaps that is why gardeners and even some farmers tend to overwork their soil to its detriment—it *just feels good to get your hands in the dirt!*

Soil science is one of the most exciting, interesting, and rewarding disciplines, largely because it is multi-dimensional and interdisciplinary, basic and applied. It affects all of our lives. Thus, I encourage you to peruse these volumes and discover all kinds of interesting aspects of soil in the environment.

April Ulery

Former President, Soil Science Society of America
Professor, Soil & Environmental Sciences
New Mexico State University
Las Cruces, NM
The United States

PREFACE

This second edition of the *Encyclopedia of Soils in the Environment* follows in the footsteps of a highly rated first edition overseen by Professor Daniel Hillel as editor in chief, ably assisted by six editors. That edition provided a very valuable resource on the “state-of-the-art” in soil science in the context of the environment in the closing years of the twentieth century and the first three years of the current century. It set a standard that the editorial team for the second edition has aimed to match. The soil and environmental sciences have made huge strides in the almost 20 years since the first edition was published, and the new team has had to embrace these developments while retaining the basic components of these sciences. The project has taken 4 years to complete. It began with the appointment of the two editors-in-chief, professors Michael J. Goss and Margaret A. Oliver, followed by that of the nine subject editors and a pedology adviser.

The project was in its infancy (a few months only) when COVID-19 struck. The editors-in-chief discussed the format and structure of the second edition around the time that the first “lockdowns” around the world were being introduced. The first edition was alphabetically organized, whereas we decided to have a section-oriented approach in this edition, which we hope gives a more structured feel for the various topics embraced. The principal sections are soil biology, soil chemistry, the soil environment and its management, pedology (the origins and development of soil), pedometrics (the application of mathematics and statistics to soil), and soil physics. We were able to give our full attention to this development and to the selection of the scientists to join us on the road ahead. Furthermore, we were fortunate to choose a team that has maintained its enthusiasm despite the many tribulations we have faced along this path. The team comprises Jane Blagodatskaya and Adrian Unc (biology and microbiology), Siobhan Staunton and Baoshan Xing (chemistry), Karin Müller and Christina Siebe (environment and management), Rupert Bäumler (pedology), Uta Stockmann and Margaret Oliver (pedometrics), and Paul Hallett and Dani Or (physics). The editorial board has been mindful in their choice of subjects of the need for basic soil science and for state-of-the-art developments, with the aim that the audience for these volumes can be as large as possible.

During the gestation period of the Encyclopedia, methods of working for those in academia were turned on their heads. The change from “face-to-face” teaching to formulating and establishing online courses, initially for lockdown periods but then for much of the next 2 years, had huge consequences for academic staff worldwide. This meant that they had less time to devote to outside activities, such as writing chapters for an encyclopedia. As a result, some authors were unable to complete their chapters in time, leaving a few unfortunate gaps. The whole team has done as much as they can to limit any adverse effects, but we ask readers to appreciate the struggle that authors have had with illness (personal, within their families and work colleagues) and a great increase in their workload. Despite this, we are presenting a comprehensive document that should stand the test of the next 20 years of soils in the environment.

This century has seen an almost unprecedented interest in the soil and environmental sciences. It has been triggered not only by issues that the Earth is facing from global climate change, increasing population pressure, soil and land degradation, and food insecurity but also by the many new and sophisticated methods available, such as the DNA determination of microbial populations and their functional capabilities, proximal sensing, machine learning methods, and an ever-increasing array of technological tools. These developments have greatly increased our knowledge and appreciation of the complexity of soils, which has changed our understanding of some well-established aspects of the science. The soil and its environment are at the critical interface of the atmosphere, biosphere, hydrosphere, and lithosphere. They deserve to be taken seriously by the international community in the same way that air and water have been. Governments are now paying attention

to the soil and environment, and a legal and governance framework is emerging to look after these parts of our life support resources.

We have aimed, through our editing of chapters, to make the language of the chapters accessible to all, regardless of educational background. This does not mean that we have diminished the scientific content, but that readers should be able to gain the foothold they desire in these subjects to progress to greater knowledge and understanding of them. We have retained the overview of soil written by Daniel Hillel for the First Edition, first to honor his great contribution to soil science and second to demonstrate the significance of soil. It follows this Preface.

*Michael J. Goss
Margaret A. Oliver*

SOIL

The term “soil” refers to the weathered and fragmented outer layer of our planet’s land surfaces. Formed initially through the physical disintegration and chemical alteration of rocks and minerals by physical and biogeochemical processes, soil is influenced by the activity and accumulated residues of a myriad of diverse forms of life. As it occurs in different geologic and climatic domains, soil is an exceedingly varied body with a wide range of attributes.

Considering the height of the atmosphere, the thickness of the earth’s rock mantle, and the depth of the ocean, one observes that soil is an amazingly thin body—typically not much more than one meter thick and often less than that. Yet it is the fount of terrestrial life, within which biological productivity is generated and sustained. It acts like a composite living entity, a home to a community of innumerable microscopic and macroscopic plants and animals. A mere fistful of soil typically contains billions of microorganisms, which perform vital interactive biochemical functions. Another intrinsic attribute of the soil is its sponge-like porosity and its enormous internal surface area. That same fistful of soil may actually consist of several hectares of active surface, upon which physicochemical processes take place continuously.

Realizing humanity’s utter dependence on the soil, ancient peoples, who lived in greater closeness to nature than many of us today, actually revered the soil. It was not only their source of livelihood but also the material from which they built their homes and that they learned to shape, heat, and fuse into household vessels and writing tablets (ceramics, made of clayey soil, being the first synthetic material in the history of technology). In the Bible, the name assigned to the first human was Adam, derived from “adama,” meaning soil. The name given to the first earthling’s mate was Hava (Eve, in transliteration), meaning “living” or “life-giving.” Together, therefore, Adam and Eve signified quite literally “Soil and Life.”

The same powerful metaphor is echoed in the Latin name for human species—Homo, derived from humus, the material of the soil. Hence, the adjective “human” also implies “of the soil.” Other ancient cultures evoked equally powerful associations. To the Greeks, the earth was a manifestation of Gaea, the maternal goddess who, impregnated by Uranus (god of the sky), gave birth to all the gods of the Greek pantheon.

Our civilization depends on the soil more crucially than ever because our numbers have grown while available soil resources have diminished and deteriorated. Paradoxically, however, even as our dependence on the soil has increased, most of us have become physically and emotionally detached from it. Many of the people in the so-called developed countries spend their lives in the artificial environment of a city, insulated from direct exposure to nature, and some children may now assume as a matter of course that food originates in supermarkets.

Detachment has bred ignorance, and out of ignorance has come the delusion that our civilization has risen above nature and has set itself free of its constraints. Agriculture and food security, erosion and salination, degradation of natural ecosystems, depletion and pollution of surface waters and aquifers, and decimation of biodiversity—all these processes, which involve the soil directly or indirectly, have become abstractions to many people. The very language we use betrays disdain for that common material underfoot, often referred to as “dirt.” Some fastidious parents prohibit their children from playing in the mud and hurry to wash their “soiled” hands when the children nonetheless obey an innate instinct to do so. Thus, soil is devalued and treated as unclean though it is the terrestrial realm’s principal medium of purification, wherein wastes are decomposed and nature’s productivity is continually rejuvenated.

Scientists who observe soil closely see it in effect as a seething foundry in which matter and energy are in constant flux. Radiant energy from the sun streams onto the field and cascades through the soil and the plants growing in it. Heat is exchanged, water percolates through the soil’s intricate passages, and plant roots extract

water and transmit it to their leaves, which transpire it back to the atmosphere. Leaves absorb carbon dioxide from the air and synthesize it with soil-derived water to form the primary compounds of life. Oxygen emitted by the leaves makes the air breathable for animals, which consume and in turn fertilize plants.

Soil is thus a self-regulating bio-physio-chemical factory, processing its own materials, water, and solar energy. It also determines the fate of rainfall and snowfall reaching the ground surface—whether the water thus received will flow over the land as runoff, or seep downward to the subterranean reservoir called groundwater, which in turn maintains the steady flow of springs and streams. With its finite capacity to absorb and store moisture, and to release it gradually, the soil regulates all these phenomena. Without the soil as a buffer, rain falling over the continents would run off entirely, producing violent floods rather than sustained river flow.

Soil naturally acts as a living filter, in which pathogens and toxins that might otherwise accumulate to foul the terrestrial environment are rendered harmless. Since time immemorial, humans and other animals have been dying of all manner of diseases and have then been buried in the soil, yet no major human disease is transmitted by it. The term antibiotic was coined by soil microbiologists who, as a consequence of their studies of soil bacteria, specifically actinomycetes, discovered streptomycin (an important cure for tuberculosis and other infections). Ion exchange, a useful process of water purification, also was discovered by soil scientists studying the passage of solutes through beds of clay.

However unique in form and function, soil is not an isolated body. It is, rather, a central link in the larger chain of interconnected domains and processes comprising the terrestrial environment. The soil interacts both with the overlying atmosphere and the underlying strata, as well as with surface and underground bodies of water. Especially important is the interrelation between the soil and climate. In addition to its function of regulating the cycle of water, it also regulates energy exchange and surface temperature.

When land is cleared of vegetation and turned into a cultivated field, the native biomass above the ground is often burned and the organic matter within the soil tends to decompose. These processes release carbon dioxide into the atmosphere, thus contributing to the Earth's greenhouse effect and to global warming. On the other hand, the opposite act of afforestation and soil enrichment with organic matter, such as can be achieved through conservation management, may serve to absorb carbon dioxide from the atmosphere. To an extent, the soil's capacity to store carbon thus helps to mitigate the greenhouse effect.

Thousands of years are required for nature to create life-giving soil out of sterile bedrock. In only a few decades, however, unknowing or uncaring humans can destroy that wondrous work of nature. In various circumstances, mismanaged soils may be subject to erosion (the sediments of which tend to clog streambeds, estuaries, lakes, and coastal waters), to leaching of nutrients with attendant loss of fertility and eutrophication of water bodies, to waterlogging and impaired aeration, or to an excessive accumulation of salts that may cause a once-productive soil to become entirely sterile. Such processes of soil degradation, sometimes called "desertification," already affect large areas of land.

We cannot manage effectively and sustainably that which we do not know and thoroughly understand. That is why the tasks of developing and disseminating sound knowledge of the soil and its complex processes have assumed growing urgency and importance. The global environmental crisis has created a compelling need for a concentrated, concise, and definitive source of information—accessible to students, scientists, practitioners, and the general public—about the soil in all its manifestations in nature and in relation to the life of humans.

Daniel Hillel
May 2004

LIST OF CONTRIBUTORS FOR VOLUME 4

Wulf Amelung

Institute of Crop Science and Resource Conservation (INRES), Soil Science and Soil Ecology, University of Bonn, Bonn, Germany

Dominique Arrouays

INRAE, InfoSol, Orléans, France

Ebrahim Babaeian

Department of Soil, Water, and Ecosystem Sciences, University of Florida, Gainesville, FL, United States

David Badía-Villas

Technological College of Huesca, University of Zaragoza, Zaragoza, Spain

Rupert Bäumler

Institute of Geography, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

A Belmonte

Institute for Electromagnetic Sensing of the Environment, National Research Council (CNR-IREA), Bari, Italy

JF Berkowitz

United States Army Corps of Engineers, Engineer Research and Development Center, Vicksburg, MS, United States

Thomas FA Bishop

Sydney Institute of Agriculture, The University of Sydney, Camperdown, NSW, Australia; School of Life and Environmental Sciences, The University of Sydney, Camperdown, NSW, Australia; Sydney Informatics Hub, The University of Sydney, Camperdown, NSW, Australia

Antonio Bispo

INRAE, Info&Sols, Orléans, France

Janis L Boettinger

Department of Plants, Soils, and Climate, Utah State University, Logan, UT, United States

J Böhner

Research Section Physical Geography, Institute of Geography, Center for Earth System Research and Sustainability (CEN), Universität Hamburg, Hamburg, Germany

Gabriele Broll

Institute of Geography, University of Osnabrueck, Osnabrueck, Germany

DJ Brus

Biometris, Wageningen University and Research, Wageningen, Netherlands

C Lee Burras

Iowa State University, Ames, IA, United States

A Castrignanò

Department of Engineering and Geology (InGeo), Università degli Studi Gabriele D'Annunzio, Chieti-Pescara, Chieti, Italy

Sabine Chabrillat

Helmholtz Center Potsdam GFZ German Research Center for Geosciences, Potsdam, Germany; Leibniz University Hannover, Institute of Soil Science, Hannover, Germany

Sarah E Chadburn

College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, United Kingdom

Déborah Chavrit

INRAE, Info&Sols, Orléans, France

Mathew S Crowther

Sydney Institute of Agriculture, The University of Sydney, Camperdown, NSW, Australia; School of Life and Environmental Sciences, The University of Sydney, Camperdown, NSW, Australia

Nilton Curi

Department of Soil Science, Federal University of Lavras (UFLA), Lavras, Minas Gerais, Brazil

Dennis Dahms

Department of Geography, University of Northern Iowa, Cedar Falls, IA, United States

Martin G De Kauwe

School of Biological Sciences, University of Bristol, Bristol, United Kingdom

Seppe Deckers

*Department of Earth and Environmental Science,
University of Leuven, Leuven, Belgium*

Stefaan Dondeyne

*Department of Geography, Ghent University, Ghent,
Belgium*

Markus Egli

*Department of Geography, University of Zürich, Zürich,
Switzerland*

Richard J Ellis

*UK Centre for Ecology & Hydrology (UKCEH),
Wallingford, United Kingdom*

Patrick Filippi

*Sydney Institute of Agriculture, The University of Sydney,
Camperdown, NSW, Australia; School of Life and
Environmental Sciences, The University of Sydney,
Camperdown, NSW, Australia*

Peter A Finke

*Department of Environment, Faculty of Bioscience
Engineering, Ghent, Belgium*

Maria Gerasimova

*V.V. Dokuchaev Soil Science Institute, Russian Academy of
Sciences, Moscow, Russia*

Alexander Gruber

*Department of Geodesy and Geoinformation, Vienna
University of Technology (TU Wien), Vienna, Austria;
Department of Earth and Environmental Sciences, KU
Leuven, Heverlee, Belgium*

S Grunwald

*Pedometrics, Landscape Analysis, and GIS Laboratory,
Department of Soil, Water, and Ecosystem
Sciences, University of Florida, Gainesville, FL,
United States*

Stephan M Haefele

Rothamsted Research, Harpenden, United Kingdom

Si Yang Han

*Sydney Institute of Agriculture, The University of Sydney,
Camperdown, NSW, Australia; School of Life and
Environmental Sciences, The University of Sydney,
Camperdown, NSW, Australia*

Januar Harianto

*Sydney Institute of Agriculture, The University of Sydney,
Camperdown, NSW, Australia; School of Life and
Environmental Sciences, The University of Sydney,
Camperdown, NSW, Australia; Sydney Informatics Hub,
The University of Sydney, Camperdown, NSW, Australia*

James B Harsh

Washington State University, Pullman, WA, United States

Alfred E Hartemink

*Department of Soil Science, University of
Wisconsin-Madison, Madison, WI, United States*

Ludger Herrmann

*Department of Soil Science and Land Evaluation (310a),
Universität Hohenheim, Stuttgart, Germany*

Gerard BM Heuvelink

*ISRIC – World Soil Information and Wageningen
University, Wageningen, The Netherlands*

John Hollis

The James Hutton Institute, Aberdeen, United Kingdom

Jingyi Huang

*Department of Soil Science, University of
Wisconsin-Madison, Madison, WI, United States*

Hermine Huot

*Laboratoire Interdisciplinaire des Environnements
Continentaux, Université de Lorraine – CNRS, Nancy,
France*

Juan José Ibañez

*National Museum of Natural History (MNCN), Spanish
National Research Council (CSIC), Madrid, Spain*

Alberto Vasconcellos Inda

*Department of Agronomy, Federal University of Rio
Grande do Sul (UFGRS), Porto Alegre, Rio Grande do Sul,
Brazil*

Nicolas A Jelinski

*Department of Soil, Water, and Climate, University of
Minnesota, Saint Paul, MN, United States*

Claudy Jolivet

INRAE, Info&Sols, Orléans, France

Scott B Jones

*Department of Plants, Soils, and Climate, Utah State
University, Logan, UT, United States*

Eugene F Kelly

*Department of Soil and Crop Sciences, Colorado State
University, Fort Collins, CO, United States*

Dmitry Konyushkov

*V.V. Dokuchaev Soil Science Institute, Russian Academy of
Sciences, Moscow, Russia*

Alberto Martínez-de la Torre

*UK Centre for Ecology & Hydrology (UKCEH),
Wallingford, United Kingdom; Meteorological Surveillance
and Forecasting Group, DT Catalonia, Agencia Estatal de
Meteorología (AEMET), Barcelona, Spain*

Holger Lange

*Norwegian Institute of Bioeconomy Research (NIBIO),
Høgskoleveien, Ås, Norway*

Guy J Levy
ARO, Volcani Institute, Rishon Lezion, Israel

Zamir Libohova
USDA-ARS Dale Bumpers Small Farms Research Center, Booneville, AR, United States

Allan Lilly
The James Hutton Institute, Aberdeen, United Kingdom

Xiaobing Liu
Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Harbin, PR China

Bernhard Lucke
FAU Erlangen-Nürnberg, Institute of Geography, Erlangen, Germany

Brendan Malone
CSIRO Agriculture and Food, Butler Laboratory, Black Mountain, Acton, ACT, Australia

Stephan Mantel
ISRIC – World Soil Information, Wageningen, The Netherlands

Toby Richard Marthews
UK Centre for Ecology & Hydrology (UKCEH), Wallingford, United Kingdom

Alex B McBratney
Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Camperdown, NSW, Australia

Susan E Melzer
Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, United States

Budiman Minasny
Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Camperdown, NSW, Australia

Jean Louis Morel
Laboratoire Sols et Environnement, Université de Lorraine – INRAE, Nancy, France

Uri Nachshon
ARO, Volcani Institute, Rishon Lezion, Israel

Edward A Nater
Department of Soil, Water, and Climate, University of Minnesota, Saint Paul, MN, United States

Attila Nemes
Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway; Division of Environment and Natural Resources, Norwegian Institute of Bioeconomy Research, Ås, Norway

Stephen Nortcliff
Soil Research Centre, School of Archaeology, Geography and Environmental Science, Reading, United Kingdom

Madlene Nussbaum
Bern University of Applied Sciences BFH, School of Agricultural, Forest and Food Sciences HAFL, Berne, Switzerland

Inakwu OA Odeh
University of Sydney, Sydney, NSW, Australia

Margaret A Oliver
Department of Geography & Environmental Science School of Human & Environmental Sciences, Soil Research Centre, The University of Reading, Reading, United Kingdom

Phillip R Owens
USDA-ARS Dale Bumpers Small Farms Research Center, Booneville, AR, United States

Eswaran Padmanabhan
Department of Geosciences, University Technology PETRONAS, Seri Iskandar, Malaysia

Jian Peng
Department of Remote Sensing, Helmholtz Centre for Environmental Research—UFZ, Permoserstrasse 15, Leipzig, Germany; Remote Sensing Centre for Earth System Research, Leipzig University, Leipzig, Germany

Marco Pfeiffer
Departamento de Ingeniería y Suelos, Facultad de Ciencias Agronómicas, Universidad de Chile, La Pintana, Chile

Chien-Lu Ping
Agriculture and Forestry Experiment Station, University of Alaska Fairbanks, Fairbanks, AK, United States

Rosa M Poch
Departament de Medi Ambient i Ciències del Sòl, Universitat de Lleida, Lleida, Spain

Laura Poggio
ISRIC World Soil Information, Wageningen, Netherlands

Céline Ratié
INRAE, Info&Sols, Orléans, France

Paul F Reich
USDA Natural Resources Conservation Service, Beltsville, MD, United States

Justin B Richardson
Department of Geosciences, University of Massachusetts-Amherst, Amherst, MA, United States

Mercedes Román Dobarco
Sydney Institute of Agriculture, The University of Sydney, Camperdown, NSW, Australia; School of Life and

*Environmental Sciences, The University of Sydney,
Camperdown, NSW, Australia*

N Romano
*Department of Agricultural Sciences, AFBE Division,
University of Naples Federico II, Portici (Napoli), Italy*

Cornelia Rumpel
*CNRS, Institute of Ecology and Environmental Sciences
Paris (IEES Paris), UMR Sorbonne Université, CNRS,
INRA, IRD, UPEC, Paris, France*

Jonathan A Sandor
Iowa State University, Ames, IA, United States

Marcel G Schaap
*Department of Environmental Science, College of
Agriculture and Life Sciences, University of Arizona,
Tucson, AZ, United States*

Andreas C Scheinost
*The Rossendorf Beamline at ESRF, Grenoble Cedex,
France; HZDR Institute of Radiochemistry, Dresden,
Germany*

Heike Schimmel
*Institute of Crop Science and Resource Conservation
(INRES), Soil Science and Soil Ecology, University of
Bonn, Bonn, Germany*

Darrell G Schulze
*Agronomy Department, Purdue University, West Lafayette,
IN, United States*

Sérgio Henrique Godinho Silva
*Department of Soil Science, Federal University of Lavras
(UFLA), Lavras, Minas Gerais, Brazil*

Balwant Singh
*Sydney Institute of Agriculture, School of Life and
Environmental Sciences, The University of Sydney, NSW,
Australia*

Nicolas Soler-Dominguez
INRAE, Info&Sols, Orléans, France

Tobias Sprafke
*Kompetenzzentrum Boden (Center of Competence for
Soils), BFH-HAFL, Zollikofen, Switzerland; Institute of
Geography, University of Bern, Bern, Switzerland*

Karl Stahr
*Department of Soil Science and Land Evaluation (310a),
Universität Hohenheim, Stuttgart, Germany*

Uta Stockmann
*Commonwealth Scientific and Industrial Research
Organisation (CSIRO), Agriculture and Food Business
Unit, Canberra, NSW, Australia*

Georges Stoops
Vakgroep Geologie, Universiteit Gent, Gent, Belgium

James A Taylor
*UMR ITAP, L'Institut Agro-Montpellier, INRAE,
University of Montpellier, Montpellier, France*

John CF Tedrow
*Posthumous Contribution, formerly with Rutgers
University, New Brunswick, NJ, United States*

Arnaud JAM Temme
*Department of Geography and Geospatial Sciences, Kansas
State University, Manhattan, KS, United States*

Michael Thompson
Iowa State University, Ames, IA, United States

Mikaela Tilse
*Sydney Institute of Agriculture, The University of Sydney,
Camperdown, NSW, Australia; School of Life and
Environmental Sciences, The University of Sydney,
Camperdown, NSW, Australia*

J Torrent
*Departamento de Agronomía, Universidad de Córdoba,
Córdoba, Spain*

Markus Tuller
*Department of Environmental Science, The University of
Arizona, Tucson, AZ, United States*

G Uehara
University of Hawaii, Honolulu, HI, United States

Rota Wagai
*Institute for Agro-Environmental Science, National
Agriculture and Food Research Organization, Tsukuba,
Ibaraki, Japan*

Cornelius Wilhelm van Huyssteen
University of the Free State, Bloemfontein, South Africa

Bas van Wesemael
*Earth and Life Institute, Université Catholique de Louvain,
Ottignies-Louvain-la-Neuve, Belgium*

MJ Vepraskas
*Department of Crop and Soil Sciences, North Carolina
State University, Raleigh, NC, United States*

Alexandre MJ-C Wadoux
*Sydney Institute of Agriculture & School of Life and
Environmental Sciences, The University of Sydney, Sydney,
NSW, Australia*

Tetsuhiro Watanabe
Kyoto University, Kyoto, Japan

Richard Webster
Rothamsted Research, Harpenden, United Kingdom

Scott Werts
*Department of Chemistry, Physics, and Geology, Winthrop
University, Rock Hill, SC, United States*

Skye A Wills
*Natural Resources Conservation Service, Lincoln, NE,
United States*

Caroline M Yonker
*Department of Soil and Crop Sciences, Colorado State
University, Fort Collins, CO, United States*

Mengyang You
*School of Environmental Science and Engineering,
Guangzhou University, Guangzhou, PR China; Northeast*

*Institute of Geography and Agroecology, Chinese Academy
of Sciences, Harbin, PR China*

Yakun Zhang
*Department of Soil Science, University of
Wisconsin-Madison, Madison, WI, United States*

Yonggen Zhang
*School of Earth System Science, Tianjin University, Tianjin,
P.R. China*

CONTENTS OF ALL VOLUMES

VOLUME 1: SOIL BIOLOGY

Overview

Edaphology	1
<i>RM Goss and AL Ulery</i>	

Biodiversity	10
<i>Monica A Farfan and Diana H Wall</i>	

Soil organisms

Bacteriophage (Viruses)	16
<i>Li Bi, Ji-Zheng He, Dan-Ting Yu, Li-Li Han, and Hang-Wei Hu</i>	

Archaea	25
<i>Mikhail V Semenov</i>	

Bacteria	31
<i>Mikhail V Semenov</i>	

Soil protists	39
<i>Stefan Geisen, Nathalie Amacker, and Alexandre Jousset</i>	

Fungi including mycorrhizal fungi	49
<i>F Buscot</i>	

Soil algae	61
<i>Karen Baumann and Patrick Jung</i>	

Soil arthropods: Underfoot and all around	70
<i>Ernest C Bernard</i>	

Nematodes	105
<i>Deborah A Neher and Thomas O Powers</i>	

Earthworms: Essential ecosystem engineers providing vital ecosystem services	112
<i>Maria JI Briones</i>	

Complex structures

Soil biofilms: Evolving concepts and ecological functions	120
<i>Yichao Wu, Marc Redmile-Gordon, Peng Cai, Chunhui Gao, and Qiaoyun Huang</i>	

Biocrusts	127
<i>Matthew A Bowker, Jasmine Anenberg, Hilina Yohannes Kebede, Anita Antoninka, and Lydia Bailey</i>	

Functioning

Studying soil microbial communities: Culture-dependent approaches	139
<i>Eduardo K Mitter, Dasiel Obregon, Micaela Tosi, Heather Slinn, and Kari Dunfield</i>	

Studying soil microbial communities: Culture-independent approaches	146
<i>Tolulope Mafa-Attoye, Heather Slinn, Micaela Tosi, Evan Mayer, Dasiel Obregon, and Kari Dunfield</i>	

Studying microbial function in soil: Culture-independent techniques <i>Micaela Tosi, Dasiel Obregon, Heather Slinn, and Kari Dunfield</i>	155
Kinetics of microbial processes: General principles <i>NS Panikov</i>	168
Kinetics of microbial processes: Applications to soil biology <i>NS Panikov</i>	186
Biogeochemical cycles	
Carbon cycle in soils: Dynamics and management <i>Charles W Rice, Carlos B Pires, and Marcos VM Sarto</i>	219
Carbon storage in soils <i>Isabelle Basile-Doelsch, Tiphaine Chevallier, Marie-France Dignac, and Amandine Erktan</i>	228
Soil organic nitrogen <i>EJ Daly and KA Congreves</i>	243
Nitrification: Process, organisms and management <i>Jeanette M Norton and Yang Ouyang</i>	253
Microbial denitrification <i>Michael J Goss, Sergey Blagodatsky, and Adrian Unc</i>	261
Symbiotic nitrogen fixation <i>Florence Mus and Hsin-Hua Wu</i>	270
Nitrogen: Non-Symbiotic Nitrogen fixation in soils <i>Vadakattu VSR Gupta and Sarah S Roley</i>	283
Phosphorus in soils—Biological interactions <i>Nataliya Bilyera and Iryna Loginova</i>	293
Biological transformations of sulfur in soil <i>Chaoqun Wang and Maxim Dorodnikov</i>	303
Plant uptake of mineral nutrients <i>Angela Hodge</i>	310
Higher plants and their interactions with soil organisms	
Soil Biology: Root form and function <i>Magda Garbowski, Grégoire Freschet, Louise Jackson, Cynthia Brown, and Louise Comas</i>	321
Root mucilage: Chemistry and functions in soil <i>Yutong Jiang, Lijun Hou, Fang Ding, and Joann K Whalen</i>	332
Root exudates and microorganisms <i>Aidan De Sena, Krisztina Mosdossy, Joann K Whalen, and Chandra A Madramootoo</i>	343
Root interactions with the microbiome from the rhizoplane to the bulk soil: An overview <i>Joann K Whalen and Shamim Gul</i>	357
Microorganisms & biological processes	
Soil respiration <i>Jeremiah D Vallotton, Sergey Blagodatsky, and Adrian Unc</i>	369
Anaerobic soil processes <i>PW Inglett, D Said-Pullicino, and K Ramesh Reddy</i>	379
Mineral-organic-microbial interactions <i>Balwant Singh, Sabina Yeasmin, and Donald L Sparks</i>	387

Microbial ecology & management

Habitat for soil organisms and their functions <i>Alexandra Kravchenko and Andrey Guber</i>	407
Plant growth promoters <i>Elisa Gamalero, Gustavo Santoyo, and Bernard R Glick</i>	418
Inoculation with microorganisms <i>Maureen O'Callaghan and David Wright</i>	428
Biocontrol of soil-borne plant diseases <i>Mehrdad Mohammadiani and Matthew Bakker</i>	439
Management of soil biology for crop protection from biotic and abiotic stresses <i>Michael J Goss, Isabel Brito, and Mário Carvalho</i>	447
Trees, hedges, agroforestry and microbial diversity <i>Josiah B Judson, Joseph Holden, Pippa Chapman, and Marcelo Valadares Galdos</i>	469
Urban soil microbiome: Activity, diversity and functioning <i>Maria Korneykova, Ekaterina Kozlova, Kristina Ivashchenko, Maria Vasilieva, and Viacheslav Vasenev</i>	480
Microbiology of extreme soil environments <i>Patrice Dion</i>	494
Microbial mobility and transport in soils <i>Mohammad Reza Mosaddeghi, Adrian Unc, Ali Akbar Mahboubi, and Azadeh Safadoust</i>	512

Biology of soil amendments & contaminant degradation

Biology of compost <i>Heribert Insam, Thomas Klammersteiner, and María Gómez-Brandón</i>	522
Landfarming and its microbiology <i>Naga Raju Maddela, Balasubramanian Ramakrishnan, Kadiyala Venkateswarlu, and Mallavarapu Megharaj</i>	533
Biodegradation of organic contaminants <i>PB Hatzinger and JW Kelsey</i>	547
Biological aspects of manure management <i>Natasha Rayne and Bee Khim Chim</i>	558
Septic systems and municipal waste disposal on land <i>Charles P Gerba, Michael J Goss, and Adrian Unc</i>	567

Climate change

Climate change impacts on soil biology <i>Alberto Canarini, Lucia Fuchslueger, François-Xavier Joly, and Andreas Richter</i>	578
---	-----

VOLUME 2: CHEMISTRY**Basic Soil Chemistry**

Acidity and acidification <i>Siobhan Staunton</i>	1
pH <i>Siobhan Staunton</i>	8
Buffering capacity <i>Siobhan Staunton and Ravi Naidu</i>	15
Soil specific surface area <i>Emmanuel Arthur, Markus Tuller, and Kurt D Pennell</i>	22

Soil Analytical Chemistry

Cation exchange capacity and reactions <i>Donald S Ross and Ulf Skyllberg</i>	32
Testing for nutrients in the soil and the environment <i>Antonio P Mallarino</i>	43
Metal extraction and analysis <i>Guo Wang and Honghong Li</i>	50
Soil organic matter: Composition <i>Markus Kleber</i>	58
Physical fractionation techniques <i>Sam J Leuthold, Michelle L Haddix, Jocelyn Lavallee, and M Francesca Cotrufo</i>	68
The use of stable isotopes in soil science: Low atomic number elements <i>Chiara Pistocchi, Isabelle Bertrand, Kristýna Kantnerová, Astrid Oberson, Federica Tamburini, and Roland A Werner</i>	81
The use of radioactive isotopes to study soil pools and processes <i>Timothy I McLaren</i>	96
The use of stable isotopes in soil science: Metals <i>Cécile Quantin and Damien Guinoiseau</i>	107

Periodic Table

Alkali metals in soils <i>Guy J Levy and Asher Bar-Tal</i>	116
Alkaline earth metals in soil <i>Justin B Richardson</i>	126
Transition metals in soil <i>Siobhan Staunton</i>	135
Nitrogen and ammonia in soils <i>Nicholas T Girkin and Hannah V Cooper</i>	142
Phosphorus in soil <i>Luisella Celi, Maria Martin, and Elisabetta Barberis</i>	152
Aluminum in soils <i>Syuntaro Hiradate</i>	167
Iron in soils <i>Trolard Fabienne and Guilhem Bourrié</i>	176
Sulfur and selenium in soil <i>Dongli Liang, Fei Zhou, and Mingxing Qi</i>	192
Arsenic in soil <i>Md Imran Ullah Sarkar, Afsana Jahan, Md Tofail Hossain, Mohammad Mahmudur Rahman, and Ravi Naidu</i>	201
Non-transition elements: Metals and metalloids in soils <i>Natasha, Sana Khalid, Muhammad Shahid, Faiz Rabbani, and Camille Dumat</i>	214
Lanthanides, actinides and radon <i>Franco Ajmone Marsan and Mery Malandrino</i>	225
Halogens in soils <i>Svensson Teresia, Thiry Yves, Bueno Maïté, and Oelmann Yvonne</i>	234
Transformation of metals and metalloids by microorganisms <i>Susan M Glasauer, Euan P Burford, and Geoffrey Michael Gadd</i>	243
Dynamics of radionuclides <i>Siobhan Staunton</i>	256
Silicon and other non-metal elements from Group 14 <i>Bruno Delvaux and Zimin Li</i>	263

Soil Physical Chemistry

Chemical equilibria	274
<i>Arthur Paul Schwab</i>	
Dissolution and precipitation processes	284
<i>Zhenqing Shi, Shiwen Hu, Rong Li, and Yang Lu</i>	
Reactions, redox potential and kinetics	297
<i>Trolard Fabienne and Guilhem Bourrié</i>	
Sorption of organic chemicals in soil	315
<i>Joseph J Pignatello</i>	
Sorption—Oxyanions	327
<i>CP Schulthess, H Wijnja, W Yang, Anahita Khosravi, and Baoshan Xing</i>	
Sorption—Metals	336
<i>Daniel G Strawn and Donald L Sparks</i>	
Sorption–desorption kinetics	343
<i>Gautier Landrot and Donald L Sparks</i>	
Sorption-desorption hysteresis	355
<i>Wenhao Wu, Hongxia Zhu, and Kun Yang</i>	
Surface charge and zero-charge points	368
<i>J Chorover</i>	
Flocculation and dispersion	376
<i>Guy J Levy and Isaac Shainberg</i>	

Soil Carbon Chemistry

Soil organic matter dynamics	387
<i>Yu Luo and Jianming Xu</i>	
Soil proteins	396
<i>Siobhan Staunton</i>	
Biochar: Properties, mechanisms, and interactions in the soil	406
<i>Balwant Singh, Shamim Mia, and Marta Camps-Arbestain</i>	
Dissolved organic carbon	421
<i>Jason James and Rob Harrison</i>	
Black carbon, pyrogenic carbon	432
<i>Heike Knicker</i>	
Fire effects on soil	448
<i>Stefan H Doerr, Cristina Santín, and Jorge Mataix-Solera</i>	
Interfacial reactions of microorganisms with minerals and organic matter	458
<i>Rota Wagai, Puu-Tai Yang, and Klaus Kaiser</i>	
Pesticide fate in soils	470
<i>Pierre Benoit, Laure Mamy, Carole Bedos, and Enrique Barriuso</i>	
Soil organic matter in molecular simulations	483
<i>Martin H Gerzabek, Daniel Tunega, Edgar Galicia-Andrés, and Chris Oostenbrink</i>	

Soil Physical Analysis

Raman spectroscopy	497
<i>Anupam Das and Huiyuan Guo</i>	
Fourier transform infrared spectroscopic methods of soil analysis	510
<i>Derek Peak</i>	
Electron paramagnetic resonance spectroscopy: Part I Historic Perspectives	519
<i>Giorgio S Senesi and Nicola Senesi</i>	

Electron paramagnetic resonance spectroscopy: Part II the view forward <i>Robert L Cook, Emmanuel Ampiah, Phillip T Sprunger, and Randall W Hall</i>	533
Fluorescence Spectroscopy: Part I Principles <i>Yong Zhang, Shuai Guo, and Chaoxian Wei</i>	544
Fluorescence spectroscopy: Part II practical applications <i>Yong Zhang, Chaoxian Wei, and Shuai Guo</i>	552
Near and mid infrared soil spectroscopy <i>Wartini Ng, Brendan Malone, Budiman Minasny, and Sangho Jeon</i>	567
Electron microscopic methods (TEM, SEM and energy dispersal spectroscopy) <i>Jian Zhao and Xia Liu</i>	575
Application of atomic force microscopy (AFM) on environmental interfaces <i>Saikat Ghosh, Zhang Di, Bo Pan, and Baoshan Xing</i>	589
Applications of nuclear magnetic resonance for the study of soils <i>Daniel H Lysak, Myrna J Simpson, and Andre J Simpson</i>	605
Neutron-gamma analysis of soil elements <i>Aleksandr Kavetskiy, Galina Yakubova, Nikolay Sargsyan, Stephen A Prior, and H Allen Torbert</i>	625
X-Ray Diffraction <i>Chao Wang and Cheng Gu</i>	642
Soil analysis using laser-induced breakdown spectroscopy <i>Débora Marcondes Bastos Pereira Milori, Ladislau Martin-Neto, Hero T Gollany, Diego Victor de Babos, Luis Carlos Leva Borduchi, and Paulino Ribeiro Villas-Boas</i>	654
Calorimetric analysis <i>Xingmin Rong and Qiaoyun Huang</i>	662
Understanding of soil processes at the microscale—Use of NanoSIMS in soil science <i>Carsten W Mueller, Carmen Hoeschen, and Ingrid Koegel-Knabner</i>	670
Nanoparticles in soil <i>Di Lang, Quan Chen, Yang Liu, and Bo Pan</i>	681
Overview	
Soil chemistry and climate change <i>Jianming Xu, Yong Li, and Zhongmin Dai</i>	692
VOLUME 3: ENVIRONMENT AND MANAGEMENT	
Soil functions, potential uses of soil and ecosystem services	
Soil functions, soil potential uses and ecosystem services <i>Karl Stahr</i>	1
Farming systems to conserve land for nature <i>Rattan Lal</i>	12
Geoarchaeology and soils <i>Carlos E Cordova</i>	20
Civilizations and the role of soil, agriculture, and climate change <i>Jeffrey A Homburg, Eric C Brevik, and Jonathan A Sandor</i>	29
Ecosystem services provided by soils in highly anthropized areas (SUITMAs) <i>Jean Louis Morel, Geoffroy Séré, Viatcheslav Vasenev, and Thomas Nehls</i>	37
Structure and function of forested soils <i>Peter Clinton and Jen Owens</i>	56
Agroforestry <i>NV Thevathasan, AM Gordon, and PKR Nair</i>	68

Rangeland management <i>Alexandra Sandhage-Hofmann</i>	88
Effects on soil of grassland management for pasture, hay and silage <i>Carmen Segura, Laura Cardenas, Richard McDowell, Sarah Morgan, and Martin SA Blackwell</i>	102
Dryland farming: Technological and management options for sustainable agriculture and food systems <i>Ch Srinivasarao, G Ranjith Kumar, Ravula Manasa, Kiran Pilli, Samaresh Sahoo, S Rakesh, Sumanta Kundu, K C Nataraj, KV Rao, JVNS Prasad, and S Malleswari</i>	113
The effect of crop rotations on soil <i>Caio Fernandes Zani, Arlete Simões Barneze, Rogério Peres Soratto, and Charles Andrew Francis</i>	125
The contribution of organic farming systems to soil fertility—A systems perspective <i>Bernhard Freyer, Pierre Ellssel, Jürgen K Friedel, and Kurt Möller</i>	135
Soil functions to support natural plant communities and crops: Soil multifunctionality <i>Leidivan Almeida Frazão, Luiz Arnaldo Fernandes, Adriana Marcela Silva Olaya, and Igor Costa de Freitas</i>	146
Methane and nitrous oxide emissions from land <i>Joanna M Cloy and Keith A Smith</i>	153
Soil and land degradation	
Overview chapter on soil degradation <i>K Müller, Margaret A Oliver, and C Siebe</i>	165
Soil monitoring <i>Ben Marchant and Nicolas Saby</i>	172
Soil quality and health key indicators <i>Edoardo AC Costantini and Simone Priori</i>	181
Water induced soil erosion <i>R Jane Rickson</i>	193
Erosion by wind <i>John Tatarko</i>	208
Human exposure to aeroparticles emitted from unpaved roads due to traffic resuspension and wind erosion <i>Arón Jazcilevich and Ann G Wellens</i>	223
Soil salination processes and management <i>Uri Nachshon and Guy J Levy</i>	236
Degradation of soil structure and soil organic matter <i>Andrew S Gregory</i>	246
Drainage, soil water storage, buffering and filtering <i>Jeffrey Strock and Norm Fausey</i>	257
Environmental effects of irrigation <i>Sevilay Topcu</i>	266
Trade-offs of wastewater irrigation <i>Manzoor Qadir and Christopher A Scott</i>	277
Pollution and risks of trace elements in the soil environment <i>Peng Wang, Fang-Jie Zhao, and Steve P McGrath</i>	288
Pollutants—Persistent organic <i>Rai S Kookana and Divina A Navarro</i>	296
Pollution by microplastic in agricultural areas <i>Nicolas Beriot and Esperanza Huerta-Lwanga</i>	309

Municipal solid waste management: Production, management, and environmental effects <i>Bryant M Serre and Lynda H McCarthy</i>	321
Remediation of metals and organic contaminants in soil <i>Peter Sanderson, Md Mezbaul Bahar, Bhabananda Biswas, and Ravi Naidu</i>	333
Soil change under different scenarios	
Soil change under different scenarios <i>Joann K Whalen</i>	344
The effect of global change on the soil body <i>Sophie Cornu, David Montagne, and Antra Boca</i>	360
The effects of adequate and excessive application of mineral fertilizers on the soil <i>Bijay-Singh and Tek B Sapkota</i>	369
Foliar Application of Nutrients <i>Thomas Eichert and Victoria Fernández</i>	382
Mulches and cover crops part I: Types <i>KK Bandyopadhyay, CL Acharya, and KM Hati</i>	392
Mulches and cover crops part II: Role in soil health and climate resilient agriculture <i>KK Bandyopadhyay, CL Acharya, and KM Hati</i>	401
The potential agronomic and environmental applications of biochar: Prospects and challenges <i>Yu-Jun Wang, Ping Wu, Nanthi S Bolan, and Hai-Long Wang</i>	414
Soil management for carbon sequestration <i>Christopher Brock, Uwe Franko, and Martin Wiesmeier</i>	424
Soil tillage <i>Mathieu Lamandé, Lars Juhl Munkholm, and Trond Børresen</i>	433
Weed and soil management: A balancing act <i>Trevor Kenneth James and Charles Norman Merfield</i>	439
Terraces as traditional agricultural landscapes. Stability amidst change <i>Gerardo Bocco</i>	450
Soil quality protection policies and plans to ensure sustainability <i>Ram Swaroop Meena and Sandeep Kumar</i>	457
Construction of Technosols for green roofs and vertical green system <i>Eleonora Flores-Ramírez, Stefan Abel, Elizabeth Chávez-García, and Geoffroy Séré</i>	473
Management of park areas, sport fields, and school yards, including golf courses and public right of ways <i>Maxine J Levin and Anna Paltseva</i>	488
Soil-society relations	
Overview chapter on soil-society relations <i>C Siebe, Margaret A Oliver, and K Müller</i>	505
Soil art: Sensory and symbolic engagement with soils <i>Alexandra R Toland and Daniel Wolter</i>	509
Soil science in the criminal justice system <i>Lorna Anne Dawson</i>	521
Soil governance at the international, regional and national level <i>Harald Ginzky</i>	532
The economics of soils' contribution to human well-being <i>Bartosz Bartkowski and Julian R Massenber</i>	547

Soil and human health	555
<i>Margaret A Oliver and Eric C Brevik</i>	
Soil Science Education, Awareness and Outreach in the Modern World	572
<i>Laura Bertha Reyes-Sánchez and Eric C Brevik</i>	
Soils in landscape planning	580
<i>Seth Denizen</i>	

VOLUME 4: PEDOLOGY

Pedology

Pedology—Basic principles	1
<i>Rupert Bäumler, Uta Stockmann, and Margaret A Oliver</i>	

Factors of soil formation, pedogenic processes and mineralogy

Overview chapter on factors of soil formation	7
<i>Rupert Bäumler, Margaret A Oliver, and Uta Stockmann</i>	
Factors of soil formation: Climate	14
<i>Stefaan Dondeyne, Stephan Mantel, and Seppe Deckers</i>	
Factors of soil formation: Soil biota and organic matter as a factor for soil formation and weathering processes	25
<i>Scott Werts</i>	
Soil forming factors: Parent material	34
<i>Stephen Nortcliff</i>	
Parent materials: Loess	42
<i>Tobias Sprafke</i>	
Time	48
<i>Susan E Melzer, Caroline M Yonker, and Eugene F Kelly</i>	
Factors of soil formation—Human impacts	54
<i>Jonathan A Sandor, C Lee Burras, Michael Thompson, and Skye A Wills</i>	
Dynamic soils	69
<i>Markus Egli and Dennis Dahms</i>	
Interactions between soils and climate change	82
<i>Cornelia Rumpel</i>	
Primary minerals	91
<i>Balwant Singh</i>	
Mineral components of the soil clay fraction	109
<i>Darrell G Schulze</i>	
Short-range ordered aluminosilicates	121
<i>Tetsuhiro Watanabe, James B Harsh, and Rota Wagai</i>	
Metal oxides	135
<i>Andreas C Scheinost and Balwant Singh</i>	
Soil morphology	149
<i>Phillip R Owens and Zamir Libohova</i>	
Soil micromorphology	171
<i>Rosa M Poch and Georges Stoops</i>	
Soil Classification and distribution	
History and principles of soil classification	185
<i>Maria Gerasimova and Dmitry Konyushkov</i>	

Soil classification systems: World reference base and soil taxonomy	197
<i>David Badía-Villas</i>	
World reference base for soil resources (WRB)	206
<i>Stephan Mantel, Stefaan Dondeyne, and Seppe Deckers</i>	
World soil map based on soil taxonomy	218
<i>Eswaran Padmanabhan and Paul F Reich</i>	
Soils as archives of city history	232
<i>Hermine Huot and Jean Louis Morel</i>	
Classification of soils by ecozones	
Ecozones and soils – An introduction	246
<i>Rupert Bäumler</i>	
Cold Region Soils: Part I—Distribution, Properties and Processes	250
<i>Nicolas A Jelinski, Chien-Lu Ping, and John CF Tedrow</i>	
Cold region soils: Part II – Genesis and classification	263
<i>Nicolas A Jelinski, Chien-Lu Ping, and John CF Tedrow</i>	
Boreal forest soils	280
<i>Gabriele Broll</i>	
Soils of humid cool temperate regions	289
<i>Nicolas A Jelinski, Justin B Richardson, and Edward A Nater</i>	
Grassland soils in the cool–arid–temperate ecozone (Steppe, grassland)	299
<i>Mengyang You, Xiaobing Liu, Rupert Bäumler, and Margaret A Oliver</i>	
Soils from a warm-temperate semi-tropical ecozone (Mediterranean) with humid winter months	319
<i>J Torrent, Marco Pfeiffer, and Juan José Ibañez</i>	
Arid and semi-arid subtropical and tropical ecozone (II)	332
<i>Bernhard Lucke</i>	
Humid subtropical ecozone with emphasis on Acrisols and Alisols	343
<i>Nilton Curi, Sérgio Henrique Godinho Silva, and Alberto Vasconcellos Inda</i>	
Soils of the humid and sub-humid tropics	353
<i>Karl Stahr and Ludger Herrmann</i>	
Classification of soils by properties and environmental settings	
Sodic soils	368
<i>Guy J Levy and Uri Nachshon</i>	
Organic soils	383
<i>Heike Schimmel and Wulf Amelung</i>	
The pedology of anaerobic soils	398
<i>John Hollis and Allan Lilly</i>	
Alluvium and alluvial soils	417
<i>Janis L Boettinger</i>	
Hydric Soils	424
<i>MJ Vepraskas and JF Berkowitz</i>	
Wetland soils	436
<i>Cornelius Wilhelm van Huyssteen</i>	
Paddy soils: Characteristics, productivity and management	447
<i>Stephan M Haefele</i>	
Volcanic soils and non-volcanic andic soils	458
<i>Rupert Bäumler and G Uehara</i>	

PEDOMETRICS**Pedometrics**

- Overview of Pedometrics 471
Alexandre MJ-C Wadoux, Inakwu OA Odeh, and Alex B McBratney

Statistics and soil data analysis

- Statistics in soil science 486
Richard Webster

- Multivariate analysis for soil science 499
Si Yang Han, Patrick Filippi, Mercedes Román Dobarco, Januar Harianto, Mathew S Crowther, and Thomas FA Bishop

- Machine learning and processing of large data 509
Madlene Nussbaum

- The issue of scale and change of support in the spatial analysis of environmental data 521
A Castrignanò, A Belmonte, and N Romano

Soil measurements in space and time

- Sampling for soil survey 535
DJ Brus

- Handling of soil samples in the field, in transport to the laboratory and in the laboratory 551
Claudy Jolivet, Céline Ratié, Déborah Chavrit, Nicolas Soler-Dominguez, and Antonio Bispo

- Digital Soil Morphometrics 568
Yakun Zhang, Alfred E Hartemink, Jingyi Huang, and Budiman Minasny

- Proximal soil sensing in the field 579
Mikaela Tilse, Uta Stockmann, and Patrick Filippi

- Proximal sensing of soil moisture 591
Markus Tuller, Ebrahim Babaeian, and Scott B Jones

- Proximal sensing of land surface temperature 600
Ebrahim Babaeian and Markus Tuller

- Proximal sensing of evapotranspiration 610
Ebrahim Babaeian and Markus Tuller

- Remote sensing of soil moisture 618
Alexander Gruber and Jian Peng

- Remote Sensing of Soil Organic Carbon 631
Bas van Wesemael and Sabine Chabrilat

- Pedotransfer functions and their application to soil water dynamics 642
Marcel G Schaap, Yonggen Zhang, and Attila Nemes

Spatial and temporal soil variation

- Geostatistics: Modelling spatial variation 655
Richard Webster

- Uncertainty assessment of spatial soil information 671
Gerard BM Heuvelink and Richard Webster

- Digital soil mapping: Evolution, current state and future directions of the science 684
Brendan Malone, Dominique Arrouays, Laura Poggio, Budiman Minasny, and Alex B McBratney

- Geographical information systems (GIS) and soils 696
S Grunwald and J Böhner

- Precision agriculture 710
James A Taylor

Foundations, measurements and trends in pedodiversity <i>Juan José Ibañez and Marco Pfeiffer</i>	726
Soil landscape modelling	
Modeling soil development in a landscape context <i>Peter A Finke, Budiman Minasny, and Arnaud JAM Temme</i>	739
Climate and land surface models: Role of soil <i>Toby Richard Marthews, Holger Lange, Alberto Martínez-de la Torre, Richard J Ellis, Sarah E Chadburn, and Martin G De Kauwe</i>	751
VOLUME 5: SOIL PHYSICS	
Soil structure & mechanics	
Soil structure <i>Steffen Schlüter and John Koestel</i>	1
Aggregation <i>Teamrat A Ghezzehei</i>	8
Porosity and pore-size distribution <i>John R Nimmo, Sheela Katuwal, and Maik Lucas</i>	16
Pore-scale modelling of flow and transport phenomena in soils <i>Kirill M Gerke and Marina V Karsanina</i>	25
Soil texture: Its relevance, measurement and representation internationally <i>Attila Nemes</i>	35
Soil tilth <i>Thomas Keller, Douglas L Karlen, and Paul D Hallett</i>	48
Crusts and seals: Structural <i>Shmuel Assouline and Sally Thompson</i>	57
Soil mechanics <i>Markus Berli and Paul D Hallett</i>	69
Compaction <i>JJH van den Akker, L ten Damme, M Lamandé, and Thomas Keller</i>	85
Bioturbation—Physical processes <i>Siul Ruiz, Paul D Hallett, and Dani Or</i>	100
The reinforcement of soil by plant roots <i>Alexia Stokes and Anil Yildiz</i>	115
Physical effects of soil salinity <i>David Russo</i>	123
Shrink-Swell processes in soils <i>Pascal Boivin</i>	139
Water retention	
Water properties <i>N Lu, C Zhang, and D Hillel</i>	146
Water potential <i>Dani Or, Markus Tuller, and Jon M Wraith</i>	159
Water content and water potential measurement <i>Wolfgang Durner</i>	168
Soil water retention and characteristic curve <i>Markus Tuller and Dani Or</i>	187
Soil water repellency <i>Jörg Bachmann and Marc-Oliver Goebel</i>	203

Flow & transport processes

Aeration	216
<i>Rodney T Venterea and Dennis E Rolston</i>	
Capillarity	225
<i>Dani Or and Markus Tuller</i>	
Darcy's law	240
<i>Conrad Jackisch and Eva Kröner</i>	
Diffusion	247
<i>TKK Chamindu Deepagoda and Shoichiro Hamamoto</i>	
Hydrodynamics in soil	256
<i>Ty Ferre and Art Warrick</i>	
Thermodynamics of soil water	262
<i>Pieter H Groenevelt</i>	
Immiscible Fluids	271
<i>Yukie Tanino and Dubravka Pokrajac</i>	
Overland flow	287
<i>Tammo S Steenhuis, Rajith Mukundan, and Seifu A Tilahun</i>	
Direction-dependency of soil hydraulic and mechanical properties	294
<i>Steffen Beck-Broichsitter, Heiner Fleige, Jose Dörner, and Rainer Horn</i>	
Macropores and macropore flow	306
<i>Nicholas Jarvis and Mats Larsbo</i>	
Unstable preferential flow in porous media	314
<i>Naaran Brindt, J-Yves Parlange, and Tammo S Steenhuis</i>	
Leaching processes	321
<i>Brent Clothier and Steve Green</i>	
Colloid transport	336
<i>Yan Jin, Chongyang Shen, and Volha Lazouskaya</i>	
Scaling of transport properties	353
<i>Robert Ewing and Behzad Ghanbarian</i>	
Soil influence on groundwater	364
<i>David Rudolph and Edwin Cey</i>	

Thermal properties and behaviour

Temperature effects of soil hydraulic properties	379
<i>Jan W Hopmans and Robert Horton</i>	
Heat and vapor transfer	385
<i>Sunny Eng Giap Goh, Joshua L Heitman, and Robert Horton</i>	
Radiation Balance	393
<i>Jerry L Hatfield, Kaitlin Togliatti, Thomas J Sauer, and John H Prueger</i>	
Freezing and thawing cycles	400
<i>Alan W Rempel</i>	
Arctic permafrost	410
<i>Julia Boike, Andrey Abramov, Katrina E Bennett, and Lars Kutzbach</i>	
Permafrost and cold region hydrology	419
<i>Masaki Hayashi and Ylva Sjöberg</i>	

Sensing Techniques

Soil water sensing by neutron scattering	427
<i>Steven R Evett</i>	

Time-domain reflectometry	436
<i>GC Topp and Ty Ferre</i>	
Geophysical methods for soil applications	444
<i>Sarah Garré, Guillaume Blanchy, David Caterina, Philippe De Smedt, Alejandro Romero-Ruiz, and Nataline Simon</i>	
Non-invasive imaging of soil processes	459
<i>Craig J Sturrock</i>	
Hydrology & water balance	
Soil processes in the hydrologic cycle	469
<i>Cosimo Brogi, Harry Vereecken, Heye Reemt Bogena, and Luca Brocca</i>	
Global water cycle from a soil perspective	482
<i>Michael H Young and Dani Or</i>	
Water evaporation from bare soil	490
<i>Peter Lehmann</i>	
Evapotranspiration	501
<i>Yuting Yang</i>	
Plant available water	509
<i>Quirijn de Jong van Lier, Sally D Logsdon, Everton Alves Rodrigues Pinheiro, and Paulo Ivonir Gubiani</i>	
Plant-water relations	516
<i>Ian C Dodd, Andrew D Hirons, and Jaime Puértolas</i>	
Water use efficiency: A review of spatial and temporal variability	527
<i>Simone Fatichi, Athanasios Paschalis, Sara Bonetti, Gabriele Manoli, and Christoforos Pappas</i>	
Root water uptake and myco-rhizosphere hydraulic properties	543
<i>Andrea Carminati, Maria Marin, Valentin Couvreur, Mohsen Zarebanadkouki, Stéphane Declerck, and Mathieu Javaux</i>	
Plant-soil-water relations	552
<i>Jan Vanderborght, Quirijn de Jong van Lier, Andrea Schnepf, and Mathieu Javaux</i>	
Crop and landscape water requirements	571
<i>Luis S Pereira, Isabel Alves, and Paula Paredes</i>	
Water harvesting	593
<i>Wolde Mekuria and Desalegn Tegegne</i>	
Irrigation methods	608
<i>Shmulik P Friedman</i>	
Soil–plant–atmosphere continuum	624
<i>John M Norman, Martha C Anderson, and William P Kustas</i>	
The Penman-Monteith equation	634
<i>Richard G Allen, Ayse Kilic, Guido D'Urso, and Simona Consoli</i>	
Infiltration	646
<i>Ty Ferre and Art Warrick</i>	
Solute transport	652
<i>MC Sukop and M Dentz</i>	
Lysimetry	667
<i>Thomas Pütz and Jannis Groh</i>	
Recognition of notable past soil scientists	680
<i>Michael J Goss, Evgenia Blagodatskaya, Ariena HC van Bruggen, Monika Joschko, Stefan Schrader, and Christoph Tebbe</i>	
<i>Index</i>	691

Pedology—Basic principles

Rupert Bäumler^a, Uta Stockmann^b, and Margaret A Oliver^c, ^aInstitute of Geography, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany; ^bCommonwealth Scientific and Industrial Research Organisation (CSIRO), Agriculture and Food Business Unit, Canberra, ACT, Australia; ^cDepartment of Geography & Environmental Science School of Human & Environmental Sciences, Soil Research Centre, The University of Reading, Reading, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

This is an update of MJ Singer, PEDOLOGY | Basic Principles, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 151–156, ISBN: 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00001-1>

Introduction	1
Pedology	1
Soil formation (pedogenesis)	2
Importance of pedological research	5
Concluding remarks	5
References	6

Abstract

This chapter encompasses the basic principles of pedology, the study of soils in their natural setting. It provides basic information about the formation and properties of soils from any kind of parent material, the factors that influence soil formation at the Earth's surface, the induced processes, and the role of soils in the environment.

Key points

- Factors of soil formation—climate, landscape, vegetation, parent material and human influences.
- Translocation and transformation processes.
- Origin, development, properties, soil threats.

Introduction

The interest in soils goes back to Mesolithic and Neolithic times when humans recognized that soils provide food for life, but are different from place to place. Our ancestors observed that plants grow well at some sites, but not at others. They became aware of differences in soil colors, particle sizes, or moisture conditions at these sites, and tried to identify advantages. With increasing population in different regions of the world, with environmental fluctuations, and in combination with the domestication of wild animals and sedentism the pressure on soils and the need to gain knowledge about soils gradually increased.

Pedology is the study of soils in their natural setting and has at times been equated with the study of soil genesis and soil classification. The people who study soils are pedologists. Pedologists integrate understanding of landscapes, vegetation patterns, climate, and human activity into knowledge about soils, their distribution, mode of formation, and value in different contexts. Pedologists also extrapolate information from one scale to the next, to create a coherent model of soils from microscopic observations to the landscape scale. Integration and extrapolation are necessary because soils form a highly variable continuum over the Earth's surface.

Pedology

The term pedology derives from the ancient Greek terms 'πέδον' which means soil (pedo) and '-λογία' now 'ology' meaning the study of something. The International Union of Soil Scientists (IUSS) interprets pedology as the overall scientific study of the soil. Pedology embraces the scientifically based research of soils that provides information about soil morphology, the processes involved during soil formation, the physical, chemical and biological properties to increase our overall knowledge on soils. The

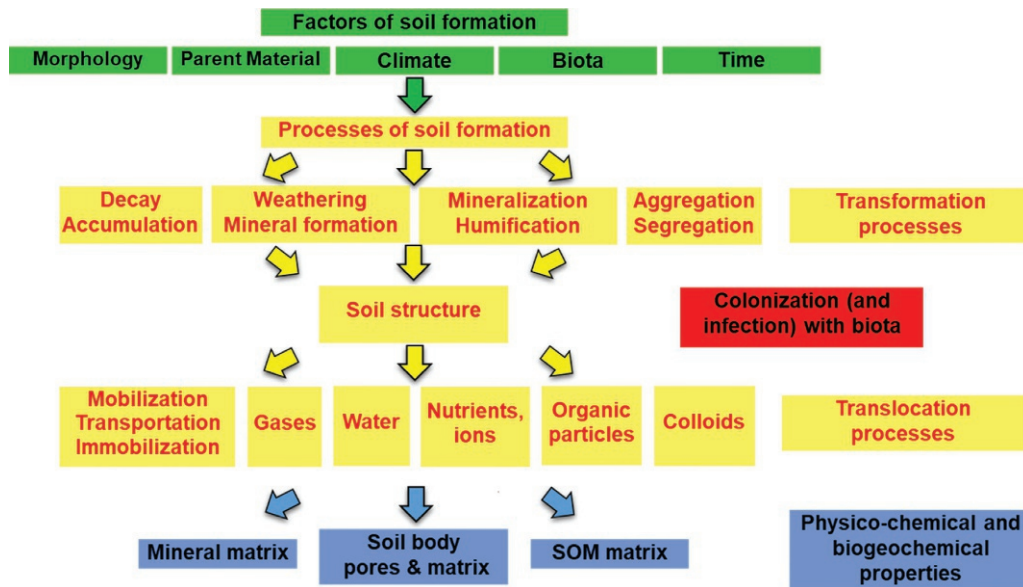


Fig. 1 Illustration of the pathways of soil formation, from the factors to processes and finally to properties of soils.

importance of soil and its study is emphasized by the fact that soils belong to the essential natural resources for all life on our planet that also include water and air.

Pedology includes ‘pedogenesis’, the study of the origin and formation of soil as a natural body, the spatial and vertical distribution of soil in the landscape which includes classification, and interpretation of soil properties for soil use and its management. The topic emerged in the natural sciences in the 19th century. A German scientist Albert Fallou introduced the term in 1862 (Singer, 2005). Early pedology focused on the description and mapping of soil, whereas today it embraces a complex subject covering many sub-disciplines such as soil chemistry, physics, biology, microbiology, ecosystem services, pollution and contamination, and agricultural management and sustainability.

Basic questions on the study of soils are: What are soils? How do they form? What kind of processes are present? What are its properties? What is the role of soils in the environment? What are the threats to the soil resource?

Many attempts have been made to define soils (Hartemink, 2016). Soils are the thin layer on top of the Earth’s surface including semi-terrestrial and sub-hydric soils. Their global mean thickness is about 60 cm. Soils represent the interface of the lithosphere, hydrosphere, atmosphere and biosphere. They serve as the habitat of plants, animals, insects, bacteria, fungi, and humans. Soils are the conversion product of mineral and organic matter under the influence of highly variable environmental factors like climate, biota, landscape morphology and dynamics, parent materials, and time (Fig. 1).

These main factors of soil formation vary from place to place and with time. Any soil profile (i.e., the vertical surface of a soil) and any sample of soil material or soil horizon is therefore unique in their properties. Soils are multi-dimensional objects with overlapping and interacting solid, liquid and gaseous phases, which explain the complexity and heterogeneity of and within soils. The complexity of soils makes their classification, mapping, or modeling of processes challenging (Phillips, 2017; Mikhailova et al., 2021). Classification, for example, is an essential tool of scientific disciplines. It enables soil information to have a structure so that people in many spheres can work with soils, to create soil maps, to assess the land-use potential not only for food and crop production but for many other wildlife and human needs, and to identify threats.

The complexity and polygenetic nature of soils is shown in Fig. 2. The soil profile is an assemblage of fossil soils developed in pyroclastic materials of Holocene volcanic eruptions covering a morainic soil of Late Pleistocene age. Each fossil soil represents a period of soil development before interruption and burial with new material. The soil profile is therefore a good example of the complexity, but also the ‘memory’ of soils that allows reconstruction of the history of volcanic eruptions, landscape evolution, and time-related periods of soil formation (Targulian and Bronnikova, 2019).

Soil formation (pedogenesis)

Soil formation is a dynamic process towards quasi-equilibrium stages of soil development. The input of energy and matter is essential to transform parent materials to soils. Parent materials include solid rocks, mineral and or organic sediments, or (artificial) material of human origin like waste or concrete (Howard, 2021). Considering a block of solid rock like granite, the block is at an equilibrium stage without changes as long as it is surrounded by granite. Weathering starts as soon as the block is exposed at the Earth’s surface to a completely different environment with temperature fluctuations, moisture fluctuations, and exposure to oxygen. The rate at which weathering occurs is influenced by the chemical and mineral properties of the parent material itself. For example,



Fig. 2 Example demonstrating the complexity of soils: Gelic Andosol on top of Late Pleistocene moraine deposits in S-Kamchatka (Photograph by R. Bäumler).

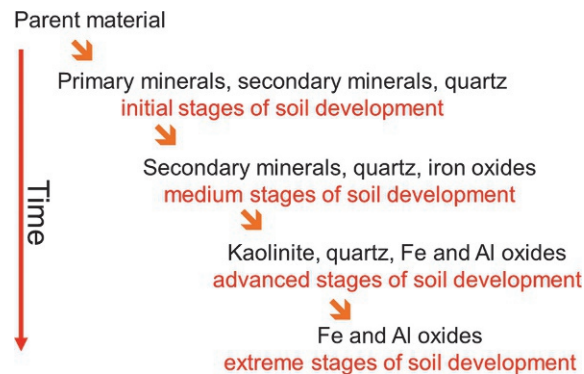


Fig. 3 Schematic diagram of soil development stages under a leaching environment from initial to advanced soil development or weathering (*Fe, iron and Al, aluminium).

porosity, mineral size, crystal structures or specific surface area of parent materials are important factors. Combined, physical and chemical weathering are both highly effective to drive soil formation, with physical weathering causing the disintegration of parent materials such as rocks without chemical changes, and chemical weathering causing the fractional or total dissolution of minerals. Therefore, soil formation might be seen as an equilibration of matter under a dynamic environment at the Earth's surface.

Under a leaching environment (i.e., precipitation exceeds evaporation), for example, pedogenesis proceeds until steady-state conditions from initial stages to advanced or extreme stages of soil development are reached (Fig. 3). The first step in soil development is the disintegration or decomposition of primary minerals that form new minerals, so-called secondary minerals. They form from the dissolution products such as clay minerals and pedogenic oxides, which accumulate with time as long as the soil forming processes are active. Interruptions to these processes might be induced by tectonics, climate change, soil erosion, land-use, man-made or natural coatings with new material including concrete sealing.

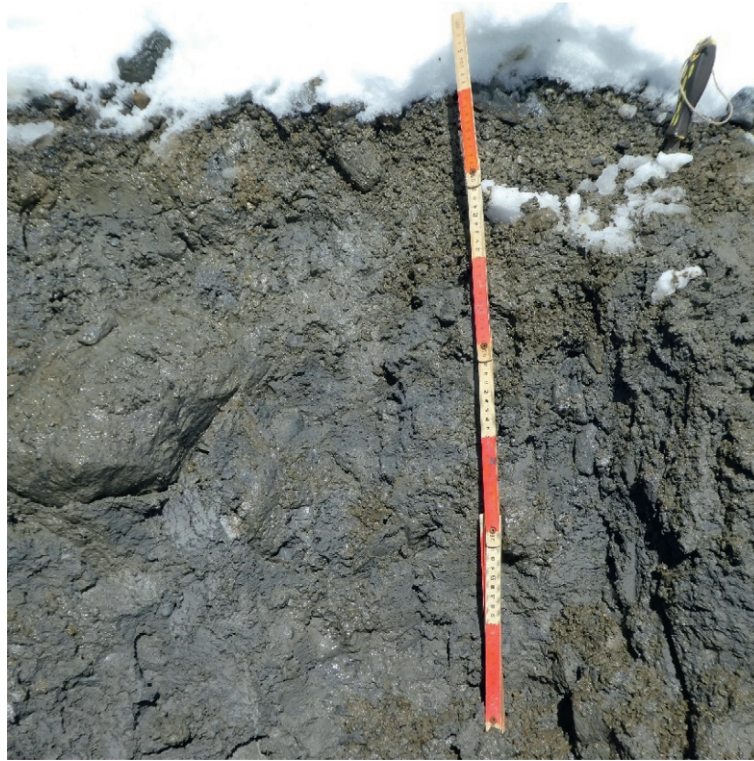


Fig. 4 Initial soil development in moraine deposits of the F urkeleferner glacier, Martell-Valley, South Tyrol, N-Italy (Photograph by R. B aumler, taken 2019).

Fig. 4 shows an example of initial soil development in moraine deposits directly after retreat of the F urkeleferner glacier situated in the Martell Valley in South Tirol, Italy. The ablation moraine parent material shows signs of extreme wetness from the recently melting ice. Hydromorphic features caused by oxidation processes can be seen as initial signs of soil development in the top few centimeters of the soil profile. It is noticeable that the soil is still free of organic matter.

Fig. 5 shows an example of a well-developed soil. The soil profile shown represents an intensively weathered fossil Ferralsol/Oxisol of at least Middle Pleistocene age developed in phyllite, and covered by a Cambisol of 70-cm depth developed from eolian deposits of Late Pleistocene origin. Optically stimulated luminescence (OSL) dating of the eolian deposits at 65 cm of depth categorized these to be of a minimum age of 46 ± 5 ka (GI-32; Dorji, 2016). The near-surface overall enrichment of oxidic iron and aluminium compounds and clay minerals of the kaolinite group that form lateritic bauxite deposits might be representative of the final stages of weathering and soil formation. However, a final soil formation equilibrium stage is possibly questionable, because soils represent biogeochemical reactors with process-response-reactions that can result in reversible and irreversible changes in a dynamic environment (Lin, 2011).

Transformation and translocation processes of matter and energy are the main processes that occur during the development of a soil profile. Transformation processes include, for example, the decay and accumulation of organic matter, the weathering of primary minerals into secondary minerals and the formation of pedogenic minerals from weathering products. Organic compounds are added to the soil profile through litter, and are transformed to soil organic matter mainly by soil biota to gain nutrients and energy. Organic compounds might be completely decayed through processes of mineralization with the release of CO_2 , nutrients and water, or might be partly decayed and accumulated in the soil profile as organic matter by the process of humification.

Interactions of pedogenic minerals, organic compounds, biota exudates and water result in a unique soil structure, a complex three-dimensional mixture of matter with voids in-between, so-called pores, and also result in the formation of unique soil layers known as soil horizons (Hartemink et al., 2020; Matus, 2021). Soil pores can be filled with soil solution (i.e., the liquid phase with water and solutes) or gases (i.e., the gaseous phase). Coarse and middle-sized pores, in turn, are the living environment of soil biota, and are responsible for drainage (coarse pores) or water storage (medium pores). All soil forming processes can proceed consecutively or together at the same time. Changes in any of the regulating factors will immediately induce transformations within soils and the surrounding environment.

Soils are open systems. Matter and energy can be mobilized, transported by leaching, interflow, capillary rise or diffusion. The soil system includes gases, ions (nutrients), water, organic particles, colloids, pollutants. It may lead to a total loss or to a gain by input or immobilization and accumulation of matter within specific parts of the soil such as the formation of eluvial and illuvial horizons during the process of podzolization, argillization, or the accumulation of salts and the formation of soil crusts in dry areas.

Humans (biota) are considered part of the soil forming factors. Human activities have influenced soils and soil formation since the so-called Anthropocene (Amundson, 2021; Howard, 2021), with most soils and their characteristics being changed by human

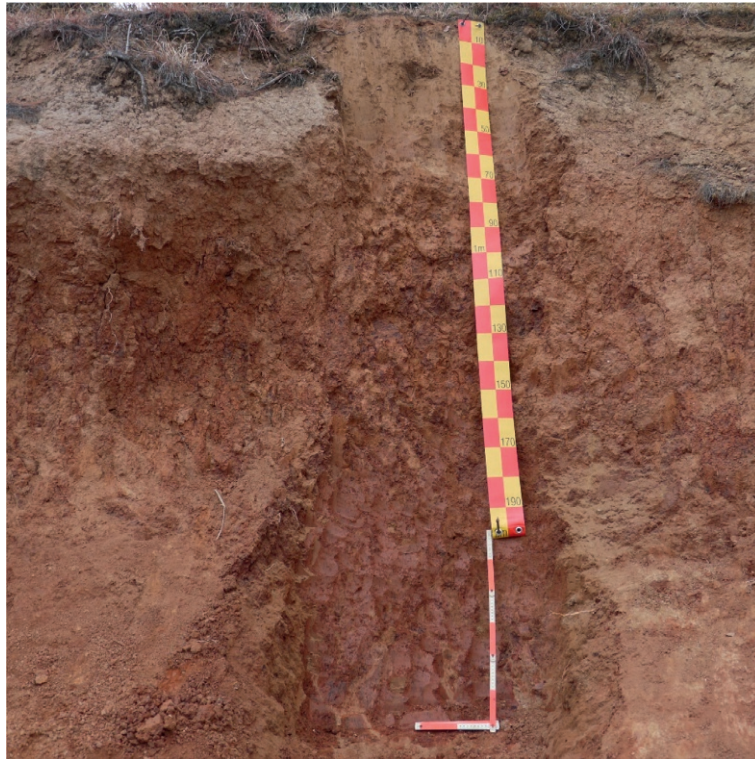


Fig. 5 Fossil soil developed from phyllite and covered by eolian deposits, Paro valley, Bhutan (Photograph by R. Bäumler).

influence. Negative human influences are the result of intensive land use (soil compaction, monocultures, misuse of pesticides, salinization by irrigation mismanagement), the production of greenhouse gases, deforestation and over-grazing inducing desertification and soil degradation, urbanization (sealing of soils), legal and illegal mining activities, waste production and deposition.

Importance of pedological research

Soils are essential for terrestrial biogenic life. More than 90% of all insects, for example, have a geobiotic growth stage. Knowledge about processes and properties of soils are therefore essential with respect to the role of soils in the environment, i.e. energy, water and element (nutrients) cycling, home of life. Pedological research provides information about soil-landscape interactions, soil-plant interactions, soil-human interactions, nutrient supply or deficiency, the behavior of solutes and solids including harmful compounds like neonicotinoids and other pesticides, elements and organic compounds at toxic concentrations. Those toxic compounds might be generated during the natural biodecay of soil organic matter of originally non-toxic compounds. Pedological research provides the theoretical framework for classification, mapping, modeling, land-use potential, food supply, risk assessment, ecosystem management in a changing environment. Information is needed about temporal and spatial aspects with respect to the behavior of hazardous compounds like greenhouse gases. Almost all important parameters and properties can be measured or observed in the field or in the laboratory, or with the help of remote sensing techniques to answer questions about source, storage, decomposition, translocation or sorption in any of the soil compartments.

Concluding remarks

Soils are a complex multi-dimensional natural materials with still many unknown variables. Recent studies of soils go beyond qualitative analysis towards quantitative research and modeling of processes (Minasny et al., 2008) including energy and mass transfer (Amundson, 2021). Organic compounds with complex structures are still partly unknown. Only a fraction of soil microorganism species and their role in the pedosphere is well known. Manifold threats affect soils like extensive land-use, overgrazing, deforestation, salinization, acidification, soil erosion, desertification, contamination, or ground sealing. Increasing pedological knowledge may help to gain greater insight into soils and avoid or reduce such threats.

Soil protection should therefore be one of the most important social and political responsibilities of the 21st century for a sustainable use of soils, a material, which—once destroyed—is not reproducible even within multiple human life spans.

References

- Amundson R (2021) Factors of soil formation in the 21st century. *Geoderma* 391. <https://doi.org/10.1016/j.geoderma.2021.114960>.
- Dorji KD (2016) *Soils as proxies of environmental fluctuations at the southern slopes of the Bhutan Himalayas*. Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, urn:nbn:de:bvb:29-opus4-72774.
- Hartemink AE (2016) The definition of soil since the early 1800s. *Advances in Agronomy* 137: 73–126.
- Hartemink AE, Zhang Y, Bockheim JG, Curi N, Silva SHG, Grauer-Gray J, Lowe DJ, and Krasilnikov P (2020) Soil horizon variation: A review. *Advances in Agronomy* 160(1): 125–185.
- Howard JL (2021) Urban anthropogenic soils—A review. *Advances in Agronomy* 165: 1–57.
- Lin H (2011) Three principles of soil change and pedogenesis in time and space. *Soil Science Society of America Journal* 75: 2049–2070.
- Matus FJ (2021) Fine silt and clay content is the main factor defining maximal C and N accumulations in soils: A meta-analysis. *Nature Scientific Reports* 11: 6438 17 pages. <https://doi.org/10.1038/s41598-021-84821-6>.
- Mikhailova EA, Zurlani HA, Post CJ, Schlautmann MA, and Post GC (2021) Soil diversity (Pedodiversity) and ecosystem services. *Landscape* 10(3): 34 pages. <https://doi.org/10.3390/land10030288>.
- Minasny B, McBratney, Alex B, and Salvador-Blanes S (2008) Quantitative models for pedogenesis—A review. *Geoderma* 144(1–2): 140–157.
- Phillips JD (2017) Soil complexity and pedogenesis. *Soil Science* 182(4): 117–127.
- Singer MJ (2005) Pedology—Basic Principles. *Encyclopedia of Soils in the Environment*, 1st edn, vol. 4, 151–156. <https://doi.org/10.1016/B0-12-348530-4/00001-1>.
- Targulian VO and Bronnikova MA (2019) Soil memory: Theoretical basics of the concept, its current state, and prospects for development. *Eurasian Soil Science* 52(3): 229–243.

Overview chapter on factors of soil formation

Rupert Bäumler^a, Margaret A Oliver^b, and Uta Stockmann^c, ^aInstitute of Geography, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany; ^bDepartment of Geography & Environmental Science School of Human & Environmental Sciences, Soil Research Centre, The University of Reading, Reading, United Kingdom; ^cCommonwealth Scientific and Industrial Research Organisation (CSIRO), Agriculture and Food Business Unit, Canberra, ACT, Australia

© 2023 Elsevier Ltd. All rights reserved.

Introduction	7
Parent material	8
Climate	8
Biota	10
Relief, topography	10
Time	11
The human factor	11
References	13

Abstract

Dokuchaev first introduced formally the concept of factors of soil formation, which act and interact at different spatial and temporal scales. This chapter provides an overview of these factors, which is followed by chapters on most of the specific factors. Climate has a major effect on soil formation because it affects the weathering of rocks and the mineral soil, and vegetation and organisms, which play a major role in changing rocks and sediments to soil. Parent material underpins soil particle-size distribution, chemistry and physics. Relief affects soil hydrology and human use of the soil. Finally, human influence on soil formation has now been recognized as a factor.

Key points

- The four factors originally suggested by Dokuchaev have now become six
- The original factors were climate, parent material, vegetation and biota, and time
- The next development added relief
- Finally, it is now considered that a further factor, humans, should be included to reflect major changes to soil by humans

Introduction

This chapter provides an overview of the chapters in the section on factors of soil formation. Soil, the thin outer skin of the Earth's land, is a critical and fragile natural resource. It is the basis for almost all global agriculture and the medium in which most terrestrial biological activity occurs. Here, we introduce the five forming factors of soil originally suggested more than a century ago (parent material, time, climate, topography, and organisms) and updated over the years to add human activity as the sixth forming factor, namely anthropogenesis (Dror et al., 2022). The major soil forming factors have a profound effect on the types, processes, properties and regional distribution of the soil resource that we observe. The diversity of soils over the Earth's surface results from the site-specific interaction of these factors during soil development.

Dokuchaev (1883), the founding father of soil science as we know it today expressed the relation between the factors of soil formation simply as:

$$s = f(cl, p, v)^t,$$

where s = soil, cl = climate, p = parent material, v = vegetation and biota. The superscript ' t ' indicates that these relations change with time. Nortcliff (2023) notes in his chapter in this section that Hilgard (1914) in the USA and Crowther (1930) in the United Kingdom continued to consider the factors of soil development based on environmental interactions. In addition, Shaw (1930) from the USA also considered a model of soil formation, but did not develop his model further before he died suddenly in 1939. Jenny (1941) then developed these ideas further with the concept that the state of the soil system depends on the interaction of independent variables or soil forming factors. These are represented in his now famous general 'clorpt' formula:

$$S = f(cl, o, r, p, t, \dots),$$

where 'S' is a property of the soil, (f) is a function of: 'cl' is climate, 'o' now represents vegetation and living organisms (macro- and micro-biota), 'r' is relief or topography (this will affect the hydrological regime), 'p' is parent material (material on which the factors have acted to form soil) and 't' is time (how old the soil is). With the increase in human activity and humanity's effects on soil development, a factor related to mankind in the era of the Anthropocene has now been considered to be added to the factors of soil formation.

The factors and their interactions define the state of the soil system. The soil is an open system with incoming and outgoing fluxes of energy and matter in all directions. These include incoming and outgoing radiation, gases moving by mass flow, water in liquid and solid forms, solids dispersed in air or dissolved or dispersed in water, and organisms that migrate through the system.

In addition to the inherent complexity of each of the soil formation factors, they act and interact at different spatial and temporal scales. For example, there are broad climate belts that operate at the scale of tens to hundreds of kilometers, but there are also microclimates that act over tens of centimeters. These differences in scale further transmit to vegetation and organisms, and might also impinge on the time factor. It is notable that different types of soil may form from the same parent material depending, in particular, on the climate and vegetation. Soil development is a slow process with changes taking place over long time periods, which has implications for the damage caused by human activities. Once destroyed, soils are considered a non-renewable resource within the timeframe of generations.

Practical problems related to the complexity of the interactions between factors and scale issues include difficulty in the development of soil classifications that can communicate information at different spatial scales, climates, vegetation and countries. There are many worldwide classification systems and they are often limited by scale issues. In the early Russian and North American soil classifications, the different scales of influence of the soil forming factors were identified in the broad structure of the classification systems (Nortcliff, 2023). Increasingly, the interrelationships between the soil forming factors and their influence on the nature of soils have been recognized as more complex and variable with more subtle relationships at scales from hundreds to tens of kilometers, and at even smaller scales as understanding progressed.

Parent material

Within the suite of soil factors, parent material forms the basis from which soils develop and is considered the initial state of the soil system. Thus, parent material is one of the key factors that underpins soil formation (Nortcliff, 2023). Most parent materials are mineral matter, although there are distinctive soils that developed from organic materials. The first stages of soil development involve the weathering of minerals, and in young soils parent material is a major influence on their nature that is independent of other factors. In most soils this influence reduces through time. The parent materials may be in situ geological materials or eroded and redeposited materials. Parent materials might therefore be solid rock, unconsolidated unweathered or pre-weathered materials or sediments, with or without carbonates. Mineral structure (crystallinity), grain sizes (surface area) and chemical composition in turn influence resistance to weathering or rate of soil development, and almost all physical and chemical soil properties such as the content and release of nutrients. Fig. 1 shows a Leptosol (WRB)/Entisol (Soil Taxonomy) from diabase rock under a temperate climate in Middle Europe. The photograph clearly indicates the weathering front along the rock cracks.

One of the most important groups of agricultural soils, Mollisols (Soil Survey Staff, 2022) or Chernozems (World Reference Base for soil resources, WRB, 2022), have formed largely on redeposited wind-blown material known as loess (Sprafke, 2023). Loess has formed largely at the margins of the world's major ice sheets of the Quaternary period and covers about 10% of the Earth's surface, now mainly in the temperate zone. Silicate silt predominates over carbonate, together with clay and some fine sand, making loess a favorable soil parent material.

Climate

Climate (Dondyene et al., 2023) is an important factor; it determines the rates of weathering, chemical reactions, transport of matter and biological activity in soils. Therefore, it is not surprising that the earliest soil scientists categorized and mapped soils according to climatic zones or conditions (Hartemink and Bockheim, 2013). This led to the concept of 'zonal soils,' i.e., soils that are in equilibrium with climate and vegetation in different latitudinal zones, which was introduced by Dokuchaev in the 1880s. Solar radiation, precipitation, temperature, and wind are the key climatic components in relation to the formation, properties and hydrology of soils. Solar radiation determines the net energy influx and varies according to latitude and season, and determines the temperature. Temperature and temperature fluctuations influence physical weathering (frequency of frost changes; Fig. 2) and the kinetics of geochemical and biochemical reactions (catalytic accelerator). Wind affects the fluxes through soil and can cause soil erosion (abrasion), transport, and deposition of fine-grained particles. Evapotranspiration potential depends on temperature and wind; it includes the water transpired by plants, water that evaporates from water bodies, surfaces and the soil, and sublimation from snow or ice. The net water flux in soil, i.e., the balance between infiltrating water and evapotranspiration, determines the movement of soluble elements and ultimately the soil's water content and the nutrition of plants. Understanding the soils' potential for agriculture, environmental functioning, and climate change mitigation and adaptation requires insight into their relation to climate.



Fig. 1 Leptosol (WRB)/Entisol (US Soil Taxonomy) from diabas weathering, Harz, Germany. Photograph by R. Bäumler.



Fig. 2 Frost pattern Leptosol (Entisol) induced by intensive frost changes, Snæfellsjökull volcano, Central Iceland. Photograph by R. Bäumler.



Fig. 3 Kastanozem developed from Loess near Bordea (Eastern Romania). The subsoil is characterized by precipitation of secondary carbonates and intensive bioturbation (krotowinas). Photograph by R. Bäumler.

Biota

In terrestrial environments, soils are the primary interface between geology and biology (Werts, 2023), and without the biological component the surface deposits cannot be classified as soil in a pedo-ecological sense. Biology has various influences on soil properties with depth resulting in a notable transition of biological activity from the top- to sub-soil along horizon boundaries (Fig. 3). Biological activity leads to the incorporation of organic matter in the topsoil and the amount gradually decreases with depth, in general. Much of the organic matter in soil derives from the input of above- and below-ground vegetation, the latter in the form of root systems. The initial breakdown of material can begin with microorganisms such as fungi, bacteria and lichens, and by vertebrate species such as birds, reptiles or mammals and invertebrate detritivores, herbivores, and carnivores such as earthworms, millipedes, termites, insects, snails and so on. In addition, organic matter derives from root exudates and necromass from microorganisms. The abundance and influence of each of these groups varies with climate in terrestrial environments and provides the foundation for the changes in ecosystems that we can observe in the major ecozones. Consequently, biological influences also change with temperature, latitude, water availability, and human influence.

The huge number of different organic compounds in the soil dwarfs the variety of minerals present and can often be very difficult to identify (Werts, 2023). The soil biology also has a major effect on the plant nutrients present, soil structure and hydrology. It is probably the most complex of the soil factors because of the many components, their interactions with soil minerals, water, roots, gases and themselves. There has been a huge increase in interest in soil biology with the relatively new omics techniques and more feasible methods to determine the DNA of soil biota. These techniques should enable the identification of new antibiotics that are still unknown from the potentially large reservoir provided by soil microorganisms and their exudates.

Relief, topography

The influence of topography on soil formation from macro to microscales is not described in detail in a separate chapter despite its significance especially in hilly and mountainous areas. Finke et al. (2023; in this Encyclopedia) consider landscape evolution models as a means of understanding the relations between relief and soil formation. Relief or topography plays an important role in soil formation through steepness of the slope, degree of departure from the horizontal and elevation, and aspect, direction the slope faces relative to the sun. Relief affects both vertical and horizontal soil processes and their rates of action. Slope affects the degree of movement of soil across the surface, fluxes through the soil and soil development. Where the slope is gentle or almost level, the soil shows the greatest relative development. The steeper is the slope, the greater is the movement of material along it by gravity. Heavy rain or permafrost conditions may cause erosion or solifluction of the soil at hilltop and slope positions downhill with accumulation at the foot of the slope or valley bottom. These mobilization processes depend strongly on the material properties of the soil particles and are additionally enhanced by increasing slope angles, a lack of vegetation, and insolation of the slopes. Slope and elevation also affect soil drainage and depth to the water table, which in turn affect soil formation. This is reflected in the nature and extent of horizon development. A series of soils that develops along a hillslope on the same parent material and climate is described as a Catena. Along a Catena, deeper well-drained soils occur in summit positions, thinner well-drained soils are present on the slope and deeper sometimes poorly-drained soils can be found in the footslope positions. Aspect affects insolation so that in the northern

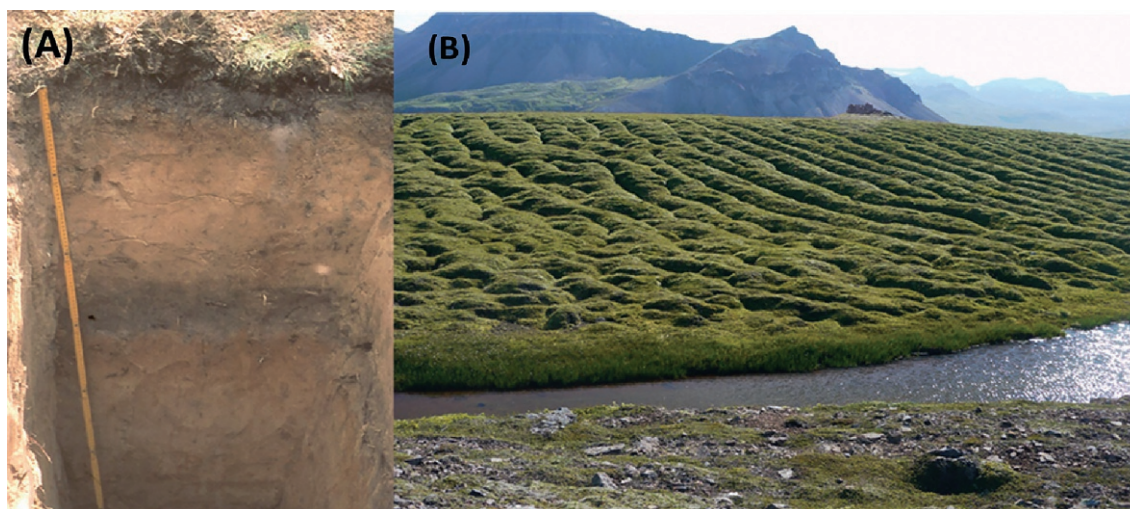


Fig. 4 Colluvisol from Late Pleistocene deposits near Berlin, Germany (A), and (B) solifluction after thufur (icelens) formation under the influence of permafrost at low slope angles, northeastern Iceland. Photographs by R. Bäumler.

hemisphere the north side of slopes tends to be cooler and wetter with slower rates of soil chemical and biological processes, and more vegetation growth, whereas south facing slopes tend to be warmer and drier with faster rates of soil processes. The reverse is the case in the southern hemisphere.

Fig. 4 gives two examples of soils influenced by topography. Fig. 4A is a so-called Colluvisol, a fossil soil successively covered by eroded soil from upslope during the Holocene, and Fig. 4B shows a soil displaced by solifluction under permafrost conditions in northeastern Iceland at low slope inclination.

Time

The environmental conditions under which soils and their properties develop, are governed by combinations of the above factors, which operate over time (Melzer and Yonker, 2023). Climate, biota, topography, and parent material have a direct effect on soil-forming processes, whereas time has an indirect effect. Time affects soil-forming processes and the degree of soil development by controlling their duration. Thus, we can differentiate between ‘young,’ ‘intermediate’ and ‘old’ soil in each ecosystem based on degrees of chemical and physical weathering and the resulting expression of soil properties (Fig. 5). The amounts and distribution of soil organic carbon (SOC), carbonate, and other key elements, and the amount and distribution of clay and its mineralogy change as a function of time in a generally consistent and predictable manner (Koop et al., 2020).

The amounts and distribution of the soil constituents above are often used to determine the relative age of a soil with sufficient accuracy despite the heterogeneous soil matter (Dorji et al., 2009). Older soils in general have more fine particles and clay minerals, larger contents of pedogenic oxides and a greater accumulation of SOC than younger soils. The accumulation of SOC eventually reaches a steady state where inputs equal outputs, and there is no further net accumulation with increasing soil age (time) (Lawrence et al., 2015; Wang and Amundson, 1996; Merritts et al., 1991). The length of time needed for SOC to reach a steady state varies among environments, but is within a relatively short time frame of 10^2 – 10^3 years (Buol et al., 1997).

The accumulation of calcium carbonate (CaCO_3) in soils over time is limited in general to soils of semiarid and arid regions, where carbonates precipitate in dry soil environments, with high pH, and high evapotranspiration (see also Fig. 3). Unlike SOC, CaCO_3 accumulation may not reach a steady state, but continues to increase through time (Wang et al., 2016; Monger and Rachal, 2013).

The most abundant elements in soils, excluding oxygen and carbon, are silicon, aluminum, iron, magnesium, calcium, sodium, and potassium, and their relative abundance is a direct function of the degree of soil weathering (Dixon and Weed, 1989), which increases or decreases through time and soil development (young, intermediate, old).

The human factor

An important question now arises because of the human influence on soils that has been ongoing for many centuries. The question is “Are humans part of the ‘Biota’ factor or should they be considered an independent additional ‘anthropogenesis’ factor of soil formation as suggested by Dror et al. (2022), especially if we acknowledge the Anthropocene as the present era” (Sandor et al., 2023).



Fig. 5 Fossil soil covered by a Cambisol of about 65 cm depth developed from terminal moraine deposits of the Djol Djilga glacier. The soil formation was stopped 8000 years BP, the fossil A horizons was dated by ^{14}C . Photograph by R. Bäumler.

Soil formation can be strongly influenced by any kind of anthropogenic land-use activities or land-use changes that interact with or can override the natural factors of soil formation. This influence can extend to a total interruption of soil formation, for example, by soil sealing. Human activities that influence the state of the soil include, for example, mining, urbanization, agriculture and horticulture, forestry or deforestation, road construction, military training areas, terracing and many other activities. These result in the input of matter (industrial and organic fertilizers, lime, pesticides, technogenic materials, or irrigation water), mixing of the natural sequence of soil horizons (Fig. 6), the loss and degree of re-deposition of soil by wind or water erosion, the deposition of any kind of material on top of the natural soil (construction debris, disposed waste, landfill), a human induced loss or decay of soil



Fig. 6 A deep ploughed site (trenching) in eastern Romania for amelioration, afterwards covered with spoil debris. Photograph by R. Bäumler.

organic matter, or the neoformation of soils exclusively by human activities (Anthrosols, Terra preta; Glaser and Woods, 2004). Another issue to consider under the human factor is the contamination of soils with compounds mainly generated by chemical industries (pesticides, microplastics, herbicides), either localized or over larger areas, or more diffuse. These inputs might be intentional or accidental, and are often toxic or with unknown toxicity, which negatively affects processes and properties of soils and soil biodiversity.

References

- Buol SW, Hole FD, McCracken RJ, and Southard RJ (1997) *Soil Genesis and Classification*, 4th edn. Ames, IA: Iowa State University Press. 527 pp. ISBN 0-8138-7464-2.
- Crowther EM (1930) The relationship of climate and geological factors to the composition of soil clay and the distribution of soil types. *Proceedings of the Royal Society B* 107: 1–30.
- Dixon JB and Weed SD (1989) *Minerals in the Soil Environment*, 2nd edn. Madison, WI: Soil Science Society of America.
- Dokuchaev VV (1883) *Russian Chernozem. Selected Works of V.V. Dokuchaiev*. vol. 1. Israel Program of Scientific Translations, Jerusalem. (translated 1967).
- Dondeyne S, Mantel S, and Dekkers S (2023) Factors of soil formation: Climate. In: *Encyclopedia of Soils in the Environment*, 2nd edn. Elsevier.
- Dror I, Yaron B, and Berkowitz B (2022) The human impact on all soil-forming factors during the Anthropocene. *ACS Environmental Au* 2(1): 11–19.
- Finke P, Minasny B, Vanwaghem T, Salvador S, and Stockmann U (2023) Soil formation models. In: *Encyclopedia of Soils in the Environment*, 2nd edn. Elsevier.
- Glaser B and Woods WI (2004) *Amazonian Dark Earths: Explorations in Space and Time*. Berlin: Springer-Verlag.
- Hartemink A and Bockheim J (2013) Soil genesis and classification. *Catena* 104: 251–256.
- Hilgard EW (1914) *Soils*. New York, NY: The Macmillan Company.
- IUSS Working Group WRB (2022) *World Reference Base for Soil Resources. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*, 4th edn.
- Jenny H (1941) *Factors of Soil Formation: A System of Quantitative Pedology*. New York: McGraw-Hill. 281 p.
- Koop AN, Hirmas DR, Sullivan PL, and Mohammed AK (2020) A generalizable index of soil development. *Geoderma* 360. <https://doi.org/10.1016/j.geoderma.2019.113898>.
- Lawrence CR, Harden JW, Xu X, Schulz MS, and Trumbore SE (2015) Long-term controls on soil organic carbon with depth and time: A case study from the Cowlitz River Chronosequence, WA USA. *Geoderma* 247–248: 73–87.
- Melzer SE and Yonker CM (2023) Soil forming factors: Time. In: *Encyclopedia of Soils in the Environment*, 2nd edn. Elsevier.
- Merritts DJ, Chadwick OA, and Hendricks DM (1991) Rates and processes of soil evolution on uplifted marine terraces, northern California. *Geoderma* 51: 241–275.
- Monger HC and Rachal DM (2013) Soil and landscape memory of climate change-how sensitive, how connected? *SEPM Special Publication* 104: 63–69.
- Nortcliff S (2023) Soil forming factors: Parent material. In: *Encyclopedia of Soils in the Environment*, 2nd edn. Elsevier.
- Sandor JA, Burras CL, Thompson M, and Wills SA (2023) Factors of soil formation: Human impacts. In: *Encyclopedia of Soils in the Environment*, 2nd edn. Elsevier.
- Shaw CF (1930) Potent factors in soil formation. *Ecology* 11(XI): 239–245.
- Soil Survey Staff (2022) *Keys to Soil Taxonomy*, 13th edn USDA-Natural Resources Conservation Service.
- Sprafke T (2023) Parent materials: Loess. In: *Encyclopedia of Soils in the Environment*, 2nd edn. Elsevier.
- Dorji T, Caspari T, Bäumler R, Veldkamp A, Jongmans A, Tshering K, Dorji T, and Baillie I (2009) Soil development on late quaternary river terraces in a montane valley in eastern Bhutan. *Catena* 78(1): 48–59.
- Wang Y and Amundson R (1996) Radiocarbon dating of soil organic matter. *Quaternary Research* 45: 282–288.
- Wang J, Monger C, Wang X, and Leinauer MSB (2016) Carbon sequestration in response to grassland- shrubland- turfgrass conversions and a test for carbonate biomineralization in desert soils, New Mexico, USA. *Soil Science Society of America Journal* 80: 1591–1603.
- Werts S (2023) Factors of soil formation: Biota. In: *Encyclopedia of Soils in the Environment*, 2nd edn. Elsevier.

Factors of soil formation: Climate

Stefaan Dondeyne^a, Stephan Mantel^b, and Seppe Deckers^c, ^aDepartment of Geography, Ghent University, Ghent, Belgium; ^bISRIC – World Soil Information, Wageningen, The Netherlands; ^cDepartment of Earth and Environmental Science, University of Leuven, Leuven, Belgium

© 2023 Elsevier Ltd. All rights reserved.

Introduction	14
Soils in humid and sub-humid tropical climates	16
Soils in semi-arid and arid climates	17
Soils in warm temperate climates	18
Soils in cold temperate climates	20
Soils in sub-arctic and arctic zones	21
Soils in changing climates	22
Conclusions	24
References	24

Abstract

Climate determines the rates of chemical reaction, substance transport and biological activity in soils, therefore, the earliest soil scientists categorized soils according to climatic zones. Understanding the soils' potential for agriculture, environmental functioning, and climate change mitigation and adaptation requires insight into their relation to climate. The Soil Orders (Soil Taxonomy) and the Reference Soil Groups (World Reference Base for soil resources) that typically occur in each climatic zone are discussed and underlying processes of soil formation are explained. Finally, the effects of global climate change on soils are discussed, as well as evidence in soils of past climatic changes.

Key points

- Climate is a key factor in the formation of soils
- Major soils and soil-forming processes are presented for each climatic zone
- Understanding the relationship between climate and soils is key to assessing land-use potential, including their potential for carbon sequestration, climate change mitigation and adaptation
- Soils are recorders of past climate changes

Introduction

Climate is a key factor in the formation of soils and it also determines the potential for land uses. It controls the rates of chemical reaction and substance transport in soils, and directly affects the flora and soil fauna. The climate influences human land uses by regulating the duration of growing seasons and the kind of crops that can be planted. Furthermore, soils also serve as climatological archives, as their properties not only reflect current but also past climatic conditions. These examples demonstrate the complex relationship between soils and climate. While the human world population is expected to increase further from the 7.7 billion reached in 2019, to over 10.9 billion by 2100 (United Nations and Department of Economic and Social Affairs, 2019), global climate change is posing unprecedented challenges to human societies (IPCC, 2022). In this context, insight into the relationship between climate and soils is important, to understand the soils' role in agricultural and environmental functioning and, to be able to assess their potential for climate change mitigation and adaptation.

Solar radiation, precipitation, temperature, wind, and evapotranspiration are the key climatic components in relation to soils. Solar radiation determines the net energy influx and varies according to latitude and season, and determines the temperature. Temperature and temperature fluctuations influence physical weathering and the kinetics of geochemical and biochemical reactions. Wind contributes to processes of soil erosion and deposition and, together with temperature, also determines the evapotranspiration potential. Evapotranspiration is the water transpired by plants combined with water that evaporates from water bodies, surfaces and the soil, as well as the sublimation from snow or ice. The net water flux of water in the soil, the balance between infiltrating water and evapotranspiration, determines the movement of soluble elements and ultimately the soil's water content.

Given the strong influence of the climate on soil properties, it is not surprising that the first attempts to classify and map soils were referring to climatic conditions (Hartemink and Bockheim, 2013; Sibirtsev et al., 1898) (Table 1). The concept of 'zonal soils,'

Table 1 Soil zones of Europe following Dokuchaev's concepts and with corresponding Soil Orders of Soil Taxonomy (Soil Survey Staff, 2014) and Reference Soil Groups of WRB (IUSS Working Group WRB, 2015).

Soil zone	Soil order	Reference soil group
Polar-Tundra soils	Gelisols	Cryosols
Sod-Podzolic soils of taiga forest	Spodosols, Histosols, Glossudalfs	Podzols, Histosols, Retisols
Grey forest soils	Alfisols, Glossudalfs, Mollisols	Luvisols, Retisols, Phaeozems
Chernozems and steppe soils	Mollisols	Chernozems, Kastanozems
Desert soils	Aridisols	Solonchaks, Solonetz, Gypsisols, Calcisols
Red-earth laterite soils	Ultisols	Chromic Luvisols, Lixisols, Acrisols, Nitisols

Soil zones adapted from Sibirtsev NM, Ferkhmin AR, and Tanfiliev GI (1898) *Schematic Soil Map of European Russia* [in Russian].

i.e., soils that are in equilibrium with climate and vegetation in different latitudinal zones, was introduced in the 1880s by V.V. Dokuchaev (Rusakova et al., 2022). It was also realized that at the fringes of each zone, soils with intermediate characteristics occurred which were referred to as 'intrazonal' soils. Furthermore, where soil formation was limited, soils occur that are not particularly typical for any zone, e.g. soils of recent river or wind deposits were called 'azonal' soils.

Between the Tropics of Cancer and Capricorn the sun's rays reach Earth at a steep angle, up to an angle of 90° at noon twice per year. The intertropical, low latitude, zone benefits from a radiation surplus, while the polar, high latitude, zones experience a deficit. As a result, seasonal thermal variations are low in the intertropical zone, while these are high in polar and mid-latitude zones. These variations in temperature, in conjunction with air and oceanic currents, lead to a zonal pattern of climate and vegetation varying from warm and humid around the equator where rainforests occur, to warm and arid around the Tropics of Cancer and Capricorn where deserts occur, and cold in the sub-arctic and arctic zones with tundra vegetation. In the European part of Russia, Belarus and Ukraine, the fairly uniform gradient of sediments, ranging from sandy in the north to silty in the south, coincides with a temperature and precipitation gradient. In the late 19th Century, V.V. Dokuchaev observed that this climatic gradient is reflected in a sequence of vegetation zones, to which particular soils are associated (Table 1). On his map of the soils of the northern hemisphere, Dokuchaev indicated five zones: boreal, taiga, Chernozem, aerial and lateritic zones (Rusakova et al., 2022), with which he applied the principal of zonation for mapping and classifying soils at world level in direct relation to climate.

The vegetation and soils in North America, however, do not occur in such latitudinal zones. Although the parent material is also quite homogenous over fairly large areas, the precipitation decreases from the east coast to the interior, i.e. perpendicular to the north-south temperature gradient. As Chernozems, Chestnut brown soils and Desert soils all occur at the same latitude as Brown forest soils and Podzols, it is difficult to apply the concept of zonal soils in North America. During soil surveys in the USA in the 1930s soil groupings were based on soil properties and broad environmental conditions. Only the highest category (Category VI, Solum Composition Groups), the distinction between the Pedocal (soils with accumulated lime) and Pedalfer groups (lacking secondary lime) reflects a relation with climate: where precipitation is lower than evapotranspiration, secondary lime will accumulate versus where precipitation is higher than evapotranspiration, lime will dissolve and leaches.

In Soil Taxonomy, the American classification system introduced in the 1970s, soils are classified based on directly observable and measurable properties. At the second level of classification, climatic characteristics are taken into account because it is based on soil moisture and soil temperature regimes. These soil climatic criteria correlate with the global climatic zones. For example the xeric soil moisture regime (i.e., total soil moisture control section is, in normal years, dry for 45 or more consecutive days in the 4 months following the summer solstice and moist for 45 or more days in the 4 months following the winter solstice) is typical for Mediterranean climates (moist and cool winters, dry and warm summers). Similarly, the (iso)hyperthermic soil temperature regime (i.e., mean annual soil temperature at 50-cm depth is 22 °C or higher; 'iso-' indicating a temperature fluctuation of less than 6 °C which is typical for tropical climates).

The relation between climate and soils is further presented following climatic zones based on the aridity/humidity index (AHI) and combined with four temperature regimes (Fig. 1). We discuss the soils that typically occur in each climatic zone following the 12th edition of *Soil Taxonomy* (ST) (Soil Survey Staff, 2014) and the third edition of the *World Reference Base for Soil Resources* (WRB) (IUSS Working Group WRB, 2015). The AHI, as proposed by the United Nations Environmental Programme (UNEP, 1992), is calculated as the ratio between the average annual precipitation (P) and the reference (or potential) evapotranspiration (ET₀). Six climatic zones can be recognized using the aridity/humidity index, ranging from hyper-humid (AHI > 2), predominantly around the equator, to hyper-arid (AHI < 0.05) mostly around the tropics. Agro-ecologically it is more relevant to define the intertropical zone as the area where average monthly temperature corrected to sea level, is always more than 18 °C (isomesic to isohyperthermic in ST), rather than just considering the latitude (FAO, 1978). Similarly the Arctic and subarctic can be defined as the area where monthly temperatures are above 10 °C for at least one and at most 3 months of the year (subgelic to hypergelic, in ST) (<https://arcticportal.org/maps/download/arctic-definitions>). The mid-latitudinal zones can be further divided into a cold temperate zone, where during summer the soil temperature does not exceed 8 °C (frigid and cryic, in ST) and a warm temperate zone where soil temperature does not exceed 22 °C (mesic to (hyper)thermic, in ST).

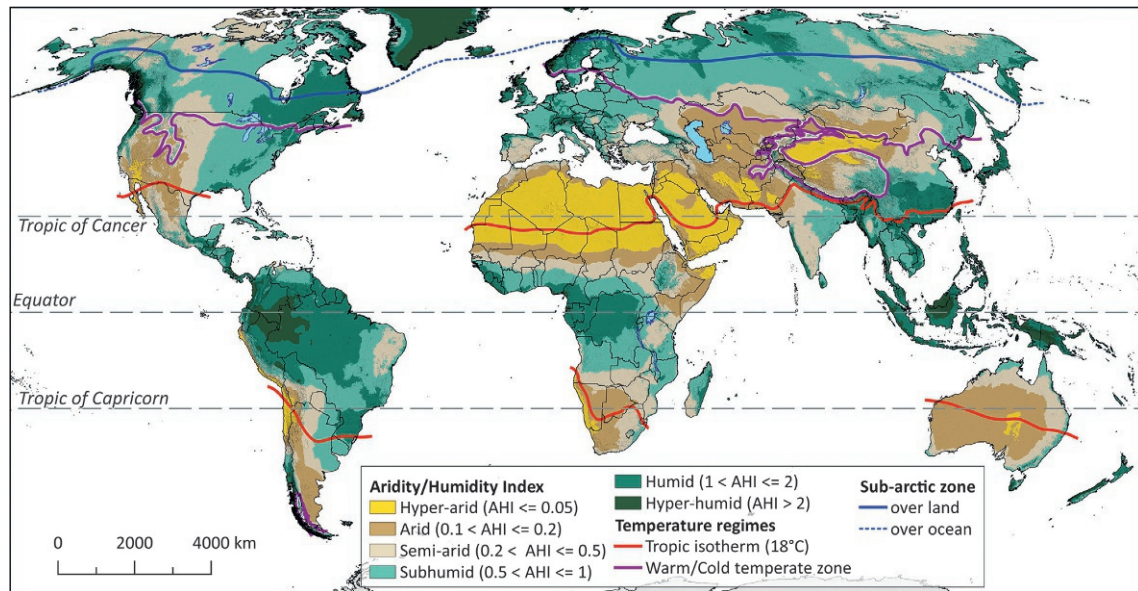


Fig. 1 Climatic zones of the world based on the UNEP aridity/humidity index. Note the north–south zonation in Eurasia, Asia and Africa vs the east–west zonation in the Americas. The tropical isotherm shows the area where average monthly temperature is more than 18°C, when corrected to sea level. In the subarctic, temperatures are above 10°C for at least one and at most three months of the year. The temperature zone can be divided into a cold zone (soil temperature is <8°C) and a warm zone. Authors' map based on Global Potential Evapotranspiration and Global Aridity Index - <https://cgicarsi.community/>.

Soils in humid and sub-humid tropical climates

Tropical climates have an average temperature, corrected to sea level, above 18 °C in every month (isomesic, iso(hyper)thermic, in ST). In most of the areas around the equator, annual rainfall is high and exceeds the annual evaporation (Fig. 1).

High precipitation combined with high temperature in the hyper-humid and humid tropical climates, results in intense, deep chemical weathering and strong leaching of solutes in the soil. Silicates are weathered by water and CO₂ (carbon dioxide) that produces dissolved silica (H₄SiO₄) and bicarbonate, which are removed with percolating water. This process is called desilication (Van Breemen and Buurman, 2002). Primary minerals such as feldspars and micas are dissolved or transformed into kaolinite (Al₂Si₂O₅ · 2H₂O), goethite (α-FeOOH), and gibbsite (α-Al(OH)₃). In some cases lepidocrocite (γ-FeOOH) may be formed. This process of ferralitization is widespread in the warm hyper-humid, and humid tropical regions particularly on old geomorphic surfaces. It gives rise to acidic soils with low cation exchange capacity (CEC < 16 cmol_c kg⁻¹ clay) due to the prevalence of low-activity clay, and soils with variable charge occur where iron and or aluminum hydroxides are dominant. Such soils are known as Oxisols (ST) or Ferralsols (WRB) (Fig. 2A), and kandic Great Groups of the Ultisols (ST), which in part fall under the Acrisols (WRB). On relatively younger deposits, and younger, more dynamic geomorphic surfaces, chemical weathering of primary minerals produces large amounts of exchangeable aluminum, resulting in Paleudults (ST) or Alisols (WRB). In sandy deposits, strong leaching can lead to the formation of giant Spodosols (ST) or Podzols (WRB) (Fig. 2B).

In hyper-humid and humid climates, often in combination with oligotrophic conditions, undecomposed or incompletely decomposed organic matter accumulates in wet conditions forming peats, or Histosols (ST and WRB). Histosols typically have layers containing more than 20% soil organic carbon that together are at least 40 cm thick in the first meter below ground surface. While the majority of Histosols are located in cool climates, they also occur in tropical swamp forests and papyrus wetlands. The Cuvette Centrale depression in the Congo Basin is one of the world's most extensive swamp forests with Histosols (Dargie et al., 2017). Other major areas with Histosols are in Kalimantan and Sumatra in Indonesia and in the Amazon basin (Miettinen et al., 2016; Xu et al., 2018).

In subhumid tropical climates, annual precipitation is equal to, or less than, the reference evapotranspiration (ET₀), but still greater than half the ET₀ (0.5 < AHI ≤ 1). In this zone a pronounced dry season occurs around the winter solstice (ustic soil moisture regime in ST). Weathering and leaching of minerals are less intense compared with the (hyper)humid climates. Less rainfall leads to some upward capillary movement of solutes during the dry period when evapotranspiration exceeds precipitation. Moreover, the soils are often enriched with airborne particles as desert dust can be transported over long distances, such as by the Harmattan winds in West Africa, which frequently obscure the sunlight during daytime. Processes of desiccation play an important role: when soils dry out, completely hydrated forms of iron and aluminum hydroxides may lose at least part of their crystal water. Goethite (α-FeOOH) is transformed into hematite (α-Fe₂O₃), following the reaction

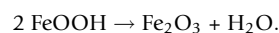




Fig. 2 Soils of humid tropical climates: (A) a Xanthic Haplustox (ST) or Xanthic Ferralsol (WRB) on the Makonde plateau, Tanzania. The yellowish color is due to goethite. (B) a giant Spodosol (ST) or Podzol (WRB) in Malaysia. The white sand in the upper part has been leached under which iron-humus complexes have accumulated. If the iron-humus enriched horizon occurs at more than 2-m depth, the soil will be a Spodic Quartzipsamment (ST) or Albic Arenosol (Bathypodic) in the WRB. Photographs: (A) Stefaan Dondeyne, (B) ISRIC.

The dehydration of goethite and the formation of hematite, are responsible for the striking difference in colors between soils formed in humid tropical climates and those of the subhumid zones. In humid tropical climates soils are dominantly yellow or yellowish-brown (the color of goethite) (Fig. 2A), whereas they are predominantly red when they form in subhumid climates due to the presence of hematite (Fig. 3A).

Because of the less intense weathering and leaching, and the addition of dust, soils of the subhumid tropical zone are less acidic and have higher base saturation than their counterparts under humid climates. Typical soils are Kandiuustalfs (ST) or Lixisols (Fig. 3B) and Nitisols (*pro parte*) (WRB), together with Paleustalfs and Haplustalfs (ST) or Luvisols (WRB). Oxisols (ST) or Ferralsols (WRB) often occur here too, they probably formed under past more humid climates. In flat areas, where heavy rainfall alternates with pronounced dry seasons, seasonal stagnation of surface water occurs over large areas, leading to the formation of Albaqualfs (ST) or Planosols (WRB) (Fig. 3B). Where soils have developed from basic rocks (such as basalts, or limestone) and where there is a pronounced dry season they result in the formation of shrink-swell clays, (montmorillonite) Vertisols (ST and WRB).

Soils in semi-arid and arid climates

In semi-arid and arid climates annual precipitation is equal to, or lower than, $ET_0/2$ ($AHI \leq 0.5$). When the soil is moist, water will dissolve soluble minerals. When evapotranspiration exceeds precipitation, water will move upwards and the soluble minerals will precipitate. This process leads to the enrichment of the soil with secondary calcium carbonate ($CaCO_3$), gypsum ($CaSO_4 \cdot 2H_2O$), and in the most dry climates soluble salts such as mirabilite ($Na_2SO_4 \cdot 10H_2O$), thenardite (Na_2SO_4), hexahydrite ($MgSO_4 \cdot 6H_2O$), nahcolite ($NaHCO_3$), soda ($Na_2CO_3 \cdot 10H_2O$), trona ($Na_3CO_3HCO_3 \cdot 2H_2O$), and halite ($NaCl$). The accumulation of soluble salts is known as 'salinization.'

Soils of sand dunes, which are conspicuous landforms of semi-arid and arid climates, are classified as Torripsamments (ST) or Arenosols (WRB). Other soils do not show much development either, apart from the accumulation of carbonates, gypsum, or salts. Following Soil Taxonomy these soils are Aridisols in particular Calcids (ST) or Calcisols (WRB), Gypsisols (ST) or Gypsisols, and Durids (ST) or Durisols (WRB). Other common soils include Natrustalfs (ST) and Solonetz (WRB), both of which have a high sodium content in the subsoil. Where salts have accumulated at the soil surface Salids (ST) or Solonchaks (WRB) are found. Where



Fig. 3 Soils of subhumid climates: (A) Kandiusalf (ST) or Lixic Nitisol (WRB) from Tasmania, Australia. The reddish color indicates the dominance of hematite, the formation of which requires dry seasons. (B) An Albaqualf (ST) or Solodic Planosol (WRB) from Victoria, Australia. The whitish color in the upper 40 cm illustrates the seasonal stagnation of water. Photographs: Stefaan Dondeyne.

cold ocean currents meet desert shores, such as the Benguela current along the west coast of southern Africa, mist drifts inland, carrying moisture and solutes from the sea. This gives rise to the deposition of sulfates. In this area, up to 150 km inland, the vast areas of accumulated gypsum can be attributed to the influence of mist.

Because of the scarcity, or total lack of vegetation, wind plays a large role in these climates and has a considerable impact on the soils. Wind contributes to the speed of evaporation, but also removes fine particles (as dust storms), and also leads to the formation of dunes. Where fine particles are removed, a gravelly surface is often left behind, known as a desert pavement (Fig. 4A). The gravel stones at the surface are shaped by scouring sand particles creating 'ventifacts' and can develop a shiny surface, the 'desert varnish' (Fig. 4B).

Soils in warm temperate climates

At mid-latitudes, the seasonal variation of the sun's position results in summer and winter seasons. In maritime regions, such as in western Europe, the influence of warm oceanic currents contributes to the occurrence of mild winters. In continental locations, such as in Ukraine, southern Russia or the Midwest of the USA, climate is much drier leading toward semi-arid. Summers are hot, but winters are also much colder than in the maritime regions (mesic to thermic in ST). In most of the temperate zone, precipitation is evenly distributed over the year but dry summers alternate with wet winters around the Mediterranean Sea. This climate corresponds to the xeric soil moisture regime (ST) that also occurs, e.g., in California (USA), South Africa and southern Australia.

Under warm humid temperate climates excess precipitation gives rise to chemical weathering and leaching. With the large fluctuation in temperature and varying rainfall patterns, rates of chemical weathering are also variable. However, the net transport of substances, either in solution or as colloids, from the upper parts of the soils to the lower parts or to the groundwater is the common denominator for most of the soils in warm temperate climates.

Chemical weathering of non-calcareous parent materials leads to the release of free iron hydroxides (amorphous iron hydroxide, $\text{Fe}(\text{OH})_3$; goethite, $\alpha\text{-FeOOH}$) and the formation of clay, particularly illite. Clay formation mainly results from the weathering of muscovite, accompanied by the partial or total loss of K^+ ions. Chemical weathering of calcareous parent materials starts with the dissolution and removal of carbonates. It is only after decalcification that other processes, such as 'brunification' start. This process involves the release of free iron hydroxides that give the soils in warm temperate climates their characteristic yellowish-brown color. In warm temperate climates with distinct dry summers, dehydration of the hydroxides leads to more red colors (rubefaction) due to the formation of hematite ($\alpha\text{-Fe}_2\text{O}_3$) (Fig. 5A).



Fig. 4 Soil features of arid climate: (A) A Natric Argicryid (ST) or a Salic Solonetz (WRB) with a desert pavement in the Gobi desert (Mongolia). The white streaks are accumulations of lime and salts. (B) A stone that has been shaped by wind abrasion (ventifact). The shiny surface is the so called desert varnish (Gobi desert, Mongolia). Photographs: Stefaan Dondeyne.

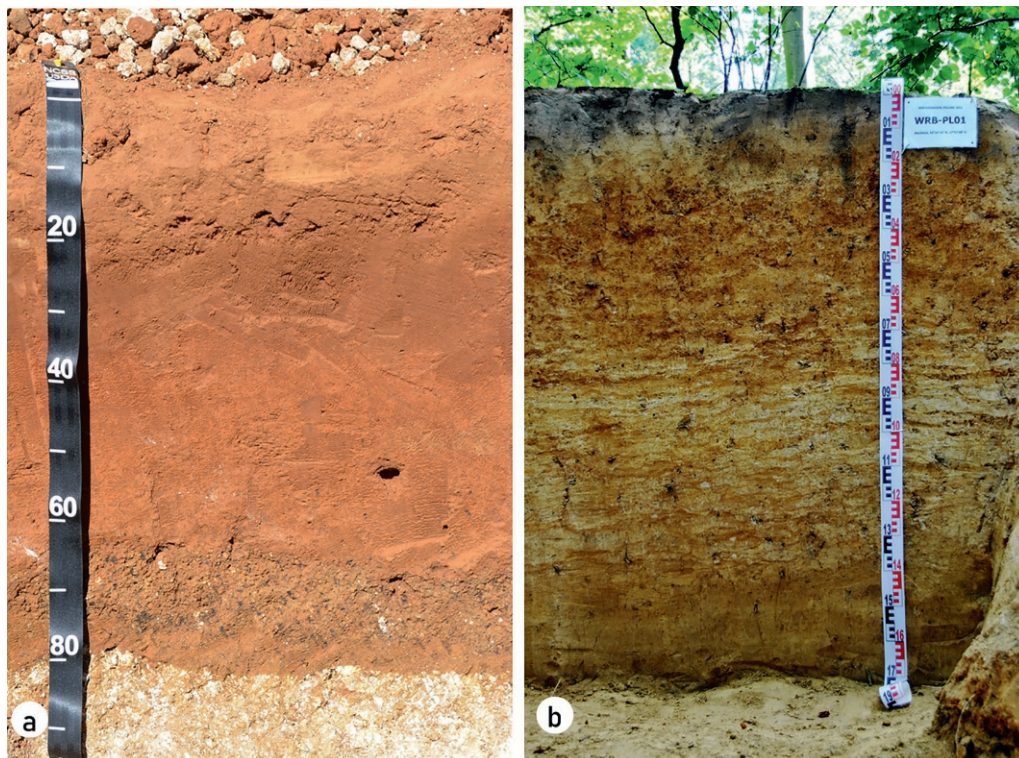


Fig. 5 Soils of temperate warm climates: (A) A Calcic Haplustalf (ST) or Luvic Calcisol (WRB). Under the brownish plough layer (0–30 cm), a reddish clay enrichment horizon occurs (30–65 cm). Manganese concretions together with part of the parent material (Andesite) is still found as a gravel layer (65–80 cm), below which a hardpan of calcium carbonate is formed (South Africa). (B) Lamellic Hapludalf (ST), or an Albic Luvisol (WRB). Directly under the humus rich surface horizon (0–10 cm) first there is a bleached eluvial horizon (10–25 cm), followed by a brownish clay enrichment horizon (25–70 cm). Below this horizon, more clay eluviation and illuviation occur in bands (70–125 cm) (Poland). Photographs: Stefaan Dondeyne.

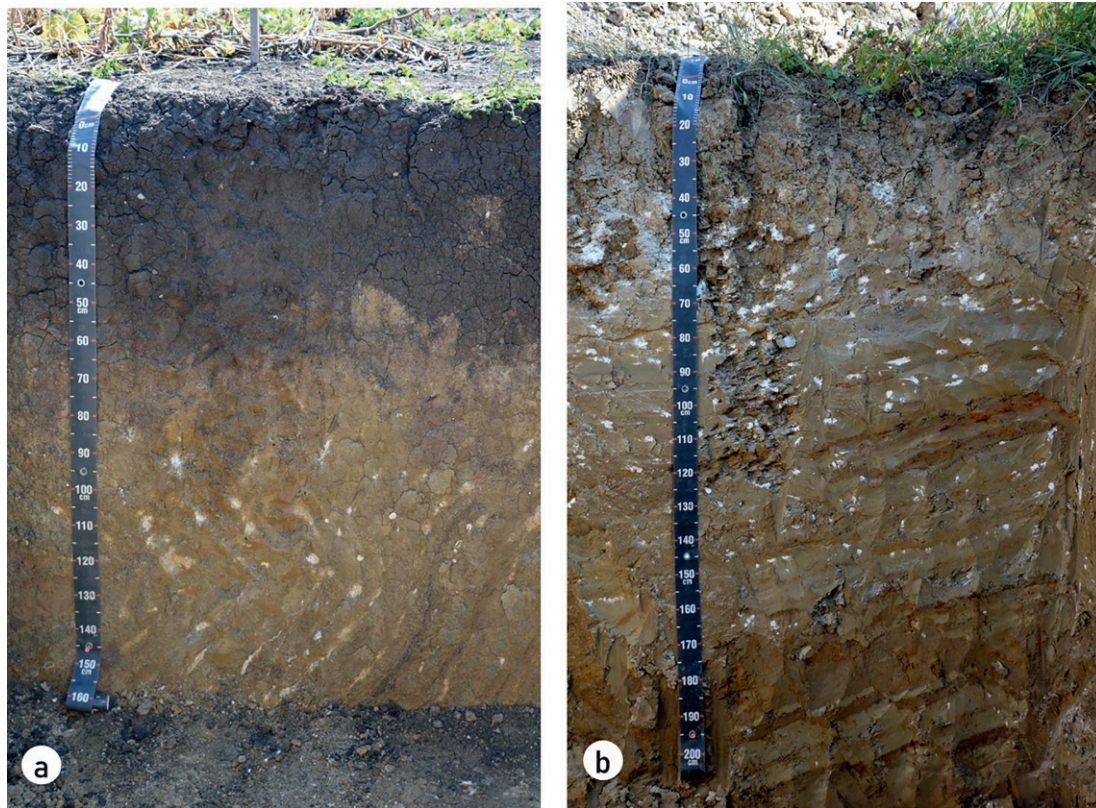


Fig. 6 Soils of temperate cool climates: (A) A Pachic Calciustoll (ST) or Calcic Chernozem (Poland). Due to the cool climate organic matter decomposes only slowly and accumulates in the surface horizon. (B) Typical Calciustept (ST) or Cambic Calcisol (WRB) (Romania). Soil development is limited to the development of a cambic horizon (40–60 cm). Due to the dry climate calcium carbonate precipitates along cracks and soil structural elements which appear as white concretions. Photographs: Stefaan Dondeyne.

The physical translocation of clay particles by gravity water from the upper soil horizons to the lower soil horizons is called clay illuviation. It results in clay-enriched, subsurface horizons, known as argillic (ST) or argic (WRB) horizons (Fig. 5B). The process is enhanced by a soil environment with relatively small content of exchangeable Ca^{2+} and Al^{3+} ions, both of which tend to flocculate the clay particles, thus inhibiting the dispersion necessary for clay translocation. Furthermore, it is favored by desiccation of the soil through which larger pores and cracks are created, and by the occurrence of heavy rain showers, creating the low-electrolyte environment needed for clay translocation.

In warm humid temperate climates Alfisols (ST) or Luvisols (WRB) are typically the dominant soils (Fig. 5). In drier climates, soils with accumulations of calcium carbonate occur (Fig. 6). Toward the boundary with cool and drier temperate climates, organic matter builds up in the surface horizons where Mollisols (ST) or Chernozems (WRB) are encountered.

Soils in cold temperate climates

The mean soil temperature of cold temperate climates (mostly, frigid and cryic in ST) is above freezing point, but is less than 8°C , therefore, no permafrost occurs here. However, at higher elevation as, for example, on the Tibetan plateau, there are areas where permafrost occurs. Large parts of this climatic zone have been formerly glaciated. This has resulted in the deposition of glacial and periglacial deposits such as boulder clays, outwash materials, fluvial deposits of braided river systems, lake sediments, loess, and cover sands. The heterogeneity of these sediments, together with the conditions of cool temperate climates, results in a variety of soil-forming processes.

Of importance in the cold temperate climates are the long, cold winters, during which the soils are frozen and covered by snow. In areas with slowly permeable parent materials (boulder clays, lake sediments, etc.) meltwater stagnates during springtime, causing waterlogging and temporary reducing conditions. This gives rise to dissolution and removal of iron, decline in soil structure in the surface horizons, and acidification. On better-drained parent materials (cover sands, braided river deposits, etc.), water stagnation may still be present due to the frozen subsoil, but not for as long and not as severe as on poorly drained parent materials. Under coniferous forest, acidic soils develop. On sandy deposits, the process of cheluviation takes place, whereby organic acids, released from the vegetation debris, complexed with aluminum and iron are subsequently leached downwards and deposited in the lower

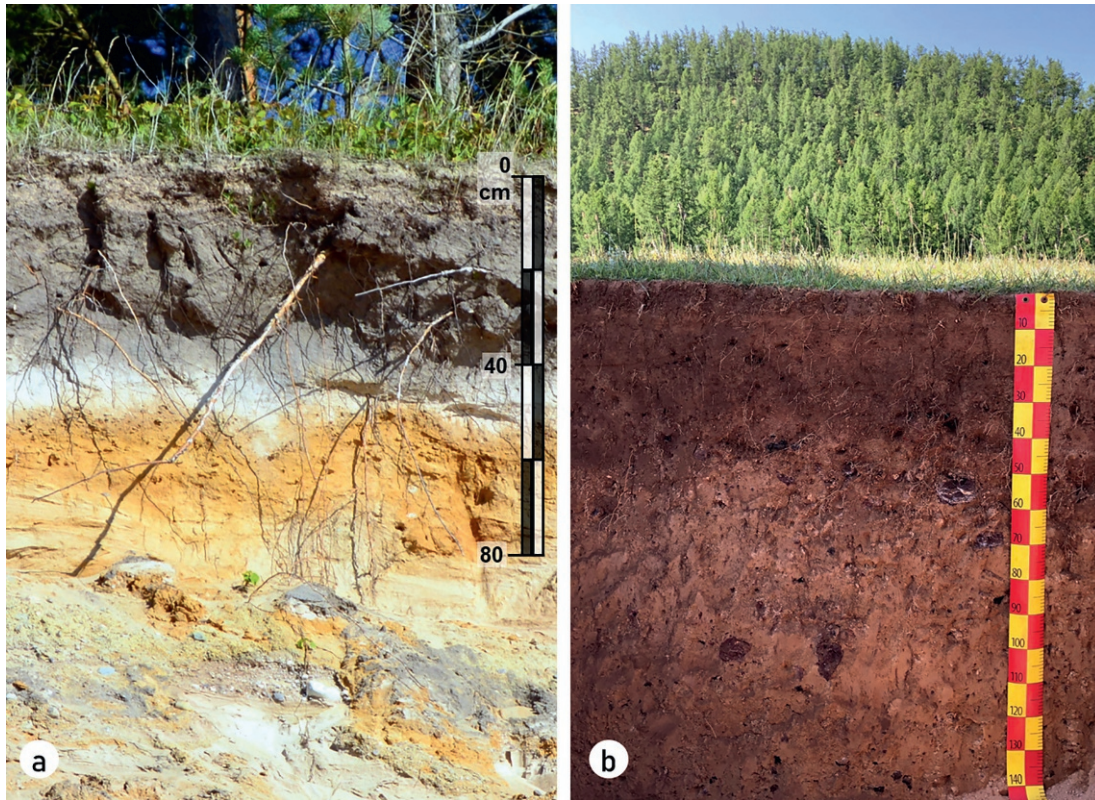


Fig. 7 Soils of the cold temperate climates: (A) Typical Haplocryods (ST) or an Albic Rustic Podzol from Latvia. Acidification of the soil, due to the humid climate and enhanced by coniferous forest, activates the leaching of complexes of iron and aluminium resulting in a white, bleached horizon (30–50 cm) and subsequent deposition to form a reddish horizon (50–80 cm) (Latvia); (B) Typical Haplocryoll (ST) or an Haplic Kastanozem from Mongolia. The upper 40 cm is rich in organic matter, and the dominant brown color is typical of soils of the drier climate under short steppe vegetation. Photographs: (A) Stefaan Dondeyne; (B) Stephan Mantel.

parts of the profile. This results in the development of Spodosols (ST) or Podzols (WRB) (Fig. 7A). The low temperature also hampers biological activity in these soils, consequently, humification and bioturbation are slow.

During the warm summers soil life becomes active and contributes to the process of homogenization of soil material. Where summers are sufficiently long and warm, and winters not very severe, biological activity in the soils contributes greatly to the processes of homogenization and humification. This is especially so under grasslands with a large amount of organic matter. Burrowing soil animals homogenize the soils to a considerable depth, sometimes up to 2 m, and thoroughly mix the organic matter with the mineral soil constituents, which is called bioturbation.

Soils in sub-arctic and arctic zones

In the sub-arctic or boreal zone, the average temperature is above 10 °C for one to at most 3 months; in the Arctic zone, it is even colder. Maritime regions may be hyper-humid to humid ($AHI > 1$), for example, Iceland and northern Norway, whereas, continental regions may be semi-arid to arid ($AHI \leq 0.5$) as in central Yakutia in Russia. Weathering under the low temperatures in the subarctic and Arctic zones is almost exclusively physical; chemical weathering is limited and, if any, it is restricted to the short periods of thawing. In addition, biological activity is low due to the low temperatures.

The effects of freezing and thawing are characteristic for the soils in sub-arctic and Arctic climates and, in many cases, there is permafrost (permanently frozen subsoil with ice segregation). The annual freezing and thawing of the soils result in thermal contraction and expansion of water and minerals in rocks, which slowly weaken the rock structure and permits water to enter in cracks. The freezing of water further widens the cracks, as ice has a lower bulk density (hence more volume per weight) than water (0.9 versus 1.0 kg dm^{-3}), and over time the rock will disintegrate into angular fragments. The annual freezing and thawing results in the process of ‘cryoturbation,’ giving rise to features such as horizon involution (Fig. 8), frost heave, cryogenic sorting, and patterned grounds (e.g. stone polygons).

In permafrost areas, soils are dominantly Gelisols (ST) or Cryosols (WRB). These soils have, by definition, permafrost occurring within 100-cm depth, or within 200-cm depth in association with gelic materials above it. Alternatively, they have one or more cryic horizons within 100 cm of the soil’s surface.

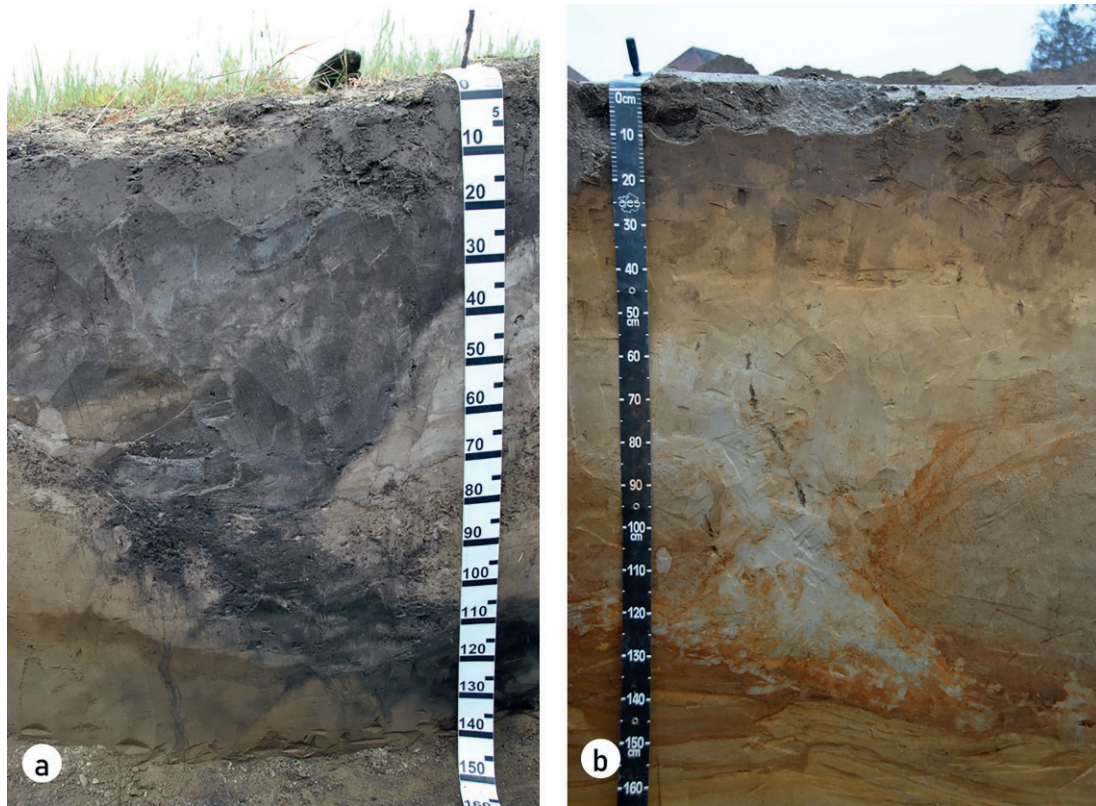


Fig. 8 Soils of sub-arctic climates: (A) Active cryoturbation pulls organic rich material from the surface downwards in a wedge shape pattern when thermokarst mounds are formed (central Yakutia, Russia); (B) a similar wedge shape pattern in a sandy soil in northern Belgium, where it is obviously a relict feature. Photographs: Stefaan Dondeyne.

The low rate of humification of organic matter and the wet conditions during the thawing periods result in the accumulation of peat at the surface. Peat soils are among the most widespread soils in the (sub)arctic regions. Approximately half of the Earth's peat soils, i.e. Histosols (ST and WRB), occur in the circumpolar region of the northern hemisphere.

Soils in changing climates

The ever-increasing net emission of greenhouse gases is resulting in an unprecedented global warming of the climate (IPCC, 2021). A comprehensive assessment of the potential impact of climate change on soils, and of all the feedback loops that can be expected between climate and soil processes, is beyond the scope of this contribution. However, based on the insights into the relations between climate and soil formation some critical points can be highlighted. When soils under permafrost are included, the soil organic carbon content is estimated to be 3200–4100 Gt C, which is approximately seven times as much as that stored in the above-ground biomass (450–650 Gt) and about four times that in the atmosphere (860 Gt C) (Friedlingstein et al., 2020).

Soils have an enormous potential for carbon sequestration where organic matter can accumulate, as in soils under permafrost (Gelisol ST, or Cryosols WRB), in peatlands (Histosols), and in soils typical of grasslands (Mollisols in ST, or Chernozems and Kastanozems in WRB). Recent research has also highlighted historical losses from the soil organic carbon pool, and pointed to the added risk of future accelerated losses under a warming climate (Bossio et al., 2020). The Arctic permafrost, which stores roughly 1700 Gt C, is of great concern. Rapid thawing and thermokarst processes might release significant amounts of carbon into the atmosphere in a few days or years. While carbon dioxide emissions in the Arctic are proportionally higher than other greenhouse gases, methane emissions are expected to rise as anoxic conditions in thawing permafrost and soils increase (Miner et al., 2022). Inhabitants of sub-arctic and Arctic regions are already confronted with the consequences of global warming because the warming is especially intensive in permafrost areas. In central Yakutia, the thawing of arable land leads to the formation of new thermokarst mounds (Fig. 8A). In the process, Gelisols (ST) or Cryosols (WRB) are transformed in soils with seasonally stagnating water (Aquic Haplogelols, in ST, or Stagnic Cambisols, WRB) or become salt affected soils (Natric Argiorthels in ST, Calcic Solonetz in WRB). These rapid changes in the soil render much of the former arable land unusable for agriculture (Desyatkin et al., 2021).

The increase in temperature will also lead to greater evapotranspiration and higher rates of both geochemical and biochemical processes. For example, soil N (nitrogen) cycling will shift from microbial immobilization to enhanced mineralization, nitrification and denitrification across global terrestrial ecosystems (Dai et al., 2020). Therefore, whereas soil carbon sequestration is recognized as a way to reduce atmospheric carbon dioxide, climate warming may also hasten loss of the soil's carbon stock. It has been demonstrated that soil carbon in particulate organic matter and in mineral-associated organic matter has different vulnerabilities to climate change. In Europe, soils under arable land and coniferous forest contain the largest but also the most vulnerable carbon stocks (Lugato et al., 2021). Given the complex interactions between all these processes, it remains important to appreciate the different processes of soil formation. Research in California, for example, showed that while forests are frequently thought to be important for carbon sequestration, they may be in peril where extreme heatwaves, drought, and wildfires become more common. In such conditions, grasslands – that typically occur on Mollisols (ST), or Chernozems and Kastanozems (WRB) – are more resilient to climate change and serve as important belowground carbon sinks (Dass et al., 2018).

Soil forming processes such as cryoturbation, rubefaction and ferralitization only occur under very specific climatic conditions. Other processes such as brunification, decalcification, clay illuviation, and cheluviation occur under a wider range of climatic conditions. Some soil properties may persist after changes in the climate, while others, such high salt content or secondary calcium carbonates, can quickly be removed when climate becomes more humid. Persistent soil properties can be good indicators of past climatic changes. The relict wedge shape pattern in the sandy profile of northern Belgium (Fig. 8B), illustrates that at the end of the last glaciation similar processes of thermokarst must have occurred which are now active in sub-arctic regions (Fig. 8A) as described by Desyatkin et al. (2021).

Frost action in periglacial conditions also result in the formation of ice plates and wedges following polygonal structures. Water stagnation in the layers above the permafrost leads to formation of a bleached horizon. This white (albic) material will seep into the cracks left by ice plates and wedges after thawing. Such a reticular pattern of white tongues (albeluvic glossae), is the most plausible process leading to the formation of Glossudalfs (ST) or Retisols (WRB) (Dondeyne and Deckers, 2019; van Vliet and Langohr, 1981). The presence of these soils in current warm temperate climates is a testimony of the cold climatic conditions that prevailed during the last glaciation (Fig. 9A).

Rubefaction, the process of red coloring as a result of the formation of hematite, is associated with subhumid climates that have a long, hot dry season. The occurrence of such soils in warm temperate climates (Fig. 9B) is attributed to subhumid warm climates during the Mesozoic Era.



Fig. 9 Soils in changing climates: (A) Glossaqualf (ST) or Retic Stagnosol (WRB) in Belgium. Note the white tongues (albeluvic glossae) that extend from the white eluviation horizon (40–60 cm) and cut across the brownish clay-enrichment horizon. These white tongues follow a polygonal pattern and are remnants of periglacial processes of the last glaciation. (B) The strong reddish colors of this Hapludult / Haplic Alisol under a coniferous forest in Luxembourg shows that the parent material has been weathered in a tropical subhumid climate. Photographs: Stefaan Dondeyne.

Conclusions

Soils bear a strong stamp of both present and past climates, consequently early soil classification systems integrated climatic characteristics explicitly. Though modern international classification systems are based on observable or measurable soil properties as these classification systems rely on concepts of soil formation, major classification units inevitably correspond to variations in climate. Temperature is a key climatic factor because it not only regulates the speed of weathering: higher temperatures lead to faster chemical processes, but also more biological activity which plays a crucial role in the breakdown or the accumulation of organic matter. Furthermore, temperature, together with wind, are also key to the rate of evapotranspiration. The ratio between precipitation and evapotranspiration determines whether soil water fluxes are predominantly downwards or upwards. Downward fluxes result in weathering and eluviation of soluble minerals and colloids. Where strong evapotranspiration occurs, soluble minerals as carbonates, gypsum and salts are redistributed in the soil and accumulate at or near the surface. Raising temperature and changing precipitation patterns will strongly affect all of these soil processes.

References

- Bossio DA, Cook-Patton SC, Ellis PW, Fargione J, Sanderman J, Smith P, Wood S, Zomer RJ, von Unger M, Emmer IM, and Griscom BW (2020) The role of soil carbon in natural climate solutions. *Nature Sustainability* 3: 391–398. <https://doi.org/10.1038/s41893-020-0491-z>.
- Dai Z, Yu M, Chen H, Zhao H, Huang Y, Su W, Xia F, Chang SX, Brookes PC, Dahlgren RA, and Xu J (2020) Elevated temperature shifts soil N cycling from microbial immobilization to enhanced mineralization, nitrification and denitrification across global terrestrial ecosystems. *Global Change Biology* 26: 5267–5276. <https://doi.org/10.1111/gcb.15211>.
- Dargie GC, Lewis SL, Lawson IT, Mitchard ETA, Page SE, Bocko YE, and Ifo SA (2017) Age, extent and carbon storage of the central Congo Basin peatland complex. *Nature* 542: 86–90. <https://doi.org/10.1038/nature21048>.
- Dass P, Houlton BZ, Wang Y, and Warlind D (2018) Grasslands may be more reliable carbon sinks than forests in California. *Environmental Research Letters* 13, 074027. <https://doi.org/10.1088/1748-9326/aacb39>.
- Desyatkin R, Filippov N, Desyatkin A, Konyushkov D, and Goryachkin S (2021) Degradation of arable soils in central Yakutia: Negative consequences of global warming for Yedoma landscapes. *Frontiers in Earth Science* 9, 683730. <https://doi.org/10.3389/feart.2021.683730>.
- Dondeyne S and Deckers JA (2019) The ABC soil types: Podzoluvisols, albeluvisols or retisols? A review. In: Deák J, Ampe C, and Mikkelsen JH (eds.) *Soils as Recorders of Past and Present. From Soil Surveys to Archaeological Sites: Research Strategies for Interpreting Soil Characteristics*, pp. 55–64. Bruges, Belgium: Raakvlak.
- FAO (1978) Report on the agro-ecological zones project. Vol 1: *Methodology and Results for Africa*. vol. 1. Rome: FAO. World Soil Resources Report 48.
- Friedlingstein P, O'Sullivan M, Jones MW, Andrew RM, Hauck J, Olsen A, Peters GP, Peters W, Pongratz J, Sitch S, Le Quéré C, Canadell JG, Ciais P, Jackson RB, Alin S, Aragão LEOC, Arnett A, Arora V, Pirani A, Conners SL, Péan C, Berger S, ... Zhou B (eds.) *Summary for Policymakers*. Cambridge University Press.
- IPCC (2021) *Summary for Policymakers*. In: Pörtner H-O, Roberts DC, Polaczanska ES, Mintenbeck K, Tignor A, Alegría A, ... Okem A (eds.) *Climate Change 2022: Impacts, Adaptation, and Vulnerability, Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. p. v + 35.
- IUSS Working Group WRB (2015) *World Reference Base for soil resources 2014*. update 2015. International soil classification system for naming soils and creating legends for soil maps. Update 2015 Rome: FAO.
- Lugato E, Lavalée JM, Haddix ML, Panagos P, and Cotrufo MF (2021) Different climate sensitivity of particulate and mineral-associated soil organic matter. *Nature Geoscience* 14: 295–300. <https://doi.org/10.1038/s41561-021-00744-x>.
- Miettinen J, Shi C, and Liew SC (2016) Land cover distribution in the peatlands of Peninsular Malaysia, Sumatra and Borneo in 2015 with changes since 1990. *Global Ecology and Conservation* 6: 67–78. <https://doi.org/10.1016/j.gecco.2016.02.004>.
- Miner KR, Turetsky MR, Malina E, Bartsch A, Tamminen J, McGuire AD, Fix A, Sweeney C, Elder CD, and Miller CE (2022) Permafrost carbon emissions in a changing Arctic. *Nature Reviews Earth & Environment* 3: 55–67. <https://doi.org/10.1038/s43017-021-00230-3>.
- Rusakova E, Sukhacheva E, and Hartemink AE (2022) Vasily Dokuchaev – A biographical sketch on the occasion of his 175th birthday. *Geoderma* 412: 115718. <https://doi.org/10.1016/j.geoderma.2022.115718>.
- Sibirtsev NM, Ferkhmin AR, and Tanfilyev GI (1898) *Schematic Soil Map of European Russia*. [in Russian].
- Soil Survey Staff (2014) *Keys to Soil Taxonomy*, 12th edn USDA: Natural Resources Conservation Service.
- UNEP (1992) *Status of desertification and implementation of the United Nations Plan of action to combat desertification*. (Report of the Executive Director No. UNEP/GC/DEC/SS.III/1) Nairobi, Kenya: UNEP.
- United Nations, Department of Economic and Social Affairs (2019) *World Population Prospects Highlights*. 2019 revision Highlights, 2019 revision.
- Van Breemen N and Buurman P (2002) *Soil Formation*. Dordrecht, Boston: Kluwer Academic.
- van Vliet B and Langohr R (1981) Correlation between fragipans and permafrost with special reference to silty Weichselian deposits in Belgium and northern France. *Catena* 8: 137–154. [https://doi.org/10.1016/0341-8162\(81\)90002-3](https://doi.org/10.1016/0341-8162(81)90002-3).
- Xu J, Morris PJ, Liu J, and Holden J (2018) PEATMAP: Refining estimates of global peatland distribution based on a meta-analysis. *Catena* 160: 134–140. <https://doi.org/10.1016/j.catena.2017.09.010>.

Factors of soil formation: Soil biota and organic matter as a factor for soil formation and weathering processes

Scott Werts, Department of Chemistry, Physics, and Geology, Winthrop University, Rock Hill, SC, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	25
Components of soil organic matter	26
Soil ecosystems	26
Tropical soil systems	28
Savanna and grassland soil systems	29
Desert soil systems	29
Subtropical and temperate systems	30
Deciduous forest systems	30
Coniferous forest systems	30
Boreal forest systems	30
Tundra systems	32
Conclusion	32
References	32

Abstract

In terrestrial environments, soils are the primary boundary between geology and biology. Not only do biological influences in soils change with depth, they change with temperature, latitude, water availability, and human influence. The huge number of different organic compounds in the soil dwarfs the variety of minerals present and can be very difficult to identify. When considering the influence of biology on soil development on a large scale, we tend to think in trends and abundances of the various organic influences rather than specific intricacies of particular ecosystems. By separating soil systems by biological zones, we can understand better all of the variation in biological influences and development in soils across different latitudes and landscapes.

Key points

- Soil biology has a large influence on soil development
- Biological influences vary with depth in all soils
- Soil biological communities vary with variations in other soil forming factors
- Variations in soil biology affect rates of soil decomposition, nutrient cycling, and carbon storage

Nomenclature

% Carbon	Carbon content (grams of carbon per 100 g of soil)
% Clay	Clay content (total number of particles per 100 particles that are below 2 μm in size)
% Nitrogen	Nitrogen content (grams of nitrogen per 100 g of soil)
C:N	Carbon to nitrogen ratio (unitless value)
Epipedon	Diagnostic surface horizons of soils, often based on organic matter content and base saturation
pH	negative log of hydrogen ions (unitless)

Introduction

In terrestrial environments, soil is the direct interface between geology and biology. Biology has varying influences on soil properties with depth resulting in a notable transition of biological activity from top to bottom along horizon boundaries. To understand the myriad of biological components in a soil and how they affect soil development, it requires a fairly large net to be cast, which would include everything from microbes and fungi, to invertebrate detritivores, herbivores, and carnivores, vertebrate species, and, of course, the plants and their root systems. The abundance and influence of each of these groups varies along climatic boundaries in terrestrial environments and provides the foundation for the changes in ecosystems we see along these boundaries.

Components of soil organic matter

Soil organic material (SOM) is made up of an incredibly wide range of compounds originating from biological contributions both in and above the soil. The amount of variation in organic carbon and nitrogen based compounds in soils is striking compared to that of the mineralogy of the soil parent material. The dominant source of the biomass contribution to the soil in most ecosystems is from terrestrial plants, although biomass from microorganisms and animals cannot be discounted. Plants shed leaves, branches, roots, propagate pollen, root exudates, and respire from the root systems, and these contributed parts are consumed by decomposers or herbivores. The consumed material is transformed into various excrements from those species, and then other decomposers break down that material even further and into increasingly recalcitrant compounds. The rate at which this occurs is strongly related to the temperature and moisture levels in the soil environment which results in the establishment of particular suites of soil organisms.

The quality of the organic material also plays a substantial role in soil development. This decomposition is primarily mediated by the microbial community. Decomposers in the soil preferentially consume detritus with a low carbon to nitrogen ratio. This suggests that the amount of nitrogen, one of the key macronutrients in the soil, is in sufficient amounts to fuel the metabolic activity. Carbon to nitrogen (C:N) ratios provide a fairly ubiquitous and cursory assessment of soil quality as it is a predictor of rates of biological activity in the soil itself. Organic compounds that contain a relatively large carbon to nitrogen ratio are considered to be of lower quality and do not decompose nearly as fast. As decomposition of organic compounds continues, increasing amounts of nitrogen are used in metabolic systems from the material, which increases the carbon to nitrogen ratio. This older organic material becomes increasingly recalcitrant, more stable, and tends to remain in the soil for much longer periods of time (Swangjang, 2015). For example, grasslands typically accumulate organic material with a lower C:N ratio while coniferous systems generate organic material that usually has a high C:N ratio. A mixed forested environment or typical hardwood system would fall between these two end members in their C:N ratios. In Table 1, a list of average carbon/nitrogen ratios are included for dominant ecosystems. Deciduous forests generally have higher ratios than evergreen forests and drier landscapes.

Organic matter in soils can be found in several different forms, all of which can exist at the same time in various stages of decomposition. Surface plant residue resides on the surface of the soil and contains leaf litter, grass, crop or pasture material. Buried plant residue is contained within the soil and contains plant material that is greater than 2 mm in size. Particulate plant material consists of partially decomposed material that is smaller than 2 mm. All of the above portions of organic matter are considered labile in that they will decompose in the order of days to years, based on other environmental factors where they are deposited. Organic matter that lacks an organic structure and can be determined to be specifically derived from plant or animal is characterized as humus. These fractions of organic matter are usually associated with soil aggregates and are smaller than 1 mm, and are considered to be more stable. This reflects their more advanced state of decomposition and in part the physical and chemical protection from non-organic compounds in the aggregate structures. They can remain stable in the soil from years to decades. Charcoal, charred material, and ash from fire related processes create resistant organic carbon (ROC), which will remain stable in the soil from decades to thousands of years (Stockmann et al., 2013).

Soil ecosystems

Some aspects of soil biology do not change a great deal from ecosystem to ecosystem. For example, the nitrogen cycle is controlled by the microbial biomass in all systems. A relatively large number of plants establish relationships with mycorrhizal fungi in most ecosystems and utilize these relationships to channel nutrients to the rhizosphere, which has a unique set of conditions than the areas of soil further from the root zones. Soil animals in most ecosystems around the world are dominated by arthropods (mostly

Table 1 Carbon to nitrogen ratios of organic material in ecosystems.

<i>Ecosystem type</i>	<i>Soil organic matter carbon/nitrogen ratio (unitless)</i>
Evergreen needle leaf forest	16.1
Evergreen broadleaf forest	12.8
Deciduous needle leaf forest	24.8
Deciduous broadleaf forest	30.0
Mixed forest	10.1
Shrub land	19.3
Woody savannah	15.0
Savannah	15.0
Grassland	13.1
Crop land	13.2
Barren or sparse vegetation	26.8

Data after Wang et al., 2010 based on modelling from Glopnet datasets for each biome listed.

mites and springtails) and nematodes, although relative proportions of those two phyla can shift when considering grassland compared with forest soils, for example. However, the relative abundances of all these species, the rate at which they perform their roles in transforming soil physical and chemical properties, and the way they operate can change as the other influences of soil development change. With the differing chemical compounds and structures in plant biomass from ecosystem to ecosystem, it is logical to expect plants would contribute organic material to soil in different ways and in varying amounts. It is therefore necessary to organize the discussion of this wide variation of biological influences in a way that can be more easily evaluated. One way to do this is to discuss the differences in soil biological activities in various ecosystems. Here, I have organized the soils into the common climatic zones and in the general order of moving from the tropical environments at the equator, through the grasslands and deserts of the lower latitudes, through the temperate forests and into to the boreal forest and tundra ecosystems. This does not imply that all soils in a climate belt are equivalent in nature with regard to their biology. Biology is but one of five soil forming factors, variation in the other factors (climate, topography, parent material, and the age of the soil) can have a marked effect on the soil biological components as can the variation in microclimate across a landscape. With the overwhelming number of species present in soil communities, it is impossible to discuss with any degree of certainty the precise proportions of soil organisms on an ecosystem scale. However, in recent years there have been enough data to identify generalized patterns of species distributions and the SOM that they are associated with on a large scale. Interestingly, while the amount of above-ground biomass is highest at the latitudes nearest to the equator, the largest amount of soil biomass is located in the high latitudes of the northern hemisphere associated with boreal and tundra ecosystems. However, the amounts of below ground microbial biomass remain fairly consistent through all the lower and mid latitudes and it increases substantially only in the high latitudes (Crowther et al., 2019).

The amount of organic carbon and nitrogen stored in ecosystems per unit area is dominated by wet and or cold areas (wetlands, boreal forests, and tundra) (Fig. 1). While the amount of microbially associated carbon and nitrogen in ecosystems with large organic matter content is also large, the total amount of microbial biomass does not vary equally and in proportion as organic matter varies. Microbial biomass can be somewhat limited by saturated conditions, therefore, the wetland microbial population is less than that of the tundra. When considering the amount of microbial biomass as a function of depth and amount of land covered by the tundra ecosystem, a different pattern emerges. Boreal forests, due to the amount of land, depth of the soil, and large organic matter accumulation, have the largest amount of microbially associated biomass at depth, followed by mixed forests, tropical and

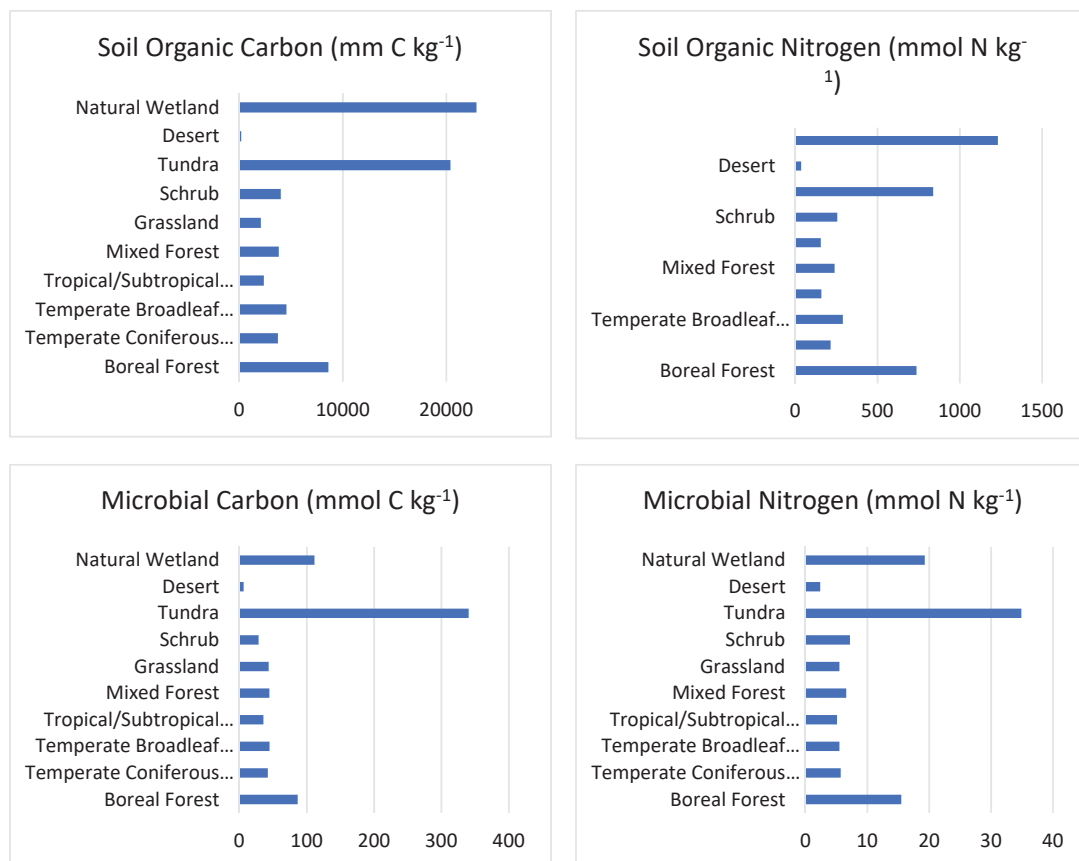


Fig. 1 Average carbon and nitrogen concentrations of soil organic material and microbial biomass in various ecosystems. Graph of data published in Xu et al., (2013).

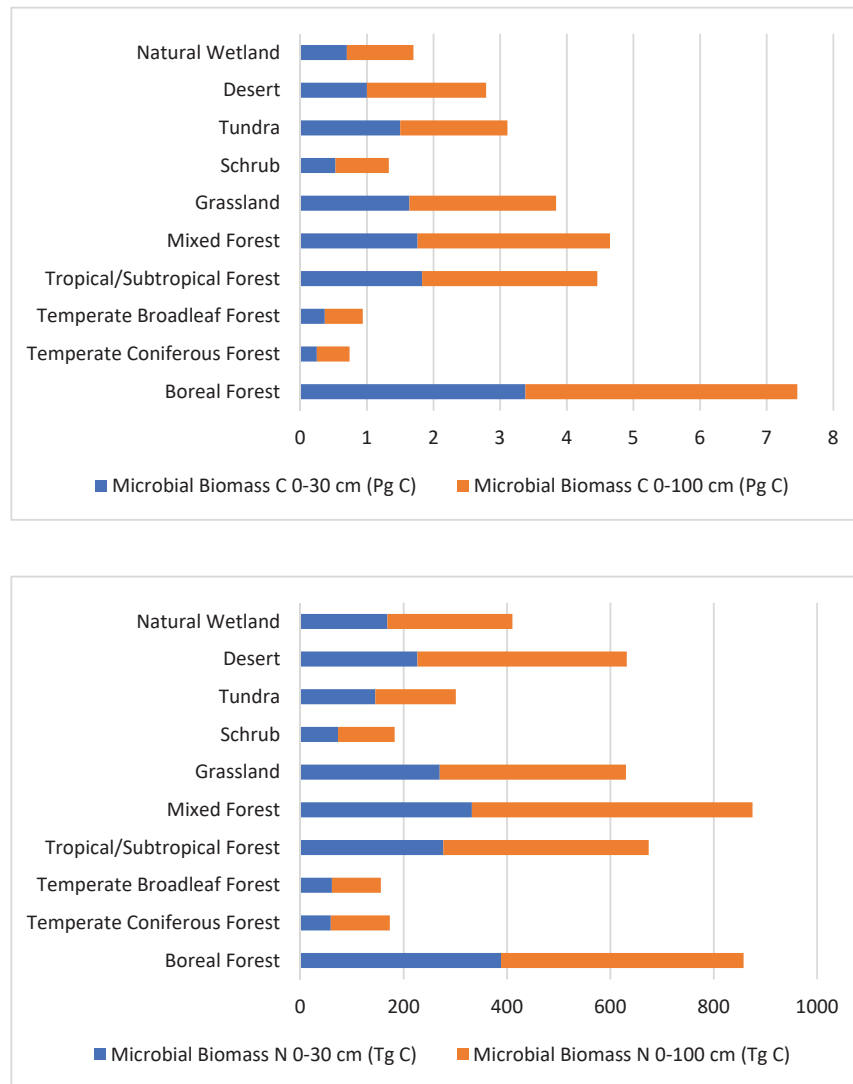


Fig. 2 Average microbial biomass in various biomes as a function of depth. Graph of data published in Xu et al. (2013).

subtropical forests, and grasslands (Fig. 2). Wetlands and tundra display a smaller overall share of the world's microbial biomass because of the smaller amount of land covered by wetlands and the shallower soil depth of soil in the tundra.

In addition, the discussion of soil biology in this chapter will exclude agricultural systems, which can be present in nearly all the natural ecosystems described below. With the wide variation in types of crops grown, amendments added to the soils, and human driven alterations of other physical and chemical properties on farmland, it can be difficult to consider soil biology in this context with any degree of accuracy. A great deal of research has gone into the structure and behavior of soil organisms under various forms of agriculture in most regions of the world (Bünemann et al., 2006; Bender et al., 2016; Hendrix et al., 2020). It is relatively common that the original soil community will remain intact for several years after a landscape is converted to agriculture. But over time, the below ground community will adjust and begin to reflect ecosystems more similar to those related to the crop being grown.

Tropical soil systems

The rate of decomposition in tropical soils is the fastest of all the ecosystems. This is primarily due to higher soil temperatures which increase biological activity and metabolic rates, and the lack of seasonality in plant growth (Krishna and Mohan, 2017). The soil community overall, despite their overall tendency to work intensively and fast at their various ecosystem functioning roles, is highly sensitive to changes in moisture and temperature. The community itself does not develop in such a way that it can deal with diverse survival strategies for harsh environments and is thus subject to major disruptions caused by drought, unexpected cold conditions, or fire events, for example.

This region is noted for incredibly high biological productivity on the surface and, in many cases, deep and highly weathered soils. While there is a great deal of organic matter in the soils of the tropics, these soils tend to have small amounts of biologically labile fractions, a smaller microbial biomass, but a very large fraction of organic material associated with clay minerals. The humus organic carbon fraction dominates, in part due to the high rate of clay formation related to chemical weathering activities. This keeps the SOM stable for long periods of time and results in the carbon being buried deep in the soils. This will, over time, develop into humin–mineral complexes with organic materials being bound in some way to inorganic material. Complexed material in soil is not exclusive to tropical zones, but is a common characteristic of the deeper horizons. This complexed material can help to hold soil water and supply deeper rooting systems with nutrients for longer periods of time and prevent increasing compaction from the weight of the thick soil above, which will allow microorganisms to function more easily (Bruun et al., 2010).

There can be considerable variability in soil communities across the tropical landscape depending on smaller scale variations of other soil forming factors. Environments that are poorly drained, for example, can result in peat systems, such as those in Indonesia. These high moisture and high productivity situations will result in histic epipedons with organic horizons extending beyond several meters in depth. In contrast, those areas along slopes with nutrient poor parent material can have very small amounts of organic materials and a very different above- and below-ground communities from those of the surrounding areas.

Savanna and grassland soil systems

The amount of organic matter contributed to the soil by grasses is greater than that of the deciduous forest environments because grasses contribute the majority of their biomass to the soil at the end of each growing season. In a typical grassland, there is a relatively equal distribution of biomass above- and below-ground in the soil from the grasses. When grasses die, they contribute all of their biomass either on or below the surface which contributes to the SOM of the soil system throughout the epipedon and into some of the subsurface horizons.

The greater proportion of woody material derived from either shrubs or trees in savannas makes the structure of their soil ecosystem somewhat different, especially in the warmer areas of the Earth. The A horizon is typically more nutrient poor than at higher latitude because of higher rates of decomposition and more litter with a large carbon to nitrogen ratio. In addition, some of these areas are dominated by ants and termites in the soil macrofauna rather than earthworms and springtails which are frequently more dominant in other biomes.

In traditional grasslands of the more temperate zones that have much less woody plant material, organic material is more concentrated in the first meter of soil due to the volume of biomass turned over annually by the grasses and their root systems. The leachate in these areas commonly has a higher pH than in the savannas, and the organic matter itself has a lower carbon to nitrogen ratio which encourages decomposition. The resulting epipedon is fairly soft and dark with a high base saturation.

Desert soil systems

The amount of biological activity and biomass in desert soils is lower than that of other ecosystems, as one would expect. The growth rate above the surface is slow due to the lack of water, which reduces the amount of food source available for the living biology below ground. The density of root systems is usually low because large gaps exist between plants that are competing for these limited resources. The frequency of naturally formed Entisols (Soil Survey Staff, 2014), which lack a B horizon, is much greater in these areas with the slow rate of soil development because weathering processes do not progress as rapidly and biological contributions are few. Compared with all the other ecosystem types, however, treating deserts as a single entity can be problematic because they can exist from the tropics to the poles; they are defined simply by a lack of precipitation rather than temperature. These different latitudinal locations can have a marked effect on the communities of soil dwelling fauna and flora and contradict what would be considered a standard description. For example, the strongly alkaline nature of hot desert soils and high salinity of hot desert basins would contrast markedly with the acidic nature of the non-calcareous polar deserts, and which would affect the soil dwelling biota. The incredibly short growing season in the polar deserts is also a severe restriction to above ground plant growth allowing the soils to develop in the near absence of root systems. What can be said, however, is that most deserts force biological communities to employ survival strategies for such extremes compared to the other soil ecosystems.

Although some parts of a desert landscape may appear barren at a glance, deserts can have quite active and resilient biocrusts on the surface. These crusts are created predominantly from microbial matter and take advantage of the unshaded and untrampled regions of the deserts to form a living surface across wide expanses. These communities of cyanobacteria, mosses, and lichen are usually the pioneering species in any newly exposed environments prior to soil formation. In the arid and semi-arid areas, however, they can form crusts of several millimeters in thickness on the surface and will serve to alter or enhance nutrient cycling, regulate the soil water balance and thermal energy balance of the soil system below (Bowker et al., 2018). These biocrusts will attach themselves to the surface substrate and slow erosion and, subsequently, the loss of epipedon material that the rest of the soil biological system relies on. The overall slow development of biological systems in water depleted regions is especially important to note when discussing these biocrusts and their propagation as new crusts in disturbed areas can take several hundreds to several thousands of years to re-establish themselves and return to their full previous function.

Subtropical and temperate systems

Deciduous forest systems

Deciduous ecosystems are dominant in the subtropical to temperate latitudes. Some of the soils of these ecosystems are exposed to frost during part of the year or to heavily frozen ground, therefore, the diversity and general composition of the various suites of organisms can change a great deal with latitude and elevation. In these deciduous environments, trends in soil activity occur in predictable annual pulses as the seasons change. As the leaves drop from the trees in the autumn, considerable organic material is added to the litter material on the soil surface. Decomposition of the litter can start and continue over winter, but the rate depends markedly on temperature. Large rates of decomposition in and on the soil typically do not start until spring.

Both macrofauna and bacteria are a dominant force in the decomposition of the leaf material, but fungal presence is also very high because of difficulty in the decomposition of lignin in woody material by other organisms. In deciduous forests, the fungal to bacterial ratio is usually between 5:1 and 10:1. Deciduous tree species can form relationships with either endomycorrhizae or ectomycorrhizae, which will result in an expansive network of filaments through the soil profile. This symbiotic relationship results in large rates of nutrient cycling and carbon mineralization.

Deciduous soils tend to have less organic matter in the A horizons compared with grassland systems due to the high turnover of grasses which incorporates more organic material to the system on an annual basis. However, the increase in clay content from enhanced chemical weathering processes in wetter environments in these forested soils leads to increased aggregation and physical and chemical protection of the SOM. This, combined with the greater rooting depth of the vegetation, will result in larger amounts of SOM throughout the profile than in grasslands. In addition, the presence of anecic earthworms, the deeper roots of trees and other animals that make vertical burrows cause organic material to be distributed deeper into the soil than in non-forested regions.

Coniferous forest systems

Coniferous systems contribute more plant tissue to the soil annually than any of the other ecosystem. However, the material is of very poor quality (high carbon to nitrogen ratios). This results in thick layers of dominantly fibrous detritus in the uppermost O horizon with much less organic material contributed to the A horizon than in deciduous forests. Despite this contribution, the organic matter that is incorporated is much more stable and more difficult to break down due to mineral associations and the formation of complexed compounds, which result in more organic material at depth as the soils develop. In addition, the average rooting area of conifers is wider than that of deciduous trees, contributing organic carbon in a larger perimeter in the soil from root shedding, exudates and other associations in the rhizosphere.

The overall acidic nature of coniferous systems tends to reduce the diversity of soil biological systems. This is primarily related to the pH of the surface plant residues which are deposited throughout the year. The reduced diversity is especially true of the macroinvertebrate detritivores, whose abundance declines precipitously during the transition from deciduous to coniferous forest. Earthworm density in coniferous forests is in the order of 10 times less than in deciduous forests due to the often low pH of the litter contributed to the soil there. The fungal to bacterial ratio is much higher; these can be in the order of 100:1 or 1000:1 depending on other soil forming factors (Korboulewsky et al., 2016).

A comparison of coniferous and deciduous soil systems from the piedmont of the south eastern USA with similar bedrock (gneiss) reveals some distinct differences in soils that are less than 600 m apart (Fig. 3). The acidic nature of the organic material incorporated from conifers causes the pH of the soil to decrease, which enhances the rates of chemical weathering. The formation of clay minerals from many silicate minerals by hydrolysis will produce siliceous acid, which further decreases the pH of the soils here. If the primary minerals in the soils contain metals, such as iron, this results in a much more oxidized soil with higher clay content than the soil forming below the deciduous forest.

Boreal forest systems

The suite of organisms in the boreal forest soils is often less diverse than that of the lower latitudes because of the need for the organisms to survive long periods of hard freezing and months of little to no activity in the soil. In the absence of permafrost, soil temperatures can stay at or above freezing below a few centimeters depth, which allows species to survive in the uppermost horizons. The hard freeze in the winter month causes dormancy of the organisms and most activity slows to a halt. However, in the spring during the thaw, this ecosystem can have a wide range of biological activity depending on local drainage systems in the area. The immense amount of snow melt and waterlogged areas can essentially turn forested lands to wetlands for portions of the year, which will result in pulses of biological activity from different types of organisms.

With slow rates of evaporation and perpetual shade from the evergreen species in the region, the soil can retain a great deal of moisture allowing thick beds of sphagnum moss to develop above the soil surface. Decomposition of the dead moss below is slow due to the low temperatures and high moisture content in the soil pores which encourages the rapid accumulation of organic material in the soil. As discussed above in the coniferous forest section, the high acid content and constant shedding of conifer



Fig. 3 Comparison of a soil from a coniferous ecosystem and deciduous ecosystem in Redlair, NC. The coniferous soil (A) has a lower pH and higher clay content than the deciduous soil (B). Despite these soils being located merely several hundred feet apart, the influence from the different plant and ecosystem contributions plays a role in altering both physical and chemical properties. (A) Very fine, kaolinitic, thermic Typic Kanhapludults (Soil Taxonomy) or Haplic Acrisol (WRB).

Depth (cm)	Horizon	Color	% Clay	Structure	pH
0–3	A	5YR 3/3	9	1vfsbk	6.0
3–15	Bt ₁	2.5 YR 4/3	19	2vfsbk	5.5
15–22	Bt ₂	2.5 YR 4/4	28	2fsbk	4.5
22+	Bt ₃	2.5 YR 5/4	37	3msbk	4.5

(B) Fine, kaolinitic, thermic Typic Kanhapludults (Soil Taxonomy) or Haplic Acrisol (WRB).

Depth (cm)	Horizon	Color	% Clay	Structure	pH
0–2	A ₁	7.5 YR 3/2	5	2fsbk	6.0
2–6	A ₂	7.5 YR 3/4	7	2fsbk	5.5
6–34	Bt ₁	5 YR 4/4	17	2msbk	5.0
34–52	Bt ₂	5 YR 4/3	26	3msbk	5.0
52+	C	5 YR 5/3	30	2csbk	5.0

needles contribute more biomass to the soil than any other ecosystem type allowing this boreal soil biome to store more organic carbon than all of the other soil types combined. This has had a marked impact on the overall ability for soils to sequester carbon from the atmosphere.

Bogs are another common feature of boreal lands that can result in a slightly different biome from that of the forested counterpart. As the glacial lakes in these regions slowly fill with sediment and organic detritus, the resulting peat bogs become large, treeless carbon sinks with a layer of sphagnum moss and other unique plants on the top. The anoxic layers of this soil lie just below the surface and are dominated by a community of bacteria that are both resistant to disturbance and slow working (Allison and Treseder, 2011). Nutrient cycling in these soils, as a result, is also slow, which necessitates the development of alternative nutrient sources for some primary producers, such as pitcher plants and other carnivorous plants that extract their nutrients from the insects they trap.



Fig. 4 The patterned ground of the tundra biome in the North Slope borough of Alaska results in soil organic matter inputs from vegetation restricted to mosses, grasses, and sedge species. Decomposition rates are slow due to the saturated conditions and cold temperatures resulting in high carbon sequestration potential. Photo credit: Ken Hill, National Park Service. Public Domain.

Tundra systems

The permanently frozen horizon of material in the soil system in the tundra causes diversity in the soil biology to become severely limited. In areas where the ice layer is too shallow to allow tree roots to penetrate far enough for stability, the surface of the landscape is dominated by short grass species, mosses, and lichens. Because there is a physical barrier to drainage across the landscape, much of the upper horizons of the soil remain saturated where there is sufficient precipitation throughout the year. This cold and wet soil environment results in very slow rates of decomposition and an accumulation of SOM above the permafrost (Fig. 4). For this reason, the boreal and tundra systems contain the largest amounts of SOM on the planet.

Evidence suggests that decomposition starts in early spring before the first snow melt due to microbial activity. The base of the snow pack is often at freezing point throughout the winter. As that temperature creeps above freezing, microbes will become active in the soil pores and begin to decompose material from the previous year. The largest factor in the diversity of the soil macro- and meso-fauna is the moisture content of the soil across microenvironments. While the more mesic soils have more protozoans, the mesofauna (collembola, mites, and enchytraeids) tend to have greater densities where the soil is drier (Kane, 1996).

Conclusion

The influence of biology on soil development is not an easy topic to discuss because the influence of other soil forming factors varies so widely across the landscape. The soil species diversity, density, and even the timing of activity of the species can change dramatically with latitude, altitude, and other physical and chemical influences. It is the biological system in the soil that allows carbon to be either stored or recycled to the biology above ground through conversion to carbon dioxide and back to the plant through photosynthesis, and which processes the organic contributions in such a way that allows other life forms to access the nutrients they need to survive. Plants enlist the help of all other flora and fauna within the soil to survive when obtaining some of their resources from either above or below ground, essentially bridging the line between earth and life. While our understanding of the intricate details of the biological influences on the soil is still being developed, the appreciation and importance we have for what we know the soil biology does is very large.

References

- Allison SD and Treseder KK (2011) Climate change feedbacks to microbial decomposition in boreal soils. *Fungal Ecology* 4(6): 362–374.
- Bender SF, Wagg C, and van der Heijden MG (2016) An underground revolution: biodiversity and soil ecological engineering for agricultural sustainability. *Trends in Ecology & Evolution* 31(6): 440–452.
- Bowker MA, Reed SC, Maestre FT, et al. (2018) Biocrusts: The living skin of the earth. *Plant Soil* 429: 1–7. <https://doi.org/10.1007/s11104-018-3735-1>.
- Bruun TB, Elberling B, and Christensen BT (2010) Lability of soil organic carbon in tropical soils with different clay minerals. *Soil Biology and Biochemistry* 42(6): 888–895.
- Bünemann EK, Schwenke GD, and Van Zwieten L (2006) Impact of agricultural inputs on soil organisms—A review. *Soil Research* 44(4): 379–406.

- Crowther TW, Van den Hoogen J, Wan J, Mayes MA, Keiser AD, Mo L, Averill C, and Maynard DS (2019) The global soil community and its influence on biogeochemistry. *Science* 365(6455).
- Hendrix PF, Crossley DA, Blair JM, and Coleman DC (2020) Soil biota as components of sustainable agroecosystems. In: *Sustainable Agricultural Systems*, pp. 637–654. CRC Press.
- Kane DL (1996) The impact of hydrologic perturbations on Arctic ecosystems induced by climate change. In: Oechel WC (ed.) *Global Change and Arctic Terrestrial Ecosystems*, pp. 63–81. Heidelberg: Springer.
- Korbolevsky N, Perez G, and Chauvat M (2016) How tree diversity affects soil fauna diversity: A review. *Soil Biology and Biochemistry* 94: 94–106.
- Krishna MP and Mohan M (2017) Litter decomposition in forest ecosystems: A review. *Energy, Ecology and Environment* 2(4): 236–249.
- Soil Survey Staff (2014) *Keys to Soil Taxonomy*, 12th edn. Washington, DC: United States Department of Agriculture–Natural Resources Conservation Service.
- Stockmann U, Adams MA, Crawford JW, Field DJ, Henakaarchchi N, Jenkins M, Minasny B, McBratney AB, De Courcelles VDR, Singh K, and Wheeler I (2013) The knowns, known unknowns and unknowns of sequestration of soil organic carbon. *Agriculture, Ecosystems & Environment* 164: 80–99.
- Swangjang K (2015) Soil carbon and nitrogen ratio in different land use. *International Conference on Advances in Environment Research* 87(10): 36–40.
- Wang YP, Law RM, and Pak B (2010) A global model of carbon, nitrogen and phosphorus cycles for the terrestrial biosphere. *Biogeosciences* 7(7): 2261–2282.
- Xu X, Thornton PE, and Post WM (2013) A global analysis of soil microbial biomass carbon, nitrogen and phosphorus in terrestrial ecosystems. *Global Ecology and Biogeography* 22(6): 737–749.

Soil forming factors: Parent material

Stephen Nortcliff, Soil Research Centre, School of Archaeology, Geography and Environmental Science, Reading, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

This is an update of K.R. Olson, FACTORS OF SOIL FORMATION I Parent Material, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 532-535, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00005-9>.

Introduction	34
Soil forming factors	37
Weathering	37
Influence of soil parent material on soil properties	39
Soil formation	39
Parent materials and soil classification	40
Summary	40
References	40

Abstract

The transformation of parent material is the starting point for soil development and as such it is one of the key environmental factors determining soil formation. Most parent materials are mineral matter, although there is a distinctive group of soils developed from organic materials such as peat or coral sand. In the first stages of soil development involving the weathering of minerals, the parent material is a major influence on the nature of the soil; in most soils this influence reduces through time. In addition to non-organic parent materials from in situ geological materials, many have formed from eroded and redeposited materials.

Key points

- The diversity of soil parent materials is large consisting of residual, transported and organic materials.
- In many environments soils are developed in relatively thin layers of transported superficial materials overlying the underlying solid geology.
- Reworking of soil parent materials within the landscape frequently results in complex soil patterns.
- The soils that develop at a site are the result of complex interactions of the soil forming factors.
- Soil parent material is subject to weathering; in some parent materials their distinctive features may persist, in others they may be modified and markedly different following weathering.
- Soil parent material was recognized as a major feature at the highest levels in early soil classifications; in more recent soil classifications such as Soil Taxonomy and the World Reference Base soil parent material is mostly used at lower levels to distinguish soils.

Introduction

Parent material is the material from which soils have developed. While some soils are developed from organic materials the vast majority have developed from mineral materials, and it is in this context that parent materials are most frequently discussed. For mineral soils the parent materials include both in situ consolidated and unconsolidated materials, and materials which may have been subject to soil development, transported and deposited with soil development beginning afresh at the new location. [Table 1](#) broadly summarizes this range of materials. The mineral matter derived from these materials is the starting point for soil development. In the early stages of soil development the parent material will normally have a major influence on the nature and properties of the soil developed, but with increasing development this influence frequently decreases. For organic soils the primary parent materials are decomposing materials of various plant types and macro-, meso- and micro-fauna.

Residual parent material develops in place from the underlying rock. It may have experienced intense weathering over a long period of time. In a warm and humid climate it is likely to be thoroughly oxidized and well leached with most readily soluble materials removed. Rock materials subject to intense weathering in humid climates often have a distinctive reddish or yellowish color due to the oxides of iron. In cooler climates and specifically in drier climates, weathering is less intense and oxidation and hydration of iron are less visible. In drier climates leaching is limited and alkali and alkaline earth cations may be present in larger concentrations.

Transported materials form a significant proportion of global soil materials. Colluvial debris is transported by gravity and locally by water. Under gravitational transport the materials are generally poorly or weakly sorted e.g. rock falls, talus slopes and avalanche

Table 1 Classification of soil parent materials.

Residual Parent Materials		
I	Igneous rocks and crystalline schists A. Acid rocks (gneiss, phyllite, mica schists etc.) B. Neutral rocks (diorite, microdiorite, andesite, etc.) C. Basic rocks (pyroxene, amphiboles, serpentine, etc.)	
II	Transition forms A1 } B1 } C1 }	Deposits of the coarse disintegration products of Class I. Mainly gravel moraines, arkose gravel and sandy alluvial deposits.
	A2 Quartzite B2 Shales C2 Calcareous shales	
III	Sedimentary Rocks A Quartz sandy rocks(sandstones, sands) B Clayey rocks (fine clay, clay, loam) C Carbonate rocks (limestones, dolomites, marl, etc.)	
Transported Parent Materials		
A By Gravity	- Colluvial	
B By Water	- Alluvial - Marine - Lacustrine	
C By Ice	- Glacial	
D By Wind	- Eolian	
E By Humans	- Natural and technogenic substrates, waste	

material. Where materials are moved locally by water, material is transported from upper parts of the landscape and deposited in the lower parts. The nature of these colluvial materials depends on the source of material higher in the landscape, but in general finer materials are preferentially transported.

Alluvial deposits are materials transported by streams and rivers and include flood plain deposits, alluvial fans and deltas. Flood plain deposits are laid down when the streams or rivers overtop their banks and flood surrounding areas. There is a distinct pattern of deposition, with coarser materials, sand and gravel, deposited closest to the channel and finer materials, silt and coarse clays, further away. In environments where there has been a fall in the base level of the river, the river may cut through the alluvium and the resultant features are left as terraces. Terraces are a feature of many river basins in northern hemisphere environments where there were numerous changes of base levels during the Pleistocene period. Alluvial fans occur where a river or stream leave a narrower valley with a steeper grade and descend to a broader low angle plain. The sediment is deposited as a fan, with coarser material deposited first. Deltaic deposits occur where rivers and streams enter lakes, seas or oceans. There is a reduction in velocity of the flow and hence the carrying capacity of the watercourses; the suspended material is deposited to form a delta.

Much of Northern Europe, North America and parts of Northern Asia were covered by ice sheets during the Pleistocene (10,000 to 2,500,000 years ago). The ice cover was not constant during this period and in warmer interglacial periods the ice cover was often less than 20% of their maximum extent. While there were local variations, it is generally accepted that in Europe and North America there were up to 20 distinct periods of maximum ice advance over a period of 1 to 1.5 million years separated by warmer interglacial periods when ice coverage was considerably reduced. With the advance of the ice sheets the weathered material at the surface (earlier soil development) was removed together with unweathered materials which were physically eroded from the solid rock. Frequently this material was carried considerable distances by the ice sheets and when the ice retreated these transported and eroded materials were deposited forming a veneer, known as drift, a few meters to tens of meters thick over the local rock. This material became the future parent material for soil development and is often markedly different from the underlying solid rock material. Locally, other glacial drift materials such as glacial outwash and lacustrine deposits were deposited associated with the often complex glacial landscape. As an example, Boulton et al. (1985) describe the nature of glacial geology and the associated patterns of deposition in the later mid-latitude ice sheets, providing a description of the wide range of materials present at the surface as parent materials for soil development.

Wind blown (eolian) deposits are found extensively as soil parent materials. Today the source of these materials is principally from sparsely vegetated environments often associated with deserts. In much of the Northern Hemisphere there are extensive eolian deposits, which are associated with the Pleistocene period. During this period fine unconsolidated materials, principally silt and fine sand, were picked up by wind from outwash surfaces and arid zones in locally deglaciated or periglacial zones and transported long distances before being deposited. This material is known as loess and it is estimated that there are deposits of varying thickness across 10% of the Earth's surface. These loess deposits are very variable in terms of thickness, ranging from layers tens of meters thick as on the Loess Plateau of Central China to a few tens of centimeters thick. Soil development in this latter situation may take place

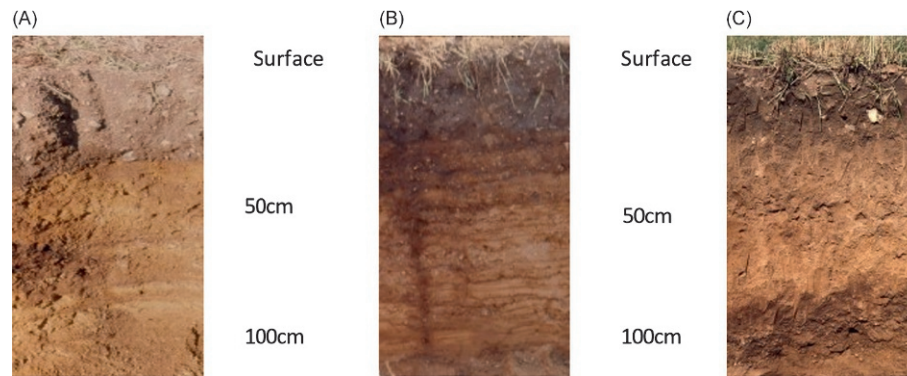


Fig. 1 (A) Hall Series – 60 cm of coverloam over morainic material; (B) Freckenham Series – Morainic material at the surface; (C) Sheringham Series – 110 cm of coverloam over morainic material. Photographs S. Nortcliff.

with two parent materials, the lower part of the profile developing from local geological materials or drift and the upper part from wind-blown deposits; these deposits have often been subject to weathering and soil formation processes prior to being transported. In some landscapes these thin veneers of wind blown material may also be subject to colluvial processes resulting in complex soil patterns. As an example, [Catt et al. \(1971\)](#) describe the distribution of wind-blown materials in the soils of northern Norfolk in the United Kingdom. These deposits are generally less than 1 m in depth and have been subject to subsequent erosion and deposition in the undulating landscape of the area.

A study by [Evans and Nortcliff \(1978\)](#) investigated the erosion and deposition patterns of the windblown deposits deposited in the Anglian Glaciation c. 450, 000 years ago that overlie gravelly morainic materials in an area of north Norfolk. This landscape consists of relatively flat crest surfaces with asymmetrical valleys ([Corbett and Tatler, 1974](#)). The soils identified as Hall Series on the flat crests and low angle slopes have a covering of between 50 and 70 cm of wind-blown fine sand and coarse silt (coverloam) overlying gravelly morainic materials ([Fig. 1A](#)). On the steeper slopes much of the windblown material has been eroded with the coarse morainic material at the surface; these soils have been identified as the Freckenham Series ([Fig. 1B](#)). In the valley bottoms, eroded windblown materials have accumulated with over 1 m-depth of fine sand and silt (coverloam) before the coarser morainic materials are encountered; these soils are identified as the Sheringham Series ([Fig. 1C](#)). These patterns are shown schematically in [Fig. 2](#). In many years these patterns can be seen from the air as crop growth is strong in the valley bottoms and exceptionally poor on the eroded slopes.

While these materials are broadly categorized as natural soil forming materials, increasingly in urban areas natural materials are moved around during development and frequently they are mixed with materials that might be broadly described as technogenic materials. Soil development begins afresh in these new materials which may be exceedingly diverse in character. Relocated natural materials arise when landscapes are levelled or excavated during construction projects, and the nature of the soil development that takes place subsequently on these relocated materials is highly dependent on their state during movement (in particular their moisture content) and how they were handled during extraction and placement at a new site. Technogenic materials result from human activity and are often used as fill materials, frequently with excavated natural materials applied as a thin veneer over them. These technogenic materials include building rubble, which often consists of brick and mortar debris, ashes from industrial processes such as coal fired power stations and incineration processes, slags from blast furnaces and steelworks, materials derived

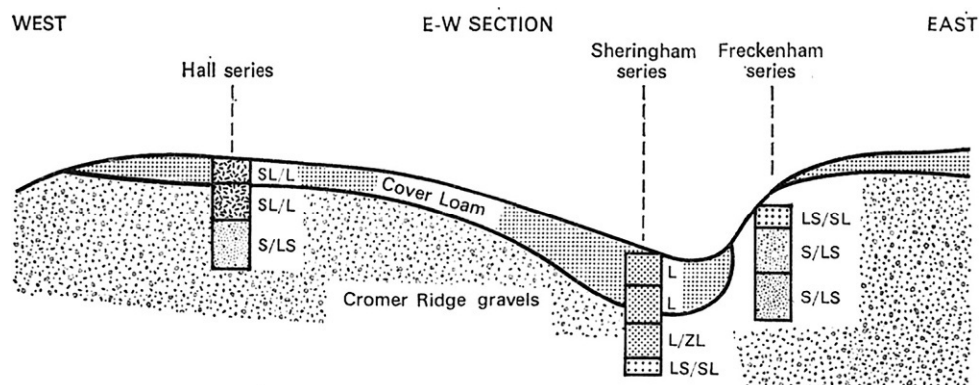


Fig. 2 A schematic diagram of the distribution of coverloam in the Cromer Ridge landscape. Soils Data © Cranfield University (NSRI) and for the Controller of HMSO 2022 used with permission.

from household waste management, sewage sludge and industrial sludges. These materials are very diverse and some such as ashes and slags may have strongly alkaline pH values and may contain large concentrations of contaminants such as metals and complex organic materials which strongly influence the soils that develop and also their use.

Soil forming factors

The diversity of soils found across the Earth's surface, even from parent materials of the same or similar nature and composition, is because of the interaction of the processes of soil development with external factors. Soil development is a slow process with changes taking place over long time periods. Dokuchaiev (1883) in Russia, Hilgard (1914) in the USA and Crowther (1930) in the United Kingdom considered the nature of soil development and the influences of environmental interactions. They also noted that the same parent material may form different types of soil depending upon the environmental conditions at the site of soil development, particularly the climate and vegetation. Dokuchaiev expressed this in a broad relationship:-

$$s = f(\text{cl}, p, v)^t \text{ where } s = \text{soil, cl} = \text{climate, } p = \text{parent material, } v = \text{vegetation/biota}$$

These relationships were changing through time^t.

Jenny (1941) developed these ideas and presented them in his seminal work *Factors of Soil Formation* presenting the concept that the state of the soil system is defined by the interaction of independent variables or soil forming factors. These were presented by Jenny in a generalized environmental formula of soil forming factors;

$$S = f(\text{cl, o, r, p, t, } \dots),$$

where S is some attribute or measurable property of soil, f() means 'as a function of,' cl is the environmental climate, o is species or organisms (often this is subdivided in to plant species, v, and living organisms (both macroscopic such as burrowing animals and microscopic such as bacteria and fungi)), r is relief or topography (including hydrological variations such as water table depth across the landscape), p is parent material (the initial state of the soil system) and t is time (age of the soil, absolute period of soil formation), with additional unspecified factors. More recently, with the increasing influence of human activity on soil development it has been considered that we should recognize an anthropic factor as a strong influence on soil development in many areas.

When considering the soil forming factors and their influence on soil formation, it is possible to consider their influences at different spatial scales. In the early attempts at soil classification in Russia and North America the different scales of influence of the soil forming factors were identified in the broad structure of the classification systems (Dokuchaiev (1883) in Russia and Marbut (1935) in USA). The highest level of these classifications distinguished Zonal Soil where the determining factors of soil formation were identified as climate and vegetation (tundra, taiga, deciduous forest, wooded steppe, steppe, semi-arid, and tropical soils). These broad zones were distinguished at the continental scale. The second level was Intrazonal Soils where the dominant factors were parent material and topography (calcareous, saline and hydromorphic soils); these soils were often distinguished at scales of tens to hundreds kilometers. Azonal Soils were the third category where soil formation was not advanced and included soils developed in recent colluvium and alluvium; these were frequently distinguished at the local scale of less than 1 km. Glinka (1927) summarized this approach presenting a global view of the characterization and classification of soils. While providing a relatively straightforward starting point, this approach was unable to account for the complex interrelationships between the soil forming factors which resulted in the tremendous diversity of soils at the Earth's surface.

Increasingly the interrelationships of the soil forming factors and their influence on the nature of soils was recognized as more complex than those at the broad continental scale which was the focus in these earlier approaches. With time these many more subtle relationships were recognized at scales from hundreds to tens of kilometers and less as understanding developed. The factors and their interactions define the state of the soil system. The soil system has been defined as an arbitrary volume of the solum, a soil body, a pedon (the smallest three-dimensional body of soil with lateral dimensions large enough to permit the study of horizon shapes and relations in it). The soil system is open to energy and matter influxes and outfluxes. Influx of energy includes solar radiation, and heat and entropy transfer; outflux of energy includes heat radiation and the reflection of light. Influx of matter involves gases entering the system by mass flow (wind), water in liquid and solid forms, solids dispersed in air or dissolved or dispersed in water, and organisms that migrate into the system.

Within this suite of properties in mineral soils, the parent material is the initial mineral material from which the soil will develop and is the initial state of the soil system.

Weathering

The fundamental process transforming mineral soil parent material into soil is weathering. Rock weathering is the in situ alteration and fragmentation of rock materials exposed to the atmosphere or hydrosphere. Weathering consists principally of the physical and chemical alteration of rocks or minerals at the Earth's surface, although it is increasingly recognized that biological processes also contribute to this alteration in many environments. These processes may occur simultaneously or in sequence. For example, the initial physical breakdown of rocks increases the surface area accessible to chemical and biological weathering.

Most rock materials when exposed at or near the Earth's surface are in an environment markedly different from that in which they were formed and when exposed are not in equilibrium with their environment, this is particularly true for igneous and metamorphic rocks. Weathering is their transformation to a form that is more stable under these near surface conditions. The weathering process results in the transformation of primary minerals into smaller particles and secondary minerals together with the release of other constituents which may be removed by leaching.

Physical weathering is the rupturing of rocks as a result of physical stress. As rocks are exposed and the confining pressures removed they may expand and rupture along fracture planes. Water may penetrate fractures and on freezing can cause pressure which may cause disintegration. Plant growth may contribute to physical weathering as a result of the pressure exerted by roots, which can sometimes rupture the rocks and force blocks apart.

Chemical weathering resulting in the transformation of the original mineral material brings about many of the changes that give the soil its distinctive characteristics. Chemical weathering occurs because exposed rocks and minerals are not at equilibrium with their environmental conditions. Weathering produces materials that are more stable in these near surface environments, but may continue to change as the pedo-environmental conditions change. There are several principal processes of chemical weathering:

Dissolution is the process whereby water soluble salts (sodium chloride, NaCl; gypsum, $\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$) dissolve through attachment of H_2O dipoles to cations and anions of the crystal framework with the removal of ions in solution. Weathering of carbonate rocks gives a distinctive suite of soils. Calcium carbonate is quite soluble under surface conditions where CO_2 levels are enhanced. In this reaction the resulting calcium and bicarbonate ions are both soluble and readily leached from the system, therefore, the properties of the soils resulting from these dissolution processes depend strongly on the nature of any insoluble fraction in the parent rock. (see Fig. 3 below).

Hydrolysis is the decomposition of salts, if formed by a strong base and a weak acid or a weak base and a strong acid; water dissociates into H^+ and OH^- ions that attack (hydrolyze) the salt. Hydrolysis is a key process in chemical weathering, particularly of silicate and aluminosilicate minerals. This process is slow and the reactions occur predominantly on the surface of the crystals. Slightly acidic water together with a variety of ions present in solution react with these minerals. For aluminosilicates the products of these processes are the clay minerals, which play a key role in determining the nature and properties of soils and their development.



Fig. 3 Rendic Leptosol, developed on Jurassic Limestone under mature beech woodland (*Fagus sylvatica* L.), Germany. Photograph: Jörg Prietzel.

Oxidation is the process by which an element loses an electron resulting in an increase in the positive valency of the element. Examples include oxidation of the reduced elements iron (Fe^{II}), silicon (Si^{II} , Si^{I}), manganese (Mn^{II}) by O_2 (oxygen) in the presence of H_2O and microorganisms to the corresponding oxidized states Fe^{III} , Si^{VI} , Mn^{III} and Mn^{IV} . Iron is the most commonly oxidized element in the soil environment, and is associated with the distinctive brown or reddish brown colors of many soils. The presence of Mn darkens the soil color.

Hydration and dehydration are processes by which water molecules are added to or removed from a mineral, resulting in the formation of a new mineral. A common reaction in soils involves the reaction of hematite with water to form goethite. In well drained soils the presence of hematite gives the soils the distinctive reddish and reddish brown colors.

Ion exchange between the soil solution and soil mineral is an important process in the transformation of one clay mineral to another.

Influence of soil parent material on soil properties

The rate at which soil development occurs depends principally on climate, including rainfall and temperature, but the other factors will also influence the nature and rate of development.

In the early stages of soil development the parent material often has an over-riding influence on the nature of the soil properties which develop, but through time the effects of other factors increase and the dominant effect of parent material is greatly diminished or eliminated. Two of the important soil properties that are inherited from soil parent material are the mineral composition of the soil and its texture. Although these properties are often altered with time during soil development, in some soils the influence of the parent material persists. Wilson (2006) compared the nature of soils developed on two common but contrasting soil parent materials, granite and basalt, through time. The general characteristics of the soils developed from these two contrasting parent materials were markedly different in the early stages of development, but through time and increased weathering many of these contrasts have persisted but some have become less marked. The sand mineralogy of soils developed from granite tends to consist of quartz, K-feldspar, Na-feldspar, muscovite, biotite and amphiboles, whereas for the basalt-derived soils it consists of Ca-feldspar, orthopyroxene, clinopyroxene, olivine, magnetite and smectite. In the early stages of soil development there would be marked differences in the clay mineralogy of the soils developed from these two parent materials. Granitic soils are dominated by illite and Al-hydroxy interlayered vermiculite as well as vermiculite itself. In contrast, soils developed on basaltic materials contain little or no illite and often contain smectite, either directly inherited from the parent rock or derived due to pedogenesis. In old well weathered soils derived from both groups of parent materials these marked differences in clay mineralogy disappear, and both groups show broad similarities with kaolinite, halloysite, gibbsite and iron oxides becoming most prevalent. There are marked differences in the soil physical characteristics of the soils from the two parent materials, differences which persist as the soils are weathered. The clay content in soils derived from basaltic rocks is consistently higher than in those developed from granitic materials and these differences persist in older well weathered soils. Young soils derived from granitic materials have higher porosity and hydraulic conductivity, but lower water retention and water availability than those derived from basaltic materials, and these differences persist with time. In terms of the soil chemical properties, the pH of young soils developed on granite parent materials tends to be below 5.5 and is often below 5.0. In contrast, the pH of young soils developed on basaltic materials is generally within the range of 5.0 to 7.5, and for older more weathered soils the pH range is normally between 5.0 and 6.5. The percentage base saturation shows similar patterns; in young granitic soils it is often <20% but >60% in young basaltic soils. In older soils these values are lower, but marked differences persist.

Where a parent material is rich in calcium carbonate its weathering and the leaching of soluble by-products usually results in distinctive soils where the small proportion of insoluble residue in the material results in soils with a small mineral fraction and a large proportion of organic matter. Fig. 3 is a Rendic Leptosol from Germany developed from Jurassic Limestone under a stand of mature beech (*Fagus sylvatica*). The soil has a shallow organic rich layer (<25 cm) including limestone fragments overlying fragmented Jurassic Limestone. Further analytical details are provided in Prietzel et al. (2021) and Prietzel et al. (2022).

Soil formation

Soils develop from parent materials that are exposed at the Earth's surface as a result of changes which occur within the soil system. These parent materials can be bedrock exposed at the Earth's surface or relatively unconsolidated materials which have been deposited in the landscape. The result over time will frequently be a soil profile exhibiting a sequence of distinctive layers, known as soil horizons. The lowest strata of the soil profile are known as the C horizons; these horizons show the least alteration from the nature of the original parent material. The process of transformation from rock materials into soil profiles is known as soil genesis.

While the processes that bring about changes in the material and development of a soil profile are complex, Simonson (1959) suggested a simple framework of broad processes:-

- Additions of materials
- Removal of materials
- Transformation of materials
- Transfers of materials within the soil

The principal influence of parent material on this development is expressed through the transformation of these materials as a result of weathering in the upper part of the profile, which has properties that are progressively different from the original parent material and with less weathered material deeper in the profile. In dynamic landscapes there may also be additions of material to the soil surface through alluvial, colluvial or eolian processes that result in complex soil profiles. Within the soil profile there may then be transfers of material, for example, clays may be transferred from upper horizons and deposited in clay enriched argillic horizons. Similarly, iron, aluminium and organic matter may be transferred from upper soil layers and deposited deeper in the soil in podzolic or spodic horizons. In some arid and semi-arid environments salts may be transferred from solutions deeper in the soil to be deposited in saline or alkaline layers near or at the soil surface.

Parent materials and soil classification

As mentioned above, much of the early work on soil classification focused upon the zonal concept which had been developed in Russia by Dokuchaiev (1883) and in the USA by Marbut (1928) and Baldwin et al. (1938). As knowledge about soils increased it became increasingly difficult to accommodate soils satisfactorily into this classification structure. There was a move away from a soil forming factor-based system of classification toward a quantitative property-based system. This culminated in 1960 with the "Seventh Approximation" (Soil Survey Staff, 1960) where the higher levels of the classification (Soil Orders) are distinguished by the presence or absence of one or more diagnostic horizon. This publication provided a basis for discussion and in 1975 Soil Taxonomy (Soil Survey Staff, 1975, 1999) was published. Soil Taxonomy is a strongly hierarchical system with 12 Orders at the highest level and a further 5 levels of subdivision with series at the lowest level. There was a shift away from the soil forming factor approach as the basis of soil classification in Soil Taxonomy, however, the factors as key elements of the classification are evident at many levels of the classification.

Soil parent material is used at various levels of Soil Taxonomy to identify soil taxa. At the highest level of Soil Order it is used to distinguish Andisols, Entisols and Histosols. Andisols are soils formed in volcanic ash and defined as containing a high proportion of glass and amorphous colloidal materials. Entisols are broadly described as soils that show very limited soil development and consequently are relatively unaltered from their parent material. Histosols are derived from organic materials. Parent material is also used to define taxa at lower levels, for example at subgroup level with modifiers including Andic, Arenic, Lithic and Paralithic.

While Soil Taxonomy is widely used there is another globally important soil classification system the World Reference Base for Soil Resources (IUSS Working Group, WRB, 2014). This latter system consists of a flat hierarchy with only two levels. At the first level, 32 soil reference groups (RSGs) are recognized, with the second level distinguished by modifiers of the RSGs. Leptosols, Regosols and Fluvisols are the RSGs which represent soils that are at the initial stages of soil development and soil parent material is a dominant influence on the nature of the soils. Andosols are soils formed in volcanic materials. Arenosols are soils where the texture is sandy because of the sandy texture of the parent materials. As with Soil Taxonomy, the WRB distinguishes Histosols as a distinct group of soils developed from organic materials.

Summary

The five soil-forming factors – parent material, climate, biota, topography, and time – determine the distribution and nature of soils over the Earth's surface. Mineral components of many soils are inherited from parent materials, while others develop in situ during weathering and pedogenesis. Morphological properties of soil develop as a result of the processes acting on parent materials. Soil scientists have attempted to show the controlling effects of parent material on soil properties to the extent that it is proposed as an independent soil-forming factor and defined as the state of the soil system at time zero of soil formation. The physical body of soil and its associated mineralogical and chemical properties provide the basis for the operation of the soil forming processes under the influence of the other soil-forming factors. Previously weathered rock or even a previous soil could be considered parent material, which can be modified over time. Parent material exerts its strongest influence on the soil-forming process of immature soils.

References

- Baldwin M, Kellogg CE, and Thorp J (1938) Soil classification. In: *Soils and Men, USDA Yearbook*. Washington, DC, 979-1001: Govt. Printing Office.
- Boulton GS, Smith GD, Jones AS, and Newsome J (1985) Glacial geology and glacial deposition of the last mid-latitude ice sheets. *Journal of the Geological Society of London* 142: 447-474.
- Catt JA, Corbett WM, Hodge CAH, Tatler W, and Wier AH (1971) Loess in the soils of North Norfolk. *Journal of Soil Science* 22: 444-452.
- Corbett WM and Tatler W (1974) *Soils in Norfolk II Sheets TG 13/14 (Barnham/Sheringham)*, Soil Survey. Harpenden: Rothamsted Experimental Station.
- Crowther EM (1930) The relationship of climate and geological factors to the composition of soil clay and the distribution of soil types. *Proceedings of the Royal Society B* 107: 1-30.
- Dokuchaiev VV (1883) *Russian Chernozem. Selected Works of V.V. Dokuchaiev Vol. 1*. Israel Program of Scientific Translations, Jerusalem (translated 1967).
- Evans R and Nortcliff S (1978) Soil erosion in north Norfolk. *Journal of Agriculture Science (Cambridge)* 90: 185-192.
- Glinka KD (1927) *The Great Soil Groups of the World and Their Development*. Translation from German by C.F. Marbut, Edwards, Ann Arbor, MI.
- Hilgard EW (1914) *Soils*. New York, NY: The Macmillan Company.
- IUSS Working Group WRB: World reference base for soil resources 2014, Schad P, Van Huyssteen C, and Micheli E (eds.) (2014) *World Soil Resources Reports 106*. Rome: FAO. Update 2015.

- Jenny H (1941) *Factors of soil formation. A system of Quantitative Pedology*. New York: McGraw Hill.
- Marbut CF (1928) A scheme for soil classification. In: *1st Int. Congr. Soil Sci. Comm. 5 Proc. And Papers*, 4 1–31.
- Marbut CF (1935) Soils of the United States. In: Baker OE (ed.) *Atlas of American Agriculture, USDA Bureau of Chemistry and Soils*, pp. 1–98. Washington, DC: U.S. Govt. Printing Office.
- Prietzl J, Klysubun W, and Colocho Hurtate LC (2021) The fate of calcium in temperate forest soils: A Ca K-edge XANES study. *Biogeochemistry* 152: 195–222.
- Prietzl J, Krüger J, Kaiser K, Amelung W, Bauke SL, Dippold MA, Kandeler E, Klysubun W, Lewandowski H, Löppmann S, Luster J, Marhan S, Puhlmann H, Schmitt M, Siegenthaler MB, Siemens J, Spielvogel S, Willbold S, Wolff J, and Lang F (2022) Soil phosphorus status and P nutrition strategies in European beech forests on carbonate compared to silicate parent material. *Biogeochemistry* 158: 39–72.
- Simonson RW (1959) Outline of a generalised theory of Soil Science. *Soil Science Society of America Proceedings* 23: 152–156.
- Soil Survey Staff (1960) *Soil Classification, A Comprehensive System, 7th Approximation*. Washington, DC: U.S. Govt. Printing Office.
- Soil Survey Staff (1975) *Soil Taxonomy. A Basic System of Soil Classification for Making and Interpreting Soil Surveys. Agricultural Handbook #436*. Washington, DC: Soil Survey Staff.
- Soil Survey Staff (1999) *Soil Taxonomy*, 2nd edn Washington, DC: USDA Natural Resources Conservation Services.
- Wilson MJ (2006) Factors of Soil formation: Parent material as exemplified by a comparison of granitic and basaltic soils. In: Certini G and Skallerghe R (eds.) *Soils: Basic Concepts and Future Challenges*, pp. 211–218. Cambridge University Press.

Parent materials: Loess

Tobias Sprafke^{a,b}, ^aKompetenzzentrum Boden (Center of Competence for Soils), BFH-HAFL, Zollikofen, Switzerland; ^bInstitute of Geography, University of Bern, Bern, Switzerland

© 2023 Elsevier Ltd. All rights reserved.

Introduction	42
Genesis	42
Distribution	44
Properties	44
Temperate-zone loess soils	45
Dust, soil, (paleo)environments	46
Conclusions	47
Acknowledgments	47
References	47

Abstract

Loess is commonly defined as a distinct body of mainly wind-blown silt and covers c. 10% of the Earth's surface, mainly in the temperate zone. Silicate silt predominates over carbonate, clay and some fine sand, making loess a favorable soil parent material. Thick loess deposits are well-mapped and important archives of Quaternary landscape and human evolution, whereas thin loess deposits and dust admixtures into soils are rarely studied, despite their usually positive influence on soil properties.

Key points

- Loess is commonly defined as the accumulation of wind-blown dust, but in situ processes are essential to provide its coherent structure.
- Loess covers c. 10% of the Earth's surface, mainly in temperate zone lowlands, outside areas glaciated during the Pleistocene.
- Loess is mainly (coarse) silt with minor clay and sand admixtures, lightly cemented by clays and carbonates.
- Loess soils are among the most fertile soils worldwide and thin covers/admixtures of mineral dust to parent materials usually have favorable effects on soil functions.
- Loess–paleosol sequences are widespread terrestrial paleoenvironmental records and can also preserve open-air Paleolithic sites.

Introduction

Loess, frequently defined as an accumulation of wind-blown dust, covers c. 10% of the earth's surface (Fig. 1; Busacca and Sweeney, 2005). This common estimate excludes thin loess deposits of a few decimeters thickness, and dust contributions to soils and soil parent materials. Loess is widely recognized as a parent material for fertile soils, geotechnically distinct ground in densely populated regions, Quaternary paleoenvironmental archive, and preserver of Paleolithic heritage.

Loess is a term initially adopted from people of the Upper Rhine Valley (SW Germany), where loess is an important element of landscapes and soils. The first scientific description of loess, originally interpreted as a water-lain deposit, dates back c. 200 years and highlights its silt loam texture and earthy structure (Leonhard, 1824). Since then, there have been controversies concerning the definition of loess, due to its intermediate position between sediment and soil (Sprafke and Obrecht, 2016). The predominance of silt results from preferential particle selection by wind, the coherent structure is due to "loessification," a weak cementation, attributed to initial pedogenesis or weak diagenesis.

Being an earth surface deposit, loess and related sediments are frequently affected by reworking and pedogenesis. It is difficult to draw boundaries between typical loess and loess-like deposits (loess structure, but no purely aeolian deposit) or loess derivatives (altered/reworked loess or loess-like deposits). By considering genetic criteria it is possible to subdivide loess sediments into several subtypes (Koch and Neumeister, 2005). Pécsi (1990) lists 10 criteria to define loess, but in most cases a simple definition of loess as a distinct body of mainly wind-blown silt is sufficient.

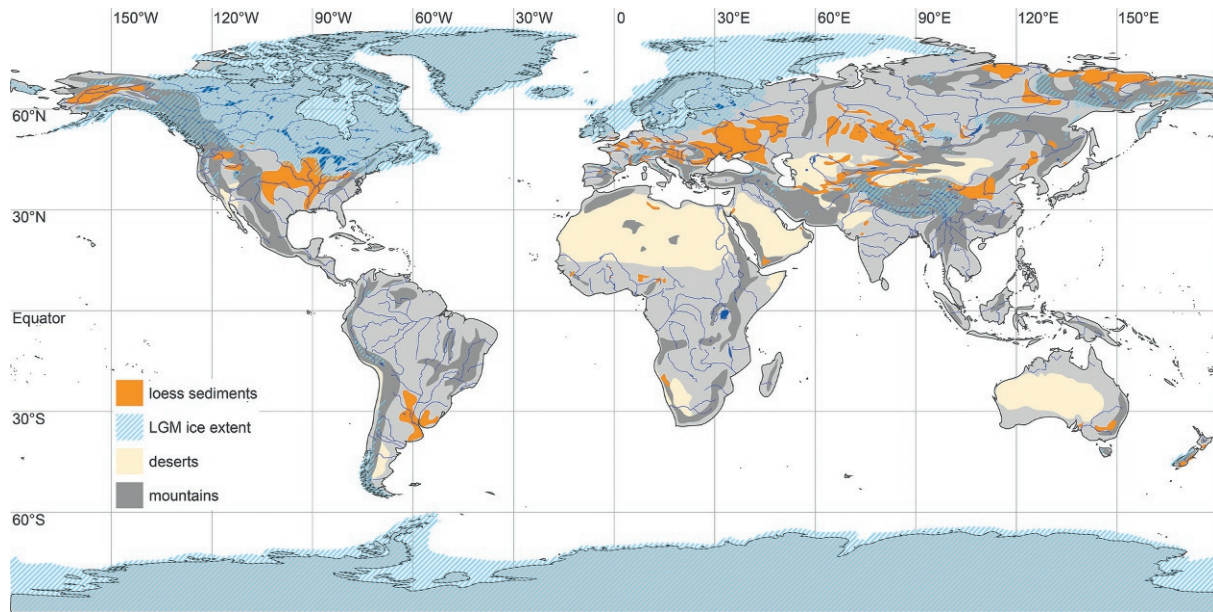


Fig. 1 Global distribution of major loess deposits, deserts and mountains and extent of ice sheets during the last glacial maximum (LGM; c. 24,000 years ago). Compiled with Adobe Illustrator[®] using shapefiles and satellite imagery (for desert and mountain extent) projected on naturalearthdata.com, assembled with mapshaper.org. Loess extent digitized from Li et al. (2020), LGM ice extent from Ehlers et al. (2011).

Genesis

The formation of loess involves silt production, transport, deposition and post-depositional processes. From a sedimentological perspective, aeolian transport and deposition are essential to create the silty texture of loess. A pedological perspective highlights in situ processes creating loess structure, which is important from a geotechnical point of view. Fig. 2 conceptually integrates both views.

Silt is the predominant particle size of loess and is produced from larger particles by a number of processes (Wright, 2001): Silt-sized mineral particles are a common byproduct of physical weathering in deserts or periglacial environments, but may also be released by chemical weathering. During fluvial or aeolian transport, silt can be formed by comminution and abrasion. Among the most efficient processes to produce large quantities of coarse silt, typical for loess, is glacial grinding (Smalley, 1995). Muhs and Bettis (2003) differentiate two main models of loess formation from particle production to aeolian deposition, the “classical” glacial and the desert model; an important difference being that the desert model requires no glacial grinding in silt formation (Lancaster, 2020). Next to silt production, recycling of older deposits contributes significantly to silt availability.

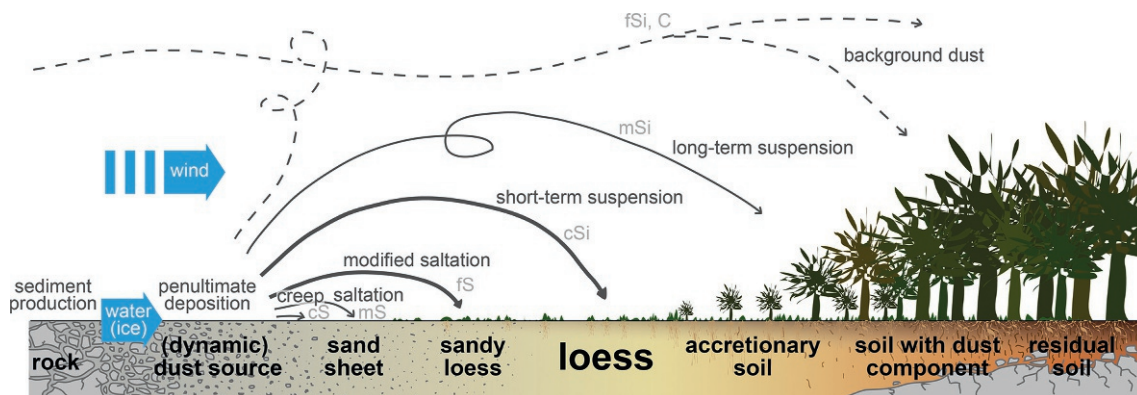


Fig. 2 Loess and the rock–sediment–soil continuum. Thick blue arrows indicate the main transport agents from the place of sediment production; the final transport is by wind. Local reworking is excluded from this simplified scheme. Thin gray arrows show aeolian transport modes adapted from Pye (1995). Common grain sizes for the respective transport mode are indicated in light gray (S: sand, Si: silt, C: clay; c: coarse, m: medium, f: fine). Note that available grain sizes differ across environments. Residual soils form from local parent materials without a significant contribution of mineral dust. These may be humid tropical soils far from deserts, but also temperate zone soils that formed from the loess of Pleistocene glacial periods.

Free silt minerals can easily be entrained and transported by wind as mineral dust, whereas sand may be too heavy and clay too cohesive. Direct aeolian mobilization of silt from production areas is possible, but in the complex sequence of processes that lead to the formation of loess, usually at least one phase of non-aeolian transport takes place before the final aeolian transport and deposition of mineral dust (Wright, 2001). This explains the location of major loess deposits close to large rivers, especially those draining centers of extensive Pleistocene glaciations (Smalley et al., 2009).

Deserts are the main contemporary source of aeolian dust, which contributes to soils in their surroundings and even to distant ecosystems. For example, Saharan dust reportedly provides nutrients to the Amazon rainforest. In semi-arid regions, partly connected to large river systems, loess forms from deposits of deflated silt (Lancaster, 2020). In periglacial environments of Alaska and Northeast Siberia, silt deflated from dry beds of braided rivers can accumulate as distinct sedimentary bodies. However, most loess deposits in the currently humid to semi-humid temperate zone formed during glacial periods when large ice masses reached far into the temperate latitudes, which were largely dominated by tundra or (cold) steppe ecosystems drained by rivers loaded with sediment.

Next to source properties, transport mode and distance determine the granulometry of aeolian deposits (Fig. 2; Pye, 1995). Close to deflation centers, sand and sand loess deposits transported by saltation and lifted by turbulent airflow are common, whereas with increasing distance over tens of kilometers, silt transported through suspension predominates. Fine silt can be transported over thousands of kilometers. Loess usually has a mixed provenance, with a predominance of regional components in the coarse silt mode and variable finer components deposited after long-range transport.

Deposition of mineral dust takes place when wind-speed is reduced, which is behind topographic obstacles and strongly supported by local vegetation. Considerable dust deposition typically occurs in polar to subtropical semi-desert/steppe environments. Herbs, grasses and cyanobacteria trap silt particles which are successively incorporated into the permanent silt deposit. The transformation of loose silt to loess is commonly referred to as “loessification,” a complex (bio)physiochemical process between pedogenesis and diagenesis that remains insufficiently studied (Sprafke and Obrecht, 2016).

Typical loess consists essentially of windblown silt. Yet, within the depositional environment, some short-scale reworking by water may occur, potentially involving the admixture of non-aeolian material, which may differ in grain size (Vandenbergh et al., 2018). Loess-like deposits are not purely aeolian deposits with a coherent loess-like structure (Sprafke and Obrecht, 2016). Reworking of loess and loess-like sediments creates loess derivatives, dominated by silt, but lacking loess structure. The distinction between different types of loess sediments is not always straightforward and is further hampered when these are overprinted by pedogenesis during phases of morphodynamic stability.

Distribution

There is no global agreement on how to define loess and related sediments, and regional loess maps rely on the standards of the geological surveys involved in mapping. By convention, loess less than 1–2 m thick is not mapped on most geological maps. This is unfortunate for pedologists and ecologists because even a few decimeters of loess strongly influence the pathways of pedogenesis and finally soil properties (Lorz et al., 2013). To derive European loess maps, Bertran et al. (2016) used topsoil silt data to model loess distribution, whereas Lehmkuhl et al. (2020) compiled data from numerous regional geological or pedological surveys. Global scale loess maps only approximate the main loess regions characterized by thick loess sediments.

Li et al. (2020) provide a global compilation of loess distribution and information on the key genetic processes of the different loess landscapes (Fig. 1). The most remarkable loess landscape is the Chinese loess plateau with more than 300 m-thick loess in the middle reaches of the Yellow River, which drains (semi-)arid and mountainous regions further west. Between lowland deserts and mountains of Central Asia, loess can reach more than 100 m thickness. Loess that is tens of meters thick drapes Western Siberia, Southeastern Europe, and Midwest USA. Continuous loess covers Eastern Europe and Northern Argentina. The European loess belt is subdivided by numerous low mountain ranges.

Rare occurrences of desert margin loess are reported from NW India, Nigeria and Namibia. Periglacial regions drained by the larger rivers in Alaska and NE Siberia provide modern analogs of Pleistocene loess formation in the temperate zone. Outside the common definitions for loess are alluvial silts that have formed by periodic flooding, e.g. in the vast Hungarian lowlands, but appear loess-like upon drying. Loess mapped in subtropical regions of China, e.g., in the lower reaches of the Yangtze River may rather be accretionary soils or even soils with aeolian admixtures. Yellowish tephra deposits intercalated by paleosols in Central Mexico (Sedov et al., 2009) may be called volcanic loess, but are never indicated on global loess maps. Tropical loess reported for NE Argentina and S Brazil has caused strong controversies, as it is most likely deep red soil (Morrás et al., 2009).

Properties

Loess sediments are characterized by the predominance of silt, with a distinct mode in the coarse silt fraction (20–60 μm), less commonly in the medium silt fraction. Sandy loess contains more than 20% sand and clayey loess more than 20% clay (Pye, 1995). Loess is usually homogenous, without any clear stratification. The mineralogical composition depends on the source region, but a predominance of quartz (40–70%) is typical for most loess regions. Common minerals include feldspars, micas, carbonates, oxides and clay minerals (e.g., kaolinite, illite). While there are regional variations in loess geochemistry, its average composition resembles the upper continental crust (Taylor et al., 1983).

From a pedological point of view, the content of carbonate in loess is specifically relevant. Calcite and to a lesser degree dolomite, are comparatively quickly leached from loess in a humid climate, facilitating advanced chemical weathering and leaching of further base cations, once the carbonate buffer is removed. Depending on the dust source, carbonate contents of loess range from 5% to >30%. Primary loess free from carbonate is rarely reported, e.g., from New Zealand (Muhs and Bettis, 2003).

Loess has a coherent structure caused by the aggregation of silt particles by clay and/or carbonate bridges. At outcrops, loess can form steep cliffs of more than 10 m in height. The massive structure is interrupted by numerous pores (mainly medium to fine), mainly created by former grass root channels that sum to 40–55% of the volume of loess. Bulk density of loess is between 1.5 and 2.0 g cm⁻³. Typically, loess has a light-yellow color, related to the presence of colorless transparent (quartz) or white minerals (feldspar, carbonate) and yellow limonite (iron oxide). Deviations from this general description are related to reworking (leading to a loss of structure) and advanced degrees of pedogenic structuring and pigmentation.

Temperate-zone loess soils

Due to its porosity and mineralogical composition, loess is easily altered into soil by transformation, turbation and translocation processes. Bioturbation and the amount of water moving through initially porous homogenous loess determine the extent of alteration and redistribution of solutes or colloidal particles. Terrestrial soils from loess are instructive examples to illustrate the influence of climate on soil properties. The concept of soil zonality was developed 150 years ago for the soils of Russian loess landscapes, with soil properties largely determined by the climatic conditions and related ecosystems (Paton et al., 1995). Soil development from loess is the subject of numerous soil science textbooks; herein it is referred to temperate-zone soil types after the IUSS Working Group WRB (2015).

Under well-drained conditions in a temperate humid climate (potentially under deciduous forest), bioturbation and humification, carbonate leaching, oxidation, silicate weathering and clay transformation as well as clay translocation lead to the formation of brown-colored soils with clay enriched subsoils (Luvisols; Fig. 3A). Reduced subsoil drainage results in enhanced redoximorphic



Fig. 3 Loess, soils and paleosols. Black and white scales in Figures A–D mark the thickness of aeolian silt. (A) Protostagnic Luvisol from >2 m-thick loess in S Germany. (B) Stagnic Luvisol from 1.35 m-thick loess overlying reworked sand loess in SW Poland. (C) Albic Stagnic Luvisol from c. 1 m-thick loess overlying weathered penultimate glacial till in NE Wisconsin, United States. (D) Albic Dystric Retisol from 0.35 m-thick mantle loam (interpreted as aeolian silt) and penultimate glacial till, NW Russia. (E) Vermic Chernozem from over 2 m-thick loess in SW Russia. (F) Fossil Reductaquic Cryosol (commonly referred to as tundra gleys) in loess profile of SE Poland. (G) 22 m-thick last glacial LPS Toshan (N Iran) with mainly weakly developed paleosols above the last interglacial pedocomplex. (H) Middle- to Lower Pleistocene interglacial paleosols of the >40 m-thick loess-paleosol sequence Stari Slankamen (N Serbia). (B) Photo by J. Waroszewski.

processes, eventually leading to the formation of Stagnic Luvisols (Fig. 3B and C) or Stagnosols. Retisols that form under cooler climate than Luvisols have characteristic pale cracks/tongues that probably relate to past periglacial processes (Fig. 3D). In a less humid, more continental climate (potentially under steppe), carbonate leaching is reduced and weathering limited, thus bioturbation and humification lead to the formation of thick dark horizons, typical of Phaeozems (decalcified) and Chernozems (Fig. 3E). Under even dryer conditions, descending transport is limited or in semi-arid regions is replaced by ascending solute transport; typical soil types are Kastanozem or Solonetz/Solonchak, respectively.

Luvisols, Phaeozems and Chernozems from loess are among the most fertile soils worldwide. While both, sandy and clayey soils provide few advantages and many disadvantages, (silty) loams are the most suitable for growing a wide range of agricultural products (Catt, 2001). The downside of millennia of agricultural production in loess landscapes is long-term soil erosion, facilitated by the considerable erodibility of silt. Its saturation with water can lead to a collapse of loess structure and rather rapid erosion by downslope transport or through piping (Busacca and Sweeney, 2005). The loss of well-structured surface soil also accelerates this problem, as water easily erodes fresh loess with its largely coherent structure. Fertile colluvial and alluvial soils in regions characterized by centuries to millennia of agriculture frequently result from downslope/downstream transport of loess soil.

Dust, soil, (paleo)environments

In (semi-)humid temperate regions, present day loess soils formed after the end of major dust accumulation around 10,000–15,000 years ago. This temporal separation of loess and soil formation is due to profound environmental changes at the transition between the Pleistocene and Holocene, resulting in a transition from morphodynamic activity to stability (Kemp, 2001). Soils of the (semi-) humid temperate zone are residual soils that form from local parent material (e.g., loess) without significant dust admixtures during their formation. The same is likely for soils of the central humid tropics; e.g., Sahara dust reaching the Amazon rainforest is not traceable in soils, but has a positive effect to the nutrient budget of local vegetation. Closer to Holocene dust sources (e.g., the Sahara desert) are soils with detectable quasi-constant aeolian admixtures (e.g., in some Mediterranean or savannah regions). These unusually have large silt contents compared to locally formed parent materials and are good indicators for dust contributions to soils, but additional results from mineralogy and geochemistry are required for reliable conclusions. Accretionary soils form by quasi-constant sediment admixture, but pedogenesis is still strong enough to be imprinted into the substrate. The boundary between accretionary pedogenesis and loess formation is poorly defined (Fig. 2).

While the effect of present-day mineral dust additions to soils and ecosystems is being studied increasingly, the effect of dust admixture to loess parent materials remains poorly recognized. In situ or reworked Pleistocene weathering products outside the mapped loess regions contain variable amounts of windblown dust that may strongly influence pathways of pedogenesis and soil properties (Lorz et al., 2013). While loess formed in temperate zone lowlands during glacial periods, surrounding highlands experienced frost weathering, cryoturbation and solifluction under periglacial conditions. Slope deposits (cover beds) active during the Pleistocene, but stabilized by vegetation and soil formation during the Holocene, contain variable admixtures of mineral dust. These may be difficult to trace if present in small quantities or strongly affected by pedogenesis. At the margins of loess areas are gradual transitions to loess-like deposits and slope deposits with clearly detectable dust admixtures. Landscape position and geological context as well as soil texture, mineralogy and geochemistry help to quantify the extent of dust admixture to parent materials (Fig. 3A–D; Waroszewski et al., 2019).

Well-mapped thick loess deposits frequently show alternations of loess units and buried soil horizons, so called loess-paleosol sequences (LPS). LPS form by the interplay of dust sedimentation and pedogenesis due to changing paleoenvironments (Kemp, 2001). The intensity of paleosol development, usually referred to as degree of chemical weathering, is foremost a function of past moisture availability and the overall duration of pedogenesis; paleosols within LPS indicate past environmental conditions, if the duration of pedogenesis is established (Bronger, 2003). Strongly reduced horizons in temperate zone LPS, commonly referred to as tundra gleys (Fig. 3F), indicate the presence of permafrost that prevented drainage during topsoil thawing in summer (Reductaquic Cryosols). Pale brown to gray horizons indicate weak weathering, typical of interstadials (Fig. 3G), whereas paleosols with a development similar or more advanced than the present day soil typically indicate interglacial periods (Fig. 3H) (Bronger, 2003).

Paleoenvironmental interpretations from paleosols are challenging if there is weak age control. In the last decades new geochronological methods (e.g., optically stimulated luminescence dating) have delivered increasingly robust estimates on the timing of sedimentation and the duration of pedogenesis. Loess paleosols with complete horizonation are rarely preserved. Topsoils may be reworked along slopes or completely lost by erosion during the transition from morphodynamic stability to activity. Careful field observations and micromorphology help to distinguish an in situ paleosol (remnant) from a pedosediment, eventually overprinted by pedogenesis. Polypedogenesis is common in areas with comparably low rates of dust sedimentation, where soil forming processes change without being separated by loess units. Only the combination of sedimentological and paleopedological approaches unravel the full potential of LPS as key archives of terrestrial paleoenvironmental evolution and dust dynamics during the Quaternary (Kemp, 2001). Additionally, loess preserves important Paleolithic open-air sites, allowing the integrated study of human and environmental evolution.

Time scales of soil formation and a potential influence of relict and fossil soils to surface soil properties remain poorly established. Compared to hard rock, loess is comparably easily transformed into soil, leading to the formation of thick steppe or forest soils in a relatively short time (1–2 m in 10,000 years). In landscapes with thick loess covers, fertile loams may appear to stakeholders as an endless resource. It is important to raise awareness about the processes and time scales of soil (neo)formation to

promote sustainable use of soil. To quantify rates of soil production, knowledge on loess formation, dust contributions to mapped lithologies or current dust admixtures to soils is essential. It is important to note that deep, fertile temperate zone soils largely result from past morphodynamics and long-term pedogenesis and may not be restored quickly under present day environments. Thin loess mantles may be quickly lost by erosion, leaving less favorable geological substrate behind.

Conclusions

Loess is among the most widespread terrestrial deposits and parent material of fertile soils. Due to its complex formation and widespread occurrence with variable properties, there exists no universally accepted definition of loess. A simple definition as a distinct body of mainly windblown silt is in most cases sufficient. Aeolian deposition of mineral dust explains the silty texture, while the coherent structure is related to weak in situ cementation. Loess is an easily erodible and weatherable substrate and therefore frequently reworked and overprinted by pedogenesis, which makes the mapping of loess sediments challenging. Thin loess mantles remain frequently overlooked, despite their usually positive influence on soil properties. Thick loess deposits and intercalated paleosols are important archives of quaternary climate, landscape and human evolution.

Acknowledgments

The thoughtful comments of Ian Smalley, Jarosław Waroszewski, and the editors Margaret A Oliver and Rupert Bäumler on earlier versions of this manuscript are gratefully acknowledged.

References

- Bertran P, Liard M, Sitzia L, and Tissoux H (2016) A map of Pleistocene aeolian deposits in Western Europe, with special emphasis on France. *Journal of Quaternary Science* 31(8): 844–856.
- Bronger A (2003) Correlation of loess-paleosol sequences in East and Central Asia with SE Central Europe: Towards a continental Quaternary pedostratigraphy and paleoclimatic history. *Quaternary International* 106: 11–31.
- Busacca AJ and Sweeney MR (2005) Loess. In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*, pp. 364–373. Oxford: Elsevier.
- Catt JA (2001) The agricultural importance of loess. *Earth-Science Reviews* 54(1–3): 213–229.
- Ehlers J, Gibbard PL, and Hughes PD (eds.) (2011) *Quaternary Glaciations—Extent and Chronology—A Closer Look*. Amsterdam: Elsevier.
- IUSS Working Group WRB (2015) *World Reference Base for Soil Resources 2014, Update 2015. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps, World Soil Resources Reports*. Roma: FAO.
- Kemp RA (2001) Pedogenic modification of loess: Significance for palaeoclimatic reconstructions. *Earth-Science Reviews* 54(1–3): 145–156.
- Koch R and Neumeister H (2005) Zur Klassifikation von Lösssedimenten nach genetischen Kriterien. *Zeitschrift für Geomorphologie* 49(2): 183–203.
- Lancaster N (2020) On the formation of desert loess. *Quaternary Research* 96: 105–122.
- Lehmkuhl F, Nett JJ, Pötter S, et al. (2020) Loess landscapes of Europe—Mapping, geomorphology, and zonal differentiation. *Earth-Science Reviews* 215: 103496.
- Leonhard KCV (1824) *Charakteristik der Felsarten—Dritte Abteilung. Trümmer-Gesteine. Lose Gesteine. Kohlen*. Heidelberg: Joseph Engelmann.
- Li Y, Shi W, Aydin A, Beroya-Eitner MA, and Gao G (2020) Loess genesis and worldwide distribution. *Earth-Science Reviews* 201: 102947.
- Lorz C, Frühauf M, Mailänder R, Phillips JD, and Kleber A (2013) Chapter 3—Influence of cover beds on soils. In: Kleber A and Terhorst B (eds.) *Developments in Sedimentology*, pp. 95–125. Amsterdam: Elsevier.
- Morrás HJM, Moretti L, Piccolo G, and Zech W (2009) Genesis of subtropical soils with stony horizons in NE Argentina: Autochthony and polygenesis. *Quaternary International* 196(1–2): 137–159.
- Muhs DR and Bettis EA III (2003) Quaternary loess-Paleosol sequences as examples of climate-driven sedimentary extremes. *Geological Society of America, Special Paper* 370: 53–74.
- Paton TR, Humphreys GS, and Mitchell PB (1995) *Soils: A New Global View*. New Haven CT: Yale University Press.
- Pécsi M (1990) Loess is not just the accumulation of dust. *Quaternary International* 7–8: 1–21.
- Pye K (1995) The nature, origin and accumulation of loess. *Quaternary Science Reviews* 14(7–8): 653–667.
- Sedov S, Solleiro-Rebolledo E, Terhorst B, et al. (2009) The Tlaxcala basin paleosol sequence: A multiscale proxy of middle to late Quaternary environmental change in central Mexico. *Revista Mexicana De Ciencias Geológicas* 26: 448–465.
- Smalley IJ (1995) Making the material: The formation of silt-sized primary mineral particles for loess deposits. *Quaternary Science Reviews* 14(7–8): 645–651.
- Smalley IJ, O'Hara-Dhand K, Wint J, et al. (2009) Rivers and loess: The significance of long river transportation in the complex event-sequence approach to loess deposit formation. *Quaternary International* 198: 7–18.
- Sprafke T and Obrecht I (2016) Loess: Rock, sediment or soil—What is missing for its definition? *Quaternary International* 399: 198–207.
- Taylor SR, McLennan SM, and McCulloch MT (1983) Geochemistry of loess, continental crustal composition and crustal model ages. *Geochimica et Cosmochimica Acta* 47: 1897–1905.
- Vandenbergh J, Sun Y, Wang X, Abels HA, and Liu X (2018) Grain-size characterization of reworked fine-grained aeolian deposits. *Earth-Science Reviews* 177: 43–52.
- Waroszewski J, Sprafke T, Kabala C, et al. (2019) Tracking textural, mineralogical and geochemical signatures in soils developed from basalt-derived materials covered with loess sediments (SW Poland). *Geoderma* 337: 983–997.
- Wright JS (2001) "Desert" loess versus "glacial" loess: quartz silt formation, source areas and sediment pathways in the formation of loess deposits. *Geomorphology* 36(3–4): 231–256.

Time

Susan E Melzer, Caroline M Yonker, and Eugene F Kelly, Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of E.F. Kelly, C.M. Yonker, FACTORS OF SOIL FORMATION I Time, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 536-539, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00015-1>.

Introduction	49
Key soil properties that vary with time	49
Studies of time and soil formation	51
Conclusion	52
References	52

Abstract

The environmental conditions under which soils and their properties develop are governed by combinations of the climate, biota, geology, and topography, which operate over time. Time affects soil-forming processes by controlling their duration and degree of soil development. Thus, we can differentiate between ‘young’ and ‘old’ soil in each ecosystem based on degrees of chemical and physical weathering and the resulting expression of soil properties. The amounts and distribution of biotic and abiotic soil constituents are often used to determine the relative age of a soil and differences in their properties can be explained as a function of time.

Glossary

Calcium carbonate Carbonate mineral that forms coatings on soil that react with an acid to form carbon dioxide gas.

Evaporation Water on the Earth’s surface or in the soil absorbs heat from the sun to the point that it vaporizes or evaporates and becomes part of the atmosphere.

Horizon An individual layer within the soil which has its own unique characteristics (such as color, structure, texture, or other properties) that make it different from the other layers in the soil profile.

In situ Latin for ‘the original position’.

Soil horizons An identifiable soil unit due to color, structure, or texture.

Soil profile The ‘face’ of a soil when it has been cut vertically that shows the individual horizons and soil properties with depth.

Soil texture The way soil ‘feels’ when it is squeezed between the fingers or in the hand. The texture depends on the amount of sand, silt, and clay in the sample (particle-size distribution), as well as other factors.

Translocation The movement or transfer of material from one part of the soil profile to another.

Transpiration Water in plants escapes or transpired into the atmosphere as the leaf stomata open to exchange carbon for oxygen.

Key points

- The environmental conditions under which soils and their properties develop are governed by unique combinations of the climate, biota, geology, and topography, which operate over time.
- Time affects soil forming processes by controlling their duration and degree of soil development.
- Relative soil age may be determined by the extent of chemical and physical weathering and the resulting expression of soil properties.
- The amounts and distribution of soil organic carbon, carbonate, and various other elements, and the amount and distribution of clay and its mineralogy are often used to determine the relative age of a soil.
- Soil chronosequences provide an opportunity to determine relative ages of landscapes that are 100,000 to 1,000,000 years old and a chronologic framework to contextualize numerical age by use of absolute dating techniques.

Introduction

Thousands of soil types at the Earth's surface are composed of unique biological, chemical, and physical properties. These properties are the direct result of a complex set of soil-forming processes that are conditioned by distinctive environmental factors. The environmental conditions under which soils and their properties develop over time are governed by combinations of climate, biota, geology, and topography (Jenny, 1941).

Climate affects soil development, as moisture and temperature control the rate of biological, chemical, and physical processes. The effect of biota is primarily related to controls on inputs and outputs of organic and inorganic materials; the effect of topography relates to the changes in temperature and precipitation due to differences in elevation and slope angles. The chemical and mineralogical composition of the parent material influences such key soil properties as nutrient status, alkalinity or acidity, and texture (Hillel, 2008; Jenny, 1980).

Climate, biota, topography, and parent material have a direct effect on soil-forming processes, whereas time has an indirect effect. Time affects soil-forming processes by controlling duration and, therefore, the degree of soil development (Yaalon, 1983). We can differentiate between 'young' and 'old' soil in each ecosystem on the basis of degrees of chemical and physical weathering and the resulting expression of soil properties.

Key soil properties that vary with time

The amount and distribution of soil organic carbon (SOC), carbonate, and other key elements, and the amount and distribution of clay and corresponding mineralogy change as a function of time in a generally consistent and predictable manner (Koop et al., 2019) (Fig. 1). The amounts and distribution of these soil constituents are often used to determine the relative age of a soil.

Once vegetation is established in an ecosystem, biotically-related properties such as SOC are continually added through both above- and below-ground inputs of living and dead plant tissues. Older soils are expected to have a greater accumulation of SOC than younger soils (Gile, 1970). The accumulation of SOC eventually reaches a steady state where inputs equal outputs, and there is no further net accumulation with increasing soil age (time) (Lawrence et al., 2015; Wang and Amundson, 1996; Merritt et al., 1991). The length of time needed for SOC to reach a steady state varies among environments but is within a relatively short time frame of 10^2 – 10^3 years (Buol et al., 1997).

The accumulation of abiotic properties such as calcium carbonate (CaCO_3) in soils over time is generally limited to soils of semiarid and arid regions, where precipitation of carbonates occurs under conditions typical of dry soil environments, high pH, and high evapotranspiration. The increased production of carbonic acid (H_2CO_3) under humid conditions results in the dissolution of CaCO_3 and a decrease in CaCO_3 over time (Lal et al., 2021). Sources of calcium in arid and semiarid soils may be from in situ weathering of calcium-bearing minerals or atmospheric deposition. Unlike SOC, CaCO_3 accumulation may not reach a steady state, but continues to increase through time (Wang et al., 2016; Monger and Rachal, 2013).

The most abundant elements in soils, excluding oxygen and carbon, are silicon, aluminium, iron, magnesium, calcium, sodium, and potassium. Although these elements have varying solubilities, their relative abundance is a direct function of the degree of soil weathering (Dixon and Weed, 1989), which increases through time and soil development (young, intermediate, old). In general, the simple soluble cations, magnesium, calcium, sodium, and potassium, are readily released from the soil through weathering processes and are easily leached (Fig. 2a). Unless there is an input of these elements to the soil system from the atmosphere or some

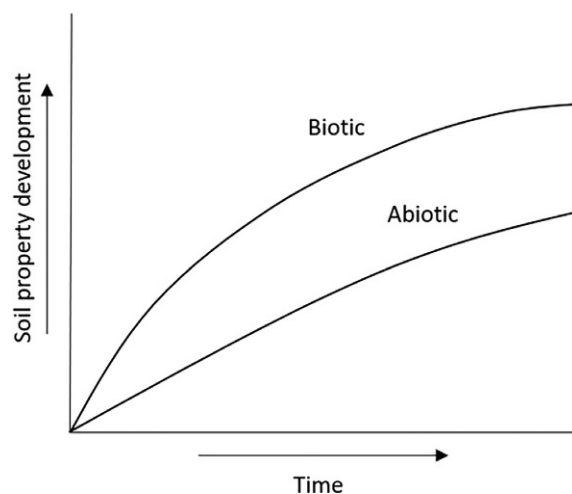


Fig. 1 The effect of time on the development of biotic (e.g., soil organic carbon) and abiotic (e.g., calcium carbonate) soil properties.

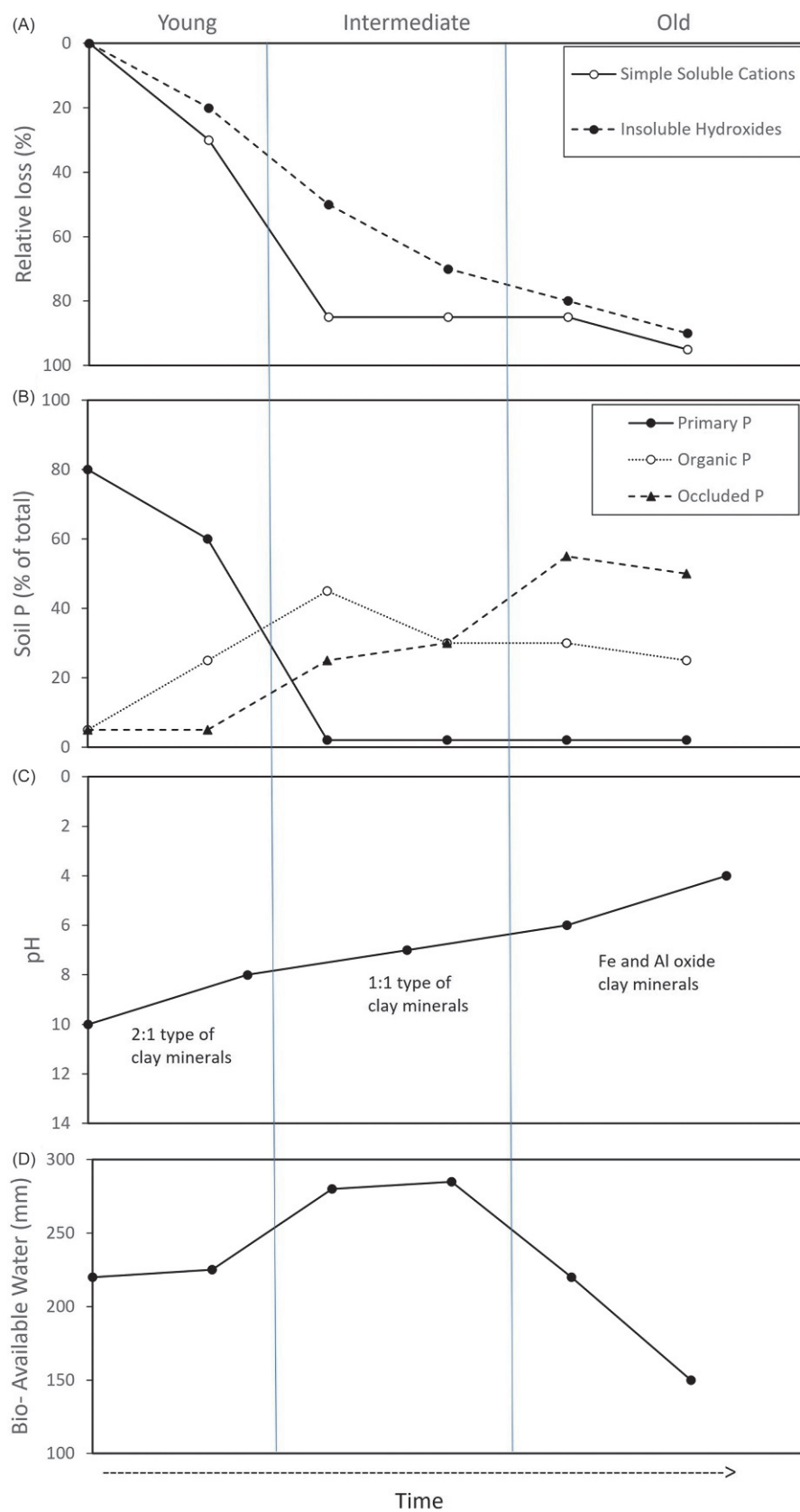


Fig. 2 The effect of time on soil development: (a) the loss of chemical soil constituents, (b) the transformation of phosphorous, (c) pH and clay mineralogy, (d) and biologically available water.

other mechanism, their quantities steadily decrease as a function of time (Chadwick et al., 2022). Soil silicon, aluminium, and iron exist primarily as relatively insoluble hydroxides and are leached at a slower rate than the simple soluble cations (Fig. 2a).

Phosphorus is an element which exists in soils in a variety of organic and inorganic forms, the distribution of which indicates the relative weathering of the soil profile, and therefore, soil age. As soil development begins and the soil matures, primary mineral sources are depleted, and organic and secondary occluded mineral forms of phosphorus accumulate (Turner and Laliberté, 2015; Crews et al., 1995; Walker and Syers, 1976) (Fig. 2b).

The accumulation and distribution of soil clay minerals are indicators of soil age (Arduino et al., 1986). Clays accumulate through time as a function of the atmospheric deposition of dust, translocation from superjacent soil horizons, and in situ formation of secondary clay minerals. Layers of accumulated clay can form in as little as a few hundred years if there are significant atmospheric inputs in the form of dust or mud rains. Rates of accumulation are largely a function of translocation and weathering in areas of decreased atmospheric input. Clay-enriched horizons can develop within 10^3 – 10^4 years, depending on environmental gradients from humid to arid.

Clay mineralogy is partly a function of soil age because, over time, soils become increasingly depleted in soluble framework cations and silicon, major constituents of clay minerals. Clay minerals, therefore, often form in a recognized sequence. The youngest stages of soil development commonly consist of 2:1 type clay minerals (i.e., illites and/or chlorites) that weather early in the sequence to other 2:1 type clays (i.e., smectites and/or vermiculites). Weathering of 2:1 clay minerals to 1:1 clay minerals (e.g., kaolinite) occurs in the intermediate to late sequence of soil development. Iron and aluminium oxide clay minerals result in the oldest soils that have undergone the longest time of soil formation (Arduino et al., 1986) (Fig. 2c).

Biologically available water is optimized during the intermediate stage of soil development (Salley, 2015) (Fig. 2d) where the accumulation of clay and organic compounds enhances aggregation. Aggregates provide the foundation for soil structure which influences productivity and organic matter additions through its effect on aeration and root proliferation. While the weathering process over time results in clay accumulation, water is held too tightly in the micropores of clay for plants to access it (Easton and Bock, 2016). Thus, water available for transport and plant uptake depends on the macropores created during aggregation where water is more likely to infiltrate (Beven and Germann, 2013). Loamy soils have the greatest biologically available water because they drain excess water quickly and store it in medium- and small-sized pores and in organic matter (Easton and Bock, 2016). Over time, the soil will accumulate clay and organic matter, and develop structure (i.e., aggregation) to enhance porosity that supports water infiltration, storage, and bioavailability.

The preceding discussion theoretically describes the effect of time on soil development when climate is a constant variable. Given the same duration of soil formation, a soil forming in a warm, humid environment will have soil properties that are significantly more developed, and which appear to be older than a soil forming in a cool, dry environment. Increasing precipitation causes increased chemical and physical weathering, effectively mimicking the effect of a long period of soil formation in a relatively short time period. In humid areas, where precipitation exceeds potential evapotranspiration, biological and chemical weathering processes rapidly transform organic and mineral matter. Leaching transfers materials down, and eventually through, the soil profile. The same soil-forming processes operate in drier areas where precipitation is less than potential evapotranspiration; however, the availability of water limits the intensity and duration of these processes (Birkeland, 1999).

Studies of time and soil formation

The differences in properties among soils can be explained as a function of time if those soils are forming in the same climate, maintain the same type of vegetation, have developed from the same or similar parent materials, and are located at the same topographic position. In other words, time is the only factor that varies between the two soils being compared. Soil properties, developed as a function of time, are used as a proxy of age (Markewich et al., 2017). Climate, biota, parent material, and relief can be held constant under a limited number of circumstances, including the comparison of soils developing on stream terraces sequentially incised by a single river system (Haan, 2014). In this case, the soils of the floodplain are developing in the most recently deposited material and are the youngest; the soils on the highest stream terrace relative to the floodplain are the oldest (Loadholt, 2002) (Fig. 3). Such a series of related terrace soils, which differ only with respect to their duration of formation, are called a chronosequence (Stevens and Walker, 1970). Soils developing on chronosequences show broadly similar trends in morphology with time, including: (1) a thickening zone of organic matter accumulation; (2) increasing pedon thickness; (3) increasing horizon complexity; and (4) development of zones of secondary mineral accumulation. Trends 1–4 continue until the soil begins to degrade due to weathering. Many soils will not follow this predictable model which correlates dated terraces to the development of soil forming processes and properties over time; instead, unpredictable geomorphic processes result in non-linear soil development that does not correlate to absolute age (Huggett, 1998; Haan, 2014).

Soil chronosequences determine relative ages of landscapes that are 100,000–1,000,000 years old and also provide a chronologic framework to contextualize numerical age (by use of absolute dating techniques) (Markewich et al., 2017). Greater relevance to the chronosequence concept is provided where material for absolute dating techniques is not available. Ideally, absolute dating of the parent material anchors (to a maximum soil age) the relative age assessment determined by soil development indices. Absolute dating methods include optical luminescence dating (Markewich et al., 2017) and radiogenic techniques, such as radiocarbon analysis (soils <40 kyr) and U-series dating (soil >40 kyr). The latter methods require the availability of stable soil organic matter and pedogenic calcium carbonate, respectively. Other methods include dating by biostratigraphic age and amino acid racemization which, according to Wehmiller and Miller (2000), can be used as a semi-quantitative tool to estimate the relative age of molluscs in

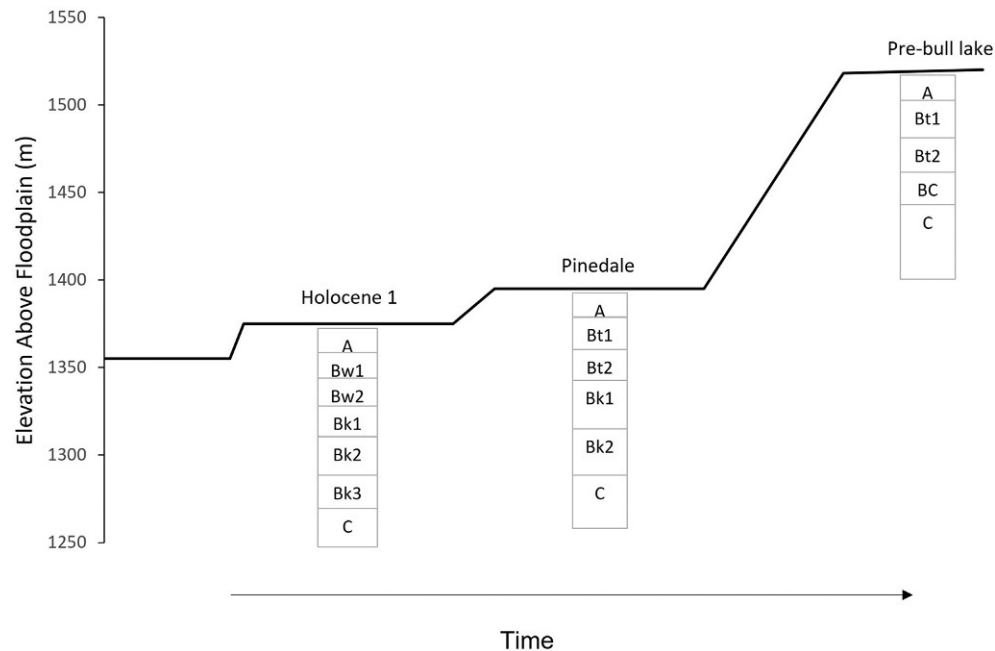


Fig. 3 Soil chronosequence developing in a semiarid ecosystem. The youngest soil is at the lowest elevation above the floodplain; the oldest is at the highest elevation. The soil ages are, Holocene (<10,000 years), Pinedale (13000–600,000 years), Pre-Bull Lake (300,000–600,000 years).

marine terrace deposits (Haan, 2014). Meteoric cosmogenic ^{10}Be has the potential to record soil residence time and numeric soil age (soils >40 kyr) when no other chronometer is available; this method estimates the minimum required time for a soil to form (Markewich et al., 2017; Willenbring and von Blanckenburg, 2010).

Conclusion

Chronosequences are problematic in that, over millennial time scales, a constant climate cannot be assumed. Numeric soil age, or duration of soil development, can also be difficult to establish. Despite these obstacles, chronosequences provide an excellent opportunity to study the influence of time on soil formation. The combination of relative and absolute dating techniques is promising for the reconstruction of landscape history.

References

- Arduino E, Barberis E, and Ajmone Marsan F (1986) Iron oxides and clay minerals within profiles as indicators of soil age in northern Italy. *Geoderma* 37: 45–55.
- Beven K and Germann P (2013) Macropores and water flow in soils revisited. *Water Resources Research* 49: 6.
- Birkeland PW (1999) *Soils and Geomorphology*, 3rd edn. New York: Oxford University Press.
- Buol SW, Hole FD, McCracken RJ, and Southard RJ (1997) *Soil Genesis and Classification*, 4th edn. Ames, IA: Iowa State University Press.
- Chadwick OA, Chorover J, Chadwick KD, Bateman JB, Slessarev EW, Kramer M, Thompson A, and Vitousek PM (2022) Constraints of climate and age on soil development in Hawaii. In: *Biogeochemistry of the Critical Zone*. Springer.
- Crews TE, Kitayama K, Fownes JH, Riley RH, Herbert DA, Mueller-Dombois D, and Vitousek PM (1995) Changes in soil phosphorus fractions and ecosystem dynamics across a long chronosequence in Hawaii. *Ecology* 76: 1407–1424.
- Dixon JB and Weed SB (eds.) (1989) *Minerals in the Soil Environment*. 2nd edn In: *Book Series No. 1*, 2nd edn. Madison, WI: Soil Science Society of America.
- Easton ZM and Bock E (2016) Soil and soil water relationships. In: *Virginia Cooperative Extension*. Virginia Tech/Virginia State University. Publication BSE-194-P.
- Gile LH (1970) Soils of the Rio Grande valley border in southern New Mexico. *Soil Science Society of America Proceedings* 34: 465–472.
- Haan M (2014) *Evaluating Different Dating Methods on Middle Pleistocene to Holocene Marine Sediments in Southern Italy for Potentials in Soil Chronosequences*. Research Gate.
- Hillel D (2008) *Soil in the Environment*. New York, NY: Goddard Institute for Space Studies, Columbia University, 320 pp.
- Huggett RJ (1998) Soil chronosequences, soil development, and soil evolution: A critical review. *Catena* 32: 155–172.
- Jenny H (1941) *Factors of Soil Formation: A System of Quantitative Pedology*. New York: McGraw-Hill, 281 p.
- Jenny H (1980) *The Soil Resource*. New York: Springer-Verlag.
- Koop AN, Hirmas DR, Sullivan PL, and Mohammed AK (2019) A generalizable index of soil development. *Geoderma* 360: 113898.
- Lal R, Monger C, Nave L, and Smith P (2021) The role of soil in regulation of climate. *Philosophical Transactions of the Royal Society B* 376: 20210084.
- Lawrence CR, Harden JW, Xu X, Schulz MS, and Trumbore SE (2015) Long-term controls on soil organic carbon with depth and time: A case study from the Cowlitz River Chronosequence, WA USA. *Geoderma* 247–248: 73–87.
- Loadholt SW (2002) *Soil Physiography of a Terrace Chronosequence in the Pawnee National Grasslands, Colorado*. PhD Dissertation Colorado State University, 198 p.
- Markewich HW, Pavich MJ, and Wysocki DA (2017) Soils as relative- age dating tools. In: Richardson D, Castree N, Goodchild MF, Kobayashi A, Liu W, and Marston RA (eds.) *The International Encyclopedia of Geography*. John Wiley & Sons, Ltd.

- Merrittz DJ, Chadwick OA, and Hendricks DM (1991) Rates and processes of soil evolution on uplifted marine terraces, Northern California. *Geoderma* 51: 241–275.
- Monger HC and Rachal DM (2013) Soil and landscape memory of climate change: How sensitive, how connected? *Society of Sedimentary Geology Special Publication*. vol. 104, pp. 63–70. SEPM Special Publications.
- Salley SW (2015) *A Study of Long-Term Soil Moisture Dynamics*. Ph.D. Dissertation Colorado State University, 111 p.
- Stevens PR and Walker TW (1970) The chronosequence concept and soil formation. *Quarterly Review of Biology* 45: 333–350.
- Turner BL and Laliberté E (2015) Soil development and nutrient availability along a 2-million-year coastal dune chronosequence under species-rich Mediterranean shrubland in southwestern Australia. *Ecosystems* 18: 287–309.
- Walker TW and Syers JK (1976) The fate of phosphorus during pedogenesis. *Geoderma* 15: 1–19.
- Wang Y and Amundson R (1996) Radioactive dating of soil organic matter. *Quaternary Research* 45: 282–288.
- Wang J, Monger C, Wang X, and Leinauer MSB (2016) Carbon sequestration in response to grassland- shrubland- turfgrass conversions and a test for carbonate biomineralization in desert soils, New Mexico, USA. *Soil Science Society of America Journal* 80: 1591–1603.
- Wehmiller JF and Miller H (2000) Aminostratigraphic dating methods in quaternary geochronology. In: Stratton Noller J, Sowers JM, and Lettis WR (eds.) *Quaternary Geochronology: Methods and Applications*, pp. 187–222. Washington DC: The American Geophysical Union.
- Willenbring JK and von Blanckenburg F (2010) Meteoric cosmogenic Beryllium-10 adsorbed to river sediment and soil: Applications for Earth-surface dynamics. *Earth Science Reviews* 98(1–2): 105–122.
- Yaalon DH (1983) Climate, time and soil development. In: Wilding LP, Smeck NE, and Hall GF (eds.) *Pedogenesis and Soil Taxonomy: I. Concepts and Interactions*, pp. 233–251. New York: Elsevier.

Factors of soil formation—Human impacts

Jonathan A Sandor^a, C Lee Burras^a, Michael Thompson^a, and Skye A Wills^b, ^aIowa State University, Ames, IA, United States;

^bNatural Resources Conservation Service, Lincoln, NE, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	55
Scales and scope of human impacts on soil formation	55
Evaluating human impacts on soil formation	56
Changes in soil properties and processes resulting from human activities	58
Agriculture	58
Cities and industry	60
Mining	63
War	66
Climate change and soil formation	67
Significance and future of human impacts on soil formation	67
References	68

Abstract

Soils form the dynamic foundation for terrestrial life on Earth. In contrast to the slow rate of natural soil formation in diverse landscape settings, soil change by humans is often more rapid and far-reaching. Land use changes soil properties and processes both directly and indirectly through agriculture, urbanization, industry, mining, war, and climate change. People can enhance soils for uses such as agriculture and engineering, but their actions often degrade soil health, leading to lower productivity and ecosystem function. Knowledge about anthropogenic transformations of soil is essential to foster soil conservation, remediate degraded soil, and promote sustainable land use.

Key points

- Objective: to introduce readers to anthropogenic soil change and its importance.
- Humans have far-reaching impacts on soil resources through agriculture, urbanization, industry, mining, war, and other land use.
- People also change soils indirectly by altering the factors of soil formation, such as through climate change.
- Knowledge about anthropogenic soil transformation is essential in efforts toward soil conservation, fostering soil health and soil security, and sustainable land use.

Technical Nomenclature

A horizon	Surface mineral horizon with some organic matter accumulation
Argillic horizon	Subsurface diagnostic horizon with significant clay accumulation
B horizon	Subsurface mineral soil horizon (several kinds)
C horizon	Relatively unweathered, unconsolidated mineral horizon or layer (or soft bedrock), usually beneath the active zone of soil formation
Fragipan	Dense, brittle subsurface horizon
Irragic horizon or soil	Anthropogenic horizon formed by long-term irrigation and accompanying additions of sediment
Paleosol	Soil formed in a past geologic period
Petroplinthite	Hardened or cemented horizon of highly weathered, iron-rich soil
Plaggen horizon or soil	Thick, organic-matter rich anthropogenic horizon formed during long-term cultivation by additions of manure, sod, and other materials.
Spodic horizon	Refer to Spodosol below

The following terms are nine out of the 12 soil orders of the US soil classification system (Soil Taxonomy)

Alfisol	Mineral soil that is relatively low in organic matter, relatively high in base cations, and with subsurface clay accumulation
Entisol	Undeveloped mineral soil
Gelisol	Mineral or organic soil of cold regions, with permafrost
Histosol	Organic soil
Inceptisol	Mineral soil with initial, but relatively weak development
Mollisol	Mineral soil with thick, dark, organic matter-rich surface horizon, high in base cations
Oxisol	Intensively weathered mineral soil, typically developed on old land surfaces in intertropical regions and distinguished by residual accumulations of kaolinite, iron oxides, and aluminum oxides.
Spodosol	Mineral soil with subsurface accumulation of organic matter, aluminum, and commonly iron
Ultisol	Highly weathered mineral soil that is relatively low in organic matter, relatively low in base cations, and with subsurface clay accumulation

Introduction

Humans, like many other organisms, are active agents of soil formation. Because soils comprise the dynamic, vibrant skin of the Earth's terrestrial surface, people have always interacted with, and therefore changed, soils and the course of their formation. While soils are subject to major change and even destruction by natural forces on the scale of geologic time, changes resulting from human activity usually occur on a much shorter time scale. Human impact on the Earth's land, water, atmosphere, and life has become so pervasive that the current geologic age is recognized as the Anthropocene (Brown et al., 2017; Richter, 2020).

People have had an impact on the soil in a multitude of ways and extents, through farming, building, mining, war, and recently climate change (Capra et al., 2015; Dazzi and Lo Papa, 2015; Howard, 2017). In some cases, human activities enhance soils for particular uses such as agriculture and engineering. However, in other cases, the interplay between humans and soils has resulted in degradation of soil health, leading to lower productivity and environmental function. Recognizing that soil use is literally and figuratively the ecological basis for civilizations and that soil resources are essentially nonrenewable on the human time scale, understanding soil formation and its transformation by humans is imperative for developing agricultural and natural resource sustainability, and for protecting environmental quality.

Scales and scope of human impacts on soil formation

To understand the effects of humans on soil, it is helpful first to provide a frame of reference by considering soil-formation processes and their rates in natural environments. From the perspective of pedology, the study of soil formation, soil is a complex assemblage of mineral and organic materials formed at the Earth's surface. In spite of the increasing impact of human activities on soil and the likelihood that all of the Earth's ecosystems have been influenced to some degree by humans, many soils still retain their basic morphology imparted by natural processes and environmental factors. A hallmark of soil formation is the differentiation of horizons (layers). Soils are developed and organized into horizons by complex, interrelated physical, chemical, and biological processes that are determined by the factors of soil formation: climate, organisms, geology, topography, and time. The morphology of every soil is the expression of those fundamental processes interacting with one another over multiple spatial scales (from micrometers to kilometers) and temporal scales (from days to millennia). Horizons at the surface (A horizons) are often enriched with organic matter, while deeper horizons (B horizons) may have accumulations of clay minerals, calcium carbonate, metal-organic complexes, or other materials. Formation of A horizons is relatively rapid because organic matter accumulates in a few centuries to a millennium. Formation of B horizons commonly takes many thousands of years to become fully expressed. In this sense, soils can range from young to old. As soil formation progresses, soils generally tend to become increasingly anisotropic as they differentiate into greater kinds and numbers of horizons.

Landscape stability is a prerequisite for the development of soil horizons to proceed. Even so, disturbance and change are integral to the functions of all natural ecosystems and their soils. Soils, landscapes, and associated biological communities are changed by disturbances ranging from minor perturbations such as low-intensity fires to major events such as volcanism, as well as long-term climatic and environmental change. On the geologic time scale, soils are subject to major alteration, destruction, and renewal. In contrast to the relatively slow pace of natural soil development, soil changes resulting from human activity are often more rapid and far-reaching. Human-caused change may be so fast and irreversible that the affected soil bears little resemblance to its original form, such as in cases of severe erosion or soil excavation and filling in urban engineering. Human impacts often reverse the anisotropic trend of soil formation, making soils simpler, less organized, and more homogeneous.

A wide array of land use and other human activities have altered the paths of formation in many soils (Table 1). Corresponding soil changes also vary greatly in kind, intensity, time and spatial scale, and significance for soil and environmental quality

Table 1 Human agents of change in soil properties and processes.

<i>Human agents of soil change</i>	<i>Examples</i>
Agriculture and forestry	
• Crop and animal agriculture	Tillage, fertilization, cropping, flooding, irrigation, drainage, terracing, grazing
• Forestry	Altered vegetation type and cover, harvesting operations
• Off-site effects	Erosion and sedimentation, causing soil contamination and burial
Cities and industry; buildings, roads, and other structures	Excavation, urban cover, artificial fills, land leveling, paving, pollution
	Dams and reservoirs, polders, dikes, mounds, other artificial landforms and fills
Mining and other earth and water resource extraction	Soil removal for pits and quarries, erosion and sedimentation associated with excavation, mixing and inversion of earth materials, acid drainage, land disturbance to build pipelines and well pads, groundwater pumping, reclamation
War and military operations	Bomb craters, military transport and engineering, tunneling, contamination by chemicals and radioactive elements
Human-linked environmental change	
• Global climate change	Increased atmospheric CO ₂ and other greenhouse gases, climate warming, increased climate variability and extremes
• Desertification	Vegetation loss or change, accelerated wind and water erosion, salinization
• Acid rain	Soil acidification and other chemical changes

(Tables 2 and 3). Human actions that change soil may act directly or indirectly by changing both soil morphology and underlying soil-forming processes. They may be intentional changes based on land-management strategies to increase soil productivity, or they may be rearrangements of landscapes and soil materials to construct buildings, roads, and other structures. Conversely, they may be unintended changes that can lead to degradation such as soil erosion and burial under sediment. Indirect soil change may result from off-site processes physically remote from the affected soil, such as subsidence following groundwater pumping or downstream sedimentation. In land uses such as agriculture, surface horizons are usually more altered than subsurface horizons, while engineering activities often change the entire sequence and character of natural soil horizons. Engineering activities lead to some of the most intensive effects by removing soils entirely and replacing them with different earth materials, as well as by reshaping landforms or constructing new ones. Agriculture and engineering merge in management practices such as in paddy rice production and terracing of slopes, which also involve major geomorphic and soil change.

Some human activities such as agriculture have influenced soil processes and morphology for thousands of years (Macphail and Goldberg, 2018; Winiwarter, 2014), while others such as global climate change have come to the forefront recently (Horwath and Kuzyakov, 2018). The impacts of anthropogenic climate change on soil formation and distribution are large because climate is a major determinant of soil formation over time, acting through multiple pathways. Some anthropogenic impacts on the environment, such as acid rain and other pollutants, lead to less visible chemical and biological changes, while processes such as desertification result in more sweeping changes through loss of vegetative cover, desiccation, wind and water erosion, and salt accumulation.

In their response to human actions, soils differ in their resistance to change and resilience, that is, their ability to rebound toward their original state. Processes of human-caused soil change may be reversible or irreversible, and the resulting changes may be ephemeral to permanent (Johnson and Lewis, 2007). Response depends on both external factors, such as the type of impact and environmental conditions, and on internal soil properties like texture and mineralogy. For example, soils with uniform textures or that are rich in organic matter tend to be resistant to compaction. After many thousands of years of natural formation, some subsurface soil horizons disrupted by deep plowing may return to their original condition in a few years (e.g., clay-rich argillic horizons), whereas more indurated horizons such as silica-cemented duripans remain fragmented for longer.

Evaluating human impacts on soil formation

Human-caused soil change has been studied by pedologists in several ways (Dudal et al., 2002; Wills et al., 2017). Hans Jenny and others framed their views in the context of the factors of soil formation, placing humans within the biotic factor, while recognizing the unique role of human culture as distinct from other organisms (Amundson and Jenny, 1991). Jenny showed that human land use often alters soils through changes in the other soil-forming factors. For example, irrigation alters arid land soils by changing the climate factor, effectively increasing precipitation (Bidwell and Hole, 1965). Other researchers set humans apart from the natural factors of soil formation to emphasize the extraordinary scale and rate of anthropogenic effects on soils (Yaalon and Yaron, 1966). In working with drastically altered soils such as mine soils and urban soils, pedologists have defined new soil taxonomic classes, because these soils bear little or no resemblance to their original state (Soil Survey Staff, 2014).

How is soil change detected and measured? While large impacts on soil may be obvious, such as wholesale change or destruction of soils and landscapes in mining, terrace building, or urban development, it is still difficult to quantify changes to soil forming factors that may be longer-term. Soil changes due to land use conversion or management practices are often more subtle and

Table 2 Scope of anthropogenic soil change and response. Bracketed terms illustrate range end members.

Soil Change Characteristics and Range	Examples
Causes of soil change may be	
- Direct - Indirect	Compaction Slower or diverted water movement and reduced soil water storage after compaction
- Deliberate - Unintended	Fertilization, liming, irrigation Nutrient depletion or imbalance; salinization
- Constructive - Destructive	Plaggen soil; terraced soil construction Accelerated erosion, sulfide oxidation with drainage or exposure, salt/sodium buildup
Magnitude and extent	
- Low impact - High impact	Conservation tillage; minor soil change from low-intensity fires Urban expansion; land filling
- Soil property or horizons - Whole soil effect	Thinning of A horizon by erosion Removal of entire soil by intense erosion
Duration and rate	
- Short-term/ephemeral/reversible - Long-term/permanent/irreversible	Alleviation of plow pans by freeze-thaw; liming Urban soil; mine soil; Laterite/plinthite hardening with exposure
- Slow rate - Fast rate	Agric horizon development Oxidation following drainage
Response of soil to human impact	
- Susceptible - Resistant	Low-pH soil susceptible to acid rain Neutral to high-pH soil buffered against acid rain
Change outcomes for soil health and use	
- Beneficial - Neutral/benign - Degradation (of productivity and environment)	Organic matter addition Fertilization that balances crop removal of nutrients A horizon erosion

Table 3 Spatial scales of soil change.

Soil components	Approximate spatial scale (m)	Examples of impacts
Physical, chemical, and biological components	10^{-10} – 10^{-4}	Retention of pollutants such as heavy metals, pesticides, industrial solvents, and human and livestock drugs
Morphological properties	10^{-3} – 10^{-2}	Soil structure degradation; changes in texture, color, porosity and pore distribution
Horizons	10^{-1} – 10^0	A horizon erosion
Whole soils (pedon)	10^0 – 10^1	Plaggen soils; salinized soils; liming that changed Ultisols to Alfisols
Soil-watersheds-landscapes-biosphere	10^2 – 10^7	Broad changes in Mollisols, Histosols, Gelisols and other soil orders due to processes such as accelerated erosion and soil organic carbon loss, as well as climate change

difficult to document than dramatic wholesale landscape changes. Histories of soil change may be difficult to reconstruct, complicated by the physical, chemical, and biological legacy of multiple land use activities and changing environmental conditions. Both inherent properties and disturbances are spatially and temporally variable, making it difficult to quantify the kinds and amounts of soil change across scales.

The most straightforward approach to assess anthropogenic soil change is to monitor soil in one location over time. Soil changes from agriculture have been monitored using controlled experiments at a few locations such as Rothamsted Experimental Station (United Kingdom) and Morrow Plots (Illinois, United States) for more than a century. An alternative approach is called “space-for-time substitution,” in which soil change by humans is measured at multiple locations either in reference to undisturbed soils that serve as benchmarks, or by comparing soils inferred to have been the same originally but that have now diverged due to different land use histories (Wills et al., 2017). For example, Fenton and others in Iowa, United States correlated declines in crop productivity with soil properties that were originally similar but experienced varying degrees of erosion (slight to severe, resulting from agriculture) at different locations (Fenton et al., 2005). Finding truly comparable sets of reference and altered soils is challenging and imprecise, and may not be possible in some situations. Pedogenic process models can also be used to understand and evaluate soil change with land use and management. When paired with careful site selection and carefully constructed hypotheses, these models provide powerful tools for linking soil formation and soil properties to disciplines such as soil management, biogeochemistry, and geophysics.

Recently evolving techniques in proximal sensing and remote sensing allow the detection and synthesis of soil change from molecular to global scales. New sensor technology provides rapid, inexpensive detection and resolution by many wavelengths of the electromagnetic spectrum. At sample and field scales, these sensors allow data collection at many more locations than traditional laboratory analyses. At continental scales, geographic information systems and remote sensing allow coverage of wide areas and detection of change through repeated data collection over time. For example, global models of soil carbon stocks have been improved through increased use of mid-infrared assessment of soil carbon concentrations, and those values have been extrapolated through remotely sensed images and geographic models using predictive digital soil mapping techniques.

Evaluating ancient agricultural soils up to about 1000–2000 years in age has extended the time perspective on anthropogenic soil change (Sandor and Homburg, 2017). Much can be learned about long-term human effects on soils, and about successes and failures in soil use and management, from studies of past cultures at archeological sites, as well as studies of contemporary traditional cultures who have lived on the same land for many generations. Investigations of past and present traditional land use contribute information for evaluating soils on longer time scales that are inherent in the concept of sustainable land use, as well as for modeling and predicting the future condition of land resources.

Changes in soil properties and processes resulting from human activities

To facilitate an appreciation of the many ways people alter soil, and depend on soil for sustenance and support (Goudie, 2013), soil changes are presented here in the context of the type of land use or other human activity. Examples from agriculture, engineering for urban development, mining, waste disposal, war, and global environmental change are selected to represent the myriad of ways people change soil. It is important to recognize that human effects on soil are complex because soils are dynamic systems with interrelated components; they are themselves part of ecosystems within the Earth’s biosphere. Soils vary in their sensitivity to disturbance and thresholds for change. Individual land-use practices can affect many soil properties in a cascading process, and different land uses may initiate similar processes of change in soils. Soils may be changed by more than one land-use practice simultaneously or asynchronously. Land uses and their variations often differ in the way they impact soils in terms of magnitude, intensity, spatial extent, rate, and duration.

Agriculture

Agriculture has profoundly and extensively affected soils since its inception over 10,000 years ago (Kuzyakov and Zamanian, 2019; Winiwarter, 2014). In this discussion, agriculture is framed broadly to include all plant and animal production for food, feed, fiber, and fuel, including crop and livestock farming, as well as forestry. All of these land uses rely on many kinds of soil worldwide.

Some soils are so impacted by agriculture that their original horizons are wholly transformed or buried. Examples of truly anthropogenic soils are plaggens of northern Europe that mostly date to approximately 1000 years ago, during the Middle Ages (Fig. 1). These constructed surface horizons, which can be a meter thick, are the product of centuries of cultivation and additions of organic materials such as manure and sod. Soils that have been similarly transformed through management practices such as terracing and long-term applications of fertilizing materials are found in other regions with long histories of intensive agriculture such as Southeast Asia and the Andes. In the Amazon Basin, anthropogenic Amazonian Dark Earth (or terra preta) soils were created over long periods through incorporation of biochar and other organic materials. They have much higher fertility and crop productivity than the more common acidic and highly weathered soils (Oxisols and Ultisols) of the region.

In contrast to both the intent and effect of constructed soils, many soils on slopes have been damaged by erosion accelerated by agriculture, some to the point of obliteration or deep burial under eroded sediments through mismanagement. Surface horizons, which generally contain the most plant nutrients and organic matter, are the most immediately vulnerable to erosion. Those soils with well-developed, organic matter-rich surface horizons such as Mollisols are especially subject to major change. The

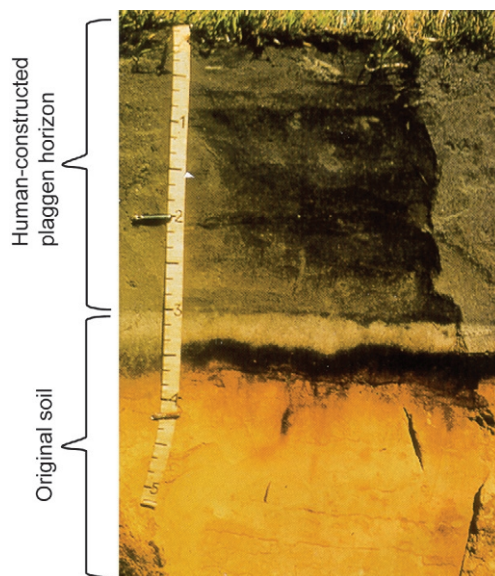


Fig. 1 Plaggen horizon (approximately 1 m thick) from The Netherlands, created by long-term organic matter additions and cultivation, burying a natural Spodosol. Reproduced from Soil Survey Staff (1999) *Soil Taxonomy*, 2nd edn. U.S. Department of Agriculture-Natural Resources Conservation Service. Washington, DC: US Government Printing Office.

transformation of Mollisols by erosion to soils classified in other orders such as Alfisols, Inceptisols, and Entisols has been documented for ancient and contemporary agriculture. A key reason for increased soil erosion is the loss of protective vegetation cover as natural ecosystems such as forests or prairies are converted to agricultural use. While accelerated erosion is difficult to quantify on regional to global scales, estimated average erosion rates on cropland range from about 12 (United States) to 30–40 (Africa, Asia, South America) metric tons per hectare annually, which is much higher than natural rates of erosion. Human-induced water and wind erosion are estimated to have resulted in the major degradation of approximately one-quarter of the world's arable land.

As surface horizons are eroded, subsurface horizons or bedrock effectively rise closer to the surface and may become exposed. Their incorporation into the topsoil under moderate to severe erosion alters soil properties such as color, structure, and texture, often causing a decline in soil quality. If subsurface horizons are problematic for plant growth, serious degradation of productivity can ensue. In some tropical regions such as West Africa and Southeast Asia, exposure of horizons of cemented, iron-rich clay (petroplinthite) through erosion has resulted in abandonment of agricultural land. Erosion of Ultisols with highly weathered, acidic horizons of clay accumulation in warm, temperate to tropical environments has significantly reduced soil productive potential. Even in thick sedimentary parent materials that are considered “forgiving” to erosion, productivity can diminish through exposure or the occurrence of dense or cemented horizons and layers at shallow depths that limit rooting, decrease available water capacity, or greatly slow drainage. For example, thick deposits of loess (wind-blown sediment rich in silt) such as in China and the Palouse and Midwest regions of the United States are favorable for crops but are particularly susceptible to erosion. Erosion rates in loess can reach 100 metric tons per hectare per year, exceeding natural erosion rates by one to two orders of magnitude. This has resulted in the loss of A horizons, followed by the exhumation of buried soils (paleosols) or dense, brittle fragipans at shallower depths that reduce productivity.

Erosion also alters soil formation away from the eroded sites. As eroded sediments are transported downslope, soils in the lower terrain of watersheds may become buried by sediment that does not reach streams. Studies in several regions have measured burial of original soils by sediment that is centimeters to meters thick. Also, the freshly deposited sediment constitutes another parent material in which soil formation may begin anew. An opposite situation occurs when natural sedimentation processes and renewal of soil fertility are impaired by dam construction, as on the Nile River where the Aswan Dam impedes annual flooding and deposition of nutrient-rich sediment. The extensive erosion and sedimentation that occur in many agricultural landscapes indicate that massive transformation of many soils is continuing. This situation underscores the need for documenting and responding to changes that undermine soil quality and productivity.

Because water deficiency or excess is of major concern in most agricultural systems, many soils have been altered by water-management practices. Some changes are direct, such as the effects of flooding soils for wet rice production, creating paddy soils with distinctive anthropogenic characteristics (Fig. 2; Zhang and Gong, 2003). “Irragric” soils in central Asia and other arid to semiarid regions have been highly altered by long-term irrigation, including the addition of suspended sediment in the irrigation water. Unintended chemical effects from irrigation in some regions have led to severe soil degradation through salt accumulation and transformation to saline or sodic soils. In contrast to paddy soil management, conversion of wetlands to row crop production by artificial drainage in regions such as the Midwest USA has changed soils from dominantly anaerobic to aerobic states, leading to



Fig. 2 Sample (approximately 5 cm thick) of paddy soil managed for wet rice production in Southeast Asia. Anthropogenic features include variegated colors from flooding and alternating reduction/oxidation, and vesicles formed from gases trapped beneath platy, puddled surface layers. Reproduced with permission from Moorman F and van Breemen N (1978) *Rice: Soil, Water, Land*. Los Baños, Philippines: International Rice Research Institute.

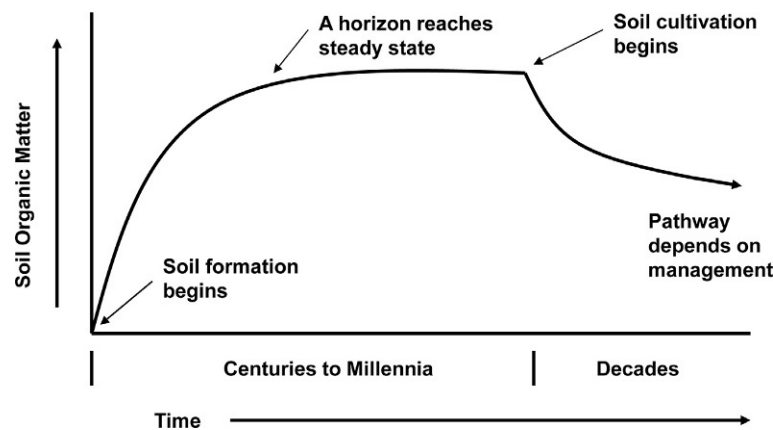


Fig. 3 Accumulation of organic matter in topsoil (A horizon) during long-term natural soil formation, and subsequent rapid decrease following cultivation. Based on the work of Hans Jenny.

changes such as organic matter oxidation. In some areas of the world, many meters of organic soil thickness have vanished through oxidation processes, resulting in significant land subsidence. Such losses of sequestered soil organic carbon have contributed significantly to increased atmospheric CO_2 and to climate change.

More subtle agricultural impacts may not completely alter the original soil, yet over time cause significant change. For example, since the mid-19th century, conventional cultivation of organic matter-rich, prairie-derived soils (Mollisols) throughout the Midwest USA has led to marked decreases in organic matter in the upper soil horizons (Fig. 3; Jenny, 1984). Many Mollisols have lost approximately one-third to one-half of their original organic matter. In addition to erosion, a principal cause is the disruption of soil aggregates by cultivation, making previously protected humus accessible to microbial decomposition. Compared with native prairie ecosystems that are characterized by abundant organic matter and conservative nutrient cycling, agricultural soils have lower inherent fertility and are “leakier” with respect to nutrients such as nitrogen. Soil organic matter loss has a cascading effect on soil properties such as structure, by reducing aggregate stability and making soils more prone to compaction. Soil health can be restored through better management and regenerative farming techniques that use conservation tillage, diverse crop rotations, and cover crops to rebuild soil biological function and diversity, organic matter, and structural stability.

Cities and industry

The growth of cities and industry has profound impacts on soils and their continued formation (Bryant and Galbraith, 2003; Levin et al., 2017). Some of the impacts are direct; others are indirect. The creation of constructed urban soils is perhaps the most easily recognized direct effect of city growth on soils (Fig. 4). Constructed soils have distinct profiles that are typically the product of the cheapest engineering fill available at the moment of creation. Pioneering urban soil scientist Phillip J. Craul wrote that the history of urban development has been based upon the premise “dirt is dirt and it’s cheap.” These human-created soils may be highly stratified



Fig. 4 Three views of urban soils at Iowa State University (United States). Upper left: urban soil being created after the natural soil has been removed. Note natural soil profile in background. Upper right: older urban soil. Note several layers and artifacts (pipe, ceramics). Scale is 10 cm. Lower right: new urban soil. (Upper left) Photo by Julie McLaughlin. (Upper right) Photo by Jon Sandor. (Lower right) Photo by Lee Burras.

and composed of very different types of fill in terms of their texture, mineralogy, and chemistry. Total fill—and consequently soil—thicknesses can reach many meters, with a soil's maximum age corresponding with the age of the city. That is, a constructed urban soil in Rome, Athens, Mexico City, and Beijing can be several thousands of years old, whereas a similar soil in New York or Buenos Aires must be less than 500 years. Given the massive ongoing rise of new urban and suburban settings, many new hectares of urban soils are being actively created each year in every city. Many ancient cities occur on “tells,” human-created mounds or hills constructed over time as a city is rebuilt on top of its predecessor. In addition to unconsolidated fill, other construction materials used in urban soils include composted sewage sludge, municipal solid wastes, heavy metals, glass, plastic, and metal. Still others consist of clean sand mined from nearby rivers or topsoil transported from local farms. In other words, urban soils have an exceptionally high degree of disorganized variability.

A second direct impact of city development on soil formation is at the landscape scale as houses, gardens, parks, and streets replace the natural terrain. The result is fragmented landscapes, where soil-forming processes become controlled by constructed topographic, hydrologic, and ecologic factors. Although the morphology of many of these soils does not show the dramatic changes of their downtown counterparts, their soil-forming processes are considerably altered. This can be illustrated by considering two hypothetical, adjacent gardens: one growing turf that is heavily fertilized and irrigated, while the other is planted to conifer shrubs and trees but receives no fertilizers and extra water. Over time, the properties of the soils under these two gardens will diverge in response to their different microclimates (wet versus dry), biota, and chemical inputs. Like their constructed counterparts, they will be characterized by disorganized variability, although in this case the disorganization occurs at the landscape (i.e., garden-to-garden) scale.

A more drastic direct effect of urbanization on soils is “urban soil sealing,” when soils are covered with buildings, roads, and other constructions. Such soils are commonly altered by removal of surface horizons, by the addition of coarse particles or exogenous fill, and by compaction before being covered by impervious material. These alterations effectively arrest the normal processes of soil formation. Urban soil sealing reduces biodiversity, inhibits soil water infiltration and internal movement, and exacerbates problems of excessive runoff and flooding in urban areas.

The impact of cities and their continued growth on soil formation is increasing. Prime farmland and wild, often fragile, lands are the two most commonly converted rural lands (Fig. 5). In the United States the area of urban land increased from 6 to 26 million

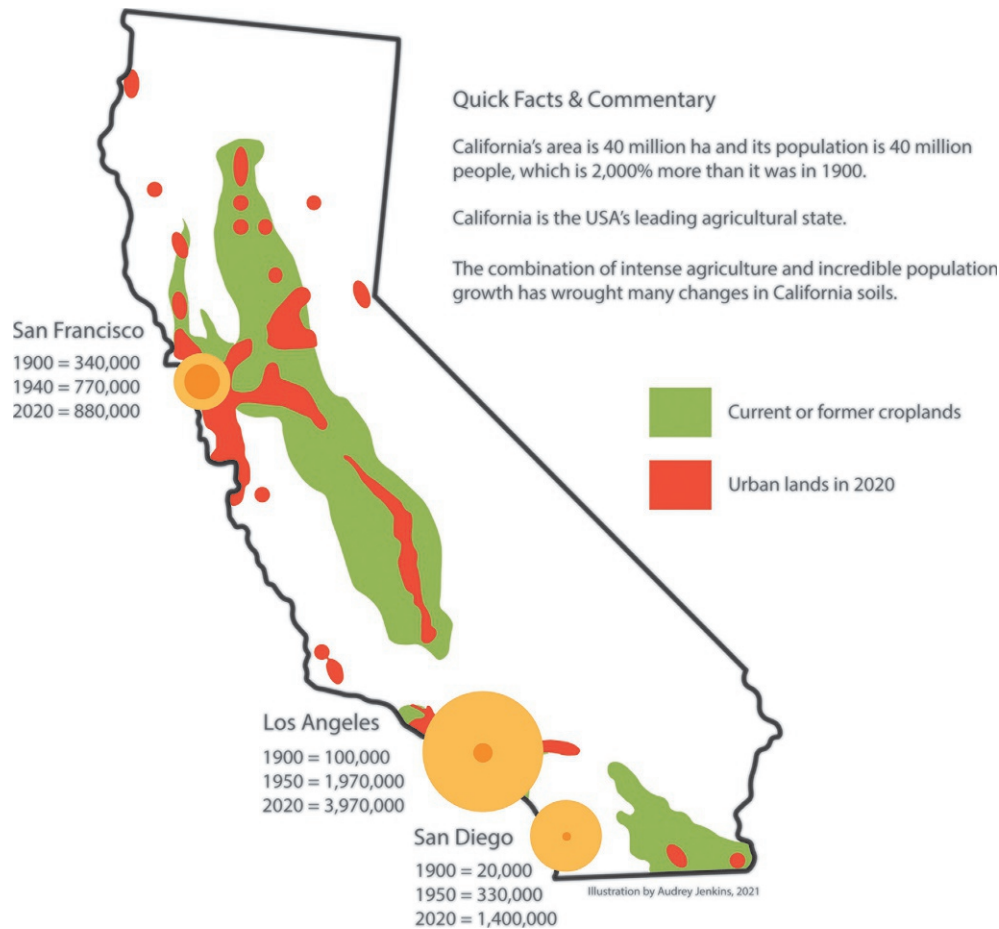


Fig. 5 California: an example wherein soils are increasingly a product of human activities such as agriculture and urbanization. Compiled from a variety of sources, including US Census Bureau and US Census of Agriculture data and maps for c. 1900, 1950, 2000 and 2020 as well as several atlases and documents from the Great Valley Center and US Geological Survey's "Preliminary Assessment of Urban Growth in California's Central Valley." Updated by Audrey Jenkins.

hectares between 1945 and 1992 and is currently estimated to be about 120 million hectares due to dramatic increases in per capita urban land conversion. In a number of locations in California and the Eastern Seaboard "100-mile cities" are common with urban areas interwoven through express highways and suburbs. Urbanization in China and India is expanding even more actively than in the United States although the most affected areas today are in developing nations due their rapidly growing and urbanizing populations. The World Resources Institute notes urban land creation results in not only in-place soil loss but also creates regional environmental impacts that reduce food and water security. There is also distortion of real estate values within and outside the cities. On the positive side, many urban areas are experiencing growth in urban gardening and locally produced food on soils that only recently were considered agriculturally useless.

Industry, which often but not always overlaps with urban areas, directly and indirectly affects soil formation. Land excavation and drainage are two important direct agents of change. Dredging of wetlands as well as sediments from beneath shallow waters to improve harbors, build canals, and meet other commercial needs is an example of industrial excavation. The US Army Corps of Engineers and the Environmental Protection Agency estimate that more than 1 billion cubic meters of sediment are dredged annually around the globe with over 300 million cubic meters being dredged annually just in the United States. Dredging has three direct impacts with respect to soils and their formation. First, the dredging itself destroys whatever natural soil existed. Second, the dredged material—or "spoil"—is stockpiled and later used in soil construction. Soil created from dredge spoil may be used for buildings, roads, parks, farms, or even wild areas. Over time it will develop a sequence of horizons in response to its environment. The third pedological impact of dredging is oxidation of mineral and organic matter. This occurs because spoil that was excavated from below the water table is now directly exposed to the oxygen-rich atmosphere. In many cases, the oxidation is fairly benign in terms of soil properties and formation. A noteworthy problem occurs when the spoil contains sulfides. Exposure of sulfidic compounds to air results in the rapid formation of large quantities of sulfuric acid (H_2SO_4) in these soils. Such acid sulfate soils are also created when naturally sulfidic soils are drained. It is estimated there are potentially 24 million hectares of acid sulfate soils globally, with many located in prime settings for agricultural and urban development.

The indirect impacts of industry on soil formation are as manifold as the indirect impacts of urbanization. One of these is the addition to soil of contaminants such as lead, cadmium, and other heavy metals. In some locations these metals are deposited on soils as airborne contaminants that originated in factory or vehicle exhausts. In other cases they are added to soil directly and intentionally as constituents of fly ash and other wastes. “Land farming” refers to disposing of waste by applying and incorporating it into soil a few tons per hectare at a time. These soils are subsequently cropped, although the crops are not normally used for human consumption. Compounds that contain sodium comprise another class of industrial contaminant that changes soil morphology—ultimately creating sodic soils. This process is increasingly noted worldwide because of the huge number of industrial uses for sodium and sodium compounds (e.g., coolant in nuclear processes; fumigants; solvents in manufacturing of brass, paper, ceramic glazes, textiles, and fertilizers; and in food processing) and the frequency at which sodium wastes are applied to land. Radionuclides are another class of industrial contaminants that alter soil processes. For example, 8.4 million hectares around Chernobyl, Ukraine, remain contaminated by ^{90}Sr and ^{137}Cs from the 1986 nuclear reactor disaster there.

The influence of most industrial processes on soil formation is proportional to the intensity of industrial activity and distance of the soil from the industry. This means that soils in areas having numerous factories, automobiles, and other industrial activities are more highly enriched in contaminants than soils where there is little industry. Likewise, the magnitude of impact radiates out with distance from the source of impact.

New tools for identifying and mapping urban and industrially impacted soils have resulted from scientific advances in the geospatial sciences and soil classification systems such as Soil Taxonomy (United States) and the World Reference Base for Soil Resources.

Mining

Humans have mined the Earth for metals, minerals, and fuel for thousands of years. Since the 19th century, in particular, mining activities have radically transformed landscapes and altered natural soil development processes. Currently, about 80% of the Earth's mined area is located in 15 countries, and about 50% occurs in China, Australia, the United States, Russia, and Chile. Mining operations are estimated to cover approximately 57,000 km². Mining normally requires that the overburden of soil and its parent materials be displaced so that the product (e.g., mineral, rock, or coal) can be removed, concentrated, and transported off-site as economically as possible. Mining often results in large pits, high walls, and large accumulations of waste spoil, that is, unconsolidated overburden, smelting residue, and finely ground tailings from which the product has been extracted. At the Earth's surface, these landscape features and waste materials pose significant environmental risks due to subsidence, settlement, and slope instability. Spoil may also be contaminated with soluble metals. In addition, when some types of mine waste (overburden and tailings) are exposed to air and percolating water, oxidation reactions produce strongly acidic water with large metal concentrations, severely restricting subsequent biological activity. Unfortunately, abandoned mining sites are abundant in many countries. Their physical hazards and impacts on surface water and groundwater resources make remediation very expensive and severely limit return of the land to productive ecosystems.

In the late 20th century concern about mines and their waste prompted many nations to enact environmental-quality laws. For example, since enactment of the Resource Conservation and Recovery Act of 1976 and the Surface Mining Control and Reclamation Act of 1977 in the United States, planning for and completing reclamation of mined land has been part of every mining operation. Thus a number of practices are actively followed to allow a mined site to be used again for other purposes and to ensure that surface water and groundwater resources are protected. Reconstruction of soil at reclaimed mine sites is an example of the most radical of human impacts on soil formation.

Successful reclamation of mined land must address a number of large-scale environmental issues while at the same time creating a near-surface environment that promotes vegetative growth and movement of water through the watershed. Reclamation efforts may include topographic reconstruction, chemical treatment of spoil to remove or mitigate hazards, burial of acid-producing overburden and tailings, establishment of new vegetation, or construction of wetlands. Organic matter and plant-available nutrients are commonly lacking in spoil, so application of biosolids or other organic amendments to the reclaimed surface is often necessary. Applications of plant nutrients such as nitrogen, phosphorus, and potassium may also be required. A major problem in reclamation of coal and metallic ore mines is the oxidation of sulfide minerals and subsequent drainage and runoff of acidic water (Fig. 6). Acidic spoil, surface water, and seepage must be treated with large amounts of calcite or hydrated lime to neutralize acidity before vegetation can be established. Revegetation in the form of grasslands is a common approach to stabilize a reclaimed site quickly. However, reforestation may be more appropriate for a particular site, depending on its ecological context and plans for long-term use.

Large-scale topographic reconstruction of mined land is expensive, but it is essential to ensure that reconstructed soils maintain ecological and environmental quality. In some reconstructions, it is possible to approximate the general contours of the pre-mining landscape (Fig. 7). During the process of consolidating, grading, and shaping the new landscape, it is critical to control settlement, limit slope instability, and restrict surface erosion to minimize off-site damage of sediments and drainage.

Where the original materials were not systematically stockpiled during mining, reclaimed soils often display extreme variability in particle size and composition. Problems of heterogeneous mixtures of coarse and fine particles include both subsidence and inadequate water-holding capacity for plants to grow. For this reason, reclamation approaches may call for stockpiling excavated material to be restored in their natural sequence (e.g., bedrock-derived spoil placed below till-derived spoil that is placed below organic matter-rich topsoil). Reconstruction of the landscape includes designing local watersheds to direct surface water



Fig. 6 Water that drains through spoil from mines is often highly acidic because of the oxidation of sulfur and iron. This water is toxic to plants and aquatic organisms. Acid Mine Drainage at the Minnesota Ridge Mine, South Dakota, United States. This site was remediated in 2001. Photo by Tom Durkin. Reproduced with permission from the South Dakota Department of Environment and Natural Resources, Minerals and Mining Program.

appropriately to higher-order watersheds. It also requires planning for drainage of subsurface water. Installation of wetlands on the site may afford opportunities for long-term treatment of acidic or metal-contaminated water that passes through the reclaimed spoil.

Rates of soil development in reconstructed mine soils are generally slow. Initially, reclaimed mine soils tend to have “AC” horizon sequences, in which incipient topsoil is underlain by unconsolidated layers lacking soil structure or other pedogenic features (Sencindiver and Ammons, 2000; Howard, 2017). The first steps of pedogenesis include physical weathering of coarse rock and clasts and chemical weathering of minerals. These processes can be accelerated where the spoil is affected by acidic drainage or where it is poorly consolidated and allows for rapid movement of air and water. Organic matter accumulates near the surface as newly established plants and their roots grow and decay. The development of pedogenic structure in both surface and subsurface zones depends on the activity of roots, microorganisms, and fauna that gradually return to the soil. The frequency of wet–dry and freeze–thaw cycles will also regulate the formation of peds. In both reclaimed coal mine spoil of the humid-continental climate of eastern United States and in reclaimed limestone mine spoil in tropical southeastern Brazil, recognizable A horizons in the upper 10 cm of have been observed within 5–6 years, and incipient B horizons have been identified after periods of 20 years. Reclaimed mine soils are typically classified as Entisols in the US Soil Taxonomy and as Technosols in the World Reference Base for Soil Resources.

Similar to mining, the extraction of oil and gas from geologic deposits by traditional drilling methods, hydraulic fracturing (fracking), and excavation of oil sands impacts soil properties and soil processes. For example, significant local land disturbance occurs during the construction and use of access roads, pipelines, and well pads. Loss of vegetative cover at the site may lead to increased soil erosion and concomitant increases in suspended sediments and dissolved solids in local streams. Moreover, the leakage from pipelines or storage ponds or direct disposal of fluid wastes can contaminate soil, surface water, and groundwater with salts, metals and metalloids, naturally occurring radioactive materials, and organic additives. Additional effects of the extraction on soils can be either direct or indirect. For example, deep subsurface injections of fluids during hydraulic fracturing operations and injections of liquid waste from oil production have led to increased numbers of earthquakes in some parts of the central United States. To enable recovery of natural gas, groundwater can be first pumped to the surface from deep coal deposits. Although that water may be saline, it has been used to irrigate nearby cropland, leading to increased sodicity and salinity in the affected soils.

Recent extraction methods, coupled with regulatory oversight and remediation guidelines, have helped to reduce environmental degradation near the sites of oil and gas extraction. For example, because of innovations in the drilling technology associated with hydraulic fracturing, several horizontal wells can be drilled from the same well pad, reducing the number of well pads and access



Fig. 7 The Kennecott Flambeau Mine near Ladysmith, Wisconsin, United States was mined for copper, gold, and silver in the 1990s. At the end of mining, the open pit was completely backfilled with spoil material and native vegetation communities were established. The site now includes wetlands, native prairie, woodlands, and nature trails. Photos before mining (1991), during mining (1996), and after restoration (2001) by T-B0 and courtesy of the Kennecott Flambeau Mining Company.

roads and their collective impacts on soils and vegetation. Also, treatment of saline wastewater with gypsum (CaSO_4) and sulfuric acid or simply diluting it with higher quality water before it is used for irrigation may limit the effects of salinity on the stability of soil structure in irrigated fields. Stockpiling of displaced soil during construction of pipelines and roads facilitates the reconstruction and revegetation of soil at the site.

War

War, like mining, may have rapid and long-lasting impacts on soils (Certini et al., 2013). Examples of soil disturbance associated with warfare, preparation for warfare, and associated military operations include (1) deep soil compaction due to occupation by personnel and movement of heavy machinery and armaments, (2) excavation of trenches to limit movement of people, machinery, and weapons across the soil surface, (3) excavation of mass gravesites, (4) contamination with radionuclides from the explosion of nuclear weapons as in Japan during World War II, (5) contamination by radionuclides near sites of the manufacture of nuclear weapons, (6) contamination by fuel spills and storage of hydrocarbons, and (7) mass slope movements induced by the loss of vegetation from intentionally set fires or incendiary bombing.

The United States' war in Vietnam (1965–72) provides several specific examples of the impacts of war on soil properties and processes (Westing and Pfeiffer, 1972). There, it has been estimated that bombs created some 21 million craters between 1965 and 1971, covering more than 50,000 ha (Fig. 8). The explosion of a 2400-kg bomb typically created a hole approximately 10 m in diameter and 5 m deep. Metal fragments from each bomb could be spread over an area of 0.5 ha. Some craters exposed soil horizons of iron oxide accumulation that subsequently irreversibly hardened to petroplinthite. In local areas, soils were also intentionally disturbed by the creation of extensive underground tunnels that were used to move, hide, and store personnel and materiel. Tunnels were constructed in highly weathered Ultisols and Oxisols where the water table allowed excavation between the depths of 1.5 and 20 m below the soil surface. Once exposed to air, iron oxides in the soil hardened, physically stabilizing the tunnels and their associated underground rooms (Olson and Speidel, 2020).

Similarly, programs to clear vegetation in Vietnam altered soil processes. Widespread and concentrated application of herbicides used to defoliate both crops and native forests, as well as bulldozing of vegetation, resulted in extensive soil erosion, loss of organic matter-rich surface horizons, and increased flooding. Native mangrove forests did not return to the drastically disturbed or herbicide-treated lands. More than 50 years after the application of herbicides ended, dioxin, a manufacturing contaminant in the herbicides, remains in the soil and sediments of the countryside and near former military bases, endangering both water supplies and the human food chain (Olson and Morton, 2019). Soils affected by the Vietnam war effectively began a new phase of pedogenesis with new topography, less organic matter, disturbed parent material, and vegetation altered from the previous state of dynamic equilibrium.

In addition to the persistence of war-related changes on soil properties and processes, there may be indirect effects that have a cascading environmental impact caused by displacement of civilian, agrarian populations. In rural Vietnam, soils in bomb craters were contaminated with metal fragments, unexploded ordnance, and residual dioxin. War-related destruction of water-diversion structures has also slowed agricultural and forestry recovery for inhabitants of Vietnam. In response, some farmers have moved to less-productive, marginal land that is more susceptible to erosion. Others have moved to cities, increasing the environmental pressures on urban areas. Thus the impacts of war on soil properties and processes can lead to long-term land-use changes that affect not only soil but also people and cultures who depend on it.



Fig. 8 A cratered mangrove forest in Gia Dinh province, Vietnam in August 1971. Bomb craters created during the war in Vietnam disrupted agriculture and most other land uses. Photo courtesy of Arthur Westing. Reproduced from Westing AH and Pfeiffer W (1972) The cratering of Indo-China. *Scientific American* 226: 20–29.

Climate change and soil formation

Soil formation is closely tied to climate, and to the extent that human activities alter local or regional climate variables such as temperature and effective precipitation, soil properties inevitably change also (Horwath and Kuzyakov, 2018; Ruddiman, 2010). At the global scale, climate change is driven by increasing atmospheric concentrations of “greenhouse gases” such as carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O). As a result, the mean temperature of the Earth’s surface is rising and could increase by approximately 2–4 °C by 2100. Although the local impacts of global warming are difficult to predict with certainty, significant shifts in the mean and seasonal variation of air temperature as well as in the total and seasonal distribution of rainfall are changing natural and agricultural ecosystems. The effects of global increase in temperature are unevenly spread in both space and time. Some of the impacts include: (1) new precipitation patterns in which today’s wet regions become wetter and dry regions become drier; (2) migration northward of boreal and hardwood forests; (3) melting of ice sheets, sea ice, and glaciers, with concomitant rise in sea level that results in salt water intrusion, flooding, and submergence of coastal soils; and (4) melting of permafrost in northern latitudes and oxidation of organic matter now stored in cold-region soils. For example, thawing from temperature increase is shrinking the extent of approximately 11 million square kilometers of Gelisols (about 9% of global land surface), the cold-region soils with permafrost, where an estimated 30% of the world’s soil organic carbon resides (measured to 3-m depth). The release of significant amounts of CO_2 and CH_4 from organic matter-rich Gelisols when permafrost thaws would result in a positive feedback loop between climate warming and soil change.

Human activities over the last 200 years especially, mainly burning of fossil fuels, have thus led to increasingly rapid climate change (Ruddiman, 2010). Both regionally and locally, climate change directly and indirectly impacts soil processes by altering vegetation patterns, encouraging increased water and wind erosion, favoring mass movement, increasing the leaching of nutrients and organic matter through soils, and favoring microbial oxidation of organic matter in surface horizons.

Human-induced climate change is also altering soil by increasing the intensity and extent of fire on the landscape. Fire has always influenced soil formation and properties through mineral weathering, organic matter decomposition, and nutrient cycling. Fire plays key roles in ecosystem functions, and has also been used through human history as a management tool to increase soil fertility and to manage vegetation. In recent years, however, highly destructive “megafires” of unprecedented intensity and magnitude resulting mainly from climate change and fire suppression policy have become common in forest and grassland ecoregions that are most vulnerable to drought and warming such as in the Mediterranean, Australia, and western North America. Such extreme wildfires harm soil health by direct biological sterilization and combustion of organic matter, and indirectly through accelerated erosion following the elimination of protective vegetative cover and the development of water-repellent layers in soils. With increased aridity and temperature, and degraded soils with reduced organic matter and ability to retain and transmit water, compromised ecosystems are unable to recover. Human-induced underground fires in organic soils such as Histosols may burn uncontrolled for years, significantly increasing greenhouse gases.

Although it is difficult to predict how fast soil properties will be altered as a result of global and regional climate change, it is likely that some types of soil horizons will be lost (for example, surface horizons composed entirely of organic matter) and some types will develop where they did not previously exist (for example, spodic horizons under new boreal forests that replace tundra). As human populations adapt agricultural practices to new climate conditions, further changes in soils may be accelerated. For example, salinization may increase due to more extensive irrigation, and landscape instability may become widespread as more forest land is cleared for the production of crops and livestock. Furthermore, patterns of cultivation and grazing will shift with changing climate and weather patterns.

Significance and future of human impacts on soil formation

Soil is a critical, dynamic natural resource and vital component of ecosystems, but one that is often neglected. This is in part because soil lies hidden beneath the surface and so is not as familiar to people as water, plants, or animals. As a natural resource, soil is crucial for food-, fiber-, and fuel-production systems, for construction materials and foundations, for replenishing and maintaining the quality of surface water and groundwater, and for waste processing and containment. Natural variation in pathways of soil formation and soil properties make for a diverse range of suitabilities for different land uses and sensitivities to anthropogenic change. Conservation of soil resources and soil quality is a critical priority globally because soil is fundamentally nonrenewable on a human time scale.

Human activities dictate that soil change is inevitable and necessary. Anthropogenic soil change has a long history (Winiwarter, 2014), but it has increased exponentially in intensity, spatial extent, and rate since the 19th century. Some changes are advantageous and improve soil functionality. However, many soil changes induced by past and present land use have resulted in environmental degradation. In developed and developing nations alike, accelerated erosion, compaction and disruption of structural aggregates, decreased fertility, and contamination by pollutants continue to be sources of serious concern. Because soil serves as a filter, substrate, and reservoir linked to other land, water, and biological resources, degradation is not just detrimental to the soil internally and locally, but it extends to all parts of the larger hydrologic, geologic, and biological systems. Some soils, like plants and animals and the habitats and ecosystems with which they are closely connected, have become endangered and in some cases extinct (Drohan and Farnham, 2006). Soils are a key part of global environmental degradation such as desertification, acid precipitation, and loss of tropical forest ecosystems. Human-altered soil is a significant factor in global environmental change as both a cause and a

consequence of change. For example, oxidation of soil organic matter by cultivation is a source of increased atmospheric CO₂ and climate change, which leads to further soil change. Yet improved management techniques and restoration of soil health provide promise of carbon sequestration in soils and ecosystems.

What can be done to counter the negative impacts of human activities on soils and the environment and to support goals of conservation, soil security, and more sustainable land use (Amundson et al., 2015; McBratney et al., 2014)? Current efforts to document soil biodiversity, assess soil health in a more holistic way, and promote regenerative agricultural practices can help to achieve these goals (Lal, 2020). Improved knowledge of soil formation and soil change is an important component of these efforts. By recognizing more clearly the extent of soil change and understanding its causes, mechanisms, and consequences, the better our chances become to develop and implement management practices that can sustain soil resources and restore damaged land. Traditionally, soil maps have portrayed soils in their relatively original, undisturbed state. However, because of the unprecedented scale of soil change and massive transformation, it is imperative that anthropogenic soil change be documented, monitored, and acted upon to a much greater degree (Tugel et al., 2005). Efforts to deal seriously with the extent and significance of human impact on soil formation and distribution have led to expanded classes of human-altered soils (Bryant and Galbraith, 2003; Soil Survey Staff, 2014) and to a proposal to include Artesols (connotes artifacts) in the US Soil Taxonomy, and the recognition of Anthrosols and Technosols at the highest level in the current international soil classification system (World Reference Base for Soil Resources; IUSS, 2014). Improved knowledge of soil formation processes in relation to natural and human-altered pathways is essential to the restoration of ecosystems and the development of sustainable land use. The future of human society and the Earth we inhabit depend on soil formation processes, and increasingly, on how we respond to the changes human actions cause in soils and the biosphere.

References

- Amundson R and Jenny H (1991) The place of humans in the state factor theory of ecosystems and their soils. *Soil Science* 151: 99–109.
- Amundson R, Berhe AA, Hopmans JW, Olson C, Sztein AE, and Sparks DL (2015) Soil and human security in the 21st century. *Science* 348(Issue 6235): 1261071.
- Bidwell OW and Hole FD (1965) Man as a factor of soil formation. *Soil Science* 99: 65–72.
- Brown AG, Tooth S, Bullard JE, et al. (2017) The geomorphology of the Anthropocene: Emergence, status and implications. *Earth Surface Processes and Landforms* 42: 71–90.
- Bryant RB and Galbraith JM (2003) Incorporating anthropogenic processes in soil classification. In: Eswaran H, Rice T, Ahrens R, and Stewart BA (eds.) *Soil Classification: A Global Desk Reference*, pp. 57–66. Boca Raton, FL: CRC Press.
- Capra GF, Ganga A, Grilli E, Vacca S, and Buondonno A (2015) A review on anthropogenic soils from a worldwide perspective. *Journal of Soils and Sediments* 15: 1602–1618.
- Certini G, Scalenghe R, and Woods WI (2013) The impact of warfare on the soil environment. *Earth-Science Reviews* 127: 1–15.
- Dazzi C and Lo Papa G (2015) Anthropogenic soils: General aspects and features. *Ecocycles* 1: 3–8.
- Drohan PJ and Farnham TJ (2006) Protecting life's foundation: A proposal for recognizing rare and threatened soils. *Soil Science Society of America Journal* 70: 2086–2096.
- Dudal R, Nachtergaele FO, and Purnell MF (2002) The human factor of soil formation. In: 17th World Congress of Soil Science, Bangkok, Thailand: International Union of Soil Sciences.
- Fenton TE, Kazemi M, and Lauterbach-Barrett MA (2005) Erosional impact on organic matter content and productivity of selected Iowa soils. *Soil and Tillage Research* 81: 163–171.
- Goudie AS (2013) *The Human Impact on the Natural Environment: Past, Present, and Future*, 7th edn, Chichester, UK: Wiley-Blackwell.
- Horwath WR and Kuzakov Y (eds.) (2018) *Climate Change Impacts on Soil Processes and Ecosystem Properties*. Elsevier.
- Howard J (2017) *Anthropogenic Soils*. Springer.
- IUSS (International Union of Soil Sciences) Working Group WRB (2014) (2014 and 2015 update) *World Reference Base for Soil Resources*, 3rd edn. Rome, Italy: Food and Agriculture Organization of the United Nations.
- Jenny H (1984) The making and unmaking of a fertile soil. In: Jackson W, Berry W, and Colman B (eds.) *Meeting the Expectations of the Land*, pp. 42–55. San Francisco: North Point Press.
- Johnson DL and Lewis LA (2007) *Land Degradation: Creation and Destruction*. Rowman & Littlefield.
- Kuzakov Y and Zamanian K (2019) Reviews and syntheses: Agropedogenesis—Humankind as the sixth soil-forming factor and attractors of agricultural soil degradation. *Biogeosciences* 16: 4783–4803.
- Lal R (ed.) (2020) *The Soil-Human Health-Nexus*. CRC Press.
- Levin MJ, Kim K-HJ, and Morel JL, et al. (eds.) (2017) *Soils Within Cities. Global Approaches to Their Sustainable Management—Composition, Properties, and Functions of Soils of the Urban Environment*. Stuttgart, Germany: Schweizerbart Science Publishers.
- Macphail RI and Goldberg P (2018) *Applied Soils and Micromorphology in Archaeology*. Cambridge, UK: Cambridge University Press.
- McBratney A, Field DJ, and Koch A (2014) The dimensions of soil security. *Geoderma* 213: 203–213.
- Olson KR and Morton LW (2019) Long-term fate of Agent Orange and dioxin TCDD contaminated soils and sediments in Vietnam hotspots. *Open Journal of Soil Science* 9: 1–34.
- Olson KR and Speidel DR (2020) Review and analysis: Successful use of soil tunnels in Medieval and modern warfare and smuggling. *Open Journal of Soil Science* 10: 194–215.
- Richter DD (2020) Game changer in soil science. The Anthropocene in soil science and pedology. *Journal of Plant Nutrition and Soil Science* 183: 5–11.
- Ruddiman WF (2010) *Plows, Plagues, and Petroleum: How Humans Took Control of Climate*. Princeton, New Jersey: Princeton University Press, Princeton Science Library Edition.
- Sandar JA and Homburg JA (2017) Anthropogenic soil change in ancient and traditional agricultural fields in arid to semiarid regions of the Americas. *Journal of Ethnobiology* 37: 196–217.
- Sencindiver JC and Ammons JT (2000) Minesoil genesis and classification. In: Barnhisel RI, Darmody RG, and Daniels WL (eds.) *Reclamation of Drastically Disturbed Lands*. Madison, WI: Soil Science Society of America.
- Soil Survey Staff (2014) *Keys to Soil Taxonomy*, 12th edn. Washington, DC: USDA-Natural Resources Conservation Service.
- Tugel AJ, Herrick JE, Brown JR, Mausbach MJ, Pickett W, and Hipple K (2005) Soil change, soil survey, and natural resources decision making: A blueprint for action. *Soil Science Society of America Journal* 69: 738–747.
- Westing AH and Pfeiffer EW (1972) The cratering of Indochina. *Scientific American* 226: 20–29.
- Willis S, Williams C, Duniway M, Veenstra J, Seybold C, and Presley D (2017) Human-induced soil change. In: West L, Singer M, and Hartemink A (eds.) *Soils of the USA*, pp. 351–371. Springer.
- Winiwarter V (2014) Environmental history of soils. In: Agnoletti M and Serneri SN (eds.) *Environmental History*, pp. 79–120. Springer.
- Yaalon DH and Yaron B (1966) Framework for man-made soil changes—An outline of metapedogenesis. *Soil Science* 102: 272–277.
- Zhang G and Gong ZT (2003) Pedogenic evolution of paddy soils in different soil landscapes. *Geoderma* 115: 15–29.

Dynamic soils

Markus Egli^a and Dennis Dahms^b, ^aDepartment of Geography, University of Zürich, Zürich, Switzerland; ^bDepartment of Geography, University of Northern Iowa, Cedar Falls, IA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	69
Soil accretion	71
Principles	71
Soil evolution	72
Aeolian influx	72
Organic matter	74
Erosion	75
Evolution of erosion rates in natural soils	75
Present-day and past erosion rates	76
Thresholds in soil genesis	76
Physical explanations and consequences for soils and landscape	77
Conclusion and outlook	79
References	80

Abstract

Soil characteristics and processes are dynamic and may change distinctly during pedogenesis. Rather rarely, soil evolution occurs only progressively. Phases of stagnation or regression can be observed. Under undisturbed conditions, the soils evolve by showing mostly a humped to exponential function. With time, rates of soil production, erosion, and soil organic matter sequestration strongly decrease. With aeolian additions, this trend is more linear owing to a constant rejuvenation of the soil material. Soil formation is strongly governed by the availability of water and infiltration rate. More recently, the link between soil formation and the Earth's water cycle has become clearer with the recognition that in situ weathering rates are mostly flux-limited.

Key points

- Soil formation can be modeled by a humped, exponential or power law function.
- The trajectories of pedogenesis are often characterized by progressive and regressive phases.
- Organic carbon sequestration rates decrease strongly with increasing pedogenesis.
- Aeolian soils and soil formation in deserts are markedly different.
- Soil production and chemical weathering are primarily limited by water infiltration.
- Threshold response to external stimuli or internal change suddenly and irreversibly affects soil processes.

Introduction

Soil is a transition phase in the continuum of the Earth's surface weathering processes as minerals and rocks alter to more stable chemical states. Understanding the processes and rates of alteration of materials from their original state to phases that are more stable at the Earth's surface is basic to defining and implementing the sustainable use of soils.

The need for such understanding of the chemistry and physics of the solid surface environment is increasing. This is at least partially due to the realization of the impact(s) that human activity has on the global, regional and local cycling of chemical constituents that directly and indirectly affect all kinds of surface weathering processes.

New rock material is exposed to surface weathering whenever a glacier retreats, a landslide descends, a volcano erupts, a sand dune is removed or a lake is drained. Each event marks "time zero" so that the initial state of the resulting soil is defined with respect to the corresponding parent material, topography and climate. In pedology we are interested in the dynamics of soil characteristics and processes and how these change over time. Soil formation and the rates of soil formation/production depend primarily on (i) the original lithology of the parent materials (e.g., highly reactive minerals such as carbonates and sulfates vs. crystalline rocks), (ii) the interaction with vegetation and other organisms, (iii) the rate of supply of fresh regolith through physical weathering and erosion on various geomorphic surfaces, (iv) the age of exposure, and (v) the character of the hydrological system. This harkens back to the fundamental concept of Dokuchaev (1883) and, in an extended form, of Jenny (1941), both cited in Huggett (2021). With respect to the dynamics of soil development, the terms soil production and soil formation are frequently used and require brief explanation: "Soil production" designates the gross production while "soil development (or soil formation)" describes the net effect.

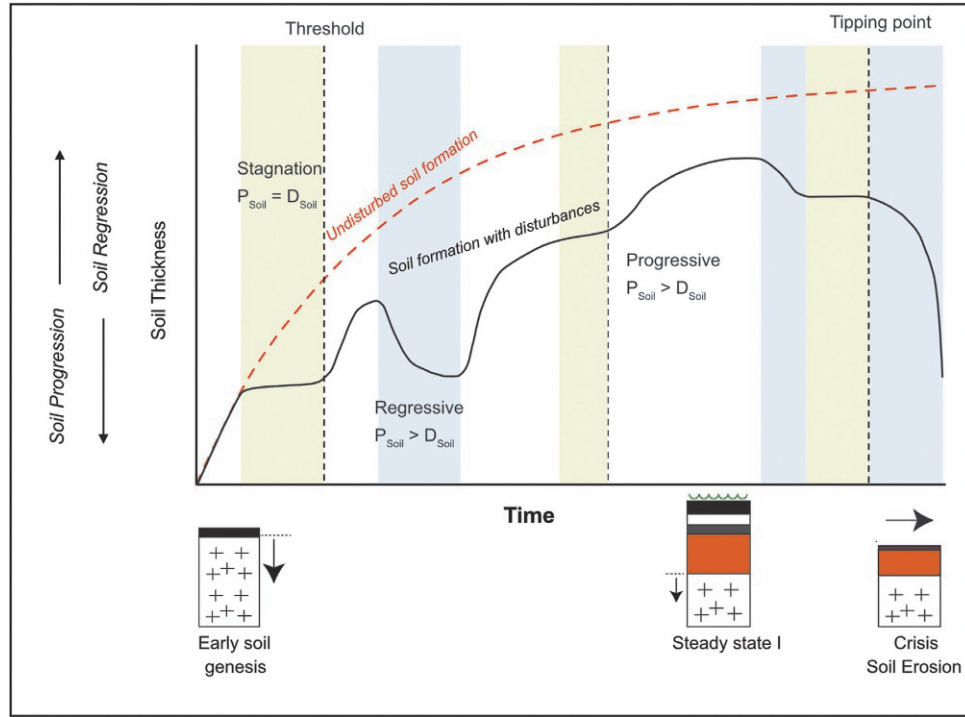


Fig. 1 A conceptual soil evolution model, showing progressive and regressive phases and phases of stagnation compared to the dominant soil evolution pathway (dashed red line) under undisturbed conditions (Johnson and Watson-Stegner 1987, Phillips, 1993; both cited in Huggett, 2021). Threshold effects may accelerate or decelerate soil formation and strong human impacts (tipping point) may drastically increase soil degradation leading to a soil crisis.

- *Soil production*: In general, soil production describes the transformation of the parent material into a soil due to chemical/physical weathering and mineral transformation resulting in the lowering of the bedrock (or parent material)—soil boundary. It can thus be paraphrased as gross regolith production.
- *Soil formation* (or soil development; see e.g., Jenny (1941), Minasny et al. (2008)): Pedogenesis can be progressive or regressive. Progressive pedogenesis includes processes that promote differentiated profiles leading to horizonization (anisotropy), leaching, developmental accretion, and soil deepening (Fig. 1; Minasny et al., 2008) while soil erosion dominates during regressive pedogenesis. Thus, soil formation is the net increase in soil thickness over time that includes the net effect of progressive (accretion) and regressive processes (erosion).

Soils and landscapes are known to form in complex and non-linear ways over thousands of years. Periods of dominantly progressive soil processes (e.g., soil deepening, regolith weathering) alternate with periods of dominantly regressive processes (e.g., erosion) due to rapid and substantial changes in many of the controlling factors of the various weathering processes. Soils can exist only if the rate of soil production is higher than denudation (erosion). Soil formation (F_{Soil}) can be therefore be expressed by

$$F_{\text{Soil}} = P_{\text{Soil}} - D_{\text{Soil}} \quad (1)$$

where P_{Soil} = soil production and D_{Soil} = soil denudation. Progressive phases occur when soil production is greater than soil denudation and regressive phases occur when soil denudation rates are greater than soil production. Denudation (D) consists of the sum of chemical weathering fluxes (W_{Soil}) and physical erosion fluxes (E_{Soil}) with

$$D_{\text{Soil}} = W_{\text{Soil}} + E_{\text{Soil}} \quad (2)$$

Thus, chemical and physical weathering and mineral transformation contribute to progressive soil formation, whereas strong erosion leads to regressive development and thereby to surface lowering and denudation (Raab et al., 2018). Phillips (1993) developed a model that incorporates progressive and retrogressive pedogenesis and related it to soil development (Fig. 1). Phillip's analysis revealed that soil development may be chaotic, with the soil profile state at any time being unique that depends on the interplay of progressive and regressive soil-forming processes (Huggett, 2021). According to Phillips (1993) (cited in Huggett, 2021) soil development can be expressed as

$$\frac{ds}{dt} = \frac{dP}{dt} - \frac{dR}{dt} = c_1 e^{-k_1 s} - c_2 e^{-k_2 s} \quad (3)$$

where s is soil development (degree of alteration and profile development), c_1 is the maximum rate of progressive pedogenesis P , c_2 is the maximum rate of regressive pedogenesis R , k_1 and k_2 are coefficients describing the decrease in P and R rates as soil develops (Huggett, 2021).

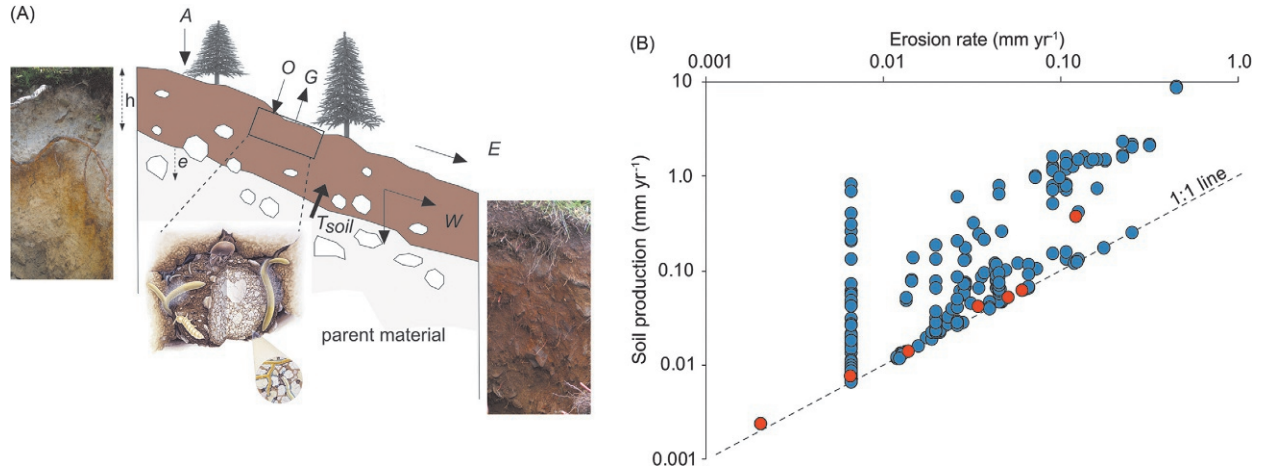


Fig. 2 (A) Schematic overview of soil forming (and, schematically, biotic processes giving rise to aggregates) along a slope. The symbol *e* refers to the weathering front and, thus, the interface between bedrock (or parent material) and soil; *h* is the soil thickness. Atmospheric input is given by *A* and the transformation of parent material into soil by *T_{soil}*. The input of organic matter (*O*) and gas output due to decay (*G*) is mediated by biotic processes (schematically drawn in the image magnification, where organic matter is decomposed, mixed with the mineral matrix, and pores are created). (B) Comparison between rates of soil production (alpine and Mediterranean areas; Egli et al., 2018) and erosion. The red circles indicate global catchment-averaged rates (Chile, northern and southeastern Australia, California, Oregon Coast Range). Data from Heimsath AM, DiBiase RA and Whipple KX (2012) Soil production limits and the transition to bedrock-dominated landscapes. *Nature Geoscience* 5: 210–214.

It has been generally acknowledged that soils evolve over time and, under more or less stable topographic conditions, increase their thickness (Fig. 2). Based on a global data compilation, soil production and denudation (erosion) positively correlate to each other (Heimsath et al., 2012; Fig. 2). A high rate of denudation rejuvenates the soils, exposes fresh minerals to the surface and thus increases the rates of weathering.

McFadden (2013) argued that in some regions, such as arid rangelands and deserts, the processes of accretion imbedded in the soil “production” model are not appropriate to explain how the observed soils developed. In these areas, the genesis of soils is controlled by processes associated with a model of *accretionary-inflationary profile* (AIP) development. The soil thickness model was elaborated by Johnson et al. (2005; cited in Huggett, 2021). In principle, the following processes contribute to soil thickness (*T*):

$$T = (W + B) + (A + O + V) - (E + L + C_{\text{surf}} + C_{\text{sub}}) \quad (4)$$

Accretionary processes in the expanded soil thickness model include bedrock weathering (*W*), bioturbation (*B*), sediment accretion (*A*), rate of organic matter accumulation (*O*), and volume expansion (*V*). Removal processes include surficial erosion and mass wasting (*E*), leaching (*L*), and surface consumption by fire, uptake, decay and harvesting (*C_{surf}*), and subsurface removals (*C_{sub}*) by the same processes.

As a consequence of these considerations, the mass balance of a soil is the result of the difference between soil material production from the parent material, aeolian input and erosion (or denudation). For all soils, changes in soil material mass can be expressed by:

$$\frac{\partial(\rho_w h_w)}{\partial t} = \rho_{\text{UA}} Q_{\text{UA}} - \nabla(\rho_w Q_D) = (\rho_r Q_r + \rho_A Q_A + O) - \nabla(\rho_w Q_D + C) = \phi - \nabla q_e - \nabla q_w - \nabla C \quad (5)$$

where ρ = bulk density (*w*: soil; *r*: parent material), h_w = soil material thickness, Q_{UA} = soil material mass production rate per unit area, Q_A = aeolian input contributing to soil formation, Q_D = denudation mass flux rate per unit area, ϕ = soil production rate, q_e = soil erosion rate per unit area, q_w = chemical weathering rate per unit area and C = gain/loss of organic matter.

Due to steep slopes and in many cases extreme climate conditions, mountain soils may experience particularly high rates of erosion and chemical weathering, and consequently, production rates. Soil production and transport play essential roles in controlling the spatial variation of soil depth and therefore hillslope hydrological processes, distribution of vegetation, and soil biological activity.

Soil accretion

Principles

Soil chronosequences are commonly used to analyze the temporal course of pedogenesis. A time-dependent sequence of soils formed on progressively older landscapes is considered a chronosequence, while the quantitative variability of a characteristic along these landscapes is a chronofunction (Bockheim, 1980). Using chronosequences, soil formation or even soil production rates can be

calculated. A time series of soils that have developed under constant environmental conditions is very rarely achieved. In general, soils that started forming at subsequent points in time over the Holocene or Pleistocene may have been under the influence of different climatic conditions for some parts of their lifetime (Zwanzig et al., 2021). Despite efforts to keep all soil forming factors constant except time when establishing soil chronosequences, various differences in vegetation or parent material characteristics, etc., may occur even in the same study area (Zwanzig et al., 2021). Nonetheless, soil chronosequences can provide very useful information about weathering regimes and soil morphogenesis and they have greatly advanced our understanding of short- to long-term landscape processes. Soil development has been successfully correlated to land surface age in several regions, although in some cases such studies were complicated by erosion and sedimentation as well as possible tectonic activity (Sauer et al., 2010). Thus, the study of chronosequences has a long tradition, and both ecologists and soil scientists use the concept of substituting space for time to understand ecosystem and soil evolution.

Soil evolution

Soils form only when accretionary processes exceed denudation rates (Eq. 1). Therefore, soil trajectories must be dominated by progressive processes.

Birkeland (1984) argues that most chronofunctions are fitted best by a power model, but some are fitted better by linear or exponential models. Although problems exist in the purity of the chronosequences because other soil-forming factors are not completely constant, duration of pedogenesis seems to be the most significant factor in explaining the trends. When a regolith cover is lacking, soil production and formation stall at the beginning. As soon as physical weathering starts to break up the rock, chemical weathering increases up to a maximum and then decreases exponentially. This description depicts a ‘humped’ function and was first described by Gilbert (1877; in Riebe et al., 2017). The humped and negative exponential soil production function can be described as follows (Riebe et al., 2017; Fig. 3):

$$P = P_0 e^{-bh} (1 + ch) \quad (6)$$

Here P is the soil production rate, h is soil thickness, and P_0 is the production rate under zero soil thickness. The parameter b is a scaling factor for the decrease in rate of production with increasing h , and c designates the thickness at which soil production is maximized. For $c = 0$, the humped function turns into a negative exponential, with a maximum production rate of P_0 at $h = 0$ (Riebe et al., 2017). Soil production rates attain a maximum at an intermediate soil thickness. Today, it is widely thought that other processes, such as root wedging and animal burrowing, could contribute to humped soil production functions in landscapes.

The exponential phenomenology for the soil production function obscures the connections between solute transport, chemical weathering, C and N sequestration rates, and soil production processes (Hunt and Ghanbarian, 2016). The cause of this difficulty is that chemical weathering and non-Gaussian transport (Hunt and Ghanbarian, 2016) are described using power laws, rather than an exponential phenomenology (see below). It is increasingly recognized that weathering processes and thus soil formation are transport limited, i.e., by the water infiltration rate.

Aeolian influx

Aeolian materials affect soil development in many ways. Aeolian influx is usually considered to be associated with “progressive” soil development (Fig. 1), since new materials are added to the soil that build up the profile thickness. Such external influences can completely alter the “direction” of soil development and are considered to be external threshold effects. Loess and dust deposition occurs worldwide. Both glacial and non-glacial processes produce fine-grained particles that can be transported by the wind (Muhs, 2013). Silt and dust production is caused by glacial grinding, by fluvial comminution, from volcanic activity, by frost shattering, chemical weathering, aeolian abrasion, by salt weathering in desert basins, etc. Silt and dust may then be transported over short to long distances and deposited giving rise to a new substrate for soil formation.

In desert areas with large amounts of aeolian influx, soil development is often cumulative as it is associated with the development of desert pavement forming over continually thickening vesicular A horizons (Fig. 4). Soil profile anisotropy then develops over time within the accumulated materials rather than as a result of a purely bedrock-to-soil transformation as often proposed by a soil “production” model. This contrast is especially noted in the hillslope soils of arid rangelands (McFadden, 2013). McFadden (2013) suggests that the model for soil formation in deserts is ‘markedly different’ from the CLORPT-based A/B/C model of soil development generally attributed to Dokuchaev (1883) and Jenny (1941) and the soil production function model. McFadden (2013) suggests that a model of *accretionary and inflationary profile* (AIP) development (an “upward growth” model) is more appropriate for areas where entrapped dust becomes translocated beneath desert pavement features (Fig. 4).

Over time, as aeolian influx of mineral materials thicken surface horizons (unless surface erosion is greater than the rate of influx), the aeolian deposits change the soil physical/chemical characteristics by adding new materials of various mineralogies to the grain size distribution. The physical changes may then affect a cascade of physical and chemical processes. With proceeding soil evolution, the predominantly silt-sized particles are translocated deeper into soil profiles through water percolation and freeze-front activity. These conditions are often seen in coarse-textured arctic and alpine soils. The additional materials change the infiltration characteristics of subsurface horizons, so that under unsaturated conditions, the textural difference(s) caused by the accumulation of silt-sized particles may temporarily retard deeper infiltration. Under seasonally saturated conditions with a high gravitational head,

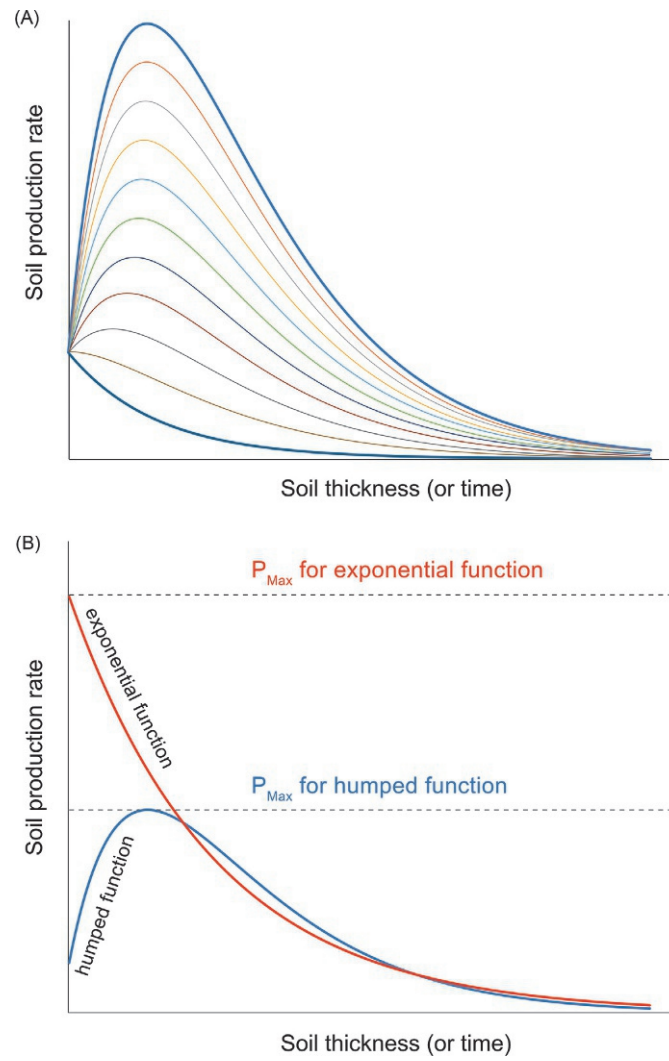


Fig. 3 Chronofunctions according to Eq. (6) describing the transition between a negative exponential to a humped function.

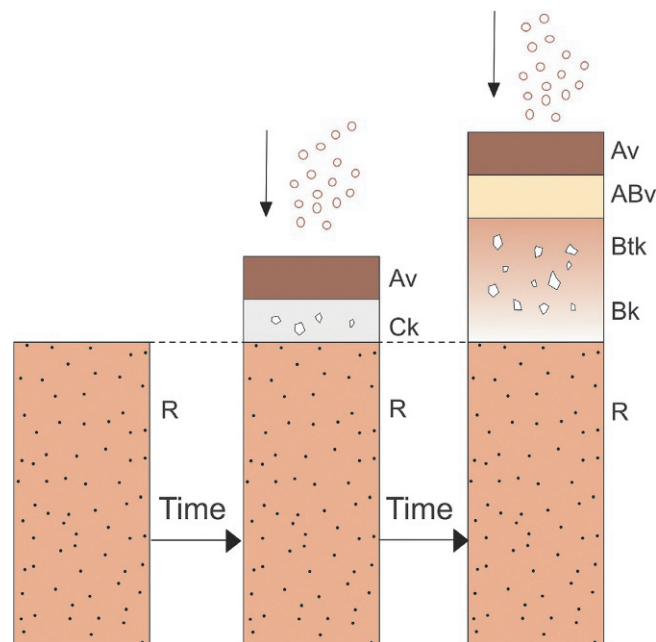


Fig. 4 Conceptual model of soil development dominated by accretion (according to McFadden, 2013; Muhs, 2013).

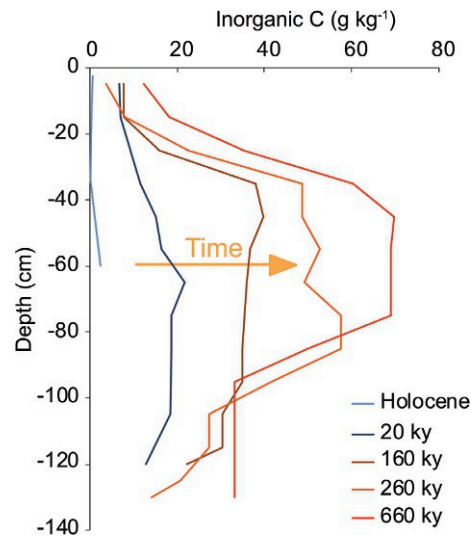


Fig. 5 Carbonate accumulation in granitic soils (outwash terraces) over time due to dust input. From Dahms D and Egli M (2016) Carbonate and elemental accumulation rates in arid soils of mid-to-late Pleistocene outwash terraces, southeastern Wind River Range, Wyoming, USA. *Chemical Geology* 446: 147–162.

however, water is “flushed” deeper into the profile. When the water table is seasonally high, water may be translocated upward by capillary action into horizons that ordinarily are unsaturated. Over time, these seasonally saturated conditions contribute to the formation of “redoximorphic” features in deeper soil horizons. Since the addition of the aeolian parent material is ultimately responsible here, the textural change that creates the conditions for redox features is considered to be an “external” pedogenic threshold (Muhs, 1984; see below).

The purely accretionary model of McFadden often does not strictly apply and, particularly in less dry regions, aeolian input is admixed into the parent material or existing soil (see the model given in Muhs, 2013). Aeolian influx may change soil characteristics through changes in weathering pathways. Minerals supplied through aeolian influx will become actively involved in common weathering reactions (solution/hydrolysis/hydration/redox) that introduce ‘new’ soluble minerals to subsurface horizons (e.g., CaCO_3 , CaSO_4 , etc.). If these new minerals were absent from the original parent material complex, their presence constitutes a threshold, after which other processes may occur that would not have previously occurred. Examples of such a process are seen in the U.S. western mountains where carbonate dust is commonly added to granitic parent materials of alpine soils (Dahms and Egli, 2016; Fig. 5). As soil moisture dissolves the dust, it moves into the subsurface where it retards the translocation of clays and impedes the development of argillic horizons over time.

Reheis (1990; cited in Dahms and Egli, 2016) reported that carbonate-rich aeolian dust affects the major-element geochemistry of soils on a chronosequence of granitic outwash terraces in Wyoming in a manner that varies according to moisture conditions. The B horizons of the moist soils developed at logarithmic rates while those of the most arid soil developed at linear rates. The variation is apparently due to differences in leaching rates and the addition of aeolian silt, clay, CaCO_3 and gypsum, which act to increase the water-holding capacity in the moist-area profiles. In a similar manner, the clay minerals of the B horizons varied according to moist–dry conditions. Vermiculite, illite-smectite and smectite form in B horizons of the moist chronosequence, while smectite and palygorskite form in the dry B horizons. Amorphous clays appear to be the main clays in the aeolian complex (rather than crystalline forms), which are only later transformed into crystalline clays in these profiles. In other parts of the world, e.g., Middle Europe, soils in mountainous areas are also strongly influenced by aeolian silt. On slopes, cover beds have often developed. Waroszewski et al. (2018) showed that many mountainous soils have an aeolian contribution with a different geologic origin. The incorporation of aeolian silt is the main agent in initiating and stimulating clay translocation and the formation of an argic horizon within mixed zones or just below the boundary between the silt mantle and coarse/fine basal strata.

Organic matter

If the evolution of soils is not disturbed too much by erosion and accumulation processes, then organic matter stocks and rates of accumulation of carbon and nitrogen depend greatly on the age of the surface (or soil, respectively). In a review of Holocene chronosequence studies in a variety of ecosystems, Schlesinger (1990; in Egli et al., 2012) found a much higher and more variable rate of C accumulation in relatively young soils. In Alpine environments, soils developed on silicate parent material can accumulate up to $\sim 40 \text{ kg organic C m}^{-2}$, with frequent values between 15 and 25 kg C m^{-2} (Egli et al., 2012). Within about 500 years, the rates of carbon accumulation decrease markedly to very low values (Fig. 6). Based on an investigation of 10 glacier forelands distributed on four continents, Khedim et al. (2021) found that the rate of SOM accumulation in topsoils is accelerated by higher temperatures

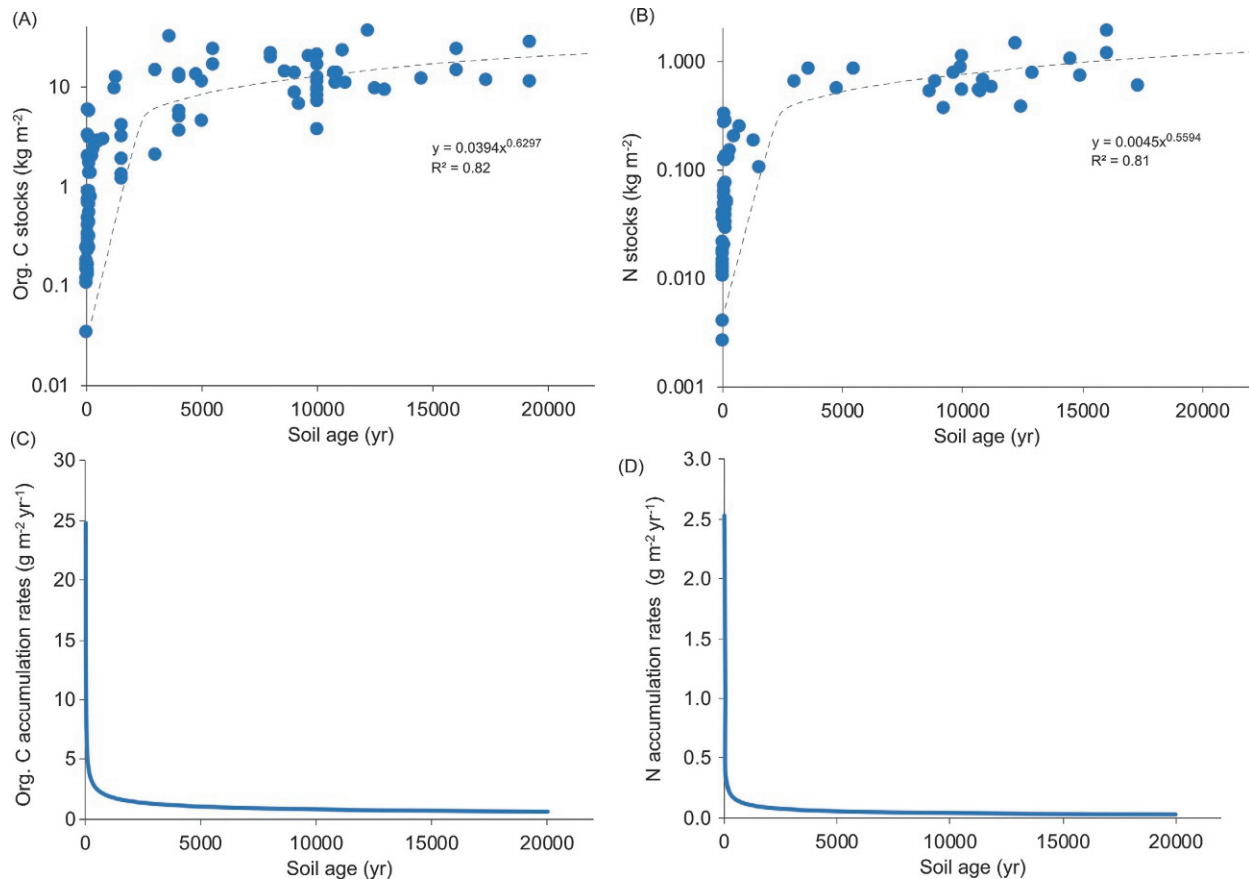


Fig. 6 Compilation of org. C (A) and N (B) stocks as a function of soil age and calculated rates of sequestration over time for C (C) and N (D) for alpine areas worldwide. Data from Egli M, Favilli F, Krebs R, Pichler B and Dahms D (2012) Soil organic carbon and nitrogen accumulation rates in cold and alpine environments over 1 Ma. *Geoderma* 183–184: 109–123.

during the warmest quarter, suggesting that climate and time are key factors of SOM build-up during the initial stages of topsoil development after glacier retreat. The stability of SOM decreases with soil age at all sites, indicating that SOM storage is dominated by the accumulation of labile SOM during the first centuries of soil development, suggesting plant carbon inputs to soil (SOM depleted in nitrogen, enriched in hydrogen and in aromatic carbon). A steady state of the amounts of stable organic C and N compounds is often reached within about 1000 years (Egli et al., 2012).

Erosion

Evolution of erosion rates in natural soils

The surface of the continental Earth is constantly reworked at different scales of time and space by weathering and erosion. Natural rates of soil formation counteract soil erosion and can be used as a basis for setting tolerable rates of soil erosion (Verheijen et al., 2009).

Since soil properties change over time, one has to expect that soil erosion rates also change in parallel. At the beginning of soil formation, e.g., due to glacier melt and exposure of fresh material, the vegetation cover is often scarce, which promotes soil erosion. Such areas are prone to strong erosion. With time, vegetation can establish and increasingly covers the surface and therefore minimizes the erosive power of rainfall (Fig. 7).

In the mid-European Alps, the erosion rates are strongly reduced after about 3 kyr of soil formation, whereas under drier conditions (e.g., Wind River Range of the Rocky Mountains, US) it may take longer until slopes stabilize. Depending on the relief and vegetation development, it takes up to at least 10 ka to reach soil stability (Musso et al., 2020).

Several recent studies have attempted to compare erosion rates over various periods of time in order to estimate rates of different geomorphic processes. Qualitative or quantitative estimates of erosion rates over different time-scales have been reported for river sediment yields, lake and marine sedimentation (Bajard et al., 2017) or for speleothem growth rates. Raab et al. (2018) and Mariotti et al. (2021) report that (soil) erosion and denudation rates have varied considerably in the Mediterranean and European Alps over the last ca. 100 kyr. High-resolution cores collected in the Mediterranean Sea provided a near-continuous reconstruction of

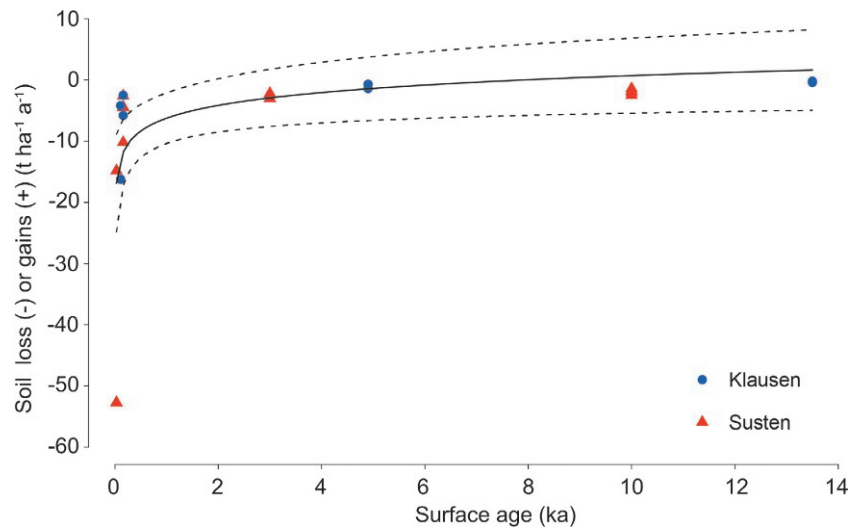


Fig. 7 Temporal evolution of erosion rates in two proglacial areas of the Alps (Musso et al., 2020).

denudation in the Southern Alps since 75 kyr ago (ka). These high-resolution palaeo-denudation records were compared with well-constrained climatic variations over the last glacial cycle. Mariotti et al. (2021) concluded that denudation rates were approximately twice as great as during the Last Glacial Maximum LGM (26.5–19 ka). Raab et al. (2018) showed that denudation rates were about five times higher in the Mediterranean directly after the LGM and reached values of about $\sim 150\text{--}200 \text{ t km}^{-2} \text{ year}^{-1}$ (compared to $30 \text{ t km}^{-2} \text{ year}^{-1}$ before the LGM and during Holocene). Bajard et al. (2017) showed that the erosion rates in an Alpine catchment were mostly very low during the Holocene (about $1 \text{ t km}^{-2} \text{ year}^{-1}$), but increased sharply during the last ~ 3 kyr.

Present-day and past erosion rates

In many agricultural landscapes, soil erosion rates are an order of magnitude higher than possible soil production rates. As shown by Bajard et al. (2017) and Raab et al. (2018) and several others, erosion rates strongly increased after humans started to use soils more intensively. Often, erosion rates increased even more than one order of magnitude compared to natural rates leading to strong soil degradation. It has become obvious that soil erosion in the Anthropocene mainly occurs as a consequence of a combination of natural and human-induced soil erosion processes. In an increasing number of case studies anthropogenic soil erosion processes have become dominant (Poesen, 2018). The first signs of increased erosion due to human impact apparently occurred in the Bronze age and became increasingly obvious about 2 ka BP (Bajard et al., 2017). Taking steps to preserve the quality and quantity of global soil resources should require no justification (Poesen, 2018).

Thresholds in soil genesis

Soil properties sometimes change suddenly and irreversibly in a threshold response to external stimuli or internal change in soil processes (Chadwick and Chorover, 2001). The concept of the “threshold” was not used explicitly for studies of soil genesis until recently. Muhs (1984) noted that, unlike traditional geomorphic studies of landscape development, studies of soil genesis have not explicitly used the threshold concept. Yaalon (1971) presented examples of research that applied the concept, but they refer to it only indirectly as, respectively, a “metastable state”, an “environmental” threshold or a change in one of the “state” factors (Jenny, 1941).

Two types of thresholds are normally recognized: *Extrinsic* and *Intrinsic*. Extrinsic thresholds involve alterations that are associated with changes in external variables that are outside the immediate soil system to explain variations or changes in soil development/genesis. Examples of external thresholds that change the trajectory of soil evolution might be changes in climate, temperature/precipitation gradient, parent material or vegetation succession. These are easily seen in alpine environments where, with similar parent materials, soils develop from Inceptisols -to- Mollisols -to- Spodosols. These developments occur because various of these gradients are crossed with increasing altitude that cause changes in organic matter content, water percolation and associated changes in soil geochemistry. Dixon et al. (2016) show that soil chemical changes occur rapidly across a narrow portion of a rainfall gradient, indicating that water availability creates thresholds for soil weathering. In a temperate to cool-temperate climate zone (New Zealand Alps), Dixon et al. (2016) measured soil elemental geochemistry along a large precipitation gradient ($400\text{--}4700 \text{ mm year}^{-1}$). They found evidence of an important pedogenic threshold that coincided with a mean annual precipitation of $\sim 800 \text{ mm year}^{-1}$, which was associated with cation weathering and release, and the rapid mobilization of trivalent metallic ions.

Intrinsic thresholds are those that can be crossed only after a prior change in a physical/chemical property alters the direction of soil genesis *in the absence of environmental change* (Muhs, 1984). The accumulation of authigenic clays in argillic horizons that eventually decreases internal drainage conditions (Yaalon, 1971; Muhs, 1984) is an example of an intrinsic pedogenic threshold. In such a case the valence state of Fe can be changed from ferric to ferrous. This change may then trigger a cascade of geochemical processes beginning with the replacement of ferrous Fe in dioctahedral-smectite that subsequently decreases the content of dithionite-extractable Fe over time. Bockheim (1980) observed Fe in B horizons to reach a maximum value and then to decrease over time.

A primary suite of chemical reactions in soil involves the production of acids by biodecay or atmospheric input and their depletion through the weathering of minerals (Chadwick and Chorover, 2001). Soil development can be characterized as a long-term acid–base reaction in which acids react with bases in the form of rock minerals to form secondary minerals and soil solution alkalinity. Depending on the buffering capacity and type of buffer range, there is little change in the soil pH when acids are added because weathering mechanisms are able to buffer these acids.

Additional examples of intrinsic thresholds where no change in external variables is involved include the impediment/movement of clays in the presence of Ca^{++} or Na^{++} (Muhs, 1984). The flocculating effect of Ca^{++} may be cancelled by sufficient leaching of free carbonates, thus allowing clays to begin to move. Likewise, in soils developing on parent materials rich in marine shales, as an appropriate amount of Na^{++} is released by weathering of the parent material, it enhances the ability of clay to move and rapidly accumulate. However, this has a limit which becomes another intrinsic threshold; once clays accumulate to an extent that pore spaces become too small to allow adequate percolation, little clay movement can occur subsequently (Muhs, 1984).

Physical explanations and consequences for soils and landscape

Weathering, soil formation, vegetation succession and its decay and integration into the soil are all processes that are interconnected and driven by (solar) energy and moisture availability. Soil formation and vegetation growth are both processes with significant impact on global C cycling, although the primary influences differ at widely varying time scales. Individual organisms, populations, community distributions and ecosystem composition and function strongly depend on water availability, the energy balance and organic matter. The organisms supply the CO_2 for the weathering and soil formation processes. Together with the pore separation (particle size) that defines the fractal dimensionality of a soil system, the water flow rate determines the rate of soil formation (Hunt and Ghanbarian, 2016).

Recent modeling and comparisons with field results showed that soil formation by chemical weathering, either from bedrock or unconsolidated material, is more often limited by solute transport than by reaction kinetics. Predicting what drives the transition from ‘non-soil’ to a soil-mantled landscape (or from bedrock to a “developed” soil mantle) is, therefore, a significant challenge for models of landscape evolution and for “critical zone” studies. If predictions for the rates of soil formation and vegetation growth, at both regional and global scales, can be made accurately, effects of future climate change on alpine systems can be assessed more reliably.

Although precipitation is abundant in many regions, the availability and flux of water through the soil is the prime factor that controls weathering intensity. This understanding derives from the percolation theory, that both vegetation growth and soil formation processes are proportional to the water fluxes remaining after direct evaporation processes are excluded, including interception (e.g., Hunt and Ghanbarian, 2016). Thus, the rate of reaction is proportional to the water throughflow within the Earth’s skin (i.e., the soil or regolith), and quantifying the partitioning of water at the terrestrial surface–atmosphere interface becomes key to understanding rates of weathering around the world (Egli et al., 2018).

Water fluxes in soils depend on evapotranspiration, overland flow, subsurface flow and vertical percolation. These partitions of water and in particular the soil infiltration rates often vary greatly as a function of topography. Polar-facing sites are usually characterized by lower evapotranspiration rates and thus have higher rates of infiltration than equatorial-facing slopes (Fig. 8; Pelletier et al., 2018). As a consequence, soil characteristics are often developed better at polar-facing sites where more water infiltrates and so, chemical weathering is less limiting (Figs. 8 and 9). The conceptual model of Pelletier et al. (2018) can be related to weathering rates, soil formation, soil biology, organic matter and wood decomposition. They all indicate that infiltration rates and thus water predominantly control the rate of reactions (Fig. 8; Hunt and Ghanbarian, 2016). This system, however, is valid only to a certain point, from which temperature, and thus a higher thermal gradient, becomes more important. This holds for very cold environments in which lower temperatures on polar-facing slopes limit vegetation cover and reverse many of the feedbacks associated with the water-limited case (Pelletier et al., 2018).

Hunt and Ghanbarian (2016) proposed upper bounds for three processes described by scaling relationships: soil depths, natural vegetation, and crop heights, each as a power of time. The bounds are related to maximum water flow rates. The general scaling relationship can be represented as

$$x = x_0 \left(\frac{t}{t_0} \right)^P \quad (7)$$

with the primary distinction among these cases being the value of the exponent P . For both natural vegetation and intensively managed crops, the variable x represents the height or root radial extent; for soils, x represents depth. The variable t represents the

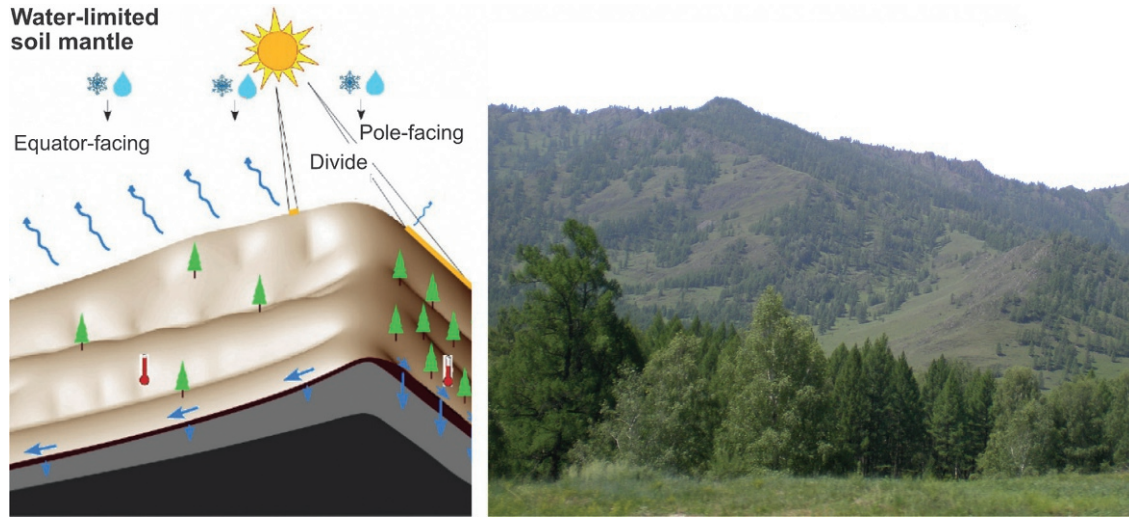


Fig. 8 Conceptual models for a water-limited soil-mantled case. Soils are less developed on equator-facing hillslopes, due to the limiting water storage potential. Adapted from Pelletier JD, Barron-Gafford GA, Guitérrez-Jurado H, Hinckley E-LS, Istanbuluoglu E, McGuire LA, Niu G-Y, Poulos MJ, Rasmussen C, Richardson P, Swetnam TL and Tucker GE (2018) Which way do you lean? Using slope aspect variations to understand critical zone processes and feedbacks. *Earth Surface Processes and Landforms* 43: 1133–1154; photograph: Altai mountains, Russia.

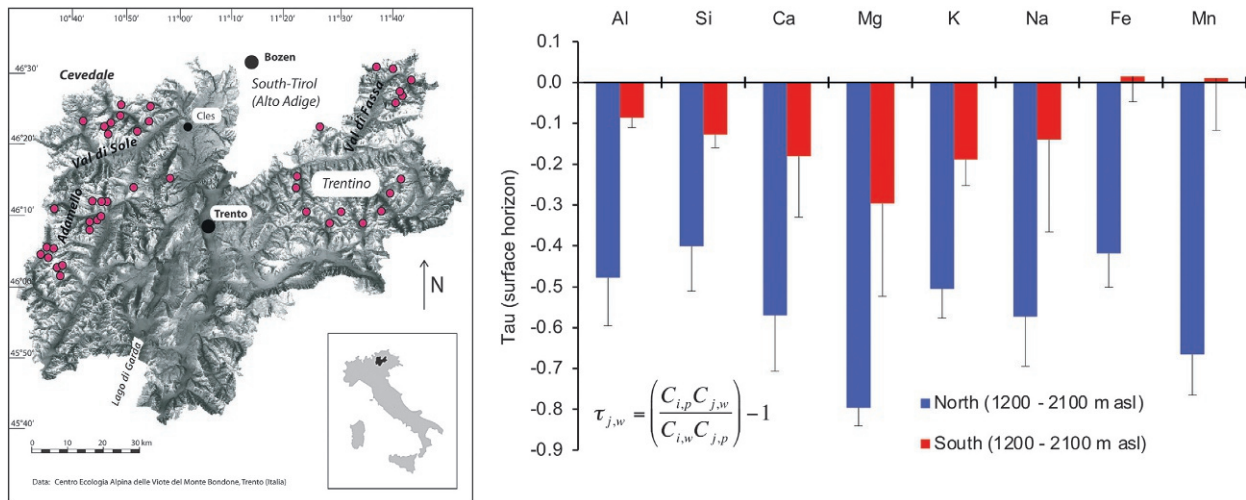


Fig. 9 Relative element losses due to weathering and chemical leaching in the Alps where $\tau_{j,w}$ denotes the open-system mass transport function, i the immobile element, $C_{i,p}$ (g kg^{-1}) is the concentration of element j in the unweathered parent material, and $C_{j,w}$ is the concentration of element j in the weathered product (g kg^{-1}). Chemical losses are distinctly higher at north-facing sites (moister conditions) than at south-facing sites.

time since germination for plant growth; for soil depth, it is the time since exposure at the surface. The spatial scale, x_0 , relates to a grain or pore (or plant xylem) diameter, and x_0/t_0 is the water flow rate, v_0 , underlying the particular process.

It has been shown that, under the assumption that chemical weathering is limited by solute transport (and thus water flow into the soil), the process of soil production may be predictable (Hunt and Ghanbarian, 2016; Fig. 10). Such predictions appear successful when applying percolation theory for solute transport in porous media. The percolation theory can be used when the medium is highly disordered and chemical weathering in situ is transport-limited (Hunt and Ghanbarian, 2016). Variations in soil depth and vegetation characteristics with topography appear to be best understood when we consider the effects of aspect on rates of erosion and the effect(s) of water supply and energy balance on soil production (Fig. 10).

Soil depth represents the integration of soil production and denudation (Fig. 2). Denudation (D) is the sum of erosion (E) and chemical leaching (W): $D = W + E$.

Temporal differentiation of Eq. (1) for soil depth generates the soil production function:

$$\frac{dx}{dt} = R - E(t) = \frac{1}{1.87} \frac{x_0}{t_0} \left(\frac{t}{t_0} \right)^{-0.87} - E(t) = \frac{1}{1.87} \frac{I(t)}{\phi} \left(\frac{x}{x_0} \right)^{-0.87} - E(t) \quad (8)$$

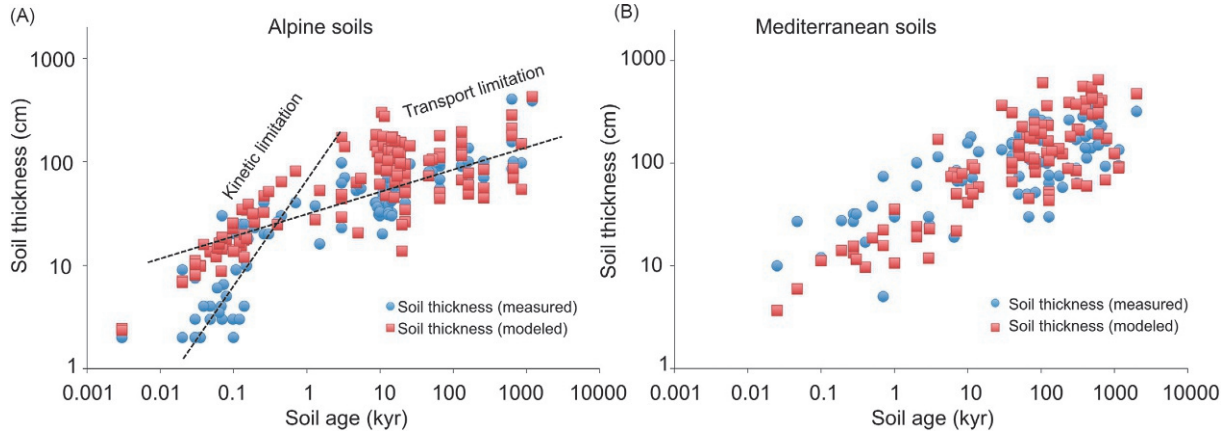


Fig. 10 Time trends of modeled (using the percolation theory, Eq. 8) and measured soil thickness for distinct geographic regions, (A) alpine soils, (B) Mediterranean soils. In alpine soils, mineral weathering seems to be kinetically limited until about 1 kyr.

with $I(t)/\phi$ as the net infiltration rate that varies, like E , over time. From this equation it is evident that solute transport through soil (similar to erosion of soil matrix) limits the formation of soils. Solute transport limitations reflect the necessity for flowing water to bring reagents to (chiefly CO_2), and to remove reaction products from, the reaction zone (the bottom soil boundary), preventing equilibration. Thus, depending on local/regional climate forcing, soil depth implicitly reflects slope processes in terms of erosional and biological processes (Figs. 2, 7 and 8).

Under steady-state conditions, soil production and rates of soil erosion are equal. Soil production is proportional to the rate of water infiltration, meaning that processes such as evaporation and interception reduce the water available for replacing soil lost by erosion, thus also placing a limit on the rate of erosion (Fig. 9). At longer time scales, carbon and nitrogen sequestration rates appear to have the same time-dependence as rates of soil formation (Hunt and Ghanbarian, 2016). Similarly, steady-state in biologic processes, presumed relevant for a given climatic regime and for near steady-state soil conditions, would imply an equality of production and consumption of carbon stocks in the soil. The higher rates of decay on north-facing slopes can be explained by the greater moisture content. Under conditions of climate or land-use change, however, steady-state conditions will be disrupted, leading to rates of soil erosion (and soil carbon depletion rates), for example, that are larger than rates of soil production (and carbon sequestration). Such effects are also expected to be aspect-dependent, with the most pronounced impacts on equator-facing slopes because water infiltration is hypothesized to be lower and erosion rates higher compared to north-facing sites giving rise to shallower soils. At high altitudes (above the treeline), we expect a reversal of these processes.

Transport limitation and *kinetic limitation* are often processes that determine the speed of silicate weathering. A transport-limited weathering regime exists when the supply of water, inorganic and organic acids and ligands are large relative to the supply of silicate (West et al., 2005). The supply of fresh minerals limits the weathering rates. Old and flat topographies are typically in this category. A *kinetic* limitation is then given when the main control on chemical weathering ω depends on the kinetic rate of mineral dissolution W , the supply of material (e.g., by erosion) e , and the time t available for reaction (West et al., 2005)

$$\omega = W \cdot e \cdot t \quad (9)$$

Mineral dissolution (W) depends on environmental conditions such as temperature and runoff (or precipitation). The rate of weathering of a mineral or rock decreases with the time the mineral spends in the weathering environment (West et al., 2005). Apparently, young surfaces or surfaces exhibiting a high erosion rate (leading to a continuous rejuvenation of the surface and soils) are in this category (Fig. 9).

The percolation theory model therefore fails when simulating soil formation on young surfaces (Fig. 9). This seems to be the case up to about 100 years, but sometimes up to 1000 or more years of soil formation (Fig. 9; Egli et al., 2018). During early soil formation, chemical weathering is more limited by reaction kinetics than by solute transport. Thereafter, solute transport seems dominant. Thus, the reaction rate is proportional to the water throughflow within the Earth's skin, and quantification of the partitioning of the water at the interface between the terrestrial surface and the atmosphere becomes key to understanding the rates of weathering around the world (Hunt and Ghanbarian, 2016).

Conclusion and outlook

Soil formation and production rates in general tend to decrease distinctly over time and can be best modeled using a humped, an exponential function or a power model. Often undisturbed conditions are not encountered and soil formation can even be regressive. The trajectories are influenced by a number of internal and external processes (among which also anthropogenic

activities can be mentioned). Soil formation in deserts is “markedly different” and rather tends towards a linear development. As soil chronosequences are crucial to the study of soil genesis and since chronosequence studies are based on the concept of *time* as an independent (i.e., external) variable, the passage of time can be seen to enable any number of internal thresholds. That is, it takes time for silicate weathering to produce enough authigenic clays to then be organized in sufficient concentration to decrease the movement of soil water; it takes *time* for a soil solution to remove enough Ca^{++} to allow clays to move and accumulate in argillic horizons. Thus, the use of soil chronosequences is important to studies of soil genesis and landscape evolution, since only by seeing the effect of time the presence and the effect(s) of internal thresholds can be illuminated.

Although it seems obvious that the availability of water and its transport through the soil determines the rate and trajectories of soil formation, a number of questions remain open. It is still not fully clear why the rates of weathering of minerals measured in the field are several orders of magnitude less than those obtained from laboratory experiments. According to Fig. 10, a phase of kinetic limitation is detectable in alpine soils, but not in Mediterranean soils (with the available datasets). Therefore, it would be interesting to know under which conditions a kinetic control is dominant and when exactly the change into transport limitation occurs. Given the fact that soil production rates (Fig. 2) vary over 3–4 orders of magnitude (Hunt and Ghanbarian, 2016), the soils have a robust tendency to be about 1-m deep in temperate to boreal climates, with the topsoil and the typical rooting depth about one third of that (Hunt and Ghanbarian, 2016; Egli et al., 2018). Aeolian, (semi) desert soils seem to exhibit a linear evolution trend with time. If and when a quasi-steady state will be established is so far uncertain. The characteristics of threshold in soil evolution are known, however not their influence on the rate of soil production and organic matter sequestration.

Weathering of silicate rocks removes CO_2 from atmosphere. With rising levels of atmospheric CO_2 , the understanding of rates of weathering is increasingly important—also when related to enhanced weathering of silicates as a removal technology by adding powdered silicate rock to agricultural soils (Andrews and Taylor, 2019)—to estimate the true ability of weathering to sequester carbon.

References

- Andrews MG and Taylor LL (2019) Combating climate change through enhanced weathering of agricultural soils. *Elements* 15: 53–258.
- Bajard M, Poulencard J, Sabatier P, Develle A-L, Giguet-Covex C, Jacob J, Crouzet C, David F, Pignol C, and Arnaud F (2017) Progressive and regressive soil evolution phases in the Anthropocene. *Catena* 150: 39–52.
- Birkeland PW (1984) Holocene soil chronofunctions, Southern Alps, New Zealand. *Geoderma* 34: 115–134.
- Bockheim JG (1980) Solution and use of chronofunctions in studying soil development. *Geoderma* 24: 71–85.
- Chadwick OA and Chorover J (2001) The chemistry of pedogenic thresholds. *Geoderma* 100: 321–353.
- Dahms D and Egli M (2016) Carbonate and elemental accumulation rates in arid soils of mid-to-late Pleistocene outwash terraces, southeastern Wind River Range, Wyoming, USA. *Chemical Geology* 446: 147–162.
- Dixon JL, Chadwick OA, and Vitousek PM (2016) Climate-driven thresholds for chemical weathering in postglacial soils of New Zealand. *Journal of Geophysical Research* 121: 1619–1634.
- Egli M, Favilli F, Krebs R, Pichler B, and Dahms D (2012) Soil organic carbon and nitrogen accumulation rates in cold and alpine environments over 1 Ma. *Geoderma* 183–184: 109–123.
- Egli M, Hunt A, Dahms D, Raab G, Derungs C, Raimondi S, and Fang Y (2018) Prediction of soil formation as a function of age using the percolation theory approach. *Frontiers in Environmental Science* 6: 108.
- Heimsath AM, DiBiase RA, and Whipple KX (2012) Soil production limits and the transition to bedrock-dominated landscapes. *Nature Geoscience* 5: 210–214.
- Huggett RJ (2021) Soil as a system: A history. In: Hunt AG, Egli M, and Faybishenko B (eds.) *Hydrogeology, Chemical Weathering, and Soil Formation*. AGU Geophysical Monograph Seriespp. 3–20. Hoboken, USA: John Wiley & Sons.
- Hunt AG and Ghanbarian B (2016) Percolation theory for solute transport in porous media: Geochemistry, geomorphology, and carbon cycling. *Water Resources Research* 52: 7444–7459.
- Khedim N, Cécillon L, Poulencard J, Barré P, Baudin F, Marta S, Rabatel A, Dentant C, Cauvy-Fraunié S, Antheime F, Gielly L, Ambrosini R, Franzetti A, Azzoni RS, Caccianiga MS, Compostella C, Clague J, Tielidze L, Messenger E, Choler P, and Ficitola GF (2021) Topsoil organic matter build-up in glacier forelands around the world. *Global Change Biology* 27: 1662–1677.
- Mariotti A, Blard P-H, Charreau J, Toucanne S, Jorjy SJ, Molliex S, Bourlès DL, Aumaître G, and Keddadouche K (2021) Nonlinear forcing of climate on mountain denudation during glaciations. *Nature Geoscience* 14: 16–22.
- McFadden LD (2013) *Strongly Dust-Influenced Soils and What They Tell us About Landscape Dynamics in Vegetated Aridlands of the Southwestern United States*. Special Paper #500 Boulder, CO: Geological Society of America.
- Minasny B, McBratney AB, and Salvador-Blanes S (2008) Quantitative models for pedogenesis—A review. *Geoderma* 144: 140–157.
- Muhs DR (1984) Intrinsic thresholds in soil systems. *Physical Geography* 5: 99–110.
- Muhs DR (2013) The geologic records of dust in the quaternary. *Aeolian Research* 9: 3–48.
- Musso A, Ketterer M, Greinwald K, Geitner C, and Egli M (2020) Rapid decrease of soil erosion rates with soil formation and vegetation development in periglacial areas. *Earth Surface Processes and Landforms* 12: 2824–2839.
- Pelletier JD, Barron-Gafford GA, Gutiérrez-Jurado H, Hinckley E-LS, Istanbuluoglu E, McGuire LA, Niu G-Y, Poulos MJ, Rasmussen C, Richardson P, Swetnam TL, and Tucker GE (2018) Which way do you lean? Using slope aspect variations to understand critical zone processes and feedbacks. *Earth Surface Processes and Landforms* 43: 1133–1154.
- Poesen J (2018) Soil erosion in the Anthropocene: Research need. *Earth Surface Processes and Landforms* 43: 64–84.
- Raab G, Scarciglia F, Norton K, Dahms D, de Castro Portes R, Christl M, Ketterer M, Ruppli A, and Egli M (2018) Denudation variability of the Sila massif upland (Italy) from decades to millennia using ^{10}Be and $^{239+240}\text{Pu}$. *Land Degradation & Development* 29: 3736–3752.
- Riebe CR, Hahm WJ, and Brantley SL (2017) Controls on deep critical zone architecture: A historical review and four testable hypotheses. *Earth Surface Processes and Landforms* 42: 128–156.
- Sauer D, Wagner S, Brückner H, Scarciglia F, Mastronuzzi G, and Stahr K (2010) Soil development on marine terraces near Metaponto (Gulf of Taranto, southern Italy). *Quaternary International* 222: 48–63.
- Verheijen FGA, Jones RJA, Rickson RJ, and Smith CJ (2009) Tolerable versus actual soil erosion rates in Europe. *Earth Science Reviews* 94: 23–38.

- Waroszewski J, Sprafke T, Kabala C, Muszyńska E, Łabaz B, and Woźniczka P (2018) Aeolian silt contribution to soils on mountain slopes (Mt. Ślęza, southwest Poland). *Quaternary Research* 89: 702–717.
- West AJ, Galy A, and Bickle M (2005) Tectonic and climatic controls on silicate weathering. *Earth and Planetary Science Letters* 235: 211–228.
- Yaalon DH (1971) Soil-forming processes in space and time. In: Yaalon DH (ed.) *Paleopedology—Origin, Nature and Dating of Paleosols*, pp. 29–39. Jerusalem: Israel Universities Press.
- Zwanzig L, Zwanzig M, and Sauer D (2021) Outcomes of a quantitative analysis of 48 soil chronosequence studies in humid mid and high latitudes: Importance of vegetation in driving podzolization. *Catena* 196: 104821.

Interactions between soils and climate change

Cornelia Rumpel, CNRS, Institute of Ecology and Environmental Sciences Paris (IEES Paris), UMR Sorbonne Université, CNRS, INRA, IRD, UPEC, Paris, France

© 2023 Elsevier Ltd. All rights reserved.

Introduction	82
Climate change predictions	83
Impact on soil properties and biogeochemical cycling	84
Impact on soil functions and ecosystem services derived from soil	85
Impact on land-degradation processes	86
Soil erosion	86
Desertification	87
Salinization	87
Fire	88
Soil-based climate change mitigation and adaptation options	88
Conclusion, summary, outlook	89
References	89

Abstract

There is little doubt that anthropogenic activities have led to increasing atmospheric carbon dioxide (CO₂) concentrations, global temperature increases, changes in precipitation regimes and more frequent occurrence of extreme events. These global change effects alter soil properties and their biogeochemical functioning. In this chapter, I will present the global impacts and likelihood of these change and their influence on soil functioning, biogeochemical cycling and ecosystem services in different environments. Feedbacks between the effects of climate change and soil will be presented and soil-based strategies for climate change mitigation and adaptation will be discussed.

Key points

- Increasing CO₂ concentrations, increasing temperature and altered precipitation regime affect organic and mineral soil properties, biogeochemical cycling and provision of ecosystem services.
- These effects may be direct and or indirect through their influence on vegetation and plant growth.
- Extreme events affect soil degradation processes through increasing soil erosion, salinization, desertification and frequency of fire. Their impact on organic and mineral soils is contrasting.
- Feedbacks between soil degradation and climate change occur through CO₂ uptake and release from soils.
- Sequestration of organic and inorganic carbon in soils may be options for climate change mitigation and adaptation. They require large scale implementation through region- specific strategies.

Introduction

Soils form through the interaction of five factors, including (1) parent material, (2) relief or topography, (3) organisms (including humans), (4) climate, and (5) time. The nature of this interaction varies in the different regions of the world, resulting globally in several thousand soil subtypes and transition phases. Soil formation is a long process and consequently it takes thousands of years for a soil to form, while in comparison global changes in the anthropocene have occurred only recently after the beginning of industrialization. Even rapid changes in any of the soil-forming factors following global changes will affect directly and indirectly biogeochemical cycling and functioning of soils, with important implications for their development and use. As a result, human activity impacts soil formation through perturbations of several earth system processes at the planetary scale. Global changes affecting soil are related to the increasing atmospheric CO₂ concentrations leading to climate change, loss of biosphere integrity including genetic and functional biodiversity, and perturbations of the biochemical flows by massive use of N (nitrogen) and P (phosphorus) fertilizers for agricultural production.

This chapter focuses on the increasing atmospheric CO₂ concentrations inducing global changes in climatic conditions. These include temperature and moisture, which may influence soils and the ecosystem services they provide. Depending on their nature, soils are likely to respond in various ways to climate change. For example increasing global temperatures may lead to increased (micro-) biological activity and change microbial community composition, thereby affecting soil organic matter (SOM) storage and biogeochemical cycling. Temperature and precipitation changes also alter the rates of mineral weathering. Moreover, extreme

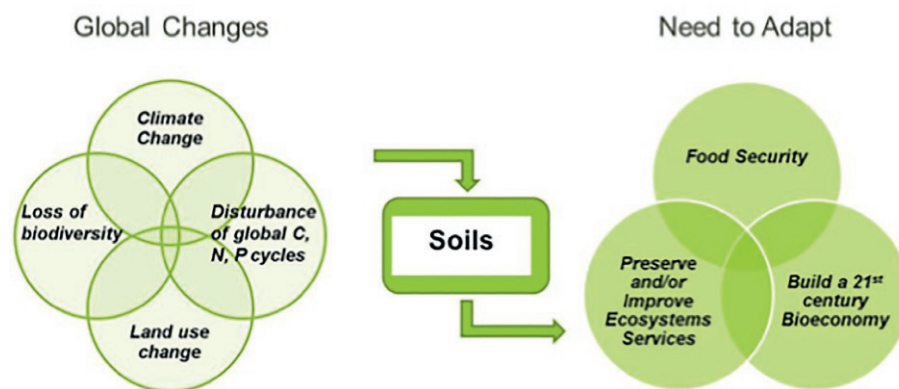


Fig. 1 Global change pressures on soils and the reasons for requiring adaptation.

events, such as drought and flooding are likely to increase under climate change and may not only have various effects on soil biota and biogeochemical cycling, but could also influence soil salinity due to irrigation measures (Corwin, 2021) and rates of soil erosion (Routschek et al., 2014). Other climate change impacts include rising sea levels, which may affect soils in coastal areas (Corwin, 2021) and increasing frequency and size of wildfires (Keyser and Westerling, 2019), which may disturb soils under forest ecosystems. All these climate change effects will most probably alter the functioning of natural as well as managed ecosystems. However, the effects of climate change will not operate in isolation, as perturbations of earth system processes may interact. To secure ecosystem service provision, food security and to build the bioeconomy of the 21st century under global changes, system adaption will be necessary (Fig. 1).

While climate change may thus affect a number of soil properties, soils are also responding to climate change through the release and uptake of CO₂ and other greenhouse gases. These processes occur as positive and negative feedback mechanisms following increasing temperatures and/or after changes in landuse and management. In the past, these alterations mainly led to the release of CO₂, N₂O (nitrous oxide) and CH₄ (methane) to the atmosphere following intensive agricultural practices, which favored SOM decomposition leading to a carbon loss of more than 100 Gt since the beginning of industrialization (Sanderman et al., 2017). However, it was recognized that sustainably managed soils may be carbon sinks and they were thus cited as a nature-based solution to mitigate greenhouse gas emissions (Bossio et al., 2020), thereby fighting climate change.

This chapter introduces the main climate change predictions and their likely effects on soil properties and their consequences for soil functioning and biogeochemical cycling. The possibility of mitigating these effects by sustainable management of soils to contribute to climate change mitigation, adaptation and food security will also be discussed. Organic and mineral soils will be distinguished, as their response to climate change may be substantially different (Rumpel, 2019).

Climate change predictions

Estimates of global climate change are continually being revised as models are being improved and new data are collected. The current main sources of information to inform on the likely extent of climate change are the Sixth Assessment Reports of the Intergovernmental Panel on Climate Change. The latest report of working group one on the physical science bases, which integrates scientific knowledge based on 14,000 publications, concluded that human influence has warmed the climate at a rate that is unprecedented in at least the last 2000 years. The increase in global surface temperature relative to 1850–1900 (IPCC, 2021) was +1.1 °C in 2020. As a consequence, northern ice masses are melting leading to sea level rise and permafrost thawing. In addition, with warmer air temperatures the atmosphere can hold more water and more water will be evaporated and returned to land with heavier and more frequent precipitation. Global warming consequently leads to extreme events such as heavier rainfall and flooding, the intensification of dry seasons in some regions with more frequent and more intense heat waves, droughts and more frequent fire events. Some of the projected environmental changes and their likelihood are presented in Table 1. These events driven by climate change have already been recorded in every inhabited region of the world and there is medium to high confidence in the human contribution to the observed changes (IPCC, 2021). As a result, the Sixth Assessment Reports include assessment of climate change mitigation and adaptation options.

Soil is addressed in the IPCC's special report on land and climate change, which was released in 2019. This report indicated that warming over land has occurred at a faster rate than the global mean and that these warmer temperatures and associated changing patterns of precipitation have altered the start and end of the growing seasons, and reduced the availability of freshwater (IPCC, 2019). These changes affect soil through their impact on vegetation and this is likely to increase in the future. However, there is also considerable confidence in options for land based mitigation and adaptation involving soil through sustainable land management practices that could both reduce emissions and increase carbon uptake. These options have co-benefits in the form of enhanced crop productivity, plant nutrient status and/or biodiversity.

Table 1 Projected climate change induced alterations of the global environment.

<i>Changes in environmental conditions</i>	<i>Confidence in projections</i>
Global temperatures will continue to increase until 2050	High confidence
Global water cycle will continue to intensify with temperature increase	High confidence
Global area experiencing increases or decreases in seasonal precipitation will increase	Medium confidence
Drought and or flooding events will intensify	High confidence
Monsoon precipitation is projected to increase	High confidence
With increasing CO ₂ emissions, sink efficiency on land and ocean decreases	High confidence
Mountain and polar glaciers will continue melting for decades or centuries	Very high confidence
Global mean sea level will continue to rise over the 21st century	Virtually certain

Impact on soil properties and biogeochemical cycling

Increased CO₂ concentrations and associated temperature may affect physicochemical soil properties and in particular weathering processes. Indeed, rising atmospheric CO₂ concentration was found to increase soil acidity, although carbonic acid is only a weak proton source. This effect may be even stronger under nitrogen fertilizer application. Thus increased atmospheric CO₂ concentrations may enhance rates of chemical weathering directly because of increased dissolution of CO₂ in soil water, which leads to the weathering of silica, depending on the lithology, ambient temperature and moisture conditions. Indirectly weathering may be affected by climate change through its effect on plant and soil (micro-)organisms as they are strongly involved in these processes (Rosenstock et al., 2019). Weathering processes, may lead to the removal of CO₂ from the atmosphere through pedogenic carbonate formation as a negative feedback mechanism. Therefore, it has been suggested that artificial acceleration of this carbon sink by the distribution of pulverized silicate rocks across the Earth's surface, especially in tropical regions, where weathering rates are high, could help to offset anthropogenic CO₂ emissions (Taylor et al., 2016). However, the authors also indicate that this solution has major barriers due to costs, social acceptance and unanticipated consequences.

In addition, the increase of CO₂ concentrations and associated higher air temperatures have strong impacts on soils and their properties through their effects on the soil–plant continuum. Alteration of both parameters influences vegetation growth and plant physiology. As a result, increasing atmospheric CO₂ concentrations were hypothesized to lead to greater carbon input into soil following more intensive photosynthesis, thereby increasing the terrestrial carbon sink. However, evidence from field experiments and observations is contradictory. Consequently, other factors, such as temperature, nutrient and or water availability may also play a role and strongly alter the plants' response to elevated CO₂. Plants may respond by several mechanisms to changes in the physicochemical properties of their environment including those of resource allocation and stoichiometry with consequences for soil processes and soil carbon residence times (Walker et al., 2021). A recent literature review indicated that greater plant input belowground may be compensated for by increased soil organic carbon (SOC) turnover. Indeed, high plant-derived C input may affect microbial biomass and its activity, metabolism and respiration. It may also induce positive priming of SOC and nutrient pools. A recent study concluded that following elevated CO₂, pool sizes may be unaltered while process rates could be increased thereby accelerating soil biogeochemical cycling and organic matter turnover (Kuzakov et al., 2019).

Soil microbial activity, such as growth and enzyme production, depends on temperature. Much research interest has been given to the temperature sensitivity of SOM decomposition, which is critical for forecasting the fate of soil (organic) carbon under a warming climate. In theory, under ideal conditions with non-limiting substrate availability, SOM decomposition follows Arrhenius kinetics and increases exponentially with temperature (Arrhenius, 1889). In these cases, the relative temperature sensitivity increases with chemical recalcitrance of the organic substrate and the stable, slow cycling SOM pool will be more vulnerable than the fast cycling pool (Davidson and Janssens, 2006). This might be the case for organic soils of peatlands, which formed under oxygen deficiency resulting from high water tables. Such soils are present on only 3% of the land area mainly in the boreal and temperate regions. Nevertheless, they are important contributors to the global carbon budget as they store 21% of the global SOC stocks (Leifeld and Menichetti, 2018).

In mineral soils, with less than 20% organic carbon in the surface layer, the situation is different as the residence times of SOM are less controlled by its chemical structure than by its location within the soil matrix. Under such conditions, the temperature sensitivity of SOM decomposition may be determined by the nature of physicochemical protection mechanisms rather than substrate properties (Von Lützow and Kögel-Knabner, 2009). Evidence from recent field studies suggested that physical protection of SOM constrains the temperature sensitivity of its decomposition. Consequently, temperature sensitivity of SOM decomposition in mineral soils was found to decrease with increasing SOM stability, which is in direct contrast to the Arrhenius theory (Moinet et al., 2020).

More than half of the organic carbon in soil is often stored below the topsoil and possibly also affected by global warming, as the world's soils are assumed to warm by 4.5 + 1.1 °C down to 1 m by the end of the 21st century under the Representative Concentration Pathway (RCP) 8.5 (IPCC, 2014). Therefore, recently attention has been paid to biogeochemical processes occurring in whole soil profiles subjected to warming. A better understanding of the response of the deep soil carbon reservoir to warming would probably contribute to a decrease in the large uncertainties in model predictions of soil responses to warming. A recent study

showed that warming by 4 °C of profiles of forest mineral soil down to 100 cm augmented CO₂ emissions from all depths increasing annual soil respiration by 34–37% (Hicks Pries et al., 2017). This is considerably more than observed in previous *in situ* studies taking into account only the surface soil. Similar temperature sensitivities were recorded at all soil depths and related to the degradation of decadal-aged SOC. In the forest soil, this warming thus induced an increase in decomposition that led to the loss of subsoil organic carbon and nitrogen stocks and a concomitant increase in topsoil (0–20 cm) SOC stocks (Soong et al., 2021).

Soils of the permafrost regions are likely to be strongly affected by global warming due to thawing. The unique pedogenic processes in these regions on the one hand accumulated large organic carbon stocks and permafrost soils account for approximately 50% of the global SOC pool (Hugelius et al., 2014). The accumulation of organic matter in permafrost-affected soils occurs in the active layer due to their high moisture content and low temperatures, which affect decomposition (Ping et al., 2015). Warming of these soils will lead to changes in their hydrology and microbial functioning. A recent study indicated large SOC losses from high carbon soils affected by permafrost not only as CO₂ but also through lateral hydrological export (Plaza et al., 2019). However, the vulnerability of the accumulated SOC stocks to climate change will depend on several parameters, including its form and protection mechanisms, vegetation establishment and the soil microbial consortium (Ping et al., 2015). On the other hand, vast permafrost regions have mineral soils with sparsely distributed vegetation and shallow soils without prior accumulation of thick organic layers. In these regions, warming leads to changes in soil chemical, physical and (micro-)biological properties, which may favor plant life and soil development (Doetterl et al., 2021). In particular, rates of (bio-)chemical weathering of mineral surfaces may be enhanced under warmer and wetter conditions (Hall et al., 2002) and global warming may therefore accelerate soil formation through increasing the kinetics of chemical reactions and vegetation establishment. These processes could lead to soil carbon sequestration thereby partly compensating for carbon loss from thawing organic permafrost layers (Doetterl et al., 2021).

In addition to temperature, moisture availability is a major factor determining microbial substrate access, SOM decomposition and soil biogeochemical cycling at several scales. As climate change leads to extreme climatic events such as drought and flooding, fluctuations in water availability will strongly influence microbial community composition and activity with effects on soil physicochemical properties and its organic constituents.

Drought, for example, has direct negative consequences for microbial activity and substrate access by causing osmotic stress through increased salt concentrations in the soil solution and consequently ionic strength. This can alter chemical environments at the pore scale and influence the sorption and desorption of stabilized organic molecules at mineral surfaces. Drought leads to reduced access to nutrients and organic matter by limiting substrate diffusion within the soil pore network. It also has impacts on leaching losses of elements and rates of mineral weathering (Schlesinger and Bernhardt, 2013). On the other hand, flooding may lead to anaerobic conditions directly affecting the microbial community composition. It also has consequences for the soils' redox potential and the chemical form of mineral soil constituents such as Fe (iron) and Al (aluminium), which in reduced forms may be less able to stabilize organic matter and may adversely affect plant growth. To investigate the changes in soil processes and biogeochemical cycling *in situ* long-term experiments that generate empirical data in multiple environments and climatic contexts are necessary. Long-term monitoring sites are also necessary to benchmark model projections. Improved biogeochemical models are needed to increase confidence in the magnitude of projected soil–climate feedbacks. Model improvement should be done by including recent advances in mechanistic understanding of soil biogeochemical cycling (Bradford et al., 2016).

Impact on soil functions and ecosystem services derived from soil

Soils are fundamental to land-based ecosystems and, in association with climate, govern the nature and distribution of the world's ecosystems. They are multifunctional natural bodies providing water, nutrients, and habitat for a number of organisms. Soil functions can be influenced by management and may be altered by climate change. Although most ecosystems are naturally variable with time and adapt reasonably well to small climatic fluctuations, changes to soils as a result of climate change are likely to influence their function and this will also be reflected in a change in ecosystems and their services (Fig. 2). The form the change will take depends much on the degree of climate change, and the possibility of adaptation through management.

The most recognized function of soils is their production function and the implications it has for feeding the global population. Indeed, production of food and other commodities relies over 90% on soils and food security of a growing population puts pressure on soils, which are already degraded to a large extent (FAO and ITPS, 2015). Climate change effects on the physical, chemical, and biological properties of soils alter their nutrient recycling function. This impacts the soil biogeochemical cycles and biodiversity with major implications for their fertility and thus their production function (Smith et al., 2015). Soil properties are also important in controlling the fate and behavior of water once it reaches the surface. Therefore, related soil functions likely to be influenced by climate change are water storage and filtration. Consequently, global warming and resulting changes in the precipitation regime may affect the ability of crops to reach maturity and in extreme cases may lead to crop failure.

There are likely to be major changes to the distribution of water under a changed climate. Reductions in amounts of soil water may be compensated for by irrigation, but scarcity of water may preclude the use of (fresh) water for irrigation. Reduced vegetation growth, heavy rainfall events and use of water with high electrolyte content may increase the risk of land degradation, whether in the form of soil erosion, desertification, and or salinization (see below). Current predictions of climate change indicate a likely shift in agricultural zones, requiring a movement of populations, new farming methods, development of new skills, or some combination of these. Soil management will be central to adapting to these new conditions.

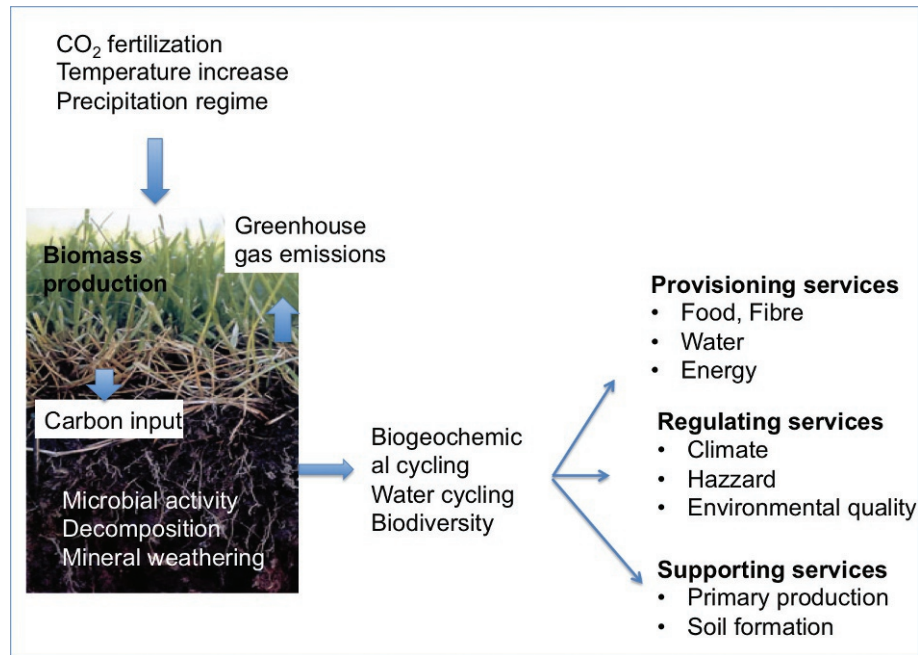


Fig. 2 Effects of global changes on soil processes, biogeochemical cycling and ecosystem services.

Soil also plays an important part as a medium supporting buildings and other infrastructure, and in protecting our archeological heritage. Current infrastructure will almost certainly need to change to meet the new climatic conditions.

Finally, climate change concerns associated with the amounts of greenhouse gases in the atmosphere have focused on the soils' function to store carbon, and their ability to act as a source and or sink for CO₂ and other greenhouse gases. The fact that the soil can act as both a source and sink for several greenhouse gases, e.g., CO₂, N₂O, and CH₄, has recently become an important issue in the quest to reduce the impacts of climate change (see below).

Impact on land-degradation processes

Land degradation is already one of the major problems affecting the world's population. The estimated global extent of the degraded area ranges between <10 to >60 million km² (Gibbs and Salmon, 2015). Currently the estimated amount of soil eroded annually ranges between 35.9 and 75 Pg (Borrelli et al., 2013), desertification affected 270 million ha of global drylands between 1982 and 2015, and salinization is affecting some 833 million hectares of global land area (FAO, 2021). Land degradation through damage to the soil is a serious problem and its causes are often complex and interwoven. Severe damage has already been done to the world's soils, and the impact of climate change needs to be considered in addition to existing pressures on the land. It is difficult to separate these various effects, and their cumulative impact on soils is often greater than a simple summation. Land degradation processes are affecting a large percentage of the global population of emerging economies vulnerable to climate change. Despite the fact that there is confidence that land degradation processes are likely to be exacerbated with ongoing climate change and act as a positive feedback (Fig. 3), there are large uncertainties concerning its extent, severity and links to the climate change (IPCC, 2019).

Soil erosion

Soil erosion is a natural phenomenon, which is controlled by biophysical factors and strongly influenced by land use and especially vegetation cover and agricultural activities. The main erosion process leading to important soil loss in most global environments is hydrological erosion. As global precipitation and especially rainfall amount, intensity and spatiotemporal distribution changes and extreme events, such as heavy rainfall may increase under climate change (IPCC, 2021), soil loss through water erosion may be directly affected (Li and Fang, 2016). There are strong region-specific differences in these impacts, with high and mid-latitude wet regions being most affected. However, hydrological erosion following heavy rainfall events may also increase in dryland areas with sparse vegetation. Moreover, soil erosion may be affected indirectly by climate change through its effects on vegetation and changing landuse practices. Indeed, climate change through its multiple consequences for plant growth and risks including a greater abundance of pests and extreme climatic events leading to crop failure and vegetation dieback may lead to bare soils, which consequently are more exposed to erosion events (Li and Fang, 2016). Moreover, landuse practices are likely to change thereby

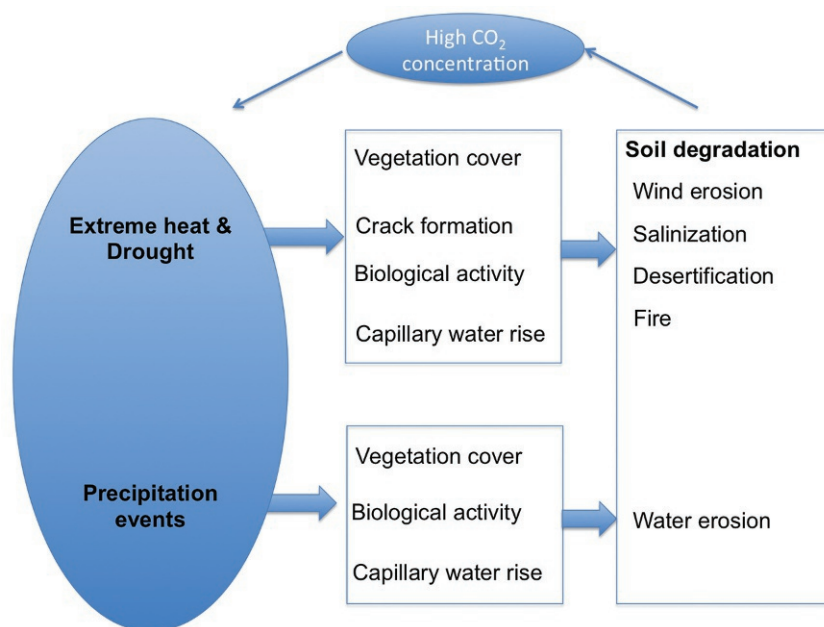


Fig. 3 Relationship between climate change induced extreme events and land degradation.

affecting rates of soil erosion. Another indirect effect leading to increased rates of erosion following climate change is increasing soil salinity (see below). High salt content leads to clay dispersion, which may consequently make the soil more prone to be exported by water erosion. The combination of all of these factors increases uncertainty of the impact of climate change on soil erosion.

Quantification of global change effects on soil erosion and their inclusion into earth system models is complex. As a result, the database of erosion responses to climate change is sparse. In fact, depending on the region, soil type and management, effects of climate change may be positive, negative or non-existing (Li and Fang, 2016). However, using a global database average erosion rates in most cultivated soils were already found to exceed the tolerable rate of $1 \text{ t ha}^{-1} \text{ year}^{-1}$ (García-Ruiz et al., 2015). Climate change is likely to exacerbate soil erosion in regions where it is already severe (Li and Fang, 2016). Therefore, soil protection measures are required including afforestation, conservative tillage regimes and the maintenance of permanent soil cover (Routschek et al., 2014).

Desertification

Desertification is the process of ecological degradation by which economically productive land becomes less productive, in some cases leading to the development of a desert-like landscape. It occurs mainly in hyperarid, arid, semiarid, and subhumid climatic zones, ranging from precipitation: potential evapotranspiration indices (P: PET) of less than 0.05 to 0.70 (IPCC, 2019). The area of land occupied by these four zones in which true or induced deserts can occur is 47% of the planet. With the increased temperatures and evaporation due to climate change, this percentage could increase. There is a huge literature on the nature and causes of desertification, some of which indicates that human impacts arising from overgrazing, overcultivation, and deforestation are primarily responsible for the process, while others lay the blame on the impact of extended droughts over the last few decades. In reality, desertification is likely to be due to a combination of drought and mismanagement of land, particularly where there is a lack of harmony between land use and management on the one hand and prevailing climate on the other. Disentangling the roles of the different factors was identified as a key knowledge gap by the three UN conventions (UNCCD, UNFCCC and UNCBD).

Salinization

The accumulation of salts in soils can negatively affect soil properties and processes. It can lead to degradation of soil structure, decline in porosity, and increase in bulk density, and adversely affect nutrient dynamics by impeding nutrient-holding capacity. It leads to a decline in productivity and can cause land to become unsuitable for agricultural production. Soil salinity is already a major global problem in Australia, Africa, Latin America, and the Near and Middle East (FAO, 2021). It is typically found in areas where evaporation exceeds precipitation and salts accumulate in soil. In the USA alone, some 50 million hectares of cropland and

pasture are affected by salinity, and the area of land so affected is growing by some 10% annually. Salinization is likely to be prevalent in three situations:

- Coastal zones. In these situations salinity is likely to depend primarily on the sea level and its tidal, seasonal, and long-term fluctuations. Given that the global mean sea level is projected to rise by 15–30 cm regardless of emission reductions until the middle of the century (IPCC, 2021), there is likely to be a territorial extension of coastal salinity under the direct and indirect effects of saline seawater. Impacts are likely to include flooding of coastal plains by saline seawater, including low-lying coastal fringes, marshlands and swamps, deltas and estuaries of some of the world's largest rivers, increase in storm tides affecting areas several meters in elevation around the coast, with penetration of saline and brackish water inland, and rapid erosion of coastlines. For example, the soils of the Bangladesh delta, the world's biggest delta, are strongly threatened by salt water intrusion following sea level rise affecting the livelihood of 160 million people.
- Continental salt transport and salt accumulation processes. There are three principal mechanisms of this form of salinization: salt accumulation, seepage, and wind deposition. Salinization by salt accumulation happens when leaching is reduced and salt accumulates at or near the surface. It can also occur when salt is leached into a perched water table and the water moves through the landscape to areas of lower elevation, where the water evaporates, leaving the salts. Given suitably exposed deposits of salts, wind erosion can transfer the salts elsewhere in the landscape. Global salinization hotspots are South America, southern and western Australia, Mexico, southwest United States, and South Africa (Hassani et al., 2021).
- Management induced salinization. Agricultural management, in particular the excessive use of mineral fertilizers and saline irrigation water, can lead to salt accumulation. For example, potassium is applied to soils as potash in the form of KCl (potassium chloride) and has a strong effect on salinization as Cl^- remains in soil pore water or is leached to groundwater (Buvaneshwari et al., 2020). Poor irrigation water quality and or the reuse of degraded water (e.g., drainage water or municipal waste water) are another source for salt addition to agricultural soil. Climate change, especially in dry areas or under drought conditions can lead to an increase of degraded water reuse (Corwin, 2021).

All of these causes of salinity are likely to be enhanced with climate change. Zones subject to rising temperatures during the summer will experience more evapotranspiration and aridity and thus higher concentrations of salt in the soil solution. There may also be enhanced capillary rise from shallow water tables with the potential to carry salt into the overlying soil horizons. Where climate change leads to increasing wetness as well as temperature, there will be a reduction in the concentration of salts in the salt solution. However, where periodic drying is also involved, capillary upward transport of salts can take place from shallow groundwater to overlying soil horizons.

Given the predictions for climate change, the extent of salt-affected soils is likely to increase globally, although there may be a small improvement to a few of the existing saline areas of the world.

Fire

The frequency of fire will increase following climate change due to the occurrence of drought and heatwaves, leading especially in forests to the accumulation of high fuel loads. Mineral soils have poor heat conductivity and fire usually has little effect on their properties. In contrast, fires have strong effects on the degradation of organic soils, which are lost with combustion. Organic soils occur in large extents under boreal forests, constituting the carbon sink recorded for these environments. A recent study showed that shortening of the fire return intervals following climate change may lead to legacy carbon loss, thereby transforming boreal forests from a carbon sink into a carbon source (Walker et al., 2019).

Soil-based climate change mitigation and adaptation options

Options for soil-based climate change mitigation are focused on soil carbon sequestration. Carbon accumulation in soils may occur in organic and also inorganic forms. While SOC sequestration refers to biological processes involving plants and carbon capture through photosynthesis, accumulation of inorganic carbon may occur through silicate weathering and involve pedogenic carbonate formation.

In terms of SOC sequestration, it has been suggested that increasing the global SOC stocks by just a few per mil could effectively counteract the increasing atmospheric CO_2 concentrations (Balesdent and Arrouays, 1999). Because SOC sequestration, a so-called nature based solution for climate change mitigation, has many advantages and is technically ready to be implemented without requiring spare land, it has received a lot of policy attention at the highest level (Rumpel et al., 2020). It led to the launch of the 4 per 1000 initiative at the 21st conference of parties of the United Nations Framework convention for Climate Change, a global initiative with the aim to increase soil carbon to mitigate and adapt to climate change and to increase food security. All three goals may be met because SOC sequestration comes with co-benefits for many ecosystem services and may greatly contribute toward achieving several sustainable development goals. However, quantitative evidence about the relation of soil carbon increases and specific soil properties is scarce although it is well accepted that organic matter enrichment in soil increases its quality. Moreover, SOC

sequestration may have trade offs, such as other greenhouse gas emissions and or water and nutrient requirements (Rumpel et al., 2020).

Therefore, increasing standing SOC stocks should be achieved by the implementation of sustainable soil management practices, which reduce environmental effects such as greenhouse gas emissions, and at the same time ameliorate soil properties, ultimately benefitting soil productivity. Such practices are known, and include best management practices such as permanent soil cover, conservation tillage, organic fertilizer use, conservative fertilization practices, improved grazing management and agroforestry (Paustian et al., 2016). On the other hand, inorganic carbon sequestration in soils may be favored by specific technologies, such as enhanced weathering, and may be able to decrease atmospheric CO₂ concentrations (see above).

It has been pointed out that region-specific soil carbon sequestration practices need to be implemented at scale by millions of farmers to have a global impact as a climate change mitigation strategy (Amelung et al., 2020). To take into consideration the requirements of all actors involved including policy makers, practitioners, scientists, NGOs, businesses and the general public, multistakeholder collaboration is necessary (Rumpel et al., 2018).

Conclusion, summary, outlook

Increasing atmospheric CO₂ has led to changes in global temperature and precipitation regimes. These alterations lead to increasing frequency of extreme events such as drought, heat waves and flooding. All of these changes affect organic and mineral constituents and alter soil properties, biogeochemical cycling and ecosystem services. Their effects may be direct, for example enhanced weathering rates and microbial activity. Others are indirect through plant activity and vegetation growth. Climate change effects are region specific. Whereas in arid climates, salinization and desertification may lead to land degradation as a result of climate change, areas with high precipitation may show increases in soil loss through water erosion. Increased temperatures may increase rates of mineral weathering thereby increasing soil inorganic carbon sequestration. They may also lead to salinization of coastal areas due to sea level rise. Because of the negative consequences of climate change for soil functioning and ecosystem services, especially in vulnerable regions, a global effort is needed to reduce atmospheric CO₂ concentrations. This would need immediate and aggressive emission reduction and negative emission technologies such as SOC sequestration, which if implemented at scale may contribute to combatting climate change while providing many co-benefits because of its positive effect on soil quality.

References

- Amelung W, Bossio D, de Vries W, Kögel-Knabner I, Lehmann J, Amundson R, Bol R, Collins C, Lal R, Leifeld J, Minasny B, Pan G, Paustian K, Rumpel C, Sanderman J, van Groenigen JW, Mooney S, van Wesemael B, Wander M, and Chabbi A (2020) Towards a global-scale soil climate mitigation strategy. *Nature Communications* 11: 1–10.
- Arrhenius SZ (1889) Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren. *Zeitschrift für Physikalische Chemie* 4: 226–248.
- Balesdent J and Arrouays D (1999) Usage des terres et stockage de carbone dans les sols du territoire français. Une estimation des flux nets annuels pour la période 1900–1999. *C. R. Acad. Agric.* 85: 265–277.
- Borrelli P, Robinson DA, Fleischer LR, et al. (2013) An assessment of the global impact of 21st century land use change on soil erosion. *Nature Communications* 8.
- Bossio D, Cook-Patton SC, Ellis PW, et al. (2020) The role of soil carbon in natural climate solutions. *Nature Sustainability* 3: 1–8.
- Bradford M, Wieder W, Bonan G, et al. (2016) Managing uncertainty in soil carbon feedbacks to climate change. *Nature Climate Change* 6: 751–758.
- Buvaneshwari S, Riotte J, Sekhar M, et al. (2020) Potash fertilizer promotes incipient salinization in groundwater irrigated semi-arid agriculture. *Scientific Reports* 10: 3691.
- Corwin DL (2021) Climate change impacts on soil salinity in agricultural areas. *European Journal of Soil Science* 72: 842–862.
- Davidson EA and Janssens IA (2006) Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* 440: 165–173.
- Doetterl S, Alexander J, Fior S, Frossard A, Magabosco C, Van de Broek M, and Westergaard KB (2021) Will accelerated soil development be a driver of Arctic Greening in the late 21st century? *Journal of Plant Nutrition and Soil Science*. <https://doi.org/10.1002/jpln.202100334>.
- FAO (2021) *Global Map of Salt-Affected Soils*. Rome: FAO.
- FAO and ITPS (2015) *Status of the World's Soil Resources (SWSR) – Technical Summary*. Rome, Italy: Food and Agriculture Organization of the United Nations and Intergovernmental Technical Panel on Soils.
- García-Ruiz JM, Beguería S, Nadal-Romero E, González-Hidalgo JC, Lana-Renault N, and Sanjuán Y (2015) A meta-analysis of soil erosion rates across the world. *Geomorphology* 239: 160–173.
- Gibbs HK and Salmon JM (2015) Mapping the world's degraded lands. *Applied Geography* 57: 12–21.
- Hall K, Thorn CE, Matsuoka N, and Prick A (2002) Weathering in cold regions: Some thoughts and perspectives. *Progress in Physical Geography* 26(4): 577–603.
- Hassani A, Azapagic A, and Shokri N (2021) Global predictions of primary soil salinization under changing climate in the 21st century. *Nature Communications* 12: 6663.
- Hicks Pries CE, Castanha C, Porras R, and Torn MS (2017) The whole-soil carbon flux in response to warming. *Science* 355: 1420–1423.
- Hugelius G, et al. (2014) Estimated stocks of circumpolar permafrost carbon with quantified uncertainty ranges and identified data gaps. *Biogeosciences* 11: 6573–6593.
- IPCC (2014) In: Core Writing Team, Pachauri RK, and Meyer LA (eds.) *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, p. 151. Geneva, Switzerland: IPCC.
- IPCC (2019) An IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems. In: Shukla PR, Skea J, Calvo Buendia E, Masson-Delmotte V, Pörtner H-O, Roberts DC, Zhai P, Slade R, Connors S, van Diemen R, Ferrat M, Haughey E, Luz S, Neogi S, Pathak M, Petzold J, Portugal Pereira J, Vyas P, Huntley E, Kissick K, Belkacemi M and Malley J (eds.) *Climate Change and Land*, 874 pp.
- IPCC (2021) In: Masson-Delmotte V, Zhai P, Pirani A, Connors SL, Péan C, Berger S, ... Zhou B (eds.) *Climate Change (2021) The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. In Press.
- Keyser AR and Westerling AL (2019) Predicting increasing high severity area burned for three forested regions in the western United States using extreme value theory. *Forest Ecology and Management* 432: 694–706.
- Kuz'yakov Y, Horwath WR, Dorodnikov M, and Blagodatskaya E (2019) Review and synthesis of the effects of elevated atmospheric CO₂ on soil processes: No changes in pools, but increased fluxes and accelerated cycles. *Soil Biology and Biochemistry* 128: 66–78.
- Leifeld J and Menichetti L (2018) The underappreciated potential of peatlands in global climate change mitigation strategies. *Nature Communications* 9: 1071.

- Li Z and Fang H (2016) Impacts of climate change on water erosion: A review. *Earth Science Reviews* 163: 94–117.
- Moinet GYK, Moinet M, Hunt J, et al. (2020) Temperature sensitivity of decomposition decreases with increasing soil organic matter stability. *The Science of the Total Environment* 704: 135460.
- Paustian K, Lehmann J, Ogle S, et al. (2016) Climate-smart soils. *Nature* 532: 49–57.
- Ping CL, Jastrow JD, Jorgenson MT, et al. (2015) Permafrost soils and carbon cycling. *The Soil* 1: 147–171.
- Plaza C, Pegoraro E, Bracho R, et al. (2019) Direct observation of permafrost degradation and rapid soil carbon loss in tundra. *Nature Geoscience* 12: 627–631.
- Rosenstock NP, van Hees PAW, Fransson PMA, Finlay RD, and Rosling A (2019) Biological enhancement of mineral weathering by *Pinus sylvestris* seedlings – Effects of plants, ectomycorrhizal fungi, and elevated CO₂. *Biogeosciences* 16: 3637–3649.
- Routschek A, Schmidt J, and Kreien Kamp F (2014) Impact of climate change on soil erosion — A high-resolution projection on catchment scale until 2100 in Saxony/Germany. *Catena* 212: 99–109.
- Rumpel C (2019) Soils linked to climate change. *Nature* 572: 442–443.
- Rumpel C, Amiraslani F, Koutika L-S, et al. (2018) Put more carbon in soils to meet Paris climate pledges. *Nature* 564: 32–34.
- Rumpel C, Amiraslani F, Chenu C, et al. (2020) The 4p1000 Initiative: opportunities, limitations and challenges for implementing soil organic carbon sequestration as a sustainable development strategy. *Ambio* 49: 350–360.
- Sanderman J, Hengl T, and Fiske GJ (2017) Soil carbon debt of human land use. *Proceedings of the National Academy of Sciences* 114: 9575–9580.
- Schlesinger WH and Bernhardt ES (2013) *Biogeochemistry: An Analysis of Global Change*. Waltham, MA: Elsevier/Academic Press.
- Smith P, Cotrufo MF, Rumpel C, et al. (2015) Biogeochemical Cycles and Biodiversity as key drivers of ecosystem services. *Soil* 1: 665–685.
- Soong JL, Castnaha C, Hicks-Pries CE, et al. (2021) Five years of whole-soil warming led to loss of subsoil carbon stocks and increased CO₂ efflux. *Science Advances* 7: eabd1343.
- Taylor L, Quirk J, Thorley R, et al. (2016) Enhanced weathering strategies for stabilizing climate and averting ocean acidification. *Nature Climate Change* 6: 402–406.
- Von Lützow M and Kögel-Knabner I (2009) Temperature sensitivity of soil organic matter decomposition—What do we know? *Biology and Fertility of Soils* 46: 1–15. <https://doi.org/10.1007/s00374-009-0413-8>.
- Walker XJ, Blatzer JL, Cumming SG, et al. (2019) Increasing wildfires threaten historic carbon sink of boreal forests. *Nature* 572(7770): 520–523.
- Walker AP, De Kauwe MG, Bastos A, et al. (2021) Integrating the evidence for a terrestrial carbon sink caused by increasing atmospheric CO₂. *New Phytologist* 229: 2413–2445.

Primary minerals

Balwant Singh, Sydney Institute of Agriculture, School of Life and Environmental Sciences, The University of Sydney, NSW, Australia

© 2023 Elsevier Ltd. All rights reserved.

This is an update of P.M. Huang, M.K. Wang, MINERALS, PRIMARY, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 500-510, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00464-1>.

Introduction	91
Silica	93
Quartz	93
Solubility of silica minerals	94
Silica as an indicator of soil parent material	96
Feldspars	96
Weathering of feldspars	99
Micas	100
Weathering of micas	101
Common accessory primary minerals	102
Olivines	102
Weathering of olivine	102
Pyroxenes and amphiboles	103
Weathering of pyroxenes and amphiboles	106
Conclusions and future perspectives	107
References	107

Abstract

Major primary minerals in soils include feldspars, quartz, micas, and olivines, amphiboles and pyroxenes as accessory minerals. Weathering of these minerals serves as a template for various secondary minerals and to supply elements for the crystallization of secondary minerals in soil environments. Primary minerals are the source of many essential plant nutrients and also provide pH buffering capacity to soils. The weatherability of primary minerals in soils depends on their structural (e.g., points of weakness in the structure) and chemical properties (e.g., presence of Fe²⁺ iron), and environmental conditions. The structure and composition of common primary minerals and their weathering products in soils are described in this chapter.

Key points

- Feldspars, quartz, and micas are major primary minerals in soils.
- Olivines, pyroxenes and amphiboles are accessory minerals.
- The weathering of primary minerals produces important secondary minerals in soils.
- Several essential nutrients are released from the weathering of primary minerals.
- Primary minerals contribute to soil pH buffering capacity and some easily weatherable minerals such as forsterite has potential to contribute to carbon sequestration in soils

Introduction

Primary minerals are formed at elevated temperatures and inherited from igneous and metamorphic rocks, sometimes through a sedimentary cycle. These minerals have not been altered chemically in soils since their crystallization and occur as residuals of weathering processes. In most soils, primary minerals exist in the sand and silt fractions, however, they may be present in the clay fraction of weakly weathered soils. Commonly, primary minerals are minor constituents of the clay fraction of most agricultural soils. The weathering of primary minerals is an important process for the global hydrogeochemical cycling of elements. The process is equally vital in soils, because it releases many essential plant nutrients, such as calcium (Ca), magnesium (Mg), potassium (K), phosphorus (P), iron (Fe), manganese (Mn) and boron (B). The dissolution of minerals and subsequent release of elements determines the chemical composition of soil, soil solutions and terrestrial ecosystems.

Primary minerals are the main source of base cations and their weathering buffer pH and neutralize acid input in the soil. The weathering of primary silicate minerals consumes carbon dioxide (CO₂) through the generation of bicarbonate alkalinity in stream water. In the last few years there has been an increased focus on the potential use of primary minerals such as forsterite to accelerate CO₂ sequestration in the soil to mitigate climate change (Schuiling and Krijgsman, 2006; Köhler et al., 2010).

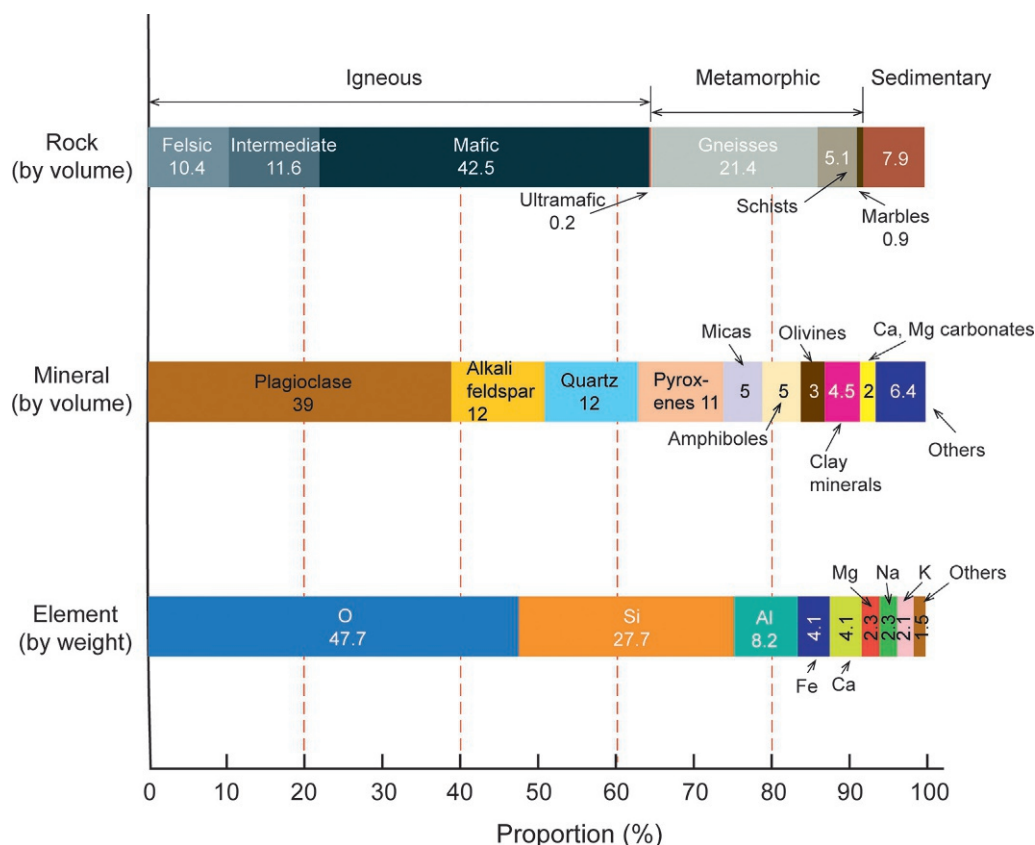


Fig. 1 Abundance of rocks, minerals and elements in the Earth's crust (rocks and minerals data from Ronov and Yaroshevsky, 1969; and element composition from Bowen, 1979).

Primary minerals in soils are derived from their parent materials, such as rocks, river-borne sediment (alluvium), wind-borne sediment (eolian), hillslope sediment (colluvium), lakebed sediment (lacustrine), materials moved by glacial advance and retreat (glacial till), and volcanic ash. The main source of most unconsolidated soil parent materials is one of the rocks. The abundance of rocks, minerals and elements in the Earth's crust is given in Fig. 1. Igneous and metamorphic rocks represent over 92% (on volume basis) of the crust. The clays, shales and to a lesser extent carbonates and sands are the main constituents in the thin veneer of sediments present at the Earth's surface (Wyllie, 1971). Feldspars and quartz are the dominant minerals constituting 63% (on volume basis) of the crust, whereas pyroxenes, amphiboles and olivines comprise 19% and micas 5%. Other minor constituents in the crust include 4.6% clay minerals (including chlorite), 2.0% Ca/Mg carbonates, 1.5% magnetite (and titanomagnetite) and 4.9% other minerals. Oxygen (O), silicon (Si) and aluminum (Al) are three most abundant elements, constituting 47.7, 27.7 and 8.2%, respectively (on weight basis) of the crust. Other elements present in significant amounts are Fe, Ca, Mg, sodium (Na) and K with an average concentration of 4.6, 4.1, 2.3, 2.3 and 2.1%, respectively.

Similar to their abundance in the Earth's crust, feldspars and quartz are the most abundant primary minerals in soils. Both feldspars and quartz are tectosilicates, with a continuous three-dimensional framework of tetrahedra. Quartz is essentially pure SiO_2 , that contains tetrahedra formed by Si and O only, and it is the main form of free silica in soils. Feldspars are anhydrous aluminosilicates in which both Si and Al tetrahedra share all four O atoms with neighboring tetrahedra. Feldspars contain varying amounts of Na, K, and Ca, and occasionally of other cations such as barium (Ba), which balance the negative charge resulting from the substitution of Al^{3+} for Si^{4+} in their structure.

Feldspars occur in virtually all sediments and soils in quantities that vary with the nature of the parent materials and stage of weathering. The weathering of feldspars is an important terrestrial geochemical process. Feldspars are an important source of the plant nutrients K and Ca; they play a substantial role in overall K dynamics in soils.

Micas are 2:1 phyllosilicates with tightly held, non-hydrated, interlayer cations (mostly K) balancing a high layer charge. They are largely of primary origin, occur extensively, and serve as precursors for expansible 2:1 phyllosilicates such as vermiculites and smectites in soil environments. Micas are often present in soils as components of particles that have been only partially transformed to expansible 2:1 minerals through inter-stratification with the other minerals or the formation of mica cores that are surrounded by expanded zones. Micas are important reservoirs of K for plants in soils. The K-supplying power of a soil depends to a large extent on the nature and amount of micas and their dynamics of K release.

Olivines, pyroxenes, and amphiboles are important accessory primary minerals that occur in small but significant amounts. Olivines are olive-green nesosilicates in which Mg and Fe^{2+} are octahedrally coordinated by O atoms. Pyroxenes and amphiboles,

which are inosilicates, are ferromagnesian minerals with single- and double-chain structures of linked silica tetrahedra, respectively. The variety of isomorphous substitution in olivines, pyroxenes, and amphiboles and their relative ease of weathering make them excellent source minerals for Ca, Mg, and trace elements in soils.

Silica

Silica is a chemical term or expression for silicon dioxide (SiO_2) and is often used to describe a range of structures and properties. Silica occurs in nature in at least 15 polymorphs including α -quartz, β -quartz, α -cristobalite, β -cristobalite, α -tridymite, β -tridymite, coesite, stishovite, moganite, keatite, lechatelierite (silica glass), cryptocrystalline silica (chalcedony) and opal (Deer et al., 1992; Heaney, 1994; Guthrie and Heaney, 1995; Götze et al., 2021). Some of the silica polymorphs are metastable and amorphous.

Important properties of common forms of silica are presented in Table 1. Quartz is the most abundant of the silica polymorphs in soil environments. Disordered cristobalite commonly occurs in soils, but tridymite is rare and usually associated with siliceous volcanic rocks. Coesite, stishovite, and lechatelierite, which are formed at high pressures, are rare polymorphic forms in soils. Opal is a generic term for hydrated and amorphous silica, is of biogenic origin, and although minor it is a ubiquitous constituent of soils.

Quartz

Quartz occurs as two polymorphs – α -quartz (low quartz) and β -quartz (high quartz). The β -quartz is only stable at temperatures greater than about 573 °C. Below this temperature it converts to α -quartz, which is stable under surface conditions, therefore, only α -quartz (from here on quartz) occurs in soil. Quartz is one of the last minerals to crystallize from magma. Therefore, it is formed under conditions closer to those present on the Earth's surface than minerals crystallized earlier (at higher temperatures), which contributes to its high thermodynamic stability. Dense packing of tetrahedra in the crystal structure and high activation energy required to alter the Si—O—Si bond are the main factors that contribute to the strong stability of quartz compared to other silica polymorphs. Quartz is in almost all soils and often represents the major portion of all sand and silt fractions. It is also a significant component of the coarse clay fraction of most soils.

The structure of quartz can be visualized as a spiral network of silica tetrahedra about the Z-axis (Fig. 2A). Each tetrahedron is repeated in the network by a rotation of 120° and a translation of one-third about the Z-axis. The tetrahedra are linked to form a hexagonal structure in quartz. The silica tetrahedron of quartz is almost symmetrical and has a Si—O distance of 0.161 nm. The Si—O bond is approximately equally ionic and covalent in nature. The compact structure of quartz suggests that there is limited

Table 1 Some properties of silica minerals found in soil environments.

Mineral Group and name	Formula unit	Crystal system and unit cell	Specific gravity (Mg m^{-3})	Major X-ray diffraction peaks (nm) and intensity (%) in parentheses	Features and presence in soil
α -Quartz	SiO_2	Trigonal $a = 0.4913 \text{ nm}$ $c = 0.5404 \text{ nm}$ $Z = 3$	2.65	0.426 (35), 0.334 (100), 0.246 (12), 0.228 (12), 0.182 (17)	Colorless, white or variable black, purple, green with impurities. Ubiquitous in all soil types, often constitutes major portions in the sand and silt size fractions.
Cristobalite	SiO_2	Trigonal $a = 0.497 \text{ nm}$ $c = 0.692 \text{ nm}$ $Z = 4$	2.33	0.404–0.411 (100), 0.313–0.318 (12), 0.283–0.285 (14), 0.248–0.250 (20)	Colorless, white or yellowish. Occurs in volcanic rocks and co-exist with tridymite.
Tridymite	SiO_2	Orthorhombic $a = 0.504 \text{ nm}$ $b = 0.874 \text{ nm}$ $c = 0.824 \text{ nm}$ $Z = 8$	2.26	0.430–0.433 (90), 0.409–0.411 (100), 0.380–0.382 (50), 0.296–0.298 (25)	Colorless or white. Typically occurs in acid volcanic rocks and abundant in rhyolitic tuffs.
Opal-CT (cristobalite/tridymite)	$\text{SiO}_2 \cdot n\text{H}_2\text{O}$	About 40% non-crystalline structures mixed with crystalline ones related to tridymite.	2.20–2.38	~0.430 (30–40), 0.404–0.411 (100), 0.313–0.318 (12), 0.283–0.285 (14), 0.248–0.250 (20)	Varied, cryptocrystalline blades aggregated into lepispheres that gives color and translucency. Commonly associated with volcanic deposits.
Opal-A (amorphous)	$\text{SiO}_2 \cdot n\text{H}_2\text{O}$		1.50–2.30	Diffuse, broad peak centered at ~0.41 nm	Opaque, opal-A in soils is of biogenic (organic) origin.

The data from Drees et al. (1989), Deer et al. (1992), and Fröhlich (2020).

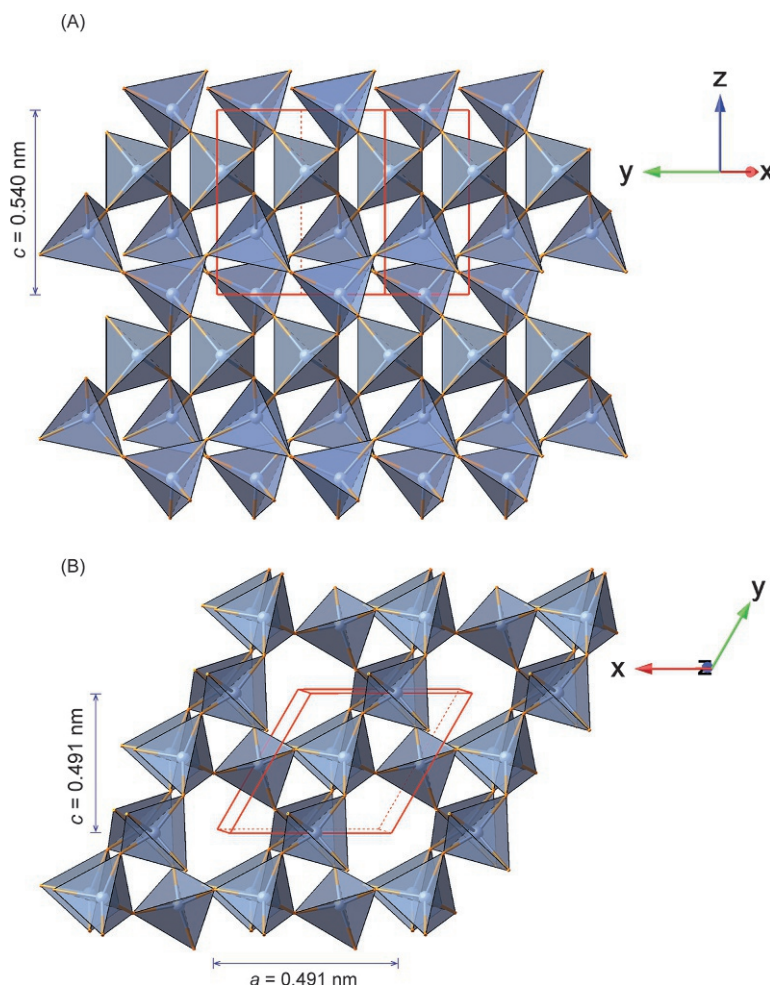
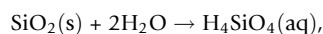


Fig. 2 The structures of low quartz (α -SiO₂) showing spiral network of Si—O₄ tetrahedra. (A) the structure projected along the Z-axis, each tetrahedron is repeated in the network by a rotation of 120° and the linked tetrahedra form a hexagonal structure in quartz. (B) the intertwined chains when projected parallel to the Z-axis show open channels that are wide enough to accommodate cations such as Al³⁺ and Na⁺.

opportunity for the incorporation of impurity elements, and these usually occur in traces to very minor amounts (Wilson, 2020). However, Fig. 2B shows that the structure is characterized by channels running parallel to the Z-axis, which are wide enough to accommodate other cations either interstitially or as substituents. Aluminum is the most abundant impurity in natural quartz, with concentrations ranging from <13 to 15,000 mg kg⁻¹ (Smith and Steele, 1984). Its presence in quartz relates to its common occurrence in the Earth's crust and the similar ionic radii of Si⁴⁺ and Al³⁺ (Götze, 2009 in Götze et al., 2021). Trace amounts of Ti⁴⁺ (titanium) and Fe³⁺, with comparably large ionic radii, may also occur in marginal parts of quartz crystals. Other monovalent cations, such as H⁺ (hydrogen), Li⁺ (lithium), Na⁺, K⁺, Cu⁺ (copper) or Ag⁺ (silver) can be incorporated in interstitial positions, sometimes acting as charge-compensating ions typically maintaining electrical neutrality. Lithium is frequently incorporated into the quartz structure, whereas Na and K may be hosted by fluid inclusions (Gotze et al., 2004 in Götze et al., 2021). Smith and Steele (1984) observed a range for Na from 9 to 1420 mg kg⁻¹ and for K from 3 to 290 mg kg⁻¹. Plots of the contents of Na + K against Al + Fe showed a 1:1 correspondence (Graetsch et al., 1987).

Solubility of silica minerals

The dissolution of silica in water can be described as follows:



where s and aq represent solid and aqueous phases.

The solubility of silica minerals decreases as a function of increasing packing density of the silica tetrahedra and long-range crystal order. Solubility of quartz at ambient temperatures and neutral pH is commonly 3–7 mg Si L⁻¹, which is much lower than that of amorphous silica (50–60 mg Si L⁻¹). Freshly ground quartz commonly shows abnormally high solubility (37 mg Si L⁻¹).

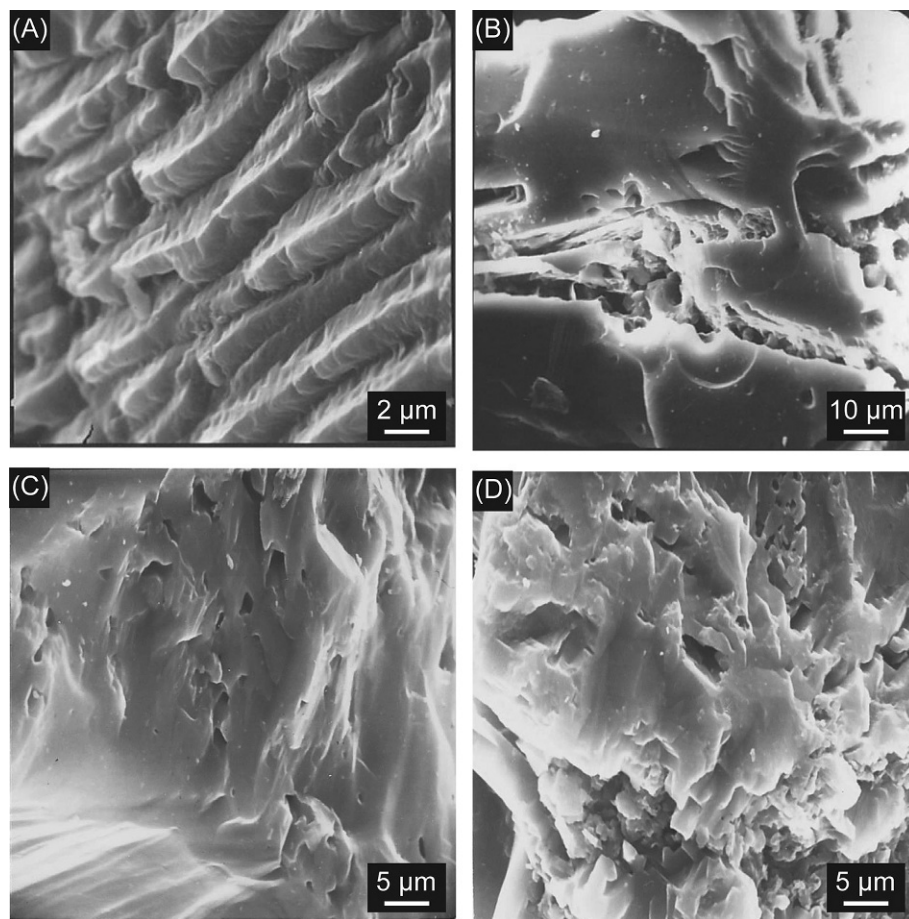


Fig. 3 SEM images of quartz grains showing (A) distinctive etched steps in the Eg (eluviated or bleached) horizon, and (B) deeply etched pits and linear trenches in the B/C horizon of from a peaty gleyed Spodosol developed on fluvio-glacial sands in north-east Scotland. (C) quartz grains showing intensive etching and (D) etched and disintegrating grains from the Cg (strong gleying) horizon of an imperfectly drained iron Spodosol developed on granitic material. From Wilson MJ (2020) Dissolution and formation of quartz in soil environments: A review. *Soil Science Annual* 71(2): 99–110.

This has been attributed to the formation of a disrupted surface layer, which is believed to be amorphous or microcrystalline. These solubility determinations were made using bulk methods so that the dissolution mechanisms were deduced indirectly. Since the 1980s, however, emphasis has been on the processes, such as development of etch pits at structural dislocations and defects on the grain surfaces, which might have a major control on the overall dissolution kinetics of the mineral (Wilson, 2020). Many researchers have observed abundant etch pitting of quartz where significant dissolution has occurred in the upper part of the saprolite, which was related to undersaturation of solute silica with respect to quartz solubility in the pore water. There is a critical concentration of dissolved silica in the external solution, above which etch pits will not form or form at a reduced rate. However, Wilson (2020) observed extensive corrosion and etching of quartz at all depths within soil profiles from north-east Scotland (Fig. 3), which could have been inherited from the effects of former weathering cycles.

The dissolution of quartz in soils is a function of temperature, pH, and particle size. It increases with increasing temperature and the rate of dissolution also increases with decrease in particle size, i.e., increasing specific surface area. With respect to pH, the minimum rate of dissolution at 25 °C ($10^{-12.7}$ mol quartz $\text{m}^{-2} \text{s}^{-1}$) occurs in the pH range 4–6, which is close to the point of zero charge of quartz (Brantley, 2005). The rate of dissolution increases linearly under both acid and alkaline conditions, with the rate being $10^{-12.0}$ mol quartz $\text{m}^{-2} \text{s}^{-1}$ at pH 2 and $10^{-11.0}$ mol quartz $\text{m}^{-2} \text{s}^{-1}$ at pH 10, indicating that quartz is more soluble under alkaline conditions.

In addition, the concentrations of alkali and alkaline-earth cations, organic acids and Al and Fe species in the external solution also affect the dissolution of quartz. The increased rate of dissolution of quartz in aqueous solutions containing small concentrations of common alkali and alkaline-earth cations (such as Na^+ , K^+ , Ca^{2+} and Mg^{2+}) has been attributed to increased rates of solvent exchange at the mineral–solution interface and rupture of Si—O bonds (Dove and Nix, 1997; Dove and Han, 2007; Dewan et al., 2013). The enhanced dissolution of quartz by organic acids has been suggested to be caused by the effective complexation and mobilization of silica, which decreases the activity of dissolved silicic acid and leads to a far-from equilibrium system. The dissolved species of Al and Fe have been shown to decrease the solubility of quartz, perhaps by forming coatings on quartz surfaces.

Silica as an indicator of soil parent material

Oxygen-isotope composition of minerals is a useful tool for determining the genesis and history of soil parent materials. The oxygen isotope composition of quartz depends on the temperature of formation, with delta values ($\delta^{18}\text{O}$) decreasing as the temperature of formation increases. The delta value ($\delta^{18}\text{O}$) is calculated as follows:

$$\delta^{18}\text{O} = \left[\frac{(\delta^{18}\text{O} - \delta^{16}\text{O})_{\text{sample}}}{(\delta^{18}\text{O} - \delta^{16}\text{O})_{\text{SMOW}}} - 1 \right] \times 1000,$$

where SMOW is standard mean ocean water.

Quartz in igneous rocks has $\delta^{18}\text{O}$ values in the order of 8–13‰, whereas quartz from metamorphic rocks has somewhat higher $\delta^{18}\text{O}$ values, with a wider range that results from varying degrees of metamorphism. Oxygen–isotope ratios of quartz in shales range from 15 to 24‰; the ratios of quartz in sandstones range from 10 to 16‰. Silica minerals, such as cherts, diatoms, plant opal, which are formed at ambient temperatures have considerably higher $\delta^{18}\text{O}$ values ranging up to 35‰ (Allègre, 2008). The intermediate values of sedimentary deposits are attributed to the mixed origins of sediments. Delta values of quartz from soils range from 9 to 30‰, which are generally similar to those of sedimentary rocks. The large value ($\sim 30\text{‰}$) suggests that quartz may have formed in the soil as a secondary mineral (Drees et al. (1989). There is some evidence in the literature that supports this at ambient temperatures in soils (Drees et al., 1989; Wilson, 2020).

Quartz is exceedingly resistant to O-isotope exchange; therefore, its isotopic composition in soils, dusts, and sediments has been used as an indicator of provenance. The main use of isotopic composition of quartz in pedologic studies has been for identification of eolian additions to soils.

Feldspars

Feldspars comprise 595 g kg^{-1} of igneous rock, 300 g kg^{-1} of shale, and 115 g kg^{-1} of sandstone on average. Many metamorphic rocks also contain large amounts of feldspars. Therefore, feldspars are present in virtually all sediments and soils; their quantities vary with the nature of parent material and the extent of weathering of the parent material. Feldspars play a substantial role in the overall dynamics of macronutrients K and Ca in soils.

The structure of feldspars is a three-dimensional framework of corner-sharing SiO_4 and AlO_4 tetrahedra, with sufficient opening in the framework to accommodate K, Na, Ca, or Ba to maintain electroneutrality. In the structure of the framework, four-membered rings of tetrahedra are the basic units. Chains of four-membered rings of tetrahedra are parallel to the X-axis and cross-linked to adjacent parallel chains by shared O atoms (Fig. 4). One zigzag chain is formed by superimposing four-membered rings of tetrahedra that share some of their vertices to form new, four-membered rings. The remaining vertices form cross-links between

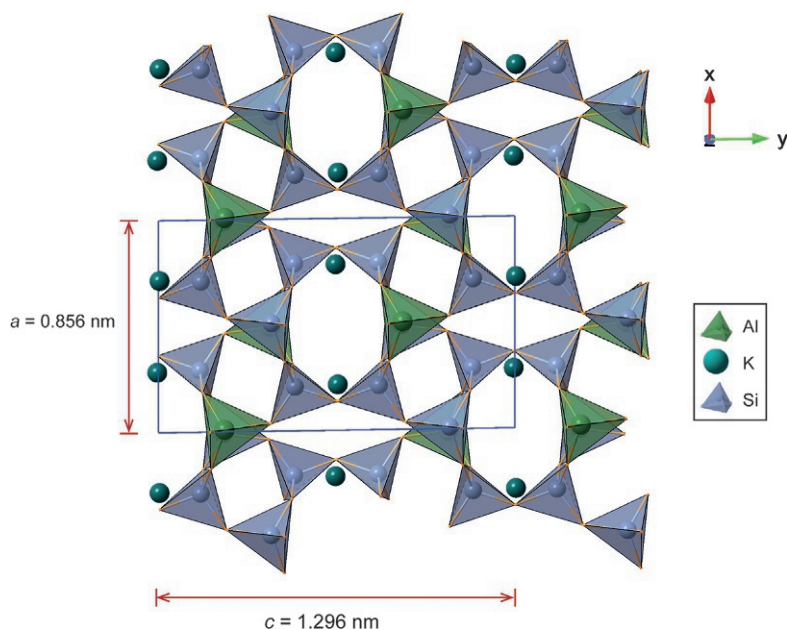


Fig. 4 The structure of microcline feldspar projected along the (001) plane. Silicon and Al tetrahedra are shown in light blue and green colors, respectively, the corners of tetrahedra represent O, and the bluish-green color spheres represent K.

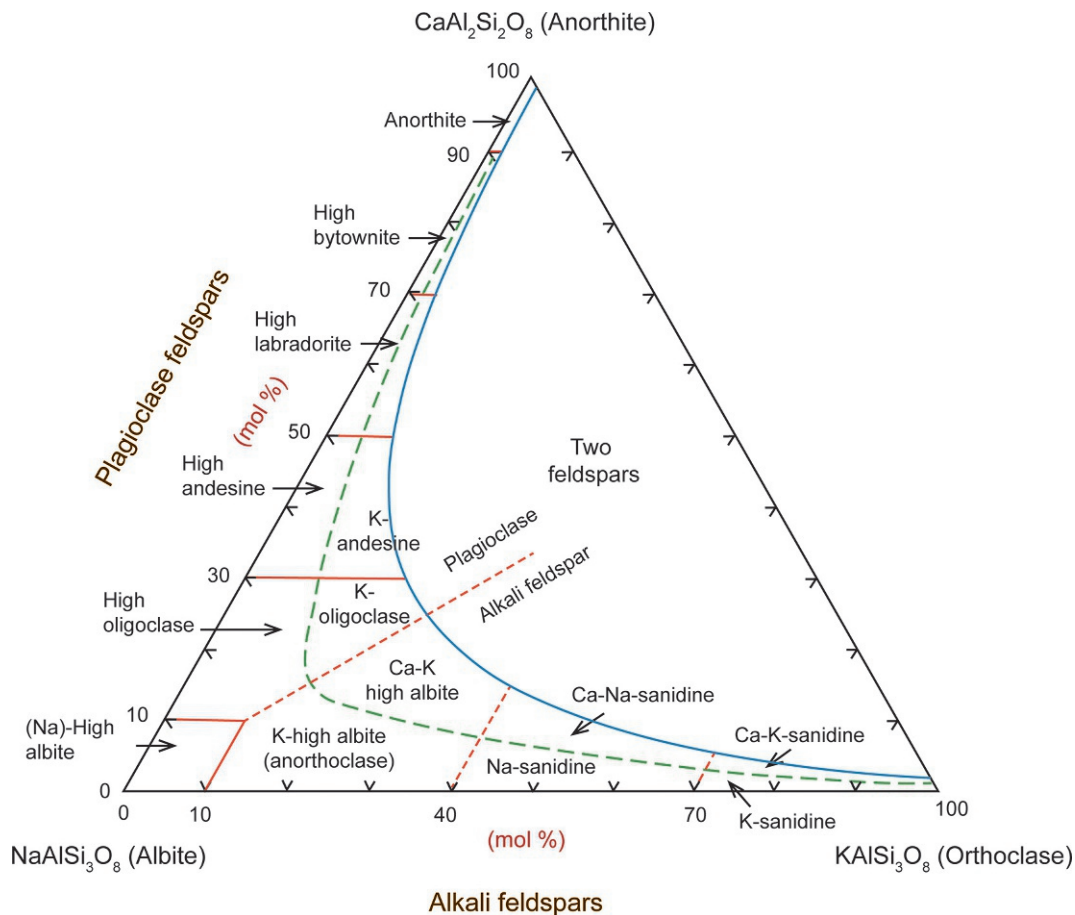


Fig. 5 Ternary diagram illustrating composition and nomenclature of feldspars in the orthoclase–albite–anorthite – names based on their chemical compositions, i.e., relative amounts of K, Na, and Ca, respectively. The points of the triangle are so-called ‘end member’ compositions. From Smith JV (1974) *Feldspar minerals: I. Crystal Structure and Physical Properties*. Berlin: Springer-Verlag.

chains. The repeat periodicity along the chain, i.e., the *a*-dimension of the unit-cell is 0.856 nm in the microcline structure (Fig. 4). The *a*-dimension (along the *X* axis) of different feldspars varies from 0.81 to 0.86 nm.

Feldspars represent solid-solutions between three end members orthoclase (potassium feldspar, $K(AlSi_3O_8)$), albite (sodium feldspar, $Na(AlSi_3O_8)$) and anorthite (calcium feldspar, $Ca(Al_2Si_2O_8)$) as shown in Fig. 5 (Smith, 1974). Albite and anorthite form a completely miscible series called plagioclase. The plagioclase feldspars are arbitrarily divided into six minerals based on their mol (%) proportion of albite (Ab) and anorthite (An) (Fig. 5): albite ($Ab_{100} - Ab_{90}An_{10}$), oligoclase ($Ab_{90}An_{10} - Ab_{70}An_{30}$), andesine ($Ab_{70}An_{30} - Ab_{50}An_{50}$), labradorite ($Ab_{50}An_{50} - Ab_{30}An_{70}$), bytownite ($Ab_{30}An_{70} - Ab_{10}An_{90}$), and anorthite ($Ab_{10}An_{90} - Ab_0An_{100}$). These boundaries have no structural significance. Their use is justified because plagioclase feldspars are the most common and abundant minerals in the Earth’s crust (Fig. 1) and their composition is rather predictable. In the sandine field in Fig. 5, the feldspars ($Or_{50}-Or_{80}$ or $< Or_{50}$) have monoclinic symmetry and K-high albite are triclinic at room temperature (Table 2). Pure albites, both low- and high-temperature forms, have triclinic symmetry. Sodic plagioclase (oligoclase) feldspars are more common in granite, Ca-rich feldspars (labradorite) in mafic rocks (e.g., gabbro) and intermediate plagioclase such as andesine exists more in intermediate igneous rocks (e.g., andesite).

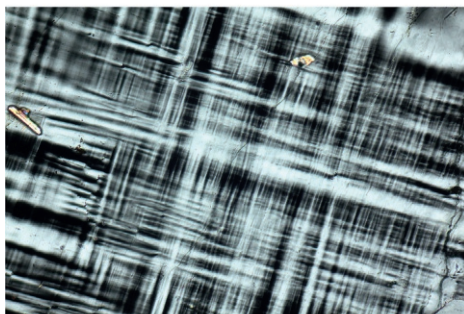
Alkali feldspars have a range of chemical compositions between the two end members, i.e., $KAlSi_3O_8$ and $NaAlSi_3O_8$. The K-feldspar polymorphs, sanidine, orthoclase, microcline, and adularia have identical chemical compositions. Both, orthoclase and sanidine are monoclinic K-feldspars. Sanidine occurs in volcanic and subvolcanic rocks, whereas orthoclase is more common in plutonic rocks. Microcline is a triclinic K-feldspar and has a characteristic cross-hatched tartan pattern of twinning observed under a petrographic microscope (Fig. 6A). Adularia occurs in low-temperature hydrothermal veins. Anorthoclase (K-high albite) is the only alkali feldspar that is not a K-feldspar. It is a Na-rich feldspar that is triclinic and shows very fine cross-hatched twinning.

The high-temperature series of mixed crystals is quite continuous between the end-member orthoclase and albite, and also between albite and anorthite. However, under low-temperature conditions, these phases undergo structural changes and tend to segregate. The segregation results in the formation of lamellar aggregates whose composition is similar to that of the end members. The lamellar aggregates, which are composed of considerable amounts of alkali feldspar and albite, are referred to as ‘perthite.’ Except for authigenic K-feldspar, which is formed through reconstitution and precipitation, most alkali feldspars usually contain

Table 2 Unit-cell dimensions, chemical formulae and specific gravity of end members of feldspars.

Mineral name	Ideal chemical formula	Crystal system and unit cell	Specific gravity (Mg m^{-3})
Orthoclase	KAlSi_3O_8	$a = 0.856 \text{ nm}$, $b = 1.300 \text{ nm}$, $c = 0.719 \text{ nm}$ $\alpha = 90^\circ$, $\beta = 116.01^\circ$, $\gamma = 90^\circ$	2.56
Sanidine	KAlSi_3O_8	$a = 0.860 \text{ nm}$, $b = 1.303 \text{ nm}$, $c = 0.718 \text{ nm}$ $\alpha = 90^\circ$, $\beta = 116^\circ$, $\gamma = 90^\circ$	2.56
Microcline	KAlSi_3O_8	$a = 0.859 \text{ nm}$, $b = 1.297 \text{ nm}$, $c = 0.722 \text{ nm}$ $\alpha = 90.6^\circ$, $\beta = 115.9^\circ$, $\gamma = 87.7^\circ$	2.56
Low albite	$\text{NaAlSi}_3\text{O}_8$	$a = 0.814 \text{ nm}$, $b = 1.279 \text{ nm}$, $c = 0.716 \text{ nm}$ $\alpha = 94.3^\circ$, $\beta = 116.6^\circ$, $\gamma = 87.7^\circ$	2.61
High albite	$\text{NaAlSi}_3\text{O}_8$	$a = 0.816 \text{ nm}$, $b = 1.287 \text{ nm}$, $c = 0.711 \text{ nm}$ $\alpha = 93.5^\circ$, $\beta = 116.4^\circ$, $\gamma = 90.3^\circ$	2.61
Anorthite	$\text{CaAl}_2\text{Si}_2\text{O}_8$	$a = 0.818 \text{ nm}$, $b = 1.288 \text{ nm}$, $c = 1.417 \text{ nm}$ $\alpha = 93.2^\circ$, $\beta = 115.8^\circ$, $\gamma = 91.2^\circ$	2.76

(A) Cross-hatched twinning in microcline



(B) Perthite in microcline

**Fig. 6** (A) Cross-hatched twinning in microcline. $10\times$ (Field of view = 2 mm). (B) Visible perthitic exsolution lamellae of albite in microcline. Evje, Norway. Width of sample 6 cm. (A) From: <http://www.alexstrekeisen.it/english/pluto/alkalifeldspar.php>. (B) From: <https://www.sandatlas.org/feldspar/>.

varying amounts of Na in their structure and are thus more or less perthitic. The morphology of soil perthite crystals with a second phase, most likely albite, in a host crystal of K-feldspar is shown in Fig. 6B. The chemical composition of feldspars varies widely. Reasonable amounts of elements other than Na, Ca and K, may be present in feldspars. Other elements that occur in trace to small amounts in feldspars include Ti (titanium), iron (both in Fe^{3+} and Fe^{2+} forms), Mn, Mg, Ba and Sr (strontium).

Feldspars are present in nearly all soils, but their quantities vary with the intensity and capacity factors of weathering reactions. They are absent or in only small quantities in strongly weathered soils, even though the parent material contains considerable quantities of those minerals. Some feldspars may occur in humid tropical soils that contain relatively fresh rock materials due to erosional and depositional processes. Feldspars are common in the silt and sand fractions of young to moderately developed soils representing various soil parent materials and soil-forming conditions. Alkali feldspars are even present in the clay fraction of moderately weathered soils.

Weathering of feldspars

The weathering of feldspars occurs at structural defects, particularly at the boundary of twinned and untwinned domains. Both nonstoichiometric and stoichiometric dissolution behaviors have been reported in controlled studies using different feldspars, and depend on the mineral (such as, defects, exsolution) and solution properties (e.g., pH and ionic strength) (Casey et al., 1989; Muir et al., 1989; Nesbitt et al., 1991; Welch and Ullman, 1993; Stillings and Brantley, 1995). Casey et al. (1989) showed nonstoichiometric dissolution of labradorite in acid solutions (pH 1 to 3), where H^+ ions infiltrated to a depth of several tens of nanometers, selectively removed Na^+ , Ca^{2+} and Al^{3+} and produced an amorphous Si-rich layer at the mineral surface. At higher pH values, however, only limited penetration of H^+ ions was observed, and dissolution occurred at the immediate mineral surface. With secondary ion mass spectrometry (SIMS) and X-ray photoelectron spectroscopy (XPS), Muir et al. (1989) produced similar conclusions from labradorite dissolution in solutions of pH 3.5 and 5.7. Stillings and Brantley (1995) observed stoichiometric dissolution for microcline and bytownite and nonstoichiometric for oligoclase and labradorite at 25 °C and pH 3.0. The preferential release of Ca and Al in oligoclase was attributed to the impurity phases, and apparent nonstoichiometric dissolution in labradorite possibly occurred due to its exsolution texture, that caused the preferential Ca and Al release. Chou and Wollast (1984) studied the dissolution of albite with a continuous flow reactor and observed three steps in the dissolution of fresh feldspar. The first step involves instantaneous and reversible rapid exchange of Na^+ with H^+ , which is followed by the build-up of a residual layer that is depleted in Na and also Al under acidic conditions. Finally, at a steady-state a congruent dissolution stage is formed, where the rate is controlled by a surface reaction between the residual layer and the solution.

The rates of dissolution for the main feldspar minerals in the acidic pH range from the data compiled by Blum and Stillings (1995) are presented in Fig. 7. There is a strong dependence of mineral dissolution on solution pH, which reflects the importance of a proton-promoted dissolution mechanism. There is general agreement that feldspar dissolution occurs by protonation of Al surface sites in the tetrahedral framework. The dissolution behavior of feldspars is consistent with Goldich's relative order of stability of the feldspar minerals to natural weathering: anorthite < bytownite < labradorite < andesine < oligoclase < albite < K-feldspars. The dissolution data show the importance of the chemical characteristics of minerals, with Ca, Al-rich plagioclases dissolving faster than Na, Si-rich plagioclases in similar solutions.

The rates of experimental dissolution of albite and K-feldspar are quite similar (Fig. 7A and B), which contrasts with observations in most soils, including acidic soils, where K-feldspars commonly weather more slowly than sodic plagioclase. This has been explained by the soil porewater chemistry, which tends to be closer to saturation with respect to K-feldspars than albite (Blum and Stillings, 1995).

Organic acids have been shown to have enhanced feldspar dissolution in soils, however, there is still some debate on this subject. Ligands of complexing organic acids may play a significant role in feldspar dissolution, particularly their ability to complex Al from the minerals. At concentrations of organic acids that might exist in natural systems, their effect on feldspar dissolution is generally small. Drever and Stillings (1997) observed that the effect of 1 mM oxalic acid on feldspar dissolution ranged from none to

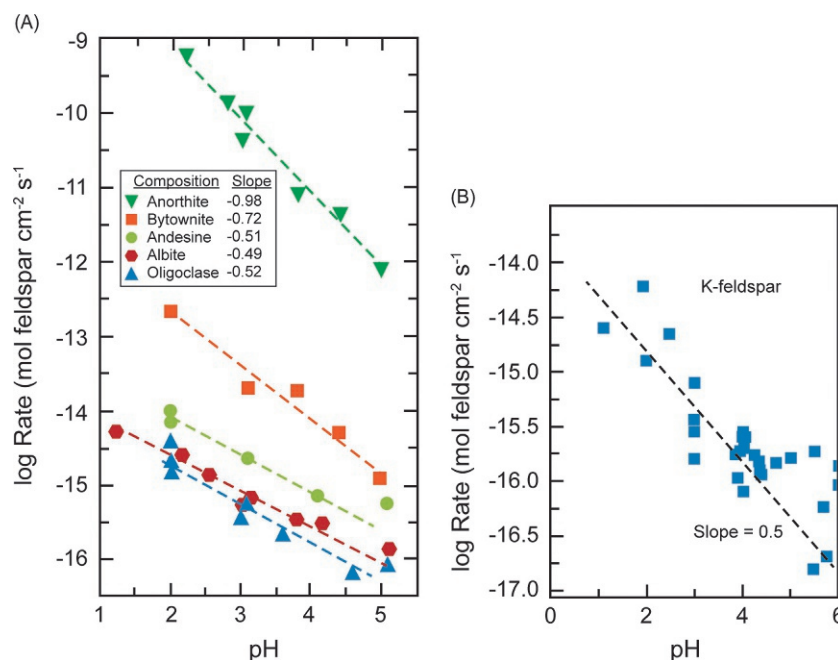
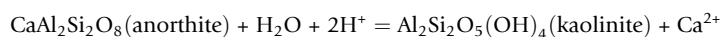
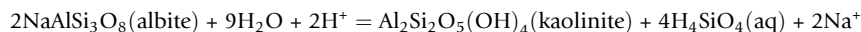


Fig. 7 Rate of dissolution of feldspars as a function of pH (A) Plagioclase feldspars of different compositions and (B) K-feldspars. The data for both feldspars were compiled from several studies. From Blum AE and Stillings LL (1995) Feldspar dissolution kinetics. In: White AF and Brantley SL (eds.) Chemical Weathering Rates of Silicate Minerals, Reviews in Mineralogy, Vol. vol. 31, pp. 291–346. Washington, DC: Mineralogical Society of America.

enhancement by a factor of 15, and that humic acids do not significantly affect feldspar dissolution. [van Hees et al. \(2002\)](#) studied the effects of naturally occurring organic acids in soils on microcline and labradorite dissolution, and at pH 5 the rate increased from 2.4 to 5.7 times from aqueous dissolution at the same pH, which was considered to be a significant enhancement. Organic acids, such as oxalic and citric acids, at concentrations likely to occur in soils may increase the rates of dissolution of feldspars by less than an order of magnitude, through their ability to complex Al around pH 5 ([Wilson, 2004](#)). At lower pH values increased dissolution effects are more attributable to the higher proton concentration rather than to the complexing nature of the acid. Aluminum-rich feldspars (e.g., Ca-rich plagioclase) are more susceptible to attacks by organic acids than K-feldspars due to preferential adsorption of organic acids at Al sites on mineral surfaces that cause ligand promoted dissolution.

In the natural environment, the steady state dissolution step is the main process controlling the weathering of feldspars. If the residence time of the solution in contact with the rocks is sufficiently long, the dissolved ion concentrations may reach an oversaturation stage that may cause the precipitation of secondary minerals. Kaolinite and halloysite are the most common products of feldspar weathering in soils and saprolite (Robertson and Eggleton, 1991 and Jeong, 1998 in [Wilson, 2004](#)). Weathering reactions of feldspars to kaolinite can be written as:



Gibbsite has also been reported as a weathering product at early stages in soils and saprolites. Few studies have identified smectite formation from feldspar weathering in more closed and alkaline conditions ([Wilson, 2004](#)).

Micas

Micas are 2:1 phyllosilicates with tightly held, non-hydrated, interlayer cations balancing a high layer charge. Primary attention has been given to those micas with K^+ as the interlayer cation because they are the abundant and important micas in most soils. Micas are abundant in many rocks, e.g., shales, slates, phyllites, schists, gneisses, and granites and in sediments derived from these and other rocks. The micas in soils are mainly inherited from soil parent materials. Micas serve as precursors for expandable 2:1 phyllosilicates, i.e., vermiculites and smectites, to which micas are transformed by replacement of the interlayer cations (usually K^+) by hydrated cations. Micas serve as an important source of K for plant growth in soils.

The 2:1 layer of micas comprises an octahedral sheet sandwiched between two tetrahedral sheets as depicted in [Fig. 8](#). Micas can be classified as dioctahedral or trioctahedral, depending on the types and locations of cations in the octahedral sheets ([Table 3](#)). In trioctahedral micas such as biotite, all three octahedral positions are filled with divalent cations (e.g., Mg^{2+} , Fe^{2+}), whereas in dioctahedral micas such as muscovite ([Fig. 8](#)), only two out of three octahedral cations are filled with trivalent cations (e.g., Al^{3+} , Fe^{3+}). Isomorphous substitution in the mica structure creates excess negative charge, which results in a strong coulombic attraction for charge-compensating interlayer cations such as K, which is not readily exchanged in standard cation exchange capacity analysis. The negative layer charge in micas arises from combinations of three mechanisms: (1) substitution of Si^{4+} in tetrahedral positions by R^{3+} (primarily Al or Fe), (2) substitution of R^{3+} in octahedral positions by R^{2+} , and (3) vacancies in octahedral positions. The resultant layer charge may originate entirely within the tetrahedral sheet or may originate entirely within the octahedral sheet in some species or may come partly from both sheets. The layer charge of micas is expressed on a formula unit basis ([Table 3](#)).

Muscovite is the most abundant dioctahedral primary mica. In muscovite, excess negative layer charge results from the substitution of Si^{4+} by Al^{3+} in one out of each of four Si in the tetrahedral sheet positions ([Fig. 8](#)). Biotite is the most common trioctahedral mica. Approximately three-quarters of the cation locations in the tetrahedral sheets of biotite are filled with Si^{4+} and about one-quarter is filled with Al^{3+} . The excess negative charge generated by substitution of Si^{4+} by Al^{3+} results in the negative

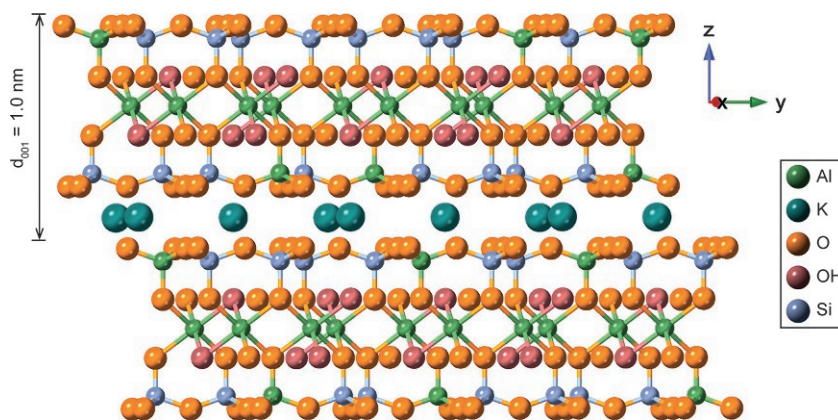


Fig. 8 Schematic structure of muscovite, two tetrahedral sheets and an octahedral sheet make up the 2:1 layer. The interlayer region is occupied by K^+ ions, the d_{001} spacing in 1.0 nm.

Table 3 Ideal formulae and layer charge of some dioctahedral and trioctahedral micas.

Mineral	Layer charge (mole per formula unit)	Chemical formula
Dioctahedral micas		
Muscovite	1	$K[Si_3Al](Al_2)O_{10}(OH)_2$
Paragonite	1	$Na[Si_3Al](Al_2)O_{10}(OH)_2$
Margarite	2	$Ca[Si_2Al_2](Al_2)O_{10}(OH)_2$
Illite	0.75	$K[Si_{3.5}Al_{0.5}](Al_{1.75}(MgFe(II)_{0.25}))O_{10}(OH)_2$
Phengite	1	$K[Si_{3.5}Al_{0.5}](Al_{1.5}(MgFe(II)_{0.5}))O_{10}(OH)_2$
Glauconite	0.91	$K_{0.81}Ca_{0.05}[Si_{3.75}Al_{0.25}](Al_{0.36}Fe(III)_{0.98})(Mg_{0.45}Fe(II)_{0.24})O_{10}(OH)_2$
Trioctahedral micas		
Biotite	1	$K[Si_3Al](Mg,Fe(II)_3)O_{10}(OH)_2$
Phlogopite	1	$K[Si_3Al](Mg_3)O_{10}(OH)_2$
Clintonite	2	$Ca[SiAl_3](Mg_2Al)O_{10}(OH)_2$
Lepidolite	1	$K[Si_3Al](Li,Al)_3O_{10}(OH)_2$

charge that is localized in the basal O atoms of the tetrahedral sheets where Al substitution occurs. In soils and sedimentary rocks, mica in the clay fraction is usually identified as illite. Illite has been called 'hydrous mica,' 'clay mica,' and 'sericite.' It is commonly dioctahedral, has poorer crystallinity, lower K content and greater H₂O content than muscovite. Glauconite is a green, ferric iron-rich micaceous mineral that forms mainly in marine, estuarine, and occasionally nonmarine sediments. Glauconite structure has many similarities to that of illite, with comparably little minor substitution of Si by Al in the tetrahedral sheet, but it contains considerably more Fe than illite.

After feldspars and quartz, micas are the third most extensive group of minerals in granitic and felsic rocks in general, but less extensive in most mafic rocks. Muscovite and biotite are the most extensive micas in igneous and metamorphic rocks. Phlogopite occurs as a product of the metamorphism of magnesian or dolomitic limestones and in serpentine rocks. Micas are generally more extensive in fine-grained sediments and sedimentary rocks (e.g., shales) than in coarse-textured sedimentary rocks (e.g., sandstones). Shales and slates are usually rich in the fine-grained, illitic-type micas, which are also important minerals in limestones. Illite in sedimentary rocks is apparently detrital, but some might also have been developed or affected by authigenic or diagenetic processes. The contents and species of micas in more recent, unconsolidated parent materials such as glacial tills, loess, and alluvium vary with their origin. Mica group minerals are easily identified by X-ray diffraction of a basally oriented specimen, with two intense peaks at 1.0 and 0.33 nm that remain unaffected by glycerol or heat treatments.

Micas represent a major source of the macronutrient K in soil and thus play an important role in plant nutrition. If the soil solution is low in K⁺ activity, K near the edges of micaceous minerals may be released into solution by exchange with H⁺, Na⁺, Mg²⁺ or Ca²⁺. When the readily exchangeable K is exhausted, fixed K from the interlayer position can also be released. Estimating the capacity of micaceous minerals to supply K to the soil solution is critical for predictions of plant nutrition and growth.

Micas exhibit a strong preference for large weakly hydrated cations such as Cs⁺ (cesium), which has an important environmental application in determining the fate of some radioactive wastes. Radioactive isotopes of Cs have been added to soils for waste disposal and also by deposition of fallout from both nuclear weapon testing and accidents at nuclear energy facilities. The mobility and bioavailability of radiocesium in mineral soils is limited by nearly instantaneous sorption of ¹³⁷Cs⁺ ions at the frayed edges of mica particles followed by their slower fixation by diffusion into interlayer sites within mica crystals. Therefore, micas play a role in restricting the contamination of ¹³⁷Cs⁺ in the terrestrial food chain and the impact on human health.

Structural Fe contents of micas may be substantial. For example, Fe accounts for approximately 23% by mass of ideal biotite (i.e., has a chemical composition close to the pure form). Structural Fe can participate in electron-transfer reactions in soils. Three examples of mica involvement in environmentally significant electron-transfer reactions are the reduction of NO₃⁻ (nitrate) to NH₄⁺ (ammonium), the reduction of Cr(VI) (chromium VI) to Cr(III), and the reductive degradation of chlorinated hydrocarbons in soils and groundwater.

Weathering of micas

The weathering of micas has been well studied because of their importance as a potential source of K for plants in the soil. It has long been known that there was a sharp contrast in the susceptibility to weathering of biotite compared with muscovite. Biotite is known to weather very easily and in contrast, muscovite is often extremely resistant to weathering and persists even in highly weathered soils (Wilson, 2004). The simple transformation of micas to expansible 2:1 phyllosilicates (e.g., vermiculites and smectites), involves exchange of the interlayer K by hydrated cations (e.g., Ca²⁺, Mg²⁺ and Na⁺), solution and reduction of negative charge in the mineral structure. The transformation can occur by layer weathering (most common) and edge weathering or both (Fanning et al., 1989). Layer weathering is more common in the finest illite particles, while larger particles may release K through edge weathering. The former leads to the formation of a regularly interstratified mica-vermiculite or mica-smectite minerals (see Coleman et al., 1963 and Wilson, 1970 in Wilson, 2004). In edge weathering, interlayer K ions are exchanged from several interlayers along the edges of mica particles producing frayed edges, while some of the interlayers remain closed. During the transformation of micas to expansible 2:1 clay minerals there is also a reduction of layer charge in the mineral structure (Norris,

1973). This reduction can result from the oxidation of structural Fe^{2+} to Fe^{3+} , ejection of octahedral cations and proton incorporation into the structure (see Raman and Jackson, 1966 and Gilkes et al., 1972a,b in Fanning et al., 1989 and Scott and Amonette, 1988 in Wilson, 2004).

Mica weathering in the soil can also result from the uptake of interlayer K by plants and decomposition by organic acids excreted by plant roots, fungi or bacteria. Transformation of micas into vermiculite and interstratified minerals has been observed in controlled and field experiments (Hinsinger and Jaillard, 1993; Singh and Goulding, 1997).

The simple transformation of micas produces vermiculite, smectite and interstratified minerals that contain regular or random layers of these minerals. Other minerals, including kaolinite, halloysite, goethite, gibbsite and Al-hydroxy interlayered vermiculites can also result from the weathering of micas (Wilson, 2004).

Common accessory primary minerals

Olivines, pyroxenes and amphiboles are the accessory minerals of soils and are present in the heavy specific-gravity fractions. Weathering of these minerals is one of the major processes in the geochemical cycles of Mg and Fe. The variety of isomorphous substitution in these minerals and their relative ease of weathering make them excellent source minerals for Ca, Mg, and trace elements in soils.

Olivines

Olivines are nesosilicates, with structures consisting of independent SiO_4 tetrahedra. Olivines forms a complete solid solution series between forsterite (Mg_2SiO_4) and fayalite (Fe_2SiO_4) in nature, and have $(\text{Mg, Fe})_2\text{SiO}_4$ as a general formula. Members of the forsterite–fayalite series comprise the most abundant naturally occurring olivines. The structure of forsterite consists of individual Si tetrahedra linked by Mg (and Fe) atoms that are in octahedral coordination (Fig. 9). The O atoms lie in approximately hexagon closely-packed sheets parallel to the (100) plane. The SiO_4 tetrahedra point alternately in opposite directions along the X- and Y-axis. Half of the available octahedral voids are occupied by Mg or Fe atoms and one-eighth of the available tetrahedral voids by Si atoms. Replacement of Mg^{2+} and Fe^{2+} by Mn^{2+} and Ca^{2+} may occur in minerals of the olivine group, for example – Tephroite (Mn_2SiO_4), Knebelite ($(\text{Fe, Mn})_2\text{SiO}_4$), Monticellite (CaMgSiO_4). The most common composition of olivine found on the surface of the Earth is a forsterite-rich olivine with $\sim 10\%$ Fe^{2+} . The Mg-rich olivines are predominant in basic rocks and the Fe-rich varieties in intermediate and acid igneous rocks. Tephroite occurs mainly in metasomatic rocks and Mn deposits, and knebelite is largely restricted to Fe–Mn ore deposits and to metamorphosed manganese sediments.

Olivine in soils weathers rapidly, therefore and it is rarely found in soils. Delvigne et al. (1979) identified three situations for olivine's existence: (i) in rock fragments and sand fraction in the vicinity of large basic or ultramafic outcrops, fresh or moderately weathered rocks that are actively eroding; (ii) in soils formed from recent volcanic deposits, and (iii) in older soils, where olivine has partly transformed into secondary minerals and the olivine core is protected by a cortex of iddingsite or a reaction rim of pyroxene or garnet.

Weathering of olivine

Olivine releases Mg, Fe and other nutrients and potential toxic elements (e.g. nickel, chromium) on weathering. Iddingsite (a mixture of trioctahedral smectite and Fe oxides) has been reported as an alteration pseudomorph after olivine (Eggleton, 1984

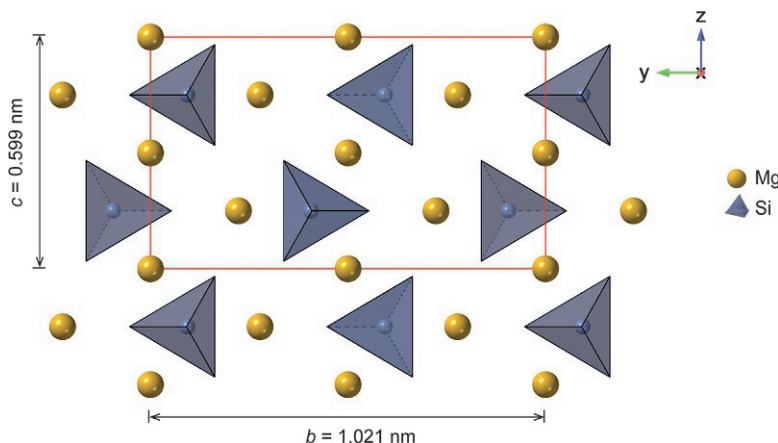
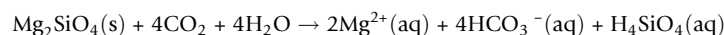


Fig. 9 The structure of forsterite (Mg_2SiO_4) parallel to [100] plane. The Si—O bonds are shown in the isolated tetrahedra, which point alternatively in opposite direction along the X and Y axes. The unit-cell is outlined in the red rectangle.

and Smith et al., 1987 in Wilson, 2004). The extent of weathering of these minerals indicates the intensity of chemical weathering of soils in cool humid regions. Where leaching is only moderate, olivines may alter to serpentine or trioctahedral smectite (usually of the saponite type), nontronite, and various ferric hydrates and gels. In an environment where leaching is intense, the degradation products are mixtures of poorly crystallized smectite, kaolinite, halloysite, and Fe oxides such as goethite, hematite, or the noncrystalline precursors of these minerals (Wilson, 2004). In addition to abiotic chemical weathering, olivine can be weathered by microbial activity and by organic acids.

In the last few years, olivine dissolution has received considerable attention because its enhanced weathering has been considered in carbon dioxide (CO₂) removal technology to mitigate anthropogenic climate change (Oelkers et al., 2018). Enhanced weathering involves the dissolution of silicate minerals and Ca and Mg are released, which can react with dissolved CO₂, (HCO₃⁻ (bicarbonate), CO₃²⁻ (carbonate)) and precipitate as carbonate minerals, such as calcite (CaCO₃), dolomite (MgCO₃), and magnesite (CaMg(CO₃)₂). Olivine (forsterite) dissolution implies the sequestration of 4 mol of CO₂ for each mole of olivine:



With both the absence of covalent Si—O—Si bonds and relative weakness of the ionic divalent metal–oxygen bonds in its structure, olivine is among the fastest dissolving silicate mineral, and is considered one of the key minerals in enhanced mineral weathering technology for CO₂ sequestration (Oelkers et al., 2018). The dissolution behavior of forsterite as a function of pH in aqueous solutions containing only Mg, Si, Fe, and simple inorganic acids or bases is depicted in Fig. 10 (Rimstidt et al., 2012; Oelkers et al., 2018). The rate of forsterite dissolution decreases with increasing pH that is typical for the Mg-silicates, and contrasts with that of Al-silicates, which tend to increase with increasing pH under alkaline conditions (Oelkers et al., 2018). To make olivine a viable option for CO₂ sequestration and to enhance weathering, Oelkers et al. (2018) suggested increasing the mineral surface area and decreasing solution pH. Organic ligands generally have a limited influence on rates of forsterite dissolution in acidic solutions. However, under neutral to alkaline conditions organic ligands could potentially absorb and remove Mg that is abundant at the forsterite surface and accelerate mineral dissolution rates.

Pyroxenes and amphiboles

Pyroxenes and amphiboles are two groups of ferromagnesian minerals that are inosilicates with long chains of linked SiO₄ tetrahedra. Pyroxenes consist of single chains of linked SiO₄ tetrahedra, each of which shares two O atoms with its neighboring tetrahedra (Fig. 11A). The top vertices of the tetrahedra of each chain always face the same direction and subsequent chains in the structure have opposite polarity of these vertices, facing alternatively up and down as shown in Fig. 11A. The tetrahedral chains are separated by layers of octahedral sites.

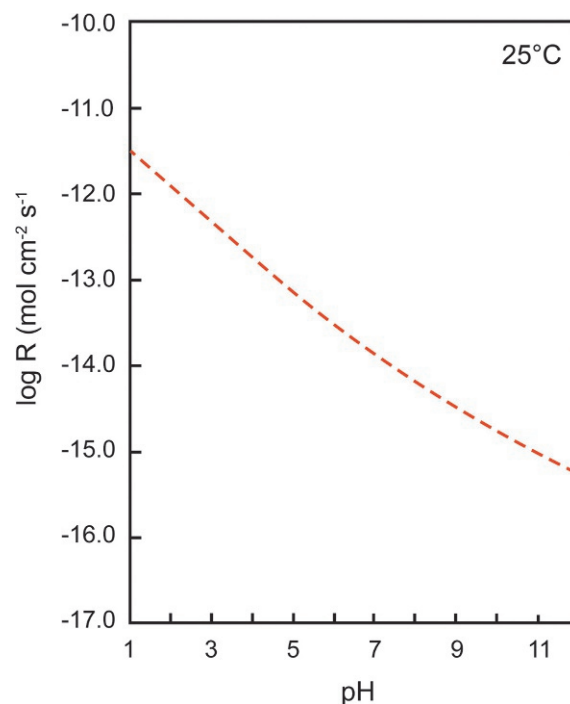
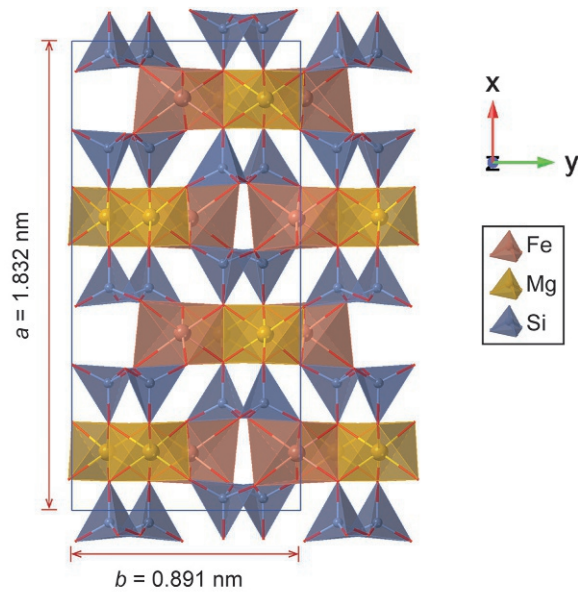
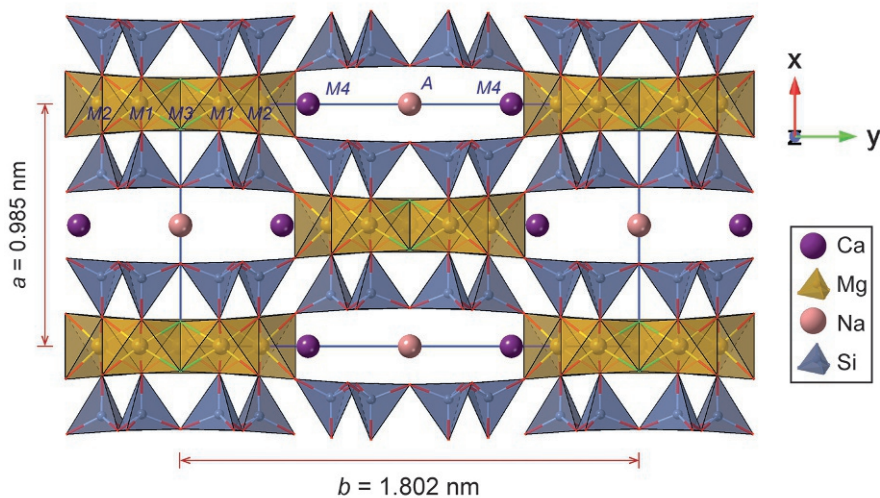


Fig. 10 Steady-state dissolution rate of forsterite at 25 °C as a function of pH, the curve is based on the statistical analysis of published data from several studies by Rimstidt et al. (2012) and drawn based on Oelkers et al. (2018).

(A) Orthopyroxene



(B) Hornblende - viewed along 001 plane



(C) Hornblende - viewed along 101 plane

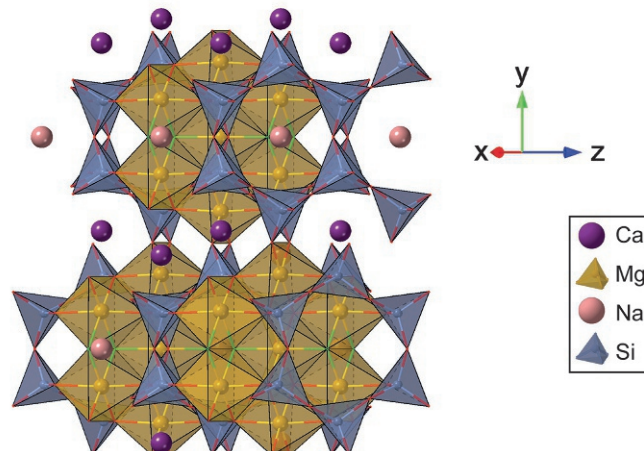


Fig. 11 (A) The structure of an orthopyroxene (a pyroxene with orthorhombic crystal symmetry, with general composition $(\text{Mg,Fe})\text{SiO}_3$) viewed along the Z-axis, i.e., parallel to [001] plane, showing a single chain of tetrahedra. (B) The structure of hornblende (an amphibole) viewed parallel to [001] plane and (C) the hornblende structure viewed along [101] plane showing double chain of tetrahedra. Octahedral sites (M1, M2, M3, and M4) and sites (A) that exist as large gaps between tetrahedra are indicated in (B).

Table 4 Ideal formulae and unit cell parameters of pyroxenes.

Mineral	Chemical formula	Unit cell parameters			
		a (nm)	b (nm)	c (nm)	β (°)
Pyroxenes					
Orthopyroxene					
Enstatite	Mg ₂ Si ₂ O ₆	1.823	0.884	0.519	90
Ferrosilite	Fe ₂ Si ₂ O ₆	1.830	0.913	0.520	90
Clinopyroxene					
Clinoenstatite	Mg ₂ Si ₂ O ₆	0.961	0.882	0.517	108.3
Clinoferrosilite	Fe ₂ Si ₂ O ₆	0.953	0.921	0.515	107.63
Pigeonite	(Mg,Fe ²⁺ ,Ca) ₂ Si ₂ O ₆	0.970	0.884	0.527	106.97
Calcium pyroxenes					
Diopside	CaMgSi ₂ O ₆	0.975	0.890	0.525	105.63
Hedenbergite	CaFe ²⁺ Si ₂ O ₆	0.984	0.902	0.524	104.80
Augite	(Ca,Mg,Fe ²⁺ ,Al) ₂ (Si,Al) ₂ O ₆	0.970	0.884	0.527	106.97
Calcium-sodium pyroxenes					
Omphacite	(Ca,Na)(Mg,Fe ²⁺ ,Fe ³⁺ ,Al) Si ₂ O ₆	0.959	0.878	0.526	105.85
Aegirine-augite	(Ca,Na)(Mg,Fe ²⁺ ,Fe ³⁺) Si ₂ O ₆	0.966	0.880	0.529	107.42
Sodium pyroxenes					
Jadeite	NaAlSi ₂ O ₆	0.944	0.857	0.523	107.58
Aegirine	NaFe ³⁺ Si ₂ O ₆	0.966	0.880	0.529	107.43
Kosmochlor	NaCrSi ₂ O ₆	0.957	0.871	0.526	107.49
Lithium pyroxenes					
Spodumene	LiAlSi ₂ O ₆	0.947	0.841	0.522	110.05

The general chemical formula (formula unit) for pyroxenes is $M_2M_1T_2O_6$, where M_2 refers to cations in a generally distorted octahedral coordination, M_1 refers to cations in a regular octahedral coordination, and T to tetrahedrally coordinated cations. Common pyroxenes and minerals of end members of various groups are listed in Table 4. Magnesium-Fe pyroxenes are a solid-solution between enstatite (Mg₂Si₂O₆) and ferrosilite (Fe₂Si₂O₆) and are also known as orthopyroxenes due to their orthorhombic crystal structure ($\beta = 90^\circ$). Clinoenstatite and clinoferrosilite, the low-temperature monoclinic ($\beta > 90^\circ$) polymorphs of the orthopyroxenes, exist in nature but are rather uncommon. Calcium pyroxenes represent a solid solution between the monoclinic diopside (CaMgSi₂O₆) and hedenbergite (FeMgSi₂O₆), also known as clinopyroxenes. Magnesium-Fe pyroxenes and some of the Ca pyroxenes (e.g., augite) are the most common rock forming pyroxenes.

The structure of amphiboles consists of indefinite double chains along the [001] plane of tetrahedra that are linked together by O atoms, and the tetrahedra share two and three O atoms alternately (Fig. 11B). Double chains are joined together by strips of octahedral sites (M sites). There are four different types of octahedral sites (M1, M2, M3, and M4) in the mineral structure that vary in size. Three smaller size sites (M1, M2, M3) occur sandwiched between two double chains of tetrahedra, and the larger M sites (M4) occupy a peripheral position with respect to the chains (Fig. 11B). Fig. 11B shows an amphibole structure with a series of sandwiches of tetrahedral chains with octahedral sites in between, and each island consists of 5 octahedral sites with every 8 tetrahedra, four on each side. The large gaps (A sites) that are evident in the structure may be empty or partially filled by large cations, such as Na, and are points of weakness in the mineral structure. The bond between the O atoms and the cations linking the tetrahedral chains is weaker than that between the O and Si. Therefore, cleavage takes place diagonally through the crystal structure and does not rupture the Si—O chains.

The general formula of amphiboles may be written as $A_{0-1}B_2C_5T_8O_{22}(OH,F)_2$, where A, B and C are the atoms at sites A, M4, and M1 + M2 + M3, respectively (site as shown in Fig. 11B), and T cations at tetrahedral sites (Deer et al., 1992). The most common cations at each structural site are A— Na⁺, K⁺; B— Na⁺, Ca²⁺, Mg²⁺, Fe²⁺; C— Mg²⁺, Fe²⁺, Al³⁺, Fe³⁺; T— Si⁴⁺, Al³⁺, and hydroxyls (OH⁻) can be replaced by Cl⁻ and F⁻. The composition of amphiboles varies considerably, but four groups can be distinguished based on the cations occupying the M4 sites (or B in the general formula): calcic amphiboles (B = Ca), sodic amphiboles (B = Na), sodic-calcic amphiboles (B = both Na, and Ca in appreciable contents), and iron-magnesium-manganese amphiboles (B = Fe, Mg, Mn). Amphiboles that exist in soils are presented in Table 5.

Many amphiboles are absent in primary igneous rocks or may occur as secondary minerals. The main primary amphiboles of igneous rocks are the Mg-poor riebeckites and the Si-poor basaltic and common hornblendes. Hornblendes are widely distributed in igneous rocks from syenite and granite to gabbro, and in metamorphic rocks such as gneiss, hornblende schist, and amphibolites.

Pyroxenes and amphiboles are largely confined to the sand and silt fractions of soils. Some may occur in the clay fraction of soils that are developed from glacial rock flour not subjected to intensive weathering. The variety of isomorphous substitution in these minerals and their easy weatherability make them excellent sources of trace elements and Ca and Mg in soils.

Table 5 Chemical compositions of amphiboles classified based on the cation occupancy of the B (or M4) position in the mineral structure^a.

Group	Mineral	Cation sites ^b				Anions	Symmetry
		A ₀₋₁	B ₂	C ₅	T ₈	O ₂₂ (OH) ₂	
Iron-magnesium	Anthophyllite	—	(Mg,Fe) ₂	(Mg,Fe) ₅	Si ₈	O ₂₂ (OH) ₂	Orthorhombic
	Gedrite	—	(Mg,Fe) ₂	(Mg,Fe) ₃ Al ₂	Si ₆ Al ₂	O ₂₂ (OH) ₂	Orthorhombic
	Cummingtonite-Grunerite	—	(Mg,Fe) ₂	(Mg,Fe) ₅	Si ₈	O ₂₂ (OH) ₂	Monoclinic
Calcic	Tremolite-actinolite	—	Ca ₂	(Mg,Fe) ₅	Si ₈	O ₂₂ (OH) ₂	Monoclinic
	Hornblende	Na, K ₀₋₁	Ca ₂	(Mg,Fe ²⁺ , Fe ³⁺ , Al) ₅	(Si, Al) ₈	O ₂₂ (OH) ₂	Monoclinic
	Kaersutite	Na	Ca ₂	(Mg,Fe ²⁺) ₄ Ti	Si ₆ Al ₂	O ₂₂ (OH) ₂	Monoclinic
Sodic-calcic	Richterite	Na	NaCa	(Mg,Fe) ₅	Si ₈	O ₂₂ (OH) ₂	Monoclinic
	Katophorite	Na	NaCa	(Mg,Fe ²⁺) ₄ Fe ³⁺	Si ₇ Al	O ₂₂ (OH) ₂	Monoclinic
Sodic	Glaucophane-riebeckite	—	Na ₂	(Mg,Fe ²⁺) ₃ (AlFe ³⁺) ₂	Si ₈	O ₂₂ (OH) ₂	Monoclinic
	Eckermanite-arfvedsonite	Na	Na ₂	(Mg,Fe ²⁺) ₄ (AlFe ³⁺)	Si ₈	O ₂₂ (OH) ₂	Monoclinic

^aBased on Nasse (2000) and Deer et al. (1992).^bSite A – refers to large space that may be empty or partly occupied by large cations (Na and K), B denotes peripheral octahedral, C represents octahedral site sandwiched between adjacent tetrahedra, and T refers to the tetrahedral sites; the subscript number indicates the number of atoms in these sites.

Weathering of pyroxenes and amphiboles

Brantley and Chen (1995) reviewed the rates of dissolution of pyroxenes and amphiboles from published data; they are shown for selective minerals in the acidic pH range in Fig. 12. The rates of dissolution decrease with increasing solution pH for all minerals, however, there are substantial differences between minerals. The dissolution rates of spodumene and jadeite are much faster than those of other minerals and this may be partly because the data were not normalized to the minerals' final surface area. Nevertheless, the figure shows that dissolution rates of Fe-free amphiboles (e.g., anthophyllite) are slower than comparable pyroxenes (e.g., enstatite and diopside). It has also been noted that the dissolution rate constants of the comparable ortho-, and single-chain and double-chain inosilicates decrease with increasing polymerization of the silica tetrahedral structure (Brantley and Chen, 1995).

Both pyroxenes and amphiboles have been shown to weather to hydrous layer silicates involving a topotactic mechanism, i.e., a high degree of structural inheritance by the secondary phase from the primary mineral (Eggleton, 1975 and Eggleton and Boland, 1982 in Wilson, 2004). However, few studies have found that the weathering occurs via the dissolution and precipitation pathway (Nahon and Colin, 1982, and Singh and Gilkes, 1993 in Wilson, 2004).

Pyroxenes tend to weather to chlorite or smectite or both, through partial dissolution of Mg, Ca, and Fe²⁺. Calcite may develop in the environment where the rate of dissolution of Ca is greater than the rate of total breakdown of the pyroxene. As chemical weathering proceeds, all the Ca and Mg, and part of the Si are lost, resulting in progressive enrichment of the residue in kaolinite, ferric oxides and oxyhydroxides, and anatase. Amphiboles broadly follow a weathering sequence similar to pyroxenes. Velbel

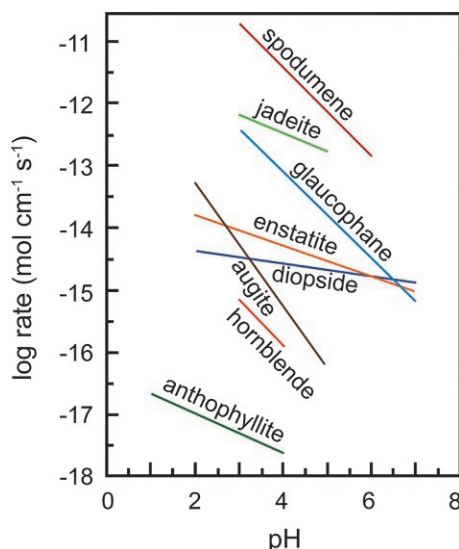


Fig. 12 Dissolution rates for some common pyroxenes and amphiboles at 25 °C in the acidic pH range. From Brantley SL and Chen Y (1995) Chemical weathering rates of pyroxenes and amphiboles. In: White AF and Brantley SL (eds.) Chemical Weathering Rates of Silicate Minerals, Reviews in Mineralogy, vol. 31, pp. 119–172. Washington, DC: Mineralogical Society of America.

(1989) observed stoichiometric dissolution of amphibole that was followed by the precipitation of goethite, gibbsite and kaolinite within and around the mineral.

Some other minerals that are not as abundant as the accessory minerals discussed above, but are still important in soils, are apatite, $\text{Ca}_{10}(\text{PO}_4)_6(\text{F},\text{OH},\text{Cl})_2$ (and its weathering products), tourmaline (a complex boron silicate mineral with a variable chemical composition), epidote ($\text{Ca}_2(\text{Al}_2\text{Fe}^{3+})\text{O}(\text{OH})(\text{Si}_2\text{O}_7)(\text{SiO}_4)$), rutile (TiO_2), ilmenite (FeTiO_3), magnetite (Fe_3O_4), zircon (ZrSiO_4). They are often present in small concentrations (<1%) and not identified in routine soil characterization. Some of these minerals are important sources of nutrient elements to soil, for example, apatite for Ca and phosphorus (P) and tourmaline for boron (B), and others such as zircon are important in tracing the origin of soil parent materials or determining the extent of soil weathering.

Conclusions and future perspectives

Primary minerals are present in all soils and sediments, and they are abundant in the sand and silt size fractions. The weathering of primary minerals is a key process for the terrestrial ecosystem because it is the principal source of many essential plant nutrients, such as Ca, Mg, K, P, Fe, Mn, B. Primary minerals serve as a template for secondary soil minerals or release elements for the precipitation of secondary minerals in soils. The weathering of primary minerals occurs through combinations of physical, chemical and biological processes. Chemical weathering, however, is most important in the ultimate breakdown of primary minerals.

Primary minerals can dissolve congruently or there can be the preferential removal of certain elements from a mineral structure, depending upon the mineral properties and conditions of weathering. Minerals often dissolve at points of weakness in their structure such as dislocations, defects and crystal edges. In ferro-magnesian minerals, such as biotite, olivine, pyroxenes and amphiboles, the oxidation (and even release) of structural Fe^{2+} is a key process of mineral weathering. The weathering of primary silicate minerals is crucial for the release of plant nutrients to maintain a healthy terrestrial ecosystem. Primary minerals buffer acid input in the soil. Forsterite has been considered a potential mineral for CO_2 sequestration in enhanced weathering technology.

Much of the current information about the properties of primary mineral is based on limited studies in young soils and further research using advanced techniques is required to gather knowledge about their occurrence, properties, dissolution behavior in different soil types and weathering environments. There is also little knowledge about the properties and dissolution behavior of many accessory primary minerals that occur in small quantities, which requires more research.

References

- Allègre C (2008) *Isotope Geology (Translated by C. Suttle)*. Cambridge: Cambridge University Press.
- Blum AE and Stillings LL (1995) Feldspar dissolution kinetics. In: White AF and Brantley SL (eds.) *Chemical Weathering Rates of Silicate Minerals. Reviews in Mineralogy*, vol. 31, pp. 291–346. Washington, DC: Mineralogical Society of America.
- Bowen HJM (1979) *Environmental Chemistry of the Elements*. London: Academic Press.
- Brantley SL (2005) Reaction kinetics of primary rock-forming minerals under ambient conditions. In: Drever JI (ed.) *Surface and Ground Water, Weathering and Soils. Treatise on Geochemistry*, vol. 5, pp. 73–117. Amsterdam: Elsevier.
- Brantley SL and Chen Y (1995) Chemical weathering rates of pyroxenes and amphiboles. In: White AF and Brantley SL (eds.) *Chemical Weathering Rates of Silicate Minerals. Reviews in Mineralogy*, vol. 31, pp. 119–172. Washington, DC: Mineralogical Society of America.
- Casey WH, Westrich HR, Arnold GW, and Banfield JF (1989) The surface chemistry of dissolving labradorite feldspar. *Geochimica et Cosmochimica Acta* 53: 821–832.
- Chou L and Wollast R (1984) Study of the weathering of albite at room temperature and pressure with a fluidised bed reactor. *Geochimica et Cosmochimica Acta* 48: 2205–2217.
- Deer WA, Howie RA, and Zussman J (1992) *An Introduction to the Rock Forming Minerals*. Harlow: Longman.
- Delvigne J, Bisdom EBA, Sleeman J, and Stoops G (1979) Olivines, their pseudomorphs and secondary products. *Pédologie* 29: 247–309.
- Dewan S, Yeganeh MS, and Borguet E (2013) Experimental correlation between interfacial water structure and mineral reactivity. *Journal of Physical Chemistry Letters* 4: 1977–1982.
- Dove PM and Han N (2007) Kinetics of mineral dissolution and growth as reciprocal microscopic surface processes across chemical driving force. *American Institute of Physics Conference Series* 916: 215–234.
- Dove PM and Nix CJ (1997) The influence of the alkaline earth cations, magnesium, calcium and barium on the dissolution kinetics of quartz. *Geochimica et Cosmochimica Acta* 61: 3329–3340.
- Drees LR, Wilding L, Smeeck NE, and Senkayi AL (1989) Silica in soils: Quartz and disordered silica polymorphs. In: Dixon JB and Schulze DG (eds.) *Soil Mineralogy with Environmental Applications*, pp. 331–378. Madison, WI: Soil Science Society of America Inc.
- Drever JI and Stillings LL (1997) The role of organic acids in mineral weathering. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* 120: 167–181.
- Fanning DS, Keramidas VZ, and El-Desoky MA (1989) Micas. In: Dixon JB and Schulze DG (eds.) *Soil Mineralogy with Environmental Applications*, pp. 551–634. Madison, WI: Soil Science Society of America Inc.
- Fröhlich F (2020) The opal-CT nanostructure. *Journal of Non-Crystalline Solids* 533: 119938.
- Götze J, Pan Y, and Müller A (2021) Mineralogy and mineral chemistry of quartz: A review. *Mineralogical Magazine* 85: 639–664.
- Graetsch H, Flörke OW, and Miehe G (1987) Structural defects in microcrystalline silica. *Physics and Chemistry of Minerals* 14: 249–257.
- Guthrie GD Jr. and Heaney PJ (1995) Mineralogical characteristics of silica polymorphs in relation to their biological activities. *Scandinavian Journal of Work, Environment & Health* 21(2): 5–8.
- Heaney PJ (1994) Structure and chemistry of the low-pressure silica polymorphs. In: Heaney PJ, Prewitt CT, and Gibbs GV (eds.) *Silica - Physical Behavior, Geochemistry and Materials Applications*, pp. 1–38. Washington DC: Mineralogical Society of America.
- Hinsinger P and Jaillard B (1993) Root-induced release of interlayer potassium and vermiculization of phlogopite as related to potassium depletion in the rhizosphere of ryegrass. *Journal of Soil Science* 44: 525–534.
- Köhler P, Hartmann J, and Wolf-Gladrow DA (2010) Geoengineering potential of artificially enhanced silicate weathering of olivine. *Proceedings of the National Academy of Sciences of the USA* 107: 20228–20233.
- Muir IJ, Bancroft MG, and Nesbitt HW (1989) Characteristics of altered labradorite surfaces by SIMS and XPS. *Geochimica et Cosmochimica Acta* 53: 1235–1241.

- Nesbitt HW, Macrae ND, and Shetye W (1991) Congruent and incongruent dissolution of labradorite in dilute acidic salt solutions. *Journal of Geology* 99: 429–442.
- Nesse WD (2000) *Introduction to Mineralogy*. New York: Oxford University Press.
- Norrish K (1973) In: Serratosa JM (ed.) *Factors in the weathering of mica to vermiculite*. Proceedings of the International Clay Conference, pp. 419–432. Madrid: CSIC.
- Oelkers EH, Declercq J, Saldi GD, Gislason SR, and Schott J (2018) Olivine dissolution rates: A critical review. *Chemical Geology* 500: 1–19.
- Rimstidt JD, Brantley SL, and Olsen AA (2012) Systematic review of forsterite dissolution rate data. *Geochimica et Cosmochimica Acta* 99: 159–178.
- Ronov AB and Yaroshevsky AA (1969) Chemical composition of the Earth's crust. In: Hart PJ (ed.) *The Earth's Crust and Upper Mantle. Geophysical Monograph 13*, Washington, DC: American Geophysical Union.
- Schilling RD and Krijgsman P (2006) Enhanced weathering: An effective and cheap tool to sequester CO₂. *Climate Change* 74: 349–354.
- Singh B and Goulding KWT (1997) Changes with time in the potassium content and phyllosilicates in the soil of the Broadbalk continuous wheat experiment at Rothamsted. *European Journal of Soil Science* 48: 651–659.
- Smith JV (1974) *Feldspar minerals: I. Crystal Structure and Physical Properties*. Berlin: Springer-Verlag.
- Smith JV and Steele IM (1984) Chemical substitution in silica polymorphs. *Neues Jahrbuch für Mineralogie* 3: 137–144.
- Stillings LL and Brantley SL (1995) Feldspar dissolution at 25°C and pH 3: Reaction stoichiometry and the effect of cations. *Geochimica et Cosmochimica Acta* 59: 1483–1496.
- van Hees PAW, Lundström US, and Möhrh CM (2002) Dissolution of microcline and labradorite in a forest O horizon extract: The effect of naturally occurring organic acids. *Chemical Geology* 189: 199–211.
- Velbel MA (1989) Weathering of hornblende to ferruginous products by a dissolution-reprecipitation mechanism: Petrography and stoichiometry. *Clays and Clay Minerals* 37: 515–524.
- Welch SS and Ullman WJ (1993) The effect of organic acids on plagioclase dissolution rates and stoichiometry. *Geochimica et Cosmochimica Acta* 57: 2725–2737.
- Wilson MJ (2004) Weathering of the primary rock-forming minerals: Processes, products and rates. *Clay Minerals* 39: 233–266.
- Wilson MJ (2020) Dissolution and formation of quartz in soil environments: A review. *Soil Science Annual* 71(2): 99–110.
- Wyllie PJ (1971) *The Dynamic Earth: Textbook in Geosciences*. New York: John Wiley & Sons Inc.

Mineral components of the soil clay fraction

Darrell G Schulze, Agronomy Department, Purdue University, West Lafayette, IN, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of D.G. Schulze, CLAY MINERALS, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 246–254, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00189-2>.

Introduction	110
Major element composition of clay minerals	110
Basic structural concepts	110
Tetrahedra and octahedra	110
Representing crystal structures	110
Tetrahedral and octahedral sheets	111
The tetrahedral sheet	111
The octahedral sheet	112
Phyllosilicate minerals common in soil clays	113
1:1-Type minerals	113
The 1:1 layer structure	113
Kaolinite	114
Halloysite	114
2:1-Type minerals	115
The 2:1 layer structure	115
Pyrophyllite	115
Micas	115
Vermiculites	116
Smectites	116
Chlorites	117
Hydroxy-interlayered vermiculite and smectite	117
Interstratification in phyllosilicates	117
Palygorskite and sepiolite	118
Other minerals that occur in the clay fraction of soil	119
Zeolites	119
Allophane and imogolite	119
Aluminum hydroxide minerals	119
Iron oxide minerals	119
Manganese oxide minerals	120
Titanium oxide minerals	120
Carbonates, sulfates, and soluble salts	120
Summary	120
Acknowledgments	120
References	120

Abstract

The clay fraction of soils consists of mineral particles $<2\ \mu\text{m}$ in diameter, with the most abundant component being clay minerals based on sheets of tetrahedral and octahedral structural units containing silicon (Si), aluminum (Al), and oxygen (O). These exceedingly small particles crystallize in the aqueous environment at the Earth's surface and are responsible for many of the most important physical and chemical properties of soils such as cation exchange, water holding capacity, ability to supply plant nutrients, shrink–swell properties, and the ability to attenuate pollutants.

Key points

- The clay fraction of soils consists primarily of a variety of phyllosilicate minerals that form as the result of weathering at the Earth's surface.
- The most common and abundant clay minerals are based on 1:1 layer and 2:1 layer structures consisting mainly of Si, Al, and O.
- Structural differences between different clay minerals result in their different chemical and physical properties.
- Differences in the types of soil clay minerals account for much of the chemical and physical variation in soils.

Introduction

The clay fraction of soils consists of mineral particles that are less than 2 μm in equivalent spherical diameter. This is the realm of exceedingly small, crystalline particles dominated by planar arrays of SiO_4 (silicate) structural units and many structural hydroxyls and water. These ‘clay minerals’ crystallize in the aqueous environment at the Earth’s surface from constituent ions released by dissolving (weathering) ‘primary minerals’ such as olivines, pyroxenes, feldspars, micas, quartz, and others that were formed in igneous and metamorphic rocks under extreme heat and pressure within the Earth. Clay minerals are responsible for many of the soil’s most important and characteristic physical and chemical properties. Fundamental soil properties such as cation exchange capacity and shrink–swell properties, and practical considerations such as how well a particular soil will attenuate a specific pollutant, or how much fertilizer phosphorus will be fixed and unavailable to crops, are all influenced by molecular-scale differences in soil clay minerals. Clay minerals are distinguished by their different crystal structures, and there is a close relationship between the crystal structure and the corresponding bulk physical and chemical properties of a particular type of clay mineral.

This chapter draws heavily from [Dixon and Weed \(1989\)](#) and [Dixon and Schulze \(2002\)](#), which contain individual chapters on all the major clay minerals discussed below. Unless otherwise noted, the reader is referred to these texts for additional information. I begin by considering some of the properties of the major chemical elements that make up the clay minerals.

Major element composition of clay minerals

Most of the mass and volume of the Earth’s crust is made up by only a few chemical elements. Oxygen (O) and Si (silicon) alone account for almost 75% of the mass, with most of the remainder, in order of decreasing abundance, consisting of Al (aluminum), Fe (iron), Ca (calcium), Na (sodium), K (potassium), Mg (magnesium), Ti (titanium), H (hydrogen), P (phosphorus), and Mn (manganese) ([Klein and Dutrow, 2008](#)). On a volume basis, oxygen alone accounts for more than 90% of the total volume. Oxygen, as O^{2-} , is the only abundant anion, while the other abundant elements are all cations. Most of these cations have only one stable oxidation state at the Earth’s surface (Al^{3+} , Ca^{2+} , Na^+ , K^+ , Mg^{2+} , Ti^{4+} , H^+ , P^{5+}); Fe (Fe^{2+} , Fe^{3+}) and Mn (Mn^{2+} , Mn^{3+} , Mn^{4+}) are the exceptions. The O^{2-} anions are much larger than most of the positively charged cations. The Earth’s crust, therefore, can be characterized as consisting of large O atoms in an approximately close-packed arrangement held together by attraction to smaller cations located in the interstitial space.

Most of the elements in the crust and in soils occur in minerals, and the elements listed above are major constituents of the most abundant minerals, including clay minerals. Building on the concept of packing O atoms in space, I will consider atoms as rigid spheres, realizing that this is an oversimplified but convenient model for developing the key structural concepts.

Basic structural concepts

Tetrahedra and octahedra

Two distinct structural features occur within the crystal structures of soil clay minerals as a consequence of packing the large O^{2-} ions together in space. The first consists of four O^{2-} ions packed closely together and can be described as three O^{2-} ions arranged in a triangle with the fourth O^{2-} occupying the dimple formed by the other three ([Fig. 1](#)). The centers of the four O^{2-} ions form the apices of a regular tetrahedron, and the small space in the center is called a ‘tetrahedral site.’ Cations located in tetrahedral sites are in fourfold or tetrahedral coordination because they are surrounded by and bonded to four O^{2-} ions.

The second structural feature consists of six closely packed O^{2-} ions. Three of them are arranged in a triangle in one plane, and the other three, also in a triangle but rotated 60° relative to the first three, are in a second plane so that the two triangular groups intermesh ([Fig. 1](#)). The centers of the six O^{2-} ions form the apices of a regular octahedron, and the small space in the center is called an ‘octahedral site.’ Cations located in the octahedral site are said to be in sixfold or octahedral coordination because they are surrounded by and bonded to six O^{2-} ions.

Tetrahedral and octahedral sites differ in an important way. The space that can be occupied by a cation in a tetrahedral site is smaller than the space that can be occupied in an octahedral site. Since cations vary in size, smaller cations tend to occur in tetrahedral sites, somewhat larger cations tend to occur in octahedral sites, and the largest cations must fit into spaces that are larger than octahedral sites. Cations with sizes intermediate between the optimum for the two sites can occur in either site. The Al^{3+} ion, for example, can occur in either octahedral or tetrahedral sites. [Table 1](#) summarizes the structural sites in which cations tend to occur in clay minerals.

Representing crystal structures

Octahedra and tetrahedra are commonly represented using different types of models, each of which portrays the same concept, but highlights different structural features. Sphere-packing models give an impression of the space occupied by the atoms, ball-and-stick

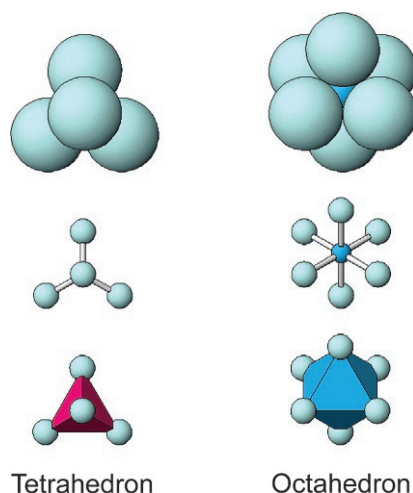


Fig. 1 Spheres closely packed to form a tetrahedron and an octahedron. Note the appearance for the three different ways of drawing the model: as a sphere-packing model (top row), ball-and-stick model (middle row), and a polyhedral model (bottom row). Adapted from Schulze DG (2002) An introduction to soil mineralogy. In: Dixon JB and Schulze DG (eds) *Soil Mineralogy with Environmental Applications*, pp. 1–35. Madison, WI: Soil Science Society of America, with permission.

Table 1 Type of structural sites in which common cations tend to occur in phyllosilicate mineral structures.

Type of site	Cation
Tetrahedral only	Si^{4+}
Tetrahedral or octahedral	Al^{3+} , Fe^{3+}
Octahedral only	Mg^{2+} , Ti^{4+} , Fe^{2+} , Mn^{2+}
Interlayer sites	Na^+ , Ca^{2+} , K^+

models highlight the bonds, while polyhedral models emphasize the tetrahedral and octahedral units (Fig. 1). There is some ambiguity associated with each representation, and it is important to understand the correspondence between, and limitations of, each type of model. Polyhedral models are used here, along with some sphere-packing models.

Tetrahedral and octahedral sheets

The most common and abundant clay minerals belong to a group of minerals called phyllosilicates (from Greek ‘phyllon,’ meaning ‘leaf’) or sheet silicates. A common structural theme in all phyllosilicates is the presence of SiO_4 tetrahedra arranged into sheets. Octahedra arranged into sheets are also present in the structures of phyllosilicates, and in some hydroxide minerals. Different combinations of the two sheets give rise to the different clay mineral structures.

The tetrahedral sheet

The tetrahedral sheet consists of SiO_4 tetrahedra arranged such that three of the four O^{2-} ions of each tetrahedron are shared with three nearest-neighbor tetrahedra (Fig. 2). These shared O^{2-} ions are all in the same plane and are referred to as basal oxygens. Note that adjacent tetrahedra share only one O^{2-} between them (the tetrahedra share apices or corners). The fourth O^{2-} ion of each tetrahedron is not shared with another SiO_4 tetrahedron and is free to be part of other polyhedral elements. These unshared O^{2-} ions are referred to as apical oxygens. Since each basal oxygen contributes a charge of -1 to each Si^{4+} ion, the addition of H^+ ions to the apical oxygens to form hydroxyls should result in an electrically neutral tetrahedral sheet. Such individual tetrahedral sheets do not form stable mineral structures by themselves and only occur in combination with octahedral sheets, as described below.

Fig. 2 shows all the apical oxygens pointing in the same direction, namely, out of the plane of the paper toward the reader in the upper half of the figure. This is the most common arrangement, but structures also occur in which the apical oxygens point alternately in opposite directions.

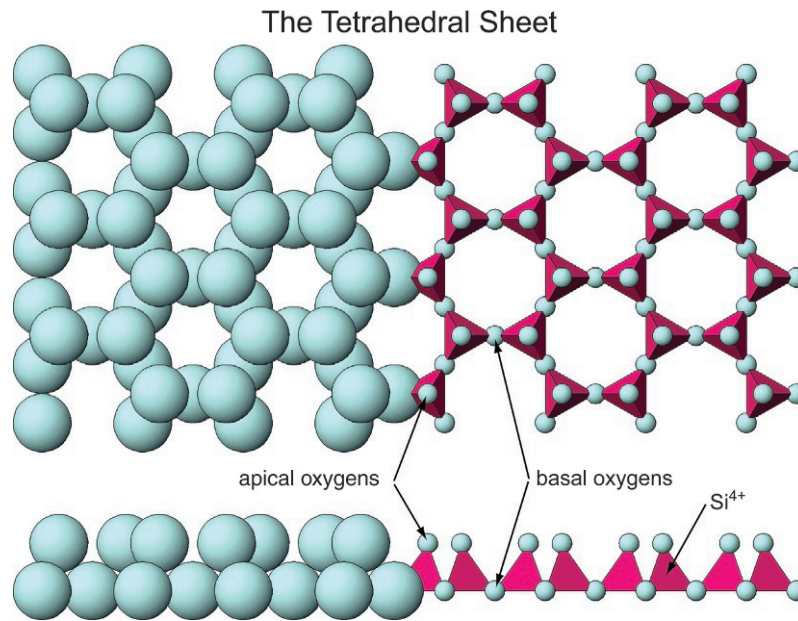


Fig. 2 The tetrahedral sheet as a sphere-packing model (left half) and a polyhedral model (right half). The upper half of the figure shows the view looking down onto the plane of the sheet, while the bottom half shows the view after rotating the sheet 90° to view the edge of the sheet. Adapted from Schulze DG (2002) An introduction to soil mineralogy. In: Dixon JB and Schulze DG (eds) *Soil Mineralogy with Environmental Applications*, pp. 1–35. Madison, WI: Soil Science Society of America, with permission.

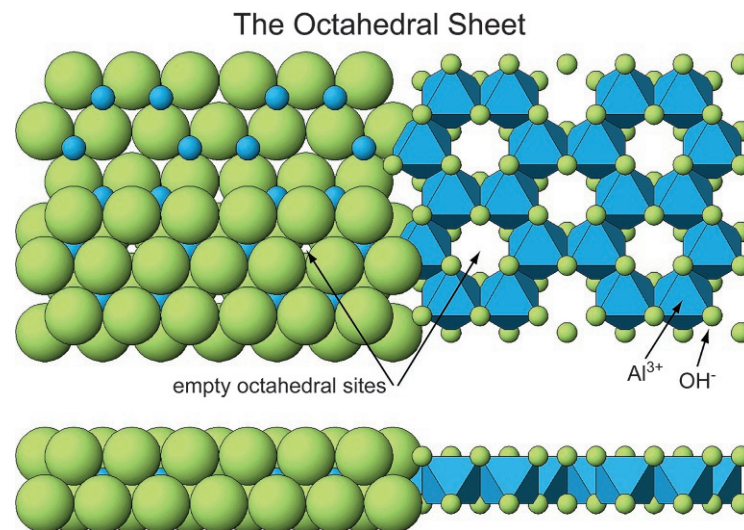


Fig. 3 The octahedral sheet as a sphere-packing model (left half) and a polyhedral model (right half). The top three rows of spheres have been omitted from the sphere-packing model so that the underlying cations, represented here as Al^{3+} , can be seen more easily. The upper half of the figure shows the view looking down onto the plane of the sheet, while the bottom half shows the view after rotating the sheet 90° to view the edge of the sheet. The dioctahedral arrangement is shown here, in which two-thirds of the possible octahedral sites are filled with cations and one-third are empty. Adapted from Schulze DG (2002) An introduction to soil mineralogy. In: Dixon JB and Schulze DG (eds) *Soil Mineralogy with Environmental Applications*, pp. 1–35. Madison, WI: Soil Science Society of America, with permission.

The octahedral sheet

Analogous to the tetrahedral sheet, we can consider the octahedral sheet illustrated in Fig. 3 as an assemblage of octahedra in which adjacent octahedra share two oxygens with one another. In other words, adjacent octahedra share edges. For the arrangement of octahedra shown in Fig. 3, the octahedral sites are occupied by trivalent cations, typically Al^{3+} , and for charge balance, a proton (H^+) must be associated with each O^{2-} . (The H^+ takes up very little space, and the OH^- ion can be considered a sphere of roughly the same size as an O^{2-} ion.) Each OH^- contributes one-half a negative charge to each cation because each OH^- is shared between two octahedra. Each Al^{3+} cation is therefore effectively surrounded by $6 \times 0.5 = 3$ negative charges and the sheet is electrically neutral. Note the pattern of empty and filled octahedral sites. Two of every three possible octahedral sites are filled when trivalent cations are

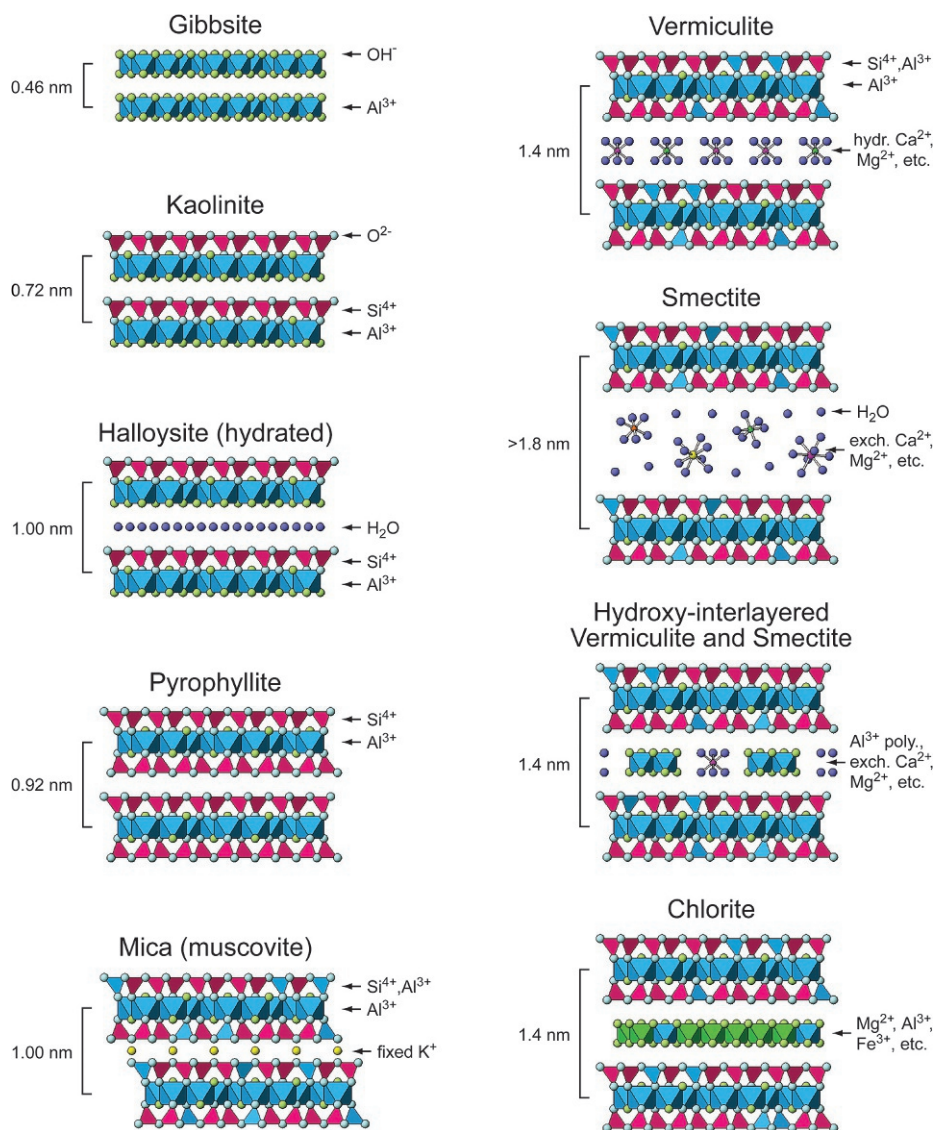


Fig. 4 Structural scheme of soil minerals based on tetrahedral and octahedral sheets. Adapted from Schulze DG (2002) *An introduction to soil mineralogy*. In: Dixon JB and Schulze DG (eds) *Soil Mineralogy with Environmental Applications*, pp. 1–35. Madison, WI: Soil Science Society of America, with permission.

present in the octahedral sites. This arrangement is called dioctahedral and is the most common in soil clay minerals. If the octahedral sites are filled with divalent cations such as Mg^{2+} , then every possible octahedral site must be occupied to produce an electrically neutral structure. This arrangement is called trioctahedral.

Octahedral sheets, stacked one on top of the other and held together by hydrogen bonds, make up the structure of gibbsite, $\text{Al}(\text{OH})_3$ (Fig. 4), an aluminum hydroxide mineral that occurs in intensively weathered soils. The structure of gibbsite is the simplest in a series of structures containing octahedral sheets.

Phyllosilicate minerals common in soil clays

Phyllosilicates are divided into two groups, 1:1- and 2:1-type minerals, based on the number of tetrahedral and octahedral sheets in the layer structure.

1:1-Type minerals

The 1:1 layer structure

The 1:1 layer structure consists of a unit made up of one octahedral and one tetrahedral sheet, with the apical O^{2-} ions of the tetrahedral sheets being shared with (and part of) the octahedral sheet (Fig. 5A). There are three planes of anions (Fig. 5B). One

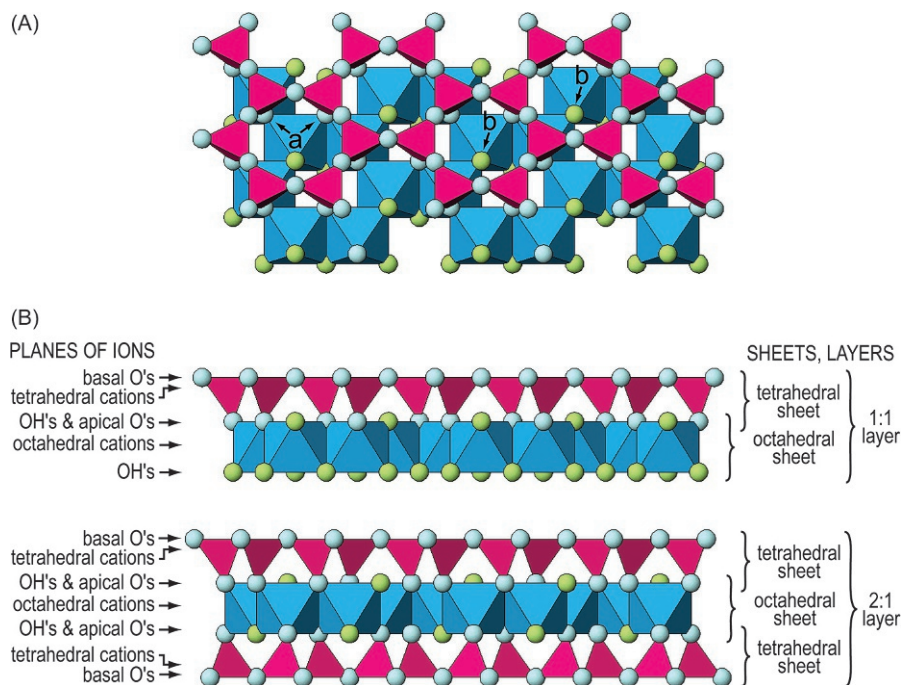


Fig. 5 (A) Oblique view of the 1:1 layer structure illustrating the relationship between the tetrahedral and octahedral sheets. Note that adjacent apical oxygens of the tetrahedral sheet (arrows 'a') also define corners of octahedra in the octahedral sheet. Arrows marked 'b' point to OHs that lie directly in the center of the hexagonal rings of tetrahedra, although they appear off-center in this oblique view. (B) Edge view of the 1:1 and 2:1 layer structures, illustrating phyllosilicate nomenclature. Adapted from Schulze DG (2002) An introduction to soil mineralogy. In: Dixon JB and Schulze DG (eds) *Soil Mineralogy with Environmental Applications*, pp. 1–35. Madison, WI: Soil Science Society of America, with permission.

plane consists of the basal O^{2-} ions of the tetrahedral sheet, the second consists of O^{2-} ions common to both the tetrahedral and octahedral sheets (marked 'a' in Fig. 5A) plus OH^- belonging to the octahedral sheet ('b' in Fig. 5A), and the third consists only of OH^- belonging to the octahedral sheet.

Kaolinite

The structure of kaolinite consists of 1:1 layers stacked one above the other. Kaolinite contains Al^{3+} in the octahedral sites and Si^{4+} in the tetrahedral sites (Fig. 4). The 1:1 layer is electrically neutral and adjacent layers are held together by hydrogen bonding between the basal oxygens of the tetrahedral sheet and the hydroxyls of the exterior plane of the adjacent octahedral sheet.

Kaolinite is a common mineral in soils and is the most common member of this subgroup. It tends to be particularly abundant in more weathered soils such as Ultisols and Oxisols (Soil Taxonomy, Soil Survey Staff, 2022) and Acrisols, Ferralsols, and Nitisols (IUSS Working Group WRB, 2022). Kaolinite has very little isomorphous substitution in either the tetrahedral or octahedral sheets and most kaolinite has close to the ideal formula $Al_2Si_2O_5(OH)_4$. The 1:1 layer has little or no permanent charge because of the low amount of substitution. Consequently, cation exchange capacities and surface areas are typically low. Kaolinite in soils typically occurs as thin, rounded, pseudo-hexagonal crystals that can be up to $\sim 2 \mu m$ across (Fig. 6A) and therefore is often concentrated in the coarse clay fraction of a soil.

Kaolinite can form in soils from Al and Si released by the weathering of primary and other secondary minerals. For example, feldspars often weather to kaolinite in soils formed from igneous rocks. Kaolinite can also be inherited from clayey, sedimentary soil parent materials. Soils high in kaolinite are generally less fertile than those in which 2:1 clay minerals dominate. This is because of the low cation exchange capacity of kaolinite, and because the long and/or intense leaching conditions that result in kaolinite becoming the most abundant clay mineral also result in the destruction of most of the weatherable minerals that could provide nutrients to plants.

Halloysite

Halloysite has a 1:1 layer structure similar to that of kaolinite except that the 1:1 layers are separated by a layer of H_2O (water) molecules when fully hydrated (Fig. 4). This water is probably present as hydration shells around a small number of interlayer cations (cations that reside between two adjacent 1:1 or 2:1 layers), although the presence of interlayer cations and the existence of a layer charge to attract them has been difficult to confirm. Halloysite often occurs as tubular (Fig. 6B) or sometimes spherical particles.

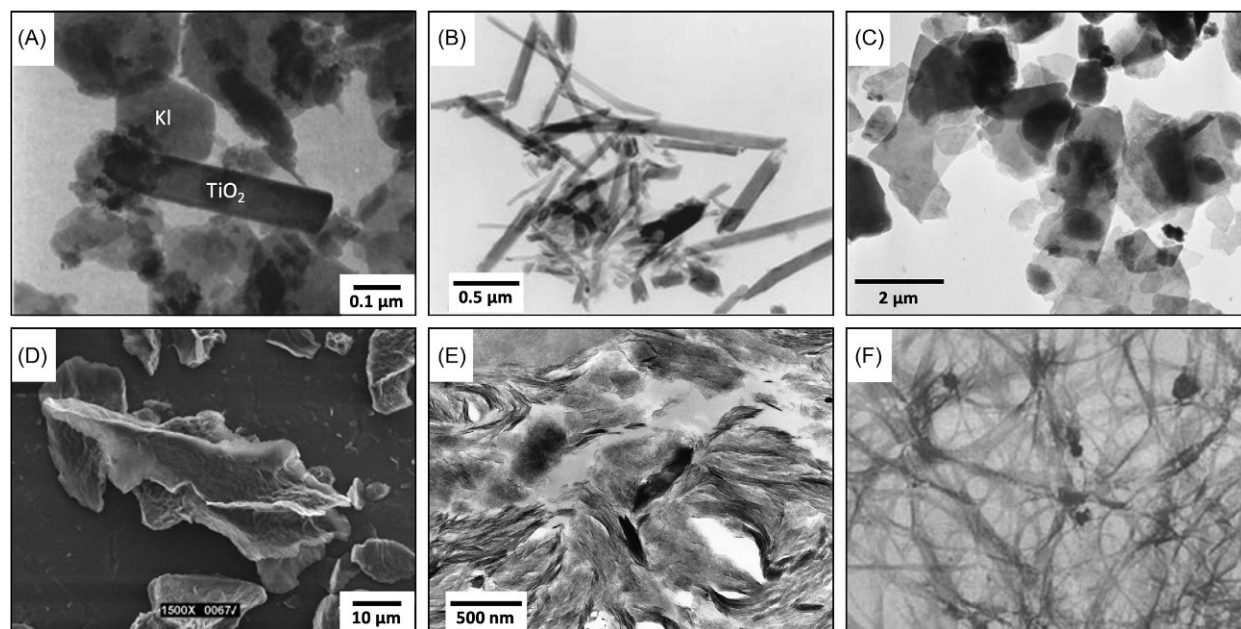


Fig. 6 Transmission (TEM) and scanning (SEM) electron micrographs of some soil clay minerals. (A) Pseudo-hexagonal kaolinite crystals (KI) predominate in this TEM image of the 2–0.2 μm fraction of an Oxisol (ST) or Ferralsol (WRB) from Sete Lagoas, Brazil. The rectangular particle labeled TiO_2 is either anatase or rutile, while the darker diffuse clusters are iron oxides, either goethite or hematite. (B) Halloysite tubes from the saprolite below an Ultisol in the Piedmont Region of Alabama, USA. (C) A mixture of 2:1 clay minerals predominate in this TEM image of the $<2 \mu\text{m}$ clay fraction of the Drummer silt loam soil [fine-silty, mixed, superactive, mesic Typic Endoaquolls (ST) or Gleyic Phaeozems (Siltic) (WRB)] from the Purdue Agronomy Center for Research and Education, Tippecanoe County, Indiana. The X-ray diffraction analysis shows the presence of illite, vermiculite, smectite, and hydroxy-interlayered vermiculite together with kaolinite. (D) SEM image of soil smectite quasicrystals. See text for details. (E) TEM image of a $< 100 \text{ nm}$ thin slice of an aggregate from an Iowa Mollisol (ST) or Phaeozem (WRB) showing smectite quasicrystals in situ. See text for details. (F) TEM image of thread-like bundles of imogolite tubes in a natural sample. Thread widths are 10–30 nm across. (A) From de Brito Galvão TC and Schulze DG (1996) Mineralogical properties of a collapsible lateritic soil from Minas Gerais, Brazil. *Soil Science Society of America Journal* 60:1969–1978. (B) TEM image from White GN and Dixon JB (2002) Kaolin-serpentine minerals. In Dixon JB and Schulze DG. *Soil Mineralogy with Environmental Applications*, pp. 389–414. Madison, WI: Soil Science Society of America. (C) Image by D. G. Schulze. (D) Image courtesy of D. A. Laird, downloaded from <https://www.minersoc.org/images-of-clay.html>. (E) Image courtesy of M. L. Thompson. (F) From Harsh J, Chorover J, and Nizeyimana E (2002) Allophane and imogolite. In Dixon JB, and Schulze DG. *Soil Mineralogy with Environmental Applications*, pp. 291–322. Madison, WI: Soil Science Society of America.

Halloysite is usually found in soils formed from volcanic deposits, particularly volcanic ash and glass, and it is a common clay mineral in Andisols (ST) or Andosols (WRB). Halloysite forms early in the process of weathering of primary minerals, but it is less stable than kaolinite and transitions to kaolinite with time.

2:1-Type minerals

In contrast to the 1:1 phyllosilicate minerals, which are represented in soils by only two major minerals, the 2:1 minerals are structurally more diverse and are represented by several common mineral species.

The 2:1 layer structure

The 2:1 layer structure consists of two tetrahedral sheets, with one bound to each side of an octahedral sheet (Fig. 5B). There are four planes of anions. The outer two consist of the basal oxygens of the two tetrahedral sheets, while the two inner planes consist of oxygens common to the octahedral sheet and the two tetrahedral sheets, plus the hydroxyls belonging only to the octahedral sheet.

Pyrophyllite

The simple structure of pyrophyllite is a good starting point for discussing 2:1 structures. Pyrophyllite consists of 2:1 layers stacked one above the other (Fig. 4). The tetrahedral sheets contain only Si^{4+} and the octahedral sheet contains only Al^{3+} , resulting in the ideal formula $\text{Al}_2(\text{Si}_4)\text{O}_{10}(\text{OH})_2$. The charge is balanced completely within the 2:1 layer, making the layer electrically neutral, and adjacent 2:1 layers are held together only by weak van der Waals forces. Pyrophyllite occurs only rarely in soils, usually only when it is inherited from low-grade metamorphic rocks.

Micas

Mica minerals have the 2:1 layer structure described for pyrophyllite but with two important differences. First, instead of having only Si^{4+} in the tetrahedral sites, one-quarter of the tetrahedral sites are occupied by Al^{3+} . Because of this substitution, there is an

excess of one negative charge per formula unit in the 2:1 layer. Second, this excess negative charge is balanced by monovalent cations, commonly K^+ , that occupy interlayer sites between two 2:1 layers (Fig. 4). This gives an ideal formula of $KAl_2(AlSi_3)O_{10}(OH)_2$ for a mica mineral with Al in the octahedral sites.

The octahedral sheet can contain either Al^{3+} (the dioctahedral case; Fig. 3) or Mg^{2+} (the trioctahedral case). There are several different mica species because micas can be either dioctahedral or trioctahedral, Fe^{2+} and Fe^{3+} can substitute for Mg^{2+} and Al^{3+} in the octahedral sheet, and Na^+ and Ca^{2+} can substitute for K^+ in the interlayer.

Muscovite, biotite, and phlogopite are the three most common mica group minerals in rocks, and consequently in soils. All three contain K in the interlayer, but they differ in the composition of the octahedral sheet and whether they are di- or trioctahedral. Mica in the clay fraction of soils and sediments differs somewhat from the macroscopic muscovite mica it closely resembles. This clay-size dioctahedral mica is often referred to as illite. Glauconite is another mica mineral that is similar to illite, but it contains more Fe and less Al in its octahedral sheet than illite.

Micas in soils are usually inherited from the parent rock and occur widely in soils derived from various igneous and metamorphic rocks, in soils derived from illitic sedimentary rocks, and from sediments derived from them. In glaciated regions, mica in soils may be a mixture of micas from igneous and metamorphic rocks together with mica or illite from sedimentary rocks. As a result of soil processes, micas can be transformed (weathered) to other minerals, particularly to vermiculites and smectites, and the K^+ released during weathering is an important source of K for plants. The dioctahedral micas such as muscovite are usually more resistant to weathering than trioctahedral micas because the interlayer K^+ ions are held more tightly in the interlayer sites of dioctahedral micas than in trioctahedral micas. Thus, muscovite tends to be the most common mica mineral in soils. The thin, irregularly shape plates in Fig. 6C are a mixture of mica together with vermiculite, smectite, and hydroxy-interlayered vermiculite.

Vermiculites

Vermiculite has a 2:1 layer structure as described for mica, but instead of having a layer charge of ~ 1 per formula unit and K^+ in interlayer positions, vermiculite has a layer charge of 0.9–0.6 per formula unit and contains hydrated exchangeable cations, primarily Ca and Mg, in the interlayer (Fig. 4). A typical formula for an idealized vermiculite weathered from muscovite is $M_{0.75}^+Al_2(Si_{3.25}Al_{0.75})O_{10}(OH)_2$, where M^+ represents charges from exchangeable cations. The lower charge per formula unit of vermiculite, compared to micas, allows the interlayer to open, allowing exchangeable cations to move in and out. The layer charge of 0.9–0.6 per formula unit gives vermiculite a high cation exchange capacity, but a strong affinity for weakly hydrated cations such as K^+ , NH_4^+ (ammonium), and Cs^+ (cesium). Fixation of K^+ in the interlayer positions of vermiculite can be significant in soils that have a large vermiculite content and have not received large amounts of inorganic fertilizers.

Vermiculites in soils are believed to form almost exclusively from the weathering of micas and chlorites. The weathering of micas to vermiculite (or smectite) is believed to occur by replacement of K^+ in the interlayer sites with hydrated exchangeable cations. The integrity of the 2:1 layer is preserved, but there is a reduction in the layer charge. Vermiculite does not swell as extensively as smectite, and this is shown in Fig. 4 by the presence of only two planes of water molecules surrounding the hydrated cations in the interlayer space.

Smectites

The smectite group consists of minerals with the 2:1 structure already discussed for mica and vermiculite, but with a still lower charge per formula weight, namely 0.6–0.2. As for vermiculite, the interlayer contains exchangeable cations (Fig. 4). An idealized formula for a common soil smectite, the mineral beidellite, is $M_{0.33}^+Al_2(Si_{3.67}Al_{0.33})O_{10}(OH)_2$, where M^+ represents charges from hydrated, exchangeable cations, typically Ca^{2+} and Mg^{2+} .

The most common smectite minerals range in composition between three end-members: montmorillonite, beidellite, and nontronite. All are dioctahedral, but they differ in the composition of the tetrahedral and octahedral sheets. Since the layer charge of smectite is low compared to vermiculite, smectites do not fix K^+ as readily as does vermiculite, but since the interlayer cations are not as strongly attracted by the 2:1 layers, they attract more water molecules and as a result, smectite swells more extensively upon wetting than does vermiculite.

The morphology of soil smectite crystals is not well documented. The typical procedures used to separate and concentrate soil clay fractions for examination by electron microscopy rely on dispersion of the clay particles, usually by saturation of the cation exchange sites with Na^+ and rigorous mechanical agitation via sonication, followed by centrifugation to obtain the clay fraction, and then flocculation of the separated clay for further study. Minerals like kaolinite, halloysite, and mica are usually not greatly altered by these procedures and occur as distinct, well-defined crystals (Fig. 6A–C). Smectite, however, can be greatly altered, even to the point of delamination of the individual 2:1 smectite layers during dispersion, and then reforming them upon flocculation in a form that may not have been present in the original sample.

Work by Laird and Thompson (Laird et al., 1991; Laird and Thompson, 2009) sheds light on the morphology of smectite in soils. They studied specimens from the surface horizon of the Webster silty clay loam soil [fine-loamy, mixed, superactive, mesic Typic Endoaquoll (ST); Gleyic Phaeozems (Loamic) (WRB)] on the Wisconsin Age till plain of southern Minnesota and north-central Iowa, USA. The $<0.02 \mu m$ fine clay fraction of this soil is overwhelmingly dominated by smectite and randomly interstratified smectite/illite. Fig. 6D shows ‘reformed smectite quasicrystals’ from the $<0.02 \mu m$ e.s.d (equivalent spherical diameter) fine clay fraction that had been saturated with Na^+ ions, dispersed, centrifuged to obtain the $<0.02 \mu m$ fraction, and then flocculated with $CaCl_2$. A drop of suspension was dried and imaged by scanning electron microscopy. These particles are described as ‘reformed’ smectite quasicrystals because they are believed to have been formed upon flocculation of the delaminated Na^+ -saturated clay

(D.A. Laird, personal communication, Sept. 2022). Most of the flocculated particles are 20–50 μm across, which is considerably larger than the $<0.02 \mu\text{m}$ e.s.d. based on the fractionation conditions. This is attributed to the slow settling velocity of the platy, low density, hydrated, dispersed smectite particles which settle considerably slower in a centrifugal field than more equidimensional particles like quartz and kaolinite with particle densities of $\sim 2.65 \text{ g cm}^{-3}$ (Laird and Thompson, 2009). The delaminated Na^+ -saturated 2:1 layers may be as thin as $\sim 1 \text{ nm}$, but several micrometers in diameter. When the Na^+ ions are replaced with Ca^{2+} ions, the planar surfaces of the 2:1 layers are attracted to the Ca^{2+} ions and water molecules in between. The structure that results is called a *quasicrystal*; there is order between adjacent 2:1 layers, but the structure overall does not have large-scale periodicity like a true crystal. The large, floppy quasicrystals crumple and bend into the randomly shaped particles shown in Fig. 6D.

Fig. 6E shows that similar quasicrystals of smectite occur in soils. For this image, a small ($\sim 2 \text{ mm}$ diameter) field moist, minimally disturbed soil aggregate from the surface horizon of the Webster soil was examined using techniques usually used for biological samples. In brief, the sample was dewatered by solvent exchange, stained, imbedded in epoxy resin, and sectioned with a diamond knife to produce slices $<100 \text{ nm}$ thick. These were then examined by scanning transmission electron microscopy (Laird and Thompson, 2009). Although the focus of the study was on imaging clay–humic complexes, the clay fraction dominated images such as Fig. 6E. Note that the entire field of view is only $\sim 3 \mu\text{m}$ across. The areas labeled QC are interpreted as places where the crumpled and bent quasicrystals of smectite have been sliced perpendicular to the plane of the 2:1 layers, while the more diffuse gray areas are areas where the quasicrystals have been sliced parallel or oblique to the plane of the 2:1 layers. Some of the gray areas may also be organic matter.

Smectites are important minerals in many temperate-region soils. Plant nutrient cations (Ca, Mg, K, and others) are held in a bioavailable form on the cation exchange sites of soil smectites. Soils rich in smectite tend to be very effective at attenuating many organic and inorganic pollutants because of the large surface area and adsorptive properties of the smectites. Smectite quasicrystals shrink upon drying and swell upon wetting. This shrink–swell behavior is most pronounced in Vertisols (ST, WRB) and in vertic subgroups of other soil orders (ST) or Reference Soil Groups with Vertic qualifiers (WRB). The shrink–swell properties lead to cracking and shifting problems when houses, roads, and other structures are built on smectitic soils.

Chlorites

Like mica, chlorite minerals have a 2:1 layer structure with an excess of negative charge. In contrast to mica, however, the excess charge is balanced by a positively charged interlayer hydroxide sheet (Fig. 4), rather than K^+ . The interlayer hydroxide sheet is an octahedral sheet as shown in Fig. 3 and can be either di- or trioctahedral. Instead of being electrically neutral as in gibbsite, the hydroxide sheet has a positive charge caused by substitution of higher-valence cations for lower-valence ones, for example, $\text{Mg}_2\text{Al}(\text{OH})^+_6$. Either octahedral sheet – the one that is part of the 2:1 layer or the interlayer hydroxide sheet – can be dioctahedral or trioctahedral, and can contain Mg^{2+} , Fe^{2+} , Mn^{2+} , Ni^{2+} (nickel), Al^{3+} , Fe^{3+} , and Cr^{3+} (chromium), resulting in many different mineral species.

Chlorite minerals in soils are often primary minerals inherited from either metamorphic or igneous rocks. They may also be inherited from sedimentary rocks such as shales, or from hydrothermally altered sediments. Chlorites are rather infrequent minerals in soils, and when present they generally occur in small amounts. Chlorite weathers to form vermiculite and smectite, and the ease with which chlorites break down makes them sensitive indicators of weathering.

Hydroxy-interlayered vermiculite and smectite

Hydroxy-interlayered vermiculite and smectite can be considered a solid solution with vermiculite or smectite as one end-member and chlorite as the other. Hydroxy-interlayered minerals form as Al^{3+} released during weathering hydrolyzes and polymerizes to form large polycations with a postulated formula of $\text{Al}_6(\text{OH})_{15}^{3+}$ (or similar) in the interlayers of vermiculite and smectite. These polycations balance some of the charge of the 2:1 layer. The combination of a 2:1 layer with hydroxy Al in the interlayer gives a structure similar to that of chlorite (Fig. 4). Thus, these minerals have also been called secondary chlorites. The degree of filling of the interlayer with hydroxy Al can vary from none to almost complete, with properties of the clay varying accordingly. The interlayer hydroxy Al is not exchangeable with other cations and its presence reduces the cation exchange capacity of smectite or vermiculite almost linearly as a function of the amount of Al adsorbed in the interlayer.

Interlayer hydroxy Al prevents smectite from shrinking and swelling as it normally would because the interlayer surfaces are blocked, and the movement of exchanging cations and water molecules is restricted. In vermiculite, interlayer hydroxy Al reduces K^+ fixation by lowering the exchange capacity and by preventing the interlayer from collapsing around the K^+ . The positively charged hydroxy interlayers also provide potential sites for anion adsorption. Hydroxy-interlayered vermiculite and smectite are most common in Alfisols and Ultisols (ST) and Luvisols, Acrisols, Retisols, Lixisols, and Alisols (WRB). Within a given soil profile, they tend to be most abundant near the soil surface where weathering is most intense.

Interstratification in phyllosilicates

Because of the structural similarities of all the phyllosilicate minerals just discussed, phyllosilicates in soils do not always occur as discrete particles of mica, vermiculite, smectite, chlorite, or kaolinite. For example, instead of being made up of a stack of identical 2:1 vermiculite layers, one physically discrete particle may consist of a mixture of both mica and vermiculite layers instead. Such minerals are referred to as ‘mixed-layer’ or ‘interstratified minerals.’

Different types of interstratified minerals have been identified. Two-component systems include: mica–vermiculite, mica–smectite, mica–chlorite, kaolinite–smectite, and others. Three-component mixed layer systems can also occur. The sequence

of layers can be either regular or random. A regularly interstratified mineral consisting of two types of layers denoted by A and B could have a sequence such as ABABAB ..., or ABBABBABB ..., or any other repeating sequence. In a randomly interstratified mineral, the sequence of layers is random – for example, ABBABAABBAAA Random interstratification of layer-silicates is more common in soils than regular interstratification, but regular interstratification, especially in weathering micas, is not rare.

Partial removal of interlayer K from micas or of interlayer hydroxide from chlorite is one way that interstratified minerals can form in soils. Other possibilities include fixation of adsorbed K^+ by some vermiculite layers to give mica-like layers, and the formation of hydroxide interlayers to produce chlorite-like layers. There is also the possibility that random interstratification may be an artifact of the laboratory methods used to disperse and isolate the soil clay fraction for identification and characterization.

Palygorskite and sepiolite

Palygorskite and sepiolite are considered phyllosilicates but are distinct structurally from the typical 1:1 and 2:1 layer structures. Both minerals have continuous tetrahedral sheets, but adjacent bands of tetrahedra within one tetrahedral sheet point in opposite directions rather than in one direction as in the 1:1 and 2:1 structures. The result is a structure that can be described as ribbons of 2:1 layers joined at their edges, as illustrated in Fig. 7. Water molecules occur in the spaces between the ribbons. The 2:1 ribbons are wider in sepiolite than in palygorskite. Both minerals have a fibrous morphology in contrast to the platy morphology of most 1:1 and 2:1 minerals.

Palygorskite and sepiolite are often found in soils of arid and semiarid environments, either inherited from the soil parent material or as products of pedogenesis. The stability of both minerals requires alkaline conditions and strong Si and Al activity in solution, conditions that are most likely to occur in arid and semiarid soil environments.

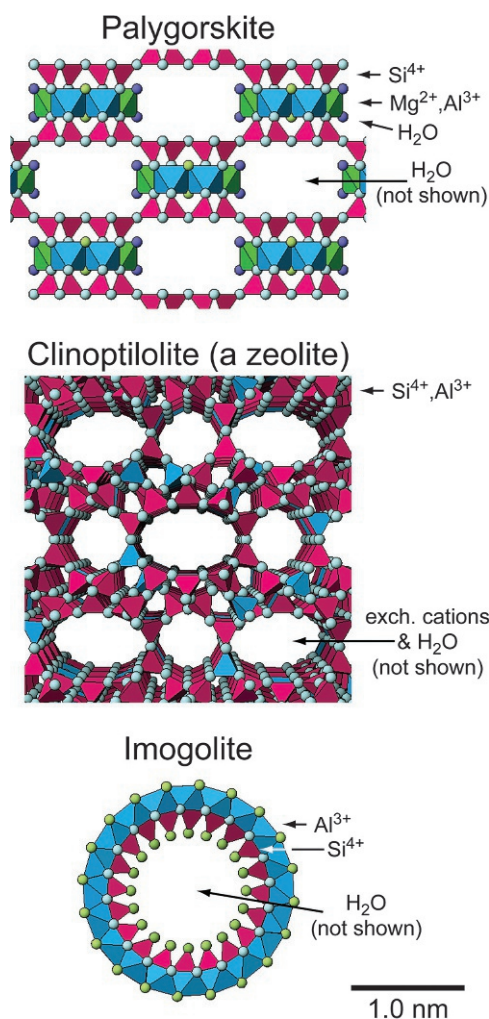


Fig. 7 Structural models of representatives of three other aluminosilicate mineral groups that occur frequently in soils. All three are drawn to the same scale. Adapted from Schulze DG (2002) An introduction to soil mineralogy. In: Dixon JB and Schulze DG (eds) *Soil Mineralogy with Environmental Applications*, pp. 1–35. Madison, WI: Soil Science Society of America, with permission.

Other minerals that occur in the clay fraction of soil

The phyllosilicate clay minerals described above are the most abundant and common in most soils, and they are the minerals that are usually considered to make up the 'clay minerals' group. Phyllosilicates, however, are not the only minerals that occur in the soil clay fraction, nor are they the only minerals that form as primary minerals from igneous and metamorphic rocks weather in soils. Several additional minerals or mineral groups require mention also. Some occur only in particular soils, others occur in many different types of soils, but usually only at low concentrations.

Zeolites

Zeolites are a large group of aluminosilicate minerals that consist structurally of SiO_4 tetrahedra arranged in ways that result in large amounts of pore space within the crystals (Fig. 7). Aluminum substitutes for Si in the tetrahedral sites and, as a result, the (Si,Al) O_4 framework has a net negative charge. The charge is balanced by cations that reside in the channels and pores together with water molecules. Because the cations are exchangeable, zeolites have cation exchange properties similar to the phyllosilicates, but, because the tetrahedral framework of the zeolites is rigid and the size of the pores is fixed, small cations can move into and out of the pores freely, while larger cations are excluded. Thus, zeolites are often referred to as 'molecular sieves' because of their very selective cation exchange properties. Zeolites are relatively rare in soils because they weather and dissolve easily in humid regions, but they occur in some soils of arid regions.

Allophane and imogolite

The aluminosilicate minerals discussed above have three-dimensional crystal structures, with atoms packed together in a more or less regular manner over relatively long distances (e.g., 10s of nanometers). They exhibit long-range order. Two other aluminosilicates, allophane and imogolite, exhibit short-range (or local) order. Structures with short-range order exhibit order over several nanometers, but on a larger scale the structure is disordered.

Allophane consists chemically of variable amounts of O^{2-} , OH^- , Al^{3+} , and Si^{4+} , and is characterized by short-range order and a predominance of Si—O—Al bonds. It consists of small (3.5–5.0 nm) spheres. The spheres clump together to form irregular aggregates. Imogolite consists of tubes several micrometers long with an outer diameter of 2.3–2.7 nm and an inner diameter of approx. 1.0 nm. The tubes consist of a single dioctahedral sheet with the inner surface OH replaced by SiO_3OH groups (Fig. 7). Several individual tubes are arranged in bundles 10–30 nm across to give thread-like particles several micrometers long (Fig. 6F).

Allophane and imogolite usually occur as weathering products of volcanic ash and are important minerals in Andisols (ST) and Andosols (WRB). Imogolite has also been identified in the Bs horizons of Spodosols (ST) and Podzols (WRB). Allophane and imogolite can adsorb many specific inorganic and organic compounds. Andisols (ST) and Andosols (WRB), for example, usually fix large amounts of phosphate, making it unavailable to plants, and the large amounts of organic matter common in Andisols (ST) and Andosols (WRB) may be due, in part, to adsorption of organic molecules by allophane and imogolite. Soils containing large amounts of allophane and imogolite usually have unique physical properties such as a low bulk density, high water-holding capacity, high liquid and plasticity limits, and a thixotropic consistence.

Aluminum hydroxide minerals

Gibbsite, which has already been mentioned as simply a stack of octahedral sheets containing Al^{3+} in the octahedral sites (Fig. 4), occurs in situations where weathering and leaching have been intense or long. Gibbsite often occurs in Ultisols and Oxisols (ST) and Acrisols, Ferralsols, and Nitisols (WRB), and can be important in the fixation of phosphate fertilizers.

Iron oxide minerals

Iron is almost as abundant as Al in the Earth's crust, and one or more strongly colored iron oxide minerals are ubiquitous accessory minerals in almost all soil clays. Goethite (FeOOH) is the most common soil iron oxide mineral and accounts for the yellowish-brown colors of many soils. Hematite (Fe_2O_3) is common also and accounts for the red colors of many soils. Lepidocrocite (FeOOH) (orange) occurs in noncalcareous hydromorphic soils, while maghemite (Fe_2O_3) (reddish brown) occurs in some soils of the tropics and subtropics. Two poorly crystalline minerals, ferrihydrite ($\sim\text{Fe}_5\text{HO}_8 \cdot 4\text{H}_2\text{O}$) (reddish brown) and schwertmannite [$\text{Fe}_8\text{O}_8(\text{OH})_6\text{SO}_4$] (yellowish brown), occur in surface environments when waters rich in Fe^{2+} are exposed and rapidly oxidized in the presence of organic compounds (ferrihydrite) or sulfate (schwertmannite). Lithogenic magnetite (Fe_3O_4) (black) is often found in the sand fraction of soils.

The surfaces of iron oxide minerals are highly reactive and can sorb and fix phosphate, many transition metals, and organic compounds. Oxidation–reduction reactions are an important aspect of soil iron oxide mineralogy. Iron oxide minerals can be dissolved under reducing conditions and then be reprecipitated, often in a different mineral form, when oxidizing conditions return. The spatial distribution of iron oxide minerals that develop under redoximorphic conditions are important indicators of the patterns of water saturation and redox conditions that occur in soils over time.

Manganese oxide minerals

Manganese is about 50 times less abundant in the Earth's crust than Fe, so minerals containing large amounts of Mn are considerably less abundant. Nevertheless, some Mn oxide minerals occur in almost all soils. Mn oxides in soils are often formed in response to microbially mediated oxidation of Mn. These crystals have structures and chemical formulas similar to those of birnessite $((\text{Na,Ca})(\text{Mn}^{3+}, \text{Mn}^{4+})_7\text{O}_{14} \cdot 2.8\text{H}_2\text{O})$. The Mn^{4+} and Mn^{3+} in manganese oxide minerals is reduced to soluble Mn^{2+} under relatively mild reducing conditions, making Mn oxide minerals important in many soil oxidation–reduction reactions, including enzyme reactions associated with soil carbon cycling. Like Fe oxides, Mn oxides have a large capacity to adsorb both trace metals and phosphate ions in soils.

Titanium oxide minerals

Titanium is slightly more abundant than Mn in the Earth's crust, and two Ti minerals, anatase and rutile (both TiO_2), occur widely in soil clays. The Ti-oxide minerals, however, are relatively inert chemically and have a negligible impact on most soil properties. Fig. 6A shows a rectangular Ti-oxide mineral in association with kaolinite in the clay fraction from an Oxisol (ST) or Ferralsol (WRB).

Carbonates, sulfates, and soluble salts

In soils of semiarid to arid climates, calcite (CaCO_3), gypsum ($\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$), and an array of evaporite minerals (minerals with solubilities greater than or equal to gypsum) often occur co-associated with the aluminosilicate and oxide minerals discussed above. The chemical and physical impacts of these minerals on soil processes varies greatly, depending on the amount and type of mineral present.

Summary

The clay fraction of soils consists of a variety of fine-grained secondary minerals that are formed as primary minerals in igneous and metamorphic rocks weather in the aqueous environment at the Earth's surface. The most common soil clay minerals are a variety of aluminosilicate minerals with chemical structures based on planar sheets of Si tetrahedra and Al octahedra. These aluminosilicate clay minerals are co-associated with a variety of Fe, Mn, and Ti oxide and hydroxide minerals, as well as carbonate, sulfate, sulfide, and chloride minerals, depending on the particular soil environment. Minerals in the soil clay fraction impart many of the physical and chemical properties that we associate with soils including cation exchange capacity, water holding capacity, shrink-swell capacity, and other properties.

Acknowledgments

I thank Michael Thompson and Margaret Oliver for helpful comments for improving the manuscript. I also thank Curtis Monger for checking my translations of soil taxonomic classifications from *Soil Taxonomy* into the *World Reference Base for Soil Resources* (WRB).

References

- Dixon JB and Schulze DG (2002) *Soil Mineralogy with Environmental Applications*. Madison, WI: Soil Science Society of America.
- Dixon JB and Weed SB (1989) *Minerals in Soil Environments*. Madison, WI: Soil Science Society of America.
- IUSS Working Group WRB (2022) *World Reference Base for Soil Resources. International soil classification system for naming soils and creating legends for soil maps*, 4th edn. Vienna, Austria: International Union of Soil Sciences (IUSS).
- Klein C and Dutrow B (2008) *Manual of Mineral Science*, 23rd edn. John Wiley & Sons.
- Laird DA and Thompson ML (2009) The ultrastructure of clay-humic complexes in an Iowa Mollisol. In: Laird DA and Cervini-Silva J (eds.) *CMS Workshop Lectures. Carbon Stabilization by Clays in the Environment: Process and Characterization Methods*, vol. 16, pp. 96–118. Chantilly, VA: The Clay Minerals Society.
- Laird DA, Barak P, Nater EA, and Dowdy RH (1991) Chemistry of smectitic and illitic phases in interstratified soil smectite. *Soil Science Society of America Journal* 55: 1499–1504.
- Soil Survey Staff (2022) *Keys to Soil Taxonomy*, 13th edn. Washington, DC: USDA-Natural Resources Conservation Service.

Short-range ordered aluminosilicates

Tetsuhiro Watanabe^a, James B Harsh^b, and Rota Wagai^c, ^aKyoto University, Kyoto, Japan; ^bWashington State University, Pullman, WA, United States; ^cInstitute for Agro-Environmental Science, National Agriculture and Food Research Organization, Tsukuba, Ibaraki, Japan

© 2023 Elsevier Ltd. All rights reserved.

This is an update of J. Harsh, Amorphous Materials, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 64–71, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00207-1>.

Introduction	121
Formation and distribution	122
Identification	124
Structure and charge	127
Chemisorption and ligand exchange	129
Reaction with soil organic matter	130
Aluminum solubility	131
Physical properties	131
Implications of short-range ordered aluminosilicates for agricultural productivity	132
Outlook	133
References	133

Abstract

Short-range ordered aluminosilicates, represented by allophane and imogolite, occur in large quantities in Andosols and to a lesser degree in Podzols. With their nano-scale crystallinity, these minerals contribute to a large specific surface area and reactive functional groups and, therefore, strongly influence the physicochemical properties of soils, such as large sorption capacity of phosphate and organic matter, high water retention, and low bulk density. In this chapter, the formation and distribution, identification, and fundamental chemical and physical properties of these minerals are described, together with their agricultural and environmental importance.

Key points

- The formation and distribution of short-range ordered (SRO) aluminosilicates (allophane and imogolite) are explained.
- Identification and structural properties of SRO aluminosilicates, and their reaction with anions, cations, and organic matter are described.
- The environmental and agricultural importance of SRO aluminosilicates is discussed, including nutrient availability, 'heavy' metal sorption, organic matter preservation, and agricultural productivity.

Introduction

Short-range ordered (SRO) aluminosilicates in soils consist primarily of allophane and imogolite, which have nano-sized spherical (allophane, Fig. 1) and tubular (imogolite, Fig. 2) structures, respectively. They also include imogolite precursors (proto-imogolite). These SRO aluminosilicates have been variously described as amorphous, non-crystalline, poorly crystalline, nano-crystalline, and SRO materials, minerals and aluminosilicates. However, 'amorphous' and 'non-crystalline' are not technically correct terms for allophane and imogolite as they possess SRO structures; allophane was previously assumed to be amorphous (non-crystalline) because it does not show clear X-ray diffraction patterns. 'Nano-crystalline' is not widely used but may be a more suitable descriptor than SRO because imogolite has long-range order in the longitudinal direction.

Large amounts of SRO aluminosilicates are predominantly found in soils derived from volcanic ash parent materials, but they also occur in other soils under a variety of parent materials (Parfitt, 2009). Their abundance is commonly estimated by selective dissolution using acid oxalate, as acid oxalate-extractable Si (Si_o) and Al (Al_o) are derived from SRO aluminosilicates and organo-Al complexes (see 'Identification' section below). Organo-Al complexes are estimated by sodium pyrophosphate extraction (Al_p), and hence the $\text{Al}_o - \text{Al}_p$ fraction is attributed to SRO aluminosilicates. The Al/Si (aluminum/silicon) ratio, calculated by $(\text{Al}_o - \text{Al}_p)/\text{Si}_o$, is 2 for imogolite and ranges from 1 to 2 for allophane. Allophane with an Al/Si ratio of 2 is called Al-rich or proto-imogolite allophane, whereas that with an Al/Si ratio of 1 is termed Si-rich allophane.

The SRO aluminosilicates are characterized not only by their poor crystallinity, but also by their small particle size which imparts a large specific surface area, and their variable surface charge properties. These characteristics make them important ion exchange

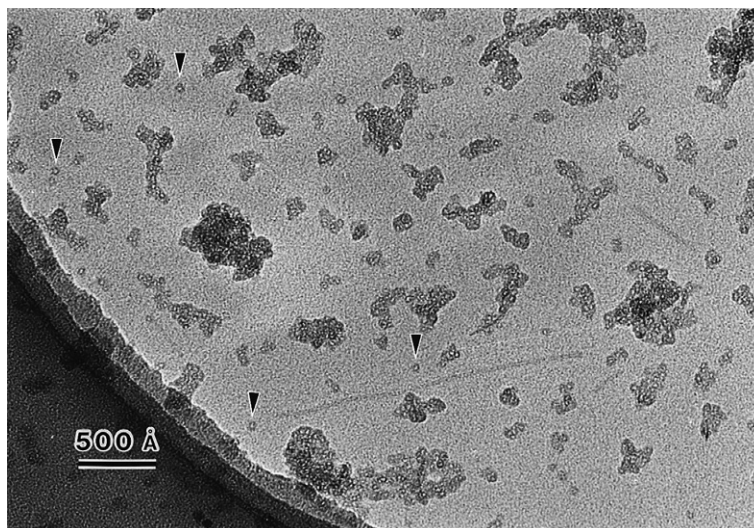


Fig. 1 Transmission electron micrograph of natural allophane particles (negatively stained with LaCl_3 , lanthanum(III) chloride). Individual (black arrows) and partially-aggregated allophane particles isolated from weathered pumice (Kitakami, Iwate, Japan) dispersed on hydrophilic carbon film. Reproduced from Wada SI and Wada K (2014) Visualization of the hollowness in unit particles of allophane and imogolite, *Journal of the Faculty of Agriculture Kyushu University*, 59(2): 369–372, permission obtained from publisher.



Fig. 2 Transmission electron micrographs of natural imogolite with an external diameter of about 2 nm. By courtesy of N. Yoshinaga.

materials even when present at relatively low concentrations. The SRO aluminosilicates also play important roles in regulating Al^{3+} and H_4SiO_4 (orthosilicic acid) activities of the soil solution, chemical and physical stabilization of organic matter (OM) and soil physical properties, such as bulk density and water retention. Consequently, SRO aluminosilicates take part in a number of environmental- and agricultural-related processes, including solute transport, 'heavy' metal behavior, soil OM accumulation, and agricultural productivity.

Formation and distribution

Short-range ordered aluminosilicates (allophane and imogolite) form as metastable phases in solutions with high H_4SiO_4 activity (X-axis of Fig. 3) and/or high Al^{3+} activity relative to pH (3 pH + log Al^{3+} Y-axis of Fig. 3). High H_4SiO_4 and Al^{3+} activities are frequently achieved in soils derived from volcanic ejecta, which dissolves at a high rate due to its small particle size, glassy nature,

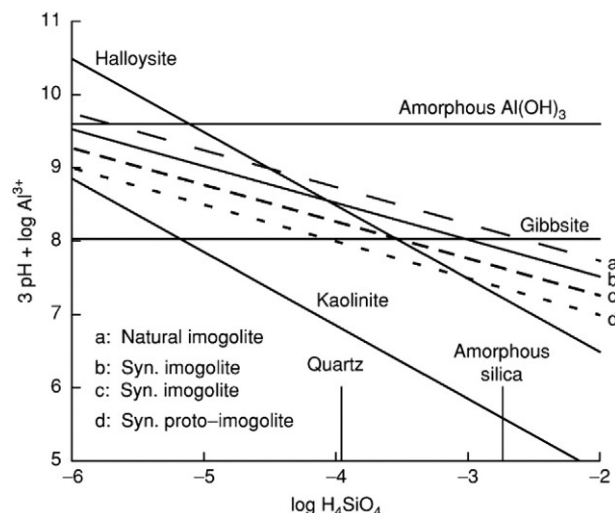


Fig. 3 Stability diagram for selected Al hydrous oxide and aluminosilicate solid phases. Syn, synthetic.

and high porosity (Ugolini and Dahlgren, 2002). Nevertheless, SRO aluminosilicates can be found in soils from other parent materials if composition of the solution allows for their formation. The favorable pH range for allophane and imogolite formation is mildly acidic to neutral conditions ($\text{pH} = 5\text{--}7$; Dahlgren et al., 2004). At more acidic pH values, these minerals tend to dissolve or not form. Under supersaturated conditions, nucleation of allophane and imogolite is kinetically favored over more crystalline minerals (e.g., kaolinite) because of the lower solid–solution interfacial tension of more soluble phases (i.e., SRO aluminosilicates) (Stumm, 1992). Based on nuclear magnetic resonance (NMR) analysis, the process of formation of allophane from volcanic glass under soil conditions is proposed as the reaction of a gibbsite-like sheet that is formed from Al initially released from volcanic glass and H_4SiO_4 that is subsequently released from a silica gel-like polymer in volcanic glass (Hiradate and Wada, 2005). Because of the high solubilities of allophane and imogolite, they dissolve, releasing Si and Al, which then re-precipitate as less soluble, but more thermodynamically stable crystalline minerals (e.g., kaolinite, halloysite, and gibbsite) with time; the Ostwald ripening process.

Short-range ordered aluminosilicates have a disproportionate influence on soil properties, as explained later in this chapter. Their content is a key criterion for Andosols (IUSS Working group WRB, 2022), corresponding to Andisols in USDA Soil Taxonomy (Soil Survey Staff, 2022) and Podzols (WRB, Spodosols in USDA Soil Taxonomy), as well as organo-Al/Fe complexes and ferrihydrite. The $\text{Al}_0 + 1/2\text{Fe}_0$ quantity, where Al_0 originates from dissolution of SRO aluminosilicates and organo-Al complexes and Fe_0 from ferrihydrite and organo-Fe complexes, should be 20 g kg^{-1} (or 4 g kg^{-1} for young volcanic soils) or more for andic soil properties and 5 g kg^{-1} or more for a spodic horizon.

Soils derived from volcanic ejecta frequently have large contents of SRO aluminosilicates, which allows them to be classified as Andosols. The materials comprising volcanic ejecta, such as volcanic glass, quickly dissolve leading to supersaturated soil solutions and the formation of allophane and imogolite. Because of the limited translocation of Al^{3+} at pH values >5 and the low dispersibility of allophane and imogolite, they are generally believed to form in situ and show minimal translocation within the soil profile of Andosols (Dahlgren et al., 2004).

Depending on H_4SiO_4 activity in the soil solution, Al-rich allophane (elemental ratio of Al:Si $\approx 2:1$), Si-rich allophane (Al:Si $\approx 1:1$) and halloysite are commonly formed from volcanic ejecta. Under low H_4SiO_4 activity ($< \sim 10^{-3.45}$; Churchman and Lowe, 2012), Al-rich allophane, which is the most common type, forms, whereas greater H_4SiO_4 activity ($> \sim 10^{-3.45} \text{ mol L}^{-1}$) favors the formation of Si-rich allophane and/or halloysite (Churchman and Lowe, 2012). The rate of tephra deposition, climate and soil depth, among others, influence H_4SiO_4 activity, which is high in drier sites with limited leaching and in deeper soil layers, and low in humid regions and well-drained sites (Churchman and Lowe, 2012).

Andosols can be divided into silandic and aluandic subgroups based on the abundance of SRO aluminosilicates. Silandic Andosols ($\text{Si}_0 \geq 6 \text{ g kg}^{-1}$) are dominated by SRO aluminosilicates, whereas aluandic Andosols ($\text{Si}_0 < 6 \text{ g kg}^{-1}$) have a predominance of organo-Al complexes. Aluandic Andosols are also called non-allophanic Andosols. Soil pH is considered a key factor leading to the formation of SRO aluminosilicates with the polymerization of Al^{3+} at mildly acid to neutral pH values. Allophane and imogolite formation is favored under low OM content and higher pH conditions (5–7) which may result from the rapid dissolution of colored volcanic glass (a base-rich glass) and/or lower precipitation and leaching. On the contrary, allophane and imogolite may dissolve under low pH conditions (< 5 ; Dahlgren et al., 2004) because of high precipitation and/or the low base element content of non-colored volcanic glass. The Al^{3+} released under acidic conditions favors complexation of Al^{3+} with OM leading to formation of organo-Al complexes. In general, silandic Andosols exhibit a neutral to mildly acidic nature, whereas aluandic Andosols have an extremely acid to acid nature.

The content of SRO aluminosilicates in volcanic soils can change over time with a significant influence of temperature, moisture and degree of chemical weathering (Lyu et al., 2018, 2022; Watanabe et al., 2023). Short-range ordered aluminosilicates are formed

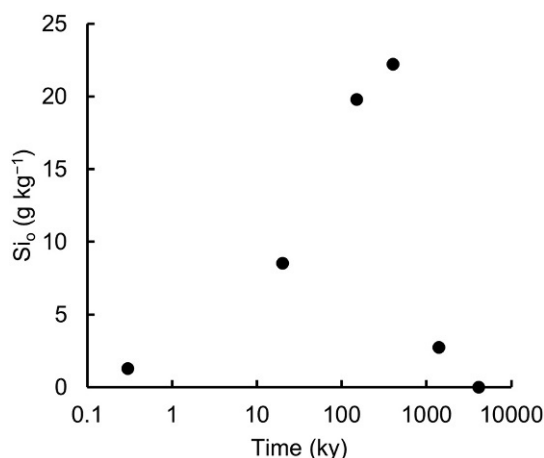


Fig. 4 Time course of short-range ordered aluminosilicate content represented by oxalate extractable Si (Si_o) in B horizon soils in Hawaii derived from basaltic volcanic materials under mean annual temperature and precipitation of 16 °C and 2500 mm, respectively. Data taken from Chorover J, Amistadi MK and Chadwick OA (2004) Surface charge evolution of mineral-organic complexes during pedogenesis in Hawaiian basalt, *Geochimica et Cosmochimica Acta*, 68(23): 4859–4876.

from volcanic ejecta and increase until the rapidly weathered primary minerals and volcanic glass are depleted resulting in lower Al^{3+} and H_4SiO_4 activities that are necessary to support their formation (Fig. 4). Later in soil development, the content of SRO aluminosilicates decreases as Ostwald ripening transforms the metastable phases to more thermodynamically stable, crystalline minerals (e.g., kaolinite and gibbsite). Sufficient moisture seems to be critical for the formation of SRO aluminosilicates because most Andisols occur in udic (63%) and ustic (30%) soil moisture regimes (Dahlgren et al., 2004). Soils with less excess precipitation (precipitation minus potential evapotranspiration) contain a limited content of SRO aluminosilicates (Fig. 5a). Dry conditions lead to slower rates of weathering, higher H_4SiO_4 activity due to less leaching, and promotion of crystallization, favoring halloysite formation over allophane and imogolite. Furthermore, high temperature seems to inhibit the formation of SRO aluminosilicates due to enhanced dissolution and precipitation reactions, which results in the preferential formation of more thermodynamically stable crystalline clays that consume metastable SRO aluminosilicates (Fig. 5b).

Other soil-forming factors can influence the formation and distribution of SRO aluminosilicates, primarily by affecting Al^{3+} and H_4SiO_4 activities in the soil solution. Organic matter supplied by vegetation complexes with Al^{3+} in the soil solution results in high organo-Al complexes and opaline silica contents, but little or no allophane in surface horizons (Harsh et al., 2002). Similarly, 2:1 type layer silicates hinder the formation of SRO aluminosilicates, preserving Al^{3+} as hydroxy-Al interlayer polymers. This influence of OM and 2:1 type layer silicates competing for Al^{3+} has been referred to as the ‘anti-allophanic effect’ (Dahlgren et al., 2004), which can lead to the formation of Aluandic Andosols. Other factors include topographic position which controls leaching intensity and residence time of water and, therefore, affects pH and H_4SiO_4 activity; and parent materials that control rates of chemical weathering (e.g., colored vs. non-colored volcanic glasses) and thereby the pH and Al^{3+} and H_4SiO_4 activities in the soil solution.

Short-range ordered aluminosilicates are commonly found in the illuvial horizons of Podzols, in particular Bs horizons rather than organic-rich Bh horizons. In contrast to Andosols, where SRO aluminosilicates precipitate in situ, Si and Al translocate from upper horizons and some may precipitate as SRO aluminosilicates in the illuvial horizons. Several mechanisms are proposed for the translocation of Si and Al and the accumulation of SRO aluminosilicates in subsoil horizons. They include the formation of organo-metal complexes and inorganic colloidal sols (e.g., proto-imogolite) in eluvial horizons that are subsequently translocated to illuvial horizons where they accumulate due to an increase in the metal to OM ratio caused by the addition of metal ions, microbial decomposition of organo-metal complexes, and/or a higher pH range that facilitates the formation of SRO aluminosilicates (Sauer et al., 2007). Alternatively, formation of SRO aluminosilicates in Podzol Bs horizons has been ascribed to in situ weathering as the organic acid acidity and Al^{3+} -complexing organic ligands in surface horizons (e.g., E horizon) are attenuated in the underlying Bh horizons, and carbonic acid becomes the primary proton donor at pH values greater than 5 in the underlying Bs horizons. The dominant mechanism for their translocation and/or accumulation in the subsoil is still open to debate, but may differ according to the soil-forming conditions (Sauer et al., 2007). Soil groups other than Andosols and Podzols can also contain SRO aluminosilicates if the soil-forming factors contribute to the composition of a solution suitable for their formation.

Identification

The poorly crystalline nature of allophane and imogolite defies their easy identification in soils by X-ray diffraction, which is the method of choice for the phyllosilicate clay minerals and many oxides. Rather, a combination of methods is used to identify and characterize these materials, including electron microscopy, selective chemical dissolution, Fourier transform infrared (FTIR) spectroscopy, and differential thermal analysis. Synthetic allophane and imogolite or purified samples from soils can be further

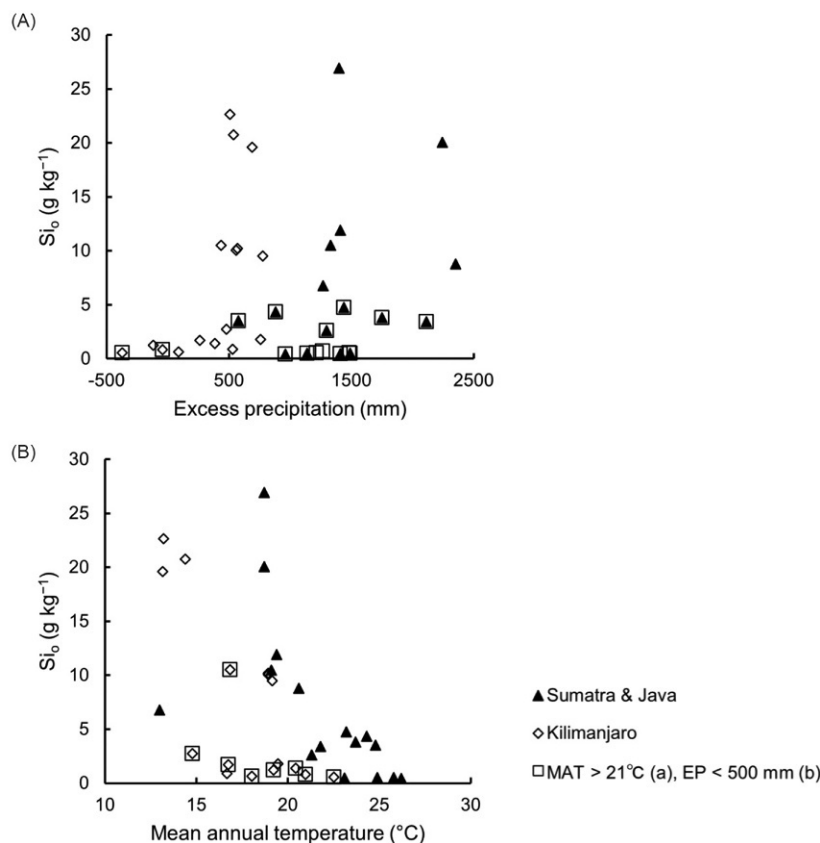


Fig. 5 Relationship between short-range ordered aluminosilicate content represented by oxalate extractable Si (Si_o) vs. excess precipitation (mean annual precipitation – potential evapotranspiration) (a) and vs. mean annual temperature (b) in B horizon soils. The soils were derived from volcanic ejecta in tropical regions having relatively constant temperatures throughout the year. Data taken from Lyu H, Watanabe T, Kilasara M and Funakawa S (2018) Effects of climate on distribution of soil secondary minerals in volcanic regions of Tanzania, *Catena*, 166: 209–219, Lyu H, Watanabe T, Ota Y, Hartono A, Anda M, Dahlgren RA and Funakawa S (2022) Climatic controls on soil clay mineral distributions in humid volcanic regions of Sumatra and Java, Indonesia, *Geoderma*, 425: 116058.

characterized by NMR and X-ray absorption spectroscopy (XAS). Electron microscopy is useful for defining the morphology of individual particles and aggregates of SRO minerals, whereas the other methods give information on the structural and chemical bonding properties.

Allophane and imogolite are not truly non-crystalline or amorphous minerals and do exhibit X-ray diffraction patterns, albeit with diffuse and often weak diffraction peaks. These patterns can be used to identify imogolite and distinguish it from Al-rich (proto-imogolite) allophane (Fig. 6), something not readily accomplished with FTIR. All allophanes, regardless of the Al/Si ratio, display two broad peaks centered around 0.33 and 0.25 nm. The imogolite peaks result from both the long-range order along the longitudinal direction of the imogolite fiber and the close packing structure of fiber bundles.

The clearest way to distinguish between allophane and imogolite is with electron microscopy. Transmission electron micrographs reveal the fibrous bundles of imogolite (Fig. 2) that are readily identifiable even in complex soil mixtures. The imogolite fibers have cylindrical morphology, consisting of a gibbsite-like sheet of aluminum hydroxide octahedra pulled into a cylinder by individual silica tetrahedra. A two-dimensional (2-D) ball-and-stick model in Fig. 7 represents a cross-section of the imogolite tube. Allophane appears amorphous under the electron microscope, appearing gel-like (Fig. 1) or as aggregates of small spherules varying from 3.5 to 5.5 nm. The structure, composition, and morphology of allophane depend largely on its Al/Si ratio.

The quantity of SRO minerals is most commonly estimated by selective dissolution of these materials from soil with some notable cautions. Extraction with acid ammonium oxalate in the dark mainly removes Al and Si from poorly crystalline aluminosilicates and iron (Fe) oxides (i.e., ferrihydrite, nano-crystalline goethite) and Al from complexes with OM (Shang and Zelazny, 2008). The latter Al phase (organo-Al complexes) can be removed by sodium pyrophosphate extraction, which does not dissolve a significant amount of SRO aluminosilicates (Shang and Zelazny, 2008). Thus, the Al/Si ratio of allophane can be estimated from the extracted Al by acid oxalate (Al_o) and pyrophosphate (Al_p) using the formula $(Al_o - Al_p)/Si_o$. Together with sodium pyrophosphate solution to remove organo-Al and -Fe complexes, reductive dissolution with sodium dithionite (e.g., dithionite-citrate bicarbonate (DCB) solution) is applied to remove iron oxides for purifying SRO aluminosilicates; DCB can dissolve a considerable amount of SRO aluminosilicates (Rennert, 2019). It should be noted that the interpretation of extracted Al and Si fractions by these selective extractions has inherent uncertainty, such as Al_o includes Al in the lower-crystallinity iron oxides, the hydroxy-Al interlayer of 2:1

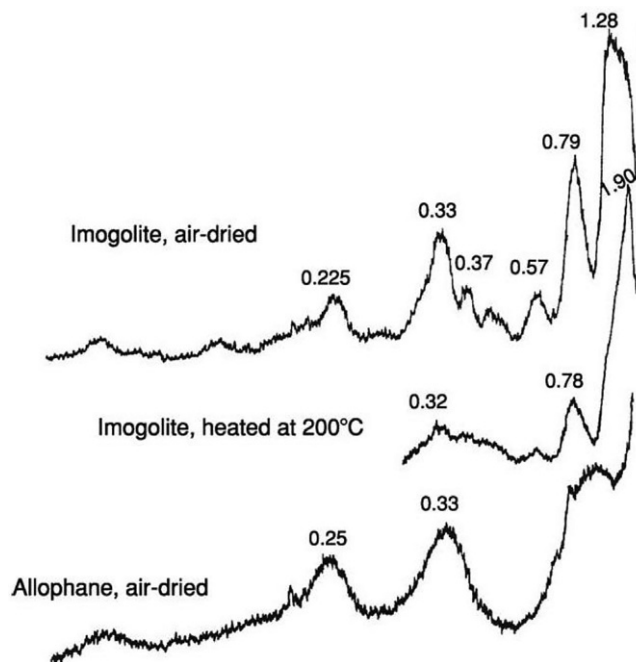


Fig. 6 X-ray diffraction patterns of naturally occurring imogolite and allophane. Reproduced from Wilson MJ (1987) X-ray powder diffraction methods in Wilson MJ, *A Handbook of Determinative Methods in Clay Mineralogy*, Glasgow, UK: Blackie & Son.

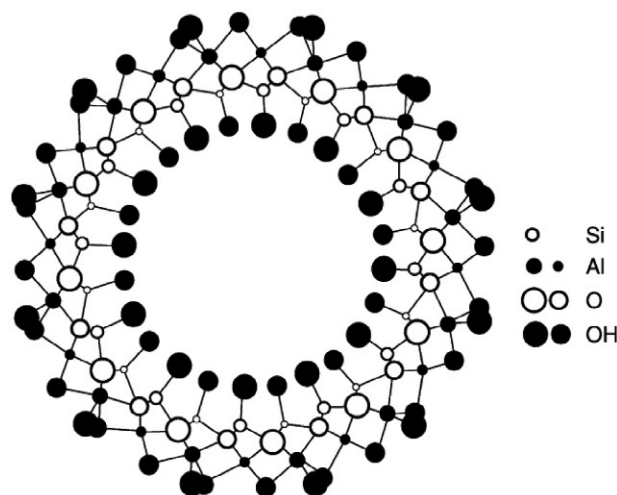


Fig. 7 Crystallographic representation of the cross-section of an imogolite tube. Reproduced from Cradwick PD, Wada K, Russell JD, Yoshinaga N, Masson CR and Farmer VC (1972) Imogolite, a hydrated aluminum silicate of tubular structure, *Nature Physical Science*, 240(104): 187–189.

type layer silicates, and Al_p may include small particles of Al hydroxides (colloids) that are not bound to OM (Kaiser and Zech, 1996; Shang and Zelazny, 2008; Rennert, 2019). In addition, Si_o values less than 1 to 2 g kg⁻¹ should not be considered significant because acid oxalate may attack the surfaces of crystalline minerals (Dahlgren, 1994). Once the amount of Al and Si coming solely from SRO minerals has been estimated (e.g., $Al_o - Al_p$ and Si_o), the percentage of allophane and imogolite in the soil can be estimated from:

$$\% \text{ allophane in soil} = Si / (-0.067 Al / Si + 0.27), \quad (1)$$

where Si is the percentage Si from SRO aluminosilicates. This formula is based on an average Al/Si ratio for allophanes found in soils.

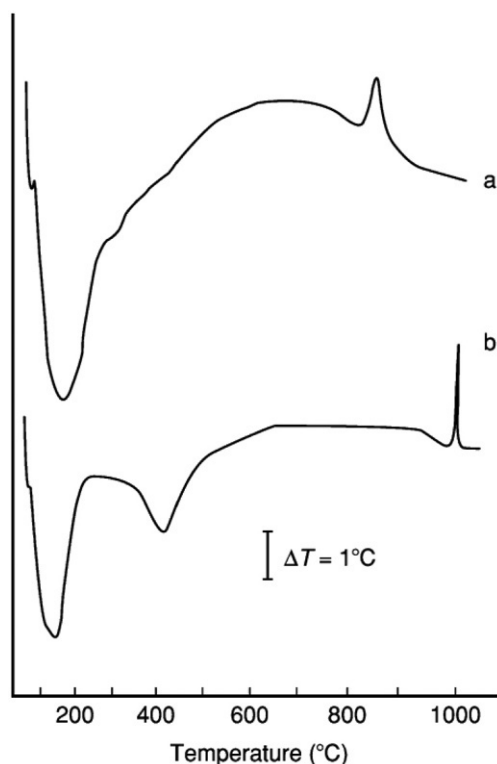


Fig. 8 Differential thermal analysis curves of (a) Al-rich allophane and (b) imogolite.

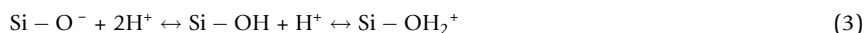
Imogolite can be distinguished from allophane and its concentration in soil estimated by both differential thermal analysis (DTA) and FTIR. Imogolite gives rise to a unique endothermic peak at 400 °C due to structural water loss (Fig. 8). The area of this peak is proportional to the quantity of imogolite in the soil. An FTIR band at 348 cm⁻¹ indicates the structure of both imogolite and Al-rich (proto-imogolite) allophane. This peak can be normalized to provide the quantity of 'imogolite-like structures' in a given soil.

Structure and charge

Infrared and NMR spectra have been instrumental in helping to infer the structure of both imogolite and allophane. The IR band at 940 cm⁻¹ is assigned to the unshared OH (hydroxyl) groups of the orthosilicate groups in imogolite and Al-rich allophane. For all other allophanes, the Si–O (oxygen) stretching band shifts from 975 to 1020 cm⁻¹ as the Al/Si ratio decreases (Fig. 9). This indicates that polymerization of the silicate groups increases with increasing Si concentration. The -79 ppm NMR shift also indicates Si coordinated to three Al–O groups and one OH (Fig. 10). The negative shift in the NMR spectrum as the Al/Si ratio decreases supports increased oxygen-sharing among Si atoms.

Polymerization is accompanied by an increase in tetrahedrally coordinated Al, as indicated by ²⁷Al NMR and XAS spectra, inferring that the Si-rich allophanes have a feldspathoid-like structure. The feldspathoids are network silicates that, like zeolites, are characterized by high structural negative charge density. Such a charge also increases in allophanes as the Al/Si ratio decreases.

In Al-rich allophane and imogolite, the surface charge arises almost exclusively from variably charged aluminol (Al–OH) and silanol (Si–OH) groups. Chemically, the surface charging reactions can be represented as protonation–deprotonation reactions on these groups:



Aluminol groups in gibbsite have no net charge at around pH 9, whereas silanol groups in quartz have a net neutral charge near pH 2. Consequently, it is assumed that the acidic silanol groups contribute a negative charge to the SRO aluminosilicates, and aluminol groups primarily bear a positive charge because, typically, the pH range of soils falls between 3 and 9.

Fig. 11 shows the variable-charge behavior of an Al-rich allophane. The data points represent the amount of Na (sodium) and Cl (chlorine) adsorption determined as a function of pH and ionic strength. The magnitude of adsorption increases with increasing

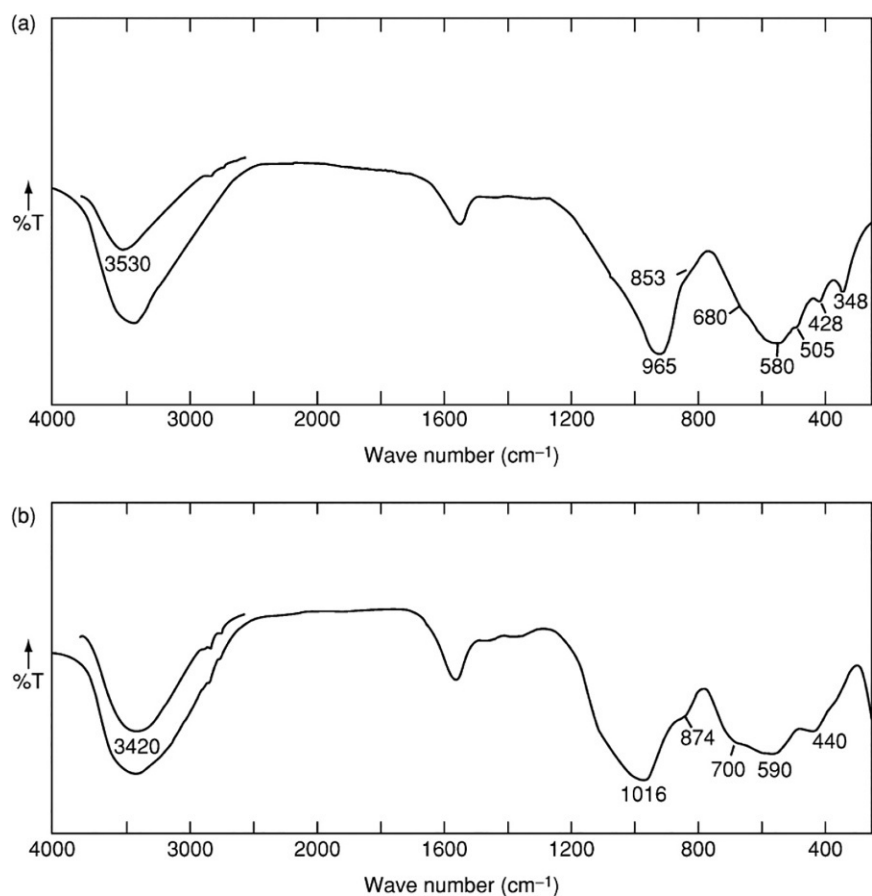


Fig. 9 Infrared spectra of (a) Al-rich (protoimogolite) and (b) Si-rich allophanes. The y-axis is %T, which represents percentage of transmittance.

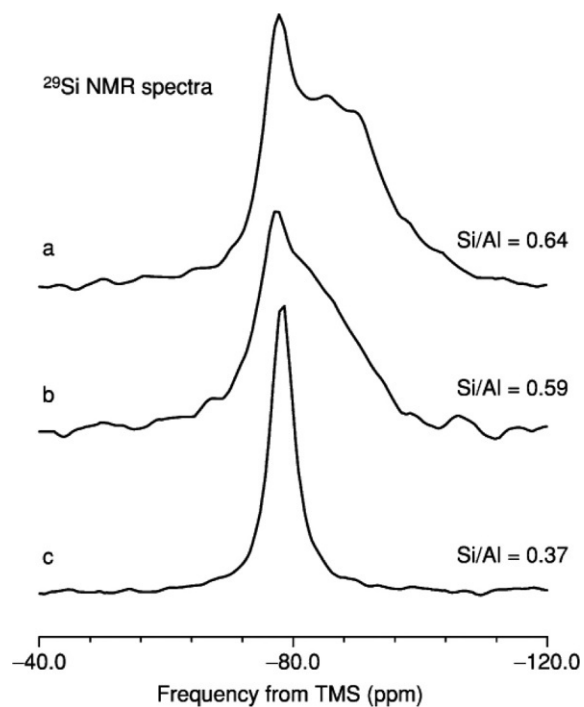


Fig. 10 The ²⁹Si NMR spectra of three allophanes with decreasing Si/Al molar ratio: A, 0.64; B, 0.59; C, 0.37.

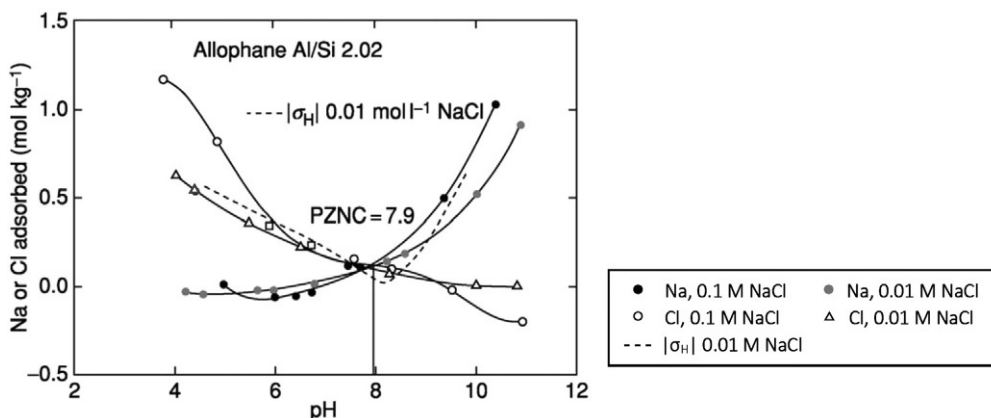


Fig. 11 Ion adsorption and proton charge (σ_H) of synthetic Al-rich allophane in NaCl. Reproduced with permission from Su CM, Harsh JB and Bertsch PM (1992) Sodium and chloride sorption by imogolite and allophanes, *Clays and Clay Minerals*, 40(3): 280–286.

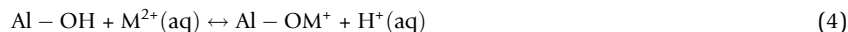
ionic strength, whereas the sign of the charge depends on pH. The pH at which Na and Cl adsorption is equal (i.e., no net surface charge) is known as the point of zero net charge (PZNC). The dotted line in Fig. 11 represents the surface charge determined by adsorption of H^+ and OH^- (σ_H). The pH at which σ_H is zero is known as the point of zero proton charge (PZNPC). The PZNC and PZNPC are equivalent when there is no permanent structural charge in the allophane. The fact that these two are almost equivalent in Fig. 11 also shows that Al is not substituting for Si at tetrahedral sites, and therefore not giving rise to structural charge.

It is evident from Fig. 11 that Al-rich allophane is positively charged in all but calcareous and other alkaline soils. The amount of charge is significant, rivaling smectites in magnitude of charge in both acid and highly alkaline soils. Allophane and imogolite may be among the most important anion exchangers in soils. The use of soils containing allophane as filters for anions such as arsenate, iodide, and technetium has been considered. Silica-rich allophanes, on the other hand, have a significant negative structural charge and will function as important cation exchangers.

Chemisorption and ligand exchange

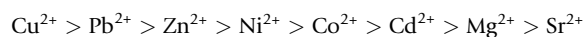
In addition to the nonspecific cation and anion exchange reactions that occur on positively and negatively charged sites, respectively, metals and ligands can chemisorb to the reactive aluminol and silanol groups. The nonspecific interactions on charged surfaces are characterized by the fact that the ions can be easily exchanged with other ions of like charge from the soil solution. The adsorbed ions remain hydrated (outer-sphere complexes), are kinetically labile, and equivalently balance the surface charge. Chemisorbed ions, on the other hand, bond directly to the aluminol and silanol groups (inner-sphere complexes), are not easily or rapidly exchanged, and do not require an oppositely charged surface group to adsorb. These interactions can govern the transport, stability, and bioavailability of solutes including toxic metals, essential plant nutrients, and soil OM.

Chemisorption of 'heavy' metals such as Zn (zinc), Cu (copper), Pb (lead), Co (cobalt), and Cd (cadmium) generally involves exchange with a proton from a surface silanol or aluminol group. An inner-sphere complex is formed as the metal bonds directly with the oxygen of the surface group. For example, on an aluminol group on allophane:



represents the chemisorption of a divalent metal cation (M^{2+}) where one proton is exchanged. The result, in this example, is a more positively charged surface, a reduction in the mobility of the metal cation, and a decrease in soil pH. Given the high specific surface areas of SRO aluminosilicates and the relatively large number of reactive surface groups, allophane and imogolite can play an important role in metal behavior in soils, even when these minerals are present in low concentration.

The cation exchange behavior of metals depends primarily on their charge and hydrated radius. The selectivity for a negatively charged site increases with charge and decreases with size of the hydrated ion. In the case of the chemisorbed metals, however, the selectivity depends on the nature of the reaction with the aluminol or silanol group. The selectivity series for metals on aluminol groups follows the order:



This series represents the position of the adsorption edge when the amount of metal adsorbed is plotted against pH. It implies that the metals such as Cu and Pb will be chemisorbed to aluminol groups even in acid soils, whereas the alkaline metals like Mg (magnesium) and Sr (strontium) are not chemisorbed in slightly alkaline to acid soils, but only undergo cation exchange. Chemisorption to silanol groups is similar.

Just as allophane and imogolite can be important anion exchangers in soil, they can also chemisorb anions and neutral solutes through ligand exchange:



In the above reaction, an anion (A^-) exchanges with a water molecule, increasing the negative charge on the surface. The anion bonds directly with the surface Al or Si ion, forming an inner-sphere complex. Several solutes are known to chemisorb to allophane and imogolite through ligand exchange, including phosphate, fluoride, selenite, and organic acids.

The sorption of phosphate to SRO minerals has a major effect on the availability of phosphate to crops grown in allophanic soils. Although it was long assumed that this reaction was governed strictly by ligand exchange, it is more likely that phosphate reacts with allophane to form insoluble Al-phosphate phases similar to variscite ($\text{AlPO}_4 \cdot 2\text{H}_2\text{O}$) (Nanzyo, 1987). The strong retention of phosphate may limit fertility of both agricultural and forested soils. Solutions to this potential problem have included not only large applications of inorganic fertilizers, but also the addition of silica and OM to compete with phosphate for reactive sites.

Fluoride (F^-) is strongly adsorbed to both aluminol and silanol groups and results in the release of OH^- by the reactions:



These reactions have been used as a test for the presence of SRO aluminosilicates in soil. The large concentration of reactive groups can easily result in a $\text{pH} > 10$ when a few grams of an allophanic soil are reacted with 0.1 mol L^{-1} NaF, initially at neutral pH. Other SRO minerals, such as ferrihydrite, nano-crystalline goethite and organo-Al complexes, also contribute to this reaction. As in the reaction with phosphate, insoluble metal fluorides can replace allophane.

Reaction with soil organic matter

Soils high in SRO aluminosilicates and other SRO minerals (e.g., ferrihydrite) tend to store large amounts of OM because (i) the environmental conditions favoring their formation ensure relatively large OM input from plants, and (ii) their physical and chemical properties cause a strong affinity for OM. High OM storage capacity and the old C age of Andosols is well-known (Torn et al., 1997). The term 'Andosol' originated in Japan and refers to 'dark soils.' In Andosols and other soils with volcanic parent materials, bulk soil C contents often show a significant positive correlation with pyrophosphate-extractable Al and Fe, and to a lesser extent, with oxalate-extractable Al and Fe (e.g., Percival et al., 2000). Similar positive correlations have been shown for spodic horizons. More recent studies have shown that the Al and Fe which are extractable by these two extractions significantly controlled soil OM content in a wide range of soil types beyond Andosols and Podzols (Ashida et al., 2021; Rasmussen et al., 2018). While the abundance of clay-sized minerals (clay %) has been considered to be the main control on soil OM storage and turnover time, a growing number of studies are showing greater control by SRO mineral abundance, specifically oxalate-extractable Al and Fe, than clay percent in non-arid soils around the world (Rasmussen et al., 2018). Similarly, subsoil OM storage under wetter climates is strongly controlled by SRO mineral abundance due mainly to its sorptive capacity for dissolved organic carbon (Kramer and Chadwick, 2018). Soil C persistence inferred from radiocarbon analysis can be controlled by oxalate- and especially pyrophosphate-extractable Al and Fe contents in soils developed under relatively wet climatic regimes (Lawrence et al., 2021), although the quantitative contribution of specific metal phases such as SRO aluminosilicates remains to be elucidated.

The SRO minerals protect OM from microbial degradation by three general mechanisms as follows. (i) SRO minerals can adsorb large amounts of dissolved organic matter through ligand exchange because of the high specific surface area of allophane and imogolite ($700\text{--}900$ and $900\text{--}1100 \text{ m}^2 \text{ g}^{-1}$, respectively; Harsh et al., 2002) and high surface reactivity (i.e., high density of hydroxyl group per unit surface area). Allophane and imogolite sorb up to $61\text{--}167 \text{ mg C g}^{-1}$ in batch experiments (Inoue and Wada, 1968), which is 1–2 orders of magnitude greater sorptive capacity than well-crystallized silicate clays. (ii) Al and Fe released from SRO minerals (and from other minerals) can be complexed with organic ligands to form mono- or poly-nuclear complexes that are relatively stable against microbial degradation (Boudot et al., 1989; Takahashi and Dahlgren, 2016). Complexation of OM with Al in or released from SRO minerals not only slows OM decomposition but may inhibit the crystal growth of allophane and imogolite (Levard et al., 2012). The formation of organo-Al complexes with little or no SRO aluminosilicates is typically found in the surface horizons with a large OM content and low pH. In this case, it is the Al complexes themselves that impart high reactivity to the soils, such as high phosphate sorption and Al^{3+} solubility. (iii) SRO minerals act as binding agents to form stable aggregates that physically protect OM therein reducing the access of microbes, extracellular enzymes, water and oxygen. Allophanic Andosols have particularly strong physical stability and can form irreversible microaggregates upon dehydration. Across a range of soils, both SRO minerals and OM were co-localized in the meso-density ($1.8\text{--}2.2 \text{ g cm}^{-3}$) fraction, mainly as shaking-resistant, organo-mineral microaggregates where the amounts of residing OM often exceed the level accountable by sorptive stabilization (Wagai et al., 2020). Dissolved Fe, Al, and Si released during mineral weathering can readily interact with dissolved OM to form nano-sized coprecipitates (Tamrat et al., 2019) that contain both sorptive associations and precipitates of organo-metal complexes. In soils, the three mechanisms take place concurrently and their relative importance changes depending on pedogenic conditions (Mikutta et al., 2009).

Aluminum solubility

The activity of Al in soils is an important consideration from an environmental standpoint because Al is toxic to various plant species, especially agricultural crops, and to aquatic organisms. Aluminum activity increases 1000-fold for every unit decrease in pH when Al solubility is controlled by solid phases such as gibbsite, kaolinite, and imogolite. Among the soils for which a solid-phase dissolution mechanism is likely to control Al activity in the soil solution are those containing SRO aluminosilicates. Solution compositions of the B horizons of Podzols, in particular, near equilibrium with an imogolite-like phase are determined in both field and laboratory studies. As mentioned above, the SRO aluminosilicates are metastable with respect to more crystalline minerals such as kaolinite; however, they are persistent in soils and approach equilibrium solubility fairly rapidly.

The dissolution reaction for imogolite can be written as follows:



and the log solubility product ($\log K_{\text{sp}}$) is:

$$\log K_{\text{sp}} = \log(\text{Al}^{3+}(\text{aq})) + 0.5 \log(\text{H}_4\text{SiO}_4(\text{aq})) + 3\text{pH}, \quad (9)$$

where parentheses denote activities.

Aluminum solubility can then be obtained by rearranging:

$$\log(\text{Al}^{3+}) = \log K_{\text{sp}} - 0.5 \log(\text{H}_4\text{SiO}_4(\text{aq})) - 3\text{pH} \quad (10)$$

Thus, the activity of Al^{3+} in a soil solution controlled by imogolite depends negatively on $\text{H}_4\text{SiO}_4(\text{aq})$ activity and pH.

A stability diagram based on some of the published values for imogolite and Al-rich (proto-imogolite) allophane in relation to more crystalline minerals is shown in Fig. 3. The two imogolite lines are from two different investigations. The more stable of the two synthetic imogolites crosses the halloysite stability line at an aqueous H_4SiO_4 activity near $10^{-3.5}$, consistent with the formation of halloysite in soils with higher values. However, the formation of halloysite in preference to imogolite may be kinetic, requiring significant supersaturation. Although the solubilities of these minerals are not well defined, Fig. 3 still shows the metastability of SRO minerals with respect to kaolinite at nearly all H_4SiO_4 activities: the relative stability of gibbsite being higher at low ($\text{H}_4\text{SiO}_4(\text{aq})$) and of halloysite being higher ($\text{H}_4\text{SiO}_4(\text{aq})$).

The kinetics of Al^{3+} and Si release under acidic conditions from allophanic soils containing a variety of SRO aluminosilicates and organo-Al complexes along with pure imogolite gel films from Kitami Pumice beds showed very rapid rates of Al^{3+} and Si release (Fig. 12). The rapid Al^{3+} release kinetics from allophanic soils was attributed in part to a simultaneous Al^{3+} equilibrium that occurs among exchangeable Al^{3+} , organo-Al complexes, SRO minerals and gibbsitic phases that commonly occur in allophanic soils. The rapid release of Al^{3+} from exchange and organo-Al complexes maintains a short-term equilibrium, whereas slower Al^{3+} release from allophane, imogolite and gibbsitic phases maintains the longer-term equilibrium. In soils with the co-existence of imogolite and hydroxy-Al polymers (a gibbsite-like phase) within the interlayer of 2:1 layer silicates, the hydroxy-Al polymer appears to set the ultimate Al^{3+} solubility, whereas the simultaneous Al^{3+} equilibrium with imogolite sets the H_4SiO_4 activity at approximately 10^{-4} (Fig. 3). In soils with considerable 2:1 layer silicates and soil OM, preferential incorporation of Al^{3+} into hydroxy-Al polymers and organo-Al complexes may inhibit the formation of allophane and imogolite, which is a theoretical explanation of the anti-allophanic effect mentioned earlier.

Physical properties

Soils high in SRO minerals generally have very low bulk densities, typically less than 0.85 g cm^{-3} . This is a defining characteristic of Andosols or soils classified as having 'andic' properties (Bäumler et al., 2005). However, individual allophane and imogolite particles have densities of $2.7\text{--}2.8 \text{ g cm}^{-3}$ and $2.6\text{--}2.75 \text{ g cm}^{-3}$, respectively. The low bulk densities result from high microporosity

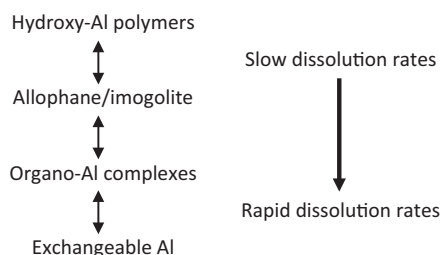


Fig. 12 The active Al fractions in Andosols appear to form a simultaneous equilibrium between the various solid-phase constituents. Reproduced from Dahlgren RA, Saigusa M and Ugolini FC (2004) The nature, properties and management of volcanic soils in Sparks DL., *Advances in Agronomy*, 113–182.

brought about by particle–particle and aggregate–aggregate associations (Bäumler et al., 2005), and from the strong association between allophanes and soil OM, discussed above. Pure allophane particles form associations with one another that flocculate into microporous aggregates (Fig. 1) when near the point of zero charge or the point where the particles have zero mobility in an electric field. Even in the presence of OM and other minerals, allophanic soils are abundant in highly microporous aggregates that are characterized by strong physical stability (resistant to 5–10 fold greater sonication energy compared to non-andic soils) and having a hierarchical structure (Asano and Wagai, 2014).

The surfaces of allophane and imogolite are assumed to be largely accessible to the soil solution. Specific surface areas determined by a polar molecule such as ethylene glycol monoethyl ether are around $10^3 \text{ m}^2 \text{ g}^{-1}$ (Harsh et al., 2002), which is greater than crystalline clay silicates found in soils including smectites. Specific surface areas of allophanes determined with N_2 (nitrogen) are significantly lower, ranging from 0.3 to $0.6 \times 10^3 \text{ m}^2 \text{ g}^{-1}$ (Vandickelen et al., 1980).

The physical characteristics of SRO minerals and associated OM have a profound effect on soils containing these materials. In addition to the low bulk densities, such soils have unusually large water retention capabilities at both 33 kPa ('field capacity') and 1500 kPa ('permanent wilting point') suction. Saturated water retention has been reported as high as 1.8 kg kg^{-1} . This can be attributed to the high microporosity and leads to favorable conditions for plant growth, particularly in forested soils. Meso- and macro-porosity are also high in andic soils as a result of the granular structure brought about by the aggregation of particles and OM. Consequently, saturated hydraulic conductivity is generally high in allophanic soils.

Both physical and chemical alteration of these soils can change hydraulic conductivity and other physical properties. When soils containing allophane and imogolite are air-dried, there is an irreversible increase in void volume, which results in higher saturated hydraulic conductivity and lower water retention. It is evidently difficult to rehydrate all surfaces and micropores after drying. In the field, mechanical disturbance that enhances drying or disrupts through shearing can lead to significant lowering of soil plasticity, whereas moist andic soils are considered highly plastic. Under the wetting and drying conditions common in cropping systems, these soils tend to resist compaction by machinery and mechanical disruption by rainfall. Finally, adjusting soil pH to values higher or lower than the near-neutral point of zero charge of Al-rich allophanes disturbs particle–particle associations and can reduce hydraulic conductivity by as much as 75% or more.

Implications of short-range ordered aluminosilicates for agricultural productivity

In terms of overall agricultural productivity, silandic (allophanic) Andosols generally have a greater agricultural potential than aluandic (non-allophanic) Andosols. The SRO minerals strongly contribute to several physical, chemical and biological properties of soils that greatly affect agricultural productivity. The SRO aluminosilicate minerals impart excellent physical properties for agricultural production, such as large water retention, high porosity, high permeability, good friability, low bulk density, excellent soil tilth and resistance to soil compaction and erosion. These properties are conducive to plant growth because of the large available water content which facilitates seedling emergence and root penetration. These favorable properties largely result from the formation of stable aggregates facilitated by SRO minerals in association with soil OM.

The chemical properties of Andosols change considerably as chemical weathering progresses and leads to an increasing abundance of SRO minerals. Young volcanic deposits have an abundance of rock-derived nutrients, including P, because of the rapid weathering of glassy and porous volcanic deposits. Nitrogen is generally the most limiting nutrient in young volcanic deposits as N fixation and OM accumulation proceed over the first several hundred years of pedogenesis. As previously discussed, SRO minerals contribute to the accumulation of soil OM and its associated nutrients by chemical and physical stabilization mechanisms. The amount of N released by mineralization is strongly correlated with the amounts of soil OM, especially the fraction defined as the easily decomposable OM pool.

The different charge characteristics between allophanic and non-allophanic Andosols because of their contrasting colloidal mineralogy is the most important factor regulating chemical fertility attributes (Table 1). As weathering progresses and the SRO mineral content increases, strong P retention by SRO minerals may result in P-limitation to several agricultural crops. However, the amount of P adsorbed per mole of active Al was higher in non-allophanic Andosols than allophanic Andosols (0.13 and $0.09 \text{ mol mol}^{-1}$, respectively; Takahashi and Dahlgren, 2016). Silicon, as an important nutrient for crops such as rice (*Oryza sativa* L.) and sugar cane (*Saccharum officinarum* L.), is strongly affected by the colloidal mineralogy. For example, the presence of halloysite generally supports an H_4SiO_4 activity of approximately $10^{-3.5}$ compared to allophane and imogolite which support an H_4SiO_4 activity of about 10^{-4} (Fig. 3). Notably, sheep grazing on silica-rich volcanic soils in New Zealand wears down their teeth much more rapidly than those grazing on non-volcanic soils due to the larger silica content of forage on volcanic soils that abrades the teeth.

Compared to non-allophanic Andosols, allophanic Andosols have only trace amounts of KCl-extractable Al^{3+} and tend to buffer the soil pH(H_2O) values near to 5, resulting in relatively small soluble Al^{3+} concentrations. Consequently, Al^{3+} toxicity is predominantly observed in non-allophanic Andosols containing large amounts of KCl-extractable Al^{3+} derived primarily from 2:1 layer silicates, but not in allophanic Andosols whose major clay minerals are SRO minerals. Larger extractable and soluble Al^{3+} concentrations are shown to inhibit several soil-borne plant pathogens (e.g., common potato scab, rice seedling damping-off disease). Consequently, crops growing on allophanic Andosols are more susceptible to these soil-borne pathogens than non-allophanic Andosols having larger KCl-extractable Al^{3+} concentrations.

Table 1 Comparison of selected soil properties between allophanic Andosols and non-allophanic Andosols.

Property	Allophanic andosols	Non-allophanic andosols
Major clays	Allophane and imogolite	Hydroxy-Al interlayered 2:1 layer silicates
Source of active Al	Allophane and Al-humus complexes	Al-humus complexes
Constant negative charge	None	High
Variable positive charge	Some	None
Soil acidity	Weakly acid	Strongly acid
KCl-extractable Al	Low	High
Al-toxicity	Rare	Common
Consistency—LL, PL, PI ^a	High	Low
Stickiness	None—slightly	Moderately
Plasticity	None—slightly	Moderately
Compaction potential	Slight	Moderate

^aLL is liquid limit, PL is plastic limit; PI is plasticity index.

Reproduced from Dahlgren RA, Saigusa M and Ugolini FC (2004) The nature, properties and management of volcanic soils' in Sparks, D. L., ed. *Advances in Agronomy*, 113-182.

Outlook

The SRO aluminosilicates have significant influences on environmental and agricultural processes because of their large surface areas and reactive functional groups. Their abundance often covaries with other mineral phases (e.g., organo-metal complexes, amorphous gibbsite, pedogenic Fe oxides) in soil environments, leaving some uncertainty as to their specific independent roles in regulating soil physiochemical properties. Advanced spectroscopic techniques such as scanning transmission electron microscopy with energy dispersive X-ray spectroscopy or electron energy loss spectroscopy may allow their identification and quantification from intact natural soils. Furthermore, as their influence is evident even at small concentrations, such as on OM preservation and phosphate sorption, future investigations on the marginal SRO containing soils, for example, the distributions of SRO minerals in soils other than Andosols and Podzols, their persistence in highly weathered volcanic soils under different temperature and moisture conditions, and their forms in the soils (i.e., imogolite, allophane and proto-imogolite as well as organo-Al complexes) would enhance our understanding of the soil biogeochemical cycles (e.g., C, N, P, 'heavy' metals).

References

- Asano M and Wagai R (2014) Evidence of aggregate hierarchy at micro- to submicron scales in an allophanic Andisol. *Geoderma* 216: 62–74.
- Ashida K, Watanabe T, Urayama S, Hartono A, Kilasara M, Ze ADM, Nakao A, Sugihara S, and Funakawa S (2021) Quantitative relationship between organic carbon and geochemical properties in tropical surface and subsurface soils. *Biogeochemistry* 155(1): 77–95.
- Bäumler R, Caspari T, Totsche KU, Dorji T, Norbu C, and Baillie IC (2005) Andic properties in soils developed from non-volcanic materials in Central Bhutan. *Journal of Plant Nutrition and Soil Science* 168: 703–713.
- Boudot JP, Brahim ABH, Steiman R, and Seigle-Murandi F (1989) Biodegradation and synthetic organo-metallic complexes of iron and aluminum with selected metal to carbon ratios. *Soil Biology and Biochemistry* 21: 961–966.
- Churchman GJ and Lowe DJ (2012) Alteration, formation, and occurrence of minerals in soils. In: Huang PM, Li Y, and Sumner ME (eds.) *Handbook of Soil Science Properties and Processes*, 2 ed, p. 20-1-20-72. Boca Raton, FL: CRC Press.
- Dahlgren R (1994) Quantification of allophane and imogolite. In: Amonette JE and Stucki JW (eds.) *Quantitative Methods in Soil Mineralogy*, pp. 430–451. Madison, WI: Soil Science of America.
- Dahlgren RA, Saigusa M, and Ugolini FC (2004) The nature, properties and management of volcanic soils. In: Sparks DL (ed.) *Advances in Agronomy*, 113–182.
- Harsh J, Chorover J, and Nizeyimana E (2002) Allophane and imogolite. In: Dixon JB and Schulze DG (eds.) *Soil Mineralogy with Environmental Applications*, pp. 291–322. Madison, WI: Soil Science Society of America.
- Hiradate S and Wada S (2005) Weathering process of volcanic glass to allophane determined by ²⁷Al and ²⁹Si solid-state NMR. *Clays and Clay Minerals* 53: 401–408.
- Inoue T and Wada K (1968) Adsorption of humified clover extracts by various clays. In: *Transactions of 9th International Congress of Soil Science*, pp. 289–298. South Australia: Adelaide.
- IUSS Working Group WRB (2022) *World Reference Base for Soil Resources. International soil classification system for naming soils and creating legends for soil maps*, 4th edition Vienna, Austria: International Union of Soil Sciences (IUSS).
- Kaiser K and Zech W (1996) Defects in estimation of aluminum in humus complexes of podzolic soils by pyrophosphate extraction. *Soil Science* 161: 452–458.
- Kramer MG and Chadwick OA (2018) Climate-driven thresholds in reactive mineral retention of soil carbon at the global scale. *Nature Climate Change* 8(12): 1104–1109.
- Lawrence CR, Schulz MS, Masiello CA, Chadwick OA, and Harden JW (2021) The trajectory of soil development and its relationship to soil carbon dynamics. *Geoderma* 403: 115378.
- Levard C, Doelsch E, Basile-Doelsch I, Abidin Z, Miche H, Masion A, Rose J, Borschneck D, and Bottero JY (2012) Structure and distribution of allophanes, imogolite and proto-imogolite in volcanic soils. *Geoderma* 183: 100–108.
- Lyu H, Watanabe T, Kilasara M, and Funakawa S (2018) Effects of climate on distribution of soil secondary minerals in volcanic regions of Tanzania. *Catena* 166: 209–219.
- Lyu H, Watanabe T, Ota Y, Hartono A, Anda M, Dahlgren RA, and Funakawa S (2022) Climatic controls on soil clay mineral distributions in humid volcanic regions of Sumatra and Java, Indonesia. *Geoderma* 425: 116058.
- Mikutta R, Schaumann GE, Gildemeister D, Bonneville S, Kramer MG, Chorover J, Chadwick OA, and Guggenberger G (2009) Biogeochemistry of mineral-organic associations across a long-term mineralogical soil gradient (0.3–4100 kyr), Hawaiian Islands. *Geochimica et Cosmochimica Acta* 73(7): 2034–2060.

- Nanzyo M (1987) Formation of noncrystalline aluminum phosphate through phosphate sorption on allophanic and soils. *Communications in Soil Science and Plant Analysis* 18(7): 735–742.
- Parfitt RL (2009) Allophane and imogolite: role in soil biogeochemical processes. *Clay Minerals* 44(1): 135–155.
- Percival HJ, Parfitt RL, and Scott NA (2000) Factors controlling soil carbon levels in New Zealand grasslands: Is clay content important? *Soil Science Society of America Journal* 64(5): 1623–1630.
- Rasmussen C, Heckman K, Wieder WR, Keiluweit M, Lawrence CR, Berhe AA, Blankinship JC, Crow SE, Druhan JL, Pries CEH, Marin-Spiotta E, Plante AF, Schadel C, Schimel JP, Sierra CA, Thompson A, and Wagai R (2018) Beyond clay: towards an improved set of variables for predicting soil organic matter content. *Biogeochemistry* 137(3): 297–306.
- Rennert T (2019) Wet-chemical extractions to characterise pedogenic Al and Fe species—A critical review. *Soil Research* 57(1): 1–16.
- Sauer D, Sponagel H, Sommer M, Giani L, Jahn R, and Stahr K (2007) Review Article Podzol: Soil of the Year 2007 A review on its genesis, occurrence, and functions. *Journal of Plant Nutrition and Soil Science* 170(5): 581–597.
- Shang C and Zelazny L (2008) Selective dissolution techniques for mineral analysis of soils and sediments. In: Ulery AL and Drees LR (eds.) *Methods of Soil Analysis Part 5 Mineralogical Methods*, pp. 33–80. Madison, Wisconsin, USA: Soil Science Society of America, Inc.
- Soil Survey Staff (2022) *Keys to Soil Taxonomy*, 13th ed USDA Natural Resources Conservation Service.
- Stumm W (1992) *Chemistry of the Solid–Water Interface: Processes at the Mineral–Water and Particle–Water Interface in Natural Systems*. NJ, USA: John Wiley & Sons.
- Takahashi T and Dahlgren RA (2016) Nature, properties and function of aluminum-humus complexes in volcanic soils. *Geoderma* 263: 110–121.
- Tamrat WZ, Rose J, Grauby O, Doelsch E, Levard C, Chaurand P, and Basile-Doelsch I (2019) Soil organo-mineral associations formed by co-precipitation of Fe, Si and Al in presence of organic ligands. *Geochimica et Cosmochimica Acta* 260: 15–28.
- Torn MS, Trumbore SE, Chadwick OA, Vitousek PM, and Hendricks DM (1997) Mineral control of soil organic carbon storage and turnover. *Nature* 389: 170–173.
- Ugolini FC and Dahlgren RA (2002) Soil development in volcanic ash. *Global Environmental Research* 6(2): 69–81.
- Vandickelen R, Deroy G, and Vansant EF (1980) New-Zealand allophanes: A Structural study. *Journal of the Chemical Society-Faraday Transactions I* 76: 2542–2551.
- Wagai R, Kajiura M, and Asano M (2020) Iron and aluminum association with microbially processed organic matter via meso-density aggregate formation across soils: Organo-metallic glue hypothesis. *The Soil* 6(2): 597–627.
- Watanabe T, Ueda S, Nakao A, Ze AM, Dahlgren RA, and Funakawa S (2023) Disentangling the pedogenic factors controlling active Al and Fe concentrations in soils of the Cameroon volcanic line. *Geoderma* 430: 116289.

Metal oxides

Andreas C Scheinost^{a,b} and Balwant Singh^c, ^aThe Rossendorf Beamline at ESRF, Grenoble Cedex, France; ^bHZDR Institute of Radiochemistry, Dresden, Germany; ^cSydney Institute of Agriculture, School of Life and Environmental Sciences, The University of Sydney, NSW, Australia

© 2023 Elsevier Ltd. All rights reserved.

This is an update of A.C. Scheinost, METAL OXIDES, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 428-438, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00194-6>.

Introduction	135
Iron oxides	136
Goethite (α -FeOOH)	137
Hematite (α -Fe ₂ O ₃)	140
Ferrihydrite (Fe ₅ HO ₈ ·4H ₂ O)	140
Lepidocrocite (γ -FeOOH)	140
Magnetite (Fe ₃ O ₄) and maghemite (γ -Fe ₂ O ₃)	141
Rare Fe oxides	141
Manganese oxides	142
Phyllosulfates (layer structures)	142
Tectosulfates (tunnel structures)	143
Chain structures	143
Aluminum oxides	143
Gibbsite	144
Rare Al oxide minerals	144
Aluminum hydrolysis and poorly crystalline Al hydroxides	145
Silicon oxides	145
Quartz and cristobalite	145
Opal	146
Titanium oxides	146
Conclusions and the future	147
References	147

Abstract

Oxide minerals (include oxides, hydroxides, oxyhydroxides and hydrated oxides) are primary and secondary minerals of Si, Fe, Mn, Al and Ti. Secondary oxides, particularly of Fe, Al and Mn, are perhaps the most reactive and important components in many soils due to their large specific surface area and strong sorption capacity for many essential and potentially toxic elements. Iron and Mn oxides also play key roles in many redox reactions and have a major influence on the transformation and availability of organic and inorganic contaminants in soils. This chapter provides an overview of the oxide minerals in soils, with an emphasis on their structure and composition, environmental conditions for their formation and their properties.

Key points

- Metal oxides are important constituents of most soil types.
- Iron, Al and Mn oxides contribute to many chemical reactions in soils.
- In highly weathered soils, Fe oxides control most chemical reactions, also contribute to soil physical and microbial properties.
- Both Fe and Al oxides contribute to the preservation of organic matter in the soil.
- Iron and Mn oxides participate in redox reactions in soils and change the availability of various elements and contribute to the degradation of organic contaminants.

Introduction

The oxide minerals typically present in soils comprise oxides, hydroxides, oxyhydroxides and hydrated oxides of silicon (Si), iron (Fe), manganese (Mn), aluminum (Al) and titanium (Ti). With the exception of the Si oxide quartz and some Ti oxides, which are predominately inherited from parent rock (i.e. primary minerals), most oxides form in soil (i.e. secondary minerals) (Kämpf et al.,

2000). The metal cations, Fe^{2+} , Mn^{2+} , Ti^{4+} , Al^{3+} and Si^{4+} , of oxides are released with the weathering of silicates. The divalent cations Mn^{2+} and Fe^{2+} oxidize and hydrolyze, and precipitate almost exclusively as oxide minerals. The more abundant cations and those with smaller ionic radii, i.e. Al^{3+} and Si^{4+} , have a strong tendency to form secondary aluminosilicate minerals; the formation of Al and Si oxides requires special conditions (Waychunas, 1991). All oxides have low solubility at common soil pH values, and are therefore enriched during pedogenesis. Highly weathered soils in tropical and sub-tropical regions, which have lost a substantial part of their alkali metals, alkaline earths and Si, may contain as much as 50 wt-% metal oxides.

Oxides of Fe, Mn and Al in soil generally exhibit a high specific surface area with reactive surface sites which bind oxyanions strongly (e.g. phosphate (PO_4^{3-}), arsenate (AsO_4^{3-}), selenate (SeO_4^{2-}), molybdate (MoO_4^{2-}) and metal cations (e.g. lead (Pb^{2+}), copper (Cu^{2+}), nickel (Ni^{2+}), zinc (Zn^{2+}), cadmium (Cd^{2+})) thereby affecting the availability and mobility of plant nutrients and toxic metals (Cornell and Schwertmann, 2003). X-ray absorption fine structure (XAFS) spectroscopy has been widely used to probe directly the nature of adsorption reactions of various ions on metal oxides (Ginder-Vogel and Sparks, 2010). Such studies have enabled determination of the exact nature of adsorption complexes, such as, inner- vs. outer-sphere surface adsorption complexes, monodentate or bidentate linkage to the oxide surfaces, mononuclear vs. multinuclear metal surface complexes, and adsorbed vs. co-precipitated metal species (Waychunas et al., 2005).

In the last two decades, it has been recognized that Fe and Al oxides play a crucial role in the stabilization of soil organic matter (SOM) (Kaiser and Guggenberger, 2007; Kleber et al., 2015; Coward et al., 2017; Heckman et al., 2018; Kirsten et al., 2021). Soil organic matter can be stabilized by various processes, such as by forming chemical complexes on the reactive surfaces of Fe and Al oxides, coprecipitation in the mineral structures, forming aqueous complexes and by incorporation within soil aggregates (Kleber et al., 2015).

Some Mn oxides have a strong oxidizing power and may degrade organic pesticides, decrease or increase the toxicity of trace metals, and contribute to the formation of soil organic matter (Grangeon et al., 2020). In contrast, sand-sized crystals of the Si oxide quartz are chemically inert.

The basic structural unit of Fe, Mn, Al and Ti oxides are cationic metal centers bound to six oxygens or hydroxyls. If one drastically reduces the size of the oxygen in relation to the metal cations, this arrangement can be visualized as an octahedron, which is the basic structural unit of these oxides (Fig. 1). The octahedra may be linked to each other in three ways, sharing oxygen corners (1 O), sharing edges (2 O) or faces (3 O). Many oxide minerals like goethite contain only the first two types of connections, while the Fe oxide hematite contains all three. In addition, different structures of Fe, Mn, Al and Ti oxides result from differences in the stacking arrangements of oxygen and hydroxy planes, continuity of the structure in the direction of stacking, and the extent and way the octahedral sites are filled. Silicon is bound to four oxygens, forming a tetrahedron. These tetrahedra are connected via corners only. We will use this polyhedral approach in the text and in figures to introduce the mineral species and highlight their major differences and similarities. For each oxide group, we will give an overview including general formation pathways and their influence on soil properties. This overview is followed by a description of the minerals, their occurrence in soil, and specific conditions of their formation.

Iron oxides

Iron is released by the weathering of Fe(II)-containing silicates (e.g. biotite, pyroxenes, amphiboles, olivines). After oxidation to Fe^{3+} and hydrolysis, most Fe precipitates as Fe(III) oxides. A small proportion of the Fe becomes the structural part of phyllosilicates such

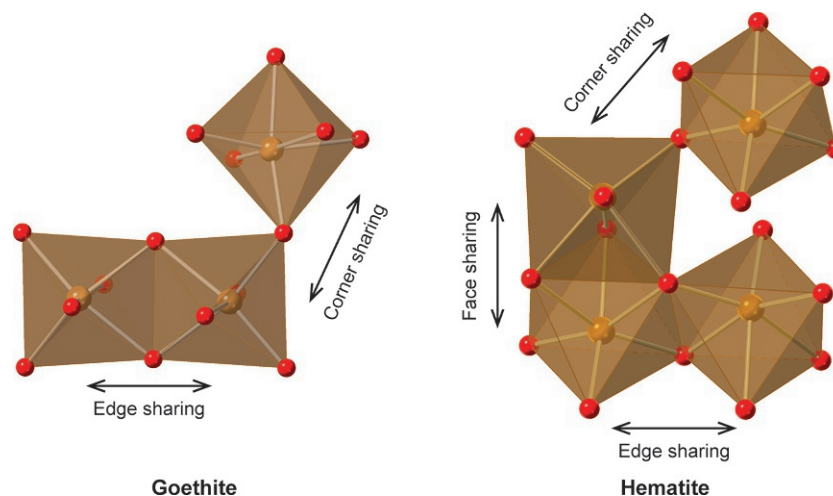


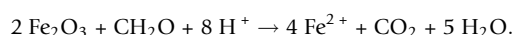
Fig. 1 Linkage of metal octahedra by corners, edges and faces, involving one, two, and three oxygen ligands, respectively. Small structural sub-units are shown of the two most common Fe oxides - goethite and hematite.

Table 1 Iron oxides found in soil environments.

Mineral name Formula	Structural type	Median color Range of hue	Typical crystal shape	Occurrence	Figure
Goethite α -FeOOH	diaspore	strong yellowish-brown 7.3YR – 1.5Y	needles, laths, rod, sub-rounded and rounded plates	Ubiquitous in soils of all climates	2A
Lepidocrocite γ -FeOOH	boehmite	moderate orange 4.9 YR - 7.9 YR	laths	Seasonally anaerobic, noncalcareous soils in cool-temperate climate	2D
Akaganéite β -FeOOH Cl	hollandite	strong brown 1.2 YR - 6.8 YR	spindle-shaped rods	Hot springs, volcanic deposits	2G
Feroxyhyte δ' -FeOOH		brownish orange 2.8 YR - 9.2 YR	spheres		2F
Schwertmannite Fe ₈ O ₈ (OH) ₆ SO ₄	hollandite	dark orange yellow 6.2 YR - 0.3 Y	hedge-hog aggregates	Acid-sulfate springs and mine drainages	2G
Ferrihydrite Fe ₅ HO ₈ · 4H ₂ O		strong brown 3.7 YR - 5.4 YR	plates, needles	Frequent in seasonally wet and/or organic-matter rich soils in cool-temperate climate	2C
Hematite α -Fe ₂ O ₃	corundum	moderate reddish brown 3.5 R - 4.1 YR	hexagonal plates, rods, spheres	Frequent in soils of tropical, subtropical and Mediterranean climate	2B
Magnetite Fe ^{II} Fe ^{III} ₂ O ₄	inverse spinel	dark yellowish brown 6 YR - 9.5YR	laths, cubes	Inherited from magnetite-containing bedrock, microbial origin possible	2E
Maghemite γ -Fe ₂ O ₃	defect spinel	black	octahedra	Common in highly-weathered, tropical and subtropical soils	2E
Green rust Fe ^{II} Fe ^{III} (OH) ₄ Cl	hydrotalcite (pyroaurite)	Bluish-green	plates	Permanently wet subsoils	2H

Bold letters indicate minerals commonly occurring in soil.

as smectite. Under oxic conditions, Fe oxides are very insoluble and are enriched during pedogenesis. They dissolve readily, however, under reducing soil conditions:



This reaction is controlled by microorganisms, which metabolize biomass (represented by CH₂O in the equation above), and transfer freed electrons to Fe³⁺, thereby reducing it to Fe²⁺. Dissolution of Fe oxides is therefore linked to microbial activity. Depending on parent material, degree of weathering and redox conditions, the amount of Fe oxides in soils varies between <1 and > 500 g kg⁻¹. Iron oxides are commonly the strongest pigments in soils, and small quantities may impart vivid colors of red, orange, yellow, brown and even blue-green to soils (Table 1) (Scheinost and Schwertmann, 1999). Depending on pedogenesis, they can occur evenly distributed in the soil matrix, concentrated in certain horizons, or locally enriched as skins, mottles, concretions. Soil horizons indurated by Fe oxides form ferricretes or laterites (Cornell and Schwertmann, 2003).

Soil Fe oxides commonly exist as very small crystals (2–50 nm) and have a large specific surface area (50–450 m² g⁻¹). Due to their circum-neutral point of zero charge (PZC), they contribute to the anion exchange properties of soils. In addition, their surfaces have a strong affinity for oxyanions and many metal cations. In highly weathered soils, there is a significant structural substitution of Al for Fe³⁺, which can be as high as 35 mol% Al in goethite and 18 mol% Al in hematite and maghemite (Singh and Gilkes, 1992). Finally, other metal cations with an ionic diameter similar to Fe³⁺ (e.g., Ni²⁺, Ti⁴⁺, Mn⁴⁺, Co³⁺ (cobalt), Cr³⁺ (chromium), Cu²⁺ (copper), Zn²⁺, V³⁺ (vanadium)) may also partially replace Fe in the crystal structure of Fe oxides, including goethite and hematite (Kaur et al., 2009a, b; Singh et al., 2010; Singh and Gilkes, 1992). Iron oxides interact strongly with organic matter in soils and play an important role in the preservation of organic matter and soil aggregation. Because of these properties, Fe oxides play a significant role in the environmental cycling of plant nutrients such as N and P, of transition metals relevant for metabolic reactions (e.g. Mn, Zn, Co), and of toxic elements like arsenic (As) and uranium (U). The 10 Fe oxide minerals observed in soils or sediments at the Earth's surface, and some of their diagnostic criteria are listed in Table 1. Crystal structures are shown in Fig. 2. Schematic formation and transformation pathways of the most common Fe oxides are shown in Fig. 3.

Goethite (α -FeOOH)

With its strong stability, goethite is the most ubiquitous Fe oxide mineral in soils. In the absence of hematite (see below), its yellowish color is responsible for the yellowish-brown hue of many soils (7.4 YR – 3.5 Y). Goethite formation is favored by the slow hydrolysis of Fe³⁺ hydroxy cations at low temperature (Bigam et al., 2002; Schwertmann and Taylor, 1989). The even distribution of goethite in a soil horizon indicates therefore that the soil has formed under aerated, temperate, humid conditions. The crystal structure of goethite consists of double chains of edge-shared octahedra that are joined to other double chains by sharing corners and by hydrogen bonds (Fig. 2A).

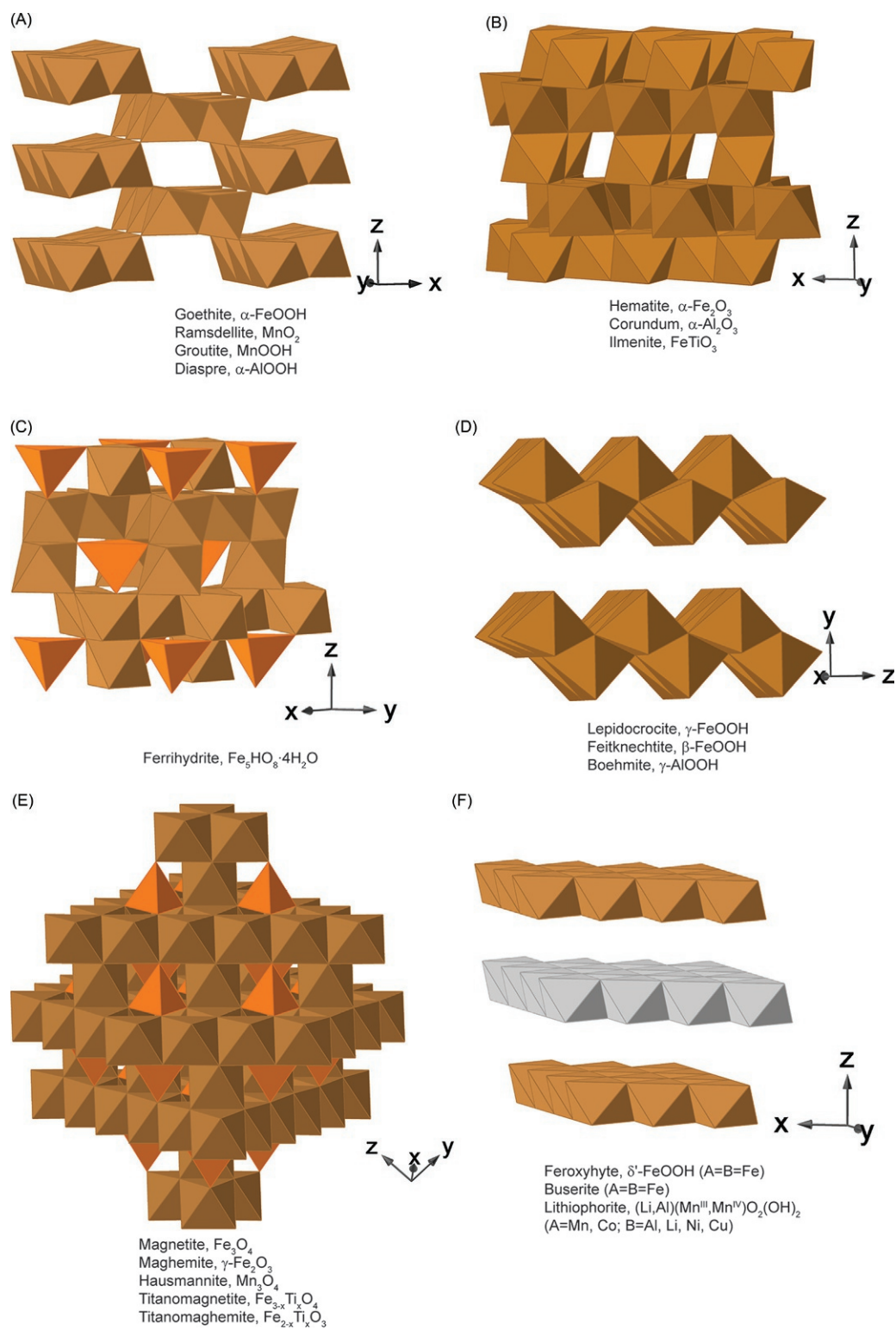


Fig. 2 Crystal structures of common oxide minerals in soils.

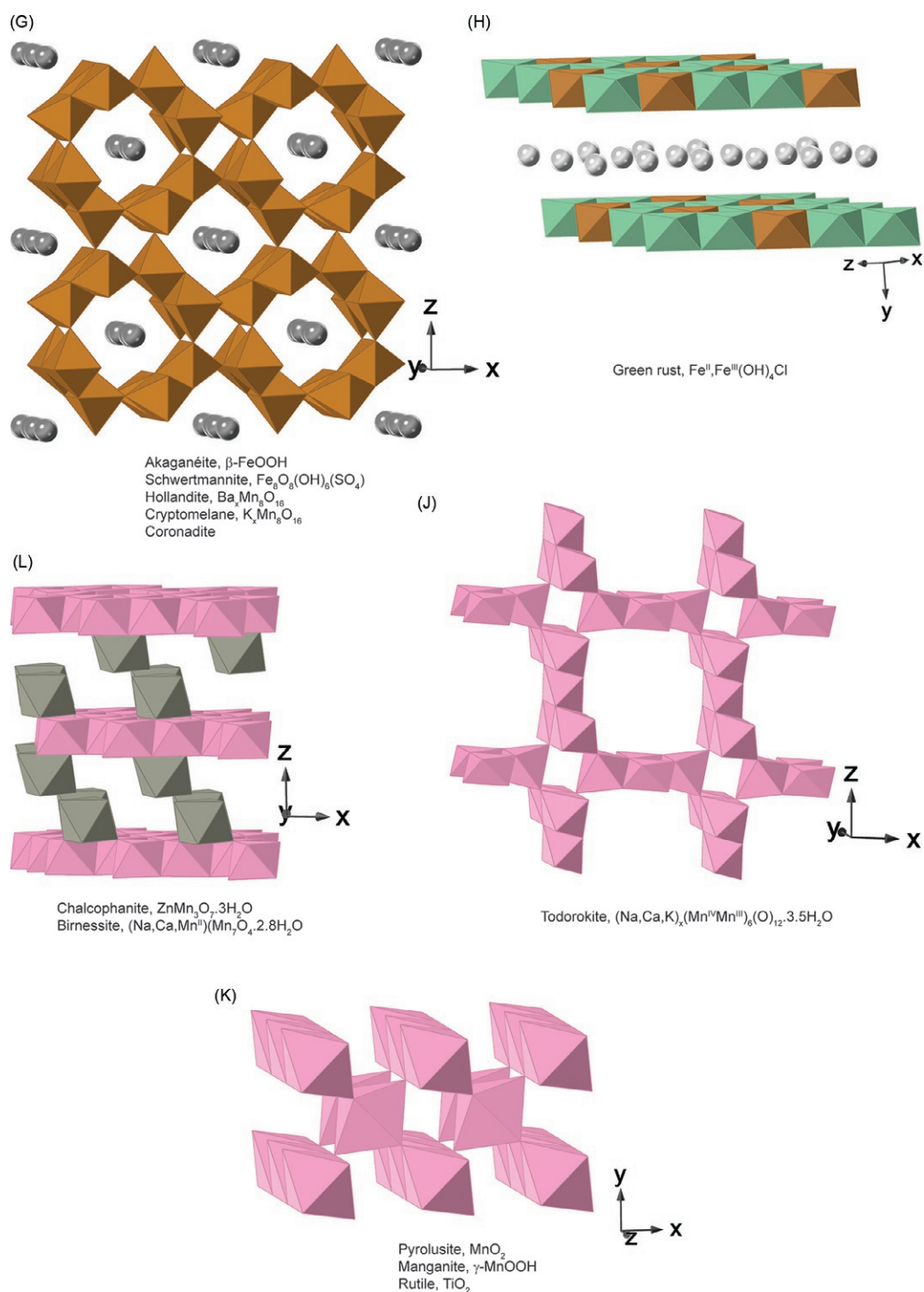


Fig. 2—Cont'd

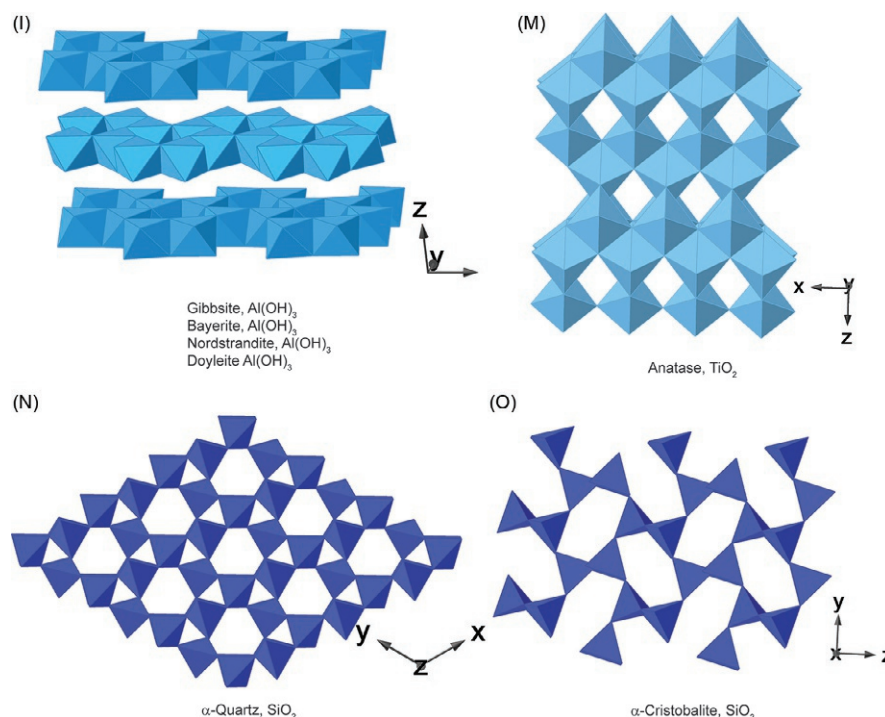


Fig. 2—Cont'd

Hematite ($\alpha\text{-Fe}_2\text{O}_3$)

Soils forming in subtropical or tropical climates often have reddish or even purplish hues (8.0 R – 1.4 Y), derived from blood-red hematite. Hematite usually coexists with goethite, but due to its greater tinting strength it masks the yellowish color of goethite even at low concentrations. Hematite forms by dehydroxylation of ferrihydrite (see below) via a solid-state mechanism or topotactic transformation involving aggregation and crystallization or dehydration with an aqueous-aided 2-D growth mechanism (Fischer and Schwertmann, 1975; Schwertmann and Murad, 1983; Vu et al., 2008). In comparison to goethite, the formation of hematite is favored by neutral pH, increasing soil temperature and decreasing water activity (Schwertmann and Taylor, 1989). The crystal structure of hematite consists of sheets of edge-sharing octahedra. The sheets are connected by edge- and face-sharing octahedra (Fig. 2B). The unique face-sharing arrangement and the resulting proximity of neighboring Fe centers is responsible for the red color of this mineral.

Ferrihydrite ($\text{Fe}_5\text{HO}_8 \cdot 4\text{H}_2\text{O}$)

This meta-stable and poorly crystalline, hydrous oxide of Fe commonly occurs in soils formed under cooler conditions and with changing redox conditions. Its formation is favored by the rapid oxidation and hydrolysis of Fe^{3+} from solution. The transformation into thermodynamically more stable goethite may be hindered by adsorption of silicate, phosphate and organic anions, making ferrihydrite a common constituent of younger soils in association with goethite. The brownish-orange color of ferrihydrite often marks gradients between anoxic and oxic soil regions, e.g. at the transition between water-saturated and aerated soil horizons, and along root channels or other macropores. Ferrihydrite is also responsible for the brownish-orange color in podzol B horizons. Because of the very small crystal size and lack of long-range structural order, ferrihydrite's structural model continues to remain a contentious issue. However, there is mounting evidence that at an early stage the ferrihydrite structure may contain up to 20% Fe in the tetrahedral coordination and the remainder in two different types of octahedral coordination (Fig. 2C) (Funnell et al., 2020; Gilbert et al., 2013).

Lepidocrocite ($\gamma\text{-FeOOH}$)

This vividly orange, less abundant mineral occurs in seasonally anaerobic, clay-rich, carbonate-free soils in humid-cool climates in association with the more stable goethite. Lepidocrocite forms through relatively slow oxidation of Fe^{2+} at low concentrations of carbonate. Its crystal structure contains double chains of octahedra which are joined by shared edges, resulting in corrugated sheets of octahedra. These corrugated sheets are stacked one on top of the other and are held together by hydrogen bonds (Fig. 2D).

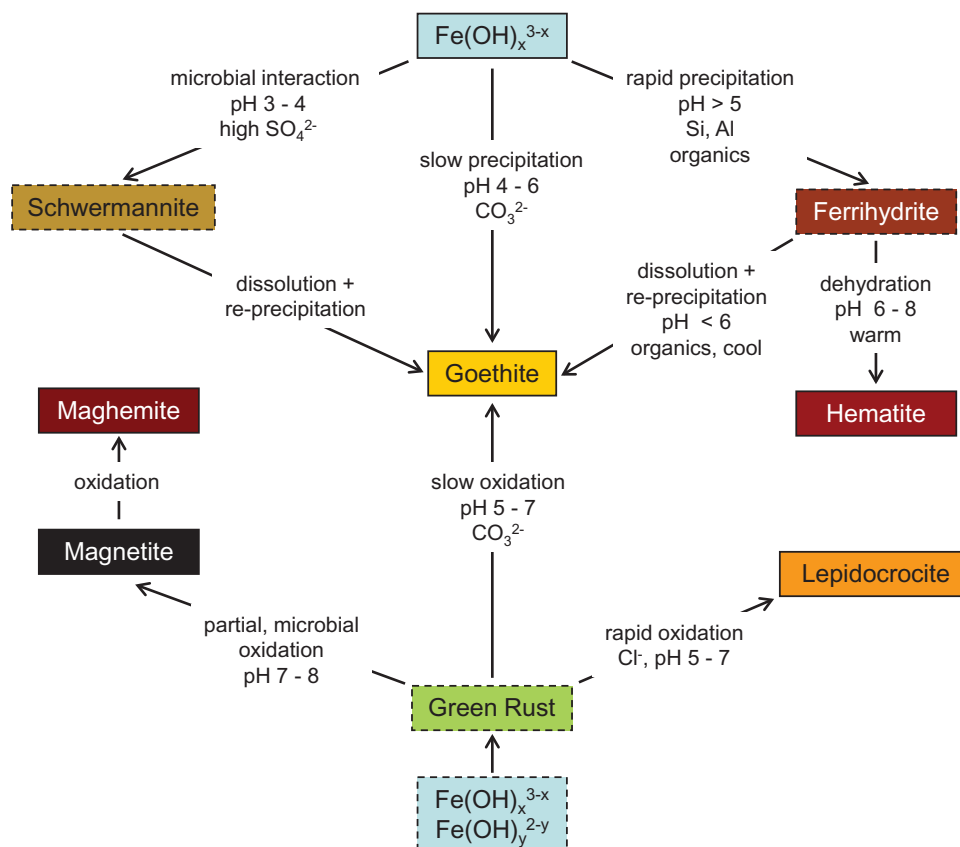


Fig. 3 Schematic formation and transformation pathways of common Fe oxides in soils. The precipitation of Fe oxides proceeds from either ferric hydrolysis species (top) or from mixed ferric-ferrous hydrolysis species (bottom). Meta-stable phases are shown in dotted lined boxes. Figure modified after Bigham JM, Fitzpatrick RM, and Schulze DG (2002) Iron oxides. In: Dixon JB and Schulze DG (eds) *Soil Mineralogy with Environmental Application*, pp. 323–366. Madison, WI: Soil Science Society of America; and Cornell RM and Schwertmann U (2003) *The Iron Oxides: Structure, Properties, Reactions, Occurrence and Uses*, 2nd edn. Weinheim, Germany: VCH Verlagsgesellschaft.

Magnetite (Fe_3O_4) and maghemite ($\gamma\text{-Fe}_2\text{O}_3$)

Blackish magnetite is usually inherited from the parent material, but may also form in soil through biotic processes. The brownish-red maghemite is common in soils of tropical and subtropical climates, but also occurs occasionally in temperate soils that have been exposed to forest or heath fires. Maghemite forms either through the oxidation of magnetite, or by heating (300–425 °C) other Fe oxides in the presence of organic compounds (forest fires, campfires, fires in the hearth and other human induced or natural fires). Both magnetite and maghemite have the same inverse spinel structure, with octahedral and mixed tetrahedral or octahedral layers stacked along the crystal direction [111] (Fig. 2E). In magnetite, which contains both Fe^{2+} and Fe^{3+} ions, Fe^{3+} occupies both tetrahedral and octahedral sites, while the larger Fe^{2+} ion occupies octahedral sites only.

Rare Fe oxides

Several minerals have rarely or never been observed in soils. They may form in extreme natural environments (akaganéite, schwermannite), in polluted surface waters (schwermannite, feroxyhyte), or transform rapidly into more stable phases (green rust). *Feroxyhyte*, $\delta'\text{-FeOOH}$, was observed in rusty precipitates of rapidly flowing Fe^{2+} -containing water which was quickly oxidized. It consists of sheets of edge-sharing octahedra (Fig. 2E).

Akaganéite ($\beta\text{-FeOOH Cl}$) has been found in environments with large chloride concentrations (0.1 M), low pH (3–4), and high temperature (60 °C), e.g. in hot springs, inland acid sulfate soils and volcanic deposits (Bibi et al., 2011; Bigham et al., 2002). Akaganéite consists of 2×2 channels (referring to the width of the channels in octahedral units) formed by edge-sharing double chains (Fig. 2G). These tunnels are 0.5 nm in cross section and contain Cl^- (chloride) ions that stabilize the structure.

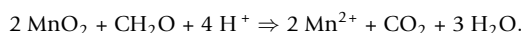
Schwermannite, $\text{Fe}_8\text{O}_8(\text{OH})_6\text{SO}_4$, is isostructural with akaganéite, but SO_4^{2-} occupies the tunnels (Fig. 2G). Schwermannite has been found in strongly acid sulfate waters associated with mining activities (acid mine drainage) or pyritic rock outcrops (Bigham et al., 1990, 2002).

Green rust, also called fougérite $[\text{Fe}^{\text{II}}_{1-x}\text{Fe}^{\text{III}}_x(\text{OH})_2]^{x+} \cdot (x/n)\text{A}^{n-} \cdot m\text{H}_2\text{O}$ (A: anions like Cl^- , NO_3^-), has been observed rarely in reducing, weakly acid to weakly alkaline subsoils. Its greenish-blue color changes rapidly to yellowish brown on exposure to the air,

indicating its instability against oxidation. Green rust consists of sheets of edge-sharing octahedra, which are occupied by both Fe^{2+} and Fe^{3+} . The resulting positive charge of the layers is balanced by hydrated anions in the interlayer space (Fig. 2H). Because of its anion exchange capacity, redox reactivity and high surface area, even small amounts of green rust may play an important role in aquifers and the vadose zone. Since the mineral forms as a metastable corrosion product of metallic iron under anaerobic conditions, green rust is also relevant for barriers of scrap iron engineered to clean up groundwater.

Manganese oxides

Manganese is released by weathering of Mn(II)-containing silicates (biotite, pyroxenes, amphiboles). After oxidation of the soluble Mn^{2+} to Mn^{3+} and Mn^{4+} , brownish-black Mn oxides of low solubility precipitate, which re-dissolve only under reducing soil conditions. The reaction equation assuming an Mn(IV) oxide can be written in an analogous way to the dissolution of Fe oxides



As for the Fe oxides, dissolution depends on microbial activity. While oxidation of Mn^{2+} is catalyzed by mineral surfaces, microbial oxidation seems to be the more dominant process in soil environments. Therefore, both formation and dissolution of Mn oxides are intimately linked to microbial activity in soils (Banfield and Nealson, 1998; Grangeon et al., 2020).

Manganese oxides may occur evenly distributed in the soil matrix like the stable Fe oxides goethite and hematite. With their relatively small concentration (about 1/50 of the concentration of Fe oxides) and their poor crystallinity, however, they are extremely difficult to identify. Therefore, most soil Mn oxides have been identified in local enrichments, like black coatings on peds, dendrites or nodules, which form along redox gradients. The blackish colors of Mn oxides are often separated from the brownish-orange color of Fe oxides (ferrihydrite) by extending further toward the higher redox potential, indicating the greater mobility of Mn^{2+} compared to Fe^{2+} .

Manganese oxides can accumulate a wide range of other elements such as lithium (Li), barium (Ba), As, Pb and almost all 3d (group 3 in the d-block of the periodic table) transition metals. These elements are either part of the crystal structure, or they are tightly adsorbed to the large surface area of Mn oxides. Manganese oxides have an oxidizing potential stronger than that of O_2 . Therefore, they can oxidize inorganic ions such as Co^{2+} , Cr^{3+} and As^{3+} , thereby either increasing (e.g., Cr^{6+}) or decreasing (e.g., Co^{3+} , As^{5+}) their mobility and toxicity. They are also able to oxidize organic molecules, thereby enhancing the degradation of anthropogenic compounds such as pesticides. Furthermore, Mn oxides catalyze condensation reactions like the Maillard reaction. Hence, Mn oxides may play an important role in the abiotic formation of humic substances.

In spite of their importance for the geochemistry of metals and carbon, the knowledge of Mn oxides and their formation in soils remains relatively poor. Manganese oxides often occur as nanoparticles, typically in the 5–100 nm size range, with poor crystallinity and small concentrations, all of these and the occurrence of Mn in different valence forms complicates the determination of their exact mineralogical nature in soils (Bargar et al., 2009; Grangeon et al., 2020). The structures of Mn oxides have been determined using synchrotron-based X-ray diffraction and XAFS spectroscopy (Silvester et al., 1997; Post et al., 2002). More recently, Raman spectroscopy has been shown to be an effective technique for routine identification and characterization of layer- and tunnel-structure Mn oxide phases (Post et al., 2020, 2021). Manganese can occur in Mn(II), Mn(III), and Mn(IV) oxidation states depending on the Eh/pH conditions prevailing in the soil environment. Iltona et al. (2016) developed a procedure using X-ray photoelectron spectroscopy (XPS) to quantify the Mn oxidation state in hydrous multivalent Mn oxides such as birnessite.

Based on the arrangement and linkage of their metal octahedra, the Mn oxides have been categorized into three major groups: (1) phylломanganates or layer structures, (2) tectomanganates or tunnel structures, and (3) chain structures. The most common Mn oxide minerals in soils are listed in Table 2, their crystal structures are shown in Fig. 2 (Waychunas, 1991).

Phylломanganates (layer structures)

Phylломanganates are commonly observed after microbially catalyzed oxidation of Mn^{2+} and they are ubiquitous in soils.

Birnessite forms in a wide variety of soils. It consists of sheets of edge-sharing MnO_6 octahedra with vacancies. The resulting negative layer charge is compensated by cations like Na^+ , K^+ , Ca^{2+} , Mg^{2+} , and Mn^{2+} , which may cover the vacancies. The interlayer region contains water, resulting in a d-spacing of 7 Å (Fig. 2I). Three other minerals, vernadite, chalcophanite and busserite, are structurally very similar to birnessite.

Vernadite is frequently observed as an initially amorphous oxidation product of microorganisms and has been found in soils and in freshwater lakes. Its structure may be similar or identical to turbostratic H^+ -birnessite.

Chalcophanite has been observed in Zn-contaminated soils. In comparison to birnessite, it has a greater number of vacancies that are covered by Zn in octahedral or tetrahedral coordination (Fig. 2J).

Busserite forms in hydrothermal environments. Compared with the former phyllosilicates, it has a larger d-spacing of 10 Å, probably due to a larger amount of interlayer water. After dehydration, busserite transforms into birnessite.

Lithiophorite (a basic oxide of Li, Mn and Al) has been frequently identified in soils. Its layer structure consists of sheets of edge-sharing MnO_6 octahedra which alternate with sheets of (Al, Li)(OH)₆ octahedra (Fig. 2F). The cation sites in the Mn—O octahedral sheet are occupied with Mn^{4+} and Mn^{3+} . The net negative layer charge is counterbalanced by the positive charge of the (Al, Li)-OH octahedral sheets. Lithium is frequently replaced by Ni, Cu and Zn, while Co substitutes for Mn within the octahedral Mn sheets.

Table 2 Manganese oxides found in soil environments.

Mineral name (synthetic equivalent)	Formula	Structural type	Figure
Birnessite	$(\text{Na,Ca,Mn}^{2+})\text{Mn}_7\text{O}_4 \cdot 2.8 \text{ H}_2\text{O}$	layer structure	2I
Vernadite	$\text{MnO}_2 \cdot n\text{H}_2\text{O}$	layer structure	—
(δ - MnO_2)			
Chalcophanite	$\text{ZnMn}_3\text{O}_7 \cdot 3 \text{ H}_2\text{O}$	layer structure	2I
Buserite	$\text{Na}_4\text{Mn}_{14}\text{O}_{27} \cdot 21 \text{ H}_2\text{O}$	layer structure	—
Lithiophorite	$\text{LiAl}_2\text{Mn}_2^{4+}\text{Mn}^{3+}\text{O}_6(\text{OH})_6$	layer structure	2F
Todorokite	$(\text{Na,Ca,K})_x(\text{Mn}^{4+}\text{Mn}^{3+})_6\text{O}_{12} \cdot 3.5 \text{ H}_2\text{O}$ ($x = 0.3\text{--}0.5$)	3×3 tunnel	2J
Romanechite	$\text{Ba}_{0.66}(\text{Mn}^{4+}\text{Mn}^{3+})_5\text{O}_{10} \cdot 1.34\text{H}_2\text{O}$	3×2 tunnel	—
Hollandite	$\text{Ba}_x(\text{Mn}^{4+}\text{Mn}^{3+})_8\text{O}_{16}$ ($x = 1$)	2×2 tunnel	2G
(α - MnO_2)			
Cryptomelane	$\text{K}_x(\text{Mn}^{4+}\text{Mn}^{3+})_8\text{O}_{16}$ ($x = 1.3\text{--}1.5$)	2×2 tunnel	2G
Coronadite	$\text{Pb}_x(\text{Mn}^{4+}\text{Mn}^{3+})_8\text{O}_{16}$ ($x = 1\text{--}1.4$)	2×2 tunnel	2G
Pyrolusite	MnO_2	rutile	2K
(β - MnO_2)			
Manganite	MnOOH	rutile	2K
(γ - MnOOH)			
Ramsdellite	MnO_2	diaspore	2A
Groutite	MnOOH	diaspore	2A
(α - MnOOH)			
Feitknechtite	MnOOH	boehmite	2D
(β - MnOOH)			
Hausmannite	Mn_3O_4	inverse spinel	2E

Bold letters indicate minerals commonly occurring in soil.

Asbolane forms in soils with large Co and Ni concentrations. Its structure is very similar to that of lithiophorite, with alternating sheets of Mn^{4+} -O octahedra and Co—Ni—OH octahedra (Fig. 2F). The Co—Ni sheet may be discontinuous (island-like).

Tectomanganates (tunnel structures)

Tectomanganates are formed from double or triple chains of MnO_6 octahedra, that form tunnels (or channels). These tunnels are occupied by large foreign cations and water molecules.

Todorokite is the most frequently observed soil tectomanganate. It consists of triple chains of edge-sharing MnO_6 octahedra linked to form large 3×3 tunnels. These tunnels contain Na, Ca, K, Ba, strontium (Sr) and water molecules (Fig. 2J). The occurrence of variable tunnel widths (3×2 , 3×3 , 3×4 , and 3×5) observed in HRTEM images suggests that todorokite represents a family rather than a single mineral.

Hollandite, *cryptomelane*, and *coronadite*, sometimes grouped as α - MnO_2 , have rarely been observed in soils. They consist of double chains of edge-sharing MnO_6 octahedra linked to form 2×2 tunnels (Fig. 2G). The tunnels contain water molecules together with cations, which are primarily Ba^{2+} in hollandite, K^+ in cryptomelane, and Pb^{2+} in coronadite. Natural samples usually contain a variety of cations in the tunnels. The large cations are located at specific tunnel sites, and their presence is necessary to prevent the structure from collapsing. The charges of the tunnel cations are balanced by the substitution of Mn^{4+} by Mn^{3+} in the tunnel walls.

Chain structures

Manganese oxides of this group have been synthesized in the laboratory at ambient temperature and pressure, and at moderately alkaline pH. However, they have rarely been observed in soil. Pyrolusite (MnO_2) and Manganite (γ - MnOOH) consist of single chains of edge-sharing MnO_6 octahedra which form 1×1 pseudo-tunnels (Fig. 2K). Ramsdellite (MnO_2) and Groutite (α - MnOOH) are isostructural with goethite (Fig. 2A); Feitknechtite (β - MnOOH) has a structure similar to lepidocrocite (Fig. 2D). Hausmannite (Mn_3O_4) has a disordered-spinel structure analogous to magnetite (Fig. 2E).

Aluminum oxides

In contrast to Fe and Mn, most Al released by the weathering of Al-containing silicates is taken up by secondary, pedogenic phyllosilicates. As long as the Si concentration in the soil solution is $>0.5 \text{ mg L}^{-1}$, the formation of phyllosilicates out-competes the formation of Al oxides. Furthermore, the complexation of Al^{3+} by organic matter prevents Al oxide precipitation in soils. Therefore, Al oxides are commonly observed only in organic matter-poor soils of tropical and subtropical climates, where intense weathering

Table 3 Aluminum oxides found in soils.

<i>Mineral</i>	<i>Formula</i>	<i>Figure</i>
Gibbsite	$\gamma\text{-Al(OH)}_3$	2L
Nordstrandite	Al(OH)_3	2L
Bayerite	Al(OH)_3	2L
Doyleite	Al(OH)_3	2L
Diaspore	$\alpha\text{-AlOOH}$	2A
Boehmite	$\gamma\text{-AlOOH}$	2D
Corundum	$\alpha\text{-Al}_2\text{O}_3$	2B

Bold letters indicate minerals commonly occurring in soil.

causes secondary phyllosilicates to dissolve and Si is lost by leaching. The (colorless) Al oxides are either evenly distributed within the soil matrix or enriched in concretions (pisolithes). Horizons cemented by Al oxides are called alcretes.

Aluminum oxides have a large specific surface area (up to $600 \text{ m}^2 \text{ g}^{-1}$), a high PZC (pH 8–10), and a pH-dependent surface charge. Similar to Fe and Mn oxides, Al oxides strongly adsorb heavy metals (Cu, Pb, Zn, Ni, Co, Cd) and anions (phosphate, silicate, molybdate, sulfate, catechol) by forming inner-sphere complexes (Hsu, 1989; Huang et al., 2002). The adsorption is related to the reactive surface area (singly coordinated OH^- groups at crystallite edges) rather than the total surface area. Aluminum oxides, in addition to Fe oxides, are responsible for soil aggregation of Ferralsols (Oxisols) and Alisols or Acrisols (Ultisols), but the mechanisms of aggregation are unclear. Aluminum oxides have a low solubility at slightly acidic and neutral pH, but at low pH soluble species form, particularly Al^{3+} , that are toxic to plants and animals. These soluble species affect food production on naturally acidic soils and contribute to forest decline in areas influenced by acid rain.

Of the six different Al oxide minerals, predominately gibbsite and, less commonly, boehmite, nordstrandite and bayerite form under soil conditions (Table 3). Diaspore and corundum are occasionally found in bauxite deposits, but are rare in soils. Aluminum also forms highly reactive, poorly crystalline precipitates and colloids (Huang et al., 2002).

Gibbsite

Of the four Al(OH)_3 polymorphs (Table 3), gibbsite ($\gamma\text{-Al(OH)}_3$) is most common in soils. It consists of sheets of edge-sharing octahedra stacked along the c axis (Fig. 2L). Only two-thirds of the available octahedral sites are filled with Al^{3+} ions analogous to the dioctahedral sheets of phyllosilicate minerals. In soil, gibbsite forms thick, platy crystals. In saprolites, feldspars (plagioclase) may transform directly into gibbsite while maintaining the original shape of feldspar crystals (pseudomorphs).

Gibbsite is found in Ferralsols (Oxisols) on old, stable upland landscapes in association with Fe oxides (goethite, hematite) and kaolinite. Intense weathering and leaching under humid tropical and subtropical climates over long periods of time (millions of years) have depleted these soils of the alkali metals, alkali earth metals and Si, thereby enhancing the formation of gibbsite. The gibbsite content of Ferralsols may be as large as 600 g kg^{-1} , but some Oxisols do not contain gibbsite at all. In Alisols or Acrisols (Ultisols) and Cambisols or Inceptisols of tropical to temperate climates, gibbsite is derived from the underlying saprolite, which has formed by weathering of igneous and metamorphic rocks. The gibbsite content generally decreases from the saprolite to the soil surface. Finally, gibbsite has been observed in Andosols or Andisols, where it formed from the rapid weathering of volcanic ash. Although these soils are often very young (hundreds of years) and Si-rich, sufficient Si has been leached locally to explain the formation of gibbsite.

Gibbsite formation is conditioned by the intensity of leaching which, in turn, is affected by rainfall, temperature, parent rock, topography, ground-water table, vegetation, and time. Environments with warm temperatures, high rainfall and free drainage favor desilication and leaching of ions, as well as mineralization of organic matter. Anions such as sulfate, carbonate, phosphate and silicate, together with organic ligands like citric, malic, tannic, aspartic, and fulvic acids have a strong affinity for Al^{3+} , and they may interfere with the crystallization of Al(OH)_3 . This explains why there is either none or only small amounts of gibbsite in the A horizon of acidic soils with considerable organic matter and exchangeable Al, while it is found in greater quantities in deeper horizons. However, gibbsite may also decrease with soil depth. A possible explanation is an inverse gradient of Si(OH)_4^0 (silica) activity: rain water infiltrating at the soil surface is not yet in equilibrium with silicates and therefore has a lower silica activity than the soil water at a greater depth, which has had more time to equilibrate. The heterogeneous distribution of gibbsite in a landscape may be explained similarly by the silica activity of percolating waters.

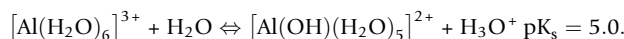
Rare Al oxide minerals

The difference between gibbsite and its other polymorphs bayerite, nordstrandite and doyleite consist of a slightly different arrangement of hydroxyl groups. While gibbsite formation is favored by the slow hydrolysis of Al and by a pH <6, nordstrandite and bayerite form at neutral to alkaline pH under fast hydrolysis. Correspondingly, their rare occurrences are related to limestone materials. Boehmite ($\gamma\text{-AlOOH}$), which is isostructural with lepidocrocite (Fig. 2D), has been identified in lateritic materials and in bauxites. It may form from gibbsite by diagenesis or hydrothermal alteration. The oxyhydroxide diasporite ($\alpha\text{-AlOOH}$) is

isostructural with goethite (Fig. 2A). It has been identified as a surface weathering product formed by the desilication of kaolinitic clay. The rarely found oxide corundum (α - Al_2O_3), which is isostructural with hematite (Fig. 2B), may be derived from corundum-containing parent rock, or form by heating of the soil by bush fires.

Aluminum hydrolysis and poorly crystalline Al hydroxides

The Al^{3+} cations released from Al-bearing minerals are initially coordinated with 6 water molecules forming the hexaquo complex $\text{Al}(\text{H}_2\text{O})_6^{3+}$. This hexaquo complex hydrolyses in several steps, the first and most important one in the soil solution is



At higher Al^{3+} or OH^- concentrations, the mononuclear complexes start to polymerize to form polynuclear species. The smallest unit identified so far is $\text{AlO}_4\text{Al}_{12}(\text{OH})_{24}(\text{H}_2\text{O})_{12}^{7+}$. This so-called Al_{13} or Keggin polymer consists of a highly symmetrical, tetrahedrally coordinated Al center enclosed in a cage-like structure of 12 octahedrally coordinated Al centers. Further polymerization of this colloidal species leads to the precipitation of a poorly crystalline solid phase.

Silicon oxides

Silicon oxides differ from the oxides of Fe, Mn and Al in several aspects. The most prevalent Si oxide, quartz (SiO_2), is a primary mineral in many rocks, contributing $\sim 120 \text{ g kg}^{-1}$ to the average composition of the Earth's upper mantle. In soils, dissolved Si associates with Al to form aluminosilicate clay minerals. However, secondary Si oxide precipitates may also form in soil, sometimes to such an extent that they cement the soil matrix to form duricrusts and silcretes. Primary quartz crystals are very pure, hard and recalcitrant, and occur in soil as large particles (sand). Their small specific surface area, their small number of reactive bonds at the surface, and the lack of isomorphic substitution makes them chemically very inert. Quartz has the lowest ion exchange capacity ($\text{CEC} = 1\text{--}2 \text{ cmol}_c \text{ kg}^{-1}$) of all soil minerals. The solubility of Si oxides is low at usual soil pH (2–8), but strongly increases at pH >8. The main soluble species, $\text{Si}(\text{OH})_4^0$ (monosilicic acid), is taken up by higher plants, especially grasses and paddy rice, to increase the mechanical strength and resistance against fungal pathogens and herbivores, and is later recycled back into the soil as biogenic opal. Radiolarians, diatoms, silica sponges and cyanobacteria are known to accumulate large amounts of Si oxides in fresh and marine waters, and microorganisms may exert a similar function in soils, competing with the inorganic formation of aluminosilicate clay minerals.

The Si oxides are tectosilicates. The repeating unit is a SiO_4 tetrahedron in which each O is linked to Si atoms of adjacent tetrahedra, forming a strong three-dimensional framework structure. This contrasts with the Fe, Al, Mn and Ti oxides in which the basic unit is a cation in an octahedral coordination. In the following, only α -quartz, α -cristobalite and opal, which commonly occur in soil, will be described (Table 4) (Heaney et al., 1994; Monger and Kelly, 2002).

Quartz and cristobalite

The α -Quartz consists of paired helical chains of corner sharing SiO_4 tetrahedra that spiral along the c-axis. The intertwined chains produce open, hexagonal channels (Fig. 2N). In cristobalite, the tunnels are larger (Fig. 2O), and the density is smaller (2.32 g cm^{-3}) compared to quartz (2.65 g cm^{-3}).

Quartz is ubiquitous in soil environments, while cristobalite is restricted to soils developed from volcanic materials. The quartz in soils is mainly a primary mineral, inherited from the parent material. In soil solution, quartz may form from soluble $\text{Si}(\text{OH})_4^0$ by polymerization, precipitation of amorphous, hydrous SiO_2 , that slowly transforms into crystalline quartz. Direct precipitation from solution seems to be unlikely. Quartz is generally concentrated in the sand and silt fractions of soils, with a lower frequency in the coarse-clay fraction ($0.2\text{--}2 \mu\text{m}$). Quartz may comprise >90% of the inorganic fraction of Quartzipsamments. By contrast, highly

Table 4 Silicon oxides found in soils.

Mineral	Formula	Structural type	Occurrence	Figure
α-Quartz	SiO_2		ubiquitous	2N
α -Tridymite	SiO_2			–
α-Cristobalite	SiO_2		volcanic deposits	2O
Opal-C	$\text{SiO}_2 \cdot x\text{H}_2\text{O}$	well-ordered α -cristobalite	volcanic deposits	–
Opal-CT	$\text{SiO}_2 \cdot x\text{H}_2\text{O}$	disordered α -cristobalite and α -tridymite	cherts, porcelanites, fossil wood, silcretes	–
Opal-A	$\text{SiO}_2 \cdot x\text{H}_2\text{O}$	nearly amorphous	biogenic ubiquitous; inorganic in strongly weathered soils, duripans, silcretes	–

Bold letters indicate minerals commonly occurring in soil.

weathered and leached Oxisols have only very small amounts of quartz left. The solubility of quartz ranges from 1 to 4 mg L⁻¹ Si depending on particle size.

Opal

Opal is classified into three structural groups, opal-C (well-ordered α -cristobalite), opal-CT (disordered α -cristobalite, α -tridymite) and opal-A (highly disordered, nearly amorphous). Opal contains 40–90 g kg⁻¹ water, and the specific gravity ranges from 1.5 to 2.3 g cm⁻³.

Opal-C and α -cristobalite occur in Andepts and other soils derived from volcanic material. Opal-CT has been identified in soils derived from bentonites, siliceous shales, and indurated silicates. Opal-CT is also a primary constituent of many cherts, porcelanites, fossil wood, silcretes, and some duripans. Opal-A may form by both organic and inorganic processes in pedogenic environments. Biogenic opal-A originates from Si accumulated by plants and aquatic organisms, and thus occurs under a wide range of environmental conditions. Biogenic opal is an almost ubiquitous constituent of soils. Opal phytoliths are mostly derived from grasses, while sponge spicules, diatoms, and radiolaria, which form in aquatic environments, may reach soils by weathering of sedimentary bedrocks. Amounts of opal phytoliths in soils commonly range from <1 to 30 g kg⁻¹, usually occurring in the 5–50 μ m size fraction, and decreasing with soil depth. Inorganic opal-A forms from supersaturated soil solutions. It occurs as nodules and as the primary cement of indurated soil horizons (duripans) and silcretes.

At ambient temperature and neutral pH, the solubility of amorphous Si is approximately 50–60 mg Si L⁻¹, and therefore much higher than that of quartz. The presence of organic molecules greatly enhances the rates of dissolution of Si oxides. Coatings of chemisorbed Al and Fe, however, reduce the rates of dissolution rates. Oxides of Fe and Al, by acting as a sink for soluble Si, greatly increase the rates of dissolution of amorphous Si oxides and of reactive, uncoated quartz surfaces. Dissolution (weathering) of Si oxides, which starts below approximately 3 mg of Si L⁻¹, is induced by the reduction of Si levels in the soil solution by leaching and plant uptake.

Titanium oxides

Titanium oxides are common in many igneous and metamorphic rocks and sediments. Because of their resistance to weathering, they are often inherited in soils. However, Ti oxides also form in soil after weathering of less resistant, Ti-bearing minerals (biotite, amphiboles). Titanium oxide minerals are relatively heavy (density > 2.9 g cm⁻³). Their concentration in soils is usually small; because of their small specific surface area, they do not contribute significantly to soil reactivity. In tropical soils only, where they become increasingly enriched during pedogenesis, has some influence on phosphate and arsenate retention has been observed. Old sand dunes, e.g., in the southeastern United States or Senegal, may contain enough Ti oxides to be mined as a commercial source for TiO₂, which is used to produce white paint. Because of their recalcitrant nature, Ti oxides can be used as reference minerals to study weathering and soil genesis.

Titanium occurs primarily in octahedral coordination, and like the Fe and Mn oxides, the structures of the various Ti oxide minerals can be described in terms of the different arrangements of the Ti-containing octahedra. The principal Ti minerals are listed in Table 5 (Fitzpatrick and Chittleborough, 2002).

Rutile (TiO₂) is isostructural with pyrolusite and manganite, consisting of single chains of edge-sharing TiO₆ octahedra (Fig. 2K). Rutile is a common residual mineral occurring in the sand and silt fractions of a variety of soils. Anatase consists of edge and corner-sharing octahedra that outline a three-dimensional framework, rather than distinct chains (Fig. 2M). The platy crystals found in soils are believed to be a weathering product of titanite (CaTiSiO₅) and Ti-bearing silicates. Anatase is easily synthesized at room temperature. Ilmenite (FeTiO₃) is almost isostructural with hematite, with one half of the octahedral centers occupied by Fe²⁺, the other half by Ti⁴⁺ (Fig. 2B). By oxidation of the structural Fe²⁺ to Fe³⁺, ilmenite weathers to pseudorutile (Fe₂O₃·nTiO₂·mH₂O; 3 < n < 5 and 1 < m < 2), a structurally disordered and poorly characterized mineral. Pseudorutile has been widely identified in

Table 5 Titanium oxides found in soils.

<i>Mineral</i>	<i>Formula</i>	<i>Figure</i>
Rutile	TiO ₂	2K
Anatase	TiO ₂	2M
Ilmenite	Fe ²⁺ TiO ₃	2B
Pseudorutile	Fe ₂ O ₃ ·nTiO ₂ ·mH ₂ O (n = 3–5, m = 1–2)	—
Ulvöspinel	Fe ₂ TiO ₄	2E
Titanomagnetite	Fe _{3–x} Ti _x O ₄	2E
Titanomaghemite	Fe _{2–x} Ti _x O ₃	2E

Bold letters indicate minerals commonly occurring in soil.

soils, primarily as the weathering product of ilmenite. Ilmenite itself is relatively unstable in soils, because it weathers easily to pseudorutile and mixtures of rutile, anatase, and Fe oxides.

Titanomagnetites ($\text{Fe}_{3-x}\text{Ti}_x\text{O}_4$) are solid solutions of magnetite with ulvöspinel (Fe_2TiO_4), with Ti^{4+} occupying only octahedral sites of the magnetite structure (Fig. 2E). Their formation is probably related to that of magnetite (see there). Titanomaghemitites ($\text{Fe}_{2-x}\text{Ti}_x\text{O}_3$) form by weathering of titanomagnetites. Their composition is between those of magnetite, maghemite and ulvöspinel. Many aspects of the Ti oxides, such as whether the minerals are relict or pedogenic, and the conditions responsible for their weathering and formation in soils, are still unresolved.

Conclusions and the future

Metal oxides are important soil components, and they often dominate the clay fraction of highly weathered soils in tropical and sub-tropical regions. Iron oxides are the most abundant and frequently occurring metal oxides in soil environments, and they exert a significant influence on the chemical, physical and microbial properties of soils. The structure and formation of metal oxides, which exist in small amounts in soils (e.g., Mn and Ti oxides) have not been studied adequately and remain poorly understood.

Metal oxides, particularly Fe, Al and Mn oxides, exist in poorly crystalline to well crystalline forms that range from nano- to micro-sized particles in soils. These oxides contribute significantly to the sorption of nutrients and contaminants in soil environments. In the last two decades the role of Fe and Al oxides in the preservation of soil organic carbon (SOC) has been recognized, however, the actual mechanisms and stability of the SOC associated with oxides are not well understood. Soil microbes are involved in the formation and dissolution of some of the Mn and Fe oxides, and many of the processes involved in their crystallization and dissolution are not well understood in soil environments, particularly under field conditions.

There is a good understanding of the sorption behavior and mechanisms of ions of nutrient elements and contaminants on the synthetic or pure metal oxides, however, limited research has been done on the interactive behavior of metal oxides in soil systems involving sorption of ions that exist in mixtures. Thus, the modelling predictions in the soil systems remain uncertain.

References

- Bandfield JF and Nealson KH (eds.) (1998) Geomicrobiology: Interactions between Microbes and Minerals. In: *Reviews in Mineralogy*, vol. 35. Washington, D.C.: Mineralogical Society of America.
- Bargar JR, Fuller CC, Marcus MA, Brearley AJ, Perez De la Rosa M, Webb SM, and Caldwell WA (2009) Structural characterization of terrestrial microbial Mn oxides from Pinal Creek, AZ. *Geochimica et Cosmochimica Acta* 73: 889–910.
- Bibi I, Singh B, and Silvester E (2011) Akaganéite (b-FeOOH) precipitation in inland acid sulfate soils of south-western New South Wales (NSW), Australia. *Geochimica et Cosmochimica Acta* 75: 6429–6438. <https://doi.org/10.1016/j.gca.2011.08.019>.
- Bigham JM, Fitzpatrick RW, and Schulze DG (2002) Iron oxides. In: Dixon JB and Schulze DG (eds.) *Soil Mineralogy with Environmental Applications*, pp. 323–366. Madison: Soil Science Society of America.
- Bigham JM, Schwertmann U, Carlson L, and Murad E (1990) A poorly crystallized oxyhydroxysulfate of iron formed by bacterial oxidation of Fe(II) in acid mine waters. *Geochimica et Cosmochimica Acta* 54: 2743–2758.
- Cornell RM and Schwertmann U (2003) *The iron oxides: Structure, properties, reactions, occurrences and uses*. Weinheim: Wiley-VCH Verlag GmbH & CoKGa.
- Coward EK, Thompson AT, and Plante AF (2017) Iron-mediated mineralogical control of organic matter accumulation in tropical soils. *Geoderma* 306: 206–216.
- Fischer WR and Schwertmann U (1975) The formation of hematite from amorphous iron (III) hydroxide. *Clays and Clay Minerals* 23: 33–37.
- Fitzpatrick RW and Chittieborough DJ (2002) Titanium and zirconium minerals. In: Dixon JB and Schulze DG (eds.) *Soil Mineralogy with Environmental Applications*. Madison: Soil Science Society of America.
- Funnell NP, Fulford MF, Inoue S, Kletetschka K, Michel FM, and Goodwin AL (2020) Nanocomposite structure of two-line ferrihydrite powder from total scattering. *Communications Chemistry* 3.
- Gilbert B, Erbs JJ, Penn RL, Petkov V, Spagnoli D, and Waychunas GA (2013) A disordered nanoparticle model for 6-line ferrihydrite. *American Mineralogist* 98: 1465–1476.
- Ginder-Vogel M and Sparks DL (2010) The impacts of X-ray absorption spectroscopy on understanding soil processes and reaction mechanisms. In: Singh B and Gräfe M (eds.) *Synchrotron-Based Techniques in Soils and Sediments*, pp. 1–26. Burlington: Elsevier.
- Grangeon S, Bataillard P, and Coussy S (2020) The nature of manganese oxides in soils and their role as scavengers of trace elements: Implication for soil remediation. In: van Hullebusch E, Huguenot D, Pechaud Y, Simonnot MO, and Colombano S (eds.) *Environmental Soil Remediation and Rehabilitation*, pp. 399–429. Cham: Springer.
- Heaney PJ, Prewitt CT, and Gibbs GV (eds.) (1994) *Silica: physical behavior, geochemistry and materials applications*. In: *Reviews in Mineralogy*, vol. 29. Washington, D.C.: Mineralogical Society of America.
- Heckman K, Lawrence CR, and Harden JW (2018) A sequential selective dissolution method to quantify storage and stability of organic carbon associated with Al and Fe hydroxide phases. *Geoderma* 312: 24–35.
- Hsu PA (1989) Aluminum oxides and oxyhydroxides. In: Dixon JB and Weed SB (eds.) *Minerals in Soil Environments*, pp. 331–378. Madison: Soil Science Society of America.
- Huang PM, Wang MK, Kämpf N, and Schulze DG (2002) Aluminum hydroxides. In: Dixon JB and Schulze DG (eds.) *Soil Mineralogy with Environmental Applications*, pp. 261–289. Madison: Soil Science Society of America.
- Iltona ES, Post JEB, Heaney PJ, Ling FT, and Kerisit SN (2016) XPS determination of Mn oxidation states in Mn (hydr)oxides. *Applied Surface Science* 366: 475–485.
- Kaiser K and Guggenberger G (2007) Sorption stabilization of organic matter by microporous goethite: Sorption into small pores vs. surface complexation. *European Journal of Soil Science* 58: 45–59.
- Kämpf N, Scheinost AC, and Schulze DG (2000) Oxide minerals in Soils. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. F125–F168. Boca Raton: CRC Press.
- Kaur N, Gräfe M, Singh B, and Kennedy BJ (2009a) Simultaneous incorporation of Cr, Zn, Cd, and Pb in the goethite structure. *Clays Clay Minerals* 57: 234–250. <https://doi.org/10.1346/CCMN.2009.0570210>.
- Kaur N, Singh B, and Kennedy BJ (2009b) Copper substitution alone and in the presence of chromium, zinc, cadmium and lead in goethite (α -FeOOH). *Clay Minerals* 44: 293–309.
- Kirsten M, Mikutta R, Vogel C, Thompson A, Mueller CW, Kimaro DN, Bergsma HLT, Feger K-H, and Kalbitz K (2021) Iron oxides and aluminous clays selectively control soil carbon storage and stability in the humid tropics. *Scientific Reports* 11(5076): 1–12. <https://doi.org/10.1038/s41598-021-84777-7>.

- Kleber M, Eusterhues K, Keiluweit M, Mikutta C, Mikutta R, and Nico PS (2015) Mineral–organic associations: Formation, properties, and relevance in soil environments. In: Sparks DL (ed.) *Advances in Agronomy*. vol. 130, pp. 1–140. Amsterdam: Elsevier.
- Monger HC and Kelly EF (2002) Silica minerals. In: Dixon JB and Schulze DG (eds.) *Soil Mineralogy with Environmental Applications*, pp. 611–636. Madison: Soil Science Society of America.
- Post JE, Heaney PJ, and Hanson J (2002) Rietveld refinement of a triclinic structure for synthetic Na-birnessite using synchrotron powder diffraction data. *Powder Diffraction* 17(3): 218–221.
- Post JE, McKeown DA, and Heaney PJ (2020) Raman spectroscopy study of manganese oxides—Tunnel structures. *American Mineralogist* 105: 1175–1190.
- Post JE, McKeown DA, and Heaney PJ (2021) Raman spectroscopy study of manganese oxides: Layer structures. *American Mineralogist* 106: 351–366.
- Scheinost AC and Schwertmann U (1999) Color identification of iron oxides and hydroxysulfates - Use and limitations. *Soil Science Society of America Journal* 63: 1463–1471.
- Schwertmann U and Murad E (1983) Effect of pH on the formation of goethite and hematite from ferrihydrite. *Clays and Clay Minerals* 31: 277–284.
- Schwertmann U and Taylor RM (1989) Iron oxides. In: Dixon JB and Weed SB (eds.) *Minerals in Soil Environments*, second ed., pp. 379–438. Madison: Soil Science Society of America.
- Silvester E, Manceau A, and Drits VA (1997) Structure of synthetic monoclinic Na rich birnessite and hexagonal birnessite: II. Results from chemical studies and EXAFS spectroscopy. *American Mineralogist* 82: 962–978.
- Singh B and Gilkes RJ (1992) Properties and distribution of iron oxides and their association with minor elements in the soils of south-western Australia. *Journal of Soil Science* 43: 77–98. <https://doi.org/10.1111/j.1365-2389.1992.tb00121.x>.
- Singh B, Gräfe G, Kaur N, and Liese A (2010) Applications of synchrotron-based X-ray diffraction and X-ray absorption spectroscopy to the understanding of poorly crystalline and metal-substituted iron oxides. In: Singh B and Gräfe M (eds.) *Synchrotron-Based Techniques in Soils and Sediments*, pp. 199–254. Burlington: Elsevier.
- Vu HP, Shaw S, and Benning LG (2008) Transformation of ferrihydrite to hematite: an in situ investigation on the kinetics and mechanisms. *Mineralogical Magazine* 72: 217–220.
- Waychunas GA (1991) Crystal chemistry of oxides and oxyhydroxides. In: Lindsley DH (ed.) *Petrologic and magnetic significance. Reviews in Mineralogy*, vol. 25, pp. 11–68. Washington, D.C.: Mineralogical Society of America.
- Waychunas GA, Kim CS, and Banfield JF (2005) Nanoparticulate iron oxide minerals in soils and sediments: unique properties and contaminant scavenging mechanisms. *Journal of Nanoparticle Research* 7: 409–433.

Soil morphology

Phillip R Owens and Zamir Libohova, USDA-ARS Dale Bumpers Small Farms Research Center, Booneville, AR, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of P.R. Owens, E.M. Rutledge, MORPHOLOGY, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 511-520, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00002-3>.

Introduction	150
Site characterization	150
Horizons	150
Transitional and combination horizons	151
Soil texture	153
Soil color	153
Redoximorphic features and mottles	155
Redoximorphic features	155
Mottles	157
Soil structure	157
Shape	159
Size	159
Grade	161
Compound structure	161
Soil consistence	161
Roots and pores	161
Chemical response	161
pH	161
Effervescence	162
Reaction to alpha,alpha'-dipyridyl	162
Special features	163
Pedon features	163
Ped surface features	163
Location	164
Amount	164
Horizon or layer features	164
Boundary	164
Proximal sensing	166
Digital soil profile description (digital soil morphometrics)	166
Soil micromorphological description	166
Conclusions	166
References	169

Abstract

Soil morphology focuses on standardizing descriptions of soil properties and features in the field. Although often qualitative and empirical, these descriptions aid in the interpretation of soil properties measured in the laboratory through chemical, biological and physical methods. Soil morphology is also one of the foundations for soil taxonomy. In the past, soil morphology has relied mostly on human sensors to detect features like color, consistency, texture, horizons, roots, pores, etc. Developments in digital sensors, remote and proximal, now offer more sophisticated tools to describe soils. In this chapter, we provide brief descriptions of the morphological concepts and tools used by soil scientists.

Key points

- Soil morphology is used to describe soil properties and features in the field using simple tools and preceptory human sensors.
- Soil micromorphology uses tools such as microscopes to characterize features like pores, mineralogy, etc.
- Proximal sensors and soil digital morphometry bring together morphology, micromorphology and analytical chemical, physical and biological methods using software and machine learning techniques.

Introduction

Soil morphology is defined as the branch of soil science that deals with the description of in situ spatial organization and physical properties of soil regardless of potential applications (Wilding et al., 1983). Precise descriptions that use standard terminology are needed in all areas of soil science. The basic terminology has been developed over the past 50 years and continues to expand. Prior to the use of standard terminology, soils were described as clayey, sandy, stony, sedimentary, saline, marshy, dry or moist, heavy or light, soft or compact, fatty, friable, or lean. Soil horizons were described by terms such as 'gray watery sand' or 'rusty brown clay' (Buol et al., 1997; Soil Science Division Staff, 2017). While these terms illustrate a fundamental understanding of soil properties, they do not impart any specific knowledge about the soils in question and cannot be compared to descriptions of other soils by other scientists. An objective, complete description of the soil is essential because it serves as a basis for soil identification, classification, correlation, mapping, communication, and interpretation (Soil Survey Staff, 2010; U.S. Department of Agriculture, Natural Resources Conservation Service, 2022).

The basic tools used by field soil morphologists are very simple and usually include a knife, tape measure, water bottle, Munsell color book, 10× hand lens, 10% HCl (hydrochloric acid), a clinometer or Abney level, and often GPS to locate the sites in the geographic domain. The development of proximal sensors has provided additional tools for soil scientists in the field that has led to the emergence of digital soil morphometrics (Hartemink and Minasny, 2014). One such tool is a handheld portable X-ray diffractometer (pXRF), used to map the distribution of metals in the soil profile. Another sub-discipline of soil science, digital soil morphometrics, is adding to soil morphologic techniques by using multiple proximal sensing techniques such as pXRF, image acquisition and processing of soil profiles (Hartemink and Minasny, 2014).

Soil scientists describe soil profiles, vertical cross-sections of the soil representative of a three-dimensional pedon. A pedon is the smallest volume that is recognized as a soil, and it is best to determine field morphology of a soil profile using a recently opened pit (or a fresh road cutting) (Soil Science Division Staff, 2017). The soil scientist begins by organizing and subdividing the soil profile into reasonably distinct layers or pedogenic horizons. A soil description includes a site characterization and horizon depth intervals, horizon boundary characteristics, color, texture, structure, consistence, roots and pores, pH, effervescence, and special features of each horizon (Soil Science Division Staff, 2017).

Site characterization

The soil pedon features at a site result from the combined action of five soil-forming factors: parent material, biota, climate, topography, and time (Jenny, 1941). It is important to identify the parent material because some soil properties are inherited (Brewer, 1976). Knowledge of parent material combined with the observable features in the landscape conveys information about the formation and expected behavior of a soil. Descriptions need to be as precise as possible, including the geomorphic information, physiographic information, and surface morphometry. The physiographic information includes: (1) physiographic division, (2) physiographic province, (3) physiographic section, (4) state physiographic area, and (5) local physiographic or geographic name. The geomorphic information includes: (1) landscape, (2) landform, (3) microfeatures, and (4) anthropogenic features. The surface morphometry information includes: (1) elevation, (2) slope aspect, (3) slope gradient, (4) slope complexity, (5) slope shape, (6) geomorphic components, and (7) microrelief. The scientist describes the vegetation type, listing the common names of plants and their scientific names, when possible. Latitude and longitude, in degrees, seconds, and minutes, are most commonly used to record the geographical location of the soil as precisely as possible. Site characterization is organized as a hierarchy, typically starting with general setting of the site (physiography, landform, landform position, microfeatures, exposure, etc.) followed by a detailed description of the soil pedon or profile. The order may change depending on the local survey conditions. The hierarchical approach provides a better understanding of the observed pedon and horizon features and also connections with geographic, climatic, physiographic, and landscape settings (Schoeneberger et al., 2012).

Horizons

Soil horizons may be the most distinctive visual features of a soil profile. Both the *Soil Survey Manual* (Soil Science Division Staff, 2017) and World Reference Base for Soil Resources (IUSS Working Group WRB, 2015), refer to horizon(s) as a layer(s) approximately parallel to the surface of the soil, distinguishable from adjacent layers by a distinctive set of properties produced by the soil-forming processes. In the case of US Soil Taxonomy (Soil Survey Staff, 2010) the term 'layer,' rather than 'horizon,' is used if all the properties are believed to be inherited from the parent material. For example, recent loess or alluvial deposits that have not undergone in situ soil development or pedogenesis are considered soil layers. Horizons are typically described in two dimensions, vertically and horizontally.

The soil scientist first determines the horizon designation (label) and records the horizon thickness. Horizons are labeled according to diagnostic features (i.e., color, structure, texture, rock fragment content, etc.) and interpretations (i.e., how the horizon formed and how it responds to anthropogenic influence). The master horizon label represents characteristic properties. In the classification system developed by the U.S. Department of Agriculture (USDA) (Soil Survey Staff, 2010) capital letters (O, A, E, B, C, R, and V) indicate master horizons. Similar terminology for naming the main horizons is also used by the World Reference Base for

Soil Resources (IUSS Working Group WRB., 2015). As an example, the designations according to the U.S Soil Taxonomy (Soil Survey Staff, 2010) are as follows:

horizons and layers—Horizons or layers dominated by organic material. Some are saturated with water for long periods of time or were once saturated but are now artificially drained; others have never been saturated.

A horizons—Mineral horizons that formed at the surface or below an O horizon which exhibit obliteration of all or much of the original rock structure and that show one or more of the following: (1) an accumulation of humified organic matter intimately mixed with the mineral fraction and not dominated by properties characteristic of E or B horizons, or (2) properties resulting from cultivation, pastures, or similar kinds of disturbance.

E horizons—Mineral horizons in which the main feature is loss of silicate clay, iron, aluminum, or some combination of these, resulting in a concentration of sand and silt particles.

B horizons—Horizons that formed below an A, E, or O horizon, have soil structure, and show one or more of the following: (1) illuvial concentration of silicate clay, iron, aluminum, humus, carbonates, gypsum, or silica, alone or in combination; (2) evidence of the removal of carbonates; (3) residual concentration of sesquioxides; (4) coatings of sesquioxides that make the horizon color conspicuously lower in value, higher in chroma, or redder in hue than overlying and underlying horizons without apparent illuviation of iron; (5) alteration that forms granular, blocky, or prismatic structure if volume changes accompany changes in moisture content; or (6) brittleness.

C horizons or layers—Horizons or layers, excluding hard bedrock, that are little affected by pedogenic processes and lack properties of O, A, E, and B horizons. The material of C layers may be either like or unlike that from which the solum (horizons above the C) presumably formed.

R layers—Bedrock or strongly cemented to indurated. Examples are granite, basalt, quartzite, limestone, and sandstone. These layers are difficult to excavate and dig with handheld tools like spades but can be destroyed with heavy equipment.

V horizons—Mineral horizons that formed at the soil surface or below a layer of rock fragments (e.g., desert pavement), a physical or biological crust, or recently deposited eolian material.

Transitional and combination horizons

Horizons may result from two master horizons or layers transitioning into each other or contain portions of two distinct master horizons. Transitional soil horizons may be dominated by properties of one master horizon but have subordinate properties of another; for these horizons, symbols of master horizons may be combined (i.e., AB, EB, CB, BC, etc.). Combination horizons have materials that can be described separately regarding color, texture, and structure; a virgule (/) is used in their symbol to indicate the two source horizons (i.e., Bw/A, Bt/A, C/Bt, etc.).

Designations of master soil horizons are supplemented with suffixes or subscripts to convey more information about formation processes. Description systems vary slightly. Table 1 lists the suffixes used by the USDA (Soil Survey Staff, 2010). The World Reference Base for Soil Resources (IUSS Working Group WRB., 2015) uses similar qualifiers to describe different horizons and their associated pedogenetic processes.

Table 1 Suffix symbols used by the USDA for master horizons.

<i>Suffix</i>	<i>Horizon</i>
a	Highly decomposed organic material
b	Buried genetic horizon
c	Concretions or nodules
co	Coprogenous earth
d	Physical root restriction
di	Diatomaceous earth
e	Organic material of intermediate decomposition
f	Frozen soil or water
ff	Dry permafrost
g	Strong gleying
h	Illuvial accumulation of organic matter
i	Slightly decomposed organic material
j	Accumulation of jarosite
jj	Evidence of cryoturbation
k	Accumulation of carbonates
kk	Engulfment of horizon by secondary carbonates
m	Cementation or induration
ma	Marl
n	Accumulation of sodium

(Continued)

Table 1 (Continued)

<i>Suffix</i>	<i>Horizon</i>
o	Residual accumulation of sesquioxides
p	Tillage or other disturbance
q	Accumulation of silica
r	Weathered or soft bedrock
s	Illuvial accumulation of sesquioxides and organic matter
se	Presence of sulfides
ss	Presence of slickensides
t	Accumulation of silicate clay
u	Presence of human-manufactured materials (artifacts)
v	Plinthite
w	Development of color or structure
x	Fragipan
y	Accumulation of gypsum
yy	Dominance of horizon by gypsum
z	Accumulation of salts more soluble than gypsum
Suffix	Horizon
a	Highly decomposed organic material
b	Buried genetic horizon
c	Concretions or nodules
co	Coprogenous earth
d	Physical root restriction
di	Diatomaceous earth
e	Organic material of intermediate decomposition
f	Frozen soil or water
ff	Dry permafrost
g	Strong gleying
h	Illuvial accumulation of organic matter
i	Slightly decomposed organic material
j	Accumulation of jarosite
jj	Evidence of cryoturbation
k	Accumulation of carbonates
kk	Engulfment of horizon by secondary carbonates
m	Cementation or induration
ma	Marl
n	Accumulation of sodium
o	Residual accumulation of sesquioxides
p	Tillage or other disturbance
q	Accumulation of silica
r	Weathered or soft bedrock
s	Illuvial accumulation of sesquioxides and organic matter
se	Presence of sulfides
ss	Presence of slickensides
t	Accumulation of silicate clay
u	Presence of human-manufactured materials (artifacts)
v	Plinthite
w	Development of color or structure
x	Fragipan
y	Accumulation of gypsum
yy	Dominance of horizon by gypsum
z	Accumulation of salts more soluble than gypsum

Reproduced from Soil Survey Staff (2010) *Keys to Soil Taxonomy*, 11th edn. USDA Natural Resources Conservation Service. Washington, DC: U.S. Government Printing Office.

Types of soil horizons and layers were developed from soil formation and pedogenic concepts based on field investigations and laboratory analysis. Properties can vary greatly between soil types and horizons, with varying degrees of continuity that can be described as very abrupt, abrupt, clear, gradual, or diffuse (Soil Survey Staff, 2010). There are no hard rules in labeling horizons. Labeling is an interpretation based on observations by the scientist. The classification of the soil, however, is based on the description and measured soil properties and follows strict criteria.

Soil texture

Soil texture is one of the most common features used by scientists and laymen to describe soils. Texture refers to the relative abundance, by weight, of the size fractions—sand, silt, and clay. Particles less than or equal to 2 mm make up the ‘fine-earth fraction.’ This is one of the most important fractions for determining chemical, biological and some physical properties (Soil Survey Staff, 2014). In all textural classification systems, the sum of all these particles (by weight) equals 100% (Fig. 1). The several textural classification systems have varying ranges of particles sizes (Fig. 2).

Textural descriptions such as gritty, smooth, and sticky are commonly used because determining the percentage of sand, silt, and clay can be cumbersome. Commonly, field estimates of sand, silt, and clay are used, and the soils are placed into textural classes. These textures are verified by laboratory analysis. The mineralogy of the clay may affect estimates of clay in field textures. The kaolinitic clays expand little and tend to be harder to discern than smectitic clays by ‘feel texture,’ resulting in an underestimate of the clay percentage.

Soil texture relates to many of the ways a soil performs. If a soil is coarse (sandy), water tends to move through it without much impediment and the soil may not retain adequate amounts of water for plant growth. If a soil is clayey, water will probably move more slowly and be retained for plant growth; however, not all the water may be available or easily accessed by plant roots because clay can bond to the water. The shrink–swell potential of clayey soils varies depending on the type of clay minerals. If the soil is high in smectitic clay minerals, it will shrink and swell as the moisture content of the soil changes. This shrinking and swelling must be considered in the designs of infrastructures, such as roads and foundations. Clayey soils that primarily consist of kaolinite and other low-activity minerals have fewer limitations on designs. If more than 15%, by volume, of the soil consists of particles larger than 2 mm, prefixes or modifiers are added to the textural classes to indicate the amounts and type of fragments. These separates are classed as gravel, cobbles, stones, or boulders (Fig. 2).

Soil color

Soil colors can infer pedogenic processes in soils. The main pigmenting agents in soils are organic matter, iron, and, to a lesser extent, manganese and sulfur compounds; they cover the natural color of mineral grains (most of which are gray). Fig. 3 shows two

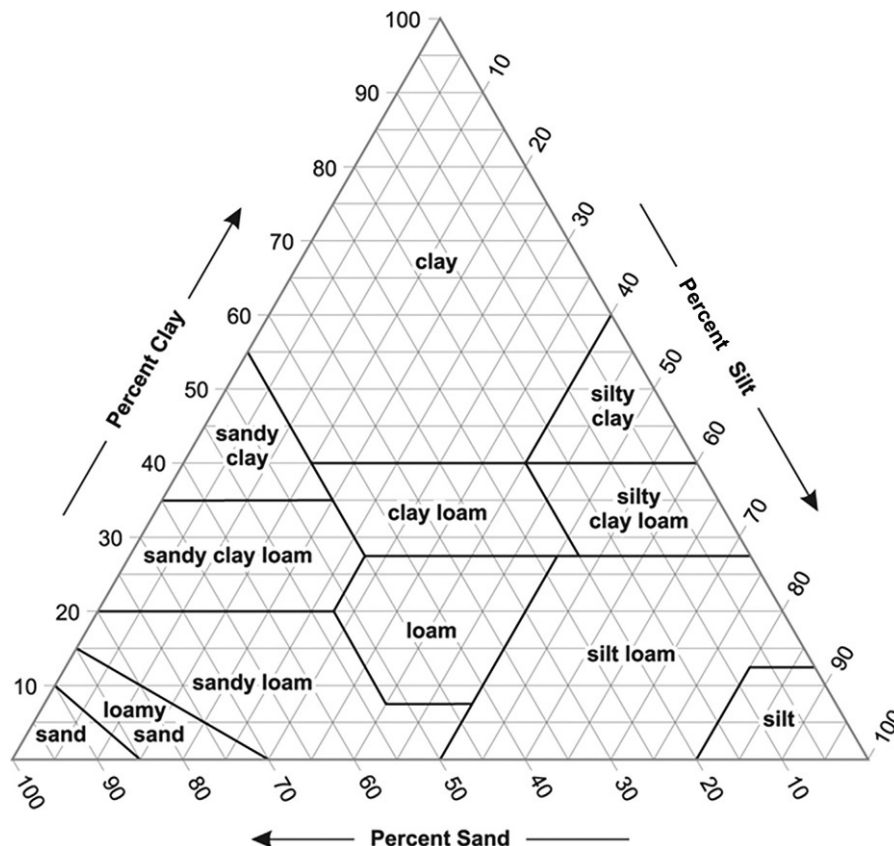


Fig. 1 The USDA soil textural triangle showing the percentages of clay, silt, and sand in the 12 basic texture classes. Reproduced from Soil Science Division Staff (2017) Soil Survey Manual. Ditzler C, Scheffe K, and Monger HC (eds) *USDA Handbook 18*. Washington, DC: Government Printing Office.

Comparison of Particle Size Classes in Different Systems

USDA ¹	FINE EARTH										ROCK FRAGMENTS													
	Clay ²		Silt		Sand						channers			6" ¹⁵⁰	15" ³⁸⁰	24" ⁶⁰⁰ mm	boulders							
											flagst	stones	stones	boulders										
	fine	co.	fine	co.	v.fi.	fi.	med.	co.	v.co.	fine	medium	coarse	Cob- bles	Stones	Boulders									
millimeters:	0.0002	.002 mm	.02	.05	.1	.25	.5	1	2 mm	5	20	76	250 mm	600 mm										
U.S. Standard Sieve No. (opening):					300 ³	140	60	35	18	10	4	(3/4")	(3")	(10")	(25")									
Inter- national ⁴	Clay	Silt	Sand						Gravel	Stones														
			fine	coarse																				
	millimeters:	.002 mm	.02	.20	2 mm	20 mm																		
U.S. Standard Sieve No. (opening):				10	(3/4")																			
Unified ⁵	Silt or Clay				Sand			Gravel		Cobbles	Boulders													
					fine	medium	co.	fine	coarse															
	millimeters:		.074	.42	2 mm	4.8	19	76	300 mm															
U.S. Standard Sieve No. (opening):		200	40	10	4	(3/4")	(3")																	
AASHTO ^{6, 7}	Clay	Silt	Sand		Gravel or Stones			Broken Rock (angular), or Boulders (rounded)																
			fine	coarse	fine	med.	co.																	
	millimeters:	.005 mm	.074	.42	2 mm	9.5	25	75 mm																
U.S. Standard Sieve No. (opening):		200	40	10	(3/8")	(1")	(3")																	
phi #:	12	10	9	8	7	6	5	4	3	2	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-12	
Modified Wentworth ⁸																								
millimeters:	.00025	.002	.004	.008	.016	.031	.062	.125	.25	.5	1	2	4	8	16	32	64	128	256	4092 mm				
U.S. Standard Sieve No.:																								

Fig. 2 Particle-size classes according to five particle-size class systems. Reproduced from Schoeneberger PJ, Wysocki DA, Benham EC, and Soil Survey Staff (2012) *Field Book for Describing and Sampling Soils, Version 3.0*. USDA Natural Resources Conservation Service, National Soil Survey Center: Lincoln, NE. Fi.: fine; co.: coarse; v.Fi.: very fine; v.Co.: very coarse; med.: medium.



Fig. 3 A well-developed Minnieville soil (Prince William County, Virginia) in a well-drained convex position and a poorly drained Coxville soil (Greenville, North Carolina) in a concave position. Modified from the John A. Kelley Collection, USDA Natural Resources Conservation Service (<https://www.flickr.com/photos/soilscience/albums>). Shared according to Creative Commons License terms at <https://creativecommons.org/licenses/by/2.0/>.

soils that have similar texture and structure yet differ in color. If a soil horizon has more than one color, the dominant color, by volume, is considered the matrix color.

Soil colors can be conveniently identified using a soil color chart. Soil scientists generally use Munsell color charts (Munsell Soil Color Charts, 1994). The charts arrange color by hue, value, and chroma. A combination of these three variables is used for all colors. Hue is the dominant spectral (rainbow) color; it is related to the dominant wavelength of the light. Value refers to the relative lightness of the color; it is a function (approximately the square root) of the total amount of light. Chroma is the relative purity or strength of the spectral color; it increases in value as grayness decreases.

Each soil color chart represents a specific hue, indicated in the upper right-hand corner (Fig. 4). The colors become successively lighter from bottom to top as the value increases. They increase in chroma to the right and become grayer to the left. The value and

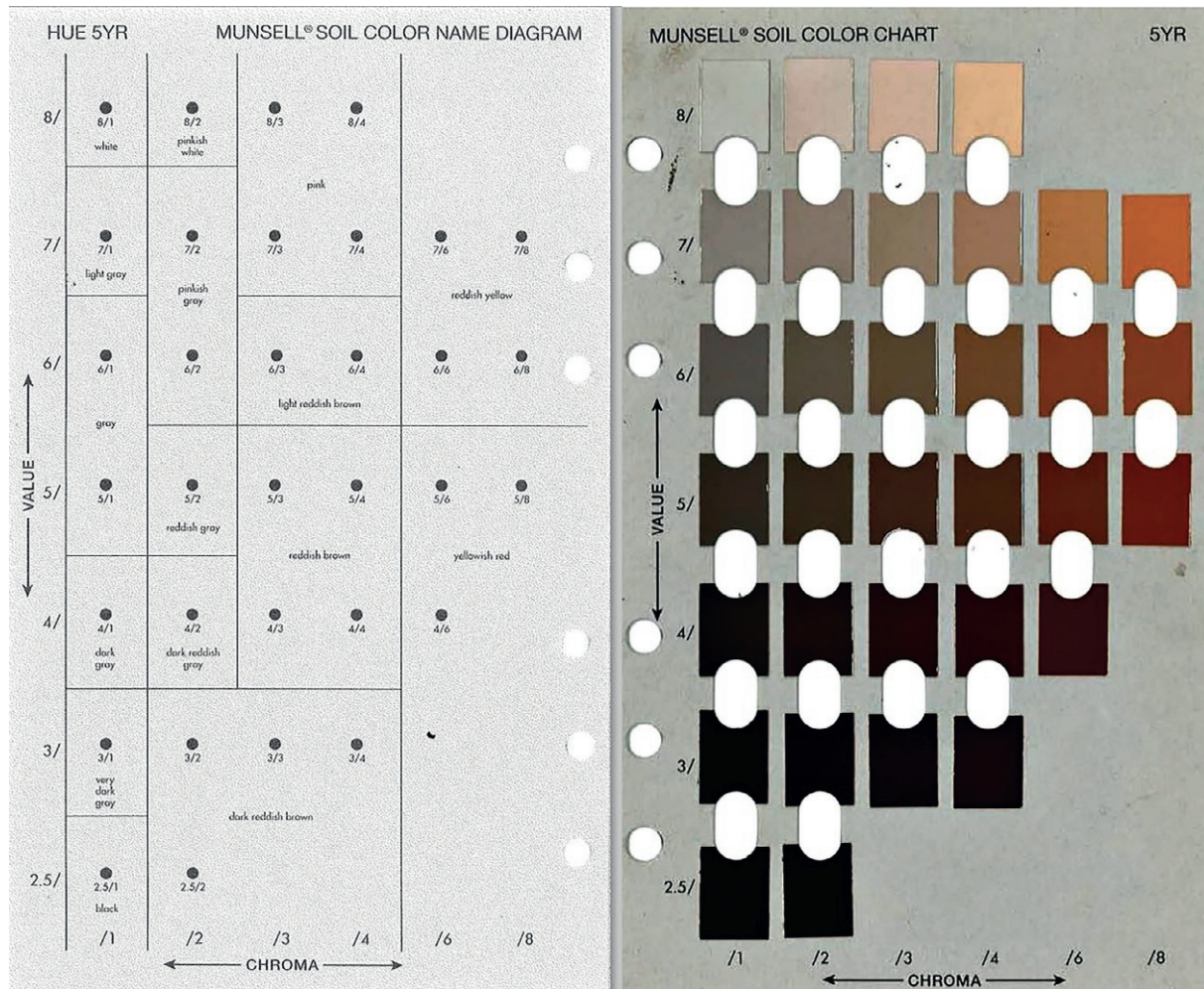


Fig. 4 A Munsell soil color chart for hue 2.5YR, value 2.5–8, and chroma 1–8. Reproduced from Munsell Soil Color Charts (1994) Revised Edition, Macbeth Division of Kollmorgen Instruments Corporation, 405 Little Britain road, New Windsor, NY, 12553.

chroma of each color are shown beneath the color chip. As arranged in the chart, the colors form three scales: (1) radial, or from one card to the next, in hue; (2) vertical in value; and (3) horizontal in chroma (Fig. 5).

The matrix color (the dominant color) is the first color described in a soil horizon or layer. The moisture status (moist or dry) of the soil must be considered when describing soil color. When determining soil color, a soil scientist uses a fresh ped interior, to eliminate the possibility of describing the color of a ped surface coating (Fig. 6). While the Munsell color system continues to be widely used in the field, the use of spectral sensors has led to the use of other colors schemes like CIE (Commission Internationale de l'Eclairage—CIE | International Commission on Illumination/Comission internationale de l'Eclairage/Internationale Beleuchtungskommission), XYZ (Red, Green, Blue—RGB) color space (CIE, 1932). The CIE color space links the wavelength in the visible spectrum quantitatively with color visions (like Munsell color charts) perceived by humans. These links allow for quantitative analysis of soil colors in the visible range which is very useful in digital soil morphometrics (Hartemink and Minasny, 2014).

Redoximorphic features and mottles

A lack of uniformity in soil color can provide insight into the pedogenesis of the soil. In the USDA description terminology, the generic term 'mottle' was changed to 'redoximorphic feature' to describe soil features resulting from reduction and oxidation processes. The term 'redoximorphic feature' better describes the drainage status of the soil. The term 'mottles' does not infer a pedogenic process.

Redoximorphic features

Soil colors are greatly affected by the amount of oxygen in the soil. If there is sufficient oxygen for microbial activity, the soil is typically brown or red and frequently has a single dominant color. However, when the soil becomes wet (saturated with water), the

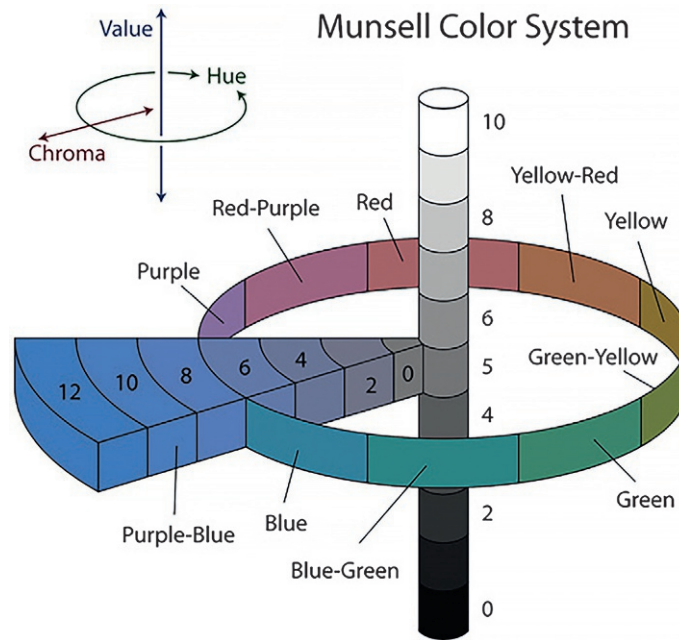


Fig. 5 A diagram showing relationships between hue, value, and chroma in the Munsell color system. Reproduced from <https://commons.wikimedia.org/wiki/File:Munsell-system.svg>. Shared according to Creative Commons Attribution ShareAlike 3.0 license (Accessed 26 January, 2022).



Fig. 6 A freshly broken ped with the exposed interior is held next to a Munsell soil color chart. Reproduced from John A. Kelley Collection, USDA Natural Resources Conservation Service (<https://www.flickr.com/photos/soilscience/albums>). Shared according to Creative Commons License terms (<https://creativecommons.org/licenses/by/2.0/>), 2009; 113: 6694–6698.

oxygen supply cannot flow as freely and may become insufficient for many microbes. Insufficient oxygen in the soil causes reduction—a process by which iron and manganese gain electrons. Reducing conditions in soils are precursors to redoximorphic features (Vepraskas, 1992). Reduction of iron and manganese greatly increases their solubility, allowing them to dissolve into solution and move with the water (i.e., become mobile). Soil features caused by reduction and subsequent reoxidation are also redoximorphic features.

When the iron in the soil dissolves in solution, the coating is removed from the surface of the mineral grains, resulting in zones called 'depletions.' The color of depletions is that of the mineral grains, usually gray. When air reenters the soil, the iron and

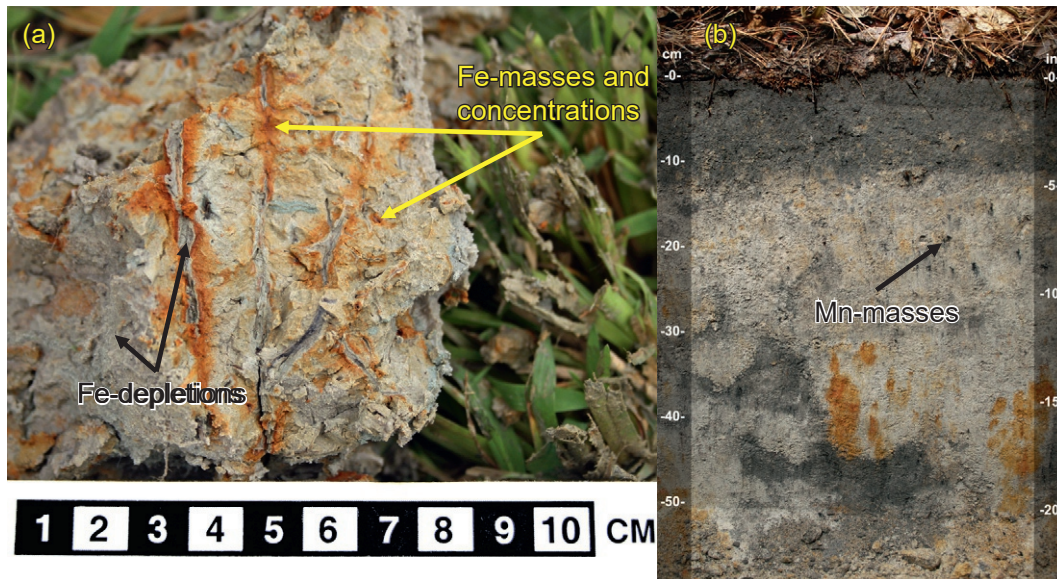


Fig. 7 (a) Soil with redoximorphic features associated with iron (Fe) depletions (gray colors) and iron concentrations (reddish colors), in the matrix and along soil pores. (b) Soil with manganese (Mn) concentrations (black colors). Modified from John A. Kelley Collection, USDA Natural Resources Conservation Service (<https://www.flickr.com/photos/soilscience/albums>). Shared according to Creative Commons License terms (<https://creativecommons.org/licenses/by/2.0/>).

manganese are oxidized and precipitate into brownish, reddish, or blackish forms called ‘concentrations.’ Thus, depletions are caused by removal of iron coatings, and concentrations are caused by oxidation and precipitation of previously reduced forms of iron and manganese. Both depletions and concentrations require reduction, but only concentrations require subsequent reoxidation.

Concentrations and depletions are described by kind, size, abundance, and color (Schoeneberger et al., 2012). Concentrations are usually browner, redder, or blacker than the matrix (Fig. 7). There are several kinds of concentrations, including finely disseminated materials, masses (non-cemented), nodules, and concretions (cemented). Concretions and nodules are both cemented accumulations of iron and manganese but differ in their internal structure: concretions have concentric rings whereas nodules do not (Fig. 7). Depletions are usually lighter colored than the matrix (Fig. 3, Coxville soil). There are two types of depletions: (1) iron depletions, which are primarily a lighter color than the matrix; and (2) clay depletions, which have the light gray color but also have lost clay and thus are grainier.

Mottles

Mottles are areas that have a different color from the matrix. They commonly have lithochromic or lithomorph colors. These features typically are inherited from the geologic parent material rather than the result of pedogenesis. Relict redoximorphic features associated with mottles are features that formed in a pluvial environment or in landscape that has been naturally or anthropogenically altered and permanently affects the drainage of the soil (Fig. 8). Description of mottles may include abundance, size, contrast (to matrix), color, and name.

Soil structure

Soil structure refers to units composed of primary particles. The cohesion within the units is greater than the adhesion among them. Soil structure results from processes that aggregate, cement, compact, or unconsolidated soil particles. Under stress, the soil mass tends to rupture along predetermined planes or zones. The peds form repetitive, quasi-uniform shapes varying in size from millimeters to centimeters and result from the aggregation of soil particles held together by the electrical charges on the surfaces of the minerals and organic matter (Fig. 9). An undisturbed soil has several levels of structure, which can be divided into primary, secondary (compound), and tertiary. Morphological descriptions identify the primary structure. Soil structure can be defined by: (1) the size, shape, and arrangement of peds, (2) the voids between peds, and (3) the associated interpedal pedological features in a soil material.

Clods and fragments differ from peds. Peds are a result of soil-forming processes, whereas clods and fragments are not. Clods result from a disturbance, such as plowing or digging that molds the soil to a transient mass which slakes with repeated wetting and drying. Fragments include: (1) units of undisturbed soil with bounding planes of weakness that are formed on drying without application of external force and which do not appear to have predetermined bounding planes; (2) units of soil disturbed by



Fig. 8 Mottles and relict redoximorphic features associated with different parent material deposits in alluvium and fluvial systems. Modified from the John A. Kelley Collection, USDA Natural Resources Conservation Service (<https://www.flickr.com/photos/soilscience/albums>). Shared according to Creative Commons License terms (<https://creativecommons.org/licenses/by/2.0/>).

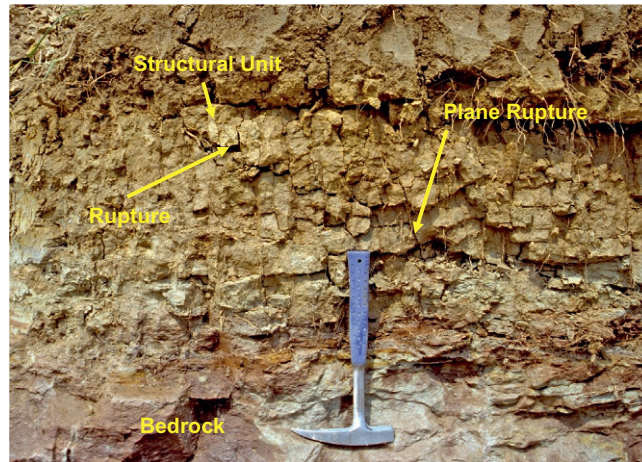


Fig. 9 A clay soil with repetitive blocky structural units (peds) showing ruptures along predetermined planes or zones. Modified from John A. Kelley Collection, USDA Natural Resources Conservation Service (<https://www.flickr.com/photos/soilscience/albums>). Shared according to Creative Commons License terms (<https://creativecommons.org/licenses/by/2.0/>).

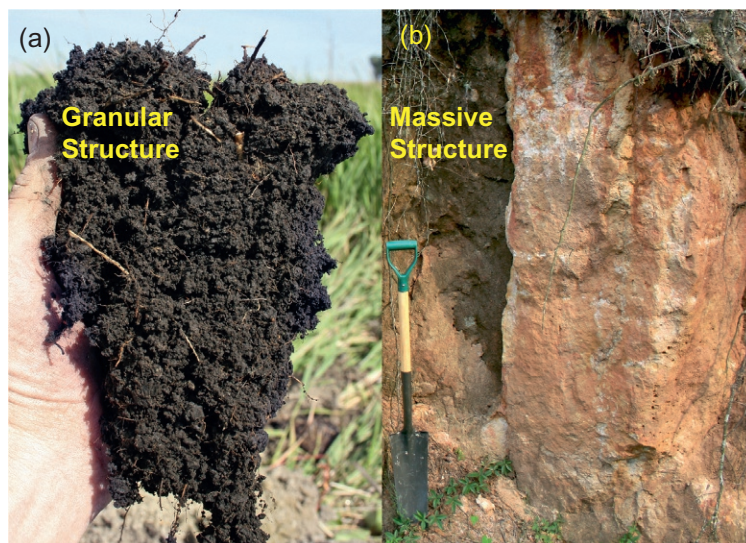


Fig. 10 Granular structure in a sample of a surface soil horizon (a) and a profile with massive structure (b). Modified from John A. Kelley Collection, USDA Natural Resources Conservation Service (<https://www.flickr.com/photos/soilscience/albums>). Shared according to Creative Commons License terms (<https://creativecommons.org/licenses/by/2.0/>).

mechanical means but without significant rearrangement to a denser configuration; and (3) pieces of soil bounded by planes of weakness caused by pressure exerted during examination, with size and shape highly dependent on the manner of manipulation.

Some soils lack structure and are referred to as structureless. When structureless soils are ruptured, soil fragments, single grains, or both may result. Structureless soil may be either 'massive' or 'single grain.' Massive soils tend to form fragments when disturbed, and single grain soils are typically sandy and loose in 50% or more of the volume (Fig. 10).

Shape

Several basic shapes are recognized for soil structural units. Supplemental statements about variations in shape of individual peds are needed in detailed descriptions of some soils. Table 2 and Fig. 11 identify the possible shapes of soil structure.

Size

Five classes are used to describe the size of structural units: very fine, fine, medium, coarse, and very coarse. The size limits vary according to the shape of the unit. They refer to the smallest dimension of plates, prisms, and columns.

Table 2 Types of soil structure.

Structure	Description
<i>Natural soil structural units (pedogenic structure)</i>	
Granular	Small peds with rounded or very irregular faces
Angular blocky	Peds with faces that intersect at sharp angles
Subangular blocky	Peds with sub-rounded shapes and no sharp angles
Platy	Flat, plate-shaped units
Wedge	Elliptical, interlocking lenses that terminate in acute angles and are bounded by slickensides
Prismatic	Vertically elongated peds with flat tops
Columnar	Vertically elongated peds with rounded tops
<i>Structureless</i>	
Single grain	Absence of structural units; entirely noncoherent material; e.g., loose sand
Massive	Absence of structural units; coherent mass of material

Reproduced from Soil Science Division Staff (2017) Soil Survey Manual. In: Ditzler C, Scheffe K, and Monger HC (eds) *USDA Handbook 18*. Washington, DC: Government Printing Office.

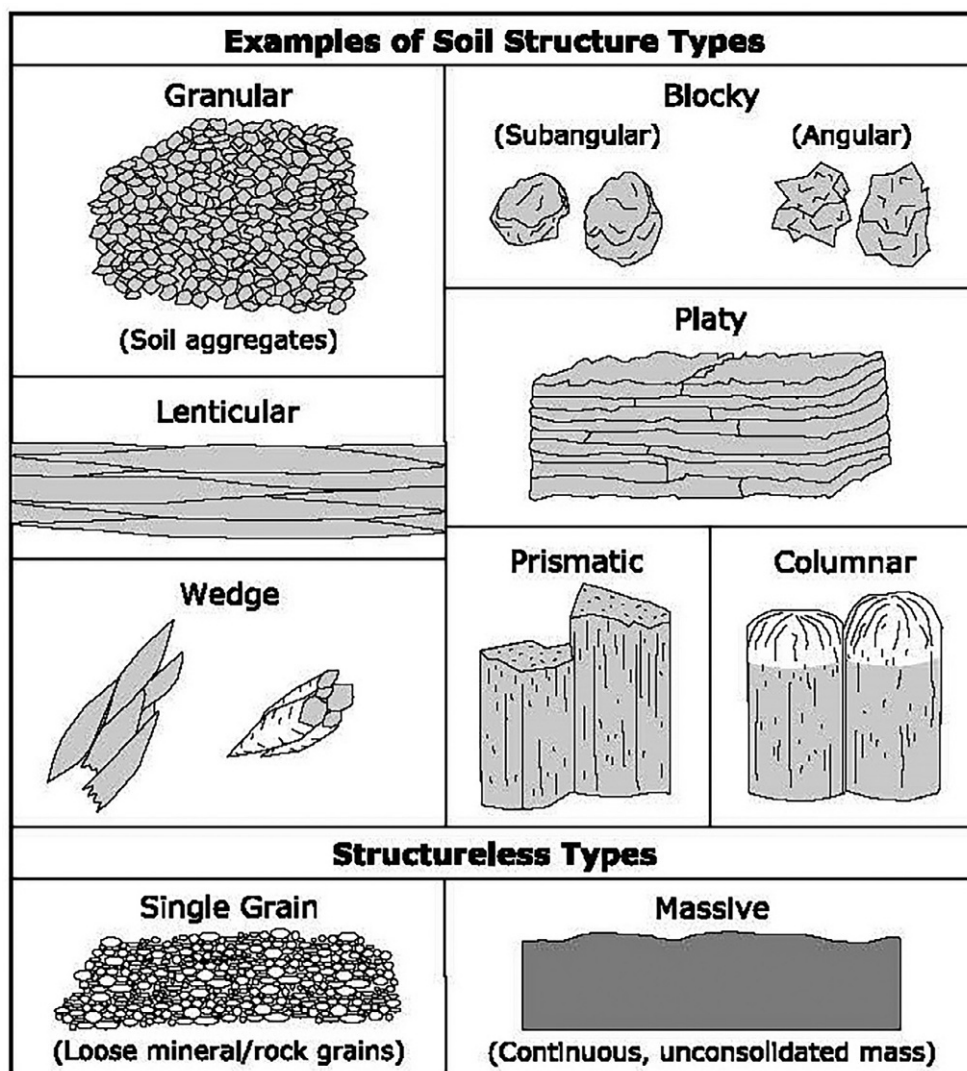


Fig. 11 Shapes of soil structure. Reproduced from Schoeneberger PJ, Wysocki DA, Benham EC, and Soil Survey Staff (2012) *Field Book for Describing and Sampling Soils, Version 3.0*. USDA Natural Resources Conservation Service, National Soil Survey Center: Lincoln, NE.

Grade

The distinctness of soil structural units is described by grades based on the relationship between cohesion within units to adhesion between units (i.e., the ease with which the units separate and the proportion of units that hold together when the soil is handled). There are three grade classes:

Weak—Units are barely observable in situ. When gently disturbed, the soil material parts into a mixture of whole and broken units and mainly exhibits no planes of weakness. Faces that indicate persistence through wet-dry cycles are evident.

Moderate—Units are well formed and evident in undisturbed soil. When disturbed, the soil material parts into a mixture of mostly whole units, some broken units, and material that is not units. Peds part from adjoining peds to reveal nearly entire faces that have properties distinct from those of fractured surfaces.

Strong—Units are distinct in undisturbed soil. They separate cleanly when the soil is disturbed. When removed, the soil material separates mainly into whole units.

The degree of disturbance required to determine structure grade depends largely on moisture content and percentage and mineralogy of clay. Slight disturbance only may be necessary to separate the units of moist sandy loam with strong granular structure, while considerable disturbance may be required to separate units of moist clay loam with strong blocky structure.

The designation of structure grade, size, and shape can be modified with other appropriate terms when necessary to describe other characteristics. Surface characteristics of units are described separately. Special structural units, such as wedge-shaped units of soils with a high shrink-swell potential, are described if present.

Compound structure

‘Compound structure,’ another term for ‘secondary structure,’ refers to smaller structural units that are held together to form larger units. Grade, size, and shape can be described for both sets of units as well as the relationship of one set to the other. Examples of compound structure are ‘strong medium blocks within moderate coarse prisms’ and ‘moderate coarse prismatic structure parting to strong medium blocky.’

Soil consistence

Consistence refers to the cohesion and adhesion within a soil or its resistance to deformation or rupture. Soil consistence largely depends on soil moisture. It can be expressed as rupture resistance, surface crust formation, manner of failure, stickiness, plasticity, and penetration resistance.

Consistence is typically described by field evaluation (i.e., rupture resistance). A soil sample 25–30 mm long is held between the thumb and first finger and pressure is applied. The moisture content of the sample is an important part of the calculation.

Soils in humid and subhumid climates may form brittle subsurface horizons. The degree and percentage of brittleness of the soil are described. Soils may also be hardened due to cementation. The accumulated materials, such as carbonates, gypsum, humus, iron, and or silica, are listed. In addition, the percentage of cementation and its strength are noted.

Roots and pores

Soils are penetrated by roots that grow underneath the surface that anchor plants while absorbing water and nutrient from the soil. The larger soil roots and pores can be observed in the field (Fig. 12). The larger pores (macropores?) are mostly noncapillary pores, i.e., pores that cannot hold water against the force of gravity. They are important for the movement of water and air within the soil. After a major wet event (rainfall, snowmelt, irrigation, etc.), these pores transmit the water throughout the soil to the smaller capillary pores. The noncapillary pores redistribute the water in a few days; as they drain, oxygen (air) is distributed into the soil. If soil oxygen is depleted for more than a few days, many plants will die or be harmed.

Soils with a mixture of noncapillary and capillary pores are best suited to plant growth because they both drain well and retain water. The roots and pores are described based on abundance and size classes (for example, common fine roots, many medium pores, etc.).

Chemical response

pH

Much information can be inferred from the acidity or alkalinity of a soil. Soil pH can be estimated with a simple field test kit. Each horizon or layer within a pedon is tested because pH may vary significantly due to pedogenic or lithogenic properties. There are several soil pH class systems that categorize soil pH.

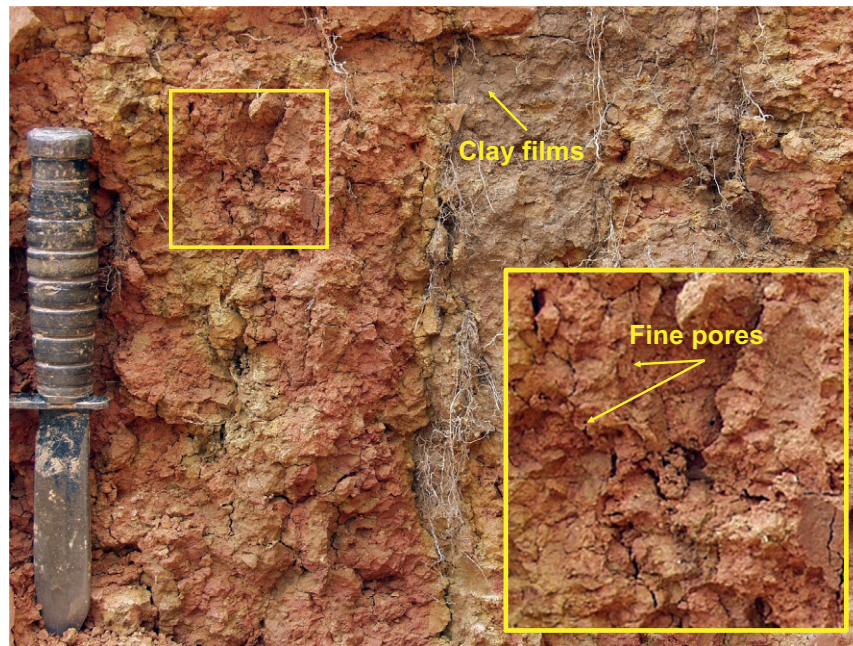


Fig. 12 A soil with root distribution showing small pores (enlarged inset) and clay films. Modified from John A. Kelley Collection, USDA Natural Resources Conservation Service (<https://www.flickr.com/photos/soilscience/albums>). Shared according to Creative Commons License terms (<https://creativecommons.org/licenses/by/2.0/>).

Effervescence

A simple field technique may be used to check for the presence of free carbonates. A 10% solution of HCl added to the soil produces a gaseous response when CaCO_3 (calcium carbonate) is present. The gaseous response will vary depending on the amount of carbonates in the soil. The degree of effervescence is noted in the description.

Reaction to alpha,alpha'-dipyridyl

In soils that are actively reducing at an Eh (redox potential) or pE (relative electron activity) low enough to produce ferrous iron, a 0.2% concentration of alpha,alpha'-dipyridyl (also listed as 2,2'-bipyridine) may be used (U.S. Department of Agriculture, Natural Resources Conservation Service, 2022; Childs, 1981). This solution is applied to the soil within minutes of exposure to air but in the absence of sunlight. If Fe^{2+} (ferrous iron) is present, the alpha,alpha'-dipyridyl will become red or pink; if only Fe^{3+} (ferric iron) is present, the solution will remain clear (Fig. 13).

Because this solution will react with just a few parts per million of Fe^{2+} , care is needed when extracting the soil. Ferrous iron can be transferred from a knife blade or spade and produce a false-positive response. The solution should only be used when reducing conditions are suspected.

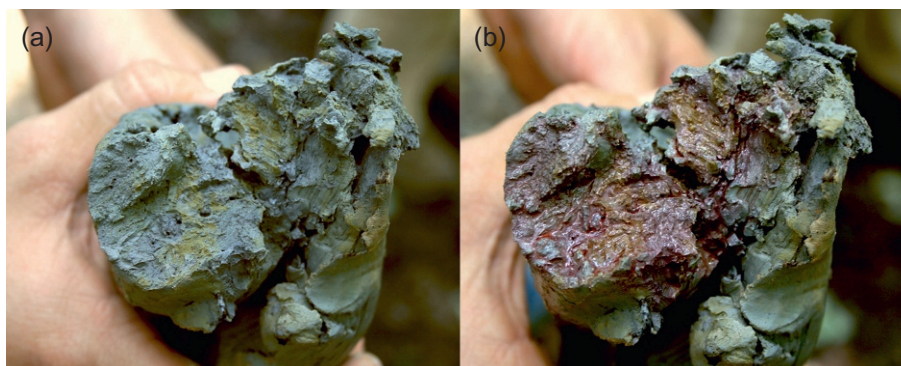


Fig. 13 A soil sample before (a) and after (b) being treated with alpha,alpha'-dipyridyl. Ferrous iron is detected (reddish areas in right image). Reproduced from John A. Kelley Collection, USDA Natural Resources Conservation Service (<https://www.flickr.com/photos/soilscience/albums>). Shared according to Creative Commons License terms (<https://creativecommons.org/licenses/by/2.0/>).

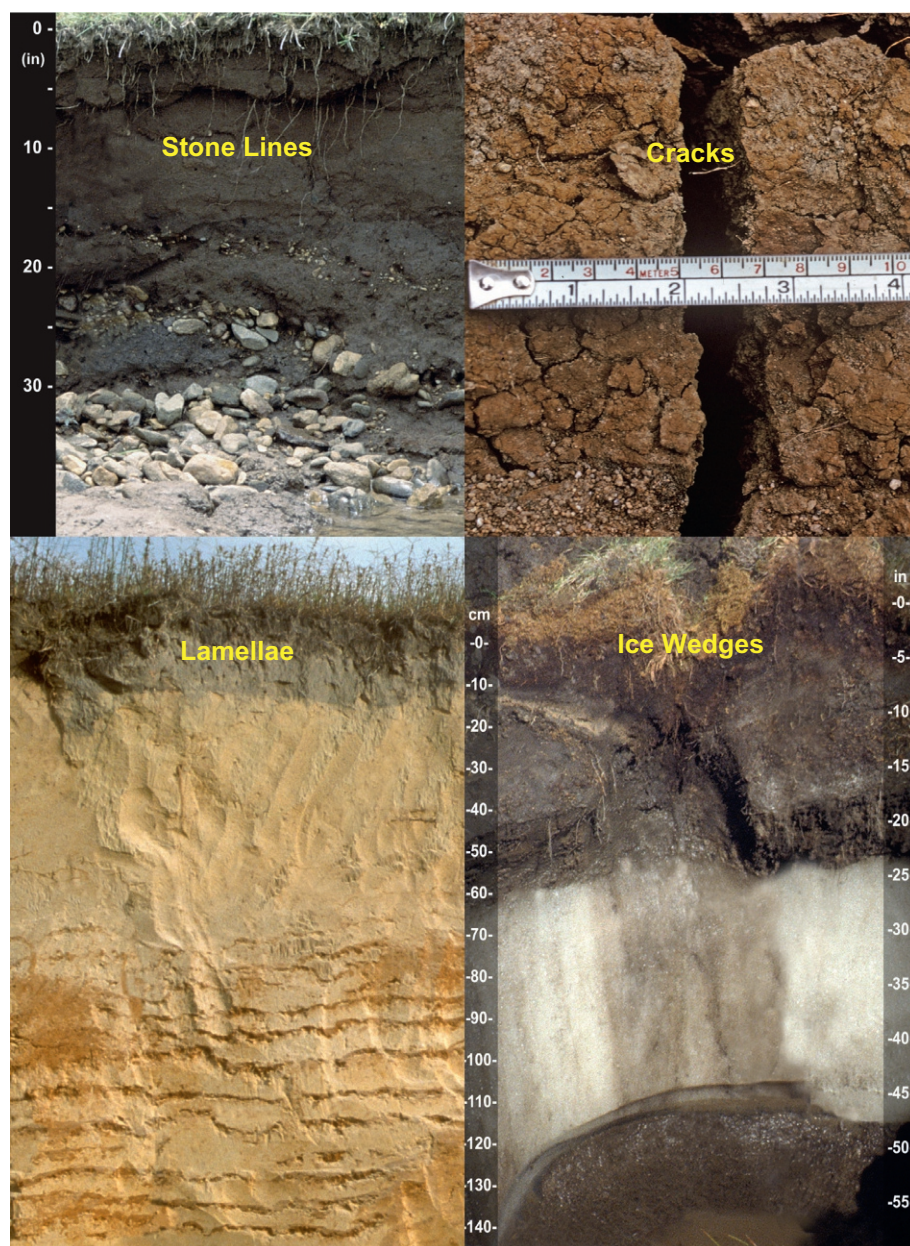


Fig. 14 Examples of kinds of pedon features (stone lines, cracks, lamellae, ice wedges). Reproduced from John A. Kelley Collection, USDA Natural Resources Conservation Service (<https://www.flickr.com/photos/soilscience/albums>). Shared according to Creative Commons License terms (<https://creativecommons.org/licenses/by/2.0/>).

Special features

Pedon features

Pedon features occur throughout soils and may cross adjacent horizons (Fig. 14). These features are described using standard technical terms. In some cases, additional information is needed in the description (Table 3).

Ped surface features

Surface features form from a variety of pedogenic processes. They differ from the matrix soil in composition, orientation, or packing. These features include: (1) coats of a variety of substances unlike the adjacent soil material that cover part or all of surfaces; (2) material concentrated on surfaces by the removal of other material; and (3) stress formations in which thin layers at the surface have undergone reorientation or packing by stress or shear. Descriptions of surface features may include kind, location, amount, continuity, distinctness, and thickness. Table 4 lists common ped surface features. Selected examples are shown in Fig. 15.

Table 3 Common pedon features.

<i>Pedon feature</i>	<i>Description</i>
Cracks	Fissures other than those attributed to soil structure
Desert pavement	Natural concentration of polished gravel and stones at the soil surface, primarily in arid and semiarid climates
Ice crystals, lenses, layers (veins), wedges	Features occurring in soils of Arctic, Antarctic, and Subarctic regions
Jarosite mottles	Yellow jarosite mineral accumulations resulting from oxidation of sulfidic accumulations
Krotovinas	Irregular tubular streaks of soil material transported from one horizon or layer to another by biologic activity
Lamellae	Thin banded illuviated clay
Microbiotic crust	Thin surficial zone exhibiting more mechanical continuity of the soil fabric than the subjacent soil
Stone line	Natural accumulation of coarse fragments below the surface that extends horizontally through the pedon
Tongues of albic material	Lobes of eluviated soil that extend into a soil horizon from overlying horizons

Reproduced from Schoeneberger PJ, Wysocki DA, Benham EC, and Soil Survey Staff (2012) *Field Book for Describing and Sampling Soils, Version 3.0*. USDA Natural Resources Conservation Service, National Soil Survey.

Table 4 Common ped surface features.

<i>Ped feature</i>	<i>Description</i>
Clay films	Layers of oriented, translocated clay (clay skins)
Clay bridges	Links between mineral grains occurring primarily in sandy soils
Stress surfaces, pressure faces	Smooth, shiny, or smeared surfaces formed through rearrangement of particles due to shear forces
Slickensides	Stress surfaces that are polished and striated; direction of movement can be determined from striations
Other features	Coats of iron, aluminum, or manganese oxides, organic matter, salts, or carbonates

Reproduced from Schoeneberger PJ, Wysocki DA, Benham EC, and Soil Survey Staff (2012) *Field Book for Describing and Sampling Soils, Version 3.0*. USDA Natural Resources Conservation Service, National Soil Survey.



Fig. 15 Clay bridges (smooth yellowish areas) coating individual quartz grains (white areas); clay films on the surface of a ped; and a shiny slickenside. Modified from John A. Kelley Collection, USDA Natural Resources Conservation Service (<https://www.flickr.com/photos/soilscience/albums>). Shared according to Creative Commons License terms (<https://creativecommons.org/licenses/by/2.0/>). Slickenside image (right), Reproduced from Soil Science Division Staff (2017) Soil Survey Manual. In: Ditzler C, Scheffe K, and Monger HC (eds) *USDA Handbook 18*. Washington, DC: Government Printing Office.

Location

Surface features may vary in the coverage of peds or may coat other soil constituents such as gravel or rocks.

Amount

The number of ped surface features is noted as a percentage or a percentage range. If a horizon has more than one surface feature, each surface feature is noted with the percentages of that feature.

Horizon or layer features

Boundary

Soil horizons or layers generally are oriented parallel to the surface. In some soils, however, they are oriented differently because of pedogenesis. In most description systems, the bottom boundary of the soil horizon or layer is described as smooth, wavy, irregular, or broken (Fig. 16). The distance over which the boundary grades into another horizon or changes is also usually described. Boundaries can be identified by classes: abrupt (0–2 cm), clear (2–5 cm), gradual (5–15 cm), and diffuse (more than 15 cm).

(a)

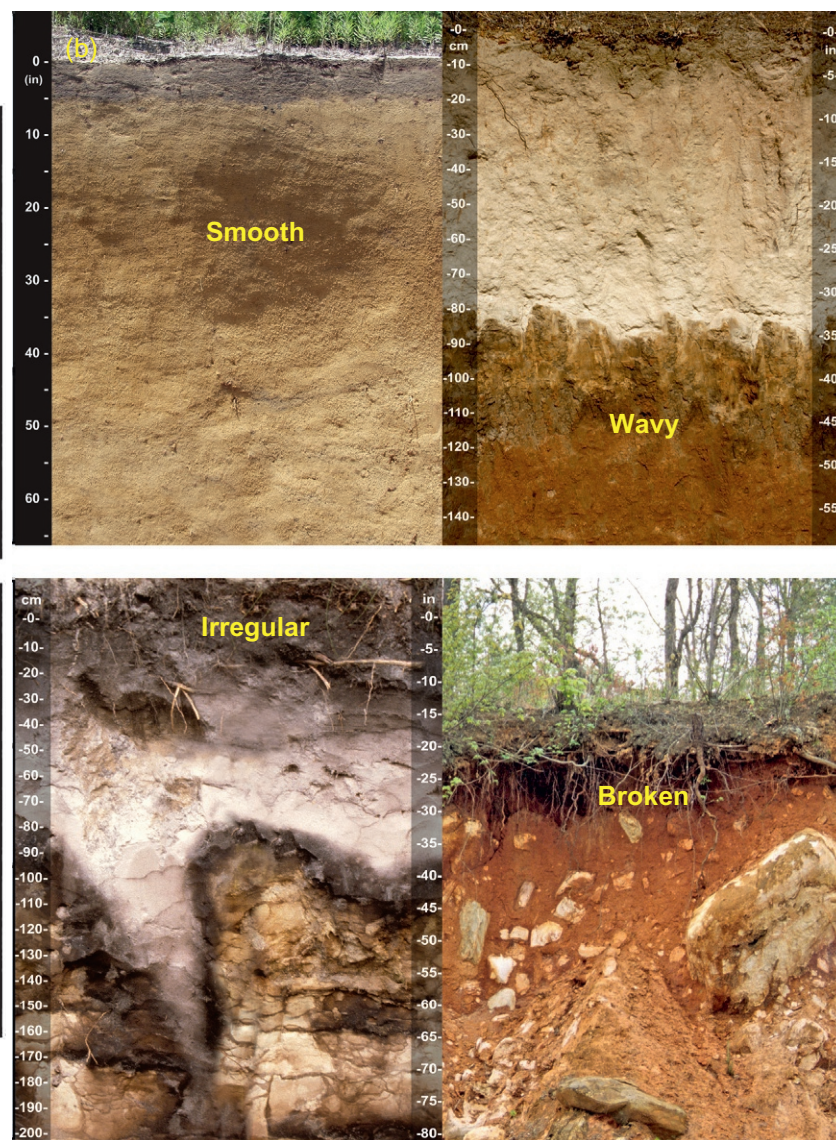
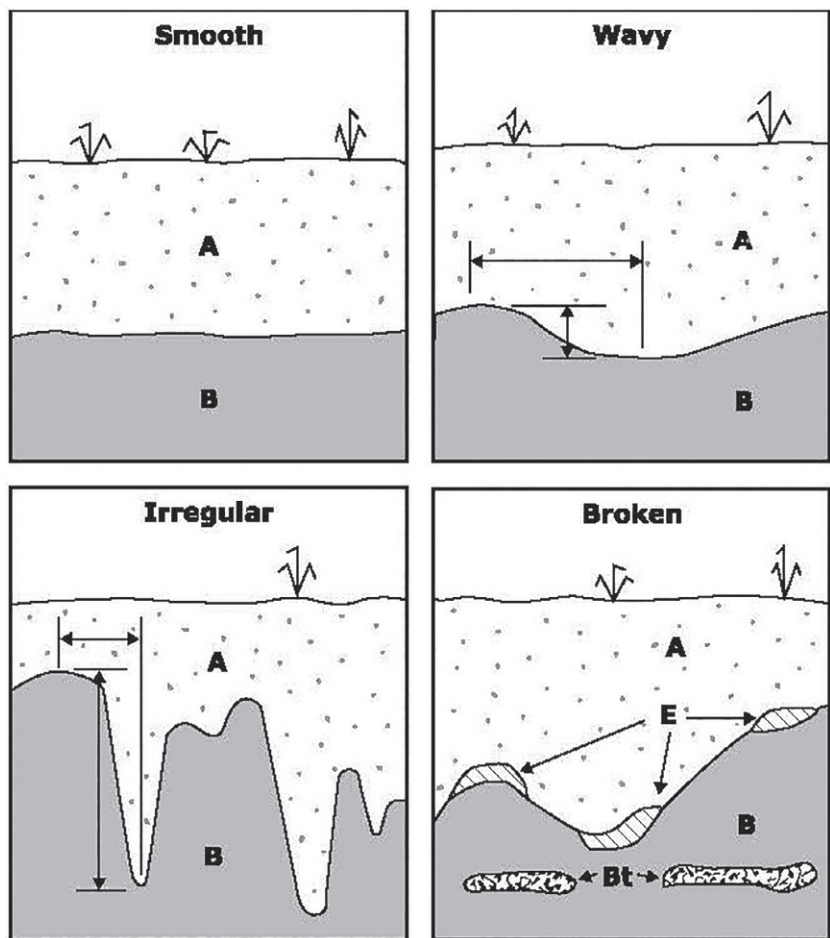


Fig. 16 Boundary nomenclature of soil horizons. (a) Reproduced from Schoeneberger PJ, Wysocki DA, Benham EC, and Soil Survey Staff (2012) *Field Book for Describing and Sampling Soils, Version 3.0*. USDA Natural Resources Conservation Service, National Soil Survey. (b) Examples of boundaries observed in the field. Modified from John A. Kelley Collection, USDA Natural Resources Conservation Service, <https://www.flickr.com/photos/soilscience/albums>. Shared according to Creative Commons License terms (<https://creativecommons.org/licenses/by/2.0/>).



Fig. 17 (a) A portable XRF spectrometer is used in the field to scan exposed faces of soil pits or surfaces. Reproduced from Soil Survey Manual. Ditzler C, Scheffe K, and Monger HC (eds) *USDA Handbook 18*. Washington, DC: Government Printing Office. (b) A mid-infrared (MIR) instrument for determining soil properties based on soil reflectance (personal contribution).

Proximal sensing

Field observations and measurements, whether based on human sensors (sight, touch, etc.) or chemical sensors (acid and alkaline solutions, etc.), can be time-consuming or costly, or both. Proximal soil sensing is a faster and often cheaper technology (Viscarra Rossel et al., 2011). It uses sensors close to, or in direct contact with, the soil surface and measures soil properties directly or indirectly. Two methods used to document soil properties at specific locations (point data) are portable X-ray fluorescence (P-XRF) and optical reflectance (Fig. 17).

Portable X-ray fluorescence (P-XRF) uses high-energy X-ray photons to destabilize electrons, which emit X-ray energy when moving from outer to inner orbits (shell). Because different chemical elements emit specific fluorescence, they can be identified and quantified. The p-XRF has many applications, including (i) to differentiate soil horizons based on the concentration of different metals; (ii) to quantify the calcium content and percent of gypsum; (iii) to assess differences in weathering indices, mineralogy, and geologic signatures related to soil development; and (iv) to characterize differences in mineralogy and lithology.

Optical reflectance is based on the soil's ability to reflect light in different parts of the electromagnetic spectrum. The spectrum includes ultraviolet (100–400 nm), visible (400–750 nm), near infrared (750–2500 nm), and mid-infrared (2500–25,000 nm) wavelengths. Measurements obtained using optical sensors can be related to many soil properties, such as soil mineral composition, texture (sand, silt, and clay content), soil color, moisture, organic carbon content, pH, cation-exchange capacity, and carbonates.

Digital soil profile description (digital soil morphometrics)

Digital soil morphometrics is based on proximal sensing tools and techniques to measure and quantify soil properties and could complement existing description and analytical methods. In addition, it may offer new insights in soil horizon development and classification (Hartemink and Minasny, 2014). One of the main expected benefits of digital soil morphometrics is the quantification of soil profile descriptions to create continuous soil depth functions that can be related to other laboratory measured soil properties. Digital soil morphometrics can be used for soil horizons, colors, texture, structure, coarse fragments, pores, and roots, etc., (Gallegos et al., 2021). Analysis of the observations and images from the soil profile use image analysis and pattern recognition software and multiple steps for processing (Figs. 18–20).

Soil micromorphological description

Most of this chapter has focused on the macromorphological description of soils in the field, but micromorphological descriptions can be made by laboratory analysis of soil thin sections (Bullock et al., 1985). Micromorphological descriptions can supplement and verify soil features and are valuable for classifying soils and understanding pedogenesis. Fig. 21 illustrates the use of thin sections to quantify pore size, shape, and distribution, information important in understanding water movement through the soil profile (Libohova et al., 2015).

Conclusions

Traditionally, soil morphology has provided the first clues about soils, properties and pedogenesis under field conditions as observed by soil scientists using simple tools like a knife, tape measure, water bottle, Munsell color book, 10× hand lens, 10% HCl

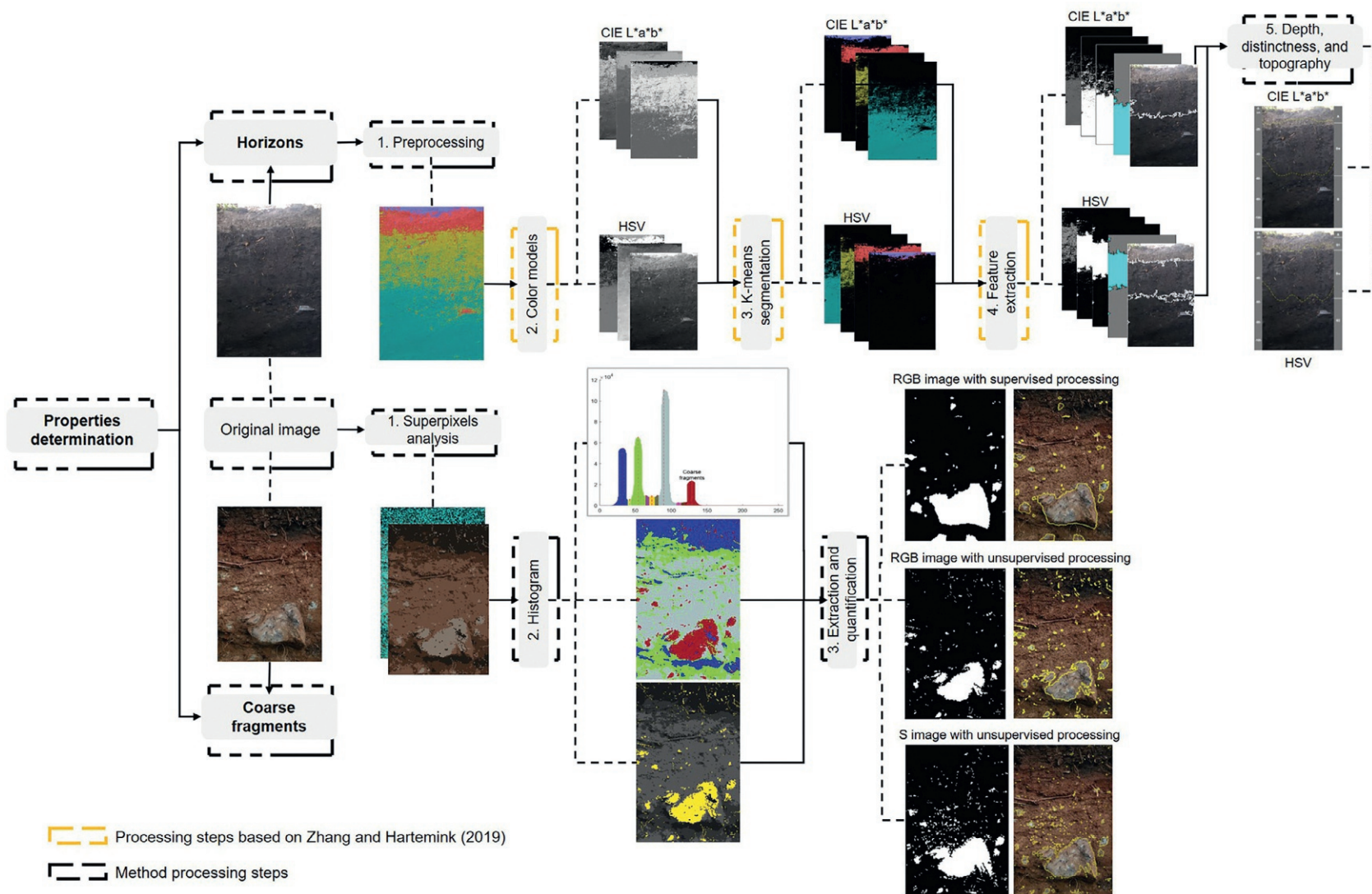


Fig. 18 A flowchart showing step by step analyzes of soil profile images to identify, describe, and quantify features such as color, coarse fragments, horizon boundaries, etc. Reproduced from Gallegos A, García-Oliva F, Pereira-Corona A, Bautista F (2021) Digital soil morphometrics of coarse fragments and horizon delineation in soil profiles from Central Mexico. *Geoderma Regional* 26: doi: 10.1016/j.geodrs.2021.e00403.

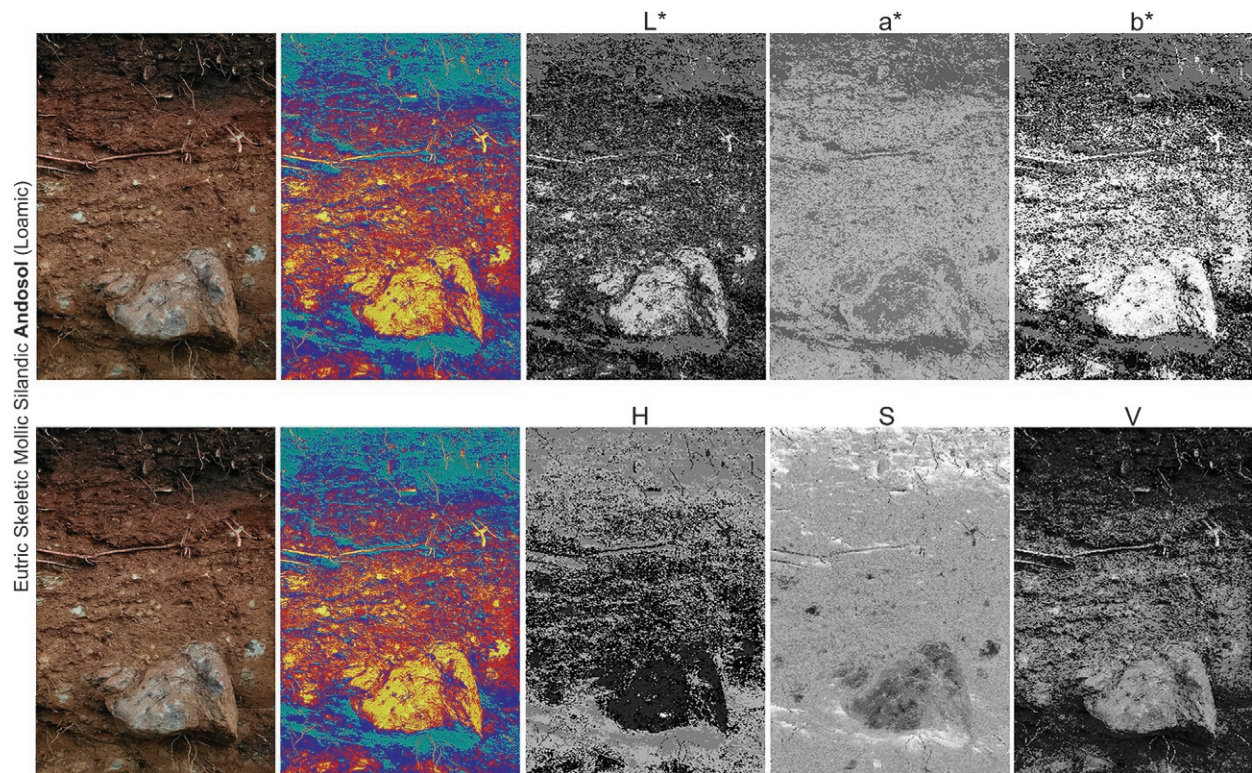


Fig. 19 The CIE $L^*a^*b^*$ and HSV color systems used to extract information from the soil profile. Reproduced from Gallegos A, García-Oliva F, Pereira-Corona A, Bautista F (2021) Digital soil morphometrics of coarse fragments and horizon delineation in soil profiles from Central Mexico. *Geoderma Regional* 26. doi: <https://doi.org/10.1016/j.geodrs.2021.e00403>.

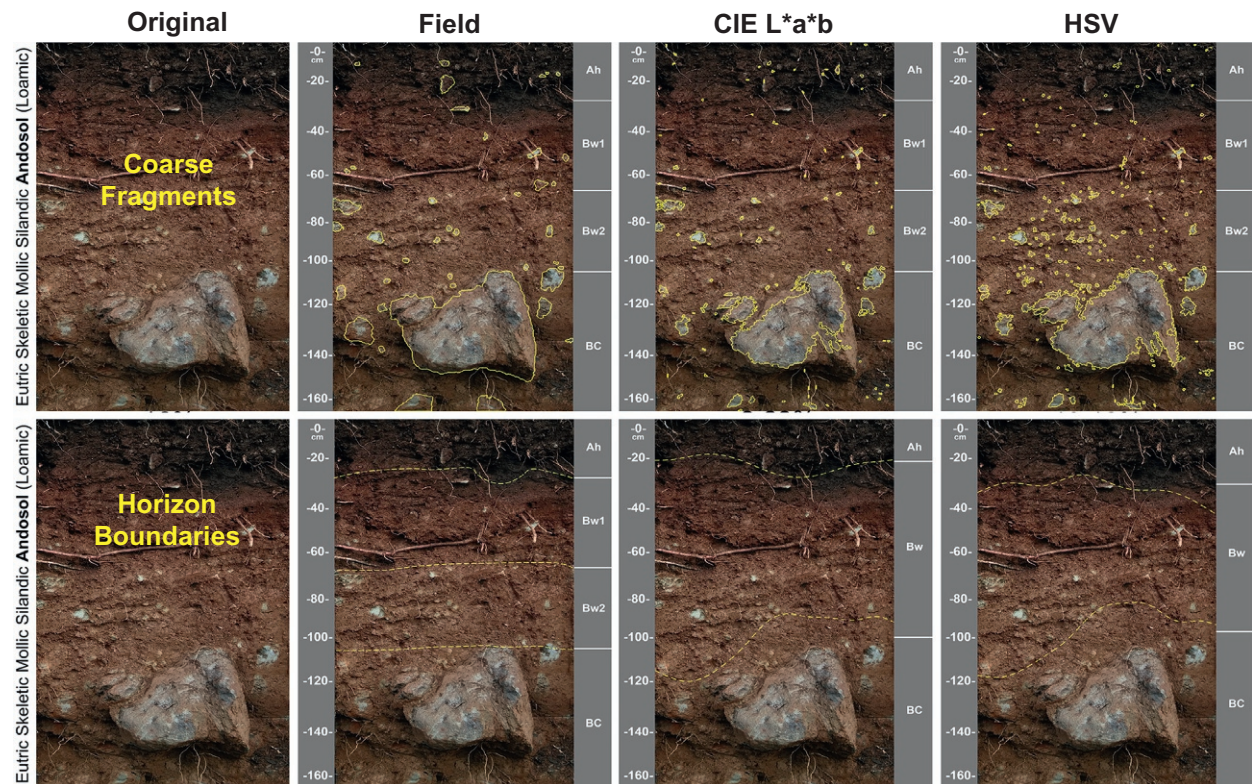


Fig. 20 CIE $L^*a^*b^*$ and HSV color systems used to extract information from the soil profile for further processing and identification of features such as coarse fragments and soil horizon boundaries. Modified from Gallegos A, García-Oliva F, Pereira-Corona A, Bautista F (2021) Digital soil morphometrics of coarse fragments and horizon delineation in soil profiles from Central Mexico. *Geoderma Regional* 26. <https://doi.org/10.1016/j.geodrs.2021.e00403>.

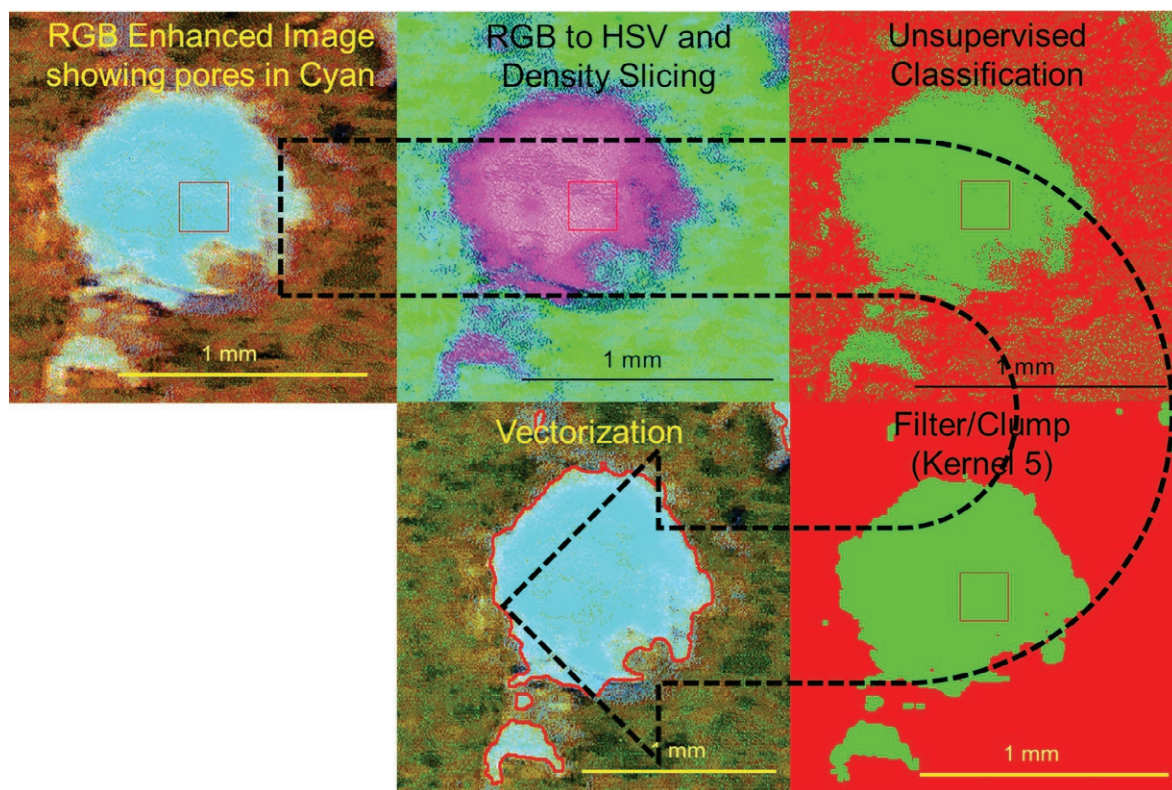


Fig. 21 A soil thin section processed with different techniques to quantify pore size, shape, and distribution. Reproduced from Libohova Z, Schoeneberger P, Owens PR, Wills S, Wysocki D, Williams C, and Seybold C (2015) Comparative analysis of saturated hydraulic conductivity (Ksat) derived from image analysis of soil thin. In: Hartemink AE and Minasny B (eds) *Digital Soil Morphometrics. Progress in Soil Science*. Springer International Publishing: Switzerland. doi: https://doi.org/10.1007/978-3-319-28295-4_10.

(hydrochloric acid), a clinometer etc. Soil micromorphology added polarizing light microscopes and thin sections among other tools for describing and quantifying soil features such as pore size and distribution, mineralogy, etc. The development of proximal sensors such as portable X-ray fluorescence (pXRF), mid-infrared spectroscopy (MIR), etc., has taken soil morphology and micromorphology all together to the new era of digital soil morphometrics. The continuous development of new tools and sensors for observing soils, on one hand, brings the soil morphology and micromorphology closer to biological, physical, and chemical soil analysis. On the other hand, digital soil morphometrics promises quantitative analysis and evaluations of complex soil systems through a combination of the relevant tools and sensors with powerful software and machine learning algorithms.

References

- Brewer R (1976) *Fabric and Mineral Analysis of Soils*. Huntington, NY: Kreiger.
- Bullock P, Federoff N, Jongerius A, Stoops G, and Tursina T (1985) *Handbook for Thin Section Description*. Wolverhampton, UK: Waine Research.
- Buol SW, Hole FD, McCracken RJ, and Southard RJ (1997) *Soil Genesis and Classification*, 3rd edn. Ames, IA: Iowa State University Press.
- Childs CW (1981) Field tests for ferrous iron and ferric organic complexes (on exchange sites or in water-soluble forms) in soils. *Australian Journal of Soil Research* 19: 175–180.
- CIE (1932) *Commission Internationale De l'Eclairage Proceedings, 1931*. Cambridge: Cambridge University Press.
- Gallegos A, García-Oliva F, Pereira-Corona A, and Bautista F (2021) Digital soil morphometrics of coarse fragments and horizon delineation in soil profiles from Central Mexico. *Geoderma Regional* 26. <https://doi.org/10.1016/j.geodrs.2021.e00403>.
- Hartemink AE and Minasny B (2014) Towards digital soil morphometrics. *Geoderma* 230–231: 305–317. <https://doi.org/10.1016/j.geoderma.2014.03.008>.
- IUSS Working Group WRB (2015) *World Reference Base for Soil Resources 2014, Update 2015 International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*. World Soil Resources Reports No. 106. Rome: FAO.
- Jenny H (1941) *Factors of Soil Formation—A System of Quantitative Pedology*. McGraw-Hill.
- Libohova Z, Schoeneberger P, Owens PR, Wills S, Wysocki D, Williams C, and Seybold C (2015) Comparative analysis of saturated hydraulic conductivity (Ksat) derived from image analysis of soil thin sections, pedotransfer functions, and field-measured methods. In: Hartemink AE and Minasny B (eds.) *Digital Soil Morphometrics. Progress in Soil Science*. Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-28295-4_10.
- Munsell Soil Color Charts (1994) *Macbeth Division of Kollmorgen Instruments Corporation, 405 Little Britain Road*. New Windsor, NY: Munsell Soil Color Charts. 12553 (Revised Edition).
- Schoeneberger PJ, Wysocki DA, Benham EC, and Soil Survey Staff (2012) *Field Book for Describing and Sampling Soils, Version 3.0*. Lincoln, NE: USDA Natural Resources Conservation Service, National Soil Survey Center.
- Soil Science Division Staff (2017) Soil survey manual. In: Ditzler C, Scheffe K, and Monger HC (eds.) *USDA Handbook 18*. Washington, DC: Government Printing Office.
- Soil Survey Staff (2010) *Keys to Soil Taxonomy*, 11th edn. Washington, DC: U.S. Government Printing Office. USDA Natural Resources Conservation Service.

- Soil Survey Staff (2014) Kellogg Soil Survey Laboratory Methods Manual. In: Burt R and Soil Survey Staff (eds.) *Soil Survey Investigations Report No. 42, Version 5.0*. U.S. Department of Agriculture, Natural Resources Conservation Service.
- U.S. Department of Agriculture, Natural Resources Conservation Service (2022) *National Soil Survey Handbook, Title 430-VI*. http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/ref/?cid=nrcs142p2_054242 (accessed 25 01, 2022).
- Vepraskas MJ (1992) *Redoximorphic Features for Identifying Aquic Conditions*. North Carolina Agriculture Research Service Technical Bulletin 301. Raleigh, NC: North Carolina State University.
- Viscarra Rossel RA, Adamchuk VI, Sudduth KA, McKenzie NJ, and Lobsey C (2011) Proximal soil sensing: An effective approach for soil measurements in space and time, chapter 5. *Advances in Agronomy* 113: 237–283.
- Wilding LP, Smeck NE, and Hall GF (eds.) (1983) *Pedogenesis and Soil Taxonomy. Concepts and Interactions, Part A*. New York: Elsevier.

Soil micromorphology

Rosa M Poch^a and Georges Stoops^{b,*}, ^aDepartament de Medi Ambient i Ciències del Sòl, Universitat de Lleida, Lleida, Spain; ^bVakgroep Geologie, Universiteit Gent, Gent, Belgium

© 2023 Elsevier Ltd. All rights reserved.

Introduction	172
History	172
Techniques	173
Sampling and thin section preparation	173
Observation techniques	173
Study of thin sections	174
Microstructure and groundmass	174
Pedofeatures	176
Applications	178
Soil genesis and paleopedology	178
Pedogenic formation processes in arid soils: calcium carbonate, gypsum and salt dynamics	178
Processes in acid sulfate soils	178
Clay illuviation	180
Redox processes	180
Bioturbation	180
Frost action	180
Argilloturbation	181
Podsolization	181
Formation of oxic or ferrallitic materials	181
Soil classification	181
Geomorphology and quaternary geology	182
Agronomy	182
Archaeology and geoarchaeology	183
Forensic science	183
Conclusions and perspectives	183
References	184

Abstract

Soil micromorphology is the technique in soil science that enables us to observe how soil constituents are organized at a microscopic scale, i.e., between field and submicroscopic observations. Special terms are used to study and describe soil microfabrics viewed in soil thin sections, which have evolved from a genetic to a more morphoanalytical approach. The applications of soil micromorphology extend to most fields of the earth and archaeological sciences, but it is in soil genesis, paleopedology and geoarchaeology where it has been most useful to determine relative chronologies of processes.

Key points

- In the continuum of soil observation, soil micromorphology fills the gap between field and submicroscopic scales.
- Special concepts and terms are needed to describe and understand the organization of soil constituents at the microscopic scale.
- The interpretation of micromorphological features needs to establish relationships between features and processes that lead to them.
- Soil micromorphology has been applied to soil genesis, soil management, soil classification, geomorphology, paleopedology, archaeology and forensic sciences among other disciplines within earth sciences.

*Deceased.

Introduction

Soil micromorphology is the observation and interpretation of undisturbed, oriented samples of soil material by microscopic and ultramicroscopic techniques to identify their constituents (including voids) and to determine their mutual relations in space and time. In the continuum of observation of soil cover: pedological systems—horizons—materials—elementary organizations (FAO/ISRIC/ISSS, 1998), soil micromorphology deals with the latter two levels. They require special techniques to observe and study them. Soil micromorphology allows us to deduce genetic and chronological relationships between constituents; this is soil genesis in its widest sense, comprising natural and also human induced processes, e.g., by drainage, irrigation or management. Other techniques, such as in soil chemistry, physics and mineralogy often deal with homogenized bulked samples where heterogeneity and anisotropy are lost, whereas micromorphology enables us to observe and measure the micro-heterogeneity and anisotropy in undisturbed samples. For example, chemistry provides results on the total content of calcium carbonate, whereas micromorphology distinguishes between lithogenic or pedogenic calcite or aragonite present as coatings or nodules.

At the outset the term ‘micropedology’ was commonly used, which seemed to narrow the approach to pedogenetic studies only. It was soon replaced by ‘soil micromorphology,’ to emphasize its broader use.

To keep the samples undisturbed, special sampling techniques have been developed. Most micromorphological studies are made with a petrographic microscope, using polarized light with thin sections prepared from resin impregnated blocks, sometimes completed by ultramicroscopic techniques and/or microanalyses.

In the second half of the 20th century soil micromorphology was mainly used as a tool for pedogenic research and to support the development of soil classifications. When funds for these disciplines decreased, and therefore the use of micromorphology by soil scientists, this was compensated for by an increase in interest by palaeopedologists, Quaternary geologists and especially geoarchaeologists.

History

Soil thin section analyses were recognized as a valuable technique in soil science after the publication of Kubiěna’s book “Micropedology” in 1938. It introduced concepts and terms to distinguish between coarse constituents (‘fabric skeleton’) and fine constituents and their state (‘fabric plasma,’ peptized or pectized) and their configurational relationship (‘elementary fabrics’). In his later publications Kubiěna (1938) no longer considered individual constituents, but the overall aspect of the thin section, naming it after the soil type in which it characteristically occurred, such as Braunlehm, Braunerde or Rotlehm. This system was in use until the end of the 1960s for soils in Europe, North America and partly for humid tropical soils.

The interpretation of soil thin sections was then mainly limited to a mention of the corresponding soil-type described by Kubiěna (Braunlehm, Rotlehm, etc.), thus as a basis for classification only. This implied that the pedogenesis of these materials described by Kubiěna was also accepted, although his interpretation was not based on other evidence such as chemical, physical or mineralogical analyses. Therefore, its use was applied more to European soils, than to the as yet not well understood soils of the tropics.

Early in the 1960s the technology for thin section preparation considerably improved with the use of cold-setting polystyrene for impregnation, which avoided the artefacts often observed in the thin sections of soils made with resins that needed heating, as done before. Moreover, it enabled the preparation of larger samples (e.g., thin sections of 90 mm × 180 mm, compared to 28 mm × 48 mm previously) and the analysis of larger fabrics and structures.

The Australian geologist Roy Brewer published several papers, starting from 1960, on a new analytical approach to micromorphology, which were compiled in 1964 in his famous book ‘Fabric and Mineral Analysis of Soils’. He was the first scientist to make a clear distinction between the undifferentiated groundmass (called ‘s-matrix’), composed of immobile coarse constituents (the ‘skeleton grains’) and the mobile, fine mass (‘plasma’), and the organized special pedogenic features (‘pedological features’). Pores (voids) were treated as equally important fabric constituents. A weak point was that organic material was totally ignored. This system soon became used worldwide in the second half of the 1960s. However, as many new soils were discovered and studied in the arid and humid tropics, it became clear that the system was incomplete and contained contradictions. This resulted in additions by several authors, and also in an uncontrolled proliferation of terms.

In this approach the individual constituents and features were identified, which made it possible to quantify them. In the first part of the 1960s this was done by manual point counting, as used in petrography, but later in that decade the first tools for image analysis became available, and were mainly used for porosity studies. The first scanning electron microscopes and microprobes also became commercially available in that period for in situ study of fabrics and constituents at high magnification.

To come to a coherent, internationally accepted morphoanalytical system, a working group was created by the ISSS (now IUSS) in 1969. This resulted in the publication of a handbook by Bullock et al. (1985) with precisely defined concepts and terms that were partly based on Brewer’s system, but with important contributions of sedimentary petrography and botany. The system was applied worldwide as a standard for soil thin section analysis. When the book was out of print at the end of the 1990s, the Soil Science Society of America asked Georges Stoops to prepare a revised and updated version (Stoops, 2003), which contained keys for the morphometric identification of several features. After this was out of print at the end of the 2010, an updated version was prepared (Stoops, 2021).

In the former Soviet Union terms borrowed from traditional Russian sedimentary petrography, completed with terms from Kubiěna, were still used for some time, but were gradually replaced by the system of Bullock et al. (1985) and Stoops (2003).

Much effort was made to synthesize systems for thin section descriptions, but this was not the case for interpretation. Apart from the books by FitzPatrick (1984, 1993) based on his own experience, and rare review papers on specific soil types, scientists had to rely on papers in journals and proceedings that dealt with local soils. Until 1965 fewer than 40% of the papers were published in English (Stoops, 2014). The first overall review paper on natural soils was published only 10 years ago (Stoops et al., 2010, 2018), whereas for archaeological materials the following books became available: Courty et al. (1989), Nicosia and Stoops (2017) and Macphail and Goldberg (2018).

For more information and references of the publications cited in this entry, the reader is referred to Stoops (2009).

Techniques

Sampling and thin section preparation

By definition, soil micromorphology concerns the study of undisturbed samples, therefore, special sampling techniques are necessary. During the whole process from sampling in the field through to the different sample treatments, it is essential to keep the soil constituents and microstructure as they were originally, and in such a way that they can be clearly observed by microscopic techniques. Therefore, any change in structure (crushing, swelling/shrinking, collapse) or composition (dissolution, mineralogical changes) should be avoided, as well as introducing any component that may interfere with the observation.

In the case of materials with enough coherence, blocks can be carved out of the profile and carefully packed with newspaper or plastic wrap, ensuring that the top and bottom orientations are clearly indicated. For less stable and loose soil materials the so called Kubiěna boxes were developed. These are metal boxes consisting of a frame that can be opened, and with two covers (top and bottom). The frame is gently inserted into the profile wall by pressing on a pre-carved block. After removal the box is closed with the two covers, and the sample number and orientation are marked on the frame. As cheaper alternatives recycled cardboard boxes (tetra-brick type) or special plastic boxes are sometimes used. Pre-carved blocks can be stabilized by enveloping them in plaster bandages or by spreading plaster of Paris inside the box before encasing the soil block, to give cohesion to it before its removal. The blocks are then stored in boxes after hardening. For more detail on sampling techniques and strategies see Nicosia and Stoops (2017), ch. 35.

Most soil micromorphology research is done on soil thin sections, although some can be done on polished block faces, whole blocks or on soil aggregates. To prepare a thin section, a hard material is needed. This can be obtained by impregnating the undisturbed soil sample, once removed from the box, under vacuum conditions with a cold-setting resin, mostly a diluted polystyrene. As the resin is incompatible with water, the latter has to be removed, either by air or oven drying, or by acetone replacement. This is done by submerging the sample in an acetone-water bath, and replacing the diluted acetone daily by a dilution progressively richer in acetone, until all water is extracted (about 1 week), or by putting the wet sample on a plate over a disk with acetone in a closed desiccator. An equilibrium will be established between water vapor and acetone vapor in the atmosphere. By replacing daily the acetone diluted by the water of the sample until it becomes pure acetone, all water will be gradually extracted (duration about 1 month). This procedure has the advantage that disturbance by shrinking, as a result of drying, is avoided.

After hardening, a slice of the impregnated block is cut with a diamond saw cooled with a special oil (to avoid solution of constituents by water). The slice is then ground and mounted on a support glass of corresponding sizes (e.g., 48 mm × 28 mm for a petrographic size thin section, up to 180 mm × 120 mm for a so called 'mammoth' thin section). By further sawing and grinding the thickness of the slice is reduced to 20–30 µm, the standard thickness of a thin section. At that thickness, most minerals are transparent, and can be determined by optical techniques. For grinding, a slurry of fine corundum or diamond powder and oil is used. Thin sections are sometimes covered with a very thin cover-glass for protection, unless staining tests, cathodoluminescence or microprobe studies are planned. For more detail and references see Murphy (1986).

Some differences to the basic procedure include, for example, percolation tests with dyeing agents as a pre-treatment to identify connected porosity, or the addition of fluorescent dyes to the resin so that porosity images can be obtained with UV (ultraviolet) light.

Observation techniques

The most common observation technique is optical microscopy, using a polarizing microscope (petrographic microscope). This requires at least a basic knowledge of optical crystallography (Stoops, 2021).

For specific research a range of additional techniques is available. The use of oblique incident light is needed to distinguish between opaque constituents and to determine their internal fabric. Incident UV or blue light can be used to detect specific minerals such as phosphates (e.g., bone splitters) and cellulose. When a fluorochrome (a photoreactive chemical) is added to the impregnating resin, observation in UV or blue light shows a clear image of the porosity. For cathodoluminescence studies an uncovered thin section is mounted in a small vacuum chamber with two glass windows and an electron gun is mounted on the stage. As the cathodoluminescence of some minerals, such as calcite, depends upon the presence of trace elements in the lattice, information on the different origins or pedogenic stages can be obtained (for more information see Nicosia and Stoops, 2017, ch 36).

Uncovered thin sections can also be treated with selective staining solutions (e.g., to distinguish between carbonates) or can be subjected to selective extraction treatments (e.g., for amorphous, organic bound or free iron oxides) (Stoops, 2021).

Since the 1970s, ultramicroscopic techniques, such as TEM (Transmission Electron Microscopy) on ultrathin slices or SEM (Scanning Electron Microscopy) on fracture surfaces, are used as complementary techniques. Microprobe studies of carbon coated uncovered thin sections are used to determine and map the chemical composition of constituents; μ -XRD (micro X Ray Diffraction) and μ -FTIR (micro Fourier Transform Infrared Spectroscopy) are other accessory techniques used on thin sections (Nicosia and Stoops, 2017).

Scans of the total thin section are also useful to observe soil structure and bioturbation by macrofauna. Photographic polarizing films can be used as polarizer and analyzer when scanning thin sections.

The X-ray computed tomography permits the visualization and quantification of soil porosity through the three-dimensional analysis of intact soil cores. Micron-scale X-ray CT (μ CT) is the preferred method to analyze intact soil blocks, which enables analysis of their structure, while allowing other tests on the same sample afterwards. Samples for this method do not need to be dried before analyses (Taina et al., 2008).

Study of thin sections

At present soil thin sections are generally analyzed and described according to internationally accepted concepts and terms proposed by Bullock et al. (1985) and later by Stoops (2003, 2021); these are summarized below. The terminology is available in 19 languages on <https://www.isric.org/explore/ISRIC-collections/micromulti> (accessed January 2022).

Analysis of the soil fabric comprises the description of the shape, relative size and frequency of the soil constituents and their spatial arrangement. The latter includes their orientation and distribution patterns.

Microstructure and groundmass

Microstructure deals with the size, shape and arrangement of primary particles (e.g., sand grains) or aggregates, and voids. Voids can include 'packing voids' (e.g., between grains of sand or between aggregates), 'planes' (e.g., between blocky peds), 'channels', 'chambers' (equidimensional voids interconnected by channels), 'vesicles' (spherical voids), 'vughs' (irregular equidimensional voids) and 'mouldic' voids (resulting from total dissolution of a constituent). Various types of 'peds' are considered: (sub)angular blocky, platy, lenticular, crumb and granular. Massive horizons often have a channel, vuggy or vesicular microstructure.

A distinction is made between the groundmass, composed of coarse and fine constituents and associated packing voids, and pedofeatures, discrete fabric units distinguishable from the adjacent groundmass by a difference in concentration of one or more constituents (e.g., clay coatings, iron nodules, gypsum crystals) or by a difference in fabric (e.g., a crescent arrangement due to biological activity). Three basic constituents are distinguished: coarse and fine material (either mineral or organic), and voids. The size limit between coarse and fine ('c/f limit') is not fixed and is a function of the type of material studied and the magnification used, but it is mainly between 2 μ m and 10 μ m. The coarse fraction comprises mineral grains (e.g., quartz, feldspar, mica), rock fragments (e.g., basalt, volcanic glass, sandstone), inorganic residues of biological origin (e.g., phytoliths, diatoms, shells, bones), anthropogenic elements (e.g., brick fragments, furnace slag) and larger organic remains. Their size, shape and abundance and degree of alteration need to be mentioned. The fine fraction contains, depending upon the c/f limit chosen, clay (e.g., clay minerals, short range order minerals, cryptocrystalline goethite or hematite, fine organic material) and fine silt particles, including microcrystalline calcite.

Organic material is important in horizons such as A1 and Bh. Depending upon the size and complexity one can recognize organ, tissue and cell residues, in addition to fine organic material such as opaque punctuations and amorphous material. In addition, the degree of decomposition is determined based on color, birefringence and sometimes fluorescence.

The way that the coarse and fine constituents are arranged is determined by their c/f (coarse/fine) related distribution pattern. Seven types are distinguished (Fig. 1): 'coarse monic': only coarser particles are present, 'chitonic': the fine particles occur as a coating around the larger ones, 'gefuric': the finer particles form bridges between the coarser, 'enaulic': the finer particles occur as aggregates between the coarser, 'porphyric': the coarser particles are embedded in a finer mass, and 'fine monic': only finer particles are present. Intergrades between these types commonly occur. Different types of microstructures and c/f related distribution patterns are shown in Fig. 2.

The fabric within the micromass cannot be observed directly because the constituent particles are below the resolution of the optical system. However, clay particles in soil form packages called 'domains' of parallel oriented flakes, which behave optically as single crystals, and therefore visible between crossed polars. Their arrangement enables identification of the birefringence fabric or b-fabric of the micromass. The following types of b-fabric are recognized: (i) 'undifferentiated': no interference colors visible between crossed polars (e.g., in allophanic or ferralitic materials); (ii) 'speckled': small equant domains occur (e.g., in most A1 horizons); (iii) 'striated': elongated domains (striae) that can occur on pore walls ('porostriated,' e.g., on slickensides), on grain boundaries ('granostriated'), as single streaks ('monostriated'), as parallel lines (parallel striated), as intersecting lines ('cross striated'), as rings ('circular striated'), as concentric rings ('concentric striated'), as bowlike features ('crescent striated') and without any specific pattern ('random striated'); (iv) 'crystallitic': the interference colors are caused by very small crystals, mostly of calcite, for example in arid soils. Striated fabric generally indicates the presence of swelling clay minerals, e.g., in Vertisols.

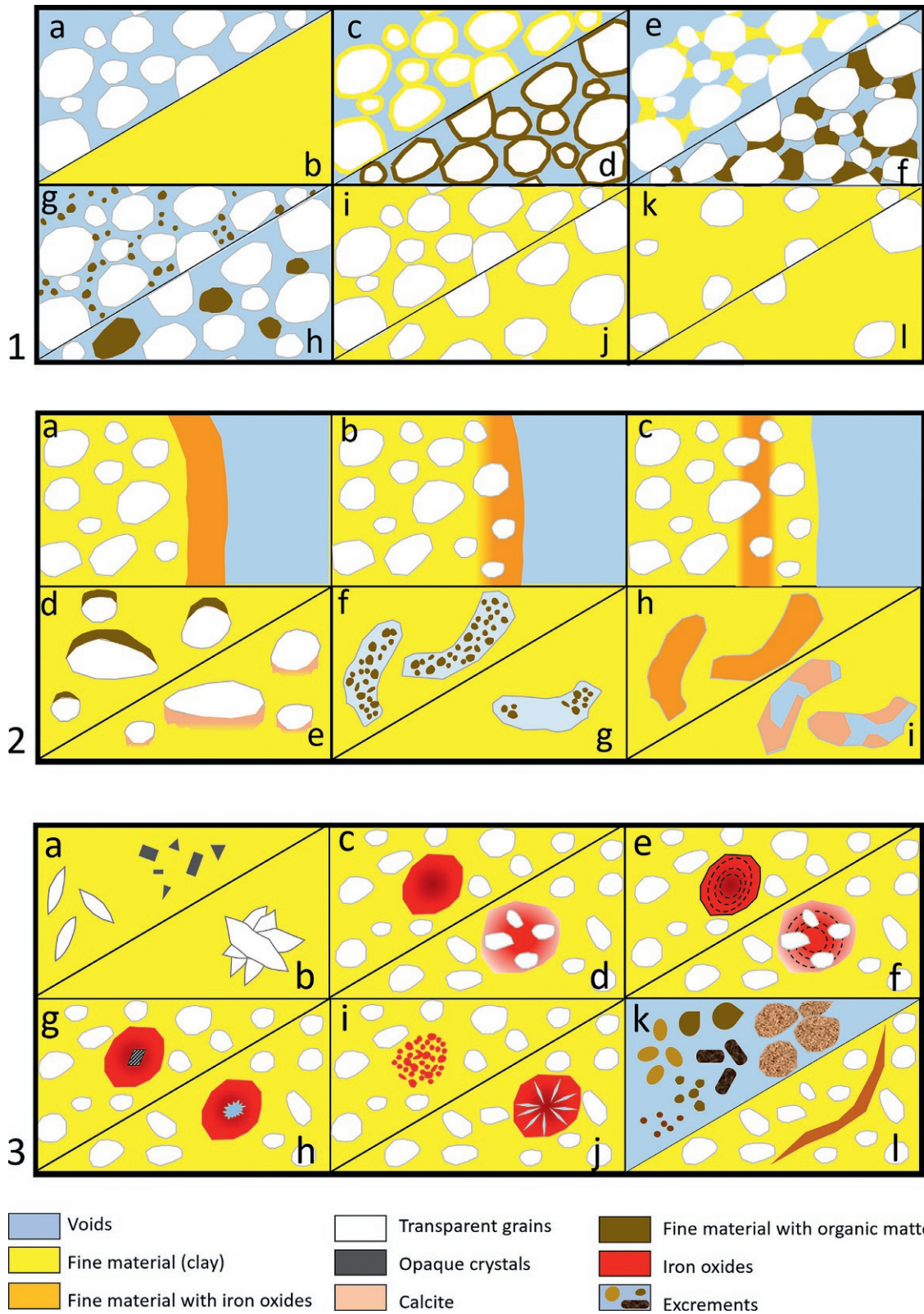


Fig. 1 Coarse/fine (c/f) related distributions: (1a) coarse monic; (1b) fine monic; (1c) chitonic with clay coatings (as in Bt); (1d) chitonic with organic coatings (as in spodic B); (1e) concave gefuric; (1f) convex gefuric; (1d) fine enaulic (aggregates smaller than coarse grains); (1e) equal enaulic (aggregates same size as coarse grains); (1f) close porphyric (grains are touching, although not always in the plane of the section); (1g) single spaced porphyric; (1h) double spaced porphyric; (1i) open porphyric. Pedofeatures: (2a) typical coating of iron-rich clay; (2b) hypocoating of iron-oxides; (2c) iron-oxide quasiccoating; (2d) organic rich capping on coarser grains; (2e) calcite pendants below coarser grains; (2f) loose continuous infilling; (2g) loose discontinuous infilling; (2h) dense complete infilling; (2i) dense incomplete infilling; (3a) lenticular gypsum crystals (left) and cubic pyrite crystals (right); (3b) crystal aggregate of gypsum; (3c) intrusive typical iron rich nodule; (3d) impregnative iron rich nodule; (3e) intrusive concentric nodule; (3f) impregnative concentric nodule; (3g) nucleic nodule; (3h) geodic calcite nodule; (3i) iron rich aggregate nodule; (3j) calcitic septaric nodule; (3k) different shapes of excrements; (3l) iron rich clayey intercalation.

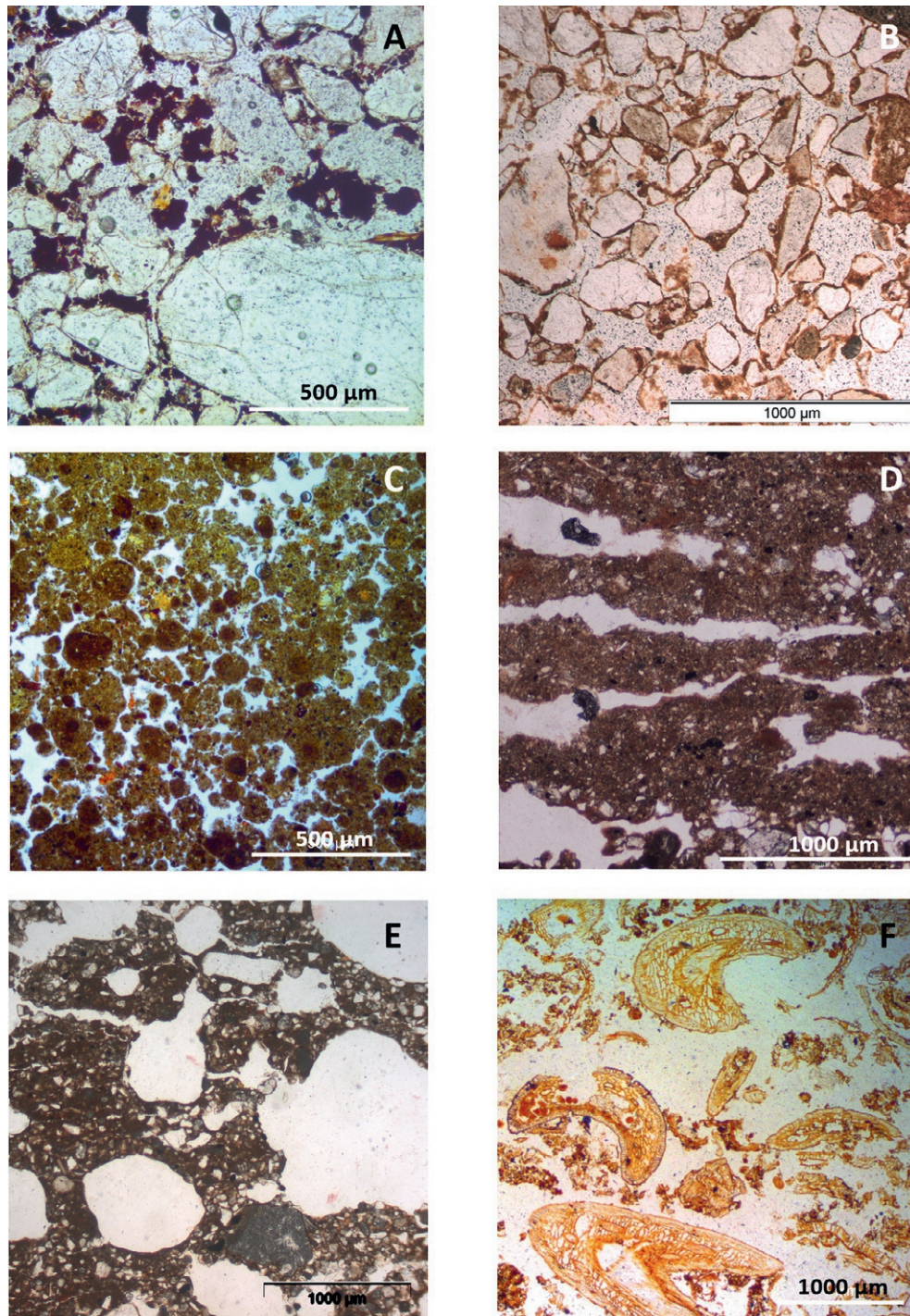


Fig. 2 Coarse and fine related distribution patterns (c/f r.d.p.) and microstructures. (A) Intergrain microaggregate microstructure (fine enaulic c/f r.d.p.). B_{21h} horizon of sandy podzol, Campine area, Belgium. (B) Pellicular grain microstructure (chitonic c/f r.d.p.) in sandy soil, Greece. (C) Granular microstructure in soil with andic properties, Isla Santa Cruz, Galápagos. (D) Lenticular microstructure in a soil on calcareous fine colluvium. Inside the ped a porphyric c/f related distribution pattern, Catalonia. (E) Vesicular microstructure in surface layer of arid soil, Israel. (F) loose packing of conifer needles. Note the presence, inside some needles, of small, orange homogeneous ellipsoidal excrements. Litter layer under dune forest, Belgian coast.

Pedofeatures

Pedofeatures are subdivided according to their morphology, either related to surfaces or voids, or unrelated. The first group (Fig. 2) comprises 'coatings' and 'infillings' and the second includes (Fig. 3) 'nodules,' 'crystals' and 'crystal intergrowths,' 'intercallations' and 'excrements.' A second criterion distinguishes between 'matrix pedofeatures,' visible as an overprint over the groundmass (e.g., diffuse iron oxide impregnations, spots of calcite depletion) and 'intrusive pedofeatures' having no characteristics of the groundmass.

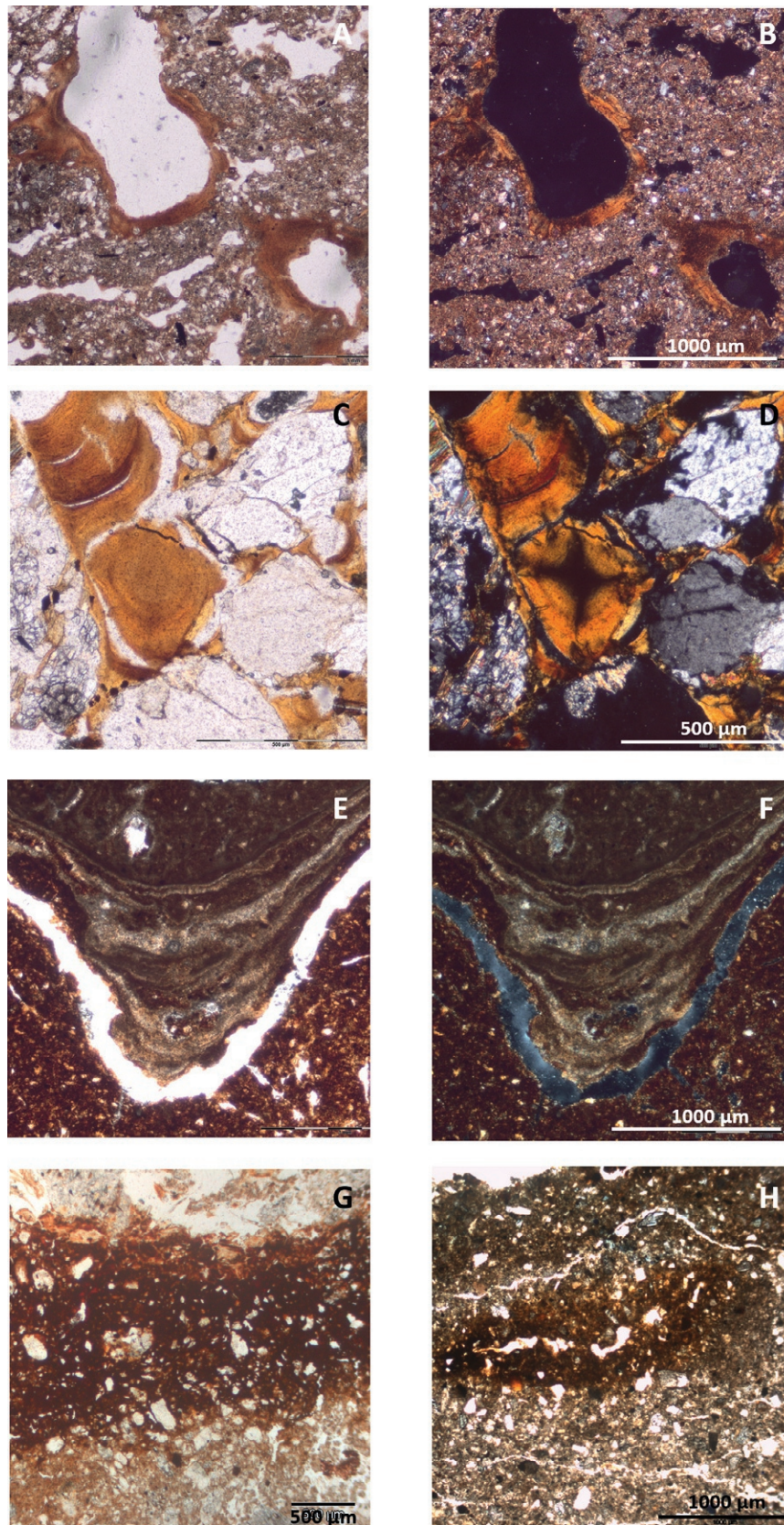


Fig. 3 Coatings and infillings. (A) Illuvial coating of coarse clay in channel, soil on calcareous fine colluvium, Catalonia. PPL (parallel polars). (B) Same in XPL (crossed polars). Note weak continuous orientation of clay in coating, and calcitic crystallitic b-fabric in the groundmass. (C) Complete dense infilling by fine, limpid illuvial clay with crescent pattern. Bt horizon on reworked till from weathered granite, Pyrenees. PPL. (D) Same, XPL. Note interference colors due to parallel orientation of the fine clay particles, and extinction lines parallel to polarizer and analyzer. (E) Multi-layered calcite pendent below a limestone fragment. Bk horizon in a doline, Mallorca. PPL. (F) Same in XPL. Note calcitic crystallitic b-fabric in groundmass. (G) Iron oxide pan resulting from the impregnation of the groundmass with reddish brown iron compounds. Bir horizon, Ireland. (H) Iron oxide hypoc coating on planar void in young volcanic ash deposit, Indonesia. Note presence of several parallel planar voids.

The features related to surfaces are subdivided into three groups: (i) 'coatings': intrusive pedofeatures occurring on the surface outside the groundmass (e.g., clay, calcite or goethite coatings); (ii) 'hypocoatings': matrix pedofeatures occurring directly on a surface, but inside the groundmass (e.g., microcrystalline calcite surrounding a root channel, iron depletion on a ped surface); (iii) 'quasicoatings': matrix pedofeatures related to a surface, but occurring at some distance from it (compare with Liesegang rings). According to their morphology, several types can be distinguished, e.g., typic, cappings (covering a larger constituent), pendants (underneath a larger constituent), crusts, pans. 'Infillings' are intrusive pedofeatures in void spaces. They are subdivided into 'loose continuous,' 'loose discontinuous' (e.g., excrements), 'dense complete' (e.g., illuvial clay) and 'dense incomplete' (e.g., calcite crystallization). Fig. 3 illustrates several types of infillings and coatings.

Pedofeatures not related to surfaces or voids can be subdivided in three groups (i) 'crystals and crystal intergrowths': these are euhedral or subhedral pedogenic crystals in the groundmass (e.g., lenticular gypsum in desert roses, cubic pyrite, vivianite); (ii) 'nodules': discrete bodies that can be intrusive or matrix features that can be either sharp or diffuse; they are subdivided according their internal fabric into 'typic' (undifferentiated), 'concentric,' 'nucleic' (with central core), 'geodic' (with a central void), 'septaric' (with concentric and radial cracks), 'aggregate' and 'dendritic' (composed of smaller nodules) and 'alteromorphic' (displaying the fabric of a weathered mineral or rock fragment); (iii) 'intercallations': elongate, undulating pedofeatures unrelated to natural surfaces or voids, not consisting of crystals. Fig. 4 shows some microphotographs of pedofeatures not related to surfaces or voids.

The last type of pedofeatures are 'Excrements' (mainly of the mesofauna), subdivided according to their shape, such as spheres, ellipsoids, cylinders, tuberoses or mamillated. Their description also comprises their degree of disintegration or coalescence and aging.

'Compound pedofeatures' are a combination of two or more features that can be 'juxtaposed' (e.g., a calcite coating over a clay coating) or superimposed (e.g., an Fe-oxide nodule developed in a clay coating). Pedofeatures often are fragmented (e.g., by freezing), deformed (e.g., by vertic movements) or partly dissolved.

Applications

The wide range of applications of soil micromorphology in earth sciences is variable in time. It is shifting from mainly pedogenesis and classification till the 1980s, to palaeopedology, quaternary geology and geoarchaeology last decennia. Also its practical application in agronomy is getting more important. Fig. 5 shows the evolution of the topics treated in papers containing micromorphology. For more detail, see [Stoops \(2014\)](#).

Soil genesis and paleopedology

Since its beginning, most of the research using soil micromorphology techniques has aimed to describe and understand the processes of soil and regolith formation. This results from the main paradigm of soil science that the morphological characteristics of the soils we observe today are mainly the result of particular soil formation processes that depend on the combination of various factors of soil formation. Therefore, the observation of their constituents and of their spatial relationships at a microscopic scale will provide information about the transformation, addition, removal, and translocation of soil constituents in the soil during its formation.

In most of the cases it is not possible to establish direct relationships between processes of soil formation and micromorphological features. For example, crumb microstructures can originate from faunal activity, but also by frost action. Likewise, silt cappings and coatings are characteristic of certain natric horizons, but they are also found in buried plough layers. Nevertheless, it is possible to assign sets of characteristic features to the outcomes of given soil formation processes, sometimes with the confirmation of additional laboratory analyses. With increasing research these relations have reached a consensus for some kinds of soil materials, while for others there is still some controversy. A compilation of the state-of-art interpretation of micromorphology for soil genesis is given in [Stoops et al. \(2018\)](#). Some of these processes with a summary of their resulting characteristics are listed below.

Pedogenic formation processes in arid soils: calcium carbonate, gypsum and salt dynamics

In contrast to other analytical techniques, micromorphology enables the distinction between lithogenic (inherited) and pedogenic (neoformed) constituents. The accumulation of pedogenic calcium carbonate, gypsum and salts occurs under arid climates because these minerals are not leached out of the profile. It results in calcitic, gypsic or saline pedofeatures, for example as nodules, coatings or infillings, which can be distinguished from the calcite or gypsum as lithogenic groundmass constituents. In the later stages of development, the maximum expression of this process is their generalized accumulation, with calcite, gypsum or soluble salt minerals becoming the main constituents of the groundmass. Additional characteristics, such as biocalcification, carbonate-depletion features, gypsum-mouldic voids or the presence of companion minerals like celestite are indicators of specific soil formation processes.

Processes in acid sulfate soils

Soils containing sulfidic or sulfuric materials are common in some coastal lowlands as well as in specific inland environments. Their dynamics result from alternating wetting and drying, water table oscillation and the mixing of continental and sea water, as well as

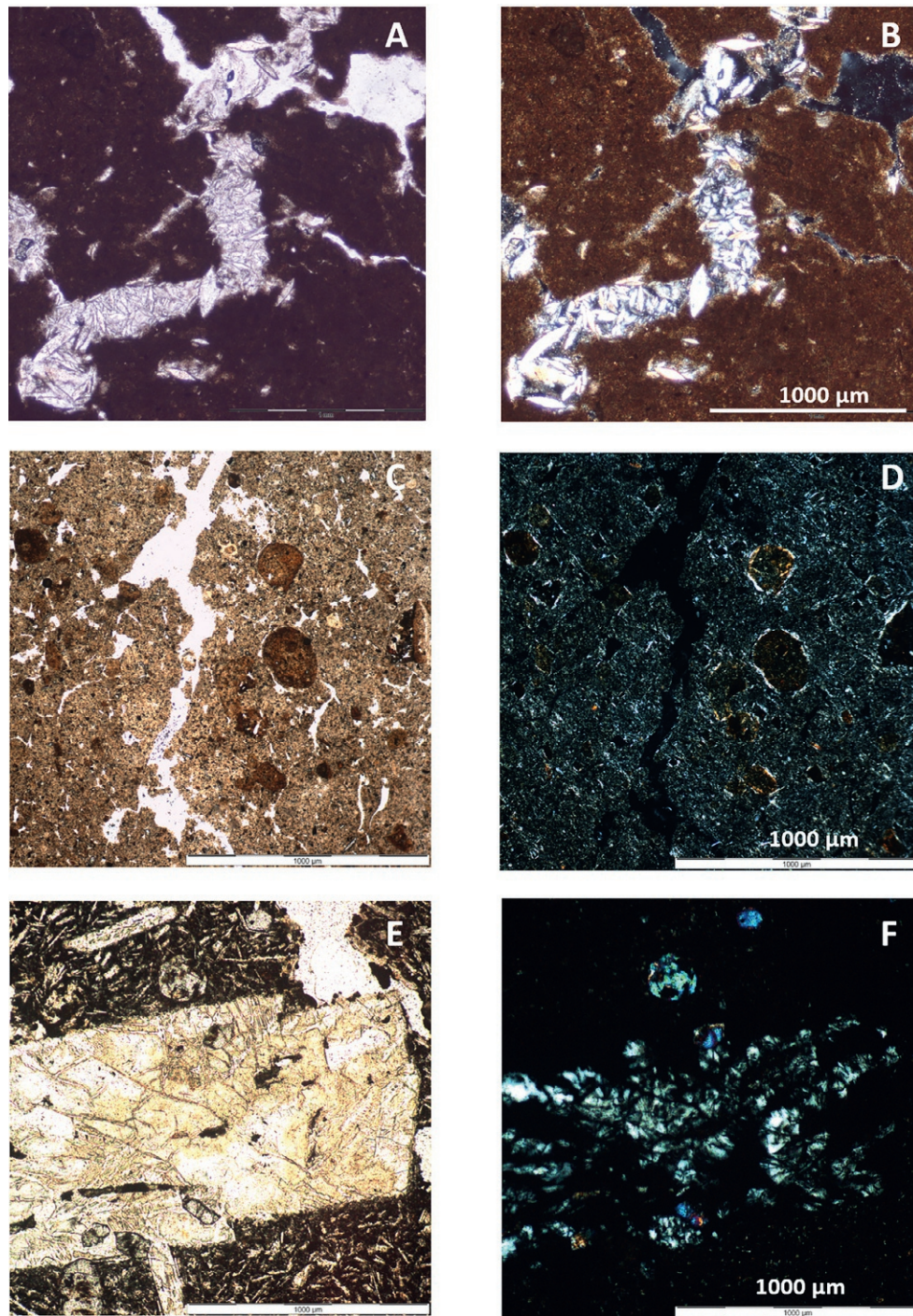


Fig. 4 Crystals and nodules. (A) Loose continuous lenticular gypsum infillings in a soil on alluvial fan deposits near Lleida, Catalonia. PPL. (B) Same in XPL. Note calcitic crystallitic b-fabric in groundmass. (C) Iron nodules in groundmass with porphyric c/f related distribution pattern. Clayey soil on basalt, Galápagos. PPL. (D) Same in XPL. Note the granostriated b-fabric around nodules, created by parallel orientation of the clay particles due to pressure. (E) Plagioclase in basaltic scoria altered to Si-Al colloids. Soil with andic properties, Galápagos. (F) Same, XPL. It is clear that in the phenocryst the colloid is partly crystallized to gibbsite, not in the microlites.

anthropic changes in drainage. The processes involve redox transformations of Fe and S, dissolution of Ca-bearing constituents, and the neoformation of minerals such as pyrite, jarosite, various iron oxides or gypsum, among other changes. Micromorphology enables the identification of these transformations and constituents, and of their spatial relations to organic matter and to the pore system, which is not possible with other techniques.

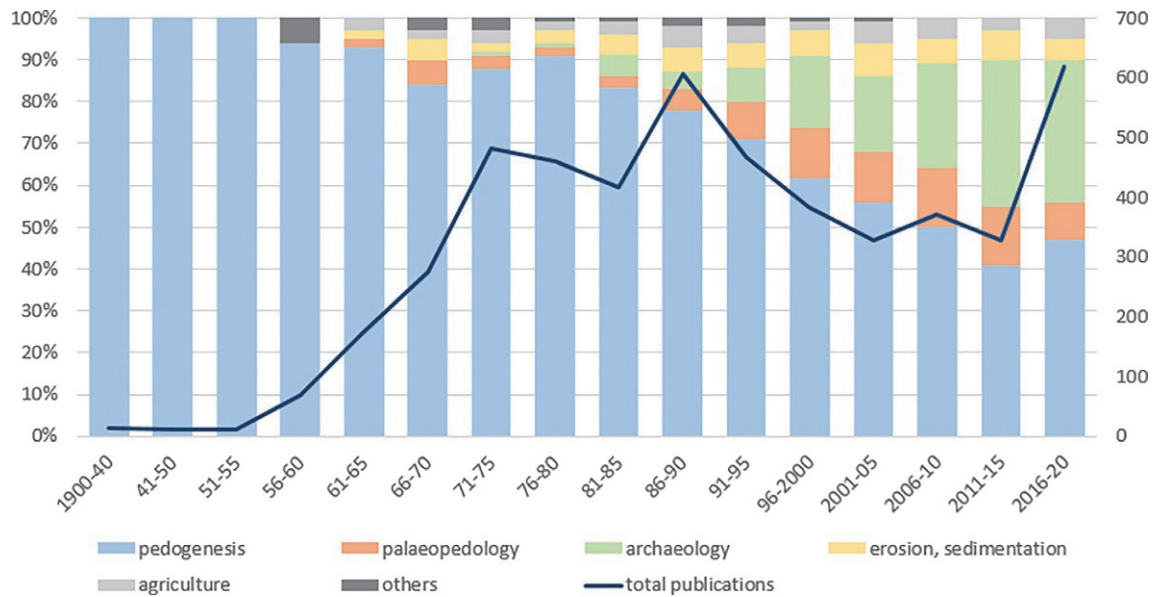


Fig. 5 Evolution of the application fields of micromorphology expressed in % and/or the total number of publications (partly based on the database of Stoops, 2014).

Clay illuviation

This process was the first to be identified through micromorphology, because the main characteristic feature of this process, microlaminated clay coatings, is easily identified in thin section. Coatings of microlaminated clay around sand grains or along pore walls that originated from the deposition of illuviated clay show interference colors under crossed polars. They result from the continuous parallel orientation of the clay flakes, which is different from the unoriented clay in the groundmass. The micromorphology of clay illuviation can indicate different types of clay coatings resulting from different conditions, such as the clay size, sorting or layering within the coating. It also shows their spatial relations with the pore system, which indicates whether clay illuviation is active or relict, and whether they are altered due to changing conditions. Soil micromorphology enables us to distinguish coatings around pores or ped faces from pressure faces due to swelling and shrinking of clays, for example, which is not possible in field profile descriptions. It is also the only technique that can distinguish between clay coatings formed by physico-chemical processes, and those resulting from gravitational transport, such as flood coatings.

Redox processes

The identification of redoximorphic features is an important part of routine soil profile descriptions, and is a diagnostic feature of drainage conditions. The morphological changes due to the redox status of the soil constituents mainly affect Fe (iron) and Mn (manganese). Micromorphology can help to identify the neoformation of minerals containing these elements and their relation with the pore system, which can be alternately aerated or filled with water. There are many redox pedofeatures (e.g. different types of nodules, coatings, hypo- and quasi- coatings, Fe-depletion features), which either alone or in combination give information about the intensity of the redox process, whether it is active or relict, or about its temporal relation with other soil forming processes.

Bioturbation

The result of faunal activity in thin sections is seen either by the presence of specific types of pores due to animal excavation (channels, chambers), by passage (fabric) pedofeatures resulting from their circulation through the soil, by the different types of excrements, formed by mineral and/or organic soil constituents, or by specific types of aggregates such as crumb or granular microstructures. The expression of bioturbation is extremely variable in size, morphology and related distribution to organic residues depending on the taxa; its complete description can give information about their feeding and displacement habits. Micromorphology can determine the extent of bioturbation by the fauna responsible for porosity and structure, mostly macro- and meso-fauna, and also the degree of mixing between mineral and organic constituents. Although it is possible in only a few cases to assign some features unequivocally to a given taxon (e.g., some biocalcifications by earthworms, or herbivores), in most cases the possible organisms producing excrements with a specific size and morphology can be determined with some degree of precision.

Frost action

The formation of ice in soils, either present or past, imprints specific morphologies in soils mainly due to the volume increase of water when passing from liquid to solid, to the desiccation associated to ice segregation, and also to the change in drainage conditions when a complete horizon is frozen, either temporarily or permanently. Some microfabrics are specific to soils subjected to glacial or periglacial conditions, such as the lenticular microstructure produced by ice lensing, sand and or silt cappings with

inverted sorting, and some injection features or vertical alignment of coarse fragments. In other cases, it will be the combination of several pedofeatures that will serve as diagnostic of frost action, such as silt cappings on 'floating' coarse fragments (underlain by a pore), ice-fragmented constituents, vesicular porosity, grano- and poro-striated b-fabrics due to stress by increased ice volume or some non-biogenic granular structures. The identification of cryogenic microstructures as in cave sediments has been useful for geoarchaeologists and quaternarists to determine past cold climates.

Argilloturbation

Clay shrinking and swelling due to moisture changes in soils is expressed in the field by a characteristic structure and fissure pattern, and by slickensides, which are indicators of vertic properties and vertisols. The most common micromorphological features are angular blocky microstructures with well-accommodated (recent) planes, and striated b-fabrics (mainly poro-, grano-, cross-striated). Other additional features can be granular microstructures at the surface (corresponding to the self-mulching), evidence of rotation and fragmentation of coarse fragments, nodules and other pedofeatures, and incorporation of clay coatings in the micromass (e.g., in Mediterranean soils).

Podsolization

The translocation and accumulation of complexed organic matter, Fe and Al (aluminium) in soil is expressed by different morphologies depending on their position in the profile. Micromorphology can identify these processes when they are not evident in the field, mainly in spodic soil materials. The accumulation of amorphous (isotropic, opaque) organic matter with Fe and Al in a sand results in different types of c/f related distributions depending on whether the OM coats grains (chitonic), forms bridges or aggregates between them (gefuric or enaulic) or completely fills the spaces between them (porphyric). The internal fabric of the organic constituents (monomorphic or polymorphic) gives clues on their origin, either translocated or in situ. Because of the isotropic nature of organic matter, additional analyses are necessary for the complete characterization of spodic soil materials.

Formation of oxic or ferrallitic materials

These result from extreme weathering under moist, hot tropical conditions. Only the most resistant minerals (quartz, zircon) will survive, together with a variety of Fe and Al oxides and hydroxides with kaolinite as dominant clay. The most common micromorphological characteristics are a reddish-yellowish groundmass with a granular microstructure, an undifferentiated (due to Fe oxides) or stipple-speckled b-fabric, and the dominance of quartz in the coarse fraction, which may show incipient weathering stages (runiquartz).

Finally, paleopedology, defined as the discipline of soil genesis deals with soils formed in past environments, benefits considerably from micromorphology for identifying the soil forming processes that no longer relate to present conditions. Polygenetic soils, buried or not, formed under a succession of different soil forming factors and will show superposition of those features that are characteristic to each of them, as if they were medieval palimpsests. In this way it is possible to establish the relative chronology of the processes and possibly the sequence of soil forming factors that led to them. Together with the appropriate dating techniques, it is possible to reconstruct paleosedimentary records, especially in Quaternary formations. Loess-paleosol sequences are perhaps the best known examples of buried soils since the Late Tertiary, because the accretion of deposits and their homogeneous composition allow soil formation processes to be clearly expressed and differentiated at depth. The observed features result from the succession of cold/mild/warm and humid/dry climates together with the associated vegetation during the Quaternary. The features include clay illuviation, calcitic and gypsic pedofeatures, accumulation of organic matter in buried A horizons, rubefaction, and different types of cryogenic features the main paleoenvironmental markers.

Soil classification

With the older, natural soil classifications, such as that of Kubiěna (1953), there was a close link between soil types and their micromorphological aspects. This link was lost with the morpho-analytical thin section descriptions and current soil classification systems, mostly based on quantitative chemical, physical, morphological, mineralogical and climatic data. Therefore, soil type can no longer be derived from its micromorphology alone, because the relationships between the micromorphology and their overall classification are no longer possible because of the many characteristics considered when classifying them.

Nevertheless, the two international classification systems (Soil Taxonomy and the World Reference Base for Soil Resources) were initially conceived as morphogenetic systems, where morphology and micromorphology were used as markers for soil genesis. Micromorphology played a role in the development of the American Soil Taxonomy, especially for argillic, spodic and cambic horizons. Micromorphological features were in general not used as criteria, but as a guide for selecting criteria because of their value in determining the direction of pedogenic processes.

At present, micromorphology is only explicit in the definition of argillic (Soil Taxonomy, Soil Survey Staff, 2014), argic (World Reference Base for Soil Resources, IUSS Working Group WRB, 2015), and natric horizons (both systems), for which evidences is required of illuvial clay, i.e., oriented clay, clay coatings and or "thin sections with oriented clay bodies in more than 1% of the section" in both systems. Even with a hand lens in the field it is impossible to discern whether observed coatings are composed of clay or of coarser fractions, but micromorphology enables this and is the most used technique to diagnose clay illuviation for soil classification (Stoops et al. 2018, ch. 14 and early papers on soil micromorphology).

Some diagnostic soil characteristics for classification require the use of micromorphology, such as for the calcic horizon, defined as an “illuvial horizon, in which secondary calcium carbonate or other carbonates have accumulated to a significant extent”. Secondary carbonates are best identified and characterized in thin section, as calcitic pedofeatures, which can be distinguished from groundmass calcite belonging to the parent material (Stoops et al., 2018, ch. 9), whereas chemical analyses of carbonate content cannot differentiate between them.

Similarly, gypsic diagnostic horizons are defined as those horizons where “gypsum has been accumulated or transformed to a significant extent” in the classification systems. The identification of gypsic pedofeatures in thin section becomes an unequivocal proof, regardless of whether the quantitative requirements are fulfilled. The World Reference Base for Soil Resources (IUSS Working Group WRB, 2015) explains, in detail, the micromorphology of petrogypsic horizons in its additional characteristics.

Redoximorphic features, one of the indicators of aquic conditions, are morphological features that can be observed in the field, but it is in thin sections that they can be described in detail. Redox concentrations, redox depletions and reduced matrices can be characterized not only as presence or absence, but in relation to the pore system and biological activity, and also determine whether the redoximorphic features are currently active or relict (Stoops et al., 2018, ch. 15). Micromorphology can identify gleyic or stagnic redoximorphic patterns (IUSS Working Group WRB, 2015) since they are defined in relation to present porosity. Stoops et al. (2018, ch. 15), established seven degrees of redox conditions according to different combinations of redoximorphic pedofeatures observed in thin section.

The classification of organic layers as humus profiles relies mainly on morphological and laboratory data, but combined with their micromorphological characterization it is possible to assign sets of micromorphological features to given humus types (Zaiets and Poch, 2018).

Geomorphology and quaternary geology

The most important geomorphological processes studied using micromorphology are weathering, erosion and sediment deposition (for more detail see Stoops et al., 2018).

Micromorphological studies combined with chemical and mineralogical analyses allow a better understanding of present or past weathering processes and their relative chronology. Several features that can be detected only by micromorphology are the relation between porosity and weathering patterns, the development of new porosity (e.g., mouldic voids), or the successive phases of mineral alteration, including the transformation of earlier alteration products (e.g., feldspar → Si-Al-colloids → gibbsite).

Erosion is detectable in only a few cases by microscopic analysis (e.g., tillage erosion), but it can be indirectly deduced from the sedimentary deposits. In colluvial deposits laminated and non-laminated fabrics are distinguished. Mass-wasting deposits comprise solifluction deposits, landslides, debris-flow deposits and grain-flow deposits. Interpretation is based partly on micromorphological studies during field or laboratory experiments, such as artificial rain plots or wind tunnels. Their internal fabric provides information on the conditions under which the processes were active (e.g., frost) and the material transported (e.g., eroded soil). Locally, wind deposits are also important in landscape formation (e.g., sand and loess deposits) and can be detected in thin sections. Post depositional alteration may yield information on later processes including interruptions during which pedogenesis took place.

Since the early 1990s micromorphology has often been used in Quaternary geology. It has proved to be very efficient for the analysis of glacial and fluvioglacial deposits (see Van der Meer and Menzies, 2011). A terminology based on Brewer, (1964) is mostly used.

Agronomy

Agricultural practices that can leave an imprint on soil micromorphology are tillage, irrigation, addition of materials to increase soil fertility, burning, waste recycling, clearance and husbandry (Stoops et al., 2018, ch. 26). Some of the effects can be detected through morphological (field), chemical or physical analyses, but micromorphology can provide explanations on the soil's behavior, even when the processes are incipient or are not yet reflected in other properties (Stoops et al., 2018, ch. 1). These effects are mainly explained by (1) changes in microstructure, (2) presence or absence and location of some constituents (both as coarse or fine constituents and as pedofeatures), and (3) changes in the activity of soil biota.

One of the main questions in the interpretation of features due to agriculture is whether they are unequivocally due to it, or whether the same features can be caused by natural processes. This is the case, for example with silt cappings and coatings on pores, which can be result from both periglacial processes and disturbance by ploughing. Therefore, their interpretation must always be made with caution.

Tillage practices change the amount, type and distribution of pores and aggregates. The clearest effects of ploughing and tillage are a temporal increase of porosity in the topsoil, the formation of dense plough pans with laminar structure (Stoops et al., 2018, ch. 26) or compaction by traffic at the surface. While these effects can be quantified by physical methods, micromorphology also provides information about types and functionality of pores, and changes in type and size of aggregates or degree of pedality. For example, tilled soils compared with untilled, natural or forested soils may show changes from a crumb microstructure due to bioturbation to a subangular blocky structure because of a reduction in faunal activity. In frost-affected environments, the interaction between tillage and freeze-thaw processes was studied by Sveistrup et al. (2005) who reported an increase in bioturbation when forest changed to cropland.

Surface crusting and sealing is another area where micromorphology has contributed considerably to explaining the mechanisms of formation, because of the inherent difficulties in sampling for other types of analyses. Thin sections enable the distinction between structural and depositional crusts (Stoops et al., 2018, ch. 19), and can provide information about the causes of crust formation, such as the slaking of aggregates. Fragments of surface crusts and of silt coatings around pores formed after the collapse of aggregates has been used as a diagnostic feature for identifying buried Ap horizons.

Soil micromorphology can help to explain the effects of the application of fertilizers and amendments in soils from changes in aggregation and microstructure or an increase of faunal activity. The application of organic amendments as animal manure has a clear influence on soil microstructure, mainly with the increase of bioturbation (Yagüe et al., 2016). Chemical fertilizers may also have such an effect with the increase in nutrients or changes in pH, for example from liming.

Irrigation practices may change soil porosity patterns and create textural pedofeatures derived from the migration of fine particles with water after aggregate collapse or change in the composition of the exchange complex (Stoops et al., 2018, ch. 26). Redistribution of gypsum, carbonates and salts are other features commonly reported. Processes such as the silting of drains in irrigated sodic soils can be described and explained through thin section analysis. The analysis of redoximorphic pedofeatures in paddy soils may give information on the nature and intensity of redox processes.

Some particular soils with long-term anthropic influence have merited special attention among micromorphologists. For example, deeply disturbed soils and the effects of old tillage practices, on the Amazonian dark earths ('terras pretas do índio'), (Nicosia and Stoops, 2017, ch. 33); plaggen soils (Davidson & Carter, 1998; Múcher et al., 1990), bench-terraced soils in the Mediterranean (Boixadera et al. 2016), and soils on 'camellones' in the Bolivian Moxos plains (Boixadera et al. 2019).

Archaeology and geoarchaeology

Soil micromorphology offers a wide range of applications in archaeology (see Nicosia and Stoops, 2017) and its use by geoarchaeologists has increased since the end of the 1970s.

In paleoenvironmental studies micromorphology provides information on processes affecting archaeological sites (e.g., earthy constructions such as Tells), and on colluvial processes during and after occupation of the sites, often responsible for their burying. The materials filling manmade structures, such as ditches yield information on the processes involved and their relative chronology.

An important application is the reconstruction of past soil management practices, such as vegetation clearing and tillage, trampling and traffic effects and the formation of the so-called 'dark earths,' mainly studied in European towns and in South America.

The study of living or occupation floors, e.g., in houses or caves, yields information on activities such as metal working, manufacture of stone tools or butchery. Stable floors contain elements on the type of animals and their feeding habits.

One of the oldest applications of micromorphology in archaeology was the study of materials not commonly studied in natural soils, such as excrements of larger animals, bones, eggshells, phytoliths, charred material and plant ash and anthropogenic constituents such as plasters and mortars, ceramic materials (e.g., bricks, tiles and pottery) and metal working residues. The fabric study of man-made constituents helps in reconstructing the manufacturing processes, e.g., the preparation of mortars: what is the origin of the coarse material, was it sieved, what proportions coarse/fine were used, which type of limestone or shells were used to prepare the quick-lime, among other questions.

Forensic science

The application of soil science to forensic sciences, also referred as forensic soil science, is a fairly recent discipline that applies information related to the properties of soils in matters that may come before a court of law. It is commonly used to find evidence to link a suspect to a crime scene or to determine the state of decay (taphonomy) of any biological materials in the soil, including human remains.

A systematic approach to forensic soil examinations involves a thorough description and analysis of soil morphology (Munsell color, texture, structure, consistence), together with mineral and organic composition, for which micromorphological techniques are valuable (Fitzpatrick et al., 2009). The concept behind it is that soil composition can be used as a predictive tool to identify where that soil originated.

Soil science is very useful in the taphonomy of buried organic remains because the evolution of soil organic matter depends on soil characteristics (composition, drainage status, temperature, pH). For example, some researchers have tried to establish micromorphological indicators as presence or absence of certain compounds to characterize the time from the burials.

Conclusions and perspectives

There are several challenges in soil science where soil micromorphology has a great potential to contribute. One is the assessment of risks associated with new soil management techniques, such as the addition of residues as amendments. In particular, micromorphology can reveal interaction mechanisms of microplastics in soils, or the effects of the application of different types of sludges, how they evolve and how they interact with soil constituents or with soil microstructure.

Soils are important carbon sinks. In the framework of climate change, micromorphology can help to describe the dynamics of organic and inorganic carbon in soils, for example the evolution and transformation processes of organic matter, or the influence of irrigation on the dynamics of calcium carbonate as a source or sink of carbon.

We have some knowledge of the soil organisms, their functions and interrelations, but the interactions of soil biota with soil mineral constituents are still largely unknown. Experimental pedology, which attempts to reproduce the habitats for particular soil organisms under controlled conditions, is a promising research field for the interpretation of biotic features using micromorphological techniques.

As soil analytical methods evolve and new techniques are developed, there will remain a role for soil micromorphology in field and landscape surveys because soil morphology at any scale, including micromorphology, provides the spatial framework within which any further observations or analyses can be considered.

References

- Boixadera J, Riera S, Vila S, Esteban I, Albert RM, Llop JM, and Poch RM (2016) Buried A horizons in old bench terraces in les Garrigues (Catalonia). *Catena* 137: 635–650.
- Boixadera J, Esteban I, Albert RM, and Poch RM (2019) Anthropogenic soils from Llanos de Moxos (Bolivia): Soils from pre-Columbian raised fields. *Catena* 172: 21–39.
- Brewer R (1964) *Fabric and Mineral Analysis of Soils*. New York-London-Sydney: Wiley.
- Bullock P, Fedoroff N, Jongerius A, Stoops G, Tursina T, and Babel U (1985) *Handbook for Soil Thin Section Description*. Wolverhampton, UK: Waine Research Publications.
- Courty MA, Goldberg P, and Macphail R (1989) *Soils and Micromorphology in Archaeology*. Cambridge, UK: Cambridge University Press.
- Davidson DA and Carter SA (1998) Micromorphological evidence of past agricultural practices in cultivated soils: The impact of a traditional agricultural system on soils in Papa Stour, Shetland. *Journal of Archaeological Science* 25: 827–838.
- FAO/ISRIC/ISSS (1998) *World Reference Base for Soil Resources. World Soil Resources Report No. 84*. Rome: FAO. 88 pp.
- FitzPatrick EA (1984) *Micromorphology of Soils*. London, UK: Chapman and Hall.
- FitzPatrick EA (1993) *Soil Microscopy and Micromorphology*. Chichester, UK: John Wiley & Sons.
- Fitzpatrick RW, Raven MD, and Forrester ST (2009) A systematic approach to soil forensics: Criminal case studies involving transference from crime scene to forensic evidence. In: Ritz K, Dawson L, and Miller D (eds.) *Criminal and Environmental Soil Forensics*, pp. 105–127. Dordrecht: Springer.
- IUSS Working Group WRB (2015) World reference base for soil resources 2014, update 2015. International soil classification system for naming soils and creating legends for soil maps. In: *World Soil Resources Reports No. 106*. Rome: FAO.
- Kubiěna WL (1938) *Micropedology*. Ames, Iowa, USA: Collegiate Press.
- Kubiěna WL (1953) *The Soils of Europe*. London, UK: Thomas Murby & Co.
- Macphail RI and Goldberg P (2018) *Applied Soils and Micromorphology in Archaeology. Cambridge Manuals in Archaeology*. Cambridge, UK: Cambridge University Press.
- Mücher HJ, Slotboom RT, and ten Veen WJ (1990) Palynology and micromorphology of a man-made soil. A reconstruction of the agricultural history since late-medieval times of the Posteles in The Netherlands. *Catena* 17: 55–67.
- Murphy CP (1986) *Thin Section Preparation of Soils and Sediments*. Berkhamsted, UK: AB Academic Publishers.
- Nicosia C and Stoops G (eds.) (2017) *Archaeological Soil and Sediment Micromorphology*. Chichester, UK: John Wiley & Sons, Ltd.
- Soil Survey Staff (2014) *Keys to Soil Taxonomy*, 12th edn Washington, DC: USDA-Natural Resources Conservation Service.
- Stoops G (2003) *Guidelines for Analysis and Description of Soil and Regolith Thin Sections*. Madison, Wisconsin, USA: Soil Science Society of America.
- Stoops G (2009) Seventy Years “Micropedology” 1938–2008. The Past and Future. *Journal of Mountain Science* 6: 101–106.
- Stoops G (2014) The “fabric” of soil micromorphological research in the 20th century—A bibliometric analysis. *Geoderma* 2013: 193–202.
- Stoops G (2021) *Guidelines for Analysis and Description of Soil and Regolith Thin Sections*, 2nd edn Madison, Wisconsin, USA: Soil Science Society of America and Wiley.
- Stoops G, Marcelino V, and Mees F (eds.) (2010) *Interpretation of Micromorphological Features of Soils and Regoliths*. Amsterdam: Elsevier.
- Stoops G, Marcelino V, and Mees F (eds.) (2018) *Interpretation of Micromorphological Features of Soils and Regoliths*, 2nd edn., p. 982. Amsterdam: Elsevier.
- Sveistrup TE, Haraldsen TK, Langohr R, Marcelino V, and Kværner J (2005) Impact of land use and seasonal freezing on morphological and physical properties of silty Norwegian soils. *Soil and Tillage Research* 81: 39–56.
- Taina IA, Heck RJ, and Elliot TR (2008) Application of X-ray computed tomography to soil science: A literature review. *Canadian Journal of Soil Science* 88: 1–19.
- Van der Meer JJM and Menzies J (2011) The micromorphology of unconsolidated sediment. *Sedimentary Geology* 238: 213–232.
- Yagüe MR, Domingo-Olivé F, Bosch-Serra AD, Poch RM, and Boixadera J (2016) Dairy cattle manure effects on soil quality: Porosity, earthworms, aggregates and soil organic carbon fractions. *Land Degradation and Development* 27: 1753–1762.
- Zaiets O and Poch RM (2018) Use of micromorphology for humus characterization and classification in some mediterranean calcareous soils. *Applied Soil Ecology* 123: 672–681.

Relevant websites

<https://www.isric.org/explore/ISRIC-collections/micromulti> <https://www.iuss.org/newsroom/newsletters/soil-morphology-and-micromorphology-newsletters-commission-11/>—ISRIC.
<https://www.iuss.org/organisation-people/organisation/commissions/>—IUSS.

History and principles of soil classification

Maria Gerasimova and Dmitry Konyushkov, V.V. Dokuchaev Soil Science Institute, Russian Academy of Sciences, Moscow, Russia

© 2023 Elsevier Ltd. All rights reserved.

Introduction. Challenges of soil classification	185
Indigenous soil knowledge and first soil classification systems	186
Dokuchaev's immediate predecessors in the field of soil classification	187
Dokuchaev's paradigm, the genetic approach to soil classification	188
Post-Dokuchaev early period. Diversification of classification decisions and dissemination of Dokuchaev's ideas over the world	189
Post-Dokuchaev late period. Divergence of approaches. National systems. Authors' systems. Old and new Russian systems	191
Conclusions	195
References	196

Abstract

The chapter describes the first attempts in Antiquity to differentiate between lands with certain soils, followed by the major developments in Dokuchaev's time when soil itself became the focus of classification. Soil-forming agents, soil processes and properties were interrelated in his system, and this was the source of two trends in choosing priorities by his successors. There were many individual author's classification systems and the criteria ranged from soil physicochemical properties to soil profile morphology and soil genesis. Almost all modern soil classification systems have been greatly influenced by Dokuchaev. Quite different systems have been developed on this basis throughout the world and are discussed briefly.

Key points

- Need for soil classification increases, and a regular updating of existing systems is required.
- First attempts to classify soils for applied purposes appeared in Antiquity. Then in the 18–19th centuries, a new outburst of classification activity may be noted.
- Dokuchaev's paradigm was a milestone in pedology. His predecessors and followers created an array of concepts on soil classification.
- Diversity of natural systems and divergence in classification principles: priority of soil-forming factors versus soil properties.
- However, in all systems, pedogenetic processes were accounted for although in different ways: minimal as diagnostic criteria, maximal as differentiating taxa.
- Development of an International system started in 1970s.
- Russian soil classification system is a properties-oriented one and has inherited Dokuchaev's pedogenetic approaches.

Introduction. Challenges of soil classification

"The general problem of Classification... may be stated as follows: To provide that things shall be thought of in such groups, and those groups in such an order, as will best conduce to the remembrance and to the ascertainment of their laws.... Each science... forms its classification of things according to the properties which fall within its special cognizance, or of which it must take account in order to accomplish its peculiar practical end" (Mill, 1882, pp. 870–874). In other words, classification is a contrivance to organize knowledge in our mind.

Knowledge about soils has been accumulated by humans since prehistoric times, and various soil classification systems have been created over time. They are based on the data already acquired; therefore, these systems considered in the historical perspective reflect the evolution of soil science and also possibly the related sciences of biology, chemistry, and physics. For example, the father of the American *Soil Taxonomy* Guy Smith stated: "classification is a creation of man and is a reflection of the state of knowledge at that time and the uses that were intended at that time" (Smith, 1986, p. 111). There is a similar saying of the famous Austrian pedologist Walter von Kubiřna (1948): "Show me your classification, and I will tell you what the state of your science is."

Soil classification systems differ in their purposes. Some of them address particular applied problems, e.g., soil suitability for septic tank construction, or soil suitability for grain crop growing, or soil physical behavior (classification of soil textures). Such classification systems consider a relatively narrow range of soil properties pertinent to a given purpose. These are technical, or applied, or artificial (as defined by J.S. Mill) classifications. Other systems address a much broader challenge to understand relationships between different soil groups, to map their geographic distribution, and to make justified assumptions about the actual and potential functions and uses of soils. They were named *Natural classification systems* by J.S. Mill, and are now referred to as fundamental or 'basic systems', or taxonomies. These are mostly hierarchical systems.

By giving appropriate definitions and names to soil taxa, classification ensures a common language and common understanding of soils, and eases communication between specialists in soil science, as well as in different fields of learning and working for people from different countries. A good soil classification system should have a predictive power, i.e., it should make it possible to predict unknown soils, their properties and functions. It should also be a guide in the search for new knowledge by addressing controversial relationships and finding appropriate criteria to differentiate soils.

As knowledge on soils is continuously accumulating, soil classification systems cannot remain static; they should be regularly updated or completely revised. New facts and data that cannot be explained within the framework of the existing classification serve as impetus for its change: from minor adjustment to a radically new paradigm. The development of classification is an endless process with certain stages. According to Vladimir M. Fridland, the intervals between such stages should be 10–15 years.¹

Almost all modern soil classification systems have been greatly influenced by a revolutionary perception of soil shaped in the last quarter of the 19th century by the Russian geologist Vasilii V. Dokuchaev: “soil is an independent (individual, separate, of its own) natural body, the product (later, function) of soil-forming factors (climate, relief, biota, parent material, and time).” This was a radically new definition of soil, simple and yet very comprehensive; it provided an excellent basis for classifying soils in accordance with either the soil forming agents, or soil properties, or both.

Different soil classification systems have been developed on this basis in Russia and throughout the world and are briefly discussed below; similarities and differences in the approaches and structure of these systems are analyzed. It is argued that, although recent major soil classification systems used in different countries have much in common, direct correlation between them is impossible. The *World Reference Base for Soil Resources* (WRB), which was initially designed as an umbrella for national systems and a bridge between them, remains a partial solution to the current incompatibility of soil classification systems (IUSS Working Group WRB, 2015).

Thus, pedology is still dreaming about creating a new universal soil classification system. The need for such a system is fostered by new demands related to the growing awareness of the key role of soils to the solution of pressing environmental problems and in sustaining life on our planet.

Indigenous soil knowledge and first soil classification systems

People live on soil and cannot but pay attention to this substrate in the support of their lives. It is probable that ancient hunters and gatherers knew different physical states of soils (wet, dry, loose, hard, soft, sticky, stony, etc.) and types of environments (hilly and plain areas, flooded areas, lowlands and highlands, mires, forests, steppes, etc.). The Neolithic Revolution about 10,000 BP and the birth of agriculture in the area now known as the Middle East shifted the focus of soil cognition to soil fertility and soil suitability for crop growing. It was also important to predict the response of soil to various ameliorative measures, including application of manure and irrigation. Numerous local soil names appeared in different parts of the world; many of them are still in use and have entered modern national and international soil classification systems. A new science, ethnopedology, has emerged recently at the interface of cultural anthropology and soil science to study folk soil knowledge (Krasilnikov et al., 2009).

In this context, almost all (ancient) human cultures and natural religions all over the world have soils/earth as essential element of worship in general represented as feminine “deity” directly indicating fertility. The German term “Mutterboden” is a good example.

Vernacular soil names reflect specific features of soils in particular localities with a focus on soil color (dark, light, whitish, gray, red), soil texture (clay, sand, silt) and physical conditions (hard, loose, cemented, excessively moistened, dry, cold), soil salinity, specific environments (soils of moors, meadows, of pine, oak or spruce stands), and soil suitability for growing particular crops (wheat (*Triticum aestivum* L.), barley (*Hordeum vulgare* L.), or oats (*Avena sativa* L.)). The notion and nomenclature of soil types in the Russian soil classification systems are often derived from the local folk soil names, such as Chernozem, Podzol, Solonchak, Solonetz, Gley. However, the same local soil name (e.g., Chernozem or black earth, or Belozem or white earth) could mean different soils in different places, which often led to misunderstanding (Krasilnikov et al., 2009).

The diversity of local soil names and the obviously different productive capacities of soils necessitated a formal classification of soils for agricultural and land inventory purposes based on soil productivity. This challenge required a broader geographic extent of soil examination that had already been realized in large ancient empires. One of the oldest systems was developed more than 4000 years ago in China, during the Yao dynasty (2357–2261 BCE) (Lee, 1921). The best grade of soils were the yellow, soft soils of Yung Chow (Shensi and Kansu provinces, semiarid to arid Northwest China), whereas the next best were the red, rich clayey soils of Su Chow (Shantung, Kiangsu, and Anhwei provinces, humid East China). The five-colored “sacrificial altar” established in 1421 CE, during the Ming dynasty, at Zhongshan Park in Beijing demonstrates clear knowledge of the geographic distribution of soils in China: the central yellow color is characteristic of loess soils on the Loess Plateau, blue color to the east corresponds to gley, black color to the north denominates chernozems in northeastern China, white color to the west signifies the development of saline and light-colored desert soils, and red color to the south is typical of tropical Ferralsols (Gong et al., 2003; Krupenikov, 1992).

In Ancient Greece, soil was one of the basic philosophical categories—basic simple elements along with fire, water, and air for Empedocles (494–434 BCE) and Aristotle (384–322 BCE). These elements were shaped from the same prime matter by a spirit endowed with reason and were carriers of active (heat and cold) and passive (dry and wet) qualities. For Theophrastus (371–287

¹The American system is a good example: its main issues were in: 1954, 1960, 1975, 1999, 2014.

BCE), who was named the father of botany by Carl Linnaeus (1707–1778), soil was a more “earthy” matter. He made a distinction between soils of plains, mountains, valleys, seashores; differentiated soils by color and thickness; paid attention to soil textures; and considered soil suitability for particular crops (Strzemska, 1975).

In the Roman empire, a classification of arable soils based on farming utility with nine major classes and 21 subclasses was described by Marcus Porcius Cato (234–149 BCE) in his book *De Agri Cultura*. As noted by M. Strzemska (1975), the works of Roman agriculturalists Cato, Marcus Terentius Varro (116–27 BCE), Lucius Junius Moderatus Columella (4–70 CE), and Gaius Plinius Secundus (23/24–79 CE) were “the pillars of soil science not only in the antiquity but also in the Middle Ages and early modern times.” Thus, Cato considered the suitability of soils for farming and their productive potential, Varro paid attention to the physical composition (texture, stoniness) of soils, Columella focused on the physical properties of soils, and Plinius regarded minerals and rocks as soil-forming materials.

As shown by B. Williams (2006), good knowledge of soils was also gained by the Aztecs. Many other interesting stories about the accumulation of soil knowledge over the millennia throughout the world are collected in the same book edited by B. Warkentin.

The European Middle Ages and Renaissance periods were marked by growing knowledge of soils and methods of soil management (Krupenikov, 1992; Strzemska, 1975). However, no considerable new concepts in the field of soil classification were achieved. An important innovation noted by M. Strzemska was the division of soils into (a) natural (vulgar) and (b) artificial (precious) by W. Folkingham in his *Feudigraphia* (1610) with a further subdivision of natural soils into simple and composite according to their main constituents.

Dokuchaev's immediate predecessors in the field of soil classification

‘A new wave’ in the study of soils, acquisition of new knowledge on them, laboratory experiments, and extensive geographic exploration of little known lands arose in Europe in the 18th–19th centuries. It was greatly influenced by the advances in chemistry. Fruitful debates on the major sources of plant nutrition, either organic (humus) or mineral, and the development of methods to determine the contents of different humus substances and the chemical elements in soils provided abundant data. The German agronomist Albrecht Thaer (1752–1828) considered soil as an unconsolidated substrate comprising mineral substances and humus. His agronomic classification of soils (1810) was based on their division into classes according to texture and mineralogy with an indication of a soil’s suitability for major crops (wheat, barley, oats). It was widely applied in European countries. In particular, it was used in questionnaires sent to landowners and local authorities, on the basis of which the first soil maps of European Russia were produced in the middle of the 19th century. In the soil map compiled by Vasilii Chaslavsky (1834–1878) in 1875 and based on data from local sources, the legend included 32 soils.

After the sudden death of Chaslavsky (in 1878), Vasilii Dokuchaev (1846–1903) was entrusted to prepare explanatory notes for this map. This was realized in a fundamental Dokuchaev’s publication *Cartography of Russian Soils* (1879), in which the framework of the new concept of soil was suggested. Dokuchaev emphasized the complicated and in some cases chaotic pattern of soil distribution on the map, although zonal regularities could be traced, and he argued that a new soil classification system was needed to deal with the real soil diversity.

In his review of major soil classification systems (1886), Dokuchaev analyzed the “chemical” classification of soils by the German agronomist Wilhelm Knopp (1817–1891) with a major division into silicate, calcareous, and sulfate soils and the “petrographic” classification by the German lawyer and tax assessor Friedrich Albert Fallou (1794–1877), who also coined the term “pedology”. Fallou stated in his monograph (1862) that soil is the product of mineral and rock weathering. He analyzed manifestations of weathering processes in different kinds of rock. At the highest level, soils were classified into primary (sedentary) and transferred soils with further division into petrographic genera (soils of quartz, clay, mica, feldspar, marly rocks, etc.). In 1882, Ferdinand von Richthofen (1833–1905) suggested a somewhat different system with the division into eluvial (primary) and transferred soils at the highest level and to different kinds of parent materials at the second level. The distinction between in situ (sedentary, eluvial, non-transferred, primary) and transferred soils (sediments) was most important for the German geological school of thought in soil science.

The agrogeological approach to soil classification advanced by German scientists quickly spread throughout the world in the 1860s–1890s. In the United States, this spread was greatly facilitated by the fact that the first work on soils was conducted by geologists. In a review by George Coffey (1875–1967), the reports by Amos Eaton (1776–1842) and Theodric Romeyn Beck (1791–1855) on the soils of Albany county (1820) and then on Rensselaer County, New York (1822) and districts adjoining the Erie Canal (1824) are mentioned as the first attempt to classify the soils of the United States with a major distinction between unmoved soil and soil conveyed by water (Coffey, 1911).

Perhaps, the most significant contribution to shaping American soil science in that period belongs to Eugene Hilgard (1833–1916). Being educated in Europe and investigating geological conditions, topography, and agriculture in Mississippi (1860) and then in Michigan, California, northwestern states, and the Pacific Coast, Hilgard indicated the relationships between soils and climate. He understood the difference in the nature of soil formation in arid and humid regions, and explained the appearance of salinization in arid regions (Hilgard, 1892). In his views on soil classification, Hilgard followed the agrogeological approach separating soils into divisions of (1) sedentary (residual) and (2) transported soils; in the latter, he distinguished between colluvial and alluvial soils and assumed the presence of various intergrades between these classes.

Another characteristic feature of the American approach to soil classification in that period was a keen attention to soil texture and related physical properties. Their importance for agriculture was emphasized by Milton Whitney (1860–1927) in his paper of

1892. In 1890, regular soil surveys on a large scale were initiated by the Division of Soils (in 1901–1927, Bureau of Soils) headed by Whitney. Initially, the soils of surveyed areas were grouped into soil types based on texture. Then, the concept of soil series as a set of soil types derived from the same kind of geologic material came into force. In the first national soil map (Whitney, 1909), soil series were grouped into 14 provinces based on geology and climate (Simonson, 1989; Brevik and Hartemink, 2012).

Thus, the agrogeological approach to soil classification predominated in the world at the turn of 20th century; it was partly due to the international activity of German scientists. Max Fesca (1846–1917) from Germany worked as a foreign advisor to the Meiji Government in Japan in 1882–1894 and conducted the first soil surveys there; parent material, texture, and the organic matter content were the main criteria for differentiating soil units (Hirai and Hamazaki, 2004). Ferdinand von Richthofen brought the German experience to China. The first two international soil science congresses in Bucharest (1909) and Stockholm (1910) were named Agrogeological congresses.

Dokuchaev's paradigm, the genetic approach to soil classification

As noted in the review of existing soil classification systems by Dokuchaev (1886), these systems: (i) consider soil from different viewpoints (chemical, physical, geological, agricultural) instead of giving the definition of soil per se as an individual natural body; (ii) do not pay attention to the fact that many soil properties are tightly linked to particular climatic, relief, geological, and vegetative conditions; and (iii) give preference to some artificially selected soil properties instead of classifying soils on the basis of the integrity of properties. He criticized the focus of the agronomic approach to soils on the upper fertile layer only, as well as the focus of agrogeologists on the parent material.

Dokuchaev suggested that a natural scientific classification of soils as individual natural bodies should be based on their genesis as related to soil-forming factors. He advanced the concept of the soil profile as the A–B–C system of horizons. His ideas predetermined the further development of soil classification in Russia and abroad. Initially (1879), Dokuchaev subdivided the universe of soils into two groups: (Group A) 'normal soils': soils formed in situ, and (Group B) 'abnormal soils': soils beyond the place of their origin. This division was similar to that suggested by agrogeologists. However, further division was based on the pedogenic properties. Normal soils were subdivided into classes of terrestrial 'vegetative' soils (soils developing under normal aeration conditions and not affected by groundwater) and terrestrial 'swampy' soils (soils developing under poor aeration conditions; soils affected by groundwater). Abnormal soils were subdivided into classes of outwashed (eroded) and inwashed (aggraded, sedimentary) soils. Several soil types with different properties were separated within the class of terrestrial vegetative soils. In 1886, Dokuchaev modified this system and added the group of transitional soils (Table 1).

Thus, at the highest level, Dokuchaev paid attention to the manner of soil occurrence, namely, sedentary or translocated soils, which depended on the relationships between pedogenesis and geodynamic processes. At the second level, soil hydromorphism was taken into account for normal soils and the trend of geodynamic processes for transitional and abnormal ones. At the third level, the nature of pedogenesis manifested itself in soil color, morphological and chemical soil features served as differentiating criteria.

In 1899, Dokuchaev forwarded the idea of natural horizontal and vertical zonality in the world of soils. It was a brilliant geographical concept. The concept of natural zonal regularities in the pedosphere was very attractive for many specialists in the Earth sciences. Dokuchaev's disciple Nikolay Sibirtsev (1860–1900) introduced the zonal perception of soils into the classification scheme. At the highest level, he divided all soils into classes: (i) 'zonal soils', or full-profile mature fine-earth soils forming continuous zones on the Earth's surface and developing under the impact of zonal (climate and vegetation) soil-forming agents; (ii) 'intrazonal soils' that appear within soil zones due to the effect of some specific local factors, such as parent material; and (iii) 'azonal', or immature soils with profiles that are not fully developed.

The following types of zonal soils formed the system of Sibirtsev (1895, 1900): (1) Lateritic (red earths); (2) Aeolian—dust (loess); (3) Desert-Steppe; (4) Chernozemic; (5) Gray Forest; (6) Soddy-Podzolic; and (7) Tundra soils. Intrazonal soils comprised

Table 1 Upper levels of the Dokuchaev's system, 1886.

<i>Group A. Normal soils</i>	<i>Group B. Transitional soils</i>	<i>Group C. Abnormal soils</i>
Class 1. Terrestrial "vegetative" soils: a. Light-gray northern soils b. Gray soils (forest soils) c. Chernozemic soils d. Chestnut soils e. Brown solonchic soils Class 2. Terrestrial "swampy" soils (meadow soils) Class 3. Typical swampy (bog) soils	Class 4. Outwashed (eroded) soils Class 5. Partly inwashed (aggraded) soils	Class 6. Aggraded soils (alluvial or aeolian sediments)

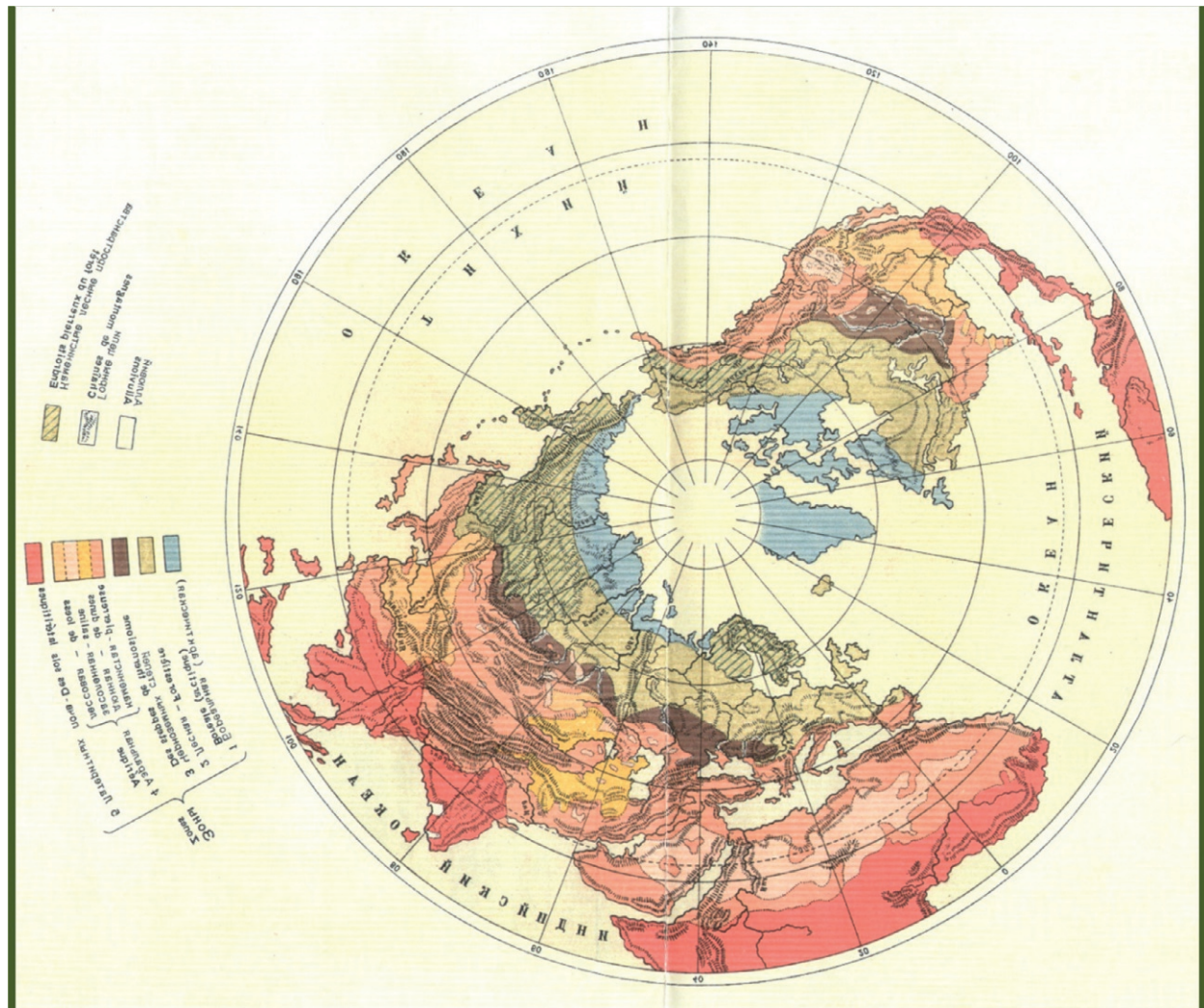


Fig. 1 Schematic map of soil zones of the northern hemisphere.

the types of: (1) Solonchaks; (2) Bog (Peat); and (3) Mucky-Calcareous (Rendzic) soils. Azonal soils were separated into subclasses of non-riparian soils with the types of skeletal and coarse-textured (dune) soils and floodplain soils with the type of Alluvial soils. Thus, at the highest level, this classification puts emphasis on soil-forming factors (mainly climate) and on the geographic pattern of soil distribution rather than on soils themselves. It can be referred to as the first factor-genetic system, and this principle was inherited by many soil classifications in Russia. The next step was Dokuchaev's classification scheme for soils of the northern hemisphere with concise descriptions of typical zonal soil-forming conditions and characteristics of zonal pedogenetic processes. In this scheme, the initial concepts of normal, transitional, and abnormal soils were considered as synonyms for those of zonal, intrazonal, and azonal soils, respectively. This system was implemented in a map developed by Dokuchaev for the World Exhibition in Paris in 1900 (Fig. 1) and became the 'trademark' of Russian small-scale soil mapping.

Post-Dokuchaev early period. Diversification of classification decisions and dissemination of Dokuchaev's ideas over the world

The first decades of the 20th century were marked by an impressive diversification of soil classification decisions in Russia. In fact, almost all prominent Russian pedologists suggested their own soil classification systems. Some ambiguity in the factor-genetic approach to soil classification was quite obvious to the followers of Dokuchaev. There were several attempts to make the classification "closer to the soil", i.e., to basic classification derived from soil properties rather than based on soil-forming factors.

Konstantin Glinka (1867–1927), the leader of soil geography in Russia, in his first textbook (1903–1904), had stressed that the concepts of zonal, intrazonal, and azonal soils should be considered as geographical rather than classification concepts. He noted that some isolated areas of Chernozem soils could occur in the forest zone and, in this case, should be placed into the category of intrazonal soils. Thus, the same soil (Chernozem) can be found in two different groups of soils (zonal and intrazonal) at the higher level of ‘zonal’ classification, which is inconsistent with the rules of any classification. Glinka was the first to revive J.S. Mill’s ideas on the theory of classification and introduce them into pedology. The German edition of Glinka’s textbook in 1914 (Glinka, 1914) and its English translation by Curtis Marbut (1863–1935) published in the United States in 1927 greatly facilitated the worldwide dissemination of Dokuchaev’s ideas.

In 1903–1908, Glinka proposed the subdivision of soils by soil-forming agents into ectomorphic soils, whose features are specified by external (climatic) conditions, and endomorphic soils, whose features are very strongly influenced by the properties of the parent material. In his latest publications (1921–1927) Glinka placed the types of pedogenesis at the highest taxonomic level: Lateritic, Podzolic, Steppe, Bog, and Solonetzic types of pedogenesis; they were further subdivided into 25 soil types. This was the first profile-genetic classification.

Georgiy Vysotskiy (1865–1940) tried to improve Sibirtsev’s classification by introducing the factors of relief and parent materials. He suggested that zonal soils should be defined as soils of particular climatic zones developing on flat surfaces composed of loamy sediments. Intrazonal soils were subdivided into subclasses: (i) soils that become zonal in neighboring climatic zones; (ii) absolutely intrazonal soils (affected by groundwater or by surface water that has stagnated in depressions); (iii) skeletal soils. Azonal immature soils were subdivided into subclasses of denuded (eroded) and aggraded (accumulative, sedimentary) soils. Vysotskiy (1906) distinguished between the soils of flat well-drained warm sites (soils of southern slopes that become zonal in a more arid climate) and poorly drained cold sites (soils of northern slopes that become zonal in a more humid climate). This was a factor-genetic classification with emphasis on topography and climate.

The idea of analogous soil series under different climatic conditions was developed by Yakov Afanasiev (1887–1937). Afanasiev’s classification was based on the idea that analogous combinations of parent rocks, topography, and vegetation can occur in different climatic zones, resulting in the development of analogous series of soils in these zones. Climatic zones were differentiated by temperature (cold, temperate cold, temperate warm, subtropical, and tropical) and by climate continentality. The resulting cells of the classification table were then subdivided into columns with different vegetation types (forest, forest-meadow, grasses) and rows presenting different kinds of parent material (silicate, calcareous, saline). It was a climate-oriented factor–genetic classification. The English translation of Afanasiev’s paper on the Russian soil classifications (Afanasiev, 1927) was an important event because it made it possible for English-speaking researchers to be aware of novel Russian approaches to classifying soils.

Sergey Neustruev (1874–1928) developed an original classification of elementary pedogenetic processes; he also specified processes as automorphic and hydromorphic (affected by groundwater) at the highest level. Automorphic processes were subdivided into three classes by the intensity of transformation of the mineral mass of soils (strong, moderate, and weak). At the third level, the fate of residual products was considered. Hydromorphic processes were subdivided into subgroups with (1) the accumulation of precipitates from the groundwater and (2) the stagnation of water and development of anaerobic reducing conditions. This was a process-oriented substantive-genetic classification.

Konstantin Gedroits (1872–1932) tried to substantiate the separation of soils based on the properties of the soil adsorption complex (SAC). At the highest level, he distinguished between the groups of base-saturated soils (without H^+ ions) and unsaturated soils (with H^+ ions). The major types of pedogenesis (Chernozemic, Solonchakous, Solonetzic, Solodic, Podzolic, and Lateritic) were characterized with respect to the composition of exchange cations, the state of the SAC, and the redistribution of pedogenetic products in the soil profile (Gedroits, 1925). Gedroits believed that the further development of physicochemical studies would make it possible to expand his system, which can be referred to as a physicochemical substantive-genetic classification.

Petr Kossovich (1862–1915) and Boris Polynov (1877–1952) combined the substantive-genetic principle of soil classification with geochemical ideas. Kossovich subdivided soils at the highest level into classes of geochemically independent (autonomous) and dependent (subordinate, heteronomous) soils and characterized seven autonomous types of pedogenesis and four types of genetically related subordinate pedogenesis (Kossowitsch, 1911). Polynov (1933) emphasized the idea of consecutive stages of soil evolution in both classes. Autonomous (eluvial, automorphic) soils were subdivided into the groups of alkaline carbonate-containing soils and acidic soils. The types of pedogenesis within these groups were believed to be linked by evolution. For example, the group of alkaline eluvial types of pedogenesis included Primitive Alkaline, Pre-Chernozemic, and Chernozemic. Later, Polynov linked soil development to evolutionary stages of residual and accumulative crusts of weathering (Polynov, 1936). This was an evolutionary geochemical-genetic soil classification.

The zonal paradigm was strongly supported by American soil scientists in their ‘middle’ Curtis Marbut period, although the uppermost category in Marbut’s classification was properties-oriented—Pedalfers and Pedocals, meaning soils dominated by the behavior of iron compounds and those rich in calcium carbonates, respectively (Marbut, 1935). The next, most familiar American system introduced by Baldwin et al. (1938), was zonal with three classes at the highest level—zonal, intrazonal, azonal, thus repeating the early Dokuchaev–Sibirtsev scheme. The 36 great soil groups, arranged in the zonal order comprised soils similar to the Russian ones, and they had the same names (solonetz, solonchak, brown forest, tundra, chestnut soil, etc.). Some soil names included landscape components (desert, prairie, meadow, Alpine, semi-bogged).

Despite this diversity of soil classification systems, none of them was detailed enough to meet the needs of large-scale soil surveys that started in the former Soviet Union in the 1930s. For the readers worldwide, these studies remained little known. However, as shown below, gradual dissemination of Dokuchaev’s genetic paradigm over the world occurred in that period.

Post-Dokuchaev late period. Divergence of approaches. National systems. Authors' systems. Old and new Russian systems

The second half of the 20th century may be named the 'Golden age' of soil classification: the impetus given by Dokuchaev and his successors continued to affect the development of classification activity in many countries. These were France, Germany, the United States, USSR, Canada, Great Britain, Australia, China, Brazil and some other countries. The possibility of dual interpretation of Dokuchaev's paradigm was the reason for developing the factor-genetic and properties-oriented approaches, and they were implemented in diverse classification systems in the 1960s–1990s. This was the period of active discussions of approaches triggered by the appearance of a markedly new American system in the form of several approximations and two 'Soil Taxonomy' books (1975, 1999), the development of national and individual authors' systems, and the birth of the first International soil classification—the legend to the FAO World Soil Map (1974)—transferred later (in 1998) into the World Reference Base (WRB) for Soil Resources.

National systems were created, updated, correlated, and their number increased. They were supported by extensive terrain and analytical research on soils, soil surveys, and interpreting their results for diverse purposes, both scientific and applied. The latter was more obvious in the American system, specifically in its early versions (1954, 1960) while in the European systems this trend was weaker. In the European and Soviet Union systems, priority was given to soil genesis, soil-forming processes and factors; of less importance were soil properties.

The German pedological school is frequently described in textbooks as having an agrogeological background. However, it was particularly perceptive to new Russian ideas. Hermann Stremme (1879–1961) was personally acquainted with K. Glinka and helped him with the German edition of Glinka's book (1914). Stremme's upper-level classes, or groups of soils (biogenic, hydrogenic, lithogenic and orogenic) correlated with Dokuchaev's soil-forming agents. The later system of Eduard Mückenhausen (1965) was based on the ideas of Walter Kubiřna (1953) and included subaqueous soils together with terrestrial, hydromorphic, and bog soils, and these groups of soils comprise the upper level—Abteilungen (Orders). The system has seven taxonomic categories and is defined as a morpho-genetic classification (*Working Group on Soil Classification of the German Soil Science Society, 1997*). The second category—suborder (Classes) includes soil properties and soil dynamics, the result of the main soil-forming processes, which produce genetic horizons; their assemblages identify soil types, the central units of the system. Subtypes are intergrades between types or deviations from the central type, varieties are qualitative modifications of subtypes, and subvarieties are their quantitative variants. Texture is a criterion at the lowest level. The structure of the system, names of categories and even some differentiating criteria have much in common with those used in Russian systems (types, subtypes, varieties). Much attention was always paid by German soil scientists to gley (hydromorphic) soils, among which Stagnogley, Pseudogley and even Hangogley, Oxigley and Bleichogley (bleached gley) were identified (there are in total four types and 27 subtypes in the class of Gley soils), to semi-terrestrial and sub-hydric soils, to solifluction covers, and to coastal soils under the influence of tides. Arable and cultivated variants of soils are included in many soil types.

The German system is regarded as still incomplete, and its new version is expected. Urban soils and their classification received much attention in German soil science in early 2000s, and seem to be included in the updated system.

The history of the French classification is rather unusual: two systems with quite different approaches, principles, hierarchy, new soil names, and construction were created. The first 'old' system of Aubert and Duřhaufour (1956) was quite individual in addressing soil evolution, which is recorded in the type of soil profile development. It was represented schematically as a sequence of horizons: profiles AC → A(B)C → ABC → AEBC. Pedogenetic processes underlie the system, and soil-forming agents, such as parent material or climate, are used to differentiate and even name soils; some of them have also zonal names. The French system, like the German one, is hierarchical and has seven categories; although, unlike it, the differentiating criteria are diverse at the upper level. They include humus forms, mineralogical composition, hydromorphism, and salinity. Soil diagnostics at the type level are based on profile morphology rather than specific horizons. The new French system also designated for soils of the world was published in 1995 and 2008, it was entitled *Référenciel Pédologique*; its authors did not consider it a classification *sensu stricto*. However, the system is very detailed and is defined as non-hierarchical, although it has two levels. The first level is based on soil properties implemented into the diagnostic reference horizons, whose combinations create central images (*grands ensembles de références*). The genetic approach is clearly present at this level, but it is more evident at the lower level, where numerous qualifiers are added to central images—reference solums. For qualifiers, diverse soil properties were used including types of humus, analytical and landscape characteristics such as alpine, subalpine, and also human effects on soils. Both French systems may be regarded as flexible and oriented or underlain by pedogenetic concepts. The new system enables detailed qualification of soils by compiling unlimited assemblages of qualifiers for soil types.

In the United States, in early 1950s, as a reaction to the overestimated zonal bias in soil classifications of Marbut (1935), and of Baldwin, Kellogg and Thorp (1938), Soil Taxonomy was born, first as a series of approximations; its Seventh Approximation of 1960 became well-known in the world. However, at first it was not accepted unanimously, and the main criticisms were: formal approach, absence of pedogenetic concepts, strange and difficult nomenclature.² Nevertheless, an important achievement of Soil Taxonomy was the system of diagnostic tools, horizons, properties and materials, which was accepted by many developers of soil classification.

²In the Introduction to *Référenciel Pédologique* of 1995 soil names were considered as a reminder of "Alice in the Wonderland" by Lewis Carroll.

In the Soviet Union, several individual or authors' systems appeared in the second half of the 20th century. Together with the national systems for soils of the USSR and Russia, several classification systems were proposed by eminent soil scientists. Most of them concern the upper taxonomic levels, where soil scientists implemented their personal perceptions of the world of soils; these systems ended at the level of 'genetic soil type' (solonetz, chernozem, gray forest soil). They may be regarded as updated approaches originally put forward by scientists of the Early Dokuchaev period.

Maria Glazovskaya revised the ideas of Polynov and suggested a geochemical-genetic system (Glazovskaya, 1983), and on this basis in the 1990s she proposed a series of applied classifications of soils by their vulnerability and tolerance towards pollutants ('heavy metals', acid rain).

Vladimir Volobuev (1973) suggested a system of soil grouping based on the recognition of 11 types of soil humus (in terms of its humate or fulvate nature, base saturation and the rate of plant residue decomposition) combined with categories of mass transformation of the mineral soil (from bedrock to ferrallitic weathered mantle). These entries provided a separation of the classes of soil formation. Criteria for the next level were the types of biological cycle and the geochemical position of the soil. The further subdivision indicated zonal types of soil formation. His approaches were defined as priority of 'organomineral reactions'.

Victor Kovda combined the evolutionary and geochemical approaches to soil grouping. He tried to arrange a system of soils by the balance of substances in the course of soil evolution from the hydroaccumulative (subaquatic) stage, via various hydromorphic states to the automorphic one; at higher level, soils were grouped into soil-geochemical formations by several characteristics embracing soil properties, pedogenetic processes and environment (Kovda et al., 1967).

Innokentiy Gerasimov trying to support and promote the genetic trends in soil classifications suggested a system based on elementary pedogenetic processes, thus reviving the early scheme of his teacher Sergey Neustruev. Pedogenetic processes were arranged into major groups: (I) transformation of organic material, (II) transformation of the soil mineral mass, (III) migration and accumulation of mineral, organomineral, and organic products of weathering and pedogenesis, (IV) cementation, and deformation (V). Major groups comprised elementary processes (25 in total), which received their own codes, and soil horizons were characterized by the sets of processes shaping them; combinations of horizons produced soil types, and each type was presented by 'process codes' and 'profile codes', i.e., profile formula (Gerasimov, 1975). Thus, soil properties and pedogenetic processes served as the framework of the system, and profile formulae for soil types were later implemented into the new Russian system.

Vladimir Fridland developed a classification system as supervisor of the State project on the compilation of a soil map of Russia. This work started with a detailed program to describe the soil map; the text provided descriptions of the 205 soils in the legend (1972). Soils were described in terms of preliminarily defined horizons and properties. This huge cartographic project resulted in a basic soil classification comprising three components: profile-genetic, regime, and mineralogical-petrographic (Fridland, 1982). Full soil names were composed of three groups of qualifiers for each component; soil types remained central units in this system and were identified by the sets of horizons. The profile-genetic component of Fridland's system became a basis for the new Russian soil classification of 1997–2004–2008.

In Russia, there are now two classification systems in use. The official system of 1977³ embraces soils of agricultural and potentially agricultural lands of the USSR, while the new classification includes soils of the whole country as its object (Fig. 2).

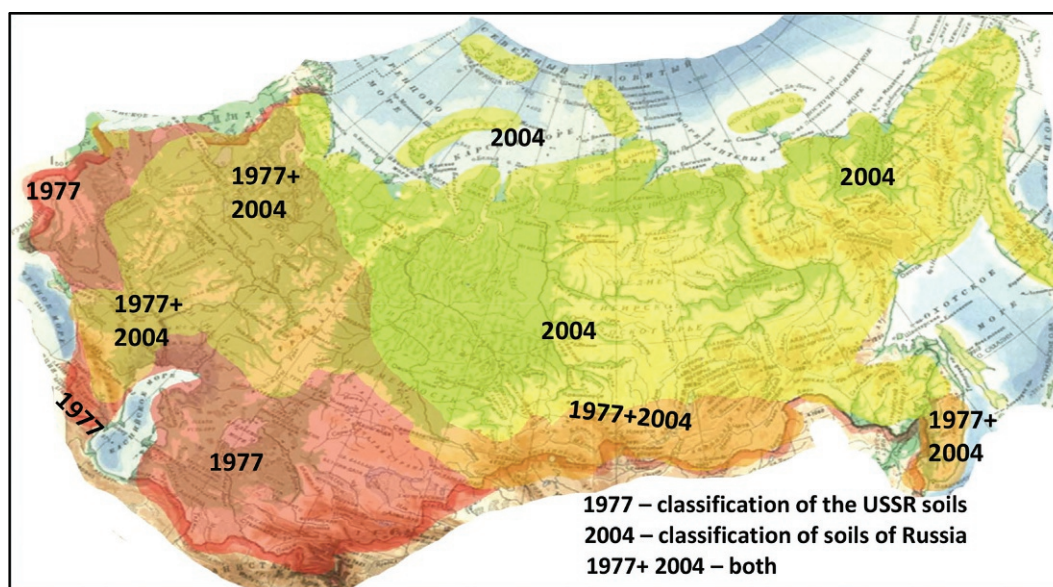


Fig. 2 Areas in the Soviet Union/Russia showing which soils are presented in two soil classification systems.

³Its English translation was first published in Egorov et al. (1986).

The new system was published as consecutive approximations in 1997,⁴ 2004, and 2008. Both systems were developed in the V.V. Dokuchaev Soil Science Institute, both are hierarchical, and differ in principles and structure.

The prototype of the Soviet system of 1977 was created by I. Gerasimov, A. Zavalishin, and Ye. Ivanova in 1939. The uppermost and main taxonomic unit ‘soil type’, was defined as “a group of soils developing in similar conditions and characterized by common origin and common processes of the transformation and migration of substances, similar level of fertility”. Criteria for the categories below the type level were officially adopted in 1958 by the Ministry of Agriculture of the USSR, as well as the list of 110 soil types (1966) for large-scale soil surveys.

The ideology of the system comprised: (i) the zonal (‘bioclimatic’) approach presuming a determined position of soil types in the coordinate system of bioclimatic parameters, (ii) the degree of soil hydromorphism (automorphic, semihydromorphic and hydromorphic soils). These principles were very strict; some deviations (e.g., the appearance of soluble salts in a soil of the humid climate) were shifted down to the lower level, if not omitted as incompatible with zonal conditions. It was a very consistent and fully developed system.

The 1977 system has many advantages: it is detailed; quantitative criteria for soil subdivision at the lower classification levels were thoroughly developed; addressed soil-forming agents was valuable for soil survey, and soils of croplands were subdivided by the degree of their cultivation. It was applied for large-scale soil mapping of agriculturally suitable areas. A large block presented aridic and irrigated soils of Central Asian republics. However, it did not include soils of the vast cold and permafrost-affected regions of the Russian North and Siberia (Fig. 2), and this was one of the reasons for developing a new system.

The new Russian soil classification system (RSCS) inherited many elements of Fridland’s basic system (1982), namely: principles, genesis and properties, taxonomic structure, broad range of soils, priority of diagnostic horizons and properties; it is open, hierarchical, and should be updated every 10–15 years. The first version of RSCS appeared in 1997, and the next one was published in 2004. Advice on the contents (not principles!) of both versions have been taken into account for a shorter version in 2008 “Field Guide for Russian Soils”. Updating of the RSCS is now in progress (Khitrov and Gerasimova, 2021, 2022).

Unlike other soil classifications in Russia/USSR, the RSCS starts with the definition of soil, which is important for delineating the objects classified, in particular, artificial human-made, urban, and subaqueous soils. As an extension of the sphere of objects, a special group of solid-phase bodies is mentioned in the RSCS under the name “Technogenic Surface Formations”. In order to emphasize their non-soil nature, unusual clumsy names were deliberately given to them, for example, “artificrat” meaning an artificially made (fabricated) pseudo-soil construction. Addressing non-traditional objects coincided in time with the birth of the International Working group SUITMA⁵ (1998).

In the RSCS soils are classified according to their properties which have been formed by pedogenetic processes; environmental conditions may be considered only for explanations and descriptions. Hence, the basic principle of RSCS is similar to the background of the International WRB system, as well as the diagnostic tools for soil identification, diagnostic horizons and diagnostic properties. In the RSCS, soil types are central soil units, and lower categories remain almost unchanged compared with the system of 1977. Two higher categories were introduced: trunks and orders (Table 2). At the uppermost level, trunks, soils are grouped according to the ratio of pedogenesis to lithogenesis (i.e., parent material formation; Fig. 3) The majority of soils belong to the postlithogenic trunk; the synlithogenic trunk embraces alluvial, volcanic, and urban soils, while all kinds of thick peat soils comprise the organogenic trunk; primary pedogenesis is weakly manifested on hard rocks or artificial substrates.

Table 2 Taxonomic categories and criteria for classes in the new Russian soil classification system.

Category	Criteria	Example of a full soil name
Trunk	Pedogenesis-to-lithogenesis (peat formation included) ratio	Synlithogenic
Order	Similar major elements of the profile, similarity of pedogenesis	Alluvial soils
Type	System of diagnostic horizons	<i>Dark-humus^a</i> alluvial soil
Subtype	Modifications of diagnostic horizons, intergrades between types	Dark-humus alluvial <i>quasigleyic</i> soil
Genus	Base (un)saturation, salinity, presence of carbonates and/or gypsum	Dark-humus alluvial quasigleyic <i>carbonate-rich</i> soil
Species	Degree of diagnostic property manifestation	<i>Deep</i> dark-humus alluvial quasigleyic carbonate-rich soil
Variety	Texture, amount of stones	Deep dark-humus <i>shallow</i> alluvial quasigleyic carbonate-rich loamy soil
Phase	Parent material, depth of the solum	Deep dark-humus shallow alluvial quasigleyic carbonate-rich soil <i>on finely stratified alluvium</i>

Example of alluvial soil (Fluvisol).

^aItalicized are qualifiers added to the soil name at the corresponding category.

⁴English translation edited by R.W. Arnold with some corrections was published in Shishov et al. (2001), prior to the main versions.

⁵SUITMA—Soils of Urban, Industrial, Transport, Mining/Military Areas.

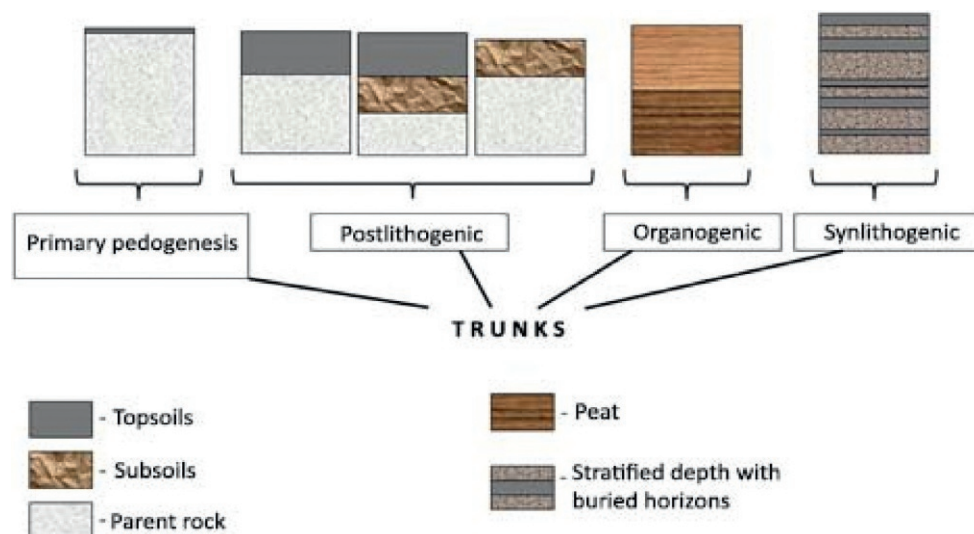


Fig. 3 Soil horizons and parent materials in the trunks—upper category in the RSCS.

Table 3 Groups and number of diagnostic horizons in RSCS, WRB and US Soil Taxonomy (horizons of tropical soils in the WRB and Soil Taxonomy are excluded in this calculation).

Groups of horizons	Classification systems			
	Russian, 2008	Russian, proposed (2021, 2022)	WRB, 2014	Soil Taxonomy, 1999
Upper organomineral, humus- accumulative	11	10	5	4
Upper organic, accumulation of organic material	6	7	2	2
Transitional subsurface, destruction and removal of matter	5	5	–	2
Middle-profile, accumulation and transformation of matter	17	14	22	15
Hydrogenic, effect of ground water	4	2	–	–
Halomorphic, soluble salts in the solum	2	2	1	1
Human-modified and anthropogenic	6	6	5	2
Reclamation by filling	–	2	1	–
TOTAL	51	48	36	26

Criteria for the next category, order, are soil-forming processes in the broad sense. They are recognized by the character of the profile composition, or by a diagnostic horizon common to all soils of the order. Thus, orders of texturally differentiated, humus-accumulative, and alluvial soils may serve as examples. The classes of the third category are 'genetic soil types'. Soil type is defined as a soil body with a certain sequence of diagnostic horizons, or profile formula; consequently, each soil type has its own formula written in Roman capitals. These formulae are complemented by symbols (Roman lower-case letters) corresponding to diagnostic features serving as criteria for the fourth category, subtypes. There are four lower categories after the subtype, and they retain their names given in 1977: genus, species, variety, phase (Table 2).

Diagnostic horizons are important in the new system. Soil genesis underpins the choice of differentiating criteria and control of horizon definitions, which is similar to the WRB and US Soil Taxonomy (Soil Survey Staff, 1999), whereas the horizons themselves are close in their nature (Table 3) but not identical.

Diagnostic features are mostly modifications of genetic horizons produced by the overlay of several mechanisms responsible for horizon development, imprints of present-day soil regimes, or inheritance of former pedogenesis, including the human-affected events. To facilitate their application, they are grouped in accordance with their origin: process-derived, intergrades, evolutionary, rock-inherited, substrate-accumulative, etc.

Special attention is paid in the RSCS to human-modified soils, which are considered to be the result of soil evolution under the impact of human activities, mostly agricultural. Several stages of agrogenic evolution are specified; they are identified by the occurrence of human-modified horizons and features, and are reflected by the position of soils in the system. Thus, strongly modified soils form individual orders (second level); however, they are not analogues of Anthrosols. Moderately modified soils (with one new horizon and the remainder of the natural profile) are qualified as types within natural soil orders, whereas weakly modified ones (no changes in the sets of horizons) are included as subtypes into natural type classes (Fig. 4).

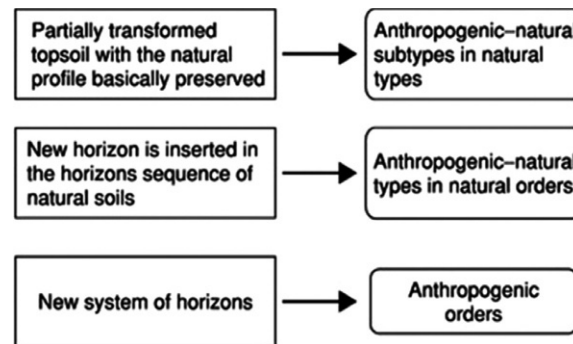


Fig. 4 Major elements of profile composition and taxonomic level of human-modified soils.

Traditional *soil names* were mostly retained in the new classification with two exceptions. First, components of soil names indicating environmental conditions (e.g., forest soils, meadow soils) were excluded. Second, new names were introduced for ‘new’ soils that were absent in the former system, for example, cryozems, rzhavozems. For plowed soils, the prefix ‘agro’ was proposed for the agrogenic soil types: agro-chnozem (PU-AU-BCA-Cca), agro-soddy-podzolic soil (P-EL-BEL-BT-C); for strongly eroded arable soils whose profiles comprise an arable horizon and subsoil, the name ‘agrozem’ is applied with the subsoil name, e.g., alfe-humus agrozem (P-BHF-C), dark-humus carbonate-accumulative agrozem (PU-BCA-Cca).

Theoretical aspects of soil classification were further developed by Ilia Sokolov. In 1978, he suggested the idea of a basic polycomponent soil classification that had much in common with Fridland’s (1982) proposal. Sokolov distinguished between stable (conservative) soil features reflecting the long history of soil development (soil memory) and dynamic soil characteristics reflecting current conditions of soil functioning (soil regimes) and argued in favor of their separate classification with further combination of these two components on a coordinate basis (Sokolov, 1978). Among stable features, it was proposed that the lithogenic heritage of soils derived from the parent material be classified separately from the pedogenic heritage. Another novelty introduced by Sokolov (1991) concerned the definition of the objects of classification. The entire universe of soils and soil-like bodies performing ecological functions of soils in the biosphere was embraced under a new term: Ecosols. At the highest level, Ecosols were subdivided into Natural, Natural-Artificial (anthropogenic), and Artificial kingdoms; each of them was classified in its own hierarchical system. The priority of the distinction between natural and artificial Ecosols was important in the context of the growing anthropogenic footprint in soils. A clear distinction was also made between soils (bodies shaped by pedogenesis from the already formed parent material) and pedoliths (bodies shaped by synchronous pedogenetic and geomorphic processes, such as aeolian, fluvial, or volcanic sedimentation; turbation; and hydrogenic accumulation of substances). Soils proper (normal soils of Dokuchaev) were separated at the fifth level of this system and further subdivided at four hierarchical levels. These two papers by Sokolov remained untranslated into English and had little impact on the development of soil classification ideas beyond Russia. Their theoretical significance has yet to be analyzed.

In the recent decade, key elements of Dokuchaev’s paradigm have been extended beyond the Earth to cover all loose surface formations (‘exons’) of lithospheric planets developing under the impact of exogenous factors. At the highest level, exons are subdivided into formations transported by lateral processes (‘transons’), in situ formations (‘sitons’), and transitional formations (‘transsitons’). In situ processes acting in sitons and transsitons result in differentiation of their vertical profiles. Such soil-like bodies have been named ‘soloids’ (Targulian et al., 2010, 2017). Soloids can be formed under the impact of cosmic, solar, lithospheric, atmospheric, hydrospheric, biospheric, and anthropospheric processes acting together or separately. This approach developed by Victor Targulian might be helpful in unveiling the history of Martian regolith and the search for Martian ‘soils’ (paleobiotic soloids). It can be considered ‘the revised edition’ of the agrogeological approach.

The search for better soil classification decisions continues throughout the world within Dokuchaev’s paradigm of soil science. It is now possible to analyze big soil and soil-related data. Will artificial intelligence confront or confirm Dokuchaev’s principles? How will they change? Pedologists are eager to know.

Conclusions

In concluding this brief overview on soil classifications created in the second half of the last century we may note that two aspects of Dokuchaev’s paradigm exerted a dual effect on soil classifications: (a) involving genesis as an open or concealed criterion for discriminating between classes of soils and (b) directly addressing soil-forming agents for aggregating soils into zonal entities. The WRB is an example of the first trend, and the USSR-1977 system of the second. The new Russian system has much in common with the WRB, which is being continuously updated: its new version has appeared in 2022.

References

- Afanasiev JN (1927) *The Classification Problem in Russian Soil Science*. Leningrad: Publishing Office of the USSR Academy of Sciences, p. 63.
- Aubert G and Duçhaufour PH (1956) Project de classification des sols. In: *Transactions Third International Congress of Soil Science (Paris)*, Vol. E (Commission V) 597–604.
- Baldwin M, Kellogg CE, and Thorp J (1938) Soil classification. In: *Soils and Men: Yearbook of Agriculture*, pp. 979–1001. Washington DC: USDA.
- Brevik EC and Hartemink AE (2012) Soil Maps of the United States of America. *Soil Science Society of America Journal* 77(4): 1117–1132. <https://doi.org/10.2136/sssaj2012.0390>.
- Coffey GN (1911) The development of soil survey work in the United States with a brief reference to foreign countries. *Proceedings of the American Society of Agronomy* 3: 115–129.
- Egorov VV, et al. (1986) *Classification and Diagnostics of Soils of the USSR. Russian Translation Series 42*. New Delhi: Amerind Publishing.
- Fridland VM (1982) *Main Principles and Elements of the Basic Soil Classification and Development of Its Program*. Moscow: Dokuchaev Soil Science Institute. [in Russian].
- Gedroiz KK (1925) The soil absorbing complex and the absorbed cations of the soil as a basis for a genetic soil classification. *Nosov Agr. Exp. Bull. 38. Leningrad. 1925. Kolloidchem. Beih.* 29: 149–260.
- Gerasimov IP (1975) Experiment in genetic diagnosis of the soils of the USSR on the basis of elementary soil processes. *Soviet Soil Science* 7: 257–263.
- Glazovskaya MA (1983) *Soils of the World. Soil Families and Soil Types*, vol. 1. Rotterdam: A.A. Balkema.
- Glinka KD (1914) *Die Typen der Bodenbildung, ihre Klassifikation und geographische Verbreitung*. Berlin, 582 pp.
- Gong Z, Zhang X, Chen J, and Zhang G (2003) Origin and development of soil science in ancient China. *Geoderma* 115: 3–13. [https://doi.org/10.1016/S0016-7061\(03\)00071-5](https://doi.org/10.1016/S0016-7061(03)00071-5).
- Hilgard EW (1892) *A Report on the Relations of Soil to Climate*. Washington, DC: Weather Bureau.
- Hirai H and Hamazaki T (2004) Historical aspects of soil classification in Japan. *Soil Science & Plant Nutrition* 50(5): 611–622. <https://doi.org/10.1080/00380768.2004.10408519>.
- IUSS Working Group WRB (2015) *World Reference Base for Soil Resources 2014, Update 2015. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps. World Soil Resources Reports 106*. Rome: FAO.
- Khitrov NB and Gerasimova MI (2021) Diagnostic horizons in the classification system of Russian soils: Version 2021. *Eurasian Soil Science* 54(8): 1131–1140.
- Khitrov NB and Gerasimova MI (2022) Diagnostic properties and soil forming materials in the classification system of Russian soils: version 2021. *Eurasian Soil Science* 55(1): 1–10.
- Kossowitsch P (1911) *Die Bodenbildungsprozesse und die Hauptprinzipien der Bodenklassifikation*. 2. Internat. Agrogeol. Konf. Stockholm Springer, pp. 232–254.
- Kovda VA, Lobova YEV, and Rozanov BG (1967) Classification of the world's soils. *Soviet Soil Science* 7: 851–863.
- Krasilnikov P, Tabor J, and Arnold R (2009) In: Krasilnikov P, Ibáñez Martí J-J, Arnold R, and Shoba S (eds.) *Ethnopedology and folk soil classifications, in A Handbook of Soil Terminology, Correlation and Classification*, pp. 337–404. London, Sterling, VA: Earthscan.
- Krupenikov IA (1992) *History of Soil Science: From Its Inception to the Present*. New Delhi: Oxonian Press.
- Kuběřna WL (1948) *Entwicklungslehre des Bodens*. Wien: Springer - Verlag.
- Lee MP-H (1921) *The Economic History of China: With Special Reference to Agriculture*. Columbia University. Studies in History, Economics and Public Law, No. 225. New York: Columbia University.
- Marbut CF (1935) Soils of the United States. In: *Atlas of American Agriculture. Part III*, pp. 1–98. Washington DC: USDA.
- Mill JS (1882) *A System of Logic*, 8th edn New York: Harper & Brothers.
- Polynov BB (1936) Principles of the genetic classification of soils. Transaction of the XIII Congress of International Society of Soil Science, vol. 3, pp. 158–161. London: Thomas Murby & Co.
- Shishov LL, Tonkonogov VD, Lebedeva II, and Gerasimova MI (2001) *Russian Soil Classification System. English translation edited by R Arnold*. Moscow: Dokuchaev Soil Science Institute.
- Simonson RW (1989) *Historical Highlights of Soil Survey and Soil Classification With Emphasis on the United States, 1899–1970*. ISRIC Technical Paper 18: The Netherlands.
- Smith GD (1986) *The Guy Smith Interviews: Rationale for Concepts in Soil Taxonomy*. Washington DC: SMSS Technical Monograph No. 11.
- Soil Survey Staff (1999) *Soil Taxonomy. A Basic System of Soil Classification for Making and Interpreting Soil Surveys*. 2nd edn *Handbook of Agriculture No. 436*, 2nd edn, Washington DC: United States Government Publishing Office.
- Sokolov IA (1978) On a basic classification of soils. *Pochvovedenie* 3: 113–123. [in Russian].
- Sokolov IA (1991) Basic substantive-genetic soil classification. *Pochvovedenie* 3: 107–121. [in Russian].
- Strzemiński M (1975) *Ideas Underlying Soil Systematics*. (Translated from Polish). Warsaw
- Targulian V, Mergelov N, Gilichinsky D, Sedov S, Demidov N, Goryachkin S, and Ivanov A (2010) Dokuchaev's soil paradigm and extraterrestrial soils. In: *Proceeding of 19th WCSS Soil Solutions for a Changing World (Brisbane, Australia). Published on DVD*.
- Targulian VO, Mergelov NS, and Goryachkin SV (2017) Soil-like bodies on Mars. *Eurasian Soil Science* 50(2): 185–197.
- Whitney M (1909) *Soils of the United States. Bureau of Soils Bull. 55* Washington DC: USDA US Gov Print Office.
- Williams BJ (2006) Aztec soil knowledge. Classes, management, and ecology, in footprints in the soil. In: Warkentin BP (ed.) *People and Ideas in Soil History*. Amsterdam: Elsevier.
- Working Group on Soil Classification of the German Soil Science Society (1997) Soil classification of the Federal Republic of Germany. *Mitteilungen der Deutschen Bodenkundlichen Gesellschaft* 84: 253–275.

Soil classification systems: World reference base and soil taxonomy

David Badía-Villas, Technological College of Huesca, University of Zaragoza, Zaragoza, Spain

© 2023 Elsevier Ltd. All rights reserved.

Introduction	197
Background of WRB and ST systems	198
The ‘Cornerstones’ for classifying soils: Genetic and diagnostic horizons	198
Genetic horizons	198
<i>Diagnostic horizons for mineral and organic soils</i>	199
The rules for classifying soils: How it works	199
Example profile	203
Summary	204
References	204

Abstract

This chapter considers the characteristics of the two soil taxonomies that have made them world references: World Reference Base, WRB (IUSS Working Group WRB, 2015) and Soil Taxonomy, ST (Soil Survey Staff, 2014). Diagnostic horizons that serve to classify soils are mentioned, after recalling the genetic horizons. The first hierarchical levels of both taxonomies are described: the Reference Soil Groups of the WRB and the Orders of the Soil Taxonomy. Finally, the procedure for classifying and naming a soil is described using, as an example, the field observations and the analytical data from a soil profile of Toledo, Spain.

Key points

- Soils with similar properties have to have the same name.
- WRB and Soil Taxonomy as world reference systems.
- The basic rules for classifying soils are explained.

Introduction

This chapter introduces the use of the two reference soil taxonomies in the world: the World Reference Base, WRB (IUSS Working Group WRB, 2015) and Soil Taxonomy, ST (Soil Survey Staff, 2014). This chapter will be useful to soil scientists and those working in related and unrelated fields who struggle to understand how soil taxonomy works. Most users of soil information (from maps, reports or scientific articles), will probably never classify a soil themselves. But understanding the steps followed in the classification process may help to improve the interpretation of soils already classified. Classifying soils requires some knowledge of soil-forming processes and an ability to interpret the morphological description and the physical and chemical properties of the soils obtained from field and laboratory work (Arnold, 2005; Switoniak et al., 2018).

According to Soil Survey Staff (1999), a soil is defined as “A natural body comprised of solids (minerals and organic matter), liquid, and gases that occurs on the land surface, occupies space, and is characterized by one or both of the following: horizons, or layers, that are distinguishable from the initial material as a result of additions, losses, transfers, and transformations of energy and matter or the ability to support rooted plants in a natural environment”. For classification purposes, the lower depth limit of the soil in Soil Taxonomy is 2 m. Similarly, the object classified in the WRB is: “any material within 2 m of the Earth’s surface that is in contact with the atmosphere, excluding living organisms, areas with continuous ice not covered by other material, and water bodies deeper than 2 m” (IUSS Working Group WRB, 2015).

Why give names to soils? Classification lets us organize and transmit knowledge about our soils, making it possible to transfer the experience acquired in a given geographic area to other areas. This requires that soils with similar properties have the same name. Currently, the classification of soils is an essential instrument in research, and can be applied in many topics such as the rehabilitation of degraded areas, contaminated soils, land grabbing, (digital) mapping, and land evaluation and land assessments for different uses, such as vineyard zoning, irrigation projects of rainfed soils, etc. (Porta et al., 2003). In field research projects, it is usual to provide a description of the area of study: the geomorphology, the use of the land, the climate, the plant species, etc. In the same way, the name of the soil should be provided. As Kubiiena (1952) wrote: “For the rationalization of research in a natural science there is no other way than to rely on a systematic system as developed as possible. Can someone get any benefit from an experiment with plant material if the author cannot define exactly the name of the plant that he has used? Analogously, if the indications on the soil are missing, the total experiment remains in the air”.

Throughout the history of Soil Science, multiple national-based classification systems have been proposed and coexist, e.g., Australian, Brazilian, Canadian, Chinese, French, German, and others. But the two systems that are currently used as world references are: World Reference Base, WRB (IUSS Working Group WRB, 2015) and Soil Taxonomy (Soil Survey Staff, 2014).

Background of WRB and ST systems

Let us see how these most recent and current classifications (IUSS Working Group WRB, 2015; Soil Survey Staff, 2014) have been arrived at.

The International Society of Soil Sciences (ISSS) at its Sixth Congress, at Madison, Wisconsin (United States) in 1960 recommended that information about soils of the world should be compiled. The FAO-UNESCO (Food and Agriculture Organization of the United Nations), beginning in 1961, created a 1:5,000,000 soil map of the world (FAO, 1974), whose legend began to be used as if it were a true classification system. This legend was later revised (FAO, 1988) to be able to use it at more detailed scales and it would eventually become the current World Reference Base for soil resources (FAO-ISRIC-ISSS, 1998), which is successively updated (IUSS Working Group WRB, 2007, 2015). The WRB is intended to serve as a basis for correlation between existing classification systems, a communication tool for compiling global soil databases, and for the inventory and monitoring of the world's soil resources (Bridges et al., 1998; Deckers et al., 1998; Nachtergaele, 2005). The International Soil Reference and Information Centre (ISRIC) collects information on the world's soils and manages it with this taxonomy (<https://soilgrids.org>).

On the other hand, the Soil Taxonomy system developed from the need to order the several thousand series of soils recognized at the beginning of twentieth century (Brown, 2005). In 1951 the Soil Survey Staff (United States Department of Agriculture) began to work on a classification system that developed into many drafts or approximations. The Seventh Approximation, which was presented to the above mentioned Sixth Congress of the ISSS led to the official version of Soil Taxonomy edited in 1975 (Soil Survey Staff, 1975). Since then, several committees have periodically reviewed and added modifications that are reflected in the "Keys to Soil Taxonomy", whose first edition (Soil Survey Staff, 1983) has been successively improved until the latest version (Soil Survey Staff, 2014) and that the institution itself is updating (<https://www.nrcs.usda.gov>).

The 'Cornerstones' for classifying soils: Genetic and diagnostic horizons

Genetic horizons

The entry of matter and energy on the soil surface generates the formation of layers, whose upper and lower limits are often parallel to it and are known as soil horizons. The horizons are subdivisions of a soil into volumes considered to be homogeneous, which makes them the basic unit for describing and sampling a soil. Individual soils and their horizons are identified and described, in practice, by means of a soil profile, which may be seen, for example, in the exposed face of a soil pit or cutting. In fact, we can say that soil classification starts with differentiation of the horizons in the field and their complete and accurate description and sampling (FAO, 2006; Schoeneberger et al., 2012), later completed with analytical data.

The presence of these horizons results from the evolution or genesis of the soil from the parent material hence they are called genetic horizons, which applies to both the major classifications. These horizons are assigned by a combination of letters (uppercase and lowercase) and numbers, indicating the following:

1. The main horizons (or layers) are designated by a capital letter (master horizons), which informs about the position they occupy in the soil: A: top; B: middle; C: bottom.
2. The main process responsible for the formation of the horizon is indicated by lowercase letters that are added to the main horizon as a subscript or as a second character, i.e., Ap indicates that the surface horizon has been ploughed; Bt represents a subsurface horizon with a significantly higher percentage of phyllosilicate clay than the overlying soil material resulting from translocation (movement within the profile); Ck is a bottom horizon where secondary calcium carbonate has accumulated, etc.
3. With numerical indices the position of the horizon within the soil is designated, for example: Bt1, Bt2, etc.
4. Transition horizons are designated by the combination of letters of the relevant horizons, without indices. It differs between:
 - 4.1 no bar, if the properties of the horizon are intermediate although with a certain predominance of the characteristics of one horizon over those of the other, for example: AB or BA.
 - 4.2 with a bar, for example A/B, if the horizons are clearly identifiable, mixing one of them (A) with the other (B).
5. Lithic discontinuities are expressed by placing an Arabic numeral before the horizon letter, sequentially, for example: A-2C.

Although the use of this nomenclature is very similar in both the WRB and Soil Taxonomy, some differences remain, so the reader is redirected to the original sources.

Many master horizons and layers can have one or two (rarely three) lowercase letter suffixes, that must be cited according to some rules (FAO, 2006). The listing of genetic horizons ordered from the soil surface downwards is called a sequence. For example, Ap-Bt-Ck-C could be the sequence of an evolved profile of the Mediterranean region, whose details we will provide later.

Diagnostic horizons for mineral and organic soils

The description of genetic horizons is merely qualitative and is impractical for identifying and classifying soils, therefore, diagnostic horizons have been defined. They are defined morphometrically, with quantifiable physical and chemical properties measured in the field and laboratory, and usually related to soil-forming processes. Both WRB (IUSS Working Group WRB, 2015) and Soil Taxonomy (Soil Survey Staff, 2014) are based on diagnostic horizons, although they maintain some differences in their definition. In order to qualify for a particular soil grouping, a soil has to possess the measured criteria dictated in this definition.

Two types of diagnostic horizons are specified according to their positions in the soil profile: at the surface (surface horizons, named epipedons by the ST) and at depth below the surface (subsurface horizons). Between surface and subsurface horizons, 37 diagnostic horizons are currently defined by the WRB and 26 by the ST. Additionally, other diagnostic properties and materials are also used to classify soils, usually at a lower hierarchical level. Describing in detail the total requirements for diagnostic horizons, materials, and properties is beyond the scope of this chapter, so the reader is referred to original sources (IUSS Working Group WRB, 2015; Soil Survey Staff, 2014).

It is remarkable how Soil Taxonomy, unlike the WRB, uses climate as an important diagnostic characteristic for the classification of both mineral and organic soils, taking account of both the soil moisture and temperature regimes because of their strong influence on the soil formation processes (additions and losses of components, transformations and translocations within the profile). In this way, one soil order, specifically Aridisols, and most of the suborders are governed by soil moisture regimes, while soil temperature regimes become important at lower levels of Soil Taxonomy. Another example is the soil order Gelisols, which is based on soil features like permafrost that is strongly correlated with atmospheric climate. Climate features, however, are not considered in the soil definitions and classification by the WRB. The latter excludes climate for several reasons: to avoid subordinating soil classification to the availability of climate data and to avoid changes in the soil name by global or local climate change.

One of the most important differences between Soil Taxonomy and WRB systems focuses on the use or not of climate data

The characteristics that have allowed the WRB and the ST to be world reference soil classification systems are summarized below :

- The classification is based on soil characteristics defined in terms of diagnostic horizons, as well as properties and climate regime in ST and diagnostic properties and materials in WRB, which sustain relationships with soil-forming processes.
- They consider characteristics that, generally, are of significance for soil management.
- It can be applied objectively as it is based on quantified information.
- They have a hierarchical organization, modeled on biological taxonomy (Brown, 2005), in which different levels are defined with precision, so that it can be used at both a detailed and general level; Soil Taxonomy has six levels and WRB has two levels.
- Taxa are mutually exclusive.
- Once the nomenclature rules are understood, the terminology is clear and does not require translation into different languages and allows users to glean basic soil information from the name alone. The terminology used by ST is self-explanatory, whereas that of WRB retains names used traditionally.

The rules for classifying soils: How it works

In practice, the exposed soil pit (the profile) showing the horizons or layers serves as the basic unit for description and sampling in the field. The completed description of the profile and laboratory analytical data of each horizon, thereafter, allow classification.

To place a soil in its taxonomic class requires the use of the most recent version of the published Key (IUSS Working Group WRB, 2015; Soil Survey Staff, 2014). Users should be familiar with the meanings of the terms for describing the diagnostic criteria. The first step in classification consists of checking whether any of the genetic horizons or sub-horizons fulfill some of the criteria for diagnostic horizons, properties and materials. Once this is verified, the soil can be allocated first into the highest hierarchical level (Table 1) followed in subsequent steps to go down to the lower levels using the keys. All the keys in both taxonomies are designed in such a way that the user can determine the correct classification of a soil by going through them systematically.

(1) For the WRB classification, we need to identify diagnostic horizons and properties as well as diagnostic materials. Diagnostic horizons (anthraquic horizon, argic horizon, calcic horizon, etc.) and properties (abrupt textural difference, albeluvic glossae, andic properties, etc.) have a combination of attributes linked to processes of soil formation. Diagnostic materials are materials that influence pedogenetic processes or are indicative of them, i.e., albic material, artefacts, calcaric material, etc.

Detecting diagnostic elements (horizons, properties and materials) is the first step in allocating the soil into the appropriate Reference Soil Group (RSG), the first level of classification (Table 2). The user should go through the Key systematically, starting at the beginning and excluding all RSGs, one by one, for which the specified requirements are not met. The soil belongs to the first RSG for which it fulfils the criteria. The most extensive RSGs in the world are shown graphically in Fig. 1.

For the classification of a soil to the second level of the WRB, two types of qualifiers must be added to the name of the RSG:

- principal qualifiers, which are ranked and given in an order of importance, and
- supplementary qualifiers, which are simply arranged in alphabetical order.

Table 1 List of the first hierarchical levels of taxonomies: Reference Soil Groups (RSGs) of the WRB and the Orders of the Soil Taxonomy. The correlation between both is merely indicative.

Main characteristic	WRB (IUSS Working Group WRB, 2015)	STS (Soil Survey Staff, 2014)
1. Soils with thick organic layers	Histosols	Histosols
2. Soils with strong human influence		
With long and intensive agricultural use	Anthrosols	
Containing significant amounts of artefacts	Technosols	
3. Soils with limitations to plant growth		
Permafrost-affected	Cryosols	Gelisols
Arid climate	—	Aridisols
Thin or with many coarse fragments	Leptosols	(Entisols)
With a high content of exchangeable Na	Solonetzs	
High concentration of soluble salts	Solonchaks	
With shrink-swell clays	Vertisols	Vertisols
4. Soils distinguished by Fe/Al chemistry		
4.1 affected by groundwater or stagnating water		
Groundwater-affected, underwater and in tidal areas	Gleysols	
Stagnating water, structural difference and/or moderate textural difference	Stagnosols	
Stagnating water, abrupt textural difference	Planosol	
4.2 affected by Fe/Al chemistry		
Allophanes or Al-humus complexes	Andosols	Andisols
Subsoil accumulation of humus and/or oxides	Podzols	Spodosols
Accumulation and redistribution of Fe	Plinthosols	
Low-activity clay, P fixation, many Fe oxides, strongly structured	Nitisols	
Dominance of kaolinite and oxides	Ferralsols	Oxisols
5. Pronounced accumulation of organic matter in the mineral topsoil		
Very dark topsoil, secondary carbonates	Chernozems	Mollisols
Dark topsoil, secondary carbonates	Kastanozems	
Dark topsoil, no secondary carbonates, high base status	Phaeozems	
Dark topsoil, low base status	Umbrisols	
6. Accumulation of substances		
Accumulation of, and cementation by, secondary silica	Durisols	(Inceptisols)
Accumulation of secondary gypsum	Gypsisols	
Accumulation of secondary carbonates	Calcisols	
7. Soils with clay-enriched subsoil		
Interfingering of coarser-textured, lighter colored material into a finer-textured, stronger colored layer	Retisols	Ultisols, Alfisols
Low-activity clays, low base status	Acrisols	
Low-activity clays, high base status	Lixisols	
High-activity clays, low base status	Alisols	
High-activity clays, high base status:	Luvissols	
8. Soils with little or no profile differentiation		
Moderately developed	Cambisols	Inceptisols
Sandy	Arenosols	Entisols
Stratified fluvatile, marine and lacustrine sediments	Fluvisols	
No significant profile development	Regosols	

Note: this table is not to be used as a key. For full definitions, please refer to original sources (IUSS Working Group WRB, 2015; Soil Survey Staff, 2014).

Unlike ST, in the WRB there are no additional keys to qualify for the second level. Both types of qualifiers are listed in the key for each of the RSGs (IUSS Working Group WRB, 2015). The principal qualifiers are added before the name of the RSG without brackets and without commas and sequenced from right to left. The supplementary qualifiers are added in brackets after the name of the RSG, separated by commas. Both qualifiers types must start with capital letters. This soil, whose steps for its classification we will develop at the end of this chapter, can serve as an example:



The RSGs are related to the primary processes that condition soil formation while the second level qualifiers focus on specific physical and chemical properties of the soil, usually with a significant effect on land use.

Table 2 RSG formative elements and main characteristics.

RSG	Formative element	Connotation
Acrisols	From Latin <i>acer</i> , very acid	Strongly weathered acid soils with low base saturation and argic with low-activity clays
Alisols	From Latin <i>alumen</i> , alum	Strongly weathered acid soils with low base saturation and argic with high-activity clays
Andosols	From Japanese <i>an</i> , dark, and <i>do</i> , soil.	Dark soils developed on glass-rich volcanic material (andic or vitric properties)
Anthrosols	From Greek <i>anthropos</i> , human being.	Soils modified by long-continued cultivation or addition of material (anthropogenic diagnostic horizons).
Arenosols	From Latin <i>arena</i> , sand	Sandy soil, with a textural class of loamy sand or coarser (and no skeletal)
Calcisols	From Latin <i>calx</i> , lime	With substantial accumulation of secondary lime (calcic or petrocalcic horizon)
Cambisols	from Late Latin <i>cambiare</i> , to change.	soils with at least an incipient subsurface soil formation (cambic horizon)
Chernozem	From Russian <i>cherniy</i> , black, and <i>zemlya</i> , earth or land	Blackish humus-rich surface horizon (chernic) with accumulation of secondary carbonates with depth (protocalcic properties or calcic horizon)
Cryosols	From Greek <i>kryos</i> , cold	Frost-affected soils
Durisols	from Latin <i>durus</i> , hard	Soils with hardened secondary silica
Ferralsols	From Latin <i>ferrum</i> , iron, and <i>alumen</i> , alum	Red and yellow tropical soils with a high content of sesquioxides
Fluvisols	From Latin <i>fluvius</i> , river	Soils developed in recent fluvial deposits (fluvic material)
Gleysols	From Russian <i>gley</i> , mucky mass	Soils with clear signs of groundwater influence (gleyic properties)
Gypsisols	From Greek <i>gypsos</i> , gypsum	Soils with substantial accumulation of secondary gypsum (gypsic or petrogypsic horizon)
Histosols	From Greek <i>histos</i> , tissue	Peat and muck soils (with histic horizon)
Kastanozems	From Latin <i>castanea</i> and Russian <i>kashtan</i> , chestnut, and Russian <i>zemlya</i> , earth or land	Dark brown, humus-rich surface horizon (mollic) with accumulation of secondary carbonates with depth (protocalcic properties or calcic horizon)
Leptosols	From Greek <i>leptos</i> , thin	Thin soils with no diagnostic horizons
Lixisols	From Latin <i>lixivia</i> , washed-out substances	soils with high base saturation and argic with low-activity clays
Luvisols	From Latin <i>eluere</i> , to wash	soils with high base saturation and argic with high-activity clays
Nitisols	From Latin <i>nitidus</i> , shiny	red tropical soils with a clayey (nitic) horizon shiny aggregate faces
Phaeozems	From Greek <i>phaios</i> , dusky, and Russian <i>zemlya</i> , earth or land	Dark brown, humus-rich surface horizon (mollic) without accumulation of secondary carbonates with depth
Planosols	From Latin <i>planus</i> , flat	With a coarse-textured surface horizon abruptly over a dense and finer textured subsoil (abrupt textural difference), typically in seasonally waterlogged flat lands
Plinthosols	From Greek <i>plinthos</i> , brick	Soils with plinthite, petroplinthite or pisoliths; that is with plinthic, petroplinthic or pisoplinthic horizons
Podzols	From Russian <i>pod</i> , underneath, and <i>zola</i> , ash	Soils with an eluvial horizon (albic material), which has the appearance of ash on an illuvial spodic horizon
Regosols	From Greek <i>rhegos</i> , blanket	Weakly developed soils in unconsolidated material with no diagnostic horizons
Retisols	From Latin <i>rete</i> , net	Clay illuviation horizon (argic) with retic properties, that is an interfingering of coarser-textured albic material into the illuviation horizon forming a net-like pattern
Solonchaks	From Russian <i>sol</i> , salt and <i>chak</i> , salt-affected area	Saline soils (salic horizon)
Stagnosols	From Latin <i>stagnare</i> , to flood	With perched water (stagnic properties)
Solonetz	From Russian <i>sol</i> , salt and <i>etz</i> , extremely	Soils with a high content of exchangeable Na (natric horizon)
Technosols	From Greek <i>technikos</i> , skilfully made	Artificial soil, with artefacts or technic hard material, built by the human activity
Umbrisols	From Latin <i>umbra</i> , shade	With dark brown umbric surface horizon
Vertisols	From Latin <i>vertere</i> , to turn	Heavy clay soils with a high proportion of swelling clays (vertic horizon)

Additionally, qualifiers that have depth requirements may be combined with specifiers (Epi-, Endo-, Amphi-, Ano-, Kato-, Panto-, Bathy-, etc.) to create subqualifiers (e.g., Epicalcic), showing the depth (or intensity) to which the property is occurring.

(2) In Soil Taxonomy, the highest hierarchical level are the Orders (12), see Table 3, followed by Suborders (64), Great groups (>300), Subgroups (>2400), Family and Series. The first four levels are primarily governed by climate and soil genesis, with the family and series levels capturing important physical and chemical properties for crop growth, engineering, and land management (Brown, 2005). To classify soils down to the Subgroup level, the following diagnostic criteria are required: materials (organic and mineral), horizons and diagnostic characteristics, and climate (moisture and temperature regime). To classify at the family level, it is additionally required to identify classes of: particle size, mineralogical constituents, soil depth, consistency, coatings and cracks. Series name of the soil is usually designated by the local name of the site where it is dominant, or where it was described for the first time.

Once we have identified which diagnostic criteria fulfill the horizons and layers of our soil, the next step consists of assigning the soil to the first hierarchical level: the Order (Table 2). The user must start at the beginning of the “Key to Soil Orders” and eliminate, one by one, all classes that include criteria that do not fit the soil in question. Soils are keyed out in order, starting with Gelisols and ending with Entisols. The soil belongs to the first class listed for which it meets all the required criteria. Once the Order has been assigned, the next step will be the “Key to Suborders” and so the process is repeated until the subgroup level is reached.

Below some representative soil profiles of various orders of Soil Taxonomy and their RSG equivalents (WRB) are shown (Fig. 2).

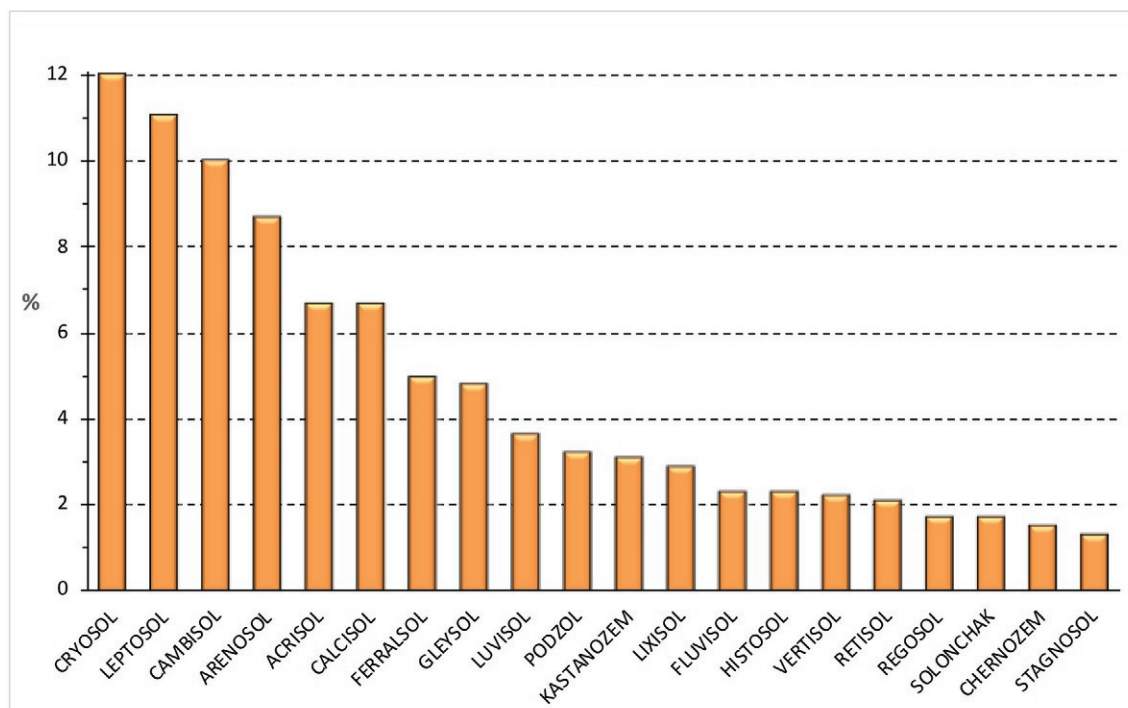


Fig. 1 Land surface (%) occupied by the main RSGs. From FAO (2001) Lecture Notes on the Major Soils of the World (With CD-ROM), by P Driessen, JA Deckers, O Spaargaren and F Nachtergaele (eds). *World Soil Resources Report No. 94*. pp. 334, Rome: FAO.

Table 3 Formative elements in the Orders of ST and main characteristics.

Orders	Formative element & origin	Connotation
Alfisols	Alf, from Pedalfer (Marbut classes)	No acid soils with an argillic, natric or kandic subsurface diagnostic horizon
Andisols	And, From Japanese <i>an</i> , dark, and <i>do</i> , soil.	Dark soils developed on glass-rich volcanic material (andic properties)
Aridisols	Id, from Latin <i>aridus</i> , dry	Soils with aridic moisture regime
Entisols	Ent, from English <i>recent</i> ,	Young soils: with ochric epipedon and no subsurface diagnostic horizon
Gelisols	El, from Latin <i>gelu</i> , ice	Soils that have permafrost or gelic materials
Histosols	Ist, from Greek <i>histos</i> , (plant) tissue	soils organic soil materials saturated with water for 30 days or more per year (peatbog)
Inceptisols	Ept, from Latin <i>inceptum</i> incipient	Soil with an incipient weathering revealed with some diagnostic horizon as cambic, calcic, gypsic, etc., without aridic moisture regime
Mollisols	Oll from Latin <i>mollis</i> , soft	With epipedon mollic with a base saturation of 50% or more throughout
Oxisols	Ox from French <i>oxide</i>	Strongly weathered acid soils with an oxic horizon
Spodosols	Od, from Greek <i>spodos</i> , ash	Soils with an illuvial spodic horizon, usually underling an eluvial albic horizon (<i>L. albus</i> , white)
Ultisols	Ult, from Latin <i>ultimus</i> , latest	Very acid soils with an argillic, kandic or fragipan subsurface diagnostic horizons and very low base saturation (BS <35%)
Vertisols	Ert, from Latin <i>vertere</i> , to turn.	Heavy clay soils with slickensides or other vertic characteristics, such as wedge-shaped peds and surface cracks, that open and close periodically, due to the high proportion of swelling clays

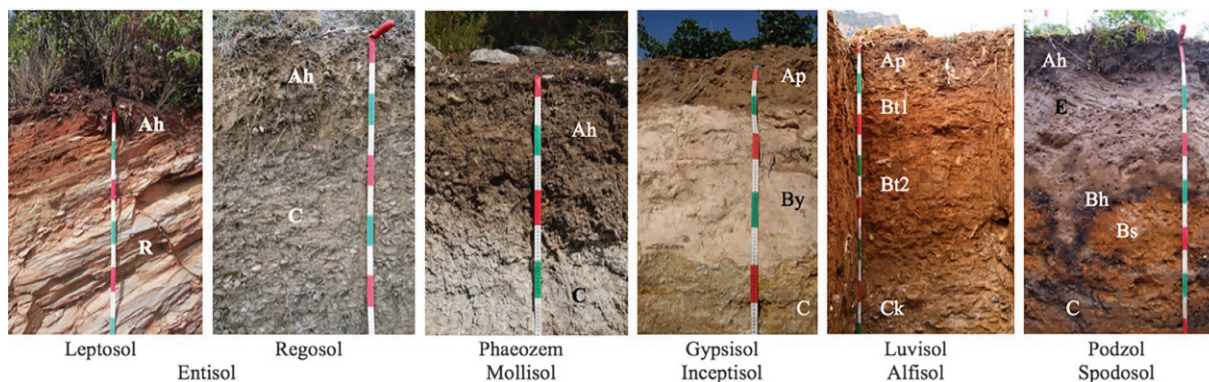


Fig. 2 Some of the most common soil morphologies showing their genetic horizons. The RSG (WRB) and most probable order (ST) is indicated.

In Fig. 2 starting at the left there are soil profiles of less development (with simple sequences of horizons A-R or A-C) and towards the right profiles of greater development (A-E-B-C); there is a gradient of processes ranging from incipient weathering and melanization to gypsification, calcification, argilluviation and podzolization. To a greater or lesser degree, these processes and its resulting properties are incorporated into the denomination of these soils.

Example profile

The site description, and essential field description and analytical properties data for a profile in the center of Spain are provided in Table 4. Using this information, the key diagnostic horizons are outlined. This example shows that the depths of the horizons described in the field do not always coincide with the depths required for the diagnostic horizons. The top 18 cm (Ap1 and top Ap2) fulfill the requirements for an ochric horizon (STS). The thick Bt-horizon described in the field meets the requirement for an illuvial horizon (observed clay coatings) and increase in phyllosilicate clay for an argillic/argic horizon (STS/WRB, respectively). It should be noted that the increase requested by both taxonomy systems is different ($\times 1.2$ STS and $\times 1.4$ WRB), which is not uncommon for other criteria in other diagnostic horizons and properties. The Ck horizon has the overall quantity of CaCO_3 equivalent required (15% or more, by weight, in the fine-earth fraction) and much evidence of its illuviation: its CaCO_3 equivalent is 5% or more (absolute) higher than that of the underlying C-horizon; moreover, 5% or more (by volume) are identifiable secondary carbonates.

Table 4 La Higuera Soil Profile: Calcic Haploxeralf (Soil Survey Staff, 2014), Calcic Chromic Luvisol (Loamic, Aric, Cutanic, Hypereutric, Ochric).

Site description										
Location: Experimental farm "La Higuera" (CSIC), Toledo, Spain						Stoniness: no stones on the surface or within the profile				
Altitude: 460 m above sea level						Moisture regime: Xeric				
Landform: Plain (1% slope), East facing						Temperature regime: Thermic				
Parent material: Arkosic sandstones and clays (Miocene)						Drainage: Moderately well drained				
Land use: rainfed winter cereals						Clay minerals in Bt horizon: Illites, smectites and kaolinites				

	Sand %	Silt %	Clay %	EC 1:5 dS/m	CaCO_3 eq %	Base saturation %	pH 1:2.5	Org C %	CEC $\text{cmol} + \text{kg}^{-1}$	CEC/clay $\text{cmol} + \text{kg}^{-1}$
Horizon										
Ap1 (0–10 cm)	76	5	19	0.2	0	47	5.9	0.8	11	58
Ap2 (10–36 cm)	69	9	22	0.1	0	53	6.6	0.3	15	68
Bt (36–78 cm)	50	12	39	0.3	0	61	7.5	0.3	36	92
Ck (78–108 cm)	37	25	36	0.1	19	100	7.8	0.1	41	120
C (>108 cm)	41	21	38	0.1	11	100	7.9	0.1	30	80
	10YR 6/4 (dry) and 10YR 5/4 (wet), clay loam, massive structure, few (<5% v/v) soft accumulations of secondary carbonates									

		DIAGNOSTIC HORIZONS			
		STS:	OCHRIC	ARGILLIC	CALCIC
	Genetic horizons	WRB:	None	ARGIC	CALCIC
	Ap1 0–10 cm		(18 cm) Value w = 4.5>3 Value dry= 6 > 5 Org C=0.5%<0.6 % (w/w) cutoff	Elluvial horizons 22 x 1.2=26.4 % cla ST cutoff 22 x 1.4=30.8 % clay WRB cutoff	CaCO ₃ None
	Ap2 (10–36 cm)				
Bt (36–78 cm)		39% clay >> ST and WRB cutoff Common (10% v/v) clay coatings Not natric	CaCO ₃ None		
Ck 78–108 cm			CaCO ₃ =19% (w/w) > 15% cutoff Second. CaCO ₃ 20% (v/v) > 5 % cutoff		
C >108 cm			CaCO ₃ =11% w/w		

These diagnostic horizons, related to processes of formation, will be essential to classify soils at the first hierarchical levels of both taxonomies. For the WRB key, with the diagnostic horizons obtained (an argic not permeated with secondary carbonates above a calcic one) we can place the soil profile at the first level within the Luvisol Reference Soil Group. To classify to the second level of the system (IUSS Working Group WRB, 2015) it is necessary to use not only diagnostic horizons but also diagnostic properties and diagnostic materials. We need to check if the soil conforms to the definitions given to be used as principal and supplementary qualifiers. Reviewing the lists of the principal qualifiers, we can choose the following:

- Chromic (by the reddish color of a subsurface horizon), and
- Calcic (by the presence of the “calcic” diagnostic horizon)

Between the supplementary qualifiers, we can choose:

- Loamic (by the loamy texture),
- Aric (being ploughed to a depth of ≥ 20 cm from the soil surface),
- Cutanic (by the abundance of clay coatings),
- Hypereutric (by the relatively high base saturation between 20 and 100 cm) and.
- Ochric (by the moderate organic C content in the surface horizon, Ap).

Therefore, the profile used here as an example is classified as a Chromic Calcic Luvisol (Loamic, Aric, Cutanic, Hypereutric, Ochric) at the second level of the WRB system (IUSS Working Group WRB, 2015). Note that, the principal qualifiers are added before the name of the RSG without brackets and without commas. The sequence is from right to left, i.e., the lowermost qualifier in the list (here Calcic) is placed closest to the name of the RSG. The supplementary qualifiers are added in brackets after the name of the RSG always placed in the order of the alphabet and separated from each other by commas.

To classify this soil by ST (Soil Survey Staff, 2014), we also need to know the horizons and diagnostic characteristics but, moreover, we need to have some climate data, even for the higher categories, especially the soil moisture regime. In the profile taken as an example in this section, we have: a surface diagnostic horizon or epipedon (ochric), and two subsurface diagnostic horizons (argillic and calcic) with a xeric soil moisture regime, which allows us to classify the soil as Calcic Haploxeralf, at subgroup level (Soil Survey Staff, 2014). Pay attention to the fact that this nomenclature picks up the hierarchical structure of the system. Soil names in Soil Taxonomy are formulated from right to left, with formative elements (segments or syllables) drawn from the diagnostic horizons previously listed. This nomenclature system can be illustrated for this Calcic Haploxeralf:

<i>Hierarchical level</i>	<i>Order</i>	<i>Suborder</i>	<i>Great Group</i>	<i>Subgroup</i>
Name	Alfisol	Xeralf	Haploxeralf	Calcic Haploxeralf
Formative elements	Alf + i + sol	Xer + alf	Haplo + xeralf	Qualifier + Group
<i>Segment or syllable</i>	<i>Calcic</i>	<i>Haplo +</i>	<i>xer +</i>	<i>alf</i>
Hierarchical level	Subgroup	Great group	Subgroup	Order

Summary

This chapter highlights how to classify soils according to their properties, which are generally related to soil formation and management, following the world benchmarks: the WRB and Soil Taxonomy systems. These classifications are dynamic, and change over time to the extent that our knowledge about the soil does. Naming the soil synthesizes all this knowledge through a few qualifiers. On the contrary, interpreting the name of a soil requires being able to identify its components, properties and capabilities for different uses (agricultural, pastoral, forestry, etc.). For as long as we continue to cultivate the land, to carry out scientific experiments on the soils and to transfer the knowledge obtained in them, the classification of soils will be required.

References

- Arnold RW (2005) Classification of soils. In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*. vols. 1–4, pp. 204–210. Oxford/St Louis, MO: Elsevier. 2119 pp.
- Bridges EM, Batjes NH, and Nachtergaele FO (eds.) (1998) *World Reference Base for Soil Resources: Atlas*, 79. ISRIC-FAO.ISSS. Acco ed., Leuven, Belgium.
- Brown DJ (2005) Classification systems: USA. In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*. vols. 1–4, pp. 235–245. Oxford/St Louis, MO: Elsevier. 2119 pp.
- Deckers JA, Spaargaren FO, and Nachtergaele F (1998) *World Reference Base for Soil Resources: Introduction*. 165. ISRIC-FAO.ISSS. Acco ed., Leuven, Belgium.
- FAO (1974) *Soil map of the World*. vol. 1, Rome: FAO-UNESCO59. Legend. E. 1:5,000,000.
- FAO (1988) Soil map of the world: Revised legend. *World Soil Resources Report*. vol. 60. Rome: FAO-UNESCO.
- FAO (2006) *Guidelines for Soil Description*, 4th edn, Rome: FAO97.

- FAO-ISRIC-ISSS (1998) World reference base for soil resources. In: *World Soil Resources Report*, p. 84. Rome: Food and Agriculture Organization of the United Nations, International Soil Reference and Information Centre, and International Society of Soil Science.
- IUSS Working Group WRB (2007) *World Reference Base for Soil Resources 2006, first update 2007*. World Soil Resources Reports No. 103 Rome: FAO.
- IUSS Working Group WRB (2015) *World Reference Base for Soil Resources 2014, Update 2015. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*. World Soil Resource Reports, No 106 Rome: FAO. <http://www.fao.org/soils-portal/soil-survey/soil-classification/world-reference-base/en/>.
- Kubiens WL (1952) *Claves sistemáticas de suelos - Diagnóstico y sistemática ilustrados de los suelos más importantes de Europa con sus sinónimos más usuales*. Madrid, Spain: Instituto de Edafología- Consejo Superior de Investigaciones Científicas 388.
- Nachtergaele FO (2005) Classification systems: FAO. In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*. Vols. 1–4, pp. 216–222. Oxford/St Louis, MO: Elsevier. 2119 pp.
- Porta J, López-Acevedo M, Roquero C (2003) *Edafología para la agricultura y el medio ambiente*. 3ª edición. Mundi-Prensa: Madrid.
- Schoeneberger PJ, Wysocki DA, Benham EC, and Soil Survey Staff (2012) *Field Book for Describing and Sampling Soils, Version 3.0*. Lincoln, NE: Natural Resources Conservation Service, National Soil Survey Center.
- Soil Survey Staff (1975) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 1st edn, Washington, DC: US Government Printing Office. USDA–NRCS, Agriculture Handbook No. 436.
- Soil Survey Staff (1983) *Keys to Soil Taxonomy*, 1st edn. Washington, DC: US Department of Agriculture, Natural Resources Conservation Service.
- Soil Survey Staff (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn, Washington, DC: Natural Resources Conservation Service 869. U.S. Department of Agriculture Handbook 436.
- Soil Survey Staff (2014) *Keys to Soil Taxonomy*, 12th edn. Washington, DC: USDA-Natural Resources Conservation Service.
- Switoniak M, Kabala C, Karklins A, and Charzynsky P (2018) *Guidelines for Soil Description and Classification: Central and Eastern European Student's Version*. Ed. Torun, Poland: Polish Society of Soil Science 286.

Relevant websites

- <http://www.fao.org/soils-portal/soil-survey/soil-classification/world-reference-base/en/>—FAO.
- <http://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/faunesco-soil-map-of-the-world/en/>—FAO.
- https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/class/taxonomy/?cid=nrcs142p2_053580—USDA.
- www.cienciadelsuelo.es—Universidad de Zaragoza.
- www.suelosdearagon.com—Universidad de Zaragoza.
- <https://soilgrids.org>—International Soil Reference and Information Centre.

World reference base for soil resources (WRB)

Stephan Mantel^a, Stefaan Dondeyne^b, and Seppe Deckers^c, ^aISRIC – World Soil Information, Wageningen, The Netherlands;

^bDepartment of Geography, Ghent University, Ghent, Belgium; ^cDepartment of Earth and Environmental Science, University of Leuven, Leuven, Belgium

© 2023 Elsevier Ltd. All rights reserved.

Historical background	206
Development of the WRB	207
Structure	207
The reference soil groups	207
The qualifiers	207
The diagnostics	208
Naming soils	208
An example	213
Buried soils	214
Major changes in the WRB 2022 edition	214
The WRB and soil taxonomy	215
Application of the WRB	215
Creating map legends	216
The WRB in education	216
References	216

Abstract

The World Reference Base (WRB) is an international soil classification system that evolved from the FAO–UNESCO Legend of the Soil Map of the World. It is intended to name both natural and human-made soils within the first two meters of the Earth’s terrestrial surface. Soils are classified by diagnostic features, as in all modern classification systems. The WRB uses two hierarchical levels only: 32 Reference Soil Groups (RSGs) at the first level and an undetermined number at the second level as soils are named by combining the RSG with qualifiers. The WRB is continuously being tested and updated, with the latest edition presented in 2022 at the 22nd World Congress of Soil Science in Glasgow.

Key points

- WRB evolved from the FAO-UNESCO Legend of the Soil Map of the World
- WRB allows naming any part of the upper 2 m from the terrestrial Earth’s surface and facilitates correlation between national classification systems
- Soils are classified based on diagnostic features (horizons, properties, and materials)
- WRB has only two hierarchical levels of classification with 32 Reference Soil Groups (RSGs) at the highest level
- At the second level, soils are named by adding qualifiers to the RSG; the qualifiers are selected from lists making it a flexible and open system.
- How WRB correlates with Soil Taxonomy is explained
- Examples of applications from WRB are presented

Historical background

The international soil classification system, the World Reference Base for Soil Resources (WRB), evolved from the FAO-UNESCO legend of the Soil Map of the World (FAO, 1988; FAO-UNESCO, 1971). The FAO–UNESCO soil maps were published in the 1970s–1980s. After the Second World War, out of concern to meet world food demands, many countries invested in the survey of soils in support of the agricultural sector. With the lack of an international soil classification system, individual countries elaborated upon national soil survey protocols and classification systems which hampered global or even regional comparisons.

In 1960, the Seventh Congress of the International Soil Science Society (ISSS, now the International Union of Soil Sciences – IUSS) in Madison, Wisconsin, USA, recommended that soil maps of large regions and continents should be produced. The FAO and UNESCO decided to create a Soil Map of the World in 1961. It took worldwide collaboration between innumerable soil scientists to produce the maps over about 20 years, and this was the first global inventory of natural resources at a scale as detailed as 1:5 million (FAO–UNESCO, 1974). The ISSS recommendations for the Soil Map of the World project were substantiated with six

objectives (FAO, 1988; Nachtergaele, 2005): to (i) make a first appraisal of the world's soil resources; (ii) supply a scientific base for the transfer of experience between areas with similar environments; (iii) promote the establishment of a generally accepted soil classification system and nomenclature; (iv) establish a common framework for more detailed investigations in developing areas; (v) prepare a document for educational, research, and development activities, and (vi) strengthen international contacts in the field of soil science. The third objective, promotion of a nomenclature and soil classification system that is widely accepted, aligns with the WRB's evolution into an international classification system.

Development of the WRB

As a first step toward the creation of a global soil classification system, an international group of soil scientists launched the "International Reference Base for Soil Classification" project in 1980. The project was endorsed by the ISSS in 1982 and the International Reference Base for Soil Classification (IRB) working group was established within Commission V of the ISSS. In 1992, the IRB was renamed to World Reference Base for Soil Resources (WRB), and its working group was created at the Fifteenth World Congress of Soil Science in Acapulco in 1994.

The prime objective of the working group was to provide scientific depth and background to the Revised FAO Legend and to ensure that the latest knowledge relating to global soil resources and interrelationships was incorporated in a world-wide soil reference system (Spaargaren and Deckers, 1998). The first edition of the WRB (ISSS-ISRIC-FAO, 1998) together with explanatory texts (ISSS Working Group RB, 1998a, b) were presented at the Sixteenth World Congress of Soil Science in Montpellier, France. The great interest in the WRB transpires from the fact that it was translated into 15 languages. Furthermore, the Lecture Notes on the Major Soils of the World (Driessen et al., 2001), based on the WRB, were widely used. A second edition of the WRB was published in 2006, with an update in 2007 (IUSS Working Group WRB, 2007), a third update in 2014, and one in 2015 (IUSS Working Group WRB, 2015). At the World Congress of Soil Science in Glasgow, July–August 2022, the fourth edition was presented (IUSS Working Group WRB, 2022).

The WRB has grown beyond a system for correlation between national classification systems. The European Union adopted WRB for soil maps and for land information databases and also published a Soil Atlas of Europe based on it (Jones et al., 2005). Several countries use the WRB as a soil classification system and for mapping, or for updating their legacy soil survey data as, for example, in Belgium (Dondeyne et al., 2014), Hungary (Dobos et al., 2019) and Nepal (Lamichhane et al., 2021).

The WRB working group is part of the IUSS Division 1 (Soils in Space and Time) and falls under Commission 1.4. (Soil Classification). It consists of an international group of soil scientists testing, revising, and improving WRB as an academic service to society. Comments and suggestions from those applying WRB are greatly appreciated and field excursions are organized annually to test the system further. These excursions provide the basis for reviewing and revising the rules and definitions and are essential for updating and improving the system.

Structure

At the highest level of classification the WRB has 32 Reference Soil Groups (RSGs) which are identified through a key. The RSGs group soils that share dominant soil forming processes and which are convenient for making generalized global soil maps (Fig. 1). At the second level, soil names are constructed by adding qualifiers to the RSGs. The qualifiers represent morphologic, physical, or chemical characteristics that strongly affect the soil's functioning. Some qualifiers are specific to a few or even just one RSG while other qualifiers are foreseen to be combined with various RSGs.

The reference soil groups

The basic philosophy of the WRB is that Reference Soil Groups must represent a minimal geographical coverage and reflect some dominant soil-forming processes (Spaargaren and Deckers, 1998). In Table 1, the RSGs are aggregated into nine sets according to dominant soil forming processes and identifiers. Striking characteristics, and limitations relevant for land husbandry are also indicated for each RSG.

The qualifiers

What sets WRB apart from other soil classification systems is that at the second, most detailed, level soils are named according to a flexible, open system. In the current (4th) edition of WRB (IUSS Working Group WRB, 2022) 281 qualifiers have been defined for naming soils at the second level. Soils are differentiated according to features resulting from secondary soil-forming processes that have altered the primary characteristics and that affect land-uses, or, more generally, soil functions. Some qualifiers are specific to one or only a few RSGs while others can be combined with many RSGs. The qualifiers are divided into principal and supplementary qualifiers. Principal qualifiers are regarded as the most pertinent in relation to soil functioning and are ranked according to their perceived importance. Supplementary qualifiers provide additional details about the RSG. They are not ranked but listed in alphabetical order (except for the supplementary qualifiers pertaining to the texture which are given first). Many qualifiers are mutually exclusive, and, in practice, most classified soils have fewer than 10 qualifiers. When a soil is classified for which a qualifier would apply that is not foreseen in the list, it can still be added as a supplementary qualifier. In this way the classification system is flexible and open.

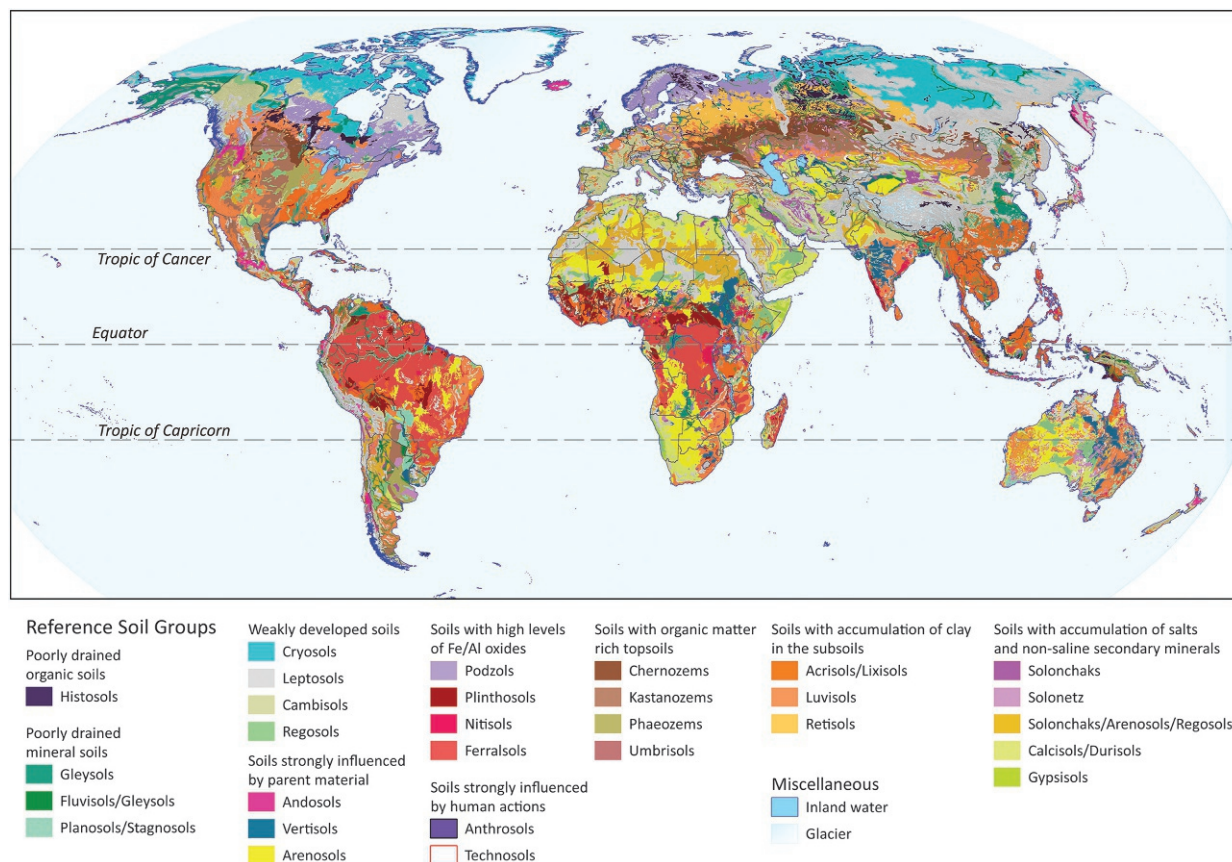


Fig. 1 The Reference Soil Groups represent soils that share some dominant soil forming processes and which are convenient for making generalized global soil maps [authors' mapped based on the Harmonized Soil Map of the World (FAO-IIASA-ISRIC-ISSACS-JRC, 2012)].

The diagnostics

The classification of soils in the WRB is based on soil features defined as diagnostic horizons (Table 2), diagnostic properties, and diagnostic materials (Table 3), together called the diagnostics. The WRB diagnostics require a minimum expression of one or more characteristics (or a combination of them), occurring within a specified depth. These characteristics are assessed in the field or may require laboratory analyses.

A diagnostic horizon is a soil layer, with distinct properties, or consisting of particular materials with a minimum thickness. A vertic horizon, for example, must have at least 30% clay, and must have at least 20% by volume of wedge-shaped soil aggregates or slickensides on more than 10% of the surfaces of soil aggregates, and must have shrink-swell cracks, and must have a thickness of 25 cm or more. Typical slickensides, wedge-shaped aggregates, and shrink-swell cracks of a Vertisol are shown in Fig. 2.

Diagnostic properties reflect soil-forming processes, or soil-forming conditions, but are not defined by the thickness of the particular layer. Diagnostic materials describe the nature of the soil material, without referring to the parent material. Table 3 presents the diagnostic properties and materials of the WRB.

Diagnostic properties can be observed or measured in the field or laboratory and require a minimum or maximum expression to qualify as diagnostic. Diagnostic materials significantly influence soil-forming processes. Their characteristics refer (as for all diagnostics) to the fine earth fraction (<2 mm), unless stated otherwise. Table 3 presents the diagnostic properties and materials of the WRB.

Within a soil profile a layer may have overlapped or diagnostics that coincide as illustrated by the soil profile in Fig. 3. The surface horizon (0–45 cm) is relatively thick (>20 cm), dark-colored and well structured, and has a large content of organic matter ($\geq 0.6\%$) and a high base saturation ($\geq 50\%$), as revealed by laboratory analyses. It therefore fulfills the criteria of a Mollic horizon. As the profile is developed in volcanic material, it contains volcanic glass, so it has vitric properties, and because it contains short-range-order minerals, it also has andic properties.

Naming soils

In naming a soil in the WRB the first step is to specify the diagnostics. These are identified by checking all the diagnostic criteria. A complete soil profile description is required to detect the diagnostics for a soil. To ensure that the recorded observations and data

Table 1 Reference Soil Groups aggregated according to dominant identifier and striking properties representing potentials and constraints for land husbandry. Note this sequence does not follow the key exactly. For the exact definition of the RSGs see (IUSS Working Group WRB, 2022).

Set	RSGs (code)	Dominant identifier and properties
1	Accumulation of organic matter	
	Histosols (HS)	Soils with very large organic matter contents that are water saturated or occur in areas with very high precipitation. Important for C sequestration, and biodiversity, and highly susceptible to subsidence and resulting into large emissions of CO ₂ and CH ₄ when drained
2	Mineral soils strongly influenced by human actions	
	Anthrosols (AT)	Ancient agricultural land; often important for C sequestration
	Technosols (TC)	Large amounts of artifacts; soil sealing, potentially high levels of contaminants and pollutants
3	Shallow and weakly developed mineral soils	
	Cryosols (CR)	Permafrost affected soils; important for C sequestration but with strong potential of CO ₂ and CH ₄ emission
	Leptosols (LP)	Rocky areas, or thin soils with many coarse fragments with usually very limited potential for agricultural production but valuable for nature conservation or forestry
	Regosols (RG)	Soils with no discernible profile development, of very varying potential for land-use and conservation
	Cambisols (CM)	Soils with the beginning of soil development, of very varying potential for land husbandry
4	Mineral soils with impeded drainage or formed from river, lake or sea deposits	
	Gleysols (GL)	Soils with shallow groundwater, mostly of good chemical fertility
	Planosols (PL)	Seasonally water saturated soils with lighter textured topsoil over heavy textured subsoil. Limited by poor drainage
	Stagnosols (ST)	Soils affected by seasonal stagnation of surface water but without strong textural differentiation. Limited by poor drainage
	Fluvisols (FL)	Soils formed from recent river, lake or marine deposits, subject to regular flooding but generally of high chemical fertility
5	Mineral soils strongly influenced by parent material	
	Andosols (AN)	Soils formed in volcanic materials, with mostly good chemical fertility but highly susceptible to soil erosion
	Vertisols (VR)	Soils with shrink-swell clay minerals of alternating wet and dry conditions, with mostly good chemical fertility but highly susceptible to gully erosion
	Arenosols (AR)	Soils of residual and shifting sands, with limited or no soil profile development, with mostly low chemical fertility and low water holding capacity
6	Mineral soils with high levels of Fe and Al oxides	
	Podzols (PZ)	Soils with accumulation of humus and or Fe–Al oxides in the subsoil. Soils with low fertility and strong acidity
	Plinthosols (PT)	Soils marked by the presence of iron-rich clayey material (plinthite) that hardens irreversibly upon drying, or that has hardened into ironstone (petroplinthite). Limited by drainage, stoniness or soil depth
	Nitisols (NT)	Deep, well-drained, reddish or yellowish soils with more than 30% clay with characteristic shiny, blocky or polyhedral structural elements. Good infiltration capacity, but susceptible to landslides
	Ferralsols (FR)	Deeply weathered reddish or yellowish soils with high content of Fe and Al oxides, and very low cation exchange capacity. Soils of low chemical fertility but often with good structural properties
7	Mineral soils with pronounced accumulation of organic matter in the topsoil	
	Chernozems	Soils with very dark organic matter rich topsoil and with secondary carbonates; soils with good infiltration capacity and highly fertile
	Kastanozems	Soils with brownish organic matter rich topsoil and with secondary carbonates; soils with good infiltration capacity and highly fertile
	Phaeozems	Soils with dark or brownish organic matter rich topsoil, without secondary carbonates, but with high base status and highly fertile
	Umbrisols	Soils with dark or brownish organic matter rich topsoil, with low base status (acidic soils)
8	Mineral soils with accumulation of salts, and non-saline secondary minerals	
	Solonchaks (SC)	Soils with high content of soluble salts in the topsoil. Very restricted plant growth, and highly susceptible to water and wind erosion
	Solonetz (SN)	Soils with high content of Na in the clay-enriched subsoil. Limited water infiltration capacity and susceptible to slaking and water erosion. High Na content is harmful to many plants
	Durisols (DU)	Soils with accumulation and cementation of secondary silica. Free draining soils mostly limited by shallow soil depth
	Gypsisols (GY)	Soils with accumulation of secondary gypsum. Limited by plant nutrient imbalance, soil depth, and stoniness
	Calcisols (CL)	Soils with accumulation of secondary carbonates. Can be highly productive, particularly under irrigation but can be constrained by limited soil depth and risk for salinization
9	Mineral soils with accumulation of clay in the subsoil	
	Retisols (RT)	Soils with a clay-enriched subsoil horizon that is interfingering by coarser-textured, lighter colored material. Can be limited by water infiltration capacity
	Acrisols (AC)	Soils with a clay-enriched subsoil horizon of low-activity clays and with low base status. Soils limited by low chemical fertility
	Lixisols (LX)	Soils with a clay-enriched subsoil horizon of low-activity clays and with high base status. Soils with relatively high chemical fertility
	Alisols (AL)	Soils with a clay-enriched subsoil horizon of high-activity clays and with low base status. Soils limited by high acidity
	Luvissols (LV)	Soils with a clay-enriched subsoil horizon of high-activity clays and with high base status. Soils with high chemical fertility

Authors' synthesis based on (Driessen et al., 2001; IUSS Working Group WRB, 2022).

Table 2 Descriptive overview of diagnostic horizons^a.

<i>Diagnostic horizons</i>	
<i>Name</i>	<i>Simplified Description</i>
1. Anthropogenic diagnostic horizons (all are mineral)	
Anthraquic horizon	In paddy soils: the layer comprising the puddled layer and the plow pan, both showing a reduced matrix and oxidized root channels
Hortic horizon	Dark, large content of organic matter and P, high animal activity, high base saturation; resulting from long-term cultivation, fertilization and application of organic residues
Hydragric horizon	In paddy soils: the layer below the anthraquic horizon showing redoximorphic features and or an accumulation of Fe and or Mn
Irragic horizon	Uniformly textured, at least moderate content of organic matter, high animal activity; gradually built up by sediment-rich irrigation water
Plaggic horizon	Dark, at least moderate content of organic matter, sandy or loamy; resulting from application of sods and excrements
Pretic horizon	Dark, at least moderate content of organic matter and P, large contents of exchangeable Ca and Mg, with black carbon; including Amazonian Dark Earths (also known as <i>Terra Preta</i>)
Terric horizon	Evidence of addition of substantially different material, at least moderate content of organic matter, high base saturation; resulting from adding mineral material (with or without organic residues) and cultivation
2. Diagnostic horizons that may be organic and or mineral	
Calcic horizon	Accumulation of secondary carbonates, not continuously cemented
Cryic horizon	Perennially frozen (visible ice or, if not enough water, < 0 °C)
Salic horizon	Large amounts of readily soluble salts
Thionic horizon	With sulfuric acid and a very low pH
3. Organic diagnostic horizons	
Folic horizon	Organic layer, not water-saturated and not drained
Histic horizon	Organic layer, water-saturated or drained
4. Surface mineral diagnostic horizons	
Chernic horizon	Thick, very dark-colored, high base saturation, moderate to high content of organic matter, well developed soil structure or structural elements created by cultivation, high biological activity (special case of the mollic horizon)
Mollic horizon	Thick, dark-colored, high base saturation, moderate to high content of organic matter, at least some soil structure or structural elements created by cultivation
Umbric horizon	Thick, dark-colored, low base saturation, moderate to high content of organic matter, at least some soil structure or structural elements created by cultivation
5. Other mineral diagnostic horizons related to the accumulation of substances due to (vertical or lateral) migration processes	
Argic horizon	Subsurface layer with distinctly higher clay content than the overlying layer without a lithic discontinuity and or presence of illuvial clay minerals (with or without a lithic discontinuity)
Duric horizon	Concretions or nodules, cemented by secondary silica, and/or fragments of a broken-up petroduric horizon
Ferric horizon	≥ 5% reddish to blackish concretions and or nodules or ≥ 15% reddish to blackish coarse masses, with accumulations of Fe (and Mn) oxides
Gypsic horizon	Accumulation of secondary gypsum, not continuously cemented
Limonic horizon	Accumulation of Fe and or Mn oxides in a layer that has or had gleyic properties; at least partially cemented
Natric horizon	Subsurface layer with distinctly higher clay content than the overlying layer without a lithic discontinuity and or presence of illuviated clay minerals; high content of exchangeable Na
Petrocalcic horizon	Accumulation of secondary carbonates, relatively continuously cemented
Petroduric horizon	Accumulation of secondary silica, relatively continuously cemented
Petrogypsic horizon	Accumulation of secondary gypsum, relatively continuous cemented
Petroplinthic horizon	Consists of oximorphic features inside (former) aggregates that are at least partially interconnected and have a yellowish, reddish and or blackish color; large contents of Fe oxides at least in the oximorphic features; relatively continuously cemented
Pisoplinthic horizon	≥ 40% at least moderately cemented yellowish, reddish, and or blackish concretions and/or nodules, with accumulation of Fe oxides
Plinthic horizon	Has in ≥ 15% of its exposed area oximorphic features inside (former) soil aggregates that are black or have a redder hue and a higher chroma than the surrounding material; large contents of Fe oxides, at least in the oximorphic features; not continuously cemented
Sombric horizon	Subsurface accumulation of organic matter other than in spodic or natric horizons; not a buried surface horizon
Spodic horizon	Subsurface accumulation of Al with Fe and or organic matter
Tsitelic horizon	Lateral accumulation of Fe, usually derived from Planosols and Stagnosols further upslope
6. Other mineral diagnostic horizons	
Albic horizon	Light-colored; loss of colored substances (e.g. oxides, organic matter) due to soil-forming processes
Cambic horizon	Evidence of soil-forming processes; not meeting the criteria of diagnostic horizons that indicate stronger alteration or accumulation processes
Cohesic horizon	Massive or subangular blocky structure, root penetration restricted, drainage normally free, rich in kaolinite, poor in organic matter
Ferrallic horizon	Strongly weathered; dominated by kaolinites and oxides
Fragic horizon	Structure that roots and percolating water penetrate only along interped faces; non-cemented
Nitic horizon	Rich in clay minerals and Fe oxides, moderate to strong structure, shiny soil aggregate surfaces
Panpaic horizon	Buried mineral surface horizon with a significant content of organic matter formed before having been buried
Protovertic horizon	Influenced by swelling and shrinking clay minerals
Vertic horizon	Dominated by swelling and shrinking clay minerals

^aThe actual definitions are more detailed and specific. For diagnostic criteria please refer to WRB, IUSS Working Group WRB (2022).

Table 3 Diagnostic properties and materials of the WRB^a.

<i>Diagnostic feature</i>	<i>Description</i>
1. Diagnostic properties related to surface characteristics	
Takryic properties	Fine-textured surface crust with a platy or massive structure, under arid conditions in periodically flooded soils
Yermic properties	Combination of desert features: desert pavement, varnishing, ventifacts, vesicular pores, platy structure
2. Diagnostic properties defining the relationship between two layers	
Abrupt textural difference	Very sharp increase in clay content within a limited depth range
Albeluvic glossae	Interfingering of coarser-textured and lighter colored material into an argic horizon forming vertically continuous tongues (special case of retic properties)
Lithic discontinuity	Differences in parent material
Retic properties	Interfingering of coarser-textured and lighter colored material into an argic or natric horizon
3. Other diagnostic properties	
Andic properties	Short-range-order minerals and or organo-metallic complexes
Anthric properties	Applying to soils with mollic or umbric horizons, if the mollic or umbric horizon is created or substantially transformed by humans
Continuous rock	Consolidated material (excluding cemented pedogenetic horizons)
Gleyic properties	Saturated with flowing or upwards moving groundwater (or upwards moving gases), permanently or long enough that reducing conditions occur
Protocalcic properties	Carbonates derived from the soil solution and precipitated in the soil (secondary carbonates), less pronounced than in calcic or petrocalcic horizons
Protogypsic properties	Gypsum derived from the soil solution and precipitated in the soil (secondary gypsum), less pronounced than in gypsic or petrogypsic horizons
Reducing conditions	Low rH value and or presence of sulfide, methane or reduced Fe
Shrink–swell cracks	Open and close due to swelling and shrinking of clay minerals
Sideralic properties	Relatively low CEC
Stagnic properties	Saturated with surface water (or intruding liquids), at least temporarily, long enough that reducing conditions occur
Vitric properties	≥ 5% (by grain count) of volcanic glasses and related materials, and containing a limited amount of short-range-order minerals and or organo-metallic complexes
4. Diagnostic materials related to the concentration of organic carbon or to organic artifacts	
Mulmic material	Developed from water-saturated organic material after drainage; 8–20% soil organic carbon
Mineral material	< 20% soil organic carbon and < 35% (by volume) organic artifacts
Organic material	≥ 20% soil organic carbon
Organotechnic material	< 20% soil organic carbon and ≥ 35% (by volume) organic artifacts
Soil organic carbon	Organic carbon that does not meet the diagnostic criteria of artifacts
5. Diagnostic material related to color	
Claric material	Light-colored fine earth, expressed by high Munsell value and low chroma
6. Technogenic diagnostic materials	
Artifacts	Created, substantially modified or brought to the surface by humans; no subsequent substantial chemical or mineralogical change of properties
Technic hard material	Consolidated and relatively continuous material resulting from an industrial process
7. Other diagnostic materials	
Aeolic material	Deposited by wind
Calcaric material	≥ 2% calcium carbonate equivalent, at least partially inherited from the parent material
Dolomitic material	≥ 2% of a mineral that has a ratio $\text{CaCO}_3/\text{MgCO}_3 < 1.5$
Fluvic material	Fluviatile, marine or lacustrine deposits with evident stratification
Gypsic material	≥ 5% gypsum, at least partially inherited from the parent material
Hypersulfidic material	Containing sulfides and capable of severe acidification
Hyposulfidic material	Containing sulfides and not capable of severe acidification
Limnic material	Deposited in water by precipitation (possibly with sedimentation), or derived from algae, or derived from aquatic plants with subsequent transport or subsequent modification by aquatic animals or microorganisms
Ornithogenic material	Excrements or remnants of birds or bird activity
Solimovic material	Heterogeneous mixture that has moved down a slope, suspended in water or by plowing; dominated by material that has undergone soil formation at its original place.
Tephric material	≥ 30% (by grain count) volcanic glass and related materials

^aFor definitions and diagnostic criteria, please refer to WRB, IUSS Working Group WRB (2022).

accord with the criteria and definitions of the WRB, in the 4th edition of the WRB (IUSS Working Group WRB, 2022) a field guide has been included. Once the diagnostics have been determined, they are compared with the key to identify the Reference Soil Group. For the second level of classification, qualifiers are put before and after the name of the RSG. The qualifiers that apply to the RSG are listed in the key. The principal qualifiers are added before the name of the RSG starting with the first principal qualifier that applies closest to the RSG name, and then the second and next ones consecutively left of the previous qualifier. The supplementary

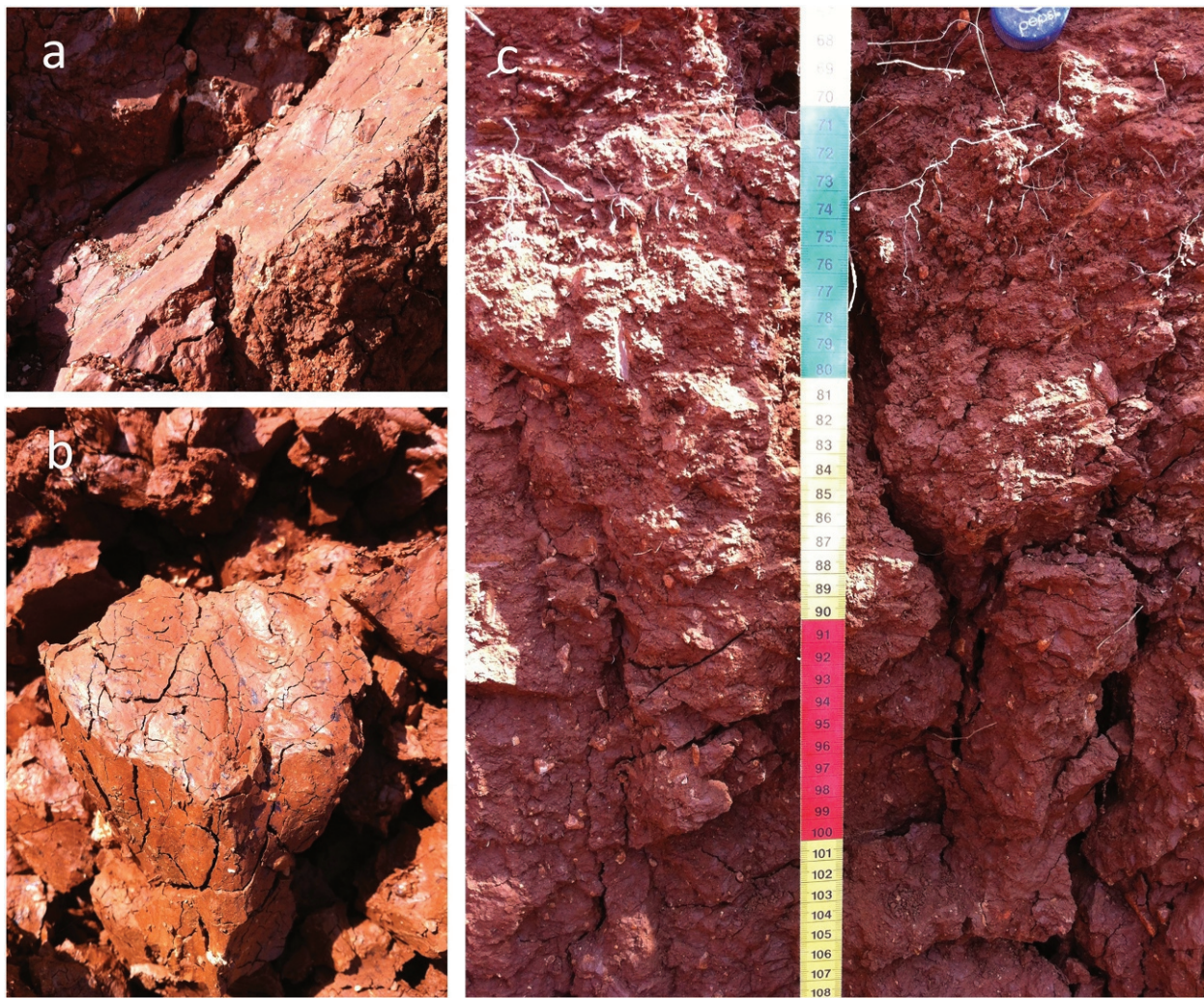


Fig. 2 Example of a diagnostic horizon: a vertic horizon is a clay-rich (>30%) subsurface horizon with slickensides and wedge-shaped soil aggregates that result from shrinking and swelling: (A): tilted shear plane of a slickenside; (B): wedge-shaped element, both A and B are formed by swelling of the soil under moist conditions. (C): deep shrink cracks formed when the soil is shrinking under dry conditions. Vertisol, Jordan. Images: Stephan Mantel.



Fig. 3 Diagnostics of a volcanic soil in Goðafoss, Iceland: a mollic horizon (0–45 cm) with vitric properties (volcanic glass 20–91 cm), and with andic properties (0–100 cm), with allophane, imogolite and similar minerals predominating (Silandic). The soil classifies in WRB2022 as a Mollic, Vitric, Silandic Andosol (Siltic, Eutrosilic, Humic) and in Soil Taxonomy as a Medial, amorphic Eutric Pachic Fulvicryand. Image: Stephan Mantel.

qualifiers are placed between brackets to the right of the name of the RSG. Qualifiers related to the texture, if applicable, are the first in the list. All other supplementary qualifiers follow these in alphabetical order and are separated by commas. Subqualifiers can be created, by addition of specifiers to the qualifier name, and added for specific conditions. For example, to specify that a particular qualifier occurs in the depth range of 0 to 50 cm, the specifier Epi- (from Greek *epi*, over) can be added. Other examples of specifiers are those related to buried conditions (Thapto-), properties occurring at greater depth (Bathy-) or weaker expression (Proto-). The position of the subqualifier in the name of the soil is regulated by specific rules.

Apart from the addition of a field guide for soil description, the 4th edition of the WRB differs from previous editions by the addition of new definitions such as: fine earth, litter layer, soil surface, mineral soil surface, soil layer and soil horizon. The field guide also includes an example field description sheet that is made available online. Many of the definitions in the new field guide were sourced from the USDA (United States Department of Agriculture) Soil Survey Manual (Soil Science Division Staff, 2017) and the NRCS (Natural Resources Conservation Service) Fieldbook (NRCS–USDA, 2012), which brings the WRB and Soil Taxonomy closer together.

An example

An example of a full classification in WRB 2022 is the soil in Fig. 4 from the steppes in Mongolia, that was described and classified during the WRB field workshop in 2019.

Name: Cambic **Kastanozem** (Pantoloamic, Endoskeletal, Endoraptic).

Description and diagnostics:

0–20 cm: Ah, mollic horizon, 10YR 3/3, sandy loam, pH 7.14;

20–40 cm: Bw, cambic horizon, sandy loam, pH 7.36;

40–55 cm: Ck, sandy loam with gravels, protocalcic, pH 7.61;



Fig. 4 A Cambic Kastanozem under short-grass steppe vegetation in Mongolia. Image: Stephan Mantel.

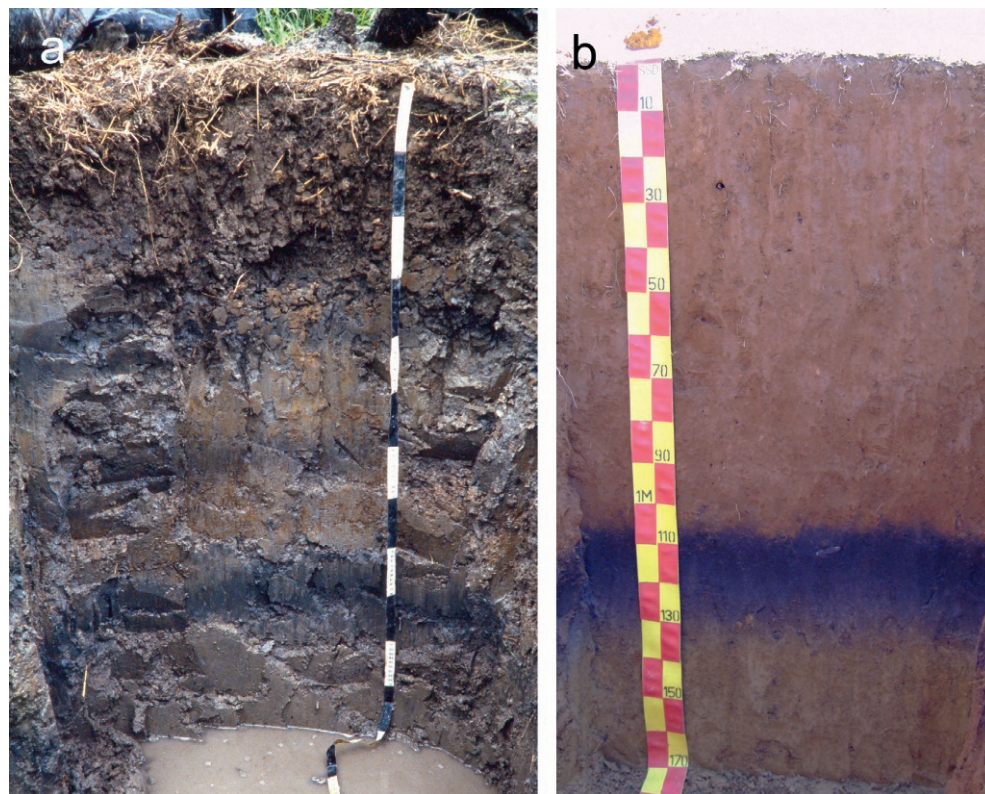


Fig. 5 Polygenetic soils, (A): multiple panpaic horizons (buried A-horizons) in a backswamp soil (Eutric Gleysol) in the Netherlands, (B): a panpaic horizon in an Andosol on a hillslope in Buthan. Images: ISRIC World Soil Information (l), Thomas Caspari (r).

55–90 cm: C, sandy loam, pH 7.85;
 90–140 cm: 2C, sandy loam, skeletal;
 140– (180) cm: 3C, loamy sand.

Buried soils

A special case in soil description and classification is that of buried soils. A soil is buried when it is covered by younger deposits. This may even be in repeated cycles. Fig. 5 shows examples of polygenetic soils with buried surface horizons consisting of mineral material and a significant amount of organic matter. In the 4th edition of WRB a buried organic-matter-rich A horizon is recognized as a panpaic horizon. This diagnostic horizon occurs in various landscapes and is the result of geomorphological processes, i.e. deposition of transported material.

In the case of buried soils, the following rules apply. The overlying material and the buried soil are classified as one soil if together they qualify as a Histosol, Anthrosol, Technosol, Cryosol, Leptosol, Vertisol, Gleysol, Andosol, Planosol, Stagnosol, Arenosol, Fluvisol or Regosol. Otherwise, the overlying material is classified with preference, if it is ≥ 50 cm thick (unless the overlying soil is a Regosol). In all other cases, the buried soil is classified with preference. If the overlying soil is classified with preference, and the underlying soil is not a Regosol or Leptosol, and shows a complete horizon sequence, and one soil does not influence the pedogenic processes in the other soil, then the name of the buried soil is placed after the name of the overlying soil adding the word 'over' in between, e.g. Skeletic Umbrisol (Siltic) over Albic Podzol (Arenic). Otherwise, a buried diagnostic horizon or a buried layer with a diagnostic property is added with the Thapto- specifier to the name of the overlying soil. If the buried soil is classified with preference, the overlying material is indicated with the Novic qualifier.

Major changes in the WRB 2022 edition

In the 4th edition of the WRB some diagnostics have been removed, some new ones added, and definitions were improved where needed. Among the diagnostics that have been taken out are the fulvic and the melanic horizons (dark topsoils associated with allophane), as these concepts have not proven to be useful for understanding soil organic matter dynamics or soil quality (Wander, 2004). The aridic properties, a soil condition that is actually defined by climate characteristics, have also been taken out. The geric property is not retained as diagnostic, but can be expressed as a qualifier. The reformulation of the hyper- and hyposulfidic material

makes the term sulfidic redundant as a diagnostic. Newly introduced are: albic horizon (light colored and eluviated), cohesic horizon (dense subsurface horizon dominated by kaolinite), limonic horizon (with accumulation of Fe by capillary rise in groundwater soils, traditionally referred to as bog iron), panpaic horizon (buried A horizon), tsitelic horizon (has accumulation of Fe by subsurface flow), protogypsic properties (accumulation of secondary gypsum, not sufficient for a gypsic or petrogypsic horizon), aeolic material (deposited by wind), mulmic material (mineral material with a high content of soil organic carbon, derived from organic material), organotechnic material (containing large amounts of organic artifacts and relatively small contents of soil organic carbon in the fine earth).

Some diagnostic materials received new names, such as 'claric material' instead of 'albic material' (to avoid confusion with the re-introduced albic horizon), 'solimovic material' instead of 'colluvic material' (colluvium appeared to refer to many different concepts; see [Miller and Juilleret, 2016](#)). Various criteria in the diagnostics, in the key and in the definitions of the qualifiers were reformulated and refined.

The WRB and soil taxonomy

Both the WRB and Soil Taxonomy (ST) are endorsed by the IUSS as the two international soil classification systems. The WRB and Soil Taxonomy have a common basis in that the classification hinges on diagnostic features relating to soil morphology, and physical and chemical properties. Even though the diagnostics reflect soil formation processes, soils are classified based on precisely defined diagnostic features rather than on soil-forming processes, which was the case in older classification systems. A major difference between the WRB and Soil Taxonomy is that soil climate is not used as a differentiating characteristic in the WRB.

Furthermore, with six levels of soil classification Soil Taxonomy has a more elaborate hierarchical structure than WRB. The 12 soil orders at the highest level are further subdivided into suborders, great groups, subgroups, family, and series. Because of differing criteria, definitions and concepts between the RSGs and Soil Orders, it is not possible to have a one-to-one conversion between the two systems. However, it is possible to indicate similarities between some of the RSGs and Soil Orders; there is a good match between the Histosols (WRB/ST), Cryosols (WRB) and Gelisols (ST), Vertisols (WRB/ST), Andosols (WRB) and Andisols (ST), and Cambisols (WRB) and Inceptisols (ST). The Chernozems, Kastanozems and Phaeozems of the WRB, which are soils with different expressions of dark, organic-matter-rich A horizons, correspond to the Mollisols in Soil Taxonomy. The highly weathered tropical soils such as Ferralsols and Plinthosols have similarity with Oxisols in Soil Taxonomy (Plinthaquox for Plinthosols). The soils related to arid climates and saline soils in WRB, such as the Solonetz, Solonchaks, Gypsisols, Calcisols and Durisols, are mostly accommodated in the order of the Aridisols of Soil Taxonomy, or correspond to some suborder level, such as: Solonetz (WRB) with Natrargids (ST), Solonchaks (WRB) with Salorthids (ST), Gypsisols (WRB) with Gypsids (ST), Calcisols (WRB) with Calcids (ST) and Durisols (WRB) with Durids (ST). The soils with clay illuviation and different clay mineralogy and base saturation in the WRB, the Nitisols, Retisols, Acrisols, Lixisols, Alisols, and Luvisols, correspond to different suborders of the Ultisols and Alfisols in Soil Taxonomy. The young, little developed soils group in the WRB, the Leptosols, Fluvisols, Arenosols and Regosols, correspond to the Entisols of Soil Taxonomy. A special case is the Umbrisols (soils with a dark topsoil and low base status) which in Soil Taxonomy will mostly be classified as Inceptisols. Anthropogenic soils, or soils with human-transported material in Soil Taxonomy, are recognized with a combination of taxa at the subgroup and family levels ([Galbraith and Shaw, 2017](#)), such as Plaggic Anthroisol (WRB) corresponding to the Plagganthrept and Plaggept (ST).

Application of the WRB

The term Reference Base indicates the common denominator function of the WRB, whereby the RSGs have sufficient width to facilitate harmonization and correlation with national classification systems. In addition, the WRB also serves as a communication tool for compiling global soil databases and for the inventory and monitoring of the world's soil resources. The nomenclature used to distinguish soil groups retains terms that have been used traditionally or that can be introduced easily into common language. The wide acceptance of the WRB can be attributed to three factors: (1) it is not intended to replace any existing system, but rather to facilitate communication and correlation between national and local systems; (2) it does not attempt to provide names for detailed maps and so does not interfere with local initiative; (3) it is based on extensive field experience and thus forms a useful stratification of the soil universe at high to medium categorical levels ([Rossiter, 2007](#)).

The WRB is used in several countries and projects as a basis for soil classification and for soil mapping. Several countries have produced a subnational or national soil map based on the WRB for exchange and correlation, e.g. the Flemish region ([Dondeyne et al., 2014](#)), Hungary ([Dobos et al., 2019](#)). The WRB is accepted as the common classification system for the European Union. Other examples of production of WRB maps at the national scale in addition to the existing national soil maps are the [national soil map for England and Wales](#) and [Scotland](#), the soils of Italy that have been mapped in the WRB at different scales ([Costantini and Dazzi, 2013](#); [Napoli et al., 2005](#)) and of Romania at 1:200.000 scale ([Vlad et al., 2012](#)). Maps at continental and regional level have been produced of which the soil atlases edited by the Joint Research Centre of the European Commission are good examples; for Latin America and the Caribbean ([Gardi et al., 2014](#)), for Europe ([Jones et al., 2005](#)), the [Northern Circumpolar Region](#) ([Jones et al., 2010](#)) and for Africa ([Jones et al., 2013](#)).

Legacy maps and reports are also increasingly becoming available in digital format. The online libraries of [ISRIC](#) (International Soil Reference and Information Centre, now ISRIC World Soil Information) the [WOSSAC](#) (World Soil Survey Archive and Catalog) library of Cranfield University, The [library of the European Soil Data Centre](#) (ESDAC) of the Joint Research Centre and the online [FAO soils portal](#) all have digitized legacy documents accessible. These maps and data are increasingly being used as input in digital soil mapping approaches. For example, for mapping soil units (RSGs with principal qualifiers) at district level in Ethiopia ([Leenaars et al., 2020](#)) but also at global level in the [Soilgrids](#) global digital soil mapping system ([Hengl et al., 2017](#); [Poggio et al., 2021](#)). Alternatively, WRB soil mapping units have proven useful covariates for predicting soil organic carbon hotspots in Belgium ([Ottoy et al., 2022](#)). Various databases contain information on the FAO-revised legend and or WRB classification for soil profiles, such as the global database of the World Soil Information Service ([WoSIS](#)), the Africa Soil Profiles Database ([AfSP](#)), the Harmonized World Soil Database ([HWSD](#)), the Soil and Terrain Databases ([SOTER](#)). These databases have also been used to predict e.g. soil hydraulic properties of the Zambezi River Basin ([Kalumba et al., 2022](#)) or soil nutrient status in Africa ([Hengl et al., 2021](#)), for example.

Creating map legends

In classifying individual soils, all applicable (sub)qualifiers are used to name the soil. For the creation of map legends, the number and nature of qualifiers depend on the scale and purpose of the map. For maps at continental or regional scales only the dominant Reference Soil Group may be used, without qualifiers. For next larger map scales, the RSG plus the first applicable principal qualifier can be used. For more detailed maps, the RSG plus the first two applicable principal qualifiers can be considered. The principal qualifiers are indicated before the RSGs, just as for naming a soil. Further qualifiers may be added, depending on the purpose and scope of the map. These may be principal qualifiers from further down the list, not already used in the soil name, or supplementary qualifiers placed according to rules for supplementary qualifiers.

The WRB in education

Various courses are provided on the WRB, and educational materials have been compiled. Examples include courses provided by Nicolaus Copernicus University (Department of Soil Science and Landscape Management) in Toruń that have made available a series of [atlases and guidelines](#). Various courses are offered, often with field excursions and practice, such as a Spring School course by ISRIC World Soil Information, and the summer [WRB Soil Classification Intensive Field Workshop](#) organized in Toruń. Many Universities offer lectures dedicated to, or that include the WRB. The [Lecture Notes on the Major Soils of the World](#) ([Driessen et al., 2001](#)) has proven to be a great resource over many years for both students and lecturers as a guide to soils of the world, and has possibly contributed to the use and acceptance of WRB in teaching and research. The recently published textbook ‘Soils of the World’ ([Zech et al., 2022](#)) presents the world soil geography following the 3rd edition of the WRB. The World Soil Museum has made the soil reference collection of ISRIC World Soil Information accessible via components of the virtual soil museum, such as a [soil explorer](#) that provides access to high resolution images of soil monoliths and to the [ISIS](#) (ISRIC Soil Information System) soil reference database (including soil and site descriptions, physical and chemical data and field images) and a [world map](#) with soil profiles that can be viewed at the sample location (allowing one to zoom in at field level on remote sensing images). Online visitors can explore the World Soil Museum via an online [indoor tour](#). The University of Leuven has developed an [educational platform](#) for soils and landscapes of the world drawing on data from their own collections and from the ISRIC Soil Information System. The KU Leuven has also developed, with contributions from ISRIC, a Massive Open Online Course (MOOC) on [soils of the Tropics](#), in which various RSGs from WRB are highlighted. The Technical University of Munich has developed [instructional videos](#) on how to describe a soil according to the FAO Guidelines for Soil Description ([FAO, 2006](#)) and how to do a soil classification with the WRB (2014, update 2015).

References

- Costantini E and Dazzi C (2013) *The Soils of Italy*. vol. 1. Dordrecht: Springer. 354. <https://doi.org/10.1007/978-94-007-5642-7>.
- Dobos E, Vadrnai P, Kovács K, Láng V, Fuchs M, and Michéli E (2019) A novel approach for mapping WRB soil units—A methodology for a global SOTER coverage. *Hungarian Geographical Bulletin* 68: 157–175. <https://doi.org/10.15201/hungeobull.68.2.4>.
- Dondeyne S, Vanierschot L, Langohr R, Van Ranst E, and Deckers JA (2014) *The soil map of the Flemish region converted to the 3rd edition of the World Reference Base for soil resources* (41 map sheets at scale 1:40 000, 1 map sheet at 1:250 000). KU Leuven, Universiteit Ghent: Vlaamse Overheid, Brussels.
- Driessen PM, Deckers JA, Spaargaren OC, and Nachtergaele FO (eds.) (2001) *Lecture Notes on the Major Soils of the World*. Rome: FAO. World Soil Resources Reports no. 94.
- FAO (1988) *FAO-UNESCO Soil Map of the World. Revised Legend with Corrections*. Reprinted, with corrections, as Technical Paper 20, ISRIC, Wageningen, 1994 Rome: FAO.
- FAO (2006) *Guidelines for Soil Description*. Rome: Food and Agriculture Organization of the United Nations. 97.
- FAO-IIASA-ISRIC-ISSACS-JRC (2012) *Harmonized World Soil Database-Version 1.2*. Rome, Italy; Laxenburg, Austria: FAO & IIAS. 43.
- FAO-UNESCO (1971) *Soil Map of the World*. vol. 1–10. Rome, Italy: Food and Agriculture Organization of the United Nations.
- FAO-UNESCO (1974) *Soil Map of the World*. 1–5 000 000. Volume 1: Legend. Paris: UNESCO.

- Galbraith JM and Shaw RK (2017) Human-altered and human-transported soils. In: Ditzler C, Scheffe K, and Monger HC (eds.) *Soil Survey Manual*. 4th edn, *USDA Handbook 184th* edn, pp. 525–554. Washington, D.C.: USDA-Natural Resources Conservation Service. Government Printing Office. https://www.nrcs.usda.gov/wps/portal/nrcs/detailfull/soils/ref/?cid=nrcs142p2_0542.
- Gardi C, Angelini M, Barceló S, Comerma J, Cruz Gaistardo C, Encina Rojas A, ... Vargas R (eds.) (2014) *Atlas de suelos de América Latina y el Caribe*, p. 176. L-2995, Luxembourg: Comisión Europea - Oficina de Publicaciones de la Unión Europea.
- Hengl T, Mendes de Jesus J, Heuvelink GBM, Ruiperez Gonzalez M, Kilbarda M, Blagotić A, Shangguan W, Wright MN, Geng X, Bauer-Marschallinger B, Guevara MA, Vargas R, MacMillan RA, Batjes NH, Leenaars JGB, Ribeiro E, Wheeler I, Mantel S, and Kempen B (2017) SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One* 12(2): e0169748. <https://doi.org/10.1371/journal.pone.0169748>.
- Hengl T, Miller MAE, Křížan J, et al. (2021) African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Scientific Reports* 11(6130): 1–18. <https://doi.org/10.1038/s41598-021-85639-y>.
- ISSS, ISRIC, FAO (1998) *World Reference Base for Soil Resources*. World Soil Resources Report, 84 Rome: FAO. 88.
- ISSS Working Group RB (1998a) Introduction. In: Deckers JA, Nachtergaele FO, and Spaargaren OC (eds.) *World Reference Base for Soil Resources*, 1st edn. Leuven: ISRIC-FAO-ISSS. Acco.
- ISSS Working Group RB (1998b) Atlas. In: Bridges EM, Batjes NH, and Nachtergaele FO (eds.) *World Reference Base for Soil Resources*, 1st edn. Leuven: ISRIC-FAO-ISSS. Acco.
- IUSS Working Group WRB (2007) A framework for international classification, correlation and communication. In: *World Reference Base for Soil Resources 2006. First Update 2007*. Rome: FAO.
- IUSS Working Group WRB (2015) International soil classification system for naming soils and creating legends for soil maps. Update 2015. In: *World Reference Base for soil resources 2014. Update 2015*. Rome: FAO.
- IUSS Working Group WRB (2022) World reference base for soil resources. In: *International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*, 4th edn. International Union of Soil Sciences (IUSS): Vienna, Austria.
- Jones A, Montanarella L, and Jones R (eds.) (2005) *Soil Atlas of Europe*. Luxembourg: European Commission, Publications Office of the European Union.
- Jones A, Stolbovoy V, Tarnocai C, Broll G, Spaargaren OC, and Montanarella L (eds.) (2010) *Soil Atlas of the Northern Circumpolar Region*. Luxembourg: European Commission, Publications Office of the European Union.
- Jones A, Breuning-Madsen H, Brossard M, Dampha A, Deckers JA, Dewitte O, ... Zougmore R (eds.) (2013) *Soil Atlas of Africa*. Luxembourg: European Commission, Publications Office of the European Union.
- Kalumba M, Nyirenda E, Nyambe I, Dondeyne S, and Van Orshoven J (2022) Machine learning techniques for estimating hydraulic properties of the topsoil across the Zambezi river basin. *Land* 11: 591. <https://doi.org/10.3390/land11040591>.
- Lamichhane S, Kumar L, and Adhikari K (2021) Updating the national soil map of Nepal through digital soil mapping. *Geoderma* 394: 115041. <https://doi.org/10.1016/j.geoderma.2021.115041>.
- Leenaars JGB, Elias E, Wösten JHM, Ruiperez González M, and Kempen B (2020) Mapping the major soil-landscape resources of the Ethiopian Highlands using random forest. *Geoderma* 361. <https://doi.org/10.1016/j.geoderma.2019.114067>.
- Miller B and Juilleret J (2016) *Defining Colluvium and Alluvium: An Experiment to Discuss and Consolidate Perspectives*. <https://doi.org/10.13140/RG.2.1.4903.9769>.
- Nachtergaele FO (2005) Classification systems. In: *Encyclopedia of Soils in the Environment*, pp. 216–222. FAO.
- Napoli R, Costantini EAC, Castellani F, and Gardin L (2005) New proposals toward a WRB system for soil cartography: The soil map at 1: 250000 scale of the Tuscany region (Central Italy). *Eurasian Soil Science* 38: S20–S26.
- NRCS-USDA (2012) *The Field Book for Describing and Sampling Soils, version 3.0*. Schoeneberger, Wysocki, Benham, and Soil Survey Staff (2012). National Soil Survey Center, Natural Resources Conservation Service, US Department of Agriculture. Sept 2012, Reprint 2021.
- Otto S, Truyers E, De Block M, Lettens S, Swinnen W, Broothaerts N, Hendrix R, Van Orshoven J, Verstraeten G, De Vos B, and Vancampenhout K (2022) Digital mapping of soil organic carbon hotspots in nature conservation areas in the region of Flanders, Belgium. *Geoderma Regional* 30: e00531. <https://doi.org/10.1016/j.geodrs.2022.e00531>.
- Poggio L, de Sousa LM, Batjes NH, Heuvelink GBM, Kempen B, Ribeiro E, and Rossiter D (2021) SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *The Soil* 7: 217–240. <https://soil.copernicus.org/articles/7/217/2021/>.
- Rossiter DG (2007) Classification of urban and industrial soils in the world reference base for soil resources. *Journal of Soils and Sediments* 7(2): 96–100.
- Soil Science Division Staff (2017) *Soil survey manual*. In: Ditzler C, Scheffe K, and Monger HC (eds.) *USDA Handbook 18*. Washington, D.C.: Government Printing Office.
- Spaargaren OC and Deckers JA (1998) The world reference base for soil resources. In: Schulte A and Ruhayat D (eds.) *Soils of Tropical Forest Ecosystems*, pp. 21–28. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-03649-5_2.
- Vlad V, Florea N, Toti M, Daniela R, Seceleanu I, Vintila R, Cojocaru G, Cotet V, Dumitru S, Eftene M, Gherghina C-A, Ignat P, Mocanu V, and Vranceanu A (2012) Definition of the soil units of the 1:200,000 Soil Map of Romania using an extended terminology of the WRB system. *Annals of the University of Craiova XLII*: 615–639.
- Wander M (2004) Soil organic matter fractions and their relevance to soil function. In: Magdoff F and Weil R (eds.) *Soil Organic Matter in Sustainable Agriculture*. CRC Press. <https://doi.org/10.1201/9780203496374.ch3>.
- Zech W, Schad P, and Hintermaier-Erhard G (2022) *Soils of the World*. Berlin, Heidelberg: Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-540-30461-6>.

World soil map based on soil taxonomy

Eswaran Padmanabhan^a and Paul F Reich^b, ^aDepartment of Geosciences, University Technology PETRONAS, Seri Iskandar, Malaysia;
^bUSDA Natural Resources Conservation Service, Beltsville, MD, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	218
Soil classification	218
Global distribution	219
Gelisols	226
Histosols	226
Andisols	226
Spodosols	227
Oxisols	227
Vertisols	227
Aridisols	227
Ultisols	229
Mollisols	229
Alfisols	229
Inceptisols	229
Entisols	230
Conclusion	230
Acknowledgments	230
References	230

Abstract

Soil Taxonomy, a soil classification system, was created to help organize knowledge about soils. Soils are grouped into 12 orders in the system based on soil forming processes and soil attributes. Soil Taxonomy incorporates climate by using the soil moisture and the temperature regimes to define lower categories. The system enables soil–landscape relationships to be established and accounts for all known soils in addition to facilitating soil management in a sustainable manner. Soil Taxonomy is a powerful tool for soil quality assessments and mitigation measures.

Key points

- To enable the reader to be familiar with the basics of Soil Taxonomy
- To show how soil classification is linked to soil management
- To illustrate the importance of soil classification in maintaining soil health

Introduction

Soil is commonly described as the mantle or the skin covering the landmass of the Earth. For most people, soil is the natural medium for vegetation growth and its thickness determined by the rooting depth of plants. It is composed of solids (mineral and organic), liquids and gases. The upper limit of the soil is the contact between soil and air or shallow water, to a depth of several meters. Soil on the landscape is a continuum that eventually grades into non-soil materials such as rock, ice, or deep water.

Pedology is the study of soils which is conducted by pedologists or soil scientists. Conventionally, pedologists study the zone of the soil from its upper limit to a depth of about 200 cm, unless non-soil material occurs at a shallower depth. This is the zone where maximum biological or pedological processes occur. Occasionally, pedologists analyze the soil to depths of 20 m or more for special studies. There are five factors of soil formation: climate, topography, parent material, time, and biological activities (Jenny, 1941). Variations in these factors can result in a range of soils that differ in their physical, chemical, biological, and mineralogical properties. Therefore, it is important to know how soils form and what differences exist among them as further explored in Buol et al. (2011). Further information about soil forming processes and characteristics is available in other chapters of the Encyclopedia.

Soil classification

Soil classification systems were created to help organize knowledge about soils (Arnold et al., 1997; Ahrens and Arnold, 2000). An important difference between the classification of soils and that of plants and animals is that unlike the latter, soil is not discrete but a continuum and therefore the boundaries between named soils are abrupt and artificial. Depending on the objectives for which the classification was created, the class definitions and the structure of the classification system may vary (Arnold, 1990). Many countries have national soil classification systems but the two that are widely used and employed throughout the world are Soil Taxonomy (Soil Survey Staff, 1975, 1999; Wilding et al., 1983a,b) and the World Reference Base, IUSS Working Group WRB (2015).

Soil surveys enable the creation of maps of soil characteristics that provide information on the management of soil properties. These surveys provide the description, classification, and delineation of soils across the landscape. Specialized maps with soil classifications (Bartelli, 1978; Orvedal and Edwards, 1941) are produced for specific objectives (Benchmark Soils Project, 1982; USDA-SCS, 1991; Libohova et al., 2020). Soil maps are also produced at different scales: detailed maps (1:10,000 or larger scale) to show soils on farms and general maps (1:100,000 or smaller scale) to depict soils over large areas such as countries, continents, or the world. Map units appear as polygonal areas and group soils based on their similarities (Soil Science Division Staff, 2017).

To appreciate the variety of soils that occur on the planet, it is necessary to understand the classification system being used since many classification systems exist. These systems provide information on their physical and chemical properties from which an understanding of their potential specific uses can be based. Two classification systems that are widely used are the World Reference Base (Deckers et al., 1998; IUSS Working Group WRB, 2015; Schad, 2017) for soil classification that was initiated by the Food and Agriculture Organization of the United Nations (FAO), while Soil Taxonomy (Soil Survey Staff, 2014) is a system developed in the United States with the collaboration of international soil scientists and is a natural classification system. Nearly half of the world's countries use Soil Taxonomy as their National Soil Classification system, and many others use it in conjunction with other systems and include Soil Taxonomy in technical reports. In addition, Soil Taxonomy has been accommodating other soil classification systems by serving as a common platform for communication and exchange of technology for more than half a century.

Soil Taxonomy has six categories as listed in Table 1. The soil series is the lowest category and is defined with the maximum amount of detailed information. Information is generalized in the definition of each higher category. Consequently, each category can only be used for specific purposes. When employed to designate map units, the scale of the maps limits the category that is represented as shown in Table 1. The highest category, the order, has only 12 classes. There are 64 classes in the suborder, 300 classes in the great group, and 2400 classes in the subgroup (Soil Survey Staff, 2014). In the United States, 24,858 soil series have been established but there are no estimates for the total number of soil series in the world. Therefore, at each level, the amount and quality of information delivered for mapping and interpretative purposes is largely controlled by the scale of observation. Onsite soil management is done at the series level but as we move higher in the Taxonomy, the dominance of soil forming factors in defining the soil orders, for example, assists in defining appropriate strategies for management at that level.

Fig. 1 shows the global distribution of the 12 soil orders in Soil Taxonomy, while Fig. 2 shows soil profile examples of each order, and how remarkably different they are from each other. Soil forming processes (Table 1) and the attributes they provide for the soil are used to define the soil. Major differentiating criteria for each soil order have been generalized and listed in Table 2. Soils are generally composed of mineral and organic materials. In general, in Soil Taxonomy, it is the presence or absence of subsurface horizons and their characteristics that are used to separate the orders. For example, dominance of short-range order minerals such as allophane and imogolite in Andisols, smectitic clays in Vertisols, organic matter in Histosols, and low-activity clays in Oxisols, each impart characteristics that are used to define specific groupings of soils. The second group of soil processes listed in Table 2, have horizons of accumulation or depletion. The next group of soils (Inceptisols and Gelisols) also develop from soil forming processes, but the manifestations are slight or mainly physical. Finally, the last group contains the Entisols. These are soils where the original material is least altered or not modified by soil forming processes. Each order of soils also has specific subordinate properties that are used to define the lower categories.

Global distribution

Climate or hydrology controls many of the soil forming processes, therefore, each of these major kinds of soils has a specific place on the landscape (Fig. 3). However, each soil may have properties transitional to other soils and such secondary features are used to define the intergrades of the lower categories. Soil Taxonomy incorporates climate by using the soil moisture regime (SMR) as shown in Fig. 3 and the soil temperature regime (STR) to define lower categories. Soils with an aridic SMR (deserts) occupy about 30% of the total global land area. These soils are very dry and are only suitable for agriculture if there is a source of irrigation. Soils with a xeric SMR (Mediterranean like climate) occur in areas where precipitation occurs in winter, and like those with a ustic SMR (semi-arid) have moisture stress for prolonged periods during the year. These two groups of soils occupy about 28% of the land and form the semi-arid zones where crops require supplemental irrigation. The soils with an aquic SMR represent wetlands, where the soil is saturated for long enough to cause anaerobic conditions. About 24% of the total land is relatively free from moisture stress (udic and perudic SMRs) although some areas have low temperatures. The very cold permafrost (permanently frozen) soils, which occupy about 5.4% of the land, are the tundra regions. Soils with intermittent permafrost (interfrost) are generally under boreal forest, but are cultivated in some areas with an annual grain crop of short duration. Soils with mesic, thermic, isomesic, and

Table 1 Defining characteristics of the categories in Soil Taxonomy.

Category	Definition	Functions	Potential uses
Order	Soils having properties (marks) or conditions, resulting from major soil forming processes that are sufficiently stable in a pedologic sense and that help to delineate broad zonal groups of soils.	Depict zones where similar soil conditions have occurred for general understanding of global patterns of soil resources; Establish global geographic areas within which more specific factors and processes result in the diversity of soils.	General global/continental/regional assessment; Global climate change studies. AMS: <1:10,000,000 MSD: >40,000
Suborder	Soils within an order having additional properties or conditions that are major controls or reflect such controls, on the current set of soil forming processes and delineate broad ecosystem regions.	Demarcate broad areas where dominant soil moisture conditions generally result from global atmospheric conditions; Delineate contiguous areas with similar natural resource endowments; Ecosystems with distinct vegetation affinities usually determined by limiting factor/s of soil moisture or conditions.	Demarcate areas in regions or large countries for assessment and implementation of economic development; Analysis of international production and trade patterns; Priority setting for multi-purpose uses of land resources. AMS: 1:1,000,000 to 1:10,000,000 MSD: 4000–40,000
Great Group	Soils within a suborder having additional properties that constitute subordinate or additional controls, or reflect such controls, on the current set of soil forming processes including landscape-forming processes.	To demarcate contiguous areas with similar production systems or performance potentials.	Development of strategic plans for regional development; Basis for coordinating national resource assessment and monitoring programs; Infrastructure development to assure equity in development; AMS: 1:250,000 to 1:1,000,000 MSD: 250–4000
Subgroup	Soils within a great group having additional properties resulting from a blending or overlapping of sets of processes in space or time that cause one kind of soil to develop from, or toward, another kind of soil: <i>Intergrades</i> show the linkage to the great group, suborder, or order level; <i>Extrgrades</i> have sets of processes or conditions that have not been recognized as criteria for any class at a higher level, including non-soil features; The soil is considered as the “typic” member of the class if the set of properties does not define intergrades or extrgrades.	To demarcate production land units with similar land use and management requirements.	Targeting research and development for specific land use or cropping systems; Implementing conservation practices and ecosystem-based assistance; Community development projects and monitoring sustainability. Basis for diversification of agriculture and uses of land resources; Basis for implementing environmental management programs and modeling ecosystem performance. AMS = 1:100,000 to 1:250,000 MSD: 40–250
Family	Soils within a subgroup having additional properties that characterize the parent material and ambient conditions; The three most important properties are particle size, mineralogy, and the soil temperature regime.	Demarcate resource management domains characterized by similar management technology and production capabilities.	Basic units for extension/technology transfer; Modeling cropping systems’ performance; Addressing socioeconomic concerns. AMS: 1:25,000 to 1:100,000 MSD: 2.5–40
Series	Soils within a family having additional properties that reflect relatively narrow ranges of soil forming factors and processes, determined by small variations in local physiographic conditions, that transform parent material into soil.	To delineate land units for site-specific management of farms.	Implementing soil-specific farming; Designing farm-level conservation practices. AMS: >1:25,000 MSD: <2.5

AMS, appropriate map scale.

MSD, minimum size delineation; smallest area that can be delineated on a map with a legible identification; in hectares.

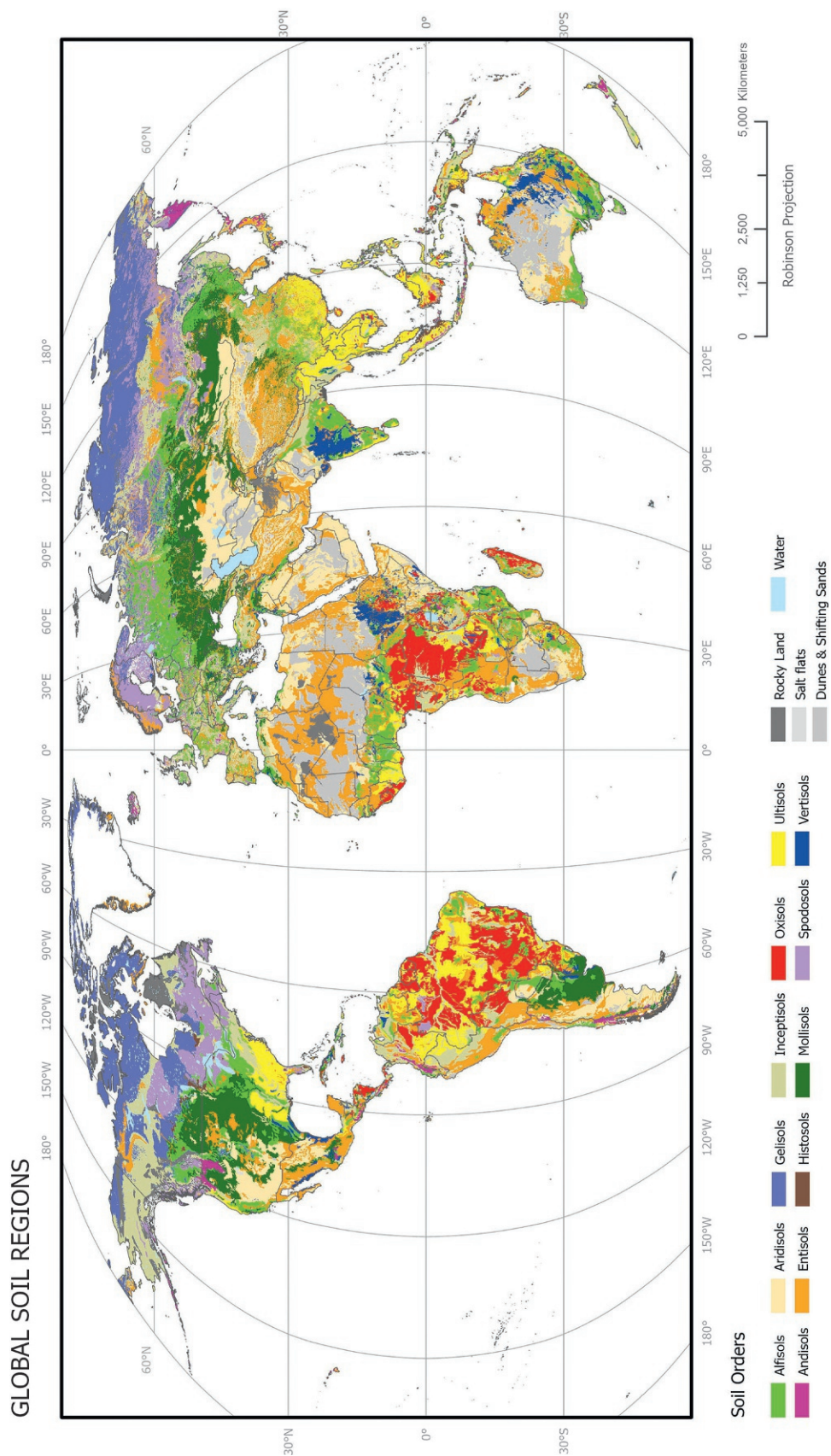


Fig. 1 A generalized soil map of the world.

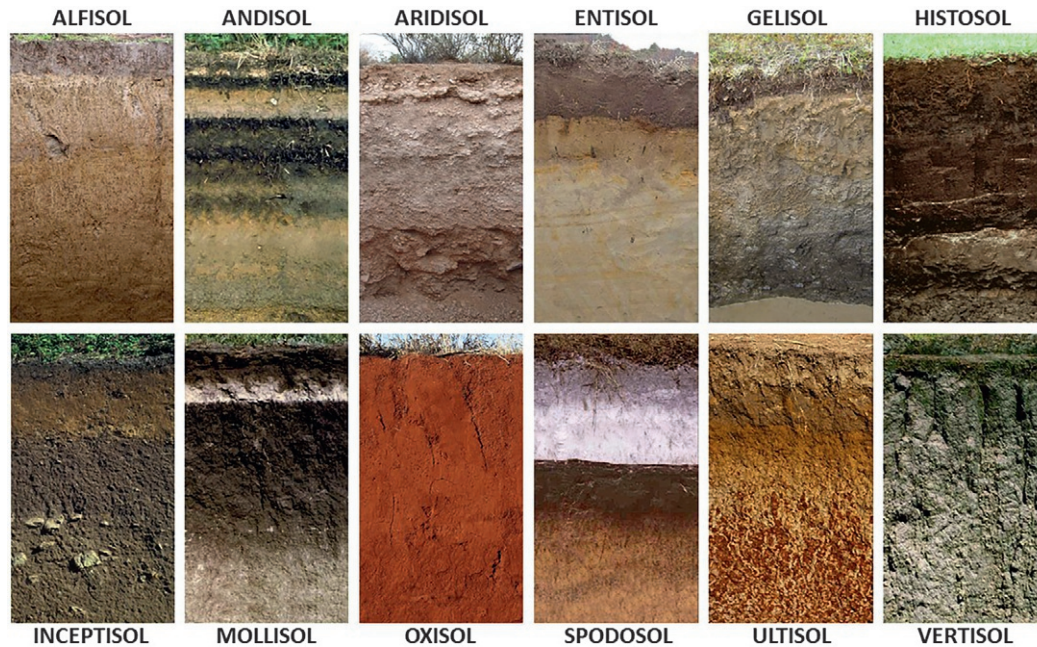


Fig. 2 Examples of the major soil orders.

Table 2 Grouping of soils based on major differentiating criteria.

<i>Defining properties determined by the soil material</i>	<i>Soils with a horizon resulting from accumulation or depletion of some soil component</i>	<i>Soils with a horizon formed by alteration of primary minerals or by slight changes in aberrant properties</i>	<i>Soils with only minimal development; they may retain the original features of the sediment (stratification)</i>
Histosols	Alfisols	Inceptisols	Entisols
Andisols	Ultisols	Gelisols	
Oxisols	Mollisols		
Vertisols	Spodosols		
	Aridisols		

isothermic STRs are generally those with the highest agricultural potential if adequate soil moisture is available. The humid tropical soils are represented by the combination of an isohyperthermic or isothermic STR and udic, perudic or aquic SMRs. These humid tropical soils occupy about 9% of the land surface and are not only some of the poorest soils for agriculture but are also subject to heavy incidence of pests and diseases. Each combination of SMR and STR represents a unique pedo-environment that supports the development of distinct soil properties.

The global soil regions map (Fig. 1) was created using the Harmonized World Soil Database (HWSD) (FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012) and modeled SMRs (Fig. 3) using the Newhall Simulation model (USDA-NRCS, 2011) with climatic data from <https://WorldClim.org> (Hijmans et al., 2005). Soil Taxonomy orders and suborders were assigned to each combination of the HWSD soil class and SMRs.

Estimates of the distribution of the soil orders and suborders across SMRs is presented in Table 3. Entisols occupy about 20% of the land, making it the largest single order in the world. Inceptisols occupy about 12% of the land. Even though these two soil orders are present in a range of climatic conditions, it appears that the Inceptisols are mostly dominated by those that occur in wet areas, whereas most of the Entisols occur under drier moisture conditions. The Aridisols (11% of the land) are dominated by soils that have special horizons such as calcic, petrocalcic, argillic, or natric horizons. The Histosols, Andisols, and Vertisols collectively occupy less than 4% of the total land area. These are soils formed in special materials and occupy specific geographic areas. The Gelisols (6.5%) are confined to the very cold regions of the world. Due to intense weathering and leaching in hot humid environments, most of the Oxisols (6%) and Ultisols (8%) are in tropical or subtropical zones. Although they are more extensive in the temperate regions, Alfisols (9%) and Mollisols (8%) are found under a wide range of SMRs and STRs. Inceptisols and Entisols, as mentioned earlier, are present under a range of climatic environments.

The properties used in Soil Taxonomy classification are easily observed in the field using reliable techniques such as the Munsell soil color chart to characterize the soil colors even in mottled horizons, measurements of soil temperature, permeability and even acidity. This includes properties that are seasonally dynamic such as fluctuations in soil moisture or temperature conditions,

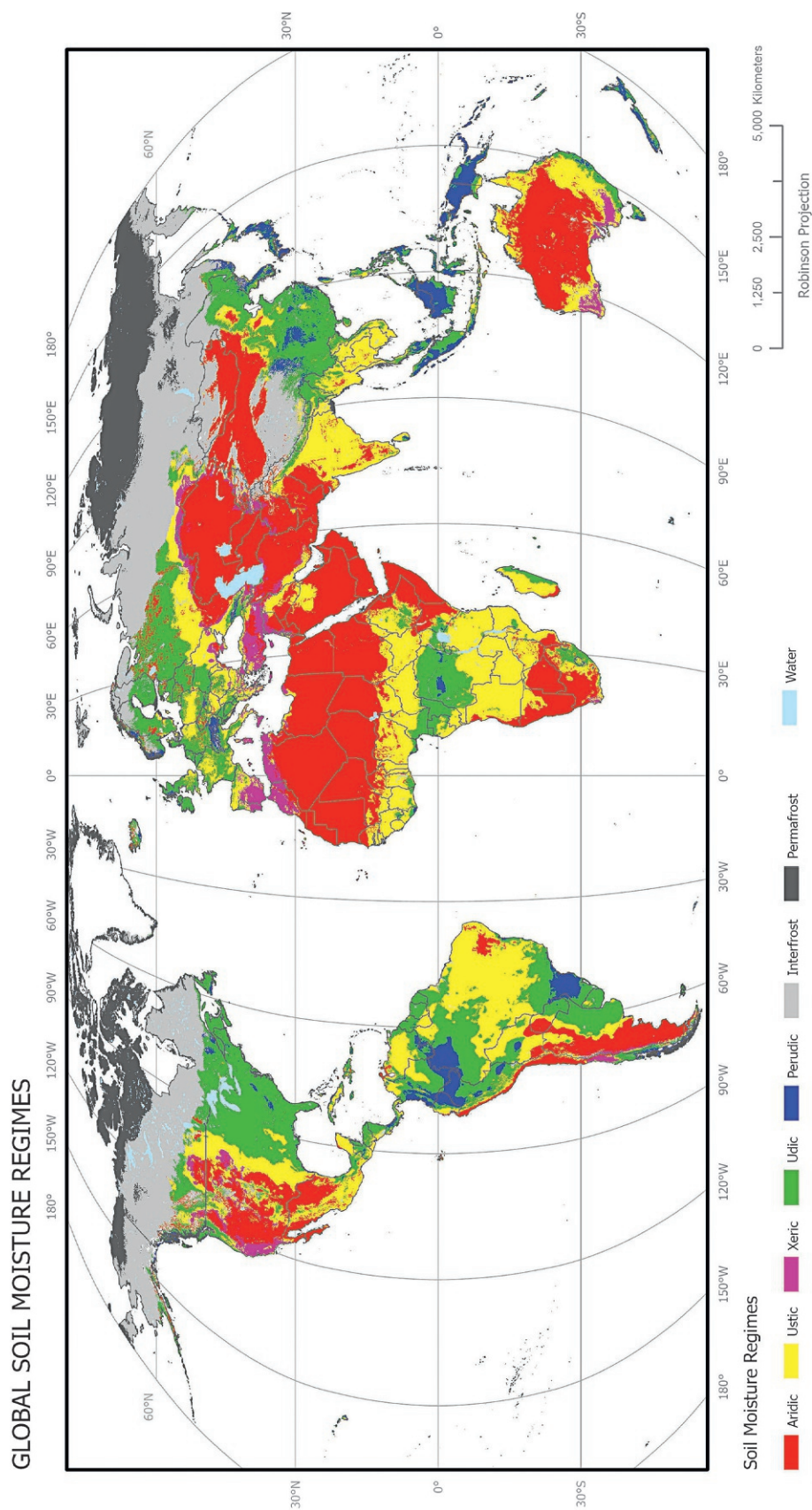


Fig. 3 Global soil moisture regimes map.

Table 3 Estimates of area occupied by major soils of the world.

Order	Suborder	Global		Aridic		Ustic		Xeric		Udic		Perudic		Interfrost		Permafrost	
		km ²	%	km ²	%	km ²	%	km ²	%	km ²	%	km ²	%	km ²	%	km ²	%
Gelisol	Histels	1,891,779	1.45	93	0.00	15,960	0.05	426	0.01	108,479	0.42	—	—	1,683,339	9.74	83,482	1.17
	Turbels	5,402,168	4.15	—	—	—	—	—	—	—	—	—	—	1478	0.01	5,400,691	76.00
	Orthels	1,162,214	0.89	729	0.00	911	0.00	—	—	5280	0.02	1105	0.02	758,297	4.39	395,892	5.57
		8,456,161	6.50	822	0.00	16,871	0.05	426	0.01	113,759	0.44	1105	0.02	2,443,114	14.13	5,880,064	82.75
	Folist	20,715	0.02	—	—	1912	0.01	—	—	18,803	0.07	—	—	—	—	—	—
	Fibrists	201,476	0.15	2218	0.01	10,153	0.03	312	0.01	178,361	0.69	10,432	0.20	—	—	—	—
	Saprist	368,878	0.28	8201	0.02	44,450	0.14	—	—	185,247	0.71	130,980	2.56	—	—	—	—
Histosol	Hemists	175,636	0.14	—	—	32,567	0.10	677	0.02	139,317	0.54	3076	0.06	—	—	—	—
		766,706	0.59	10,420	0.03	89,082	0.28	988	0.03	521,729	2.01	144,487	2.83	—	—	—	—
	Aquods	132,910	0.10	2420	0.01	6276	0.02	—	—	92,539	0.36	4079	0.08	27,596	0.16	—	—
	Gelods	3,940,476	3.03	—	—	—	—	—	—	—	—	—	—	3,196,478	18.49	743,998	10.47
	Cryods	1,648,389	1.27	198,887	0.51	13,893	0.04	—	—	1,342,061	5.17	93,547	1.83	—	—	—	—
	Humods	140,579	0.11	—	—	53,583	0.17	10,455	0.29	57,921	0.22	18,621	0.36	—	—	—	—
	Orthods	790,940	0.61	30,075	0.08	134,774	0.42	26,152	0.73	500,705	1.93	99,234	1.94	—	—	—	—
Spodosol		6,653,294	5.12	231,382	0.60	208,525	0.65	36,607	1.02	1,993,226	7.68	215,481	4.22	3,224,074	18.65	743,998	10.47
	Cryands	229,542	0.18	9466	0.02	88,948	0.28	1297	0.04	118,305	0.46	11,526	0.23	—	—	—	—
	Torrands	11,139	0.01	11,139	0.03	—	—	—	—	—	—	—	—	—	—	—	—
	Xerands	11,083	0.01	—	—	—	—	11,083	0.31	—	—	—	—	—	—	—	—
	Vitrands	143,645	0.11	10,309	0.03	49,992	0.15	—	—	50,094	0.19	33,250	0.65	—	—	—	—
	Ustands	62,402	0.05	—	—	62,402	0.19	—	—	—	—	—	—	—	—	—	—
	Udands	332,074	0.26	—	—	—	—	—	—	183,016	0.70	149,058	2.92	—	—	—	—
Andisol	Gelands	150,423	0.12	—	—	—	—	—	—	—	—	—	—	149,905	0.87	517	0.01
		940,308	0.72	30,914	0.08	201,342	0.62	12,380	0.34	351,415	1.35	193,834	3.79	149,905	0.87	517	0.01
	Aquox	917,536	0.71	2026	0.01	541,278	1.68	180	0.01	372,499	1.43	1552	0.03	—	—	—	—
	Torrox	59,909	0.05	59,909	0.15	—	—	—	—	—	—	—	—	—	—	—	—
	Ustox	3,089,620	2.38	—	—	3,087,452	9.56	2168	0.06	—	—	—	—	—	—	—	—
	Perox	832,531	0.64	—	—	—	—	—	—	270,871	1.04	561,660	10.99	—	—	—	—
	Udoox	3,320,360	2.55	—	—	—	—	—	—	3,299,948	12.71	20,412	0.40	—	—	—	—
Oxisol		8,219,955	6.32	61,935	0.16	3,628,730	11.24	2347	0.07	3,943,318	15.18	583,624	11.42	—	—	—	—
	Cryerts	9957	0.01	—	—	—	—	49	0.00	9584	0.04	188	0.00	135	0.00	—	—
	Xererts	131,463	0.10	—	—	—	—	131,463	3.66	—	—	—	—	—	—	—	—
	Torrerts	922,574	0.71	922,574	2.38	—	—	—	—	—	—	—	—	—	—	—	—
	Usterts	2,024,847	1.56	—	—	2,001,034	6.20	—	—	22,598	0.09	1215	0.02	—	—	—	—
	Uderts	214,705	0.17	—	—	—	—	—	—	213,704	0.82	1001	0.02	—	—	—	—
		3,303,546	2.54	922,574	2.38	2,001,034	6.20	131,512	3.66	245,887	0.95	2404	0.05	135	0.00	—	—
Vertisol	Cryids	376,987	0.29	363,968	0.94	3999	0.01	7180	0.20	1841	0.01	—	—	—	—	—	—
	Salids	1,461,051	1.12	1,460,669	3.77	—	—	382	0.01	—	—	—	—	—	—	—	—
	Gypsid	1,236,853	0.95	1,151,913	2.97	72,038	0.22	5623	0.16	—	—	—	—	7279	0.04	—	—
	Argids	3,507,926	2.70	3,318,374	8.57	132,720	0.41	50,254	1.40	6578	0.03	—	—	—	—	—	—
	Calcids	5,701,082	4.38	5,181,402	13.38	320,316	0.99	149,251	4.16	6033	0.02	—	—	44,079	0.25	—	—
	Cambids	2,004,355	1.54	2,004,355	5.18	—	—	—	—	—	—	—	—	—	—	—	—
		14,288,254	10.99	13,480,680	34.81	529,073	1.64	212,691	5.92	14,452	0.06	—	—	51,359	0.30	—	—
Aridisol	Aquults	1,103,444	0.85	—	—	602,252	1.87	7880	0.22	456,297	1.76	37,015	0.72	—	—	—	—

	Humults	698,928	0.54	—	—	85,332	0.26	13,359	0.37	396,207	1.53	204,030	3.99	—	—	—	—
	Udults	4,750,569	3.65	—	—	—	—	—	—	3,648,146	14.05	1,102,423	21.58	—	—	—	—
	Ustults	4,248,453	3.27	—	—	4,248,453	13.16	—	—	—	—	—	—	—	—	—	—
	Xerults	2118	0.00	—	—	—	—	2118	0.06	—	—	—	—	—	—	—	—
Ultisols		10,803,512	8.31	—	—	4,936,037	15.29	23,357	0.65	4,500,651	17.33	1,343,467	26.30	—	—	—	—
	Albolls	33,337	0.03	24,193	0.06	—	—	8429	0.23	73	0.00	642	0.01	—	—	—	—
	Aquolls	266,532	0.20	295	0.00	43,375	0.13	27,640	0.77	151,625	0.58	7856	0.15	35,742	0.21	—	—
	Rendolls	263,618	0.20	—	—	37	0.00	—	—	202,457	0.78	61,125	1.20	—	—	—	—
	Gelolls	1,266,618	0.97	—	—	—	—	—	—	—	—	—	—	1,264,895	7.32	1722	0.02
	Cryolls	1,897,123	1.46	562,151	1.45	603,112	1.87	290,917	8.10	381,845	1.47	59,097	1.16	—	—	—	—
	Xerolls	683,395	0.53	—	—	—	—	683,395	19.03	—	—	—	—	—	—	—	—
	Ustolls	4,568,875	3.51	1,942,581	5.02	2,626,294	8.14	—	—	—	—	—	—	—	—	—	—
	Udolls	1,733,681	1.33	—	—	—	—	—	—	1,720,927	6.63	12,754	0.25	—	—	—	—
Mollisols		10,713,179	8.24	2,529,220	6.53	3,272,818	10.14	1,010,381	28.13	2,456,927	9.46	141,474	2.77	1,300,637	7.52	1722	0.02
	Aqualfs	2,155,814	1.66	—	—	715,552	2.22	137,010	3.81	831,437	3.20	11,775	0.23	460,039	2.66	—	—
	Cryalfs	2,044,083	1.57	—	—	190,887	0.59	28,160	0.78	1,197,389	4.61	45,289	0.89	582,359	3.37	—	—
	Ustalfs	4,884,166	3.76	—	—	4,884,166	15.13	—	—	—	—	—	—	—	—	—	—
	Xeralfs	571,052	0.44	—	—	—	—	571,052	15.90	—	—	—	—	—	—	—	—
	Udalfs	2,592,985	1.99	—	—	—	—	—	—	2,410,341	9.28	182,644	3.58	—	—	—	—
Alfisols		12,248,099	9.42	—	—	5,790,604	17.94	736,222	20.50	4,439,167	17.09	239,708	4.69	1,042,398	6.03	—	—
	Aquepts	4,810,104	3.70	—	—	949,473	2.94	21,043	0.59	1,434,050	5.52	242,473	4.75	2,163,064	12.51	—	—
	Gelepts	2,480,174	1.91	—	—	—	—	—	—	—	—	—	—	2,480,174	14.34	—	—
	Cryepts	606,816	0.47	—	—	107,807	0.33	40,193	1.12	297,406	1.15	161,410	3.16	—	—	—	—
	Ustepts	2,720,613	2.09	—	—	2,720,613	8.43	—	—	—	—	—	—	—	—	—	—
	Xerepts	681,942	0.52	—	—	—	—	681,942	18.99	—	—	—	—	—	—	—	—
	Udepts	3,878,102	2.98	—	—	—	—	—	—	2,936,321	11.31	941,781	18.43	—	—	—	—
Inceptisols		15,177,752	11.67	—	—	3,777,893	11.70	743,179	20.69	4,667,777	17.97	1,345,665	26.34	4,643,239	26.85	—	—
	Aquents	306,627	0.24	14,036	0.04	225,533	0.70	—	—	43,148	0.17	23,841	0.47	69	0.00	—	—
	Psamments	2,887,757	2.22	123,694	0.32	2,209,488	6.85	19,611	0.55	423,070	1.63	74,235	1.45	37,659	0.22	—	—
	Fluvents	2,839,752	2.18	681,788	1.76	929,085	2.88	121,705	3.39	552,159	2.13	86,228	1.69	468,785	2.71	—	—
	Orthents	19,879,772	15.28	9,914,413	25.60	4,349,652	13.48	525,826	14.64	1,684,373	6.49	704,252	13.78	2,701,256	15.62	—	—
Entisols		25,913,908	19.92	10,733,932	27.72	7,713,757	23.90	667,143	18.57	2,702,750	10.41	888,557	17.39	3,207,770	18.55	—	—
	Dunes and Shifting sand	9,594,833	7.38	9,517,724	24.58	46,326	0.14	1782	0.05	1689	0.01	6005	0.12	20,722	0.12	586	0.01
	Rocky land	2,844,756	2.19	1,061,365	2.74	66,611	0.21	11,323	0.32	16,166	0.06	3047	0.06	1,207,099	6.98	479,146	6.74
	Salt flats	144,095	0.11	142,804	0.37	—	—	1291	0.04	—	—	—	—	—	—	—	—
Misc.		12,583,685	9.67	10,721,893	27.69	112,937	0.35	14,395	0.40	17,855	0.07	9052	0.18	1,227,821	7.10	479,732	6.75
Total ice-free land		130,068,358	100	38,723,771	100	32,278,703	100	3,591,627	100	25,968,914	100	5,108,858	100	17,290,451	100	7,106,034	100

changes in nutrient properties, water-table levels, soil organic matter content, soil carbon contents and freeze–thaw cycles. The structure of Soil Taxonomy was created to classify soils that currently exist and to have the capability to be modified to account for new soil types. To do this, the definitions and nomenclature used must always deliver exactly the same meaning to everyone.

Gelisols

These soils commonly occur in areas where the mean annual soil temperature is less than 0 °C. They are frozen for long periods of the year and thaw out during the short warmer spells (Bockheim and Tarnocai, 2000). The freezing and thawing processes promote physical changes and mixing of the soil called cryoturbation or frost churning. If there is sufficient water and the warm period is long enough, vegetation can grow, and organic matter accumulates on the soil, leading to the development of organic-rich soils or peat. These arctic soils have unique features such as ice lenses, ice wedges, or other frozen features below the surface. The presence of permafrost and other soil features from freezing and thawing cause irregular or broken horizons and incorporation of organic materials into the soil profile. This process increases soil volume and reduces bulk density. In cold arid areas, the soil particles are frozen together by dry permafrost. Dry permafrost occurs when the soil temperature remains below 0 °C but there is insufficient moisture to be cemented by ice.

The three suborders in Gelisols are:

Histels (1.45% of land area): These soils have an organic surface (histic epipedon) and may be underlain by ice wedges, permafrost, or bedrock.

Turbels (4.15% of land area): Contain disrupted horizons caused by cryoturbation or frost churning that mixes the soil through freeze-thaw process. Fig. 2 includes a Turbel soil profile from Alaska.

Orthels (0.89% of land area): Other Gelisols that do not meet other subgroup criteria and are mostly found in North America and in some parts of Europe and Asia.

Histosols

Histosols are soils that dominantly consist of organic soil materials (Rabenhorst and Swanson, 2000). Histosols comprise a small portion of the world's land area (0.59%) but are widely distributed irrespective of climate. They develop where the rates of organic matter accumulation exceed decomposition and removal. These soils are commonly referred to as bogs, moors, marsh, swamps, or fens.

The suborders of Histosols are separated by the degree of decomposition of the organic matter, and climate. These are:

Folists (0.02% of land area): Are not saturated with water for more than a few days per year; they consist of horizons formed by the decomposition of leaves, twigs, and branches at the soil surface. The organic materials accumulate in aerobic conditions where the climate is cool and humid enough to reduce decomposition. They occupy a small area mainly in the northern latitudes.

Fibrists (0.15% of land area): Consist largely of plant remains so slightly decomposed that rubbing does not destroy plant fibers and their botanical origin can be easily determined. They occupy large areas in the northern latitudes.

Hemists (0.14% of land area): Consist of organic materials that are so decomposed that the botanic origin of as much as two-thirds of the plant materials cannot be readily discerned or the fibers can be largely destroyed by rubbing between the fingers. They occupy small areas in the northern latitudes and may also be found in the tropics.

Saprist (0.28% of land area): Consist of almost completely decomposed plant remains and their botanic origin cannot be determined. Most of these soils are found in warmer regions with mesic, thermic, and hyperthermic STRs. Fig. 2 includes a Saprist soil profile from Florida.

Wassists (not included in this assessment): Have a positive water potential at the soil surface for more than 21 h of each day in all years and are permanently submerged soils in fresh and saltwater environments (Soil Survey Staff, 2014).

Andisols

Andisols are soils formed on volcanic ash that have andic soil properties derived from aluminum–humus complexes (Kimble et al., 2000). The Andisols have a mineralogical composition comprising volcanic glass, short-range order minerals such as allophane and imogolite, and variable amounts of halloysite. This mineralogical association results in unique soil properties, such as high phosphate fixing capacity, low cation retention, and a high water holding capacity. They are dominantly found along the circum-Pacific belts but occur sporadically elsewhere (0.72% of the landmass).

Eight suborders are recognized in the Andisols. These are:

Aquands (land area too small for map scale): These soils experience water saturation for prolonged periods during the year.

Cryands (0.18% of land area): These soils have a cryic STR (mean annual soil temperature between 0 °C and 8 °C).

Gelands (0.12% of land area): These soils have a gelic STR (mean annual soil temperature less than 0 °C).

Torrands (0.01% of land area): These soils have an aridic SMR.

Xerands (0.01% of land area): These soils have a xeric SMR.

Vitrands (0.11% of land area): These soils are recent ash or cinder deposits where some weathering and soil formation has taken place.

Ustands (0.05% of land area): These soils have an ustic SMR. They are widespread on the lee-ward side of tropical volcanoes and are commonly used for agriculture.

Udands (0.26% of land area): These soils have an udic SMR and are widely used for agriculture. Fig. 2 includes an Udand soil profile from Japan.

Spodosols

These soils have a black, reddish brown to strong brown spodic subsurface horizon (spodic horizon) that is high in illuvial organic matter, aluminum (Mokma and Evans, 2000) and may or may not contain iron. It is often overlain by a gray to light gray eluvial horizon.

Five suborders are recognized in Spodosols. These are:

Aquods (0.10% of land area): These soils have an aquic SMR and are generally found in depressions. Fig. 2 includes an Aquod soil profile from Florida.

Gelods (3.03% of land area): Spodosols that have a gelic STR. These are the most extensive Spodosols (3.9 million km²) and are dominant in cold temperate areas of the northern latitudes.

Humods (0.11% of land area): Have a spodic horizon enriched with elevated amounts of organic matter. These soils are generally found in cool climates including high elevations in tropical mountains.

Cryods (1.27% of land area): Have a cryic STR.

Orthods (0.61% of land area): Other Spodosols that do not meet the above criteria.

Oxisols

Oxisols are highly weathered soils and often have a reddish, yellowish, or grayish color. They are commonly located on the gentle undulating surfaces of geologically old surfaces in tropical and subtropical regions where they occupy about 8.2 million km² (Beinroth et al., 2000). They contain few weatherable minerals and tend to have high levels of iron and aluminum. These soils are very infertile and require fertilizers for cropping.

Oxisols are present in every SMR. Natural vegetation ranges from tropical rainforests to savannas.

The suborders of the Oxisols are based on the SMR:

Aquoxs (0.71% of land area): Have an aquic SMR and are the very wet Oxisols.

Torrox (0.05% of land area): Have an aridic SMR. Such soils occur in parts of South America and Africa.

Ustoxs (2.38% of land area): Have an ustic SMR. These are the second most extensive suborder of Oxisols.

Peroxs (0.64% of land area): Have a perudic SMR. These represent the most typical Oxisols of the humid zones in parts of South America, Central Africa, and Asia.

Udoxs (2.55% of land area): Have an udic SMR. These are the most extensive of the Oxisols with large areas in South America, Central Africa, Asia, and some of the Pacific Islands. Fig. 2 includes a Udox soil profile from Brazil.

Fig. 4 shows the distribution of Oxisol suborders in Africa. It is evident that the main types of Oxisols at this level of Taxonomy include the Udic, Ustic, Aquic and Perudic moisture regimes. Other factors that control this distribution include relief and geology.

Vertisols

Vertisols are clayey soils, that shrink and swell and have deep, wide cracks that open and close depending on moisture conditions (Coulombe et al., 2000). Most Vertisols are well suited for mechanized farming if there is sufficient rainfall or irrigation water and if appropriate management practices are followed.

Six suborders are recognized in Vertisols based on the SMR or STR.

Aquerts (land area too small for map scale): Wet Vertisols with an aquic SMR.

Cryerts (0.01% of land area): Cold Vertisols with a cryic or gelic STR.

Torrerts (0.71% of land area): Vertisols with an aridic SMR. Fig. 2 includes a Torrert soil profile from Botswana.

Xererts (0.10% of land area): Vertisols with a xeric SMR. These are the typical reddish brown Vertisols that have a Mediterranean climate.

Usterts (1.56% of land area): Vertisols with an ustic SMR. Extensive areas are found in South Asia and Africa.

Uderts (0.17% of land area): Vertisols with an udic SMR. Large areas of Uderts are found in North and South America.

Aridisols

Aridisols are soils that do not have water available to plants for long periods (Southard, 2000). During most of the time when the soil is warm enough for plants to grow in Aridisols, soil water is held at potentials less than permanent wilting point. There is no period of 90 consecutive days when moisture is continuously available for plant growth.

The seven suborders are:

Salids (1.12% of land area): Characterized by the accumulation of salts more soluble than gypsum. These are the typical soils of the playas or desert depressions or closed basins.

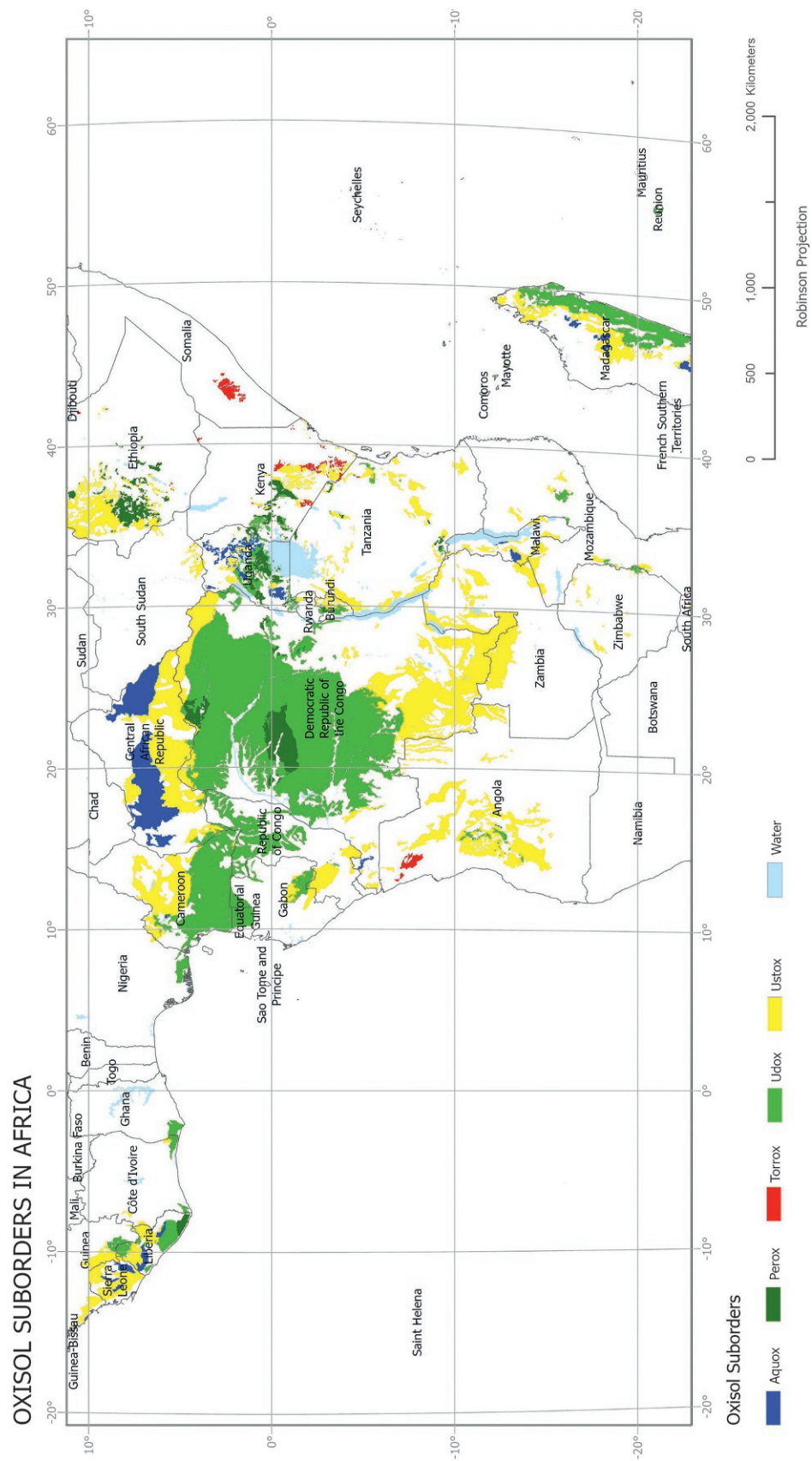


Fig. 4 Oxisol suborders in Africa map.

Durids (land area too small for map scale): Characterized by a subsurface cemented layer.

Gypsis (0.95% of land area): Characterized by an accumulation of gypsum. Such soils are extensive in the Middle East.

Argids (2.70% of land area): Characterized by an accumulation of clay by translocation.

Calcids (4.38% of land area): Characterized by an accumulation of carbonates, the soils have a calcic, petrocalcic, or a hypercalcic horizon; these soils are the most extensive of the Aridisols. Fig. 2 includes a Calcid soil profile from New Mexico.

Cambids (1.54% of land area): These soils have a cambic horizon that is characterized by physical or chemical transformations of material.

Cryids (0.29% of land area): Aridisols that have a cryic STR.

Ultisols

Ultisols are similar to Alfisols having a subhorizon of clay accumulation (West and Beinroth, 2000) but have a lower base saturation of cations. Most Ultisols are acid although some may have a high pH in the surface horizons due to agricultural liming.

The suborders of the Ultisols are:

Aquults (0.85% of land area): These are the wet Ultisols with an aquic SMR.

Humults (0.54% of land area): These soils have accumulations of organic matter and usually occur in areas with an udic SMR and cooler temperatures.

Udults (3.65% of land area): These soils have an udic SMR. Fig. 2 includes an Udult soil profile from Vietnam.

Ustults (3.27% of land area): These soils have an ustic SMR.

Xerults (<0.01% of land area): These soils have a xeric SMR.

Mollisols

Presence of a thick dark humus-rich surface horizon (mollic epipedon) is the key feature of Mollisols (Bell and McDaniel, 2000). Most Mollisols are cultivated and sustain high yields of grain crops in the breadbasket regions of the world such as in the United States, Russia, Ukraine, and Argentina.

Eight suborders are recognized in the Mollisols.

Albolls (0.03% of land area): These soils have a bleached subsurface horizon (albic horizon) that is usually wet. Fig. 2 includes an Alboll soil profile from Kansas.

Aquolls (0.20% of land area): These are the wet Mollisols and may have a histic epipedon.

Rendolls (0.20% of land area): These are formed in highly calcareous parent materials and have a udic SMR or cryic STR.

Gelolls (0.97% of land area): Have a gelic STR.

Cryolls (1.46% of land area): Have a cryic STR.

Xerolls (0.53% of land area): Have a xeric or border an aridic SMR.

Ustolls (3.51% of land area): Have a ustic or border an aridic SMR.

Udolls (1.33% of land area): Have an udic SMR.

Alfisols

Alfisols are commonly forested, with moderate to high base saturation (Hallmark and Franzmeier, 2000). Typically, they have a light-colored layer (albic horizon) over a layer of silicate clay accumulation (argillic horizon). They are generally found in climates favorable to crop production; in warm moist climates, they are used to grow many crops.

The five suborders of Alfisols are:

Aqualfs (1.66% of land area): These are the wet Alfisols and have an aquic SMR. Fig. 2 includes an Aqualf soil profile from China.

Cryalfs (1.57% of land area): Have a cryic STR.

Ustalfs (3.76% of land area): Have a ustic SMR.

Xeralfs (0.44% of land area): Have a xeric SMR.

Udalfs (1.99% of land area): Have an udic SMR.

Inceptisols

Inceptisols are soils that are either in the early stages of soil formation and have little soil development (Hudnall et al., 2000) or are soils with well-developed diagnostic horizons that merely fail the criteria of the other soil orders. Most Inceptisols have a cambic horizon and an ochric epipedon.

Six suborders are recognized in the order of Inceptisols.

Aquepts (3.70% of land area): These are the wet Inceptisols that have an aquic SMR.

Gelepts (1.91% of land area): Have a gelic STR.

Cryepts (0.47% of land area): Have a cryic STR.

Ustepts (2.09% of land area): These soils have an ustic SMR.

Xerepts (0.52% of land area): These soils have a xeric SMR.

Udepts (2.98% of land area): These soils have an udic SMR. Fig. 2 includes an Udept soil profile from Idaho.

Entisols

The Entisols have very little or no evidence of soil formation (Nordt et al., 2000). They are most extensive on young alluvial plains or on steep slopes where there is a lack of water movement through the soil. The rate of soil formation may be reduced for several reasons, but generally time has not elapsed since deposition of the material for soil forming processes to act.

Six suborders are recognized in the Entisols:

Aquepts (0.24% of land area): These are the wet Entisols.

Wassents (not included in this assessment): Have a positive water potential at the soil surface for more than 21 h of each day in all years.

Psamments (2.22% of land area): These are the sandy Entisols. Fig. 2 includes a Psamment soil profile from New Hampshire, USA.

Fluvents (2.18% of land area): These are recent alluvial soils showing an irregular decrease in organic matter.

Orthents (15.28% of land area): These are often found on recent erosional surfaces but include all other soils that do not meet the above criteria or criteria for the other soil orders.

Conclusion

Although developed primarily for the soils in the U.S., Soil Taxonomy is applicable world-wide. This system was designed with an open structure to enable modifications to the classification system and can account for new soil properties and soil types, e.g., the non-volcanic andic soils or soil properties. For over six decades, soil scientists from around the world have contributed to the development and application of Soil Taxonomy for soil survey, mapping and interpretation. Soil Taxonomy enables soil-landscape relationships to be identified quite easily and caters for all soils that exist on earth. This allows the information obtained to be used constructively for designing proper decision support systems that result in good practices being advocated to maintain and enhance soil quality. This is particularly important since food security is a major concern for many countries around the world. Therefore, Soil Taxonomy provides an intelligent, open-structured and logical means of maintaining crop production at the desired levels and is flexible enough to cater for inclusions or omissions from the system as deemed necessary. The process of amending Soil Taxonomy requires the existence of the pedon to be established first. After that, the level of placement of the pedon in Soil Taxonomy will be decided based on the main characteristics. The tremendous effort of soil scientists world-wide over the last six decades or so has brought about a comprehensive Soil Taxonomy system that covers all known soil types on Earth. The knowledge pool that has been acquired is very large and has the potential to be applied in new areas as required by respective national space agencies.

A wide diversity of soils exists and each has its own potential and limitations. The performance-related attributes result from the properties and the prevailing environmental conditions. Understanding the soil and its relation to the rest of the environment is important for its sustainable use. Detailed soil surveys provide such information, but most countries of the world lack this information. Assessment and monitoring of the health of our soil resources are essential to their judicious use and management, especially with respect to global environmental changes.

Soil Taxonomy serves as an indispensable tool for decision makers on managing the renewable soil resource in a sustainable manner. This classification system remains a powerful tool for soil quality assessments and mitigation measures.

Acknowledgments

Hari Eswaran coauthored the first edition of this chapter and created the conversion matrix used to determine the global distribution of Soil Taxonomy suborders. Thomas Reinsch provided modifications to the conversion matrix used for this updated chapter. John Kelley provided soil profile images used in Fig. 2. Dan Wing and Charles Kome provided valuable review and suggestions to improve the chapter.

References

- Ahrens RJ and Arnold RW (2000) Soil taxonomy. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E-117–E-135. Boca Raton: CRC Press.
- Arnold RW (1990) Soil Taxonomy, a tool of soil survey. In: Rozanov BG (ed.) *Soil Classification*, pp. 94–111. Moscow: Publ. Centre for International Projects, USSR State Committee for Environmental Protection.
- Arnold RW, Ahrens RJ, and Engel RJ (1997) Trends in Soil Taxonomy—a shared heritage. *Communications of the Austrian Soil Science Society* 55: 167–170.
- Bartelli LJ (1978) Technical classification system for soil survey interpretation. In: Brady NC (ed.) *Advances in Agronomy*, vol. 30, pp. 247–289. New York: Academic Press.
- Beinroth FH, Eswaran H, Uehara G, and Reich PF (2000) Oxisols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E-373–E-392. Boca Raton: CRC Press.
- Bell JC and McDaniel PA (2000) Mollisols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E-286–E-307. Boca Raton: CRC Press.
- Benchmark Soils Project (1982) *Assessment of Agrotechnology Transfer in a Network of Tropical Soil Families. Progress Report 3 (July 1979–September 1982)*. Department of Agronomy and Soil Science, College of Tropical Agriculture and Human Resources, University of Hawaii, and Department of Agronomy and Soils, College of Agricultural Sciences, University of Puerto Rico. 104 pp.
- Bockheim JG and Tamocai C (2000) Gelisols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E-256–E-269. Boca Raton: CRC Press.
- Buol SW, Southard RJ, Graham RC, and McDaniel PA (2011) *Soil Genesis and Classification*, 6th edn, John Wiley & Sons, Inc. 543 pp.

- Coulombe CE, Wilding LP, and Dixon JB (2000) Vertisols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E-269–E-286. Boca Raton: CRC Press.
- Deckers JA, Nachtergaele FO, and Spaargaren OC (eds.) (1998) *World Reference Base for Soil Resources: Introduction*. Leuven, Belgium: International Society of Soil Science (ISSS), International Soil Reference and Information Centre (ISRIC) and Food and Agriculture Organization of the United Nations (FAO).
- FAO/IIASA/ISRIC/ISS-CAS/JRC (2012) *Harmonized World Soil Database (Version 1.2)*. Rome, Italy; Laxenburg, Austria: FAO; IIASA. URL: <https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/harmonized-world-soil-database-v12/en/>.
- Hallmark CT and Franzmeier DP (2000) Alfisols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E-338–E-358. Boca Raton: CRC Press.
- Hijmans RJ, Cameron SE, Parra JL, Jones PG, and Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25: 1965–1978. URL <http://www.worldclim.org/>.
- Hudnall WH, West LM, Benham EC, and Wilding LP (2000) Inceptisols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E-242–E-255. Boca Raton: CRC Press.
- IUSS Working Group WRB (2015) *World Reference Base for Soil Resources 2014, Update 2015 International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*. *World Soil Resources Reports No. 106*. Rome: FAO.
- Jenny H (1941) *Factors of Soil Formation—A System of Quantitative Pedology*. McGraw-Hill.
- Kimble JM, Ping CL, Sumner ME, and Wilding LP (2000) Andisols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E-209–E-224. Boca Raton: CRC Press.
- Libohova Z, Martín-López JM, da Silva M, Lagoueyte C, Cruz J, Drohan P, Maximova S, Guiltinan M, Ferruzzi MG, Guarín D, Reich P, Kome C, Zapata YP, Gallego-Sánchez G, Quintero C, Botero C, Winters NP, and Robotham M (2020) *Soil and Cacao Genomics Survey of Sierra Nevada de Santa Marta Region, Colombia*. United States Department of Agriculture, Natural Resources Conservation Service; International Center for Tropical Agriculture (CIAT); and Pennsylvania State University.
- Mokma DL and Evans CV (2000) Spodosols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E-307–E-321. Boca Raton: CRC Press.
- Nordt LC, Collins ME, Fanning DS, and Monger HC (2000) Entisols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E-224–E-242. Boca Raton: CRC Press.
- Orvedal AC and Edwards MJ (1941) General principles of technical grouping of soils. *Soil Science Society of America Proceedings* 6: 386–391.
- Rabenhorst MC and Swanson D (2000) Histosols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E-183–E-209. Boca Raton: CRC Press.
- Schad P (2017) Technosols in the World Reference Base for Soil Resources – history and definitions. *Soil Science and Plant Nutrition* 64(2): 138–144. <https://doi.org/10.1080/00380768.2018.1432973>.
- Soil Science Division Staff (2017) Soil survey manual. In: Ditzler C, Scheffe K, and Monger HC (eds.) *USDA Handbook 18*. Washington, D.C.: Government Printing Office.
- Soil Survey Staff (1975) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*. U.S. Dept. Agric. Handbook 436. Govt. Printing Office. 754 pp.
- Soil Survey Staff (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn. U.S. Dept. Agric. Handbook 436. Govt. Printing Office. 869 pp.
- Soil Survey Staff (2014) *Keys to Soil Taxonomy*, 12th edn. Washington, DC: USDA-Natural Resources Conservation Service.
- Southard RJ (2000) Aridisols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E-321–E-338. Boca Raton: CRC Press.
- USDA-NRCS (2011) JAVA Newhall Simulation Model—Update to a traditional soil climate simulation model. In: *Presented at the Soil Sci. Soc. Am. Int. Annual Meeting*. San Antonio, TX. 18 Oct 2011. URL: http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/class/?cid=nrcs142p2_053559.
- USDA-SCS (1991) *Soil Management Support Services Final Report, FY 1980 to FY 1990*.
- West LT and Beinroth FH (2000) Ultisols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. E-358–E-372. Boca Raton: CRC Press.
- Wilding LP, Smeck NE, and Hall GF (eds.) (1983a) *Pedogenesis and Soil Taxonomy. I. Concepts and Interactions*. Amsterdam: Elsevier.
- Wilding LP, Smeck NE, and Hall GF (eds.) (1983b) *Pedogenesis and Soil Taxonomy. II. The Soil Orders*. Amsterdam: Elsevier.

Soils as archives of city history

Hermine Huot^a and Jean Louis Morel^b, ^aLaboratoire Interdisciplinaire des Environnements Continentaux, Université de Lorraine – CNRS, Nancy, France; ^bLaboratoire Sols et Environnement, Université de Lorraine – INRAE, Nancy, France

© 2023 Elsevier Ltd. All rights reserved.

This is an update of J.L. Morel, C. Schwartz, L. Florentin, C. de Kimpe, URBAN SOILS, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 202–208, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00305-2>.

Introduction	233
Footprints of past human activities in SUITMAs	233
Soils as archives of human–nature relations	233
Markers of human activities in SUITMAs	234
Deciphering memory of past human activities in SUITMAs	235
Urban soil stratification as record of the city history	236
Soils of industrial brownfields as archives of past industrial activities	236
Human-induced modifications on factors of evolution	237
Parent materials	239
Topography and hydrology	239
Climate	239
Organisms	240
Time	240
Evolution of SUITMAs	241
Main soil-forming processes observed in SUITMAs	241
Accumulation and transformation of organic matter	241
Mineral transformations	241
Transfers of soluble compounds and particles	242
Development of soil structure	242
Differentiation of soil horizons	242
Specific pedogenetic trends in SUITMAs	242
Prediction of the evolution of SUITMAs	243
Direction of the pedogenesis	243
Spatial variation and soil diversity	244
Temporal discontinuities	244
Concluding remarks	244
References	244

Abstract

Humans have modified soils through time. Soil composition and organization can record information about the history of land use. The investigation of human-influenced soils as archives requires multidisciplinary and multiscale approaches. Human activities have the potential to modify all soil-forming factors and thus influence the evolution of these archives. If the main pedogenic processes observed in human-affected soils are similar to those occurring in natural soils, human-modified soil-forming conditions may reorientate the trajectories of these soils. Improving the knowledge about soil quality and evolution in inhabited areas is necessary to manage the soils of these areas in a sustainable way.

Key points

- Human-modified soils can record information about history of land use.
- Soil composition and organization are markers of human activities.
- Reconstruction of the evolution of these soils requires multidisciplinary and multiscale approaches.
- Human activities modify soil-forming conditions and may create some specific pedogenic trends.
- Knowledge about long-term soil evolution would be of help for managing inhabited areas sustainably.

Introduction

For several millennia, humans have shaped and are shaping the soils in the areas where they live. Through history, human activities have successively modified the soil characteristics and the environmental conditions of soil formation, both intentionally or unintentionally. By modifying the soil properties, human activities alter in the short or long term the capacity of soils to fulfill essential functions and to provide services necessary for the development of human societies (Fig. 1). Soil non-ecological functions include the protection of cultural heritage. Human-modified soils can be considered as recorders of the history of human activity (Urbańska and Charzyński, 2021). For a long time, soils of human-affected areas have received little attention in soil science because of the greater focus on agricultural and forest soils. The investigation of urban, industrial or mining soils started in the 1970s, mainly in relation to the growing public concern about environmental issues and human health. In 1998, a working group dedicated to “Soils of Urban, Industrial, Traffic, Mining and Military Areas” (SUITMAs) was created within the framework of the International Union of Soil Science (IUSS) (De Kimpe and Morel, 2000; Burghardt et al., 2015). The SUITMAs encompass a large diversity of soils with varying degrees of human influence depending on the history of their land use. Nowadays, inhabited areas face many issues mainly related to the increased population living in cities, the needs of food, energy and goods, and to the adaptation to climate change. In this context, the potential of SUITMAs to provide services such as support of green spaces and food production, regulation of the water cycle, or a reservoir of biodiversity should be considered better in land planning (Morel et al., 2015). Sustainable management of inhabited areas requires understanding of how human activities change soil properties and their ability to fulfill essential functions, and also to predict the evolution of soils in such environments. This chapter focuses on the study of SUITMAs as archives with the aim of improving the knowledge about human history, but also about the conditions and processes of soil evolution in inhabited areas.

Footprints of past human activities in SUITMAs

The succession of land use has left multiple traces in soils, which can be altered or erased over time. The investigation and reconstruction of the trajectories of these complex soils require multidisciplinary and multiscale approaches to decipher the successive mixed human and environmental contributions through time.

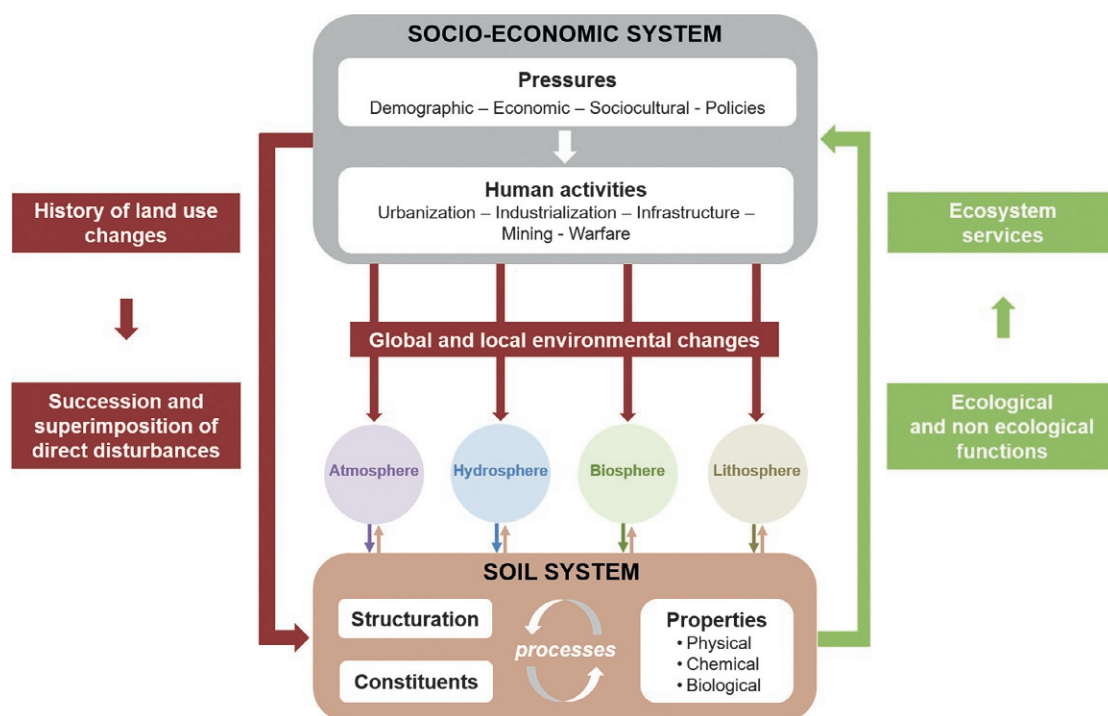


Fig. 1 Human–soil relationships in human-affected areas. Through history, human activities have successively disturbed the soils and the environmental conditions of soil formation. By modifying the soil characteristics (structure, constituents, properties), human activities alter their capacity to fulfill essential ecological functions and provide services necessary for the development of human societies in the short or long term.

Soils as archives of human–nature relations

Soils have the ability to record their past by storing information about factors and processes of formation. Several conceptual models have been established to explain the formation of natural soils. The classic factorial approach defines the state of the soil system with five factors: parent material, relief, climate, living organisms and time. Parent material and topography determine the initial state of the system, which evolves over time through the influence of climate and biological activity (Jenny, 1941). The process-based approach considers that the parent materials weather under the action of a combination of processes including additions, losses, transfers and transformations. These lead to the development of a soil profile and the differentiation of soil horizons with specific characteristics. The nature of the soil is governed by the relative importance and balance between the wide range of processes, which leave their imprint on soil characteristics (Simonson, 1959). The concept proposed by Rode in 1947 suggests that the repetition of incompletely closed cyclic processes involving mainly liquids, gases and biota (functioning processes) generates residual solid products that accumulate and differentiate in the long term (pedogenetic processes). The soil memory relies on the direct link between factors, functioning processes, pedogenetic processes and residual solid features at different scales (memory carriers), such as amorphous compounds, aggregates, horizons and profile (Targulian and Bronnikova, 2019).

Human activities have transformed soils and their conditions of formation and can have a long-term impact on soil composition and properties. Some soil features can be specific to human land use and related practices, like regular ploughing, fertilization, pollution or dumping of waste and construction debris. Thus, soils can contribute to store buried elements of human heritage, and can be considered as valuable recorders of human-induced changes of the environment. Human-modified soils have been proposed as markers of the start of the Anthropocene, which could be around 2000 BP when the terrestrial surface of the planet was noticeably modified by organized civilizations through the exploitation of natural resources, such as land cultivation, deforestation or mining (Certini and Scalenghe, 2011).

In that context, SUITMAs can be considered as archives of past urban, industrial or mining activities, and are of interest for educational and touristic purposes (Urbańska and Charzyński, 2021). Analyses of soil features can document the history of land use in correlation with other sources of historical information. In addition, such investigations of the past to reconstruct the trajectory of soils will enhance the understanding of the impact of land use practices on soils and will provide recommendations for future land management and planning (Riddle and Levin, 2017). In that way, SUITMAs can be helpful educational supports to raise awareness about the impact of human activities on soils and more generally on the environment, and to promote more sustainable management and protection of the limited soil resource.

Markers of human activities in SUITMAs

Depending on land use and the intensity of induced changes, human activities may alter the composition and properties of existing soils or greatly modify the organization of the soil profile, which may lead to the formation of new soils.

For example, SUITMAs may display disturbed profiles, with partial or complete disorganization of the horizons initially formed by natural soil development. Soil profiles can be modified by the truncation of upper layers, by mixing of the horizons during earthworks, by surface sealing or by burying of infrastructures or impermeable geomembranes in the soil (Fig. 2). Native soils can be buried under human-transported materials, either earthy materials brought from surrounding areas as backfill or human-made materials generated by urban, industrial or mining activities. Soils can also be intentionally amended or even constructed to fulfill specific functions, such as to support vegetation. Such soil constructions consist of mixed organic and mineral by-products from urban or industrial activities to obtain the required soil properties, which are deposited in layers to mimic natural soil horizons

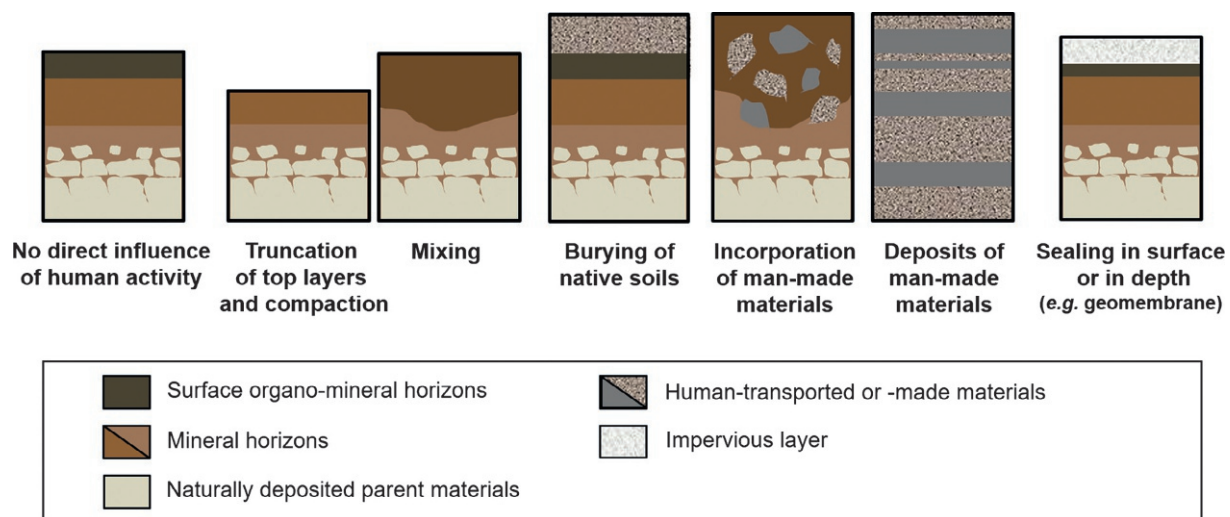


Fig. 2 Possible human-induced changes of soil organization along the profiles of soils of urban, industrial, traffic, mining and military areas (SUITMAs).

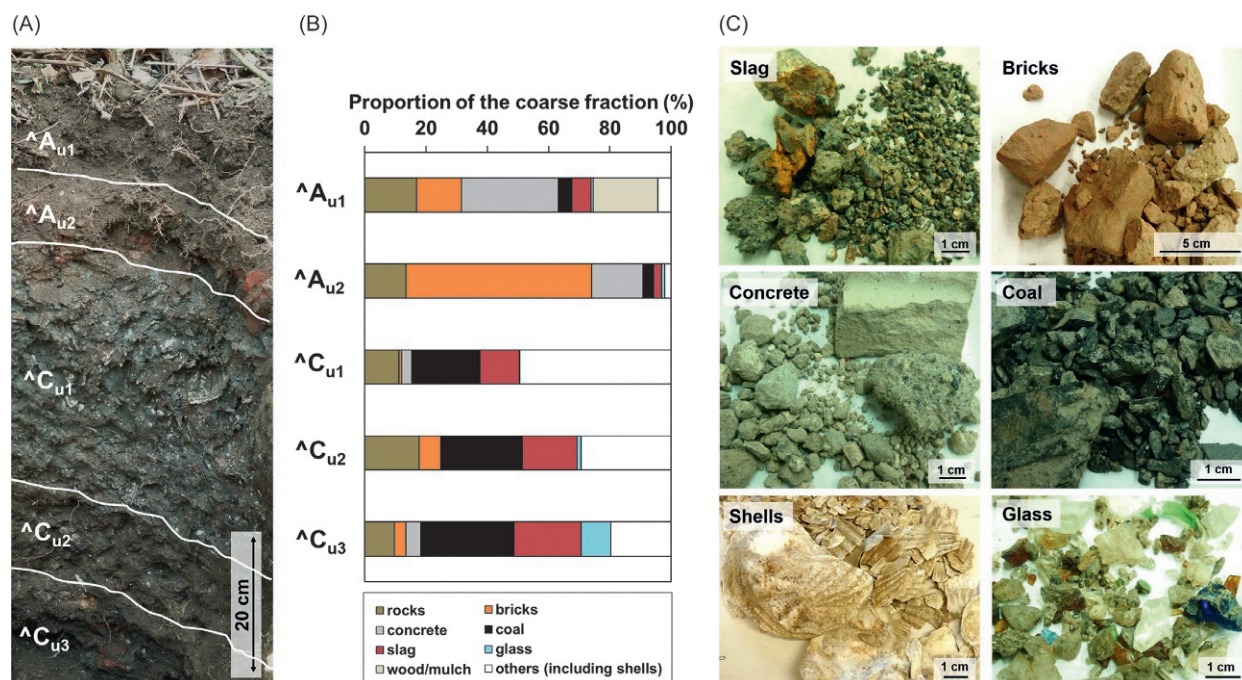


Fig. 3 Example of urban soil with a large content of artifacts: (A) profile of a soil formed in loamy fill with construction debris in New York City (Laguardia series in the city-wide soil map completed by USDA-NRCS Soil Survey group; <http://websoilsurvey.sc.egov.usda.gov/App/HomePage.htm>), (B) distribution of the types of artifacts in the coarse fragments (> 2 mm) and (C) pictures of the main types of artifacts. This profile was described in Huot H, Joyner J, Córdoba A, et al. (2017) Characterizing urban soils in New York City: profile properties and bacterial communities. *Journal of Soils and Sediments* 17: 393-407.

(Séré et al., 2010). As a result, SUITMA profiles may display a complex polygenetic organization with the superimposition of layers reflecting the succession of human transformations. The burying by new inputs may inhibit evolution of the buried soil materials, leading to a 'book-like' type of memory, typical of sedimentary systems. When the human disturbances stop, the 'new' soil gradually evolves under the natural environmental factors with a 'palimpsest' type of memory (multiple records on the same 'page'), typical of natural soils. Thus, profiles of SUITMAs can record land use changes over time as well as the conditions of soil formation (Prokofeva and Poputnikov, 2010).

One of the main footprints of human activities in soils is the presence of artifacts, i.e. materials created or modified by industrial or artisanal manufacturing processes or brought to the surface by human activity from a depth where they were not influenced by soil-forming processes. Artifacts include construction debris and waste of human settlements, industrial waste, mine spoil, dredged sediments, excavated bedrocks and oil products (IUSS Working Group WRB, 2014). The presence of construction debris is common in urban soils because of long accumulation of debris from regular demolition and reconstruction of houses as well as destruction by wars or fire (Fig. 3). Construction debris was often disposed of at or near the site of demolition, so ancient cities are generally built on their own waste (De Kimpe and Morel, 2000).

At the end of the Second World War, many European cities lay in ruins. In Berlin, 75 million m³ of debris was deposited throughout the city. Part of the debris that was not reusable as building materials, was used as foundations or disposed of in heaps (up to 100 m in elevation) in public parks and the outskirts of the city, and were revegetated to create recreational areas. Soils developing on debris contain large amounts of coarse fragments (> 2 mm) and artifacts such as bricks, mortar, slag and ashes, metal and electrical scrap, which bear witness to the city's past. The nature and size of these anthropogenic components influence the soil physical, chemical and biological properties, and also the quality of water percolation. For instance, sulfates have been leached from WWII debris into ground water in Berlin (Wessolek et al., 2011).

In addition to the incorporation of human-made materials, atmospheric depositions from industrial activities and traffic or regular additions of amendments and gardening practices contribute to alter the soil geochemical composition. Consequently, phosphorus, metals like copper (Cu), zinc (Zn), cadmium (Cd) or lead (Pb), and also organic compounds resulting from combustion (black carbon) or from petroleum products are often enriched in SUITMAs. They are likely to contain large concentrations of potentially toxic compounds that need to be identified and monitored for their fate (De Kimpe and Morel, 2000). These soils can also be highly compacted due to earthworks and the use of heavy machinery.

Deciphering memory of past human activities in SUITMAs

The preservation of soil memory depends on the weathering conditions and the evolution of natural and anthropogenic soil-forming factors. Consequently, it is necessary to take a multiscale, multidisciplinary and multi-proxy approach to decipher

anthropogenic signatures and traces of pedogenic processes, and to establish hypotheses of reconstruction. In particular, documentation about the history of land use at sites studied and knowledge on artisanal and industrial processes used to produce materials present in the soils, are a useful complement to the soil analyses. This is illustrated through two examples of studies of soils as archives of past urban and industrial activities.

Urban soil stratification as record of the city history

Urban areas encompass a large diversity of soils, from soils which have preserved the characteristics of natural regional soils, for example in urban forests, to soils strongly transformed by successive human land uses (e.g. land clearance, cultivation, building and infrastructure construction, artisanal and industrial activities) during city development. In ancient towns, accumulation of anthropogenic wastes, partially transformed by pedogenic processes, has led to the formation of 'cultural layers' comprising the superimposition of weakly developed buried soils overlying the initial natural soils (Alexandrovskiy et al., 2012). Cultural layers are a source of historical information about city development and the interactions between humans and their environment (Alexandrovskaia and Alexandrovskiy, 2000).

The study of such urban soils benefits from an interdisciplinary approach linking archeology and geosciences. First, the organization of the soil and distribution of artifacts are observed at the meter scale in excavations. Stratigraphic units are delimited in the soil profile depending on the variation in characteristics of the materials (e.g. color, texture, presence of artifacts). The units are then sampled and characterized to obtain further information about the nature of the different inputs and processes of accumulation, erosion and transformation. Methods used to analyze urban strata include micromorphology, physicochemical analyses (e.g. particle size analysis, organic matter, nutrients, metals and carbonates), geophysics, quantitative and qualitative study of the artifacts (e.g. charcoal, pottery), and archeozoological and botanical investigations (phytoliths, pollen, seeds). Artifacts or organic compounds in the soil can be dated, for instance using radiocarbon analysis (Borderie et al., 2019).

In ancient European cities, urban strata often include thick poorly stratified layers of dark color and contain important artifacts. These urban 'Dark Earths' occur in a wide range of environmental and historical contexts, and recur between the strata of antiquity (before 4th century) and the Middle Ages (11th century) in northwestern Europe. Characteristics of such layers depend on the pedoclimatic context and the different inputs linked with the land uses. But they generally include significant concentrations of organic matter with a high C/N ratio due to the large presence of charcoal and organic fragments, thereby contributing to carbon storage in urban soils. The diversity of artifacts indicates the nature and origin of the inputs, such as building materials (mortar, bricks, tiles, plaster), domestic activities (manure, bones) or craft activities (ash, charcoal, slag, glass). Geochemically, Dark Earths often show an enrichment in phosphates related to the addition of manure, bones or ashes and in metals as a result of past or more recent polluting activities. They are likely to have neutral to alkaline pH with the presence of carbonated artifacts from construction materials. If these deposits show an apparent homogeneity at the macroscale, the study of micromorphological features gives valuable information about the processes of accumulation and transformation leading to the formation of Dark Earths (see example in Fig. 4). In particular, bioturbation contributes to homogenization of the deposits, which can also be due to mechanical soil cultivation. Percolation and waterlogging also affect mineral weathering and organic matter preservation in these soils. The main interpretations of the origin of Dark Earths included gardening practices, waste deposition or abandonment of some urban areas with subsequent development of vegetation and homogenization by strong biological activity. More recent investigations using a geoarcheological approach revealed that Dark Earths result from a complex combination of processes depending on environmental conditions and practices associated with the multiple uses of these urban spaces. In this way, these soils store information about the medieval conditions of life and their study contributes to a new perception of the urban spaces at this period (Borderie et al., 2019).

Soils of industrial brownfields as archives of past industrial activities

With the decline of heavy industry at the end of 20th century, some areas in industrialized countries include numerous 'brownfield' sites that were abandoned following the closure or even demolition of plants. Spontaneous vegetation growth has generally occurred on these sites, sometimes leading to the establishment of forested ecosystems. The rehabilitation of these sites depends on the site constraints (e.g. pollution, infrastructures, land pressure), regional redevelopment programs and stakeholders involved. Brownfields may have a heritage value which can be recognized and preserved. Analyses of their soils and of the artifacts they contain can complement the work of historians and archeologists about the history of these industrial sites (Urbańska and Charzyński, 2021). Documentation of the site history (e.g. aerial views, former maps, administrative documents) and information about the industrial processes are also useful for such sites for risk assessment and reconstruction of their trajectory of evolution.

Investigations of such sites need to combine observations at different scales (site, profile, microstructure) with soil and artifact analyses, including geochemical and mineralogical ones. The nature and distribution of major and trace elements, organic compounds and minerals can be used as tracers of industrial by-products and their subsequent changes in the soil. Enrichment or depletion in some elements compared to the regional background concentrations can reveal anthropogenic inputs and or activities. Molecular markers, based on the distribution of some classes of aliphatic, aromatic or polar compounds, can be used to distinguish natural or anthropogenic origin of soil organic matter and different sources of organic pollutants (e.g. combustion, petroleum products). The study of mineral phases together with knowledge about the industrial processes and conditions of formation and weathering of these minerals (e.g. temperature, pH, redox potential) helps to establish hypotheses about the origin and transformations of the minerals.

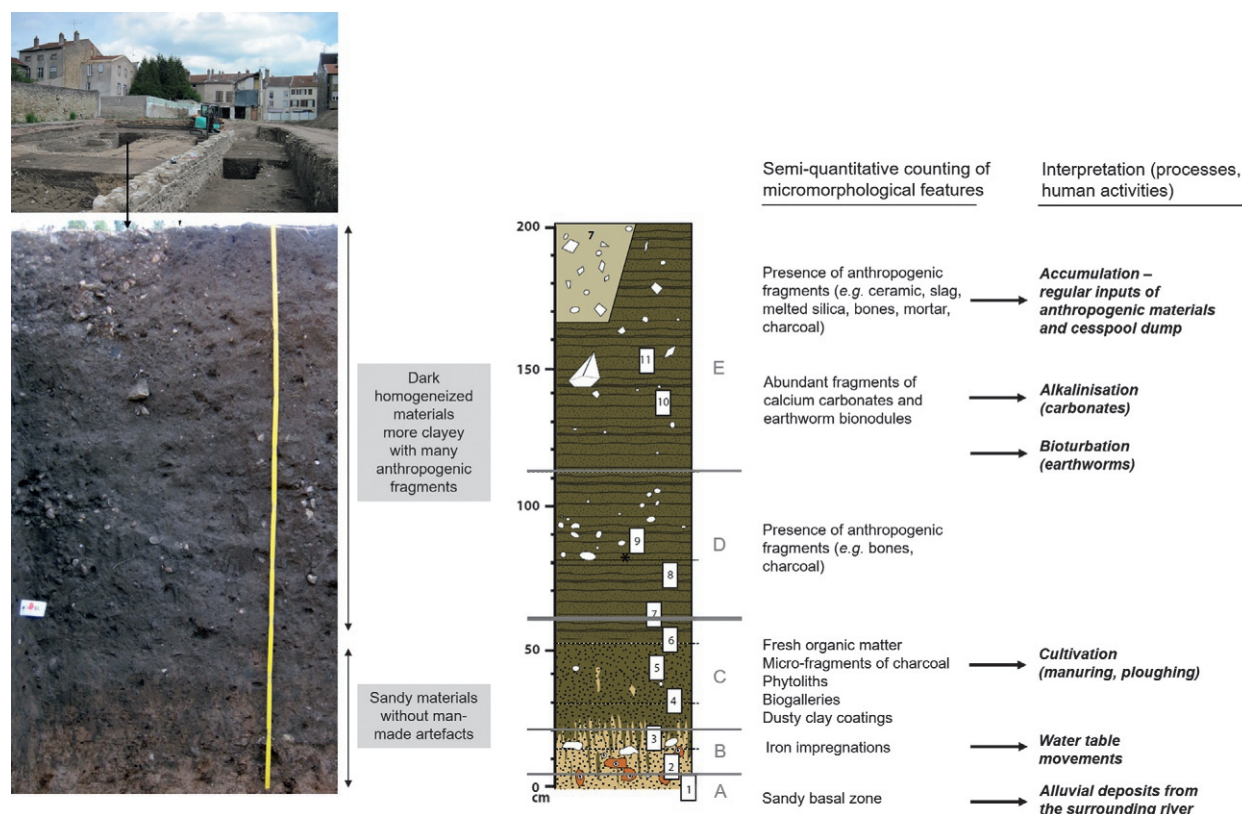


Fig. 4 Example of late medieval dark earth soil formation at the ancient convent garden in Lunéville (North-East France): dark earth (layers D and E) overlying a cultivated alluvial soil (layers A, B and C), based on description of the soil profile, the distribution of artifacts and micromorphological features observed in thin sections along the profile (1 to 11). Source: A. Gebhardt, French national institute for Preventive Archaeological Research INRAP.

As an example, a soil developed on a former settling pond storing blast furnace sludge from the iron industry was investigated using a multi-technique approach (Fig. 5; Huot et al., 2014a,b; 2015). It showed a particular geochemistry dominated by manganese (Mn), calcium (Ca), silicon (Si), iron (Fe), carbon (C) and aluminium (Al) with very large concentrations of potentially toxic metals such as Pb or Zn, which are related to composition of the raw materials and the enrichment of volatile elements in the blast furnace. From a mineralogical point of view, some phases are inherited from the raw material used in the blast furnace, such as Fe oxides from the iron ore. Other phases have been more or less (trans)formed in the blast furnace under high temperature, such as the formation of spherical aluminosilicate glass. Their weathering into poorly crystalline aluminosilicates or the precipitation of secondary calcium carbonates at greater soil depth indicate mineral (trans)formations under the influence of environmental factors in the settling pond. The molecular markers indicated organic compounds of vegetal origin in the surface horizon and of anthropogenic origin in the underlying layers with a signature of oil products and a contribution of combustion processes. The compounds inherited from the iron-making process are particularly well preserved in the soil, probably in relation to the soil mineralogy and physicochemical properties. This soil can be considered an archive of past industrial activities (Huot et al., 2014b). The study of micromorphological features revealed traces of physical, chemical and biological processes, such as the presence of cracks due to wetting–drying cycles, the precipitation of manganic coatings in the cracks related to changes in redox potential or the development of organo–mineral associations at the surface under the action of soil fauna (Huot et al., 2015; Fig. 5). Such studies are useful to propose scenarios of evolution, and to assess the risks of dispersion of contaminants in these industrial soils.

Human-induced modifications on factors of evolution

The preservation of these soils and materials they contain as archives of past human activities will depend on the context of their evolution. In human-affected areas, soils undergo the effects of both natural and anthropogenic factors, creating unique conditions of soil evolution. Human activity can be integrated in the soil-forming factor ‘organisms’, but given the intensity of human-induced soil changes in the Anthropocene, it has been proposed to consider humans as a sixth factor of soil formation (Dudal, 2005). Human impact can be direct or indirect through the modifications of the five natural soil-forming factors.

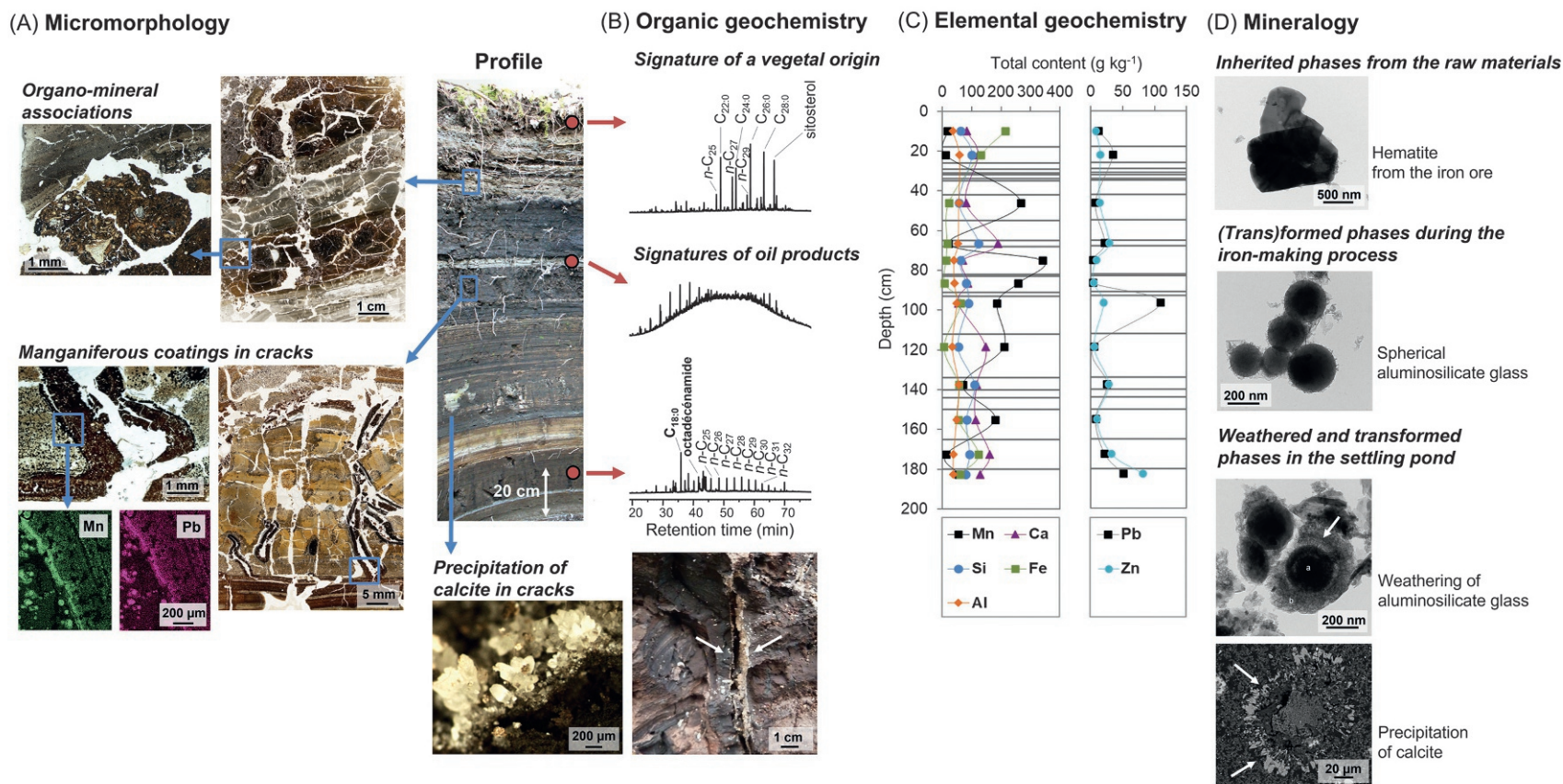


Fig. 5 Multi-technique approach for the reconstruction of the formation of a soil developing on a former settling pond of an iron industry under a forest cover by combining observations and analyses at different scales of the soil profile as tracers of the origin and the evolution of the materials: (A) micromorphological features, (B) molecular markers of the sources of organic matter, (C) distribution of the contents of major and trace elements along the profile and (D) nature and weathering of mineral phases.

Parent materials

The role of the parent materials is crucial in soil formation, especially during the early stages. The degree and rate of weathering depends on the nature and reactivity of parent materials and the environmental conditions. As shown above, SUITMAs encompass soils with varying degrees of human influence on the nature and mode of deposition of the parent materials. As a result, these soils may develop on materials with specific mixtures of natural and anthropogenic constituents, quite different from the composition of natural bedrocks or sediments. The persistence of artifacts and stratigraphic evidence of human activities depends on both the soil environment, in which they are buried, and on the nature of the constituents and conditions of their production. For instance, the nature of the clay and firing temperature influence the robustness of ceramics in soils (Kibblewhite et al., 2015). The susceptibility of anthropogenic materials to weathering is also affected by some of their properties, such as the granulometry, porosity, specific surface area, molecular structure of organic compounds or the degree of crystallinity of minerals. Fragmentation during the manufacturing process or their deposition may favor the weathering of man-made or -transported materials. The main soil properties influencing the dissolution or degradation of archeological artifacts include hydrology which determines the levels of dissolved oxygen, soil acidity or alkalinity, concentrations of solutes and of dissolved organic matter in soil solution, and soil stiffness (measure of resistance to deformation) that condition the physical protection of brittle artifacts. Stratigraphic evidence is better preserved in soils receiving regular inputs of materials, subjected to little or no erosion, with wet anaerobic conditions or with low bioturbation (Huot et al., 2015; Kibblewhite et al., 2015). Human-induced modifications on the exposure of the materials (e.g. burying, sealing) and the soil's environmental conditions (e.g. pollution, acid rain, use of deicing salts, water table fluctuations, fertilizer application) can favor or limit weathering of the parent materials. For example, the study of buried artifacts after one century in an urban soil showed that the stability of artifacts to weathering (glass > bone > mortar > plaster > paint) accorded with the solubility products of the main mineral constituents of these artifacts. Their good state of preservation could be due to carbonates in the soil environment and compaction limiting the fluxes of water and air into and within the soil. The activity of earthworms can help to preserve artifacts by burying them beneath the layer created by casting activity, but could also have the opposite effect by creating deep vertical burrows that increase air and water circulation in the soil. Interactions between weathering processes of diverse artifacts may also influence the rates of weathering. For instance, the leaching of phosphates from bones or carbonates from mortar and the precipitation of protective coatings contribute to the preservation of iron artifacts (Howard et al., 2015).

Topography and hydrology

Human activities can greatly disturb the topography and hydrology. Earthworks related to urban management, mining and quarrying or the effects of bombing in war zones strongly modify the local topography and generate typical anthropogenic landforms. The U.S. Soil Taxonomy distinguishes (i) constructional anthropogenic landforms or microfeatures on land that has been elevated or raised by human activity (e.g. burial mounds, dumps, spoil banks from dredging, earthworks, sanitary landfills, spoil banks, reclaimed land, terraces, embankments) and (ii) destructional anthropogenic landforms or microfeatures on land that has been cut or lowered without being backfilled (e.g. cuts, dredged channels, earthworks, leveled land, open pit mines, quarries, ditches, furrows, terraces, impact craters) (Riddle and Levin, 2017). Human-modified landscapes may have steep slopes that are subjected to severe erosion, runoff and landslides. Remediation of these sites can encompass geomorphology in the design to promote soil formation and vegetation development.

These modifications of the topography have impacts on the groundwater level. In addition, surface and subsurface sealing, buried infrastructures or drainage alter the interlinking between soils and the water regime in human-affected areas (Burghardt, 1994). For instance, the groundwater table rose in urban soils in Moscow because of the accumulation of technogenic materials and gradual leveling of the topography (Alexandrovskaya and Alexandrovskiy, 2000). Furthermore, soils with geomembranes or on green roofs or walls are disconnected from the groundwater.

Water balance in urban and traffic areas shows considerable spatial variation with areas of extreme infiltration or evapotranspiration adjacent to areas with almost no infiltration or evapotranspiration (Burghardt, 1994). For example, roadside soils have a high rate of infiltration compared to those further from the road because of runoff from the road (Wessolek et al., 2011).

Climate

As for natural soils, the climate, mainly the dynamics of precipitation, temperature, and freeze–thaw cycles, is a factor in the evolution of SUITMAs. The climatic conditions also influence the development of vegetation and soil biota, which are crucial soil-forming factors. For instance, a study of soils developed on the tailings of coal mines of different ages from humid to arid conditions in Siberia showed the formation of profiles specific to the different climatic zones (Sokolov et al., 2015).

Human activities affect the climate globally and locally. Urban areas usually have higher temperatures than the surrounding rural areas, especially at night, thereby forming urban heat islands. Climate change might induce significant modifications to temperature and precipitation regimes, with potential augmentation of the intensity and frequency of extreme weather events, such as droughts, heavy rains, cyclones or heatwaves. This will affect the evolution of all soils around the world, including SUITMAs and the anthropogenic materials they contain.

In human-affected areas, modification of the interface between the soil and atmosphere, and artificial hydrologic management moderate the direct effect of natural climatic conditions on SUITMAs. Kodešová et al. (2014) showed that the nature of the soil cover (e.g. sand, bark chips, gravel, concrete paving) influences water and heat regimes in urban soil profiles. Soil sealing is common in urban and traffic areas leading in general to higher soil temperatures, limited water infiltration and gas diffusion depending on the degree of sealing (Scalenghe and Marsan, 2009). Moreover, structural modifications of soils (e.g. compaction, lithological disturbances, presence of artifacts) as well as the burying of pipes may modify both heat and water flow within such soil profiles. Soils covering capped landfills may have higher temperatures due to microbial activity in the refuse layer. Effluxes of gases (mainly methane (CH₄) and carbon dioxide (CO₂)) through cracks in the capping system may change the soil atmosphere and create anaerobiosis (Handel et al., 1997). Regarding water management, irrigation can create similar soil moisture conditions in different climatic zones in highly managed soils, such as golf courses (Obear et al., 2014).

Organisms

Plants, soil microorganisms and fauna play a major role in soil evolution and functioning of SUITMAs as in more natural soils. Human activities modify the habitat of organisms through the alteration of chemical and physical soil properties or through management of the vegetation cover. Consequently, the abundance, diversity, structure and activity of biological communities may be different in SUITMAs, thereby affecting soil development.

Some SUITMAs have unfavorable chemical and physical conditions (e.g. nutrient deficiency, water stress, salinity, extreme pH, compaction, large contents of coarse fragments) and/or problems of toxicity depending on the availability of contaminants they contain, which inhibit organism growth and activities. Soil sealing reduces soil biodiversity. Microbial biomass is reduced under impervious surfaces due to low carbon concentrations, moisture and aeration (Piotrowska-Długosz and Charzyński, 2015). Plant root development can be constrained by debris or compacted layers which act as a mechanical barrier. Urban trees generally have restricted volumes of soil to grow in with limited water and nutrient storage, which reduce the extension of their root system (De Kimpe and Morel, 2000). In cities, habitats are fragmented by the construction of buildings and roads, and green areas are dominated by exogenous ornamental plant species. Consequently, urbanization may lead to biological homogenization (Scalenghe and Marsan, 2009).

Recent studies, however, have revealed unexpected high levels of soil biodiversity in some SUITMAs. For instance, based on communities of soil microarthropods, the biological quality of urban garden soils and industrial brownfields was similar to that of forest soils and higher than that of agricultural soils; this contradicts what might be expected from the physicochemical soil properties (Joimel et al., 2017). Important factors for the development of soil biota are the diversity and density of plant cover, the quantity and quality of organic matter including exogenous organic inputs, and the intensity of land management and frequency of human interventions. The spatial heterogeneities of substrates and of physical and chemical properties, and changes in the microtopography create a mosaic of microhabitats with varying microclimates (e.g. temperature, moisture) in SUITMAs, which favors high biodiversity (Auclerc, 2017). Organisms may take advantage of these heterogeneities to develop in preferential areas like depressions or cracks. For example, in well-drained soils developing on debris, deep rooting to reach layers with more water supply depends on the size and compaction of the stones. A high density of roots and root hairs has been observed on bricks, probably showing their ability to use the water and nutrients stored in them (Nehls et al., 2013; Wessolek et al., 2011).

Regarding the dynamics with time, primary successions can occur on bare substrates, such as mine tailings, industrial by-products or abandoned infrastructure, whereas secondary successions take place in ecosystems after disturbances, such as compaction, erosion, contamination. Pioneer species on such bare substrates modify the habitat for other species. Remediation practices (e.g. leveling, addition of amendment, planting) may help to improve conditions for the development of organisms and to accelerate processes of natural succession. However, remediation techniques could also result in uniform conditions, thereby decreasing the diversity of such ecosystems (Auclerc, 2017).

All these abiotic and biotic constraints on biological communities in SUITMAs may generate specific structures of communities compared with native soils. For example, in temperate climates, a lack of earthworms was observed in various SUITMAs because of sharp artifacts, and unfavorable physical properties (texture, moisture) or toxicity (e.g. Burghardt, 1994).

Time

Soil formation depends on the time from deposition of the parent materials (time zero). In human-influenced areas, however, the period of pedogenesis is related to land use history and degree of human disturbances. Soils subjected to few or indirect influences of human activities (pseudonatural soils) can be as old as surrounding natural soils. However, the input of natural or anthropogenic materials or strong alterations of naturally developed soil profiles may create the conditions to reset soil development to a new time zero. Consequently, soils markedly affected by human activities are relatively young, from several years or decades to thousands of years in ancient cities or mines. The evolution of these soils is controlled by the succession of human activities, which is typical for old settlement places (Alexandrovskiy et al., 2012).

In summary, SUITMAS develop under varying conditions from pseudonatural to highly disturbed and managed environments. Human activities can modify each of the five soil-forming factors and their interactions. This may lead to specific conditions of soil evolution in human-influenced areas (Fig. 6).

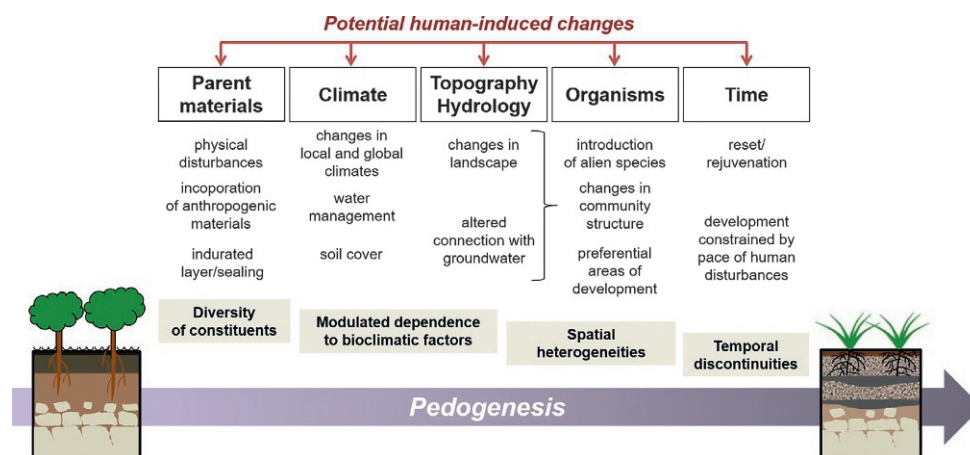


Fig. 6 Potential conditions of soil evolution in human-affected areas. Transport, mixing and incorporation of substrates result in a unique mixture of diverse constituents from both natural and anthropic origins. Alterations of local topography and hydrology, modifications of plant and soil cover and water management modulate the dependence of soils on the bioclimatic factors. These soils are characterized by spatial heterogeneities related to the deposition of parent materials, alteration of local topography and development of organisms in preferential areas. Soil development is controlled by the pace and degree of human disturbances, which can reset the pedogenesis.

Evolution of SUITMAs

In view of the conditions for the evolution of SUITMAs, the question of their long-term evolution arises. What are the consequences of human-induced changes of factors on the soil-forming processes? How can the evolution of such soils and their anthropogenic materials be predicted?

Main soil-forming processes observed in SUITMAs

Based on studies of the pedogenesis of SUITMAs, a list of the main processes of accumulation, transfer and transformation reported in the literature is given below (Huot et al., 2015).

Accumulation and transformation of organic matter

The accumulation of organic matter (OM) at the surface is one of the most commonly reported soil-forming processes in all types of SUITMAs, except for sealed ones. It mainly relates to vegetation and biological activity, and exogenous organic inputs. In urban soils developing on debris, organic matter accumulates over time at an increasing rate initially before a steady state is reached (Wessolek et al., 2011). It can be enhanced by inputs of organic amendments within the scope of remediation of brownfields or mine tailings, or by horticultural and gardening practices (e.g. Howard et al., 2015). In SUITMAs, organic matter may accumulate in deeper layers with the input of organic waste and or by burying the organic horizons of native soils, as in urban Dark Earths. Soil sealing decreases organic matter and nutrient concentrations compared to open sites, because of the removal of topsoil during construction and the absence of organic debris inputs through the impervious covers (Kida and Kawahigashi, 2015).

Degradation, transformations and migration of organic substances depend on many biotic and abiotic properties, such as the nature and recalcitrance of the organic compounds, the microbial and faunal communities and the soil conditions (e.g. pH, moisture, mineralogy, porosity). For example, OM degradation may be retarded in some SUITMAs due to waterlogged conditions (Alexandrovskiy et al., 2012) or physical protection in small pores, and stabilization by interaction with reactive mineral surfaces (Huot et al., 2014b). Soil sealing induces a decrease in microbial biomass and enzyme activities, contributing to a decrease in OM turnover (Piotrowska-Długosz and Charzyński, 2015). The lack of earthworms observed in different types of SUITMAs may contribute to limiting incorporation and mixing of OM in the soil.

Mineral transformations

The evolution of SUITMAs is marked by the weathering and neoformation of minerals, mainly controlled by changes in pH and redox conditions in relation to climatic and biologic factors.

Oxidation of sulfides and the subsequent formation of sulfates and iron oxides commonly occur when materials brought to the surface by human activities from reducing environments (e.g. sulfide-containing mine tailings, dredged sediments, excavated rocks) are in contact with oxygen and water, and subjected to the action of some microorganisms. This reaction releases sulfuric acid, leading to the formation of acid mine drainage rich in potentially toxic metals. The degree of sulfide weathering increases with time and can be an indicator of material transformations (Uzarowicz and Skiba, 2011).

Changes in redox conditions in relation to wetting-drying cycles trigger the dissolution, translocation and precipitation of Fe and Mn oxides in soils formed in iron-rich materials (e.g. iron industry by-products, ferrous artifacts) or fertilized with Fe

compounds. These processes lead to the formation of Fe—Mn concretions (Sokolov et al., 2015), coatings in cracks and on mineral surfaces (Huot et al., 2014a), crusts associated with iron artifacts (Howard et al., 2015) or cemented layers (Obear et al., 2014).

Changes in pH are also an important driver of mineralogical modifications. Acidification resulting from the decomposition of organic matter, percolation of water or sulfide oxidation promote the dissolution and leaching of carbonates from calcareous anthropogenic materials. This may cause the alkalization of buried natural acidic soils, such as volcanic soils under road sub-base layers (Kida and Kawahigashi, 2015) or podzolic soils under historical deposits of construction debris (Alexandrovskiy et al., 2012). Precipitation of secondary calcite as crusts, coatings in cracks or nodules has been observed in various SUTIMAs formed in industrial by-products, mine dumps or construction debris (e.g. Huot et al., 2015; Howard et al., 2015). Acid reactions in anthropogenic materials may also promote the weathering of primary silicates into clay minerals (Uzarowicz and Skiba, 2011). Alkaline pH values are conducive to the weathering of aluminosilicate glass resulting from industrial processes (Huot et al., 2015).

Transfers of soluble compounds and particles

The leaching of soluble compounds (sulfates, carbonates or chlorides) is often reported in various SUTIMAs developing on anthropogenic materials, such as construction debris or industrial by-products. For example, the leaching of sulfates from WWII debris has led to contamination of the groundwater of Berlin (Wessolek et al., 2011). Even under asphalt surfaces, water infiltration through cracks enables dissolution of calcium cations from subbase materials and accumulation of calcium carbonates in deeper horizons (Kida and Kawahigashi, 2015). In addition, translocation of fine particles that filled cracks was observed in some SUTIMAs, for instance in urban soils from the construction debris deposits into the underlying soil in cracks (Alexandrovskiy et al., 2012).

Development of soil structure

The development of soil structure in SUTIMAs is mainly controlled by the inputs of organic matter and the activity of roots, microorganisms and soil fauna, but also depends on the deposition of parent materials and particle mixing related to human activities. The formation of granular structure under the activity of roots and fauna has been observed in different anthropogenic substrates. Aggregates involving anthropogenic mineral and organic components, such as hydrocarbons, green waste or sewage sludge, has been reported in brownfield soils or in constructed soils (e.g. Séré et al., 2010).

Differentiation of soil horizons

Profiles of SUTIMAs are generally little differentiated because of their young age and their organization is marked by the deposition of materials and traces of human intervention. However, a rapid morphological differentiation in the layers occurred in a soil constructed by the deposition of different layers of anthropogenic materials (Séré et al., 2010). Accumulation of organic matter at the surface contributes to the formation of organic (O) and organo–mineral (A) horizons with increasing thickness over time (Bini and Gaballo, 2006). The (O) –A–C profiles can evolve towards (O) –A–B–C profiles, with the development of a cambic horizon or accumulations in layers at depth resulting from the leaching and precipitation of sulfates, carbonates (e.g. Howard et al., 2015) or Fe oxides (Obear et al., 2014).

Specific pedogenetic trends in SUTIMAs

Most of the pedogenetic processes observed in SUTIMAs seem similar to those in soils in natural environments (Huot et al., 2015). However, specific human-induced conditions of evolution may generate some particular features in the combination, localization and temporality of the processes in SUTIMAs.

Concomitant processes governing the pedogenesis of different types of natural soils have been observed in some SUTIMAs, probably related to the diversity of their constituents. For instance, the study of soil developing on by-products of the iron industry; Fig. 5 shows the weathering of glass and neoformation of poorly ordered aluminosilicates as in volcanic soils, segregation of Fe and Mn as in hydromorphic soils and migration and precipitation of carbonates as in calcareous soils (Huot et al., 2015). Processes which are usually typical of different climatic regions may also occur. For example, the leaching of sulfates from construction debris or industrial by-products was observed in temperate humid regions (Séré et al., 2010; Wessolek et al., 2011), whereas this process is generally more often encountered in arid regions. This could result from the presence of anthropogenic materials with a distinctive composition from the natural materials in this climatic region. Soil sealing and management may also reduce the dependence of the soil-forming processes on climatic condition. For instance, Fe-cemented layers observed in golf course soils may develop independently from the climate because all golf course soils are fertilized with Fe, engineered with textural discontinuities and irrigated to provide similar soil moisture conditions (Obear et al., 2014).

Frequently, SUTIMAs display considerable spatial and vertical variation from materials deposited and human disturbances. This can favor local processes, such as preferential precipitation of minerals in cracks (Huot et al., 2014a) and or at a textural discontinuity (Obear et al., 2014), and preferential biological development, and associated processes like rhizosphere effects. This may lead to variable rates and orientations of pedogenesis over a short distance (Bini and Gaballo, 2006).

Rapid and intense weathering has been often reported in the early stages of pedogenesis of SUTIMAs. This could be attributed to the disequilibrium between anthropogenic materials and environmental conditions (Séré et al., 2010). By contrast, soil development can be hindered by restricted biological activity, sealing or recalcitrance of some anthropogenic materials. As in alluvial soils, the degree of development of soils in human-influenced areas depends on the rate of deposition of materials and human

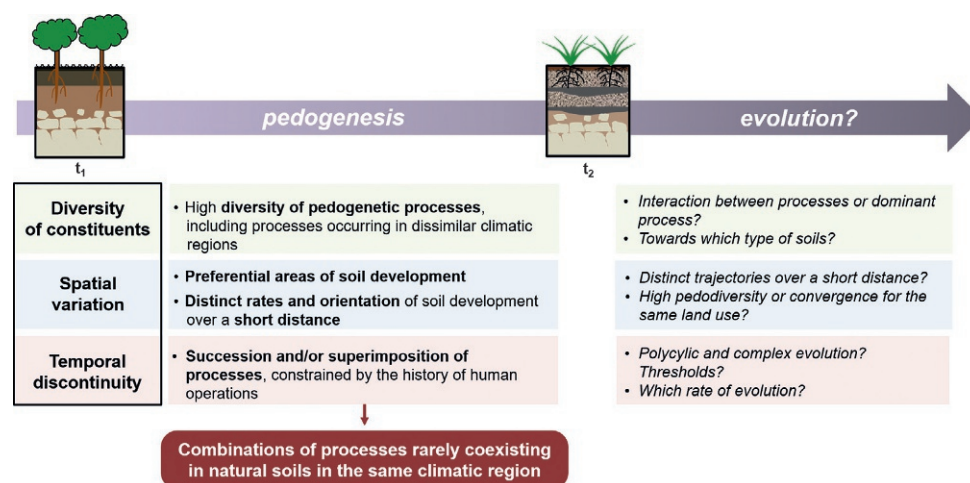


Fig. 7 Specific pedogenic trends and hypotheses of evolution of SUITMAs. The diversity of constituents, large spatial heterogeneity, and development constrained by the pace of human operations may create conditions for a combination of pedogenetic processes in SUITMAs, which rarely coexist in natural soils developing under the same climatic conditions. These trends raise several questions about the possible evolution of SUITMAs, including the direction of pedogenesis, diversity or polycyclic evolution of such soils. Adapted from Huot H, Simonnot M-O, and Morel JL (2015) Pedogenetic trends in soils formed in technogenic parent materials. *Soil Science* 180: 182–192.

disturbances. Urban soils of Moscow contain thin organic layers corresponding to the development of soils under herbaceous vegetation and trees when the time span between consecutive inputs of construction debris was greater than 10 or 20 years (Alexandrovskaya and Alexandrovskiy, 2000). The new top layers may interrupt pedogenesis and preserve the features acquired during previous surface weathering in buried layers.

In summary, the diversity of constituents, relative independence from the bioclimatic factor, the spatial heterogeneity, and soil development constrained by the pace of human interventions may create conditions for a combination of pedogenetic processes in SUITMAs, which rarely coexist in natural soils developing under the same climatic conditions (Fig. 7; Huot et al., 2015).

Prediction of the evolution of SUITMAs

Based on the specific trends in the pedogenesis of human-influenced areas, several questions arise about the possible evolution of SUITMAs (Fig. 7). The large diversity of SUITMAs and their relatively young age hamper the prediction of their long-term evolution. Modeling of pedogenesis can be a predictive tool to determine the partial influence of the main factors controlling soil development. However, quantitative pedogenesis models adapted to human-affected soils still need to be developed. Integrating the human influence into existing pedogenetic models requires consideration of the presence and type of anthropogenic materials and the considerable variation in soil physico-chemical properties, spatial heterogeneities, and temporal changes at different scales (Legu  dois et al., 2016). In this respect, it seems essential to improve the quantification of soil-forming processes, together with the monitoring and simulation of pedogenesis under the influence of human activities. The investigation of older SUITMAs and reconstruction of their past evolution with the history of the sites can help us to understand human-induced changes on the paths of evolution.

Direction of the pedogenesis

The potential combinations of soil-forming processes in anthropogenic soils raise the question as to whether the development of such soils will be controlled by the interactions between these processes or by a dominant process that will determine the direction of pedogenesis and the type of soil towards which these systems will evolve. If short-term laboratory experiments can be used to simulate the weathering processes of anthropogenic materials under various controlled conditions, they should be complemented by long-term field observations of SUITMAs and analogous natural soils (Scholtus et al., 2014). For soils developing on excavated geological materials, the comparison with natural soils formed on outcrops of similar parent materials under similar climatic conditions could be of value to predict long-term evolution. For example, the study of a young soil formed on excavated calcareous clayey rock, and of agricultural and forest soils formed on similar parent materials has shown that development of the anthropogenic soil will be controlled in the short term by the leaching of salts and accumulation of organic matter, and in the long term by decarbonization and brunification processes. This will lead to the formation of Cambisols (Scholtus et al., 2014). For soils developing in human-made materials, it can be assumed that their influence on pedogenesis will decrease with time, depending on the nature and quantity of anthropogenic materials and the conditions of weathering. Some predictions have been made by analogy with regional soils formed in natural parent materials with similar properties to the anthropogenic ones and under the same climatic conditions (e.g. S  r   et al., 2010; Sokolov et al., 2015). Sequences of evolution have been proposed from the study of

soils developing on anthropogenic materials abandoned at different times (chronosequences) over centuries or even millennia compared with adjacent natural soils (Bini and Gaballo, 2006).

Spatial variation and soil diversity

Spatial variation in the composition of parent materials and/or of hydrology, and the presence of hotspots of biological activity may generate the potential for distinct soils to develop over short distances in human-influenced areas. For instance, soil sealing can lead to the formation of soils that are distinct from the native and surrounding soils solely by limiting some processes, such as OM accumulation (Kida and Kawahigashi, 2015). This could increase pedodiversity at the scale of one industrial or mining site or in a city. However, the relative homogenization of anthropogenic processes across the world could also lead to a reduction of soil diversity in human-affected areas compared to natural sites (Burghardt, 1994). Human activities could contribute to the convergence of soil properties depending on the land use with actions taken to limit the constraints related to environmental factors, such as irrigation, leveling of the topography, use of pesticides, or fertilization. This hypothesis of convergence of urban ecosystems has been evaluated for some soil chemical properties in five cities in the world across four biomes. It showed that some urban soil characteristics such as organic C and N concentrations converged across these cities under the influence of soil management practices. But other properties (e.g. phosphorus (P) and potassium (K) concentrations) associated with local pedogenic factors such as parent materials were distinct depending on the cities (Pouyat et al., 2015). From a historic point of view, the investigation of cultural layers in ancient cities of Russia in boreal and steppe zones showed a zonality of the urban deposits related to changes in the bioclimatic conditions and associated processes, but also related to the nature of the anthropogenic inputs; a more organic character in forest zones and a more mineral character in the steppe zones. However, the evolution of construction techniques in Russian cities in the last centuries has led to more similar deposits of construction debris between cities that are independent of the bioclimatic area (Alexandrovskiy et al., 2012). Such results take us to a new hypothesis of a greater similarity between SUITMAs with the same land use from various geographic positions than between them and surrounding natural soils.

Temporal discontinuities

The SUITMAs are also likely to be polycyclic from the successive inputs of materials, leading to the development of incipient soils and burying of the native soils (De Kimpe and Morel, 2000; Bini and Gaballo, 2006). The superimposition of distinct types of materials could lead to separate weathering trajectories in successive incipient soils. However, the evolution of anthropogenic materials and the underlying buried soils may also be interconnected. For instance, the leaching of carbonates from the layers of construction debris may cause the alkalization of buried native acidic soil (Alexandrovskiy et al., 2012). The development of anthropogenic soils is likely to be marked by thresholds in response to changes of external natural or anthropogenic factors or intrinsic changes of soil properties. Understanding and modeling their evolution requires integration of the periodic changes in the short-term and changes in trend in the long-term. The decadal scale is appropriate to monitor human-induced changes of various soil properties (Legu  dois et al., 2016). Monitoring data at this time scale is necessary to detect the transition points in the soil's evolution trajectory and to improve our knowledge about rates of soil-forming processes in SUITMAs, especially by chronosequence studies (Bini and Gaballo, 2006).

Concluding remarks

The soils and human-made or human-transported materials they contain contribute to information stored about past human activities and the development of inhabited areas. In this way, their investigation constitutes a valuable complement to historical and archeological studies. Conversely, the characterization of these complex soils and the reconstruction of their formation requires a multidisciplinary approach that couples history, archeology and geosciences. Such studies provide information about the fate of human-made materials and potentially toxic compounds, and the possible paths of soil development. Sustainable management of inhabited areas requires us to consider the potential of soils in land planning. The challenge is to provide indicators of soil quality for land-use planning. In such areas, it can be assumed that the quality of soils depends on the historical evolution of the land uses it underwent. Based on this assumption, mapping of the land-use history and associated soil disturbances can be a valuable tool to help decision making about the preservation of areas of interest for green spaces, urban agriculture or biodiversity (Franck-N  el et al., 2015). Future research is needed to link history of land uses and soil use potential better. In addition, longer-term monitoring of these soils and further characterization of the artifacts and their reactivity would improve the prediction of their evolution in the context of global change. Thus, the soils of highly human-influenced areas open up a new path for soil science, i.e. prediction of their pedogenesis, which is essential for the sustainable management of urban ecosystems, based on an in-depth analysis of their past.

References

- Alexandrovskaya EI and Alexandrovskiy AL (2000) History of the cultural layer in Moscow and accumulation of anthropogenic substances in it. *Catena* 41: 249–259.
 Alexandrovskiy AL, Dolgikh AV, and Alexandrovskaya E (2012) Pedogenetic features of habitation deposits in ancient towns of European Russia and their alteration under different natural conditions. *Bolet  n de la Sociedad Geol  gica Mexicana* 64: 71–77.

- Auclerc A (2017) Regulating services provided by urban soils: Biodiversity. In: Levin MJ, Kim K-HJ, Morel JL, Burghardt W, Charzyński P, Shaw RK, and edited on behalf of IUSS Working Group SUITMA (eds.) *Soils Within Cities—Global Approaches to Their Sustainable Management—Composition, Properties, and Functions of Soils of the Urban Environment*, pp. 213–220. Germany: Stuttgart.
- Bini C and Gaballo S (2006) Pedogenic trends in anthrosols developed in sulfidic mine spoils: A case study in the Temperino mine archaeological area (Campiglia Marittima, Tuscany, Italy). *Quaternary International* 156: 70–78.
- Borderie Q, Devos Y, Nicosia C, Cammas C, and Macphail RI (2019) Chapter 18. Dark Earth in the geoarchaeological approach to urban contexts. In: Arnaud-Fassetta G and Carcaud N (eds.) *La Géochronologie Française au XX^{ème} siècle*, pp. 245–255. Paris: CNRS Alpha, CNRS Éditions.
- Burghardt W (1994) Soils in urban and industrial environments. *Zeitschrift für Pflanzenernährung und Bodenkunde* 157: 205–214.
- Burghardt W, Morel JL, and Zhang GL (2015) Development of the soil research about urban, industrial, traffic, mining and military areas (SUITMA). *Soil Science and Plant Nutrition* 61(sup1): 3–21.
- Certini G and Scalenghe R (2011) Anthropogenic soils are the golden spikes for the Anthropocene. *The Holocene* 21: 1269–1274.
- de Kimpe CR and Morel JL (2000) Urban soil management: A growing concern. *Soil Science* 165: 31–40.
- Dudal R (2005) The sixth factor of soil formation. Presented at International Conference on Soil Classification, August 3–8 2004, Petrozavodsk, Russia. *Eurasian Soil Science* 38: S60–S65.
- Franch-Néel C, Borst W, Diome C, and Branchu P (2015) Mapping the land use history for protection of soils in urban planning: What reliable scales in time and space? *Journal of Soils and Sediments* 15: 1687–1704.
- Handel SN, Robinson GR, Parsons WFJ, and Mattei JH (1997) Restoration of woody plants to capped landfills: Root dynamics in an engineered soil. *Restoration Ecology* 5: 178–186.
- Howard JL, Ryzewski K, Dubay BR, and Killion TW (2015) Artifact preservation and post-depositional site-formation processes in an urban setting: A geoarchaeological study of a 19th century neighborhood in Detroit, Michigan, USA. *Journal of Archaeological Science* 53: 178–189.
- Huot H, Simonnot M-O, Watteau F, et al. (2014a) Early transformation and transfer processes in a Technosol developing on iron industry deposits. *European Journal of Soil Science* 65(4): 470–484.
- Huot H, Faure P, Biache C, et al. (2014b) A Technosol as archives of organic matter related to past industrial activities. *Science of the Total Environment* 487: 389–398.
- Huot H, Simonnot M-O, and Morel JL (2015) Pedogenetic trends in soils formed in technogenic parent materials. *Soil Science* 180: 182–192.
- IUSS Working Group WRB (2014) *World Reference Base for Soil Resources 2014. International soil classification system for naming soils and creating legends for soil maps*. (No. 106), World Soil Resources Reports Rome: FAO.
- Jenny H (1941) *Factors of Soil Formation: A System of Quantitative Pedology*. New York: McGraw-Hill Book Company Inc.
- Joimel S, Schwartz C, Hedde M, et al. (2017) Urban and industrial land uses have a higher soil biological quality than expected from physicochemical quality. *Science of the Total Environment* 584–585: 614–621.
- Kibblewhite M, Tóth G, and Hermann T (2015) Predicting the preservation of cultural artefacts and buried materials in soil. *Science of the Total Environment* 529: 249–263.
- Kida K and Kawahigashi M (2015) Influence of asphalt pavement construction processes on urban soil formation in Tokyo. *Soil Science and Plant Nutrition* 61: 135–146.
- Kodešová R, Fér M, Klement A, et al. (2014) Impact of various surface covers on water and thermal regime of technosol. *Journal of Hydrology* 519, Part B: 2272–2288.
- Leguédols S, Séré G, Auclerc A, et al. (2016) Modelling pedogenesis of technosols. *Geoderma* 262: 199–212.
- Morel JL, Chenu C, and Lorenz K (2015) Ecosystem services provided by soils of urban, industrial, traffic, mining, and military areas (SUITMAs). *Journal of Soils and Sediments* 15: 1659–1666.
- Nehls T, Rokia S, Mekiffer B, et al. (2013) Contribution of bricks to urban soil properties. *Journal of Soils and Sediments* 13: 575–584.
- Obear GR, Hartemink AE, and Soldat DJ (2014) Soils with iron-cemented layers on golf courses in the USA. *Geoderma* 232–234: 198–207.
- Piotrowska-Długosz A and Charzyński P (2015) The impact of the soil sealing degree on microbial biomass, enzymatic activity, and physicochemical properties in the Ekranic Technosols of Toruń (Poland). *Journal of Soils and Sediments* 15: 47–59.
- Pouyat RV, Yesilonis ID, Dombos M, et al. (2015) A global comparison of surface soil characteristics across five cities: A test of the urban ecosystem convergence hypothesis. *Soil Science* 180: 136–145.
- Prokof'eva TV and Poputnikov VO (2010) Anthropogenic transformation of soils in the Pokrovskoe-Streshnevo park (Moscow) and adjacent residential areas. *Eurasian Soil Science* 43: 701–711.
- Riddle RL and Levin MJ (2017) Anthropogenic soil criteria, identification and classification of human-altered and human-transported materials. In: Levin MJ, Kim KHJ, Morel JL, Burghardt W, Charzynski P, Shaw RK, and edited on behalf of IUSS Working Group SUITMA (eds.) *Soils Within Cities, Global Approaches to their Sustainable Management—Composition, Properties, and Functions of Soils of the Urban Environment*, pp. 27–34. Germany: Stuttgart.
- Scalenghe R and Marsan FA (2009) The anthropogenic sealing of soils in urban areas. *Landscape and Urban Planning* 90: 1–10.
- Scholtus N, Echevarria G, and Florentin L (2014) Expected evolution of a technosol derived from excavated Callovo-Oxfordian clay material. *Journal of Soils and Sediments* 15: 332–346.
- Séré G, Schwartz C, Ouvrard S, et al. (2010) Early pedogenic evolution of constructed technosols. *Journal of Soils and Sediments* 10: 1246–1254.
- Simonson RW (1959) Outline of a generalized theory of soil genesis. *Soil Science Society of America Journal* 23: 152–156.
- Sokolov DA, Androkhonov VA, Kulizhskii SP, Domozhakova EA, and Loiko SV (2015) Morphogenetic diagnostics of soil formation on tailing dumps of coal quarries in Siberia. *Eurasian Soil Science* 48: 95–105.
- Targulian VO and Bronnikova MA (2019) Soil memory: Theoretical basics of the concept, its current state, and prospects for development. *Eurasian Soil Science* 52: 229–243.
- Urbańska M and Charzyński P (2021) SUITMAs as an archive of the human past: Educational implications. *Journal of Soils and Sediments* 21: 1928–1937.
- Uzarowicz L and Skiba S (2011) Technogenic soils developed on mine spoils containing iron sulphides: Mineral transformations as an indicator of pedogenesis. *Geoderma* 163: 95–108.
- Wessolek G, Kluge B, Toland A, et al. (2011) Urban soils in the vadose zone. In: Endlicher W (ed.) *Perspectives in Urban Ecology*, pp. 89–133. Berlin Heidelberg: Springer.

Ecozones and soils – An introduction

Rupert Bäumler, Institute of Geography, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

© 2023 Elsevier Ltd. All rights reserved.

Soils and Ecozones	246
Azonal soils	246
Intrazonal soils	248
Concluding remarks	248
References	249

Abstract

The pedosphere and biosphere together, despite their marked differences, form the skin of the Earth's terrestrial ecosystems. The ecozonal concept is one of the first concepts to enable classification of both soils and the biosphere. Climatic factors are the main trigger that creates the differences within soils or soil types, within ecosystem sub-types, and between both. This chapter provides a short introduction to the chapters with a special focus on soils and soil types that follow a zonal concept, but also on soils which follow the law of exception, both are within the Pedology Section of the second edition of the Encyclopedia of Soils in the Environment.

Key points

- Zonal concept of soil classification
- Regional climate as main trigger of soil formation
- Impact of environmental fluctuations
- Azonal and intrazonal soils

Soils and Ecozones

The idea and concept of classification and mapping of soils goes back to the famous Russian Soil Geographer Vasily Dokuchaev. He first introduced the classification of soils in their natural environment based on his research and geographical observations of specific soil types and sequences of soil horizons that dominate in specific climatic areas. Typical examples are permafrost affected soils and patterned ground (Gelisols, Cryosols) in polar and subpolar areas, or a dominance of Podzols (IUSS Working Group, WRB, 2022)/Spodosols (Soil Survey Staff, Soil Taxonomy, 1999) with their specific sequence of eluvial and illuvial horizons in the boreal ecozone, or soils of steppe ecosystems (Fig. 1), or soils of the Ferralsol- and Plinthosol-type in tropical environments with intensive chemical weathering (Fig. 2). Fig. 1 is a Chernozem (WRB)/Mollisol (ST) from northeastern Romania near Scânteia developed in eolian deposits of Pleistocene origin, and Fig. 2 shows an intensively weathered Ferralsol/Oxisol site from limestone weathering in Northeastern Thailand. Dokuchaev's great advantage was the existence of the sequence of almost all ecozones in Russia from subtropical climatic conditions in the south through the dry mid-latitude ecosystems to cold environments north of the Arctic Circle. He identified specific climatic and geographic patterns as the most important driving factors of soil formation, i.e. mean annual temperature and seasonal to daily temperature variations, or the pattern of distribution of precipitation from dry to wet with or without a distinct rainy season. In combination with oxygen, temperature and water (precipitation) are the catalysts of weathering and soil formation and are independent of the influence of other soil forming factors. These are important not only for recent processes of soil formation, but also for paleosols, which provide information about past environmental fluctuations (Fig. 3).

With an increasing interest in soil properties and processes, researchers have recognized that zonal concepts do not work at sites dominated by factors other than climate.

Azonal soils

Azonal soils are young soils with only the initial stages of soil development and weak profile differentiation. Those soils cannot be assigned directly to a specific climate zone and might be found almost everywhere. They are typical for lowlands near rivers characterized by periodic or episodic flood events (alluvial soils; Fig. 4), and for morphologically or tectonically active mountainous landscapes (Zink, 2015; Leptosols/Entisols).



Fig. 1 Mollic Chernozem (Mollisol) developed from Pleistocene eolian deposits, Scânteia, NE Romania, under agriculture and characterized by intensive bioturbation. Photograph by R. Bäumler.



Fig. 2 Intensively weathered Ferralsol (Oxisol) on limestone parent material, NE Thailand, under tropical climate conditions. Photograph by R. Bäumler.

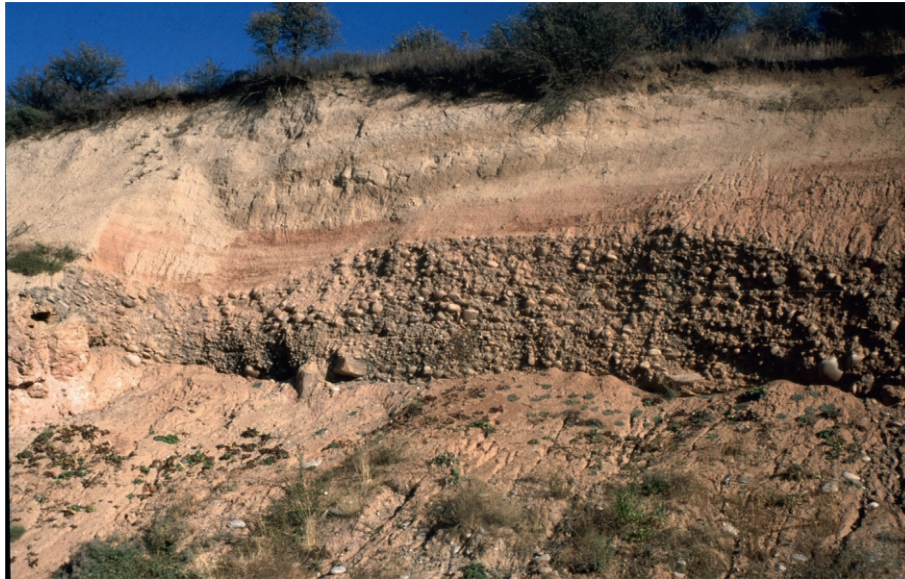


Fig. 3 Loess-Paleosol-sequence above old Quaternary gravels near Samarkand, Uzbekistan. Photograph by R. Bäumler.



Fig. 4 An episodically flooded Fluvisol on top of the lowest terrace of the Seti Khola near Puranchaur, Central Nepal. The bedding indicates the flood events with subsequent initial soil formation. Photograph by R. Bäumler.

Intrazonal soils

Intrazonal soils are characterized by the dominance of factors other than climate with respect to soil formation, like parent material (including organogenic material), hydric conditions, or anthropogenic soils. Typical soil representatives are Gleysols (WRB) or Stagnosols (WRB), Andosols (WRB)/Andisols (ST; Fig. 5), Vertisols (WRB and ST), Terraes calcis, Histosols (WRB and ST), and Anthrosols (WRB) with or without technogenic matter like Technosols (WRB), Terra preta (pretil horizon; WRB), Hortisols (hortic horizon; WRB), soil formation after mining activities or in disposed waste (Howard, 2021; Zech et al., 2022).

Concluding remarks

Deep understanding of soils involves knowledge of its factors of formation and surrounding disciplines, e.g. geology or lithology, geomorphology, climatology or atmospheric science, hydrology or the hydrosphere, flora and fauna or biosphere, and their mutual



Fig. 5 Histric Andosol/Andisol near Gunnarsholt, S-Iceland. Photograph by R. Bäumler.

interactions. Soil knowledge is the key to understanding environmental interdependency, resilience and protection in our fast-changing environment.

Soils no longer seem to be in balance with their biotic and abiotic natural surrounding conditions. Soil degradation by erosion, deforestation, desertification, sealing, soil compaction, misuse of pesticides, such as the neonicotinoids, are some of the problems we are confronted with. The protection of soils, one of our essential natural resources, should be one of our most important challenges in the 21st Century.

Zonal concepts have their justification (Schultz, 2005). Nevertheless, scientific concepts such as the ecozonal concept of soil classification have to be questioned at all times. Ecosystems are open dynamic systems characterized by fluxes of matter and energy. This will consequently lead to blurring of the effects and fluent transitions in combination with small-scale three-dimensional diversity (Mikhailova et al., 2021). This ultimately leads to some defining questions. For example, how can boundaries be defined between adjacent ecozones, or within altitudinal gradients in high mountain areas? Some boundaries such as terrestrial–semi-terrestrial or terrestrial–marine are not in accord with a climate-based ecozonal concept. Is a zonal concept of soil classification or soil formation a congruent synthesis of climate, soils and vegetation? Are ecozones a conceptual abstraction only – or based on natural constants? How should relict properties be embedded? For example are permafrost soils always part of the boreal ecozone? Is the human influence in the Anthropocene part of a natural causal network (Amundson, 2021)?

References

- Amundson R (2021) Factors of soil formation in the 21st century. *Geoderma* 391. <https://doi.org/10.1016/j.geoderma.2021.114960>.
- Howard JL (2021) Urban anthropogenic soils – A review. *Advances in Agronomy* 165: 1–57.
- IUSS Working Group WRB (2022) *World Reference Base for Soil Resources. International soil classification system for naming soils and creating legends for soil maps*, 4th edn, Vienna, Austria: International Union of Soil Sciences (IUSS).
- Mikhailova EA, Zurqani HA, Post CJ, Schlautmann MA, and Post GC (2021) Soil diversity (Pedodiversity) and ecosystem services. *Land* 10(3), 34 pages. <https://doi.org/10.3390/land10030288>.
- Schultz J (2005) *The Ecozones of the World: The Ecological Divisions of the Geosphere*. Berlin, Heidelberg: Springer.
- Soil Survey Staff (1999) Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys. In: *U.S. Department of Agriculture Handbook*, 2nd edn, p. 436. Natural Resources Conservation Service.
- Zech W, Schad P, and Hintermaier-Erhard G (2022) *Soils of the World*. Berlin, Heidelberg: Springer.
- Zink JA (2015) Relationships between geomorphology and pedology: Brief review. *Geopedology: An Integration of Geomorphology and Pedology for Soil and Landscape Studies* 11–26.

Cold Region Soils: Part I—Distribution, Properties and Processes

Nicolas A Jelinski^a, Chien-Lu Ping^b, and John CF Tedrow^c, ^aDepartment of Soil, Water, and Climate, University of Minnesota, Saint Paul, MN, United States; ^bAgriculture and Forestry Experiment Station, University of Alaska Fairbanks, Fairbanks, AK, United States; ^cPosthumous Contribution, formerly with Rutgers University, New Brunswick, NJ, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of JCF Tedrow, Polar Soils, Editor(s): Daniel Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 239–249, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00010-2> and CL Ping, Cold-Region Soils, Editor(s): Daniel Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 268–276, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00024-2>.

Introduction	250
Definition and distribution of cold region soils	251
Soil temperature and moisture regimes	251
Distribution of cold region soils	251
Characteristics of cold region soils	252
Physical processes and properties	252
Permafrost	252
Freezing point depression and water movement in freezing soils	253
Morphological properties	253
Structure	253
Cryoturbation	254
Redoximorphic features	257
Patterned ground	257
Biological and chemical properties	260
Effects of land use and climate change on cold region soils	260
Climate change	260
Engineering and development	260
Agriculture	261
Conclusion	261
References	261

Abstract

Cold region soils are increasingly impacted by anthropogenic activities and climatic change. Despite their cold character, these unique soils are highly dynamic. Low temperatures depress but do not stop biochemical reactions. This results in high rates of carbon accumulation, and slow but significant chemical modification. Volumetric expansion and contraction due to water movement and phase change occurs in both permafrost and non-permafrost-affected soils, and dramatically affects their morphological, physical and chemical properties. Conserving and managing cold region soils sustainably depends on understanding their distribution, properties, and the unique set of physical and biochemical processes that occur in them.

Key points

- Cold region soils are physically and biochemically dynamic.
- Cold region soils are influenced by the presence or absence of permafrost, water movement, ice formation, redoximorphic processes, cryoturbation, and patterned ground.
- These dynamic aspects of soils in cold regions make them susceptible to climatic, environmental, and anthropogenic disturbances.

Introduction

The soils of cold regions are widely distributed and highly unique. Cold region soils may or may not be permafrost-affected, but permafrost-affected soils occur exclusively in cold regions. One of the major physical processes occurring in both permafrost and non-permafrost-affected cold region soils is volumetric expansion and contraction resulting from phase changes of water. Liquid water exists even in soils below 0 °C because of freezing point depression. This water can move towards the freezing front, dramatically influencing ground ice content and morphology as well as the formation of soil structure. The ubiquitous low temperature in cold regions slows the rates of biochemical reactions, which accounts for high rates of soil carbon accumulation

in these regions. However, microbial decomposition of organic matter does still occur, and biotic and abiotic reactions can result in significant modifications to soil parent materials over time. Where permafrost persists, the morphology and properties of cold region soils are extremely heterogeneous, influenced by the movement of soil materials due to freeze-thaw processes (cryoturbation) and linked closely to patterned ground formation. As cold regions continue to experience environmental change and pressures from development on a global scale, understanding the processes that occur in cold region soils and the effect that these factors have on their morphology and properties has never been more important.

Definition and distribution of cold region soils

Soil temperature and moisture regimes

Cold region soils can be defined as soils formed in areas with a mean annual soil temperature (MAST) of less than 8 °C (at 50 cm below the soil surface or at the top of a root limiting layer) and having soil formation controlled at least in part by cryogenesis or cryogenic processes, the physical, biological, and chemical processes that occur during freezing and thawing. This definition of cold region soils corresponds to the frigid, cryic (MAST between 0 °C and 8 °C) and gelic (MAST less than 1 °C or 0 °C) soil temperature regimes (STRs) in US Soil Taxonomy (Soil Survey Staff, 2014). The gelic STR includes the subgelic (MAST of 1 °C to −4 °C), pergelic (MAST of −4 °C to −10 °C), and hypergelic (MAST less than −10 °C) soil temperature classes (Soil Survey Staff, 2014) (Fig. 1). Despite their unifying cold character, these soils experience a wide range of annual temperatures and even wider ranges of annual soil moisture balance. Soil moisture regime varies significantly in cold regions, ranging from arid conditions to humid or aquic (saturated) conditions. These soil temperature and moisture regimes and classes correspond broadly to boreal or subarctic (Dfc, Dwc), arctic (Dfd, Dwd), and high arctic or polar (ET, EF) Köppen climate zones.

Distribution of cold region soils

The northern circumpolar region (north of 45°N latitude) contains the largest extent of cold soils in the world, upwards of 16.6 million km² of ice-free land (Jones et al., 2010). In North America, most of Alaska, Canada (between 48°N latitude in the east and 55°N latitude in western Canada), and the ice-free portions of Greenland are considered cold region soils. In Eurasia, the northern cold soils belt ranges from the lands north of approximately 50°N latitude in the east (near the Kamchatka Peninsula and in Siberia) to 60–65°N latitude in maritime influenced western Europe (Fig. 1).

The southern circumpolar region includes Antarctica and its outlying islands and archipelagos, and Chilean and Argentinian Patagonia south of 46°S. Despite having a large land mass (14 million km²), Antarctica contains a large proportion of ice-covered land, and therefore has a much smaller extent of ice-free cold soils (approximately 49,500 km², or 0.35% of the land area of Antarctica (Bockheim, 2014)). These southern circumpolar cold soils are largely confined to maritime regions, with large extents in

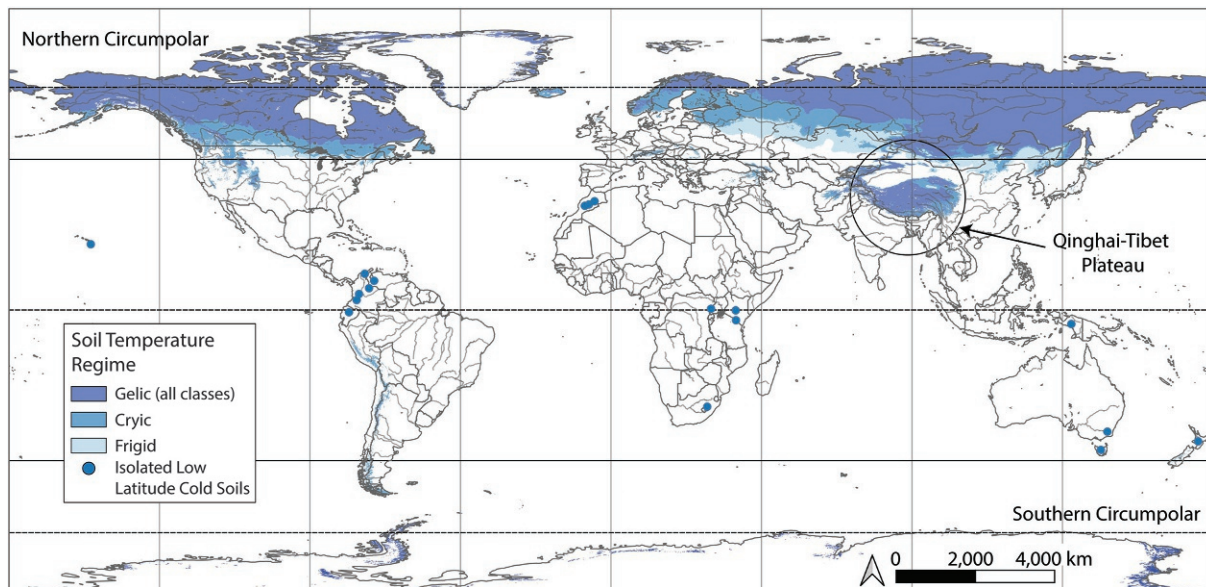


Fig. 1 Distribution of cold region soils defined by soil temperature regime (STR). The global distribution of soils with frigid, cryic, and gelic (inclusive of the subgelic, gelic, and hypergelic soil temperature classes) is represented (Soil Survey Staff, 2014). Data on soil temperature regime are from USDA-NRCS (2016). Bounded regions show areas of focus, including (1) the northern circumpolar region (north of 45 degrees), (2) the southern circumpolar region (south of 45 degrees), and (3) the Qinghai-Tibetan Plateau (QTP). The Arctic Circle (66°33'N), Antarctic Circle (66°33'S) and Equator (0° latitude) are shown as dotted lines.

the McMurdo Dry Valleys in Victoria Land, the Antarctic Peninsula, and outlying islands such as the South Georgia, South Sandwich, and South Shetland Islands (Fig. 1).

Increases in altitude allow cold region soils to persist at lower latitudes. Although of moderate extent, southern extensions of cold region soils exist in the North American Rocky Mountains to latitudes as low as 36°N, and in Europe they persist throughout the European Alps and in isolated pockets of the Caucasus Mountains. In Asia, the largest expanse of altitude-driven cold region soils occurs in the Altai Mountains of Russia and Mongolia, the Tian Shan, Pamir, and Hindu Kush Mountains of China, Kyrgyzstan, Tajikistan, Pakistan and Afghanistan, and across nearly the entire extent of the Qinghai-Tibetan Plateau, a land area of ~1.5 million km² (Fang et al., 2015), reaching the southern flank of the Himalayas in India, Nepal, and Bhutan to 28°N. Low latitude cold soils occur in South America throughout the Andes mountains, stretching from Tierra del Fuego to as far north as Peru. Isolated pockets of cold soils exist on Mauna Loa and Mauna Kea on the island of Hawaii, and on the African continent near the summits of Mt. Kenya and Kilimanjaro, the tallest of the Rwenzori mountains in the Democratic Republic of Congo, and the High Atlas of Morocco (Fig. 1).

Characteristics of cold region soils

Physical processes and properties

Of critical importance to understanding the properties and development of cold region soils are the physical processes of heat transfer, water movement, and phase changes of water, all of which affect the rate and depth of thaw in cold region soils, the presence or absence of permafrost, and influence soil morphology and soil properties. The thermal properties of soil materials (in particular thermal conductivity) are important because the degree and intensity to which soils experience cryogenic processes in cold regions depends in large part on the materials that they are comprised of. The importance of the thermal properties of soil materials on soil temperature is largest at lower latitudes and decreases as latitude increases and the climate becomes colder. Snow depth and composition are additional dynamic factors that can affect the thermal regime and mean annual soil temperature. Snow is an excellent thermal insulator, and soils in areas that receive large, early snowfalls may remain relatively warm throughout the winter. In contrast, soils in areas that receive little to no snow or late snowfalls will freeze back much more rapidly in the winter.

Permafrost

A unifying characteristic of cold regions as defined here is the presence of sporadic to continuous permafrost, which is defined as the thickness of soil, surficial deposits, or bedrock below the surface of the Earth where temperature has remained below 0 °C for at least two consecutive years. In the northern circumpolar region, permafrost is generally ice-cemented. There are, however, locations in mountainous areas, on the Tibetan Plateau, and Antarctica where water or ice is present only at the contact points of the soil particles as hygroscopic water, a condition referred to as dry permafrost (Bockheim and Tarnocai, 1998).

The shallowest depth below the soil surface where materials have remained below 0 °C for at least two consecutive years is considered the top of the permafrost, or permafrost table. During the growing season, heat is transferred from the atmosphere into the soil, and the thawing front proceeds downward. This heat transfer continues until the end of the growing season, when freeze-back can occur both from the soil surface and from the permafrost table below. The active layer is the thickness of the soil above the permafrost table that is subjected to seasonal freeze and thaw. In the northern hemisphere, seasonal thaw usually begins in snow-free areas in April through to early June, but by late September seasonal thaw has largely reached its maximum, even though soils may remain unfrozen well into October or even later under certain conditions. The depth of this seasonal thaw may be less than 30 cm in the far north, but within the northern fringes of forested areas it may extend to depths of 150 to 250 cm or more. In parallel, seasonal thaw in the southern hemisphere begins around October or November and reaches a maximum by February before the onset of the southern winter, although in some ice-free regions of Antarctica temperature may not rise above 0 °C at any time during the year (Kimble et al., 2021). At some depth below the surface, the temperatures of soils, sediments, and rock begin to rise due to geothermal energy. The depth at which materials are no longer below 0 °C is considered the permafrost base.

Permafrost varies significantly in both thickness and extent. It underlies nearly all of the Arctic and Antarctic lands (continuous permafrost), including adjacent off-shore sectors, but underlies only some of the alpine, boreal, subarctic and subantarctic regions, depending largely on soil and ecological conditions (discontinuous or sporadic permafrost). Permafrost can extend hundreds of kilometers south of the tree line, and still much further south into high mountain areas at lower latitudes. The thickness of permafrost varies greatly, reaching thicknesses of more than 1200 m at a few locations in Siberia. However, most research in the field of soil science and soil survey is concerned only with the presence or absence of ‘near-surface’ permafrost, typically defined in soil survey and soil taxonomic systems as permafrost within 2 m of the soil surface (Soil Survey Staff, 2014; IUSS Working Group WRB, 2015).

Permafrost can thaw or be degraded by increased heat flow into the soil. In the Pleistocene it existed across nearly all cold regions, but natural and anthropogenic climate warming during the Holocene, and human land use in cold regions have resulted in accelerated permafrost degradation (French, 1999; Biskaborn et al., 2019). All else being equal, soils with thin surface organic materials, coarser (sandier) textures, large amounts of coarse fragments, and wetter conditions experience increased temperatures and thus are less likely to have near-surface permafrost. Soils with finer textures, thicker and less decomposed organic materials, fewer coarse fragments, and lower moisture content will tend to have colder mean annual soil temperatures or shallower permafrost (if permafrost is present). At the landscape scale, however, these conditions rarely exist independently. For example, wet soils often have thick surficial organic materials, but also increased moisture and extensive subsurface hydrologic flow. The influence of these multiple factors ultimately controls the soil thermal balance and determines whether or not near-surface permafrost is present.

Freezing point depression and water movement in freezing soils

At the end of the growing season, the freezing front moves downward through the active layer and the solid and liquid components of the soil begin to cool. Although the freezing point (the temperature at which a substance turns from liquid to solid) of pure, free water is 0 °C, the freezing point of water in soils is depressed below 0 °C. The phenomenon of freezing point depression in soils depends largely on solute concentration of the soil water (related to soil salinity), the antecedent soil water content, and pore-size distribution of the soil.

Increasing solute concentration in the soil water depresses the freezing point because ions disrupt the nucleation and organization of water into a solid, crystalline phase. Soil pore-size distribution affects the freezing point of soil water because water molecules adsorbed to soil particles have lower free energy. This means that soils with smaller pores (finer textured soils) will, in bulk, experience greater freezing point depression than soils with larger pores (coarse textured soils) where a significant proportion of the water molecules exist far from the surfaces of solid particles and therefore are at higher energy states. Antecedent water content influences freezing point indirectly through the distribution of soil water in pores of different sizes. The collective result of these phenomena is that soils with high salinity and a greater proportion of smaller pores can experience freezing point depression to temperatures of less than –20 °C. Thus, under almost all natural conditions, it is possible for a significant amount of unfrozen water to exist in soils at temperatures below 0 °C.

The presence of liquid water in soils below 0 °C results in several important phenomena. First, this liquid water provides an opportunity for microbial life to continue slowly on in an otherwise static, solid matrix. Second, it means that it is possible for water to move even when soil temperatures are below 0 °C. When soil freezes, water migrates towards the freezing front, and this results in the formation of segregated ice (ice inclusions in soil that separate or segregate soil particles)—which can take the form of ice veins, lenses, bands, or crusts—across a wide range of spatial scales (French and Shur, 2010). This movement of liquid water towards the freezing front results in the growth of segregated ice that is much larger than can be accounted for by the volumetric expansion of water alone. The growth of segregated ice bodies also means that, in ice-cemented permafrost, the upper permafrost tends to be more ice-rich than both the soil above it, which is often desiccated from water movement towards the permafrost table, and the soil well below the permafrost table, which does not generally experience water influx.

Morphological properties

The physical processes described above result in both permafrost and non-permafrost soils in cold regions having a unique set of morphological properties that differ significantly from those of temperate and tropical regions due to freeze-thaw cycles and the presence of either seasonally frozen ground or permafrost. Permafrost has a profound effect on soil morphology because it acts as a barrier to the movement of water and other materials and contributes to the increased intensity of cryoturbation (the physical translocation of soil materials by freeze-thaw processes). In addition, in ice-cemented permafrost, a saturated zone forms above the permafrost table during the growing season, which results in reducing conditions that lead to grayish (depletions) or bluish (gleyed) redoximorphic features in the upper soil.

Structure

When soil freezes, soil water moves to the freezing front forming a thin layer of ice called an ice lens. Because water moves towards the freezing front, the progressive freezing of soils can create a stratified fabric that consists of alternate layers of soil and ice lenses. Thus, a distinctive, platy, cryogenic soil structure can form in the early stages of this process, and after repeated freeze-thaw cycles finer soil particles become oriented along the plate surfaces. With time, the expansion of ice deforms the flat plates into discontinuous, small, curved plates referred to as 'lenticular' fabrics by geocryologists (French and Shur, 2010). In both seasonally frozen and permafrost-affected soils, a crusty or crumb-like structure at the soil surface is also common caused by needle-ice formation during early winter when water is extracted from below and freezes at the soil surface. Ice formation during freezing cycles can also cause a granular cryogenic soil structure to form in surface mineral materials (Fig. 2).

In cold region soils without permafrost, the effect of cryogenic processes on soil structure is largely limited to the topsoil and upper subsoil. Freeze-thaw cycles can fragment larger soil aggregates, increase soil pore connectivity, or consolidate and reinforce existing soil aggregates. In addition, under some conditions the annual formation of ice lenses in the upper subsoil can lead to the development of platy soil structure, which is maintained or reinforced by annual near-surface ice lens formation, provided that biological activity does not disrupt them during the growing season.

One of the most striking morphological components of permafrost-affected soils is the presence of visible, excess ice in the soil profile (Ping et al., 1998). In soils with ice-cemented permafrost, it is important to distinguish between cryostructures (patterns made by inclusion of visible ice that segregate the soil particle) and post-cryogenic soil structure (the structure of soil aggregates formed by cryogenic processes) (Ping et al., 2004). In regions with prevalent near-surface permafrost, soils generally have more strongly expressed cryogenic structure. Commonly an ice-rich layer occurs just below the permafrost table in soil which used to be part of the active layer because of water movement and subsequent desiccation. At the permafrost table, there is often a zone with horizontal and vertical cracks like an ice net (Fig. 3) (Ping et al., 2008). Here, an angular, blocky structure forms and, in the lower portions, the ice-rich layer often contains more than 70% ice. The mineral soil blocks seem to be 'floating' or suspended in the ice matrix (French and Shur, 2010; Ping et al., 2008). When the permafrost table drops and the soil becomes unsaturated, a rectangular-shaped lattice can form reticulate structures (Fig. 3).

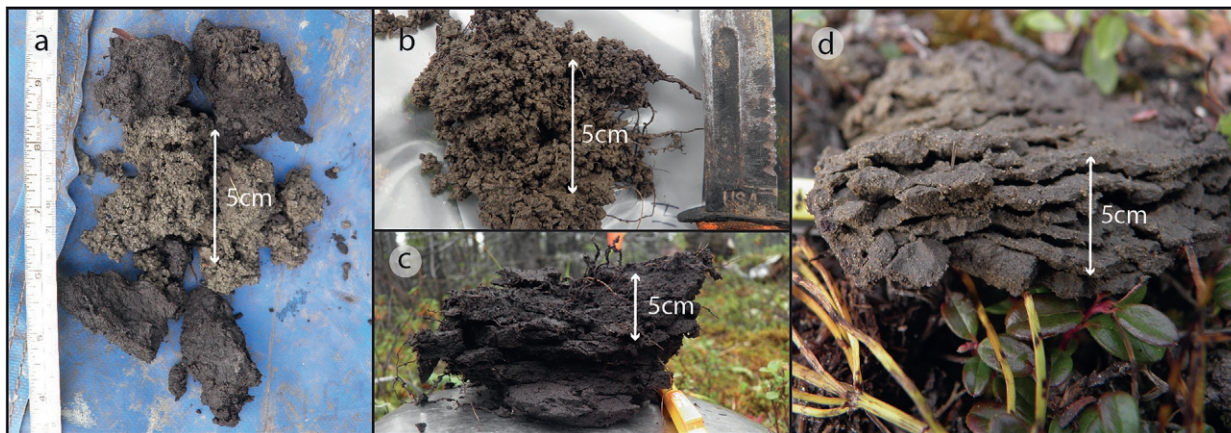


Fig. 2 Representative examples of cryogenic soil structures observed in the active layer in soils with near-surface permafrost: (A) Example of contrasting cryogenic soil structures in active layer materials; A horizons (top and bottom of image)—moderate coarse subangular blocky, and C horizons in lacustrine material (center of image)—strong very fine to fine angular blocky structure. (B) Strong very fine to fine angular blocky cryogenic soil structure in the lower active layer. (C) Moderate to strong very thick platy structure in an A horizon. (D) Strong medium to thick platy structure remains after ice lenses melt in a frozen sample following removal of a core from the lower active layer. All photos by N. Jelinski - figure reprinted with permission from Jelinski N, Sousa M, Williams A, GreyBear E, Finnesand K, Mulligan D, Cole C, Stillinger M, Feinberg J (2019) Cryoturbation and carbon stocks in Gelisols under late-successional black spruce forests of the Copper River Basin. *Alaska Soil Science Society of America Journal* 83: 1760-1778, <https://doi.org/10.2136/sssaj2019.07.0212>.

Cryoturbation

Cryoturbation, by strict definition, refers to the mixing (-turbation) of soil materials by freeze-thaw (cryo-) processes. In practice, the term cryoturbation is used as an umbrella to refer to the wide range of specific physical mechanisms that can result in unique morphological patterns observed in cold region soils, such as irregular or broken horizons, involutions, the accumulation of organic matter at the permafrost table, oriented rock fragments, and silt caps on rock fragments (Soil Survey Staff, 2014). Many of these physical mixing mechanisms do not truly homogenize the soil. Indeed, unlike other heavily turbated soils in temperate or tropical regions that have been homogenized by the shrinking and swelling of clays (Vertisols) or by extensive biological activity (through the action of macroarthropods or ground-dwelling mammals), cryoturbated soils retain distinct ‘packages’ of organic and mineral materials that have been physically translocated in a wide range of patterns and orientations (Ping et al., 2015). Cryoturbation is a common phenomenon in permafrost-affected areas but can also occur in non-permafrost cold region soils, for example in the alpine zone, where the soils are subjected to strong freeze-thaw cycles but lack permafrost.

The major mechanisms that result in the morphological patterns identified as cryoturbation are frost cracking, soil deformation and warping from the growth of ground ice, increased pore water pressures exerted on unfrozen soil as a result of confinement by frozen materials, the flow of less dense materials through and over more dense materials (diapirism, Swanson et al., 1999), differential frost heave (including frost jacking and frost sorting), and the movement of dissolved organic matter or retinization of saturated or dissolved organic materials (Dimo, 1965). Although these processes overlap and may operate together in cold region soils, environmental factors dictate their prevalence and intensity which influence the morphological manifestation of cryoturbation features in the soil. Because many of these mechanisms result in the net downward transfer of surface organic-rich materials to the upper permafrost, cryoturbation is a critical process for generating the large stocks of soil organic carbon that are found in cold regions in general, and particularly in northern latitude permafrost-affected soils (Koven et al., 2009) (Fig. 4).

- Frost-cracking and warping from the growth of ground ice or segregation ice.

In many areas affected by permafrost, the ground cracks due to cold and dry conditions. These cracks provide important opportunities for water and surface organic materials to enter the subsurface, particularly if seasonal ice occupies the cracks and, once thawed, saturated organic materials can move downward through viscous flow into the subsoil. Frost cracking in a primary sense is therefore an important mechanism for generating cryoturbation features in soils. However, the evolution of cracks in newly exposed parent materials, reinforced over long time periods, also dictates ground ice patterns which can warp and distort soil horizons. Once an initial crack forms, subsequent cracks curve to intersect it at near right angles, which maximize strain energy release. Repeated opening and closing of cracks (occurring with annual freeze-thaw cycles in cold environments) results in curving of the initial intersections, resulting, eventually in a unique polygonal pattern on the land surface (French and Shur, 2010). Each year liquid water enters the cracks and freezes in the winter, resulting in the growth of vertical ice veins, which eventually thicken and become ice wedges. The formation of ice wedges increases the internal volume of the soils, and the soils buckle up at the edge of the polygons. In the process, the horizontal soil horizons become warped and distorted (Fig. 4).

- Increased pressures exerted on unfrozen soil as a result of confinement by frozen materials (hydrocryostatic model) and the flow of less dense materials through and over more dense materials (diapirism).



Fig. 3 Representative examples of cryostructures observed in ice-cemented permafrost-affected soils of the Subarctic and Low Arctic. (A) Ataxitic (suspended) cryostructure with 2–3 cm ice belts (Imnavait Creek, Alaska, N.A. Jelinski); (B) Parallel planar layered cryostructure (Utqiagvik, Alaska, C.-L. Ping); (C) Latent micro-lenticular cryostructure (Copper River Basin, Alaska, N.A. Jelinski). (D) Microporphyrific cryostructure (Copper River Basin, Alaska, N.A. Jelinski); (E) Coarse lenticular/ataxitic (suspended) cryostructure (Copper River Basin, Alaska, N.A. Jelinski); (F) Irregular reticulate cryostructure (Franklin Bluffs, Alaska, C.-L. Ping). (G) Micro-braided cryostructure in organic materials (Copper River Basin, Alaska, N.A. Jelinski); (H) Very fine non-parallel wavy and medium crustal cryostructure (Copper River Basin, Alaska, N.A. Jelinski); (I) Wedge ice with threads or foliations of organic materials (Franklin Bluffs, Alaska, C.-L. Ping).

The re-freezing of soils from the top down at the end of the growing season can result in moist, unfrozen materials being squeezed between a freezing front and the permafrost. This can generate confining pressure that can push, deform, or extrude plastic or viscous materials through preferential planes of weakness or cracks in the soil. In extreme form, this hydrocryostatic process can result in the extrusion of subsoil materials at the soil surface. In addition, inversions of bulk density are common in permafrost-affected soils. The general prevalence of ice-rich upper permafrost (due to water movement towards the freezing front



Fig. 4 Morphological manifestations of cryoturbation processes in the Subarctic and Low Arctic. (A) Example of extreme variability in organic layer thickness and morphology due to differential frost-heave in association with a patterned ground feature (in this case stripes) (Imnavait Creek, Alaska, N.A. Jelinski); (B) Deep post-fire cryoturbation of organic-rich materials overprinted with recently developed lenticular cryostructures (Copper River Basin, Alaska, N.A. Jelinski). (C) Ice-rich footslope soil showing mixing of organic-rich materials and underlying mineral materials in diapiric forms (Imnavait Creek, Alaska, N.A. Jelinski); (D) Downward movement of organic materials through vertical cracks in the upper permafrost (Sagwon Hills, Alaska, N.A. Jelinski); (E) Cryoturbation in a shallow permafrost soil showing movement of organic-rich materials due to diapiric and solutioning mechanisms (Copper River Basin, Alaska, N.A. Jelinski); (F) Deformation of alternating organic and mineral layers due to ground ice (wedge ice) (Sagwon Hills, Alaska, N.A. Jelinski); (G) Mixing of organic-rich and mineral materials in the active layer (Copper River Basin, Alaska, N.A. Jelinski); (H) Mixing of organic-rich and mineral materials in the permafrost (Copper River Basin, Alaska, N.A. Jelinski).

and desiccation of the lower active layer) can result in less dense, water-rich materials 'below' more dense materials in upper soil horizons, setting up conditions for diapirism (Swanson et al., 1999), the buoyant upward movement of less dense materials through the planes of weakness in more dense materials above them. This results in unique, swirling and or mushroom-like diapiric forms that can range from micro- to macro-scales (Fig. 4).

- Differential frost heave (including frost jacking and frost sorting).

Due to differences in the thermal conductivities of various soil materials and coarse fragments, freezing and thawing fronts can penetrate the soil at dramatically different rates. Mineral soils, particularly those with a high proportion of coarse fragments, will experience accelerated rates of freezing or thawing front penetration relative to soils that have thick organic materials at the surface. Differential frost heave occurs when soils with little to no surface organic materials freeze back faster at the end of the growing season than immediately adjacent soils with thick organic cover. This situation occurs most commonly in patterned ground configurations such as non-sorted circles, sorted circles, and stripes. When mineral soils or those rich in coarse fragments begin to freeze, liquid water migrates upwards and inwards to join the freezing front. This results in the growth of ice lenses, which can significantly increase the volumetric expansion of the soil, resulting in heave of the ground surface (frost heave) relative to the adjacent, organic covered soil. Differential frost-heave can therefore increase the elevation of mineral materials relative to adjacent organic materials and generate micro-colluvial movement which can 'bury' or 'roll over' the adjacent organic materials, subducting them to the permafrost table over many cycles. This mechanism of cryoturbation is perhaps one of the most widespread at the landscape scale and has been used to explain the ubiquitous organic rich materials associated with the permafrost table in the Arctic. The differential frost heave process also can play a role in situations where fine-earth mineral soils and coarse fragments interact. For example, frost sorting of coarse fragments into sorted circles or stone nets and the frost jacking or orientation of individual coarse fragments also occurs due to differences in freezing front penetration and ice growth between soil material in contact with the coarse fragments and those further away.

- *Leaching or retinization of saturated or dissolved organic materials.*

As soils thaw during the growing season, surface or subsurface organic materials which are saturated can become viscous and mobile. This allows them to move downward through macropores between soil aggregates (solutioning). In addition, stagnant water interacting with organic-rich soil horizons in the upper horizons can pick up significant amounts of dissolved organic matter. As the soil thaws, this water proceeds downwards to the permafrost table, where it stops. This downward movement of dissolved organic matter to the permafrost table has been termed 'retinization' in the Russian literature (Dimo, 1965).

Redoximorphic features

If soil drainage is impeded during thaw and snow melt due to underlying seasonal frost or permafrost, the topsoil remains saturated for longer than the underlying zone. These episaturated conditions can produce reduced and gleyed features near the top of the profile because of the physical effect of frozen materials on impeding the downward movement of water, which is unique to cold regions (United States Department of Agriculture, Natural Resources Conservation Service, 2018). In addition, in soils with permafrost, the permafrost table acts as a barrier to water and roots during the growing season, thus creating a reduced matrix because the water table perches above the permafrost. Thus, the lower active layer in permafrost-affected soils is often gleyed (Fig. 5). Iron reduction in anoxic conditions in soils is a biochemical process, and the presence of significant redoximorphic features in cold region soils is direct evidence of the effect of psychrophilic anoxic microorganisms continuing to function near or below the freezing point of water.

Patterned ground

Cryogenic processes result in self-organized micro landforms known as patterned ground. This self-organization of vegetation and coarse fragments vary across these micro landforms, which may appear "sorted" or "unsorted". 'Patterned ground' is a group term for symmetrical forms such as circles, polygons, hummocks, nets, steps, and stripes. Various terms have been used to describe these patterns, such as 'rutmark,' 'ra~mark,' 'frost polygon,' 'struk- turboden,' 'polygonboden,' 'ice-wedge polygon,' 'stone polygon,' 'soil circle,' 'mud circle,' 'stone ring,' 'stone stripe,' 'hummocky ground,' 'frost boil,' 'ice mound,' 'earth hummock,' and many others. One of the most common forms of patterned ground in Arctic regions are ice-wedge polygons. The ground develops a series of ice wedges that result from soil frost cracking in winter and water freezing in cracks. The ice wedges tend to increase in size year after year and exert strong lateral forces that result in a buckling of the polygons. Changes in elevation of the soil surface adjacent to ice wedges results in formation of the so-called low-center polygons in which the center of the polygon is depressed. Other polygon types, such as high-centered polygons, can form when ice wedges degrade and the soil adjacent to them subsides, leaving a high, dry polygon center with lower outer troughs.

Non-sorted circles or frost boils are common throughout permafrost regions, especially where the soils are of medium to clayey texture (Walker et al., 2004). Frost boils appear on the ground surface as areas of the mineral soil material 'boiled' from underneath the surface, forming circles (Fig. 6). The vegetation cover is most abundant on the edge and outside of the frost boil. Although there are many theories as to how frost boils form, field morphology suggests that it is due to differential frost heave and diapirism. In a well-developed soil profile of a frost-boil, it is evident that organic matter has been cryoturbated into the lower soil horizons (Fig. 6). Even though the processes of formation involved in earth hummocks and frost boils may not be the same, they all result in



Fig. 5 Redoximorphic features in permafrost-affected soils of the Subarctic and Low Arctic. (A) Strong near-surface gleying (indicating biochemical iron reduction) in an otherwise oxidized upland soil—these conditions are induced when thaw depth is shallow early in the season and the soil becomes saturated due to perched water (Imnavait Creek, Alaska, N.A. Jelinski); (B) Reduced pore linings in an upland soil due to seasonal saturation and iron reduction (Fairbanks Area, Alaska, N.A. Jelinski); (C) Strongly gleyed material formed due to saturation above the permafrost table underlying more oxidized upper horizons in a core from a permafrost-affected soil (Copper River Basin, Alaska, N.A. Jelinski) (D) Low chroma, high value (gray) depleted matrix underlying a surface organic horizon in a permafrost-affected soil with a shallow saturated zone (Copper River Basin, Alaska, N.A. Jelinski); (E) Example of rapid color change in organic materials of a Low Arctic drainage basin upon exposure to air—the bright orange material in the center of the photo represents frozen organic materials upon immediate excavation from a soil pit before exposure to oxygen, while the darker materials above and below are the same frozen organic materials after exposure to oxygen for approximately 3 h. This rapid color change is likely due to redox reactions involving manganese and sulfate (Arctic Foothills, Alaska, N.A. Jelinski).

similar soil morphology: cryoturbated organic matter in the lower horizons and warped and broken soil horizons (Walker et al., 2004). Other sets of cryogenic processes related to soil freezing and thawing can result in the development of stone rings (sorted circles). These stone rings can form entirely in coarse fragments, or from an upwelling of the mineral material in the form of a core of raw, earthy material and an outward displacement of the coarser rock fragments (Fig. 6).

On sloping land, there are many elongated forms of patterned ground that result from at least two processes: frost sorting and solifluction (the slow gravity-driven downhill movement of saturated soil). Patterns include stone garlands, stone stripes, soil stripes, and others (Fig. 6). Where solifluction is intense, the genetic soil horizons become highly deformed and often completely obliterated in the upper solifluction lobe.

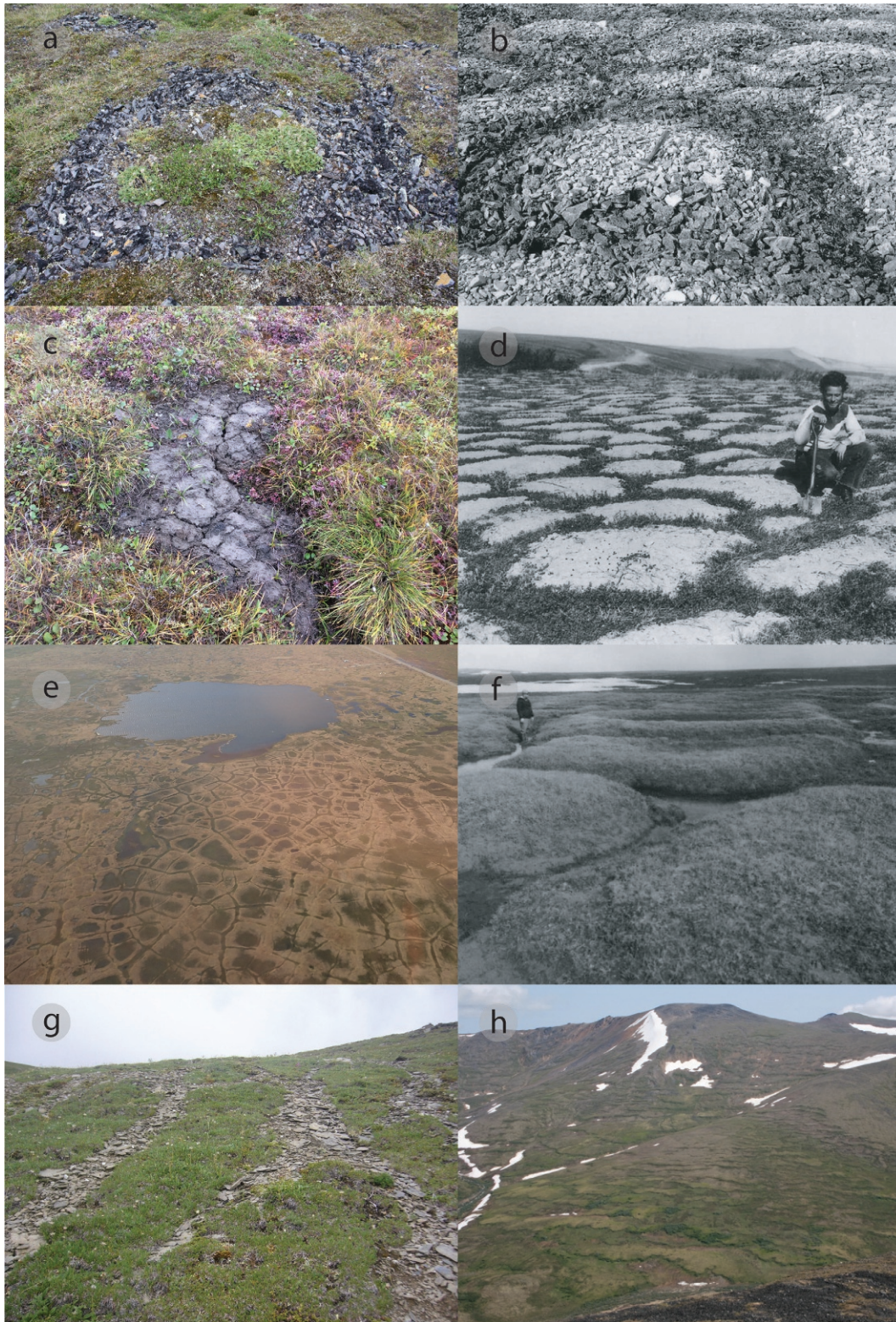


Fig. 6 Representative examples of patterned ground features in Arctic environments. (A) Sorted circle with fine-earth center and coarse-fragment dominated stone outer ring (Brooks Range, Alaska, N.A. Jelinski); (B) Sorted circles (stone rings) formed solely in coarse fragments (Prince Patrick Island, Canada, J.C.F. Tedrow); (C) Mud boil in glacial sediments (Atigun Valley, Alaska, N.A. Jelinski); (D) Non-sorted circles (frost boils) in northern Alaska (J.C.F. Tedrow) (E) Low-centered ice-wedge polygons (Arctic Coastal Plain, Alaska, G. Iwahana); (F) High-centered ice wedge polygons (Prince Patrick Island, Canada, J.C.F. Tedrow) (G) Stone stripes (Brooks Range, Alaska, N.A. Jelinski); (H) Solifluction lobes (Bering Land Bridge National Preserve, D.K. Swanson).

Biological and chemical properties

On an aerial basis, the accumulation of organic carbon in soils of cold regions (particularly the northern circumpolar) exceeds that of any other soil regions on earth. Cold region soils may contain 30–40% of the global stock of soil organic carbon to a depth of 3 m (Hugelius et al., 2013; Ping et al., 2015). The generally cold and moist conditions in cold region soils slow microbial respiration and the decomposition of organic matter. These conditions, in addition to cryoturbation processes, explain why cold region soils store large amounts of carbon even though inputs from aboveground biomass are small.

Temperature and moisture affect the accumulation and distribution of C stocks in the soil profile through controls on vegetation, cryogenic soil processes (such as cryoturbation) and the presence of or depth to permafrost. In addition, the landscape-scale effects of cold temperatures and reduced evapotranspiration in cold regions has resulted in the development of large peatland complexes. Most of these organic soils are in the boreal regions in the Northern Hemisphere: northern Russia, the Canadian Shield, and interior Alaska. Based on recent studies, average organic carbon stocks stored in the peatlands of the northern circumpolar region may exceed 100 kg C m^{-2} (Hugelius et al., 2020). Additionally, in the mineral soils of the Arctic, nearly 50% of the total organic carbon may be stored in the upper permafrost as the result of cryoturbation (Ping et al., 2015). Thus, cryoturbation contributes to the sequestration of atmospheric carbon into the upper permafrost (Koven et al., 2009).

The character of soil organic matter (SOM) in cold soils is influenced by its cryogenic environment, its position in the soil profile, the surface, subsurface soil active layer, and the presence of permafrost. In general, the SOM in cold soils is less decomposed than that in temperate and tropical regions. The SOM in arctic soils can be separated into extractable (EF) and non-extractable fractions (NEF) based on alkali solubility. The NEF, containing cell wall and related constituents, such as hemicellulose and soluble fibers, dominate the SOM of cold soils (Dai et al., 2002). The largest proportions of NEF occur in the SOM of the upper permafrost and this fraction may have a greater potential to influence carbon cycling than EF in these ecosystems with climate warming.

Chemical weathering of most soil parent materials in cold regions tends to be shallower and less intense than in temperate and tropical regions. Permafrost and seasonal freeze-thaw cycles mean that leaching, chemical transformation, and biochemical weathering is limited in most cold soils, relative to soils of temperate and tropical regions. Most cold region soils tend to be relatively unweathered and rich in 2:1 phyllosilicates such as micas, illites, etc. (Ito and Wagai, 2017). Chemical transformation and clay neoformation from weathering rarely extends beyond the formation of mixed 1:1 and 2:1 minerals. Phyllosilicate minerals that indicate more advanced weathering such as kaolinite can occur in cold regions, however they are often inherited from the parent materials rather than produced from mineral weathering in situ (Borden et al., 2010). Although chemical transformations are limited, given sufficient time and appropriate environmental conditions, measurable weathering of soil minerals has been observed even in the coldest regions on Earth (Gibson et al., 1983). Iron oxidation in particular is prevalent across all cold regions, including the cold desert zones of Antarctica (Ugolini et al., 1990; Bockheim, 1997). These biochemical weathering processes are greatest at lower latitudes in cold regions and gradually decrease at higher latitudes.

Effects of land use and climate change on cold region soils

Cold region soils present a special challenge to land use and management owing to the effects of frost heave, the presence of massive ground ice, and accelerated thaw due to natural and anthropogenic climate warming. These issues have caused problems for construction, development and maintenance of transport, structure and foundation instability, and challenges to agricultural development (Hjort et al., 2022). With cold temperatures and short growing seasons, vegetation takes longer to recover after disturbance, and the unique species of wildlife and ecosystems associated with cold soils are globally unique and highly sensitive to ecological change. Natural resources extraction, mainly oil, gas, and minerals, has increased pressure from development and disturbance in cold regions. Optimum solutions for ensuring environmental sustainability lie within our understanding of these cold soils and their environment.

Climate change

The global extent of permafrost and cold region soils has generally decreased throughout the Holocene (Li et al., 2021), however accelerated warming and permafrost degradation is changing cold regions on a massive scale (Biskaborn et al., 2019). This has major implications for cold region soil conservation and development. Cold region soils and permafrost-affected soils in particular hold a significant proportion of the total terrestrial soil carbon stock (Hugelius et al., 2013). Warming may result in the increased release of carbon from permafrost to the atmosphere as microbes become more active and degrade the soil carbon formerly stabilized by cold ground temperatures. This carbon release may potentially offset any major gains in the reduction of anthropogenic carbon emissions and it is therefore of great importance that these cold region terrestrial emissions are constrained (Natali et al., 2021). In addition, the impacts of surface subsidence in regions with massive ground ice (thermokarst) may have far-reaching ecological and biogeochemical consequences, and pose a major threat to cold region infrastructure (Murtin, 2009).

Engineering and development

Infrastructure development in cold regions generally leads to an increased active layer thickness and deeper permafrost, and this results in changes in leaching, oxidation and reduction, and cryogenic processes. In the boreal regions, where permafrost is discontinuous and thin, general practice is either to thaw the permafrost or to insulate it to minimize the effects of frost damage. But in the Arctic region, the best practice is to insulate the permafrost to avoid thawing. Thick beds of gravel are applied for

insulation, and a lighter-color pavement is used to decrease absorption of solar radiation. For permanent structures and houses, gravel pads or insulated foundations are commonly used in soils containing only ice lenses. But in areas where the soils have a large ice content, piling is used for structures and houses. In some cases, the piling or foundation is refrigerated to avoid heat conduction to the permafrost.

Vegetation tends to be slow to recover from disturbance in cold regions because of short growing seasons and cold temperatures. This is especially true in the Low and High Arctic and in the Antarctic, where the result of surface disturbances such as compaction and tracks from vehicles, even when travelled in the winter, can have long lasting effects on vegetation cover, soil thermal regime, the soil biological community, and localized hydrology (Kevan et al., 1995).

Agriculture

Soil affected by permafrost accounts for more than 10% of the exposed global land surface (Obu, 2021). Most of these areas are, in their natural state, tundra, boreal forest, cold deserts, and bogs. However, some areas have been cultivated for a long time. Cold soils present a great challenge to agricultural development not just because of the low temperatures, but because of the presence of ice. In the subarctic or boreal regions, farms newly cleared from native forest with underlying permafrost often face subsidence. The land surface may become hummocked because of melting of ice wedges and even thermokarst sinkholes, where large volumes of ice are present. Such problems are more severe on Pleistocene surfaces because of the larger ice content (Desyatkin et al., 2021). Holocene terraces are generally more favorable for land-clearing and farming.

The native peoples of the circumpolar regions have used Arctic lands for animal production (both wild and managed) for millennia. These people include the Samis in northern Scandinavia, the Inuit in northern Alaska and Canada, and the Chukchi people in northern Russia. In the alpine and plateau regions, the Tibetans raise yaks, sheep, and cattle on the Tibetan Plateau; the native peoples of the Andes have long raised llamas and other domesticated animals. In recent years there have been concerns of overgrazing in these fragile ecosystems. Once the organic layers or the surface vegetation covers are disturbed or destroyed, the quality of the land deteriorates owing to topsoil erosion.

Agriculture in the circumpolar region is made possible because of the long daylight hours in the summer that enable crops to mature in a shorter time than in temperate regions. In the boreal regions, where the soil conditions are favorable, small grains, especially barley (*Hordeum vulgare* L.), can mature within 110 days, and oats (*Avena sativa* L.) within 130 days. Vegetables, such as carrots (*Daucus carota* L.), lettuce (*Lactuca sativa* L.), and cabbages (*Brassica oleracea* L.) can be harvested in less than 60 days. Potatoes (*Solanum tuberosum* L.) can be grown more than 200 km north of the Arctic Circle, where the permafrost table persists at less than 1 m from the surface. Such intensive agriculture can only exist in high-latitude environments because of the prolonged daylight and warmer summer temperatures, however, and not in alpine and plateau environments that experience strong diurnal temperature changes; the temperature can reach more than 20 °C at midday but drop below freezing at night.

The greatest concern of farming in soils affected by permafrost is subsidence. These soils contain variable amounts of ice, therefore, once the soils are stripped of their natural vegetation cover the thermal regime changes and the soils will warm. In soils with a large ice content and ice wedges, thermokarst will occur, and cleared fields become hummocky, resulting in difficult management. Another concern is the nature of the substratum. When the land is cleared, water from the melted ice must be drained or waterlogging or ponding will occur if there is a hydrologically restricting layer, or the land is on a topographic low. Thus, good understanding of the soil and the geographic environment is the key to successful farming in cold soils.

Conclusion

The properties of soils in cold regions are distinctive, and controlled by unique physical and biochemical processes. These processes have resulted in soils that are generally high in organic carbon and have experienced limited chemical weathering. This accumulation of organic carbon has resulted from slow microbial decomposition from cold temperatures, and the physical translocation of organic rich surface materials downward through cryoturbation. However, despite these cold temperatures, biochemical reactions continue and have significant effects on soil morphology and properties. The heterogeneous nature of cold region soils over short distances is linked closely to patterned ground formation and ground ice content, particularly in permafrost regions. These large stocks of organic carbon, the formation of ground ice, and unique biochemical properties make cold region soils highly sensitive to environmental and anthropogenic change. As climate change and human land use accelerate in cold regions it is critical that the body of knowledge on cold region soils continues to be synthesized, improved, and utilized to maximize ecological health and sustainability.

References

- Biskaborn BK, Smith SL, Noetzi J, et al. (2019) Permafrost is warming at a global scale. *Nature Communications* 10: 264. <https://doi.org/10.1038/s41467-018-08240-4>.
- Bockheim JG (1997) Properties and classification of cold desert soils from Antarctica. *Soil Science Society of America Journal* 1997(61): 224–231.
- Bockheim JG (2014) Antarctic soil properties and Soilscales. In: Cowan D (ed.) *Antarctic Terrestrial Microbiology*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-45213-0_16.
- Bockheim JG and Tarnocai C (1998) Nature, occurrence and origin of dry permafrost. In: *Proceedings of the Seventh International Conference on Permafrost, Yellowknife, Canada. Collection Nordicana No. 55*, 57–63.

- Borden PW, Ping CL, McCarthy PJ, and Naidu S (2010) Clay Mineralogy in Arctic Tundra Gelisols, Northern Alaska. *Soil Science Society of America Journal* 74: 580–592.
- Dai XY, Ping C-L, and Michaelson GJ (2002) Characterizing soil organic matter in Arctic tundra soils by different analytical approaches. *Organic Geochemistry* 33: 407–419. [https://doi.org/10.1016/S0146-6380\(02\)00012-8](https://doi.org/10.1016/S0146-6380(02)00012-8).
- Desyatkin RV, Filippov N, Desyatkin A, Konyushkov D, and Goryachkin S (2021) Degradation of arable soils in Central Yakutia: Negative consequences of global warming for Yedoma landscapes. *Frontiers in Earth Science*. <https://doi.org/10.3389/feart.2021.683730>.
- Dimo VN (1965) Formation of a humic-illuvial horizon in soils in permafrost. *Soviet Soil Science* 9(1013): 1021.
- Fang H, Zhao L, Wu X, et al. (2015) Soil taxonomy and distribution characteristics of the permafrost region in the Qinghai-Tibet Plateau, China. *Journal of Mountain Science* 12: 1448–1459. <https://doi.org/10.1007/s11629-014-3133-y>.
- French HM (1999) Past and present permafrost as an indicator of climate change. *Polar Research* 18(2): 269–274. <https://doi.org/10.3402/polar.v18i2.6584>.
- French HM and Shur Y (2010) The principles of cryostratigraphy. *Earth-Science Reviews* 101: 190–206.
- Gibson EK, Wentworth SJ, and McKay DS (1983) Chemical weathering and diagenesis of a cold desert soil from Wright Valley, Antarctica: An analog of Martian weathering processes. *Journal of Geophysical Research* 88.
- Hjort J, Streletskiy D, Doré G, et al. (2022) Impacts of permafrost degradation on infrastructure. *Nature Reviews Earth & Environment* 3: 24–38. <https://doi.org/10.1038/s43017-021-00247-8>.
- Hugelius G, Loisel J, Chadburn S, Jackson R, Jones M, MacDonald G, Marushchak M, Olefeldt D, Packalen M, Siewert M, Treat C, Turetsky M, Voigt C, and Yu Z (2020) Large stocks of peatland carbon and nitrogen are vulnerable to permafrost thaw. *Proceedings of the National Academy of Sciences* 117(34): 20438–20446. <https://doi.org/10.1073/pnas.1916387117>. In this issue.
- Hugelius G, Tarnocai C, Broll G, Canadell JG, Kuhry P, and Swanson DK (2013) The Northern Circumpolar Soil Carbon Database: Spatially distributed datasets of soil coverage and soil carbon storage in the northern permafrost regions. *Earth System Science Data* 5: 3–13. <https://doi.org/10.5194/essd-5-3-2013>.
- Ito A and Wagai R (2017) Global distribution of clay-size minerals on land surface for biogeochemical and climatological studies. *Scientific Data* 4: 170103. <https://doi.org/10.1038/sdata.2017.103>.
- IUSS Working Group WRB (2015) *World Reference Base for Soil Resources 2014. Update 2015 International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*. Rome: FAO. World Soil Resources Reports No. 106.
- Jones A, Stolbovoy V, Tarnocai C, Broll G, Spaargaren O, and Montanarella L (2010) *Soil Atlas of the Northern Circumpolar Region*, 144. Luxembourg: Publications Office of the European Union.
- Kevan PG, Forbes BC, Kevan SM, and Behan-Pelletier V (1995) Vehicle tracks on high arctic tundra: Their effects on the soil, vegetation, and soil arthropods. *Journal of Applied Ecology* 32(3): 655–667. <https://doi.org/10.2307/2404660>.
- Kimble JM, Seybold C, Harms D, Aislabie J, McLeod M, Paetzold R, Sletten R (2021) *Soils and soil climate dataset*, Antarctica since 1999. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/climate/?cid=nrcs142p2_053772
- Koven CD, Friedlingstein P, Ciais P, Khvorostyanov D, Krinner G, and Tarnocai C (2009) On the formation of high-latitude soil carbon stocks: The effects of cryoturbation and insulation by organic matter in a land surface model. *Geophysical Research Letters* 36: L21501. <https://doi.org/10.1029/2009gl040150>.
- Li TY, Baker JL, Wang T, et al. (2021) Early Holocene permafrost retreat in West Siberia amplified by reorganization of westerly wind systems. *Communications Earth & Environment* 2: 199. <https://doi.org/10.1038/s43247-021-00238-z>.
- Murton JB (2009) Global warming and thermokarst. In: Margesin R (ed.) *Permafrost Soils. Soil Biology*, vol. 16. Berlin, Heidelberg: Springer https://doi.org/10.1007/978-3-540-69371-0_13.
- Natali SM, Holdren JP, Rogers BM, Treharne R, Duffy PB, Pomeroy R, and MacDonald E (2021) Permafrost carbon feedbacks threaten global climate goals. *Proceedings of the National Academy of Sciences of the United States of America*, 118.
- Obu J (2021) How much of the earth's surface is underlain by permafrost? *Journal of Geophysical Research - Earth Surface* 126(5): e2021JF006123. <https://doi.org/10.1029/2021JF006123>. In this issue.
- Ping CL, Bockheim JB, Kimble JM, Michaelson GJ, and Walker DA (1998) Characteristics of cryogenic soils along a latitudinal transect in arctic Alaska. *Journal of Geophysical Research* 103(D22): 28917–28928.
- Ping CL, Jorgeson TM, and Shur Y (2004) Soil cryogenic structures and their implications for land use in Alaska. *Journal of Glaciology and Geocryology* 26: 1–7.
- Ping C, Michaelson GJ, Kimble JM, Romanovsky VE, Shur Y, Swanson DK, and Walker DA (2008) Cryogenesis and soil formation along a bioclimate gradient in Arctic North America. *Journal of Geophysical Research*, 113.
- Ping CL, Jastrow JD, Jorgenson MT, Michaelson GJ, and Shur YL (2015) Permafrost soils and carbon cycling. *The Soil* 1: 147–171. <https://doi.org/10.5194/soil-1-147-2015>.
- Soil Survey Staff (2014) *Keys to Soil Taxonomy*, 12th edn. Washington, DC: USDA-Natural Resources Conservation Service.
- Swanson DK, Ping CL, and Michaelson GJ (1999) Diapirism in soils due to thaw of ice-rich material near the permafrost table. *Permafrost and Periglacial Processes* 10: 349–367.
- Ugolini FC, Sletten RS, and Marrett DJ (1990) Contemporary pedogenic processes in the Arctic: Brunification. *Science Du So* 28(4): 333–348.
- United States Department of Agriculture, Natural Resources Conservation Service (2018) In: Vasilas LM, Hurt GW, and Berkowitz JF (eds.) *Field Indicators of Hydric Soils in the United States, Version 8.2*. USDA, NRCS. In cooperation with the National Technical Committee for Hydric Soils.
- USDA-NRCS (2016) *Soil Climate Regions Map*. Washington, DC: Soil and Plant Sciences Division. <https://www.nrcs.usda.gov/wps/portal/nrcs/main/soils/use/worldsoils/>.
- Walker D, Epstein H, Gould W, Kelley A, Kade A, Knudson J, Krantz W, Michaelson G, Peterson R, Ping CL, Reynolds M, Romanovsky V, and Shur Y (2004) Frost-boil ecosystems: Complex interactions between landforms, soils, vegetation and climate. *Permafrost and Periglacial Processes* 15: 171–188. <https://doi.org/10.1002/ppp.487>. In this issue.

Cold region soils: Part II – Genesis and classification

Nicolas A Jelinski^a, Chien-Lu Ping^b, and John CF Tedrow^c, ^aDepartment of Soil, Water, and Climate, University of Minnesota, Saint Paul, MN, United States; ^bAgriculture and Forestry Experiment Station, University of Alaska Fairbanks, Fairbanks, AK, United States; ^cPosthumous Contribution, formerly with Rutgers University, New Brunswick, NJ, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of J.C.F. Tedrow, POLAR SOILS, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 239–249, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00010-2> and C.-L. Ping, COLD-REGION SOILS, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 268–276, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00024-2>.

Introduction	263
Genesis and classification of cold region soils	263
Northern Circumpolar soils	264
Parent materials	264
Boreal and subarctic regions	264
Low Arctic regions	267
High Arctic soils	268
Soils of the southern circumpolar region	270
Soils of the Subantarctic Forest and Subantarctic Tundra	270
Soils of the Antarctic Polar Desert	273
Soils of the Antarctic Cold Desert	274
Alpine and Plateau soils	275
Conclusion	278
References	278

Abstract

Cold region soils are strikingly beautiful, highly diverse, globally important and increasing affected by anthropogenic activities and climatic change. In this chapter, we examine the global distribution and unique and diverse characteristics of these soils through a focus on their genesis and classification in each of the three major zones of cold region soils (the Northern Circumpolar, Southern Circumpolar, and Qinghai–Tibet Plateau). By highlighting the variable formation of cold region soils around the world, we can identify gaps in our knowledge of them and enable better prediction of the future trajectory of these soils as they respond to change.

Key points

- Cold region soils are globally distributed and highly diverse
- Soils of the three major cold regions (the Northern Circumpolar, Southern Circumpolar, and Qinghai–Tibet Plateau) share some similarities but differ in critical ways.
- Soil forming factors in cold regions are aligned along latitudinal gradients (in the Northern and Southern Circumpolar) and with elevation (in low latitude Alpine regions and the Qinghai–Tibetan Plateau), allowing a general understanding of the distribution and properties of cold region soils.

Introduction

The soils of cold regions are highly diverse. They hold globally important terrestrial carbon stocks, affect local and regional hydrology, and support ecosystem functions at high latitudes and altitudes. An understanding of their distribution and properties can only come from detailed study of the unique set of factors that has shaped these soils both within and across cold regions. Although united by similar sets of biophysical and chemical processes that operate in cold environments, the genesis of cold soils is far from homogeneous. The beauty and complexity of these soils becomes apparent when taking a global perspective on their distribution, genesis, and classification.

Genesis and classification of cold region soils

General frameworks for understanding the genesis of cold region soils have been constructed by evaluating the common processes of their formation and spatial trends by climatic or geographical zones (Ugolini, 1986; Bockheim and Ugolini, 1990). In this

chapter we take a zonal approach, similar to previous authors, to identify broad trends in soil genesis and soil forming factors between cold regions. It is important to note that the zonal approach applies primarily to well-drained soils under representative climate and vegetation types, and that regional soil diversity is always much greater than implied by zonal characterizations (Tedrow, 1977). Therefore, we also attempt to illustrate some of that regional variation within the zonal framework. Despite the inherent limitations of the zonal approach, it provides a framework to understand the major trends in pedogenesis necessary for a global perspective. When using taxonomic names in this discussion, we provide both the taxonomic name in the US Soil Taxonomy (Soil Survey Staff, 2014) and World Reference Base for Soil Resources (WRB, IUSS Working Group, 2015), with the WRB name provided in parentheses.

Northern Circumpolar soils

The Northern Circumpolar Region (NCR) contains more than 17 million km² of ice-free cold region soils, which is the largest area of ice-free cold region soils in the world. We refine the generalized NCR pedogenic zones of Tedrow (1977) and Ugolini (1986) with the addition of spatial information from the Circumpolar Arctic Vegetation Map (CAVM) bioclimate zones (Walker et al., 2005), and include Boreal regions with frigid STRs (soil temperature regimes, US Soil Taxonomy, Soil Survey Staff, 2014) to 45° latitude. We subsequently group the NCR into the Boreal and Subarctic zone (including frigid STRs to the Boreal/temperate boundary), the Low Arctic zone (CAVM bioclimate zones D and E) and the High Arctic zone (including the polar and subpolar deserts (Ugolini, 1986, Tedrow, 1977), and CAVM bioclimate zones A, B, and C)) (Fig. 1).

Parent materials

Despite the large ice sheets which covered all of Canada, Fennoscandia and western Europe during the last glacial maximum, many parts of the NCR remained ice-free throughout much of the Pleistocene. These ice-free areas include the Alaskan Interior and Alaska North Slope as well as the Lena river basin, Cherskiy and Verkhoyansk mountains, and Kolyama Lowland in northeast Russia. In areas covered by continental or mountain glaciers, NCR soils have formed predominantly in deposits such as glacial till, glaciolacustrine or glaciofluvial materials, while in areas not covered by ice during the last glacial maximum, soils are often blanketed by loess, pedis sediment, or colluvium of varying thickness or formed in alluvium or residual materials (Börker et al., 2018). Additionally, marine sediments exposed subaerially due to isostatic rebound are common in maritime regions of the northern coasts of Alaska, Canada, and Siberia.

Boreal and subarctic regions

Boreal and Subarctic soils are associated with frigid, cryic, and gelic mean annual soil temperatures (MAST; US Soil Taxonomy, Soil Survey Staff, 2014). The Boreal and Subarctic regions experience the warmest temperatures and highest precipitation of the NCR zones. Most of the southern Boreal is permafrost-free, while northern Boreal areas have isolated and sporadic permafrost, and the Subarctic is largely characterized by discontinuous permafrost.

- **Climate:** Climate in Boreal regions ranges from cold ultra-continental climates (such as mid-Canada and Siberia), to more moderate cold maritime-influenced climates in coastal areas. In the southern continental parts of the Boreal, average high temperatures may exceed 20 °C for several summer months, while for at least 3 months during the winter, average high temperatures are often below 0 °C in the same areas. In the wettest parts of the Boreal, annual average rainfall can exceed 800 mm year⁻¹ and significant snowfall (>2000 mm), with total precipitation as water equivalent exceeding 1000 mm. The continental Subarctic may experience average high temperatures >20 °C in summer from less than 1 to 3 months, and average high temperatures of -10 °C to -20 °C for much of the winter. Precipitation tends to be less than in the Boreal, with rainfall ranging between 200 and 400 mm, with an additional 50–200 mm water equivalent as snowfall. In general, two-thirds to over 90% of total precipitation in the Boreal and Subarctic regions occurs as rain during the growing season. In maritime influenced regions of the Boreal and Subarctic temperatures are moderated and precipitation is generally higher.
- **Vegetation.** Most upland soils in the Boreal and Subarctic have formed under coniferous forest dominated by one or two of four major Boreal tree genera (*Picea* sp. L. (spruce), *Abies* sp. L. (fir), *Pinus* sp. L. (pine), or *Larix* sp. L. (Larch/Tamarack), with ground flora dominated by cryptogams (mosses (bryophytes) and lichens). Coniferous needles and moss or lichen litter tend to be acidic, which reduces microbial decomposition and leads to the widespread accumulation of organic materials at the soil surface. The prevailing soil moisture, drainage, and temperature conditions dictate the thickness and decomposition state of these surface organic materials. They are excellent thermal insulators, are critical for regulating soil temperatures, and protect discontinuous, sporadic, and isolated near-surface permafrost across Subarctic and Boreal regions where permafrost would not otherwise persist.

For this reason, permafrost-affected soils in Boreal and Subarctic regions are highly sensitive to disturbance, and fire is an important factor in controlling soil properties and vegetation patterns. The Boreal forest burns and is subject to frequent lightning-ignited fires during the summer. In soils recently affected by fire, vegetation is commonly altered to willow (*Salix* sp. L.) shrub and aspen (*Populus* sp. L.) or mixed white spruce–aspen forest. While pH in the uppermost organic layers of Boreal and Subarctic soils is typically acidic, following fire it can increase to near-neutral or slightly alkaline due to the concentration of base cations released from the combusted organic materials. These pH values typically decline as the organic layer regrows and re-acidifies. Combustion and consolidation of the insulating surface organic materials can warm underlying soils, and where there is near-surface permafrost it can lower the permafrost table. Under ‘stable’ climatic conditions, soil and ecosystem recovery occurs over centuries, with return to pre-burn conditions as the surface organic layers recover and the permafrost table aggrades.

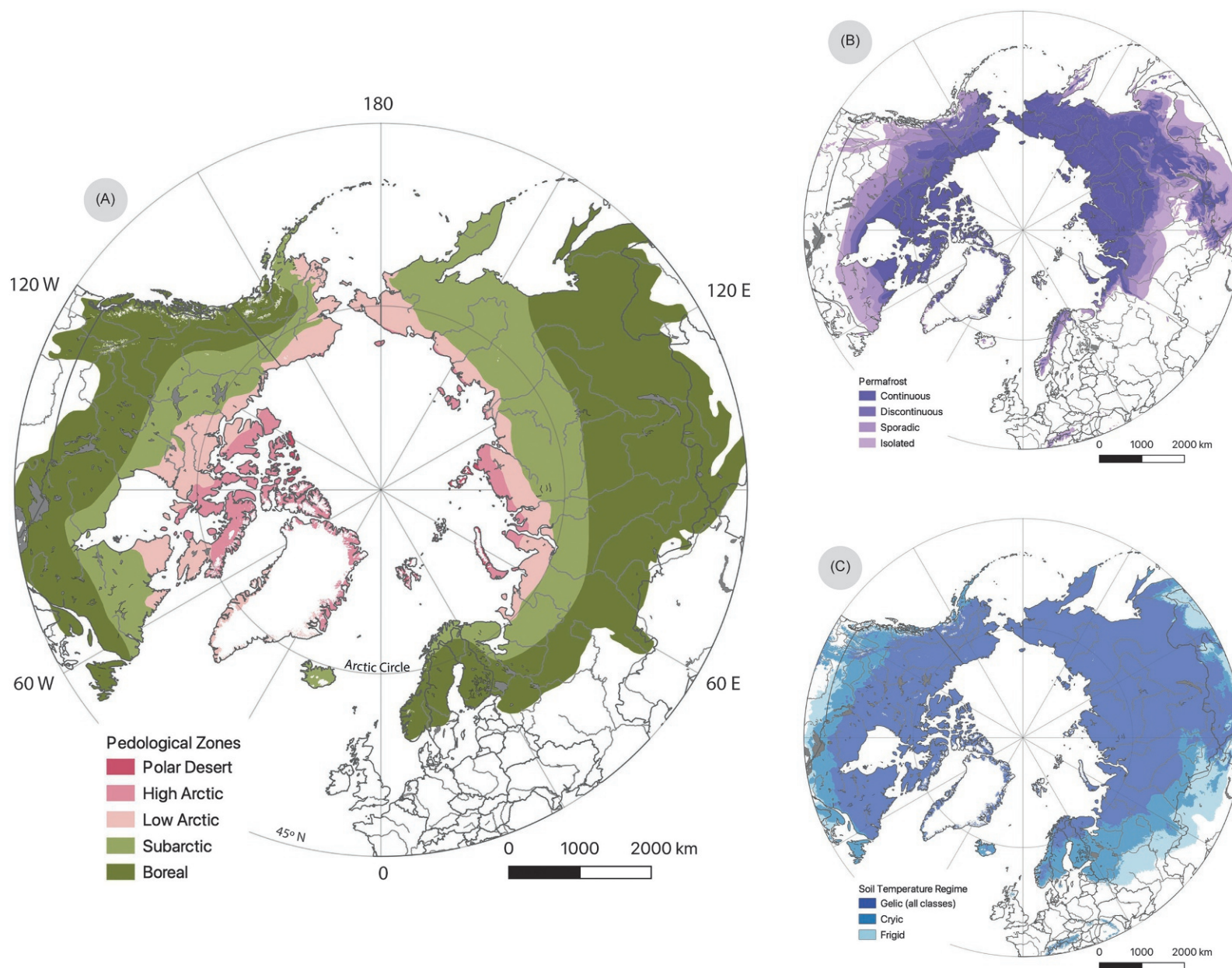


Fig. 1 Pedological zonation, permafrost extent, and soil temperature regimes of the Northern Circumpolar region. (A) Pedological zones are re-interpreted from [Tedrow \(1977\)](#) and [Ugolini \(1986\)](#) with addition of the spatial distribution of Arctic bioclimate Subzones from the Circumpolar Arctic Vegetation Map (CAVM) project ([Walker et al., 2005](#)). The distribution of polar desert (a portion of the High Arctic) corresponds with Arctic bioclimate subzone A, while the rest of the High Arctic includes bioclimate subzones B and C. The Low Arctic includes CAVM bioclimate subzones D and E. The Subarctic–Boreal boundary is interpreted from [Ugolini \(1986\)](#). The southernmost extent of the Boreal zone is interpreted with the southern Boreal boundary from the Boreal Forest monitoring project data ([Potapov et al., 2008](#)). (B) Permafrost zonation is from the Circum-Arctic Map of Permafrost and Ground-Ice Conditions ([Brown et al., 2002](#)). (C) Soil temperature regime data are from [USDA-NRCS. \(2016\)](#).

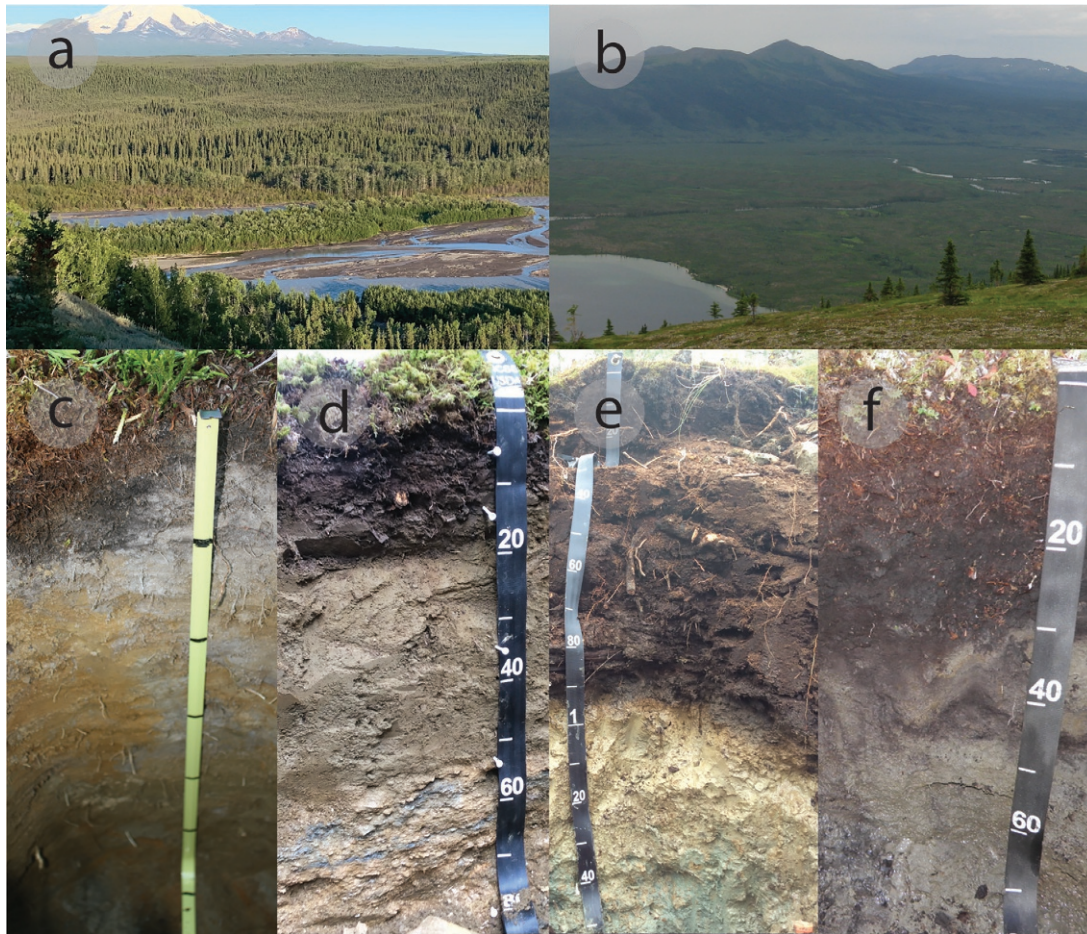


Fig. 2 Representative landscape and soils of the Boreal and Subarctic regions. (A) Boreal landscape and vegetation (southeast Alaska, N.A. Jelinski); (B) Subarctic landscape on the southern flanks of the Brooks Range, Alaska, USA, N.A. Jelinski); (C) Spodosol (Podzol) formed in glacial outwash, Walker Lake, Alaska (N.A. Jelinski), (D) Dystricryept (Cambisol) formed in loess over residuum, Fairbanks Area, Alaska (N.A. Jelinski), (E) Hemist (Histosol) over glaciolacustrine sediments in northern Minnesota, USA (N.A. Jelinski), (F) Histoturbel (Cryosol), formed in glaciolacustrine sediments, Copper River Basin, Alaska (N.A. Jelinski).

- *Soil formation.* In upland soils of the southern and maritime influenced regions of the Boreal and Subarctic, temperatures are warm and infiltration intense in well drained soils without permafrost. Pedogenesis is largely controlled by excess moisture and free drainage, resulting in the leaching (eluviation) and downward movement (illuviation) of dissolved organic matter and iron oxy-hydroxides (podzolization), fine clays (argilluviation), and carbonates (decalcification). These processes are similar to those in northern parts of the humid temperate zone, except for the more intense seasonal freeze–thaw cycles in cold regions (Fig. 2). The most common mechanism of soil formation in well-drained coarse-textured parent materials is podzolization, caused by the action of weak acids from acidic litter. Podzolization is common in the Boreal where rainfall exceeds 500 to 700 mm year⁻¹. Argilluviation is rare in the Subarctic and sporadic in the Boreal, whereas it dominates soil development in humid temperate regions. Finally, iron oxidation and acidification in upper soil horizons leads to brunification, resulting in well developed, oxidized upper soil horizons.

The differences between soil properties on opposing aspects of the same hillslope in the Boreal and Subarctic can be quite dramatic and are amplified with increasing slope due to major differences in effective solar radiation, which affects seasonal moisture balance and soil temperature. For example, in Subarctic Interior Alaska, acidic soils with near-surface permafrost and thick surface organic materials are common on northeast facing slopes which receive little direct solar radiation during the growing season. In contrast, on southwest facing slopes receiving greater solar radiation, warmer, drier soils develop, with thin surface organic materials and carbonates higher in the soil profile.

In poorly drained lowlands of the Subarctic and Boreal landscape, soils have developed under forested or non-forested wetlands dominated by sedges (*Carex* sp. L.), mosses and lichens. The combination of cold temperatures and anaerobic conditions in these wetlands can result in the accumulation of a thick deposits of organic materials (Fig. 2). The peatlands of the Boreal and Subarctic lowland regions of Canada and western Siberian Lowlands (the Hudson Bay Lowlands and the Ob and Yenisei Lowlands) and the Kolyma Lowland of eastern Siberia hold some of the largest carbon stocks of any soils in the world (Hugelius et al., 2020).

Saturation and iron reduction (gleization) is a major soil forming process in poorly drained soils (which includes both lowland soils and permafrost-affected upland soils). Many soils in the Boreal and Subarctic, particularly those with near-surface permafrost, undergo gleization due to long-term saturation and iron reduction at some depth. In addition, seasonal gleization closer to the soil surface can also occur in pockets in otherwise oxidized upland soil, when, during snowmelt, seasonal frost or permafrost does not enable downward percolation of excess water and the upper soil remains saturated for some time.

Cryoturbation occurs in Boreal and Subarctic soils, but typically only in those with near-surface permafrost, and at a lower intensity than in the Low and High Arctic (Ping et al., 2008a; Jelinski et al., 2019). Where patterned ground does occur it is often dominated by vegetated earth hummocks that can often be relict (Walker et al., 2005; Ping et al., 2008a,b). Frost cracking does not play a major role in the Boreal and Subarctic due to warmer temperatures, greater rainfall, and well vegetated protected surfaces. Because frost cracking is less intense than at higher latitudes, lenticular cryostructures from ice lens formation are the predominant cryostructures (Jelinski et al., 2019).

- **Classification and carbon stocks.** Soils of the Boreal and Subarctic are dominated by Inceptisols (Cambisols) and Gelisols (Cryosols), which cover approximately two thirds of the ice-free land area (36% and 32%, respectively). Spodosols (Podzols) are the next most prevalent soil order (15%), followed by Alfisols (Luvisols/Albeluvisols) in the southern Boreal (5%), and Histosols (Histosols) (~4%), with minor components of Entisols (Regosols), Mollisols (Chernozems) and Andisols (Andosols). Although Andisols comprise a minor component of Boreal and Subarctic regions, they can be the dominant soils along the Pacific 'ring of fire' in the Boreal zone. Carbon stocks (to 1-m depth) in the Boreal and Subarctic can be high, ranging from 10 kg C m⁻² in well-drained mineral soils to >100 kg C m⁻² in poorly drained peatlands (Hugelius et al., 2020).

Low Arctic regions

Low Arctic soils are associated with gelic and pergelic MASTs and have continuous near-surface permafrost across more than 90% of the landscape. The treeline which marks the boundary between the Subarctic at the northern edge of forested regions and the Low Arctic is one of the most striking vegetative transitions on Earth. With colder temperatures and decreased precipitation, chemical weathering and biochemical processes are generally less intensive in the Low Arctic than in the Boreal and Subarctic.

- **Climate.** Much of the Low Arctic lies close to the Arctic Ocean and lacks the major climatic distinctions between continental and maritime regions which occur in the Subarctic and Boreal. Summer high temperatures average < 10 to 15 °C, while winter highs are typically -10 to -20 °C. In the Low Arctic, approximately 50 to 75% of the total precipitation occurs as rain during the growing season. Rainfall ranges from 100 to 250 mm year⁻¹, with an additional 50–150 mm of water equivalent as snow. Although precipitation is less than in the Subarctic and Boreal, poorly drained soils dominate the Low Arctic because of ubiquitous near-surface permafrost, which acts as an aquitard leading to a perched water table (Ping et al., 2008a,b). This, together with low annual rates of evapotranspiration, results in predominantly hydric conditions in Low Arctic soils (Tedrow, 1977).
- **Vegetation.** Soils with near-surface permafrost and in lowlands of the Low Arctic are characterized by wet cotton grass and sedge meadows, dominated by *Carex* sp., *Eriophorum* sp., and *Kobresia* sp. Closed to open canopy low (40 cm to >2 m) and dwarf (< 40 cm) shrub tundra, (dominated by *Betula* sp., *Salix* sp. and *Alnus* sp.) occupies much of the remaining landscape (Walker et al., 2005). Vegetative cover in most non-mountainous areas of the Low Arctic is almost continuous, and all soils have surface organic horizons of varying thickness. The major exception to continuity of vegetation cover is bare soil associated with patterned ground features. Fire disturbance does occur naturally and has an effect on ecosystem structure and function, but it is not common and return intervals are much longer than in the Boreal and Subarctic.
- **Soil formation.** Decomposition of organic materials (melanization, humification, mineralization) in Low Arctic soils is significantly slower than at lower latitudes because of colder and wetter conditions. There is widespread accumulation of organic matter (paludification), however due to the absence of acidic needleleaf litter and less intense eluviation, acidification does not occur as deeply as in the Subarctic and Boreal. The Low Arctic includes large areas of acidic, permafrost-affected peatland soils (Fig. 3), but on the coastal plains, fens dominate and the soils can be slightly to moderately alkaline (Ping et al., 2008a,b). On lower hillslope positions (footslopes, toeslopes, and basins), ice content and the accumulation of organic materials can increase markedly, with organic materials reaching several meters in thickness. Permafrost-affected mineral soil materials across much of the landscape show morphological characteristics of saturation and iron reduction (gleization). The near-surface pH values of wet mineral soils in the Low Arctic are 4.5–6.5, but, where there is carbonate-rich loess cover, the soils are alkaline (Ping et al., 2005, 2008a,b).

There are also smaller areas of well-drained soils in the Low Arctic. These soils typically form on upper hillslope positions of broad glacial moraines and develop on sites where there is free drainage to a depth of at least 60 cm or soils shallow to bedrock. They are characterized by oxidized, brownish colors in the upper solum, similar to upland soils of the Subarctic and Boreal. Well-drained soils can also be found in bare spots on patterned ground where the mineral soil is exposed at the surface.

Low Arctic soils are commonly characterized by patterned ground and cryoturbation. Ice wedge polygons and non-sorted circles are characteristic of medium to fine-textured soils and low gradient landscapes of the Low Arctic, while sorted circles and stone stripes are common in soils with considerable coarse fragments or where slope gradients are high. These features result in 'spotted tundra', with co-occurring well vegetated and sparsely vegetated areas (Ping et al., 2008a,b). Bare, exposed areas typically experience high evapotranspiration and accumulate soluble salts, have higher soil pH, and lower organic matter and carbon

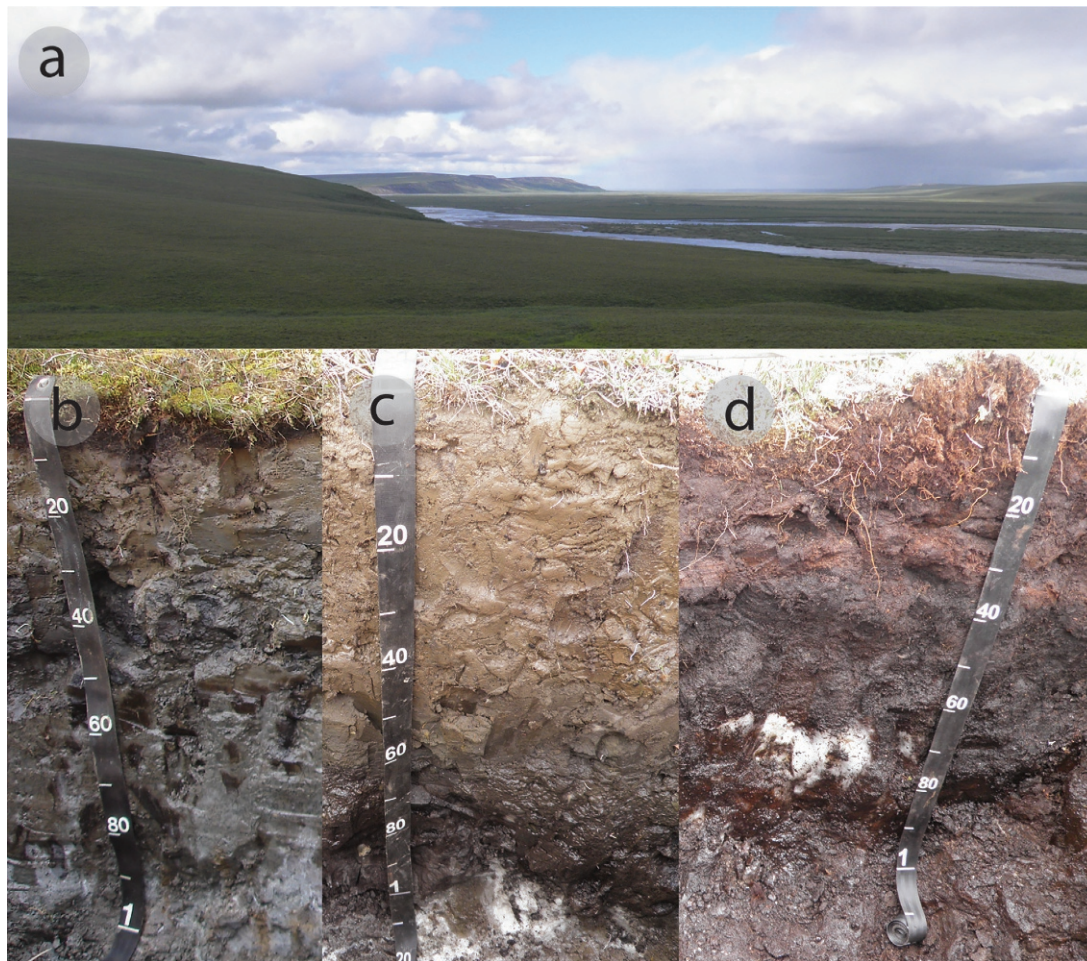


Fig. 3 Representative landscape and soils of the Low Arctic. (A) Sagavanirktok River Valley, Sagwon Hills, northern Alaska, USA (N.A. Jelinski); (B) Aquiturbel (Cryosol) formed in marine sediments near Utqiagvik, Alaska on the Alaskan Coastal Plain (J. D. Jastrow), (C) Haploturbel (Cryosol) formed in loess over glacial till in the Sagwon Hills, Alaska (N.A. Jelinski), (D) Hemistel (Cryosol) formed in organic materials in the Sagwon Hills, Alaska (N.A. Jelinski).

stocks, whereas protected micro depressions in frost cracks are well vegetated, have thick organic matter at the surface, are more acidic, and have larger carbon stocks.

Frost-cracking in fine- to medium-textured sediments is largely responsible for patterned ground formation and the accumulation of extensive ground ice. Ground ice content in the Low Arctic, with the exception of mountainous areas, is generally higher than that of the Subarctic and Boreal. Due to colder and drier conditions and extensive frost cracking, reticulate cryostructures, with vertical cracks or ice nets, ice veins and wedges, as well as suspended (ataxitic) cryostructures and massive ground ice are common (Ping et al., 2008a,b). Cryoturbation processes are dominated by cracking and warping, hydrocryostatic and cryostatic pressures, diapirism, and differential frost heave.

- **Classification and Carbon stocks.** Gelisols (Cryosols) cover 90% of the ice-free land area in the Low Arctic. Most remaining soils of the Low Arctic are Entisols (Regosols), Inceptisols (Cambisols), or Mollisols (Chernozems). These minor soils typically occur on floodplains, steep, rocky slopes, or in coarse textured soils on colluvium where near-surface permafrost is not present, despite cold annual average temperatures. Extensive cryoturbation as well as burial due to eolian or alluvial activity has led to the development of large carbon stocks in mineral soils of the coastal plains and river deltas of the Low Arctic. The SOC stocks in Low Arctic mineral soils to 1-m depth typically range from 10 kg C m^{-2} to $>80 \text{ kg C m}^{-2}$ (Ping et al., 2008a,b), and the soils of permafrost-affected peatlands in the Low Arctic can exceed $>100 \text{ kg C m}^{-2}$.

High Arctic soils

The High Arctic cold regions (Arctic bioclimate zones A, B, and C, Walker et al., 2005) are associated with pergelic and hypergelic MASTs, and near-surface permafrost is continuous across the landscape (Ping et al., 2008a,b). Naturalists have long had a variety of names for the High Arctic sectors, such as 'high polar,' 'polar desert,' 'cold steppe,' 'arctic wastes,' 'rock desert,' 'polar tundra,' 'barren-ground tundra,' 'fell field' (desert-like), 'desert glacial,' 'nival zone,' and 'dry periglacial zone.' The High Arctic can be broadly differentiated from the Low Arctic in that southern air masses almost never reach the High Arctic, and shrubs, where they do

occur, are prostrate and creeping. In Arctic bioclimate subzone A (often identified as the 'Arctic polar desert'), woody species do not occur at all.

- **Climate.** In High Arctic subzones B and C, average high temperatures in summer are typically between 5° and 10 °C, and in winter –10 to –30 °C. Annual average rainfall ranges from <100 to 150 mm, with an additional 50–200 mm of snow water equivalent. In the High Arctic subzone A, average high temperatures in summer are rarely above 5 °C and in winter average highs are typically between 20 and 30 °C. Annual average rainfall in subzone A is <50 mm, with an additional 50–200 mm of snow water equivalent. Approximately 25–75% of the total annual precipitation falls as rain in subzones B and C, whereas in subzone A less than 30% of total annual precipitation falls as rain during the growing season. Low precipitation in the High Arctic leads to several desert-like features. However, there are major differences between High Arctic environments and true low latitude deserts; importantly permafrost results in sub-irrigation of soils during the growing season and the hyper cold temperatures and short growing season results in low evapotranspiration. Therefore, evapotranspiration and precipitation are more closely balanced in the High Arctic than in true deserts (Bliss, 1977).
- **Vegetation.** The High Arctic regions have a discontinuous, sparse vegetative cover (Fig. 4). Vascular plants do not compete for open space as they do in the Low Arctic zone. Vegetative cover decreases polewards in the High Arctic. In subzones B and C, the vegetation of mesic sites is dominated by hemiprostrate and creeping dwarf shrubs (*Salix* sp. L. and *Dryas* sp. L., < 40 cm in height), but cover is interrupted frequently by bare soils associated with patterned ground. There are some wetland plant communities in subzones B and C dominated by the tussock sedges *Carex*, *Kobresia*, and *Eriophorum* sp. In subzone A, there are no woody species and most plants have a compact, 'cushion-type' growth form (Walker et al., 2005). The plant cover consists mainly of lichens, mosses, liverworts, and a scattering of Arctic–Alpine herbaceous plants; sedges are rare, and wetlands, where they do exist, typically lack organic layers. Ecological and pedological differences between well drained and poorly drained sites

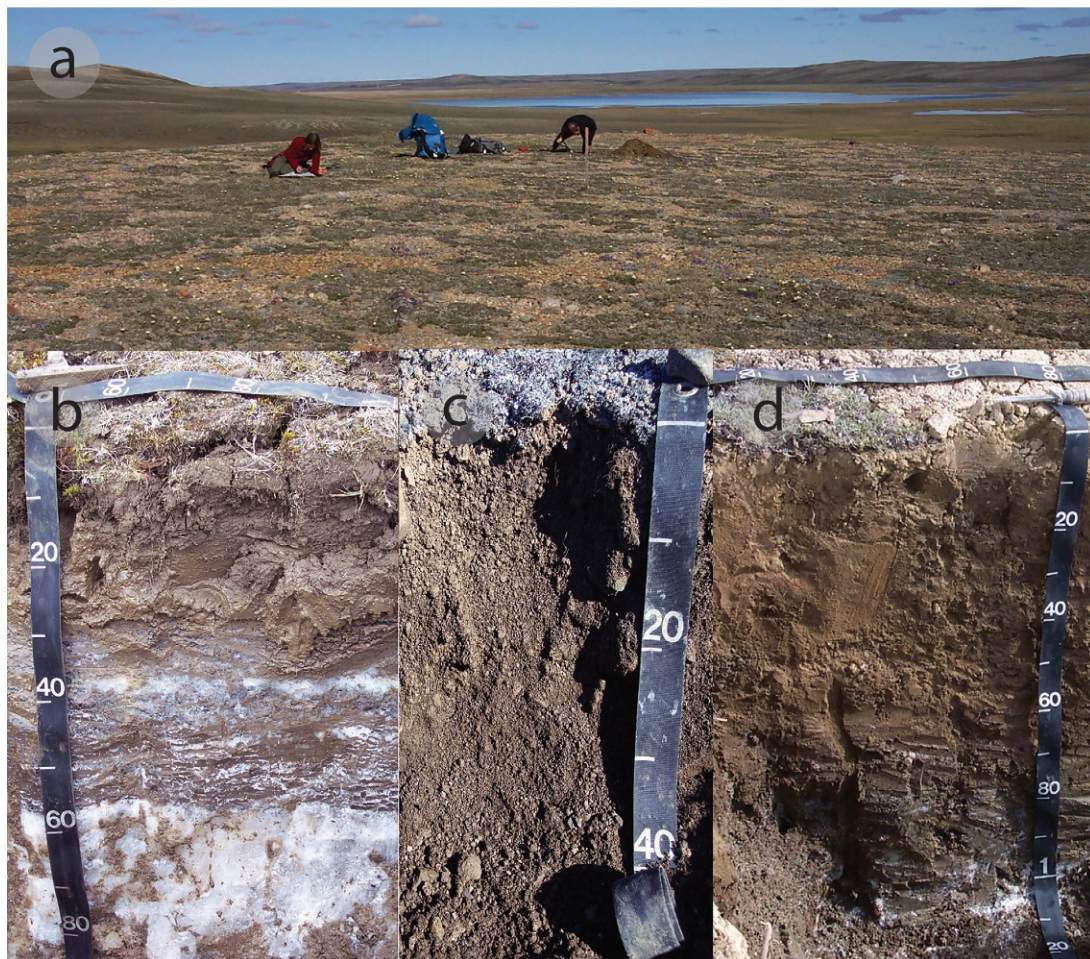


Fig. 4 Representative landscape and soils of the High Arctic. (A) Banks Island, Canada (Canadian Arctic Archipelago) landscape and land cover (C.-L. Ping); (B) Poorly drained Glacial Aquiturbel (Cryosol) under hummocks, Isachsen, Ellef Ringnes Island, Canadian Arctic Archipelago (C.-L. Ping), (C) Well-drained Orthel (Cryosol) on a ridgetop under lichen barrens, Isachsen, Ellef Ringnes Island, Canadian Arctic Archipelago (C.-L. Ping), (D) Orthel (Cryosol) in Mould Bay, Prince Patrick Island, Canadian Arctic Archipelago (C.-L. Ping).

are minimal, and there are large areas of bare, exposed soil where vegetation is often confined to cracks and depressions in patterned ground or in areas in between hummocks.

- **Soil formation.** Well-drained soils of the High Arctic have several properties associated with deserts such as saline conditions, desert pavements, and wind-abraded rocks. Secondary carbonates and calcite pendants are common. Desert pavements (formed from deflation of fine particles at the surface in dry and windy conditions) usually occur where coarse fragments are abundant in the substrate. In some well-drained soils, there may be a thin oxidized zone in the upper profile. The pH of well drained High Arctic soils varies from ~4.4 to 7.9. Salt crusts of Thenardite (Na_2SO_4) and other minerals can form on some well-drained to xeric sites (Tedrow, 1977). Horizon differentiation is minor other than that induced by the distribution of iron coatings on soil particles in oxidized zones (Fig. 4). The organic matter content of well-drained High Arctic soils is generally low (often <1% SOC), but does increase where vascular plant cover is more continuous. Unlike the Low Arctic and Subarctic/Boreal, fire disturbance is non-existent.

In lowland soils of the High Arctic, organic matter contents in surface horizons may reach 5%, with greater carbon accumulations in small poorly drained areas with peat at the surface. These soils typically lack carbonates in the upper profile and may have slightly acidic reactions in surface horizons. Hydric soils are also less prevalent (Bliss, 1977) because biological activity is slow (due to cold temperatures) and organic matter contents are low – therefore iron reduction and gleization take much longer to occur. Patterned ground in the High Arctic is largely related to frost-cracking and rock sorting because there is not enough moisture for large amounts of ground ice, except in isolated wetland conditions, where ice wedge polygons can form (Bliss, 1977).

- **Classification and Carbon stocks.** Over 99% of the ice-free High Arctic landscape is underlain by near-surface permafrost and nearly all soils are classified as Gelisols (Cryosols). Controversy remains as to whether carbon stocks have been over- or under-estimated in High Arctic soils, but it is clear that carbon stocks are lower in the High Arctic than in the Low Arctic, Subarctic, and Boreal. Carbon stocks to 1-m depth, regardless of topographic position and hydrologic setting, rarely exceed 25 kg C m⁻² (Ping et al., 2008a,b).

Soils of the southern circumpolar region

The soils and landscapes of the Southern Circumpolar Region (SCR) are far from a mirror image of the NCR. While the NCR comprises an oceanic core (the Arctic Ocean) surrounded by islands and continents, the SCR has a continental core (Antarctica) with outlying islands surrounded by open ocean. These differences make the soils of the SCR unique in several ways. First, there is much less ice-free land at latitudes greater than 45° in the SCR than in the NCR (<50,000 km⁻² and >17 million km⁻², respectively), and in addition, of the 2 to 3% of continental Antarctica that is free of ice and snow, only a small percentage is underlain by unconsolidated materials; the remainder of the ice-free land consists of bedrock and boulder fields. Second, the continentality of the SCR core leads to much colder temperatures than in the NCR (Tedrow, 1991). Third, the soils of the SCR are predominantly maritime influenced, because of the large system of islands and coastal areas. One of the most important effects of this is the major role of colony-forming sea birds that result in unique soil properties. Although such colonies exist in the NCR (Otero et al., 2018), the effects are more intense in the SCR because the ratio of bird populations to ice-free land area in the southern hemisphere is much higher (Otero et al., 2018). The effects that these colonies have on SCR soils are tremendous, and result in unique soil formation processes (collectively known as ornithogenesis), which can lead to increased carbon and phosphorus stocks (Rodríguez et al., 2021), and generally greater acidity and chemical weathering.

The cold soils of the SCR can be broadly divided into four zones: the Subantarctic Tundra (the northern Antarctic peninsula, south Shetland islands, all of the Subantarctic islands south of the Antarctic convergence), and Subantarctic Forest (islands south of the subtropical front and the Antarctic convergence and maritime southern coastal Patagonia (the Chilean Fjords south of 46°S), Tierra del Fuego), the Polar Desert (most of maritime western Antarctica and the central portion of the Antarctic Peninsula between 65 and 68°S), and the Cold Desert (the ice free portions of Antarctica south of ~68°), (Tedrow, 1977; Bockheim and Ugolini, 1990) (Fig. 5).

Soil parent materials. Almost all of ice-free Antarctica and most of the SCR was glaciated during the Pleistocene, except for a few areas which remained ice-free such as the Friis Hills of Antarctica. Most of western Patagonia, western Tierra del Fuego and many of the Subantarctic islands were completely glaciated during the early- to mid-Pleistocene, but glaciation was more limited at the last glacial maximum (Rainsley, 2019). The Subantarctic islands are largely volcanic, but have been modified by ice sheets during previous glacial periods. The South Georgia Islands were completely covered by ice during the last glacial maximum (Clapperton et al., 1989), whereas the Campbell, Falkland and Auckland Islands, for example, may have experienced only limited glacial cover in the late Pleistocene. Soil parent materials are therefore dominated by glacial deposits (till, outwash, glaciolacustrine materials), with eolian sands, colluvium and alluvium occupying most of the rest of the landscape. Unlike the NCR, loess is largely absent from the SCR.

Soils of the Subantarctic Forest and Subantarctic Tundra

The Subantarctic Forest and Subantarctic Tundra belong to a maritime system, consisting solely of coastal regions and islands. The Subantarctic Forest includes much of Tierra del Fuego and the western Patagonian fjords south of 45° S. The Subantarctic Tundra includes islands south of the Antarctic convergence. In latitudes 45 to 60° S of the Subantarctic zone, permafrost is largely non-existent, and average temperatures exhibit little annual fluctuation.

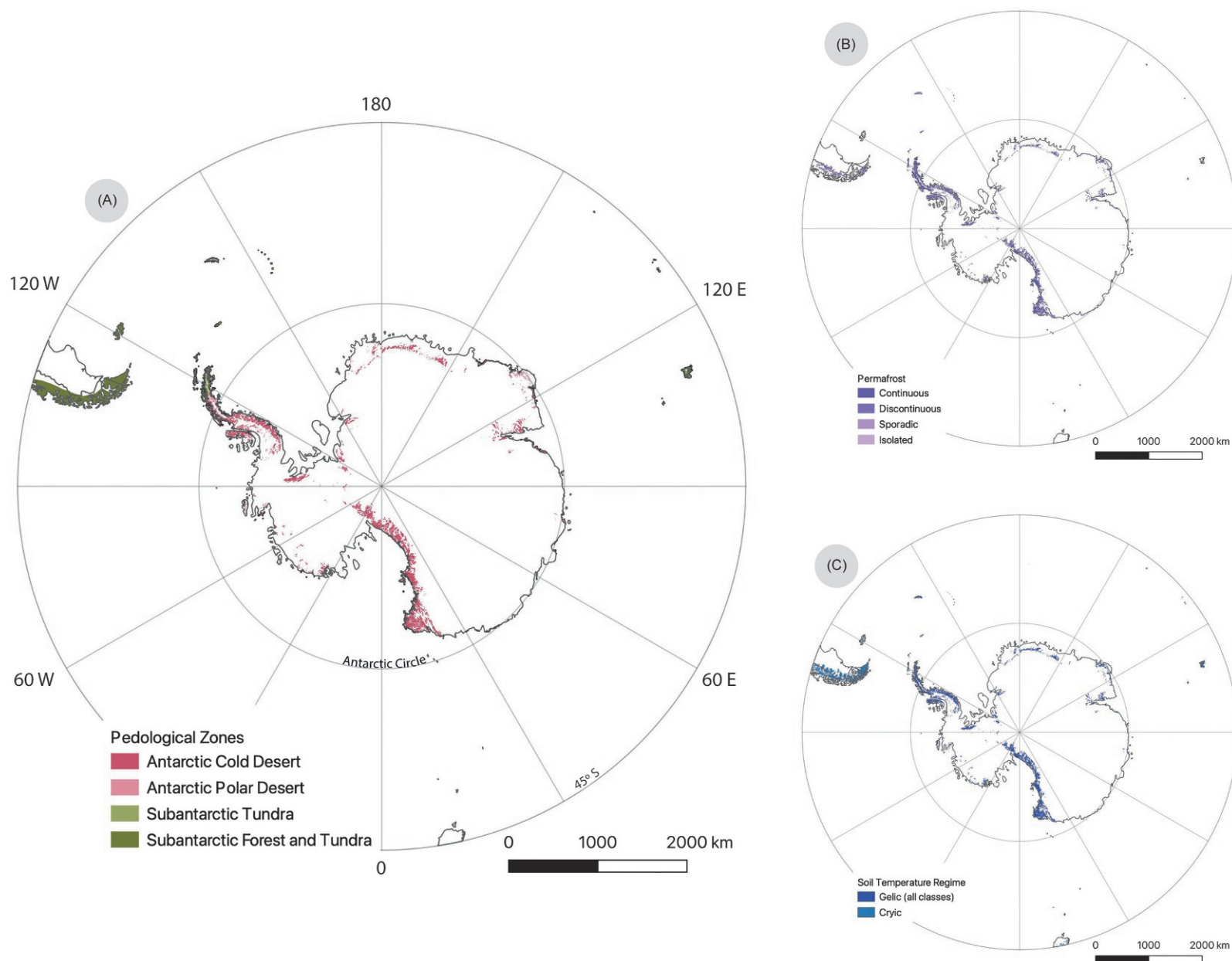


Fig. 5 Pedological zonation, permafrost extent, and soil temperature regimes of the Southern Circumpolar region. (A) Pedological zones are re-interpreted from Tedrow (1977) and Bockheim and Ugolini (1990). The northernmost Subantarctic Forest boundary here includes the maritime portions of western Patagonia and Tierra del Fuego south of 45° latitude. Ice-free areas only are shown and the distribution of ice-free areas is taken from Obu et al. (2019a). (B) Permafrost extent in ice-free areas is derived from (Obu et al., 2019a), assuming permafrost probabilities of 0–9% are equivalent to isolated permafrost, 10–29% are equivalent to sporadic permafrost, 30–89% are equivalent to discontinuous permafrost, and >90% are equivalent to continuous permafrost. (C) Soil temperature regimes for southern South America, New Zealand, and major Subantarctic islands are derived from USDA-NRCS. (2016), while soil temperature regimes for the Antarctic continent are derived from mean annual ground temperature (MAGT) predictions (temperature at the top of the permafrost) from Obu et al. (2019a).

- *Climate.* Average high temperatures are often above freezing for the entire year, and precipitation often exceeds 1000 mm per year. Rain makes up >75% to 90% of total precipitation. In the higher latitudes of the Subantarctic Tundra (South Shetland Islands and northern Antarctic Peninsula, 60 to 65°S), conditions are transitional to the Polar Desert zone. Permafrost becomes sporadic to discontinuous, and annual average temperatures show larger annual variations. Annual precipitation is lower in the southern Subantarctic Tundra, with rain accounting for a smaller proportion of the total.
- *Vegetation.* Subantarctic Forest is limited largely to coastal southern Patagonia, Tierra del Fuego, and the Auckland Islands in New Zealand. The vegetation is dominated by *Nothofagus* sp. in South America and *Metrosideros* sp. in the Aucklands. Most of the Subantarctic region is dominated by tundra, characterized by tussock grasses (*Poa foliosa*) and megaherbs in mesic sites. Natural fire disturbance in the Subantarctic is rare. Vegetated wetlands in the Subantarctic are dominated by rushes, sedges and cryptogams. In the southern latitudes of the Subantarctic Tundra (northern Antarctic peninsula, South Georgia, South Orkney and the Shetland Islands), which are transitional to the Polar Desert, vegetative cover becomes sparse, and consists largely of mosses, lichens, and algae.
- *Soil formation.* The absence of permafrost, relatively mild annual temperatures, and nearly constant precipitation results in excess moisture in the Subantarctic, leading to leaching at well-drained and mesic sites and the accumulation of organic materials in low gradient or poorly drained areas (Fig. 6). Soils are acidic, the accumulation of organic materials is common, and podzolization is common in well-drained soils (Wilson, 2001). Carbonates are largely absent from Subantarctic soils due to intensive leaching processes.
- *Vegetation.* Vegetation cover and biomass production decrease from the northern to southern Subantarctic, while the relative importance of ornithogenesis on soil properties increases. Soils under current or former rookeries tend to be more acidic (pH 3.5–4.5), have higher carbon stocks, and P concentrations up to 10 times greater than adjacent, non-ornithogenic soils. With more acidic conditions and greater carbon and nutrient availability, vegetative growth and microbial activity are elevated in soils under rookeries and biochemical weathering is more rapid than in surrounding areas, leading to changes in mineralogy and chemical properties.

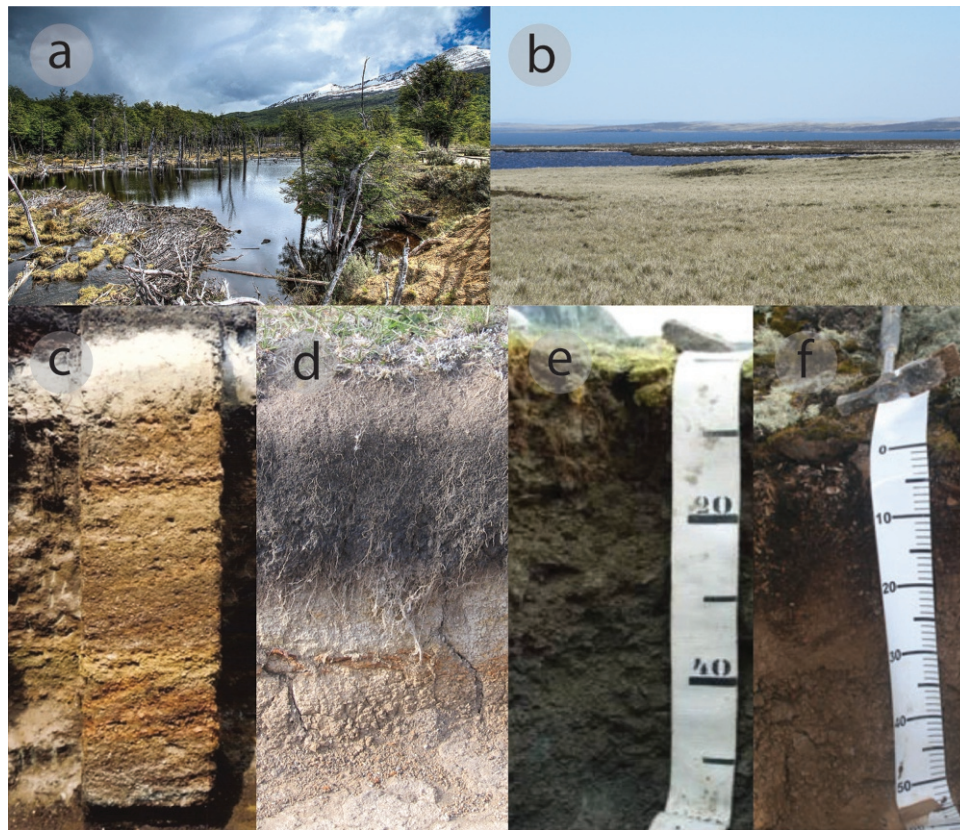


Fig. 6 Representative landscapes and soils of the Subantarctic. (A) Subantarctic Forest in Tierra del Fuego National Park, Chile (Courtesy of A. Dirkse); (B) Acidic grasslands and dwarf shrub heath in the Falkland Islands (Courtesy of S. Carter, South Atlantic Environmental Research Institute); (C) Spodosol (Podzol) in Tierra del Fuego Province, Argentina (Courtesy of F. Pereyra and P. Bouza); (D) Spodosol (Podzol) from the Falkland Islands (R. Burton); (E) Soil with incipient ornithogenic influence due to Skua nesting from King George Island (Courtesy of F. Simas, C. Schaefer, M. Francelino, and R. Michel); (F) Melanized soil on uplifted marine terrace on the Barton Peninsula, King George Island (Courtesy of F. Simas, C. Schaefer, M. Francelino, and R. Michel).

- *Classification and carbon stocks.* Permafrost is largely absent from the northern Subantarctic regions, which are dominated by Histosols (Histosols), Spodosols (Podzols), and Inceptisols (Cambisols). Cold Mollisols (Chernozems), Inceptisols (Cambisols) and some Andisols (Andosols) are also regionally common, for example, in Tierra del Fuego (Rodríguez et al., 2019). Some of the Subantarctic islands may have large areas of Histosols (Histosols) (as thick organic deposits in poorly drained areas and across uplands as blanket peat), and other mineral soils without permafrost. In the southern Subantarctic, permafrost is largely absent at elevations less than 20 m asl, but rapidly becomes sporadic to discontinuous at elevations greater than 50 m asl. Consequently, the southern Subantarctic is a mixture of cold Entisols (Regosols) and Inceptisols (Cambisols) without permafrost at lower elevations and Gelisols (Cryosols) with near-surface permafrost at higher elevations. Data on soil carbon stocks in the Subantarctic are sparse, but estimates suggest that under well vegetated sites in the northern Subantarctic, SOC stocks to 1 m can exceed 30 kg m⁻² (Smith and Karlsson, 2017), and under deep peatlands it may be >100 kg m⁻². In the southern Subantarctic, where vegetation is more sporadic and peat accumulation is rare or highly localized, carbon stocks to 1 m in non-ornithogenic soils are much smaller, often <3 kg C m⁻² (Simas et al., 2007). The SOC stocks in adjacent ornithogenic soils to 1-m depth can be 5–10 kg C m⁻².

Soils of the Antarctic Polar Desert

The Polar Desert zone includes the Antarctic Peninsula (excluding the transitional south Shetland and south Orkney islands) and the ice-free regions of coastal west Antarctica (Enderby land, MacRobertson land, and Wilkes land). Landforms include low hills and moraines with thin to moderately thick glacial drift over bedrock. At elevations less than 300 m, bare, exposed coarse-textured soils and soils shallow to bedrock can be warm enough during the summer to have active layers deeper than 1 m (see Mergelov et al., 2015 Chapter 5 in Bockheim, 2015a,b). At higher elevations, or in most other situations, near-surface permafrost is continuous across the landscape.

- *Climate.* High temperatures in summer (December to February) may reach 5 °C, with mean annual air temperatures between –10 °C and –20 °C. Precipitation is limited almost completely to snow, with total precipitation ranging from 100 to 250 mm year⁻¹ of water equivalent.
- *Vegetation.* Vegetative cover is sporadic, consisting solely of cryptograms (lichens, mosses), and algae. It is most extensive in areas protected from the wind such as drainage pathways and hollows, and under sea-bird colonies. The most extensive vegetation cover occurs as moss cushions, which are closely related to the presence of rookeries. Ornithogenesis is ecologically important in the Polar Desert zone and has a major effect on soil properties at local scales, however, rookeries are less widespread in the Polar Desert zone than in the Subantarctic (Otero et al., 2018).
- *Soil formation.* Most soils of the Polar Desert zone exhibit weakly to well-developed desert pavements (Tedrow, 1977), which are largely absent from most soils in the Subantarctic zones farther north. Continental Antarctica is characterized by high, sustained katabatic winds throughout the year from the action of cold air flowing down from the interior Antarctic plateau. Desert pavements develop from the combination of dry conditions and the action of wind deflation, where wind picks up and removes fine particles, leaving behind a surface covered by coarse fragments. This surface, dominated by coarse fragments protects the soil, preventing further deflation of the materials below. Unconsolidated materials below the desert pavement retain a reasonable amount of moisture during the growing season because of snow melt and thawing of ice in the active layer (see Zazovskaya et al., 2015 Chapter 3 in Bockheim, 2015b).

The lack of vascular flora and sporadic vegetation results in low soil organic matter concentrations in most Polar Desert soils and limited to the upper portions of the soil (see Mergelov et al., 2015 Chapter 5 in Bockheim, 2015a, b). The SOC concentrations are often elevated only at depths less than 5 cm from the soil surface and are commonly <1%, but can range from 1 to 5% (see Zazovskaya et al., 2015 Chapter 3 in Bockheim, 2015a, b). In localized depressions or under more continuous moss cover, SOC contents in the top 5 cm may reach 5–15%, but decline rapidly to <0.5% below 5-cm depth, except under rookeries where SOC contents may remain elevated to depths of 30 cm or more. Shallow peat accumulation (<20 cm thick) can occur in isolated, protected drainage pathways that are better vegetated and accumulate snowmelt during the summer (see Dolgikh et al., 2015 Chapter 4 in Bockheim, 2015a, b).

Soils are sub-irrigated and can be moist in the summer due to snow melt and protection from the desert pavement. There is little to no eluviation, illuviation, and downward movement of soil constituents in the Polar Desert. Thin weakly to strongly oxidized subsoil horizons can occur, but these are typically limited to the upper 20 cm (see Dolgikh et al., 2015 Chapter 4 in Bockheim, 2015a, b). Secondary carbonates occur locally on the underside of coarse fragments, and salts more soluble than carbonates mainly from marine aerosols exhibit patterns consistent with distance from the coast and may occur as salt crusts or efflorescence. Redoximorphic features rarely form, even under frequently saturated conditions because of cold temperatures and low SOC concentrations.

- *Classification and carbon stocks.* Nearly all soils are classified as Gelisols (Cryosols) in the Polar Desert zone. At elevations less than 300 m asl with coarse fragments and little vegetation on exposed areas, the active layer may be deeper than 1 m and soils may be classified as Entisols (Regosols). The SOC stocks are generally low; most soils have <2 kg m⁻² to 1-m depth, while isolated soils with peat at the surface as well as ornithogenic soils may have carbon stocks of 10–20 kg C m⁻² (see Blume and Bölker, Chapter 6 in Bockheim, 2015a, b).

Soils of the Antarctic Cold Desert

Major ice-free areas in the Cold Desert zone include Ellsworth Land, Victoria Land, and the Transantarctic Mountains. Permafrost is continuous, but most surface deposits are not ice cemented because of low moisture conditions which produces dry permafrost (Bockheim, 1997). Active layers are generally very shallow, ranging from <10 cm inland to >90 cm near the coast (Kimble et al., 2021).

- **Climate.** Climatic conditions in the Cold Desert zone are some of the harshest on Earth with total precipitation in water equivalent of 50 to 150 mm year⁻¹, with all precipitation occurring as snow. Summer high temperatures rarely exceed 0 °C, with mean annual air temperatures ranging from –17 to –24 °C (Kimble et al., 2021).
- **Vegetation.** Vegetation cover is sporadic, and where it occurs it is largely algae and lichens. One interesting aspect of Cold Desert zone soils is that there is almost no effective organic component except for local accumulations of guano under penguin rookeries (Tedrow, 1991) (Fig. 7). Orthogenesis is less important in the Cold Desert zone than in the more northerly Polar Desert and Subantarctic zones because rookeries are fewer.
- **Soil formation.** Accumulation of calcite, gypsum, magnesium sulfate, and other soluble minerals is widespread in soils of the Cold Desert. Highly soluble forms of calcium, magnesium, sodium, potassium, chloride, sulfate, and nitrate have also been reported. Nearly all soils of the Cold Desert are alkaline, with pH values 7–10 (Campbell and Claridge, 1987). Closed basins, depressions, and other sites commonly have saline soils and accumulations of salt encrustations. Horizonation is somewhat blurred, with a lack of structural aggregates (Fig. 7). The 'active' layer is less than 1-m thick, but because of the ultra-dry, loose condition, depth to the frost table is often only approximated. There is evidence that there is ionic migration in the soil, even when it is dry and frozen (Campbell and Claridge, 1987). Accumulation of salts is common in the soil in bedrock and boulders.

Physical weathering in the Cold Desert zone is intense. Boulders have shells of exfoliation that can be pried off. Wind polishing of rocks is intense with the abrasion caused by blowing ice particles because it may be as hard as quartz at very low temperatures. Desert varnish is also evident in many places, especially on older land surfaces. The temperature gradient in rocks

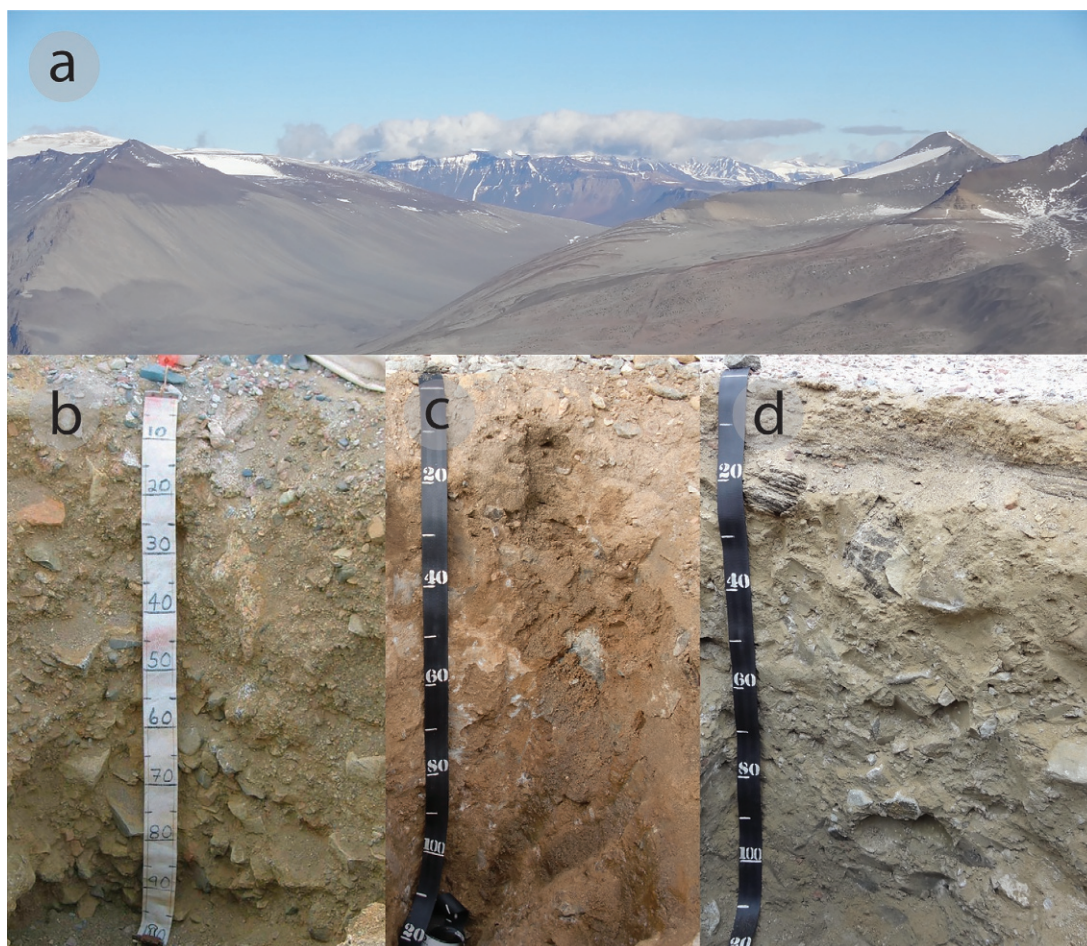


Fig. 7 Representative landscape and soils from the Antarctic Cold Desert. (A) Bull Pass, Wright Valley, McMurdo Dry Valleys, Antarctica (C.A. Seybold); (B) Anhyorthel (Cryosol), McMurdo Dry Valleys (M.R. Balks), (C) Anhyorthel (Cryosol), McMurdo Dry Valleys (C.A. Seybold), (D) Typical Anhyorthel, Wright Valley, Antarctica (C.A. Seybold).

on sunny days results in both physical and chemical weathering. Garnets split, feldspars weather, and clay substances accumulate in the fractured areas. Although chemical weathering is less important than physical weathering in this zone, it can still occur over longer timeframes. Increasing oxidation of surface deposits of various ages can occur even in some of the harshest Cold Desert environments (Bockheim, 1997).

- **Classification and carbon stocks.** All soils in the Cold Desert zone are classified as Gelisols (Cryosols). Many soils, due to the dry permafrost conditions, are Anhyorthels (Bockheim, 2015a,b; Kimble et al., 2021). Patterned ground is limited largely to sorted circles and other sorted coarse fragment forms on the soil surface due to the lack of ground ice. Cryoturbation can occur, although it is limited to near-surface horizons. Organic materials do not occur because SOC concentrations are very low, often <1%. Research on SOC stocks in the Cold Desert zone is lacking, however, it is likely that SOC stocks to 1 m do not exceed 1–2 kg m⁻² under most conditions (Bockheim, 2015a,b).

Alpine and Plateau soils

The broad climatic influence of latitude on cold regions declines rapidly beyond the Boreal and Subantarctic regions and requires gains in elevation to achieve cold conditions. As an approximation, every 10 degrees movement south in the northern hemisphere (or north in the southern hemisphere) in latitude requires an additional 1000 m in elevation to maintain similar climate conditions (Bockheim, 2015a,b). Therefore, cold region soils between the 45th parallels are associated with mountainous and alpine environments. These cold areas in otherwise temperate and tropical regions are dominated by the high altitude Qinghai–Tibetan Plateau (QTP, often called the third pole”), a large alpine plateau in the rain shadow of the Himalaya mountains. The remainder of temperate and tropical cold region soils lies predominantly in the US Rockies, Hindu Kush and Himalaya, Andes, Alps and Caucasus mountains, and in a few isolated tropical high mountains in Hawaii, East Africa, and Indonesia.

In these high relief landscapes in otherwise warm latitudes, rapid landscape and vegetation change results in complex and often abrupt pedological and ecological boundaries. Slope and aspect are extremely important in determining solar radiation, soil temperature, and effective soil moisture. With the exception of relatively low gradient landscapes in some areas of the QTP, alpine cold region soils are generally geomorphically active, with high rates of soil erosion and colluvial movement. Parent materials in alpine regions are dominated by gravitationally re-worked glacial sediments and colluvium. Because these soils are in otherwise warm regions of the globe, they are also some of the most rapidly changing areas on Earth.

- **Climate.** The climate of alpine mountainous cold soil regions at latitudes less than 45° N or 45° S is very spatially variable. In many subtropical and tropical cold mountainous regions, diurnal temperature variations exceed those of annual variations in magnitude. Diurnal temperature variation can often exceed 20 to 30 °C, whereas in high latitude cold regions average diurnal temperature fluctuations are more commonly <20 °C. Permafrost in alpine and cold plateau regions is also spatially variable, ranging from non-existent to continuous (Obu et al., 2019b). Although mean annual soil temperatures are often below 0 °C, active layers can be deep due to strong diurnal temperature fluctuations.

Cold plateau soils exist below the snow-line at high elevations on the QTP, the Andes, and mountains of Central Asia. Annual evaporation can exceed 3000 mm in some areas, and the climate can be extremely dry, with aridity indices (the ratio of potential evaporation to precipitation) exceeding 20–40 (Ping et al., 2008a,b). Although mean annual ground temperature (MAGT) in these areas can be less than –3 °C (–2.7 to –1.2 °C at the permafrost table), active layers range from 1 m to more than 3 m thick because of strong diurnal temperature fluctuations, dry conditions, and coarse-textured soils.

- **Vegetation.** Vegetation in alpine and cold plateau regions is generally continuous, except at very high altitudes, where only sparse, discontinuous moss and lichen cover may persist. Forested alpine soils occur in areas with excess moisture (due to regional climate) or in favorable local microclimates (for example on northeast facing slopes in the northern hemisphere), while alpine grasslands typically exist in drier climates. On the QTP, three major vegetation types dominate in order of increasing aridity: alpine meadow, alpine steppe, and alpine desert (Wang et al., 2016) (Fig. 8). The alpine meadows predominate in the eastern portion of the QTP at lower elevations and are characterized by widespread sedge meadows (dominated by *Kobresia* sp.) with minor shrub and forb components. At higher elevations, in the central and western portions of the QTP, alpine steppe dominates the landscape. Vegetation cover in the alpine steppe is less continuous than in the alpine meadows and is generally dominated by cold and dry grasslands (dominated by *Stipa* sp., *Koeleria* sp., and *Festuca* sp.) with minor forb components. Finally, alpine desert occurs at the highest altitudes in the northwestern QTP. These are characterized by sparse vegetation cover, dominated by winterfat (*Ceratoides latens*), thistle (*Ajania fruticulosa*), mustard (*Christolea crassifolia*) and needle grass (*Stipa* sp.).
- **Soil formation.** Cold soils formed under high altitude forests have morphological characteristics typical of excess moisture, such as podzolization, argilluviation and decarbonation, and are moderately acidic, with pH values ranging from 4.8–6.0. In soils under alpine grasslands, podzolization and argilluviation are less common and melanization is more prevalent, resulting in thick, dark colored surface horizons (Munroe, 2012) (Fig. 9).

The alpine meadow zone lies below the high mountain zone and occupies mountain slopes and some lowlands in the alpine zone. The climate is cold, semiarid to subhumid; with MAST of –0.2 to –6.0°C, and MAP of 300–400 mm, with 85–90% falling between June and September. The annual potential evaporation can exceed 2,000 mm and thus the aridity index is often more than 10. The active layer is generally more than 1.5-m thick. In addition to seasonal freeze–thaw, the area also experiences strong diurnal freeze–thaw cycles. The well-drained soils have a dense, black surface horizon, with many fine grass roots overlying a strong brown

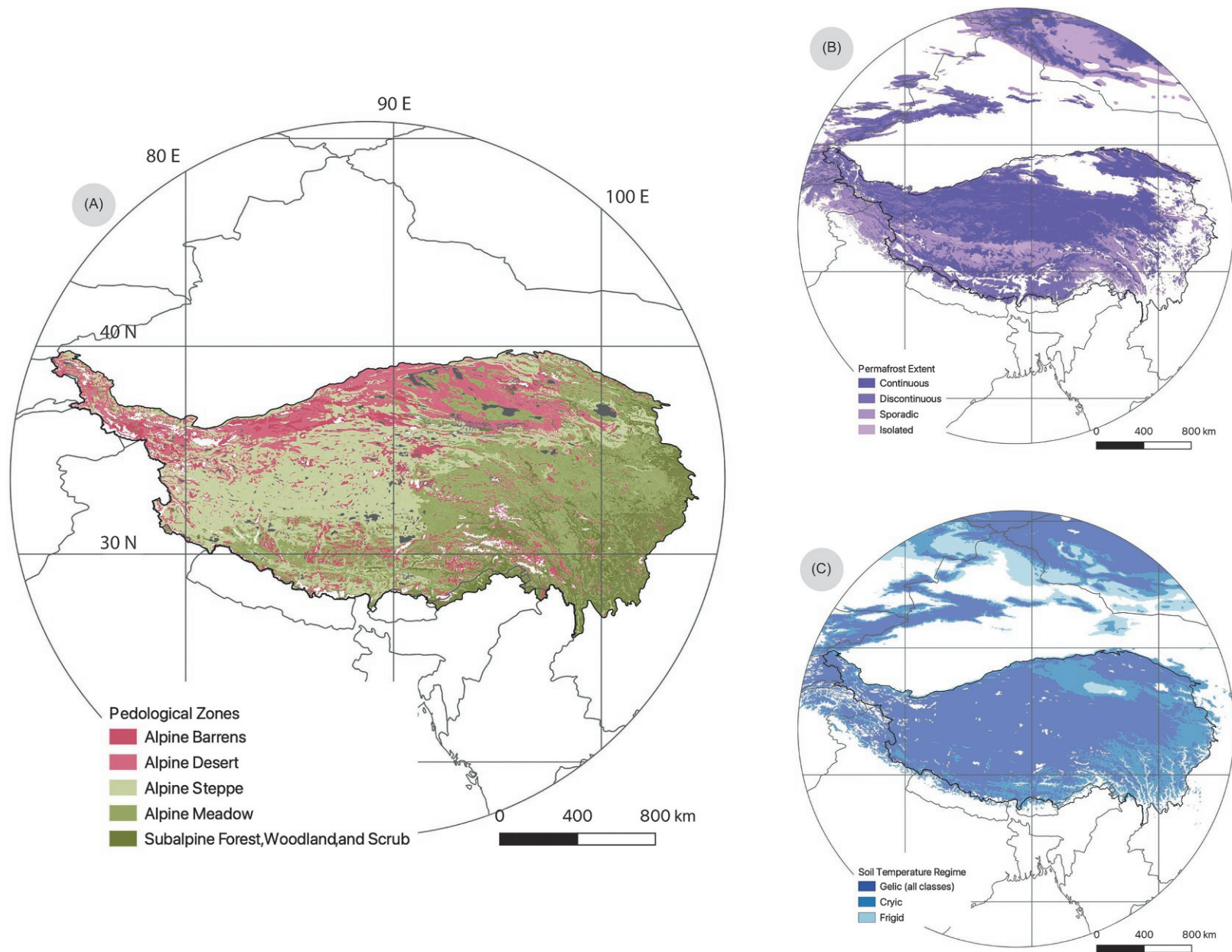


Fig. 8 Pedological zonation, permafrost extent, and soil temperature regimes of the Qinghai-Tibet Plateau (QTP). The QTP boundary was derived from Zhang et al. (2014). (A) Pedological zones are based on broad vegetation groupings interpreted from the 1:1 million vegetation map of China (Hou, 2019). (B) Permafrost extent in ice-free areas is derived from Obu et al. (2018), assuming permafrost probabilities of 0–9% are equivalent to isolated permafrost, 10–29% are equivalent to sporadic permafrost, 30–89% are equivalent to discontinuous permafrost, and >90% are equivalent to continuous permafrost overlain on generalized permafrost categories from Brown et al. (2002) to display more localized variation. (C) Soil temperature regimes are derived from USDA-NRCS (2016).

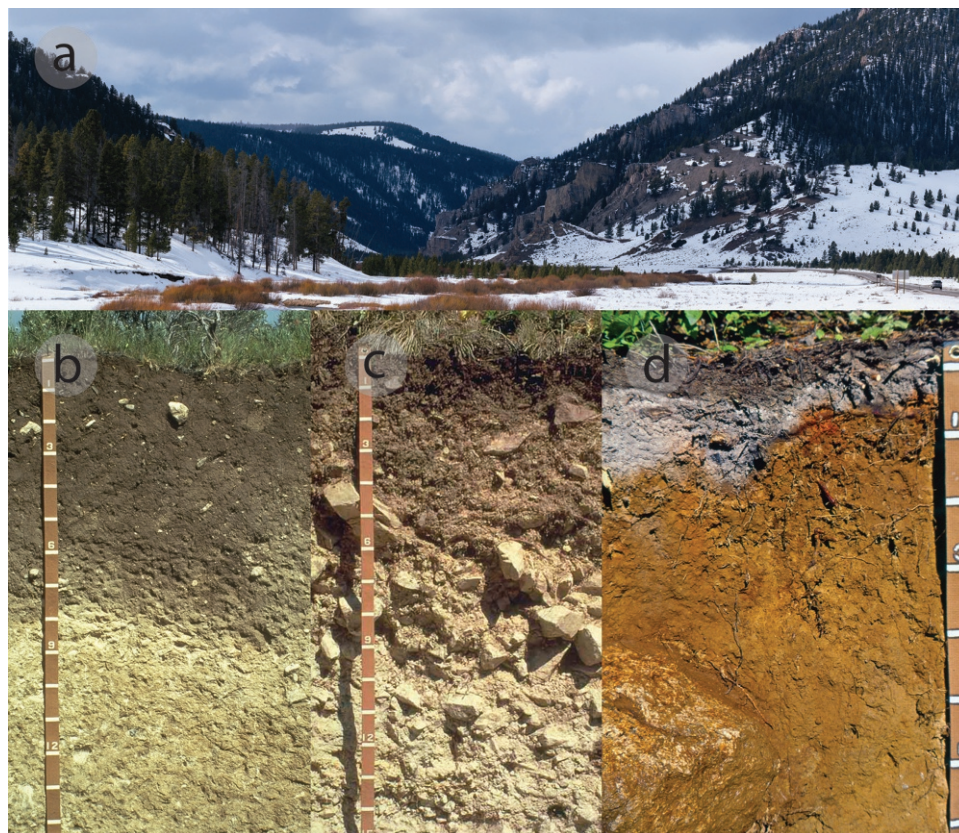


Fig. 9 Representative landscape and soils from low latitude (between 45°N and 45°S) cold alpine regions (excluding the Qinghai–Tibetan Plateau). (A) The headwaters of the Gallatin River, Bozeman, Montana, USA (USDA–Natural Resources Conservation Service. J. Gibson photographer); (B) Pachic Argicryoll (Chernozem), Idaho, USA (Image courtesy of University of Idaho); (C) Xeric Dystricrypt (Cambisol), Idaho, USA (Image courtesy of University of Idaho); (D) Andic Haplocryod (Podzol), Idaho, USA (P. McDaniel).

subsoil, often with an accumulation of carbonates, Fe, manganese oxides, and humus (to a lesser extent) coatings on soil particles, and well-developed reticular and fine, sub-angular blocky structures (Fig. 10). Strong freeze–thaw cycles result in frost-heaved grass mounds, frost cracks, and gelifluction lobes. In lowlands and depressions of the alpine meadow zone, where drainage is limited, gleization features are prominent in the sub-surface horizons. Organic soils form in depressions with restricted drainage.

In the alpine steppe and alpine desert regions, leaching is less extensive due to the arid conditions. In the high mountains and plateau, such as the northwestern Qinghai–Tibet Plateau the aridity index exceeds 10 and can be 40 in some areas. Thus these soils are alkaline and carbonates, gypsum, and soluble salts often accumulate on the soil surface. Organic matter accumulation in alpine steppe and alpine desert soils is less than in the alpine meadow because of colder temperatures and drier conditions, however the grass-dominated ecosystems result in increased OM relative to forested alpine systems (Fig. 10).

Soils formed in alpine and high plateau regions have characteristic morphological features. Strong and frequent diurnal and seasonal freeze–thaw cycles can result in well-developed granular structure in surface horizons due to needle ice formation and reticulate or platy structures in subsurface horizons. Porous cryptogamic crusts are common. These soils can also have cryogenic fabrics such as lenticular and reticulate structures in mountains with high precipitation. On the QTP, however, with its general arid environment, these structures are more abundant in the lower active layers. Rock fragments are not only subject to physical weathering because of freeze–thaw, but also grind against each other during the process, which produces finer particles, especially silt. Thus, many rock fragments in the soil profile are silt-capped. Cryoturbation occurs in these soils and patterned ground such as sorted circles (particularly on exposed ridges and slopes) is common. In addition, these soils can have such strong diurnal temperature fluctuations that freeze–thaw cycles are common even during the summer.

Classification and carbon stocks. Cold soils of alpine and plateau regions are very diverse and are commonly classified as Mollisols (Chernozems), Alfisols (Luvisols), Spodosols (Podzols), Inceptisols (Cambisols), Andisols (Andosols), and Entisols (Regosols), with minor components of Histosols (Histosols). On the QTP, soils under alpine meadow vegetation without near-surface permafrost are commonly Mollisols (Chernozems) Inceptisols (Cambisols), or Entisols (Regosols), while those with near-surface permafrost are Gelisols (Cryosols). Alpine steppe and alpine desert soils verge on aridic soil moisture regimes and may be classified as Aridisols (Calcisols, Gypsisols, or Durisols) or Gelisols (Cryosols). Carbon stocks in alpine and plateau soils are also highly variable, spanning a wide range from less than 4 kg C m⁻² to greater than 30 kg C m⁻².

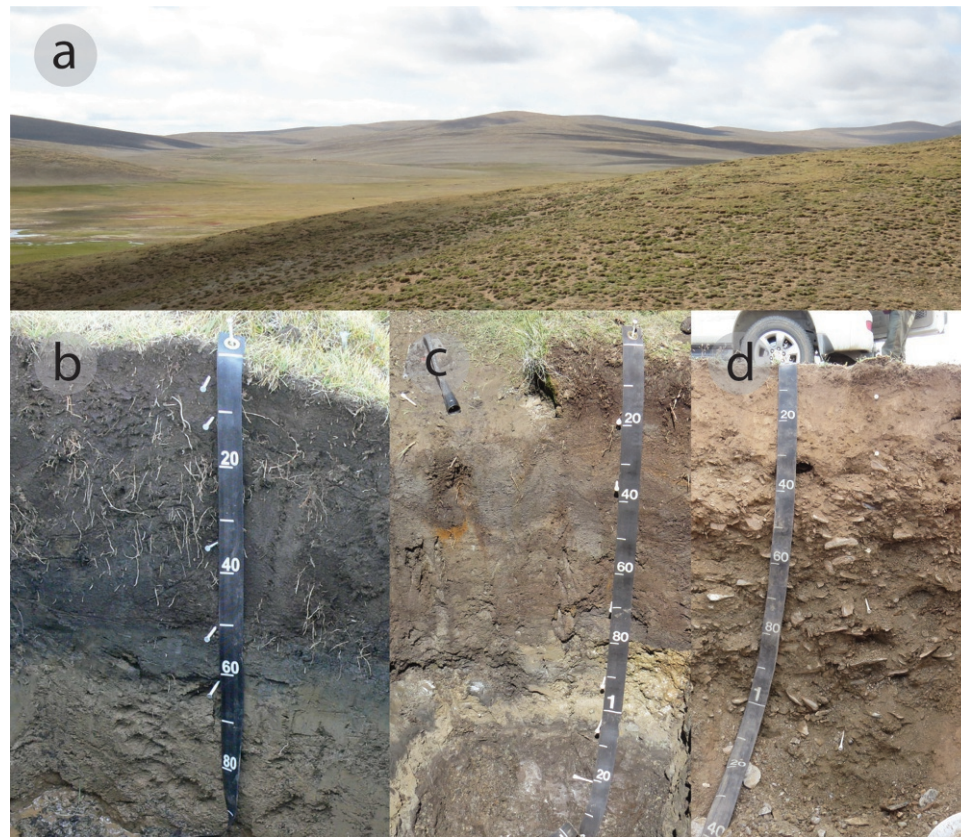


Fig. 10 Representative landscape and soils of the Qinghai-Tibet Plateau (QTP). (A) Alpine dry meadow-steppe landscape (C.-L. Ping); (B) Alpine meadow soil near Qilian, QTP (C.-L. Ping); (C) Soil under *Kobresia* meadow, QTP (C.-L. Ping), (D) Alpine steppe soil, QTP (C.-L. Ping).

Conclusion

Understanding cold region soil genesis and distribution is an important part of improving global carbon models and projecting the impact of environmental and ecosystem change throughout the 21st century. Although research on cold region soils has been locally extensive, there are many areas in the NCR and SCR as well as the QTP and low latitude alpine regions where data on soil properties and distribution remain extremely limited. For example, in the US state of Alaska, less than 15% of the ice-free land area (as of 2021) is covered by a modern detailed soil map that can be used for land use and management decisions at a local scale. Additionally, the SCR and Antarctica are often not represented in global maps of soil classes and soil properties. It is critical that ongoing research addresses these gaps in our knowledge because of the importance and urgency of understanding cold region soils in the global context of our changing environment.

References

- Bliss LC (1977) *Truelove Lowland, Devon Island, Canada: A High Arctic Ecosystem*. Edmonton: University of Alberta Press, 714.
- Bockheim JG (1997) Properties and classification of cold desert soils from Antarctica. *Soil Science Society of America Journal* 1997(61): 224–231.
- Bockheim JG (2015a) Global distribution of cryosols with mountain permafrost: An overview. *Permafrost and Periglacial Processes* 26: 1–12.
- Bockheim JG (ed.) (2015b) *The Soils of Antarctica*. In: *World Soils Book Series* Switzerland: Springer International Publishing.
- Bockheim JG and Ugoili FC (1990) A review of pedogenic zonation in well-drained soils of the Southern Circumpolar Region. *Quaternary Research* 34: 47–66.
- Börker J, Hartmann J, Amann T, and Romero-Mujalli G (2018) Terrestrial sediments of the earth: Development of a global unconsolidated sediments map database (GUM). *Geochemistry, Geophysics, Geosystems* 19(4): 997–1024.
- Brown J, Ferrians O, Heginbottom JA, and Melnikov E (2002) *Circum-Arctic Map of Permafrost and Ground-Ice Conditions, Version 2. Permafrost Extent*. Boulder, Colorado USA. NSIDC: National Snow and Ice Data Center <https://doi.org/10.7265/skbg-kf16>. [Date Accessed].
- Campbell IB and Claridge GGC (1987) Antarctica: Soils, weathering processes and environment. *Developments in Soil Science*. vol. 16. Amsterdam, Netherlands: Elsevier. 368 pp.
- Clapperton Chalmers M, Sugden David E, Birnie Jacqueline, and Wilson Mandy J (1989) Late-Glacial and Holocene Glacier Fluctuations and Environmental Change on South Georgia, Southern Ocean. *Quaternary Research* 31: 210–228.
- Hou X (2019) *1:1 million vegetation map of China*. National Tibetan Plateau Data Center. <https://data.tpdc.ac.cn/en/data/eac4f2cf-d527-4140-a35d-79992957f043/>.
- Hugelius G, Loisel J, Chadburn S, Jackson RB, Jones M, MacDonald G, Marushchak M, Olefeldt D, Packalen M, Siewert MB, Treat C, Turetsky M, Voigt C, and Yu Z (2020) Large stocks of peatland carbon and nitrogen are vulnerable to permafrost thaw. *Proceedings of the National Academy of Sciences* 117(34): 20438–20446.

- IUSS Working Group WRB (2015) *World Reference Base for Soil Resources 2014, update 2015*. International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports No. 106 Rome: FAO.
- Jelinski NA, Sousa MJ, Williams A, GreyBear E, Finnesand K, Mulligan D, Cole C, Stilling MD, and Feinberg JM (2019) Cryoturbation and carbon stocks in Gelisols under late-successional black spruce forests of the Copper River Basin, AK. *Soil Science Society of America Journal* 83: 1760–1778. <https://doi.org/10.2136/sssaj2019.07.0212>.
- Kimble JM, Seybold C, Harms D, Aislable J, Mcleod M, Paetzold R, and Sletten R (2021) Soils and soil climate dataset, Antarctica since 1999. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/climate/?cid=nrcs142p2_053772.
- Munroe JS (2012) Physical, Chemical, and Thermal Properties of Soils across a Forest-Meadow Ecotone in the Uinta Mountains, Northeastern Utah, U.S.A. *Arctic, Antarctic, and Alpine Research* 44(1): 95–106. <https://doi.org/10.1657/1938-4246-44.1.95>.
- Obu J, Westermann S, Kääb A, and Bartsch A (2018) *Ground Temperature Map, 2000–2016, Northern Hemisphere Permafrost*. Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, PANGAEA. <https://doi.org/10.1594/PANGAEA.888600>.
- Obu J, Westermann S, Kääb A, and Bartsch A (2019a) *Ground Temperature Map, 2000–2017, Antarctic*. University of Oslo, PANGAEA. <https://doi.org/10.1594/PANGAEA.902576>.
- Obu J, Westermann S, Kääb A, and Bartsch A (2019b) *Ground Temperature Map, 2000–2016, Andes, New Zealand and East African Plateau Permafrost*. University of Oslo, PANGAEA. <https://doi.org/10.1594/PANGAEA.905512>.
- Otero XL, De La Peña-Lastra S, Pérez-Alberti A, et al. (2018) Seabird colonies as important global drivers in the nitrogen and phosphorus cycles. *Nature Communications* 9: 246. <https://doi.org/10.1038/s41467-017-02446-8>.
- Ping CL, Michaelson GJ, Kimble JM, and Walker DA (2005) Soil acidity and exchange properties of cryogenic soils in Arctic Alaska. *Soil Science & Plant Nutrition* 51(5): 649–653.
- Ping C, Michaelson GJ, Kimble JM, Romanovsky VE, Shur Y, Swanson DK, and Walker DA (2008a) Cryogenesis and soil formation along a bioclimate gradient in Arctic North America. *Journal of Geophysical Research* 113.
- Ping CL, Michaelson G, Jorgenson M, et al. (2008b) High stocks of soil organic carbon in the North American Arctic region. *Nature Geoscience* 1: 615–619. <https://doi.org/10.1038/ngeo284>.
- Potapov P, Hansen MC, Stehman SV, Loveland TR, and Pittman K (2008) Combining MODIS and Landsat imagery to estimate and map boreal forest cover loss. *Remote Sensing of Environment* 112(9): 3708–3719. <https://glad.geog.umd.edu/projects/gfm/boreal/data.html>.
- Rainsley E, et al. (2019) Pleistocene glacial history of the New Zealand subantarctic islands. *Climate of the Past* 15: 423–448. <https://doi.org/10.5194/cp-15-423-2019>.
- Rodrigués WF, Oliveira FS, Schaefer C, Leite MG, and Pavinato PS (2021) Phosphatization under birds' activity: Ornithogenesis at different scales on Antarctic Soilscapes. *Geoderma* 391: 114950.
- Rodríguez D, Schulz GA, Aleksa A, and Tenti Vüegen L (2019) *Distribution and classification of soils. The Soils of Argentina*. Springer. 63–80.
- Simas FNB, Schaefer CEGR, Melo VF, Albuquerque-Filho MR, Michel RFM, Pereira VV, Gomes MRM, and Costa LM (2007) Ornithogenic cryosols from maritime Antarctica: phosphatization as a soil forming process. *Geoderma* 138: 191–203.
- Smith SW and Karlsson S (2017) High stocks, but slow recovery, of ecosystem carbon in southern oceanic tussock grasslands. *Polar Biology* 40: 1617–1628. <https://doi.org/10.1007/s00300-017-2084-5>.
- Soil Survey Staff (2014) *Keys to Soil Taxonomy*, 12th edn Washington, DC: USDA-Natural Resources Conservation Service.
- Tedrow JCF (1977) *Soils of the Polar Landscapes*. New Brunswick, NJ: Rutgers University Press. 639 pp.
- Tedrow JCF (1991) Pedogenic linkage between the cold deserts of Antarctica and the polar deserts of the High Arctic. *Contributions to Antarctic Research II: Antarctic Research Series* 53: 1–17.
- Ugolini F (1986) Pedogenic zonation in the well-drained soils of the Arctic regions. *Quaternary Research* 26(1): 100–120. [https://doi.org/10.1016/0033-5894\(86\)90086-4](https://doi.org/10.1016/0033-5894(86)90086-4).
- USDA-NRCS. (2016) *Soil Climate Regions Map*. Washington, D.C.: Soil and Plant Sciences Division. <https://www.nrcs.usda.gov/wps/portal/nrcs/main/soils/use/worldsoils/>.
- Walker DA, Raynolds MK, Daniëls FJ, Einarsson E, Elvebakk A, Gould WA, Katenin AE, Kholod SS, Markon CJ, Melnikov ES, Moskalenko NG, Talbot SS, Yurtsev BA, and The other members of the CAVM Team (2005) The Circumpolar Arctic vegetation map. *Journal of Vegetation Science* 16: 267–282. <https://doi.org/10.1111/j.1654-1103.2005.tb02365.x>.
- Wang Z-W, Wang Q, Zhao L, Wu X-D, Yue G-Y, Zou D-F, Nan Z-T, Liu G-Y, Pang Q-Q, Fang H-B, Wu T-H, Shi J-Z, Jiao K-Q, Zhao Y-H, and Zhang L-L (2016) Mapping the vegetation distribution of the permafrost zone on the Qinghai-Tibet Plateau. *Journal of Mountain Science* 13(6): 1035–1046. <https://doi.org/10.1007/s11629-015-3485-y>.
- Wilson P (2001) Rate and nature of podzolisation in aeolian sands in the Falkland Islands, South Atlantic. *Geoderma* 101: 77–86.
- Zhang Y, Li B, and Zheng D (2014) Datasets of the boundary and area of the Tibetan Plateau[J/DB/OL]. *Digital Journal of Global Change Data Repository* 2014. <https://doi.org/10.3974/geodb.2014.01.12.V1> also <http://www.geodoi.ac.cn/weben/doi.aspx?id=135>.

Boreal forest soils

Gabriele Broll, Institute of Geography, University of Osnabrueck, Osnabrueck, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	280
Soils of the boreal forest ecozone	280
Genesis, properties, distribution and classification	280
Retisols (US: Alfisols)	282
Podzols (US: Spodosols)	283
Gleysols (US: Soils with aquic conditions)	284
Histosols (US: Histosols)	284
Cryosols (US: Gelisols)	285
Carbon of boreal forest soils and climate change	286
Land use and disturbances in boreal forests and their impact on soils	287
Conclusions	287
References	288

Abstract

The boreal ecozone is a continuous vegetation belt dominated by coniferous trees located between latitudes 50° and 70° N in the northern hemisphere. Dominating soil types in this ecozone are Podzols, Retisols, Gleysols, Histosols and Cryosols. Many boreal forest sites are characterized by disturbances such as forest fires. Human impacts result in the effects of unsustainable forestry and many contaminants. All of these disturbances affect the soils. Furthermore, the boreal forest plays an important role in the global carbon cycle. It is a great carbon sink and this aspect could be seriously affected by climate change. The northern high latitudes have experienced considerable warming, which especially affects the permafrost soils.

Key points

- Genesis, properties, distribution and classification of soils of the boreal forest ecozone
- Carbon of boreal forest soils and climate change
- Land use and disturbances in boreal forests and their impact on soils

Introduction

The boreal forest ecozone in the northern hemisphere (boreal means north in the Greek language) covers approximately $15 \times 10^6 \text{ km}^2$. Previously, it was also referred to as the taiga (Russian term). This vegetation belt is dominated by coniferous trees and is located between latitudes 50° and 70° N (Fig. 1). In North America the boreal forest is widespread in Alaska and especially in Canada (Fig. 2). A vast boreal forest zone also exists in Russia and China, whereas the European part is comparatively small (Fig. 3) (McLaren and Turkington, 2013; DeAngelis, 2019).

The boreal biome is characterized by a diversity related to climate (maritime or continental), hydrology (dry, mesic or wet), plant cover and soil type. The climate of the boreal forest is characterized by long-lasting winter snowpack and short summers. Therefore, the growing season is less than 6 months. The amplitude between summer and winter temperatures is usually extreme in continental parts and small in maritime regions of the boreal ecozone. In those areas where the soil temperature is below 0 °C for more than 2 years permafrost occurs and affects the hydrology of the forest causing stagnant water. In addition, the amount of precipitation is associated with the location in the ecozone. Precipitation is low in continental regions and high in maritime regions. A large proportion of the annual precipitation usually falls as snow (McLaren and Turkington, 2013; DeAngelis, 2019).

Many boreal forest sites are characterized by disturbances such as forest fires, snow breakage, and windthrow. Human impacts are unsustainable forestry and various types of contamination. Moreover, the boreal forest is a great carbon sink and is therefore greatly affected by climate change.

Soils of the boreal forest ecozone

Genesis, properties, distribution and classification

The following soil types (named after IUSS Working Group WRB, 2022) occur in the boreal forest: Podzols, Retisols, Histosols, Cryosols and Gleysols (Jones et al., 2010; Thiffault, 2019). Following the Soil Survey Staff, 2014 the soils are named: Spodosols,

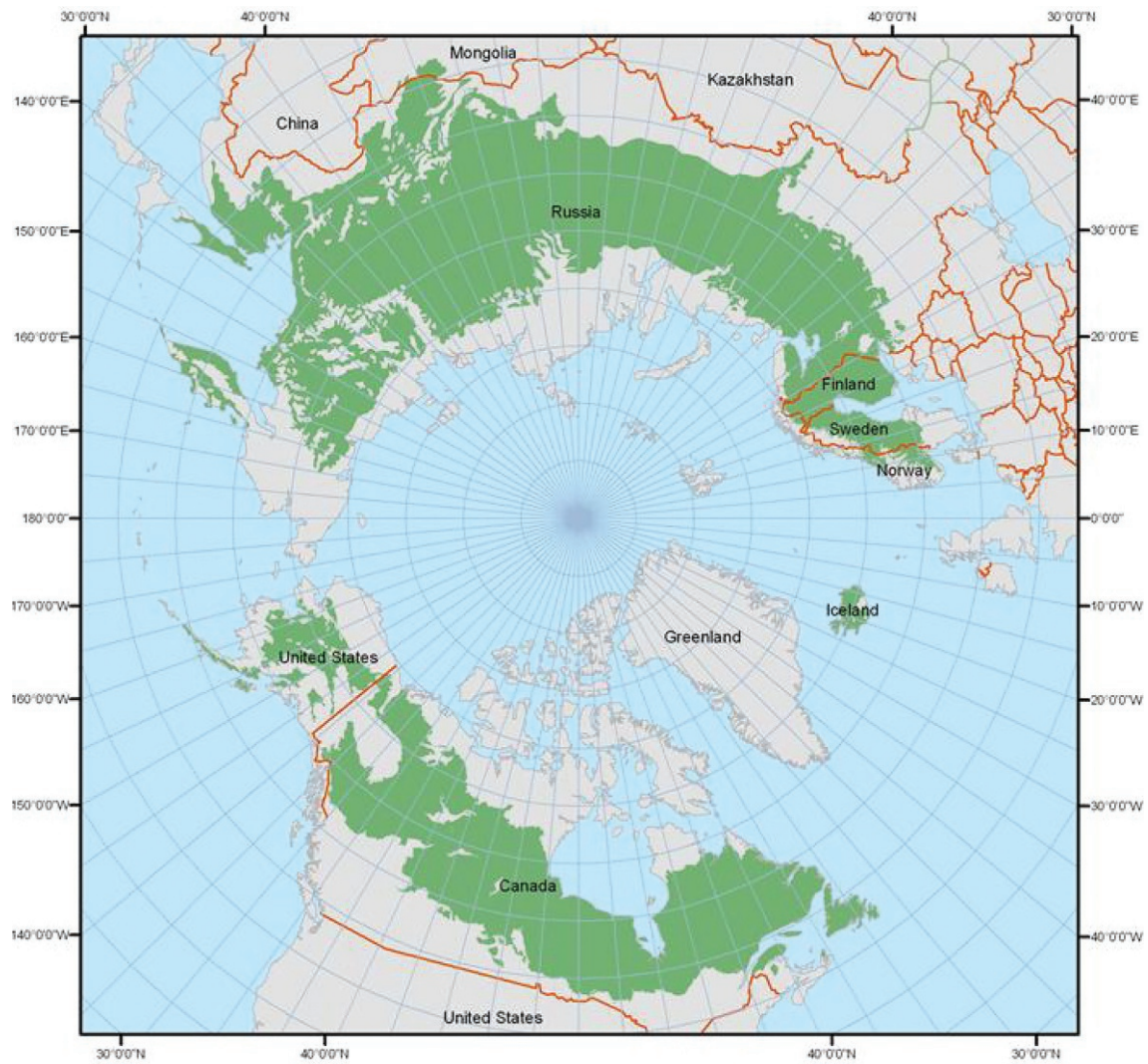


Fig. 1 Worldwide distribution of the boreal forest (shown in green). Source: WWF/ESRI. Map: Hatfield Consultants. <http://www.ramp-alberta.org/river/boreal/distribution.aspx>.



Fig. 2 Typical plant cover in the boreal forest, Yukon Canada.



Fig. 3 Typical plant cover in the boreal forest, Northern Finland.

Alfisols, Histosols and Gelisols. Soils influenced by groundwater like Gleysols are referred to in the US Soil Taxonomy only at the suborder level as under aquic conditions. In addition, a few more soil types occur in the boreal forest that are not dominant (in detail: [Zech et al., 2014](#)).

It is the interaction of the pedogenetic factors of climate, geological substrate, topography and vegetation which explains why Podzols, Retisols, Gleysols, Histosols and Cryosols prevail in the boreal ecozone. The climate is characterized by relatively low temperatures, low evaporation and sufficient precipitation to allow infiltration either to the groundwater or to a barrier in the subsoil such as impermeable clay layers, bedrock or permafrost. Therefore, well-drained soils exhibit the effects of leaching processes while poorly drained soils show the effects of water saturation.

The underlying materials of large areas of the boreal zone in North America and Eurasia are characterized by a mix of pre-Quaternary igneous and metamorphic rocks partly covered by Quaternary sediments from various glaciations. In general, this geological substrate is acidic and poor in nutrients.

Thus, the original coniferous forest is closely associated with the climatic conditions and nature of the parent material. The vegetation produces slowly decomposing organic matter which accumulates and produces acids. Thus, most soils are characterized by low pH values. In general exposed parent material like gneiss or granite is covered by cryptogams which cause intense chemical weathering of the rocks. Lichens seem to be more effective than mosses in altering feldspars, for example, to clay minerals ([Jackson, 2015](#)). One reason could be the production of strong organic acids by lichens.

Large areas in the boreal ecozone are relatively level or characterized by gently rolling uplands. Mountainous topography occurs in some areas only. Level areas are characterized by a varying microtopography. Thus, a mosaic of well-drained Podzols and poorly drained Gleysols or Histosols occurs. In the northern or more continental parts of the boreal forest permafrost-affected soils (Cryosols) appear because of the very low winter temperatures. Retisols are common for the southern part of the boreal forest with a thick cover of often loamy Quaternary sediments which have characteristic leaching processes.

Retisols (US: Alfisols)

Retisols, the former *Albeluvisols*, are characterized by the leaching of clay from the topsoil to the subsoil. Retisols have a clay illuviation horizon with an interfingering of bleached coarse textured soil material into the fine textured illuvial horizon. Clay and iron oxides are partly removed from the coarse textured material. In the boreal ecozone they are distributed on level to undulating plains mostly associated with glacial till as the parent material ([IUSS Working Group WRB, 2022](#)).

Podzols (US: Spodosols)

Podzols are widespread in the boreal ecozone. They are usually distributed under coniferous forest and on non-calcareous nutrient poor parent material. Podzols are characterized by a spodic (gray, leaching of iron (Fe), aluminum (Al) and organic matter) horizon (IUSS Working Group WRB, 2022). In addition to the coniferous litter, the role of litter derived from the understory vegetation has recently been shown to be involved in these processes for boreal forest stands (Hensgens et al., 2020). Thus, the typical podzolization processes in boreal forests with leaching of dissolved organic matter and elements are possibly mainly driven by decomposition of the litter provided by the understory vegetation like *Vaccinium myrtillus*. Podzols have reddish and or dark horizons in the subsoil below the leached gray horizon (Fig. 4). Depending on the climatic conditions, Podzols that are highly enriched with organic matter in the subsoil occur in regions with high precipitation, whereas those with strong enrichment of iron oxides in the subsoil have formed in regions of low precipitation (Sanborn et al., 2011). In many northern regions the soil horizons of Podzols are very shallow. Sometimes, vertical leaching along stones can be found (Fig. 5). In those regions where Podzols have been influenced by frequent intense freeze–thaw cycles in past soil horizons often show the effects of cryoturbation (Fig. 6).



Fig. 4 Podzol profile, Finland.



Fig. 5 Podzol mini profile, Finland.

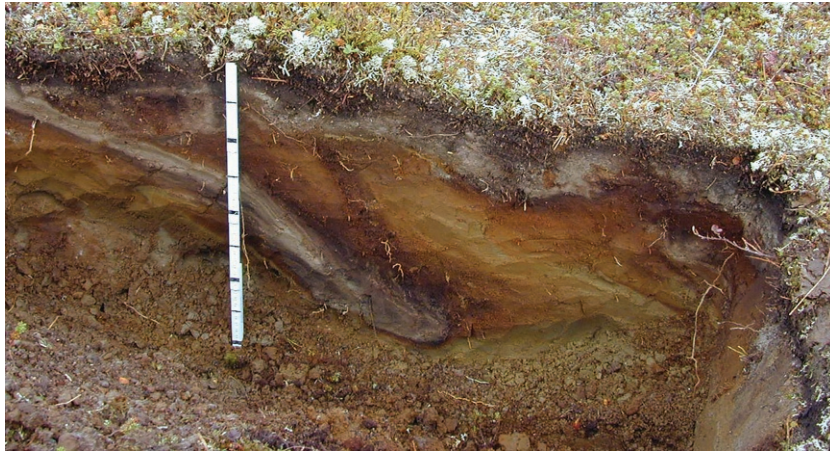


Fig. 6 Podzol profile, cryoturbated, Finland.

Gleysols (US: Soils with aquic conditions)

Gleysols are those affected by changes in groundwater depths resulting in periods of waterlogging and periods of drying, which lead to the formation of redoximorphic features. The upper part of the soil (reddish color) is characterized by oxidizing conditions while reducing conditions prevail in some parts of every sublayer (grayish color) (Zech et al., 2014; IUSS Working Group WRB, 2022). In the boreal ecozone Gleysols can be found around lakes (Fig. 7) or in floodplains. They are sometimes characterized by cryoturbation.

Histosols (US: Histosols)

In the boreal zone, Histosols (Figs. 8 and 9) comprise soils formed from organic material mostly accumulated as groundwater peat (fen) or rainwater peat (bog) (IUSS Working Group WRB, 2022). Most of the world's peatlands with the soil type Histosol occur in the boreal zone (Wieder and Vitt, 2006). Therefore, Histosols play an important role in this biome because they serve as a carbon sink.



Fig. 7 Gleysol at a lake, cryoturbation indicated by patterned ground, Finland.



Fig. 8 Peatland landscape, Finland.



Fig. 9 Histosol profile with cryoturbation features, Finland.

Cryosols (US: Gelisols)

Permafrost-affected soils (Cryosols) occur in the north and in the continental part of the boreal forest (Jones et al., 2010; Zech et al., 2014). The subsurface layers (cryic horizon), and if present, soil water are permanently frozen. Cryogenic processes (frost heave, cryogenic sorting of solid material, thermal cracking, ice segregation, patterned ground, gelifluction) are the dominant soil-forming processes in most Cryosols (Fig. 10) (IUSS Working Group WRB, 2022). Land use such as agriculture or fire events that remove the vegetation cover has a great impact on Cryosols, because the permafrost table drops down quickly in the profile (Fig. 11). Therefore, decomposition will be enhanced because of the increasing oxygen supply.



Fig. 10 Turbic Cryosol profile, Yukon Canada, red marks at the top of the permafrost table.

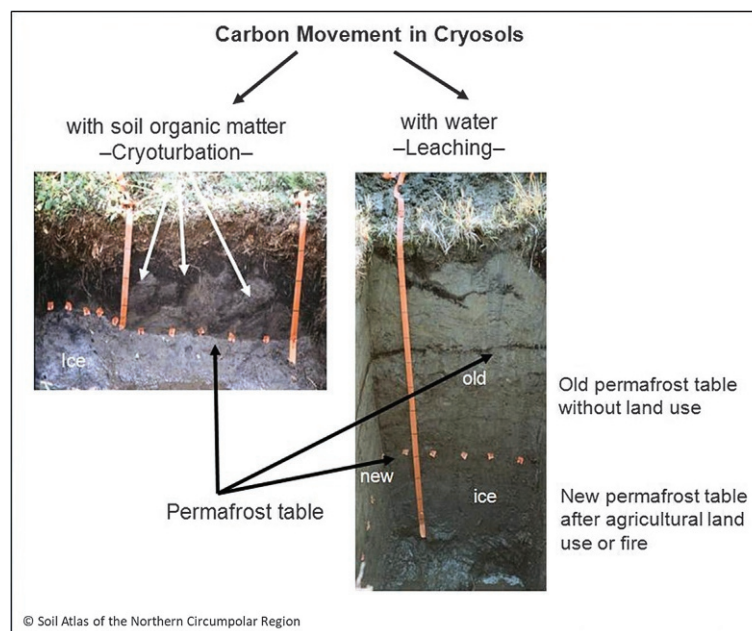


Fig. 11 Carbon movement in Cryosols under the influence of land use change.

Carbon of boreal forest soils and climate change

The Intergovernmental Panel on Climate Change, [IPCC \(2018\)](#) stresses that “boreal forests are particularly at risk of climate change-induced degradation and loss, with woody shrubs already encroaching into the tundra (high confidence) and this will proceed with further warming. . . .” This invasion of plants to the north or to subalpine zones and the effects on genesis and properties of soils have been shown for many forest-tundra ecotones of the northern hemisphere ([Moscatelli et al., 2017](#); [Holtmeier and Broll, 2018](#)).

The boreal forest plays an important role in the global carbon cycle. It contains about 30% of the world’s forest carbon stocks and acts as an important carbon sink. The northern high latitudes are experiencing considerable warming. Therefore, boreal forests are becoming more vulnerable to drought, fire and insect outbreaks, for example ([DeLuca and Boisvenue, 2012](#); [Hugelius et al., 2013](#);

Buermann et al., 2014). Permafrost-affected soils are very sensitive to climate warming (Tarnocai et al., 2009; Ping et al., 2010; IPCC, 2018; Stuenzi et al., 2021).

Concern is growing that due to the possible transition from a carbon sink to a carbon source old soil organic matter could be re-introduced into the modern carbon cycle. This risk for mobilization of ancient carbon as dissolved organic carbon and its re-introduction into the modern carbon cycle has been investigated in detail in Podzols of the Swedish boreal forest (Hensgens et al., 2021). Dissolved organic carbon derived from plant litter plays an important role in the carbon balance and the biogeochemistry of forests in general. In addition, the forest floor and thus the carbon stocks of boreal forests in Canada are also sometimes modified by invasive earthworms. The enhanced bioturbation increases the decomposition of the litter (Lejoly et al., 2021).

Peatlands (unfrozen and frozen organic soils) of the boreal region are most affected by climate change because these areas contain large amounts of organic carbon (Wieder and Vitt, 2006). Peatlands have acted as a carbon sink in sequestering carbon during the Holocene. They are also an important global source of methane. Human-induced climate change will change the ecohydrological conditions of the peatlands. These changes are considered to be the primary driving force for future changes of the peatland carbon cycle (Wieder and Vitt, 2006). Approximately 56% of the organic carbon mass in all Canadian peatlands will be severely affected by climate change. It will result in serious degradation of perennially frozen peatlands in the Canadian Subarctic and boreal regions. The degradation in these areas will result in wetter conditions than before during summer. Emissions of methane will be enhanced in this way. On the other hand severe drying of peatlands can occur in the southern parts of the boreal region. Drying of peatlands means increasing decomposition with the consequence of high emissions of carbon dioxide (Tarnocai, 2009; Laamrani et al., 2020; Olefeldt et al., 2021).

Land use and disturbances in boreal forests and their impact on soils

Several types of disturbance in boreal forest ecosystems influence forest soils and therefore biogeochemical processes. Natural disturbances are snow breakage, windthrow, fires and diseases as well as outbreaks of leaf-eating insects and or bark beetles, which may cause severe damage to trees. In mountainous regions mass movements on slopes also affect forest soils by erosion (e.g., Holtmeier and Broll, 2018).

Natural wildfires have been the principle disturbance in boreal forests for a long time. Large amounts of carbon are emitted to the atmosphere because of the combustion of organic matter from the forest floor and the topsoil. Climate warming and drying of the soil have led to an increase in the frequency, extent, intensity and severity of wildfires which could result in the burning of soil organic matter to greater depth. Thus, boreal forests may change from a net carbon sink to a net carbon source, which is a potentially serious problem in relation to the global carbon cycle (Turetsky et al., 2011; Walker et al., 2019).

In addition to the influence of human-induced climate change further human effects on boreal forest ecosystems are related to land use and contamination. Large boreal areas are used by forestry (Wieder and Vitt, 2006), which can affect soil chemical properties such as increasing acidity and soil physical properties like soil compaction can be induced. Meanwhile forest management to preserve ecosystem services and for ecological restoration in relation to carbon sequestration have been focused upon (Safford and Vallejo, 2019; Jørgensen et al., 2021; Kellomäki, 2022).

Boreal forest soils have also become increasingly contaminated. Acidic emissions and heavy metals play a role at the regional level, for example, the copper–nickel industry on the Kola Peninsula has a considerable impact on the plant cover and soils in Finland and Russia (Kashulina, 2017). The contamination affects soil chemical processes and also soil organisms, for example, the contamination with mercury in Swedish forest soils (Xu et al., 2019). In the Canadian boreal forests mining also affects the soils at the regional scale such as the mining of zinc (Hamilton et al., 2016). Mobilization of contaminants from acid sulfate soils should also be mentioned (Yli-Halla et al., 2017). Sulfate is released into the pore water after oxidation of the sulfidic materials in the soil, e.g., after drainage of the soils. Then, the simultaneously produced acidity dissolves metals from the soil matrix. Acid sulfate soils occur in mining areas with peat soil and also in the coastal area around the Baltic Sea. Large amounts of zinc have been leached in these acid sulfate soils (Yli-Halla et al., 2017). Furthermore, contamination by hydrocarbons can have toxic effect on soils, especially in oil mining areas. Depending on the intensity of the contamination Gleysols and Podzols in Siberia can be considered as Technosols (Bykova et al., 2021).

Conclusions

Soils in the boreal zone show great pedodiversity belowground despite the dominant plant cover of coniferous trees aboveground. This is mainly due to the spatial variability in topography and thus drainage conditions. Consequently, in this landscape soils can be dry like Podzols or even waterlogged like Histosols within short distances affecting the chemical and biological soil properties. Soils in the boreal zone are also greatly affected by climate change. Therefore, the impact on the carbon cycle is considerable. The northern and southern ecotones of the boreal zone are indicators of global warming. The southern border is increasingly characterized by deciduous forest and also by agricultural activities. Both kinds of land use change soil properties. The northern border of the boreal zone is characterized by invading trees to the north and also by thawing of the permanently or seasonally frozen ground. This strong anthropogenic impact on boreal soils will also increase the area of contaminated soils which have been localized areas until now, but will be enlarged by increasing industrial activities in the north.

References

- Buermann W, Parida B, Jung M, et al. (2014) Recent shift in Eurasian boreal forest greening response may be associated with warmer and drier summers. *Geophysical Research Letters* 41: 1995–2002.
- Bykova MV, Alekseenko AV, Pashkevich MA, and Drebenstedt K (2021) Thermal desorption treatment of petroleum hydrocarbon-contaminated soils of tundra, taiga, and forest steppe landscapes. *Environmental Geochemistry and Health* 43: 2331–2346.
- DeAngelis DL (2019) The boreal forest ecosystem. In: Fath B (ed.) *Encyclopedia of Ecology*, 2nd edn., pp. 479–483. Amsterdam: Elsevier.
- DeLuca T and Boisvenue C (2012) Boreal forest soil carbon: Distribution, function and modelling. *Forestry* 85: 161–184.
- Hamilton JG, Farrell RE, Chen N, et al. (2016) Characterizing zinc speciation in soils from smelter-affected boreal forest ecosystem. *Journal of Environmental Quality* 45: 684–692.
- Hensgens G, Laudon H, Peichl M, et al. (2020) The role of the understory in litter DOC and nutrient leaching in boreal forests. *Biogeochemistry* 149: 87–103.
- Hensgens G, Laudon H, Johnson MS, and Berggren M (2021) The undetected loss of aged carbon from boreal mineral soils. *Scientific Reports* 11: 6202.
- Holtmeier F-K and Broll G (2018) Subalpine forest and treeline ecotone under the influence of disturbances: A review. *Journal of Environmental Protection* 9: 815–845.
- Hugelius G, Tarnocai C, Broll G, et al. (2013) The Northern Circumpolar Soil Carbon Database: Spatially distributed datasets of soil coverage and soil carbon storage in the northern permafrost regions. *Earth System Science Data* 5: 3–13.
- IPCC (2018) Summary for policymakers. In: Masson-Delmotte V, Zhai P, and Pörtner H-O, et al. (eds.) *Global Warming of 1.5° C. An IPCC Special Report on the Impacts of Global Warming of 1.5° C Above Pre-Industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty*. Geneva: World Meteorological Organization.
- IUSS Working Group WRB (2022) *World Reference Base for Soil Resources 2022: International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*. Rome: Food and Agriculture Organization of the United Nations.
- Jackson TA (2015) Weathering, secondary mineral genesis, and soil formation caused by lichens and mosses growing on granitic gneiss in a boreal forest environment. *Geoderma* 251–252: 78–91.
- Jones A, Stolbovoy V, and Tarnocai C, et al. (eds.) (2010) *Soil Atlas of the Northern Circumpolar Region*. Luxembourg: Publications Office of the European Union.
- Jørgensen K, Granath G, Lindahl BD, and Stenborg J (2021) Forest management to increase carbon sequestration in boreal Pinus sylvestris forests. *Plant and Soil* 466: 165–178.
- Kashulina GM (2017) Extreme pollution of soils by emissions of the copper-nickel industrial complex in the Kola Peninsula. *Eurasian Soil Science* 50: 837–849.
- Kellomäki S (2022) *Management of Boreal Forests: Theories and Applications for Ecosystem Services*. Berlin, Heidelberg: Springer.
- Laamrani A, Valeria O, Chehboni A, and Bergeron Y (2020) Analysis of the effect of climate warming on paludification processes: Will soil conditions limit the adaptation of northern boreal forests to climate change? A synthesis. *Forests* 11: 1176.
- Lejoly J, Quideau S, and Laganière J (2021) Invasive earthworms affect soil morphological features and carbon stocks in boreal forests. *Geoderma* 404: 115262.
- McLaren JR and Turkington R (2013) Boreal forest ecosystems. In: Levin SA (ed.) *Encyclopedia of Biodiversity*, 2nd edn., pp. 626–635. Elsevier.
- Moscatelli MC, Bonifacio E, Chiti T, et al. (2017) Soil properties as indicators of treeline dynamics in relation to anthropogenic pressure and climate change. *Climate Research* 73: 73–84.
- Olefeldt D, Heffernan L, Jones MC, et al. (2021) Permafrost thaw in northern peatlands: Rapid changes in ecosystem and landscape functions. In: Canadell JG and Jackson RB (eds.) *Ecosystem Collapse and Climate Change*, pp. 27–67. Cham: Springer.
- Ping C-L, Michaelson GJ, Kane ES, et al. (2010) Carbon stores and biogeochemical properties of soils under black spruce forest, Alaska. *Soil Science Society of America Journal* 74(3): 969–978.
- Safford HD and Vallejo VR (2019) Ecosystem management and ecological restoration in the Anthropocene: Integrating global change, soils, and disturbance in boreal and Mediterranean forests. In: Busse M, Giardina C, Morris D, and Page-Dumroese D (eds.) *Global Change and Forest Soils*, 1st edn., pp. 259–308. Amsterdam: Elsevier.
- Sanborn P, Lamontagne L, and Hendershot W (2011) Podzolic soils of Canada: Genesis, distribution and classification. *Canadian Journal of Soil Science* 91: 843–880.
- Soil Survey Staff (2014) *Keys to Soil Taxonomy*, 12th edn. Washington, DC: USDA-Natural Resources Conservation Service.
- Stuenzi SM, Boike J, Gädeke A, et al. (2021) Sensitivity of ecosystem-protected permafrost under changing boreal forest structures. *Environmental Research Letters* 16: 084045.
- Tarnocai C (2009) The impact of climate change on Canadian peatlands. *Canadian Water Resources Journal* 34: 453–466.
- Tarnocai C, Canadell JG, Schuur EA, et al. (2009) Soil organic carbon pools in the northern circumpolar permafrost region. *Global Biogeochemical Cycles* 23: GB2023.
- Thiffault E (2019) Boreal forests and soils. In: Busse M, Giardina C, Morris D, and Page-Dumroese D (eds.) *Global Change and Forest Soils*, 1st edn., pp. 59–82. Amsterdam: Elsevier.
- Turetsky MR, Kane E, Harden J, et al. (2011) Recent acceleration of biomass burning and carbon losses in Alaskan forests and peatlands. *Nature Geoscience* 4: 27–31.
- Walker XJ, Baltzer JL, Cumming SG, et al. (2019) Increasing wildfires threaten historic carbon sink of boreal forest soils. *Nature* 572: 520–523.
- Wieder RK and Vitt DH (eds.) (2006) *Boreal Peatland Ecosystems. Ecological Studies* 188. Berlin, Heidelberg: Springer.
- Xu J, Buck M, Eklöf K, et al. (2019) Mercury methylating microbial communities of boreal forest soils. *Scientific Reports* 9: 518.
- Yli-Halla M, Virtanen S, Mäkelä M, et al. (2017) Abundant stock and mobilization of elements in boreal acid sulfate soils. *Geoderma* 308: 333–340.
- Zech W, Schad P, and Hintermaier-Erhard G (2014) *Böden der Welt*, 2nd edn. Berlin, Heidelberg: Springer.

Soils of humid cool temperate regions

Nicolas A Jelinski^{a,*}, Justin B Richardson^b, and Edward A Nater^a, ^aDepartment of Soil, Water, and Climate, University of Minnesota, Saint Paul, MN, United States; ^bDepartment of Geosciences, University of Massachusetts-Amherst, Amherst, MA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction and definition of humid cool temperate soils	289
Geographic scope and environmental characteristics, and land use	290
In the 'sweet spot'—the biogeochemical middle ground of the temperate zone	291
Excess moisture: Eluviation, illuviation and the leaching of dissolved solutes	292
Groundwater recharge and wetlands in the humid cool temperate zone	292
A predominance of youthful, transported parent materials	293
Progressive mineral weathering, alteration, and transformation	294
The temperate plant-soil interface: Impacts of forest and humid grasslands	295
Classification of humid cool temperate soils	296
Impacts of land use and management on humid cool temperate soils	297
Conclusion	297
References	297

Abstract

Humid cool temperate soils are widely distributed and are some of the most intensively managed soils on the planet due to their favorable properties. The genesis of humid cool temperate soils is dominated by biogeochemical processes operating in moderate temperatures with a pronounced cool season and excess moisture throughout the growing season, resulting in downward leaching of soil constituents and groundwater recharge. In combination with transported soil parent materials, progressive chemical weathering, and distinctive intergrading biomes, humid cool temperate soils represent a unique set of highly fertile soils that are of critical importance to human food, animal feed and fiber requirements.

Key points

- Humid cool temperate ecozones experience excess moisture throughout the year under moderate temperatures with a pronounced cool season.
- These soils are widely distributed on all continents except Antarctica.
- Major trends influencing the development of humid cool temperate soils include, excess moisture, groundwater recharge, youthful parent materials, progressive mineral weathering, and distinctive intergrading forest and grassland biomes.
- Humid cool temperate soils play a disproportionate role in food, animal feed and fiber production.

Introduction and definition of humid cool temperate soils

Humid cool temperate (HCT) ecozones are geographic regions characterized by continental climates with excess moisture and significant seasonal differences in temperature. These ecozones correspond broadly with Köppen-Geiger warm temperate (C) and snow (D) climates, with fully humid (f) and winter dry (d) precipitation regimes and hot summer (a) and warm summer (b) temperature categories. The HCT ecozones are dominated by broadleaf and mixed deciduous forest, grading towards evergreen coniferous forest at the colder boreal limits, humid grassland vegetation at the drier limits and tropical savannah vegetation at the warm subtropical limits. The U.S. Soil Taxonomy (Soil Survey Staff, 2014) and the World Reference Base for Soil Resources (IUSS Working Group WRB, 2015) are useful for defining and classifying soils of the HCT ecozones based on their properties. Where possible, we utilize terminology from both of those systems in this chapter.

Soil temperature and soil moisture regimes in U.S. Soil Taxonomy (STR and SMR, respectively) can define climatic regions more appropriately with regard to pedological (soil forming) processes. The HCT soils have frigid, mesic, or thermic STRs, which correspond to annual average soil temperatures (at a depth of 50 cm from the soil surface) of: 0 °C to <8 °C, with warm summer temperatures (frigid), 8 °C to <15 °C (mesic), and 15 °C to <22 °C (thermic) (Soil Survey Staff, 2014). Within these STRs, HCT soils are additionally restricted to udic (humid) or perudic (aquic) SMRs. The definition of HCT soils as those soils with frigid, mesic or thermic STRs and udic or perudic SMRs notably excludes soils of semi-arid regions that are dominated by mixed- and short-grass prairies in the Great Plains of the United States or the steppes of Eurasia.

This definition of HCT boundaries results in a broad set of geographically distributed soils that have a number of unique pedologic characteristics. Specifically, HCT soils are characterized by: (1) moderate temperatures with a pronounced cool season

that regulate the rate of soil biochemical processes, (2) excess moisture throughout the course of the growing season, which results in downward leaching of soil constituents, (3) surface and groundwater recharge, leading to regionally extensive areas of wetlands, particularly in low-gradient landscapes; (3) a predominance of youthful and transported soil parent materials such as glacial till, loess, or alluvium; (4) progressive chemical weathering processes favoring mineral dissolution and alteration, and (5) an assemblage of distinctive, intergrading biomes dominated by deciduous broadleaf forests, which uniquely influence processes at the plant-soil interface, soil morphology, and specific pedogenetic pathways.

Although HCT ecoregions account for less than 15% of global land area, HCT soils have played a major role in the modern understanding of soil formation and soil genesis. Therefore, soil development in HCT ecoregions has, in many ways, come to represent a central, archetypal conceptual model of pedogenesis. This is partially due to the geographical bias inherent in the development of currently accepted models of pedogenesis (Simonson, 1999; Bockheim et al., 2005) (developed and focused in large part on the eastern United States, western Europe, and Eastern Russia). Despite this historical framing, however, HCT pedogenesis is central only in that it represents an environment between the extremes, and results in changes in soil properties that are readily observable due to gradients in MAT and MAP in HCT regions that align along clear latitudinal and longitudinal gradients. In this chapter we review the distribution, classification, and major pedogenetic themes that unify HCT soils and discuss the impacts of human land management and global environmental change on their properties.

Geographic scope and environmental characteristics, and land use

The HCT soils as defined above fall mainly in the mid-latitude (~ 30 – 60 degrees latitude) regions of the globe (Fig. 1), and are the principal resource for many of the world's most productive and intensively managed agricultural (West et al., 2010) and forest lands (Curtis et al., 2018). Large areas of HCT soils are located across North America (Eastern United States), Europe (southern United Kingdom and France to eastern Russia), Eastern Asia (Eastern/southeastern China and Japan), Oceania (southeastern Australia and New Zealand), and South America (southern Brazil and northeastern Argentina, and southern Chile), ranging in elevation from sea level to ~ 2000 m.a.s.l. (USDA-NRCS, 2016). Soils in HCT ecozones experience humid summers, with adequate precipitation throughout the growing season (in Köppen zones Cwa, Cwb, Dwa and Dwb), or evenly distributed precipitation throughout the entire year (in Köppen zones Cfa, Cfb, Dfa, Dfb).

The HCT ecozones are bounded by cooler boreal regions, where winter air and soil temperatures are lower and growing seasons shorter, and by subtropical and tropical regions, where temperatures are warmer and frosts rarely, if ever, occur. They are also bounded by semiarid and Mediterranean climate regions, where soil moisture becomes insufficient to support the growth of most crops without irrigation during the growing season. Smaller areas of HCT soils may occur outside the dominant latitudes within other ecozones. High elevations and maritime influences are particularly important for generating HCT conditions in tropical and subtropical regions by modifying annual temperature ranges, for example, in mountainous regions of Central and South America,

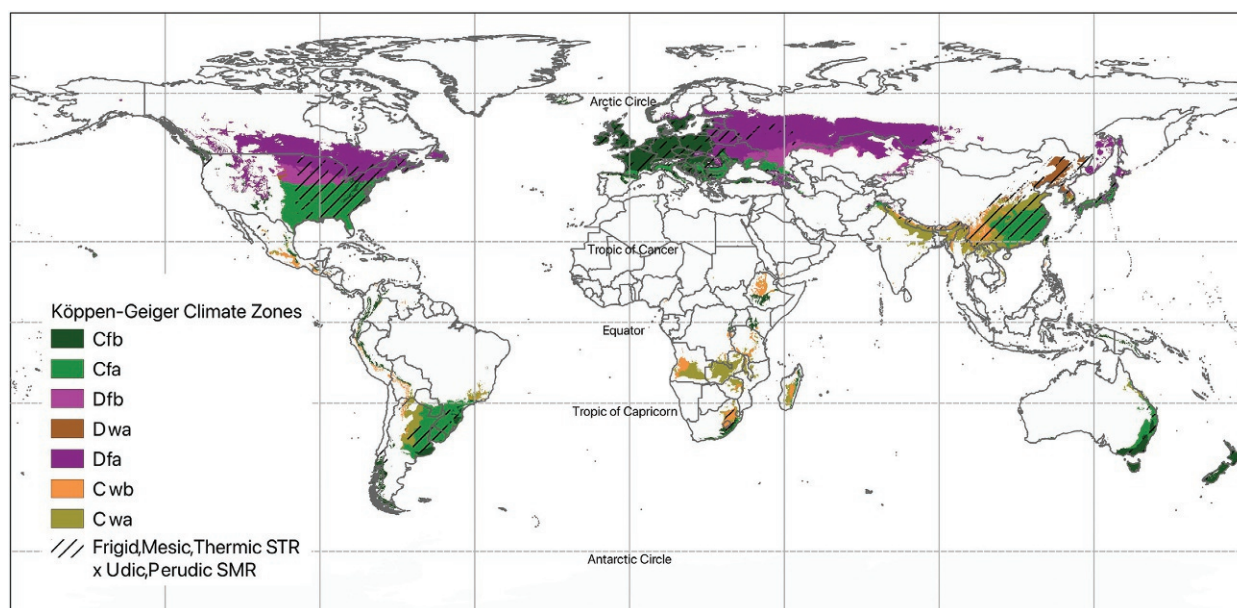


Fig. 1 Distribution of humid cool temperate zone soils (defined as soils with frigid, mesic or thermic soil temperature regimes and udic or perudic soil moisture regimes, shown by crosshatching) are in the humid temperate Köppen-Geiger climate zones. Soil temperature and moisture regime data are from USDA-NRCS (2016) *Soil Climate Regions Map*. Soil and Plant Sciences Division: Washington, DC <https://www.nrcs.usda.gov/wps/portal/nrcs/main/soils/use/worldsoils/>, and Köppen-Geiger climate zone boundaries are from Beck H, Zimmermann N, McVicar T et al. (2018) Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific Data* 5: 180214. doi: 10.1038/sdata.2018.214.

Sub-Saharan Africa, and Southeast Asia. The HCT regions encompass significant spatial and temporal gradients in temperature, precipitation, and atmospheric moisture deficit that are broadly aligned with latitude (temperature) and longitude (precipitation). The relative timing of these factors is an important determinant for vegetative growth in both natural and managed ecosystems, and has a profound effect on evapotranspiration, infiltration, soil moisture status, and plant available moisture.

The growing season length and temperature in HCT ecoregions is sufficient for the production of major cereal and legume crops such as maize (*Zea mays* L.), barley (*Hordeum vulgare* L.), soybean (*Glycine max* (L.) Merr.), and alfalfa (*Medicago sativa* L.), with accumulated annual growing degree days generally ranging from 2000 to 5000. Mean annual precipitation in HCT ecozones ranges from 700 mm yr⁻¹ to greater than 1500 mm yr⁻¹ and precipitation always exceeds evapotranspiration, resulting in excess moisture in upland soils and groundwater recharge (Fick and Hijmans, 2017). In HCT regions, this precipitation is distributed throughout the growing season and is sufficient for the growth of crops without irrigation, though precipitation may decrease towards the end of the growing season in the drier parts of this region, thus necessitating early-maturing crops like winter wheat (*Triticum aestivum* L.) (You et al., 2014).

In the ‘sweet spot’—the biogeochemical middle ground of the temperate zone

Temperature influences the rates of soil biogeochemical processes, particularly organic matter decomposition, soil mineral weathering, and the rate of oxidation and reduction reactions in soils (Chorover et al., 2007). Typically, a 10 °C increase in temperature will produce a two or threefold increase in the rate of many biochemical reactions (Davidson and Janssens, 2006). The HCT soils, by definition, have adequate moisture throughout the year for biochemical reactions to proceed unhindered, and fall between the temperature extremes of the high and low latitudes, which results in moderate rates of biochemical reactions relative to cold and or arid regions (where soil biochemical reaction rates are generally slower), and subtropical or tropical equatorial regions (where soil biochemical reaction rates are faster).

The HCT soils experience a cool season—even in the southern reaches of their extent there is a pronounced slowing in biochemical reactions during the winter as temperatures cool. These intermediate temperatures lead to differences in the rate of microbial decomposition of soil organic matter compared to higher and lower latitudes, which, in turn, influences organic matter accumulation and soil organic carbon (SOC) stocks in HCT soils. The intermediate character of HCT soil temperature regimes generally means that organic carbon stocks are higher than in soils of similar drainage classes in the subtropics and tropics (Poggio et al., 2021). In contrast, SOC stocks tend to be lower in HCT soils than in soils of similar drainage classes at higher latitudes in colder environments. This same pattern holds true for comparisons of rates of HCT soil mineral weathering with those of lower and higher latitudes in similar drainage classes. Clay mineral assemblages in HCT soils tend to be mixed, intermediate between the relatively unweathered soils of higher latitudes and the highly weathered soils of the tropics and subtropics.

Within the HCT region, these trends are reinforced by the occurrence of large areas of younger parent materials in the cooler parts of the region due to the direct or indirect effects of continental glaciation (Fig. 2). Furthermore, the soil surface may be frozen for a

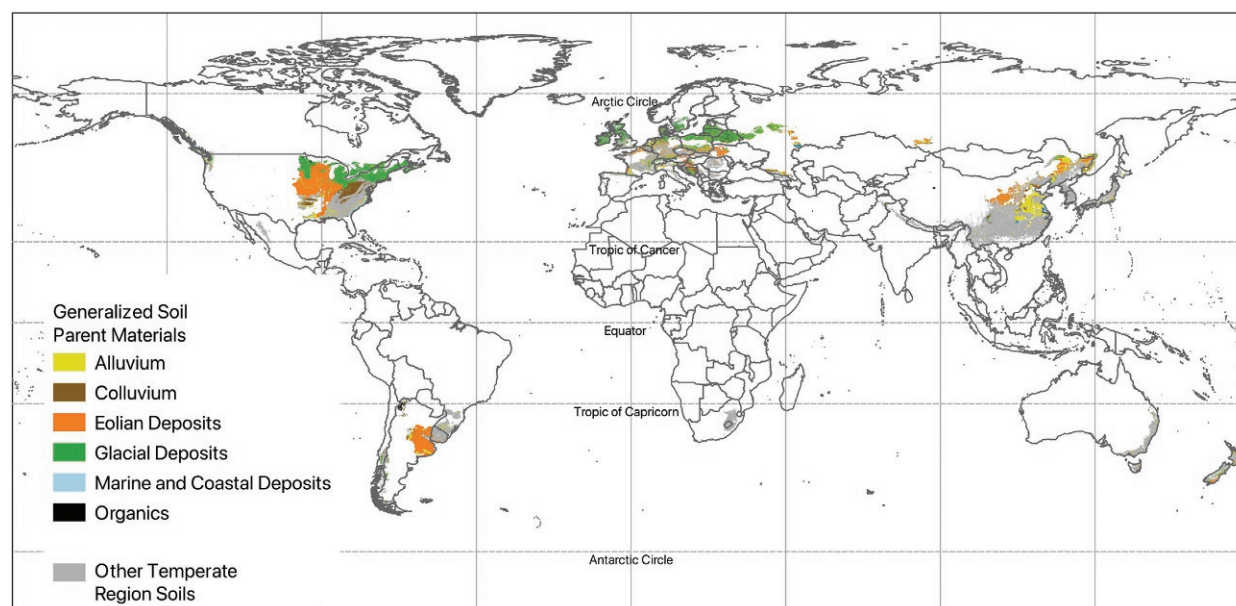


Fig. 2 Distribution of glacial sediments, loess, and alluvial soil parent materials within humid cool temperate ecozone boundaries. Soil temperature and moisture regime data are from USDA-NRCS (2016) *Soil Climate Regions Map*. Soil and Plant Sciences Division: Washington, DC <https://www.nrcs.usda.gov/wps/portal/nrcs/main/soils/use/worldsoils/>, and parent material data are derived from Beck H, Zimmermann N, McVicar T et al. (2018) Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific Data* 5: 180214. doi: 10.1038/sdata.2018.214.

significant portion of the year in the higher latitudes of the HCT zone, thus sharply reducing the duration of action of many biologically- and hydrologically-driven pedogenic processes. Consequently, soils in warmer parts of the HCT region tend to be more weathered, with lower SOC stocks because of increased rates of organic matter decomposition (Crowther et al., 2019), while soils in cooler parts of the region tend to be less weathered, with higher SOC stocks due to decreased rates of organic matter decomposition.

Excess moisture: Eluviation, illuviation and the leaching of dissolved solutes

In HCT ecozones, precipitation exceeds evapotranspiration losses on an annual basis, which allows the downward movement of excess water through the soil in all but the most poorly drained conditions (where water may be ponded or stagnated). When this downward moving water brings with it soluble soil constituents, the result is a process known as leaching or eluviation. Due to excess moisture conditions, eluviation is one of the most common processes in HCT soils and occurs throughout the year during the period that soils are unfrozen. Most HCT regions experience evenly distributed annual precipitation, however in cooler parts of the HCT with sufficiently cold winters, the accumulation of snow as stored water and its sudden release in the spring can result in bursts of infiltration and leaching in the spring as the soil thaws.

Excess moisture and eluviation in HCT soils results in the dissolution and leaching of ions, dissolved organic matter, fine clays, and soluble ionic compounds (calcium and sodium carbonates (CaCO_3 and Na_2CO_3 , respectively), gypsum (CaSO_4), and more soluble salts). Depending on the average depth and extent to which excess water moves downward over long timescales, these materials then either accumulate in the subsoil (a process known as illuviation) or are removed completely from the soil to the groundwater. Any ionic salts inherited from the soil parent material are rapidly dissolved and leached under HCT conditions. Consequently, HCT soils have been completely leached of salts more soluble than calcium carbonate (CaCO_3 , lime) and, in most parts of the region, are free of calcium carbonate as well. Although of minor importance across the entire HCT region, carbonates may still be found in soils near the semi-arid border of the region, in some isolated, younger, fine-textured soils where restrictive subsoils have not allowed downward leaching over sufficient time periods, and in the rims of wetland basins near the semi-arid border that receive discharge flow.

There are a small number of dominant cations (positively charged ions) that account for the majority of dissolved or adsorbed (attached by electrostatic attraction to soil clays and organic matter) ions in soils. Of these, H^+ (hydrogen) and Al^{3+} (aluminum) are considered acidic cations (Al^{3+} is considered acidic because when released into solution it hydrolyzes, generating additional H^+) and Na^+ (sodium), K^+ (potassium), Ca^{2+} (calcium), Mg^{2+} (magnesium) are considered non-acidic ('base') cations, because they do not hydrolyze when released into solution and thus do not generate additional H^+ (Foth and Ellis, 1997). All else being equal, soils with a greater proportion of acidic cations will have lower pH values ($\text{pH} < 6$), and those with a greater proportion of non-acidic will be closer to neutral ($\text{pH} 6\text{--}7$).

Natural topsoil pH values in the HCT region are generally moderately acidic to neutral, ranging from pH values of 4.5–7 (Poggio et al., 2021). Under the influence of excess moisture, non-acidic cations are lost deeper in the subsoil or completely to the groundwater as acid generating processes from carbonic acid in rainwater, organic matter decomposition, and plant root ion exchange contribute H^+ to the soil. Over time, these H^+ ions undergo exchange with adsorbed non-acidic cations which are lost downwards through leaching, and the topsoil is acidified. Depending on the extent and intensity of leaching, this acidification process can reach great depths, penetrating deep into the subsoil given enough effective moisture and time. This means that the topsoils in HCT regions are, as a rule, more acidic than their underlying subsoils. Because carbonates are the major contributor of OH^- to soil solutions and are responsible for raising soil pH above 7, true alkaline soil conditions ($\text{pH} > 7$) in HCT soils are rare except in special situations described above.

The downward movement of fine clays due to leaching and eluviation and their deposition in the subsoil as laminar coatings on the surfaces of soil aggregates (a morphological phenomena known as 'clay films' or 'argillans' and a process known as 'argilluviation'), is one of the major hallmarks of HCT soils (Bockheim and Hartemink, 2013). Nearly all HCT soils (except those in geologically young and undeveloped deposits of alluvium or colluvium) have clay enriched subsoils that have formed due to argilluviation. This often results in a 'peak' in clay content in the subsoil before the clay content decreases again as the unaltered parent material is approached. This phenomenon is widespread in HCT soils, and these soils have assumed such a central place in the conceptual framework of soil genesis that it is often assumed that all soils become 'clayier' with depth.

A final process associated with excess moisture and eluviation occurs under conditions of intense and rapid infiltration in soils with acidic organic horizons at the surface. Under these conditions, dissolved organic acids can chelate iron oxides that coat soil minerals and transport them to the subsoil where they can accumulate. This process, known as podzolization, results in distinct morphological properties, forming a white or grey leached layer (an 'E' or 'eluviated' horizon) just below the soil surface where iron oxide nanoparticles have been stripped from mineral grains. Below this eluviated zone is often a dark or bright colored zone of illuviation, where dissolved organic matter and chelated iron oxides have been deposited in the subsoil.

Groundwater recharge and wetlands in the humid cool temperate zone

Mineral wetlands, bogs, and fens are common components of many HCT ecozones due to excess moisture, which leads to groundwater recharge and saturation in low-lying areas. Mineral wetlands are particularly common in recently glaciated regions

due to the lack of well-developed dendritic surficial drainage networks and the presence of numerous closed depressions. These wetlands typically form in depressions where runoff from higher elevations collects, particularly if the depressions have restricted drainage or if they intersect the groundwater table. Furthermore, they usually undergo periodic saturation of the soil during the growing season. In non-glaciated parts of the HCT, or in areas of higher relief, mineral wetlands tend to be present only along the margins of surface waters (lakes, streams).

Where saturation is more permanent and the rate of organic matter decomposition is depressed, organic materials accumulate and, if thick enough, can form peat. Although peatlands are not as prevalent across the HCT region as they are in boreal regions or in the vast lowlands of the equatorial tropics, they are of regional importance in many areas. They are also prevalent on low-gradient coastal plains where the groundwater table is near to the soil surface. Ombrotrophic bogs form in areas where groundwater inputs to the organic layers are minimal and nearly all nutrients are supplied by precipitation and inputs from surface runoff. Over time the surface elevation of the center of the bog will increase due to the formation of peat at the surface, producing a raised dome that is not affected by runoff or groundwater inputs. Bogs have low floristic diversity (mosses, adapted tree and shrub species, few forbs) due to their poor nutrient status. Fens have significant groundwater inputs and, therefore, higher nutrient inputs and greater floristic diversity (grasses (*Poa* sp.), sedges (*Carex* sp.), ericaceous shrubs, and a wider range of forbs and herbs) than ombrotrophic bogs. Fens may also form on mild to steep slopes where seepage contributes a steady flow of water downslope causing saturation at the soil surface. Organic horizons tend to form in this saturated environment and, if thick enough, are described as 'hanging fens.' Although not as common, blanket peatlands (extensive areas of continuous peat across the land surface) can form on low relief landscapes in the northern portions of the HCT region, such as in the northern United States, Scotland, and other HCT areas of northern Europe (Xu et al., 2018).

In addition to the process of organic material accumulation, oxidation and reduction (redox) reactions are the major biogeochemical process of importance in HCT wetland soils. Broadly, redox reactions are defined as any chemical reactions that involve the transfer of electrons between two chemical species. In soils, the most prominent redox reaction under aerobic (oxygen-rich) conditions is aerobic respiration by microbes, which oxidizes organic molecules to CO_2 (carbon dioxide) gas; in turn, oxygen (O_2) is used as a terminal electron acceptor and is reduced to water vapor (H_2O). This is an exothermic process resulting in net energy output to the microorganism, the same function that aerobic respiration serves in macro-organisms such as humans. However, when oxygen becomes depleted in soil due to saturated conditions (oxygen diffuses 10,000 times slower through water than through air), microorganisms in soil must use alternate electron acceptors as they continue to decompose organic molecules for energy. After oxygen, the next most favorable electron acceptor is nitrate (NO_3^- , which is reduced to N_2 (nitrogen) gas in a process known as denitrification), followed by manganese (Mn), and then ferric iron (Fe^{3+}). Depending on prevailing temperatures and soil chemical parameters, which can vary in both space and time, it takes approximately 2 weeks to 1 month of saturation for microorganisms to exhaust available oxygen, nitrate, and manganese and begin utilizing iron as an alternative electron acceptor (Fiedler and Sommer, 2004).

Iron is particularly important because it is highly abundant in most soils (2–8% by weight) and oxidized, ferrous iron compounds (iron oxyhydroxides) are responsible for the brown, orange, red, and yellow earth colors—caused by iron oxide nanoparticles—that coat the surfaces of otherwise drab mineral grains. When microorganisms use iron as a terminal electron acceptor in aerobic respiration, ferric iron (Fe^{3+}) is reduced to ferrous iron (Fe^{2+}). Except in unique conditions of high ferrous iron concentrations, ferrous iron does not form stable solid compounds as does ferric iron, so once reduced, the iron simply acts as a dissolved ion and does not affect soil color. The iron oxide nanoparticle 'paints,' which give soils the distinctive earth tone colors, are dissolved, leaving the underlying mineral grains uncoated. Because the most dominant soil minerals are grey in color, this process leaves the soil with a grey or dull appearance. If oxygen is reintroduced into the soil following a period of saturation, any remaining ferrous iron will rapidly and abiotically re-oxidize to ferric iron, re-forming iron oxide nanoparticles, but the reoxidation is never homogeneous. It leaves a splotchy, mottled pattern of brightly colored iron oxide concentrations and dull colored depletions. Collectively, these are known as redoximorphic features (mottling related to the biochemical reduction of iron under conditions of saturation, and the abiotic reoxidation of iron under fluctuating saturated/unsaturated conditions), and they serve as major morphological indicators of soil wetness timing, duration and the seasonal fluctuation of the saturated zone.

A predominance of youthful, transported parent materials

Soil 'age' refers to the time since exposure of a geomorphic surface or soil parent material and, as such, is a measure of the length of time a soil or landscape has been affected by agents that drive pedogenic development in HCT environments. Some landscapes in HCT ecoregions are composed almost entirely of (geologically) young parent materials formed from processes such as continental glaciation, loess deposition, and glacio-fluvial transport (Ehlers et al., 2011; Börker et al., 2018). Other areas are affected by other rejuvenating processes such as alluvial or colluvial deposition. Overall, more than a third of temperate soils are derived directly from recent glacial deposits, loess or alluvium (Fig. 2). If these criteria were expanded to include materials deposited in early and mid-Pleistocene glaciations, thin loess caps that do not comprise the entire parent formation and minor alluvial deposits, this proportion would likely reach beyond 50%, explaining the wide prevalence of high natural soil fertility in the temperate region.

The HCT soils are derived from a broad array of parent materials, ranging from glacially-derived sediments and wind-blown loess, to residuum, alluvium, and colluvium. Despite this wide diversity in parent materials, some generalities can be observed. In the colder parts of temperate regions, particularly in the northern hemisphere, a significant proportion of soil parent materials

were derived either directly or indirectly from processes associated with late Quaternary glaciations (Börker et al., 2018). Many of these soil landscapes formed on glacial drift, glaciofluvial, or glaciolacustrine sediments. Extra-glacial areas near glacial margins were strongly affected by periglacial processes and may also have experienced extensive wind erosion or fluvial re-working. Still other HCT landscapes, often far removed from the glacial margins, were blanketed by varying thicknesses of loess deposits or eolian materials derived mainly from riverine sediment sources associated with the retreat of continental ice sheets (Li et al., 2020). These materials are geologically young and relatively unweathered, with a higher proportion of non-acidic cations and a mineral assemblage containing significant amounts of primary minerals.

In warmer parts of the region, in landscapes not affected by recent glaciation or loess deposition, soils have formed mainly on alluvium (parent materials transported by streams and rivers), colluvium (the material that has accumulated at the base of hillslopes through creep, slope wash, and erosion), or residuum (soil parent material developed from underlying bedrock without undergoing off-site transport) (Börker et al., 2018). Alluvial deposition may result in the abrupt stratification of materials with differing chemical and physical properties and the burial of topsoil horizons, resulting in irregular decreases of organic carbon in some alluvial soils. In contrast, soils formed in colluvium are typically well mixed, poorly developed, and may have a high percentage of coarse fragments. The physical and chemical properties of residual soils (formed from the physical and chemical weathering of the underlying bedrock, also known as residuum) are initially closely linked to those of the underlying bedrock and vary with local to regional lithology. For example, soils derived from fine-grained sedimentary rocks can inherit clay minerals and trace metals (Dere et al., 2013), and calcareous rocks may initially result in soils with higher pH and calcium concentrations. As soils develop in HCT regions over time, however, their chemical and physical properties deviate increasingly from their parent bedrock. In particular, coarse-grained granites and metamorphic rocks (e.g., schists) will give rise to secondary clay minerals and oxides of Al and Fe.

Progressive mineral weathering, alteration, and transformation

Initially, soil mineralogical properties in HCT soils strongly reflect the parent materials from which they form. Mineral assemblage fractions in relatively young soils contain a high proportion of primary minerals such as framework silicates (feldspars, quartz), chain silicates (pyroxene and amphibole), and primary phyllosilicates (micas such as biotite and muscovite). Soils can also inherit secondary phyllosilicates (such as illite or smectite) or other mineral constituents such as soluble salts (e.g., CaSO_4 , calcium sulfate), carbonates (e.g., CaCO_3), phosphates (e.g., CaPO_4) or sulfides (e.g., FeS). In the soil environment, these primary minerals are exposed to more free water, higher oxygen concentrations, and lower temperatures and pressures than the conditions under which they were formed in the earth's upper mantle. This results in thermodynamic instability, and mineral weathering begins. The two major pathways of progressive mineral weathering in soils are transformation (the re-arrangement of bonds in mineral structures without complete dissolution) and neoformation (the in-situ formation of nano-crystalline minerals from the previously dissolved weathering products in the soil water) (Wilson, 1999).

As soils age in HCT environments, mineral weathering through transformation is progressive and occurs by hydration of oxygen bonds, acid-driven dissolution, and oxidation of redox sensitive elements (e.g., Fe^{2+} to Fe^{3+}). Although these are often conceptualized as geochemical processes, they are actually more appropriately biogeochemical processes—driven in large part by the acidity and organic ligands produced by plant litter decomposition, plant root ions and gas exchange, and the microbially-mediated dissolution of minerals. In HCT environments, any salts or ionic compounds are dissolved relatively quickly and removed from the soil over timescales ranging from centuries to millennia. Over longer timescales (thousands to tens of thousands of years), these mechanisms transform inherited mineral assemblages from less thermodynamically favored primary silicates to moderately stable 2:1 phyllosilicates such as illites, smectites and vermiculites. Finally, over timescales of tens of thousands to hundreds of thousands of years, the further alteration of these mineral assemblages results in large accumulations of the most thermodynamically stable secondary minerals, kaolinite and iron and aluminum oxides.

The neoformation of nano-crystalline, clay-sized mineral particles occurs through precipitation of mineral species from the dissolved weathering products of other minerals. There are two major pathways of phyllosilicate neoformation. The first leads to the neoformation of 2:1 expanding phyllosilicates such as smectite, and is favored in semi-arid conditions with large base cation concentrations and seasonal wet-dry cycles. The second leads to the neoformation of the stable 1:1 phyllosilicate kaolinite, and is favored by well drained acidic conditions, where base cation concentrations are lower and silicic acid and aluminum dominate the dissolved ions. In HCT soils, due to conditions favoring leaching and acidification, this second mechanism is the most dominant, meaning that the major pathway of clay mineral neoformation in HCT soils leads to kaolinite (Wilson, 1999).

Therefore, in well-drained HCT soils, initially contrasting parent materials can end up with similar mineral assemblages through both transformation and neoformation processes operating over sufficiently long periods of time. Although all HCT soils lie somewhere on this progressive mineral alteration trajectory, there is wide variability in HCT mineral assemblages due to differences in parent material, age and climatic factors. In the cooler, higher latitude regions of the HCT, where soils are more likely to be formed in more recently transported or glacially derived parent materials of ages ranging from thousands to tens of thousands of years, mineral assemblages are likely to be mixed, with the clay fraction dominated by 2:1 phyllosilicates such as illite, vermiculite, and smectite. In contrast, in lower latitude, warmer areas of the HCT region, where the soils of broad ridgetops and interfluvies have

been weathering in places for tens of thousands to hundreds of thousands of years, kaolinite is often the dominant phyllosilicate mineral in the clay fraction together with iron and aluminum oxides such as hematite and gibbsite (Dere et al., 2013). These broad trends are tempered by significant local variability, depending on parent material composition and drainage class. On balance, HCT soil mineral assemblages represent a middle ground between the intensive weathering processes that occur in the tropics and subtropics, and the much more limited weathering that occurs in cold and arid climates.

The temperate plant-soil interface: Impacts of forest and humid grasslands

Soils and the plant assemblages that grow on them are co-dependent and generate predictable patterns in soil physical and chemical properties, which are a key part of the classical 'organisms' state factor of soil formation. The majority of the native vegetation in HCT ecoregions is dominated by deciduous and mixed deciduous-coniferous forests, followed by humid grasslands and wetland plant communities (Hengl et al., 2018). Other ecosystem types (such as temperate evergreen rainforests) are locally important but are less globally dominant in HCT ecozones. Biomes bounding this region are chaparral or desert scrub ecosystems in more arid regions; annual short- and mid-grass prairies and savannas in Mediterranean regions; coniferous forests in boreal regions; and broadleaf evergreen forests or savannas in the tropics and subtropics.

Deciduous and mixed deciduous-coniferous forests have roots that are coarser, distributed closer to the soil surface, and undergo a much slower rate of root turnover than ecosystems dominated by grasses (Gill and Jackson, 2000). Therefore, the majority of organic inputs in forests are from leaf litterfall and root decomposition, which predominantly occur at or near the soil surface. The incorporation of surface leaf litter and organic matter in forest soils thus requires physical mixing by soil invertebrates such as insects, earthworms, and gastropods. Where leaf materials are not incorporated, the soil typically develops a layer of partially decomposed organic materials at the soil surface, most often termed the forest floor or litter layer, which ranges in composition from nearly intact (fibrous) materials (Oi horizons), to intermediately decomposed hemic (Oe) materials and well decomposed sapric (Oa) materials. These organic (O) horizons play a critical role in seedling recruitment, and serve as a key germination site for seedlings and understory plants. Moreover, the acidity and production of organic ligands by these organic materials can drive the formation of leached layers (E horizons) immediately below them, formed by the eluviation of clays and the chelation and leaching of iron, manganese, and aluminum (hydr)oxides and other coatings from the surface of silt and clay grains in the soil matrix. If earthworms or other invertebrates are present and incorporate the forest floor into the underlying mineral soil, then a relatively thin, light-colored mineral topsoil (ochric epipedon) is observed at the surface. Introduction of non-native and invasive earthworm species into forested landscapes in the northern US has all but eliminated the thick O horizons once common in many of these North American HCT ecozone forests (Hale et al., 2005; Chang et al., 2021).

Soils formed under coniferous forests or mixed deciduous/coniferous forests in HCT regions with coarse, sandy parent materials may have subsoil characteristics resulting from podzolization processes (Lundström et al., 2000). Coniferous leaf litter is more acidic than that of most broadleaf deciduous species, and releases large quantities of dissolved organic acids when decomposed and leached. These organic compounds can chelate Fe, Mn, and Al from metal (hydr)oxide coatings on the soil grains. In relatively coarse-textured soils, the organic acid-metal chelates are transported with the soil solution as long as they remain soluble, producing a distinct, light colored E horizon where surficial metal (hydr)oxides have been removed. The chelated metals are deposited and accumulate lower in the profile due to changes in pH, increases in calcium or clay contents, or other factors, producing distinctive red or black horizons of accumulation of organic matter and Fe, Mn, and Al chelates. The illuvial accumulation of organic matter, iron and aluminum in the subsoil is defined as a spodic diagnostic horizon for the purposes of soil classification (Soil Survey Staff, 2014).

In contrast, soils formed predominantly under deciduous forests in HCT regions with finer parent material display prominent clay accumulation in the subsoil due to the process of argilluviation. Broadleaf deciduous leaf litter is not as acidic as coniferous leaf litter, and therefore true podzolization rarely occurs under broadleaf deciduous vegetation. The downward transport of fine clays is amplified by stemflow phenomena common to deciduous forests, which may concentrate a high proportion of the precipitation falling on a tree directly under its bole, thus producing a localized effect similar to increased precipitation. Therefore, diagnostic subsurface horizons generated by transport of clays or iron oxides are commonly observed. Clay-enriched subsoil horizons dominated by 2:1 phyllosilicates or mixed clay mineral assemblages are defined as argillic horizons in Soil Taxonomy (argic horizons in WRB) and are some of the most common subsoil horizons in the HCT region. A second diagnostic type of clay enriched subsoil formed by argilluviation are the kandic horizons (also argic horizons in WRB), dominated by 1:1 clay minerals such as kaolinite. Kandic horizons tend to be present in soils formed on acidic parent materials, or in more highly-weathered soils in warmer regions of the HCT region (Shaw et al., 2010).

Landscapes of low relief at the drier limits of the HCT regions are often characterized by humid grasslands (locally known as prairies, steppes, Pampas or veld), which are dominated by tallgrass species. Because excess moisture in HCT ecozones favors forest growth, these humid grasslands thrive on fire or grazing disturbance to outcompete the forests, and grasslands can withstand and more readily regenerate from frequent fires than most forests. Fires that start in low relief landscapes are often windswept and can cover very large areas, whereas higher relief landscapes tend to slow or stop the spread of fires.

The formation of thick, dark, organic rich mineral A horizons (mollic epipedons in both US Soil Taxonomy and WRB) is common under humid grassland or humid savanna type vegetation type with a grass-dominated understory (Thorp, 1948). Grasses

have deep fibrous root systems and a rapid rate of root turnover relative to forests (Gill and Jackson, 2000) which, in combination, produce significant inputs of organic matter at depth; no translocation is required to move the organic matter deep into the solum. In addition, the efficient base-cycling of grassland ecosystems leads to a large concentration of non-acidic cations, increasing pH to near-neutral levels in climates where more extensive acidification would be expected under forest vegetation. Subsoil horizon development can be minimal in many recently glaciated grassland soils, which may have no subsoil development at all or have subsoil horizons that have undergone only a moderate degree of development in color or structure relative to the parent material. Other older, more well-developed humid grassland soils may have argillic horizons, but these subsoil horizons are not nearly as common under grasslands as they are under forest vegetation. Several factors interact to inhibit fine clay dispersal and movement in grassland soils, including the presence of organic coatings on mineral grains in the topsoil that tend to bind clays together. Exchangeable Ca contents and pH are also generally higher in grasslands, both of which limit clay dispersion and promote its flocculation.

Classification of humid cool temperate soils

Both U.S. Soil Taxonomy (Soil Survey Staff, 2014) and the World Reference Base for Soil Resources (WRB, UN-FAO, 2015) differentiate HCT soils at the highest levels based on major developmental pathways related to predominant vegetation, climatic and parent material influences on pedogenic processes during soil development. The dominant soil groups in HCT regions are Inceptisols (WRB: Cambisols) and Alfisols (WRB: Luvisols and Alisols), which comprise 23.6% and 20.7% of the total land area, respectively (Table 1). The next most common HCT soil groups are Ultisols (WRB: Acrisols and Lixisols) (16.4%), Mollisols (15.0%) (WRB: Chernozems), Entisols (WRB: Regosols) (13.9%), and Spodosols (WRB: Podzols) (4%). All other soil orders or reference groups make up less than 10% of the region (USDA-NRCS, 2005).

Alfisols (high-base status soils with clay-enriched argillic subsurface horizons), Ultisols (low base status soils with clay-enriched kandic subsurface horizons) and Spodosols (soils with organic matter and iron/aluminum oxide enriched subsurface horizons) are associated mainly with forested landscapes and Mollisols (soils with thick, carbon rich topsoils) mainly with grasslands. Alfisols are more common under broadleaf deciduous forests in cooler, drier climates, and younger landscapes where weathering, leaching, and removal of bases is not as extensive. Ultisols occur under broadleaf deciduous forests in warmer, wetter, and more stable landscapes of the region where soils are more weathered and acidic cations dominate. Spodosols occur throughout the HCT region under coniferous forest vegetation in coarse-textured parent materials.

Entisols (morphologically undeveloped soils) and Inceptisols (minimally developed soils that have some color or structural change in the subsoil) are interspersed throughout the HCT region at erosional or depositional landscape positions, in particular shoulder and back slopes where soil development is limited or where new parent materials have been exposed or deposited by geomorphic processes, such as on floodplains. Entisols and Inceptisols are also common in wetter portions of the landscape where soil development processes are slowed due to anaerobic conditions. Histosols (peatland soils formed in large deposits of organic materials) occur on poorly-drained landscape positions, particularly in glaciated regions, where closed depressions abound.

Table 1 Areal extent and proportion of ice-free land surface for soil orders (Soil Survey Staff, 2014) and approximate WRB Soil Reference Groups (IUSS Working Group WRB, 2015) worldwide and in humid-cool temperate regions.

Soil Order (WRB Reference Group)	Extent		Proportion	
	Earth's Surface (10 ⁶ km ²)	Temperate Regions (10 ⁶ km ²)	Earth's Ice-Free Land Surface (%)	Temperate Ice-Free Land Surface (%)
Inceptisols (Cambisols)	12.8	5.3	9.8	23.6
Alfisols (Luvisols, Albeluvisols)	12.6	4.7	9.7	20.7
Ultisols (Acrisols, Alisols)	11.1	3.5	8.5	16.4
Mollisols (Chernozems, some Phaeozems)	21.1	3.1	6.9	15.0
Entisols (Regosols)	9.0	3.3	16.2	13.9
Spodosols (Podzols)	3.4	0.8	2.6	4.0
Vertisols (Vertisols)	3.2	0.6	2.4	2.7
Andisols (Andosols)	0.9	0.2	0.7	1.0
Oxisols (Ferralsols)	9.8	0.2	7.5	1.0
Aridisols (Gypsisols, Durisols, some Calcisols and Kastanozems)	15.7	0.2	12.0	<0.1
Histosols (Histosols)	1.5	0.1	1.2	<0.1
Gelisols (Cryosols)	11.3	—	8.6	—
Shifting Sands	5.3	0.1	4.1	<0.1
Rock	12.1	10.0	10.0	<0.1

Soil order distribution data from USDA-NRCS (2005) *Global Soil Regions Map*. Based on FAO-UNESCO, *Soil Map of the World*, Digitized by ESRI. Soil Science Division, World Soil Resources: Washington, DC. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/use/worldsoils/?cid=nrcs142p2_054010.

Impacts of land use and management on humid cool temperate soils

Humans have had a profound influence on HCT soils, both directly through agricultural and forestry activities and indirectly through the alteration of global biogeochemical cycles. Although temperate region soils (as defined above) make up only 16% of the world's ice-free terrestrial surface, roughly 1/4 (27%) of the world's croplands lie in HCT regions, and even more of the area is under pasture or other food, animal feed, and fiber production (Yu et al., 2020). Many of the world's most intensively managed forestlands also occur in HCT ecozones (Curtis et al., 2018). This is due, in large part, to the favorable properties of HCT soils and the favorable environments in which they occur.

At regional scales, the long-term human agricultural use of soils in HCT regions associated with the addition of organic materials through heavy manuring has resulted in phosphorus-rich topsoils (plaggen epipedons), which are typically higher in organic matter and absolute elevation (due to upbuilding through the addition of organic materials) than adjacent soils. In North America, Indigenous peoples of the Mississippi River Valley utilized extensive areas of HCT soils for crop production and earth-building activities (Woods, 2004). In the past century, HCT soils have also undergone extensive change at larger scales due to the high density of intensive, mechanized agriculture. Agricultural tillage has caused accelerated erosion on many HCT landscapes through the net downslope movement of soil material disturbed by tillage implements and by leaving soil surfaces poorly protected against raindrop impact. This accelerated erosion can increase the variability of soils across these landscapes; for example, in some parts of the United States Corn Belt the erosion of material from upper slope positions and decomposition of soil organic carbon has resulted in a change from mollic to ochric epipedons, which effectively results in the co-dominance of Alfisols or Inceptisols and Mollisols where Mollisols were once the primary soil order (Jelinski and Yoo, 2016). The fact that these changes have taken place over a time period of less than 200 years in the United States accentuates the tremendous capacity of human activities to change soil properties. At still larger scales, the indirect effects of industrialization, fossil fuel burning (which produces atmospheric deposition of both nitric and sulfuric acid) and the widespread production and application of nitrogenous fertilizers has resulted in the anthropogenically accelerated acidification of many soils in HCT regions.

The artificial drainage of wetlands or seasonally wet soils has been a widespread practice in intensively cropped soils in the HCT zone. The installation of artificial drainage (including both ditching and tile drainage) is most prevalent on a global scale across HCT regions in the northern Hemisphere (Feick et al., 2005), but tile drainage in particular is most prevalent in the flat, glacially modified landscapes dominated by fine textured soils in the corn belt of the central United States (Valayamkunnath et al., 2020) and western Europe (areas such as eastern Denmark (Motarjemi et al., 2020)). Widespread drainage can significantly affect local hydrology and, together with agricultural tillage, can increase aeration of the solum which may enhance the rate of organic matter oxidation and release of soil carbon to the atmosphere in the form of CO₂. The magnitude of this process and its impact on the global carbon budget are, however, currently topics of intense debate.

The HCT forests cover 1.04 billion hectares or 25% of the world's forested areas (Nabuurs et al., 2007) and extensive areas of forest have been replaced by agricultural or urban land uses across North America, Europe, and Asia, particularly between 1800 and the early 1900s. The global logging rate for temperate forests has increased (2–6 million ha yr⁻¹) but is more constant than the dramatic increase in logging in tropical forests from 0.3 but to over 8 million ha yr⁻¹ (Runyan and D'Odorico, 2016) due to intensive, cyclical harvest management in HCT zones. Deforestation and agroforestry directly affect the morphology of forest soils by diminishing surface organic materials on the forest floor and organic matter inputs from roots. Moreover, the conversion of forest impacts the biogeochemical cycling of C, N and other elements tied with tree uptake and storage in HCT forest soils.

Conclusion

The soils of humid cool temperate regions, although highly diverse, are predominantly formed in transported parent materials under forest, grassland, and wetland vegetation. Because they tend to be highly fertile and occur in favorable climates, these soils have been significantly altered by human activities and play a large role in global food security and the global carbon cycle. The sustainable management of these soils depends on our understanding of the processes responsible for their formation and the accurate prediction of contemporary pedogenic trajectories.

References

- Bockheim JG and Hartemink AE (2013) Distribution and classification of soils with clay-enriched horizons in the USA. *Geoderma* 209: 153–160.
- Bockheim JG, Gennadiyev AN, Hammer RD, and Tandarich JP (2005) Historical development of key concepts in pedology. *Geoderma* 124(1–2): 23–36.
- Börker J, Hartmann J, Amann T, and Romero-Mujalli G (2018) Global unconsolidated sediments map database v1.0 (shapefile and gridded to 0.5° spatial resolution). *PANGAEA*. <https://doi.org/10.1594/PANGAEA.884822>.
- Chang CH, Bartz MLC, Brown G, et al. (2021) The second wave of earthworm invasions in North America: Biology, environmental impacts, management and control of invasive jumping worms. *Biological Invasions* 23: 3291–3322. <https://doi.org/10.1007/s10530-021-02598-1>.
- Chorover J, Kretzschmar R, Garica-Pichel F, and Sparks DL (2007) Soil biogeochemical processes within the critical zone. *Elements* 3(5): 321–326. <https://doi.org/10.2113/gselements.3.5.321>.
- Crowther TW, van den Hoogen J, Wan J, Mayes MA, Keiser AD, Mo L, Averill C, and Maynard DS (2019) The global soil community and its influence on biogeochemistry. *Science* 365(6455), eaav0550. <https://doi.org/10.1126/science.aav055031439761>.

- Curtis PG, Slay CM, Harris NL, Tyukavina A, and Hansen MC (2018) Classifying drivers of global forest loss. *Science* 361(6407): 1108–1111. <https://doi.org/10.1126/science.aau3445>. 30213911.
- Davidson E and Janssens I (2006) Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* 440: 165–173. <https://doi.org/10.1038/nature04514>.
- Dere A, White TS, April RH, Reynolds B, Miller TE, Knapp EP, McKay L, and Brantley SL (2013) Climate dependence of feldspar weathering in shale soils along a latitudinal gradient. *Geochimica et Cosmochimica Acta* 122: 101–126.
- Ehlers J, Gibbard PL, and Hughes PD (2011) *Quaternary Glaciations—Extent and Chronology: A Closer Look*. Amsterdam/Boston: Elsevier.
- Feick S, Siebert S, and Döll P (2005) *A Digital Global Map of Artificially Drained Agricultural Areas*. Frankfurt Hydrology Paper 04 Institute of Physical Geography, Frankfurt University.60. https://www.uni-frankfurt.de/45217762/FHP_04_Feick_et_al_2005.pdf.
- Fick SE and Hijmans RJ (2017) WorldClim 2: New 1km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* 37(12): 4302–4315.
- Fiedler S and Sommer M (2004) Water and redox conditions in wetland Soils—Their influence on pedogenic oxides and morphology. *Soil Science Society of America Journal* 68: 326–335. <https://doi.org/10.2136/sssaj2004.3260>.
- Foth HD and Ellis BG (1997) *Soil Fertility*, 2nd edn USA: Lewis CRC Press LLC290.
- Gill RA and Jackson RB (2000) Global patterns of root turnover for terrestrial ecosystems. *New Phytologist* 147: 13–31. <https://doi.org/10.1046/j.1469-8137.2000.00681.x>.
- Hale CM, Frelich LE, Reich PB, et al. (2005) Effects of European earthworm invasion on soil characteristics in northern hardwood forests of Minnesota, USA. *Ecosystems* 8: 911–927. <https://doi.org/10.1007/s10021-005-0066-x>.
- Hengl T, Walsh MG, Sanderman J, Wheeler I, Harrison SP, and Prentice IC (2018) Global mapping of potential natural vegetation: An assessment of machine learning algorithms for estimating land potential. *PeerJ* 6: e5457. <https://doi.org/10.7717/peerj.5457>.
- IUSS Working Group WRB (2015) *World Reference Base for Soil Resources 2014, Update 2015. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*. World Soil Resources Reports No. 106 Rome: FAO.
- Jelinski NA and Yoo K (2016) The distribution and genesis of eroded phase soils in the conterminous United States. *Geoderma* 279: 149–164. <https://doi.org/10.1016/j.geoderma.2016.05.015>.
- Li Y, Shi W, Aydin A, Beroya-Eitner MA, and Gao G (2020) Loess genesis and worldwide distribution. *Earth-Science Reviews* 201: 102947.
- Lundström US, van Breemen N, and Bain DC (2000) The podzolization process. A review. *Geoderma* 94(2–4): 91–107. [https://doi.org/10.1016/S0016-7061\(99\)00036](https://doi.org/10.1016/S0016-7061(99)00036).
- Motarami SK, Möller AB, Plauborg F and Iversen BV (2020) Predicting tile drainage discharge using machine learning.
- Nabuurs GJ, Masera O, Andrasco K, Benítez-Ponce P, Boer R, Dutschke M, Elsidig E, Ford-Robertson J, Frumhoff P, Karjalainen T, Kränkka O, Kurz WA, Matsumoto M, Oyhanbcal W, Ravindranath NH, Sanz Sanchez MJ, and Zhang X (2007) Forestry. In: Metz B, Davidson OR, Bosch PR, Dave R, and Meyer LA (eds.) *Climate Change 2007: Mitigation. Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge/New York, NY: Cambridge University Press.
- Poggio L, de Sousa LM, Batjes NH, Heuvelink GBM, Kempen B, Ribeiro E, and Rossiter D (2021) SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *The Soil* 7: 217–240.
- Runyan C and D'Odorico P (2016) *Global Deforestation*. Cambridge University Press.
- Shaw JN, Hajek BF, and Beck JM (2010) Highly weathered mineralogy of select soils from Southeastern U.S. Coastal Plain and Piedmont landscapes. *Geoderma* 154: 447–456.
- Simonson RW (1999) Origin and acceptance of the term pedology. *Soil Science Society of America Journal* 63: 4–10. <https://doi.org/10.2136/sssaj1999.03615995006300010002x>.
- Soil Survey Staff (2014) *Keys to Soil Taxonomy*, 12th edn Washington, DC: USDA-Natural Resources Conservation Service.
- Thorpe J (1948) How soils develop under grass. In: *Yearbook of Agriculture, USDA*, pp. 55–66. Washington, DC: U.S. Govt. Printing Office.
- USDA-NRCS (2005) Global soil regions map. In: *Based on FAO-UNESCO, Soil Map of the World, digitized by ESRI*. Soil Science Division, World Soil Resources: Washington D.C. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/use/worldsoils/?cid=nrcs142p2_054010.
- USDA-NRCS (2016) *Soil Climate Regions Map*. Washington, DC: Soil and Plant Sciences Division. <https://www.nrcs.usda.gov/wps/portal/nrcs/main/soils/use/worldsoils/>.
- Valayamkunnath P, Barlage M, Chen F, et al. (2020) Mapping of 30-meter resolution tile-drained croplands using a geospatial modeling approach. *Scientific Data* 7: 257. <https://doi.org/10.1038/s41597-020-00596-x>.
- West PC, Gibbs HK, Monfreda C, Wagner J, Barford CC, Carpenter SR, and Foley JA (2010) Trading carbon for food: Global comparison of carbon stocks vs. crop yields on agricultural land. *Proceedings of the National Academy of Sciences, USA* 107(46): 19645–19648. <https://doi.org/10.1073/pnas.1011078107>. 21041633. PMC2993385.
- Wilson MJ (1999) The origin and formation of clay minerals in soils; past, present and future perspectives. *Clay Minerals* 34(1): 7–25.
- Woods WI (2004) Population nucleation, intensive agriculture, and environmental degradation: The Cahokia example. *Agriculture and Human Values* 21: 255–261. <https://doi.org/10.1023/B:AHUM.0000029398.01906.5e>.
- Xu J, Morris PJ, Liu J, and Holden J (2018) PEATMAP: Refining estimates of global peatland distribution based on a meta-analysis. *Catena* 160: 134–140.
- You L, Wood S, Wood-Sichra U, and Wu W (2014) Generating global crop distribution maps: From census to grid. *Agricultural Systems* 127: 53–60. <https://doi.org/10.1016/j.agsy.2014.01.002>.
- Yu Q, You L, Wood-Sichra U, Ru Y, Joglekar AKB, Fritz S, Xiong W, Lu M, Wu W, and Yang P (2020) A cultivated planet in 2010—Part 2: The global gridded agricultural-production maps. *Earth Systems and Scientific Data* 12: 3545–3572. <https://doi.org/10.5194/essd-12-3545-2020>.

Grassland soils in the cool–arid–temperate ecozone (Steppe, grassland)

Mengyang You^{a,b}, Xiaobing Liu^b, Rupert Bäumler^c, and Margaret A Oliver^d, ^aSchool of Environmental Science and Engineering, Guangzhou University, Guangzhou, PR China; ^bNortheast Institute of Geography and Agroecology, Chinese Academy of Sciences, Harbin, PR China; ^cInstitute of Geography, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany; ^dDepartment of Geography & Environmental Science School of Human & Environmental Sciences, Soil Research Centre, The University of Reading, Reading, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction	299
Grassland soil formation	300
Environment of pedogenesis in grasslands of the cool–arid–temperate ecozone	301
Mineral weathering, clay translocation, and secondary carbonate mineral formation	301
Organic matter accumulation, bioturbation, and structure development	301
Dynamics of organic matter after cultivation	303
Grassland soil hydrology and its influence on pedogenesis	304
Classification of the cool–arid–temperate grassland soils	305
Chernozems–Mollisols	305
Phaeozems (Udolls and Aquolls)	306
Kastanozems–Ustolls and Borolls of the Mollisols	307
Distribution of Mollisols–Chernozems	308
Distribution in North America	308
Distribution in South America	309
Distribution in Europe	310
Distribution in Northeast China	313
Concluding comments	316
Acknowledgments	317
References	317

Abstract

Most Mollisols–Chernozems are recognized as inherently productive and fertile soils that have formed under prairie–steppe grassland ecosystems. Their characteristics result from their formation under grassland and a cool–arid–temperate climate in the continental interiors of the northern and southern hemispheres. The effective moisture, mineral weathering, clay translocation, and secondary carbonates are the main factors that determine the environment of pedogenesis in prairie–steppe grassland regions. The distinctive characteristics of Mollisols–Chernozems are a black or dark topsoil with a large organic matter content. They have experienced major land-use change from natural grassland to intensive agriculture, which has led to a decrease in soil organic carbon (SOC) content; they are the breadbaskets of the world.

Key points

- Mollisols–Chernozems are the most fertile and productive soils in the world;
- Mollisols–Chernozems develop from prairie–steppe grassland ecosystems or grasslands in the cool–arid–temperate ecozone;
- Commonly have a topsoil with a large organic matter content of and show considerable bioturbation;
- Conversion of Mollisols–Chernozems from grassland to arable land has depleted soil organic carbon.

Introduction

The grassland soils of the cool–arid–temperate ecozone, which have formed in the prairie, steppe ecosystems of the world, have very distinctive properties related to their pedogenesis. The distinctive characteristics of these grassland soils partly reflect the influence of the climate under which grassland vegetation predominates. Most soils that have formed under prairie–steppe grassland ecosystems or grasslands recently converted to agricultural use (Barthold et al., 2013) are classified as Chernozems (Russia, IUSS Working Group WRB, 2022), Kastanozems and Phaeozems (WRB), or Mollisols (Soil Taxonomy, Soil Survey Staff, 2022), and Isohumosols or Black Soils (China). The Canadian system of soil taxonomy does not specifically identify Mollisols, but the Chernozemic, Solonchic and Vertisolic Great Groups are those which most closely fit the diagnostic criteria (Liu et al., 2012; Sorokin et al., 2021). They are widely known as Chernozems or Mollisols, and in this chapter we more or less use these terms interchangeably.

Mollisols–Chernozems are identified by a thick, dark-colored, humus and base rich surface horizon (mollic epipedon) with a high base saturation. They have formed mainly on unconsolidated base rich materials, largely wind-blown, such as loess and glacial till. These soils were, and still are, crucial to the agriculture of their respective countries. Other soil orders that occur in grasslands can form in a range of other ecozones and ecosystems.

This chapter comprises much of what was in the original chapter on grassland soils and for which we acknowledge the original authors. Here, we now focus specifically on the grassland soils of the cool-arid-temperate ecozone; the rangeland soils of more arid regions and the pasture soils of wetter regions are covered in other chapters of this edition of the Encyclopedia. First, we consider the formation of grassland soils and the pedogenetic features of the cool-arid-temperate ecozone where climate plays a major role. These soils have also largely formed at the equatorial and eastern margins of the major ice sheets of the Quaternary period, which means that their parent materials are mainly unconsolidated basic materials such as loess and glacial till. The Mollisols–Chernozems are important agricultural soils worldwide and the pressures on these from intensive agriculture are considered. In the classification section we include terms from both Soil Taxonomy (Soil Survey Staff, 2022) and the World Reference Base for Soil Resources (IUSS Working Group WRB, 2022). It becomes clear that this group of soils varies considerably depending on their location, previous vegetation and local climate. Last, the widespread distribution of this group of soils is illustrated with maps and descriptions of their form in various countries and continents. The soils that formed under prairie (steppe) ecosystems were central to the pioneering work by the Russian scientist V.V. Dokuchaev in the 1870s (Kravchenko et al., 2010; Vysloužilová et al., 2015), and later by Guy D. Smith, the originator of Soil Taxonomy.

Grassland soil formation

The natural areas, where temperate grassland soils occur, are the prairies and steppes of the cool-arid-temperate ecozones of the world. From a global perspective, the distribution of grassland is largely a function of climate, particularly climatic factors that influence soil moisture availability. Effective moisture (annual precipitation minus annual potential evapotranspiration) is a useful index of those factors, although it neglects important seasonal variations in moisture supply. At the broad, continental scale, grasslands occupy regions with intermediate effective moisture, between the humid regions that support forest and arid regions dominated by desert shrub communities and vegetation covers of less than 50%. The soil–water balance in grasslands varies widely, but, in the most extensive grassland regions such as the steppes of Russia and Central Asia or the North American Great Plains, potential evapotranspiration frequently exceeds precipitation both on an annual basis and during the growing season. Large interannual variability in precipitation is common and droughts can cause a marked reduction in plant productivity as well as changes in species composition. Geologic records also indicate longer-term shifts between wetter and drier climates during the Quaternary (the last 1,700,000 years); thus, most grassland soils have formed under a range of climatic conditions and plant communities (Anderson, 2006).

Wind-blown sand and loess are probably the most common geologic materials underlying grasslands, mainly as a consequence of the climatic setting and geographical location. Eolian sand and dust were generated locally within grasslands during past periods of dry climate, for example at the ice margins in the Pleistocene, or are derived from neighboring arid regions. In both cases, the wind-blown sediment can subsequently be stabilized by grassland vegetation and preserved over long time periods (Hanberry, 2021). Grassland ecosystems also thrive on a wide range of other geologic materials, from weathered bedrock to glacial till and alluvium. At the local scale, factors other than regional climate become important in determining boundaries between grassland and other vegetation. Fire, ignited by lightning or by humans, is believed to have maintained grassland or savanna in regions where the climate would otherwise have allowed forest development (Andrade et al., 2022). Recurrent fires generally favor grasses with ground-level or underground growth points over woody plants and shrubs that have most of their biomass aboveground, and in addition the accumulation and stabilization of organic matter in prairie soils (Soong and Cotrufo, 2015).

The distinctive characteristics of grassland soils, with their thick dark-colored, humus and base rich surface horizon, in part reflect the influence of the climatic conditions under which grassland vegetation predominates. The frequent soil-moisture deficits in many grassland regions inhibit the weathering of silicate minerals and the oxidation of organic matter, and favor precipitation of secondary carbonates within the soil profile. Calcium is generally the main exchangeable cation throughout the profile. Ecologic processes also influence grassland soil development. Net primary productivity (biomass production by photosynthesis minus biomass consumed through respiration by the photosynthesizers themselves) is relatively modest in most grasslands, relative to many forest ecosystems. In many grasslands, underground plant biomass is greater than aboveground biomass, in some cases by a factor of 5 or more. As consequence, the fraction of total organic matter input that is added to the soil belowground is larger in grasslands than in forest, where most input occurs in the form of surface litter. Consistent with the relative importance of belowground biomass, burrowing activity by animals appears to be greater in grasslands than in other ecosystems, although the difference is difficult to quantify. There are many observations of remarkably extensive burrowing by earthworms, insects (especially ants and cicadas), and rodents in North American and Eurasian grasslands. Geomorphic processes that interact with pedogenesis may also differ between grasslands and other ecosystems. As noted above, eolian sand and loess deposits are widespread in grassland regions, indicating the potential for past truncation of soil profiles by wind erosion or accretionary processes by dust deposition. The latter can lead to the formation of paleosols, which provide a record of terrestrial sedimentary history (Retallack, 1986; Tabor et al., 2017), and represent a vast archive of information about Earth history (Alexandrovskiy et al., 2022).

Environment of pedogenesis in grasslands of the cool–arid–temperate ecozone

Mineral weathering, clay translocation, and secondary carbonate mineral formation

The typical climatic setting of grassland soils of the cool–arid–temperate ecozone does not favor rapid silicate mineral weathering, although some studies have demonstrated that measurable chemical weathering does occur over periods of thousands to tens of thousands of years. The soil–water balance in this setting does not allow large volumes of water flux through soil; thus, the products of weathering are not effectively removed from the soil profile. In temperate grassland soils, in particular, the cation exchange complex remains dominated by Ca^{2+} (calcium) and Mg^{2+} (magnesium), in contrast to soils with greater throughput of water in which these cations are depleted over time. The base saturation requirement for mollic epipedons is consistent with this observation. The predominance of Ca^{2+} , Mg^{2+} , and other base cations buffers soil pH in the circumneutral range, in which silicate mineral weathering is generally slowest. Divalent cations also inhibit clay particle dispersion and translocation. Ecosystem characteristics and also climate limit weathering and clay translocation in these grassland soils. This is indicated by rapid changes in soil morphology across the grassland–forest transition. With increasing forest density, more prominent eluvial E horizons develop, with a more abrupt downward transition to clay-rich B horizons. Forest soils are reported to have lower pH in upper horizons (under natural vegetation), larger rates of silicate mineral weathering, and increased depletion of total phosphorus than nearby grassland soils. These differences can be explained by contrasting organic matter dynamics, weathering intensity, and clay mobility under the two vegetation types. Organic matter decomposition in grassland tends to produce organic compounds that form strong organoclay complexes and inhibit downward translocation of base cations and clays. Less total organic matter is available under forest to produce resistant organoclay complexes, and these effects combine to facilitate downward translocation of clay. Soils near the transition between present-day forest and prairie (steppe) often show evidence of both ecosystems. This is because this boundary is not static and will shift in relation to changes in climate and vegetation.

Clay translocation in grassland soils is a controversial subject. Argillic horizons (zones of clay enrichment through translocation) are uncommon in grassland soils of Eurasia, except near the grassland–forest boundary. This is consistent with limited clay translocation under conditions of circumneutral pH and abundant exchangeable Ca^{2+} and Mg^{2+} . The German pedologist Arnt Bronger has proposed that these clay-rich horizons represent paleosols buried by loess with much lower clay content. This explanation has been received with skepticism by some North American researchers, but clear evidence for ongoing Holocene loess deposition at some sites suggests that Bronger's hypothesis is plausible (Bronger, 1991; Lucke et al., 2019). Precipitation of secondary calcite (CaCO_3) is almost ubiquitous in fine-textured grassland soils of temperate regions, except in the most humid areas near the grassland–forest boundary. However, secondary carbonates are less common in soils formed in sandy parent materials. In well-drained soils, the processes of secondary carbonate formation are straightforward and are clearly controlled by climate. Soil water contains dissolved Ca^{2+} derived from rainwater, dry dust deposition, or weathering of carbonate or silicate minerals. During major infiltration events, the Ca^{2+} is translocated to deeper soil horizons, but the amount of infiltrating water is insufficient to remove Ca^{2+} completely from the profile. When the soil subsequently dries, the soil water becomes saturated with calcite, which precipitates (the required anion, HCO_3^- , is ubiquitous in soil water). As effective moisture decreases across a grassland region, the depth of frequent wetting decreases and secondary carbonates are concentrated at a shallower depth. When soils are considered across a wide range of climatic conditions, a systematic decrease in depth to the upper limit of secondary carbonates is evident with decreasing effective moisture. At more local scales, however, depth to secondary carbonates is surprisingly variable. This variation may reflect differing soil hydraulic properties, or local topographically controlled differences in the partition of precipitation between runoff and infiltration.

Organic matter accumulation, bioturbation, and structure development

One of the most distinctive aspects of soil formation under grassland in temperate climates, i.e., large organic matter accumulation, was recognized by Dokuchaev (1883) in his early classic of pedology, *Russian Chernozem*. The term 'chern' is Russian for black, therefore 'black earth,' is an apt description of the thick, dark-colored zone of organic matter accumulation, or A horizon, in soils of relatively humid cool temperate grasslands in Russia and the Ukraine (Sorokin et al., 2021). The organic matter derives almost entirely from the decomposition of plants, root exudates and microbial necromass, and an important feature is its long maturation (Vysloužilová et al., 2015). Moving south from the 'chernozem zone' in Russia, effective moisture decreases and A horizons become thinner and lighter-colored, reflecting smaller organic carbon content. Similar trends in carbon content and A horizon color are also observed along east-to-west transects of decreasing effective moisture in the grasslands of central North America (Figs. 1–3) and are likely to be present in other regions. Local anomalies occur within these broad regional trends, in areas where A horizons are being built upwards by ongoing eolian or alluvial deposition, and soil carbon is therefore being sequestered in deeper horizons beyond the depth of most active decomposition (Fig. 3b). A temperature effect that is independent of the soil–water balance is also evident. In North America, soils in the cold grasslands of the northeastern Great Plains store more organic carbon than grassland soils further south that receive similar effective moisture, but have formed under warmer temperatures. Organic carbon content also increases as temperature decreases and precipitation increases, from low- to high-elevation grasslands. The spring and early summer provide the best conditions for humification with the increase in soil moisture from snowmelt (Vysloužilová et al., 2015). The seasonal nature of the climate also results in the formation of stable molecules of large molecular weight that have a long mean residence time in the soil.

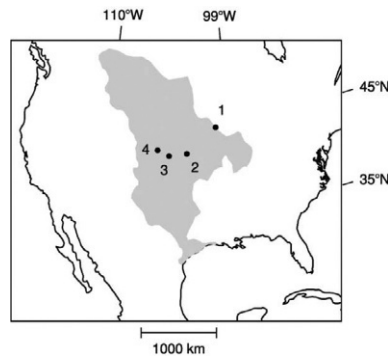


Fig. 1 Locations of grassland soil profiles shown in Figs. 2 and 3, within the grasslands of central North America (shaded). 1, Caledonia, Minnesota; 2, Seward, Nebraska; 3, Beaver City, Nebraska; 4, Brule, Nebraska. Boundaries of the grassland region are approximate and encompass numerous outlying patches of forest and savanna, particularly near the eastern margin of grassland.

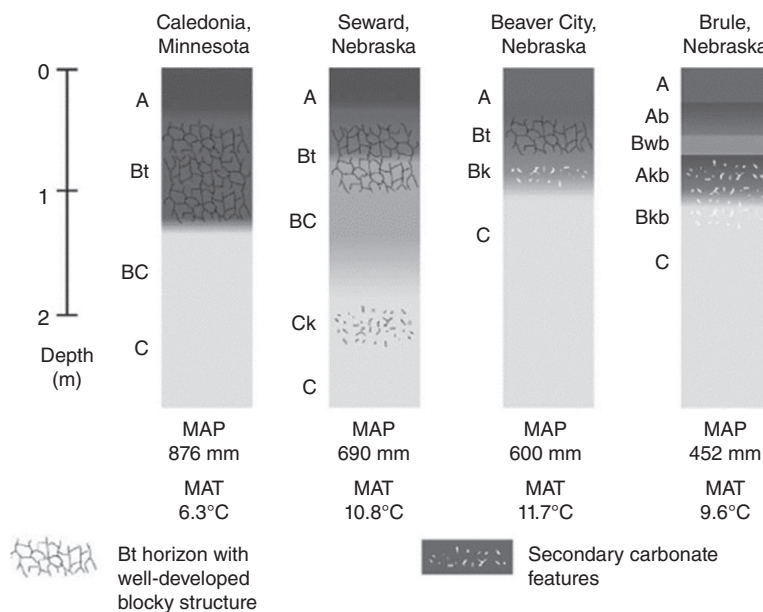


Fig. 2 Schematic view of soil profiles from humid to semiarid grasslands in central North America (locations shown in Fig. 1), showing typical trends associated with climatic gradients. Climatic normal indicated below each column (MAP, mean annual precipitation, 1961–90; MAT, mean annual temperature, 1961–90). The climate at each site has varied significantly over geologic time but the regional climate gradients are likely to have been maintained throughout the period of soil development. Major soil horizons are indicated. Note that 't' indicates clay accumulation, 'k' secondary carbonates, and 'b' buried horizons covered by younger sediment. All profiles have formed in late Pleistocene loess deposited 25,000–12,000 years ago, except for the soil at Brule, Nebraska. The latter formed in late Pleistocene loess at greater than 75-cm depth and in at least two increments of Holocene loess at less than 75 cm. Thickness of dark-colored upper horizons decreases westward as rainfall decreases from Caledonia to Beaver City. The Brule profile deviates from this trend because renewed loess deposition has buried the upper Ab horizon, and organic matter then accumulated in the newly deposited loess, in effect producing an overthickened A horizon. Thickness of the Bt horizon and depth to secondary carbonates also decrease with decreasing rainfall.

These trends with respect to climate can be interpreted using a simple model, in which total soil organic carbon content is in a steady state, with additions equal to losses. The rate of addition is a function of primary productivity, which increases with both increasing effective moisture and increasing temperature. For example, in the central Great Plains of North America, primary productivity increases by two- to threefold in an east-to-west traverse where annual precipitation increases from 250 to 750 mm. Each year's addition of organic matter is offset by loss through oxidation, largely through microbial activity, of a fraction of the total carbon pool. The fraction oxidized also increases with temperature and effective moisture, but not as fast as primary production. Therefore, a much larger pool of soil carbon is required to maintain a steady state in cool, humid grasslands than in areas that are warmer and/or drier. This model is simplistic, because the climate in most grasslands is too variable to allow a true steady state to be maintained, and various forms of organic matter are oxidized at rates that differ by orders of magnitude. Modeling of soil organic carbon dynamics considers multiple pools of soil carbon that turn over at varying rates (Falloon and Smith, 2010). Nevertheless, the results are still consistent with observed trends of soil carbon content in grassland in relation to climate.

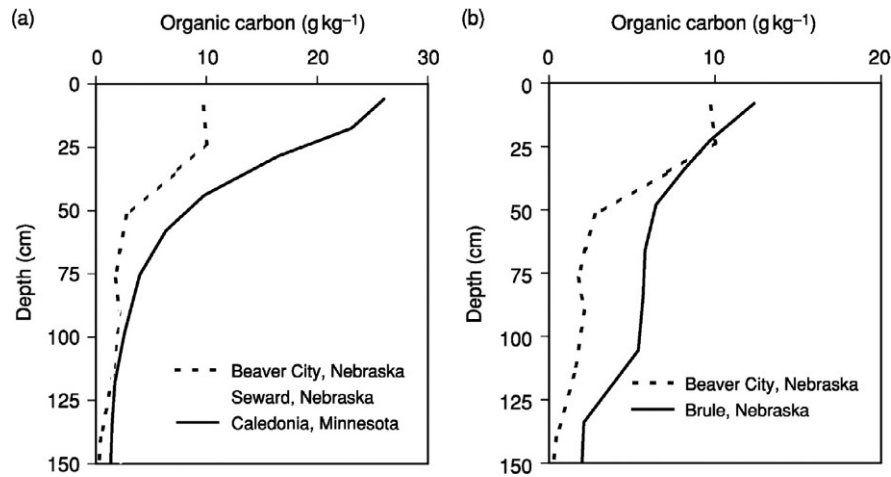


Fig. 3 (a) Profiles of organic carbon content in three of the soils shown in Fig. 2. Total organic carbon storage decreases with decreasing mean annual precipitation. The Caledonia profile also formed in a somewhat colder climate, which further enhances organic carbon accumulation; (b) effect of ongoing loess deposition on carbon storage. The Brule profile, where 75 cm of loess has accumulated in the past 10,000 years contains more organic carbon to a greater depth than the Beaver City profile, where mean annual precipitation is somewhat greater but loess deposition in the same time interval has been less than 40 cm.

Soil texture also strongly influences organic matter content. Regression models to predict soil organic carbon content often include clay content or some other textural index as a key predictor. This effect is related to the formation of complexes containing both humus and mineral particles, which can enhance resistance to organic matter decomposition. Organic carbon associated with coarse clay and fine silt is more stable than the more rapidly turned over portion associated with fine clays. Strongly aromatic humic acids are associated with the coarser clays. Original resistance, complexation with phenolic polymers, and absorption to clay can maintain humic materials in the soil for thousands of years.

Locally, the vertical distribution of organic matter within the soil profile can change rapidly across the transition from grassland to forest. Scattered trees have limited influence on surface soil characteristics, but as the forest becomes denser the addition of organic matter to the soil is increasingly dominated by litter deposited on the soil surface rather than that formed belowground by grass roots. Furthermore, a portion of tree root biomass comprises large, suberized and lignified roots that are longer lived and less readily decomposed than fine grass roots, and therefore contribute less organic matter to the soil on an annual basis. Addition of organic matter, primarily at the surface, and limited bioturbation lead to much thinner A horizons under closed forest than under nearby grasslands. Soils with intermediate A-horizon thickness occur in the grassland–forest transition zone. Enhanced moisture in the litter layer of a shaded forest floor environment results in greater rates of decomposition than in prairie and steppe settings. This increases the production of organic acids which can move through the profile, with the effect of reducing total carbon storage in near-surface horizons under forest (Fig. 4).

Both bioturbation and formation of mineral–organic matter complexes play a role in the development of the fine, exceptionally stable structure of many A horizons in temperate grasslands. Macroscopically, soils under undisturbed natural grassland often have moderate to strong granular structure. Microscopically, these soils display a spongy fabric made up of rounded aggregates with abundant interaggregate pore space. Many of the most stable and well-rounded aggregates are the excrements of soil fauna. Others are produced through extensive burrowing activity and propagation of the fine root network of grasses. An extreme case is represented by soils that are predominantly made up of earthworm casts (Vermustolls and Vermudolls, Soil Taxonomy or Vermic Chernozems, WRB). Aggregates in undisturbed grassland soils are stabilized by mineral–organic matter complexes (particularly microbial polysaccharides that interact with clays of submicron size), fungal hyphae, and other bonding agents, which makes them resistant to physical disruption and slaking or dispersion in water.

Conversion of grasslands to agricultural use results in a reduction of the organic carbon content by 20–50% over 20–40 years. Agronomic tillage practices break up soil aggregates, distributes oxygen and moisture through the plow layer resulting in a loss of organic carbon. This reduces soil water-holding capacity, though effects are small, and structural stability, and may reduce cation exchange capacity.

Dynamics of organic matter after cultivation

The conversion of land-use type from grassland to arable can profoundly change SOC storage through its effects on soil C turnover and distribution in the soil profile. To meet the rapid growth in population and the need for food security, Mollisols have experienced excessive land-use change from natural grassland to intensive agriculture, which has led to a decrease in SOC content. This has led to a deterioration in soil physical properties, greater soil erosion, and an increase in soil degradation. It has been estimated that over the past 150 years of cultivation, SOC content has decreased by almost 50% (Food and Agriculture Organization (FAO) of the United Nations, 2022). Intensive crop management has also caused a loss of plant nutrients, which has caused a

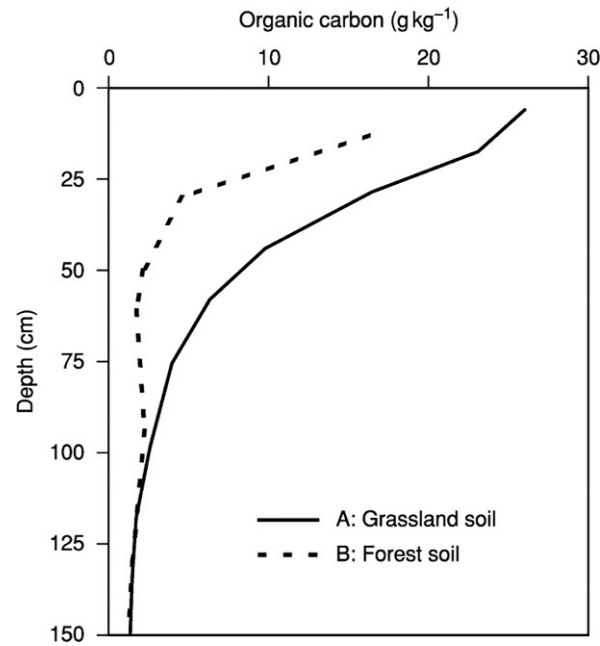


Fig. 4 Soil profiles of contrasting organic carbon content in Minnesota, United States, that were under grassland or forest at the time of public land surveys in the 19th century. The soils formed under identical parent materials and climatic conditions and are in similar topographic settings. Locations shown in Fig. 1.

decline in soil fertility. The major Mollisol regions in Northeast China have experienced a downward trend in carbon source that could further degrade the soil if unsustainable land use and management are disregarded. Long-term traditional continuous cropping and tillage in Northeast China are the major factors that have led to the reduction in SOC content and adverse effects on soil properties (Xu et al., 2020).

Mollisols are inherently productive and fertile soils. They are extensively used for cereal and legume production, and are also important soils in pasture, range and forage systems. The majority of the world's Mollisols are affected by water or wind erosion and also by physical or chemical degradation, including loss of SOC, aggregate breakdown, crusting, plow pan development, and acidification due to their long agricultural history and improper management. Agricultural practices and policies are required to guarantee the sustainable use of Mollisols worldwide. Practices, which have been recommended, include increasing the acreage under minimum or restricted tillage, returning plant residues and adding organic amendments, such as animal manure, to maintain or increase soil organic matter content, and the more systematic use of inorganic amendments, such as agricultural limestone to replenish soil calcium reserves. The Food and Agricultural Organization (FAO) of the United Nations has recently published a report on the 'Global Status of Black Soils' (Food and Agriculture Organization (FAO) of the United Nations, 2022) in which concerns about their intensive use and the aim to establish a global agreement for the sustainability of these soils are raised.

Grassland soil hydrology and its influence on pedogenesis

Closed depressions are common in many midlatitude grassland regions. These include regions that were glaciated during the most recent glacial period, and where there has been insufficient time for a connected drainage system to form. Closed depressions of eolian origin are common in dune fields and loess-mantled landscapes. Depressions can serve either as points of focused groundwater recharge or as discharge areas for groundwater flow systems. Some depressions alternate between these two roles on a seasonal basis. Precipitation and snowmelt in recharge depressions is removed by both percolation and evapotranspiration (unless artificial drainage has been installed). Grassland landscapes containing closed depressions frequently illustrate the concept of a soil *catena*. A catena is a sequence of soils along a slope that have different profile characteristics related to topographic position and different subsurface hydrologic systems. It shows a process-response relation to differences in morphology, erosion, deposition, transport of surface material and leaching. Well-drained soils on broad summits are drier and have moderately deep, dark brown surface colors reflecting more oxidizing conditions. Rates of decomposition are also slow in the relatively cool, wetter soils in depressions or on lower slopes. Soils in some closed depressions display a distinct, light-colored E horizon, which is generally lacking in other grassland soils. Grassland soils in areas of groundwater discharge often accumulate soluble salts, gypsum, and/or secondary carbonate minerals. These minerals form from ions transported to the site by groundwater flow. As water is extracted from the soil by root uptake and surface evaporation, the concentration of ions exceeds the solubility of the minerals and they precipitate, often in order of increasing solubility, along the upward groundwater

flow path. That is, carbonates precipitate first, followed by gypsum and soluble salts; this occurs in Kastanozems (equivalent to Xerolls and Ustolls in Soil Taxonomy). Both the vertical and horizontal distributions of secondary minerals are markedly sensitive to the local topographic and hydrologic setting. In other settings, there is a vertical sequence of secondary mineral formation, from calcite at greater depths through a gypsum enriched zone to soluble salts in shallow horizons. This sequence can lead to large concentrations of sodium (Na) in near-surface soil horizons, either in the form of sodium salts or as Na^+ ions held on cation exchange sites of clay and organic matter (Bischoff et al., 2018). Provided that the soluble salt content is large, soil structure is stable. If the upper part of the soil profile periodically experiences downward fluxes of infiltrated water (e.g., in a depression that is a site of either discharge or recharge depending on the season), then salts will be moved to greater depths. The Na^+ -saturated clay and organic matter particles are then easily dispersed and translocated to form a dense, nearly-impermeable B horizon. Such sodium-affected soils often occur in patches associated with subtle microrelief, such as shallow depressions a few meters to tens of meters in diameter.

Classification of the cool–arid–temperate grassland soils

Soil classification was established by the Russian scientist V.V. Dokuchaev in the 1870s (Kravchenko et al., 2010; Vysloužilová et al., 2015). Dokuchaev devised a genetic classification, which was further refined by Russian, European and American scientists. He identified the five major factors that control soil formation, namely: the nature of the parent material (chemical and mineralogical composition), the climate (especially temperature and precipitation), the influences of organisms—flora, fauna and humans, the topography of the area (relief), and the length of time over which soil formation had occurred (Whalen and Sampedro, 2009). Dokuchaev's approach remains important in most, if not all, of the world's major soil classification systems, such as USDA Soil Taxonomy, which recognizes 12 soil orders (Soil Survey Staff, 2022) of which Mollisols is one, and the World Reference Base for Soil Resources (IUSS Working Group WRB, 2022) which has 32 soil reference groups (RSGs) of which three relate to Mollisols.

The two major world soil classification systems use different terminology for the soils of the cool–arid–temperate ecozone, but they embrace similar soil types. Soil Taxonomy (Soil Survey Staff, 2022) uses the term Mollisol, which is one of the 12 soil orders. Mollisols are variable and the differences are indicated at the suborder level by eight groups—the main ones are Udolls, Cryolls, Xerolls, Aquolls and Ustolls. The WRB has three representative groups that cover this ecozone: Chernozems (equivalent to Mollisol *sensu stricto*), Kastanozems (equivalent to Ustolls and Cryolls of Mollisols) and Phaeozems (equivalent to Udolls and Aquolls).

Dark-colored Chernozems (Fig. 5) occupy the humid end of the climatic range of grassland ecosystems, whereas lighter-colored Chestnut soils occur in the drier grasslands and shrub steppes. The WRB system maintains this subdivision (Chestnut soils become Kastanozems), though with more rigorous definitions. The USDA Soil Taxonomy subdivides most well-drained Mollisols on the basis of climatic setting rather than intrinsic soil properties. Many Mollisols of cold climates, such as those formed under high-altitude grasslands, are included in the Cryoll suborder (Mason and Zanner, 2005). More specifically, Mollisols are most extensive in mid-latitude, mid-continental locations, but are also identified in some colder and drier climates.

Chernozems–Mollisols

The International Soil Reference and Information Centre (ISRIC, Wageningen, The Netherlands) indicates that Chernozems cover an area of ~230 million hectares (see relevant websites below). Both Mollisols and Chernozems belong to the tall grass steppe (for example the Haplic Chernozem, Fig. 6) of the mid-continental regions of the mid-latitudes of Eurasia and North America with forest on the northern margin. These areas have a continental climate with cold winters, hot summers and rain mainly in summer. These soils have a thick dark humus-rich surface horizon (known as a mollic epipedon in Soil Taxonomy and a chernic horizon in the WRB) which is the key feature of Mollisols (Bell and McDaniel, 2000). The topsoil horizon is thick (at least 20 cm, but can be up to 200 cm), very dark-colored, has high base saturation >50%, a moderate to large content of organic matter, has strong biological activity, and a well-developed crumb structure. The texture tends to be sandy loam, and secondary carbonates occur within 2 m of the soil surface. There should be no petrocalcic horizon between 25 cm and 200 cm, no secondary gypsum, and no uncoated sand or silt grains on ped faces. The organic matter content can be up to 10–16%, topsoil pH of 6.5–7.5 and subsoil pH of 7.5–8.5. The large porosity means that the soils are well drained and that there is some translocation to 50–200 cm during wet periods. With the increase in clay content, the structure becomes blocky or weakly prismatic. Water holding capacity is good at ~33% at field capacity and 13% at wilting point. The C:N ratio is typically ~10. https://www.isric.org/sites/default/files/major_soils_of_the_world/set8/ch/chernoze.pdf.

Most Chernozems–Mollisols are naturally fertile and are widely cultivated. They produce a large proportion of the world's grain crops, and are known as the 'breadbasket' regions of the world, such as in the United States, Russia, Ukraine, and Argentina. On their northern borders and at higher altitudes they grade into Kastanozems (WRB) or Ustolls and Cryolls (Soil Survey Staff, ST), and on their southern borders and warmer wetter regions they grade into Phaeozems (WRB) or Udolls or Aquolls (ST).



Fig. 5 Photograph of a Russian Chernozem taken by Professor Peter Schad. The labels refer to the horizons and the scale is given by the 10-cm rule.
<https://www3.lis.tum.de/boku/wrb-working-group/pictures/>

Phaeozems (Udolls and Aquolls)

The Phaeozems of the WRB classification (linked to the Udolls and Aquolls of Soil Taxonomy) cover approximately 190 million hectares worldwide, with ~70 million hectares in the eastern Great Plains of North America, ~50 million hectares in the subtropical pampas of Argentina and Uruguay, and ~18 million hectares in NE China (see relevant websites below). They belong to the warm and wet steppe or prairie and forest prairie. The rates of weathering and leaching are greater than for the Chernozems or Kastanozems that embrace the Mollisol group of Soil Taxonomy, and may contain uncoated sand and silt particles on ped faces. They are less rich in bases than Chernozems and Kastanozems, but they still exceed 50% in the upper meter where there are also no secondary carbonates. These soils still contain sufficient bases and plant nutrients for agriculture, have a large biomass and active faunal activity. They have developed on loess, glacial till and other basic unconsolidated parent materials. These soils have a dark topsoil, a mollic horizon that is thinner than the chernic horizon over a cambic or argic horizon. https://www.isric.org/sites/default/files/major_soils_of_the_world/set8/ph/phaeozem.pdf.

Phaeozems (Figs. 7 and 8) are well drained and fertile, therefore they are widely cultivated for cereals or soybean (*Glycine max* (L.) Merrill), and are also used for cattle ranching.

The characteristics of Phaeozems relate strongly to the greater rainfall (up to 1200 mm per annum) and higher temperatures (up to an average temperature of 18 °C) in the regions where they form. The mollic horizon is less dark and thick (30–50 cm) than that of the Chernozem, is well aerated and is more leached, however it has good water holding capacity especially related to clay enrichment at depth. The organic matter content of the topsoil is ~5%, the C:N ratio is 10–12, and pH is 5–7. The topsoil has a well-developed crumb to blocky structure, and at depth the structure becomes more blocky.



Fig. 6 Photograph of a Haplic Chernozem taken in Russia by Professor Peter Schad. In: *Essentials of Soil Science*, W. Blum, P. Schad, S. Nortcliff, 2018, IUSS, Borntraeger Science Publishers, Stuttgart, Germany. <https://www.iuss.org/soils-4-u/world-of-soils/>

Kastanozems–Ustolls and Borolls of the Mollisols

This component of the Mollisol order (Soil Survey Staff) covers ~465 million hectares (see relevant websites below). These are the soils of the short grass steppe belt, south of the Eurasian tall grass steppe, the western part of the North American Great Plains, and drier parts of the South American pampas region (Argentina, Uruguay and Paraguay). They have a similar profile to Chernozems, but the humus rich topsoil horizon is less deep and less black. They have a more chestnut color (see above) and contain a greater accumulation of carbonates. The mollic horizon is at least 20-cm deep over a cambic or argic horizon, secondary carbonates are within 100 cm, and secondary gypsum might also be present at depths greater than 150 cm (Figs. 9 and 10). The main parent material is loess, but they also form on other unconsolidated materials. They form under a drier climate than Chernozems or Phaeozems with hot summers and cold winters. Kastanozems are fertile soils, but agriculture usually requires irrigation or the use of dry farming methods. There is also a need to avoid salinization by the careful management of water. The main crops are small grains and vegetables, and large areas are also used for grazing. https://www.isric.org/sites/default/files/major_soils_of_the_world/set8/ks/kastanoz.pdf.

The intermittent water regime means that Kastanozems can dry out to a great depth in the dry season, and they might not be completely wetted during wetter periods. They are more prone to soil erosion than the other two types of Chernozem. The physical structure is less favorable than for Chernozems because of the weaker aggregation. The organic matter content of the topsoil is now only 2–4% and rarely exceeds 5%. The base saturation remains above 50% as for the other two types, and the C:N ratio is ~10. The structure is granular to fine blocky, and because it is weaker than for the other two types the aeration and water holding capacity are less. The pH is ~7, but can reach 8.5 at depth.



Fig. 7 Photograph of a Phaeozem taken in Mexico by Professor Peter Schad. The scale is given by the 20-cm rule. <https://www3.lis.tum.de/boku/wrb-working-group/pictures/>

Distribution of Mollisols–Chernozems

Mollisols, Black Soils or Prairie Soils make up about 916 million ha worldwide, which is ~7% of the world's ice-free land surface. Their distribution strongly correlates with native prairie ecosystems, but is not limited to them. They are most prevalent in the mid-latitudes of North America, Eurasia, and South America. In North America, they cover 200 million ha of the United States, more than 40 million ha of Canada and 50 million ha of Mexico. Across Eurasia they cover around 450 million ha; 148 million ha in southern Russia, 34 million ha in Ukraine, and 35 million ha in northeast China. They are common in the South American countries of Argentina and Uruguay, and cover about 89 and 13 million ha, respectively. Mollisols are often recognized as inherently productive and fertile soils. They are extensively and intensively farmed, and increasingly dedicated to cereals production, which needs significant inputs of fertilizers and tillage.

On a world-wide basis, soil scientists generally speak of four major regions of Mollisols (Table 1). One is in central North America radiating across the central plains of the United States, eastern Mexico and southern Canada. The next two appear as a discontinuous belt, which extends across southeastern Europe and central Asia. The western belt begins in the sub-humid steppes of south-central Europe and extends across Russia and into the eastern belt, which is best represented in northeast China. The fourth major region corresponds to the Pampas of South America, covering most of central-eastern Argentina and almost all of Uruguay. Thus, most of the world's 916 million ha of Mollisols occur in three regions of the northern hemisphere and one region south of the equator, the Parana-La Plata basin of South America. The Global Status of Black Soils Report by Food and Agriculture Organization (FAO) of the United Nations (2022, Fig. 2.3) provides a recent global map of the distribution of black soil. Understanding of this distribution, and also the formation, uses and risks of these soils are crucial given that these four major Mollisol regions collectively form the world's natural granary. Intensive cultivation, loss of soil organic carbon (SOC), soil erosion and associated yield suppression are still the problems that threaten future sustainability of agriculture in the different Mollisol regions.

Distribution in North America

Mollisols in North America are most extensive in the central plains, extending from the prairie provinces of Canada across the USA's Great Plains and into eastern Mexico's semi-arid grasslands (Fig. 11). Sizeable areas of Mollisols are present in the more



Fig. 8 Photograph of a Luvic Phaezoem taken in the USA by Professor Peter Schad. The scale is given by the rule in centimeters. In: *Essentials of Soil Science*, W. Blum, P. Schad, S. Nortcliff, 2018, IUSS, Borntraeger Science Publishers, Stuttgart, Germany. <https://www.iuss.org/soils-4-u/world-of-soils/>

mountainous, semi-arid western USA and there are even locally important regions of Mollisols in the humid southeast USA, especially in Florida. The distribution across much of North America suggests their formation requires only two features: (a) fertile parent material, and (b) biota that result in humus formation to a depth of 25 cm or more. The former feature is most commonly associated with Pleistocene-aged sediments, especially till, loess and alluvium. The latter feature usually occurs in prairie ecosystems, which are known for their perennial, deep, fibrous root systems with slow decomposition to form metastable Ca-humic materials.

Distribution in South America

The South American Mollisols extend principally over the Chaco-Pampa region together with patches in Patagonia and Mesopotamia regions in Argentina (8.87 million ha, 32% of the territory), Brazil (Pinheiro Junior et al., 2022) and most of the Uruguayan territory (1.30 million ha, 75% of total land, Fig. 12). The majority of South American Mollisols are in the temperate zone; the mean annual temperature ranges from 14 °C at the southern limit of the Argentine Pampa to 19 °C in northern Uruguay. The Pampa region is a vast and continuous plain, rising slightly and gradually to the north and to the west where level areas alternate with slightly undulating plains and rolling landscapes. The most recent Quaternary deposits at the surface and therefore the parent material of Pampean Mollisols are loess and loessoid sediments to the east and eolian sands to the west of the region. The loess here differs from the typical loess in the northern hemisphere where it comprises 50–70% quartz; the light mineral suite in Pampean loess is generally dominated by an abundance of volcanic glass, with more pyroxenes and amphiboles in the heavy fraction. With regard to the clay fraction there is a clear dominance of illite in the west and an increase in smectite and interstratified illite-smectite to the east, close to the Paraná River.

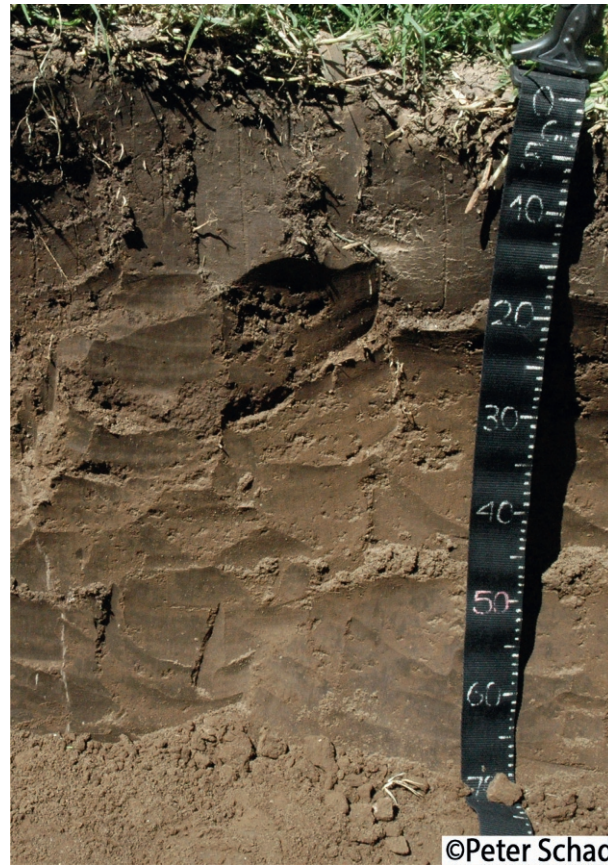


Fig. 9 Photograph of a Haplic Kastanozem taken in Ecuador by Professor Peter Schad. The scale is given by the 10-cm rule. <https://www3.lis.tum.de/boku/wrb-working-group/pictures/>

All the Pampean (Argentine) Mollisols have a thermic temperature regime although some border on mesic. The moisture regimes are mostly udic and ustic. The Great Group of Argiudolls is undoubtedly the most characteristic of the Argentinean Pampa as well as the most productive, and where agricultural activity is more intense.

Mollisols are the most representative soils in the Pampa grasslands although they are also widely distributed associated with several other dominant soil orders in the northern Chaco and Mesopotamia subtropical regions, and in the cold-temperate parts of Patagonia in the extreme south of Argentina. All Suborders of Mollisols are here; the most extensive being Udolls and Ustolls.

Uruguay has zones of pure Mollisols, pure Vertisols and Mollisol-Vertisol mixtures (Fig. 13). The Vertisols are mentioned because they have many properties similar to Mollisols in Uruguay, such as large organic matter content. Mollisols are distributed all over Uruguay, especially in the western half of the territory. Those in the western region are more fertile, richer in organic matter, and more resistant to water erosion than those in the eastern part of the country. Mollisols in Uruguay have developed from a wide variety of mainly fine or medium textured sediments from basalt. On the acid crystalline rocks of the Precambrian shield, Mollisols have a reddish brown or red argillic horizon, a sandy loam A horizon overlying a clay loam or clay B horizon. A stone line between the two horizons is often evident. They have neither swelling clays nor free carbonates and are more acid with depth. Base saturation does not increase significantly in the lower horizons (B and C) as it does in other Mollisols. Albolls, Aquolls and Udolls Suborders of Mollisols occur in Uruguay.

Distribution in Europe

Russian Chernozems are distributed between 50° to 54° N and extend from the border with Ukraine in the west to the Yenisei River in the east (Fig. 14). Podzolized and leached Chernozems typically cover 45 million ha, while meadow-Chernozems (semi-hydromorphic soils) cover 13.5 million ha. In the steppe region, 52 million ha are typical Chernozems, and 11.5 million ha are meadow-chernozems. In addition, 11 million ha of Kastanozems (Chestnut Soils) and 15 million ha of Humic Cambisols could also be added to the Chernozems of this region. The total area of Chernozems accounts for 52.6% of arable lands in the Russian Federation and approximately 72% of these soils have been cultivated (Sorokin et al., 2021).

Ukrainian Chernozems developed on loess and loess-like sediments cover 25.6 million ha, while those developed on eluvium from limestone, sandstone and slates cover 1.8 million ha (Fig. 15). Humic Cambisols with chestnut properties cover 6.6 million



Fig. 10 Photograph of a Calcic Endogypsic Kastanozem taken in Mexico by Professor Peter Schad. The scale is given by the 20-cm rule. In: Essentials of Soil Science, W. Blum, P. Schad, S. Nortcliff, 2018, IUSS, Borntraeger Science Publishers, Stuttgart, Germany. <https://www.iuss.org/soils-4-u/world-of-soils/>

Table 1 Distribution of Mollisols.

	<i>Location</i>	<i>Soil organic carbon</i>	<i>Temperature</i>	<i>Precipitation</i>
Russia	lat. 50° to 54°N and long. 33–102° E	80–220 Mg ha ⁻¹	–4 to 8 °C	350–700 mm
Ukraine	lat. 51°18'N to 44°41'N and long. 24°18'E to 40°12'E	30–120 g kg ⁻¹	7.0–7.7 °C	490–550 mm
Argentina	lat. 22–39°S and long. 52–65°W	16–55 g kg ⁻¹	14 °C	500 mm
Uruguay	lat. 32–35°S and long. 52–58°W	27 g kg ⁻¹	19 °C	1500 mm
Northeast China	lat. 38°43'–53°33'N and long. 115°31'–135°05'E	28 g kg ⁻¹	–2.5 to 5.7 °C	300–600 mm
Poland	N 50°59' E 16°56', N 50°49' E 16°51'	12–21 g kg ⁻¹	8.0–8.5 °C	580–620 mm
Slovakia	N 48°17' E 17°33', N 48°17' E 21°22'	19–21 g kg ⁻¹	9.0–10 °C	500–580 mm
Hungary	N 46°02' E 18°03', N 47°19' E 18°47'	12.0–16.0 g kg ⁻¹	10.6–11 °C	560–610 mm

ha. There are 34 million ha of cultivated Chernozems, which account for 62% of all agricultural lands in Ukraine and approximately 78% of these soils have been cultivated. These soils are in a zone of 737 km north–south (51°18'N to 44°41'N) and 1144-km east–west (24°18'E to 40°12'E). The Ukrainian Chernozems are on a plain, but it is a geologically nonuniform region. The podzolized and leached Chernozems are distributed across the well-drained uplands of the forest steppe and watersheds. The typical Chernozems are widespread on upland plateaus between river valleys and terraces. They also occur everywhere in the northern subzone of the steppe, covering the watershed plateaus. 'Southern' Chernozems are common across the Black Sea lowlands and mid-Crimean peninsula, and on the flat plateaus of the southern steppe. Ukrainian Chernozems are distributed in the forest–steppe and steppe zones of Ukraine, where deciduous forests are the native vegetation in the west (pre-Carpathians) to the forest zone in the north, and ending along the Black and Azov sea coasts. The climate in the forest–steppe zone with podzolized,

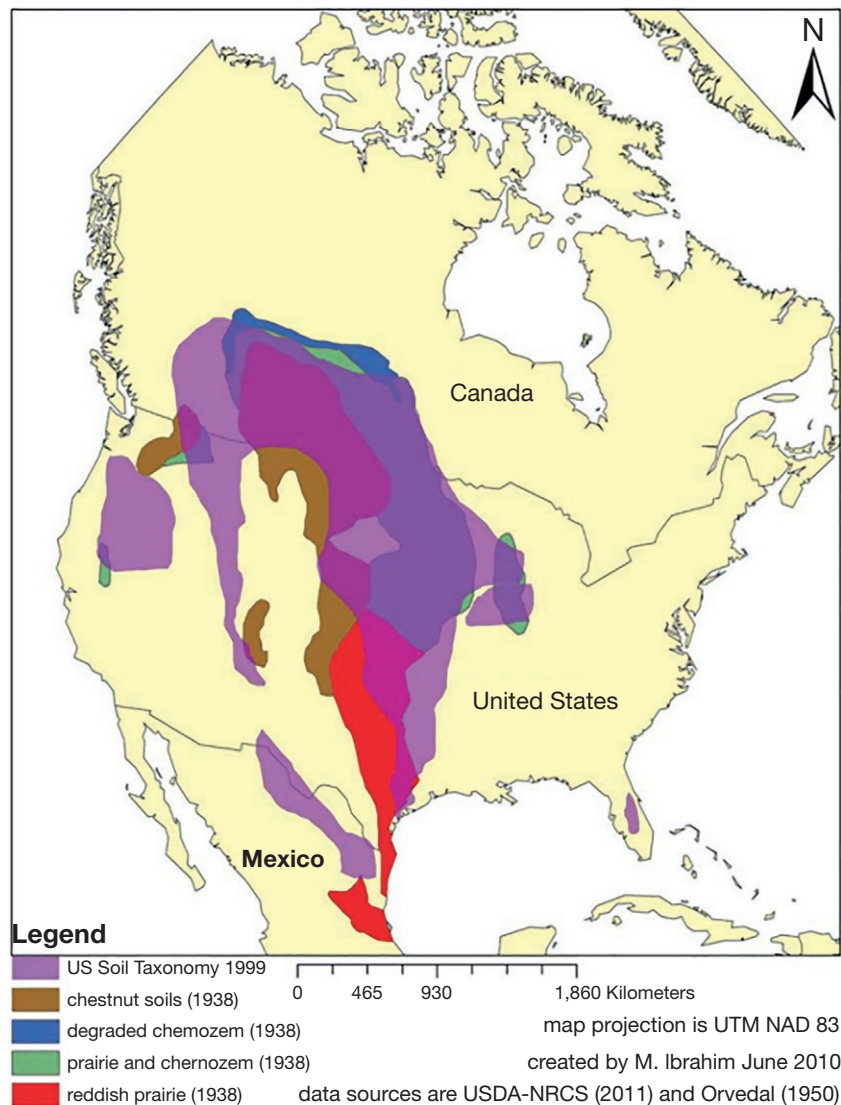


Fig. 11 Distribution of Mollisols across North America. From Liu XB, Burras, CL, Kravchenko, YS, Duran, A, Huffman, T, Morras, H, Studdert, G, Zhang, X, Cruse, RM, and Yuan, X (2012) Overview of Mollisols in the world: Distribution, land use and management. *Canadian Journal of Soil Science* 92: 383–402. <https://cdnsiencepub.com/doi/full/10.4141/cjss2010-058>

leached, and typical Chernozems is slightly warmer and milder than in the forest zone. Fig. 16 shows the type of Chernozem described above from nearby Romania (Scheia, near the Moldovan border). The soil has developed on Late Pleistocene loess over calcareous sandstone. The land has the characteristic undulating form with a well-developed Chernozem in the foreground, but a lighter calcareous rich soil in the background with clear evidence of soil erosion on the convex slope. The land in the area is used for cereals and grassland, and there is also deciduous woodland in places.

In Central Europe, Chernozemic soils at various stages of formation are placed in the RSGs of Chernozems, Phaeozems or Kastanozems, with respective qualifiers, e.g., Cambic, Luvic, Salic or Protosalic, Sodic or Protosodic, etc. Classification of these soils in the WRB depends on the presence of chernic or mollic horizons, depth of secondary carbonate accumulation and depth of gleyic or stagnic properties, and may vary from Gleyic or Stagnic Chernozems or Phaeozems to Mollic Gleysols or Stagnosols.

Chernozemic soils developed from glacial tills with naturally poor drainage and waterlogging occur on the moraine plateaus in northern Poland. In SW Poland (the Lower Silesia region), two types of chernozemic soils have developed; the naturally well-drained ('dry') and the waterlogged or artificially drained ('hydromorphic') forms. Both 'dry' and 'hydromorphic' chernozemic soils have developed in SW Poland from thick loess or loess-like materials, and typically have a texture of silt loam throughout the profile, but sometimes they are underlain with glacial till or glaciofluvial sand (Kabala et al., 2019).

Slovakia, like other Central European countries, has two types of chernozemic soils differing in moisture regime, i.e. dry or well-drained and hydromorphic. Both forms occur mainly in the Danube lowlands of southwest Slovakia. Hydromorphic Chernozems mainly occur on flat land in the Danube lowlands, on loess sediments mixed with loamy calcareous alluvial deposits,

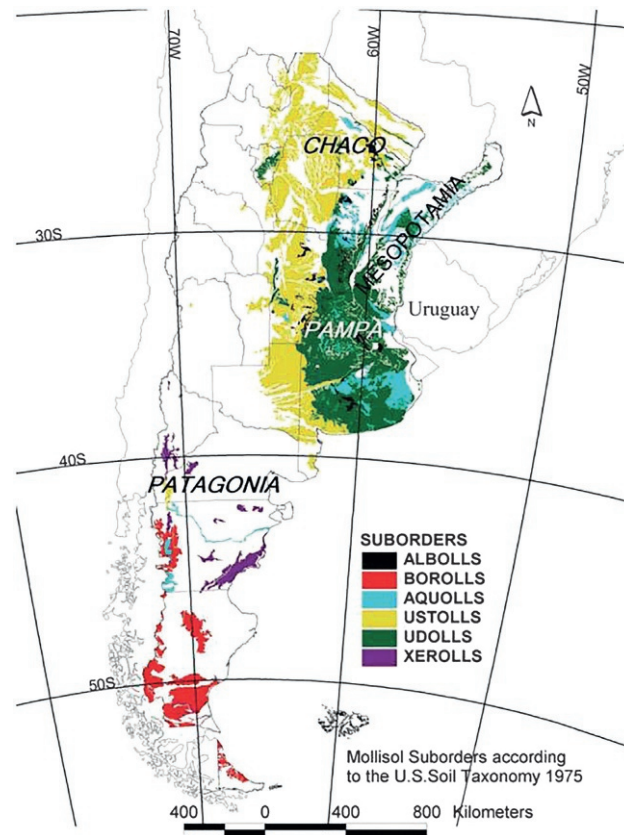


Fig. 12 Geographical distribution of Mollisol suborders in Argentina (SAGyP-INTA, 1990) based on 1975 edition of the US Soil Taxonomy. From Liu XB, Burras, CL, Kravchenko, YS, Duran, A, Huffman, T, Morras, H, Studdert, G, Zhang, X, Cruse, RM, and Yuan, X. (2012) Overview of Mollisols in the world: Distribution, land use and management. *Canadian Journal of Soil Science* **92**: 383–402. <https://cdnsiencepub.com/doi/full/10.4141/cjss2010-058>

whereas dry Chernozems occur mainly on the hilly areas of the Danube and usually have a parent material of loess. Dry Chernozems in SW Slovakia have also developed from loess and loess-like materials, and have a loamy or silty-loam texture, and secondary carbonates accumulation starts within the upper 100 cm of the soil profile. The topsoil humus horizon often fulfills the criteria for a chernic horizon according to the WRB classification, therefore these soils typically meet the requirements for Calcic Chernozems.

Chernozems or chernozem-like soils, according to the genesis-based current Hungarian soil classification system, cover 22.4% of Hungary; it is one of the most important soil types in terms of agricultural productivity. They are considered as zonal soils of the Hungarian lowland steppe and forest steppe vegetation. Chernozems are widespread on flat and slightly sloping hilly areas and foothills, where the limiting soil forming factors, such as shallow consolidated parent material, shallow (saline) groundwater and light (sandy) or heavy (shrink–swell clays) textures are absent. The WRB approximate equivalents of the Hungarian chernozemic soils are Phaeozems, Kastanozems and Chernozems, or in some cases they are interpreted as Mollic Vertisols and Mollic Gleysols. Chernozemic soils, in many places, have been cultivated since the Neolithic age (ca. 8000 years BP), and have been affected by profound anthropogenic, climatic and land use influences over the past few thousand years. A good overview about Chernozem formation and distribution in Central Europe is given in Eckmeier et al. (2007).

Distribution in Northeast China

The Mollisols of Asia are primarily located in China, especially in northeast China (Fig. 17). Northeast China includes the provinces of Heilongjiang, Jilin, and Liaoning, and the Hulunbeier League of the Inner Mongolian Autonomous Region. The total area is 1.299 million hectares. The region is 1600 km long on the east–west axis, and 1400 km wide on the north–south axis. Mollisols have developed mostly from steppe, prairie to forest–steppe vegetation in this region. Geomorphologic landscapes in the region are plain terraces and tablelands. Most of region has undulating to rolling topography with gentle slopes. The elevation is around 150–400 m. Mollisols have developed from Quaternary loess-like sediments. The thickness of loess-like materials is 10–15 m. The soils and their parent materials are generally slightly acid to alkaline. The soil texture ranges from loam to light clay, with most sola being predominantly clay loam. The main clay mineral is illite, although some chlorite and montmorillonite occur in certain profiles. The accumulation of soil organic matter is the main process of Mollisol formation. Many other processes are also involved,

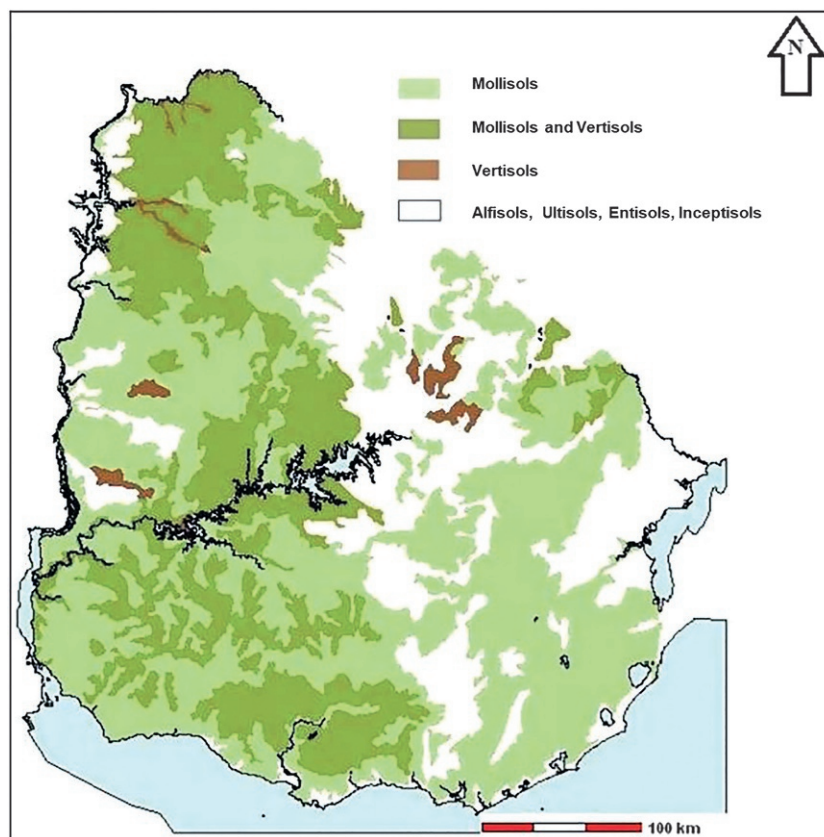


Fig. 13 Geographical distribution and relative importance of Mollisols and Vertisols in Uruguay. From Liu XB, Burras, CL, Kravchenko, YS, Duran, A, Huffman, T, Morras, H, Studdert, G, Zhang, X, Cruse, RM, and Yuan, X. (2012) Overview of Mollisols in the world: Distribution, land use and management. *Canadian Journal of Soil Science* **92**: 383–402. <https://cdnsiencepub.com/doi/full/10.4141/cjss2010-058>

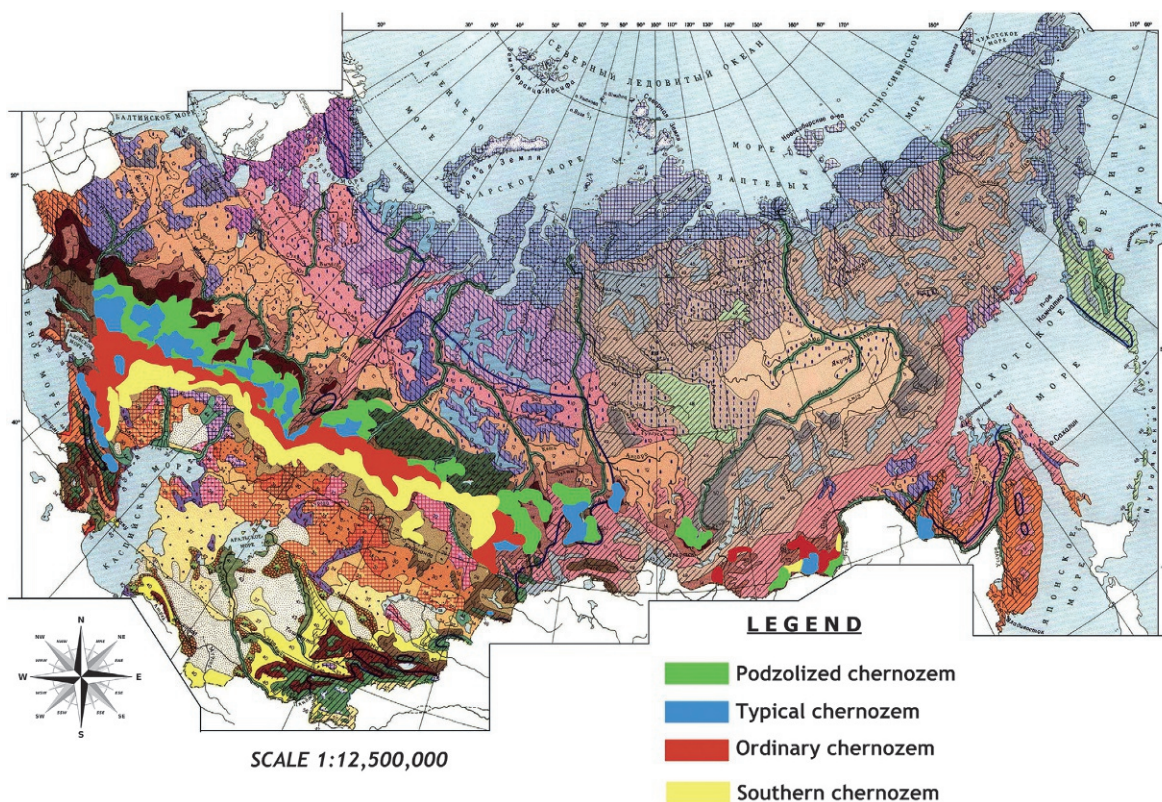


Fig. 14 Distribution of Russian Chernozems. From Liu XB, Burras, CL, Kravchenko, YS, Duran, A, Huffman, T, Morras, H, Studdert, G, Zhang, X, Cruse, RM, and Yuan, X. (2012) Overview of Mollisols in the world: Distribution, land use and management. *Canadian Journal of Soil Science* **92**: 383–402. <https://cdnsiencepub.com/doi/full/10.4141/cjss2010-058>



Fig. 15 Distribution of Subtypes in Ukrainian Chernozems. From Liu XB, Burras, CL, Kravchenko, YS, Duran, A, Huffman, T, Morras, H, Studdert, G, Zhang, X, Cruse, RM, and Yuan, X. (2012) Overview of Mollisols in the world: Distribution, land use and management. *Canadian Journal of Soil Science* **92**: 383–402. <https://cdnsiencepub.com/doi/full/10.4141/cjss2010-058>



Fig. 16 A Chernozem developed on Late Pleistocene loess overlying calcareous sandstone in Romania at Scheia near the border with Moldovia. There is a well-developed Chernozem in the foreground, but a lighter calcareous rich soil in the background with clear evidence of soil erosion on the convex slope. (Photograph taken by R. Bäumler).

such as clayization, (aka, neoformation), eluviation and illuviation, leaching and re-precipitation of carbonate minerals, oxidation and reduction of iron and manganese, salinization, desalinization and albic bleaching. These processes divide the Mollisols into different Suborders and Great groups. Four Mollisol suborders are common to northeast China. Albolls are distributed in Heilongjiang and Jilin provinces only. They account for 31% of the total cropland in the two provinces. Albolls that occur in lowlands are best suited for rice (*Oryza sativa* L.) growing if water is available, whereas those in hilly regions are prone to soil erosion and should be used for afforestation. The Aquolls occur where soil temperature is cooler and have larger organic matter contents

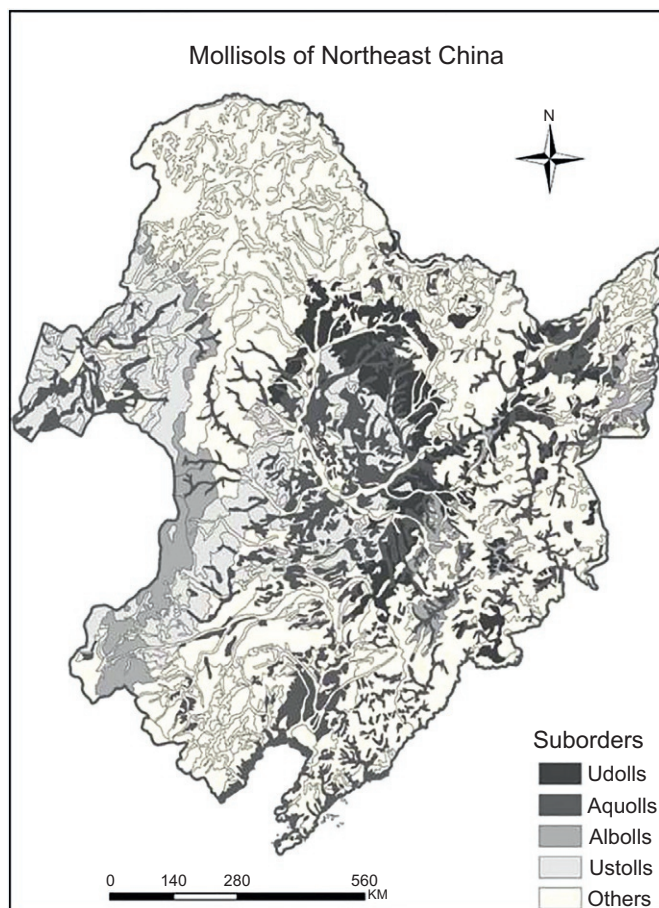


Fig. 17 Distribution of Mollisols in Northeast China. From Liu XB, Burras, CL, Kravchenko, YS, Duran, A, Huffman, T, Morras, H, Studdert, G, Zhang, X, Cruse, RM, and Yuan, X. (2012) Overview of Mollisols in the world: Distribution, land use and management. *Canadian Journal of Soil Science* **92**: 383–402. <https://cdnsiencepub.com/doi/full/10.4141/cjss2010-058>

than their better-drained analogs in the region. The Ustolls are distributed in sub-humid and sub-arid regions, and their land use and management vary. In sub-humid regions with fertile soil and reasonable precipitation during the growing season, Ustolls are used for crop production such as maize (*Zea mays* L.), soybean (*G. max* (L.) Merrill) or wheat (*Triticum aestivum* L.), while most of the Ustolls in semiarid regions are used for livestock production and breeding. Some of the Ustolls continue to have native vegetation consisting of lush and high-quality grasses. These are ideal grazing lands. Udolls cover the center of the Chinese Mollisol area. They are the most important soils for croplands (Liu et al., 2021).

Although the total area of Chinese Mollisols (Black Soils in the Chinese soil classification) cover 1.289 million hectares, accounting for 13.5% of the world total, the continuous area of pure Mollisols in northeast China is 361,920 million ha. Over half of these occur in Heilongjiang Province (i.e., 54.5% of China's total area of Mollisols). The Hulunbeier District of the Inner Mongolian autonomous region and Jilin Province have 22.5% and 19%, respectively, of China's Mollisols, while Liaoning Province accounts for only 4.0% of the total. Although some areas of Black Soils areas were cultivated 200 years ago, most of these soils have been under agricultural production for about 50 years only.

Concluding comments

The cool-temperate grassland soils are an iconic set of soils with Mollisols–Chernozems at their center with variations (Phaeozems–Udolls and Aquolls; Kastanozems–Ustolls and Borolls) related to differences in climate and vegetation. The main areas where these soils develop are the prairies and steppes of the northern hemisphere and the pampa of South America. The climate tends to be continental, with cold winters often with sub-zero temperatures, hot summers and moderate rainfall, and plays an important role in the formation of these soils. The original vegetation was tall or short grassland with forest margins. Mollisols–Chernozems have often formed on parent materials of loess, glacial till and other unconsolidated basic materials. They are characterized by a large and stable organic matter content which makes them inherently productive and fertile. They have a stable well developed granular structure in the topsoil and are well drained. The areas of Mollisols–Chernozems are often

known as the ‘bread baskets’ of the world because of their widespread use for cereals. They have become important for legumes such as soybeans, and for grazing animals in the drier areas. With their intensive use for crops and grazing they have become prone to degradation with the loss of organic matter, aggregate disruption, crusting, development of plow pans, and acidification. These effects have led to increased wind and water erosion which further degrade the soil. With the intensification of management because of the increased demand for food from a growing world population, the pressures on Mollisols–Chernozems is ever increasing. There is an urgent need for continued research to mitigate the effects of management and potential climate change to ensure future sustainability of these soils.

Acknowledgments

The authors acknowledge that some of this chapter contains material from a chapter in the First Edition of the Encyclopedia of Soils in the Environment with the title “Grassland Soils” by J. A. Mason and C. W. Zanner. This chapter is partially supported by the project of Degradation mechanism and control technology of world Mollisols granted by the Chinese Academy of Sciences (131323KYSB20210004).

References

- Alexandrovskiy AL, Chendev YG, and Yurtaev AA (2022) Soils with the second humus horizon, paleochernozems, and the history of pedogenesis at the border between forest and steppe areas. *Eurasian Soil Science* 55(2): 127–146.
- Anderson RC (2006) Evolution and origin of the Central Grassland of North America: Climate, fire, and mammalian grazers. *Journal of the Torrey Botanical Society* 133(4): 626–647.
- Andrade BO, Dallmann JD, Volesky JD, Schacht WH, and Guretzky JA (2022) Grassland plant community response to interacting disturbances and temporal variability. *Restoration Ecology* 30(1), e13495. <https://doi.org/10.1111/rec.13495>.
- Barthold FK, Wiesmeier M, Breuer L, Frede H-G, Wu J, and Blank FB (2013) Land use and climate control the spatial distribution of soil types in the grasslands of Inner Mongolia. *Journal of Arid Environments* 88: 194–205.
- Bell JC and McDaniel PA (2000) Mollisols. In: Sumner ME (ed.) *Handbook of Soil Science*, pp. 286–307. Boca Raton, Florida: CRC Press.
- Bischoff N, Mikutta R, Shibistova O, Dohrmann R, Herdtle D, Gerhard L, Fritzsche F, Puzanov A, Silanteva M, Grebennikova A, and Guggenberger G (2018) Organic matter dynamics along a salinity gradient in Siberian steppe soils. *Biogeosciences* 15: 13–29.
- Bronger A (1991) Argillic horizons in modern loess soils in an Ustic soil moisture regime: Comparative studies in forest-steppe and steppe areas from eastern Europe and the United States. *Advances in Soil Science* 15: 41–90.
- Dokuchaev VV (1883) *Russian Chernozem*. Jerusalem, Israel: Israel Program for Scientific Translation.
- Eckmeier E, Gerlach R, Gehrt E, and Schmidt MWI (2007) Pedogenesis of chernozems in Central Europe—A review. *Geoderma* 139(3–4): 288–299.
- Falloon P and Smith P (2010) Modelling soil carbon dynamic. In: Kutsch W, Bahn M, and Heinemeyer AA (eds.) *Soil Carbon Dynamics: An Integrated Methodology*, pp. 221–224. Cambridge: Cambridge University Press.
- Food and Agriculture Organization (FAO) of the United Nations (2022) *Global Status of Black Soils*. Roma: FAO <https://doi.org/10.4060/cc3124en>.
- Hanberry BB (2021) Wind-bounded grasslands of North America. *Ecological Indicators* 129: 107925. <https://doi.org/10.1016/j.ecolind.2021.107925>.
- IUSS Working Group WRB (2022) World Reference Base for soil resources. In: *International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*, 4th edn Vienna, Austria: International Union of Soil Sciences (IUSS).
- Kabala C, Charzyński P, Czigány S, Novák TJ, Saksa M, and Owitoniak M (2019) Suitability of World Reference Base for Soil Resources (WRB) to describe and classify chernozemic soils in Central Europe. *Soil Science Annual* 70: 244–257.
- Kravchenko Y, Perenko L, and Zhang XY (2010) Ukrainian chernozems: Genesis, properties and amendment. In: Liu XB, Song CY, Cruse RM, and Huffman T (eds.) *New Advances in Research and Management of World Mollisols*, pp. 3–24. Harbin: Northeast Forestry University Press.
- Liu XB, Buras CL, Kravchenko YS, Duran A, Huffman T, Morris H, Studdert G, Zhang X, Cruse RM, and Yuan X (2012) Overview of Mollisols in the world: Distribution, land use and management. *Canadian Journal of Soil Science* 92: 383–402.
- Liu B, Zhang G, Xie Y, Shen B, Gu Z, and Ding Y (2021) Delineating the black soil region and typical black soil region of northeastern China [东北黑土区和东北典型黑土区的范与划界]. *Xekue Tongbao/Chinese Science Bulletin* 66(1): 96–106.
- Lucke B, Sandler A, Vanselow KA, Bruins HJ, Abu-Jaber N, Bäuml R, Porat N, and Kouki P (2019) Composition of modern dust and holocene aeolian sediments in archaeological structures of the southern levant. *Atmosphere* 10(12): 762. <https://doi.org/10.3390/ATMOS10120762>.
- Mason JA and Zanner CW (2005) Grassland soils. In: *Encyclopedia of Soils in the Environment*, pp. 138–145. Elsevier.
- Pinheiro Junior CR, Tavares TR, Oliveira FSD, Santos OAQD, Demattê JAM, Garcia AC, Anjos LHCD, and Pereira MG (2022) Black soils in the Araripe basin, Northeast Brazil: Organic and inorganic carbon accumulation in a Chernozem–Kastanozem–Phaeozem sequence. *Journal of South American Earth Sciences* 116: 103789.
- Retallack GJ (1986) The fossil record of soils. In: Wright VP (ed.) *Paleosols: Their Recognition and Interpretation*, pp. 1–57. Oxford: Blackwell Scientific Publications.
- Soil Survey Staff (2022) *Keys to Soil Taxonomy*, 13th edn USDA–Natural Resources Conservation Service.
- Soong JL and Cotrufo MF (2015) Annual burning of a tallgrass prairie inhibits C and N cycling in soil, increasing recalcitrant pyrogenic organic matter storage while reducing N availability. *Global Change Biology* 21(6): 2321–2333.
- Sorokin A, Owens P, Láng V, Jiang Z-D, Michéll E, and Krasilnikov P (2021) “Black soils” in the Russian Soil Classification system, the US Soil Taxonomy and the WRB: Quantitative correlation and implications for pedodiversity assessment. *Catena* 196: 104824.
- Tabor NJ, Myers TS, and Michel LA (2017) Sedimentologist’s guide for recognition, description, and classification of paleosols. In: *Terrestrial Depositional Systems*, pp. 165–202. Elsevier.
- Vysloužilová B, Ertlyen D, Schwartz D, and Šefrna L (2015) Chernozem. From concept to classification: A review. *AUC Geographica* 51: 85–95.
- Whalen JK and Sampedro L (2009) *Soil Ecology and Management*. Cambridge, MA: CABI International.
- Xu XR, Pei J, Xu Y, and Wang J (2020) Soil organic carbon depletion in global Mollisols regions and restoration by management practices: A review. *Journal of Soils and Sediments* 20: 1173–1181.

Relevant websites

https://www.isric.org/sites/default/files/major_soils_of_the_world/set8/ch/chernoze.pdf.
https://www.isric.org/sites/default/files/major_soils_of_the_world/set8/ph/phaeozem.pdf.
https://www.isric.org/sites/default/files/major_soils_of_the_world/set8/ks/kastanoz.pdf.
https://www.researchgate.net/publication/40120040_Lecture_Notes_on_the_Major_Soils_of_the_World.

Soils from a warm-temperate semi-tropical ecozone (Mediterranean) with humid winter months

J Torrent^a, Marco Pfeiffer^b, and Juan José Ibañez^c, ^aDepartamento de Agronomía, Universidad de Córdoba, Córdoba, Spain; ^bDepartamento de Ingeniería y Suelos, Facultad de Ciencias Agronómicas, Universidad de Chile, La Pintana, Chile; ^cNational Museum of Natural History (MNCN), Spanish National Research Council (CSIC), Madrid, Spain

© 2023 Elsevier Ltd. All rights reserved.

Soil-forming factors	319
Climate	319
Vegetation	321
Lithology and relief	322
Time	322
Soil-forming processes	323
Mass additions and losses	323
Leaching and redistribution of calcite and silica	323
Clay translocation	324
Rubefaction	324
Development of vertic features	326
Major soil types	326
Physical, chemical, and mineralogical properties	328
Physical properties	328
Mineralogical properties	329
Chemical properties	329
Main threads	330
References	331

Abstract

The term ‘Mediterranean’ is applied to a specific climate and kind of vegetation that is typical of areas surrounding the Mediterranean Sea but also found in five other world regions, namely California, central Chile, Southwest and South Australia, and the Cape region of South Africa. Because soil-forming factors vary widely among Mediterranean areas, the resulting soils share neither morphology nor genesis. Despite their considerable diversity, the body of properties of soils in the Mediterranean areas makes them significantly different from soils in subtropical and other temperate climatic regions. This justifies separate consideration of the genesis and properties of ‘Mediterranean soils,’ a term that is used here with no intended specific genetic meaning and encompasses all soils in areas with a Mediterranean climate.

Key points

- This chapter compiles the main aspects that explain the genesis and distribution of soils in the five Mediterranean climates of the planet.
- The main characteristic of Mediterranean soils is the discordance between high temperatures and high moisture, allowing the occurrence of a soil leaching season (winter) and soil drying season (summer).
- This particular alternation between seasons leads to coexistence of pedogenic processes such as rubefaction, illuviation and the formation of silica and calcite rich horizons.

Soil-forming factors

Climate

In Köppen’s classification scheme, the Mediterranean climate is designated as a warm, temperate, rainy climate with dry summers (Cs), the wettest winter month receiving at least three times as much rain as the driest summer month. Areas with a Mediterranean climate lie mostly around the Mediterranean sea, and at latitudes of 30–45° on the west coasts of North and South America, and the SW and South coasts of South Africa and Australia. They form transitional zones between the influence of the westerlies of higher latitudes and subtropical high-pressure cells. The winter weather is dictated by the fronts of the westerlies, whereas the dry summer reflects the dominance of the midtropospheric anticyclone. A characteristic common to Mediterranean climates is their location in

transition zones between temperate humid climates and subtropical deserts, experiencing a northward or southward shift in response to the glacial-interglacial cycles during the last three million years (Dallman, 1998). There is a general consensus that the climate has become progressively drier in the summer in most regions with a Mediterranean climate throughout the Quaternary period. These areas are also very susceptible to more frequent and intense droughts as a result of climate change (Safford and Vallejo, 2019).

The Mediterranean climate encompasses an interesting variety of subclimates, as illustrated by the temperature and rainfall data from selected stations listed in Table 1. The mean annual precipitation ranges from 300 to more than 1500 mm, with one or two maxima in the rainy season. Summer drought lasts from less than 2 to approximately 5 months. The mean annual temperature ranges from about 11 °C (e.g., in regions of Anatolia and central Spain) to approximately 19 °C (e.g., in coastal areas of Cyprus), and absolute maxima above 45 °C and minima below –20 °C have been recorded at some stations. The difference between the mean temperatures of the warmest and coldest month is less than 8 °C in some coastal areas affected by cold ocean currents (e.g., San Francisco Bay, California) but exceeds 20 °C in inland areas with little oceanic influence. The difference at stations by the Mediterranean Sea is typically 13–16 °C. From the standpoint of pedoclimate, the contrast in moisture content between winter and summer was used as the basis for defining the *xeric* moisture regime, an important diagnostic criterion in the Soil Taxonomy classification system. Under this regime, the soil zone affecting plant growth is dry for more than 45 days in summer and moist for more than 45 days in winter.

The influence of the Mediterranean climate on soil development can be gleaned by considering two simple parameters, namely: the degree of leaching (*L*) and the actual evapotranspiration (*AET*). The annual value of *L* is calculated as the sum of the excess precipitation (*P*) over potential evapotranspiration (*PET*) for those periods in which $P > PET$ less the soil water-holding capacity (*WHC*). The annual value of *AET* is equivalent to $P - L$. Table 2 shows the water balance for a station with a warm continental Mediterranean climate (Córdoba, Spain) and includes the mean monthly and annual values of *P*, *PET*, *AET*, and *L*. Winter leaching (37 mm) is low, consistent with other Mediterranean stations with mean annual *P* below 600–700 mm, if mean monthly *P* data are used in the calculations. As can be seen from Table 2, the summer drought leaves the soil with little water to meet plant needs and enable mineral weathering and biological activity. The intensity of these processes peaks in autumn and late spring when soil is moist and temperatures are mild. In summary, the relative lack of coincidence between high soil temperatures and high moisture contents precludes strong weathering and continuous biological activity (Torrent, 1995).

The variation in annual rainfall is high in the Mediterranean region. For example, the annual rainfall in Córdoba for the period 1901–2000 ranged from less than 250 mm to more than 1250 mm, i.e., by more than five times. The annual leaching for the same period ranged from 0 to more than 650 mm (at *WHC* 125 mm), with a moderately bimodal distribution (Fig. 1). Because of the

Table 1 Climatic data from selected stations in the Mediterranean region.

Month													
Station	J	F	M	A	M	J	J	A	S	O	N	D	Year
<i>Athens</i>													
Temperature (°C)	9	10	11	15	20	24	27	27	23	19	15	11	18
Precipitation (mm)	62	36	38	23	23	14	6	7	15	51	56	71	402
<i>Rome</i>													
Temperature (°C)	8	9	11	14	18	22	24	24	22	17	13	10	16
Precipitation (mm)	83	73	52	50	48	18	9	18	70	110	113	105	749
<i>Cuenca (Spain)</i>													
Temperature (°C)	3	4	7	10	13	18	22	21	18	12	7	4	12
Precipitation (mm)	43	41	70	48	72	50	19	26	43	55	49	55	571
<i>Beirut (Lebanon)</i>													
Temperature (°C)	10.5	11.3	13.7	16.8	20.7	23.8	25.9	26.1	24.5	21.5	16.4	12.4	18.6
Precipitation (mm)	147	143	106	48	19	3	1	1	8	33	80	137	726
<i>San Francisco</i>													
Temperature (°C)	9	11	12	13	15	16	17	17	18	16	13	10	14
Precipitation (mm)	102	88	68	33	12	3	0	1	5	19	40	104	475
<i>Perth (Australia)</i>													
Temperature (°C)	23	24	22	19	16	14	13	14	15	16	19	22	18
Precipitation (mm)	7	12	22	52	125	192	183	135	69	54	23	15	889
<i>Cape Town (South Africa)</i>													
Temperature (°C)	20	20	19	17	14	13	12	12	14	16	18	20	16
Precipitation (mm)	15	17	20	41	69	93	82	77	40	30	14	17	515
<i>Santiago (Chile)</i>													
Temperature (°C)	21	20	18	14	11	8	8	9	11	15	18	20	14
Precipitation (mm)	0	1	3	10	42	70	87	52	22	13	9	2	342

Mean monthly and annual temperature and precipitation values are rounded to the next unit.

Table 2 Soil water balance (mm) for Córdoba (southern Spain; 90 m, 38°51' N, 4°50' W), on the basis of the agricultural year, September–August (mean of 100 years: 1901–2000 period), for a soil with a water-holding capacity of 125 mm.

Month													
Variable	S	O	N	D	J	F	M	A	M	J	J	A	Year
<i>P</i>	27	74	91	87	73	73	79	61	45	19	3	3	635
<i>PET</i>	134	78	46	33	38	48	75	108	140	178	218	194	1288
<i>P – PET</i>	–106	–4	45	54	35	25	4	–47	–95	–159	–215	–191	–653
Soil storage ^a	0	0	45	99	125	125	125	78	0	0	0	0	0
<i>L</i>	0	0	0	0	8	25	4	0	0	0	0	0	0
<i>AET</i>	27	75	46	33	38	48	75	108	123	19	3	3	598

P, mean precipitation; *PET*, potential evapotranspiration according to Penman; *L*, leaching; *AET*, actual evapotranspiration.

^aFor a soil with a water-holding capacity of 125 mm.

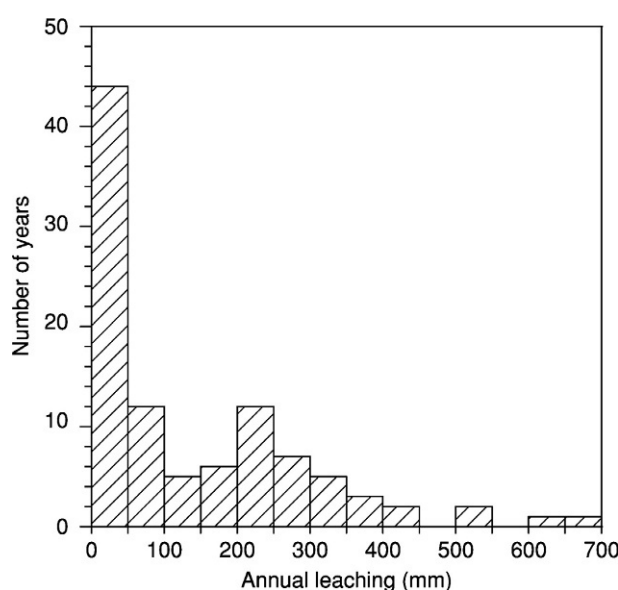


Fig. 1 Frequency distribution of annual leaching in Córdoba, southern Spain, for the period 1901–2000, as calculated from monthly rainfall data for the period, the mean monthly potential evapotranspiration (Table 2), and a soil water-holding capacity of 125 mm.

considerable variation in rainfall, *L* has relatively large values in some years. Mean climatic data thus lead to gross underestimation of the leaching potential of the Mediterranean climate, which is generally significant if the mean annual *P* exceeds 500 mm.

Vegetation

Most Mediterranean areas were once covered with a sclerophyllous forest, which adapts itself readily to both summer drought and light winter frost. The exact nature of this forest in the areas bordering the Mediterranean Sea is difficult to ascertain, but the native vegetation probably included coniferous and broad-leaved evergreen trees. Among the latter, holm oak (*Quercus ilex* L. and *Q. rotundifolia* Lam.), cork oak (*Q. suber*), and Kermes oak (*Q. coccifera* L.) are widespread. Typically sclerophyllous species include wild olive (*Olea europaea* L.), carob (*Ceratonia siliqua* L.), and lentisk (*Pistacia lentiscus* L.). Isolated specimens of stone pine (*Pinus pinea* L.) and large areas of Aleppo pine (*P. halepensis* Mill.) are also commonplace. Degradation of the original forest has often resulted in the establishment of scrub communities that are given local names (*garriga* in Catalonia, *maquis* in France, *macchia* in Italy, etc.) and include cistus (*Cistus* spp.), heathers (*Erica* spp.), gorse (*Ulex* spp.), broom (*Genista* spp.), and thyme (*Thymus* spp.). Some scrub communities probably constitute the original vegetation in areas where either low precipitation or pervious rocks and wind exposure gave rise to habitats that were too dry for the sclerophyllous forest to survive (Dallman, 1998).

In North America, the sclerophyllous forests of western California are rich in species of *Quercus* and *Cupressus* similar to those of the Mediterranean basin (e.g., *Q. agrifolia* Née. resembles *Q. ilex*). In areas with mean annual rainfall below 500 mm, a formation similar to the maquis, namely the *chaparral*, reaches 3 m in height and encompasses bush oaks in addition to species of *Ceanothus* and *Arctostaphylos*. In California, vegetation has been shown to rely on the existing deep water of the subsurface throughout the

vadose zone in the dry season, profiting from water at depths up to several meters reaching the weathering front. This demonstrates the role that deeply weathered rock resources can play during periods of water stress (Dawson et al., 2020). In South America, the Chilean sclerophyllous forest exhibits a completely different floristic composition, but that is remarkably similar in appearance to analogous formations in California; *Quillaja saponaria* Molina, *Lithraea caustica* (Molina) Hook. & Arn., and *Peumus boldus* Molina are especially commonplace. Uncultivated land in the Cape region of South Africa is now occupied by sclerophyllous scrub called *fynbos*, which varies between 1 and 4 m in height and exhibits a bluish-green hue; the only tree is *Leucodendron argenteum* (L.) R. Br., which is accompanied by scrub species pertaining to the Proteaceae, Brassicaceae, and Rosaceae families. Finally, in the Southwest and South Australia Mediterranean areas, the vegetation typically consists of *Eucalyptus* species with coriaceous leaves. The *jarrah* forest with *E. marginata* Donn ex Sm. is typical of areas with a mean annual rainfall of 600–1200 mm; on the other hand, the drier (500–600 mm) *wandoo* zone has a less dense forest of *E. redunca* Schauer (Dallman, 1998).

Lithology and relief

The strong tectonic activity present in Chile, California and the Mediterranean basin generate high mountains, which were partially glaciated during the ice ages. Here sedimentary rocks predominate, but plutonic, volcanic, and metamorphic rocks are well represented in most areas. In the Mediterranean region proper, limestones and dolomites constitute a fairly high proportion of the mountains. Frequent erosion and deposition episodes since the Upper Tertiary have led to the formation of large surfaces covered by fine-grained sediments at the foot of the mountains. These tectonically active regions tend to have soils that are richer in nutrients than those found in tectonically stable Mediterranean areas of Australia and South Africa (Casanova et al., 2021). While the landscape of the South African Cape province comprises ancient low altitude ranges, with valleys between where agriculture is developed, in Australia the extreme flatness of its landscape stands out (Fey, 2010; McKenzie et al., 2004). Here, ancient and stable cratons exposed rocks to millions of years of weathering during periods of different climatic conditions than nowadays, resulting in highly developed soils that have nothing to do with the current climatic conditions.

The abundance of steep slopes in mountainous regions has contributed to make water erosion the principal process in slope development, which was particularly active in drier periods during the Quaternary (e.g. interglacial) with scarce or uneven rainfall and soils without a continuous vegetation cover. In periods of more even rainfall, slope development under a continuous vegetation cover was probably more markedly influenced by soil creep and subsurface water erosion, similar to slopes in humid temperate regions. The strong influence of water erosion is apparent in the slope profiles of many lithologically uniform areas, which, as predicted by geomorphic models, consist of a long, concave element at the foot, a straight middle segment, and a short convex summit.

Structural reliefs with parallel-retreating slopes are commonplace in areas of sedimentary rocks (particularly in the drier regions) and tectonically 'passive' regions. These slopes commonly merge into pediments. Depositional pediment surfaces, which are more common in Mediterranean areas than in humid temperate regions, generally have slopes of less than 1–2% and occasionally extend over tens of square kilometers. Some such surfaces are very old and are covered by strongly weathered soils. Frequently, the reverse of pediment surfaces form with V-shaped creeks or valleys when those surfaces have been dissected to a greater or lesser extent by a central stream and its tributaries.

Studies of soil production on hillslopes in Chile and California indicate that Mediterranean climates, which are usually in a transition zone between arid and humid climates, have higher weathering rates (hence soil production) than their climatic neighbors. This results in a lower residence time (<100 ka) of regolith and therefore a greater contribution of nutrients from bedrock decomposition. In this sense, it has been shown that forests in California and Chile at mid-range rainfall (400–700 mm year⁻¹) are highly dependent on the flow of nutrients from the weathering of the underlying rock (Schaller et al., 2018).

Alluvial plains and their contiguous terraces dominate landscapes over stretches of tens of kilometers in the middle and lower parts of major river valleys, where they support irrigated agriculture. River terrace systems provide excellent soil chronosequences for the study of the major soil-forming processes as the systems formed over long periods during the Quaternary. For instance, the oldest river terraces reach heights of more than 200 m above the present river channel in the river Guadalquivir valley (southern Spain) (Pena and Torrent, 1984).

Time

Because the Mediterranean regions worldwide underwent very limited Quaternary glaciation, their landscapes exhibit geomorphic surfaces of widely different ages. Young Holocene landscapes are found widely in fluvial plains, dune fields and mountainous areas around the Mediterranean basin, California and Chile. In these areas, these younger deposits mesh with older surfaces of Pleistocene age, which are well represented in sequences of fluvial and marine terraces. In some places, those surfaces date back to the Pliocene (e.g. the *raña* surfaces in several parts of Spain), where well developed soils of polygenetic origin are found (Gallardo, 2015). Many features present in Mediterranean soils are attributed to wetter periods during glaciations such as rubefaction in *Terra rossa* soils around the Mediterranean basin and clay illuviation in *Duplex* soils in South Africa, especially on older flat surfaces (Fedoroff and Courty, 2013; Fey, 2010). Both processes mentioned above, would take hundreds of thousands of years to form in Mediterranean climates. Australia is at the extreme end in terms of age of soils in Mediterranean areas, with some landscapes in which the soil cover has been in the same place for millions of years. A considerable part of the Australian area with a Mediterranean climate corresponds to the Yilgarian craton of Precambrian age. This ancient non-tectonically active craton has created a landscape with soil patterns that

respond to millions of years of weathering in diverse climates (McKenzie et al., 2004). An intricate mosaic of soil formation times is therefore the rule rather than the exception in most areas with Mediterranean climate. The rapid development in recent years of techniques for dating surfaces and deposits (e.g. cosmogenic nuclides; optically stimulated luminescence) has made it possible to determine the ages of soils, and therefore better constrain the time factor in soil development at a pace that was unthinkable a couple of decades ago (Watchman and Twidale, 2002).

Soil-forming processes

Mass additions and losses

The strong influence of water erosion results in marked catenary differentiation in mountainous and hilly areas. Thin, occasionally skeletal soils on the upper part of the slope give way to increasingly thick soils as the footslope is approached. Increasing thickness is usually associated with increasing textural fineness and with changes in the mineralogical assemblage of the clay fraction.

Most Mediterranean areas lie at the fringe of deserts (e.g., the Sahara, Mojave, Atacama, Great Victoria, Namib and Negev deserts). This has resulted in more or less frequent episodes of eolian dust addition. Direct measurement of the dust carried by 'mud rains' and detailed mass balances in soils close to the Mediterranean sea have yielded or supported figures of a few grams per square meter and per year. Dust accretions, gradually assimilated by the soil during the Quaternary, may account for up to 50% of the mass of some soils. In the Mediterranean region the textural classes of eolian deposits of silty textures are assimilated into the classic definition of loess soils, nonetheless coars silt and fine sand are also present (Coudé-Gaussen, 1990). The deposition of atmospheric dust must be considered as an important factor in the formation of certain Mediterranean soils, such as Terra rossa (Yaalon, 1997).

Leaching and redistribution of calcite and silica

Weak-to-moderate winter leaching enables weathering of silicates and significant translocation of basic cations. Thus, mature soils that developed on base-poor parent materials exhibit acid pH values in the surface horizons, with pH and base saturation generally increasing with depth. If the parent material contains calcite, dolomite, or Ca-rich weatherable minerals and winter leaching is scarce to moderate, Ca precipitates as secondary calcium carbonate (in the form of calcite) in some soil horizons. Calcic horizons with carbonate nodules of different forms and sizes occur in soils in areas with $P < 700\text{--}800\text{ mm}$ and are replaced by indurated calcrete (petrocalcic) horizons where $P < 400\text{--}500\text{ mm}$. Long stretches of the dry Mediterranean areas (e.g., La Mancha region in central Spain) are covered by soils with calcrete (Torrent and Cabedo, 1986). The depth at which carbonate accumulates is a function of not only the climate and parent material, but also relief and, particularly, time (Fig. 2). Additionally, palustrine calcretes also develop in lowlands of semi-enclosed drainage basins, constituting what the literature describes as azonal soils. Based on ages obtained by absolute dating, the formation of calcrete takes several tens of thousands of years to reach its most advanced stages. While Ca-rich

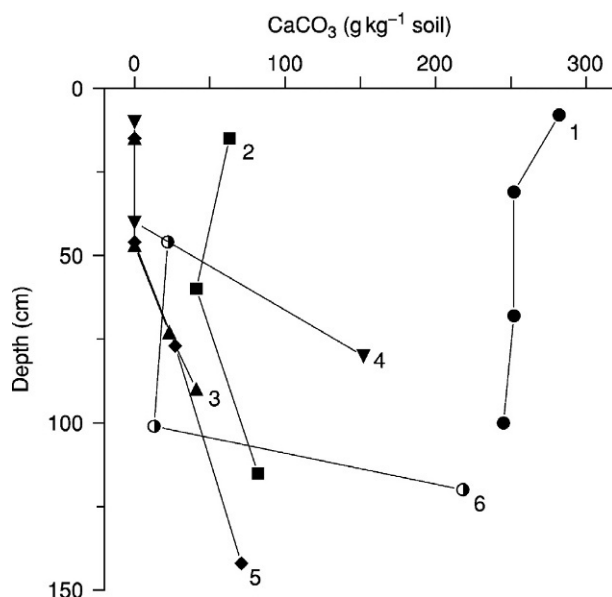


Fig. 2 Depth distribution of CaCO_3 in soils of a river terrace sequence (Henares River, central Spain). Numbers denote relative soil age (1, recent alluvium; 6, highest terrace). Data from Díaz MC (1986) *Estudio de los Procesos de Rubefacción de los Suelos de las Terrazas de la Cuenca del R/ó Henares y del R/ó Jarama*. PhD thesis, Universidad Politécnica de Madrid, Spain.

soils are widespread in the Mediterranean region and some areas of South Africa and Australia, these are not as common in the Mediterranean areas of Chile and California.

While silica and mixed silica–carbonate horizons are less abundant than carbonate accumulations in the Mediterranean basin, they are present in all other Mediterranean areas with their distributions being remarkable in some areas of Australia and South Africa. The release of silica is favored by weathering of volcanic materials or a large presence of sodium additions by rain, fog or dust. Silica is slightly soluble at pH above 8.5, and it is translocated during winter and precipitated as hydrated amorphous silica (opal) in the form of nodules or crusts (duripans) during the summer. Its presence is associated with old stable land surfaces in dry areas with a winter rainfall patterns. Examples of this are the San Joaquin duripan in California of fluvial origin, the durbank soils of South Africa, where marine spray plays an important role, and the deeply weathered landscapes of southwestern Australia, where in many cases they preserve old surfaces from erosion (Torrent et al., 1980b). In Australia silcretes or duripans can be the result of many cycles of landscape and soil formation over millions of years under different climates.

Clay translocation

Clay translocation (illuviation) has been, and continues to be, a major process in the genesis of Mediterranean soils, particularly in the more humid areas and periods. Because of the strong flocculating effect of polyvalent cations like Ca or Mg, decalcification is required for clay particles to disperse and be carried by percolating water; however, translocation of clay has occasionally been observed in calcareous soils. Summer drought seems to favor clay translocation, probably because, when the soil dries, cracks develop and clay on the freshly exposed surfaces is easily dispersed and transported by the water percolating after the first autumn rains (Sauer, 2010). Clay translocation is generally a slow process, but soils with horizons rich in illuviated clay (argillic horizons) have formed within 2000–3000 years on the alluvium of some rivers of Spain. However, well developed argillic horizons with notable clay films could take over 50,000 years to form under Mediterranean climate conditions. Summer drought generally contributes to disruption of the clay cutans resulting from clay translocation (Fig. 3). Thus, evidence for clay translocation is provided by the presence of undisturbed cutans in deep soil horizons, or by the increase in the fine-to-total clay ratio with increasing depth (Fig. 4), since the fine clay particles are assumed to be more mobile than the coarse ones (Torrent et al., 1980a).

Rubefaction

Rubefaction (i.e., reddening due to pedogenic hematite formation) is quite common in the Mediterranean region. It follows decalcification because the high pH of calcareous soils hinders weathering and Fe release from primary minerals. Mediterranean climate conditions are particularly favorable for hematite formation, as during winter there is sufficient moisture for Fe to be released by weathering. Iron precipitates as ferrihydrite, which is dehydrated and transformed to hematite during the dry and warm Mediterranean summer. Appropriate environmental conditions for the formation of hematite from ferrihydrite include a high temperature, dry conditions, low organic matter content, near-neutral pH, and large concentrations of Ca^{2+} and Mg^{2+} . In the absence of one or more of these conditions, goethite (a yellow Fe oxide) is formed (Sauer, 2010).

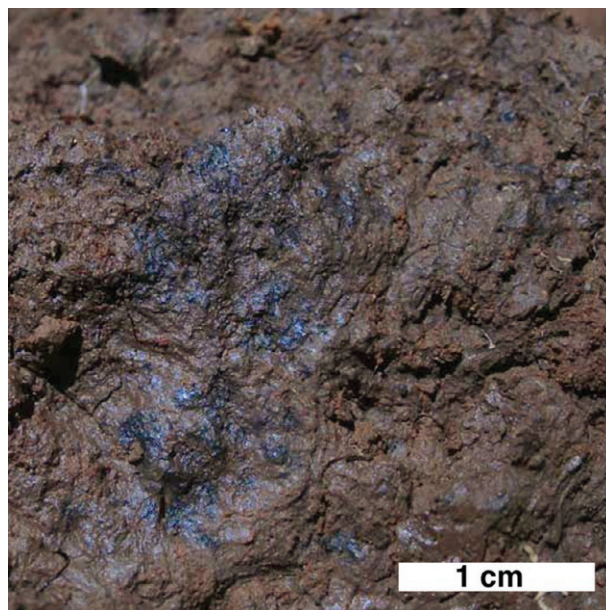


Fig. 3 Clay cutans (films) on a ped from a Rhodic Paleudult (Ultisol) in the Nahuelbuta mountain range, at the southern edge of the Mediterranean climate region of Chile. Photograph credit: Marco Pfeiffer.

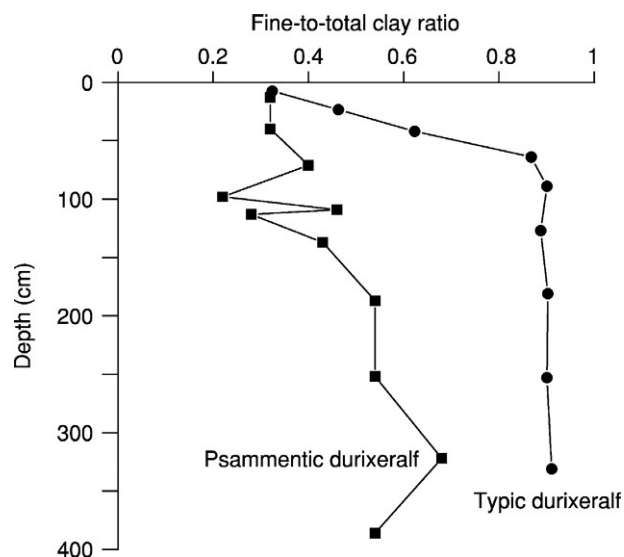


Fig. 4 The increase with depth in the fine-to-total clay ratio in two deep soils of southern California is taken as evidence for clay translocation. Data from Torrent J, Nettleton WD, and Borst G (1980a) Clay illuviation and lamellae formation in a Psammentic Haploxeralf of Southern California. *Soil Science Society of America Journal* 44: 363–369; and Torrent J, Nettleton WD, and Borst G (1980b) Genesis of a Typic Durixeralf of southern California. *Soil Science Society of America Journal* 44: 575–582.

The vivid red hue of some Mediterranean soils has attracted the attention of pedologists, geologists, naturalists, and travelers. Terms such as the Italian ‘Terra Rossa’ or ‘red Mediterranean soils’ have been incorporated into some classification schemes and used (and abused) by pedologists to denote different red-colored Mediterranean soils. Indeed, pedogenic hematite possesses a strong pigmenting power, so a concentration of more than 10 g kg^{-1} soil often results in reddish brown or redder hues. Redness ratings derived from the Munsell color notation and the soil hematite content are well correlated (Fig. 5).

Soil chronosequences indicate that rubefaction and clay illuviation develop at similar rates once the upper soil horizons have been decalcified. These three processes play a crucial role in soil development on stable geomorphic surfaces. This is shown in Fig. 6, which depicts an idealized development sequence on calcareous alluvium under an ‘average’ Mediterranean climate. Soils comparable with members of this sequence are present on widely different parent materials that range from hard igneous rocks to unconsolidated sediments. Two related development sequences must be considered here. One has soils with no carbonate

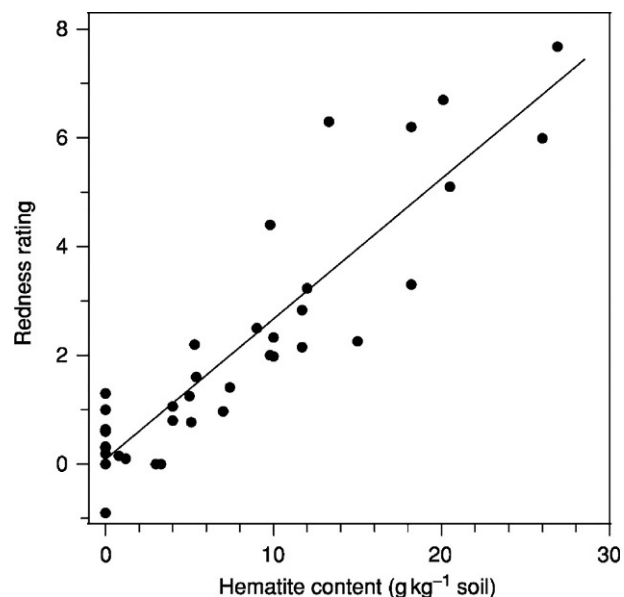


Fig. 5 Relationship between a redness rating derived from the Munsell color notation and the hematite content in a group of Mediterranean soils. Data from Peña F and Torrent J (1984) Relationships between phosphate sorption and iron oxides in Alfisols from a river terrace sequence of Mediterranean Spain. *Geoderma* 33: 283–296; Torrent J and Cabedo A (1986) Sources of iron oxides in reddish brown soil profiles from calcarenites in southern Spain. *Geoderma* 37: 57–66.

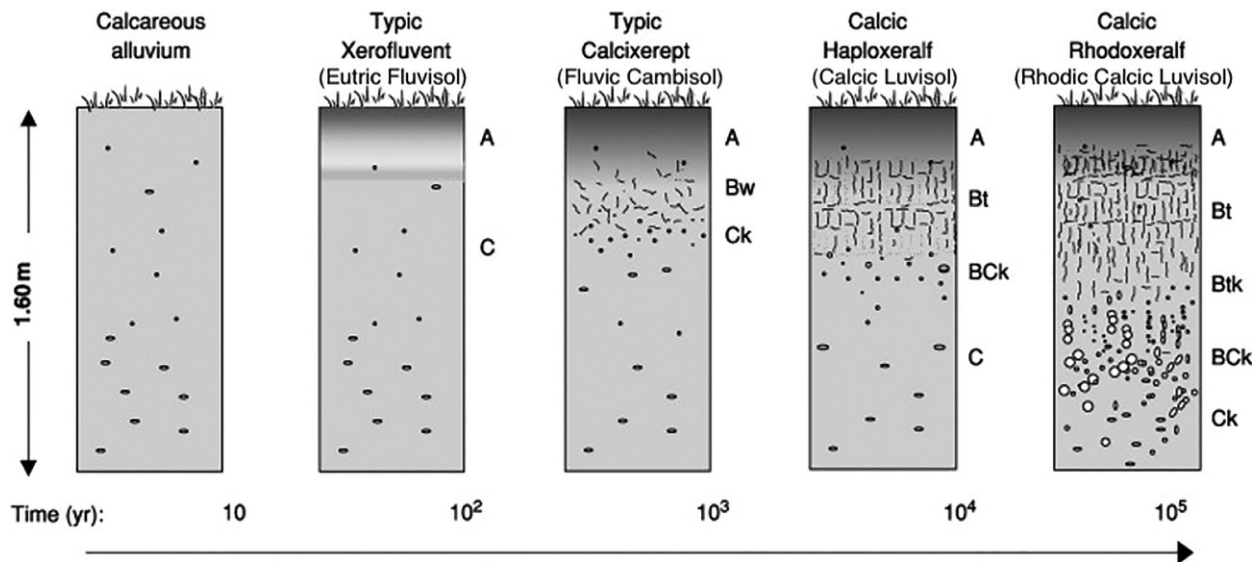


Fig. 6 Idealized soil development sequence from a calcareous loamy alluvium in a typical Mediterranean environment. Soil Taxonomy names above, in parenthesis are WRB names.

accumulation, so it must be assigned to base-poor rocks. In the other, the old soils exhibit a seasonally perched water table that develops as a result of pore clogging by illuviated clay. In this case, redoximorphic features appear and the red hue is lost due to reductive hematite dissolution.

Development of vertic features

The xeric regime favors development of the vertic features typical of Vertisols and other clayey soils that usually experience several major shrink–swell cycles each year. Vertic features seem to develop over short pedogenic times in clayey sediments as recent alluvium or Tertiary marls. Furthermore, vertic features are enhanced by smectite formation in the soils of footslopes and closed basins as a result of lateral illuviation of silica and bases. For this reason, Vertisols occupy the lowest topographic positions of some Mediterranean soilscapes, even though the parent material may not be clayey.

Major soil types

Many pedologists would subscribe to the idea that the Mediterranean climate dictates the predominance of some specific soil-forming processes and thus justifies the term ‘Mediterranean soils.’ However, pedodiversity is probably higher in the Mediterranean region than in any other climatic zone (Ibáñez et al., 2013; Ibáñez et al., 1998). How can these two conflicting views be reconciled? Probably by remembering that: (1) the Mediterranean climate encompasses a rich variety of subclimates and (2) the effect of the Mediterranean climatic and biotic environments is modulated by the considerable lithological and geomorphic diversity resulting from the complex tectonic and climatic history of Mediterranean regions.

Various classification schemes have been used – some only at the national level – to encompass the Mediterranean pedodiversity. Some of the older schemes incorporated geographical and color terms (e.g., ‘Black Andalusian Earths’) with somewhat diffuse definitions. Most pedologists have by now adopted the FAO *World Reference Base for Soil Resources* and the Soil Taxonomy System developed by the Soil Survey Staff of the United States Department of Agriculture (USDA). The latter is used here for a cursory description of the main soil types.

Eleven of the 12 orders in Soil Taxonomy are represented in the Mediterranean regions. Soil orders in decreasing area of coverage are Entisols, Aridisols, Inceptisols, Alfisols, Mollisols, Andisols, Vertisols, Spodosols and Ultisols. While Histosols and Oxisols are rare and very localized (Fig. 7).

Entisols are widespread in all Mediterranean regions, with an important presence in Mountainous areas and young fluvial deposits along the Mediterranean basin, and in dune fields in South Africa and Australia. Entisols, which exhibit minimal pedogenic alteration, are particularly abundant in the mountainous or hilly areas where water erosion and other mass-wasting processes keep the soil permanently thin. These soils are usually classed as Xerorthents. Less abundant, but agriculturally important, are Xero-fluvents, which occupy alluvial plains and some colluvial areas. Psamments (sandy Entisols) are confined to alluvial, coastal dunes, and sandy continental deposits, and Aquents (wet Entisols) are confined to some depressions, deltas, and alluvial flats. It is quite common for the soils of these two suborders to underlie ecologically significant rather than agriculturally valuable areas.

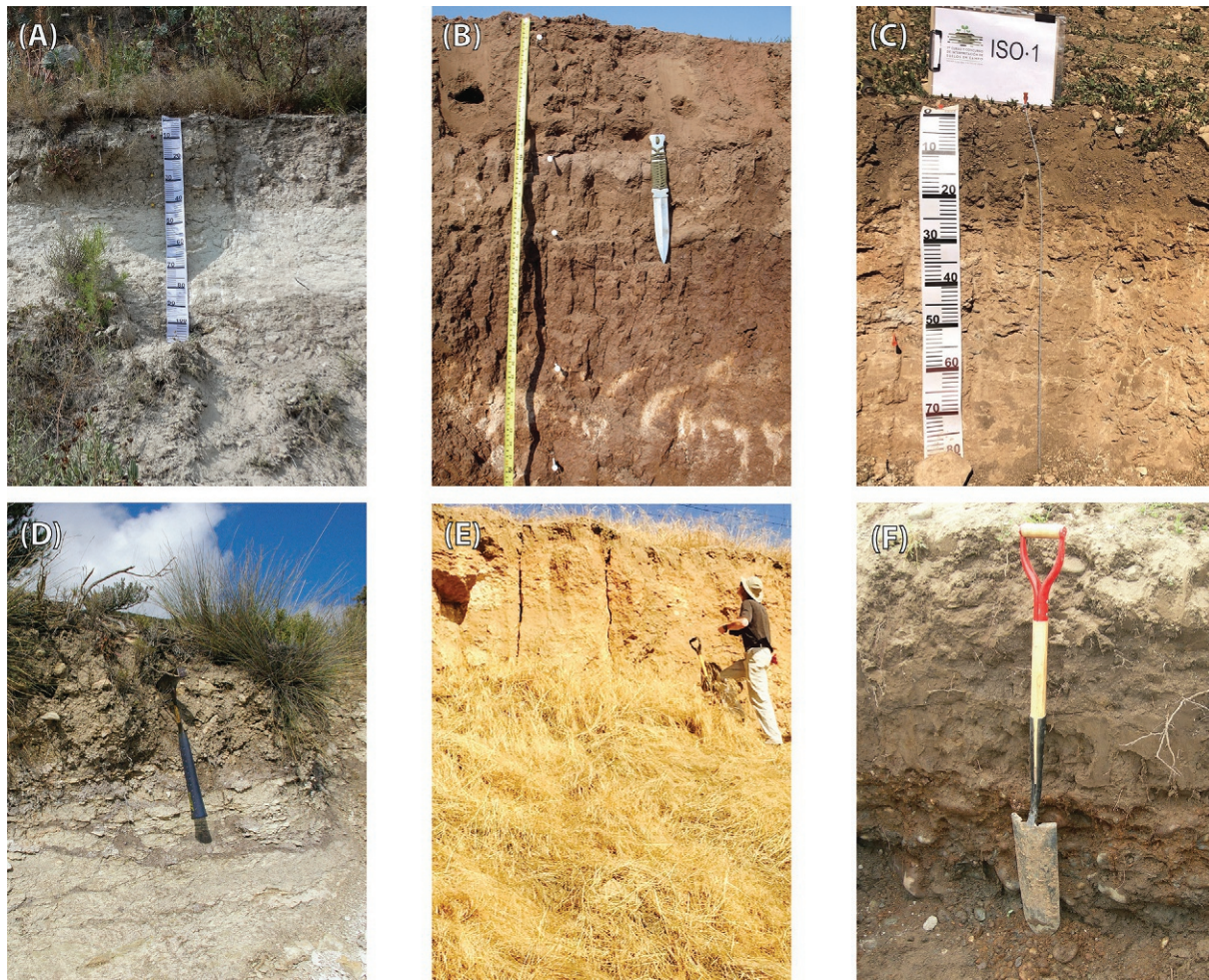


Fig. 7 Selection of major Mediterranean soil types. (A) Entisol (Typic Xerorthent) from Sierra de Mariola, Alicante, Spain; (B) Aridisol (Typic Natrargid) from Huentelauquén, Coquimbo, Chile; (C) Inceptisol (Petrocalcic Calcixerapt) from Tresp, Lleida, Spain; (D) Inceptisol (Calcixerapt) from Onil, Alicante, Spain. (E) Alfisol (Abruptic durixeralf) from San Joaquin, California, USA; (F) Mollisol (Typic Calcixeroll) from Central Chile. Photograph credits: A, C and D Jorge Mataix-Solera; B, E and F, Marco Pfeiffer.

In contrast to what was previously thought, Aridisols cover vast areas in the driest parts of the Mediterranean regions with an important presence in North Africa and Australia. In total, they cover about a quarter of the areas with Mediterranean climate. The suborder Calcid is the most common and it usually occupies flat to undulating plains. They can form from weathered limestone rocks, eolian blown and lacustrine sediments. Calcids also can occur at different stages of development from carbonate masses at calcic horizons to well-developed petrocalcic horizons forming calcretes. Argids, tend to form on the old stable geomorphic surfaces and they are mostly attributed to wetter paleoclimates that allowed clay illuviation. Occupying much smaller areas, Cambids can be locally important for agriculture. Gypsis, Cryids and Salids are not abundant, however the latter frequently host ecologically unique areas in deltas, enclosed depressions, and coastal plains.

Inceptisols in Mediterranean areas often represent steps in evolutionary sequences ranging from raw soils (Entisols) to Alfisols or Vertisols, but they are truly mature soils in other cases. Most are Xerepts, but Udepts, Ustepts and Aquepts are also present to a limited extent. Xerepts include soils with calcic or petrocalcic horizons (Calcixerpts) and acidic soils (Dystroxerepts). They occur in all types of landscapes and climatic areas, and include agricultural and forest soils. Aquepts are restricted to alluvial and coastal plains, and closed basins; in addition, like Aquents, some are environmentally significant and can be turned into good agricultural soils after drainage.

Alfisols are widespread and correspond to most of the development stages of the chronosequence in Fig. 6 (and similar chronosequences). Alfisols are generally found on stable geomorphic surfaces over a wide variety of parent materials and are represented by the Xeralf, Ustalf and Aqualf suborders. Aqualfs are locally significant in depressions, closed basins, and plateaus, and they tend to exhibit a perched rather than a permanent water table. Among the Xeralfs, Durixeralfs (with duripan) are generally restricted to volcanic areas, and Fragixeralfs (with fragipan) and Plinthoxeralfs (with plinthite) are relatively rare. Rhodoxeralfs, which have vivid red colors, are commonplace – but not so much so as the Haplo- and Pale-xeralfs – and cover a variety of well-drained parent materials (calcarenes, hard limestones, arkoses, sandstones, colluvial of acid rocks, basalts, etc.). Terra rossa

sensu stricto (i.e., a red, clayey soil on pure, hard limestone) generally belongs to the Rhodoxeralf group. Reddish-brown Xeralfs not meeting the stringent color requirements for Rhodoxeralfs are also commonplace and are generally classed as Haploxeralfs. Palexeralfs ('old' Xeralfs) generally occupy old geomorphic surfaces such as higher river or marine terraces and ancient pediments. Some possess a petrocalcic horizon; others tend to support a perched water table. Haploxeralfs, however, which are regarded by many as the 'ragbag' of Xeralfs, constitute the most widespread great group.

The presence of a dark, well-structured (Mollic) epipedon (a requirement for the Mollisol order) was probably common at some time under the typical sclerophyllous vegetation. Man's activities (deforestation, plowing, grazing, etc.) have resulted in substantial losses of organic matter and deterioration of the good structural properties of the epipedon. Assessing the extent of such changes, however, is a difficult task. At present, less than 10% of the Mediterranean region is covered by Mollisols (mostly Xerolls), which tend to be concentrated in mountainous areas, man-made grasslands, and other uncultivated areas.

Andisols are mostly present in Italy and Chile (Xerands) and are located in areas where intense volcanic activity meets the climatic conditions that allow the process of andosolization to happen, namely the formation of allophanic nanominerals and Al-humus complexes. In flat areas, these soils play an important role in agriculture as they have very good physical and chemical properties. In areas with steep slopes Andisols are at high risk of landslides due to their particular mechanical/thixotropic properties.

Vertisols form readily where the parent material is clayey or the local conditions favor the formation of smectite. Mediterranean Vertisols (Xererts) occupy relatively large areas on Tertiary marls and Quaternary clays (e.g., in the Guadalquivir river valley in Spain and some plains of western Morocco, where smectite is largely inherited from the parent material ('lithogenic' Vertisols)). In closed basins, footslopes, and some poorly drained areas, where smectite is neoformed ('topogenic' Vertisols), Xererts alternate with Aquerts and gradually merge into Alfisols at some higher topographic locations. Among Xererts, Calcixererts are probably more abundant than Haplo- and Duri-xererts, because the low hydraulic conductivity hinders decalcification in Vertisols.

Ultisols exhibit low base saturation in the subsoil, so they can rarely develop under the rainfall regime imposed by the average Mediterranean climate. However, they do occur in areas where at least two of the following conditions are met, namely: (1) high (more than 200–300 mm) winter leaching; (2) base-poor parent materials (e.g., slates or quartzites); and (3) a stable geomorphic surface. These three conditions are fulfilled, for instance, by Xerults and Aquults in Pliocene *raña* formations, which occupy more than 10,000 km² in Spain. They are generally associated with Ultisols on these old surfaces, soils with plinthic properties appear sporadically, and even bauxites. The presence of Nitisols has also been documented. Many of these began to form in the Pliocene and have undergone various climatic changes over time, which is why we can speak of polycyclic paleosols.

Finally, only the coincidence of unusual local conditions favors the occurrence of Spodosols, Andisols, and Histosols. For instance, Spodosols in western Portugal form on coastal dunes under pine forests in a subhumid Mediterranean environment. Leached podzols in the South African Cape Region, south Western Australia and Californian marine terraces have their location on old landscapes in common, while the particular conditions of their formation are not fully understood. Small patches of spodosols occur locally in high mountains under acidifying vegetation and humid environmental conditions, as has been reported in Spain.

Table 3 shows the probable extent for the most common soils in the Mediterranean region. These figures are only tentative, given the relative scarcity of detailed soil maps for some Mediterranean countries.

Physical, chemical, and mineralogical properties

Many areas bordering the Mediterranean have been deforested and cultivated for centuries and even millennia. Man-induced water erosion and organic matter depletion have affected extensive areas, down to unsustainable limits, i.e. badlands. Soil degradation in other Mediterranean areas has been less marked, but current risks from agriculture, forestry, industrial activities, or modern urban needs are considerable in most areas. Soil properties must be considered from the standpoint of their significance and evolution in different human-influenced scenarios, which change – sometimes dramatically – with time.

Physical properties

In general, the quality of Mediterranean soils is more influenced by physical than by chemical properties. Moreover, some physical properties cannot be greatly modified (e.g., texture) or changed without much effort (e.g., structural stability).

Mediterranean soils tend to be more clayey than soils from cool temperate areas. Textural contrasts such as those exhibited by Alfisols are common and generally result in water-holding capacity increasing with an increase in depth, which favors growth of rain-fed crops. Stoniness is widespread and generally reduces crop production; however, surface stoniness helps to reduce surface evaporation, protects soil against splash erosion, and delays runoff.

The weak structural stability of many cultivated soils has been attributed to their low organic matter content. Various studies, however, have revealed other factors (Fe oxides, Ca saturation, clay type) to be equally important with a view to creating good structure. Soil crusting affects many soils rich in silt and fine sand, where it hinders the emergence of small seedlings and triggers runoff. Hardsetting A horizons, which combine structural instability and high strength when the soil is dry, have been reported in the Xeralfs of Southwest Australia and some countries bordering the Mediterranean.

Low water availability in summer limits the yield potential of Mediterranean soils, even though they possess, on average, a water-holding capacity exceeding that of soils in humid temperate regions. Soil crusting and poor infiltration may offset the latter advantage, however.

Table 3 Soil types most commonly found in Mediterranean areas.

<i>Order^a</i>	<i>Suborder^a</i>	<i>Group^a</i>	<i>WRB^b</i>	<i>Probable extent (%)</i>	
Entisols	Orthents	Xerorthents	Leptosols, Regosols	20–40	
	Fluvents	Xerofluvents	Fluvisols	1–3	
	Psamments	Quartzipsamments	Arenosols	<1	
		Xeropсамments	Arenosols	1–3	
Aridisol	Calcid	Haplocalcid	Calcisol	8	
		Petrocalcid	Calcisol	4	
	Argids	Haplargid	Luvisol	6	
		Paleargid	Luvisol	3	
	Cambid	Haplocambid	Cambisol	<1	
	Salid	Haplosalid	Solonchak	<1	
	Inceptisols	Xerepts	Calcixerepts	Calcisols	5–10
			Dystroxerepts	Cambisols	3–5
		Haploxerepts	Cambisols	5–10	
Alfisols	Aquepts		Gleysols	<1	
	Xeraifs	Haploxeraifs	Luvisols	3–5	
		Palexeraifs	Luvisols	3–5	
		Rhodoxeraifs	Luvisols	<1	
		Durixeraifs	Durisols	<1	
	Aqualfs		Luvisols	<1	
	Mollisols	Xerolls	Calcixerolls	Kastanozems	2–4
			Haploxerolls	Phaeozems	2–4
Andisol	Xerand		Andosol	<1	
Ultisols	Xerults	Haploxerults	Acrisols	<1	
		Palaxerults	Acrisols	<1	
Vertisols	Xererts	Calcixererts	Vertisols	<1	
		Haploxererts	Vertisols	<1	
	Aquerts		Vertisols	<1	

^aAccording to Soil Survey Staff (1999) Soil taxonomy, 2nd edn. *Agricultural Handbook 436*. Washington, DC: USDA.

^bIUSS Working Group WRB (2014) *World Reference Base for Soil Resources*. Rome: FAO.

Mineralogical properties

The lack of coincidence of high temperatures and high moisture contents prevents intensive weathering in the Mediterranean environment. Therefore, highly weathered soils occur only on old, stable geomorphic surfaces in subhumid regions. Generally, because of the low-to-moderate degree of weathering, many of the silicate clays are inherited from the parent material, or produced by transformation of primary micas rather than neoformed from solution. Illite comprises approximately 50% of the clay fraction on average, as it abounds in sedimentary rocks and is produced by the transformation of muscovite.

The overall sequence of abundance of clay minerals is illite > smectite, illite–smectite mixed-layers > kaolinite > chlorite, vermiculite, hydroxy-interlayered vermiculite >> halloysite, palygorskite, pyrophyllite, talc. The composition of the soil solution in many soils falls near the lines bounding the kaolinite and smectite stability fields. There is indeed evidence that kaolinite in the more leached soil environments and smectite in solute-rich environments are neoformed in substantial amounts. Neo-formation of illite has been documented in K-rich solutions, such as those in the microcracks of potassium feldspar grains. The presence of different microenvironments, leaching gradients in the soil profile, and lateral eluviation combine to create differences in clay mineralogy with depth and along *catenas*. Fig. 8 depicts an idealized spatial distribution of the minerals present in the clay fraction along a typical catena.

Iron oxides are the most abundant among non-silicate clay minerals; they typically account for 5–10% of the clay fraction. The Mediterranean pedoenvironment favors crystallization of ferrihydrite to goethite and hematite, so that the ratio of poorly crystalline to crystalline Fe oxides is generally low. Aluminum oxides are rare, but gibbsite is present in some excessively drained soils formed on coarse-grained, acidic plutonic rocks.

Among the abundant carbonates, calcite and magnesium calcite are either inherited or secondary, whereas dolomite is always inherited. Calcite crystals can be dispersed in the soil matrix or aggregated in nodules, tubules, or crusts. Crystals range in size from clay to fine sand, but the fine silt fraction is generally the most abundant.

Chemical properties

The pH of Mediterranean soils generally lies within the slightly acidic to moderately alkaline range and, because of the occasionally steep vertical leaching gradient, it can increase substantially with depth. Thus, the idealized mature soils of Fig. 6 typically have a pH

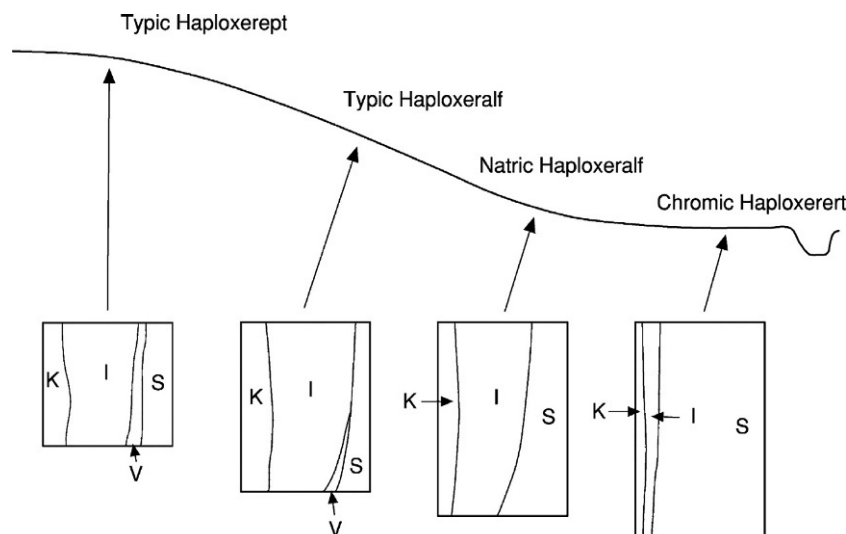


Fig. 8 Relative abundances of silicate clays in the clay fraction of soils of a hypothetical catena on a granodiorite saprolite in a typical Mediterranean environment. K, kaolinite; I, illite; V, vermiculite; S, smectite.

of around 6 on the surface and around 8 in the horizons where carbonate accumulates. However, plowing and accretion of dust from nearby or distant sources often result in recalcification of surface horizons. Consequently, the highly acidic soils formed on old geomorphic surfaces from acidic rocks are usually far from areas with calcareous soils and rocks.

The characteristic mineral assemblages of Mediterranean soils endow them with interesting quantitative adsorption features. Thus, the cation exchange capacity of the clay fraction generally ranges from 40 to 60 $\text{cmol}_c \text{ kg}^{-1}$ as a result of the prevalence of 2:1 clay minerals. Specific ion adsorption by hydroxylated mineral surfaces largely arises from Fe oxides and, to a lesser extent, from the edges of clay minerals and aluminum oxides. Calcite surfaces also play a prominent role in adsorption and heterogeneous precipitation processes. Quantitative relationships between adsorption capacity and simple mineralogical or chemical properties can be established for most of the inorganic adsorbates of interest (e.g., phosphate).

In general, the adsorptive properties of Mediterranean soils depend strongly on depth as a result of steep gradients in pH, particle-size distribution, and mineralogical composition. Consequently, the appropriate 'filters' for the different nutrients and pollutants lie at different depths – and make morphological and mineralogical characterization of the soil profile necessary.

From the standpoint of chemical fertility, Mediterranean soils compare favorably with soils in other geographic areas. After N, P is the most limiting nutrient in native soils; fortunately, most soils are acceptably responsive to fertilization with P. Most soils are well supplied with K, because of the high illite content, and contain adequate levels of Ca and Mg. However, Fe, Cu, Mn, Zn, and B deficiencies frequently arise in some horticultural crops cultivated on calcareous soils.

Salinization, occasionally accompanied by specific ion toxicity problems, has made its way in both old and recently irrigated areas. Even though the principles for good management of irrigation have been known for decades, salinization continues to be a major issue; the scarcity of good irrigation water in most areas, and the use of water-saving irrigation systems, contribute to the problem.

Main threads

Mediterranean regions suffer from environmental pressures that mainly derive from its greatest attribute: their privileged climate makes them attractive places to settle as well as for the development of agriculture. Most Mediterranean areas have suffered from long and intensive exploitation for agriculture, which competes with human settlements for fertile soils. The rapid population increase over the last Century has resulted in a continuous loss of natural areas and fierce competition for the scarce soil and water resources of these areas. In addition to this, human impacts in Mediterranean regions are accentuating the effect of climate change through resources capture, intensive land use change, soil degradation and an increasing occurrence of forest fires.

One of the biggest challenges that threaten these particular climatic regions is a substantial decrease in water availability, as reports indicate that climate change will dramatically affect the water balance through an increase in temperature (and therefore evapotranspiration) and decrease in precipitation (Deitch et al., 2017; IPCC, 2013). These will place increased stress on ecosystems and increase the conflict between water users. This together with increasing land degradation because of poor soil and water management, a decrease in soil organic carbon, salinization of water and depletion of aquifers among other factors (Ryan et al., 2006) could directly impact local populations. These factors could also affect access to good quality drinking water and food security through the increase in production costs and the moving away from cultivation to urban activities.

References

- Casanova M, Salazar O, Seguel O, Tapia Y, Pfeiffer M, and Eslamian S (2021) Combined agroforestry and rainwater harvesting to reduce soil degradation in Mediterranean zones. In: Eslamian S (ed.) *Handbook of Water Harvesting and Conservation: Case Studies and Application Examples*, pp. 81–102. John Wiley and Sons: London.
- Coudé-Gaussen. (1990) The loess and loess-like deposits along the sides of the western Mediterranean Sea: Genetic and palaeoclimatic significance. *Quaternary International* 5: 1–8.
- Dallman PR (1998) *Plant Life in the World's Mediterranean Climates: California, Chile, South Africa, Australia, and the Mediterranean Basin*. Place of publication: University of California Press.
- Dawson TE, Hahn WJ, and Crutchfield-Peters K (2020) Digging deeper: What the critical zone perspective adds to the study of plant ecophysiology. *New Phytologist* 226: 666–671.
- Deitch MJ, Sapundjef MJ, and Freirer ST (2017) Characterizing precipitation variability and trends in the World's Mediterranean-climate areas. *Water* 9: 259.
- Fedoroff N and Courty M-A (2013) Revisiting the genesis of red Mediterranean soils. *Turkish Journal of Earth Sciences* 22(3): 359–375.
- Fey M (2010) *Soils of South Africa*. Cambridge University Press.
- Gallardo JF (2015) *The Soils of Spain*. Springer.
- Ibáñez JJ, De-Alba S, Lobo A, and Zucarello V (1998) Pedodiversity and global soil patterns at coarse scales (with discussion). *Geoderma* 83(3–4): 171–192.
- Ibáñez J, Zinck J, and Dazzi C (2013) Soil geography and diversity of the European biogeographical regions. *Geoderma* 192: 142–153.
- IPCC (2013) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*.
- IUSS Working Group WRB (2014) *World Reference Base for Soil Resources. International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports No. 106*. Rome: FAO.
- McKenzie N, Jacquier D, Isbell R, and Brown K (2004) *Australian Soils and Landscapes: An Illustrated Compendium*. CSIRO publishing.
- Pena F and Torrent J (1984) Relationships between phosphate sorption and iron oxides in Alfisols from a river terrace sequence of Mediterranean Spain. *Geoderma* 33(4): 283–296.
- Ryan J, de Paw E, Gómez H, and Mrabet R (2006) Drylands of the Mediterranean zone: Biophysical resources and cropping systems. *Dryland Agriculture, Agronomy Monography*. 2nd edn, vol. 23, pp. 577–624. Madison, WI: American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc.
- Safford HD and Vallejo VR (2019) *Ecosystem Management and Ecological Restoration in the Anthropocene: Integrating Global Change, Soils, and Disturbance in Boreal and Mediterranean Forests, Developments in Soil Science*. vol 36, Elsevier, pp. 259–308.
- Sauer D (2010) Approaches to quantify progressive soil development with time in Mediterranean climate—I. Use of field criteria. *Journal of Plant Nutrition and Soil Science* 173(6): 822–842.
- Schaller M, Ehlers TA, Lang KAH, Schmid M, and Fuentes-Espoz JP (2018) Addressing the contribution of climate and vegetation cover on hillslope denudation, Chilean Coastal Cordillera (26°–38°S). *Earth and Planetary Science Letters* 489: 111–122.
- Torrent J (1995) *Genesis and Properties of the Soils of the Mediterranean Regions*. Departamento di Scienze Chimico-Agrarie. Università di Napoli Federico II.
- Torrent J and Cabedo A (1986) Sources of iron oxides in reddish brown soil profiles from calcarenites in southern Spain. *Geoderma* 37(1): 57–66.
- Torrent J, Nettleton W, and Borst G (1980a) Clay illuviation and lamella formation in a Psammentic Haploxeralf in southern California. *Soil Science Society of America Journal* 44(2): 363–369.
- Torrent J, Nettleton W, and Borst G (1980b) Genesis of a typic durixeralf of southern California. *Soil Science Society of America Journal* 44(3): 575–582.
- Watchman AL and Twidale CR (2002) Relative and 'absolute' dating of land surfaces. *Earth-Science Reviews* 58(1–2): 1–49.
- Yaalon DH (1997) Soils in the Mediterranean region: what makes them different? *Catena* 28(3–4): 157–169.

Arid and semi-arid subtropical and tropical ecozone (II)

Bernhard Lucke, FAU Erlangen-Nürnberg, Institute of Geography, Erlangen, Germany

© 2023 Elsevier Ltd. All rights reserved.

This is an update of H.C. Monger, J.J. Martinez-Rios, S.A. Khresat, TROPICAL SOILS I Arid and Semiarid, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 182–187, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00021-7>.

Introduction	332
Climate: Definition and significance for soil development	333
Processes of soil formation	334
The role of aeolian sediments	335
Desert pavements	335
Desert fringe loess	336
Aeolian transport	337
Aeolian sediments and soil formation in semi-arid regions	337
Soil colors	338
Soil fertility, land use, and desertification	338
Soil fertility	338
Land use	338
Desertification	340
Summary and conclusion	341
References	342

Abstract

Soil formation in arid and semi-arid regions is dominated by physical processes and largely determined by available water. Moisture enters and leaves mainly at the soil surface, which leads to incomplete leaching and precipitation of secondary minerals. Soil types are characterized by initial or weak soil development, and parent materials often comprise a significant proportion of aeolian sediments. Dust fixation processes play a role for soil properties because aeolian sediments may be partly re-mobilized. If sufficient water is available, soils can be fertile as base saturation can reach 100%. Soil quality is dominated by physical properties such as particle-size distribution.

Key points

- Soil formation in arid and semi-arid ecosystems is governed by physical processes
- Deep and intense soil formation is rare due to minimal chemical weathering
- Soils are characterized by soil types with initial or weak soil development
- Dust and aeolian sediments contribute significant proportions of parent materials and may create deep soil covers (desert fringe loess)
- Sediment fixation processes play a role for soil properties because aeolian sediments may be partly re-mobilized
- Soils in desert fringes have been used agriculturally since earliest civilizations and can be very productive under well-adapted land use strategies

Introduction

About 30% of the earth's land surface is covered by arid and semi-arid areas (Monger et al., 2005). Due to the unfavorable climate causing a lack of water, human land use and settlement are of limited size and restricted to relatively small parts of this area. Drylands are considered important land reserves for future generations which can be utilized if sustainable, well-adapted land use strategies that address the restricted resources are developed and applied. Arid and semi-arid areas are usually covered by sparse vegetation, and the intensity of soil development is confined accordingly. Many soils in drylands are shallow, stony, or affected by high salt concentrations. This is reflected by the main characteristic soil types, which indicate limited soil development (Table 1). However, despite these natural limitations, desert fringes host the oldest agricultural soils of the world which were intensively used by ancient civilizations. In this context, deep and very productive soils of aeolian origin are present in some areas, mainly at the transition of semi-arid to arid climate (desert fringe loess). Their origin and formation are still debated. This also applies to clay-rich soils on limestones (mostly of reddish color: Terra rossa), which dominate many areas of semi-arid (Mediterranean) climate in the subtropics.

Table 1 Prevailing soil types in arid and semi-arid climates of the subtropics and tropics according to Soil Taxonomy (USDA, 2014) and World Reference Base of Soil Resources (WRB, 2015).

Climate zone	Soil Order (USDA)	Reference Soil Group (WRB)	Description
Arid	Entisols	Leptosols	Weakly developed soils, either very shallow (Leptosols) or without characteristic pedogenic horizon development (Regosols)
		Regosols	
	Aridisols	Arenosols	Soils with aridic (i.e., dry) moisture regime, weak soil development, often enrichment of calcium carbonate and/or gypsum or other salts
		Solonchaks	
		Gypsisols	
Semi-arid	Inceptisols	Calcisols	Moderately developed soils, often of reddish color, with clay-rich B-horizon, and frequently rich in primary or secondary calcium carbonate
		Cambisols	
		Calcisols	
	Vertisols (Alfisols)	Vertisols	Deep cracking soils, often of reddish color, with high clay content. Vertisols shrink and swell with moisture variations (mixing the solum), whereas Alfisols and Luvisols are characterized by clay maxima in the subsoil due to clay illuviation (rare in drylands)
		(Luvisols)	

Climate: Definition and significance for soil development

Arid regions are commonly understood to receive less than 250 mm of mean annual rainfall, whereas semi-arid areas receive between 250 and 500 mm. The terms desert and steppe are used to describe such arid and semi-arid lands, accordingly. However, precipitation alone is insufficient to determine how much water is available to plants. Therefore climate classification systems, such as the Köppen–Geiger system, include mean annual temperature as a second property (Peel et al., 2007). The consideration of precipitation in the context of temperature leads to different demarcations of desert and steppe areas depending on the type of precipitation, e.g., summer or winter precipitation, because subsequent evapotranspiration depletes soil moisture to a larger degree during summer than during winter (Fig. 1).

However, climate zone classifications based on temperature and precipitation are useful only on large spatial scales. Variation at scales <100 km, in particular in transitional areas, are usually not reflected by these systems. Therefore, potential evapotranspiration (PET) has been used with mean annual precipitation P to calculate an aridity index P/PET (UNESCO, 1979). However, PET is not

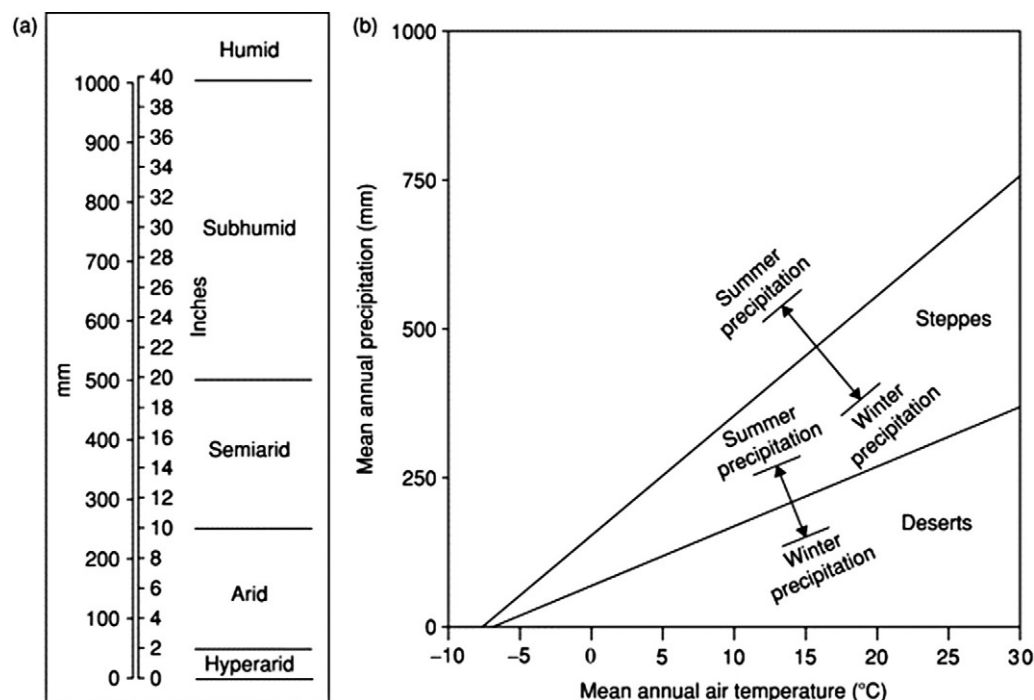


Fig. 1 (a) Climate categories based on mean annual precipitation. (b) Desert and steppe boundaries as a function of both mean annual precipitation and temperature according to the Köppen–Geiger system. Arrows show how boundaries shift according to whether precipitation falls mainly in the summer or winter. From Monger HC, Martinez-Rios JJ, & Khresat SA (2005) Tropical soils. Arid and semi-arid. In Hillel D (Ed.) *Encyclopedia of Soils in the Environment*. 1st edn, pp. 182–187. Amsterdam: Elsevier, 183, Fig. 1.

easily determined as most meteorological stations do not systematically record all properties needed for its calculation. The Penman approach to calculate PET, for example, includes net solar radiation, wind speed, and humidity of the air.

The Thornthwaite approach is a simplified method to calculate PET and is based on only two variables: mean monthly temperature and average number of daylight hours per month. It has been applied to the Negev desert in southern Israel and enabled sub-division of the area into semi-arid, arid, and hyper-arid parts, whereas the Köppen–Geiger system characterizes the whole region uniformly as arid (B)¹ (Bruins, 2012). The more detailed subdivisions based on P/PET are very important for any assessment of land use and soil development because effective soil moisture is more significant to vegetation than annual precipitation and temperature alone. In addition, soil water potential includes the effect of particle-size distribution and salts, which may lead to locally varying soil development processes. Therefore it is difficult to predict or assess soil development or land use potential in drylands solely on the basis of climate classifications.

Processes of soil formation

Jenny (1941) outlined how soil formation is a function of five main factors:

- (1) climate
- (2) organisms
- (3) topography
- (4) parent material
- (5) time

These interact in specific ways to produce arid and semiarid soils (Fig. 2). Climate determines water supply and is therefore the most relevant factor. Water governs mineralogical transfers and transformations, and is essential for vegetation, both with regard to direct use and nutrient supply. (Monger et al., 2005) Vegetation translocates ions by root transfer and bioaccumulation, adds organic matter, and provides ground cover and roots which protect the soil from erosion. Furthermore, vegetation feeds animals which catalyze seed dispersal and affect soils by bioturbation.

The limited availability of water means that nearly all moisture in arid soils and much in semiarid soils enters and leaves via the soil surface. Only in specific topographies, that is in low-lying areas which receive runoff water (e.g., playas or runoff-irrigated terraces), can water permeate to deeper parts of the soil profile. Such soils are mostly not saline if the water table is deep and if the soils have good permeability, because leaching prevents the accumulation of salts. However, salts accumulate in depressions where soils have low permeability (e.g., if clay contents are large or when compaction has occurred), or where shallow groundwater tables are present (typically leading to the presence of salt pans). If a capillary bridge to shallow groundwater is established, evaporation will transport ions to the soil surface where they precipitate as salts and impede the growth of most plants.

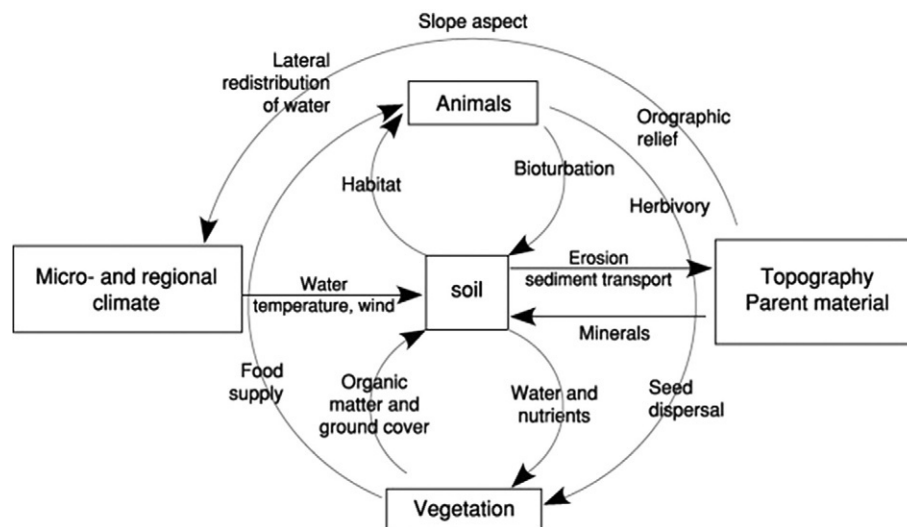


Fig. 2 Illustration of the links among the soil-forming factors of climate, biota (vegetation and animals), topography, and parent material. The fifth factor, time, affects all of the illustrated links in proportion to the duration over which the factors operate. From Monger HC, Martinez-Rios JJ, & Khresat SA (2005) Tropical soils. Arid and semiarid. In Hillel D (Ed.) *Encyclopedia of Soils in the Environment*, 1st edn, pp. 182–187. Amsterdam: Elsevier, 184, Fig. 2.

¹The classification of the Negev as arid (B) according to the Köppen–Geiger System includes steppe climate BSh and desert climate BWh (where mean annual temperatures are $>18^{\circ}\text{C}$), and steppe climate BSk (where mean annual temperatures are $<18^{\circ}\text{C}$).

The movement of water and solutes is not mainly directed to the bottom of soil profiles as in humid areas, but largely toward the soil surface. Leaching is mostly incomplete which favors the precipitation of secondary minerals, such as calcite, silica, and gypsum. These accumulate in the profile, leading to the formation of calcic, petrocalcic, gypsic, and petrogypsic horizons, and of duripans (Monger et al., 2005).

With the limited amount of water, chemical weathering and dissolution of bedrock do not play a major role in arid and semi-arid soils. Rock disintegration, however, can be intense, resulting from the crystallization of salts at the rock surface and from thermal fluctuations. The detritus derived from rocks is transported by gravity, runoff, or wind and contributes to soil formation. In particular wind transport, i.e., the deposition of aeolian sediments, plays a major role in soil development of arid and semi-arid areas.

The role of aeolian sediments

Typical aeolian landscapes, i.e., landforms showing obvious formations of aeolian origin such as dunes, are relatively rare. Deserts do not always host sand dunes (Ergs), but are often characterized by a cover of rock fragments (Hammadas). Aeolian transport is nevertheless an important process of soil development and sediment deposition in most drylands.

Desert pavements

The importance of aeolian processes can be illustrated by desert pavements, which are present in many arid areas. They are characterized by a dense cover of stones and detritus at the soil surface and an accumulation of the fine fraction (mostly of silt and fine sand size, i.e., $<200\ \mu\text{m}$) below. Desert pavements have often been interpreted as the result of intense wind erosion which removed fines, originally from much larger deposits of sediments, leaving the larger size fraction in place until a dense stone cover formed that protected the remaining fines from further wind erosion.

Recent research, however, suggests that sometimes the opposite could be true: various desert pavements might be products of dust accumulation (Fig. 3). This is due to the processes of dust fixation: Aeolian sediments accumulate below stones as these provide wind shadow and preserve moisture, which contributes to the fixation of fines. Dust gathers over time at stone bottoms and is covered by surface crusts of biological and or chemical origin, i.e., formed by lichens, bacteria, mosses and liverworts or secondary minerals such as CaCO_3 (calcium carbonate). Formation of surface crusts is favored by arid conditions as evaporation leads to formation of secondary minerals at the soil surface. In addition, vegetation competes with biological crusts because plant cover reduces solar radiation and suppresses the growth of biological crusts in better-watered areas. These crusts stabilize substrates and are impermeable to air. Therefore, water entering during rain events forces soil air into gas-filled voids, expanding soil volume and creating vesicular layers. These can be found in topsoils below desert pavements. The formation of air-filled voids leads to the uplift of detritus and, over time, to the formation of stone pavements (Turk, 2012). The flat and apparently 'empty' areas in deserts which



Fig. 3 Desert pavement at the Neolithic site of Harrat Juhayra on a basalt field near Ma'an in southern Jordan. An archeological excavation trench exposes a Neolithic wall, and illustrates that silt-dominated sediments were deposited after construction of the wall, burying parts of it. The silt-rich sediments were themselves completely covered by basalt rock detritus, suggesting that this desert pavement is of the accretionary type, characterized by an uplift of the stone cover during the accumulation of fines. Person gives scale.



Fig. 4 Calcareous, silt-dominated sediment covered by detritus in a depression of sandstone rock close to the summit of a mountain near Petra, southern Jordan. GPS gives scale. Note that fine sediments and detritus are bound together by a crust developed on the surface. Such sediments are only present in areas where water gathers, illustrating the role of fixation for dust deposition processes.

are covered by desert pavements are thus landscapes largely formed by aeolian processes, either wind erosion or deposition of aeolian sediment.

The potential role of dust aggradation for the formation of desert pavements underlines the importance of sediment fixation for soil development in drylands (Fig. 4). In this context, the material moving through the atmosphere is not necessarily identical to that of the substrates of soils formed from aeolian sediments, as deposition partly depends on fixation processes. Biological crusts, for example, have been found to fix particles of silt and clay size (Danin and Ganor, 1991), but not sand, which might subsequently be re-mobilized.

Desert fringe loess

Transitional areas between the semi-arid and arid ecozones are sometimes covered by deep, silt-dominated soils, which have properties very similar to loess. Most of them formed during the Pleistocene. However, as no nearby glaciers are known that could have produced the silt dominating the substrate, the origin of these soils is debated. The discussion focuses largely on the sources of silt that could have provided the parent material of desert fringe loess. One theory that has received much attention is that abrasion of moving sand dunes might produce silt-sized particles, including quartz. This could resemble the main parent material of desert fringe loess as many of these soils are downwind from Ergs (Smalley and Vita-Finzi, 1968). Smaller abrasion sediments that were transported further may have contributed to soil development in the semi-arid ecozone (Crouvi et al., 2018). Soil formation could therefore have corresponded to specific meteorological conditions, for example, high wind speeds at some time during the Pleistocene.

Loess deposits, however, are often older than the Ergs which presumably produced their parent material (Roskin et al., 2011). In addition, they are also present in areas not located downwind from Ergs. In this context, various other processes such as fluvial comminution, insolation weathering, salt weathering, volcanism, frost shattering, and deep weathering of saprolite can produce silt-sized detritus. In particular flash floods, which frequently occur in drylands after erratic but intense rains, can lead to substantial abrasion of silt-sized material. Fluvial comminution was found to be the most effective short-time process of silt production, even more effective than glacial grinding (Wright et al., 1998). Therefore, the formation of desert fringe loess might be less a question of parent material, as silt-sized aeolian sediments are omnipresent in drylands, but of the fixation of this material.

Similar to desert pavements, it might be the presence of sediment-fixing agents such as vegetation or crusts in soils that allow aeolian sediments to settle, and prevent re-mobilization (Lucke and Bäuml, 2021). Therefore, properties of the resulting soils might partly reflect sedimentation and fixation processes, which can be illustrated with analogies to dust collectors. Depending on design, artificial dust collectors harvest different components of the material moving through the air. For example, wet samplers were found to collect more clay-sized material than dry collectors. Similar mechanisms could act in natural soils and explain why, for example, aeolian sediments in archeological ruins in the Negev desert and the desert of southern Jordan were found to be statistically similar (Fig. 5). It was possible to model them as one typical deposit, although sediments in Jordan were characterized by more sand contributed from nearby sandstone rocks which are absent in the Negev. Desert fringe loess might therefore be mainly a product of enhanced sediment fixation, possibly connected with specific meteorological conditions such as snowfall, and not to certain dust sources (Lucke et al., 2019).

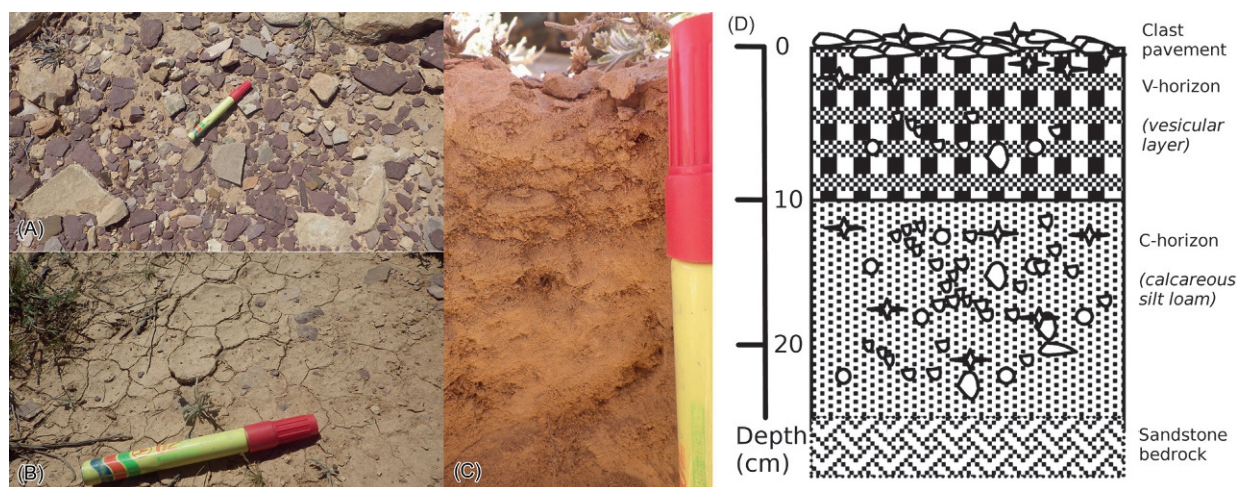


Fig. 5 Composite image showing soil development in hilltop archeological ruin near Petra in southern Jordan. A: Clast cover on soil surface. B: Surface crust below clast cover (after careful removal of clasts). C: Close-up of the upper part of the profile. Pen gives scale. D: Schematic drawing of the soil profile, classified as Calcaric Leptosol (Protic). Picture assembled from Lucke B, & Bäuml R. (2021) Holocene dust dynamics archives in archaeological ruins in arid and semi-arid environments in the Southern Levant. *Arabian Journal of Geosciences* 14(23): 1–15, Figs. 7 and 8, p. 8, edited, CC-BY.

Aeolian transport

Identifying and discriminating between sources of global dust transport, and estimating their contribution to soil development, has been a major research question with important implications for various fields of geosciences. Satellite images have made observations of dust plumes possible, showing that deserts, and in particular the Sahara, are major sources of aeolian sediments. Saharan dust was found to have contributed significantly to soil development in remote areas (Muhs, 2013). Although the main trajectories of dust transport are well known, identification of major source areas, for example, inside the Sahara, remains difficult (Goudie and Middleton, 2001). This is connected with the rather erratic aeolian transport of particles, which depends on the changing situation of the ground surface and the occurrence of sufficient wind speeds. In addition, local dust sources may include re-mobilized older dust deposits, which can partly consist of remote material, or which might be (pre-)weathered to a certain extent before re-mobilization and subsequent deposition.

Grains are transported according to their size and weight: sand moves a few hundred meters, but silt is carried high into the atmosphere. Clay-sized particles remain in suspension in the atmosphere until washout takes place. Most of the dust that is transported over long distances consists of these clay-sized particles, which are permanently present in the high atmosphere over deserts. Sometimes 'giant' sand-sized particles have been transported over long distances of thousands of kilometers, but the reasons for this phenomenon are so far unknown.

To overcome obstacles in the identification of dust sources, some studies have focused on suspended dust in the atmosphere, assuming that it comprises mainly material transported from distant sources. But samples collected in Israel had the same mineralogical and chemical composition of dust particles from several synoptic conditions that carry desert dust. The variable aspects were somewhat larger amounts of palygorskite and Mg in dust that originated from the Arabian deserts (Kalderon-Asael et al., 2009).

Local, regional, and remote sources play a role in the composition of the silt and clay fractions (i.e., of material $<63 \mu\text{m}$) of aeolian sediments. Mixing of particles occurs with distance, and the variability of single dust storms obscures the contribution of individual sources. Aeolian transport homogenizes the material: the longer sediments remain in suspension in the atmosphere, the more difficult it becomes to discriminate sources. Sampling is also a challenge because the type of dustfall plays a role in dust composition. Samples collected during wet dustfall (rain and, even more pronounced, snow) in the southern Levant showed enrichment by calcareous silt, and also enhanced presence of most major and trace elements compared with samples from dry dustfall (Lucke et al., 2019). It seems possible that this is connected with coatings of suspended clay-sized particles on calcareous silt particles (Mahowald et al., 2014), which essentially means that larger particles may 'harvest' suspended dust during wet dustfall events.

Aeolian sediments and soil formation in semi-arid regions

Chemical weathering and thus in-situ pedogenesis might alter aeolian sediments after deposition in semi-arid areas, which adds to the complexity of identifying the contributions of parent material from remote sources. In this context, many limestones under (Mediterranean) semi-arid climate are covered by clay-rich soils of reddish or reddish-brown colors (Chromic Cambisols according to WRB (2015), or Chromoxererts or Xerochrepts according to Soil Taxonomy (USDA, 2014), but better known as Terra rossa or Terrae calcis according to older classification systems). The origin and genesis of these soils is debated: the two main hypotheses are that they are mainly derived from insoluble bedrock residue (i.e., clays trapped by the growing reefs that later formed the

limestones), or that they mainly represent long-range dust transport. The main argument against the bedrock dissolution theory is that the limestones contain very little insoluble residue, requiring vast amounts of limestone to be dissolved over long periods. In this context, available ages from the southern Levant suggest formation mainly during the Late Pleistocene (Crouvi et al., 2018), which is relatively young and contradicts an origin from bedrock dissolution. Available geochemical data supports the notion that a remote dust fraction must have contributed to soil development, but a contribution from local rocks was also evident. Exogenous grains of medium sand size were found in hilltop soils. Nearby rock outcrops on valley slopes could be identified as sources of these grains which suggests that this contribution from local rocks might have occurred largely in the form of aeolian sedimentation. As significant energy is necessary for the aeolian uplift of such large grains, their deposition was possibly associated with major storm events. This is supported by reconstructions of transport and deposition histories from grain microsurfaces which indicate an important role of intense dust storms involved in soil formation (Lucke, 2017).

Soil colors

In arid regions with limited chemical weathering, soil colors largely reflect parent materials. These are mainly yellow and brown and dominated by iron oxides. They are inherited from parent materials, which may however include (aeolian) sediments from pre-weathered older soils.

In semi-arid regions, however, in-situ weathering might take place and apparently leads mainly to formation of hematite, which gives soils a reddish color. The preferential oxidation of iron to hematite seems to be connected with rapid dehydration during summer drought. Reddish colors were sometimes found associated with enhanced magnetic susceptibilities which could arise from the formation of maghemite as a precursor of hematite (Lucke, 2017). A transect covering semi-arid to arid soils in northern Jordan showed that soil surface colors corresponded to the current hygric gradient, suggesting that the intensity of redness in this area was connected with in-situ oxidation of iron-bearing minerals (Sahwan et al., 2021). Organic matter contents, in contrast, played no significant role for soil colors.

The limited presence of vegetation and organisms, which is connected with the lack of available water, means that organic matter contents of soils in arid and semi-arid regions are small (usually <1%). Soil colors in drylands are therefore in general not related to vegetation and organic matter. This is also true for the border between Israel and Egypt, which can be identified on satellite imagery from the darker soil color in Israel, even though the geology and soils do not differ on the two sides of the border. Although measures to protect the vegetation from grazing on the Israeli side has led to a substantial recovery of shrub vegetation, the main reason for the difference in reflectance spectra seems to be the widespread, undisturbed presence of biological crusts on soil surfaces in the Israeli Negev (Tsoar et al., 1995).

Soil fertility, land use, and desertification

Soil fertility

Base saturation in most arid and many semi-arid soils can reach 100% because of limited leaching. This means that in contrast to humid soils, lack of cations is usually not a growth-limiting factor for plants. In fact, it is more likely that excess of cations (i.e., salinity) is a problem. The most frequently deficient ion is N, which is connected to the small organic matter contents. Supply of most nutrients is good provided that enough water is available for solution and transport to plants. Water availability is a limiting factor for the application of manure, as an organic fertilizer, and to an even larger degree inorganic fertilizers because they need water to become available for plants.

Physical soil properties are more important for soil fertility in drylands than chemical ones. This applies first and foremost to the soils' capacity to store plant-available water, which is why soil depth and particle-size distribution are of paramount importance. Due to small organic matter contents, cation exchange capacity is largely determined by grain sizes and depends mainly on clay minerals. This implies that many indices for assessment of soil quality or productivity that were tested and developed in temperate regions are not well-suited for arid and semi-arid regions. There, fertility assessments need to focus on physical soil properties (Husein et al., 2021). This also means that mass movements leading to erosion of topsoil are less relevant for soil degradation in drylands than in humid areas.

Land use

Semi-arid regions are the natural habitats of many plants that were domesticated for agriculture, such as cereals (*Triticum* spp.) and olives (*Olea* spp.). Domestication first took place in the Middle East during the Neolithic period some 10,000 years ago (Neolithic revolution). It was connected with the development of irrigation technologies, which also made the production of surpluses possible and led to the emergence of first civilizations. The steppes enabled rainfed agriculture, but irrigation improved yields and made expansion into arid areas possible. Rain-fed agriculture also expanded into Mediterranean forests, but probably at a later stage when fire management and metal tools were available that permitted the cutting of the often very resistant Mediterranean woods.

The Neolithic revolution induced not only domestication of plants, but also of animals which led to the emergence of pastoralism. The borders between farmers and herdsmen were probably fluid as most tribes would practice a mixed economy. Nomads who rely fully on their animals, which enables them to penetrate deeply into waterless arid areas, are a later specialized



Fig. 6 Irrigated palm groves and gardens in an oasis in south Morocco. Availability of groundwater near the surface made long-term, traditional agricultural land use possible which applies sustainable recycling of organic matter. The limiting factor of productivity is the availability of water.

development. Farming and grazing exploit different ecological niches and potentially benefit each other. In the past, the emergence and fall of civilizations has been interpreted as the 'battle between the desert and the sown,' i.e., a fight between civilized farmers and uncivilized nomads. However, it became increasingly clear that this reflects a foreign view because of lack of familiarity with the real social and economic situation in drylands. For example, archeological evidence suggests that part of the ancient civilizations' wealth was based on surpluses from pastoral land use, i.e., that some of the ancient palaces had been owned by early Bedouins. Pastoralism and agriculture were often practiced in parallel by members of the same tribal group (Lucke, 2017).

The main challenge of land use is the unpredictable water supply. For agriculture, it could be addressed by diverting water from springs and streams for irrigation (Fig. 6). In addition, rainwater could be collected in cisterns or by runoff diversion, an important technology. When sufficient water supply is secured, preventing salinization of soils is a major problem of irrigated agriculture. It is addressed by using sufficient water of good quality (i.e., not brackish or saline), and by draining excess water, which flushes out salts. Grazing of animal herds enables advantage to be taken of the spatial variation in water supply because of mobility of the animals which can move to areas with a sufficient water supply. This often follows seasonal patterns: for example, tribes in northern Jordan move their flocks eastwards to arid lands for winter pasture, and in summer to the better-watered Mediterranean areas in the west. After harvest, the grazing of stubble was a common practice that also fertilized soils. If droughts led to insufficient yields, the animals were fed on the poorly developed plants on the fields, thus still making use of them.

Agriculture even expanded to arid areas without permanent water resources, such as the Negev desert, where very fertile desert fringe loess could be cultivated. Water supply relied on runoff from the few rain events during winter time, which was diverted to terraced fields. The key to successful runoff diversion was not high average annual rainfall, but the occurrence of a few major storms that provided enough rain for the development of significant runoff. The diverted water was then stored in soils behind the terrace walls. However, only relatively small areas can be cultivated in this way, as fields need runoff collection areas about 20 times larger than the cultivated area.

These systems harvested not only water, but also aeolian sediments, and could therefore be applied even in rather hostile environments such as the arid sandstone mountains of Petra in southern Jordan (Fig. 7). There, it created fertile soils on previously barren rocks (Lucke et al., 2019). Furthermore, it reduced the risk of flash flood damage as water was diverted on a large scale at the head of the catchment, controlling flows before dangerous runoff concentrations could develop. Modern experiments in the Negev desert show that such systems could still be operated today (Evenari et al., 1971). They might be a way forward to long-term sustainable agriculture in drylands as runoff-irrigated farming can be extremely productive in years with good rainfall. The reasons for such record yields are so far unknown. However, the rather frequent occurrence of droughts and subsequent bad harvests or crop failures lead to overall mean yields comparable to rain-fed agriculture. The main disadvantage of runoff-irrigated terrace agriculture is that they require an enormous input of manual work: terraces and runoff diversion systems need to be repaired every year, while small fields do not permit the use of heavy machinery. Therefore this land use strategy currently cannot compete with world market prices of agricultural goods.

Modern technology has benefitted land use in drylands mainly by providing additional and advanced means of water provision. In contrast, advances in the field of mechanization or manure have played only minor roles due to limitations posed by the often dissected terrain, small field sizes, and lack of available water. The advances in water provision rely to a large degree on the pumping of fossil groundwater from deep geological reservoirs, which has made Saudi-Arabia, for example, one of the main wheat exporting states in the world. However, these water reserves will probably soon be depleted. The excessive use of water resources during the



Fig. 7 Remains of runoff-irrigated terraces near Petra, southern Jordan. Car gives scale. Note the small cereal plants and evidence of plowing testifying to current re-use of these installations by Bedouins inhabiting the area.

recent past has led to the deterioration of water quality, and only 'gray water' is currently available for agriculture in many arid and semi-arid regions. This is wastewater after treatment in sewage plants, which may still, however, contain harmful substances such as hormones, microplastics, or 'heavy' metals. Future land management strategies will have to develop concepts to deal with the challenges posed by these substances. If a way could be found to remove or inactivate them, this might allow the use of sewage sludge as manure, which could become a precious source of organic matter, N, and P for dryland soils.

It seems probable that non-conventional or counter-intuitive concepts will be of growing importance in the future. For example, intensive fish breeding in tanks proved very successful in the deserts of Israel, as these animals need limited water and even thrive in saline groundwater, while solar energy can be used to operate the tanks. In this context, the harvesting of renewable energy from wind and solar radiation is another promising field of future land use in drylands, provided that problems such as cleaning solar collectors from dust can be solved without excessive use of the limited water resources.

Desertification

Both variation in climate and poorly adapted land use can lead to vegetation damage and soil degradation in drylands, with the potential consequences of long-term reduced vegetation growth and diminished soil productivity. Negative impacts of climate change are mainly associated with prolonged droughts, but also excess rain that concentrates during runoff events can have adverse consequences, causing gullying for example. Poorly adapted land use is first and foremost a problem of modern practices, in particular the intense use of bulldozers, concrete construction, and the exploitation of non-renewable (fossil) water resources. When bulldozers became available, they were used to create false terraces, often at the landscape scale, or to conduct other major modifications of the natural relief. While these might be beneficial to farmers in the short term, they threaten landscape stability in the long term, aggravating the effect of flashfloods, for example (Grove and Rackham, 2003). Traditional land use such as plowing and grazing may also have negative consequences, in particular deforestation and soil erosion. With limited vegetation cover and often weakly developed soil structure, damage to plants (in particular roots) can rapidly make soils susceptible to water erosion. Plowing destroys crusts and surface cover which can trigger wind erosion.

Traditional land use in arid and semi-arid regions has been associated with Bedouins and other poorly educated inhabitants of drylands, who were sometimes considered to be the causes of desertification. However, the environmental impacts of traditional land use are smaller than modern practices, simply because it relies mostly on human and animal power and lacks modern machinery. There are also historic examples of well-managed land use strategies in drylands, such as the construction of runoff-irrigated terraces, which suggest caution with general negative assessments of traditional land use practice. The association of desertification with 'poorly developed' traditional societies could at least partly result from cultural bias (Lucke, 2017).

For more than 50 years, desertification is a topic of intense scientific debate, both with regard to defining and assessing the problem, and with regard to possible mitigation or solution strategies (Sterk and Stoorvogel, 2020). Research faces various methodological challenges, for example, determining effective numerical estimates of actual soil erosion (Grove and Rackham, 2003). Installations to record soil losses might not show any significant movements of soil for years until an extreme rainfall event hits the area. Then, again, meaningful measurements might not be possible, for example, if the remains of the installation have to be searched for and are recovered from meters of debris at the lower end of the catchment. Indirect measurements of mass movements,



Fig. 8 Camel herd grazing an *Acacia* spp. tree in the Atlas mountains of south Morocco. These animals can penetrate deeply into arid areas and feed on various dryland plants.

such as from dating sediments in depressions, can also be problematic because these might have been deposited (or re-deposited) mainly discontinuously during rare but intense rainfall events (Lucke, 2017).

Grazing, especially with flocks of goats, has often been assessed to have a general negative impact on vegetation, in particular on woody plants. It was assumed that browsing would rapidly cause overgrazing and irreversible degradation of plants and soils (Fig. 8). Animals and their herders were therefore banned from various semi-arid areas which were re-forested. However, this has sometimes made the situation worse: reforestation by tree monocultures has caused lowering of the groundwater table in some regions, while the lack of browsing of brushwood has led to an enormous increase in fire damage. Ecologically, the removal of grazing animals from forested areas has decreased biodiversity and moved the grazing pressure to more sensitive arid areas. Modern drilling of wells or moving water and herds by trucks has converted previous reserve areas (perhaps extensively grazed by camels) to intensively exploited lands. This replacement of traditional land management with its system of reserve lands by industrial land use, often been made possible only by the intense use of fossil water, has therefore sometimes caused more harm than good.

Alternative views of grazing are that in most semi-arid and arid regions, the vegetation and wild herbivores evolved in parallel and are thus adapted and interdependent (Savory and Butterfield, 2016). The impact of grazing on vegetation and soils is often quickly reversed, as many dryland plants have strong regeneration capacity and browsing does not make a significant difference to the already low organic matter contents of soils. In areas with a long evolutionary history of grazing and low net primary production, moderate grazing pressure during short time periods can result in increased vegetation growth. This is associated with some of the positive effects of limited disturbances by the animals who may act as agents of seed dispersal while their excrements supply N and P to soils. Therefore, it seems possible that historic grazing strategies mirrored to some degree the effect of browsing by wild herds, and had net positive effects for the ecosystem. This is, for example, supported by an expansion of forest in the Mediterranean areas of Jordan during periods of Bedouin domination (Lucke, 2017).

To summarize, desertification today is perceived less as a global catastrophic development, but a more local-scale degradation problem and not necessarily an irreversible process. Land use plays a crucial role as it can either stimulate or decrease degradation processes, which emphasizes the importance of long-term sustainable land use practice.

Summary and conclusion

The limited availability of water in arid and semi-arid regions largely determines soil development. Climate is the most important soil-forming factor in these areas because other factors, such as the development of vegetation and organic matter, or the development of processes linked to topography, such as runoff formation, are directly or indirectly connected with available water. However, the water balance depends not only on precipitation, but also on potential evapotranspiration (PET). The latter is linked with various other climatic properties such as solar radiation intensity, day length, or wind speed. The type of precipitation also matters, for example, whether it occurs mainly during summer or winter, and whether it is distributed in many small or few intense rainfall events. In addition, local factors such as relief, soil depth, or particle-size distribution can cause small-scale variation in moisture availability. Climate classifications are therefore useful for a first evaluation of probable soil formation on large spatial scales, but cannot be used to predict local soil development.

Moisture in arid and semiarid soils enters and leaves mainly at the soil surface. The movement of water and solutes is therefore less directed to the bottom of soil profiles, but merely toward the soil surface. Leaching is often incomplete which leads to the precipitation of secondary minerals and formation of respective soil horizons, such as secondary carbonates in calcic horizons. Due to limited chemical weathering, soils in arid and semi-arid regions are mainly characterized by soil types with only initial or weak development.

Soil formation is governed by physical processes, which can, however, be intense. Thermal fluctuation and crystallization of salts can produce significant amounts of detritus which are moved by gravity, water, and wind, and contribute to soil development. In particular, the deposition of aeolian sediments plays a major role for soil development in arid and semi-arid areas. They are both sources and sinks of dust. Although major dust trajectories can be observed on satellite images and are well-researched, it is difficult to determine specific dust sources because local, regional, and remote sources mix in the atmosphere. Local sources may also include re-mobilized, earlier aeolian sediments which may have originated from remote areas.

Because of the limited water resources, semi-arid and arid lands are largely not yet used by man and may therefore represent important reserve areas for the growing human population. Water management remains the key aspect for any successful land use in these regions, and it seems likely that non-conventional strategies such as the breeding of fish in tanks or the harvesting of renewable energies will offer future opportunities to support human life in these regions.

Future research should focus less on single variables, such as the provenance of the parent materials of dust, but more on the complexity of geomorphological processes that shape soils and landscapes in drylands. For example, dust fixation mechanisms have rarely been considered but may play an important role in the composition of soil substrates. Furthermore, dust transport in the atmosphere probably interacts with synoptic conditions and the presence or absence of precipitation, which, however, has so far remained difficult to determine. Last but not least, historic experiences of millenia of successful human land use in drylands—despite harsh and strongly changing environmental conditions, which sometimes appear hostile to life—may provide important examples for the development of future sustainable land use strategies in drylands.

References

- Bruins HJ (2012) Ancient desert agriculture in the Negev and climate-zone boundary changes during average, wet and drought years. *Journal of Arid Environments* 86: 28–42.
- Crouvi O, Barzilai O, Goldsmith Y, Amit R, Matskevich Z, Porat N, and Enzel Y (2018) Middle to late Pleistocene shift in eolian silts contribution into Mediterranean soils at the fringe of the Negev loess, Israel. *Quaternary Science Reviews* 191: 101–117.
- Danin A and Ganor E (1991) Trapping of airborne dust by mosses in the Negev Desert, Israel. *Earth Surface Processes and Landforms* 16: 153–162.
- Evenari M, Shanan L, and Tadmor N (1971) *The Negev. The Challenge of a Desert*. Cambridge, MA: Harvard University.
- Goudie AS and Middleton NJ (2001) Saharan dust storms: Nature and consequences. *Earth-Science Reviews* 56(1–4): 179–204.
- Grove AT and Rackham O (2003) *The Nature of Mediterranean Europe: An Ecological History*. New Haven: Yale University Press.
- Husein HH, Lucke B, Bäumler R, and Sahwan W (2021) A contribution to soil fertility assessment for arid and semi-arid lands. *Soil Systems* 5(3): 42.
- Jenny H (1941) *Factors of Soil Formation: A System of Quantitative Pedology*. New York: McGraw-Hill. Re-published in 1994. New York: Dover publications.
- Kalderon-Asael B, Erel Y, Sandler A, and Dayan U (2009) Mineralogical and chemical characterization of suspended atmospheric particles over the east Mediterranean based on synoptic-scale circulation patterns. *Atmospheric Environment* 43(25): 3963–3970.
- Lucke B (2017) *Landscape Transformations in the Context of Soil Development, Land Use, and Climate: A Comparison of Marginal Areas in Jordan, Mexico, and Germany*. Stuttgart: Schweitzerbart Science Publisher.
- Lucke B and Bäumler R (2021) Holocene dust dynamics archives in archaeological ruins in arid and semi-arid environments in the Southern Levant. *Arabian Journal of Geosciences* 14(23): 1–15.
- Lucke B, Sandler A, Vanselow KA, Bruins HJ, Abu-Jaber N, Bäumler R, Porat N, and Kouki P (2019) Composition of modern dust and holocene aeolian sediments in archaeological structures of the Southern Levant. *Atmosphere* 10(12): 762.
- Mahowald N, Albani S, Kok JF, Engelstaeder S, Scanza R, Ward DS, and Flanner MG (2014) The size distribution of desert dust aerosols and its impact on the Earth system. *Aeolian Research* 15: 53–71.
- Monger HC, Martinez-Rios JJ, and Khresat SA (2005) Tropical soils. Arid and semiarid. In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*, 1st edn., pp. 182–187. Amsterdam: Elsevier.
- Muhs DR (2013) The geologic records of dust in the Quaternary. *Aeolian Research* 9: 3–48.
- Peel MC, Finlayson BL, and McMahon TA (2007) Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences* 11: 1633–1644.
- Roskin J, Porat N, Tsoar H, Blumberg DG, and Zander AM (2011) Age, origin and climatic controls on vegetated linear dunes in the northwestern Negev Desert (Israel). *Quaternary Science Reviews* 30(13–14): 1649–1674.
- Sahwan W, Lucke B, Sprafke T, Vanselow KA, and Bäumler R (2021) Relationships between spectral features, iron oxides and colours of surface soils in northern Jordan. *European Journal of Soil Science* 72(1): 80–97.
- Savory A and Butterfield J (2016) *Holistic Management. A Commonsense Revolution to Restore Our Environment*, 3rd edn. Washington D.C.: Island Press.
- Smalley I and Vita-Finzi C (1968) The formation of fine particles in sandy deserts and the nature of 'desert' loess. *Journal of Sedimentary Petrology* 38(3): 766–774.
- Sterk G and Stoorvogel JJ (2020) Desertification—scientific versus political realities. *Land* 9(5): 156.
- Tsoar H, Goldsmith V, Schoenhaus S, Clarke K, and Karnieli A (1995) Reversed desertification on sand dunes along the Sinai/Negev border. In: Tchakerian VP (ed.) *Desert Aeolian Processes*, pp. 251–268. Dordrecht: Springer.
- Turk JK (2012) *Vesicular Horizon Distribution, Properties, and Pedogenic Processes in Deserts of the Western United States*. Ph.D. Dissertation Riverside, CA, USA: UC Riverside. Available online <https://escholarship.org/uc/item/325854wj>, accessed on 26 September 2022.
- UNESCO (1979) *Map of the World Distribution of Arid Regions. Man and the Biosphere (MAB) Technical Notes 7*. Paris: UNESCO.
- USDA (2014) Soil survey staff. In: *Keys to Soil Taxonomy*, 12th edn. Washington D.C.: United States Department of Agriculture (USDA) & Natural Resources Conservation Service (NRCS).
- WRB (2015) *World Reference Base for Soil Resources 2014. Update 2015. World Soil Resources Reports 106*. Rome: FAO.
- Wright J, Smith B, and Whalley B (1998) Mechanisms of loess-sized quartz silt production and their relative effectiveness: Laboratory simulations. *Geomorphology* 23(1): 15–34.

Relevant websites

<https://escholarship.org/uc/item/325854wj>.

Humid subtropical ecozone with emphasis on Acrisols and Alisols

Nilton Curi^a, Sérgio Henrique Godinho Silva^a, and Alberto Vasconcellos Inda^b, ^aDepartment of Soil Science, Federal University of Lavras (UFLA), Lavras, Minas Gerais, Brazil; ^bDepartment of Agronomy, Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul, Brazil

© 2023 Elsevier Ltd. All rights reserved.

Introduction	343
Brief characterization of the humid subtropical ecozone	343
Factors of formation of Acrisols and Alisols in humid subtropical conditions	344
Processes of formation of Acrisols and Alisols	346
Properties of Acrisols and Alisols	349
Sustainable use and management of Acrisols and Alisols	351
Conclusion	352
References	352

Abstract

Acrisols and Alisols are found in different parts of the world and under different climatic conditions. Both soils have an argic horizon in the subsurface, although their physical, chemical and mineralogical properties are contrasting. Within the humid subtropical ecozone, these soils tend to occur mostly in hilly landscapes. This chapter focuses on characterizing Acrisols and Alisols of the humid subtropical ecozone, providing details on their properties, sites of occurrence and sustainable management systems under different degrees of management.

Key points

- Genesis of Acrisols and Alisols and their regions of occurrence within the humid subtropical ecozone are discussed;
- Differentiation of Acrisols and Alisols regarding their physical, chemical, and mineralogical properties is provided;
- Sustainable management practices are suggested for these soils to avoid soil degradation in the humid subtropical ecozone.

Introduction

Soils vary across the globe and this is mainly controlled by the five soil forming factors (climate, organisms, relief, parent material and time) and several pedogenic processes. For a better interpretation about each soil class and its importance in the environment, evaluation of soils within certain climatic conditions is necessary.

Considering all the soil types of the world, there are two somewhat similar types in the humid subtropical ecozone: Acrisols and Alisols. Both soil classes have an argic horizon in the subsurface, but the former is more weathered and leached than the latter. These soils are widely used for agriculture, although care should be taken with their management since they are susceptible to both water and wind erosion. The undulating or steeper relief on which they are commonly found, in addition to the high mean annual precipitation of the humid subtropical ecozone, make them especially prone to water erosion.

This chapter focuses on characterizing Acrisols and Alisols of the humid subtropical ecozone, providing detail on their properties, regions of occurrence and sustainable management systems under different types of management.

Brief characterization of the humid subtropical ecozone

The world has multiple combinations of climate, geology, native vegetation, altitude and relief that govern several aspects related to the environmental sciences. Each of these combinations, for example a flat area under a humid climate compared with a flat area under an arid climate, is capable of creating different conditions that control the establishment and evolution of site-specific fauna, flora, soils, and others.

For a better characterization of these different environments and their individualization regarding their ecological aspects, the so-called ecozones, also named ecoregions, have been created and defined according to the environmental characteristics intrinsic to each ecozone. Ecozones were defined by Omernik (2004) as areas that present coincident geographical phenomena, i.e., climate, vegetation, geology, soils, fauna, hydrology, physiography, impacts of human activities. According to the FAO (2012), the world is divided into 20 different ecozones, varying from tropical deserts (e.g., Sahara desert in Africa), to tropical forest (e.g., Amazonian

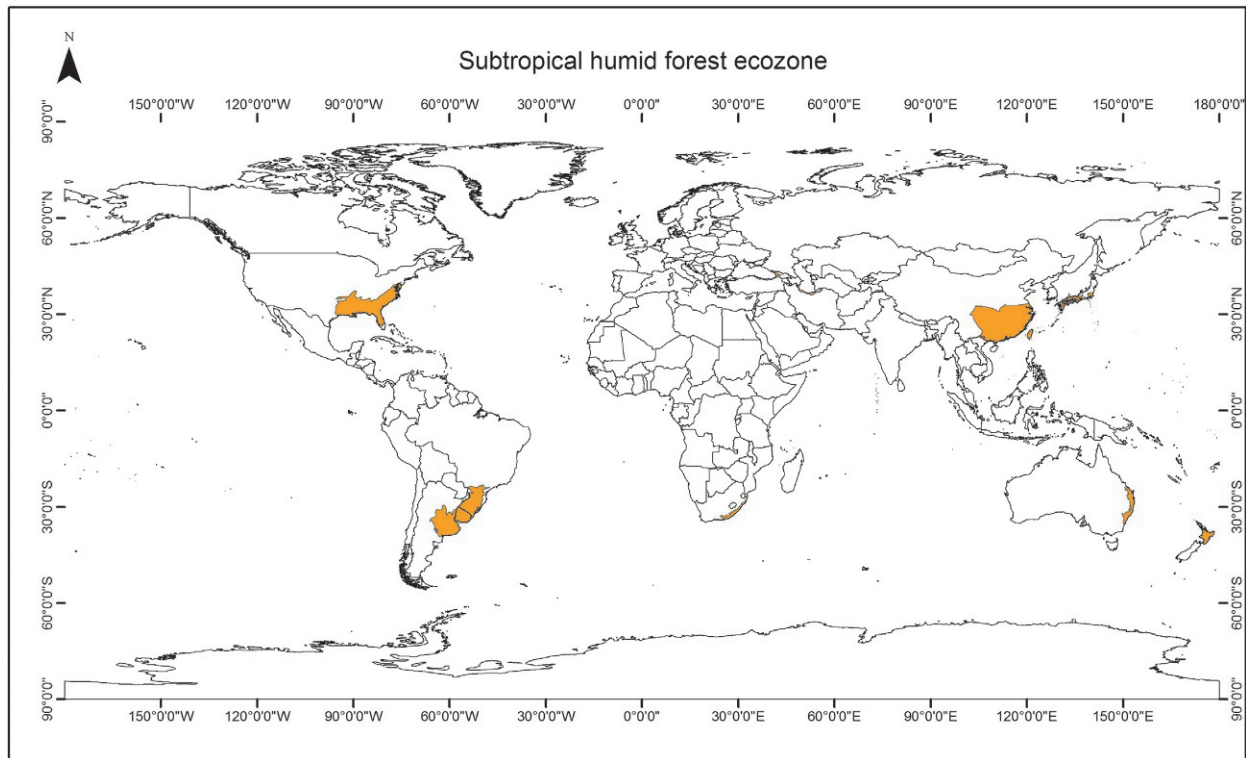


Fig. 1 Worldwide distribution of the humid subtropical ecozone, according to the FAO (2012).

forest in South America), temperate continental forest (e.g., central European region), and Boreal coniferous forest (e.g., parts of the northern Europe and North America continents) to name a few.

According to the Food and Agriculture Organization (FAO, 2012), the subtropical domain encompasses areas that have at least 8 months of the year with mean temperatures above 10 °C and situated between 25° and 40° of latitude. For both northern and southern hemispheres, this domain generally occupies the southeastern parts of the continents. The subtropical region is divided into five ecozones: subtropical desert, subtropical steppe, subtropical mountain system, subtropical dry forest and subtropical humid forest. In other systems of climate classification, the humid subtropical zone is classified as a type of warm-temperate climate. This chapter focuses on the humid subtropical forest zone, code SCf, which is included within humid subtropical climatic criteria, and characterized by well distributed rainfall during the whole year with all months being humid. Specifically, its mean annual precipitation is commonly greater than 1000 mm.

The native vegetation of the humid subtropical ecozone is characterized by evergreen broadleaved forests, evergreen coniferous forests, and winter-deciduous forests. Examples of vegetation types in the humid subtropical ecozone include *Eucalyptus*--*Nothofagus* (*Eucalyptus globulus* Labill. – *Nothofagus* Blume.) forests in southeastern Australia and Tasmania, *Castanopsis* (*Castanopsis* (D. Don) Spach, 1841) forests in southeastern China, *Araucaria* (*Araucaria* Juss.) forests in southern Brazil, and Longleaf-Slash pine (*Pinus elliottii* Engelm.) forests in southeastern USA. In mountainous areas of the ecozone, rainfall increases and temperature decreases with altitude. This combination of low temperature and a large amount of rainfall favors the establishment of these native forest types compared with the areas surrounding the places where such forests occur.

This ecozone is mainly distributed in southeastern USA, southern Brazil, southeastern tip of Africa, southeastern Australia, and southeastern China (Fig. 1).

Factors of formation of Acrisols and Alisols in humid subtropical conditions

The paradigm of the factors of soil formation furnishes a conceptual structure for the understanding of soil formation. One should recall that soils are formed due to the interaction of climate (cl), organisms (o), relief (r), parent material (p) and time (t), in addition to some unspecified subsidiary factors (...) that control soil formation (Eq. 1). Climate and organisms act on the parent material (material from which soils are formed, such as bedrock or mineral or organic sediments), under the influence of relief over time. Thus, the different combinations of these so-called factors of soil formation lead to the occurrence of different soil classes and with different properties across the globe.

$$S = f(\text{cl}, \text{o}, \text{r}, \text{p}, \text{t}, \dots). \quad (1)$$

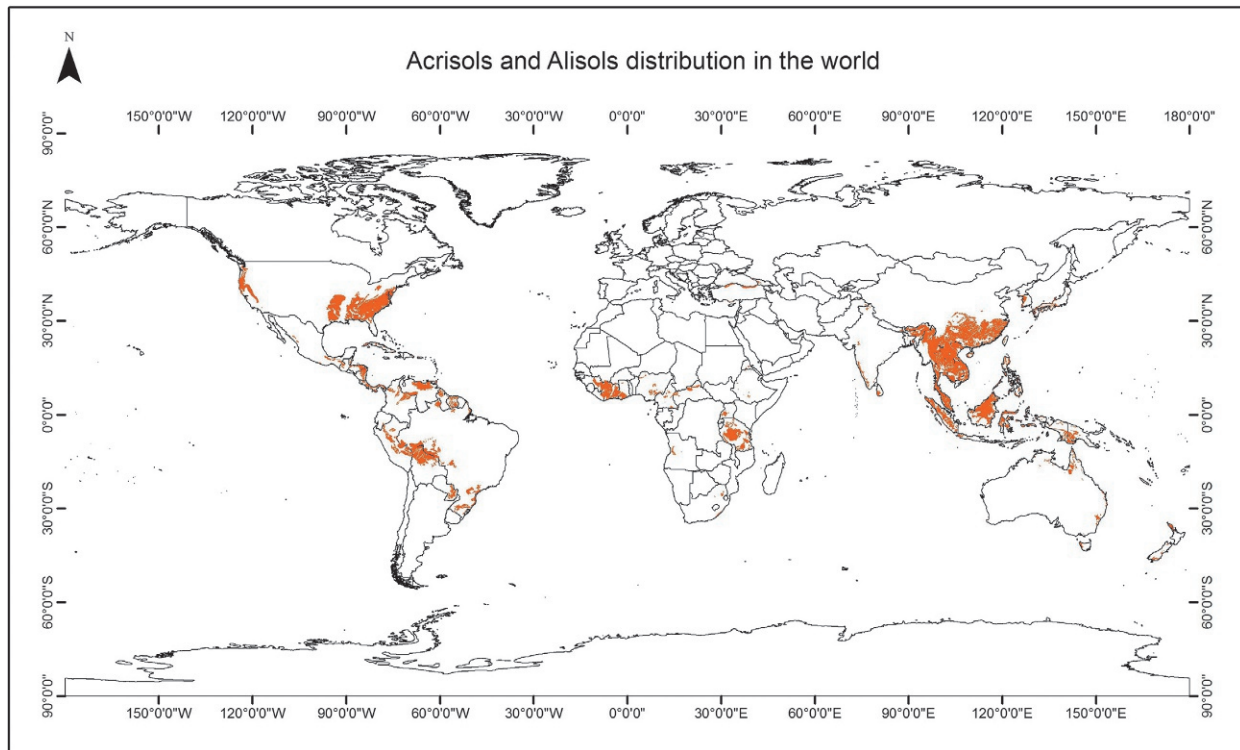


Fig. 2 Occurrence of Acrisols and Alisols across the world. Adapted from FAO-UNESCO (1974) *Soil map of the world 1:5000000* Vol I. Legend, I. Rome: FAO-UNESCO.

The humid subtropical conditions mainly refer to the climate factor, but the other factors of soil formation can also vary under the similar climatic condition. For instance, by maintaining the other factors constant, variation in parent material may be the factor that determines the formation of different soil classes and properties. Different parent materials may result in contrasting soils in relation to their classification, physicochemical properties and sustainable management practices. This fact helps to explain the widespread occurrence of different soil types under the same climatic conditions. However, among the soils commonly found in areas under humid subtropical conditions, two soil classes can be highlighted: Acrisols and Alisols (The World Reference Base according to the FAO/WRB soil classification system, [IUSS Working Group, 2015](#)). These soils are equivalent to Ultisols with low-activity clays and Ultisols with high-activity clays, respectively, in the USDA Soil Taxonomy ([Soil Survey Staff, 2014](#)).

In humid subtropical regions, Acrisols are most extensive in southeast USA, the southeast tip of Africa and south Brazil ([IUSS Working Group, 2015](#)), whereas Alisols occur mostly in southeast USA and south Brazil ([FAO-UNESCO, 1974](#); [IUSS Working Group, 2015](#); [Streck et al., 2018](#)) (Fig. 2). In Brazil, these soils are found in the Paraná, Santa Catarina and Rio Grande do Sul states ([Resende et al., 2021](#)).

The climate and altitude factors under which both Acrisols and Alisols form tend in general to result in a udic soil moisture regime and a thermic soil temperature regime according to the US Soil Taxonomy ([Soil Survey Staff, 2014](#)). The udic soil moisture regime applies to soils in which the soil is not dry in the active section of the soil (2 m in these soils) for as long as 90 cumulative days in regular years. For such conditions, rainfall either has to be well distributed throughout the year (humid subtropical conditions) or there is enough rainfall in the summer so that stored soil moisture plus rainfall equal or exceed the evapotranspiration. The thermic soil temperature refers to soils in which the mean annual soil temperature is 15 °C or higher, but lower than 22 °C, and the difference between mean summer and mean winter soil temperatures is 6 °C or more.

In humid subtropical conditions with well-distributed rainfall there is intense leaching (removal of bases, Ca^{2+} (calcium), Mg^{2+} (magnesium), K^+ (potassium), and Na^+ (sodium), and silica as water percolates downwards through the soil profile), which is accompanied by a change in the mineralogy of clay-sized particles, mainly from smectite to kaolinite. Smectite, a 2:1 phyllosilicate mineral, occurs in greater quantity in less weathered-leached soils, which is not the case for Acrisols and Alisols. Along with weathering and leaching processes, smectite tends to break down with changes in soil chemistry, mainly low pH, which releases the elements that comprise part of its crystalline structure. Under these intense leaching conditions, typical of humid subtropical environments, most such bases (Ca^{2+} , Mg^{2+} , K^+ , and Na^+) and silica are removed from soil ecosystems, causing a change in the former conditions that favored the formation of smectite. Thus, other minerals such as kaolinite tend to form ([Kämpf et al., 2012](#)). Moreover, the formation of pedogenetic iron (but not aluminum) oxide minerals is favored under these new soil conditions. Of the main iron oxide minerals in these soils, goethite formation is favored over that of hematite under these climatic conditions, i.e., lower temperatures and greater water content. Humid subtropical conditions also favor the accumulation of larger amounts of

organic compounds, which may complex with iron (Fe) and reduce its availability for the formation of hematite (Schwertmann and Taylor, 1989). In addition, the formation of gibbsite, an aluminum hydroxide also found in very weathered–leached soils, is not favored because in most cases the silica in solution prevents its formation in Acrisols and Alisols (Kämpf et al., 2012).

The humid subtropical environmental conditions under which Acrisols and Alisols form cause mineralogical transformations from easily-weatherable primary minerals (EWPM), e.g., feldspars and micas, to kaolinite mainly in Acrisols (more weathered–leached than Alisols). With these mineralogical transformations, other soil properties are consequently influenced, causing low natural soil fertility (Acrisols and Alisols) and low cation exchange capacity (CEC) (Acrisols). For example, the CEC of smectite tends to be at least 10 times greater than that of kaolinite (Resende et al., 2021), which markedly reduces the negative electrical charges of Acrisols and the capability of their clay particles to adsorb bases and other cations.

In the case of Alisols, the conspicuous presence of hydroxy-interlayered vermiculite (HIV) in the clay fraction is responsible for high CEC values, but these soils still have low natural fertility. This fact indicates that the soil uses and management practices considered for Alisols need to be different from those for Acrisols.

For the ‘organisms’ factor, one may consider both vegetation and fauna present in soils. Together with other factors, they are responsible for the transformation of parent rocks into sediments and, hence, soils, favoring the development of an A horizon. Moreover, while microorganisms are extremely important for the cycling of nutrients by promoting the decomposition of organic residues, vegetation provides organic matter to the soil surface, creates preferential water pathways as the plant roots are decomposed, increases the content of organic matter at depth, and contributes to nutrient cycling by the uptake of nutrients by roots at depth and their eventual incorporation into the plant tissues. Later, these nutrients will be detached from the plant tissues, incorporated into topsoil, and then leached or eroded or taken up again by vegetation. In addition to being a factor of soil formation, vegetation is also influenced by soil properties (together with the climate) that control the type of climax vegetation to be established on it. In areas where Acrisols and Alisols occur under humid subtropical conditions, native forests are common on the various continents. This, together with the lower temperatures of humid subtropical ecozones, favor the accumulation of soil organic matter and its compounds.

Relief may control soil drainage (steep areas tend to have lower water infiltration than flat areas), may enhance erosion, especially on steep slopes with soils deprived of or with little vegetation cover, and resulting in redeposition of sediments, soil organic carbon, nutrients and other elements at lower landscape positions, including floodplains. Therefore, these consequences of relief may increase environmental risks, such as soil degradation through the removal of topsoil and deposition of sediments in water bodies. The predominant relief where Acrisols and Alisols tend to occur is dominantly hilly or undulating, which affects soil depth: moderately deep to deep soils (*solum* thickness ranging from 50 cm to 2 m) but not very deep soils >2 m, which is common with Ferralsols or Oxisols.

Acrisols occur on a great variety of parent materials, mainly from the weathering of acid parent materials such as granite and rhyolite. These rocks have high contents of silica (SiO₂) and are dominated by quartz and feldspars in addition to accessory minerals, e.g., micas and amphiboles. Although Alisols may also develop in different parent materials, they are mainly formed from the weathering of mafic rocks such as basalt, gabbro and diabase, and from unconsolidated sediments. Mafic rocks are Fe- and Mg-rich, mainly composed of Ca- and Na-bearing feldspars (labradorite) in addition to pyroxenes (augite), amphiboles, olivine, and magnetite. Differences in the main parent materials of Acrisols and Alisols tend to promote greater clay contents in Alisols because of the reduced contents of quartz in their original parent material, i.e., mafic rocks. Consequently, Alisols tend to be a less permeable soil system, especially considering that neither Acrisols nor Alisols have a microgranular structure in the B horizon like Ferralsols (Oxisols). In Ferralsols, soil structure is more important than soil texture in terms of physico-hydraulic behavior (Ferreira et al., 1999).

In terms of time, the formation of soils having subsurface argic horizons was estimated to about 29 thousand years (Butler, 1958). This is for general reference only, as it is well-known that rates of soil formation are influenced by multiple factors, such as resistance of bedrock to weathering and climatic conditions, relief (that may promote variable erosion and water infiltration), soil mineralogy, and amount and type of soil organic matter, among others.

Processes of formation of Acrisols and Alisols

The multiple pedogenetic processes that explain soil formation are divided into four main categories: additions, removals, transformations and transfers. Each of these processes affects both the mineral and organic components present in soils, causing differences in soil evolution.

Concerning the formation of Acrisols and Alisols, additions are mainly related to organic matter that accumulates in these soils, mainly but not exclusively in the surface horizons. The B horizons of some Alisols may have larger contents of organic carbon in their upper part. This is probably a consequence of cooler temperatures under which Alisols form and that reduce the rate of organic matter decomposition by microorganisms. Moreover, these cooler temperatures coupled with abundant and well-distributed rainfall during the year favor the formation of dense natural vegetation cover, such as the native forests commonly found in areas with these two soil classes.

Losses are related to the loss of plant nutrients by an intense degree of leaching of these soils, causing low base saturation values and, hence, low natural soil fertility. Other elements (not nutrients) are also leached from these soils. Moreover, the subsurface argic horizon, which reduces water infiltration, makes these soils susceptible to superficial (runoff) losses of mineral material, water,

organic carbon, nutrients and other elements by water erosion, especially on undulating or steeper relief, which is common for both these soil classes. Losses from the interflow of percolation water can also occur, and it may also reach the vadose zone.

Transformations within the profile are mainly caused by the conversion of easily weatherable primary minerals (EWPM), such as feldspars and micas, into kaolinite and other clay-sized particles, such as iron oxides, e.g., goethite, maghemite, and hematite. This is typical for Acrisols. On the other hand, Alisols are characterized by the substantial presence of hydroxy-interlayered vermiculite (HIV) together with kaolinite in the clay fraction.

Transfer reactions can be regarded as the most typical formation process of Acrisols and Alisols. They encompass the translocation of clay minerals from surfacer horizons (A and E horizons) and their accumulation in the argic subsurface horizons (Bt horizon). Consequently, this latter horizon will be enriched with clay compared to the A, E, and C horizons. Clay skins on the surface of peds and pores of this illuvial horizon are common resulting from the deposition and orientation of clay particles eluviated from surface horizons. It is important to mention that some soil fauna (e.g., termites and ants), however, can easily destroy these features in Brazilian Acrisols and Alisols, which is in contrast to the fact that the soil formation processes took millions of years. Moreover, the translocation of clay particles from the surface horizons induces a larger concentration of residual sand-sized particles, causing a texture gradient between the Bt and surface horizons. Translocation results in a series of soil properties that have important effects on both environmental and agricultural issues. Some examples are: reduction in rate of water infiltration and degree of macroporosity (causing some restriction in water percolation); greater difficulty of root growth in the Bt horizon, mainly for perennial crops; and elongated root systems associated with the blocky structure and increase in bulk density, reduced diffusion and consequent uptake of phosphorus, etc. (other consequences are better explained in the following sections). Conversely, the texture gradient of these soils reduces the loss of water by capillary rise because of discontinuity in the soil pores from the Bt horizon (greater clay content and smaller size of pores) in comparison to the greater continuity in pores of the surface horizons (greater sand content and larger pore size).

From the multiple to the specific pedogenetic processes of soil formation, some of the latter considered important for the formation of Acrisols and Alisols are:

- Leucinization: paleness of the lower part of A horizons by the loss or transformation of SOM, producing E horizons (Fig. 3).
- Melanization: darkening of mineral material by mixing with SOM, producing thick A horizons and darkened upper argic subsurface horizons mostly in Alisols (Fig. 4D and F).
- Desilification: release and partial removal of silica from the soil.



Fig. 3 Lessivage, leucinization and ferrolysis (shown in detail in the zoomed portion) with degradation of the top of the Bt horizon in an Acrisol. Source: Prof. Nestor Kämpf's personal archive.

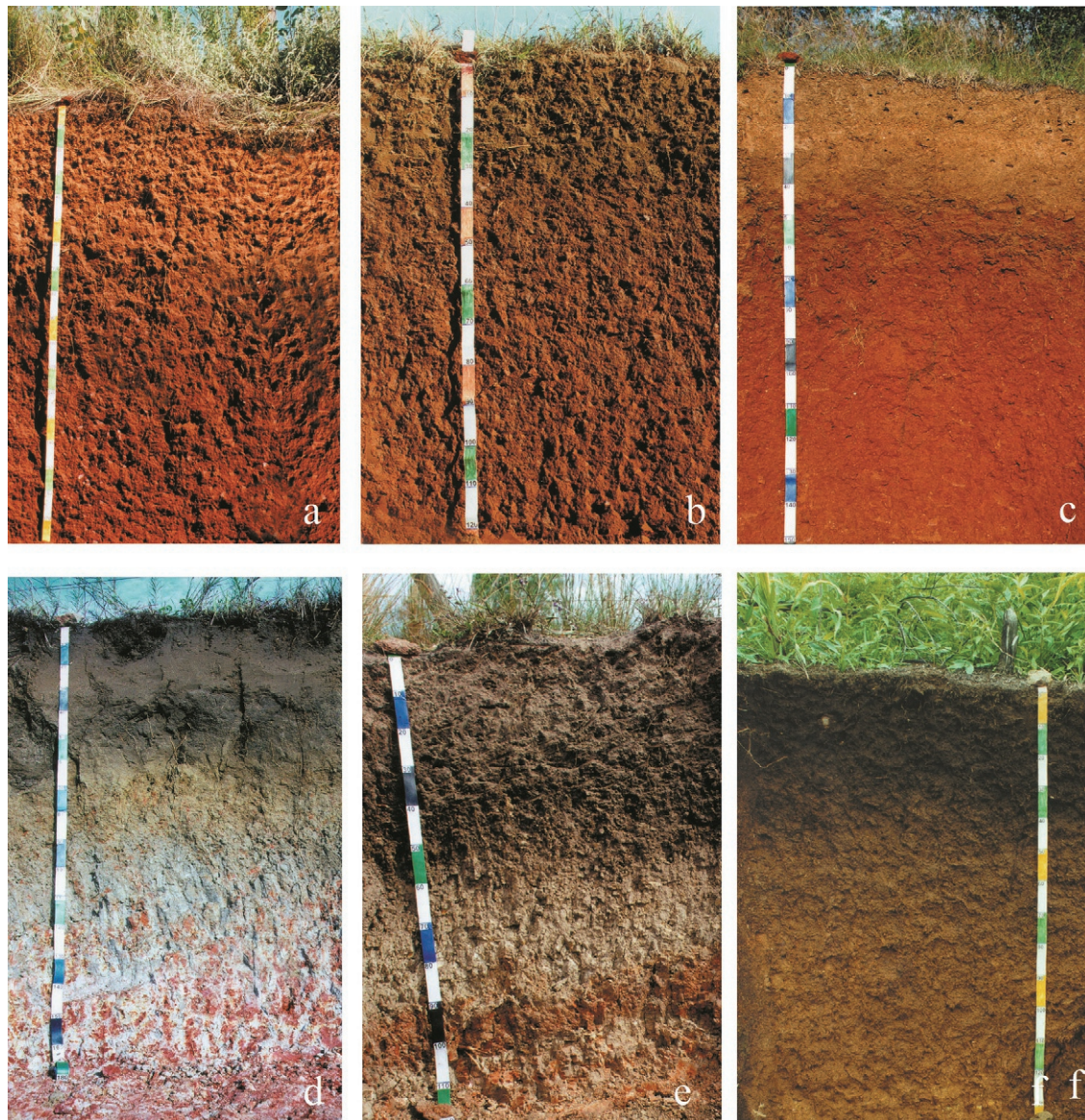


Fig. 4 Variation in soil profiles of Acrisols (A, B and C) and Alisols (D, E and F). Adapted from Streck E, Kämpf N, Dalmolin RSD et al. (eds.) (2018) *Solos do Rio Grande do Sul* (3rd edn). Porto Alegre: UFRGS-EMATER/RS-ASCAR, with permissions granted.

- Brunification, rubefication, ferruginization: release of Fe^{2+} from primary minerals, oxidation and dispersion of Fe oxide minerals, imposing brownish, yellow and red–yellow colors mainly to the matrix of Alisols (Fig. 4D–F), and red colors mainly to the matrix of Acrisols (Fig. 4A–C).
- Argilluviation or lessivage: it is related to the translocation or migration of clay-sized particles from the A and E horizons (eluvial horizons) to the B horizon (illuvial), promoting the formation of a Bt horizon (Fig. 3). Argilluviation is related to the classic vertical migration of clay-sized particles toward Bt horizons (translocation), whereas elutriation or appauvrissement (van Breeman and Buurman, 1998) is caused by lateral movement (multiple process) of clay (and fine silt) particles from the surface horizons by lateral subsurface water flow. This process is facilitated when the impact of water drops previously accumulated on the leaves of native forest trees or direct impact of raindrops on bare soil detach clay particles in a very thin layer of the soil surface, similar to the first step for water erosion. However, instead of being carried away by surface runoff (erosion), these dispersed clay particles are moved from the surface into the first few centimeters of the soil surface horizon. Eventually these dispersed particles can be moved laterally within the soil horizons (i.e., at the contact of the A with the Bt horizon) as subsurface water moves according to the landscape toward lowland areas, including floodplains. In this process, water removes clay-sized particles from the upper horizons, resulting in the accumulation of clay in the Bt horizon.
- Ferrollysis: this process includes reduction, oxidation and hydrolysis. It occurs under seasonal hydromorphic conditions where alternating reduction and oxidation reactions take place during repeated wetting and drying cycles. This reduces the Fe^{3+} present

in phyllosilicate minerals of the surface or upper part of B horizons (much more prone to wetting and drying cycles than the middle and lower part of the B and C horizons). Phyllosilicate minerals are then disintegrated by hydrolysis, which reduces the clay content in the surface and upper B horizons (Fig. 3). This type of pedogenic process explains the presence of the texture gradient between the B and A horizons (Almeida, 1992).

- Xanthization: Yellowing of the surface horizons by SOM complexing with Fe, preventing the precipitation of ferrihydrite and consequent formation of hematite. Therefore, goethite is the dominant form of iron in the upper part of the profile of Acrisols (Fig. 3C), which may result in a bichromic soil profile.
- Erosion: removal of soil material, water, organic carbon, plant nutrients and other elements (not plant nutrients) by water erosion, with (accelerated erosion) or without human interference (geologic or natural erosion). Erosion has varied in importance in features of the Brazilian landscape, mainly in the case of polygenetic soils (Rezende, 1980). In certain regions some Oxisols have been eroded, forming Inceptisols from the material of ancient Oxisols due to the decrease in thickness of the B horizon (Rezende, 1980).

Properties of Acrisols and Alisols

In general, Acrisols and Alisols have some common properties such as the texture gradient related to a greater clay content in the Bt horizon (Fig. 4). However, they also have some important differences. Acrisols, as defined by the World Reference Base for soil resources (IUSS Working Group, 2015), have an argic horizon starting within the first 100 cm from the surface, and a CEC (by 1 M NH_4OAc , pH 7) less than $24 \text{ cmol}_c \text{ kg}^{-1}$ clay in some part of the argic horizon within ≤ 50 cm below its upper limit. These soils also have a base saturation lower than 50% between 50 and 100 cm from the mineral soil surface. Acrisols might also have a hard or cemented or indurated layer in the lower half of the mineral soil above continuous rock, starting below 100 cm from the soil surface. The main adjectives to describe these layers include: Abruptic, Fragic, Leptic, Petroplinthic/Pisoplinthic/Plinthic, Hydragric/Anthraquic/Pretic/Terric, Gleyic, Stagnic, Nudiargic, Lamellic, Albic, Ferric, Rhodic/Chromic/Xanthic, Skeletic, and Haplic. The supplementary adjectives are: Alumatic, Andic, Arenic/Clayic/Loamic/Siltic, Aric, Colluvic, Cutanic, Densic, Differentic, Epieutric, Fractic, Gibbsic, Humic, Magnesic, Nechic, Nitic, Novic, Oxyaquic, Profundic, Ruptic, Sombric, Technic, Toxic, Transportic, Vertic, Vitric (IUSS Working Group, 2015).

Many Acrisols can be correlated with Red–Yellow Podzolic soils (e.g., in Indonesia), Argisols (Brazil) (SiBCS., 2018), Kurosols (Australia), Sols ferralitiques fortement ou moyennement désaturés (France), and Ultisols with low-activity clays (US Soil Taxonomy) (Soil Survey Staff, 2014). Acrisols do not have retic properties, i.e., coarser-textured albic material mixed into a finer-textured argic horizon. Moreover, their name connotes to the Latin word *acer* (very acid), which is a consequence of their relatively strong degree of weathering (IUSS Working Group, 2015, with adaptations).

Alisols also have a base saturation lower than 50% between 50 and 100 cm from the mineral soil surface and an argic horizon, but, conversely to Acrisols, the argic horizon of Alisols starts at a depth greater than 100 cm from the soil surface. They might also have a hard or cemented or indurated layer in the lower half of the mineral soil above continuous rock, starting in the first 100 cm from the mineral soil surface (IUSS Working Group, 2015). The main adjectives of these layers are: Abruptic, Fragic, Leptic, Petroplinthic/Pisoplinthic/Plinthic, Hydragric/Anthraquic/Plaggic/Pretic/Terric, Gleyic Stagnic, Nudiargic, Lamellic, Albic, Ferric, Rhodic/Chromic, Skeletic, and Haplic. The supplementary adjectives are: Alumatic, Andic, Arenic/Clayic/Loamic/Siltic, Aric, Neocambic, Colluvic, Cutanic, Densic, Differentic, Epieutric, Fluvic, Fractic, Gelic, Humic, Hyperallic, Magnesic, Nechic, Nitic, Novic, Oxyaquic, Profundic, Ruptic, Protospodic, Technic, Toxic, Transportic, Turbic, Vertic, and Vitric.

Alisols have high-activity clays (e.g., smectite and vermiculite) throughout the argic horizon. Similar to Acrisols, Alisols do not have retic properties (explained above) by definition.

Many Alisols can be correlated with Parabraunerden (Germany), Argisols (Brazil) (SiBCS., 2018), Ultisols with high-activity clays (US Soil Taxonomy) (Soil Survey Staff, 2014), Kurosols (Australia), and Fersialsols and Sols fersiallitics très lessivés (France) (IUSS Working Group, 2015). Moreover, their name connotes to the Latin *alumen* (aluminum) related to their large exchangeable Al^{3+} content.

Both Acrisols and Alisols may lose Fe oxide minerals, other clay minerals, soil organic matter and other components, resulting in the formation of a bleached E horizon above the argic horizon (IUSS Working Group, 2015) (Fig. 3). As previously mentioned, both soils can also show lateral loss of the clay fraction by elutriation, contributing to the generation or increment of a texture gradient between the Bt and A horizons. This helps to explain the relative increment of the clay fraction in the Bt horizon observed in cases when clay skins do not exist.

The main differences between Acrisols and Alisols are derived from differences in their parent material and degree of weathering (relatively high in Acrisols and moderate in Alisols). Consequently, Acrisols have low-activity clays and Alisols high-activity clays throughout the argic horizon. Their different clay mineralogies increase bulk density and reduce permeability in Alisols (consequence of larger clay content, controlled mainly by the much smaller content of quartz in the mafic parent materials – basalt, gabbro, and diabase) compared with Acrisols (lower clay content, controlled mainly by the much greater content of quartz in the felsic parent materials – granite and rhyolite) (Fig. 5). Table 1 shows the differences in attributes of an Acrisol and an Alisol from southern Brazil.

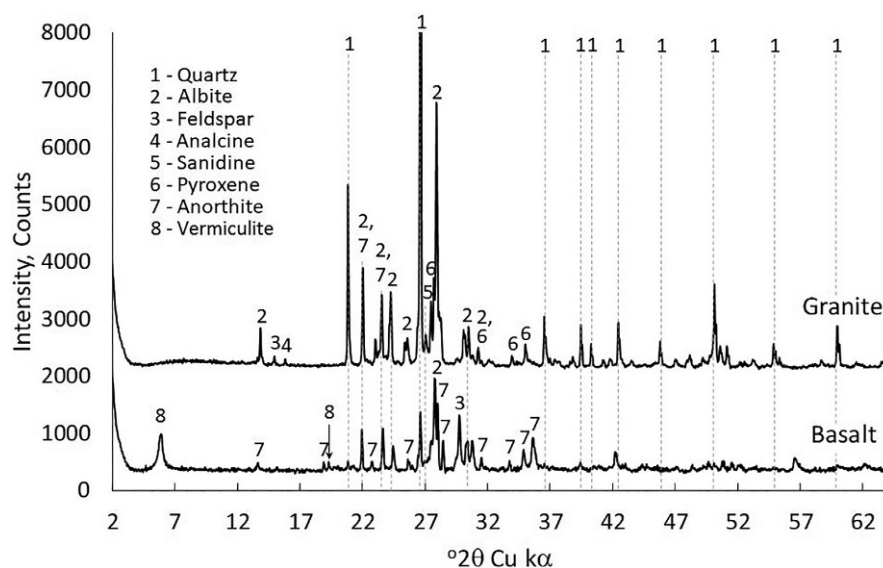


Fig. 5 X-ray diffractograms of granite and basalt.

Table 1 Examples of differential attributes of an Acrisol and an Alisol from southern Brazil.

Soil	Horizon	Clay content (%)	Exchangeable Al^{3+} ($cmol_c\ kg^{-1}$)	CEC ^a ($cmol_c\ kg^{-1}$)	Al saturation (%)	SOM ^b (%)
Acrisol	A	8	0.45	2.6	17	0.99
	Bt	58	2.15	9.8	22	1.21
Alisol	A	52	1.7	16.2	24	4.06
	Bt	65	5.9	9.3	86	0.36

Data from Acrisol were selected from Almeida (1992) and data from Alisol were selected from Curi (1975).

^aCEC – Cation exchange capacity.

^bSOM – Soil organic matter.

Kaolinite tends to predominate in the clay fraction of both Acrisols and Alisols. However, in the latter HIV is also present in its clay fraction together with a smaller kaolinite content than in Acrisols (Fig. 6). Quartz is present, but gibbsite has not been detected in the X-ray diffractograms of both soils (Figs. 5 and 6), although it can be present in very small amounts only in some samples.

Some explanations for the above features are related to the fact that Alisols tend to occur at higher altitudes than Acrisols, at least in southern Brazil; therefore, Alisols are associated with less weathering than Acrisols because of the lower temperatures at higher altitudes. The mafic parent material of Alisols also has much less quartz content and tends to produce soils with higher clay content than the felsic parent material of Acrisols (Fig. 5). Moreover, this difference in parent materials tend to form Alisols that have greater density and are more impermeable than Acrisols. These differences in parent material help to explain the presence of high-activity clays in the Bt horizon of Alisols and the low-activity clays in the Bt horizon of Acrisols.

Hydroxy-interlayered vermiculite (HIV) is a pedogenic clay mineral with Al-hydroxy interlayers formed in acid environments by the intercalation of Al polymers in the interlayer space of the 2:1 structures of vermiculite (more common) and smectite (less common). The Al-hydroxy polymers modify substantially the physico-chemical properties of these clay minerals. The extent of such alteration is proportional to both the degree to which the interlayers are filled with Al-hydroxy polymers and the stability of these components (Kämpf et al., 2012).

The conditions that favor the formation of interlayer Al-hydroxy laminae in vermiculite and smectite are: active weathering to release Al^{3+} ions; moderately acid environment, around pH 5; low soil organic matter content to prevent Al complexation; and frequent wetting and drying cycles (Rich, 1968). Many authors, for example Barnhisel and Bertsch (1989), have demonstrated experimentally the processes of formation of such clay minerals and changes in the physical and chemical properties of the original vermiculite and smectite.

Hydroxy-interlayered vermiculite tends to occur mainly in the coarse clay fraction (2–0.2 mm) of soils and has increased concentration in the surface horizons of Alisols (Curi et al., 1984), indicating its pedogenic formation and strong resistance to weathering. The HIV is more stable thermodynamically than kaolinite (Karanthanas, 1988). The Si biocycling can also contribute to its stability in the surface horizons of soils (Kämpf et al., 2012).

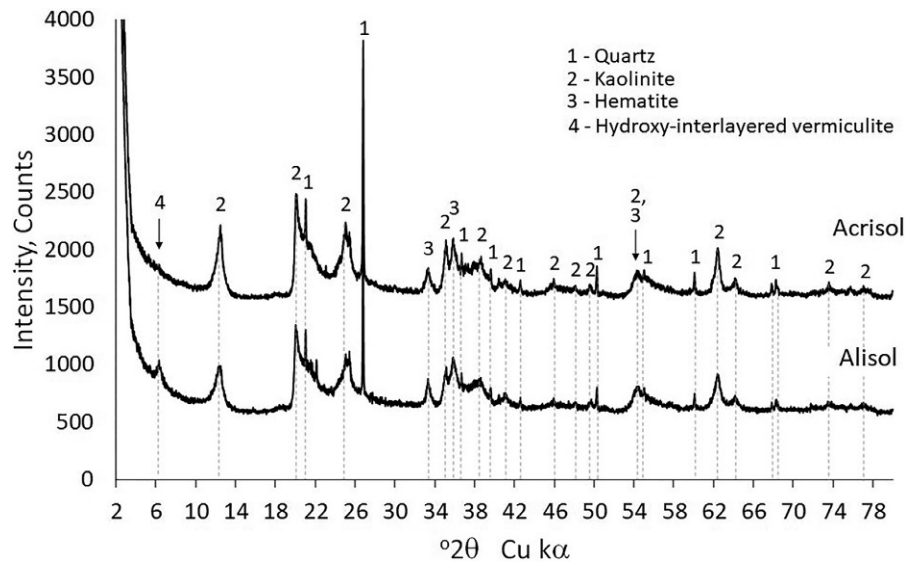


Fig. 6 X-ray diffractograms of the clay fraction of an Acrisol and Alisol.

In terms of environmental importance, the filling of vermiculite and smectite interlayers by Al—OH (common) and Fe—OH (less common) produces significant changes in the chemical and physical properties of these clay minerals: increase in variable charge, and reduction in internal specific surface area and of swelling–shrinkage capacity. The fixation of anions (phosphates, borates, sulfates) and heavy metals has been attributed to these Al-hydroxy interlayers (Barnhisel and Bertsch, 1989). Moreover, this fixation is the cause of very large exchangeable Al^{3+} content under certain soil environmental conditions (Marques et al., 2002).

Sustainable use and management of Acrisols and Alisols

In humid subtropical conditions, soil and water conservation are crucial for agriculture on Acrisols and Alisols because of the undulating or steeper relief on which they generally occur. Considering low and intermediate technological management systems, they are most suited for use with perennial crops because of the small number of annual interventions. Within this scope, extensive pasture (low technological systems) and forestry (intermediate technological systems) constitute sustainable land uses.

Under advanced technological management systems, including the control of water erosion and soil compaction, liming or the use of agricultural gypsum and fertilizer applications, crop succession/rotation, appropriate genetics and mechanization in all phases of the production process, Acrisols and Alisols can achieve high productivity for maize (*Zea mays* L.), soybeans (*Glycine max* (L.) Merrill, wheat (*Triticum aestivum* (L.), etc., in particular under no-till management systems (Peixoto et al., 2020).

Gradient terracing (prepared with a small degree of unevenness transverse to the greatest slope) is recommended for soil and water conservation, to take account of the decrease in permeability in the Bt horizon of Acrisols and Alisols. The poor natural soil fertility is common for both soil types due to their relatively high (Acrisols) and moderate (Alisols) degrees of weathering, with a need for liming and fertilization. Larger contents of exchangeable Al^{3+} are common in Alisols, which require substantial amounts of lime or gypsum for sustainable cultivation of the most sensitive crops. When Alisols are managed appropriately for soil acidity and are adequately fertilized, the crops can benefit from their high activity clays (high CEC) and adequate water-holding capacity; they behave similarly to Luvisols which have high activity clays and a large base saturation (IUSS Working Group, 2015). Extensive pastures, acid-tolerant crops and forestry (with eucalyptus (*Eucalyptus* L'Hér.) and pines (*Pinus* L.)) are suggested in the adoption of low and intermediate technological systems for soil management and land cultivation.

The nature of HIV clay minerals has resulted in their dynamic transformation by soil management practices (Kämpf et al., 2012). In this context, management systems that favor the accumulation of organic compounds in soil, (for example, no-till) can, in the long term, destabilize and remove this Al-hydroxy interlayer from vermiculite (Inda et al., 2010). This results in an increase in acidification, which needs appropriate amendments based on soil analyses and monitoring.

Fink et al. (2014), based on a 10-year experiment in an Acrisol under humid subtropical conditions, tested the following treatments: native field (NF), crop under a no-till system without the application of poultry litter (NPL), crop under no-till system with the application of poultry litter (PL), and native pasture with the application of manure and cattle droppings (CD). They concluded that the change in land use to CD promoted dissolution of the 2:1 clay minerals (including HIV), which was shown by the decrease in reflection intensity of these minerals and their area in the X-ray diffractograms, and by the greater activity of Al and Si in the soil solution, corroborating the dynamic nature of these clay minerals under different soil management systems.

A comparison of phosphorus (P) adsorption in Acrisols and Alisols with Ferralsols and Nitisols by Oliveira et al. (2020) showed that it was positively correlated with soil clay and organic matter contents, and specific surface area. The clay minerals most strongly influencing P adsorption were goethite, followed by hematite and gibbsite. These clay minerals represent an increasing order of crystal size (Kämpf et al., 2012), and as the Acrisols and Alisols have smaller contents of the Fe oxide minerals than the Ferralsols and Nitisols, their P retention is much smaller. This results in a potential saving of P-fertilizers for sustainable crop production under moderate and high technological management systems.

Conclusion

Acrisols and Alisols in the humid subtropical ecozone occur on different continents, mostly in their southeastern and southern regions. The main differences between Acrisols and Alisols are derived from differences in their parent materials and degree of weathering (greater in Acrisols than in Alisols). However, both soils have an argic horizon in the subsurface, which can increase the risks of water erosion because of delayed infiltration. These processes are exacerbated because of the undulating or steeper relief on which they generally occur. This reinforces the need to adopt sustainable conservation practices, such as gradient terracing, to avoid losses of soil material, water, organic carbon, plant nutrients and other elements (non-nutrient). With proper land use and management systems, both Acrisols and Alisols can be highly productive and sustainable.

References

- Almeida JA (1992) *Degradação do topo do horizonte B de um Podzólico Vermelho-Amarelo abrupto da Planície Costeira do RS*. Porto Alegre: Universidade Federal do Rio Grande do Sul.
- Barnhisel RI and Bertsch PM (1989) Chlorites and hydroxy-interlayered vermiculite and smectite. In: Dixon JB and Weed SB (eds.) *Minerals in Soil Environments*, pp. 729–788. Madison: Soil Science Society of America.
- Butler BE (ed.) (1958) Depositional Systems of the Riverine Plain of South-eastern Australia in Relation to Soils, *Soil publication No 10*. Melbourne: CSIRO.
- Curi N (ed.) (1975) *Relações genéticas e geomórficas em solos das encostas inferior e superior do nordeste, no Rio Grande do Sul*. Porto Alegre: Universidade Federal do Rio Grande do Sul.
- Curi N, Kämpf N, and Resende M (1984) Mineralogia, química, morfologia e gemorfologia de solos originados de rochas efusivas das Encostas Superior e Inferior do Nordeste, no Rio Grande do Sul. *Revista Brasileira de Ciência do Solo* 8: 269–276.
- FAO-UNESCO (1974) In: Vol I and Legend I (eds.) *Soil map of the world 1:5000000*. Rome: FAO-UNESCO.
- Ferreira MM, Fernandes B, and Curi N (1999) Mineralogia da fração argila e estrutura de Latossolos da região sudeste do Brasil. *Revista Brasileira de Ciência do Solo* 23: 507–514.
- Fink JR, Inda AV, Almeida JA, et al. (2014) Chemical and mineralogical changes in a Brazilian Rhodic Paleudult under different land use and managements. *Revista Brasileira de Ciência do Solo* 38: 1304–1314.
- Food and Agriculture Organization [FAO] (2012) *Global Ecological Zones for FAO Forest Reporting: 2010 Update*. Rome: FAO.
- Inda AV, Torrent J, Barrón V, and Bayer C (2010) Aluminum hydroxy-interlayered minerals and chemical properties of a subtropical Brazilian Oxisol under no-tillage and conventional tillage. *Revista Brasileira de Ciência do Solo* 34: 33–41.
- IUSS Working Group (2015) World reference base for soil resources 2014 International soil classification system, World Soil Resources Reports No. 106. Rome: FAO.
- Kämpf N, Marques JJ, and Curi N (2012) Mineralogia de solos brasileiros. In: Ker JC, Curi N, Schaefer CEGR, and Vidal-Torrado P (eds.) *Pedologia - Fundamentos*, pp. 81–146. Viçosa: Sociedade Brasileira de Ciência do Solo.
- Karathanasis AD (1988) Structural and solubility relationships between Al-hydroxyinterlayered soil smectites and vermiculites. *Soil Science Society of America Journal* 52: 1500–1508.
- Marques JJ, Teixeira WG, Schulze DG, and Curi N (2002) Mineralogy of soils with unusually high exchangeable Al from the western Amazon Region. *Clay Minerals* 37: 651–661.
- Oliveira JS, Inda AV, Barrón V, et al. (2020) Soil properties governing phosphorus adsorption in soils of Southern Brazil. *Geoderma Regional* 22: e00318.
- Omernik JM (2004) Perspectives on the nature and definition of ecological regions. *Environmental Management* 34: 27–38.
- Peixoto DS, Silva LCM, Melo LBB, et al. (2020) Occasional tillage in no-tillage systems: A global meta-analysis. *Science of the Total Environment* 745: 140887.
- Resende M, Curi N, and Poggere GC, et al. (eds.) (2021) *Pedologia, fertilidade, água e planta: Inter-relações e aplicações*, 2nd edn. Lavras: Editora UFLA.
- Rezende SB (ed.) (1980) *Geomorphology, Mineralogy and Genesis of Four Soils on Gneiss in Southeastern Brazil*. West Lafayette: Purdue University.
- Rich D (1968) Hydroxy interlayers in expansible layer silicates. *Clays and Clay Minerals* 16: 15–30.
- Schwertmann U and Taylor RM (1989) Iron oxides. In: Dixon JB and Weed SB (eds.) *Minerals in Soil Environments*, pp. 379–438. Madison: Soil Science Society America.
- SIBCS. (2018) *Sistema Brasileiro de Classificação de Solos*, 5th edn. Rio de Janeiro: Embrapa Solos.
- Soil Survey Staff (ed.) (2014) *Keys to Soil Taxonomy*, 12th edn Washington: USDA-NRCS.
- Streck EV, Kämpf N, and Dalmolin RSD, et al. (eds.) (2018) *Solos do Rio Grande do Sul*, 3rd edn. Porto Alegre: UFRGS-EMATER/RS-ASCAR.
- van Breeman N and Buurman P (eds.) (1998) *Soil Formation*. Dordrecht: Kluwer Academic Publishers.

Soils of the humid and sub-humid tropics

Karl Stahr and Ludger Herrmann, Department of Soil Science and Land Evaluation (310a), Universität Hohenheim, Stuttgart, Germany

© 2023 Elsevier Ltd. All rights reserved.

This is an update of S.W. Buol, Tropical Soils I Humid Tropical, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, pp. 187–195, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00511-7.s>.

Introduction: Delineation of tropical soils	353
Factors governing soil development and fertility	355
Moisture and temperature conditions (climate)	355
Old continents and young landscapes (time and parent material)	357
Lowlands, uplands and mountain ranges (relief)	357
Biotic diversity and interrelation with soils (termites and their multiple roles)	357
Lateritic weathering as central concept for the tropical soil and landscape development?	358
Soils, processes and landscapes	361
Humid zone river soilscapes (Fluvisol, Gleysol, Histosol, Podzol, Terra preta; India, Congo, Brazil)	361
The Terra preta story	362
Humid tropical uplands (Ferralsol, Plinthosol, Planosol; Brazil, Benin, Thailand)	363
Seasonal river soilscapes (Fluvisol, Cambisol, Vertisol, Gleyic Solonchak; Amazon, Congo, Niger)	363
Sub-humid tropical uplands (Acrisol, Luvisol, Nitisol, Cambisol, Planosol; Benin, Niger, Ghana)	364
Young volcanic territories (Leptosol, Regosol, Andosol, Phaeozem; Ecuador, Ethiopia, Indonesia)	364
Human influences on soil fertility	365
Outlook and research questions	366
References	366

Abstract

The tropics are between the equator and Tropics of Cancer and Capricorn. They change with latitude from perhumid to seasonal rainfall. The vegetation period extends to the entire year around the equator down to at least 90 days at higher latitudes. Parent materials and age of the landscapes also differ. They include old strongly weathered landscapes (e.g., Southern India), young river basins (e.g., Amazon Basin) and active volcanic areas (e.g., Indonesia). The older fluvial landscapes are rather levelled, but in tectonically active and volcanic landscapes relief energy can be large. Characteristic zonal soils are Ferralsols in the perhumid and Acrisols in sub-humid climate zones. In addition, depending on parent rock and relief Nitisols, Luvisols, Plinthosols, Andosols and Vertisols also occur.

Key points

- In the humid and sub-humid tropics there is a contrast between old table lands and low lying river soilscapes.
- Lateritic weathering occurs in the tropics, but it is only found in landscapes with long lasting continuous weathering.
- Desilification is a prominent process in acid tropical soils.
- Clay illuviation occurs mainly in landscapes with seasonal climate.
- The “Terra preta” represents an anthropogenic improvement of chemically poor soils.

Introduction: Delineation of tropical soils

The tropics are the area where the sun reaches the zenith position, i.e., an angle position relative to the ground of 90° during at least 1 day of the year. This is exactly to a latitude of 23° 26′ 5″ north and south from the equator at present. At the continental scale climate is the major variable for explaining soil variation, but at higher resolution (i.e., regional) parent material and landscape age (i.e., time for undisturbed soil development) become more important. At the landscape scale, relief is a major explanatory variable.

In total, we find about 10 soil zones around the world (Table 1 and Fig. 1). Three of these occur in the tropics (Mohr et al., 1972), namely the rainforest Ferralsol-Fluvisol-Gleysol zone, the Savannah Nitisol-Acrisol-Vertisol zone, and the semi-desert Regosol-Calciisol-Solonchak zone (FAO, 2015; IUSS, 2022). The latter leads over to the subtropics, is less relevant for land use, and therefore, will not be covered here.

The tropics can be subdivided into three major landscape types: 1. The old remains of the big southern supercontinent Gondwana in Brazil, Central Africa and Southern India, which are tectonically stable and thus have been subject to weathering

Table 1 Soil zones and their characteristics including the tropics with increasing latitude.

<i>Soil zone</i>	<i>Climate (Köppen)</i>	<i>Natural vegetation</i>	<i>Suitability</i>	<i>Soil fertility</i>	<i>Natural productivity</i>	<i>Minimum factors</i>
1. Ferralsol–Fluvisol–Gleysol	Af	Tropical rainforest	Forest, agriculture	Low	High	Nutrients (K, P, N)
2. Acrisol–Nitisol–Plinthosol–Vertisol	Aw	Savanna	Forest, agriculture, pasture	High	Medium-high	Water, nutrients (P, N, K)
3. Arenosol–Leptosol–Gypsisol–Solonchak	Bsh/BWh	Extreme desert	–	Low	Very low	Water
4. Regosol–Calcisol–Solonchak	BS	Semi-desert	Pasture, agriculture	Medium	Low	Water, salt
5. Chromic Lixisol–Acrisol–Calcic Cambisol–Planosol	Csa	(Evergreen) sclerophytic forest	Agriculture	Medium-high	High-very high	Degradation, nutrients (N, P)
6. Kastanozem–Chernozem–Phaeozem–Solonetz	D/BS	Steppe	Pasture, agriculture	Very high	Medium	Water, heat
7. Cambisol–Luvisol–Gleysol	Cfb, (D)	Deciduous forest	Agriculture, pasture	Medium	Medium	Heat, nutrients (N), water
8. Podzol–Retisol–Cambisol–Histosol	Df	Taiga	Forest, (pasture)	Low	Low	Heat, nutrients (N)
9. Gelic Regosol–Cryosol–Gellic Gleysol	ET	Tundra	Pasture	Low	Very low	Heat
10. Ice-covered–Gellic Leptosol	EF	Cold desert	–	Very low	Lacking	Heat, water

Meaning of Köppen abbreviations: Capital letters: A equatorial, B arid, C warm temperate, D boreal, E polar, F polar frost, S steppe, W desert. Lower case letters: a hot summer, b warm summer, f fully humid, h hot arid, s summer dry.

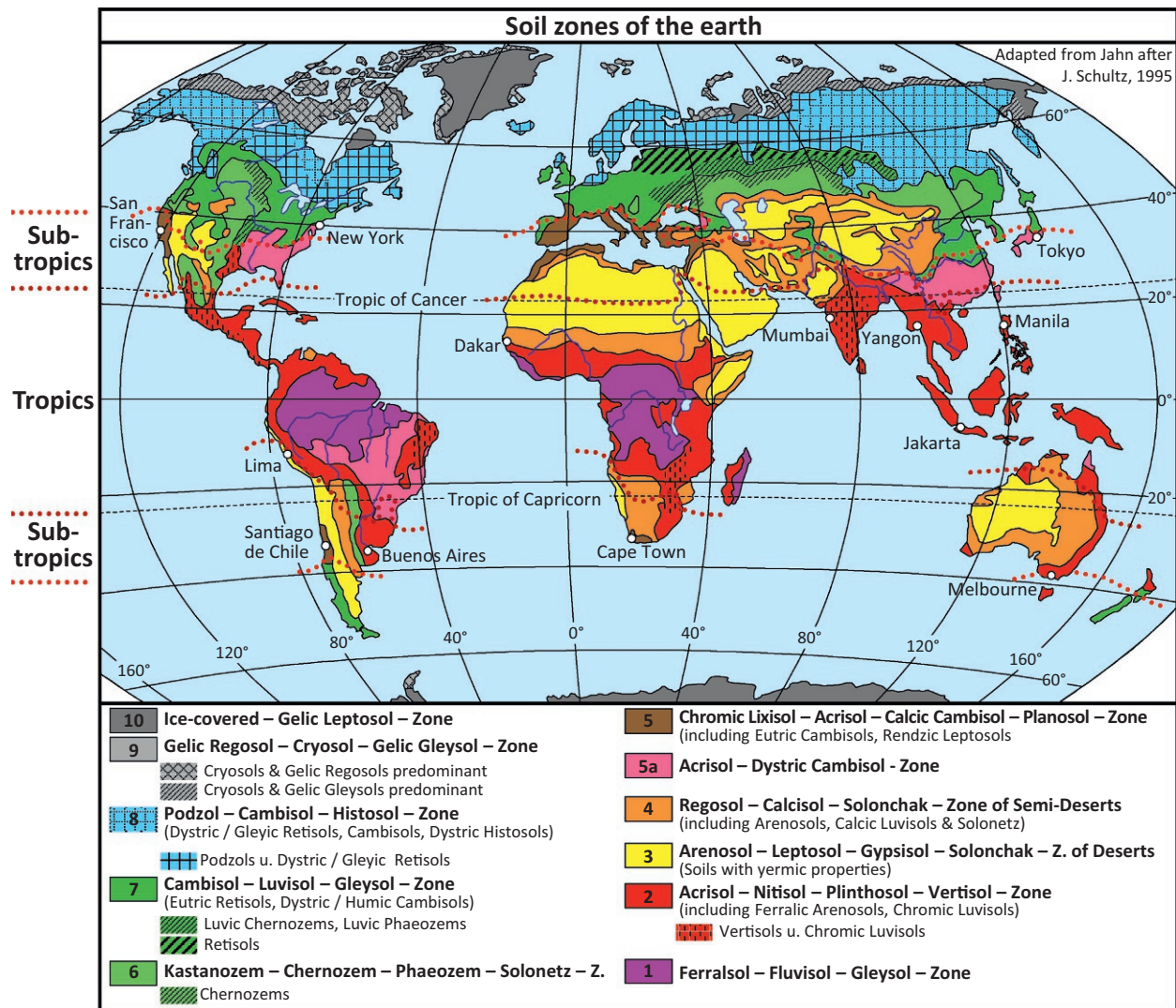


Fig. 1 Map of global soil zone distribution. Adapted from Schultz, J. (1995) *Ökozonen der Erde*, 2nd edn. Ulmer, Stuttgart.

and soil development since the Jurassic period; 2. The large river floodplains of the Amazon and the Congo; 3. The young volcanic regions of Central Africa, Indonesia, the Philippines and South America. Towards the sub-humid and semi-arid areas, there are increasingly younger sedimentary environments (Tertiary to Quaternary in age) that are remnants of climate change in the past.

Factors governing soil development and fertility

The water budget at a site, i.e., the ratio of rainfall to evapotranspiration determines the direction of water movement in the soil and thus the rates of weathering and mineral neoformation. The latter two are modified by the parent material, i.e., fabric, mineral and geo-chemical composition, and these latter components affect the potential nutrient status. The time taken for soil formation determines the rate and state of weathering, and the fertility status. Therefore, soils cannot be understood without considering these three major factors: climate, parent material and time. Relief resulting from tectonic processes, surface erosion and sedimentation also play a role at the landscape scale, and recently human influence has become important. Therefore, these aspects are elucidated in more detail in the following.

Moisture and temperature conditions (climate)

A globally accepted approach to categorize the climate effect on soil has been developed by Soil Survey Staff (1999) on the basis of the Köppen climate classification (1936). It differentiates so-called soil moisture regimes (SMR), and soil temperature regimes (STR).

Soil-moisture regimes (SMRs) are defined to classify a soil's ability to supply water to annual plants and crops without irrigation. In soils where the groundwater table is not reached by the roots of these plants, the SMR is determined by the seasonal distribution of rainfall in "normal" years. Most food crops require a reliable supply of water for at least 90 consecutive days.

The perudic SMR has precipitation that exceeds potential evapotranspiration every month of normal years. Although this may seem desirable, these areas have weed, insect, and disease problems, and the constantly humid conditions make it difficult to harvest mature grain crops.

The udic SMR has fewer than 90 cumulative days when water is not available in the rooting zone in normal years. It is possible to grow food crops any time of the year without irrigation when the temperature is warm enough for the particular crop. Available water is less reliable during some parts of the year, and farmers often select more drought-tolerant crops, but perennial plants are adequately supplied with water throughout most years in the udic SMR areas.

The ustic SMR is borderline to the common concept of humid. In normal years, soils with an ustic SMR have at least 90 consecutive days when moisture is available, but more than 90 cumulative days when water is not available in the rooting zone. Natural vegetation is either seasonal rain forest or savanna. At least one crop can be grown reliably each year, and it is possible to have two harvests per year on some soils, but there is a seasonal dry period of 90 days or more when crop production is not possible without irrigation. The reliable dry season of the ustic SMR is a distinct advantage for weed and disease control and grain-crop harvest in isohyperthermic and isothermic STRs.

Soils that have a reliable moisture supply for fewer than 90 consecutive days in normal years have an aridic SMR and are excluded from the concept of this chapter.

The above SMRs are used to classify only the well-drained upland soils. Groundwater often saturates the rooting zone of some soils within all areas of the tropics. These soils are identified as having an "aquic soil moisture condition."

Mean annual soil temperatures are 2–4 °C warmer than mean annual air temperatures. Soil-temperature regimes (STRs) are defined by two criteria, mean annual soil temperature and seasonal temperature difference. Seasonal soil-temperature difference is determined as the mean soil temperature of June to August compared with the mean soil temperature of December to February. Almost all of the soils within the geographic tropics have soil temperatures that differ seasonally by less than 6 °C and are identified by the prefix "iso" placed before the name of the STR. Except for a few soils near the northern and southern extremes of the geographic tropics in Africa, this characteristic is the only soil property that is almost universal with the concept of tropical soil. The practical aspect of this soil property is that seasonal soil temperatures seldom have to be considered when planting food crops.

Soils with mean annual soil temperatures of 22 °C or higher are classified as isohyperthermic; soils with mean annual soil temperatures of 15–22 °C are identified as isothermic. Freezing conditions are seldom a problem in the isothermic and isohyperthermic STRs except for high mountain areas. Higher elevations in tropical areas have isomesic STRs, with mean annual soil temperatures less than 15 °C; crop growth is slow, night-time freezing is common.

Throughout the areas concerned the temperature is warm or hot all year round. No plants or animals living there need to have any adaption to reduced heat. In the rainforests the air and soil temperature do not change by more than 1 degree centigrade during the year. However, in the savannas during the dry season the temperature in the topsoil may become very hot, and therefore organisms need heat adaption. From both zones, an example of a typical climate has been selected (Figs. 2 and 3). Manaus is a typical case of the humid tropics and has, according to the Köppen–Geiger system, an Af-climate (Köppen, 1936). The average temperature at almost sea level is 26.6 °C and the annual rainfall amounts to 2272 mm. More important, there is no dry season. Consequently, there is a good water supply all year round. For the other case, Brasilia is taken as the example. Here at an altitude of 1158 m a.s.l. the average temperature is 20.7 °C with 1555 mm annual rainfall. This is classified as an Aw-climate. Because of the dry season the growing season is restricted to around 9 months. In other areas in the savanna the dry season may be up to 6 months, and the temperature may rise above 30 °C. There is a need for organisms to adapt to drought and heat.

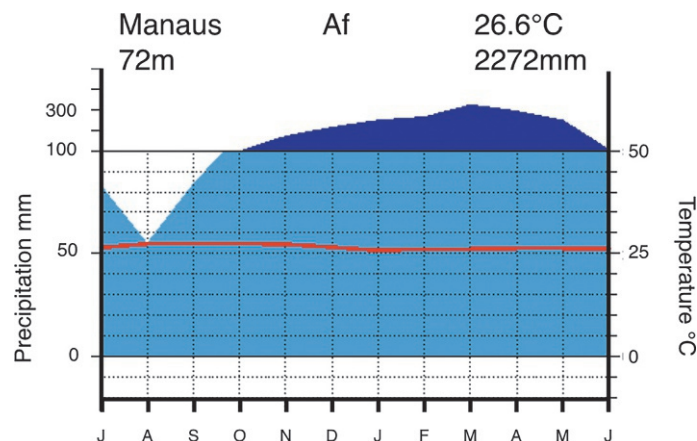


Fig. 2 Climate diagram for Manaus, Brazil in the rainforest zone. Example for a perudic soil moisture and isohyperthermic soil temperature regime. Af represents the climate classification: equatorial, perhumid.

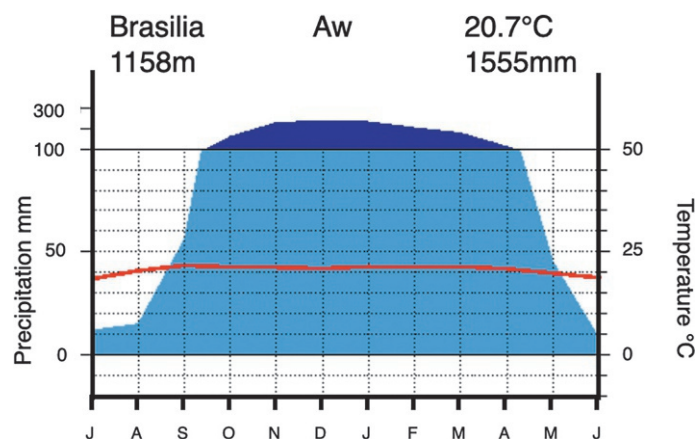


Fig. 3 Climate diagram for Brasilia, the capital of Brazil, situated in a typical savanna region. Example for an udic soil moisture and isohyperthermic soil temperature regime. Aw represents the climate classification: equatorial, “dry winter.”

Old continents and young landscapes (time and parent material)

Soil landscapes in the humid and sub-humid tropics have developed under two extremely different geological and geomorphological situations. The Amazon and Congo basin as well as the Indonesian Archipelago are very young landscapes under strong fluvial or volcanic influence. The others such as Southeast Brazil, West Africa, India and Western Australia are old consolidated continents where soils developed from Palaeozoic or Precambrian sediments, as well as metamorphic and plutonic rocks. Even in young landscapes, deep solum development through podsolization in sandy substrates or peat formation in endorheic depressions is possible. Nevertheless, weathering of the solum and solute transport may reach 50-m depth on the old continental landscapes, leading to the deepest weathering transformations in the world. It is important to realize that most of these soils formed under a tropical climate, but equally important is that they are very old weathering remnants. They may date back to the Jurassic period. During these long periods, continental movement plays a role and variation in climate over time needs to be considered. For example, one can find what is perceived as the results of “tropical” weathering, i.e., plinthite and bauxite in the nowadays non-tropical southwest Australia (Kew et al., 2008).

Lowlands, uplands and mountain ranges (relief)

Tropical landscapes can be roughly divided into three major units. First, the old remains of the big southern super-continent Gondwana in Brazil, Western and Central Africa and Southern India. These generally form levelled upland areas. They are strong and deeply weathered, and often also large in spatial extent. Second, the large river systems, for example of the Amazon, Congo as well as Niger can form huge floodplains, inland deltas, and lowlands ending in peri-marine marshes. These latter ones provide the opposite situation: young landscapes and soils with current weathering. These soils are generally in equilibrium with the actual environmental conditions. Finally, there are younger undulating landscapes mostly generated by younger volcanism such as in the Americas, Central and Eastern Africa and building up Indonesia and entirely forming the Philippines. Here soil formation may start on newly solidified lava, but may have occurred over a much longer time. These three units provide totally different weathering environments, since they differ with respect to age, parent material properties, topography and water residence times. Therefore, these three completely different landscapes dominate the different soilscapes.

Biotic diversity and interrelation with soils (termites and their multiple roles)

While earthworms (Lumbricidae) and ants (Formicidae) are more important soil organisms in the sub-tropics and temperate climates, termites (Isoptera) appear to be the more relevant soil animals in tropical soils due to their structural diversity. While earthworms reach dominance in specific landscapes (e.g., wet lowlands with sufficiently high pH-conditions), termites appear in all types of environments and even within the vegetation cover.

Termites are so widespread since they have a wide range of feeding behavior. When living in symbiosis with different microorganisms, they are able to digest totally stable organic material such as lignin and cellulose (Vargo and Husseneder, 2009). More than 2900 species have been identified (Parman and Vargo, 2008) and their sizes range from 2 to 20 mm. They mostly live in state-like organizations with up to 3 million individuals, and build nests and mounds that greatly influence physical, chemical and biological soil properties. By preferentially transporting particles and aggregates of specific sizes in horizontal as well as vertical directions, they contribute to bioturbation, sorting and other soil formation processes. This cannot be underestimated in relation to soil development because of their activity rates over long time periods.



Fig. 4 Termites cover litter with soil aggregates to prepare it for consumption, hidden from predators. Surface of an Acrisol near Bhubaneswar (India) Photograph by Stahr.

An obvious effect of *Macrotermes* species, for example, is to cover dead plants with soil material (Fig. 4). Together with the plants they are protected from predators, and plant material can be transported later to their mounds for further digestion and consumption. Roughly, we can estimate that the belowground volume of a termite mound as depicted in Fig. 5 is at least as big as the aboveground volume. Termite mounds are nutrient accumulation sites since plant material is actively incorporated. The changes in soil properties and structure can still be detected decades after the mounds have been abandoned.

In particular, aboveground as well as belowground mound building termites change the soil hydrological properties. Apart from the mound itself, termites can build galleries that reach down to where water can be collected, i.e., several meters, and they can construct larger hollows for intermittent storage (Fig. 6). All these aspects underline the importance of soil organisms for tropical soil morphology, genesis and properties.

Lateritic weathering as central concept for the tropical soil and landscape development?

Laterite is derived from the Latin word for “later,” which means brick. The term was first used by Francis Buchanan in 1807 (cited in Aleva, 1994) when he described field observations in Madras: “*the most proper English name would be Laterite, from lateritis, the appellation, which may be given to it in science.*” Laterite, with the meaning of a hard outcrop (often seen as rock), is often observed in the tropics at the surface due to erosion. It is perceived as a product of intensive long-term weathering of diverse parent rocks.



Fig. 5 Termite mound incorporating a living tree. Mound surrounded by silt material eroded from the mound (Central Somalia; Luvisol). Photograph by Stahr.



Fig. 6 Acrisol with termite caves in the E-Horizon (Bhubaneswar, India) Photograph by Stahr.

Under these conditions most minerals weather and some more easily soluble elements are depleted such as sodium (Na), potassium (K), calcium (Ca), magnesium (Mg) and silicon (Si). This often results in a residual accumulation of iron (Fe) and aluminum (Al), which is known as ferralitization. Hereby residual quartz and new formations of kaolinite, goethite, hematite, maghemite, gibbsite and other pedogenic Al-oxides can be enriched (Aleva, 1994).

There is a wide range of observations, which have been called laterite (Aleva, 1994). Among these have been rocks, land surfaces and soils. As a result, the definition of laterite became too ambiguous. Consequently, the terms “laterite” and “latosol” disappeared in modern soil systematic considerations and have been replaced by more specific ones.

A typical sequence of horizons has been described in different continents such as Asia (India), Africa (Nigeria) and America (Brazil) (compare Figs. 7 and 8). The complete sequence starts with a Haplic Ferralsol (according to the World Reference Base for Soil Resources, FAO, 2015), or Oxisol (Buol and Eswaran, 2000). This upper part of the sequence is often disturbed by erosion, translocation and sedimentation. The second part is enriched in iron and residual aluminum. It often hardens irreversibly upon drying, and is then called plinthite (“ironstone” Figs. 7 to 9).

In the plinthite and in the following mottled horizon (Fig. 8c) residual biostructures are often observed. The mottled horizon has pale and red parts, which are evenly distributed. The iron content is much lower than that of the horizon above. Below the mottled horizon, there is a bleached zone (Figs. 7 and 8a). Here iron is merely absent and there are residual aluminum compounds and quartz. The enrichment of aluminum may be sufficiently strong that it can be used as an ore. It is then known as bauxite, namely derived from the occurrence and former mining at Les Beaux-de-Provence in southern France (Valeton, 1972; Fig. 9). Saprolite follows further down the profile. This is the first zone, where the structure of the parent rock is preserved, but the material becomes soft. Silicates, however, show isovolumetric weathering and can often be totally transformed into kaolinite. Sometimes underneath this layer there is an abrupt boundary at which the unweathered parent rock appears. Each of the zones described has a thickness of several meters, and therefore the sequence has a total thickness of 10–50 m.

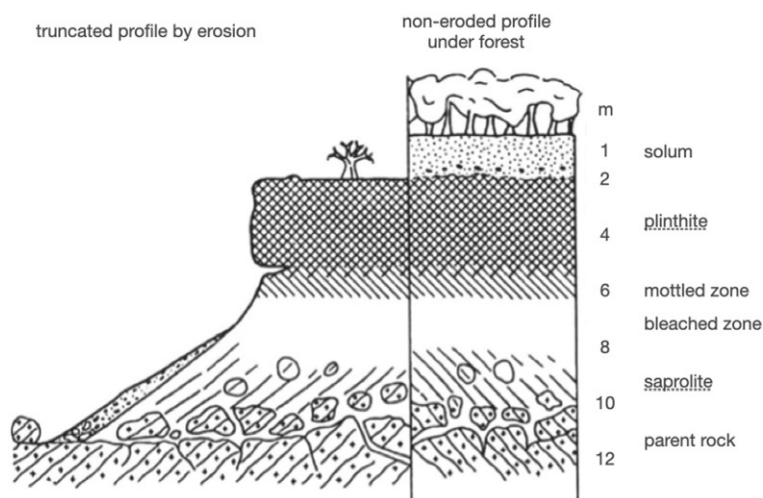


Fig. 7 Idealized laterite profile found under a seasonal tropical climate. After Ollier, C.D. and Sheth, H.C. (2008) The High Deccan duricrusts of India and their significance for the laterite issue. *Journal of Earth System Science* 117 (5): 537–551.

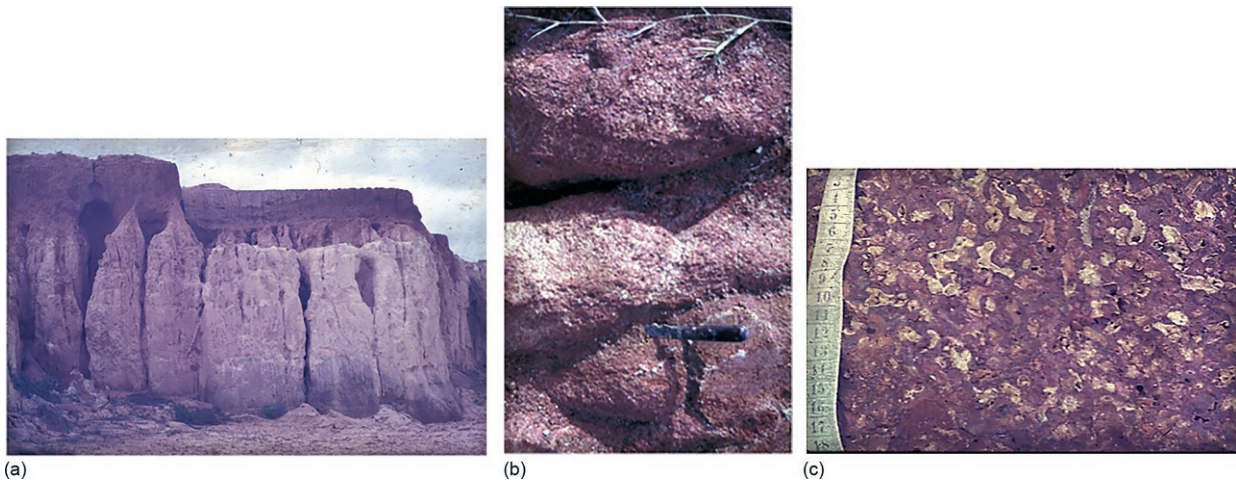


Fig. 8 (a): Photograph of a characteristic lateritic weathering sequence at Bangalore, Southeast India (ca. 30 m). Upper soil eroded, plinthite, mottled zone, bleached (bauxitic) zone, saprolitic zone, parent rock not exposed. (b): Typical plinthitic ironstone Central Benin. (c): Plinthite with many hardened tubes of an ancient termite mound in India near Bhubaneswar. (a) From Blume, H.-P. and Röper, H.-P. (1974) Zur Genese lateritischer Roterden. *Mitteilungen der Deutschen Bodenkundlichen Gesellschaft* 20: 80–86, Göttingen. (b) and (c) Photographs by K. Stahr.

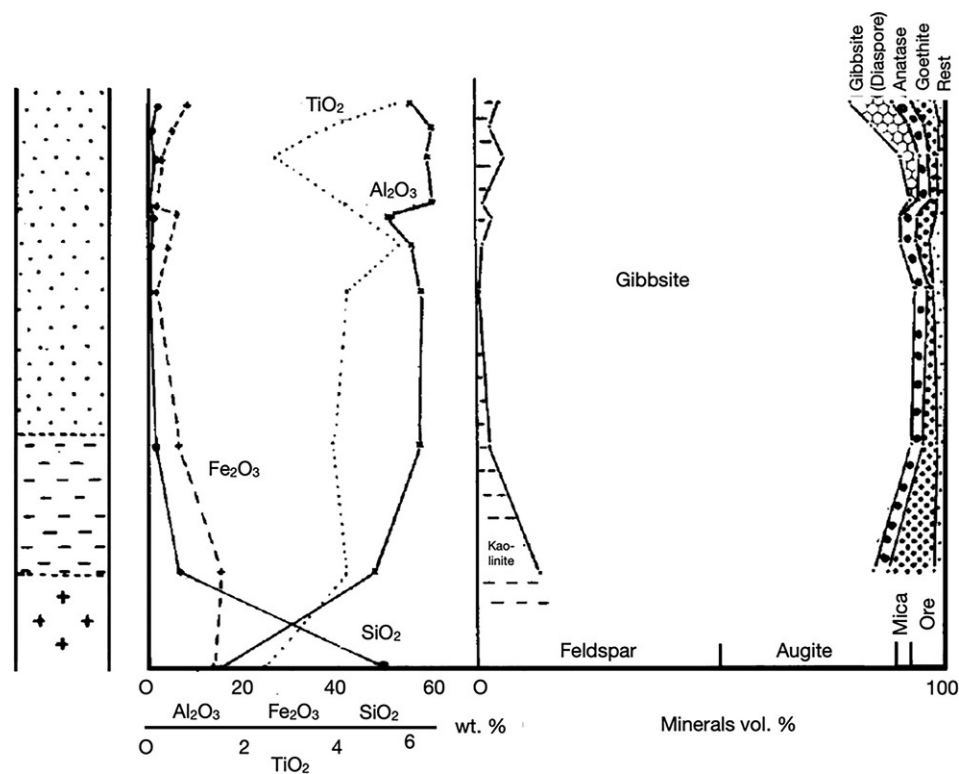


Fig. 9 Vertical distribution of elements and minerals in a lateritic weathering profile in southern India. Modified after Valeton I (1972) *Bauxites*. Amsterdam: Elsevier.

Leaching of bases and desilification are prominent soil forming processes that are observed throughout. The rates of residual accumulation depend on the different weathering zones.

With respect to residual minerals, they concentrate abundantly like quartz and others such as anatase, titanite, ilmenite, staurolite, tourmaline, and zircon. Further, more typical neoformed minerals are present: kaolinite dominates, and in addition gibbsite, goethite, hematite, maghemite and also secondary quartz occur (compare Fig. 9). In particular, Si, Fe, and Al show distinct depth trends. While Fe and Si are completely lost in the bleached zone, Al is residually enriched. Silica is lost by lateral transport from the landscape. Some of the iron can accumulate in the upper parts of the soil profile (plinthite, mottled zone). Together with

the iron, some “heavy” elements may be enriched, either because they are also redox sensitive or because they are incorporated into minerals that hardly weather and enrich residually.

Iron minerals such as goethite, hematite and maghemite are all neoformed. Through relative and absolute enrichment two types of ores are formed (Schellmann, 1994). In the plinthite zone we can find lateritic or plinthitic iron ores (Figs. 8a–c and 9), while in the bleached zone bauxitic aluminum ores (Figs. 8a and 9) often occur.

The frequent occurrence of lateritic weathering profiles under different environmental conditions has led to different hypotheses about their formation. The most important ones are presented in the following:

- (1) The one that is probably the most supported is the residual accumulation hypothesis: This is based on the idea that due to permanent hydrolytic weathering and leaching an absolute loss of mainly K, Na, Mg, Ca, and Si occurs. The result is a residual enrichment of elements and stable minerals, i.e., Al and Fe in oxides, Ti (titanium) in rutile, and Zr (zirconium) in zircon. This approach leaves the complex horizonation unexplained as well as the differences in the Fe to Al ratio.
- (2) The second one is the groundwater hypothesis: During long periods iron was transported downslope into depressions and became enriched in soils where the groundwater table fluctuated (Bl horizon). Below this the groundwater is permanently present (Br horizon) and leads to a reduced bleached zone. The laterally moving groundwater exports elements out of the landscape. Later after (relative) uplift of the landscape long lasting erosion processes led to subsidence of the surrounding landscape and incision by the rivers. In this way, the iron-enriched plinthite became aerated and transformed into petroplinthite. This process conserved the former Bl horizons. The continuous lateral erosion finally resulted in an inversion of the relief, i.e., the former lowlands now form the plateaux.
- (3) According to the Stagnosol–Gleysol hypothesis, three phases occur. In the first phase mottles and or concretions are formed due to water stagnation. In a second phase, alternating groundwater levels occur that increase over time and depth, and lead to deep soil development and enrichment of iron in the concretionary horizon. Finally, the groundwater table continues to lower.
- (4) Following the slope gley hypothesis, in the central part of a plateau water stagnation occurs. There, iron is mobilized and transported laterally to the fringes. There it becomes oxidized and is precipitated. The typical element distribution, as shown in Fig. 9, cannot be totally explained by this process description.
- (5) The floating ore hypothesis is based on the observation that in some plinthites rock fragments have been observed. Understanding of the process is as follows: first phase, ferrallitic soil development results in a relative enrichment of Fe and Al; second phase, the ferrallitic material is eroded and transported along the slope. It is deposited down the slope and then cemented. In fact, in many landscapes of western Africa several laterite generations at different elevations can be observed within one landscape, supporting this hypothesis. However, the original occurrence at the highest landscape position cannot be explained in such a way.
- (6) Finally, we have the karst hypothesis. This relates to the fact that plinthite can be observed in limestone areas (e.g., at Les Beaux-de-Provence, southern France). The concept is that decalcified material accumulated in large karstic hollows. Then, following desilification leads to the enrichment of Fe and Al. Understanding of this process more or less follows the residual accumulation hypothesis, but related to a specific parent material only.

In conclusion, lateritic weathering profiles cannot be explained fully by one of the above hypotheses alone. A combination of the first two is the most probable; it is supported by the data and appears reasonable given the long time spans necessary for the genesis of laterite during which climate fluctuations definitely occurred.

Soils, processes and landscapes

Humid zone river soils (Fluvisol, Gleysol, Histosol, Podzol, Terra preta; India, Congo, Brazil)

Note: In the following we use the World Reference Base for Soil Resources (FAO, 2015) terminology in the text. Where useful, Soil Taxonomy (Soil Survey Staff, 1999) terms are given in brackets.

The major river basins in the tropics include the Amazon, Orinoco, and Paraguay river basins in South America, and the Congo basin in Africa (Monger, 2005). The basins are composed of sedimentary materials derived from surrounding uplands and mountains. The western part of the Amazon basin and the central part of the Congo basin have udic and perudic SMRs and an Af-climate, while the eastern portion of the Amazon basin grades into an ustic SMR. Ferralsols (Oxisols) predominate in the Congo and eastern Amazon basin. The sediments of the latter derive from the Guyana and Brazilian shields.

Fluvisols and Gleysols are widespread in the lowlands (Fig. 10). In particular, settlements are traditionally found on Fluvisols, because drinking water occurs close to the surface, and the rivers provide transport arteries and offer protein-rich food, i.e., fish. Sustainable horticulture and fruit cultivation can be practiced here because water in the root zone rarely runs short, and the steady water flow delivers nutrients. Gleysols and Histosols are less favorable for arable farming, except for crops such as paddy rice (*Oryza sativa* L.) and sugar cane (*Saccharum officinarum* L.) that grow well in marshy areas.

Nitisols and Acrisols (Ultisols) predominate in the western part of the Amazon basin, where the sediments derive from the Andes and are consequently slightly more nutrient-rich than those towards the east. Both Acrisols and Ferralsols are by definition nutrient-poor soils. However, local inclusions of more nutrient-rich parent materials are present. Luvisols and Eutric Nitisols (Alfisols) have formed in these more base-rich sediments of western Brazil and on the eastern borders of the Congo basin. Many areas with these more fertile soils in the chemical sense, are being cleared of forest for crop production and cattle grazing.

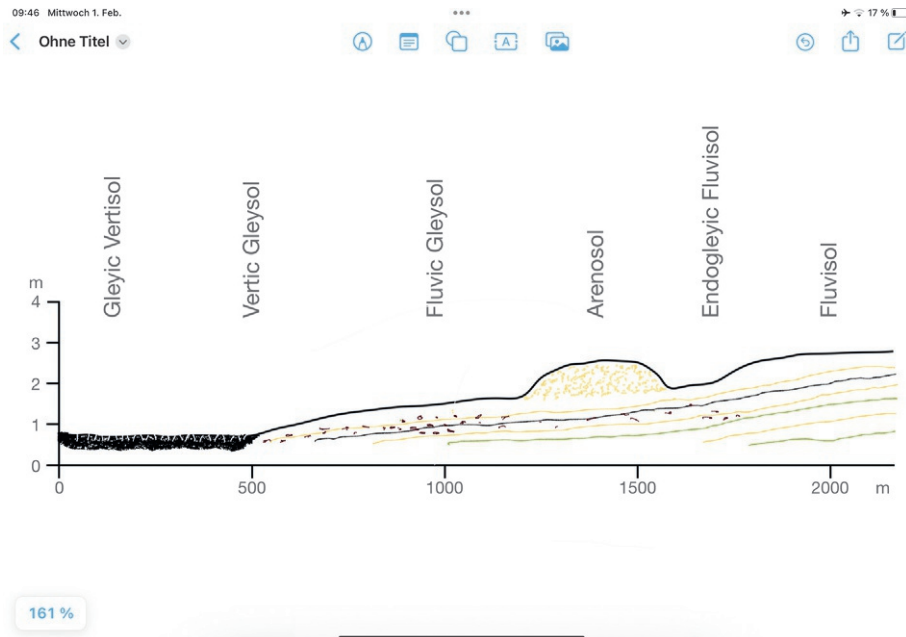


Fig. 10 Cross section through a floodplain at Rio Babahoyo in the lowlands of Western Ecuador. Graph by Stahr and Herrmann.

Some of the sediments in the basins are sandy. Where the water table occurs at depth, soils formed from such sediments are chemically infertile Arenosols (Quartzipsamments) or (giant) Podzols (Spodosols). It has been estimated (Klinge, 2006) that about 10% of the Amazon basin is covered by these kinds of soils, e.g., in the headwaters of the Rio Negro in the northwestern Amazon basin, where it merges with the Orinoco watershed. Histosols and Gleysols are also widespread there. Smaller areas of these sandy soils are scattered throughout the basins, often detectable by coffee-colored “black” waters of the rivers that drain these watersheds. So-called white sands ecosystems have recently captured more interest in the marshy areas (Klinge, 2006; Bleackley and Khan, 1963; Atwell et al., 2023).

The Terra preta story

Many tropical soils are considered to be degraded, though the definition of what degradation in this context means is seldom defined. In particular, in subsistence-oriented farming systems, fertilizer input amounts are small and fallow times are decreasing. This eventually leads to a decrease in soil organic matter and plant available nutrient concentrations.

To increase soil fertility, much attention has been paid to what is called “Terra preta.” These anthropogenic soils (WinklerPrins, 2014) were first discovered in South America; they have a deep topsoil enriched in organic matter with a greater share of so-called black carbon. The latter is an organic matter fraction that is rather resistant to microbial decay (Glaser and Woods, 2004). The Terra preta (Pretic Anthrosol) concept has much in common with the Hortisols or Plaggenesch (Hortic or Plaggic Anthrosols) in northern Germany and the Netherlands.

The major question is whether Terra preta is, in general, a suitable concept to increase the chemical fertility status of tropical mineral soils. To answer this question one has first to acknowledge that Terra preta are sites of accumulation; they cannot develop from the biomass production at the sites themselves, but depend on the input of organic matter from other sites. This means that in areas with a high population density, and where all surfaces are already under agricultural use, sufficient biomass as a source of input is rarely available. This is particularly true for the Sahelian Arenosol belt south of the Sahara desert.

Where sufficient external biomass is available, the question arises as to whether it would be better to use the biomass for soil C and thus nutrient cycling rather than increasing the C-stock. Increasing the C-stock makes particularly good sense in soils with low cation exchange capacity, e.g., Arenosols. But as described above, in the areas where the latter are prominent, the external biomass available is insufficient to build up Terra preta. And if one looks at the rates of application in scientific publications, the inputs are far from what is manageable in a subsistence context (compare also Feil et al., 1995).

An additional problem is the effort required to transform biomass, into compost or Bokashi, (inoculated fermented compost), for example, or the application of pyrolysis. The latter needs additional technology that is costly and may lead to byproducts that are potentially harmful, depending on the process management.

In conclusion, application of the Terra preta concept will, in the long run, be restricted to small areas where topsoil cation exchange capacity and organic matter content are low, where sufficient biomass waste products are available from other sources, and sufficient labor is available under the condition of low alternative income possibilities (“opportunity costs”).

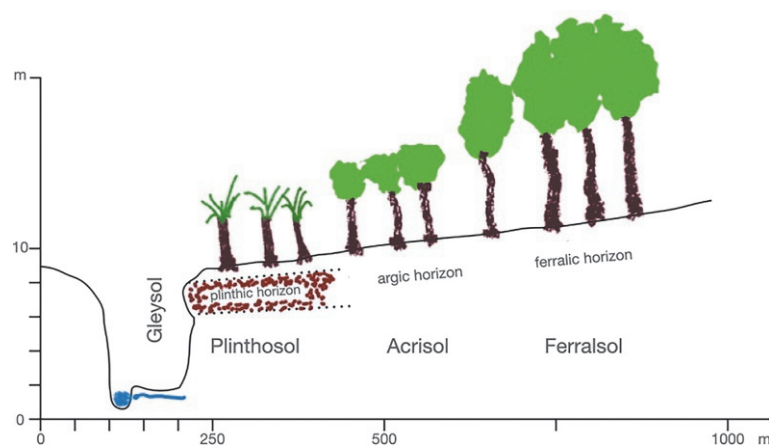


Fig. 11 Cross section through a soilscape close to Rio Preto in the southern Amazon basin of Brazil. Redrawn after Blume H-P, Brümmer GW, Fleige H, Horn R, Kandeler E, Kögel-Knabner I, Kretschmar R, Stahr K, Wilke B-M (2016) Scheffer/Schachtschabel soil science. pp. 399–402. Chap. 8.4 Soil Zones of the World. Springer-Verlag Berlin Heidelberg.

Humid tropical uplands (Ferralsol, Plinthosol, Planosol; Brazil, Benin, Thailand)

The remains of the old continents are dominated by deeply weathered ferralitic laterite profiles (Figs. 7 and 8a). The river systems generally cut into the saprolite or even into the unweathered rock. They also have udic and perudic SMRs and an Af-climate and grade at their outside into ustic SMRs, and the influence of groundwater is reduced in comparison to river soilscares above. The plateaux are covered with strongly and deeply weathered Ferralsols (Fig. 11). At the beginning of the slopes they can change to Acrisols or Plinthosols. The latter often form from the hardened plinthite, a morphologically visible step. Beneath the ferralitic material can be found again. In contrast to the Rhodic Ferralsols on top of the weathering sequence these often show xanthic (yellow or golden) color patterns or geric properties. Gleysols occur where streams flow in narrow floodplains. Ferralic Arenosols or even Podzols occur where the parent material is Palaeozoic sandstone (Blume et al., 2016).

Landscapes derived from “acid igneous rock” occur more extensively in Southeast Asia, parts of the eastern highlands in South America, West Africa, the central highlands of East Africa, and eastern India. In these progressively weathered areas, there are Ferralsols on hilltops and in plateau positions, together with Planosols in micro-topographic depressions. Plinthosols with partly petro-plinthic horizons occur on slopes and in slightly eroded places. These topographic positions indicate the influence of lateral transport processes as well as erosion that trigger hardening of the plinthite.

All the soils mentioned above are difficult to use for agriculture because of their poor nutrient status resulting from the intense weathering and leaching. In places, where landscapes were less stable, and consequently soil formation is younger or results from finer-grained basement rocks Acrisols and Dystric Cambisols are prominent, which have a moderate fertility status, but are still commonly deficient in phosphorus (P) and calcium (Ca). In most areas there are inclusions of more base-rich sedimentary rock where Luvisols have formed. In former times agriculture in these areas was limited to slash and burn unless the infrastructure for markets and fertilizers existed (Monger, 2005).

Under certain conditions, however, nutrient status can be high (Singh et al., 1995). On the Deccan plateau of western India, for example, chemically fertile Vertisols have formed on volcanic, i.e., basaltic rocks (Bronger, 2007). The Andean chain along the west coast of South America was formed by tectonic uplift and associated volcanic activity. Some of the uplifted materials are limestones, rich in calcium, while others are granite and schist, relatively poor in calcium but rich in potassium. Tectonic instability, steep slopes, and glaciers at the highest elevations have resulted in rapid erosion of surface materials and deposits of relatively fertile sediments in intermountain valleys where fertile Eutric Cambisols, Regosols and Leptosols occur, but are partly limited by water availability.

Seasonal river soilscares (Fluvisol, Cambisol, Vertisol, Gleyic Solonchak; Amazon, Congo, Niger)

Soils formed in river floodplain sediments and deltas are often fertile (Monger, 2005) either due to their mineral and geochemical composition or they obtain nutrients that are delivered steadily by the flowing groundwater. In the first case, fertility is derived from eroded (top)soil of the surrounding uplands. Consequently, where the soil is relatively rich in essential elements, people can grow and harvest food with little or no attention to the addition of mineral fertilizer. Crop yields are low to medium inter alia depending on the depth to the groundwater, but they are reliable. The soil moisture regime is still udic, the climate, however, is Aw and can have dry periods. Most soils formed on floodplains are Fluvisols, when flooding is still frequent, or Cambisols in more elevated positions. Clay rich sediments are the parent material of widespread Vertisols (Hole, 1961). In addition, Gleysols occur where groundwater appears at a shallow depth. Despite the physical dangers from flooding, human settlements are concentrated in these areas, attracted by the relatively elevated soil fertility, access to water, transport, and protein sources, i.e., fish. Subsistence agriculture

is often practiced, unless flood hazards and groundwater table can be controlled. Narrow areas of floodplains are also present along coastlines throughout the humid tropics with similar types of soil occurring. In the Amazon floodplains long-term used soil and the organic matter-enriched Terra preta occur around traditional villages (see Terra preta section above). They are Anthrosols that have seen intended long-term accumulation of organic inputs.

Sub-humid tropical uplands (Acrisol, Luvisol, Nitisol, Cambisol, Planosol; Benin, Niger, Ghana)

The sub-humid seasonal tropics exhibit a broad range of different soils depending on rock type, relief and particularly the age of the stable land surface (Blume et al., 2016). Strongly weathered and kaolinite-rich Ferric Acrisols and Lixisols occur on the older land surfaces in West Africa (Fritz, 1996), Eastern Brazil and Northern Borneo. Plinthosols are also present there. Younger landscapes with more active morphodynamics (SE Asia), and landscapes with longer dry periods are covered with Nitisols and Ferric Planosols, whereas depressions have gleyic and stagnic variants. In these areas the moisture regime is generally ustic and temperature regime is hyperthermic. The Köppen climate is Aw bordering on the dry side to BS.

Vertisols dominate in areas derived from clay or mudstone with strongly alternating wetting and drying cycles, i.e., pronounced seasonality (Thakur et al., 1999). Consequently, Vertisols can occur at the drier end of the seasonal climates, i.e., in semi-arid landscapes, e.g., in Venezuela and India, and Ethiopia (Bronger, 2007, Murthy et al., 1982). Parent materials with lower clay contents lead to the formation of Lixisols in drier climates, and Nitisols in more humid climates (Fig. 12). Under terrestrial conditions Vertic Cambisols and Chromic Luvisols occur, and on relatively old surfaces Plinthosols are widespread. In Asia, the coastal regions especially are characterized by Gleysols and Fluvisols that have been modified by long-term rice cultivation (Anthrosols).

Plinthosols and Acrisols have very poor nutrient stocks and are therefore often under shifting cultivation in South America and West Africa (e.g., Igue, 2000). Alternatively, they can be used for Al-tolerant perennial crops such as tea (*Camellia sinensis* (L.) Kuntze), rubber (*Hevea brasiliensis* (A.Juss.) Muell. Arg.) or oil palm (*Elaeis guineensis* Jacq.). In contrast, Lixisols and particularly Nitisols can be highly productive if adapted nutrient management is practiced (Bronger and Bruhn, 1989).

Water erosion is a general problem associated with arable farming and particularly in seasonal climates where high rain intensities occur. Despite poor workability, either too dry or too wet, Vertisols are in general nutrient rich (Blockhuis, 1993) due to their high cation exchange capacity and their frequent clay–humus complexes. As fertile sites, they are traditionally used for agriculture, if not too wet for prolonged periods. In drier areas additional irrigation can be favorable, otherwise they also provide good pasture sites.

Young volcanic territories (Leptosol, Regosol, Andosol, Phaeozem; Ecuador, Ethiopia, Indonesia)

A small proportion of the soils in humid tropical areas are formed in relatively nutrient-rich material derived from volcanic activity. A typical representative of weathering from volcanic ash and basaltic materials of various ages are Andosols (Andisols). In a chronosequence, Tephric Regosols are precursors, and Luvisols or Acrisols follow Andosols. Historically, these soils have supported large human settlements. Java, southern Sumatra, parts of the Philippines and other Pacific islands, Central America, the Andean mountains in South America and the Rift Valley in Africa are major areas with volcanic material. Many of these volcanic areas are mountainous and the higher elevations are characterized by isothermic or isomesic STRs where crop performance is reliable, but cool temperatures can slow down crop growth at higher elevations. The volcanic ash regions are naturally the most fertile (Fig. 13). Ash compounds weather easily and therefore deliver nutrients in large amounts to the developing soil. In addition, these soils have low bulk density and therefore good rooting capabilities, and their structure is quite stable enabling high rates of water infiltration. The converse is phosphorus fixation, and in the case of relatively mono-mineral composition there are specific nutrient deficiencies.

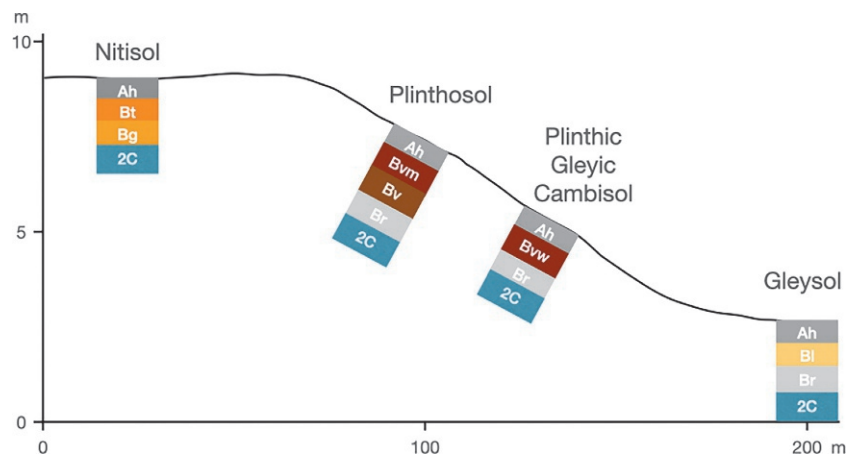
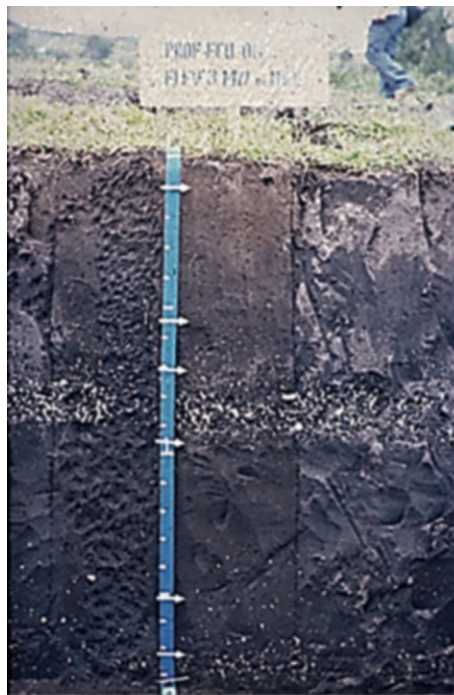


Fig. 12 Cross section of a soilscape at the WARDA research center, Ivory Coast, Africa (redrawn from excursion guide).



(a)



(b)

Fig. 13 (a) Tephric Andosol in the Andies of Ecuador ca. 3000 m a.s.l. under grassland. (b) Silandic Andosol under palm trees in western Ecuador on the footslope of the Andies. Photographs Stahr, 1984.

Human influences on soil fertility

The predominant soil limitation in humid tropical regions is low chemical fertility (Monger, 2005). Major soils are formed from parent materials that contain very limited quantities of essential nutrients such as phosphorus, potassium, calcium, or magnesium. The most infertile soils are the uplands in the interiors of Africa and South America. Most of these soils are physically deep and have reliable moisture for one or two harvests per year. The most infertile ones do not even support enough natural

biomass to sustain slash-and-burn agriculture. However, where reliable infrastructure has been developed, sustained commercial agriculture can now be practiced. Reliable markets, fertilizer, lime, and fuel supplies are essential. A sufficient amount of phosphorus must be applied and mixed into the topsoil to saturate the iron and aluminum oxide anion exchanger sites to an extent that sufficient phosphorus can be made available for the crops. Coarse-grained carbonates or gypsum need to be applied to raise pH and or inactivate the extractable aluminum. Nitrogen fertilizer is always required for non-legume crops. Potassium becomes a limiting factor if high yields are envisaged. Micronutrients such as copper, zinc, boron, and molybdenum are needed in certain places and for specific crops. After an initial investment is made to alter the less favorable chemical conditions, annual fertilizer requirements are greater than in other grain-growing areas around the world. Modern soil-testing technology should be applied to avoid blanket recommendations for fertilizer application rates.

Somewhat less infertile soils with dense forest vegetation can be and are being used for slash-and-burn subsistence agriculture. The rate at which essential nutrients are released from the minerals is too slow to replenish those removed in rapidly growing food crops. The slow-growing forest vegetation that regenerates in abandoned fields can accumulate enough nutrients in its biomass that—when cut and burned—one or two seasonal food crops can be grown, but only every 10–30 years. These soils have the potential for more intense food crop production if chemically fertilized, which is seldom possible without a market infrastructure.

Human populations in the humid tropics are concentrated on soils formed from materials relatively rich in phosphorus, calcium, magnesium, and potassium. Such soils are primarily located in floodplains and deltas, in areas of volcanic activity, and in upland areas underlain by limestone, base-rich igneous, and metamorphic rocks.

Importing essential elements to compensate for those exported with human food can increase crop yields from all soils. Applying animal and human wastes wherein the essential elements have been gathered from sites distant from the food-growing area is a partial solution in many cultures. However, the concentration of essential elements in organic residue and waste materials is small, transport is difficult, and the rate at which the residues decompose to release inorganic ions for crop uptake is often too low. The most efficient fertilizer strategy is the one that combines organic and mineral fertilization.

Outlook and research questions

We have to differentiate between the two great topics of soil genesis and adapted soil use. With respect to soil genesis and the distribution of soils the large divisions are clear, but detailed processes and their rates are still vague. In particular, the processes of desilification and plinthite formation are still poorly understood and require more attention.

In the humid tropics, the typical conditions provide a growing period and ongoing soil formation that lasts the whole year. Both are reduced in the subhumid tropics of the savanna zone, but there is still a long growing period and strong soil formation, which may have lasted for millions of years. The dynamics of soil formation do not cease because of the consistently high temperatures throughout the year. One of the key processes is the loss of silica. In general, silica is removed from the landscapes and is found in the oceans. On the other hand desilification is a process that needs to be described better in young and old landscapes.

Young soils such as Regosols mostly occur on young or strongly eroded land surfaces. On the other hand, there are old soils, which seem to have a typical tropical history like Ferralsols. There are also specific types of soils that might be not related to the actual processes and relief situation. They often occur in landscapes that have undergone lateritic transformation like Plinthosols, deep Ferralsols and even Cambisols formed from deep and old weathered saprolites in one landscape. To explain such occurrences we have to search for further explanations in the future.

Nutrient recycling and replenishment are crucial for the agricultural use of tropical soils. In many cases, increasing organic matter stock is seen as the solution. However, the question remains as to where these materials can be resourced. In most cases the biomass produced has many different uses, particularly for subsistence farmers, so they are unwilling to leave them on the field. Therefore first, biomass production needs to be increased, which is hardly imaginable without primary mineral fertilizer input. Second, the efficiency of nutrients in the system needs to be increased. Here, placing rather than broadcasting of fertilizer is a solution, in particular for crops with a larger sowing distance. New fertilizer resources need to be exploited, including human urine and feces to increase the rate of recycling and sustainable land use. And finally, the nutrients transported to the cities from agricultural products need to be returned to the farms.

References

- Aleva GJJ (ed.) (1994) *Laterites. Concepts, Geology, Morphology and Chemistry*. Wageningen, Netherlands: ISRIC. ISBN: 90-6672-053-0. 165 pp.
- Atwell MA, Wuddiviriy MN, Fiedler S, Oatham M, Herrmann L, Glasner B, Vetter V, and Jungkunst HF (2023) Influence of soil geomorphic factors on vegetation patterns in a model white sands ecosystem complex. *Catena* 225: 107044.
- Bleackley D and Khan EJA (1963) Observations on the white sand areas of the Berbice formation, British Guiana. *European Journal of Soil Science* 14(1): 44–51.
- Blockhuis WA (1993) *Vertisols in the Central clay plain of the Sudan*. Dissertation Wageningen: Agricultural University.
- Blume H-P, Brümmer GW, Fleige H, Horn R, Kandeler E, Kögel-Knabner I, Kretschmar R, Stahr K, and Wilke B-M (2016) Scheffer/Schachtschabel soil science. In: *Soil Zones of the World*, pp. 399–402. Berlin Heidelberg: Springer-Verlag. Chap. 8.4.
- Bronger A (2007) Bodengeographie der wechselfeuchten Tropen am Beispiel Indiens. Kap.3.4.5.6. (1996 ff.) In: Blume, et al. (ed.) *Handbuch der Bodenkunde*. Weinheim: Wiley-VCH.
- Bronger A and Bruhn N (1989) Relict and recent features in tropical Alfisols from South India. In: Bronger A and Catt J (eds.) *Paleopedology—Nature and Application of Paleosols*, *Catena Suppl.* vol. 16, 107–128.
- Buol SW and Eswaran H (2000) Oxisols. *Advances in Agronomy* 68: 151–195.

- FAO (2015) World reference base for soil resources 2014, update 2015. In: *World Soil Resources Report 106*. Rome: FAO.
- Feil PR, Lamers JPA, and Herrmann L (1995) Knowledge transfer in the field: Solving crop residue problems in Niger. *The Journal of Technology Transfer* 20: 31–41.
- Fritz C (1996) Boden- und Standortmuster in geomorphen Einheiten Süd-Benins (Westafrika). *Hohenheimer Bodenkundliche Hefte*. vol. 29. Stuttgart: Univ. Hohenheim.
- Glaser B and Woods WI (2004) *Amazonian Dark Earths—Explorations in Space and Time*. Berlin: Springer.
- Hole FD (1961) A classification of pedoturbation and some other processes and factors of soil formation in relation to isotropism and anisotropism. *Soil Science* 91: 375–377.
- Igue AM (2000) The use of a soil and terrain data base for land evaluation procedures—Case study of Central Benin. *Hohenheimer Bodenkundliche Hefte*. vol. 58. Stuttgart: Univ. Hohenheim.
- IUSS Working Group WRB (2022) *World Reference Base for Soil Resources. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*, 4th edn. Vienna, Austria: International Union of Soil Sciences (IUSS).
- Kew GA, Gilkes R, and Mathison C (2008) Nature and origins of granitic regolith in bauxite mine floors in the Darling Range, Western Australia. *Australian Journal of Earth Sciences* 55(4): 473–492.
- Klinge H (2006) Podzol soils in Amazon basin. *European Journal of Soil Science* 16(1): 95–103.
- Köppen W (1936) Das Geographische System der Klimate. In: Köppen W and Geiger R (eds.) *Handbuch der Klimatologie*. vol. I, Part C. Berlin: Gebrüder Borntraeger.
- Mohr E, van Baren F, and van Schuylenborgh J (1972) *Tropical Soils*, 3rd edn. Den Haag: Mouton.
- Monger HC (2005) Tropical soils—Arid and semiarid. In: *Encyclopedia of Soils in the Environment*, 1st edn., pp. 182–187. Amsterdam: Elsevier.
- Murthy RS, Hirekerur LR, Deshpande SB, and Venkata Rao BV (1982) Benchmark soils of India. In: *Morphology, Characteristics and classification for Resource Management*. Nagpur: Nat. Bur. of Soil Survey and Land Use Plan. (ICAR).
- Parman V and Vargo EL (2008) Population density, species abundance, and breeding structure of subterranean termite colonies in and around infested houses in central North Carolina. *Journal of Economic Entomology* 101(4): 1349–1359.
- Schellmann W (1994) Geochemical differentiation in laterite and bauxite formation. *Catena* 21: 131–143.
- Singh SK, Das K, Shyampura RL, Giri JD, Singh RS, and Seghal JL (1995) Genesis and taxonomy of black soils from basalt and basalt alluvium in Rajasthan. *Journal of the Indian Society of Soil Science* 43: 430–436.
- Soil Survey Staff (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn. Washington, DC: USDA–NRCS. Agricultural Handbook No. 436.
- Thakur DS, Bapat PN, Dubey DD, and Gupta GP (1999) Clay properties, mineral stability and mineralogy of Vertisols in central India. *Journal of the Indian Society of Soil Science* 47: 781–788.
- Valeton I (1972) *Bauxites*. Amsterdam: Elsevier.
- Vargo EL and Husseneder C (2009) Biology of subterranean termites: Insights from molecular studies of Reticulitermes and Coptotermes. *Annual Review of Entomology* 54: 379–403.
- WinklerPrins AMGA (2014) Terra Preta. In: *The Soil Underfoot: Infinite Possibilities for a Finite Resource*. CRC Press. 235 p.

Sodic soils

Guy J Levy and Uri Nachshon, ARO, Volcani Institute, Rishon Lezion, Israel

© 2023 Elsevier Ltd. All rights reserved.

Introduction	369
Definitions of sodic soils	369
Origin and distribution of sodic soils	370
Mechanisms responsible for sodic behavior	371
Clay swelling	371
Clay dispersion	371
Structural stability of sodic soils	372
Sodicity and aggregate stability	372
Role of organic matter	373
Water flow in sodic soils	373
Hydraulic conductivity (HC)	373
Sodicity and electrolyte concentration	373
Soil properties affecting sodic behavior	374
Sodicity and conditions prevailing in the soil	375
Infiltration rate	376
Soil erosion	377
Sodicity and soil productivity	377
Crop tolerance to sodicity	377
Plant nutrition in sodic soils	378
Reclamation and management of sodic soils	378
Chemical amelioration	378
Chemical amendments	378
Successive dilutions of high salt water	378
Amendment with organic wastes	379
Tillage	379
Remediation with vegetation	379
Environmental aspects of sodic soils	380
Erosion and suspended sediments	380
Salination and pollutants	381
Carbon (C) sequestration	381
Concluding comments	381
References	381

Abstract

Sodic soils are of major concern, especially in semi-arid and arid regions. Soil sodicity results from both natural and anthropogenic processes. Many studies have focused on the adverse effects, that sodic soils have on soil degradation, soil productivity, and associated harmful effects on the agricultural and natural environments. Here, basic processes related to soil sodicity are presented, and their impact on soil physical and hydraulic properties and productivity is discussed. The chapter also presents key insights regarding common practices used to cope with soil sodicity and the reclamation of sodic soils.

Key points

- Soil sodification is a major cause of soil degradation, especially in semi-arid and arid regions.
- Soil sodicity results from both natural and anthropogenic processes.
- Clay swelling and dispersion are the main mechanisms that control sodic behavior in soils.
- The magnitude of sodic behavior depends closely on the electrolyte concentration of the soil solution.
- Sodic soils exhibit poor soil structure and hydraulic properties.

Introduction

The presence of sodium (Na) as an adsorbed cation on the soil exchange phase is found in many soils worldwide. Sodium-affected soils exhibit poor soil water and air ratios which adversely affect water movement in the soil, root growth and plant production, and make the soil difficult to farm when either dry or wet. Consequently, in many cases, such soils have had to be abandoned.

Sodium-affected soils have been referred to in the past as alkali soils. This may have originated from the traditional separation of sodium-affected soils into two groups (Szabolcs, 1979): (1) saline soils (soils affected by neutral Na salts, mainly NaCl (sodium chloride) and Na₂SO₄ (sodium sulfate)); and (2) alkali soils (soils affected by Na salts capable of alkaline hydrolysis, including NaHCO₃ (sodium bicarbonate), Na₂CO₃ (sodium carbonate) and Na₂SiO₃ (sodium silicate)). However, at present, Na-affected soils are referred to as sodic soils, with the realization that Na concentration in the soil solution and its amount on the exchange complex, in conjunction with the total electrolyte concentration, determines whether or not a given soil is sodic, and not the type of the salt.

Problems associated with sodic soils are expected to increase. That is because of the need to provide food to an ever expanding population coupled with the increasing demand for good quality water for the urban and industrial sectors, result in both poorer quality water (e.g., Na-rich treated effluents or saline water) and soils being used for food production. Understanding the behavior of sodic soils and its effects on agriculture and adjacent ecosystems is, therefore, essential to manage sodic soils properly to secure crop production and maintain sustainable agriculture, worldwide.

Definitions of sodic soils

Prior to defining sodic soils it is necessary to present two important properties by which sodicity is evaluated. The first is the exchangeable sodium percentage (ESP) which describes the fraction of adsorbed Na from the cation exchange capacity (CEC) of the soil, and is defined as:

$$\text{ESP} = (\text{Exchangeable Na} / \text{Cation exchange capacity}) \cdot 100 \quad (1)$$

where the CEC is normally determined at a reference pH (7.0 or 8.2). However, instead of CEC it is also possible to use the sum of exchangeable cations (Ca + Mg + Na + K). The second property is the sodium adsorption ratio (SAR) that indicates the sodicity level of the irrigation water or soil solution, and is defined as:

$$\text{SAR} = (\text{Na}) / ((\text{Ca} + \text{Mg}) / 2)^{0.5} \quad (2)$$

where () indicates cation concentration in mmol_c L⁻¹. Thus, SAR has units of (mmol_c L⁻¹)^{0.5}. A more accurate calculation of the SAR is obtained when cation activities are used rather than cation concentrations. Based on soil chemistry considerations, Sposito and Mattigod (1977) in Levy (2012) introduced corrections for cation concentrations due to ion pairs or complexes and arrived at the following empirical relation between true (SAR_t) and measured (SAR_p) SARs:

$$\text{SAR}_t = 0.08 + 1.15\text{SAR}_p \quad (3)$$

In arid and semiarid soils, irrigation may lead to dissolution or precipitation of CaCO₃, depending on soil pH and the CO₂ partial pressure (pCO₂) of the system. Thus, the SAR of irrigation water (SAR_{iw}) may differ from that of the drainage water (SAR_{dw}), with the latter being considered a better indicator for sodicity hazard. It has been further suggested that SAR_{dw} be calculated based on the concentration of Na and Mg in the irrigation water, leaching fraction, and Ca concentration in equilibrium with CaCO₃ at a given pCO₂. However, for soils containing gypsum (CaSO₄), Callaghan et al. (2016) proposed the use of equilibrium geochemical modeling for adjusting SAR of saturated paste for gypsum dissolution.

Recently, Rengasamy and Marchuk (2011) proposed an alternative index for water quality instead of SAR, namely the cation ratio for soil structural stability (CROSS) index. The CROSS index considers the effects of all four major cations present in the irrigation water:

$$\text{CROSS} = (\text{Na} + 0.56\text{K}) / (\text{Ca} + 0.6\text{Mg})^{0.5} \quad (4)$$

where all cation concentrations are expressed in mmol L⁻¹. Further generalization of the CROSS index based on both clay dispersing and flocculating abilities have also been proposed.

Exchange reactions take place between the soil solution and the exchangeable phase. Thus, soil ESP can be estimated from the SAR of saturated paste extracts using the following empirical relationship (USSL Staff, 1954):

$$\text{ESP} = 100(-0.0126 + 0.01475 \cdot \text{SAR}) / (1 + [-0.0126 + 0.01475 \cdot \text{SAR}]) \quad (5)$$

or from the monograph presented in USSL Staff (1954). When more dilute extracts are used, such as 1:5 soil:water ratio, then a different relationship holds (Rengasamy et al., 1984):

$$\text{ESP} = 1.95\text{SAR} + 1.8 \quad (6)$$

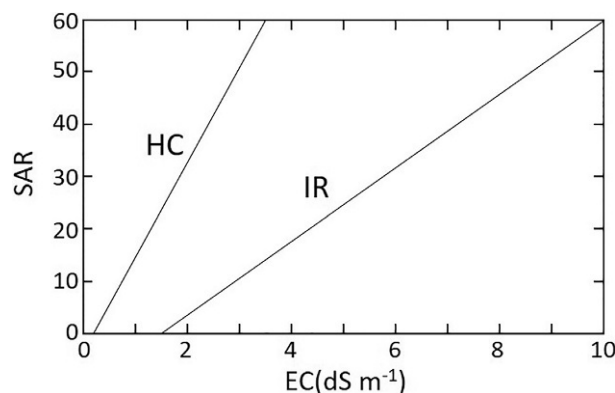


Fig. 1 Threshold values of sodium adsorption ratio (SAR) and electrical conductivity (EC) for hydraulic conductivity (HC) and infiltration rate (IR) associated with the likelihood of substantial losses in permeability. Adapted from Oster JD and Jayawardane NS (1998). *Agricultural management of sodic soils*, p. 143–165. In Sumner ME and Naidu R (eds.) *Sodic Soils*. Oxford University Press.

The USSL Staff (1954) empirical method has gained worldwide popularity because of its simplicity and has been used extensively to predict ESPs of Na affected soils, even in recent years. However, Kopittke et al. (2006) in Levy (2012) stated that the relationship between SAR and ESP is not constant because preference for Na varies depending on the type of soil clay mineralogy and ionic strength; hence, this relationship should be determined directly for the soil of interest.

To date, there is no widely accepted definition of a sodic soil. The USSL Staff (1954) defined sodic soils as those whose physical properties are adversely affected by the presence of Na and suggested that an ESP of 15 should separate sodic from non-sodic soils. The USSL Staff (1954) added the reservation that this limit must be regarded as somewhat arbitrary and tentative. McNeal and Coleman (1966), supported use of $\text{ESP} > 15$ for defining sodic soils, but added that the electrolyte concentration in the percolating solution must exceed $3 \text{ mmol}_c \text{ L}^{-1}$ for this to hold true. Others (e.g., Levy et al., 2005), advocated that soils with $\text{ESP} > 5$ either exposed to solutions with total electrolyte concentration (TEC) of $0.7 \text{ mmol}_c \text{ L}^{-1}$ or being calcareous should be considered as sodic soils. The above critical ESP levels were based on hydraulic conductivity measurements, but studies on soil susceptibility to aggregate slaking (Levy, 2012), seal formation, runoff and soil loss (Kazman et al., 1983) have revealed that soils can exhibit sodic behavior at ESPs of 2–4. In view of the continuous effect of Na on soil behavior, from low to high ESP and TEC levels, no satisfactory decision or critical level of ESP can be recommended. Hence, Rengasamy et al. (1984) suggested that the “non-sodic” group of soils may be better classified as “low-sodic,” considering that dispersive effects may be present.

Moreover, as will be shown later, the interrelationship between ESP of the soil or SAR of its equilibrium solution, and the TEC of the soil solution, is important in dictating whether sodic or non-sodic behavior will be observed. This approach was used by Oster and Jayawardane (1998) to assess soil hydraulic conductivity and infiltration rate when exposed to waters with different TEC values (expressed in terms of electrical conductivity [EC]) when applied to soils with SAR values in the range of 0–60 (Fig. 1).

Origin and distribution of sodic soils

For most sodic soils, sodicity is a natural phenomenon related to the type of parent material and subsequent pedogenic processes. However, there are also sodic soils where sodicity arises from anthropogenic processes, which is termed secondary sodification. Irrigation without proper drainage, forest clearing, and other land management practices that can lead to waterlogging are the main human activities that cause rapid secondary sodification. It is expected that secondary sodification of soils will increase considerably in the future.

Data compiled by Wansik et al. (2016) indicates that sodic soils cover 581 million ha worldwide. The largest areas of sodic soils are concentrated in Australia (340 million ha) followed by north-central Asia (120.1 million ha) South America (59.6 million ha), Africa (27 million ha), Europe (22.3 million ha) and North America (9.6 million ha).

The spatial distribution of sodic soils demonstrates that they occur under a range of climates. Among the environmental conditions that promote the formation of sodic soils are the presence of shallow saline groundwater, the occurrence of perched watertables within 1 m of the soil surface, impeded drainage, low slope gradients, and textural discontinuities during deposition of sediments such as eolian, glacial or alluvial materials (Bui et al., 1998, in Levy, 2012).

The distribution of sodic soils and the conditions promoting their formation are based on a pedological or morphological approach. By contrast, a more useful definition of sodic soils is based on land management considerations. Many agricultural soils that exhibit sodic behavior may not fit into the classical sodic soil group, because typical morphological features including columnar subsurface structure are absent. Consequently, the large areas that these soils cover remain essentially unaccounted for.

Mechanisms responsible for sodic behavior

Among the various soil constituents, the fraction that determines the physical behavior of soils is mainly colloidal clay. Compared with other soil constituents, the clay fraction possesses a large specific surface area (per gram of clay), which combined with its charge make it a very reactive soil constituent in physico-chemical processes, mainly swelling and dispersion. These two processes largely determine the microstructure of the soil, thus significantly affecting many of its physical properties (Shainberg and Letey, 1984).

Clay swelling

Swelling takes place on wetting in smectite type clays, represented most commonly in soils by montmorillonite. Repulsive forces (swelling pressure) that can be predicted from the diffuse double layer (DDL) theory are responsible for the swelling (van Olphen, 1977, in Shainberg and Letey, 1984); their magnitude depends on the TEC of the ambient solution and degree of sodicity. When saturated with Na ions and in the presence of a dilute solution, a diffuse (wide) DDL forms creating high swelling pressures between the clay platelets, resulting ultimately in dispersion (see discussion below) of the platelets (Shainberg et al., 1971 in Shainberg and Letey, 1984). Conversely, low swelling pressures are observed between Ca-saturated platelets because strong forces of electrical attraction between Ca and the clay surfaces (not considered by the DDL theory) prevent complete swelling of Ca montmorillonite, even in deionized water. Instead, Ca platelets aggregate into tactoids or quasicrystals consisting of 4–9 platelets each. Moreover, in mixed Na–Ca systems, the ions are not randomly distributed throughout the system as was previously thought, but are located preferentially on inner and outer surfaces, respectively (Shainberg and Otoh, 1968 in Shainberg and Letey, 1984). This phenomenon is called ion demixing. Shainberg et al. 1971 in Shainberg and Letey (1984) demonstrated that clay swelling is limited at $ESP < 15$ because of demixing; as ESP increases further, a sharp increase in macroscopic swelling occurs as Na enters the tactoids.

Clay dispersion

Unlike swelling, a continuous and reversible process, clay dispersion, the separation of clay tactoids to single platelets, is not continuous and may even occur at low ESP values if TEC of the ambient solution is lower than a critical flocculation concentration (CFC) (Shainberg and Letey, 1984). Dispersion is an irreversible process because flocculation by increasing the TEC does not result in the original particle associations and orientations. A schematic illustration of the difference between dispersion and swelling is presented in Fig. 2.

Among the main clay minerals in soils, illite is the most sensitive to dispersion. The TEC required to prevent Na-illite from dispersing is $40\text{--}50\text{ mmol}_c\text{ L}^{-1}$, while for Na-montmorillonite a TEC of $12\text{--}16\text{ mmol}_c\text{ L}^{-1}$ is needed. For kaolinite systems no single TEC value can be used as a threshold value for preventing dispersion. Sodium saturated kaolinite will not disperse even in a salt-free solution provided solution-pH < 7 . With increasing pH, the positive charge on the edges of kaolinite particles is gradually eliminated and the kaolinite system becomes increasingly dispersive (Levy, 2012 and references cited therein).

Clay dispersion is affected not only by the presence of Na but also by the type of divalent cation. The presence of Mg enhances clay dispersion more than Ca in soils with mixed mineralogy (Ali et al., 1987 in Levy, 2012), and in kaolinitic and illitic soils.

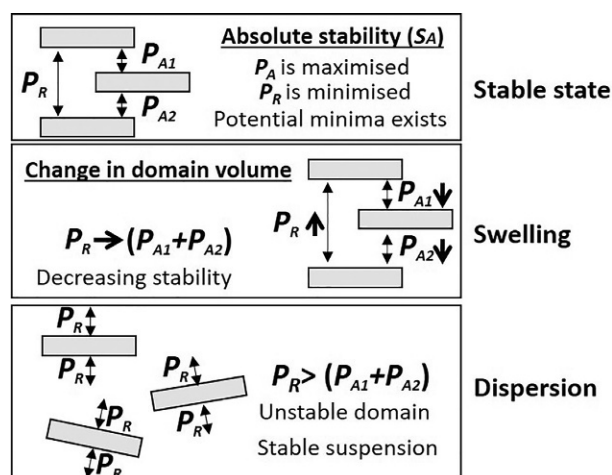


Fig. 2 Changes in the structure of a clay domain from a stable state to swelling state and to dispersion. Adapted from Bennett J, Marchuk A, Marchuk S and Raine SR (2019) Towards predicting the soil-specific threshold electrolyte concentration of soil as a reduction in saturated hydraulic conductivity: The role of clay net negative charge. *Geoderma* 337: 122–131.

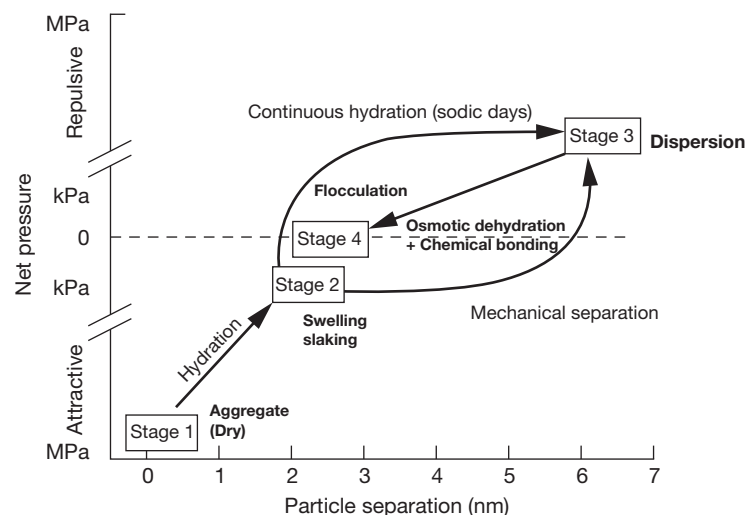


Fig. 3 Schematic illustration of the processes and intensity of attractive and repulsive forces involved when a dry soil aggregate is wetted, based on (Rengasamy and Sumner, 1998). Reprinted from Qadir M, Oster JD, Schubert S, Noble AD and Sahrawat KL (2007) Phytoremediation of sodic and saline-sodic soils. *Advances in Agronomy* 96: 197–247.

Furthermore, aggregates saturated with Na and Mg disperse at lower ESPs than those saturated with Na and Ca (Ali et al., 1987 in Levy, 2012).

Swelling and dispersion of soil clays differ significantly from that of pure clay systems, possibly because soil clays usually occur as mixtures and because of their association with other minerals, oxides and organic matter in the soil. It has been demonstrated that the TEC required for preventing dispersion of soil clays is 2–10-fold higher than that for pure clays (Frenkel et al., 1992). On the other hand, the presence of hydrous oxides or sparingly soluble minerals such as CaCO_3 has resulted in lower soil clay dispersivity. Thus, extrapolation of sensitivity to dispersion from pure clay systems to soil systems may often prove problematic.

The classical theory of DDL repulsive forces and van der Waals attractive forces developed for colloidal systems cannot satisfactorily explain clay dispersion following the disintegration of soil aggregates on wetting. Thus, a schematic 4-stage model was proposed by Rengasamy and Sumner (1998) to explain swelling and dispersion of soil clay particles (Fig. 3). Initially, when dry aggregates are wetted hydration of the adsorbed cations leads to swelling. Further hydration leads, in the case of highly sodic soils, to spontaneous dispersion, while for soils containing predominantly divalent cations mechanical separation is needed to disperse the clay. Dehydration of the dispersed system and buildup of osmotic pressure by increasing the electrolyte concentration leads to flocculation of the dispersed particles (Fig. 3).

Structural stability of sodic soils

Soil structural stability is the ability of the soil to retain its arrangement of aggregates and pores when exposed to different stresses. One of the most common properties used to characterize soil structure stability, especially against forces applied by water, is aggregate stability.

Aggregate breakdown by water may result from several physical and physico-chemical mechanisms. These include: (i) slaking, i.e., breakdown caused by compression of entrapped air during fast wetting; (ii) breakdown by differential swelling during fast wetting; (iii) breakdown by impact of raindrops; and (iv) physico-chemical dispersion due to osmotic stress upon wetting with low electrolyte water (Levy, 2012 and references cited therein).

Sodicity and aggregate stability

The effects of exchangeable Na in enhancing aggregate breakdown may occur by weakening the covalent associations between organic materials and soil minerals, and/or by increasing the osmotic/hydration forces that can cause particle repulsion during wetting.

Direct measurements of the effects of sodicity on aggregate stability have yielded inconclusive results. Some studies covering a wide range of soil texture (e.g., Alaboz, 2021) observed no correlation between aggregate stability and level of soil sodicity. Other studies have suggested that macroaggregate disintegration was affected only by $\text{ESP} > 10$, yet Crescimanno et al. (1995) in Levy (2012) noted that $\text{ESP} \leq 6$ could already have an adverse effect on aggregate stability, especially in fine-textured soils. Conversely, in the presence of Al and Fe oxides high ESP and low EC have only a small effect in promoting aggregate breakdown and clay

dispersion. Ruiz-Vera and Wu (2006) suggested that the response of aggregate stability to sodicity depends on clay mineralogy and content, and the rate at which the aggregates are wetted.

Role of organic matter

Soil organic matter (OM) comprises about 1–10% of a typical soil's mass of which 1–5% is microbial biomass (Nelson and Oades, 1998). Organic matter acts both as a bonding and dispersing agent in soils. The balance between these opposing effects depends upon the nature of the organic materials and their types of interaction with inorganic colloids. Because of the importance of organic matter in controlling soil structural stability, its involvement in aggregate breakdown under sodic conditions deserves a specific discussion.

Large amounts of soil-borne or added OM seem to promote resistance to sodic conditions. An inverse relationship has been noted between soil instability (inferred from modulus of rupture) and soil OM content at a wide range of ESP (Clark et al., 2009), especially for soils containing 10–20% clay. The favorable effects of organic agents on aggregate stabilization can be explained by nonionic bonding mechanisms by which roots, hyphae and natural polysaccharides stabilize aggregates, and which are not always directly affected by interactions between water, clay and exchangeable cations. An additional mechanism proposed for improving aggregate stability by SOM in sodic soils was the accumulation of cations such as Ca^{2+} , Mg^{2+} and K^+ and the increased leaching of Na^+ leading to low ESP in such soils.

The optimal dose of compost or other organic amendments to be applied to sodic soil ($\text{ESP} > 20$) for recovery of soil physical, chemical, and biological properties ranges from 10 to 50 t ha⁻¹ for different organic amendments (Leogrande and Vitti, 2019 and references cited therein). However, high rates of organic amendments (greater than 50 t ha⁻¹), even if useful to reduce soil sodicity, may cause increases of soluble salt concentrations and the accumulation of heavy metals or organic pollutants in the subsoil and/or groundwater (Leogrande and Vitti, 2019).

The more persistent SOM components including polyanionic colloidal materials (plant remains and microbial products), often referred to as humic substances, may adversely affect the stability of smaller scale aggregates (microaggregates) through a variety of mechanisms, but especially through their dispersing power. Organic anions can increase clay dispersivity, in particular when variable charge minerals are present mainly by increasing the negative charge density of the soil colloid fraction (Frenkel et al., 1992). Because humic substances form chelates with Ca, the effective SAR of the soil solution increases, and clay dispersion is promoted. There are also indications that the addition of humic fractions from a variety of sources (permanent pasture soil, cattle manure, and raw oxidized coal) could improve aggregate stability even in highly sodic soils ($\text{ESP} = 17$) (Imbufe et al., 2005 in Levy, 2012).

Water flow in sodic soils

Sodicity related degradation of soil structure leads to deterioration in soil water transmission properties. Water flow in the soil profile is commonly characterized by its hydraulic conductivity (HC), while water flow into the soil is characterized by its infiltration rate (IR). A distinction between HC and IR is made because a thin seal forms at the soil surface when water flows into the soil due to considerable disturbance from external forces, e.g., rain drop impact. Consequently, properties of the seal (or crust) differ from those of the underlying soil, and its thickness is difficult to measure, which makes the determination of its HC extremely difficult compared with a direct measure of its IR. Accordingly, the impact of sodicity on HC and IR are discussed separately.

Hydraulic conductivity (HC)

The HC of a soil depends on its properties and those of the percolating fluid. Several models have been proposed to describe the relations between the structure of a porous material, reflected by its pore size distribution, and its permeability. These types of models have failed to predict soil HC adequately, largely because they do not take into account the effect of fluid properties on the soil matrix geometry. Swelling and dispersion of clay in response to sodic conditions and the level of TEC in the percolating solution change the size of the conducting pores and hence affect the HC (Levy, 2012 and references cited therein).

Sodicity and electrolyte concentration

The realization that soil HC depends on both the SAR and TEC of the percolating solution, led to the development of the “threshold concentration” concept (Quirk and Schofield, 1955 in Shainberg and Letey, 1984) for defining the TEC required to prevent a decrease of usually >25% in the HC for a given soil ESP or SAR of the percolating solution (Fig. 4). This concept has been tested and validated on a large variety of soil types (e.g., Bennett et al., 2019) exposed to water of low salinity (e.g., rain or snow water), and explains decreased HC in soils, even Ca-soil and soils of low ESP. The decrease in HC is because salt concentration in the soil solution is not sufficient to prevent swelling and clay dispersion. Conversely, the Israeli experience in commercial fields showed that for Na-affected soils, irrigation with water having an SAR value as high as 26 with an electrical conductivity of 4.6 dS m⁻¹ exhibited no permeability problems because the electrolyte concentration in the irrigation water was sufficient to prevent the dispersive effect of Na (Shainberg and Letey, 1984). However, on applying electrolyte free water (to simulate rain water in winter), soil HC decreased

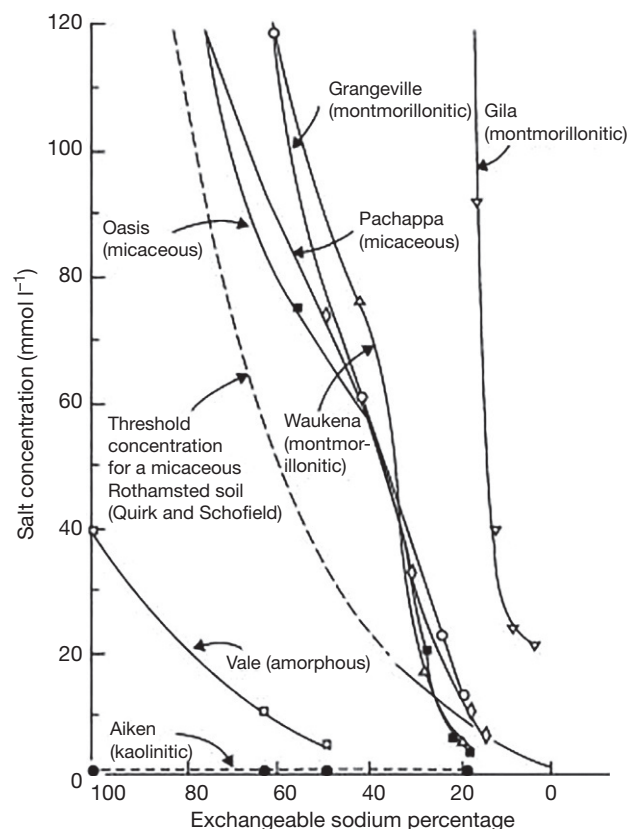


Fig. 4 Combinations of salt concentration and exchangeable sodium percentage required to produce a 25% reduction in hydraulic conductivity for selected soils. From McNeal BL and Coleman NT (1966) Effect of solution composition on soil hydraulic conductivity. *Soil Science Society of America Proceedings* 30: 308–312.

to a small fraction of its initial high value. Similar problems in HC arising from a decrease in TEC of the percolating water have been reported for soils mainly from semi-arid and arid zones. It was noted, however, that maintaining TEC of $10 \text{ mmol}_e \text{ L}^{-1}$ was not enough in many cases to prevent a decrease in HC for solutions with $\text{SAR} \sim 5$ (Suarez et al., 2008 in Levy, 2012).

A distinction should be made between effects of clay swelling and those of clay dispersion on the HC of soils. Soils begin to swell as the TEC of the soil solution is reduced, but the effect of swelling on HC becomes apparent only at SAR values above 10 in medium- and heavy-textured soils. In light textured soils the effect of swelling on reducing the conducting pores and hence HC is hardly noticeable. Furthermore, the swelling process is reversible, thus an increase in HC can be noted following an increase in TEC of the soil solution. Clay dispersion, on the other hand, can take place at low ESP levels provided the TEC is below a critical level. When clay dispersion occurs, the dispersed clay particles move down the soil profile and may even cause a complete blockage of the conducting pores, and hence an irreversible change in the HC (Shainberg and Letey, 1984). Consequently, the effects of reducing TEC on the HC are likely to be more severe when clay dispersion rather than swelling, is the predominant cause.

Soil properties affecting sodic behavior

The response of soils to sodic conditions varies among soils, mainly because soil properties have a significant effect on the dependence of the HC on sodicity. A major factor causing differences among various sodic soils in their susceptibility to hydraulic failure when leached with water of low TEC was their rate of salt release from mineral dissolution. Mineral dissolution determines the electrolyte concentrations of percolating solutions and hence soil HC. Sodic soils containing minerals (CaCO_3 and a few primary minerals) that readily release soluble electrolytes will not readily disperse when leached with deionized water (DW) at moderate ESP values, because a sufficiently high electrolyte concentration ($\sim 3 \text{ mmol}_e \text{ L}^{-1}$) could be maintained to prevent clay dispersion. In addition, ESP values will decrease, because most of the cations released are Ca and Mg. Conversely, soil solution concentrations in soil lacking readily weatherable minerals are likely to be below the critical TEC required for preventing dispersion, making the soils more susceptible to clay dispersion and reduction in HC (Shainberg and Letey, 1984).

Clay content, through its effect on pore-size distribution, contributes significantly to the response of soils to sodic conditions. The sensitivity of HC to sodic conditions is generally greater for soils with medium and high clay content. When clay content is high, clay swelling may significantly reduce the size of water conducting pores and lead to low HC. Similarly, under dispersive conditions the dispersed clay particles clog the pores and cause a reduction in soil HC. Conversely, in sandy soils, due to the low clay content, clay swelling has only a limited effect on the HC. Conditions supporting clay dispersion in these soils may cause HC first to decline

and subsequently to recover. This phenomenon is explained by the change in flow pattern from a solution flowing through a matrix of sand covered with clay, which starts to disperse and reduce HC to the flow of a clay suspension in a pure sandy matrix with large pores resulting in a subsequent increase in HC (Pupisky and Shainberg, 1979 in Shainberg and Letey, 1984). More recently, Levy et al. (2005) noted that the response of soils' HC to the combined impact of clay content and soil ESP on the HC is difficult to predict. Clay content acts as a cementing agent, hence higher clay content supports more stable aggregates and prevents the decrease in HC. Conversely, the greater the clay content the more effective is clay swelling in narrowing the water conducting pores. Changes in soil ESP affect soil swelling and dispersion. The reliance of HC on both factors depends on the relative weight of each of these factors in determining the mechanism that affects HC (Levy, 2012).

The role of soil OM and, especially its labile fraction, water extractable OM (commonly termed dissolved organic matter [DOM]), in determining soil sensitivity to sodicity has recently gained much interest. The reason for that lies in the increasing use of treated wastewater for irrigation of cultivated lands. Unlike fresh water, TWW contains DOM, the amount of which depends on the treatment the TWW had been subjected to, and higher SAR than its original fresh water. The relation between soil solution sodicity (SAR) and DOM are complex. Concentration of DOM and specific ultra-violet absorbance (SUVA) were significantly higher at SAR 20 than at SAR < 3 in combination with low TEC (Mavi et al., 2012). In turn, increased concentration of dissolved organic molecules may complex more Ca^{2+} and Mg^{2+} cations that lead to an increase in the effective SAR of the soil solution, which would finally contribute to a decrease in soil HC. Reports of differences of 48% between SAR and effective SAR were attributed to a greater affinity of divalent cations to organic ligands. The direct effect of DOM on water flow in soil was considered to be equivalent to an increase in SAR of 2–3 units (Suarez and Gonzalez-Rubio, 2017). Yet, in soils containing available Ca (e.g., calcareous soils), the affinity of the soil to Na decreases with the increase in DOM concentration, suggesting that DOM diminishes the adverse effects of Na on the soil (Assouline et al., 2016).

Other soil properties that may affect the response of soils to sodic conditions include:

1. Presence of aluminum and iron oxides—these oxides commonly occur in highly weathered soils and enhance, via cementing action, soil stability, and substantially decrease its susceptibility to sodic conditions.
2. Clay mineralogy—soil with high content of 2:1 layer silicates (e.g., montmorillonite) are the most labile ones while those high in kaolinite and sesquioxides are the least labile.
3. Exchangeable Mg—although in many soils the exchangeable Ca/Mg molar ratio is high, but in others the reverse is true. Effect of adsorbed Mg on soil hydraulic properties seems to depend on soil type. In calcareous soils, the presence of Mg enhanced the dissolution of CaCO_3 producing electrolytes that prevent clay dispersion and HC decay, while in noncalcareous soils Mg causes a decrease in the HC of the soils beyond that of the Na/Ca system. Exchangeable Mg can cause a direct (“specific”) effect and decrease in HC, whereas Mg in irrigation water cannot counter the accumulation of exchangeable Na in the soil.

Sodicity and conditions prevailing in the soil

It has been recognized recently that soil response to sodic conditions depends not only on soil properties but also on the conditions prevailing in the soil such as antecedent moisture content, wetting rate, and wetting and drying cycles (Hamilton et al., 2007). Soil conditions affect aggregate stability/breakdown and thus pore-size distribution and consequently the HC. Aggregate breakdown results from dispersion and slaking during wetting, whose rate determines the degree of slaking. It was reported that the HC of sodic soils leached with DW decreased more steeply and to lower values with an increased rate of wetting. Fast wetting, which increases aggregate slaking, increased the susceptibility of soil to sodic conditions. Conversely, slow wetting decreased significantly the effects of sodicity on soil HC, especially in fine-textured soils (Fig. 5).

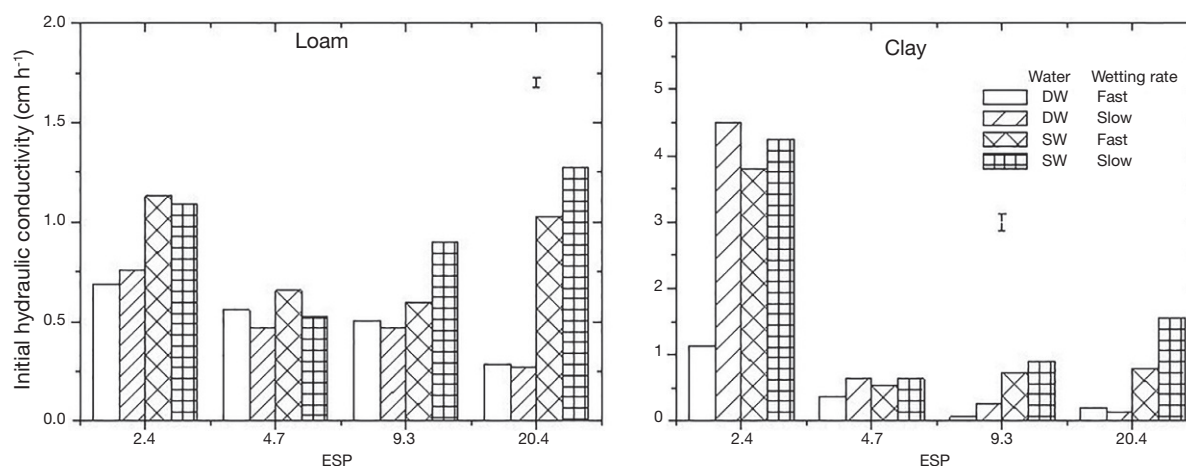


Fig. 5 Initial hydraulic conductivity of a loam and a clay as a function of the exchangeable sodium percentage (ESP), wetting rate and water quality. Bars indicate a single confidence interval at $P = .05$. DW = deionized water, SW = saline water. Adapted from Levy GJ, Goldstein D and Mamedov AI (2005). Saturated hydraulic conductivity of semi arid soils: Combined effects of salinity, sodicity and rate of wetting. *Soil Science Society of America Journal* 69: 653–662.

Infiltration rate

Soil infiltration rate (IR) is defined as the volume flux of water flowing into the profile per unit of soil surface area under any set of circumstances. Often, when water is supplied to the soil the IR decreases from its initial high rate due to the formation of a thin seal (<2 mm) at the soil surface. A structural seal, usually caused by raindrop impact, should be distinguished from a depositional seal formed by translocation and deposition of fine soil particles at a certain distance from their original location. The discussion to follow will focus on structural seals.

Sensitivity of seal formation and IR of a smectitic loamy sand to ESP when exposed to deionized water under laboratory simulated rain is presented in Fig. 6. Even at the lowest sodicity (ESP 1.0), a seal was formed and the IR decreased from an initial value >100 to a final infiltration rate (FIR) of 7.0 mm h⁻¹. An ESP value of 2.2 was sufficient to cause a further decrease in the final IR of the sandy loam to 2.4 mm h⁻¹. The amount of rain required to approach the FIR (i.e., the rate of seal formation), was also affected by sodicity (Fig. 6). As the energy from raindrop impact was the same in all experiments, the differences in IR curves for the various treatments were the result of chemical dispersion of the soil clay caused by sodicity. The sensitivity of the soil surface to low ESP values is explained by (i) the mechanical impact of the raindrops, which enhances chemical dispersion, and (ii) the almost total absence of electrolytes in the applied deionized water (Shainberg and Letey, 1984). With respect to (i), it has been noted that in the lower range of ESP the chemical mechanism needs some activation energy to start operating at the soil surface, which, in the case of rain is provided by raindrop impact. Conversely, in the absence of mechanical impact, the chemical mechanism does not come into effect at low ESP levels (<5) (Shainberg and Letey, 1984).

In cases where soil is exposed to overhead irrigation (e.g., sprinklers, moving irrigation line), IR is sensitive to both the TEC and SAR of the water applied. The lower the concentration of the electrolyte solution, the faster is the rate at which the IR decreases. For similar electrolyte concentrations, increasing soil ESP leads to a sharper decrease in and a lower final IR. Increasing TEC greatly increased the IR even at low ESPs of 1–2. For example, in Californian soils final IR increased from 2 to 28 mm h⁻¹ as cation concentration in the applied water with SAR <5 increased from 5 to 28 mmol_cL⁻¹. These results, and others, indicate that IR is far more sensitive than HC to both SAR and electrolyte concentration of the applied water (Levy, 2012).

Although Ca and Mg are both divalent cations, their effects as the complementary cation to Na, on IR may differ. At high energy rain (>20 kJ m⁻³) exchangeable Mg had an effect similar to that of Ca, but at low to medium energy (8.0–12.5 kJ m⁻³), the IR in Mg–Na soil was lower than that in the Ca–Na soil. These findings suggest that the adverse effect of Mg on IR is pronounced only under conditions where chemical dispersion is dominant in controlling IR, i.e., raindrops with low to medium kinetic energy (Keren, 1990 in Levy, 2012).

The susceptibility of sodic soils to seal formation and reduction in IR depends to a large extent on clay content. Soils with 10–30% clay are perceived as the most susceptible to seal formation and have the lowest IR. With increasing clay content, soil structure becomes more stable and seal formation diminishes. In soils with lower clay contents (<10%), the amount of clay available to disperse and clog soil pores is limited and, as a result, a poorly developed seal is formed (Levy, 2012).

Similar to HC, it has recently been recognized that the IR of sodic soils depends not only on soil properties but also on the prevailing conditions such as rate of aggregate wetting, antecedent moisture content (AMC) in the aggregates and aging duration (e.g., Kjaergaard et al., 2004 in Levy, 2012). This dependence is, however, greater in fine textured soils than in coarse ones.

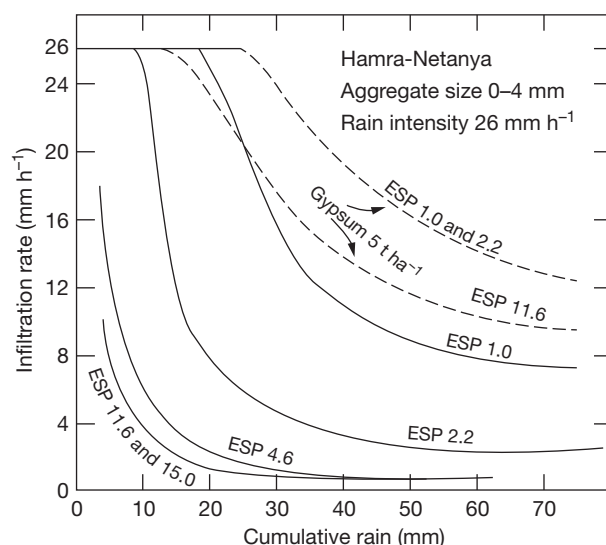


Fig. 6 The effects of the soil exchangeable sodium percentage (ESP) and phosphogypsum application on the infiltration rate of the sandy (Hamra Netanya, Israel) soil as a function of the cumulative rain. From Kazman Z, Shainberg I and Gal M (1983) Effect of low levels of exchangeable Na and applied phosphogypsum on the infiltration rate of various soils. *Soil Science* 35: 184–192.

Soil erosion

A strong relation between dispersible clay, IR and soil loss has been reported in numerous studies (Levy, 2012). The IR for soils from different regions of the world decreases as SAR and ESP increase, making more water available for runoff which, in turn, tends to increase erosion (Sumner, 1993, in Rengasamy and Sumner, 1998). Recently, Mamedov and Levy (2019) suggested that under sodic conditions, erosion by water increases exponentially with the increase in runoff compared with a linear relation between these in non-sodic soils. Levy et al. (1994) in Levy (2012), however did not observe a direct relationship between soil ESP, IR and erosion. They argued that with an increase in ESP the IR decreases, but the resulting erosion depends on the balance between the larger volume of runoff and a more dense and structurally more resistant seal; thus the amount of erosion cannot be predicted based on runoff volume only. Enhanced clay dispersivity associated with sodic conditions in the soil can lead not only to surface but also to subsurface erosion processes such as piping and tunneling, which may lead to the formation of deep gullies that further enhance erosion, mainly in kaolinitic and illitic soils.

Sodicity and soil productivity

Productivity of sodic soils is limited, mainly because the deterioration in their physical properties which causes poor aeration, restricts root development and enhances root diseases. These not only decrease crop yields but also limit the choice of crops that can be grown. For a detailed review of crop response to sodicity and sodicity-related fertility and nutritional constraints, readers are referred to Curtin and Naidu (1998).

Crop tolerance to sodicity

Research on crop tolerance to sodicity has been conducted under two types of condition, in the presence or absence of sodicity-related adverse soil physical conditions. Studies on crop tolerance to sodicity based merely on nutritional factors have indicated that at relatively low ESP values (2–10) some crops show signs of sodium toxicity (Pearson, 1960); however, most field crops are classified as moderately tolerant or tolerant to sodicity (Table 1). Conversely, studies carried out under conditions where soil structure was allowed to be affected by sodicity reported grasses as being tolerant, wheat (*Triticum aestivum* L.) and barley (*Hordeum vulgare* L.) being moderately tolerant, and peas (*Pisum sativum* L.) and beans (*Phaseolus vulgaris* L.) being sensitive to sodicity (Gupta and Abrol, 1990). Recently, Garg et al. (2005) observed that fennel (*Foeniculum vulgare* Mill) was tolerant to sodicity to ESP of 25. Use of the concept where the soil is allowed to be affected by sodicity may, however, cause results to be site specific and hinder their applicability to other locations and soil types. Thus, ranking of crops for sodicity tolerance depends on the conditions under which the response was determined.

Table 1 Tolerance of various crops to exchangeable sodium.

Tolerance to ESP ^a	Crop
Extremely sensitive (ESP 2–10)	Deciduous fruits Nuts Citrus Avocado (<i>Persea Americana</i> Mill)
Sensitive (ESP 10–20)	Beans (<i>Phaseolus vulgaris</i> L.)
Moderately tolerant (ESP 20–40)	Clover (<i>Trifolium</i> L.) Oats (<i>Avena sativa</i> L.) Tall fescue (<i>Festuca arundinacea</i> Shreb.) Rice (<i>Oryza sativa</i> L.) Dallis grass (<i>Paspalum dilatatum</i> Poir.)
Tolerant (ESP 40–60)	Wheat (<i>Triticum aestivum</i> L.) Cotton (<i>Gossypium hirsutum</i> L.) Alfalfa (<i>Medicago sativa</i> L.) Barley (<i>Hordeum vulgare</i> L.) Tomatoes (<i>Solanum lycopersicum</i> L.) Beets (<i>Beta vulgaris</i> L.)
Most tolerant (ESP >60)	Crested and fairway wheatgrass (<i>Agropyron cristatum</i> L.) Tall wheatgrass (<i>Thinopyrum ponticum</i> , (Podp.) Z.-W. Liu & R.-C. Wang) Rhodes grass (<i>Chloris gayana</i> Kunth)

^aESP, exchangeable sodium percentage.

Reproduced from Pearson GA (1960) Tolerance of crops to exchangeable sodium. *US Department of Agriculture Information Bulletin* 216.

Plant nutrition in sodic soils

Field fertility studies carried out in sodic soils often yield inconclusive results, mainly due to considerable spatial variation. Consequently, most dependable data on sodicity-related nutritional aspects have been obtained from greenhouse and laboratory studies (Levy, 2012).

Nitrogen, phosphorus, and potassium are the most important nutritional elements. Phosphorus and potassium, generally do not pose a nutritional problem in sodic soils. Conversely, nitrogen (N) can limit plant growth in many sodic soils because of low rates of N mineralization, denitrification under restricted aeration conditions, volatilization of NH_3 under alkaline conditions, and inhibition of NO_3 uptake by Cl or SO_4 ions which are usually abundant (Grattan and Grieve, 1999).

Plants growing in sodic soils may also suffer from impaired Ca uptake. The ratio of Ca to total cation concentration (TCC) is considered a better indicator of Ca availability to plants than actual Ca concentration. Carter and Webster (1990) found that Ca/TCC ratios <0.1 often occur in saturated paste extracts of sodic soils, which for many crops would result in Ca deficiency. In addition, high Mg/Ca ratios, often found in sodic soils, can also cause Ca deficiency.

Furthermore, extreme ratios of Na:K, Mg:Ca, and $\text{Cl}:\text{NO}_3$ (Grattan and Grieve, 1992 in Levy, 2012) are likely to be found in the soil solution of sodic soils. Such ratios, rather than absolute values, may lead to salt-stressed plants, specific-ion toxicity, and nutritional disorders.

Reclamation and management of sodic soils

Management of sodic soils, involving the improvement in soil structure (to alleviate problems of water permeability and logging, erosion, inadequate aeration, compaction, etc.), is essential if long-term productivity is to be maintained. Several site-specific approaches involving the use of chemical amendments, organic waste materials, tillage, crop diversification, have been used to ameliorate sodic soils.

Chemical amelioration

The most common concept for reclamation of sodic soils advocates the addition of divalent cations to the soil to replace exchangeable Na. There are two common practices: (i) use of chemical amendments that release Ca ions to the soil, and (ii) application of successive dilutions of a high salt water containing divalent cations.

Chemical amendments

Typical amendments are those containing a source of soluble Ca or that dissolve Ca upon reaction in the soil. Common amendments include gypsum, lime, CaCl_2 , sulfuric acid, and sulfur (Levy, 2012 and references cited therein).

Gypsum ($\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$), both from geological and non-geological sources (industries that produce CaSO_4 by-products) is the most commonly used amendment for the reclamation of sodic soil primarily because of its low cost, reasonable solubility and availability. Gypsum added to a sodic soil can increase aggregate stability and permeability by both TEC and cation-exchange effects. If gypsum is used to maintain high permeability at the soil surface during the application of good quality water, i.e., rainwater or other sources of water with low TEC, then the electrolyte effect is important. In this case, surface application of gypsum may be worthwhile, and the dissolution rate of gypsum dictates whether the rate of electrolyte supply to the water is sufficient to prevent clay dispersion and swelling. Thus, because of its higher rate of dissolution, gypsum from industrial sources such as by-products of the phosphate (phosphogypsum) and power generation (flue gas desulfurization gypsum) industries is preferable to mined gypsum. In this case, the amount of gypsum required depends on the amount of high quality water applied and the rate of gypsum dissolution, and is relatively independent of the amount of exchangeable Na in the soil profile (Levy, 2012).

Cation exchange effect is important in soils where treating high ESP of the soil profile is important. The amount of gypsum required depends then on that of exchangeable Na in a selected depth of soil. Amount of exchangeable Na to be replaced per unit land area (Q_{na} , mole_c ha⁻¹) during reclamation depends on the initial and desired final exchangeable Na fraction, cation exchange capacity of the soil, its bulk density and depth of soil to be reclaimed. Thus, the amount of gypsum needed to reclaim the soil, termed "gypsum requirement" (GR) is calculated as follows (Oster and Jayawardane, 1998):

$$\text{GR} = (8.61 \times 10^{-5}) Q_{\text{na}} \quad (7)$$

Efficiency and rate of exchange, namely percentage of Ca that exchanges for adsorbed Na, varies with ESP, being much greater at initial high ESP values.

Successive dilutions of high salt water

In areas where water is not a limiting factor, reclamation of sodic soils can be achieved by leaching the soil with successive dilutions of water with a high salt content containing divalent cations. In the early stage of the leaching, the high TEC in the applied water prevents clay dispersion and induces flocculation of the soil colloids. Simultaneously, the Ca ions in the water decrease sodicity by replacing exchangeable Na. Upon dilution of the saline water with high quality water, the SAR of the water is reduced. To ensure successful soil reclamation, depth of water added should be at least nine times the depth of soil to be reclaimed.

Problems associated with this method include (i) use of water containing inadequate concentration of divalent cations, (ii) the need to develop facilities that collect, convey and treat the saline water, and (iii) the need to dispose of the highly saline-sodic drainage water in such a way that soil, and surface and groundwater contamination are avoided (Qadir et al., 2007).

Amendment with organic wastes

In addition to chemical ameliorants, soil amendments can include organic wastes (fresh or composted) such as farm manure and plant residues that contribute to the stabilization of soil aggregates. Addition of these organic ameliorants is a widely recognized practice for managing cultivated sodic soils. Application of organic matter is most effective in ameliorating alkali (soils with high pH) sodic soils. However, its ameliorative effects are generally restricted to the surface horizons where organic matter inputs tend to be concentrated (Nelson and Oades, 1998). Because successful amelioration with organic materials must involve leaching, irrigation must be frequent and in small amounts. Under rainfall, amelioration with organic wastes can be expedited by minimizing runoff and evaporation, and increasing infiltration.

Furthermore, biologically ameliorated sodic soils must be cropped continuously. Rice is commonly used especially in the initial stages of reclamation because of its tolerance to sodicity, and also waterlogged conditions favor reclamation. When cropping is not possible, pastures, green manure and agroforestry can also contribute to soil reclamation (Levy, 2012).

Tillage

Tilling sodic soils, which often results in improvements in physical and hydraulic properties, reduces bulk density and increases total porosity (especially macroporosity), improves internal drainage and saturated and unsaturated HC. Consequently, aeration is improved and plant growth is enhanced (Sadiq et al., 2007). However, the beneficial effects of tillage on sodic soils are often not long lasting because they tend to reconsolidate under their own weight when wetted, and the hydraulic properties deteriorate once more.

Deep tillage (1–2 m) or ripping breaks up hardpans and cemented layers, mixes soil layers and causes, in general, a permanent improvement in soil structure and physical properties. To obtain optimal results plowing should be carried out at the correct soil moisture content that may differ according to soil texture. Deep tillage can sometimes lead to water logging and infiltration problems by bringing sodic subsoil to the surface. Conversely, shallow tillage (disking or chiseling) is effective in reducing soil permeability problems associated with surface crusts, hardsetting and compaction. However, shallow tillage may require repetition from time to time.

A combined treatment of tillage and gypsum application, termed “gypsum slotting” has been developed in Australia. Gypsum is added to narrow slots (about 150-mm wide, 0.4-m deep and spaced about 1–2-m apart) created by tillage at rates sufficient to reclaim the soil in the tilled slot. The advantages include efficient removal of excess water from the surface layer and its storage in the subsoil for use during dry spells, reclamation of the soil in the slots, maintaining adequate macroporosity and increasing yields (Jayawardane et al., 1987 in Levy, 2012).

Remediation with vegetation

Qadir et al. (2007) have extensively reviewed amelioration of sodic soils using vegetation. Vegetative remediation treatments are effective largely through their effects on both the physical and chemical properties of sodic soils (Fig. 7). Crop roots can stimulate changes in physical properties of the root zone in several ways such as removal of entrapped air from larger conducting pores, generation of alternate wetting and drying cycles, and creation of macropores. Aggregate stability is enhanced because of in situ production of polysaccharides and fungal hyphae in conjunction with differential dewatering at the root-soil interface.

Instead of deep plowing, which at times could aggravate rather than improve soil conditions, vigorous rooted rotation crops can also alleviate subsoil compaction. An alternative to subsoil amelioration is the use of ridge instead of bed farming techniques. With ridges, the plants are ~0.25 m above the neighboring furrows, and aeration is improved, especially in heavy textured soils suffering from water logging in the subsoil (Hearn and Constable, 1984 in Levy, 2012).

Cropping in sodic soils affects the chemistry of the root zone environment, especially in calcareous soils. The following processes are involved: (i) enhancing the $p\text{CO}_2$, (ii) dissolution of CO_2 in water to form carbonic acid (H_2CO_3) and its immediate dissociation to proton (H^+) and bicarbonate (HCO_3^-), (iii) reaction of H^+ with soil CaCO_3 to produce Ca, (iv) Na–Ca exchange at the soil's cation exchange sites (v) leaching of the exchanged Na^+ in the percolating water and (vi) subsequent reduction in soil sodicity (Fig. 7).

Removal of aboveground biomass of plant species, used for phytoremediation of sodic and saline-sodic soils, removes salts and Na taken up by the plants and accumulated in their shoots (Qadir et al., 2007). Highly salt-resistant species such as halophytes may accumulate quite large amounts of salts and Na in their shoots. The efficiency of various plant species for ameliorating sodic soils varies, possibly due to differences in their tolerance to salinity which often accompanies sodicity. A summary of lysimeters and field experiments showed that vegetative remediation was somewhat less effective than chemical amelioration in reducing soil sodicity. This finding could be explained, at least in part, because crops used on sodic soils may experience oxygen deficiency especially when excessive irrigations are used to leach the soil. Thus, evaluation of crops for the amelioration of sodic soils should consider also their resistance against hypoxia (Qadir et al., 2007).

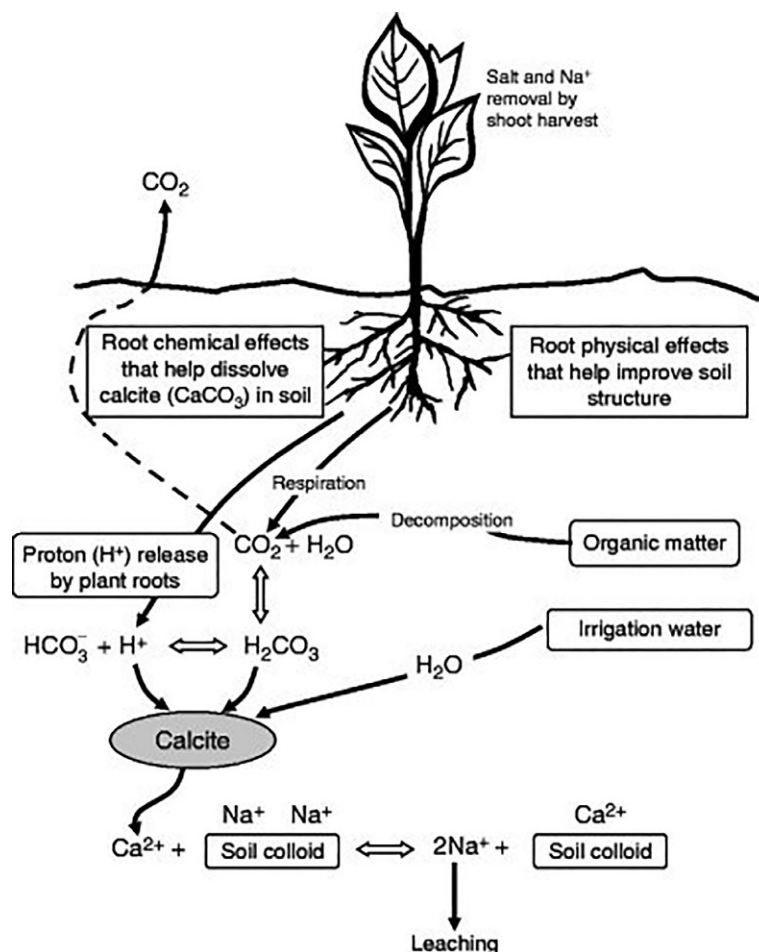


Fig. 7 Schematic illustration of driving forces for phytoremediation of calcareous sodic and saline-sodic soils. Reprinted from Qadir M, Oster JD, Schubert S, Noble AD and Sahrawat KL (2007) Phytoremediation of sodic and saline-sodic soils. *Advances in Agronomy* 96: 197–247.

Environmental aspects of sodic soils

Assessment of sodium related soil degradation has usually been carried out in the context of agriculture and soil productivity. However, sodic behavior of cultivated soils also poses numerous environmental hazards which are discussed below. Sodic soils may also have some environmental bearings on engineering and urban applications. Yet, they also present some opportunities for enhanced carbon sequestration in the soil.

Erosion and suspended sediments

Sodic soils are susceptible to clay dispersion and surface sealing, which generate runoff and cause erosion that poses a serious problem for crop production. About 40% or more of the sediments carried by runoff leave cultivated fields and reach streams or other water bodies, thus causing substantial negative effects to the environment. Moreover, large sediment and salt losses can occur in runoff from saline-sodic soils, even at low slopes and from apparently stable soils, with adverse consequences for downstream water quality. The sediments removed clog drainage ditches and irrigation canals, and decrease the capacity of water reservoirs, thus increasing maintenance costs. In addition, sediments damage or entirely destroy fish spawning areas. Sedimentation of rivers and streams reduces their capacity to carry water and increases the danger of flooding and consequent damage to property and to human health. It has been estimated that the cost of offsite damage from cropped land in the United States is in the range of \$2.2–13.5 billion annually (Tegtmeier, and Duffy, 2004).

The presence of suspended sediments in water, originating from eroded sodic soils, increases its turbidity. Turbid waters are less attractive for recreational activities and allow less sunlight to pass through, thus negatively affecting aquatic productivity. In addition, turbid water, which is denser than clean water, is more costly to pump. High turbidity in water bodies used for domestic consumption leads to extra costs in purification.

Salination and pollutants

Sodic soils commonly contain large amounts of salts that are transferred to nearby water bodies at a relatively fast rate when the soil is cultivated. Shallow saline–sodic watertables contribute significantly to the salt load of rivers. In addition, sodic soils often require proper drainage systems to remove water that is much more saline than the irrigation water. When this drainage water enters surface water, the quality sometimes deteriorates rendering it unsuitable for further use in irrigation, livestock watering or cities.

Salt accumulation resulting from sodic soil reclamation is an additional source for soil salination. Most reclamation activities require the addition of Ca salts (e.g., gypsum). Following exchange reactions in the soil, Na salts become dominant. In general, Na salts dissolve readily and are easily displaced downward into soil strata and groundwater beneath irrigated fields. Extensive reclamation can also cause shallow saline watertables which often require drainage. The saline drainage water ends up in surface water bodies, increases their salinity, and reducing their value. Therefore, reclamation of sodic soils may have some negative environmental consequences that should be carefully considered when planning and executing reclamation activities are considered.

Transport of various contaminants by soil colloidal materials suspended in runoff water to ground or surface water bodies poses a serious environmental hazard. Metals that are usually fairly immobile in soil and have strong affinity for soil colloids move together with the mobile colloids. Similarly, insoluble pesticides, herbicides, and some nutrients, e.g., P (Mamedov et al., 2020) have a strong affinity for soil colloids and can thus be transported with suspended solids to adjacent water bodies and groundwater. The increased concentration of suspended colloids in runoff water due to sodic conditions greatly increases the danger of higher than permitted loadings of these pollutants in ground and surface water bodies.

Carbon (C) sequestration

Sodic soils have lost a large fraction of their original C pool. The magnitude of the loss may range between 10 and 30 Mg C ha⁻¹, depending on the antecedent pool and the severity of degradation (Qadir et al., 2007). Amelioration of sodic soils may enhance C accumulation in the soil in both organic and inorganic forms.

Moreover, sodic soils commonly occur in arid and semi-arid regions where they generally contain large pools of primary (lithogenic) and secondary (pedogenic) inorganic carbonates (Lal, 2002). Dissolution of the carbonates in the root zone could result from the accumulation of high pCO₂ (root respiration and decomposition of organic matter) or by acidic root exudates. The resulting HCO₃⁻ can then be leached, especially under irrigation management, down the soil profile and its subsequent precipitation is a possible pathway for C sequestration in an inorganic form. It is estimated that C sequestration through this mechanism ranges between 0.25 and 1.0 Mg C ha⁻¹ year⁻¹.

Concluding comments

Sodium is present in many soils worldwide in amounts that may negatively affect their behavior. The area occupied by sodium-affected soils is expected to grow due to increased use of poor quality soils and water for agricultural production. The most important processes that occur in sodic soils, swelling and dispersion, adversely affect soil structure, and soil physical and hydraulic properties. The adverse effects of sodicity are inversely related to TEC of the irrigation water and soil solution. Soil sodicity enhances runoff, soil erosion and transport of contaminants that cause both on-site, (cultivated fields), and off-site (environmental) damage. However, sodium-related soil degradation and loss of productivity are, to a certain degree, reversible because sodic soils can be reclaimed. In addition, realization of the hazards of sodicity and consequent adoption of suitable management practices allow safe and sustainable use of sodium-affected soils, reducing the likelihood of their abandonment.

References

- Alaboz P (2021) Selecting soil properties for assessment of soil aggregation using principal component and clustering analyses. *Soil Research* 59: 170–178.
- Assouline S, Narkis K, Gherabli R, and Sposito G (2016) Combined effect of sodicity and organic matter on soil properties under long-term irrigation with treated wastewater. *Vadose Zone Journal* 15: 1–10.
- Bennett J, Marchuk A, Marchuk S, and Raine SR (2019) Towards predicting the soil-specific threshold electrolyte concentration of soil as a reduction in saturated hydraulic conductivity: The role of clay net negative charge. *Geoderma* 337: 122–131.
- Callaghan MV, Cey EE, and Bentley LR (2016) Adjustment of soil saturated paste extract electrical conductivity and sodium adsorption ratio for excess gypsum dissolution using equilibrium geochemical modeling. *Soil Science Society of America Journal* 80: 878–887.
- Carter MR and Webster GR (1990) Use of calcium-to-total cation ratio as an index of plant-available calcium. *Soil Science* 149: 212–217.
- Clark GJ, Sale PWG, and Tang C (2009) Organic amendments initiate the formation and stabilization of macroaggregates in high clay sodic soils. *Australian Journal of Soil Research* 47: 770–780.
- Curtin D and Naidu R (1998) Fertility constraints to plant production. In: Sumner ME and Naidu R (eds.) *Sodic Soils*, pp. 125–142. New York, NY: Oxford University Press.
- Frenkel H, Fey MV, and Levy GJ (1992) Organic and inorganic anion effects on reference and soil clay critical flocculation concentration. *Soil Science Society of America Journal* 56: 1762–1766.
- Garg VK, Singh PK, and Pushpangadan P (2005) Exchangeable sodium induced changes in yields, water relation and cation composition of fennel (*Foeniculum vulgare* Mill). *Journal of Environmental Biology* 26: 335–340.
- Grattan SR and Grieve CM (1992) Mineral element acquisition and growth response of plants in saline environments. *Agriculture, Ecosystems & Environment* 38: 275–300.

- Gupta RK and Abrol IP (1990) Salt-affected soils: Their reclamation and management for crop production. *Advances in Soil Sciences* 11: 223–288.
- Hamilton AJ, Stagnitti F, Xiong X, Kreidl SL, Benke KK, et al. (2007) Wastewater irrigation: The state of play. *Vadose Zone Journal* 6: 823–840.
- Jayawardane NS, Blackwell J, and Stapper M (1987) Effects of changes in moisture profiles of a transitional Red Brown earth due to surface and slotted gypsum applications. *Australian Journal of Agricultural Research* 38: 239–251.
- Kazman Z, Shainberg I, and Gal, M. (1983) Effect of low levels of exchangeable Na and applied phosphogypsum on the infiltration rate of various soils. *Soil Science* 35: 184–192.
- Lal R (2002) Carbon sequestration in dryland ecosystems of west Asia and north Africa. *Land Degradation & Development* 13: 45–59.
- Leogrande R and Vitti C (2019) Use of organic amendments to reclaim saline and sodic soils: A review. *Arid Land Research and Management* 33: 1–21.
- Levy, G.J. (2012). Sodicty. In: Huang, M.P., Li, Y., and Sumner, M.E. (eds.). *Handbook of Soil Sciences: Resource Management and Environmental Impacts* (2nd edn.). CRC Press, Boca Raton, FL, p. 18-1 to 18-28.
- Levy GJ, Goldstein D, and Mamedov AI (2005) Saturated hydraulic conductivity of semi arid soils: Combined effects of salinity, sodicty and rate of wetting. *Soil Science Society of America Journal* 69: 653–662.
- Mamedov AI and Levy GJ (2019) Soil erosion–runoff relations on cultivated land: Insights from laboratory studies. *European Journal of Soil Science* 70: 686–696.
- Mamedov AI, Bar-Yosef B, Levkovich I, Fine P, Silber A, and Levy GJ (2020) Physicochemical mechanisms underlying soil and organic amendment effects on runoff P losses. *Land Degradation & Development* 31: 2395–2404.
- Mavi MS, Marschner P, Chittleborough DJ, Cox JW, and Sanderman J (2012) Salinity and sodicty affect soil respiration and dissolved organic matter dynamics differentially in soils varying in texture. *Soil Biology and Biochemistry* 45: 8–13.
- McNeal BL and Coleman NT (1966) Effect of solution composition on soil hydraulic conductivity. *Soil Science Society of America Proceedings* 30: 308–312.
- Nelson PN and Oades JM (1998) Organic matter, sodicty and soil structure. In: Sumner ME and Naidu R (eds.) *Sodic Soils*, pp. 67–91. New York, NY: Oxford University Press.
- Oster JD and Jayawardane NS (1998) Agricultural management of sodic soils. In: Sumner ME and Naidu R (eds.) *Sodic Soils*, pp. 143–165. New York, NY: Oxford University Press.
- Pearson GA (1960) *Tolerance of Crops to Exchangeable Sodium*. USDA Inf. Bull 216.
- Qadir M, Oster JD, Schubert S, Noble AD, and Sahrawat KL (2007) Phytoremediation of sodic and saline-sodic soils. *Advances in Agronomy* 96: 197–247.
- Rengasamy P and Marchuk A (2011) Cation ratio of soil structure stability (CROSS). *Soil Research* 49: 280–285.
- Rengasamy P and Sumner ME (1998) Processes involved in sodic behavior. In: Sumner ME and Naidu R (eds.) *Sodic Soils*, pp. 51–67. New York, NY: Oxford University Press.
- Rengasamy P, Greene RSB, Ford GW, and Mehammi AH (1984) Identification of dispersive behaviour and the management of red-brown earths. *Australian Journal of Soil Research* 22: 413–431.
- Ruiz-Vera VM and Wu L (2006) Influence of sodicty, clay mineralogy, prewetting rate and their interactions on aggregate stability. *Soil Science Society of America Journal* 70: 1825–1833.
- Sadiq M, Hassan G, Mehdi SM, Hussain N, and Jamil M (2007) Amelioration of saline-sodic soils with tillage implements and sulfuric acid application. *Pedosphere* 17: 182–190.
- Shainberg I and Letey J (1984) Response of soils to sodic and saline conditions. *Hilgardia* 52: 1–57.
- Suarez DL and Gonzalez-Rubio A (2017) Effects of the dissolved organic carbon of treated municipal wastewater on soil infiltration as related to sodium adsorption ratio and pH. *Soil Science Society of America Journal* 81: 602–611.
- Szabolcs I (1979) *Review of Research on Salt Affected Soils*. Paris, France: UNESCO.
- Tegtmeier EM and Duffy MD (2004) External costs of agricultural production in the United States. *International Journal of Agricultural Sustainability* 2: 1–20.
- USSL Staff (1954) Diagnosis and improvement of saline and alkali soils. *Handbook of Agriculture*, vol. 60. Washington, DC: USDA.
- Wansik S, Siddikee MA, Joe MM, Benson A, Kim K, Selvakumar G, Kang Y, Jeon S, Samaddar S, Chatterjee P, Walitang D, Chanratana M, and Sa T (2016) Halotolerant plant growth promoting bacteria mediated salinity stress amelioration in plants. *Korean Journal of Soil Science and Fertilizer* 49: 355–367.

Organic soils

Heike Schimmel and Wulf Amelung, Institute of Crop Science and Resource Conservation (INRES), Soil Science and Soil Ecology, University of Bonn, Bonn, Germany

© 2023 Elsevier Ltd. All rights reserved.

This is an update of D.L. Mokma, ORGANIC SOILS, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 118-129, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00034-5>.

Introduction	383
Classification	385
Formation of organic soils	386
Formation of Folic Histosols	386
Formation of peatlands	386
Distribution	387
Tundra	387
Taiga	388
Temperate zone	388
Tropics	388
Mountain regions	389
Characteristics and properties	390
Folic Histosols	390
Peat soils	391
Fens (Rheic Histosols)	392
Bogs (Ombric Histosols)	392
Utilization	392
Consequences of anthropogenic use	393
Restoration and sustainable use of organic soils	395
Outlook	396
Acknowledgment	396
References	396

Abstract

Organic soils (Histosols) are formed by accumulation of incompletely decomposed organic matter. This chapter introduces the classification, genesis and properties of organic soils and potential management. Organic soils comprise histosols with folic horizons formed in cold climates as well as fens and bogs with or without contact with minerotrophic groundwater (mainly Rheic and Ombric Histosols, respectively). The soils have low bulk density and high water storage capacity, but may be limited by nitrogen, phosphorus, and other inorganic trace elements, particularly when plants lack contact with groundwater. Histosols cover 3% of the global land surface, but store 21% of terrestrial soil carbon.

Key points

- Organic soils form by the incomplete decomposition of plant remains.
- Organic soils comprise well aerated Folic Histosols as well as bogs and fens.
- Cold temperatures, terrestrialization and paludification are the main mechanisms of the formation of organic soils.
- Drainage and use of organic soils induce irreversible subsidence and carbon emissions.
- Restoration and paludiculture are sustainable land use options for organic soils.

Introduction

Organic soils store with 650 Gt carbon (C) 21% of the global soil organic C stock (Yu et al., 2010) and they can sequester another 0.1 Gt C annually (Frolking et al., 2011). However, they cover only 3% of the Earth's land surface. Nevertheless, areas with organic soils provide a unique habitat for many specialized and endangered species. Hence, organic soils are a key component of terrestrial feedbacks to biodiversity, global water, energy and matter cycles and thus to our climate system.

When biomass production exceeds decay, organic soil horizons and finally organic soils can be formed. Organic soils therefore cover a wide range of soil types under conditions promoting organic matter accumulation and inhibiting its decomposition. Hence,

organic soils are abundant in regions with cold temperatures, with water saturation or both. They are typically found in the subarctic and boreal regions of the northern hemisphere, in depressions and marshlands of the temperate zone, and in tropical lowlands. The organic horizons mainly comprise plant residues of different but incomplete stages of decomposition. The organic remains originate from various plant sources, including grasses, sedges, mosses or trees, with some additional contributions from, e.g., microbial necromass and black carbon. Due to the lack of stabilizing minerals, organic soils are vulnerable to changes in climatic conditions and management. The anthropogenic use of organic soils, for example, has been estimated to lead to almost 2 Gt of CO₂-eq emitted from drained organic soils only (Leifeld and Menichetti, 2018).

The classification system derived from the FAO (WRB; World Reference Base; IUSS Working Group WRB, 2015) defines organic soils as Histosols (Greek *histos*: tissue) with a diagnostic folic or histic horizon. These two diagnostic organic horizons may occur in other major reference soil groups, e.g., Cryosols, Leptosols, Podzols, where, however, they appear only as a single principle qualifier, which is not sufficient to categorize them as an organic soil. This is because, depending on clay content, the folic or histic horizons must also contain at least 12–18% organic C and have a minimum depth of either 10 cm when overlying rock or ice, or of 40 cm, when overlying mineral soil. In both cases the mineral matter does not influence soil properties significantly. Soils with a folic horizon (Folic Histosols) are well aerated and occur mostly in mountainous alpine areas and high latitudes where low temperatures reduce rates of decay. The vegetation period is short and biomass production is limited, so that folic horizons are frequently shallow. The organic matter composition usually differs from that of histic horizons because the vegetation cover is often different. Histic horizons are generally water-saturated for at least one or more months. Organic soils with a histic horizon are usually found in peatlands and can be several meters thick. There are two major types of peatlands: fens receive water and nutrients from groundwater and surface streams, while bogs are rain-fed and poor in nutrients (Fig. 1) (Box 1). The decomposition of organic

Box 1

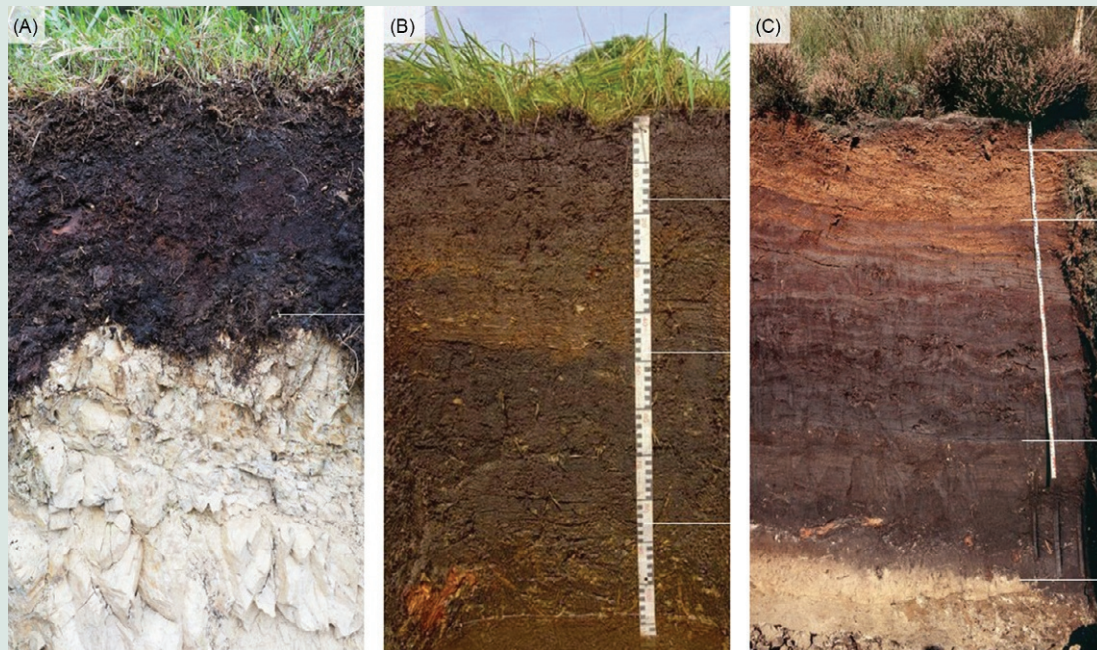


Fig. 1 Organic soil profiles with indicated horizons; (A) Folic Histosol (© Bayerisches Landesamt für Umwelt (LfU), Robert Traidl), (B) Fen profile (© Daniel Devecioglu), (C) Bog profile (© Landesamt für Bergbau, Energie und Geologie, LBEG). From (A) R. Traidl, Bayerisches Landesamt für Umwelt <https://www.lwf.bayern.de/wissenstransfer/forstliche-informationsarbeit/188945/index.php>; (B) Daniel Devecioglu, Bundesverband Boden e.V., <https://www.bodenwelten.de/content/niedermoor-boden-des-jahres-2012>; (C) Landesamt für Bergbau, Energie und Geologie. https://www.bgr.bund.de/DE/Themen/Boden/Bilder/LABO_Moorbodenschutz/Bod_Profil_BB17_g.html.

- Histosols with a *folic horizon* (Folic Histosols) overly bedrock or fragments and are naturally well drained. They receive their water and nutrient supply mainly from precipitation and are poor in nutrients and often acidic. They are found in cold regions in high latitudes or altitudes and the vegetation is mainly grasses, heathers and mosses while trees also grow on thicker Folic Histosols.
- Fens (e.g., Rheic, frequently also Eutric Histosols) are formed under high groundwater levels and can exhibit a wide range of pH and nutrient availability, depending on the type of bedrock. Fens with low pH and low nutrient availability (mesotrophic) support vegetation such as mosses and have a C:N ratio of 20–33, while nutrient-rich (eutrophic) fens have elders, sedges and reeds and a low C:N ratio of 10–20.
- Bogs (e.g., Ombric, frequently also Dystric Histosols) are fed solely by rainwater; thus, they are very acidic with pH values between 2.5 and 4. They are oligotrophic (nutrient-poor) and the C:N ratio is usually higher than 33. Highly specialized plant genera such as peat mosses (*Sphagnum*) and sundews (*Drosera*) live in bogs in temperate and boreal zones. In tropical bogs, the vegetation consists mainly of tall trees which have adapted to the low nutrient availability and anaerobic conditions.

material in peatlands is decelerated by water saturation and poor aeration in fens and in addition by acid conditions and limited nutrient availability in bogs.

All organic soils have distinct properties that distinguish them from mineral soils, such as large organic matter content, low bulk density, large pore volume and thus, large water storage capacity. Therefore, the agricultural use of organic soils is difficult and frequently unsustainable. Nevertheless, they are widely used for agriculture, forestry and peat extraction after drainage, which alters their physical and chemical properties, diminishes C storage and converts them into large C sources. In contrast to the conventional use, wet biomass cultivation (paludiculture) is one sustainable management option for organic soils.

Classification

There are non-uniform classifications for organic soils, some of which refer only to peat soils. Many countries use their own classification systems, which differ mainly in the thickness of the organic horizon and also in the required content of organic C. Here, we refer to the classification of the World Reference Base for Soil Resources (IUSS Working Group WRB, 2015) and also present Soil Taxonomy (United States Department of Agriculture, USDA). Both call organic soils Histosols (Table 1).

In the WRB classification, they are defined as soils having organic material which

- (a) starts at the soil surface, has a minimum thickness of 10 cm and directly overlies ice, continuous rock or coarse fragments of which the interstices are filled with organic material. Or
- (b) the organic material starts within the first 40 cm from the soil surface and has within the first meter of the soil a combined minimum thickness of either 60 cm if $\geq 75\%$ of the organic material is moss-derived or 40 cm in other organic material (IUSS Working Group WRB, 2015).

A further diagnostic criterion for Histosols is the content of organic C. The soil has to contain at least 20% organic C in the fine earth (by mass) or $(12 + [\text{clay percentage in the mineral fraction} \times 0.1])$ % or more organic C in the fine earth or at least 18% organic C in the fine earth if it is saturated with water for at least 30 consecutive days per year. Associated horizons with these properties are folic and histic horizons. Folic horizons are well aerated and water-saturated for less than 30 consecutive days per year. Histic horizons are, on the other hand, organic horizons which are poorly aerated and water-saturated for more than 30 consecutive days per year.

There are certain terms to describe the degree of decomposition of peat and other organic soils. 'Fibric' material is slightly decomposed, with 2/3 recognizable plant tissues. Amorphous organic material with less than 1/6 recognizable plant tissues is called sapric. 'Hemic' organic material is intermediate between fibric and sapric.

Unlike Histosols in the WRB classification, Histosols in the USDA Soil Taxonomy do not have permafrost or gelic materials within the first 100 cm. In addition, the suborder distinguishes between 'Folists' and 'Fibrists', 'Hemists' and 'Sapristis'. Folists represent freely drained organic soils where the organic matter overlies rock or fills its interstices. Fibrists, Hemists and Sapristis describe water-saturated organic soils with different degrees of decomposition, similar to the principal qualifiers fibric, hemic and sapric in the WRB classification. There, at least half of the material in the upper 80 cm is organic. If the bulk density is less than 0.1 g cm^{-3} , at least $\frac{3}{4}$ of the material needs to be organic (Soil Survey Staff, 1999).

Table 1 Criteria used for classification of organic soils in World Reference Base (IUSS Working Group WRB, 2015) and Soil Taxonomy (Soil Survey Staff, 1999).

Order	Principal Qualifiers/Suborder	Criteria
World Reference Base		
Histosols	Muusic/Rockic/Mawic	Organic soils overlying ice/continuous rock/coarse fragments
	Cryic	Presence of permafrost
	Thionic	Presence of sulfuric acid
	Folic	Naturally well-aerated organic soil
	Floatic/Subaquatic/Tidalic	Organic matter floating on water/submerged in lakes/affected by tidal waters
	Fibric/Hemic/Sapric	Degree of decomposition of organic matter
	Leptic	Having continuous rock in the first 100 cm
	Murshic/Drainic	Naturally drained with cracks and blocky structure/artificially drained organic soils
	Ombic/Rheic	Rain-fed/groundwater-fed wet organic soils
	Hyperskeletal/Skeletal	Organic soils with stones in the first 75 cm/100 cm
	Andic	Presence of weathered pyroclastic deposits
	Vitric	Presence of volcanic glass
	Calcic	Having a calcic horizon in the first 100 cm
	Dystric/Eutric	pH in water $<5.5/\geq 5.5$
Soil Taxonomy		
Histosols	Folists	Freely drained organic soils
	Fibrists	Wet organic soil with weakly decomposed organic matter
	Hemists	Wet organic soil with moderately decomposed organic matter
	Sapristis	Wet organic soil with highly decomposed organic matter

Formation of organic soils

Organic soils form under conditions that reduce the decay of organic material. While cold temperatures are the main determinant for folic horizons, water saturation is the key determinant for histic horizons. In principle it is also possible that organic matter accumulates due to dryness (Xeropreservation), but this does not form organic horizons because of insufficient biomass production. Slow rates of decomposition also occur under unfavorable, e.g., acidic soil conditions. The resulting organic layers in, for example, gymnosperm forests are part of Folic Podzols and rarely grow to a thickness that would allow reclassification as Histosols.

Formation of Folic Histosols

Histosols with a folic horizon develop on bedrock or ice in flat areas in high latitudes and on slopes in high mountain regions. The regions are characterized by cold temperatures which inhibit plant growth and decomposition of organic matter. Also, with reduced biological activity, there is little mixing of the soil through bioturbation. The succession sequence is initially colonization by undemanding vegetation such as lichens and mosses on mostly solid bedrock. Then, first grasses, such as *Carex firma*, and herbs, ferns and heathers establish. With increasing thickness of the organic horizon, dwarf mountain pines (*Pinus mugo*) grow (Fig. 2). Their litter decomposes slowly and leads to further growth of the organic horizon and increases the nutrient stock and availability. With increasing thickness and root space, the mountain pine can be replaced by spruce (*Picea* spp.). Their dead wood also plays a role in organic matter accumulation. Above the tree-line and under deciduous forests, the Folic Histosols are often shallower due to the more easily decomposable organic matter. In the tundra, they reach thicknesses of more than 30 cm, while they can be 1 m thick under coniferous forest. Their rate of accumulation and growth are consequently slow, with average growth rates between 0.1 and 0.3 mm per year. However, a large and rapid supply of poorly decomposable litter can also favor faster accumulation.

Formation of peatlands

Under anaerobic conditions, the decomposition of organic material is low because oxygen (O) diffuses 10,000 times more slowly through water than air. Peat accumulates because more biomass is produced than decays. As a result, there is vertical growth of these organic soils through time. Peat can form directly over wet mineral soil or via terrestrialization or paludification. Organic soils

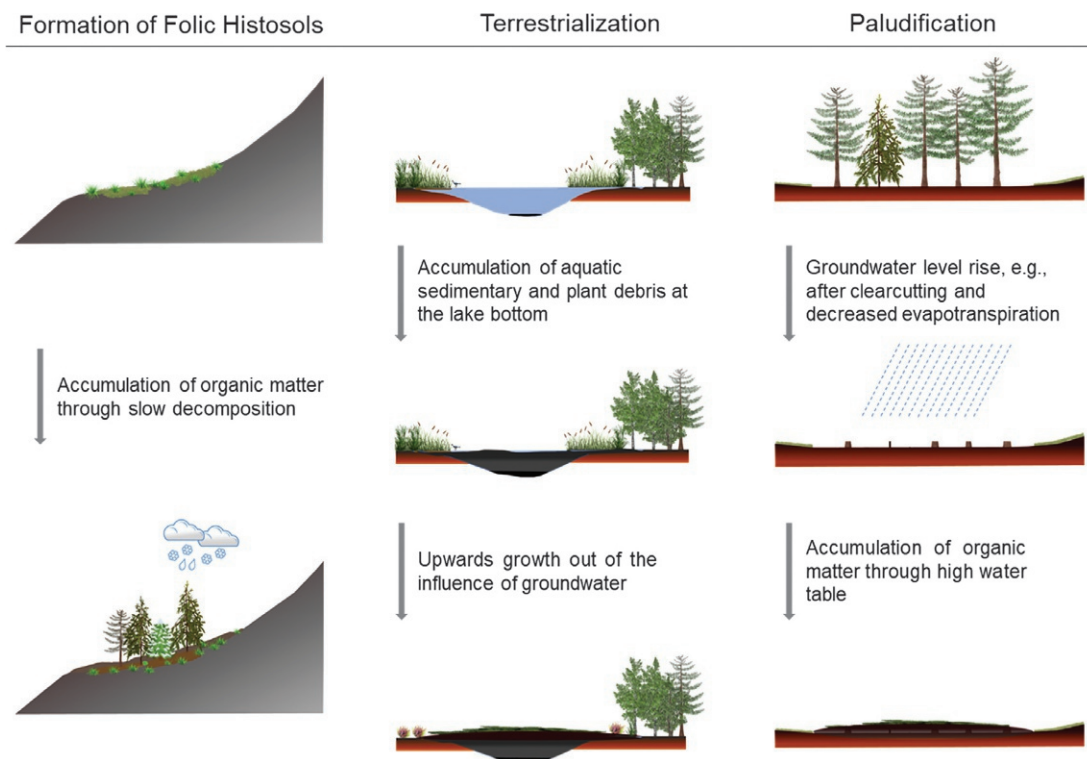


Fig. 2 A conceptual illustration of the main processes in the formation of Histosols. Folic Histosols form through slow decomposition under cold temperatures. Peat soils develop through terrestrialization, which is the development of peat in open water or paludification which is the development of peat over formerly drier mineral soil, e.g., by water table rise through clearcutting. Paludification may also involve other processes causing groundwater level to rise, such as on indurated soil horizons or after melting of permafrost (not illustrated here).



Fig. 3 (A) peat moss *Sphagnum pulchrum* (© Monika Koperski); (B) weakly decomposed *Sphagnum* peat (© Corinna Schulz, Ron Meier-Uhlherr). From (A) Koperski M (2016) In: *Niedersächsisches Ministerium für Umwelt, Energie, Bauen und Klimaschutz, 2016. Programm Niedersächsische Moorlandschaften. Grundlagen, Ziele, Umsetzung.* Hannover; (B) Meier-Uhlherr R, Schulz C, Luthardt V (2015) *Steckbrief Moorsubstrate. 2., unveränd. Aufl.* HNE Eberswalde, Berlin.

formed by terrestrialization develop by the accumulation of peat in open water systems, either at the bottom of lakes from the accumulation of dead plant biomass or in floating mats. Plant debris from lakes, e.g., sedges and reeds, only partly decompose and accumulate in the lake margin area. The lake might gradually fill from the margins with organic matter. When completely filled, the peat can grow upwards, thereby losing contact with the groundwater and forming rain-fed bogs (Fig. 2). Paludification, on the other hand, describes peat accumulation directly over former drier mineral soil where water-saturated conditions occur through changing external factors such as indurated soil layers in Podzols or melting permafrost. Clearcutting of forest can also lead to peat formation because of the considerable reduction in evapotranspiration with the lack of trees, resulting in a rise in water table (Fig. 2). Consequently, peatlands formed by paludification often have woody peat with tree stumps at the bottom, while terrestrialization peatlands show aquatic sediments at the bottom (Rydin and Jeglum, 2006).

Water saturation inhibits the decay of the peat, but aeration at the soil surface accelerates organic matter decomposition, with potential losses of 80–90% of dry mass. This rate of decay depends strongly on the plant genera, for example, *Sphagnum* moss (Fig. 3) is more resistant to decomposition than parts of vascular plants, e.g., *Molinia* species. Consequently, some plant species are more likely to form peat although it can form from almost all plant genera including trees, mosses, heathers, grasses. Northern peat soils are more likely to be formed by mosses, grasses and reeds, whereas roots of trees provide the main materials for forming peat soils in the tropics.

Organic peat soils grow slowly and need hundreds of years to reach a thickness of several dm. The average rate of growth of peat is 1 mm per year, but it can vary from 0.3 to 13 mm a⁻¹, depending on the region. Most organic soils in the northern hemisphere have formed since the end of the last ice age about 11,000 years ago and reach depths of more than 10 m, but many are shallower. Tropical peatlands can reach greater thicknesses with depths exceeding 20 m. It is estimated that peat soils have an average thickness of 1.3–1.4 m worldwide.

Distribution

Organic soils are predominant in the boreal and subarctic regions of the northern hemisphere. They cover vast areas in Russia, Canada, Alaska, Scandinavia, Great Britain, the Baltic countries, Patagonia and the regions of Central Europe influenced by oceanic climates. Histic organic soils also occur on tropical lowlands, especially in Indonesia, Central Africa and the Amazonas region. Overall, around 3/4 of all organic soils are in the boreal and subarctic regions, another 10% in the mild, rainy, oceanic regions of the mid-latitudes and in mountain regions, and 1/6 in the tropics (Fig. 4; Page et al., 2011). Organic soils occur in almost every region, but their inventory is largely incomplete. In particular, the distribution of tropical peatlands and Follic Histosols has not yet been adequately studied. Follic Histosols occur mostly on a small scale and are often close to other Histosols. The extent of Follic Histosols is estimated to be 0.15% of the Earth's land surface; they are mainly in mountainous regions of the temperate zone. In the absence of precise data, it is estimated that organic soils cover between 2.75 and 4.4 million km² (Yu et al., 2010).

Tundra

Organic soils with both follic and histic horizons are abundant in the subarctic climate zone, namely the tundra which is a biome characterized by open landscapes where tree growth is inhibited by low temperatures. It occurs in high latitudes beyond the northern tree line and in Patagonia and the Falkland Islands. The vegetation is composed mainly of mosses, lichens, grasses and

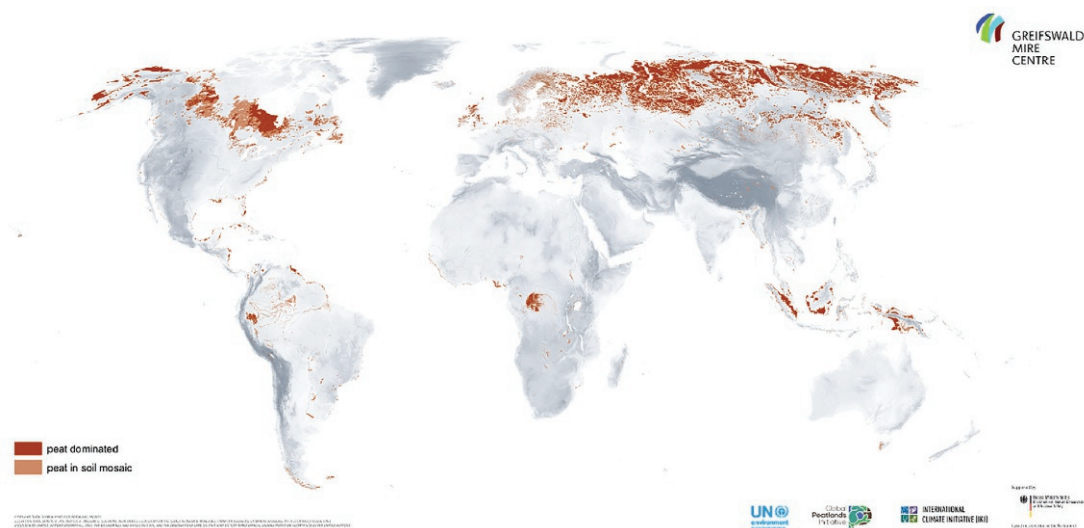


Fig. 4 Global distribution of peatlands (Based on data from the Global Peatland Database/Greifswald Mire Centre (2021)). The distribution of Follic Histosols is not shown due to their small-scale occurrence and the lack of data. Greifswald Mire Centre <https://greifswaldmoor.de/global-peatland-database-en.html>.

shrubs. The tundra is dominated by permafrost which also makes it impossible for trees to grow. Melting of the upper part of the permafrost in the summer months causes waterlogging in the soil above the frozen ground, which can lead to the formation of shallow peatlands. These polar peatlands are characterized by freeze-thaw cycles and thus form patterned ground, such as polygon mires or palsa mires (Fig. 5 (B)). In other parts of the tundra, however, organic soils have formed without water saturation because of the very slow decomposition of organic matter under low temperatures.

Taiga

The taiga is also referred to as boreal forest and covers large areas of the northern hemisphere. It is particularly abundant in Scandinavia, Russia, the Baltic countries, Canada and Alaska. The taiga is also affected by permafrost and freeze-thaw cycles, which form large mire complexes covering vast areas. The proportion of peatlands in the boreal zone is greater than in any other biome on earth, covering 15% of the boreal land area (Xu et al., 2018).

Temperate zone

In the temperate zone, organic soils occur largely as peatlands and form where sufficient precipitation or high groundwater is available. In oceanic regions such as Ireland or Great Britain, peat soils cover entire landscapes, regardless of their relief (Fig. 5 (E)). These 'blanket bogs' are formed by high precipitation throughout the year and the relatively cool temperatures. They are relatively shallow and develop mainly from grasses, dwarf shrubs and mosses.

In temperate, western European regions like north-west Germany, the Netherlands and Denmark, rain-fed bogs were predominant and widely distributed (Fig. 5C). Nowadays, their occurrence is rather fragmentary as many of the bogs have been drained and destroyed over the last century. Many of these bogs developed after the last ice age in natural depressions and dead ice holes through terrestrialization, where bog peat has developed from fen peat and the influence of groundwater. In addition to bogs, fens are widespread; they form mainly on river floodplains (Fig. 5 (D)) and in terrestrialization zones of lakes. Some local fens form near springs. The widest distribution of peatlands in the temperate zone is on the flat areas of northwest Europe and southern Canada, mainly Québec. Peatlands also form in low mountain ranges and occur on slopes as well as on plateaus and in valleys. There is also a wide distribution of peatlands in the foothills of the Alps in the natural depressions formed during the ice ages.

Tropics

Tropical histic organic soils account for about 16% of all organic soils (Page et al., 2011). Most of the tropical peatlands are in South East Asia, namely Indonesia and Malaysia. However, they also occur in Africa and South and Central America, and the area of Amazonian peatlands is estimated to be similar to or even larger than those in Indonesia (Ribeiro et al., 2021). There is still a scarcity of data, however, on tropical peatlands.

The largest areas of peat soils are in coastal lowlands where they are formed mainly by tree remains, such as branches, roots and trunks, and can reach maximum depths of 40 m. Under high temperature and evapotranspiration conditions, the high rainfall leads to sufficient moisture for peat development. Thus, plant productivity is large combined with slow decomposition through

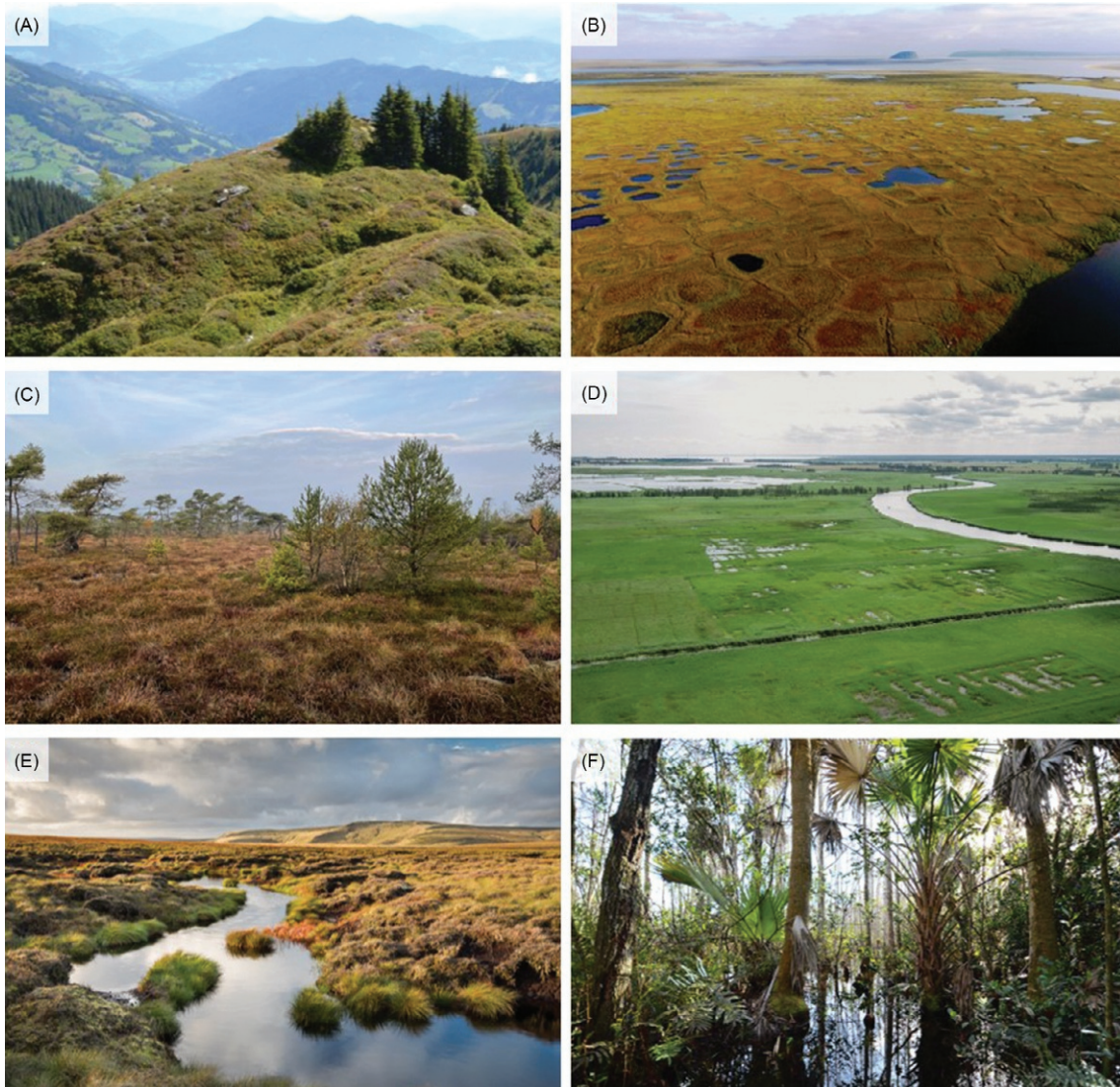


Fig. 5 Landscapes with organic soils: (A) Mountainous region with Folic Histosols (© Gerhard Milbert, Kuratorium Boden des Jahres), (B) Polygon mire in the tundra (© Sebastian Zubrzycki), (C) Temperate raised bog (© David Horlemann), (D) River valley fen (© Tobias Dahms/AESA aerial), (E) Blanket bog (© Moors for the Future Partnership) and (F) Tropical peat swamp forest (© Alec Batts). (A) Kuratorium Boden des Jahres <https://www.bmel.de/DE/themen/landwirtschaft/pflanzenbau/bodenschutz/boden2018.html>; (B) Dr. Sebastian Zubrzycki <https://blogs.egu.eu/network/geosphere/files/2015/08/785fa58f960cddb9509ed4413609cd56.jpg>; (C) David Horlemann; (D) Stephan Busse, Tobias Dahms/AESA aerial <https://www.uni-greifswald.de/en/university/information/current-news/details/n/wiedervernaesste-moore-sind-neuartige-oekosysteme/>; (E) <https://twitter.com/moorsforfuture/status/1369994766369255424>; (F) Ramsar

water-saturated soil. Under these conditions, tropical peat swamp forests form (Fig. 5 (F)), which are mainly dome-shaped bogs that rise above the influence of groundwater. Their hydraulic conductivity is higher than in temperate raised bogs, which makes them more vulnerable to drainage because they are likely to collapse after the water is removed.

Mountain regions

Organic soils are also formed in mountain regions at all latitudes, e.g., the Tibetan plateau, the Alps or the Andean highlands. Under cold temperatures but aerobic conditions, Folic Histosols can form and cover slopes (Fig. 5 (A)), protecting them from erosion. Biological activity is decreased by the cold temperatures, and different organic horizons form with more or less easily identifiable boundaries between Oi ('fibric,' organic horizon with a low degree of organic matter decomposition; plant structures are still intact), Oe ('hemic,' organic horizon with a moderate degree of organic matter decomposition; only fragments of plant residues left, basic morphology, however, still intact), and Oa horizon ('sapric,' 'humified' organic matter, with high degree of decomposition, usually

blackish in color, no macroscopic identification of plant structures possible). Once stored on the cold bedrock, radiocarbon ages may easily exceed several hundreds of years. Nevertheless, even after 1500 years Wang et al. (2019), for instance, failed to discover new organic forms of organic P species, thus questioning a significant role of the humification processes but rather supporting the idea of temperature-controlled preservation and slow microbial synthesis contributing to C and P accrual. This accrual was probably supported by the stabilizing effects of bridging calcium (Ca), stemming from calcareous parent materials. The spatial extent of these soils, however, can be small and isolated, e.g., in cracks between rock fragments but they can also cover several hectares.

Increased rainfall in the mountains through orographic precipitation leads to increased water availability, which, combined with the lower temperatures, enables the formation of histic horizons. Peat soils are formed in valleys and on slopes, but also in the headwaters of rivers and as mires around springs. In mountainous regions, these peat soils are often shallow through the low biomass production at high altitudes. Their distribution and condition are as yet incompletely studied because organic soils are often considered to be mineral soils and used as pastures. Furthermore, they are often in remote areas, which hampers their investigation and mapping.

Characteristics and properties

Organic soils have several properties that distinguish them from mineral soils. One of the most striking is their large C reservoir. By definition, they contain more than 12% C but that often exceeds 50%. Their large organic matter content leads to very low bulk densities which range from 0.01 to 0.2 g cm⁻³ (Rezanezhad et al., 2016), but which increase naturally with increasing mineral content, with soil use and drainage and subsequent compaction (Liu et al., 2022).

Folic Histosols

Organic soils with a folic horizon are freely drained but they have high water retention which promotes organic C accrual. Less decomposed organic matter can hold more water due to larger pore space than strongly decomposed organic matter. Consequently, Oa horizons with sapric organic matter have lower water content and water availability is also less because the small pores of highly decomposed organic matter hardly drain at low suction. The Oi horizons with weakly decomposed organic matter, on the other hand, have good water availability for plants and a higher water holding capacity.

Litterfall decomposition is low and consequently, the organic matter has long turnover times. The nutrient availability is generally low. The concentration and availability of phosphorus (P) is small in Folic Histosols and C:P ratios can exceed 1000 in the upper 50 cm. Further, mobile iron (Fe) and manganese (Mn) have small concentrations in folic horizons, compared to adjacent mineral soils. Nitrogen (N) mineralization is rather small with the inhibited microbial activity. Folic Histosols contain on average 1.6% N and have a C:N ratio of 20–40 which decreases with ongoing decomposition (Vaughan and McDaniel, 2009). With enhanced decomposition, the contents of easily decomposable carbohydrates decrease and also those of lignin, fats and waxes.

The pH depends on the plant community and bedrock, but Folic Histosols have a rather low pH because of the input of acidic litter from coniferous plants and heathers with values ranging from 2.9 to 5. The base saturation (BS) of Folic Histosols is generally low, except for Folic Histosols with a high mineral content or aeolian input, or for those overlying calcareous bedrock. The cation exchange capacity (CEC) is on average 100 cmol_c kg⁻¹, while greater decomposition leads to a higher CEC. The anion exchange capacity (AEC) is generally low at 16 cmol_c kg⁻¹. Both, CEC and AEC, increase with increasing C content (Vaughan and McDaniel, 2009; Table 2).

Table 2 Main properties of the three main types of Histosols.

	<i>Folic Histosols</i>	<i>Rheic Histosols</i>	<i>Ombric Histosols</i>
Genesis	Cold temperatures	Water saturation (ground and surface water)	Water saturation (rainwater)
Vegetation	Grasses, mosses, dwarf trees	Sedges, reeds, alder trees	Mosses, heathers
pH	3.5–5	>4	<4
C:N ratio	20–40	10–33	>33
Cation exchange capacity (CEC)	100 cmol _c kg ⁻¹	80–200 cmol _c kg ⁻¹	70–80 cmol _c kg ⁻¹
Trophic conditions	Mesotrophic to oligotrophic	Eutrophic to mesotrophic	Oligotrophic
Nutrient availability	High N stock Low N availability Low P and K stock and availability Low in Fe, Mn	Moderately to well supplied with nutrients and base cations; depending on groundwater composition	High N stock Low N availability Low P and K stock and availability Low in Ca, Cu, Fe, Mn

Peat soils

Fens and bogs have differences in their properties due to their different water sources. Other properties are, however, very similar and apply to both types of peatland, such as their great affinity to take up and hold water. The water content at saturated conditions ranges from 86% to 94% and undecomposed peat can hold up to 20–30 times its own weight of water. The large water contents make them susceptible to instability, low bearing capacity and shear strength. In general, organic soils are soft and highly compressible, with sinking and slip failure resulting from loads (Fig. 8). Drainage is a way to increase the bearing capacity of peat soils, but this also leads to considerable shrinkage of the soil. Depending on the degree of decomposition, histic horizons can shrink 70% or more upon drying and have a high porosity that can cause irreversible physical collapse.

The peat body can be divided into at least two parts, which differ in the degree of decomposition and in the availability of oxygen. The upper part of the peat body, which consists mainly of living plants and is temporarily aerated, is called acrotelm. The catotelm is the lower part of the peat body, which is permanently anaerobic and consists of dead plant material (Clymo, 1984). This concept has been widely accepted, but also criticized for simplifying the hydrological processes. The boundary between the two layers is not sharp but continuous, and the acrotelm may have developed differently, for example, it depends more on vegetation structure and is usually shallower in fens (Morris et al., 2011). However, some general properties of peat bodies can be derived from the degree of decomposition, such as hydraulic conductivity and composition of the organic matter that usually also change with depth.

The upper, weakly decomposed peat has a high fiber content and the pores are large, open and interconnected, leading to a high hydraulic conductivity, low tortuosity and low flow resistance (Fig. 6). These large pores represent the preferential flow path of solutes, water and colloids (Rezanezhad et al., 2010). This weakly decomposed peat is highly compressible and due to its high hydraulic conductivity and large active porosity, it can shrink and swell with drying and wetting, enabling a water table close to the surface that is independent of water availability (Rezanezhad et al., 2016). With ongoing decomposition of the plant debris into smaller fragments, the pore space and proportion of large pores are reduced (Fig. 6). Consequently, the active porosity decreases because the decomposed peat consists mainly of small, partially closed, dead-end pores. These pores have low interconnection, resulting in restricted flow with low velocity, a large tortuosity and low compressibility (Rezanezhad et al., 2016). The hydraulic conductivity accordingly decreases from undecomposed, fibric peat to hemic to amorphous, sapric peat.

Decomposition processes in peat occur mainly in its upper part as the water table fluctuates, leading to partly aerobic conditions. There, the peat is weakly decomposed but still undergoes intense and ongoing transformations. Almost the whole decomposition process takes place in the upper part of the peat body, which leads to the release of humic acids. The organic matter is dominated by labile C compounds and oxygenated functionalities, like carbohydrates. The O-substituted organic matter plays an important role in microbial humification and decomposition. Thus, during decomposition and with depth, the O contents decrease while the H and C contents increase. The contents of easily degradable carbohydrates and carboxyl-C also decrease, whereas the proportion of more stable structures, such as aliphatic-C, aromatic-C and microbial mucilage-C increases with ongoing decomposition. In the lower part, which is almost permanently water-saturated, decomposition is very slow with the absence of oxygen. Because minor

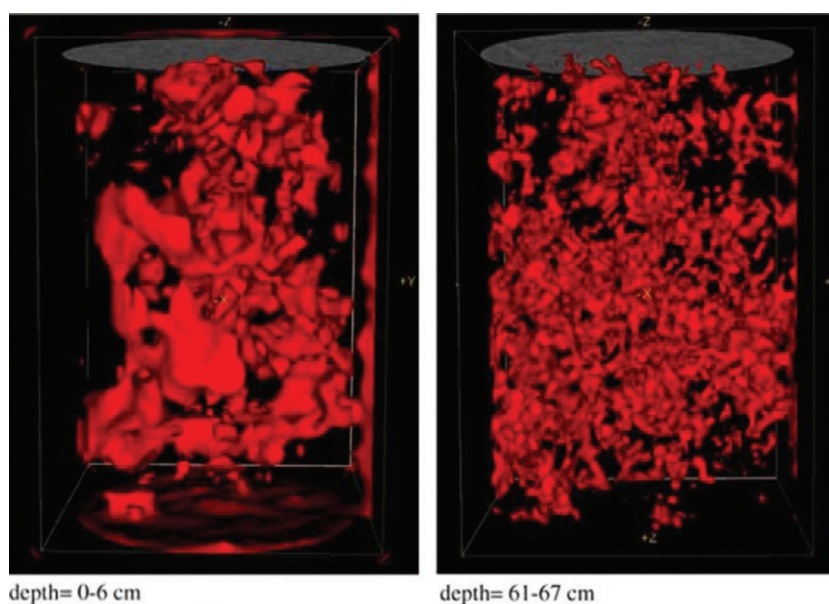


Fig. 6 Visualization of air-filled pores (red) in *Sphagnum* peat from two different depths, taken by X-ray computed tomography. In the upper layer, the pores are interconnected and larger than in the deeper layer. Modified from Rezanezhad F, Quinton WL, Price JS, et al. (2010) Influence of pore size and geometry on peat unsaturated hydraulic conductivity computed from 3D computed tomography image analysis. *Hydrological Processes* 24, 2983–2994. <https://doi.org/10.1002/hyp.7709>.

transformation processes only take place, this part of the peat body is more homogenous and highly decomposed with a higher proportion of aromatic humic materials (Cocozza et al., 2003).

Other peatland properties, however, do not primarily depend on the degree of decomposition but on the source of water. Therefore, properties, such as pH, CEC and nutrient availability, can be distinguished between groundwater-fed fens and rainwater-fed bogs.

Fens (Rheic Histosols)

Fens vary considerably in their characteristics and properties because they are fed by ground and surface water which is influenced by the bedrock and mineral soil with which it has been in contact. Poor and moderately rich fens have pH values between 4 and 6.4, while rich, alkaline fens contain calcium carbonate and have pH values >6.4. The BS and CEC also depend on pH and values increase naturally from poor to rich fens. At lower pH values of 4–6.4, the CEC and BS range from 80 to 120 cmol_c kg⁻¹ and from 35% to 70%, respectively. At pH 7, the CEC reaches 160–200 cmol_c kg⁻¹ and BS is >70% (Table 2). Consequently, alkaline fens have a good supply of cations, especially Ca and magnesium (Mg), whereas poor fens have small concentrations of both cations (Bourbonniere, 2009).

Fens have moderate to good availability of nutrients, such as P, potassium (K) and others. The concentration of K is highly variable depending on the composition of the water. The availability of P depends strongly on pH because it is often bound to Fe at pH <5 and to Ca at pH >6.5 (Bourbonniere, 2009). Yet, Kooijman et al. (2020) showed that Fe-rich fens are also P-rich with a high P-availability, presumably due to weak P sorption to Fe-OM complexes. The Fe-poor fens, however, had small P-contents and low P-availability. In general, fens have greater nutrient availability than bogs because of their connection to groundwater, but it depends on the composition of the groundwater.

Bogs (Ombric Histosols)

The properties of bogs are less variable than those of fens. Bogs are generally acidic with pH values <4 because rain is their sole source of water. *Sphagnum* mosses further acidify their environment as nutrients are taken up by cation exchange and H⁺ ions are released. Their CEC and BS are also lower than in fens with 70–80 cmol_c kg⁻¹ and 5–35%, respectively. Consequently, the concentration and availability of Ca and Mg is low in bogs because the water has not had contact with mineral soil or bedrock (Bourbonniere, 2009; Wang et al., 2015).

Nutrients are generally scarce in bogs. Bogs contain only small amounts of P, which is mainly organically bound, and the concentration of K is similarly low. The C:P and C:K ratios consequently often exceed 1000 in bogs. In the upper peat, the ratios are lower but increase with depth, indicating a fast recycling of P and K at the surface (Wang et al., 2015). Nitrogen contents usually vary between 1% and 2% (Table 2) and in contrast to P, N accumulates in bogs and the C:N ratio decreases with depth. Even though C:N ratios are large, the main limiting nutrient in bogs is frequently P (Wang et al., 2015).

Utilization

The use of organic soils varies mainly with its properties. Follic Histosols may be used for grazing, but in general they are frequently not used, whereas fens and bogs are often used for agriculture and forestry. Worldwide, 75% of all peatlands are still in a pristine condition, but 25% have already been destroyed or degraded anthropogenically with a further 1% degraded annually. Agriculture is responsible for 50% of this degradation and use, while 30% of peatlands have been anthropogenically afforested.

Follic Histosols as well as fens and bogs are commonly used for grazing. Low-density grazing occurs naturally (e.g., by deer) which maintains the natural vegetation, but commercial grazing includes drainage and some fertilization. Consequently, the vegetation changes and organic soils become prone to compaction, erosion and mineralization. The use as pasture is the most common agricultural use of fens and bogs in Europe.

Crop production on fens and bogs requires drainage, fertilization and some liming. In some regions, other techniques are used to make the wet soils suitable for cultivation. In western Germany and the Netherlands, for example, the cultivation of peat soils started in the 16th century and several techniques were used whose remnants can still be found. One practice was to plough up sand from 2 m to 3 m depth to form diagonally layered sand and peat layers, which were partly covered with a mineral layer (Fig. 7 (A)). In Indonesia, for example, the Mega Rice Project turned native peat swamp forests into rice paddies in 1996. The peatlands were drained and deforested on a large scale. Tropical peatlands are now also frequently drained and partly burned for agricultural plantations.

In the tropics and boreal region, fens and bogs are partly naturally forested. In other regions, the temperate raised bogs or blanket bogs are naturally almost treeless. However, many fens and bogs have been artificially afforested, especially in Fennoscandia, Russia and United Kingdom. Afforestation includes drainage, partial ploughing to create a dry environment, and fertilization to overcome the paucity of nutrients.

Other uses are infrastructure, such as streets or wind farms, and peat extraction, each accounting for 10% of the land use. Peat is mainly extracted for horticultural use and fuel. Its extraction mainly occurs on bogs and includes drainage of large peatland areas. Peat is either extracted in dried sods or milled on the surface. The main peat extracting countries are the Baltic countries, Germany and Canada. Peat extraction leaves shallow peat deposits with cropland, grassland, afforestation, abandonment or restoration as common after-uses.



Fig. 7 (A) Cultivation of mixed sand and organic soils in northwest Germany (© Helmut Burghardt), (B) Diminishing organic soil under crop cultivation, Switzerland. The white line indicates the boundary between organic soil and fluvial sediments (© Agroscope (Gabriela Brändle, Urs Zihlmann), LANAT (Andreas Chervet)). (A) H. Burghardt (OFD Niedersachsen), in: *Moore in Schleswig-Holstein. Geschichte – Bedeutung – Schutz*. Published by: Landesamt für Landwirtschaft, Umwelt und ländliche Räume des Landes Schleswig-Holstein (LLUR). ISBN: 978-3-937937-77-9; (B) Agroscope, © Gabriela Brändle <https://houseofswitzerland.org/de/swissstories/umwelt/schweizer-boeden-schaetze-die-es-zu-bewahren-gilt>.

The use of peatlands varies greatly from region to region. In countries and regions with a high population density such as Western Europe, up to 99% of the peatlands have already been degraded, whereas in large parts of Russia or Northern Scandinavia the peatlands are still largely pristine.

Consequences of anthropogenic use

Organic soils naturally have limited capacity for anthropogenic use, due to high acidity, low nutrient availability and low bulk density, and the vulnerability of organic matter to decompose when aerated. Specific measures may be required such as covering or mixing with sand to reduce wind erosion and risks of fire. When drained, peat shrinks and compacts and aerobic conditions promote oxidative decay of the organic material. Consequently, shrinkage, consolidation and mineralization lead to a loss of height of several centimeters per year, which is further enhanced by wind and water erosion, and possibly burning. The rate of subsidence of peat soils is highest in the first year of up to 1 m and remains at 1–2 cm a⁻¹ in temperate climates and at 3–5 cm a⁻¹ in the tropics in subsequent years (Wösten et al., 1997). In England, for example, a drained peat soil subsided 3.9 m in 130 years, while in Malaysia 2.8 m subsided in 28 years (Hutchinson, 1980; Wösten et al., 1997). Consequently, several meters of peat soil can be lost within a few decades and as a consequence of sustained drainage, the bedrock can come to the surface (Fig. 7 (B)).

Subsidence by mineralization and physical compaction poses additional risks for infrastructure. A famous example occurred in 2017 in the Eastern part of Germany, where highways collapsed on subsiding peatlands (Fig. 8).

Changes in water regime are largely irreversible, as the hydraulic and physical properties change. Consolidation after drainage leads to a higher bulk density which affects hydraulic conductivity and also the plant-available water. Liu et al. (2022) showed, that the available water content increases with increasing bulk density up to 0.2 g cm⁻³, but decreases with further increasing bulk density. Very strongly decomposed peat under agricultural use has low available water contents in the upper decimeters. The small, inactive pores and low porosity of drained and degraded peat mean that plant-available water is lower than in near-natural, less decomposed peat and the crops are more susceptible to drought.

Secondary pedogenic processes are initiated by artificial drainage and use which alter the physical properties of peat soils. Decreased water table depths (10–80 cm) lead to aeration of the upper decimeters and consequently, enhanced decomposition. Weather-related swelling and shrinking in combination with the activity of soil organisms such as earthworms form a crumbly peat structure. This process of enhanced decomposition, the activity of earthworms, and swelling and shrinking, resulting in a crumbly structure, are summarized together under the term earthification. Ongoing and deeper drainage (>80 cm) leads to further decomposition of the peat where the crumbly structure of earthified peat disintegrates into a fine, grainy partly powdery mass when dry which is at risk of wind erosion. When wet, the peat has a greasy, muddy consistence. However, it can also be hydrophobic and when wetted, the peat does not swell but floats on the water surface. The so-called moorshyfted peat is dark brown to black and a further development of earthified peat. Both processes only occur through intensive anthropogenic use with deep drainage.

In addition to secondary pedogenic processes, the anthropogenic use of peatlands affects greenhouse gas (GHG) emissions. The total rates of emission depend strongly on the type of peatland use and the trophic conditions. In boreal and temperate zones, afforested peatlands emit between 0.9 and 9.5 t CO₂ ha⁻¹ a⁻¹, whereas croplands emit on average 29.0 t CO₂ ha⁻¹ a⁻¹ (IPCC, 2014).



Fig. 8 The highway was built on peat soils and subsided due to the insufficient load bearing capacity of peat soils (© Alexander Müller). Alexander Müller, <https://www.ostsee-zeitung.de/Vorpommern/Greifswald/A20-Loch-voellig-ungeeignete-Gruendung>

Aerobic decomposition is strongly influenced by temperature, and CO_2 emissions in drained tropical peatlands are generally higher than in boreal and temperate regions. Depending on water table depth, tropical plantations on peat soil lead to emissions of up to $86.0 \text{ t CO}_2 \text{ ha}^{-1} \text{ a}^{-1}$ (Hooijer et al., 2010) and tropical croplands emit between 48.0 and $51.3 \text{ t CO}_2 \text{ ha}^{-1} \text{ a}^{-1}$ (Hooijer et al., 2010; IPCC, 2014).

Carbon dioxide (CO_2) accounts for the largest emissions from drained peatlands, but nitrous oxide (N_2O), dissolved organic carbon (DOC) and methane (CH_4) also play an important role in the GHG inventory. Nitrous oxide emissions in pristine peatlands are negligible unless affected by nitrate inputs from adjacent farmland, but drainage, subsequent oxidation and fertilization increase the microbial activity and thus, nitrification processes. The N_2O emissions in boreal and temperate zones range on average from 0.22 to $13.0 \text{ kg N}_2\text{O-N ha}^{-1} \text{ a}^{-1}$, while they are generally higher from nutrient-rich sites (IPCC, 2014). Data for N_2O emissions in drained tropical peatlands are still scarce, but overall are more variable than in boreal peatlands (IPCC, 2014). Under warmer climates, N_2O emissions are higher, but they also depend strongly on N availability in the soil. In general, N_2O emissions from organic soils can be predicted by the C:N ratio of the topsoil; soils with a low C:N ratio have higher N_2O emissions (Pärn et al., 2018).

Through the ongoing decomposition of organic matter in drained peat soils, DOC is frequently released and transported through drainage ditches. Under use, DOC export is increased on average by 60% and ranges from 20 to $110 \text{ mg DOC L}^{-1}$ (IPCC, 2014), but several studies have reported much higher values (e.g., Frank et al., 2014). Increased microbial activity through aeration and fertilization is a main factor in the large export of DOC. Dissolved organic carbon is the major component of waterborne C loss in peat soils and also an important factor for further CO_2 and CH_4 emissions from drainage ditches. When drained, peat soils emit negligible amounts of CH_4 or even act as small CH_4 sinks through the uptake by plants. Nevertheless, drainage ditches can form new CH_4 hotspots and contribute to the overall GHG balance of drained peatlands (Köhn et al., 2021).

Anthropogenic change through drainage and forest clearance, and prolonged droughts through global warming increase the risk of peat fires, which contribute to the GHG balance. Peatlands are usually naturally protected against fire by their hydrological self-regulation. However, a deeper water table through drainage exposes the dried, easily flammable topsoil to an enhanced risk of fire. Burning peat is characterized by smoldering combustion, a flameless form of combustion that can persist under cold temperatures, high moisture contents and low O_2 concentrations and can continue over several months. In Indonesia, for instance, major peat fires were intensified by land-use changes and drainage of the peatlands. These large peat fires contribute considerably to the GHG balance of peatlands in certain years. However, exact estimates of GHG emissions from burning peat are difficult to determine because of the often unknown depth of the fire.

The data on emissions from Follic Histosols are scarce because these soils tend to occur on a small scale, are often considered as mineral soil and are mainly used as pasture. There are also inaccuracies in their classification due to different classification systems in different countries. However, Eickenscheidt et al. (2015) showed that the emission factors from C-rich soils can be attributed to land-use type rather than C content, indicating that degradation and use of Follic Histosols can lead to similarly large emissions as the anthropogenic use of peatlands.

Restoration and sustainable use of organic soils

To prevent organic matter losses, related nutrient release and GHG emissions, organic soils should be preserved in their pristine state. This is often no longer possible, therefore the rewetting of degraded peatland sites is recommended (Günther et al., 2020; IPCC, 2014). Rewetting by drain-blocking can restore a high water table and the peatlands' typical vegetation. Some peatlands have undergone partly irreversible changes due to degradation, e.g., in their bulk density, nutrient status and C stability, which still differ from natural peatlands after decades of rewetting (Kreyling et al., 2021; Schimmel et al., 2021).

With rewetting, however, rates of decomposition slow down, promptly reducing CO₂ and N₂O emissions, and DOC loss through drainage waters (Wilson et al., 2016). The direct reduction in GHGs through rewetting ranges between 8 t CO₂-eq ha⁻¹ a⁻¹ in nutrient-poor forestry and 26 t CO₂-eq ha⁻¹ a⁻¹ on former cropland (Wilson et al., 2016). After several years of rewetting, organic matter may re-accumulate, leading to a net C sequestration of 0.12 (±0.40) t C ha⁻¹ a⁻¹, which depends strongly on site and climatic conditions. Boreal peatlands have the greatest potential for C sequestration with rewetting (IPCC, 2014; Wilson et al., 2016). The rewetting of peatlands is an important measure in GHG reduction even if CH₄ emissions increase (Günther et al., 2020).

Rewetting and subsequent anaerobic conditions, nevertheless, favor methanogenesis and CH₄ emissions can increase up to 40-fold (mean: 8.0 ± 10.5; IPCC, 2014; Wilson et al., 2016). There are methods to reduce CH₄ emissions from rewetted sites. First, the water table should be raised gradually and stepwise to avoid inundation in order to reduce CH₄ emissions. Second, the removal of drained and fertilized topsoil prior to rewetting reduces CH₄ emissions by a factor of 30–400 because these are mainly caused by large nutrient concentrations in the peat and decaying labile plant material (Huth et al., 2020).

To rewet peatlands, former drainage ditches need to be refilled. Material is needed for this and if this comes from the soil surface CH₄ emissions are concentrated in a smaller area and reduced overall; peatland restoration goes hand in hand with climate change mitigation. Topsoil removal also reduces the eutrophication of adjacent freshwater and marine ecosystems (Zak et al., 2018). With P mobilization and denitrification of nitrates on rewetting, the large nutrient load in surrounding waters can be an unwanted side effect of peatland restoration (van de Riet et al., 2013). The extent of P release depends on the degree of peat decomposition and redox potential: the more decomposed is the peat by drainage and use, the higher is the P concentration in the upper peat. With rewetting and subsequent decline in redox potential within a few days, P is mobilized from Fe(III)-oxyhydroxides. Even ammonium shows higher rates of release in highly than in weakly decomposed peat (van de Riet et al., 2013). Even though it seems counterintuitive, the removal of fertile topsoil can help in protecting adjacent water bodies from N and P loads, which is also positive for peatland restoration (Zak et al., 2018).

Many sites that have been used conventionally do not meet the objectives of nature conservation, which primarily aims to preserve near-natural peatlands. These peatlands can still be used economically, but with a peat-conserving and sustainable form of land use that can be achieved by paludiculture. Paludiculture is a responsible management option that allows biomass cultivation in wet conditions on rewetted fen and bog soils. As for rewetted peatlands, the high water contents promote C sequestration and reduce GHG emissions (Günther et al., 2020). Depending on the site conditions, different plants are grown on the rewetted sites. The cultivation of reeds, cattail, sedges and reed canary grass is possible on fens (Fig. 9). Rewetted fen soils can also be used as grazing land for, e.g., water buffalo. In formerly drained swamp forests, alder swamp forests can be established through rewetting,



Fig. 9 Reed harvesting with a caterpillar vehicle on a paludiculture site (© C. Schröder, with permissions from Springer Nature). Wichtmann W, Holsten T, Nordt A, Dahms T, Schröder C (2014) Think rural! Think paludicultural! In: Dünkel F, Herbst M and Schlegel T (eds.), *Think Rural!* Springer Fachmedien Wiesbaden: Wiesbaden, pp. 215–231. doi: 10.1007/978-3-658-03931-8_19.

with subsequent use of the wood. On bog soils, peat mosses can be grown on former extraction sites; *Sphagnum* mosses are the main peat-forming plants in raised bogs when the water table is high enough and can serve as a peat substitute and growing medium for other plants.

Outlook

Organic soils cover only minor areas of the land's surface but play a major role in nature conservation and climate change mitigation. So far, there has been insufficient research on organic soils, especially on Follic Histosols and tropical peatlands, whose distribution, depth, condition and properties are largely unknown. With the valuable functions of organic soils for C and water storage and unique habitat for rare species, more research is needed on sustainable restoration and management options. The large C storage potential and emissions from organic soils indicate an urgent need for rewetting to reduce global greenhouse gas emissions.

Acknowledgment

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—SFB 1502/1-2022—Project number: 450058266.

References

- Bourbonniere RA (2009) Review of water chemistry research in natural and disturbed peatlands. *Canadian Water Resources Journal* 34: 393–414. <https://doi.org/10.4296/cwrj3404393>.
- Clymo RS (1984) The limits to peat bog growth. *Philosophical Transactions of the Royal Society, B: Biological Sciences* 303: 605–654. <https://doi.org/10.1098/rstb.1984.0002>.
- Cocozza C, D'Orazio V, Miano TM, and Shoty W (2003) Characterization of solid and aqueous phases of a peat bog profile using molecular fluorescence spectroscopy, ESR and FT-IR, and comparison with physical properties. *Organic Geochemistry* 34: 49–60. [https://doi.org/10.1016/S0146-6380\(02\)00208-5](https://doi.org/10.1016/S0146-6380(02)00208-5).
- Eickenscheidt T, Heinichen J, and Drösler M (2015) The greenhouse gas balance of a drained fen peatland is mainly controlled by land-use rather than soil organic carbon content. *Biogeosciences* 12: 5161–5184. <https://doi.org/10.5194/bgd-12-5201-2015>.
- Frank S, Tiemeyer B, Gelbrecht J, and Freibauer A (2014) High soil solution carbon and nitrogen concentrations in a drained Atlantic bog are reduced to natural levels by 10 years of rewetting. *Biogeosciences* 11: 2309–2324. <https://doi.org/10.5194/bg-11-2309-2014>.
- Frolking S, Talbot J, Jones MC, et al. (2011) Peatlands in the Earth's 21st century climate system. *Environmental Reviews* 19: 371–396. <https://doi.org/10.1139/a11-014>.
- Günther A, Barthelmes A, Huth V, et al. (2020) Prompt rewetting of drained peatlands reduces climate warming despite methane emissions. *Nature Communications* 11: 1644. <https://doi.org/10.1038/s41467-020-15499-z>.
- Hooijer A, Page S, Canadell JG, et al. (2010) Current and future CO₂ emissions from drained peatlands in Southeast Asia. *Biogeosciences* 7: 1505–1514. <https://doi.org/10.5194/bg-7-1505-2010>.
- Hutchinson JN (1980) The record of peat wastage in the East Anglian Fenlands at Holme Post, 1848–1978 A.D. *The Journal of Ecology* 68: 229. <https://doi.org/10.2307/2259253>.
- Huth V, Günther A, Bartel A, et al. (2020) Topsoil removal reduced in-situ methane emissions in a temperate rewetted bog grassland by a hundredfold. *Science of the Total Environment* 721: 137763. <https://doi.org/10.1016/j.scitotenv.2020.137763>.
- IPCC (2014) *2013 Supplement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories: Wetlands: Methodological Guidance on Lands With Wet and Drained Soils, and Constructed Wetlands for Wastewater Treatment*. Geneva, Switzerland: IPCC.
- IUSS Working Group WRB (2015) *World Reference Base for Soil Resources 2014, Update 2015: International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*. World Soil Resources Reports No. 106 Rome: FAO.
- Köhn D, Welpelo C, Günther A, and Jurasinski G (2021) Drainage ditches contribute considerably to the CH₄ budget of a drained and a rewetted temperate fen. *Wetlands* 41: 71. <https://doi.org/10.1007/s13157-021-01465-y>.
- Kooijman AM, Cusell C, Hedenäs L, et al. (2020) Re-assessment of phosphorus availability in fens with varying contents of iron and calcium. *Plant and Soil* 447: 219–239. <https://doi.org/10.1007/s11104-019-04241-4>.
- Kreyling J, Tanneberger F, Jansen F, et al. (2021) Rewetting does not return drained fen peatlands to their old selves. *Nature Communications* 12: 5693. <https://doi.org/10.1038/s41467-021-25619-y>.
- Leifeld J and Menichetti L (2018) The underappreciated potential of peatlands in global climate change mitigation strategies. *Nature Communications* 9: 1071. <https://doi.org/10.1038/s41467-018-03406-6>.
- Liu H, Rezanezhad F, and Lennartz B (2022) Impact of land management on available water capacity and water storage of peatlands. *Geoderma* 406: 115521. <https://doi.org/10.1016/j.geoderma.2021.115521>.
- Morris PJ, Waddington JM, Benscoter BW, and Turetsky MR (2011) Conceptual frameworks in peatland ecohydrology: Looking beyond the two-layered (acrotelm-catotelm) model. *Ecohydrology* 4: 1–11. <https://doi.org/10.1002/eco.191>.
- Page SE, Rieley JO, and Banks CJ (2011) Global and regional importance of the tropical peatland carbon pool: Tropical peatland carbon pool. *Global Change Biology* 17: 798–818. <https://doi.org/10.1111/j.1365-2486.2010.02279.x>.
- Pärn J, Verhoeven JTA, Butterbach-Bahl K, et al. (2018) Nitrogen-rich organic soils under warm well-drained conditions are global nitrous oxide emission hotspots. *Nature Communications* 9: 1135. <https://doi.org/10.1038/s41467-018-03540-1>.
- Rezanezhad F, Quinton WL, Price JS, et al. (2010) Influence of pore size and geometry on peat unsaturated hydraulic conductivity computed from 3D computed tomography image analysis. *Hydrological Processes* 24: 2983–2994. <https://doi.org/10.1002/hyp.7709>.
- Rezanezhad F, Price JS, Quinton WL, et al. (2016) Structure of peat soils and implications for water storage, flow and solute transport: A review update for geochemists. *Chemical Geology* 429: 75–84. <https://doi.org/10.1016/j.chemgeo.2016.03.010>.
- Ribeiro K, Pacheco FS, Ferreira JW, et al. (2021) Tropical peatlands and their contribution to the global carbon cycle and climate change. *Global Change Biology* 27: 489–505. <https://doi.org/10.1111/gcb.15408>.
- Rydin H and Jeglum JK (2006) *The Biology of Peatlands*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198528722.001.0001>.

- Schimmel H, Braun M, Subke J-A, Amelung W, and Bol R (2021) Carbon stability in a Scottish lowland raised bog: Potential legacy effects of historical land use and implications for global change. *Soil Biology and Biochemistry* 154: 108124. <https://doi.org/10.1016/j.soilbio.2020.108124>.
- Soil Survey Staff (1999) *Soil Taxonomy—A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn, U.S. Department of Agriculture Handbook, Natural Resources Conservation Service, 436.
- van de Riet BP, Hefting MM, and Verhoeven JTA (2013) Rewetting drained peat meadows: Risks and benefits in terms of nutrient release and greenhouse gas exchange. *Water, Air, and Soil Pollution* 224: 1440. <https://doi.org/10.1007/s11270-013-1440-5>.
- Vaughan KL and McDaniel PA (2009) Organic soils on basaltic lava flows in a cool, arid environment. *Soil Science Society of America Journal* 73: 1510–1518. <https://doi.org/10.2136/sssaj2008.0257>.
- Wang M, Moore TR, Talbot J, and Riley JL (2015) The stoichiometry of carbon and nutrients in peat formation. *Global Biogeochemical Cycles* 29: 113–121. <https://doi.org/10.1002/2014GB005000>.
- Wang L, Amelung W, Prietzel J, and Willbold S (2019) Transformation of organic phosphorus compounds during 1500 years of organic soil formation in Bavarian Alpine forests—A ³¹P NMR study. *Geoderma* 340: 192–205. <https://doi.org/10.1016/j.geoderma.2019.01.029>.
- Wilson D, Blain D, and Couwenberg J (2016) Greenhouse gas emission factors associated with rewetting of organic soils. *Mires and Peat* 1–28. <https://doi.org/10.19189/MaP.2016.OMB.222>.
- Wösten JHM, Ismail AB, and van Wijk ALM (1997) Peat subsidence and its practical implications: A case study in Malaysia. *Geoderma* 78: 25–36. [https://doi.org/10.1016/S0016-7061\(97\)00013-X](https://doi.org/10.1016/S0016-7061(97)00013-X).
- Xu J, Morris PJ, Liu J, and Holden J (2018) PEATMAP: Refining estimates of global peatland distribution based on a meta-analysis. *Catena* 160: 134–140. <https://doi.org/10.1016/j.catena.2017.09.010>.
- Yu Z, Loisel J, Brosseau DP, Beilman DW, and Hunt SJ (2010) Global peatland dynamics since the Last Glacial Maximum. *Geophysical Research Letters* 37. <https://doi.org/10.1029/2010GL043584>.
- Zak D, Goldammer T, Cabezas A, et al. (2018) Top soil removal reduces water pollution from phosphorus and dissolved organic matter and lowers methane emissions from rewetted peatlands. *Journal of Applied Ecology* 55: 311–320. <https://doi.org/10.1111/1365-2664.12931>.

The pedology of anaerobic soils

John Hollis* and Allan Lilly, The James Hutton Institute, Aberdeen, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction	399
Soil forming factors resulting in anaerobic horizons	399
Climate	399
Relief	399
Type of parent material	400
Synthesis of the key factors for anaerobic soil development	400
Morphology of anaerobic soils	401
Anaerobic soil resulting from saturation by a rising local water table	404
Local and regional groundwater bodies	404
Tidal saltmarshes and mangrove swamps	406
Seepage	406
Anaerobic soils resulting from superficial saturation	406
Superficial saturation resulting from impeded vertical drainage	406
Superficial saturation resulting from excessive rainfall	410
Superficial saturation resulting from paddy cultivation	411
Artificial field drainage	411
Classification of anaerobic soils	413
United States Department of Agriculture (USDA) soil taxonomy	413
Food and Agriculture Organization (FAO)/International Union of Soil Science (IUSS) World Reference Base	414
Geography of anaerobic soils	415
Conclusions	415
References	416

Abstract

This chapter describes how the interaction between the soil forming factors of climate, relief, organisms, and parent material type results in the formation of anaerobic soil layers due to periodic and sustained waterlogging. The interaction is complex but results in three types of waterlogging regime with differing gley morphology: groundwater encroachment into the soil profile; impedance of downward percolating soil water by slowly permeable soil layers and or substrates; superficial saturation, either by large amounts of rainfall in cold, upland areas, or the deliberate, annual flooding of fields for example, during rice cultivation. The morphology of each regime is described and the way this is used in soil classification at the global level summarized.

Key points

- Most anaerobic soils are the result of permanent or periodic/seasonal water saturated soil conditions.
- For anaerobic conditions to develop fully requires at least about 30 consecutive days of saturation in temperate climates but can require fewer days at saturation in tropical areas.
- Interaction between soil forming factors of climate, relief, organisms and parent material give rise to three different types of waterlogging regime.
- The three types of waterlogging regime are associated with different types of gley morphology and soil profile horizon sequences.
- Gley morphology is some combination of greyish, ferruginous, black, greenish or bluish colors resulting from differences in the color and mobility of iron and manganese under aerobic (oxidized) and anaerobic (reduced) conditions.
- Specific differences in patterns of gley morphology identify the processes that result in soil waterlogging.
- The differences can be used in soil classification to differentiate soil types according to the specific soil waterlogging regime responsible for anaerobic conditions.

*Retired.

Introduction

What do we mean by the Pedology of Anaerobic soils? Pedology is the study of pedogenesis—how soils develop their inherent characteristics driven by environmental weathering processes determined by the dynamic interaction between the variables of climate, organisms and relief acting on parent material over time (Jenny, 1941). In the late 19th Century, the Russian geographer Vasily Dokuchaev, was the first to postulate that soils were natural bodies developed in response to various environmental factors acting at the Earth's surface. He recognized that specific types of soil with a similar sequence of horizontally aligned layers, or soil horizons, occur in areas with distinct types of climate and 'climax' vegetation, such as the Eurasian Steppe and Boreal Forests. This became known as the 'Zonal Soil' concept but it soon became clear that these zones were not as well expressed in the southern hemisphere, or in maritime and equatorial zones where the soil pattern appeared more irregular. It was therefore expanded to include 'Intrazonal' and 'Azonal' Soils which are found in areas where factors other than climate and the climax vegetation, have an overriding influence on the soil characteristics. Anaerobic soils are Azonal and this chapter focuses on describing those factors that result in the formation of anaerobic soil layers, the distinctive types of soil morphology associated with them and their global classification and distribution.

Soil forming factors resulting in anaerobic horizons

Anaerobic conditions in soil develop where oxygen becomes depleted, usually because oxygen uptake by soil biota is not replenished quickly enough by atmospheric diffusion into, or through, the soil profile. This is mainly caused by soil waterlogging but can also be the result of human agricultural mis-practices such as compaction by heavy farm machinery, particularly at times when the soil is wet (the issue of compaction is covered elsewhere in the Encyclopedia), and the unjudicial application of organic manure slurries or ammoniacal fertilizers as well as spillages of petroleum hydrocarbons. The latter four causes can be corrected and will not be discussed further. An excess of rainfall over evapotranspiration means that excess precipitation must be redistributed through the soil profile if waterlogging is to be avoided. This is an important soil function governed by the local interaction between climate, relief, and type of parent material. A final factor, organisms in the form of humans, also creates soil waterlogging through deliberate annual flooding of fields during paddy cultivation.

Climate

Whenever there is an excess of rainfall over evapotranspiration there is the potential for soil waterlogging, however briefly that may last. For the development of anaerobic conditions in soil layers however, waterlogging must be long enough to saturate significant parts of the soil matrix, as well as much of the larger, drainable macropore space, for soil biota to sufficiently deplete the zone of oxygen. The timing for this clearly depends on factors such as biological activity, rates of oxygen diffusion and soil drainage. Studies of soils with morphology associated with anaerobic soil conditions suggest that it is likely to require about 30 consecutive days of waterlogging in most years in temperate zones (Hollis, 1989) and somewhat less than this in tropical zones. Climates with the potential for anaerobic conditions to develop must therefore have, on average, at least 20–30 consecutive days in which rainfall exceeds evapotranspiration. Seasonality is another climatic factor to consider. Tropical climates with warm temperatures all year round are likely to have consistently high rates of evapotranspiration. Therefore, rainfall must be greater than evapotranspiration for prolonged periods for significant waterlogging to occur. In contrast, climates at increasingly high latitudes, both north and south of the equator, will have increasingly longer periods where evapotranspiration is seasonally low so that rainfall during these times is more likely to produce excess water within the soil. In summary, the number of consecutive days when rainfall exceeds evapotranspiration and there is no significant soil moisture deficit (termed the 'field capacity period'), is an important determinant for the likely development of anaerobic soil conditions and the resulting land use.

Fig. 1 shows the distribution of the USDA NRCS soil moisture regimes used in their Soil Taxonomy (Soil Survey Staff, 2014). The classes are based on the average annual duration when the soil, within defined limits, is dry, rather than moist or wet. The exact definitions are complicated and combine the seasonality of rainfall and evapotranspiration with soil temperature, but, for simplicity, those areas with Ustic, Udic and Perudic water regimes have climates in which the soil is moist (or wet) in some part for at least 180 cumulative days or at least 90 consecutive days (45 days where the annual soil temperatures are low). The change from Ustic to Udic to Perudic is based on increasing duration of moist (or wet) conditions, with Udic regimes occurring in humid climates with well distributed rainfall and Perudic regimes occurring where rainfall exceeds evapotranspiration in all months of normal years. Soils with these moisture regimes are not necessarily waterlogged for significant periods but do indicate areas where anaerobic conditions could occur. Aridic and Xeric moisture regimes are restricted to areas unlikely to contain soil with anaerobic layers resulting from climatic factors.

Relief

Once rainfall exceeds transpiration, then the excess water must be redistributed through the soil or over its surface. Slope is clearly an important relief factor here as steeper slopes remove excess water more quickly by surface runoff and lateral flow through upper soil

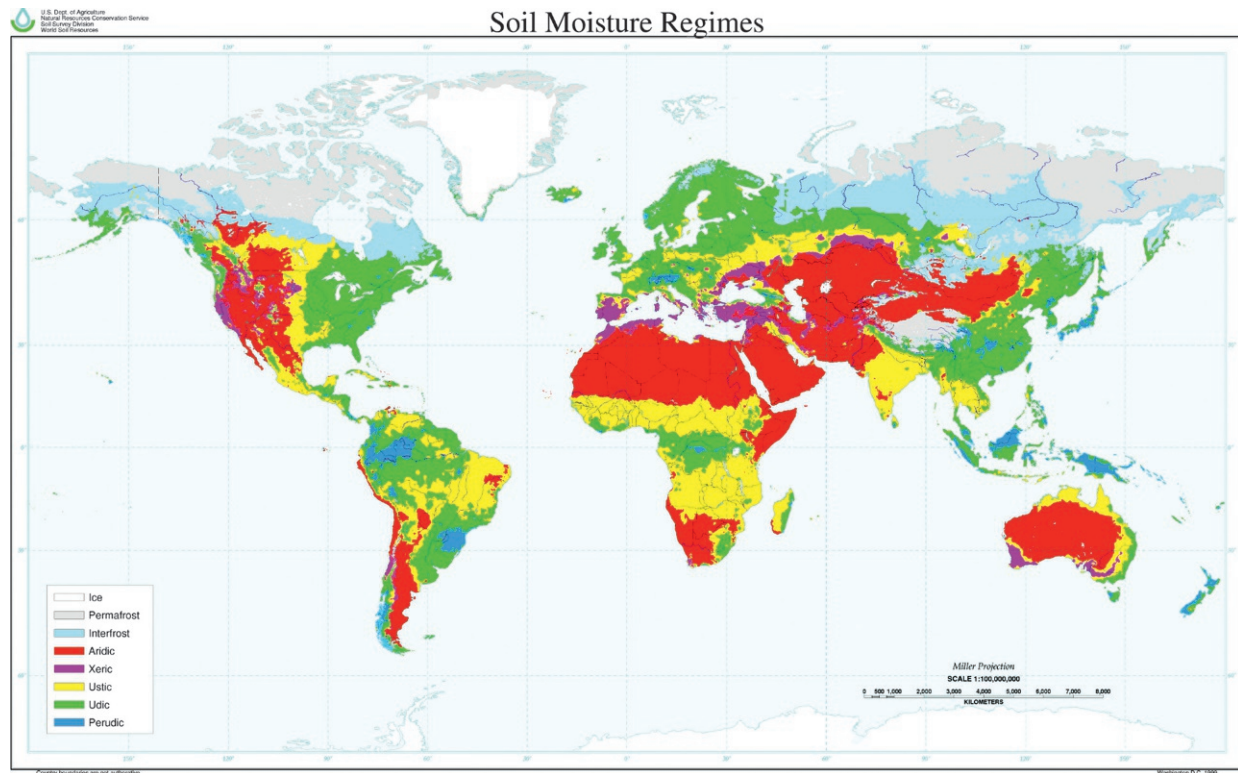


Fig. 1 Global distribution of USDA Soil Taxonomy Soil Moisture Regimes. Courtesy US Department of Agriculture, Natural Resources Conservation Service, Soil Survey Division, World Soil Resources.

horizons. Topography is the other main factor. Footslopes and low-lying ground such as local hollows and basins, especially where enclosed, are likely to receive more excess water via runoff and drainage than slopes, plateaux, and interfluvies. Anaerobic soils are thus more likely to be found on concave footslopes and low-lying, level ground. Nevertheless, on slowly permeable parent materials (see below), level plateau areas may also become seasonally waterlogged and anaerobic, whilst slope angle will influence the duration and depth at which anaerobic conditions develop (Fig. 2).

Type of parent material

Permeability of the soil and its parent material is the final factor determining the likely development of long-term anaerobic soil conditions. Hollis (1989) reviewed various systems for classifying saturated soil hydraulic conductivity and found that most recognized one or more slow categories with conductivities of $<10 \text{ cm day}^{-1}$. Avery (1980) used this value to define the upper limit of a 'slowly permeable layer' that impedes the downward percolation of excess soil water. Soil and substrate layers with saturated hydraulic conductivities of between 10 and about 100 cm day^{-1} are moderately permeable and likely to conduct excess water sufficiently rapidly to avoid significant waterlogging, particularly in climatically drier areas. More permeable soil and substrate layers with saturated hydraulic conductivity $>100 \text{ cm day}^{-1}$ should be permeable enough to allow infiltration of excess soil water without causing any saturation, unless they are waterlogged by rising water from local groundwater bodies or have underlying, less permeable layers.

Synthesis of the key factors for anaerobic soil development

The local interaction between the three main variables of climate, relief and parent material is complex but results in three broad types of soil waterlogging regime (Fig. 3): groundwater encroaching into the soil profile, impedance of downward percolating soil water by slowly permeable soil layers and/or substrates, and superficial saturation either by large amounts of rainfall in cold upland areas or the deliberate annual flooding of fields, for example, during rice paddy cultivation. These two latter conditions both lead to the development of organic-rich topsoils that retain large amounts of water and release it only slowly so that the upper soil layers remain wet and mostly anaerobic.

As with most things related to soil, there are intergrades between all three regimes. However, by looking at trends in each of the three controlling factors, it is possible to illustrate the likelihood of anaerobic conditions developing for each waterlogging regime (Fig. 4).

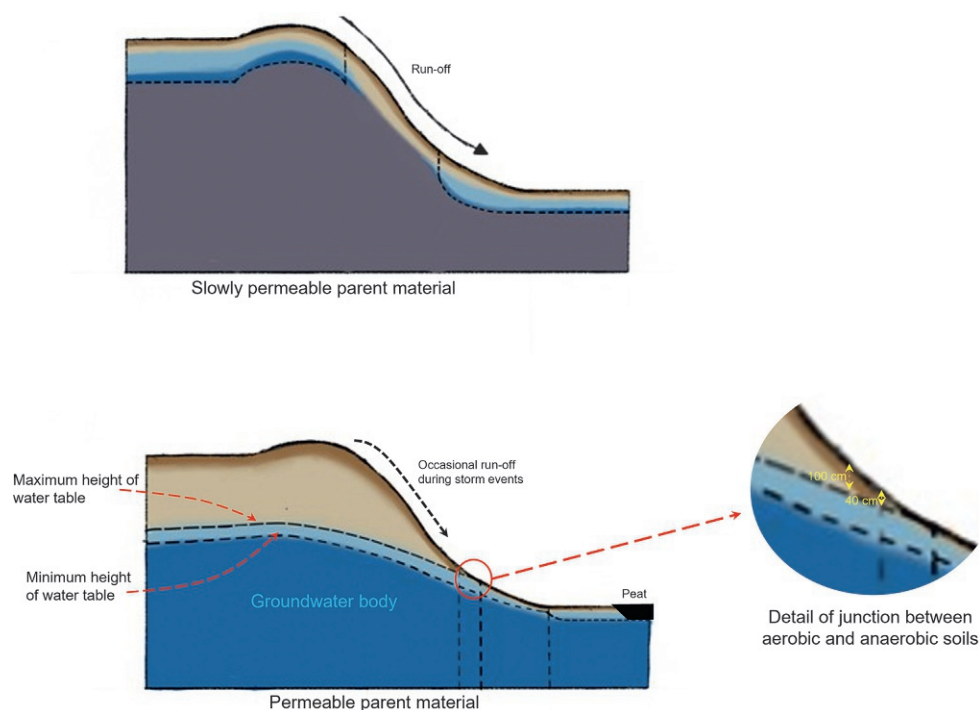


Fig. 2 Influence of relief on the distribution of anaerobic soils with permeable and slowly permeable types of parent material.

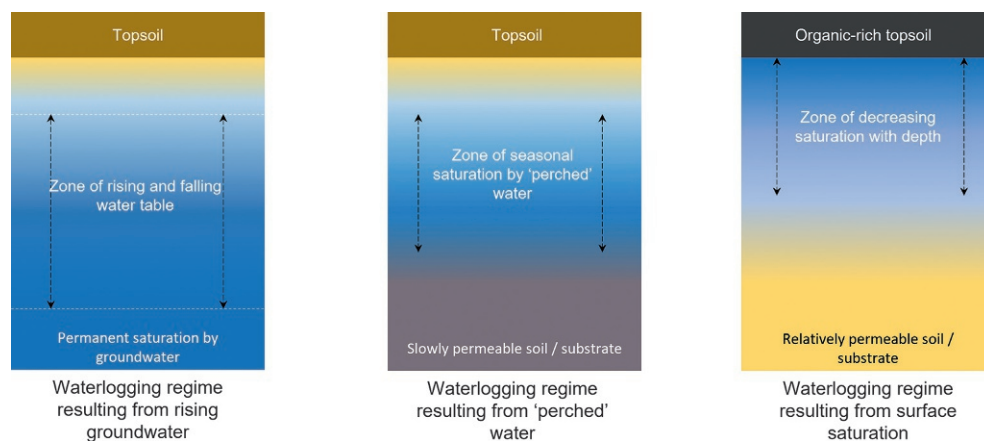


Fig. 3 Three different types of soil water logging regime resulting in anaerobic soil conditions.

Morphology of anaerobic soils

Mineral soil horizons that experience anaerobic conditions for a significant period in most years develop a distinctive color morphology. This is normally termed 'gley morphology' a term derived from the Ukrainian word 'glei' meaning clay, and used to describe a particular type of wet sticky bluish grey clay. At its simplest, gley morphology comprises a variegated appearance of obviously contrasting grey, ferruginous (rusty) and black colors as shown in Fig. 5. Such distinct mottling is caused by the different ionic forms taken by iron and manganese under periodic aerobic (oxidized) and anaerobic (reduced) conditions. In their reduced form, iron (Fe^{2+}) and manganese (Mn^{2+}) become soluble in water and thus mobile in the soil solution. In this form, they are attracted to negatively charged clay particles in the soil matrix, becoming concentrated in specific areas. When these parts of the matrix become re-aerated, the ions take on their oxidized Fe^{3+} and Mn^{4+} forms, precipitating out as iron oxide and manganese oxide segregations, the pre-cursors of the former often being ferrihydrite. In contrast, those areas of the soil matrix from which the reduced Fe and Mn ions have migrated, are depleted of these elements, and take on a pale greyish color. The gleying process is facilitated by the ability of some microbes to survive anaerobic soil conditions by developing metabolic pathways of respiration using compounds other than oxygen, including transition metals such as Fe^{3+} and Mn^{4+} (Lecomte et al., 2018). This ability initiates the reduction of iron and manganese in waterlogged soil layers. These anaerobic soil processes are explained in more detail elsewhere in the Encyclopedia.

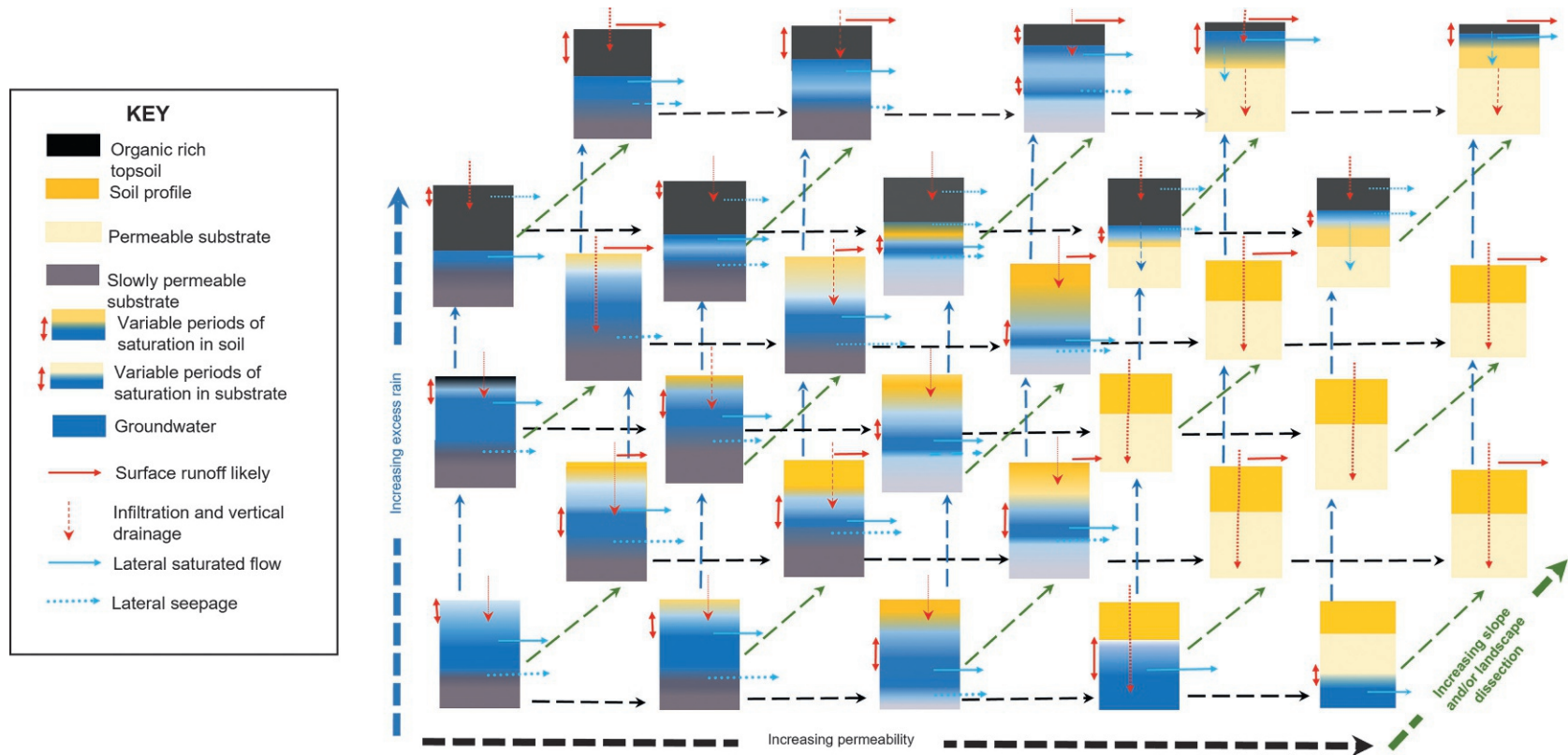


Fig. 4 Relationship between soil waterlogging regimes and trends in excess rain, parent material permeability and relief.

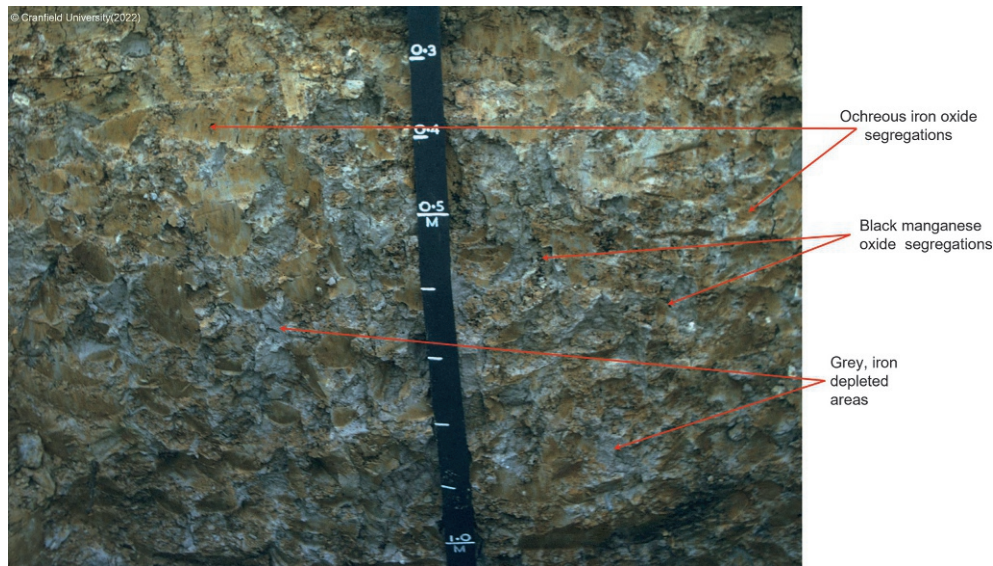


Fig. 5 Common gley morphology resulting from alternating periods of anaerobic and aerobic soil conditions. Courtesy Centre for Environmental and Agricultural Informatics, Cranfield University, College Road, Cranfield, United Kingdom.

Development of this basic gley morphology clearly depends on the weathering of iron compounds in the soil and thus the mineralogy of the soil parent material. Hematite forms of iron are particularly stable and resistant to reduction. Consequently, gley morphology is weakly expressed in soils formed on reddish parent materials characterized by this type of mineralogy (Thomasson and Robson, 1967). Similarly, soils formed in sedimentary parent materials that have small contents of ‘free’ iron (extractable by sodium dithionite) have small amounts of ‘reduceable’ iron (Van Bodegom et al., 2003) and weakly expressed gley morphology. In both cases, the color contrast between parent material colors and any oxidized ferruginous segregations or iron-depleted zones is muted, even where anaerobic conditions are prolonged.

Reduced conditions are effectively permanent where a soil layer never dries out enough to allow air into the soil matrix and, in mineral soils, this is usually indicated by the presence of predominantly greenish grey, greyish green or bluish grey colors indicating the presence of iron in its reduced form, and few, if any, ferruginous mottles (Fig. 6).

This is the basic gley morphology of mineral anaerobic soils, but the exact patterns and relationship of the contrasting colors, both within the matrix and down the soil profile, depend on the different processes of saturation. Such differences are important as they enable us to infer the cause of waterlogging that results in anaerobic conditions within the soil profile.

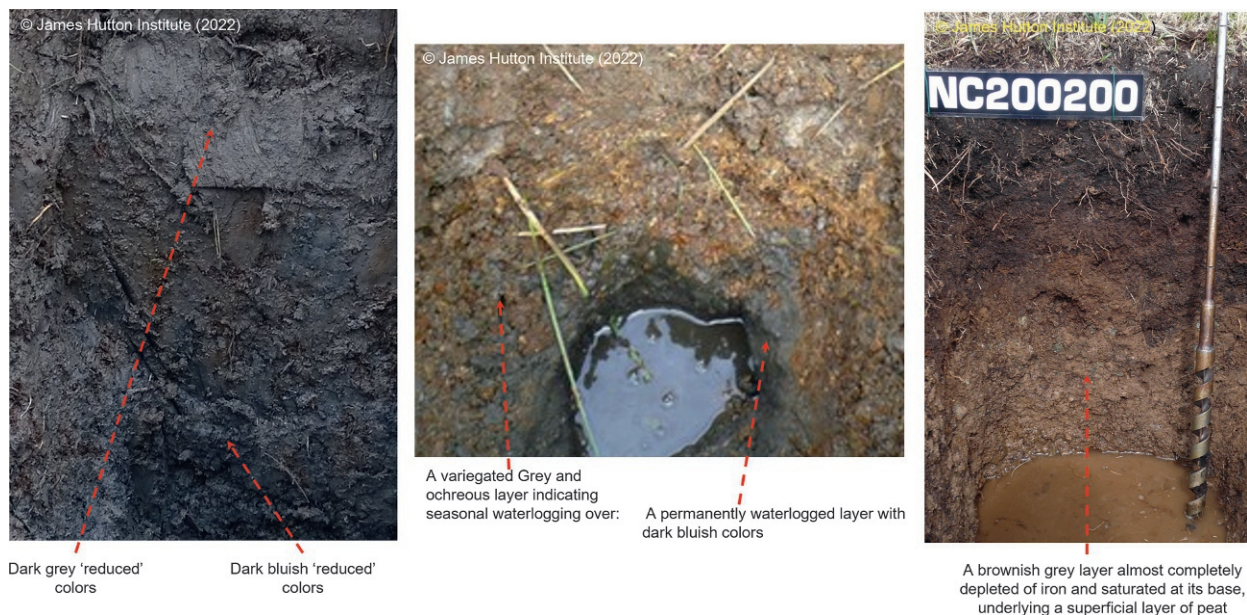


Fig. 6 Different types of gley morphology associated with anaerobic soil conditions. Courtesy James Hutton Institute, Aberdeen, United Kingdom.

Table 1 Horizon Nomenclature used in this chapter.

<i>Master Horizons</i>	<i>Description</i>	<i>Same as Soil Taxonomy?</i>
A	A mineral horizon formed at the surface.	Yes
B	Subsoil mineral layers in which pedogenetic processes have obliterated any original sedimentary bedding or rock structure.	Yes
C	Layers, excluding hard rock that show little effect of pedogenetic processes.	Yes
E	Subsoil layers in which eluvial processes have resulted in the loss of silicate clay, iron, aluminium, or some combination of these.	Yes
H	An organic layer saturated, or formerly saturated with water for prolonged periods	No
W	A waterlogged soil layer	Yes
Suffixes to Master Horizons	Description	Same as Soil Taxonomy?
a	An Organic horizon with highly decomposed organic matter	Yes
d	A dense soil layer that restricts roots	Yes
e	An Organic horizon with intermediate decomposition of organic matter	Yes
g	A mineral horizon with gley morphology indicative of waterlogging resulting from a slowly permeable layer	No
h	A mineral soil layer with accumulation of organic matter	No
i	An Organic horizon with little decomposition and many recognizable fibrous plant remains	Yes
l	A mineral horizon with gley morphology characteristic of the capillary fringe above a (W) groundwater table	No
m	An horizon that is cemented by pedogenetic processes	Yes
o	A mineral horizon with residual accumulation of sesquioxides, including ferruginous segregations	Yes
p	A surface cultivated layer	Yes
r	A dark greyish, bluish or bluish-green soil layer indicating strong reduction and the presence of iron in its reduced state	No
s	A mineral soil layer with illuvial accumulation of sesquioxides	Yes
w	Used with a mineral B horizon to indicate the development of color and/or structure but no pedogenetic features	Yes
x	A compact subsoil horizon with Fragipan characteristics of firmness, brittleness and often platy structure with high bulk density	Yes

Modified from Chapter 5 and Table 85 of FAO (2006) Guidelines for Soil Description. 4th edn. Food and Agriculture Organization of the United Nations: Rome.

The ultimate expression of anaerobic soil conditions is the development and accumulation of superficial, black, or dark reddish brown organic material called peat. This situation occurs where almost continuous anaerobic conditions, often exacerbated by acid or cold conditions, inhibit the biological breakdown of superficial dead vegetative material which starts to accumulate. Where these conditions persist over centuries, peat layers can extend to a thickness of many meters. Different peat horizons within the soil profile can be distinguished by the extent to which their plant remains have been broken down into amorphous organic matter, a process known as humification and, where identifiable, their main constituent plant species.

To differentiate specific types of morphology in individual soil horizons, it is convenient to use a shorthand nomenclature. Whilst a number of these types of nomenclature have been developed, the one used here is that of the Food and Agriculture Organization of the United Nations (FAO, 2006). Most of its notations are similar to those used in the USDA Soil Taxonomy (Soil Survey Staff, 2014) and its Field Handbook for describing and sampling soils (Schoeneberger, et al., 2012). The nomenclature is an alpha-numeric one in which capital letters designate so-called 'Master' horizons or layers, lower case letter suffixes identify specific characteristics of the master horizons, and numbers placed either before or after the letters indicating, respectively, either a lithological change from the overlying horizon, or a sequence of successive horizons with the identical nomenclature. Table 1 summarizes the nomenclature as it applies to anaerobic conditions.

The following sections summarize the specific controlling factors and associated morphology of each of the three waterlogging regimes illustrated in Fig. 3.

Anaerobic soil resulting from saturation by a rising local water table

There are three main situations where a rising water table results in temporally fluctuating anaerobic conditions in the local soil cover. Each situation has soils with different morphology and horizon sequence.

Local and regional groundwater bodies

Groundwater bodies form in geological strata where rainfall infiltrates through the surface layers to a depth where it is unaffected by evaporation or extraction forces exerted by transpiring vegetation. It then continues to percolate through any underlying permeable sediments until it reaches an impermeable layer such as massive clay or hard dense rock where it starts to back up, gradually filling the air spaces in the strata through which it has travelled. Depending on the local topography and the seasonal balance between

rainfall and evapotranspiration, the resulting saturated groundwater body rises near to the land surface creating streams, rivers and wetlands in the lowest parts of the landscape. Such surface water features are often initiated as springs. Significant groundwater bodies only form in relatively permeable material and their upper levels depend on the balance between rainfall and evapotranspiration. When rainfall exceeds evapotranspiration for a significant time, recharge of the water body occurs and its upper surface rises. When the reverse is the case, its upper surface falls. In higher latitudes, both north and south of the equator, the groundwater surface tends to fall during late spring, summer, and early autumn, whereas in tropical latitudes the surface will rise during the rainy season and fall during the dry season.

During the year therefore, there are two zones associated with groundwater bodies, an upper zone that experiences alternating periods of anaerobicity and aeration, and a lower zone that is permanently saturated and anaerobic. Finally, because of surface tension forces in the porous medium of the groundwater body its upper surface is overlain by a thin, superficial capillary zone in which the pores are partially saturated, the extent of saturation decreasing upwards. How far these zones affect the soil profile depends on local topography as illustrated in Fig. 2.

Mineral soils in which full or periodic groundwater saturation rises to immediately below the topsoil can be recognized from their distinctive sequence of soil horizons (see Fig. 7). Under natural or semi-natural vegetation, there is a superficial A or H horizon (see Table 1), usually with a moder or mor humus form (Zanella et al., 2011), depending on the average duration of saturation, or, if farmed and artificially drained, an A horizon with mull humus form. The immediately underlying subsoil is either a Bo horizon showing the mottled grey and ferruginous colors indicating periodic reduction, segregation and re-oxidation of iron resulting from a fluctuating groundwater table, or an E horizon in which periodic saturation and reduction of iron are followed by removal of the reduced iron by leaching, either laterally or vertically. At the base is a BCl or Cl horizon with a mottling pattern characteristic of the capillary zone. Ferruginous iron oxide segregations are concentrated on ped and pore surfaces or around root fibers, indicating those preferential pathways by which air enters the horizon as it de-saturates. In contrast, capillary surface tension means the finer pores remain saturated so that the bulk of the soil remains greyish and anaerobic. This mottled zone passes down into a wet, predominantly greyish, or bluish colored Cr horizon which in its upper parts may also show thin, ferruginous iron oxide segregations lining pores or root fibers.

Anaerobic peat soils that develop because of saturation by a rising local water table are usually restricted to discrete enclosed hollows in the landscape which may contain open water. Here, the vegetation fringing waterbodies comprises water tolerant reed (*Phragmites*), rush (*Juncaceae*) and sedge (*Cyperaceae*) species, the remains of which constitute the main plant remains although woody fragments of trees such as willow (*Salix alba* L.) and alder (*Alnus glutinosa*) may also be present. As the water body fills with sediment and the vegetation spreads, peat accumulates, and occasionally obliterates the open water area. Once the accumulated

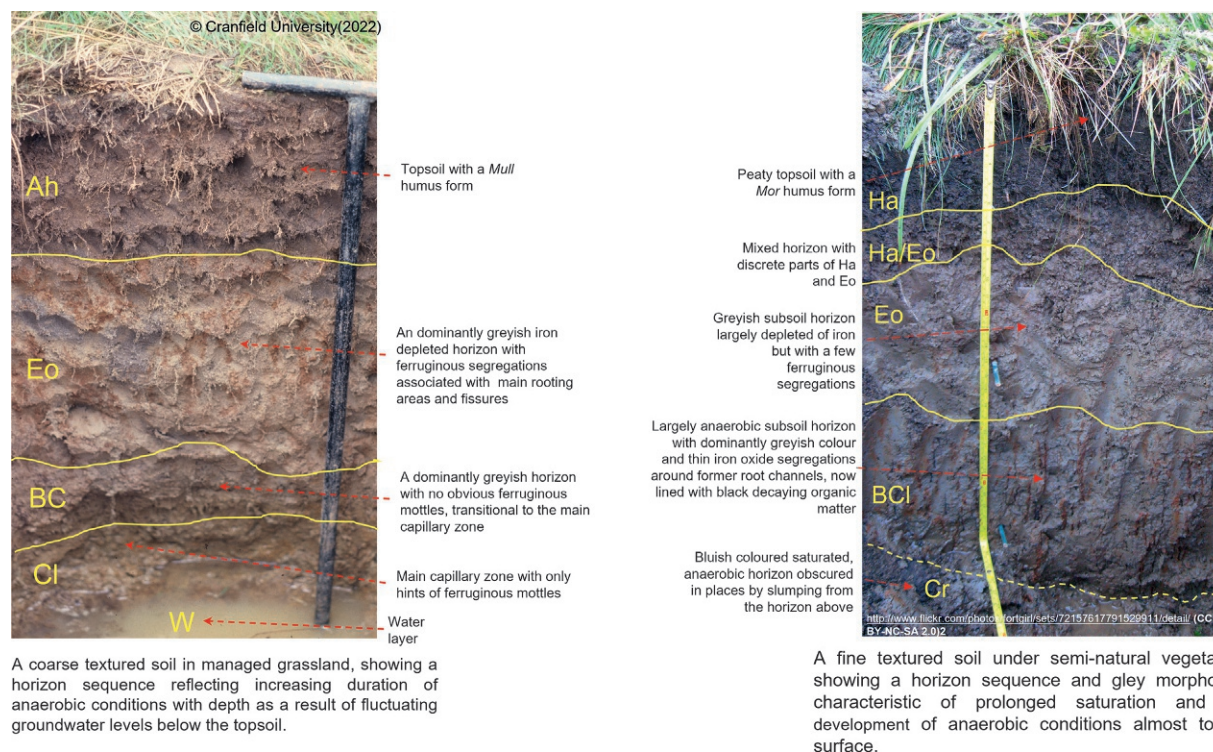


Fig. 7 Gley morphology associated with the capillary zone of two soils under different land use, affected by a rising groundwater table. Courtesy Centre for Environmental and Agricultural Informatics, Cranfield University, Cranfield, United Kingdom; <http://www.flickr.com/photos/fortgirl/sets/72157617791529911/detail/> (CC BY-NC-SA 2.0)2.

peat rises above the level of ground waterlogging, then intermittent aeration affects the highest parts, accelerating peat humification to an extent where a balance between accumulation and humification is reached and peat growth ceases. Where rainfall is large enough, a final phase of peat accumulation occurs. At this stage the saturated upper layers are maintained solely by rainfall and the vegetation starts to be dominated by wet-tolerant species such as *Sphagnum* moss, 'cotton grass' (*Eriophorum* species) and tussock grass (*Poa flabellata*). Continued rainfall-generated growth of peat layers dominated by these plants enables the peat layers to rise above those maintained solely by groundwater saturation and this type of climax lowland peat is known as a 'raised bog', characterized by acid conditions deficient in nutrients, especially nitrogen and phosphorus.

Tidal saltmarshes and mangrove swamps

Coastal fringe areas where low-lying land is saturated by the rising tide have a specialized vegetation and a distinctive soil biomass adapted to the daily inundations. It usually includes marginal marine crustaceans. The inundation cycle means that, although upper horizons are exposed to the atmosphere for some hours, they do not usually become fully aerated, remaining wet but not saturated. The resulting soil profile has a semi terrestrial hydric, organic (H), or organo-mineral (Ah) humus form, the latter usually being massive with very little, if any, structural development because of its permanent wetness. Horizon development below this is indistinct, with a dark, organic (He), or dark greyish to bluish organo-mineral BC, C or Cl horizon with muted dark reddish-brown areas that may represent ferrihydrite formed during the short period of daily exposure to air. Lower parts are effectively an 'unripened' W layer comprising loose sediments that are permanently waterlogged with very low bearing strength (Fig. 8). If the land is reclaimed by isolation from the tidal cycle and installation of land drainage, these layers start to dry out, irreversibly losing some water, becoming more consolidated and starting to develop the classic gley morphology associated with profiles affected by seasonally fluctuating groundwater.

Seepage

Local seepage areas, often termed 'flush' sites, occur on mid- to foot-slopes where there is a significant reduction in the permeability between overlying porous strata and underlying impervious strata. Water bodies that build up over the junction with less pervious material seep out at the surface, causing prolonged periods of saturation following rainfall. Such sites are usually localized and easily identified by the distinctive presence of species tolerant of wet conditions (Fig. 9).

The soil morphology of these sites is conditioned by the extent and duration of local waterlogging, as determined by the rainfall frequency or seasonality, and the relief around the seepage site. Anaerobic soils will only develop at seepage sites that are waterlogged regularly enough to ensure the development of their characteristic vegetation. Those where rainfall is seasonal or infrequent are likely to become aerated for significant periods, at least in their upper parts. Here the upper subsoil horizons usually exhibit the grey and ferruginous mottling characteristic of Bo horizons affected by a fluctuating groundwater table. Similarly, if the local slopes around the site enable some surface run off and or lateral drainage to remove excess water gradually, then the profile again exhibits Bo subsoil horizons. In contrast, where rainfall is more frequent and or local relief is subdued, then saturated conditions persist for much of the time and a slightly different profile morphology develops (see Fig. 9). In cold, acid, usually upland areas that receive large amounts of rainfall, a peaty, semi-fibrous He topsoil horizon with a mor humus form develops. Where slopes allow at least some lateral seepage through the upper layers, the peat is underlain by a predominantly greyish subsoil E horizon, almost completely depleted of iron. This is because leachate from the peat topsoil is acid and contains soluble organic compounds that form complexes with any reduced iron in the horizon, which is then removed during continuous lateral flow. Below this wet grey horizon is a permanently waterlogged Cr horizon where seepage, whether lateral or sub-vertical, is not sufficiently rapid to remove its excess water. Here, greyish colors predominate with a few areas of darker reddish brown iron segregation, possibly composed of ferrihydrite.

Anaerobic soils resulting from superficial saturation

The development of anaerobic conditions caused by superficial saturation within a soil profile depends on the dynamics between excess water input, the saturated vertical and lateral hydraulic conductivity of layers within the soil profile or its immediate substrate and the amount of any downslope surface run-off. Where rainfall frequency and magnitude are such that the rate of excess water moving into the soil exceeds the saturated hydraulic conductivity of any soil or substrate layer, then waterlogging is likely to occur. If this situation continues for a significant period, anaerobic conditions will develop, their duration and vertical extent within the soil profile depending on the duration of saturation. All soils subject to this type of waterlogging regime have a distinct gley morphology in that there is a greater degree of gleying in their upper portions, in contrast to the morphology associated with waterlogging by rising local water tables.

Again, there are three main situations that give rise to this type of waterlogging regime.

Superficial saturation resulting from impeded vertical drainage

As described in the section on soil parent material, soil horizons or substrates with a saturated lateral hydraulic conductivity of less than about 10 cm day^{-1} , are likely to impede the downward percolation of excess water, causing it to accumulate and saturate the impeding layer and, normally, the overlying material. Such impeding horizons are dense and either massive, or with coarse prismatic or blocky peds and less than about 50% total porosity, the latter having a distinctly bi-modal size distribution. A small proportion only of the total porosity comprises coarse pores large enough to allow excess water to drain vertically (at least

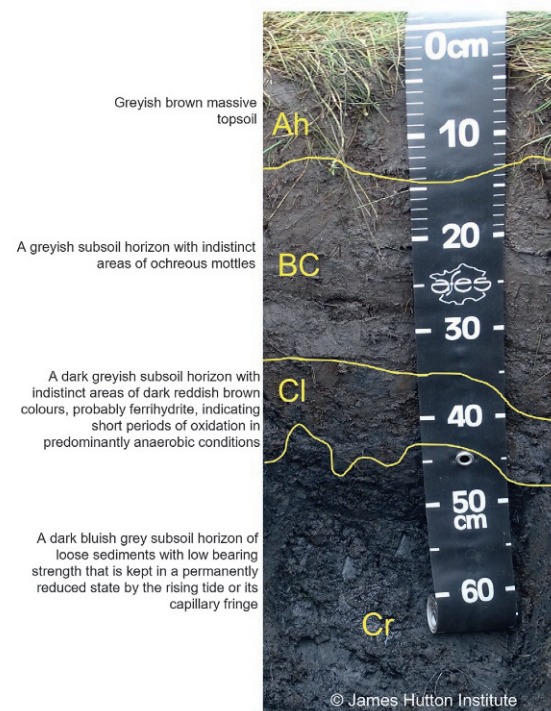


Aerial view Fobbing Marshes, the Thames Estuary Southeast England



Landscape at the Carse of Delnies, North Scotland

Two views of Saltmarsh areas in the UK showing the surface water tidal network and water levels.



A Saltmarsh soil profile showing a diffuse horizon sequence resulting from daily inundation by rising tides and associated anaerobic conditions.

Fig. 8 Landscapes and soil profile of a typical saltmarsh soil. Courtesy Contains OS data © Crown copyright and database rights 2020; Ian R Maxwell, Saltmarsh on the Carst of Delnies/CC BY-SA 2.0 geograph.org.uk; James Hutton Institute, Aberdeen, Scotland, United Kingdom.

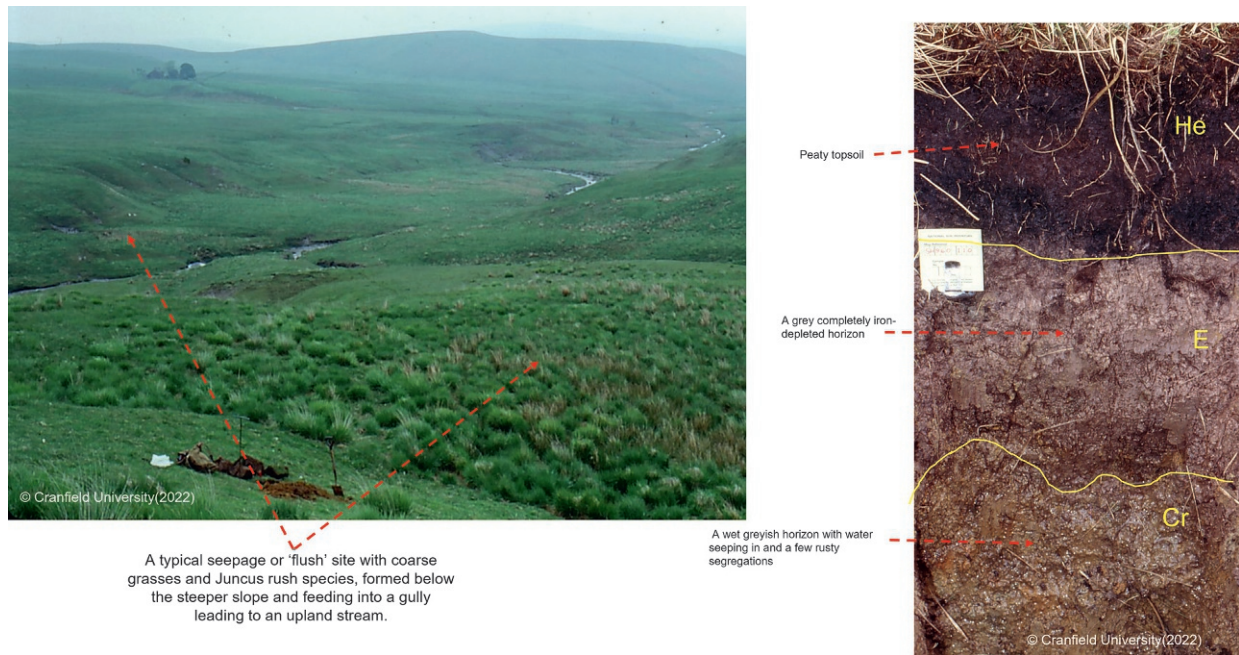


Fig. 9 Landscape and soil profile of a typical seepage site. Courtesy Centre for Environmental and Agricultural Informatics, Cranfield University, Cranfield, United Kingdom.

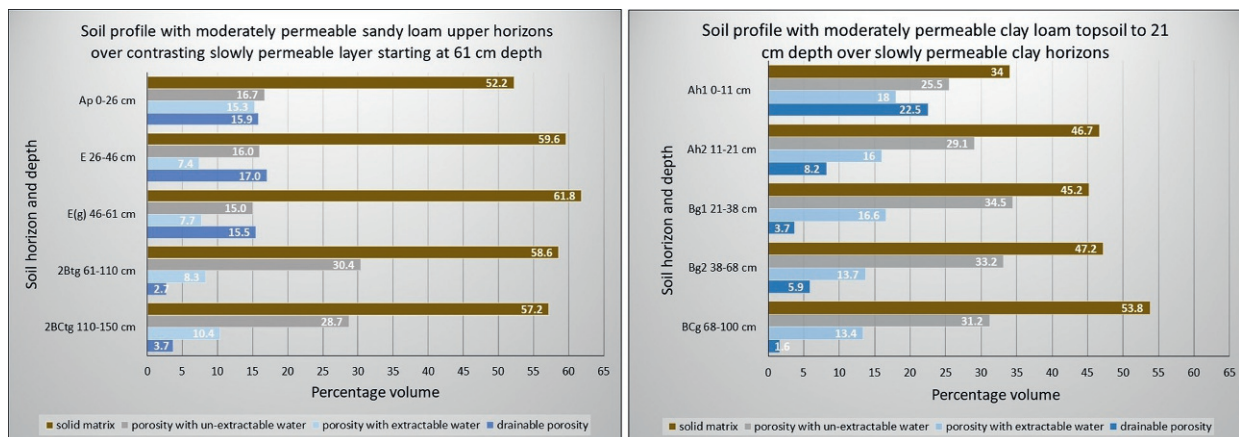
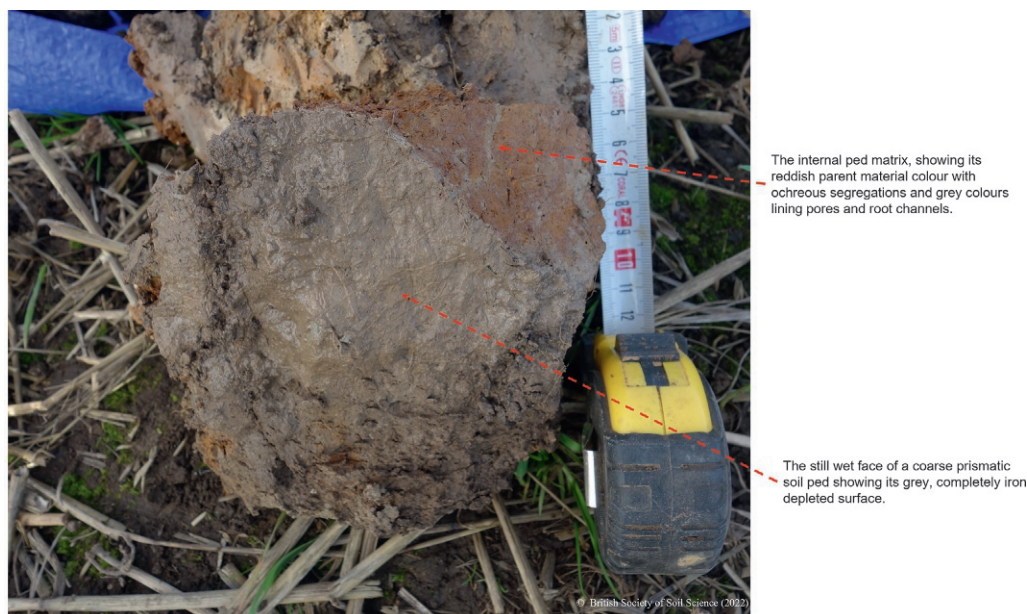


Fig. 10 Porosity and water retention of three profiles with slowly permeable layers starting at different depths.

approximately 0.06 mm average diameter), and these pore sizes usually represent 5% or less of the total volume of the soil layer (Fig. 10). In addition, very small pores which hold water so strongly that plants are unable to extract it, usually comprise at least half of the total porosity.

The characteristics of these slowly permeable horizons give rise to a distinctive gley morphology. As excess water moves into the horizon, the air spaces in the small fraction of coarse pores, the junctions between soil peds and the root channels become saturated relatively quickly. From these areas, soil water diffuses very slowly into the ped matrices. The saturated, coarse pore network quickly becomes anaerobic, reducing any iron on the mineral interfaces. This reduced iron, now mobilized, is either removed by vertical or lateral flow, or diffuses into the ped. Once within the ped, the reduced iron is preferentially attracted to, and retained by, negatively charged clay particles. In this way, the edges of peds, pores and root channels become depleted of iron whilst, within the ped, iron becomes segregated. Once the horizon starts to dry out and aerate, then the iron-depleted ped and pore faces remain grey and the ped interiors take on the characteristic variegated grey and ferruginous segregation resulting from alternating periods of anaerobic and aerobic conditions (Fig. 11). This type of gley morphology is denoted in horizon nomenclature by a g suffix.

Such gley morphology is the basic characteristic of mineral soil profiles with anaerobic conditions resulting from impeded vertical drainage, but there are two distinct soil horizon sequences that can be recognized depending on the duration of saturation



A broken very coarse prismatic ped taken from a slowly permeable horizon about 12 hours after significant rain.

Fig. 11 Detail of the gley morphology associated with a coarse prismatic ped taken from a slowly permeable horizon. Courtesy British Society of Soil Science, Building 42a, Cranfield University, Cranfield, www.soils.org.uk.

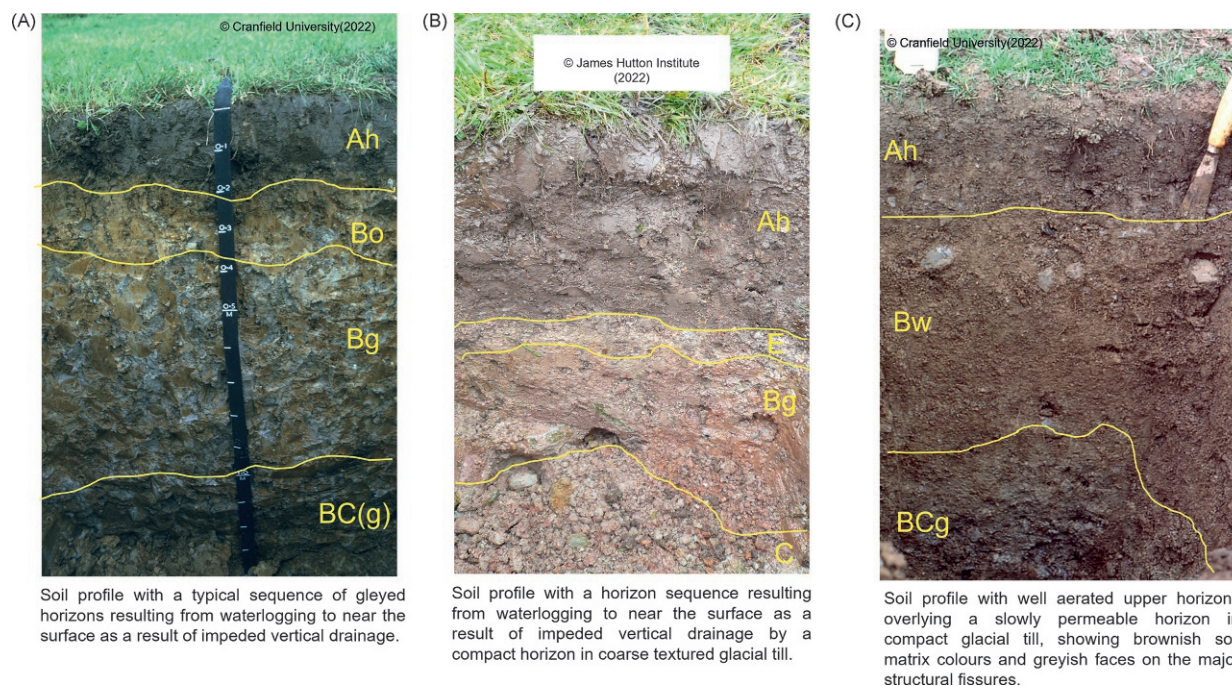
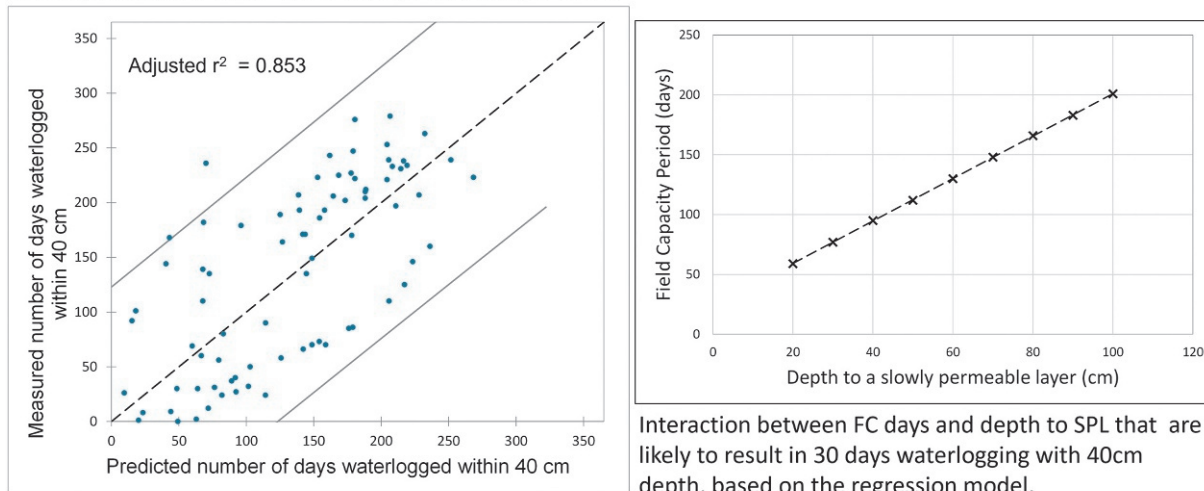


Fig. 12 Three soil profiles with anaerobic gley morphology resulting from impeded vertical drainage. Courtesy James Hutton Institute, Aberdeen, United Kingdom. Centre for Environmental and Agricultural Informatics, Cranfield University, Cranfield, United Kingdom.

that occurs within about 50 cm of the soil surface. Profiles which are saturated within this depth for significant periods have an organo-mineral Ah, or occasionally an organic semi-fibrous He, topsoil horizon, below which is the variegated grey and ferruginous mottled Bo horizon common to many anaerobic soils (Fig. 12, profile a). Where the horizon is more permeable, or the profile is on a significant slope, then there is no Bo horizon and, instead, a wet, iron-depleted E horizon is usually present (Fig. 12 profile b). In either case, these horizons overlie the defining slowly permeable Bg horizon. In contrast to anaerobic soils resulting from a rising groundwater table, the gley morphology of the Bg horizon becomes less well defined with depth, gradually merging into what

Source	Value	Standard error	t	Pr > t	Lower bound (95%)	Upper bound (95%)
Intercept	0.000					
FC days	1.288	0.097	13.293	<0.0001	1.095	1.481
Depth to SPL	-2.291	0.436	-5.254	<0.0001	-3.159	-1.423

No of days wet within 40cm = 1.288*FC days - 2.291*Depth to SPL



Results of the stepwise multiple regression analysis.

Fig. 13 Relationship between field capacity days, depth to slowly permeable layer and duration of soil waterlogging within 40 cm depth.

appears to be a less saturated BC(g) horizon in which structure is often defined by a few grey colored semi-vertical fissures and a matrix that has lost less of its inherent color but has patches of ferruginous iron oxide segregations and fewer greyish colors. Eventually, where deep enough from the surface to be unaffected by weathering processes, the unaltered C horizon is reached. Alternatively, profiles in which there is no prolonged saturation within 50-cm depth have a modified horizon sequence in which the two upper horizons are a mineral Ah horizon over a relatively well-aerated Bw horizon showing little, if any, gley morphology merging downwards into a lower sequence showing some combination of Bo, Bg, BC(g) and C horizons formed below 50 cm depth (Fig. 13 profile c).

The extent to which the upper parts of the soil profile become anaerobic depends on the depth at which the slowly permeable layer occurs, the duration of field capacity and the magnitude of excess rain during this period, although these latter two are usually strongly positively correlated. The data described by Hollis (1989) include measurements of the duration of waterlogging within both 70-cm and 40-cm depths for up to 3 years at a large number of sites across England and Wales. The two recording depths were chosen because they are used in the definitions of soil wetness classes for England and Wales (Hodgson, 1997) and indicated how soils functioned with respect to crop growth. At each site, duration of the field capacity period for each year was recorded, as well as the depth to the slowly permeable layer. These data have been analyzed by stepwise linear regression to examine the influence of the independent variables of field capacity days (FCD) and depth to slowly permeable layer (DSPL) on the duration of waterlogging within 40-cm depth for soils that exhibited gley morphology starting within 100-cm depth. The analysis showed that both FCD and DSPL were highly significant explanatory variables and that the derived regression equation had an adjusted R^2 value of 0.853 (Fig. 13). With this relationship, it is possible to plot the combinations of FCD and DSPL that would give a duration of waterlogging of 30 days which is the probable minimum required to result in the gley morphology associated with anaerobic soil conditions in temperate zones (Fig. 13). This suggests that, for those soils with a slowly permeable layer starting at shallow depth (20–30 cm), anaerobic conditions would be present within 40-cm depth in areas with as few as 60–70 field capacity days. For soils in which a slowly permeable layer did not start until deep in the profile (80–90 cm), anaerobic conditions would only encroach within 40-cm depth in areas with 160–180 days at field capacity.

Superficial saturation resulting from excessive rainfall

In climates where rainfall frequency and magnitude exceed the ability of even moderately permeable superficial soil horizons to transmit the excess water either vertically or laterally, then they tend to become partially or completely waterlogged. Where this situation extends for prolonged periods, anaerobic conditions will develop. Such situations are particularly prevalent in uplands where cold, wet, and acid conditions inhibit the breakdown of superficial organic material, and an organic-rich or a peaty topsoil develops. On the wettest upland plateau areas and basins, peaty H horizons dominated by plant species tolerant of waterlogged conditions, can accumulate to the extent that thick peat soils blanket the landscape.

Where rainfall is slightly less, or the land slopes more, then, even though the largely anaerobic, water-retentive peaty topsoil is present, it overlies mineral subsoil horizons into which its excess water is continually seeping. The uppermost mineral subsoil

horizons are thus variably saturated and only periodically aerated. The distinguishing characteristic of this type of soil is a change from anaerobic conditions (with large reducing capacity) in upper horizons to aerobic conditions (with large oxidizing capacity) in lower horizons.

The wet, acid nature of these types of anaerobic soils means that they often develop spodic horizons in the upper parts of their well-aerated subsoil zone. Such profiles have a typical horizon sequence with a peaty H topsoil of variable thickness and a mor humus form, overlying a greyish, acid Eo horizon with a few ferruginous segregations formed during short periods of aeration. In places this horizon is abruptly terminated by a thin, irregular, dark reddish brown to black cemented Bsm 'iron pan', the lateral coherence of which can be variable, and elsewhere there is a more gradual transition to a spodic Bs or a non-podzolic Bw horizon (Fig. 14). These horizons merge downwards into the relatively unaltered C horizon, usually through a transitional BC.

Superficial saturation resulting from paddy cultivation

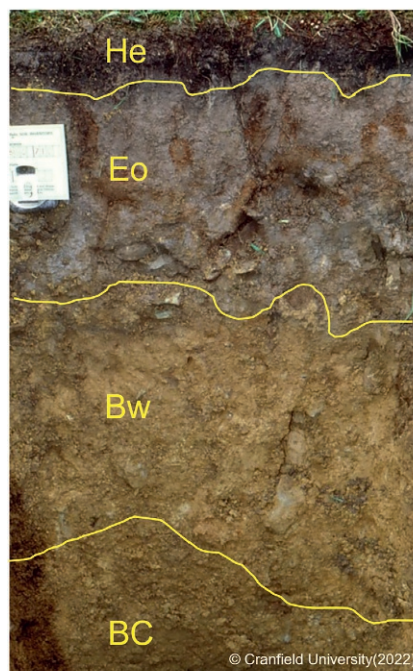
Soils under rice paddy field cultivation probably constitute one of the most extensive types of anaerobic soils in the world. They are subject to a specific anthropic type of waterlogging regime resulting from the deliberate creation of a compacted lower topsoil that is intended to prevent downward percolation of surface water channeled into fields by a system of peripheral ditches and small sluice gates (Fig. 15). Their gley morphology and full horizon sequence takes time to develop following the initiation of paddy cultivation; the exact period varies with the local soil textural and structural characteristics. In many cases it is likely to be more than 100 years (Kögel-Knabner et al., 2010). The characteristic features of the profile comprise three horizons developed by the annual paddy creation process. The first two of these are the result of cultivation when the field surface is partially wet. This creates a compacted 'plough pan' at the base and a superficial wet 'puddled layer'. The latter is a dark-colored, often organic-rich Ap horizon that, when not flooded, has a loose consistence with assorted small-sized peds. Underlying this is the compacted Ad plough pan horizon that impedes downward percolation when the overlying topsoil is waterlogged. It has a massive or platy structure, with a matrix usually showing the paler colors created by prolonged reduced conditions. If ferruginous segregations are present, they occur mainly around root channels or coating any structural fissures. These two layers overlie a structured Bg horizon of variable thickness, depending on the permeability of the soil parent material. Its gley morphology is similar to that of subsoil horizons in soils with superficial saturation resulting from impeded vertical drainage. It is characterized by peds with greyish iron depleted faces and an internal matrix dominated by ferruginous segregations but with greyish-colored pore surfaces often spreading a few millimeters inwards. Where the soil parent material is permeable, the Bg horizon merges downwards into an aerobic BC or C horizon; the change from anaerobic to aerobic conditions is similar to that of soils where superficial saturation results from excessive rainfall. In contrast, profiles with less permeable substrates, show a gradual downwards change from the Bg horizon to a BC horizon in which the greyish and ferruginous colors become less pronounced and the inherent color of the parent material dominates.

Artificial field drainage

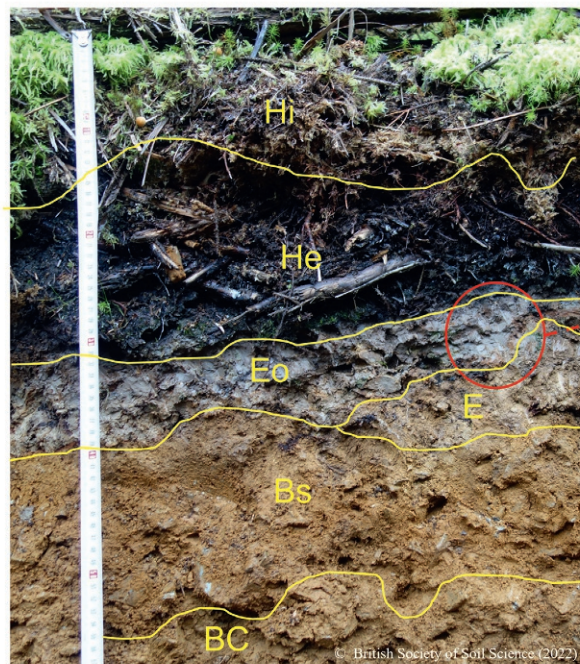
An additional anthropogenic soil forming factor affecting anaerobic soils is the installation of field drainage systems designed to ameliorate waterlogged conditions quickly in the upper parts of the soil to facilitate crop growth. The successful installation of such systems enhances drainage from waterlogged upper layers, aerating them or, at the least, significantly reducing the duration of anaerobic conditions. The installation and management of these artificial drainage systems has been the subject of a considerable amount of research and is covered elsewhere in the Encyclopedia.

In many areas where anaerobic soils result from superficial saturation by excess rainfall, productive agriculture is precluded because of climatic conditions, making the cost of effective artificial drainage uneconomic. Elsewhere, anaerobic soils resulting from rising ground water tables require different types of drainage systems from those resulting from impeded vertical drainage. Where effective, the installation of field drainage increases the aeration of upper soil horizons that were formerly either permanently anaerobic, or experienced alternating anaerobic and aerobic conditions. This is likely to increase microbial activity resulting in increased organic carbon turnover and loss of carbon from upper horizons. Former moder or mor humus forms will change to mull forms, especially if the drainage is accompanied by cultivation. However, subsoil gley morphology is unlikely to change significantly in Eo or Bg and BCg horizons because any ferruginous segregations are already in their oxidized form and grey areas are already depleted of iron. Only where soil bioturbation by faunal or vegetative activity is intense enough to mix the different components of these horizons gradually, will gley morphology be lost, starting from the base of the topsoil. Because the reason for installing effective field drainage is to enable more intensive agricultural or forest production, such bioturbation is unlikely to be intense enough to change the morphology significantly. This means that some gley morphology may be a relic feature and not representative of the current soil water regime, but resulting either from recent agricultural drainage, or previous climatic conditions.

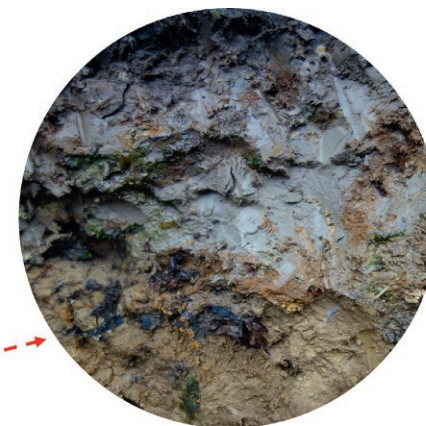
Where permanently saturated horizons extended to at least the base of the topsoil, then if field drainage is effective, and the upper horizons become aerated, the characteristic gley morphology is likely to change. Air initially infiltrates into the horizons by pore networks and root channels resulting in oxidation of the reduced iron on their surfaces forming a network of ferruginous segregations; the same as that of the capillary fringe horizon associated with fluctuating groundwater tables. Eventually, as air infiltrates into the soil, structural formation and brunification of the horizon will begin. The rapidity with which this process happens depends on the effectiveness of the drainage and the length of time it persists.



The horizon sequence in a profile with a thin peaty topsoil with *Mor* humus form over an acid, intermittently anaerobic greyish horizon with a few ferruginous segregations, over horizons in well-aerated material



The horizon sequence in a profile with a thick peaty topsoil with *Mor* humus form, the upper part of which is dominated by *Sphagnum* mosses, over an acid, intermittently anaerobic greyish horizon with a few ferruginous segregations, over a well-aerated *Spodic* horizon which, in places, has a superficial, eluvial E horizon developing

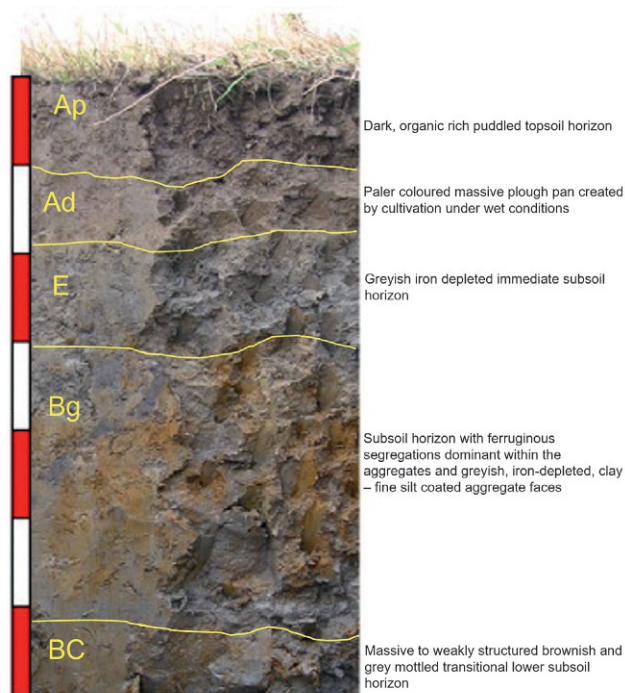


Close up of the transition from a greyish upper horizon with a few ferruginous segregations resulting from intermittently anaerobic acid conditions to a lower brownish, aerobic Bw or Bs horizon



Close up of a fragment of thin iron pan (**Bsm**) at the base of an Eo horizon. It has fallen from profile face just left of the section shown in the right hand picture above

Fig. 14 Two examples of soil profiles with anaerobic conditions resulting from excessive rainfall. Courtesy Centre for Environmental and Agricultural Informatics, Cranfield University, Cranfield, United Kingdom.; British Society of Soil Science, Building 42a, Cranfield University, Cranfield, www.soils.org.uk.



A typical horizon sequence in a profile resulting from paddy cultivation over hundreds of years. (Adapted from Figure 2 of Kögel-Knabner *et al.*, 2010)



Fig. 15 Landscape and typical soil profile associated with rice paddy cultivation. Adapted from Fig 4 of (2010) *Geoderma* 157(1–2): 1–14; Basile Morin, CC BY-SA 4.0 <<https://creativecommons.org/licenses/by-sa/4.0/>>, via Wikimedia Commons; Manojk, CC BY-SA 3.0 <<https://creativecommons.org/licenses/by-sa/3.0/>>, via Wikimedia Commons.

Classification of anaerobic soils

Anaerobic soils are recognized in most national systems of soil classification but, understandably, the terminology used, and level of importance attached to them varies depending on the environmental characteristics of each country. At the global level however, there are two principal classification systems that can be applied: The USDA NRCS Soil Taxonomy (Soil Survey Staff, 2014) and the FAO/IUSS World Reference Base (WRB) for soil resources (IUSS Working Group WRB, 2015). Both systems are hierarchical and have evolved through several iterations as understanding of the Pedosphere, its soil forming processes and their controlling factors increased. Both also define soil types by the presence or absence of diagnostic characteristics that may be horizons, properties, or materials. This concept was introduced in the first iteration of Soil Taxonomy and was an important advance, enabling scientists to identify soil horizons with properties of sufficient importance to influence soil behavior. The classification of anaerobic soils with the types of waterlogging regime defined above are set out in Table 2.

United States Department of Agriculture (USDA) soil taxonomy

The USDA Soil Taxonomy system has five hierarchical levels: Orders; Suborders; Great Groups; Subgroups; Families. At the highest level, there are 12 Orders defined by the presence of important characteristics that are diagnostic of important soil processes. Each Order is increasingly subdivided at each level depending on the presence of other specific diagnostic characteristics. The names of different soil types at different levels are a combination of part of the Soil Order name and prefixes based on the names of specific diagnostic characteristics. Soil profiles with significant periods of anaerobic conditions caused by waterlogging are defined as having an Aquic moisture regime, and Aquic conditions are a diagnostic property used at the Suborder, Great Group and Subgroup levels, depending on the depth at which saturation starts within the soil profile and the type of saturation present. Three types of Aquic Conditions are defined:

- Endosaturation, where the profile is saturated to at least 200-cm depth in all layers from the upper boundary of saturation.
- Episaturation, where the soil is saturated in one or more layers within 200-cm depth, but also has at least one layer below the saturated layer that is unsaturated within 200 cm.
- Anthric saturation, where the profile has a tilled surface layer, an underlying slowly permeable layer and a subsurface horizon with a gley morphology indicative of alternating periods of anaerobic and aerobic conditions (greyish areas and areas of ferruginous segregations).

Table 2 Classification of Anaerobic soils.

Waterlogging Regime	Upper level (cm)	World Reference Base		Soil Taxonomy			
		1 ^a	2 ^a	1 ^a	2 ^a	3 ^a	4 ^a
Saturation by a rising local or regional water table or a seepage site	Base of topsoil or within 40 cm	Gleysols Histosols	Rheic	–	Endoaqu-	–	–
	<75	–	Gleyic	–	–	–	Aquertic Aquic -Aqu- Oxyaquic
	76–100	–	–	–	–	–	–
Diurnal saturation by tidal regimes	Base of topsoil or within 40 cm	Gleysols or Solonchaks	Tidalic	Entisols	Endoaqu-	–	–
Superficial saturation resulting from impeded vertical drainage	Base of topsoil	Planosols Stagnosols Solonetz Vertisols	Stagnic Stagnic	–	Epiaqu-	–	–
	<75	–	Stagnic	–	–	–	Aquertic Aquic -Aqu- Oxyaquic
	76–100	–	–	–	–	–	–
Superficial saturation resulting from excessive rainfall	Soil surface	Stagnosols ^b Podzols	Histic Umbric Stagnic ^b Histic	–	Epiaqu-	–	–
Superficial saturation resulting from Paddy cultivation	Soil surface	–	Hydragric Anthraquic	–	Epiaqu-	–	–

^aThese numbers denote the hierarchical level at which the specified diagnostic characteristics are applied within each classification system. The symbol – in a column means that it applies to more than one class within the specified level.

^bThe Stagnic properties that are diagnostic for this waterlogging regime do not have the gley morphology related to slowly permeable horizons (greyish colors dominant on structure and pore faces) but still qualify as Stagnic through the criteria 1a, 2 or 3.

Organic (peat) soils of the Histosol Order that have Aquic conditions for at least 30 days during normal years are subdivided at the Suborder level into Fibrists, Hemists, or Sapristis depending on the extent of humification in the profile, with Fibrists being least humified and Sapristis most humified.

Mineral soils with anaerobic conditions are distinguished at the Suborder level using the term Aqu-, denoting the presence of an Aquic moisture regime, added to a component of the Order name. Thus, in the Order of Alfisols, Suborders with an Aquic moisture regime are termed Aqualfs, and within this Suborder, the terms Epiaqualfs and Endoaqualfs are used to characterize Great Groups with Episaturation and Endosaturation, respectively. Although the diagnostic condition of Anthric saturation is recognized, it doesn't appear to be used, at least in the top five hierarchical levels. This means that, at present, it must be included with one of the other two Aquic Conditions, probably that of Episaturation.

In Soil Taxonomy therefore, mineral anaerobic soils resulting from a rising local water table would be classed as Endoaqu -, whereas those resulting from superficial saturation would be classed as Epiaqu -, the - indicating a combination of letters derived from the Order within which the soil profile has been classified.

Food and Agriculture Organization (FAO)/International Union of Soil Science (IUSS) World Reference Base

The FAO/IUSS WRB system has only two hierarchical levels: a set of 32 reference soil groups (RSGs) that represent the major soil types of the World and approximate to the zonal soil concept; a second level which identifies additional specific characteristics of each RSG using 'qualifying' terms, which can be considered as adjectives that describe the presence of important diagnostic characteristics in addition to those used to characterize the RSG.

Peat soils developed because of prolonged waterlogging and associated anaerobic conditions are included in the Histosol reference soil group and qualified as either Ombric Histosols, where saturation is the result of excessive rainfall, or Rheic Histosols, where it is the result of rising local water tables. Paddy soils showing a full expression of anaerobic conditions in terms of the upper puddled and plough pan layers, recognized as a diagnostic Anthraquic horizon overlying a greyish and ferruginous mottled diagnostic Hydragric horizon are classified as Hydragric Anthrosols.

Mineral soils subject to anaerobic conditions within 50-cm depth are recognized within three other reference soil groups:

The Gleysol reference group is characterized by gleyic properties resulting from saturation by a local rising groundwater table. Specific differences in gley morphology that reflect differences in the intensity of reductive saturated conditions are used to qualify the group as follows: Reductigleyic Gleysols are those that do not have any significant ferruginous segregations (mottles) within

40-cm of the soil surface, whilst Oxygleyic Gleysols do not have any layers that lack significant ferruginous segregations within 100-cm of the soil surface. In addition, soils in saltmarsh areas that are inundated to near the surface at high tide are differentiated as Tidalic Gleysols.

The Planosol and Stagnosol reference groups are recognized by the presence of diagnostic stagnic properties exhibiting the characteristic gley morphology of slowly permeable horizons. The former reference group is differentiated by an abrupt change in clay content between coarser textured upper horizons and finer textured underlying horizons, but their diagnostic gley morphology is similar and no further qualifiers related to differences in anaerobic patterns are made.

Those soils where anaerobic conditions occur only lower in the profile and do not affect the upper soil horizons where the bulk of vegetative roots occur, are recognized using qualifying adjectives in soil classes other than those mentioned above. Two main terms are used. Gleyic reference group qualifiers are applied where Gleyic properties diagnostic of saturation by a rising local water table, are present within 75-cm but not to the base of the topsoil. Stagnic reference group qualifiers are applied where stagnic properties diagnostic of saturation caused by slowly permeable horizons are present within 75-cm, but not to the base of the topsoil.

Finally, in situations where the degree of waterlogging that resulted in anaerobic gley morphology has significantly reduced, but the morphology remains unchanged, especially where effective field drainage has been installed, two qualifying terms are used. Relictigleyic or Relictistagnic are used with reference groups other than Gleysols, Planosols or Stagnosols, to indicate the presence of gleyic or stagnic property morphology within 75-cm depth but no reducing conditions. In Gleysols, Planosols and Stagnosols with an artificial field drainage system that may have effectively aerated the upper soil horizons but where reducing conditions still occur within 75-cm depth, the Drainic qualifier is used.

Geography of anaerobic soils

Anaerobic soils are globally widespread but in geographically distinct areas. There are no current maps at the global level that adequately show the different types of waterlogging regimes, but the following gives an overview of their geography:

- Most anaerobic soils that have developed because of saturation by a rising local water table occur in and fringing the river floodplains of continental drainage basins and in coastal saltmarsh and mangrove swamps.
- A few are found in small local lower- and foot-slope seepage areas or confined hollows in flatter areas, both where permeable strata overlie impermeable strata. These areas rarely cover more than a few tens of square meters, but can occur anywhere with a suitable combination of relief and geological conditions and sufficient rainfall to generate regular groundwater recharge.
- Those anaerobic soils that have superficial saturation caused by excessive rainfall are mainly confined to upland plateau, brows, and upper slopes, and or areas where rainfall exceeds evapotranspiration for much of the year.
- Anaerobic paddy soils occur mainly in the monsoon areas of southeast Asia where 90% of the world's rice is grown (Dong and Xiao, 2016, Fig. 4). More restricted areas are globally diverse, occurring in parts of California and southeast USA, central America and the Caribbean, Peru, Chile, Brazil, Uruguay, Spain, Italy, west Africa, Madagascar and southeast Australia.
- Anaerobic soils that have superficial saturation cause by impeded vertical drainage in slowly permeable material are not well mapped at the global scale, mainly because their distinctive gley morphology has only recently been acknowledged in many countries. They occur wherever climates have an excess of rainfall over evapotranspiration for significant periods of the year and slowly permeable material such as weakly consolidated massive clay or glacial till, within about 200-cm depth. The Soil Atlas of the Northern Circumpolar Region (Jones et al., 2009) shows that Stagnosols are scattered within a relatively narrow zone extending from Ireland in the west into central Asia in the east, corresponding mainly to the Udic and to a lesser extent the Ustic moisture regime areas (see Fig. 1). It is likely that these types of soils, including Planosols will also occur in all areas with similar soil moisture regimes, where slowly permeable parent materials occur.

Conclusions

The impact of soil anaerobic conditions affects both the composition of the soil biological population and the ionic forms of some common soil elements, particularly iron and manganese. Most anaerobic soil conditions are the result of waterlogging in soil layers and the duration and frequency of such waterlogging results in a distinctive 'gley morphology' of variegated greyish, ferruginous, and black (manganiferous) colors that is easily recognized in a soil profile. However, few soils remain permanently waterlogged to the surface, most experiencing alternating periods of anaerobic 'reducing' conditions and more aerated 'oxidizing' conditions. Only those soil horizons that remain permanently saturated have consistently greyish, dark bluish or bluish green colors with no ferruginous areas. The different ways that soil waterlogging develops, whether from rising groundwater levels or surface saturation, determine how the soil matrix and its pore spaces wet up and re-aerate. These different waterlogging regimes produce distinct differences between the exact location of greyish, ferruginous, and black areas and this enables us to infer the type of waterlogging regime causing the anaerobic conditions. Such differences can be utilized in soil classification to inform us of soil pedogenesis at any site. They can also enable us to devise the best way to ameliorate anaerobic conditions that inhibit plant growth, through specific types of field drainage, or, if wetland restoration is the desired aim, what level of management is needed to increase its likely success.

References

- Avery BW (1980) *Soil Classification for England and Wales [Higher Categories]*. Soil Survey Technical Monograph No. 14. Harpenden, UK
- Dong J and Xiao X (2016) Evolution of regional to global paddy rice mapping methods: A review. *ISPRS Journal of Photogrammetry and Remote Sensing* 119: 214–227.
- FAO (2006) *Guidelines for Soil Description*, 4th edn, Rome: Food and Agriculture Organization of the United Nations, p. 2006.
- Hodgson, J.M. 1997. *Soil Survey Field Handbook*. Soil Survey Technical Monograph No. 5. pp. 116, Cranfield.
- Hollis JM (1989) *A Methodology for Predicting Soil Wetness Class From Soil and Site Properties*. SSLRC Research Report for MAFF Project CII, Silsoe, pp. 18.
- IUSS Working Group WRB (2015) *World Reference Base for Soil Resources 2014, Update 2015. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*. World Soil Resources Reports No. 106 Rome: FAO.
- Jenny H (1941) *Factors of Soil Formation: A System of Quantitative Pedology*. New York: McGraw-Hill.
- Jones A, Stolbovoy V, Tarnocai C, Broll G, Spaargaren O, Montanarella L, and (eds.). (2009) *Soil Atlas of the Northern Circumpolar Region*. Luxembourg: European Commission, Office for Official Publications of the European Communities, p. 142.
- Kögel-Knabner I, Amelung W, Fiedler S, Frenzel P, Jahn R, Kalbitz K, and Schloter M (2010) Biogeochemistry of paddy soils. *Geoderma* 157(1–2): 1–14.
- Lecomte SM, Achouak W, Abrouk D, Heulin T, Nesme X, and Haichar FZ (2018) Diversifying anaerobic respiration strategies to compete in the rhizosphere. *Frontiers in Environmental Science* 6: 139. <https://doi.org/10.3389/fenvs.2018.00139>.
- Schoeneberger PJ, Wysocki DA, Benham EC, and Soil Survey Staff (2012) *The Field Book for Describing and Sampling Soils, Version 3.0*. Lincoln, NE: Natural Resources Conservation Service, USDA, National Soil Survey Center.
- Soil Survey Staff (2014) *Keys to Soil Taxonomy*, 12th edn, USDA-Natural Resources Conservation Service.
- Thomasson AJ and Robson JD (1967) The moisture regimes of soils developed on Keuper Marl. *Journal of Soil Science* 18: 329–340.
- Van Bodegom PM, Van Reeve J, and Van Der Gon HACD (2003) Prediction of reducible soil iron content from iron extraction data. *Biogeochemistry* 64(2): 231–245.
- Zanella A, Jabiol B, Ponge JF, Sartori G, de Waal R et al. (2011) European Humus Forms Reference Base. fhal-00541496v2.

Alluvium and alluvial soils

Janis L Boettinger, Department of Plants, Soils, and Climate, Utah State University, Logan, UT, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of J.L. Boettinger, ALLUVIUM AND ALLUVIAL SOILS, Editor(s)0029: Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 45–49, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00008-4>.

Introduction	417
Alluvium	417
Mineralogical composition	418
Particle-size distribution	418
Landforms	419
Alluvial soils	420
Recent alluvial soils	420
Older alluvial soils	421
Classification	422
Human uses of alluvial soils	422
Conclusions and outlook	423
References	423

Abstract

Alluvial soils, formed in alluvium deposited by streams and rivers, are useful and productive soil resources. Alluvium has varied particle-size distributions and occurs in all climates. It underlies geomorphic surfaces ranging in age from the present day to millions of years. As a consequence of the range of conditions for their formation, alluvial soils are extremely diverse. They are important for understanding landscape evolution, and for flood control and food security.

Key points

- Alluvium and alluvial soils are very diverse, with properties depending on the ultimate rock type source, climate, and geomorphic surface age.
- Alluvial soils occur on alluvial fans and terraces, and floodplains.
- Alluvial soils are important for understanding landscape evolution and soil formation, for flood control, and for the production of food and fiber.
- The loss of prime agricultural land from urbanization and industrialization is a serious threat to food security.

Introduction

Alluvial soils are some of the world's most useful and productive soil resources. Alluvium, the parent material of alluvial soils, is the sediment deposited by fluvial systems such as rivers and streams. Alluvium includes a wide variety of mineralogical compositions and particle-size distributions, depending on the source of geologic materials and depositional environment. Alluvium occurs in all climates and underlies geomorphic surfaces ranging in age from present day to millions of years. Therefore, alluvial soils that have formed in alluvium are extremely diverse. Alluvial soils are also important for use by humans such as for understanding landscape evolution, flood control, and food security.

Alluvium

Alluvium is ultimately derived from the weathering and erosion of igneous, sedimentary, and or metamorphic rocks, such as basalt, limestone, or gneiss, respectively. Alluvium may also be derived from the erosion of other unconsolidated sedimentary deposits, including colluvium (gravity-deposited), lacustrine (lake-deposited), loess or eolian sand (wind-deposited), or older alluvium, that has been deposited in a fluvial system. Important characteristics of alluvium, which ultimately influence the properties of alluvial soils, include chemical and mineralogical composition and particle-size distribution. The type of landform composed of alluvium and the stability and age of its surface (geomorphic surface) also influence soil formation and use and management by humans.

Mineralogical composition

The mineralogical composition of alluvium depends on the lithology, or type of rock material, from which the sediments are ultimately derived. Alluvium derived from sedimentary rocks such as limestone and dolomite will be calcareous (contain calcite, CaCO_3). Alluvium derived from gypsum-bearing shale will be gypsiferous (contain gypsum, $\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$). Alluvium derived from igneous rocks lacks carbonates and gypsum. Alluvium from mafic igneous rocks, such as basalt and gabbro, will be rich in magnesium- and iron-bearing silicate minerals, whereas alluvium from siliceous igneous rocks, such as rhyolite and granite, will be rich in quartz and silica-rich feldspars. Alluvium derived from mixed lithology will have mixed mineralogy. The north-south trending San Joaquin Valley of central California, USA, for example, has alluvium derived from the granitic rocks of the Sierra Nevada mountain range on the east side and alluvium derived from the marine sedimentary rocks of the Coastal Ranges on the west side. Alluvium in the east is dominated by silicate minerals including quartz and silica-rich feldspar with minor amounts of biotite and hornblende. In contrast, alluvium in the west contains varying amounts of carbonate, sulfate, and silicate minerals, reflecting the shale, siltstone, sandstone, and limestone sedimentary rocks of the Coastal Ranges.

Particle-size distribution

The particle-size distribution of alluvium depends primarily on the energy of the fluvial environment in which sediments are deposited. High-energy fluvial systems, such as channels of braided streams, can carry and ultimately deposit relatively large (coarse) particles. Glacial outwash, a type of alluvium deposited by high-discharge glacial meltwater streams, typically has a large volume of rock fragments (gravel and cobbles) and sand. Low-energy fluvial environments transport and deposit smaller (finer) particles. As water rises in a stream and overflows its banks, coarser particles are deposited adjacent to the channel, eventually building up natural levees, and finer particles are deposited by the overbank flood waters. Overbank flood deposits often exhibit a fining of upward particle-size distribution as coarser particles (sands) settle faster from suspension than finer particles (silt and clay). An example of gravely channel deposits and silty overbank flood and distal fan deposits is shown in Fig. 1.

Alluvium is generally considered to be well sorted—each layer of the alluvial deposit exhibiting a relatively narrow range of particle-size distribution. However, different components of the stream channel vary in water flow velocity, which affects the erosion, transport, and deposition of sediment. Stream channels can migrate across the landscape in response to changes in water discharge and bed loads. Therefore, alluvium can be stratified, i.e., composed of alternating layers with different particle-size ranges (e.g., Fig. 2). The stratigraphy, or study of the relative positions, of alluvial deposits can provide clues to the history of landscape evolution.

The lithology and physical weathering characteristics of the source of alluvium also influence the particle-size distribution. Alluvium derived from sandstone is typically sandy, as the cement between sand grains is weathered and the sand grains are eroded, transported and deposited in fluvial systems. Alluvium derived from granitic rocks is typically sandy to fine gravely—the size range of the individual mineral grains—because the granitic rock fabric typically shatters into individual grains upon weathering. Alluvium derived from finer textured rocks and sediments is finer textured. For example, alluvium derived from the erosion of fine-textured lacustrine sediment typically lacks rock fragments and is rich in silt and clay.

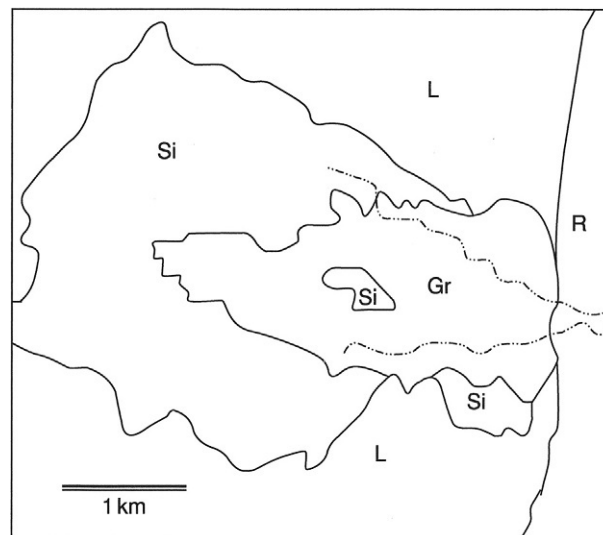


Fig. 1 Schematic representation of the Holocene Green Canyon alluvial fan cut into and deposited on lacustrine deposits of Pleistocene Lake Bonneville, Cache Valley, northern Utah, USA. Streams draining rocky limestone and dolomite uplands (R) deposited gravely channel alluvium (Gr) and silty overbank and distal fan alluvium (Si). These Holocene soils are the Green Canyon series (loamy-skeletal Typic Haploxerolls) formed in gravely alluvium, and the Millville series (coarse-silty Typic Haploxerolls) formed in silty alluvium. (Information derived from the Soil Survey of Cache Valley Area, Utah).



Fig. 2 Holocene alluvial soil in the arid climate of the Circle Cliffs area of the Colorado Plateau, Garfield County, southern Utah, USA, showing stratification typical of alluvium. This soil is mapped as the **Radnik** series (coarse-loamy Ustic Torrifluvents), and the profile is about 1.2 m thick.

Landforms

Alluvium occurs in deposits that comprise a variety of different landforms, such as alluvial fans, alluvial terraces, alluvial plains, floodplains, and footslopes or toeslopes where hillslopes transition into alluvial fans. Pleistocene glacial outwash occurs on outwash terraces.

Streams draining mountainous terrain that discharge into wide valleys or basins deposit sediments as fan-shaped landforms termed alluvial fans (Fig. 3a). Alluvial fans are common in low-lying areas adjacent to tectonically active mountain ranges, including those in western North America, Central America, and South America; central Asia; southern Europe; the Middle East, and eastern and southern Africa. Components of alluvial fans include channels, floodplains, fan terraces, and dissected fan remnants. Alluvial fans emanating from adjacent streams can coalesce and form bajadas (Fig. 3b). Finer textured distal fan deposits can form alluvial plains or alluvial basins on the valley floor. Overbank flood deposits can form inter-fan basins between low-gradient alluvial fans. Alluvial fans generally slope away from the sediment source, with steeper slope gradients occurring higher and gentler slope gradients occurring lower on the fan.

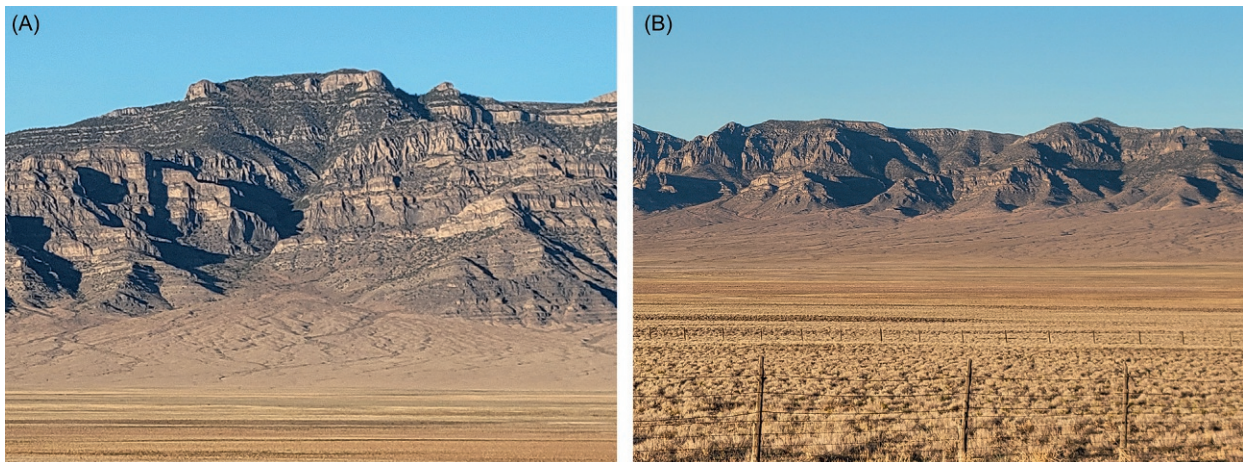


Fig. 3 (A) Alluvial fan emanating from a stream; (B) Alluvial fans coalescing to form a bajada; west side of the Wah Wah Mountains of the Basin and Range, Millard County, western Utah, USA. Photos courtesy of Eugene W. Schupp, Utah State University.



Fig. 4 Photograph of head-trained Zinfandel wine grape vineyard on alluvial terraces on the west side of Dry Creek Valley, Sonoma County, California, USA. Reddish surface soil (*Manzanita* series, fine-loamy Ultic Palexeralfs) is mapped on the gently sloping higher terrace in the foreground formed in the oldest alluvium, intermediate aged soil (*Yolo* series, fine-silty Mollic Xerofluvents) is mapped in the middle ground on a lower nearly level alluvial terrace, and the youngest soil (*Cortina* series, loamy-skeletal Typic Xerofluvent) is mapped on the lowest floodplain adjacent to Dry Creek (marked by trees). Cobb Mountain is the peak visible in the center background.

Floodplains and alluvial terraces occur parallel to the stream or river, as illustrated in Dry Creek Valley of northern California (Fig. 4), and by the extensively studied alluvial terrace system developed parallel to the Guadalquivir River, southern Spain (Calero et al., 2008, 2009). Recent alluvium occurs on almost level floodplains and recent alluvial terraces adjacent to streams and rivers. Older alluvium occurs on alluvial terraces, also known as stream terraces or river terraces, that are generally higher than the associated floodplains and are not subject to frequent flooding. Higher terraces can be subject to upland erosion and dissection as the valley is incised by the river, as demonstrated in the Tagus and Duero river valleys of central Spain (Roquero et al., 2015).

Alluvial soils

The morphological, physical, chemical, and mineralogical properties of alluvial soils depend greatly on the characteristics of the alluvium (parent material) in which the soils formed, especially when the soils are young. As alluvial soils develop over time, the other soil-forming factors (climate, organisms/biota, and relief) influence the resulting soil properties. As alluvium is often well sorted, the particle-size distribution and stratification of layers influence the movement of water and air through the soil.

Recent alluvial soils

Soils on active floodplains receive deposits of new alluvium with each flooding episode. The amount of alluvium deposited during each event will vary. Small amounts of material deposited on the soil surface can be barely perceptible and sediments can be incorporated rapidly into the underlying surface horizon, the rate of which depends on climate and biota. Plants adapted to frequently flooded soils may survive and grow up through the newly deposited sediment, with soil fauna facilitating mixing of the newer sediments into the soil. Larger amounts of alluvium deposited in thicker layers may completely bury underlying soils, setting soil-forming time back and resetting the geomorphic surface age to zero. If there is sufficient time between flooding events for the geomorphic surface to stabilize, plants to establish, and soil organic matter to accumulate, A and/or O horizons may develop.

Recent alluvial soils affected by periodic flooding and sediment typically have somewhat elevated concentrations of organic carbon at depth. New alluvium is often derived from the eroded A or O horizons of upland and/or upstream soils. If the deposit of



Fig. 5 Photograph of a fine-textured alluvial soil with a buried A horizon on a recent (Holocene) alluvial terrace of the Little Bear River, Cache County, Utah, USA. The alluvium was derived from the erosion of older sediments deposited by Pleistocene pluvial Lake Bonneville. The soil profile is about 1.4 m thick.

new alluvium is significant, landform stabilization and soil formation can result in a soil with one or more buried A and/or O horizons. (Fig. 5). Thus, soils with buried A and or O horizons demonstrate an irregular decrease in organic carbon with increasing depth.

Climate and biota further influence the properties of recent alluvial soils. The soils in humid climates generally support dense vegetation, thus developing A and or O horizons more rapidly than in drier climates. Highly soluble minerals such as gypsum and salts more soluble than gypsum (halite, NaCl; soda, Na_2CO_3 , etc.) present in the parent alluvium, will be rapidly dissolved and leached in humid climates, whereas these constituents are often retained (and can accumulate) in arid and semiarid climates.

Older alluvial soils

In general, older alluvial soils develop when they are no longer subject to periodic events that involve flooding and sediment deposition. Geomorphic surfaces and landforms are more stable and thus able to support a mature vegetation cover. Organic carbon in the subsoil eventually decomposes and the soil develops a regular distribution of organic carbon with increasing depth.

Climate, which influences vegetation and associated biota, further modifies the properties of alluvial soils. In cold climates with permafrost, cryoturbation (mixing of soil by freeze–thaw) can disrupt stratification of alluvial layers. In warmer climates, the available precipitation influences the resulting soil properties because water is required for soil-forming processes such as chemical weathering of minerals, decomposition of organic matter, translocation of clay, leaching of soluble constituents, oxidation and reduction of iron, etc. In humid climates, older alluvial soils are commonly leached of carbonates and other more soluble minerals, and resulting in a low base saturation. With time, B horizons develop in the subsoil and accumulate constituents such as silicate clay, free iron oxides, and or metal–humus complexes. Soils on different-aged alluvial terraces illustrate how alluvial soils develop over time. In Mediterranean-type climates soils typically increase in clay content, have redder Munsell color hue (e.g., Calero et al., 2008), decrease in pH and carbonates, and increase in iron and aluminum oxides with increasing age (e.g., Calero et al., 2009; Martín-García et al., 2016). Similar trends are observed in humid temperate climates, including changes in soil properties associated with decreasing inherent fertility, such as decreasing cation exchange capacity and decreasing extractable micronutrients with increasing age (Shaw et al., 2003). In subhumid, semiarid, and arid climates, alluvial soils may be incompletely leached and B horizons commonly accumulate carbonates (e.g., Sion et al., 2022) and or silica (e.g., Kendrick and Graham, 2004). In more arid climates, gypsum, and or salts more soluble than gypsum can accumulate, depending on mineralogy of the alluvium and the landscape setting (e.g., Moghiseh and Heidari, 2012).

Older alluvial soils have often been subject to changes in climate during their development. This is particularly the case in soils formed in alluvial valleys or basins over hundreds of thousands to millions of years, such as the valleys and basins of the Basin and Range physiographic province of western North America. Very old alluvial soils can exhibit the imprinting of several climate regimes, e.g., well developed, clay-rich B horizons that have been engulfed by calcium carbonate accumulations such that clay skins (clay films, argillans) are no longer visible in the soil morphology observed in the field, but are evident in the soil micromorphology observed with a petrographic or electron microscope.

Alluvium, alluvial landforms, and the associated alluvial soils can be subject to active fluvial processes at one time or another. Therefore, all or a portion of an alluvial soil may be eroded and new sediment can be deposited adjacent to and or overlying the

older alluvial soil (e.g., Gile and Hawley, 1966; Eghbal and Southard, 1993). There may be erosion only of the upper soil horizons overlying a B horizon that is relatively resistant to erosion, such as a carbonate-rich and or clay-rich B horizon. Such a truncated soil may have the B horizon exposed at the soil surface. Recent alluvial deposits may bury a truncated soil. As the deposit and the geomorphic surface stabilize and soil development occurs, soil-forming processes alter the recent overlying alluvium. If the recent alluvium is thin, soil formation can extend into the older buried alluvial soil, which is sometimes referred to as having a welded profile.

Alluvial soils often retain at least a partial record of the Earth's history through alluvial deposition, landform and geomorphic surface stabilization, soil formation, and subsequent erosion. The stratigraphy of soils across the landscape records the geomorphic and soil-forming processes influenced by the environment. Therefore, alluvial soils in their landscape context are useful for soil-geomorphic studies. The degree of soil development on alluvial surfaces has been used to understand landscape evolution (e.g., Eghbal and Southard, 1993), correlate geomorphic surfaces of similar ages (e.g., Gile and Grossman, 1979; Harden, 1987; Roquero et al., 2015), decipher paleoclimate, and assess the tectonic activity of faults that cut alluvial surfaces. Understanding landscape evolution and the formation of alluvial soils also facilitates the mapping of soils (e.g., Fig. 1).

Classification

Recent alluvial soils that lack significant development of surface or subsurface diagnostic features are classified into the order of Entisols (U.S. Department of Agriculture Soil Taxonomy; Soil Survey Staff, 2014) or reference soil group of Fluvisols (World Reference Base; IUSS Working Group WRB, 2015). In Soil Taxonomy, the formative element 'fluv' indicates the alluvial origin and the stratified nature of recent alluvial soils. Alluvial Entisols little affected by saturation and reducing conditions that have relatively large organic carbon concentrations (greater than or equal to 0.2%) at depth (125 cm) or an irregular decrease in organic carbon with depth (see Fig. 5) classify into the suborder of Fluvents. Fluvents are further subdivided into 'great groups' by soil moisture regimes (e.g., Udifluvents with a udic [humid] soil moisture regime, Torrifluvents with an aridic [arid] soil moisture regime, etc.). Similar Entisols that are saturated for prolonged periods during a normal year classify into the great group of Fluvaquents. Alluvial soils in drier climates may classify in the Entisols suborder of Orthents (e.g., Torriorthents, Xerorthents, Ustorthents, and Udorthents), which vary in soil moisture regime. In the World Reference Base, soils with fluvic material, with evidence of obvious stratification or relatively large organic carbon concentrations (greater than or equal to 0.2%) and an irregular decrease in organic carbon with depth, classify as Fluvisols.

With greater landform stability and soil development, alluvial soils can develop into a myriad of different soil types and be classified into several different orders of Soil Taxonomy and into several reference soil groups of World Reference Base. In Soil Taxonomy, alluvial soils in arid climates that have subsurface diagnostic horizons (e.g., argillic horizon, calcic horizon, gypsic horizons, duripans, petrocalcic horizons, etc.) classify as Aridisols. Alluvial soils with mollic epipedons (thick, dark, soft, A horizons) and high base saturation throughout the soil depth are Mollisols. Other alluvial soils with ochric or umbric epipedons classify as Spodosols, Oxisols, Ultisols, Alfisols, Inceptisols, depending on the diagnostic subsurface horizons and properties present and the degree of leaching indicated by base saturation. Gelisols, Vertisols, and Andisols are also possible, depending on climate and diagnostic horizons and properties. In the World Reference Base, older, more developed, alluvial soils can be classified into more than 20 reference soil groups, depending on their diagnostic horizons, properties, and materials. Examples include Gleysols, Podzols, Ferralsols, Chernozems, Calcisols, Lixisols, Luvisols, and Cambisols. For illustration, the five soils in the Guadalquivir fluvial chronosequence were classified according to the World Reference Base/Soil Taxonomy as Haplic Fluvisol/Typic Xerofluvent (0.3 ka), Haplic Calcisol/Typic Calcixerept (7 ka), Cutanic Luvisol/Haplic Palexeralf (70 ka), Lixic Calcisol/Calcic Haploxeralf (300 ka), and Cutanic Luvisol/Haplic Palexeralf (Calero et al., 2008).

Human uses of alluvial soils

Alluvial soils have been the sites of the earliest agriculture. For millennia, humans have used alluvial soils, especially those that are young and less developed, for the production of food and fiber. New alluvium rich in organic matter and nutrients provide fertile soils for agriculture. Alluvial soils are often deep, occur on nearly level land, and are relatively easy to cultivate. In drier climates, the occurrence of alluvial soils near sources of water for irrigation of crops facilitated the development of agriculture and civil society. For example, early agricultural societies with rich cultures developed along the Tigris and Euphrates rivers in the Middle East, the Nile in Egypt, and the Ebro in the Iberian Peninsula. However, irrigated alluvial soils can be subject to salinization.

Indigenous cultures throughout the world raise crops on alluvial soils, particularly on recent alluvial soils that are productive and easy to cultivate. Periodic flooding results in the rejuvenation of soil fertility by depositing organic-matter-rich sediment and nutrient-rich water on soil surfaces. Zuni and other indigenous peoples of the arid and semiarid southwestern U.S. used the periodic deposition of runoff and sediments eroded from hillslope soils onto traditional agricultural fields on alluvial fans to grow crops and sustain soil quality (Norton et al., 2007).

Recent alluvial soils on floodplains are subject to periodic flooding and urban and industrial development should be avoided. These soils are important for flood control and can provide other ecological benefits. For example, the Yolo Bypass in the Central Valley of California near Sacramento diverts floodwaters away from urban and high-value agricultural land while providing wildlife habitat.

Even today, alluvium and alluvial soils underlie the most productive agricultural regions of the world. For example, the Central Valley of California supports a vibrant, multi-billion-dollar agricultural industry. However, many of the soil and landscape properties that make alluvial soils attractive for agriculture are also attractive for residential, commercial, and industrial development. Alluvial soils are often deep and well-drained; occur in wide, nearly level valleys or plains; and are easy to excavate. Therefore, thousands of hectares of productive agricultural lands on alluvial soils are being lost each year to urbanization and industrialization. Alluvial soils can also provide the gravel and sand for road base and for concrete for construction.

Conclusions and outlook

Alluvium and alluvial soils are important for understanding recent Earth history through landscape evolution and soil formation, for flood control, for construction materials for roads and buildings, and for the production of food and fiber. The loss of prime agricultural land to urbanization and industrialization; the degradation of alluvial soils by erosion and salinization; and the dwindling supply and quality of irrigation water for agriculture in a rapidly warming climate are serious threats to food security on local, regional, and global scales.

Given changes in climate and the importance of these soils to humans, future research on the impacts of climate change on alluvial soils is needed. Because alluvial soils occur on different aged geomorphic surfaces, in different parent materials, and in different climates, further research on properties and genesis could provide insight on potential soil and land use change as climate changes.

References

- Calero J, Delgado R, Delgado G, and Martín-García JM (2008) Transformation of categorical field soil morphological properties into numerical properties for the study of chronosequences. *Geoderma* 145(3–4): 278–287. 2008.
- Calero J, Delgado R, Delgado G, and Martín-García JM (2009) SEM image analysis in the study of a soil chronosequence on fluvial terraces of the middle Guadalquivir (southern Spain). *European Journal of Soil Science* 60: 465–480.
- Eghbal MK and Southard RJ (1993) Stratigraphy and genesis of Durorthids and Haplargids on dissected alluvial fans, western Mojave Desert, California. *Geoderma* 59: 151–174.
- Gile LH and Grossman RB (1979) *The Desert Project Soil Monograph: Soils and the Landscapes of a Desert Region Astride the Rio Grande Valley, Near Las Cruces, New Mexico*. Washington, DC: USDA Soil Conservation Service.
- Gile LH and Hawley JW (1966) Periodic sedimentation and soil formation on an alluvial fan piedmont in southern New Mexico. *Soil Science Society of America Proceedings* 30: 261–268.
- Harden JW (1987) Soils developed in Granitic Alluvium Near Merced, California. In: *US Geological Survey Bulletin 1590-G*. Washington, DC: Government Printing Office. pp. A1–65.
- IUSS Working Group WRB (2015) World Reference Base for Soil Resources 2014, update 2015 International soil classification system for naming soils and creating legends for soil maps. In: *World Soil Resources Reports No. 106*. Rome: FAO.
- Kendrick KJ and Graham RC (2004) Pedogenic silica accumulation in chronosequence Soils, Southern California. *Soil Science Society of America Journal* 68(4): 1295–1303.
- Martín-García JM, Sánchez-Marañón M, Calero J, Aranda V, Delgado G, and Delgado R (2016) Iron oxides and rare earth elements in the clay fractions of a soil chronosequence in southern Spain. *Soil Science* 67(6): 749–762.
- Moghiseh E and Heidari A (2012) Polygenetic saline gypsiferous soils of the Bam region, Southeast Iran. *Journal of Soil Science and Plant Nutrition* 12(4): 729–746.
- Norton JB, Sandor JA, and White CS (2007) Runoff and sediments from hillslope soils within a Native American agroecosystem. *Soil Science Society of America Journal* 71(2): 476–483.
- Roquero E, Silva PG, Goy JL, Zazo C, and Massana J (2015) Soil evolution indices in fluvial terrace chronosequences from central Spain (Tagus and Duero fluvial basins). *Quaternary International* 376: 101–113.
- Shaw JN, Odom JW, and Hajek BF (2003) Soils on quaternary terraces of the Tallapoosa River, central Alabama. *Soil Science* 168(10): 707–717.
- Sion BD, Harrison BJ, McDonald EV, Phillips FM, and Axen GJ (2022) Chronofunctions for New Mexico, USA soils show relationships among climate, dust input, and soil development. *Quaternary International* 618: 35–51.
- Soil Survey Staff (2014) *Keys to Soil Taxonomy*, 12th edn. Washington, DC: USDA-Natural Resources Conservation Service.

Relevant websites

- Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Official Soil Series Descriptions. Available online. Accessed [January 30, 2023]. <https://www.nrcs.usda.gov/resources/data-and-reports/official-soil-series-descriptions-osd>
- Yolo Bypass Foundation <https://www.yolobasin.org/> (Accessed October 14, 2022).

Hydric Soils

MJ Vepraskas^a and JF Berkowitz^b, ^aDepartment of Crop and Soil Sciences, North Carolina State University, Raleigh, NC, United States;
^bUnited States Army Corps of Engineers, Engineer Research and Development Center, Vicksburg, MS, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of GW Hurt, Hydric Soils[☆], *Reference Module in Earth Systems and Environmental Sciences*, Elsevier, 2013, ISBN 9780124095489, <https://doi.org/10.1016/B978-0-12-409548-9.05186-1>.

Introduction	424
Hydric soils in soil classification systems	425
Formation of hydric soil features	427
Types of morphological features of hydric soils	428
Organic-carbon based morphological features	428
Iron and manganese-based morphological features	429
Redox concentrations	430
Redox depletions	430
Reduced matrix	430
Gley soil colors	430
Formation of redoximorphic features	430
Field indicators of hydric soils	432
The hydric soil technical standard	433
Summary	434
References	435

Abstract

Hydric soils are found in wetlands. They develop in areas that are saturated and anaerobic for periods long enough for redoximorphic features to form, and/or for organic C to accumulate, all within 30 cm of the soil surface. Hydric soil field indicators are soil layers having well-defined colors, specific quantities of redoximorphic features, and organic C contents, that are used to identify hydric soils in the field. The Hydric Soil Technical Standard describes how a hydric soil without field indicators can be identified using measurements to determine if the soil becomes saturated and anaerobic.

Key points

- Hydric soils form in wetland environments subject to prolonged saturation and anaerobic conditions.
- Hydric soils were defined to enforce wetland protection laws in the United States.
- They are identified using field indicators of hydric soils that document characteristic morphological features.
- The Hydric Soil Technical Standard provides a mechanism to link hydric soil form and function.

Introduction

Hydric soils are those that form in wetlands. The term 'hydric soil' was defined in the United States as part of the national effort to protect wetlands from being destroyed through drainage or the placement of fill materials (Vepraskas, 2016). The concept of hydric soils and the methods used to identify them follow specific requirements in the United States, and these will be summarized in this chapter. How hydric soils are identified in other countries will be illustrated by discussing some of the major soil classification systems that are used around the world.

Hydric soils are defined as "soils that formed under conditions of saturation, flooding, or ponding long enough during the growing season to develop anaerobic conditions in the upper part." (Federal Register, 1994). The 'upper part' is loosely defined to be the soil between depths of 0 to 30 cm in soils with a loamy or clayey texture, and between the depths of 0–15 cm in sands. When organic soil materials are present at the soil surface, the upper part may be only 1 cm thick or less depending on the land region where the soil is found. This is the official definition that was incorporated into a number of federal statutes and as a result it cannot be easily changed. Vepraskas et al. (2016) provides a complete history of the concept and definition of hydric soils.

The term 'hydric soil' was used for the first time in the wetlands classification system of Cowardin et al. (1979) to describe the soils found in frequently saturated or inundated ecosystems. The Cowardin system was used by the United States Fish and Wildlife Service to develop the National Wetlands Inventory (NWI) that mapped wetlands across the United States. It defined wetlands as: "...lands transitional between terrestrial and aquatic systems where the water table is usually at or near the surface or the land is covered by

Table 1 U.S. Government statutes that protect wetlands.

<i>Federal Act</i>	<i>Year Passed</i>	<i>Lands Affected</i>	<i>Federal Agency for Enforcement</i>	<i>Actions Prevented</i>
Clean Water Act	1977	Wetlands on non-agricultural land and forested wetlands on farms	U.S. Army Corps of Engineers and the Environmental Protection Agency	Filling or draining wetlands without a permit
National Food Security Act	1985	Wetlands on agricultural land except those in forests	U.S. Department of Agriculture-Natural Resources Conservation Service	Draining wetlands for agricultural production

shallow water. For purposes of this classification wetlands must have one or more of the following three attributes: (1) at least periodically, the land supports predominantly hydrophytes, (2) the substrate is predominantly hydric soil, and (3) the substrate is non-soil and is saturated with water or covered by shallow water at some time during the growing season of each year."

Hydric soils are now identified primarily to enforce water protection laws in the United States that were enacted through Congressional Acts. The two major acts shown in [Table 1](#) prevent wetlands from being filled in or drained without a permit first being obtained from the federal agency enforcing the laws. These laws required the boundaries of wetlands be identified on maps of individual parcels of property ([Hurt and Noble, 2016](#)). Each wetland is expected to meet separate requirements for wetland hydrology, hydric soils, and hydrophytic vegetation.

Hydric soils in soil classification systems

Hydric soils were defined independently of any scientific soil classification system, but similar soils can be found in most soil classifications containing saturated and chemically-reduced (i.e., anaerobic) soils. For example, in the United States soil classification system *Soil Taxonomy*, mineral soils that are saturated and reduced, either continuously or periodically, have 'aquic conditions' ([Soil Survey Staff, 2014](#)). The presence of aquic conditions can be determined in the field from direct measures of saturation with piezometers or wells, and documenting reducing conditions through either measurements of oxidation-reduction (redox) potentials, use of dipyrindyl dye, or by 'Indicator of Reduction in Soil' (IRIS) tubes (i.e., iron oxide coated pipes, [Berkowitz et al., 2021](#)). However, aquic conditions are more commonly identified using readily observable redoximorphic features characterized by mottled gray or red colors that form in response to prolonged or repeated periods of chemical reduction ([Vepraskas and Vaughan, 2016](#)).

Of the 12 soil orders defined in *Soil Taxonomy*, 10 have suborders that require aquic conditions be present "for some time in normal years" ([Soil Survey Staff, 2014](#)). [Table 2](#) shows that the occurrence of aquic conditions is identified in the taxonomic name at the suborder level when they occur within 50 cm of the soil surface. Aquic conditions are not identified in Aridisols, which are found in deserts, but some soils in this group do saturate periodically and may be hydric soils. Histosols are organic soils that require at least 30 days of saturation in normal years, but do not require the soils to be reduced or anaerobic. In reality, most Histosols probably do develop reducing conditions when saturated at temperatures when microbes are active, and so there is no need to recognize aquic conditions in a separate taxonomic group in Histosols.

All hydric soils have aquic conditions, but not all mineral soils with aquic conditions are hydric soils. [Table 2](#) shows that to be classified as having aquic conditions at the suborder level, a soil must have redoximorphic features (i.e., mottled soil color patterns) within 50 cm of the surface. Hydric soils, on other hand, must have redoximorphic features within 30 cm of the surface for loamy and clayey materials, and within 15 cm of the surface for sandy soils. In addition, the saturated and reduced conditions needed for aquic conditions can occur at any time during the year, while for hydric soils they must occur during the frost-free growing season when microbial activity is occurring in the soil.

The World Reference Base for Soil Resources (WRB) is an international soil classification system that also includes saturated and reduced soils ([IUSS Working Group WRB., 2015](#)). Like *Soil Taxonomy*, it does not recognize hydric soils as a distinct group, but includes such soils in at least four major 'reference soil groups': Gleysols, Histosols, Planosols, and Stagnosols. The characteristics of each of these groups are summarized in [Table 3](#). The reference soil groups are further defined for specific soils using principal and supplemental qualifiers, examples of which are also shown in [Table 3](#). A hydric soil might be classified as a Sapric Histosol (Hyperorganic, Sulfidic) for example.

Examples of soil orders or the highest levels in the classification systems used by Russia, China, Australia and Brazil that contain hydric soils are shown in [Table 4](#). Additional details for the lower categories of these classification systems were summarized by [Buol et al. \(2011\)](#). Because not all systems use the term 'soil order' we will refer to these levels simply as categories. These classification systems were selected to show how terms for the classification categories that contain hydric soil can vary around the world. Hydric soils occur in both mineral and organic soils for each of the classification systems shown in [Table 4](#). As with *Soil Taxonomy* and the WRB systems, organic soils are designated as a separate soil category. Unlike *Soil Taxonomy*, mineral soils that are saturated for long periods, such that they develop low chroma or gleyed colors (redoximorphic features), are separated out into their own categories (e.g., Gley soils).

Table 2 Taxonomic groups from *Soil Taxonomy* which contain hydric soils.

<i>Soil Order</i>	<i>Aquic Suborder</i>	<i>Examples of Aquic Great Groups</i>	<i>Depths of Aquic Conditions or Redoximorphic Features cm</i>	<i>Examples of Soil Series with Hydric Soils^a</i>	<i>General Comments</i>
Alfisols	Aqualf	Glossaqualf Endoaqualf	≤50	Sebring	Soils with accumulation of clay in subsurface, higher base saturation and pH than Ultisols
Andisols	Aquand	Epiaquand Endoaquand	40–50	Onomea	Soils commonly formed in volcanic materials
Aridisols	NA ^b	NA	NA	NA	Desert soils
Entisols	Aquent	Sulfaquent Fluvaquent	≤50	Chincoteague	Young, undeveloped soils
Gelisols	NA	Aquorthel Aquiturbel	≤50	Mankomen	Soils frozen at least part of the year
Histosols	NA	NA	≤40 for some classes	Ponzer	Soils formed from organic materials
Inceptisols	Aquept	Fragiaquept Humaquept	40–50	Chastain	Soils showing more development than Entisols, but less than other soil orders
Mollisols	Aquoll	Argiaquoll Cryaquoll	40–50	Espelie	Soils with thick, black surfaces
Oxisols	Aquox	Haplaquox	≤50	Moteado	Weathered, frequently iron-rich soils, usually in tropics
Spodosols	Aquod	Alaquod Duraquod	≤50	Immokalee	Generally sandy soils formed in forests
Ultisols	Aquult	Kandiaquult Albaquult	≤50	Rains	Soils with accumulation of clay in subsurface, more acidic than Alfisols
Vertisols	Aquert	Sulfaquent Salaquent	≤50	Sharkey	Clayey soils that swell on wetting and shrink on drying

^aDescriptions of the soil series can be found at: <https://soilseries.sc.egov.usda.gov/osdname.aspx>^bNA is not applicable.**Table 3** Taxonomic groups from the World Reference Base soil classification system which may contain hydric soils.

<i>Reference Soil Group</i>			
<i>Name</i>	<i>Description</i>	<i>Examples of Principal Qualifiers</i>	<i>Examples of Supplemental Qualifiers</i>
Histosols	≥ 40 cm organic material with ≥ 20% organic carbon	Fibric/Hemic/Sapric—organic materials also called peat, mucky peat, and muck, respectively, which vary by percentage of visible fibers present. Only one used.	Hyperorganic—organic material ≥ 200 cm thick
Gleysols	Saturated with groundwater long enough to become reduced within 40 cm of surface	Stagnic—saturated with surface water long enough to become reduced and show redoximorphic features Reductigleyic—gley colors present but no features with oxidized iron (no redox concentrations)	Sulfidic—contains inorganic sulfides Humic—having ≥ 1% organic carbon to depth of 50 cm Limnic—contains organic or mineral material deposited by organisms such as diatoms or algae
Planosols	Abrupt textural change within 100 cm of surface capable of perching water, with redoximorphic features	Histic—organic material saturated with water ≥ 30 consecutive days most years Gleyic—saturated with groundwater long enough to get reduced, gley Munsell hues may be present	Capillaric—layer ≥ 25 cm thick that has so few macropores that water saturation of capillary pores causes reducing conditions Plinthic—a layer with ≥ 15% Fe nodules or nearly cemented Fe concentrations
Stagnosols	Within 25 cm of surface, layer(s) with redoximorphic features in 50% of volume, or reducing conditions	Thionic—extremely acid subsurface horizon (pH < 4) contain sulfuric acid produced from oxidation of sulfides Anthraquic (rice paddy) near surface with reduced matrix and oxidized channels	Ferric—has ≥ 15% redox concentrations within 100 cm of surface Nitric—subsurface horizon with ≥ 30% clay

IUSS Working Group WRB (2015) *World Reference Base for Soil Resources 2014*. Update 2015 International Soil Classification System for Naming Soils and Creating Legends for Soil Maps. World Soil Resources Reports No. 106. FAO: Rome.

Table 4 Summary of the classification systems used in selected countries which may contain hydric soils similar to those defined in the U.S.

Country	Soil Type	Soil Order or Equivalent	Description	Hydric Soil Occurrence
Russia	Mineral	Gley soils	Gley horizon with low chroma colors and iron mottles	When horizon is within 25 cm of surface and is at least 15 cm thick
		Hydrometamorphic soils	Hydrometamorphic horizon with olive hues without iron mottles	When horizon is within 30 cm of surface
	Organic	Peat soils	Oligotrophic, eutrophic, and dry peat soils	Most organic horizons ≥ 50 cm will be hydric
China	Mineral	Reclaimed peat soils	Drained peat soils with some alteration	Most organic horizons ≥ 50 cm will be hydric
		Gleysols	Soils with gleyic features within 50 cm of surface	Will be hydric if low chroma matrix colors occur within 25 cm of surface in horizon at least 15 cm thick in most cases
	Organic	Histosols	Organic soils similar to Histosols in <i>Soil Taxonomy</i>	Will be hydric in most cases
Australia	Mineral	Hydrosols	Most of the soil profile saturated for 2–3 months in most years	Will be hydric if low chroma matrix colors occur within 25 cm of surface in a layer at least 15 cm thick
	Organic	Organosols	Have more than 40 cm of organic materials in upper 80 cm	Most will be hydric soils
	Brazil	Gleissolos	Includes most soils with aquic suborders as defined in <i>Soil Taxonomy</i>	Will be hydric if low chroma matrix colors occur within 25 cm of surface in a layer at least 15 cm thick
		Planossolos	Similar to the Alfisols and Ultisols in <i>Soil Taxonomy</i> with aquic conditions but probably saturated for shorter periods than Gleissolos	Will be hydric if low chroma matrix colors occur within 25 cm of surface in a layer at least 15 cm thick
		Organossolos	Includes the Histosols as defined in <i>Soil Taxonomy</i>	Will be hydric in most cases

Data from McBride MB (1994) *Environmental Chemistry of Soils*. Oxford University Press: New York. Fig. 7.5, maximum Eh at which the given elements are reduced based on laboratory experiments.

In the Russian soil classification system, the saturated mineral soils are included in two categories—Gley soils and Hydro-metamorphic soils (Table 4) (Shishov et al., 2005). These orders probably differ in the durations of their saturation periods. Hydric soils as defined earlier can occur in both categories if the color patterns diagnostic of hydric soils occur within the depths shown in Table 4.

The Russian system is unique in that it identifies two categories of organic soils, essentially drained and undrained, where hydric soils can be found. A soil that has been drained is still considered a hydric soil, under the criteria used in the United States, if it retains the soil characteristics that developed when it was saturated and anaerobic. Hydric soil features associated with mineral soils such as the gray matrix colors and redoximorphic features can persist after drainage for several years because the features remain intact under aerobic soils unless physically disturbed (Hayes Jr. and Vepraskas, 2000). Conversely, organic surface layers can oxidize overtime, and if they were the features used to designate the soil as hydric, then the soil may no longer be considered hydric following organic matter oxidation (Ewing and Vepraskas, 2006). However, it is likely in such cases that the deeper mineral layers in the soils would have also developed the characteristics diagnostic of hydric soils and could be used to support a hydric designation for the soil if they occur ≤ 25 cm from the surface after the original organic layers have disappeared.

The soil classification systems used in China and Australia are similar in that hydric soils can be found in a single category in both the mineral and organic soil groups (Gerasimova, 2010; Isbell, 2002). Most of the organic soils will be hydric under these classification schemes. The mineral soils will be hydric as long as low chroma colors are found within 25 cm of the surface in layers that are at least 15-cm thick. The Brazilian soil classification system also has two categories of mineral soils, Gleissolos and Planossolos. These are similar to those described for Gleysols and Planosols in the World Reference Base system (Table 3). The Organossolos are similar to the Histosols defined in *Soil Taxonomy* and the World Reference Base systems. However, the amount of organic C required for organic soil materials in the Brazilian system is about one-half that required for hydric soils, so not all Organossolos will be hydric soils (AcervoLima, 2022).

Despite the differences that occur across these classification systems, several themes are found in all in that hydric soils are associated with: (1) the persistence of water tables near the soil surface, (2) the accumulation of organic materials, and (3) the presence of low chroma colors and/or redoximorphic features that can be used to readily identify soils formed in wetlands.

Formation of hydric soil features

Oxidation-reduction reactions govern many of the chemical processes occurring in saturated soils and sediments (McBride, 1994). Four conditions are needed for a soil to become anaerobic through reducing chemical reactions: (1) the soil must be saturated or

inundated to exclude atmospheric O₂ (oxygen); (2) the soil must contain accessible organic tissues (organic matter) that can be oxidized or decomposed; (3) a microbial population must be respiring and oxidizing the organic tissues; and (4) the water should be stagnant or moving very slowly. Saturation or inundation is needed to keep the atmospheric O₂ out of the soil. Exclusion of atmospheric O₂ is the major factor that determines when reduction can occur in the soil. Presence of oxidizable organic tissues is probably the most important factor determining whether or not reduction occurs in a saturated soil (Beauchamp et al., 1989). Some soils are known to be saturated yet do not display any signs that reducing reactions such as Fe³⁺ reduction to Fe²⁺ have occurred. In most instances, such soils simply lack the oxidizable organic tissues needed to supply the electrons used in reducing reactions (Couto et al., 1985).

A respiring microbial population is essential to the formation of reduced soils. Bacteria are widespread, abundant, varied, and adapted to function in the climates in which they occur. As reducing chemical reactions have been studied more extensively in the field, it has become clear that they occur more frequently than originally thought (Megonigal et al., 1996). Lastly, stagnant or nearly-stagnant water is needed for reducing reactions to occur (Gilman, 1994). Fast-moving water, particularly at the surface, retards the onset of reduction by supplying oxygen to the soil. While the water is in motion, its O₂ content is difficult to deplete.

Types of morphological features of hydric soils

The redox reactions that create hydric soils produce indicators, signs that are easily seen or smelled, showing that they have occurred in the soil. A 'reduced' soil is one in which redox reactions have caused reduced forms of O, N, Mn, Fe, or S (sulfur) to be present in the soil solution. 'Reduced' is a general term that implies that some elements in addition to O₂ are present in their reduced electron valence states. Common reduced forms of elements or compounds that are found in hydric soils include: H₂O (water), N₂ (nitrogen), Mn²⁺ (manganese), Fe²⁺ (iron), and H₂S (hydrogen sulfide), while their oxidized counterparts are O₂, NO₃⁻ (nitrate), MnO₄⁻ (permanganate), Fe(OH)₃ (ferric hydroxide), and SO₄²⁻ (sulfate), respectively.

Morphological features of hydric soils include specific color patterns, odors, color changes that occur on exposure to air, or a specific kind of organic material. These features can occur at any depth in the soil, and the abundance of a given feature is variable. The principal reducing reactions that form hydric soils involve four elements (O, Mn, Fe, and S), which are used as electron acceptors in anaerobic microbial respiration. Table 5 shows that there are also four basic groups of features that are associated with the principal reducing chemical reactions. These feature groups will be identified by the major element related to their formation: (a) organic-C based features, (b) Mn-based features, (c) Fe-based features, and (d) S-based features. Selected examples of morphological features of reduction are listed in Table 5 and are described below.

Organic-carbon based morphological features

Organic-carbon based features consist of one of three kinds of materials: organic, mucky mineral, or mineral soil materials with a black color. These materials form either distinct horizons (O or A horizons) or occur as aggregates of organic-rich material. The O horizon thickness and state of decomposition must be considered when identifying hydric soils. The O horizons with a thickness ≥ 20 cm and a black or very dark gray color, regardless of the state of decomposition, are in most cases found only in soils that were periodically anaerobic and reduced (United States Department of Agriculture—Natural Resources Conservation Service (USDA-NRCS), 2018). Thinner layers can also be found in seasonally reduced soils, but their thickness and state of decomposition requirements vary for different textural groups and geographic areas. The A horizons consist of mineral (occasionally mucky mineral) soil material, and those that formed in reduced soils have moist Munsell colors with a value of 3 or less and a Chroma of 2 or less (Fig. 1A). The dark color is a direct result of relatively large amounts of organic matter that accumulated under reduced conditions. However, not all soils having dark colors necessarily experience periods of saturation and reduction. This is particularly

Table 5 Reducing reactions associated with the creation of morphological features found in hydric soils.

Reducing Reaction	Approximate Eh (pH 7) ^a	Morphological Feature Formed	
	mV	Group Name	Examples
O ₂ + 4e + 4H ⁺ → 2H ₂ O ^b	600	Organic C-based features	Oa and some black A horizons
MnO ₂ + 2e + 4H ⁺ → Mn ²⁺ + 2H ₂ O	300	Mn-based features	Mn concentrations, some depletions
Fe(OH) ₃ + e + 3H ⁺ → Fe ²⁺ + 3H ₂ O	175	Fe-based features	Fe concentrations and Fe depletions
SO ₄ ²⁻ + 8e + 10H ⁺ → H ₂ S + 4H ₂ O	-200	S-based features	Odor of rotten eggs

^aData from McBride MB (1994) *Environmental Chemistry of Soils*. Oxford University Press: New York.

^bReaction develops anaerobic conditions, and features form after O₂ depleted.

Modified from: United States Department of Agriculture – Natural Resources Conservation Service (USDA-NRCS). (2018) *Field Indicators of Hydric Soils in the United States (Version 8.2)*. Vasilas LM, Hurt GW and Berkowitz JF (eds.) Washington, DC: USDA-NRCS, in cooperation with the National Technical Committee for Hydric Soils.

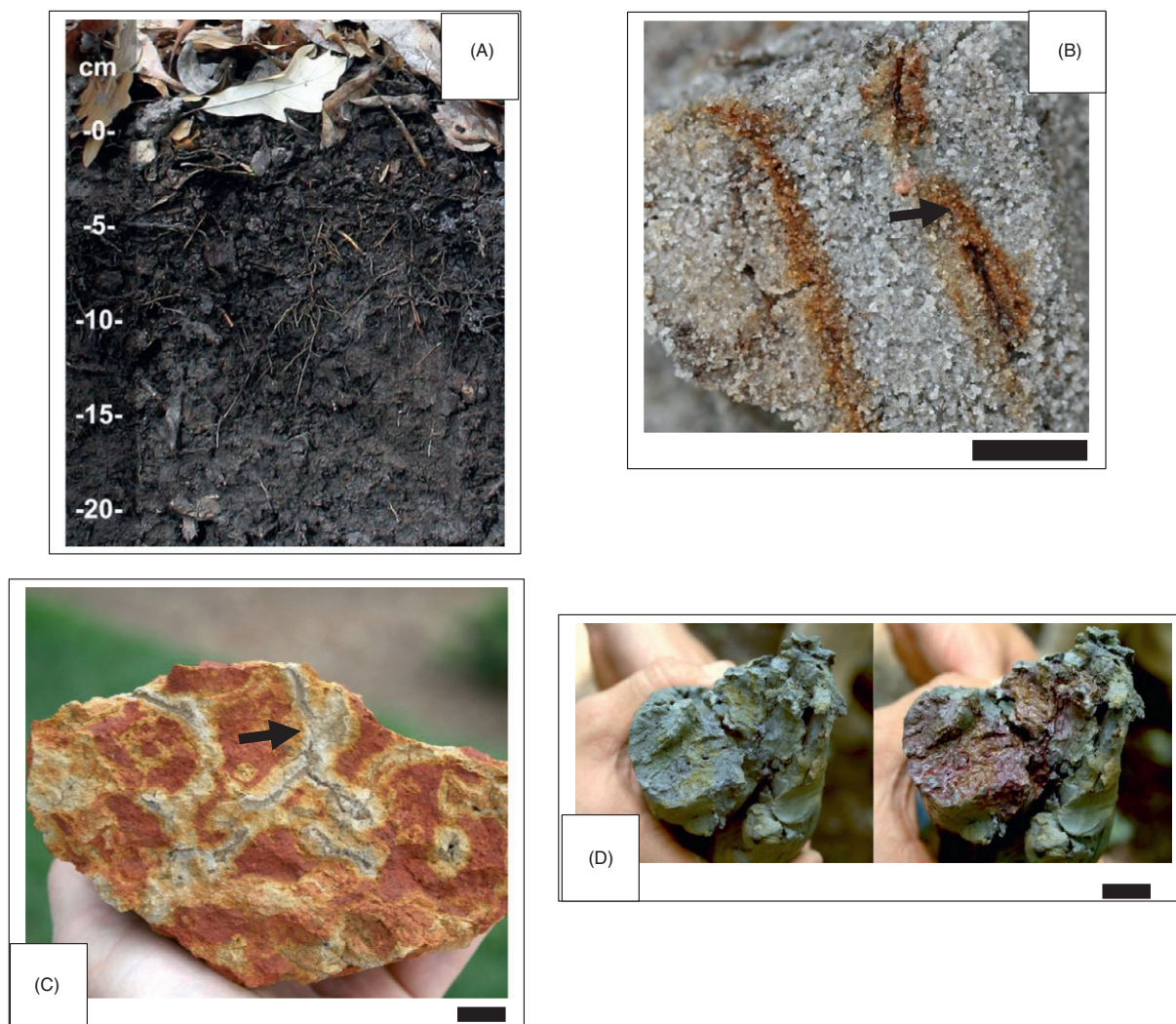


Fig. 1 Example of the morphological features that form hydric soil field indicators. (A) Black mineral horizon formed by C accumulation creating a C-based feature. (B) Fe-based features, or redoximorphic features in the form of an Fe pore lining (redox concentration, arrow) with a gray Fe depletion in the matrix. Scale bar is 1 cm long. (C) Fe depletions along root channels (arrow) with redox concentrations (Fe masses) in the matrix. Formation of these Fe-based features is illustrated in Fig. 2. Scale bar is 1 cm long. (D) A gley-colored soil material (bluish gray color) on left as found in the soil representing a reduced matrix, Fe-based indicator. Sample on right turned red when dipyrindyl dye was applied indicating that reduced Fe was present. Scale bar is 1 cm long. Photos courtesy of John Kelley, used with permission.

true of the Mollisols found in the Midwestern United States, Ukraine, and Northeastern China, and other ecosystems characterized by grassland prairies (Bell and Richardson, 1997).

Iron and manganese-based morphological features

Redoximorphic features are formed by the reduction, movement, and oxidation of iron (Fe) and manganese (Mn) compounds (Vepraskas, 2015; Vepraskas and Vaughan, 2016). These features form the gray, red, yellow, brown, or black mottled color patterns that are normally associated with saturated and reduced soils. Redoximorphic features are the most widely observable morphological features formed by reduction, and are most commonly used to identify the spatial extent of hydric soils in the landscape. There are three basic kinds of redoximorphic features: redox concentrations, redox depletions, and the reduced matrix.

Redox reactions affect Fe and Mn similarly, and the two elements frequently occur together. Manganese minerals are black in their oxidized form, whereas oxidized Fe minerals are generally yellow, brownish red, or red. Iron is usually in greater abundance than Mn, but small quantities of Mn can cause some redox concentrations to appear black (Gallaher et al., 1974). Manganese can be abundant in certain soils, such as those with a pH >7, or in some clay soils having Munsell hues of 5YR or redder (e.g., Moreland series reported in Hudnall et al., 1990). When Mn is abundant, it can prevent the reduction of Fe and the formation of gray soil

colors because it is reduced before Fe (McBride, 1994). The remainder of this discussion will focus on Fe, but Mn should also be assumed to be included because it is subject to the same processes.

Redox concentrations

Redox concentrations are features formed when Fe oxides or hydroxides have accumulated at a point or around a large pore such as a root channel (Fig. 1B). They have been defined as “bodies of apparent accumulation of Fe-Mn oxides and hydroxides” (Vepraskas, 2015). This means that they appear to have formed by Fe or Mn moving into an area, oxidizing, and precipitating. Redox concentrations contain more Fe³⁺ oxides and hydroxides than were found in the soil matrix originally. Three kinds of redox concentrations have been defined: Fe pore linings, Fe masses, and Fe nodules and concretions. These differ in their hardness and also in where they occur in the soil.

Pore linings are accumulations of Fe oxides and hydroxides that lie along ped surfaces or root channels (Fig. 1B). These features occur in the soil and not directly on the root. They are similar to oxidized rhizospheres, but these are thought to form on root tissue while the root is alive (Mendelsohn et al., 1995), whereas pore linings do not need a live root in order to form.

Iron masses are simply soft accumulations of Fe³⁺ oxides and hydroxides that occur in the soil matrix, away from cracks or root channels (Fig. 1C). They can be of any shape. The masses are easily crushed with the fingers because the concentration of Fe is not great enough to cement the soil particles into a solid mass. Pore linings differ from iron masses only in where they occur in the soil: masses occur in the matrix, while the pore linings must be along root channels or cracks. The colors of the two features are similar.

Nodules and concretions are hard, generally spherical-shaped bodies made of soil particles cemented by Fe oxides or hydroxides. They range in size from less than 1 mm to over 5 cm in diameter. When broken in half and examined, the concretions are seen to consist of concentric layers like an onion, while no layers are seen in nodules. Most people seem to use the two terms interchangeably, and there is no special significance attached to the layered structure other than it shows that the concretion formed in episodes over time.

Redox depletions

Redox depletions are zones formed by the loss of Fe and other components (Fig. 1C). They have been defined as “bodies of low Chroma (<2), having values of 4 or more where Fe-Mn oxides alone have been stripped out or where both Fe-Mn oxides and clay have been stripped out” (Vepraskas, 2015). This definition is used to meet the soil classification requirements for aquic conditions as set out in *Soil Taxonomy* (Soil Survey Staff, 2014). Redox depletions in principle could form with Chromas >2 as long as they developed in a soil horizon whose matrix lost Fe by chemical reduction processes.

Reduced matrix

The reduced matrix has been defined as a soil matrix that has a “low chroma color *in situ* because of the presence of Fe²⁺, but whose color changes in hue or chroma when exposed to air as the Fe²⁺ is oxidized to Fe³⁺” (Vepraskas, 2015). The time needed for the color change to occur was set for practical reasons at 30 min, because waiting longer would interfere with the completion of field work. However, there are few data on how long such color changes require and changes often occur following only a few minutes of exposure to oxygen. The *United States Army Corps of Engineers* (U.S. Army Corps of Engineers (USACE), 2012) provides guidance for identifying a reduced matrix in the field. It can also be detected by spraying a field moist sample of soil with a dye such as α , α' -dipyridyl that reacts with Fe²⁺ (Fig. 1D). If a positive reaction with the dye occurs, then it should be assumed that a reduced matrix is present.

Gley soil colors

Gley colors consist of Munsell hues found on the “Gley Pages” of a Munsell Color Chart that display a value of 4 or more. If the color of the soil does not change on air drying, the material with the gley hue is probably a redox depletion (e.g., Fe depletion). This occurs when the Fe is reduced, translocated and therefore removed so that it is no longer present in the soil matrix. If a color change (e.g., reddening of the matrix) occurs on drying, the material is probably a reduced matrix. A reduced matrix indicates the Fe is reduced but not translocated from the soil matrix so that with exposure to oxygen, the iron is oxidized to a colored form.

The gley hues can, in some cases, be unique minerals that contain a reduced form of Fe that is combined with an anion of phosphate, sulfate, carbonate, or other compound (Schwertmann and Taylor, 1989). These hues are a morphological indicator of reduction when their values are 4 or more. When values are less than 4, the colors are probably too dark to separate gley hues accurately from other kinds of soil features.

Formation of redoximorphic features

Redoximorphic features are found in most mineral hydric soils. They form after one or more of the following three processes have occurred: (1) Fe³⁺ cations in oxides or hydroxides have been reduced, (2) the solubilized Fe²⁺ has moved to another portion of the soil, and (3) the Fe²⁺ has been oxidized to form an Fe mass, pore lining, or nodule. A reduced matrix requires that only the first step occurs. Iron depletions require that the first two steps occur, while formation of redox concentrations requires that all three steps occur.

Redox depletions: The process of redox depletion formation is shown in Fig. 2. Prior to forming depletions the soil matrix had a uniform brown color throughout. The color came from Fe³⁺ oxide or hydroxide minerals that coated soil particles, such as

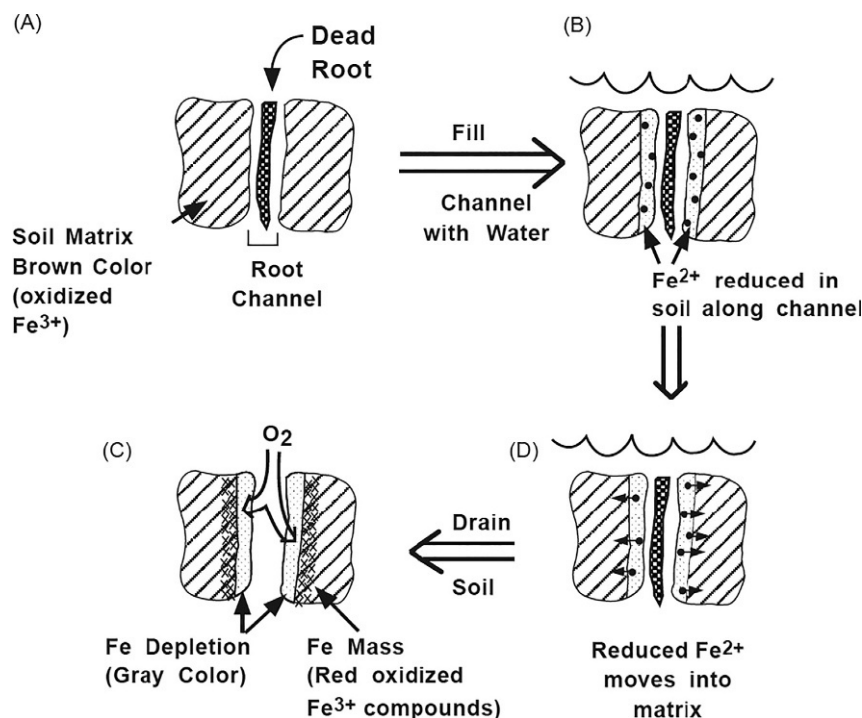


Fig. 2 Formation of redox depletions (Fe depletions) around a root channel. Initially (A) the soil is uniformly brown throughout its matrix. The channel contains a dead root, which is being decomposed by bacteria. Upon flooding (B), the channel is filled with water, and all O₂ dissolved in the water is reduced. When the water is anaerobic, the bacteria reduce Fe oxides and hydroxides in the soil surrounding the channel. The Fe²⁺ dissolves and moves away from the channel (C) leaving the soil particles around the channel stripped of the Fe coatings. If the stripped grains are largely composed of quartz and/or uncoated clay minerals, they will be gray in color (D). The Fe²⁺ may subsequently oxidize to form a reddish brown Fe mass. Adapted from Vepraskas MJ (2015) *Redoximorphic Features for Identifying Aquic Conditions*. Technical Bulletin 301, North Carolina Agricultural Research Station: Raleigh, NC.

sand, silt, and clay grains. Each of these grains had a coating of an Fe³⁺ compound that effectively painted the particle surface brown.

The soil shown in Fig. 2A has a channel containing a dead root that is being decomposed by bacteria. If the channel is filled with air, oxidation of the organic matter releases electrons which are used to reduce O₂ to water. Provided that the channel contains air, the only reduction that takes place is that of O₂ to water, and there is no change in the color of the soil around the root.

If the channel fills with water, however, the supply of O₂ from the atmosphere is cut off (Fig. 2B). Microbial respiration still occurs, and the organic tissue continues to be oxidized. The O₂ dissolved in the soil water is quickly depleted because it is the primary electron acceptor used in the respiration process. After O₂ is depleted, the water becomes anaerobic, and electrons generated during microbial respiration begin to interact with alternative electron acceptors (e.g., Fe³⁺), which induces a change in soil color.

When the electrons are transferred to Fe³⁺ atoms in oxides or hydroxides, the reduction of the Fe causes two changes in the minerals: they lose their color and the Fe²⁺ dissolves. The dissolved Fe²⁺ moves off the particle surfaces and may diffuse through the soil matrix, or it may be carried to other parts of the soil in moving water (Fig. 2C). As Fe²⁺ leaves the particle surfaces, the color of the soil around the channel changes and the soil gradually becomes gray in color. In Munsell terminology the soil Chroma decreases and Value increases until all Fe has been removed from the surfaces of particles lying near the channel. When the soil drains and O₂ enters the soil, the newly formed Fe depletion feature will retain its gray color because it is the color of the uncoated mineral soil grains. Oxygen penetrating into the matrix may oxidize the Fe²⁺ and cause the formation of pore linings or Fe masses elsewhere in the soil profile (Fig. 2D).

Stripping the soil particles of Fe³⁺ mineral coatings changes the soil's color to the natural color of the soil minerals (i.e., the 'naked grain'). Normally this is a gray color when the particles consist of quartz or a clay mineral such as kaolinite. This gray color is relatively permanent and is not affected appreciably by additional periods of saturation and reduction. The only way the color of the gray soil particles could become brown again is if Fe²⁺ were moved onto the particle surfaces and reoxidized to recoat the particle surfaces with more of the 'paint' composed of Fe³⁺ minerals.

This is the basic process that forms redox depletions. It is easiest to see in the field when the organic matter occurs as a root (Fig. 1C). In such cases the gray depletions form cylindrical features around root channels. However, if roots are closely spaced along a crack or ped surface, then the depletions will have a planar shape as they coat the surface of the crack, or they may occupy entire layers.

In A horizons the organic matter that starts the depletion formation can be a piece of leaf tissue or a fragment of some other part of the plant. This is an isolated source of organic tissue. The depletions that form around this tissue tend to be spherical. The processes that form the depletion are the same as those shown in Fig. 2.

Field indicators of hydric soils

The chemical reactions that create anaerobic conditions as well as Fe reduction in near surface layers, create unique hydric soil characteristics that can be seen, felt, or smelled as listed in Table 6. These characteristic features are used to identify hydric soils at wetland sites by using *Hydric Soil Field Indicators* (United States Department of Agriculture—Natural Resources Conservation Service (USDA-NRCS), 2018). They are applied in conjunction with indicators of hydrophytic vegetation and wetland hydrology to identify wetlands for a variety of natural resource management applications (U.S. Army Corps of Engineers (USACE), 2012; Tiner, 2016).

The field indicators of hydric soils specify combinations of soil horizon thickness, color, and redoximorphic feature characteristics, known to develop when soils form in areas subject to prolonged saturation and anaerobic conditions, therefore linking soil morphology with the definition of hydric soils. The field indicators also include information about the soil texture and the region of the United States where the field indicator can be applied. For example, the field indicator S5—Sandy Redox states the following:

“S5—Sandy Redox. For use in all LRRs [Land Resource Regions], except for W, X, and Y. A layer starting at a depth ≤ 15 cm (6 inches) from the soil surface that is at least 10 cm [4 inches] thick and has a matrix with 60% or more chroma of 2 or less and 2% or more distinct or prominent redox concentrations occurring as soft masses and/or pore linings.”

The field indicators of hydric soils were developed for three different textural groups indicated by the alphanumeric designation (e.g., S5) associated with each indicator. Field indicators designated with an ‘A’ (i.e., all soil types) can be applied to soil layers (or a combination of layers) composed of organic materials or any mineral soil textures; indicators designated with an ‘S’ (i.e., sandy soil) are only applicable in mineral soil layers composed of USDA soil texture classes sand, fine sand, very fine sand, loamy coarse sand, loamy sand, and loamy fine sand; and indicators designated with an ‘F’ (i.e., fine soil) are applicable to all mineral soil textures not addressed by the ‘S’ indicators.

This system reflects soil texture and organic matter contents, and is not directly related to any soil classification system. It accounts for the different water holding capacities and capillary fringe potentials that vary with soil texture, and allows for the application and use of multiple field indicators of hydric soils within soil profiles composed of varying soil textures. The field indicator names provide insight into the general characteristic(s) of the indicator. For example, S5—Sandy Redox requires a sandy soil texture and the presence of redoximorphic features.

The geographic area where a given field indicator of hydric soils can be applied in the United States is defined using Land Resource Regions (LRRs) and Major Land Resource Areas (MLRAs) (United States Department of Agriculture, Natural Resources Conservation Service, 2006). For example, the field indicator S5—Sandy Redox can be applied in all areas of the United States except in Alaska (LRRs W, X, Y). This approach recognizes the differences in pedogenic processes and morphology related to varying climate, parent materials and other factors that occur across the United States. Field indicators approved for cold regions with short growing seasons where microbial activity is depressed require thicker organic matter layers compared with field indicators applied in hot, tropical climates where the year-round growing season and increased microbial respiration limit the accumulation of carbon in most soils. The LRRs are based largely on soil temperature regimes, soil moisture regimes, and physiography, and can be extrapolated to other parts of the world using the definitions of the regions given in United States Department of Agriculture, Natural Resources Conservation Service (2006).

The guidance document “Field Indicators of Hydric Soils in the United States” currently identifies 45 field indicators of hydric soil approved for use, as well as several field indicators approved for testing. Indicator “F3—Depleted matrix” is the field indicator

Table 6 Common field indicators of hydric soils in the United States.

Indicator	General diagnostic characteristics
A1—Histosol	Soil layers composed of organic material ≥ 40 cm thick occurring within the upper 80 cm of the soil surface
A11—Depleted Below Dark Surface ^a	A gray or gley soil layer with a depleted matrix ≥ 15 cm thick beginning within ≤ 30 cm of the soil surface. Soil layers above the depleted matrix must be dark (i.e., Munsell value ≤ 3 , Chroma ≤ 1 for sands; Value ≤ 3 , Chroma ≤ 2 for all other textures)
A4—Hydrogen Sulfide Odor	A hydrogen sulfide odor encountered within ≤ 30 cm of the soil surface
F3—Depleted Matrix ^a	A gray or gley loamy/clayey soil layer with a depleted matrix either ≥ 5 cm thick beginning within ≤ 10 cm of the soil surface, or ≥ 15 cm thick beginning within ≤ 25 cm of the soil surface
F13—Umbric Surface ^a	A dark loamy/clayey soil layer ≥ 15 cm thick beginning within ≤ 15 cm of the soil surface with Munsell value 3 or less and Chroma 1 or less, and immediately underlain by a layer ≥ 10 cm thick with Chroma of ≤ 2
S5—Sandy Redox ^a	A sandy soil layer ≥ 10 cm thick beginning within ≤ 15 cm of the soil surface with Chroma ≤ 2 and $\geq 2\%$ redoximorphic concentrations

^aField indicators commonly observed at the hydric soil boundary (i.e., often used for hydric soil delineation)

Modified from Berkowitz JF, Vepraskas MJ, Vaughan KL and Vasilas LM (2021) Development and application of the Hydric Soil Technical Standard. *Soil Science Society of America Journal* 85: 469–487.

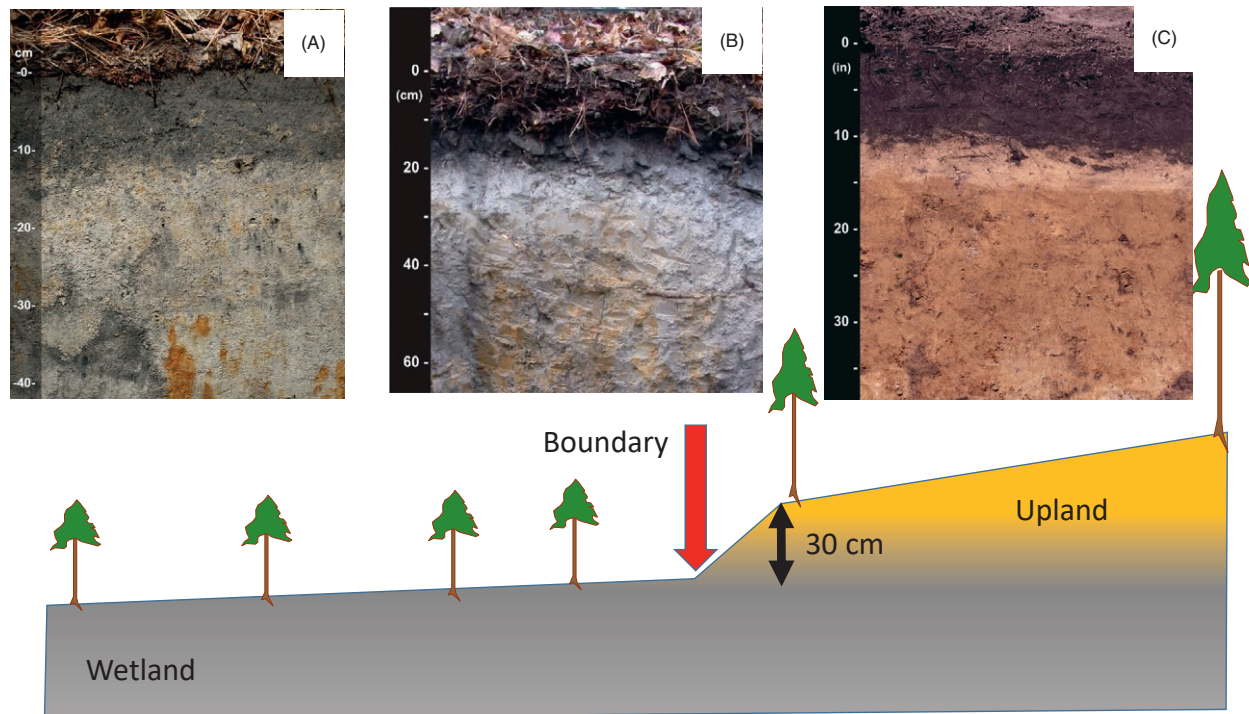


Fig. 3 Example of how hydric soil field indicators are used to identify hydric soil boundaries. The Coastal Plain landscape is nearly level with an elevation change of approximately 30 cm. The soil scientist begins by examining soil near the center of the wetland, which in this example is shown in (A). This soil meets the hydric soil field indicator Umbric Surface. Having confirmed the site had a hydric soil, the soil scientist moves toward the edge near the upland. Near the edge, the soil described is shown at (B), and this met the hydric soil field indicator Depleted Matrix. From this point the soil scientist walks up the slope into the upland and examines the soils. In this landscape, the soil at (C) occurred, and this did not meet a hydric soil field indicator. The hydric soil boundary occurred between soils (B) and (C). The boundary was placed at the base of the slope, because soil borings showed that field indicators were not met on the slope.

most frequently used to identify hydric soils and delineate their boundaries (Berkowitz, 2011). As its name implies, the depleted matrix has a matrix formed by redox depletion with a high Munsell Value (≥ 4) and low Chroma (≤ 2), causing the soil to appear gray. Redox concentrations are also required in A and E horizons.

To delineate the hydric soil boundary using the field indicators of hydric soils, practitioners typically begin in wetter areas, identify the diagnostic characteristics associated with one or more of the field indicators, then move upslope in the landscape along a line toward drier areas, examining the soil periodically until the diagnostic criteria are no longer met. For example, when applying a field indicator that requires a Munsell Chroma ≤ 2 and $\geq 2\%$ redoximorphic features, the hydric soil boundary occurs at the location where either the chroma exceeds 2.0 or the soil first exhibits $<2\%$ redoximorphic features.

This approach is illustrated in Fig. 3 which shows a landscape described by Hayes Jr. and Vepraskas (2000). The Coastal Plain landscape is almost level with small changes in elevation that separate wetland areas from uplands as shown. Soil A in Fig. 3 occurs in the wetland center and meets the hydric soil field indicator called “Umbric Surface” which is characterized by a black A horizon (10YR 3/1) beginning at the mineral surface. It is underlain by material with a Chroma of 2 or less. At the edge or hydric soil boundary, soil B meets the hydric soil field indicator termed the ‘Depleted Matrix’ characterized by a layer at a depth of 20–35 cm with a color of 10YR 4/2 and approximately 2% redox concentrations seen as orange mottles. Between soils B and C in Fig. 3, no soil meets a hydric soil field indicator because the subsoils lose their low (≤ 2) chroma color. The boundary of the hydric soils in this landscape would be placed where the elevation starts to rise to approximately 30 cm (i.e., base of slope) and the soil color changes occur.

The hydric soil technical standard

Over the past several decades, research in soil science has linked the unique biogeochemical processes occurring in wetlands with the distinct morphological patterns occurring in hydric soils (Vepraskas and Vaughan, 2016). This research facilitated development of the field indicators of hydric soils described above, which are approved by the National Technical Committee for Hydric Soils

Table 7 Summary of the sampling techniques and minimum technical requirements of the Hydric Soil Technical Standard.

Measurement Technique	Property Criteria
Anaerobic conditions criteria	
Alpha-alpha dipyridyl dye	A positive dye reaction during two applications that occur 2 weeks apart. The reaction occurs within a 10 cm thick layer in the upper 30 cm for most soils, a 6.25 cm thick layer within the upper 12.5 cm in sandy soils, or a 5 cm layer within the upper 10 cm in soils that flood or pond.
IRIS devices	At least three of IRIS tubes must show a minimum of 30% iron removal within a zone 15 cm or more thick. The zone of removal must begin within 15 cm of the soil surface. Tubes should be deployed for a minimum of one wet-dry cycle not to exceed one growing season.
Pt electrodes	A minimum of 3 of 5 Pt electrodes display Eh values below the threshold defined by $Eh < 595 - 60(\text{pH})$. Pt electrodes are installed 25 cm for most soils, 12.5 cm in sandy textured soils, and 10 cm in soils that inundate by flooding or ponding.
Soil saturation, flooding, or ponding criteria	
Shallow groundwater monitoring wells	Soil saturation occurs within 25 cm of the soil surface for ≥ 14 consecutive days in most soils. Vertisols in Louisiana and Texas require soil saturation for ≥ 7 consecutive days with saturation occurring $\geq$ days annually.
Precipitation normality criteria	
Analysis of precipitation normality	Analysis using the Direct Antecedent Rainfall Evaluation Method, Moving Total Antecedent Rainfall Method, or Adjusted Moving Total Antecedent Rainfall Method to determine whether the study occurred during a period of normal, above-normal, or below-normal precipitation based on the 30th and 70th percentile averages from 30-year precipitation records (Sumner et al., 2009).

Adapted from Berkowitz et al. (2021).

(NTCHS) for application in the United States. To ensure that field indicators of hydric soils are defensible, the NTCHS developed a scientifically based and repeatable approach to evaluate the hydric status if a soil known as the Hydric Soil Technical Standard (HSTS). The HSTS provides a mechanism to: (1) initiate new field indicators of hydric soils, (2) alter existing field indicators, (3) identify hydric soils in areas lacking an approved field indicator, and (4) evaluate the current functional status of a hydric soil. For example, Berkowitz and Sallee (2011) utilized the HSTS to propose a new field indicator for an area that contained hydric soils but that did not meet a currently approved field indicator, and also expanded the applicable region of several other field indicators with HSTS data. The HSTS documents that a soil is saturated and anaerobic, thus demonstrating that the soil meets the definition of a hydric soil noted earlier (Berkowitz et al., 2021). As a result, the HSTS requires that three technical criteria be met: (1) anaerobic conditions; (2) soil saturation in the upper 25 cm of the soil profile for at least 14 consecutive days for most soils, or for 7 consecutive days occurring for a minimum of 28 days annually for Vertisols in Louisiana and Texas; and (3) the saturation and anaerobic conditions criteria must be met following a period of normal or drier than normal precipitation when soil microbes are active.

Under the HSTS framework, practitioners document the durations of anaerobic conditions and saturation. Precipitation must also be recorded to determine if the anaerobic and saturated conditions occur during wet, normal, or dry years using a variety of techniques (Sumner et al., 2009). The HSTS documents whether or not a soil currently functions as a hydric soil and does not address the conditions under which the soil formed. As a result, the HSTS is not designed or intended to be used to overrule or invalidate a hydric soil determination based on the presence of one or more field indicators of hydric soils. The field indicators reflect natural processes that develop over time during soil formation and generally provide the best evidence that hydric soils are present on a site. However, the NTCHS acknowledges that some hydric soils do not display approved field indicators of hydric soils for a variety of reasons (e.g., problematic parent materials, recently deposited or disturbed soils). In such cases, the HSTS is a valuable tool to evaluate the current hydric status of those soils. In addition, application of the HSTS improves our understanding of hydric soil function and morphology across a variety of environmental gradients, management scenarios, and disturbance regimes (Table 7).

Summary

Hydric soils form in wetlands where saturation persists long enough to create anaerobic conditions. The presence of anaerobic conditions induces changes in soil morphology that have been used to develop diagnostic criteria in support of wetland protection laws in the United States and elsewhere. Soil classification systems provide useful corollaries to identify areas where hydric soils are likely occur, based on the presence of aquic conditions, carbon accumulation, or the transformation and translocation of redox active elements such as Fe and Mn. The hydric soil field indicators incorporate a variety of characteristics such as redoximorphic features to help locate wetland boundaries in support of natural resource management. In cases where field indicators of hydric soils are absent, the Hydric Soil Technical Standard provides a mechanism to link observed hydric soil morphologies with biogeochemical processes and ecological functions.

References

- AcervoLima (2022) *Organosol*. <http://wiki.acervolima.com/organosol/>. Accessed 27 January 2022.
- Beauchamp E, Trevors JT, and Paul JW (1989) Carbon sources for bacterial denitrification. *Advances in Soil Science* 10: 113–142.
- Bell JC and Richardson JL (1997) Aquic conditions and hydric soil indicators for Aquolls and Albolls. In: Vepraskas MJ and Sprecher SW (eds.) *Aquic Conditions and Hydric Soils: The Problem Soils*. Madison, WI: Soil Science Society of America. SSSA Special Publication No. 50.
- Berkowitz JF (2011) Recent advances in wetland delineation—Implications and impact of regionalization. *Wetlands* 31(3): 593–601.
- Berkowitz JF and Saltee JB (2011) Investigating problematic hydric soils using water table measurements, IRIS tubes, soil chemistry, and application of the Hydric Soils Technical Standard. *Soil Science Society of America Journal* 75(6): 2379–2385.
- Berkowitz JF, Vepraskas MJ, Vaughan KL, and Vasilas LM (2021) Development and application of the Hydric Soil Technical Standard. *Soil Science Society of America Journal* 85(3): 469–487.
- Buol SW, Southard RJ, Graham RC, and McDaniel PA (2011) *Soil Genesis and Classification*, 6th edn, Chichester, UK: J. Wiley & Sons.
- Couto W, Sanzowicz C, De A, and Bacellos O (1985) Factors affecting oxidation-reduction processes in an Oxisol with a seasonal water table. *Soil Science Society of America Journal* 49: 1245–1248.
- Cowardin LM, Carter V, Golet FC, and LaRoe ET (1979) *Classification of Wetlands and Deepwater Habitats of the United States*. FWS/OBS-79-31 Washington, DC: U.S. Fish and Wildlife Service.
- Ewing JM and Vepraskas MJ (2006) Estimating primary and secondary subsidence in an organic soil 15, 20, and 30 years after drainage. *Wetlands* 26: 119–130.
- Federal Register (1994) *Changes in Hydric Soils of the United States*. Washington, DC. (Definition of hydric soils)
- Gallaher RN, Perkins HF, and Tan KH (1974) Classification, composition, and mineralogy of iron glauclites in a Southern Coastal Plain soil. *Soil Science* 117: 155–164.
- Gerasimova MI (2010) Chinese soil taxonomy: between the American and international classification systems. *Eurasian Soil Science* 43: 945–949.
- Gilman K (1994) *Hydrology and Wetland Conservation*. New York, NY: John Wiley & Sons.
- Hayes WA Jr. and Vepraskas MJ (2000) Morphological changes in soils produced when hydrology is altered by ditching. *Soil Science Society of America Journal* 64: 1893–1904.
- Hudnall WH, Szogi A, Touchet BA, Diagle J, Edwards JP, and Lynn WC (1990) Guidebook for Louisiana. In: *VIII International Soil Correlation Meeting: Classification and Management of Wet Soils*, Agronomy Department, Louisiana State University, Baton Rouge.
- Hurt GW and Noble CV (2016) Delineating hydric soils. In: Vepraskas M and Craft C (eds.) *Wetland Soils: Genesis, Hydrology, Landscapes, and Classification*. CRC Press.
- Isbell RF (2002) *The Australian Soil Classification System*, Revised Edn Collingwood, VIC: CSIRO Publishing.
- IUSS Working Group WRB (2015) *World Reference Base for Soil Resources 2014. Update 2015 International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*Rome: FAO. World Soil Resources Reports No. 106.
- McBride MB (1994) *Environmental Chemistry of Soils*. New York, NY: Oxford University Press.
- Megonigal JP, Faulkner SP, and Patrick WH Jr. (1996) The microbial activity season in southeastern hydric soils. *Soil Science Society of America Journal* 60: 1263–1266.
- Mendelsohn IA, Kleiss BA, and Wakeley JA (1995) Factors controlling the formation of oxidized root channels: A review. *Wetlands* 15(1): 37–46.
- Schwertmann U and Taylor RM (1989) Iron oxides. In: Dixon JB and Weed SW (eds.) *Minerals in Soil Environments*, 2nd ed. Madison, WI: Soil Science Society of America.
- Shishov LL, Tonkonogov VD, Gerasimova MI, and Lebedeva II (2005) New classification system of Russian soils. *Eurasian Soil Science* 38(supplement 1): S35–S43.
- Soil Survey Staff (2014) *Keys to Soil Taxonomy*, 12th edn, Washington, DC: United States Department of Agriculture, Natural Resources Conservation Service.
- Sumner JP, Vepraskas MJ, and Kolka RK (2009) Methods to evaluate normal rainfall for short-term wetland hydrology assessment. *Wetlands* 29(3): 1049–1062.
- Tiner RW (2016) *Wetland Indicators: A Guide to Wetland Formation, Identification, Delineation, Classification, and Mapping*. CRC Press.
- U.S. Army Corps of Engineers (USACE) (2012) In: Wakeley JS, Lichvar RW, Noble CV, and Berkowitz JF (eds.) *Regional Supplement to the Corps of Engineers Wetland Delineation Manual: Northcentral and Northeast Region (Version 2.0)*. Vicksburg, MS: U.S. Army Engineer Research and Development Center. ERDC/EL TR-08-28.
- United States Department of Agriculture, Natural Resources Conservation Service (2006) *Land Resource Regions and Major Land Resource Areas of the United States, the Caribbean, and the Pacific Basin*. U.S. Department of Agriculture Handbook 296. Available at: https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_050898.pdf (Verified 6 May 2022).
- United States Department of Agriculture—Natural Resources Conservation Service (USDA-NRCS) (2018) In: Vasilas LM, Hurt GW, and Berkowitz JF (eds.) *Field Indicators of Hydric Soils in the United States (Version 8.2)*. Washington, DC: USDA-NRCS. In cooperation with the National Technical Committee for Hydric Soils.
- Vepraskas MJ (2015) *Redoximorphic Features for Identifying Aquic Conditions*. Technical Bulletin 301 Raleigh, NC: North Carolina Agricultural Research Station.
- Vepraskas MJ (2016) History of the concept of hydric soils. In: Vepraskas MJ and Craft CB (eds.) *Wetland Soils: Genesis, Hydrology, Landscapes, and Classification*. Boca Raton, FL: CRC Press.
- Vepraskas MJ and Vaughan KL (2016) Morphological features of hydric and reduced soils. In: Vepraskas MJ and Craft CB (eds.) *Wetland Soils: Genesis, Hydrology, Landscapes, and Classification*. Boca Raton, FL: CRC Press.
- Vepraskas MJ, Polizzotto M, and Faulkner SP (2016) Redox chemistry of hydric soils. In: Vepraskas MJ and Craft CB (eds.) *Wetland Soils: Genesis, Hydrology, Landscapes, and Classification*. Boca Raton, FL: CRC Press.

Wetland soils

Cornelius Wilhelm van Huyssteen, University of the Free State, Bloemfontein, South Africa

© 2023 Elsevier Ltd. All rights reserved.

Introduction	437
Ramsar convention	437
Classification	437
Estuarine wetlands	438
Palustrine wetlands	438
Ecosystem services	438
Chemistry	439
Reducing conditions	440
Morphology	440
Hydric soils	443
Measurement	444
Hydrology	444
Redox potential and pH	445
Indicators	445
Conclusions	445
References	445

Abstract

Wetland ecosystems occur at the interface between aquatic and terrestrial ecosystems. The soils in wetlands are periodically or permanently saturated with water that excludes oxygen from the soil environment. Therefore, only vegetation adapted to growth in these saturated conditions are able to survive in wetland ecosystems. Wetland soils have unique chemistry due to the deficiency of oxygen, where microorganisms utilize electron acceptors other than oxygen in respiration. The reduced elements thus produced, specifically Mn^{2+} (manganese) and Fe^{2+} (iron) become mobile and are redistributed, resulting in unique redox morphology. It is this morphology that is used to characterize wetland soils.

Glossary

Aerobic An environment in which molecular oxygen is present (rH greater than 20).

Anaerobic An environment in which molecular oxygen is essentially absent (rH less than 20).

Drained wetland A wetland where ground or surface water was artificially removed.

Flooded soil Where the soil surface is temporarily covered with water from any source, e.g., from streams, upslope, tides, or a combination thereof.

Frequently flooded, ponded or saturated soil Flooding, ponding, or saturation is likely to occur often under the normal weather conditions (i.e., >50% probability in any year, or >50 times in 100 years).

Gleyic color pattern Bright (Munsell color chroma >2) macro pores that form due to oxidation, with grey (Munsell color chroma ≤2) matrices that form due to reduction, indicative of longer periods of ground water saturation.

Hydric soil A soil that formed under conditions of water saturation, flooding, or ponding long enough during the growing season to develop anaerobic conditions.

Long duration A single event that lasts from 7 days to 1 month.

Stagnic color pattern Grey (Munsell color chroma ≤2) macro pores that form due to reduction, with bright matrices (Munsell color chroma >2) that form due to oxidation, indicative of shorter periods of surface water saturation.

Very long duration A single event that lasts longer than 1 month.

Key points

- Wetlands form at the interface between aquatic and terrestrial ecosystems.
- Water saturation in wetland soils excludes oxygen from these environments.
- Only vegetation adapted to these water saturated conditions can therefore survive.
- Anaerobic microorganisms additionally reduce elements such as Mn^{2+} (manganese) and Fe^{2+} (iron).
- These processes can lead to unique soil chemistry and morphology, known as redox morphology.

Introduction

Wetlands are defined as ecosystems that occur at the interface between the terrestrial (upland) and the aquatic (water) ecosystems. Therefore, they occur where land and water meet, and thus have alternating aerated and water saturated conditions in their soils. The habitat in wetlands is therefore neither fully aquatic nor fully terrestrial. The soils in wetlands thus have too much water, resulting in the exclusion of air that causes a deficiency in oxygen. The net effect is that only plants that have adapted physiologically and morphologically to growth in water saturated soil environments can survive. Because wetlands occur at the interface between terrestrial and aquatic ecosystems they are typically found lower in the landscape along rivers, lakes, bays, and deltas. They can, however, also occur higher up in the landscape in depressions where artesian water typically accumulates.

Globally, inland and coastal wetlands comprise more than $12.1 \times 10^6 \text{ km}^2$ of the land surface, of which 54% is permanently inundated while 46% is seasonally inundated. The areal extent of natural wetlands has declined by about 35% since 1970, whereas the areal extent of human-made wetlands (mainly for paddy field rice production and water reservoirs) have almost doubled now to comprise 12% of all wetlands.

Traditional attitudes to wetlands are quite negative, as reflected in words such as 'bogged down,' 'swamped,' 'muddied,' 'mired,' 'murky,' 'morass,' 'malarial,' and 'quagmire.' Wetlands have thus historically been treated as wastelands that need to be drained or filled in order to be farmed or to be built upon (the latter is a probable source of the word 'landfill' constructed on a 'wasteland'). Drainage schemes, from large government initiatives to small local endeavors, to rid wetlands of 'excess' water have therefore dominated the historic approach of humans to wetlands. Recently however, human perception of wetlands has changed to recognize the many vital functions and ecosystem services that they provide within the broader environment. These wetland ecosystem services include providing a wildlife habitat, flood water and erosion control, aesthetics and recreation facilities, areas for carbon sequestration, and high biological productivity.

It is estimated that 50% of all the wetlands in the United States have been drained since colonial times and that 90% of all present-day agricultural lands in the United Kingdom were former wetlands. Wetlands therefore remain under threat either from direct or indirect causes. Indirect causes such as sea-level rise can inundate coastal wetlands, while direct causes include changes in wetlands' hydrological regimes due to upstream water extraction, construction of dams, dikes, or levees, which all reduce the water supply to downstream wetlands.

Wetlands are typically defined by the presence of wetland vegetation, wetland soils, and or topographical position. It is, however, in the soil where water saturation occurs and controls the alternating oxidizing and reducing environments. It is further possible that the wetland vegetation can be destroyed, leaving soil and topography as the only two permanent wetland indicators. The excessive water in wetland soils causes anaerobic conditions that lead to a decrease of the redox potential. It is the latter that results in differential reduction of certain soil constituents, leading to unique soil morphology that can be utilized in the identification of wetlands. These morphological features can persist for centuries in the soil, even after a wetland has been drained.

The Ramsar Convention (Ramsar, 1971) defines wetlands as "areas of marsh, fen, peatland or water, whether natural or artificial, permanent or temporary, with water that is static or flowing, fresh, brackish or salt, including areas of marine water the depth of which at low tide does not exceed six meters". It is, however, also important to note that the definition of wetland differs between, and often within, countries. These definitions may further vary in complexity from legislative ones to loose anecdotal concepts.

Ramsar convention

The Ramsar Convention on Wetlands (since it was signed in 1971 in the Iranian city of Ramsar) came into force in 1975 and is the only international agreement that is focused primarily on wetlands (Ramsar, 1971). To date, 170 countries have joined the Ramsar Convention as Contracting Parties, incorporating 2300 sites that cover about $250 \times 10^6 \text{ ha}$ (Ramsar Convention on Wetlands, 2018). Obligations of the Contracting Parties include: Conservation and wise use of all wetlands (Pillar 1), Designating and conserving at least one Wetland of International Importance, known as a Ramsar Site (Pillar 2), and Cooperation across national boundaries on transboundary wetlands, shared wetland systems, and shared species (Pillar 3).

The Ramsar Convention further provides a framework for wise use to ensure that wetlands are on the global agenda, to secure the sustainable development thereof, and that support initiatives for biodiversity increase, and to ensure the mitigation of climate change, disaster risk, and land degradation. These initiatives are also in accord with the United Nations' Sustainable Development Goals, the Aichi Biodiversity Targets, the Paris Agreement on Climate Change, and other related international commitments.

The ecological character of wetlands ("the combination of the ecosystem components, processes and benefits/services that characterize a wetland at a given point in time") is a further key Ramsar concept. Ramsar Site wetlands need to be actively maintained to preserve their ecological character and therefore the participating countries have to report any negative human impacts on such sites to the Ramsar Secretariat and must further take remedial action to restore them to their former state.

Classification

Various wetland classifications exist, but most of these divide wetlands based on their relationship to water bodies. The Ramsar Convention (Ramsar Convention on Wetlands, 2018) classifies 42 wetland types into three main categories, namely marine and

Table 1 The Ramsar Convention classification of wetland categories and types.

<i>Inland wetlands</i>	<i>Coastal/marine wetlands</i>	<i>Human-made wetlands</i>
River Stream	Salt Marsh	Reservoir
Lake	Mangrove	Rice Paddy
Peatland	Seagrass	Wet Grass
Marsh Swamp	Coral Reef	Waste Ponds
Underground	Shellfish Reef	Salinas
	Lagoon	Aqua Ponds
	Kelp	

Adapted from Ramsar Convention on Wetlands, 2018.

coastal wetlands, inland wetlands, and human-made wetlands (Table 1). These wetland categories are further subdivided into wetland types. Many of the wetland types, for example, the Marine, River Stream, and Lake Wetlands are primarily aquatic and the soil is therefore permanently covered with water of up to six meters deep. These permanently inundated soils therefore do not experience the alternating cycles of oxidation and reduction, typically associated with terrestrial wetlands. The terrestrial wetlands e.g., Marsh and Peatland Wetlands contain hydric soils that form in the interface between oxidizing and reducing environments and are discussed later.

The Cowardin classification (Cowardin et al., 1979) classifies wetlands into five environments, namely Marine, Riverine (along rivers and streams), Lacustrine (along lakes), Estuarine (along tidal coasts and rivers), and Palustrine (inland freshwater forested and non-forested) wetlands. The first three (Marine, Riverine, and Lacustrine) wetlands classes are predominantly aquatic and therefore would have little or no soil development. It is thus only the Estuarine and Palustrine wetlands that would contain wetland soils.

Estuarine wetlands

Estuarine Wetlands refer to inter-tidal salt and brackish low and high marsh zones, non-vegetated tidal flats, brackish waters of coastal rivers and coastal areas, and mangrove swamps, comprising about 80 different tree species. The estuarine wetlands are subdivided into the emergent wetlands and the shrub wetlands. The estuarine emergent wetlands are dominated by herbaceous grasses and are referred to as salt marsh, low marsh (intertidal marsh), high marsh, and brackish tidal marsh, normally found in protected coastal areas or behind barriers where water movement has slowed. Many were formed due to sea level rise. There is typically a strong zonation of plant species in salt marshes, depending on the frequency and duration of inundation. The estuarine shrub wetlands have woody vegetation such as marsh elder or high tide bush, typically adapted to a saline environment. Mangrove swamps are included in this category.

Palustrine wetlands

The majority of freshwater wetlands falls under the Palustrine Wetlands and includes marshes, bogs, swamps, and ponds, with water bodies <2 m deep also included. Palustrine emergent wetlands are dominated by herbaceous vegetation with different species of grasses, rushes, and sedges. They are also known as freshwater marshes, wet meadows, and fens. The marshes may be infrequently to permanently flooded with water. Palustrine wetlands are subdivided into the emergent, shrub, and forested wetlands. The palustrine emergent wetlands have a variety of habitats, including peatlands and freshwater marshes. The palustrine shrub wetlands are characterized by woody vegetation <6 m tall, such as bog laurel (*Kalmia polifolia* Wangegh), labrador tea (*Rhododendron groenlandicum*), cranberry (*Oxycoccus palustris*), and leatherleaf (*Chamaedaphne calyculata*), and with stunted trees such as black spruce (*Picea mariana*), larch (*Larix laricina*), and balsam fir (*Abies balsamea*). Palustrine forested wetlands have trees >6 m tall, with tree species such as red maple (*Acer rubrum*), pin oak (*Quercus palustris*), sweet gum (*Liquidambar styraciflua*), black spruce (*P. mariana*), and larch (*L. laricina*).

Ecosystem services

The value of wetlands to human well-being has traditionally been ignored. It is only recently that this value has been recognized and categorized as provisioning, regulating, cultural, and supporting services (Table 2).

Wetland provisioning ecosystem services (Table 2) consist of food, fish, water, fiber, and fuel production. The regulating services determine climate and hydrological regimes, and reduce pollution, and the risk of disaster. The services of cultural importance include spiritual and inspirational services, as well as aesthetic spaces for recreation and education. Supporting services consist of biodiversity, soil formation, nutrient cycling, and pollination.

The value of wetland ecosystem services far exceeds that of terrestrial ecosystems, with the total economic value of inland wetlands being almost five times higher than that of tropical forests, the next most valuable terrestrial habitat. Estimated economic losses from 1997 to 2011 due to loss of wetland ecosystem services were US\$7.2 trillion from tidal marshes and mangroves, US\$2.7

Table 2 Summary of the wetland ecosystem services.

<i>Provisioning services</i>	<i>Regulating services</i>	<i>Cultural services</i>	<i>Supporting services</i>
Food	Climate	Spiritual and inspirational	Biodiversity
Fresh water	Hydrological	Recreational	Soil formation
Fiber and fuel	Pollution control	Aesthetic	Nutrient cycling
Biochemical products	Erosion protection	Educational	Pollination
Genetic materials	Natural hazards		

trillion from swamps and floodplains, and US\$11.9 trillion from coral reefs (Costanza et al., 2014). These estimates were made by assessing the trade-offs required to achieve a common goal in the absence of wetlands (Farber et al., 2002; Costanza et al., 2014).

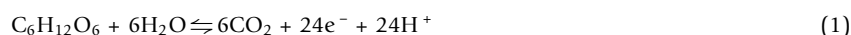
Chemistry

The major difference between terrestrial and wetland soils is that wetland soils experience alternating periods of water saturation, resulting in alternating aerobic and anaerobic conditions, while terrestrial soils typically experience primarily aerobic conditions (Terrestrial soils may experience short periods of anaerobic conditions, but these are normally too short to affect the soil morphology.). Wetland soils therefore have unique chemistry because they are in environments where oxygen concentrations fluctuate between aerobic and anaerobic conditions (Mitsch and Gosselink, 2015; Ponnampereuma, 1972; Richardson and Vepraskas, 2000).

In aerobic terrestrial soils, the aerobic soil microorganisms utilize oxygen as an electron acceptor in the oxidation of organic matter to extract the latent energy from their organic matter food source. Conversely, in wetland soils that lack oxygen due to water saturation, anaerobic and facultative anaerobic microorganisms can utilize other electron acceptors in the oxidation of organic matter to extract energy. These electron acceptors are utilized in a fixed sequence that ensures maximum energy extraction (Table 3). The result is that these electron acceptors are reduced in a sequence that releases increasingly less energy, with each reduction half reaction providing increasingly less energy to the microorganisms, since the reaction has a lower redox potential. This reduction sequence is therefore determined by the redox potential of each reduction half reaction and therefore governs the possible energy released from the oxidized organic matter. The presence of large quantities of a preceding electron acceptor might therefore prevent the reduction of an electron acceptor that would release less energy, buffering or poisoning the soil system at a specific redox potential. The reduction half reactions possible in soil are similarly limited by oxygen at the higher and by hydrogen at the lower redox potential; oxygen will not be limiting in an oxidized environment, while hydrogen will not be limiting in a water-saturated environment.

In soil reduction reactions, organic matter is always the electron donor (the entity being oxidized). The organic matter being oxidized in the oxidizing half reaction originates from that produced through photosynthesis which enters the soil where it is then humified through various processes. The sources of this organic matter are variable, but can include dead plant roots, plant debris (leaves, pieces of roots), dissolved organic carbon, and/or the excreta of organisms. Oxidation of organic matter in aerobic or anaerobic respiration therefore comprises the final step in the humification process.

The simplified oxidation half reaction:



Since the reduction reactions take place in a specific sequence, it follows that oxidation of the reduced species will take place in the reverse sequence, with the last species to be reduced being the first to be oxidized. In the oxidation of the reduced elements, oxygen (O_2 , present in aerated environments) is the oxidizing agent that is in turn being reduced. The elements being reduced and the sequences of these reduction and oxidation reactions are the main causes of the unique morphology observed in wetland soils that are addressed in the next section.

Table 3 The sequence of reducing (left to right) and oxidizing (right to left) half reactions possible in wetland soils.

<i>Element</i>	<i>Half reaction</i>	<i>Expected soil redox potential (Eh, mV)</i>
Oxygen	$\text{O}_2 + 4\text{H}^+ + 4\text{e}^- \rightleftharpoons 2\text{H}_2\text{O}$	600–400
Nitrogen	$2\text{NO}_3^- + 12\text{H}^+ + 10\text{e}^- \rightleftharpoons \text{N}_2 + 6\text{H}_2\text{O}$	500–200
Manganese	$\text{MnO}_2 + 4\text{H}^+ + 2\text{e}^- \rightleftharpoons \text{Mn}^{2+} + 2\text{H}_2\text{O}$	400–200
Iron	$\text{Fe}_2\text{O}_3 + 6\text{H}^+ + 2\text{e}^- \rightleftharpoons 2\text{Fe}^{2+} + 3\text{H}_2\text{O}$	300–100
Sulfate	$\text{SO}_4^{2-} + 10\text{H}^+ + 8\text{e}^- \rightleftharpoons \text{H}_2\text{S} + 4\text{H}_2\text{O}$	0 to –150
Carbon	$\text{CO}_2 + 8\text{H}^+ + 8\text{e}^- \rightleftharpoons \text{CH}_4 + 2\text{H}_2\text{O}$	–150 to –220
Hydrogen	$2\text{H}^+ + 2\text{e}^- \rightleftharpoons \text{H}_2$	–150 to –220

Adapted from Bohn et al., 2002.

Based on the discussion above, there are four prerequisites for reduction in a wetland soil, namely: anaerobic conditions (the absence of oxygen), oxidizable organic matter, anaerobic microorganisms, and suitable electron acceptors. These are all required for a wetland soil to reduce, and absence of any one of these will therefore prevent reduction. It is thus quite possible for soils to be saturated with water, but not to undergo reduction because one or more of these prerequisites is missing.

Reducing conditions

Reducing conditions can be calculated theoretically for wetland soils by considering the following equation:

$$m \text{ (oxidant)} + m \text{ H}^+ + n \text{ e}^- = m \text{ (reductant)} \quad (2)$$

And through the Nernst equation:

$$Eh = E^\circ - \left(\frac{0.059}{n} \right) \log \left(\frac{\text{Reductant}}{\text{Oxidant}} \right) - 0.059 \left(\frac{m}{n} \right) \text{ pH}. \quad (3)$$

where: Eh = Electrode potential (Volt); E° = Standard electrode potential (Volt); m = Number of protons involved in the reaction; n = Number of electrons involved in the reaction.

Eq. (3) can essentially be reduced to:

$$\text{rH} = \frac{Eh}{29} + 2 \text{ pH}, \quad (4)$$

where, rH = Negative logarithm of hydrogen partial pressure; Eh = Electrode potential (Volt).

In soils, the rH would range from 20.78 in extremely oxidizing conditions to 0 in strongly reducing conditions. When the $\text{rH} < 20$, there is almost no molecular oxygen present in the soil environment and it is thus interpreted as signifying a reducing environment.

The presence and stability of the various oxidized and reduced species possible in wetland soils are given by pH/ Eh stability diagrams (Bohn et al., 2002), but is not elaborated on here.

Morphology

Iron and manganese ions are present in the reduced form as silicate minerals in rock and as such are fairly colorless. With chemical weathering, these iron and manganese ions are segregated, oxidized and impart a bright brown, yellow, red and or black colors to the soil. The iron and manganese minerals that form through these chemical weathering processes are immobile in oxidized soils. The iron minerals are predominantly hematite and goethite, with pyrolusite as the dominant manganese mineral. In wetland soils, the oxidized iron and manganese ions are reduced and become soluble and thus mobile, and largely lose their color. These properties of iron and manganese, i.e., brightly colored and immobile in the oxidized state and colorless and mobile in the reduced state, give rise to the unique color patterns that are observed in wetland soils.

Reduction, segregation, and relocation of iron and manganese in wetland soils give rise to redoximorphic features in the form of redox concentrations, redox depletions, and or reduced matrices (Richardson and Vepraskas, 2000; Schoeneberger et al., 2012). Redox concentrations (also known as an oximorphic color pattern) refer to brightly colored (Munsell chroma > 2) accumulations of oxidized iron and manganese to form masses (also known as mottles), pore linings, and or nodules and concretions. Redox depletions (also known as a reductimorphic color pattern) refer to zones where iron and manganese have been reduced and removed forming reduced (grey color with Munsell chroma ≤ 2) pore linings, depleted (grey color with Munsell chroma ≤ 2) and or reduced (with bluish or greenish colors, also known as grey colors) soil matrices. Soil matrix refers to the solid phase constituents of the soil (Fig. 1).

The color patterns that develop through the processes described above can be related to the soil water regime (Vasilas et al., 2010; Vepraskas, 2004). These different color patterns have found application in the WRB stagnic color pattern (Fig. 2) and gleyic color pattern (Fig. 3), and in the USDA definition of hydric soils. A stagnic color pattern is defined as one where the macro pores have a reductimorphic (grey) color, while the soil matrix has an oximorphic (bright) color. The converse is true for a gleyic color pattern: the macro pores have an oximorphic (bright) color, while the matrix has a reductimorphic (grey) color. Stagnic color patterns are associated with shorter periods of water saturation, where water primarily enters the soil from the soil surface along macro pores and cracks, while gleyic color patterns are associated with long periods of water saturation, where the water primarily enters the soil through ground water saturation (from the bottom upwards).

Oximorphic mottles are bright and vary from black to red to brown and yellow-orange that form when reduced iron and or manganese encounter oxygenated zones, are oxidized and then precipitate as iron and or manganese oxides. These precipitates typically coat soil structural units, pores, or root channels where oxygen can readily enter the soil. Reductimorphic mottles are grey (with a Munsell color chroma ≤ 2) that develop where iron and or manganese oxides are reduced and removed to reveal the grey underlying silicate minerals.

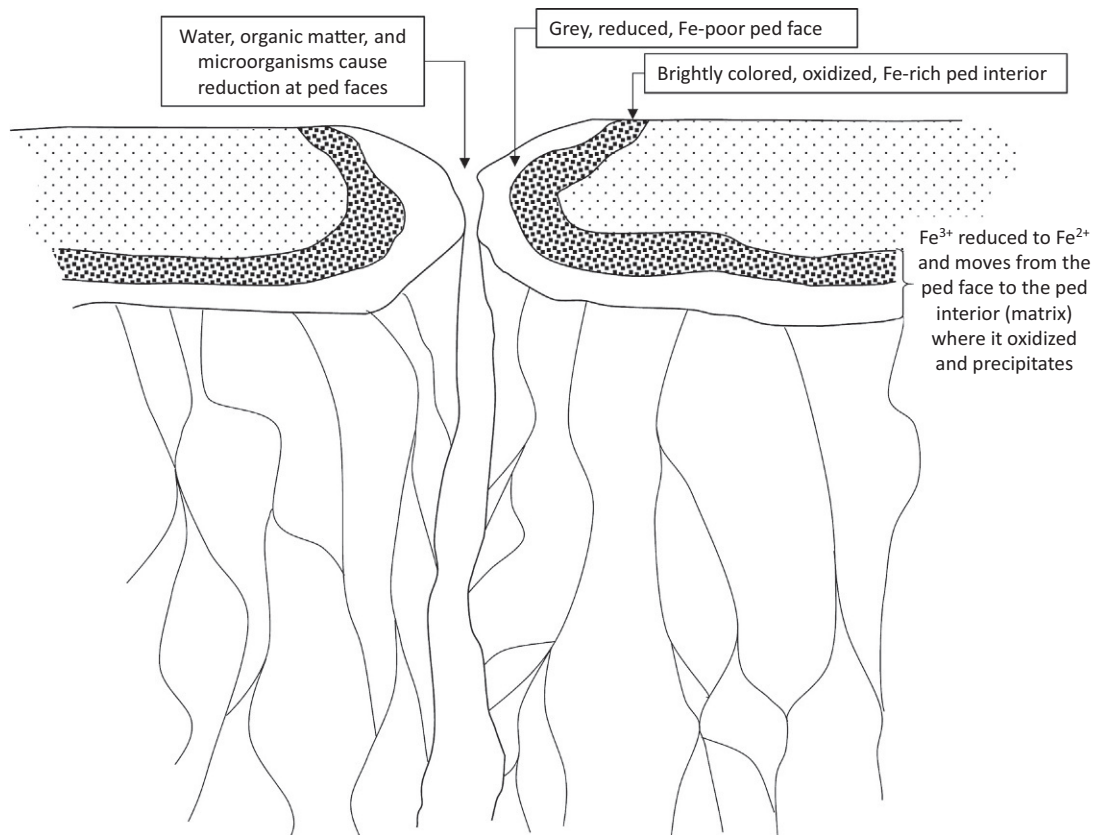


Fig. 1 Formation of a stagnic color pattern through water saturation in the macro (inter-ped) pores, where oxidized iron (Fe^{3+}) is reduced to Fe^{2+} , forming grey reductomorphic colors and is then relocated (as Fe^{2+}) into the oxidized matrix where it precipitates as Fe^{3+} oxides, forming bright, oximorphic colors.

The extent of gleying and mottle abundance, size, and color depends on the extent of reduction, which is governed by the hydrology and microbial activity. The microbial activity is, in turn, governed by the presence of readily oxidizable organic matter, microorganism species, and temperature (there is little microbial activity if the soil is too cold). Redoximorphic features can remain visible even after wetlands have been drained and thus offer a useful tool in determining the original wetland extent. With increasing wetness, organic matter accumulation, gley colors (chroma <1), as well as nodule and concretion occurrences are expected to increase (Vepraskas and Craft, 2016).

Exceptions can occur when red mottles are masked in soils with a much redder matrix that have formed, for example, from red glacial till parent material. Similarly, it is almost impossible for grey mottles to develop in a grey soil that formed from grey quartzitic parent material. In these cases soils will not exhibit an oximorphic or reductomorphic color despite experiencing water saturation and anaerobic conditions. Other wetland indicators such as measurements of water levels and or the presence of hydrophytic vegetation should then be used in the identification of wetland soils. Redoximorphic features can remain visible even after wetlands have been drained and thus offer a useful tool in determining the original wetland extent.

Soils that experience extreme variations in redox potential, through changes in water saturation, tend to have a lower clay content. This typically occurs along the zones that experience the longest periods of water saturation, such as macropores, where the clay minerals are degraded from 2:1 to 1:1 clays to quartz. Brinkman (1970) described this process that leads to a texture gradient due to water saturation, as ferrollysis (Fig. 3). Ferrollysis is thought to occur when the ferrous iron (Fe^{2+}) concentration in the soil solution increases dramatically when the soil is subjected to reduction. Ferrous iron then replaces the basic cations on the exchange complex of clays, and may or may not be removed from the soil. With oxidation, the ferrous iron is oxidized and removed from the soluble phase and thus also from the exchange complex. Oxidation of ferrous iron also leads to an increase of hydrogen (H^+), and a concomitant decrease in pH. If the pH decreases far enough, the clay minerals will disintegrate to release basic cations into the soil solution. Repeated oxidation–reduction cycles are required for ferrollysis to occur and would lead to a reduced pH and lowering of the overall clay content through clay destruction.

It is quite possible that the observed color patterns are not in phase with the current soil water regime because they might have developed under a previous period of water induced reduction, for example in drained wetlands. The following arguments can, however, be used to prove or disprove the contemporary nature of the observed soil morphology:

Gradient: First, one would anticipate a wetness and thus morphological gradient in topography. Indicators of water saturation are assumed to increase down slope. Second, a gradient in wetness and morphology is expected from the soil's surface to deeper in the profile. Last, a gradient should be observable within the peds as a logical sequence of the juxtaposition of colors (e.g., black—red—yellow—grey), as depicted for example in Figs. 1 and 2.

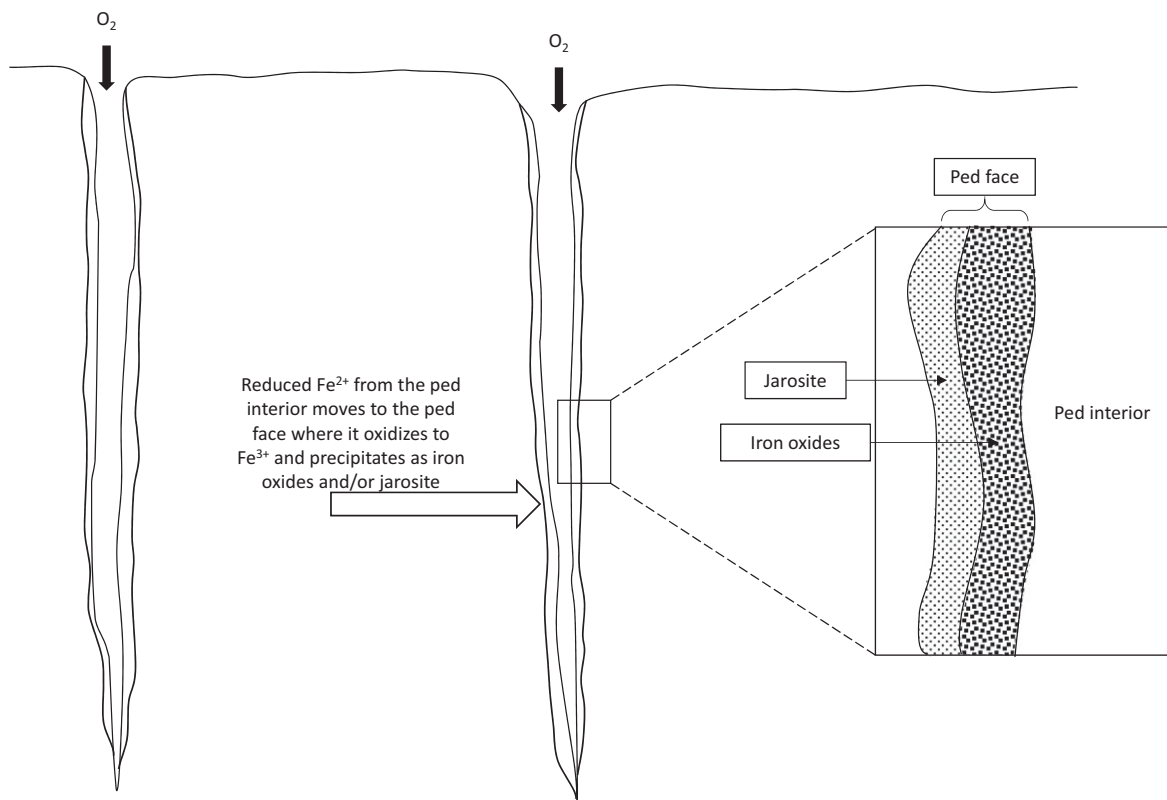


Fig. 2 Formation of a gleyic color pattern through the precipitation of reduced iron (Fe^{2+}) on the oxidized macro (inter-ped) pores as iron (Fe^{3+}) oxides, forming bright, oximorphic colors. The Fe^{2+} is formed through reduction in the reduced matrix, from where it moves under a concentration gradient to the oxidized ped face.

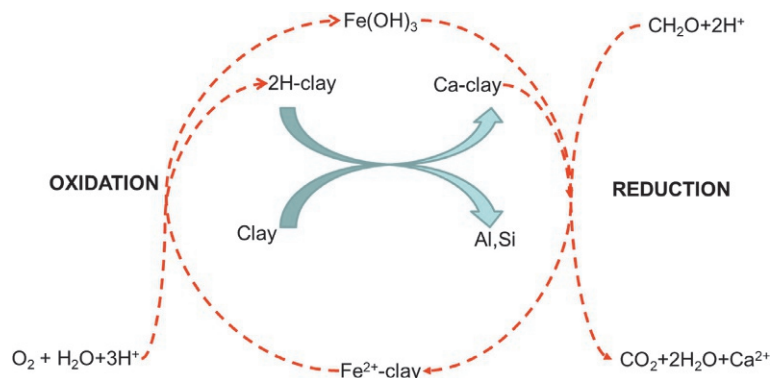


Fig. 3 Schematic representation of the ferrollysis process occurring in soils subjected to alternating oxidized and reduced conditions. Adapted from Brinkman, 1970.

Mottles: Actively forming mottles typically have diffuse boundaries, while relict mottles have more abrupt ones. This is because the former has an accumulation of iron and manganese that does not always move to the same area, while degradation is more probable in the latter, removing iron and manganese from the outside of the mottle inwards. It is also assumed that mottles would have greater accumulations of iron and manganese at the bottom of the mottle, rather than at the top because water saturation is expected to occur from the bottom upwards in the soil. In addition, mottle abundance is expected to increase with depth where water is actively saturating soils.

Juxtaposition: The sequencing of colors (juxtaposition) expected in actively saturated soils are normally black—red—yellow—grey, either from the inside out or from the outside in along a ped face (Figs. 1 and 2), reflecting the dominance of pyrolusite, hematite, goethite, and quartz respectively, as governed by the oxidation and reduction sequence (Table 3).

Nodules and concretions: Although nodules and concretions can actively form in contemporary saturated soils, they are not considered indicators of wetland hydrology because they are indurated and tend to persist within the soil and in the landscape. It is therefore possible for nodules and concretions to exist in both contemporary and relict saturated soils. [Nodules and concretions are

both hard, cemented concentrations of iron or manganese oxides. Concretions have an internal symmetry organized around a point, line, or plane, while nodules have no orderly internal organization.]

Conclusion: All the indicators discussed above are not necessarily present in all actively saturated soils. However, the more that these indicators occur, the more likely it would be that a soil experiences active water saturation and reduction and that the observed soil morphology is therefore not relict.

Hydric soils

The United States Army Corps of Engineers' Wetlands Delineation Manual (Environmental Laboratory, 1987) define wetlands as "Those areas that are inundated or saturated by surface or ground water at a frequency and duration sufficient to support, and that under normal circumstances do support, a prevalence of vegetation typically adapted for life in saturated soil conditions. Wetlands generally include swamps, marshes, bogs, and similar areas." Wetlands are therefore identified by the presence of hydrophytic vegetation, hydric soils, and wetland hydrology. At least one positive indicator from each of these parameters must be present to warrant the identification of a wetland. Hydrophytic vegetation is defined as vegetative species that are adapted to growth in saturated soil conditions. Hydric soils are those that have reducing soil conditions. Wetland hydrology requires that the area is permanently or periodically inundated with mean water depths <2.0 m, or that the soil is saturated with water to the surface for some time during the prevalent vegetation's growing season.

The USDA-NRCS [United States Department of Agriculture, Natural Resources Conservation Service] (2018) defines hydric soils (Table 4) as soils that are "saturated, flooded, or ponded long enough during the growing season to develop anaerobic conditions in the upper part of the soil." For identification, the hydric soils are divided into organic and mineral soils. Organic hydric soils are characterized by a thick organic layer that has accumulated over time because the microbial oxidation of plant residue is severely hindered under saturated conditions. The organic matter content in organic soils has to be greater than 20% and the soil must be at least 0.6 m to more than 9 m deep. Organic soils typically develop in areas where anaerobic conditions prevail (e.g., in depressions, bogs or marshes). Based on the nature of the accumulated organic matter, organic soils can be further divided into mucks (saprist or WRB Sapric qualifier), peats (fibrist or WRB Fibric qualifier), and hemists (mucky peat, peaty muck or WRB Hemic qualifier) as follows: Rub the wet soil material gently between the forefinger and thumb. If the material feels gritty after the first or second rub, then it is mineral soil material. If the material feels greasy after the second rub, then it is either mucky mineral or organic soil material. Rub the soil material gently two or three more times. If it now feels gritty or plastic, then it is mucky mineral soil material, and if it still feels greasy, then it is organic soil material. With the large organic matter content of organic soils, the bulk density is much lower ($0.9\text{--}1.2\text{ Mg m}^{-3}$) and they also have a much larger water holding capacity than mineral soils.

Mineral hydric soils have less than 20% organic material (based on the dry weight) and have visual redoximorphic features. These redoximorphic features can vary from very dark grey to very light (low chroma ≤ 2) matrix colors, with or without mottles. Mottles can either be zones of iron and or manganese oxide accumulation (oximorphic color pattern) or zones of iron and or manganese oxide depletion (reductimorphic color pattern). Mineral hydric soils form where moving or stagnating water accumulates, permanently or periodically, and are therefore typically found in low-lying positions, depressions, or on slopes where groundwater seeps out (Fig. 1). These soils can receive water through surface runoff or from groundwater seepage, and may have a subsoil layer that restricts water percolation (e.g., fragipan, clay pan, hardpan or bedrock) (Fig. 4).

Table 4 Summarized criteria for hydric soils (USDA-NRCS [United States Department of Agriculture, Natural Resources Conservation Service], 2018, [with approximate IUSS Working Group WRB (2015) equivalent classification in brackets]).

All Histels [Cryic Histosols] except Folistels [Folic Histosols]; or all Histosols [Histosols] except Folists [Folic Histosols]; or Soils in Aquic suborders [Stagnic or Gleyic qualifiers], great groups [Stagnosols & Gleysols], or subgroups, Albolls suborder [Stagnic or Gleyic Chernozems, Kastanozems, Phaeozems or Umbrisols], Historthels great group [Histic Cryosols], Histoturbels great group [Turbic Histic Cryosols], or Andic [Andic qualifier], Cumulic [cf Hyperhumic qualifier], Pachic [Pachic qualifier], or Vitrandic [Vitric qualifier] subgroups that:
a. Based on the range of characteristics for the soil series, will at least in part meet one or more field Indicators of Hydric Soils, or
b. Show evidence that the soil meets the definition of a hydric soil; or
Soils that are frequently ^a ponded for long durations ^b or very long durations ^c during the growing season ^d that:
a. Based on the range of characteristics for the soil series, will at least in part meet one or more field Indicators of Hydric Soils, or
b. Show evidence that the soil meets the definition of a hydric soil; or
Map unit components that are frequently ^a flooded for long durations ^b or very long durations ^c during the growing season ^d that:
a. Based on the range of characteristics for the soil series, will at least in part meet one or more field Indicators of Hydric Soils, or
b. Show evidence that the soils meet the definition of a hydric soil.

From: https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/use/hydric/?cid=nrcs142p2_053959

^aFrequently is flooding, ponding, or saturation that is likely to occur often under usual weather conditions (>50% chance in any year, or more than 50 times in 100 years).

^bLong duration is a single event lasting from 7 to 30 days.

^cVery long duration is a single event lasting longer than 30 days.

^dGrowing season is the portion of the year when soil temperatures are above biologic zero (5 °C) at 50 cm.

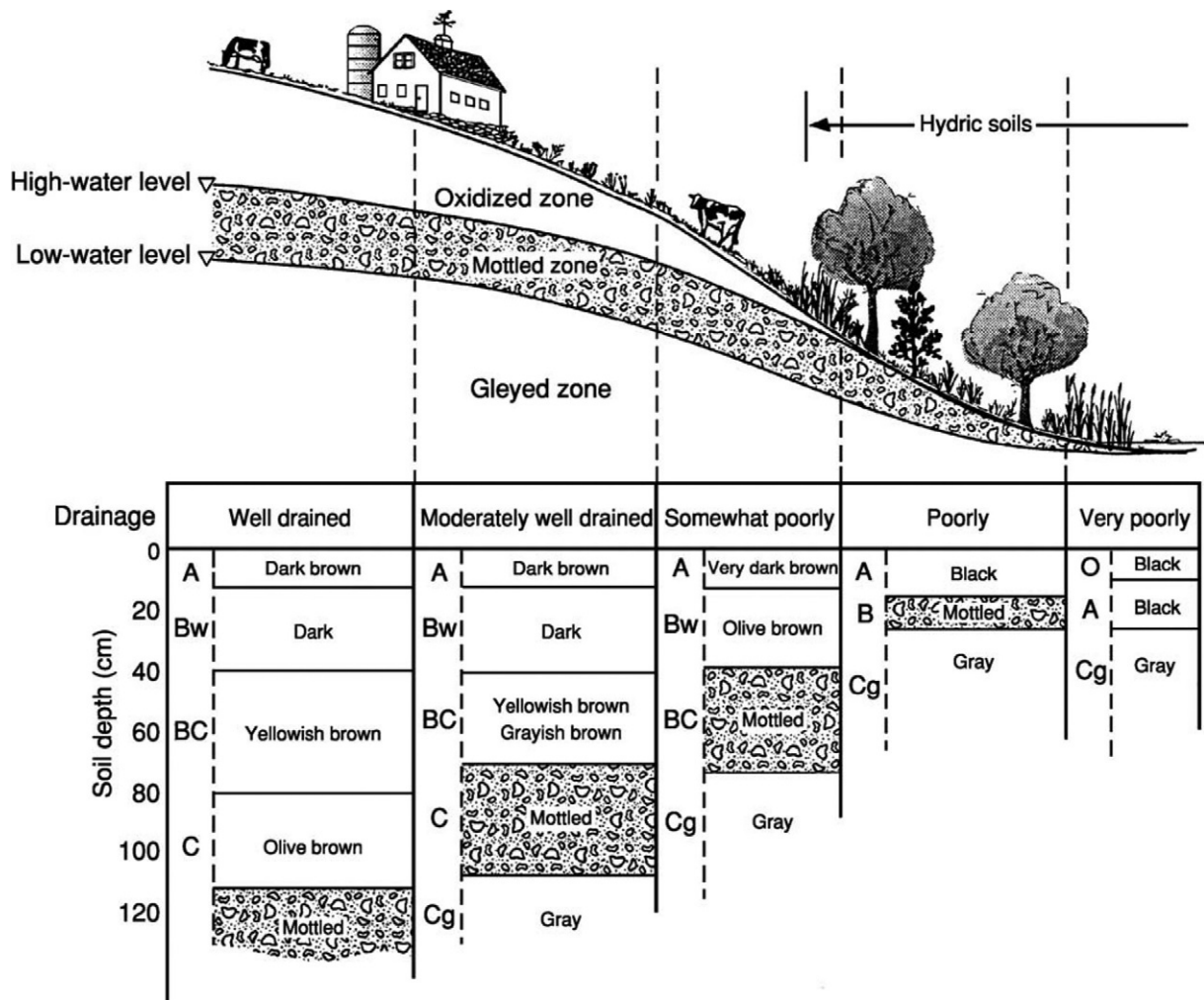


Fig. 4 Soil properties along a wetland-to-upland gradient. Note that the seasonal high-water table, indicated by the mottled zone, occurs at increasingly shallower depths and only reaches the surface at the lower elevations. Adapted from [Tiner and Burke \(1995\)](#).

Measurement

The prevalent conditions in the root environment of wetland soils can be described qualitatively as ‘anaerobic’, ‘reduced’, ‘flooded’, ‘saturated’, or ‘waterlogged’. Wetlands are also defined rather qualitatively based on the presence of hydrophytic vegetation, hydric soils, and wetland hydrology. However, measurements of water table position, redox potential, pH, and reduced iron (Fe^{2+}) need to be made to define and characterize wetlands quantitatively.

Hydrology

Water content measurements are probably the most direct approach to ascertain soil water saturation. This can be done through water level measurements in piezometers or in wells or by measuring the soil water content with a neutron water meter, tensiometer, time domain reflectometry (TDR) or other similar technology. Water level measurements in piezometers or wells are typically expressed as hydroperiod, seasonal high water level, or the duration of water saturation within a specific horizon. Hydroperiod is, however, more typically modelled from precipitation, temperature, and evapotranspiration. In the case of soil water content measurements, these need to be converted to degree of saturation (water volume as a fraction of the pore volume that is in turn calculated from the bulk density) in order to estimate water saturation in the soil or at a specific depth. Irrespective of the method used to determine soil water saturation, it should be longer than 7 days during the growing season—i.e., when soil temperatures $>5^\circ\text{C}$ at 50 cm. It has also been noted that soils can be reduced if the water saturation reaches 70% of porosity ([Van Huyssteen et al., 2004](#)).

Redox potential and pH

It is quite difficult to measure oxygen concentration and the rate of oxygen diffusion in wetland soils. Measuring the soil Eh is, however, a viable alternative in the laboratory and in the field. Soil Eh varies from -300 to $+700$ mV in flooded soils and is $< +350$ mV in permanently waterlogged soils, compared to $> +400$ mV in aerated soils. The Eh must, however, be measured in conjunction with pH to characterize fully the wetland biogeochemical environment (see the discussion on reducing conditions).

Redox potential measurements are done by installing platinum electrodes to the desired soil depths. These electrodes are then linked to a volt meter and a standard calomel reference electrode. The pH measurements are done similarly, but with a pH electrode instead of the platinum one. Soils are considered reduced when the rH < 20 .

In situ redox potential and pH measurements, however, suffer from quite a few limitations. First, the redox couples are not necessarily in equilibrium with their soil environment, except in highly reduced soils because the soil system is very heterogeneous and considerable differences can occur over short distances (less than millimeter scale). Second, respiration is actively taking place in all soil environments and electrons are therefore being added and removed continuously, and thus equilibrium is seldom, if ever, reached. Third, redox potential measurements cannot be interpreted on their own because they are interrelated with the pH, necessitating both Eh and pH measurements. Last, the surface of the platinum electrode can be contaminated through the precipitation of oxide coatings, sulfides and/or other impurities, especially in in situ installations.

Indicators

As discussed above, the direct measurement of reduction in soils is time consuming, tedious, and suffers from equipment issues. Furthermore, characterization of redox potential must be done by measuring Eh and pH, which are then evaluated against Eh-pH stability diagrams. Taking only a single measurement would also not reflect the annual or inter-annual variation. Various indicators can, however, be used to determine the presence of reduced elements in the soil.

Indicators, such as α , α' dipyridyl, can be used to test for the presence of reduced ions (Childs, 1981), such as Fe^{2+} in this case. These indicators are quick and easy to use, but might be difficult to obtain, are hazardous to use, and also only provide a snapshot in time.

Indicator of Reduction In Soil (IRIS) tubes (Castenson and Rabenhorst, 2006) are PVC tubes covered with synthetic ferrihydrite, goethite and hematite, installed vertically into the upper 50 cm of soil for a period of 2–4 weeks. In a reduced soil more than 30% of the ferrihydrite paint is removed from the IRIS tube. These tubes therefore offer an innovative, cheap and easy method to determine temporal soil reduction. See for example Rabenhorst (2012), Mutiti et al. (2018) and LeFevre et al. (2021).

Conclusions

Wetland soils are unique environments that occur at the interface between aquatic and terrestrial ecosystems. Traditionally, humans have a rather pessimistic relationship with wetlands, as depicted in terms such as 'bogged down', 'swamped' and 'quagmire'. It is therefore no surprise that wetlands have been drained and filled in to 'improve' the land. The net effect has been that the hidden ecosystem services of wetlands such as provisioning, regulating, cultural, and supporting services, were overlooked. Fortunately, this trend has recently been reversed and wetlands now occupy an increasingly important position on local, national and global agendas.

Wetlands are predominantly classified based on their relationship to water bodies. It is, however, only the terrestrial wetlands such as marsh and peatland or the estuarine and palustrine wetlands where the soils occur at the interface between oxidizing (aerobic) and reducing (anaerobic) environments. It is these alternating conditions that cause the unique soil morphology that is used for wetland identification and delineation.

In aerobic soil environments the microorganisms utilize oxygen as electron acceptors in the oxidation of organic matter to obtain energy. However, in anaerobic environments the microorganisms must utilize other electron acceptors such as nitrate, manganese, and iron that result in them obtaining increasingly less energy. These reactions cause a reducing environment (lowering of the electrode potential), which causes the manganese and iron coloring agents to be reduced, become soluble, and lead to the development of different oximorphic and reductimorphic color patterns. It is these color patterns that have been used to define stagnant and gleyic color patterns and hydric soils. These redox color patterns find application in various soil classification systems and in wetland identification and delineation.

References

- Bohn HL, Myer RA, and O'Connor GA (2002) *Soil Chemistry*. New York: John Wiley & Sons.
- Brinkman R (1970) Ferrololysis, a hydromorphic soil forming process. *Geoderma* 3: 199–206.
- Castenson KL and Rabenhorst MC (2006) Indicator of reduction in soil (IRIS) Evaluation of a new approach for assessing reduced conditions in soil. *Soil Science Society of America Journal* 70: 1222–1226.
- Childs CW (1981) Field tests for ferrous iron and ferric-organic complexes (on exchange sites or in water-soluble forms) in soils. *Soil Research* 19: 175–180. <https://doi.org/10.1071/SR9810175>.

- Costanza R, de Groot R, Sutton P, Van der Ploeg S, Anderson SJ, et al. (2014) Changes in the global value of ecosystem services. *Global Environmental Change* 26: 152–158.
- Cowardin LM, Carter V, Golet FC, and LaRoe ET (1979) *Classification of Wetlands and Deepwater Habitats of the United States*. Fish and Wildlife Service, FWS/OBS-79/31. Washington, DC: US Government Printing Office.
- Environmental Laboratory (1987) *Corps of Engineers Wetlands Delineation Manual, Technical Report Y-87-1*. Vicksburg, MI: U.S. Army Engineer Waterways Experiment Station.
- Farber SC, Costanza R, and Wilson MA (2002) Economic and ecological concepts for valuing ecosystem services. *Ecological Economics* 41: 375–392.
- IUSS Working Group WRB (2015) International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports No. 106 In: *World Reference Base for Soil Resources 2014, Update 2015*. Rome: FAO.
- LeFevre OV, Knappenberger T, Shaw JN, and Olshansky Y (2021) Camera illustration of Indicator of Reduction in Soils (IRIS) reduction dynamics. *Agricultural & Environmental Letters* 6: e20051.
- Mitsch WJ and Gosselink JG (2015) *Wetlands*. John Wiley & Sons.
- Mutiti C, Mutiti S, and VandeVoort A (2018) Developing methods for quantifying IRIS tube data. *Georgia Journal of Science* 76: 60.
- Ponnamperuma FN (1972) The chemistry of submerged soils. *Advances in Agronomy*. vol. 24, pp. 29–96. Academic Press.
- Rabenhorst MC (2012) Simple and reliable approach for quantifying IRIS tube data. *Soil Science Society of America Journal* 76: 307–308.
- Ramsar (1971) *Convention on Wetlands of International Importance, Especially as Waterfowl Habitat*. Iran: Ramsar.
- Ramsar Convention on Wetlands (2018) *Global Wetland Outlook: State of the World's Wetlands and Their Services to People*. Gland: Ramsar Convention Secretariat.
- Richardson JL and Vepraskas MJ (2000) *Wetland Soils: Their Genesis, Morphology, Hydrology, Landscapes and Classification*. Boca Raton, FL: CRC Press.
- Schoeneberger PJ, Wysocki DA, and Benham EC (2012) *Field Book for Describing and Sampling Soils*. Government Printing Office.
- Tiner RW and Burke DG (1995) *Wetlands of Maryland. US Fish and Wildlife Service, Hadley/Maryland Department of Natural Resources*. Annapolis, MD: US Fish and Wildlife Service.
- USDA-NRCS [United States Department of Agriculture, Natural Resources Conservation Service] (2018) In: Vasilas LM, Hurt GW, and Berkowitz JF (eds.) *Field Indicators of Hydric Soils in the United States, Version 8.2*. USDA, NRCS. Cooperation With the National Technical Committee for Hydric Soils.
- Van Huyssteen CW, Hensley M, and Le Roux PAL (2004) The relationship between soil morphology and soil water regime: Preliminary results in the Weatherley catchment. *South African Journal of Plant and Soil* 21: 240–244.
- Vasilas LM, Hurt GW, and Noble CV (2010) *Field Indicators of Hydric Soils in the United States*. Washington, DC: United States Department of Agriculture, Natural Resources Conservation Service. Cooperation With the National Technical Committee for Hydric Soils.
- Vepraskas MJ (2004) *Redoximorphic Features for Identifying Aquic Conditions*. Technical Bulletin 301 Raleigh, North Carolina: North Carolina State University.
- Vepraskas MJ and Craft CB (2016) *Wetland Soils: Genesis, Hydrology, Landscapes, and Classification*. CRC Press.

Paddy soils: Characteristics, productivity and management

Stephan M Haefele, Rothamsted Research, Harpenden, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

This is an update of C Witt, SM Haefele, Paddy Soils, Editor(s): D Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 141–150, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00286-1>

Introduction	447
Cropping environments in Asia's lowlands	447
Distribution and important characteristics of paddy soils	448
Paddy soil fertility and nutrient management	450
Lowland rice systems and declining water availability	454
Sustainability and environmental issues	454
Conclusions	456
References	456

Abstract

Rice is a very important crop for human nutrition globally, and most rice soils and their processes are exceptional because they are flooded for a considerable part of the year. The rice production system underwent big changes following the Green Revolution, with much greater irrigated areas, more crops grown per year, and the increasing use of agro-chemicals to achieve much higher yields. This article describes the major cropping environments, outlines the distribution and main characteristics of “paddy soils”, presents the basics and advances of soil fertility and nutrient management, and discusses important sustainability problems of rice soils and cropping systems.

Key points

- Rice soils cover more than 160×10^6 ha and are mostly located in the tropics and sub-tropics around the world.
- Their particular characteristics are due to flooding for part of the year, both in irrigated and rainfed cropping systems.
- Flooding of paddy soils has marked effects on electrochemical, chemical and microbial processes, which affect nutrient availability and soil organic matter decomposition.
- The quality of rice soils can be maintained even where rice is continuously grown in monocropping at high production levels.
- Problems of rice soils and systems include water shortages, excessive N application and losses to the environment, and considerable methane emissions.

Introduction

Wetland rice or paddy soils provide the staple diet for nearly half the world's population. Worldwide, the total harvested rice area covers about 164 million ha, and about 85% of the rice area is situated in Asia, with a similar contribution to rice production. The flooded soil environment is one of the specific characteristics associated with rice cultivation in the traditional rice-growing areas in the river deltas of China, Southeast Asia and the Indian subcontinent (Moormann and van Breemen, 1978). The term “paddy soil” does not describe a specific soil type but rather denotes soils in lowland rice production systems having a prolonged period of submergence. Water regime is used to classify the environments of paddy soils into broad ecosystems, distinguishing irrigated and rainfed rice grown predominantly in the lowlands, upland rice grown in the uplands under aerated soil conditions, and deep-water rice in flood-prone river deltas and coastal areas. Paddy soils need to be discussed in the context of their production systems, considering trends, opportunities and challenges, and here we focus on lowland rice production systems of Asia, because of their prominent role in global food security, and in forcing global climate change by the production and release of greenhouse gasses.

Cropping environments in Asia's lowlands

With an increasing demand for rice from a growing population, rice cropping and paddy soil management in Asia underwent substantial intensification since the first modern high-yielding varieties with shorter growth periods, reduced height and better

fertilizer response were released in the early 1960s. Rice production tripled from 1948 to 1990 because of a larger physical land area cropped to rice, an increased number of crops grown per year, and greater yields as a result of Green Revolution technologies, such as improved germplasm, irrigation infrastructure, water and pest management, and increased use of inorganic fertilizer (GRiSP (Global Rice Science Partnership), 2013). The intensification was accompanied by changes in chemical, biological and physical processes in the soil-floodwater system. The duration of soil submergence and tillage operations increased, especially in the expanding double- and triple-rice cropping systems. In such systems, the periods of aerobic soil conditions are shorter and mostly restricted to dry fallows. The duration of the dry fallow period between two rice crops varies from a few weeks up to two or three months in the typical annual double-rice cropping system in Asia.

In the mid-1990s, the harvested area in Asian rice ecosystems included 73×10^6 ha of irrigated rice, 46×10^6 ha of rainfed lowland production, 9×10^6 ha of upland rice, and around 4×10^6 ha of flood-prone rice ecosystems (Huke and Huke, 1997). Although the overall physical rice area was expected to change only little in the future, the forecast (Rosegrant et al., 1995) was that total rice production in Asia would have to increase by about 25% from 530×10^6 t in 1999 to 662×10^6 t in 2020 to meet the increasing rice demand in Asia (Table 1). This forecast was generally met, with total paddy production in Asia reaching 676×10^6 t in 2020. However, contrary to the forecast, this goal was achieved by a considerable increase of the harvested area of rice to 141×10^6 ha whereas the average rice yield remained below the target of 4.9 t ha^{-1} (Table 1), only reaching 4.8 t ha^{-1} in 2020. The ongoing changes in rice demand, environment and socio-economic conditions, and their effects on productivity and management of paddy soils in Asia's intensive rice systems are depicted in Fig. 1. Most of further production increase will have to come from grain yield increases in the lowland rice ecosystems, and technologies developed in recent years provide substantial opportunities for farmers to increase productivity and profitability as rice demand continues to increase. With sufficient investment, rice production can undergo an intensification that is environmentally sound and sustains the soil resource base. However, in some areas, crop management may need to change in response to environmental concerns, resource availability or low profitability, with possible adverse effects on rice yields and soils.

Distribution and important characteristics of paddy soils

Paddy soils constitute an important group within the wetland soils where water saturation dominates soil development and the types of plant and animal communities (IRRI, 1978). Thus, paddy soils are defined as wetland soils with standing water in banded (Plate 1) and leveled fields (Plate 2) used for cultivation of rice (bunds are small dikes around the field to keep the water from running off). Paddy soils may support an upland, non-rice crop in the dry season when soils are aerated, but insufficient drainage capacities often make rice the only possible crop choice in the lowlands of the humid tropics. Free surface water may occur naturally or rainfall, runoff or irrigation water may be retained by field bunds or compacted subsoil layers.

According to the FAO World Reference Base for Soil Resources (IUSS Working Group, 2015), most paddy soils are designated as Anthrosols (having a puddled surface layer and a plow pan) or as Gleysols, Fluvisols, Planosols, Plinthosols and Histosols. Smaller areas of paddy soils fall within the gleyic soil units of Arenosols, Andosols, Cambisols, Solonetz, Solonchaks, Luvisols, Lixisols, Acrisols and Alisols. Although Vertisols, Nitisols and Ferralsols have no gleyic soil units, these soils may be artificially flooded and used for rice cultivation. The U.S. Department of Agriculture Soil Taxonomy does not recognize wetland soils at the level of soil orders, but classifies soils with aquic conditions at the suborder level and soils with hydromorphism at the subgroup level (USDA, 1999). Most paddy soils are therefore assigned to the aquic suborders of Andosols, Oxisols, Vertisols, Ultisols, Mollisols, Alfisols, Inceptisols and Entisols. A global assessment of rice soil quality and constraints was conducted by Haefele et al. (2014).

Natural or artificial flooding of paddy soils has marked effects on electrochemical, chemical and microbial processes, which in turn affect soil fertility and thereby crop growth in a dynamic manner. The main electrochemical changes include a decrease in redox potential and changes in soil and solution pH, microbially mediated processes which are mainly controlled by organic matter and reducible Fe contents of the soil (Kirk, 2004). With decreasing redox potential, oxidized forms of soil redox systems serve as electron

Table 1 Yield, area and production of irrigated and non-irrigated rice for Asia in 1999, and required changes in yield based on projected changes in area and production requirements until 2020 (Dawe, IRRI, unpublished).

System	Grain yield			Area			Production		
	(t ha^{-1})		$\Delta\%$	(10^6 ha)		$\Delta\%$	(10^6 t)		$\Delta\%$
	1999	2020		1999	2020		1999	2020	
Irrigated	5.36	6.72	+25%	76.0	78.4	+3.2%	406	527	+30%
Non-irrigated	2.08	2.39	+14%	59.5	56.5	-5.0%	124	135	+9%
Total	3.92	4.90	+25%	135.2	135.2	0.0%	530	662	+25%

Projected changes in area and production are based on Rosegrant MW, Sombilla MA, Perez N (1995) *Global Food Projections to 2020: Implications for Investment. Food, Agriculture and the Environment*. Discussion Paper No. 5. International Food Policy Research Institute: Washington, DC. pp. 1–54.

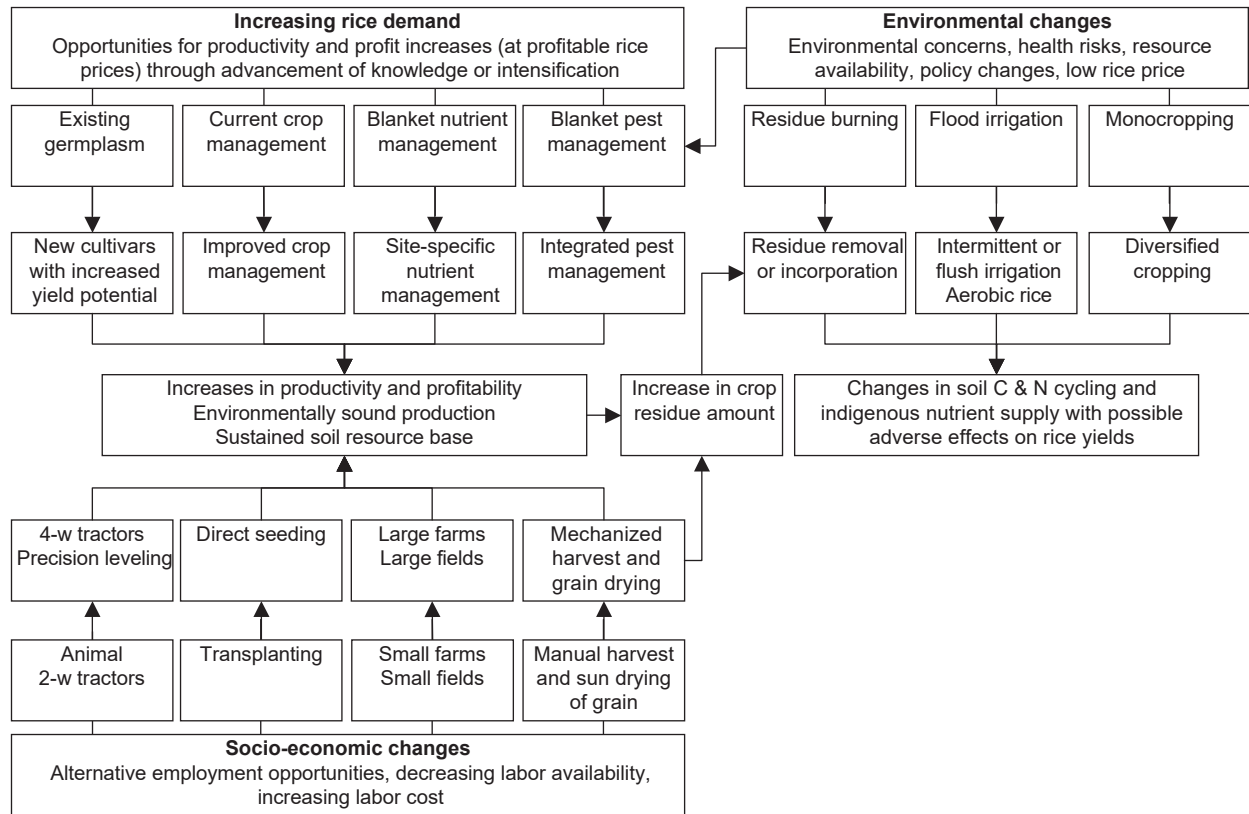


Fig. 1 Expected changes in rice demand, environment and socio-economic conditions affecting the productivity and management of paddy soils in intensive rice systems in Asia. Some of these changes have already occurred or are in progress in certain areas.



Plate 1 Flooded rice fields at transplanting (Bangladesh, 2005).



Plate 2 Transplanted rice in the field with standing water (Bangladesh, 2005).

acceptors in microbial respiration. The general sequence of reduction is O_2 - NO_3^- - $Mn(III,IV)$ - $Fe(III)$ - SO_4^{2-} - CO_2 - H^+ . Some of these redox systems are overlapping (e.g., O_2 , NO_3^- and $Mn(III,IV)$), whereas others are not (e.g., O_2 and $Fe(III)$; SO_4^{2-} and CO_2). Reduction of CO_2 causes the emission of methane from flooded paddy soils in fresh- water systems. The electrochemical changes induce a pH increase in most acid soils, whereas the pH decrease upon flooding of calcareous and sodic soils is due to the accumulation of carbon dioxide. Two distinctly different topsoil layers develop: a thin aerobic surface layer with a depth of few millimeters caused by O_2 diffusion through the aerobic flood-water layer (and around the rice root surface due to O_2 transport through the rice plant's aerenchyma) and the anaerobic soil layer below with a variable depth. Most roots are concentrated in the first 20 or 30 cm below the soil surface (Kyuma, 2004). The presence of aerobic and anaerobic soil compartments causes considerable soil N losses through mineralization-nitrification-denitrification reactions (Fig. 2). In most soils, flooding increases the plant availability of P, K, Ca, Mg, Na, Mn, Fe and Si, at least temporarily. Reduction of Fe and Mn-oxides releases sorbed and occluded elements such as P, Zn, Cu, B and Mo. Increasing pH reduces plant availability of Zn and Cu, whereas the solubility of Mo increases. Reduction of SO_4^{2-} reduces plant-available S but SO_4^{2-} is a common accompanying anion in commercial mineral fertilizers.

A widespread and traditional soil preparation practice in lowland production systems is puddling (Plate 3). Puddling refers to the practice of mixing surface soil with water to make it soft and leveled for transplanting, the traditional crop establishment method for rice. Other short-term puddling effects are the destruction of soil aggregates, irrigation water savings due to reduced percolation, reduced soil bulk density, increased soil water-holding capacity, decreased hydraulic conductivity, an acceleration of the electro-chemical processes described above, and good weed control. A long-term effect of puddling is the development of a hardpan (plow pan, traffic pan) in the subsoil below the puddled layer, often associated with the accumulation of Fe, Mn and Si (Kawaguchi and Kyuma, 1977). The puddled surface layer is often characterized by a coarser texture caused by clay mineral decomposition or transport processes. Hardpan formation may take from three to 200 years, depending on soil type, climate, hydrology and puddling frequency.

Paddy soil fertility and nutrient management

The nutrient-supplying capacity of paddy soils in the traditional rice-growing environments was largely determined by their parent material, and the alluvial deposits in the river deltas and flood plains. Natural nutrient inputs derived from sedimentation, irrigation and rainwater, organic residues or manure, biological N_2 -fixation, and carbon assimilation by floodwater flora and fauna (Greenland, 1997). Efforts to maintain soil fertility have historically centered around maintaining soil organic matter (SOM) through the addition of green manure, animal waste and crop residues, but their significance has declined continuously in the last decades (Dawe et al., 2003). The total annual C input in a double-rice cropping system may currently range from 4000 to 7000 kg C ha^{-1} , including inputs of 1000–2000 kg C ha^{-1} from the photosynthetic and heterotrophic biomass in rice fields. Despite high

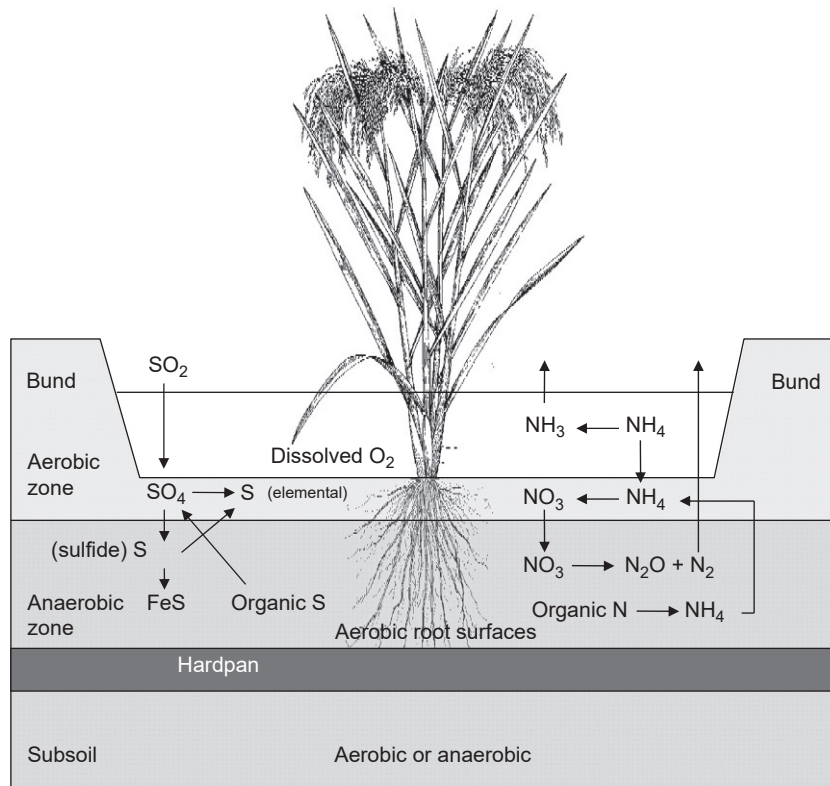


Fig. 2 Aerobic and anaerobic soil compartments in a flooded paddy soil, and related sulfur and nitrogen dynamics. Redrawn from Greenland DJ (1997) *The Sustainability of Rice Farming*. International Rice Research Institute (IRRI) and CABI Publishing: Manila/Wallingford. pp. 1–273.



Plate 3 Flooded rice fields after puddling (Bangladesh, 2005).

temperatures in the tropical soil floodwater environment and rapid initial breakdown of organic inputs, average SOM concentrations in paddy soils are higher than in aerated soils, mostly because of high organic matter inputs and a slower (anaerobic) decomposition rate of humus. The typical soil for irrigated lowland rice maintains an average carbon concentration of about 1.5–2% and a nitrogen concentration of about 0.15–0.2%; paddy soils are possibly an exception to the rule that the shift from natural vegetation to arable use always causes a large loss of soil organic carbon. Crop intensification with prolonged soil flooding changes the nature of SOM towards more aromatic compounds (chemical structures with six C-atoms arranged in a ring) because of changes in the breakdown of organic matter under increasingly anaerobic conditions. These changes appear to have limited effects on yield stability under rice monocropping, but were hypothesized to affect systems with intensive triple cropping and nearly continuous soil submergence (Kirk and Olk, 2000). However, an analysis of yield trends in long-term experiments with intensive rice-based cropping systems in tropical and subtropical regions of Asia suggested that yields in most experiments have remained stable with recommended fertilizer N, P and K addition (Dawe et al., 2000). There is also evidence that organic matter application did not improve grain yield trends in rice-based systems vis-à-vis a balanced mineral fertilizer application. But the application of organic matter is often profitable if used as a nutrient source complementary to inorganic fertilizer, also in part because of the recycling of often deficient micronutrients.

A unique contribution to the indigenous N supply of paddy soils derives from biological N_2 -fixation (BNF), providing 15–50 kg N ha⁻¹ per cropping cycle from blue-green algae in the floodwater and heterotrophic bacteria in the rice rhizosphere. The mineralization of algae and bacteria provides a constant supply of ammonium (NH₄⁺), the dominating inorganic-N form available to plants under anaerobic soil conditions. Because of efficient plant uptake of mineralized N, larger amounts of extractable NH₄⁺ are found only after soil flooding during the initial decomposition of residues or if fertilizer-N is applied (Fig. 3). The major N-loss mechanism during the cropping season is ammonia (NH₃) volatilization from the flooded soil, particularly when NH₄⁺ concentrations in the floodwater rise after N-fertilizer application. Nitrogen is also leached, denitrified or immobilized by soil microbial biomass and SOM. About 70% of the fertilizer-N applied by farmers is currently not recovered by crop N uptake, but N-recovery efficiencies of applied fertilizers can reach 50% or more in farmers' fields with adequate real-time N-management strategies (Dobermann et al., 2004). Considerable amounts of nitrate only accumulate in the aerated soil during fallow periods or during upland cropping cycles in rice-based rotations. Almost all nitrate is lost after soil flooding because of leaching and denitrification. Including upland crops in rice-based systems affects mineralization, plant availability and loss of inorganic-N. Soil aeration also accelerates the turnover of SOM and changes its composition towards a less aromatic nature, with possible adverse effects on the amount of indigenous soil-N supplied to rice. However, few long-term data are available from experiments

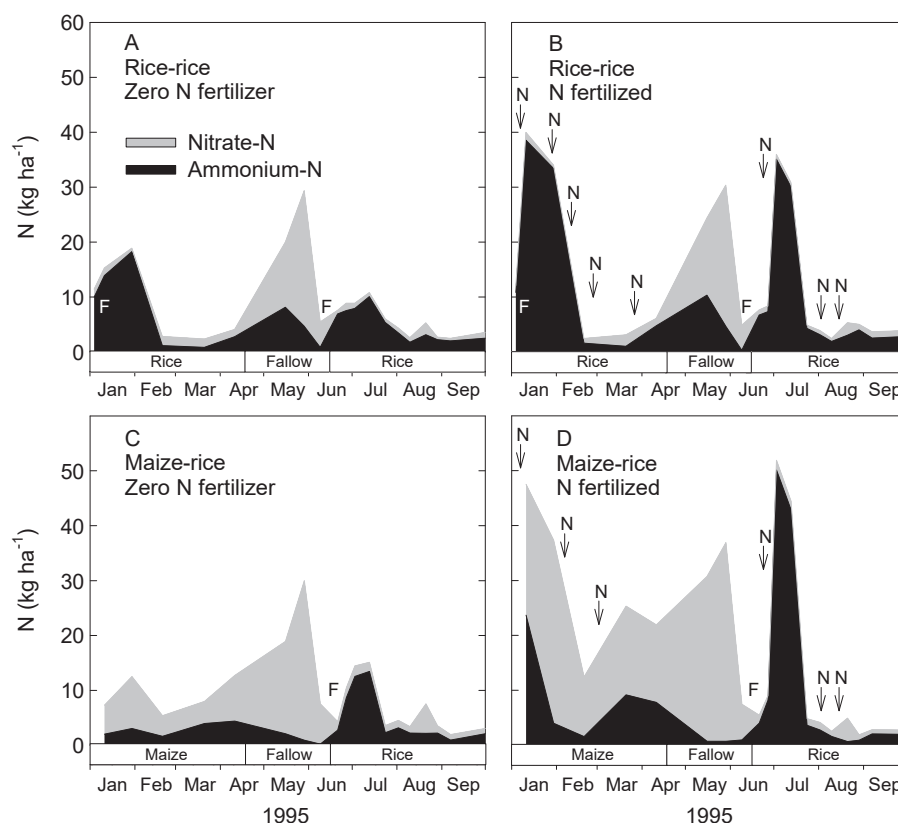


Fig. 3 Nitrate (NO₃⁻) and ammonium (NH₄⁺) dynamics in a rice-rice and a maize-rice cropping experiment with and without fertilizer N application, IRRI experimental farm, 1995. Fertilizer N was applied in several split applications with a total of 190 kg N ha⁻¹ in the dry season (January to mid-April) and 100 kg N ha⁻¹ in the wet season (mid-June to early October). F = flooding; N = fertilizer N split application.

that directly compare crop rotations but declining soil organic carbon concentrations were repeatedly reported from systems combining rice with upland crops. Other anticipated future changes in crop management include mechanization of soil preparation and harvest operations, changes in water management towards increasing soil aeration (see below), and larger amounts of crop residues remaining in the field for farmers to deal with (Fig. 1). These changes will affect soil C and N cycling in paddy soils, and research currently aims to develop a portfolio of technology and management options to mitigate potential adverse effects (Ladha et al., 2003).

Rice requires an adequate supply of nutrients from various sources to produce high and sustainable yields. A rice crop with a yield of 6 t ha⁻¹, for example, accumulates about 100 kg N, 18 kg P, 100 kg K, 0.3 kg Zn, 11 kg S, 480 kg Si, 21 kg Mg, 24 kg Ca, 3 kg Fe and less than 0.1 kg Cu and B. Without fertilizer, the indigenous supply of N limits yield in most paddy soils. Other nutrients, which have to be commonly supplied with fertilizer in intensive rice systems, include P and K, and in some areas also S or Zn. Required fertilizer rates are determined by the deficit between plant nutrient requirement and indigenous nutrient supply (Dobermann et al., 2004). This is particularly relevant for N, because residual effects of fertilizer-N application are small and long-term trends show little change in the indigenous supply of N with time in farmers' fields. For nutrients other than N, fertilizer application is often recommended even in the absence of a yield response (maintenance rates) to avoid depletion of soil nutrients particularly at elevated yield levels when nutrient removal with grain and straw is large. However, a comparison of current fertilizer rates used by farmers in seven major irrigated rice areas in Asia showed that fertilizer use is often not in congruence with the deficit between yield level and indigenous supply of nutrients (Table 2). On average, farmers used 90–140 kg N, 10–25 kg P and 10–55 kg K ha⁻¹ to achieve grain yields of 4.4–6.2 t ha⁻¹ in the typical annual double- or triple-rice cropping systems of South and Southeast Asia. This is about 50–75% of the climate-adjusted, genetic yield potential of 8.3 t ha⁻¹ that has been estimated for existing germplasm across sites, seasons and years. In general, yields and fertilizer rates are larger in the typical dry season with greater solar radiation than in the typical wet season with greater cloud cover. The indigenous supply of N, P and K in irrigated rice in Asia was on average 41–65 kg N ha⁻¹, 12–19 kg P ha⁻¹ and 61–102 kg K ha⁻¹, but varied widely with site, year, season, farm and field. In rainfed lowlands, yields and fertilizer applications are generally smaller, decreasing from favorable to unfavorable lowlands with increasing production risk.

Blanket fertilizer recommendations still prevail in most of Asia's lowland rice areas without considering amount and spatial distribution of indigenous nutrient supplies. There is also concern that fertilizer-rates of P and K are not optimally adjusted to long-term needs. A typical input-output balance for P and K in farmers' fields is given in Table 3. On-farm measurements in irrigated rice have confirmed for a wide range of soils and growing environments that farmers generate slightly positive P balances with current fertilizer P management, whereas their fertilizer-K application is insufficient to balance K removal in grain and straw. Modern site-specific nutrient management (SSNM) strategies have been developed in recent years to manage nutrients more efficiently and supply fertilizers as required by the crop, considering indigenous supplies of nutrients (Dobermann and Fairhurst, 2000) and have been shown to be beneficial for smallholder farmers (Chivenge et al., 2021). Estimates of indigenous supplies of nutrients need to be verified if general crop management of paddy soils changes (see Fig. 1). Changes in fallow period and residue management or system diversification are likely to introduce changes in indigenous supplies that are sufficiently large to require fertilizer adjustments.

Deficiency of Zn is the most widespread micronutrient disorder in rice, not only on soils with a traditionally small supply of Zn (e.g., neutral and calcareous soils containing a large amount of bicarbonate), but increasingly also on other soils under intensive rice cropping (Dobermann and Fairhurst, 2000). Sulfur deficiency occurs mainly in upland soils containing allophane (an aluminosilicate with primarily short-range structural order, which occurs as exceedingly small spherical particle especially in soils formed from volcanic ash), soils low in organic matter, highly weathered soils containing large amounts of Fe oxides, and sandy soils with high percolation. Lowland rice areas with common S deficiencies are located in Bangladesh, China, India, Myanmar, Pakistan, Sri Lanka and Thailand. Other nutrient disorders are usually only of local importance.

Table 2 Average seasonal indigenous N, P and K supply, fertilizer use, and grain yield in intensive rice systems of South and Southeast Asia.

Site	Indigenous nutrient supply (kg ha ⁻¹)			Fertilizer (kg ha ⁻¹)			Yield (t ha ⁻¹)
	N	P	K	N	P	K	
Jl	60–79	18–24	97–141	147–192	14–23	41–75	5.3–6.7
HA	49–68	12–19	61–102	90–118	13–30	50–80	5.5–6.7
AD	43–59	13–20	69–109	93–137	17–27	22–54	5.3–6.6
TH	30–44	14–19	47–65	83–124	9–25	27–45	4.6–6.2
SU	36–74	10–14	65–91	86–148	0–17	0–0	3.6–5.4
MA	43–65	10–18	65–100	75–142	9–20	8–31	3.9–5.3
OM	27–49	11–18	55–94	96–123	13–25	9–37	3.4–5.5
Mean	41–65	12–19	61–102	92–141	10–25	11–56	4.4–6.2

Jl: Jinhua, Zhejiang (China); HA: Hanoi, Red River Delta (Vietnam); AD: Aduthurai, Old Cauvery Delta, Tamil Nadu (India); TH: Thanjavur, New Cauvery Delta, Tamil Nadu (India); SU: Sukamandi, West Java (Indonesia); MA: Malingaya, Central Luzon (Philippines); OM: Omon, Mekong Delta (Vietnam). Ranges refer to interquartile ranges (25th, 75th percentiles) of on-farm data reported by Dobermann et al. (2003).

Table 3 Typical input-output balance for P and K during one cropping cycle in intensive irrigated rice systems in South and Southeast Asia.

Parameters for P and K balance	Unit	P	K
Input with inorganic fertilizer	kg ha ⁻¹	10–25	10–55
Input with irrigation water	kg ha ⁻¹	0–1	10–30
Input with rainwater	kg ha ⁻¹	0–1	0–10
Losses with percolation	kg ha ⁻¹	0–1	5–15
Removal with grain and straw	kg ha ⁻¹	10–15	50–65
Nutrient balance	kg ha ⁻¹	0 to +11	–35 to +15

Based on Dobermann A, Witt C, Abdulrachman S, Gines HC, Nagarajan R, Son TT, Tan PS, Wang GH, Chien NV, Thoa VTK, Phung CV, Stalin P, Muthukrishnan P, Ravi V, Babu M, Simbahan GC, Adviento MAA, Bartolome V (2003) Estimating indigenous nutrient supplies for site-specific nutrient management in irrigated rice. *Agronomy Journal* 95: 924–935. doi: 10.2134/agronj2003.9240.

Lowland rice systems and declining water availability

Today, nearly 90% of the fresh water diverted for human use in Asia goes to agriculture and, of this, more than 50% is used to irrigate rice. With the growing demand for water for non-agricultural uses (domestic, municipal, industrial and environmental), the proportion available for agriculture is projected to decline from 90% to 73% in developing countries in the near future. Indicators of a looming water crisis include falling groundwater tables in regions where mainly tube wells are used for irrigation and severe water shortages downstream of some major rivers in Asia (e.g., Yellow River, Indus and Ganges). Fig. 4 gives the distribution of irrigated and rainfed lowland rice in Asia, and major ‘hot spots’ of groundwater depletion and reduced river water flows. The productivity of intensive rice systems depends strongly on the availability of sufficient irrigation water, and production is likely to decrease with water shortage unless suitable mitigation options are developed (Tuong and Bouman, 2003). By 2020, physical water scarcity was projected for wet-season irrigated rice in North China (2.5 million hectares), Pakistan (2.1 million hectares) and north and central India (8.4 million hectares), as well as on 2 million hectares of dry-season rice in central India.

Water productivity at the field level can be increased by (i) increasing yield per unit cumulative evapotranspiration through the use of high-yielding, short-duration varieties and optimal, site-specific agronomic practices, (ii) more effectively using rainwater by using dry-seeding rice technologies to synchronize the start of the cropping season with the start of the rainy season, and (iii) reducing unproductive water losses (seepage, percolation, evaporation) through adequate soil and water management. Examples of the latter include saturated soil culture (daily flooding with about 0.01 m water depth) or alternate wetting-and-drying regimes (flooding every 2–7 days with 0.02–0.05 m water depth). These measures frequently cause a reduction in yield and profitability, although they often result in an increased water productivity. A more radical approach is to grow rice under permanently aerobic soil conditions by keeping the soil water content in the root zone always at or below field capacity, a technology called aerobic rice. Contrary to upland rice, this system is intended to be a high-yielding system, but more research is required to evaluate long-term effects on soil organic matter and nutrient cycling. Such systems may gain local importance where water is limiting and prices for rice are high, but yields may not reach those achieved in well-managed flooded systems. The adoption of such water-saving irrigation technologies will cause a shift from mainly anaerobic soils to partly or even completely aerobic soils, but most research examining this issue and possible consequences is still inconclusive. Changes in the soil moisture regime of paddy soils will have profound effects on soil organic matter turnover, nutrient dynamics, carbon sequestration, soil productivity, weed ecology and greenhouse gas emissions in irrigated systems. Water-saving technologies must therefore develop into integrated natural resource management approaches to maintain the productivity, profitability and sustainability of the evolving systems.

Water productivity in rainfed lowland systems can be increased through more efficient and effective use of all inputs. A general increase in productivity will also reduce the pressure for more intensive management and area expansion of irrigated lands. Direct-seeding techniques offer further opportunities to achieve greater (water) productivity by timely crop establishment. Avoiding drought stabilizes yields and in some situations a second, upland crop can be grown after rice using the residual soil moisture content. Such crop intensification must be accompanied by balanced fertilizer application and adequate residue management on the often-poor-quality soils in rainfed systems to maintain or build up SOM and thereby the indigenous supply of nutrients.

Sustainability and environmental issues

In general, rice cultivation is sustainable, even where rice is continuously grown in monocropping at high levels of production (Greenland, 1997). The quality of the soil resource base can be maintained through adequate management practices, including proper crop residue and water management, tillage operations, soil aeration during fallows and balanced fertilizer application (Yanai et al., 2022). The sustainability of rice cropping may be at risk, however, where management-induced changes strongly influence soil characteristics and the soil environment. This may include prolonged soil submergence in intensive systems with three

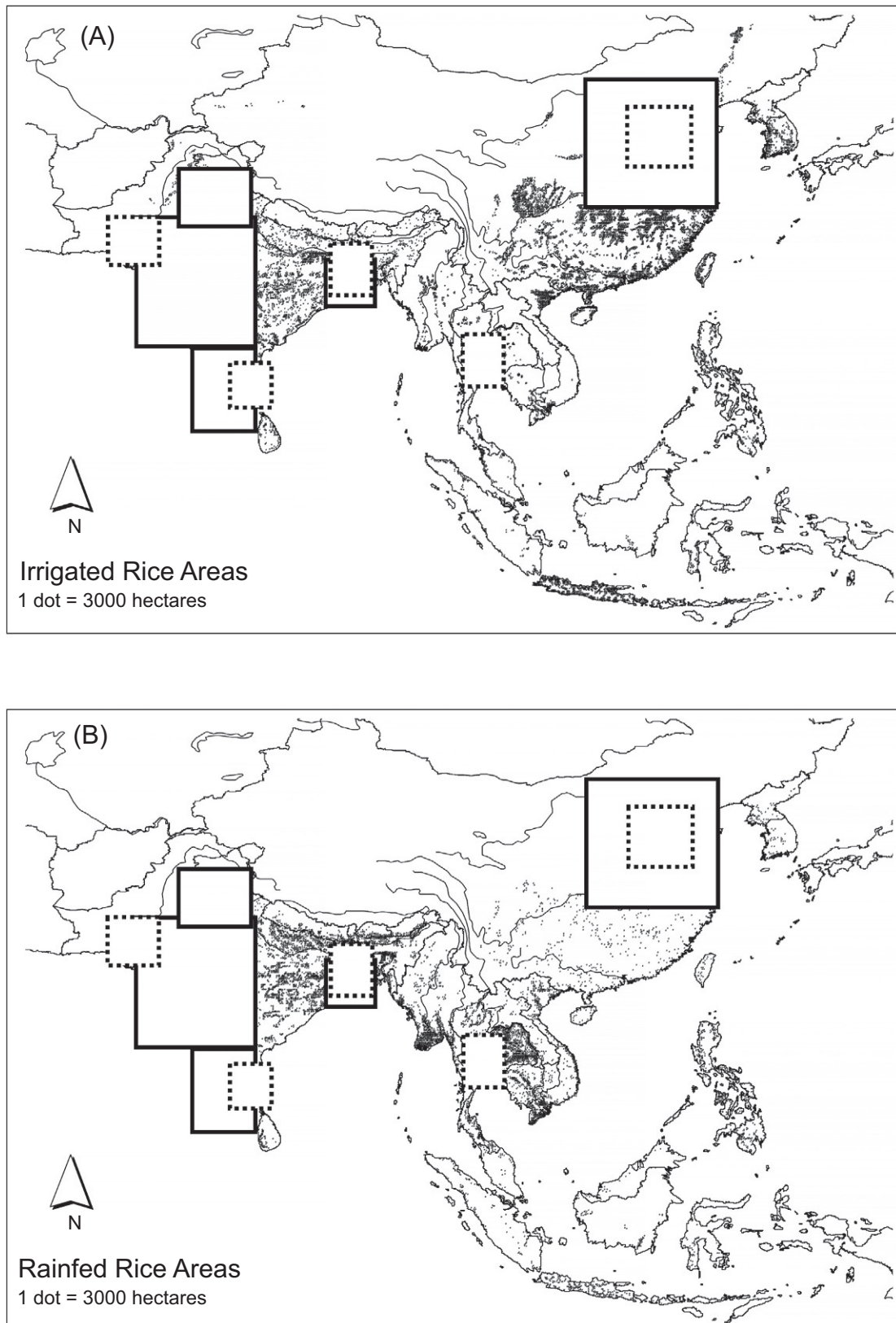


Fig. 4 Current distribution of irrigated (A) and rainfed lowland rice (B) in Asia (based on Haefele SM, Nelson A, Hijmans RJ (2014) Soil quality and constraints in global rice production. *Geoderma* 235–236: 250–259.), and major ‘hot spots’ of groundwater depletion (squares with solid lines) and reduced river water flows (squares with broken lines).

rice crops per year as well as prolonged soil aeration because of crop diversification (Fig. 1). In the following, the focus is on selected issues that are of general importance for sustainable paddy soil management and the environmentally sound production of rice.

When crops are grown with irrigation in dry regions, the sustainability of agriculture is potentially threatened by several water-related processes, such as salinization and alkalization. In the semiarid tropics, irrigation may cause salinization of soils by rising saline groundwater tables. Alkalization can occur when irrigation water with a positive residual alkalinity is used. However, ample irrigation can leach salts from the topsoil through percolation if sufficient drainage is available. This even makes irrigated rice a reclamation crop for such soils. A very different water-related problem is accidental flooding caused by inadequate irrigation control in some parts of tropical Asia, taking valuable land out of agricultural production: for example, this accounts for about 6 million ha of arable land in India annually.

Other soils that pose problems for management in the lowlands include Thionic Fluvisols or acid sulfate soils, which are widespread in Thailand, Vietnam and Indonesia. Potential acid sulfate soils strongly acidify upon drainage, whereas actual acid sulfate soils are characterized by pH values around four. Rice cropping under submergence offers one of the few sustainable uses of these soils. Management strategies include (i) complete soil oxidation through drainage and leaching of acidity from the crop root zone, which offers a permanent solution but may cause a contamination of surface water affecting crops, soils and aquatic organisms in surrounding regions, or (ii) the limitation of soil oxidation by maintaining a permanent, high groundwater table, which requires sufficient irrigation water.

Environmental concern is increasing because the intensification of rice cropping in the past has led to an excessive use of fertilizer-N and pesticides in many intensive rice-growing areas in Asia. This has been recognized, and much progress has been made in recent years in developing and disseminating adequate management strategies, such as real-time N management and integrated pest management. Current changes in many cropping environments include a shift from irrigated transplanted rice to direct seeding and the use of water-saving technologies. This causes greater weed pressure and often leads to an increased use of herbicides, particularly at sites with further intensification and mechanization because of labor shortages. Research efforts aim to better integrate water, nutrient, weed and pest management approaches to enhance direct environmental and economic benefits for farmers.

Paddy soil environments have also caused global environmental concerns because methane (CH_4) is second in importance to CO_2 as a greenhouse gas and flooded paddy soils contribute about 10% of global CH_4 emissions (Wassmann et al., 2000). Although the CH_4 production of paddy soils is much smaller than originally assumed, it is still an important contribution. Methods to reduce this share include water-saving technologies as described above, improved crop residue management and direct seeding instead of transplanting. Care has to be taken to avoid significant increases in N_2O -emissions, another greenhouse gas which is ten times more “effective” than methane. Soil moisture regimes keeping the soil redox potential between -100 and $+200$ mV combined with efficient N fertilizer management could minimize both methane and nitrous oxide emissions.

Conclusions

Paddy soils are defined through their production environment, the typically irrigated and rainfed lowland rice cropping systems. Cropping intensification in the last 40 years has greatly increased rice yield and production, and further intensification is needed to achieve food security at the national and household level. Productivity increases at the farm level could be mainly achieved in two ways. One option is the continuing optimization of the existing rice-based cropping systems with two to three annual rice crops and short fallow periods. There is strong evidence that rice double cropping in Asia can provide high, sustainable yields despite an increase in the duration of soil submergence and a shift from organic to inorganic nutrient sources in the last 40 years. Sufficient inputs of external nutrients with fertilizer are necessary to match the insufficient supply of nutrients from indigenous sources, but inefficient fertilizer use, nutrient imbalances and excessive pesticide use have become widespread and must be overcome through more site- and season-specific crop, soil and water management to efficiently manage all inputs. The second option for farmers is to turn to crop diversification, adding more profitable non-rice crops to the rice-based cropping system or grow rice under aerobic soil conditions. Conditions imposed in the non-rice season will influence important characteristics of paddy soils and remove typical paddy soil characteristics temporarily or permanently. A key issue is whether the high productivity of current rice cropping systems can be sustained in such diversified systems. The other big uncertainty is how changing rain patterns will affect water availability or flooding patterns in many rice environments, which might soon change cropping systems and paddy soils in many regions of Asia and beyond.

References

- Chivenge P, Saito K, Bunquin MA, Sharma S, and Dobermann A (2021) Co-benefits of nutrient management tailored to smallholder agriculture. *Global Food Security* 30. <https://doi.org/10.1016/j.gfs.2021.100570>.
- Dawe D, Dobermann A, Moya PF, Abdulrachman S, Singh B, Lal P, Li SY, Lin B, Panaullah G, Sariam O, Singh Y, Swarup A, Tan PS, and Zhen Q-X (2000) How widespread are yield declines in long-term rice experiments in Asia? *Field Crops Research* 66: 175–193.
- Dawe D, Dobermann A, Ladha JK, Yadav RL, Lin B, Lal P, Panaullah G, Sariam O, Singh Y, Swarup A, and Zhen Q-X (2003) Does organic matter improve the sustainability and profitability of intensive rice systems? *Field Crops Research* 83: 191–213.

- Dobermann A and Fairhurst T (2000) *Rice: Nutrient disorders and nutrient management*. Singapore/Los Baños: Potash & Phosphate Institute (PPI), Potash & Phosphate Institute of Canada (PPIC), and International Rice Research Institute (IRRI), pp. 1–191.
- Dobermann A, Witt C, Abdulrachman S, Gines HC, Nagarajan R, Son TT, Tan PS, Wang GH, Chien NV, Thoa VTK, Phung CV, Stalin P, Muthukrishnan P, Ravi V, Babu M, Simbahan GC, Adviento MAA, and Bartolome V (2003) Estimating indigenous nutrient supplies for site-specific nutrient management in irrigated rice. *Agronomy Journal* 95: 924–935. <https://doi.org/10.2134/agronj2003.9240>.
- Dobermann A, Witt C, and Dawe D (eds.) (2004) *Increasing Productivity of Intensive Rice Systems Through Site-Specific Nutrient Management*, pp. 1–410. Enfield, NH/Los Baños: Science Publishers, Inc./International Rice Research Institute (IRRI).
- Greenland DJ (1997) *The Sustainability of Rice Farming*. Manila/Wallingford: International Rice Research Institute (IRRI) and CABI Publishing, pp. 1–273.
- GRISP (Global Rice Science Partnership) (2013) *Rice Almanac*, 4th edn. Los Baños: International Rice Research Institute, p. 283.
- Haefele SM, Nelson A, and Hijmans RJ (2014) Soil quality and constraints in global rice production. *Geoderma* 235–236: 250–259.
- Huke RE and Huke EH (1997) *Rice Area by Type of Culture: South, Southeast, and East Asia A Revised and Updated Data Base*. International Rice Research Institute (IRRI), pp. 1–59.
- IRRI (1978) *Soils and Rice*. Los Baños: International Rice Research Institute (IRRI), pp. 1–825.
- IUSS Working Group (2015) World Reference Base for Soil Resources 2014, Update 2015. In: *International Soil Classification System for Naming Soils and Creating Legends for Soil Maps*. Rome: Food and Agriculture Organization. World Soil Resources Reports No 106.
- Kawaguchi K and Kyuma K (1977) *Paddy Soils in Tropical Asia: Their Material Nature and Fertility*. Honolulu: The University Press of Hawaii, pp. 1–258.
- Kirk GJD (2004) *The Biogeochemistry of Submerged Soils*. Chichester: John Wiley, pp. 1–291.
- Kirk GJD and Oik DC (eds.) (2000) *Carbon and Nitrogen Dynamics in Flooded Soils*, pp. 1–188. Los Baños: International Rice Research Institute (IRRI).
- Kyuma K (2004) *Paddy Soil Science*. Kyoto/Melbourne: Kyoto University Press, Trans Pacific Press, pp. 1–280.
- Ladha JK, Hill JE, Duxbury JM, Gupta RK, and Buresh RJ (eds.) (2003) *Improving the productivity and sustainability of rice-wheat systems: Issues and impacts, Proceedings of an International Symposium Cosponsored by ASA, IRRI and CIMMYT at Charlotte, NC, USA, 22 October 2001, ASA Special Publication Number 65*, pp. 1–231.
- Moormann FR and van Breemen N (1978) *Rice: Soil, Water, Land*. Los Baños: International Rice Research Institute (IRRI), pp. 1–185.
- Rosegrant MW, Sombilla MA, and Perez N (1995) *Global Food Projections to 2020: Implications for Investment*. Food, Agriculture and the Environment. Discussion Paper No. 5. Washington, DC: International Food Policy Research Institute, pp. 1–54.
- Tuong TP and Bouman BAM (2003) Rice production in water-scarce environments. In: Kijne JW, Barker R, and Molden D (eds.) *Water Productivity in Agriculture: Limits and Opportunities for Improvement. The Comprehensive Assessment of Water Management in Agriculture Series*, vol. 1, pp. 53–67. Wallingford: CABI Publishing.
- USDA (1999) Soil taxonomy. In: *A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn., p. 871. United States Department of Agriculture.
- Wassmann R, Lantin RS, and Neue HU (eds.) (2000) *Methane Emissions from Major Rice Ecosystems in Asia*. In: *Reprinted From Nutrient Cycling in Agroecosystems*, vol. 58, pp. 1–394. Dordrecht, the Netherlands: International Rice Research Institute (IRRI) and Kluwer Academic Publishers. Nos. 1–3.
- Yanai J, Tanaka S, Nakao A, Abe SS, Hirose M, Sakamoto K, Masai F, Saito H, Kajiwara N, Dejhimon K, Sriprachote A, Kanyawongha P, Lattirasuvan T, Timbas N, Medina S, Paing TN, Yusoff KHM, Wakatsuki T, and Kyuma K (2022) Long-term changes in paddy soil fertility in tropical Asia after 50 years of the Green Revolution. *European Journal of Soil Science* 73(1), e13193. <https://doi.org/10.1111/ejss.13193>.

Volcanic soils and non-volcanic andic soils

Rupert Bäumler^a and G Uehara^b, ^aInstitute of Geography, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany;
^bUniversity of Hawaii, Honolulu, HI, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	459
Rock weathering	460
Soil formation	460
Vertisols	461
Oxisols or Ferralsols	461
Ultisols or Acrisols	464
Andisols or andosols of volcanic origin	464
Andosols of non-volcanic origin	464
Human interactions with volcanic soils	469
Conclusions	469
References	470

Abstract

The chapter informs about the formation and the variety of soils from volcanic parent materials, their pathways of weathering, their processes of soil formation, and their most important properties. A special focus is given on the classification of soils of the Andosol/Andisol type developed from non-volcanic parent materials, as their formation is not yet fully understood.

Glossary

Allophane An aluminium(Al)-hydroxo-silicate of the clay fraction that cannot be identified by X-ray diffraction because of its amorphous or poorly crystalline structure (short-range-order minerals). Allophanes are common weathering products of volcanic ash, and are characterized by surface areas up to 1000 m² g⁻¹.

Alu-andic Andic properties mainly caused by Al-hydroxo-organic complexes.

Andic properties Soil properties characterized by short-range-order minerals and or metal–organic complexes as a result of the weathering of volcanic ejecta. Common features are large contents of soil organic matter throughout the profile, low bulk density, thixotropic behavior, high P-retention.

Andosol or Andisol Soils commonly developed from volcanic ejecta having an intermediate stage of soil development with andic properties characterized by short-range-order minerals and or metal–organic complexes. Andosols, however, may also develop from non-volcanic parent materials under specific site conditions.

Cryptopodzols Soils fulfilling all classification criteria of Podzols (spodic properties), but without visible eluvial and illuvial horizons.

Ferro-andic Andic properties mainly caused by Fe(iron)-hydroxo-organic complexes.

Imogolite An aluminium-hydroxo-silicate of the clay fraction with tubular structures commonly found in Andosols as weathering products of volcanic ejecta.

Sil-andic Andic properties caused by Si-hydroxo-organic complexes. The acid-oxalate extractable silica content should be 0.6% or more.

Spodic properties Soils (Podzols) characterized by very low pH values and leaching of metal–organic complexes and or oxidic Fe- and Al- compounds from topsoil to subsurface horizons commonly in coarse-textured parent materials under cool-humid climatic conditions (podzolization).

Thixotropic features Abrupt change of the viscosity of materials under shear stress. It may happen in soil material with andic properties.

Vitric Vitric soil properties represent an early stage of soil development between fresh pyroclastic deposits containing volcanic glass and andic properties.

Key points

- Soil formation on parent materials of volcanic origin in general
- Andosols or Andisols of volcanic origin
- Andosols and soil with andic properties on non-volcanic parent materials

Introduction

Molten **magma** extruded as liquid lava or ejected as cinders or ash from deep within the Earth adds to the variety of parent materials and soils we find on landscape surfaces. Unlike their plutonic counterparts that cool and solidify slowly underground and are subject to weathering after uplift and erosion, **volcanic rocks** cool and solidify immediately upon exposure to the atmosphere. The faster cooling results in the formation of amorphous/glassy and finer rock textures in volcanic rocks (Fig. 1). Plutonic and volcanic rocks with similar or different chemical composition will lead to the formation of different soils depending on time of soil development, i.e. age volcanic rock exposure, and site conditions. This is especially true for finely divided pyroclastic **ejecta** that cool even before deposition (Fig. 1). The porosity and texture of such cindery or ashy materials differ greatly from those of other parent materials, and vitric volcanic ash develops into soils with andic soil properties, i.e. the Andosol or **Andisol** type of soils.

In addition to rock texture and porosity, volcanic materials vary in chemical and mineral composition. The chemical composition of the magma determines, to a large extent, the texture and porosity of the extruded material. Vitric and porous pyroclastic rocks are more common in high-silica magma that forms acidic rocks than the less viscous, basic magma that erupts less explosively and produce basaltic lavas (Nockolds, 1954; Fig. 1).

The higher viscosity of silicic magma results in the formation of steep-sided, cone-shaped **volcanoes** exemplified by Mount Fuji-san in Japan, whereas the basic, low-viscosity magma forms broader, shield volcanoes exemplified by the volcanoes of the Galapagos Islands. Both types of volcanoes produce **pyroclastic materials** and lava, but the silicic magmas have more frequent explosive eruptions than shield volcanoes. The 1980 eruption of Mount St. Helens in the U.S. illustrates the explosive nature of intermediate and silicic eruptions often forcing nearby residents to evacuate the area (Christiansen and Peterson, 1981). So-called phreatomagmatic eruptions occur when lava gets in contact with water or ice leading to violent eruptions by abrupt water vaporization, thus producing large volumes of volcanic ash. The 2021 Fagradalsfjall eruption in Iceland, however, is a good example of molten lava pouring out from vents and fissures along **rift zones** or shield volcanoes.

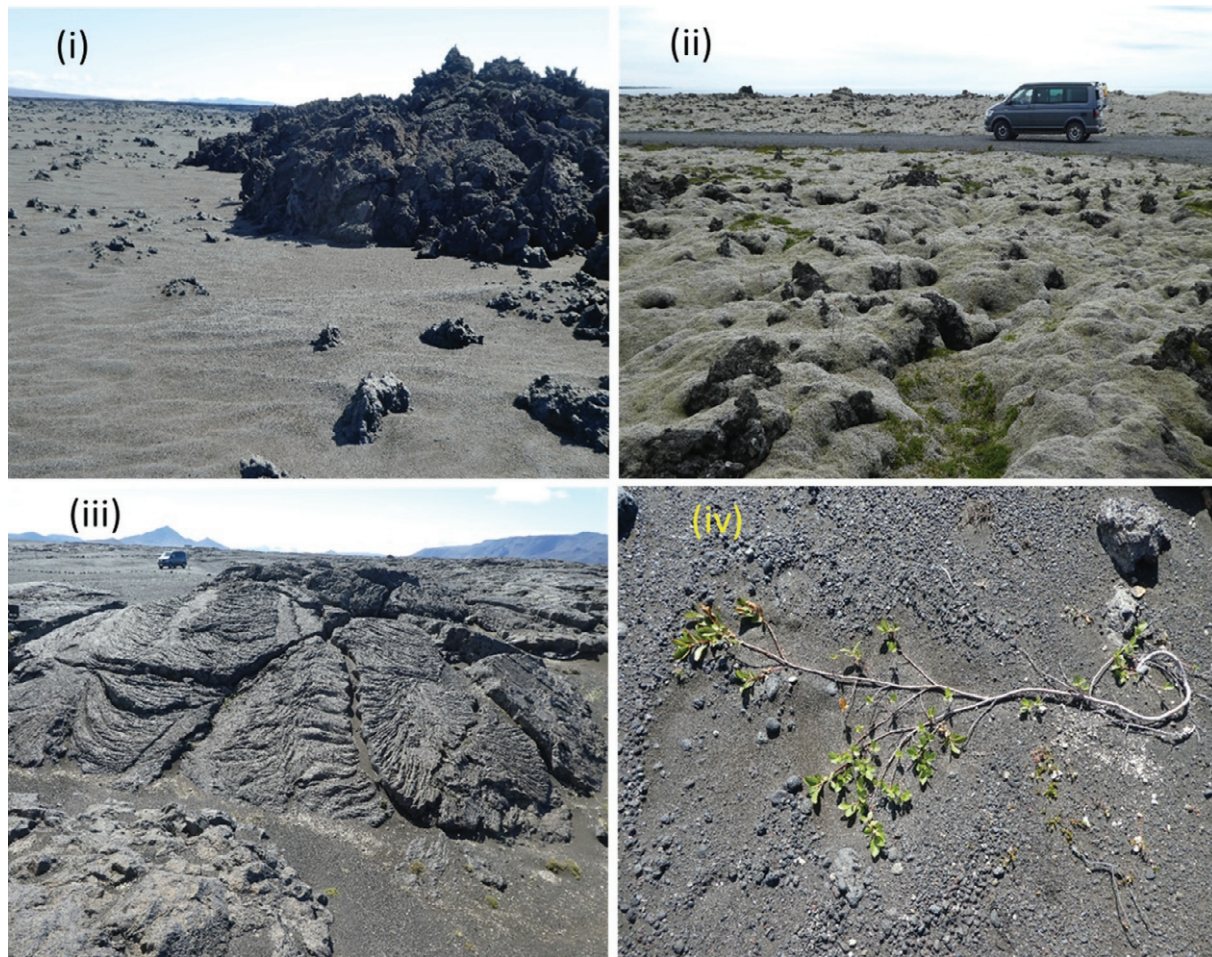


Fig. 1 Common parent materials of volcanic origin (i) A'a lava from the Holuhraun lava field in Central Iceland, (ii) lava flow covered with moss from SW Iceland near Krysvík, (iii) Pahoehoe lava near Herðubreið, Central Iceland, (iv) pyroclastic ejecta with *Salix spec.* Pioneer plant, Central Iceland. Photographs by R. Bäumlér.

Table 1 Chemical composition of volcanic rocks (Nockolds, 1954).

	<i>Alkali rhyolite</i>	<i>Dacite</i>	<i>Andesite</i>	<i>Tholeiitic olivine basalt</i>	<i>Nephelinite</i>
Silicon dioxide	74.57	63.58	54.2	47.9	39.07
Titanium dioxide	0.17	0.64	1.31	1.65	3.86
Aluminium oxide Al ₂ O ₃	12.58	16.67	17.17	11.84	12.82
Ferric oxide Fe ₂ O ₃	1.30	2.24	3.48	2.32	8.75
Ferrous oxide FeO	1.02	3.00	5.49	9.8	6.39
Manganese oxide MnO	0.05	0.11	0.15	0.15	0.26
Magnesium oxide MgO	0.11	2.12	4.36	14.07	6.14
Calcium oxide CaO	0.61	5.53	7.92	9.29	14.20
Sodium oxide Na ₂ O	4.13	3.98	3.67	1.60	4.09
Potassium oxide K ₂ O	4.73	1.40	1.11	0.54	2.07

Table 1 shows the range of silica content of volcanic rocks. As silica decreases the content of iron (Fe), magnesium (Mg), calcium (Ca), and sodium (Na) increases. These differences in chemical composition influence the rocks' susceptibility to weathering. In general, the basic rocks weather more rapidly owing to the ease with which Na-, Ca-, Mg-, and Fe-bearing silicate minerals undergo hydrolysis and dissolution. The rate of weathering is even faster in fine-grained pyroclastic material (Bonatotzky et al., 2021).

Rock weathering

The rate of weathering of volcanic rocks is influenced by their chemical and mineral composition, glass content, and particle size (Lowe, 1986; Hatano et al., 2021). Volcanic ash, for example, weathers comparably fast owing to its high glass content and large specific surface area associated with smaller particle sizes (Bonatotzky et al., 2021; Candra et al., 2021). Volcanic glass, however, varies in chemical composition. Felsic glass is high in silicon (Si), potassium (K), and sodium and is more resistant to weathering than mafic glass, which is high in iron and magnesium.

Artifacts such as arrowheads made of obsidian show little signs of weathering even after centuries of burial, primarily because they are felsic in composition and have low surface-area-to-mass ratios. It is a different matter with ash particles, called shards, which geologists and soil scientists often use as stratigraphic markers and evidence of ancient volcanic eruptions. Shards have high surface-area-to-mass ratios so that volcanic ash particles still identifiable as shards after deposition and buried for thousands of years are more likely to be felsic, as mafic shards would long have been weathered to clay minerals (Bleeker and Parfitt, 1974).

Soil formation

From a taxonomic standpoint, 'volcanic soil' is not a particularly useful term because a great variety of soils can have volcanic origins and or andic properties. Soils characterized by clay migration can form from volcanic rocks where rainfall is sufficiently high to leach bases, and conditions exist that favor activation and illuviation of clay. It is not unusual to find two soil orders, in close proximity that have formed from the same parent rock.

With the high porosity and large pores of most lavas and ash deposits, runoff is rare, and plants are able to establish quickly on these deposits (Kuzera et al., 2005). Nitrogen-fixing lichens and moss are the first to colonize lava flows, and, under warm and humid climates, dense forests can cover lava fields in less than a century. In humid areas, the soils that form first are organic soils of the Histosol Reference Soil Group. Histosols on volcanic parent material are often well drained and consist primarily of an organic surface horizon resting on rocks or fragmented materials (Hatano et al., 2021). If the underlying rock is A'a lava (Hawaiian lava flow of broken and rough surface) the soils are often Typic Folist or Folic Histosols, but if the rock is Pahoehoe lava (Hawaiian lava flow of smooth to ropy surface) the soil is commonly a Lithic Folist or Leptic Histosols. This distinction between them is crucial from an agricultural standpoint. Folic Histosols with their 'clinkery' A'a substrate are ideal for use as orchards, but their Leptic counterpart, with an unbroken and smooth rock substrate, is not. Fig. 1 shows some of the common parent materials of volcanic origin with or without pioneer plants or vegetation cover.

In the case of volcanic ash, plants may readily send roots deep into the loose ash deposit, and a rapid accumulation of organic matter occurs throughout the rooting depth. The first recognizable soil would most likely be a Vitric Andosol/Vitrand. As the names imply, those soils are Andosols/Andisols consisting mainly of unaltered or less altered glass fragments (Dethier et al., 1981).

Histosols and Vitric Andosols do not remain unchanged for very long. While soils change little in a human lifespan, one can, given the right circumstances, understand how volcanic soils change with time (Vilmundardóttir et al., 2015; Candra et al., 2021). Such circumstances exist in almost all volcanically active regions, e.g. Iceland, Japan, Indonesia, the Hawaiian Islands, New Zealand, which arise from the movement of tectonic plates over stationary hotspots, at subduction zones or mid-ocean ridges, rift zones (Arnalds, 2004; Bonatotzky et al., 2019; Saigusa et al., 1978; Stevens and Vucetich, 1984). The Hawaiian Islands for example formed



Fig. 2 A highly weathered [Andosol](#), an Acrudoxic Hydrudand. Photograph by G. Uehara.

from outpouring of lavas from the hotspot and are displaced in a northeasterly direction; they form a chain of islands that makes it almost perfect for studying volcanic soils. The age of the parent materials in the island chain ranges from recent on the island of Hawaii to 5 million years on the island of Kauai. Ten of the 12 soil orders of Soil Taxonomy occur in the island chain. Unlike Follic Histosols and Vitric Andosols that can form in less than a century, and are common on the youngest island of Hawaii, [Oxisols](#) and Ultisols that require prolonged weathering and leaching to develop, dominate on Kauai, the oldest island, while young Follics and Vitrandis do not occur there ([Beinroth et al., 1974](#)). The greatest diversity of soils exists on the island of Oahu, midway between Hawaii and Kauai. Here soils found on the older islands dominate, but young, post-erosional [volcanic eruptions](#) have rejuvenated part of the landscape. Soils that do not occur on the young islands, but peak on the intermediate islands, are [Vertisols](#). Their abundance on Kauai (the oldest island) is slightly diminished. The Vertisols form in valley floors on the dry, leeward side of the older islands. There, rainfall is lower and the bases that accumulate in the valley floor provide the right conditions for [smectite](#) clay formation. Vertisols do not occur on the young islands because valleys have not yet formed there. On the oldest islands in the wetter zones, Vertisols lose their bases and eventually become kaolinitic.

The soils developed from volcanic ash age much faster ([Bonatutzky et al., 2021](#)). Vitrandis quickly become ‘Haplic Andosols’ in the drier zone and hydromorphic Andosols in the wetter areas ([Bonatutzky et al., 2019](#)). The Hydrudands are some of the lightest soils in the world, with bulk densities of less than $0.5 \text{ cm}^{-3} \text{ g}$ and field water content by weight generally exceeding 100%. Owing to their porous nature, they are readily leached of bases and silica, leaving a residue high in hydrated [iron and aluminum oxides](#). These highly weathered, prematurely old soils occur on the youngest islands of Hawaii ([Fig. 2](#)).

Vertisols

A combination of basic parent material and a low-leaching environment often serves as a necessary condition for Vertisol formation. Sediment deposited by streams and rivers cutting through basaltic rock often provides the first condition, and the semiarid tropics that circle the globe north and south of the equatorial [humid tropics](#) provide the second.

[Table 2](#) gives the chemical [properties](#) of a Vertisol from Hawaii. They have low [organic carbon](#) contents and their main chemical and physical properties are determined by the large clay content ([Shoji et al., 1982](#)).

Oxisols or Ferralsols

A key feature of Ferralsols/Oxisols is the low [cation exchange capacity](#) (CEC) of their clay content. This condition is most likely to occur when the clay fraction is low in layered [silicate](#) clay and rich in iron oxides. [Volcanic rocks](#) low in silica and rich in ferromagnesium minerals weather rapidly to Oxisols under warm and humid conditions. While time and intense leaching can produce Oxisols from a variety of materials, volcanic materials with porous characteristics produce Oxisols with greater ease. The chemical composition and selected properties of an Oxisol formed from basic volcanic rock are given in [Table 3](#). The extreme weathering of the soil is indicated by its placement in the Acrudoxic subgroup (US Soil Taxonomy) and also by the increase in the

Table 2 Chemical composition and properties of a Gypsic *Vertisol* developed from basaltic *alluvium* (Lualualei series, island of O'ahu: fine, smectitic, isohyperthermic, Typic Gypsite; (USDA-SCS, 1978)).

Depth (cm)	Chemical composition (% of whole soil)											Chemical properties								
												Exchangeable bases (cmol kg ⁻¹)							pH (1:5)	
	SiO ₂	TiO ₂	Al ₂ O ₃	Fe ₂ O ₃	MnO ₂	MgO	CaO	Na ₂ O	K ₂ O	P ₂ O ₅	LOI	C (%)	N (%)	Ca	Mg	Na	K	CEC (cmol kg ⁻¹)	H ₂ O	1 mol l ⁻¹ KCl
0–3	30.54	7.71	19.29	26.26	0.30	1.50	1.57	0.26	0.07	0.50	11.63	0.66	0.08	17.1	15.2	0.8	1.4	34.1	7.1	5.9
3–25	31.74	7.59	20.34	25.81	0.33	1.07	1.72	0.28	0.08	0.48	10.81	0.42	0.06	15.1	15.1	1.3	0.4	32.9	7.2	5.4
25–35	33.91	7.52	19.04	26.28	0.45	1.57	0.62	0.29	0.05	0.51	10.87	0.21	0.05	14.6	13.4	2.5	0.2	32.9	7.2	5.4
35–75	33.54	7.59	19.57	25.82	0.43	1.72	0.34	0.37	0.06	0.36	10.54	0.17	0.04	16.0	10.5	3.9	0.2	32.1	6.8	5.2
75–123	32.25	6.18	20.02	24.24	0.33	1.58	2.56	0.44	0.04	0.38	11.39	0.17	–	73.1	9.3	7.2	0.2	29.8	5.6	4.9
>123	31.52	6.57	20.39	22.37	0.38	1.50	4.39	0.49	0.07	0.29	11.56	0.16	–	73.8	10.7	8.4	0.3	30.6	5.8	5.0

LOI: loss on ignition; CEC: cation exchange capacity; C: organic carbon; N: organic nitrogen; P₂O₅: phosphorus pentoxide.

Table 3 Chemical composition and properties of an [Acric](#) Ferralsol developed from [nepheline basalt](#) at Kaua'i island (Kapaa series: very fine, sesquic, isohyperthermic, Anionic Acrudox; (USDA-SCS, 1978)).

Depth (cm)												Chemical properties									
	Chemical composition (% of whole soil)											Exchangeable bases (cmol kg ⁻¹)								pH (1:5)	
												C (%)	N (%)	Ca	Mg	Na	K	CEC (cmol kg ⁻¹)	H ₂ O	1 mol l ⁻¹ KCl	
	SiO ₂	TiO ₂	Al ₂ O ₃	Fe ₂ O ₃	MnO ₂	MgO	CaO	Na ₂ O	K ₂ O	P ₂ O ₅	LOI										
0–30	6.9	7.6	24.6	38.5	0.06	0.32	–	0.03	0.28	0.63	21.7	3.92	0.22	0.7	0.4	0.20	0.40	15.9	5.0	4.5	
30–40	4.4	7.2	28.1	39.3	0.09	0.25	–	0.03	0.11	0.55	19.9	1.46	0.07	0.2	0.0	0.10	0.20	3.7	5.5	5.7	
40–63	4.1	8.1	28.8	38.5	0.08	0.08	–	0.03	0.07	0.58	19.7	1.09	0.04	0.4	0.1	0.20	0.10	2.6	5.5	5.7	
63–90	4.3	7.5	32.9	34.3	0.17	0.13	–	0.03	0.07	0.63	20.0	0.64	0.02	0.1	0.1	0.20	0.10	1.9	5.8	5.7	
90–123	7.1	7.5	30.9	35.2	0.15	0.11	–	0.03	0.07	0.57	18.9	0.46	0.02	0.8	0.1	0.20	0.10	3.7	5.3	5.4	
123–150	13.2	8.4	24.8	36.8	0.15	0.38	–	0.03	0.07	0.61	15.5	0.48	0.02	1.0	0.1	0.30	0.10	6.1	5.5	4.9	

LOI, loss on ignition; CEC, [cation exchange capacity](#).

1 mol l⁻¹ KCl pH in the ferralic or oxic horizons midway in the profile. In the ferralic or oxic horizons the pH measured in 1 mol l⁻¹ KCl is higher than that measured in water, indicating that the net charge of the soil material is positive. Soils with these characteristics are infertile, but can be made productive with phosphorus fertilization, as well as liming with [calcium silicate](#) rather than [calcium carbonate](#) (De Souza et al., 2008; Montalvo et al., 2015).

Ultisols or Acrisols

Ultisols or Acrisols and Oxisols or Ferralsols have developed synchronously on the same geomorphic surface on Kauai in the Hawaiian Islands. The Ultisols, however, always occur, where [soil creep](#) activates and mobilizes clay along sheer planes in the [subsoil](#) (Beinroth et al., 1974). The chemical composition and properties of such an Ultisol are given in [Table 4](#). The same shearing action leads to the formation of Alfisols in drier locations, where leaching is less intense and basic cations are retained in the profile.

Andisols or andosols of volcanic origin

If one had to choose a soil to represent the class of soils we call volcanic, it would be an Andisol/Andosol, a soil with andic properties. Those soils have two characteristics, which make them unique. They contain large amounts of short-range-order and non-crystalline materials, and have low bulk densities. Two short-range-order minerals commonly found in highly leached and weathered soils from pyroclastic parent materials are [allophane](#) and [imogolite](#) (Bleeker and Parfitt, 1974; Parfitt, 1975; Shoji et al., 1982). These minerals have extremely high specific surfaces (approximately equal to 1000 m² g⁻¹) and pH-dependent charge. The high specific surface of the inorganic colloids enables the soil to sequester large quantities of organic matter. The prefix 'Ando' is the Japanese word for 'dark,' used to describe the low chroma of organic-rich [volcanic ash soils](#) of Japan. In Japan, where the ash is generally high in silica and low in iron, organic matter imparts a nearly black color to the soil (Shoji et al., 1982; Hatano et al., 2021).

The chemical composition and selected chemical properties of two Andosols, one developed in a perudic (precipitation >> evapotranspiration; water saturated soil conditions long enough to cause oxygen deficiency), and the other in a [ustic](#) (seasonally dry, but moist during the vegetation period) moisture regime, are given in [Tables 5 and 6](#). Both soils are excellent agricultural soils; the wetter soil requires careful nutrient management and the drier one requires irrigation for profitable farming. However, the low bulk densities of Andosols may lead to severe soil erosion under intensive land-use ([Fig. 3](#)).

Andosols of non-volcanic origin

Since the late 1950s an increasing number of soils are described as having both andic and spodic or podzolic properties developed in non-volcanic parent materials ([Schönhals, 1957](#)). An overview is given in [Bäumler et al. \(2005\)](#). Based on their physico-chemical properties they were named 'Lockerbraunerden' (very friable Cambisols), Cryptopodzols, or non-volcanic or non-allophanic Andosols because they could not be classified properly ([Bargon, 1960; Blaser et al., 1997; Kleber et al., 2003; Dümig et al., 2008; Iamarino and Terribile, 2008; Mileti et al., 2017; Terribile et al., 2018](#)).

Common facts at all sites described so far, are.

- (i) acid parent materials,
- (ii) leaching environment,
- (iii) comparably large amounts of oxidic Fe- and Al-compounds in the soil matrix, and high Fe- and Al-contents in the parent materials,
- (iv) large contents of soil organic matter throughout the soil matrix,
- (v) a lack of short-range-order clay minerals (non-allophanic),
- (vi) extremely low bulk densities, partly <0.5 g cm⁻³ despite clay contents of >50%,
- (vii) high porosity and strong P-retention.

In general, non-volcanic andic soils are also characterized by both andic and spodic/podzolic properties, but without visible eluvial and illuvial horizons ([Blaser et al., 1997; Bäumler et al., 2005](#)). They lack coarse fragments indicating possible eolian deposits or a considerable degree of soil development, or both. Typical features are thixotropy of A and B horizons, as well as extremely stable (micro-)aggregates resembling a pseudosand-like structure. These aggregates were highly resistant to conventional dispersion treatments before particle-size analysis ([Bäumler et al., 2005](#)). The soil properties were specified as silandic, aluandic, transitional alu-sil-andic, and recently as ferro-andic ([Bäumler et al., 2005](#)). Most sites are located in undisturbed mountain environments with native forests or non-intensive land-use. Non-volcanic eolian additions mixed up with the in-situ weathered soil material could be identified for some of the soils described ([Mileti et al., 2017](#)). Radiocarbon dating of the soil organic matter at 1.5-m depth of the soil shown in [Fig. 4](#) gave ages of more than 15,000 years despite recent rooting, leaching of dissolved organic matter, and biodecay ([Bäumler et al., 2005](#)). Age calculations based on soil development indices provided ages of 60–70 ka at the Bhutan sites ([Fig. 4](#)), and up to more than 200 ka for a site in the Khumbu Himal, eastern Nepal ([Dorji, 2016](#)).

Table 4 Chemical composition and properties of a Humic Acrisol developed from basalt (Haiku Series, island of Maui: very fine, isohyperthermic, Ustic Palehumult; (USDA-SCS, 1978)).

Depth (cm)	Chemical composition (% of whole soil)											Chemical properties								
												Exchangeable bases (cmol kg ⁻¹)							pH (1:5)	
	SiO ₂	TiO ₂	Al ₂ O ₃	Fe ₂ O ₃	MnO ₂	MgO	CaO	Na ₂ O	K ₂ O	P ₂ O ₅	LOI	C (%)	N (%)	Ca	Mg	Na	K	CEC (cmol kg ⁻¹)	H ₂ O	1 mol l ⁻¹ KCl
0–18	16.8	11.6	7.6	46.3	0.15	0.97	–	0.06	1.15	0.55	15.9	3.08	0.26	–	0.6	0.1	0.3	15.9	5.1	4.1
18–33	16.8	11.2	11.1	43.0	0.12	0.77	–	0.07	1.23	0.51	15.5	2.79	0.23	–	0.6	0.1	0.2	14.3	5.0	4.0
33–45	14.1	7.9	15.2	42.7	0.07	0.66	–	–	0.99	0.57	17.6	1.97	0.16	0.2	0.8	0.2	0.1	15.5	4.9	4.1
45–70	11.9	6.5	23.2	36.9	0.09	0.78	–	0.02	0.65	0.52	19.1	1.78	0.13	0.6	0.7	0.4	0.1	12.2	5.2	4.4
70–98	18.9	5.2	27.4	28.4	0.12	0.98	–	–	0.31	0.39	18.1	1.08	0.08	1.0	0.8	1.0	0.1	12.7	5.1	4.0
98–155	17.3	5.7	30.1	26.4	0.12	1.62	–	–	0.11	0.42	17.6	0.91	0.06	0.6	0.2	0.8	0.1	12.0	5.0	4.0
155–175	17.0	5.2	32.9	24.7	0.11	1.54	–	–	–	0.32	18.0	0.74	0.04	0.4	0.3	>0.8	0.1	12.4	4.9	4.0

LOI, loss on ignition; CEC, cation exchange capacity.

Table 5 Chemical composition and properties of a Eutric Andosol developed from basaltic ash (Waikaloa series, island of Hawai'i: medial, amorphic, isothermic, Typic Haplotorrands; (USDA-SCS, 1978)).

Depth (cm)												Chemical properties								
	Chemical composition (% of whole soil)											Exchangeable bases (cmol kg ⁻¹)							pH (1:5)	
												C (%)	N (%)	Ca	Mg	Na	K	CEC (cmol kg ⁻¹)	H ₂ O	1 mol l ⁻¹ KCl
	SiO ₂	TiO ₂	Al ₂ O ₃	Fe ₂ O ₃	MnO ₂	MgO	CaO	Na ₂ O	K ₂ O	P ₂ O ₅	LOI									
0–13	32.3	3.7	21.1	14.5	0.44	3.30	3.44	1.76	0.80	0.63	18.3	7.53	0.32	27.9	7.8	0.40	4.80	81.0	6.6	5.7
13–25	32.9	3.5	23.9	15.0	0.42	2.86	3.55	2.24	0.87	0.61	14.7	3.69	0.16	28.0	6.4	0.40	5.60	78.0	7.1	6.2
25–50	36.1	3.2	24.6	15.4	0.43	2.03	2.45	2.26	0.88	0.44	11.7	1.88	0.11	32.6	7.1	0.80	4.80	88.0	7.3	6.2
50–63	38.6	3.4	24.0	15.1	0.47	2.07	3.44	2.25	0.82	0.38	10.0	1.02	0.09	34.6	9.8	1.10	3.00	97.0	7.5	6.3
63–78	40.4	4.3	22.5	14.8	0.43	2.04	3.17	2.41	0.92	0.36	9.2	0.77	–	33.3	12.8	1.60	0.60	93.0	7.6	6.4
78–98	44.8	3.2	20.9	13.9	0.40	2.08	2.11	2.42	1.42	0.26	8.6	0.31	–	33.0	15.9	2.20	0.30	>100	7.8	6.5
98–125	51.9	2.5	20.4	6.7	0.29	1.00	2.61	4.20	2.70	0.09	8.1	0.15	–	34.0	12.4	5.40	0.50	>100	8.1	6.8
125–163	55.7	1.6	21.1	4.8	0.31	0.77	1.68	4.19	2.66	0.09	7.6	0.11	–	33.3	14.4	6.10	1.30	>100	8.2	6.9

LOI, loss on ignition; CEC, cation exchange capacity.

Table 6 Chemical composition and properties of an Acroxic Andosol developed from basaltic ash (Hilo series, island of Hawai'i: medial over hydrous, ferrihydritic, isohyperthermic Acrudoxic Hydrudands; (USDA-SCS, 1978)).

Depth (cm)	Chemical composition (% of whole soil)											Chemical properties								
												Exchangeable bases (cmol kg ⁻¹)							pH (1:5)	
	SiO ₂	TiO ₂	Al ₂ O ₃	Fe ₂ O ₃	MnO ₂	MgO	CaO	Na ₂ O	K ₂ O	P ₂ O ₅	LOI	C (%)	N (%)	Ca	Mg	Na	K	CEC (cmol kg ⁻¹)	H ₂ O	1 mol l ⁻¹ KCl
0–40	12.90	5.24	24.32	27.56	0.15	–	–	–	–	–	27.30	5.30	0.41	2.0	1.8	0.1	0.1	67.6	5.8	5.6
40–53	8.78	5.58	34.40	26.26	0.25	–	–	–	–	–	25.82	3.06	0.22	2.2	0.6	–	0.1	68.4	6.1	6.2
53–65	8.58	5.13	35.92	26.10	0.34	–	–	–	–	–	24.90	2.61	0.17	2.1	0.3	–	–	60.4	6.3	6.5
65–80	8.54	5.30	35.04	26.62	0.37	–	–	–	–	–	25.50	2.67	0.19	1.2	0.6	–	–	62.9	6.3	6.5
80–123	10.82	5.38	33.78	26.38	0.36	–	–	–	–	–	24.46	2.80	0.20	1.8	0.4	–	–	60.5	6.3	6.4
123–140	10.73	5.60	32.56	25.97	0.25	–	–	–	–	–	24.13	2.57	0.16	1.8	0.5	–	–	65.5	6.4	6.4
140–168	10.68	5.68	30.04	28.95	0.25	–	–	–	–	–	24.43	2.24	0.15	2.5	0.5	–	–	71.0	6.3	6.4

LOI, loss on ignition; CEC, cation exchange capacity.



Fig. 3 Extreme erosion of Andosols developed from pyroclastic ejecta in Southern Iceland. Remnants of the old soil and land surface can be seen in the upper right. Photograph by R. Bäumler.



2.5-0 cm. O. Partially fragmented +decomposed litter.

0-17 cm. A. Very dark greyish brown (2.5Y3/2) clay, fine and medium granular to subangular blocky, 52 % clay, 0.49 g cm⁻³, 11.7 % Corg.

17-24 cm. AB. Olive brown (2.5Y4/4) silty clay, medium granular to subangular blocky, 50 % clay, 0.59 g cm⁻³, 5.8 % Corg.

24-49 cm. B1. Brownish yellow (10YR6/8) silty clay loam, weak discontinuous clay skins, 38 % clay, 0.44 g cm⁻³, 2.3 % Corg.

49-101 cm. B2. Yellowish brown (10YR5/8) clay loam, weak discontinuous clay skins, 28 % clay, 0.69 g cm⁻³, 2.2 % Corg.

101-148 cm. B3. Yellowish brown (10YR5/8) silty clay loam, coarse subangular blocky, no stones up to here, 37 % clay, 0.84 g cm⁻³, 1.2 % Corg.

148-200 cm. CB. Brownish yellow (10YR6/8) sandy clay loam, massive to granular, abundant stones, 22 % clay, 1.24 g cm⁻³, 0.5 % Corg.

200+cm. C. Mica schist, slightly weathered, massive.

Fig. 4 Non-volcanic andic soil from Central Bhutan: 90°44'E, 27°32.2'N; 3025 m asl; 7.6 °C annual temperature, 1100 mm annual precipitation; mica schist parent material. Photograph: R. Bäumler.

The well-established international classification systems address andic soils as commonly dark colored soils of volcanic areas typically developed from pyroclastic materials, and characterized by a moderate extent of soil development containing short-range-order minerals and or organo-metallic complexes. Andic properties from non-volcanic materials are mentioned as occurring marginally (USDA-SCS, 1999; IUSS Working Group WRB, 2015). Researchers working on non-volcanic andic types of soil commonly agree that their distribution is widely underestimated, however (for example Fig. 5 shows the probability of their occurrence on the southern slopes of the Himalayas). At present their classification across the globe is unsatisfactory. A re-definition of their classification is proposed towards andosolization as the dominant soil forming process (in non-volcanic mountain areas or undisturbed ecosystems) independently of the origin of the parent materials, or even a new soil type has been suggested that clearly separates these from Andosols or Andisols *sensu strictu* (Kleber et al., 2003; Bäumler et al., 2005; Iamarino and Terribile, 2008).

There are few options to classify non-volcanic andic soils at present, as spodic horizons may have andic properties, but Andosols should not have a spodic horizon (IUSS Working Group WRB, 2015). Non-volcanic andic soils with spodic horizons are therefore classified as Podzols/Spodosols, because of the presence of a spodic horizon (unless buried deeper than 50 cm) is given preference over the occurrence of andic properties. Non-volcanic andic soils missing a single spodic criterion have to be classified as Andosols/Andisols.

Questions remain at present about the role of non-volcanic eolian additions (Mileti et al., 2017), and or multiple freeze-thaw cycles inducing at least part of the specific properties. Moreover, it is debated that andosolization could be considered a common soil forming process independent of the parent material, but considerably endangered by erosion, and therefore not well preserved

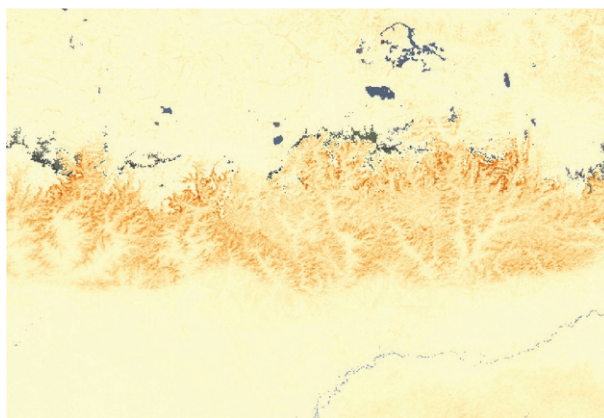


Fig. 5 Occurrence of alu-andic Andosols on the (non-volcanic) southern slopes of the Himalayas (Nepal, Sikkim/India, Bhutan, Tibet/China) between 2000 and 3500 m asl. On the map, the darker the orange color, the higher is the probability of occurrence (SoilGrids 2.0: Poggio et al., 2021).

(McDaniel et al., 2018). And in addition, almost all research papers indicate an under-estimated role of polyvalent Fe ions and/or oxidic iron compounds for ferro-andic properties (Bäumler et al., 2005).

Human interactions with volcanic soils

Volcanic soils may have special meaning to people living in humid regions where the soil would be less fertile than under normal circumstances, but are maintained by periodic additions of nutrient-rich ash from a neighboring volcano (Lubis et al., 2021). The advantage of having an adequate supply of water and nutrients in contrast to having one or the other was, for our ancestors, the difference between being able to form creative, sedentary communities in contrast to simply surviving by hunting and gathering. Thus, it may not be the soil itself but the presence of an active volcano nearby that makes volcanic soils what they are.

Our human ancestors may have avoided extinction and evolved by adapting to the diverse and ever-changing volcanic environments. The soils rich in mineral nutrients, constantly rejuvenated by sprinklings of volcanic ash support large herds of herbivores and their predators, and some of the highest nonurban human population densities (Small and Naumann, 2001).

When modern humans left their native habitats for any reason, then crossed Asia and the Bering Straits, they encountered the North, Central, and South American segment of the 'Ring of Fire' that circles the Pacific. Like the rich [alluvial soils](#) of the Nile, Tigris, Euphrates, and Indus rivers that enabled humans to shift their energies from hunting and gathering to creative endeavors, the fertile and well-watered [volcanic soils](#) all around the Ring of Fire provided new settlers with the means to do the same (Willerslev and Meltzer, 2021).

Even today, the influence of volcanic soils on human society is evident. In Indonesia, for example, nearly half of its population live on the volcanic islands of Indonesia. Efforts to move whole communities from densely populated areas to the larger and less densely populated neighboring islands have been difficult. Villagers accustomed to farming [Andosols](#) with characteristics similar to the one shown in [Table 5](#) found it difficult to survive on the less fertile [Ultisols](#) of other islands with features similar to those shown in [Table 4](#).

Today [science and agro-technology](#) enable farmers to understand and overcome soil constraints, but only a century ago [soil fertility](#) determined to a large degree the outcomes of human interactions with the environment. This is still true in many developing countries, where science-based soil management practices have yet to be adapted. Volcanic soils of Africa, Asia, and Latin America that helped people in the past now need new ideas and technologies to perform and contribute to human well-being in a fast changing environment.

Conclusions

Soils developed from volcanic parent materials are found all over the world in present or past tectonically and/or volcanically active areas. The story of their pedodiversity is recorded by the chemical composition, the mineral structure and/or the surface area of the volcanic parent materials. The specific soil properties are further explained by the period of soil development, respective age of the soils, and the specific site conditions. The site conditions are in turn controlled by the climate, the relief, the volcanic or tectonic activity, and human impact. Volcanic soils are in general very fertile. Land-use intensity is high provided that the moisture conditions are not too wet or too dry, and as long as the relief is not too high. Volcanic soils are often severely endangered on steep slopes by thixotropic features inducing landslides, especially in Andosols developed from volcanic ash. A focus of future research should be on soil conservation measures against erosion at sites of intensive agriculture. Moreover, the processes of soil formation and the development of andic properties from non-volcanic parent materials are yet not fully understood.

References

- Arnalds O (2004) Volcanic soils of Iceland. *Catena* 56(1-3): 3–20.
- Bargon E (1960) Über die Entwicklung von Lockerbraunerden aus Solifluktionsmaterial im vorderen Odenwald. *Zeitschrift für Pflanzenernährung und Bodenkunde* 90: 229–243.
- Bäumler R, Caspari T, Totsche KU, Dorji T, Norbu C, and Baillie IC (2005) Andic properties in soils developed from non-volcanic materials in Central Bhutan. *Journal of Plant Nutrition and Soil Science* 168: 703–713.
- Beinroth FH, Uehara G, and Ikawa H (1974) Geomorphic relationships of Oxisols and Ultisols on Kauai, Hawaii. *Soil Science Society of America Proceedings* 38: 128–131.
- Blaser P, Kernebeek P, Tebbens L, Van Breemen N, and Luster J (1997) Cryptopodzolic soils in Switzerland. *European Journal of Soil Science* 48: 411–423.
- Bleeker P and Parfitt RL (1974) Volcanic ash and its clay mineralogy at Cape Hoskins, New Britain, Papua New Guinea. *Geoderma* 11: 123–135.
- Bonatzotky T, Ottner F, Erlendsson E, and Gísladóttir G (2019) The weathering of volcanic tephra and how they impact histosol development. An example from South East Iceland. *Catena* 172: 634–646.
- Bonatzotky T, Ottner F, Erlendsson E, and Gísladóttir G (2021) Weathering of tephra and the formation of pedogenic minerals in young Andosols, South East Iceland. *Catena* 198. <https://doi.org/10.1016/j.catena.2020.105030>.
- Candra IN, Gerzabek MH, Ottner F, Wriessnig K, Tintner J, Schmidt G, Rechberger MV, Rampazzo N, and Zehetner F (2021) Soil development and mineral transformations along a one-million-year chronosequence on the Galápagos Islands. *Soil Science Society of America Journal* 85(6): 2077–2099.
- Christiansen RL and Peterson DW (1981) Chronology of the 1980 eruption activity. In: Lipman PW and Mullineaux DR (eds.) *The 1980 eruptions of Mount St. Helens*, 1st edn, pp. 17–30. Washington, DC: US Government Printing Office. Geological Survey Professional Paper 1250.
- De Souza RF, Faquin V, Carvalho R, Torres PRF, and Pozza AAA (2008) Soil chemical properties influenced by the substitution of calcium carbonate by calcium silicate. *Revista Brasileira de Ciência do Solo* 32(4): 1563–1572.
- Dethier DP, Pevear DR, and Frank D (1981) Alteration of new volcanic deposits. In: Lipman PW and Mullineaux DR (eds.) *The 1980 Eruptions of Mount St. Helens*, pp. 649–665. Washington, DC: Geological Survey Professional Paper 1250.
- Dorji KD (2016) *Soils as proxies of environmental fluctuations at the southern slopes of the Bhutan Himalayas*. Dissertation Friedrich-Alexander-Universität Erlangen-Nürnberg. urn:nbn:de:hbz:29-opus4-72774.
- Dümig A, Schad P, Kohok M, Beyerlein P, Schwimmer W, and Kögel-Knabner I (2008) A mosaic of nonallophanic Andosols, Umbrisols and Cambisols on rhyodacite in the southern Brazilian highlands. *Geoderma* 145(1-2): 158–173.
- Hatano R, Shinjo H, and Takata Y (eds.) (2021) *The Soils of Japan*. In: *World Soils Book Series*. Singapore: Springer Nature.
- Iamarino M and Terribile F (2008) The importance of andic soils in mountain ecosystems: a pedological investigation in Italy. *European Journal of Soil Science* 59: 1284–1292.
- IUSS Working Group WRB (2015) World Reference Base for Soil Resources 2014, update 2015. In: International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports, 106, Rome: Food and Agriculture Organisation of the United Nations.
- Kleber M, Zikeli S, Kastler M, and Jahn R (2003) An andosol from eastern Saxony, Germany. *Journal of Plant Nutrition and Soil Science* 166: 533–542.
- Kuzera K, Rogan J, and Ronald Eastman J (2005) Monitoring vegetation regeneration and deforestation using change vector analysis: Mt. St. Helens study area. In: *American Society for Photogrammetry and Remote Sensing - Annual Conference 2005 - Geospatial Goes Global: From Your Neighborhood to the Whole Planet*, 1 588–595.
- Lowe DJ (1986) Controls on the rate of weathering and clay mineral genesis in airfall tephra: A review and New Zealand case study. In: Coleman SM and Dethier DP (eds.) *Rates of Chemical Weathering of Rocks and Minerals*, pp. 265–330. New York: Academic Press.
- Lubis RL, Juniarti, Rajmi SL, Armer AN, Hidayat FR, Zulhakim H, Yulanda N, Syukri IF, and Fiantis D (2021) Chemical properties of volcanic soil after 10 years of the eruption of Mt. Sinabung (North Sumatera, Indonesia). *IOP Conference Series: Earth and Environmental Science* 757(1). <https://doi.org/10.1088/1755-1315/757/1/012043>. art. no. 012043.
- McDaniel P, Ross M, Jimenez J, Strawn DG, Valerio M, Kimsey MJ, Campbell S, and Falen A (2018) Pedogenic pathways in andic soils of the northern Rocky Mountains (USA). *Soil Science Society of America Journal* 82: 1308–1318.
- Mileti FA, Vingiani S, Manna P, Langella G, and Terribile F (2017) An integrated approach to studying the genesis of andic soils in Italian non-volcanic mountain ecosystems. *Catena* 159: 35–50.
- Montalvo D, Degryse F, and McLaughlin MJ (2015) Agronomic effectiveness of granular and fluid phosphorus fertilizers in andisols and oxisols. *Soil Science Society of America Journal* 79(2): 577–584.
- Nockolds SR (1954) Average chemical compositions of some igneous rocks. *Bulletin of the Geological Society of America* 65: 1007–1032.
- Parfitt RL (1975) Clay minerals in recent volcanic ash soils from Papua New Guinea. *Bulletin of the Royal Society of New Zealand* 13: 241–245.
- Poggio L, de Sousa LM, Batjes NH, Heuvelink GBM, Kempen B, Ribeiro E, and Rossiter D (2021) SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *Soil* 7: 217–240.
- Saigusa M, Shoji S, and Kato T (1978) Origin and nature of halloysite in Ando soils from Towada tephra, Japan. *Geoderma* 20: 115–129.
- Schönhals E (1957) Spätglaziale äolische Ablagerungen in einigen Mittelgebirgen Hessens. *Eiszeitalter u. Gegenwart* 8: 5–17.
- Shoji S, Fujiwara Y, Yamada I, and Saigusa M (1982) Chemistry and clay mineralogy of Ando soils, Brown forest soils, and Podzolic soils formed from recent Towada ashes, northeastern Japan. *Soil Science* 133: 69–86.
- Small C and Naumann T (2001) The global distribution of human population and recent volcanism. *Environmental Hazards* 3(3): 93–109.
- Stevens KF and Vucetich CG (1984) Pedogenic weathering of Upper Quaternary tephra in New Zealand. Isovolumetric geochemical evidence of cation movement. *Chemical Geology* 47: 285–302.
- Terribile F, Iamarino M, Langella G, Manna P, Mileti FA, Vingiani S, and Basile A (2018) The hidden ecological resource of andic soils in mountain ecosystems: Evidence from Italy. *Solid Earth* 9(1): 63–74.
- USDA-SCS (1978) *Soil Survey Laboratory Data and Description for Some Soils of Hawaii*. Soil Survey Investigations Report No. 29. Washington, DC: USDA Soil Conservation Service/Hawaii Agricultural Experimental Station/Hawaiian Sugar Planters' Association.
- USDA-SCS (1999) *Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys*, 2nd edn. Washington, DC: Natural Resource Conservation Service.
- Vilmundardóttir OK, Gísladóttir G, and Lal R (2015) Between ice and ocean; soil development along an age chronosequence formed by the retreating Breiðamerkjökull glacier, SE-Iceland. *Geoderma* 259: 310–320.
- Willerslev E and Meltzer DJ (2021) Peopling of the Americas as inferred from ancient genomics. *Nature* 594: 356–364.

PEDOMETRICS

Overview of Pedometrics

Alexandre MJ-C Wadoux^a, Inakwu OA Odeh^{b,*†}, and Alex B McBratney^a, ^aSydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Camperdown, NSW, Australia; ^bUniversity of Sydney, Sydney, NSW, Australia

© 2023 Elsevier Ltd. All rights reserved.

*Formerly.

†Deceased.

Introduction	471
Methods and problems	472
Pedometrics: Principles and hypotheses	472
Soil measurements and properties	473
Field properties: Hard and fuzzy descriptors?	473
Soil spectroscopy	473
Pedotransfer functions	474
Quantitative soil geography	474
Soil classification and soil map units	474
Fuzzy classes and the breakdown of the classification paradigm	476
Pedodiversity	476
Space and time prediction	477
The clorpt approach	477
The scorpan approach	477
Geostatistics	478
The hybrid of scorpan and geostatistics	478
Machine learning	479
Space-time prediction	479
Sampling design	481
Soil utility and quality	481
Land evaluation	481
Soil functions	481
Quantitative pedogenesis	483
Soil processes	483
Soil-landscape evolution	483
Conclusion and outlook	485
References	485

Abstract

Pedometrics is concerned with the application of mathematical and statistical methods to the study of the distribution and genesis of soils. Here, we describe the main areas that pedometric research addresses: distribution of the soil pattern in character space, spatial and spatio-temporal soil variation, quantitative evaluation of the utility and quality of soil, and quantitative pedogenesis. To these main areas akin to the problems of pedology, pedometrics considers and represents uncertainty. Pedometric research is undeniably in an expansion phase and has now many areas of application at the interface with many questions relevant to the sustainable growth of our societies.

Key points

- Succinctly describes the four main areas of pedometrics.
- Pedometric research can contribute to many areas of soil science.

Introduction

The term 'pedometrics', first coined by McBratney in the late 1980s, 'is a neologism derived from the Greek roots pedos or πεδος (soil) and metron or μετρον (measurement)'. It is used in a similar fashion to other words such as biometrics, psychometrics,

Table 1 Types of quantitative models used for soil studies with examples and sources of uncertainty.

Causality and/or uncertainty	Model		
	Deterministic	Stochastic	Nonstatistical
Empirical	Generalized linear models, numerical taxonomy, Jenny functional relations and canonical ordination	Time series, spatial processes, temporal and spatial variation of soil properties, Markov chain models	Fuzzy systems, Machine learning
Mechanistic	Flow and diffusion of soil plastic materials, profile and landscape development	Indeterminate	Indeterminate
Uncertainty	Imprecise measurements, model uncertainty	Random process, probability	Vagueness, ambiguity, and fuzzy geostatistics

Courtesy AB McBratney and CJ Moran.

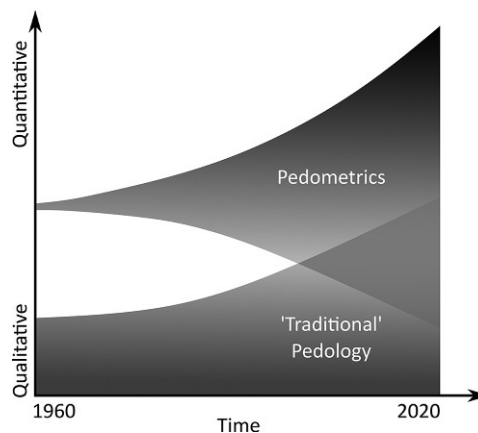


Fig. 1 A timeline of the growth of pedology and pedometrics. Modified from McBratney AB, Odeh IOA, Bishop TF, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. *Geoderma* **97**: 293–327.

econometrics, chemometrics, and, the oldest of all, geometry. Furthermore the term covers two main ideas—the ‘soil’ or ‘pedo’ part, which corresponds to that branch of soil science we call pedology, and the ‘metrics’ part, which is restricted to mathematical and statistical methods. Simply defined, pedometrics is the use of quantitative methods for the study of soil distribution and genesis, and as a sustainable resource. Another problem-oriented definition is ‘soil science under uncertainty.’ In this sense pedometrics deals with uncertainty in soil models that describe deterministic or stochastic variation, vagueness, and lack of knowledge of soil properties and processes (Table 1). Thus, mathematical, statistical and numerical methods could be applied to resolve the uncertainty and complexity inherent in a soil system model, including numerical approaches to classification, which deals with supposedly deterministic variation (Webster, 1994).

Pedometrics is not new, as mathematical and statistical methods have been applied to soil studies since the 1900s (statistical research by R. Fisher at Rothamsted) and more intensively since the 1960s and 1970s. However, it is now a technical branch of soil science complementing traditional pedology. Over time the use of computers has increased in both fields, and the difference between the two has decreased and, in some cases, overlapped. Due to new demands for quantitative soil information required for global-scale models, regional environmental planning, and field-scale agricultural land management, traditional pedology has become more quantitative through the increased use of digital soil data in computerized soil information systems. Concurrently, pedometrics has emerged as a collection of quantitative tools, which are increasingly being used to account for conceptual pedological models of soil variation. As of 2021, there is a strong and growing overlap and synthesis between ‘traditional’ pedology and quantitative pedology or pedometrics (as shown in Fig. 1).

Methods and problems

Pedometrics: Principles and hypotheses

As the definition implies, pedometrics encompasses the whole quantitative approach to the study and description of soil. It also involves the use of mechanistic, stochastic, deterministic, and empirical models (Table 1). A more restricted form of pedometrics is one that involves applications of statistical and probabilistic modelling to soil variation. Pedometric principles are therefore embodied in (geo-) statistical modelling of pedological processes and the associated phenomena as they influence soil variation, and the associated model uncertainty.

Pedometrics is therefore aimed at resolving or testing a number of questions or hypotheses (McBratney and Lark, 2018):

1. What is the pattern of soil distribution in character space, i.e. soil taxonomy?
2. How does the soil vary spatially and temporally?
3. What are the utility and quality of soil?
4. What are the causal factors of soil formation, i.e., soil genesis?
5. What is the uncertainty of modelling the spatial pattern of soils, i.e., is a model an accurate representation of reality?

With the exception of (5), these hypotheses dovetail well with the requirements of conventional pedology. Although developments in pedometrics in the past were restricted to statistics and probability, increasingly the utility of soil is the subject of pedometric research and mechanistic models are being applied and developed to model soil processes.

Soil measurements and properties

To model soil quantitatively, we need to perceive it, i.e., sense it or measure some quality or quantity of it. In other words, how do we represent soil properties numerically to model their variation mathematically and/or statistically? Ever since the advent of modern environmental science, technological advances have improved immensely the way we characterize or measure the quality and/or quantity attributes of soil. Soil measurement techniques include direct measurements, visual observations, and remote sensing (including proximal, airborne, and space-based modes of sensing).

Field properties: Hard and fuzzy descriptors?

Other than color, many soil morphological properties, required in routine soil surveys, are described in somewhat vague terms. For example, soil structure grade is described as 'structureless' or 'weak' or 'moderate' or 'strong' in many national soil survey handbooks. To make the grades more amenable to quantitative analysis (e.g., multivariate statistical analysis) these terms are usually coded as 0, 1, 2, and 3, respectively. But because of the vagueness in the linguistic characterization, fuzzy coding, which accounts for uncertainty or lack of clear boundary between the grades, has been used. Other morphological descriptors such as soil aggregate (structure) size classes, e.g., 'fine,' 'medium,' and 'coarse,' may be given linguistic variables which could be fuzzified to determine the degree of truth that a given soil layer is characterized by. Simple examples of fuzzy coding are presented in Table 2.

Soil spectroscopy

Measurement of soil properties has gained tremendously from advances in technology, particularly in spectroscopy. Soil spectroscopy is loosely defined as the study of the spectral signature of a soil material. The spectral signature of a soil is primarily a function of its mineral and organic composition since the soil is formed from the transformation of rocks, organic matter from plants and organism activity. Thus, a soil type has a unique spectral reflectance curve. Spectroscopic measurements can be proximal or remote, when mounted on a satellite, plane or drone. Proximal measurements might be done in situ on the soil profile using portable field spectrometers or in the laboratory. Because of the large number of soil characteristics that a soil spectrum contains, it is said that spectroscopic measurements are cost-effective and fast, compared to conventional methods of soil analysis (Nocita et al., 2015).

The visible (400–700 nm) and infrared (400–25,000 nm) range of the electromagnetic spectrum is of particular interest to soil scientists. Infrared spectra have encoded information on both organic and inorganic soil material. The mid-infrared range (2500–25,000 nm; 400–4000 cm⁻¹), contains distinct absorption features for soil organic matter and mineralogy. These absorption features have overtones and combinations in the visible and near-infrared (vis-NIR) range (400–2500 nm), which means that soil organic and mineral characteristic are also reflected in the vis-NIR range but with fewer, broader, and more complex (i.e. non-specific) absorption bands. Fig. 2 shows three vis-NIR spectra for soil samples with different compositions.

Table 2 Fuzzy coding of structural type.

Type	Horizontality	Verticality	Flatness	Accommodation
Platy	1.0	0.1	1.0	1.0
Lenticular	1.0	0.3	0.3	1.0
Prismatic	0.2	1.0	1.0	1.0
Columnar	0.2	1.0	0.9	0.9
Angular blocky	1.0	1.0	1.0	1.0
Subangular blocky	0.7	0.7	0.5	0.5
Granular	0.2	0.2	0.0	0.1
Massive	0.0	0.0	0.0	1.0
Single grain	0.0	0.0	0.0	0.0

Reproduced with permission from Odeh IOA, McBratney AB, and Chittleborough DJ (1991) Elucidation of soil-landform interrelationships by canonical ordination analysis. *Geoderma* 49: 1–32.

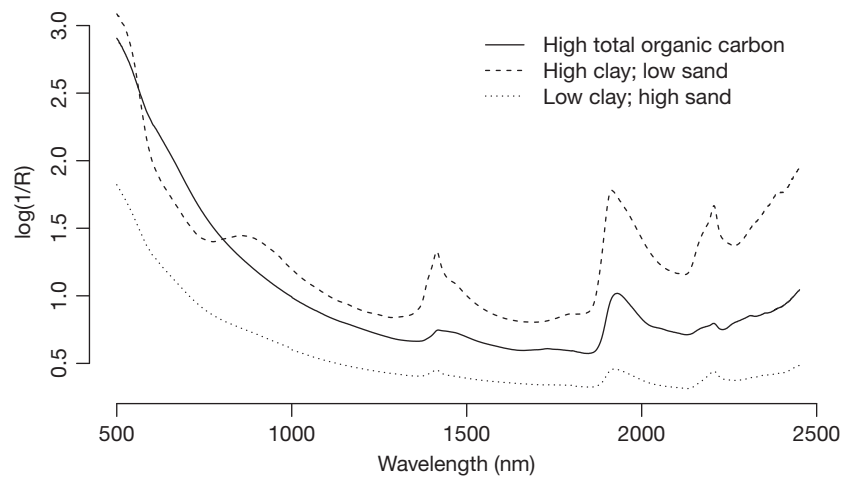


Fig. 2 Vis-NIR spectra of three soil samples from wheatbelt of southern New South Wales, Australia: a soil with 12% total organic carbon and two soil samples with low total organic carbon content (<1%) of which one has 74% clay and 19% sand and the other 5% clay and 91% sand.

Pedometricians have extensively investigated the predictive power of such spectra to replace and complement soil analyses. Mathematical transfer functions are used to correlate the spectral wavelengths (the independent variable) to laboratory-measured values of soil properties (the dependent variables). Relatively simple models can be used, for example linear regression on user-defined wavelengths, but more complex models exist, such as principal component analysis and partial least-squares regression. Both models consider the whole spectra for estimating their parameters and make use of reduction dimensionality (principal component analysis). More recently, pedometricians have also considered machine learning models, with some success. Once calibrated, the model is used to predict the soil properties using spectral information only. These soil property estimates are stored together with the soil spectral information into spectral libraries. Lately, efforts have been made to combine and standardize soil spectra from various instruments into large, regional and global soil spectral libraries. Research is also on-going to fuse multiple sensors and instruments to obtain more precise estimates of soil properties.

Pedotransfer functions

The term ‘pedotransfer function’ (PTF) is generally defined as translating the raw soil data into more useful information. It can also be defined as predictive functions of certain soil properties difficult to obtain from other easily, routinely, or cheaply measured properties (Van Looy et al., 2017). The most readily available data come from soil survey, such as field morphology, texture, structure, and pH. The PTFs add value to the basic soil data by translating them into predictors of other more laborious and expensively determined soil properties. These functions fill the gap between the available soil data and other properties that are more useful or required for a particular model or quality assessment. A simple illustration of how to apply PTFs to practical problems is shown in Fig. 3.

Soil spectral inference systems also rely on PTFs to populate soil databases. Spectral inference systems combine soil spectral data to infer a set of basic soil properties, and PTFs to estimate another set of soil properties more difficult to measure or without spectral response, such as permanent soil wilting point or field capacity. The PTFs are found in the literature and combined into a network structure to create an inference system. One of the main features of soil spectral inference systems is the quantification and propagation of uncertainty, using methods such as model ensemble by bootstrap and aggregation (See chapter on this topic in Encyclopedia).

Quantitative soil geography

Soil classification and soil map units

The pattern of soil in the landscape is as a result of effects of soil-forming factors, namely: climate, organisms, parent material, topography, and time (usually referred to as ‘clorpt,’ but recently extended to ‘scorpan’—see Eq. (2), below). These factors, particularly climate and parent material, produce broad patterns of soil in geographic space. The complex combination of these factors is what causes repetitive patterns of soil in the landscape, which form the basis for soil classification, identification, and mapping. In traditional soil classification, both the conceptual and real classes are qualitative and have no clear-cut boundary. Pedologists have, in the past, attempted to avoid these by applying numerical classification to complex soil data. Soil classification, in both geographical and taxonomic space, is a simple representation of complex, and sometimes repetitive patterns of soil in the landscape. Class identification, on the other hand, is aimed at matching classes that are conceptually described (usually in a classification system) with reality—represented by data obtained by soil morphological description and measurements.

A soil map unit is a collection of areas that are relatively homogeneous in soil constitution or miscellaneous areas that constitute different soil components or both. In a given survey region, a map unit differs in some respect from all others and can be uniquely identified on a soil map. A soil map delineation, on the other hand, is an individual area on the map, bounded by other delineations or the map boundary. Usually a delineation contains the dominant components in the map unit name, but it may

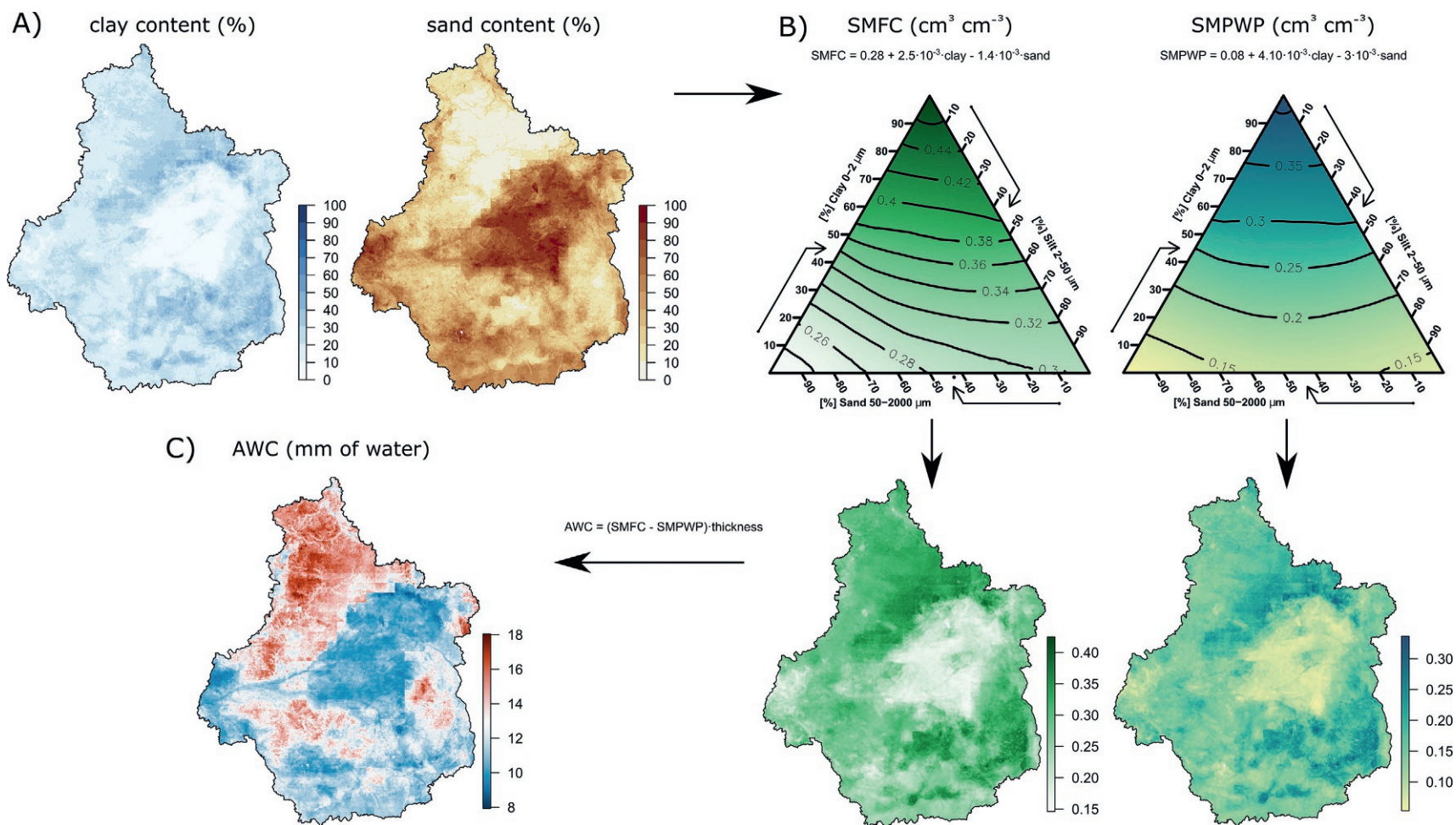


Fig. 3 A simple example of pedotransfer functions (PTFs) for the Centre-Val de Loire region (France): (A) particle-size fractions were interpolated on to a fine grid from a set of point measurements; (B) horizon (5–15 cm) soil water at field capacity (SMFC) and soil moisture content at permanent wilting point (SMPWP) were in turn estimated through a PTF expressed as a function of particles sizes; (C) available water capacity is estimated by multiplying the SMFC subtracted from the SMPWP by the soil layer thickness (in mm), assuming all soil is fine earth. Modified from Román Dobarco, MR, Bourennane, H, Arrouays, D, Saby, NP, Cousin, I, and Martin, MP (2019) Uncertainty assessment of GlobalSoilMap soil available water capacity products: A French case study. *Geoderma* **344**: 14–30.

not always contain a representative portion of each kind of inclusion. Soil boundaries can seldom be shown with complete accuracy on soil maps; hence parts of adjacent polypedons are inadvertently included or excluded from delineations. This problem has been the stimulus for paradigm shifts in soil classification in the latter part of the 1990s; hence the applications of numerical hard and fuzzy classification systems in soil science.

From the 2010s onwards, there was a renewed interest in numerical soil classification, after nearly two decades of meagre activity and progress. Two classification systems, the World Reference Base (WRB) and Soil Taxonomy, are considered global soil classification systems. Both systems, however, also have several shortcomings for local applications. A better communication between regional soil classification systems was attempted, for example by standardization using taxonomic distance. All soil classifications can be evaluated in the multi-dimensional space of the soil data with dimensionality reduction, and by computing centroids for each class of any system. Numerical methods applied to soil classifications have other advantages. For example, they enable the quantification of similarity between soil profile descriptions or between genetic horizons (Hughes et al., 2018).

Fuzzy classes and the breakdown of the classification paradigm

Traditionally most classification systems are composed of mutually exclusive classes in order to conform to discontinuous soil variation embedded in the traditional concept of soil map units. But soil variation is more continuous than discrete. The pioneer work in pedometrics, involving computer-based numerical classification, was designed to address this limitation, among others (McBratney et al., 2000). While the applications of numerical soil classification to soil studies are, to some extent, based on continuous representation of soil in space, their results are still interpreted in terms of 'hard' classes. There is also lack of any spatial coherence for the classes to be mapped, despite some attempts with, for example, spatially weighted classification. Recent advances are based on fuzzy sets to optimize the quality of prediction of the resulting classification, and which take cognizance of the continuous nature of soil variation and nonlinearity in the inter-attribute relationships.

The first application of fuzzy set theory in soil survey was principally for classification. Two different but complementary approaches to grouping individuals into fuzzy classes in soil science are: (1) fuzzy *c*-means (FCM, also known as 'fuzzy *k*-means'), and (2) the semantic import model (SI).

The FCM algorithm for improved predictive classification by providing for membership to an extragrade class is based on the objective function, defining the within-class sum-of-square errors, J_E , expressed as (McBratney and Odeh, 1997):

$$J_E(M, c) = \alpha \sum_{i=1}^n \sum_{j=1}^c m_{ij}^{\varphi} d_{ij}^2 + (1 - \alpha) \sum_{i=1}^n m_{i*} \sum_{j=1}^c d_{ij}^{-2}, \quad (1)$$

where c is the number of classes, n is the number of individuals or pedons; m_{ij} is the membership of an individual i in class j ; φ is the fuzziness exponent ($1 < \varphi < \infty$); d_{ij} is the character space between the feature value of an individual, i , and the feature centroidal value for class j ; α is the parameter that determines the mean value of m_{i*} , which is the membership value of an individual, i , in the extragrade class.

The SI model was developed primarily for land evaluation. The need to use fuzzy set theory in land evaluation, such as that defined in the Food and Agriculture Organization of the UN (FAO) framework, arises because basic soil information used for land evaluation is mostly described by seemingly vague terms such as 'poorly drained,' 'slightly susceptible to soil erosion,' and 'moderate nutrient availability.' Not even when these terms are defined precisely is the qualitative ambiguity removed. Usually, the land evaluator's aim is to produce a set of clearly defined classes of land qualities based on specified land use requirements. These subsequently provide the means of transferring information about the soil and its use. As land qualities are complex attributes that are derived from land characteristics such as topography, soil, water, or biological and human activity, subsequent Boolean logical operations in the process of land evaluation tend to throw away much useful information.

Pedodiversity

Measuring pedodiversity within an area is a way of measuring soil variation, usually using pre-classified soil entities such as taxa or properties. The diversity of pedological entities such as pedotaxa, pedogenetic horizon and soil properties, as well as their spatial and temporal patterns influences several ecosystem functions that the soil provides. A more diverse soil is also more resilient against disturbance and stress caused by unsound soil management practices.

The concept of diversity contains two fundamental components which apply to pedodiversity (Ibáñez and Bockheim, 2013): (i) the richness, i.e., the number of soil types (soil classes) in the area of interest or within a level of the taxonomic system and (ii) the evenness, i.e., the relative number of soil types or classes. Soil scientists have contributed to developing indices of pedodiversity, most of them being similar to the well-known indices of ecological and biological diversity developed by ecologists and biologists. Such indices encompass species richness and abundance models, few of them incorporate the two into a single index, the most popular of which is the Shannon diversity index, defined by:

$$\text{Shannon diversity index} = - \sum_{i=1}^n p_i \times \ln p_i, \quad (2)$$

where p_i is the proportion of the i th soil type relative to the total number of soil types n in the area of interest or level of the taxonomic system. When the area is composed of a single soil type, it equals zero and is unbounded otherwise as the number of soil types increases. When applied to existing soil classification systems, several authors have found clear relationships between the number of soil types (i.e., the pedodiversity) and the size of the area surveyed. Such indices were applied at the global scale, where it was found that some continents are more diverse than others.

Pedometricians contended that taxonomic distance between soil types should be included in the computation of pedodiversity indices because the similarity differs among taxa. These can be included by indices of quadratic and taxonomic entropy.

Space and time prediction

The clorpt approach

As previously stated, the 'clorpt' methods are based on the empirical-deterministic models that originated from Hans Jenny's *Factors of Soil Formation*. Jenny's state-factor equation, in its extended form, is expressed as (McBratney et al., 2003):

$$S = f(s, c, o, r, p, a, n), \quad (3)$$

where S is some soil properties or soil type as a function f of state factors: s as some other soil property at a point, c as climate, o as organisms, r as relief, p as parent material, a as time or age, and n as space or spatial position. Soil spatial variability is therefore considered as being causative realizations of the complex combinations of soil-forming processes as influenced by the soil-forming factors. The scorpan function (Eq. (3))—in its original form, the clorpt—earlier in the twentieth century stimulated numerous studies, mainly quantitative prediction of soil attributes, which we shall treat in a later section. Until recently, because data on soil classes were non-quantitative, the scorpan methods were only restricted to predicting soil attributes measured on a continuous scale.

Models such as that expressed in Eq. (3) can be derived for predicting soil classes or soil types using various techniques, for example logistic regression, classification trees and other machine learning models. Logistic regression is designed specifically for situations in which we have a dichotomous (nominal or ordinal) dependent variable (e.g., soil classes), in comparison to classical linear regression typically used for continuous dependent variables (e.g., soil pH value). Several problems arise when a dependent variable is binary: the error terms are non-normal and their variance is non-constant. Although the errors are not normal, this method still provides unbiased regression estimators that are approximately normal if the sample size is large. Classification tree algorithms and other machine learning models, on the other hand, search for combinations of values of independent variables (e.g., Jenny's state factors or attributes derived from them) that best predict the value of the dependent variable (e.g., unordered factors such as soil classes). Prediction quality is based on criteria such as the within-group noises, variance, or statistical significance of the conditional frequency distribution of values for the dependent variable, conditioned on the answers to questions asked. A digital map of soil suborder classes produced by the classification tree model is shown in Fig. 4.

The scorpan approach

The 'scorpan' approach (Eq. 3) earlier in the twentieth century stimulated numerous studies in quantitatively predicting soil attributes measured on a continuous scale. Many of the earlier studies, and indeed some recent examples, were based on general and bivariate-simple linear regression, although multiple polynomial regression models were occasionally applied. But due to

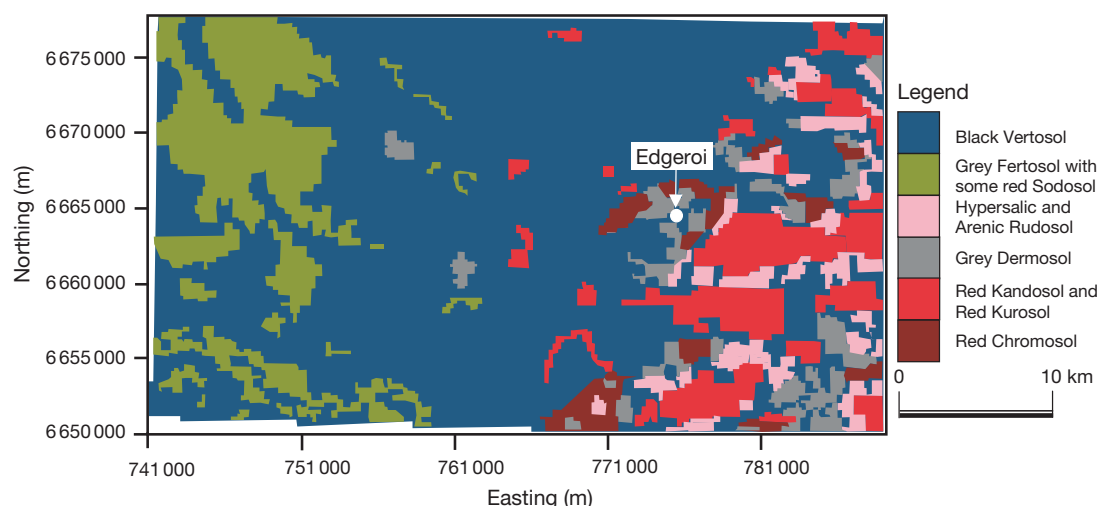


Fig. 4 Predicted soil classes for the Edgeroi area in accordance with Australian Soil Classification. Reproduced with permission from McBratney AB, Odeh IOA, Bishop TF, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. *Geoderma* 97: 293–327.

nonlinearity in the correlations among many soil variables, and indeed of many of the soil variables with ancillary variables, robust methods such as generalized linear models (GLMs), generalized additive models (GAMs), and random forest (RF) have been developed and applied. Another development is the artificial neural network models, which are nonparametric modelling techniques that mimic the neural networks of the brain. The networks are composed of processing units, or neurons, which are organized into layers, i.e., input, hidden, and output layers. We shall describe mapping based on machine learning in the next subsection.

The problem is that while the classic models or the more robust methods may take care of the deterministic relations, they do not account for spatial autocorrelations of soil properties, especially at the local level. To solve this problem, the pioneer pedometricians initiated the application of geostatistics (which was primarily developed for the mining industry).

Geostatistics

Matheronian geostatistics is based on the theory of regionalized variables, which allows us to consider spatial variability of a soil property or even soil types, if quantified, as a realization of a random function represented by a stochastic model. The generic geostatistical method of spatial interpolation is termed 'kriging' in its various forms: simple, ordinary, lognormal, and disjunctive kriging (Webster and Oliver, 2007).

In geostatistics, the variogram is a primary requirement for spatial prediction or kriging of a target geographical feature. The variogram can be obtained by computing the spatial correlation or covariance between pairs of samples at certain distances apart in a data set to produce empirical semivariances. The plot of the semivariance and the corresponding distance or lag at which the pairs are separated produce the variogram. The latter describes the magnitude, spatial scale, and the general form of variation of a given variable (Heuvelink and Webster, 2001). An example of an experimental variogram is shown in Fig. 5, together with several exponential models based on different fitting methods.

The first major applications of kriging, in its ordinary and univariate form, for soil studies were in the early 1980s. Since then ordinary kriging has been widely used in various subfields of soil science, including soil reclamation, soil classification, and soil pollution. Major limitations of the univariate geostatistics technique of kriging are due to the assumptions of stationarity, which are not often met by the field-sampled data sets and, of course, the often-cited requirement of large amounts of data to define the spatial autocorrelation. However, with increasing availability of ancillary information, the lack of adequate samples has been partially solved. The univariate use of kriging is also limiting in situations of complex terrain where the soil-forming processes are themselves complex. In such situations there is the need to model both the structured and the spatially dependent components of the soil variable. Also there are economic and logistic reasons for including the ancillary variables influencing soil variability, especially if the latter are more readily and cheaply available. As both the soil and exogenous factors are multivariate, the most obvious choices are appropriate combinations of multivariate or univariate analysis using the scorpan factors and geostatistical methods. These combinations constitute the hybrid techniques.

The hybrid of scorpan and geostatistics

The hybrid techniques for soil survey and mapping are based on various combinations of the geostatistical and multivariate or univariate scorpan methods. Let us suppose that a data vector describing a soil property is a random variable Z , determined at locations s in a region $\mathfrak{A}$, $s_i (i = 1, \dots, N; s_i \in \mathfrak{A})$, and consisting of three components as:

$$Z(s) = m(s) + Z_1(s) + \varepsilon(s), \quad (4)$$

where m is the local mean for the region, Z_1 is the spatially dependent component, and ε the residual error term, spatially independent. Now there may be situations where m varies and is dependent on some exogenous factors such as the scorpan factors

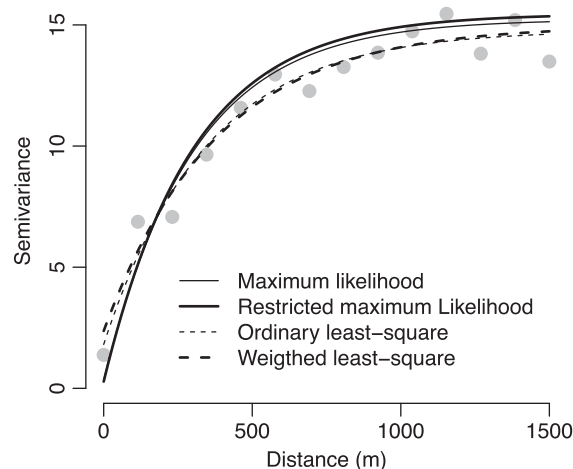


Fig. 5 An experimental variogram fitted with exponential models based on different fitting techniques.

at spatial location s . In other words it is deterministically related to some causative factors (in geostatistics parlance, the variable is said to exhibit a trend). Wherever trend exists, ordinary univariate kriging is inappropriate. Several methods have been designed to accommodate the trend.

Universal kriging has been the commonly used method to accommodate the trend or 'changing drift,' as it is sometimes known, in a soil variable. Universal kriging is a combination of the standard model of multiple-linear regression and the geostatistical method of ordinary kriging, which is also analogous to combining scorpan methods with univariate kriging, but only using the geographic coordinates for determining the drift (the n factor of the scorpan model). Another approach is to use an intrinsic random function of order k (IRF- k), which has been used to accommodate the varying nature of the trend in a regionalized soil variable. The term ' k ' represents the order of polynomial trends: $k = 0$ means constant drift, and the IRF- k is equivalent to the ordinary kriging system of equations. If $k = 1$, we have linear drift; $k = 2$ yields quadratic drift; but where there is no trend but deterministic relationships are with some known or readily available and inexpensive covariates (scorpan factors) or other easy-to-measure soil variables, co-kriging has played a major role in efficiently predicting the target soil variable. Universal co-kriging is also possible when considering the trend and covariation with one or more secondary variable.

Co-kriging is the multivariate extension of kriging that allows the inclusion of more readily available and inexpensive attributes in the prediction process. There are many instances in soil survey where the scorpan factors such as topography, time, and variable parent material, are easily discernible or are either readily available and/or are cheap to obtain. The most efficient way to predict the expensive-to-measure target soil variable, the variation of which is affected by the scorpan factors, is to supplement the information of the sparsely sampled target variable with more densely available information from a cheap to obtain variable.

Regression-kriging (RK) is another hybrid method that combines either a simple/multiple-linear regression model or machine learning (for example a variant of GLM, GAM, RT or RF) with ordinary, or simple, kriging of the regression residuals. The assumption here is that the deterministic component (m in Eq. 4) of the target (soil) variable is accounted for by the regression model, while the model residuals represent the spatially varying but dependent component (Z_1 in Eq. 4). If the exogenous variables used in the regression equation are available at more dense locations than the target variable, the equation can then be used to predict m on to those locations. The Z_1 can also be predicted to the same locations by the simple kriging system of equations, and then added to the m to obtain Z^* . A variant of RK is kriging with uncertainty, which introduces regression residuals (as representing model uncertainty) into the kriging system used to predict the target soil variable. This reduces the extrema of the target soil variable and therefore produces a smoother function of the predicted values. Another variant of RK, kriging with external drift (KED), is a hybrid technique that integrates the universality conditions into the kriging system using one or more of the ancillary drift variables. It is similar to universal kriging, but using an ancillary variable to represent the trend (Fig. 6). As shown in Fig. 6A, a regional digital map of cation exchange capacity (CEC) has been produced using KED with elevation as the external drift. Comparing this map with the one produced using RK (Fig. 6B) indicates a slightly more smoothed map produced by KED than RK.

Machine learning

Machine learning (ML) has emerged since the 1990s as a set of algorithms useful for soil mapping, but it is in the past decade only that the use of ML algorithms in soil science have burgeoned. Machine learning refers to a large set of non-parametric and data-driven algorithms developed for data mining and pattern recognition purposes, and now frequently used in all fields of soil science for classification and regression purposes (Wadoux et al., 2020). The increase in spatial prediction with ML corresponds with the recent availability of large digital soil databases and computer resources. In the previous Section we discussed some limitations of geostatistical mapping of soil, such as stationarity, normal distribution of the residuals, and difficulties for non-linear modelling of soil properties using numerous and cross-correlated covariates, among others. Machine learning can handle most of the limitations of geostatistical models when the amount of data to calibrate the model is sufficiently large. Indeed, ML algorithms do not make assumptions on the distribution of the soil properties and can accommodate numerous, categorical and continuous, and correlated covariates as predictors.

Examples of ML algorithms are, for example, regression tree, random forest (RF), support vector machine, genetic algorithms, and, perhaps the most popular of all, artificial neural networks (ANN). Variants of ANN are the convolutional neural network, which uses images as input and long short-term memory for space-time prediction. Despite the wide adoption of such models for mapping, it is still unclear if interpretation of these complex models is possible, and if it could reveal the importance of the environmental covariates used as predictors. Conversely, including pedological knowledge on the soil processes into these highly empirical models is, to date, very limited. These constitute current challenges in pedometric research (Wadoux et al., 2021).

Space-time prediction

Space-time prediction is concerned with the prediction of soil properties that vary both in space and time, such as soil moisture, soil available water capacity or soil organic carbon. Space-time prediction was first developed as an extension of classical geostatistics that considers spatial variation, but non-parametric models such as machine learning models also have potential extensions for space-time prediction. Extension to space-time prediction of a soil property is a challenge because variation in space might be different from variation in time. Two approaches are usually considered to extend spatial prediction to space-time. The first adds a time dimension by mapping a spatial variable repeatedly over time. Often this approach is preferred when the temporal correlation between time steps is weak, such as when mapping short-term change in soil texture. The second approach treats time as an additional dimension to be

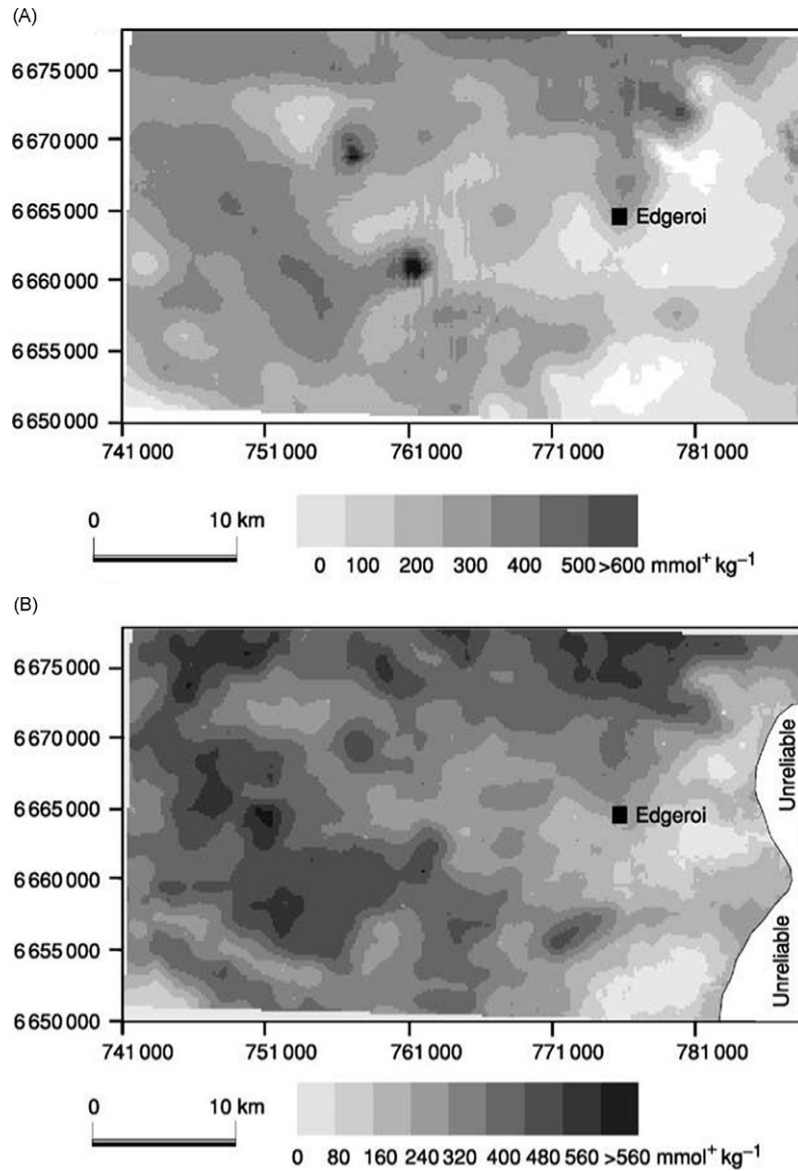


Fig. 6 Predicted cation exchange capacity (CEC; millimoles per kilogram) for the Edgeroi area using (A) kriging with elevation as external drift and (B) regression-kriging. Reproduced with permission from McBratney AB, Odeh IOA, Bishop TF, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. *Geoderma* **97**: 293–327.

considered. This approach is taken in space-time geostatistics by extending the methods of classical geostatistics to the space-time domain, such as the variogram. Many contributions from statisticians have been made to develop space-time variogram functions. Consider the model of spatial variation in Eq. (4), it is extended to the space-time domain by Heuvelink et al. (2017):

$$Z(s, t) = m(s, t) + Z_1(s, t) + \varepsilon(s, t), \quad (5)$$

where $Z(s, t)$ is a spatio-temporal random field, m is a space-time trend, often modelled as a function of known environmental covariates, and $Z_1(s, t)$ is a spatially and temporally dependent component usually assumed to be (multivariate) normally distributed and characterized by a covariance function. An example of variogram modelling the spatio-temporal covariance function is presented in Fig. 7. In space-time geostatistics, measurement can be repeated in time at the same spatial locations or vary between time steps. Predictions are made at unobserved locations in space and time.

Extension of machine learning models to the space-time domain is also possible but is still at the preliminary stages of academic research and has not yet reached a large audience. We expect this to change in the future with the development of new software implementations.

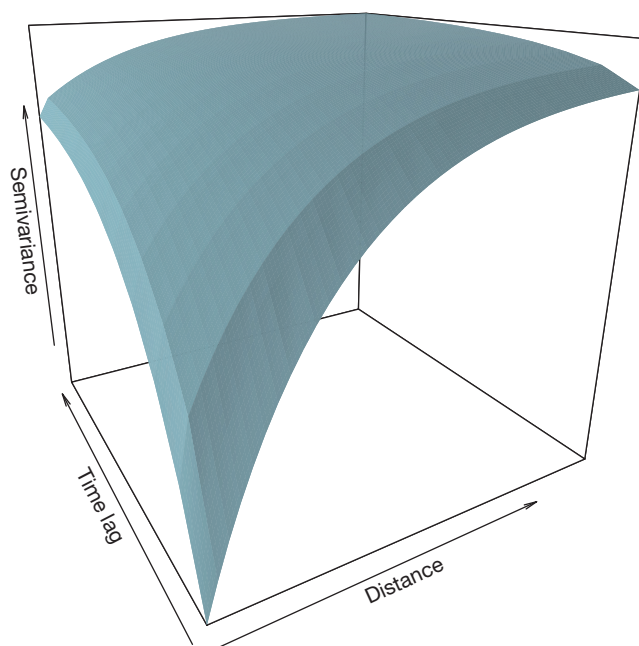


Fig. 7 A fitted spatio-temporal variogram using a separable function. This is the space-time extension of the spatial variogram presented in Fig. 5.

Sampling design

The sampling design used to calibrate and validate the spatial or space-time prediction models play a key role in the final accuracy. Pedometricians have contributed to the development of sampling approaches for mapping. Sampling strategies can be divided into two main approaches, namely the design- or model-based approaches (De Gruijter et al., 2006). The design-based approach relies on sampling theory where the sampling locations are selected randomly, either with simple random sampling, or by more advanced probability sampling designs such as stratified sampling. Generally, non-probability sampling approaches (i.e., model-based sampling) are more suitable for local mapping of the soil. Examples of such designs are regular grid sampling, feature space coverage or spatial coverage sampling. A method much used in digital soil mapping is conditioned Latin hypercube sampling (cLHS). Geostatistical techniques also enable sampling optimization, in which the fitted variogram and optimization algorithms enable the optimal spatial locations for mapping to be found. Much research has been done in this area, and developments along these lines are being made for machine learning. All these techniques apply for selecting the calibration sample, but it has been acknowledged that whenever possible the soil maps should be validated by independent, probability sampling (Brus et al., 2011).

Soil utility and quality

Land evaluation

Pedometrics has played a major role in advancing the role of land evaluation and, more recently, the quantification of soil quality for land management and sustainable use of land resources (Rossiter et al., 2018). Although 'soil quality assessment' is often used as a misnomer for 'land evaluation,' both could be regarded as the interpretative phase of soil survey. While land evaluation is concerned with the assessment of land performance when used for specified purposes, soil quality is defined as 'the capacity of a specific kind of soil to function, within natural or managed ecosystem boundaries, to sustain plant and animal productivity, maintain or enhance water and air quality, and support human health and habitation'. In considering productivity, environmental quality, and human health as major functions of soil, this definition requires that values be placed on specific soil functions as they relate to the overall sustainability of alternate land-use decisions. Pedometric techniques (based on 'hard' and/or fuzzy techniques) are being used to quantify soil quality and associated uncertainty. For example soil quality can be assessed in terms of requirement for different specific uses such as shown in Fig. 8. Many of the soil quality indicators are required at various scales and spatial extents in which pedometrics is playing a major role in modelling them spatially for incorporation into regional, catchment, and field-scale modeling. Another approach to soil quality analysis may be based on the concept of pedodiversity (see also previous section).

Soil functions

Pedometrics is also playing a role in quantifying and mapping soil functions. Soils provide services in terms of agricultural production, water protection and filtration, carbon storage or biodiversity habitat, among others. The ability of a soil to provide

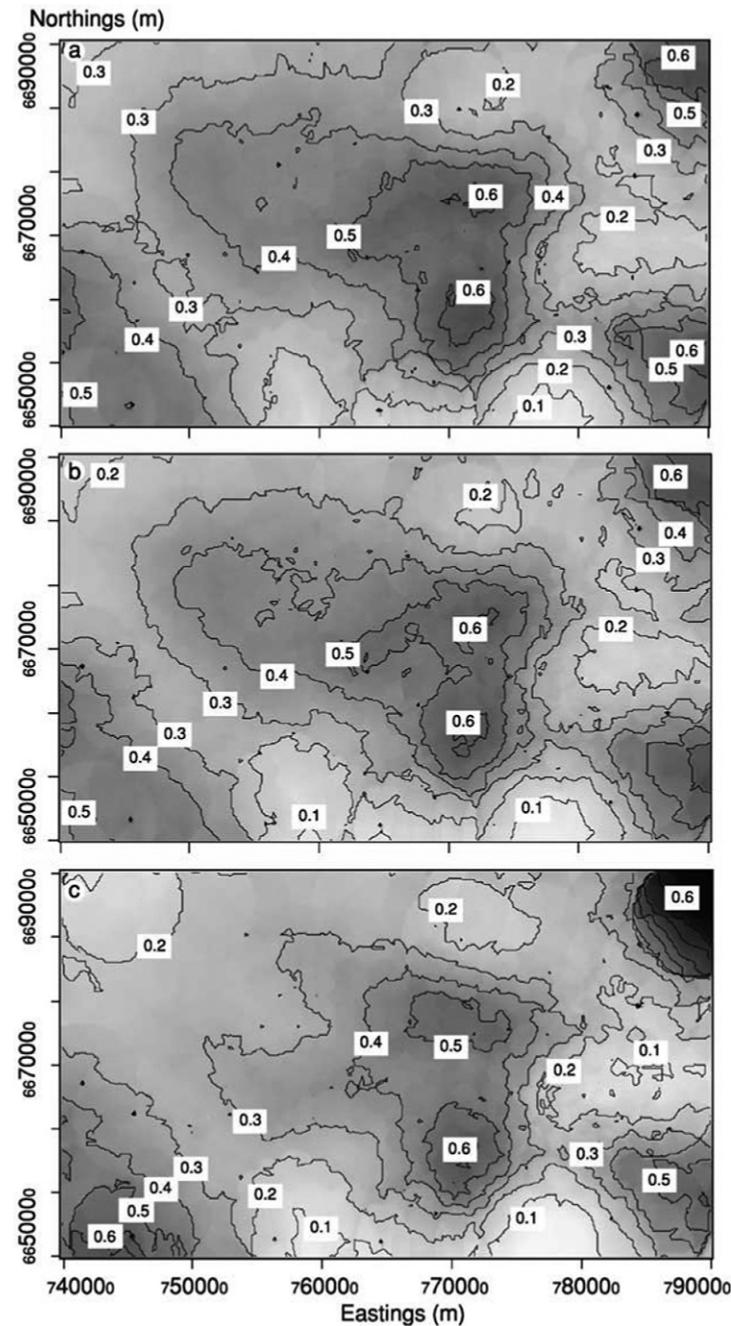


Fig. 8 Interpolated suitability of (A) wheat, (B) sorghum and oats, and (C) dryland cotton. Based on Triantafyllis J, Ward WT, and McBratney AB (2001) Land suitability assessment in the lower Namoi Valley of Australia using a continuous model. *Australian Journal of Soil Research* 39: 273–290.

a set of ecosystem services is determined by its functions, which can be evaluated in various ways. One way is to use empirical models based on a set of indicators (i.e., basic soil properties) and combine them into soil functions. The choice of indicators is based on correlation statistics between properties, on pedological knowledge as well as on a context in which the functions will be used. Another way to model functions is to use biophysical models that better account for the soil management practices and environmental factors. In both ways, the different soil functions are interdependent since they are based on similar basic soil indicators. A change in agricultural practices to improve one function (e.g., production) can simultaneously affect other functions positively or negatively (e.g., the habitat of microorganisms). Because of these synergies and antagonisms between functions, a soil can never mobilize to its full potential each of the functions it can support. The need to study multifunctionality was the basis for the concept of soil security, or other concepts such as soil health. Converting basic properties into a product that better matches end-user demand was also the objective of land evaluation. Pedometric techniques are being used to quantify soil functions, to

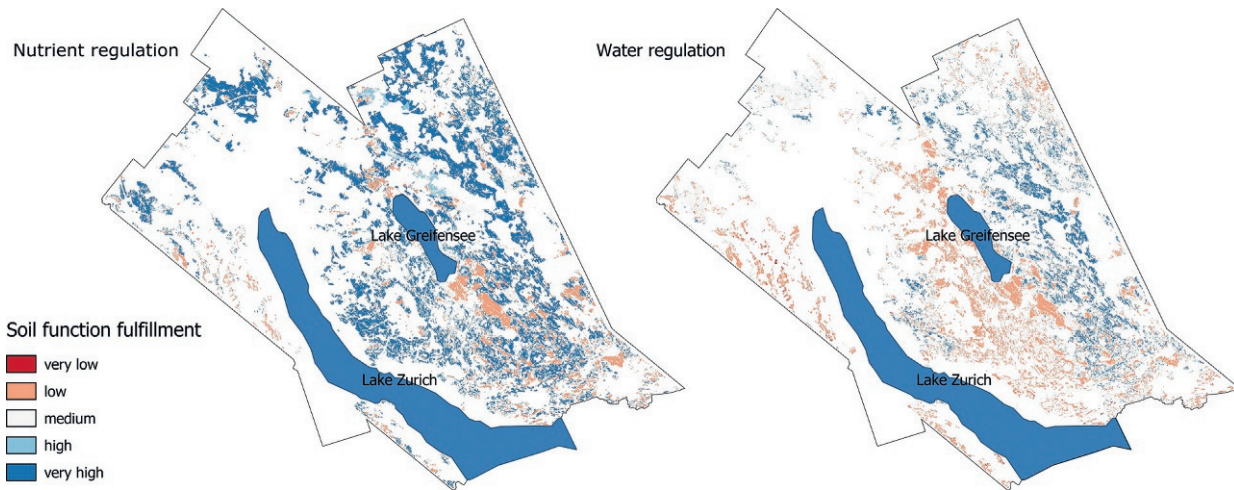


Fig. 9 Soil function maps of water regulation function and nutrient storage capacity for the agricultural land in the Greifensee study area, Switzerland. Reproduced with permission from Greiner, L, Nussbaum, M, Papritz, A et al. (2018). Assessment of soil multi-functionality to support the sustainable use of soil resources on the Swiss Plateau. *Geoderma Regional* **14**: e00181.

account for their spatial and temporal variation and to quantify and propagate their uncertainty. For example, in Fig. 9 two maps of soil functions are shown.

Quantitative pedogenesis

As stated above, soil variability is a function of factors of soil formation. Since the publication of Jenny's state factors, pedologists have attempted to develop pedogenetic (mathematical) functions that could explain soil variability and, in many instances, predict or even simulate the soil. This has resulted in models of pedogenesis based on chemical and physical processes. The formulation of the process models and their applications depend on the scales in both the temporal and spatial dimensions. This modelling of the space-time continuum is determined by the complexity of the modelling process, which can be characterized in several ways. The first is based on computational complexity of the model, which ranges from purely qualitative (mental models) to highly quantitative, with the latter involving complex computer coding; the second is based on the complexity of the model structure, which distinguishes between mechanistic (highly complex) and empirical (simplified functional) models; and the last, but not least, is based on organizational hierarchy, which determines at which level a model is used to simulate the soil system. The hierarchical levels range from $i - 4$ (molecular-level processes) to $i + 6$ (global level). Each level is a subsystem of the level above it, therefore allowing room for upscaling.

The three characterizations above are well illustrated in Fig. 10. It should be noted that pedometrics has played a dominant role in the quantitative half of the organizational hierarchy, particularly from and above the $i - 3$ level, which is governed by material fluxes in the pores between the primary particles. Recent developments in process modelling focus on mechanistic stochastic simulations, particularly at levels $i - 3$ to i (Fig. 10) and on how the latter can be upscaled to any of the higher hierarchies.

Soil processes

Soil processes that are responsible for differentiation of the soil profile have been modeled in various ways (Stockmann et al., 2018). Soil transformation processes such as soil physical and chemical weathering are modelled through empirical chronofunctions or by mass balance models linking the chemical composition of bedrock to soil profile properties. Translocation processes such as eluviation/illuviation and soil mixing are quantified by mathematical functions based on soil textural properties. Overall, these models require considerable input of measurement data, for example radionuclide data, and have many parameters to estimate. Other problems are that these models do not provide uncertainty, and have been derived using data from soil profiles or at the pedon scale (i.e., they have scaling issue to larger area). Pedometricians have contributed to solving some of these challenges by the spatialization of soil profile process models.

Soil-landscape evolution

Process modelling has enabled the development of soil-landscape evolution models. These models are quantitative models of the short- and long-term processes of soil formation and which predict the present-day variation of soil properties. Often these models

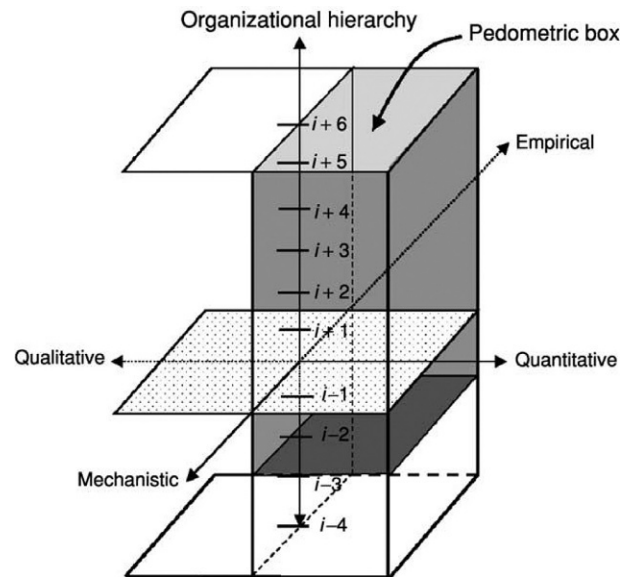


Fig. 10 Pedometric box in the organizational hierarchy of pedogenetic modeling approach. Adapted from Hoosbeek MR and Bryant RB (1992) Towards quantitative modelling of pedogenesis—A review. *Geoderma* 55: 183–210, with permission.

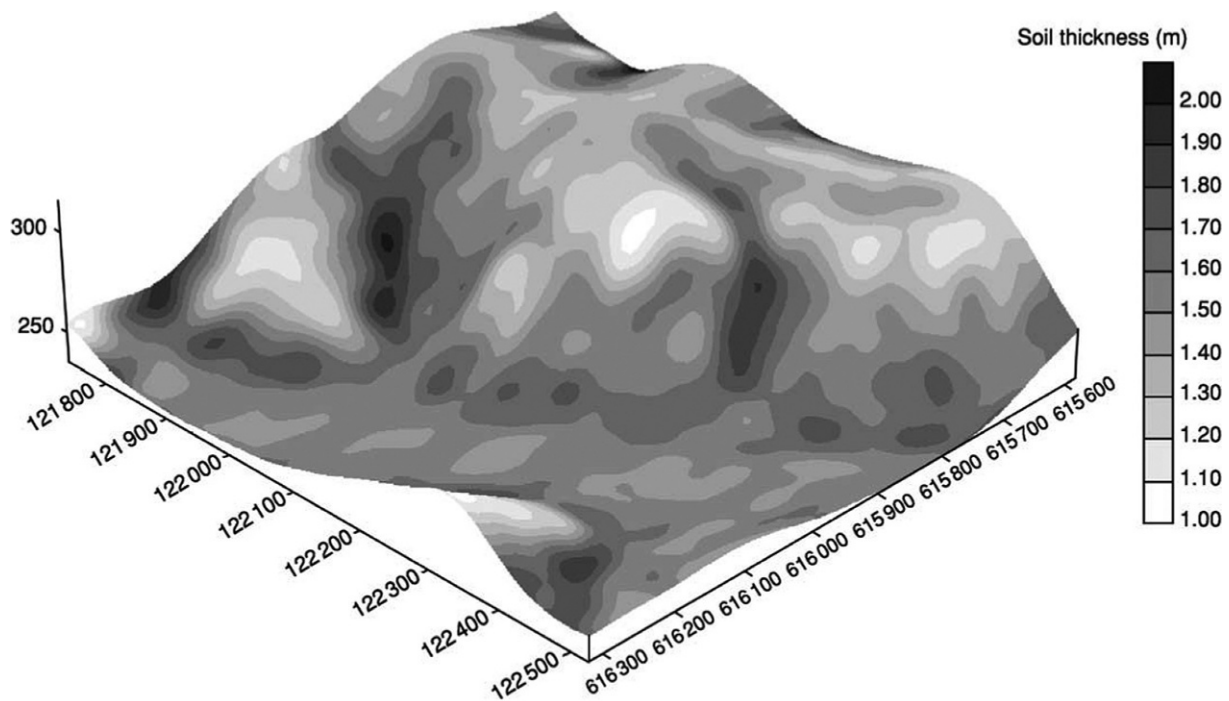


Fig. 11 Soil formation in a landscape after 10,000 years. Reproduced with permission from Minasny B and McBratney AB (2001) A rudimentary mechanistic model for soil formation and landscape development: II. A two-dimensional model incorporating chemical weathering. *Geoderma* 103: 161–179.

are a combination of the understanding of process and empirical modelling. For example: a mechanistic pedogenetic model, based on a digital elevation model (DEM, Burrough and McDonnell, 2000), was used to simulate pedogenesis by a combination of several submodels: (1) rate of physical weathering as an exponential decline function of soil thickness; (2) rate of chemical weathering represented as a negative exponential function of both soil thickness and time; and (3) the movement of material as characterized by the diffusion transport model. The result of upscaling such an analysis is illustrated in Fig. 11, whereby, after 10,000 years, soil accumulation is predominant in the gullies compared with the ridges, where soil erodes. See also Minasny et al. (2015) for further reading (Fig. 11).

Recent research has shown that the main impediments of soil process modelling discussed above (i.e., data requirement, mechanistic understanding, number of parameters to estimate, scaling problems, uncertainty) may be tackled using statistical (data-driven) modelling, for example with state-space modelling and the space-time Kalman filter (Webster and Heuvelink, 2006). With increasing environmental concerns, the development and applications of dynamic landscape models under different land-use scenarios may well dominate pedometric research in the next few decades.

Conclusion and outlook

Stepping back in time, pedometrics appeared from a need for quantitative and digital methods in soil science and emerged from research on numerical soil classifications and geostatistical modelling of spatial variation. During the past two decades, the considerable demand for soil information has motivated pedometricians to address other soil questions from a quantitative point of view: digital soil mapping, soil monitoring, quantitative pedogenesis and soil utility, soil functions and security and to diversify the tools and techniques that they use to carry out such research: soil spectroscopy, machine learning, spatio-temporal soil inventories and quantitative process-based soil-landscape models, among others. These efforts go beyond the scope of this short introductory chapter, but were described in 2018 in a book (McBratney et al., 2018) that attempts to cover almost all aspects of pedometric research.

While pedometric research is undeniably in an expansion phase and has now many areas of application, some gaps in knowledge and challenges need to be tackled. Wadoux et al. (2021) describe 10 recent and longstanding pedometric challenges. They are classified into three aspirational categories, (i) a better understanding of soil formation, (ii) an improvement in methods for obtaining relevant digital soil data, and (iii) an improvement of our ability to address demands by soil users. These challenges show that pedometrics can contribute to many areas of soil science, and is at the interface with many questions relevant to the sustainable growth of our societies.

References

- Brus DJ, Kempen B, and Heuvelink GBM (2011) Sampling for validation of digital soil maps. *European Journal of Soil Science* 62(3): 394–407.
- Burrough AB and McDonnell RA (2000) *Principles of Geographical Information Systems*. New York: Oxford University Press.
- De Grujter JJ, Brus DJ, Bierkens MF, and Kotters M (2006) *Sampling for Natural Resource Monitoring*. Berlin: Springer.
- Heuvelink GBM and Webster R (2001) Modelling soil variation: past, present, and future. *Geoderma* 100(3–4): 269–301.
- Heuvelink GBM, Pebesma E, and Gräler B (2017) Space-time Geostatistics. In: Shekhar S, Xiong H, and Zhou X (eds.) *Encyclopedia of GIS*. Cham: Springer.
- Hughes P, McBratney AB, Huang J, et al. (2018) A nomenclature algorithm for a potentially global soil taxonomy. *Geoderma* 322: 56–70.
- Ibáñez JJ and Bockheim JG (2013) *Pedodiversity*. Boca Raton: CRC Press.
- McBratney AB and Lark RM (2018) Scope of pedometrics. In: Minasny B, McBratney AB, and Stockmann U (eds.) *Pedometrics*, pp. 7–39. Cham: Springer.
- McBratney AB and Odeh IO (1997) Application of fuzzy sets in soil science: Fuzzy logic, fuzzy measurements and fuzzy decisions. *Geoderma* 77(2–4): 85–113.
- McBratney AB, Odeh IOA, Bishop TFA, Dunbar MS, and Shatar TM (2000) An overview of pedometric techniques for use in soil survey. *Geoderma* 97: 293–327.
- McBratney AB, Mendonça Santos ML, and Minasny B (2003) On digital soil mapping. *Geoderma* 117: 3–52.
- McBratney AB, Budiman M, and Stockmann U (2018) *Pedometrics*. Cham: Springer.
- Minasny B, Finke P, Stockmann U, Vanwalleghe T, and McBratney AB (2015) Resolving the integral connection between pedogenesis and landscape evolution. *Earth-Science Reviews* 150: 102–120.
- Nocita M, Stevens A, van Wesemael B, et al. (2015) Soil spectroscopy: An alternative to wet chemistry for soil monitoring. *Advances in Agronomy* 132: 139–159.
- Rossiter DG, Hewitt AE, and Dominati EJ (2018) Pedometric valuation of the soil resource. In: Minasny B, McBratney AB, and Stockmann U (eds.) *Pedometrics*, pp. 521–546. Cham: Springer.
- Stockmann U, Salvador-Blanes S, Vanwalleghe T, Minasny B, and McBratney AB (2018) One-, two- and three-dimensional pedogenetic models. In: Minasny B, McBratney AB, and Stockmann U (eds.) *Pedometrics*, pp. 555–593. Cham: Springer.
- Van Looy K, Bouma J, Herbst M, et al. (2017) Pedotransfer functions in Earth system science: challenges and perspectives. *Reviews of Geophysics* 55(4): 1199–1256.
- Wadoux AMJ-C, Minasny B, and McBratney AB (2020) Machine learning for digital soil mapping: applications, challenges and suggested solutions. *Earth-Science Reviews* 210: 103359.
- Wadoux AMJ-C, Heuvelink GBM, Lark RM, et al. (2021) Ten challenges for the future of pedometrics. *Geoderma* 401: 115155.
- Webster R (1994) The developments of pedometrics. *Geoderma* 62: 1–15.
- Webster R and Heuvelink GBM (2006) The Kalman filter for the pedologist's tool kit. *European Journal of Soil Science* 57: 758–773.
- Webster R and Oliver MA (2007) *Geostatistics for Environmental Scientists*, 2nd edn. Chichester: John Wiley & Sons.

Statistics in soil science

Richard Webster, Rothamsted Research, Harpenden, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction—Why statistics?	487
Population, units and samples	487
Replication and randomization	487
Descriptive statistics	489
The mean	489
Characteristics of variation	489
Histogram	489
Variance	490
Estimation variance, standard error and confidence	490
Coefficient of variation	491
Additivity of variances	491
Statistical significance	491
Transformations	492
Analysis of variance	494
Fixed effects, random effects and intra-class correlation	495
Covariance, correlation and regression	496
Covariance	496
Correlation	496
Spearman rank correlation	496
Regression	497
Further matter	498
References	498

Abstract

Statistics embodies numerous techniques for analyzing data to discover meaning in those data against a background of variation that is of no interest, i.e., of separating signal from noise. At its heart is the notion of variance, the expected squared difference between the mean of a population and any member of it. Estimation of a variance requires replicate observations; randomization is necessary to avoid bias in estimates. Summary statistics of sets of observations comprise means, medians, variances, standard deviations and skewness coefficients. Distributions are best displayed by histograms and Q–Q plots. Analysis of variance enables investigators to estimate the effects of treatments of soil in designed experiments and the confidence they can have in their results. The analysis must match the design. The same forms of analysis can be applied to survey data in which several classes are distinguished. Relations among two or more variables are expressed by Pearson correlation coefficients calculated from measurements on those variables. The same basic calculations underlie regression in which one variable depends on one or more others and is predicted from them, and the strength of the predictions can again be assessed by analysis of variance.

Key points

- Soil varies from place to place; statistics is a technology for describing that variation quantitatively and separating signal from noise.
- Replicate sampling is required to estimate the variation in experimental material and geographic regions.
- Randomization is needed to estimate mean values and their variances, and to place confidence limits on means.
- Data can be summarized by means, medians, variances, standard deviations and skewness coefficients.
- Distributions of data are best illustrated by histograms and Q–Q plots.
- Analysis of variance apportions the variance in a set of data to sources in the designs of experiments and surveys; any actual analysis must fit the design by which the data have been obtained.
- Relations among two or more measured variables are expressed by Pearson correlation coefficients.
- A variable may be predicted from one or more others on which it depends by regression.

Introduction—Why statistics?

Soil varies from place to place at all scales from the global to the infinitesimal. Much of the variation is natural, arising from variation in its parent material and processes such as erosion and deposition on the land surface. Soil also varies from time to time as it responds to weather, plant growth, and processes in the rhizosphere engendered by that growth. Farmers have added to the variation by their enclosing, reclaiming, clearing, and fertilizing the land, though within fields they have removed some by cultivation and drainage. Further sources of variation are mineral extraction and subsequent reclamation, dumping of waste, and pollution of many kinds. Data on soil embody variation from those many sources, and the aim of statistics at its simplest is to express quantitatively that variation, to separate contributions from different sources and to describe relations.

Investigators design experiments and surveys in such a way that they can estimate the variation from particular sources such as imposed treatments or strata in a population and the differences between them. Variation in data also arises from the way observations are made; from the people who make the observations, from the imprecision of instruments, from imperfect laboratory technique, and from sampling fluctuation. One may like to think that the contributions from the laboratory are negligible, though ‘ring’ tests have often revealed them not to be so. In general, however, sampling fluctuation in the field, arising from the spatial variation there, is much the larger.

Any one measurement of a soil property is influenced by contributions from at least some of these sources. It cannot be taken as an absolutely accurate representation of the truth therefore; rather it must be seen in relation to the likely error.

Statistics are needed in soil science to estimate and express this error and to apportion it to the different sources. In this way sense (signal) can be separated from meaningless or uninteresting variation (noise) in comparisons between classes of soil, in expressing relations, in assessing the effects of treatments in experiments, and in prediction. The statistical repertoire is huge, and this article describes the basics and the fairly elementary techniques that soil scientists most often need.

The techniques can be divided into two groups, namely, description and analysis. They also have two fairly distinct fields of application: survey and experiment. In the first investigators observe and record the soil as it is on samples, and descriptive techniques tend to dominate. In the second they deliberately control some of the variation so that they can assess the effects of changes in one or a few factors that are of specific interest by analysis.

There are numerous text books on the statistics summarized in this chapter, and many more tutorial articles in scientific journals and on the World-Wide Web. I recommend strongly three: [Snedecor and Cochran \(1989\)](#), [Sokal and Rohlf \(2012\)](#) and [Welham et al. \(2015\)](#).

Population, units and samples

The soil is regarded for statistical purposes as a *population* comprising elements or *units*. The units are the bodies of soil on which measurements are made. They are more or less arbitrary and determined largely by convenience and practicality. They may be individual cores of soil, pits or pedons, each may be the volume of soil occupied by the roots of a single plant or that deformed under the wheel of a tractor, they may be pots in a greenhouse experiment, lysimeters, plots in a field experiment, or whole fields or farms. The variation among them depends very much on the size of the units; the larger they are the more variation they encompass within them and the less there is between them to be revealed in data. The size, shape, and orientation of the units, known as the “support” in survey, must be defined and adhered to throughout an investigation. The population is then all such units in a specified region or falling within some other definition for the purposes of the investigation. It is an operational definition, often known as the “target population.” In a more restricted sense the population and the units comprising it may be the values of a particular soil property in the defined supports.

Populations in surveys are typically very large, in many instances infinite, or hypothetical, and in some they are poorly defined. Measurement is feasible only on small subsets, i.e., on samples, and these subsets must properly represent their populations for the measurements to apply to the larger populations. The units in an experiment, in contrast, are typically a few dozen, though ideally the experimental material should be representative of some larger population.

Replication and randomization

Estimating the mean in a population and its associated error from measurements on a sample requires both replication and randomization. The need for replication is evident; a single value can contain no information on variation. Randomization is needed to avoid bias. At its simplest it means choosing units such that all have the same chance of being selected.

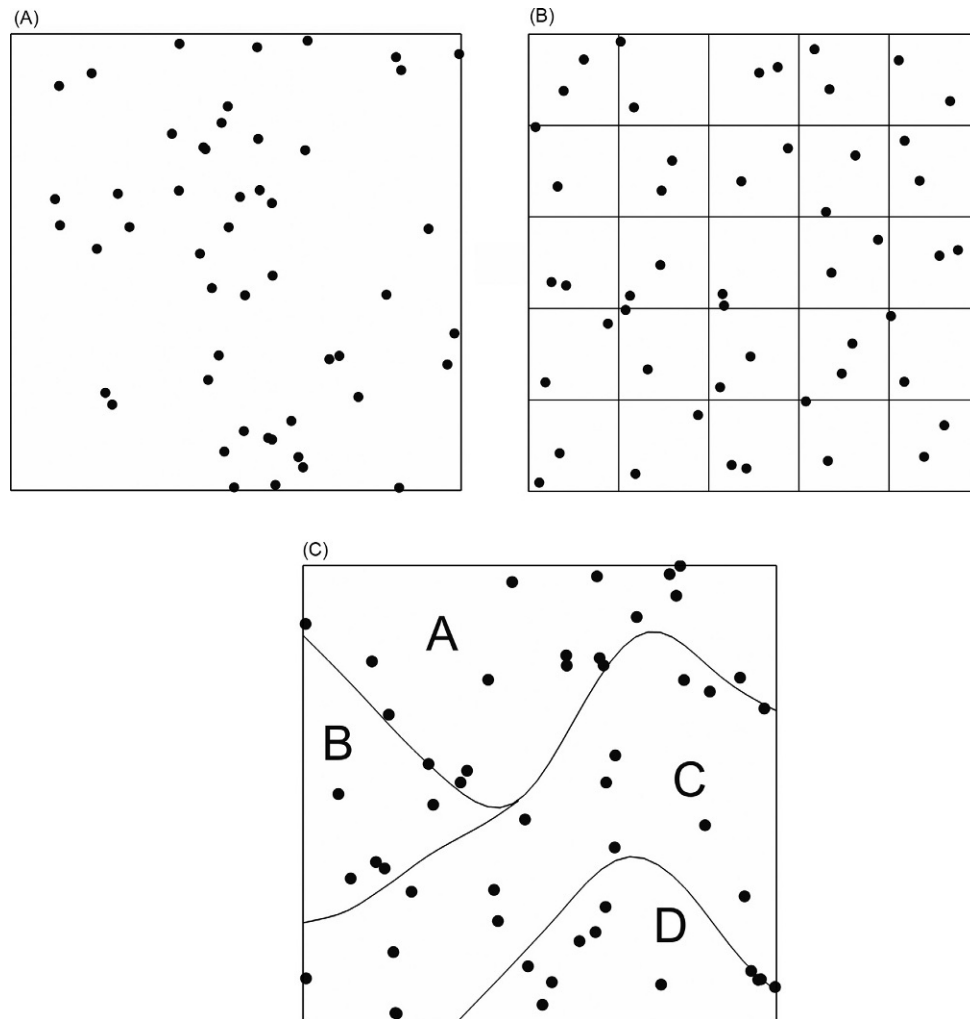


Fig. 1 Probabilistic sampling of a region; (A) simple random sampling; (B) stratified random sampling with the region divided into 25 square cells (strata) and two points per cell; (C) stratified random sampling with four mapped classes of soil as strata.

In survey selection can take the form of simple random sampling. To apply it requires (1) that each unit can be identified uniquely, and (2) a rule for the selection. Simple random sampling, as in Fig. 1A, tends to be inefficient in that it takes no account of anything known about the population beforehand, such as its spatial dependence, a soil map of the region, the relations between soil and vegetation or physiography, or the farming history. To take advantage of such knowledge the population is first stratified either into small grid cells, as in Fig. 1B, or by type of soil, vegetation, physiography, or farming as above, as in Fig. 1C. The soil is then sampled at random within each stratum independently; this is stratified random sampling, and it enables the variation due to the strata to be distinguished from variation within them.

Brus (2021) has drawn attention to a growing notion that random sampling is to be avoided where there is spatial correlation. As Brus points out, that notion is false. See also de Gruijter et al. (2006). Random sampling guarantees that estimates of means and variances are unbiased, regardless of any spatial correlation, and estimating them needs to be distinguished from mapping for which other sampling designs are more efficient. Do not be misled by publications that state otherwise; stick to the authoritative texts mentioned above and to Cochran's (1977) standard text on sampling.

In experiments treatments are allocated to the units at random. The units may be arranged completely randomly, as in Fig. 2A. In the field and greenhouse they are usually grouped into blocks such that each block contains one unit of each treatment, as in Fig. 2B; long-range variation then appears in the variation among blocks. There are many more elaborate designs.

(A)					(B)				
W	0	F	W	D	W	D	F	W	F
F	D	W	F	0	F	0	D	F	0
W	0	W	0	D	0	W	W	0	D
D	F	F	D	0	D	F	0	D	W

Fig. 2 Layout of a randomized field experiment with four treatments, 0 (control), D (dung), W (industrial waste), and F (NPK fertilizer), and fivefold replication; (A) completely randomized; (B) with the replicates arranged in five blocks.

Descriptive statistics

The mean

In almost all investigations mean values are of prime interest. Provided sampling has been properly randomized the mean of the data, denoted $\bar{z}_1, \bar{z}_2, \dots, \bar{z}_N$,

$$\bar{z} = \frac{1}{N} \sum_{i=1}^N z_i \quad (1)$$

estimates without bias the mean, μ , in the population from which the sample was drawn. How well it does so depends on the variation it embodies.

Characteristics of variation

Histogram

The variation in a set of measurements, if there are sufficient of them, can be displayed in a histogram. The scale of measurement is divided into segments of equal width, or “bins,” the values falling in each bin are counted, and bars of height proportional to the counts are drawn. Fig. 3 is an example; it summarizes graphically the way in which the frequency is distributed over the range of the data.

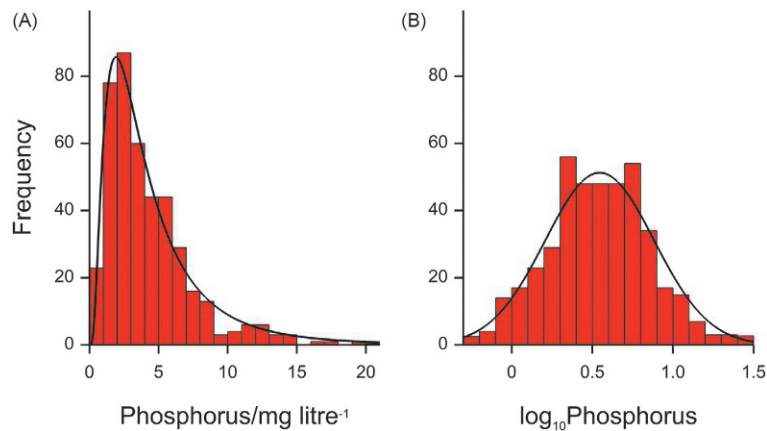


Fig. 3 Histograms of 434 data on available phosphorus (A) in mg L^{-1} , and (B) transformed to common logarithms. The curve in (B) is that of the fitted normal distribution, and that in (A) shows the lognormal distribution.

Variance

Variation is best expressed quantitatively as variance. For a set of N data it is the average squared difference between the observations and their mean:

$$S^2 = \frac{1}{N} \sum_{i=1}^N (z_i - \bar{z})^2 \quad (2)$$

Its square root, S , is the standard deviation, which is often preferred because it is in the same units as the measurements and is more intelligible therefore.

Ideally this variance should estimate the variance of the larger population, of which the N observations are a sample. This variance of a population is by definition

$$\sigma^2 = E[(z - \mu)^2] \quad (3)$$

where μ is mean of z in the population and E denotes expectation. Eq. (2) gives a biased result, however; S^2 is a biased estimator. The reason is that $\bar{z}$ in the equation is itself more or less in error as the estimate of μ . To remove the bias N must be replaced by $N - 1$ in the denominator:

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (z_i - \bar{z})^2 \quad (4)$$

The result, s^2 , is now unbiased, provided the sampling was unbiased in the first place.

Estimation variance, standard error and confidence

Both S and s measure the dispersion in the observations; neither expresses the reliability of the estimate of μ . The reliance one can place in an estimate of the mean can be expressed in terms of the expected squared deviation of it from the true mean, i.e., $E[(\bar{z} - \mu)^2]$. Provided that the sampling points are drawn independently of one another, as for example in simple random sampling, its estimate is derived from s^2 by

$$s^2(\bar{z}) = s^2/N \quad (5)$$

This is the estimation variance, and its square root is the standard error, which is the standard deviation of means of samples of size N . The larger is the sample the smaller is this error, other things' being equal, and the more confident we can be in the estimate. So, distinguish between the standard deviation, which describes the variation in a sample, and the standard error, which expresses the confidence associated with a mean. Standard errors should accompany means in tables compiled from replicated measurements, and they may be shown by error bars on graphs.

Eq. (5) gives the estimation variance for a simple random sample of size N . If the population has been divided into strata then s^2 , the population variance on the right hand side of the equation, can be replaced by s_w^2 , the variance within strata. The latter is generally less, often much less, than s^2 , and so stratified sampling is more precise than simple random sampling by the factor $s_{\text{random}}^2(\bar{z})/s_{\text{stratified}}^2(\bar{z})$. It also means that one can achieve the same precision with a smaller sample, and in this sense stratified sampling is more efficient. This efficiency can be expressed as $N_{\text{random}}/N_{\text{stratified}}$. The within-stratum variance, s_w^2 , can be estimated by the analysis of variance, see below.

The measures of error above represent uncertainty, which in the current context can be translated into confidence, that is the confidence we can have that a true mean, μ , deviates no more than some value from our estimate. Limits of confidence of our choosing can be obtained from the standard errors for known kinds of distribution. If the data are drawn from a normal (Gaussian) distribution and there are many of them then an interval of $2s/\sqrt{N}$ centered on $\bar{z}$ spans a symmetric confidence interval of approximately 68%. We can readily calculate wider confidence intervals by multiplying by factors, η , such as

Confidence/%	80	90	95	99
η	1.28	1.64	1.96	2.58

The factor η is the value of the standard normal deviate for the desired confidence. Evidently, the more certain, or more confident, we want to be that μ lies within some limits the wider that confidence interval must be.

When data are few ($N < 30$) the above factors need to be replaced by Student's t for the number of degrees of freedom, f . The lower and upper symmetric confidence limits about a mean $\bar{z}$ are

$$\bar{z} - t_f s / \sqrt{N} \text{ and } \bar{z} + t_f s / \sqrt{N} \quad (6)$$

Values of t are available in tables and in most statistical computer packages. For $30 < N \leq 60$ t converges to η , and for $N > 60$ the values in the above table can be used.

There are formulae for calculating the confidence limits for other theoretical distributions. In many instances, however, it is easier to transform data to approximate a normal distribution and subsequently analyze the transformed values.

Coefficient of variation

The coefficient of variation (CV) is the standard deviation divided by the mean, i.e., $s/\bar{z}$. It is often multiplied by 100 and quoted as a percentage. Its merit is that it expresses variation as pure numbers independent of the scales of measurement. It enables investigators and those reading their reports to appreciate quickly the degree of variation present and to compare one region with another and one experiment with another. It should not be used to compare variation in different variables, especially ones having different dimensions.

The coefficient is sensible only for variables measured on scales with an absolute zero. Otherwise the arbitrary choice of the zero affects it. Examples for which it should not be used are soil temperature in degrees Celsius (arbitrary zero at 273 K), color hue (which is approximately circular), and soil acidity on the pH scale (arbitrary zero equivalent to $-\log_{10}[\text{H}^+]$ with H^+ expressed in mol L^{-1}).

For some soil properties physics sets limits on the utility of the CV. For example, the minimum bulk density of the soil is determined by the physical structures that keep particles apart. Particles must touch one another, otherwise the soil collapses. For mineral soils on dry land a working minimum bulk density is around 1 g cm^{-3} . At the other end of the scale the bulk density cannot exceed the average density of the mineral particles, approximately 2.7 g cm^{-3} . So the CV of bulk density is fairly tightly constrained.

The sensible use of the coefficient of variation for comparing two variables y and z relies on the assumption that they are the same apart from some multiplying factor, b , thus:

$$y = bz \quad (7)$$

Then the mean of y is $\bar{y} = b\bar{z}$, its variance $s_y^2 = b^2 s_z^2$, and its standard deviation is $s_y = b s_z$. From there simple arithmetic shows that their CVs are the same. This principle offers a means of comparing variation with logarithms of the observations. Eq. (7) becomes

$$\log y = \log b + \log z \quad (8)$$

The logarithm $\log b$ is a constant, and so the variances, $s_{\log y}^2$ and $s_{\log z}^2$, are equal, as are their standard deviations. The resulting measure of variation is therefore independent of the original scale of measurement.

The measure can be used to compare variation in two groups of observations. Consider again soil acidity. To compare the variation in acidity of a class A with that in another, class B, we treat the hydrogen ion concentration as the original variable, transform it to pH, and compute the variances of pH. Whichever has the larger variance is the more variable, regardless of the mean. Further, we can make a formal significance test by computing $F = s_{\log y}^2 / s_{\log z}^2$ and compare the result with F for $N_y - 1$ and $N_z - 1$ degrees of freedom, see below.

Additivity of variances

Variances are additive; those from two or more independent sources in an investigation sum to the total in the data. Their square roots, the corresponding standard deviations, are not. To obtain an average variation on the original scale of measurement from several sets of data compute the arithmetic mean of their variances, weighted as appropriate by the numbers of degrees of freedom, and then take the square root of it to give an "average" or pooled standard deviation. This divided by the mean will give an "average" CV. More generally, the additive nature of variances confers great flexibility in analysis, enabling investigators to distinguish variation from two or more sources and estimate their components according to the design as by the analysis of variance.

Statistical significance

Significance in a statistical context means distinguishing a signal or the effects of some imposed treatment or detecting differences between strata against a background of "noise." It is matter of separating the variance due to the signal, treatments, or strata from that from other sources that are of no interest. The question being addressed is as follows; given the magnitudes of the several components of variance, is the signal so strong or are the differences observed so large that they are highly unlikely to have occurred by chance. If the answer is "yes," then the result is said to be significant.

A significance test is prefaced by a hypothesis. This is usually that there is no real difference between populations or treatments and that any differences among the means of observations are due to sampling fluctuation. That is the "null hypothesis," often designated H_0 , and the test is designed to reject it; *not to confirm it*. The alternative that there is a difference is denoted either H_1 or H_A .

To judge, for example, whether two means differ one computes from the sampling error the probability, P , of obtaining the observed difference if the true means were identical, assuming that one knows the form of the distribution. If P is small (conventionally <0.05) the null hypothesis is rejected, and the difference is judged significant. If P is large then the null hypothesis is likely to be correct, but we have no measure of the probability that the two means are indeed identical; instead we take the view that we have too little evidence to conclude that the difference observed applies to the population from which the sample was drawn.

One can still draw mistaken conclusions as the result of significance tests. Mistakes can be of two kinds, denoted Type I and Type II. The first occurs when we reject the null hypothesis on the basis of our sample evidence, i.e., declare a difference significant, when

the populations do not differ. The second is when we accept the null hypothesis, i.e., state that we have insufficient evidence for a difference, when the populations do differ.

One can diminish the likelihood of drawing wrong conclusions by increasing the sensitivity of the test, and that depends on the precision with which the means have been estimated, i.e., on their estimation variances or on the estimation variance of their difference. The latter is given by

$$s_{\text{diff}}^2 = \frac{s_W^2}{n_1} + \frac{s_W^2}{n_2} \quad (9)$$

where n_1 and n_2 are the numbers of observations from which the means in classes 1 and 2 derive, and s_W^2 is the variance within the classes, assumed to be common. If $n_1 = n_2$ then the variance of the difference is simply twice the estimation variance of the individual means.

Eq. (9) shows two features. One is that the larger is the variance within the populations the larger is the variance between the means and the less likely is one to establish a difference as significant. The second is that larger samples result in smaller variances and hence more sensitive comparisons. If the samples are large enough one can establish that any soil is different from almost any other for whatever property of interest.

The significance test is valuable in preventing false claims on inadequate evidence. Thus, a result might be summarized as follows.

The mean measured pH of the topsoil was 5.7 compared with 6.7 in the subsoil; but because the samples were small [or because the variances were large] the difference was not statistically significant.

However, when a difference is deemed statistically significant because the null hypothesis is rejected that does not mean that it is important or physically or biologically meaningful. For example, the difference between an observed mean pH of 5.7 in the topsoil and 5.9 in the subsoil would be of little consequence, whatever the probability of rejecting the null hypothesis. Also, while an investigator might regard a difference as significant if $P \leq 0.05$, a reader may be willing to recognize one if $P < 0.1$ or, more stringently, only if $P \leq 0.01$. The critical value of $P = 0.05$ for significance was originally suggested of R.A. Fisher; it was not a rule to be followed slavishly. Now that one can calculate probabilities so readily in computer packages professional statisticians, led by the American Statistical Association, recommend that we be guided by those values of P (Wasserstein and Lazar, 2016). If they are published instead of statements such as " $P \leq 0.05$ " along with the standard errors then readers can reach their own judgments.

Note finally that the null hypothesis is highly implausible when horizons and different types of soil are being compared; they are different.

Transformations

It is often desirable to transform data to their square roots, or logarithms or by other more elaborate functions. One reason for doing so is to obtain a new variate that approximates some known distribution, preferably normal (Gaussian) so that the usual parametric tests of significance can be applied.

The most serious departure from normality usually encountered with soil data is skewness, i.e., asymmetry, as in Fig. 3A. The normal distribution is symmetric, its mean is at its center (its mode), and the mean of the data estimates this central value without ambiguity. The mean of data from a skewed distribution does not estimate the mode, nor does the median (the central value in the data). The meaning of the statistics can be uncertain therefore. A second feature of skewed data is that the variances of subsets depend on their means. If the data are positively skewed (again the usual situation) then the variances increase with increasing mean. This is undesirable for comparisons. Third, estimation is "inefficient" where data are skewed; that is the errors are greater than they need be or, put another way, more data are required to achieve a given precision than would be if the distribution were normal. Transforming data to approximate normality overcomes these disadvantages. We achieve symmetry, and hence remove ambiguity concerning the center. We stabilize the variances. And we make estimation efficient. The second of these is perhaps the most important.

No real data are exactly normal; all deviate more or less from normality. We have therefore to decide whether to transform them. This is best done by judicious exploration of the data aided by graphic display.

Draw a histogram. If it looks symmetrical then superimpose on it a normal curve computed from the mean and variance of the data. The normal curve has the formula

$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{(z - \mu)^2}{2\sigma^2} \right\} \quad (10)$$

where y is the probability density. Scale it to fit the histogram by multiplying by the number of observations and by the width of the bins. If the curve fits well then there is no need to transform the data.

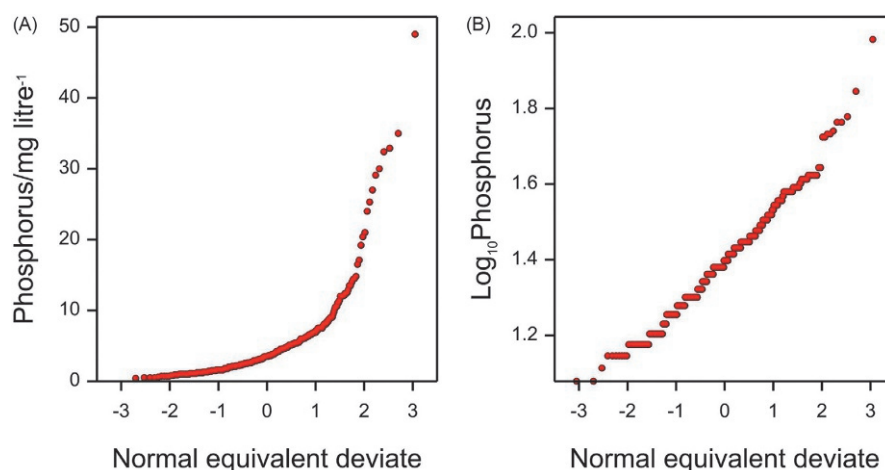


Fig. 4 Q–Q plots of 434 data on available phosphorus (A) in mg L^{-1} , and (B) transformed to common logarithms. The points in (A) lie close to a curve, signifying strong departure from normal, whereas those in (B) lie close to a straight line characteristic of a normal distribution.

Another popular way of examining data is a Q–Q plot. This is a graph of the cumulative distribution of the data plotted against the quantiles of a standard normal distribution. If the data are approximately normal then the points lie close to a straight line. Any appreciable curve signifies strong departure from normal. Fig. 4 shows the Q–Q plots for the same data as the histograms in Fig. 3.

If the histogram is skewed then compute the skewness coefficient in addition to the mean and variance. This dimensionless quantity can be obtained via the third moment of the data about their mean:

$$\gamma_1 = \frac{1}{N S^3} \sum_{i=1}^N (z_i - \bar{z})^3 \quad (11)$$

A symmetric histogram has $\gamma_1 = 0$. Values of γ_1 greater than 0 signify positive skewness, i.e., long upper tails to the distribution and a mean that exceeds the median, which is common. Negative values of γ_1 signify negative skewness, and are unusual.

If γ_1 is positive and less than 0.5 then there should be no need to transform the data. If $0.5 < \gamma_1 \leq 1$ then it might be desirable to convert the data to their square roots; and if $\gamma_1 > 1$ then a transformation to logarithms is likely to give approximate normality.

The following example illustrates the situation. The data, which are summarized in Table 1, are 433 measurements of available phosphorus, P, in the topsoil. Their skewness coefficient is 3.95, i.e., they are strongly positively skewed, and this is apparent in their histogram, Fig. 3A. Transforming to logarithms makes the histogram, Fig. 3B more nearly symmetric, and as the skewness in the logarithms is now only 0.34; the transformation seems satisfactory. Further, the normal curve appears to fit well. Fig. 4, the Q–Q plots of the data, conveys the same message: the curve on which the points lie in Fig. 4A becomes a straight line for the transformed data in Fig. 4B.

There are statistical tests for normality, but they are unhelpful. If data are numerous then small departures from normality lead to supposed significance; if they are few the results are inconclusive.

Fig. 5 shows how the transformation stabilizes the variances. In Fig. 5A the variances of subsets of 44 from the full set of data on phosphorus are plotted against the means. Evidently, the variance increases strongly with increasing mean. Converting the data to their logarithms produces a result in which there is virtually no relation, Fig. 5B.

These simple functions change only the general form of the distribution; they do not change the detail. Normalizing the detail requires a more elaborate normal score transform.

Table 1 Summary statistics of 433 values of available phosphorus measured in a survey of topsoil.

Statistic	Scale	
	P/mg L^{-1}	$\text{Log}_{10} \text{P}$
Mean	4.86	0.546
Variance	26.52	0.1142
Standard deviation	5.15	0.338
Skewness	3.95	0.23
$\chi^2_{\text{df}18}$	368.2	26.7

Values of χ^2 , with 18 degrees of freedom, are added for the hypothesis that the data or their logarithms are from a normal distribution.

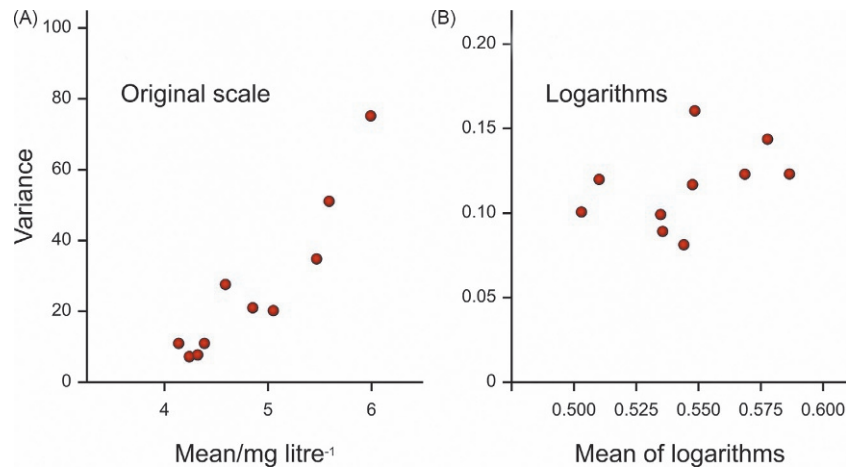


Fig. 5 Graphs of variance against mean for 10 sub-sets of 44 phosphorus data (left) on the original scale in mg L^{-1} and (right) after transformation to common logarithms.

Analysis of variance

The analysis of variance is at once one of the most powerful and elegant techniques in statistics. It and regression analysis are what may be regarded as the “bread-and-butter” of statistics in soil research. Its basis is that variances are additive and that the total variance in a population is the sum of the variances contributed by two or more sources. Working from the design of an investigation it analyzes the data by separating the contributions from those sources and estimating the variances in them. Designs vary from the simple to highly complex, but all embody the same principle.

Here we consider two of the simplest simple designs, such as appear in Fig. 2. Investigators want to know how manuring improves crop yield in the field. They have several (k) treatments, say, nothing (0), dung (D), industrial waste (W), and a complete artificial (NPK) fertilizer (F). They replicate each m times by assigning the treatments completely at random to plots of equal size. Fig. 2A shows how the experiment might be layed out with $m = 5$ replicates. They apply the treatments, grow the crop, and measure the yield at harvest, designated z .

The total variance in the yields in the experiment, s_T^2 , comprises variance from two sources, namely that between the treatments, s_B^2 , and that within them, s_W^2 , and $s_T^2 = s_B^2 + s_W^2$. The total variance is estimated by Eq. (4). The variance within any one treatment is estimated by the same formula but for only those data in that treatment. Pooling estimates for all treatments gives s_W^2 . To complete the analysis we compute also a quantity B :

$$B = \frac{1}{k-1} \sum_{i=1}^k n_i (\bar{z}_i - \bar{\bar{z}})^2 \quad (12)$$

where n_i is the number of plots in the i th treatment, $\bar{z}_i$ is the mean of the i th treatment, and $\bar{\bar{z}}$ is the general mean of the data. The computations are set out in Table 2. Finally, the analysis leads to a test of significance. We compute the ratio $F = B/W$, the distribution of which has been worked out and tabulated for degrees of freedom $k - 1$ in the numerator and $N - k$ in the denominator. If F exceeds the tabulated value at probability $P = 0.05$ (or $P = 0.01$ or $P = 0.001$, according to choice) we judge the treatments to have produced significant differences. As above, however, now that values of P can be calculated so readily we should its value for the ratio F .

In the simple experiment illustrated in Fig. 2A all n_i are equal to 5, so that n_i can be replaced by $n = 5$, and $N = mk = 5 \times 4 = 20$. Things do not always go as planned, however, and if some of the plot yields are lost then the n_i can vary from treatment to treatment, and the more general formulae in Table 2 will take care of that.

Table 2 Table for the analysis of variance for a completely randomized design with a single classification.

Source	Degrees of freedom	Sum of squares	Mean square	F
Treatments	$k - 1$	$\sum_{i=1}^k n_i (\bar{z}_i - \bar{\bar{z}})^2$	$\frac{1}{k-1} \sum_{i=1}^k n_i (\bar{z}_i - \bar{\bar{z}})^2 = B$	B/W
Residual	$N - k$	$\sum_{i=1}^k \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2$	$\frac{1}{N-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_i)^2 = W$	
Total	$N - 1$	$\sum_{i=1}^k \sum_{j=1}^{n_i} (z_{ij} - \bar{\bar{z}})^2$	$\frac{1}{N-1} \sum_{i=1}^k \sum_{j=1}^{n_i} (z_{ij} - \bar{\bar{z}})^2 = T$	

Table 3 Table for the analysis of variance for a balanced randomized block design with a single set of treatments.

Source	Degrees of freedom	Sum of squares	Mean square	F
Treatments	$k - 1$	$\sum_{i=1}^k n_i (\bar{z}_i - \bar{z})^2$	$\frac{1}{k-1} \sum_{i=1}^k n_i (\bar{z}_i - \bar{z})^2 = B$	B/W
Blocks	$m - 1$	$\sum_{l=1}^m n_l (\bar{z}_l - \bar{z})^2$	$\frac{1}{m-1} \sum_{l=1}^m n_l (\bar{z}_l - \bar{z})^2 = M$	
Residual	$(k - 1) \times (m - 1)$	$T(N - 1) - B(k - 1) - M(m - 1)$	$\frac{T(N - 1) - B(k - 1) - M(m - 1)}{(k - 1)(m - 1)} = W$	
Total	$N - 1$	$\sum_{i=1}^k \sum_{j=1}^m n_{ij} (z_{ij} - \bar{z})^2$	$\frac{1}{N-1} \sum_{i=1}^k \sum_{j=1}^m n_{ij} (z_{ij} - \bar{z})^2 = T$	

The soil might vary across the experiment systematically, so that there is trend, or in an apparently random way at a coarse scale. This variation could swell the residual variation and so mask that due to the treatments. It can be taken into account by blocking. The m replicates are now arranged in m blocks such that in each block every treatment appears once and once only. Fig. 2B shows an example in which five blocks are laid out side by side. The analysis follows the same procedure as in Table 2 except that there is an additional line for the blocks in which the sum of squares is that of deviations from the block means, as in Table 3. The residual sum of squares is diminished by this quantity, and although the residual degrees of freedom are also diminished the residual mean square, i.e., the residual variance is usually less, and the experiment more sensitive therefore.

Experiments should be planned with some hypotheses in view. The treatments should be chosen and arranged in designs to test them. The analysis that follows any experiment must fit its design; no other will do (Webster and Lark, 2018). Further, any comparisons between means should be ones that were set out at the start of the investigation; ideally they should be orthogonal to one another. The common practice of comparing all possible pairs of treatments afterwards is to be deprecated. Despite an investigator's designing an experiment well and analyzing the results in accord with nature may upset the investigator's unforeseen assumptions. Webster and Lark (2019) guide readers on what to do in those circumstances by transformations of the data.

In soil survey different classes of soil replace treatments—see Webster and Lark (2013). In the simplest cases each class is sampled at random, and the n_i are rarely equal, either because it is difficult to obtain equal representation or because we deliberately sample in proportion to area so as to maintain a fairly constant sampling density. Effectively the classes are weighted in proportion to the areas they cover. We can still compute $F = B/W$ and test for significance, but as above this is less interesting than the differences between the means.

The variance between treatments or classes, s_B^2 , can be obtained from B . The latter combines variation both from between treatments or classes and within them:

$$B = ns_B^2 + s_W^2 \quad (13)$$

if $n_i = n$ for all i . Rearranging then gives

$$s_B^2 = (B - s_W^2)/n \quad (14)$$

If the n_i are unequal then n is replaced by n^* , given by

$$n^* = \frac{1}{k-1} \left(N - \frac{\sum_{i=1}^k n_i^2}{N} \right) \quad (15)$$

and

$$s_B^2 = (B - s_W^2)/n^* \quad (16)$$

Fixed effects, random effects and intra-class correlation

The value s_B^2 obtained as above estimates $\sum_{i=1}^k (\mu_i - \mu)^2 / (k - 1)$, where μ_i is the expected value of treatment or class i , and μ is the expected value in the whole population. In designed experiments, in which the effects are fixed by the experimenter, it is of little interest. In soil survey, however, where it is often a matter of chance which classes are actually sampled, the differences $\mu_i - \mu$ are subject to random fluctuation. In this event s_B^2 estimates the variance, σ_B^2 , among a larger population of means and is termed a component of variance, and it is of considerable interest.

The between-class variance expressed as a proportion of the total variance, $\sigma_B^2 + \sigma_W^2$, is the intraclass correlation, ρ_i :

$$\rho_i = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} \quad (17)$$

which is estimated from the analysis variance table by

$$r_i = \frac{s_B^2}{s_B^2 + s_W^2} = \frac{B - W}{B + (n * -1)W} \quad (18)$$

The intra-class correlation has a theoretical maximum of 1 when every class is uniform. In practice there is always some variation within classes, and so ρ_i never attains 1. The minimum of ρ_i is zero, when $s_W^2 = 0$. The calculated estimate of ρ_i can be negative, because $B < W$, and is usually best explained by sampling fluctuation.

Covariance, correlation and regression

The relations between two variables can be expressed by correlation and regression.

Covariance

The covariance of a pair of variables, y and z , is estimated from data in a way analogous to the estimation of the variance, Eq. (4), by

$$\hat{c}_{y,z} = \frac{1}{N-1} \sum_{i=1}^N \{y_i - \bar{y}\} \{z_i - \bar{z}\} \quad (19)$$

where $\bar{y}$ and $\bar{z}$ are the means of y and z respectively. It is not easy to envisage, especially if y and z have different dimensions. The relation may be standardized by its conversion to correlation, below.

Correlation

The correlation between two variables, strictly the linear correlation, or the product-moment coefficient of linear correlation, is a dimensionless quantity, usually denoted by ρ for the population parameter. Its estimate, r , is obtained from the covariance by

$$r_{y,z} = \frac{\hat{c}_{y,z}}{\sqrt{s_y^2 s_z^2}} \quad (20)$$

s_y^2 and s_z^2 are the estimated variances of y and z . It was proposed by Karl Pearson, and for that reason it is often called the Pearson correlation coefficient.

The coefficient is effectively a standardized version of the covariance. It measures the extent to which the data when plotted as one variable against the other in a scatter graph depart from a straight line; see Fig. 6. It may vary between +1, signifying perfect positive correlation, and -1 for perfect negative correlation. Intermediate values indicate departures from the straight line, as in Fig. 6, for which $r = 0.699$. In general positive values of r indicate the tendency of y and z to increase together, whereas negative values arise when y increases as z decreases. A value of 0 represents no linear relation.

Notice that the statistic refers specifically to linear correlation. The relation between two variables might be curved; the absolute value of r would then be necessarily less than 1 regardless of any scatter about the curve.

When the data, $y_i, z_i, i = 1, 2, \dots, N$, are from a sample then $c_{y,z}$ and $r_{y,z}$ estimate corresponding population parameters, $\text{cov}_{y,z}$ and $\rho_{y,z}$. If the data can be assumed to be drawn from a bivariate normal distribution then one can test r for significance. One computes Student's t with $N - 2$ degrees of freedom:

$$t_{N-2} = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \quad (21)$$

The probability of this value's occurring on the null hypothesis that $\rho = 0$ can then be computed or looked up in a table of t .

Spearman rank correlation

Where the distributions of the underlying variables are far from normal the Pearson coefficient can be replaced by Spearman's rank correlation coefficient, usually denoted r_s . The values of each variable are ranked from smallest to largest and given new values $1, 2, \dots, N$. The correlation coefficient is then computed by applying Eqs. (19) and (20). Alternatively, one may take the differences, $d_i, i = 1, 2, \dots, N$, between the ranks and compute

$$r_s = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2 - 1)} \quad (22)$$

Many soil variables observed in the field, such as grade of structure and frequency of mottles, are recorded as rankings rather than measured. In these circumstances the correlations between them can be expressed by the Spearman coefficient, whereas

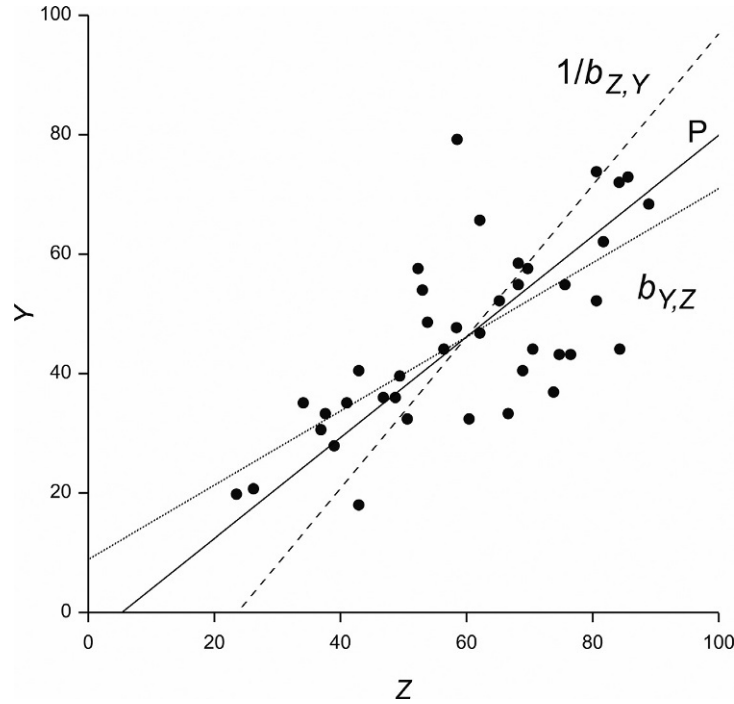


Fig. 6 Scatter graph showing the relations between two variables Y and Z for which the correlation coefficient, r , is 0.699. The symbols are the observed values, the solid line, labeled P , is the principal axis with gradient 0.844, equivalent to an angle of 40.1° to the horizontal. The dashed line (---) shows the regression of Y on Z with gradient $b_{Y,Z} = 0.621$, and the dotted line (....) shows that of Z on Y , for which $b_{Z,Y} = 0.788$, giving it a gradient in the figure of $1/b_{Z,Y} = 1.269$.

the Pearson coefficient would be inappropriate. Also in these circumstances tied ranks in any large set of data are inevitable, and Eq. (22) must be elaborated. The coefficient can be calculated in various ways, but from Eqs. (19), (20) and (22) we can derive

$$r_s = \frac{\sum_{i=1}^N (y_i - \bar{y})^2 - \sum_{i=1}^N T_{i,y} + \sum_{i=1}^N (z_i - \bar{z})^2 - \sum_{i=1}^N T_{i,z} - \sum_{i=1}^N d^2}{2\sqrt{\left\{ (y_i - \bar{y})^2 - \sum_{i=1}^N T_{i,y} \right\} \left\{ (z_i - \bar{z})^2 - \sum_{i=1}^N T_{i,z} \right\}}} \quad (23)$$

in which

$$T_i = \frac{t_i(t_i^2 - 1)}{12} \quad (24)$$

where t_i is the number of observations tied at rank i .

For small samples r_s is somewhat less sensitive than the Pearson coefficient in that larger values are necessary to establish statistical significance.

Regression

Regression, which is dealt with comprehensively by Draper and Smith (1981), treats the relation between two variables in a somewhat different way by designating one of them, y , as depending on the other, z , represented by the equation:

$$y = \beta_0 + \beta_1 z \quad (25)$$

The underlying rationale is often physical. For example, we may wish to discover how the soil's strength is changed by additions of gypsum. We could add known amounts of gypsum to the soil, measure the resultant changes in strength, and from the data estimate by how much on average the strength is changed by each increment in gypsum added. We adopt the Gauss linear model for the purpose:

$$y_i = \beta_0 + \beta_1 z_i + \varepsilon_i \quad (26)$$

where y_i is the value of the random dependent variable, Y , here strength, in unit i , z_i is that of the independent variable, gypsum added, and ε_i is random error term that is uniformly and independently distributed with variance σ_ε^2 . The quantities β_0 and β_1 are parameters of the model and are estimated as follows:

$$\hat{\beta}_1 = \frac{\hat{c}_{y,z}}{s_z^2} \quad (27)$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{z} \quad (28)$$

Eq. (27) gives the rate at which strength changes in response to increments in gypsum. The quantity $\hat{\beta}_0$ in Eq. (28) is the intercept at $y = 0$ and is likely to be of subsidiary interest. Together they may be inserted into Eq. (25) for predicting unknown values of Y if we know those of z :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_{1z} \quad (29)$$

The procedure minimizes the sum of the squares of the differences between the measured values y_i ; $i = 1, 2, \dots, N$, and those expected from Eq. (25), $\hat{y}_i$:

$$s_{y,z}^2 = \frac{1}{N-2} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (30)$$

A somewhat different situation is common in soil survey. In it we sample the soil without knowing in advance what values we shall obtain for the variables of interest. For example, we may measure both cation exchange capacity, CEC, and clay content, and we may regard CEC as depending in a physical sense on clay content. The data on both variables now contain random components, and for this reason we designate them with the capital letters, Y and Z , respectively. We may express the relation as the regression of CEC on clay and estimate the parameters in the same way as for the Gauss linear model. In doing so we assign all the error to CEC and minimize the sum of squares, $s_{y,z}^2$, as in Eq. (30). The purpose is now prediction, i.e., prediction of CEC, knowing the clay content. We could equally well compute the regression of clay on CEC. The roles of Y and Z are reversed, and we minimize

$$s_{z,y}^2 = \frac{1}{N-2} \sum_{i=1}^N (z_i - \hat{z}_i)^2 \quad (31)$$

where

$$\hat{z} = \hat{\beta}'_0 + \hat{\beta}'_{1y} y \quad (32)$$

The primes attached to β'_0 and β'_{1y} signify that these quantities refer to the regression of Z on Y and that they are different from those for the regression of Y on Z . In other words, the line defined by Eq. (29) differs from that defined by Eq. (32), as Fig. 5 shows.

The correct line to choose depends on which variable is to be predicted. To predict y from z use Eq. (29); to predict z from y use Eq. (32).

Further matter

Research into the soil and modern developments in instrumentation are leading to ever more advanced statistical understanding and technique. In recent years pedologists have mastered the theory of regionalized variables to model and map spatially correlated soil variables using geostatistics; see the chapter on this topic. Soil biologists are measuring numerous properties of the soil's microora and its genetics and are analyzing the data by various multivariate techniques to identify relations, to group organisms and their functions. Multivariate analysis is also treated as a topic in the encyclopedia. Sensing in the visible and infrared parts of the electromagnetic spectrum, from the air and satellites, from close to the soil surface (proximal sensing) and in the laboratory is producing huge quantities of multivariate data, much of which is redundant. Techniques, including machine-learning, have been developed to compress and smooth the spectral data and extract features that have mineralogical, chemical and biological meaning, and to predict properties that are much more time-consuming and expensive to measure by earlier conventional methods. Machine learning is covered in another chapter. Note, however, that these techniques do not replace or invalidate the sound experimental design, random sampling and the appropriate analysis of the data thereby obtained.

References

- Brus DJ (2021) Statistical approaches for spatial sample survey: Persistent misconceptions and new developments. *European Journal of Soil Science* 72: 686–703.
- Cochran WG (1977) *Sampling Techniques*, 3rd edn. New York: John Wiley & Sons.
- de Groot JJ, Brus DJ, Bierkens MFP, and Kotters M (2006) *Sampling for Natural Resources Monitoring*. Berlin: Springer-Verlag.
- Draper NR and Smith H (1981) *Applied Regression Analysis*, 2nd edn. New York: John Wiley & Sons.
- Snedecor GW and Cochran WG (1989) *Statistical Methods*, 8th edn. Ames, Iowa: Iowa State University Press.
- Sokal RR and Rohlf FJ (2012) *Biometry: the Principles and Practice of Statistical Research*, 4th edn. San Francisco: W.H. Freeman.
- Wasserstein RL and Lazar NA (2016) The ASA's statement on p-values: Context, process and purpose. *The American Statistician* 70: 129–133.
- Webster R and Lark RM (2013) *Field Sampling for Environmental Science and Management*. London: Routledge.
- Webster R and Lark RM (2018) Analysis of variance in soil Research: Let the analysis fit the design. *European Journal of Soil Science* 69: 126–139.
- Webster R and Lark RM (2019) Analysis of variance in soil research: Examining the assumptions. *European Journal of Soil Science* 70: 990–1000.
- Welham SJ, Gezan SA, Clark SJ, and Mead A (2015) *Statistical Methods in Biology: Design and Analysis of Experiments and Regression*. Boca Raton, Florida: CRC Press, Taylor and Francis Group.

Multivariate analysis for soil science

Si Yang Han^{a,b}, Patrick Filippi^{a,b}, Mercedes Román Dobarco^{a,b}, Januar Harianto^{a,b,c}, Mathew S Crowther^{a,b}, and Thomas FA Bishop^{a,b,c}, ^aSydney Institute of Agriculture, The University of Sydney, Camperdown, NSW, Australia; ^bSchool of Life and Environmental Sciences, The University of Sydney, Camperdown, NSW, Australia; ^cSydney Informatics Hub, The University of Sydney, Camperdown, NSW, Australia

© 2023 Elsevier Ltd. All rights reserved.

Introduction	499
Dimension reduction	501
Principal components	501
Theory	501
Application	501
An alternative	502
Supervised classification	503
Linear discriminant analysis	503
Theory	503
Application	504
An alternative	505
Unsupervised classification	505
<i>k</i> -means clustering	505
Theory	505
Applications	505
An alternative	506
Conclusions	506
References	507

Abstract

With the growing surplus of data available for modelling in soil science, multivariate analyses offer a path to understand the patterns in large datasets with 100s or 1000s of variables. In this chapter, we explore three classes of multivariate techniques and analysis, and a case study using one technique for each class: dimension reduction with principal component analysis (PCA), supervised classification with linear discriminant analysis (LDA) and unsupervised classification with *k*-means clustering. While by no means comprehensive, we demonstrate potential uses for multivariate analysis in soil in agricultural and ecological contexts. In addition, we provide some examples of alternatives for each of the classes.

Key points

- Multivariate analyses are increasingly needed by soil scientists due to the increasing ease of data collection.
- Principal component analysis can be used for dimension reduction to simplify spectral data or a datasets with many variables and nullify multicollinearity.
- Discriminant analysis is used for supervised classification to differentiate soil orders or soil types.
- *k*-means clustering is a form of unsupervised classification commonly used to group or partition soil spectra in libraries, or similar types of soil which enables stratified soil sampling.

Introduction

We live in an age where it has never been easier to measure properties on a soil sample, whether in the field or in the laboratory. For example, routine wet chemistry-based laboratory testing of soil for agronomic purposes can generate reports outlining well over 10 soil properties for consideration. This can be further accentuated with soil spectroscopy in a field or laboratory setting with technologies such as visible near-infrared (vis-NIR) spectroscopy that generate 100s of spectral measurements at very fine wavelengths (Fig. 1). More recently, advances in soil biological analysis have made it possible to generate 1000s of DNA measurements based on soil microbes.

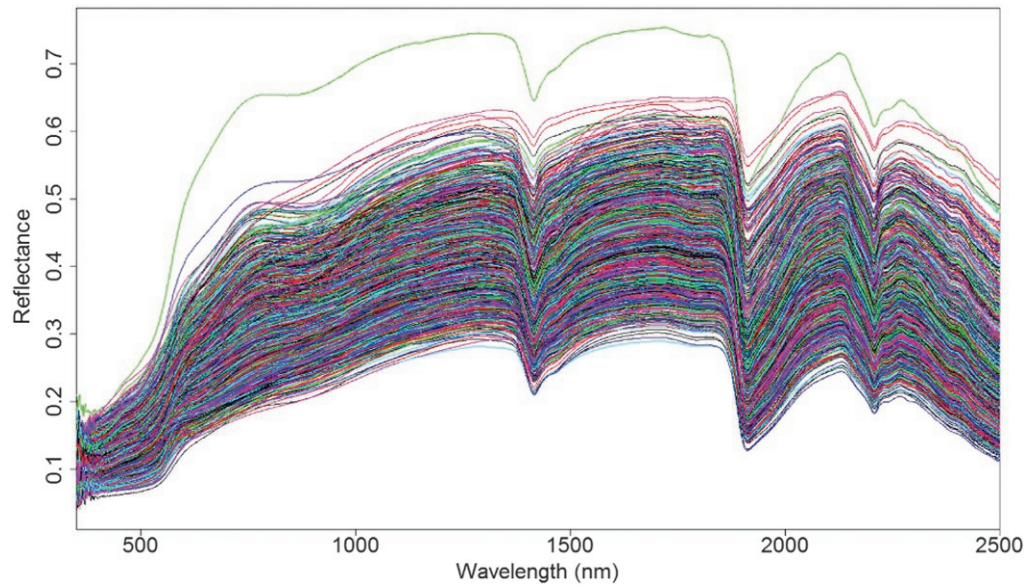


Fig. 1 An example of soil spectral vis-NIR measurements for 480 soil samples.

All of these data can be generated from one soil sample, but in many cases data are collected at depth and spatially, further increasing the amount of data that are available for analysis and interpretation from one study area. In the spatial context there is also the increase in readily available geospatial datasets, which can be related to soil. Key examples using these in spatial soil studies are digital terrain and remote sensing data. Nowadays, these datasets are free and openly available to the public for most of the globe. Key examples include a global digital elevation model (DEM) from a Shuttle Mission in 1999 and more recently the Copernicus Programme DEM, as well as the Landsat and Sentinel remote sensing products. From one digital elevation model, 10 or more terrain attributes can be generated to represent soil variation. The same is true for remote sensing data, which have individual bands but also spectral indices such as the normalized difference vegetation index (NDVI) that can be used. This means once again for each satellite product we can have 10 or more attributes available for analysis.

For the results from conventional wet chemistry laboratory analysis, it is relatively easy to interpret the measurements. The situation becomes more abstract, however, when trying to find meaning, relationships and patterns in datasets such as soil reflectance obtained from vis-NIR spectroscopic measurements due to both the size of the dataset and because the measurement is not directly related to one soil property only. The same difficulty exists when trying to relate terrain attributes or remote sensing measurements to soil properties. This is where multivariate analysis can be useful as a way to understand patterns in large multivariate soil datasets by reducing the number of variables in what is called dimension reduction and also for grouping samples together in what is called classification.

In many cases the soil scientist wishes to go beyond exploratory data analysis and aims to build some sort of predictive model; two examples are soil spectral inference and digital soil mapping. In soil spectral inference the starting point could be 200+ bands per sample to create a model. In digital soil mapping we can easily start with 40+ predictors to consider when designing a sampling scheme or building a spatial model. Not all of the data are useful for understanding the soil condition and variability, and while simple correlation analysis is useful, as the number of variables increases it becomes more difficult to identify useful patterns and variables, especially as many are correlated. In these cases, it is important to select useful features from the dataset, which is where dimension reduction techniques are advantageous and the model builder has fewer variables to consider, and each is independent of the other.

In this chapter we present key multivariate methods of analysis that focus on three general approaches:

- 1) dimension reduction, where the soil scientist aims to simplify a multivariate dataset of (most likely) correlated variables to a small set of variables, all of which are independent;
- 2) unsupervised classification, where the soil scientist aims to group multivariate observations into groups of strong similarity but does not have a variable such as soil type that can act as a natural grouping variable;
- 3) supervised classification, where the soil scientist has a variable such as soil type that can act as a natural grouping variable but aims to build a classification model based on a set of predictor variables that are typically easier to acquire than the grouping variable.

For each approach we present the basic theory and examples of applications from the literature, and also analysis with a soil dataset to illustrate the practical implementation of the method. We conclude each section with commentary about less used multivariate methods that may be useful for soil scientists.

Dimension reduction

Principal components

Theory

Principal component analysis (PCA) is generally used as an ordination technique in exploratory data analysis, but can be used to make decisions in predictive models. One common use for PCA is when predictors are strongly correlated, which can cause issues of multicollinearity if modelled. Principal component analysis can be used to transform the set of x correlated continuous variables over y samples to a set of p uncorrelated variables over the same samples.

The variables are standardized before deriving a covariance matrix, from which eigenvectors and their corresponding eigenvalues are then calculated. The eigenvectors are the directions of the axes, and the eigenvalues are the coefficients that reflect the proportion of the total variance each axis explains. A PCA involves rotating the original axes of the data to maximize the variance in the dataset. The data projected on to the new, rotated axes are principal component (PC) scores, and each dimension is an uncorrelated variable or PC. The first PC has the largest eigenvalue and explains the most variance, and each subsequent coordinate is orthogonal or perpendicular (i.e. uncorrelated) to the last. Thus, the most informative PC is the first, and the least informative is the last. As such, less informative PCs can be culled according to a scree test (Fig. 2), resulting in dimension reduction. The first few PCs combine and compress the information of all the variables, albeit at the expense of the interpretability of each predictor to the prediction of the response variable.

Principal component analysis does not require statistically normal data, though outliers and missing data should be removed, or substituted with the mean or median value of a variable to ensure sufficient data remain for analysis. The assumptions for PCA are relatively easily met. The first assumption, often called sphericity, requires the variables to have some degree of correlation for the analysis to be useful. This assumption is true if the problem of multicollinearity is the reason for performing PCA. The second assumption is that there should be an adequate number of samples per variable to ensure reliable and stable results, and many multivariate textbooks cite a 20:1 ratio. In the age of big data, multivariate datasets are often substantial and meeting this assumption is not a problem. The third assumption is that the determinant of the correlation matrix must be positive. In general it means that there should be unpredictability in the data such that it cannot be exactly predicted in the linear combination of variables. This assumption is rarely tested since failure to meet this assumption will result in failure to compute PCs, leading to an error.

Application

In the following example, we analyzed soil samples from various sites across The University of Sydney farm *L'lara*, an 1850 ha mixed farming property with Gray or Brown Vertosols (Isbell, 2016) in northern New South Wales (NSW), Australia. The farm was sampled in three separate campaigns in July 2017, July 2018, and August 2019 (Llara 1–3); Filippi et al. (2019) describe the earlier

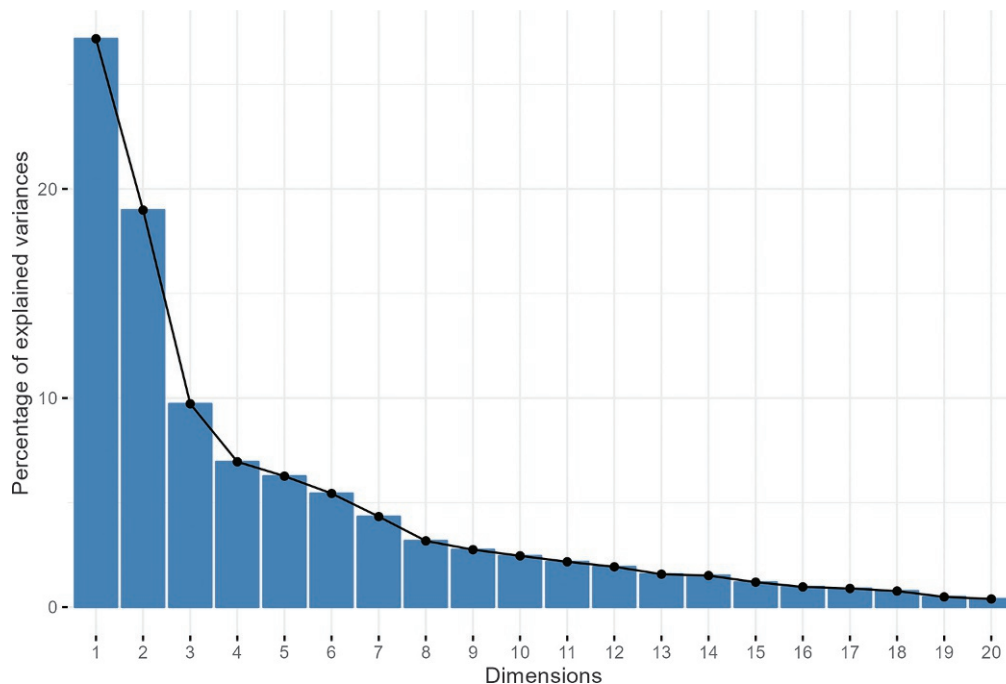


Fig. 2 Elbow or scree plot of the eigenvalues of the first 20 principal components.

Table 1 List of publicly available national datasets and proximally sensed imagery available for The University of Sydney farm L'lara, NSW, Australia.

Dataset	Data Source	Resolution	Term
Terrain Attributes	CSIRO Data Portal (Gallant and Austin, 2015)	30 m	Digital elevation model (DEM_30), aspect, multi-res ridge top flatness, multi-res valley bottom flatness, slope percent, plan curvature, profile curvature, slope in degrees (SlopeDeg), slope in percent, slope relief, topographic position index, total wetness index
Radiometrics	Geoscience Australia (Minty et al., 2009)	100 m	Radiation adsorbed dose; gamma radiometric potassium (K) (rad_k), uranium (U), thorium (Th), Th:K, U:Th ratio, U ² :Th ratio and U:K ratio
Weather	CSIRO Data Portal (Xu and Hutchinson, 2011)	30 m	ANUCLIM annual mean temperature 1976–2005, ANUCLIM annual mean rainfall 1976–2005
Satellite Imagery	USGS (Landsat 7)	30 m	2000–2020 normalized difference vegetation index percentiles (NDVI_5, NDVI_50, NDVI_95)
Other	Gray et al. (2016)	100 m	Silica index
Proximal Sensing Imagery		5 m	Gamma radiometrics potassium (gamma_k), apparent electrical conductivity (EM)

The abbreviations given are used in Fig. 3.

dataset and Jones et al. (2021) the later ones. The dataset also includes a range of proximally sensed as well as national and publicly available predictor variables, listed in Table 1.

A total of 27 numerical predictors were extracted for the L'lara dataset (Table 1). Taking its size into account, this dataset contains too many predictors for a regression model and could cause model overfitting. In addition, a correlation matrix of these predictors had multiple Pearson correlations above 0.80, namely the gamma radiometrics with their respective ratios, and the national and proximally sensed gamma radiometrics.

After applying a PCA on all 27 predictors, the subsequent PCs were uncorrelated with each other. There were, however, still 27 PCs and there are several methods to determine how many PCs to keep or discard.

An elbow or scree plot (Fig. 2) of the eigenvalues of each PC can be used to determine the number of PCs to keep. In this case the eigenvalues, or explained variances, are small for the 10th PC onward (i.e. the forearm of the curve) and would offer minimal additional value to a model.

A less subjective selection process is to retain all PCs with an eigenvalue above 1 (Kaiser's Criterion), which in this case includes the first 7 PCs. This is because an eigenvalue greater than 1 accounts for more than one variable's amount of variance. Another approach is to set a threshold of cumulative variance percent, commonly 75%, 80% or 90%, and select the number of PCs to retain accordingly. In this case keeping the first 11 PCs explains 92% of the total variance in the predictors, and a further reduction to 7 PCs retains 80% variance and 6 PCs retains 75%. Even with a conservative approach, a dimension reduction of 27–11 predictors can be carried out with PCA, while maintaining 90% of the total predictor variance.

Principal component analysis can also be used as a standalone data exploration technique. Three separate sampling campaigns were conducted for the L'lara farm property. Llara 1 and Llara 2 samples were taken from cropped land only, whereas the scope of the Llara 3 sampling campaign also covered grazing land. By conducting a PCA analysis, we can also determine whether the locations of each sampling campaign were covered by a similar range of values for the numerous predictor variables, or predictor variable space. This is important for cases where we wish to use one dataset for developing a spatial model and another for validating the spatial model independently; this determines the occurrence, or to some degree, the extent of extrapolation. We therefore need to ensure that we are testing the model in its appropriate prediction domain.

For this example, a simpler set of 6 predictor variables (labeled in Fig. 3) was analyzed with PCA. The first 3 PCs explained over 90% of the cumulative variance of the variables. Fig. 3 is a biplot which shows the scores for PC1 and PC3 as points and the 95% confidence interval for each sampling campaign's group centroid as ellipses. Overlapping ellipses indicate the samples from the Llara 1 and 2 campaigns (red and green in Fig. 3, respectively) were covered by roughly the same covariate space, whereas Llara 3 (seen in blue in Fig. 3) had a wider attribute space. This suggests if the Llara 3 dataset is used for independent validation, it could result in extrapolation beyond what the model is capable of. Poor data splitting for validation is a common issue in large scale spatial modelling, which can be identified with a PCA.

Overall, PCA is a flexible data exploration technique that can be used to counter multicollinearity, achieve dimension reduction and compare datasets in terms of the attribute space they cover. There are countless ways to use the output of the analysis to observe trends, clusters and outliers.

An alternative

An alternative to PCA is non-metric multidimensional scaling (nMDS) which produces an ordination based on a distance or dissimilarity matrix. Unlike methods which attempt to maximize the variance between objects in an ordination (e.g. as in PCA), nMDS attempts to represent, as closely as possible, the pairwise dissimilarity between objects in a reduced-dimensional space. It is also a rank-based approach (hence non-metric). This means that the original distance data are transformed into rank data. It also means that information about the magnitude of distances between objects is no longer available, but as a rank-based technique, it is robust for non-linear data or data with any statistical distribution, i.e. no assumptions about the data are necessary. It can (1) tolerate missing pairwise distances, (2) be applied to a (dis)similarity matrix formed with any (dis)similarity measure and (3) use quantitative, semi-quantitative, qualitative, or mixed variables. The output of nMDS, sample scores, can be visualized with a biplot, similar to PCA, to highlight patterns between variables.

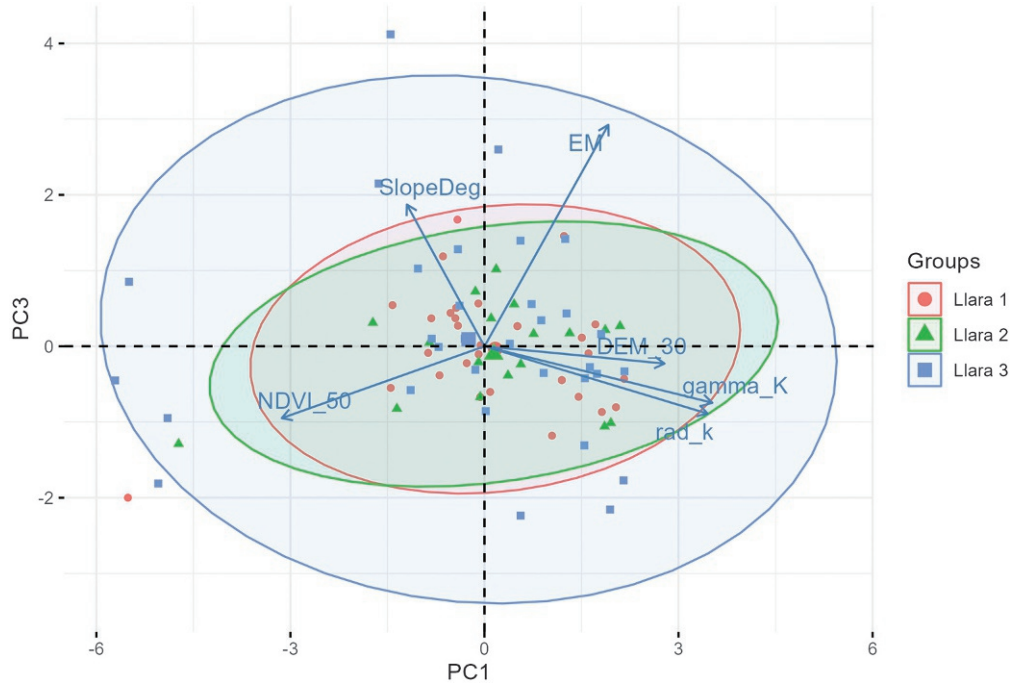


Fig. 3 Biplot of PC1 and PC3 grouped by sampling campaigns on The University of Sydney farm L'lara, NSW, Australia. The ellipses are the boundaries of the 95% confidence interval for each campaign. Descriptions for variables are in Table 1.

Supervised classification

Linear discriminant analysis

Theory

Soil scientists may be interested in identifying soil attributes that best explain differences between two soil classes or entities, or the environmental factors that explain the patterns of distribution of soil classes. In this case, the interest relies on explaining differences between well-defined classes. On the other hand, we may be interested in assigning or predicting the membership of an individual (e.g., soil profile or soil sample) of unknown membership into a predefined group. Discriminant analysis can be used for dimension reduction, descriptive analysis or classification, but here we emphasize its use for classification.

Linear discriminant analysis (LDA) is a generalization of 'Fisher's linear discriminant functions' (Fisher, 1936). Fisher formulated discriminant functions as linear functions of attributes that best discriminate individuals from two different populations or two classes. The discriminant functions can be represented as:

$$\mathbf{X} = a_1 \mathbf{x}_1 + a_2 \mathbf{x}_2 + \dots + a_d \mathbf{x}_d \quad (1)$$

Where $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_d$ are measurements of d attributes or features across the total number of individuals, and $a_1, a_2, a_3, \dots, a_d$ are the coefficients of the linear functions. The coefficients of the linear functions maximize the ratio of the difference between the class means (class centroid) to the standard deviation within the class, that is, the between-class sum of squares to the within-class sum of squares. In the two-class case, we begin by computing the mean of each class, $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$. Then, the between-class scatter matrix is calculated as $\mathbf{S}_B = (\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1)^T$, and also the within-class scatter matrix $\mathbf{S}_W$:

$$\mathbf{S}_W = \sum_{i \in C_1} (\mathbf{x}_i - \bar{\mathbf{x}}_1)(\mathbf{x}_i - \bar{\mathbf{x}}_1)^T + \sum_{i \in C_2} (\mathbf{x}_i - \bar{\mathbf{x}}_2)(\mathbf{x}_i - \bar{\mathbf{x}}_2)^T, \quad (2)$$

where C_1 and C_2 are classes 1 and 2 respectively. The solution to the coefficients is found by maximizing the quotient between the between-class variance to the within-class variance:

$$J(\mathbf{a}) = \frac{\mathbf{a}^T \mathbf{S}_B \mathbf{a}}{\mathbf{a}^T \mathbf{S}_W \mathbf{a}}, \quad (3)$$

which leads to

$$\mathbf{a} = \mathbf{S}_W^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2). \quad (4)$$

Rao (1948) generalized the method for multiple classes ($k > 2$) using $k-1$ discriminant functions. As with MANOVA (multivariate analysis of variance), there are some assumptions concerning the data for performing LDA: (1) the independent variables must have

a normal distribution at each class level (multivariate normality), and (2) the multiple variance-covariance matrices are homogeneous (homoscedasticity). The equality of covariances is commonly tested with Box's M test (*F*-ratio test) (Box, 1949) or a similar multivariate test of equal group dispersions. When the covariances are not considered equal, quadratic discriminant analysis can be used in place of LDA. Another limitation of LDA is that it does not handle collinearity of predictor variables well, but a dimension reduction technique like PCA can be used as a preliminary step to create latent independent variables.

Samples for classification LDA should represent a single random sample or a mixture of two or more populations (classes), or several independent, random samples of two or more populations (McGarigal et al., 2013). The ideal sample size will depend on the homogeneity of the soils and soil properties, so that the dispersion is estimated accurately. The minimum rule is that there should be at least two more observations than predictor variables, and at least two observations per class. Williams and Titus (1988) recommended several observations per group, at least three times the number of predictor variables.

Application

The first application of discriminant analysis in soil science was by Cox and Martin (1937), who used it to find the relevant soil properties for predicting the presence of *Azotobacter* in the soil. Oertel (1961) developed Fisher's discriminant functions to allocate soil profiles into either one of two soil classes. Several studies applied discriminant analysis to establish and test differences between soil classes (Hughes and Lindley, 1955) or allocate soil profile observations into predefined soil classes in soil survey studies (Webster and Burrough, 1974). Linear discriminant analysis is often used as a classification model for digital soil mapping (Sun et al., 2011). Soil spectral data (e.g., vis-NIR) are also used as independent variables for discriminating soil types (Lazaar et al., 2019). In chemometrics, LDA can be applied for validating the forensic classification of soil samples, as a tool in forensic soil science for linking evidence to crime scenes (Lee et al., 2012). Canonical variate analysis, or multiple discriminant analysis, has been applied to discriminate the legacy of vegetation and land use on soils using soil near-infrared (NIR) (Ertlen et al., 2015) and mid-infrared (MIR) spectral data (Dudek et al., 2021).

Forest species influence the quantity and chemical composition of the organic matter that enters the soil, which in interaction with environmental factors and the microbial community, and the successive transformations of soil organic matter may lead to distinctive spectral signatures of forest species. In the following example, we use data from Román Dobarco et al. (2021a) to test whether LDA can differentiate and predict the overstory species for the origin of the soil organic matter.

The dataset consists of 70 spectra of soils sampled at 0–15 cm depth beneath trembling aspen (*Populus tremuloides* Michx.) ($n = 27$), mixed ($n = 23$) and conifer (*Abies lasiocarpa* (Hook.) Nutt. and *Pseudotsuga menziesii* (Mirbel) Franco) ($n = 20$) stands. Soil samples were subject to fractionation to separate soil organic carbon associated with mineral particles in the fine fraction ($<53 \mu\text{m}$), and a light fraction (LF) (Román Dobarco et al., 2021a). The LF is predominantly composed of free and intra-aggregate particulate organic matter, consisting mainly of partly decomposed plant material. The LF samples were scanned with Fourier transform infrared spectroscopy in the MIR range ($4000\text{--}500 \text{ cm}^{-1}$). A scaled PCA was first applied to remove the multicollinearity from the spectral data, and the scores of the first 6 principal components that explain 96% of the variance were used as predictor variables.

The LDA had an overall accuracy of 63% and a kappa statistic of 0.43, which indicates some agreement between predictions and observations. In the confusion matrix (Table 2) we observe that most of the misclassified aspen LF samples were assigned to the mixed class ($n = 6$). The user's accuracy was similar for aspen and conifer (67%), indicating that the probability of the predictions for representing the real class was reasonably acceptable for these two classes. Mixed LF had the lowest user's and producer's accuracy, suggesting that the chemistry of the LF under mixed canopies may too heterogeneous and variable to discriminate from the other two classes, or too similar to the LF of aspen and conifer as it originates from both vegetation types (Fig. 3). User's accuracy refers to the probability that a prediction from the LDA model represents that category, or a measure of commission error. It is calculated as the total number of correct predictions of a class divided by the total number of predictions in that category (e.g., 12 predicted and observed mixed/22 predicted mixed). Producer's accuracy is calculated as the total number of correct predictions of a class divided by the total number of observations of that class and indicates the probability of an observation being correctly classified, or a measure of omission error (e.g., 12 predicted and observed conifer/20 observed conifer). The LDA model did not completely succeed in distinguishing the forest species that contributed to the LF (Fig. 3), suggesting that despite initial differences between the chemical composition of aspen litter and conifer needles, microbial activity and site differences (e.g., microclimate, soil dominant mineralogy and soil type) had caused some convergence in the chemistry of the LF at that stage of decomposition.

Table 2 Confusion matrix with the results of a LDA for classification of forest species-derived light fraction.

		Observed			Total	User's Accuracy (%)
		Aspen	Mixed	Conifer		
Predicted	Aspen	20	6	4	30	67
	Mixed	6	12	4	22	55
	Conifer	1	5	12	18	67
	Total	27	23	20	70	
	Producer's Accuracy (%)	74	52	60		

An alternative

Here, we have presented LDA as a supervised classification problem but there are many alternatives that can be used, especially with the easily accessible data science tools available to the modern soil scientist (multiple platforms and software for data storage, querying, analysis and visualization). One example is random forests, which are based on a forest of decision trees, and unlike LDA are non-parametric methods. Decision trees can be used for continuous outcomes called regression trees or categorical outcomes, called classification trees. As a whole, they are called classification and regression tree (CART) models. As techniques, they are easy to use and interpret, but their power becomes evident when they are used in ensemble models, such as bagged (bootstrap aggregating) trees and random forests. Compared to LDA, they are particularly useful when the relation between the response and predictor variables is complex, and includes non-linearities and interactions between predictors. For a classification problem it is best to compare different methods for a particular study.

Unsupervised classification

k-means clustering

Multivariate hard *k*-means clustering is a common technique implemented in soil science. This approach is used to allocate samples or data into groups based on their characteristics. Samples within each group are based on having similar characteristics. One common example of applying *k*-means clustering is to create spatial zones of similarity across a study area, such as an agricultural field, or a hydrological catchment or watershed. Another common example is in the field of soil spectroscopy, where soil samples are grouped based on their spectral reflectance derived from sensors such as vis-NIR, and MIR.

Theory

The *k*-means is a hard cluster algorithm that partitions data points into clusters, where data points in each cluster are more similar to each other than data points in different clusters (MacQueen, 1967). It uses the clustering error criterion, which calculates the squared distance for each point from the cluster center, and then sums these distances for all points in the data set (Likas et al., 2003). The goal of the objective function is to minimize the sum of squares error of the Euclidean distance over all data points (Heil et al., 2019).

If a random point is $\mathbf{x}_i = (x_1, x_2)$ and the initial cluster center is $\mathbf{c}_i = (c_1, c_2)$, then the Euclidean distance between the point and cluster center, $d(\mathbf{x}_i, \mathbf{c}_i)$, can be given by;

$$d(\mathbf{x}_i, \mathbf{c}_i) = \sqrt{(c_1 - x_1)^2 + (c_2 - x_2)^2}. \quad (5)$$

The cluster center, $\mathbf{c}_i$, assigned to a set of points is recalculated after points are reassigned to clusters by the minimum sum of squares error distance. The algorithm continues until this is minimized for a particular number of clusters.

The *k*-means clustering requires that a specific number of clusters to be created is specified (Román Dobarco et al., 2021b). A common way to select this is the 'elbow method'. In this approach, the optimal number of clusters is indicated by the turning point in the curve of total within-cluster sum of squared errors (WCSS) with regard to the number of clusters, after which the marginal effect on decreasing the WCSS with an increasing number of classes is minor (Han et al., 2012).

Applications

Spectroscopy

One common application of *k*-means clustering in soil science is to partition samples based on their spectral information, such as from vis-NIR or MIR sensors. Because soil spectral information can be recorded cost-effectively and rapidly, spectral libraries can cover very large areas, encompassing soils with very different characteristics (Heil et al., 2019). These large soil spectral libraries are often complex and have many non-linear relationships (Araújo et al., 2014). This is a challenge, and classification algorithms such as *k*-means can be applied to identify subsets or patterns of similar samples prior to further analysis (Heil et al., 2019). Partitioning soil samples into smaller subsets based on spectra can significantly improve model predictions (Heil et al., 2019). A study by Araújo et al. (2014) found that the division of a large vis-NIR soil spectral library into clusters substantially reduced errors and improved model predictions of clay and organic matter. This is because clustering divided the global dataset into more mineralogically uniform clusters, regardless of the geographical origin of samples. These results clearly demonstrate the value of clustering soil spectral data into subsets compared to using large global spectral libraries (Araújo et al., 2014).

Precision agriculture and sampling site selection

Another common application of *k*-means clustering is to create spatial zones across a study area for precision agriculture (PA) or soil sampling site selection.

In precision agriculture, *k*-means clustering is often used to identify similar areas within an agricultural field based on a number of characteristics. Common data sources include crop yield monitor data, soil apparent electrical conductivity (ECa), and elevation (Whelan et al., 2003). The clustering process enables subareas of fields to be managed separately by farmers to suit the particular characteristics of the zone (Gavioli et al., 2019). An example of clustering was its application to several seasons and years of digital

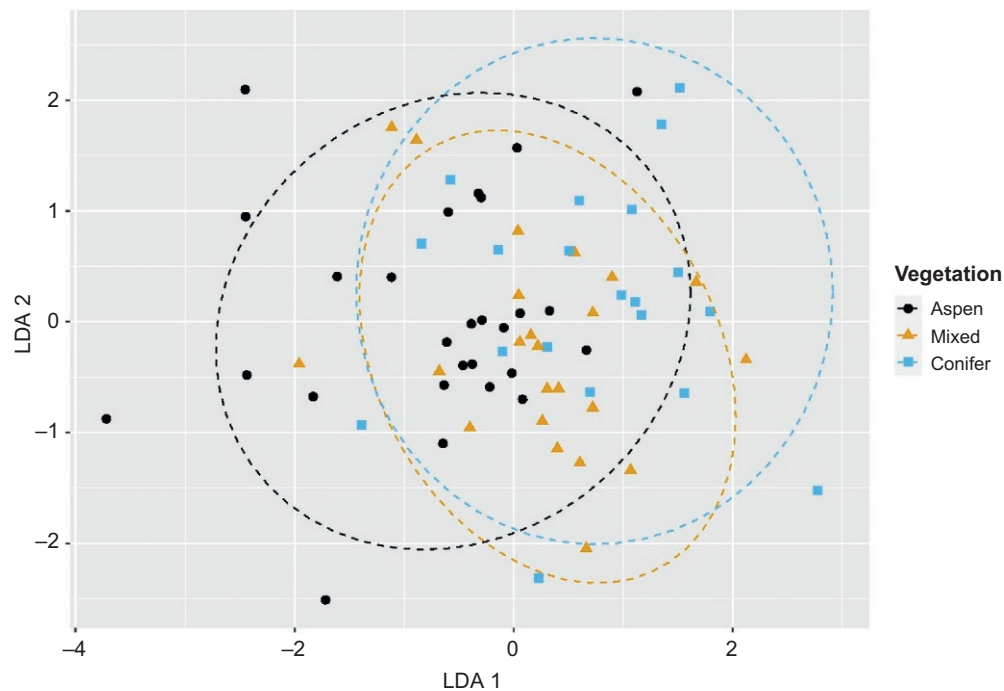


Fig. 4 Scores plot of the linear discriminant analysis (LDA) of the light fraction of soil organic matter under aspen, mixed, and conifer overstory. There is some separation between aspen and conifer along the first axis, with the first discriminant function achieving a percentage of separation of 92%. Light fraction under mixed overstory overlap with both aspen and conifer. The second discriminant function did not achieve a good discrimination, with only 7% separation.

yield maps to create different zones of potential crop yield, and then differing amounts of fertilizer were applied to match the potential yield in each zone.

The *k*-means clustering is also commonly used to decide where to take soil samples from across a study area. The process typically involves clustering a suite of existing data layers to create zones of similarity, and then applying a stratified random sampling approach to select sites. Fig. 4 shows a typical example of spatial clustering across several agricultural fields for The University of Sydney farm property L'lara in Australia from Filippi et al. (2019). In this example, digital soil maps, a digital elevation model, air-borne gamma radiometric data, and remotely sensed satellite imagery were used in the clustering process (refer to Table 1), with the optimal number of clusters determined by the elbow method. The proportion of the total study area that each cluster covered was then determined, and the number of samples to extract within each class was based on this. The overall number of sites to be sampled was pre-determined based on cost and practicality, as is typically done. In this case study, 30 sites were selected and the number per cluster was proportional to the cluster area. The sites were then selected randomly to determine the exact location of samples (Fig. 5).

An alternative

There are many algorithms available to group or cluster data points, and the choice of method depends on the dataset size, structure, and the goal of the analysis. However, *k*-means clustering is consistently reported to perform well and is fast, robust, can deal with large datasets, and is easily implemented (Cebeci and Yildiz, 2015; Román Dobarco et al., 2021b). These are key reasons why the algorithm has been so popular and continues to be to the present (Heil et al., 2019).

However, it is acknowledged that *k*-means clustering does not deal with scattered or overlapping clusters well, and fuzzy clustering approaches are better at accounting for this (Viscarra-Rossel and Webster, 2011). In fuzzy clustering, every data point belongs to every cluster to a certain degree of membership. Because of this, fuzzy clustering is more suited to more complex and high dimensional datasets, such as soil spectra (Viscarra-Rossel and Webster, 2011).

Conclusions

Advances in technology enable us to measure a multitude of attributes related to soil, many of which are difficult to interpret in their raw form. Given the array of possibilities, knowledge of multivariate techniques is an essential part of the soil scientist's toolkit to help their understanding of the state and function of soil. In this chapter we have presented one common approach for three classes of multivariate techniques, namely dimension reduction, supervised and unsupervised classification. Each is illustrated with an example to improve understanding. In addition, for each approach we described an alternative that soil scientists should consider

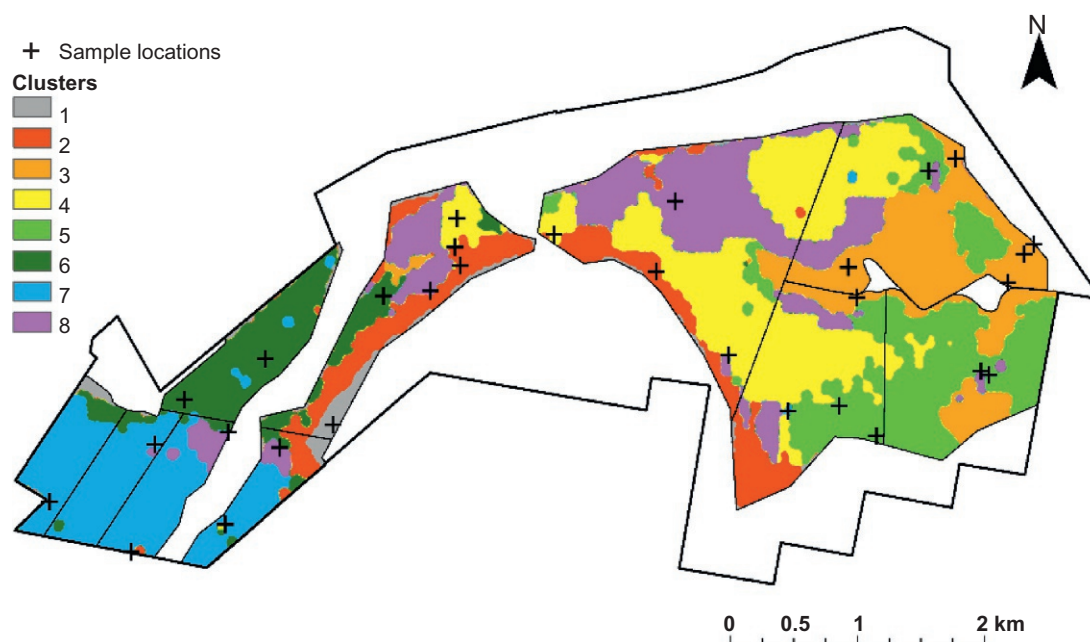


Fig. 5 The distribution of eight clusters derived from a *k*-means cluster analysis, and the locations of the soil sampling sites randomly selected within each cluster across The University of Sydney farm L'ara, NSW, Australia (Filippi et al., 2019).

because it could be more useful in some cases. This chapter is by no means comprehensive. Other fields such as computer science, data science and statistics are also developing new approaches so the reader should consider this chapter to be the start of a journey and keep their mind open to explore and test new approaches as they become available.

References

- Araújo SR, Wetterlind J, Demattê JAM, and Stenberg B (2014) Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. *European Journal of Soil Science* 65: 718–729.
- Box GEP (1949) A general distribution theory for a class of likelihood criteria. *Biometrika* 36: 317–346.
- Cebeci Z and Yildiz F (2015) Comparison of *k*-means and fuzzy *c*-means algorithms on different cluster structures. *Agrarinformatika/Journal of Agricultural Informatics* 6: 13–23.
- Cox GM and Martin WP (1937) Use of a discriminant function for differentiating soils with different *Azotobacter* populations. *Iowa State College Journal of Science* 11: 323–331.
- Dudek M, Kabala C, Labaz B, et al. (2021) Mid-Infrared spectroscopy supports identification of the origin of organic matter in soils. *Landscape* 10: 215.
- Ertlen D, Schwartz D, Brunet D, et al. (2015) Qualitative near infrared spectroscopy, a new tool to recognize past vegetation signature in soil organic matter. *Soil Biology and Biochemistry* 82: 127–134.
- Filippi P, Jones EJ, Ginns BJ, et al. (2019) Mapping the depth-to-soil pH constraint, and the relationship with cotton and grain yield at the within-field scale. *Agronomy* 9: 251.
- Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Annals of Human Genetics* 7: 179–188.
- Gallant JC and Austin JM (2015) Derivation of terrain covariates for digital soil mapping in Australia. *Soil Research* 53: 895–906.
- Gavioli A, de Souza EG, Bazzi CL, Schenatto K, and Betzek NM (2019) Identification of management zones in precision agriculture: An evaluation of alternative cluster analysis methods. *Biosystems Engineering* 181: 86–102.
- Gray J, Bishop TFA, and Wilford J (2016) Lithology and soil relationships for soil modelling and mapping. *Catena* 147: 429–440.
- Han J, Kamber M, and Pei J (2012) 10-cluster analysis: Basic concepts and methods. *Data Mining*. Vol. 3, pp. 443–495. Boston, MA: Morgan Kaufmann.
- Heil J, Häring V, Marschner B, and Stumpe B (2019) Advantages of fuzzy *k*-means over *k*-means clustering in the classification of diffuse reflectance soil spectra: A case study with West African soils. *Geoderma* 337: 11–21.
- Hughes RE and Lindley DV (1955) Application of biometric methods to problems of classification in ecology. *Nature* 175: 806–807.
- Isbell R (2016) *The Australian Soil Classification*. CSIRO publishing.
- Jones EJ, Filippi P, Wittig R, et al. (2021) Mapping soil slaking index and assessing the impact of management in a mixed agricultural landscape. *The Soil* 7: 33–46.
- Lazaar A, Mahyou H, Gholizadeh A, et al. (2019) Potential of VIS-NIR spectroscopy to characterize and discriminate topsoils of different soil types in the Triffa plain (Morocco). *Soil Science Annual* 70: 54–63.
- Lee CS, Sung TM, Kim HS, and Jeon CH (2012) Classification of forensic soil evidences by application of THM-PyGC/MS and multivariate analysis. *Journal of Analytical and Applied Pyrolysis* 96: 33–42.
- Likas A, Vlassis N, and Verbeek JJ (2003) The global *k*-means clustering algorithm. *Pattern Recognition* 36: 451–461.
- MacQueen J (1967) Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* vol. 1, 281–297. No. 14.
- McGarigal K, Cushman SA, and Stafford S (2013) *Multivariate Statistics for Wildlife and Ecology Research*. Berlin, Germany: Springer Science and Business Media.
- Minty B, Franklin R, Milligan P, Richardson M, and Wilford J (2009) The radiometric map of Australia. *Exploration Geophysics* 40: 325–333.
- Oertel AC (1961) Chemical discrimination of terra rossas and rendzinas. *Journal of Soil Science* 12: 111–118.
- Rao CR (1948) Tests of significance in multivariate analysis. *Biometrika* 35: 58–79.
- Román Dobarco M, Jacobson AR, and Van Mieghem H (2021a) Chemical composition of soil organic carbon from mixed aspen-conifer forests characterized with Fourier transform infrared spectroscopy. *European Journal of Soil Science* 72: 1410–1430.

- Román Dobarco M, McBratney AB, Minasny B, and Malone BP (2021b) A modelling framework for pedogenon mapping. *Geoderma* 393: 115012.
- Sun XL, Zhao YG, Zhang GL, et al. (2011) Application of a digital soil mapping method in producing soil orders on mountain areas of Hong Kong based on legacy soil data. *Pedosphere* 21: 339–350.
- Viscarra-Rossel RA and Webster R (2011) Discrimination of Australian soil horizons and classes from their visible-near infrared spectra. *European Journal of Soil Science* 62: 637–647.
- Webster R and Burrough PA (1974) Multiple discriminant analysis in soil survey. *Journal of Soil Science* 25: 120–134.
- Whelan BM, Cupitt J, and McBratney AB (2003) Practical definition and interpretation of potential management zones in Australian dryland cropping. In: Proceedings of the 6th International Conference on Precision Agriculture and Other Precision Resources Management, Minneapolis, MN, USA, 14–17 July, 2002, pp. 395–409. American Society of Agronomy.
- Williams BK and Titus K (1988) Assessment of sampling stability in ecological applications of discriminant analysis. *Ecology* 69: 1275–1285.
- Xu T and Hutchinson M (2011) *ANUCLIM Version 6.1 User Guide*. Canberra: The Australian National University, Fenner School of Environment and Society.

Machine learning and processing of large data

Madlene Nussbaum, Bern University of Applied Sciences BFH, School of Agricultural, Forest and Food Sciences HAFL, Berne, Switzerland

© 2023 Elsevier Ltd. All rights reserved.

Introduction	510
What is machine learning?	510
Supervised learning	511
Should we choose a simple or a complex model?	511
Strategies of machine learning algorithms	512
Regularization	512
Component-wise gradient boosting	512
Decision trees	513
Bagging	513
Support vector machines (SVMs)	514
Artificial neural networks (ANNs)	514
Ensembles	514
Model selection	514
Model assessment and uncertainty	515
Model interpretation	515
Unsupervised learning	516
Clustering	516
Dimension reduction	517
Feature engineering and data fusion	517
Processing of large data	517
Typical applications	518
Spatial and spatio-temporal machine learning	518
Transfer functions at the level of soil samples	518
Integration of sensors in soil related decision making	519
Summary and future requirements	519
Acknowledgments	519
References	519

Abstract

Machine learning refers to a set of tools to establish models of linear or non-linear relations or other previously unknown relationships in complex data. In soil science it is widely used to create soil maps, to derive difficult to measure soil properties from sensor data or to provide information for soil related decision-making in agriculture. Machine learning algorithms based on decision trees are among the most popular techniques. Models with good predictive power are complex, but often outperform simpler models from classical statistics. Interpretation with algorithms that subsequently analyze the models is possible yet challenging.

Key points

- Machine learning encompasses a wide range of approaches that enlarge a soil scientist's statistical toolbox.
- Machine learning refers to methods to establish models of linear, non-linear, interaction type relations or other previously unknown relationships in complex data.
- Often a model with good prediction power is complex and difficult to interpret. Tools applied after the model building provide insight into the modeled relationships.
- Soil science applications mostly use supervised machine learning to predict unknown soil information when auxiliary information is available.
- Algorithms based on decision trees are among the most popular techniques.
- Machine learning is widely used to create soil maps, to derive difficult to measure soil properties from sensors data or to provide information for soil related decision-making in agriculture.

Introduction

Investigations of soil–environment relationships to understand and predict soil better have a long tradition in soil science. The quantitative analysis of these relationships is studied within the field of Pedometrics. Increasingly large datasets on soil itself or on environmental aspects describing soil are available. For instance, sensors that record soil moisture at frequent intervals or the spectral reflectance of a soil sample, aerial images with large numbers of spectral bands, elevation models showing detailed terrain features or harmonized databases assembled from various archives provide valuable information on soil. The analysis of such abundant data with classical statistical approaches, for example with a linear regression, has become increasingly challenging. The data embrace several types of data which are often correlated making it difficult to choose relevant variables for a certain goal.

Imagine that your aim is to generate a map of topsoil organic carbon for land management and climate mitigation. Your budget allows you to take and analyze only a limited number of samples from the study area. Therefore, you plan to use existing information of the study area, for example relating images of two satellites with four spectral bands each to your topsoil organic carbon measurements to support mapping. But how do you know which satellite you should use? Should you calculate vegetation indices from the bands? Which bands or indices are relevant variables to include in such a model? Should you exclude the bands that are correlated? Is the relationship linear or do you need to transform your bands by some function? If yes, which function?

To answer these questions, you decide to start fitting a linear regression model and to check each band by adding and removing them from your model. Then, you verify if your predictions according to a validation measure become better or worse. You retain a variable that improves the validation measure, otherwise you remove it. Because of the large number of combinations, it would take you a considerable amount of time, but eventually you would find a possible model to predict topsoil organic carbon for your study area. In this way, you let yourself be guided by the data to ‘learn’ a model structure by a stepwise algorithm where each step was decided by you.

Such a process can be tedious, time consuming and in the end, you cannot be sure if you stopped too early to search for the best model. Furthermore, if more samples become available you have to start again. To overcome numerous constraints while exploring statistical models soil scientists started to use machine learning. Machine learning refers to a set of algorithms that basically follow the same learning procedure, just in a more formalized and statistically sound way.

Unlike geostatistics, e.g. kriging approaches – which have been fostered by many soil scientists or statisticians working on soil mapping – machine learning tools have developed outside soil science. Several advances have led to successful adoption of machine learning. In addition to the large soil and environmental datasets available, progress in computer science and statistics, open source and easy applicable software and affordable computational power have made widespread use of large data and application of machine learning possible.

What is machine learning?

Machine learning refers to a set of tools to establish models of linear, non-linear, interaction type relations or other previously unknown relationships in complex data. It is an interdisciplinary field of computer science and statistics and is considered a subfield of artificial intelligence (AI). Depending on whether a computer scientist, a statistician or mathematician teaches machine learning the vocabulary and focus differ. Computer sciences define machine learning from an algorithmic position as the science of “how to program computers so they can learn from data” (Géron, 2017, p. 2). The often synonymously used term of statistical learning stresses the fact that machine learning methods are not just ad-hoc algorithms but have a sound statistical background (Hastie et al., 2009).

The boundary between classical statistics and machine learning is blurred. Classical statistics analyzes data to infer knowledge, for example by calculating p -values or searching for easily explained models and relationships. Often prior hypotheses about relationships are formulated and parametric methods are used, meaning that assumptions are made about the form of the relationship. The use of a predetermined form of the functions used in the model (e.g. linear relations between variables y and x) often assumes a specific distribution of the data analyzed (e.g. a normal distribution). An important part of the modeling therefore is to verify the assumptions of the model used to avoid a suboptimal model fit or incorrect conclusions. This can be done by analyzing the residuals that were not explained by the model (Gareth et al., 2017).

On the other hand, a machine learning method derives the relationships in the data by an algorithmic procedure, hence it is data driven. Machine learning focuses mostly on prediction, and it is only recently that their importance to gain knowledge, e.g. on soil processes has been stressed (Wadoux et al., 2021). Machine learning is often non-parametric. This means no prior exact form of the relationship is given and no assumption about the distribution of the data is implied. There are, however, also parametric machine learning methods where algorithms create models on base procedures that have a predefined functional form. Machine learning can handle high-dimensional data even if there are more explanatory variables than observations (many multivariate methods cannot do this). Correlated variables are dealt with so that they do not disturb the model fit and models are built with little to no user interaction.

Machine learning approaches can be classified into three main types. For soil scientists the most important type is supervised learning, whereas unsupervised learning is less widespread at present but has many meaningful applications in soil science. Reinforcement learning, another variant of machine learning, is rarely applied. It is a step-by-step improvement in learning that

is used, for example, in robotics where direct sensor feedback is processed to learn for future decisions. In crop production, it has been applied to sensors controlling irrigation (Sun et al., 2017). Here, I focus first on supervised learning then on unsupervised learning. For each type, popular methods are briefly explained.

Supervised learning

Assume that we observe a quantitative response variable Y and at the same n locations or on the same n soil samples (observations) we observe p potential predictor variables $X = X_1, X_2, X_3, \dots, X_p$. Suppose there is some relation between Y and X in the form of:

$$Y = f(X) + \varepsilon, \quad (1)$$

where f is an unknown function of X and ε is a random error term and represents the part, also called residual, that cannot be explained by $f(X)$. Supervised machine learning refers to a set of approaches to estimate f from the given pairs of Y and X (Gareth et al., 2017). The topsoil carbon measurements in the above example ‘supervise’ the building of the model, also termed model training. Topsoil carbon Y is then called the response variable or short response, as used in this chapter, or also dependent, output or target variable. The values X of the satellite image bands at the sampled locations are auxiliary information synonymously named independent, explanatory or input variables, covariates, features or predictors, the latter is used in this chapter. The whole set of X is often referred to as feature space. Like geographical space it can be imagined as a high-dimensional space that covers the thematic aspects of the observations.

The goal of model building is to find f that fits best to the data. A form for the function f needs to be defined and the optimal parameters of this function are estimated from the data. If Y is a continuous numeric response variable estimation of $\hat{f}$, it is a regression problem, if Y is categorical with two or more levels it is called a classification. Most machine learning approaches can include mixed data with X being continuous and categorical. A prediction task, also called forecasting if the goal is to predict about the future, aims to find $\hat{f}(X)$ that allows good predictions of $\hat{Y}$ to be computed for new locations or samples where just the predictors $\hat{X}$ are known. Estimated $\hat{f}(X)$ can be analyzed to gain insight into the relation of $\hat{Y}$ and X to improve model building or to learn about soil processes (Fig. 1).

Should we choose a simple or a complex model?

Some methods are more restrictive and less flexible regarding the shape that the function f can take while others allow f to be more flexible. For example, a linear regression is a relatively inflexible approach because it can produce only linear terms compared to smooth spline functions which can become highly non-linear. One could reasonably ask: Since soil processes and soil–environment relations are profoundly complex why should we ever choose a restrictive model with very little flexibility?

There are reasons why we should restrict flexibility, but to decide the optimal degree of restriction is challenging. If we are mainly interested in gaining knowledge about the relations between X and Y then a simpler model is more accessible to interpretation. More complicated shapes of f make it difficult to understand how individual predictors are related to the response variable. A model that is too simple might, however, lead to an oversimplification in the interpretation of the complex nature of soils.

If the aim is to predict, one often uses much more complex models to capture all possible relations in the data. Nevertheless, for prediction we have to restrict the model flexibility because the most flexible model will not create the best prediction. It will fit well to the training data, but with little ability to generalize beyond the current dataset and predict the response variable $\hat{Y}$ well for newly

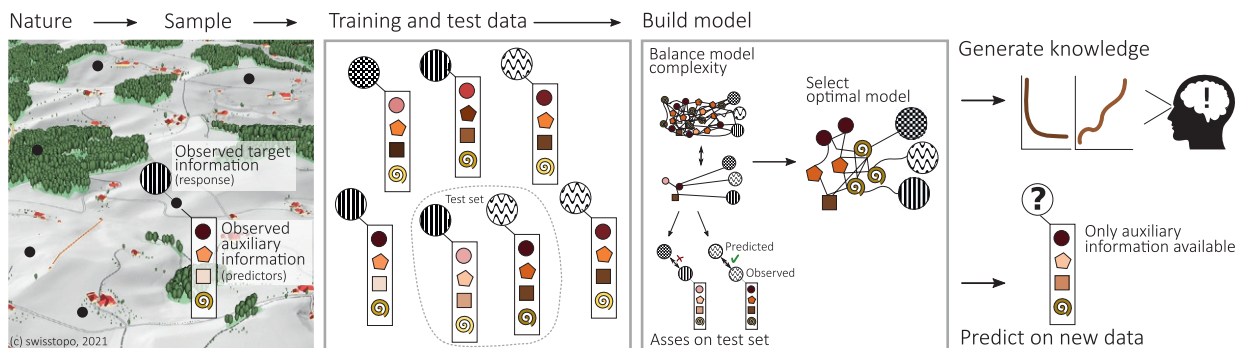


Fig. 1 Schematic view of a supervised learning task for the generation and prediction of soil knowledge. The sketch shows a categorical response variable (e.g. soil structure class) and four numeric predictors (e.g. slope, clay content) that are used to train a non-linear model searched for by an algorithm. An optimal model is determined by calculating predictions for a test set not used to create the model and by comparing them to the observed values. Only if this assessment is satisfactory, may one use the model to gain insight or predict at locations or for soil samples where the target information is missing.

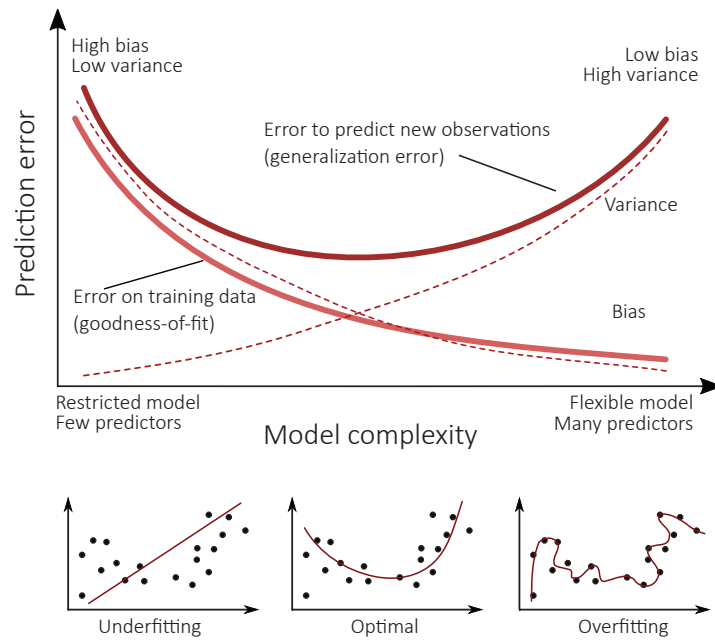


Fig. 2 Prediction error on training data (bright solid line) and new observations (dark solid line) in relation to model complexity illustrating the trade-off between model simplicity and prediction accuracy. In addition, the generalization error is decomposed into model bias and variance (dashed lines). The small bottom panels display a scatterplot of a response and predictor variable with three functions of different flexibility fitted to the data points demonstrating under- and over-fitting.

observed $\tilde{X}$. This situation is called overfitting. If a model is not flexible enough to map the basic structure in the data, then it is underfitting and its performance is weak on unseen data (Fig. 2).

The dotted lines in Fig. 2 show how the prediction error can be decomposed into model bias and model variance (Hastie et al., 2009). Variance refers to the degree by which f would change if we used a different dataset to estimate it. Large variance is observed if only small changes in the dataset, for instance by adding a few new soil samples, result in large changes in f . Bias, i.e. systematic distortion in the estimation of f , is introduced by approximating the usually very complex soil–environment relationships by a much simpler model. More flexible models generally result in less bias, but in an increase of variance. In a prediction setting, the challenge is to find a balance that allows enough model flexibility to capture the main structure to reproduce the general characteristics. At the same time the model needs to be restricted to avoid over-adaptation to the training data. Successful algorithmic procedures for prediction allow flexibility, but at the same time use strategies to reduce variance.

Strategies of machine learning algorithms

To find a balance between bias and variance, machine learning algorithms for supervised learning follow different strategies. The most often used approaches and associated methods are described in this section (Gareth et al., 2017; Géron, 2017; Hastie et al., 2009; Tutz, 2012).

Regularization

A stepwise forward selection for linear regression includes relevant predictors by testing them sequentially. Each predictor is tested and only the predictor that decreases the model error most is added. Adding of predictors is stopped when model improvement falls below a threshold. A linear model becomes more flexible the more predictors are included. In stepwise forward selection, and also in stepwise backward or best subset selection, model flexibility is controlled by using a subset only of all p predictors.

A more elaborate way to control the predictors in a linear model is regularization of model coefficients. Instead of a predictor being fully in or fully out of the model fit, it is used partially. All p predictors enter the model, but the regression coefficients are constrained. Instead of just minimizing the sum of squared errors an additional penalty is imposed on the size of coefficients. If the penalty increases, the coefficient estimates shrink toward zero. Some coefficients might even become zero which removes a predictor from the model. Ridge regression or lasso are the two best known techniques for regularization. They result in linear models where coefficients are estimated differently from ordinary least squares regression. Compared with the latter, ridge regression and lasso can even be applied when there are more predictors p than observations n .

Component-wise gradient boosting

Another way to control the inclusion of predictors and the size of coefficients is the gradient boosting algorithm. It selects predictors in a forward stagewise manner but compared to stepwise forward selection for linear regression it is not greedy, i.e. it develops the

model only slowly. It also uses an adaptive way to reduce model variance step by step. Instead of fitting the models to the response variable, at each iteration the model is fitted to the residuals of the previous iteration in choosing the best predictor to explain the remaining variance. The model is updated sequentially by adding the best performing predictor with a partial effect only, e.g. by shrinking it to 10% of its original size. Because of the small step size many iterations are needed. Each predictor can be added multiple times to ensure slow learning of its effect without overfitting the training data. The components tested in each boosting iteration are called 'base learners'. A base learner is a basic function and can take any form such as linear regression, a spline function or decision tree. The latter are the most popular base learners for boosting, the algorithm is known as 'gradient boosted trees.'

Decision trees

Tree-based methods split the predictor space into a set of rectangles by a procedure called binary recursive partitioning. The data are split into two portions by a predictor. Thresholds are tested to find the best split criterion for each predictor, and the predictor and its optimal threshold are chosen to produce the smallest error statistic calculated on the training data.

Each of the subsets is split further in two portions with the same or another predictor until a stopping criterion is met. Because the splitting rules to form this type of model can be summarized as tree diagrams, they are called decision tree methods. The most popular variant are classification and regression trees (CART). For continuous responses the squared sum of the residuals is minimized at each split, and for categorical responses the Gini index that measures the ratio of wrongly classified observations in the training data is often used. For each of the resulting final observation subsets (tree end nodes or tree leaves) decision trees use a simple model to compute the predictions. In CART the mean or the majority 'vote' of all observations inside each tree leaf is used. Other approaches like Cubist compute linear regressions at each end node.

Decision trees are easy to interpret and to apply and allow for non-parametric model estimation. Unfortunately, they often have poor prediction accuracy. Fully grown trees tend to overfit the data, so cropping (pruning) of the trees was introduced. Nevertheless, decision trees remain procedures with a large model variance, but low bias. The machine learning community introduced several strategies to overcome the drawback like combining many decision trees in component-wise boosting as introduced above.

Bagging

Another strategy to prevent procedures with large variance from overfitting is bootstrap aggregation or in short bagging. Bootstrapping is a widely applied and powerful statistical tool. In its non-parametric form, it repeatedly draws random samples from the training data. This so-called resampling creates numerous new training datasets that have identical statistical properties in terms of data distribution compared to the original training data. No parametric assumptions about the distribution of the data are required. Resampling imitates having multiple sampling campaigns in the same study. The model of interest is fitted to each of these samples and predictions are computed. The numerous models provide approximate error distributions or balance for high variance of models by aggregating the predictions, usually by their mean.

A popular method that applies bagging to decision trees is random forest (Fig. 3). It uses the bootstrap idea twice: on the observations Y and on the p predictors in X . Many trees are fitted to bootstrap samples of the original data. To ensure further that the trees are dissimilar, m of p predictors are randomly sampled at each split and only these are tested as split candidates. The trees to create random forests are fully grown and predictions are averages of all trees.

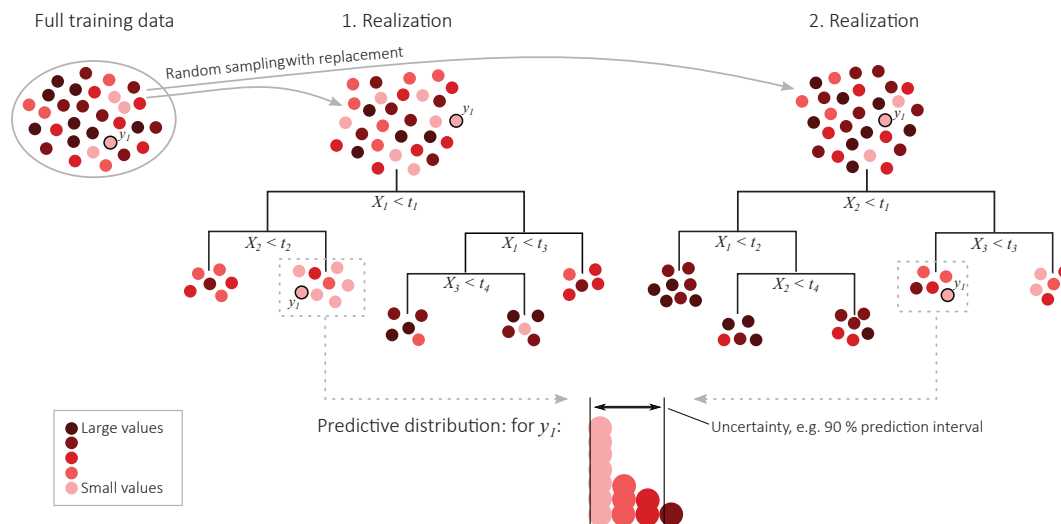


Fig. 3 Visualization of the bootstrap with two repetitions using decision trees as in random forest. $X_1, 2, 3$ are the predictors used to partition the data. $t_1, 2, \dots, n$ represent the individual thresholds optimized for each tree. For observation y_i the estimation of uncertainty with a quantile regression forest is illustrated. For each tree leaf (tree end node) the observations are kept. All the observations of leaves where y_i is found are pooled to obtain the full distribution. This predictive distribution allows to compute prediction intervals for new observations with the same values in $X_1, 2, 3$.

Support vector machines (SVMs)

Another method to approach complex datasets are models based on so called support vectors. They originated as a classification tool, and were then expanded for regression tasks also. The models are built by identifying characteristic observations within a dataset that are used to split the data into the classes of the response variable. These observations are used to define decision boundaries for the splits and are called support vectors. The shape of the decision boundaries to partition the data may be linear, but non-linear functions can also be used (e.g. radial kernel functions that result in smooth rounded boundaries).

Data can rarely be split perfectly along a decision boundary; therefore, a certain number of observations is allowed to be wrongly assigned. The decision boundaries are not crisp but are surrounded by a so-called soft margin that allows some observations to disregard the boundary. In addition to the shape of boundaries, the width of these soft margins is a further way to control model flexibility. Observations close to the decision boundaries only influence estimation of the model, hence support vector machines are not sensitive to outliers in the data. This property and the good control of model flexibility often results in good predictive performance.

Artificial neural networks (ANNs)

The development of artificial neural networks was inspired by the biological neurons in our brains. This analogy is not universally supported nowadays, however, these networks can be explained by artificial neurons that operate like switches or nodes as for neurons in the brain. Each neuron takes one or more binary inputs and has a binary output. A neuron becomes active if the rule to switch it on was fulfilled, for example at least one of the two previous neurons needs to be active. An artificial neural network consists of multiple so-called hidden layers of such neurons (computational steps) that transform model input into predictions. Neurons may also store numeric values in addition to an on-off switch, and the rules to transform the values from one layer to the next may be a complex function (activation function). Moreover, weights are applied to the values handed over from the previous to the next neuron. Simply, the goal of a neural network is to extract linear combinations of the predictors to obtain new predictors and then model the response variable as non-linear functions of these predictors.

The setup of hidden layers and structure of how rules update the neurons (function to activate neurons) is called the architecture. Neural network models take 'normal' data tables as input. But complete images may also be passed to specific implementations. Convolutional neural networks derive multiple layers of new features from the input images by applying different filters (augmenting step) before these transformations are used in a network as described above. This type of neural network is often referred to as deep learning.

Artificial neural networks perform especially well with complex prediction tasks and when a huge quantity of data is available. Model flexibility is balanced by layers that dropout and deactivate a predefined number of neurons in a layer. Furthermore, overfitting is controlled by a slow learning rate as in boosting (see above) applied to each learning iteration and by using only a randomly chosen subset to train the network and optimize weights.

Ensembles

Component-wise boosting and bagging described above combine many models of the same type to increase prediction accuracy, hence they already form model ensembles. To merge predictions from several models is not limited to using the same model type, for example in random forest. Model ensembles can be formed by producing new predictions from the predictions $\hat{y}$ of two or more models that follow a different approach. Such ensembles often outperform the predictions of a single machine learning approach because they capture different types of relations in the data.

There are several methods that merge multiple model predictions into one. A simple model is often used such as a simple or weighted mean of all predictions (model averaging). Overall prediction accuracy or local uncertainty can be considered to define weights. Predictions of well performing models receive larger weights, whereas those of poorly performing models have smaller weights. Another strategy is stacked generalization also called stacking, whereby a new model is trained using the predictions as explanatory variables. Constrained linear regressions are mainly used and optimal coefficients (= weights for each model) are chosen based on a new training dataset.

Model selection

So far, we have learned many approaches that might all seem feasible to apply to our data, but which should we choose? Unfortunately, there is no 'free lunch' in statistics. No method outperforms all others over all possible datasets. We do not know a priori which approach, which algorithm settings (tuning parameters) or which selection of predictors perform best on our data to achieve our goal (Hastie et al., 2009). Moreover, two models of completely different structure might perform equally well (model multiplicity, Breiman, 2001).

Soil scientists chose machine learning methods based on knowledge from previous studies that performed well on a similar task, or they test multiple approaches and select the best performing one. Predictors are not always selected. Many machine learning algorithms use all the predictors that are offered, some use just covariates that improve the model at a certain step, for example boosting or regularization where exclusion is part of the learning. Machine learning can handle many irrelevant predictors. With the strategies explained above the exclusion of covariates from the model is not necessary from a statistical learning point of view.

(Hastie et al., 2009). For ease of interpretation soil scientists often prefer a model with fewer covariates and processing becomes easier if only a subset is used to compute the model predictions.

One possibility is to pre-screen the set of potential predictors prior to applying machine learning. The predictors are compared to the response variable one by one and only if they pass a certain criterion, i.e. the significance level of a statistical test, are they used in the subsequent machine learning algorithm (univariate filtering). For correlated predictors only one is retained based on a similarity criterion like a correlation coefficient or the variance inflation factor (VIF). An approach using the machine learning model itself is called recursive backward elimination. To achieve a smaller set of explanatory variables, the importance of each variable in the machine learning model fit is evaluated. The least important variable is then excluded, and the model is fitted again. This procedure is repeated until an optimal number of variables based on the best performing model is obtained. A similar approach is the Boruta algorithm. It adds so-called shadow predictors to the model that were permuted randomly. Only predictors more important than the randomly shuffled predictors are retained in the model.

In addition to the selection of relevant predictors, machine learning algorithms have tuning parameters that can be optimized with the current dataset. They are settings that control the number of iterations, such as the number of covariates tested at each step in a random forest, the degree of shrinkage of coefficients in lasso or the smoothness of a kernel function in support vector machines. A grid search is usually applied to find optimal values for the tuning parameters. A grid or matrix with tuning parameters at different intervals is defined and the machine learning algorithm is applied for each parameter combination. The best performing combination for a dataset is chosen to compute the final model.

Model assessment and uncertainty

It is now time to discuss how 'best performing' or 'optimal' parameters are evaluated, how the 'best' model is chosen and what we mean by 'good' predictions. Model assessments have two goals:

1. to estimate the performance of various models and choose the best one by evaluating its 'goodness-of-fit' to the data or its capacity to generalize beyond the training dataset,
2. with a final model, estimate its prediction or generalization error if applied to new data not seen by the model before.

If our goal is to predict $\hat{Y}$ on new observations $\tilde{X}$ the best model $\hat{f}(X)$ has the smallest generalization error. This requires application of the model to new observations. To select the best model the data are often split, and the different models are trained on a subset only. Predictions are calculated on the remaining data and the prediction error is evaluated. If this process is done repeatedly, it is called cross-validation. The data are split in, for example, 10 equally sized subsets and each time one of the 10 sets is removed from model training.

The final model is evaluated ideally by a new and independent dataset. These data are often called test data, and the process independent or external validation. Additional soil sampling is often not feasible; therefore, data splitting is used by assigning roughly 20% of the soil samples to a test data set that is not used in the model selection process.

By computing predictions $\hat{Y}$ for the test data set we obtain a set of pairs of observations and predictions. From these we can calculate differences and descriptive statistics. In regression problems prediction error is often calculated as the root of the squared differences between observations and predictions (root mean squared error, RMSE). If we choose to predict classes, the prediction error can be calculated as the percentage of wrongly classified individuals (error rate). Both represent an overall error. There are many possible criteria (Wilks, 2011) and depending on the goal we might want to optimize something more specific. For example, in a task to map sites with soil pollution we would like to identify all the polluted locations, therefore, we might prefer to avoid false negative predictions and optimize on the 'hit rate,' i.e. the percent of correctly identified polluted sites.

Assessment of model accuracy results in overall estimates of mean prediction error for a map or pedotransfer function, for example. In addition, it is important to quantify how uncertain our predictions of $\hat{Y}$ are at new observation sites compared with sampled sites. Uncertainty originates from the estimation of $\hat{f}(X)$, because it is only an approximation of the true underlying process that produced the data. Moreover, $\hat{f}(X)$ does not account for all the variance in Y , described by in ε in Eq. (1).

Parametric models such as ordinary least squares regression assume distributions for f and ε and enable direct quantification of the dispersion of the two quantities. Most machine learning methods lack assumptions about the statistical distribution of model parameter and errors. One solution is to simulate the combined distribution of $\hat{f}$ and $\hat{\varepsilon}$ by a resampling tool, i.e. bootstrapping. Fig. 3 illustrates this process for decision trees which results in one combined distribution for uncertainty in estimating $\hat{f}$ and $\hat{\varepsilon}$. Another versatile approach to quantify prediction uncertainties empirically is based on partitioning of the feature space. Clustering (see below) is used to divide the predictor space, and accompanying local uncertainties are assigned to the predictions belonging to the same cluster (Malone et al., 2017).

Model interpretation

Unfortunately, accuracy of prediction and model simplicity with ease of interpretation are incompatible. Most machine learning models with good predictive power are highly complex; they establish response–predictor relationships non-linearly in various layers or with a multitude of single models. In a simple linear regression, the estimated coefficient β can be interpreted directly as the increase in the response variable $\hat{y}$ by the amount β if we increase the predictor by one unit. The many model parameter estimates

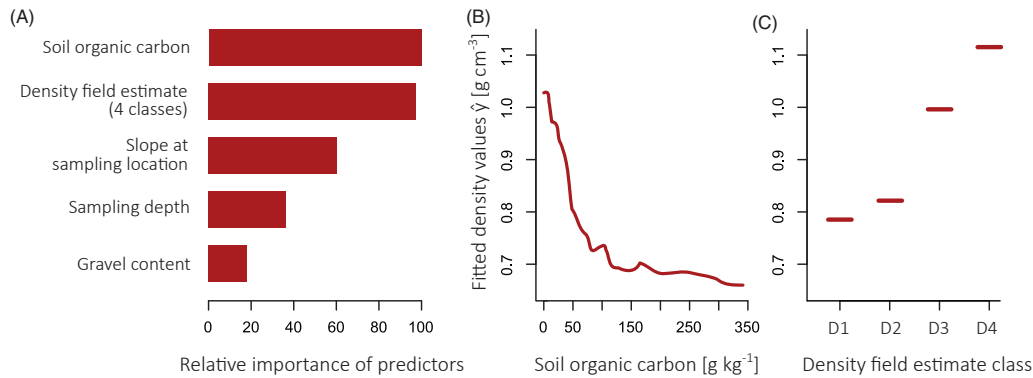


Fig. 4 Relative importance of five variables to predict soil bulk density sampled in forest soils of Switzerland by a random forest model (A). The partial dependence plots for the two most important predictors reveal non-linear effects (continuous predictor [B] and a categorical predictor [C]; based on data from Nussbaum et al., 2016).

resulting from machine learning have no such simple interpretation nor the optimized tuning parameters. On first sight, machine learning appears to be a black box. Post-processing tools applied after model training provide insight to move toward the understanding of modeled relations. Independent of the machine learning algorithm, importance of the variable and partial dependence plots can be applied.

Model-agnostic relative variable importance can be obtained by randomly permuting the predictor of interest. The model structure and all other predictors are kept fixed. Predictions are then re-computed with one predictor shuffled and the decrease in prediction accuracy is evaluated. The larger the decrease, the more important is this predictor relative to the others (Fig. 4).

After the most important predictors have been identified a graphical display of the partial or average dependence helps us to understand their approximate effects. To compute the partial dependence $\hat{f}(X_s)$ of a predictor s on the fitted values of a response $\hat{y}$ values from all other predictors are averaged. Such displays, however, have a more complex interpretation if there are strong interactions or correlations between the predictors.

Unsupervised learning

Unsupervised learning is a technique where the structure of X with p variables at n observed locations or samples is modeled without having a response variable Y . Unsupervised learning is often used for exploratory purposes to learn about the data, but it also supports sampling designs (Brus, 2019) or uncertainty mapping. Two popular sets of approaches are clustering and tools to reduce the dimensions of X .

Let us look at an example: Imagine you received a small project extension to measure aggregate stability on 10 of your 50 samples on which you had already measured carbon. You create 10 piles of the sample bags with similar soils based on their visual properties like soil color, soil structure or gravel content and size. Then you take one sample per pile that seems typical for this group. The search for a subsample that represents our soil samples well is an unsupervised problem because we do not yet know the aggregate stability. With many samples or more properties to evaluate this manual process becomes confusing. However, the choice of samples could be done by statistical clustering resulting in a reproducible and less subjective choice of samples.

Clustering

Clustering is a broad set of techniques to find subgroups in the data. Observations are assigned to groups, i.e. clusters, so that a cluster encompass similar observations while being quite dissimilar to the observations in other groups. Similarity is quantified by relative distances to the content of a set of chosen variables.

Clustering approaches differ mainly in how the distances between the observations are calculated and how the clusters are formed. Fuzzy clustering methods allow observations to become a member of several of the subgroups formed. Hierarchical algorithms either start with the full data set and iteratively split it into smaller groups (divisive) or start by combining similar observations and continue merging them to larger groups (agglomerative). The hierarchical structure can be visualized by a tree-like diagram or dendrogram which helps to find an optimal number of clusters.

The k -means clustering approaches need the number of clusters to be pre-specified. The center of each cluster is then searched for by moving them toward the optimal location during several iterations. To choose the samples in the above example, a k -means clustering with $k=10$ could be used. The approach is known as feature space coverage sampling design because it spreads the samples over the multivariate distribution of the p variables used.

Dimension reduction

Clustering seeks to find homogeneous subgroups among the observations based on selected variables p , whereas dimension reduction approaches aim to find a low-dimensional representation of p that explains a large fraction of the variance in the data. Principal components analysis (PCA) is often used for this purpose. With p variables there is a p -dimensional feature space, but not all dimensions are equally interesting. There is often redundancy in the p variables because many might be correlated. The axes of the p -dimensional feature space are transformed by linear combinations of p so that the common variance of the variables is best represented. The new axes (directions), i.e. principal components are chosen to be orthogonal to each other and to embrace as much variance in the data as possible. The first few principal components explain the most variance in descending order and provide insight about the most important variables.

The principal components (PCs) can be interpreted by the contribution (eigenvector values) made by each variable to the component and the correlation structure of p can be visualized by plotting the eigenvectors in the plane of pairs of the leading PCs. Moreover, principal components can be used to build supervised models. For example, partial least squares regression (PLSR) is an approach that produces principal components in a supervised way to use as linear predictors.

Feature engineering and data fusion

So far, we have learned about methods that can handle large numbers of predictors X to build a model $f(X)$ to predict or explain a response variable Y or learn about the (complex) structure of X . Imagine a small study area where topsoil was sampled, and clay content measured. In addition, a digital elevation model is available for the area. Now we read in the elevation at the sampled location and relate elevation with clay content using one of the numerous tools described in the supervised learning section. With just elevation to explain topsoil clay would give us a very narrow picture of the soil forming processes. Digital elevation models (DEM), however, are straightforward to process and terrain derivatives such as slope, aspect, curvature, or the position in the landscape can be generated from these. Since we are expanding our predictors or features, this is sometimes called feature engineering. This preparatory task allows us to attribute contextual information to the topsoil clay samples, i.e. if a sample was taken at the valley bottom, in a small local depression or on a hilltop. However, none of the approaches mentioned above includes such information. In convolutional neural networks a local extract from a DEM around the sampling location can be included in the model and then processed in the augmenting step.

Preprocessing and the expansion of predictors is a standard procedure, however, for spatial analysis the scale(s) of the soil forming processes are generally not known beforehand. Machine learning approaches can deal with many correlated predictors, and derivatives of elevation models are often computed at multiple scales (e.g. 30-, 100- and 250-m pixel resolution). Model selection helps to identify the relevant scale for a certain response variable (Behrens et al., 2018; Nussbaum et al., 2018). Knowledge about soil forming processes, for example by consulting local surveyors, might form a further basis to derive additional informed predictors.

In addition to elevation data, remote sensing images from satellites or drones or signals from ground sensors are preprocessed in similar and sometimes more complex ways. In remote and proximal sensing literature this is often referred to as image or data fusion. Data fusion stresses the fact that several data sources are combined to improve efficiency to reach the goal (Grunwald et al., 2015). To combine satellite images, pixel-based approaches are widespread where single bands are combined into vegetation or soil reflectance indices, or characteristics over an image time-series are derived, e.g. maximum of a spectral index over the vegetation period (Demattê et al., 2018). Feature-based fusion approaches process each image separately, for example by unsupervised learning, and then combine the results. In addition, image fusion encompasses processing steps to combine different spectral and spatial resolutions. For example, a colored satellite image with low resolution might be improved i.e. sharpened by a greyscale image at a finer resolution (Chang and Bai, 2018).

Data fusion is not limited to imagery, i.e. spatial data covering a surface, but is often applied to ground sensor data measured at single points. No single sensor can approximate soils optimally. Sensor fusion integrates several signals, for example from spectral measurements done at different bandwidths that are optimal to approximate soil organic carbon, others for clay minerals. Proximal measurements may also be fused with wet chemical laboratory measurements into combined datasets for subsequent statistical analysis.

Processing of large data

In connection with machine learning the 'buzzword' Big Data is often used. Big Data in its narrower definition encompasses massive, unstructured datasets that cannot be stored and analyzed in traditional ways and that are often collected and processed in real time (Sagioglu and Sinanc, 2013). One dataset alone is often larger than the disk space of a personal laptop.

Large file sizes originate either from having many variables, many observations or both. In soil science very large datasets originate from genomic analysis, computer tomography on soil samples, various sensor data collected in precision agriculture or from global satellite imagery time series, abundant drone imagery and digital elevation models at high resolution. Although file sizes are large, and processing becomes challenging the data are structured and their size is not the magnitude of Big Data.

Nevertheless, the processing of large datasets is demanding and has only become widespread with the massive increase in computational resources (Rossiter, 2018).

It is possible to train machine learning models on small datasets using only the computing power of a personal laptop. One random forest fit with 1000 observations and 200 predictors, for example, takes a few seconds to complete allowing for multiple runs to optimize the tuning parameters. Models fitted to datasets with many observations, feature engineering, especially on spatial data and computing predictions for large study areas becomes more time consuming and require high performance computing (HPC) facilities. Companies offer such computing infrastructure of various magnitudes and set-up plans to rent for a restricted amount of time, known as cloud computing.

Computations on high performance computers are run in parallel. This means that a job can be broken down into smaller pieces and then the sub-jobs are passed to more than one computer node to evaluate them at the same time (Malone et al., 2017). Parallel computing is now implemented directly in many software packages. If not yet the case, the data soil scientist will have to employ a small extra implementation effort because he or she will have to evaluate which computations are independent and send them to separate nodes. Independent tasks are, for example, model runs with different tuning parameters or when a digital elevation model is cut into subsections and each section is processed separately. Assembling results from parallel runs or balancing of border effects between the subsections then become additional tasks in the workflow. Apart from the increase in computation time, the demand on computer memory becomes challenging. Either software needs to be used that includes memory save implementations by reading only the data currently needed from the disk or the data needing to be split and processed one piece at a time.

Data preparation and the application of machine learning algorithms is often done with programming languages such as R or Python (Géron, 2017; Malone et al., 2017). Both have many add-on packages that provide easy-to-use high-level functions of the newest implementations of tools to process large datasets and apply machine learning. Basic skills in writing code are required to call on computational tasks and to access high performance computing in the sense of giving orders one by one. No advanced skills in object-oriented programming or software engineering are needed nowadays.

Typical applications

In soil science there are numerous applications for the approaches described in this chapter. They can roughly be categorized into applications in space or space-time, at the level of soil samples or sensor data to support soil related decisions.

Spatial and spatio-temporal machine learning

Digital soil mapping exploits the relationship between observed soil information at the level of a soil pedon (point observation, e.g. soil profile or sample) and soil forming factors such as relief, climate, or geology. The latter are represented by spatial data available for the whole target area and enable predictions at unsampled locations and hence provide spatially continuous spatial information (Malone et al., 2017).

Response variables are basic numerical soil properties like subsoil organic carbon content, topsoil moisture or rates of soil erosion, categories such as soil types or ordered categories like drainage classes. Many predictors (20–300) are often prepared and included in models. Predictors are derived from elevation models, remote sensing imagery, geological maps, or existing legacy soil maps. If soil information from multiple sampling times is available, the data can be related to corresponding satellite images to allow a temporal component to be included in the model. Tree-based machine learning approaches like Cubist or random forest are among the most successful models used (Wadoux et al., 2020). The resulting maps display the most likely class or value of the modeled response variable. Moreover, uncertainty maps are provided showing the standard deviation or a 90%-interval of each prediction, for example.

There are many variants of digital soil mapping (DSM) adapting workflow to the current data availability in an area. Sometimes responses are derived from legacy soil maps at coarse scales. Such an approach, termed map disaggregation allows us to increase map resolution with more accurate predictors (Malone et al., 2017). Moreover, the validation strategy to estimate the generalization error of the maps depends on the individual situation regarding previous work to build on, budget and knowledge of staff in a mapping project. A fully independent dataset cannot always be sampled and validation by (repeated) data splitting can be done.

Transfer functions at the level of soil samples

Many soil laboratory measurements involve time-consuming sampling and costly analytical procedures, but they provide valuable information. It is a long tradition in soil science to approximate measurements by rules based on expert knowledge. If such rules are formalized to allow direct calculations, they are called pedotransfer functions. For example, soil bulk density is correlated with soil organic carbon which can be used as a proxy to predict bulk density for samples where organic carbon only was measured (Fig. 4). Furthermore, the use of correlated but less expensive field or laboratory measurements with spectral sensors, for example measuring visible to near or mid-infrared reflectance, enables many soil properties to be approximated accurately.

For (spectral) pedotransfer functions, models are trained at the level of soil samples and predictions are then calculated for samples where only the proxy measurement was done. Ordinary linear regression is a popular approach for pedotransfer functions with one or few correlated ancillary measurements as predictors. Spectral instruments often produce many correlated reflectance bands that are available as predictors. Dimension-reduction approaches like partial least squares regression or machine learning that

handles many correlated predictors such as random forest are often used to predict soil properties from spectral data. From large collections of potential spectral training data (spectral library) usually a sub-sample only is used for model training. The library sub-sample is selected to be similar to the samples to be predicted in terms of feature space.

Integration of sensors in soil related decision making

A major field of application for machine learning approaches is agriculture. Efficient methods for in-field soil property determinations are used to infer appropriate inputs to optimize crop performance and minimize environmental effects (Sharma et al., 2020). Precision agriculture relies on spatio-temporal mapping and transfer functions as mentioned above. Various types of sensors are used that scan vegetation or soil surfaces of complete fields and the data can be combined by sensor fusion. For example, sensors mounted on fertilizer or pest control machinery determine local nutrient deficiencies or diseases by classifying plant images by neural network algorithms. Machine learning based computer vision is further used for weed detection and elimination by farm robots. Arable fields are split into zones with similar soil properties by unsupervised machine learning by, for example, electro-magnetic conductivity or a bare soil drone image. Farmers use such management zones to adapt sowing width and fertilizer quantities or to develop water saving irrigation plans, together with their agricultural knowledge of their land.

Summary and future requirements

Machine learning is not a new technology. The first description of an architecture for an artificial neural network dates to 1943, classification and regression trees have existed since 1984 and aggregating models with bootstrap resampling technique was published in 1993 (Géron, 2020; Hastie et al., 2009). In soil science, only within roughly the last 10 years have complex machine learning procedures become widespread and approaches established. Emergence of easy-to-use open-source software, advances in computational power, increasingly freely available environmental data, newly assembled large soil data bases have all contributed to the increase in machine learning applications. Machine learning is not replacing classical statistics, but rather expanding the statistical toolbox of soil scientists. The task at hand and the available soil and predictor data decide the optimal choice of approach, which might be machine learning, classical statistics, or no statistics at all.

Currently, the following challenges regarding machine learning and soil science are being researched (Wadoux et al., 2021):

- How can soil knowledge be derived from the complexity of machine learning models? For each dataset there is likely to be more than one model (approach, set of predictors, selected parameter configuration) that predicts a response variable similarly well. The so-called model multiplicity needs to be considered for model interpretation, i.e. a not included predictor might be relevant if another predictor were to be dropped.
- How can available soil knowledge be introduced explicitly into the modeling process? Models learn from the data but cannot know about restrictions or combinations of values that clearly follow from soil processes.
- Models are often built independently for each response variable in a study area, for example if topsoil and subsoil carbon are to be predicted. However, responses are often related, and better predictions could possibly be achieved by including two or more responses in multivariate models.

Acknowledgments

The comments of Stéphane Burgos and Tobias Sprafke on earlier versions of the chapter and the support of the editors are greatly acknowledged.

References

- Behrens T, Schmidt K, MacMillan RA, and Viscarra Rossel R (2018) Multi-scale digital soil mapping with deep learning. *Scientific Reports* 8: 15244.
- Breiman L (2001) Statistical modeling: The two cultures. *Statistical Science* 16: 199–231.
- Brus DJ (2019) Sampling for digital soil mapping: A tutorial supported by R scripts. *Geoderma* 338: 464–480.
- Chang N-B and Bai K (2018) *Multisensor Data Fusion and Machine Learning for Environmental Remote Sensing*. Boca Raton, FL: CRC Press.
- Demattê JAM, Fongaro CT, Rizzo R, and Safanelli JL (2018) Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. *Remote Sensing of Environment* 212: 161–175.
- Gareth J, Witten D, Hastie T, and Tibshirani R (2017) An introduction to statistical learning: With applications in R. In: *Springer Texts in Statistics*, 8th edn. New York, Heidelberg, Dordrecht, London: Springer.
- Géron A (2017) *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Safari Tech Books Online. Beijing, Boston, Farnham: O'Reilly.
- Grunwald S, Vasques G, and Rivero R (2015) Fusion of soil and remote sensing data to model soil properties. *Advances in Agronomy* 131: 1–109.
- Hastie T, Tibshirani R, and Friedman J (2009) The elements of statistical learning. In: *Springer Texts in Statistics*, 2nd edn. New York, Heidelberg, Dordrecht, London: Springer.
- Malone BP, Minasny B, and McBratney AB (2017) *Using R for Digital Soil Mapping*. Basel: Springer International Publishing Switzerland.
- Nussbaum M, Papritz A, Zimmerman S, and Walthert L (2016) Pedotransfer function to predict density of forest soils in Switzerland. *Journal of Plant Nutrition and Soil Science* 179(3): 321–326.

- Nussbaum M, Spiess K, Baltensweiler A, et al. (2018) Evaluation of digital soil mapping approaches with large sets of environmental covariates. *The Soil* 4: 1–22.
- Rossiter DG (2018) Past, present & future of information technology in pedometrics. *Geoderma* 324: 131–137.
- Sagiroglu S and Sinanc D (2013) Big data: A review. In: *2013 IEEE International Conference on Collaboration Technologies and Systems (CTS)*, 42–47. San Diego, CA, USA.
- Sharma A, Jain A, Gupta P, and Chowdary V (2020) Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access* 9: 4843–4873.
- Sun L, Yang Y, Hu J, et al. (2017) Reinforcement learning control for water-efficient agricultural irrigation. In: *2017 IEEE International symposium on parallel and distributed processing with applications and 2017 IEEE International conference on ubiquitous computing and communications (ISPA/IUCC)*, pp. 1334–1341.
- Tutz G (2012) *Regression for Categorical Data. Cambridge Series in Statistical and Probabilistic Mathematics*. New York: Cambridge University Press.
- Wadoux AM-C, Minasny B, and McBratney AB (2020) Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews* 210: 103359.
- Wadoux AM-C, Heuvelink GBM, Lark RM, et al. (2021) Ten challenges for the future of pedometrics. *Geoderma* 401: 115155.
- Wilks Daniel S (2011) *Statistical methods in the atmospheric sciences*, 3rd ed. Academic Press.

Relevant websites

<https://mlr3book.mlr-org.com/>—*mlr3 book*.

<https://cran.r-project.org/web/views/MachineLearning.html>—CRAN Task View: Machine Learning & Statistical Learning.

<https://www.kaggle.com>—Kaggle: Your Machine Learning and Data Science Community.

<https://keras.io>—Keras: Deep learning for humans.

<https://topepo.github.io/caret/>—The caret Package.

<https://geocompr.robinlovelace.net/>—*Geocomputation with R*.

<https://christophm.github.io/interpretable-ml-book/>—*Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.

<https://scikit-learn.org>—Scikit-learn 1.0.1: Machine Learning in Python.

<https://www.tensorflow.org>—TensorFlow: An end-to-end open source machine learning platform.

The issue of scale and change of support in the spatial analysis of environmental data

A Castrignanò^a, A Belmonte^b, and N Romano^c, ^aDepartment of Engineering and Geology (InGeo), Università degli Studi Gabriele D'Annunzio, Chieti-Pescara, Chieti, Italy; ^bInstitute for Electromagnetic Sensing of the Environment, National Research Council (CNR-IREA), Bari, Italy; ^cDepartment of Agricultural Sciences, AFB Division, University of Naples Federico II, Portici (Napoli), Italy

© 2023 Elsevier Ltd. All rights reserved.

Introduction	521
Aims and research significance	522
Scale, scaling, and the change of support	522
Spatial models of scale	523
Geostatistics as a coherent framework for scale modeling and scale change	523
Multi-scaling	524
Change of support	525
Geostatistical fusion of heterogeneous spatial data	527
Application of geostatistics to multi-scaling and change of support in spatial modeling	527
Case study	527
Acquisition of reflected electromagnetic radiation with UAV	528
Geostatistical analysis	528
Results	529
Concluding remarks	531
References	533

Abstract

The concept of spatial scale is of fundamental importance in the analysis of spatial data. Issues such as scaling, multi-scale, data fusion, and change of support when analyzing spatial data are currently not given due attention. There is no unique or only one approach to deal with scaling or change of scale. Geostatistics offers several tools to deal with these problems, such as the regularization technique that evaluates the effect of scaling by operating on spatial functions, such as the covariance and variogram, rather than on data. In this chapter, a case study is presented to illustrate how geostatistics can deal with the issue of spatial scale.

Key points

- Scaling is related to the transfer of information across spatial scales
- In multi-scaling, the scales at which the analysis should be conducted have to be identified.
- Scaling may involve a change of support when the variable is estimated on a different volume than that of measurement.
- Geostatistics allows the multi-scaling and change-of-support problems to be addressed with variogram regularization and block (co)kriging.
- Geostatistical data fusion is an effective way to combine heterogeneous spatial data that takes into account change of support.

Introduction

Nowadays, we are inundated with an ever-increasing amount of data distributed in both space and time, i.e. the variables and parameters that describe the dynamics of a physical process in an environmental system. Several disciplines tackle environmental problems, therefore, this information has a wider range of applications, even with neighboring sciences such as economics and epidemiology. Advances in spatial information technology (SIT), such as the global positioning system (GPS), remote and proximal sensing, and geographic information systems (GIS), have contributed to the proliferation of spatial data and related Information Technology (IT) products.

There are several undoubted potential benefits of SIT, but it is important to shed more light on some theoretical and practical problems that can affect the various applications of spatial IT. Most of them include the issues of scale, geographical representation, spatial data acquisition and integration, and uncertainty assessment. In particular, the concept of scale is of fundamental importance in environmental investigations as it involves various aspects of spatial analysis. Although different definitions of scale exist, an effective framework considers scale, as a dimension that schematically comprises the triplet of 'support', 'spacing', and

'extent.' This view was mostly conceived in hydrology but is now accepted in almost all environmental disciplines (Blöschl and Sivapalan, 1995; Loague and Corvin, 2007). One of the advantages of the 'scale triplet' concept is that it can be profitably used for either monitoring activities or computer simulations, and can be applied in either the spatial or the temporal context (McBratney, 1998; Romano, 2014). An additional advantage of applying the triplet scheme to the spatio-temporal analysis of environmental variables and parameters is that a mismatch among the components of 'scale' can be highlighted more easily. A mismatch in scale in environmental studies often occurs between the measured and modeled attributes, and between the spatial scale of observations and that of simulated scenarios of a process that are provided to an end-user (i.e. what we wish to obtain) (see, for example, Teuling et al., 2006; Gentine et al., 2012; Nimmo et al., 2021).

Research indicates that the scale of a spatial analysis should match, as closely as possible, that of the process dynamics under study (Legendre et al., 2002; Wu and Li, 2009). However, a given process can manifest itself and evolve across multiple spatial scales, with complex scale dependencies of the relevant spatial patterns. Non-linear interactions between scales may also occur (Stahl and Hisdal, 2004; Nimmo et al., 2021).

The above concerns contribute to the need to introduce scale, in particular change in scale, in spatial data modeling, and the search for scale dependence and scale invariance (Sposito, 1998).

Aims and research significance

The purpose of this chapter is mainly twofold: (i) to address the concept of scale when analyzing geospatial data to extract spatial information and relevant patterns, and (ii) to discuss the problems of change of support and multiscale spatial modeling with complex and integrated environmental datasets obtained from different types of sensing systems. This chapter draws the reader's attention to two critical questions that could be overlooked when designing a sampling scheme or mapping spatial variables, especially when covariates are used.

The first question refers to up- or down-scaling (also referred to as aggregating or disaggregating) the available observations, i.e. the use of an adequate change-of-support model (Malone et al., 2013). The second question concerns the fact that spatial (but also temporal) processes evolve over different scales and show various levels of complexity (Kavvas, 1999). Variography is a robust tool to investigate the changes in spatial variance over different scales.

The few theoretical issues underpinning this chapter are primarily based on our experiences with the use of geostatistical tools. An illustrative case study serves as a practical example of how those concepts can be approached.

Scale, scaling, and the change of support

Theoretically, scale (a spatial scale, in particular) refers to a characteristic length of the problem at hand and is related to the concept of spatial dependence, which is an intrinsic property of many environmental processes and their spatial representations. Spatial dependence refers to the property of neighboring observations being similar and influencing each other (Tobler's law, 1970). Consequently, neighboring observations cannot be considered distributed independently and randomly, but rather are characterized by various degrees of spatial association. It is the existence of spatial dependence that underpins spatial interpolation that, in turn, makes scaling issues come into play (Grayson and Blöschl, 2000).

Scaling refers to the transfer of information across scales. In this chapter, scaling is considered to be the change in scale from the observations to the predictions and is required whenever estimates are needed at a scale different from that at which the data were collected or when spatial processes occur over a broad spectrum of scales that can show nested or hierarchical structures (Wu and Li, 2009). In multi-scaling, the scales and a set of scaling thresholds should be identified to determine at which scale(s) the analysis should be conducted. Scale sensitivity analysis and robust scaling methods help make this choice and develop spatial models that enable scale effects to be compensated for (Behrens et al., 2019). There is evidence from recent investigations that the use of multiple scales, when modeling an environmental process, has the potential to enhance the accuracy and precision of estimates (Behrens et al., 2019; Biswas et al., 2013a). There are cases, however, where it is unnecessary to take all scales into account and it suffices to concentrate on the dominant process and the associated scale (Behrens et al., 2019). In practice, the choice of a scale is usually influenced by an already pre-selected scale in the system under study and the objectives of the investigation (De Boer, 1992). A few scales only are likely to be important, nevertheless those that are less influential should not be completely neglected. It is crucial to understand which dominant process governs a given phenomenon, and if the spatial models corresponding to the various scales are sufficiently different to be considered independent (Phillips, 1988). The problem of identifying the optimal measurement scale for an investigation to be carried out remains.

Even though scaling involves changes in all three components of the previously mentioned triplet (Malone et al., 2013), in this chapter we will focus on the change of support. Spatial data are often acquired by sampling the area of interest at several positions and monitoring several state variables. In some cases, the measurement locations can be considered point-like (e.g. the rainfall amount measured by a rain gage or soil moisture measured by a probe embedded to a certain soil depth), whereas other sampling activities refer to units with finite geometric dimensions larger than a point (e.g. sample plots in a forest inventory or pixels in remote-sensing images). Finite-sized supports are common in environmental analyses based on areal units, such as forest stands or

agricultural plots. When spatial data are aggregated, the support of a variable is changed and a transformed variable is generated, which is related to the previous one but has different statistical and spatial characteristics (Castrignanò and Buttafuoco, 2020). The change of support is a key topic in quantifying scale effects and will be discussed in more detail in the section on geostatistics.

Spatial models of scale

In recent years, SIT has steadily progressed to become a more cross-cutting discipline which results not only in the production of huge spatial data sets but also in a variety of acquisition tools. An unprecedented amount of Earth surface data is currently provided not only from satellites and unmanned aerial systems (UASs) but also from critical zone observatories (CZO) and open-air laboratories. Moreover, these datasets are now available at a relatively fine spatial resolution, even at submetric scales with UASs, which can be used as auxiliary variables to enhance the information contents of sparse ground-based data (Qu et al., 2022).

Therefore, there is a need to define a set of techniques capable of analyzing a huge amount of multi-source data with different spatio-temporal and statistical characteristics. These techniques should take into account scale dependencies and the diversity of data sources, and be able to manage them efficiently and effectively to extract the maximum possible information. In other words, a coherent and efficient system of techniques is required for fusing large quantities of geospatial data. The scale, or scales, in a dataset should first be assessed and then the processes that produced those data should be modeled. After that, the spatial model should adapt the scale of estimation to that of the data. The many techniques available for this purpose include fractals, geographical variance, local variance, wavelets, and spatial statistics. For a given investigation these techniques should make it possible to identify the appropriate scales of measurement, extract information, and perform a variety of spatial procedures on the data. All the above-mentioned techniques assume that the scale associated with the largest variance represents that at which most processes operate (Wu and Li, 2009). Fourier and spectral analyses as descriptors of scale have been widely used in topography and landscape analysis (Perron et al., 2008). Unlike Fourier analysis, the wavelet transform method (Biswas et al., 2013b) uses a wavelet function that can be adapted to the appropriate scale for a given data set. For example, to determine the most appropriate scale at which pedotransfer functions (PTFs) can estimate spatially-variable soil hydraulic properties, Pringle et al. (2007) used the wavelet technique to explore how the variances and correlations of both observed and estimated variables varied with the spatial scale. These authors showed that when a simplified method is used to estimate a certain variable, it is likely that the estimates are scale-dependent. A simplified method can mimic the actual spatial variation of the target variable when working at a relatively large scale (i.e. when the spatial fluctuations are small), whereas this capability becomes weak when considering the spatial variation at a relatively small scale (i.e. when the spatial fluctuations become rapid).

There are different ways to assess the similarity between neighboring observations, including spatial autocorrelation (Cliff and Ord, 1973). Before using traditional statistics, one has to check whether or not the available data are uncorrelated. In the case of spatial correlation, the covariance or correlation coefficient between two variables are no longer expressed by a number but by a mathematical function of the distance between pairs of observations such as the variogram and covariogram in geostatistics (Webster and Oliver, 2007).

Geostatistics as a coherent framework for scale modeling and scale change

Spatial dependence is an intrinsic feature of many spatial properties. The Earth's surface is characterized by varying degrees of spatial variability, such that no portion of it can be taken as representative of the whole. Spatial dependence and spatial heterogeneity are two sides of the same coin that interact with each other. A coherent and efficient system of spatial analysis is required to (1) quantify the scales of variation, (2) decompose the total spatial variation into the component attributable to the measurement and the process-specific component to determine the corresponding scales; (3) evaluate the differences in spatial variability at different scales, and (4) characterize the impact of information transfer across multiple scales. This section illustrates how geostatistics can respond to the above questions.

In geostatistics, scale specifically associated with the measurement process is called 'spatial support.' It is worth emphasizing that support should not be confused with resolution (Malone et al., 2013). The latter term is used in sparse sampling to address the average sampling distance and, in the case of raster data, is the size of the grid mesh or pixel. Here, the measurement scale will refer only to the support. Both support and resolution (depending on the density and arrangement of samples) affect estimation and the estimation variance. However, in this chapter, the focus is on support and not on sampling.

In many applications of spatial data analysis, the input dataset comprises measurements done on point supports, whereas in others they are carried out over a finite area or volume. Here, we will discuss the problem of how the spatial variation of an attribute relates to that of the same attribute when it is estimated on a different support (change-of-support problem). It is fundamental to understand how the spatial support relates to the scales of natural spatial variation, which are detectable through such a measurement but which also depend on the particular sampling pattern adopted. It is a complex subject. In this chapter, we restrict our analyses to how the sample support may affect the determination of intrinsic spatial variation. In practice, the values of properties are obtained from measurements on a given support, which represents a finite element of the space of a given size,

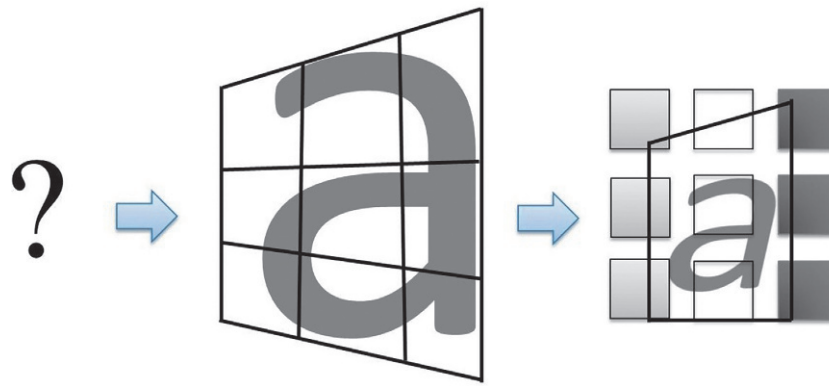


Fig. 1 View of reality through the sampling screen. Modified from Atkinson PM, Tate NJ (2000) Spatial scale problems and geostatistical solutions: A review. *The Professional Geographer* 52(4): 607–623.

geometry, and orientation. Assuming a continuous random field (Webster and Oliver, 2007), an observation $z_v(\mathbf{x}_0)$ can be considered a realization of a random variable (RV), $Z_v(\mathbf{x}_0)$, which is the spatial average or integral of Z over a finite support v centered on the point $\mathbf{x}_0$. Formally, it can be expressed as follows:

$$Z_v(\mathbf{x}_0) = \frac{1}{v} \int_{v(\mathbf{x}_0)} Z(\mathbf{y}) d\mathbf{y}, \quad (1)$$

where $Z(\mathbf{y})$ is the property Z defined on a point support. The spatial variation of Z is then averaged over the given support to produce a single mean value. It is important to distinguish between two fundamental but different scales: the scale or scales of measurement and the scale or scales of spatial variation that can be identified from the measurements.

Unfortunately, these two types of scales do not always coincide because ‘reality’ can only be observed through the implementation of a measurement system, consisting of the particular equipment used and the sampling pattern (number, locations, and support of samples), that produces spatial data that can subsequently be analyzed (Atkinson and Tate, 2000). This leads to two important considerations (Atkinson and Tate, 2000): we cannot observe ‘reality’ directly but only through the sampling screen (Fig. 1), so what we observe is not the reality (unknown) but always a filtered version of it.

Further, the scales of spatial variation can only be obtained from the analysis of data sampled at the given spatial scale(s) of measurement. The latter highlights how the scales of spatial variation obtained from the analysis of the data are intrinsically linked to the measurement scales of the sampling framework in which those data were collected. In the following, we will only assess the influence of support on spatial variation using geostatistical tools.

Multi-scaling

For a description of spatial variation with geostatistics, the interested reader is referred to manuals such as Goovaerts (1997), Wackernagel (2003), and Webster and Oliver (2007). Here, we will deal only briefly with those concepts most closely related to support. The dispersion variance $D^2(v, V)$ represents the variation of all data on a support v within a domain V , while the covariance function $C(\mathbf{h})$ represents the variation between pairs of data on support v [v, v_h] whose locations in the sampling frame are separated by the distance vector $\mathbf{h}$, called the lag.

Behrens et al. (2019) presented a method to define the effective scale interval with defined lower and upper scale limits for spatial modeling. They used the range of the variogram of a soil property of interest to determine the upper limit of the relevant range of scales, and the product of the nugget/sill ratio by the range to approximate the lower limit of the scales. Fig. 2 illustrates this with an example of how, according to the authors, the possible range of variation of scales can be determined. This lower limit is based on the assumption that the spatially uncorrelated variation of the soil property, represented by the nugget effect can be converted to the spatial variation at a given scale if the nugget is relatively large, which mainly results from too coarse a sampling scheme.

Furthermore, an experimental variogram may embrace more than one spatial scale and a nested model should be fitted that comprises a combination of multiple authorized models, each with its range representing a given scale of spatial variation (Webster and Oliver, 2007). An important consequence of the above is that the variogram can be used as a useful exploratory spatial tool to determine if multiple scales of spatial variation are present in a set of data (Fig. 3). Fig. 3 shows the nugget variance (A), the short-scale variation (B), the longer-scale variation (C), and the combination of the three components (D).

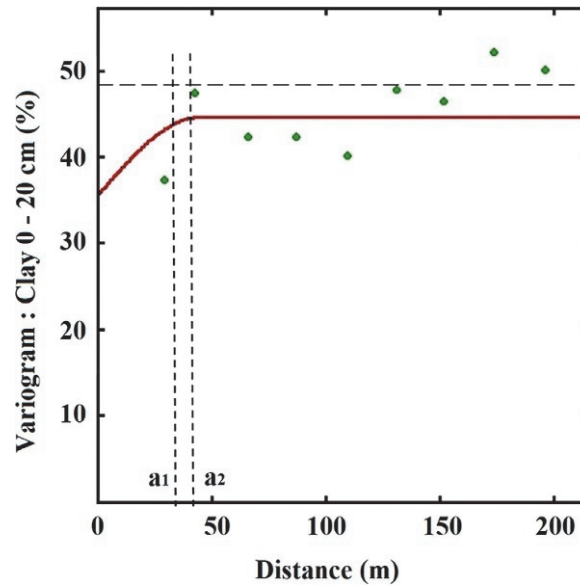


Fig. 2 An example to illustrate the way to obtain the relevant range of scales using a variogram model according to Behrens et al. (2019): the lower limit (a_1) and the upper limit (a_2) of the relevant range of scales.

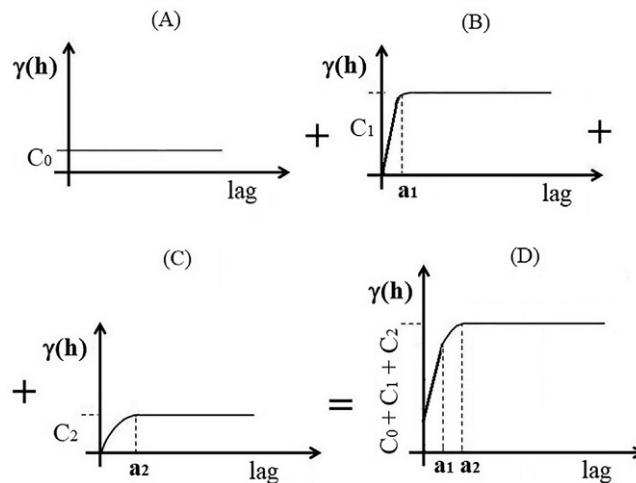


Fig. 3 Composition of a nested variogram.

Change of support

In this section, we consider in more detail how the change of support affects the scales of spatial variation that are observed. It is common in remote sensing, for example, to compare point on-the-ground measurements of a variable, i.e. measured on supports that are small compared to the pixel size of data from a raster image whose support is the pixel. If the aim is to model the relationship between the two variables, e.g. through regression or techniques in machine learning (Jin et al., 2021), and then use the relation to predict the point variable at unsampled locations, the accuracy of the prediction is expected to be quite small. The relationships between the variables are strongly scale-dependent (King, 1991), therefore such a strategy is not recommended. A proper procedure would involve rescaling or upscaling the point measurements (i.e. the small scale at which the variables are observed) to the pixel scale (i.e. the larger scale where a model applies) before a comparison is made with the raster variable. To obtain mean values of the point variable at the pixel scale, one or more samples per pixel should be available, which is rarely the case. One way of solving the problem could be to use interpolation by punctual kriging (Goovaerts, 1997), enabling the variable to be estimated at a very fine scale while preserving the original point support. The kriged point estimates could be averaged over the larger support of the raster variable. However, this is unsatisfactory because the averaging process changes the cumulative density function (CDF) of the ground-measured variable and thus also the bivariate distribution function (BDF) with the remotely sensed variable. The CDF calculated on the larger support will have a smaller variance than the 'true' CDF and this will also affect the BDF.

The first- and second-order statistical moments of a random variable on support B , $Z(B)$, are related to those of a variable on a point support, $Z(s)$ in a domain D , by the following relations (Castrignanò and Buttafuoco, 2020):

$$E(Z(s)) = \mu, \text{ then } E(Z(B)) = \mu \quad (2)$$

where μ is the expected value of point and block variables, and

$$\bar{C}(B_i, B_j) = \text{cov}(Z(B_i), Z(B_j)) = \frac{1}{|B_i||B_j|} \int_{B_i} \int_{B_j} C(s, u) ds du \quad (3)$$

where $C(s, u) = \text{cov}[Z(s), Z(u)]$, $u, s \in D$, is the covariance of the variable on a point support.

The covariance on the block support, $\bar{C}$, depends not only on the covariance on the point support, but also on the support B . The relationship between the covariance on block support B and that on the point support s is a fundamental issue in the solution of the change of support problem. Eq. (3) was theoretically derived by Journel and Huijbregts (1978) and allows the transfer from the point variogram to the block variogram. The latter is calculated at large distances from the former by subtracting a term that can be interpreted as the variance of the block:

$$\bar{\gamma}_B(h) = \gamma(h) - \bar{\gamma}(B, B) \quad (4)$$

where $\bar{\gamma}_B(h)$ is the variogram on block support, $\gamma(h)$ is that on point support, and $\bar{\gamma}(B, B)$ is the mean value of the variogram within the block and is calculated as the average value of the point-support variogram between two arbitrary points within block B . To obtain the block support covariances (or semivariances), the point support covariances $C(s, u)$ (or semivariances) must be known or estimated from the available point-support data. The block-averaged semivariance is obtained by discretizing the block into points and then calculating the arithmetic mean of the point semivariance. The process of deriving the block semivariance or variogram from the point variogram is called regularization, while the reverse process (from block to point) is called de-regularization (see Fig. 4).

When point-support data are not available, one can resort to an iterative procedure (Goovaerts, 2008), which searches for the point support model by minimizing the difference between the theoretical regularized variogram model and the model adapted to the data on block support.

Once a satisfactory point model of the variogram has been fitted, it can be regularized to any pixel size (Eq. 4). This will enable one to jointly analyze (data fusion) a virtually infinite number of covariates from heterogeneous sources of information with any support. Eq. (4) thus provides a geostatistical method for rescaling the variogram model rather than the data. The model fitted to the regularized variogram may differ from that for point support in both type and parameters. Therefore, it is possible to evaluate the effect of change in size, shape, and orientation of the support on the nature and scale of the estimated spatial variation. Moreover, Eq. (4) shows that the effect of regularization is to remove the within-block variance, i.e. small-scale or high-frequency variation. If there is considerable variation at a small scale compared to the support, then much variation will be removed by upscaling. Conversely, if the variation at a scale greater than the support is large, only a small proportion of the total variance will be suppressed. It is worth emphasizing that any variation in the measurement support has a quantifiable effect on the understanding of the process.

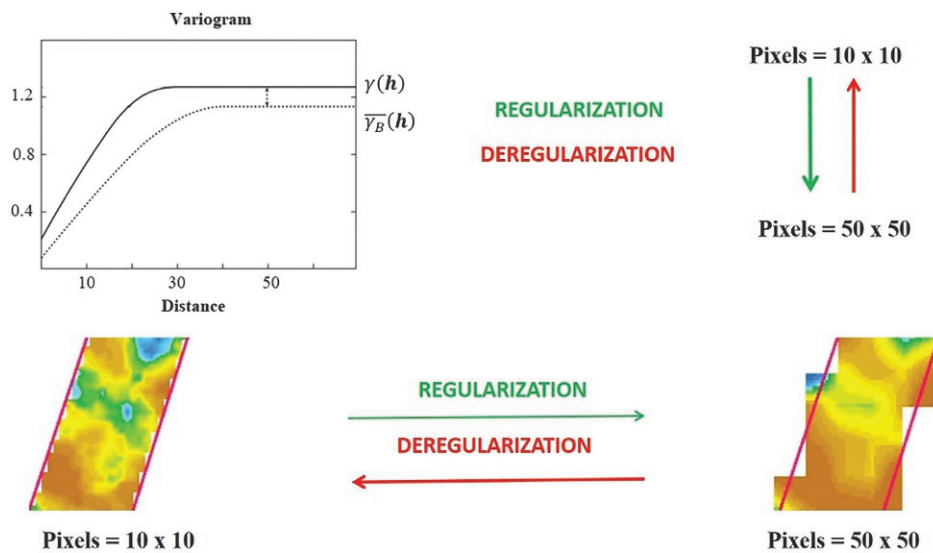


Fig. 4 An example to illustrate how change of support affects the variogram.

Geostatistical fusion of heterogeneous spatial data

The modern era is marked by a large availability of data from different sources with diverse spatial, temporal, and statistical characteristics. It is therefore very common for spatial models to use a variety of data as inputs, among which the ones recorded with different proximal and or remote sensors. Multi-source data imply the presence of multiple-scales as intrinsic to the information flow (Zhang et al., 2015), however, the process of integrating multiple scales and fusing heterogeneous information is not at all trivial. To take advantage of the complementary features of the different types of data, they should be combined in an efficient and statistically robust way. A solution to this problem can be sought by resorting to multivariate geostatistics, although this method is not the only one that can be found in the literature.

Multivariate spatial correlations can be modeled within the scope of the linear model of co-regionalization (LMC) (Webster and Oliver, 2007), which considers all n studied variables as the manifestation of the same independent physical processes, acting over a number (N_S) of spatial scales u . The $n(n+1)/2$ simple and cross variograms of the variables are modeled by linear combinations of N_S standardized variograms of unit sill, $g^u(h)$ which are the same for all variables for each scale u .

Spatial modeling with a change of support is commonly treated in geostatistics by applying block kriging, which is one of the traditional geostatistical interpolation methods to tackle this problem. From point support observations, block kriging can be used to predict the average values at a larger support, accounting for the block's attributes (size, shape, and orientation). The main difference between block kriging and point kriging is the calculation of the block-to-point covariance:

$$\overline{C}(B, \mathbf{x}_z) = \text{cov}(Z(B), Z(\mathbf{x}_z)) = \frac{\int_B C(\mathbf{u}, \mathbf{x}_z) d\mathbf{u}}{|B|}, \quad (5)$$

where $|B|$ is the measure of the block (spatial support), $\mathbf{x}_z$ the point location and $C(\cdot)$ is the point-to-point covariance. For multivariate data sets, cokriging rather than kriging is used, which enables co-estimation of poorly sampled variables from secondary more exhaustive information related to the target variables (Webster and Oliver, 2007).

Although cokriging equations are valid regardless of the support of the data, it is essential to consider the different supports in calculating the autocovariance and cross-covariance functions for valid inference (Castrignanò et al., 2018). Once a consistent model for the point covariance functions is estimated, block cokriging can be applied by replacing the covariances in punctual cokriging with their block-averaged counterparts through regularization of the covariance (variogram).

Multivariate factorial kriging is a variant of cokriging (Castrignanò et al., 2000; Wackernagel, 2003), which aims to estimate the different sources of spatial variation common to a set of variables. This amounts to calculating N_S (number of scales in the LMC) sets of ortho-normal regionalized factors, one set for each spatial scale, as a linear combination of the n variables for each factor. Mapping the regionalized factors is a very effective way to synthesize the major features of the multivariate data set at the different spatial scales identified in nested experimental variograms (Castrignanò et al., 2017, 2018).

Application of geostatistics to multi-scaling and change of support in spatial modeling

The previous sections have shown the existence of a relationship between spatial dependence, as described by a variogram, and a set of spatial scales relevant to spatial modeling. Therefore, the variogram(s) of any variable, or set of variables in a multivariate approach, can be used as an analytical tool for multiscale spatial modeling in geostatistics. A key question in spatial modeling is whether it is more appropriate to use a single most relevant spatial scale or rather provide multiple representations of the same scene at different scales, corresponding to those identified by the spatial data analysis. We believe that it is more reliable to produce multiscale descriptions because the spatial relationships among variables vary as a function of scale. Once the overall spatial variance has been decomposed into its various spatial components, the one specific to the phenomenon under study could be selected (Manziona and Castrignanò, 2019; Castrignanò et al., 2021; Buttafuoco et al., 2021).

Case study

The case study aims to explore the use of multispectral data from an unmanned aerial vehicle (UAV) on-board radiometer combined with sparse point data from a ground-based sample to improve the prediction accuracy of soil textural components. The site was in a relatively flat olive grove in southern Italy consisting of centenarian trees. The planting pattern is wide and somewhat irregular leaving large portions of soil bare or occupied by weeds as the weather evolves. This field has loose gravel-free soil, classified as a Calcic Haploxeralf, fine, mixed, thermic (USDA classification) and Calci-Chromic (Endoleptic) L (WRB classification).

Regular plowing and mulching were carried out for weed control, and no irrigation was applied. Topsoil (0–0.20 m) samples were collected in three different places near each of the 61 olive trees, within a 1-m by 1-m square, and mixed to produce a single composite sample (Fig. 5). Soil texture was determined with the pipette method in the laboratory to determine the silt and clay contents (%) of the soil samples. Sand content was calculated as the complement to 100 of the sum of silt and clay contents.



Fig. 5 Locations of topsoil samples (red dots) collected near each of the 61 trees of the olive grove.

Acquisition of reflected electromagnetic radiation with UAV

The aerial survey was conducted with a very light quadcopter (DJI Mavic Pro drone), equipped with an on-board global navigation satellite system (GNSS) and able to perform programmed missions over waypoints, on August 2nd, 2019. This date was chosen because the soil appeared dry and mostly free of weeds. A custom-designed payload tray carried the multispectral sensor Parrot Sequoia camera, which captures four-band spectral images in green (550 ± 20 nm), red (660 ± 20 nm), red-edge (735 ± 5 nm), and near-infrared (790 ± 20 nm) at a very fine spatial resolution of 0.07-m.

Before geostatistical processing, the multispectral pixels referring only to the soil were selected for use as auxiliary high-density information in the data fusion process. The selection of soil pixels was done by applying a supervised classification with ENVI (ENvironment for Visualizing Images) 5.1 software (Harris Geospatial).

Geostatistical analysis

To use the selected multispectral data as exhaustive (i.e., known everywhere) information, they were interpolated beforehand to the nodes of a predefined estimation grid.

To merge the multispectral data (0.07 m $\times$ 0.07 m support) with the soil sample data (1-m by 1-m support), it was necessary to transform them to the same support through the change of support procedures. Therefore, block cokriging was used as the interpolation technique for the four-band data. All the multivariate geostatistical analyses were undertaken using Gaussian transformed data standardized to mean 0 and variance 1. The spatial correlation structure of the multi-band UAV data was analyzed by adapting a linear model of co-regionalization (LMC) to the matrix of the four-band reflectance experimental variograms. Each variogram on a point (centimetric) support was then regularized to a corresponding variogram on the 1-m by 1-m block support, by discretizing each block into a number (10) of equal cells.

Block cokriging was used on the multi-band transformed data to produce estimates both at the nodes of a regular 1-m mesh interpolation grid, and at the locations of the soil samples to create the complete coregionalization data set. Both types of estimates were needed in the subsequent data fusion phase.

Multi-collocated cokriging was applied to the fused data set of the drone and soil sample data using the gridded auxiliary variables previously interpolated. Multi-collocated cokriging is a simplified version of full cokriging that uses the auxiliary gridded variables only at the interpolation point and at all locations where the soil variables are available. This approach is required when the auxiliary variables are particularly numerous for which resolution of the cokriging system is very time-consuming or even impossible.

The same LMC fitted for multi-collocated cokriging was also used in multi-collocated factorial cokriging analysis to extract a few synthetic scale-dependent indicators to be used for soil partitioning. Factorial cokriging is very effective in decomposing the total spatial variance of the fused data into its spatial components and for identifying the main scales of variation.

Mapping the retained scale-dependent regionalized factors, which explain most of the variance at the specific scales, provides a powerful and synthetic way to show the spatial relationships among several variables at the different spatial scales. Each map is a snapshot of the study scene at the particular scale chosen, disregarding all others (Castrignanò et al., 2017). The approach can be

used in soil science to produce a scale-dependent partition of the study area into homogeneous management zones. All geostatistical analyses were performed using the software ISATIS (Geovariances, France, 2018).

Results

There were 2,032,972 multi-spectral pixels selected as attributable to the soil only and to be fused with soil sample data. Table 1 gives the basic statistics of the four-band reflectance. They all have a skewness value greater than 1, which indicates that the data distributions were not centered.

To calculate the experimental variograms of the Gaussian transformed drone data, it was necessary to reduce the number of pixels by superimposing a regular 1-m grid on the study area and by choosing one point within each cell. The procedure retained 12,974 pixels. An LMC was jointly fitted to the ensemble of direct and cross-experimental variograms of the drone-transformed data; it comprised three basic spatial structures: a nugget effect, a spherical model with a range of 7.78 m, and a spherical model with a range of 76.59 m. Cross-validation showed that the model fitting was reasonable for all four variables. The total spatial variance was then divided into three components: the spatially uncorrelated error, accounting for 43%, and two spatially correlated components: the former at a shorter range accounting for 43% of the total spatial variance, and the latter at a longer range, accounting for 14%. These results show that almost all of the variation described by the drone data alone (approximately 86%) is attributable to variation over distances less than 8 m, approximately the average canopy size of an olive tree. The high-density information was used for downscaling the information on soil texture components from the sparse sampling at 61 locations to a sampling scale of about 15 m on average.

To fuse the drone data with soil sample data, all point experimental direct and cross-variograms were regularized on a 1-m $\times$ 1-m block. A block LMC was then fitted (Fig. 6), including the following basic spatial structures: a nugget effect, a spherical model with a range of 10.93 m, and a spherical model with a range of 77.59 m. The proportions of total spatial variance explained by the three components were: 41% for the nugget effect; 33% for short-range structure; 26% for long-range structure.

A comparison of the block LMC with the corresponding point LMC shows that the mathematical form of the basic structure has not changed, with only a slight increase in the ranges. However, the proportions of total variance explained by nugget and short-range structures were reduced in favor of an increase in the variance explained by the longer-range structure.

An expected significant reduction occurred in the total spatial variance from 5.26 for the point model to 2.93 for the block model (about 56%). The effect of upscaling was therefore a decrease in the spatially uncorrelated error (nugget effect) and total spatial variance.

From this point onwards all geostatistical analyses were based on the complete coregionalization data set, comprising the clay and silt variables, and the block cokriged estimates of the drone-transformed variables at the 61 locations. The variables were transformed to standardized Gaussian variables for multivariate analysis and an LMC was fitted to the matrix of experimental variograms including: nugget effect, a spherical model with a range of 16.79 m, and a spherical model with a range of 98.7 m, which explained 38, 30 and 32% of the total variance, respectively.

The spatially uncorrelated component is still quite large despite fusion with the drone data. To reduce this there is a need to intensify point sampling of the soil. The two spatial scales identified might be related to factors associated with the olive trees and intrinsic variation associated with the granulometric and structural soil properties.

Fig. 7 shows that all the direct variograms are reasonably structured, although for silt and clay the nugget effect remains rather large due to the coarse sampling. However, the spatial correlations between the textural and drone variables (cross-variograms) are quite weak, which can be deduced from the distance between the model (bold line) and the dashed line representing the intrinsic correlation (maximum spatial correlation) (Wackernagel, 2003).

There may be two main causes for the above situation. The first is the large difference between the pixel size of the drone data of 0.07 m, although upscaled to 1 m, and the sampling interval of the soil data of the order of tens of meters. The second cause may be the choice of broadband radiometer, which explored a wavelength range (approximately 530–810 nm) (vis–NIR), that does not correlate strongly with the soil textural properties, which mostly affect the 1000–2500 nm range (SWIR) of soil reflectance spectra (Riefole et al., 2022).

Fig. 8 shows the maps of silt and clay obtained by cokriging and that of sand obtained by the difference. A common feature of these maps is the prevalence of short-range variation. However, it is possible to distinguish some degree of spatial association in the

Table 1 Basic statistics of the four band reflectances (NIR, near-infrared).

Variable	Count	Minimum	Maximum	Mean	Median	Std. Dev.	Skewness	Kurtosis
Green	2,032,972	0.027	0.256	0.098	0.094	0.018	1.56	7.73
Red	2,032,904	0.021	0.362	0.133	0.128	0.030	1.35	6.33
Red_edge	2,032,972	0.028	0.401	0.183	0.177	0.034	1.15	5.78
NIR	2,032,972	0.032	0.558	0.235	0.228	0.041	1.08	5.65

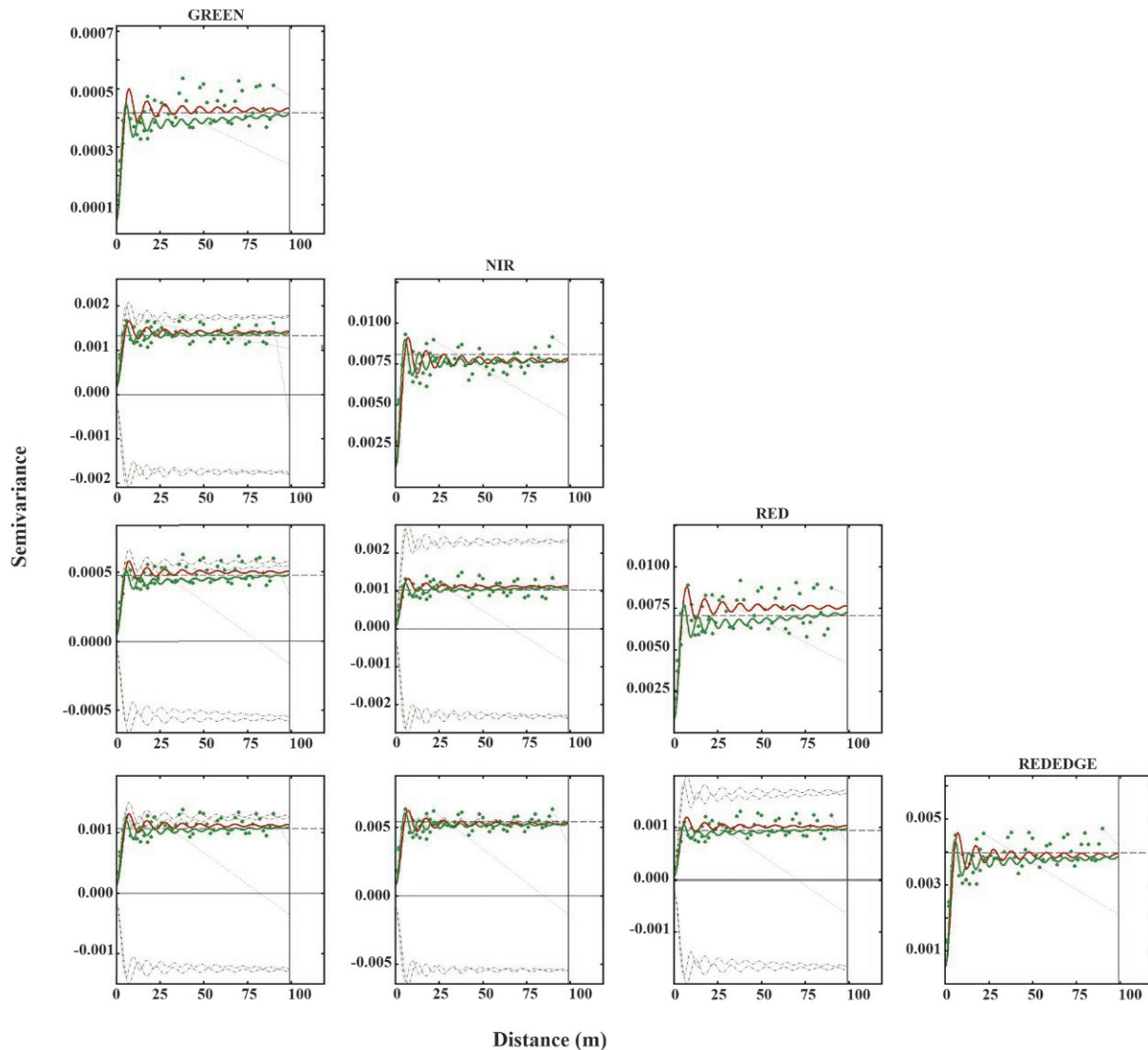


Fig. 6 Block LMC: dots: experimental variogram; bold line: model; dashed line: intrinsic correlation.

texture data whereby the particle size is predominantly finer in the north whereas the south has the largest sand content. This partitioning of the field was formerly detected with a GPR survey that revealed a large central–southern zone in which the reflected radar signal was stronger (Riefole et al., 2022).

To partition the soil into homogenous units, the first regionalized factor for the longer scale only (>98 m) was retained because it had an eigenvalue greater than 1 (2.06). In Fig. 9 this factor is represented using an isofrequency color scale with three classes.

The silt and clay variables have a negative influence on the factor, therefore, the red areas are those with finer grain composition, while the larger sand contents occur in the southeast and west.

For comparison, Fig. 10 shows the silt and clay maps obtained by a univariate approach.

Effectively, the latter seems less variable than those obtained by the fusion approach. The differences are not large because of the weak spatial correlation between the drone and sample data. The short-range variation introduced by the radiometric data could be due to properties (such as the presence of weeds) not strictly related to soil texture. Therefore, we recommend that for data fusion auxiliary variables should be chosen appropriately with supports not very different from those of the variable of interest. Otherwise, techniques of change of support should be applied. In addition, if radiometer data are used, the spectral ranges should include those bands that experimental results have verified to be correlated with specific (chemical and or physical) soil properties. Remote and proximal sensing data can be used as auxiliary variables to effectively support sparse soil sample data, but they cannot completely replace data from direct sampling.

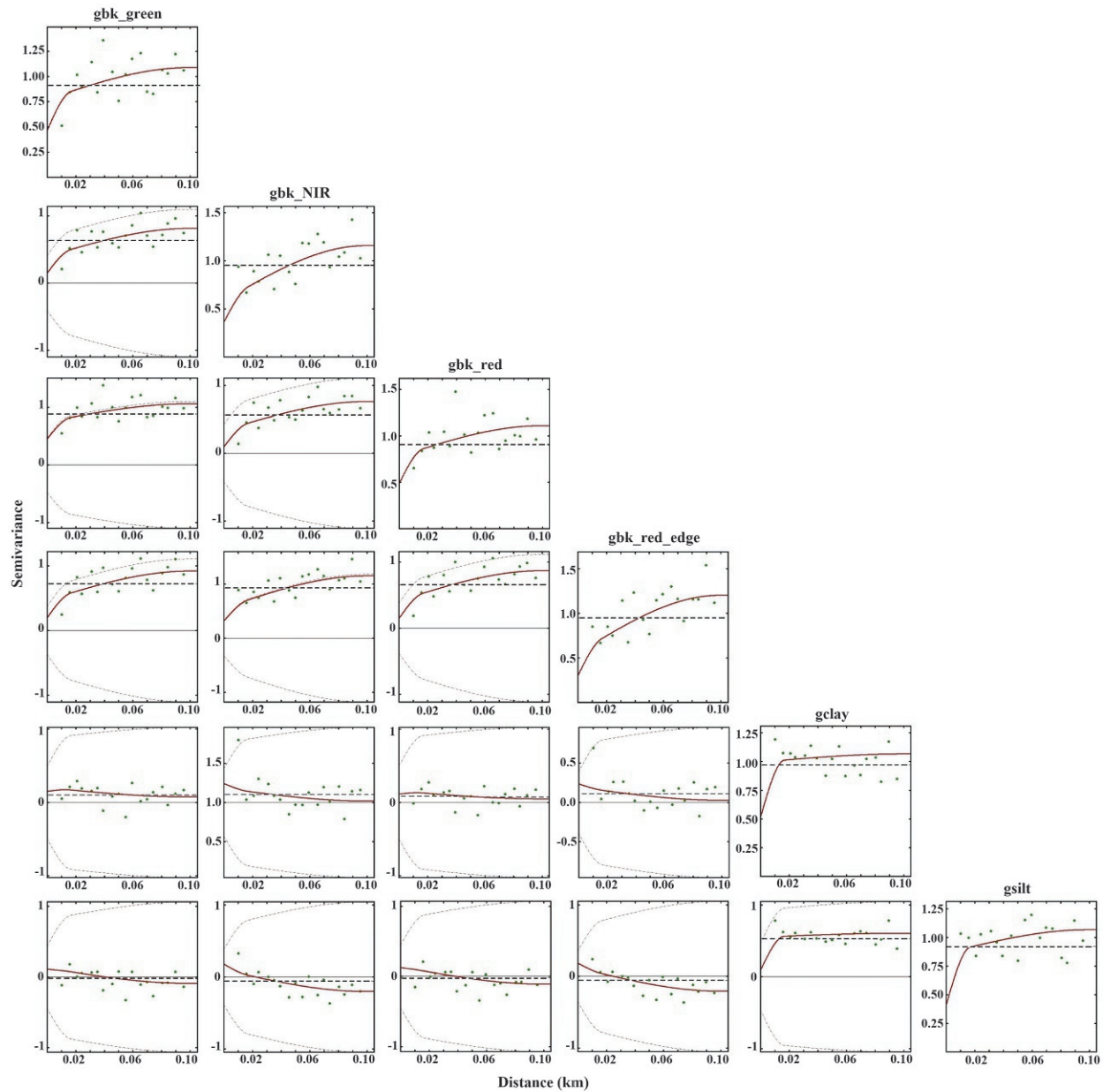


Fig. 7 The LMC of the whole data set of coregionalization (fusion).

Concluding remarks

Scale is a topic of great interest to various categories of environmental scientists for a multitude of reasons. Researchers often select a spatial scale, or rather a range of scales of measurement, to provide information on spatial variation. It is crucial to gain knowledge about the spatial structure of the property(ies) under study. This can be achieved by calculating the direct experimental variogram or the set of direct and cross experimental variograms in the multivariate approach, to determine the scale(s) of spatial variation in the sample data. If the size of sample support does not match the supports of other data sources, it is possible, once an LMC (linear model of correlation) has been fitted, to regularize (upscaling) or de-regularize (downscaling) the spatial model geostatistically to predict variation in the variable(s) at its/their actual scale(s) (change of support). This approach means that it is possible to obtain such information without having to repeat the measurements with different support, with the need only to modify the spatial functions (covariance/variogram).

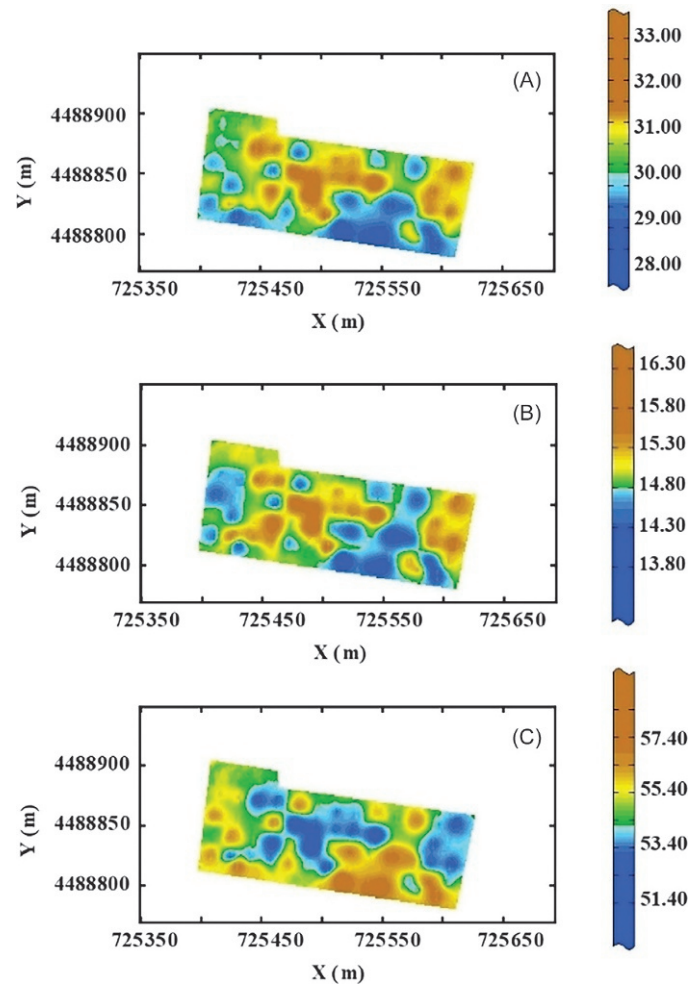


Fig. 8 The cokriging maps of silt (A), clay (B), and that derived for sand (C).

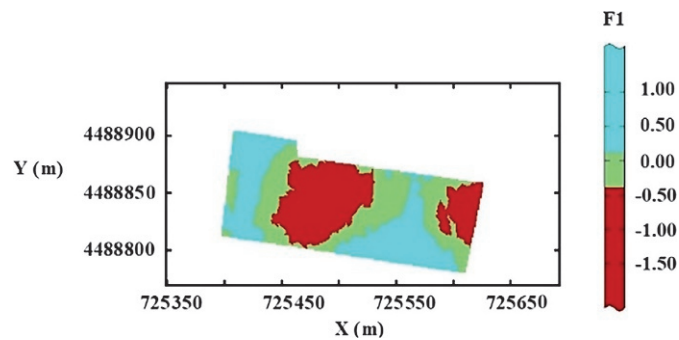


Fig. 9 The map of the first regionalized factor for the longer scale (98 m).

When the physical relations among variables are known, we recommend that they are integrated into the spatial model. This can be achieved by combining geostatistics with process-based modeling so that spatial dependence and spatial heterogeneity can be incorporated into mechanistic models, which describe the dynamics of the underlying processes that control spatial information.

An effective integration would improve understanding of the interaction between spatial dependence and the factors controlling multiple scales of spatial variation. Geostatistics and process models should form a unified framework for scale modeling and change of scale in controlling the flow of spatial information from measurement through data analysis to applications.

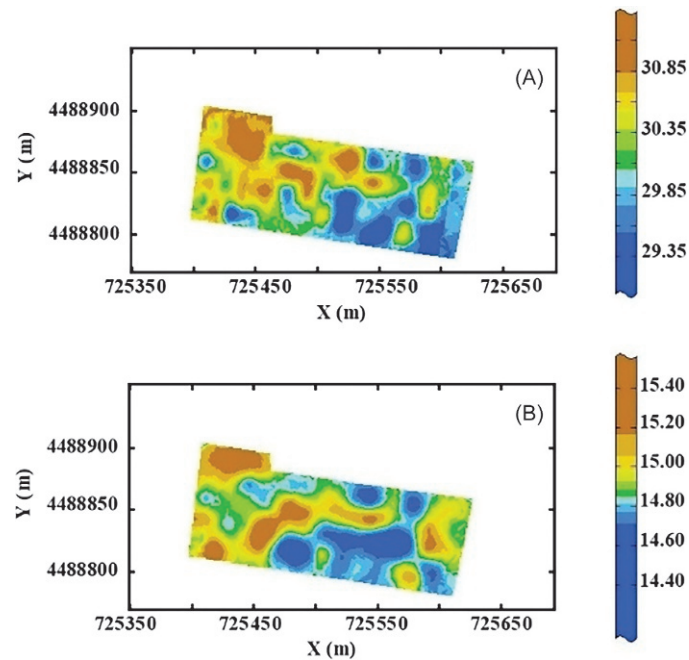


Fig. 10 The maps of silt (A) and clay (B) obtained by the univariate approach.

Last, but certainly not least, is the problem of uncertainty and its relation to scale (Heuvelink, 2014). Uncertainty is present everywhere in spatial information and it has the effect of weakening and causing bias in spatial relationships. Uncertainty has complex interactions with scale, as demonstrated in geostatistics where the variance of (co)kriging estimates depends explicitly on the nested variogram model. It has also been shown how the variance of a variable is a function of support.

Uncertainty in models has traditionally been treated with methods of error analysis based on error propagation. These assume spatial independence in error distribution, while spatial dependence is characteristic of spatial data and spatial errors. A better strategy for modeling error is based on stochastic simulation (Chilès and Delfiner, 2012), which makes it possible to generate a very large number of equally probable realizations in which the intrinsic spatial structure of the scene under study is reproduced, and the sample data are honored at their locations in the case of conditional simulation. In this respect, geostatistical simulation with its various procedures can make a valuable contribution to the assessment of the interaction between scale and uncertainty.

References

- Atkinson PM and Tate NJ (2000) Spatial scale problems and geostatistical solutions: A review. *The Professional Geographer* 52(4): 607–623. <https://doi.org/10.1111/0033-0124.00250>.
- Behrens T, Viscarra Rossel RA, Kerry R, MacMillan R, Schmidt K, Lee J, Scholten T, and Zhu A-X (2019) The relevant range of scales for multi-scale contextual spatial modelling. *Scientific Reports* 9(14800): 1–9.
- Biswas A, Cresswell HP, Chau HW, Rossel RAV, and Si BC (2013a) Separating scale-specific soil spatial variability: A comparison of multi-resolution analysis and empirical mode decomposition. *Geoderma* 209–210: 57–64. <https://doi.org/10.1016/j.geoderma.2013.06.003>.
- Biswas A, Cresswell HP, Viscarra Rossel RA, and Si BC (2013b) Characterizing scale- and location-specific variation in non-linear soil systems using the wavelet transform. *European Journal of Soil Science* 64: 706–715. <https://doi.org/10.1111/ejss.12063>.
- Blöschl G and Sivapalan M (1995) Scale issues in hydrological modelling: A review. *Hydrological Processes* 9: 251–290.
- Buttafuoco G, Quarto R, Quarto F, Conforti M, Venezia A, Vitti C, and Castrignanò A (2021) Taking into account change of support when merging heterogeneous spatial data for field partition. *Precision Agriculture* 22: 586–607. <https://doi.org/10.1007/s11119-020-09781-9>.
- Castrignanò A, Belmonte A, Antelmi I, Quarto R, Quarto F, Shaddad S, Sion V, Muolo MR, Ranieri NA, Gadaleta G, et al. (2021) A geostatistical fusion approach using UAV data for probabilistic estimation of *Xylella fastidiosa* subsp. *pauca* infection in olive trees. *Science of the Total Environment* 752.
- Castrignanò A and Buttafuoco G (2020) Data processing. In: Castrignanò A, Buttafuoco G, Khosla R, Mouazen AM, Moshou D, and Naud O (eds.) *Agricultural Internet of Things and Decision Support for Precision Smart Farming*, 1st edn, pp. 139–182. Cambridge, MA, USA: Academic Press. ch. 3.
- Castrignanò A, Giugliarini L, Risaliti R, and Martinelli N (2000) Study of spatial relationships among soil physical-chemical properties using Multivariate Geostatistics. *Geoderma* 97: 39–60.
- Castrignanò A, Buttafuoco G, Quarto R, Vitti C, Langella G, Terribile F, and Venezia A (2017) A combined approach of sensor data fusion and multivariate geostatistics for delineation of homogeneous zones in an agricultural field. *Sensors* 17: 1–20.
- Castrignanò A, Buttafuoco G, Quarto R, Parisi D, Viscarra Rossel RA, Terribile F, Langella G, and Venezia A (2018) A geostatistical sensor data fusion approach for delineating homogeneous management zones in Precision Agriculture. *Catena* 167: 293–304.
- Chilès JP and Delfiner P (2012) *Geostatistics: Modeling Spatial Uncertainty*, 2nd edn. Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Cliff AD and Ord JK (1973) *Spatial autocorrelation*. London: Pion.

- De Boer DH (1992) Hierarchies and spatial scale in process geomorphology: A review. *Geomorphology* 4(5): 303–318.
- Gentine P, D'Pdorico P, Linter BR, Sivandran G, and Salvucci G (2012) Interdependence of climate, soil, and vegetation as constrained by the Budyko curve. *Geophysical Research Letters* 39: L19404. <https://doi.org/10.1029/2012GL053492>.
- Goovaerts P (1997) *Geostatistics for Natural Resources Evaluation*. New York, NY, USA: Oxford University Press, p. 483.
- Goovaerts P (2008) Kriging and semivariogram deconvolution in the presence of irregular geographical units. *Mathematical Geosciences* 40: 101–128. <https://doi.org/10.1007/s11004-007-9129-1>.
- Grayson R and Blöschl G (2000) *Spatial Patterns in Catchment Hydrology: Observations and Modeling*. (404 p) Cambridge: Cambridge University Press.
- Heuvelink G (2014) Uncertainty quantification of GlobalSoilMap products. *GlobalSoilMap: Basis of the Global Spatial Soil Information System*. In: *Proceedings of the 1st GlobalSoilMap Conference* 335–340. <https://doi.org/10.1201/b16500-62>.
- Jin Y, Ge Y, Liu Y, Chen Y, Zhang H, and Heuvelink GBM (2021) A machine learning-based geostatistical downscaling method for coarse-resolution soil moisture products. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14: 1025–1037. [9247250] <https://doi.org/10.1109/JSTARS.2020.3035386>.
- Journel AG and Huijbregts CJ (1978) *Mining Geostatistics*. London: Academic Press.
- Kavvas ML (1999) On the coarse-graining of hydrologic processes with increasing scales. *Journal of Hydrology* 217: 191–202.
- King AW (1991) Translating models across scales. In: Turner MG and Gardner RH (eds.) *Quantitative Methods in Landscape Ecology, Ecological Studies*, 82nd edn, pp. 479–517. New York: Springer Verlag.
- Legendre P, Dale MRT, Fortin M-J, Gurevitch J, Hohn M, and Myers D (2002) The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography* 25: 601–615.
- Loague K and Corvin DL (2007) Scale issues. In: Delleur JW (ed.) *The Handbook of Groundwater Engineering*, 2nd edn. Boca Raton (FL, USA): CRC Press, Taylor & Francis Group. ch. 25. 25-1/25-21.
- Malone BP, McBratney AB, and Minasny B (2013) Spatial scaling for digital soil mapping. *Soil Science Society of America Journal* 77: 890–902. <https://doi.org/10.2136/sssaj2012.0419>.
- Manzoni RL and Castrignanò A (2019) A geostatistical approach for multi-source data fusion to predict water table depth. *Science of the Total Environment* 696: 133763. <https://doi.org/10.1016/j.scitotenv.2019.133763>.
- McBratney AB (1998) Some considerations on methods for spatially aggregating and disaggregating soil information. *Nutrient Cycling in Agroecosystems* 50: 51–62.
- Nimmo JR, Perkins KS, Plamper MR, Walvoord MA, Ebel BA, and Mirus BB (2021) Rapid-response unsaturated zone hydrology: small-scale data, small-scale theory, big problems. *Frontiers in Earth Science* 9: 613564.
- Perron JT, Dietrich WE, and Kirchner JW (2008) Controls on the spacing of first-order valleys. *Journal of Geophysical Research* 113(F4). <https://doi.org/10.1029/2007JF000977>.
- Phillips JD (1988) The role of spatial scale in geomorphic systems. *Geographical Analysis* 20: 308–317.
- Pringle MJ, Romano N, Minasny B, Chirico GB, and Lark RM (2007) Spatial evaluation of pedotransfer functions using wavelet analysis. *Journal of Hydrology* 333: 182–198.
- Qu C, Boubin J, Gafurov D, Zhou J, Aloysius N, Nguyen H, and Calyam P (2022) UAV swarms in smart agriculture: Experiences and opportunities. In: *IEEE 18th International Conference on e-Science, Salt Lake City, UT, USA* 148–158. <https://doi.org/10.1109/eScience55777.2022.00029>.
- Riefole C, Belmonte A, Quarto R, Quarto F, Ruggieri S, and Castrignanò A (2022) Potential of GPR data fusion with hyperspectral data for precision agriculture of the future. *Computers and Electronics in Agriculture* 199: 107109. <https://doi.org/10.1016/j.compag.2022.107109>.
- Romano N (2014) Soil moisture at local scale: Measurements and simulations. *Journal of Hydrology* 516: 6–20.
- Sposito G (1998) Scale dependence and scale invariance: Preface. In: *Scale Dependence and Scale Invariance in Hydrology*, pp. xi–xiii. New York, USA: Cambridge University Press.
- Stahl K and Hisdal H (2004) Hydroclimatology. *Developments in Water Science*. vol. 48, pp. 19–51. Amsterdam, the Netherlands: Elsevier Science B.V.
- Teuling AJ, Uijlenhoet R, Hupet F, van Loon EE, and Troch PA (2006) Estimating spatial mean root-zone soil moisture from point-scale observations. *Hydrology and Earth System Sciences* 10: 755–767.
- Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46: 234–240.
- Wackernagel H (2003) *Multivariate Geostatistics: An Introduction with Applications*. London, UK: Springer Nature. ISBN 13 9783540441427.
- Webster R and Oliver MA (2007) *Geostatistics for Environmental Scientists*, 2nd edn. Chichester: Wiley.
- Wu H and Li Z (2009) Scale issues in remote sensing: A review on analysis, processing and modeling. *Sensors (Basel, Switzerland)* 9: 1768–1793.
- Zhang Y, Du Y, Ling F, Wang X, and Li X (2015) Spectral–spatial based sub-pixel mapping of remotely sensed imagery with multi-scale spatial dependence. *International Journal of Remote Sensing* 36(11): 2831–2850. <https://doi.org/10.1080/01431161.2015.1047048>.

Sampling for soil survey

DJ Brus, Biometris, Wageningen University and Research, Wageningen, Netherlands

© 2023 Elsevier Ltd. All rights reserved.

Introduction	536
Probability sampling versus non-probability sampling	536
Probability sampling for estimating population parameters	537
Sampling designs	537
Estimators	538
Horvitz-Thompson estimator	538
Hansen-Hurwitz estimator	538
Ratio estimator	538
Model-assisted estimators	538
Simple random sampling	539
Stratified simple random sampling	539
Allocation of sample size to strata	540
Stratification	540
Systematic random sampling	540
Cluster random sampling	541
Balanced and well-spread sampling	542
Balanced sampling	542
Well-spread sampling	544
Sample size determination	544
Bayesian approach for sample size determination	544
Model-based optimization of probability sampling designs	545
Analytical approach	545
Simulation approach	545
Bayesian approach	545
Non-probability sampling for mapping	545
Geometric sampling designs	546
Adapted experimental designs	546
Model-based sampling designs	547
Optimization of grid spacing	547
Optimization of sampling pattern	548
Sampling for semivariogram estimation and mapping	549
Future research	549
References	550

Abstract

A sound sampling design is essential for the collection of data to support reliable scientific inference and decision making for management and policy. What counts as a sound design depends on the problem of interest, and the nature of the statistical inference that is required. Two broad aims can be distinguished, estimating the mean or total of a population or several subpopulations and mapping. For estimating (sub)population means or totals probability sampling and design-based or model-assisted estimation is most appropriate, whereas for mapping probability sampling is not required and there is scope to optimize the sampling pattern.

Key points

- Probability sampling is required for design-based estimation and model-assisted estimation of (sub)population parameters.
- For mapping probability sampling is not required, but still can be advantageous.
- Inclusion probabilities of population units need not be equal.
- Covariates can be exploited at the stage of designing a probability sample and/or at the estimation stage.
- Balanced sampling and well-spread sampling have potentials for soil survey.
- Geometric sampling designs for mapping do not require a prior model of the spatial variation.
- A prior model of the spatial variation can be used to optimize the sample size and/or the sampling pattern.

Introduction

To survey the soil data must be collected from the region of interest. We cannot afford to observe the soil everywhere, so in one way or another we must select locations where these observations will be done. Questions then, among others, are: (i) How do we define the elementary sampling units, are these points or areal sampling units, and in the case of areal sampling units, what is their shape and size? (ii) To what depth do we want to sample? (iii) How many locations should we select?, and (iv) How should we select these locations? This chapter focuses on the last two questions: the most relevant sampling designs for soil survey and methods for the determination of sample size are reviewed. For a more comprehensive text on how to design a soil sampling scheme I refer to [de Gruijter et al. \(2006\)](#). For more details on random sampling designs I recommend the following textbooks: [Cochran \(1977\)](#), [Särndal et al. \(1992\)](#), and [Lohr \(1999\)](#). [Brus \(2022\)](#) describes how the sampling designs and the estimation procedures can be implemented in the programming language R.

Probability sampling versus non-probability sampling

Sampling locations can be selected randomly or non-randomly. The term random sampling is often used for arbitrary or haphazard sampling, i.e. sampling without a specific purpose in mind. To avoid confusion in the sampling literature the terms probability sampling and non-probability sampling are used. Probability sampling is random sampling using a (pseudo) random number generator, such that all population units have a positive probability of being selected. Furthermore, these probabilities, referred to as the inclusion probabilities, must be known, at least for the selected units. As we will see hereafter, the inclusion probabilities are used in estimating population parameters such as the population mean, population total, or population proportion.

In non-probability sampling no such random number generator is used. Examples of non-probability sampling are (i) convenience sampling, i.e. sampling at places that are easy to access, e.g. along roads; (ii) arbitrary sampling, i.e. sampling without a specific purpose in mind; and (iii) targeted sampling, e.g. at sites suspected of soil pollution or at sites so that some criterion is minimized.

The choice between probability or non-probability sampling is closely connected with the choice between the design-based, model-assisted, or model-based approach for sampling and statistical inference. In the design-based approach and model-assisted approach units are selected by probability sampling ([Table 1](#)). Statistical inference is based on the sampling distribution of an estimator, which is determined by the sampling design. In the design-based approach no model is used in estimation (design-based inference). In the model-assisted approach a model is used to derive an efficient estimator, but statistical inference is not based on the model rather on the sampling distribution of the estimator. In the model-based approach a statistical model is used in prediction, i.e. a model with a random error term, for example a regression model or a kriging model. As the model already contains a random error term, probability sampling is not required in this approach. Statistical inference is fully based on the model distribution of the predictor. In contrast to the model-assisted approach, the model-based approach fully relies on the statistical model for statistical inference. This approach is also referred to as the model-dependent approach.

Which statistical approach is best depends largely on the aim of the survey. In general the following aims can be distinguished:

1. estimating parameters for the population;
2. estimating parameters for several subpopulations; and
3. mapping the study variable.

When the aim is to map the study variable, a model-based approach is the most natural option. This implies that for this aim probability sampling is not necessarily required. In principle, both approaches are suitable for estimating (sub)population parameters. The more subpopulations are distinguished, the more attractive a model-based approach becomes. Model-based estimates of the subpopulation means or totals are potentially more accurate (depending on how good the model is) than model-free design-based estimates. On the other hand, an advantage of design-based estimation is that an objective assessment of the uncertainty of the estimated mean or total is warranted, and that the coverage of confidence intervals is (almost) correct.

A probability sample can also be used in model-based inference. This flexibility can be attractive when we have a dual aim, mapping as well as estimation of parameters of (sub)populations. The balanced sample with geographical spreading shown in [Fig. 8](#) might be a good candidate for this dual aim. When units are not selected by probability sampling, model-free design-based estimation is impossible, and model-based prediction is the only option.

Table 1 Statistical approaches for sampling and inference.

<i>Approach</i>	<i>Sampling</i>	<i>Inference</i>
Design-based	Probability sampling required	Based on sampling distribution of estimator; no model used
Model-assisted	Probability sampling required	Based on sampling distribution of estimator; model used to derive estimator
Model-based	Probability sampling not required	Based on model distribution of predictor

Finally a warning. Do not be misled by publications stating that the design-based approach assumes that data are independent, and for that reason should not be used to sample populations showing spatial correlation. This is a persistent misconception: no such assumption is made and probability sampling is perfectly fine to sample such populations. See Brus (2021) for the details.

Probability sampling for estimating population parameters

Sampling designs

To estimate a population parameter probability sampling is most appropriate. The probability that a unit is included in the sample, in short the inclusion probability of that unit, can be calculated as the sum of the selection probabilities over all samples that can be selected with a given sampling design and that contain this unit. The formula is given by:

$$\pi_k = \sum_{S \ni k} p(S), \quad (1)$$

where $S \ni k$ indicates that the sum is over all samples that contain unit k , and $p(S)$ is the selection probability of sample S . $p(\cdot)$ is called the sampling design. It is a function that assigns a probability to every possible sample (subset of population units) that can be selected with a given sample selection scheme (sampling algorithm). For instance, it can be shown that for simple random sampling without replacement from a finite population the inclusion probability of every population unit is n/N , with n the sample size and N the total number of units in the population.

A common misunderstanding is that with probability sampling the inclusion probabilities must be equal. Sampling with unequal inclusion probabilities can be more efficient than with equal probabilities. Unequal probability sampling is not a problem as long as the inclusion probabilities are known and proper formulas are used for estimation.

There are many schemes for selecting a probability sample. The following sampling designs are described and illustrated in this chapter:

1. simple random sampling;
2. stratified random sampling;
3. systematic random sampling;
4. cluster random sampling;
5. balanced and well-spread sampling.

In stratified random sampling, balanced sampling, and well-spread sampling one or more ancillary variables are used in the random selection of the units. If these ancillary variables are related to the study variable, these sampling designs may provide a more precise estimate of a population parameter than simple random sampling at equal sample sizes.

The first four sampling designs will be illustrated with the Voorst study area in the Netherlands. Samples of the topsoil were collected at 132 points, which were analyzed for soil organic matter (SOM) concentrations (in g kg^{-1} dry soil) in a laboratory. The data were used in conditional geostatistical simulation of natural logarithms of the SOM concentration on a $25 \text{ m} \times 25 \text{ m}$ grid, followed by backtransformation. Fig. 1 shows the simulated map of SOM. This map is treated as the groundtruth in sampling experiments. At any location in Voorst the true SOM concentration is known, so that for any probability sample the mean SOM concentration can be estimated and compared to the true mean SOM concentration.

In the next sections I assume that the population units are points. The population is then infinite as the soil in an area is a continuum. With areal sampling units, for example circular sampling plots, that are allowed to overlap the population is also infinite. With infinite populations the distinction between without and with replacement sampling becomes immaterial and inclusion probabilities become inclusion probability densities. For instance, in simple random sampling of points from an infinite population the inclusion probability density is n/A for all points, where A is the area of the study area.

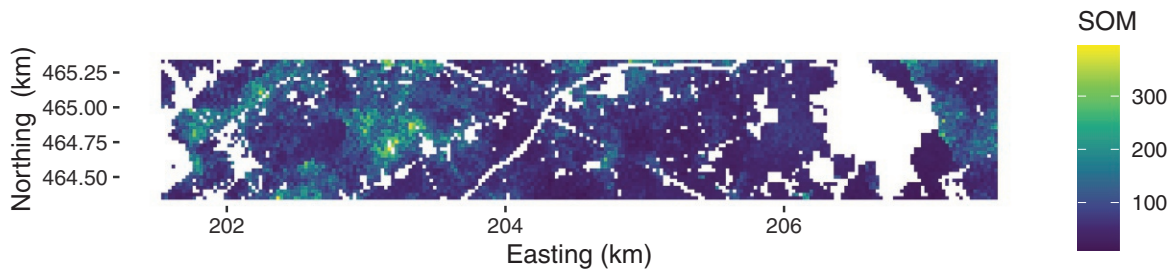


Fig. 1 Simulated map of the SOM concentration (g kg^{-1}) in Voorst.

Estimators

To estimate a population parameter several estimators are available: (i) the Horvitz-Thompson estimator; (ii) the Hansen-Hurwitz estimator; (iii) the ratio estimator; and (iv) model-assisted estimators.

Horvitz-Thompson estimator

For any probability sampling design the population total can be estimated as a weighted sum of the observations (measurements) of the study variable on the selected population units:

$$\hat{t}_{\pi}(z) = \sum_{k \in S} w_k z_k, \quad (2)$$

with S the sample, z_k the observed study variable for unit k , and w_k the design weight attached to unit k :

$$w_k = \frac{1}{\pi_k}, \quad (3)$$

with π_k the inclusion probability density of unit k . Hereafter the Horvitz-Thompson estimator is referred to in short as the π estimator. The π estimator for the population mean is simply obtained by dividing it by the size of the population, N for finite populations, A for infinite populations.

Hansen-Hurwitz estimator

In sampling finite populations, units can be selected with or without replacement. In sampling with replacement after each draw the selected unit is replaced. As a consequence a unit can be selected more than once. Sampling with replacement is less efficient than sampling without replacement. If a population unit is selected in a given draw, there is no additional information in this unit if it is selected again. One reason that sampling with replacement is still used is that it is easier to implement.

The most common estimator used for sampling with replacement is the Hansen-Hurwitz estimator, referred to as the pwr estimator hereafter. With direct unit sampling, i.e. sampling of individual population units, the pwr estimator of the population total is

$$\hat{t}_{\text{pwr}}(z) = \frac{1}{n} \sum_{k \in S} \frac{z_k}{p_k}, \quad (4)$$

with p_k the draw-by-draw selection probability of population unit k . In the pwr estimator observations of the selected units are expanded by the draw-by-draw selection probability. For example, in simple random sampling with replacement the draw-by-draw selection probability p of each unit is $1/N$. If we select only one unit k , the population total can be estimated by the observation of that unit divided by p , $\hat{t}(z) = z_k/p_k = Nz_k$. If we repeat this n times, it results in n estimated population totals. The pwr estimator is the average of these n elementary estimates. If a unit occurs multiple times in the sample S , this unit provides multiple elementary estimates of the population total.

Sampling with replacement can also be applied at the level of clusters of population units as in cluster random sampling and two-stage cluster random sampling. If the clusters are selected with probabilities proportional to their size and with replacement, estimation of the population mean, proportion or total is rather simple. This is a second reason why sampling with replacement can be attractive. With cluster sampling the pwr estimator of the population total is

$$\hat{t}_{\text{pwr}}(z) = \frac{1}{n} \sum_{j \in S} \frac{t_j(z)}{p_j}, \quad (5)$$

with $t_j(z)$ the total of the cluster selected in the j th draw.

Ratio estimator

With some sampling designs the sample size is not fixed but varies among samples selected with that design. An example is systematic random sampling from an irregular area. In this case an alternative estimator of the population mean is the ratio estimator. In the ratio estimator the π estimator of the population total is divided by the π estimator of the population size:

$$\hat{z}_{\text{ratio}} = \frac{\sum_{k \in S} w_k z_k}{\sum_{k \in S} w_k}. \quad (6)$$

The ratio estimator can also be efficient for sampling designs in which the points are selected with probabilities proportional to an ancillary variable.

Model-assisted estimators

The estimators described in the previous subsections use only data of the study variable. Over the past decades numerous large data sets have become available, such as those obtained by satellites and proximal sensors. These data sets may contain valuable information about study variables, so that they can be used to estimate global or local means and totals for these variables. Various

estimators have been developed that exploit these ancillary variables. They are jointly referred to as model-assisted estimators. A well-known example of this is the regression estimator:

$$\hat{z}_{\text{regr}} = \hat{z}_{\pi} + \sum_{j=1}^J \hat{b}_j (\bar{x}_j - \hat{x}_{j,\pi}), \quad (7)$$

with $\hat{z}_{\pi}$ and $\hat{x}_{j,\pi}$ the π estimator of the study variable and the j th covariate, respectively, $\bar{x}_j$ the population mean of the j th covariate, and $\hat{b}_j$ the estimated slope coefficient associated with the j th covariate, estimated by

$$\hat{\mathbf{b}} = \left(\sum_{k \in S} \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \sum_{k \in S} \mathbf{x}_k z_k. \quad (8)$$

A wide range of model-assisted estimators is available, see Brus (2022) for an overview. In situations where we have numerous covariates, and we do not want to rely on a linear relation between the study variable and these covariates, model-assisted estimators that build on machine learning techniques can be a good option.

An interesting application of model-assisted estimators is the estimation of the mean (total) of “small” subpopulations, i.e. subpopulations with a sample size that is too small to obtain a reliable estimate of the mean (total).

Simple random sampling

Simple random sampling is the most basic type of sampling design. All points have the same inclusion probability density. In addition, all samples of n points have the same probability of being selected. Fig. 2 shows a simple random sample of 40 points from the Voorst study area.

Working out the π estimator (Eq. 2) for simple random sampling yields the unweighted sample mean $\bar{z}_S$. The π estimate for the selected simple random sample of Fig. 2 is 93.3 g kg⁻¹.

The variance of the π estimator of the mean can be estimated by

$$\hat{V}(\hat{z}) = \frac{\hat{S}^2(z)}{n}, \quad (9)$$

with $\hat{S}^2(z)$ the estimated population variance:

$$\hat{S}^2(z) = \frac{1}{n-1} \sum_{k \in S} (z_k - \bar{z}_S)^2. \quad (10)$$

An estimate of the standard error is obtained by taking the square root of the variance estimate. The standard error has the same dimension as the study variable, and therefore is more intelligible than the variance. However, contrary to the variance estimator the estimator of the standard error is not unbiased. For the selected sample the estimated standard error is 9.604 g kg⁻¹.

Stratified simple random sampling

In stratified random sampling the study area is partitioned into subareas, referred to as strata. If the strata explain part of the spatial variation of the study variable, stratified random sampling is more efficient than simple random sampling, i.e. the π estimator of the population mean or total is more precise than the π estimator with simple random sampling and an equal number of points.

Fig. 3 shows a stratified random sample from the Voorst study area. Here, strata are combinations of soil types and land use classes. The points within the strata are selected by simple random sampling, so this is an example of a stratified simple random sample.

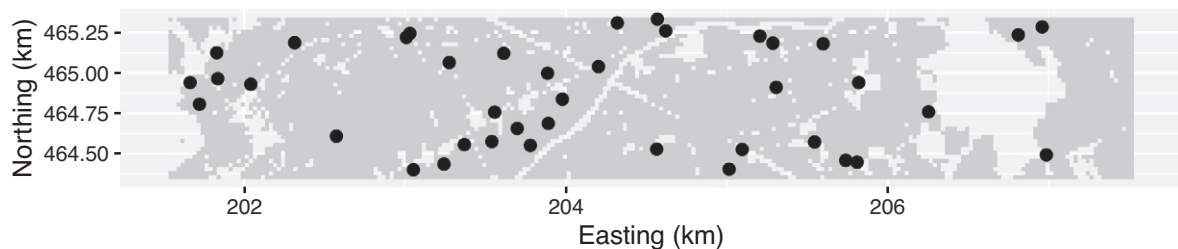


Fig. 2 Simple random sample of size 40 from the Voorst study area, The Netherlands.

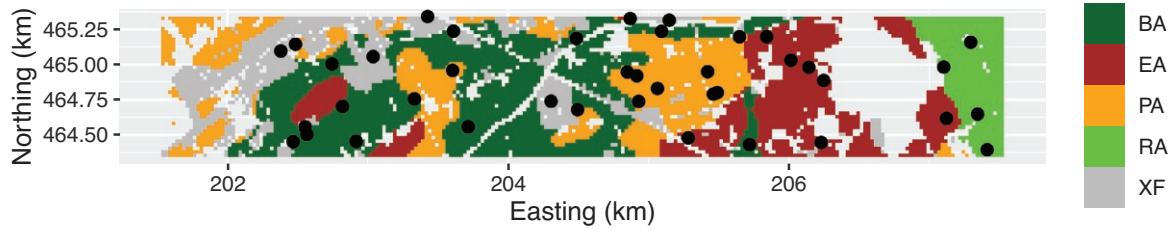


Fig. 3 Stratified simple random sample of size 40 from the Voorst study area, The Netherlands. Strata are combinations of soil class and land use.

Working out the π estimator of the population mean for this sampling design yields

$$\hat{z} = \sum_{h=1}^H w_h \hat{z}_h, \quad (11)$$

where H is the number of strata, w_h is the relative size (area) of stratum h (stratum weight): $w_h = A_h/A$ (A_h is the area of stratum h), and $\hat{z}_h$ is the estimated mean of stratum h , estimated by the sample mean for stratum h .

The variance of the π estimator of the mean can be estimated by

$$\hat{V}(\hat{z}) = \sum_{h=1}^H w_h^2 \frac{\hat{S}_h^2(z)}{n_h}, \quad (12)$$

with $\hat{S}_h^2(z)$ the estimated variance of z within stratum h and n_h the sample size of stratum h .

The design effect, defined as the ratio of the variance of the estimator of the mean with a design under study to the variance with simple random sampling, quantifies the relative efficiency of a design compared to simple random sampling. For design effects smaller than 1, the design under study is more efficient than simple random sampling. In the case of the Voorst study area, for stratified simple random sampling and proportional allocation (see hereafter) the design effect is 0.763. This implies that with simple random sampling we need $1/0.763 = 1.310$ times more sampling points than with stratified simple random sampling to obtain the same standard error.

Allocation of sample size to strata

After we have decided on the total sample size n , we must decide how to apportion the points to the strata. It is reasonable to allocate more sampling points to large strata and fewer to small strata. The simplest way to achieve this is proportional allocation, in which the stratum sample sizes are proportional to A_h .

If we have prior information on the variance of the study variable within the strata, then it makes sense to account for differences in variance. Heterogeneous strata should receive more sampling points than homogeneous strata. This leads to Neyman allocation in which the stratum sample sizes are proportional to $A_h S_h(z)$, with $S_h(z)$ the standard deviation of the study variable in stratum h .

Finally, the costs of sampling may differ among the strata. It can be relatively expensive to sample nearly inaccessible strata, and we do not want to sample many points there. This leads to optimal allocation in which the sample sizes are proportional to $(A_h S_h(z))/\sqrt{c_h}$, with c_h the cost per sampling point in stratum h .

Stratification

When we have a raster map of a quantitative covariate related to the study variable, strata can be constructed with the cum-root-f method using this covariate as a stratification variable. Raster cells with similar values for the covariate are grouped into a stratum. In cum-root-f stratification it is assumed that (after linear transformation) the covariate values are almost perfect predictions of the study variable, so that the prediction errors do not affect the stratification. Under this assumption the stratification is optimal.

If we have multiple maps of quantitative covariates, an option is to construct first homogeneous groups, referred to as clusters, of raster cells. The raster cells within a cluster are more similar to each other than to those in other clusters. Various clustering techniques are available, one of which is hard k -means.

Another option is to use the covariates as predictors in a regression model. The predictions of the study variable can then be used in cum-root-f stratification. The prediction errors are disregarded. The prediction error variances and covariances can be accounted for by the optimal spatial stratification method of de Gruijter et al. (2015).

In the absence of covariates we may still construct compact geographical strata. This can be done with the spatial coordinates of the raster cells as clustering variables in k -means. The geostrata improve the spreading of the sampling locations throughout the study area.

Systematic random sampling

A simple way of drawing a probability sample whose units are spread uniformly over the study area, is systematic random sampling. Systematic random sampling from a two-dimensional spatial population entails sampling on a regular grid, such as a square or

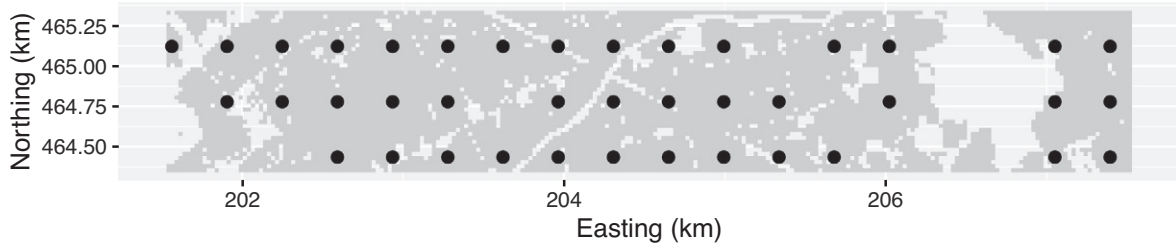


Fig. 4 Systematic random sample (square grid) from the Voorst study area, The Netherlands.

triangular grid. Quite a few countries in the world monitor the soil on a regular grid. The sample is well-suited for mapping. Furthermore, the sample data can be used in design-based estimation of (sub)population parameters. The geographical spreading of the sampling locations enhances the precision of the estimators. However, a disadvantage is that there is no unbiased estimator of the variance of the estimator of the population parameter.

Fig. 4 shows a systematic random sample of 40 points from the Voorst study area, The Netherlands. The origin of the square grid is selected randomly, the orientation is fixed. With systematic random sampling the sample size is random. We can control the expected sample size, i.e. the average sample over repeated random sampling, but for an individual sample the size may differ considerably from this expected sample size. With the grid spacing of Fig. 4 the sample size varies from 20 to 48.

With systematic random sampling the population mean can be estimated best by the ratio estimator, Eq. (6). For systematic random sampling the weights w_k are equal to $A/E[n]$, with $E[n]$ the expected sample size. With equal weights the ratio estimator is equal to the sample mean:

$$\hat{z}_{\text{ratio}} = \frac{1}{n} \sum_{k \in S} z_k. \quad (13)$$

The variance can be approximated only, for example by treating the sample as a simple random sample. This leads in general to overestimation of the variance. Alternatively, the sampling variance can be approximated by treating the systematic random sample as if it were a stratified simple random sample. The sampling units are clustered on the basis of their spatial coordinates into $H = n/2$ clusters (n even) or $H = (n-1)/2$ clusters (n odd). A third option is Matérn's method, see (Brus, 2022).

Systematic random sampling (in combination with the ratio estimator) is more efficient than simple random sampling. For the Voorst study area the design effect, quantified as the ratio of the variance of the ratio estimator of the mean with systematic random sampling to the variance with simple random sampling, equals 0.885. With simple random sampling we need 1.129 more sampling points to achieve an equal precision of the estimated mean.

Cluster random sampling

Figs. 5 and 6 show two cluster random samples from the Voorst study area. In Fig. 5 the clusters are E-W oriented transects within 1 km squares with an inter-point spacing of 100 m, whereas in Fig. 6 the clusters are 0.5 km squares, built-up areas, roads etc. are excluded. There is another important difference between the two cluster samples. In the first case all points of a selected cluster (transect) are observed, whereas in the second case a cluster is observed at a sample of points only. The first cluster sample is therefore also referred to as a one-stage cluster sample, and the second as a two-stage cluster sample. In two-stage cluster sampling the clusters are also referred to as primary sampling units (PSUs), and the points as secondary sampling units (SSUs).

In both cases the size of the clusters, i.e. the number of points, varies among the clusters. It is convenient to select the clusters with probabilities proportional to their size and with replacement (ppswr). Estimation of the population parameter is then straightforward (see hereafter). In the two-stage cluster random sample of Fig. 6 each time a PSU is selected, 10 points are selected by simple random sampling.

With ppswr sampling of clusters the population total can be estimated by the pwr estimator (Eq. 5). The draw-by-draw selection probability of a cluster for one-stage cluster random sampling is given by M_j/A , with M_j the number of points of cluster j , and A_j/A in

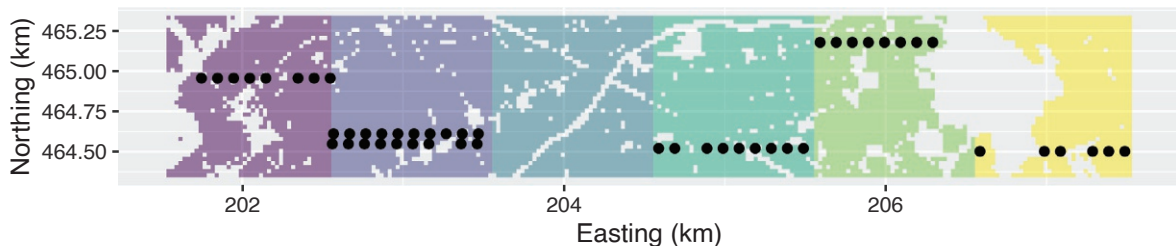


Fig. 5 Cluster random sample from the Voorst study area, The Netherlands, selected with probabilities proportional to size and with replacement (ppswr).

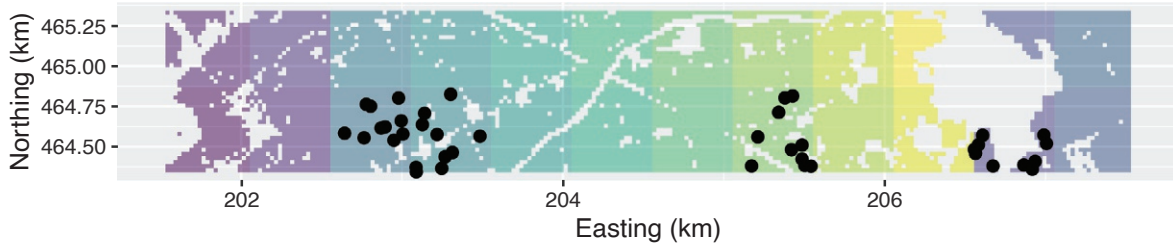


Fig. 6 Two-stage cluster random sample from the Voorst study area, The Netherlands. The primary sampling units (PSUs) are 0.5 km squares, built-up areas, roads, etc. are excluded. Four times a PSU is selected by ppswr. Each time a PSU is selected, 10 secondary sampling units (points) are selected from that PSU by simple random sampling.

two-stage cluster random sampling. This yields for one-stage cluster random sampling the following estimator for the population mean.

$$\hat{\bar{z}} = \frac{1}{n} \sum_{j \in S} \bar{z}_j, \quad (14)$$

with n the number of cluster draws and $\bar{z}_j$ the mean of cluster j . For two-stage cluster random sampling the cluster means $\bar{z}_j$ are replaced by the estimated cluster means $\hat{\bar{z}}_j$.

The variance of the pwr estimator of the mean can be estimated by

$$\hat{V}(\hat{\bar{z}}) = \frac{\hat{S}^2(\bar{z})}{n}, \quad (15)$$

where $\hat{S}^2(\bar{z})$ is the estimated variance of cluster means (the between-cluster variance):

$$\hat{S}^2(\bar{z}) = \frac{1}{n-1} \sum_{j \in S} (\bar{z}_j - \hat{\bar{z}})^2. \quad (16)$$

For two-stage cluster random sampling the estimated variance of cluster means $\hat{S}^2(\bar{z})$ is replaced by the estimated variance of estimated cluster means $\hat{S}^2(\hat{\bar{z}})$.

In both cases the sampling points are strongly spatially clustered. In general this leads to some redundant information about the population parameter. For the Voorst study area, -with one-stage cluster random sampling the design effect equals 2.82, and with two-stage cluster random sampling 3.19. So both designs are far less efficient than simple random sampling. For the Voorst study area cluster random sampling is certainly not a good idea. It can be an attractive sampling design if substantial time is needed to access the sampling points, and can be saved by clustering the points, think for instance of a continent-wide survey of the soil.

Balanced and well-spread sampling

Balanced sampling

In balanced sampling one or more covariates that are related to the study variable are exploited. A sample is balanced on a covariate if the sample mean equals the population mean of the covariate. A balanced sampling design is one that selects balanced samples. Simple random sampling is not a balanced sampling design because for many simple random samples the sample mean of the covariate is not equal to the population mean.

Fig. 7 shows a simulated population with a spatial trend from the lower-left corner to the upper-right corner. The study variable is therefore correlated with Easting and Northing. The figure shows two balanced samples of four units, one on Easting, and one on Easting and Northing. The sample mean of Easting is 10 (the population mean of Easting), and for the second balanced sample the mean of Northing also equals 10. I repeated the sampling 1000 times, using equal inclusion probabilities, and computed the variance of the estimated mean of the study variable, estimated by the sample mean. Using Easting as a balancing variable strongly reduced the variance of the π estimator (Table 2). Adding Northing as a balancing variable further reduced the variance.

In the above example equal inclusion probabilities are used, but this is not a requirement. A more general definition of a balanced sampling design is as follows. A sampling design is balanced on variable x when for all samples generated by the design the π estimate of the mean of x equals the true mean of x . Similarly to the regression estimator balanced sampling exploits the linear relation between the study variable and one or more covariates. When using the regression estimator this is done at the estimation stage, whereas balanced sampling does so at the sampling stage. With a perfectly balanced sample the second term in the regression estimator (see Eq. 7) which adjusts the π estimator, equals zero.

Fig. 8 shows a sample of size 100 from the Hunter Valley, NSW, Australia, balanced on topographic wetness index (TWI). Equal inclusion probabilities are used. The population mean of TWI equals 5.41, the sample mean equals 5.40.

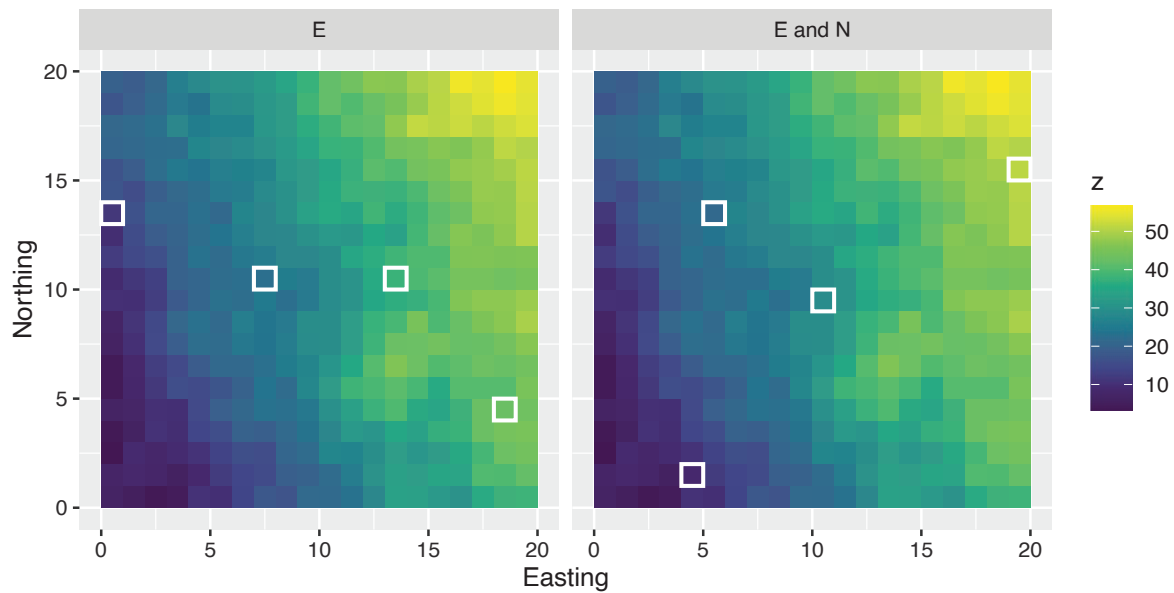


Fig. 7 Sample balanced on Easting (E) and on Easting and Northing (E and N).

Table 2 Sampling variance of the n estimator of the mean for simple random sampling (SI) and balanced sampling of four units.

Sampling design	Balancing variables	Variance
SI	–	39.7
Balanced	Easting	14.4
Balanced	Easting and Northing	9.77

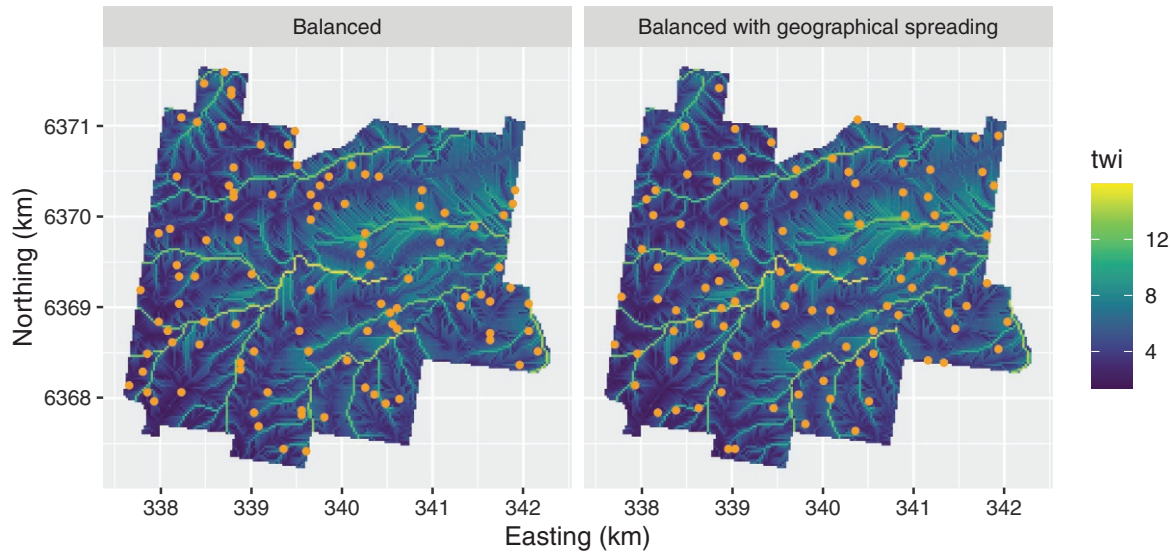


Fig. 8 Sample of 100 points from the Hunter Valley, NSW, Australia, balanced on topographic wetness index, without (left subfigure) and with geographical spreading.

Well-spread sampling

With balanced sampling the spreading of the sampling points in the space spanned by the balancing variables can be poor. For instance, in Fig. 7 the Easting coordinates of all units of a sample balanced on Easting can be equal or close to the population mean of 10. So, in this example balancing does not guarantee a good geographical spreading. Stated more generally, a balanced sample can be selected that shows strong clustering in the space spanned by the balancing variables. This clustering may inflate the standard error of the estimated population total and mean. Various sampling algorithms have been developed that avoid clustering of the sampling points in geographical or covariate space, such as generalized random tessellation stratification and the local pivotal method. Grafström and Tillé (2013) developed an algorithm for selecting probability samples that are both balanced and well-spread. Fig. 8 shows a sample obtained by using TWI in balancing and Easting and Northing as spreading variables.

Sample size determination

An important decision in designing sampling schemes is the number of points to select. If a certain budget is available for sampling, we can determine the affordable sample size from this budget. A costs model is then needed.

The alternative to deriving the affordable sample size from the budget is to start from a requirement based on the quality of the survey result obtained by statistical inference. The required sample size depends on the type of sampling design. With stratified random sampling we expect that we need fewer sampling units than for simple random sampling to estimate the population mean with the same precision, whereas with cluster random sampling in general we need more sampling units. To compute the sample size given some quality requirement, we may start with computing the required sample size for simple random sampling and then correct this sample size to account for the design effect.

The quality requirement can be expressed, for example, in terms of the standard error of the estimator or the length of a confidence interval. If we require that the length of the confidence interval of the population mean may not exceed a given limit $l_{\max}$, the required sample size for simple random sampling can be computed with

$$n = \left(u_{\alpha/2} \frac{S^*(z)}{l_{\max}/2} \right)^2. \quad (17)$$

with $u_{\alpha/2}$, the $(1 - \alpha/2)$ quantile of the standard normal distribution and $S^*(z)$ a prior estimate of the standard deviation of the study variable in the population.

A similar equation can be derived for the population proportion. Sample size determination (SSD) in this case requires a prior estimate of the population proportion itself.

The required sample size for testing a population mean with a two-sided alternative hypothesis can be computed by

$$n = \frac{S^2(z)}{\Delta^2} (u_{\alpha/2} + u_{\beta})^2, \quad (18)$$

with Δ the smallest relevant difference of the population mean from the test value, α the tolerable probability of a type I error, i.e. the probability of rejecting the null hypothesis when the population mean is equal to the test value, β the tolerable probability of a type II error, i.e. the probability of not rejecting the null hypothesis when the population mean is not equal to the test value, $u_{\alpha/2}$ as before, and u_{β} the $(1 - \beta)$ quantile of the standard normal distribution. The quantity $1 - \beta$ is the power of a test: the probability of correctly rejecting the null hypothesis. For a one-sided test, $u_{\alpha/2}$ must be replaced by u_{α} .

The SSD for testing a population proportion is not entirely straightforward. The power of a binomial test does not increase monotonically with the sample size. The graph shows a saw-toothed behavior. There are two options for sample size determination. The first is to compute the smallest sample size for which the power is larger than or equal to the required power $1 - \beta$. The alternative is to compute the smallest sample size for which the power is larger than or equal to $1 - \beta$ for all sample sizes larger than this.

Given an estimate of the design effect of an alternative sampling design, d_p , the required sample size for that design can be computed by

$$n_p = \sqrt{d_p} n_{SI}, \quad (19)$$

with n_{SI} the sample size for simple random sampling.

Bayesian approach for sample size determination

A serious drawback of the classical frequentist approach of SSD explained in the previous subsection is that the required sample sizes are sensitive to the so-called design parameters S^* and p^* (p^* is the prior estimate of the population proportion). We are rather uncertain about these parameters, and therefore it is attractive to replace a single value for these parameters by a probability distribution. This leads to the Bayesian approach for sample size determination. This Bayesian approach also offers the possibility of accommodating existing information about the population mean or proportion. Various Bayesian criteria for SSD are available, among other the average length criterion (ALC). This is the length of the highest posterior density interval with a coverage of $1 - \alpha$,

averaged over all possible data vectors that can be simulated with the prior distribution of the design parameter. The ALC is the Bayesian counterpart of the length of the confidence interval of the frequentist approach for SSD. For details I refer to Brus (2022).

Model-based optimization of probability sampling designs

If a statistical model is available, it can be used to predict the variance of an estimator of the population mean for any probability sampling design. This can be useful for SSD, optimizing parameters of a sampling design, for example the allocation of the total sample size to the strata with stratified simple random sampling, and for comparing alternative sampling design types. There are three approaches: analytical, simulation, and Bayesian. The variance predicted with a statistical model is sometimes referred to as the anticipated variance.

Analytical approach

In the analytical approach we assume that the mean is the same everywhere. The sampling variance of the π estimator of the mean is derived from mean semivariances within the study area and mean semivariances within the sample. Assuming isotropy, the mean semivariances are a function of the separation distance between pairs of points.

For any probability sampling design the sampling variance of the π estimator of the population mean can be predicted by

$$E_{\xi}\left\{V_p\left(\hat{\bar{z}}\right)\right\}=\bar{\gamma}-E_p\left(\mathbf{\lambda}^T\mathbf{\Gamma}_S\mathbf{\lambda}\right), \quad (20)$$

where $E_{\xi}(\cdot)$ is the statistical expectation over realizations from the model ξ , $E_p(\cdot)$ is the statistical expectation over repeated sampling with sampling design p , $V_p\left(\hat{\bar{z}}\right)$ is the variance of the π estimator of the population mean over repeated sampling with sampling design p , $\bar{\gamma}$ is the mean semivariance of the random variable at two randomly selected locations in the study area, $\mathbf{\lambda}$ is the vector of design-based weights of the units of a sample selected with design p , and $\mathbf{\Gamma}_S$ is the matrix of semivariances between the units of a sample S selected with design p .

This generic procedure is computationally demanding, but is the only option for complex spatial sampling designs. For basic sampling designs computing time can be saved by working out Eq. (20). For instance, for stratified simple random sampling the sampling variance can be predicted by

$$E_{\xi}\left\{V_{\text{STSI}}\left(\hat{\bar{z}}\right)\right\}=\sum_{h=1}^H w_h^2 \bar{\gamma}_h / n_h, \quad (21)$$

with H , w_h and n_h as before, and $\bar{\gamma}_h$ the mean semivariance of stratum h .

Simulation approach

The simulation approach for predicting the variance can also accommodate models in which the mean is a linear combination of covariates and/or spatial coordinates. Furthermore, this approach can be used to predict the sampling variance of the estimator of the mean of a trans-Gaussian variable, i.e. a random variable that can be transformed to a Gaussian variable. In the simulation approach many maps of the study variable are simulated, and each is used to compute the variance of an estimator of the population mean for the sampling design under study. The average of the variances over all simulated maps is a Monte Carlo approximation of the anticipated variance.

Bayesian approach

The model-based prediction of the variance of the design-based estimator of the population mean for a given sampling design is conditional on the model. If we change the model type or the model parameters, the predicted sampling variance also changes. In most situations we are quite uncertain about the model. Instead of using the best estimated model to predict the sampling variance as done in the previous sections, we may prefer to account for the uncertainty about the model parameters. This can be done through a Bayesian approach, in which legacy data are used to update a prior distribution of the model parameters to a posterior distribution. A sample from the posterior distribution of the model parameters is used one-by-one to predict the sampling variance. This results in a distribution of sampling variances which reflects our uncertainty about the sampling variance of the estimator of the population mean due to uncertainty about the model parameters. The mean or median of the distribution of sampling variances can be used as the predicted sampling variance.

Non-probability sampling for mapping

When the aim is map a variable using a statistical model of the spatial variation, probability sampling is not strictly required. This offers the opportunity of optimizing the sample pattern, i.e. the locations of the sampling points in the geographical and/or covariate space. Three groups of sampling designs for mapping can be distinguished: (i) geometric; (ii) adapted experimental; and (iii) model-based sampling designs. Examples of these three groups will be presented in the following sections.

Geometric sampling designs

Square and triangular grids are examples of geometric sampling designs; the sampling points show a regular, geometric spatial pattern. Contrary to systematic random sampling, there is no need to place the grid randomly on the study area. In other geometric sampling designs the spatial pattern is not perfectly regular. Yet these are classified as geometric sampling designs when the samples are obtained by minimizing some geometric criterion, i.e. a criterion defined in terms of distances between the sampling points and the nodes of a fine prediction grid discretizing the study area. If we use the mean of the squared shortest distance of the sampling points to the discretization nodes as a minimization criterion the clustering algorithm *k*-means can be used to compute the optimal spatial pattern. Fig. 9 shows a spatial coverage sample obtained by using the spatial coordinates of the raster cells in clustering, and a covariate space coverage sample using TWI, normalized difference vegetation index (NDVI) and elevation in clustering. In the latter case the covariates are scaled so that the covariates have equal weights in clustering.

Legacy data are easily accommodated in spatial coverage sampling and covariate space coverage sampling by using them as fixed cluster centers.

Adapted experimental designs

Experimental designs that are adapted for spatial surveys are the response surface and the Latin hypercube designs. Response surface experimental designs aim to find an optimum of the response within specified ranges of the factors. It is assumed that some type of low order regression model can be used to approximate accurately the relationship between the study variable and the factors. To make this experimental design suitable for observational research the covariates are decorrelated, using principal components analysis. In addition, spatial clustering of sampling points is avoided, so that we can safely assume that the residuals of the regression model are independent.

Latin hypercube sampling is used in designing, for instance, agricultural and computer experiments with numerous covariates and/or factors of which we want to study the effect on the output. With numerous covariates and/or levels per covariate/factor a full factorial design becomes unfeasible. A much cheaper alternative then is an experiment with for all covariates/factors just one observation per level.

Latin hypercube sampling for observational studies is coined as conditioned Latin hypercube sampling (cLHS). For each covariate a series of intervals (marginal strata) is defined. The breaks in the marginal strata are chosen such that the numbers of grid cells in these marginal strata are equal. This can be done using the quantiles corresponding to evenly spaced cumulative probabilities as stratum breaks. For instance, for five marginal strata we use the quantiles corresponding to the cumulative probabilities 0.2, 0.4, 0.6, and 0.8.

The objective function that is minimized is the weighted sum of three components. The first component is

$$O_1 = \sum_{p=1}^P \sum_{h=1}^H |n_{hp} - 1|, \quad (22)$$

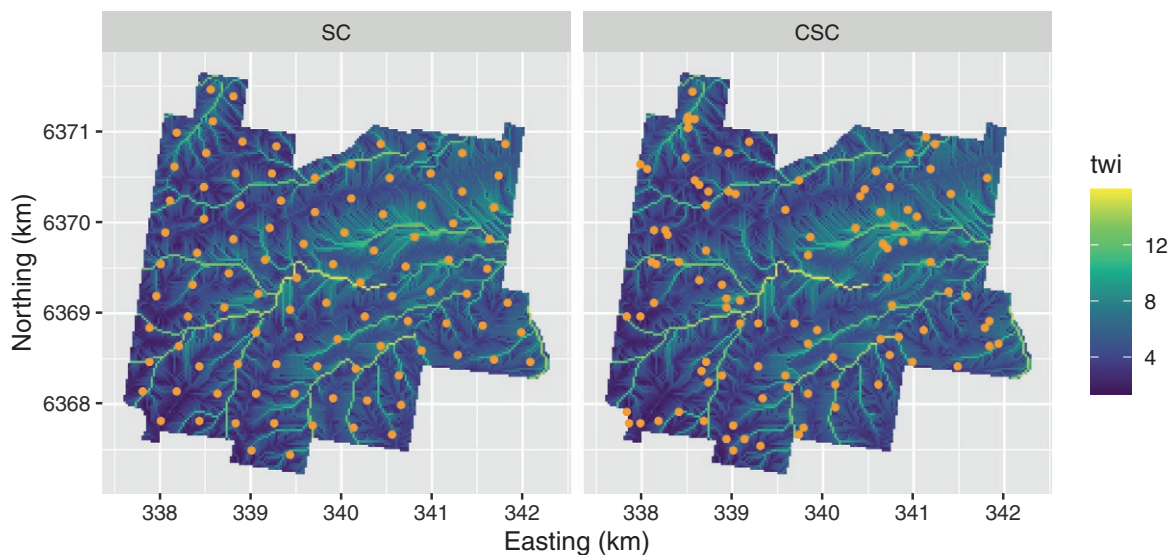


Fig. 9 Spatial coverage sample (SC) and covariate space coverage (CSC) sample of 100 points from the Hunter Valley, NSW, Australia. In CSC topographic wetness index, normalized difference vegetation index and elevation are used as clustering variables.

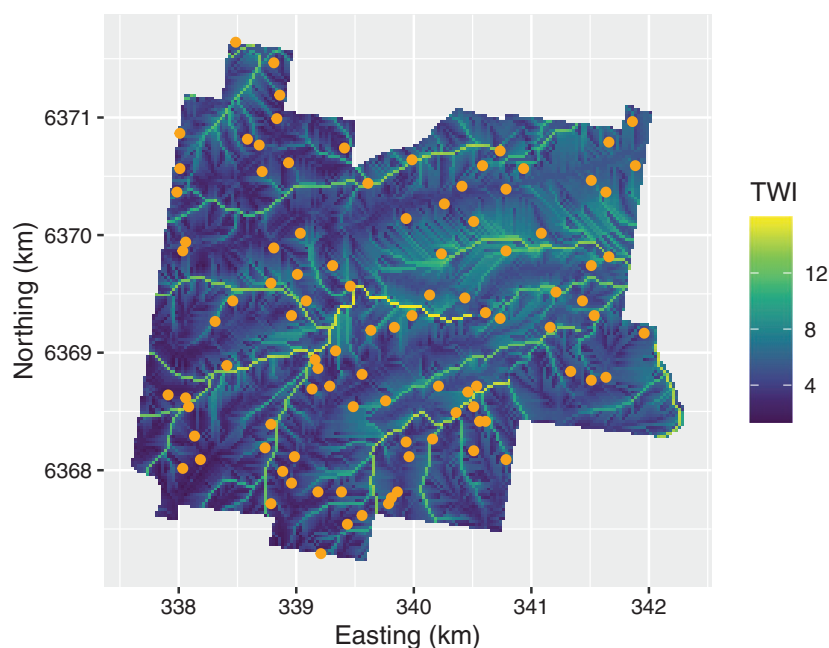


Fig. 10 Conditioned Latin hypercube sample of 100 points from the Hunter Valley, NSW, Australia, using TWI, NDVI and elevation as covariates.

with P the number of covariates, H the number of marginal strata, n_{hp} the number of sampling units in marginal stratum h of covariate p , and $|\cdot|$ the absolute value operator. A second component quantifies how much the sample correlation matrix differs from the population correlation matrix. A third criterion is involved only when we have in addition to quantitative covariates one or more categorical ancillary variables. It quantifies how much the sample sizes in the categories deviate from proportional sample sizes.

With cLHS the marginal distributions of the covariates in the sample are close to these distributions in the population. This can be advantageous for mapping methods that do not rely on linear relations, for example machine learning techniques such as classification and regression trees (CART) and random forests. Fig. 10 shows a conditioned Latin hypercube sample of 100 points from the Hunter Valley (NSW, Australia), using the same three covariates as before in covariate space sampling. Of the 300 marginal strata there are 12 with no points, and as many strata with two points.

Model-based sampling designs

In model-based sampling a model of the spatial variation of the study variable (geostatistical model) is used to design a sample. Samples are obtained by minimizing a criterion that is defined in terms of variances of prediction errors. An example is the mean kriging variance criterion, i.e. the average of the kriging variances over all nodes of a prediction grid.

Model-based sampling therefore requires prior knowledge of the model of spatial variation. Such a model must be specified and justified. Once this model is given the sample can be optimized.

Optimization of grid spacing

Given a prior geostatistical model, we can use it to optimize the spacing of a regular grid. The grid spacing is derived from a required level of accuracy of predictions for mapping. I assume that the map is constructed by kriging. A kriged prediction of the variable at an unobserved location is accompanied by a variance of the prediction error, referred to as the kriging variance. The required accuracy is a population parameter of this kriging variance, e.g. the population mean of the kriging variance or some quantile of the cumulative distribution function of the kriging variance.

Fig. 11 shows a semivariogram of the soil organic matter (SOM) concentration (in dag kg^{-1}) in the western part of the Amhara Region in Ethiopia. This semivariogram is estimated from a large convenience sample along roads.

The semivariogram is used to design a new sample, a square grid, for mapping SOM across West-Amhara. The legacy sample (convenience sample) is not used in mapping, just for estimating the semivariogram. The kriging variance is independent of the observations at the sampling points, therefore, it is not a problem that at the stage of designing a sample we do not yet have observations of the study variable. The semivariogram is used to predict a dummy variable, 'observed' at the nodes of a square grid with a given spacing, at a simple random sample of 1000 points. The median (P50), 80th percentile (P80) and 95th percentile (P95) of the 1000 kriging variances is computed. This is repeated for a series of grid spacings. Fig. 12 shows the result; it shows that the tolerable grid spacing for a median kriging variance of 0.85 is about 11.7 km. For P80 not to exceed 0.85 the spacing is reduced to 9.4 km, and for the P95 this is further reduced to 7.1 km.

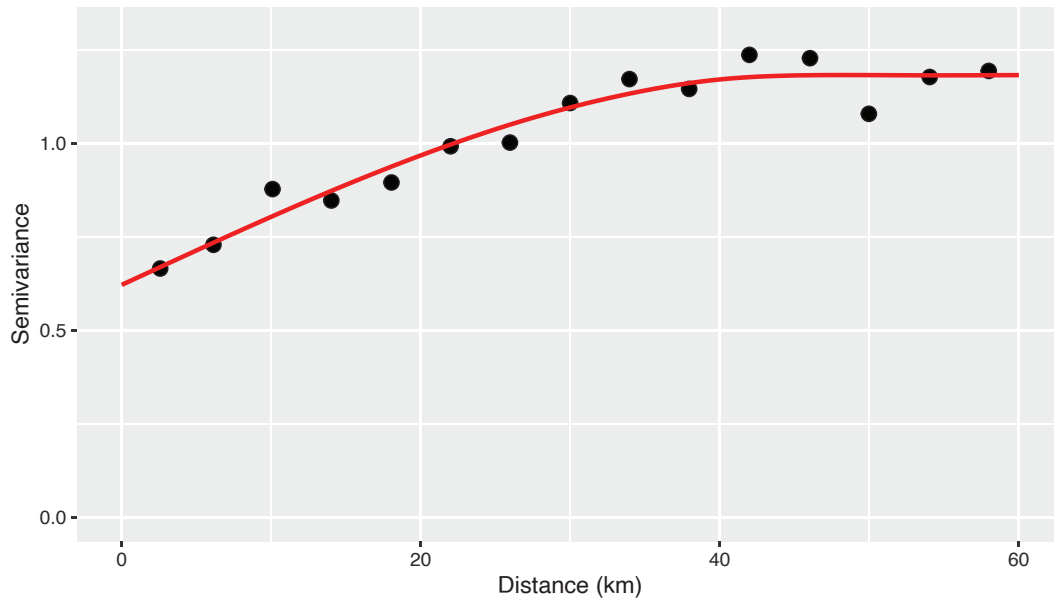


Fig. 11 Sample semivariogram and fitted spherical model of the SOM concentration, estimated from legacy data in West-Amahara of Ethiopia.

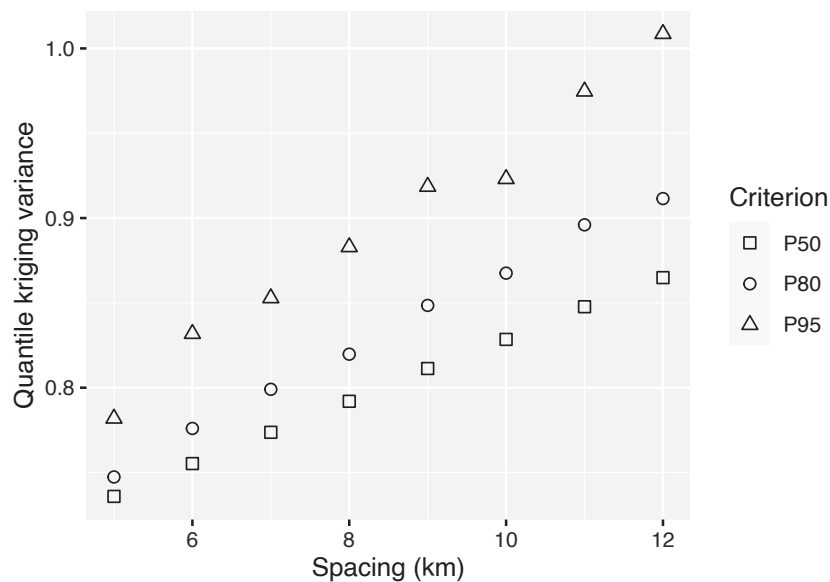


Fig. 12 Sample semivariogram and fitted spherical model of the SOM concentration, estimated from legacy data in West-Amahara of Ethiopia.

The same procedure can be used to optimize the grid spacing for kriging with external drift (KED), i.e. kriging with a geostatistical model in which the mean is a linear function of one or more covariates.

Optimization of sampling pattern

Once we have determined the tolerable grid spacing, s , we can compute the number of grid points by A/s^2 . But why should we sample on a regular grid? An irregular pattern of a sample of the same size may yield a more accurate map. A first option is to compute a spatial coverage sample, using k -means to optimize the sampling pattern as shown before. The alternative is to optimize the sampling pattern using, as before, the mean or some quantile of the kriging variance as a minimization criterion. For optimization we need some random search algorithm, for example, simulated annealing. Fig. 13 shows a model-based sample and a spatial coverage sample for an 80 ha site at the Cotton Research Farm in Uzbekistan to map soil salinity. The sampling patterns are quite similar.

Quick and dirty measurements of soil salinity can be obtained by an electromagnetic induction (EM) device. The EM measurements on transects covering the Cotton Research Farm were interpolated on a fine grid. The aim is now to design a sample at which the apparent electrical conductivity (EC) of the soil, a more accurate determination of soil salinity, will be measured. These

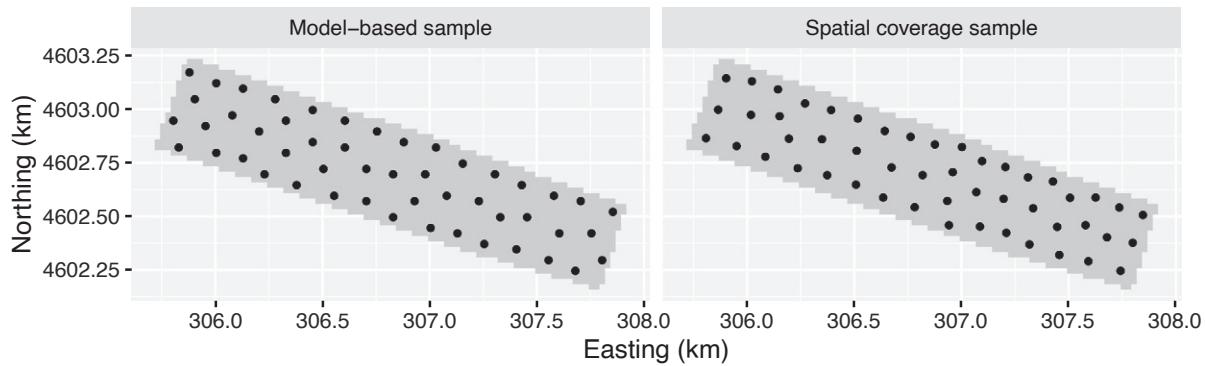


Fig. 13 Sampling pattern optimized for the mean variance of ordinary kriged predictions of $\ln \text{ECe}$ (Model-based sample) and spatial coverage sample (SCS) from the Cotton Research Farm.

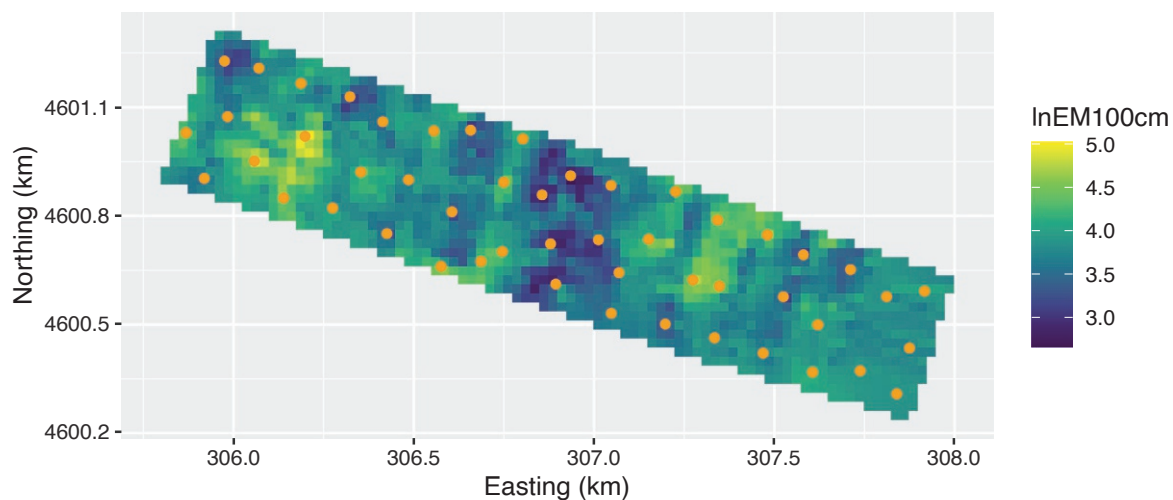


Fig. 14 Optimized sampling pattern for kriging with external drift of $\ln \text{ECe}$ at the Cotton Research Farm, using $\ln \text{EM}$ as a covariate.

EC measurements can then be used together with the interpolated EM measurements to map soil salinity by KED. Fig. 14 shows the optimized sampling pattern for kriging with external drift (KED) of natural logarithms of soil salinity with the natural logarithms of EM as a covariate. The sampling pattern is less regular. Most striking is the preferential sampling of locations with either very small or very large values of $\ln \text{EM}$. The variance of the prediction error with KED can be decomposed into a component quantifying uncertainty about the mean and a component quantifying uncertainty about the interpolated residuals. The former variance component is reduced by selecting locations with extreme values of the covariate, the second variance component is reduced by spreading the locations throughout the farm. The optimized sampling pattern automatically balances both variance components.

Sampling for semivariogram estimation and mapping

There is extensive literature on sampling for estimating the semivariogram, see references in Brus (2022). In practice, we often cannot afford a reconnaissance sample for estimating the semivariogram, followed by a survey for mapping. It is more efficient to design a single sample that is used for both estimating the semivariogram and for kriging. Quite a few minimization criteria are proposed in the literature for designing such samples. All these methods require a prior geostatistical model and various assumptions must be made. A practical solution is to select a spatial coverage sample, and to supplement this sample with a sample of points at a short distance from the spatial coverage sample points. Fig. 15 shows a spatial coverage sample of 90 points supplemented by 10 points at 20 m from a randomly selected spatial coverage sample point, in a random direction.

Future research

The most important decision in soil sampling design is the choice between probability and non-probability sampling. It determines the types of statistical inference that are feasible. With probability sampling all three types of inference, design-based, model-assisted, and model-based, are feasible, whereas with non-probability sampling model-based inference is the only option.

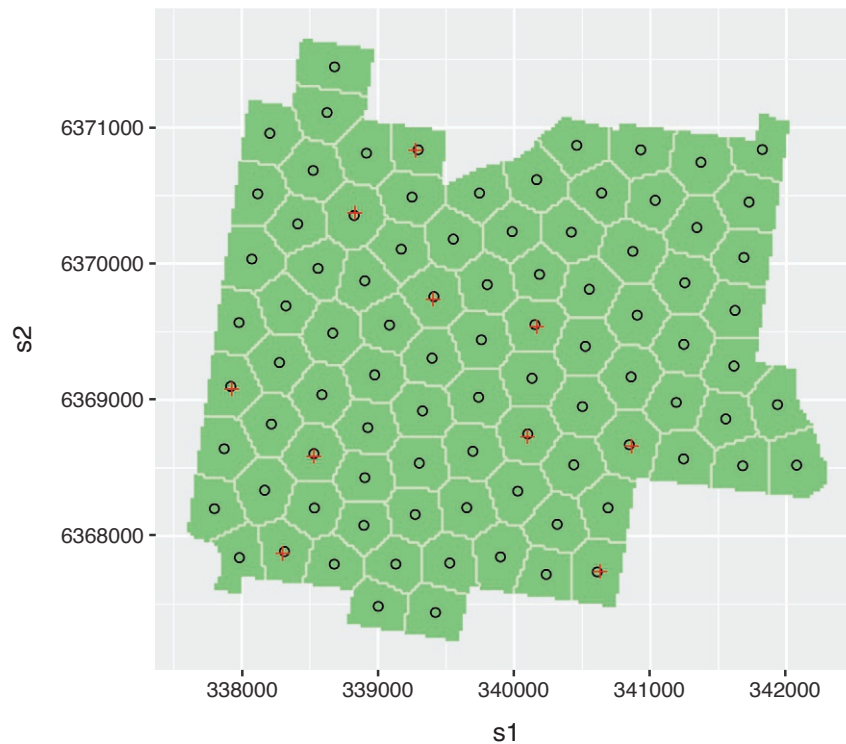


Fig. 15 Spatial coverage sample of 90 points supplemented by 10 points at short distance (20 m) from randomly selected spatial coverage point.

For estimating (sub)population parameters probability sampling is recommended, whenever feasible. Extensive literature is available on how the accuracy of estimated (sub)population parameters can be increased by using a linear or non-linear model, relating the study variable to ancillary variables, in model-assisted estimation. However, publications on how spatial dependency of the study variable or the residuals of a linear or non-linear model can be exploited in model-assisted estimation are still scarce. I have seen papers in which a spline model is used to model spatial dependency (see references in Brus (2021)), but my experience is that this strongly underestimates the variance of the estimated population mean.

For mapping through prediction at points model-based inference is the natural choice. Also for this aim I recommend probability sampling whenever feasible, using a sampling design with equal inclusion probability for all units. Equal inclusion probabilities are important because in model-based inference these probabilities are not accounted for. When units are selected with unequal inclusion probabilities this may yield strongly biased predictions.

More research is needed into efficient probability or non-probability sampling designs for mapping with machine learning techniques such as random forest.

References

- Brus DJ (2021) Statistical approaches for spatial sample survey: Persistent misconceptions and new developments. *European Journal of Soil Science* 72(2): 686–703.
- Brus DJ (2022) *Spatial Sampling with R*. New York: CRC Press.
- Cochran WG (1977) *Sampling Techniques*. New York: Wiley.
- de Gruijter JJ, Brus DJ, Bierkens MFP, and Kotters M (2006) *Sampling for Natural Resource Monitoring*. Berlin: Springer.
- de Gruijter JJ, Minasny B, and McBratney AB (2015) Optimizing stratification and allocation for design-based estimation of spatial means using predictions with error. *Journal of Survey Statistics and Methodology* 3(1): 19–42.
- Grafström A and Tillé Y (2013) Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics* 24(2): 120–131.
- Lohr SL (1999) *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.
- Särndal CE, Swensson B, and Wretman J (1992) *Model Assisted Survey Sampling*. New York: Springer.

Handling of soil samples in the field, in transport to the laboratory and in the laboratory

Claudy Jolivet, Céline Ratié, Déborah Chavrit, Nicolas Soler-Dominguez, and Antonio Bispo, INRAE, Info&Sols, Orléans, France

© 2023 Elsevier Ltd. All rights reserved.

Introduction	552
In the field	552
Optimal sampling conditions	552
Climatic conditions and soil humidity	552
Cultural constraints	552
Sampling zones	553
Avoidance of artificial contamination of samples	553
Sampling procedures	553
Collection of organo–mineral or mineral samples by augering	553
Collecting samples of humus horizons	554
Collecting samples of horizons from the soil profile of a soil pit	554
Collecting volumetric samples for bulk density determination from the soil profile	554
Collecting clod samples for the determination of available water content	558
Collecting samples to determine the coarse fraction contents of a soil horizon	560
Sampling report, field notes	560
Conditioning of samples	560
From the field to the laboratory	561
In the laboratory	561
Reception and registration of the samples	561
Soil preparation for chemical analysis	562
Avoiding any contamination and ensuring traceability of the sample	562
Weighing	562
Drying	563
Laboratory sample preparation	563
Sample sub-setting—Quartering	563
Preparation of samples from organic horizons	564
Sending for analyses and archiving	564
Sending laboratory samples for analyses	564
Archiving	566
Conclusion	566
References	567

Abstract

Based on existing standards, this chapter describes validated and recommended methods for the handling of soil samples (i) in the field (optimal sampling conditions, choice of a sampling zone, usual sampling procedures for collecting and conditioning composite samples of humus and organo–mineral, samples from the horizons of a soil pit, samples to determine bulk density, water holding capacity and coarse fragments content), (ii) from the field to the laboratory (recommendations for the transport of soil samples) and (iii) in the laboratory (reception and registration of soil samples, soil preparation for chemical analysis, sending for analysis and archiving).

Key points

- Optimal soil sampling conditions and choice of a sampling zone
- Usual methods for collecting and handling of soil samples in the field
- Recommendations for the transport of soil samples from the field to the laboratory
- Procedures for reception, registration, preparation, sending for analysis and archiving of soil samples in the laboratory

Introduction

Soil properties are highly variable in space and time. Consequently, studying soil characteristics or evolution through analytical measurements requires the use of rigorous sampling and sample handling procedures, in order to have reliable results with low uncertainty. In particular for large inventories entailing numerous field operators or for monitoring purposes featuring repeated sampling over decades, it is crucial to apply routine protocols for reliability, replicability and avoidance of bias throughout the sampling – preparation – analysis – archiving chain. Beyond these two examples of application, these recommendations are valuable overall for every soil study involving soil samples.

Here, we present general procedures to sample, handle and transport, prepare and archive different types of soil samples, according to the elements needed to be analyzed. The following sample protocols are used by the French Soil Quality Monitoring Network (“Réseau de mesures de la qualité des sols,” RMQS, Jolivet et al., 2022) and by the European Soil Samples Conservatory (Conservatoire européen des échantillons de sols, CEES, INRAE Orléans, France). All these protocols follow or are adapted from ISO the International Organization for Standardization or NF AFNOR French standards.

In the field

Optimal sampling conditions

The following instructions are adapted from [ISO 18400-102 \(2017\)](#). The quality of the observations and samples depends mainly on soil humidity and cultural constraints. For instance, when collecting several sets of soil samples by auger at different depths, it is essential to minimize the risk of mixing the layers collected to avoid any risk of cross-contamination between samples. Similarly, the quality of the bulk density measurements used to calculate element stocks in the soil is subject to the conditions under which volumetric samples are taken. Therefore, these samples must be taken in compliance with climatic (soil moisture content) and cultural (physical condition of the soil) constraints.

Climatic conditions and soil humidity

The consideration of climatic conditions takes priority over any other criterion in the choice of the sampling period, even if this interferes with organizational constraints. The ideal conditions for description and sampling are met when the soil moisture is close to field capacity (drained soil). It is thus advised not to sample when the soil is too dry or too wet. For long-term soil monitoring, it is important to keep as close as possible to similar sampling conditions to be able to compare samples from different sampling campaigns. Sampling during periods of drought or excessive moisture should be avoided, not only because of the physical difficulty of taking quality samples, but also for reasons of observational accuracy.

Dry conditions should be avoided because they make auger sampling difficult, especially in stony or sandy soils. The upper part of a dry soil sample is often truncated, which can have consequences for the validity of the contents measured. Similarly, the lower layer of the soil cannot be reached or is insufficiently sampled. Finally, the risk of cross-contamination of samples between two sampling layers increases greatly when the soil is too dry, as surface soil can easily fall into the hole.

Soil that is too dry makes it difficult nor impossible to take the volumetric samples used for bulk density determination. This is particularly true for sandy soils, where samples are too fragile, and for clay soils, where compactness may bias the sampling. A soil that is too dry is also more difficult to describe, as the structure, colors (and therefore the detection of mottles, for example) are poorly expressed.

When the soils are too wet, in addition to the damage to vegetation and soil compaction that can be caused by trampling in excessively wet conditions, waterlogging considerably hinders the quality of sampling, especially that by augering. This is particularly true when the soil is clayey, as the samples are sticky and more difficult to extract from the auger.

In waterlogged soils, observations on the profile are more difficult (some structures are closed, for example) and it is more difficult to carry out volumetric sampling with precision, in particular in the deep layers of the soil, which may be saturated by the presence of a temporary water table.

Cultural constraints

In cultivated soils, taking samples from blown, recently plowed, stubble or deep plowed soil should be avoided to circumvent any bias related to agricultural work, not only because of the difficulty of taking quality samples, but also for reasons of observational accuracy. The ideal conditions for description and sampling are met when the tilled layer is homogenous in terms of structure and compaction, for instance after seedbed preparation or after a few weeks of natural settling. These precautions are paramount for the collection of surface composites and for the determination of soil bulk density in volumetric samples. The sampling dates must therefore consider the calendar of cultivation operations, obtained from the farmers during the survey and revalidated during contact with the farmer or manager before the operation.

Information on the sampling conditions, in particular the water status of the soil, weather conditions, vegetation cover, land use, must always be noted on the sampling report or field notes (see below).

Sampling zones

The following instructions are adapted from those in NF X31-100 (2020). The choice of sampling site depends on the objectives, however, in general a site should be located in a homogeneous area in terms of occupation or management practices. Some areas should be avoided as follows:

- agricultural pounds,
- woodland edges,
- areas that straddle plots that have been consolidated or are under different crops,
- the edges of roads or stony paths, old roads that have been reconfigured (observable in particular by the presence of a slight eminence or of allochthonous pebbles),
- the immediate banks of streams and ditches,
- coal-mining rounds or former fireplaces,
- ancient human settlements (signature from the surface by shards of flint, shards, bricks, various debris, charcoal and/or black color on an abnormal thickness),
- the immediate proximity of buildings,
- small localized bumps or depressions,
- illegal dumpsites, scrap metal dumpsites.

Avoidance of artificial contamination of samples

Before any sampling, it is necessary to ensure that the equipment used (augers, knives, sampling shovels, buckets, mixing containers, etc.) does not cause any interference (pollution or loss) with the soil properties being analyzed. Thus, according to the type of analysis, equipment made of different material should be used to avoid contamination of the samples by the material, tools and containers.

In general, plastics bags, plastic trays and stainless-steel tools are used, but caution should be exercised for samples taken for metal trace element measurements, DNA analyses and determination of organic compounds:

- For samples to determine metal trace elements, equipment made of copper, brass, galvanized or chromium-plated steel or covered with paint should be avoided. Plastic or raw steel tools should be given preference. Plastic buckets and tubs (polyethylene or polypropylene) that can be carefully washed after conditioning with HCl (hydrochloric acid) should be used for sampling.
- For DNA analyses the samples must not be handled with bare hands to avoid any human DNA contamination of the samples: single-use latex gloves must be used to handle the samples.
- For the analysis of organic compounds, plastic containers or plastic bags should be avoided (ISO 18400-105, 2017) as they may adsorb contaminants.

In addition, to avoid any risk of direct or cross-contamination between different samples all sampling tools (e.g. auger, knives) and containers (e.g. trays, buckets) must be washed thoroughly with tap water then rinsed with demineralized water and then carefully dried with paper towels. This must be done systematically between two samples.

Sampling procedures

Collection of organo-mineral or mineral samples by augering

Individual sample

If the protocol does not include sampling an undisturbed soil core with a soil sampling ring or a mechanical soil coring kit, the organo-mineral and mineral samples are collected using a helical soil auger (for example EDELMAN auger). A soil auger of 7 cm in diameter is generally used to provide enough material for later use. The handle of the auger should be graduated to check the sampling depth. This can be done by attaching a piece of colored tape every 10 cm along the handle.

Sampling starts at the surface of the soil in bare soil. In the presence of dense vegetation, e.g., under grassland, plants should be removed with a knife before augering to avoid the incorporation of vegetation debris into the sample.

Systematically remove the edges of the core and its 'tip' with a knife. This precaution is imperative for two reasons: (1) to avoid any risk of cross-contamination between horizons when the core sample is in contact with the augerhole wall and (2) to collect a constant volume of soil if the objective is to form an unbiased composite sample, by mixing the different element samples in equal parts. If the soil samples are for DNA analyses, they must be handled using single-use latex gloves (see above) or with a clean knife to level the auger and place the sample in the collection bucket.

If the samples needs to be taken in several stages to reach the defined depth, the upper part of each core (about 10 cm) should also be removed. A helical auger with a diameter of 7 cm allows the removal of only 20 cm of soil when starting from the soil surface. After that, this type of auger only allows an advance of 10 cm at each 'auger turn.' The first 10 cm of the upper part of the core, which corresponds to soil that falls to the bottom of the auger hole when removing and reintroducing the auger, must be removed at each 'auger turn.' The auger should not be filled to excess. It is better to carry out the sampling in several stages to avoid the risk of biasing the sampling depth or of getting the auger stuck in the hole. Each sample is bagged separately.

Composite sample

To provide a representative sample of a plot, for instance for agronomic analysis or monitoring purposes, several individual samples are collected according to a specific sampling design, then composited. The principle of producing a composite sample is to homogenize by mixing equal parts of the material from a series of samples for each soil layer taken from the sampling area. In this case, the procedure for collecting each individual sample for the composite is repeated as many times as needed.

For monitoring purposes, two buckets per operator are used to collect each composite sample: the sample material intended for the composite sample is placed in a first bucket, which was previously cleaned (see above). The sample waste is placed in a second bucket and then eliminated to avoid contamination of the sampling surface. To prevent cross-contamination between soil layers, the buckets used for sampling are carefully cleaned.

After the sampling, the auger holes should be resealed with surface earth to prevent injury to animals and fauna (cattle with broken legs, insect and small animals falling into the auger holes).

Depending on the monitoring methods of analysis and the archiving objectives, a minimum amount of sample might be necessary. In stony or shallow soils (< 15 cm, Fig. 1, case 1), it may be difficult to collect a sufficient amount of sample in one pass. In this case, two or more samples can be taken from each plot to ensure that enough material is collected to make up the surface composite sample. The use of a smaller diameter auger (4 cm) is possible if it makes sampling easier, but in that case, the samples should be multiplied to obtain a sufficient quantity of composite sample. Caution: in the event of an occasional blockage of the auger, the upper part of the sample during a second auger coring should be discarded to avoid over-sampling the upper part of the soil profile (Fig. 1, case 2).

The soil samples are then taken from the collection buckets and placed onto a tray that has been previously cleaned. The soil is carefully swabbed using single-use latex gloves, then mixed thoroughly to form a composite sample. Depending on the amount of sample required, the whole of each composite sample can be bagged entirely, quartered, or bagged separately.

Collecting samples of humus horizons

In forest or old field soils, humus horizons can be sampled before collecting organo–mineral horizons (Fig. 2). After the litter (Oi layer) is removed or collected, in addition a sharp cylinder is pushed vertically to a depth sufficient to pass through all the O horizons (Oe + Oa layers), using a mallet if necessary. If the organic layer is thicker than the cylinder height, sampling can be done in two stages.

The cylinder is then carefully removed from the soil surface with a knife. The organic layer cut by the cylinder usually remains inside it. If not, the cylinder is removed and the previously cut organic layer is lifted out of the hole made by the cylinder. The mineral or organo–mineral fragments are then removed from the underlying horizon as needed. The Oe and Oa layers can be separated or kept as a whole. The average thickness of the whole sample or each horizon is measured and recorded in the sampling document if necessary.

The sample is placed in a plastic bag for individual analysis or in a bucket for compositing. In this case, the samples are gathered in a previously cleaned container (see above). Each sample is coarsely disaggregated and mixed carefully to form the composite. Single-use latex gloves should be used to take the samples and make up the composite sample to avoid any human DNA input.

Collecting samples of horizons from the soil profile of a soil pit

Before sampling, the pit wall should be cleaned thoroughly by working from top to bottom using a scraper (Berthelet trowel) or a knife, for example. Then, with a small shovel or a knife, the samples should be taken from the bottom upwards to avoid the step of cleaning the pit wall in between sampling. If possible, sample within each horizon (organic, organo–mineral and mineral), across the entire width of the profile wall to obtain a representative sample of the horizon.

Collecting volumetric samples for bulk density determination from the soil profile

Volumetric samples should be taken under moist soil conditions close to field capacity and in stabilized soil, i.e., not recently worked (plowed, stubble, subsoiled) (see above).

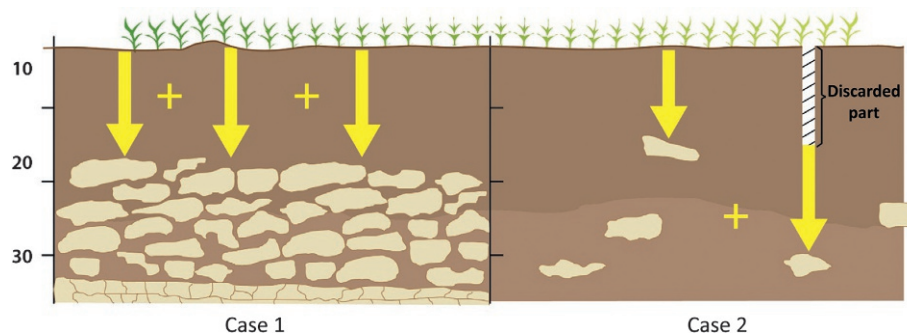


Fig. 1 Sampling of composite samples in stony and shallow soils (Jolivet et al., 2022).



Fig. 2 The different steps to take a sample of an holorganic horizon (Jolivet et al. 2022).

Three methods can be used: cylinder, water, and sand. Depending on the constraints of the soil (coarse particle load, root density, slope), it is up to the sampler to choose the most suitable method. The method used can be different between two soil layers or horizons, but only one method should be used within a soil layer or horizon.

The cylinder method (core method)

The cylinder method is suitable for soils with little or no gravel, stones or roots. It is described precisely in the standard NF EN ISO 11272 (ISO 11272, 2017). The steps of the cylinder method are described in Fig. 3.

Note: vertical sinking of the cylinder is strongly recommended. It is mandatory in surface horizons, but horizontal sinking of cylinders is also possible at depth.

The water method

This is an excavation method adapted to gravelly or rocky soils, or in the presence of a high root density. It corresponds to an adaptation of the method described in standards NF X 31-503 (NF X31-503, 1992) and NF EN ISO 11272 (ISO 11272, 2017). The methodological modifications concern the principle, the material and the field procedure. Everything related to the definitions, the expression of the results and the laboratory procedures (material and operating mode) is in accordance with the standard mentioned.

In the field, after the soil has been levelled, the procedure consists of creating a cavity, collecting all the soil to measure its mass and water content in the laboratory, placing a watertight plastic bag in the cavity and then filling the bag with water to determine the volume of the cavity. The steps of the water method are described in Fig. 4.

The main difficulty with this method is to obtain a surface that is sufficiently flat to allow the water level to be adjusted precisely to the ground surface. Therefore, it is not suitable for small volumes. On the other hand, it should be used when the cylinder method is difficult to use i.e., in the presence of gravel or stones, in the presence of roots, or in very clayey horizons. The stonier is the soil and the larger the pebbles (stones or boulders), the larger the volume of the cavity should be to improve the accuracy of the measurement: a volume close to 1500 cm³ is recommended.

Tip

A second sampling and measurement can be conducted using the first cavity made. The cavity can be deepened again, and the extracted soil collected, and the volume can be measured over the entire cavity. The volume of the second sampling will be deducted

1. The measurement can be carried out on the surface of the soil or in the deep horizons, working on the wall of a pit cut into a step. Prepare a horizontal surface using a trowel or scraper.		
2. Push in the cylinder vertically. Use a sharp cylinder and, if necessary, one or two mallets to drive it in.		
3. Gently remove the cylinder with a knife and a spatula or trowel.		
4. Gently flatten the two ends of the cylinder with a knife, using a plate on which to place the sample.		
5. Empty the cylinder into the sampling bag.		
6. Measure the depths of the sample (top and bottom) and record them on the sample bag and on the sampling report.		

Fig. 3 Steps of the cylinder method (Jolivet et al., 2022).

1. The measurement can be carried out on the surface of the soil or in the deep horizons, working on the wall of a pit cut into a step. Prepare a horizontal surface using a trowel or scraper. It must be perfectly flat (checked with the level), using the knife and the scraper, before digging the cavity.	
2. Place the template (metal or plastic) on the floor and secure it with nails. Check the flatness with the level.	
3. Dig a cavity, collecting all the soil extracted in the sampling bag, without losses (use the brush to collect any losses that have fallen onto the template and any small aggregates remaining at the bottom of the cavity). Depending on the condition of the soil, dig with a knife, spoon or ladle. The diameter of the cavity should correspond to the template and the depth of the cavity should be adjusted according to the thickness of the horizon. Be sure to take a volume of between 1,000 and 1,500 cm³. The shape of the cavity is not important, but its edges should be as regular as possible: avoid indentations and outcrops of sharp or pointed coarse elements that could puncture the bag used to measure the cavity volume.	
4. Place the plastic bag in the cavity and press it against the walls. Fill the bag with water using the measuring cylinder until the surface level is reached. Ensure that the plastic bag adheres well to the walls of the cavity by correcting the adhesion with the spoon as you fill it. The volume of soil removed corresponds to the total volume of water poured into the cavity. Record the measured volume on the sampling bag and on the sampling report.	
5. Measure the depths of each sample (top and bottom) and record them on the sampling bag and on the sampling report.	

Fig. 4 Steps of the water method (excavation method) (Jolivet et al., 2022).

4. Place the plastic bag in the cavity and press it against the walls. Fill the cavity with calibrated sand using the measuring cylinder until the surface level is reached, pouring in the sand from a height of 5 cm to ensure an even distribution. Ensure that the plastic bag adheres well to the walls of the cavity by adjusting the adhesion as you fill it. The volume of soil collected is the total volume of sand poured into the cavity as measured with the measuring cylinder. Record the measured volume on the sampling bag and on the sampling report.



Fig. 5 Measuring the volume of the cavity with the calibrated sand (step 4 of the water method) (Jolivet et al., 2022).

from the total volume measured, minus the volume of the overlying sample measured. This method does not require a flat surface. It is only feasible for two successive repetitions, due to the difficulty of extracting soil beyond 20-cm depth and a 3 L bag of water.

The sand method

This method of excavation is suitable for gravelly or stony soils, or where there is a high density of roots. The sand method is described in the standards NF X31-503 (Measuring of bulk density - Sand method) (NF X31-503, 1992) and NF EN ISO 11272 (Determination of dry bulk density) (ISO 11272, 2017). It differs from the water method in that calibrated sand is used instead of water to measure the volume of the sampling cavity (step 4). All other steps of the method (steps 1, 2, 3 and 5) are identical. Its main advantage over the water method is that it does not require a perfectly flat surface to measure accurately the volume of the cavity. This is because the sand can be adjusted to the surface of a cavity where the edges are not level. To a certain degree, corresponding to the flow limit of the sand, the measurement can be carried out on sloping ground. The steps of the sand method are described on Fig. 4 (steps 1, 2 and 3) and on Fig. 5 (step 4).

The calibrated sand can be reused without time limit. Make sure it is completely dry before each new use. Dry it at 105 °C before the start of each year's campaign.

The clod method (special cases)

Measurement of the density of certain hardened horizons (iron pan, iron crust, etc.) requires the collection of small fragments of such horizons of a few cm³ and the use of the clod method (NF EN ISO 11272). Collect and condition the samples according to the method described in the following section (See above).

Collecting clod samples for the determination of available water content

This procedure describes how to take volumes of soil from an horizon while maintaining the soil's natural macroscopic structure to estimate the available water content of a profile horizon.

Sampling conditions

The ideal soil moisture conditions for sampling correspond to those of field capacity (See above). In the field, the structural state observed largely depends on the moisture content of the horizon. Thus, an horizon rich in swelling clays, which presents a massive aspect in a wet state, is likely to shrink and divide strongly in a dry period. Therefore, structure is a property that expresses itself more, or less, and also differently over time, according to seasonal cycles or according to meteorological events.

In addition to the indications given above, sampling in conditions that are too dry would not comply with the measurement methodology, which consists of extracting water from the soil at different potentials. Conversely, sampling under conditions that are too wet, apart from the fact that it would not allow the release of ideally expressed structural elements, would lead to settlement of the collected samples during their transport.

Operating procedure

Samples are taken from the top to the bottom of the profile. The steps of the clod method for the determination of available water content are described in Fig. 6.

When packing volumes of soil into the box, manually remove as much as possible from the areas smoothed by the tines of the digging fork or by the blade of the knife.

1. Remove a layer of soil of about 2 cm on the upper limit of the horizon with a knife.	
2. Use a toothed scraper to obtain a horizontal plane.	
3. Push the blade of a knife onto the lower limit of the horizon. 4. Crumble some soil from the horizon to be sampled into the bottom of the sample box (this crumbled soil layer will somewhat dampen the shaking when the samples are transported in the box).	
5. Carry out a superficial scraping or picking of the vertical surface of the horizon to be sampled to remove any particles that have fallen from the overlying horizon. This action also enables the identification of any rodent galleries or other small burrowing animals, as well as the presence of pebbles, rocks, buried residues, etc.	
6. Position the digging fork vertically, set back 20 cm from the wall of the profile, and drive it in ± 15 cm with a rubber mallet. Gently tilt the digging fork towards you while holding the volume(s) taken.	
7. Place volumes of soil of 10 cm size into the box, if the structure allows it.	
8. The box will then be filled up with smaller volumes (bearing in mind that the smallest size on which measurements are made in the laboratory is: ± 3 cm). Note: add only soil from the horizon and do not add paper, moss or any other material that can spoil or absorb soil moisture to fill up the box, as this may compromise the validity of the usable reservoir measurement.	

Fig. 6 Steps of the clod method for the determination of available water content (Jolivet et al., 2022).

Specific cases

The collection of coherent clods may be difficult to achieve in the following cases:

- Horizons with a structure that does not offer sufficient cohesion to collect clods (e.g., particulate structure).
- Horizons with a high proportion of gravel, pebbles and or stones.
- Horizons with high root density.
- Significant presence of plant debris.

In this case, take the horizon as it appears by simply removing stones and large roots. Fill a sample box as completely as possible. Another method (fine soil protocol) will be used in the laboratory to measure the water content at different water potentials.

Collecting samples to determine the coarse fraction contents of a soil horizon

The coarse content (stones, pebbles, gravels) is a key variable in the assessment of several soil characteristics such as organic carbon stocks or available water content. Because of the considerable spatial variability of the coarse fraction of the soil, measuring stoniness with sufficiently low uncertainty requires the collection of a large amount of soil, especially in the presence of pebbles and stones.

Sampling will be facilitated if soil moisture conditions are close to those of field capacity. Sampling is generally made from a soil pit. The pit should be wide and deep enough to collect a sufficient quantity of samples to be representative of each horizon in the profile. Because of the considerable soil disturbance associated with taking these samples, they should be taken last among the different kinds of sampling of the profile.

Sampling is started from the surface. On the first horizon, a spade or digging fork is used to mark out the area to be sampled. Shovelfuls of material from the horizon are then poured into the rubble bag. After weighing the sample on the scale, the mass is noted. Between two samples, avoid mixing horizons by neatly clearing the underlying horizon. After the surface sampling, a roughly flat area should be marked out again from the upper part of the underlying horizon. Then, the procedure is repeated as for the previous horizon.

It has to be noted that the sample should not be taken during the digging of the pit. It should be carried out as described above according to the horizons described, and without sorting or discarding coarse particles to avoid bias in the determination of gravel, pebbles, stones and blocks content. Woven rubble bags can double up on the sampling bags because they are more durable.

Sampling report, field notes

The sampling report or field notes are a critical document that links the samples to the laboratory and beyond. The set of field data and their presentation in a readable and understandable form in the sampling report is an essential prerequisite for understanding the laboratory results and their presentation and interpretation in general (ISO 18400-107, 2017). The report is completed at the end of the sampling and should contain information about the sampling conditions and any deviations from the sampling plan. These aspects should include effects on the samples and on the samplers, on the observations and field measurements, any instructions for testing and analysis, comments on sources of uncertainty, comments on accuracy, fidelity and variability. More explicitly, the report should include:

- sampler names, date and purpose of sampling,
- general information about the site and its environment: weather, land use and current practices, stage of cultivation, tillage, surface condition, indications of any soil disturbance, type of soil, how to access the site, coordinates of the site. . .
- the list and details of the samples collected: origin (surface or profile), type of samples, (individual, composite, horizon), sampling depth, sampling method, number of bags or containers. . .
- any relevant comment (e.g. difficulties when sampling).

For profile samples, the sampling report may be supplemented by the coordinates of the pit, a drawing of the profile description and sample locations for a better understanding.

Conditioning of samples

Soil samples may be subject to changes between sampling and testing as a result of physical processes (for instance variation in moisture content or change in structure for samples featuring thixotropic behavior), and chemical (for instance evaporation or precipitation of volatile substances, reactions with the sample container) and biological changes (for instance organic compounds degradation due to biological activity). The extent of these changes depends on the characteristics of the sample, its temperature, its exposure to light, the nature of the container in which it is placed and the nature of the compounds to be analyzed, the time between sampling and analysis, the conditions to which it is subjected and the seasonal conditions. The significance of these changes and the prescription of the packaging process, storage, transport and delivery should be considered so that samples are still representative when delivered to the laboratory (ISO 18400-105, 2017).

Soil samples are usually packaged in collection bags, annotated with their content information. All this information is repeated on a label attached to the rubber band used to tie the bag. The clod samples are packed in plastic boxes (more details below). The samples for the coarse fraction measurements are packaged in large specific bags.

After collection and before dispatch, the samples should be stored at a temperature close to that of the sampling, i.e., a sample taken in winter should not be stored near a radiator in the office and conversely, a sample taken in summer under high temperatures should not be stored in a refrigerator or left in a vehicle in direct sunlight but in a temperate place. A rapid dispatch of samples is advised. They should not be stored in the refrigerator, except for the clods. The plastic boxes containing the clod samples are best kept in a cold room at 4 °C, or, failing that, in a cool room (to limit biological activity and condensation) and in the packaging intended for their dispatch.

To keep the integrity of the sample during transport the following tips are useful during bagging and packing of the samples:

- The air from the sample bags should be emptied, to prevent them from bursting.
- The sample bag should not be overloaded, for example, it should be filled at a maximum of 2/3 of its capacity.
- The sample bag should be put upright in the cardboard boxes.
- It is important to check that the rubber bands are in place and properly tightened.
- The gaps in the cardboard boxes should be stuffed, for example with newspaper to prevent bags from moving in the cardboard boxes.
- The plastic boxes used to pack the clods should be checked to ensure they are closed, and then taped up if necessary.

For clod samples, there are several precautions to take for their packing:

- Make sure that the lid is correctly positioned on the plastic box and that the box is tightly closed all around the box and lid contact.
- Mark on the wide adhesive tapes, affixed on the side and lid of the box, the references of the horizon regarding the following three items: Name of the site/Upper and lower depth of the horizon/Date of sampling.
- Place the plastic box in a specific cardboard box with foam inserts to prevent the plastic box from breaking.

To dispatch the samples use robust packages dedicated to heavy loads, not too large; for example double wall cardboard boxes, reinforced with tape if needed. They should be filled to a maximum of 20 kg each to facilitate the handling and respect the heavy goods handling regulations. Empty spaces in the packages should be avoided, or stuffed. Careful handling of clod and or core samples during transport is essential if samples feature thixotropic behavior (thixotropic clays or soil samples with andic properties). In this case, shaking or shearing force should be avoided.

From the field to the laboratory

The instructions below are adapted from the ISO 18400-105 (2017).

After the field, the samples should be delivered to the laboratories as soon as possible. This could be done by the field team itself or the sample could be sent by a courier. In this case, a delivery service in less than 24 h is highly recommended.

If samples are intended for routine physicochemical analysis, transport at room temperature is sufficient. For measuring organic compounds, refrigerated transport at 4 ± 2 °C is advised (ISO 18400-105, 2017) to avoid microbial activity after sampling that could alter the content of the substances to be measured.

After reception by the laboratory, the samples should be stored in a cold room (4 ± 2 °C) to minimize sample perturbations if the measurements require fresh soil (for instance for biological analysis like microbial activity measurement or DNA extraction, or for chemical analysis on specific soils like Andosols) or before drying and preparation for routine analysis, as described in the next section.

In the laboratory

The work in the laboratory includes different steps according to the sample characteristics. It always starts with the reception and registration of the samples. Then, it may include the steps of weighing, drying, sub-setting and preparation to provide a laboratory sample designed for soil analyses and lastly, archiving if required. At each step, quality assurance actions are performed to avoid any cross contamination or bias. Moreover, the analytical laboratory keeps residual samples for some time after the analysis to check the result and redo the analysis if a result is uncertain.

Reception and registration of the samples

On reception, all the samples must be checked for physical integrity, identified according to the specifications and registered in a database. In case of an uncertainty or the detection of an anomaly in a sample, it should be identified as non-compliant and put on hold until a decision is made (kept or discarded).

Each sample is assigned with a unique number to allow its identification and origin, e.g. the geographic coordinates, the date of sampling, the samplers' usernames. Each action made on a sample is also registered in order to trace the sample's life cycle. A barcode linked to the number can be used to identify a sample in the laboratory, in the archive, or at other laboratories where it is being analyzed.

Soil preparation for chemical analysis

Soil preparation is a crucial step before laboratory analyses. This step encompasses several actions that should be registered in a database. Subdivision or creation of a sub-sample should also be registered through affiliation. The samples coming from the field are the 'parents' and the created sub-samples are the 'children.' This is useful to know if an analyzed sample has been sent to the analytical laboratory in a 'raw' state (coming directly from the field) or if it has been prepared or handled beforehand in the laboratory by, for example, sub-setting (with a sample splitter such as a quartering tray).

Avoiding any contamination and ensuring traceability of the sample

Depending on the properties to be analyzed, we refer to the following standards for handling the samples:

- For metal trace elements measurements: [X31-150 \(1984\)](#)
- For physicochemical measurements: [ISO 11464 \(2006\)](#)
- For organic compounds measurements: [ISO 14507 \(2003\)](#)

In the laboratory, it is essential to follow some procedures to avoid any contamination of the samples. Tools, containers and laboratory benches are carefully cleaned between each sample preparation.

In general, cleaning materials and products are non-abrasive sponges, clean tap water, demineralized water, wipes, ultrasound baths, compressed air guns. Buckets and sample bottles are blown with compressed air. Drying and quartering trays, tweezers, foil trays, beakers and sieve receiving pans are cleaned with a non-abrasive sponge and clean tap water. They are then rinsed with demineralized water, dried, and blown with compressed air. Sieves, mortars, pestles and spoons are, in addition, put in an ultrasound bath between cleaning with clean tap water and rinsing with demineralized water.

The laboratory benches are cleaned with clean tap water and non-abrasive sponges and then wiped. The laboratory and its equipment are also cleaned once a week: vacuum-cleaning and floor mopping, cleaning of the laboratory equipment, e.g. ultrasound baths, weighing balances. When handling samples, the operator must always wear gloves to avoid artificial contamination of the samples.

To ensure traceability of the samples, all the procedures they undergo are registered in a database in which each sample is linked to a soil profile or sampling site (see above).

Weighing

Samples are weighed ([Fig. 7](#)) for several purposes. As a control procedure to know how much sample is available, and to be able to detect any loss during the laboratory process. Weighing is also used to check if soil moisture was close to field capacity during the sampling by weighing the samples before and after air drying. Samples can also be weighed according to the procedure required for future analysis, for example if they are aimed at bulk density measurements.

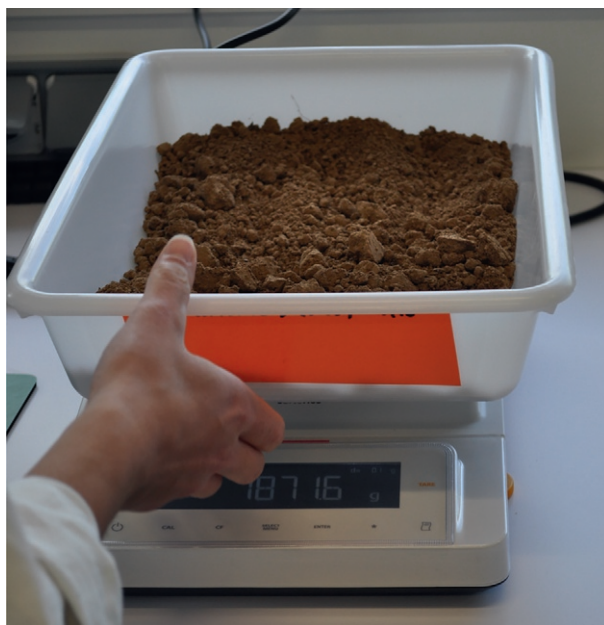


Fig. 7 Weighing of a soil sample.



Fig. 8 Soil samples drying at the European Soil Samples Conservatory.

Drying

During drying of the samples, care must be taken to keep the integrity of the sample and avoid any contamination or physico-chemical changes. The samples, stored in the cold room from their reception, are placed in clean and dry trays on a clean laboratory bench (Fig. 8). The sample identification is noted on a label pasted on each tray: ID (database or program), origin (core, profile, composite), sampling depth, tray number/total tray number. With latex gloves on, the sample is spread and mixed in the labelled tray. Each sample is divided to a sufficient number of trays to allow rapid drying. Clods are reduced to small aggregates before spreading them as a layer no thicker than 20–25 mm in each tray. Empty sampling bags and labels are then thrown away. New latex gloves are to be used between each new sample preparation.

Trays containing the samples are stored in a drying room at a controlled temperature of about 30–40 °C. Drying is over when the aggregates are dry inside and outside. The drying time is a minimum of 10 days and lasts until a constant weight is reached.

Laboratory sample preparation

The purpose of sample preparation is to ensure a representative laboratory sample of fine earth (<2 mm particle size) for soil analysis. It involves two steps: crumbling and sieving the dried sample.

Crumbling

This consists of breaking aggregates and separating the coarse fraction from the fine earth using a porcelain mortar and a pestle. Clods are ground manually without crushing the coarse fraction (e.g. gravels and stones). Care must be taken with fragile materials such as pieces of limestone or concretions that should not be crushed. If the soil is fresh, this step of crumbling can be done by hand.

Sieving

The loosened earth is then hand-sieved using a 2-mm mesh sieve to obtain the fine earth fraction (Fig. 9). Coarse particles and roots, corresponding to the residue on the top sieve, are removed using tweezers. Sieving is completed when the coarse particles are clean. The efficiency of the sieving can be judged by checking the color or face of the coarse fraction. Depending on the programs, the coarse fraction is weighed and then either eliminated or archived for further studies.

Sample sub-setting—Quartering

The protocols for quartering are adapted from ISO standards (ISO 11464, 2006; ISO 23909, 2008; NF X31-100, 2020).



Fig. 9 Sieving is a step of soil sample preparation.

The objective is to prepare an aliquot of a representative soil sample in a sufficient quantity from a soil sample. The quartering can be done for two kinds of samples:

- The diagonal method for a large sample and or containing coarse elements or a raw sample (with aggregates).
- The quartering method can be used for any sample, but it is particularly adapted for small samples (if there is only one drying tray or fine earth material).

The diagonal method (Fig. 10)

The raw sample is spread over one or several trays. After the sample has been homogenized by mixing with a scoop, similar quantities are then taken from the whole thickness of the spread sample of each tray with a scoop along the tray's diagonal. If the quantity obtained is not enough, the remaining sample is thoroughly mixed and sampling of each tray is done along the opposite diagonal. In contrast if the sample is too large, it is re-homogenized and reduced using the diagonal method until the required quantity of samples is obtained.

The quartering method (Fig. 11)

The sample, previously homogenized, is spread like a flat pie on a flat surface or in a clean and dry plastic tray. Then, it is divided into four equivalent quarters by tracing two perpendicular axes in the pie with a knife or any other appropriate tool. Two opposite quarters are excluded and the remaining two are carefully mixed. The earth is again divided into 4 quarters. This time, the two opposite quarters excluded are those in the diagonal perpendicular to one of the two previously excluded quarters. This procedure is repeated until a sufficient quantity of sample is obtained. These actions are recorded in the database according to the program.

Preparation of samples from organic horizons

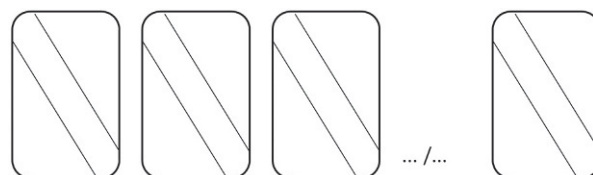
The preparation of organic horizons samples is similar to that of organo-mineral horizons (see above). The samples are weighed, then spread out in thin layers in trays for air drying for 2 to 5 days, until a constant mass is reached. Once dried, the rock fragments (stones and gravels) and living roots, mosses and other living plant material, are manually removed if present in the sample. Dead wood fragments and roots must not be removed from the sample as they are part of the organic horizon. Depending on the programs, the coarse fraction is weighed and then either eliminated or archived for further studies.

The remaining material is then ground to obtain fine organic material ($< 1 \text{ mm}$). This may be done using either a blade or bead mill provided that no contamination from the instrument can be guaranteed. An aliquot of a representative soil sample is extracted according to the diagonal or quartering method, with respect to the quantity of ground sample available and to the quantity of subsample required for soil analysis. The blade mill is cleaned between two samples. The electrical parts are blown with compressed air. The sieve and the receiving bucket are cleaned with non-abrasive sponge and tap water, then rinsed with demineralized water, wiped and blown with compressed air.

Sending for analyses and archiving

Sending laboratory samples for analyses

If samples are prepared at a location different from the one of the laboratory analyses, once prepared they will be sent for analysis in the laboratory. Aliquots aimed at laboratory measurements are typically put in sample bottles (made in PP polypropylene or PEHD



Take a sample along the diagonal of each of n trays, put the n samples in a new clean and dry tray.



Carefully mix all the takings so as to constitute a homogeneous sub-sample.



If the obtained quantity is not sufficient, homogenize again the earth of trays 1 to n and proceed again but taking the opposite diagonal. Begin again until the requested amount is reached.

Fig. 10 the diagonal method is used for the quartering of large soil samples, samples containing coarse elements or raw samples with aggregates.

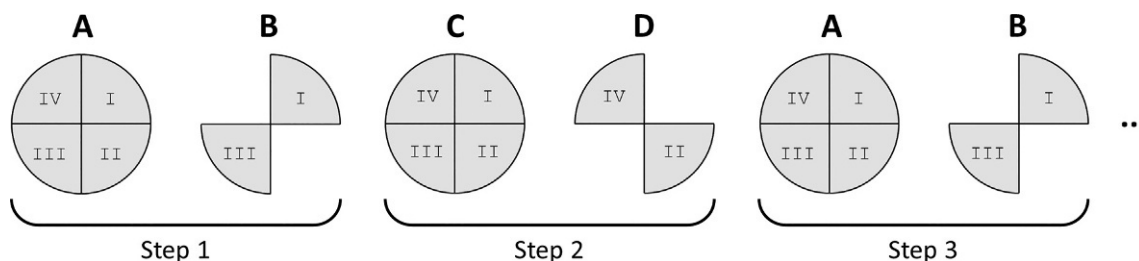


Fig. 11 Description of the quartering method for fine earth sample preparation for bulk density measurements.

polyethylene) (60, 100, 250 or 300 g) or Kraft paper bags with a plastic coating inside (Fig. 12). Before packing them, the samples will be checked to ensure that they are correctly closed and labelled. Type, material and transport are adapted according to the required analyses. Care should be taken to wedge the sample bottles in place, especially if they are made of glass, for example by stuffing the empty spaces. In the package, the samples will be accompanied by a list of all the samples contained in the package and a letter. The correspondence between the list of samples and the package content should be checked before closing it. According to the program, a request will be made to the laboratory to return the remaining samples after analysis. When samples are returned there must be a list that contains the sample database references, so that the correspondence can be established in the laboratory.



Fig. 12 Examples of soil sample containers for sending aliquots for analytical measurement and for long-term storage.

It is convenient, if not necessary, for the analytical laboratory to have the same barcode reader system as the preparation laboratory to ensure traceability of the samples and to avoid any confusion with the samples analyzed.

Archiving

Short-term storage of raw or prepared soil samples is quite common in most laboratories, often for verifying analyses. Long-term storage or archiving has different aims for example (ISO 18512, 2007):

- To take a new aliquot of a sample to check a previous analysis or a change in the method of analysis.
- To be able to determine the initial state of a sample, for example, if a significant change or some pollution has been recorded later on the sampling site.
- To perform new kinds of measurements that did not exist at the time the sample was first analyzed, which can be decades before.
- To provide a representative collection of soil samples for further research studies.

Samples can be re-used if the archiving conditions do not change their properties. The following conditions are considered optimal to prevent evolution, either for raw or fine earth samples:

- Air-dried (30–40 °C) soil samples
- Stored in food grade plastic buckets (e.g. 8 L) or thermo-weldable bags of 1, 2 or 3 L capacity (Fig. 12).
- Hermetically closed and kept in the dark
- Constant climatic conditions of storage as 18 ± 2 °C and 50% humidity.

The use of sliding shelving is a good solution to save space in the archiving facility (Fig. 13). The use of barcodes and the registration of the samples in a database allows quick identification, location and recovery of the samples when needed.

Conclusion

Soil is now back on the political agenda: soil data will be needed more than ever to report on the success of policies. Even if remote sensing techniques are developing fast for assessing and monitoring soils, sampling still remains the first step needed to report on the soil's status. Based on existing standards, this chapter describes the validated and recommended methods for handling of soil samples in the field, during transport to the laboratory and in the laboratory. Despite the constant progress of on-board analytical methods that can be deployed in the field or the development of robots for soil sample preparation in the laboratories, the human hand remains fundamental for the moment to collect and prepare quality samples that ensure repeatable and reproducible samplings to detect any evolution of soil properties without bias. The use of standardized protocols is the prerequisite for comparing data from different countries. This will eventually lead to a globally harmonized soil quality monitoring network, to support and control the efficiency of public policies toward the protection of our soil heritage, soil functions and services rendered by soil.



Fig. 13 Storage of soil samples at the European Soil Samples Conservatory.

References

- ISO 11272 (2017) *ISO 11272: Soil Quality—Determination of Dry Bulk Density*.
- ISO 11464 (2006) *ISO 11464: Soil Quality—Pretreatment of Samples for Physico-Chemical Analysis*.
- ISO 14507 (2003) *ISO 14507: Soil Quality—Pretreatment of Samples for Determination of Organic Contaminants*.
- ISO 18400-102 (2017) *ISO 18400-102: Soil Quality—Sampling—Part 102: Selection and Application of Sampling Techniques*.
- ISO 18400-105 (2017) *ISO 18400-105: Soil Quality—Sampling—Part 105: Packaging, Transport, Storage and Preservation of Samples*.
- ISO 18400-107 (2017) *ISO 18400-107: Soil Quality—Sampling—Part 107: Recording and Reporting*.
- ISO 18512 (2007) *ISO 18512: Soil Quality—Guidance on Long and Short Term Storage of Soil Samples*.
- ISO 23909 (2008) *ISO 23909: Soil Quality — Preparation of Laboratory Samples from Large Samples*.
- Jolivet C, Falcon J-LA, Berche P, Boulonne L, Fontaine M, Gouny L, Lehmann S, Maitre B, Ratié C, Schellenberger E, and Soler-Dominguez N (2022) French Soil Quality Monitoring Network Manual RMQS2: second metropolitan campaign 2016–2027. 2-7380-1451-8. <https://doi.org/10.17180/KC64-NY88><https://hal.inrae.fr/hal-03823026>.
- NF X31-100 (2020) *AFNOR NF X31-100: Qualité des sols—Échantillonnage—Méthode de prélèvement d'échantillons de sol pour analyses physico-chimiques en vue d'une interprétation agronomique*.
- NF X31-503 (1992) *AFNOR NF X31-503: Qualité des sols—Méthodes physiques—Mesure de la masse volumique apparente—Méthode au sable*.
- X31-150 (1984) *AFNOR X31-150: Sols, sédiments, matières fertilisantes, supports de culture—Préparation de l'échantillon pour la détermination d'éléments métalliques traces*.

Digital Soil Morphometrics

Yakun Zhang^a, Alfred E Hartemink^a, Jingyi Huang^a, and Budiman Minasny^b, ^aDepartment of Soil Science, University of Wisconsin-Madison, Madison, WI, United States; ^bSydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Camperdown, NSW, Australia

© 2023 Elsevier Ltd. All rights reserved.

Introduction	568
Measuring soil profile attributes	569
Soil color	569
Soil structure	569
Soil horizons and horizon boundaries	570
Coarse fragments	572
Other soil properties	572
Soil profile variation	573
Soil depth functions	573
Soil profile mapping	574
Soil sampling	575
Applications	575
Conclusion	577
References	577

Abstract

Digital soil morphometrics uses pedometric techniques in the morphological study of soil profiles. It combines tools and techniques for measuring and quantifying soil profile attributes and deriving continuous depth functions. Most soil profile attributes can be measured rapidly and quantitatively using proximal sensing techniques through image analysis methods, including soil color, soil structure, and coarse fragments. More objective and quantitative horizon delineation methods can be executed using cluster analysis of proximally sensed data, e.g., sourced from visible-near infrared and portable X-ray fluorescence spectroscopy, as well as digital images. Moreover, 1-D continuous depth functions and 2-D soil profile maps have been generated to visualize and quantify soil profile variations better. Advanced profile sampling schemes have been established to cover soil profile variation and reduce sampling effort. Digital soil morphometrics offers new quantitative and objective measures of soil profile attributes and their variability.

Key points

- Digital soil morphometrics uses proximal sensing technology for measuring and quantifying soil profile attributes.
- The tools and methods have been used to quantify soil profile attributes, including soil color, soil structure, soil horizons, and coarse fragments.
- Continuous depth functions have been developed to display the vertical distribution of soil properties.
- Soil profile maps have been generated using geostatistical methods or machine learning methods to quantify 2-D variation of soil profiles.
- Proximal sensing increased soil profile sampling density and efficiency, and advanced sampling schemes exist to investigate soil profile variation.

Introduction

Digital Soil Morphometrics was coined by Hartemink and Minasny (2014) as “the application of tools and techniques for measuring and quantifying soil profile attributes and deriving continuous depth functions”. It is a branch of quantitative pedology and investigates the quantification and variation of soil profile characteristics. Furthermore, digital soil morphometrics is extended from “soil morphology” by adding the “digital” and “metrics” parts to indicate the use of quantitative tools and methods to study morphological features of soil.

Soil morphological features have been primarily described in the field in excavated soil pits or trenches using a shovel, auger, knife, root cut, tape measure, water bottle, Munsell color book, chemicals such as Hydrochloric acid (HCl), clinometer, compass, GPS, and camera (Hartemink, 2015). Although the field description is equipped with official guidelines, such as from the *Soil Survey Manual* (Soil Science Division Staff, 2017) and *Field Book for Describing and Sampling Soils* (Schoeneberger et al., 2012), it requires

substantial pedological experience and may be subjective. Soil samples are collected from soil profiles for the analysis of physical, chemical, and biological properties in the laboratory. Because of its cost, however, only a few samples are collected to represent the whole soil profile, which may compromise the accuracy and lead to large sampling uncertainty. The increasing demand for data in decision making to support agricultural management and soil conservation and for environmental modeling has pushed forward the development of more rapid, quantitative, and objective tools and methods to obtain soil information.

Proximal soil sensing technology has been developed and adopted during the past two decades. It has improved the effectiveness of soil survey by providing rapid and reliable ways to estimate soil properties and reducing the effort and time to characterize field-scale variation with on-the-go instruments. Some hand-held proximal soil sensing tools advanced the characterization of soil profile features, as they capture the two-dimensional (2-D) soil profile variation in fine resolution with digital and hyperspectral cameras. Visible near-infrared (vis-NIR) and portable X-ray fluorescence (pXRF) spectroscopy can rapidly increase the sampling density of soil profiles. It increased observations to provide soil data that enhance the understanding of pedological and weathering processes. The high-spectral resolution data collected with proximal sensing tools have improved the objectivity of profile description of soil attributes and horizons. With the increasing accessibility of these tools to many research groups, it will become a routine method to complement soil profile description.

Measuring soil profile attributes

Soil color

Soil color is determined by mineral composition, element concentration, organic matter, and moisture content. It reflects soil properties and soil processes, and it is an important diagnostic feature for soil horizon delineation and soil classification. Soil color has been described on both dry and moist soil samples by comparing its color with the Munsell color chart. The Munsell color system describes a soil color using three attributes—hue, the basic color (e.g., 10YR, 5YR); value, the lightness ranging from 0 (black) to 10 (white); and chroma, the color intensity or saturation ranging from 0 (gray) to 12 (pure color). Soil Munsell color description is subjective, and its readings are affected by light conditions and human perception. The Munsell color chart provides discrete values which hinders its use for quantitative analysis. Therefore, other tools (e.g., digital camera, scanner, colorimeter, and spectrometer) and numeric color systems (e.g., RGB, CIE $L^*a^*b^*$, and HSV) are used to determine soil color. They are reviewed below.

The RGB color represents red (R), green (G), and blue (B), with values ranging from 0 to 255. The CIE (Commission on Illumination) $L^*a^*b^*$ color represents lightness (L^*) ranging from 0 to 100, and colorimetric values from red ($+a^*$) to green ($-a^*$) and yellow ($+b^*$) to blue ($-b^*$). The HSV color represents hue, saturation, and value (lightness). These color space models can be converted from each other by mathematical formulae with open-source software. In addition, soil numeric color systems can be converted from Munsell color or reflectance spectra in the visible range (400–700 nm). Digital cameras and cellphone cameras have been used on individual soil samples and soil profile walls to determine soil color. The RGB, CIE $L^*a^*b^*$, and HSV colors can be extracted from digital images using image analysis software (e.g., ImageJ) or computing software (e.g., R programming, Matlab). As cameras have different settings and the light conditions vary when taking photographs, soil color extracted from digital images should be calibrated using a standard color chart or standard spectrometer measurements to obtain colorimetrically accurate images. Calibration methods have been developed for both dry and moist soil conditions.

Soil colors determined by digital cameras or spectrometers have been used to predict soil properties directly related to soil colors, such as soil organic carbon (SOC), iron oxides, soil texture, and moisture content. Some mobile phone apps have been developed based on the relationship to estimate SOC from images rapidly. These relationships can be applied to predict soil properties from high-resolution soil profile images and to map the 2-D variation in soil profiles. Soil color varies between soil horizons, and therefore soil colors extracted from soil profile images can be used to differentiate horizons (Fig. 1). Other color contrast features, such as redoximorphic features, krotovinas, and rocks, have been identified and their shapes and distribution can be quantified using image analysis methods.

Soil structure

Soil structure describes the arrangement of soil particles and soil aggregates. Soil structure is an important physical property that influences water and nutrient flow, aeration to plants and microbes, and resistance to soil erosion and compaction, through which it affects plant growth. A strong soil structure can provide sufficient water, nutrients, and oxygen to support plant growth and enough space for roots to penetrate, while poor soil structure impedes root growth, water movement and drainage. Soil structure is described qualitatively in the field on the types (granular, blocky, prismatic, platy, for example), size (fine, medium, coarse), and grade (weak, moderate, strong). Likewise, describing soil structure is subjective, and a quantitative and objective method is needed for broader use.

Three-dimensional (3-D) laser scanners such as a multi-stripe laser triangulation (MLT) scanner can take 3-D objects at 360 degrees and build a 3-D structure for visualization and analysis. The surface area and volume of the 3-D object and its geometric parameters can be quantified using image analysis methods on the constructed 3-D images. This technique has been used on soil clods and soil aggregates to determine the volume and bulk density (Hirmas et al., 2013). The MLT scanner has been used to scan 2-D soil profile walls or soil monoliths with a rough surface to identify soil structure and macropores. The MLT-derived profile

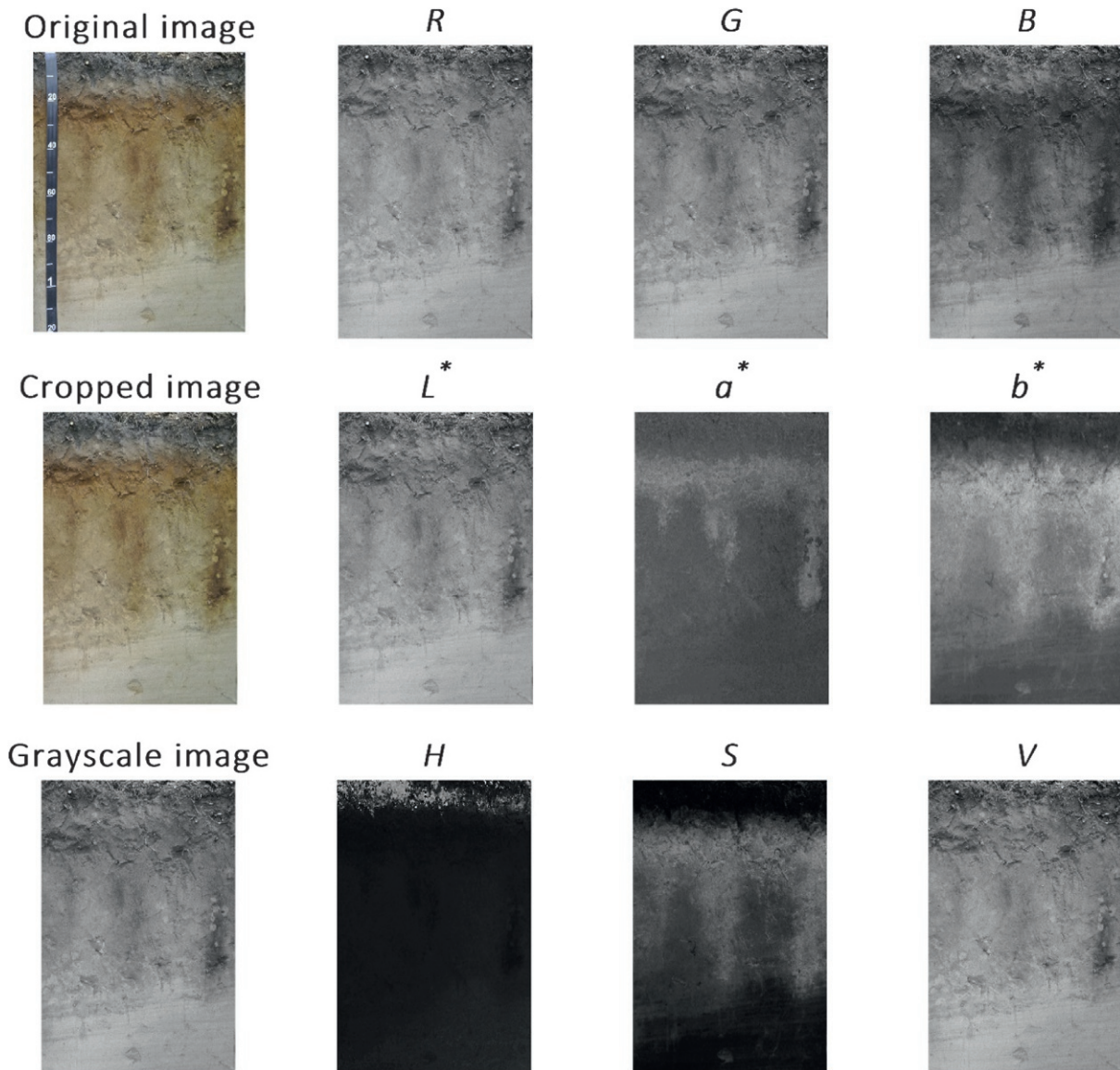


Fig. 1 A sandy, mixed, frigid Entic Haplorthods, Rousseau series from Door County, WI, United States and derived RGB, CIE $L^*a^*b^*$, and HSV colors from the soil profile image.

image reveals structural units and scan gaps representing pore spaces between the structural units, which can be used to derive soil structure indices and quantify macropores (Eck et al., 2016). It can monitor the dynamics of macropores during the drying process and be used to evaluate tillage effects on soil structure.

Soil aggregate stability is an important indicator of soil structure that measures the ability of soil aggregates to resist disintegration when forces from water, wind, and tillage are applied. It can be evaluated using the slaking method which refers to the breakdown of a large, air-dry aggregate (>2–5 mm) into small aggregates (<0.25 mm) when it is immersed in water. A mobile phone app—Slakes, has been developed to measure the area changes over time after soil aggregates are dropped into a petri dish filled with water, which can be fitted into a model to calculate soil aggregate stability indicators (Fajardo and McBratney, 2019). This method has been used to evaluate the effects of tillage on soil aggregate stability in Vertisols (Flynn et al., 2020). It provides a rapid and quantitative measure of soil aggregate and soil structure for effective soil management practices.

Soil horizons and horizon boundaries

A soil profile contains layers often parallel to the soil surface called soil horizons. A soil horizon is the result of soil-forming processes and its physical, chemical, or biological properties differ from adjacent horizons. Soil horizons are identified in the field

by visible features, mainly soil color, texture, structure, rocks, and redoximorphic features. Genetic soil horizons are described by symbols, including capital letters representing master horizons, lowercase letters representing suffixes associated with specific soil characteristics within master horizons, and numbers representing subdivision of horizon suffixes or different parent materials. Soil classification systems use diagnostic horizons, such as epipedons and diagnostic subsurface horizons in USDA Soil Taxonomy, to define soil types. Some soil horizons with distinct color or texture features are easier to identify, such as a Bh horizon in Spodosols with brownish colors stemming from increased organic matter content and sesquioxides, Bt horizon in Alfisols with illuvial accumulation of silicate clay, whereas in some profiles where haploidization processes dominate so that soil horizons are not easily discernible. Additionally, field delineation of soil horizons is subjective and novel tools and mathematical methods have been explored and evaluated to identify soil horizons quantitatively and objectively.

The vis-NIR spectroscopy measures reflectance spectra of soil samples at 350–2500 nm which reflects amounts of organic matter, mineralogy, texture, and moisture. The pXRF spectroscopy measures the concentration of total major and minor elements in soil samples, such as Al, Si, Fe, Ca, Mg, Ti, and Zr, and the spectral response of pXRF at different energy levels can also be extracted and used for further analysis. The vis-NIR spectra and pXRF element data recorded on soil cores have been used in cluster analysis to identify spectrally homogeneous parts which are regarded as soil horizons (Fajardo et al., 2016). Cluster analysis can group objects with similar attributes and separate objects with different attributes using numerical methods. The soil samples with similar characteristics in the vis-NIR spectra and pXRF data are assumed to be from the same horizon, and the transition of different horizons can be identified by calculating a digital gradient, which is considered a horizon boundary. This method has been applied on one-dimensional vertical transects or soil cores.

The 2-D soil profile images contain soil color information which is a major attribute to identify soil horizons. The soil color image or predicted soil properties of the 2-D soil profile had been used in cluster analysis to identify homogeneous regions of soil color or other properties which are regarded as soil horizons (Zhang and Hartemink, 2019a). Soil horizons are often assumed to be homogeneous, but there is substantial within-horizon variation (Hartemink et al., 2020). Fuzzy cluster analysis is advantageous over hard clustering (e.g., *k*-means clustering). It shows the membership of the soil to a specific cluster, i.e., a soil horizon instead of rigorously assigning it to a specific cluster. An example of fuzzy clustering to delineate soil horizons and cluster membership maps is shown in Fig. 2. Delineating soil horizons on 2-D soil profile images is advantageous over 1-D soil cores because the lateral distribution of horizon boundaries can be delineated. Soil horizon boundaries are described on the basis of two characteristics: (1) distinctness indicating the thickness of the transition zones, and (2) boundary topography indicating the shapes or irregularity of the boundary (Schoeneberger et al., 2012). Quantitative methods to assess the horizon boundaries by calculating tortuosity, curvature, distinctness, and topography have been successfully applied on 2-D soil profile images (Zhang and Hartemink, 2019b).

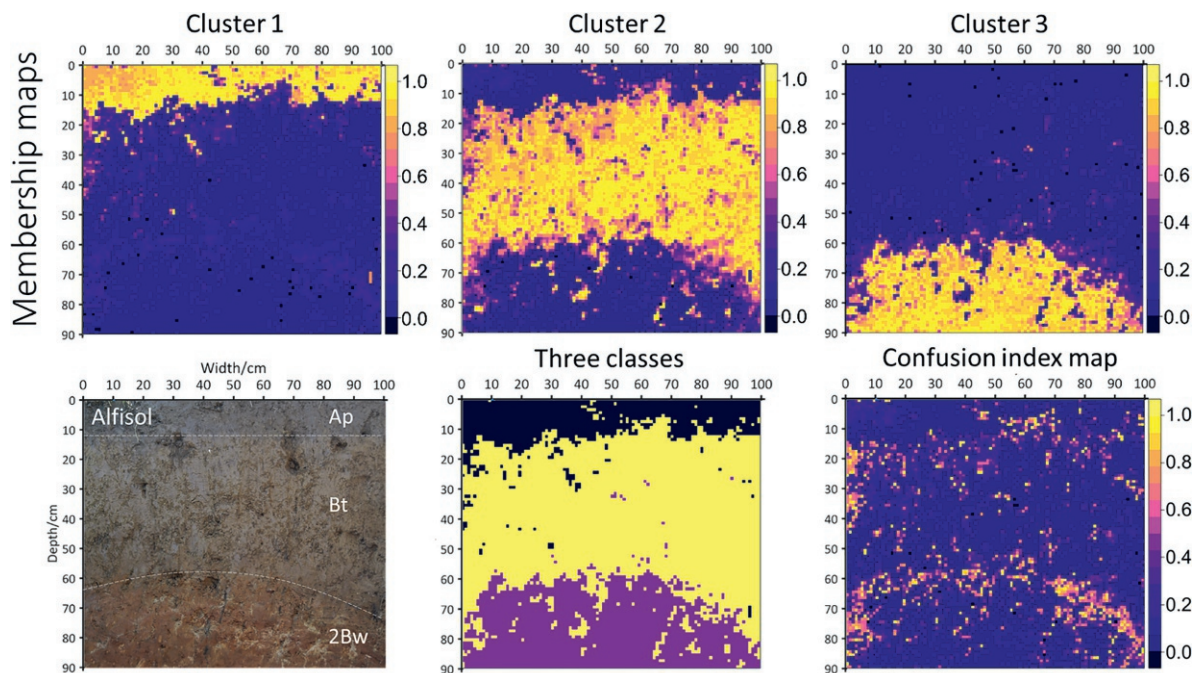


Fig. 2 An example of fuzzy clustering analysis to delineate soil horizons in an Alfisol (fine-silty, over clayey, mixed, superactive, mesic Typic Hapludalfs, NewGlarus series). The membership maps refer to the membership values ranging from 0 to 1 in each cluster corresponding to each horizon. The confusion index map provides implications on horizon boundaries. Zhang Y and Hartemink A.E. (2019a). Digital mapping of a soil profile. *European Journal of Soil Science* **70**: 27–41.

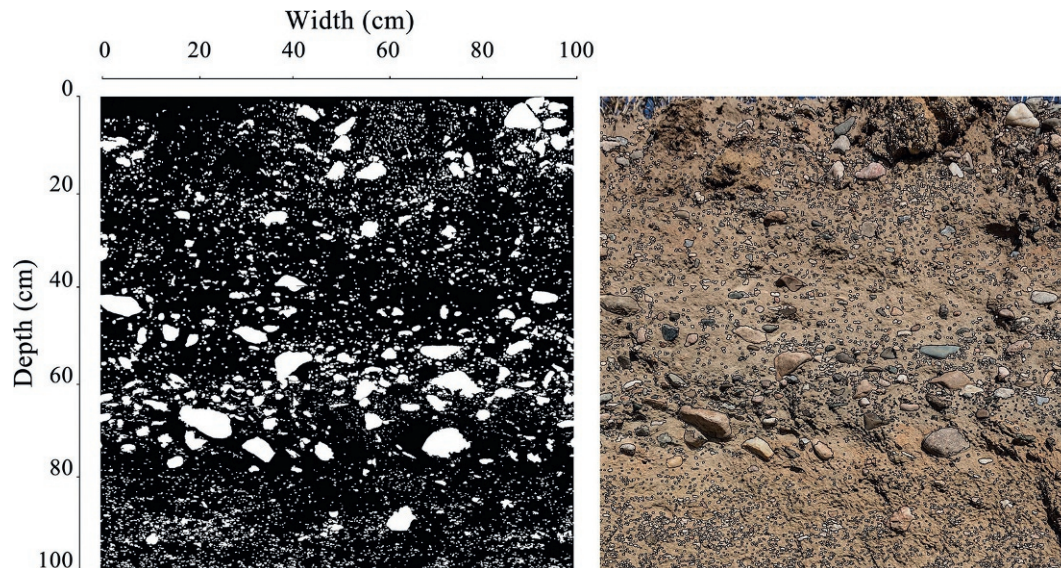


Fig. 3 An example of an image analysis method for identifying coarse fragments in a soil profile image of a coarse-sandy, mixed, superactive, mesic Typic Aquic-Alluvic Primosols. Jiang ZD, Wang QB, Adhikari K, Brye KR, Sun ZX, Sun FJ and Owens PR (2020) A vertical profile imaging method for quantifying rock fragments in gravelly soil. *Catena* **193**: 104590.

The use of proximal sensing tools (e.g., vis-NIR and pXRF spectroscopy, digital cameras) with a clustering approach provides methods to identify quantitatively and objectively homogeneous regions as soil horizons on 1-D soil cores or 2-D soil profiles. Combined with pedological knowledge, it is a powerful tool to delineate and name soil horizons accurately. Soil horizons with more distinct color and textural features, mainly master horizons, are more accurately identified than less distinct horizons.

Coarse fragments

Coarse fragments, also known as rock fragments, refer to particles with a diameter greater than 2 mm. It is an essential attribute that affects soil physical and hydrological properties, soil and water management, plant growth, and soil carbon stock assessment. Coarse fragments are primarily quantified by percentage in weight or percentage in volume in the soil sample. The volume of coarse fragments is determined by the water displacement method, which measures the increase in volume when the rocks are placed in a fixed amount of water in a graduated cylinder. In addition, its size and shape (roundness—very angular, angular, subangular, subrounded, rounded, and well-rounded) are described. Quantifying coarse fragments is essential for soil water management and accurate assessment of SOC stocks, but it can be tedious in stony soils and destructive to soil profiles. Therefore, other efficient methods, such as image analysis, have been developed.

For single gravel fragments, the MLT scanner has been used to construct the gravel's 3-D structure, which can quantify the surface area, volume, and density of the gravel. For a single soil sample, digital images can be taken on the sieved and evenly spread coarse fragments and their content and shape parameters (circularity, aspect ratio, roundness, and solidity) can be quantified using image analysis. To reduce the effort and avoid destruction of the soil profile, image analysis methods have been used to quantify coarse fragments in soil profile images (Jiang et al., 2020)—see Fig. 3. An image segmentation method was used on HSV image colors to differentiate coarse fragments from soil materials, and this method accurately identified 84.9% of the coarse fragments. To construct 3-D structure from 2-D shapes of gravels, a stereology method has been developed to construct the 3-D geometry by fitting an ellipsoid that can estimate the volume of the coarse fragments (Jiang et al., 2021). It provides an efficient way to identify and quantify coarse fragments in soil profiles, increasing the accuracy of assessing SOC stocks in the soil profile.

Other soil properties

Proximal sensing tools have improved the efficiency of measuring other attributes in soil profiles. The pXRF spectroscopy has been used to measure total elements in vertical transects in soil profiles in the field, which reduced the time of chemical analysis in the laboratory. Dense scans of a 2-D soil profile wall characterizes lateral and vertical variations. The vis-NIR and mid-infrared (MIR) spectroscopy have been combined with statistical or machine learning models to predict soil properties (e.g., organic carbon, soil texture, moisture content, carbonates, salinity, pH, etc.). When the predictive models are developed using an existing representative of soils studied, the models can rapidly estimate soil profile attributes. This fulfills the need for collecting soil profile information rapidly to study soil processes or environmental modeling. Future work should focus on improving the spectral models' predictive ability on a wide variety of soil types and developing new tools and techniques to characterize soil profile attributes.

Soil profile variation

Soil depth functions

Soil profile variation occurs from the soil surface to the parent material. As described above, the vertical variation is usually split into homogeneous regions, called soil horizons, within which the soil properties are assumed to be homogeneous. With soil samples collected on a horizon basis and analyzed for soil properties, the vertical distribution of the properties is displayed in a segmented pattern. In many cases, soil properties vary continuously with depth, and such a distribution could be presented more suitably using the continuous depth function. Jenny (1941) stated that every soil property has its own specific depth function. Depth functions with mathematical equations have been developed for several soil properties.

The SOC content generally accumulates in the topsoil and the amounts decrease with depth. The vertical distribution of SOC has been modeled using an exponential decay function with parameters representing carbon content at the soil surface and decreasing rate (Eq. 1) (Minasny et al., 2006). Some soil properties have a peak distribution where the concentration is higher in the subsoil than that in the top or at greater depth. Two peak functions with specific parameters have been developed to model such distributions by Myers et al. (2011). Soil pH in some cultivated fields showed uniform distribution in the topsoil and at greater depth in the soil profile and a transition region (either increasing or decreasing) in between. Such a distribution has been modeled using a sigmoid function with parameters representing the soil pH of the topsoil and at depth and the transition rate (Eq. 2) (Zhang et al., 2017). In addition, the non-parametric equal-area quadratic spline function has been proposed, which has the most flexible shape to model any vertical distributions and the areas above and below the fitting spline in each sampled horizon are equal (Bishop et al., 1999). Some examples of the exponential decay function, peak function, and sigmoid function are shown in Fig. 4.

$$y = C_a e^{-kx} + C_b, \quad (1)$$

$$y = s + \frac{d-s}{1 + \left(\frac{x}{\alpha}\right)^{-k}}, \quad (2)$$

where C_a (kg m^{-3}) is the difference in SOC content between the soil surface and the bottom of the profile, C_b (kg m^{-3}) is the SOC content at the bottom of the profile, $C_a + C_b$ (kg m^{-3}) is the SOC content at the soil surface, k (m^{-1}) is the decreasing rate of SOC with depth (Eq. 1). In Eq. (2), s is the soil pH at the top of the profile, d is the soil pH at the bottom of the profile, k is the hillslope parameter representing the steepness of the increasing or decreasing trend of the transition zone, and α is the inflection point representing the midpoint the transition zone. In both equations, x (m) is the soil depth, and y is either the SOC content (kg m^{-3}) or soil pH at depth x .

Minasny et al. (2016) summarized seven different types of depth functions, including uniform, gradational, exponential, wetting front, abrupt, peak, and minimum-maximum. The depth functions developed can be used to understand soil formation and processes that occurred in the soil. The uniform distribution indicates little variation in the soil profiles which can be observed in young soil profiles where soil properties are similar to the parent material attributes. As soil development intensifies, other types of distributions can be observed for different soil properties. For example, as soil acidifies in the topsoil, soil pH distribution shows an

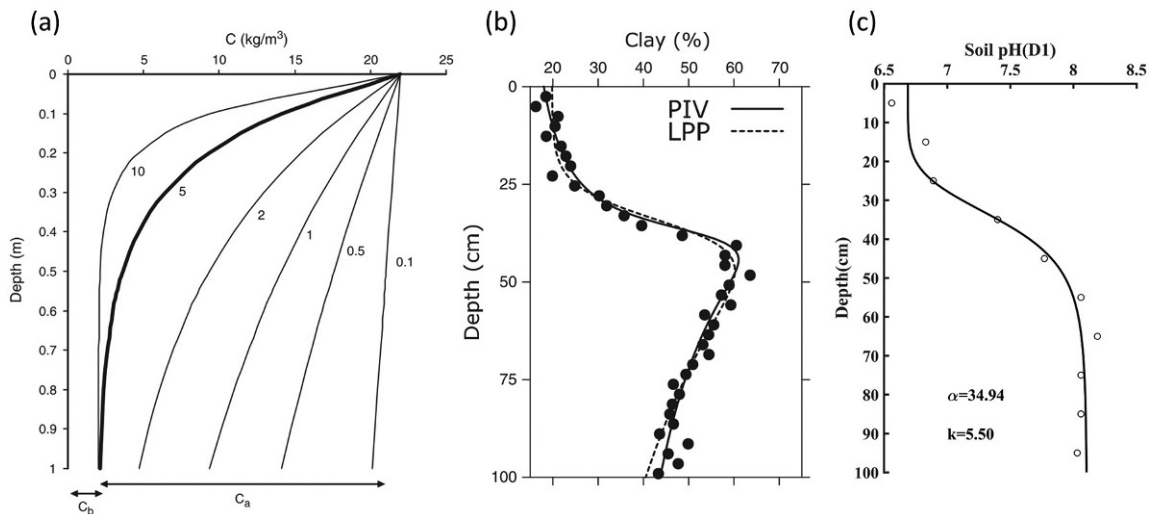


Fig. 4 Examples of depth functions. (A) An exponential decay function to model the decreasing pattern of organic carbon in soil profiles; (B) a peak function to model the distribution of clay content; (C) a sigmoid function to model the vertical distribution of soil pH in agricultural field (A) Minasny B, McBratney AB, Mendonça-Santos ML, Odeh IOA and Guyon B (2006) Prediction and digital mapping of soil carbon storage in the Lower Namoi Valley. *Soil Research* **44**: 233–244; (B) Myers DB, Kitchen NR, Sudduth KA, Miles RJ, Sadler EJ and Grunwald S (2011) Peak functions for modeling high resolution soil profile data. *Geoderma* **166**: 74–83; (C) Zhang Y, Biswas A and Adamchuk VI (2017) Implementation of a sigmoid depth function to describe change of soil pH with depth. *Geoderma* **289**: 1–10.

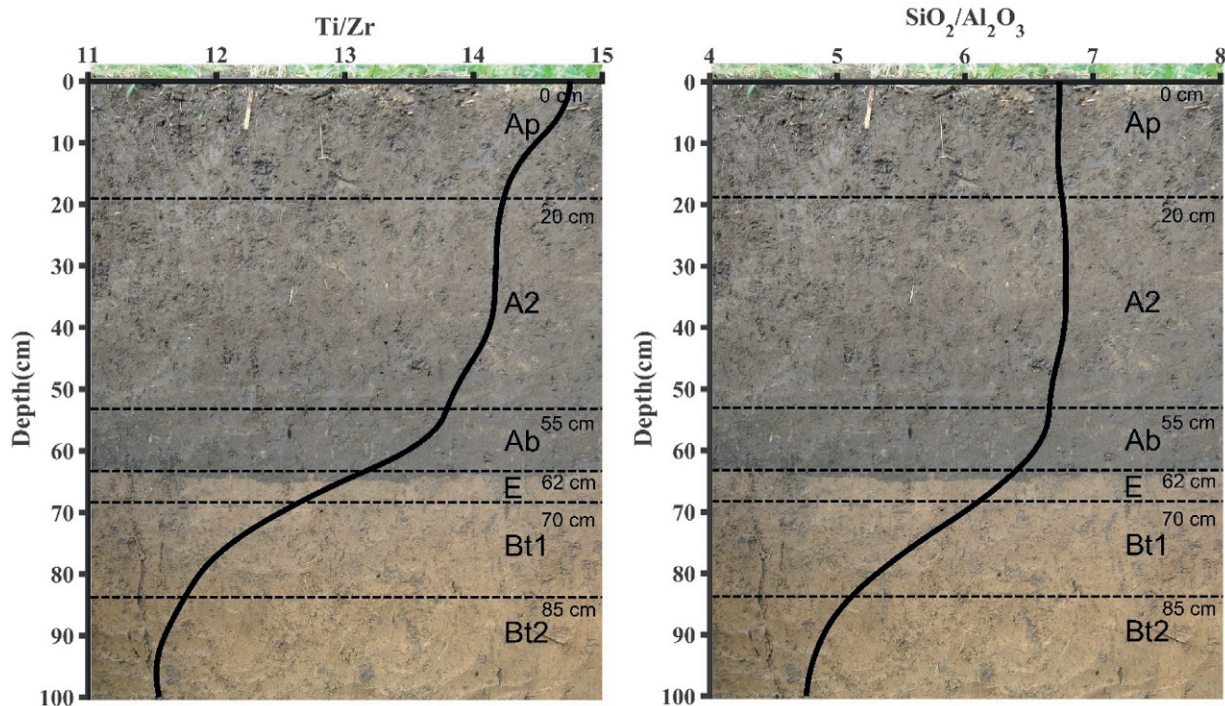


Fig. 5 The distribution of Ti/Zr and $\text{SiO}_2/\text{Al}_2\text{O}_3$ derived from pXRF elemental data in a fine-loamy, mixed, superactive, mesic Pachic Argiudolls (Troxel series) in Madison, Wisconsin. Depth functions were fitted by 12 moist samples using an equal-area quadratic spline with $\lambda = 1$. Zhang Y and Hartemink AE (2019c) Soil horizon delineation using vis-NIR and pXRF data. *Catena* **180**: 298–308.

increase with depth. When a buried topsoil horizon (Ab horizon) is formed, a peak distribution can be observed for SOC content at the depth of the buried Ab horizon. The distribution of the elemental ratio of Ti/Zr (titanium/zirconium) has been used to indicate heterogeneity of parent materials because they weather slowly and can retain the characteristics of parent materials. The distribution of the elemental ratio of $\text{SiO}_2/\text{Al}_2\text{O}_3$ (Silicon dioxide/Aluminium oxide) has been used to indicate the rate of weathering in the soil profile, as Si is a mobile element and can be easily weathered and leached, while Al is an immobile element and can accumulate during weathering. An example of the vertical distribution of Ti/Zr and $\text{SiO}_2/\text{Al}_2\text{O}_3$ in a Mollisol is shown in Fig. 5, where the soil was formed in two different parent materials in the 1-m profile—aeolian deposits (loess) over glacial outwash (Zhang and Hartemink, 2019c).

By fitting a depth function to few observations, we can predict soil properties at any depth. It has been used to model the vertical variation of soil in 3-D maps. The exponential decay function has been used for 3-D SOC mapping in which the parameters of the function are interpolated to unsampled locations to reconstruct the vertical distribution of SOC at new locations. The development of proximal sensing tools markedly increases the sampling density in soil cores or vertical transects. The pXRF spectroscopy measures elements (e.g., Si, Al, Fe, Ca, Ti, and Zr), closely related to soil weathering and parent material. The vis-NIR spectroscopy can predict soil properties related to organic components and soil mineralogy. The use of these two instruments improves the ability to derive continuous depth functions of many soil properties and to understand soil formation and processes in soil profiles. In addition, it can be potentially used to monitor the changes of depth functions of dynamic soil properties.

Soil profile mapping

Soil properties vary horizontally and vertically in the profile; however, the horizontal variation is understudied within a soil profile compared to the vertical variation. Digital soil mapping techniques have been applied on soil profile walls to map the 2-D variation in soil properties. To generate a soil profile map, soil samples are first collected from a grid with 10-cm or 20-cm intervals across the soil profile wall and analyzed for soil properties. These samples often have a spatial resolution that does not capture the variation within the 10-cm or 20-cm intervals. Geostatistical methods (e.g., ordinary kriging with point or block support to produce maps, contour mapping) can be applied on these coarsely collected samples to interpolate the soil property values to generate 2-D fine-resolution soil maps.

Along the lines of landscape-scale soil mapping, environmental covariates can be used in machine learning methods to improve the mapping accuracy and spatial resolution on soil profile mapping. Digital images of soil profiles are the most common and readily accessible covariates that have high spatial resolution and convey soil color, soil texture, carbon content, iron oxides, coarse fragments, and redoximorphic features. Soil profile images have been combined with soil properties using the random forest method to map the 2-D variation of soil carbon, soil texture, and weathering indices (Zhang and Hartemink, 2019a). Hyperspectral

images are high-resolution 2-D images where each pixel consists of a vis-NIR reflectance spectrum with a spectral range of 400–1100 or 350–2500 nm instead of RGB color (Steffens and Buddenbaum, 2013). The hyperspectral images, comparable with vis-NIR spectra, contain more information on soil properties compared to digital images and are good variables for soil profile mapping. It has been used with machine learning methods to map the variation of soil carbon, nitrogen, clay content, elemental concentrations, hematite, calcite, salt, and water content of soil profiles. An example of profile mapping of organic matter fractions, Fe, and Mn in a Luvisol (Alfisol) using the hyperspectral image is shown in Fig. 6.

The generated soil profile maps display the distribution of soil properties in the profile and improve the quantification of variation in soil horizons. These maps can be used to investigate pedological processes and weathering conditions of soil profiles and have been used in fuzzy clustering to identify soil horizons objectively and to quantify soil horizon purity and variation. These maps provide more detailed information and more accurate measurements on soil profile properties, and provide a strategy to characterize fine-resolution and deep soil features. The soil profile mapping technique has been used repeatedly on soil profile walls to monitor soil processes. For example, the hyperspectral camera has been used to take soil profile images continuously over 5 days after a rainfall event and the generated soil profile maps can monitor the movement of salt in the soil profile (Wu et al., 2018).

Many geophysical techniques have been used to characterize or map the horizontal and vertical variation of soil profile and soil transects in a non-destructive way, that is, without digging a soil pit. Electromagnetic induction instruments can measure the soil's apparent electrical conductivity (ECa) from the soil surface down to approximately 6 m. Soil ECa is influenced by many soil properties, such as soil moisture, clay, and salt, and therefore, can be used to estimate these soil properties through calibration models with individual soil properties. This technique has been used to map the soil moisture content along a 350-m transect and to delineate soil moisture dynamics and soil-drying process during evapotranspiration (Huang et al., 2017). Ground penetrating radar measures the dielectric constant of soil which is influenced by soil moisture, organic matter, bulk density, and texture. It can generate 2-D images of soil profiles for characterizing its variation, identifying different soil horizons, estimating depth-to-bedrock, and monitoring the movement of water or solutes.

Soil sampling

Soil samples collected from soil profiles for laboratory analysis are taken to represent variation within the profile. Since the variation is mostly observed in the vertical direction, the soil profile is usually divided into soil horizons or layers from where the soil samples are taken. Soil samples are either collected from horizon centers or mixed from the entire horizon to represent the soil property of the corresponding horizon. Soil samples collected in this way requires accurate delineation of soil horizons, but it does not capture the variation within soil horizons and disregards the horizontal variation. In some studies, soil samples are collected from fixed depth intervals (e.g., 10-cm interval) in soil cores. This method is convenient for recording and is often used for large-scale soil mapping. In other studies, more samples are collected at small intervals in the topsoil which is more sensitive to human activities and climate change, and fewer samples are collected in deep soils which are more static.

To capture the 2-D variation of soil profiles, other sampling methods have been developed to complement the current sampling schemes in the vertical direction. Raster sampling with a fixed interval (e.g., 10-cm, 15-cm) gives the best coverage of the soil profile. This method is mainly used to collect samples for 2-D soil profile mapping. But, it is time-consuming and labor-intensive. To reduce the sampling effort, a few vertical transects can be collected to cover the horizontal variation of the soil profile, such as collecting three transects from left, middle, and right of the profile. Jones (2017) performed a variogram analysis of proximally-sensed data of 15 soil profiles and found that the variation in soil property variance is stable with depth. This finding means that a soil profile should be sampled with equal density in the topsoil and subsoil. The lateral variance in soil properties of the soil profile is much less than the vertical variation. A 4-cm vertical increment is equivalent to the variance experienced over an entire 1-m lateral cross-section of the profiles.

Other available variables (e.g., color information from digital images) can be used to optimize the sampling scheme in soil profiles. It is based on the assumption that a good coverage in digital images (i.e., feature space) can largely guarantee a good coverage of soil profile variation. These variables can be used in conditioned Latin hypercube sampling or fuzzy *k*-means sampling to optimize the sampling scheme and reduce the sample size needed for covering the soil profile variation.

The development of proximal sensing tools reduces the sampling effort and increases the sampling density in soil profiles. Many soil properties can be rapidly predicted from proximal sensing tools, such as vis-NIR, MIR, and pXRF spectroscopy. This method expands the information obtained from soil profiles and captures more fine-scale soil variation. In some cases, soil modeling may be more accurate using densely estimated soil properties from proximal sensors compared to sparsely collected and measured soil data. Therefore, depending on different research objectives and required accuracy, one should determine the appropriate sampling schemes and proximal sensing techniques for soil profile sampling.

Applications

Digital soil morphometrics techniques and methods have improved the measurement accuracy and sampling density of soil profile attributes, and have provided more quantitative measurements of soil properties. These methods have found several practical applications.

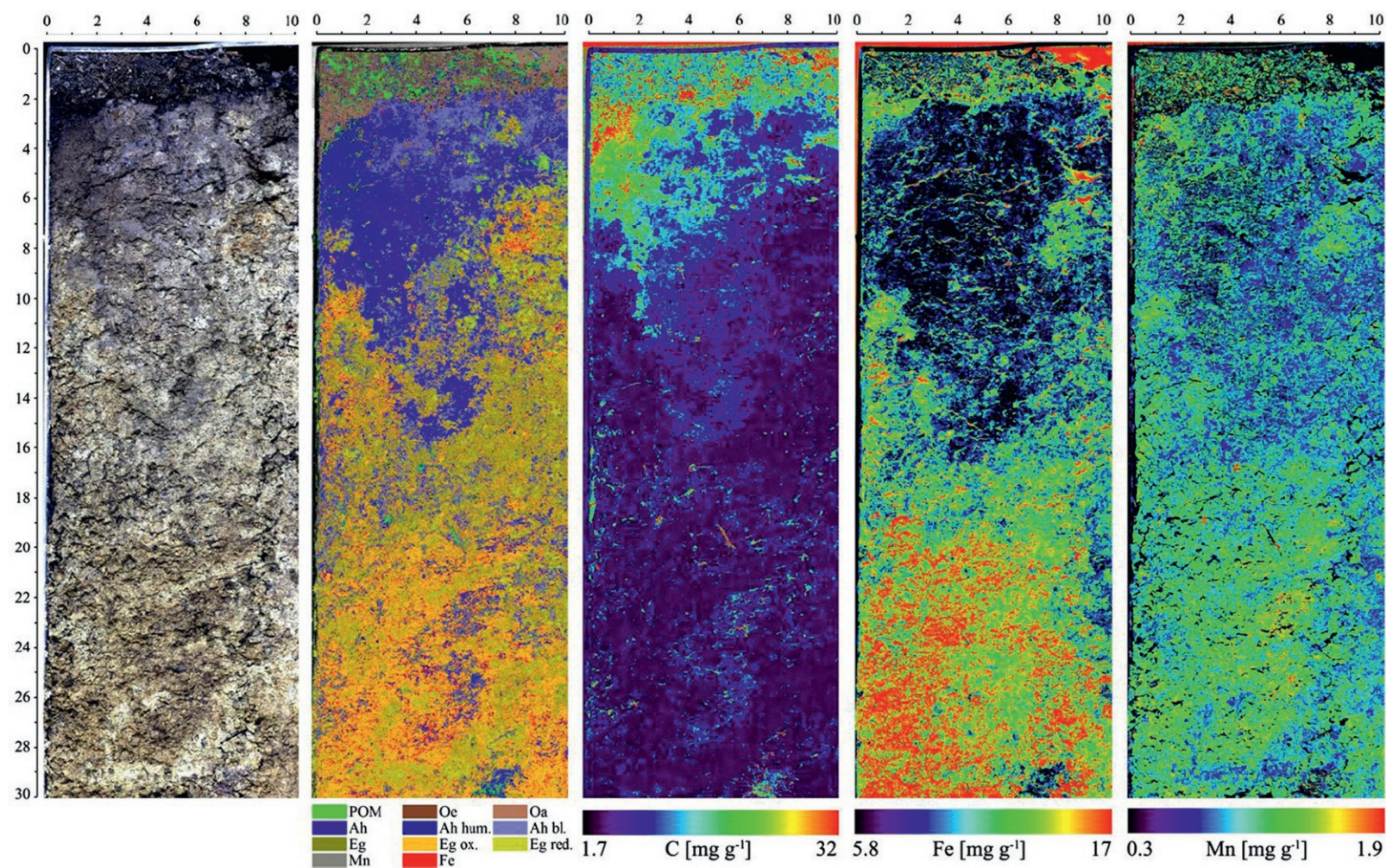


Fig. 6 Soil profile mapping of organic matter fractions, Fe, and Mn in a stagnic Luvisol, greyic, silty. Steffens M and Buddenbaum H (2013) Laboratory imaging spectroscopy of a stagnic Luvisol profile—High resolution soil characterisation, classification and mapping of elemental concentrations. *Geoderma* **195–196**: 122–132.

Accurate assessment of SOC stock in soil profiles is important due to its essential role in biogeochemical cycling, food production, and climate change mitigation. Because of the laborious nature of sampling in deep soils and its cost, many studies sample only the top 30 cm and disregard the soil carbon in deep soils that is more recalcitrant because of less disturbance, leading to underestimating SOC stock. Most studies collect one soil core from each location, ignoring the horizontal variation of soils, leading to large uncertainty in estimates. The development of proximal sensing tools and advanced profile sampling schemes can increase sampling density without adding too much cost in soil analysis and thus result in a more accurate estimation of SOC stocks. Many soils contain coarse fragments, which should be subtracted when calculating SOC stock. The proposed profile imaging method to quantify coarse fragments in stony soils provides a fast and efficient way to account for coarse fragments for accurate assessment of SOC stock in soil profiles (Jiang et al., 2021).

Cropping and tillage practices affect soil properties, such as SOC content, nitrogen content, soil structure, and soil compaction, influencing plant growth. These effects behave differently for topsoil and subsoil. Accurate evaluation of these effects on soil properties is important for performing appropriate soil management practices. Studies have shown that proximal sensing tools improved the estimation accuracy of soil properties compared to sparsely collected soil samples. Hyperspectral imaging methods have been used to generate high-resolution profile maps of SOC, nitrogen, and clay content which provided more accurate information than collected soil samples when evaluating the effects of cropping systems on soil properties (Sorenson et al., 2020). The MLT scanner has been used to evaluate how the structure changes under different tillage practices in 27 soil profiles, which is more efficient and objective than the traditional way of identifying soil structure (Bagnall et al., 2020).

Soil physical properties (e.g., macropores) and physical processes (e.g., movement of water and solutes in the profile) are important in regulating many soil processes, such as water and nutrient uptake by plants, drainage, and transport of contaminants. Proximal sensing tools can monitor soil processes occurring in the soil profiles. The soil structure feature estimated from the MLT scanner was strongly correlated with soil hydraulic conductivity and can be used as a proxy of soil hydraulic properties (Eck et al., 2016). This technique can also be used to quantify macropores of soil. The hyperspectral imaging technique can take soil profile images rapidly with spectral information correlated with many soil properties. Therefore, it can monitor the dynamics of water and salt content of soil profiles after a rainfall or irrigation event.

Conclusion

Digital soil morphometrics uses proximal sensing techniques to measure and quantify soil profile attributes. Digital cameras, vis-NIR spectroscopy, pXRF spectroscopy, hyperspectral imaging, and 3-D laser scanner have been applied to quantify soil profile properties, including soil color, soil texture, soil structure, elemental concentration, coarse fragments, etc. Novel methods have been developed to delineate soil horizons and horizon boundaries automatically using proximal sensing tools and cluster analysis. The distribution and variation of soil properties in soil profiles have been quantified through 1-D continuous depth functions or 2-D soil profile maps. Digital soil morphometrics techniques and methods greatly improve the measurement accuracy and efficiency of soil profile attributes, contributing to a more accurate assessment of SOC stocks, evaluation of cropping and tillage practices on soil properties, and monitoring soil processes.

In future work, proximal sensors should be developed and tested for quantifying other soil profile characteristics, such as redoximorphic features, pores and roots. Sensor data fusion methods have the potential to quantify the soil profile attributes more accurately. Deploying the tools and methods to field condition increases measuring efficiency and advanced methods, such as deep learning, can be evaluated for automatic soil horizon delineation. A range of diverse soil profiles should be evaluated with the existing methods and new methods for soil horizon delineation and non-invasive tools could be evaluated for delineating soil horizons. Finally, standard procedures and guidelines may be developed for pedologists as a routine quantitative and analytical method for field soil description.

References

- Bagnall DK, Jones EJ, Balke S, Morgan CLS, and McBratney AB (2020) An in situ method for quantifying tillage effects on soil structure using multistripe laser triangulation. *Geoderma* 380: 114642.
- Bishop TFA, McBratney AB, and Laslett GM (1999) Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma* 91: 27–45.
- Eck DV, Qin M, Hirmas DR, Giménez D, and Brunsell NA (2016) Relating quantitative soil structure metrics to saturated hydraulic conductivity. *Vadose Zone Journal* 15. <https://doi.org/10.2136/vzj2015.05.0083>.
- Fajardo M and McBratney AB (2019) *Slakes: A Soil Aggregate Stability Smart-Phone App [Mobile Application Software]*. Retrieved from: <https://play.google.com/store/apps/details?id=slaker.sydneyuni.au.com.slaker&hl=en> Australia: The University of Sydney.
- Fajardo M, McBratney A, and Whelan B (2016) Fuzzy clustering of Vis-NIR spectra for the objective recognition of soil morphological horizons in soil profiles. *Geoderma* 263: 244–253.
- Flynn KD, Bagnall DK, and Morgan CLS (2020) Evaluation of SLAKES, a smartphone application for quantifying aggregate stability, in high-clay soils. *Soil Science Society of America Journal* 84: 345–353.
- Hartemink AE (2015) New tools for pedologists: Digital soil morphometrics. *Soil Horizons* 56: 1–2.
- Hartemink AE and Minasny B (2014) Towards digital soil morphometrics. *Geoderma* 230: 305–317.
- Hartemink AE, Zhang Y, Bockheim JG, Curi N, Silva SHG, Grauer-Gray J, Lowe DJ, and Krasilnikov P (2020) Soil horizon variation: A review. *Advances in Agronomy* 160: 125–185.
- Hirmas DR, Giménez D, Subroy V, and Platt BF (2013) Fractal distribution of mass from the millimeter- to decimeter-scale in two soils under native and restored tallgrass prairie. *Geoderma* 207–208: 121–130.
- Huang J, Scudiero E, Clary W, Corwin DL, and Triantafyllis J (2017) Time-lapse monitoring of soil water content using electromagnetic conductivity imaging. *Soil Use and Management* 33: 191–204.

- Jenny H (1941) *Factors of Soil Formation: A System of Quantitative Pedology*. New York: McGraw-Hill.
- Jiang Z-D, Wang Q-B, Adhikari K, Brye KR, Sun Z-X, Sun F-J, and Owens PR (2020) A vertical profile imaging method for quantifying rock fragments in gravelly soil. *Catena* 193: 104590.
- Jiang Z-D, Wang Q-B, Brye KR, Adhikari K, Sun F-J, Sun Z-X, Chen S, and Owens PR (2021) Quantifying organic carbon stocks using a stereological profile imaging method to account for rock fragments in stony soils. *Geoderma* 385: 114837.
- Jones E (2017) *Proximal Sensing in Soil Profiles*. PhD Thesis Australia: The University of Sydney.
- Minasny B, McBratney AB, Mendonça-Santos ML, Odeh IOA, and Guyon B (2006) Prediction and digital mapping of soil carbon storage in the Lower Namoi Valley. *Soil Research* 44: 233–244.
- Minasny B, Stockmann U, Hartemink AE, and McBratney AB (2016) Measuring and modelling soil depth functions. In: Hartemink AE and Minasny B (eds.) *Digital Soil Morphometrics*, pp. 225–240. Dordrecht: Springer International Publishing.
- Myers DB, Kitchen NR, Sudduth KA, Miles RJ, Sadler EJ, and Grunwald S (2011) Peak functions for modeling high resolution soil profile data. *Geoderma* 166: 74–83.
- Schoeneberger PJ, Wysocki DA, and Benham EC (2012) *Field Book for Describing and Sampling Soils, Version 3.0*. Lincoln, NE: USDA Natural Resources Conservation Service, National Soil Survey Center.
- Soil Science Division Staff (2017) Soil survey manual. In: Ditzler C, Scheffe K, and Monger HC (eds.) *USDA Handbook 18*. Washington, DC: Government Printing Office.
- Sorenson PT, Quideau SA, Rivard B, and Dyck M (2020) Distribution mapping of soil profile carbon and nitrogen with laboratory imaging spectroscopy. *Geoderma* 359: 113982.
- Steffens M and Buddenbaum H (2013) Laboratory imaging spectroscopy of a stagnic Luvisol profile—High resolution soil characterisation, classification and mapping of elemental concentrations. *Geoderma* 195–196: 122–132.
- Wu S, Wang C, Liu Y, Li Y, Liu J, Xu A, Pan K, Li Y, and Pan X (2018) Mapping the salt content in soil profiles using Vis-NIR hyperspectral imaging. *Soil Science Society of America Journal* 82: 1259–1269.
- Zhang Y and Hartemink AE (2019a) Digital mapping of a soil profile. *European Journal of Soil Science* 70: 27–41.
- Zhang Y and Hartemink AE (2019b) A method for automated soil horizon delineation using digital images. *Geoderma* 343: 97–115.
- Zhang Y and Hartemink AE (2019c) Soil horizon delineation using vis-NIR and pXRF data. *Catena* 180: 298–308.
- Zhang Y, Biswas A, and Adamchuk VI (2017) Implementation of a sigmoid depth function to describe change of soil pH with depth. *Geoderma* 289: 1–10.

Proximal soil sensing in the field

Mikaela Tilse^{a,c}, Uta Stockmann^b, and Patrick Filippi^{a,c}, ^aSydney Institute of Agriculture, The University of Sydney, Camperdown, NSW, Australia; ^bCommonwealth Scientific and Industrial Research Organisation (CSIRO), Agriculture and Food Business Unit, Canberra, ACT, Australia; ^cSchool of Life and Environmental Sciences, The University of Sydney, Camperdown, NSW, Australia

© 2023 Elsevier Ltd. All rights reserved.

Introduction	579
Sensor classification	580
Advantages of proximal sensing	582
Disadvantages of proximal sensing	582
Types of sensors and their applications	582
On-the-go sensors	582
Electromagnetic induction (EMI)	582
Gamma-rays	584
COSMOS	584
Time-domain reflectometry (TDR) and ground-penetrating radar (GPR)	585
Soil compaction and strength	585
Point-based sensors	585
Visible, near and mid-infrared diffuse reflectance spectroscopy	585
Portable X-ray fluorescence spectrometry	586
Soil moisture sensors	587
Stop-and-go sensors	587
Predicting soil properties	587
Calibration and spectral libraries	587
Multi-sensor data fusion	588
Future directions of proximal sensing	588
References	589

Abstract

Proximal soil sensing in the field is the application of handheld, ground, or vehicle-based sensors to obtain data when the sensor is placed in contact, or within 1–2 m proximity, of the soil. Proximal soil sensors operate as on-the-go, stationary, or stop-and-go systems, and cover a broad range of the electromagnetic spectrum. Field-based proximal soil sensors are a cheaper, less time and labor-intensive alternative to conventional laboratory analyzes, but are an indirect method that often requires site and time-specific spectral libraries and calibration models for meaningful interpretation. Multi-sensor data fusion and multi-sensor approaches are explored, alongside future directions.

Key points

- Proximal soil sensors are defined and classified according to their model of operation, with particular focus on applications in the field.
- Advantages and disadvantages of proximal soil sensing is discussed.
- Types and applications of on-the-go, point-based, and stop-and-go field-based proximal soil sensors are described.
- The development and importance of calibration models is highlighted.
- Multi-sensor data fusion and multi-sensor approaches are discussed.
- Future directions of proximal soil sensing are explored, including shifting technologies from the research to commercialization phases.

Introduction

Soil varies spatially across landscapes and with depth, and temporally both within, and between years. Understanding this spatial and temporal variation is crucial for a range of stakeholders, including land managers, farmers, and policy makers. Timely and accurate information describing soil variability is needed at increasingly finer (spatial and temporal) scales, but obtaining a representative sample of a particular soil landscape is an ongoing challenge (Adamchuk et al., 2004). While conventional soil sampling and laboratory analysis methods have contributed greatly to our scientific understanding of the soil landscape, they are

often laborious, time-consuming, and expensive (Viscarra Rossel et al., 2011). Sensor-based technologies, which can be used in the field can overcome these shortcomings and facilitate the acquisition of high-resolution soil data, allow better understanding of the spatial and temporal dynamics of soil.

Proximal soil sensing (PSS) in the field is the application of handheld, ground, or vehicle-based sensors to obtain data on (soil) properties when the sensor is placed in contact, or within 1–2 m proximity, of the soil (Viscarra Rossel and McBratney, 1998; Viscarra Rossel et al., 2010). The signals detected can be related to specific soil properties directly, or are proxies which require calibration. This definition of PSS excludes remote sensing via airborne or satellite platforms, and laboratory measurements with benchtop instruments.

Proximal soil sensors can rapidly provide users with cost-effective, detailed and timely information about soil physical, chemical, and biological properties (Viscarra Rossel et al., 2011). While some proximal sensors may not provide results that are as accurate or precise as conventional analyzes, per individual measurement, the collection of large amounts of (spatial and temporal) data via PSS offers users with more information than ever before about the soil landscape. Technological advances, improved accessibility, and reduced costs of PSS tools are contributing to an increased shift of PSS from the research-phase, to being implemented in more applied and commercial settings. The list of available PSS systems is expanding year-on-year, although the rate varies for different sensors, and in different areas globally.

This chapter focuses on PSS in the field and does not cover the use of proximal sensors in a laboratory setting. Types of proximal soil sensors, recent advances in commercialization, and their applications will be described. Multi-sensor approaches, sensor calibration and prediction will also be explored, including the future of PSS. Numerous reviews on the principles and applications of PSS for agriculture and natural resource management can be found elsewhere and are referred to throughout.

Sensor classification

Viscarra Rossel et al. (2011) classify proximal soil sensors into four main categories according to:

- The manner in which they measure (invasive [in situ or ex situ] or non-invasive);
- Their energy source (active or passive);
- How they operate (stationary or mobile); and
- The inference used in the measurement of the target soil property (direct or indirect).

Additional characterization criteria from Gebbers (2018) also classify proximal soil sensors according to:

- Their temporal responsiveness (the time taken for measurement, processing, and subsequent decision-making);
- Their selectivity (the extent to which a sensor can determine a particular property without interference); and
- Scope of the electromagnetic (EM) spectrum that they encompass (ranging from gamma-rays to mechanical methods).

Table 1 describes the classifications of common proximal soil sensors within the EM spectrum used in the field, and their applications to a range of soil properties.

Invasive sensors require a degree of sample preparation or direct contact (e.g., insertion into or extraction of the soil). Invasive in situ sensors measure soil properties in place, usually in direct contact with the soil (e.g., soil moisture sensors), whereas invasive ex situ systems require sample extraction and placement of the sample in contact with the sensing element for analysis (e.g., ion-selective electrodes) (Adamchuk et al., 2018). Non-invasive sensors measure from a proximal distance and do not require any sample preparation or processing (e.g., electromagnetic induction, EMI, sensors). Passive sensors measure ambient sources of energy (e.g., natural radioactive decay), whereas active sensor systems provide their own source of energy (e.g., ground penetrating radar, GPR) (Viscarra Rossel et al., 2011).

Proximal soil sensors operate as either on-the-go (mobile), stationary (point-based), or stop-and-go sensors (see section on types of sensors and their applications). On-the-go (mobile) sensors continuously collect data as they move across an area (e.g., EMI; Gebbers, 2018). Stationary sensors may be handheld (e.g., handheld infrared spectrometer) or permanently fixed sensors (e.g., a soil moisture probe buried in the soil), and must remain stationary during sample collection and measurement. Stop-and-go sensors can be moved from place-to-place, but need to remain stationary during sample collection and operate automatically (e.g., Veris®; see section on Stop-and-go sensors). Sensors that can provide data immediately (e.g., soil moisture or pH sensors) can be used for real-time management decisions (e.g., irrigation scheduling). If data processing, including cleaning and calibration, is required, sensors can be used for asynchronous applications (e.g., Digital Soil Mapping, DSM) (Gebbers, 2018).

Sensor selectivity indicates how well a sensor can determine a particular property without interference (Gebbers, 2018). High selectivity sensors (e.g., pH electrodes) are preferred, however there is a trade-off between sensor selectivity, cost, and robustness (Gebbers, 2018). Non-selective sensors such as EMI sensors, which measure electrical conductivity (EC), are widely used to assess a range of soil properties. While sensors that measure a physical process (i.e., direct measures) are most desirable, these are generally more expensive, less developed, and more technically demanding (Viscarra Rossel and Lobsey, 2016). Indirect methods, which rely on a proxy, are more readily available, more mature, and are cheaper to own and operate (Viscarra Rossel and Lobsey, 2016). However, these may be less accurate as they require site-specific calibration, are inherently more subjective (as they rely on proxies), and may be subject to inference from other soil constituents (Viscarra Rossel and Lobsey, 2016).

Table 1 Commonly used proximal soil sensors in the field and their relative position on the electromagnetic spectrum.

EM range wavelength (m)		Sensor type											
		10^{-12}			10^{-10}		$10^{-8}-10^{-4}$		10^{-2}		10^1-10^6		
		Gamma-ray		X-ray	Optical		Microwave		Radio wave				
Classification	Method	TNM	Spectroscopy	XRF	Visible	NIR	MIR	TDR	FDR/Capacitance	GPR	NMR	EMI	EC/ER
	Energy ^a	A	A/P	A	A/P	A/P	A	A	A	A	A	A	A
	Interaction ^b	I	N	N	I/N	I/N	I/N	I	I	N	N	N	I
	Operation ^c	S/M	S/M	S	S/M	S/M	S	S	S/M	S/M	S/M	S/M	M
	Development status ^d	R/C	C	C	C	C/R	R	C	C	C	R	C	C
Soil property	Chemical properties												
	Organic matter content					X		X		X			
	Total Carbon	X		X		X							
	Organic carbon	X		X		X							
	Inorganic carbon	X				X							
	Nitrogen	X		X		X		X		X			
	Phosphorous	X		X		X							
	Potassium	X		X		X							
	Micronutrients, elements	X		X		X							
	Heavy metals	X		X		X							
	CEC	X				X				X			
	Soil pH	X				X		X		X			
	Salinity							X		X			
	Physical properties												
	Color			X									
	Water content	X		X		X		X		X			
	Soil texture (clay, silt, and sand)	X		X		X				X			
	Clay minerals	X		X		X							
	Bulk density	X		X		X				X			

Note: TNM: thermalized neutron methods; XRF: X-ray fluorescence; NIR: near infrared; MIR: mid infrared; TDR: time-domain reflectometry; FDR: frequency-domain reflectometry; GPR: ground-penetrating radar; NMR: nuclear magnetic resonance; EMI: electromagnetic induction; ER: electrical resistivity.

^aProximal sensors can be classified by their energy source (active sensors (A) provide their own source of energy, passive sensors (P) rely on ambient or emitted energy).

^bProximal sensors can be classified by their method of interaction with the soil (invasive sensors (I) rely on a direct contact with soil, non-invasive sensors (N) are operated without any soil distortion).

^cProximal sensors can be classified by their mode of operation (stationary operation (S) requires placing the sensor in a specific geographic location at a fixed or variable depth, mobile operation (M) allows on-the-go soil sensing).

^dProximal sensors can be classified by their development status (either remaining in the research phase (R) or has been commercialized (C)).

Adapted from Adamchuk VI, Hummel JW, Morgan MT, Upadhyaya SK (2004) On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture* **44**: 71–91. <https://doi.org/10.1016/j.compag.2004.03.002>; Viscarra Rossel RA, Adamchuk VI, Sudduth KA, McKenzie NJ, Lobsey C (2011) Proximal soil sensing: An effective approach for soil measurements in space and time. *Advances in Agronomy* **113**: 243–291. <https://doi.org/10.1016/B978-0-12-386473-4.00005-1>; Kuang B, Mahmood HS, Quraishi MZ, Hoogmoed WB, Mouazen AM, van Henten EJ (2012) Sensing Soil Properties in the Laboratory, in situ, and on-line: A review. *Advances in Agronomy*, pp. 155–223, Elsevier Inc. <https://doi.org/10.1016/B978-0-12-394275-3.00003-1>; Viscarra Rossel RA and Lobsey C (2016) *Scoping review of proximal soil sensors for grain growing*. 52. <https://doi.org/10.13140/RG.2.2.34785.51049>.

The most commonly used proximal soil sensors in research, applied, and commercial settings are electrical and EM sensors (e.g., EMI), X-ray, optical and radiometric sensors (e.g., Gamma-rays, Near and Mid-Infrared spectroscopy), and capacitance probes (e.g., soil moisture sensors) (Table 1). Others sensor types not covered in detail in this chapter include acoustic, pneumatic, and mechanical sensors.

Advantages of proximal sensing

Proximal soil sensors can facilitate the (near-instantaneous) collection of high-resolution spatial (or temporal) data using simpler, less time and labor-intensive techniques compared to conventional laboratory methods (Viscarra Rossel et al., 2011). While data from proximal sensors may not be as precise as laboratory measurements (Viscarra Rossel et al., 2011), users are offered more information than ever before about the soils condition to characterize the (horizontal and vertical) spatial and temporal variability of soil properties (Viscarra Rossel and Lobsey, 2016) with minimal disturbance.

Disadvantages of proximal sensing

Many proximal sensors are indirect measures that rely on an inference of the relationship between the sensor signal (a proxy) and the properties of interest for meaningful interpretation. These calibrations and empirical relationships are, in many cases, specific to certain soil and environmental conditions that may change over time (Adamchuk et al., 2011). Thus, the use of proximal sensors is restricted to the soil conditions it has been calibrated for. Complex interactions and interrelationships between multiple soil properties limits the number of attributes that can be measured with a single sensor, and multivariate calibrations to account for these interactions and the non-specificity of some proximal sensor outputs may be required (Stenberg and Viscarra Rossel, 2010). The performance of different sensors varies considerably due to different soil and environmental factors, including soil parent material, water content, and temperature (Kuang et al., 2012), and the deployment of proximal sensors is limited by access to the area of interest. This is a particular issue on waterlogged soils (e.g., rice paddies) or in densely forested areas.

Types of sensors and their applications

Proximal soil sensors range from simple handheld, stationary systems, through to more complex, autonomous robotic sensor platforms and are in various stages of development and commercialization (Gebbers, 2018). Systems available “off the shelf” can be purchased directly and may be ready to use, or may require some additional development or calibration before use. Others are in mature stages of research where additional testing for implementation is needed, or remain in the ‘proof of concept’ phase. Advancements in sensor capabilities, and information and communication technologies (e.g., telemetry systems and cloud-based services) have aided the development of PSS systems and are contributing to their increased adoption in-the-field.

There are issues and opportunities associated with on-the-go, stationary, or stop-and-go systems that cover a broad range of the EM spectrum (Table 1).

On-the-go sensors

Logistical constraints limit the number of samples that can be obtained via conventional sampling strategies or point-based proximal soil sensors. Improvements in the quality, accuracy, and adoption of global positioning systems (GPS) and geographical information systems (GIS) within agriculture and landscape management as well as ever increasing computing power has supported more intensive soil sampling strategies and the rise of spatial on-the-go sensing and mapping. On-the-go sensing technologies enable the collection of high-resolution soil information and can facilitate the interpretation of soil spatial patterns using a more time and cost-effective method compared to conventional analysis techniques.

On-the-go (mobile) sampling presents the “ultimate solution” (Adamchuk et al., 2018) for collecting large amounts of data for spatial interpolation and decision making. On-the-go proximal soil sensors are typically mounted to tractors, all-terrain vehicles, or other farm machinery and measurements are recorded while continuously moving across a field (Fig. 1). Most commercial on-the-go PSS tools measure soil properties indirectly from on or above the soil surface, although sensor operation varies for different systems. The information collected is often used to derive maps of soil properties, which are then used as covariates for DSM or spatial interpolation methods.

Electromagnetic induction (EMI)

Electromagnetic induction is a noninvasive, ex situ soil sampling method, and is one of the most widely-used on-the-go PSS tools worldwide. Electromagnetic induction sensors measure the apparent bulk electrical conductivity of the soil (EC_a) through the transmission and reflection of EM waves between two or more electric coils (solenoids) (Gebbers, 2018), effectively turning the soil into an EM circuit. Soil properties that (directly or indirectly) influence conductivity can be inferred from EC_a measurements (via sampling and calibration), and the most common soil properties that are correlated with EC_a are soil clay content (texture), soil moisture, and salinity (Table 2; Whelan and Taylor, 2013). Calibrations are site and time-specific, which limits the application of EMI to the soil conditions it has been calibrated for. Readings are influenced by the soil’s water and salt content, temperature, and



Fig. 1 On-the-go Dual EM-21S electromagnetic induction (EMI) meter (Dualem Inc., Ontario, Canada) and RSX-1 gamma-ray spectrometer (Radiation Solutions Inc., Ontario, Canada) can be used as part of a multi-sensor platform to map soil moisture content and salinity (via EMI), and soil type and parent material (via Gamma). Together, these complementary sensors can be used to map patterns of soil variability, alongside a high-precision GPS unit mounted to the vehicle.

Table 2 General impact of soil properties on soil apparent electrical conductivity (EC_a) readings.

	Material	Soil texture	Clay type	Moisture potential	CEC	Salinity	Nutrients
Higher EC _a	Soil	Clay	Smectite	Saturated	High CEC	High EC	High N, P, K, Ca, Mg, Na
↑	Stony soil	Silt	Illite	Field capacity			
Lower EC _a	Rock	Sand	Kaolin	Wilting point	Low CEC	Low EC	Low N, P, K, Ca, Mg, Na

Adapted from Whelan BM and Taylor J (2013) Precision agriculture for grain production systems. *International Journal of Digital Earth*. <https://doi.org/10.1080/17538947.2013.817183>.

mineral type and composition, making universal calibrations complex and difficult to obtain (Whelan and Taylor, 2013; Doolittle and Brevik, 2014).

Compared to conventional sampling, EMI is a non-invasive, relatively quick, easy-to-use, cost-effective and data-rich measure of soil variability (Doolittle and Brevik, 2014). When coupled with data logging and GPS technologies, a high density of spatial data can be gathered on-the-go by surveying the area along transects or swathes (Figs. 1 and 2). Data is logged at regular (~1–3 s) intervals from (typically) ground-based vehicles at (potentially) multiple depths down to ~3 or 6 m for the most commonly used EMIs in soil science applications (refer to Fig. 1 for an example) (Viscarra Rossel et al., 2011), with effective sensing depth largely

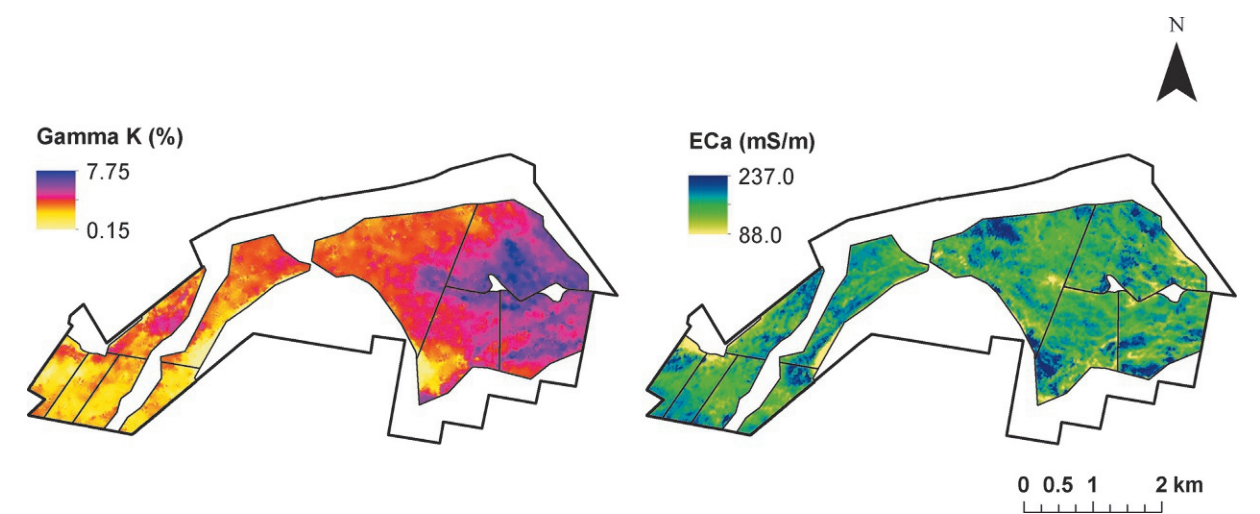


Fig. 2 Gamma radiometrics K ROI and Soil EC_a (0–3 m) maps derived from a RSX-1 gamma-ray spectrometer (Radiation Solutions Inc., Ontario, Canada) and a Dual EM-21S electromagnetic induction (EMI) meter (Dualem Inc., Ontario, Canada), respectively, across “L’lara”, a property close to Narrabri in North West NSW, Australia. Gamma radiometrics primarily relate to changes in parent material and soil type, with high Gamma K values related to soils with a low soil moisture potential (i.e., wilting point), high cation exchange capacity and K content, and a shallow depth to rock/gravel. EMI relates to soil texture, moisture, and salinity, and high EC_a values generally relate to soils with a high moisture potential (saturated), high clay (smectite) content with a high nutrient (N, P, K, Ca, Mg, Na) content and EC. Together, these on-the-go soil measurements can be used to describe soil variability across a landscape.

determined by coil spacing, orientation, frequency, and height above the ground (Gebbers, 2018). Numerous EMI sensors are commercially available for use in the field, including the DUALEM (Duaem, Inc., Milton, Ontario, Canada) and EM38 (Geonics Limited, Mississauga, Ontario, Canada) meters. Each commercial sensor has operational advantages and disadvantages depending on the depths and soil properties of interest, the preferred data collection mode, and soil condition at the time of sampling (Sudduth et al., 2003).

Gamma-rays

Mobile gamma-ray spectrometry (GRS) is a noninvasive, ex situ soil sampling method used to estimate the concentrations of naturally in the soil occurring radioelements (Potassium [^{40}K], Uranium [^{238}U], Thorium [^{232}Th], and their ratios) below the earth's surface (Reinhardt and Herrmann, 2019). This is done by measuring the gamma rays of radioactive isotopes emitted during radioactive decay. Passive and active gamma ray sensors are available, although passive gamma-ray sensors are preferred due to radiation and safety concerns associated with active sensors, which provide their own source of radiation (Adamchuk et al., 2018).

Simple gamma-ray sensors detect bulk radiation (total count), while others can distinguish different energy levels (Gebbers, 2018). Three key regions of interest (ROIs) in the gamma spectra (^{40}K , ^{238}U , and ^{232}Th) are commonly used to (directly and indirectly) infer information for understanding land forming processes and soil attribute distributions; primarily for parent material characterization and soil type differentiation (Table 3; Whelan and Taylor, 2013; Reinhardt and Herrmann, 2019). Direct, standalone total count readings may be sufficient for delineating basic patterns of soil variability, but a more detailed interpretation and relation to specific soil properties requires sophisticated data processing (Pätzold et al., 2020).

Gamma counts can be logged continuously on-the-go while driving over an area (Figs. 1 and 2). Approximately 90% of all aboveground gamma radiation originates from the upper 30–50 cm of the soil profile (Gebbers, 2018), although effective depth largely depends on bulk density and moisture content (Reinhardt and Herrmann, 2019). As for EMI, GRS enables the fast collection of dense spatial data and has a relatively well-established theoretical background (Gebbers, 2018). However, careful site- and time-specific calibrations are required to account for different geopedological conditions (Pätzold et al., 2020). The most common commercially available gamma-ray sensors for use in the field include the RS-X (Radiation Solutions Inc., Ontario, Canada) and the Medusa (Medusa Radiometrics BV, Groningen, The Netherlands) gamma-ray spectrometer series.

While EMI and Gamma-ray spectrometer readings can be used individually to create soil property maps (e.g., soil EC_a , cation exchange capacity (CEC), soil texture), these maps are often difficult for users to understand and are not useful for making management decisions. Instead, on-the-go EMI and Gamma-ray sensors are often used in combination to map variability across the soil landscape and diagnose potential constraints and limitations (Fig. 2). Gamma-ray spectrometer readings are largely impacted by shallow rocky materials, and total-count values are high in soils with a large clay (smectite), CEC and K contents, and a low soil moisture potential (i.e., wilting point). Soil EC_a readings (from EMI) are primarily related to soil texture, moisture, and salinity. High EC_a values relate to soils with high moisture potential (saturated), and large EC, clay (smectite), and nutrient (N, P, K, Ca, Mg, Na) content. The inverse response of these two soil measurements to soil moisture content, salinity status, and rocks/stones means that they can be used together to identify the causal factors of high EC_a readings (whether that be soil moisture, salinity, clay, or CEC), and identify patterns of variability. For example, low gamma counts are associated with sandy soils, and EMI can be used to determine if these sands are saline or not. These instruments are often used in tandem as part of a multi-sensor platform (Fig. 1).

COSMOS

Cosmic-ray Neutron Sensing (CRNS; Zreda et al., 2008) is a passive, non-invasive, indirect soil sensing tool used to estimate soil moisture based on the inverse relationship between the intensity of naturally occurring cosmic-ray generated neutrons, and the amount of water present in the surrounding environment (Desilets et al., 2010). While CRNS probes are used within agricultural fields and as part of larger soil moisture monitoring programs (i.e., the COsmic-ray Soil Moisture Observing System, COSMOS; Zreda et al., 2012), this stationary data does not capture the inherent spatial heterogeneity of soil moisture at a range of scales. A ground-based, mobile CRNS, called the cosmic-ray rover (Desilets et al., 2010), has been used for spatial soil moisture surveys over large areas (Fig. 3). However, there are few cosmic-ray rovers across the world, and they are not available commercially. As for all

Table 3 General impact of soil properties on soil gamma emission (total count; TC).

	Material	Soil texture	Clay type	Moisture potential	CEC	Depth to rock or gravel	Nutrients
Higher TC	Rock	Clay	Smectite	Wilting point	High CEC	Shallow	High K
	Stony soil	Silt	Illite	Field capacity			
Lower TC	Soil	Sand	Kaolin	Saturated	Low CEC	Deep	Low K

Adapted from Whelan BM and Taylor J (2013) Precision agriculture for grain production systems. *International Journal of Digital Earth*. <https://doi.org/10.1080/17538947.2013.817183>.

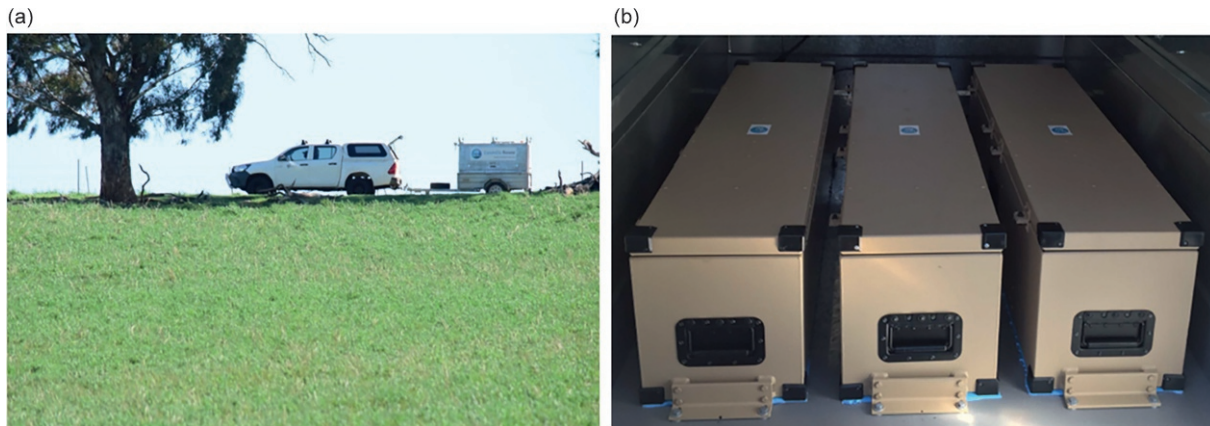


Fig. 3 Example of a mobile cosmic ray neutron sensing (CRNS) system, the 'CosmOz Rover'—CSIRO (<https://cosmoz.csiro.au/about>), operating in the field (a) and a view of the cosmic-ray neutron sensors inside the trailer. The CosmOz Rover has a counting rate equal to about 20 standard stationary cosmic-ray probes, which enables the collection of the spatial variation in neutron counting rates while driving the trailer through the landscape.

indirect soil sensing methods, site and time-specific calibrations are required to account for local conditions at the time of sampling. The calculation of volumetric soil moisture content requires information on the spatial variation of bulk density, soil organic matter, and lattice water (i.e., the amount of hydrogen held within the lattice structure of the soil minerals), which requires spatial data at an appropriate scale or additional spatial sampling to collect the necessary information. Vegetation cover and roads can also influence soil moisture readings from mobile cosmic-ray sensors (McJannet et al., 2017).

Time-domain reflectometry (TDR) and ground-penetrating radar (GPR)

Time-domain reflectometry (TDR) and ground-penetrating radar (GPR) are active sensors that operate based on the soil's dielectric constant (or the soil's ability to store electrical charge), and can be related to soil moisture. Both TDR and GPR rely on the same principle in that they measure the travel times and amplitudes of an EM pulse (Corwin, 2008; Gebbers, 2018). However, TDR uses two (or more) buried electrodes that are in direct contact with the soil, while GPR operates indirectly above the soil surface (Gebbers, 2018). TDR sensors are mainly used for stationary measurements, but prototype stop-and-go mobile TDR sensors have been developed (Thomsen et al., 2007) and more work is needed to further develop and commercialize continuous mobile TDR measurements (Gebbers, 2018). Mobile GPR sensors are commercially available and are predominately used for geophysical exploration and hydrology (Knight, 2001).

Soil compaction and strength

In addition to proximal soil sensors that operate within the EM spectrum, soil compaction, impedance or strength can be measured on-the-go or at a single point via mechanical approaches. Mechanical sensors either measure the amount of energy required to pull an implement through the soil (e.g., tine), or measure the resistance of the soil to the insertion of a probe (penetration resistance) (Adamchuk et al., 2018). Broadly, any soil tillage tool (e.g., plows, load cells) can be used to measure the spatial variability of soil strength and compaction at single or multiple depths, and are relatively inexpensive, robust, and easy-to-use. Acoustic and pneumatic sensors may also be used to assess soil structure and compaction.

Point-based sensors

Point-based, or stationary, proximal sensing systems can be operated in the field to make a single measurement, produce several measurements at different depths at a given location, or continuously monitor soil parameters when installed at a site for a period of time (Adamchuk et al., 2011). These may be handheld sensors, which can be moved from site-to-site (e.g., handheld infrared spectrometers), or be permanently fixed sensors, and must remain stationary during sample collection and analysis. Although point-based proximal sensors have a smaller spatial sampling density compared to on-the-go and stop-and-go systems, the ability to continuously monitor soil properties and understand their temporal dynamics (Adamchuk et al., 2018) means that they are an important component of proximal sensing repertoire for farmers and land managers.

Visible, near and mid-infrared diffuse reflectance spectroscopy

Diffuse reflectance infrared spectroscopy (DRS) is a proven method for rapid soil characterization under soil laboratory conditions (Stenberg et al., 2010) and with the development of field portable systems is slowly emerging for the rapid and cost-effective analysis of soil physical, chemical, and biological properties in field settings (Stockmann et al., 2018). Infrared spectroscopy is based on the reflection, absorbance, and scattering of light (photons) to measure soil properties, based on characteristic peaks or valleys at certain wavelengths in the visible (vis; 390–700 nm), near-infrared (NIR; 700–2500 nm), and mid-infrared (MIR;

2500–25,000 nm) parts of the electromagnetic spectrum. Spectral reference libraries, which are collections of known relationships between spectra and soil properties, are key to DRS. Calibration models are developed from these reference libraries to produce predictions of soil properties at varying levels of accuracy (Soriano-Disla et al., 2014). Additional pre-processing transformations may also be required to reduce noise, address the effects of field condition scanning such as soil moisture (e.g., external parameter orthogonalization, EPO; Minasny et al., 2011), and enhance more chemically-relevant peaks (Stenberg and Viscarra Rossel, 2010). Applications of DRS in research and commercial settings is increasing because it is a rapid, non-destructive, less labor-intensive, and more cost-effective alternative to conventional laboratory analyses alone. There is a growing number of portable, handheld NIR spectrometers being developed and made available commercially for soil sensing applications (Tang et al., 2020), including from Hone Carbon (<https://www.honecarbon.com/>) and Yard Stick (<https://www.useyardstick.com/>), which can both be used to measure soil carbon levels in the field (Fig. 4). However, developing representative and diverse spectral libraries for these technologies remains a challenge at present for widespread commercial uptake.

Portable X-ray fluorescence spectrometry

Portable X-ray fluorescence (PXRF) spectroscopy is a widely used, non-destructive, indirect proximal sensing tool used to determine the elemental composition of a sample, based on the excitation and release of energy from electrons. X-ray fluorescence spectrometers emit X-rays which are absorbed by electrons and shift them into an excited energy state, resulting in the release of X-ray energy which is referred to as X-ray fluorescence. This measured energy spectrum (fluorescence) contains peaks that are characteristic for each element, and this X-ray absorption can be used to identify and quantify different elements in a sample (Weindorf et al., 2014; Gebbers, 2018). Robust PXRF instruments can quantify up to 42 elements (Gebbers, 2018), but the detectable range is greater for more sophisticated instruments. XRF spectroscopy is an established laboratory method in earth sciences, and field-based, handheld PXRF instruments are increasingly becoming available at decreasing prices (Gebbers, 2018). Field PXRF spectrometers enable the rapid, relatively easy, *in-* and *ex situ* analysis of solid soil samples with minimal sample preparation compared to conventional laboratory analysis and are commonly used for land contamination assessments (Pozza et al., 2020). However, invasive site-preparation is still necessary to prepare soil pits or extract cores for scanning (Weindorf et al., 2014). As for other indirect methods, there are several sources of interference on PXRF signals mainly soil matrix heterogeneity, but also soil moisture content, and thus the degree of sample preparation (Weindorf et al., 2014), although PXRF is less impacted by soil moisture than is DRS.

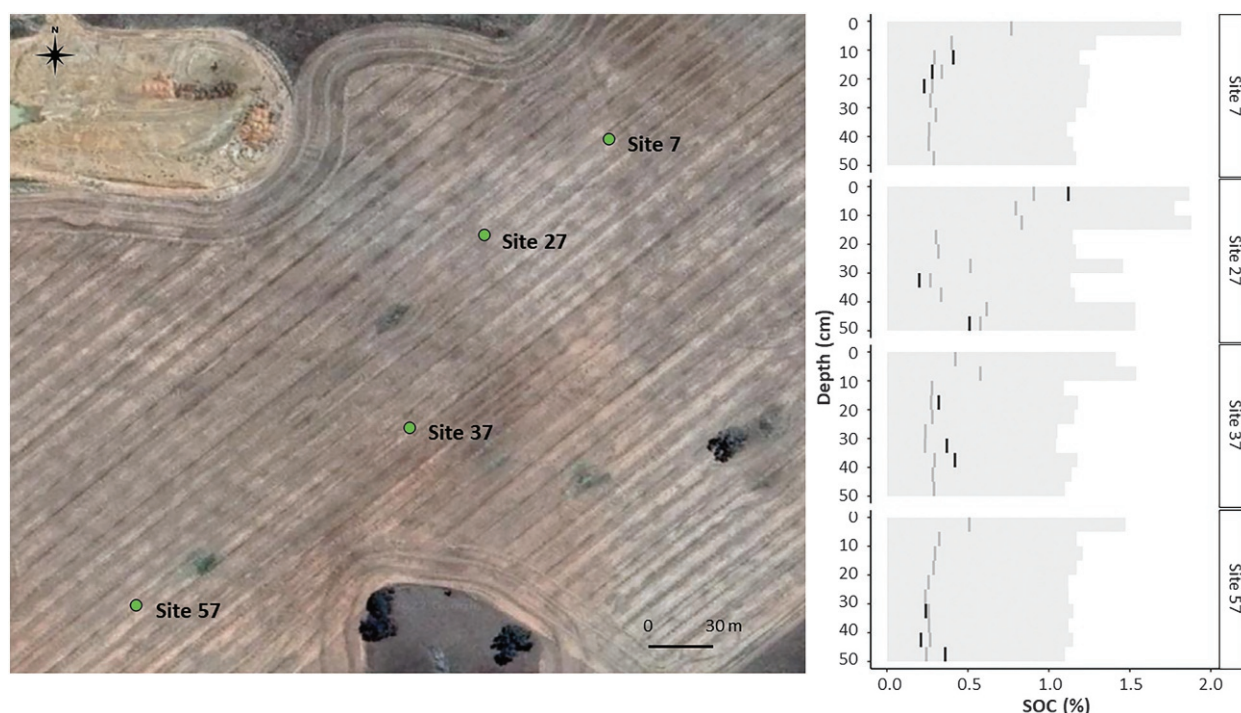


Fig. 4 Predictions of Soil Organic Carbon (SOC) % using the Hone Lab Red sensor from 0 to 50 cm at 5 cm increments (dark grey) in four locations across the CSIRO Boorowa Agricultural Research Station long-term trial site in NSW, Australia, using a local calibration model with 50 samples produced through an ensemble cubist model with 50 bags. The prediction interval (light grey) is also shown and observed SOC analytical values are presented in black. Adapted from Moloney J, Stockmann U, Karunaratne S, Malone B, Wheeler S, Brown C, Gray B, Howard S and Shiley D (2022) Australia's national soil spectral library empowering rapid soil organic carbon measurement. In: *Feasibility Study Report for the National Soil Carbon Innovation Challenge Feasibility Study Grants Project SCMICF000073*. CSIRO: Australia.

Soil moisture sensors

Proximal soil moisture sensors may directly or indirectly measure soil moisture content from above the soil surface or via direct contact. Stationary soil moisture sensors can be used to take single measurements, multi-depth measurements, or continuously monitor soil moisture at a given location. Linking numerous moisture sensors together via wireless networks can also be used to monitor the spatial and temporal dynamics of soil moisture (Adamchuk et al., 2018). Many soil moisture sensors are relatively mature technologies, and commonly used proximal soil moisture sensors in the field include capacitance probes and frequency domain reflectometry (FDR) sensors, matric potential sensors, active neutron probes and gamma-ray sensors, and passive cosmic-ray sensors.

Capacitance probes and FDR sensors are both active, invasive sensors that measure soil moisture content based on the soils dielectric constant by propagating an EM wave through the soil (i.e., using it as a dielectric medium). Matric potential sensors (e.g., tensiometers, gypsum blocks, and granular matrix sensors) measure the amount of force (suction) required to extract water from the soil (i.e., the soils matric potential), rather than the soil moisture content (Hardie, 2020). Capacitance, FDR, and matric potential sensors all require direct soil-to-sensor contact to obtain accurate readings, and the measurement area (i.e., the volume of inference around the sensor) is relatively small (Hardie, 2020).

Other field-based, stationary soil moisture sensors include active neutron probes and gamma-ray sensors, and passive cosmic-ray sensors. These noninvasive sensors all facilitate the fast, accurate, and continuous monitoring of soil moisture at a fixed point, and are particularly well-suited to hard-set, vertic (shrink-swell) soils where the installation of other soil moisture sensor types is difficult (Hardie, 2020). However, they are costly, require complex and precise calibration methods, and the measurement footprint of cosmic ray neutron sensors in particular (i.e., 200–600 m) means they are poorly suited to short-range precision agriculture treatments such as variable rate irrigation (Hardie, 2020). Further, application of active neutron and gamma-ray sensors in the field is limited due to radiation safety concerns.

Stop-and-go sensors

Stop-and-go sensors can be moved from place-to-place, but the sensor platform needs to remain stationary during sample collection. While sampling locations and intervals are typically defined by the user, the sample collection itself is often automated into the sensor platform once the vehicle has stopped. Stop-and-go platforms usually have a smaller sampling density compared to on-the-go systems but are greater than conventional sampling methods or point-based proximal sensors. Multiple sensors may be incorporated onto a single stop-and-go sensor platform for the collection of several measurements at once, and modular systems which enable the customization of sensor platforms are increasingly being adopted.

The uptake of single-vendor, modular sensing systems is increasing in applied and commercial settings. A range of on-the-go, stop-and-go, and point-based proximal soil sensors are commercially available and marketed by Veris® Technologies (Salina, USA; <https://veristech.com/>). Depending on the platform and its configuration, modular sensing systems like Veris® integrate a range of proximal sensors, such as EMI, infrared spectrometers, and penetrometers, to measure a range of soil properties in the field using a multi-sensor approach, including soil pH, EC, and soil organic matter. Sensors may be mounted on existing farm equipment, embedded within a complete sensing system, or be individual sensors that can be used in combination within a customized setup. For all Veris® sensors, sample uptake, analysis, and sensing is automated on-the-go or within a stop-and-go system. Single vendor, modular systems are widely used in research and commercial applications as they require little-to-no user calibration and configuration, provide a high-density of information to map soil properties, and are relatively easy to use and interpret.

Other automated stop-and-go proximal sensing systems are incorporating spectrometers into multi-sensor platforms. The Soil Condition Analysis System (SCANS) is an automated soil core sensing system which measures a range of soil properties (e.g., SOC, bulk density, pH, available water capacity) of an intact soil core under field condition using a complementary set of sensors covering different portions of the EM spectrum (i.e., a gamma-ray attenuation densitometer, digital camera, and a vis-NIR spectrometer) (Viscarra Rossel et al., 2017). For field deployment, soil cores are extracted and loaded manually into the SCANS platform which is mounted within a specialized field trailer, and the sample is scanned automatically along the core (Fig. 5). The modularity of SCANS can also facilitate the integration of other sensors (e.g., portable XRF and mid-IR spectrometers) as they become available (Viscarra Rossel et al., 2017).

Predicting soil properties

Calibration and spectral libraries

Many proximal sensors are indirect measures and parameter values cannot be directly deciphered from sensor signals or spectra. Instead, spectra need to be related to known reference samples, collectively known as a spectral library, through calibration of a prediction model (Stenberg and Viscarra Rossel, 2010). These reference samples must be representative of the range of soil types and conditions that the model and sensor is intended for. Commonly used calibration models include stepwise multiple linear regression, principal component analysis, partial least squares regression, and machine learning methods (e.g., regression trees). Also, additional pre-processing using mathematical functions is often used to correct for variation (e.g., normalizing spectra in respect to each other using pre-processing methods such as the Standard Normal Variate (SNV) method) and reduce noise (Stenberg and Viscarra Rossel, 2010).

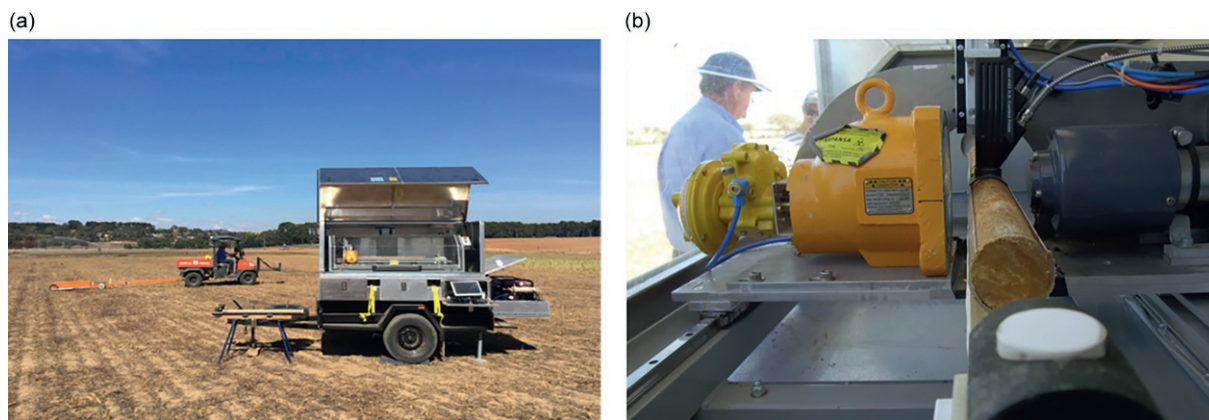


Fig. 5 (a) The trailer-mounted SCANS automated proximal sensing system at a field site. (b) The soil core sensing system collecting data down the length of a soil core through a gamma densitometer, vis-NIR spectrometer, and high-resolution RGB camera.

Spectral libraries and calibration models enable predictions on unknown samples using only spectral data, saving both time and money compared to conventional laboratory assessments. However, the construction of a representative and diverse soil spectral library still requires conventional soil sampling and analysis which is both site- and time-specific. While fewer calibrations are better in terms of logistics, cost, and error, there is a trade-off between the accuracy of this calibration data and generalization capacity (Stenberg and Viscarra Rossel, 2010). Further, sensor signals respond to multiple factors that depend on environmental conditions (e.g., soil moisture) and diverse reference samples or libraries are needed to account or correct for this variability.

Multi-sensor data fusion

No single sensor can measure all soil properties, and sensors can respond to more than one soil or landscape attribute. Using a selection of complementary sensing techniques and integrating this data, through multi-sensor data fusion or via a multi-sensor approach, presents an opportunity to provide more information about soil properties (Adamchuk et al., 2011).

Multi-sensor data fusion, or the use of two-or-more sensors and their subsequent data, can improve the quality and diversity of information that would otherwise not be available, or not be as accurate, from individual sensors (Adamchuk et al., 2011). Spectral reference libraries and calibration models are used to define relationships between sensor signals and soil attributes, including multivariate models or machine learning algorithms (Adamchuk et al., 2018). Different combinations of sensor inputs and models may be necessary depending on the target soil property, including the concatenation of spectra or model averaging of predictions from sensors across different areas of the EM spectrum (Pozza et al., 2020).

Multi-sensor approaches involve the use of multiple sensors in tandem for unique applications and output. Multi-sensor approaches can increase the robustness and validity of measurements, capture a broader range of soil attributes, and increase the density of data collected (Adamchuk et al., 2011) by using multiple sensors across the same area at the same time, often in a single pass for on-the-go or stop-and-go sensors. Within these multi-sensor platforms, sensor configuration may be complementary (i.e., information from two-or-more sensors can give a more complete image of the environment when used together), competitive (i.e., multiple sensors deliver independent measurements of the same property, often used to increase robustness or reduce error), or cooperative (i.e., use two-or-more sensors to derive information that would otherwise be unavailable from a single sensor as part of multi-sensor data fusion) (Durrant-whyte, 1990; Mitchell, 2012). Modular multi-sensor systems like Veris® (refer to section on Stop-and-go sensors) are an example of a multi-sensor approach that is used in commercial settings to gain a more complete understanding of the soil landscape.

Future directions of proximal sensing

In practice, proximal soil sensors are not a standalone tool. Instead, they form part of a complex decision-making process that draws information from a range of sources (Society of Precision Agriculture Australia (SPAA), 2017). Data integration from across multiple PSS systems and other on-farm (e.g., yield monitor data), remote sensing (e.g., LANDSAT, MODIS), and publicly available (e.g., ELVIS—Elevation and Depth—Foundation Spatial Data) data platforms can provide users with a thorough understanding of the soil landscape. Maps derived from PSS tools, such as soil gamma radiometric or EC_a maps, are often used as covariates for design-based soil sampling, DSM and spatial interpolation methods, alongside this additional auxiliary information. The acquisition, integration, and presentation of information from various sources can aid in the assessment of soils at a range of spatial and temporal scales, and can help to assess the uncertainty of proximal sensing predictions. Beyond this, PSS tools can also be used to guide future soil sampling strategies, such as delineating agricultural management zones or areas of land contamination (Society of

Precision Agriculture Australia (SPAA), 2017). However, further research on the complementary use of multi-sensor information and spatial (up and down) scaling of this data is needed (Viscarra Rossel and Lobsey, 2016).

For several soil attributes, appropriate proximal sensors remain in the research phase. However, greater commercialization potential is being realized as the cost, accuracy, and ease of use of these sensors continues to improve. To increase the uptake and application of PSS tools, better access to commercialized and integrated PSS systems and services is required (Viscarra Rossel and Lobsey, 2016). Further, integrating this information with other on-farm, industry, and publicly available data to enable meaningful interpretation needs to be supported. It should be emphasized that investment, research, and development into new and improved PSS tools should not be for the sole purpose of collecting more data. These tools should be used to collect information that is relevant to farmers, land managers, policy makers and other stakeholders that need data to support well-informed management and policy decisions. The translation of existing information into accessible tools for all end-users is an important first step to better understanding the soil landscape.

References

- Adamchuk VI, Hummel JW, Morgan MT, and Upadhyaya SK (2004) On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture* 44: 71–91. <https://doi.org/10.1016/j.compag.2004.03.002>.
- Adamchuk VI, Viscarra Rossel RA, Viscarra Rossel RA, Sudduth KA, and Lammers PS (2011) Sensor fusion for precision agriculture. In: *Sensor Fusion - Foundation and Applications*. BoD – Books on Demand. <https://doi.org/10.5772/19983>.
- Adamchuk V, Ji W, Viscarra Rossel R, Gebbers R, and Tremblay N (2018) Proximal soil and plant sensing. In: *Precision Agriculture Basics*, pp. 119–140. John Wiley & Sons. <https://doi.org/10.2134/precisionagbasics.2016.0093>.
- Corvin DL (2008) Past, present, and future trends of soil electrical conductivity measurement using geophysical methods. In: *Handbook of Agricultural Geophysics*, pp. 17–44. CRC Press. <https://doi.org/10.1201/9781420019353-8>.
- Desilets D, Zreda M, and Ferré TPA (2010) Nature's neutron probe: Land surface hydrology at an elusive scale with cosmic rays. *Water Resources Research* 46: 1–7. <https://doi.org/10.1029/2009WR008726>.
- Doolittle JA and Brevik EC (2014) The use of electromagnetic induction techniques in soils studies. *Geoderma* 223–225: 33–45. <https://doi.org/10.1016/j.geoderma.2014.01.027>.
- Durrant-whyte HF (1990) Sensor models and multisensor integration. *The International Journal of Robotics Research* 7: 73–89.
- Gebbers R (2018) Proximal soil surveying and monitoring techniques. In: Stafford JR (ed.) *Precision Agriculture and Sustainability*, pp. 29–78. London: Springer.
- Hardie M (2020) Review of novel and emerging proximal soil moisture sensors for use in agriculture. *Sensors* 20: 1–23. <https://doi.org/10.3390/s20236934>.
- Knight R (2001) *Ground Penetrating Radar for Environmental Applications*. Annual Review.
- Kuang B, Mahmood HS, Quraishi MZ, Hoogmoed WB, Mouazen AM, and van Henten EJ (2012) Sensing Soil Properties in the Laboratory, in situ, and on-line: A review. In: *Advances in Agronomy*, pp. 155–223. Elsevier Inc. <https://doi.org/10.1016/B978-0-12-394275-3.00003-1>.
- McJannet D, Hawdon A, Baker B, Renzullo L, and Searle R (2017) Multiscale soil moisture estimates using static and roving cosmic-ray soil moisture sensors. *Hydrology and Earth System Sciences* 21: 6049–6067. <https://doi.org/10.5194/hess-21-6049-2017>.
- Minasny B, McBratney AB, Bellon-Maurel V, Roger JM, Gobrecht A, Ferrand L, and Joalland S (2011) Removing the effect of soil moisture from NIR diffuse reflectance spectra for the prediction of soil organic carbon. *Geoderma* 167–168: 118–124. <https://doi.org/10.1016/j.geoderma.2011.09.008>.
- Mitchell HB (2012) *Data Fusion: Concepts and Ideas*. Berlin Heidelberg: Springer. <https://doi.org/10.1007/978-3-642-27222-6>.
- Pätzold S, Leenen M, and Heggemann TW (2020) Proximal mobile gamma spectrometry as tool for precision farming and field experimentation. *Soil Systems* 4: 1–23. <https://doi.org/10.3390/soilsystems4020031>.
- Pozza LE, Bishop TFA, Stockmann U, and Birch GF (2020) Integration of vis-NIR and pXRF spectroscopy for rapid measurement of soil lead concentrations. *Soil Research* 58: 247–257. <https://doi.org/10.1071/SR19174>.
- Reinhardt N and Herrmann L (2019) Gamma-ray spectrometry as versatile tool in soil science: A critical review. *Journal of Plant Nutrition and Soil Science* 182: 9–27. <https://doi.org/10.1002/jpln.201700447>.
- Society of Precision Agriculture Australia (SPAA) (2017) *Proximal Soil Sensing*. Adelaide, Australia: Society of Precision Agriculture Australia. <https://doi.org/10.2136/vzj2011.0105br>.
- Soriano-Disla JM, Janik LJ, Viscarra Rossel RA, MacDonald LM, and McLaughlin MJ (2014) The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied Spectroscopy Reviews* 49: 139–186. <https://doi.org/10.1080/05704928.2013.811081>.
- Stenberg B and Viscarra Rossel RA (2010) Diffuse reflectance spectroscopy for high-resolution soil sensing. In: Viscarra Rossel RA, McBratney AB, and Minasny B (eds.) *Proximal Soilsens*, pp. 29–47. Springer.
- Stenberg B, Viscarra Rossel RA, Mouazen AM, and Wetterlind J (2010) Visible and near infrared spectroscopy in soil science. In: Sparks DL (ed.) *Advances in Agronomy*, pp. 163–215. Elsevier Inc. [https://doi.org/10.1016/S0065-2113\(10\)07005-7](https://doi.org/10.1016/S0065-2113(10)07005-7).
- Stockmann U, Jones EJ, Odeh IOA, and McBratney AB (2018) Pedometric treatment of soil attributes. In: McBratney AB, Minasny B, and Stockmann U (eds.) *Pedometrics (Progress in Soil Science)*, pp. 115–153. Springer: Cham. https://doi.org/10.1007/978-3-319-63439-5_5.
- Sudduth KA, Kitchen NR, Bollero GA, Bullock DG, and Wiebold WJ (2003) Comparison of electromagnetic induction and direct sensing of soil electrical conductivity. *Agronomy Journal* 95: 472–482. <https://doi.org/10.2134/agronj2003.4720>.
- Tang Y, Jones E, and Minasny B (2020) Evaluating low-cost portable near infrared sensors for rapid analysis of soils from South Eastern Australia. *Geoderma Regional* 20: e00240. <https://doi.org/10.1016/j.geodrs.2019.e00240>.
- Thomsen A, Schelde K, Dröschner P, and Steffensen F (2007) Mobile TDR for geo-referenced measurement of soil water content and electrical conductivity. *Precision Agriculture* 8: 213–223. <https://doi.org/10.1007/s11119-007-9041-1>.
- Viscarra Rossel RA and Lobsey C (2016) *Scoping review of proximal soil sensors for grain growing*. 52. <https://doi.org/10.13140/RG.2.2.34785.51049>.
- Viscarra Rossel RA and McBratney AB (1998) Laboratory evaluation of a proximal sensing technique for simultaneous measurement of soil clay and water content. *Geoderma* 85: 19–39.
- Viscarra Rossel RA, McBratney AB, and Minasny B (2010) In: Viscarra Rossel R, Mcbratney AB, and Minasny B (eds.) *Proximal Soil Sensing*. Springer. <https://doi.org/10.2136/vzj2011.0105br>.
- Viscarra Rossel RA, Adamchuk VI, Sudduth KA, McKenzie NJ, and Lobsey C (2011) Proximal soil sensing: An effective approach for soil measurements in space and time. *Advances in Agronomy* 113: 243–291. <https://doi.org/10.1016/B978-0-12-386473-4.00005-1>.
- Viscarra Rossel RA, Lobsey CR, Sharman C, Flick P, and McLachlan G (2017) Novel proximal sensing for monitoring soil organic C stocks and condition. *Environmental Science and Technology* 51: 5630–5641. <https://doi.org/10.1021/acs.est.7b00889>.
- Weindorf DC, Bakr N, and Zhu Y (2014) *Advances in Portable X-Ray Fluorescence (PXRF) for Environmental, Pedological, and Agronomic Applications*. Elsevier. <https://doi.org/10.1016/B978-0-12-802139-2.00001-9>.

- Whelan BM and Taylor J (2013) Precision agriculture for grain production systems. *International Journal of Digital Earth*. <https://doi.org/10.1080/17538947.2013.817183>.
- Zreda M, Desilets D, Ferré TPA, and Scott RL (2008) Measuring soil moisture content non-invasively at intermediate spatial scale using cosmic-ray neutrons. *Geophysical Research Letters* 35. <https://doi.org/10.1029/2008GL035655>.
- Zreda M, Shuttleworth WJ, Zeng X, Zweck C, Desilets D, Franz T, and Rosolem R (2012) COSMOS: The cosmic-ray soil moisture observing system. *Hydrology and Earth System Sciences* 16: 4079–4099. <https://doi.org/10.5194/hess-16-4079-2012>.

Proximal sensing of soil moisture

Markus Tuller^a, Ebrahim Babaeian^b, and Scott B Jones^c, ^aDepartment of Environmental Science, The University of Arizona, Tucson, AZ, United States; ^bDepartment of Soil, Water, and Ecosystem Sciences, University of Florida, Gainesville, FL, United States; ^cDepartment of Plants, Soils, and Climate, Utah State University, Logan, UT, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	591
Definition of soil moisture	592
Proximal sensing	593
Electromagnetic soil moisture sensors	593
Neutron scattering	594
Cosmic-ray neutron probe	595
Ground penetrating radar	595
Electromagnetic induction	596
Nuclear magnetic resonance imaging	597
Microwave radiometry	598
Summary	598
References	599

Abstract

Soil moisture (SM) is of paramount importance for society and the global environment. Accurate SM information is at the core of a plethora of applications that include forecasting of weather and climate variability, projection and monitoring of drought conditions, precise agricultural irrigation management, conservation of water resources, monitoring of ecosystem response to climate change, and impact mitigation of natural disasters such as wildfires, landslides, floods, or dust storms, whose occurrence is intimately connected to the moisture status of the land surface. Variations in SM can have substantial impacts on ecosystem health, agricultural productivity, and forestry. In this chapter we present state-of-the-art and novel technologies and sensors for proximal sensing of SM together with the underlying concepts and theories.

Key points

- State-of-the-art and novel technologies and sensors for proximal sensing of soil moisture (SM) are presented.
- The underlying concepts and theories for proximal SM sensing are discussed.

Introduction

Soil moisture (SM) is a key hydrologic state variable that is of momentous importance for numerous applications in the agricultural and environmental sciences. The importance of SM is emphasized by the Global Climate Observing System (GCOS) (<https://gcos.wmo.int/>) Terrestrial Observation Panel for Climate (TOPC), who designates SM as an Essential Climate Variable (ECV) of the hydrosphere. Soil moisture is a governing variable with respect to partitioning the solar radiation impinging the land surface into latent and sensible heat fluxes, and separating precipitation into infiltration, runoff, and subsurface flow. SM controls the feedbacks between soil, vegetation, and climate and determines the prevailing vegetation cover and distribution within a region. Detailed knowledge about the state of SM and its spatial and temporal dynamics is of crucial importance for crop production, not only to prevent water stress and sustain yields, but also for the mitigation of adverse environmental impacts due to over-irrigation as well as for the conservation of water resources. The latter is especially relevant in view of the water crisis and prolonged drought conditions in many arid and semiarid regions of the world that is expected to escalate further due to the rapidly increasing human population and the changing global climate. Another area of application is the mitigation of the impact of natural disasters such as wildfires, landslides, floods, or dust storms. The potential of their occurrence is intimately connected to the SM status of the land surface. For example, when the SM diminishes below a certain threshold (i.e., the soil surface dries out) on land with sparse vegetation and the atmospheric conditions are favorable (i.e., strong winds and low relative humidity), a dust storm may develop. The frequency of dust storms in the Southwestern United States is increasing, causing chronic breathing and cardiovascular problems for humans, and exposing both humans and animals to fungal spores that can lead to potentially fatal valley fever (coccidioidomycosis). Today, a growing number of proximal techniques are available to measure and monitor the spatial and temporal SM variability. At the pedon scale, electromagnetic soil moisture sensors based on travel time-, capacitance-, or impedance-based measurement principles are used to record changes in SM. The latest generation of electromagnetic sensors incorporate microelectronic measurement

circuitry together with a microprocessor directly in the sensor head and communicate processed data to a datalogger. The cosmic-ray neutron probe (CRNP) is another promising invention that infers soil moisture for an approximately 300-m circular footprint from counting neutrons generated when naturally-occurring cosmic rays interact with hydrogen atoms in the topsoil. There are also several proximal geophysical sensing methods such as ground penetrating radar (GPR), electromagnetic induction (EMI), and nuclear magnetic resonance imaging (NMRI) deployed for field scale measurements of SM dynamics. Moreover, microwave radiometry has been successfully employed for proximal SM determination.

Definition of soil moisture

The SM or soil water content, which may be expressed in gravimetric (θ_m) or volumetric (θ_v) terms, represents the amount of water present in the soil at a given matric potential ψ_m . The ψ_m is synonymous with the combined capillary and adsorptive surface forces that hold water within the solid soil matrix and are uniquely related to SM under hydrostatic conditions. The decidedly nonlinear relationship between SM and ψ_m is termed soil water characteristic (SWC) curve or soil water retention (SWR) curve, which shows a very distinctive shape for each individual soil texture (Fig. 1).

The gravimetric SM, θ_m (kg kg^{-1}) is the ratio of the mass of water within a soil sample and the mass of the oven-dry solid material. The volumetric SM, θ_v ($\text{m}^3 \text{m}^{-3}$), defined as the volume of water within a given soil volume may be expressed in terms of θ_m as (Babaeian et al., 2019):

$$\theta_v = \theta_m \left(\frac{\rho_b}{\rho_w} \right) \quad (1)$$

where ρ_b is the dry bulk density (kg m^{-3}) of the soil, and ρ_w is the density of water (kg m^{-3}). When all soil pores are filled with water, θ_v is named saturated water content θ_s (Fig. 1). Sometimes, SM is expressed in terms of relative saturation, $S_e = \theta_v / \theta_s$, which is the volumetric SM content normalized to θ_s . Conceptually, the range for S_e is from zero, when the soil is entirely dry, to one, when all pores are filled with water. However, it is impossible in nature for soils to exhibit complete dry or saturated conditions. There is always residual moisture θ_r (Fig. 1) in soil pores under dry conditions and it is not possible to completely de-air soil because cavitation nuclei are tightly held in crevices of rough particle surfaces and air bubbles remain entrapped in dead-end pores. To account for θ_r , S_e may be expressed as:

$$S_e = \frac{\theta_v - \theta_r}{\theta_s - \theta_r} \quad (2)$$

The maximum S_e range is from approximately 0.94 to 0.97 (–) for fine-textured (clay) and coarse-textured (sand) soils, respectively.

It is convenient to express SM in length units for water balance applications to be consistent with other variables such as precipitation or evapotranspiration. This so-called equivalent depth of wetting may be obtained by multiplying θ_v with the depth of the soil layer of interest.

For agricultural irrigation management, a plant available water content (θ_{PAW}) is often defined as the difference between the water content at field capacity (θ_{FC}) and the water content at the permanent wilting point (θ_{PWP}) (Fig. 1). The θ_{FC} is synonymous

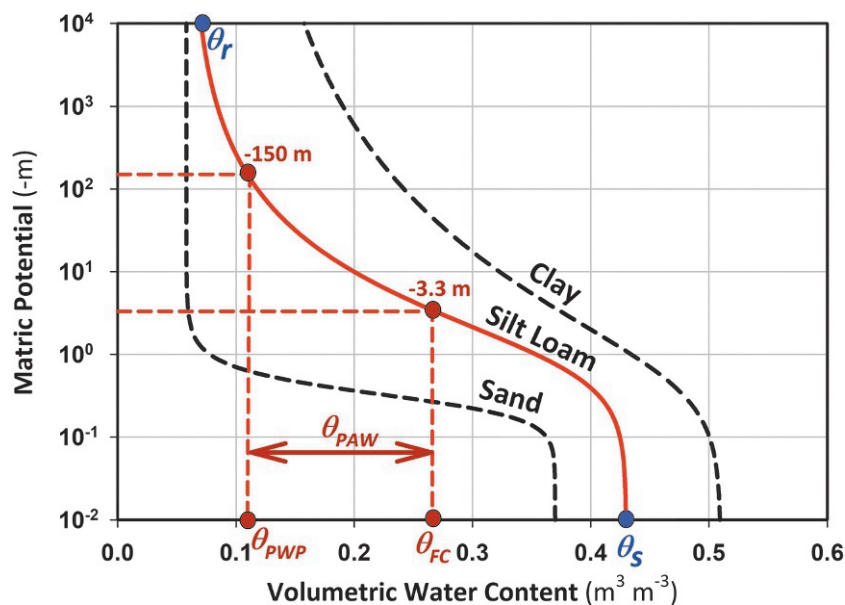


Fig. 1 The SWC curve shown for different soil textures. The markings pertain to the silt loam soil curve.

with the water content after internal redistribution of water within the soil matrix due to gravity (free drainage) and for practical purposes is often assumed to coincide with a matric potential of -3.3 m. This definition is not entirely correct as θ_{FC} depends on soil texture. The θ_{FC} for a sandy soil, for example, more likely corresponds to $\psi_m = -1$ m. Below the θ_{PWP} that is defined as the water content at $\psi_m = -150$ m, water is so tightly bound within the soil matrix that plants are no longer able to recover their turgidity and irreversibly wilt. Again, this is only an approximation, because the permanent wilting point depends on the physiology of plants. Desert plants, for example, can withstand significantly lower matric potentials (drier conditions).

Proximal sensing

Proximal sensing encompasses the acquisition of signals from the soil with sensors that are either in direct contact with or in close proximity to the soil (Viscarra Rossel et al., 2011). Proximal sensors may be further categorized based on the mode of deployment (i.e., stationary, or mobile), their source of energy (i.e., active or passive), the method of installation (i.e., invasive, or noninvasive), and based on how the soil property of interest is determined from the measured variable (i.e., direct, or indirect). An electromagnetic Time Domain Reflectometry (TDR) SM sensor, for example, is considered as invasive, stationary, active, and indirect—TDR infers θ_v from the soil's dielectric permittivity (Viscarra Rossel et al., 2011).

Electromagnetic soil moisture sensors

As early as in the 1960s electromagnetic-based instruments with 30 MHz frequency have been employed for soil moisture determination using fringe capacitance (Thomas, 1966). Soil water content sensing was truly revolutionized by Topp et al. (1980) with the introduction of time domain reflectometry (TDR). Early TDR systems were costly and initially only used in laboratory settings. Forty plus years later, the costs of TDR sensors have decreased by several orders of magnitude. At the same time, capacitance- or impedance-based circuits have also become popular for water content measurements. As a result, numerous electromagnetic (EM) sensors (Fig. 2) that use the permittivity of soil for water content determination are available today (Jones et al., 2005; Vaz et al., 2013). The permittivity or dielectric constant is the electrical potential energy of a porous material (e.g., soil) under the influence of an electromagnetic field. The sensor's electromagnetic field is directed into the soil along 2-, 3-, or 4-parallel rods or into adjacent pairs of rings. The HydraProbe (Stevens Water Monitoring Systems Inc., Portland, OR, USA), one of the most widely used EM sensors in the United States, consists of one central and three outer electrodes (Fig. 2A), and relies on a 50 MHz impedance circuit. It is deployed by U.S. federal agencies for weather and snow monitoring stations such as the Soil Climate Analysis Network (SCAN), or SNOWpack TElemetry (SNOTEL). The HydraProbe outputs both the real and imaginary parts of the complex permittivity. Considering that additional soil information may be obtained from the complex dielectric permittivity (Kellens et al., 2005), it comes as a surprise that not more effort is being invested to fully capitalize on such measurements. Recent studies aimed at obtaining dielectric spectra from the frequency domain have applied both open-ended as well as electrode-based probes using high-end as well as low-cost vector network analyzers (González-Teruel et al., 2022; Szyplowska et al., 2019). These advances provide more accurate means to determine SM and other soil properties (e.g., clay content and mineralogy) from spectral permittivity. Another impedance SM probe that operates at 100 MHz, but does not output complex permittivity, is the WET150 sensor (Delta-T Devices Ltd., Cambridge, UK) depicted in Fig. 2B.

Accurate EM-based determination of SM is feasible because of the large disparity between the dielectric permittivities of water (i.e., 80), solid soil constituents (i.e., about 4), and air (i.e., 1). The most accurate EM sensors generally operate at higher frequencies (~ 1 GHz) and measure travel time. This, in most instances, does not require soil specific calibration because soil permittivities at these higher frequencies are less susceptible to alteration from electrical conductivity, temperature and relaxation processes associated with high clay or organic matter contents (Szyplowska et al., 2019). For time-domain measurements, the electromagnetic signal travel time (t) is proportional to the relative permittivity (K_r) of the soil through which it travels:

$$t = \frac{2l\sqrt{K_r}}{c} \quad (3)$$

where l is the electrode length and c is the speed of light (3×10^8 m s⁻¹). Once the travel time is measured, K_r is computed to obtain a relationship with the volumetric moisture content (θ_v). A general volume-based mixing model (Jones et al., 2002; Roth et al., 1990) may be applied to infer θ_v :

$$\theta_v = \frac{K_r^\beta - (1 - \varphi)K_s^\beta - \varphi K_a^\beta}{K_w^\beta - K_a^\beta} \quad (4)$$

where φ is the soil's porosity and subscripts s , a and w refer to permittivities of solids (minerals), air, and water, respectively. The exponent β represents the geometry (layering) of the soil with values ranging between -1 and 1 , where $\beta = 0.5$ is applied for an isotropic mixture. Recently, GHz frequency TDR probes (Fig. 2C and D) such as the Digital True TDR-305 N sensor (Acclima, Inc., Meridian, ID, USA), and the SoilVUE™10 profile sensor (Campbell Scientific, Inc., Logan, UT, USA) have been developed using low-cost components, making them competitively priced relative to lower frequency EM sensors (e.g., HydraProbe and WET150).

All EM sensor measurements experience restricted applicability at high salinity levels due to signal attenuation. Severe attenuation of the EM signal may alter or prevent water content determination altogether due to very saline or clayey soil conditions.

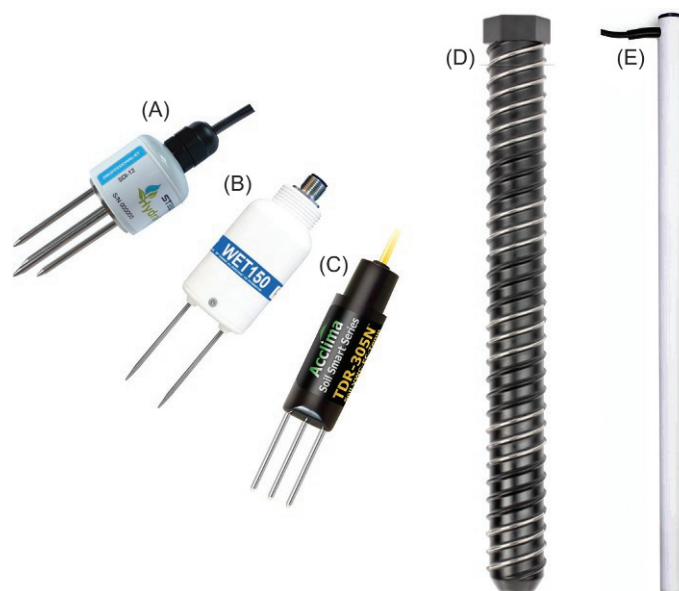


Fig. 2 Commercial electromagnetic water content sensors. (A) HydraProbe (Stevens Water Monitoring Systems, Inc., Portland, OR, USA), (B) WET150 sensor (Delta-T Devices Ltd., Cambridge, UK), (C) Digital True TDR-305 N sensor (Acclima, Inc., Meridian, ID, USA), (D) SoilVUE™10 profile sensor (Campbell Scientific, Inc., Logan, UT, USA), and (E) Drill and Drop SM sensor (Sentek Sensor Technologies, Stepney, AUS). Sensor sizes are scaled to fit side by side—the actual lengths vary.

Furthermore, soil specific calibrations are more likely to be required in soils with high clay or organic matter (OM) contents. Permittivity of drier soils with an abundance of 2:1 clay minerals (i.e., large surface area) is more susceptible to temperature fluctuations caused by the transition of bound water to a higher permittivity free water state. Permittivity is commonly lower in OM rich soils with large specific surface areas when compared with mineral soils at the same θ_v . A further limitation of EM sensors is the determination of the total water content of frozen soils due to the low permittivity of ice (Jagdhuber et al., 2014).

In recent years, borehole or downhole sensors have been developed to determine the profile SM along the length of the sensors. These sensors include the SoilVUE™10 profile sensor (Campbell Scientific, Inc., Logan, UT, USA), and the Drill and Drop SM sensor (Sentek Sensor Technologies, Stepney, AUS) depicted in Fig. 2D and E. The SoilVUE™10 sensor is available in 0.55 or 1.05 m lengths and outputs volumetric water content θ_v for 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5 m depths—the longer version additionally provides θ_v for 0.6, 0.75, and 1.0 m depths. The Drill and Drop sensor is available in lengths from 0.3 to 1.2 m and outputs θ_v every 0.1 m with the option of an integrated Bluetooth antenna that enables data readout with a mobile phone.

The Serial Digital Interface 1200 baud (SDI-12) communication protocol continues to gain popularity for environmental sensors and instruments given the capability to easily multiplex dozens of sensors from a single digital datalogger port. Sensor addressing may use single digit numbers and both upper- and lower-case letters, allowing more than 60 addresses per port on a given data logging device. Most EM sensors support SDI-12 communication.

Neutron scattering

Portable neutron probes (NP) were first deployed for measuring soil moisture (SM) in the 1950s. The concept behind the neutron scattering method is founded on the propensity of hydrogen atoms to thermalize (slow down) high-energy (2–4 MeV) neutrons emitted from a radioactive source to approach the characteristic speed of particles at ambient temperature with a corresponding energy of approximately 0.03 eV. A neutron probe is typically comprised of a radioactive neutron source (e.g., americium-241 and beryllium) and a detector (Fig. 3) to determine the flux of thermalized neutrons that form a cloud of nearly constant density near the probe. Neutrons lose various quantities of energy when colliding with different atoms. The highest energy loss is because of collisions with atoms of comparable mass, such as hydrogen. Because the quantity of hydrogen in the soil is mainly dependent on water content, a calibration equation relating the number of detected slow neutrons and soil water content can be established. To prevent entanglement with plant roots or the borehole wall, the radioactive source is commonly lowered into the soil to the desired measurement depth via an access tube (Fig. 3) made of PVC or aluminum. For accurate SM measurements calibration for each soil type and access tube material is essential. The measurement volume (sphere of influence) of a NP depends on the water content and chemical composition of the soil as well as on the strength of the radioactive source (Fig. 3). In general, reliable SM measurements can only be obtained below a depth of about 0.2 m, because close to the soil surface some of the neutrons escape into the atmosphere. To allow SM estimation closer to the soil surface, correction factors may be established, or the probe geometry modified (Hanna and Siam, 1980). Further details about the theoretical and practical aspects of the neutron scattering method are available in Sadeghi et al. (2018).

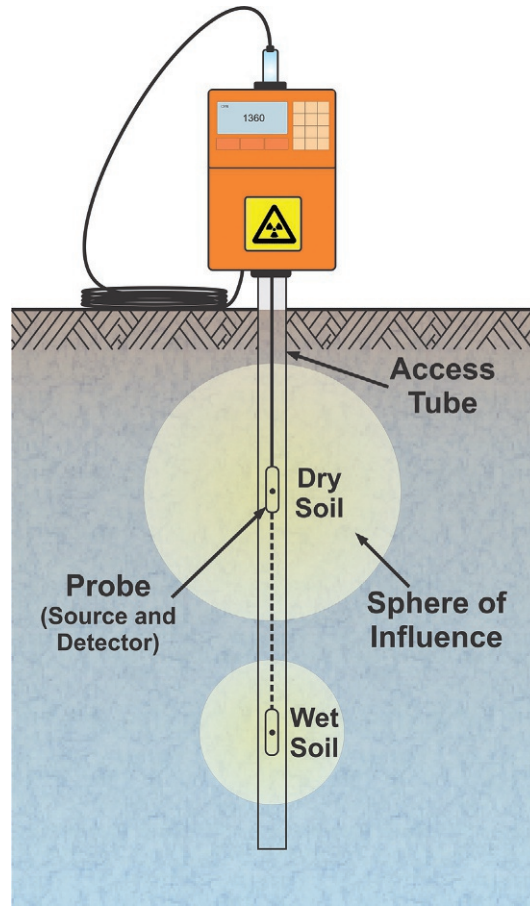


Fig. 3 Sketch of a neutron probe.

Cosmic-ray neutron probe

The cosmic-ray neutron probe (CRNP) (Fig. 4) estimates the area-averaged soil moisture (SM) by counting the neutrons within the topsoil and air that are generated by cosmic rays (Zreda et al., 2008). Similar to the standard NP, a calibration function can be established to relate the change in the number of low-energy neutrons to the change of water content. A theoretical calibration function that relates the volumetric water content (θ_v) and the relative neutron count (N/N_0) has been developed by Desilets et al. (2010):

$$\theta_v(N) = \frac{0.0808}{\left(\frac{N}{N_0}\right) - 0.372} - 0.115 \quad (5)$$

where N is the neutron count rate (count h^{-1}) normalized to a reference atmospheric pressure and solar activity level, and N_0 is the count rate (count h^{-1}) over dry soil under the same reference conditions. A detailed description of the cosmic-ray technique is provided in Zreda et al. (2012) and Andreasen et al. (2017). Because cosmic rays propagate further in thin air, the extent of the lateral measurement footprint (up to 300 m in diameter) is inversely proportional to the vapor pressure of the atmosphere. The vertical penetration depth depends on the SM content and ranges from about 15 cm in wet soils to approximately 70 cm in dry soils (Zreda et al., 2008).

A limitation of the CRNP is its sensitivity to hydrogen sources other than SM (e.g., atmospheric water vapor, organic matter, or root biomass), which may yield θ_v values that surpass soil porosity. This is why additional adjustments are necessary to eliminate the artifacts related to these additional sources as, for example, shown in Bogena et al. (2013).

Ground penetrating radar

The ground penetrating radar (GPR) is a geophysical proximal sensing technique applicable for two- and three-dimensional mapping of soil moisture (SM). A transmitter antenna emits high frequency (50 MHz to 3.6 GHz) electromagnetic (EM) waves



Fig. 4 CRNP setup in a creosote dominated shrubland in the Santa Rita Mountains southeast of Tucson, Arizona, USA. Photo credit: Dr. Trenton Franz, University of Nebraska, Lincoln.

into the subsurface. The waves are refracted, reflected, and scattered and sensed by a single or multiple receiver antennas (Klotzsche et al., 2018). Similar to time domain reflectometry (TDR), the travel time of the EM wave between the transmitter and receiver antennas is measured and the bulk dielectric permittivity of the soil (ϵ_b) that is mainly governed by the volumetric water content (θ_v) is inferred. Dielectric mixing models (e.g., Eq. 4) (Anbazhagan et al., 2020) or empirical relationships are then used to convert ϵ_b to θ_v . The most widely applied empirical relationship between θ_v and ϵ_b was developed for mineral soils by Topp et al. (1980):

$$\theta_v = -5.3 \times 10^{-2} + 2.92 \times 10^{-2} \epsilon_b - 5.55 \times 10^{-4} \epsilon_b^2 + 4.3 \times 10^{-6} \epsilon_b^3 \quad (6)$$

Three widely used GPR configurations include on-ground (surface), off-ground, and cross-borehole measurements (Klotzsche et al., 2018). On-ground measurements are commonly conducted with GPR systems that contain the transmitter and receiver antennas at a fixed separation distance and are pushed or dragged across the land surface (Fig. 5).

Off-ground GPR systems are installed at a distinct elevation above the land surface and usually mounted on a mobile platform (e.g., ATV) that drives over the soil surface. Minet et al. (2012) demonstrated that field scale θ_v can be mapped with full-waveform inversion. Some challenges associated with off-ground GPR are the sensitivity to surface roughness and the shallower measurement depth when compared to on-ground GPR. In the cross-borehole configuration, the transmitter and receiver antennas are located in two parallel boreholes. The measured two-dimensional EM wave velocity distribution between the boreholes is recorded and converted to ϵ_b and consequently to θ_v .

Electromagnetic induction

Electromagnetic induction (EMI) proximal sensing is based on the strong correlation between the soil water content (θ_v) and the measured apparent soil electrical conductivity (EC_a) (Carroll and Oliver, 2005; Moghadas et al., 2010; Doolittle and Brevik, 2014). Electromagnetic induction enables mapping of depth-averaged EC_a at high spatial resolution. To convert the EC_a maps to θ_v , a calibration based on independently measured θ_v at strategically selected locations is required (Corwin and Lesch, 2003). An EMI sensor emits an electromagnetic (EM) field from a transmitter coil. The emitted EM field induces weak eddy current loops into the conducting soil that in turn generate a secondary electromagnetic field that differs in amplitude and phase from the primary field. The primary and secondary fields are sensed with a receiver coil and converted to EC_a (dS m⁻¹). In addition to the spacing between the transmitter and receiver coils, the distance between the coils and land surface, and the coil orientation (i.e., parallel, or perpendicular to the land surface), the magnitude of amplitude and phase differences between the primary and secondary EM fields depends on soil properties including θ_v , the concentration and types of ions in solution, the quantity and



Fig. 5 On-ground GPR system with transmitter and receiver antennas at fixed separation distance. Photo credit: DitchWitch/Flickr (CC BY-SA 2.0).

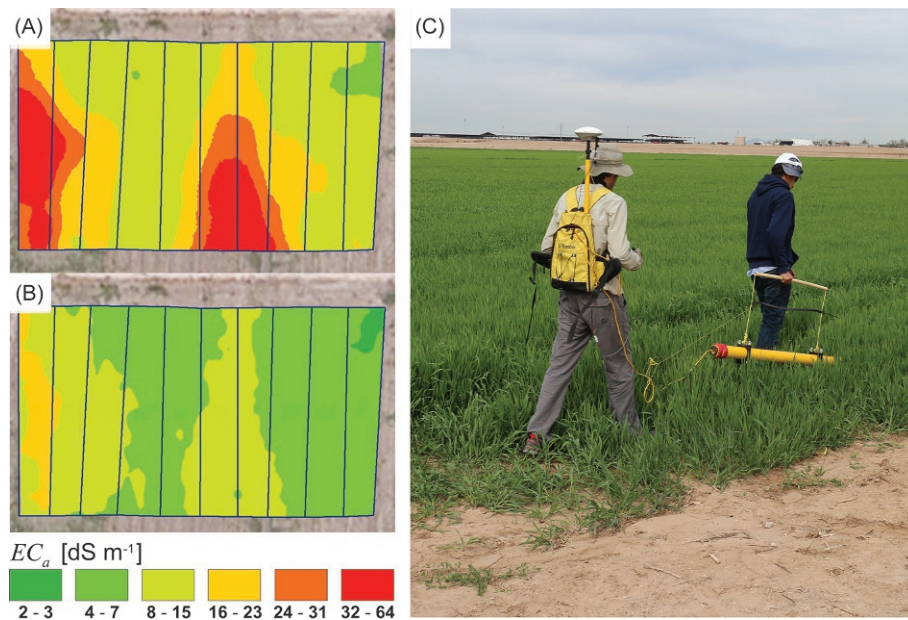


Fig. 6 EC_a mapping in Buckeye, Arizona, USA, to evaluate the salt leaching efficacy of border irrigation in an alfalfa field. EC_a maps were generated prior (A) and after (B) border irrigation. (C) The DUALEM-1S EMI sensor (Dualem Inc., Milton, Ontario, Canada) is carried over the field in 4-m spaced parallel trajectories. A GPS unit is used to georeference EC_a measurements.

mineralogy of clay constituents, and the phase and temperature of the soil water. Fig. 6 depicts mapping of EC_a with a DUALEM-1S EMI sensor in Buckeye, Arizona, USA, to evaluate the salt leaching efficacy of border irrigation in an alfalfa field.

Nuclear magnetic resonance imaging

Nuclear magnetic resonance imaging (NMRI) is based on the sensitivity of the NMRI signal intensity (echo) to soil water content (θ_v). Paetzold et al. (1985) was among the first to apply NMRI for soil moisture monitoring. The most fundamental MRI principle for θ_v determination is spin-echo imaging (Merz et al., 2016), in which an echo is created by a 180° radio-frequency pulse following an initial 90° radio frequency pulse. The echo intensity depends on the proton density, which is proportional to the water content of the soil and the relaxation times. Merz et al. (2014) tested three MRI methods for estimating near surface θ_v using the relationship:

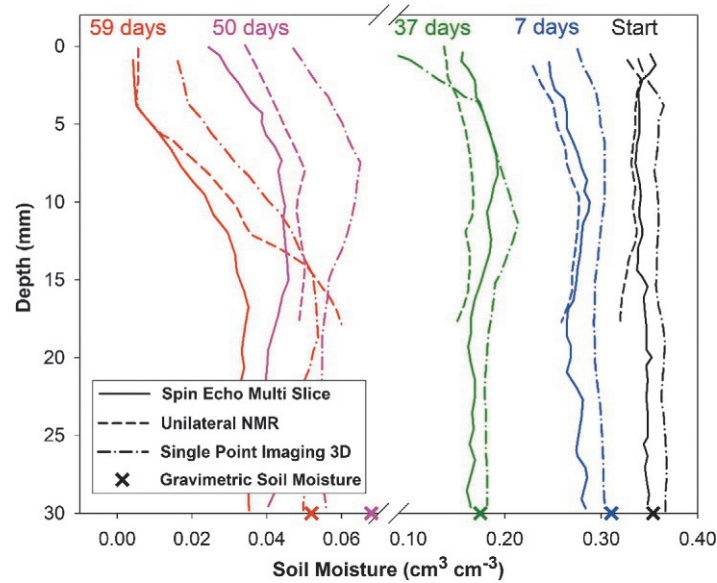


Fig. 7 Temporal near-surface θ_v profile evolution measured with different NMRI techniques. Reproduced from Merz S, Pohlmeier A, Vanderborght J, van Dusschoten D, and Vereecken H (2014) Moisture profiles of the upper soil layer during evaporation monitored by NMR. *Water Resources Research*, 50(6): 5184–5195. doi:10.1002/2013WR014809.

$$\theta_{MR} = (\theta_{a,MR} - \theta_{r,MR}) / (1 - \theta_{r,MR}) \quad (7)$$

with $\theta_{a,MR} = M(0, r) / M(0, r)_{sat}$ where $M(0, r)$ and $M(0, r)_{sat}$ denote the signal amplitude at partial and initial saturation, respectively. $\theta_{r,MR}$ is the relative magnetic resonance (MR) signal when the soil is at residual water content (i.e., dry). Fig. 7 depicts high resolution near-surface moisture profiles for medium sand monitored over time during evaporative drying with three NMRI methods (Merz et al., 2014).

Microwave radiometry

Passive microwave systems, known as radiometers, operate at long wavelengths (L-band; 1–2 GHz or 30–15 cm) and can be applied to measure thermal radiation and brightness temperature of soil and plant surfaces. Proximal L-band radiometers are commonly mounted on fixed towers or mobile platforms. The Eidgenössische Technische Hochschule (ETH) Zürich L-band Radiometer (ELBARA) is an example for a passive system that was built by the University of Bern in collaboration with ETH to measure the brightness temperature of the land surface for soil moisture retrieval at frequencies from 1.4 to 1.427 GHz (Mätzler et al., 2003). The ELBARA radiometer underwent several alterations. The ELBARA II is a newer generation of passive radiometers. It operates at 1.4 GHz and has been developed for ground calibration and validation of the Soil Moisture and Ocean Salinity (SMOS) satellite soil moisture retrieval algorithms (Schwank et al., 2010). The ELBARA-III radiometer was applied to measure freeze-thaw dynamics and soil moisture on the Tibetan Plateau (Zheng et al., 2019). Another passive radiometer is the L-band Combined Radar and Radiometer (ComRAD) simulator (O'Neill et al., 2007), which was developed by NASA's Goddard Space Flight Center (GSFC) in collaboration with George Washington University to test and refine the soil moisture retrieval algorithms for the Soil Moisture Active Passive (SMAP) satellite under controlled conditions (Srivastava et al., 2015) and to develop new radiative transfer soil moisture retrieval algorithms for vegetated terrain and small tree canopies (Kurum et al., 2011).

Summary

In many regions, the Earth's surface experiences extreme spatiotemporal variations in soil moisture that are amplified by the changing global climate. To provide water and food security for the rapidly growing human population and to aid with the conservation and sensible use of water resources, accurate spatiotemporal information about the soil moisture (SM) status is of essence. Proximal sensing techniques provide an exceedingly powerful means for characterization, monitoring, and mapping of soil moisture from the pedon to watershed scales. This chapter provides an overview of contemporary state-of-the-art proximal SM estimation methods that include electromagnetic soil moisture sensors for pedon scale measurements and geophysical techniques (i.e., GPR, EMI mapping, and NMRI) as well as the cosmic-ray neutron probe (CRNP) and passive microwave radiometry for larger scales.

References

- Anbazhagan P, Bittelli M, Palapati RR, and Mahajan P (2020) Comparison of soil water content estimation equations using ground penetrating radar. *Journal of Hydrology* 588: 125039. <https://doi.org/10.1016/j.jhydrol.2020.125039>.
- Andreasen M, Jensen KH, Desilets D, Franz TE, Zreda M, Bogaena HR, and Looms MC (2017) Status and perspectives on the cosmic-ray neutron method for soil moisture estimation and other environmental science applications. *Vadose Zone Journal* 16(8): 1–11. <https://doi.org/10.2136/vzj2017.04.0086>.
- Babaeian E, Sadeghi M, Jones SB, Montzka C, Vereecken H, and Tuller M (2019) Ground, proximal, and satellite remote sensing of soil moisture. *Reviews of Geophysics* 57: 530–616. <https://doi.org/10.1029/2018RG000618>.
- Bogaena HR, Huisman JA, Baatz R, Hendricks Franssen HJ, and Vereecken H (2013) Accuracy of the cosmic-ray soil water content probe in humid forest ecosystems: The worst case scenario. *Water Resources Research* 49(9): 5778–5791. <https://doi.org/10.1002/wrcr.20463>.
- Carroll ZL and Oliver MA (2005) Exploring the spatial relations between soil physical properties and apparent electrical conductivity. *Geoderma* 128(3–4): 354–374. <https://doi.org/10.1016/j.geoderma.2005.03.008>.
- Corvin DL and Lesch SM (2003) Application of soil electrical conductivity to precision agriculture: Theory, principles, and guidelines. *Agronomy Journal* 95(3): 455–471. <https://doi.org/10.2134/agronj2003.4550>.
- Desilets D, Zreda M, and Ferré TPA (2010) Nature's neutron probe: Land surface hydrology at an elusive scale with cosmic rays. *Water Resources Research* 46(11), W11505. <https://doi.org/10.1029/2009WR008726>.
- Doolittle JA and Brevik EC (2014) The use of electromagnetic induction techniques in soils studies. *Geoderma* 223–225: 33–45. <https://doi.org/10.1016/j.geoderma.2014.01.027>.
- González-Teruel JD, Jones SB, Robinson DA, Giménez-Gallego J, Zornoza R, and Torres-Sánchez R (2022) Measurement of the broadband complex permittivity of soils in the frequency domain with a low-cost vector network analyzer and an open-ended coaxial probe. *Computers and Electronics in Agriculture* 195: 106847. <https://doi.org/10.1016/j.compag.2022.106847>.
- Hanna L and Siam N (1980) The estimation of moisture content in the top 10 cm of soil using a neutron probe. *The Journal of Agricultural Science* 94(1): 251–253. <https://doi.org/10.1017/S0021859600028124>.
- Jagdhuber T, Stockamp J, Hajnsek I, and Ludwig R (2014) Identification of soil freezing and thawing states using SAR polarimetry at C-band. *Remote Sensing* 6(3): 2008–2023. <https://doi.org/10.3390/rs6032008>.
- Jones SB, Wraith JM, and Or D (2002) Time domain reflectometry measurement principles and applications. *Hydrological Processes* 16(1): 141–153. <https://doi.org/10.1002/hyp.513>.
- Jones SB, Blonquist JM, Robinson DA, Rasmussen VP, and Or D (2005) Standardizing characterization of electromagnetic water content sensors: Part 1. Methodology. *Vadose Zone Journal* 4(4): 1048–1058. <https://doi.org/10.2136/vzj2004.0140>.
- Kelleners TJ, Robinson DA, Shouse PJ, Ayars JE, and Skaggs TH (2005) Frequency dependence of the complex permittivity and its impact on dielectric sensor calibration in soils. *Soil Science Society of America Journal* 69(1): 67–76. <https://doi.org/10.2136/sssaj2005.0067>.
- Klotzsche A, Jonrad F, Looms MC, van der Kruk J, and Huisman JA (2018) Measuring soil water content with ground penetrating radar: A decade of progress. *Vadose Zone Journal* 17(1): 180052. <https://doi.org/10.2136/vzj2018.03.0052>.
- Kurum M, Lang RH, O'Neill PE, Joseph AT, Jackson TJ, and Cosh MH (2011) A first-order radiative transfer model formicrowave radiometry of forest canopies at L-band. *IEEE Transactions on Geoscience and Remote Sensing* 49(9): 3167–3179. <https://doi.org/10.1109/TGRS.2010.2091139>.
- Mätzler C, Weber D, Wüthrich M, Schneeberger K, Stamm C, Wyder H, and Flüher H (2003) ELBARA, the ETH L-band radiometer for soil-moisture research. In: *Proceedings of International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 3058–3060. Toulouse, France: IEEE.
- Merz S, Pohlmeier A, Vanderborght J, van Dusschoten D, and Vereecken H (2014) Moisture profiles of the upper soil layer during evaporation monitored by NMR. *Water Resources Research* 50(6): 5184–5195. <https://doi.org/10.1002/2013WR014809>.
- Merz S, Pohlmeier A, Balcom BJ, Enjilela R, and Vereecken H (2016) Drying of a natural soil under evaporative conditions: A comparison of different magnetic resonance methods. *Applied Magnetic Resonance* 47(2): 121–138. <https://doi.org/10.1007/s00723-015-0736-6>.
- Minet J, Bogaert P, Vanclooster M, and Lambot S (2012) Validation of ground penetrating radar full-waveform inversion for field scale soil moisture mapping. *Journal of Hydrology* 424–425: 112–123. <https://doi.org/10.1016/j.jhydrol.2011.12.034>.
- Moghadas D, André F, Slob EC, Vereecken H, and Lambot S (2010) Joint full-waveform analysis of off-ground zero-offset ground penetrating radar and electromagnetic induction synthetic data for estimating soil electrical properties. *Geophysical Journal International* 182(3): 1267–1278. <https://doi.org/10.1111/j.1365-246X.2010.04706.x>.
- O'Neill P, Joseph A, Nelson R, Cosh M, Jackson T, Lang R, Kurum M, and Spicknall M (2007) ComRAD active/passive microwave measurement of tree canopies. In: *Proceeding of the IEEE International Geoscience and Remote Sensing Symposium*, 1420–1423. <https://doi.org/10.1109/IGARSS.2007.4423073>.
- Paetzold RF, Matzkanin GA, and de los Santos A (1985) Surface soil water content measurement using pulsed nuclear magnetic resonance techniques. *Soil Science Society of America Journal* 49(3): 537–540. <https://doi.org/10.2136/sssaj1985.03615995004900030001x>.
- Roth K, Schulm R, Flüher H, and Attinger W (1990) Calibration of time domain reflectometry for water content measurement using a composite dielectric approach. *Water Resources Research* 26(10): 2267–2273. <https://doi.org/10.1029/WR026i010p02267>.
- Sadeghi M, Babaeian E, Arthur E, Jones SB, and Tuller M (2018) Soil physical properties and processes. In: Kutz M (ed.) *Handbook of Environmental Engineering*, pp. 137–207. Hoboken, NJ, USA: Wiley.
- Schwank M, Wiesmann A, Werner C, Mätzler C, Weber D, Murk A, Völksch I, and Wegmüller U (2010) ELBARA II, an L-band radiometer system for soilmoisture research. *Sensors* 10(1): 584–612. <https://doi.org/10.3390/s100100584>.
- Srivastava PK, O'Neill P, Cosh M, Kurum M, Lang R, and Joseph A (2015) Evaluation of dielectric mixing models for passivemicrowave soil moisture retrieval using data from ComRAD ground-based SMAP simulator. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8(9): 4345–4354. <https://doi.org/10.1109/JSTARS.2014.2372031>.
- Szyplowska A, Lewandowski A, Jones SB, Sabouroux P, Szerement J, Kafarski M, Wilczek A, and Skierucha W (2019) Impact of soil salinity, texture and measurement frequency on the relations between soil moisture and 20 MHz–3 GHz dielectric permittivity spectrum for soils of medium texture. *Journal of Hydrology* 579: 124155. <https://doi.org/10.1016/j.jhydrol.2019.124155>.
- Thomas AM (1966) In situ measurement of moisture in soil and similar substances by 'fringe' capacitance. *Journal of Scientific Instruments* 43(1): 21–27.
- Topp GC, Davis JL, and Annan AP (1980) Electromagnetic determination of soil water content: Measurements in coaxial transmission lines. *Water Resources Research* 16(3): 574–582. <https://doi.org/10.1029/WR016i003p00574>.
- Vaz CMP, Jones S, Meding M, and Tuller M (2013) Evaluation of standard calibration functions for eight electromagnetic soil moisture sensors. *Vadose Zone Journal* 12(2): 1–16. <https://doi.org/10.2136/vzj2012.0160>.
- Viscarra Rossel RA, Adamchuk VI, Sudduth KA, McKenzie NJ, and Lobsey C (2011) Chapter five—Proximal soil sensing: An effective approach for soil measurements in space and time. In: Sparks DL (ed.) *Advances in Agronomy* 113, pp. 243–291. Amsterdam, The Netherlands: Elsevier.
- Zheng D, Li X, Wang X, Wang Z, Wen J, van der Velde R, Schwank M, and Su Z (2019) Sampling depth of L-band radiometer measurements of soil moisture and freeze-thaw dynamics on the Tibetan Plateau. *Remote Sensing of Environment* 226: 16–25. <https://doi.org/10.1016/j.rse.2019.03.029>.
- Zreda M, Desilets D, Ferré TPA, and Scott RL (2008) Measuring soil moisture content non-invasively at intermediate spatial scale using cosmic-ray neutrons. *Geographical Research Letters* 35(21), L21402. <https://doi.org/10.1029/2008GL035655>.
- Zreda M, Shuttleworth WJ, Zeng X, Zweck C, Desilets D, Franz T, and Rosolem R (2012) COSMOS: the COSmic-ray soil moisture observing system. *Hydrology and Earth System Sciences* 16: 4079–4099. <https://doi.org/10.5194/hess-16-4079-2012>.

Proximal sensing of land surface temperature

Ebrahim Babaeian^a and Markus Tuller^b, ^aDepartment of Soil, Water, and Ecosystem Sciences, University of Florida, Gainesville, FL, United States; ^bDepartment of Environmental Science, The University of Arizona, Tucson, AZ, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	600
Thermal radiation	600
The basics of proximal LST sensing	601
Brightness temperature	601
Factors affecting the LST	602
Periodic variations in soil surface temperature	603
Proximal thermal infrared sensing of LST	603
Handheld TIR cameras	604
Stationary TIR sensors	604
Microwave radiometry	605
Application of LST measurements	607
References	608

Abstract

The land surface temperature (LST) governs the radiative energy budget of the Earth's surface and thus is one of the main input variables for land-surface models aimed at the estimation of soil moisture and evapotranspiration, monitoring of drought conditions and crop development, mitigation of urban heat islands, quantifying soil, vegetation, and whole ecosystem response to climate change, simulating hydrological processes, and forecasting of extreme climate events (i.e., drought, wildfires, and flooding). Because the LST affects a wide range of physical, chemical, and biological soil processes from the local to the global scales, and because of its significant spatiotemporal variations, advanced methods for measurement of the LST are of essence. Proximal sensing techniques based on thermal infrared imaging or passive microwaves are powerful means for LST determination. In this Chapter we present state-of-the-art proximal LST sensing techniques, discuss the underlying theories and algorithms, and provide examples for the application of LST information.

Key points

- State-of-the-art proximal sensing methods for land surface temperature (LST) are presented.
- The underlying theories and algorithms for LST proximal sensing are discussed.
- Examples for the application of LST information are provided.

Introduction

The Global Climate Observing System's (GCOS) Terrestrial Observation Panel for Climate (TOPC) designates the land surface temperature (LST) as an Essential Climate Variable (ECV) that is central to understand and predict climate evolution, assess risks, guide mitigation and adaptation measures, and to enable the attribution of climate events to their underlying causes. The LST drives the outgoing longwave radiation and turbulent heat fluxes at the interface between the land and atmosphere (Hulley et al., 2019), and thus is regularly used as input variable for land-surface models. The LST also influences soil microbial activity and controls soil respiration, chemical reactions, and nutrient uptake by plants (Couradeau et al., 2016). Because of the highly heterogeneous land surface that is comprised of different soil types, soil moisture regimes, vegetation covers, and topographies, the LST varies substantially in space and with time. Adequate characterization and quantification of the LST distribution and its variability necessitates measurements with detailed spatio-temporal sampling. Because of the complexity of the LST dynamics, ground-based point measurements do not provide data that are representative for large areas. Hence, indirect proximal and remote sensing techniques that measure the surface radiation energy (radiance) in the thermal infrared region of the electromagnetic spectrum or utilize passive microwaves are essential for time efficient, non-destructive, and continuous measurement of surface temperature over expansive areas of land.

Thermal radiation

Solar radiation, the amount of energy emitted by the sun as electromagnetic waves, is the primary source of energy for all processes occurring on Earth's surface. Part of the solar radiant energy is absorbed by the land surface and transformed to kinetic energy,

Table 1 Average emissivity for different land surfaces for the long-wave infrared region (8 to 14 μm).

Type of Material	Average Emissivity (ϵ)
Healthy vegetation	0.96–0.99
Dry vegetation	0.88–0.94
Wood	0.93–0.94
Sand	0.90
Dry soil	0.92
Wet soil	0.95–0.98
Water	0.98–0.99
Snow	0.98–0.99

Reproduced from Lillesand T, Kiefer RW and Chipman J (2015) *Remote Sensing and Image Interpretation*, 7th edn., New York, NY: Wiley and Sons.

thereby increasing surface temperature, while part is reflected back into the atmosphere. If the surface temperature exceeds absolute zero (i.e., -273.15°C), thermal radiation (i.e., electromagnetic energy) is emitted from the surface. The thermal radiation increases with increasing object temperature, allowing for measurement of surface temperature by quantifying the surface-emitted energy, notably in the infrared region (1 to 14 μm) of the electromagnetic spectrum.

The Stefan-Boltzmann law describes the flux of radiation energy (J_r) [W m^{-2}] received at the outer surface of the atmosphere as a function of temperature and emissivity. It implies that all objects continuously emit energy proportional to the fourth power of their surface temperature:

$$J_r = \sigma \epsilon T^4 \quad (1)$$

where σ is the Stefan-Boltzmann constant ($5.670374 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-4}$), ϵ is the emissivity [–] (varies between 0 and 1), and T is surface temperature [K]. Emissivity is defined as the ratio of the radiation energy of the object (grey body) to that of a blackbody at the same temperature, signifying the object's capacity for emitting thermal radiation. A blackbody is an idealized physical body that absorbs all incident electromagnetic radiation. For a perfect emitter or blackbody, ϵ is equal to 1, but ϵ commonly ranges between 0.8 and 1.0 for most land surfaces (Lillesand et al., 2015) (Table 1). The ϵ is mainly governed by the water content of the soil and vegetation cover, where higher water content indicates higher emissivity (Mira et al., 2010). The emissivity of healthy vegetation in the long-wave thermal infrared region exceeds 0.95 due to the high emissivity of water in leaf tissue, whereas it is below 0.95 for dry vegetation (Table 1). The land surface of natural and managed ecosystems usually comprises a mixture of soil and vegetation, both exhibiting different ϵ values. The resultant surface emissivity is a combination of soil and vegetation emissivity based on the fractions of the two surfaces. In agroecosystems, emissivity exhibits significant temporal variations due to the rapidly changing plant cover during the growth period.

The relationship between surface temperature and radiated energy of a perfect emitter, assuming that the surface emits equally in all directions, is described by Planck's law (Taylor, 1987):

$$E_r(\lambda, T) = \frac{c_1}{\lambda^5} \left[\exp\left(\frac{c_2}{\lambda T}\right) - 1 \right]^{-1} \quad (2)$$

where E_r is the spectral radiance [$\text{W m}^{-2} \mu\text{m}^{-1}$] at wavelength λ [μm] and temperature T [K] with $c_1 = 3.7417 \times 10^8 \text{ W } \mu\text{m}^4 \text{ m}^{-2}$ and $c_2 = 1.4388 \times 10^4 \mu\text{m K}$. In Eq. (1), ϵ varies with wavelength for different surfaces (e.g., soils, rocks, vegetation), hence, Eq. (2) needs to be modified to include the wavelength dependance of ϵ . Eq. (2) is inverted to solve for the surface temperature (T) of an actual surface. The relationship for E_r (Eq. (2)) as a function of surface temperature and wavelength is depicted in Fig. 1. It demonstrates that the radiative energy increases with increasing temperature, the peak of the spectrum of emitted energy shifts to shorter wavelengths as temperature increases, and the energy emitted at shorter wavelengths increases more rapidly with temperature than the energy emitted at longer wavelengths.

The basics of proximal LST sensing

Brightness temperature

The brightness temperature T_b , also referred to as radiance temperature, has been frequently used for ground- and remote sensing-based thermal infrared measurements. When the emitted radiation from a surface equals that of a blackbody, the temperature of the blackbody is defined as the T_b of the surface. It should be noted that T_b has the dimensions of temperature but lacks the physical meaning of temperature. This means that there is a difference between T_b and LST that is dependent on the amount of incident atmospheric radiation. The difference may be significant when the atmospheric emission is high, particularly in the long-wave infrared region (8 to 14 μm). Nonetheless, T_b can be used to estimate LST, when there is a single thermal infrared

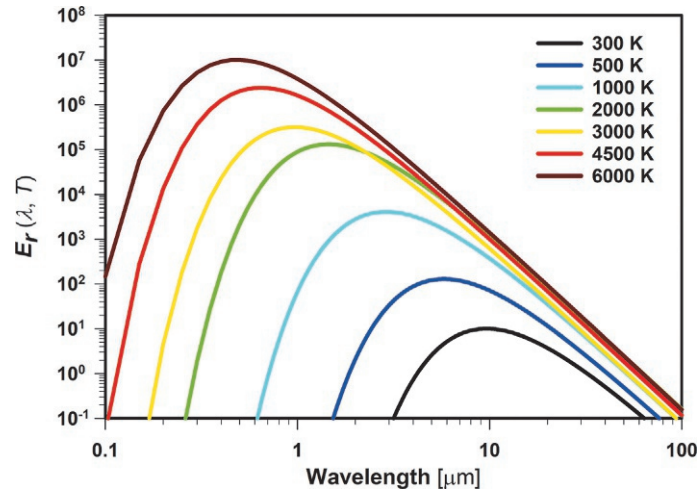


Fig. 1 The spectra of emitted radiance at various wavelengths and temperatures calculated with Eq. (2). The spectral radiance for low temperature surfaces peaks at longer wavelengths.

wavelength or channel. T_b is defined as the temperature at which a back body ($\varepsilon = 1$) in thermal equilibrium with its surrounding would emit the sensed measured temperature. It is obtained from inversion of the *Planck* equation (Eq. 2):

$$T_b = \frac{C_2}{\lambda \ln \left(\frac{C_1}{E_r(\lambda, T) \lambda^5} + 1 \right)} \quad (3)$$

For proximal and remote sensing imaging systems, λ is the central wavelength of the image channel for sensors with one or several broad thermal infrared channels. The LST is obtained by substituting Eq. (2) back into Eq. (3) and solving for T .

Factors affecting the LST

The LST is influenced by land surface and atmospheric conditions. It is well documented that the LST measured with thermal infrared imaging systems may significantly deviate from the ambient air temperature. The magnitude of the deviation varies depending on the time of day and surface conditions. For example, the difference between air temperature and LST in the middle of the day is usually larger than the difference in the morning and throughout the night. There is still a lag time between the daily maximum air temperature and the daily maximum LST. Because water has a high specific heat capacity, the LST fluctuations are smaller for moist surfaces (e.g., well-watered soil and fresh vegetation) than for dry surfaces. The size and concentration of aerosols and gases (e.g., CO_2 , H_2O) in the atmosphere may also affect LST measurements due to absorption of surface-emitted radiance, which results in underestimation of the LST.

Due to their different thermal conductivities and specific heat capacities, soil and vegetation exhibit varying temperatures. Therefore, it is critical to record soil and plant temperatures separately and identify the factors affecting soil and plant temperature variations. The amount of water in soil can affect vegetation temperature, resulting in a large difference between canopy temperature and ambient temperature above the canopy. Due to large evaporative cooling under well-watered conditions, the canopy temperature is cooler than the air temperature above the canopy. Dry soil and water-stress conditions cause the canopy temperature to exceed the ambient air temperature. Canopy temperature measurements with a proximal thermal infrared sensor may be applied for the detection of plant water stress.

It is also important to measure LST variations in space and with time. The spatial scale can affect the magnitude of the LST. At a small scale, such as a few square meters, temperature fluctuates due to heterogeneity (variability) of surface properties including soil texture, water content, and plant attributes. The fluctuations decrease with increasing footprint. In addition, the temperature of soil constituents such as fine and coarse particles, stones, organic matter vary due to difference in their thermal properties (i.e., thermal conductivity and specific heat capacity). For such conditions, an effective (average) temperature can be considered based on the fraction of each component. Periodic LST variations may cause large temperature gradients that drive soil heat flow. While the LST is not a true representation of soil temperature, not even for the upper few centimeters of the soil profile, it substantially impacts a soil's temperature regime that regulates various physical, chemical, and biological processes.

The magnitude of the LST measured with a thermal infrared camera may be additionally affected by the spatial camera resolution, the distance between the land surface and the camera, and the surface reflectivity. For a surface with high reflectivity in the thermal infrared region, the surface temperature measured with a thermal camera is a composite of the radiation from the surface of interest and the reflected thermal emission of surrounding objects that are either colder or hotter than the surface. The larger the distance between the camera and the surface, the coarser the spatial measurement resolution. For example, in a thermal image of a distant plant, the temperature of leaves, branches, and soil may be averaged into a single pixel that is inaccurate for any of

the three individual objects. Increased relative humidity in the atmosphere can attenuate the radiated energy from the object, reducing the amount of energy detected by the thermal camera (Still et al., 2019). When thermal infrared sensors are applied, measurements need to be conducted under stable conditions to exclude as many potential impacting factors as possible. These factors include the time of day, the sensor's view angle, and the spatial resolution.

Periodic variations in soil surface temperature

Soil temperature variation is characterized by periodic changes (diurnal and seasonal) that occur in response to the inherent periodicity of the atmospheric conditions that control energy inputs to the soil surface. Meteorological controls such as cloudiness, warm and cold fronts, wind speed, and precipitation disrupt diurnal and seasonal cycles and affect the soil thermal regime. A sinusoidal function can be used to quantify the diurnal effects on soil temperature regime, defined as:

$$T(z, t) = \bar{T} + A_0 e^{-\frac{z}{d}} \sin \left[\omega(t - 8) - \frac{z}{d} \right] \quad (4)$$

where z is the soil depth, t is time, $\bar{T}$ is the average temperature at the soil surface, A_0 is the amplitude of surface temperature fluctuation (i.e., the average of the minimum and maximum temperatures), d is the characteristic damping depth defined as the depth at which the amplitude A_z decreases to the fraction $1/e$ of the surface amplitude (A_0), ω is the angular frequency ($\omega = 2\pi/P$), P is the period of oscillation (e.g., P equals 365 days or 12 months), and the 8 constant is an offset in hours to the sine function to obtain maximum soil surface temperature at 2:00 PM consistent with many field measurements. The offset value may be slightly adjusted in accordance with in situ measured temperatures to achieve optimal agreement. The damping depth is defined as:

$$d = \left[\frac{2D_H}{\omega} \right]^{0.5} = \left[\frac{PD_H}{\pi} \right]^{0.5} \quad (5)$$

where D_H denotes the soil thermal diffusivity, which can be determined using soil temperature measurements taken at two separate times and depths.

$$D_H = \frac{\pi(Z_1 - Z_2)^2}{P \left[\ln \left(\frac{A_1}{A_2} \right) \right]^2} \quad (6)$$

where A_1 and A_2 are the amplitudes of temperature fluctuations for depths z_1 and z_2 (i.e., $A_1 = A_0 e^{-\frac{z_1}{d}}$ and $A_2 = A_0 e^{-\frac{z_2}{d}}$). The estimated D_H from Eq. (6) may differ due to violations of the assumptions used to derive Eq. (6). For example, if the soil is heterogeneous with depth, the measured temperature may not be well represented by a sine function.

Proximal thermal infrared sensing of LST

The physical basis of thermal infrared sensors is linked to the fact that any grey object with a temperature above absolute zero (-273.15 K) emits infrared radiation at an infrared wavelength that is dependent on the object temperature (Fig. 1). For environmental objects, radiation commonly peaks in the long-wave infrared region (8 to 14 μm).

A Thermal Infrared (TIR) camera measures the surface temperature of an object from the generated output voltage while considering the influence of the temperatures of ambient air and adjacent objects. Fig. 2 depicts the components of surface radiation together with the effect of ambient sources on the temperature detected by a TIR camera. Assuming that the received total radiation energy from a surface ($E_{r,total}$) generates an output voltage/signal (V_{cam}) that is proportional to the power input to the camera, the total signal of the camera ($V_{cam,total}$) can be calculated as the sum of emissions from the surface ($E_{r,sur}$), reflected emissions from ambient sources ($E_{r,refl}$), and the emission from the atmosphere ($E_{r,atm}$):

$$E_{r,total} = \varepsilon \tau E_{r,obj} + (1 - \varepsilon) \tau E_{r,refl} + (1 - \tau) E_{r,atm} \quad (7)$$

where ε is the surface emissivity $[-]$, $1 - \varepsilon$ is the reflectance of the surface $[-]$, and $1 - \tau$ is the emittance of the atmosphere $[-]$. The total camera output voltage can be derived by multiplying each component with a constant c and substituting cE_r with the corresponding camera signal (V_{cam}):

$$V_{cam,total} = \varepsilon \tau V_{cam,sur} + (1 - \varepsilon) \tau V_{cam,refl} + (1 - \tau) V_{cam,atm} \quad (8)$$

where $V_{cam,sur}$ is the calculated camera output voltage for a surface of temperature T_{sur} , $V_{cam,total}$ is the total camera output voltage for the actual conditions, $V_{cam,refl}$ is the theoretical camera output voltage for surface surrounding sources with temperature T_{refl} according to the calibration, and $V_{cam,atm}$ is the theoretical camera output for ambient atmosphere with temperature T_{atm} . Rearranging Eq. (8) and solving for $V_{cam,sur}$ yields:

$$V_{cam,sur} = \frac{1}{\varepsilon \tau} V_{cam,total} - \frac{1 - \varepsilon}{\varepsilon} V_{cam,refl} - \frac{1 - \tau}{\varepsilon \tau} V_{cam,atm} \quad (9)$$

Eq. (9) is the general equation used for TIR cameras for measuring surface temperature.

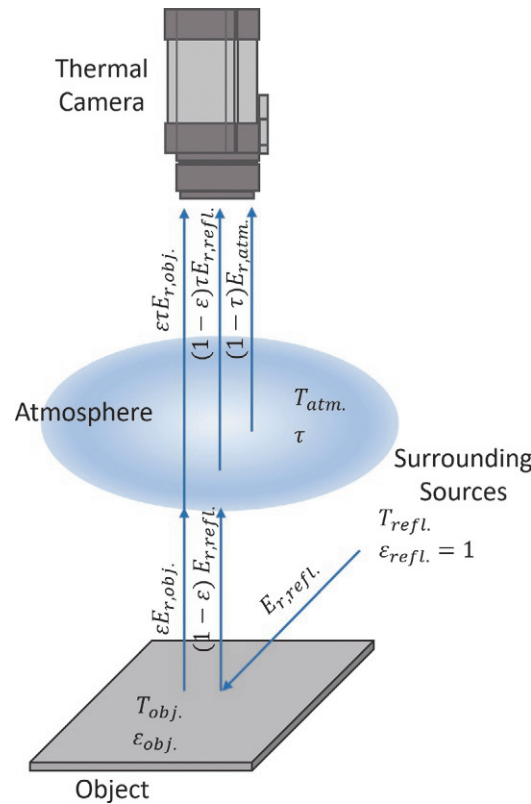


Fig. 2 Schematic depicting the effect of surrounding surfaces and atmosphere on the total radiation energy detected for an object with a TIR camera.

Handheld TIR cameras

Over the last decade, portable thermal infrared sensors have gained popularity for rapid temperature measurements in a variety of fields including environmental and agricultural sciences due to their affordability, ease of use, and ability to adjust the view field to a specific surface. In agriculture, a handheld TIR camera measures the energy radiated by the soil and plant surfaces and converts the energy received at each detector pixel to a 14- or 16-bit digital number. The raw values are transformed to temperature values via sensor-specific coefficients that are derived from the underlying physics.

Because a thermal sensor is sensitive to ambient infrared radiation, further calibration is beneficial for obtaining accurate data. To compensate for the effect of the environment, the camera requires several parameters such as emissivity, which may be obtained from tabulated values (Table 1), reflected apparent temperature, distance between the surface and the camera lens, air relative humidity, and air temperature (Maguire et al., 2021). The surface emissivity of agricultural fields changes with time due to plant growth, which should be considered for LST measurements. Optical indices such as the normalized difference vegetation index (NDVI) can be used to determine surface emissivity of vegetated areas (Brunsell and Gillies, 2002).

There are various types of handheld TIR cameras. The accuracy and temperature measurement range vary depending upon camera model. For example, a FLIR TG167 camera that operates in the 8 to 14 μm spectral range with a field of view of 19.3° can typically measure temperatures ranging from -25 to 380 °C with an accuracy of ± 1.5 °C. Fig. 3 shows an example for measurements with a FLIR TG167 camera for sorghum leaves at different dates that represent well-watered (low temperature) and water-stressed (high temperature) conditions. It is possible to detect temperature of a dry and wet soil, because of the different heat conductivities and specific heat capacities under wet and dry conditions.

It should be noted that under natural light conditions the temperature of the plant canopy is a combination of sunlit and shaded temperatures. Depending upon the height and structure of the plant canopy, the contribution of the shaded portions to total canopy temperature can be significant. Most of the literature only considers the sunlit temperature to study plant response to environmental conditions. This is because the surface temperature of sunlit canopy is known to better represent the plant conditions, for example, for water stress detection. Two approaches are commonly used to consider the shadow effects including removing shaded pixels from thermal images or using a de-shadow algorithm to correct the temperature of shaded pixels based on the temperature of sunlit pixels. Fig. 4 depicts temperature differences between sunlit and shaded portions of a corn plant (Taghvaeian et al., 2013).

Stationary TIR sensors

Over the last decade, thermal infrared imaging technology has advanced significantly, and manufacturers are now able to supply a variety of state-of-the-art TIR sensors for different temperature ranges. These include laboratory and field systems for a variety of

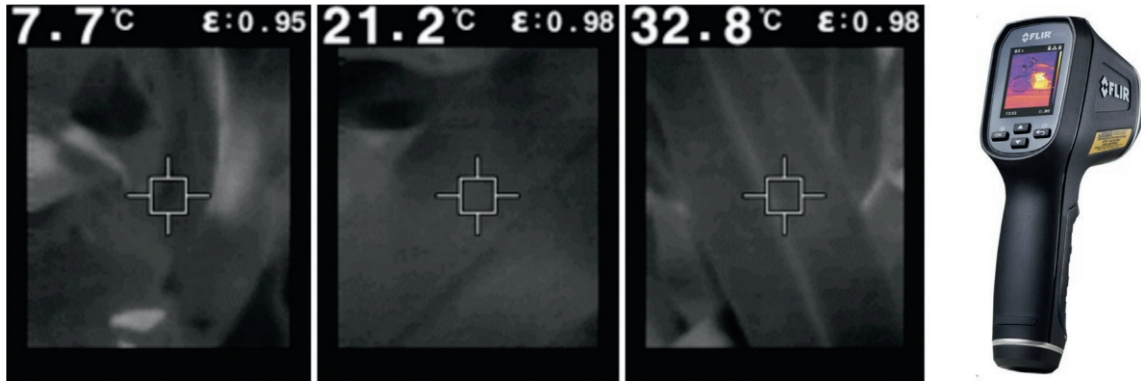


Fig. 3 Temperature ($^{\circ}\text{C}$) and emissivity (ϵ) values of energy sorghum leaves measured with a FLIR TG167 handheld TIR camera at three different dates, in Maricopa, AZ. The cursor represents the measured leaf area.

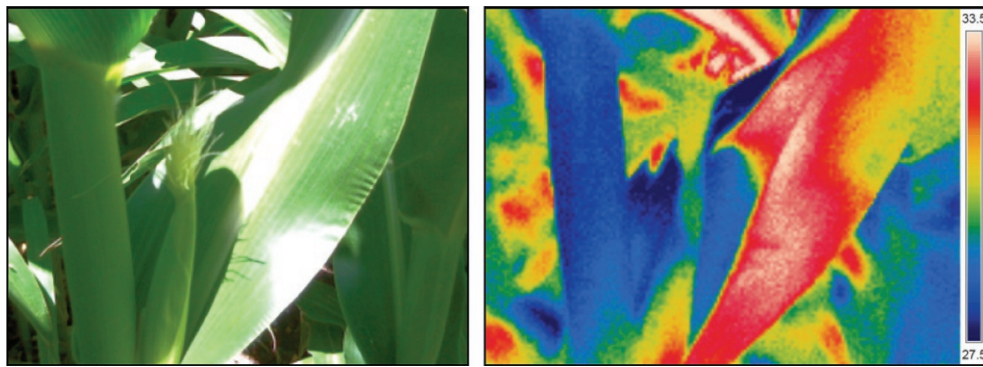


Fig. 4 A color (left) and thermal (right) image captured with a FLIR E30 camera showing temperature in $^{\circ}\text{C}$ of different parts of corn under sunlit and shaded conditions. Reproduced from Taghvaeian S, Chavez JL, Altenhofen J, Trout T and DeJonge K (2013) Remote sensing for evaluating crop water stress at field scale using infrared thermography: Potential and limitations. *Hydrology Days* doi: 10.25675/10217/201030.

applications. A thermal infrared imaging system measures the spatial distribution of surface temperature with high spatial resolution, allowing the pixels of interest to be selected for detailed analysis. For example, analysis can be limited to a portion of a full image corresponding to specific leaves of a particular plant or the soil between plants. Fig. 5 depicts a mm-scale temperature image for soil and sorghum captured with a TIR sensor mounted on a field robotic scanner (TERRA-REF, <https://terraref.org>). It is obvious that the dry soil pixels exhibit higher temperature ($\sim 320\text{ K}$) than plant pixels ($\sim 304\text{ K}$).

Microwave radiometry

Passive microwave systems, known as radiometers, operate at long wavelengths (L-band; 1–2 GHz or 15–30 cm) and can be applied to measure thermal radiation and brightness temperature (T_b) of soil and plant surfaces. At horizontal or vertical polarizations, the radiance emitted by the soil and plant canopy surfaces is a linear function of their temperature (T_s) and reflectivity (R_s). The brightness temperature measured by a radiometer oriented toward the surface is obtained from:

$$T_b = (1 - R_s)T_s + R_s T_{sky} \quad (10)$$

where $(1 - R_s)T_s$ is the upwelling brightness temperature above the surface, and T_{sky} is the downwelling sky radiation. The value of T_{sky} is determined by a radiometer oriented toward the sky.

Proximal L-band radiometers are commonly mounted on fixed towers or mobile platforms. The Eidgenössische Technische Hochschule (ETH) Zürich L-band Radiometer (ELBARA) is an example for a passive system that was built by the University of Bern in collaboration with ETH to measure the brightness temperature of the land surface for soil moisture retrieval at frequencies from 1.4 to 1.427 GHz (Mätzler et al., 2003). The radiometer is composed of a dual-mode conical horn antenna mounted on a stationary tower or the back of a truck 4–17 m above the surface. Fig. 6 shows the first generation ELBARA radiometer positioned over a plot at an experimental site in Switzerland (Mätzler et al., 2003). The footprint of the radiometer is dependent on the projection of the antenna's sensitive cone (i.e., the view angle θ), the measurement configuration defined by the observation angle φ relative to nadir, and the radiometer elevation above the land surface. The radiometer is capable of high-resolution temporal measurements of T_b in

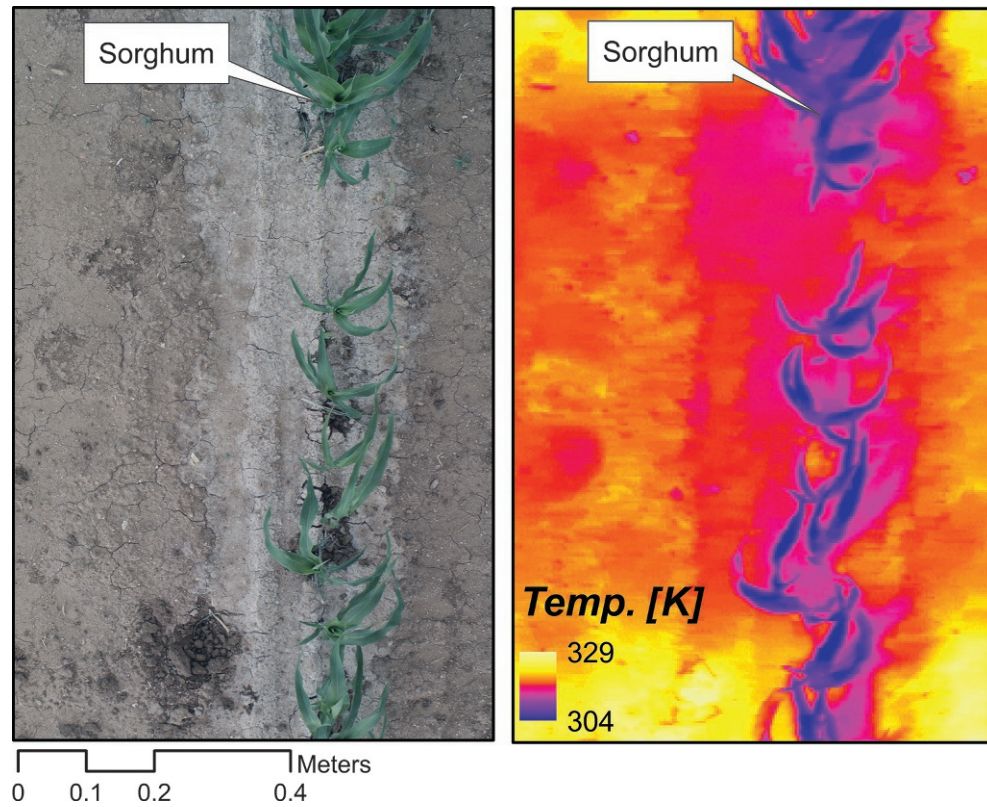


Fig. 5 A true color composite image (left) and associated TIR image with a pixel size of $2.6 \times 2.6 \text{ mm}^2$ captured with a FLIR SC615 camera across a field planted with energy sorghum in Maricopa, AZ.

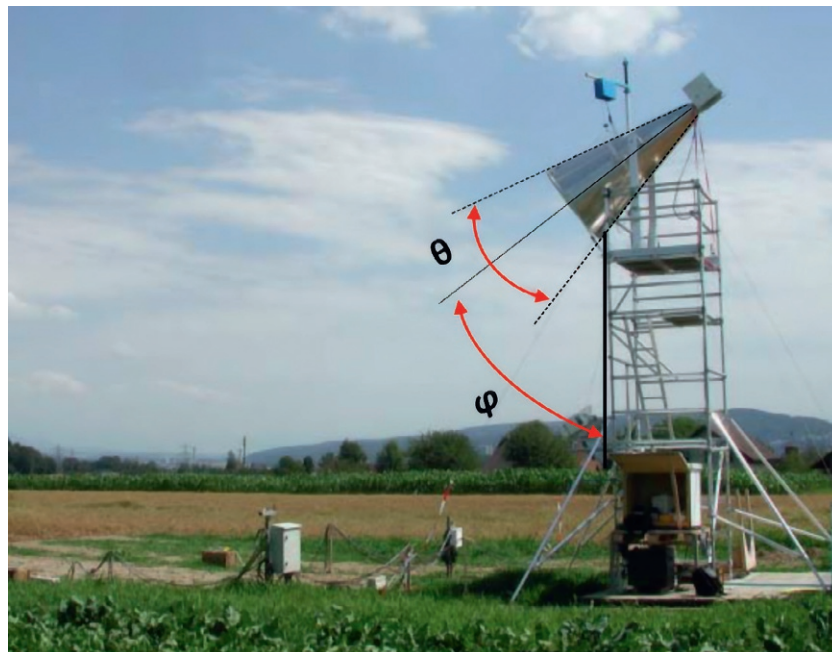


Fig. 6 ELBARA Radiometer mounted on a tower at a test site in Switzerland. θ and φ are the view and observation angles, respectively. Reproduced from Mätzler C, Weber D, Wüthrich M, Schneeberger K, Stamm C, Wydler H and Flüher H (2003) ELBARA, the ETH L-band radiometer for soil-moisture research. In: *Proceedings of International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, Toulouse, France, pp. 3058–3060.

the order of magnitude of seconds. However, multi-angular measurements (e.g., $20 < \varphi < 60$) may limit the temporal resolution due to the time required to direct the antenna at various elevation angles and the measurement integration time. The ELBARA radiometer is equipped with internal cold (273 K) and hot (338 K) loads for onsite calibration of collected data.

The ELBARA radiometer underwent several alterations. The ELBARA II is a newer generation of passive radiometers. It operates at 1.4 GHz and has been developed for ground calibration and validation of the Soil Moisture and Ocean Salinity (SMOS) satellite soil moisture retrieval algorithms (Schwank et al., 2010). ELBARA II has been deployed in several ground campaigns in Germany, Spain, Finland, and France. The brightness temperature is obtained from the radiometer raw data through an internal calibration (Eq. 11). The calibration is conducted by periodically switching between different radiometer internal reference noise sources fed to the radiometer microwave assembly, while recording the voltage responses at the output of the power detector assembly.

$$T_b = \frac{T_{RM,in} - (1 - t_{FC})T_{FC}}{t_{FC}} \quad (11)$$

where $T_{RM,in}$ is the noise temperature at the radiometer input port for H and V polarizations, which can be expressed as:

$$T_{RM,in} = \frac{U - U_{ACS}}{U_{RS} - U_{ACS}} (T_{RS} - T_{ACS}) + T_{ACS} \quad (12)$$

where T_{ACS} and T_{RS} are the noise temperatures of internal active cold sources (ACS) and resistive noise source (RS) that yield the reference response voltages U_{ACS} and U_{RS} , respectively, U is the response voltage measured by the radiometer antenna, t_{FC} and T_{FC} are transmissivity and temperature of the lossy feed cable (FC) (t_{FC} is less than 1). As Eq. (11) indicates, T_b measured by the antenna is slightly smaller than $T_{RM,in}$ due to noise added by the lossy feed cable.

The ELBARA-III radiometer was applied to measure freeze-thaw dynamics and soil moisture on the Tibetan Plateau (Zheng et al., 2019).

Another passive radiometer is the L-band Combined Radar and Radiometer (ComRAD) simulator (O'Neill et al., 2007), which was developed by NASA's Goddard Space Flight Center (GSFC) in collaboration with George Washington University to test and refine the soil moisture retrieval algorithms for the Soil Moisture Active Passive (SMAP) satellite under controlled conditions (Srivastava et al., 2015) and to develop new radiative transfer soil moisture retrieval algorithms for vegetated terrain and small tree canopies (Kurum et al., 2011). The ComRAD consists of a quad-polarized radar (1.25 GHz) and a dual-polarized radiometer (1.4 GHz) that share a 1.22 m parabolic dish antenna with a 12 degrees field of view. The antenna is mounted on a 19 m hydraulic boom truck for field deployment. Prior to measurements, the radiometer is calibrated with cold sky and ambient microwave absorber targets.

The soil characteristics (e.g., surface roughness) and canopy structure may affect T_b measurements for both horizontal and vertical polarizations. Multangular brightness temperature measurements are beneficial for separating the effect of dense vegetation cover on soil brightness temperature measurements, as dense vegetation significantly attenuates microwave signals, resulting in large volume scattering. Multangular measurements can enhance the accuracy of soil moisture retrievals over dense vegetated areas (Gherboudj et al., 2011).

Application of LST measurements

Besides for the parametrization of land-surface models, LST measurements are used in numerous applications in agriculture and environmental science.

TIR measurements are used in agriculture for detection of plant heat and drought stress and for irrigation scheduling. Plant temperature has long been recognized as a sensitive indicator of plant water stress (Jackson et al., 1981). The principle is based on the detection of changes in leaf temperature in the presence of a root zone soil water deficit. When the relative humidity of the ambient air is low (i.e., increased evaporative gradient) and the root zone is dry, the water potential in the plant decreases. As a result, the plant stomata close, which leads to a decrease in energy dissipation and an increase in plant temperature (Buckly, 2019). The difference between air and leaf temperature can provide an indicator for plant water stress. For example, the crop water stress index (CWSI) and the normalized relative canopy temperature (NRCT) (Jackson et al., 1981) are two commonly applied indices for drought stress detection in agriculture. Previous studies show that the CWSI is strongly correlated to stomatal conductance, leaf water potential, leaf transpiration rate, available soil water, leaf CO_2 exchange rate, and crop yield (Nielsen and Anderson, 1989; Irmak et al., 2000). NRCT is defined as:

$$NRCT = \frac{T_c - T_{min}}{T_{max} - T_{min}} \quad (13)$$

where T_c is the actual canopy infrared temperature, T_{min} and T_{max} are the lowest and highest temperatures that can be either measured in the field (Elsayed et al., 2015) or calculated with theoretical models with known input parameters (Moran et al., 1994). T_{min} and T_{max} can be also considered as the wet and dry bulb temperatures determined from the air temperature and the vapor pressure deficit (Jackson et al., 1981). The equation for CWSI contains similar parameters as in Eq. (13). Both the NRCT and CWSI range from 0 to 1, with higher values indicating greater water stress. To map water-stressed crops and separate plant and soil

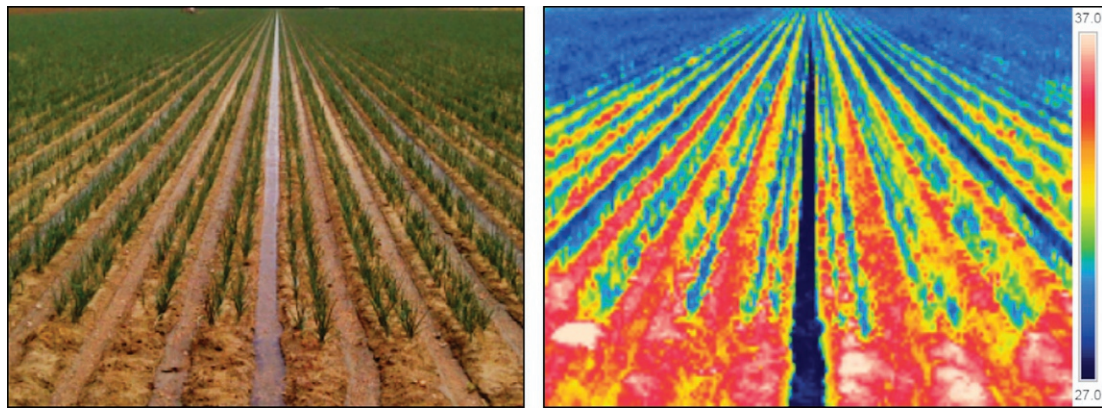


Fig. 7 Color (left) and LST (right) images of a furrow irrigated onion field. The LST [°C] was measured with a FLIR handheld camera. Reproduced from Taghvaeian S, Chavez JL, Altenhofen J, Trout T and DeJonge K (2013) Remote sensing for evaluating crop water stress at field scale using infrared thermography: Potential and limitations. *Hydrology Days* doi: [10.25675/10217/201030](https://doi.org/10.25675/10217/201030).

temperatures for irrigation scheduling, TIR measurements may be integrated with a GPS. Fig. 7 depicts the LST of a field planted with onion (Taghvaeian et al., 2013).

Surface temperature measurements are also regularly used to estimate soil moisture content. Due to the sensitivity of the surface temperature to soil and plant water content variations, TIR and optical observations may be integrated into a triangle or trapezoid model to estimate soil moisture (Moran et al., 1994). This approach has been frequently applied for large scale soil moisture estimation from satellite observations. It has the potential to be applied in conjunction with high spatial resolution imaging systems at the field scale. As discussed above, L-band radiometer brightness temperature observations are commonly used to estimate surface soil moisture and vegetation water content using either a dielectric mixing model (Srivastava et al., 2015) or a radiative transfer model (Kurum et al., 2011). A dry bare soil typically exhibits a large brightness temperature that is approximately 150 K higher than the temperature under saturated conditions. The characteristics of the soil and plant surface such as surface roughness and plant canopy structure (e.g., height and leaf area index) may influence the radiometer brightness temperature measurements.

Another area of application of LST is the estimation of evapotranspiration. When combined with vegetation indices (e.g., the normalized difference vegetation index, NDVI) measured with proximal visible and near-infrared sensors, soil and plant surface temperatures can be used for estimating actual evapotranspiration. This approach has been successfully applied at large scales using satellite thermal infrared and optical observations (Price, 1990; Carlson et al., 1994). Previous studies show that the evaporative fraction and relative soil moisture are linearly related to the surface temperature gradient Λ (Sandholt et al., 2002; Carlson, 2007) defined as:

$$\Lambda = \frac{ET_a}{ET_0} = \frac{LST_{dry} - LST}{LST_{dry} - LST_{wet}} \quad (14)$$

$$\Lambda = \frac{\theta}{\theta_s} = \frac{LST_{dry} - LST}{LST_{dry} - LST_{wet}} \quad (15)$$

where ET_0 and ET_a are the potential and actual evapotranspiration, respectively, θ and θ_s are the actual and saturated soil moisture content, respectively, LST is the land surface temperature, and subscripts 'dry' and 'wet' denote the highest and lowest surface temperatures determined via plotting NDVI versus LST for a wide range of soil and plant surface temperature and vegetation conditions. In addition, thermal imaging provides potential means for quantification of plant stomatal conductance from the temperature difference between the air and the plant leaf (Leinonen et al., 2006).

Other LST measurement applications include detection of leaf frost injury symptoms (Stegner et al., 2019) and plant pathogens (Stoll et al., 2008) as well as the assessment of aboveground biomass (Winterhalter et al., 2011) and root biomass (Lopes and Reynolds, 2010).

References

- Brunsell NA and Gillies RR (2002) Incorporating surface emissivity into a thermal atmospheric correction. *Photogrammetric Engineering and Remote Sensing* 68: 1263–1269.
- Buckly TN (2019) How do stomata respond to water status. *New Phytologist* 224(1): 21–36. <https://doi.org/10.1111/nph.15899>.
- Carlson T (2007) An overview of the "triangle method" for estimating surface evapotranspiration and soil moisture from satellite imagery. *Sensors* 7(8): 1612–1629. <https://doi.org/10.3390/s7081612>.
- Carlson TN, Gillies RR, and Perry EM (1994) A method to make use of thermal infrared temperature and NDVI measurements to infer surface soil water content and fractional vegetation cover. *Remote Sensing Reviews* 9(1–2): 161–173. <https://doi.org/10.1080/02757259409532220>.
- Couradeau E, Karaouz U, Lim HC, da Rocha UN, Northen T, Brodie E, and Garcia-Pichel F (2016) Bacteria increase arid-land soil temperature through the production of sunscreens. *Nature Communications* 7: 10373. <https://doi.org/10.1038/ncomms10373>.

- Elsayed S, Rischbeck P, and Schmidhalter U (2015) Comparing the performance of active and passive reflectance sensors to assess the normalized relative canopy temperature and grain yield of drought-stressed barely cultivars. *Field Crops Research* 177: 148–160. <https://doi.org/10.1016/j.fcr.2015.03.010>.
- Gherboudj I, Magagi R, Berg AA, and Toth B (2011) Soil moisture retrieval over agricultural fields from multi-polarized and multi-angular RADARSAT-2 SAR data. *Remote Sensing of Environment* 115(1): 33–43. <https://doi.org/10.1016/j.rse.2010.07.011>.
- Hulley GC, Ghent D, Götsche FM, Guillevic PC, Mildrexler DJ, and Coll C (2019) Land surface temperature. In: Hulley GC and Ghent D (eds.) *Taking the Temperature of the Earth*, pp. 57–127. Amsterdam, Netherlands: Elsevier. <https://doi.org/10.1016/B978-0-12-814458-9.00003-4>.
- Irmak S, Haman DZ, and Bastug R (2000) Determination of crop water stress index for irrigation timing and yield estimation of corn. *Agronomy Journal* 92(6): 1221–1227. <https://doi.org/10.2134/agronj2000.9261221x>.
- Jackson RD, Idso SB, Reginato RJ, and Pinter PJ (1981) Canopy temperature as a crop water stress indicator. *Water Resources Research* 17(4): 1133–1138. <https://doi.org/10.1029/WR017i004p01133>.
- Kurum M, Lang RH, O'Neill PE, Joseph AT, Jackson TJ, and Cosh MH (2011) A first-order radiative transfer model for microwave radiometry of forest canopies at L-band. *IEEE Transactions on Geoscience and Remote Sensing* 49(9): 3167–3179. <https://doi.org/10.1109/TGRS.2010.2091139>.
- Leinonen I, Grant OM, Tagliavia CPP, Chaves MM, and Jones HG (2006) Estimating stomatal conductance with thermal imagery. *Plant, Cell & Environment* 29(8): 1508–1518. <https://doi.org/10.1111/j.1365-3040.2006.01528.x>.
- Lillesand T, Kiefer RW, and Chipman J (2015) *Remote Sensing and Image Interpretation*, 7th edn. New York, NY: Wiley and Sons.
- Lopes MS and Reynolds MP (2010) Partitioning of assimilates to deeper roots is associated with cooler canopies and increased yield under drought in wheat. *Functional Plant Biology* 37(2): 147–156. <https://doi.org/10.1071/FP09121>.
- Maguire MS, Neale CMU, and Woldt WE (2021) Improving accuracy of unmanned aerial system thermal infrared remote sensing for use in energy balance models in agriculture applications. *Remote Sensing* 13: 1635. <https://doi.org/10.3390/rs13091635>.
- Mätzler C, Weber D, Wüthrich M, Schneeberger K, Stamm C, Wydler H, and Flüeler H (2003) ELBARA, the ETH L-band radiometer for soil-moisture research. In: *Proceedings of International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE, Toulouse, France 3058–3060.
- Mira M, Valor E, Caselles V, Rubio E, Coll C, Galve JM, Niclòs R, Sánchez JM, and Boluda R (2010) Soil moisture effect on thermal infrared (8–13 µm) emissivity. *IEEE Transactions on Geoscience and Remote Sensing* 48(5): 2251–2260. <https://doi.org/10.1109/TGRS.2009.2039143>.
- Moran MS, Clarke TR, Inoue Y, and Vidal A (1994) Estimating crop water deficit using the relation between surface-air temperature and spectral vegetation index. *Remote Sensing of Environment* 49(3): 246–263. [https://doi.org/10.1016/0034-4257\(94\)90020-5](https://doi.org/10.1016/0034-4257(94)90020-5).
- Nielsen DC and Anderson RL (1989) Infrared thermometry to measure single leaf temperatures for quantification of water stress in sunflower. *Agronomy Journal* 81(5): 840–842. <https://doi.org/10.2134/agronj1989.00021962008100050028x>.
- O'Neill P, Joseph A, Nelson R, Cosh M, Jackson T, Lang R, Kurum M, and Spicknall M (2007) ComRAD active/passive microwave measurement of tree canopies. In: *Proceeding of the IEEE International Geoscience and Remote Sensing Symposium* 1420–1423. <https://doi.org/10.1109/IGARSS.2007.4423073>.
- Price JC (1990) Using spatial context in satellite data to infer regional scale evapotranspiration. *IEEE Transactions on Geoscience and Remote Sensing* 28(5): 940–948. <https://doi.org/10.1109/36.58983>.
- Sandholt I, Rasmussen K, and Andersen J (2002) A simple interpretation of the surface temperature/vegetation index space for assessment of surface moisture status. *Remote Sensing of Environment* 79(2–3): 213–224. [https://doi.org/10.1016/S0034-4257\(01\)00274-7](https://doi.org/10.1016/S0034-4257(01)00274-7).
- Schwank M, Wiesmann A, Werner C, Mätzler C, Weber D, Murk A, Völksch I, and Wegmüller U (2010) ELBARA II, an L-band radiometer system for soilmoisture research. *Sensors* 10(1): 584–612. <https://doi.org/10.3390/s100100584>.
- Srivastava PK, O'Neill P, Cosh M, Kurum M, Lang R, and Joseph A (2015) Evaluation of dielectric mixing models for passivemicrowave soil moisture retrieval using data from ComRAD ground-based SMAP simulator. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8(9): 4345–4354. <https://doi.org/10.1109/JSTARS.2014.2372031>.
- Stegner M, Schäfermolte T, and Neuner G (2019) New insights in potato leaf freezing by infrared thermography. *Applied Sciences* 9(5): 819. <https://doi.org/10.3390/app9050819>.
- Still C, Powell R, Aubrecht D, Kim Y, Helliker B, Roberts D, Richardson AD, and Goulden M (2019) Thermal imaging in plant and ecosystem ecology: applications and challenges. *Ecosphere* 10(6): e02768. <https://doi.org/10.1002/ecs2.2768>.
- Stoll M, Schultz HR, and Berkelmann-Loehnertz B (2008) Exploring the sensitivity of thermal imaging for Plasmopara viticola pathogen detection in grapevines under different water status. *Functional Plant Biology* 35(4): 281–288. <https://doi.org/10.1071/FP07204>.
- Taghvaeian S, Chavez JL, Altenhofen J, Trout T, and DeJonge K (2013) Remote sensing for evaluating crop water stress at field scale using infrared thermography: Potential and limitations. *Hydrology Days*. <https://doi.org/10.25675/10217/201030>.
- Taylor JH (1987) Planck's radiation law. *Optics News* 13(2): 26–27.
- Winterhalter L, Mistele B, Jampatong S, and Schmidhalter U (2011) High throughput phenotyping of canopy water mass and canopy temperature in well-watered and drought stressed tropical maize hybrids in the vegetative stage. *European Journal of Agronomy* 35(1): 22–32. <https://doi.org/10.1016/j.eja.2011.03.004>.
- Zheng D, Li X, Wang X, Wang Z, Wen J, van der Velde R, Schwank M, and Su Z (2019) Sampling depth of L-band radiometer measurements of soil moisture and freeze-thaw dynamics on the Tibetan Plateau. *Remote Sensing of Environment* 226: 16–25. <https://doi.org/10.1016/j.rse.2019.03.029>.

Proximal sensing of evapotranspiration

Ebrahim Babaeian^a and Markus Tuller^b, ^aDepartment of Soil, Water, and Ecosystem Sciences, University of Florida, Gainesville, FL, United States; ^bDepartment of Environmental Science, The University of Arizona, Tucson, AZ, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	610
The energy balance of the land surface and evapotranspiration	610
The terrestrial water balance and evapotranspiration	611
Factors that control evapotranspiration	611
Energy balance-based proximal sensing of evapotranspiration	612
The eddy covariance method	612
The scintillometer method	614
Water balance-based proximal sensing of evapotranspiration	615
Summary	616
References	616

Abstract

Evapotranspiration (*ET*) is a key component of the water and energy cycles. In-depth knowledge about *ET* is of essence for a plethora of applications that include estimation of crop water consumption, irrigation scheduling and water resources management, modeling of hydrologic processes, forecasting of drought conditions, and monitoring of ecosystem response to climate change. Various proximal sensing methods may be applied in conjunction with the land surface energy balance or the terrestrial water balance to estimate *ET* at the local scale. In this chapter, we review state-of-the-art proximal *ET* sensing techniques along with their underlying theories.

Key points

- The importance of evapotranspiration (*ET*) is reviewed along with the governing soil and atmospheric variables.
- *ET* estimation by state-of-the-art proximal sensing techniques is described together with the terrestrial water and land surface energy balances.

Introduction

Evapotranspiration (*ET*) encompasses all processes related to water movement from the land surface to the atmosphere via evaporation and transpiration. *ET* consists of water evaporation into the atmosphere from the soil surface and water bodies, evaporation from the capillary fringe of the groundwater table, and transpiration — the movement of water from the soil to the atmosphere via plants. Transpiration occurs when plants take up liquid water from the soil and release water vapor into the air via their stomata. *ET* links the water, energy, and carbon cycles, hence, significantly affecting land surface-atmosphere interactions. *ET* regulates the balance of water and energy of the Earth's biosphere, hydrosphere, and atmosphere, especially in arid and semiarid ecosystems, where it may account for nearly the entire surface water budget (Fisher et al., 2017; Oki and Kanae, 2006; Morillas et al., 2013). Warming of the land and atmosphere causes fundamental changes in *ET*, because warmer air holds more water vapor evaporated from the soil surface and transpired by plants. This leads to shifts in the patterns and seasonality of precipitation and soil moisture, which modifies the balance between the water cycle components and may result in extreme weather events such as floods, droughts, wildfires, and landslides. For example, in arid and semiarid areas drought conditions are exacerbated during periods of low precipitation. Accurate estimation of *ET* is essential for the determination of agricultural and ecosystem water use, drought detection, quantification of land-atmosphere feedbacks, assessment of the anthropogenic impact on the water cycle and surface energy budget, and for validation and improvement of the parameterization of land surface models (Purdy et al., 2018). Typically, *ET* is estimated from the water or energy balance. Proximal sensing techniques, including eddy covariance and scintillometer methods, provide powerful means for indirect measurements of land surface and atmospheric variables, which may be integrated with surface energy balance to provide robust spatiotemporal *ET* estimates at the local scale.

The energy balance of the land surface and evapotranspiration

The land surface net radiation (R_n), the difference between the incoming and outgoing radiation, drives evapotranspiration (ET) and the air and soil heat fluxes as well as other processes with lower energy consumption (e.g., photosynthesis). The energy balance of the land surface is formulated as:

$$R_n = \lambda ET + H + G \quad (1)$$

where λET is the latent heat flux (W m^{-2}) with λ representing the latent heat of vaporization (2.449 MJ kg^{-1} or 585 cal g^{-1}), H is the sensible heat flux (W m^{-2}), i.e., the energy utilized to heat the air, and G is the soil heat flux (W m^{-2}), i.e., the energy used to heat the soil. The magnitude of R_n is determined by several factors, including albedo, latitude, and terrain characteristics (e.g., slope and aspect). Under wet surface conditions (i.e., wet soil and well-watered plants), most of R_n is consumed by ET . Fig. 1 depicts the components of the surface energy balance and their direction relative to the land surface.

The terrestrial water balance and evapotranspiration

The basic equation describing the water balance, which is based on the conservation of the mass of water, is given as:

$$P + I = ET + DR + RO - \Delta W \quad (2)$$

where P is precipitation, I is irrigation, ET is evapotranspiration, DR is drainage and deep percolation, RO is surface runoff, and ΔW is the change in the soil water storage (water depletion). All components of the water balance are commonly expressed in length units. W is defined as the equivalent depth of water (D_e) stored in a soil profile under consideration over a specific time period (i.e., $\Delta W = W_{t_1} - W_{t_2}$). D_e is the depth of the soil profile multiplied with the average volumetric soil water content of the profile. All components of the water balance are associated with a specific time interval (t_1 to t_2), and all water additions to the soil profile are assigned a positive sign, whereas all outputs have a negative sign. Water balance calculations may be applied to areas ranging in size from small plots (e.g., backyard garden) to watershed and global scales.

Factors that control evapotranspiration

Evaporation (E) and transpiration (T) occur simultaneously and are difficult to discern. From a physics perspective, E is the transformation of liquid water to water vapor. This process requires energy that is supplied by R_n and, to a lesser extent, by the ambient air temperature. The difference between the water vapor pressure at the evaporating surface and that of the surrounding air drives E . As E progresses, the air surrounding the evaporating surface becomes progressively saturated with water molecules, and E will eventually cease if the saturated air is not transferred to the atmosphere. Under such conditions, the replacement of humid air with dryer air is primarily determined by wind speed. Transpiration (T) is the process by which soil water is taken up by plant roots, transported through the plant xylem, and evaporated at the surface of mesophyll cells in the plant leaves before exiting through stomata. The soil water status and atmospheric water vapor content (i.e., relative humidity) affect stomata opening and closing. During precipitation or irrigation events, the plant canopy intercepts and absorbs a portion of the water. Precipitation or irrigation

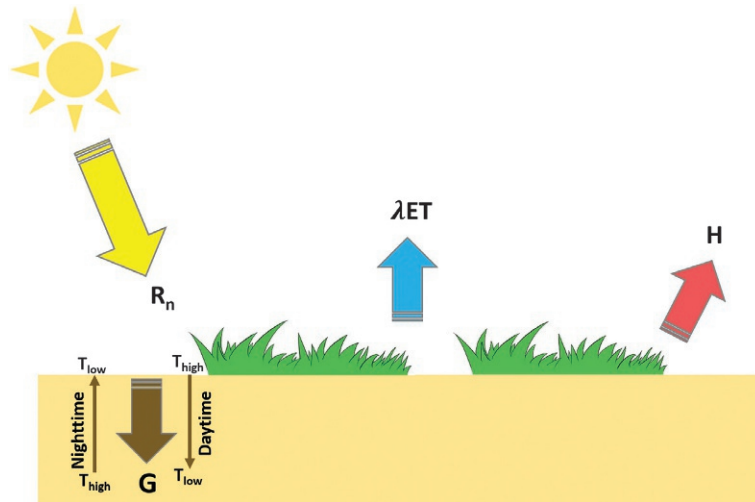


Fig. 1 Schematic of the surface energy balance components. T_{low} and T_{high} refer to soil temperature.

that is not intercepted is called throughfall and water that reaches the ground via trunks and stems is referred to as stemflow. A portion of the intercepted water evaporates at rates determined by the prevailing atmospheric conditions and the physiology of plants. In temperate humid regions, evaporation of intercepted water is an important component of the water balance. Forests commonly exhibit greater interception losses than grasslands. This is because of the greater aerodynamic roughness of the forest canopy that results in a more efficient transfer of water vapor away from the evaporating canopy surfaces.

Variations in atmospheric and soil conditions affect the rate of evapotranspiration. In general, factors such as atmospheric demand, soil water availability, soil texture and structure, and plant physiological aspects affect the rate of ET . Net radiation, air temperature, relative humidity, and wind speed are the major atmospheric parameters that control the atmospheric demand. Reference or potential evapotranspiration expresses the effect of the atmosphere on water loss. The reference evapotranspiration represents the water loss from a fully vegetated surface covered with grass or alfalfa.

The soil moisture (SM) status along with the ability of the soil to transmit water to the plant roots are significant determinants of the ET rate. SM controls the partitioning of the total available energy (i.e., R_n) into sensible (H) and latent (λET) heat fluxes at the land surface and determines near-surface ambient air conditions (relative humidity). LeMone et al. (2007) demonstrate that the SM content may significantly impact the magnitude of sensible and latent heat fluxes. ET is mainly influenced by the SM deficit and the soil type. Saturated soil conditions cause oxygen deficiency that limits root water uptake. The transpiration rate is affected by plant physiology and characteristics such as development stage, canopy roughness, canopy height, and rooting depth. Furthermore, agricultural management practices may influence the transpiration rate, especially under water stress. Nutrient status, soil salinity, the presence of impeding soil layers, plant diseases, and pests are additional factors to be considered when estimating ET , as they all may limit plant growth and thus reduce the rate of ET .

Energy balance-based proximal sensing of evapotranspiration

The eddy covariance method

The eddy covariance (EC) method is frequently applied in agriculture, hydrology, ecosystem science, and industry. EC systems measure atmospheric fluxes of H_2O , CO_2 , sensible heat (H), and latent heat (λET) that are transferred as eddies of varying size caused by turbulence within the atmospheric boundary layer. Mathematically, the EC method is based on the covariance between vertical transfer velocity measurements, the concentration of the quantity of interest, and the upward and downward air movements (Burba, 2013). As depicted in Fig. 2, at a single location on an EC system and for a single instant of time ($t = 1$), eddy #1 transports an air parcel with mass c_1 and velocity v_1 downward. At $t = 2$, eddy #2 propels an air parcel of mass c_2 upward with velocity v_2 . Because each air parcel possesses unique characteristics, the flux of interest is determined from the concentration, temperature, and humidity, among other variables. By measuring these characteristics along with the velocity of the vertical air movement, the vertical upward or downward fluxes of water vapor and gas concentrations can be determined.

Rapid fluctuating turbulent conditions pose challenges to EC measurements. The upward and downward motions and the number of molecules must be measured at high frequency, which requires sophisticated instrumentation that is capable of measuring small concentration fluctuations, densities, and temperatures with high accuracy. The EC method has distinct advantages over other ET estimation techniques that include: (1) direct measurements of sensible heat (H), and latent heat (λET) fluxes; (2) minimal disturbance of the region of interest; (3) spatially averaged measurements over large areas; and (4) automation capability for continuous long-term measurements.

The EC method determines the latent heat flux (λET) from the covariance between fluctuations in instantaneous vertical wind speed and water vapor density according to:

$$\lambda ET \text{ (W m}^{-2}\text{)} = \lambda \overline{\delta v \delta \rho_v} \quad (3)$$

where λ is the latent heat of vaporization (2260 J g^{-1}), $\overline{\delta v}$ is the mean fluctuation of the vertical wind speed (m s^{-1}) measured over a specific time period (e.g., 30 min), and $\overline{\delta \rho_v}$ is the mean fluctuation of water vapor density (g m^{-3}) for the same time period.

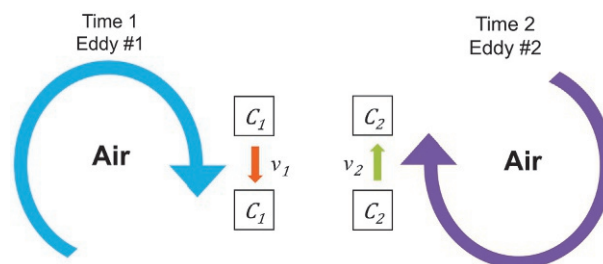


Fig. 2 Sketch illustrating the measurement principle of the eddy covariance method. Reproduced from Burba G (2013) *Eddy Covariance Method for Scientific, Industrial, Agricultural, and Regulatory Applications: A Field Book on Measuring Ecosystem Gas Exchange and Areal Emission Rates*. LI-COR Biosciences: Lincoln, Nebraska. <https://doi.org/10.13140/RG.2.1.4247.8561>.

The sensible heat flux (H) is a consequence of the direct transfer of heat from the land surface to the atmosphere due to convection and conduction processes as well as the temperature gradient between the land surface and the atmosphere. H may be calculated as:

$$H \text{ (W m}^{-2}\text{)} = \rho_a C_p \overline{\delta v} \overline{\delta T_a} \quad (4)$$

where ρ_a is the air density (g m^{-3}), C_p is the specific heat capacity of air at constant pressure ($\text{J g}^{-1} \text{ } ^\circ\text{C}^{-1}$), $\overline{\delta v}$ is the mean fluctuation of the vertical wind speed (m s^{-1}) measured over a specific time period (e.g., 30 min), and $\overline{\delta T_a}$ is the mean fluctuation of the air temperature ($^\circ\text{C}$) for the same time period.

The EC method is based on the following assumptions: (1) measurements at a specific point represent an upwind area; (2) fluxes are turbulent; (3) the terrain is horizontal and uniform; (4) measurements are conducted within the boundary layer and inside the constant flux layer; (5) air density fluctuations are negligible; and (6) instruments are capable of capturing small changes at extremely high frequencies. The extent to which some of these assumptions are valid depends on the selection of an appropriate site, the experimental design, and the atmospheric conditions.

Fig. 3 depicts a typical EC system consisting of a 3-D sonic anemometer for measuring wind speed and direction. The anemometer is commonly integrated with a gas analyzer for CO_2 and H_2O vapor concentration measurements. Wind speed and direction and CO_2 and H_2O vapor concentrations need to be measured at high frequency (i.e., 5–20 Hz). The system may further comprise of air temperature and relative humidity sensors, a net radiometer as well as soil moisture sensors and heat flux plates vertically distributed along the soil profile to measure additional soil and atmospheric variables. All measured data are automatically recorded with a data acquisition system.

The major variables measured with a basic EC system are listed in Table 1. The analysis of EC measurements is not trivial because numerous corrections are required. Burba (2013) provides an extensive review of the EC method including data analysis.

Fig. 4 shows diurnal variations of λET , H , and R_n for lettuce and wheat for mid- and end-growth stages. In the middle of the growth season, when soil water is not limited, the majority of R_n is expended by λET .

The EC method is the technique of choice for measuring surface fluxes at AmeriFlux (<https://ameriflux.lbl.gov/>) and FLUXNET (<https://fluxnet.org/>) sites. AmeriFlux is a network of more than 470 registered flux tower sites distributed across North, South, and Central America. It was established to quantify water, energy, and carbon fluxes to determine the role of the terrestrial biosphere in global climate change. FLUXNET is an international “network of networks,” connecting regional networks of earth system scientists, who use the EC method to measure the cycling of carbon, water, and energy between the biosphere and atmosphere to better understand ecosystem functioning, and to detect trends in climate, greenhouse gases, and air pollution.

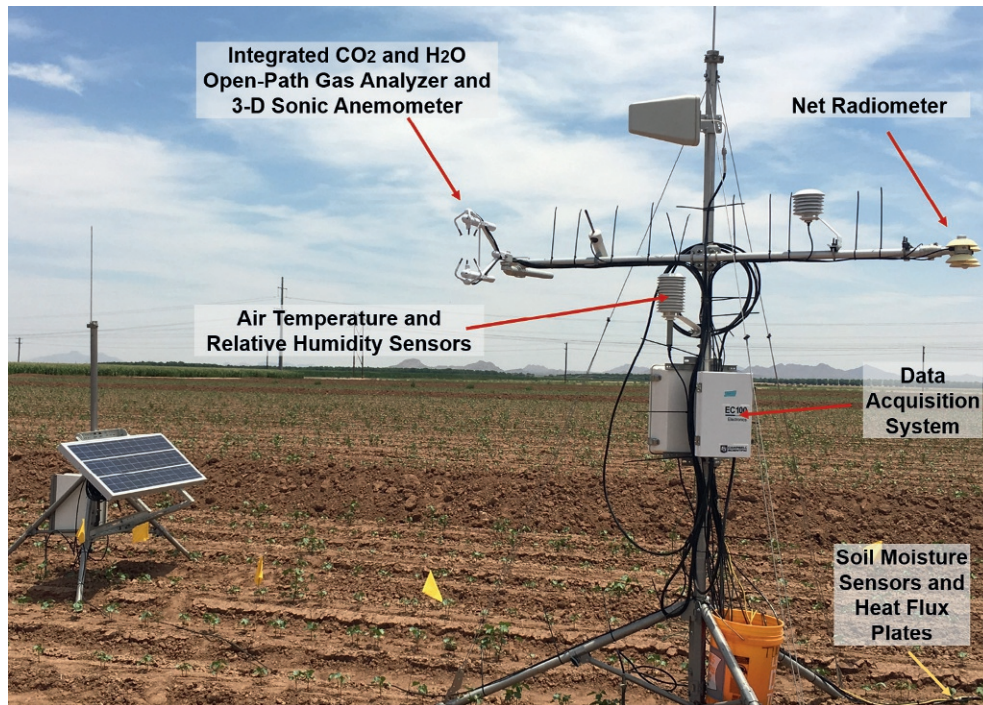
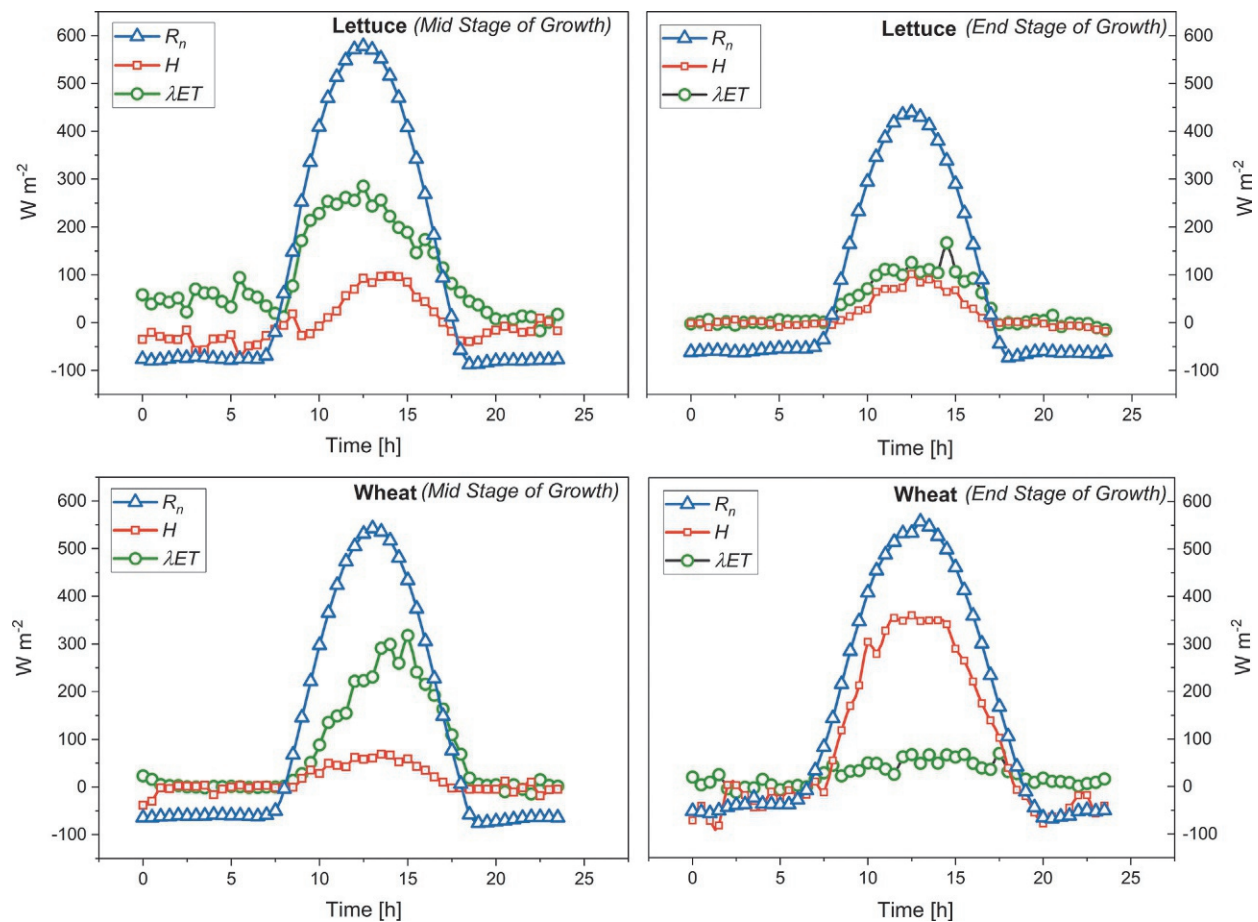


Fig. 3 Eddy covariance system installed in a cotton field in Maricopa, Arizona, USA.

Table 1 A list of major variables measured with an eddy covariance system.

Variable	Units
H ₂ O flux	$\text{g m}^{-2} \text{s}^{-1}$
CO ₂ flux	$\text{mg m}^{-2} \text{s}^{-1}$
H ₂ O concentration	$\text{g m}^{-2} \text{s}^{-1}$
CO ₂ concentration	$\text{mg m}^{-2} \text{s}^{-1}$
Friction velocity	m s^{-1}
Shear stress	$\text{kg m}^{-1} \text{s}^{-2}$
Latent heat flux	W m^{-2}
Sensible heat flux	W m^{-2}
Mean wind speed in x, y, z directions	m s^{-1}
Covariance of H ₂ O and mean wind speed in x, y, z directions	$\text{g m}^{-2} \text{s}^{-1}$
Covariance of C ₂ O and mean wind speed in x, y, z directions	$\text{mg m}^{-2} \text{s}^{-1}$

**Fig. 4** Diurnal variations of surface energy balance components for mid-season and end of season growth stages of lettuce and wheat planted in Maricopa and Yuma, Arizona, USA, in 2018, respectively.

The scintillometer method

There are different types of scintillometers, namely large aperture scintillometers (LAS) that operate at near-infrared (NIR) wavelengths ($\sim 850 \text{ nm}$), and microwave scintillometers that operate at wavelengths of about 2 mm. The LAS is designed to measure the sensible heat flux (H) over path lengths from 100 m up to 5 km. When combined with a microwave scintillometer (Fig. 5), both H and the latent heat flux (λET) can be measured directly. If the LAS is used on its own, the area-averaged λET can be determined as a residual of the surface energy balance (Eq. 1) when net radiation (R_n) and the soil heat flux (G) are additionally measured with suitable instrumentation (e.g., net radiometer and heat flux plates) (Meijninger et al., 2002).

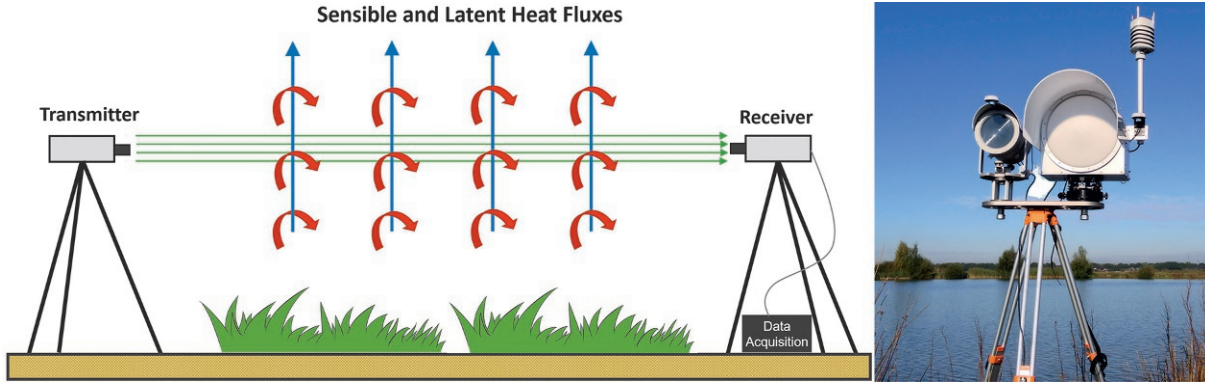


Fig. 5 Operation principle of a scintillometer (left). A Kipp & Zonen LAS MkII Scintillometer (OTT HydroMet B.V., Delft, The Netherlands) combined with an RPG-MWSC-160 microwave scintillometer (RPG Radiometer Physics GmbH Meckenheim, Germany) for direct measurement of H and λET (right).

Scintillometers measure fluctuations in turbulent intensity and display them as a path-averaged structure parameter of the air refractive index C_n^2 . For example, a LAS transmitter emits a horizontal near-infrared (NIR) beam, which is captured by the receiver that analyzes the resulting beam fluctuations caused by density and turbulence-induced atmospheric scintillations. C_n^2 is related to the temperature structure parameter C_T^2 , air pressure P (Pa), air temperature T (K), and the Bowen ratio (β):

$$C_T^2 = C_n^2 \left(\frac{T^2}{\gamma P} \right)^2 \left(1 + \frac{0.03}{\beta} \right)^{-2} \quad (5)$$

where C_T^2 and C_n^2 are expressed in units of $K^2 m^{-2/3}$, and $m^{-2/3}$, respectively, and γ is the air refractive index parameter ($-0.78 \times 10^{-6} K Pa^{-1}$). β can be determined from eddy covariance measurements, or iteratively when the difference between R_n and G is known (Solignac et al., 2009; Ezzahar et al., 2009). H is iteratively derived from C_T^2 and wind speed data based on the semiempirical Monin-Obukhov similarity theory (MOST):

$$H (W m^{-2}) = \rho_a C_p u^* T^* \quad (6)$$

where C_p is the specific heat of air ($J kg^{-1}$), ρ_a is the air density ($kg m^{-3}$), u^* is the friction velocity ($m s^{-1}$), and T^* is an air temperature scale. Details about the MOST analysis are provided in Foken (2006) and McAneney et al. (1995).

Scintillometers offer some advantages over the eddy covariance method, including average flux observations over longer distances, which are sounder for the estimation of total crop water consumption in agriculture, and require less complex setups, maintenance, and data processing. There are also some limitations associated with the scintillometer method that include effects of meteorological constraints on long-term measurements (e.g., precipitation, poor visibility, and weak turbulence), methodological constraints (e.g., signal saturation and inner-scale signal dependence), and sensitivity to vibrations of the mounting platform. Terrain characteristics, such as surface topography, type of land cover, and non-uniformity in vegetation height, may also affect the measurement accuracy.

Water balance-based proximal sensing of evapotranspiration

The evapotranspiration (ET) for agricultural fields can be calculated over a specific time period (t) as the residual of the water balance:

$$ET(t) = \sum_{i=1}^t (P + I + \Delta W - RO - DR) \quad (7)$$

where ET is evapotranspiration, P is precipitation, I is irrigation, ΔW is the change in the soil water storage (water depletion with), RO is surface runoff, and DR is drainage and deep percolation. It is common to express the components of the water balance in length units.

The accuracy of water balance-based ET estimates depends on the measurement errors associated with the methods used to determine the water balance components. P and I can be measured accurately with rain gauges and flow meters, respectively. RO is commonly considered negligible for flat surfaces (e.g., agricultural fields), and DR can be determined from the difference between P or I and ΔW short after precipitation or irrigation events.

The ΔW can be measured with state-of-the-art electromagnetic (EM) proximal soil moisture sensors that operate based on the travel time, capacitance, or impedance measurement principles. Commonly applied EM sensors include the HydraProbe (Stevens Water Monitoring Systems Inc., Portland, OR, USA), the WET150 sensor (Delta-T Devices Ltd., Cambridge, UK), the Digital True TDR-305N sensor (Acclima, Inc., Meridian, ID, USA), the SoilVUE™10 sensor (Campbell Scientific, Inc., Logan, UT, USA),

and the Drill and Drop SM sensor (Sentek Sensor Technologies, Stepney, AUS). While the HydraProbe, WET150, and Digital True TDR-305N sensors are applicable for point measurements with several sensors typically distributed vertically along the soil profile, the SoilVUE™10 and Drill and Drop probes are a new generation of profiling sensors that are available in various lengths up to 1.2 m.

ΔW can be also measured with geophysical proximal sensing methods such as ground penetrating radar (GPR) or electromagnetic induction (EMI). The GPR is applicable for two- and three-dimensional mapping of soil moisture. A transmitter antenna emits high frequency (50 MHz to 3.6 GHz) electromagnetic waves into the subsurface. The waves are refracted, reflected, and scattered and sensed by a single or multiple receiver antennas (Klotzsche et al., 2018). Similar to time domain reflectometry (TDR), the travel time of the EM wave between the transmitter and receiver antennas is measured and the bulk dielectric permittivity of the soil (ϵ_b) that is mainly governed by the volumetric soil water content (θ_v) is inferred. Dielectric mixing models (Anbazhagan et al., 2020) or empirical relationships (e.g., Topp et al., 1980) are then used to convert ϵ_b to θ_v , and subsequently to calculate ΔW .

Electromagnetic induction (EMI) proximal sensing is based on the strong correlation between the soil water content (θ_v) and the measured apparent soil electrical conductivity (EC_a) (Carroll and Oliver, 2005; Doolittle and Brevik, 2014). EMI enables mapping of depth-averaged EC_a at high spatial resolution. To convert the EC_a maps to θ_v , a calibration based on independently measured θ_v at strategically selected locations is required (Corwin and Lesch, 2003). An EMI sensor emits an electromagnetic (EM) field from a transmitter coil. The emitted EM field induces weak eddy current loops into the conducting soil that in turn generate a secondary electromagnetic field that differs in amplitude and phase from the primary field. The primary and secondary fields are sensed with a receiver coil and converted to EC_a (dS m⁻¹). In addition to the spacing between the transmitter and receiver coils, the distance between the coils and land surface, and the coil orientation (i.e., parallel, or perpendicular to the land surface), the magnitude of amplitude and phase differences between the primary and secondary EM fields depends on soil properties including θ_v , the concentration and types of ions in solution, the quantity and mineralogy of clay constituents, and the phase and temperature of the soil water.

Summary

Evapotranspiration (ET) is a major component of the water and energy cycles. It reflects the intricate interactions between soil, vegetation, and weather patterns, which are being altered by the changing global climate. To conserve and sensibly use water resources in Earth's water-starved regions, accurate spatiotemporal information about ET is crucial. Proximal sensing techniques provide powerful means for quantification of ET at the local scale. This chapter provides an overview of state-of-the-art proximal ET sensing techniques including the eddy covariance and scintillometer methods, electromagnetic sensors, ground penetrating radar and electromagnetic induction, which are applied in conjunction with either the land surface energy balance or the terrestrial water balance.

References

- Anbazhagan P, Bittelli M, Palapati RR, and Mahajan P (2020) Comparison of soil water content estimation equations using ground penetrating radar. *Journal of Hydrology* 588: 125039. <https://doi.org/10.1016/j.jhydrol.2020.125039>.
- Burba G (2013) *Eddy Covariance Method for Scientific, Industrial, Agricultural, and Regulatory Applications: A Field Book on Measuring Ecosystem Gas Exchange and Areal Emission Rates*. Lincoln, Nebraska: LI-COR Biosciences. <https://doi.org/10.13140/RG.2.1.4247.8561>.
- Carroll ZL and Oliver MA (2005) Exploring the spatial relations between soil physical properties and apparent electrical conductivity. *Geoderma* 128(3–4): 354–374. <https://doi.org/10.1016/j.geoderma.2005.03.008>.
- Corwin DL and Lesch SM (2003) Application of soil electrical conductivity to precision agriculture: Theory, principles, and guidelines. *Agronomy Journal* 95(3): 455–471. <https://doi.org/10.2134/agronj2003.4550>.
- Doolittle JA and Brevik EC (2014) The use of electromagnetic induction techniques in soils studies. *Geoderma* 223–225: 33–45. <https://doi.org/10.1016/j.geoderma.2014.01.027>.
- Ezzahar J, Chehbouni A, Er-Raki S, and Hanich L (2009) Combining large aperture scintillometer and estimates of available energy to derive evapotranspiration over several agricultural fields in a semi-arid region. *Plant Biosystems* 143(1): 209–221. <https://doi.org/10.1080/11263500802710036>.
- Fisher JB, Melton F, Middleton E, Hain C, Anderson M, Allen R, McCabe MF, Hook S, Baldocchi D, Townsend PA, Kilic A, Tu K, Miralles DD, Perret J, Lagouarde J-P, Waliser D, Purdy AJ, French A, Schimel D, Famiglietti JS, Stephens G, and Wood EF (2017) The future of evapotranspiration: Global requirements for ecosystem functioning, carbon and climate feedbacks, agricultural management, and water resources. *Water Resources Research* 53: 2618–2626. <https://doi.org/10.1002/2016WR020175>.
- Foken T (2006) 50 years of the Monin-Obukhov similarity theory. *Boundary-Layer Meteorology* 119: 431–447. <https://doi.org/10.1007/s10546-006-9048-6>.
- Klotzsche A, Jonrad F, Looms MC, van der Kruk J, and Huisman JA (2018) Measuring soil water content with ground penetrating radar: A decade of progress. *Vadose Zone Journal* 17(1): 180052. <https://doi.org/10.2136/vzj2018.03.0052>.
- LeMone MA, Chen F, Alfieri JG, Tewari M, Geerts B, Miao Q, Grossman RL, and Coulter RL (2007) Influence of land cover and soil moisture on the horizontal distribution of sensible and latent heat fluxes in southeast Kansas during IHOP_2002 and CASES-97. *Journal of Hydrometeorology* 8(1): 68–87. <https://doi.org/10.1175/JHM554.1>.
- McAneney KJ, Green AE, and Astill MS (1995) Large-aperture scintillometry: The homogeneous case. *Agricultural and Forest Meteorology* 76(3–4): 149–162. [https://doi.org/10.1016/0168-1923\(95\)02227-0](https://doi.org/10.1016/0168-1923(95)02227-0).
- Meijninger WML, Hartogensis OK, Kohsiek W, Hoedjes JCB, Zuurbier RM, and De Bruin HAR (2002) Determination of area-averaged sensible heat fluxes with a large aperture scintillometer over a heterogeneous surface—Flevoland field experiment. *Boundary-Layer Meteorology* 105: 37–62. <https://doi.org/10.1023/A:1019647732027>.
- Morillas L, Leuning R, Villagarcía L, García M, Serrano-Ortiz P, and Domingo F (2013) Improving evapotranspiration estimates in Mediterranean drylands: The role of soil evaporation. *Water Resources Research* 49(10): 6572–6586. <https://doi.org/10.1002/wrcr.20468>.
- Oki T and Kanae S (2006) Global hydrological cycles and world water resources. *Science* 313(5790): 1068–1072. <https://doi.org/10.1126/science.1128845>.

- Purdy AJ, Fisher JB, Goulden ML, Colliander A, Halverson G, Tu K, and Famiglietti JS (2018) SMAP soil moisture improves global evapotranspiration. *Remote Sensing of Environment* 219: 1–14. <https://doi.org/10.1016/j.rse.2018.09.023>.
- Solignac PA, Brut A, Selves J-L, Bêteille J-P, Gastellu-Etchegorry J-P, Keravec P, Béziat P, and Ceschia E (2009) Uncertainty analysis of computational methods for deriving sensible heat flux values from scintillometer measurements. *Atmospheric Measurement Techniques* 2(2): 741–753. <https://doi.org/10.5194/amt-2-741-2009>.
- Topp GC, Davis JL, and Annan AP (1980) Electromagnetic determination of soil water content: Measurements in coaxial transmission lines. *Water Resources Research* 16(3): 574–582. <https://doi.org/10.1029/WR016i003p00574>.

Remote sensing of soil moisture

Alexander Gruber^{a,b} and Jian Peng^{c,d}, ^aDepartment of Geodesy and Geoinformation, Vienna University of Technology (TU Wien), Vienna, Austria; ^bDepartment of Earth and Environmental Sciences, KU Leuven, Heverlee, Belgium; ^cDepartment of Remote Sensing, Helmholtz Centre for Environmental Research—UFZ, Permoserstrasse 15, Leipzig, Germany; ^dRemote Sensing Centre for Earth System Research, Leipzig University, Leipzig, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	618
Platforms and instruments	619
Tower and airborne systems	619
Space-borne systems	620
Instrument characteristics	620
Soil moisture retrieval	621
Physical retrieval models	621
Forward modeling for active remote sensing	621
Forward modeling for passive remote sensing	623
Model inversion	624
Physically-based retrieval algorithms	624
(Semi-)empirical retrieval models	624
Root-zone soil moisture estimation	624
High-resolution soil moisture estimation	625
Soil moisture retrieval validation	625
Common missions and products	627
Soil moisture remote sensing applications	627
Conclusion and outlook	627
Acknowledgments	629
References	629

Abstract

Soil moisture is an essential climate variable and knowledge about its state and dynamics is vital for numerous applications, from agricultural drought monitoring to studying land–atmosphere interactions. Remote sensing instruments mounted on satellites are currently our only way to obtain frequent observations of global soil moisture patterns. This chapter provides: (i) an overview of state-of-the-art remote sensing technology that can be used for space-borne soil moisture monitoring, (ii) an account on physical and empirical retrieval models used to generate remotely sensed soil moisture products, and (iii) an outlook into current and future cutting-edge developments.

Key points

- Remote sensing instruments suitable for soil moisture retrieval and their characteristics are introduced.
- Soil moisture retrieval theory and retrieval algorithms are described.
- Common soil moisture missions and products are outlined.
- An overview of common soil moisture remote sensing applications is provided.
- An outlook into current and future developments is presented.

Introduction

In situ sensing techniques, most commonly using dielectric probes of a few centimeters length, are considered to be the most accurate source for soil moisture information. However, in situ soil moisture sensors sample only the water content within a few cubic centimeters of soil. Since soil moisture changes rapidly in both space and time due to its strong dependence on soil texture, topography, and meteorological drivers, ground measurements are very localized and can seldom be used as a proxy for the soil moisture conditions of a larger area (a single point-scale sensor may indeed capture the temporal dynamics that result from large-scale rainfall events and area-wide post-rain dry-downs, yet the absolute magnitude of these readings is only accurate for the small soil volume in which the sensor is placed) (Brocca et al., 2012). Nevertheless, there are many applications, such as agricultural drought monitoring or climate modeling, for which the knowledge of the soil water distribution over very large areas (that is, from

the field to the continental scale) is vitally important (Dorigo et al., 2017). But because in situ measurements are so limited in their spatial representativeness and due to the fact that the global land surface is vast and in many regions hard to access, the use of proximal soil moisture sensing techniques for these applications is not feasible.

Instead, such applications rely almost exclusively on remote sensing, which is the topic of this chapter. Remote sensing refers to the process of acquiring information about an object from a distance without physical contact. The information-carrying media, in this case, are electromagnetic waves, and the sensors used to detect these waves can be placed anywhere from fairly close to the object of interest (ground-based remote sensing) all the way up to space (satellite- or space-borne remote sensing). The latter is of most interest for soil moisture monitoring as modern Earth observing satellites can provide complete scans of almost the entire Earth's surface within just 2–3 days. The downside of measuring soil moisture from space, however, is that these measurements have a rather coarse resolution (an individual measurement represents the average soil moisture conditions of an area that is typically in the order of hundreds to thousands of square kilometers, but for many applications this is not a significant limitation, as we shall see).

This chapter will provide a broad overview of the field of soil moisture remote sensing. Section “Platforms and instruments” discusses the technology used for soil moisture remote sensing and its most important characteristics, and section “Soil moisture retrieval” elaborates on the methods used to turn measurements of electromagnetic waves into soil moisture estimates. Section “Common missions and products” provides an overview of the most common state-of-the-art soil moisture remote sensing products, and section “Soil moisture remote sensing applications” outlines typical applications that rely on remotely sensed soil moisture. Lastly, section “Conclusion and outlook” gives an outlook on novel soil moisture remote sensing research and developments that can be expected in the years to come.

Platforms and instruments

A remote sensing platform is an object or vehicle that carries one or more remote sensing instruments. We can distinguish roughly between ground-based platforms (these can be measurement towers, buildings, rover vehicles, etc.), airborne platforms (i.e., airplanes and UAVs), and spaceborne platforms (i.e., satellites). The remote sensing instruments mounted on such a platform are devices that measure electromagnetic waves. Although some methods exist to retrieve soil moisture information from measurements in the visible and infrared regions of the electromagnetic spectrum, microwave observations are considered to be the most suited for soil moisture retrieval because the dielectric properties of water are distinctly different from those of almost everything else found in natural soils (Ulaby et al., 2014), as will be elaborated later in section “Soil moisture retrieval” (a second advantage of microwave measurements is that they can be taken day and night and are largely unaffected by clouds and atmospheric conditions). For this reason, virtually all modern remote sensing soil moisture products are retrieved from microwave observations, and so this chapter will be limited to this domain. The following sections will provide an overview of the different remote sensing platforms that are used for soil moisture retrieval—or for the development of soil moisture retrieval algorithms—and discuss the various characteristics of remote sensing instruments and how these determine how much information about soil water content we can extract from their measurements.

Tower and airborne systems

As mentioned above, most remotely sensed soil moisture data sets are commonly retrieved from microwave instruments mounted on satellites because they can provide frequent quasi-global soil moisture maps. Although microwave instruments can be (and are) mounted on ground-based measurement towers and airplanes, these measurements are of little practical use for soil moisture retrieval because towers cover only small areas that can be sampled with proximal sensors (which are cheaper and more accurate) and airplane flights cannot be made frequently and repeatedly, and also cover only relatively small regions. Tower-based and airborne microwave measurements are, however, vitally important for the calibration and validation of satellite instruments as well as for the development of the radiative transfer theory used for soil moisture retrieval (Montzka et al., 2020).

Calibration and validation refers to the process of comparing the raw microwave radiance measurements from the satellites to those made by instruments on a tower or on an airplane. This serves to quantify (and, where possible, account for) systematic and random errors in the satellite measurements (hereinafter referred to as biases and uncertainties, respectively). In general, this is done in a controlled laboratory environment, but because this is, for obvious reasons, not possible for space-borne instruments, we have to rely on comparisons with well-calibrated ground-based equipment (i.e., tower and airborne measurements). From these comparisons, we obtain quantitative estimates of biases and uncertainties in the satellite radiance measurements that allow us to make educated guesses about the accuracy that we can expect from soil moisture retrievals made from these measurements (soil moisture retrieval accuracy will, of course, not only be determined by radiance measurement errors, but also by the ability of the retrieval algorithm to convert these radiance measurements into soil moisture estimates; see section “Soil moisture retrieval”) (Merchant et al., 2017).

Most importantly, tower and airplane measurements allow us to create reasonably controlled field environments to study the interactions between microwaves and the land surface, that is, how exactly electromagnetic waves interact with the soil and the vegetation that covers it depending on soil texture, topography, vegetation types, and, most importantly, water content. Such experiments are utilized to derive theoretical forward models to predict the behavior of microwaves given some land surface conditions, which can then be inverted to retrieve soil moisture from actual microwave measurements, as will be discussed in detail in section “Soil moisture retrieval”.

Space-borne systems

Thousands of satellite missions are orbiting the Earth. The ones suited for soil moisture monitoring fall into the category of Earth observation satellites. These are launched into a wide range of orbits (which determine if and how often a certain region of the Earth is revisited) and vary greatly in their instrumentation depending on their intended purpose. Since many satellites are tax-funded, their data are freely available for scientific use. This has greatly fostered the development of novel data sets beyond the initial applications planned for these satellite missions. Remotely sensed soil moisture products, for example, have been retrieved from satellites since the mid-1980s even though the first dedicated soil moisture satellite mission was launched in only 2009 (see section “Common missions and products”).

As mentioned, soil moisture retrievals are predominantly made from microwave instruments. Microwave instrument-carrying satellites can be categorized into active systems (radar systems) and passive systems (radiometers). Radar systems carry antennae that actively emit microwaves and then measure the radiation that has been reflected back by the Earth’s surface, whereas radiometers passively observe the microwave radiation that is emitted by the Earth’s surface due to its physical temperature (Ulaby et al., 2014), as illustrated in Fig. 1. Radar instruments can be further divided into real-aperture radar (or “radar”, for short) and synthetic-aperture radar (SAR). The latter combine multiple successive measurements made by a relatively small radar antenna along the flight path through sophisticated processing techniques so that each measurement appears to have been taken by a much larger synthetic antenna. The reason for doing this is that the spatial resolution of radar measurements is limited by the antenna length (at best a few meters) and the satellite altitude (typically around 700–800 km). These two parameters effectively limit the measurements of real-aperture radar systems and radiometers to a spatial resolution of about 10^2 – 50^2 km², while SAR processing can yield spatial resolutions of a few square meters, albeit with a *much* lower signal-to-noise ratio. For many applications, however, the time interval between observations and their accuracy is more important than spatial resolution, which often outweighs the benefit of the higher spatial resolution of SAR instruments (see section “Soil moisture remote sensing applications”).

The orbit of a satellite together with the antenna geometry and observation strategy determine how frequently a given location is observed (radar systems and radiometers are often designed to observe as large an area as possible whereas SAR scans often have to be focused on specific regions of interest due to technical limitations such as power supply versus consumption). Most satellite missions used for soil moisture retrieval are so-called sun-synchronous polar orbiters. This means that their orbit follows the Earth’s revolution and movement around the sun, and hence that each point on the Earth is revisited at about the same time during a day, every few days apart depending on the latitude (areas near the poles are observed more frequently than areas near the equator). Modern radar systems and radiometers have typical revisit times of about 1–3 days, whereas SAR scans can be weeks to months apart, depending on the observation mode (see section “Common missions and products”).

Instrument characteristics

Satellite-based microwave instruments have three important characteristics that influence how well they are suited for soil moisture retrieval made from these instruments (Ulaby et al., 2014). These are the microwave frequency and polarization, and the antenna incidence angle.

The microwave domain of the electromagnetic spectrum covers a frequency range from about 300 MHz to 300 GHz (about 1 mm to 1 m wavelength). Various microwave bands have been used for soil moisture retrieval (Dorigo et al., 2017) including K_a-band (0.5–1.5 mm wavelength), K_u-band (1.5–2.5 cm), X-band (2.5–3.7 cm), C-band (3.7–7.5 cm), and L-band (15–30 cm). In general, the longer the wavelength, the less the radiation is affected by vegetation covering the soil and the deeper it can penetrate into the ground, thus making it more sensitive to soil moisture. In contrast, the shorter the wavelength, the better is the spatial

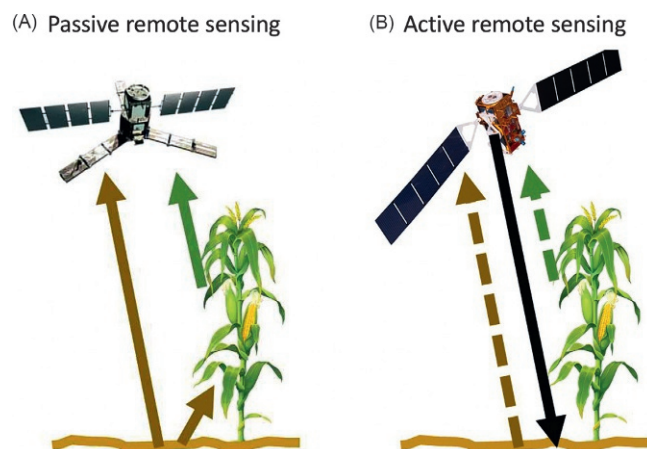


Fig. 1 Schematic illustration of (A) passive remote sensing, and (B) active remote sensing. Courtesy Gabrielle De Lannoy.

resolution. Typical spatial resolutions for different frequency bands are roughly in the order of 10^2 – 25^2 km², 15^2 – 25^2 km², 20^2 – 40^2 km², 25^2 – 50^2 km², 35^2 – 50^2 km², for K_a-, K_u-, X-, C-, and L-band, respectively. The benefit of the greater soil moisture sensitivity of longer wavelengths usually outweighs the benefit of a slightly improved spatial resolution, which is the reason why most state-of-the-art soil moisture products are retrieved from C-band and L-band measurements (see section “Common missions and products”).

Microwave antennae measure radiation at a particular polarization, that is, oscillating either parallel to the Earth’s surface (horizontally polarized; H) or perpendicular to it (vertically polarized; V). Radiometers measure H or V polarized radiation (or both) emitted from the Earth’s surface itself while radar systems and SARs emit the radiation at a chosen polarization and then measure the reflected radiation either at the same polarization (co-polarized; VV or HH) or at the other respective polarization (cross-polarized; VH or HV). Vertically (co)polarized measurements have been found to be more sensitive to soil moisture than horizontally (co)polarized measurements, and co-polarized measurements are more sensitive to soil moisture than cross-polarized measurements. Nevertheless, because cross-polarized measurements are more sensitive to vegetation cover, they can potentially be used to correct for unwanted vegetation signal components in co-polarized measurements.

Lastly, microwave instruments used for soil moisture retrieval typically point at the Earth’s surface at a particular angle, referred to as the incidence angle, which can be either fixed for all measurements (e.g., at 40 degrees) or vary across the antenna swath (e.g., 25–65 degrees). Microwave measurements exhibit a strong dependence on the incidence angle due to structural land surface characteristics, in particular soil roughness and vegetation geometry. Instruments that observe at a fixed incidence angle, therefore, have the advantage that all measurements are influenced by these characteristics in the same way, which may make it easier to account for these effects in the soil moisture retrieval algorithm. On the other hand, taking measurements at a range of incidence angles provides additional degrees of freedom that may be used to derive (and hence correct for) the incidence angle relationships.

Soil moisture retrieval

A plethora of methods exists to extract the soil moisture signal from the microwave measurements made by remote sensing instruments, referred to as soil moisture retrieval. This section provides a broad overview of the key underlying principles that most existing retrieval algorithms share. In addition, we will outline state-of-the-art methods used to push the boundaries of the soil moisture information we *can* observe from space-borne microwave remote sensing. For example, how to extract root-zone soil moisture estimates from microwave measurements that are only sensitive to the soil surface (the penetration depth of these measurements into the soil is roughly in the order of the wavelength, i.e., a few centimeters depending on the microwave band, and can vary significantly in space and time because it depends on the soil structure and the soil water content) and how to resolve sub-pixel soil moisture variation within coarse resolution satellite measurements.

Physical retrieval models

Physically-based soil moisture retrieval relies on the statistical inversion of forward models that describe what the remote sensing instruments measure (microwave radiation) as a function of the land surface conditions (soil characteristics and vegetation cover). Although forward models differ for active and passive remote sensing instruments (remember that active sensors emit their own signal and measure their return reflected by the Earth’s surface, also referred to as backscatter, whereas passive sensors measure the microwave signal naturally emitted by the Earth’s surface, referred to as brightness temperature) they comprise the same two components: one that describes what happens within the soil and one that describes what happens within the vegetation covering it.

Forward modeling for active remote sensing

Active sensors measure the area-normalized change in magnitude of a microwave signal that they emitted and that has returned to the sensor after having been reflected by the Earth’s surface (Ulaby et al., 2014). This quantity is referred to as backscatter (σ^0). The backscatter depends on the geometric and the dielectric properties of the reflecting object (i.e., the Earth’s surface) and is modeled separately for the soil and for the vegetation.

Soil backscattering

The dielectric properties of the soil (expressed in terms of the dielectric constant ϵ) determine how much of an incoming microwave signal is reflected at the soil–atmosphere interface and how much penetrates into the soil or is absorbed. Since the soil is a mixture of materials (mineral components, organic components, water, and air) its dielectric properties depend on the dielectric properties of its components. Precise theoretical modeling of the soil dielectric properties at the microscale using dielectric mixing models is possible but very complex. More importantly, given the immeasurable diversity in possible soil composition and structure, purely theory-based soil dielectric models contain a multitude of parameters and variables that are impossible to determine and parameterize in practice. Therefore, semi-empirical models have been developed that mimic the behavior of these models, but only require a few variables and parameters that are known to be the most significant drivers of the soil dielectric constant and are relatively easy to obtain (Dobson et al., 1985). These are usually the microwave frequency, soil temperature, salinity, soil texture, and, of course, soil water content. As mentioned, water has a very high dielectric constant while all other soil components have a very low one. Consequently, the dielectric constant, and hence the backscatter, of the soil are predominantly determined by soil moisture (the higher the soil water content, the greater is the backscatter). The effect of soil texture influences the soil dielectric

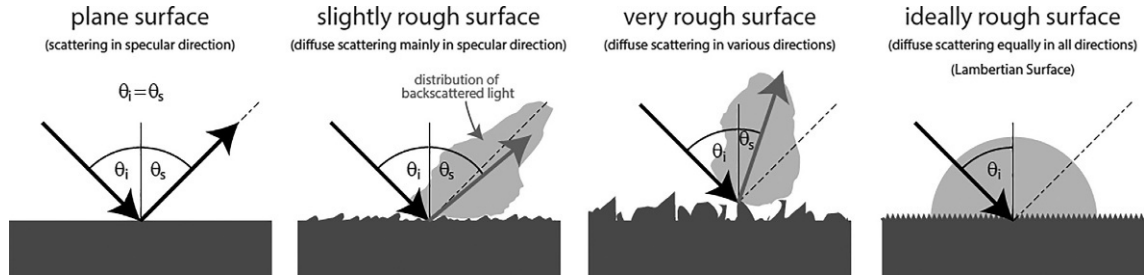


Fig. 2 Schematic illustration of rough surface scattering. Courtesy Raphael Quast.

constant indirectly by limiting its overall dynamic range. This is because sandy soils can hold much more water per volume than soils with higher clay content. The effects of soil temperature and salinity are, for soil moisture remote sensing purposes, often considered to be negligible. The general structure of these models can be written as:

$$\varepsilon = f(f, T, S, s, c, \Theta) \quad (1)$$

where f is the microwave frequency; T is the temperature; S is the salinity; s and c are the sand and clay fraction, respectively, which, together, represent the soil texture; and Θ is the soil moisture. The most commonly used soil dielectric models are the so-called Dobson model (Dobson et al., 1985), the Mironov model (Mironov et al., 2013), and the Wang and Schmugge model (Wang and Schmugge, 1980). For a detailed presentation of their exact model structures, we refer the reader to the respective references.

While the dielectric properties of the soil affect how much of the microwave signal is reflected by the soil surface, the geometric properties of the soil, more precisely its “roughness”, determine the directionality of these reflections (recall that microwave instruments typically observe at an incidence angle that can be as much as 65 degrees). The rougher the surface, the more isotropic the soil scattering becomes, as illustrated in Fig. 2. This behavior depends on the microwave frequency: the scattering will converge to an isotropic scattering behavior as the surface roughness, expressed in terms of the surface height standard deviation and its horizontal auto-correlation length, approaches the wavelength (Ulaby et al., 2014).

Taken together, the backscattering behavior of soils is described in so-called soil backscatter models as a function of dielectric and geometric properties. The only relevant physically-based soil backscattering models worth mentioning here are the Integral Equations Method (IEM) model (Fung et al., 1992), and its extensions, the Advanced IEM (AIEM) model and the Improved IEM (I²EM) model (Ulaby et al., 2014) (the soil dielectric constant used as an input to these models, however, is typically obtained from a semi-empirical soil dielectric model rather than from a physical one; see above). Such a soil backscattering model can be expressed as:

$$\sigma_s^0 = f(\theta, \varepsilon, h_s, l) \quad (2)$$

where σ_s^0 is the soil backscatter; θ is the incidence angle; and h_s and l are the standard deviation and auto-correlation length of the surface height profile, respectively. Given the highly complex, non-linear structure of these models, we omit their lengthy presentation here and refer the reader to the respective references.

Vegetation scattering

Most global ecoregions, from rainforests through savannas to boreal forests, are covered with some form of vegetation which needs to be accounted for in the backscatter forward modeling. Physically-based forward models do this by means of radiative transfer theory. Radiative transfer models (RTMs) describe the net energy change of the microwave radiation as it passes through the vegetation both while traveling from the sensor towards the soil surface and while traveling back to the sensor after having been reflected by the soil surface (as described by the soil backscatter model). This energy change occurs due to the scattering of the microwave radiation within the vegetation structure and signal absorption. Both of these effects depend strongly on the overall biomass, the structural composition of the vegetation, and the vegetation’s water content (Ulaby et al., 2014).

It is not hard to imagine that a geometrically accurate representation of a vegetation cover and its water content distribution is, for areas as large as the satellite radar footprints, virtually impossible to build. Even though attempts have been made to do so (one example is the Michigan Microwave Canopy Scattering (MIMICS) model that describes the vegetation as a collection of dielectric cylinders, representing trunks and branches, and discs, representing leaves), nearly all vegetation scattering models that are used for remote sensing purposes use a simplified description of vegetation. The most popular such description is the so-called Water Cloud Model (WCM), which is a first-order solution to the radiative transfer equations and gets its name from (figuratively) representing the heterogeneous vegetation cover of a very large area as a homogeneous layer of water droplets (Attema and Ulaby, 1978). This simplified vegetation layer is fully characterized by only two quantities, the scattering albedo (ω) and the vegetation optical depth (τ), and models only the two most dominant components of the microwave scattering behavior, which are (i) the signal that is scattered back to the sensor by the vegetation before reaching the soil surface, and (ii) the signal that penetrates through the vegetation and is reflected back to the sensor by the soil surface. So-called interaction terms, which describe microwave radiation that has been scattered multiple times within the vegetation canopy and at the soil surface before traveling back to the sensor, are

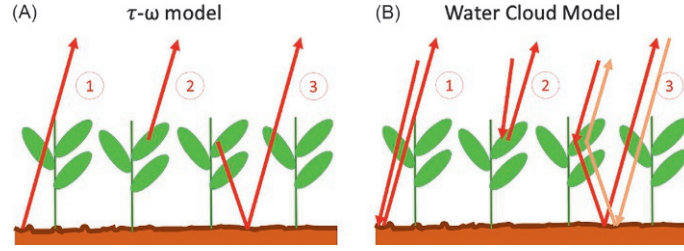


Fig. 3 Illustration of the components modeled in (A) the τ - ω model: (1) upwelling soil emission attenuated by the vegetation canopy, (2) direct upwelling vegetation emission, and (3) downwelling vegetation emission reflected by the soil and attenuated by the canopy layer; and (B) the WCM: (1) backscatter from the soil surface, (2) backscatter from the vegetation canopy, and (3) backscatter from radiation that was reflected once from both the vegetation and the soil surface, respectively. Note that the backscatter interaction term (3) of the WCM is commonly considered negligible. Courtesy Gabrielle De Lannoy.

commonly assumed to be negligible and thus usually not modeled, even though recent research has made attempts to account for these interaction terms in soil moisture retrieval algorithms (Quast et al., 2019). The WCM can be written as:

$$\sigma^0 = \frac{3\omega}{4} \cos \theta \left(1 - e^{-\frac{2\tau}{\cos \theta}} \right) + \sigma_s^0 e^{-\frac{2\tau}{\cos \theta}} \quad (3)$$

The factor $\frac{3}{4}$ arises by assuming that the vegetation exhibits Rayleigh scattering. Alternatively, the directionality of the vegetation scattering could be described more accurately by a so-called phase function (Quast et al., 2019). The scattering components modeled in the WCM are illustrated in Fig. 3.

Forward modeling for passive remote sensing

The following sections will summarize the most relevant principles for the modeling of passive microwave signatures from the land surface. The interested reader can find a detailed and more comprehensive account on that subject matter in Wigneron et al. (2017).

Soil emission

According to Planck's law, any object with non-zero physical temperature emits radiation over the entire electromagnetic spectrum. Passive microwave sensors measure the so-called brightness temperature T_b , which is the microwave radiation at a particular frequency and polarization, and can be related to the soil's effective temperature T as:

$$T_b(\theta) = e(\theta) * T \quad (4)$$

where e is the soil emissivity, which is related to the soil reflectivity (R) as:

$$e(\theta) = 1 - R(\theta) \quad (5)$$

which is related to the soil backscatter in that it is its integral over the scattering hemisphere (Ferrazzoli et al., 1989). The soil emissivity thus holds the information about soil dielectric and geometric properties and is modeled, just as the soil backscatter, by means of a soil dielectric model and a soil roughness model. While the soil dielectric models used for emissivity modeling are largely the same as those used for modeling active microwave signatures, the passive soil moisture remote sensing community most commonly uses (semi)empirical soil roughness models such as the HQN model rather than the physical models that were outlined above for soil backscatter modeling (although the active remote sensing community has spent more effort investigating these physical soil roughness models, they are not commonly used for soil moisture retrieval from active sensors either, most likely because they are highly non-linear and virtually impossible to parameterize accurately for the large areas covered by space-borne microwave antenna footprints).

Vegetation emission and scattering

Virtually all commonly used physically-based retrieval algorithms for passive microwave observations use a zero-order solution to the radiative transfer equations, which is commonly known as the τ - ω model. Like the WCM, it is fully described by the scattering albedo and the vegetation optical depth. Similar to the WCM, it represents three components contributing to the total observed brightness temperature, which are (i) the direct upwelling vegetation emission, (ii) the downwelling vegetation emission reflected by the soil and attenuated by the canopy layer, and (iii) upwelling soil emission attenuated by the vegetation canopy, as illustrated in Fig. 3:

$$T_b = (1 - \omega) \left(1 - e^{-\frac{\tau}{\cos \theta}} \right) T_c + (1 - \omega) \left(1 - e^{-\frac{\tau}{\cos \theta}} \right) e^{-\frac{\tau}{\cos \theta}} R T_c + (1 - R) e^{-\frac{\tau}{\cos \theta}} T_g \quad (6)$$

where T_c and T_g are the effective soil and vegetation temperatures, respectively. Typically, these are assumed to be equal and modeled as a single, effective temperature T_e . Note that the brightness temperature, and hence also ω , τ , and R , are different for H and V polarizations. Although the τ - ω model does not model first- and higher-order (i.e., multiple scattering) effects, it is commonly assumed that these effects can be accounted for to a sufficient degree of accuracy if the scattering albedo and the

vegetation optical depth are calibrated as “effective parameters”, that is, by specifically calibrating them to the modeled land surface environment (Wigneron et al., 2017).

Model inversion

Physically-based backscatter and brightness temperature models are too complex to be inverted analytically. Therefore, soil moisture is retrieved from these models using non-linear optimization algorithms, which aim to minimize a cost function of the following form:

$$x = \min_x \sum_i (\gamma_i^{\text{obs}}(\mathbf{x}) - \gamma_i^{\text{mod}}(\mathbf{x}))^2 \quad (7)$$

where γ^{obs} and γ^{mod} are the observed and modeled backscatter or brightness temperature, respectively, i may refer to different polarizations, incidence angles, or other parameter values that yield degrees of freedom, and the vector $\mathbf{x}$ holds the forward model variables that are being optimized, that is, soil moisture (some algorithms not only estimate soil moisture but also other state variables such as vegetation optical depth or other related dynamic parameters). These cost functions are complex, and the retrievals are thus quite sensitive to the choice of the optimization algorithm and the initial values.

Physically-based retrieval algorithms

The sections above touched upon the basic principles for the physical modeling of active and passive microwave emission and scattering. A multitude of soil moisture retrieval algorithms has been developed based on these principles. These differ in the choices made for individual model components (e.g., soil dielectric models), auxiliary data (e.g., effective soil and canopy temperatures), simplifying assumptions (e.g., neglecting polarization dependency), and inversion algorithms (e.g., solving for soil moisture only or for both soil moisture and vegetation optical depth). A complete account of these algorithms is beyond the scope of this chapter; the interested reader is referred to Wigneron et al. (2017).

(Semi-)empirical retrieval models

While most common passive-microwave-based soil moisture products are retrieved using the physically-based retrieval algorithms described above, this is not the case for active-microwave-based products. Various empirical approaches have been proposed that fit simple non-linear functions experimentally to backscatter and the quantities anticipated to be its most dominant drivers, i.e., soil moisture, roughness, radar incidence angle, soil texture, etc. (e.g., Oh et al., 1992) or even simply between backscatter and the soil dielectric constant (Loew et al., 2006). The by far most common retrieval algorithm for active-microwave-based soil moisture retrievals is the so-called TU Wien change detection method (Wagner et al., 1999), which is also the only algorithm that has been implemented for the operational production of soil moisture products, both from real-aperture radar systems and from SAR instruments. This method retrieves soil moisture as the degree of soil saturation (Θ_s) by scaling the observed backscatter on a particular day between the historically lowest and highest observed backscatter observations (referred to as dry reference, σ_{dry}^0 , and wet reference, σ_{wet}^0), which are assumed to represent completely dry and completely saturated conditions, respectively:

$$\Theta_s = \frac{\sigma^0 - \sigma_{\text{dry}}^0}{\sigma_{\text{wet}}^0 - \sigma_{\text{dry}}^0} \quad (8)$$

This algorithm further exploits multi-incidence angle observations to correct for the seasonally varying vegetation component in the backscatter signal as part of an incidence angle normalization step (Wagner et al., 1999).

Lastly, several probabilistic and machine-learning-based approaches have been tested for both active-microwave-based and passive-microwave-based soil moisture retrieval. One such approach, based on artificial neural networks, has been implemented recently for operational near-real-time soil moisture retrieval from the SMOS satellite (Rodríguez-Fernández et al., 2015).

Root-zone soil moisture estimation

As noted earlier, microwave measurements are only sensitive to the top few centimeters of the soil. Therefore, soil moisture retrievals obtained using the methods described above also only represent the water content in this shallow surface soil layer. Many applications, however, are more interested in the water content of the plant root-zone (see section “Soil moisture remote sensing applications”). Several strategies exist to derive root-zone soil moisture estimates further from surface soil moisture retrievals. Two of these strategies are worth mentioning here.

The first is the calculation of the so-called Soil Water Index (SWI; Wagner et al., 1999) which serves as a proxy for deeper-layer soil moisture, albeit without a clear definition of the exact depth of the soil layer it represents. The SWI is calculated by applying an exponential filter to the surface soil moisture estimates, which mimics the natural infiltration process:

$$SWI(t_n) = \frac{\sum_i^n \theta(t_i) e^{\frac{t_n - t_i}{T}}}{\sum_i^n e^{\frac{t_n - t_i}{T}}} \quad (9)$$

where θ is the remotely sensed surface soil moisture retrieval, and T is the so-called characteristic time length in the unit of days, which represents infiltration duration. The larger T is, the deeper is the soil layer that the SWI represents. Although the SWI has been found to correlate well with in situ measurements of deeper layer soil moisture, a specific T value cannot be associated with an exact soil depth because water infiltration properties depend strongly on soil characteristics, topography, and climatic conditions. Nevertheless, Paulik et al. (2014) have found that T values ranging from 0 to 50 represent soil layers from 0 to 50 cm in a relatively linear fashion (although with a significant spread), which can serve as a rough “rule of thumb”.

An alternative approach for obtaining root-zone soil moisture estimates from surface soil moisture retrievals is through hydrological modeling, that is, by simulating the physical processes of plant and soil evaporation, surface runoff, infiltration, and so on. However, such an approach requires a plethora of auxiliary input data such as measurements of temperature, radiation, and humidity as well as information on soil texture, topography, and land cover. This, therefore, starts to leave the realm of remote sensing and enters the field of model-data integration. The most common way to integrate satellite surface soil moisture retrievals (or raw satellite radiance measurements) with soil hydrological modeling is through data assimilation, where the satellite data are used to update soil moisture predictions of a numerical model by “pulling them in the right direction” (Reichle et al., 2002). The most popular example for an operational root-zone soil moisture product that has been generated by integrating remote sensing observations with a global land surface model is the so-called SMAP Level-4 (L4) product (Reichle et al., 2017a,b).

High-resolution soil moisture estimation

High-resolution soil moisture products are in high demand by the user community for a variety of local and regional applications. Active microwave platforms, particularly SAR sensors, provide a direct way to obtain high-resolution soil moisture estimates. Currently, Sentinel-1C-band backscatter data are used to generate high spatial resolution (1 km) soil moisture at regional/continental scales (Bauer-Marschallinger et al., 2019). Relevant operational products have also been released by the Copernicus Global Land service. Nevertheless, soil moisture retrieval algorithms for SAR data still need to be improved to account better for the effects of vegetation, surface roughness, and incidence angle. An alternative way to obtain high spatial resolution soil moisture is to downscale existing coarse-resolution soil moisture products. High-resolution optical and infrared data, radar backscatter, or prior knowledge of soil moisture variation have been used to improve the spatial resolution of various soil moisture products. The general concept of these downscaling approaches is to fuse coarse-resolution soil moisture and fine-resolution auxiliary variables by building either statistical or physical scaling models (Fig. 4; Peng et al., 2017). Several fused high-resolution soil moisture products have been recently developed and published, such as a combined SMAP/Sentinel product (1 km/3 km), and an ASCAT/Sentinel-1 product (1 km). In addition, the integration of multi-satellite platforms offers the opportunity to produce sub-daily soil moisture products, such as the combination of Metop-A, -B, and -C satellites carrying the Advanced Scatterometer (ASCAT), and the combination of optical/thermal geostationary satellites with microwave satellites. Despite these advances, novel fusing algorithms (e.g., machine learning) and modeling frameworks (data assimilation) still need to be developed to combine multi-source data sets better to generate long-term high spatial resolution soil moisture products. Future satellite missions will even provide new capabilities to provide soil moisture at both high spatial and temporal resolutions.

Soil moisture retrieval validation

The validation of soil moisture retrievals from remote sensing observations faces a particular challenge, which is the difficulty to obtain “ground-truth” estimates of the exact soil moisture conditions within the large satellite footprints (which are in the order of hundreds of square kilometers) to compare the retrievals against. Extensive field campaigns have been carried out to obtain highly accurate soil moisture proxies at the satellite footprint scale, but these are labor and time-intensive, and can only be done infrequently and over very small areas. However, retrieval accuracy varies substantially across the globe depending on the retrieval algorithm’s ability to account for specific land surface characteristics (such as vegetation, topography, landscape heterogeneity, and so on). Consequently, field campaigns or heavily equipped in situ measurement sites will never be able to provide assessments of satellite retrieval accuracy that are representative of all relevant geographic conditions.

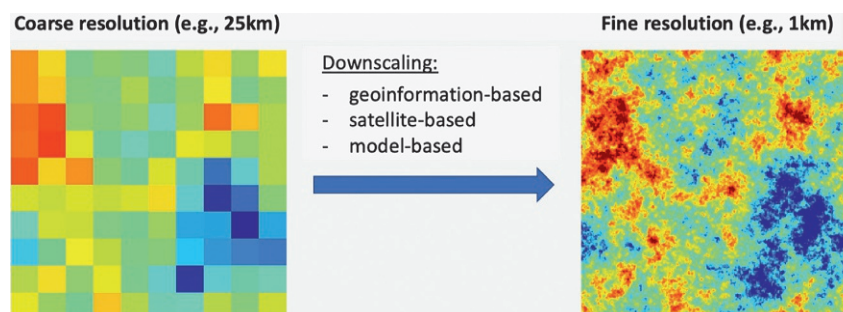


Fig. 4 Illustration of the downscaling of coarse-resolution soil moisture products using fine-resolution auxiliary data.

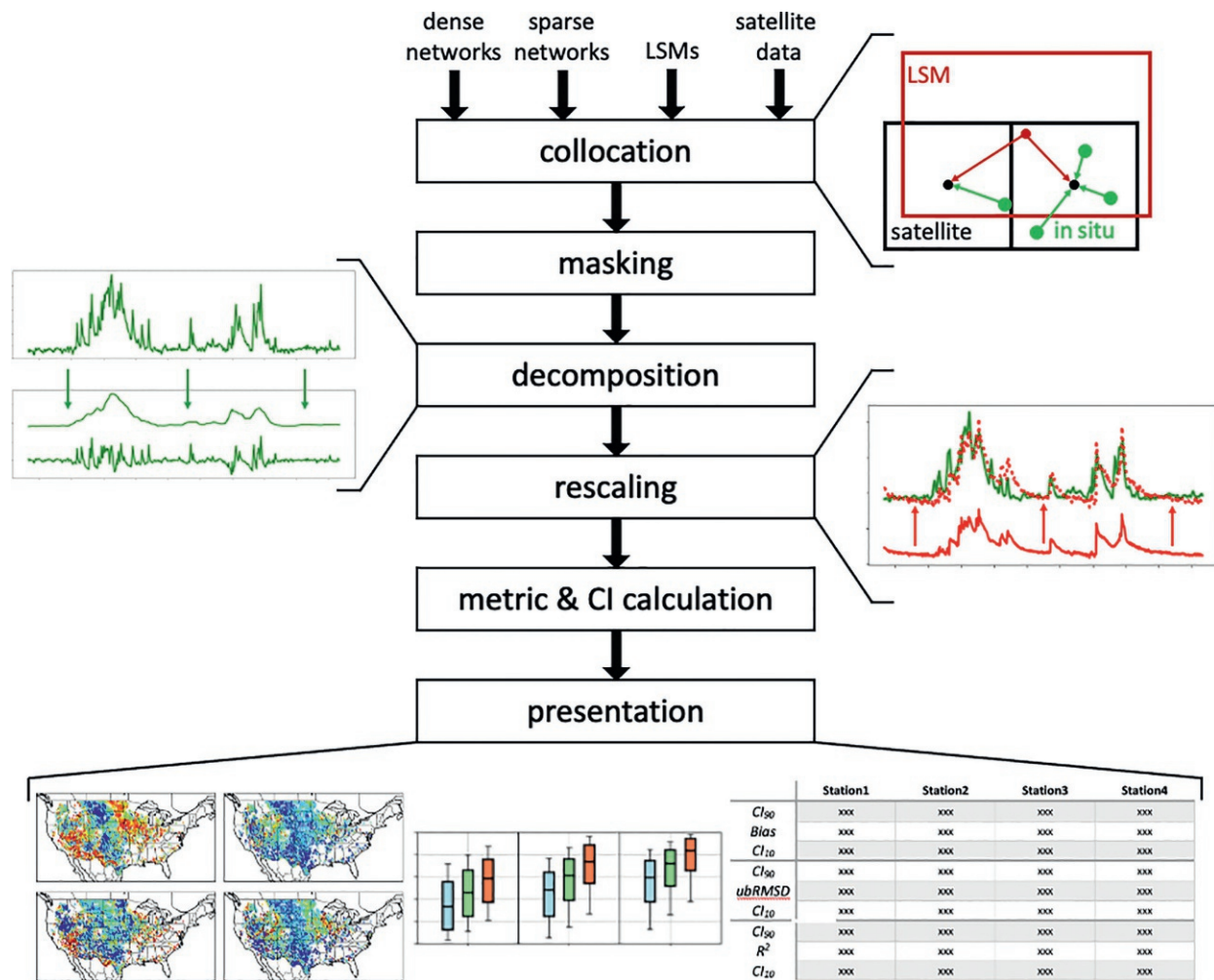


Fig. 5 Soil moisture validation protocol. Modified from Gruber A, De Lannoy G, Albergel C, Al-Yaari A, Brocca L, Calvet J-C, Colliander A, Cosh M, CrowW, Dorigo W, Draper C, Hirschi M, Kerr Y, Konings A, Lahoz W, McCol K, Montzka C, Muñoz-Sabater J, Peng J, Reichle R, Richaume P, Rüdiger C, Scanlon T, van der Schalie R, Wigneron J-P and Wagner W (2020) Validation practices for satellite soil moisture retrievals: What are (the) errors? *Remote Sensing of Environment* 244: 111806. <https://doi.org/10.1016/j.rse.2020.111806>.

The soil moisture remote sensing community has developed various strategies to tackle this issue, which are described in detail in Gruber et al. (2020). In essence, satellite soil moisture validation consists of various layers, that is, comparisons against reference data sets from different sources with varying quality and availability. These range from assessments with a few heavily-equipped ground reference sites (so-called core validation sites) that provide highly accurate estimates of accuracy yet at only a few locations worldwide, to inter-comparisons with lower-quality products that serve more as consistency checks but that can be done globally. The different characteristics of these reference data sets necessitate various pre-processing steps, such as spatial and temporal collocation or the removal of climatological differences, as illustrated in Fig. 5. For more detail about the steps involved, we refer to Gruber et al. (2020), but two vital ideas in this process are worth elaborating here.

First, the exact soil volume that a space-borne microwave instrument “sees” is more of a factitious concept rather than an actual volume due to the vast extent of the antenna footprints, which exhibit a microwave signal gradient from the center outwards, and the varying signal penetration depth. Therefore, it is difficult to obtain a reliable estimate for the absolute amount of water within an only approximately defined soil volume. But in many cases, it is sufficient to know *how much the soil moisture changes* even though the absolute amount of water is uncertain. Fortunately, soil moisture changes are not very localized. If it rains, it rains over large areas and if the sun dries out the soil, it does so over large areas. Accordingly, point measurements of single in situ stations can represent large-scale dynamics fairly accurately. Research has shown that as few as 1–5 carefully placed in situ sensors can represent satellite footprint scale soil moisture dynamics with <5% uncertainty (Brocca et al., 2012) (different locations within a satellite footprint may be more or less representative for the footprint-average soil moisture dynamics depending on the landscape and climatic conditions; locations that are more representative are commonly referred to as temporally stable locations; Vachaud et al., 1985). This allows satellite validation studies to utilize so-called sparse networks, that is, ground measurement networks with sparsely distributed in situ sensors, to assess soil moisture retrieval accuracy at least in terms of their skill in capturing temporal soil moisture

dynamics. For estimates of absolute soil moisture retrieval accuracy, the best solutions are still heavily-equipped sites or dedicated field campaigns.

The second key idea to satellite soil moisture validation that allows us to obtain spatially complete estimates of soil moisture retrieval accuracy is that one does not necessarily need a high-quality reference data set as long as one has several reference data sets whose errors are independent from each other. Specifically, a statistical method called triple collocation analysis (TCA) can estimate the individual *random* errors of three data sets x , y and z , neither of which needs to be error-free (or even highly accurate, for that matter) as:

$$\begin{aligned}\sigma_{\epsilon_x}^2 &= \sigma_x^2 - \frac{\sigma_{xy}\sigma_{xz}}{\sigma_{yz}} \\ \sigma_{\epsilon_y}^2 &= \sigma_y^2 - \frac{\sigma_{yx}\sigma_{yz}}{\sigma_{xz}} \\ \sigma_{\epsilon_z}^2 &= \sigma_z^2 - \frac{\sigma_{zx}\sigma_{zy}}{\sigma_{xy}}\end{aligned}\quad (10)$$

where σ_{ϵ_i} is the random error variance of data set i , σ_i^2 is the variance of data set i , and σ_{ij} is the covariance between data sets i and j . These individual estimates of random error variance are truly independent from those of the respective others and remain unchanged even if one or both of the other data sets are exchanged with another data source, provided that all statistical assumptions relevant to the method are respected. For a detailed description of TCA, we refer to Gruber et al. (2016). Note, however, that TCA only estimates *random* errors. There is still no method available that can accurately determine absolute biases in satellite soil moisture retrievals over larger areas.

Common missions and products

An overview of the most commonly used remotely sensed soil moisture products and information about what satellite mission/sensor they have been retrieved from, their temporal coverage and resolution, their spatial resolution, and where to access them is provided in Table 1. The following section will provide an overview of common applications for these products.

Soil moisture remote sensing applications

Satellite soil moisture products have been widely used to serve a variety of disciplines, such as hydrology, Numerical Weather Prediction (NWP), climate modeling, ecology, and agriculture (see Fig. 6; Dorigo et al., 2017). For some disciplines, the application of satellite soil moisture has become mature and widespread. For example, satellite soil moisture is widely used for the evaluation of global climate model performance and to understand land–atmosphere interactions better. Based on ASCAT and AMSR-E soil moisture products, for example, Taylor et al. (2012) demonstrated a negative feedback between soil moisture and precipitation, with afternoon rainfall being more likely to occur on dry soils across the globe. Miralles et al. (2014) found that the declining trends in global evapotranspiration and soil moisture were related to El Niño and La Niña cycles. Satellite soil moisture products, particularly ASCAT, SMOS, and SMAP, have also been used to improve NWP by assimilating them into various NWP systems (de Rosnay et al., 2020). In addition, satellite soil moisture plays an important role in a range of hydrological applications like runoff simulation, flood forecasting, and drought monitoring. However, it is still in the initial stage for some other applications (e.g., precision agriculture, forestry research), which is partly due to the limitations of current satellite products in terms of spatial and temporal resolution, and their limited ability to represent root zone soil moisture. In general, it is challenging to meet all the requirements of the broad user communities. One typical example is the accuracy requirement: large-scale climate-related applications and NWP only need accurate soil moisture temporal dynamics, while hydrological modeling and ecosystem monitoring require accurate absolute soil moisture values. Moreover, the requirements for soil moisture depth, temporal resolution, and quality information also vary with different applications (Table 2). A recent review by Peng et al. (2021) comprehensively summarizes gaps between current soil moisture products and various applications. To improve the characteristics of satellite soil moisture products further, close collaboration between data producers and various user groups is required. In addition, there is a strong demand for high spatial and temporal resolution soil moisture products in many disciplines. In the future, many applications will benefit from the availability of high-resolution soil moisture products, and more applications will emerge.

Conclusion and outlook

Remote sensing is an invaluable tool for obtaining frequent quasi-global soil moisture estimates. The number of satellites carrying instruments suitable for soil moisture retrieval has increased rapidly over the past decades, including two dedicated soil moisture missions that were launched in 2009 (SMOS) and 2015 (SMAP), respectively. Most of these satellites provide soil moisture estimates at a spatial resolution of 25²–50² km² with a global coverage every 2–3 days. Higher-resolution (1²–10² km²) products are available from synthetic-aperture radar instruments or obtained using downscaling algorithms or data assimilation techniques. Merging different satellite products can yield daily or even sub-daily soil moisture estimates almost everywhere on the globe. With

Table 1 Publicly available satellite-derived global soil moisture products.

<i>Institution</i>	<i>Temporal coverage</i>	<i>Temporal resolution</i>	<i>Grid spacing</i>	<i>Sensor</i>	<i>Data link</i>
Vrije Universiteit Amsterdam	1978–1987	2–3 days	0.25 deg	SMMR	https://www.geo.vu.nl/~jeur/lprm/
Vrije Universiteit Amsterdam	1987–1999	2–3 days	50 km	SSM/I	https://www.geo.vu.nl/~jeur/lprm/
ESA	1991–2007	1–2 days	25/50 km	ERS AMI WS	https://earth.esa.int/web/sppa/activities/multi-sensors-timeseries/scirocco/
Vrije Universiteit Amsterdam	1998–2015	2–3 days	50 km	TRMM-TMI	https://www.geo.vu.nl/~jeur/lprm/
Vrije Universiteit Amsterdam	2002–2011	1–3 days	25 km	AMSR-E	https://www.geo.vu.nl/~jeur/lprm/
NASA	2002–now	Daily	25 km	AMSR-E, AMSR2	https://nsidc.org/data/au_land/versions/1
CESBIO	2003–2011	Daily	15/25 km	SMOS, AMSR-E	https://www.catds.fr/Products/Available-products-from-CPDC
EUMETSAT H-SAF	2007–now	1–2 days	12.5/25/50 km	ASCAT	http://hsaf.meteoam.it/
CESBIO	2010–now	1–2 days	25 km	SMOS	https://www.catds.fr/Products/Available-products-from-CPDC
ESA	2010–now	1–2 days	15 km	SMOS	https://smos-diss.eo.esa.int/oads/access/
NASA	2011–2015	7 days	1 deg	Aquarius	http://nsidc.org/data/aquarius/
JAXA	2012–now	2–3 days	50 km	AMSR2	https://suzaku.eorc.jaxa.jp/GCOM_W/data/data_w_index.html
NASA	2015–now	1–2 days	3/9/36 km	SMAP	https://nsidc.org/data/smap/smap-data.html
NASA	2015–now	1–2 days	1/3 km	SMAP/Sentinel-1	https://nsidc.org/data/smap/smap-data.html
ESA	1978–2019	Daily	0.25 deg	Merged Active + Passive Microwave Sensors (ESA CCI)	http://www.esa-soilmoisture-cci.org/
NOAA	2012–now	6 h	0.25 deg	Merged Active + Passive Microwave Sensors (SMOPS)	http://www.ospo.noaa.gov/Products/land/smops/

Modified from Peng J, Albergel C, Balenzano A, Brocca L, Cartus O, Cosh MH, Crow WT, Dabrowska-Zielinska K, Dadson S, Davidson MW, de Rosnay P, Dorigo W, Gruber A, Hagemann S, Hirschi M, Kerr YH, Lovergine F, Mahecha MD, Marzahn P, Mattia F, Musial JP, Preuschmann S, Reichle RH, Satalino G, Silgram M, van Bodegom PM, Verhoest NE, Wagner W, Walker JP, Wegmüller U and Loew A (2021) A roadmap for high-resolution satellite soil moisture applications—Confronting product characteristics with user requirements. *Remote Sensing of Environment*, p. 112162, <https://doi.org/10.1016/j.rse.2020.112162>. Copyright (2021), with permission from Elsevier. Links last accessed on May 31, 2021.

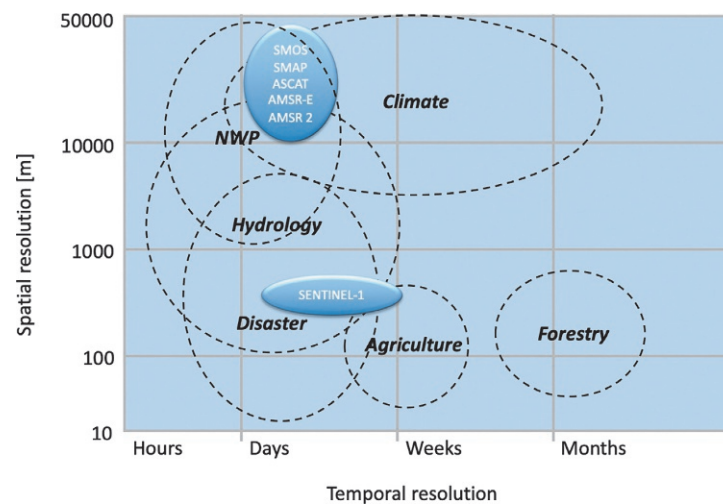


Fig. 6 Application requirements on satellite soil moisture in terms of accuracy, spatial and temporal resolutions. Reprinted from Peng J, Albergel C, Balenzano A, Brocca L, Cartus O, Cosh MH, Crow WT, Dabrowska-Zielinska K, Dadson S, Davidson MW, de Rosnay P, Dorigo W, Gruber A, Hagemann S, Hirschi M, Kerr YH, Lovergine F, Mahecha MD, Marzahn P, Mattia F, Musial JP, Preuschmann S, Reichle RH, Satalino G, Silgram M, van Bodegom PM, Verhoest NE, Wagner W, Walker JP, Wegmüller U and Loew A (2021) A roadmap for high-resolution satellite soil moisture applications—Confronting product characteristics with user requirements. *Remote Sensing of Environment*, p. 112162, <https://doi.org/10.1016/j.rse.2020.112162>. Copyright (2021), with permission from Elsevier.

Table 2 Requirements of various applications on satellite soil moisture.

Application	Usage	Accuracy	Soil moisture depth	Temporal resolution	Other
NWP	Assimilation of soil moisture or low- frequency microwave brightness temperature into NWP system	Accurate temporal dynamics	Surface and root zone	Daily or subdaily	Reliable near real-time products
Climate	Evaluation of model performance and investigation of land-atmosphere interactions	Accurate temporal dynamics	Surface and root zone	Monthly or sub-monthly	Long-term soil moisture climatology
Hydrology	Hydrological modeling and estimation of water cycle components	Accurate absolute soil moisture	Surface and root zone	Sub-daily (e.g., hourly)	Reliable quality information
Agriculture	Precision agriculture and erosion modeling	Accurate absolute soil moisture	Root zone	Weekly and sub-weekly	Reliable quality information
Ecosystem	Ecosystem monitoring and ecological modeling	Accurate absolute soil moisture	Root zone	Weekly	Reliable quality information

Reprinted from Peng J, Albergel C, Balenzano A, Brocca L, Cartus O, Cosh MH, Crow WT, Dabrowska-Zielinska K, Dadson S, Davidson MW, de Rosnay P, Dorigo W, Gruber A, Hagemann S, Hirschi M, Kerr YH, Lovergine F, Mahecha MD, Marzahn P, Mattia F, Musial JP, Preuschmann S, Reichle RH, Satalino G, Silgram M, van Bodegom PM, Verhoest NE, Wagner W, Walker JP, Wegmüller U and Loew A (2021) A roadmap for high-resolution satellite soil moisture applications—Confronting product characteristics with user requirements. *Remote Sensing of Environment*, p. 112162, <https://doi.org/10.1016/j.rse.2020.112162>. Copyright (2021), with permission from Elsevier.

the prospect of novel instrument technology of planned satellite missions and ongoing developments for retrieval, data merging, and downscaling algorithms, we can expect to approach sub-daily soil moisture estimates at an agricultural field scale with ever-improving accuracy in the years to come. This will help us to face and overcome some of the most pressing challenges of our time including climate change adaptation, food security, water resources management, and many others.

Acknowledgments

Alexander Gruber has received funding from the Research Foundation Flanders (FWO-1224320N and FWO-1530019N).

References

- Attema E and Ulaby FT (1978) Vegetation modeled as a water cloud. *Radio Science* 13(2): 357–364.
- Bauer-Marschallinger B, Freeman V, Cao S, Paulik C, Schauffer S, Stachi T, Modanesi S, Massari C, Ciabatta L, Brocca L, and Wagner W (2019) Toward global soil moisture monitoring with sentinel-1: Harnessing assets and overcoming obstacles. *IEEE Transactions on Geoscience and Remote Sensing* 57(1): 520–539. <https://doi.org/10.1109/TGRS.2018.2858004>.
- Brocca L, Tullio T, Melone F, Moramarco T, and Morbidelli R (2012) Catchment scale soil moisture spatial-temporal variability. *Journal of Hydrology* 422–423: 63–75. <https://doi.org/10.1016/j.jhydrol.2011.12.039>.
- de Rosnay P, Muñoz-Sabater J, Albergel C, Isaksen L, English S, Drusch M, and Wigneron J-P (2020) SMOS brightness temperature forward modelling and long term monitoring at ECMWF. *Remote Sensing of Environment* 237: 111424. <https://doi.org/10.1016/j.rse.2019.111424>.
- Dobson MC, Ulaby FT, Hallikainen MT, and El-Rayes MA (1985) Microwave dielectric behavior of wet soil-part II: Dielectric mixing models. *IEEE Transactions on Geoscience and Remote Sensing* 1: 35–46.
- Dorigo W, Wagner W, Albergel C, Albrecht F, Balsamo G, Brocca L, Chung D, Ertl M, Forkel M, Gruber A, et al. (2017) ESA CCI soil moisture for improved earth system understanding: State-of-the art and future directions. *Remote Sensing of Environment* 203: 185–215. <https://doi.org/10.1016/j.rse.2017.07.001>.
- Ferrazzoli P, Luzi G, Paloscia S, Pampaloni P, Schiavon G, and Solimini D (1989) Comparison between the microwave emissivity and backscatter coefficient of crops. *IEEE Transactions on Geoscience and Remote Sensing* 27(6): 772–778.
- Fung A, Li Z, and Chen K (1992) Backscattering from a randomly rough dielectric surface. *IEEE Transactions on Geoscience and Remote Sensing* 30(2): 356–369. <https://doi.org/10.1109/36.134085>.
- Gruber A, Su C-H, Zwieback S, Crow W, Dorigo W, and Wagner W (2016) Recent advances in (soil moisture) triple collocation analysis. *International Journal of Applied Earth Observation and Geoinformation* 45: 200–211. <https://doi.org/10.1016/j.jag.2015.09.002>.
- Gruber A, De Lannoy G, Albergel C, Al-Yaari A, Brocca L, Calvet J-C, Colliander A, Cosh M, Crow W, Dorigo W, Draper C, Hirschi M, Kerr Y, Konings A, Lahoz W, McCol K, Montzka C, Muñoz-Sabater J, Peng J, Reichle R, Richaume P, Rüdiger C, Scanlon T, van der Schalie R, Wigneron J-P, and Wagner W (2020) Validation practices for satellite soil moisture retrievals: What are (the) errors? *Remote Sensing of Environment* 244: 111806. <https://doi.org/10.1016/j.rse.2020.111806>.
- Loew A, Ludwig R, and Mauser W (2006) Derivation of surface soil moisture from ENVISAT ASAR wide swath and image mode data in agricultural areas. *IEEE Transactions on Geoscience and Remote Sensing* 44(4): 889–899. <https://doi.org/10.1109/TGRS.2005.863858>.
- Merchant CJ, Paul F, Popp T, Ablain M, Bontemps S, Defourny P, Hollmann R, Laverne T, Laeng A, de Leeuw G, Mittaz J, Poulsen C, Povey A, Reuter M, Sathyendranath S, Sandven S, Sofieva V, and Wagner W (2017) Uncertainty information in climate data records from earth observation. *Earth System Science Data* 9(2): 511–527. <https://doi.org/10.5194/essd-9-511-2017>.
- Miralles DG, van den Berg MJ, Gash JH, Parinussa RM, de Jeu RAM, Beck HE, Holmes TRH, Jiménez C, Verhoest NEC, Dorigo WA, Teuling AJ, and Johannes Dolman A (2014) El Niño–La Niña cycle and recent trends in continental evaporation. *Nature Climate Change* 4(2): 122–126. <https://doi.org/10.1038/nclimate2068>.
- Mironov V, Kerr Y, Wigneron J-P, Kosolapova L, and Demontoux F (2013) Temperature- and texture-dependent dielectric model for moist soils at 1.4 GHz. *IEEE Geoscience and Remote Sensing Letters* 10(3): 419–423. <https://doi.org/10.1109/LGRS.2012.2207878>.

- Montzka C, Cosh M, Bayat B, Al Bitar A, Berg A, Bindlish R, Bogaen H, Bolten J, Cabot F, Caldwell T, Chan S, Colliander A, Crow W, Das N, De Lannoy G, Dorigo W, Evett S, Gruber A, Hahn S, Jagdhuber T, Jones S, Kerr Y, Kim S, Koyama C, Kurum M, Lopez-Baeza E, Mattia F, McCol K, Mecklenburg S, Mohanty B, O'Neill P, Or D, Pellarin T, Petropoulos G, Piles M, Reichle R, Rodriguez-Fernandez N, Rüdiger C, Scanlon T, Schwartz R, Spengler D, Srivastava P, Suman S, van der Schalie R, Wagner W, Wegmüller U, Wigneron J-P, Camacho F, and Nickeson J (2020) Soil Moisture Product Validation Good Practices Protocol, Version 1.0. In: Montzka C, Cosh M, Nickeson J, and Camacho F (eds.) *Good Practices for Satellite Derived Land Product Validation, Land Product Validation Subgroup (WGCV/CEOS)*, <https://doi.org/10.5067/doc/ceoswgcv/lpv/sm.001>.
- Oh Y, Sarabandi K, and Ulaby FT (1992) An empirical model and an inversion technique for radar scattering from bare soil surfaces. *IEEE Transactions on Geoscience and Remote Sensing* 30(2): 370–381.
- Paulik C, Dorigo W, Wagner W, and Kidd R (2014) Validation of the ASCAT soil water index using in situ data from the international soil moisture network. *International Journal of Applied Earth Observation and Geoinformation* 30: 1–8. <https://doi.org/10.1016/j.jag.2014.01.007>.
- Peng J, Loew A, Merlin O, and Verhoest NE (2017) A review of spatial downscaling of satellite remotely sensed soil moisture. *Reviews of Geophysics* 55(2): 341–366. <https://doi.org/10.1002/2016RG000543>.
- Peng J, Albergel C, Balenzano A, Brocca L, Cartus O, Cosh MH, Crow WT, Dabrowska-Zielinska K, Dadson S, Davidson MW, de Rosnay P, Dorigo W, Gruber A, Hagemann S, Hirschi M, Kerr YH, Lovergine F, Mahecha MD, Marzahn P, Mattia F, Musial JP, Preuschmann S, Reichle RH, Satalino G, Silgram M, van Bodegom PM, Verhoest NE, Wagner W, Walker JP, Wegmüller U, and Loew A (2021) A roadmap for high-resolution satellite soil moisture applications—Confronting product characteristics with user requirements. *Remote Sensing of Environment* 112162. <https://doi.org/10.1016/j.rse.2020.112162>.
- Quast R, Albergel C, Calvet J-C, and Wagner W (2019) A generic first-order radiative transfer modelling approach for the inversion of soil and vegetation parameters from scatterometer observations. *Remote Sensing* 11(3): 285. <https://doi.org/10.3390/rs11030285>.
- Reichle RH, McLaughlin DB, and Entekhabi D (2002) Hydrologic data assimilation with the ensemble Kalman filter. *Monthly Weather Review* 130(1): 103–114.
- Reichle R, De Lannoy G, Liu Q, Koster R, Kimball J, Crow W, Ardizzone J, Chakraborty P, Collins D, Conaty L, et al. (2017a) Global assessment of the SMAP level-4 surface and root-zone soil moisture product using assimilation diagnostics. *Journal of Hydrometeorology* 18(12): 3217–3237. <https://doi.org/10.1175/JHM-D-17-0130.1>.
- Reichle R, De Lannoy G, Liu Q, Ardizzone J, Colliander A, Conaty A, Crow W, Jackson T, Jones L, Kimball J, et al. (2017b) Assessment of the SMAP level-4 surface and root-zone soil moisture product using in situ measurements. *Journal of Hydrometeorology* 18(10): 2621–2645. <https://doi.org/10.1175/JHM-D-17-0063.1>.
- Rodriguez-Fernandez N, Aires F, Richaume P, Kerr Y, Prigent C, Kolassa J, Cabot F, Jimenez C, Mahmoodi A, and Drusch M (2015) Soil moisture retrieval using neural networks: Application to SMOS. *IEEE Transactions on Geoscience and Remote Sensing* 53: 5991–6007. <https://doi.org/10.1109/TGRS.2015.2430845>.
- Taylor CM, de Jeu RAM, Guichard F, Harris PP, and Dorigo WA (2012) Afternoon rain more likely over drier soils. *Nature* 489(7416): 423–426. <https://doi.org/10.1038/nature11377>.
- Ulaby FT, Long DG, Blackwell WJ, Elachi C, Fung AK, Ruf C, Sarabandi K, Zebker HA, and Van Zyl J (2014) *Microwave Radar and Radiometric Remote Sensing*, vol. 4. Ann Arbor: University of Michigan Press.
- Vachaud G, Passerat De Silans A, Balabanis P, and Vauclin M (1985) Temporal stability of spatially measured soil water probability density function. *Soil Science Society of America Journal* 49(4): 822–828. <https://doi.org/10.2136/sssaj1985.03615995004900040006x>.
- Wagner W, Lemoine G, and Rott H (1999) A method for estimating soil moisture from ERS scatterometer and soil data. *Remote Sensing of Environment* 70(2): 191–207. [https://doi.org/10.1016/S0034-4257\(99\)00036-X](https://doi.org/10.1016/S0034-4257(99)00036-X).
- Wang JR and Schmugge TJ (1980) An empirical model for the complex dielectric permittivity of soils as a function of water content. *IEEE Transactions on Geoscience and Remote Sensing* GE-18(4): 288–295. <https://doi.org/10.1109/TGRS.1980.350304>.
- Wigneron J-P, Jackson T, O'Neill P, De Lannoy G, De Rosnay P, Walker J, Ferrazzoli P, Mironov V, Bircher S, Grant J, Kurum M, Schwank M, Munoz-Sabater J, Das N, Royer A, Al-Yaari A, Al Bitar A, Fernandez-Moran R, Lawrence H, Mialon A, Parrens M, Richaume P, Delwart S, and Kerr Y (2017) Modelling the passive microwave signature from land surfaces: A review of recent results and application to the I-band SMOS & SMAP soil moisture retrieval algorithms. *Remote Sensing of Environment* 192: 238–262. <https://doi.org/10.1016/j.rse.2017.01.024>.

Remote Sensing of Soil Organic Carbon

Bas van Wesemael^a and Sabine Chabrilat^{b,c}, ^aEarth and Life Institute, Université Catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium; ^bHelmholtz Center Potsdam GFZ German Research Center for Geosciences, Potsdam, Germany; ^cLeibniz University Hannover, Institute of Soil Science, Hannover, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	632
The principles of imaging spectroscopy	632
Airborne platforms	634
Unmanned aerial systems (UAS) platforms	635
Remote sensing of SOC from Orbit; towards continuous SOC maps over large areas	637
Outlook (limitations and future developments)	639
References	640

Abstract

Soil organic carbon (SOC) is one of the key indicators for soil quality and plays an important role in the global C cycle. Remote sensing techniques are rapidly evolving for mapping and monitoring SOC. Platforms range from unmanned aerial systems (UAS) to airborne and satellite systems. Multivariate analysis and machine learning techniques are used to calibrate SOC prediction models from the reflectance of bare soil surfaces in the visible, near infrared and shortwave infrared wavelengths (400–2500 nm). The principles of imaging spectroscopy, its application from UAS, airborne and satellite platforms and the challenges for the future of this rapidly evolving technique are discussed.

Glossary

Cellulose absorption index CAI $(0.5(R_{2.0} + R_{2.2}) - R_{2.1})$ describes the average depth of the cellulose absorption feature at 2100 nm, where $R_{2.0}$ is the reflectance at 2000 nm, $R_{2.1}$ at 2100 nm and $R_{2.2}$ at 2200 nm.

Chemometrics Refers to extracting information from chemical compounds using data-driven methods such as multivariate analysis, random forest and machine learning techniques.

Chromophore Is the part of the molecule responsible for its color.

Imaging spectroscopy Optical remote sensing observations and in particular reflectance spectroscopy at the remote sensing scale applied for the quantitative determination and modeling of a range of soil properties.

Normalized burn ratio2 (NBR2) Is a spectral index for multispectral satellites (such as Sentinel 2 and Landsat 8) and is calculated from the ratio of the reflectance in the SWIR bands $(SWIR1 - SWIR2)/(SWIR1 + SWIR2)$. The SWIR 1 (band 11) is at 1610 nm and SWIR 2 (band 12) at 2190 nm. For the Sentinel 2 multispectral instrument (MSI).

Normalized soil moisture index (NSMI) Is the ratio between the reflectance in the SWIR $(SWIR1 - SWIR2)/(SWIR1 + SWIR2)$, where SWIR 1 is at 1800 nm and SWIR 2 at 2119 nm. This index is currently only suitable for hyperspectral instruments.

Soil reflectance composite Mosaic of reflectance spectra from bare soil pixels collected over different years. Bare soil pixels are selected from the collection of satellite imagery based on thresholds of spectral indices enabling predominantly undisturbed soils to be separated from all other land cover types such as permanent vegetation and built up areas.

Soil spectral library Data set containing soil properties analyzed in the laboratory together with their reflectance spectra of dry, ground samples acquired with a laboratory spectrometer.

Key points

- SOC maps at the regional to continental scale are scarce and often very costly to update using conventional sampling and analysis techniques.
- Satellite data are freely available and have a relatively high spectral resolution and a revisit time of up to five 5 days. At higher resolution, multispectral cameras or mini spectrometers mounted on unmanned aerial systems (UAS) are also used.
- These technological developments have sparked the interest for SOC mapping and monitoring by remote sensing.
- Many studies still cover relatively small areas ($<200 \text{ km}^2$) within which fields have been selected that meet the optimal requirements for SOC prediction (dry, smooth topsoils without vegetation or residue cover). The number of calibration

samples is often limited and the use of large spectral libraries or soil data bases is often hindered by a lack of standardization.

- For larger areas or when considering multi-temporal reflectance composites, the topsoil conditions cannot be checked in the field. Therefore, thresholds of spectral indices are developed to select the optimal conditions.
- Both the developments in data processing and in sensor technology are rapid including geospatial processing platforms, lightweight spectrometers extending into the shortwave infrared (SWIR) for UAS systems and hyperspectral satellites with short revisit times.

Introduction

Soil organic carbon (SOC) plays a critical role in soil quality and the global C cycle. Soil organic matter and its main component SOC are dynamic properties that show large spatial variation even between fields and are frequently related to (historic) land use or land management. Unfortunately, SOC maps at the regional to continental scale are scarce and often very costly to update using conventional sampling and analysis techniques. Soil spectroscopy has become an established analytical technique for soil components such as C, N (nitrogen), clay and CEC (cation exchange capacity). Recently, similar spectroscopic techniques have been applied to optical remote sensing (RS) platforms offering the possibility of mapping large areas (Chabrilat et al., 2019). Reflectance spectroscopy at the RS scale, referred to as imaging spectroscopy (IS), or hyperspectral imaging, have been shown to be powerful techniques for the quantitative determination and modelling of a range of soil properties. Tziolas et al. (2021) reviewed the recent literature and found that the majority of RS studies dealing with soil properties focused on SOC (including soil organic matter). They showed that overall the quality of the SOC prediction models decreases from the field (median $R^2 = 0.63$) to the regional ($R^2 = 0.60$) and country/continental scale ($R^2 = 0.50$). When these models are applied to pixels with exposed soils, an up-to-date map of soil properties can be obtained that covers large areas and at the same time has a spatial resolution of 20 m (Sentinel 2) to 30 m (Landsat 8) and is detailed enough to show differences within and between fields. The satellite data are freely available and have a relatively high spectral resolution and a revisit time of 5 days (for the Sentinel 2 multispectral instrument (MSI)). For more detailed applications, such as precision agriculture multispectral cameras or mini spectrometers mounted on unmanned aerial systems (UAS) are already used (see the literature review by Angelopoulou et al., 2019). These developments have sparked the interest for SOC mapping and monitoring by remote sensing of either several fields by UAS borne sensors or for large areas by satellite imagery. Angelopoulou et al. (2019) mentioned six key advantages of remote sensing for soil property prediction: (i) a non-destructive technique, (ii) coverage of large areas, (iii) including inaccessible areas, (iv) possibility to predict several soil properties simultaneously, (v) provision of concise data and (vi) reduces the need for laborious sampling campaigns. With the increasing availability of low-cost sensors and high-quality satellite data there is an urgent need to improve automated procedures for the calibration of spectra and clearly define the conditions under which prediction models can be developed and applied for SOC mapping. Spectral models, however, only predict the SOC content for exposed topsoils and calibration models are sensitive to fractional cover by vegetation or residues, variations in moisture content and shading due to roughness. For reliable SOC maps, pixels influenced by such factors should be masked or the model should be corrected for them (Castaldi, 2021).

Here, we first discuss the principles of imaging spectroscopy, then demonstrate its application from airborne, UAS and satellite platforms and conclude with developments in the near future. Airborne applications have the longest history and so far, the highest spectral resolution. Recently, multispectral cameras and light weight spectrometers have been developed for UAS-borne applications. Although these techniques are still in full development, UAS are very flexible for SOC mapping at the field scale and have provided a level of SOC prediction comparable to one that can be obtained from traditional soil sampling and analysis (Andries et al., 2021). Satellite systems provide the benefit of a short revisit time and spatial resolution that can distinguish patterns within a field. There are already several studies that demonstrate the capability of these satellites, such as Sentinel 2 and Landsat 8, for detailed mapping and monitoring of topsoil properties (Tziolas et al., 2021). Some hyperspectral satellites such as the Italian PRISMA instrument are already operational or have recently been launched such as the German EnMAP instrument. It is expected that hyperspectral instruments on board satellite constellations will reach the same global coverage with short revisit time as the Sentinel 2 MSI in the next decades.

The principles of imaging spectroscopy

Soil spectroscopy registers the electromagnetic radiation emitted by a source and reflected by a soil surface. The wave lengths that are most often used either range from the visible light (Vis: 300–700 nm) through the near infrared (NIR: 700–1100 nm) to the short-wave infrared (SWIR: 1100–2500 nm), or cover the mid infrared region (MIR 2500–25,000 nm). Chemical and physical properties of soil compounds are related to the overall shape of the spectrum and to specific absorption bands (Stenberg et al., 2010). The SOC is a chromophore responsible for the soil color with absorption features in the visible wavelength range, (Fig. 1). Moreover, its absorption features in the SWIR range are overtones of vibrations in the MIR range due to stretching and bending of covalent bonds. In short, the correlation between spectra and soil properties is based on a chemometrics approach in which data

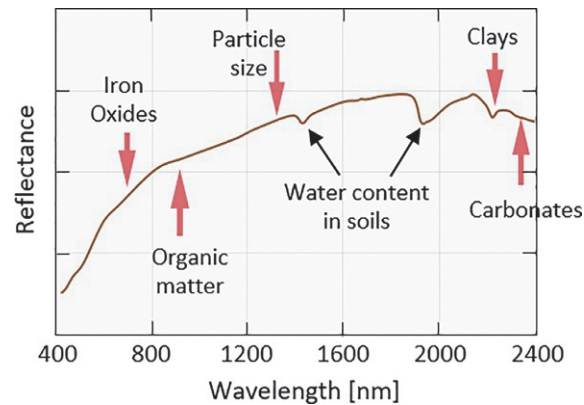


Fig. 1 A synthetic soil spectrum showing main soil chromophores.

mining techniques yield empirical relations between the reflectance at certain wavelengths and soil properties (Chabrilat et al., 2019). The multivariate analysis between the dependent variable (SOC content) and the independent variables (reflectance at different wavelengths) should take correlation between the independent variables into account. The most commonly used techniques range from multivariate regression such as partial least square regressions (PLSR) to decision trees such as random forest (RF) to machine learning techniques ranging from gradient boosting to support vector machines (SVM) and deep learning techniques such as convoluted neural networks (CNN; see for the recent literature (Tziolas et al., 2021)). These data mining techniques are trained on so-called spectral libraries containing both the spectra and the corresponding chemical or physical properties of a range of soil samples. Recently, large-scale spectral libraries covering entire countries (e.g., Australia, United States, Brazil) or continents (e.g., Europe or Africa) are becoming available.

Imaging spectroscopy has the advantage of producing a large amount of spatial soil data but at the expense of lower accuracy due to atmospheric absorption and lower spectral resolution, together with a reduced signal-to-noise ratio of the sensors (Angelopoulou et al., 2019). Compared to laboratory spectrometers, remote sensors have more difficulty in acquiring a signal from the soil surface under optimal conditions i.e., dry and smooth soils without crust, residues or vegetation. Moreover, the reflectance captured by a remote sensor originates from the surface and cannot be used to predict properties deeper in the soil profile. For temperate regions exposed soils occur mainly in cropland (Rogge et al., 2018). Cropland soils are frequently ploughed and the upper 10–30 cm are mixed creating a homogeneous SOC content. However, RS instruments capture a vegetation or residue signal for croplands throughout the largest part of the year. Vaudour et al. (2019a) investigated the performance of SOC prediction models for an agricultural area around Versailles (France) based on the spectra acquired by the Sentinel 2 MSI throughout the year. They found that most images were not suitable because of vegetation, residue cover, variation in soil moisture or roughness. A favorable time window for SOC prediction exists just after tillage and seeding (in spring and autumn for temperate regions) when the soils are exposed.

For calibration and validation geo-referenced soil property databases should be constructed or existing ones such as the European land use and coverage area frame survey (LUCAS; Tóth et al., 2013) can be used. The spectral signal acquired by the RS instrument is then matched with the soil property data from the corresponding pixel. Combining all calibration and validation data in this way will create an ad hoc spectral library with the soil property i.e., SOC content as independent variable and the reflectance (or absorbance) at the different wavelength as the dependent variables. Data mining techniques are then used to define the spectral soil property prediction model. The approach is comparable to the one used in laboratory spectroscopy with the exception that the spectra are acquired directly by the RS instrument instead of the laboratory spectrometer. To ensure validity of the calibration for the pixels of the RS instrument, it is recommended that composite samples are collected from the topsoil (0–15 cm and not deeper than the plough layer) covering an area the size of at least one pixel for satellite imagery or 3 by 3 pixels for imagery with 1–5-m resolution to take positional uncertainty into account. The three main factors affecting soil reflectance and accuracy of the model prediction have been assigned to: (1) the atmosphere; (2) a low signal-to-noise ratio of the imaging sensor data and (3) disruptive factors affecting the soil surface (partial vegetation cover, moisture, crust).

Dedicated sampling campaigns and analysis of soil properties for these ad hoc spectral libraries become the main constraints when, on the one hand, large areas that are targeted such as Sentinel 2 tiles or, on the other hand, UAS platforms are used to map several fields without a priori knowledge on the soil properties. In such cases, existing spectral libraries would offer a cost-effective alternative for calibration and validation. Unfortunately, attempts to transfer spectral models developed on laboratory spectral libraries (e.g., LUCAS) to predict a soil property using the signal of a satellite RS instrument have so far been unsuccessful. Soil standards scanned with both a master and a workhorse (slave) instrument in the laboratory have been used to predict soil properties scanned with the slave instrument using the spectral model developed with the master instrument (Kopačková and Ben-Dor, 2016). However, the transfer of models developed from a spectral library to the signal of a satellite or airborne RS instrument is not yet possible. The drawbacks are the lack of point specific correction models for atmospheric disturbance, the lack of soil standards for which a signal can be acquired by a RS instrument and the difference in spectral resolution between a laboratory spectrometer

(continuous with 1-nm band width) and a RS instrument (c. nine spectral bands mainly in visible and near infrared). For UAV borne instruments, Zhang et al. (2021) have shown that a spectral alignment procedure was needed when they predicted SOC content from the UAS spectrometer using a spectral model developed on a laboratory spectral library. Wehrhan and Sommer (2021) suggested a hybrid approach where topographic indices were used together with a reduced number of samples for the prediction of SOC content in an undulating morainic landscape. Both examples demonstrate the flexibility and effectiveness of UAS-borne instruments for mapping albeit for areas with a limited extent.

The most frequently used figures of merit for evaluating the accuracy of the SOC prediction models are the R^2 of the predicted versus observed SOC contents, the root mean square error (RMSE), the ratio of performance to deviation (RPD, i.e., the standard deviation divided by the RMSE) and the ratio of performance to interquartile range (RPIQ). Uncertainty of the SOC prediction in each grid cell is usually expressed as the ratio between the distribution of the model outcomes covering 90% of all possible cases (P95–P5) and the median model outcome. Examples of uncertainty maps are rare and the existing ones are often based on bootstrapping of the calibration dataset (Gomez et al., 2015).

Airborne platforms

The first airborne sensors were often push broom systems that collect data cubes at high spatial (2–5 m depending on the flight altitude) and spectral (200–300 spectral bands in the VNIR–SWIR) resolutions. The airborne campaigns can typically cover flight lines of 10–70 km long and up to 10 km wide (e.g., Stevens et al., 2010; Guo et al., 2019). The campaigns are generally organized when a large number of the fields have just been prepared for seeding. It is assumed that surface conditions are homogeneous and that the reflectance can be related to one or more soil properties. Overall, airborne campaigns give the best SOC prediction models among the RS platforms (Table 1). Shi et al. (2020) used a PLSR model to predict SOC content in croplands along a 66 by 7 km flight line in central Belgium using the spectra from the Swiss-Belgian Airborne Prism Experiment (APEX) sensor. They predicted SOC contents for the pixels with an NDVI (normalized difference vegetation index) ranging from 0.10 to 0.22, obtaining good results for an independent validation data set ($R^2 = 0.56$, $RMSE = 3.0 \text{ g C kg}^{-1}$, $n = 43$). Similar results were obtained by Guo et al. (2019) for a freshly ploughed field of $1.5 \times 1 \text{ km}$ in Iowa (USA) using an airborne Headwall sensor (400–1700 nm; SpecTIR LLC; $R^2 = 0.66$; $n = 55$). The spectral features related to SOC are often difficult to distinguish, in particular in the SWIR region where the signal-to-noise ratio is low. Therefore, some authors proposed ensembles of modelling approaches (Laamrani et al., 2019) in order to select the most important wavelengths for SOC prediction. Diek et al. (2016) were, to the best of our knowledge, the first to combine the spectra of flight campaigns carried out during consecutive years and thus created a multi-temporal composite. The advantage of such composites is that the area of bare soils increases, because with crop rotation different fields are exposed during the time windows when either summer or winter crops are sown. Until now, airborne sensors have provided the best spectral and spatial resolution compared to UAS and satellite platforms (Angelopoulou et al., 2019). However, flight campaigns are costly and difficult to organize as they are restricted to a limited time window when both atmospheric conditions are optimal and (a large portion of) the fields are in seedbed condition. Therefore, they are only available for limited areas and times, therefore multi-temporal composites like that of Diek et al. (2016) are extremely rare.

Table 1 Performance of SOC prediction from airborne platforms.

Sensor ^a	Spectral range (nm)	Algorithm ^b	R^2	RMSE (g kg^{-1})	RPD	References cited in Angelopoulou et al. (2019)
AHS-160	430–2540	PLSR, PSR, SVM	0.53–0.89	3.13–6.22	1.47–3.15	Stevens et al. (2010)
AHS-160	430–2540	PLSR	–	1.7	1.47	Stevens et al. (2010)
HyMap	450–2500	PLSR	0.34–0.83	0.76–1.10	1.14–2.32	Hbirkou et al. (2012)
ProSpec TIR V-S	400–2500	PLSR	0.33	3.82	1.25	Franceschini et al. (2015)
AHS-160	430–2540	PLSR	0.62	1.34	1.8	Bartholomeus et al. (2011)
AISA-Eagle	400–1000	PLSR	0.44	4.05	1.4	Vaudour et al. (2016)
AHS-160	430–2540	SLR, SMLR, PLSR	0.27–0.60	6.44–8.70	1.18–1.60	Peon et al. (2017)
AISA Dual System	400–2450	SMLR	0.73	8.4	–	Homolova et al. (2014)
APEX	400–2500	PLSR	–	4.3	2.5	Castaldi et al. (2019)
HyMap	450–2500	PLSR	0.73–0.85	0.19–0.25	1.94–2.62	Vohland et al. (2017)
APEX	400–2500	PLSR	0.56	3.0	0.83	Shi et al. (2020)
Headwall	400–1700	PLSR	0.54	3.3	1.45	Guo et al. (2019)
Nano Hyperspec	400–1000	PLSR, RF, SVM	0.35–0.72	0.35–0.47	1.3–1.8	Laamrani et al. (2019)
Nano Hyperspec	400–1000	FOD, RF	0.66	2.5	1.68	Hong et al. (2020)

^aAHS-160: Airborne Hyperspectral Sensor 160 (AHS, Caravan International Corporation, USA); HyMap (Integrated Spectronics, Sydney, Australia); ProSpecTIR V-S sensor (SpecTIR LLC, Reno, NV); Airborne Prism Experiment (APEX); AISA Eagle and AISA Dual system (Specim, Finland); Headwall sensor (Spec TIR LLC); Nano Hyperspec (Spec TIR LLC).

^bPLSR: Partial least square regression; PSR: Penalized spline regression; SVM: Support vector machine; SLR: Simple linear regression; SMLR: Stepwise multilinear regression; RF: Random forest; FOD: Fractional order derivative.

Modified after Angelopoulou T., Tziolas N., Balafoutis A., Zalidis G. and Bochtis D. (2019) Remote sensing techniques for soil organic carbon estimation: A review. *Remote Sensing* 11(6). <https://doi.org/10.3390/rs11060676>.

Unmanned aerial systems (UAS) platforms

The UAS are evolving rapidly and offer the potential for very high spatial resolution of SOC prediction in croplands. Such a resolution matches the needs for precision agriculture. At the same time, UAS systems are less sensitive to some disadvantages of space- and air-borne systems in particular atmospheric and surface conditions (Wehrhan and Sommer, 2021). Although UAS systems are constrained by payload, flight duration and legal restrictions, the organization of flight campaigns for one or several fields is quite flexible giving spatial coverage of up to a few hundreds of hectares (for a fixed wing UAS) and a spatial resolution of less than 1 m. Moreover, the flight campaigns can be easily planned to meet both optimal weather and soil conditions, minimizing atmospheric disturbance and disturbances at the soil surface such as roughness, moisture and residues. Recent literature reviews on the role of Earth Observation for SOC prediction mentions only a handful of papers using UAS borne sensors (Angelopoulou et al., 2019; Tziolas et al., 2021) (Fig. 2).

Light weight spectrometers and multispectral cameras are essential because the payload of UAS platforms is limited. Crucil et al. (2019) tested the performance of two mini spectrometers (STS-VIS and STS-NIR, Ocean Optics Inc., Dunedin, FL, United States) and two multispectral cameras (Sequoia, Parrot Drones, S.A, Paris, France) and Mini-MCA6 (Tetracam Inc., Chatsworth, CA, United States; Table 2) for their performance to predict SOC content in dried and ground soil samples. The prediction was compared to the one obtained with the ASD Fieldspec 3FR reference spectrometer (Analytical Spectral Devices Inc. Boulder, CO, United States). The tests were carried out in both the laboratory and outside under natural light. Moreover, transferability of the models between laboratory and natural light conditions was tested. The authors showed that the performance of lightweight sensors for SOC prediction was comparable to that of one of the reference instrument (ASD). Under outdoor conditions the best results were obtained using the vis-NIR range only, probably due to lower signal-to-noise ratio in the SWIR range as a result of atmospheric disturbance. Moreover, the multispectral camera gave similar or better results than the STS mini spectrometers.

Aldana-Jague et al. (2016) used a UAS-borne Tetracam Mini-MCA6 camera for mapping SOC content of the Hoosfield long term-trial at Rothamsted (United Kingdom). The soils are homogeneous with a rather strong gradient in SOC content ($8.2\text{--}46.7\text{ g C kg}^{-1}$). They corrected the camera for both systematic noise (using a dark surface in the laboratory) and for shadows along the outer edges of the images (using a white Lambertian surface). The SOC prediction model was robust with a cross-validation R^2 of 0.95 and an RMSE of 2.1 g C kg^{-1} . A good fit to the SOC patterns was obtained using an external validation

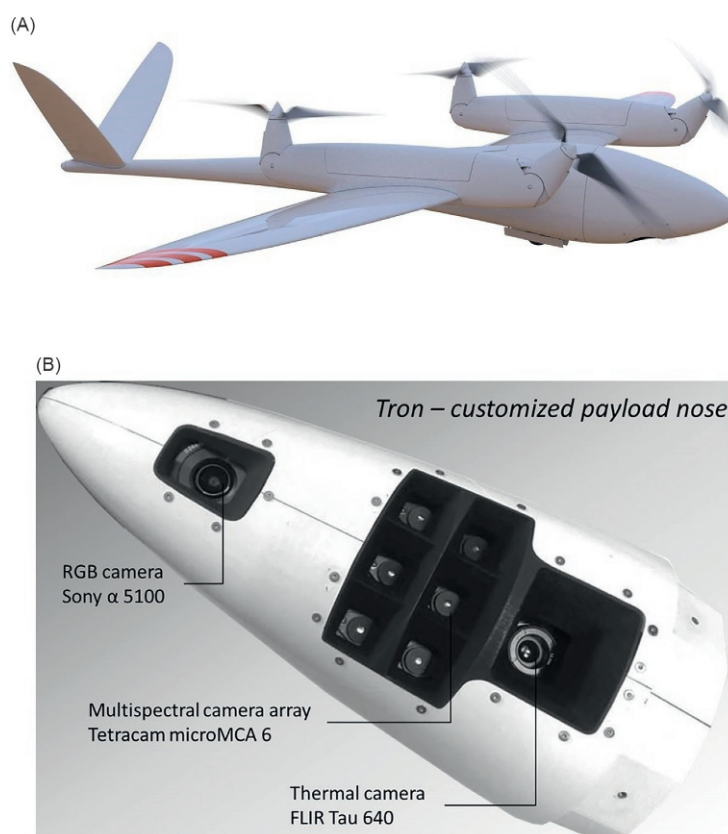


Fig. 2 (A) Example of a fixed wing UAS; (B) customized payload nose with mounted sensors. Reproduced with permission from Wehrhan M and Sommer M (2021) A parsimonious approach to estimate soil organic carbon applying unmanned aerial system (UAS) multispectral imagery and the topographic position index in a heterogeneous soil landscape. *Remote Sensing* 13(18). <https://doi.org/10.3390/rs13183557>.

Table 2 Characteristics of the spectrometers (hyperspectral) and cameras (multispectral) used for evaluating the performance of UAS systems for SOC prediction.

Instrument ^a	Type	Spectral range (nm)	Spectral resolution (nm)	Spectral windows (nm)	FWHM ^b (nm)
ASD	Hyperspectral	350–2500	1	/	/
STS-VIS	Hyperspectral	350–800	0.45	/	/
STS-NIR	Hyperspectral	650–1100	0.45	/	/
Ocean FX	Hyperspectral	350–1000	0.39	/	/
HySpex Mjolnir S-620	Hyperspectral	1000–2500	5.1	/	/
Sequoia	Multispectral	/	/	550; 660; 735; 790	40; 40; 10; 40
Mini-MCA6	Multispectral	/	/	480; 550; 670; 780; 880; 1000	10
MicaSense Altum Dual	Multispectral	/	/	475; 531; 560; 650; 668; 706; 717; 842	
MicaSense Rededge M	Multispectral	/	/	475; 560; 668; 717; 840	20; 20; 10; 40; 10

^aASD Fieldspec 3 FR (Analytical Spectral Devices Inc.); STS-VIS and STS-NIR (Ocean Optics Inc., Dunedin, FL, United States); Ocean FX (Ocean Optics Inc., Dunedin, FL, United States); HySpex Mjolnir S-620 (Norsk Elektro Optikk, NEO); Sequoia (Parrot Drones S.a.s., Paris, France); Mini-MCA66 (Tetracam Inc., Chatsworth, CA, United States); MicaSense Altum Dual and Rededge M (MicaSense Inc., Seattle, WA, United States).

^bFWHM: Full width at half maximum.

Modified after Crucil G, Castaldi F, Aldana-Jague E, van Wesemael B, Macdonald A and Oost K (2019) Assessing the performance of UAS-Compatible multispectral and hyperspectral sensors for soil organic carbon prediction. *Sustainability (Switzerland)* 11(7). <https://doi.org/10.3390/su11071889>.

dataset (RMSE = 2.6 g C kg⁻¹). They also estimated the cost of their approach, consisting of 5 h for image acquisition and the collection of 30 calibration samples, 5 h of post processing and the cost of sample analysis, totaling 800 euros for a cropland field of between 2 ha and 10 ha. More recent SOC predictions at the field scale from UAS-borne platforms used the Parrot Sequoia and the MicaSense Altum dual sensor (MicaSense Inc., Seattle, WA, United States) with nine spectral bands in the vis-NIR range (Žižala et al., 2019; Biney et al., 2021; Table 2). Both studies were carried out in the Czech Republic on fields with a smaller range in SOC contents (c. 6–29 g C kg⁻¹). Žižala et al. (2019) obtained the best SOC prediction using the Cubist model and an external validation set ($R^2 = 0.72$; RMSE = 3.1 g C kg⁻¹). The predictions of the UAS-borne system correlated well with those from a hyperspectral airborne sensor and a Sentinel 2 composite. Unfortunately, the UAS-borne images from Biney et al. (2021) were affected by banding along the flight lines and the SOC prediction models were unsatisfactory ($R^2 = 0.22$).

Structure for motion algorithms is frequently used as a technique to match overlapping points in order to mosaic images of multispectral cameras. Their very high resolution allows residue, crop or shaded pixels to be included, which makes correlation with the calibration sample uncertain. Moreover, distortion along the edges of the images, vignetting, and anisotropic behavior resulting from variation in shading according to the flight direction are sources of errors often resulting in banding along the flight lines. Zhang (2022) demonstrated the necessity for data pretreatment of mosaics from a multispectral camera including correction for anisotropy, spatial resolution and resampling together with shadow and soil crust resampling. He predicted SOC content using a UAS-borne MicaSense Rededge-M camera with five spectral bands (475, 560, 668, 717 and 840 nm) for three fields in the Belgian loam belt with a mean SOC content of 10 g C kg⁻¹. The model performance (cross-validation) clearly improved from $R^2 = 0.33$ and RMSE = 2.9 g C kg⁻¹ without correction to $R^2 = 0.67$ and RMSE = 2.9 g C kg⁻¹ after correction for anisotropy, and crust and shadow removal.

The SOC prediction models discussed in the previous paragraph were specifically calibrated for each field. Given the flexibility of UAS platforms there is the potential to fly several fields and thus increase cost efficiency because calibration could be done using samples collected from a single field or even better using an existing (local) spectral library. Wehrhan and Sommer (2021) demonstrated a parsimonious approach using a Tetracam mini MCA 6 camera (Table 2). They used only eight calibration samples and included the topographic position index (TPI) indices together with spectral bands in their SOC prediction model. The cross-validation of their model using the reflectance at 570 nm and the TPI yielded an $R^2 = 0.91$ and an RMSE = 1.1 g C kg⁻¹ for one of the fields in the morainic landscape of northern Germany with a mean SOC content of 17 g C kg⁻¹. Moreover, they showed that the SOC model developed for one field also predicted the SOC content of an adjacent field reasonably well. Zhang et al. (2021) used a UAS platform with a portable Ocean FX spectrometer with a range from 350 to 1000 nm (Ocean Optics Inc., United States) together with a high precision RTK/PKK GPS to acquire precise geo-referenced spectra for a field in the Belgian loam belt. A local calibration based on correlating the spectra with SOC content from 179 samples at the collocated points yielded a good SOC prediction performance (RMSE = 0.57 g C kg⁻¹ and RPIQ = 2.35). However, as discussed above, the large number of samples required makes this approach inefficient. Therefore, they used a local spectral library to create an SOC prediction model and applied a spectral alignment procedure to resolve the discrepancy between UAS and laboratory measurements. Although the quality of the model decreased somewhat compared to the local approach (RMSE = 0.93 g C kg⁻¹ and RPIQ = 1.45), this provides a potential for mapping the SOC contents of multiple fields efficiently within the region.

Overall, the developments in UAS borne sensors and their SOC prediction models are promising. The SOC predictions have a high spatial resolution and are generally accurate. Flight campaigns are flexible and optimal weather conditions coinciding with minimal disturbance of the soil surface can be selected. Protocols are urgently required to deal with situations of inadequate

prediction models. They need to deal with mosaicking, corrections for the vignette effect, shade, crust, and anisotropy related to the orientation of the flight line with regard to the solar position. Moreover, an efficient calibration would reduce the costs and when stable would be applicable over an entire region.

Remote sensing of SOC from Orbit; towards continuous SOC maps over large areas

Recently, the archives of medium resolution satellite sensors (10–30 m) became freely accessible. There are several multispectral instruments (Landsat8 -operational land imager (OLI) from 1984 onwards and Sentinel 2 multispectral instrument (MSI) from 2015 onwards) and two hyperspectral instruments (Hyperion, 2000–2017 and PRISMA 2020 onwards). The Landsat 8 OLI has nine bands at 30-m resolution to cover the visible (coastal, blue, green and red), near infrared (NIR), shortwave infrared (SWR1 and SWR2) and thermal infrared regions (Fig. 3). The Sentinel 2 MSI has 10 bands at 10–20 m resolution to cover the visible (coastal, blue, green and red, near infrared (Red edge 1, Red edge 2 and Red edge 3)) and short-wave infrared (SWIR 1 and SWIR 2), while another three bands are for atmospheric correction at 60-m resolution. The favorable signal-to-noise ratio and the frequent revisit time (5 days for the Sentinel 2 and 16 days for the Landsat 8) of these multispectral satellites entails a large amount of freely available data and offers potential for soil property mapping. The use of hyperspectral satellites for soil property mapping is currently limited either by their spatial coverage or their medium signal-to-noise ratio. The Hyperion sensor on the EO (Earth observation)-1 satellite platform was the only hyperspectral sensor covering the vis-NIR-SWIR range in orbit (2000–2017). The sensor provided a spatial resolution of 30-m, a spectral resolution of 10 nm, a swath of 7.5 km and a low to medium signal-to-noise ratio (around 50; Gomez et al., 2008). The hyperspectral PRISMA sensor was launched in 2019. It has 66 bands in the vis-NIR and 171 bands in the SWIR with a bandwidth of c. 12 nm with a signal-to-noise ratio of 400–500 in the NIR and part of the SWIR and smaller values at both shorter and longer wavelengths. The spatial resolution is around 40 m and the swath 30 km resulting in a revisit time of 29 days.

So far, there have been only a limited number of studies using the hyperspectral satellites for SOC prediction. They covered relatively small areas with sparse vegetation or bare soils. Gomez et al. (2008) used a PLSR approach based on Hyperion data to predict SOC in cotton (*Gossypium hirsutum* L.) fields of New South Wales, Australia, and compared the results to spectra collected in the field using an Agri Spec spectrometer (Analytical Spectral Devices, Boulder, Colorado). Overall, the SOC prediction models from the Hyperion data gave reasonable results with $R^2 = 0.51$ and RPD = 1.43, but prediction with the field spectrometer was better

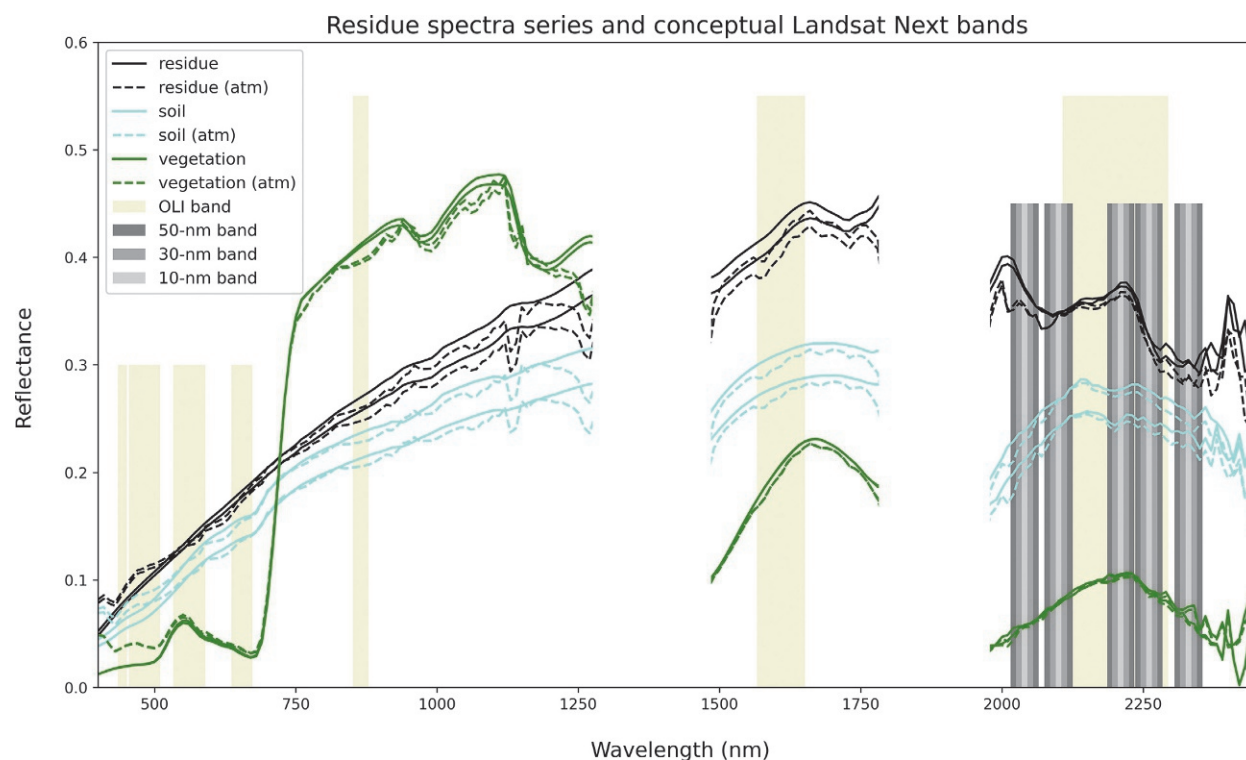


Fig. 3 Example spectra for crop residue, soil and green vegetation with and without simulated sensor noise and atmospheric correction residuals. Vertical tan bars represent the bands of the current Landsat Operational Land Imager (OLI) and the grey bars represent five shortwave infrared bands under consideration for a new version of the Landsat system. Reproduced with permission from Hively WD, Lamb BT, Daughtry CST, Serbin G, Dennison P, Kokaly RF, Wu Z and Masek JG (2021) Evaluation of SWIR crop residue bands for the landsat next mission. *Remote Sensing* 13(18). <https://doi.org/10.3390/rs13183718>.

($R^2 = 0.67$, RPD = 1.75). According to the authors, impurity of the Hyperion pixels (30 m) and the low signal-to-noise ratio explain this difference in performance. The Hyperion image of Lu et al. (2013) covered cropland in an area with a continental dry climate in China during the dormant season, when the soils were bare after wheat (*Triticum aestivum* L.) or potato (*Solanum tuberosum* L.) crops. They found that prediction performance for laboratory and simulated Hyperion spectrometry was better than that based on the spectra acquired by the satellite. The results of validation of the SOC prediction model based on Hyperion spectra ($R^2 = 0.50$; RPD = 1.45) were similar to the ones obtained by Gomez et al. (2008). Castaldi et al. (2016) evaluated the potential of future hyperspectral satellites. They used a spectral library for croplands in southern Italy ($n = 166$) and simulated the spectra based on the specifications of the PRISMA satellite and the forthcoming ones such as EnMAP and NASA-SBG (Surface Biology and Geology), the former HypsIRI mission. After considering atmospheric disturbance and the specified signal-to-noise ratio of the sensors they concluded that these forthcoming satellite missions could only marginally improve SOC predictions compared to existing hyperspectral or multispectral missions. The main constraint for the forthcoming satellites is still the low signal-to-noise ratio in the SWIR.

With the launch of Sentinel 2 MSI a large amount of spectral data at a revisit time of 5 days became available. This availability of frequent imagery offered the potential for SOC prediction in croplands where the time window during which the soil is bare is limited to a few weeks after harvesting and before seedling emergence (Diek et al., 2016). Several authors investigated the performance of SOC predictions based on Sentinel 2 images for various small regions with bare cropland soils and found that the SOC prediction models based on laboratory or airborne spectrometry were slightly better than those based on Sentinel 2 imagery (R^2 : 0.56; RPD: 1.0–2.6) (Gholizadeh et al., 2018; Castaldi et al., 2019; Vaudour et al., 2019b). Even with a slightly poorer SOC prediction performance, the SOC maps created using Sentinel 2 data display similar patterns to the maps based on airborne hyperspectral data and show coherent patterns (e.g., related to clay content).

Encouraged by the frequent revisit time of the Sentinel 2 MSI, Vaudour et al. (2019a) explored the performance of SOC prediction models established for croplands using a single Satellite 2 image at different acquisition dates. They targeted bare cropland soils with an NDVI < 0.35 and observed that performance of the PLSR models varied over time and that in general the best results were obtained for late spring when soils were probably drier and smoother after sowing (R^2 : 0.46–0.58; RPD: 1.4–1.5). The extent of bare soil varied together with development of the different crops, but for a single Sentinel 2 image it never exceeded 50% of the cropland area. Therefore, scientists are increasingly turning to multi-temporal methods of analysis, where soil reflectance composites are created by mosaicking the mean (or median) bare soil spectrum for each pixel acquired over a time span of several years. The first example of such an approach is the Geospatial Soil Sensing System (GEOS3) that creates a bare soil image after thresholding for green vegetation using NDVI and the residues with the SWIR bands (NBR2; Dematté et al., 2018). The soil composite mapping processor (ScMAP) uses annual development of the crop from exposed soil (after ploughing) to photosynthetically active vegetation cover occurring at least once per year (Rogge et al., 2018). For multi-temporal composites, they define minimum and maximum NDVI thresholds to separate cropland from other land uses.

Zepp et al. (2021) applied the SCMap soil reflectance composite approach to the Landsat archive images between 1984 and 2014 covering Bavaria. The random forest model showed the best performance when predicting SOC contents for croplands ($R^2 = 0.67$, RMSE = 1.24%, RPD = 1.77). Although the soil reflectance composites appear to provide a stable mean spectrum for predicting SOC contents, it remains challenging at the regional to country scale, and in particular for multi-temporal composites to minimize the effects of, for example green and dry vegetation, roughness, and soil moisture (Castaldi, 2021). Therefore, SOC prediction models should be formulated and applied using spectra of 'pure soil' pixels, i.e., dry soils without green or dry vegetation and reduced roughness (e.g., soil in seedbed condition; Castaldi, 2021). The spectra of green vegetation and residues are clearly different from a soil spectrum and the same applies to a dry and a wet soil spectrum (Fig. 3). Spectral indices have been developed to mask green vegetation (the well-known NDVI), crop residues (cellulose absorption index; see e.g., Hively et al., 2021) and soil moisture (normalized soil moisture index, NSMI; see e.g., Chabrilat et al., 2019). Unfortunately, the narrow SWIR bands that are required for CAI and NSMI cannot be picked up by the Sentinel 2 or Landsat multispectral instruments (Fig. 3). Nevertheless, PLSR models applied to soil reflectance composites of the spring months have shown reasonable model performance during 10-fold cross-validation for a Sentinel 2 tile T31UFU covering the area to the North and North East of Amsterdam (The Netherlands; $R^2 = 0.49$; RPD = 1.4; RMSE = 3.49 g C kg⁻¹; Fig. 4). An independent validation ($n = 28$) showed that for 74% of samples the measured SOC value was within the prediction interval of the spectral model. Regional trends in SOC contents can be observed in the reclaimed IJsselmeerpolders where croplands dominate. The Wieringermeer polder in the North is the oldest (drained in 1927) and overall has larger SOC contents than the Noordoost polder (drained in 1942) and the Flevopolder (eastern part drained in 1957 and southern part drained in 1968). The Noordoost polder includes some former peat islands and although these are mainly under grassland, the lake sediments around their edges are thin and remnants of peat have been mixed into the plough layer. This is evident in the larger SOC contents near Urk, Schokland and Lemmer (Fig. 4).

The examples of surface reflectance composites in the previous paragraph demonstrate that deviating spectra related to disturbance of surface conditions disappear in the mean spectrum. Alternatively, algorithms can be formulated to select optimal spectra only that represent 'pure' soil pixels. Time series of Sentinel 2 imagery during the growing season can help to distinguish the optimal period for SOC prediction i.e., the last image before the crop starts to develop (Dvorakova et al., 2021). This approach improves the PLSR models for SOC prediction using all cloud free Sentinel 2 images over a 2-year period in the Belgian loam belt: selecting all images: $R^2 < 0.30$, RPD < 1.4 to selecting only the greening up phase and an NBR2 < 0.07: $R^2 = 0.54$, RPD = 1.68. This strict selection targets soils under seedbed conditions. These are overall smooth, bare and the surface dries quickly. However, a limit of this method might be the restricted coverage of the total cropland area, which reached only 62% compared to 95% when only an NDVI threshold was applied.

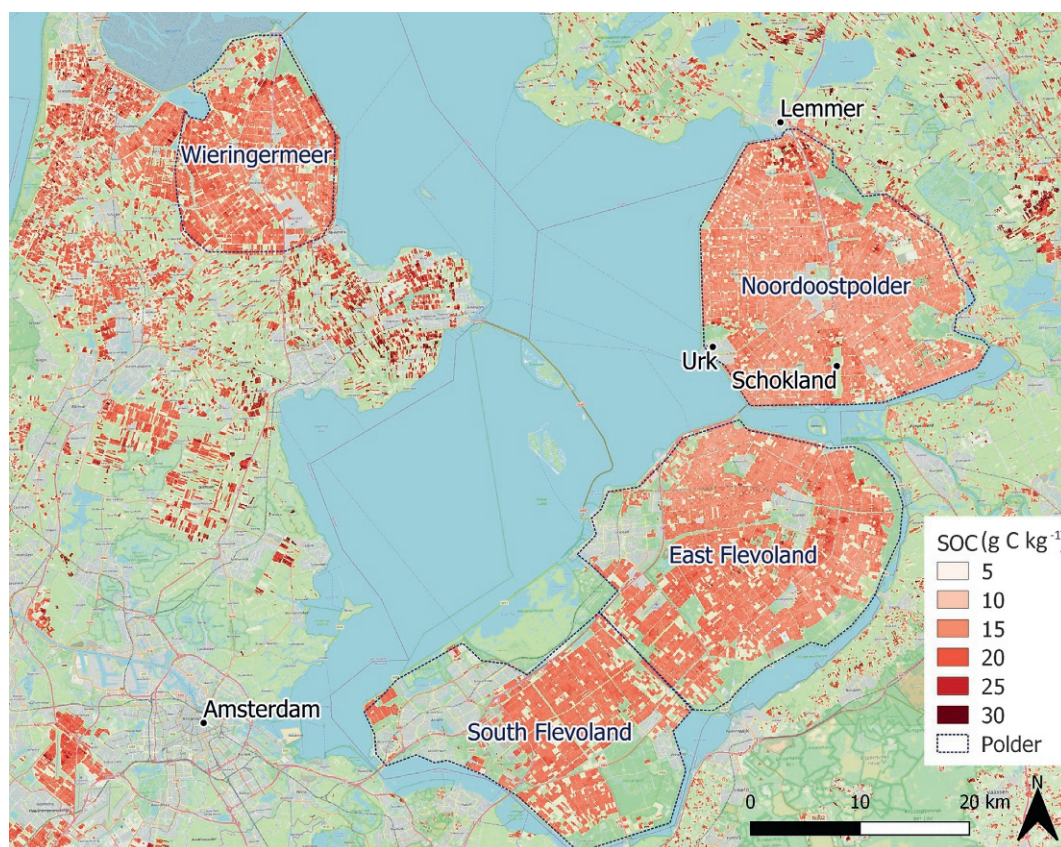


Fig. 4 Soil organic carbon map based on reflectance composites of the spring months from 2016 to 2021 of Sentinel 2 tile T31UFR covering the northwestern part of the Netherlands. Open Street Map has been used as a base map.

Outlook (limitations and future developments)

There are rapid developments in SOC mapping and monitoring using EO technology from both UAS platforms and orbit. The number of studies for UAS systems is still limited, but are increasing. The freely available Sentinel 2 and Landsat 8 archives have triggered the interest of many researchers. Several projects have started to improve SOC prediction for croplands and are moving towards operational systems for mapping and monitoring SOC contents in topsoils at continental and even global scales (e.g., World Soils financed by the European Space Agency and STEROPES financed by the European Joint Program Soil). Below we discuss the main limitations that still exist for operational SOC mapping and monitoring followed by suggestions for remediation:

- Limitation of the area covered and data to be shared:** Many studies still cover relatively small areas ($<200 \text{ km}^2$) within which fields have been selected that meet the optimal requirements for SOC prediction (dry, smooth topsoils without vegetation or residue cover). The number of calibration samples is often limited and the use of large spectral libraries or soil data bases is often hindered by a lack of standardization. Recently, there are several initiatives to foster standardization of the measurement protocols for both wet chemistry and laboratory spectral measurements (including instrument standardization). The P4005—Standards and protocols for soil spectroscopy—within the Institute of Electrical and Electronics Engineers (IEEE) standards organization was launched recently to formulate agreed protocols for spectral measurements. Such a standardization would allow data exchange between spectral libraries and thus offer the possibility of large calibration data sets and the use of more advanced machine learning techniques for EO based SOC prediction. The global soil laboratory (GLOSOLAN) initiative of the FAO has already started on the standardization of laboratory protocols for wet chemistry and has recently created a working group for soil spectroscopy. The Open Soil Spectral Library launched by Soil Spectroscopy 4 Global Good is a first step towards a freely accessible spectral library that builds on the exchange between existing continental scale libraries.
- Limitations of thresholds for bare soil detection:** For larger areas or when considering multi-temporal reflectance composites, the topsoil conditions during flight cannot be checked in the field. Therefore, thresholds of spectral indices are developed to select the optimal conditions. Unfortunately, such indices do not exist to distinguish rough soils. The indices for moisture and residues are in the SWIR. At present the multispectral satellites have only two broad bands in the SWIR that react simultaneously, but not very specifically, to moisture and residue (Fig. 3). Hively et al. (2021) suggested adding three narrow bands in the SWIR for the next generation of Landsat sensors. The cellulose absorption index (CAI) based on the reflectance of narrow bands at 2040,

2100, and 2210 nm has a strong potential to detect residues. For detecting roughness in freshly ploughed fields synthetic aperture (SAR) radar data acquired by Sentinel 1 can offer a solution (Vaudour et al., 2019a).

- **Limitations of infrastructure capabilities:** The wealth of EO data is likely to increase with the advent of hyperspectral satellite systems. The European Commission together with the European Space Agency (ESA) have developed Copernicus Data and Information Access Services (DIAS), geospatial processing platforms with a soil mapping orientation. This service provides continuous access to Copernicus, as well as cloud storage and computing resources. These services will need to be extended to cover the provision and processing of future hyperspectral satellite data. Sensors are rapidly evolving towards micro-electromechanical systems (MEMS) for UAS platforms and hyperspectral sensors for satellites. Although SWIR cameras for UAS still need to be optimized in terms of size and weight, and the auxiliary devices needed during flight have to be miniaturized, first models are just becoming available (e.g., HySpex Mjolnir S-620 from Norsk Elektro Optikk, NEO). Regarding hyperspectral satellites, the Italian PRISMA sensor has been operational since 2020 and is the first of a series of high signal-to-noise ratio instruments with a high spatial resolution and a limited coverage per day satellite imaging spectrometers. Planned missions with similar instruments are: the German EnMAP (Environmental Mapping and Analysis Program), the Japanese HISUI (Hyperspectral Imager Suite), and the Italy-Israeli SHALOM (Spaceborne Hyperspectral Applicative Land and Ocean Mission). Moreover, several global mapping missions are planned such as the NASA-SBG (Surface Biology and Geology), HySpex2 proposed as Earth Explorer Mission ESA and the Sentinel-10/CHIME satellite proposed as the ESA candidate mission (Copernicus Hyperspectral Imaging Mission for the Environment). These satellites will be a major step forward as they will improve SOC prediction capabilities of EO systems compared to the currently available multispectral satellites such as Sentinel 2 and Landsat 8.

References

- Aldana-Jague E, Heckrath G, Macdonald A, van Wesemael B, and van Oost K (2016) UAS-based soil carbon mapping using VIS-NIR (480–1000 nm) multi-spectral imaging: Potential and limitations. *Geoderma* 275. <https://doi.org/10.1016/j.geoderma.2016.04.012>.
- Andries A, Morse S, Murphy RJ, Lynch J, Mota B, and Woolliams ER (2021) Can current earth observation technologies provide useful information on soil organic carbon stocks for environmental land management policy? *Sustainability* 13(21). <https://doi.org/10.3390/su132112074>.
- Angelopoulos T, Tziolas N, Balafoutis A, Zalidis G, and Bochtis D (2019) Remote sensing techniques for soil organic carbon estimation: A review. *Remote Sensing* 11(6). <https://doi.org/10.3390/rs11060676>.
- Biney JKM, Saberioon M, Borůvka L, et al. (2021) Exploring the suitability of UAS-based multispectral images for estimating soil organic carbon: Comparison with proximal soil sensing and spaceborne imagery. *Remote Sensing* 13(2): 1–19. <https://doi.org/10.3390/rs13020308>.
- Castaldi F (2021) Sentinel-2 and landsat-8 multi-temporal series to estimate topsoil properties on croplands. *Remote Sensing* 13(17). <https://doi.org/10.3390/rs13173345>.
- Castaldi F, Palombo A, Santini F, Pascucci S, Pignatti S, and Casa R (2016) Evaluation of the potential of the current and forthcoming multispectral and hyperspectral imagers to estimate soil texture and organic carbon. *Remote Sensing of Environment* 179: 54–65. <https://doi.org/10.1016/j.rse.2016.03.025>.
- Castaldi F, Hueni A, Chabrilat S, et al. (2019) Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands. *ISPRS Journal of Photogrammetry and Remote Sensing* 147. <https://doi.org/10.1016/j.isprsjprs.2018.11.026>.
- Chabrilat S, Ben-Dor E, Cierniewski J, Gomez C, Schmid T, and van Wesemael B (2019) Imaging spectroscopy for soil mapping and monitoring. *Surveys in Geophysics* 40(3): 361–399. <https://doi.org/10.1007/s10712-019-09524-0>.
- Crucil G, Castaldi F, Aldana-Jague E, van Wesemael B, Macdonald A, and Oost K (2019) Assessing the performance of UAS-Compatible multispectral and hyperspectral sensors for soil organic carbon prediction. *Sustainability (Switzerland)* 11(7). <https://doi.org/10.3390/su11071889>.
- Demattè JAM, Fongaro CT, Rizzo R, and Safanelli JL (2018) Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. *Remote Sensing of Environment* 212: 161–175. <https://doi.org/10.1016/j.rse.2018.04.047>.
- Diek S, Schaepman ME, and de Jong R (2016) Creating multi-temporal composites of airborne imaging spectroscopy data in support of digital soil mapping. *Remote Sensing* 8(11). <https://doi.org/10.3390/rs8110906>.
- Dvorakova K, Heiden U, and van Wesemael B (2021) Sentinel-2 exposed soil composite for soil organic carbon prediction. *Remote Sensing* 13(9). <https://doi.org/10.3390/rs13091791>.
- Gholizadeh A, Žižala D, Saberioon M, and Borůvka L (2018) Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. *Remote Sensing of Environment* 218: 89–103. <https://doi.org/10.1016/j.rse.2018.09.015>.
- Gomez C, Viscarra Rossel RA, and McBratney AB (2008) Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. *Geoderma* 146(3–4): 403–411. <https://doi.org/10.1016/j.geoderma.2008.06.011>.
- Gomez C, Drost APA, and Roger JM (2015) Analysis of the uncertainties affecting predictions of clay contents from VNIR/SWIR hyperspectral data. *Remote Sensing of Environment* 156: 58–70. <https://doi.org/10.1016/j.rse.2014.09.032>.
- Guo L, Zhang H, Shi T, Chen Y, Jiang Q, and Linderman M (2019) Prediction of soil organic carbon stock by laboratory spectral data and airborne hyperspectral images. *Geoderma* 337: 32–41. <https://doi.org/10.1016/j.geoderma.2018.09.003>.
- Hively WD, Lamb BT, Daughtry CST, Serbin G, Dennison P, Kokaly RF, Wu Z, and Masek JG (2021) Evaluation of SWIR crop residue bands for the landsat next mission. *Remote Sensing* 13(18). <https://doi.org/10.3390/rs13183718>.
- Kopačková V and Ben-Dor E (2016) Normalizing reflectance from different spectrometers and protocols with an internal soil standard. *International Journal of Remote Sensing* 37(6): 1276–1290. <https://doi.org/10.1080/01431161.2016.1148291>.
- Laamrani A, Berg AA, Voroney P, et al. (2019) Ensemble identification of spectral bands related to soil organic carbon levels over an agricultural field in southern Ontario, Canada. *Remote Sensing* 11(11). <https://doi.org/10.3390/rs11111298>.
- Lu P, Wang L, Niu Z, Li L, and Zhang W (2013) Prediction of soil properties using laboratory VIS-NIR spectroscopy and Hyperion imagery. *Journal of Geochemical Exploration* 132: 26–33. <https://doi.org/10.1016/j.gexplo.2013.04.003>.
- Rogge D, Bauer A, Zeidler J, Mueller A, Esch T, and Heiden U (2018) Building an exposed soil composite processor (SCMaP) for mapping spatial and temporal characteristics of soils with Landsat imagery (1984–2014). *Remote Sensing of Environment* 205: 1–17. <https://doi.org/10.1016/j.rse.2017.11.004>.
- Shi P, Castaldi F, van Wesemael B, and van Oost K (2020) Large-scale, high-resolution mapping of soil aggregate stability in croplands using APEX hyperspectral imagery. *Remote Sensing* 12(4). <https://doi.org/10.3390/rs12040666>.

- Stenberg B, Viscarra Rossel RA, Mouazen AM, and Wetterlind J (2010) Visible and near infrared spectroscopy in soil science. *Advances in Agronomy*. vol. 107. Elsevier. [https://doi.org/10.1016/S0065-2113\(10\)07005-7](https://doi.org/10.1016/S0065-2113(10)07005-7).
- Stevens A, Udelhoven T, Denis A, et al. (2010) Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma* 158(1–2). <https://doi.org/10.1016/j.geoderma.2009.11.032>.
- Tóth G, Jones A, and Montanarella L (2013) The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. *Environmental Monitoring and Assessment* 185(9): 7409–7425. <https://doi.org/10.1007/s10661-013-3109-3>.
- Tziolas N, Tsakiridis N, Chabrillat S, et al. (2021) Earth observation data-driven cropland soil monitoring: A review. *Remote Sensing* 13(21). <https://doi.org/10.3390/rs13214439>.
- Vaudour E, Gomez C, Loiseau T, et al. (2019a) The impact of acquisition date on the prediction performance of topsoil organic carbon from Sentinel-2 for croplands. *Remote Sensing* 11(18). <https://doi.org/10.3390/rs11182143>.
- Vaudour E, Gomez C, Fouad Y, and Lagacherie P (2019b) Sentinel-2 image capacities to predict common topsoil properties of temperate and Mediterranean agroecosystems. *Remote Sensing of Environment* 223: 21–33. <https://doi.org/10.1016/j.rse.2019.01.006>.
- Wehrhan M and Sommer M (2021) A parsimonious approach to estimate soil organic carbon applying unmanned aerial system (UAS) multispectral imagery and the topographic position index in a heterogeneous soil landscape. *Remote Sensing* 13(18). <https://doi.org/10.3390/rs13183557>.
- Zepp S, Heiden U, Bachmann M, Wiesmeier M, Steininger M, and van Wesemael B (2021) Estimation of soil organic carbon contents in croplands of Bavaria from SCMap soil reflectance composites. *Remote Sensing* 13(16). <https://doi.org/10.3390/rs13163141>.
- Zhang H (2022) *Unmanned Aerial Vehicle-Based Detection of Organic Carbon With Hyper- and Multispectral Sensors in Cropland Soils*. PhD thesis Belgium: UCLouvain.
- Zhang H, Shi P, Crucil G, van Wesemael B, Limbourg Q, and van Oost K (2021) Evaluating the capability of a UAV-borne spectrometer for soil organic carbon mapping in bare croplands. *Land Degradation and Development* 32(15): 4375–4389. <https://doi.org/10.1002/ldr.4043>.
- Žižala D, Minarik R, and Zádorová T (2019) Soil organic carbon mapping using multispectral remote sensing data: Prediction ability of data with different spatial and spectral resolutions. *Remote Sensing* 11(24). <https://doi.org/10.3390/rs11242947>.

Relevant websites

- <https://sagroups.ieee.org/4005/>—Institute of Electrical and Electronics Engineers (IEEE) standards organization; P4005—Standards and protocols for soil spectroscopy.
- <https://www.fao.org/global-soil-partnership/glosolan/en/>—GLOSOLAN (Global soil laboratory network) funded by FAO.
- <https://soilspectroscopy.org/>—Soil Spectroscopy 4 Global Good.
- <https://ejpsol.eu/soil-research/steropes/>—STEROPES financed by the European Joint Program Soil.
- <https://www.world-soils.com/>—World Soils project financed by the European Space agency.

Pedotransfer functions and their application to soil water dynamics

Marcel G Schaap^a, Yonggen Zhang^b, and Attila Nemes^{c,d}, ^aDepartment of Environmental Science, College of Agriculture and Life Sciences, University of Arizona, Tucson, AZ, United States; ^bSchool of Earth System Science, Tianjin University, Tianjin, P.R. China; ^cFaculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway; ^dDivision of Environment and Natural Resources, Norwegian Institute of Bioeconomy Research, Ås, Norway

© 2023 Elsevier Ltd. All rights reserved.

Introduction	643
General overview	643
PTF categories	644
PTF output	644
Predictors	645
PTFs with hierarchical predictors	646
PTF development strategies	647
Semi-physical approaches	647
Class and continuous PTFs	648
Regression approaches	648
Machine learning approaches	649
PTF replicas and PTF ensembles	649
Calibration metrics	650
Evaluation criteria	650
Example application	650
Background and assumptions	651
Estimation and simulation results	651
Summary	653
Acknowledgments	653
References	653

Abstract

Pedotransfer functions (PTFs) are equations that use basic soil attributes, such as soil texture and soil bulk density, which are easier to measure, to predict those that are harder to determine, such as soil hydraulic parameters, for modeling relevant earth system processes. Here, we give a general overview of PTF development with a focus on soil hydraulic attributes and discuss different categories of PTFs. Frequently used metrics for PTF optimization and evaluation are summarized. We conclude the Chapter by providing an example application of how PTFs are used in modeling soil water movement and how uncertainty in soil hydraulic parameter estimates propagates to derived state variable quantities.

Glossary

Hydraulic parameter Hydraulic parameter is a parameter that quantitatively describes a soil hydraulic property. Several hydraulic parameters are used in soil hydraulic functions.

Pedotransfer function Pedotransfer function is an equation that uses basic soil attributes, such as soil texture and soil bulk density, that are relatively easy to measure, to predict those that are harder to determine, such as soil hydraulic parameters.

Soil hydraulic characteristic Soil hydraulic characteristic is the relation between soil matric potential and soil water content (also known as the water retention characteristic) and the relation among soil matric potential, water content and soil hydraulic conductivity (hydraulic conductivity characteristic), and is also called the soil hydraulic function.

Soil hydraulic conductivity Soil hydraulic conductivity is a property of porous material, describing how easy a fluid (usually water) can pass through the pore space. It depends on the intrinsic permeability of the material, the degree of saturation, and on the density and viscosity of the fluid.

Soil hydraulic property Soil hydraulic property is a property that controls the water retention, infiltration, rate of water flow, and the fate of chemicals and pollutants in soil. This property is largely affected by various intrinsic soil characteristics, such as soil texture, bulk density, porosity, structure, and the interactions of these properties with each other.

Soil water retention Soil water retention is the relationship between the water content and the soil matric potential, which is characteristic for different types of soil, and is also called the soil moisture characteristic, water retention characteristic or pF characteristic.

Vadose Zone Vadose Zone is the variably saturated zone between the ground surface and the permanent water table of the groundwater.

Key points

- An overview of pedotransfer functions (PTFs) focusing on soil hydraulic attributes is presented.
- Different categories of PTFs and evaluation criteria to evaluate the developed PTFs are discussed.
- An example application of how PTFs are used in modeling soil water movement is provided.

Introduction

Soil science, hydrology, and earth science in general require accurate descriptions of the systems under study (Vereecken et al., 2022). The relevant domains typically range from individual soils, entire critical zone systems for ecosystem services, to detailed continental to globally distributed data to assess and predict changes in water, carbon, and global climate cycles. Our capability to monitor, visualize, and detect ecosystem changes by remote sensing has seen an astounding development over the past decades. It is now sometimes possible to “see” details with publicly available remote sensing platforms at a level of detail that was previously the domain of well-funded national security agencies. Likewise, developments in computer science allow us to establish detailed global datasets of what cannot be easily seen from space. In the past decades, detailed soil databases and geographic information systems (GIS) have been established to provide a highly detailed global distribution of soil texture, organic carbon, and bulk density, among many other soil attributes relevant to ecosystem features and services (Hengl et al., 2017).

The collection, archiving, documentation and publishing of historic as well as up-to-date earth systems data are necessary but often insufficient by themselves to make assessments of past, present, and future changes in local, regional, and global environmental systems. In many cases, we can only make those assessments by using deterministic or stochastic models to obtain the requisite data relevant to scientists, policymakers, and other stakeholders. These models require data that is either unavailable or only available at an insufficient level of detail. At the same time, the models themselves may not yield immediately useful output for the scientist, stakeholder, or policymaker. In many cases, available earth systems data and earth-systems model output must be transformed or translated to be useful.

The infiltration of water in soils is a critical aspect of the terrestrial hydrological cycle and is relevant, for example, for crops and ecosystems, trafficability of land, and flooding (Vereecken et al., 2019). Infiltration is, among other factors, governed by hydraulic conductivity and can be modeled using a range from simple statistical to complex 3D dynamic simulations. While hydraulic conductivity *can* be measured in the laboratory and field, it is expensive to do so, often requiring hundreds to thousands of dollars per soil profile. Suppose there was a need to simulate infiltration for all terrestrial ecosystems at a scale of 250 m; roughly half a billion measurements of hydraulic conductivity would be needed, which would require an insurmountable investment in financial and labor resources. Intuitively, however, hydraulic conductivity can be related to soil particle-size distribution, soil bulk density, and perhaps other factors. As a simple example: it is fairly obvious that a loose (coarse-grained) sandy soil should have a greater hydraulic conductivity and, therefore, a faster infiltration rate than a dense clay soil. Global soil texture data are available at a 250 m resolution (Hengl et al., 2017). If suitable deterministic or statistical models or even simple tabulations can be established that capture the relation between hydraulic conductivity and texture, soil bulk density, and geomorphological or geological factors, we would be able to estimate hydraulic conductivity for the requisite half billion grid-cells.

In essence, pedotransfer functions address the “major challenge for soil science . . . to “translate” data we have to data we need.” (Bouma, 1989). Although the above example has a hydrological focus, it is equally well possible to extend the concept to other relevant earth system factors, such as thermal properties of soils or ecosystems (as relevant for terrestrial energy balances), soil chemical transport (nutrient and other chemical balances), as well as biological parameters (Van Looy et al., 2017).

This chapter will first focus on the general context of pedotransfer functions for estimating soil hydraulic properties and briefly review their history. Next, we will briefly discuss the different pedotransfer functions (PTFs) categories. Finally, we will provide an example of how PTFs might be used in assessing a long-term water balance of a deep vadose zone.

General overview

The term “Pedotransfer function” was first published by Bouma and van Lanen (1987) with the following definition (Bouma, 1989):

Pedotransfer functions “... relate different characteristics to one another or to land qualities. Two types of pedotransfer functions are distinguished. Those with continuous characteristics are called: continuous pedotransfer functions. Those using class characteristics are called: class pedotransfer functions.”

Land Qualities in this context are:

“... attribute of land which acts in a distinct manner in its influence on the suitability of land for a specific kind of use. Land qualities may be expressed in a positive or negative way. Examples are moisture availability, erosion resistance, flooding hazard, nutritive value of pastures, accessibility.”

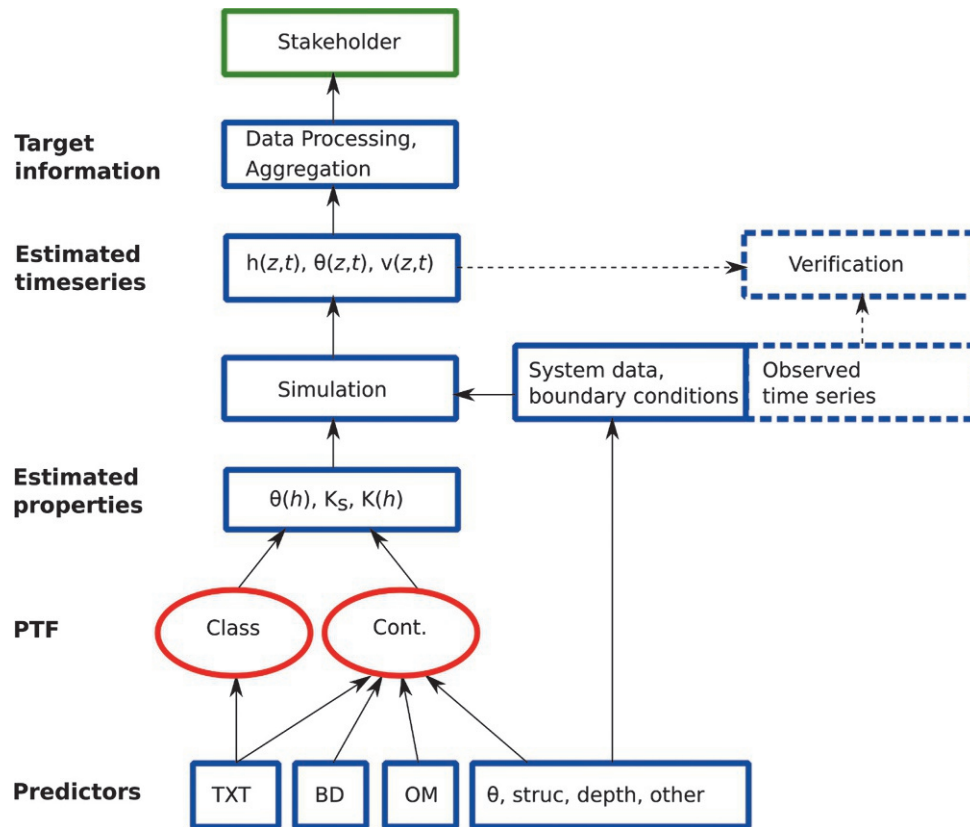


Fig. 1 PTFs are usually a small part of a larger project enabling the generation of data that is important for scientists, engineers, policymakers, or other stakeholders. They do their work by estimating difficult to measure hydraulic properties, $\theta(h)$ and $K(h)$, from more widely available basic soil properties such as soil texture (TXT), bulk density (BD), organic matter (OM) or other soil characteristics. Various types of PTF exist, and here we sketched two: one is a class PTF, the other one is a continuous PTF. Depending on the application, practitioners may prefer one over the other.

The transfer function concept outlined by Bouma and van Lanen is general and is used in a wide range of earth science applications (Van Looy et al., 2017). In soil science and hydrology, PTFs are usually employed to convert basic soil data such as soil texture, soil bulk density, and other informative data into more difficult to measure soil hydraulic properties, such as soil water retention $\theta(h)$ and hydraulic conductivity $K(h)$, where h is matric potential.

Fig. 1 (simplified after Bouma, 1989) depicts two types of PTFs, i.e., class and continuous PTFs, which predict hydraulic properties for a soil water model. The model is driven by boundary conditions and simulates time series of matric potential h , water content θ , and water flux v as a function of time t and position z , denoted as $h(t, z)$, water content $\theta(t, z)$, and water flux $v(t, z)$. Matric potential and soil water potentials in general are the physical drivers of water fluxes in soils, resulting in accumulation or depletion of soil water content. These state variables must be further aggregated to provide meaningful outputs for stakeholders.

The hydraulic properties are one component in the system depicted in Fig. 1, but are often difficult to obtain in the laboratory. However, usually available are soil survey or land classification maps, digital terrain models, and detailed interpolated maps for soil attributes, such as texture, bulk density, and organic carbon or organic matter content (Tóth et al., 2014). Hydraulic PTFs enable us to transform the soils and landscape data into the hydraulic properties needed for a simulation model (or any other application that needs hydraulic properties). There is a fairly large variety of ways by which existing PTFs can be classified: e.g., by what kind of hydraulic properties are estimated, by required input data (or predictors), by mathematical structure, and whether single or multiple models are used to make estimates.

PTF categories

PTF output

Before discussing the different methods by which PTFs can be established, we must first consider their desired output. As mentioned previously, hydraulic properties are difficult to obtain directly in the field or in the laboratory, but databases representative of a wide range of soil types of such properties and related data have been collected by various groups around the world. Such databases typically contain hydraulic properties as pairs of water content and matric potential (θ, h), hydraulic conductivity and matric potential (K, h), or hydraulic conductivity and water content (K, θ), each of which has been determined for multiple values of θ and

Table 1 Hydraulic functions: the Brooks-Corey, van Genuchten, and Kosugi retention functions and their corresponding hydraulic conductivity formulation through application in the [Mualem \(1976\)](#) pore-connectivity theory.

	Retention function $\theta(h)$	Hydraulic conductivity function $K(h)$
Brooks and Corey	$\theta = \begin{cases} \theta_r + (\theta_s - \theta_r) \left(\frac{h_0}{h} \right)^\lambda & h < h_0 \\ \theta_s & h \geq h_0 \end{cases}$	$K = \begin{cases} K_s S_e^{2/n+L+2} & h < h_0 \\ K_s & h \geq h_0 \end{cases}$ with $S_e = (\theta(h) - \theta_r)/(\theta_s - \theta_r)$
Parameters	4: $\theta_r, \theta_s, h_0, \lambda$	1: K_s with $L = 0.5$ 2: K_s, L
van Genuchten	$\theta = \begin{cases} \theta_r + \frac{(\theta_s - \theta_r)}{[1 + \alpha h ^n]^m} & h < 0 \\ \theta_s & h \geq 0 \end{cases}$	$K = \begin{cases} K_s S_e^L \left[1 - (1 - S_e^{1/m})^m \right]^2 & h < 0 \\ K_s & h \geq 0 \end{cases}$ with $S_e = (\theta(h) - \theta_r)/(\theta_s - \theta_r)$
Parameters	4: $\theta_r, \theta_s, \alpha, n, m = 1-1/n$	1: K_s with $L = 0.5$ 2: K_s, L
Kosugi	$\theta(h) = \theta_r + \frac{1}{2} (\theta_s - \theta_r) \operatorname{erfc} \left[\frac{\ln(h/h_m)}{\sqrt{2}\sigma} \right]$	$K = K_s S_e^L \left\{ \frac{1}{2} \operatorname{erfc} \left[\frac{\ln(h/h_m)}{\sqrt{2}\sigma} \right] + \frac{\sigma}{\sqrt{2}} \right\}^2$ with $S_e = (\theta(h) - \theta_r)/(\theta_s - \theta_r)$
Parameters	4: $\theta_r, \theta_s, h_m, \sigma$	1: K_s with $L = 0.5$ 2: K_s, L

Models are given with their canonical parameters.

h for a particular soil, or soil sample. In many cases the measured sets of paired values are not uniform because they were obtained with different objectives or derived from different laboratories using different measurement techniques.

To harmonize hydraulic property datasets into uniform characteristics, it is often necessary to describe the relationship between soil water content, θ , and matric potential, h , with a soil water retention function. Many formulations exist, and in [Table 1](#) we provide three of the most common functions used in PTF development: the [Brooks and Corey \(1964\)](#) model (BC), the [van Genuchten \(1980\)](#) model (VG), and the [Kosugi \(1996\)](#) model. Each of these soil water retention functions has four parameters, two of which describe the maximum (saturated) water content, θ_s , and the residual water content, θ_r , which is a minimum possible value. The other two parameters ($\{h_0, \lambda\}$, $\{\alpha, n\}$, and $\{h_m, \sigma\}$ for BC, VG, and Kosugi, respectively), are known as shape parameters, which determine the position and sharpness of the decrease of water content from saturation to residual water content as a function of matric potential. Through the application of [Mualem's \(1976\)](#) pore-size theory, a corresponding hydraulic conductivity function can be derived with at least one parameter: saturated hydraulic conductivity K_s , i.e., the hydraulic conductivity at saturation where $\theta = \theta_s$ and $h = 0$. Parameter, L , determines the rate of decrease in hydraulic conductivity and is often set to 0.5 after [Mualem \(1976\)](#). As a result, water retention and hydraulic conductivity can be described with five parameters for each of the three hydraulic functions. PTFs that predict these parameters can therefore provide a complete, though not necessarily an accurate or precise description of both hydraulic functions. We refer the reader to [Brooks and Corey \(1964\)](#), [van Genuchten \(1980\)](#), and [Kosugi \(1996\)](#) for details regarding assumptions, mathematical derivation, and the interpretation of parameters.

If standardized hydraulic measurements are available at specific matric potentials, the PTF developer has the option to build a model that predicts moisture content at these specific points. Such point-PTFs have some benefits as the fitting of hydraulic functions—which imparts some error to the estimates—can be bypassed. Point PTFs are useful if the user is interested only in field capacity or wilting point, both of which are available in many databases for PTF development ([Tóth et al., 2014](#)). When the developer decides to estimate the hydraulic parameters of one of the models in [Table 1](#), the PTF is called a parametric PTF.

Predictors

Predictors are variables that, by theoretical or statistical analysis, should be related to desired quantities. In the case of PTFs for water movement, predictors are input variables that feed into mathematical formulae, which then provide information about the values of the five hydraulic parameters identified previously. There is no rule on which predictors *should* be used, and developers often use a

heuristic approach to identify which predictors are available to them and which combination of these works best in terms of objective statistical criteria. Successful predictors usually exhibit correlations with hydraulic data in initial data exploration and most PTFs at least include soil texture class or soil particle size data as the most basic input, since these properties represent the most fundamental physical characteristics of soils.

Additional input data commonly used in PTFs includes soil bulk density and organic carbon content or organic matter content. These properties will mainly affect porosity, soil water retention, and K_s . It should be noted that bulk density and organic carbon content are often intercorrelated and that using both as predictors may not substantially improve the accuracy of PTFs. Bulk density and porosity are very dynamic properties and are affected by land use change, tillage practice, trafficking, weathering, or drainage. Soil organic carbon content is much less dynamic and affects soil adsorption properties by its large surface area and its role as gluing agents, (e.g., microbial gums and polymeric substances), and therefore, it will also affect the soil water retention and other soil hydraulic properties. Minasny and McBratney (2018) investigated the effects of soil organic matter content on the available water capacity based on a meta-analysis from more than 50,000 measurements around the globe, and concluded that organic matter content in soil has a small effect on available water capacity.

Infrequently used PTF predictors include soil chemical, soil morphological properties and soil structural information. Soil chemical properties include pH, cation exchange capacity (CEC), sodium adsorption ratio (SAR), and calcium carbonate (CaCO_3) content and may be necessary to improve PTFs (Tóth et al., 2014). Soil morphological properties, such as pedogenic horizon type, consistence, mottles, and concretions, are typically described in the soil survey but are often qualitative or semi-quantitative, which is more challenging to integrate into PTFs. Soil structure is the spatial arrangement and ordering of soil particles, determining the pore size distribution, tortuosity, and connectivity. Soil structure can be utilized to represent the biological activities (soil fauna and vegetation), abiotic factors (freezing-thawing and wetting-drying processes), or the effects of tillage practices in the soil. The role of soil structure on hydraulic processes is complex, and soil structure itself is difficult to quantify; therefore its role is typically unaccounted for when developing PTFs, which is increasingly seen as a knowledge-gap (Van Looy et al., 2017).

Some PTFs have used soil water retention data as predictors in addition to basic soil properties, such as soil texture and bulk density (for example, see Rawls et al., 1992; Schaap and Leij, 1998). This choice may seem odd because the objective of many PTFs is to estimate water retention. The input of one or two water retention points can be justified by the fact that most other predictors only show a modest correlation with the desired hydraulic properties, which can be mitigated substantially by using a single retention point as a predictor. It is found that this strategy often substantially improves PTF performance, with the drawback that water retention at specific matric potentials must be available to use the particular PTF.

PTFs with hierarchical predictors

Developers of PTFs often use a combination of intuition and statistical algorithms to select predictors that yield the “best” PTF, given a database with predictors and hydraulic parameters. In general, such approaches tend to favor PTFs that use several predictors as input (e.g., sand, silt, clay, bulk density, and organic matter content). The drawback of such an approach is that all these predictors *must* be available when the PTF is used, which potentially limits the applicability of the PTF because not all predictors may always be available.

Rather than striving for the goal of the best PTF, it is sometimes advantageous to develop PTFs with more limited sets of predictors, especially if it is known which predictors likely users of the PTFs have access to. The literature provides several of these approaches, such as Rawls et al. (1992) and Vereecken et al. (1989, 1990). More recently, Tóth et al. (2014) proposed a structured approach to develop hierarchical point and parametric PTFs for use in spatial applications in which not all data may always be available. As an example, Table 2 shows Rosetta 3 (abbreviated as R3), a hierarchical PTF model developed by Zhang and Schaap (2017) with five PTFs that uses progressively more predictors to estimate soil hydraulic parameters, including K_s and soil water

Table 2 Performance of the hierarchical Rosetta 3 PTFs in estimating van Genuchten soil hydraulic parameters (see Table 1 for the parameters), as published by Zhang and Schaap (2017).

		RMSE Calibration	RMSE Independent data	RMSE Calibration
Rosetta 3 ^a (R3)	Predictors	Soil water content (θ) ($\text{cm}^3 \text{cm}^{-3}$)	Soil water content (θ) ($\text{cm}^3 \text{cm}^{-3}$)	Log transformed saturated soil hydraulic conductivity $\log(K_s)$ (–)
R3H1	USDA Class	0.074	0.0681	0.739
R3H2	Sand, Silt, Clay	0.072	0.0628	0.726
R3H3	Sand, Silt, Clay, and Bulk density	0.062	0.0589	0.672
R3H4	Sand, Silt, Clay, Bulk density, and θ_{33}	0.040	0.0342*	0.600
R3H5	Sand, Silt, Clay, Bulk density, θ_{33} , and θ_{1500}	0.039	0.0260*	0.600

Model evaluations were performed on an independent global dataset by Zhang et al. (2020), except those marked with * which were unpublished. The global data set was selected from NCSS and included 116,641 data points from 49,855 samples, but slightly fewer for R3H4 and R3H5 as requisite data was unavailable. Zhang et al. (2020) were not able to evaluate soil hydraulic conductivity (K_s) for the Rosetta 3 PTFs. θ_{33} and θ_{1500} are water contents at -33 kPa (field capacity) and -1500 kPa (plant wilting point), respectively. RMSE is root mean square error.

^aRosetta 3 is a hierarchical set of five PTF models that use progressively more predictors to estimate soil hydraulic parameters, including saturated hydraulic conductivity (K_s) and soil water retention parameters (θ_r , θ_s , α , n).

Table 3 The Rosetta3 R3H1 Class PTF for the van Genuchten (1980) hydraulic functions.

	θ_r	θ_s	$\log_{10}(\alpha)$	$\log_{10}(n)$	$\log_{10}(K_s)$
Clay	0.131	0.457	-2.067	0.0986	1.169
Silty clay	0.124	0.473	-2.000	0.105	0.983
Sandy clay	0.147	0.382	-1.601	0.0922	1.055
Clay loam	0.107	0.429	-2.002	0.143	0.849
Silty clay loam	0.120	0.470	-2.255	0.157	1.046
Sandy clay loam	0.0934	0.380	-1.906	0.116	1.122
Loam	0.0902	0.402	-2.197	0.153	1.125
Silty loam	0.0832	0.427	-2.465	0.191	1.266
Sandy loam	0.0606	0.380	-1.785	0.163	1.573
Silt	0.0650	0.472	-2.219	0.198	1.641
Loamy sand	0.0582	0.383	-1.609	0.230	2.034
Sand	0.0546	0.363	-1.484	0.462	2.808

The parameters α , n , and K_s are given as $\log_{10}$ values because these have approximately log-normal distributions. The unit for θ_r , θ_s , α , n , and K_s are $\text{cm}^3 \text{cm}^{-3}$, $\text{cm}^3 \text{cm}^{-3}$, cm^{-1} , dimensionless, and cm day^{-1} , respectively.

retention parameters (θ_r , θ_s , α , n). The first version of Rosetta can be found under the link (<https://www.ars.usda.gov/pacific-west-area/riverside-ca/agricultural-water-efficiency-and-salinity-research-unit/docs/model/rosetta-model/>). The first model (R3H1) is simply a class PTF for each USDA texture class (see Table 3 for the Rosetta 3 class PTF to estimate the parameters of the van Genuchten hydraulic functions). The second model (R3H2) utilized the sand, silt, and clay fraction of the soil as input, with estimated soil hydraulic parameters continuously varying with soil particle fractions. From the third to the fifth model (R3H3 to R3H5), bulk density, water content at -33 kPa (θ_{33}), and water content at -1500 kPa (θ_{1500}), respectively, were included as input. The motivation for using water contents as predictors in R3H4 and R3H5 was to maximize PTF performance for the comprehensive NCSS soil database, thus extending the availability of limited water retention data to a full set of hydraulic functions. Water retention at these two matric potentials were available in nearly 400,000 records for over 60,000 soils predominantly from the USA and its territories, as well as several thousands of soils from non-USA regions.

Table 2 illustrates that increasing the number of predictors generally improves PTF accuracy, here defined in terms of root mean square errors (RMSE) for soil water content (θ) and K_s . The class and texture-only models (R3H1 and R3H2, respectively) both have similar errors, but estimates improve substantially when bulk density is added with R3H3. Adding retention data at -33 kPa yields a further improvement, indicating that not all relevant information is carried in soil texture and bulk density. Supplementing a second retention point at 1500 kPa yields little additional improvement. The above indicates that R3H4 or R3H5 should be used if their predictors are available (which they often are in the NCSS data), but that users can fall back to R3H3, R3H2, or R3H1 if constrained by the availability of input data. Because models R3H1 to R3H5 have been developed on the same database it is more likely that estimates among different models are consistent, something that cannot be guaranteed if PTFs developed on different databases are used (see Schaap and Leij, 1998). Table 2 shows a rare case when the RMSEs for independent data are smaller than for the data used to develop the model, presumably because the independent NCSS data used uniform methodologies for measuring hydraulic properties and predictors, whereas there was a large diversity in measurement methodologies used to obtain the R3 data.

PTF development strategies

PTFs can be developed using several standard statistical techniques that we categorized below into some general classes. Fig. 2 shows a generalized approach that involves a database with both predictors and hydraulic parameters, a candidate PTF with an internal structure (e.g., linear equation or machine learning), and internal empirical coefficients. These coefficients are adjusted through an iterative model development process followed by the minimization of the objective function, which calculates the sum of the squared differences between the measured and predicted properties or parameters (see Section "Calibration metrics"). The value of the objective function allows an algorithm to adjust the internal model coefficients. We also refer the reader to Tóth et al. (2014) who presented a comprehensive structured analysis of what kind of PTF to develop depending on conditions of data availability and target application.

Semi-physical approaches

Nearly all PTFs are empirical, focusing on statistical relations between predictors and hydraulic parameters without relying on theoretical concepts. Semi-physical approaches exist, however, such as Arya and Paris (1981) who proposed a direct relationship that scales the soil particle size distribution into a soil pore size distribution. The pore size distribution is then converted into capillary pressure by relating to the Young-Laplace equation, whereas volumetric soil water content is obtained by distributing the total pore volumes from the bulk volume. The calculated capillary pressure and volumetric water content are then paired to obtain the soil water retention curve. Arya et al. (1999) refined the Arya and Paris (1981) model by considering different scaling parameters that converted the soil particle size distribution into the soil water retention curve. One of the disadvantages of this method is that

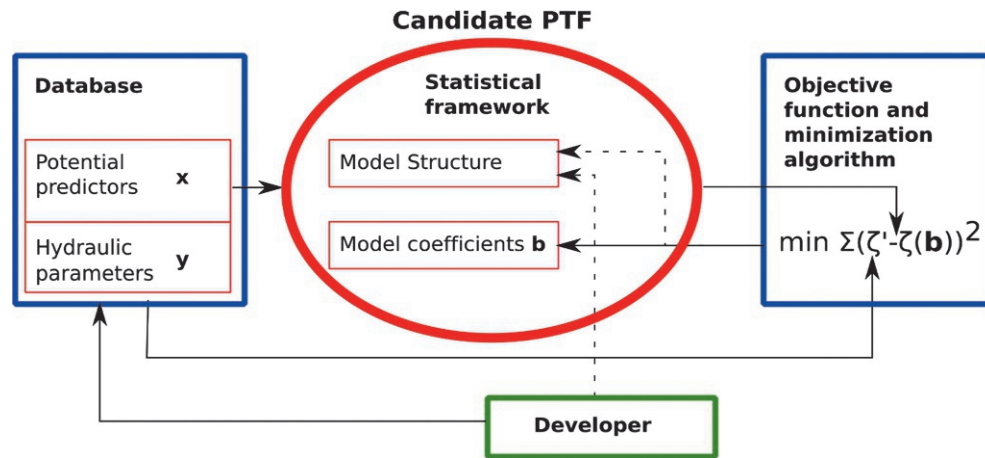


Fig. 2 Generalized PTF development schematic. There is nothing inherently magical about PTFs and their development uses well-established statistical techniques. A developer establishes a database with paired sets of predictors and hydraulic parameters. The developer also chooses a statistical framework (class PTF, regression, or machine learning) for the PTF and possibly decides upon the internal model structure (which can also be selected automatically in some frameworks). Initially, the model starts off with arbitrary model coefficients, but an iterative application of model predictions and objective function evaluations allows these coefficients to be optimized.

detailed soil particle size distribution data must be available, which makes it difficult to be applied in most practical cases. Recent developments in particle size analysis, however, may allow this approach to be revisited.

An alternative physical approach to establishing the relationship between input and output variables is to use computational fluid dynamics simulations. For example, [Zhang and Schaap \(2019\)](#) carried out a Lattice Boltzmann simulation based on a synthetic porous medium. They found that using particle size to estimate saturated hydraulic conductivity is not entirely correct and that hydraulic radius (pore volume over solid phase surface area) is more appropriate. However, this approach is computationally expensive and requires Computed Tomography (CT) and subsequent image analysis to derive the pore structure of soil samples. Similar approaches for water retention have been attempted (see [Schaap et al., 2007](#)), but the two-phase nature of this approach is even more demanding. Presently, estimates of hydraulic properties assisted by fluid dynamics simulations are still at the proof-of-concept stage without immediate usefulness for PTF users.

Class and continuous PTFs

[Bouma \(1989\)](#) identified two main types of PTFs: class and continuous PTFs (see also [Fig. 1](#)). Class PTFs are the simple yet widely used PTFs, which provide class-average soil hydraulic parameters as a look up table for each of the 12 USDA textural classes. Because of their simplicity, class PTFs have been widely used in soil hydrology and Earth system science. For example, the work by [Cosby et al. \(1984\)](#) and [Clapp and Hornberger \(1978\)](#) are employed by land surface models. Other class PTFs include the R3H1 model ([Table 3](#)), which is used in the Hydrus 1D soil water model ([Simunek et al., 2005](#)). Class PTFs are often the least demanding for data input as they can be used in conjunction with subjective data such as “hand texture” soil survey data. A distinct drawback of class PTFs is that they permit no variation of hydraulic properties within each class, but instead exhibit discrete changes at class boundaries. Class PTFs depend on pre-defined textural or other soil classes, but there is no guarantee that these class boundaries are appropriate for PTFs, as shown in a study that used k-means clustering to classify continuous PTF output ([Twarakavi et al., 2010](#)).

Continuous PTFs utilize continuous soil properties, such as sand, silt, clay fractions, bulk density, and soil organic carbon content, as input, therefore relieving the discontinuity issue of class PTFs. Continuous PTFs, however, require regression or machine learning approaches (see next two sections) to relate predictors of hydraulic parameters and are therefore more complex than class PTFs.

Regression approaches

Many of the well-established PTFs were developed based on regression approaches. [Gupta and Larson \(1979\)](#) were probably the pioneers in building a linear regression approach to estimate the soil water retention characteristics. The main advantage of the regression approach is that they can be easily implemented in spreadsheets or program code. However, the drawback of the approach is that the mathematical model structure must be chosen before calibration and that it may be difficult to capture complex non-linear relations among predictors and hydraulic parameters. The latter problem can partially be mitigated by using variable transformations prior to PTF development, or by the use of more complex or nested regression equations, as applied by, e.g., [Pachepsky et al. \(1998\)](#) using the group method of data handling (GMDH). [Table 4](#) shows the simple [Cosby et al. \(1984\)](#) PTF for estimating Brooks-Corey parameters; most other regression PTFs are considerably more complex.

Table 4 The two-variable Cosby et al. (1984) PTF for estimation of Brooks and Corey (1964) parameters.

Variable	Equation	Units
θ_r	0	$\text{cm}^3 \text{ cm}^{-3}$
θ_s	$(50.5 - 0.142 \times \text{sand} \% - 0.037 \times \text{clay} \%) / 100$	$\text{cm}^3 \text{ cm}^{-3}$
$\text{Log}_{10}(h_0)$	$1.54 - 0.0095 \times \text{sand} \% + 0.0063 \times \text{silt} \%$	cm
λ	$1 / (3.10 + 0.157 \times \text{clay} \% - 0.003 \times \text{sand} \%)$	–
$\text{Log}_{10}(K_s)$	$-0.884 + 0.0153 \times \text{sand} \%$	inch/h

Machine learning approaches

Machine learning approaches, also known as artificial intelligence, have become a powerful tool and offer exciting new opportunities in extracting patterns and insights from predictor-output relationships. Neural networks (NNs), inspired by the biological nervous systems, are probably among the most widely used. NNs typically consist of multi-input and multi-output data transformation structures containing adjustable coefficients and offer two distinct advantages over traditional regression approaches. Firstly, whereas regression approaches can only estimate one hydraulic parameter for each equation (Table 4), NNs can predict all five in a single computation step. Secondly, no mathematical model must be chosen a priori with NNs. Instead, the predictor-output relations are “learned” automatically by the NN optimization algorithm and implemented in coefficient matrices. Drawbacks of NNs are their high computation requirements and that their outputs are not easily published in a human-readable form, for example, compared with the Class PTF in Table 2 or the Cosby et al. (1984) model in Table 4. A NN essentially consists of matrices with floating point numbers, which by themselves defy interpretation: it is nearly impossible to tell what the NN is “doing.” Because of this, special care needs to be taken when establishing NN models in general as a sufficiently complex NN will be able to reproduce the calibration dataset verbatim, but have inferior performance on new data. It is therefore advisable to split calibration databases into independent calibration, validation, and testing subsets (Guo et al., 2017). The calibration subset is often used to fit the models; the validation set is utilized to estimate prediction error for selection of the calibrated model; the testing set is used for assessment and quantifying of the generalization error of the final selected model.

Support vector machines (SVMs) are an alternative method for building predictive models such as PTFs. Twarakavi et al. (2009) built SVM-based PTFs and found that SVMs outperformed NN-based PTFs, such as Rosetta (Schaap et al., 2001). In addition, other machine learning approaches, such as decision or regression trees (Tóth et al., 2015) and the k-nearest neighbor method (Nemes et al., 2006), have been widely used to develop PTFs. New machine learning technologies, such as deep neural networks, and convolutional neural networks appear to be promising future research avenues for PTF development.

PTF replicas and PTF ensembles

All PTFs have inherent (but often ignored) uncertainties in their predictions. For example, the values listed for the R3H1 class PTF in Table 2 may give the impression that these are the “true” values for each parameter in each USDA textural class. This is incorrect, and each of these parameters has a very substantial standard deviation (not listed), which for $\log_{10}(K_s)$ amounts to 0.5 or more (see Zhang and Schaap, 2017). In practice, this means that there is an uncertainty in K_s of half an order of magnitude. The source of uncertainty is inherent in the source data used to establish class or continuous PTFs: both predictors, as well as hydraulic parameters vary substantially even for the same soil.

Uncertainty for class PTFs can be computed as simple standard deviations or standard errors (as necessary for the application). Uncertainty in estimates by continuous PTFs is more difficult to assess and may need sophisticated propagation of error approaches, depending on the mathematical structure of the model. A fairly simple way to quantify uncertainty in statistical models, PTFs included, is to repeatedly subset into calibration and validation subsets, using, e.g., the Bootstrap algorithm (Schaap and Leij, 1998) or simple repeated splitting of the data using random selection without replacement. By calibrating the PTF on each of the subsets (or “replicas”), small variations in PTF estimates will result, which can be quantified by mean and variance, as well as higher order moments, provided a sufficient number of replicas are used. The R3H2 to R3H5 hierarchical models in Table 2 were calibrated on one thousand replicas, allowing the determination of correlation among predicted hydraulic parameters, as well as their uncertainty.

Applications in soil hydrology, and most Earth system models often use only one single PTF to estimate the soil hydraulic parameters (Van Looy et al., 2017). Neuman (2003) suggested that describing the complex system and quantifying the associated processes with only a single model and one parameterization often underestimates the model uncertainty and, therefore, leads to over-confidence in predictive capabilities and decision-making. Tebaldi and Knutti (2007) suggested that estimates based on the outputs from an ensemble of many models (multi-model ensemble estimates) generally increase the reliability and reduce the inconsistency of the model estimation. The improvement of predictive skills by multi-model ensemble is based on the theory that each member of the ensemble models is independent and the errors are inclined to cancel or at least be partly reduced. In addition, the uncertainty of the ensemble model should be reduced as the number of ensemble members increases, resulting in an improvement of the predictive skill relative to the constitutive ensemble members. The most important and challenging work to derive ensemble models, therefore, is to assign suitable weighting factors for each ensemble member. Guber et al. (2006) and Zhang et al. (2020) are variations on this approach.

Calibration metrics

The mathematical frameworks to establish PTFs between predictors and output variables are exceedingly diverse. With the exception of class PTFs all PTFs contain parameters that must be optimized by minimizing the differences between measured and estimated properties as

$$\chi^2 = \sum_{i=1}^N (\zeta'_i - \zeta_i(\mathbf{b}))^2 \quad (1)$$

where ζ'_i and ζ_i indicate estimated and measured hydraulic quantities, N is the number of measurements, $\mathbf{b}$ is a vector, containing the empirical PTF coefficients. ζ'_i and ζ_i can be used to indicate either the estimated and measured or fitted parameters of the soil hydraulic functions in parametric PTFs, such as the parameters of equations in Table 1, or data points of estimated and measured water retention or hydraulic conductivities in point PTFs. It is noted that some of the hydraulic parameters (e.g., hydraulic conductivity or parameters describing pore size distribution, such as n in the van Genuchten model) has to be log-transformed before optimizing Eq. (1). The motivation behind the transformation is to convert the distribution into a Gaussian distribution.

Evaluation criteria

When a PTF model is built or selected in an application, statistical criteria are available to evaluate and quantify the *accuracy* and *reliability* of PTFs in estimating soil hydraulic parameters. *Accuracy* is often used to quantify the differences between estimated and measured quantities for model building, whereas *reliability* is utilized to assess the differences between estimates and measurements on validation data. A PTF may perform accurately based on the calibration data, but the assessment of independent data may not show reliable estimation, especially when the validation datasets are significantly different from the calibration data. In addition, some factors may not be included as predictors, such as climatic conditions and geographical locations of the soils, leading to systematic errors. Minasny and Hartemink (2011) found that PTFs developed in temperate climate regions are likely to produce remarkable errors for soils in tropical regions. While this is often the case, data sets from similar climatic conditions often present different input-input and input-output correlation structures, and it has been suggested that it is the similarities or differences in data correlation structure rather than purely the climatic origin that will drive a PTFs performance for a given validation data set (Fuentes-Guevara et al., 2022).

Root mean square error (*RMSE*), mean error (*ME*), and coefficient of determination (R^2) are the most commonly used indicators to measure the model performance. The *RMSE* is defined as

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\zeta'_i - \zeta_i)^2} \quad (2)$$

ME is used to quantify the systematic bias of estimations relative to the measurements, defined as

$$ME = \frac{1}{N} \sum_{i=1}^N (\zeta'_i - \zeta_i). \quad (3)$$

ME is negative when the PTF underestimates the quantity being evaluated, and positive if it overestimates it. When both *RMSE* and *ME* are small, the PTFs performs well. However, note that *ME* is the average value over the entire evaluated set of data points, and it is possible that overestimation and underestimation for different portions of the data domain are canceled out. Therefore, *ME* should be interpreted with caution. Calculating absolute mean errors can partially help address this problem. An additional way to help diagnose the uneven distribution of residuals is by correlating estimation residuals to the model inputs and outputs. A PTF performs optimally if any such correlations are small, but this is difficult to assess due to the smoothing nature of PTFs (i.e., their general bias toward the population mean).

R^2 is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^N (\zeta_i - \zeta'_i)^2}{\sum_{i=1}^N (\zeta_i - \bar{\zeta})^2} \quad (4)$$

where $\bar{\zeta}$ is the average of ζ . R^2 normally ranges from 0 to 1, with higher values indicating a better fit.

Example application

In this section, we will provide an example of how PTFs might be used in a land surface or vadose zone study and illustrate how PTF uncertainty and variability propagates to derived quantities, such as vegetation transpiration, water storage, and groundwater recharge.

Background and assumptions

We selected three PTFs (R3H1, R3H2, and R3H3, see Table 2) from Zhang and Schaap (2017) and the PTF published by Wösten et al. (1999). More PTFs could have easily been evaluated but would have complicated this demonstration. PTF input data was derived from a previous study by Zhang et al. (2016), who studied infiltration in a deep vadose zone through a PTF-assisted model inversion. The purpose of the presented simulations is not to closely represent the conditions present at the site. Rather, we aim to illustrate the uncertainty and variability brought about by the PTFs themselves and hence, we simplified boundary conditions and some site characteristics.

The subject of the study is the Maricopa Infiltration site at the University of Arizona's Maricopa Agricultural Center (between Tucson and Phoenix in Southern Arizona). At this site, we have detailed textural and bulk density data for a 15-m deep alluvial profile (Schaap, 2013). Located in a semi-arid climate, this profile is unsaturated to a depth of 15 m where a perched groundwater table is found. A hypothetical type of vegetation was "grown" in the profile. The rooting depth of this vegetation was 100 cm, and the vegetation was assumed to be dormant between December 1st and May 1st. Vegetation coverage was assumed to be 50%, effectively meaning that the vegetation was exposed to 50% of potential evapotranspiration.

Long-term meteorological data are available through The Arizona Meteorological Network (AZMET), providing a 30-year record of daily measurements of precipitation and estimate of potential evaporation. Data for Tucson, about 100 km south, were selected as these provided the longest time series. For the purposes of this study, we increased precipitation by a factor of two, while keeping potential evaporation constant. This increase was necessary to simulate deep recharge to the groundwater. As such, the simulation did not reflect the reality present at the site's location.

The simulation was carried out with a command-line version of Hydrus-1D (Simunek et al., 2005) and input and output were generated or processed with a number of shell-scripts in a Linux environment. Hydraulic properties were generated with four PTFs for each 25 cm increment to 15 m depth (see Fig. 3). Two types of simulations were run: one for each of the four PTFs with the objective to demonstrate the variability among different PTFs. The objective of the second simulation was to illustrate the variability in simulated output brought about by uncertainty in a single PTF. To this end, we repeated the simulations for 100 of the 1000 replica estimates in the R3H3 model (see Zhang and Schaap, 2017), which contain substantial non-Gaussian bi-variate variability among the five estimated hydraulic parameters of the van Genuchten hydraulic functions (Table 1). For the purposes of this example, we set $L = 0.5$ (Mualem, 1976; see also Table 1).

Estimation and simulation results

The leftmost column of Fig. 3 shows the approximate color of the profile, as well as the texture and bulk density at each 25 cm depth increment. Immediately to the right of the profile, we graphed the estimated hydraulic parameters against depth for R3H1 (blue), R3H2 (green), R3H3 (red), and Wösten et al. (1999) (orange). For R3H3 we also plot the 10–90% probability (light gray) distribution and the 25–75% probability distribution (dark gray).

R3H1 (blue) is a class PTF with a tabulation of hydraulic parameters for each 12 USDA textural classes. For this reason, R3H1 estimates exhibit discrete jumps, as the texture at different depths belongs to different USDA classes. R3H2 (green) is a continuous function that only considers sand, silt, and clay percentages as input, which yields estimates close to those of the R3H1 class values. Bulk density is used in R3H3 and, since bulk density increases with depth, the estimated θ_s and K_s values decrease because the total porosity decreases. Because R3H1 and R3H2 do not use bulk density as the input, they overestimate the θ_s at deeper depths.

With the exception of θ_s and K_s (as discussed), the R3 hierarchical models tend to produce outputs similar to R3H1 and R3H2. This is because each of these models has been calibrated on essentially the same data, except that the number of input variables was more limited for R3H1 and R3H2. The Wösten PTF predicts rather different output (except for θ_s), simply owing to the fact that it was calibrated on data of a different source. The Wösten PTF produces lower θ_r (it is set constant at $0.01 \text{ cm}^3 \text{ cm}^{-3}$), n and K_s values, but higher values for α . We also show the parameter uncertainty for R3H3 as 10–90% (light gray) and 25–75% (dark gray) quantiles. The uncertainty is considerable and tends to increase with depth because most soil databases lack data from high-density samples, leading to uncertainty in the predictions for such soils.

Simulation results appear on the right-hand side of the figure, with the top panel displaying atmospheric boundary conditions (black: precipitation, blue: potential transpiration), root zone matric potential (second panel), total profile water storage in the third panel, and infiltration, transpiration, evaporation (offset), and drainage fluxes in the bottom panel. Colored lines, which sometimes overlap, indicate the PTFs mentioned previously. The light gray and dark gray areas are valid for the 10–90%, and 25–75% quantiles for the simulation using R3H3, with the black line representing the mean of simulated values for each of the 100 replicas.

Yearly evaporative demand is nearly 2200 mm, while this simulation only has an average of 540 mm annual precipitation. As a result, the internal dynamics of the profile are strongly controlled by root water uptake, which becomes leaf transpiration, and evaporation from the soil surface. Even with doubled precipitation, there is a strong rainfall deficit, resulting in a root zone that dries out quickly starting in May (when the vegetation becomes active) and—provided there is rainfall—wets up again in December, when the vegetation becomes dormant. As a result, the root zone matric potential is binary: it switches from a high value of around -300 cm in the winter period to -9000 cm during the growing season, at which matric potential the model was parameterized not to allow any root-water uptake. Owing to the strong atmospheric demand, there is very little substantial difference among PTFs with

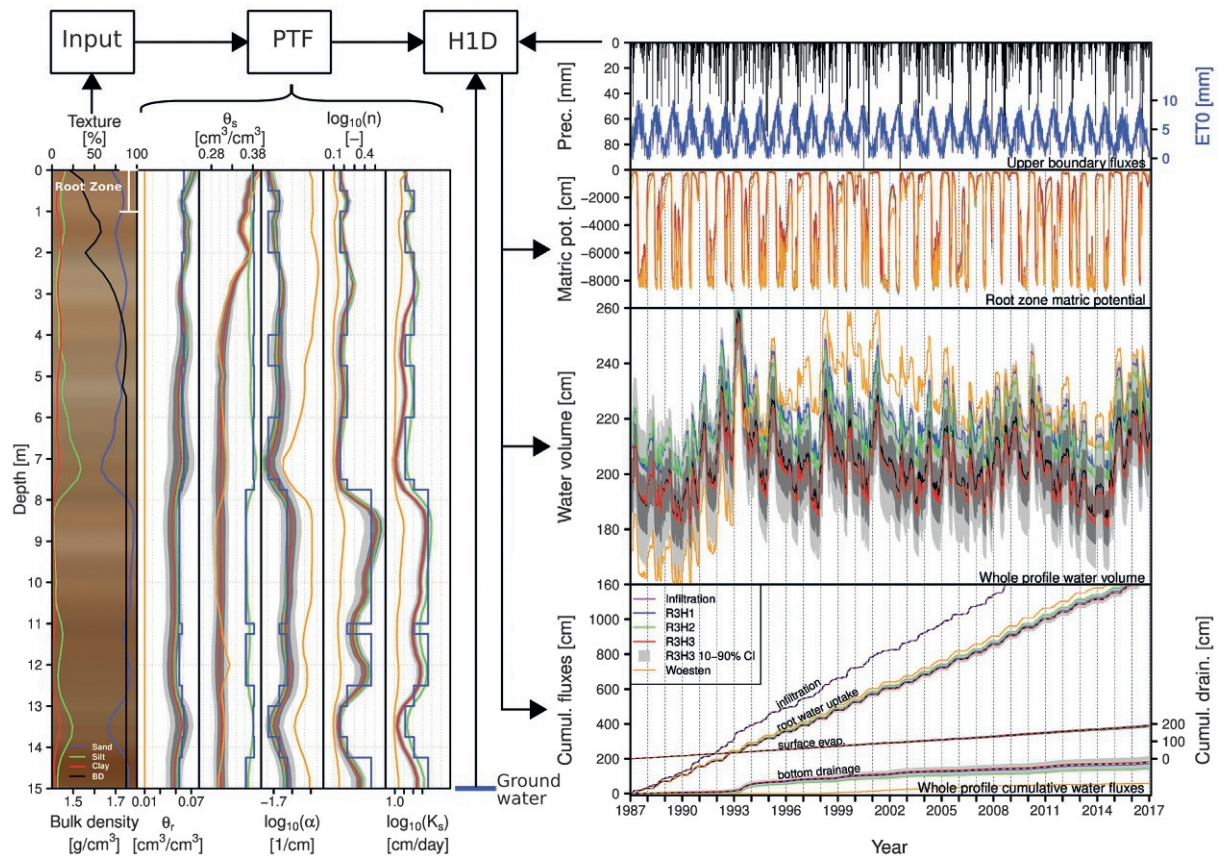


Fig. 3 Implementation of PTF estimates within a vadose zone context. The leftmost column shows the approximate coloring of a 15-m-deep vadose zone profile at the Maricopa infiltration site. Colored lines in this column indicate sand, silt and clay percentages (blue, green, and red, respectively) as well as bulk density. Estimates of the VG parameters and K_s are shown in the adjacent five columns for the R3H1 (blue) class PTF, and the R3H2 (green), R3H3 (red), and Wösten (orange) continuous PTFs. For the R3H3 we also provide uncertainty in the PTF predictions as 10–90% and 25–75% quantiles (light and dark gray, respectively). The graphs on the right-hand side indicate the input and processed results of Hydrus-1D simulations for a 30-year time series. In the top diagram we show precipitation (black) and potential evapotranspiration (blue), which are boundary conditions to the simulation, as well as a constant groundwater level at 15-m depth. The second diagram holds the mean root-zone (0–1 m depth) matric potential. The third graph provides the mean water volume for the entire 15-m profile. The bottom graph holds the cumulative water balance fluxes: infiltration (purple), root water uptake (or plant transpiration), evaporation from the soil surface, and bottom drainage to the groundwater. The colors in these diagrams correspond to those of the PTF prediction. Single simulations were performed on VG + K_s parameters predicted by the R3H1, R3H2, and Wösten PTFs; for the R3H3 model we carried out simulations for the first 100 replicas that underlie the Rosetta PTFs.

regard to matric potential. This can be understood by realizing that matric potential is a controlling factor in root water uptake, surface evaporation, and—with gravitational potential—vertical transport of liquid water.

By contrast, very substantial differences exist among the PTFs in terms of total profile water storage (third right panel of Fig. 3, note that the vertical axis does not start at zero). It appears that R3H1 typically simulates “wetter” profiles than R3H2, which is wetter than R3H3. These systematic differences are because these three hierarchical PTFs are based on different predictors (see Table 2), and therefore predict different values of hydraulic parameters (see the left-hand side of Fig. 3), which results in different simulations. The Wösten PTF, which was established on a different calibration database, has different patterns still: it starts drier than Rosetta, until the year of 1993, when it becomes wetter than any of the R3 results. This is due to an extreme rainfall event for which each of the three R3 models produces more drainage than does the simulation based on the Wösten PTF.

The third panel also shows considerable variation among the 100 replica simulations for R3H3. This implies that even for a single PTF, there is substantial uncertainty not only in the hydraulic parameters it predicts, but also in the outcome of simulations that rely on these estimates (here about 20–40%). It is also important to point out that there are differences between the red and black lines. The red line reflects the single simulation based on the mean parameters of R3H3. By contrast, the black line reflects the mean of the simulations, which is slightly wetter. The difference between the two can be the result of two factors. Firstly, it is known that the R3 model estimates have non-Gaussian distributions with heavy tails (Zhang and Schaap, 2017). Secondly, vadose zone water transport has nonlinear characteristics, which can skew simulated outcomes.

The bottom diagram in Fig. 3 shows cumulative fluxes: infiltration (purple), root water uptake, evaporation (which is offset by 200 mm for visualization purposes) and bottom of profile drainage that essentially translates to groundwater recharge. Infiltration

(precipitation) is the same for all simulations as no runoff was simulated. There is some PTF-induced variability in root water uptake (simulations based on the Wösten PTF generate higher amounts) but surface evaporation is nearly the same for all simulations. Relatively large variability (and for R3H3 large uncertainty) exists for simulated bottom-of-profile drainage. Here the simulation based on the Wösten PTF simulates the smallest amount, whereas for R3H1 to R3H3, and the replica mean are quite similar. The variability about the replica mean is around 20–30% of the final value, implying that there is uncertainty in the simulated recharge.

The above results are for only one profile in a semi-arid deep vadose zone, and different soil, vegetation, and climate conditions may yield different results. We also make no claim here as to which PTF is actually superior, but merely emphasize PTF variability and uncertainty. Overall, the above analysis illustrates a number of important points.

- Irrespective of the used PTF, the simulations yield similar (almost identical) results in terms of matric potential. This may be convenient, since many soil-vegetation-hydrology processes are driven by matric potential, although it should also be mentioned that in the present case the extreme ET demand drives all simulations to a similar state during the growing season: a dry state with limited water supply to the vegetation.
- At the same time, applications that rely on the *amount* of water in soils should carefully consider the variability displayed in the third right panel: different PTFs yield different amounts. Although the simulated temporal dynamics induced by different PTFs are approximately the same, volume (mass) based applications require careful selection of PTFs. Studies that rely on the effective soil dielectric constant or rely on variations in gravitational acceleration (e.g., the GRACE satellites) may be especially sensitive to variability among PTFs.
- The choice of PTF may also affect simulations of large-scale groundwater recharge and energy partitioning at the surface. Simulations that generate higher transpiration amounts tend to have lower recharge amounts.

Summary

Direct measurement of certain soil attributes is often difficult, expensive, time-consuming, and impossible at larger spatial scales. As an alternative, the science community has developed *pedotransfer functions* (PTFs) to estimate these from more readily available soil properties (e.g., soil texture, bulk density, and organic carbon content). This allows the estimation of soil hydraulic parameters, as well as other relevant earth system properties, such as thermal properties, solute transport, and biogeochemical parameters. Reliable soil hydraulic parameters, for example, are critical for the representation of soil hydrological processes (e.g., partitioning of water at the land surface, movement and redistribution of water within the soil profile, soil evaporation and plant transpiration, and deep drainage at the bottom of the root zone) in soil science, hydrology, and earth system science.

In this chapter we focused on PTFs for soil hydraulic parameters. A general overview of PTF development was given, and different categories of PTFs were discussed, including different desired parameters of soil hydraulic functions, predictors, and development strategies. Semi-physical approaches, such as the work done by Arya and Paris (1981), and physical approaches carried out by computational fluid dynamics were briefly analyzed. Empirical approaches were briefly introduced, including class and continuous PTFs, regression approaches, and more advanced machine learning approaches. Uncertainty is inherent to PTFs and can be quantified by model replication strategies. We highlighted developing PTF ensemble models to increase reliability and reduce the inconsistency of the PTF estimation.

An example application of how PTFs might be used in soil hydrology and how the uncertainty of the estimated soil hydraulic parameters from PTF propagate to derived state variable quantities, such as matric potential, water volume, etc., was provided. The simulation examples show that although different PTFs can reach consistent results in terms of matric potential for a vegetated semi-arid site, variable and uncertain results are obtained for simulated water volumes and fluxes. The choice of PTF can therefore significantly impact derived quantities and conclusions while conveying significant uncertainty.

Acknowledgments

Y. Zhang was supported by the National Natural Science Foundation of China (grant number: 42077168) and the National Natural Science Foundation of Tianjin (grant number: 20JCQNJC01660). M.G. Schaap was supported by USDA-NIFA project W-4188.

References

- Arya LM and Paris JF (1981) A physicoempirical model to predict the soil moisture characteristic from particle-size distribution and bulk density data. *Soil Science Society of America Journal* 45: 1023–1030.
- Arya LM, Leij FJ, van Genuchten MT, and Shouse PJ (1999) Scaling parameter to predict the soil water characteristic from particle-size distribution data. *Soil Science Society of America Journal* 63: 510–519.
- Bouma J (1989) In: Stewart BA (ed.) *Using Soil Survey Data for Quantitative Land Evaluation*, pp. 177–213. New York, NY: Springer.
- Bouma J and van Lanen HAJ (1987) Transfer functions and threshold values: From soil characteristics to land qualities. In: *Quantified Land Evaluation, Proceedings of a Workshop. ISSS/SSSA, Washington, D.C. ITC Publ. Enschede, The Netherlands.*
- Brooks and Corey (1964) *Hydraulic Properties of Porous Media. Hydrology Paper No. 3.* Fort Collins, CO: Civil Engineering Department, Colorado State University.

- Clapp RB and Hornberger GM (1978) Empirical equations for some soil hydraulic properties. *Water Resources Research* 14: 601–604.
- Cosby BJ, Hornberger GM, Clapp RB, and Ginn TR (1984) A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils. *Water Resources Research* 20: 682–690.
- Fuentes-Guevara MD, Armindo RA, Timm LC, and Nemes A (2022) Data correlation structure controls pedotransfer function performance. *Journal of Hydrology* 614: 128540.
- Guber AK, Pachepsky YA, van Genuchten MT, Rawls WJ, Šimunek J, Jacques D, Nicholson TJ, and Cady RE (2006) Field-scale water flow simulations using ensembles of pedotransfer functions for soil water retention. *Vadose Zone Journal* 5: 234.
- Guo C, Pleiss G, Sun Y, and Weinberger KQ (2017) On calibration of modern neural networks. In: *International Conference on Machine Learning*, pp. 1321–1330. PMLR.
- Gupta S and Larson WE (1979) Estimating soil water retention characteristics from particle size distribution, organic matter percent, and bulk density. *Water Resources Research* 15: 1633–1635.
- Hengl T, Mendes De Jesus J, Heuvelink GBM, Ruiperez Gonzales M, Kilibarda M, Blagotic A, Shangguan W, Wright MN, Geng X, Bauer-Marschallinger B, Guevara MA, Vargas R, Macmillan RA, Batjes NH, Leenaars JGB, Ribeiro E, Wheeler I, Mantel S, and Kempen B (2017) SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One* 12: 1–40.
- Kosugi K (1996) Lognormal distribution model for unsaturated soil hydraulic properties. *Water Resources Research* 32: 2697–2703.
- Minasny B and Hartemink AE (2011) Predicting soil properties in the tropics. *Earth-Science Reviews* 106: 52–62.
- Minasny B and McBratney AB (2018) Limited effect of organic matter on soil available water capacity. *European Journal of Soil Science* 69: 39–47.
- Mualem Y (1976) A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research* 12: 513–522.
- Nemes A, Rawls WJ, and Pachepsky YA (2006) Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. *Soil Science Society of America Journal* 70: 327–336.
- Neuman SP (2003) Maximum likelihood Bayesian averaging of uncertain model predictions. *Stochastic Environmental Research and Risk Assessment* 17: 291–305.
- Pachepsky Y, Rawls W, Gimenez D, and Watt JPC (1998) Use of soil penetration resistance and group method of data handling to improve soil water retention estimates. *Soil and Tillage Research* 49: 117–126.
- Rawls WJ, Ahuja LR, and Brakensiek DL (1992) Estimating soil hydraulic properties from soils data. In: *Indirect Methods for Estimating the Hydraulic Properties of Unsaturated Soils*, pp. 329–340. Riverside, CA: University of California.
- Schaap MG (2013) Description, analysis, and interpretation of an infiltration experiment in a semiarid deep vadose zone. *Advances in Hydrogeology* 159–183. 9781461464.
- Schaap MG and Leij FJ (1998) Database-related accuracy and uncertainty of pedotransfer functions. *Soil Science* 163: 765–779.
- Schaap MG, Leij FJ, and van Genuchten MT (2001) ROSETTA: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *Journal of Hydrology* 251: 163–176. [https://doi.org/10.1016/S0022-1694\(01\)00466-8](https://doi.org/10.1016/S0022-1694(01)00466-8).
- Schaap MG, Porter ML, Christensen BSB, and Wildenschild D (2007) Comparison of pressure-saturation characteristics derived from computed tomography and lattice Boltzmann simulations. *Water Resources Research* 43: W12S06. <https://doi.org/10.1029/2006WR005730>.
- Šimunek J, Van Genuchten MT, and Sejna M (2005) *The HYDRUS-1D Software Package for Simulating the One-Dimensional Movement of Water, Heat, and Multiple Solutes in Variably-Saturated Media*. University of California-Riverside Research Reports 3 1–240.
- Tebaldi C and Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365: 2053–2075.
- Tóth B, Makó A, and Tóth G (2014) Role of soil properties in water retention characteristics of main Hungarian soil types. *Journal of Central European Agriculture* 15: 137–153.
- Tóth B, Weynants M, Nemes A, Makó A, Bilas G, and Tóth G (2015) New generation of hydraulic pedotransfer functions for Europe. *European Journal of Soil Science* 66: 226–238.
- Twarakavi NKC, Šimunek J, and Schaap MG (2009) Development of pedotransfer functions for estimation of soil hydraulic parameters using support vector machines. *Soil Science Society of America Journal* 73: 1443–1452.
- Twarakavi NKC, Šimunek J, and Schaap MG (2010) Can texture-based classification optimally classify soils with respect to soil hydraulics? *Water Resources Research* 46.
- van Genuchten MT (1980) A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal* 44: 892–898.
- Van Looy K, Bouma J, Herbst M, Koestel J, Minasny B, Mishra U, Montzka C, Nemes A, Pachepsky YA, Padarian J, Schaap MG, Tóth B, Verhoef A, Vanderborght J, van der Ploeg MJ, Weihermüller L, Zacharias S, Zhang Y, and Vereecken H (2017) Pedotransfer functions in earth system science: Challenges and perspectives. *Reviews of Geophysics* 55: 1199–1256.
- Vereecken H, Maes J, Feyen J, and Darius P (1989) Estimating the soil moisture retention characteristic from texture, bulk density, and carbon content. *Soil Science* 148: 389–403.
- Vereecken H, Maes J, and Feyen J (1990) Estimating unsaturated hydraulic conductivity from easily measured soil properties. *Soil Science* 149: 1–12.
- Vereecken H, Weihermüller L, Assouline S, Šimunek J, Verhoef A, Herbst M, Archer N, Mohanty B, Montzka C, and Vanderborght J (2019) Infiltration from the pedon to global grid scales: An overview and outlook for land surface modelling. *Vadose Zone Journal* 18: 1–53.
- Vereecken H, Amelung W, Bauke SL, Bogaen H, Brüggemann N, Montzka C, Vanderborght J, Bechtold M, Blöschl G, Carminati A, Javaux M, Konings AG, Kusche J, Neuweiler I, Or D, Steele-Dunne S, Verhoef A, Young M, and Zhang Y (2022) Soil hydrology in the Earth system. *Nature Reviews Earth & Environment* 1–15.
- Wösten JHM, Lilly A, Nemes A, and Le Bas C (1999) Development and use of a database of hydraulic properties of European soils. *Geoderma* 90: 169–185.
- Zhang Y and Schaap MG (2017) Weighted recalibration of the Rosetta pedotransfer model with improved estimates of hydraulic parameter distributions and summary statistics (Rosetta3). *Journal of Hydrology* 547: 39–53.
- Zhang Y and Schaap MG (2019) Estimation of saturated hydraulic conductivity with pedotransfer functions: A review. *Journal of Hydrology* 575: 1011–1030.
- Zhang Y, Schaap MG, Guadagnini A, and Neuman SP (2016) Inverse modeling of unsaturated flow using clusters of soil texture and pedotransfer functions. *Water Resources Research* 52: 7631–7644.
- Zhang Y, Schaap MG, and Wei Z (2020) Development of hierarchical ensemble model and estimates of soil water retention with global coverage. *Geophysical Research Letters* 47: e2020GL088819.

Geostatistics: Modelling spatial variation

Richard Webster, Rothamsted Research, Harpenden, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction	656
Random variables and random functions	656
Estimating the variogram	658
Robust estimators	659
Models for variograms	659
Spherical	660
Exponential	660
Models with reverse curvature near the origin	660
Unbounded models	661
Anisotropy	661
Fitting the models	662
Combining trend and random fluctuation	663
Residual maximum likelihood estimation	663
Classification as external drift	663
Coregionalization—simultaneous variation in two or more variables	664
Modelling the coregionalization	666
Example	667
Spatial prediction—kriging	668
Ordinary kriging	668
Cokriging	669
Other forms of kriging	669
Universal kriging and kriging with external drift	669
Indicator and disjunctive kriging	669
Applications and new developments	669
References	670

Abstract

Geostatistics, which was developed originally for mining, is now widely used for prediction of soil conditions at unvisited sites and for mapping from sample survey data. The technique is based on models of spatial variation. Soil properties are treated as the outcomes of random spatially correlated processes, to which may be added fixed effects of geographic trend or correlation with other environmental variables. Prediction proceeds from sample data in two stages, namely (1) estimation of spatial covariances or semivariances to which are fitted plausible variogram models, and (2) kriged predictions based on the variogram models. A figure shows how comparisons between data can be binned to estimate semivariances; others show popular variogram models. The techniques for stage 1 are illustrated with examples from soil survey; equations define the kriging systems to predict without bias and minimize the prediction error variances for stage 2. Residual maximum likelihood (REML) is now best practice for estimating simultaneously coefficients of fixed effects and parameters of the variograms of random residuals for universal kriging and kriging with external drift. Thousands of published accounts attest to the merits of geostatistics for spatial prediction and mapping of soil properties.

Key points

- Geostatistics is a technology for predicting values of variables at places from more or less sparse sample data.
- It is based on the theory of regionalized variables, that is, of spatially correlated random variables.
- Crucial to any prediction is a model of the covariance structure of the random process, usually expressed in the variogram.
- The prediction itself, kriging, is an unbiased and minimum-variance weighted moving average of data with the weights determined from the variogram.
- Ordinary kriging, the basic technique, can be elaborated to take into account fixed effects such as those of geographic trend and subsidiary covariates and of coregionalized random variables.
- Geostatistics is now widely applied for mapping plant nutrients and pollutants in soil and to salinity and its remediation, and it is being combined with remotely sensed data and machine learning for spatial prediction.

Introduction

Geostatistics has its origins in the gold mines of South Africa. The main problem from a statistical point of view was to estimate from small core samples the concentrations of gold in rocks and the amounts of metal in blocks of rock still in place ahead of the mining fronts. Only if those blocks contained sufficient gold to justify extraction would they be mined. Engineer D.G. Krige realized that such estimation was a form of spatial prediction and that by taking into account the spatial correlation among cores he could predict concentrations more precisely than by the conventional methods of estimation current then, the 1950s, and without the bias that typical mine sampling could incur (Krige, 1951). Krige's procedures were put on a sound theoretical footing by G. Matheron in the 1960s (Matheron, 1965), and for the next two decades applications in mining and petroleum engineering burgeoned, aided by the authoritative text by Journel and Huijbregts (1978). The method of prediction itself came to be known as 'kriging'. During that time soil scientists became similarly concerned; they were disappointed that the predictions from soil classification failed to take into account the similarities over short distances on the ground within those classes. They would need to express those similarities in terms of spatial correlation and apply the same techniques as the miners were using for prediction at unvisited places. They also found that in many instances classification conferred no benefits. Their main aim became to interpolate and map soil properties from measurements at more or less sparse sampling points. Burgess and Webster (1980) reported the first accounts of their success, since when applications in soil science have also burgeoned.

Random variables and random functions

Geostatistics is a technology. It is based on models of spatial variation, and to understand how it works one must first understand what is meant by 'model' in this context. Soil conditions are determined by agents of the environment acting on rocks and sediments, but such is the resulting complexity that they are indistinguishable from random. Fig. 1 is an example showing how the clay content of the soil varies along a transect across Jurassic scarplands of south central England. Table 1 summarizes the data. Geostatistically-minded soil scientists, pedometricians, therefore treat soil properties as random; they have in their minds a model in which the actual value $z(\mathbf{x})$ at any point $\mathbf{x}$ on the land surface is one drawn at random from a variable $Z(\mathbf{x})$, a population with potentially an infinity of values there. In principle a point may be in one, two or three dimensions, but in almost all applications each position is in two, so that $\mathbf{x} \equiv \{x_1, x_2\}$. Further, pedometricians assume that the values of z at all such points, again potentially infinite in number, in a region are drawn from a random process $Z(\mathbf{x})$ in those same dimensions. The model is then represented as

$$Z(\mathbf{x}) = u(\mathbf{x}) + \varepsilon(\mathbf{x}). \quad (1)$$

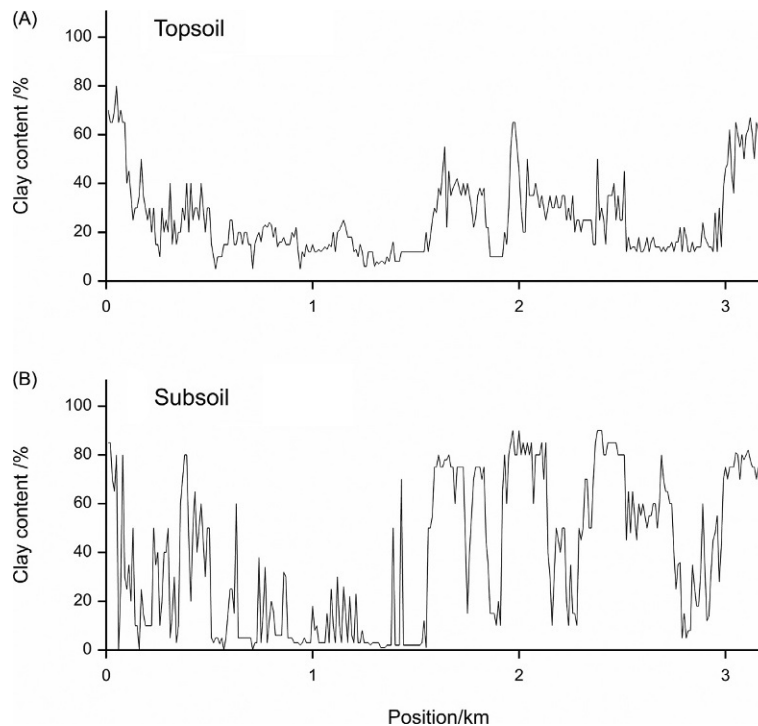


Fig. 1 Observed percentages of clay in the soil along a transect across Jurassic scarplands in central England; (A) topsoil and (B) subsoil.

Table 1 Summary of 321 observations of clay content (recorded as % by weight) in topsoil and subsoil as percentages.

	Topsoil	Subsoil
Mean	25.6	39.0
Median	20.0	36.0
Variance	255.49	936.81
Standard deviation	16.0	30.6
Skewness	1.2	0.2
Correlation coefficient	0.59	

Here $u(\mathbf{x})$ is the mean of the process at $\mathbf{x}$; it is a constant there but may vary from place to place. It is often termed 'trend' or 'drift'. The quantity $\varepsilon(\mathbf{x})$, in contrast, is a random variable with mean zero and variance σ^2 . Thus the model is a mixed one of fixed and random effects. The model is more often than not simplified to one in which $u(\mathbf{x})$ is treated as constant, the mean μ throughout the region of interest, and the model becomes

$$Z(\mathbf{x}) = \mu + \varepsilon(\mathbf{x}). \quad (2)$$

The random variable $Z(\mathbf{x})$ is then said to be stationary in the mean.

In design-based estimation, which dominated survey statistics from the 1920s and in which the sampling points are chosen independently at random, the $\varepsilon(\mathbf{x})$ are independent of one another. Their covariances are therefore zero:

$$\text{Cov} = [\varepsilon(\mathbf{x}_i), \varepsilon(\mathbf{x}_j)] = E[\varepsilon(\mathbf{x}_i), \varepsilon(\mathbf{x}_j)] = 0 \quad \text{for all } \mathbf{x}_i \neq \mathbf{x}_j, \quad (3)$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are any two places on the ground.

Random sampling in mines is not feasible, nor would it be efficient. In soil survey it is more efficient and easier to sample on grids. Grid sampling provides even coverage and limits prediction error variances. In these circumstances the $\varepsilon(\mathbf{x})$ s can no longer be treated as independent of one another; they in turn must be modelled, and in practice they are modelled as functions of their separation on the ground:

$$\text{cov} [\varepsilon(\mathbf{x}_i), \varepsilon(\mathbf{x}_j)] = c(\mathbf{x}_i - \mathbf{x}_j) = C(\mathbf{h}). \quad (4)$$

With constant mean Eq. (3) is equivalent to

$$\begin{aligned} \text{cov} [Z(\mathbf{x}_i), Z(\mathbf{x}_j)] &= E[Z\{(\mathbf{x}_i) - \mu\} \{Z(\mathbf{x}_j) - \mu\}] \\ &= C(\mathbf{h}). \end{aligned} \quad (5)$$

In two dimensions the separation $\mathbf{h}$, called the 'lag,' is a vector comprising both distance and direction. The covariance decreases towards zero as the separating distance, $|\mathbf{h}|$, increases. At the other extreme, $|\mathbf{h}| = 0$, the covariance is $\sigma^2 = C(0)$, the variance of $\varepsilon(\mathbf{x})$. Division of $C(\mathbf{h})$ by $C(0)$ gives the dimensionless spatial correlation:

$$\rho(\mathbf{h}) = C(\mathbf{h})/C(0). \quad (6)$$

If $C(\mathbf{h})$ and therefore $\rho(\mathbf{h})$ are greater than zero then the random process, $Z(\mathbf{x})$, is said to be spatially correlated, or 'autocorrelated' because z is same soil property at all places $\mathbf{x}$.

In many instances practitioners cannot assume that the mean is constant. In those circumstances they cannot assume that the covariance exists. While they may be unwilling to assume that the means are constant throughout their regions of interest they might assume the mean to be constant for small to moderate $|\mathbf{h}|$, so that differences in $Z(\mathbf{x})$ between places separated by $\mathbf{h}$ would be zero:

$$E[Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})] = 0. \quad (7)$$

This observation provoked Matheron to replace the covariances for expressions of spatial correlation by variances of the differences in Eq. (7) thus:

$$\begin{aligned} \text{var} [Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})] &= E[\{Z(\mathbf{x}) - Z(\mathbf{x} + \mathbf{h})\}^2] \\ &= 2\gamma(\mathbf{h}). \end{aligned} \quad (8)$$

Eqs. (7), (8) constitute Matheron's intrinsic hypothesis, and random variables with these properties are said to be intrinsically stationary. In combination they release practitioners from the constraints of second-order stationarity and provide a wider field of application. The quantity $\gamma(\mathbf{h})$ is called the 'semivariance' at lag $\mathbf{h}$; it is half of the variance of the difference in $Z(\mathbf{x})$ between two places. It is nevertheless the variance per point when points are considered in pairs. As a function of $\mathbf{h}$ it is the semivariogram, now usually termed simply the variogram.

For second-order stationary processes the variogram and the covariance function of $\mathbf{h}$ are equivalent:

$$\gamma(\mathbf{h}) = C(0) - C(\mathbf{h}), \quad (9)$$

and

$$\gamma(\mathbf{h}) = \sigma^2 \{1 - \rho(\mathbf{h})\}. \quad (10)$$

Estimating the variogram

Semivariances can be, and usually are, estimated from data, $z(\mathbf{x}_1), z(\mathbf{x}_2), \dots$, by the method of moments:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})} \sum_{i=1}^{m(\mathbf{h})} \{z(\mathbf{x}_i) - z(\mathbf{x}_i + \mathbf{h})\}^2, \quad (11)$$

in which $m(\mathbf{h})$ is the number of paired comparisons at lag $\mathbf{h}$. By changing $\mathbf{h}$ one obtains an ordered set, an experimental variogram, which can be displayed as a graph of $\hat{\gamma}$ against $\mathbf{h}$. Fig. 2 shows examples in which the experimental semivariances for the data in Fig. 1 are plotted as points against the lag distance, $h = |\mathbf{h}|$, for the one-dimensional transect. Where sampling is at regular intervals along a transect then it is natural to increment the lag in steps of one sampling interval. For a grid one may compute $\hat{\gamma}$ itself on a grid. Where the sampling points are irregularly scattered the actual separations must be placed into ‘bins’ with limits in separating distance and also in direction if there is more than one dimension—see Fig. 3. Gridded data may be treated similarly.

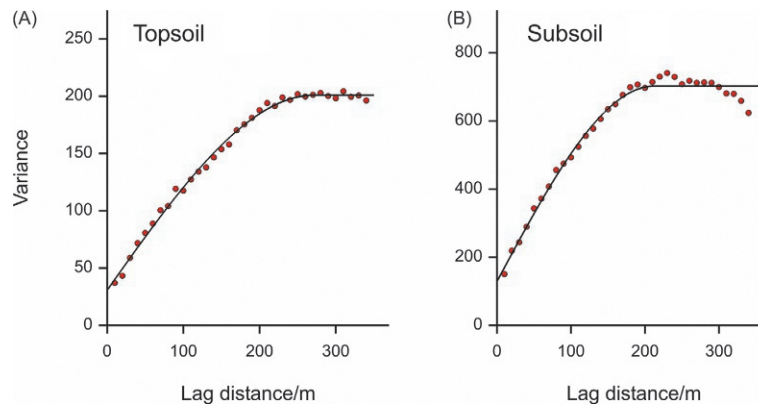


Fig. 2 Experimental variograms of clay content of soil estimated by the method of moments and spherical models fitted to them; (A) topsoil and (B) subsoil.

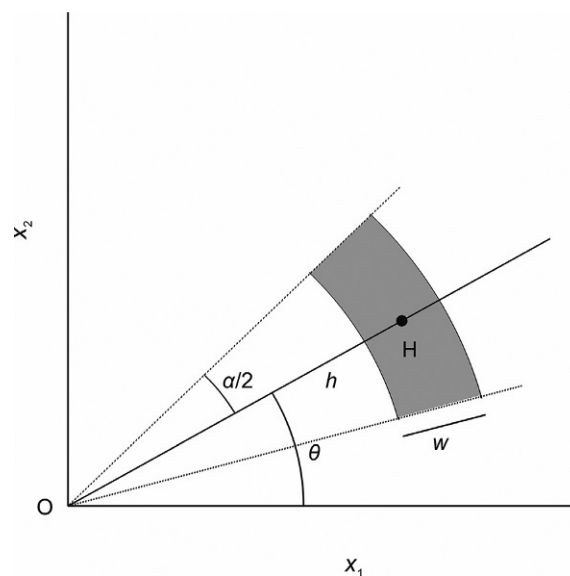


Fig. 3 Discretization of the lag in two dimensions for irregularly scattered data. All separations within the grey sector are assigned to lag distance h and direction θ .

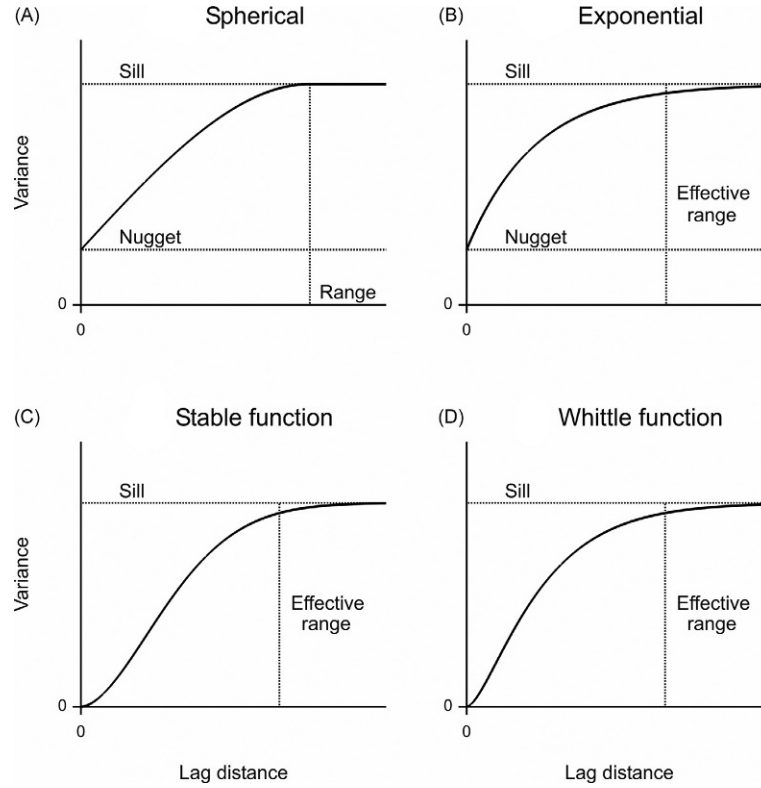


Fig. 4 Four kinds of bounded variogram: (A) spherical, (B) exponential, (C) stable function with $a = 1.75$, (D) Whittle function.

Robust estimators

Outliers and other exceptional values in the tails of distributions can carry undue weight in the estimation of semivariances. To counteract such effects robust estimators may be used. Three enjoy some popularity; they go by the names of their inventors: Cressie and Hawkins (1980), Dowd (1984), and Genton (1998). Webster and Oliver (2007) give the definitions and compare their merits; they found none to be best in all circumstances.

Models for variograms

Values of $\mathbf{h}$ define discrete points on the experimental variogram. Estimates computed by Eq. (11) at those points are more or less in error, arising from the sampling, and the sequences tend to fluctuate from point to point erratically also. The underlying variogram, Eq. (8), is a continuous function in as many dimensions as the variable $Z(\mathbf{x})$. Obtaining such a function from an experimental variogram involves a further kind of modelling; it requires one to fit a plausible function to the experimental values. The usual approach is to fit the simplest model that makes sense.

Fig. 4A shows the principal features of many, if not most, experimental variograms. They are as follows.

1. The variance increases from near the ordinate with increasing lag distance.
2. It reaches a maximum at which it remains thereafter.
3. Any simple smooth line or surface placed through the points and projected to the ordinate cuts it at some value greater than 0.

The model must also be mathematically acceptable in that it cannot give rise to 'negative variances' when random variables are combined. Let $z(\mathbf{x}_i)$, $i = 1, 2, \dots, n$, be a realization of the random variable $Z(\mathbf{x})$ with covariance function $C(\mathbf{h})$ and variogram $\gamma(\mathbf{h})$, and consider the linear sum

$$y = \sum_{i=1}^n \lambda_i z(\mathbf{x}_i), \quad (12)$$

where the λ_i are any arbitrary weights. The variable Y from which y derives is also a random variable, and its variance is given by

$$\text{var}[Y] = \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j C(\mathbf{x}_i - \mathbf{x}_j), \quad (13)$$

where $C(\mathbf{x}_i - \mathbf{x}_j)$ is the covariance of Z between $\mathbf{x}_i$ and $\mathbf{x}_j$. This variance must be positive or zero; it may not be negative. If $Z(\mathbf{x})$ is intrinsic only then the covariances do not exist, and one must re-write Eq. (13) as

$$\text{var}[Y] = C(0) \sum_{i=1}^n \lambda_i \sum_{j=1}^n \lambda_j - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j), \quad (14)$$

where $\gamma(\mathbf{x}_i - \mathbf{x}_j)$ is the semivariance of Z between $\mathbf{x}_i$ and $\mathbf{x}_j$. The first term on the right-hand side of this equation is eliminated if the weights sum to 0, and one is left with

$$\text{var}[Y] = - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{x}_i - \mathbf{x}_j). \quad (15)$$

This too must guarantee non-negative variances, and only functions that do that are admissible. They are said to be conditional negative semi-definite, CNSD, the condition being that the weights in Eq. (15) sum to zero.

There are only a few families of simple functions that satisfy the above criteria. They can be divided into those that are bounded and those that are not. In the first group are the popular spherical and exponential models. Their formulae in their isotropic forms, i.e. for $h = |\mathbf{h}|$, are as follows.

Spherical

The spherical function has the following equation:

$$\begin{aligned} \gamma(h) &= c_0 + c_1 \left\{ \frac{3h}{2r} - \frac{1}{2} \left(\frac{h}{r} \right)^3 \right\} \quad \text{for } 0 < h \leq r \\ &= c_0 + c_1 \quad \text{for } h > r \\ &= 0 \quad \text{for } h = 0. \end{aligned} \quad (16)$$

Here $\gamma(h)$ is the semivariance at lag h , and c_1 is the a priori variance of the autocorrelated process. The quantity c_0 is the intercept on the ordinate, and is known as the *nugget variance*, a term deriving from gold mining. The combined $c_0 + c_1$ is known as the *sill* of the model, and c_1 is the sill of the correlated variance. These quantities are illustrated in Fig. 4A, and Fig. 2B shows the function fitted to the experimental variogram of subsoil clay content, Fig. 1B. The values of the parameters c_0 , c_1 , and r are listed in Table 3.

The function has a finite distance parameter, r ; this is its *range*, also known as its correlation range. It marks the limit of spatial dependence; values at places closer to one another than it are more or less correlated, whereas those further apart are not. It implies that all the variance in $\mathcal{R}$ is encountered within that distance, and in this sense it corresponds to the concept of the representative elementary volume (REV).

The function gets its name from the formula for the volume of two intersecting spheres, which are of diameter r .

The semivariance at lag 0 is itself zero, and for continuous processes such as most physical and chemical properties of the soil $\gamma(h)$ should increase gradually as h increases from 0. In practice, we usually lack sufficient estimates of $\gamma(h)$ near the ordinate to fit a model through the origin, and we therefore take the conservative approach above. The nugget variance is therefore best regarded as embodying variation within the shortest sampling interval plus any measurement error.

Other functions with the same general form and finite ranges are the bounded linear function (valid in one dimension only), the circular (valid in one and two dimensions but not in three), and the pentaspherical.

Exponential

The equation for the exponential function is

$$\begin{aligned} \gamma(h) &= c_0 + c_1 \left\{ 1 - \exp\left(-\frac{h}{a}\right) \right\}, \text{ for } h > 0 \\ &= 0 \text{ for } h = 0 \end{aligned} \quad (17)$$

in which c_0 and c_1 have the same meanings as before, but now with a distance parameter, a . The exponential model approaches its sill asymptotically and has no definite range therefore. A working range is often taken as $a' = 3a$ at which point the function has reached 95% of c_1 . This model is illustrated in Fig. 4B.

Models with reverse curvature near the origin

Some variograms approach the origin with decreasing gradients. They arise from smoother variation over short distances than do the spherical and exponential functions. They may be described by the general equation:

$$\gamma(h) = c \left\{ 1 - \exp\left(-\frac{h^2}{a^2}\right) \right\}, \quad (18)$$

in which $0 < \alpha \leq 2$. If $\alpha = 2$ Eq. (18) is that of the Gaussian function. This is at the limit of acceptability and gives rise to unstable prediction—see Wackernagel (2003) for an example and discussion. It is best replaced by stable models with $\alpha < 2$. Fig. 4C with $\alpha = 1.75$ is an example.

The Gaussian function is one of another family of models, namely the Matérn functions, which have become increasingly popular recently. The equation is

$$\gamma(h) = c \left\{ 1 - \frac{1}{2^{v-1} \Gamma(v)} \left(-\frac{h}{a} \right)^v K_v \left(\frac{h}{a} \right) \right\}. \quad (19)$$

Here c is the sill variance and a is a distance parameter, as before. The equation embodies the gamma function, Γ , with parameter v and the Bessel function of the second kind, K_v , for parameter v . Parameter v describes the smoothness of variation and can vary from 0 (very rough) to infinity (very smooth). It includes as special cases the exponential function when $v = 0.5$, Whittle's elementary correlation when $v = 1$ (below) and the Gaussian function when $v = \infty$.

The attraction of Whittle's function is that it derives theoretically from diffusion in two dimensions. Its equation is

$$\gamma(h) = c \left\{ 1 - \frac{h}{a} K_1 \left(\frac{h}{a} \right) \right\}, \quad (20)$$

in which a is again a distance parameter, and K_1 is the modified Bessel function of the second kind—see Fig. 4D. To all can be added a nugget variance, c_0 , if desired.

Unbounded models

Variograms of processes that are intrinsic but not second-order stationary increase without bound as the lag distance increases. These can usually be fitted by power functions, for which the general equation including a nugget is

$$\gamma(h) = c_0 + qh^b. \quad (21)$$

The parameter q describes the intensity of the process, and the exponent, which must lie strictly between 0 and 2 (these limits are excluded), describes the curvature. If $b < 1$ then the curve is convex upwards, if it is 1 then the increase in variance with distance is strictly linear, and if $b > 1$ then the curve is concave upwards. The curve with $b = 2$ is a parabola and describes a smoothly continuous process that is not random. Fig. 5 illustrates the curves for several values of b .

Anisotropy

The variogram of a two-dimensional process is itself two-dimensional, and if the process is anisotropic then so is its variogram, which is then a function of both distance, h and direction, θ . In the simplest cases the anisotropy is geometric, meaning that it can be made isotropic by a linear transformation of the coordinates. The transformation is defined by reference to an ellipse:

$$\Omega(\theta) = \sqrt{A^2 \cos^2(\theta - \varphi) + B^2 \sin^2(\theta - \varphi)}, \quad (22)$$

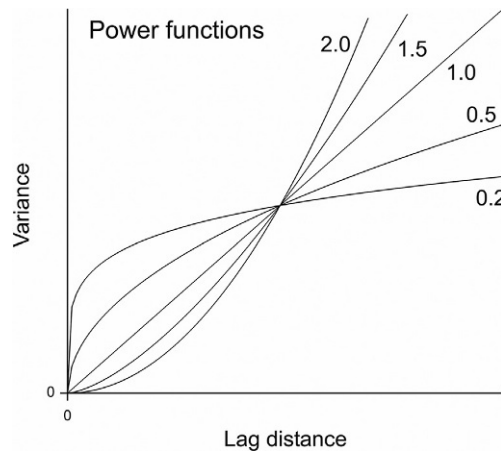


Fig. 5 Four valid power functions, Eq. (21) with exponents, b , 0.2, 0.5, 1.0 and 1.5, plus the inadmissible limiting function with $b = 2$.

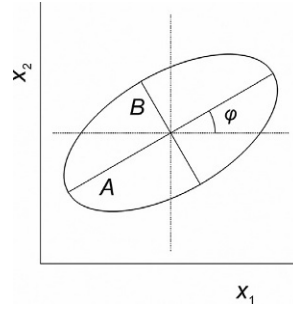


Fig. 6 Ellipse showing parameters of geometric anisotropy.

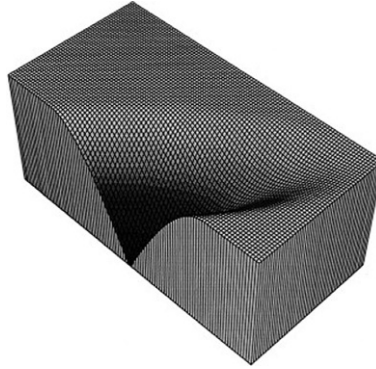


Fig. 7 Perspective diagram of the variogram surface of an anisotropic spherical function.

where A and B are the long and short diameters, respectively, of the ellipse, and φ is its orientation, i.e. the direction of the long axis; see Fig. 6. Eq. (22) is incorporated into the models as follows. For the bounded models Ω replaces the distance parameter of the isotropic variogram. So, for example, in the exponential

$$\gamma(h, \theta) = c_0 + c \left\{ 1 - \exp \left(-\frac{h}{\Omega(\theta)} \right) \right\}, \quad (23)$$

and in the linear function:

$$\gamma(h, \theta) = c_0 + \Omega(\theta)h^\alpha, \quad (24)$$

in which $\alpha = 1$. Fig. 7 shows an anisotropic spherical function. Notice how the range of the model changes with changing direction.

Fitting the models

As above, experimental variograms computed by the method of moments fluctuate more or less erratically. Any of the models represented by Eqs. (16)–(24) must be fitted to them. All have non-linear parameters, and these must be found by numerical approximation, preferably accompanied by visual inspection to ensure that the models are plausible. The individual semivariances may be weighted in proportion to the numbers of paired comparisons from which they are computed or some other function of those numbers. Finally, any chosen model may be checked by cross-validation. In this technique, each datum in turn is predicted from other data in its neighborhood by ordinary punctual kriging (see below), and the squared difference between it and its prediction is compared with its kriging variance, $\sigma_K^2(\mathbf{x}_0)$. Ideally the two should be equal. Thus, if the datum is denoted $z(\mathbf{x}_0)$ and its prediction $\hat{Z}(\mathbf{x}_0)$ then the ratio $\left\{ \hat{Z}(\mathbf{x}_0) - z(\mathbf{x}_0) \right\}^2 / \sigma_K^2(\mathbf{x}_0)$ should equal 1.0. The most suitable model might be the one for which the mean of this ratio is closest to 1.0. This ratio is variously known as the mean squared deviation ratio and the mean standardized deviation. If the variable z has a normal (Gaussian) distribution then this ratio has a skewed distribution for which the median is 0.455, and one may judge the goodness of a model for its closeness to this value.

The above procedures based on the method-of-moments or robust alternatives and subsequent non-linear least squares fitting are not the only ones for modelling variograms. Likelihood methods and specifically residual maximum likelihood (REML) are feasible for moderately large sets of data. They come into their own where there is geographic trend or correlation with other variables, and their role is dealt with in the next section.

Combining trend and random fluctuation

The above functions describe processes that are entirely random though correlated. In accord with Matheron's intrinsic hypothesis they can be represented by the model:

$$Z(\mathbf{x}) = \mu_V + \varepsilon(\mathbf{x}), \quad (25)$$

in which μ_V is the mean, i.e. constant, in some neighborhood V , and $\varepsilon(\mathbf{x})$ is the autocorrelated random variable as defined in Eq. (8). It often happens that such models are unacceptable, either because there is evident long-range trend across a region or because over short distances the variation appears smooth. In these circumstances μ_V cannot be treated as constant but must be replaced by a deterministic term, such as $u(\mathbf{x})$ in Eq. (1) and which depends on the position $\mathbf{x}$. Recall that the model is

$$Z(\mathbf{x}) = u(\mathbf{x}) + \varepsilon(\mathbf{x}).$$

If $u(\mathbf{x})$ can describe the variation over the whole of $\mathcal{R}$ simply then it tends to be called trend. If it is local only then it is more often called 'drift'. These terms are not strictly defined and are often used interchangeably. In either event $u(\mathbf{x})$ is usually represented by a low-order polynomial, so that Eq. (1) becomes

$$Z(\mathbf{x}) = \sum_{j=0}^J \beta_j f_j(\mathbf{x}) + \varepsilon(\mathbf{x}), \quad (26)$$

in which the β_j are unknown coefficients and the $f_j(\mathbf{x})$ are known functions of our choosing.

It is fairly easy, even if somewhat arbitrary, to separate any long-range trend from the short-range apparently random fluctuation, and to estimate the parameters of the two components separately. This is the basis of the technique known as regression kriging. Its shortcoming is that variograms, which are of the residuals from the estimated trends, are biased, and kriging variances (see below) are underestimated. It is not at all easy to do it where there is short-range drift. Early attempts to deal with these circumstances involved a structural analysis, which was effectively a process of trial-and-error. More recently they have been replaced by likelihood methods and specifically residual maximum likelihood (REML) in which the parameters of the variogram and the coefficients of the trend are estimated simultaneously.

Residual maximum likelihood estimation

In the simplest case of a linear trend in two dimensions Eq. (26) expands to

$$Z(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon(\mathbf{x}). \quad (27)$$

More generally and in matrix form the trend model is

$$Z(\mathbf{x}) = \mathbf{w}\boldsymbol{\beta} + \varepsilon(\mathbf{x}), \quad (28)$$

in which the vector $\mathbf{w}$ with $J + 1$ columns contains the $J + 1$ elements $1, x_1, x_2, \dots, x_J$ of the trend, and the vector $\boldsymbol{\beta}$ contains the coefficients. This is the linear mixed model. Data from a survey can be represented similarly:

$$\mathbf{z}(\mathbf{X}_d) = \mathbf{W}_d \boldsymbol{\beta} + \varepsilon(\mathbf{X}_d), \quad (29)$$

in which $\mathbf{z}(\mathbf{X}_d)$ with n rows is the vector of observed values at the corresponding set of points denoted by $\mathbf{X}_d$. Matrix $\mathbf{W}_d$ with n rows and $J + 1$ columns replaces the vector $\mathbf{w}$ of Eq. (28) and is called the 'design matrix'. The residuals $\varepsilon(\mathbf{X}_d)$ are assumed to be second-order stationary random variables jointly normally distributed with zero means and $n \times n$ covariance matrix $\mathbf{C}_d$. Being second-order stationary they have their equivalents in the variogram:

$$C(\mathbf{h}) = C(\mathbf{0}) - \gamma(\mathbf{h}).$$

Parameters for a plausible model are found by maximization of the likelihood of the residuals given the data, $L[\boldsymbol{\theta} | \mathbf{z}(\mathbf{X}_d)]$, in which the vector $\boldsymbol{\theta}$ is short-hand for the parameters of the variogram. For the spherical model, for example, $\boldsymbol{\theta} \equiv \{c_0, c_1, r\}$. The values of these parameters that maximize the likelihood are found numerically, and they are the ones that define the variogram of the residuals. Lark et al. (2006) and Webster and Oliver (2007) give the derivation of the equations in full. From those values one computes the covariance matrix $\mathbf{C}_d$ for the data points, and from it estimates the coefficients in $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{W}_d^T \hat{\mathbf{C}}_d^{-1} \mathbf{W}_d \right)^{-1} \mathbf{W}_d^T \hat{\mathbf{C}}_d^{-1} \mathbf{z}(\mathbf{X}_d) \quad (30)$$

Classification as external drift

For many years pedologists dealt with spatial variation by classification and drew boundaries between the classes when mapping them to produce what are technically called choropleth maps. In landscapes where the soil changes from one type to another fairly

Table 2 Means of subsoil clay percentage in the classes of soil over nine geological strata as estimated by ordinary least squares and REML analysis.

<i>Raw data</i>		
Mean		39.1
Median		36.0
Variance		936.81
<i>Class</i>	<i>Mean</i>	
	<i>OLS estimate</i>	<i>REML estimate</i>
Sharp's Hill Beds (clay)	61.3	75.4
Great Oolite	33.4	42.3
Lower Estuarine Beds (silt)	12.9	15.1
Chipping Norton Limestone (limestone)	9.5	10.8
Chipping Norton Limestone (sand)	15.1	25.2
Upper Lias (clay)	69.9	50.3
Pleistocene and Recent (silt and clay)	55.8	51.0
Middle Lias (ironstone)	41.4	43.4
Middle and Lower Lias (clay)	68.6	67.5

abruptly, as where it overlies boundaries between rock types, the practice has merit. Between the boundaries, however, the soil tends to vary in a more continuous spatially correlated apparently random way. The latter accords with the basic geostatistical model; classifications in contrast are fixed effects created by the pedologists. One might treat the two sources of variation separately by sampling to a random design to estimate class means, and characterize the variation within classes by variograms of the residuals from the class means. As above, that would be inefficient for mapping variation within the classes. Further, a pedologist might not know how to classify the soil sensibly until after examining survey data. The pedologist is thus typically faced with estimating class means without bias from non-random sampling and variograms of the residuals from those means. The geostatistical solution to this problem is to treat a classification as an external drift and to combine it with the apparently random variation within the classes in a single analysis of a mixed model by REML.

The data on the clay content of the subsoil along the Sandford transect displayed in Fig. 1B illustrate the solution. The transect lies across nine Jurassic geological strata in Mid-Oxfordshire, England. The figure shows some fairly sharp changes in clay content, and these correspond to boundaries between the outcrops of the rock strata. Table 2 summarizes the data overall and for the classes individually. For illustrative purposes the means calculated by ordinary least squares (OLS) are listed along with the REML estimates. Fig. 8 repeats the variogram of the raw data, shown in Fig. 2B, as the uppermost sequence of points and fitted spherical model. The lowest sequence of points in the figure is the experimental variogram of the OLS residuals with its fitted curve, also spherical. The curve in the middle is the variogram of the REML residuals from the class means as is that of the exponential model. Table 3 lists the fitting parameters.

The example shows how soil classification can affect the outcome of a geostatistical analysis and the importance of mixed modelling where data are from systematic samples. Ordinary least squares estimates of class means are evidently biased, and the variances of the residuals underestimate the true variances. The variograms over short lag distances are remarkably similar over short lag distances; as the distance increases so the variogram of the residuals from the OLS estimates departs from the best variogram based on the mixed model. Cressie (1993) forecasts exactly this behavior.

Coregionalization—simultaneous variation in two or more variables

Any two variables, say z_u and z_v , may be correlated and in particular linearly correlated. That relation is conventionally expressed by the product-moment correlation coefficient:

$$\rho = \frac{\text{cov}[uv]}{\sqrt{\text{var}[u] \times \text{var}[v]}}, \quad (31)$$

i.e. the covariance of z_u and z_v divided by the product of their standard deviations.

The two spatial random variables, $Z_u(\mathbf{x})$ and $Z_v(\mathbf{x})$, may also be *spatially* intercorrelated in that each is spatially correlated both with itself, i.e. autocorrelated, and with the other. The two variables are then said to be cross correlated. In these circumstances the two variables have autovariograms, one each, as defined by Eq. (8) and for present purposes denoted $\gamma_{uu}(\mathbf{h})$ and $\gamma_{vv}(\mathbf{h})$. They also have a cross-variogram, $\gamma_{uv}(\mathbf{h})$, defined by

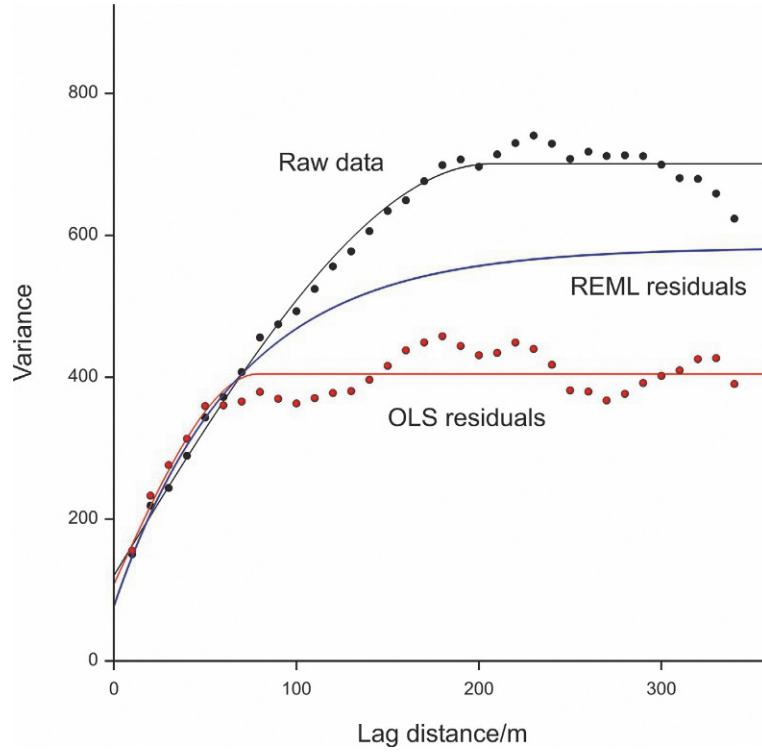


Fig. 8 Variograms of clay content of the subsoil along the transect shown in Fig. 1B; top of raw data, bottom of OLS residuals, and middle REML estimate from mixed model. Parameter estimates are listed in Table 3.

Table 3 Variogram models and their parameters of Eqs. (16) and (17) of subsoil clay on the Sandford transect, Fig. 1B.

Source	Spherical			Exponential		
	c_0	c_1	r/m	c_0	c_1	a/m
Raw data	120.6	580.3	207.0			
OLS residuals	108.4	296.4	79.2			
REML residuals				77.5	505.3	67.3

$$\gamma_{uv}(\mathbf{h}) = \frac{1}{2} E[\{Z_u(\mathbf{x}) - Z_u(\mathbf{x} + \mathbf{h})\} \{Z_v(\mathbf{x}) - Z_v(\mathbf{x} + \mathbf{h})\}]. \quad (32)$$

If both variables are second-order stationary with means μ_u and μ_v then both will have covariance functions, C_{uu} and C_{vv} , as defined in Eq. (5), and a cross-covariance:

$$C_{uv}(\mathbf{h}) = E[\{Z_u(\mathbf{x}) - \mu_u\} \{Z_v(\mathbf{x} + \mathbf{h}) - \mu_v\}]. \quad (33)$$

There is also a cross-correlation coefficient, ρ_{uv} , given by

$$\rho_{uv}(\mathbf{h}) = \frac{C_{uv}}{\sqrt{C_{uu}(\mathbf{0})C_{vv}(\mathbf{0})}}. \quad (34)$$

This is effectively the extension of the Pearson product-moment correlation coefficient of Eq. (31) into the spatial domain, and when $\mathbf{h} = \mathbf{0}$ it is the Pearson coefficient.

The cross-covariance is in general not symmetric, i.e.

$$E[\{Z_u(\mathbf{x}) - \mu_u\} \{Z_v(\mathbf{x} + \mathbf{h}) - \mu_v\}] \neq E[\{Z_v(\mathbf{x}) - \mu_v\} \{Z_u(\mathbf{x} + \mathbf{h}) - \mu_u\}]. \quad (35)$$

In words, the cross-covariance between $Z_u(\mathbf{x})$ and $Z_v(\mathbf{x})$ in one direction is different from that in the other, or expressed another way

$$C_{uv}(\mathbf{h}) \neq C_{uv}(-\mathbf{h}) \quad \text{or equivalently} \quad C_{uv}(\mathbf{h}) \neq C_{uv}(\mathbf{h}), \quad (36)$$

since

$$C_{uv}(\mathbf{h}) = C_{vu}(-\mathbf{h}).$$

One can envisage asymmetry between two soil properties at different depths on a slope as result of creep or solifluction. The subsoil will tend to lag behind the topsoil. Similarly, irrigation by flooding always from the same end of a field might distribute salts differentially in the direction of flow. Asymmetric covariances have not been reported in the soil literature as far as I know, however.

The cross-variogram and the cross-covariance function (if it exists) are related by

$$\gamma_{uv}(\mathbf{h}) = C_{uv}(\mathbf{0}) - \frac{1}{2}\{C_{uv}(-\mathbf{h}) + C_{uv}(\mathbf{h})\}, \quad (37)$$

This quantity contains both $C_{uv}(\mathbf{h})$ and $C_{uv}(-\mathbf{h})$ and in consequence loses any information on asymmetry; it is an even function, i.e. symmetric:

$$\gamma_{uv}(\mathbf{h}) = \gamma_{vu}(\mathbf{h}) \quad \text{for all } \mathbf{h}.$$

One can estimate cross-semivariances in a way similar to that of the auto-semivariances:

$$\hat{\gamma}_{uv}(\mathbf{h}) = \frac{1}{2m(\mathbf{h})} \sum_{i=1}^{m(\mathbf{h})} \{z_u(\mathbf{x}_i) - z_u(\mathbf{x} + \mathbf{h})\} \{z_v(\mathbf{x}_i) - z_v(\mathbf{x} + \mathbf{h})\}, \quad (38)$$

and form the sample cross-variogram by simply incrementing $\mathbf{h}$. There is an equivalent formula for computing the cross-covariances. Notice that there must be numerous places where both z_u and z_v have been measured.

Modelling the coregionalization

The cross-variogram can be modelled in the same way as the auto-variogram, and the same restricted set of functions is available. There is one additional constraint. Any linear combination of the variables is itself a regionalized variable, and its variance cannot be negative. This is assured by adoption of the *linear model of coregionalization*. In it the variable, $Z_u(\mathbf{x})$, is assumed to be the sum of independent (orthogonal) random variables, $Y_j^k(\mathbf{x})$:

$$Z(\mathbf{x}) = \sum_{k=1}^K \sum_{j=1}^k a_{uj}^k Y_j^k(\mathbf{x}) + \mu_u, \quad (39)$$

in which the superscript k is an index, not a power. There is a similar assumption for $Z_v(\mathbf{x})$. If the assumptions hold then the pair of variables has a cross-variogram

$$\gamma_{uv}(\mathbf{h}) = \sum_{k=1}^K \sum_{j=1}^k a_{uj}^k a_{vj}^k g^k(\mathbf{h}). \quad (40)$$

The products in the second summation can be replaced by b_{uv}^k to give

$$\gamma_{uv}(\mathbf{h}) = \sum_{k=1}^K b_{uv}^k g^k(\mathbf{h}). \quad (41)$$

The quantities are the variances and covariances, e.g. nugget and sill variances, for the independent components of a spherical model. For two variables there are the three nugget variances, b_{uu}^1 , b_{vv}^1 and b_{uv}^1 and similarly three for the sills of the correlated variances. The coefficients $b_{uv}^k = b_{vu}^k$ for all k , and for each k the matrix of coefficients

$$\begin{bmatrix} b_{uu}^k & b_{uv}^k \\ b_{vu}^k & b_{vv}^k \end{bmatrix}$$

must be positive definite. Since the matrix is symmetric it is sufficient that $b_{uu}^k > 0$ and $b_{vv}^k \geq 0$ and that its determinant is positive or zero:

$$|b_{uv}^k| = |b_{vu}^k| \leq \sqrt{b_{uu}^k b_{vv}^k}. \quad (42)$$

This is Schwarz's inequality.

Any number of regionalized variables may be embodied in the linear model of coregionalization. If there are V of them then the full matrix of coefficients, $[b_{ij}]$, will be of order V , and its determinant and all its principal minors must be positive or zero.

Schwarz's inequality has the following consequences.

1. Every basic structure, $g^k(\mathbf{h})$, present in the cross-variogram must also appear in the two auto-variograms, i.e. $b_{uu}^k \neq 0$ and $b_{vv}^k \neq 0$ if $b_{uv}^k \neq 0$. As a corollary, if a basic structure $g^k(\mathbf{h})$ is absent from either auto-variogram then it may not be included in the cross-variogram.
2. Structures may be present in the auto-variograms without their appearing in the cross-variogram, i.e. b_{uv}^k may be zero when $b_{uu}^k > 0$ and $b_{vv}^k > 0$.

Parameters of the linear model of the coregionalization with the above constraints can be fitted by iteration. The distance parameters are usually first approximated by the fitting of models independently to the experimental variograms, and good compromise values are chosen from these. Then with those values fixed the values of the b_{uv}^k are found to minimize the sums of the squares of the residuals, subject to the condition that the solution guarantees non-negative variances, i.e. is CNSD. One may check the validity of the resulting model by plotting it on a graph of the experimental cross-semivariances plus the limiting values that would hold if the correlation between the variables were perfect. These limits constitute the hull of perfect correlation, which is obtained from the coefficients b_{uu}^k and b_{vv}^k by

$$\text{hull} [\gamma(\mathbf{h})] = \pm \sum_{k=1}^K \sqrt{b_{uu}^k b_{vv}^k} g^k(\mathbf{h}).$$

The line should fit close to the experimental values for the model. It must also fall within the hull to be acceptable. If it lies close to the hull then the cross-correlation is strong; if in contrast it is far from the bounds then the cross-correlation is weak. [Wackernagel \(2003\)](#) describes this aspect of coregionalization well.

Example

The clay content of the subsoil was recorded at the same sampling points as those for topsoil in [Fig. 1B](#), and in combination with the topsoil data illustrate the coregionalization. [Table 1](#) contains a summary, and it includes the Pearson correlation coefficient, which is 0.59, indicating a modest correlation overall. The experimental auto-variograms for the two depths together with their cross-variogram are shown in [Fig. 9](#) by the black discs. Spherical models can be fitted to all of them with ranges varying between 191 and 265 m, with an average of 228 m. If this value is fixed for the coregionalization model then the nugget and sill component components of the model are obtained as described above and appear in [Table 4](#). They are somewhat different from those fitted independently, but only somewhat. The final result is shown in [Fig. 9](#) by the solid lines through the plotted points. The model evidently fits well.

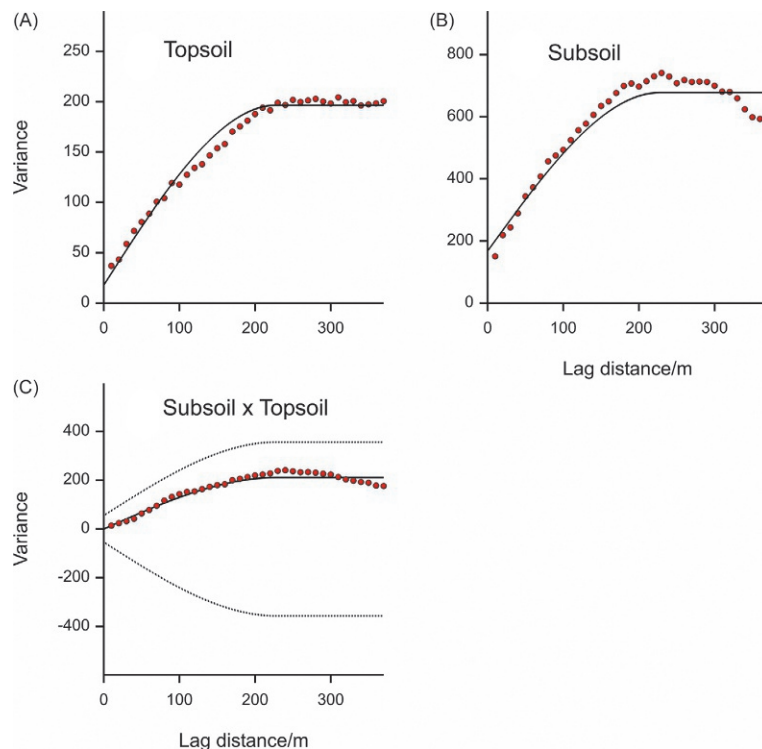


Fig. 9 Auto- and cross-variograms of clay content in topsoil and subsoil with the linear model of coregionalization fitted. The dashed lines depict the hull of perfect correlation.

Table 4 Fitted nugget and sill variances of the correlated structure, i.e. the coefficients b_{uv}^k of Eq. (41), of the spherical model of coregionalization of clay in topsoil and subsoil with a range of 228 m.

		Topsoil	Subsoil
Nugget variance	Topsoil	17.841	
	Subsoil	−0.799	167.986
Sill variance	Topsoil	178.659	
	Subsoil	211.436	509.723

The dashed lines on the graph for Subsoil $\times$ Topsoil define the hull of perfect correlation. The cross-variogram falls within the hull, as it should, and the moderate distance it keeps from the upper bound is a measure of moderate cross-correlation over the lag distances computed.

Spatial prediction—kriging

Ordinary kriging

The variogram and covariance functions are not only elegant mathematical descriptions of the real world of the soil, they are crucial for local estimation, or spatial prediction as it is better called, by kriging.

Kriging is a general term for processes of weighted averaging of data to provide unbiased local estimates of unknown values of a variable with minimum variance. In the simplest case for a place $\mathbf{x}_0$ where the mean is unknown (the usual situation) the estimate is formed as

$$\hat{Z}(\mathbf{x}_0) = \sum_{i=1}^N \lambda_i z(\mathbf{x}_i), \quad (43)$$

where the $z(\mathbf{x}_i)$, $i = 1, 2, \dots, N$ are sample data at places $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$, and the λ_i are weights. The weights sum to 1 to assure unbiasedness, i.e. $\sum_{i=1}^N \lambda_i = 1$, and the variance of the estimate is given by

$$\begin{aligned} \sigma^2 Z(\mathbf{x}_0) &= E \left[\left\{ \hat{Z}(\mathbf{x}_0) - z(\mathbf{x}_0) \right\}^2 \right] \\ &= 2 \sum_{i=1}^N \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_0) - \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j). \end{aligned} \quad (44)$$

Here $\gamma(\mathbf{x}_i, \mathbf{x}_j)$ is the semivariance between the data points $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\gamma(\mathbf{x}_i, \mathbf{x}_0)$ is the semivariance between the data point $\mathbf{x}_i$ and the target point $\mathbf{x}_0$.

This variance is minimized by solution of the $N + 1$ equations:

$$\begin{aligned} \sum_{i=1}^N \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_j) + \psi(\mathbf{x}_0) &= \gamma(\mathbf{x}_i, \mathbf{x}_j) \quad \forall j \\ \sum_{i=1}^N \lambda_i &= 1, \end{aligned} \quad (45)$$

in which the $\psi(\mathbf{x}_0)$ is a Lagrange multiplier introduced for the minimization. The solution yields the optimum weights, and these are inserted into Eq. (43) to give the required estimate at $\mathbf{x}_0$.

The minimized variance is also obtained from the solution as

$$\sigma_{OK}^2(\mathbf{x}_0) = \sum_{i=1}^N \lambda_i \gamma(\mathbf{x}_i, \mathbf{x}_0) + \psi(\mathbf{x}_0). \quad (46)$$

This particular form of the technique is ordinary punctual kriging—ordinary because it is the most used, and punctual because the estimates are for points of the same size and shape, i.e. the same *supports*, as the bodies of soil or other material on which the measurements were made. It is readily generalized for estimating blocks, B , larger than the supports of the data. Eq. (43) holds for the averaging, though with B replacing $\mathbf{x}_0$. In Eq. (45) the individual semivariances on the right-hand sides are replaced by the averages of the semivariances between the data points and the target block, B , $\bar{\gamma}(\mathbf{x}_i, B)$. Finally, the kriging variance is given by

$$\sigma_{OK}^2(B) = \sum_{i=1}^N \lambda_i \bar{\gamma}(\mathbf{x}_i, B) + \psi(B) - \bar{\gamma}(B, B), \quad (47)$$

where $\bar{\gamma}(B, B)$ is the average semivariance within B , i.e. the within-block variance.

When the kriging equations are solved it usually turns out that only the few points nearest to the target point or block carry any appreciable weight; the weights of the others are so close to 0 that they can be ignored. Kriging is thus a local weighted average. It has two other intuitively attractive features.

1. Where data points are clustered the weight of the cluster is divided among its members so that the individual weights are small compared with those of isolated points.
2. Data points lying approximately in a line between the target and more distant points screen the latter, which tend as a result to have virtually no weight, however close they are to the target.

This also has practical implications. The kriging systems, Eq. (45), need never be large. They are swiftly solved, and instabilities with matrix inversion are rare.

It is now evident why the variogram, or the equivalent covariance function, is so important; the kriging systems need values drawn from it. As above, these must not give rise to negative kriging variances, and so a valid function must be fitted to the experimental variogram.

Ordinary kriging will serve in some 90% of cases; it is the 'work horse' of practical geostatistics. It requires the fewest assumptions and the least knowledge. It has been much used to map soil properties, including concentrations of plant nutrients, salinity, trace element contents, pollutants and nematode infestation. The kriging systems are solved at close intervals on a grid, from which isarithms, 'contours,' of the estimates can be drawn by other graphics programs. This has led to its application in land reclamation and precision agriculture. The kriging variances can be mapped similarly; patches of large variance coincide with sparse sampling, and so the maps can show where denser sampling is necessary or desirable to achieve more reliable estimates.

Cokriging

Ordinary cokriging is an extension of ordinary autokriging, in which the semivariances of both the target variable and those of the subsidiary variable(s), the u and v , are drawn from the model of coregionalization, Eq. (41). Where a target variable is sampled less intensely than the subsidiary variable(s) cokriging adds spatial information and diminishes the kriging prediction variances. That is its principal merit. Where the target variable and subsidiary variable(s) are sampled at the same places, i.e. full sampling, cokriging gives the same or almost the same results as ordinary kriging of the target variable alone. Its merit then is coherence when two or more target variables are predicted from the same systems. One of the best examples of this application was for the prediction of the depths of the Upper Chalk beneath the English Channel before the rail tunnels were bored in the 1990s (Chilès and Delfiner, 2012).

Other forms of kriging

Although ordinary kriging is by far the most popular application of geostatistics in soil science there are situations in which it falls short, either by failing to take into account what is already known or because it needs taking beyond mere predictions of measured variables. There are numerous elaborations of the technique; two groups are mentioned below.

Universal kriging and kriging with external drift

Universal kriging takes into account known or estimated trend in the target variable. The underlying model is that of Eq. (26), usually with a simple low-order polynomial for the trend as in Eq. (26). The variogram of the residuals from the trend, estimated best nowadays by REML as described above, is combined with the function for trend to predict the target variable. The fixed effects of subsidiary variables, external drifts, are taken into account similarly by the combination of them and the variograms of the residuals from those drifts.

Indicator and disjunctive kriging

Authorities responsible for controlling infestation by pests and pollution, for restoring contaminated land and restricting its use set limits on the relevant variables. They then need to estimate the probabilities that these threshold values are exceeded (or not). One technique for this is indicator kriging in which the scale of a variable is divided into two or more classes each of which is modelled and kriged. Goovaerts (1997) describes the procedure fully. A more elaborate and sounder technique is disjunctive kriging. It transforms data of almost any frequency distribution to normal (Gaussian) via Hermite polynomials. The probabilities of a variable's exceeding specified thresholds are then estimated from the normal distribution—see Webster and Oliver (2007) for details.

Applications and new developments

In the two decades 1980–2000 pedometricians mastered geostatistics and adapted it to soil survey. Since then software for its implementation, especially that embodied in geographic information systems, has enabled pedologists, agronomists, environmental scientists and land managers to apply the techniques in a wide variety of situations in many parts of the world. Most have involved mapping, which is covered in the chapter on pedometrics. Insert the words 'soil' and 'kriging' into the Web of Science and

you will find more than 5000 published journal articles. Add the word ‘mapping’ and that will still produce more than 2700 articles. There are undoubtedly thousands more cases, published and unpublished, for specific projects. The attractions of unbiased predictions with minimum and known variances are compelling.

Many recent articles report straightforward applications for mapping plant nutrients, as that by Tamburi et al. (2021), for salinity control (Shaddad et al., 2020; Li et al., 2015) and organic carbon (Boubehziz et al., 2020). Many other applications concern pollutants; for example, both Kebonye et al. (2021) and Jia et al. (2021) analyzed and mapped arsenic in polluted soil. Remotely sensed data are increasingly being combined with measurements on the soil to enhance prediction either by cokriging (e.g. Shahriari et al., 2019) or with external drift.

Machine learning is also being combined with kriging (see the separate chapter on machine learning). Jia et al. (2021), for example, gathered images of the soil surface by a drone and used machine learning and kriging for their mapping. They also used simulation, another widespread geostatistical technique, to assess the uncertainties of their predictions. Szabo et al. (2019) used random forests and geostatistics to map some of the soil’s hydraulic properties, and Orton et al. (2020) combined machine learning with geostatistics to model and map soil properties in three dimensions. Fentanes et al. (2020) ventured further by combining kriging with data from a robotic cosmic-ray sensor.

References

- Boubehziz S, Khanchoul K, Benslama M, Marchetti A, Francaviglia R, and Piccini C (2020) Predictive mapping of soil organic carbon in Northeast Algeria. *Catena* 190: 104539.
- Burgess TM and Webster R (1980) Optimal interpolation and isarithmic mapping of soil properties. I. The semi-variogram and punctual kriging. *Journal of Soil Science* 31: 315–331.
- Chilès J-P and Delfiner P (2012) *Geostatistics: Modeling Spatial Uncertainty*, 2nd edn. Hoboken, New Jersey: John Wiley & Sons.
- Cressie NAC (1993) *Statistics for Spatial Data*, revised edn New York: John Wiley & Sons.
- Cressie N and Hawkins DM (1980) Robust estimation of the variogram. *Journal of the International Association Mathematical Geology* 12: 115–125.
- Dowd PA (1984) The variogram and kriging: Robust and resistant estimators. In: Verly G, David M, Journel AG, and Marechal A (eds.) *Geostatistics for Natural Resources Characterization*, pp. 91–106. Dordrecht: D. Reidel.
- Fentanes JP, Badiée A, Duckett T, Evans J, Pearson S, and Cielniak G (2020) Kriging-based robotic exploration for soil moisture mapping using a cosmic-ray sensor. *Journal of Field Robotics* 37: 122–136.
- Genton MG (1998) Highly robust variogram estimation. *Mathematical Geology* 30: 213–221.
- Goovaerts P (1997) *Geostatistics for Natural Resources Evaluation*. New York: Oxford University Press.
- Jia X, Cao Y, O’Connor D, Zhu J, Tsang DCW, Zou B, and Hou DY (2021) Mapping soil pollution by using drone image recognition and machine learning at an arsenic-contaminated agricultural field. *Environmental Pollution* 270: 116281.
- Journel AG and Huijbregts CJ (1978) *Mining Geostatistics*. London: Academic Press.
- Kebonye NM, Kingley J, Chakraborty S, Agyeman PC, Ahado SK, Eze PN, Nemecek K, Drabek O, and Borůvka L (2021) Comparison of multivariate methods for arsenic estimation and mapping in floodplain soil via portable X-ray fluorescence spectroscopy. *Geoderma* 384: 114792.
- Krige DG (1951) A statistical approach to some basic mine problems on the Witwatersrand. *Journal of the Chemical and Metallurgical Society of South Africa* 52: 119–139.
- Lark RM, Cullis BR, and Welham SJ (2006) On the prediction of soil properties in the presence of spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. *European Journal of Soil Science* 57: 787–799.
- Li HY, Webster R, and Shi Z (2015) Mapping soil salinity in the Yangtze delta: REML and universal kriging (E-BLUP) revisited. *Geoderma* 237–238: 71–77.
- Matheron G (1965) *Les variables régionalisées et leur estimation*. Paris: Masson.
- Orton TG, Pringle MJ, Bishop TFA, Menzies NW, and Dang YP (2020) Increment-averaged kriging for 3-D modelling and mapping soil properties: Combining machine learning and geostatistical methods. *Geoderma* 361: 114094.
- Shaddad SM, Buttafuoco G, and Castrignano A (2020) Assessment and mapping of soil salinization risk in an Egyptian field using a probabilistic approach. *Agronomy-Basel* 10: . article number 85.
- Shahriari M, Delbari M, Afrasiab P, and Pahlavan-Rad MR (2019) Predicting regional spatial distribution of soil texture in floodplains using remote sensing data: A case of southeastern Iran. *Catena* 182: 104149.
- Szabo B, Szatmári G, Laborezi A, Makó A, Rajkai K, and Pásztor L (2019) Mapping soil hydraulic properties using random-forest-based pedotransfer functions and geostatistics. *Hydrology and Earth Systems Sciences* 23: 2615–2635.
- Tamburi V, Shetty A, and Shrihari S (2021) Spatial variability of Vertisols nutrients of the Deccan Plateau region of north Karnataka, India. *Environment Development and Sustainability* 23: 2910–2923.
- Wackernagel H (2003) *Multivariate Geostatistics*, 3rd edn Berlin: Springer-Verlag.
- Webster R and Oliver MA (2007) *Geostatistics for Environmental Scientists*, 2nd edn. Chichester: John Wiley & Sons.

Uncertainty assessment of spatial soil information

Gerard BM Heuvelink^a and Richard Webster^b, ^aISRIC – World Soil Information and Wageningen University, Wageningen, The Netherlands;

^bRothamsted Research, Harpenden, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction	671
Defining and quantifying error and uncertainty in soil data	672
Single soil measurements	672
Statistical model of measurement error	672
Estimating the measurement error standard deviation from replicates	674
Proximal soil sensing and model errors	674
Spatial soil information	675
Statistical model of errors in spatial soil variables	675
Geostatistics and machine learning	675
Multivariate distribution of soil map errors	676
Cross-correlation, positional uncertainty and uncertainty of categorical soil variables	677
Positional uncertainty	677
Errors and uncertainty in soil class maps	677
Change of support	679
Spatial support and resolution	679
Uncertainty propagation	680
Uncertainty propagation example	680
The added value of uncertainty information	680
Conclusion	682
References	682

Abstract

Information on the soil of any region is never free of error because it is derived from imperfect measurements, imperfect models and imperfect mapping algorithms. Scientists and technologists who produce the information and those who use it cannot be certain about the true values of the soil properties and soil classes in a region, and they must acknowledge the fact. That uncertainty can best be characterized by probability distributions. Geostatistics, based on such distributions, has become fashionable in the last 30 years or so for quantifying uncertainty in spatial information, and rightly so as it quantifies the uncertainty in its predictions. Uncertainty in soil information also propagates through agronomic and environmental models. It must be communicated to end users because the validity of their decisions depends on it.

Key points

- All information about the soil embodies error.
- Acknowledgement of error translates into users being uncertain about the true values of soil properties and classes.
- Uncertainty can best be quantified by probability distributions.
- Geostatistics provides the technology to quantify uncertainty in spatial soil information.
- The propagation of uncertainty in soil information through agronomic and environmental models can be traced with Monte Carlo simulation.
- Communication of uncertainty to end users and how uncertainty can be incorporated in decision making needs more attention than it has received.

Introduction

Uncertainty is present in our daily lives. It affects our decisions on what to do. The weather forecast might tell us that there is a 60% chance that it will rain: we take umbrellas. If it says that the chance of rain is only 10% we might decide to leave our umbrellas at home and risk getting wet. More seriously, farmers want to know the likelihood of disease in their crops and the deficiencies in plant nutrients in the soil. These are matters that affect profit and loss of farm business. Agencies responsible for public health and environmental protection need to weigh the risk of doing nothing in the face of uncertain threats against the cost of acting unnecessarily to counter them when the threats are almost non-existent. There are many examples of decision making problems

involving uncertain soil information. They include the remediation of polluted soil, the prevention of soil erosion, and the mitigation of pesticide leaching. They are practical matters, not purely academic exercises in statistics.

All measurements of soil properties (and other environmental variables) contain error in the sense that they depart from the true values. That error arises from imperfections in the analytical instruments, from the people who use them and from errors that occur during the processing of the recorded data to make them suitable for storage in information databases. Short-range spatial variation is another source of error, given that soil samples are never returned to where they were taken and sampling locations have positional error. Soil taken from location s and analyzed in the laboratory might differ substantially from the soil at location $s + h$, even if $|h|$ is as small as a few decimeters. Composite soil sampling can diminish these differences, but some error inevitably persists because even such a composite is still only a sample of all the soil at that site. All this means that we can never be sure about the true state of the soil: we, the producers and users of soil information, are to some extent uncertain.

Uncertainty tends to increase when measurements of basic soil properties are used to obtain derived ones via pedotransfer functions or mechanistic models of dynamic soil processes, for example. Interpolation from measurements to create maps of soil properties adds to the errors of measurement and so too increases uncertainties. We must conclude that considerable uncertainty is often associated with the information that is stored in soil databases and presented in various forms, including maps. This does not mean that the information is of no value; uncertainty is not the same as ignorance. In many cases we do know a great deal about the soil, but we must also acknowledge that the information is not perfect.

Some numerical expression of the uncertainty is important because it is needed to determine whether the information is sufficiently accurate for the purpose that a user has in mind. Soil data of too poor a quality might lead to flawed decisions with serious undesirable consequences, both economic and environmental. For instance, the European legislation on the use of pesticides in agriculture depends crucially on the leaching potential of these substances to the ground- and surface-water, which in turn depends importantly on soil properties. In these circumstances users should be aware of the quality of the soil information so that they can be sure that it is sufficiently reliable for their purposes. Ideally they should account for the uncertainty of the information when making their decisions.

This chapter (i) provides a statistical definition of uncertainty in soil information; (ii) extends this definition to uncertainty in spatial soil information; (iii) reviews methods that are used to quantify uncertainty in soil information, while paying attention to different sources of uncertainty; (iv) shows how uncertainty in soil information propagates through subsequent analyses; and (v) explains how uncertainty information can be used in decision making. It focuses on the quantification of uncertainty of soil properties that are measured and recorded on continuous scales: properties such as pH, particle-size distribution, and soil organic matter content. The chapter also addresses uncertainty of categorical variables, such as soil type and diagnostic properties recorded as present or absent, i.e. binary variables. It begins with defining uncertainty in a single soil measurement.

Defining and quantifying error and uncertainty in soil data

Single soil measurements

The concept of error and uncertainty is perhaps best introduced with an example. Consider a soil sample that is analyzed for clay content in the laboratory. Suppose the outcome of the analysis is 38.4%. Because of imperfections in technique, equipment or operating conditions, however, the measurement is subject to error, and the true value almost certainly differs more or less from the measured value. Perhaps it is 36.1%, or perhaps 40.0%: we cannot ascribe a single value to the true clay content. We are in effect uncertain about the true value. But we are not completely ignorant: we have a good idea of the clay content. Given the measurement of 38.4% we should consider it highly unlikely that the true clay content is greater than 60% or smaller than 20%. With this in mind we should be able to list a collection of possible values of the true clay content, and to associate a probability with each of them. In other words, we should be able to characterize the true clay content or the measurement error of it with a probability distribution.

For a continuous variable such as clay content one can usually characterize the measurement error by a normal (Gaussian) distribution (Fig. 1). It is a remarkable fact that measurement errors in many fields of endeavor have this distribution, first recognized by Gauss early in the 19th century, though the distribution itself had been formulated earlier by de Moivre and Laplace. It seems to arise because measurement error comprises numerous apparently independent random effects, which when summed give rise to the normal distribution, a result known as the Central Limit Theorem. In the measurement of clay content, for example, many disturbances contribute to the error: drying, grinding, sieving, pipetting or laser diffraction, instrumental fluctuation and so on. Together these lead to an approximate normal distribution, which one can verify by plotting histograms or plotting quantiles of the observed values against those of a normal distribution (Q–Q plots), based on repeated measurements of the clay content of the same soil sample, such as is done in proficiency testing. There are also many variables for which the measurement errors are better represented by lognormal distributions; for these it seems that individual sources of error are multiplied rather than summed. Again, one can see this by plotting histograms and Q–Q plots of repeated measurements.

Statistical model of measurement error

Representing an uncertain soil property by a probability distribution effectively defines a statistical model. This model states

$$Z_T = z_M + \varepsilon, \quad (1)$$

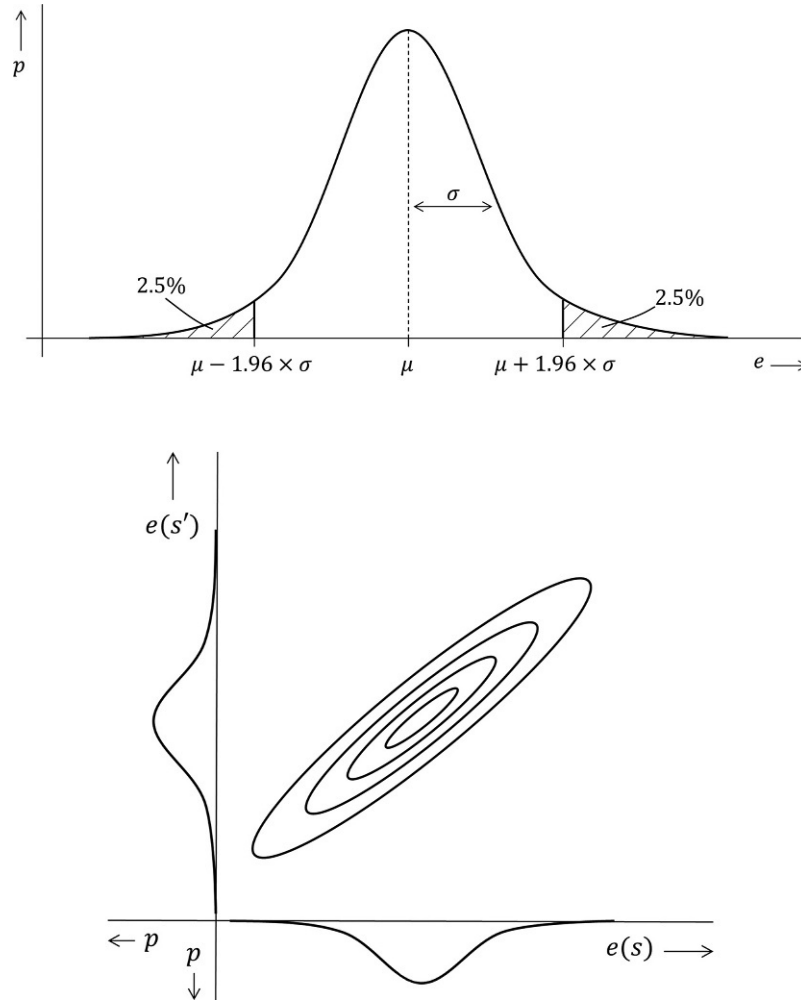


Fig. 1 Probability density of the normal distribution representing measurement or prediction error. Top: univariate case, with μ the mean (systematic error) and σ the standard deviation (measure of random error). Lower and upper limits of a 95% confidence or prediction interval are also indicated. Bottom: contour lines of a bivariate probability distribution of errors at locations $\mathbf{s}$ and $\mathbf{s}'$. Marginal densities projected on the axes. Note that the rotated ellipses indicate a positive correlation between the errors at the two locations.

where Z_T is the true soil property value, z_M a measurement of it, and ε the measurement error. Note that Z_T is written in upper case (capital letter) and z_M in lower case. This is to distinguish the stochastic Z_T (a random variable with a probability distribution) from the deterministic quantity z_M (the measured value). The measurement error ε is also stochastic, and here assumed normally distributed. Its probability density p is given by

$$p(e) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{(e - \mu)^2}{2\sigma^2} \right\}. \quad (2)$$

The normal distribution is fully characterized by two parameters: its mean μ , which in this case signifies the systematic measurement error, and its standard deviation σ , which signifies random measurement error (Fig. 1). The square of the standard deviation, i.e. σ^2 , is the variance.

A well managed laboratory aims to avoid bias, and instruments are calibrated and laboratory staff trained to achieve that. A consequence is that the mean of ε may be assumed to be zero, and its standard deviation may be estimated from replicated measurements on the same soil sample. If the same soil sample is analyzed by several laboratories then ε may be decomposed into multiple components, such as the batch, laboratory and residual error. For instance, using data from the Wageningen Evaluating Programmes for Analytical Laboratories (WEPAL, 2022), Van Leeuwen et al. (2021) estimated the standard deviation of the measurement error of total organic carbon as 0.64%, with between-laboratory variation contributing more than variation within laboratories.

We do not know what the true value in Eq. (1) is. All we can say is that it is one of an infinite number of values represented by a probability distribution; we treat it as a random variable and denote it as Z_T . It is not that we believe that the world is stochastic

(Webster, 2000). In other words, we propose a statistical model of a reality that in itself is deterministic, and this model provides us with an estimate of our uncertainty about the true value. A narrow distribution (i.e. small standard deviation) conveys the notion that we are fairly certain about the true value, whereas a wide distribution (large standard deviation) means that our uncertainty is large. In the pathological case of an infinitely wide distribution (i.e. one that has infinite variance), we should be ignorant.

Estimating the measurement error standard deviation from replicates

A complete description of a measurement error requires specification of its probability distribution, but if it is assumed that the error is normally distributed about zero, i.e. has no systematic error, then all that is needed to characterize the error is its standard deviation σ . As mentioned above, σ can be estimated from replicated measurements. Fig. 2 shows a histogram of 58 measurements of total organic carbon (TOC) of the same soil sample (WEPAL, 2022). The mean is 35.58 g kg^{-1} and the standard deviation is 6.60 g kg^{-1} . Thus, if we assume that this standard deviation is representative of the random laboratory error of measuring TOC, and if for another soil sample we had only one measurement, say 50.92 g kg^{-1} , then our estimate of the TOC of that soil sample would be its measured value, with an uncertainty expressed by a standard deviation of 6.60 g kg^{-1} . If we further assume that the measurement error is normally distributed then we can state with 95% confidence that the true TOC value is in between $50.92 - 1.96 \times 6.60 = 37.98 \text{ g kg}^{-1}$ and $50.92 + 1.96 \times 6.60 = 63.86 \text{ g kg}^{-1}$. Here, the value 1.96 is taken from the standard normal distribution, and we assume that the standard deviation of the measurement error is constant. In practice we often find that it is proportional to the measured value, meaning that σ is larger in case of a larger measured value. A decision between constant and proportional errors may be based on repeated measurements of multiple soil samples and plots of the standard deviation for each sample against its average measured value.

We can diminish uncertainty about the true TOC value of the soil sample by taking the average of replicated measurements. For instance, in this specific case where we had 58 independent measurements of the same soil sample, we are 95% confident that the true TOC lies between $35.58 - 1.96 \times 6.60/\sqrt{58} = 28.88 \text{ g kg}^{-1}$ and $35.58 + 1.96 \times 6.60/\sqrt{58} = 42.27 \text{ g kg}^{-1}$. This is because from sampling theory we know that the average of n independent measurements has standard deviation $\sigma/\sqrt{n}$; this is the standard error of the mean. So, to double accuracy (effectively to halve the uncertainty) we need to quadruple the number of measurements. In the example above with a standard deviation of 6.60 g kg^{-1} , the standard error would be 2.09, 1.04 and 0.52 g kg^{-1} for samples of size 10, 40 and 160, respectively. This 'square root law' relation between the standard deviation and standard error might seem somewhat daunting to investigators when they plan sampling and measurement campaigns, but that is how it is.

Proximal soil sensing and model errors

In recent years the use of proximal soil sensing (Viscarra Rossel et al., 2010) has grown rapidly because it is much cheaper to measure soil properties this way than by wet chemistry. However, the techniques come at a price: these measurements tend to be more uncertain. This is because the statistical methods used to convert spectral signals to values of soil properties introduce additional uncertainty. Perhaps the most often used conversion method is partial least squares regression. The residual variances of this method should be added to the variances of the wet chemistry measurements to quantify the uncertainty of the proximal soil sensing estimates (Takoutsing et al., 2022). This additional error might be small for some soil properties and sensors, but it can be

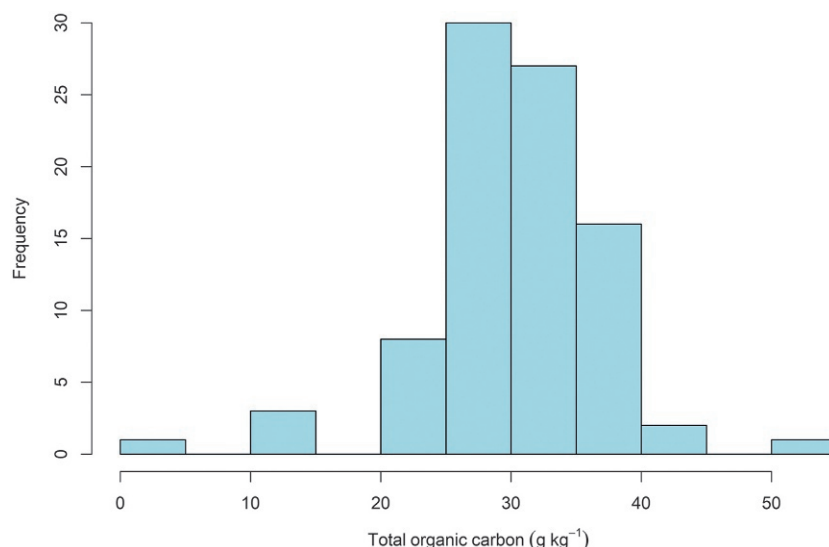


Fig. 2 Histogram of 58 replicated measurements of total soil organic carbon (g kg^{-1}) of the same soil sample. Source: WEPAL (2022). International Soil-Analytical Exchange Programme. <https://www.wepal.nl/en/wepal/Home/Proficiency-tests/Soil/Proficiency-tests/ISE.htm>.

much bigger for others. This additional uncertainty may be termed ‘model uncertainty’, though it results from both an imperfect model connecting the signals from the sensors to the soil and imperfections in the sensors themselves. Next to measurement uncertainty, model uncertainty is an important source of uncertainty in soil data. It also occurs when properties that are hard or time-consuming to measure are predicted from properties that are straightforward and relatively cheap to measure with pedo-transfer functions (Stenemo and Jarvis, 2007; Deng et al., 2009) or when soil quality indicators are derived from basic soil properties and classes (Nosrati, 2013). Mechanistic soil process models also have model uncertainty. One might quantify model uncertainty in particular circumstances, but this might lack generality. For instance, one would expect a model of soil erosion constructed for the USA and calibrated with data from the USA to serve well in the USA. But it might perform less well, even badly, in the different circumstances of, say, China.

It is beyond the scope of this chapter to provide a full treatment of model uncertainty. Interested readers are referred to Brown and Heuvelink (2005).

Spatial soil information

Farming technology has advanced apace in recent years. Machines can now be programed to vary the amounts of fertilizer in response to known or estimated local concentrations of nutrients, provided their locations are known. This is an important development in precision agriculture (Oliver, 2010). In the same vein, an environmental protection agency might want to remove soil polluted by a toxic metal if it knows where the soil’s load of that metal exceeds some statutory limit, or alternatively forbid cultivation or grazing there. In principle, both farmers and protection agencies want maps of soil conditions. Neither farmers nor protection agencies can measure the soil everywhere; they can at best sample on small supports, typically cores a few cm in diameter or small areas within which cores of soil are bulked. Yet they wish to know the concentrations everywhere, so that they can treat the soil in accord with those concentrations. They therefore turn to mapping, and for this they may use various approaches, as outlined in other chapters of the Encyclopedia. The focus of this and subsequent sections is on quantifying the uncertainties that are introduced by mapping. But before we address this we must first extend the statistical model of uncertainty of a single soil measurement to a model for uncertain spatially distributed soil properties.

Statistical model of errors in spatial soil variables

The extension of the statistical model presented in the previous sections to spatially distributed soil properties starts by our recognizing that Eq. (1) may be copied to all locations $\mathbf{s}$ in a region D :

$$Z_T(\mathbf{s}) = z_p(\mathbf{s}) + \varepsilon(\mathbf{s}) \quad \text{for all } \mathbf{s} \in D. \quad (3)$$

Here the measurement z_M in Eq. (1) has now been replaced by a soil prediction z_p to acknowledge that in this case we compare a predicted or mapped soil property to a true value. The mapped property is typically not directly measured but obtained by prediction, be it human (i.e. what was formerly convention) or statistical (i.e. digitally). This implies that spatial interpolation error will add to the uncertainty. Thus, ε is no longer solely a measurement error but rather a spatial prediction error and is likely to have a larger variance than the measurement error of Eq. (1).

Geostatistics and machine learning

Eq. (3) shows that quantification of the map error requires the specification of the probability density of the prediction error for all locations in the region. If we invoke the assumption of a normal distribution again then we shall need the mean $\mu(\mathbf{s})$ and standard deviation $\sigma(\mathbf{s})$ for all $\mathbf{s} \in D$. This might seem a formidable task, but it turns out to be a routine by-product of kriging, i.e. of mapping geostatistically. It is explained in some detail in the geostatistics chapter of the Encyclopedia and fully in text books (e.g. Webster and Oliver, 2007; Chilès and Delfiner, 2012). To krig a soil property one needs a geostatistical model, for which certain assumptions are made; these usually comprise second-order stationarity and isotropy. If the assumptions are satisfied then kriging produces unbiased predictions while minimizing $\sigma(\mathbf{s})$, the kriging standard deviation.

There are various forms of kriging that depend on the specific assumptions about the variables being predicted and mapped. They include lognormal kriging, which is used to predict variables that appear to have lognormal distributions. Indicator kriging was developed to predict and map variables that have distributions not readily parameterized; the procedure is somewhat cumbersome, and the prediction errors are no longer normally distributed (Webster and Oliver, 2007). These disadvantages are overcome by the technically more advanced disjunctive kriging. Ordinary and simple kriging are based on models in which there is no spatial trend. Trend, a fixed effect, can be accommodated in universal kriging, regression kriging and kriging with external drift, where the latter two take into account fixed effects of environmental covariates (Hengl et al., 2004). Nevertheless, whichever kriging method is used the uncertainty of the predictions is quantified. Typically, the prediction errors are small in densely sampled regions and large in sparsely sampled ones and close to the borders of the regions being mapped. In other words, the sparser is the sampling the greater is the uncertainty. The magnitude of uncertainty also depends on the spatial correlation of the target variable, as quantified by the variogram. The geostatistics chapter provides details.

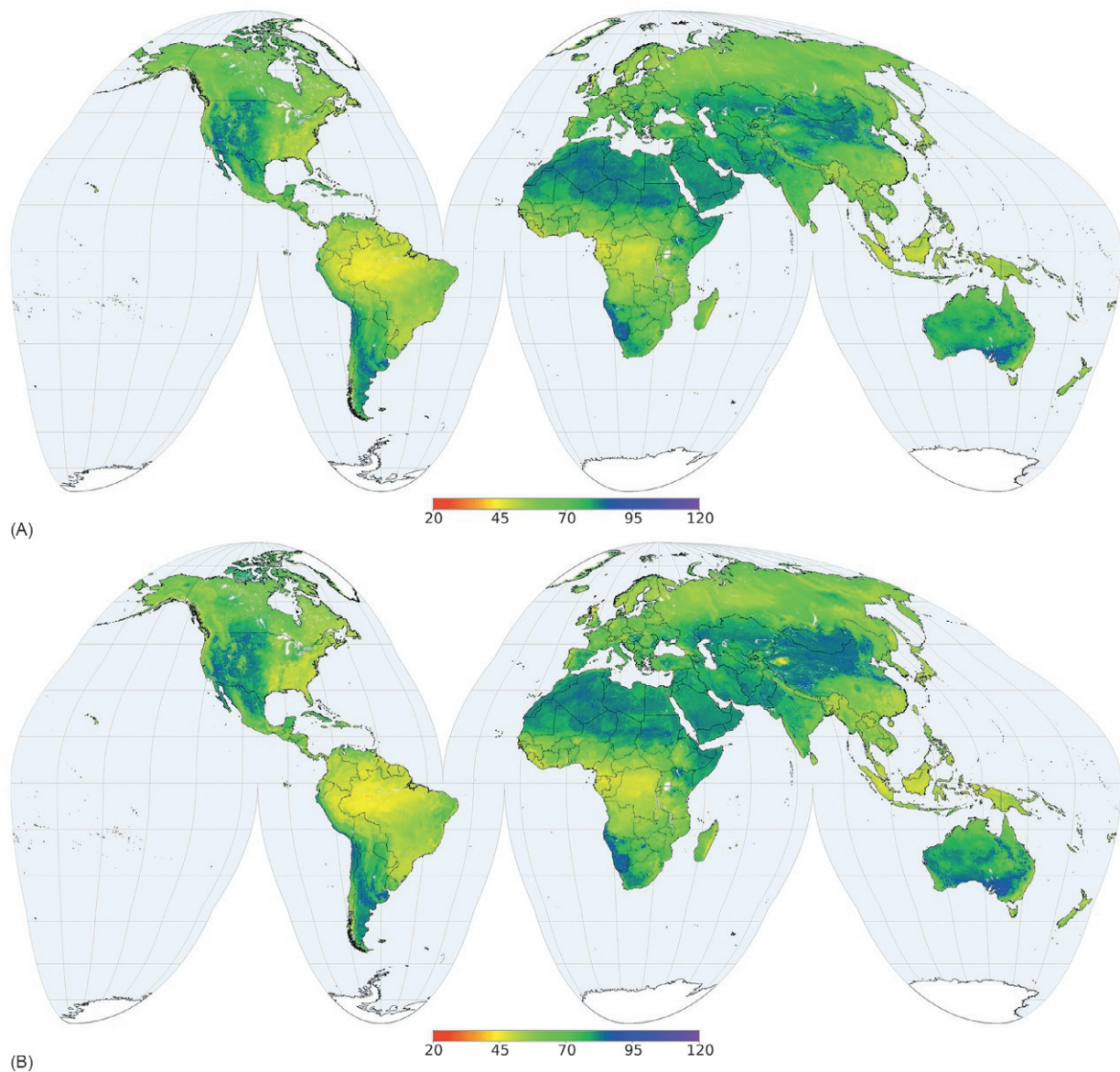


Fig. 3 Lower (A) and upper (B) limits of a 90% prediction interval of the SoilGrids 2.0 global map of pH_{water} (units in 10pH) in the 60–100 cm depth interval, as obtained using Quantile Regression Forests. Note that differences between the two maps are small, indicating high map accuracy. From Poggio L, de Sousa LM, Batjes NH, Heuvelink GBM, Kempen B, Ribeiro E, and Rossiter D (2021) SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *Soil 7*: 217–240.

If soil maps are derived by machine learning instead of geostatistics then it is more difficult to quantify the prediction error at each location in a region. The most common approach for this in digital soil mapping is Quantile Regression Forests (Meinshausen, 2006; Vaysse and Lagacherie, 2017; Poggio et al., 2021). This method requires no assumption of a normal distribution; instead it computes the quantiles of the probability distribution in a non-parametric way. All quantiles together characterize the probability distribution of the prediction error. GlobalSoilMap standards (Arrouays et al., 2014) request the 0.05 and 0.95 quantiles, which are the lower and upper limits of a 90% prediction interval, respectively. Fig. 3 shows an example.

Multivariate distribution of soil map errors

A comprehensive definition of soil map uncertainty also requires that the multivariate or joint distribution of map errors at multiple locations is specified. The bivariate probability distribution specifies the joint distribution of map errors for pairs of locations s and s' in D . An example of this bivariate distribution is given in the bottom panel of Fig. 1, where it is portrayed by 'contour' lines of equal probability density. Map errors are often positively correlated if the distance between s and s' is small, as is illustrated by the

rotated ellipses in Fig. 1. Fig. 1 also shows that the marginal, univariate distributions of the errors at the two locations are obtained by projection of the bivariate distribution on the two axes.

If we invoke the multivariate normal distribution and assume that the map errors at all locations in the region are jointly normally distributed then all that is needed to characterize this distribution is a vector of means and a variance–covariance matrix. These so-called two-point statistics are enough because for the multivariate normal distribution the higher-order moments are completely determined by them.

Kriging implicitly quantifies the correlation between map errors at any pair of locations; that is, it quantifies the correlation between the kriging prediction errors $\varepsilon(s)$ and $\varepsilon(s')$. They are usually not explicitly computed, but are taken into account when values from the joint distribution are obtained by conditional spatial stochastic simulation (Webster and Oliver, 2007, Chapter 12). Fig. 4 shows an example. These simulations or ‘possible realities’ are created with the help of a pseudo-random number generator. In principle, an infinite number of possible realities can be generated. Each of these could be the ‘true’ reality, but we do not know which one it is; we are uncertain. Thus, the differences among them express our uncertainty about the true situation at any location. In fact, the spread (represented by the standard deviation) computed from very many possible realities will equal the kriging standard deviation. Their average equals the kriging prediction, which implies that maps of the prediction errors (i.e. the differences between the true values and the kriged predictions, see bottom row of Fig. 4) average to zero over many simulations. This is a consequence of kriging’s unbiasedness. Spatial stochastic simulation and the possible realities that it produces are essential inputs for the propagation of uncertainty through Monte Carlo methods, as described below.

Machine learning methods cannot compute the joint distribution of map errors at multiple locations because they are essentially non-spatial and do not account for residual spatial correlation (Heuvelink and Webster, 2022). If one requires maps of prediction errors that take into account spatial correlation, for computing spatial averages and their uncertainty, for example, then a possible solution is to krig the residuals of the machine learning model (e.g. Szatmári et al., 2021).

Cross-correlation, positional uncertainty and uncertainty of categorical soil variables

Environmental scientists and agencies often want to consider several soil properties together and the errors arising from their analysis. If two properties are cross-correlated then one can use a similar approach to that above to characterize the joint distribution of $\varepsilon_1(s)$ and $\varepsilon_2(s)$, or that of $\varepsilon_1(s)$ and $\varepsilon_2(s')$ (Heuvelink et al., 2007). For example, cadmium and lead often occur together as pollutants and are correlated in the ordinary Pearson sense. If they are kriged to map the pollution their kriging prediction errors are likely to be positively correlated also, if only because a hot spot of high pollution may have been missed in sampling. Moreover, the interpolation error of cadmium at location s will probably be positively correlated with that of lead at location s' , if the distance between s and s' is smaller than the variogram’s range. All these correlations can be quantified using cokriging (Webster and Oliver, 2007, Chapter 10).

If measurement error is the only source of uncertainty then the correlation between errors of different soil properties may safely be assumed zero. For instance, even though organic carbon and total nitrogen in soil are usually positively correlated, their measurement errors are likely to be uncorrelated. This is because the two properties are measured in different ways and the measurement errors do not influence each other. Compositional data, such as particle-size distributions, are an exception. The proportions of sand, silt and clay measurements will be constrained to sum to 1 or to 100%. Their measurement errors are likely to be negatively correlated, therefore, because a positive measurement error for one proportion must be compensated by a negative measurement error for another.

Positional uncertainty

So far we have considered uncertainty only in the measured or predicted values of soil properties. In geographical information science this is known as ‘attribute uncertainty’. It is different from ‘positional uncertainty’, which refers to uncertainty about the geographic position of spatial objects, such as sampling locations. Positional uncertainty of soil data can be substantial, particularly with legacy data arising from sampling campaigns of many years ago. One can define positional uncertainty in a similar way to attribute uncertainty using a bivariate probability density, representing uncertainty in latitude and longitude. Positional uncertainty can be translated into attribute uncertainty if short-range spatial variation is known. When a perfect measurement at location s is taken to represent a soil property at $s + h$, with h a possible positional error, then this short-range variation means that the property at s almost certainly differs from that at $s + h$. Therefore, even though the measurement at $s + h$ is free of error, it nevertheless represents the property’s value at s with error.

Errors and uncertainty in soil class maps

For many years spatial variation in soil was represented solely by maps of soil classes, and those classes were defined by the soil in a profile regardless of any lateral variation. Errors were recognized, either because the boundaries between classes were drawn in the wrong places or because the parcels delineated contained profiles that did not belong to classes depicted, or usually both. In other words, pedologists and the clients for whom they mapped the soil were aware of the uncertainty of these maps, although it was rarely quantified.

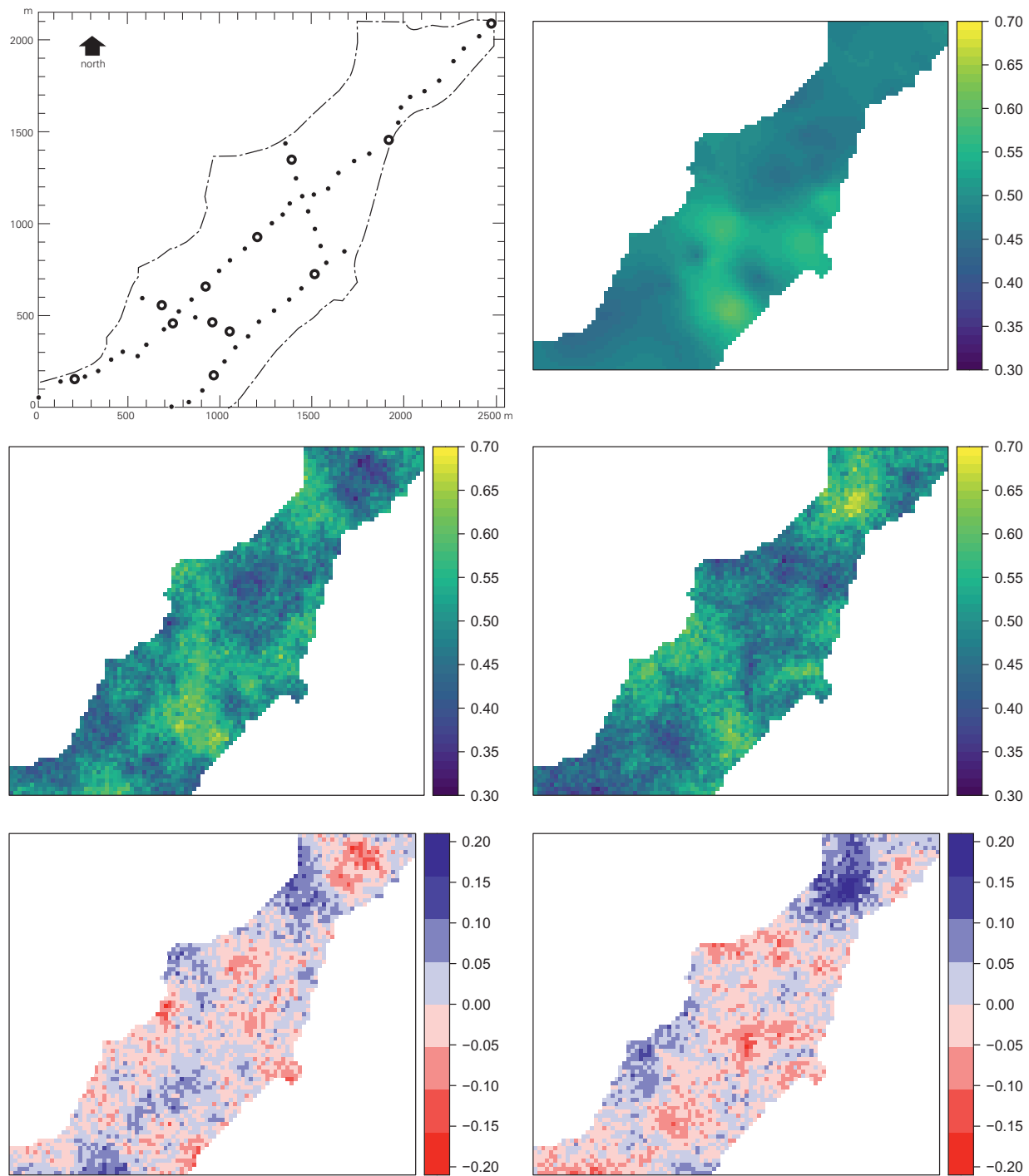


Fig. 4 Allier study area in France. Top left: stars show sampling sites of moisture content at field capacity and porosity; circled sites are those where in addition moisture content at wilting point was measured. Top right: kriging map of topsoil porosity. Middle row: two realizations of topsoil porosity, obtained using conditional spatial stochastic simulation. Bottom row: error maps corresponding with the two realizations, obtained by subtracting the kriging map from the realization. See Heuvelink (2018) for details.

Uncertainty in categorical soil variables such as soil type obviously cannot be characterized by models such as Eq. (1) and Eq. (3) nor can they be normally distributed. But we can form a discrete probability distribution by listing all possible values of a variable of interest at each location s and their associated probabilities. Clearly, these probabilities must all be non-negative and add to 1. While the definition is easy, acquiring them in practical situations for all locations in a region is more difficult. Indicator geostatistics may be used (Webster and Oliver, 2007, Chapter 11), or perhaps more advanced statistical methods such as the generalized linear geostatistical model (Diggle and Ribeiro, 2007; Kempen et al., 2012). Quantifying spatial dependence of uncertain categorical soil

variables is especially difficult. Bivariate distributions are not enough, and so-called ‘multi-point’ statistics are required, using techniques such as multi-point geostatistics (Meerschman et al., 2013).

As a last resort one may rely on expert judgment to quantify the uncertainty of categorical soil variables (O’Hagan, 2012; Lark et al., 2015). For example, to assess uncertainty of a soil class in a pit, several experts might be asked to classify the soil independently. If they all agree then the uncertainty will be small or absent, but if they do not then the distribution of the designated classes may be used to represent the uncertainty (Van Leeuwen et al., 2018). In fact, one can also use the judgment of experts for assessing uncertainty of properties measured and recorded on continuous scales, particularly when there are no replicates of field and laboratory measurements and when soil maps are made by non-statistical methods (Truong and Heuvelink, 2013).

The uncertainty of a map of soil classes as a whole can also be represented in a confusion matrix. If there are m classes then the confusion matrix, C , will have m rows, one for each of the mapped classes, and m columns, again one for each class. Each cell $C[i, j]$ of the confusion matrix contains the number of instances that the soil class at a location in the region that is mapped as class j is in fact class i . In practice, the confusion matrix is estimated from independent validation data, which are preferably placed randomly within the mapped region. The counts for each combination of map class and true class are then entered into the matrix. Counts in the diagonal of the matrix represent agreement; those in off-diagonal elements disagreement or error. The purity of the soil map may be expressed as the proportion of instances that the map was correct:

$$\text{purity} = \frac{1}{n} \sum_{i=1}^m C[i, i], \quad (4)$$

where n is the total number of observations. Note that the confusion matrix provides only a summary measure of the map accuracy. It is not spatially explicit. In contrast, indicator kriging and the generalized linear geostatistical model specify the map uncertainty for each and every location separately.

Change of support

Many users of soil maps are not interested in predictions of soil properties at points; they are much more concerned with averages of soil properties over larger areas of land. Such averages can be predicted by block kriging, as explained in the geostatistics chapter of the Encyclopedia. The blocks might be rectangular, even square, but they may take irregular shapes. They may even be as large as the entire mapped region. The geostatistics chapter also explains that the block kriging standard deviation is typically much smaller than the point kriging standard deviation, especially when the short-range variation is large. The reason for this is that within-block variation averages out when block means are predicted, and hence predicted block averages are more accurate than point predictions. So, the uncertainty of soil maps depends strongly on the support, where ‘support’ is defined as the area or volume over which a measurement or prediction is made. If the mapping did not include a change of support then the prediction support will be the same as the measurement support, which is usually treated as point support and can actually be either that of a single soil sample or a composite of several spread over some small area. Since uncertainty depends on the support, any database that reports the uncertainty of its soil maps must record the support of the map predictions as metadata.

In some circumstances users require predictions on smaller supports than those of the measurements, and if they do they may resort to area-to-point kriging (Kerry et al., 2012). This is rare for soil properties that are observed directly because the measurements are made on small supports. When properties are recorded or estimated by remote sensing, however, the supports can be substantially larger than users’ requirements; so too can those associated with the polygons of legacy soil maps (Kerry et al., 2012). Another example is the use of area-to-area kriging in the vertical dimension, to predict soil properties for layers that differ from those from which the soil sample was taken (Orton et al., 2016). In contrast to block kriging, area-to-point kriging leads to larger standard deviations because it makes predictions over smaller supports than those of the measurements.

Spatial support and resolution

Spatial support is often thought to be the same as spatial resolution. That is unfortunate; they are not; they differ in principle. For instance, the pH maps shown in Fig. 3 have a spatial resolution of $250 \text{ m} \times 250 \text{ m}$, but because the methods used to create these maps did not incorporate a change of support, the support of these maps is the same as that of the training data, that is the point support of the soil samples. The prediction interval limits shown in Fig. 3 thus refer to predictions at point support. The prediction interval would probably be narrower (smaller standard deviation) if the pH were predicted over a $250 \text{ m} \times 250 \text{ m}$ support, but this can be done only if the variogram of the map error is known and a geostatistical approach is used. One possibility is to apply block kriging to the residuals of a machine learning model. Szatmári et al. (2021) took this approach to quantify uncertainty of the change of soil organic carbon in Hungary between 1992 and 2010. They showed that uncertainty on large supports was substantially smaller than on small supports. Because of this they could show that the concentration of organic carbon in the soil had changed significantly between 1992 and 2010 for counties and the entire country, whereas for square blocks of $1 \text{ km} \times 1 \text{ km}$ and $10 \text{ km} \times 10 \text{ km}$ most predicted changes were not statistically significant, and on the point supports they found none.

Uncertainty propagation

We mentioned above the use of pedotransfer functions to predict soil properties that are difficult or time-consuming to measure from simpler basic ones. These are examples where soil information is used to compute derived properties. Soil maps are frequently used in this way—to predict by modelling crop yield, the emission of greenhouse gases, water infiltration, pesticide leaching and erosion, for example. Maps of basic soil properties are also used to derive more complex soil properties, such as soil functions and soil threats. The outputs of these models and analyses suffer from two main sources of uncertainty. One is model uncertainty, which may be split in uncertainty concerning the form of the model and uncertainty of the actual parameters of the model (e.g. Chapagain et al., 2022). Here we focus on the second one, namely uncertainty of the input variables. If the input to a model is uncertain then that uncertainty propagates into the model's output.

There are two main methods for analyzing the propagation of uncertainty through environmental models. The Taylor series method can be used for fairly simple models that are described by a continuous-differentiable mathematical function. It approximates these models locally by a linear or quadratic function, after which the propagation of uncertainty can be computed analytically (Heuvelink, 2018). The Monte Carlo method is more flexible, but it is also computationally demanding. It requires the joint probability distribution of all uncertain inputs, from which realizations or 'possible realities' are generated with a pseudo-random number generator. For spatially distributed inputs one makes use of spatial stochastic simulation, as discussed above. By repeatedly sampling from the input distribution and running the model for each realization, one obtains an equivalently sized sample from the model's output distribution. Thus, Monte Carlo simulation involves the following steps.

1. Repeat N times (typically $100 \leq N \leq 10,000$):
 - (a) Generate a realization of each uncertain model input.
 - (b) Run the model with this set of realizations and store the model output.
2. Compute parameters of the distribution of the N outputs and use these to characterize the uncertainty of the model output (e.g. mean, standard deviation, percentiles).

The Monte Carlo method is flexible in that it imposes no restrictions on the model. It can also deal with categorical uncertain inputs, whereas the Taylor series method cannot. The main difficulties are the computational burden and the need to characterize input uncertainty with a probability distribution, which can be a complex task for spatial inputs, as mentioned above. The Monte Carlo method involves an approximation error, but one can diminish this error by increasing the number, N , of Monte Carlo runs.

There are numerous applications of Monte Carlo uncertainty propagation in soil science. Some examples are Bishop et al. (2006), Van den Berg et al. (2012), and Benke et al. (2018). One can use it also to assess the contribution of individual uncertain inputs to the overall output uncertainty. One repeats the uncertainty propagation analysis with all uncertainty sources present except that for which the uncertainty contribution is required. This input is deemed certain and set to a representative value (usually the mean of the probability distribution). One then assesses the contribution of this input to the model's output uncertainty by evaluating the reduction of uncertainty that is caused by setting it to its representative value.

Uncertainty propagation example

Fig. 5 shows an example where model uncertainty and input uncertainty were propagated by the Taylor series method. The example refers to the Allier region in France shown in Fig. 4. A simple pedotransfer model that took the shape of a multiple linear regression model and that predicts the soil moisture content at wilting point from soil moisture at field capacity and soil porosity was first calibrated with data from 12 locations in the region. Model uncertainty was quantified by the variances and covariances of the regression coefficients and the residual variance; input uncertainty was derived by cokriging. Fig. 5 shows that the standard deviation of the model's output (i.e. of the soil moisture at wilting point) is smaller than that of the predicted output, although there are locations where the ratio of the two exceeds 0.5. This occurs mainly at places far from where the model inputs were measured (see top left panel in Fig. 4). The bottom panel of Fig. 5 shows that in this case the contribution of input uncertainty was much larger than that of model uncertainty. This indicates that one could obtain a more accurate map of soil moisture content at wilting point by taking more measurements of soil moisture at field capacity and soil porosity. It is remarkable that model uncertainty had a small contribution, given that the model was calibrated on only 12 paired observations.

The added value of uncertainty information

We explain above how uncertainty in soil measurements and soil maps can be quantified and used in analyses to propagate uncertainty. The quantification and propagation of uncertainty makes substantial demands for data and is not easily done. So, is it worth the effort? What is the added value of an uncertainty assessment of spatial soil information? Below we list several purposes that it serves.

1. First and foremost, any self-respecting investigator should want to check the quality of his or her results before they are made public. We should not publish soil data that have large errors nor publish poor soil maps. If they are the only data available, however, we should be honest about that and report their accuracy.

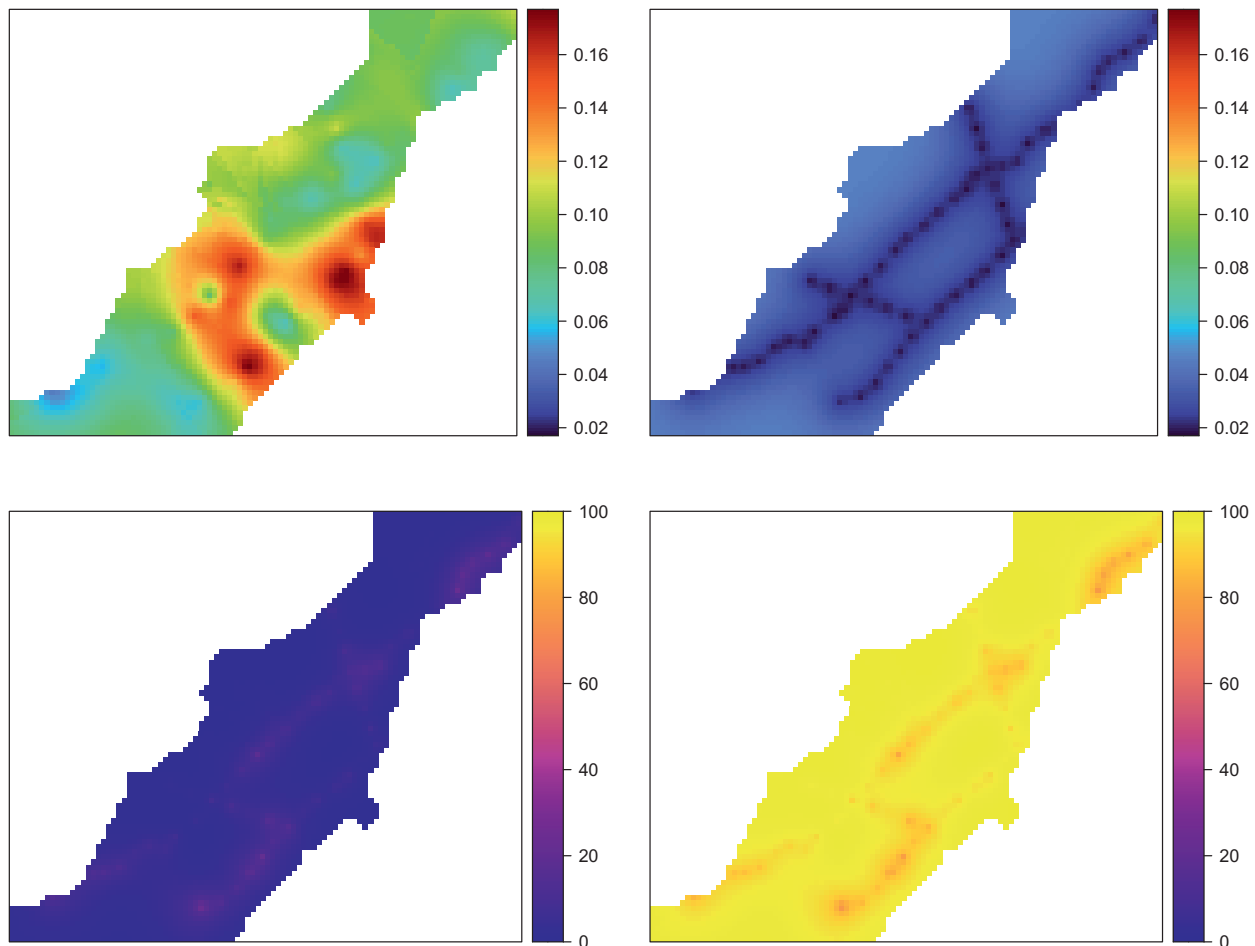


Fig. 5 Uncertainty propagation in the Allier study area. Top left: predicted soil moisture content at wilting point derived from kriged maps of soil moisture at field capacity and soil porosity using a pedotransfer function. Top right: standard deviation of soil moisture at wilting point, obtained using a Taylor series uncertainty propagation analysis where both model uncertainty and input uncertainty were included. Bottom left: Contribution (%) of model uncertainty to overall uncertainty in soil moisture content at wilting point. Bottom right: Contribution (%) of input uncertainty to overall uncertainty in soil moisture content at wilting point. See Heuvelink (2018) for details.

2. Clients and end users must know the quality of soil information so as to judge whether the information is good enough for their intended use of it. For example, there is serious regard for the soil's ability to store more carbon and mitigate climate change (Amelung et al., 2020). There have been claims of success in sequestering carbon, but these claims must be verified. One can do that by estimating the increase in organic carbon in a region between the start and end date of a project, though one would want to test the increase for statistical significance to be confident that the increase was real. And that would necessitate the quantification of the uncertainty.
3. Quantified uncertainty also helps producers of spatial soil information because it allows them to compare the performances of two or more mapping techniques, to determine which performs best, and to decide which technique to use in any particular case. For example, do many cheap, but inaccurate, observations from proximal sensors provide more accurate maps than few expensive measurements by wet chemistry? Does it pay to replace a simple empirical erosion model by a complex mechanistic model that needs many more input and calibration data but could predict erosion rates more accurately? These questions can be answered only if the uncertainty is quantified.
4. Geostatistical modelling of uncertainty enables one to see how uncertainty is influenced by a change of support. The larger are the supports at which one predicts the greater is the certainty. But that in itself is not a sufficient reason to predict on large supports. The prediction must depend on the application at hand; what size of support does the user need or demand? The farmer, for example, who wishes to apply fertilizer at a variable rate would want predictions of nutrient concentrations on a support that matched the width of his fertilizer spreader. He or she would be unlikely to want predictions on point supports from which the measurements were made. In like manner, emissions of nitrous oxide from agricultural land might be estimated for fields, farms, large regions, even whole countries, if the aim is to learn how much each is contributing to the total nitrous

oxide in the atmosphere. There is a trade-off between accuracy and spatial support, and knowing this relation is indispensable for informed decisions on the spatial scale at which to work and present results.

5. Uncertainty in soil data propagates through models and spatial analyses to output predictions and maps. Quantifying and tracking that uncertainty is valuable because it enables end-users to know how accurate the results are, and that in turn enables them to take into account that uncertainty when making decisions. They can do so, however, only if the uncertainty in the input, i.e. in the data, is quantified. Uncertainty propagation analysis can tell how much each of the uncertainty sources contributes to the output uncertainty. This helps modelers and users to take rational decisions on how to reduce uncertainty about the model output. It is usually best to tackle the main source of uncertainty (i.e., the weakest link in the chain), although this also depends on how easy or difficult it is to reduce input uncertainty. For instance, if an interpolated map of the rootable depth is the main source of uncertainty of an agronomic model that predicts potential yield in West Africa, it may not be easy to collect more point data on rootable depth in the region to improve the map.
6. Quantified uncertainty can be included in decision making, as in risk analysis. Many users want to avoid risk; they would like certainty. They cannot have it, of course, but they can be told what the uncertainty is, and that should help them when they are deciding what to do with the information provided. For instance, a farmer might be unwilling to buy fertilizers if plague, pests or drought are likely seriously to cause loss of his crop even if the expected net-return of fertilization is positive. Breure et al. (2022) showed how farmers can account for uncertainty in spatial decision making. Incorporating uncertainty information in spatial decision making has not yet received the attention it deserves by soil scientists, though as Lark et al. (2022) remarks there is much potential benefit for it.

Finally, note that for some purposes soil maps can be validated and their uncertainty quantified by design-based, probability sampling and statistical analysis and inference (Brus et al., 2011). It has several important and attractive properties. It requires no assumptions about the nature of the spatial distribution, is easily applied, and is backed by well-founded theory. It has its disadvantages, too. It can provide only summary measures of a map's accuracy; it cannot easily estimate the prediction errors for supports larger than those of the data; and it requires an independent probability sample from the population (i.e., the study area). It is a large subject well beyond the scope of this chapter, but it is comprehensively covered by De Gruijter et al. (2006).

Conclusion

Soil information is rarely if ever free of error. The main sources of uncertainty are measurement error, classification error, interpolation error and model error. That uncertainty in soil information can be adequately characterized by probability distributions, although such distributions might be complex and difficult to assess when spatial variation and cross-correlation between soil properties are involved. There are considerable advantages in quantifying uncertainty in soil information: it tells users whether the data are sufficiently accurate for their intended use, it helps producers of soil information to select the best mapping method, it allows analyses of how uncertainty in soil data propagates through environmental models, and it supports decision making under uncertainty.

Many soil databases provide no or only rudimentary information about the accuracy of the data they contain. This is certainly so for laboratory and field measurements, because laboratory errors are often not quantified or not copied into soil databases. Digital soil mappers take pride in routinely quantifying the prediction uncertainty of their maps, but not all of this information is transferred to soil spatial databases. There is also much to be done to communicate uncertainty information better to end-users and help them to use this information in decision making.

Soil science is continuously advancing. New theories and models are proposed and tested, and new techniques for acquiring data help us to improve the accuracy of soil information. The information will never be perfect, however; uncertainty is here to stay. Nevertheless, we have the tools to control and express uncertainty, and soil scientists and practitioners are urged to learn to use them to everyone's advantage.

References

- Amelung W, Bossio D, de Vries W, Kogel-Knabner I, Lehmann J, Amundson R, Bol R, Collins C, Lal R, Leifeld J, Minasny B, Pan G, Paustian K, Rumpel C, Sanderman J, van Groenigen JW, Mooney S, van Wesemael B, Wander M, and Chabbi A (2020) Towards a global-scale soil climate mitigation strategy. *Nature Communications* 11: 5427.
- Arrouays D, Grundy MG, Hartemink AE, Hempel JW, Heuvelink GBM, Hong SY, Lagacherie P, Lelyk G, McBratney AB, McKenzie NJ, Mendonca-Santos MDL, Minasny B, Montanarella L, Odeh IOA, Sanchez PA, Thompson JA, and Zhang G-L (2014) GlobalSoilMap: Toward a fine-resolution global grid of soil properties. In: Sparks DL (ed.) *Advances in Agronomy. Advances in Agronomy*, vol. 125, pp. 93–134. San Diego, CA: Academic Press.
- Benke KK, Norg S, Robinson NJ, Benke LR, and Peterson TJ (2018) Error propagation in computer models: analytic approaches, advantages, disadvantages and constraints. *Stochastic Environmental Research and Risk Assessment* 32: 2971–2985.
- Bishop T, Minasny B, and McBratney A (2006) Uncertainty analysis for soil-terrain models. *International Journal of Geographical Information Science* 20: 117–134.
- Breure TS, Haefele SM, Hannam JA, Corstanje R, Webster R, Moreno-Rojas S, and Milne AE (2022) A loss function to evaluate agricultural decision-making under uncertainty: A case study of soil spectroscopy. *Precision Agriculture* 23: 1333–1353.
- Brown JD and Heuvelink GBM (2005) Assessing uncertainty propagation through physically based models of soil water flow and solute transport. In: Anderson MG (ed.) *Encyclopedia of Hydrological Sciences*, 1181–1195.
- Brus DJ, Kempen B, and Heuvelink GBM (2011) Sampling for validation of digital soil maps. *European Journal of Soil Science* 62: 394–407.

- Chapagain R, Remenyi TA, Harris RMB, Mohammed CL, Huth N, Wallach D, Rezaei EE, and Ojeda JJ (2022) Decomposing crop model uncertainty: A systematic review. *Field Crops Research* 279: 108448.
- Chilès JP and Delfiner P (2012) *Geostatistics. Modeling Spatial Uncertainty*, 1st edn. New York: Wiley.
- De Gruijter JJ, Brus DJ, Bierkens MFP, and Knotters M (2006) *Sampling for Natural Resources Monitoring*. Berlin: Springer.
- Deng H, Ye M, Schaap MG, and Khaleel R (2009) Quantification of uncertainty in pedotransfer function-based parameter estimation for unsaturated flow modeling. *Water Resources Research* 45: W04409.
- Diggle PJ and Ribeiro PJ Jr. (2007) *Model-Based Geostatistics*, 1st ed. New York: Springer.
- Hengl T, Heuvelink GBM, and Stein A (2004) A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* 12: 75–93.
- Heuvelink GBM (2018) Uncertainty and uncertainty propagation in soil mapping and modelling. In: McBratney AB, Minasny B, and Stockmann U (eds.) *Pedometrics*, pp. 439–461. Springer International Publishing AG.
- Heuvelink GBM and Webster R (2022) Spatial statistics and soil mapping: A blossoming partnership under pressure. *Spatial Statistics* 50: 100639.
- Heuvelink GBM, Brown JD, and van Loon EE (2007) A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science* 21: 497–513.
- Kempen B, Brus DJ, and Heuvelink GBM (2012) Soil type mapping using the generalised linear geostatistical model: A case study in a Dutch cultivated peatland. *Geoderma* 189: 540–553.
- Kerry R, Goovaerts P, Rawlins BG, and Marchant BP (2012) Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale. *Geoderma* 170: 347–358.
- Lark RM, Lawley RS, Barron AJM, Aldiss DT, Ambrose K, Cooper AH, Lee JR, and Waters CN (2015) Uncertainty in mapped geological boundaries held by a national geological survey: eliciting the geologists' tacit error model. *Solid Earth* 6: 727–745.
- Lark RM, Chagumaira C, and Milne AE (2022) Decisions, uncertainty and spatial information. *Spatial Statistics* 50: 100619.
- Meerschman E, Van Meirvenne M, Van de Vijver E, De Smedt P, Islam MM, and Saey T (2013) Mapping complex soil patterns with multiple-point geostatistics. *European Journal of Soil Science* 64: 183–191.
- Meinshausen N (2006) Quantile regression forests. *Journal of Machine Learning Research* 7: 983–999.
- Nosrati K (2013) Assessing soil quality indicator under different land use and soil erosion using multivariate statistical techniques. *Environmental Modelling and Assessment* 185: 2895–2907.
- O'Hagan A (2012) Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux. *Environmental Modelling & Software* 36: 35–48.
- Oliver MA (2010) *Geostatistical Applications for Precision Agriculture*. Dordrecht: Springer.
- Orton TG, Pringle MJ, and Bishop TFA (2016) A one-step approach for modelling and mapping soil properties based on profile data sampled over varying depth intervals. *Geoderma* 262: 174–186.
- Poggio L, de Sousa LM, Batjes NH, Heuvelink GBM, Kempen B, Ribeiro E, and Rossiter D (2021) SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *The Soil* 7: 217–240.
- Stenemo F and Jarvis N (2007) Accounting for uncertainty in pedotransfer functions in vulnerability assessments of pesticide leaching to groundwater. *Pest Management Science* 63: 867–875.
- Szattmári G, Pásztor L, and Heuvelink GBM (2021) Estimating soil organic carbon stock change at multiple scales using machine learning and multivariate geostatistics. *Geoderma* 403: 115356.
- Takoutsing B, Heuvelink GBM, Stoorvogel JJ, Shepherd KD, and Aynekulu E (2022) Accounting for analytical and proximal soil sensing errors in digital soil mapping. *European Journal of Soil Science* 73: e13226.
- Truong PN and Heuvelink GBM (2013) Uncertainty quantification of soil property maps with statistical expert elicitation. *Geoderma* 202: 142–152.
- Van den Berg F, Tiktak A, Heuvelink GBM, Burgers SLGE, Brus DJ, de Vries F, Stolte J, and Kroes JG (2012) Propagation of uncertainties in soil and pesticide properties to pesticide leaching. *Journal of Environmental Quality* 41: 253–261.
- Van Leeuwen MMWJ, Heuvelink GBM, Wallinga J, de Boer IJM, van Dam JC, vanEssen EA, Moolenaar SW, Verhoeven FPM, Stoorvogel JJ, and Stoof CR (2018) Visual soil evaluation: Reproducibility and correlation with standard measurements. *Soil & Tillage Research* 178: 167–178.
- Van Leeuwen CE, Mulder VL, Batjes NH, and Heuvelink GBM (2021) Statistical modelling of measurement error in wet chemistry soil data. *European Journal of Soil Science* 73: e13137.
- Vayssié K and Lagacherie P (2017) Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma* 291: 55–64.
- Viscarra Rossel RA, McBratney AB, and Minasny B (2010) *Proximal Soil Sensing*. Dordrecht: Springer.
- Webster R (2000) Is soil variation random? *Geoderma* 97: 149–163.
- Webster R and Oliver MA (2007) *Geostatistics for Environmental Scientists*, 2nd ed. Chichester: Wiley.
- WEPAL (2022) International Soil-Analytical Exchange Programme. <https://www.wepal.nl/en/wepal/Home/Proficiency-tests/Soil/Proficiency-tests/ISE.htm>.

Digital soil mapping: Evolution, current state and future directions of the science

Brendan Malone^a, Dominique Arrouays^b, Laura Poggio^c, Budiman Minasny^d, and Alex B McBratney^d, ^aCSIRO Agriculture and Food, Butler Laboratory, Black Mountain, Acton, ACT, Australia; ^bINRAE, InfoSol, Orléans, France; ^cISRIC World Soil Information, Wageningen, Netherlands; ^dSydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Camperdown, NSW, Australia

© 2023 Elsevier Ltd. All rights reserved.

Introduction	684
Digital soil mapping concepts	685
Soil spatial prediction functions	686
Assessing digital soil map quality	687
Quantifications of uncertainty	687
Nuances of digital soil mapping	688
Knowledge-based inference and mechanistic modelling for digital soil mapping	688
Space–time modelling	689
Soil depth functions and digital soil mapping	689
Generalizing the practice of digital soil mapping	691
Operationalization of digital soil mapping	692
The GlobalSoilMap initiative	693
Uses and users of digital soil mapping	693
Conclusions	694
References	694

Abstract

Digital soil mapping (DSM) entails the creation and population of spatial soil information systems by numerical models inferring the spatial and temporal variations of soil types and soil properties from soil observation and knowledge derived from related environmental variables. This chapter summarizes the state of the science of DSM and the mechanics of its various implementations. It has moved from research to practical application and is now widely used as a method for making soil maps. The outlook for continuing research into DSM and related efforts is positive, because objective spatio-temporal assessments of the soil resources are needed and continually require updating.

Key points

- This chapter describes some of the underpinning mechanics of digital soil mapping, celebrates some of its successes, and identifies areas of the science which need further thought and work.
- Digital soil mapping has been borne out and developed alongside technological developments which include advances in computers and computer sciences, together with unparalleled growth in associated modelling technologies and widespread availability of digital Earth observation data.
- Digital soil mapping augments rather than replaces conventional soil survey and is recognized as an additive technology that enables delivery of richer spatial soil information and has the potential to deliver products customized to the needs of end-users.

Introduction

In recent times, we have witnessed the advancement of the computer and information technology ages. With such advances, vast amounts of data and tools in all fields of endeavor have become available. This has motivated numerous initiatives around the world to build spatial data infrastructures aiming to facilitate the collection, maintenance, dissemination and use of spatial information.

The science of digital soil mapping has been borne out of and developed alongside these technological developments. Digital soil mapping (DSM) is a numerical and model-based framework for inferring the spatial and temporal variation of soil types and properties from soil observation and knowledge derived from related environmental variables. Digital soil mapping has never been considered a replacement to traditional forms of soil survey and mapping, rather as a complementary tool, albeit quantitative in principle, to characterize the spatial and temporal properties of soils in ways that legacy soil mapping could not achieve.

There are many aspects to DSM, including its generalizations given different forms of soil information, diverse modelling frameworks and accommodation for both 2- and 3-dimensional vertical support, and the quantification of uncertainties, all of which are discussed in this chapter.

Digital soil mapping has matured as a science and is used as an operational tool throughout the world for the assessment and characterization of soils within both government and privately funded organizations.

Globally, the need for comprehensive and specific digital soil information has increased, particularly amidst emerging existential crises of climate change, land degradation, and unsustainable land management practices in an increasingly populous world. Digital soil mapping will continue to evolve together with the broader technological advancements. More importantly, it will also branch out further into digital soil assessment, for example in the quantification of soil function and ecosystem services. It is this derived information that is required for assessing the status of ecosystems more generally, rather than just the creation of digital maps of soil characteristics and properties. To date, we are already down that path, and well positioned to confront these environmental challenges given the large and active network of practitioners throughout the world.

Digital soil mapping concepts

The various threads in research on soil spatial prediction began to converge in the early 2000s. The development of fuzzy logic and geostatistics and then soil–landscape quantification in the 1990s, brought about a strong theoretical foundation that enabled modern digital soil mapping to emerge rapidly at the turn of the millennium. Two seminal papers were published in 2003 that summarized historical developments, albeit from different perspectives. Scull et al. (2003) wrote about past research in soil spatial prediction from a physical geography perspective and coined the term predictive soil mapping to describe the general approach. About the same time, McBratney et al. (2003) generalized the diversity of quantitative mapping approaches into a framework they called ‘scorpan.’ This uses a Jenny (1941) clorpt-like framework not for mere explanation but rather empirical quantitative description of soil–landscape relationships for spatial prediction. They called these activities digital soil mapping (DSM).

The ‘scorpan’ framework is more formally known as scorpan-SSPFe, which includes soil spatial prediction functions (SSPF) and autocorrelated errors (McBratney et al., 2003). The scorpan factors are:

- s: soil, its classes, or properties.
- c: climate, including precipitation and temperature.
- o: organisms, including vegetation, fauna, and other factors of the biotic environment.
- r: relief, or topography.
- p: parent material, including lithology.
- a: age.
- n: space, or spatial position.

The model is written as:

$$S_c = f(s, c, o, r, p, a, n) \text{ or } S_p = f(s, c, o, r, p, a, n)$$

where S_c is soil classes and S_p is soil properties or attributes. Considering soils are sampled at spatial coordinates x, y at an approximate point in time, $\sim t$, the model can be expressed explicitly as:

$$S[x, y, \sim t] = f(s[x, y, \sim t], c[x, y, \sim t], o[x, y, \sim t], r[x, y, \sim t], p[x, y, \sim t], a[x, y, \sim t], n[x, y, \sim t]).$$

Soil (s) is included as a factor because soil can be predicted from its properties, and soil properties from its class or other properties (McBratney et al., 2003). This additional soil information could be gathered from a prior soil map or from either remote or proximal soil sensing or even expert knowledge. The factor n means that soil can be predicted as a function of spatial position alone, as in the case of kriging, but it may also be predicted as a function of the distance from some landscape features such as streams, hilltops, roads or point sources of pollution, etc. Information that may represent the other scorpan factors is described in Table 1.

Table 1 Possible sources of information to represent the scorpan factors.

Scorpan factor	Possible representatives
s	Legacy soil maps, point observations, expert knowledge
c	Temperature and precipitation records, remote sensing derive climate data
o	Vegetation maps, species abundance maps, yield maps, land use maps, remote sensing derived vegetation indices
r	Digital elevation model, terrain attributes
p	Legacy geology maps, gamma radiometric information
a	Weathering indices, geology maps
n	Latitude and longitude or easting and northing, distance from landscape features, distance from roads, distance from point sources of pollution

Since the establishment of DSM in the early 2000s, there have also been developments and huge leaps forward in terms of Earth observation data from both remote and proximal sensing platforms. This has resulted in a proliferation of high-resolution environmental spatial data, and many of these can be used to represent various scorpan factors. Digital elevation models (DEMs) and satellite imagery data are two prominent examples. Subsequently, DSM techniques have been deployed to build or populate spatial soil information systems from relatively sparse datasets from the ground up. The DSM techniques have also been used to update or renew existing soil maps (e.g. [Kempen et al., 2009](#)). It is seen as a practicable framework for fulfilling the current and future demand for relevant soil information ([Searle et al., 2021](#)).

In practice, DSM starts by defining a mapping domain, and soil samples are taken based on an established sampling design. Several sampling designs for DSM have been developed, such as stratified sampling using *k*-means algorithm or conditioned Latin hypercube sampling. This is followed by intersecting the observations with a set of pedologically meaningful environmental layers. Then, fitting a mathematical or machine learning model to obtain a spatial soil prediction function. Once the model has been fitted at the observation points, it is then extended to all grid cell nodes of the raster layers giving a digital soil map. This three-component process is the hallmark of DSM ([Minasny and McBratney, 2016](#)), which entails:

1. The input which includes soil data and associated environmental covariates.
2. The modelling process, and
3. The output, i.e. the digital soil map with uncertainty.

This DSM approach is quite distinct from earlier notions of digital soil mapping that simply involved the digitization of conventional soil maps by electronic scanning. A more appropriate term for this would be digitized soil map.

Soil spatial prediction functions

The modelling step is crucial in the digital soil mapping process and fundamentally distinguishes it from digitized soil mapping. The form of the spatial soil prediction function $f()$ is usually determined at the outset. When deciding which mathematical model is the most appropriate for a given application, several factors are usually considered, including:

1. The operator's familiarity with the model.
2. The model's ease of application in the context of the project, the availability of covariates and the idiosyncrasies of the available soil information.
3. The model's complexity and its power to capture potentially complex soil–landscape relationships within the target mapping domain.

Many mathematical models are available to represent $f()$, and their development continues with advances in statistical and machine learning models. In general, some models, such as multiple linear regression or regression trees, are suited to modelling continuous variables whereas others, such as logistic regression or classification trees, are suited to modelling ordinal or nominal categorical data.

Some of the simplest models are simple linear models with either ordinary or generalized least-squares fitting. More complex models include generalized linear and additive models. Logistic regression models are a type of generalized linear model suited to modelling categorical variables. Recursive partition models such as classification and regression trees are particularly favorable because of their non-parametric structure and their capacity to deal with non-linear relationships between soil data and environmental covariates. For similar reasons machine learning algorithms (such as random forests and support vector machines to mention only two) and neural networks are often considered, as are deep convolutional neural networks, particularly in terms of incorporating spatial contextual information ([Padarian et al., 2019](#)) for which these algorithms are suited. Collectively, these model structures provide DSM practitioners with a rich and varied suite of options to explore complex soil spatial variation.

In terms of handling any spatial autocorrelation in the residual (e) that is likely to result from fitting a scorpan model, regression kriging (alternatively scorpan kriging; [McBratney et al., 2003](#)) or universal kriging could be used. Coupling a machine learning model with geostatistical modelling of the residuals is also a generalized regression kriging approach that is often used in digital soil mapping studies.

[Lark et al. \(2006\)](#) described a REML–EBLUP (residual maximum likelihood–empirical best linear unbiased predictor) model. The REML–EBLUP is intrinsically like regression-kriging in that both are mixed models where the observed data are modelled as the additive combination of fixed effects (the secondary environmental data), random effects (the spatially correlated residuals e), and independent random error. The difference is that REML estimates the parameters of the trend and covariance functions without bias. These parameters are then used in the EBLUP i.e. a general linear mixed model.

The application of formal geostatistical modelling approaches to categorical variables, such as soil classes, is relatively limited. [Kempen et al. \(2012\)](#) reminded us that the popular methods for categorical prediction of soil types to create a digital soil type map—multinomial logistic regression, classification trees etc.—are non-spatial models i.e. we do not consider spatial locational properties of the target variable as for continuous variables. Subsequently, they explored a generalized linear geostatistical model framework that addressed this issue. The method was theoretically appealing, but was computationally cumbersome and ultimately did not yield significant gains in accuracy when compared to a non-spatial multinomial logistic regression model.

Assessing digital soil map quality

In DSM, a model is created, and a map is produced, which then needs to be verified in terms of whether the resultant spatial patterns appear plausible across the landscape. To date this is a largely qualitative assessment, and usually requires expert elicitation or an assessment by the model creator who may be familiar with the area mapped and will be able to verify accordingly.

Quantitative assessments of map quality are usually done by assessing the fidelity between observations (observed data of soil phenomena) with corresponding co-located model predictions. For continuous variables, goodness of fit statistical measures are based on some form of deviation assessment such as bias and root mean square error between observations and predictions (Hastie et al., 2009). For categorical data, the measures include overall, user and producer accuracies together with indices such as the kappa coefficient.

For completely unbiased assessments of model quality, it is ideal to have an additional data set that is completely independent of the model data. When validating trained models with some sort of data sub-setting mechanism, keep in mind that the validation statistics will be biased. As Brus et al. (2011) explain, sampling from the target mapping area to be used for DSM is often from legacy soil survey. Consequently, these data are unlikely to have been collected using a probability sampling design. Therefore, the sample will be biased i.e. not a true representation of the total population. Thus, an independent probability sample is required. It is recommended that some random sampling from the target area be conducted, such as simple random sampling and stratified simple random sampling.

From an operational perspective it is usually difficult to arrange the additional costs of organizing and implementing some sort of probability sampling for determining unbiased model quality assessment. The alternative is to perform some sort of data sub-setting, such that with a data set split it into a set for model calibration and another set for validation. This type of procedure can take different forms: the two main ones being random-hold back and leave-one-out-cross-validation (LOCV). Random-hold back is where we may sample a data set by some pre-determined proportion (say 70%) which is used for model calibration. We then validate the model using the other 30% of the data. If the dataset is small, there could be a random chance that the model could overstate or falsely represent model confidence. To avoid such outcomes, this random split procedure needs to be performed many times, and the mean of such validation will provide a more representative assessment.

For k -fold validation we divide the dataset into equal sized partitions or folds, with all but one of the folds being used for the model calibration; the remaining fold is used for validation. We could repeat this k -fold process several times, each time using a different random sample from the data set for model calibration and validation. This allows one to derive distributions of the validation statistics efficiently to assess the stability and sensitivity of the models and parameters. A variant of k -fold cross-validation (CV) is spatial cross-validation where, instead of random subsets of data to act as the different folds, the data are clustered spatially into groups (which we can think of as folds to keep with the general idea).

Quantifications of uncertainty

It is relatively easy to quantify the accuracy of a digital soil map using a set of soil observations set apart from the spatial modelling process for this purpose. A range of statistical tests is available and provides a global appraisal of model performance: that is, they typically cannot tell us anything about the performance of a model at a specific location in the prediction area.

On the other hand, the local appraisal of a model's performance can be done through examination of uncertainties if they are available. Such uncertainties can be quantified at each grid cell across the prediction area and are often expressed in the form of a prediction variance or prediction interval for soil attributes, or even probability estimate for soil classes or exceedance thresholds (Fig. 1).

Uncertainties may be computed with a range of methods. For example, it is relatively straightforward to compute the kriging variance with geostatistics in the case of purely spatial models of soil variation. Minasny et al. (2011) demonstrated a model-based Bayesian approach, while Bayesian networks (Taalab et al., 2015) offer a more expert-driven approach.

Machine-learning methods, and particularly those based on iterative resampling and boosted model fitting, can provide empirical estimates of uncertainty. An example here is the quantile regression forest algorithm that was used in France by Vaysse and Lagacherie (2017).

Methods based on model perturbation through data resampling or bootstrapping have been shown to be effective for quantifying model uncertainties, particularly where the number of model parameters is high, or for very large mapping extents the use of model-based geostatistical approaches is computationally prohibitive. Malone et al. (2011) demonstrated another empirical approach which is based on data resampling and fuzzy k -means with extragrades.

In quantifying prediction uncertainties, we explicitly acknowledge that a digital soil map is not free from error. A major source of error is the sparseness of soil data, both in the landscape space and the attribute space. Often the uncertainty created by this error is of a magnitude that would preclude the use of a digital soil map in many situations where fine precision is a requirement. Consequently, decisions or policies developed based on mapping need to be made with a certain amount of risk, although the risk can often be quantified. To reduce this risk, the uncertainties may be used to prioritize resources for data collection or direct the application of alternative modelling approaches to improve digital soil maps and reduce uncertainty.

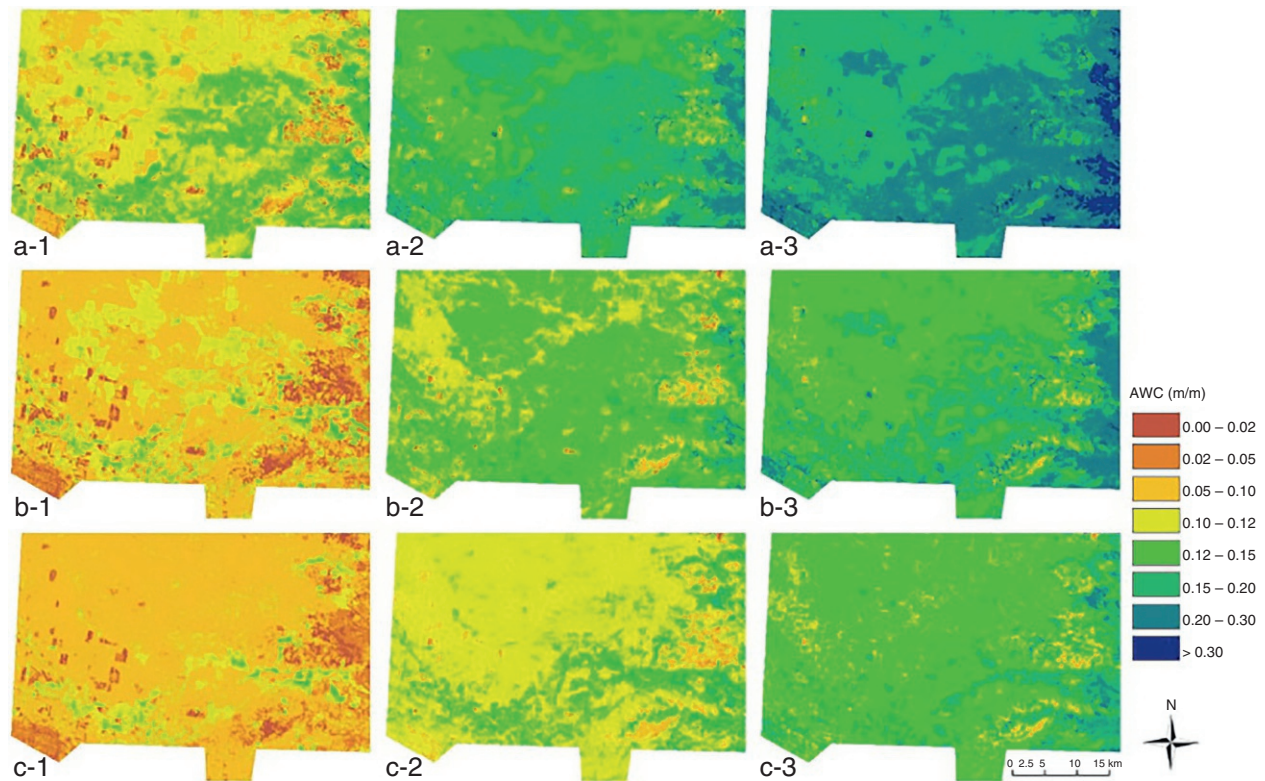


Fig. 1 Variability of AWC at 0–10 cm, 30–40 cm and 80–100 cm across the Edgeroi study area. Lower prediction limit (1), DSM final prediction (2), and upper prediction limit (3). This figure is an example of presenting digital soil map uncertainty in terms of a prediction interval with an expressed level of confidence (here 95%). These estimates are derived for every mapping pixel where one can relatively easily distinguish areas of high prediction certainty (lower prediction interval range) from low prediction certainty (high prediction interval range). From Malone BP, McBratney AB, and Minasny B (2011) Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. *Geoderma* 160: 614–626 and from <https://www.sciencedirect.com/science/article/pii/S0016706110003666>, Fig. 9.

Nuances of digital soil mapping

Knowledge-based inference and mechanistic modelling for digital soil mapping

An alternative to purely empirical modelling is knowledge-based inference. Knowledge-based inference enables the digital soil mapper to integrate expert knowledge into the mapping process. One of the best known knowledge-based tools is the Soil Land Inference Model or SoLIM (Zhu et al., 1997). SoLIM allows an expert to create membership functions manually that describe the presumed relationship between specific soil types and a range of environmental and topographic variables. With these membership functions, the expert can make predictions of soil type or properties at unobserved locations by a weighted estimate that is based on an environmental similarity score for each soil member. Subsequently, the appealing concept of fuzziness between soil objects is maintained, as well as some quantitative means for assessing the uncertainty of mapped predictions.

In similar work, Bui (2004) pointed out that the soil map legend, which is a representation of the distilled knowledge of the soil surveyors' mental model of soil variation across a mapping domain, contains valuable and important structured language that can be used for automating soil mapping if spatial cover of environmental predictors are available. Probably the closest empirical relative to this approach is the decision tree type model, where data are recursively partitioned to minimize some predictive variance.

The idea of utilizing existing soil maps in new, quantitative ways has also found application in the spatial disaggregation of soil map units, where the information contained in these map units can be downscaled to make spatial predictions of the map units' constituent soil types. This can be further enhanced by the incorporation of expert elicited rules about specific soil–landscape relationships that occur across the study area (e.g. Vincent et al., 2016).

Bayesian networks, described by Taalab et al. (2015), also enable the operator to include expert knowledge explicitly. They are also able to be tuned empirically. The methodology first requires the need to express prior knowledge, in the form of probabilities, about the interaction between a target variable and a set of environmental covariates. With these prior probabilities together with Bayesian inference, one can estimate either continuous or categorical variables and then use the derived posterior probabilities as a quantitative measure of uncertainty.

Structural equation modelling (SEM) has its roots in social sciences research and Angelini et al. (2016) have demonstrated its application to digital soil mapping. It enables the integration of empirical information with mechanistic knowledge by deriving the model equations from known causal relationships, while estimating the model parameters using the available data. This approach necessitates a thorough understanding of soil processes. Moreover, these models can predict many soil properties simultaneously, while preserving the relationships between them. The recognition that soil characteristics are not independent variables, which is how they are treated in most DSM model applications is a welcome advancement to DSM. In data science and machine learning, however, advances are continually being made and have been demonstrated by Wadoux (2019) who used a convolutional neural network model for simultaneous prediction of soil texture, organic carbon, pH, and nitrogen across France. This purely empirical model, had the advantage of quantifying prediction uncertainties, which is a general requirement for DSM and to date has not been investigated for SEM.

The future directions of DSM are likely to entail a deeper integration with soil pedology either by pure- or semi-mechanistic approaches like those described above. Ma et al. (2019) reviews the literature on this and provide insight on how both pedology and DSM can be enriched with knowledge and ideas from each science domain.

Space–time modelling

The spatio-temporal modelling of dynamic soil properties such as soil carbon and soil moisture (plant available water) necessitates an integration of process-based knowledge and empiricism and is at the frontier of DSM research. To date most spatio-temporal models of soil carbon are based around space-for-time substitution models (Gray and Bishop, 2019), where a projection of soil change because of climate or landuse change can be calculated readily. Dynamic modelling such as through RothC (soil carbon model) and biomass growth models for more explicit accounting of the biogeochemical processes underpinning soil organic carbon spatio-temporal variability have been trialled (Lee et al., 2020).

Similarly, a better understanding of the spatio-temporal controls of soil moisture variation for real time estimation and then its forecasting, and mechanistic understanding of soil–water dynamics is paramount. Ensemble Kalman filtering is one of the favored approaches for this type of work and its application in an agricultural setting has been demonstrated by Huang et al. (2017).

Soil depth functions and digital soil mapping

Soil survey, soil classification and conventional mapping of soils consider soil as a three-dimensional (3-D) entity or body. Most initial DSM research tended to focus only on the 2-D sense of this general concept where predictions of soil property variation were made for single depth intervals or horizons (and predominantly only from the topsoil).

It is important to attempt to map the variation of soil properties in both the lateral and vertical dimensions. For carbon accounting and understanding the carbon sequestration potential of soil, determining the amount of water soils can hold across a field or watershed, determining the depth to an impeding layer for crop growth across a farm, and investigations of soil acidity, will require some understanding of how soil properties vary with depth.

Ponce-Hernandez et al. (1986) in Malone et al. (2009) suggested that soil properties vary more-or-less continuously with depth. The variation is often anisotropic for certain properties such as carbon and soil texture which may be the result of land use activity or the gravitational vector of profile weathering and development or both. Exceptions to continuous soil property variation with depth is where there is strong anthropogenic (cultivation, removal, and replacement of soils), geologic (contrasting parent materials), and pedological (the development of clear and abrupt soil horizons) forcing for which sharp discontinuities in the depth distribution of soil properties will occur.

Empirical functions describing the depth distribution of soil properties include linear and polynomial functions, exponential and logarithmic functions. Asymmetric peak functions (Myers et al., 2011) and smoothing splines are further depth functions. A special case of the smoothing spline is the pycnophylactic (mass preserving) type as investigated by Bishop et al. (1999) in Malone et al. (2009). Mass preserving splines model the continuous variation of soil properties with depth while maintaining the average of the observed property through the observed horizons or layers. The proviso for a 'good fit' is that numerous observations at regular depths are required. For attributes such as soil organic carbon, it is necessary to have observations close to the surface to avoid over-estimating the contents in the very top few centimeters. Conversely, for some contaminated undisturbed soils with retained elements such as lead (Pb) observations close to the surface are necessary to avoid 'diluting' the effect of the large concentration at the surface.

The coupling of soil depth functions with DSM seems an intuitive advance toward understanding soil variation in all its spatial dimensions. A study by Minasny et al. (2006) reported in Malone et al. (2009) used the negative exponential depth function to describe variation in soil carbon concentration with depth in the Edgeroi area, Australia. The authors modelled the parameters of the exponential function using a modified neural network approach, then predicted parameters of the exponential function over the whole area, which enabled them to calculate the carbon distribution within the profile and the storage of carbon at any depth.

One of the limitations of using equivalents of the negative exponential depth function is that the function is only useful for certain soil properties, such as soil organic carbon, that naturally have that type of anisotropic variation down the soil profile. Polynomial soil depth functions also need to be considered carefully because the value at one depth will also affect the fit of the curve at other depths.

Malone et al. (2009) described a regression kriging approach using neural networks coupled with mass-preserving splines for mapping available water capacity and soil carbon stocks in Australia. The approach is performed in two steps with the first being the standardization of depth intervals given a collection of soil profile data with the spline. Integrating the spline for the specified depth intervals, for example 0–5 cm, 5–15 cm, 30–60 etc., an average value was defined for each depth for each soil profile. Not only are these values the predicted means for a specified depth interval, but they are also parameters of the spline function that can be used for subsequent refitting and soil information interrogation. After the splines are fitted, for each specified depth interval a spatial model is fitted which is then used for mapping. With such a coupling of depth function and DSM, Malone et al. (2009) were able to generate scenarios such as the depth at which the cumulative total of soil carbon was equal to 5 kg m^{-2} among other queries (Fig. 2).

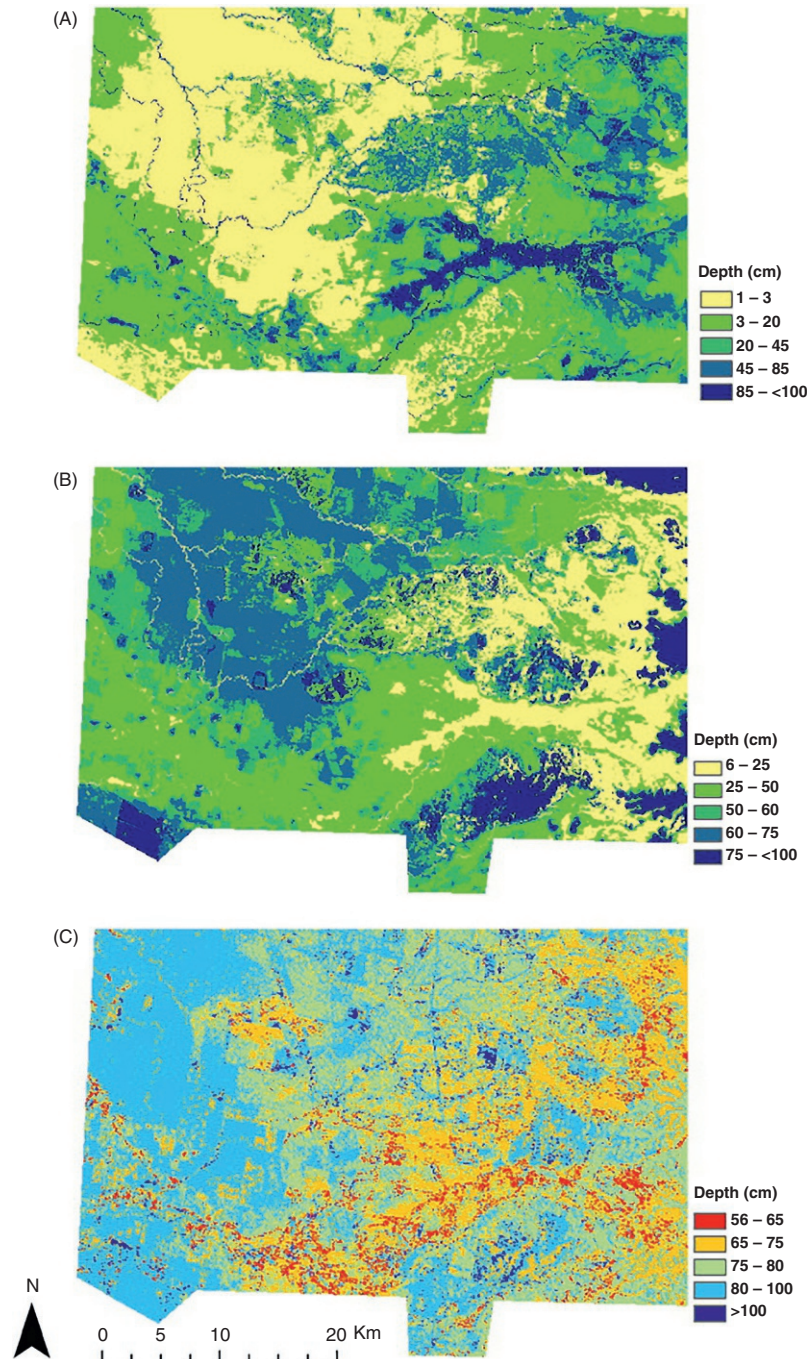


Fig. 2 Maps of scenario-based queries by Malone et al. (2009) in the Edgeroi district in northern New South Wales, Australia. (A) Depth at which soil carbon decreases to below 1%. (B) Depth at which cumulative total of soil carbon equals 5 kg m^{-2} . (C) Depth at which cumulative sum of AWC is 100 mm. From <https://www.sciencedirect.com/science/article/pii/S0016706109003255>, Fig. 9.

Poggio and Gimona (2014) describe a one-step approach to 3-D digital soil mapping using a hybrid generalized additive model (GAM) for carbon stocks in Scotland. One-step approaches are appealing because they reduce the complexity of workflow. With the Poggio and Gimona (2014) approach they avoided standardizing soil prediction depth intervals i.e. they used the observed values, and then modelled the 3-D trend in soil variation with a GAM and an associated 3-D smoother with related covariates. The one-step approach was also appealing to Orton et al. (2016), who introduced an area-to-point kriging methodology. This model could include all the data in one statistical analysis, and importantly maintained the integrity of the sample support of the soil profile data as well as providing an explicit methodology for quantifying the uncertainty of the fitted soil depth function.

Kempen et al. (2011) developed a depth function that combines general pedological knowledge with geostatistical modelling. They modelled the distribution of soil organic matter content based on typical horizons from 10 soil types. Five depth function building blocks were defined, and for each soil type the depth function structure was obtained by stacking a subset of modelled horizons. The parameters of the depth function for each of the horizons were interpolated using a geostatistical procedure that combined environmental information. While pedologically the most appealing approach, the horizon depth function method of Kempen et al. (2011) could be limiting because its application is limited to the spatial extent in which it was developed, or in similar landscape contexts or soil types.

Generalizing the practice of digital soil mapping

While most applications of DSM are based on using observed soil point data coupled with environmental covariates and a spatial predictive model function, it is certainly not restricted to this process. For example, some previous discussion of soil map disaggregation has been made where mapping units are disaggregated into their constituent classes or series whose spatial pattern is determined by a spatial prediction function.

Before exploring that concept in further detail, it is worth pointing out the various possible scenarios that could be encountered for DSM. They are described by Minasny and McBratney (2010) who presented a decision tree for DSM methodology based on the nature of available legacy soil data (Fig. 3). This tree was proposed as a general framework to aid the delivery of a digital global soil map. It can be viewed as guidance to practitioners in “what do I do?” situations. Once the practitioner has defined an area of interest and assembled a suite of environmental covariates for that area, depending on what available data there are to use, there are suggested approaches that could be implemented. The most common approach is scorpan kriging which can be performed when there are only point data, but can also be used when there are both point and map data available. The idea here is that scorpan

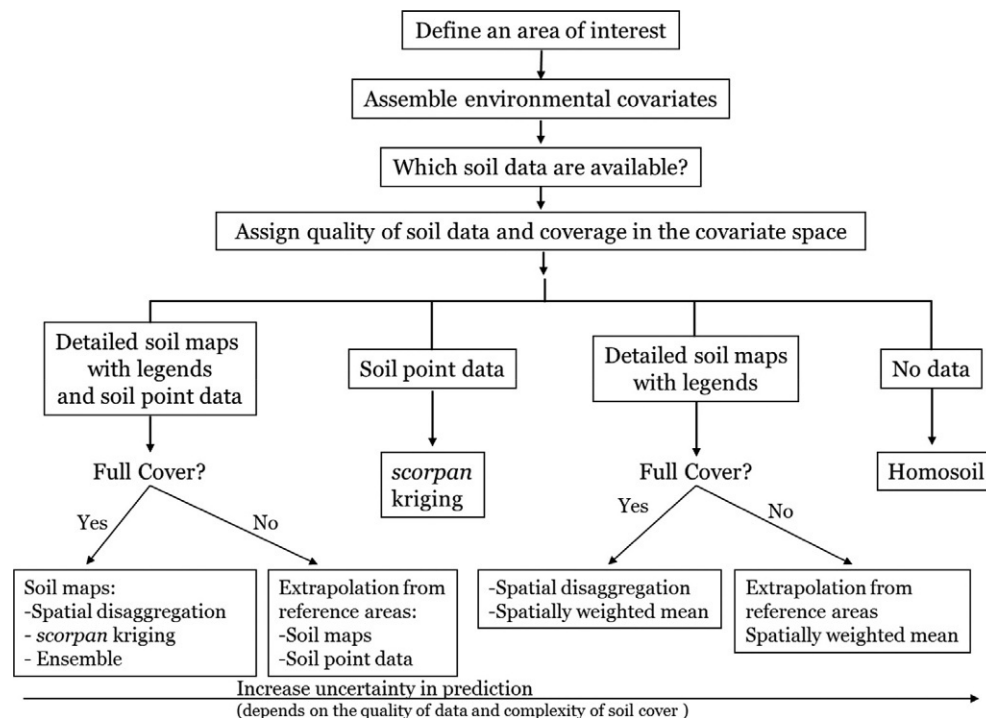


Fig. 3 Decision tree of digital soil mapping approaches. Adapted from Minasny B, and McBratney AB (2010) Methodologies for global soil mapping. In: Boettinger JL, Howell DW, More AC, Hartemink AE, and Kienast-Brown S (eds.) *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation*. London: Springer.

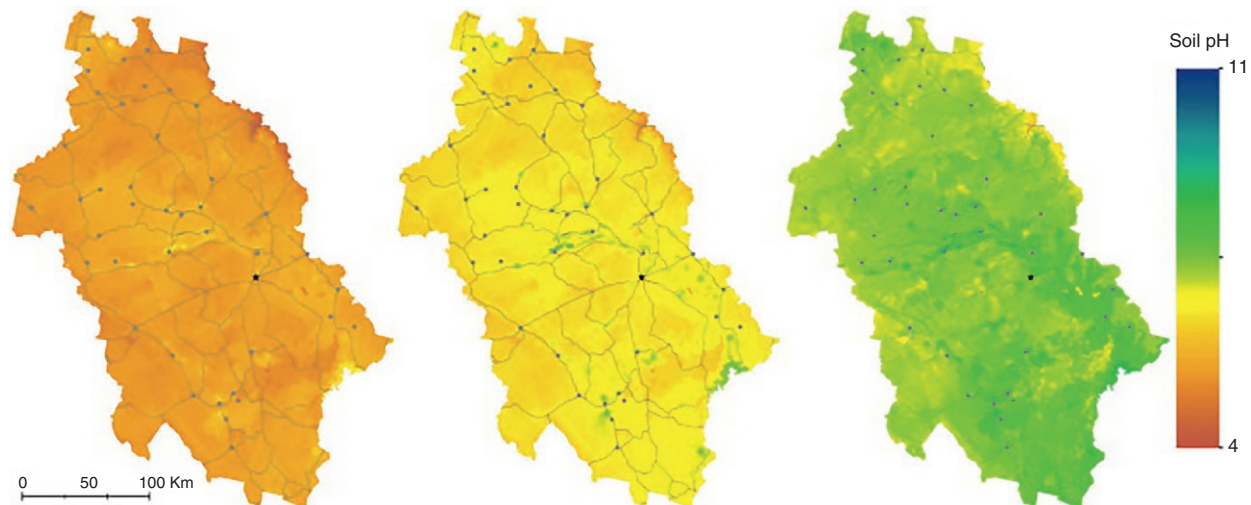


Fig. 4 Model averaging outputs from [Malone et al. \(2014\)](#). Model averaged digital soil maps of soil pH for the 0–5-cm depth interval across the Dalrymple Shire, QLD, Australia. These maps were produced using Granger–Ramanathan (GRA) model averaging, that both corrects bias and preferentially weights inputs based on comparative model accuracies. Inputs were a soil pH map with estimates from using point-based digital soil mapping (scorpan-kriging) and from disaggregated conventional map approaches (DSMART + PROPR) described in [Odgers et al. \(2014, 2015\)](#). The center map represents the combined prediction, while the left and right maps respectively are the 90% lower and upper prediction limits. From <https://www.sciencedirect.com/science/article/pii/S0016706114001906>, Fig. 5.

kriging and soil map disaggregation are fused through a combinatorial approach such as model averaging. [Malone et al. \(2014\)](#) exemplified such a procedure in an area of Queensland, Australia (Fig. 4).

The options are quite different when only soil map information is available. Consider that the quality of the soil maps depends on the scale and subsequent variation in soil cover; such that smaller scale maps e.g., 1:100,000 would be considered better and more detailed than large scale maps e.g. 1:500,000. The elemental basis for extracting soil properties from legacy soil maps comes from the central and distributional concepts of soil mapping units. For example, modal soil profile data of soil classes can be used to create soil property maps quickly. An early example of this is the multilayer soil characteristics dataset for the conterminous United States (CONUS-SOIL), which has been and continues to be used in many climate, hydrology and land surface models ([Miller and White, 1998](#) in [Odgers et al., 2014](#)). Where mapping units consist of more than one component, we can use a spatially weighted means type method, i.e. estimation of the soil properties is based on the modal profile of the components and the proportional area of the mapping unit each component covers. As a pre-processing step prior to creating soil attribute maps, it may be necessary to harmonize soil mapping units (in the case of adjacent soil maps) and or perform some type of disaggregation technique to retrieve the map unit's component information.

More recently, research into the spatial disaggregation of soil map units has emerged as a means of downscaling choropleth soil mapping. For example, the DSMART algorithm ([Odgers et al., 2014](#)) spatially disaggregates a choropleth map by iteratively resampling it to create a series of classification trees that create realizations of the potential soil class distribution. The realizations of soil class distribution are merged to estimate the probabilities of occurrence for all the soil classes in the choropleth map area. These probabilities together with modal soil profile characterizations associated with each soil class can then be integrated to derive soil attribute maps (such as pH, soil texture fractions etc.) as shown by [Odgers et al. \(2015\)](#). Modifications to DSMART by explicit inclusion of soil–landscape relationships, which assist and or constrain the allocation of likely soil types within a soil mapping unit, has been demonstrated with some success in the French region of Brittany ([Vincent et al., 2016](#)).

How do we do digital soil mapping in an area that does not have any soil data? In such areas ('recipient' areas) it may be possible to extrapolate using a model that was constructed in an area with available soil data (a 'donor' area). The rationale, called homosoil ([Mallavan et al., 2010](#)), is that if the recipient area is sufficiently homologous to the donor area with respect to the strength and expression of soil forming factors then extrapolation is feasible without an unacceptable degree of uncertainty. Identification of potential donor areas is based on computation of the similarity to the recipient area in terms of the relevant soil forming factors.

Operationalization of digital soil mapping

Since the end of the first decade of the 21st century, DSM has matured as a science and in some places worldwide is used as an operational tool for soil mapping agencies. In Australia for example, [Kidd et al. \(2020\)](#) describe the evolution of DSM from research undertakings within governmental agencies to a nationally coordinated program to deliver a [freely available, comprehensive National digital soil information system](#) and an active community of practitioners in both public and private funded organizations. Similar experiential pathways have been encountered elsewhere and new national, continental and [global](#) digital soils information facilities have also emerged similarly ([Chen et al., 2022](#)).

Not all operational work has focused on the development of digital soil maps *ab initio*. For example, government-sponsored work has led to the updating of legacy maps using digital techniques (Kempen et al., 2009). Agencies in other jurisdictions are working to integrate digital soil mapping methods with existing survey procedures. The relatively recent phenomena of national and global scale DSMs have been motivated through coordinated efforts by a series of international conferences and the GlobalSoilMap initiative (Arrouays et al., 2020a).

The GlobalSoilMap initiative

Soil mapping at the global extent is not a recent or new endeavor. The FAO–UNESCO soil map of the world was the first world soil map published (1981) that used a single map legend which accounted for the global diversity of soils. In around 2006 a proposal for a new global grid of the most important soil functional properties was made. Later that year, the [GlobalSoilMap.net](https://www.globalsoilmap.net) consortium was formed to create a new, high-resolution digital soil map of the world using state-of-the-art and emerging technologies. The consortium was inspired in part because of “policy-maker’s frustrations” about not being able to get quantitative answers to questions such as how much carbon is sequestered or emitted by soils in a particular region? Or, what is its impact on biomass production and human health? Or how do such estimates change over time?

Technical specifications of the GlobalSoilMap are given in the specifications document (Arrouays et al., 2014). This publication articulates not how digital soil maps should be created, but the standard to which they should conform to permit collation for the assemblage of a global product. This specification promotes innovation in digital soil mapping practice and recognizes that practitioners may prefer certain methods, and that different methods may perform better in different environments.

Key aspects of the specifications include the spatial entity such as the data support and resolution at which maps should be created, the soil properties to be predicted, the date associated with their prediction, and an explicit communication of the prediction uncertainty and accuracy. Other aspects include documentation standards and reproducibility, and the data release policy.

The initial organizational and collaborative structures brought about by the GlobalSoilMap initiative no longer exist, but it has left a lasting impact as exemplified by national soil mapping efforts, together with a large and active international research community which supports new, upcoming and seasoned scientists through training and education in the ‘how’ and ‘to’ of doing DSM (Kidd et al., 2020). In 2016, the International Union of Soil Sciences created a Working Group (WG) named “Global Soil Map” to maintain this active international community. The Pillar 4 “Soil Information” of the UN–FAO Global Soil Partnership actively promotes country-based DSM activities and now includes the above mentioned WG in its scientific committee (Arrouays et al., 2020a).

Uses and users of digital soil mapping

As outlined earlier, the GlobalSoilMap effort aimed to provide high-resolution soil attribute maps that can ultimately be used in a variety of applications. These maps can provide soil inputs (e.g. texture, organic carbon and soil-depth parameters) to models predicting land-cover changes in response to global climatic and human disturbances. Here, we will discuss a few examples.

Maps of more readily available soil properties can be used to estimate other more costly, or difficult to measure soil attributes. For example, Ballabio et al. (2016) used soil texture maps (%clay, %silt and %sand) created using topsoil data from the European Land Use and Coverage Area frame Survey (LUCAS) to derive maps of bulk density, USDA soil texture classes and available water capacity. These soil attributes may then be used to assess the C sequestration potential of topsoils (bulk density), assess the soil water holding potential of agricultural soils for informed soil management, or estimate soil compaction hazards (texture classes).

Similar research has been conducted in France to estimate potential C sequestration and C storage (Chen et al., 2018) and available water capacity (Román Dobarco et al., 2019). More recently, French GlobalSoilMap products were used in combination with modelling to assess the feasibility of reaching the 4 per Mille target of increasing C stocks in soils (Martin et al., 2021).

High resolution soil attribute maps can also be used for monitoring and forecasting biophysical properties. More explicitly, these soil attribute maps may be used as input soil information to drive physical systems simulation models (e.g. crop simulation models such as APSIM) or for land suitability assessment, or more broadly defined as digital soil assessment.

An example of the use of fine-resolution 3-D soil-attribute maps is land suitability assessment that contributed to the GlobalSoilMap effort in Australia (Kidd et al., 2015), and was derived for Tasmania, Australia. Here, digital soil assessment was used to provide information on the agricultural suitability of land for 20 crops (perennial horticultural, cereal and vegetable crops) in a new irrigation scheme ‘Water for Profit Program’ commissioned by the Tasmanian state government. Together with 3-D soil-attribute maps (80 m × 80 m), climate grids (e.g. growing degree-days and frost risk) were applied to a range of defined enterprise-suitability rulesets to produce enterprise suitability maps. All maps created are stored on a publicly available spatial internet portal (Land Information Services Tasmania, LISTmap <https://www.thelist.tas.gov.au/app/content/data#>) and can be used interactively to perform a range of land management based assessments, such as identifying limiting soil and climate conditions. Outputs also include an enterprise versatility index; and the highest-valued agricultural land with the highest earning potential for individual commodities can also be identified.

Soil security to maintain soil function in supporting food and fiber production together with ecosystem balance is widely reported to be under threat due to human impacts and population increase. Consequently, there is growing activity throughout the world to measure and monitor soil resources (Arrouays et al., 2021), or more specifically the functions and services that soils provide directly and indirectly (Adhikari and Hartemink, 2016 in Aitkenhead and Coull, 2019). Digital soil mapping and

assessment can provide comprehensive and objective assessments of attributes including soil carbon stocks, and assessment of the potential of soils to store more carbon, through to explicit assessment of soil ecosystem services (Aitkenhead and Coull, 2019). It is expected that practitioners of DSM will soon respond in better and more creative ways to end-user requests for soil and soil-related information for these more general assessments of ecosystem function, rather than concentrated practice of creating estimated soil property maps as an end in itself.

Conclusions

Digital soil mapping is an effective tool for producing nuanced soil information. This information can be used in various ways, from understanding the status and condition of soils in a highly granular way, to aid in the planning and deployment of strategic land management practices and policies.

Scientifically, the incorporation of mechanistic soil processes into digital soil modelling is an innovation which should be given greater focus. Recent work with soil carbon, water and texture have been important contributions, together with structural equation modelling which provide explicit characterization of soil processing into the model framework. Digital soil mapping should not just become a machine learning exercise, and there are clear examples and experiences of the pitfalls of relying solely on machine learning algorithms to do the process in a thoughtless manner, and without rudimentary quality control mechanisms (for example, does the spatial pattern of the predictions make sense in a soil–landscape pedological context) of the outputs that are generated (Arrouays et al., 2020b).

There will be a widening and deepening role of public–private partnerships in the development and application of digital soil mapping to address serious environmental issues. For this, the process of digital soil mapping should not be considered a static task, but rather viewed as a dynamic process where efforts are continually revisited, updated, and improved where possible. Efficient strategies need to be developed to ensure the best use of available funding resources for the most gain.

Ultimately, soils are our most valuable natural resource. Digital soil mapping and assessment provide the tools and objective abilities to characterize and monitor the status of this resource better. Digital soil mapping is playing its part and will continue to be engaged in efforts throughout the world to ensure that we live on this planet in a more sustainable way.

References

- Adhikari K and Hartemink AE (2016) Linking soils to ecosystem services — A global review. *Geoderma* 262: 101–111.
- Aitkenhead MJ and Coull MC (2019) Digital mapping of soil ecosystem services in Scotland using neural networks and relationship modelling. Part 2: Mapping of soil ecosystem services. *Soil Use and Management* 35: 217–231.
- Angelini ME, Heuvelink GBM, Kempen B, and Morrás HJM (2016) Mapping the soils of an Argentine Pampas region using structural equation modelling. *Geoderma* 281: 102–118.
- Arrouays D, McBratney A, Minasny B, Hempel J, Heuvelink G, Macmillan R, Hartemink A, Lagacherie P, and McKenzie N (2014) The GlobalSoilMap project specifications. In: Arrouays D, McKenzie NJ, Hempel J, De Forges AR, and McBratney AB (eds.) *GlobalSoilMap: Basis of the Global Spatial Soil Information System*. London: CRC Press.
- Arrouays D, Poggio L, Salazar Guerrero OA, and Mulder VL (2020a) Digital soil mapping and GlobalSoilMap. Main advances and ways forward. *Geoderma Regional* 21: e00265.
- Arrouays D, McBratney A, Bouma J, Libohova Z, Richer-De-Forges AC, Morgan CLS, Roudier P, Poggio L, and Mulder VL (2020b) Impressions of digital soil maps: The good, the not so good, and making them ever better. *Geoderma Regional* 20: e00255.
- Arrouays D, Mulder VL, and Richer-De-Forges AC (2021) Soil mapping, digital soil mapping and soil monitoring over large areas and the dimensions of soil security—A review. *Soil Security* 5: 100018.
- Ballabio C, Panagos P, and Monatanarella L (2016) Mapping topsoil physical properties at European scale using the LUCAS database. *Geoderma* 261: 110–123.
- Bishop TFA, McBratney AB, and Laslett GM (1999) Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma* 91: 27–45.
- Brus DJ, Kempen B, and Heuvelink GBM (2011) Sampling for validation of digital soil maps. *European Journal of Soil Science* 62: 394–407.
- Bui EN (2004) Soil survey as a knowledge system. *Geoderma* 120: 17–26.
- Chen S, Martin MP, Saby NPA, Walter C, Angers DA, and Arrouays D (2018) Fine resolution map of top- and subsoil carbon sequestration potential in France. *Science of the Total Environment* 630: 389–400.
- Chen S, Arrouays D, Mulder VL, Poggio L, Minasny B, Roudier P, Libohova Z, Lagacherie P, Shi Z, Hannam J, Meersmans J, Richer-De-Forges AC, and Walter C (2022) Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. *Geoderma* 409: 115567.
- Gray JM and Bishop TFA (2019) Mapping change in key soil properties due to climate change over south-eastern Australia. *Soil Research* 57: 467–481.
- Hastie T, Tibshirani R, and Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.
- Huang J, McBratney A, Minasny B, and Triantafyllis J (2017) Monitoring and modelling soil water dynamics using electromagnetic conductivity imaging and the ensemble Kalman filter. *Geoderma* 285: 76–93.
- Jenny H (1941) *Factors of Soil Formation, a System of Quantitative Pedology*. New York: McGraw-Hill.
- Kempen B, Brus DJ, Heuvelink GBM, and Stoorvogel JJ (2009) Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. *Geoderma* 151: 311–326.
- Kempen B, Brus DJ, and Stoorvogel JJ (2011) Three-dimensional mapping of soil organic matter content using soil type-specific depth functions. *Geoderma* 162: 107–123.
- Kempen B, Brus DJ, and Heuvelink GBM (2012) Soil type mapping using the generalised linear geostatistical model: A case study in a Dutch cultivated peatland. *Geoderma* 189–190: 540–553.
- Kidd D, Webb M, Malone B, Minasny B, and McBratney A (2015) Digital soil assessment of agricultural suitability, versatility and capital in Tasmania, Australia. *Geoderma Regional* 6: 7–21.
- Kidd D, Searle R, Grundy M, McBratney A, Robinson N, O'brien L, Zund P, Arrouays D, Thomas M, Padarian J, Jones E, Bennett JM, Minasny B, Holmes K, Malone BP, Liddicoat C, Meier EA, Stockmann U, Wilson P, Wilford J, Payne J, Ringrose-Voase A, Slater B, Odgers N, Gray J, Van Gool D, Andrews K, Harms B, Stower L, and Triantafyllis J (2020) Operationalising digital soil mapping—Lessons from Australia. *Geoderma Regional* 23: e00335.
- Lark RM, Cullis BR, and Welham SJ (2006) On spatial prediction of soil properties in the presence of a spatial trend: The empirical best linear unbiased predictor (E-BLUP) with REML. *European Journal of Soil Science* 57: 787–799.

- Lee J, Viscarra Rossel RA, Luo Z, and Wang YP (2020) Simulation of soil carbon dynamics in Australia under a framework that better connects spatially explicit data with Roth C. *Biogeosciences Discussions* 2020: 1–24.
- Ma Y, Minasny B, Malone BP, and McBratney AB (2019) Pedology and digital soil mapping (DSM). *European Journal of Soil Science* 70: 216–235.
- Mallavan BP, Minasny B, and McBratney AB (2010) Homosol, a methodology for quantitative extrapolation of soil information across the globe. In: Boettinger JL, Howell DW, Moore AC, Hartemink AE, and Kienast-Brown S (eds.) *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation*. Dordrecht: Springer Netherlands.
- Malone BP, McBratney AB, Minasny B, and Laslett GM (2009) Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma* 154: 138–152.
- Malone BP, McBratney AB, and Minasny B (2011) Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. *Geoderma* 160: 614–626.
- Malone B, Minasny B, Odgers N, and McBratney A (2014) Using model averaging to combine soil property rasters from legacy soil maps and from point data. *Geoderma* 232–234: 34–44.
- Martin MP, Dimassi B, Román Dobarco M, Guenet B, Arrouays D, Angers DA, Blache F, Huard F, Soussana J-F, and Pellerin S (2021) Feasibility of the 4 per 1000 aspirational target for soil carbon: A case study for France. *Global Change Biology* 27: 2458–2477.
- McBratney AB, Mendonça-Santos ML, and Minasny B (2003) On digital soil mapping. *Geoderma* 117: 3–52.
- Miller DA and White RA (1998) A conterminous United States multilayer soil characteristics dataset for regional climate and hydrological modeling. *Earth Interactions* 2: 1–26.
- Minasny B and McBratney AB (2010) Methodologies for global soil mapping. In: Boettinger JL, Howell DW, More AC, Hartemink AE, and Kienast-Brown S (eds.) *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation*. London: Springer.
- Minasny B and McBratney AB (2016) Digital soil mapping: A brief history and some lessons. *Geoderma* 264(Part B): 301–311.
- Minasny B, McBratney AB, Mendonça-Santos ML, Odeh IOA, and Guyon B (2006) Prediction and digital mapping of soil carbon storage in the Lower Namoi Valley. *Australian Journal of Soil Research* 44: 233–244.
- Minasny B, Vrugt JA, and McBratney AB (2011) Confronting uncertainty in model-based geostatistics using Markov Chain Monte Carlo simulation. *Geoderma* 163: 150–162.
- Myers DB, Kitchen NR, Sudduth KA, Miles RJ, Sadler EJ, and Grunwald S (2011) Peak functions for modeling high resolution soil profile data. *Geoderma* 166: 74–83.
- Odgers NP, Sun W, McBratney AB, Minasny B, and Clifford D (2014) Disaggregating and harmonising soil map units through resampled classification trees. *Geoderma* 214–215: 91–100.
- Odgers NP, McBratney AB, and Minasny B (2015) Digital soil property mapping and uncertainty estimation using soil class probability rasters. *Geoderma* 237–238: 190–198.
- Orton TG, Pringle MJ, and Bishop TFA (2016) A one-step approach for modelling and mapping soil properties based on profile data sampled over varying depth intervals. *Geoderma* 262: 174–186.
- Padarian J, Minasny B, and McBratney AB (2019) Using deep learning for digital soil mapping. *The Soil* 5: 79–89.
- Poggio L and Gimona A (2014) National scale 3D modelling of soil organic carbon stocks with uncertainty propagation—An example from Scotland. *Geoderma* 232–234: 284–299.
- Ponce-Hernandez R, Marriott FHC, and Beckett PHT (1986) An improved method for reconstructing a soil-profile from analysis of a small number of samples. *Journal of Soil Science* 37: 455–467.
- Román Dobarco M, Bourenane H, Arrouays D, Saby NPA, Cousin I, and Martin MP (2019) Uncertainty assessment of GlobalSoilMap soil available water capacity products: A French case study. *Geoderma* 344: 14–30.
- Scull P, Franklin J, Chadwick OA, and McArthur D (2003) Predictive soil mapping: A review. *Progress in Physical Geography* 27: 171–197.
- Searle R, McBratney A, Grundy M, Kidd D, Malone B, Arrouays D, Stockman U, Zund P, Wilson P, Wilford J, Van Gool D, Triantafyllis J, Thomas M, Stower L, Slater B, Robinson N, Ringrose-Voase A, Padarian J, Payne J, Orton T, Odgers N, O'Brien L, Minasny B, Bennett JM, Liddicoat C, Jones E, Holmes K, Harms B, Gray J, Bui E, and Andrews K (2021) Digital soil mapping and assessment for Australia and beyond: A propitious future. *Geoderma Regional* 24: e00359.
- Taalab K, Corstanje R, Zawadzka J, Mayr T, Whelan MJ, Hannam JA, and Creamer R (2015) On the application of Bayesian networks in digital soil mapping. *Geoderma* 259–260: 134–148.
- Vaysse K and Lagacherie P (2017) Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma* 291: 55–64.
- Vincent S, Lemerrier B, Berthier L, and Walter C (2016) Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. *Geoderma* 311: 130–142.
- Wadoux AMJC (2019) Using deep learning for multivariate mapping of soil with quantified uncertainty. *Geoderma* 351: 59–70.
- Zhu AX, Band L, Vertessy R, and Dutton B (1997) Derivation of soil properties using a soil land inference model (SoLIM). *Soil Science Society of America Journal* 61: 523–533.

Geographical information systems (GIS) and soils

S Grunwald^a and J Böhner^b, ^aPedometrics, Landscape Analysis, and GIS Laboratory, Department of Soil, Water, and Ecosystem Sciences, University of Florida, Gainesville, FL, United States; ^bResearch Section Physical Geography, Institute of Geography, Center for Earth System Research and Sustainability (CEN), Universität Hamburg, Hamburg, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	696
Soil-environmental geospatial data acquisition	697
GIS-supported digital soil mapping and modeling	705
Role of GIS in pedometrics, simulation models, and decision support systems	706
Conclusions	708
Acknowledgments	708
References	708

Abstract

Geographic information systems (GIS) are essential to facilitate digital soil mapping (DSM) and soil modeling (pedometrics) and have moved from personal desktop, client-server, to cloud-based geodata and geoprocessing engines. Soil information systems disseminate geographic soil data as vector-based map units or grids. Artificial intelligence (AI) geospatial soil models provide excellent predictive capabilities, though AI cloud-based GIS and geotechnologies have raised numerous ethical concerns. Soil and environmental geodata support pedometrics applications and soil as well as land and Earth simulation models to assess soil health, soil security, soil ecosystem services, soil erosion, global climate change and carbon sequestration, and other broad topics of global importance.

Key points

- Geographic information systems (GIS) and digital soil mapping (DSM) are closely linked.
- GIS have moved from personal desktop, client-server, to cloud-based geodata and geoprocessing engines.
- Soil information systems provide geographic soil data associated with vector-based map units or grids and soil profile data.
- Artificial intelligence (AI) soil models use machine learning and deep learning algorithms that are often superior in terms of predictive capabilities and accuracies compared to more traditional quantitative soil models derived by geostatistics, regressions, or hybrid approaches.
- Ethical concerns related to geodata privacy and AI geospatial soil models urgently need to be addressed given the fast development in geo- and agro-technologies and AI.
- Soil-environmental geodata support pedometrics applications and soil/land and Earth simulation models to assess soil health, soil security, soil ecosystem service, food security, global climate change and carbon sequestration, and other broad topics of global importance.

Introduction

The term ‘geographic information system’ (GIS) has become synonymous with a wide range of computer applications, technologies, and scientific methods related to the use of geospatial information. The origin of GIS can be traced back to early electronic mapping tools, such as the Canada Geographic Information System (CGIS) developed in the mid-1960s. However, limiting the definition of GIS to a spatial visualization tool would be to disregard its current scientific and technological capabilities, as well as the huge expansion of GIS applications in science, technology, engineering and mathematics (STEM) disciplines, public administration, and economics. A contemporary definition of GIS was provided by Bolstad (2019) stating that a GIS is a computer-based system to aid in the collection, maintenance, storage, analysis, output, and distribution of spatial data and information. The GIS system may reside on a personal computer, be implemented in the form of client-server architecture, or completely reside in the cloud (internet). Driven by the rapid growth of computer technologies on the one hand and globally increasing demands for spatial information on the other, GIS has evolved more recently from a digital mapping facility into a fast-growing spatial science technology. The amalgamation of computer-assisted cartography with database technologies and growing capabilities of analyzing data across different layers in an object-oriented programming environment has generated current GIS management technology for capturing, analyzing, and modeling spatial data. Such GIS platforms facilitate the study of geospatial environmental problems such as soil degradation, global climate change, soil carbon sequestration, soil health, soil security, and food security. Specifically, GIS has become an indispensable tool for quantitative applications in soil science, specifically in the disciplines of pedometrics and digital soil mapping (DSM).

Geospatial technologies have undergone a revolution in development in terms of data collection, operational management, mobility of public sharing of geospatial data, and interconnected GIS workspaces (Fig. 1). Advancements in high-precision and high-accuracy geo-surveying of soil and environmental data, specifically global positioning systems (GPS) that are used for geo-referencing of data (i.e., the identification of geographic coordinates such as latitude and longitude, or x and y coordinates) and satellite technology, have amplified the trend towards finer spatial resolutions of geodata products. The management of soil-environmental data collected through geotechnologies is facilitated by GIS that enables the storage and processing of vector data—including (i) point-specific data such as soil attributes for different genetic horizons or layers (Fig. 2a), (ii) pedon or site-specific data (Fig. 2b) as well as (iii) polygons such as soil map units (Fig. 2c), and (iv) voxels such as 3-D soil layers (Fig. 2d)—and raster (grid) data to capture soil properties or soil taxonomic variables. An example of gridded soil organic carbon (SOC) predictions for Florida, United States are shown in Fig. 3. Soil grids have become more common due to the ease, speed, and versatility of raster-based GIS analysis. Soil-environmental data collected as part of research projects are typically stored on personal geodata servers (PCs or organizational servers), while publicly shared geospatial data are stored and disseminated in online geodatabases or geoportals in the cloud. Examples of global geoportals are the Geospatial Google Earth Engine, Amazon Geospatial and Amazon Web Services (AWS), and the Environmental System Research Institute (ESRI) ArcGIS Online Open Data.

Earlier GIS databases and processing applications adopted a computer client architecture using commercial GIS software such as ESRI's ArcGIS Desktop, IDRISI, Earth Resources Data Analysis (ERDAS), MANIFOLD, and multiple open GIS such as the open source QGIS and the freely available System for Automated Geoscientific Analyses (SAGA) GIS. While earlier and some contemporary GIS applications run on PCs the trend is to move geoprocessing into the cloud in which an increasingly amount of geodata already reside. Examples for global cloud-geospatial processing engines are the Google Earth Engine, Amazon Elastic Cloud Compute Services (Amazon EC2), and ArcGIS Online Data. Importantly, these types of cloud-based platforms combine versatility, security, and computational speed to support advanced artificial intelligence (AI) algorithms—machine and deep learning—in support of pedometrics and DSM. Cloud-based GIS portals and processing engines also offer GIS web services and map publishing applications. Hence, the public sharing of digital soil map products and soil models facilitated by the cloud has accelerated the sharing of geospatial soil information with end-users and stakeholders. The coding of soil models in open source R and Python languages to assess soil properties, indices, services, or functions has further contributed to popularize the movement of GIS into the cloud.

The exponentially growing markets of smart phones and mobile devices have spawned the development of mobile GIS applications and mobile soil tools and 'apps,' for example the SoilWeb and mySoil apps (Fig. 1). According to Herrick et al. (2017), the benefits of mobile soil apps, such as the Land-Potential Knowledge System (LandPKS), are manifold. First, these mobile technologies make it easier to collect, store, access, integrate, and interpret the soil and land inventory and environmental monitoring data from a geographic perspective. Second, the apps provide in-built applications developed by experts that help users with little or no soils knowledge to map and interpret soils. Third, the tools and apps allow users to share and compare their data easily with others. Fourth, informed decision-making and optimal soil/land management is supported by AI and other quantitative model algorithms that are built into the mobile apps and tools. These decision-support tools often connect with larger public soil-environmental geodata sources that have accumulated in the cloud and are likely to increase as new data are continuously acquired through monitoring. This shift from personal data acquisition and privacy of data towards public and open data policies are advantageous for globalizing soil-environmental geodata and improving 'smart' decision-making on how to use and manage soils and the land efficiently to meet sustainability and food and soil security goals.

Ethics and privacy issues with regard to the public sharing of geospatial data and use of GIS have sparked a critical debate, specifically when public cadastral and land owner data are involved (Blatt, 2012). The misappropriation of private landowners' data (e.g., farm or ranch data) that are shared as public data are prone to misuse by corporate industries. Privacy and data security concerns also raise questions when soil and land data are used to overregulate common good resources, for example, by imposing environmental management regulations, taxes, or fines on landowners. Another ethical concern about open multi-user shared geospatial soil-environmental data and models pertains to the lack of protective data sharing policies of global as well as national institutions and global researcher groups that have tended to incorporate local and regional soil-environmental data and models without providing appropriate credit, acknowledgement, or compensation. This is particularly unfortunate in the soil science domain where manual labor, time, and costs are high for field data collection, investments in instrumentation and technology, laboratory analyses of soil samples, and data entry effort.

In the following sections we will explore GIS applications to soils focusing in more depth on (1) soil-environmental geospatial data acquisition, (2) GIS-supported digital soil mapping, and (3) the role of GIS in pedometrics, simulation models, and decision support systems.

Soil-environmental geospatial data acquisition

The conceptual frameworks underlying pedometrics and DSM have been rooted in soil-factorial models that relate soils and soil-environmental factors, assess the spatial variation of soils within a region, or simulate pedogenic and ecosystem processes (Grunwald, 2006). These contemporary quantitative soil-factorial models are rooted in Jenny's and Dokuchaev's mental models of soil formation that link soils to state factors, such as climate, organisms and vegetation, relief, and parent material. For all types of soil-landscape model it is essential to acquire soil-environmental data for a region or area of interest (AOI). Typically, these

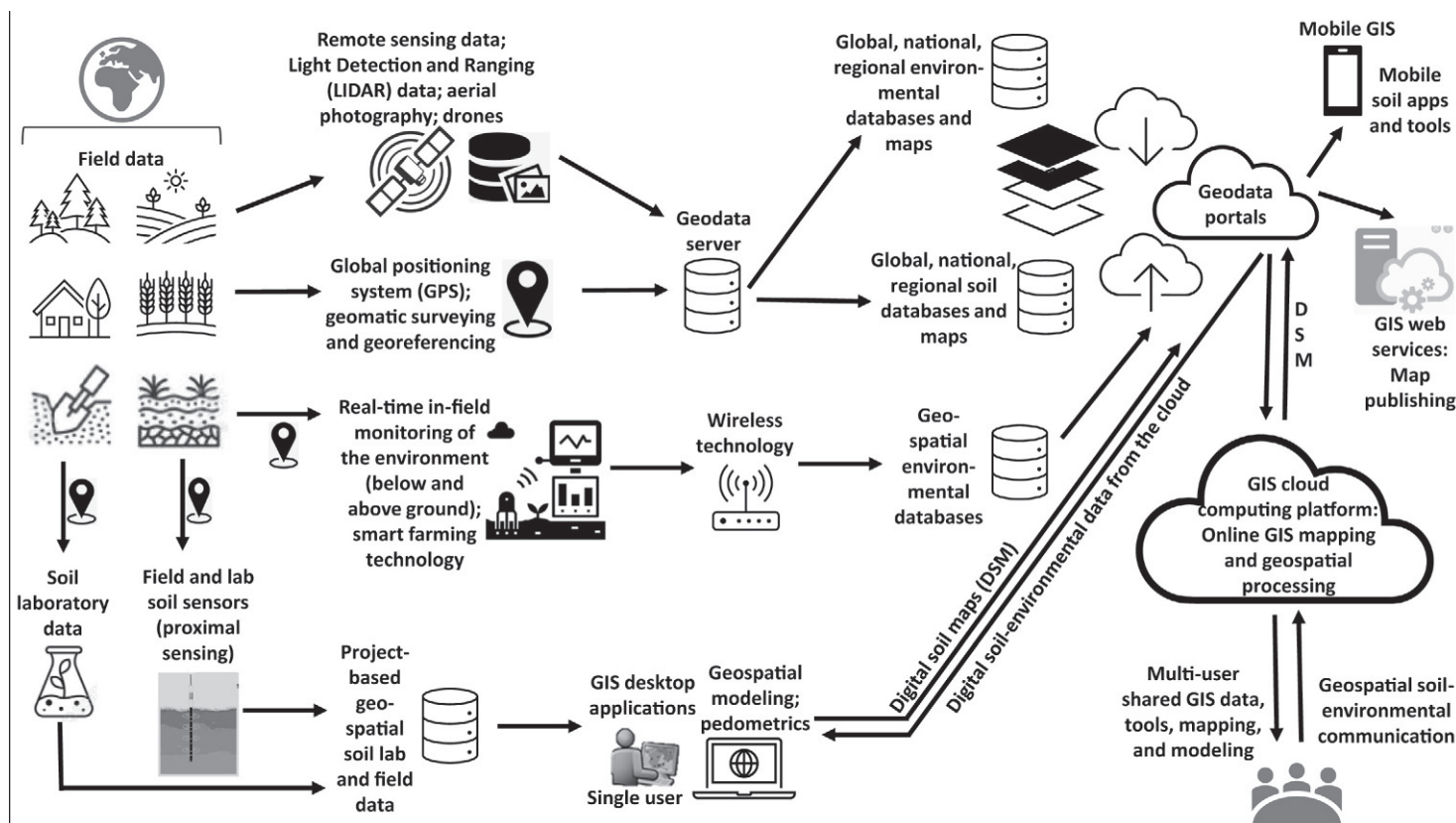


Fig. 1 Geospatial technologies used in soil science covering the spectrum from field data collection in the soil-environmental systems to the publishing and sharing of maps.

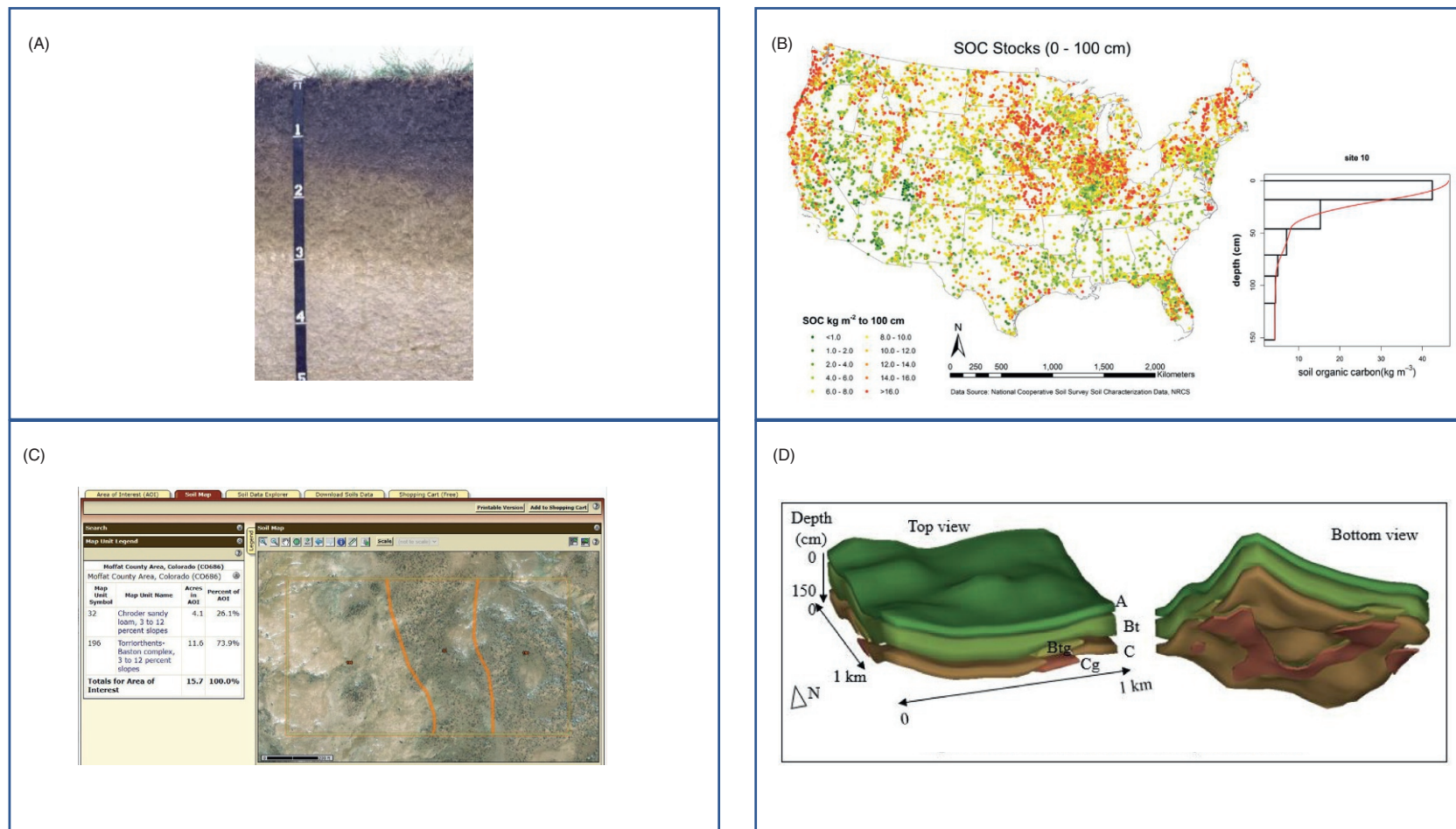


Fig. 2 Geographic representations of soils in the form of (a) site-specific pedons with genetic horizons or fixed-depth layers at specific sites, (b) continuous soil data along vertical profiles derived by splines or other interpolation methods, (c) soil taxonomic or soil property polygon maps with map units providing discrete boundaries, and (d) 3-D soil representations. Image, data, and source credits: (a) Houston soil profile, Very-fine, smectitic, thermic Oxyaquic Hapluderts. Natural Resources Conservation Service, United States Department of Agriculture. Public pedon photo gallery: https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/office/ssr7/profile/?cid=nrcs142p2_047981; (b) The map shows soil organic carbon (SOC) stock at 0–100-cm depth derived for 15,436 soil profiles in the contiguous United States of America. The SOC stocks for each profile were derived by splines with the equal-area smoothing spline function using horizon-based SOC and bulk density measurements. The graph on the left shows a fitted spline derived from SOC horizon measurements for a randomly selected site of the whole dataset (courtesy of S. Grunwald with contribution by B. Cao). Soil data source: Natural Resources Conservation Service, United States Department of Agriculture; (c) Web Soil Survey provides soil taxonomic data for soil map units with discrete boundaries for the United States of America. Available at: <https://websoilsurvey.sc.egov.usda.gov/App/WebSoilSurvey.aspx>; (d) 3-D soil model for a site in southern Wisconsin characterized by silty loam formed in loess overlaying glacial till parent material (courtesy by S. Grunwald; see: Grunwald S and Barak P (2001) The use of virtual reality modeling language (VRML) for virtual soil landscape modeling. *Systems Analysis Modelling Simulation* 41: 755–776).

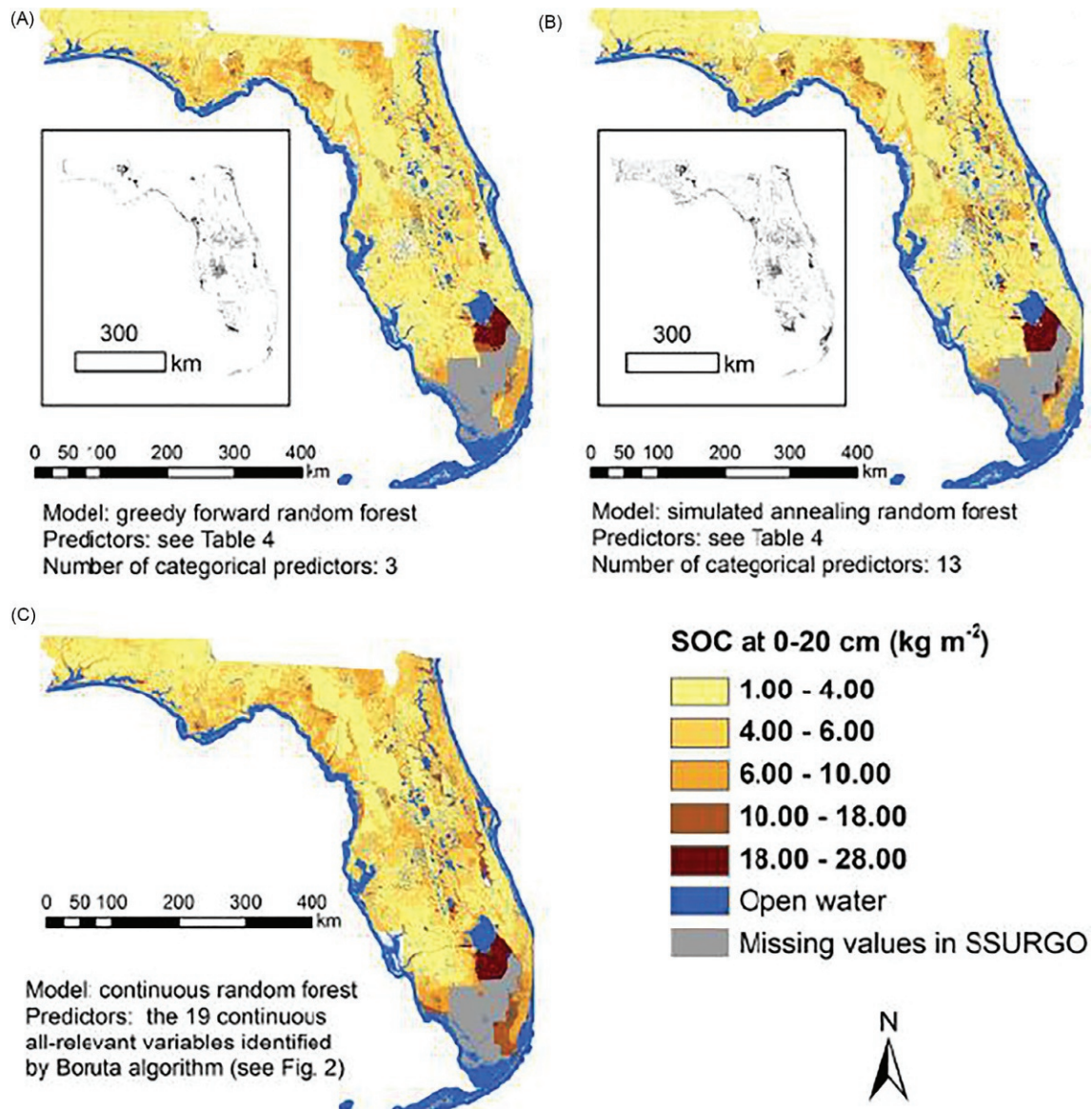


Fig. 3 Soil organic carbon (SOC) raster maps at 0–20 cm depth for Florida, United States produced by parsimonious models (a) greedy forward random forest model, (b) simulated annealing random forest model, (c) random forest model with all of the 19 continuous all-relevant variables identified by the Boruta algorithm. All the GIS layers of predictors were of the same extent and free from missing values, except the ones from the Soil Survey Geographic Database (SSURGO) which had a large data gap in southern Florida (the gray area). Reprint from Xiong X, Grunwald S, Myers DB, Kim J, Harris WG and Comerford NB (2014) Holistic environmental soil-landscape modeling of soil organic carbon. *Environmental Modelling & Software* 57: 202–215. <https://doi.org/10.1016/j.envsoft.2014.03.004>.

soil-environmental data are harmonized into a spatio-temporal data hypercube by a wide variety of geospatial technologies, remote sensing, soil surveying, proximal soil sensing, and available public geoportals (Fig. 1).

Prominent spatially and temporally explicit soil-factorial modeling frameworks that use soil-environmental hypercubes to predict or model soils are the SCORPAN model (McBratney et al., 2003; McBratney et al., 2018) and the STEP-AWBH model (Grunwald et al., 2011; Grunwald, 2021; Ma et al., 2019; Thompson et al., 2012). These modeling frameworks consider soil as an integral component of the ecosystem that can be characterized by different spatio-temporal factors (Fig. 4): (1) STEP-factors that are relatively stable across the human lifetime: Soil (S), Topography (T), Ecology (E), Parent material (P); and (2) AWBH-factors that are dynamic in space and across time: Atmosphere/climate (A), Water/hydrosphere (W), Biosphere (B), and Human (H) factors. Each of these factors can be quantified through a set of variables. For example, the S factor can be characterized by soil data such as SOC, soil texture, pH, soil taxonomic class, cation exchange capacity, etc. derived from legacy soil maps or databases, gamma ray sensing, and remotely sensed soil moisture data. The factor A may be populated by climatic data such as long-term average of mean annual precipitation, seasonal variation of minimum and maximum temperature, and long-term solar radiation, while B could be populated by satellite-derived land use/land cover maps, vegetation indices like the normalized difference vegetation index (NDVI) derived from satellite data, biodiversity, and habitat data. The factor H may be populated by variables from the social, cultural,

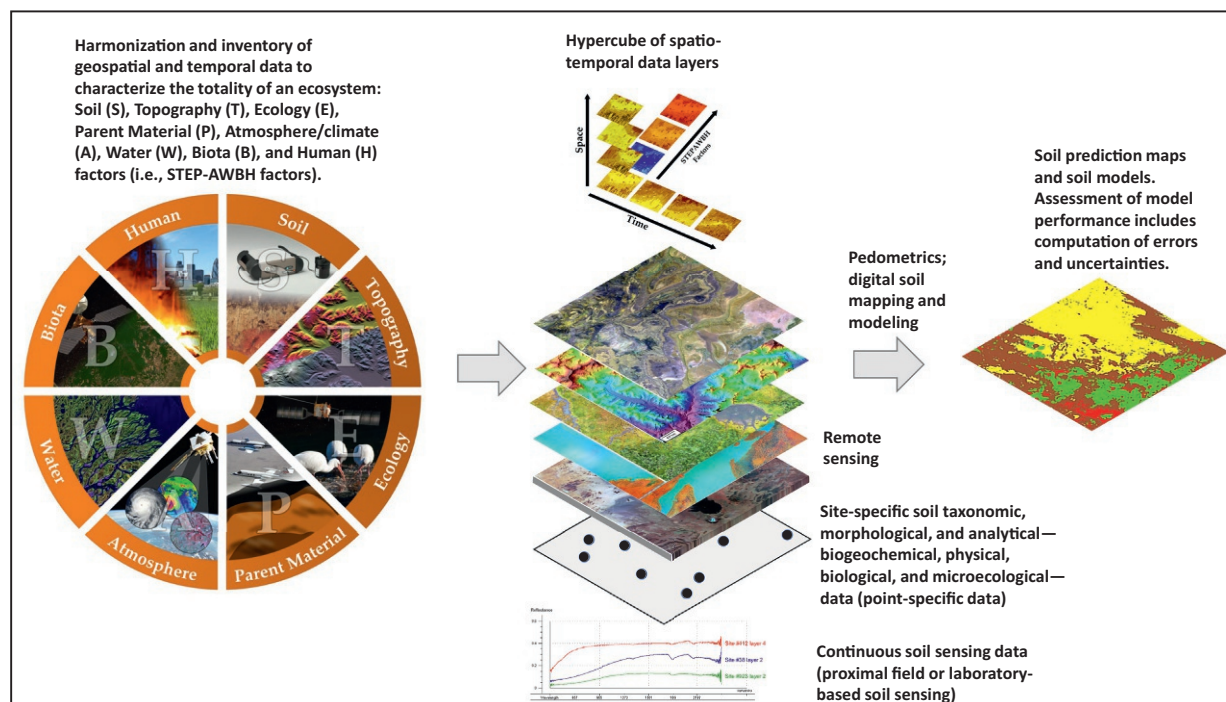


Fig. 4 Process in digital soil mapping to develop soil prediction maps and models from a hypercube of soil-environmental geodata. The soil-factorial model called STEP-AWBH (described in detail in [Grunwald et al., 2011](#) and [Grunwald, 2021](#)) provides a spatially and temporally explicit framework to characterize the STEP-AWBH factors: Soil (S), Topography (T), Ecology (E), Parent Material (P), Atmosphere/climate (A), Water (W), Biota (B), and Human (H) factors.

economic, and political domains (e.g., greenhouse gas emission data, land management data, and fertilization data). The aim is to populate the STEP-AWBH factors with soil-environmental geodata that influence pedogenesis; and thus, these factors contribute to soil formation in varying degrees. Xiong et al. (2014) exemplified the STEP-AWBH model and multiple AI-machine learning methods in Florida, United States, to develop prediction models for SOC.

The GIS are essential to populate and verify soil-factorial models, geospatial and geostatistical soil models, and (quasi) mechanistic land and Earth system simulation models to store, manage, and harmonize geospatial soil-environmental data. The observation methods described in the following paragraphs are not exhaustive but provide a brief overview of widely used sources to populate soil models.

Aerial photography, which emerged in the 1920s, is still in use for mapping planetary surface features, specifically vegetation; however, aerial imaging has been complemented and stands in competition with satellite technology and drones. The latter enable the collection of targeted ad-hoc high-resolution field data to assess soil deterioration or soil-vegetation stresses caused by multi-hazard disasters, global climate change, or others. The advent of soil and environmental data collected with a variety of fine (submeter to 5 m), medium (5–30 m), and coarse (30 m to 1 km+) spatial resolution satellite images have propelled DSM to produce local, regional, national, and global soil maps and models. There are various satellites that provide global multi-temporal data sets, for example, the Moderate Resolution Imaging Spectroradiometer (MODIS) and Landsat Enhanced Thematic Mapper (LANDSAT ETM+) by the North American Space Agency (NASA), SENTINEL images provided by the European Space Agency, Satellite Pour l'Observation de la Terre (SPOT) by the French Space Agency (Centre National D'Etudes Spatiales, CNES), and RESOURCESAT by the Indian Space Resource Organization (ISRO). Other fine-resolution commercial satellites are GeoEye, RapidEye, WorldView, and Pleiades. These passive remote sensors allow sensing of vegetation, terrain, and the thin layer of surface soil, while active remote sensors, such as RADARSAT with measurements in the microwave region and gamma-ray sensing that can measure the composition of soils and parent material, have greatly enhanced DSM. Satellite-based mapping of soil moisture has been conducted by the Soil Moisture and Ocean Salinity (SMOS) mission, NASA's Soil Moisture Active Passive (SMAP), and others. Light Detection and Ranging (LiDAR) is a remote sensing method that uses light in the form of a pulsed laser to measure ranges (variable distances) to the Earth and has been used extensively to create fine-resolution digital elevation models (DEM), some with sub-meter spatial resolution. Other global DEMs were developed by Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) satellite data, the Shuttle Radar Topographic Mission (SRTM), the high-resolution DEM of the European Copernicus program (GLO-30 and GLO-90), and the WorldDEM Neo at 5 m pixel spacing. A variety of primary and secondary terrain properties, for example, the terrain wetness index or mass balance index (see Fig. 5), can be derived from DEMs and are valuable input data for DSM (Conrad et al., 2015). Fig. 5 exemplifies the mass balance index for the parameterization of slope transfer processes, integrating the universal soil-loss equation (USLE) R-factor and different catchment area parameters (catchment area, catchment slope). Since terrain parameters based on catchment area size tend towards strongly artificial variability in hydrologically homogenous areas such as flat valley bottoms (cf. Fig. 5a), the SAGA (System for Automated Geoscientific Analyses) terrain wetness index (Fig. 5b) applies a slope-dependent, iterative modification scheme to the catchment area size which makes it suitable for the spatialization of hydromorphic soil properties (Conrad et al., 2015).

Traditionally, soil surveying entails soil observations collected with augers from genetic horizons, fixed depth layers, soil pits, or soil cores that are geo-located by GPS (Fig. 2a and b). Soil morphological and taxonomic soil characterization can be completed in the field and data recorded immediately into a geodatabase using portable devices or alternatively field notes are transferred into a geodatabase at a later time. The 'gold standard' of soil survey campaigns is to analyze samples in the laboratory for morphological, physical, chemical, biological, and microecological properties that are subsequently integrated into a geodatabase. Pedon geodatabases, such as the one developed by the Soil Research Laboratory, U.S. Natural Resources Conservation Service for the United States and ISRIC's World Soil Database, provide public access to geo-referenced soil data tabulated for different horizons (Fig. 2a and b).

Field and laboratory soil sensing based on diffuse reflectance spectroscopy, such as visible-near-infrared (VNIR) and mid-infrared spectroscopy (MIR), are now widely used for cost-effective and rapid characterization of the physical, chemical, and mineralogical composition of soils (Viscarra Rossel et al., 2011). An example of the capacity for spectral discrimination among different soils is shown in Fig. 6a. Other in-situ proximal soil sensing methods entail infrared, X-ray fluorescence sensing, field-based gamma-ray sensing, and electromagnetic induction (e.g., Fig. 6b). Hyperspectral datasets from various proximal soil sensors can be integrated into a GIS soil information system. Combining multiple soil sensor data is often beneficial to enhance soil predictive performance on new soil samples or in unsampled locations. Importantly, soil sensors collect continuous data that allow inference on multiple soil properties simultaneously. For example, laboratory grade VNIR soil sensor (e.g., the ASD spectroradiometer range by Malvern Panalytical or the Spectral Evolution spectroradiometer range) spectral data cover the spectral range from about 350 to 2500 nm with 1 nm resolution. Soil spectral models developed by machine learning and deep learning AI algorithms can predict multiple soil properties from the same soil scans. For example, a large spectral library of VNIR spectra was used to predict SOC, pH, cation exchange capacity, CaCO_3 , Fe, clay, sand, and silt content at the global scale (Viscarra Rossel et al., 2016). Soil sensors have become smaller, faster, more accurate, rugged, and 'smart' adopting wireless data transfer or collecting data with portable devices. From a geospatial perspective, each sensed soil sample or core is identified with geographic coordinates, a unique identification number (ID), time stamp, and soil attributes or soil spectral data. Even disparate soil-environmental data collected with different data density, frequency of collection, soil sampling locations and times, etc. can be integrated into relational geodatabases that store different sets of field and laboratory soil data in the form of tables that are connected by key fields (IDs). Geospatial hypercubes of soil-environmental data provide the foundation to compute soil models that are then used to derive estimates, uncertainties, and prediction errors at predicted locations.

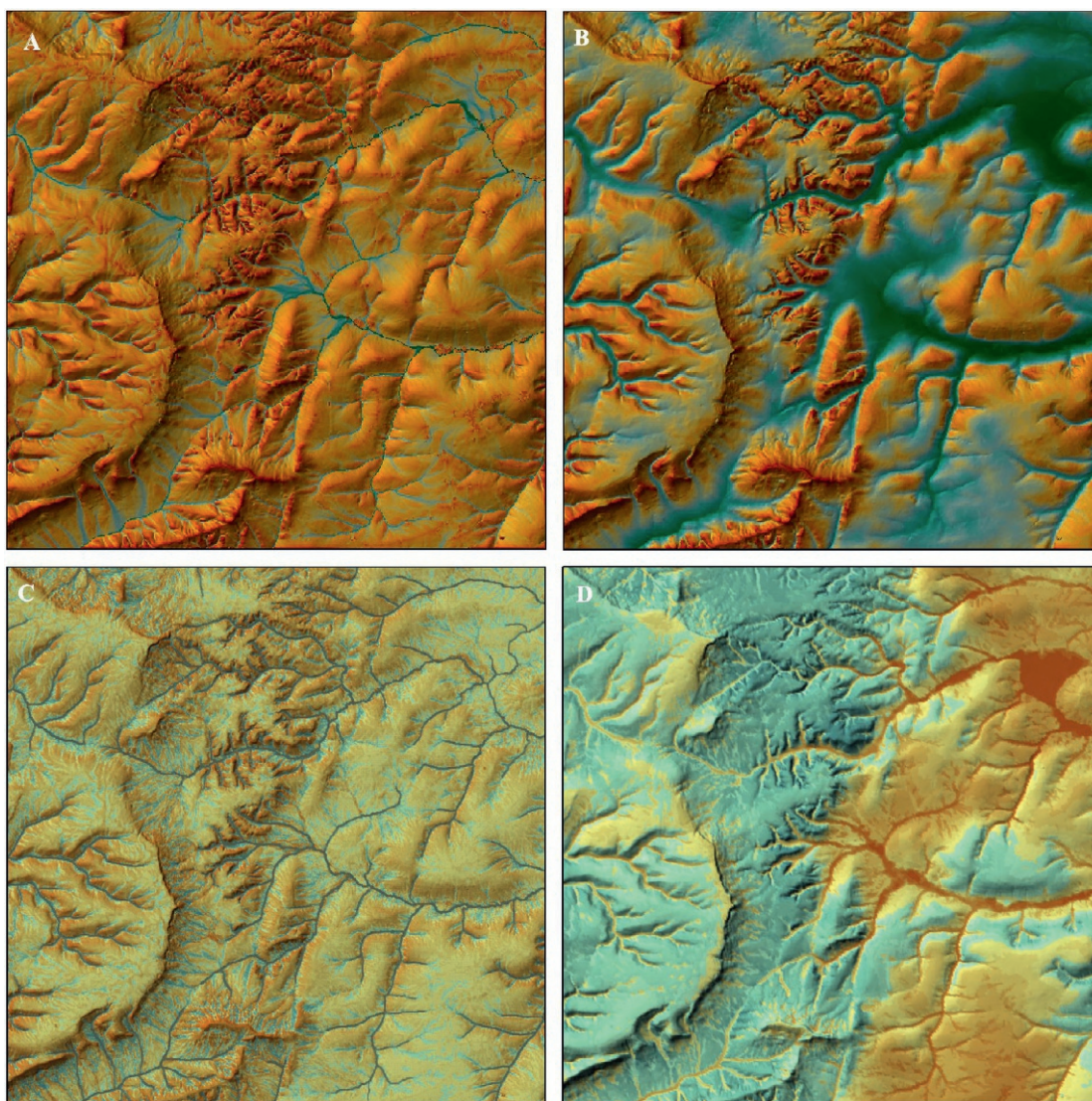


Fig. 5 The DEM-based terrain parameterization and estimated soil properties. (a) Terrain wetness index (According to Moore ID, Gessler PE and Peterson GA (1993) Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal* 57: 443–452.; (b) SAGA (system for an automated geoscientific analysis) terrain wetness index.; (c) Mass balance index, and (d) Silt content of the soil surface layer (red, high silt content; green, low silt content). Adapted from Böhner J, Bock M, Selige T and Köthe R (2014) Geographical information systems—Applications to soils. In: Elias SA (Ed.), *Reference Module in Earth Systems and Environmental Sciences*, Elsevier Inc., <https://doi.org/10.1016/B978-0-12-409548-9.09107-7>.

A review of global soil maps for Earth system models was provided by Dai et al. (2019). Global soil map products include the coarse-resolution Digital Soil Map of the World (DSMW) developed by the Food and Agriculture Organization of the United Nations (FAO/UNESCO). The map shows 4931 mapping units consisting of soil associations. For each of the soil units, estimates are provided of physical (% sand, % silt, % clay, bulk density) and chemical properties (pH, organic carbon, cation exchange capacity, base saturation, C/N ratio, CaCO_3 -content) in the topsoil and subsoil. Apart from the maps of classification units contained in the World Soil Reference Base (WSRB) and tables on soil analyses for every country of the world, an additional soil database was created for global environmental studies, including data on soil moisture storage capacity, soil-drainage classes, and effective soil depth. A corresponding data set of water-holding capacities in a 1×1 grid (latitude, longitude), drawn from original sources to meet the requirements of general circulation model (GCM) experiments. Furthermore, the Global Soil Partnership of the FAO published the Global Soil Organic Carbon (GSOC) Map on a $1 \text{ km} \times 1 \text{ km}$ grid covering a soil depth of 0–30 cm. The Harmonized World Soil Database (HWSD) served to develop the WISE30sec on a 5 by 5 arc-minutes global grid.

Examples of national and large-region soil information systems that provide soil polygon maps, soil rasters, and soil profile databases include the U.S. National Cooperative Soil Survey Soil Characterization database and Gridded National Soil Survey Geographic Database (gNATSGO), Africa Soil Information Service (AfsIS), European LUCAS Soil Database, Australian Soil Resource Information System, Chinese National Soil Profile database, Canadian Soil Information System, soil profiles from the

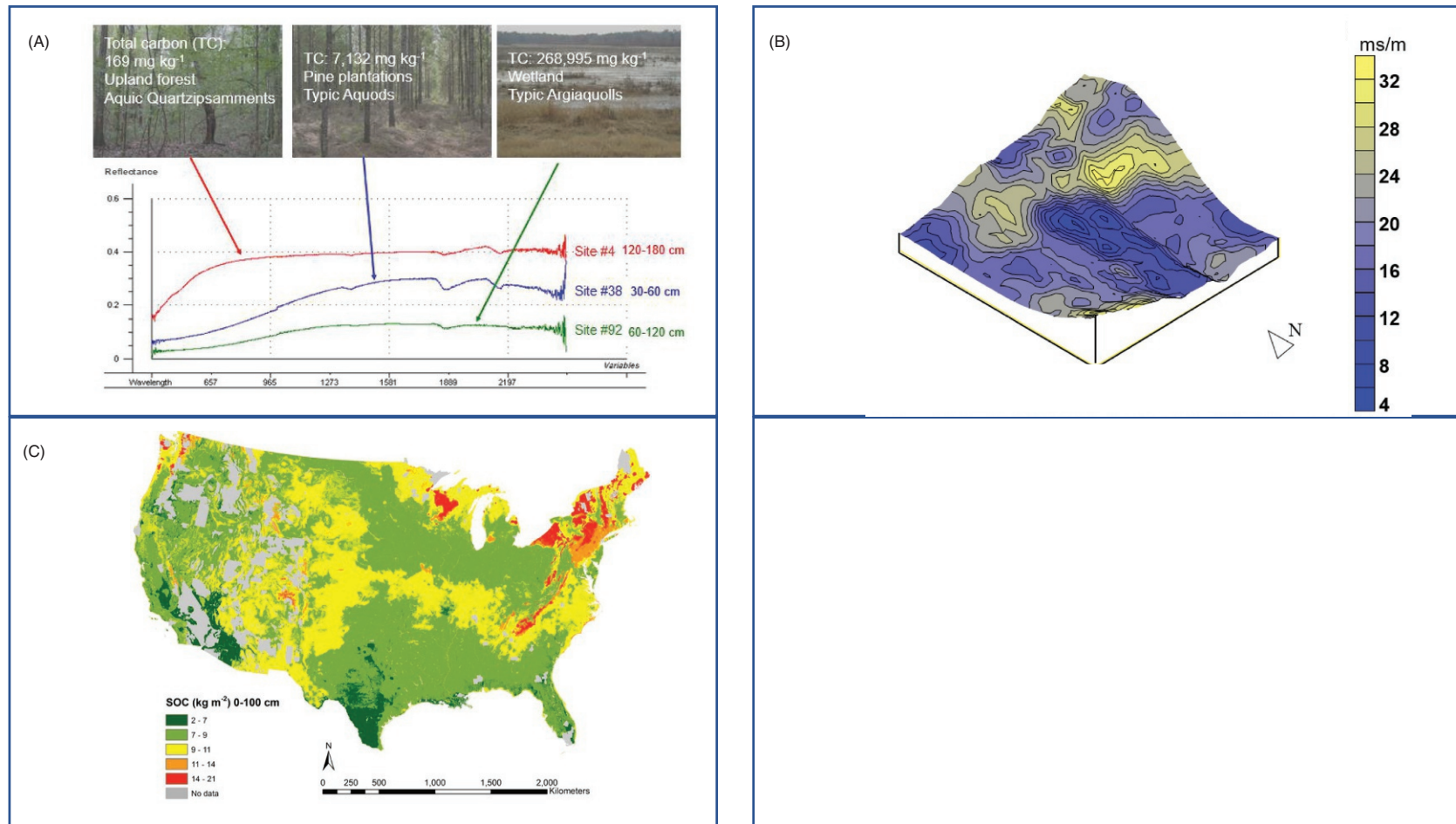


Fig. 6 (a) Visible-near-infrared spectral scans of soil samples from three different sites in Florida, United States. The spectral scans differed distinctly at sites characterized by different vegetation, soils, and soil organic carbon content. (b) Electromagnetic induction (EM) 38 soil map for a site in southern Wisconsin, Arlington Agricultural Research Station. Valley areas are associated with higher clay and silt loam content, while upslope areas show some eroded soils. Courtesy of S. Grunwald. The EC patterns may be due to the complexity of the landscape in southern Wisconsin where soils are formed in reworked loess that is silt-rich mixed with sand-rich glacial till parent material. Erosion in this landscape is a mix of wind and water erosion leading to some interesting relocation of soil material that is not solely topography driven. In this landscape both erosion and tillage operations modulate EC patterns. (c) Soil organic carbon (SOC) stock in 0–100 cm in the contiguous mainland of the United States of America. The soil grid with 1 km × 1 km spatial resolution was derived by soil predictive modeling using the Boruta algorithm and Random Forest, a machine learning algorithm (courtesy: Grunwald S, Xiong X, Patarasuk R, Ross CW, and Cao B). Soil data source: Natural Resources Conservation Service, United States Department of Agriculture.

SOTER database, and the Brazilian Soil Information System. The Global Soil Map (GSM) is a global consortium that has brought together various global soil grids standardized for six soil layers (0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm) and standardized uncertainty assessment. The SoilGrids—global gridded soil information of different versions of soil raster maps (1 km and 250 m spatial grid resolutions) were developed using state-of-the-art machine learning AI algorithms for a variety of soil properties (e.g., pH, SOC content, bulk density, coarse fragments content, sand content, silt content, clay content, cation exchange capacity, total nitrogen as well as SOC density and SOC stock) developed by ISRIC World Soil Information (<https://www.isric.org/explore/soilgrids>, last access: 14 September 2021). The Soil and Landscape Grid of Australia (SLGA) under the auspices of the GSM initiative serves as an example of consistent and relevant digital soil data at the continental scale (HYPERLINK "<https://www.cdw.csiro.au/aclep/soilandlandscapegrid/>", last accessed 2 February 2022).

The real-time below and above ground in-field monitoring of the environment is a fast-growing emergent of smart farming technology. According to Wolfert et al. (2017), Big Data in smart farming entails an integrated system consisting of smart sensing and monitoring of soil-environmental conditions of the agricultural system, sophisticated technology including high-precision GPS, satellite imaging, robots, unmanned aerial vehicles (UAVs), cloud-based event and data management, advanced geo-analysis that entails machine- and human-generated data, and advanced data analytics and interpretation. Innovation possibilities include open farm management with specific apps, remote/computer-aided advice, AI controlled decision-support systems, and regionally pooled data for scientific research. The vision for such integrated agricultural systems is to 'data mine' available geospatial data from geoportals (e.g., satellite data) and fuse them with laboratory- and field-monitoring of soils, crops, livestock, water, and climate to optimize farm management using advanced AI technology. At the core of such data and knowledge integration are geospatial and temporal domains.

Despite the potential for such integrated AI geospatial soil-environmental systems that strive towards real-time sensing of conditions and change of system components, there are also critical voices. Ethical concerns entail the amplified focus and reliance on AI technologies and machine-generated model outputs that lack knowledge discovery, human interpretation, and scientific understanding of the complexity of soil-ecosystems (Grunwald, 2021). The AI-geospatial soil models are sensitive to pseudo ('false') variables, erratic behavior of models due to outliers or misclassified pixels, and may rely on spurious relations between soil-environmental predictor variables and predicted soil properties (Grunwald, 2021). According to Liao (2020), the development of appropriate ethics and morals that value equally human- and machine-generated knowledge and understanding lags behind the rapid speed at which AI-technologies have developed. Tullis and Kar (2021) pointed out that today's large volumes of geospatial data generated by satellite sensors, GPS, and sensing systems have overlooked the provenance of information as a digital record of historical (retrospective) and potential future (prospective) geospatial processes which blurs the lineage and origins of data and subsequently makes geodata vulnerable to misappropriation. Location privacy is a somewhat nebulous geospatial concept, which has been connected to the erosion of privacy of people together with AI surveillance linked to the control and manipulation of people (Mitchell, 2019). Specifically, farmers and land managers who were traditionally considered stewards of the land and soil may face increasing infringement of the privacy of their location as more geodata of soil and the environment are collected remotely and through proximal sensing technologies. Zhao et al. (2020) asked: Can we trust AI to manage geospatial data? Or may AI be implicated in spoofing in geoscience that calls for spoofing detection principles to inform and protect people and society better from the deleterious effects of AI-managed GIS.

GIS-supported digital soil mapping and modeling

Digital soil mapping is tightly coupled to GIS, the latter enabling the management of soil-environmental data, geoprocessing, and map making. Specifically, GIS facilitate the harmonization and population of soil models to predict and or simulate soil properties, classes, indices, functions, and ecosystem services and also visualize outputs. Soil predictions and model performance metrics rooted in soil-factorial modeling based on SCORPAN or STEP-AWBH conceptual frameworks are derived either through in-built geoprocessing functions within the GIS or external applications coded in R, Python or other computer languages that interface with the geodata.

Comprehensive overviews of quantitative methods to model soil properties and soil classes were provided by Ma et al. (2019), and McBratney et al. (2003). Recent reviews of AI machine learning to model soils were provided by Khaledian and Miller (2020) and AI deep learning (artificial neural networks) to model soils by Padarian et al. (2019). A description of various hybrid (mixed stochastic-deterministic) methods (e.g., regression kriging) to model soils can be found in publications by Arrouays et al. (2017) and Keskin and Grunwald (2018).

To exemplify AI machine learning geospatial soil modeling a brief case study is presented. The goal of this study was to model SOC stock in the contiguous U.S. A total of 144,833 SOC and bulk density measurements in genetic horizons collected at 15,881 sites from the National Cooperative Soil Survey Soil Characterization Database (U.S. Natural Resources Conservation Service, NRCS) were retrieved and then interpolated with the mass preserving spline function to derive SOC stock for a depth of 0–100 cm in soil profiles. The SOC stock data were intersected with geospatial STEP-AWBH factors including soils from gSURGO (NRCS), topographic variables such as elevation, slope, curvature, and flow accumulation (National Elevation Model, U.S. Geological Survey), ecoregions (U.S. Environmental Protection Agency), parent material (rock type and gamma ray data from the U.S. Geological Survey), climate data including long-term average annual precipitation and temperature (PRISM climate data, Oregon State University), biotic data such as land cover/land use (National Land Cover Database), NDVI from MODIS, various vegetation-specific variables (Landfire dataset), and soil hydrology such as annual water table depth, available water storage

capacity, drainage class, hydrologic group, and soil moisture from gSSURGO and SMOS. Soil measurements and geospatial STEP-AWBH variables were used to train the soil model (calibration phase, 70% of data) and verify model output (validation phase, 30% of data). Random Forest (RF), one of the most popular machine learning AI algorithms, was used to model SOC stock in the contiguous U.S. The predicted raster map of SOC stock (kg m^{-2}) for 0–100 cm soil depth was produced at a spatial resolution of $1 \text{ km} \times 1 \text{ km}$ grid cells (Fig. 6c). The total SOC stock for the contiguous U.S. was 61.04 ± 10.66 (standard deviation) Pg C (0–100 cm). More specifically, the largest stocks were predicted by the RF model in three areas: Northeastern highland, northern Wisconsin, and northwest coast forest. Of the total stock of 61.04 Pg SOC about 43% (26.35 Pg) was stored in the topsoil (0–20 cm). Performance of the SOC model in calibration mode gave an R^2 of 0.88, root mean square error (RMSE) of $3.73 \text{ (kg m}^{-2}\text{)}$, and ratio of performance of inter-quartile range (RPIQ) of 1.83. In validation mode the SOC model had an R^2 of 0.43, RMSE of 6.99, and RPIQ of 0.96. Overall, of all STEP-AWBH variables SOC stock in the United States was controlled by a mix of soil, ecological, parent material, atmospheric, and water environmental variables (SEP-AW) and to lesser extent by biotic and topographic (T-B) variables.

Role of GIS in pedometrics, simulation models, and decision support systems

Geographic information systems play a similarly important role in DSM and pedogenic, land, and Earth simulation models that aim to simulate soil-ecosystem processes to model the spatial and temporal changes in soils and other environmental properties. These models view soils as a highly dynamic, reactive and controlling geographic entity rather than a static object of inventory.

A wide range of model applications exists in the context of soil carbon sequestration, soil erosion risk, soil degradation, soil health, crop yield and food security, global climate change, soil ecosystem services, and other assessments. All of them discretize a landscape into geographic model units, either polygons or grid cells. The finer is the model unit, the greater is the demand to populate models with fine-resolution soil-environmental data. The importance of GIS also changes; it is best described as a transition process, beginning with its use as a transaction-processing system and technical tool for data management. It then goes on to become an array of scientific analytical methods capable of supporting the basic demands of spatial information science, with the final stage as a decision-support system, with potential uses within a range of applications.

Examples of soil assessments entail the modeling of soil organic matter dynamics in agroecosystems at the landscape scale (Viaud et al., 2010), DayCent and its land surface submodel (Parton et al., 1998), and SOC modeling using RothC (Jebari et al., 2021). An overview of quantitative models for pedogenesis was provided by Minasny et al. (2008) and Stockmann et al. (2011).

With the dual role of soil as a production factor and an interacting and controlling layer within the atmosphere-biosphere-geosphere system, the assembly of soil data is of particular interest in the context of agro-environmental models supporting sustainable development. The momentum of this development is reflected in an increasing number of geospatial soil information systems. Although mostly embedded in primary land or environmental information systems, they show distinctly the technical and scientific specialization required for handling and providing digital soil and soil-related information targeted at sustaining soil health, soil security, and ecosystem services.

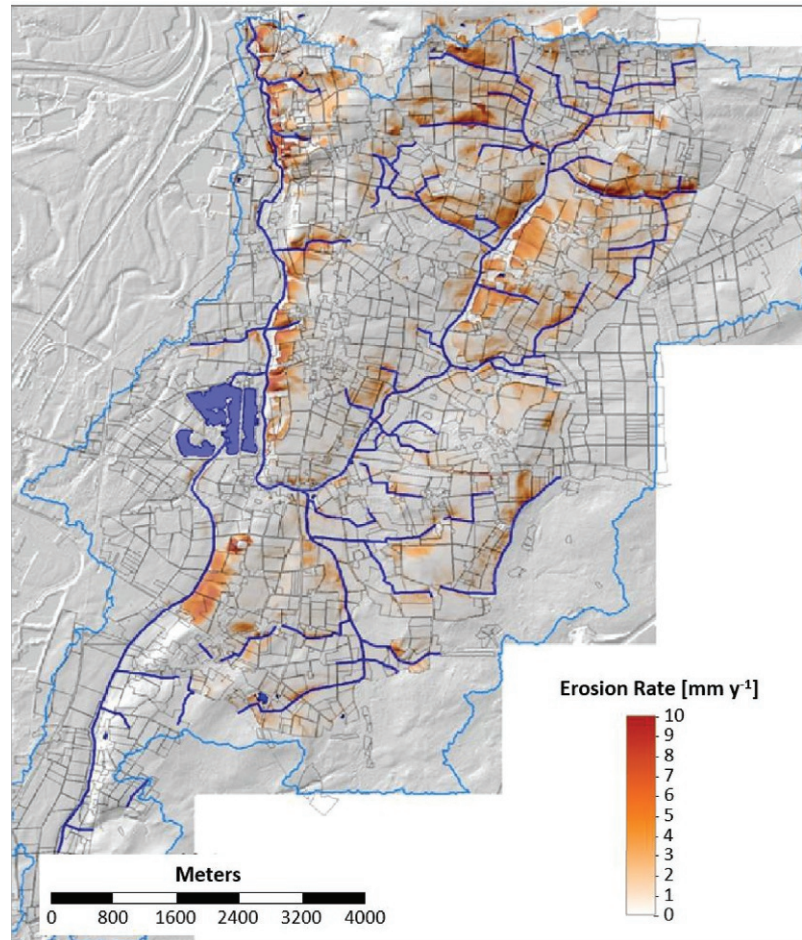
Strong dependence on the quality and parameter space of spatially explicit soil data is particularly evident in GIS-based soil erosion risk assessment. In general, models quantifying water erosion as the most severe degradation process in terms of its global distribution and impact are frequently differentiated into block models, covering agricultural plots as discrete modeling entities or spatially distributed erosion models. From an epistemological point of view, however, the differentiation between empirical bottom up and physically based top-down approaches is even more important with respect to distinctly differing data requirements.

Empirical erosion modeling traces back to the famous universal soil loss equation (USLE) developed in the late 1950s by Wischmeier and Smith (1987) to quantify rates of soil erosion on agricultural slopes as a multiplicative function of six environmental variables (the following factors: R, precipitation; L, slope length; S, slope; C, vegetation cover and tillage; K, soil-erodibility; P, erosion-protection). The parameterization of topsoil erodibility in the USLE K-factor considers soil texture (sand and silt percentages, coarse soil content), organic matter content, and classes of permeability of the soil's surface layer, e.g., entailed or delineated from soil taxonomy profiles. With its simplicity, the USLE and its numerous derivatives can be easily implemented in GIS, and until now are the established standards in practical risk assessment of soil erosion.

Empirically based models have limited transferability, therefore, more sophisticated, process-based approaches have been developed, capable of simulating water erosion in a physically consistent manner (e.g., Erosion 3D, Eurosem, and SWAT). Given that environmental threats from fluxes of water driven matter are not limited to onsite deterioration but also affect neighboring ecosystems with plant nutrients and chemical pollutants, GIS-based process modeling increasingly addresses surface runoff to support management decisions to minimize offsite risks. An example of integrated modeling of water erosion and surface runoff implemented in a client-server architecture with Web-GIS interface is given in Wendland et al. (2016). For illustration see Fig. 7.

Physically based models compared to empirical approaches are advantageous in terms of their wider applicability, but place greater demands on the spatiotemporal resolution and quality of input data. This is particularly valid for soil data, which have to be available either as physical soil attributes or as delineated parameterizations. With respect to its use in case studies, management decision support, and particularly in the context of precision farming applications, major requirements concern the spatial resolution; and thus, demand suitable procedures for the high resolution spatialization of soil properties. In this regard, Web-GIS

(A)



(B)

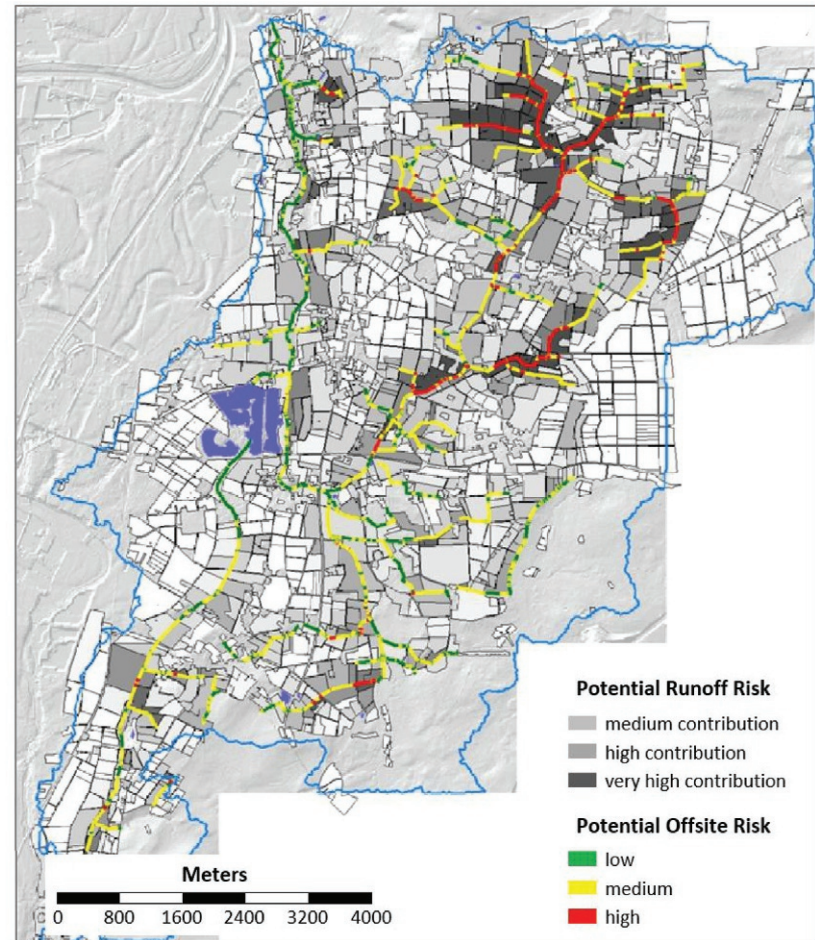


Fig. 7 Soil erosion and surface runoff in 2014 in the Gehle catchment (North Rhine-Westphalia, Germany). (a) Annual soil erosion rates. (b) Per-field classified surface runoff (shaded) and offsite risk classes (signal colors) for channel segments. Adapted from Wendland S, Bock M, Böhner J, Feise D and Lembrich D (2016) Towards the development of a GIS based Diagnosis-Tool for the spatially-explicit assessment of runoff and erosion risks on agricultural fields. *GeoÖko* 37(3–4): 139–164.

interfaces increasingly used in agriculture offer a promising option to extend existing soil data bases from legacy soil surveys through to the integration of farm level data that improve the spatial representation of soils.

Conclusions

Geographic information systems and technologies have been of critical importance in supporting the collection of soil and environmental data, database management, digital soil mapping, and pedometrics. Specifically, sensor technology, cloud-computing, and AI have accelerated the digitization of soils and simulation modeling of soil processes and change at increasingly finer spatial and temporal resolutions. These advances in technology have stressed data-driven approaches to model soils rather than the discovery of knowledge about natural soils and how people use and value them. As these trends continue in the future there are serious risks that AI and geospatial technologies could be misappropriated, used to control soil-environments, farmers, and land stewards (e.g., agro-technology), applied in unethical ways breaching data privacy issues, and overshadowing provenance and people's relation with the soil. On the other hand, digital soil information plays a profound role in addressing important global dilemmas, such as soil health and security, food security, soil erosion, climate change and carbon cycling, and degradation or even loss of soil ecosystem services and functions.

Acknowledgments

The authors thank Margaret Oliver and Uta Stockman for valuable comments on the manuscript. SAGA GIS development is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2037 'CLICCS—Climate, Climatic Change, and Society'—Project Number: 390683824, contribution to the Center for Earth System Research and Sustainability (CEN) of Universität Hamburg.

References

- Arruays D, Lagacherie P, and Hartemink AE (2017) Digital soil mapping across the globe. *Geoderma Regional* 9: 1–4. <https://doi.org/10.1016/j.geodrs.2017.03.002>.
- Blatt AJ (2012) Ethics and privacy issues in the use of GIS. *Journal of Map & Geography Libraries* 8(1): 80–84. <https://doi.org/10.1080/15420353.2011.627109>.
- Bolstad P (2019) *GIS Fundamentals: A First Text on Geographic Information Systems*. Eider Press.
- Conrad O, Bechtel B, Bock M, Dietrich H, Fischer E, Gerlitz L, Wehberg J, Wichmann V, and Böhner J (2015) System for automated geoscientific analyses (SAGA) v. 2.1.4. *Geoscientific Model Development* 8: 1991–2007. <https://doi.org/10.5194/gmd-8-1991-2015>.
- Dai Y, Shangquan W, Wei N, Xin Q, Yuan H, Zhang S, Liu S, Lu X, Wang D, and Yan F (2019) A review of the global soil property maps for Earth system models. *The Soil* 5(2): 137–158. <https://doi.org/10.5194/soil-5-137-2019>.
- Grunwald S (2006) What do we really know about the space–time continuum of soil landscapes? In: Grunwald S (ed.) *Environmental Soil-Landscape Modeling: Geographic Information Technologies and Pedometrics*, pp. 3–36. CRC Press.
- Grunwald S (2021) Grand challenges in pedometrics-AI research. *Frontiers in Soil Science - Pedometrics* 1: 1–8. <https://doi.org/10.3389/fsoil.2021.714323>.
- Grunwald S, Thompson JA, and Boettinger JL (2011) Digital soil mapping and modeling at continental scales: Finding solutions for global issues. *Soil Science Society of America Journal* 75(4): 1201–1213. <https://doi.org/10.2136/sssaj2011.0025>.
- Herrick JE, Karl JW, McCord SE, Buenemann M, Rigos C, Courtright E, Van Zee J, Ganguli AC, Angerer J, Brown JR, Kimiti DW, Saltzman R, Beh A, and Bestelmeyer B (2017) Two new mobile apps for rangeland inventory and monitoring by landowners and land managers. *Rangelands* 39(2): 46–55. <https://doi.org/10.1016/j.rala.2016.12.003>.
- Jebari A, Álvaro-Fuentes J, Pardo G, Almagro M, and del Prado A (2021) Estimating soil organic carbon changes in managed temperate moist grasslands with RothC. *PLoS One* 16(8): e0256219. <https://doi.org/10.1371/journal.pone.0256219>.
- Keskin H and Grunwald S (2018) Regression kriging as a workhorse in the digital soil mapper's toolbox. *Geoderma* 326: 22–41. <https://doi.org/10.1016/j.geoderma.2018.04.004>.
- Khaledian Y and Miller BA (2020) Selecting appropriate machine learning methods for digital soil mapping. *Applied Mathematical Modelling* 81: 401–418. <https://doi.org/10.1016/j.apm.2019.12.016>.
- Liao SM (ed.) (2020) *Ethics of Artificial Intelligence*. Oxford University Press.
- Ma Y, Minasny B, Malone BP, and McBratney AB (2019) Pedology and digital soil mapping (DSM). *European Journal of Soil Science* 70(2): 216–235. <https://doi.org/10.1111/ejss.12790>.
- McBratney AB, Mendonça Santos ML, and Minasny B (2003) On digital soil mapping. *Geoderma* 117(1–2): 3–52. [https://doi.org/10.1016/S0016-7061\(03\)00223-4](https://doi.org/10.1016/S0016-7061(03)00223-4).
- McBratney AB, Minasny B, and Stockman U (eds.) (2018) *Pedometrics*, 1st edn. Springer.
- Minasny B, McBratney Alex B, and Salvador-Blanes S (2008) Quantitative models for pedogenesis—A review. *Geoderma* 144(1–2): 140–157. <https://doi.org/10.1016/j.geoderma.2007.12.013>.
- Mitchell M (2019) *Artificial Intelligence: A Guide for Thinking Humans*. Farrar, Straus and Giroux.
- Padarian J, Minasny B, and McBratney AB (2019) Using deep learning for digital soil mapping. *The Soil* 5(1): 79–89. <https://doi.org/10.5194/soil-5-79-2019>.
- Parton WJ, Hartman M, Ojima D, and Schimel D (1998) DAYCENT and its land surface submodel: Description and testing. *Global and Planetary Change* 19(1): 35–48. [https://doi.org/10.1016/S0921-8181\(98\)00040-X](https://doi.org/10.1016/S0921-8181(98)00040-X).
- Stockmann U, Minasny B, and McBratney AB (2011) Quantifying processes of pedogenesis. *Advances in Agronomy* 113: 1–71. <https://doi.org/10.1016/B978-0-12-386473-4.00006-3>.
- Thompson JA, Roecker SM, Grunwald S, and Owens PR (2012) Digital soil mapping: Interactions with and applications for hydrogeology. In: Lin HS (ed.) *Hydrogeology—Synergistic Integration of Pedology and Hydrology*, pp. 665–709. Academic Press.
- Tullis JA and Kar B (2021) Where is the provenance? Ethical replicability and reproducibility in GIScience and its critical applications. *Annals of the American Soil Organic Carbonation of Geographers* 111(5): 1318–1328. <https://doi.org/10.1080/24694452.2020.1806029>.
- Viaud V, Angers DA, and Walter C (2010) Toward landscape-scale modeling of soil organic matter dynamics in agroecosystems. *Soil Science Soil Organic Carbonation of America Journal* 74(6): 1847–1860. <https://doi.org/10.2136/sssaj2009.0412>.

- Viscarra Rossel RA, Adamchuk VI, Sudduth KA, McKenzie NJ, and Lobsey C (2011) Proximal soil sensing: An effective approach for soil measurements in space and time. *Advances in Agronomy*. vol. 113, pp. 243–291. Elsevier. <http://linkinghub.elsevier.com/retrieve/pii/B9780123864734000051>.
- Viscarra Rossel RA, Behrens T, Ben-Dor E, Brown DJ, Demattê JAM, Shepherd KD, Shi Z, Stenberg B, Stevens A, Adamchuk V, Aichi H, Barthès BG, Bartholomeus HM, Bayer AD, Bernoux M, Böttcher K, Brodský L, Du CW, Chappell A, and Ji W (2016) A global spectral library to characterize the world's soil. *Earth-Science Reviews* 155: 198–230. <https://doi.org/10.1016/j.earscirev.2016.01.012>.
- Wendland S, Bock M, Böhner J, Feise D, and Lembrich D (2016) Towards the development of a GIS based Diagnosis-Tool for the spatially-explicit assessment of runoff and erosion risks on agricultural fields. *GeoÖko* 37(3–4): 139–164.
- Wischmeier WH and Smith DD (1987) Predicting rainfall erosion losses: A guide to conservation planning [USA]. In: *Agriculture Handbook No. 537*. Washington, DC: US Department of Agriculture.
- Wolfert S, Ge L, Verdouw C, and Bogaardt M-J (2017) Big Data in smart farming—A review. *Agricultural Systems* 153: 69–80. <https://doi.org/10.1016/j.agry.2017.01.023>.
- Xiong X, Grunwald S, Myers DB, Kim J, Harris WG, and Comerford NB (2014) Holistic environmental soil-landscape modeling of soil organic carbon. *Environmental Modelling & Software* 57: 202–215. <https://doi.org/10.1016/j.envsoft.2014.03.004>.
- Zhao B, Zhang S, Xu C, and Liu X (2020) Spoofing in geography: Can we trust artificial intelligence to manage geospatial data? In: Ye X and Lin H (eds.) *Spatial Synthesis. Human Dynamics in Smart Cities*, pp. 325–338. Springer. https://doi.org/10.1007/978-3-030-52734-1_19.

Precision agriculture

James A Taylor, UMR ITAP, L'Institut Agro-Montpellier, INRAE, University of Montpellier, Montpellier, France

© 2023 Elsevier Ltd. All rights reserved.

The emergence of precision agriculture	710
Defining precision agriculture	711
Variability, precision agriculture and the production system	712
Practical implementation of PA (particularly SSCM)	713
Evolution of global navigation satellite systems	714
Sensing systems for precision agriculture	715
What type of sensor?	715
Scientific principle - physical vs. chemical vs. biological sensors	715
Proximal vs remote-sensing	716
Active vs passive sensors	716
A sensor of a property or a response?	716
What type of data or information?	716
Soil	716
Crop	718
Animal	719
Meteorological	719
The role of interpolation, pedometrics and digital soil mapping in precision agriculture	720
Decision-making and management units in precision agriculture	721
Precision vs. digital agriculture	722
Concluding comments	724
References	724
Further reading	725

Abstract

Precision agriculture is a management strategy to increase the resolution of agricultural decision-making in response to variability in production. For crop and pasture systems, this variation in production is strongly linked to soil variability. For this reason, improving the ability to measure and map soil variation has been a key factor in Precision Agriculture. This chapter introduces the concept of precision agriculture, provides an overview of the key enabling technologies, with a focus on soil sensing, and discusses the importance of pedometrics and digital soil mapping for decision-making. On-going and emerging links with soil science are also introduced.

Key points

- Precision agriculture is a management strategy to make better decisions at finer spatial and temporal scales than was previously possible
- Site-specific crop management has always been and continues to be premised on managing the variability in soil systems to improve production
- Precision agriculture relies on pedometrics and digital soil mapping to provide the methods and the soil information needed for site-specific crop management.
- Precision agriculture continues to need soil science input to advance.

The emergence of precision agriculture

Precision agriculture (PA) is now a well-known term in global agriculture. But where did the term and concept of PA come from? The current concept of precision agriculture emerged in cropping systems in the late 1980s with the matching of soil fertility maps, which were derived from grid-based sampling of soil properties, with newly developed variable-rate (VRT) fertilizer equipment (Pierce et al., 1994; Stafford, 1996). Using dead-reckoning principles, fertilizer was applied variably to complement the variation in soil fertility. Precision agriculture at this time was often referred to as “farming by the soil.”

At this stage, crop yield monitoring technologies were still in the research phase and satellite navigation systems were unavailable to the general public. Around 1990, commercial receivers became available for civilian access to the USDA's (United States Department of Agriculture) Navstar GPS (global positioning system) enabling rapid and ‘accurate’ vehicle location and

navigation that generated a flurry of activity. This new positioning information was incorporated into variable-rate systems and agri-sensors, typified by crop yield monitors that began to be available in the commercial market (Plant, 2001). By 1993 the USDA's GPS was fully operational allowing the fine-scale monitoring and mapping of yield variation within fields with GPS-enabled on-harvester yield monitors. Linking variable yield data with maps of plant nutrient changes within a field marked the true beginning of PA in broadacre cropping.

As PA technology (crop and soil sensing) improved in response to a growing demand, to be more efficient with inputs and to reduce the environmental impact of production it became clear that grid sampling at the intensity required to characterize the variation in soil and crop properties adequately was prohibitively costly. Consequently, by the late 1990s, there was a strong shift toward 'zone' management. This approach subdivides fields into zones of similar crop response that addressed some of the issues with data resolution while trying to maximize the benefits of differential (variable) crop management (Lark and Stafford, 1997).

The success, and potential for further success, observed in the broadacre cropping industry prompted other farming industries, particularly viticultural and horticultural crops, to adopt PA (Arnó et al., 2009). From the late 1990s, there has been increasing research on non-grain crops. Advances in Global Navigation Satellite System (GNSS) technology since 1999 have opened the door for many other applications apart from variable-rate management, including machinery guidance, auto-steering and controlled-traffic farming (CTF). More recently, lessons learned from PA on data management, implementation and adoption are helping to guide the development of digital agriculture, of which PA is only a component.

Precision agriculture has typically been advocated and employed in developed agricultural systems with higher levels of mechanization (Mondal and Basu, 2009). Applications to developing agricultural systems, including subsistence farming structures exist, but are far fewer (Mizik, 2023). However, the philosophy of PA is just as applicable to small-scale and developing systems, even if the degree of technological innovation is not as great (Cook et al., 2003). While mechanization may be less, management is often more (spatially) intense and can still be spatially varied, provided suitable tools for better decision-making are available. These need to be low-cost, user-friendly and effective (Mizik, 2023). There are now globally increasing levels of disruptive technologies, notably smartphone and tablet technologies, which are providing a platform for PA applications in all types of agriculture systems (Lagos-Ortiz et al., 2018). The key is to have the correct decision-making process for each type of system and for each location, i.e. a ubiquitous platform but a site-specific decision system.

To execute, PA is no longer an agricultural science question. Precision agriculture is about developing systems to gather more or new data and information, and decision systems to process and act within agricultural production systems. It depends on advances in many other disciplines for it to advance. At the core there will always be a need for good agricultural science and good soil science, but this operates around developments in sensing technologies (the domain of electrical and chemical engineers), data management and display (computer scientists), data analytics (bioinformaticians and statisticians) and PA machinery (agricultural engineers). Precision agriculture is a truly multi-disciplinary domain, especially when the social and economic components are added to the technical aspects. With such multi- and inter-disciplinary needs, it is almost impossible for any single individual to have the complete PA skills set. Precision agriculture must therefore exist within collaborative and multi-disciplinary teams, within which diverse groups (or individuals) support and enable each other.

This provides an enormous challenge for the PA community. Agriculture is not always seen as a 'sexy' area to work in. Neither does it command the financial power of many other industries (e.g. defense, mining, finance, etc...) where data analytics and engineering skills are equally sought after. There is a challenge for the PA community to attract and to retain people capable of crossing over between engineering, data analytics and agricultural science, including soil science. This does not require such people to be experts in all areas, but rather to have a fundamental understanding of how the many aspects affect innovations in their main specialty. Biological systems, and particularly intrusive systems, such as agricultural production systems, are not intuitive systems for the application of the pure sciences. It is a highly variable and often changeable environment in which practical applications and trade-offs are needed. This is often difficult for people unfamiliar with agriculture to understand, however it is these people (engineers, computer scientists etc.) that are needed to make PA work.

Defining precision agriculture

Many definitions of PA exist and many people have had different ideas of what it should encompass. From the mid-1990s to mid-2010s, numerous and diverse definitions were proposed. In 2019, the International Society for Precision Agriculture (ISPA) (www.ispag.org) accepted a formal definition of PA after a period of consultation with PA-oriented researchers and practitioners from around the world. The ISPA definition of PA is:

Precision agriculture is a management strategy that gathers, processes and analyzes temporal, spatial and individual data and combines it with other information to support management decisions according to estimated variability for improved resource use efficiency, productivity, quality, profitability and sustainability of agricultural production.

It is important to note that the agricultural community identifies PA as an evolving 'management strategy' to support 'decision-making' with regard to resource use. It is not explicitly linked to information technology on-farm (although many new information technologies will improve decision-making). The decision-making itself could be in response to spatial (changes across a field) or temporal variation (changes through a season or seasons). The inference is that a better understanding of variation at finer spatial

and or temporal resolutions will improve decision-making and provide a wide range of benefits (economic, environmental and social) that may or may not be known or measurable at present.

This definition also refers to 'agricultural production' and not agronomy. The PA philosophy has been most strongly expounded in cropping systems; however, PA can relate to any agricultural production system. Its concepts are equally applicable to animal industries, fisheries and forestry, and in many cases PA techniques are being implemented in these domains without being identified as such. To distinguish PA in cropping systems, where soil variation becomes a more dominant issue in site-specific decision-making, the term site-specific crop management (SSCM) is often used. This is defined as "a form of PA whereby decisions on resource application and agronomic practices are improved to better match soil and crop requirements as they vary in the field" (Whelan and Taylor, 2013).

The overall aim of a PA management philosophy can be simplified as an approach to: (i) optimize production efficiency, (ii) optimize quality, (iii) minimize the environmental impact of production, and (iv) minimize the risk involved with production - all at the site-specific level (Whelan and Taylor, 2013). These objectives are not particularly new, with essays on this topic dating from the early 18th century. What is new is the scale at which these concepts can be implemented and the technology available to support them. Prior to the industrial revolution, agriculture was characterized by small fields, with farmers often having a detailed, 'site-specific,' qualitative knowledge of their production system. The mechanization of agriculture, and economies of scale, from the 1950s onwards has resulted in the latter half of the 20th century being dominated by large-scale uniform 'average' management practices in developed agricultural economies. Technological advances in the late 20th and early 21st centuries, has allowed agriculture to revert toward site-specific agriculture, by providing fine-scale, quantitative production information, while retaining the economies of scale associated with 'large' mechanized operations.

Variability, precision agriculture and the production system

Precision Agriculture depends on the existence of variability, and broadly speaking 'variability in production = PA opportunity' (Pringle et al., 2003). Variation in production exists in both a spatial or temporal context (Fig. 1, Blasch et al., 2020). Spatial variation occurs over a measurable distance, and temporal variation occurs over a measurable time period. Precision agriculture should address both of these, although to date it has focused primarily on the variation in spatial production. Both of these types of variation exist in many different forms and need to be considered from different perspectives, including the type, magnitude and distribution pattern of the observed spatial, temporal or spatio-temporal variation.

The magnitude in both types of variation is defined as the difference between the low and high values of a measured property. The distribution pattern relates to how variation is changing in either the space or time dimension. The management implications of these aspects of variation are diverse and are fundamentally linked to the property being measured. However, there are a few simple generalizations that need to be remembered (Pringle et al., 2003). First, the observed magnitude in the variation should be assessed and considered relative to a threshold level of variation below which it would be uneconomic to attempt to manage it, i.e. if the production cost of site-specific management vs uniform management is greater than the potential benefit (especially from a short-term economic perspective) then it makes little sense to apply PA (Ancev et al., 2005). However, if the environmental benefits of a PA approach could be correctly expressed (and remunerated) in a fiscal sense, then these thresholds should decrease and areas with a small magnitude of variation in production may become suitable (economically viable) for PA management. If spatial variation does not exist, then a uniform management system is both the cheaper and more effective management strategy. Second, the distribution pattern of the variation in production needs to be considered relative to possible management interventions. In spatial terms; the pattern of variation should be considered in relation to the smallest possible treatment area. For example, with variable-rate application of fertilizers this is defined by the size and reaction time of the VRT spreader used. In temporal terms, the pattern of variation relates to the timeliness of interventions at temporally varying management stages during the growing season (or the whole season if relevant). In many situations the magnitude of temporal variation may appear much greater than that of the spatial variation. If the impact of temporal variation on production overwhelms that of spatial variation, then careful consideration needs to be given to whether a uniform or differential management strategy is the optimal risk aversion strategy over time.

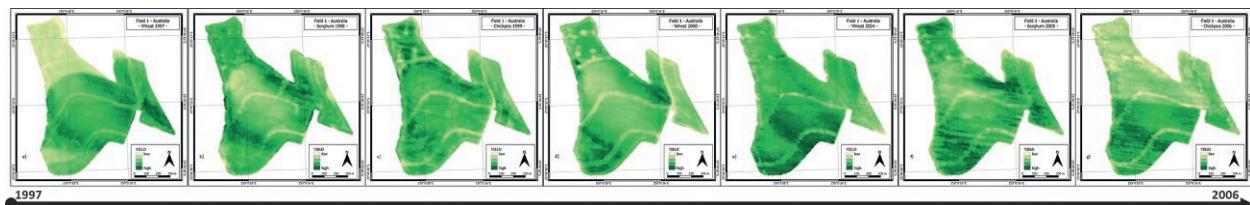


Fig. 1 A time-series of yield maps from a field in North-western New South Wales, Australia illustrating the spatial and temporal nature of yield variation within a field, which may be environmentally driven by the soil or imposed by management decisions. The 1997 and 2006 maps show distinct management patterns and the intervening years illustrate differing spatial patterns linked to soil and weather conditions in a given year. The field illustrates the difficulty that growers have in interpreting yield (and other production data) visually. Data supplied by Mr M. Smith and images adapted from Blasch G, Li Z, and Taylor JA (2020) A novel pattern recognition approach for the delineation of yield productivity-stability zones using yield map time series. *Precision Agriculture* 21: 1263–1290.

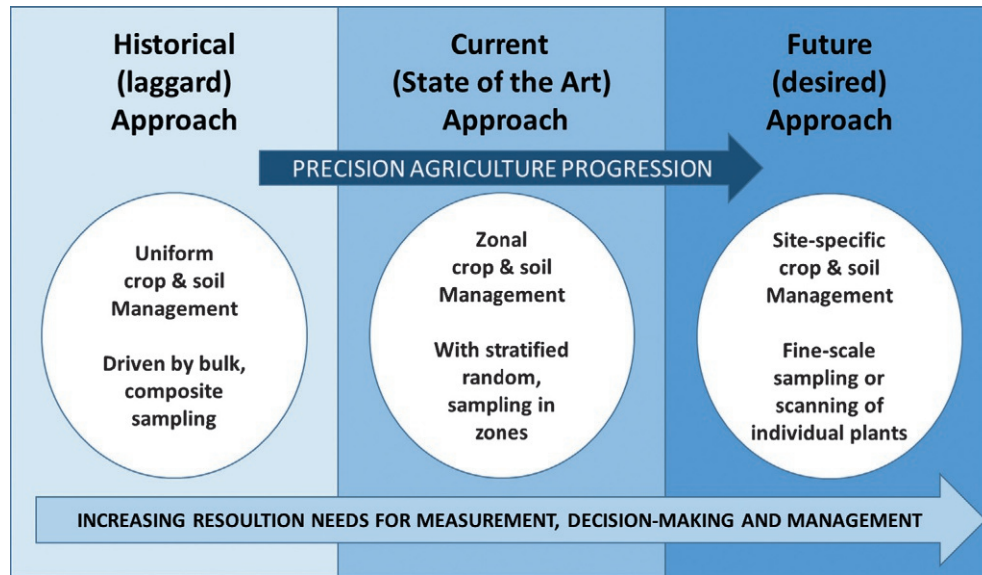


Fig. 2 A schematic diagram illustrating the shift from a historical, conventional uniform field management strategies, via management unit approaches to a true site-specific management strategy. The required increase in data with these shifts are also indicated. Adapted from an image originally produced by the Precision Agriculture Laboratory, The University of Sydney.

Based on these considerations, PA in crops at present operates mostly on a zonal rather than on a completely site-specific basis (Fig. 2). In practice, existing fields are subdivided based on production and environmental variation into more homogeneous zones (Fridgen et al., 2004). These management zones or management units are then managed uniformly, as if they were an independent field. As technology and knowledge advances PA will begin to approach a truly site-specific management regime, and this is now beginning to be observed in perennial tree cropping systems. For precision livestock farming (PLF), the animal is the real site-specific ‘management unit,’ and is typically the scale at which most PLF actions are performed.

Practical implementation of PA (particularly SSCM)

As outlined in the preceding definition, PA is a continuous management strategy, i.e. it is circular and recurring in nature. It has often been expressed in the form of a quality control cycle, based on the ability to measure production, analyze the data, make a decision and apply an intervention (Fig. 3). For agricultural systems, this cycle has been enabled and driven by the ability to potentially geo-reference any production observation, i.e. locate it in time AND space.

Geo-referencing: Global Navigation Satellite Systems (GNSS) (of which the USDA GPS (Global Positioning System) is the most widely known) are the enabling technology for PA and are now commonplace on farms and within our general lifestyle. The ability to geo-reference production activities provides producers with the option to map and visually display farm operations. More advanced systems have also enabled growers to embrace guidance and autosteer technologies on farms ahead of the general population. This permits machinery to drive along repeatable tracks that minimizes soil compaction, reduces driver fatigue over long shifts, minimizes errors and workplace accidents, and enables more precision and timeliness in production operations. For example, it minimizes overlap and ‘doubling’ of inputs or allows operations to be conducted in light-limiting conditions, such as at night or during fog events. By reducing overlaps, immediate economic gains can be achieved, which are generally in the region of 5–8% of applied inputs. The value of a timely operation is harder to quantify, but can be very high if sowing or fertilizer windows are missed due to adverse weather.

Production, Soil and Climate Monitoring: Many sensors and monitors already exist for in-situ and on-the-go measurement for a variety of production, soil and climate variables. The majority of PA research has been directed at identifying how to use the output from these sensors to improve production. These sensors can be deployed remotely (e.g. satellite systems), proximally (e.g. on farm machinery or animals) or as static systems (e.g. weather stations). They can operate on-the-go and/or log data for future analysis.

Attribute Mapping: Agri-environmental sensors often produce large, intensive data sets. These are operational data sets, often irregularly spaced and containing errors or artifacts associated with production effects. These data need to be ‘cleaned,’ processed and interpolated before they can be used for effective decision-making. Since the start of PA, statisticians have been researching accurate ways of describing and representing these spatial and temporal data, and the majority of this research has evolved from the pedometrics community (see later section). For spatial and sparse data, the challenges for attribute mapping are now generally well understood within the PA community, although generally poorly disseminated in the wider agricultural community. The mapping

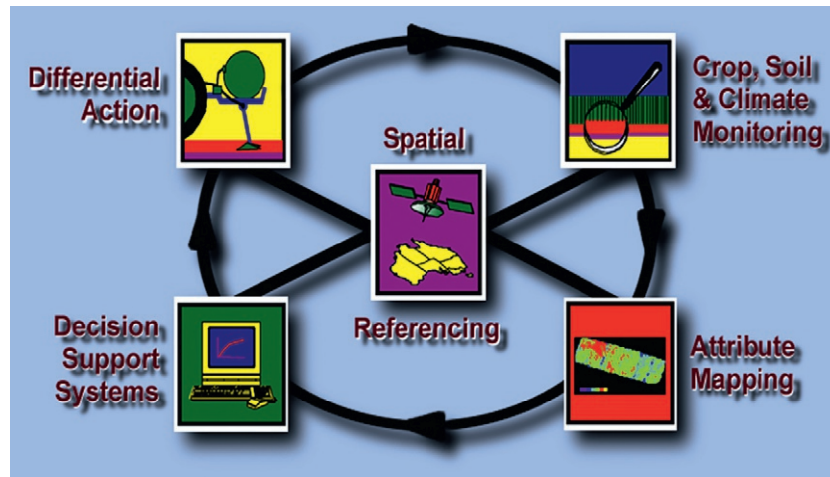


Fig. 3 The Precision Agriculture cycle indicating spatial referencing as the enabling, central technology that drives the other parts of the cycle. This follows a typical quality control cycle of observation, analysis, decision and action, which repeats to improve performance incrementally. Reproduced with permission from the Precision Agriculture Laboratory, University of Sydney.

of temporal effects remains a challenge for PA at present. Displaying either raw or processed agri-data is generally easy in agriculture with the advances in Geographical Information Systems (GIS) and computer graphics in other domains.

Decision Support Systems: The success (or failure) of PA or any management strategy depends on the quality of decisions that are generated through the strategy. This is the most important and the hardest part of the PA cycle to complete. If the wrong decision is generated, then growers will not adopt PA (Adrian et al., 2005). Agri-decisions can generally be labelled as embedded or information-intensive (Miller et al., 2018). Embedded decisions tend to be universal and do not require the grower to pass their knowledge to the process. Information-intensive decisions are the opposite and will be very dependent on grower-specific conditions. Autosteer using GNSS is an example of an embedded decision – the tractor drives straight in a repeatable line with many associated benefits without the grower needing to interact with the system. Variable-rate nitrogen application is an information-intensive decision. There is no ‘correct’ answer. The rate will depend on local environmental conditions, the producers production goal (e.g. animal feed vs premium wheat production), previous management, expected or forecast seasonal conditions, access to technology and advice, the producers’ level of acceptable risk, and so on. For these types of decision, decision support systems (DSSs) need to be as flexible, transparent and responsive to grower needs as possible (Lamb et al., 2008). A one-size-fits-all DSS system will not work for such decisions. The availability of flexible DSS for agriculture is very limited at the moment.

Differential Action: The differential application of production inputs at finer spatial or temporal resolutions is essentially an engineering and control problem that is being driven by the private sector. Many different, effective systems are commercially available to allow for differential production management on farms. Access to this technology is not a limitation, although cost may be prohibitive. The main constraint to the adoption of differential management is the quality of and trust in the decision(s) to be enacted, i.e. the availability of good agricultural DSSs.

Evolution of global navigation satellite systems

As noted previously, the game-changing technology for the evolution of PA has been access to satellite-based navigation systems. Collectively known as a GNSS, they are more ubiquitously referred to as GPS (Global Positioning System), which is the name of the GNSS operated by the government of the United States of America. The first GNSSs, namely the (then) USSR’s GLONASS and the USA’s GPS, arose from the space race and the cold war from the 1950s to 1980s, and the need for both military bodies to have more precise location information. The advantages of a global navigational system, and the need to avoid dependence on another government, has seen the EU and China also establish their own GNSS, Galileo and Beidou, respectively.

Although conceived for military purposes, and still used as such, the potential civilian and commercial usages of the technology were quickly identified and in the late 1980s the US government became active in promoting these uses via their GPS constellation. However, to stop the potential misuse of the technology, especially by foreign governments against the USA, the civilian access to the GPS signal, known as the Standard Positioning Service (SPS), contained a deliberate corruption in the signal, referred to as selective availability (SA). The US government retained a highly accurate, uncorrupted service, the Precise Positioning Service (PPS) for their military and for authorized users. The collapse of the USSR in the late 1990s led to a decline in the maintenance of the GLONASS system, and limited the Soviet/Russian ability to commercialize the system to the same extent as the USA, although GLONASS receivers for navigation were still available.

Unfortunately for agriculture, GNSS receivers used on farm vehicles were initially restricted to the lower quality SPS signal, and thus had potential errors of 100 m in the recorded coordinates because of the SA degradation. As PA developed in the 1990s, this was recognized as a major limitation to mapping and navigation, and led to the adoption of differential GPS (DGPS) systems in agriculture. These DGPSs monitored the SPS error in real-time against a known, fixed point, and transmitted a correction signal. There was a large diversity in the way that this differential correctional was transmitted but when received it enabled the position to be corrected to about a 1-m error. This made the geolocation suitable for many agricultural applications. The DGPS, however, was still not precise enough for autosteer systems or repeatable, control trafficking in fields. To achieve this, receivers that used another mode of correction, known real-time kinematic GPS (RTK-GPS) were developed to provide <0.05 m accuracy by analyzing the phase of the satellite signals and the information contained within the signal. The DGPS and in particular the RTK-GPS receivers were considerably more expensive than the stand-alone SPS receivers, but were necessary for agricultural uses.

By the late 1990s, it was clear that technological advances and commercial needs had developed effective solutions to the SA degradation and it was no longer a suitable deterrent for achieving sub-meter positional accuracy using the GPS constellation. Consequently, in May 2000, the SA degradation was removed from the SPS signal, enabling <1 m accuracy to be achieved with the uncorrected SPS signal. This SPS signal now enables accurate navigation with smartphones and is ubiquitous in our daily lives. This reduced the cost for GPS and GNSS receivers in agriculture for most applications, although for auto-steer and guidance systems, higher quality and more expensive receivers (e.g. RTK approaches) are currently still necessary for <0.1 m accuracy. To overcome this limitation, innovation continues and research into Precise Point Positioning (PPP) is ongoing to enable stand-alone GNSS receivers to resolve GNSS system errors in near real-time and to achieve high quality positioning (<0.1 m) with the SPS signal. This is being enabled by an increase in the amount of information and signals being generated by modern GNSS satellites and the access to and diversity between different constellations, i.e. the rise of multi-GNSS receivers.

Sensing systems for precision agriculture

Precision agriculture is not defined by data, it is defined by being able to make agricultural decisions at finer spatial and temporal resolutions than by current, conventional practices. However, it is certainly true that PA has been greatly enhanced by the emergence of a wide variety of sensing systems that provide agri-data and agri-information at increasingly finer spatial and temporal resolutions.

It is impossible in this short section on PA to define and list properly all of the sensing systems available for use with crop and livestock production. The intent here is to introduce the main types of agri-environmental sensing systems available and used for PA. For more information on soil-specific sensing systems in particular, interested readers are directed to other chapters and sections in the Encyclopedia.

What type of sensor?

Scientific principle - physical vs. chemical vs. biological sensors

Agri-systems are very diverse and the information desired or required by producers is equally diverse. For this reason, there have been a multitude of different scientific principles used to develop agri-environmental sensors. The most common approach is to use some part of the electromagnetic spectrum (EMS) that is known to react to specific physical or chemical properties of the target. This is particularly true for vegetation monitoring where, for example, reflectance from plants in the visible bands (red, green, blue; RGB) of the electromagnetic spectrum is related to chlorophyll content (chemical property) in the foliage, while reflectance in the adjacent near-infrared bands (NIR) is related to cellular structure (physical property) (Mulla, 2013). Optical sensors in the Visible–NIR region of the EMS are the most common sensors deployed in agriculture. However, many other parts of the EMS are also used to record agri-environmental data, for example reflectance in the radar region can also be used for biomass monitoring while measurements of naturally occurring gamma emissions are used to characterize soils. Essentially, if energy is transformed and reflected or emitted in a particular way from a target, then sensors that record the relevant part of the EMS can be used to collect data. The issue with sensors based on the EMS is that the reflectance or emission is a general response and not related directly to a specific property, i.e. the signal can be affected by multiple effects.

Alternatively, sensors can be constructed that directly interrogate a phenomenon of interest. For example, measurement of yield is based on physical sensors that either directly measure mass (load cells) or can infer it, such as impact plates that use the 'Force = mass x acceleration' principle. Similarly, when the property of interest is a chemical property, such as the soil pH, sensors based on chemical principles can be developed and deployed, provided that the sensing process is robust enough to be applied in field situations.

In addition to physical and chemical properties, sensors based on biological principles can also be deployed, although their development and commercialization is rare. For example, a sensor that measures the local microbial activity in soils has been proposed but not yet commercialized. Similarly, lateral flow devices can be used to test locally for specific chemical or biological compounds, such as enzymes or DNA sequences that are indicative of pests, weeds or diseases. In general, options for chemical and biological sensing exist, but the translation of the process into a relatively cheap and mobile platform limits their wider deployment in PA.

Proximal vs remote-sensing

In agriculture, sensors can be deployed near the target (proximally) or at a distance from the target (remotely). In many cases, especially with sensors based on the EMS, the same type of sensor can be applied at different distances or on different platforms. For example, multi-spectral optical sensing can be performed from satellite, aerial including UAVs, and terrestrial platforms. Regardless of the platform, these sensors will provide information on reflectance from the target, e.g. a crop, in the visible and NIR regions of the EMS. The difference between the sensors is in their footprint (area sensed) and temporal repeatability, i.e. their revisit time. The timing of information from satellites depends on the configuration of the constellation, while terrestrial systems can be operated when and where the operator chooses to.

Proximal sensors may be invasive or non-invasive in nature. Invasive sensors physically interact (are in contact) with the target. Soil electrical resistivity (ER) sensors are an example of an invasive, proximal sensing system that requires direct soil contact for the sensor to operate. In contrast, electro-magnetic induction (EMI) sensors for apparent soil electrical conductivity (EC_a) are an example of a proximal, non-invasive sensor, where the sensor is mounted above the soil surface. Many non-invasive sensors can also be deployed as remote sensors.

Active vs passive sensors

Agri-sensors may also be defined according to whether they are active or passive in nature. An active sensor emits some form of signal and measures the reflectance from or reaction of the target to the emission. Passive sensors, in contrast, only measure the target under ambient conditions. For example, passive optical multispectral sensors interrogate the reflectance of sunlight from the target. They cannot be deployed at night or under low-light conditions. In contrast, an active optical multispectral sensor will have its own light source and measure the reflectance of the emitted light from the target. These sensors do not depend on the presence of ambient light and can be operated at any time, including at night. Soil EMI or ER sensors are both active sensors, while a gamma-radiometer is a passive soil sensor that measures the emission of gamma radiation from the natural breakdown of isotopes within a soil.

A sensor of a property or a response?

Finally, in an agricultural context, it is important to understand if the sensor is directly measuring a soil (or agronomic) property or a response. For many agri-environmental situations, information is needed on a specific soil (or crop) property for decision-making. Some sensors measure (or estimate) a specific property but many, including the most common sensors, only measure a more general response, which is linked to a property of interest. For example, soil EMI or ER sensors are widely used to map soil variation, with larger EC_a values typically attributed to stronger textured soils (more clay). However, these sensors directly measure the response of the soil to an electro-magnetic impetus, which is only influenced by textural properties. Soil EC_a is not a direct estimate or measurement of soil particle-size distribution and is affected by other soil properties, including moisture content, temperature, soil pH etc. Consequently, the EC_a response requires calibration that will be specific locally to convert the EC_a response to a soil property (e.g. clay percentage). The EC_a cannot be assimilated directly into models or decision support systems as a surrogate for texture without such a calibration. Some sensors, e.g. on-the-go soil pH sensors do measure a soil property directly (pH in this case) and thus the sensor response could be used directly in models requiring soil pH information. When considering the use of sensor data, a clear understanding of whether the sensor measures the property of interest or is a surrogate or a more general response is important to understand the limits to the use of the sensor data in decision-making.

What type of data or information?

Soil

On-the-go, high-spatial resolution proximal soil sensing is dominated by systems that measure the apparent soil electrical conductivity (EC_a) (Corwin and Lesch, 2005). This can be achieved by either electro-magnetic induction (EMI) sensors or electrical resistivity (ER) sensors. Several commercially and widely used EMI and ER systems exist worldwide (some examples are shown in Fig. 4). An EMI sensor emits a primary electromagnetic field, which generates a secondary field within the soil system, and the interactions between the two can be used to generate an estimate of EC_a . For ER systems, dipoles are inserted into the soil (usually as a rolling disk) and a current is introduced. The drop in voltage between dipoles at fixed differences is a measure of soil resistivity, the inverse of conductivity (EC_a). By changing the distance between the emitting dipoles in either system, changes in the depth of exploration (DoE) can be made. Therefore, both systems have the potential to measure at multiple and various depths. For PA, this is particularly useful as the EC_a response of the topsoil and subsoil can be measured by fairly compact, mobile, manageable sensing systems. The advantage of the EMI sensor is that it is contact free, thus it can be elevated above the soil surface to change the DoE or to change the mounting platform, i.e. aerial systems to measure inaccessible areas or boat-mounted systems to sense along waterways are possible. The ER systems are limited by the need for good soil contact for the current to pass into the soil; this creates issues particularly in dry, hard and or stony soils. In contrast, EMI systems have problems in soils with very poor conductivity (i.e. very sandy conditions). As mentioned previously, EC_a is a soil response that is strongly influenced by texture and moisture content, but is also affected by soil temperature, pH, CEC (cation exchange capacity) etc. Mapping texture is very important for PA, to understand where production will differ.



Fig. 4 Images of some commonly used, commercially available apparent soil electrical conductivity (EC_a) sensors. (A) the Veris 3100 (Veris Technologies, Salina, KS, USA), which is a proximal, invasive sensor based on electrical resistivity; (B) the Geonics EMS8DD Mk2 sensor, which is a proximal, non-invasive electromagnetic induction (EMI) sensor (Geonics Ltd., Mississauga, ON, Canada) and (C) the DualEM-1s, which is also a non-invasive, proximal EMI sensor (DualEM Inc., Milton, ON, Canada). Note: Other, similar sensing systems are also available.

In addition to EC_a sensors, the other two available on-the-go proximal sensors are gamma-radiometers (GR) and ground penetrating radar (GPR). Gamma-radiometers are useful for mapping soil types, on the assumption that different soil types have a different chemical signature that generates a different profile of gamma emissions from the soil (Wong and Harper, 1999). In many cases this is related to the local pedology and parent material of the soil. The GR contains a crystal that collects and attenuates the gamma emissions so that they can be recorded as an energy spectrum. This spectrum is then processed to identify the total count of emissions and emissions linked to key isotopes (K (potassium), Th (thorium) and U (uranium)). Patterns in EMI and GR may be similar if the variation in soil texture aligns with the pedology, however, they often diverge as EC_a is affected by multiple soil properties, including texture, but is less affected by the chemical composition of the soil (its parent materials) (Castrignanò et al., 2012). Furthermore, different pedological processes and parent materials can generate a similar texture profile. The GPR systems measure the vertical reflectance of radar waves into the soil and can distinguish abrupt changes in density within the profile (Lombardi et al., 2022). These sensors are particularly useful in situations where the depth of a contrasting horizon is needed, for example the depth of a peat soil over the in-situ soil profile, or for mapping artifacts, such as tile drainage lines. The use of GPR to date has been limited in PA, both in research and commercially. Recent advantages in signal processing are showing more potential uses, especially for mapping subsoil constraints such as plow pans and stone lines. The ability to map the potential rooting depth, which can have a major effect on plant available water and crop yield, is the key advantage of GPR systems over other soil sensors.

Soil pH and soil organic matter are two key soil properties of interest to agricultural production. Neither of these has a strong effect on the previously listed sensors (EMI, GR or GPR) and there has been considerable research effort put into the development of dedicated soil pH and carbon (C) or organic matter (OM) sensors. Soil pH mapping has been successfully done with ISFETs (ion selective field effect transistors) or with electrodes (similar to pH measurements in a laboratory) (Archbold et al., 2023), and this latter approach has been commercialized. Soil C is typically estimated by a near-infrared (NIR) spectrometer, with the spectral response calibrated to the soil C or OM content (Kuang and Mouazen, 2013). This is typically mounted on a plow-like tine that enters the soil at a fixed depth and scans the soil as the 'tine' is pulled through the soil (Hodge and Sudduth, 2012). Again, this sensor has been commercialized. Although commercial soil pH and C sensors are available, and have been since the early 2000s, uptake is still limited. This is mainly due to reliability of the sensors, for example the soil pH sensor does not work well in hard or stony soil types, or they provide limited information, for example both sensors give only soil pH or C at one fixed depth (which is usually in the topsoil) and subsoil information, particularly for pH is unavailable. Furthermore, for pH mapping, the real agronomic decision (lime requirement) depends on the soil pH and the buffering capacity of the soil, and this latter information is usually unavailable from on-the-go soil pH sensors. These two sensors highlight some of the limitations for soil mapping in PA, namely that the on-the-go sensor does not always provide the desired agronomic information or at the right resolution or depth for production decisions.

From a remote-sensing perspective, options are available to measure and map soil from aerial and satellite systems. This has been applied predominantly to topsoil mapping using multispectral or hyperspectral sensors over bare soil (Mulla, 2013). Under bare soil conditions, good relations with soil type, texture and carbon content can be generated with local calibration and ground-truthing. This is effectively relating soil color to soil properties. Several commercial companies sell soil information based on bare soil imagery, however the ability to do this changes as crop and soil management changes. The adoption of direct-drilling into no-till and even min-till systems will limit this technology as it cannot work if the soil is not tilled. More recently, with the rise in availability of satellite-based, active radar systems and in the academic knowledge of how to process and interpret radar imagery, there is an increasing interest in mapping soil moisture from space. While this technology is not yet mature, the need for such information is likely to encourage rapid development in this area. Finally, as mentioned above, EMI and GR are passive, non-contact systems and can be mounted on aerial systems to survey large areas and in the case of EMI, to large depths (> 5 m).

In addition to on-the-go soil sensors for mapping, PA also relies on networks of static, usually low-spatial resolution soil sensors. This is particularly the case for soil moisture sensing using capacitance, time-domain reflectometry or, less commonly now, neutron-probe sensors. For PA, it is usual to use the high-resolution soil information discussed above (often together with crop data) to stratify static soil sensors into commercial production systems. This ensures that information is gathered at points relevant to production, e.g. areas of high or low productivity.

Crop

For PA, there are many sensors aimed at monitoring and measuring crop attributes, predominantly vigor or biomass or yield. The most common sensors applied to crops are optical multi-spectral sensors. These measure how visible and near infrared (NIR) light reflects from the canopy. Differences in reflectance indicate different levels of vigor. As indicated previously, these multi-spectral sensors generate a generic response of relatively higher or lower vigor. The reason for the differences in vigor (over space or over time) cannot be determined directly from a multispectral imaging system. It may be related to disease or to environmental conditions, and in nearly all cases it is likely that there are multifactorial environmental and management effects that contribute to the relative vigor of a crop at a given time in a given space.

If the spectral resolution of the sensors is increased, from a multi- to a hyper-spectral system, then there is considerably more information in the spectral signal and it becomes possible to extract some specific information pertaining to crop health and crop nutrition (Mulla, 2013). For example, hyperspectral information can be used to determine the local presence of specific crop diseases on the foliage that affect overall crop vigor. Likewise, nutrition deficiencies in the leaves may be expressed differently in hyperspectral imaging. Hyperspectral imaging systems are becoming increasingly more available (lower cost, more robust) for field use and are expected to become more common on-farm in the coming decade.

There has been a sharp rise in the use of normal (RGB) imagery, coupled with machine-learning techniques, for site-specific crop management. Imagery from everyday camera systems, including mobile phone cameras, can be processed with trained algorithms to generate useful agronomic information, and the potential applications are many and varied. A particular focus has been put on local estimation of crops with this approach to count tiller heads automatically in cereal systems (e.g.; Fernandez-Gallego et al., 2020), bunches or individual berries in grapes (*Vitis* sp.) (Nuske et al., 2014), individual fruit in arboriculture or even potato (*Solanum tuberosum* L.) tubers if plants are dug up mid-season. In addition to counts, these systems can also estimate size ranges (quality distributions) if the imagery is properly corrected (Fig. 5). Identification of defects or diseases on the crop or the foliage or just estimates of leaf area is also possible in real-time in the field (e.g. Richey et al., 2020). The technology is not yet perfect and some issues remain, particularly with errors associated with the occlusion of the crop by the foliage and a need for better standardization and or harmonization of image collection. Despite this, better estimation and site-specific estimates of yield will transform crop estimation and management. This will be especially important in high-value crops and for crops where robots will be more commonly employed.

The other dominant crop sensors deployed in PA are yield sensors. These can be either harvest-mounted sensors that audit production as it is harvested or some form of imaging system that counts or records yield near harvest (i.e. a near-harvest crop estimate as described above). On-harvester yield sensors are one of the foundation technologies in PA and particularly in arable cropping (example in Fig. 6). It was the collection of these data and the observed large spatial variation in yield that encouraged much of the initial interest in PA. Nowadays, yield sensing systems exist for most machine-harvestable crops. Tracking systems also exist for hand-harvested crops, but these have much less market penetration and acceptance than on-harvester yield monitors.



Fig. 5 Example of a next generation in-season grape sensor based on image analysis and machine-learning algorithms. The sensor (A) images the grapes on-the-go as it is driven through the vineyard. Image analysis processes are then used to count the visible berries, estimate the average berry size and to determine berry color (B). This can be used mid-season to manage crop load in vineyards, or near harvest to determine if grapes are ready for harvest (meet the market demands for quality). Images courtesy of the Efficient Vineyard project (Dr Terence Bates, Cornell) and the Carnegie Mellon Robotics group.

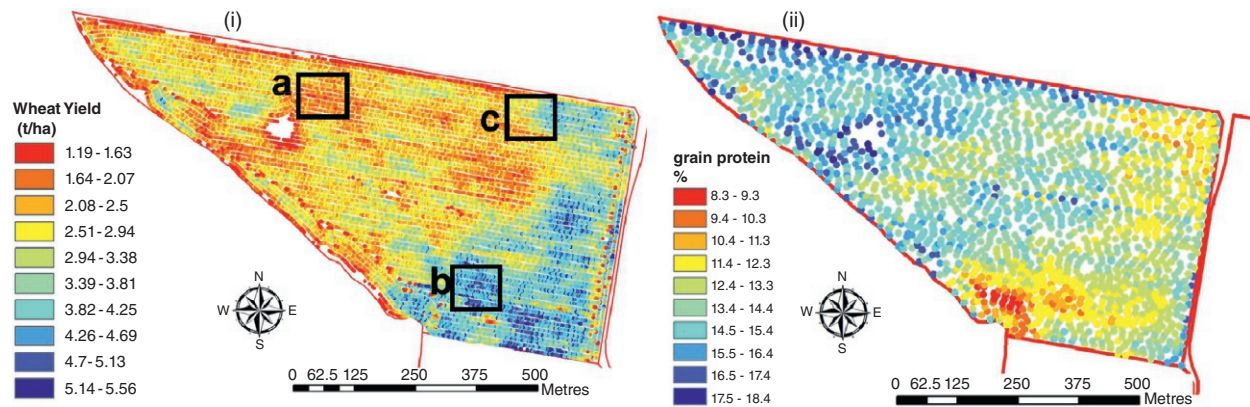


Fig. 6 Examples of yield (i) and grain protein content (ii) in a wheat field from a farm in north-west New South Wales, Australia. Both maps show the raw data and clearly show the inverse spatial relationship between yield and protein in this field in this year (high yield - low protein and vice versa). The squares in the yield map show 1-ha areas of consistently low yield (a), consistently high yield (b) and variable yield (c), highlighting the differences that can be found in 1-ha areas with regard to the amount and variation in yield. Reproduced with permission from Whelan BW and Taylor JA (2013) *Precision Agriculture for Grain Production Systems*. CSIRO Publishing, Dickson, ACT, Australia.

Many different types of yield sensors are available but they tend to be based on a few key principles. These are (i) directly weighing the crop using load cells under a conveyor belt or collection bin, (ii) measuring the force exerted by the crop on an impact plate (at a known acceleration) so that mass can be calculated, or (iii) counts associated with the disruption in a light beam(s) as the crop passes through a pipe or tube. All three approaches have advantages and disadvantages depending on the manner of harvest and type of crop, however, if properly calibrated and used, they will all generate useful data. For many harvesters, especially combine harvesters, yield monitors are standard nowadays.

Quality sensors also exist for some crops, notably for protein in cereal crops (Fig. 6), but are absent in many others, particularly for root vegetables. Grain protein sensors use NIR spectroscopy in real-time to assess grain quality on harvesters, but this form of sensing is less common than for yield (quantity) sensing (Long and McCallum, 2020).

Animal

Precision agriculture is not just limited to crops, but can be applied equally to livestock industries. Since animal production depends on feed production, many of the soil and crop technologies discussed above can also be deployed in forage and fodder production systems. While the potential exists, the commercial (and research) development has been slower. Sensing in pastures is complicated by species diversity and the temporal evolution in the dominant species over the year (in contrast to monoculture situations in arable and horticultural crops). This makes decision-making more complicated and, even with the same technologies, different decision processes are needed. The perceived value in these has had limited investment to date (relative to arable crops). There has been very little pasture-specific sensor development to support precision pasture management, except that of a yield sensor for forage harvesters. Some existing NIR sensors have been commercialized and can be calibrated to provide real-time quality (protein, sugars, fiber, etc.) information during silage or hay operations.

In addition to pasture sensing, many different types of animal-specific sensors are available to provide information on animal movements, animal feeding habits, rate of animal growth and general behavioral or well-being. The use and availability of these is increasing rapidly.

Meteorological

Agricultural production, except for controlled atmosphere systems like growth chambers and sealed greenhouses, depends on the weather. Accurate observations and forecasts of weather conditions are very important for good agriculture management. First, such information is critical to the use of many different types of models to support decision-making, e.g. epidemiological models of pest life cycles and optimal times for pest control interventions or modeling crop phenological development for the optimal timing of plant nutrition. Second, it is important to have a correct (and up to date) understanding of water supply (precipitation and evapotranspiration changes) relative to the needs of the crop for yield potential and evolving in-season crop management to meet this potential. Last, correct short-term forecasts can be critical for achieving timely operations, especially leading to extreme or large weather events.

On-farm meteorological data can be obtained from fixed weather stations, and these are often commonplace on large farms. However, on large farms there can be a considerable difference between the weather in different fields, if not in the same field. This is especially true for precipitation events that can be very localized. Therefore, a fixed weather station near the farmhouse may not be

completely relevant to the entire production system and will be unable to measure meso-climatic variations across a farm. Areas prone to frost, such as hollows, or with a different aspect, could experience very different local weather conditions. Therefore, care is needed when extrapolating some weather information across farms.

A modern alternative is to use virtual weather stations based on large-scale weather models that can provide 'local' (typically 1 km²) predictions of hourly or daily weather conditions. These observations are predictions, but very short-term (typically 1 h predictions) that have been up-dated and corrected based on recent observations at a network of connected weather stations. For some key properties, particularly temperature, they are usually very accurate and are suitable for use in many agriculture models (Launspach et al., 2017). Local predictions of precipitation can still be problematic. These virtual weather stations do provide growers with an ability to obtain weather data from different meso-climates across their production systems and to have more accurate field-specific information (compared to a single, central fixed weather station).

The role of interpolation, pedometrics and digital soil mapping in precision agriculture

As outlined in the introduction to this chapter, PA has developed from the idea of '*farming by the soil*', and was informed by an evolving understanding in agriculture that soils are variable in space (and time) and that this variation can be mapped and used to make better, spatially-variable agronomic decisions. This was predicated on the ability to observe variation (i.e. the existence of new sensors) and to deploy differential (variable-rate) management in fields (i.e. developing variable-rate machinery capabilities). In addition to these technical capabilities, there was also a need for methodologies to analyze and process properly the new spatial data structures that were available to agronomists. If these data were not correctly transformed into information layers, then effective agronomic decision-making is not possible. From the late 1980s to the mid-late 1990s, the academic and industry knowledge base to process these data lay predominantly in the soil science sector.

The knowledge that soil is a variable continuum in both time (generally long periods) and space (generally short distances) has long been understood, from the seminal work of Vasily Dokuchaev in the late 1800s, through to the formalization of soil forming factors in the mid 1900s by Hans Jenny and others, leading to modern day digital soil mapping methods (McBratney et al., 2003). It is unsurprising that the spatial analysis of soil data has been strongly advocated in the soil science and geological communities. Such spatial analyses are generally referred to as geostatistical analyses. Geostatistics relates to the application of mathematics to spatial issues. In general, this is either to describe the amount of spatial variation observed in a data set (Leroux and Tisseyre, 2019) or to interpolate spatial data by predicting at places where observations have not or could not be made. Geostatistics can be used to generate maps of soil types or properties, as well as maps of production attributes (e.g. yield). Correct use of geostatistics in this manner transforms raw data into visual maps that help growers and agronomists to understand the amount and spatial distribution of the soil and crop variation on-farm (Fig. 7).

The use of mathematics and statistics in soil science has a specific term, 'pedometrics.' It is important to note that pedometrics relates to any use of mathematics or statistics in soil science, not just to spatial analyses. However, a large part of the pedometrics community, historically and nowadays, are invested in geostatistical questions. This derives from a need to survey and to map soil for many different purposes (agriculture, hydrology, civil engineering projects, environmental habitats etc.). Initially, soil surveys and soil maps were typically done at quite broad scales as it was cost-prohibitive in most cases to collect spatially dense (high-resolution) soil samples manually. However, as sensor and information technologies have evolved over the past 40 years, the scale of mapping and the type of mapping has also evolved in response to access to high-resolution spatial soil data. Soil mapping is now done almost exclusively in digital form, termed digital soil mapping (DSM), using predominantly quantitative analytical

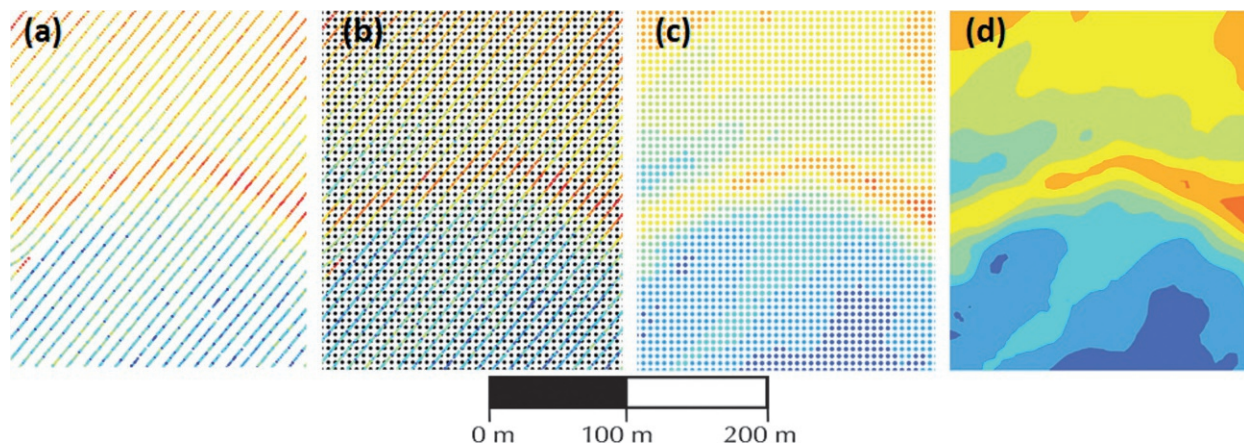


Fig. 7 The way that maps are produced for PA has been heavily influenced by developments in pedometrics and digital soil mapping. This example shows some raw data (A), which is then overlain by a prediction grid (B), and onto which values are interpolated from the raw data (C) to form a smoothed map (D). The interpolation process filters errors in the data and smooths the patterns of variation to make them visually more interpretable by agronomists and growers.

approaches applied to high-resolution data sets. This has replaced subjective, traditional soil survey and mapping techniques that often relied on expert knowledge and were informed by limited numbers of soil observations.

Pedometrics predates PA. Before the development of Earth observation satellite systems (1970s), spatial crop information was very limited and there was little knowledge or application of geostatistics to crop sciences before the 1990s. Yield monitors and satellite imagery changed this, and resulted in a need to rapidly develop the capacity for spatial crop data analysis in agriculture. These spatial yield and biomass data were complemented by high-resolution soil data and soil mapping, and it was a natural step for the geostatistical expertise applied to soil data to be shared and transferred to crop data. The result was that many leading pedometricians in the 1990s also became leading advocates and researchers in PA, notably Margaret Oliver (Reading, UK) and Alex McBratney (Sydney, Australia). During this period, it was typical for major conferences to have plenary talks from pedometricians and there were often multiple sessions dedicated to the issue of high-resolution soil measurements and mapping for PA. Soil scientists made up a large part of the original academic PA community.

As PA developed (late 1990s and early 2000s), the demand from the cropping community for better and higher resolution soil information further encouraged developments and innovations in DSM. Effectively, it changed the target scale for DSM from hundreds of meters to tens of meters. This was of great benefit for the DSM community and also exposed DSM to many different circumstances where high-resolution soil information is needed. The result was a rapid increase in research and capacity in pedometrics and DSM, and the creation of a sustainable, stand-alone academic community. Consequently, the direct role of the pedometrics community in PA has diminished over the past decade, although interactions and shared research needs persist. The mathematical and statistical knowledge applied to PA from 'traditional' geostatistics in soil science is now being replaced by machine-learning and artificial intelligence (AI). While this has brought more computer scientists into the PA community, pedometrics has also evolved to provide knowledge to PA domains, particularly through DSM applications.

Evolving, or previously under-developed areas where PA still needs considerable input from soil scientists and pedometricians are improved sensing systems of specific soil properties, especially those associated with plant nutrition and plant available water. For several soil properties, particularly those associated with soil chemistry and soil water, this information is needed both temporally and spatially. Nitrogen and water fluxes can occur quickly during a season. This is not true for all properties because physical properties such as soil texture evolve on time scales that are irrelevant for agriculture. Furthermore, better spatial information in the vertical (depth or *z*) direction is also needed. To date, spatial variation has primarily been addressed in the *x* and *y* dimensions.

Decision-making and management units in precision agriculture

Decisions in agriculture rely on the quality of data (and information) gathered, are affected by the economic and production goals of the system (farmer), are required for a spatially and temporally varying biological system, and depend on previous and future management decisions. In addition, the majority of agriculture exists within an open biological environment. When production interacts and reacts with variable environmental and biological conditions, production becomes temporally and spatially variable. In short, agri-decisions are made in an imprecise system with few situations where there is an absolute and correct decision. From a PA context, the amount and nature of variation in a production system is a key determinant of the direction and opportunity for making decisions, determining the (absolute) quality of those decisions and imposing differential management. Optimizing spatial decisions-making and demonstrating improved efficiencies is the biggest challenge to the further development, acceptance and translation of PA.

High-resolution spatial data sets are still relatively new in agriculture and the way that the data and information are presented is still relatively novel to many producers and service providers in agriculture. Interpreting maps of production attributes is not always intuitive. Individually, single biomass images, yield maps or soil maps are relatively easy for producers and agronomists to interpret and to superimpose their own knowledge on to. This is particularly true when there are clear management and/or environmental effects in the data or information, as can be seen in some of the yield maps in Fig. 1. As more layers of data and information become available for a field, the ability to process the additional data or information diminishes, i.e. interpreting and reacting to the time-series in Fig. 1 is more complicated than reacting to the pattern in any given individual year map. It becomes even more complicated if some of the raw data are poorly presented or poorly mapped.

If PA is about better decision making, then it is paramount that the way that data are presented, transformed into information layers and how they are fused into a decision process is done correctly. The more confusion that is generated at the initial stages, the harder it is to arrive at a good decision at the end. Approaches to filtering, mapping and integrating data have been key areas of research from the inception of PA (Taylor et al., 2007). Without these basic building blocks, making decisions and therefore PA adoption and translation are not possible.

Filtering and mapping approaches continue to be updated, but the transformation of raw data into information layers is relatively well understood by academic and industry stakeholders. How these data are then compressed, fused or simplified remains an issue for the industry. Methods, and well-accepted approaches exist, however on-going evolution in the amount and type of data available is constantly challenging the way that these data can be integrated and interpreted (Lyle et al., 2014). Data fusion and the transfer of relevant information into decision systems remains a key area of development for PA (Guillaume et al., 2020). The more effective such methods are, the better the resulting decisions should be.

With improvement in sensing, algorithms, decision-making and operational capabilities, the industry will become closer to real site-specific (i.e., plant or even leaf-specific) management (Fig. 2). In the interim, however, it has been widely recognized that some intermediate resolution of management is needed (Khosla et al., 2002). This is generally referred to as the management zone or management class concept. The original term proposed was ‘management units’ (Lark and Stafford, 1997); however, the term ‘management zones’ is more commonly used in PA to describe delineating a field into smaller subfield areas for differential management operations. The delineation of management units by classification or segmentation algorithms has proved to be very effective for data fusion with different agri-data sets. The most common method used has been the *k*-means classification algorithm or a variant thereof (e.g. Fridgen et al., 2004; Taylor et al., 2007). This requires data to be co-located (either sensed or measured at the same place, and or interpolated to fixed locations) which enables *k*-means classification to effectively partition multiple datasets into zones where the input data sets exhibit a similar response. The agronomic and environmental response within a delineated zone can then be interpreted with local agronomic knowledge to devise zone-specific management plans. The advantage here is that existing, local knowledge can be easily brought into the decision-making, albeit at the end of the data fusion process.

Management units (MUs) play a key role in getting producers involved in PA. When done correctly, the producers can see how the various, often complex production layers are compressed into a single, usable map that can be interpreted with their existing local knowledge of crop and soil responses (Fig. 8). Individual MUs can be treated as virtual, fenceless fields (Taylor et al., 2007). Decisions follow a field-scale process, i.e. the general agronomy process of the grower or agronomist remains unchanged, but is applied at a smaller scale (i.e. the MU instead of field scale). This achieves a higher level of spatial management while keeping the producer in a familiar decision space. If poorly constructed, MUs will be uninterpretable, the producer will be disillusioned with PA and the default uniform field management will continue to be used. As a result, a large part of PA research has been directed at issues around generating correct MUs (Betzek et al., 2018). While the majority of the work to date has focused on arable cropping systems, tools developed for MU delineation are as relevant to perennial horticultural systems as they are to arable cropping systems.

The delineation of MUs by classification was a big step forward in multi-layer data fusion. However, the current use of MUs tends toward static patterns and they are applied as a ‘one size fits all’ approach. Fixed zones, with fixed boundaries, are used for all potential variable-rate applications within a field – for example planting, chemical inputs and fertilizer. This is (was) sensible with only limited information layers available. With more information, particularly in-season temporal information, becoming increasingly available for both crop and soil properties, there is a new opportunity to tailor MUs to specific operations and times in the season, i.e. MUs should be dynamic in space and time (Fig. 9). For example, the controlling factors of variable plant establishment will probably differ from those of variable crop development in any given field. In such cases, MUs for variable-rate seeding and fertilizer application should differ. However, it is only recently that data and information availability, linked with improved connectivity, has reached a level where the potential of dynamic MUs is a reality.

The potential for dynamic MUs, generates a new paradigm for essentially the same set of questions and challenges. How are the right data chosen in an automated way for these dynamic MUs? Which method of data fusion should be used to generate MUs that evolve with decisions and or time? Currently, with fixed, static MUs, these are delineated by an expert in consultation with the producer. This ensures a rigorous selection of input data and a standardized method of delineation. A shift to dynamic MUs will make this approach economically unsustainable. Service suppliers will be unable to provide the level of personal contact needed to be altering MUs constantly with evolving seasonal conditions and for disparate management operations. It will be the producer and or an automated system created by AI and machine-learning approaches that is likely to generate these dynamic MUs in the future. The current (and future) expert knowledge of both the agronomist and producer in selecting the relevant data for specific operations, and the method of MU delineation needs to be captured within an autonomous (self-learning) model to allow producers to derive MUs with confidence by themselves (Taylor et al., 2021).

This is not an easy objective. This shifts MU delineation from an information intensive system to an embedded knowledge system. More robust and repeatable data fusion methods are needed to replace the current use of *k*-means classification (and its derivatives). Ideally, these multivariate data fusion techniques should operate directly on the raw data and include filtering and pre-processing steps. Machine-learning techniques, particularly techniques that enable soft-computing solutions, are likely to be at the forefront of this approach to MU delineation (Fig. 9).

Precision vs. digital agriculture

The increasing level of ‘connectivity’ on farms and the need for on-going, and likely increased, interactions with computer science is a reality for modern agriculture. The digital revolution is affecting all aspects of modern life and agriculture is no exception. Digital agriculture (DA) will precipitate some disruptive changes in the food production and supply chains in the near future. It is a challenge for PA researchers, and the users of PA, to understand and utilize digital agriculture innovations to improve spatio-temporal crop (and animal) management (Bellon Maurel et al., 2022).

Precision and digital agriculture are connected, and often the latter is described in a similar way to PA. However, these are different concepts. Precision agriculture is now well-defined and concerns better (higher resolution) decision-making. Digital agriculture is still new, but it relates first to the collection and the connectivity of digital data structures. It should implicitly feed data and information into PA systems to improve them (Wolfert et al., 2017). However, on its own, DA is not a holistic decision support system. It is not concerned, for example, with non-digital data that could influence farm management. Its decision support derives directly from information and communication technologies. Digital agriculture also has a much larger scope than PA. It is

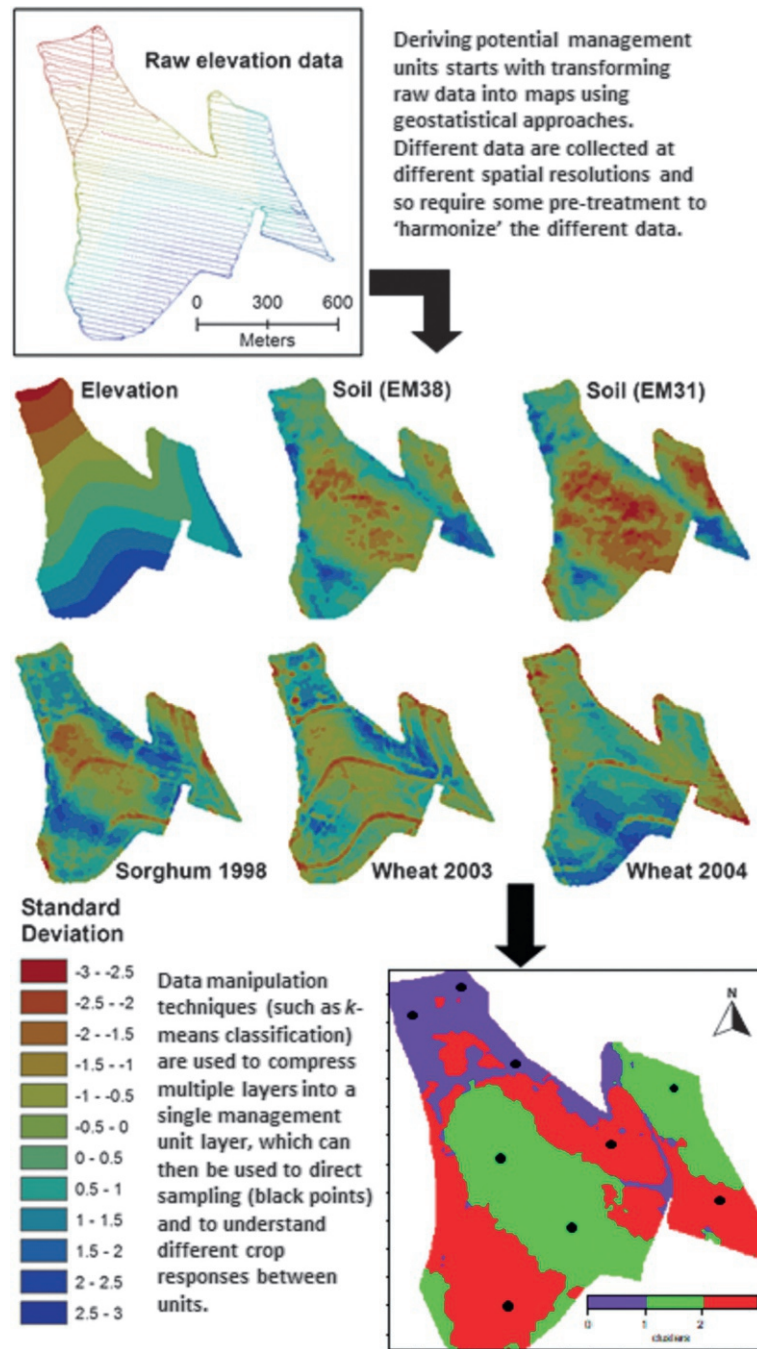


Fig. 8 A visual illustration of the general process to transform raw irregular crop and environmental data into single layer maps and the subsequent fusion of these layers into a potential management unit (MU) map (in this case using *k*-means classification). Different MUs have different yield responses. The MUs were used to target soil sampling (3 samples per MU) to explain the different yield responses. Adapted from an image first published in CSA News, which is an industry magazine that supports research published within ASA, CSSA and SSSA journals.

equally concerned with non-spatial and non-temporal data structures. This is exemplified by the fact that PA is farm and production-centric while DA is a farm to fork concept. Digital technologies in the supply chain will be a large part of the DA revolution, and will have a major role to play in evolving food security and human health issues, but these solutions may not necessarily be relevant to on-farm production management.

Finally, it is important to contextualize the role of robots and unmanned aerial vehicles (or systems) (UAV or UAS) within PA. Robots or UAVs by themselves are not PA. They are platforms from which information can be gathered or from which a decision can be implemented. Essentially, they are the next iteration in agri-machinery and are replacing existing farm machinery for certain

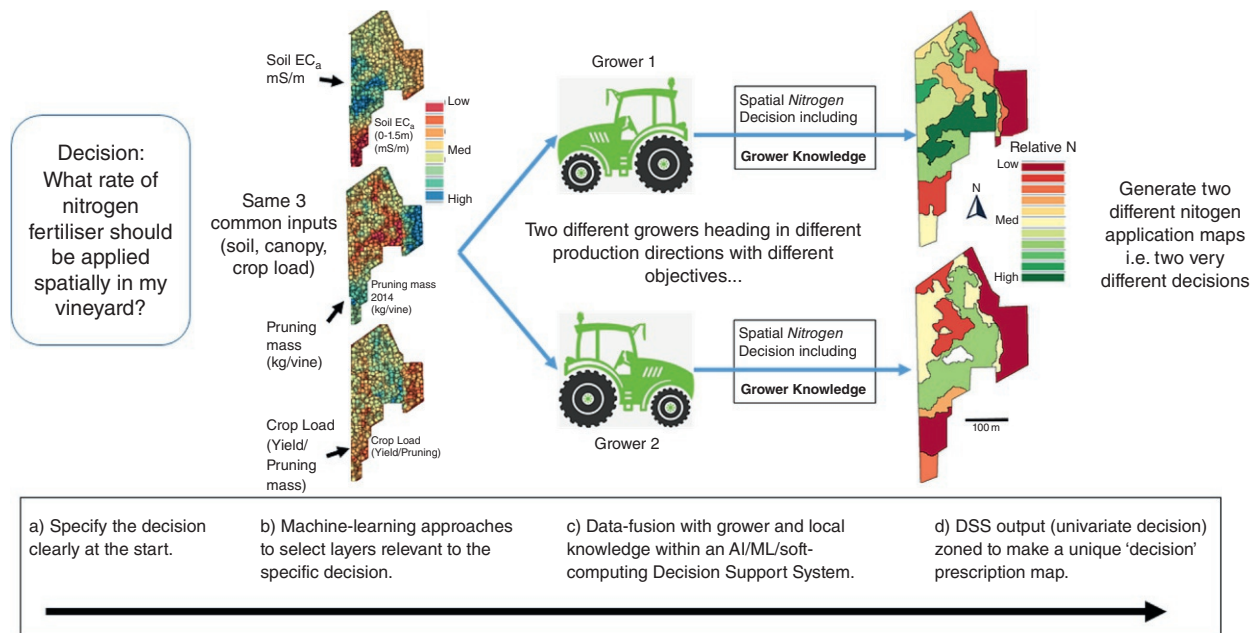


Fig. 9 A schematic diagram of how a change from fixed to dynamic decision zoning could operate. The process starts by (A) specifying the decision first, (B) then selecting data layers relevant to the decision (from all available layers), (C) before using a dedicated data-fusion tool that includes grower knowledge and (D) is 'zoned' only on the output (decision) from the decision support system. Adapted from Guillaume S, Bates TR, Lablee J-L, Betts T, and Taylor JA (2020) Combining spatial data layers using fuzzy inference systems: Application to an agronomic case study. In Proceedings of the 6th International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM 2020), 62–71, with permission.

tasks. They are and will continue to become an important part of modern and future agriculture. However, a robot could be used to implement a uniform, flat management strategy, for example for seeding a crop, as a straight swap for a tractor-based approach. This is not PA. It becomes PA if the robot enables better decision-making at a finer spatio-temporal scale than is the norm. Platforms that host sensors or enable variable-rate management are enablers of PA. Robots and UAVs are the next generation of platforms. The PA component comes from the evolving management strategies that will use the data or information collected from these new platforms that can potentially be implemented via these platforms. Autonomous robots, in particular, have a great potential to change the scale of data collection and decision-making in agricultural fields if the analyses, decisions and actions can be effectively linked 'on-board.'

Concluding comments

Precision agriculture started with the intent to 'farm to the soil,' with the recognition that soil is a spatially variable continuum and that its impacts on production are also spatially variable. This continues to be the building block on which site-specific crop management and precision pasture management have developed. To this end, soil science will always be fundamental to PA. However, as growers can understand and manage their soil variability better for their production goals to negate these effects on production variability, then other factors and interactions become more significant and have noticeable effects on production. This is the situation that PA in developed agricultural countries now finds itself. Consequently, the PA academic community has shifted from a soil science base to a computer science base over the past 10 years, particularly at the International and European PA conferences. However, in regions where PA is less developed, especially in African agriculture, the PA community and its objectives remain more strongly aligned with soil science and managing soil variability. The resurgence of soil science needs for small-holder agriculture has been a key theme of the first two African PA conferences in the 2020s.

Pedometrics and digital soil mapping have played a crucial role in the evolution of PA over the past 30 years. The continued need for more and better soil information and at a lower cost will ensure that it continues to be a part of the next evolution in PA.

References

- Adrian AM, Norwood SH, and Mask PL (2005) Producer's perceptions and attitudes toward precision agriculture technologies. *Computers and Electronics in Agriculture* 48(3): 256–271.
- Ancev T, Whelan B, and McBratney A (2005) Evaluating the benefits from precision agriculture: The economics of meeting traceability requirements and environmental targets. In: Stafford JV (ed.) *Precision Agriculture '05. Proceedings of 5th European Conference on Precision Agriculture*, pp. 985–992. Wageningen Academic Publishers.

- Archbold G, Parra C, Carrillo H, and Mouazen AM (2023) Towards the implementation of ISFET sensors for in-situ and real-time chemical analyses in soils: A practical review. *Computers and Electronics in Agriculture* 209: 107828.
- Arnó J, Casasnovas M, Ribes-Dasi M, and Rosell-Polo J (2009) Precision Viticulture. Research topics, challenges and opportunities in site-specific vineyard management. *Spanish Journal of Agricultural Research* 7: 770–779.
- Bellon Maurel V, Brossard L, Garcia F, Mitton N, and Termier A (2022) *AGRICULTURE and Digital Technology: Getting the Most Out of Digital Technology to Contribute to the Transition to Sustainable Agriculture and Food Systems*. Paris, France: INRAE-Inria185.
- Betzek NM, Souza EGD, Bazzi CL, Schenatto K, and Gavioli A (2018) Rectification methods for optimization of management zones. *Computers and Electronics in Agriculture* 146: 1–11.
- Blasch G, Li Z, and Taylor JA (2020) A novel pattern recognition approach for the delineation of yield productivity-stability zones using yield map time series. *Precision Agriculture* 21: 1263–1290.
- Castrignanò A, Wong MTF, Stelluti M, De Benedetto D, and Sollitto D (2012) USE of EMI, gamma-ray emission and GPS height as multi-sensor data for soil characterization. *Geoderma* 175–176: 78–89.
- Cook SE, O'Brien R, Corner RJ, and Oberthür T (2003) Is precision agriculture irrelevant to developing countries? In: Stafford JV and Werner A (eds.) *Proceedings of the 4th European Conference on Precision Agriculture*, pp. 115–119. Wageningen Academic Publishers.
- Corwin DL and Lesch SM (2005) Apparent soil electrical conductivity measurements in agriculture. *Computers and Electronics in Agriculture* 46(1–3): 11–43.
- Fernandez-Gallego JA, Lootens P, Borra-Serrano, et al. (2020) Automatic wheat ear counting using machine learning based on RGB UAV imagery. *The Plant Journal* 103(4): 1603–1613.
- Fridgen JL, Kitchen NR, Sudduth KA, et al. (2004) Management zone analyst (MZA): Software for sub-field management zone delineation. *Agronomy Journal* 96: 100–108.
- Guillaume S, Bates TR, Labee J-L, Betts T, and Taylor JA (2020) Combining spatial data layers using fuzzy inference systems: Application to an agronomic case study. In: *Proceedings of the 6th International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM 2020)*, 62–71.
- Hodge AM and Sudduth KA (2012) *Comparison of Two Spectrometers for Profile Soil Carbon Sensing*. ASABE, 121338240, St Joseph, MI, USA.
- Khosla R, Fleming K, Delgado JA, Shaver TM, and Westfall DG (2002) USE of site-specific management zones to improve nitrogen management for precision agriculture. *Journal of Soil and Water Conservation* 57(6): 513–518.
- Kuang B and Mouazen AM (2013) Non-biased prediction of soil organic carbon and total nitrogen with vis-NIR spectroscopy, as affected by soil moisture content and texture. *Biosystems Engineering* 114(3): 249–258.
- Lagos-Ortiz K, Medina-Moreira J, Sinche-Guzmán A, et al. (2018) Mobile applications for crops management. In: Valencia-García R, Alcaraz-Mármol G, Del Cioppo-Morstadt J, Vera-Lucio N, and Bucaram-Leverone M (eds.) *Technologies and Innovation. CITI 2018, Communications in Computer and Information Science*, vol. 883, 57–69.
- Lamb DW, Frazier P, and Adams P (2008) Improving pathways to adoption: Putting the right P's in precision agriculture. *Computers and Electronics in Agriculture* 61(1): 4–9.
- Lark RM and Stafford JV (1997) Classification as a first step in the interpretation of temporal and spatial variability of crop yield. *Annals of Applied Biology* 130: 111–121.
- Launsbach M, Taylor JA, and Wilson JM (2017) Can temperatures from an online weather forecast service be suitable for modelling growth stages using a CERES-Wheat type phenology model? *Advances in Animal Biosciences* 8(2): 684–688.
- Leroux C and Tisseyre B (2019) HOW to measure and report within-field variability: A review of common indicators and their sensitivity. *Precision Agriculture* 20(3): 562–590.
- Lombardi F, Ortuani B, Facchi A, and Lualdi M (2022) Assessing the perspectives of ground penetrating radar for precision farming. *Remote Sensing* 14(23): 6066.
- Long DS and McCallum JD (2020) Adapting a relatively low-cost reflectance spectrometer for on-combine sensing of grain protein concentration. *Computers and Electronics in Agriculture* 174: 105467.
- Lyle G, Bryan BA, and Ostendorf B (2014) Post-processing methods to eliminate erroneous grain yield measurements: Review and directions for future development. *Precision Agriculture* 15(4): 377–402.
- McBratney AB, Mendonça Santos ML, and Minasny B (2003) On digital soil mapping. *Geoderma* 117(1–2): 3–52.
- Miller NJ, Griffin TW, Ciampitti IA, and Sharda A (2018) Farm adoption of embodied knowledge and information intensive precision agriculture technology bundles. *Precision Agriculture* 20: 348–361.
- Mizik T (2023) How can precision farming work on a small scale? A systematic literature review. *Precision Agriculture* 24: 384–406.
- Mondal P and Basu M (2009) Adoption of precision agriculture technologies in India and in some developing countries: Scope, present status and strategies. *Progress in Natural Science* 19(6): 659–666.
- Mulla DJ (2013) Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems Engineering* 114(4): 358–371. <https://doi.org/10.1016/j.biosystemseng.2012.08.009>.
- Nuske S, Wilshusen K, Achar S, Yoder L, and Singh S (2014) Automated visual yield estimation in vineyards. *Journal of Field Robotics* 31(5): 837–860.
- Pierce FJ, Robert PC, and Mangold G (1994) Site-specific management: The pros, the cons, and the realities. In: *Proceedings of the International Crop Management Conference*, pp. 17–21. Ames, Iowa: Iowa State University Press.
- Plant RE (2001) Site-specific management: The application of information technology to crop production. *Computers and Electronics in Agriculture* 30(1–3): 9–29.
- Pringle MJ, McBratney AB, Whelan BM, and Taylor JA (2003) A preliminary approach to assessing the opportunity for site-specific crop management in a field, using a yield monitor. *Agricultural Systems* 76: 273–292.
- Richey R, Majumder S, Shirvaikar M, and Kehtarnavaz N (2020) Real-time detection of maize crop disease via a deep learning-based smartphone app. In: *Proceedings SPIE 11401, Real-Time Image Processing and Deep Learning 2020*. 114010A.
- Stafford JV (1996) Essential technology for precision agriculture. In: Robert PC, Rust RH, and Larson WE (eds.) *Precision Agriculture*. Proceedings of the 3rd International Conference on Precision Agriculture. pp 595–604, American Society of Agronomy, Crop Science Society of America, Soil Science Society of America.
- Taylor JA, McBratney AB, and Whelan BM (2007) Establishing management classes for broadacre grain production. *Agronomy Journal* 99(5): 1366–1376.
- Taylor JA, Bates TR, Manfrini L, and Guillaume S (2021) Zoning and data fusion in precision horticulture: Current and needed capabilities to assist decision-making. *Acta Horticulturae* 1314: 173–188.
- Whelan BW and Taylor JA (2013) *Precision Agriculture for Grain Production Systems*. Dickson, ACT, Australia: CSIRO Publishing.
- Wolfer S, Ge L, Verdouw C, and Bogaardt M-J (2017) BIG data in smart farming—A review. *Agricultural Systems* 153: 69–80.
- Wong MTF and Harper RJ (1999) USE of on-ground gamma-ray spectrometry to measure plant-available potassium and other topsoil attributes. *Australian Journal of Soil Research* 37(2): 267–277.

Further reading

- Adamchuk VI, Hummel JW, Morgan MT, and Upadhyaya SK (2004) On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture* 44(1): 71–91.
- Fountas S, Aggelopoulou K, and Gemtos TA (2016) Precision agriculture. In: *Supply Chain Management for Sustainable Food Networks*, pp. 41–65. Chichester, UK: John Wiley & Sons, Ltd.
- Gebbers R and Adamchuk VI (2010) Precision agriculture and food security. *Science* 327: 828–831.
- McBratney AB, Whelan B, Ancev T, and Bouma J (2005) Future directions of precision agriculture. *Precision Agriculture* 6(1): 7–23.
- Pierce FJ and Nowak P (1999) Aspects of precision agriculture. In: Sparks DL (ed.) *Advances in Agronomy*, vol. 67, 1–85.
- Zhang N, Wang M, and Wang N (2002) Precision agriculture—A worldwide overview. *Computers and Electronics in Agriculture* 36(2–3): 113–132.

Foundations, measurements and trends in pedodiversity

Juan José Ibañez^a and Marco Pfeiffer^b, ^aNational Museum of Natural History (MNCN), Spanish National Research Council (CSIC), Madrid, Spain; ^bDepartamento de Ingeniería y Suelos, Facultad de Ciencias Agronómicas, Universidad de Chile, La Pintana, Chile

© 2023 Elsevier Ltd. All rights reserved.

Introduction	726
The concept of diversity	727
The scientific concept of pedodiversity	727
Pedodiversity and its measurements	727
State of the art on the relations between pedodiversity and biodiversity and their preservation	728
The conundrum problem to reach full diversity inventories	728
Main mathematical tools in diversity analysis	728
Richness–area relationships	730
Abundance distribution models	731
β diversity index	732
Diversity and the nested subset theory and taxa-range size distributions	733
Nested subsets theory	733
Taxa-range size distributions	734
Diversity, fractals and multifractals	734
Diversity, fractals and multifractal: What are the relationships?	734
Diversity, fractals and multifractals in soil sciences	735
Diversity analysis and the Willis curve	735
The broad universe of the Willis curve distributions	735
Willis curve and long tailed distributions. The example of soil mapping	736
Scaling hierarchies and the ht-index	736
The logarithmic thinking: Cognition soil maps and soil taxonomies	737
Summary	737
References	738

Abstract

Pedodiversity is a neologism of soil diversity concerning to the analysis of the number of soil classes or pedotaxa using mathematical procedures. However, the same tools are shown to be useful for other purposes such as the analysis of the structure of pedological hierarchical taxonomic systems and the standards for soil surveys and maps. Pedodiversity tools are described including their relations with other mathematical approaches, including: nested subset analysis, species–range size relationships, fractals and multifractals and complex systems. One of the most relevant findings obtained in pedodiversity analysis is that the patterns/regularities are the same for pedological and ecological entities, among others.

Key points

- Why should pedodiversity be analyzed?
- What are the mathematical tools applied to diversity studies?
- What patterns and regularities are detected in studies on pedodiversity?
- Comparison between the patterns of biodiversity, pedodiversity and geodiversity.
- Applications of geodiversity studies in other fields of knowledge.
- What is the role of the human mind in the patterns and regularities detected?

Introduction

Traditionally, soil has been defined from the soil forming factors (parent material, climate, biota, physiography, time and human intervention). The evolution of life on Earth has created the soil and it is impossible to imagine our planet without this natural resource, since it is the product of the interaction between the geosphere, biosphere, atmosphere, hydrosphere and cryosphere. Soil has been the driving force of plant diversity and the origin of new species. Likewise, the soil is the habitat for soil organisms and is at the interface of above- and below-ground biodiversities. Therefore, knowledge of the diversity and variety of the world's soils (pedodiversity) is vital for maintaining the biosphere, guaranteeing food security, and protecting the climate system itself.

Currently, the community of pedometricians recognizes pedodiversity analyzes as an independent line of research requiring the attention of experts from various scientific disciplines (Wadoux et al., 2021). The first attempt to analyze pedodiversity quantitatively was carried out by the Russian pedologist Vladimir Fridland in 1976 (see Ibáñez and Bockheim, 2013 and references therein). Unfortunately, his proposal had little success within the community of soil scientists. In 1990, Ibáñez and collaborators estimated the pedodiversity of a geographic space with the tools used by biodiversity experts for more than a century. Much has since been published on pedodiversity, and this chapter summarizes both the achievements and challenges existing to date.

To estimate the diversity of a natural resource, a classification or taxonomy is required and the abundances of each class in a given geographic space. The pedologist classifies soils according to their features and properties giving rise to pedological taxonomies, i.e., soil types or pedotaxa. Soil inventories and maps show the properties of soil landscapes, their diversities or heterogeneities, and how the soil may be used for different purposes. Furthermore, the value of the tools used to study diversity go beyond what we might initially think of as useful because they can be applied in various scientific disciplines and sub-disciplines. In summary, pedodiversity analyzes the complexity and diversity of soil entities, their properties, and also how they are distributed in the landscape.

The concept of diversity

According to Huston (1994, p. 65) diversity can be defined conceptually as:

The concept of diversity has two primary components, and two unavoidable value judgments. The primary components are statistical properties that are common to any mixture of different objects, whether the objects are balls of different colors, segments of DNA that code for different proteins, species or higher taxonomic levels, or *soil types* or habitat patches on a landscape. Each of these groups of items has two fundamental properties: (1) the number of different types of objects (e.g., species, soil types) in the mixture or sample; and (2) the relative number or amount of each different type of object. The value judgments are: (1) whether the selected classes are different enough to be considered separate types of objects; and (2) whether the objects in a particular class are similar enough to be considered the same type. On these distinctions hangs the quantification of biological diversity.

We consider this to be the most comprehensive description of what diversity means, with both its advantages and disadvantages.

The scientific concept of pedodiversity

The Huston Concept of Diversity (Huston, 1994, p. 65) is valid as a first approach to the concepts of biodiversity, pedodiversity and geodiversity among many others. In all cases there is a need to define the “objects” of study precisely. In general, there are many classes of objects for all the scientific disciplines. For example, in ecology the first focus was biological species, followed by interest in the hierarchy of the “object.” The concept of biodiversity began with biological species because they were considered naively to be real entities, while other taxonomies were considered abstractions. The current terminology on biological diversity includes terms such as genetic diversity, habitat diversity, functional diversity, landscape diversity, and so on. Likewise, pedologists consider the diversity of soil types, soil horizons, soil attributes, soil biota, soil niches, etc., but also refer to functional pedodiversity. Furthermore, in pedometrics, mathematical tools are used to split the character space into classes derived by grouping the soil variables that are considered to be of interest for solving specific problems. On the other hand, geostatisticians retain continuity in geographical space to indicate spatial diversity.

Pedodiversity and its measurements

The study of pedodiversity arose because the soil cover is very complex and there is the need for an index or indices to measure it. Pedodiversity in fact embraces a range of diversities (Ibáñez et al., 1995a,b). However, the definitions mean nothing unless they are operationalized to be used in a scientifically sound way with mathematical tools.

Ibáñez and Bockheim (2013) among many others use taxonomic pedodiversity to embrace:

1. Applications of pedodiversity tools to detect soil spatial patterns, including pedodiversity–area and pedodiversity–time relations,
2. Patterns of pedological assemblages (soilscape, soil regions, etc.),
3. Pedogenetic diversity related to soil horizons in a given region,
4. Quantitative mathematical concepts and tools useful in soil geography such as quantification of soil endemisms,
5. Pedodiversity indices from the application of non-linear or complex systems, or dissipative structures in pedology such as convergent versus divergent pedogenesis,
6. Genesis of soilscape,
7. Areas selected for designing networks of natural reserves for soil preservation,

8. Other regularities in soil assemblages such as potential nesting among them, species–range size distribution, scale invariance or scale dependence of soilscape patterns,
9. Comparison of diversity patterns in space and time of different natural resources (e.g., soils, rocks, landforms, minerals, biological diversity),
10. Spatial representation of the variation in soil properties across landscapes with soil taxonomies,
11. Mathematical analysis of the structure of pedological and biological taxonomies,
12. Relations between mathematical tools of diversity and properties of free scaling laws (e.g., fractals and multifractals),
13. Use of diversity mathematical tools to analyze the structures of maps,
14. Cognitive bias in taxonomies, soil mapping and scales, as well as diversity results.

Most pedogenetic studies have mostly been concerned with the formation of the soil body over time. Jonathan Phillips (2013) and co-workers among others have analyzed this topic in depth, showing that pedodiversity tools open new perspectives on pedogenetic theory (divergent pedogenesis) from the point of view of complex or non-linear systems, challenging traditional concepts. The now popular digital soil mapping (DSM) tools offer new opportunities for pedologists with an interest in pedogenesis and pedodiversity from a spatial perspective (Fajardo and McBratney, 2018).

State of the art on the relations between pedodiversity and biodiversity and their preservation

It should be noted that the spatial and temporal patterns of the distribution of soil types in the landscape are strikingly similar to those of aboveground biodiversity such as the following macroecological patterns (see Ibáñez and Bockheim, 2013 and chapters therein): (i) species (taxa)–area relationships; (ii) local–regional richness relationships; (iii) latitudinal gradient in species richness; (iv) altitudinal gradient in species richness; (v) species–range size relationships; (vi) nestedness of species occurrence; (vii) abundance–range size distribution, and (viii) species–abundance distribution.

Several studies have shown that different soil types as well as different soil horizons have different assemblages of soil living organisms, demonstrating that they should be considered as diverse habitats. The soil microbiome appears to be self-organized, including soil aggregate habitats (e.g., Karimi et al., 2020).

It is clear that if we want to preserve biodiversity, we must also preserve pedodiversity (see Ibáñez, 2017 and references therein). For this reason, various soil scientists have proposed quantitative frameworks to create networks of soil reserves and or the inclusion of pedodiversity in other programs related to efforts to preserve nature with the aim of including a framework of pedological entities that must be preserved for their intrinsic value. These include landscape and soilscape metrics, and databases on pedotaxa in risk of extinction. The considerable degradation of the biosphere and geosphere has profoundly altered the pedosphere. Several studies have shown that in certain countries and geographical areas there has been both a loss of pedodiversity and the transformation of some pedotaxa to others (e.g., Lo Papa and Dazzi, 2013 and Bockheim and Haus, 2013 in Ibáñez and Bockheim, 2013). Human actions have transformed the pedosphere to such an extent that new anthropogenic taxa are included in modern soil taxonomies, such as the WRB–FAO Reference groups of Technosols and Anthrosols (IUSS Working Group WRB). There has been a loss of certain types of natural soil that have been replaced by others of anthropogenic origin (Zhang, 2013; in Ibáñez and Bockheim, 2013).

Ibáñez and Montanarella (2013) show that pedodiversity studies require the categorization of the different classes or types of soil. These could be universal, national or “ad hoc” classifications. Universal taxonomies allow for a comparison of the results obtained in different studies, whereas ad hoc classifications would be more useful for purpose-oriented studies. However, different soilscales might require the selection of very different variables for an adequate description of the area’s soils (e.g., permafrost versus arid land taxa). Nevertheless, each approximation has advantages and disadvantages. For instance, national taxonomies often provide better results than universal ones because their taxonomic constructs are adapted to their study area, and are therefore more precise and locally specific. However, the results cannot be compared easily with those obtained in other countries with their own national taxonomies.

The conundrum problem to reach full diversity inventories

How many biological species are found on Earth? How many mineral species are part of the Earth? How many soil types or pedotaxa are found on Earth? (and so on).

Unfortunately, no one can answer these questions in a sound scientific way. Every day new species are discovered or become extinct, every few years all taxonomic structures are changed or reorganized. Consequently, scientists must recognize that the rigorous estimation of diversities according to some premises underestimates what really exists in the world around us.

Main mathematical tools in diversity analysis

It is common in the ecological literature to distinguish between three types of diversities, which have been adopted and used in pedodiversity analysis.

- o α (Alpha) diversity richness is the number of taxa found in a particular geographical space. Usually, it is termed “richness” or “S.” If, only a sample and not the entire population can be obtained, it is necessary to distinguish between numerical richness and “object density.” In the case of a soil survey, the former would be defined as the number of different pedotaxa found in the study area. The latter would refer to the number of pedotaxa per sampling area (see Magurran, 1988).
- o β (Beta) diversity richness is the difference in taxa composition between two areas or regions. It takes into account their diversity as well as the number of distinctive species that appear in each studied area. For example: if land unit 1 has 6 taxa: a, b, c, d, e, and j (α diversity = 6) and land unit 2 has 4 taxa: a, b, k, l (α diversity = 4), to calculate the β diversity the number of overlapping species (or pedotaxa) of each α alpha diversity must be subtracted and the results summed: (6 taxa on land unit 1–2 overlapping taxa) + (4 taxa on land unit 2–2 overlapping taxa) = (6–2) + (4–2) = 4 + 2 = 6. Thus β diversity between land units 1 and 2 is 6.
- o γ (Gamma) diversity richness is a measure of the overall number of taxa or richness (the diversity) within a region or study area. It is basically the sum of all the taxa in the complete study area involved (land unit 1 + land unit 2). In this example γ would be 8.
- o Formal diversity indices are indices based on the proportional abundance of objects and are the most frequently used. They consist of two components: (i) the variety (richness); and (ii) the relative abundance (evenness or equitability) of species (Table 1). Several techniques have been proposed to measure diversity (Magurran, 1988). These are the Shannon negentropy index proposed by Shannon and Weaver in 1948 and the Brillouin index proposed by Brillouin in 1956 (Magurran, 1988).

The Shannon index is calculated by the following formula:

$$H' = - \sum_{i=1}^m p_i * \ln(p_i) \quad (1)$$

where p_i is estimated as n_i/N (n_i is the area covered by the i th pedotaxon and N is the total area under study). The index H' is usually between 1.5 and 3.5 and only rarely surpasses 4.5. The ratio of observed diversity to maximum diversity (this occurs when all elements are equally abundant) is taken as a measure of evenness “ E ” (Pielou, 1969 in Magurran 1988) (Fig. 1).

Table 1 Conversion of common indices to true diversities after Jost (2007).

Index H	Diversity in terms of H		Diversity in terms of p_i
Species Richness	$H \equiv \sum_{i=1}^S p_i^0$	H	$\sum_{i=1}^S p_i^0$
Shannon Entropy	$H \equiv \sum_{i=1}^S p_i \ln p_i$	$\exp.(H)$	$\exp \left(\sum_{i=1}^S p_i \ln p_i \ln p_i \right)$
Simpson concentration	$H \equiv \sum_{i=1}^S p_i^2$	$1/H$	$1 / \sum_{i=1}^S p_i^2$
Ginni–Simpson index	$H \equiv 1 - \sum_{i=1}^S p_i^2$	$1/(1-H)$	$1 / \sum_{i=1}^S p_i^2$
HCDT entropy	$H \equiv \left(1 - \sum_{i=1}^S p_i^q / (q-1) \right) [1 - (q-1)/H]^{1/(1-q)}$		$\left(\sum_{i=1}^S p_i^q \right)^{1/(1-q)}$
Renyi entropy	$H \equiv \left(- \ln \sum_{i=1}^S p_i^q / (q-1) \right)$	$\exp.(H)$	$\left(\sum_{i=1}^S p_i^q \right)^{1/(1-q)}$

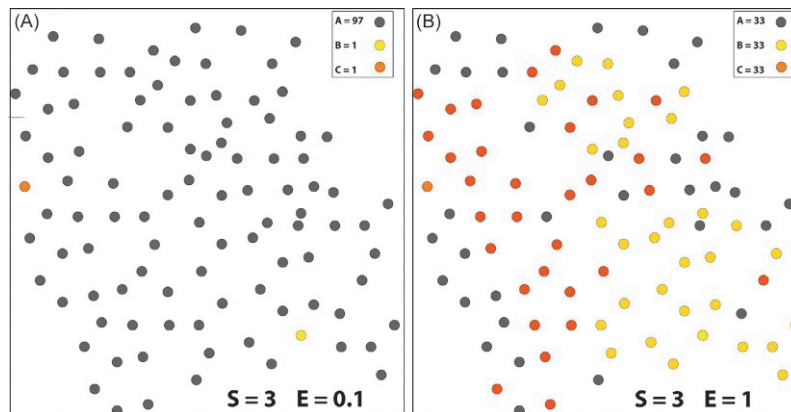


Fig. 1 The figure represents the concepts of richness and evenness. While both boxes have the same number of objects (99) and classes (3), their distribution is different, meaning their richness values are equal ($S = 3$) but their evenness varies. In box (A) evenness has a low value because most objects are concentrated in only one class, while in box (B) evenness is 1 because all objects are equally distributed among classes.

The Brillouin index is calculated as follows:

$$HB = \frac{\ln N! - \sum \ln n_i!}{N} \quad (2)$$

where N is the total number of pedotaxa and n_i is the number of individuals belonging to the i th pedotaxa.

Minasny et al. (2010) suggested that taxonomic distance or taxonomic differences between soil classes could be used to obtain a diversity index in trying to improve the estimation of pedodiversity (also see Ibáñez and Bockheim 2013).

The above formulae are commonly applied in pedodiversity and use three types of input: (i) the number of individuals of a given pedotaxon; (ii) the proportion of the area characterized by a given soil type; and sometimes (ii) the number of diagnostic horizons identified in each sample. The first two result in taxonomic pedodiversity, while the last one constitutes a first approximation to the genetic diagnostic horizon diversity or genetic pedodiversity (see Ibáñez, 2017). Ibáñez and Effland (2011) apply these mathematical tools to each level of a given hierarchical pedological taxonomy because they consider that this procedure shows the relevance of different soil forming factors that are the major determinants of each of these hierarchical levels.

Regardless of this variety of indices, with their respective virtues and imperfections, several authors show how most of them are strongly correlated (Table 2). Similarly, different evenness indices tend to be closely correlated (see; Feoli et al., 2013 in Ibáñez and Bockheim, 2013).

Richness–area relationships

The area–species relationship (SPARs) is one of the topics most widely accepted and corroborated empirically in pedodiversity studies (see Ibáñez and Bockheim, 2013; and chapters therein). The increase in the number of taxa is represented graphically against the area sampled in a double logarithmic plot and a straight line is obtained, indicating a power law (MacArthur and Wilson, 1967, in Magurran, 1988), whose simplest mathematical formulation is the following:

$$S = C \cdot A^z \quad (3)$$

where S is the number of taxa encountered, A is the area sampled, C is an empirical constant, and z is the slope of the curve in the log–log plots. Thus, C and z are empirical constants which would vary from some taxa to others and from one locality to another. The smaller the z value, the less will be the increase in the number of taxa according to the increase in area.

The SPARs is also another way to explore, quantify and compare the complexity of abiotic landscape structures in different areas and environments. According to the *Theory of Island Biogeography* (MacArthur and Wilson, 1967 in Magurran, 1988) a seminal theoretical construct used in ecology, the dependence of S (number of different taxonomic groups) on the area A (referring in particular to islands and other analogous land fragmentations approaches such as drainage basins, bioclimatic belts, etc.) can be modeled by a power law (Fig. 2A).

The SPARs have been considered the cornerstone of the biogeography and conservation biology sciences. In view of the similarities that occur with pedodiversity analyzes, Ibáñez and Effland (2011) also suggest a new potential concept *Theory of Island Pedogeography*. Likewise SPARS seems to occur in soil microbial diversity communities (e.g., Ranjard et al., 2013).

Table 2 Richness, diversity, evenness, and dominance calculations: Pros and cons (after that Ibáñez et al., 1990).

	<i>Discriminant ability</i>	<i>Sensitivity to sample size</i>	<i>Richness, evenness or dominance</i>	<i>Calculation</i>	<i>Widely used</i>
α (log series)	Good	Low	Richness	Simple	Yes
λ (log normal)	Good	Moderate	Richness	Complex	No
Q statistic	Good	Low	Richness	Complex	No
S (species richness)	Good	High	Richness	Simple	Yes
Margalef index	Good	High	Richness	Simple	No
Shannon index	Moderate	Moderate	Richness	Intermediate	Yes
Brillouin index	Moderate	Moderate	Richness	Complex	No
McIntosh index	Good	Moderate	Richness	Intermediate	No
Simpson index	Moderate	Low	Dominance	Intermediate	Yes
Berger–Parker index	Poor	Low	Dominance	Simple	No
Shannon evenness	Poor	Moderate	Evenness	Simple	No
Brillouin evenness	Poor	Moderate	Evenness	Complex	No
McIntosh D index	Poor	Moderate	Dominance	Simple	No
Power law	Moderate	Moderate	Richness	Simple	No ^a
Z value					

^a Z value is the most used in taxa–area relations, but not as an abundance distribution model. Dominance is the opposite of rarity proposed by Whittaker in 1965 (see Magurran, 1988). Modified from Magurran AE (1988) *Ecological Diversity and Its Measurement*. London, UK: Croom Helm.

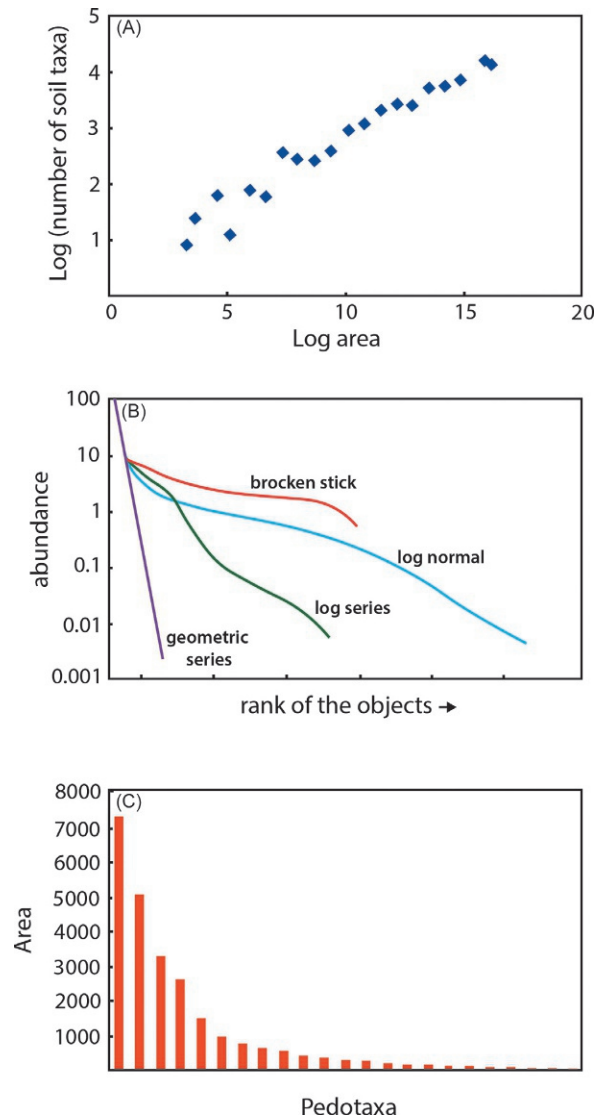


Fig. 2 Abundance distribution models. (A) Richness–area relationship by all countries using the FAO soil database; (B) a hypothetical distribution plot with the shapes of the most common abundance models; (C) ranked abundance list of pedotaxa for the Aegean archipelago showing a typical Hollow/Willi's curve. (A) Modified from Caniego J, Ibáñez JJ and San José Martínez F (2006) Selfsimilarity of pedotaxa distributions at planetary level: A multifractal approach. *Geoderma* 134: 306–317. doi: 10.1016/j.geoderma.2006.03.007, (B and C) modified from Ibáñez JJ, De-Alba S, Bermúdez FF and García-Álvarez A (1995) Pedodiversity: Concepts and measures *Catena* 24: 215–232, [https://doi.org/10.1016/0341-8162\(95\)00028-Q](https://doi.org/10.1016/0341-8162(95)00028-Q).

Abundance distribution models

Taxa abundance distribution models (ADM) use all the information gathered in a community. Taxa abundance data are described by families of distributions, whereas diversity is usually examined by four main models: geometric series, logarithmic series, lognormal distributions and MacArthur's broken stick model (Magurran, 1988). A rank/abundance plot or RAP (Fig. 2B) of the four models represents a progression ranging from the geometric and logarithmic series where a few taxa are dominant with the remainder fairly uncommon, through the lognormal distributions where species of intermediate abundance become more common and then to the conditions represented by the broken stick model in which empirical data show that species are more equally distributed (Magurran, 1988). Diversity experts display relative taxa abundance with taxa on the X-axis according to their decreasing abundance, while their relative abundance is shown on the Y-axis. Frequently, the same set of data might fit more than one ADM with (more or less) the same confidence levels, and it can be difficult to recognize which one is the "real" data distribution (e.g., Jiang, 2013). Although some ecological data fit other types of distribution, Magurran (1988) emphasizes that the four models mentioned should be used whenever possible to enable comparisons. Although the adjustment of abundance distributions rarely reaches the 95% significance level, this methodology aims to identify the best fit, with the goodness of fit of secondary importance.

Visual inspection of RAPs recommended by Mandelbrot, as well as the term “ht-index” (Jiang, 2013; Jiang and Yin 2014; Ibáñez et al., 2021) are useful for interpreting the results of a goodness of fit test. The RAPs in biological taxonomies and biodiversity inventories go back a century and are currently known as “Willis’s or Hollow curve,” which is also present in pedodiversity analysis based on pedological taxonomies (see Ibáñez and Bockheim, 2013; Ibáñez and Montanarella, 2013 for further details). The origin of this term comes from the concave shape of the frequency distribution of taxonomic assemblages when its taxa are ranked from the most to the least abundant (Fig. 2C). The Willis’s curve shows that there are many rare taxa within each assemblage and few very abundant taxa. The hollow curve distributions are well fitted by a power law or related statistical distribution models. In diversity algorithm terms this feature means that the evenness index (E) will not have the highest possible value, thus the Shannon entropy index (H) for a given number of species (S) never reaches its potential maximum value (all the taxa have an equiprobable number of individuals or cover).

Most data sets fit well with several models of distribution, for example, to both the lognormal and power-law distributions (Jiang, 2013; Ibáñez et al., 2021). This problem results from both sampling capacity and the statistical techniques. First, it is impossible to sample all taxonomic diversity in any sampling scheme because it is restricted in space and in time, which results in many “rare” or infrequent taxa not being sampled or identified. Consequently, the distribution model obtained will underestimate richness and overestimate evenness. Thus, the diversity figures are also affected. Similarly, the sample data could be fitted by a distribution model which might be different from what could be obtained with all the data from the study area (Tokeshi, 1993).

β diversity index

The β diversity measures the change in composition of taxa from one sampling area or habitat to another. The term was introduced by R.H. Whittaker in 1960 and is defined as “the extent of change in community composition, or degree of community differentiation, in relation to a complex-gradient of environment, or a pattern of environments.” In his seminal paper, Whittaker proposed several methods to measure β diversity. At its simplest, it could be defined as the ratio between gamma (regional) and alpha (local) diversities. The beta diversity quantifies numerically the number of different communities in the region. It not only considers the relation between local and regional diversity, but also informs about the degree of differentiation among pedological assemblages. A high β diversity index indicates a low level of similarity, while a low beta diversity index shows a high level of similarity. The simplest method proposed by Whittaker to quantify differences in diversity is the following

$$\beta = \gamma / \alpha \quad (4)$$

where gamma γ diversity is the total taxa diversity of a whole study area, and α diversity is the mean taxa diversity for each pedological assemblage. A pragmatic solution is to consider γ diversity as the total species diversity in the dataset and α diversity the mean species diversity per subunit. Over time diversity experts have continued to propose more mathematical expressions with varying degrees of complexity. As a result, there are now many procedures and β diversity algorithms proposed. Some of the oldest similarity coefficients are now being considered useful by diversity experts such as the Jaccard and Sørensen indices.

$$\text{Jaccard } J = j / (a + b - j) \quad (5)$$

and

$$\text{Sorensen} = 2j / (a + b) \quad (6)$$

where j is the number of taxa found at both sites, a is the number of taxa at Site A and b is the number of taxa at Site B. These indices are designed to equal “1” for complete similarity (that is where the two sets of taxa are the same) and “0” if the sites are dissimilar and have no taxa in common. These indices have the advantage of being simple. However, these coefficients do not take taxa abundance into account. Thus, all taxa have the same weight in the equation irrespective of their abundance (e.g., dominant, frequent, rare, etc.). This has led to similarity measures based on quantitative data (see Magurran, 1988).

For each assemblage of taxa, it is assumed that some taxa are replaced by others resulting from environmental or human factors. However, when these assemblages are of different sizes, it is possible that low abundance taxa could be nested subsets of the larger ones, as is proposed by the “nested subset theory” (see below). Therefore, the dissimilarity measured by β diversity could be due to substitution/replacement, nesting, human impact or all of them to a different degree. Thus, total dissimilarity can be portioned additively into replacement and nestedness components.

Ibáñez et al. (2019) note an important factor not previously taken into account in the literature: the design of the sampling areas. The authors split the soilscapes and vegetation of a given region according to two perspectives (i) drainage basins and (ii) bioclimates, and compared the results. For drainage basins, the driving force of β diversity in their study area was the nestedness, whereas for bioclimatic fragmentation it was the species replacement. The literature on Beta diversity is broad and can be analyzed from so many perspectives including numerical taxonomy and data analysis in general (see Feoli et al., 2013 in Ibáñez and Bockheim, 2013).

Diversity and the nested subset theory and taxa-range size distributions

Nested subsets theory

One of the other most common patterns detected in pedodiversity studies concerning SPARs is termed “nested subsets,” an idea proposed by Patterson and Atmar in (1986). Nestedness appears to be a near-universal property of the taxa assemblages studied by pedologists and ecologists provided that they meet certain conditions, such as belonging to geographic areas largely under similar environmental conditions allowing for β and γ diversity to be estimated. A fundamental asymmetry of ecological/pedological relationships is implied by nested subsets. This pattern arises in view of taxa that appear in smaller areas (e.g., islands or drainage basins of small size) and are also present in other larger ones of the same type. This does not, however, work both ways. In other words larger assemblages contain additional idiosyncratic pedotaxa that do not appear in the smaller ones (see Ibáñez and Bockheim, 2013 and Fig. 3A). Thus, less taxa-rich-assemblages tend to be subsets of more taxa-rich assemblages. The nested subset pattern arises because taxa differ in their distributions across heterogeneous spaces (usually with particular habitats), or imply locations with contrasting geographic histories. In pedological terms larger islands, on average, have more soil types than small ones because there is greater variety in soil forming factors, such as relief, climate and mature depositional landforms. This fact is the result of two driving forces: (i) there is a positive correlation (a power law) between the island size (or drainage basins) and relief/altitude that increases the environmental heterogeneity; (ii) larger islands and drainage basins have landforms that do not appear in the smaller ones (Ibáñez and Effland, 2011; Ibáñez et al., 2019). In pedological terms it is possible to predict the probability that some specific soil types appear according to the size of the above mentioned land units. Nested patterns have been documented for an abundance of pedological assemblages in more disparate habitats and on spatial scales, such as those induced by urban sprawl (see Zhang, 2013, in Ibáñez and Bockheim, 2013).

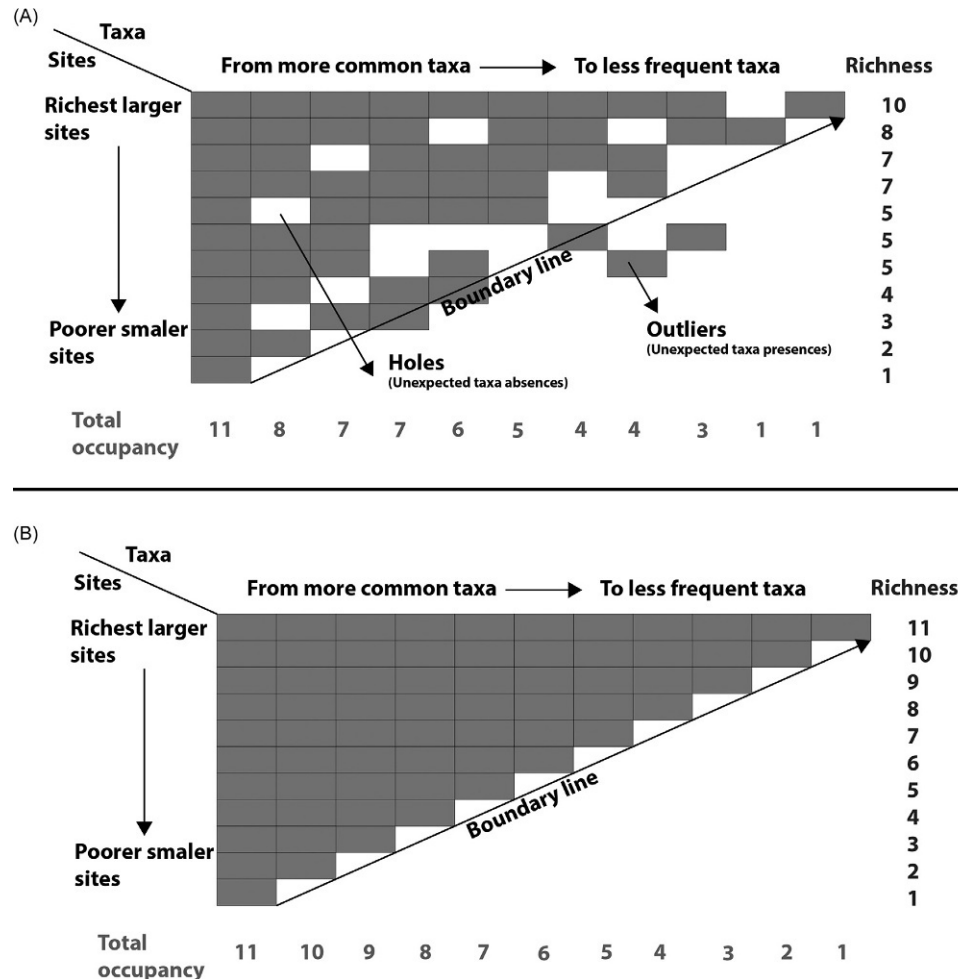


Fig. 3 Distribution matrix of pedotaxa for realistic (A) and perfect (B) nested matrices. Each horizontal and each row is organized from larger sites on top toward smaller sites at the bottom. Larger sites are expected to have greater richness in contrast to smaller sites. Each column represents the occurrence of a particular taxa, an unfilled box correspond to an absent taxa, most common taxa are expected to be present in all sites, while less abundant taxa are expected to be present at larger sites. There is a tendency for occurrences to cluster in the upper left corner of the table.

All nested patterns have a certain amount of noise. For hypothetically perfect nestedness (Fig. 3B), if a taxon is present in a given assemblage or island, it is also present in all larger islands with richer assemblages. If a taxon is absent from an assemblage, it is also absent from all smaller (i.e., less taxa-rich) assemblages. Taxa are ranked by number of occurrences among sampled habitats/sites; sampled habitats are ranked by taxa richness. The wedge-shaped pattern of occurrences in a taxa presence-absence matrix is characteristic of nested soilscape (Fig. 3B).

In a perfectly nested matrix (Fig. 3B) the hypothetical line that separates the occupied area of the matrix (i.e., the upper-left corner of the matrix) from the unoccupied portion is called the “boundary line.” Taxa absences above and to the left of the line are defined as unexpected absences, similarly to the taxa presented below and to the right that are defined as unexpected presences (Fig. 3A). When randomness is low, unexpected presences and absences cluster near the boundary line. In contrast, if randomness increases unexpectedness moves further away from it. The measurable “entropy” of the matrix is the result.

Several metrics have been devised to measure the fit of a given presence-absence matrix to the nested subset model (e.g., *Atmar and Patterson, 1995*). One metric is the “statistic T ,” which measures matrix disorder, entropy or temperature (*Atmar and Patterson, 1995*). This is calculated as the sum of squared deviations from the matrix boundary line (Fig. 3B) of unexpected presences and absences, divided by the maximum value possible for the matrix multiplied by 100. It also has the advantage of allowing for the identification of idiosyncratic taxa and sites. In other words, this procedure detects taxa or patches that contribute the most to matrix disorder by disrupting the nested pattern. The statistical significance of T is determined by comparing the index of the patterns with the random matrices generated through a Monte Carlo simulation.

Unexpectedness is similar to Boltzmann’s probabilistic description of entropy (disorder) and Shannon’s definition of information or surprise (e.g., the Shannon index). Unexpected presence or absence of taxa in island/drainage basins/habitats range from 0 °C in perfectly nested matrices to 100 °C in random matrices. Two forms of variability contribute to the temperature of a matrix: (i) the random variation of environmental stochasticity and (ii) the “coherent” variability caused by specific (bio)-pedogeographic events. Random events create a uniform gray “noise” band along the length of the boundary line. In contrast, coherent events generate idiosyncratic spikes in taxa or island entropies that are more random than the ones for the overall matrix. Random noise generates a break in the smoothness of the boundary line. The less smooth this line is, the more unexplained variety occurs in the matrix (*Atmar and Patterson, 1995*).

The ubiquity of nested subsets in soilscape analysis refutes the notion that this pattern is the result of idiosyncratic properties in biologic communities (see *Ibáñez et al., 2021*); additional mathematical information is provided by *Atmar and Patterson (1995)*.

Taxa-range size distributions

Other patterns usually detected in pedodiversity and biodiversity are termed “taxa-range size distributions,” which indicates that widespread taxa are much more abundant than taxa of restricted occurrence in a territory (see *Ibáñez and Bockheim, 2013* and references therein). Once again, a power law describes this pattern: the relationship between the abundance and their geographic range.

Diversity, fractals and multifractals

Diversity, fractals and multifractal: What are the relationships?

Power laws are the only decreasing monotonic functions that do not vary with changes of scale; they are, therefore, known as “free scaling laws” (see below). Thus, it is not surprising that some time ago many scientists started to analyze the relations between power laws and fractals (see San José and Caniego, 2013 in *Ibáñez and Bockheim, 2013*). According to many authors, biodiversity, pedodiversity and the species-area and pedotaxa-area relations are fractal and or multifractal structures (see San José and Caniego, 2013 in *Ibáñez and Bockheim 2013*).

As the mathematical explanation of fractals and multifractals is complex and extensive, we describe only the most basic concepts, specifically, those related with diversity analysis, as explained by *Caniego et al. (2006)*. For many fractal and multifractal experts, the detection of power laws (which, is not a trivial problem) is proof of the existence of underlying fractal processes or structures. If this is the case, the estimates of diversity (at least in those cases where the data sets agree with power laws) would be a manifestation of the nature of underlying fractal structures (e.g., floristic assemblages of phytocenoses, biotaxonomic schemes, soil classifications, phylogeny and habitat diversity, etc.). San José and Caniego (2013, in *Ibáñez and Bockheim, 2013*) detected fractal and multifractal patterns in pedodiversity studies (see below). There are two main approaches in multifractal analysis: (i) singularity exponents and (ii) Rényi dimensions. The Rényi spectrum of generalized dimensions summarizes some important diversity indices that are widely used in diversity analysis (*Borda-de-Água et al., 2002*).

It is remarkable that in the sphere of applied physics, relations between statistical and information entropies (e.g., Shannon Index), and fractals have also been detected, especially in developments on multifractal analysis. The most important concepts of fractal dimension are the Kolmogorov capacity D_C , the Hausdorff-dimension D_H , the correlation dimension D_{corr} , and information dimension D_I .

The different fractal dimensions mentioned above D_H , D_{corr} , D_I and D_C are not equal, but satisfy the inequality:

$$D_{corr} \leq D_I \leq D_H \leq D_C \quad (7)$$

where equality occurs if and only if the fractal is strictly self-similar and homogeneous. Since the inequality is quite tight in most cases, in most applications it is tacitly assumed that $D_{\text{corr}} = D_H$ mainly because the numerical determination of D_{corr} is much easier than that of D_H .

Diversity, fractals and multifractals in soil sciences

Power-law scaling of pedomorichness–area relationships was first conjectured by Ibáñez and De Alba in 2000 (see Ibáñez and Bockheim, 2013) for Earth soil systems and they have been reported later for different soil landscapes, pedological classifications, soil map standards and soil geography studies (see Ibáñez et al., 2021). For example, the spatial distribution of pedotaxa at a planetary and continental level or by biomes also conforms to fractal and multifractal structures. Likewise, they have been interpreted in terms of the self-similarity of the spatial distribution of pedomorichness (Caniego et al., 2006). Thus, the power law becomes the by-product of a fractal structure while the scaling exponent quantifies the regularity trend in the reproduction of the spatial pattern of pedomorichness within a certain range of scales. It is noticeable that the same occurs concerning the structure of pedological and biological classifications (e.g., see Ibáñez and Montanarella, 2013 and references therein), but also multifractal analysis can also explain diversity indices, and how they can be interpreted as pedodiversity indicators.

Caniego et al. in 2006 (see San José and Caniego chapter in Ibáñez and Bockheim, 2013) computed Rényi generalized dimensions for the abundance distribution of pedotaxa for the Earth's five landmasses and the whole world using the FAO Soil Database at the second level (FAO, 1974). Their results indicate that the complex behavior of pedodiversity distributions at the planetary scale follows a well-defined multifractal behavior (see San José and Caniego in Ibáñez and Bockheim, 2013). Therefore, multifractal parameters can be used as pedodiversity indicators and show promise for analyzing and characterizing the complexity of soil development at multiple scales. Thus, multifractal analysis could be a way to unify a framework that includes the common procedures to characterize diversity: taxa-abundance distributions, taxa–area relationships, nestedness, and diversity indices (e.g., Richness, Shannon Entropy, Shannon Equitability, Simpson Index, Berger–Parker Index) (Caniego et al., 2006), as can be seen below.

Power laws, fractal and multifractal structures also appear in the soil matrix and soil living communities (see Ibáñez, 2017 and references therein). A well-developed soil matrix structure is a self-structured or self-organized habitat with very specific and fascinating properties. Three examples will suffice to understand these features. The inorganic particles that form the soil matrix (excluding rock fragments) constitute what is called texture, consisting of particles of decreasing size: sand, silt and clay. These small particles interact and bind with particles of decomposing organic matter (that includes different types of material with particular properties) forming so-called soil aggregates (e.g., Guadagnini et al., 2013). The latter confer on the soil their most relevant properties needed to sustain life, water and air circulation, as well as regulation of the pedoclimate. It seems that both the size distribution of the soil particles as well as the aggregates and soil pores conform, more or less, to the same statistical number–size distribution previously detected in pedodiversity analysis. These are power laws, fractals or multifractals where there are far more small soil particles, aggregates, and pores and organisms than large ones (e.g., Guadagnini et al., 2013).

Similarly, the formation of most soils implies that the rocks and sediments become structured (self-organized) (e.g., see Phillips, 2017) in layers or horizons, with different chemical and physical properties, leading to the creation and increase in the number of habitats that can be inhabited by idiosyncratic assemblages of soil organisms (e.g., Crawford et al., 2012 and references therein). In other words, the increase in the number of soil horizons according to pedogenetic progress in many pedotaxa entails the genesis of new and different habitats, and as a result, the increase in self-organized soil biodiversity, since it begins to form from fresh rocks or parent materials. Furthermore, soil structures produce an abundance of microhabitats (e.g., Nunan et al., 2003, among many others). Therefore, the soil should be considered to be a cascade of nested habitats within each other, with the creation of different macro-, meso- and micro-environments with different biological communities (soilscapes, soil bodies, soil horizons, soil architecture, soil aggregates, etc.) (e.g., Nunan et al., 2003).

Various experts have shown evidence that both soil bodies and soilscapes are complex systems (e.g., Phillips, 2017; San José-Martínez and Caniego, 2013 in Ibáñez and Bockheim, 2013) and that they show fractal properties. In fact, Jonathan Phillips demonstrated that Hans Jenny's factors of soil formation respond as non-linear systems. We assume that such a link occurs between fractals and complex systems and also occur for the whole of the Earth surface system (e.g., Phillips, 2016).

Diversity analysis and the Willis curve

The broad universe of the Willis curve distributions

In previous sections we have tried to synthesize the studies carried out in the analysis of pedodiversity and its relationships with those obtained in the science of biodiversity. More recently, two meta-analyses have been published that may reveal some intriguing aspects not previously taken into account in such inquiries (Ibáñez et al., 2021).

The opportunities offered by geospatial technologies in recent decades have resulted in certain controversies in the geographical thinking and classical analysis of spatial structures. One of these involves the ubiquity of scaling laws such as power laws and fractals (Mandelbrot, 1982), as we have seen in previous sections. A simple and well-known metaphor for the spatial fragmentation of the polygons in a map (Jiang, 2015 in Ibáñez et al., 2021) is the following: “if we forcefully throw a glass bottle to the ground, it will fragment into pieces of all sizes with the abundance of any size class decreasing as the size increases, and this relationship between size and abundance will follow a power law.” There are many claims in geographical sciences suggesting that the classical

geographical analyzes and map representations are not appropriate in the analysis of geographical spaces. A paradigm shift has been proposed from our current Gaussian thinking to another, based on Paretian thinking, as has happened with other scientific disciplines (Ibáñez et al., 2021 and references therein).

Normal distributions and the Gauss curves and related current quantitative methods are commonly used in statistical analysis. However, the growing ubiquity of power laws signifies that Pareto rank/frequency distributions, fractals, and underlying scale-free theories are increasingly widespread and provide valid characterizations of natural and social structures and their dynamics. This is because data extremes of the distribution should be considered not the Gaussian statistical averages. Although it is possible that both modes of thinking are relevant (see below).

Geographical analyzes show the ubiquity of Pareto power-law distributions over Gaussian bell curves and linear fits. These well-established laws correspond more or less to the more general concept of the Willis curve or long tailed distributions. Therefore, a new type of statistical analysis that includes Bayesian thinking is needed by researchers in disparate disciplines. Within this framework, the general notion that there are far more small things than large ones across several orders of magnitude should be considered in geographical and mapping analyzes (Jiang, 2015 in Ibáñez et al., 2021), as this has been demonstrated to be the rule or conjectured as a “universal law” (Salingaros and West, 1999; in Jiang, 2013; Jiang and Yin, 2014).

In the most problematic cases, the dataset might fall into the Willis curves category. This occurs for a variety of potential reasons (small data sets, small uncertainties in data quality, etc.). This is the main reason that controversies have arisen in some studies regarding whether a given geographical feature should be considered as fractal or not. Jiang (2013) claims that the geographical space is dynamic and is changing over time. Therefore, sometimes the structures could trend toward fractals even when they do not reach their full development. Ibáñez et al. (2021) show many examples of this in the most disparate disciplines, scientific, social or economics among others (e.g., urban maps). When these circumstances, among others, occur, tested datasets might fit better with other related distributions, such as geometric, logarithmic, lognormal or exponential functions rather than a power function (Jiang, 2013). In general, power laws and fractals are linked to non-linear and or complex systems in which heterogeneity/diversity and the interaction/connectivity between their constituent elements is required (Ibáñez et al., 2021 and references therein). Therefore the Jiang (2013) conjecture seems to be corroborated beyond cartographic representations across the most disparate disciplines and scales.

Willis curve and long tailed distributions. The example of soil mapping

A map series is a set of maps that generally follow the same standards and cover an area or a country in a systematic way. The maps of a given series have the same scale, format and system of symbolization (graphic semiology). National map agencies in each country commonly have sets of map series (for example, topographic series) at different scales, based on their larger scale maps, and proceeding to smaller scale maps through processes of scale reduction and generalization. Map generalization is a very complex visual-intellectual process. Robinson (1995, in Ibáñez et al., 2021) defines the whole process in terms of four aspects: simplification, classification, symbolization and induction. When reviewing different areas of cartography, we found no reference related to fractal structures or statistical formulae that guide the content and production of maps (e.g., Ibáñez and Montanarella, 2013).

Scaling hierarchies and the ht-index

Jiang (2013) propose the “ht-index,” complementary to fractal dimension, to better understand geographic patterns, which analyzes and quantifies the texture of the Willis curves and the fractal or scaling patterns detected in geographical spaces. According to Jiang (2013), many controversies over whether a given structure is fractal or not, occurs as a result of the narrow and “orthodox definition” of fractal dimension as a measure or index for characterizing free scaling features or forms.

Jiang and co-workers proposed and tested, in different publications, a very simple but useful mathematical procedure that attempts to capture the free scaling/fractal nature of spatial structures using Willis curves or Rank Abundance Plots (RAPs), as shown in Fig. 2C. According to Jiang’s conjecture (“there are more small things than large ones”), when a structure is analyzed at different “scales,” in the sense of the latter concept being applied in fractal sciences, it is possible that the measurement of the complexity and coherency of Willis distributions, considered by these authors as fractals, correspond with their proposition of a relaxed definition of fractal (fractal complexity). These authors state that “the ht-index captures how many scales, ranging from the smallest to the largest, form a much greater number of small things in the scaling hierarchy. The ht-index is defined as one plus the recurring times of small things than large ones. In other words, a geographic feature has an ht-index of h if the pattern of far more small things than large ones recurs $(h-1)$ times at different scales. The higher the ht-index, the more complex the geographic feature.” Therefore given an RAP curve, it is easy to deliver a type of “scaling hierarchy” using the ht-index.

An RAP is the first step in obtaining an ht-index. Then the same analysis is performed once again for those data (objects) that are above the arithmetic mean of the dataset analyzed, to differentiate between small (termed “the tail”) and large (called the “head”) objects. Note that this subset is considered a nested RAP of the former RAP. If the result is “more small things than large ones” the same operation is repeated recursively until the condition is violated. The number of times that the “more small than large” pattern occurs (“scaling hierarchy”) is a measure of the complexity of the full data RAP, as a nested hierarchy or scaling. Jiang and coworkers have developed more quantitative versions of the method to estimate the ht-index. Ibáñez et al. (2021) corroborate Jiang’s proposal in pedodiversity studies. Jiang’s proposal is also applicable in non-geographic analyzes, as we have seen previously.

The logarithmic thinking: Cognition soil maps and soil taxonomies

According to some anthropologists, some aboriginal cultures (e.g., the Mundurucu indigenous peoples of the Amazon) calculate logarithmically instead of in the linear way, which is typical of western culture. Likewise experimental psychologists have shown evidence that childrens' minds begin to calculate logarithmically, changing over time and gradually (probably through cultural training) to do so in a linear way (see bibliography in Ibáñez et al., 2021). Why do we perceive the world around us logarithmically? According to some authors, it is that such a way of calculating is an adaptive advantage in the search of natural resources in contrast with Gaussian thinking.

Does the mind work logarithmically? Both biology and biological taxonomies conform to free scaling laws (see Ibáñez and Montanarella, 2013 and references therein). Soil and biological classifications were developed independently across decades and considered as natural biological taxonomies and artificial soil taxonomic constructs (Ibáñez and Montanarella, 2013). None of the developers of these taxonomies made intentional use of logarithmic thinking. Rather, soil taxonomies were intuitive/rational products of numerous groups of experts working independently over decades in different disciplines, and without considering the mathematical aspects that might be desirable to create efficient taxonomic structures. The same can be said of cartographic products and their rules for making maps, yet power laws again appear to be ubiquitous in their official standard structures (e.g., Soil Survey Staff, in all versions) as well as in the resulting structures derived from them (Ibáñez et al., 2009).

The case of fractal coupling between the soil taxonomies, soil survey and mapping procedures will serve as an example that corroborates the above mentioned theses. Some studies show that the USDA (United States Department of Agriculture) Soil Taxonomy has the same mathematical structure as some biological ones which conform to physical laws that dictate and optimize information flow in user-friendly retrieval systems according to the Jaynes' (maximum entropy) Principle (see Ibáñez and Montanarella, 2013). Likewise Ibáñez et al. (2009) demonstrate that the fractal or multifractal nature of the USDA Soil Taxonomy is strongly linked with conventional soil survey practices. In fact, most surveys are packed with power-law distributions, such as: (i) hierarchic taxonomic level used according to the map scale; (ii) minimum polygon size fits the functions to the map scale; and (iii) boundary density-scale map relationship, among others (see Ibáñez et al., 2009 and references therein). These results supported Beckett and Bie (1978), who showed that most pedological maps have attributes that conform to a power law or other related long tailed statistical distribution. Consequently, a cascade of power laws occurs in soil survey products and soil taxonomies (see Ibáñez and Montanarella, 2013 and references therein). As both activities are strongly linked, it seems as if the minds of soil surveyors and soil taxonomists create similar fractal structures. This process could be the reason for why maps devoid of legends and other information have a strong resemblance and information content, and without scales they provide a clear fractal signature. The same is true concerning the hierarchical branching of classical taxonomies. Experts in cartography know, more or less, well the rules that they apply. However, in the bibliographic material consulted, there were no references about the implications of human cognitive bias in their practice, or to mathematical relationships between classifications and map scale.

Therefore, it remains to be clarified whether Paretian and Gaussian thinking (i) exist simultaneously in our minds; (ii) if Paretian thought is a residual product in our cultural evolution, and replaced by Gaussian thinking in our civilizations over time; (iii) if we unconsciously project our logarithmic mental maps onto nature and structure the world around us in such a way that many scientific findings are more a mental product than an unchanging reality; and (iv) if our minds and the world around us and with which we interact are inescapably the product of a fractal universe (Ibáñez et al., 2021).

In summary, the systems used by soil surveyors and soil taxonomists as a whole have fractal-like structures. Our view is that the standards of many natural resource maps are similar, therefore, it could be conjectured that scale-invariant information processing is intuitive to human minds and that a more rigorous formalization of survey-taxonomy architectures may help practitioners to understand better their activities and constructs, and provide a way to improve them.

Summary

In broad terms, pedodiversity is the study of the variety of soil types within an area using mathematical tools taken from biodiversity. The most common tools used in pedodiversity are related with the measurement of richness (amount of soil taxa in a region), abundance (sum of soil types) and evenness (or the relative distribution of soil individuals). One of the most relevant findings obtained in pedodiversity analysis is that the patterns and regularities are the same for ecological and pedological entities, and possibly for many natural and artificial resources (e.g., geodiversity, land system units). This means that the tools applied in diversity studies have been found to be useful for other purposes (taxonomic structures, standards for soil surveys and maps studies, etc.) that transcend their initial objectives. Pedodiversity mathematical tools were described, as well as other mathematical approaches of relevance that include: nested subset analysis, species-range size relationships, fractals and multifractals and non-linear dynamics.

It is intriguing that the regularities previously detected in biodiversity studies were later corroborated in the study of other entities not typically considered living beings (see Ibáñez et al., 2021). However, it seems even more disconcerting that they emerge again in the analysis of structures made by human minds (e.g., pedological and biological hierarchical taxonomies, soil and vegetation mapping).

References

- Atmar, W. and Patterson, B. D. (1995). The Nestedness Temperature Calculator: A Visual Basic Program, Including 294 Presence-Absence Matrices. AICS Research, University Park, NM, and The Field Museum, Chicago, IL, <<http://aics-research.com/nestedness/tempcalc.html>>.
- Beckett PHT and Bie SW (1978) Use of soil and land-system maps to provide soil information in Australia. In: *CSIRO Division of Soils Technical Paper No. 33*. Melbourne, Australia: Commonwealth Scientific and Industrial Research Organization.
- Bockheim JG and Haus N (2013) Soil endemism and its importance to taxonomic pedodiversity. In: Ibáñez JJ and Bockheim JG (eds.) *Pedodiversity*, pp. 195–210. Boca Raton, FL: CRC Press.
- Borda-de-Água, L., Hubbell, S.P. and McAllister, M. (2002). Species-area curves, diversity indices, and species abundances distributions: A multifractal analysis. *The American Naturalist* 159, 138–155. <http://www.jstor.org/stable/10.1086/324787> (free available from http://ctfs.arnarb.harvard.edu/Public/pdfs/Borda_etal_2002_AmNat.pdf).
- Caniego J, Ibáñez JJ, Martínez SJ, and F. (2006) Selfsimilarity of pedotaxa distributions at planetary level: A multifractal approach. *Geoderma* 134: 306–317. <https://doi.org/10.1016/j.geoderma.2006.03.007>.
- Crawford JW, Deacon L, Grinev D, Harris JA, Ritz K, Singh K, and Young I (2012) Microbial diversity affects self-organization of the soil–microbe system with consequences for function. *Journal of the Royal Society Interface* 9: 1302–1310.
- Fajardo M and McBratney A (2018) Pedodiversity. In: McBratney AB, Minasny B, and Stockmann U (eds.) *Pedometrics*, pp. 491–519. Cham, Switzerland: Springer.
- FAO (1974) FAO-UNESCO Soil Map of the World. In: *Vol. I. Legend*. Paris: UNESCO.
- Guadagnini A, San José Martínez F, and Pachepsky Y (2013) Scaling in soil and other complex porous media. *Vadose Zone Journal* 12. <https://doi.org/10.2136/vzj2008.0127>.
- Huston MA (1994) *Biological Diversity the Coexistence of Species on Changing Landscapes*. Cambridge, UK: Cambridge University Press.
- Ibáñez, J. J. (2017). Diversity of Soils. Oxford Bibliographies (2nd version). Oxford University Press (article on line). <https://doi.org/10.1093/obo/9780199874002-0104>, <http://www.oxfordbibliographies.com/view/document/obo-9780199874002/obo-9780199874002-0104.xml>.
- Ibáñez JJ and Bockheim JG (eds.) (2013) *Pedodiversity*. Boca Raton, FL: CRC Press.
- Ibáñez JJ and Effland WR (2011) Toward a theory of island pedogeography: Testing the driving forces for pedological assemblages in archipelagos of different origins. *Geomorphology* 135: 215–223.
- Ibáñez, J. J., and Montanarella, L. (2013). Magic numbers: A meta-analysis for enlarging the scope of a universal soil classification system. European Commission, JRC Technical Reports. European Commission Brussels. EUR. 25849, ISBN 978-92-79-28899-9. <http://publications.jrc.ec.europa.eu/repository/bitstream/11111111/28069/1/lb-na-25-849-en-n.pdf>
- Ibáñez JJ, Jiménez-Ballesta R, and García-Álvarez A (1990) Soil Landscapes and drainage basins in Mediterranean mountain areas. *Catena* 17(6): 573–583.
- Ibáñez JJ, De-Alba S, Bermúdez FF, and García-Álvarez A (1995a) Pedodiversity: Concepts and measures. *Catena* 24: 215–232. [https://doi.org/10.1016/0341-8162\(95\)00028-Q](https://doi.org/10.1016/0341-8162(95)00028-Q).
- Ibáñez JJ, de-Alba S, and Boixadera J (1995b) The pedodiversity concept and its measurement: application to soil information systems. In: King D, Jones RJA, and Thomasson AJ (eds.) *European Land Information Systems for Agro-Environmental Monitoring*, pp. 181–195. Luxembourg: EU-JRC.
- Ibáñez JJ, Arnold RW, and Ahrens RJ (2009) The fractal mind of pedologists (soil taxonomists and soil surveyors). *Ecological Complexity* 6: 286–293.
- Ibáñez JJ, Pérez-Gómez R, Oyonarte C, and Zinck A (2019) Biogeodiversity and pedodiversity islands in arid lands of Europe (Almería Province, Spain). *Spanish Journal of Soil Science* 9: 148–168.
- Ibáñez JJ, Ramírez-Rosario B, Fernández-Pozo LF, and Brevic EC (2021) Exploring the scaling law of geographical space: Gaussian versus Paretian thinking. *European Journal of Soil Science* 72: 495–509.
- Jiang B (2013) Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *The Professional Geographer* 65: 482–494.
- Jiang B and Yin J (2014) Ht-index for quantifying the fractal or scaling structure of geographic features. *Annals of the Association of American Geographers* 104: 530–541.
- Jonathan Phillips JD (2013) Nonlinear dynamics, divergent evolution, and pedodiversity. In: Ibáñez JJ and Bockheim J (eds.) *Pedodiversity*, pp. 59–82. Boca Raton, FL: CRC Press.
- Jost L (2007) Partitioning diversity into independent alpha and beta components. *Ecology* 88: 2427–2439.
- Karimi B, Villerd J, Dequiedt S, et al. (2020) Biogeography of soil microbial habitats across France. *Global Ecology and Biogeography* 29: 1399–1411.
- Lo Papa G and Dazzi C (2013) In: Ibáñez, Pedodiversity JJ, and Bockheim J (eds.) *Repercussion of Anthropogenic Landscape Changes on Pedodiversity and Preservation of the Pedological Heritage*, pp. 153–194. Boca Raton: Raton: CRC Press Taylor & Francis Group. ISBN: 978-1-4665-8277-4.
- Magurran AE (1988) *Ecological Diversity and Its Measurement*. London, UK: Croom Helm.
- Mandelbrot BB (1982) *The Fractal Geometry of Nature*. New York: Freeman.
- Minasny B, Mcbratney AB, and Hartemink AE (2010) Global pedodiversity, taxonomic distance, and the World Reference Base. *Geoderma* 155: 132–139.
- Nunan N, Wu K, Young IM, Crawford JW, and Ritz K (2003) Spatial distribution of bacterial communities and their relationships with the micro-architecture of soil. *FEMS Microbiology Ecology* 44: 203–215. [https://doi.org/10.1016/S0168-6496\(03\)00027-8](https://doi.org/10.1016/S0168-6496(03)00027-8).
- Patterson, B D., and Atmar, W. 1986. Nested subsets and the structure of insular mammalian faunas and archipelagos. *Biological Journal of the Linnean Society* 28: 65–82., DOI: <https://doi.org/10.1111/j.1095-8312.1986.tb01749.x> (available from <http://www.aics-research.com/research/pa1986.pdf>).
- Phillips JD (2016) Complexity of Earth surface system evolutionary pathways. *Mathematical Geosciences* 48: 743–765.
- Phillips JD (2017) Soil complexity and pedogenesis. *Soil Science* 182: 117–127.
- Ranjard L, Dequiedt S, Chemidlin Prévost-Bouré N, et al. (2013) Turnover of soil bacterial diversity driven by wide-scale environmental heterogeneity. *Nature Communications* 4: 1434. <https://doi.org/10.1038/ncomms2431>. PMID: 23385579.
- Tokeshi M (1993) Species abundance patterns and community structure. *Advances in Ecological Research* 24: 111–186.
- Wadoux AMJ-C, Heuvelink GR, Lark RM, Lagacherie P, Bouma J, Mulder VL, Libohova Z, Yang L, and McBratney AB (2021) Ten challenges for the future of pedometrics. *Geoderma* 401: 115155. <https://doi.org/10.1016/j.geoderma.2021.115155>.

Modeling soil development in a landscape context

Peter A Finke^a, Budiman Minasny^b, and Arnaud JAM Temme^c, ^aDepartment of Environment, Faculty of Bioscience Engineering, Ghent, Belgium; ^bSydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Camperdown, NSW, Australia; ^cDepartment of Geography and Geospatial Sciences, Kansas State University, Manhattan, KS, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	740
Current soil-landscape modeling approaches	741
Focus on pedon	741
Focus on landscape	742
Weathering	743
Erosion and soil flux	743
Focus on soilscape	743
Space–time boundary conditions for soilscape models	745
Initial condition	745
Time-variant upper boundary condition	746
Time-variant lower boundary conditions	747
Heterogeneity	747
Identifying and minimizing uncertainty	748
Accuracy assessment	748
Soil and elevation data for calibration and accuracy assessment	748
Calibration	748
Outlook	749
References	749

Abstract

Soil development models exist at the pedon scale and the landscape scale. Both groups of models have developed along different pathways, and still differ in the processes covered, state of testing and application domain. An overview of different models indicates that their application domains are converging to simulation of global change scenarios. Prerequisites are that these soil(-scape) models take climate and land use change as input at their upper model boundary, and that they allow the simulation of soil hydrology.

Glossary

Accuracy The degree of conformation of a (simulated) property to the measured reality.

Boundary condition Input at the boundary of the modeled soil system, either varying over time at the top (soil–atmosphere interface) and the bottom (e.g., bedrock) or at time zero (initial situation).

Pedon A three-dimensional soil body that is near 1 m² at the surface and extends to the bottom of the soil. In a pedon modeling context, lateral variation is usually neglected and focus is on the (development of) vertical variation.

Soilscape A three-dimensional soil body of undefined area and extending to the bottom of the soil, the saprolite or the bedrock. Lateral variation and its development are, at least at the soil surface, studied explicitly and there is also a focus on the (development of) vertical variation.

Key points

- Soil development models exist at both landscape and pedon scales and have different development histories.
- Pedon models contain mechanistic process formulations of water, solute and heat flow and include various biological, chemical, and physical processes to quantify sources and sinks.
- Soilscape models contain mechanistic formulations of surface mass redistribution processes and functional formulations of water flow and include various biological, chemical and physical processes in a quantified but empirical way.
- Soil(-scape) model inputs at the soil–atmosphere interface can be obtained by proxy-based reconstruction, by outputs from earth system models and by definition.
- Model testing is at present more commonly applied with pedon models than with soilscape models.
- Both types of model allow the analysis of global change scenarios in terms of soil development (including soil organic carbon) if as minimum they include water flow.

Introduction

The term “soil development” is used to cover the change of soil over time in terms of its horizons and its properties. Traditionally, soil development research reconstructs the pathway from a parent material to an observed depth distribution of the biologically active zone (A-horizons), zones of export (E-horizon) and of accumulation (B-horizon) of soil components and zones of weathering (regolith). Essentially, the (often uncertain) initial status, the measured (certain) current status of the soil and process knowledge are used to hypothesize about this pathway, aided by a wide range of techniques (e.g., mineralogical analyzes, microscopy, chemical analyzes, etc.) to estimate soil properties (e.g., mineralogy, proof of biological activity, pedogenic oxides, etc.). The well-known factorial model by Jenny (1941, 1961), based on work by Dokuchaev (1886) and Hilgard (1906) (all in Minasny et al., 2015) was a first attempt towards a quantitative, functional linkage between the soil that has formed and climate, organisms, relief, parent material and time: CLORPT.

From, roughly, the 1980s onwards this changed (Fig. 1) by quantification of the dynamics of soil processes in a mechanistic sense, in the context of research on (a.o.) fate of pesticides (Wagenet and Hutson, 1986), acidification (De Vries et al., 1994) and eutrophication (Hansen et al., 1991). The focus was on short (decadal scale) temporal extents and the ambition to allow prediction into the future. This trend is still ongoing, currently often in the context of global change research. Knowledge obtained and its various implementations have been fed back into the community of soil development researchers. One of the reasons is that global change is directly linked to the factors of soil formation climate, organisms (including human activity), and indirectly linked to relief and parent material as well. In the narrower context of modeling soil development, model development itself has followed two tracks: the landscape/soilscape track and the pedon (1-D) track (Fig. 1).

Here, we will focus on how today's soil development models are organized, what system they describe and how they link to the wider environment via boundary conditions. The landscape–soilscape models and the pedon models will be described separately as they have different origins and underlying concepts, and do not produce the same output variables. As a consequence, interpretation of the model outputs towards soil horizonation differs between these models.

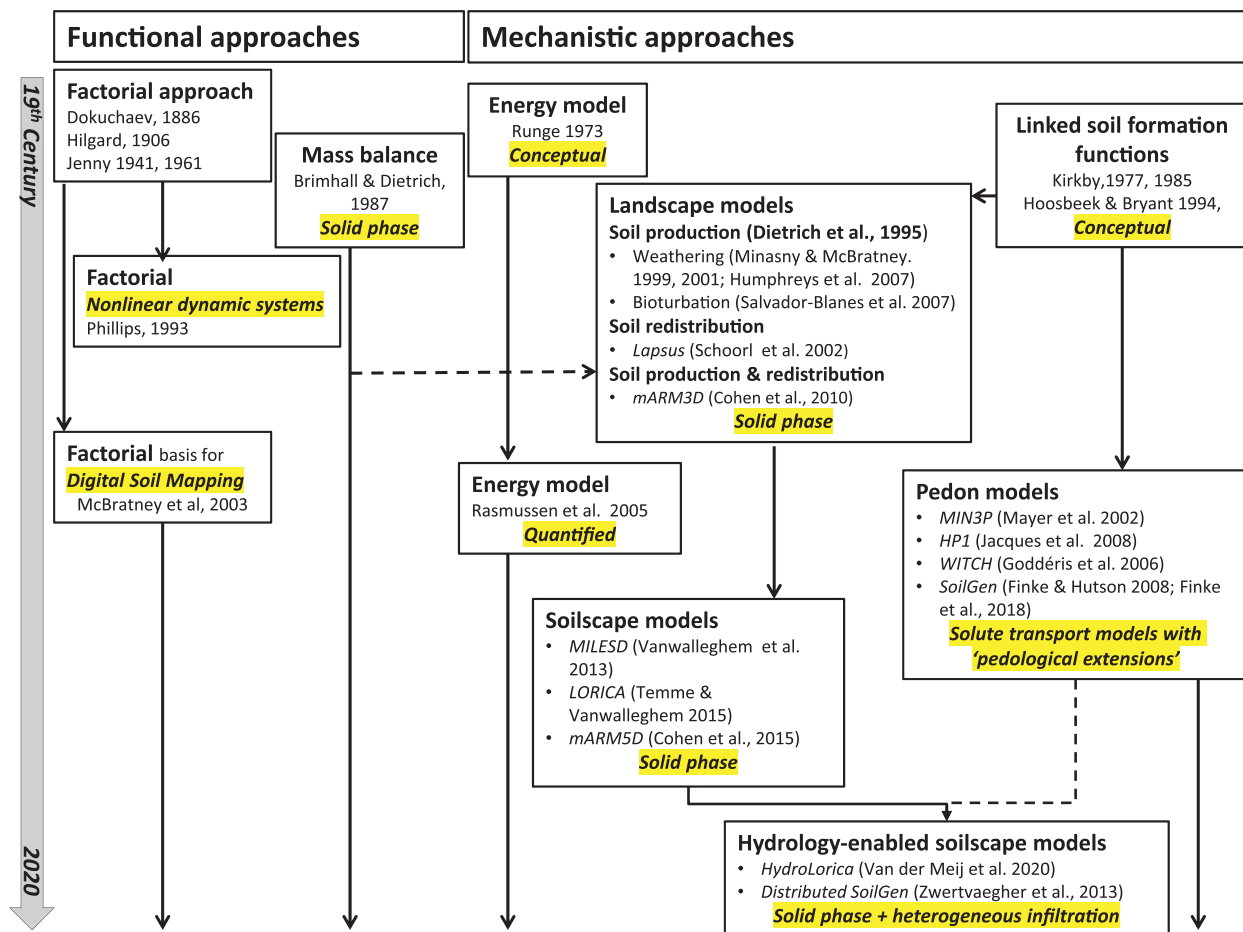


Fig. 1 History and evolution of soil development models. References before 2016 also in Minasny et al. (2015). Updated after Opolot, 2016, p.7.

Current soil-landscape modeling approaches

Focus on pedon

Kirkby and Hoosbeek and Bryant were among the first (Minasny et al., 2015) to develop process-based formulations of chemical reactions and weathering of minerals at the pedon scale and coupling these with a water and solute flow model. This allowed the development of weathering profiles and podzols to be mimicked. Others followed (Fig. 1) but it would take until the 21st century to build mechanistic models that describe more aspects of profile development such as the formation of A, E and Bt horizons by the interaction of processes of bioturbation, decalcification, clay dispersion and transport (Finke, 2012).

Fig. 2 depicts a pedon-scale soil development model. The soil profile is commonly schematized in compartments (finite difference approach), though also a finite element schematization is possible. Flow of water, heat and (dissolved or dispersed) matter between the compartments is commonly simulated by numerically solving partial differential equations (PDE) that combine flow equations with mass balance equations. The mass balance includes source/sink terms by the production or depletion of stocks. In case the soil compartments are considered a continuum of soil, water and air, the Richards equation for water flow in variably saturated soils is a well-known example of such PDEs. More advanced methods exist that take into account soil structure (e.g., macroporosity) explicitly, and simpler methods (e.g., capacity or bucket models) also exist.

For soil development models, additional modes of mass transport need to be considered such as bioturbation, plowing, cryoturbation, etc. This can be done either as a separate mass transport process or as a diffusion term in the PDE (Keyvanshokouhi et al., 2019).

As soil development models may cover millennia, transport-related properties (e.g., hydraulic and thermal conductivity) and capacities (e.g., water retention, heat capacity) cannot be considered constants. The changing stocks in a soil compartment (e.g., contents of clay, carbon, minerals, elements in various pools) (Fig. 3) can be used to feed pedotransfer functions (Link to Encyclopedia Chapter on PTFs here) to update these transport-related properties.

These pedon-scale models, essentially enhanced leaching models, need quantitative descriptions of the conditions to be met at the system boundaries, both in a temporal and in a spatial sense. This makes them responsive to the factors of soil formation that operate at the system boundaries over time T (Figs. 2 and 3): At $T = 0$, Parent material = texture + unweathered phase; along the timeline Climate = $P + PE + T$, Organisms = vegetation + bioturbation + tillage + cropping + fertilization. Pedon models, in essence, describe the vertical transport of matter, so surface mass redistribution processes are not represented and erosion and deposition are effectively also boundary conditions.

The outputs of pedon models are time-varying depth distributions of various soil properties. These still have to be converted to horizons if that is required. As an example, Finke et al. (2013) developed rules, for a Belgian loess soil that post-process time–depth patterns of contents of organic carbon, clay and calcite to time–depth patterns of A, E, Bt and C-horizons.

Strong points of pedon models at present are:

1. A fairly strong process coverage (Minasny et al., 2015), allowing simulation of the formation of various soil types;
2. Inclusion of water flow as a driving agent of soil development, making pedon models suitable to simulate consequences of global change;
3. Mechanistic process description: the physical laws in these models reduce the risk of unsound extrapolations in case of extreme boundary inputs (e.g., climate scenarios) when compared to empirical models.

Weak points of current pedon models are:

1. Assumptions on a constant volume of compartments, which limits situations with volume change (weathering, strong argillo- and bio-turbation) and restricts tillage land use scenarios;
2. Absence of redox processes;
3. Runtime.

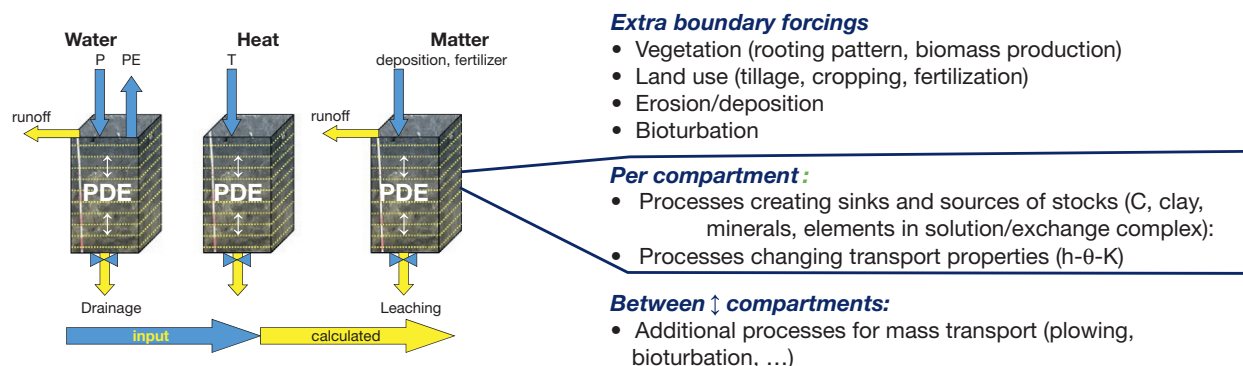


Fig. 2 Graphical representation of SoilGen, a Pedon-scale model. PDE, Partial differential equation; P, precipitation; PE, potential evapotranspiration; T, air temperature.

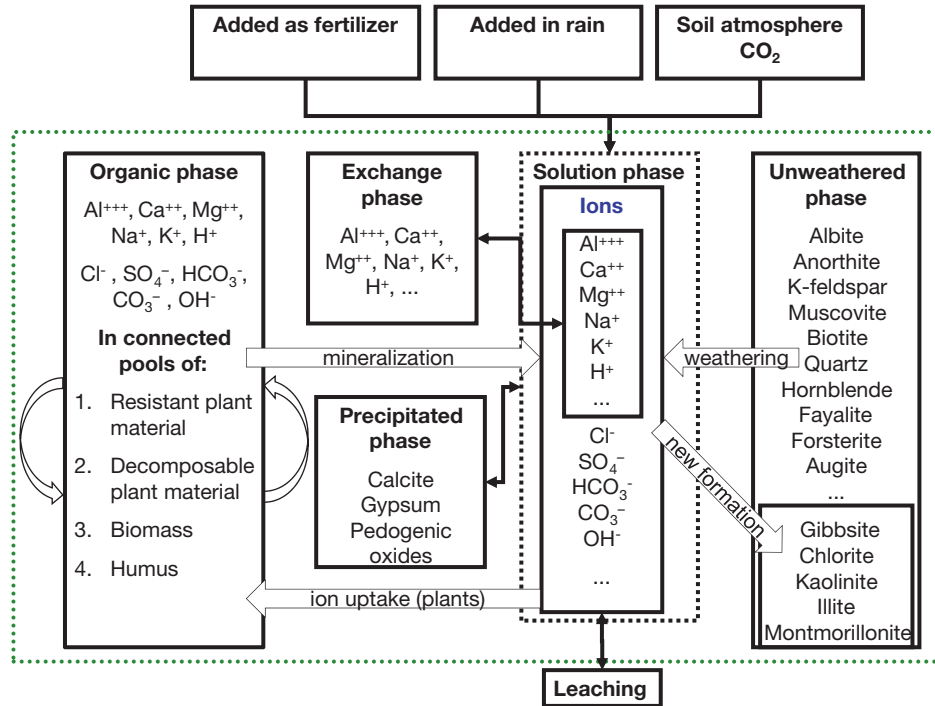


Fig. 3 Phases in which elements can occur in soils, and associated conversion processes, from the SoilGen2-model. Note that the C-cycle is incompletely represented. Updated after Finke, 2012.

Focus on landscape

As opposed to detailed storage and fluxes of materials over the full soil depth in the profile models, landscape evolution models mainly consider the change in a landscape's shape, i.e., elevation due to material redistribution (Fig. 4) at the soil surface. The simplest model (assuming no tectonic change) can be written as:

$$\frac{\partial z}{\partial t} = -\nabla q \quad (1)$$

where z is elevation [L] and q is material flux [$L^3 T^{-1}$], and ∇ is a partial derivative vector over space. This model is used widely in geomorphology studies, yet it neglects soil as part of the system. This problem was addressed by Carson and Kirkby and later by Dietrich et al. (see Minasny et al., 2015) where soil needs to be weathered first from bedrock before it can be transported. The model is written as:

$$\frac{\partial h}{\partial t} = \frac{\rho_r}{\rho_s} \frac{\partial e}{\partial t} - \nabla q \quad (2)$$

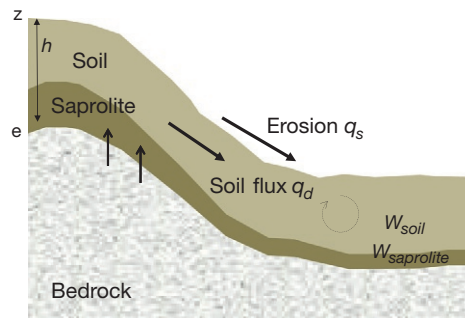


Fig. 4 Conceptual model of processes in landscape evolution. h is soil thickness, q are fluxes and W are weathering zones. Modified after Dixon and von Blanckenburg (see Minasny B, Finke P, Stockmann U, Vanwallieghem T and McBratney A (2015) Resolving the integral connection between pedogenesis and landscape evolution. Earth-Science Reviews 150: 102–120, modified Fig. 1b, p.104).

where h is soil thickness [L], e is the boundary between soil and bedrock [L], ρ_s is the density of soil [M L^{-3}] and ρ_r is the density of rock.

The term for material flux can be further decomposed into soil erosion q_s (removal of soil from the surface) and soil transport q_d (the transport of soil) (Willgoose, 2018)

$$\frac{\partial h}{\partial t} = \frac{\rho_r}{\rho_s} \frac{\partial e}{\partial t} - \nabla q_s - \nabla q_d \quad (3)$$

This type of rudimentary model was explored by Minasny and McBratney (Minasny et al., 2015) where the soil evolution in a landscape can be simulated realistically. This basic model combines three processes: weathering, surface soil erosion and transport of soil materials downslope.

Weathering

Weathering functions of rock to soil $\partial e / \partial t$ have been conceptualized, such as the weathering of rock to soil as an exponential decline with soil thickness:

$$\frac{\partial e}{\partial t} = -P_0 \exp(-kh) \quad (4)$$

where P_0 [L T^{-1}] is the potential weathering rate of bedrock, and k [T^{-1}] is an empirical constant. This model describes the weathering of rock that slows down with increasing soil thickness due to the decrease in temperature amplitude and reduction in water infiltration. Another conceptual model is the humped function of Humphreys and Wilkinson (Minasny et al., 2015) which hypothesizes that there is an optimum soil thickness for maximum soil production. Thus, soil production is not maximal at zero thickness as there is considerable instability in reaction and water is not available to accelerate the process.

Real world measurement of soil production only started with the availability of terrestrial cosmogenic nuclide (TCN) dating (Nishiizumi et al., 1991). Heimsath and coworkers (Minasny et al., 2015) measured cosmogenic nuclides of the rock–soil interface and defined rates of soil production where soil thickness is in a steady-state condition between production and erosion. There is a distinction between the rate of soil weathering and that of soil production. The weathering rate defined in pedology includes weathering of parent materials and soils due to physical, chemical, and biological processes. In contrast, the rate of soil production refers to the conversion of consolidated bedrock to soil, which results in thickening of the soil profile.

Stockmann and coworkers (Minasny et al., 2015) found that although rates of soil production vary across studies from different environments, they show an exponential decline with increasing soil thickness. And soil production in Australia appears to be similar in range to that in other parts of the world. Therefore, they were able to derive a global rate of (mm kyr^{-1}) soil production:

$$e/t = 114 \pm 11 \exp(-2.05 \times h) \quad (5)$$

where h is soil thickness in mm. Nevertheless, the production function can be defined to vary spatially (Saco et al., 2006 in Minasny et al., 2015) and should be expected to relate to properties of the local climate. Recent models now distinguish between the weathering of soil (W_{soil}), which includes chemical weathering and saprolite weathering ($W_{\text{saprolite}}$). The soil production function can be equated to saprolite weathering. In models that use this distinction, the thickness of soil and saprolite are simulated at each location.

Erosion and soil flux

In the simplest form, soil erosion (∇q_s) is a power law of slope steepness and upstream area (termed the stream power approach, Tucker and Hancock, 2010). Common refinements include the impact of sediment as both abrading agent and protector of bedrock (Whipple et al., 2000). More advanced models now use cellular-automata where the landscape can be finely discretized in 3-D (Davy and Lague, 2009). Transport of materials can be modeled explicitly in each voxel, creating a more realistic landscape that respects mass-balance.

Focus on soilscape

Recent soilscape evolution models try to combine the features of pedon-scale and landscape evolution models by incorporating both pedogenic and geomorphic processes. The potential advantage of such models is that they allow exploration of the myriad of feedbacks between soil formation and landscape evolution. A key downside is the need to deal with multiple soil components (e.g., multiple texture classes, organic matter pools, bulk density) in both vertical and lateral dimensions. This places large demands on computer memory and processing power. The first soilscape evolution model, MILESD (Vanwalleghe et al., 2013) dealt with this problem by choosing master soil horizons (A, B, C) for its vertical discretization rather than thinner (and more numerous) layers. These horizons would change in thickness over time, depending on the rates of processes that affected soils in different landscape positions. Choosing soil horizons is additionally advantageous because it allows the use of large sets of empirical data based on soil horizonation for model calibration and validation. MILESD was used to simulate soil–landscape development of a catchment in Australia, with simulations starting from a homogenous soil across a hypothetical landscape. Simulations resulted in a distribution of soils across the landscape (virtual catenas) that corresponded to expectations (Fig. 5, taken from Vanwalleghe et al., 2013).

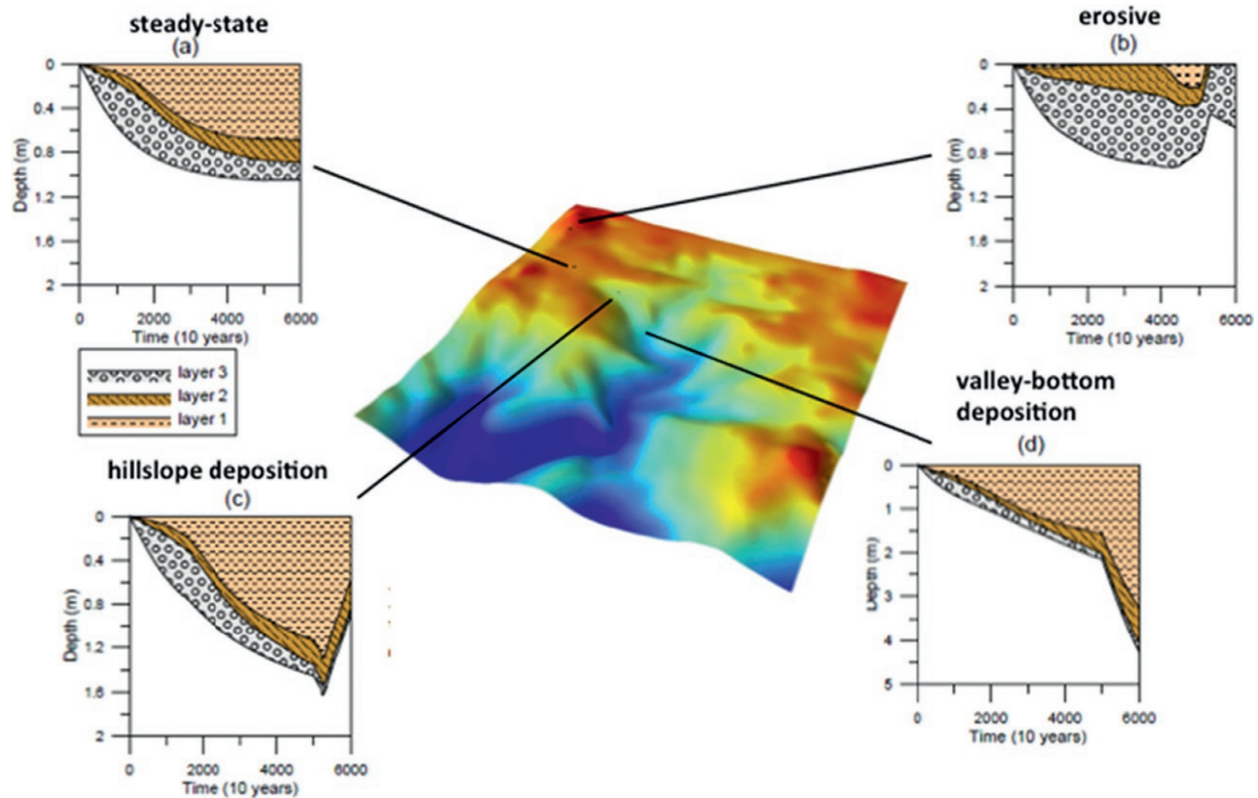


Fig. 5 Evolution of soil horizons in four different landscape positions, simulated with MILESD. Reproduced from Minasny B, Finke P, Stockmann U, Vanwallegghem T and McBratney A (2015) Resolving the integral connection between pedogenesis and landscape evolution. *Earth-Science Reviews* 150: 102–120, Fig. 7, p.104.

A downside of simulating horizons explicitly is the lumping of much soil variation within horizons, and the imposition of one set of soil horizons over the entire landscape, whereas in reality many different sets of horizons may exist in different landscape positions.

A successor model to MILESD, Lorica (Temme and Vanwallegghem, 2015), uses vertical discretization in layers thinner and more numerous than horizons that can vary in thickness in one temporal and three spatial dimensions as a result of pedogenic and geomorphic processes. Lorica also included further detail in the representation of erosion and deposition by merging a transport- and detachment-limited process (Davy and Lague, 2009) with grainsize specificity (Hjulstrom, 1935). The main benefit of these smaller layers that change thickness in response to the erosion, deposition and translocation of the materials in them is in avoiding the loss of information that would follow from fixed layer thickness (see above).

Both MILESD and Lorica mostly run at annual timescales and therefore need extensive calibration to be able to simulate pedogenic and geomorphic processes that occur at much smaller timescales. In addition, even in Lorica, layers are in general currently too thick to use the Richards equation for waterflow—even if computational requirements are met. As a result, soil hydrology is essentially absent and soil hydraulic parameters (such as the saturated water content) are hardly used. In particular, the promise of sub-surface lateral waterflow simulations remains elusive (Van der Meij et al., 2020). The benefit of that absence is that the pedogenic development at each location in the model can be calculated in parallel since locations do not depend on their neighbors other than through geomorphic processes.

A step towards a monthly consideration of surface hydrology, which is intended to enable more detailed representation of soil hydrology was recently made by Van der Meij et al. (2018). Their model HydroLorica, an adaptation of Lorica, uses different timescales to simulate realistically landscape-scale variation in infiltration rates, vertical subsurface waterflow and hence rates of some pedogenic processes (Fig. 6). This is a precondition for improving soil hydrology, particularly for subsurface lateral waterflow.

Despite the computational shortcomings that currently limit the further development of four-dimensional soilscape evolution models, their value for exploring different scenarios of climate and land use change is already clear. Van der Meij et al. (2020), for instance, used HydroLorica to show the development of natural and used landscapes under a range of climates leading to grassland and forest cover, with simultaneous resolution of the changes in landscapes due to erosion, tillage and tree fall, and the changes in soils due to erosion, tillage and translocation.

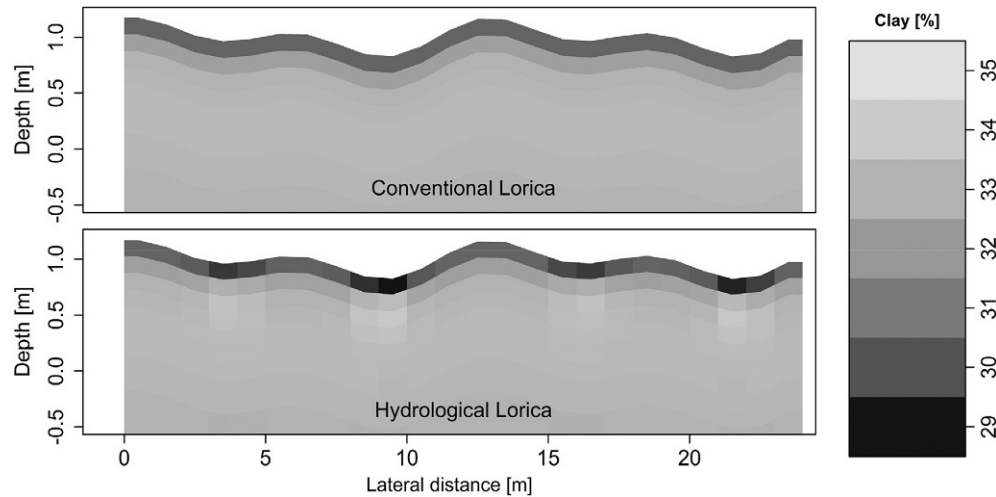


Fig. 6 Simulated clay content in an undulating transect, showing the impact of dealing with hydrology at monthly resolution and the resulting spatially variable infiltration and clay translocation. Reproduced from Van der Meij WM, Temme AJAM, Lin HS, Gerke HH and Sommer M (2018) On the role of hydrologic processes in soil and landscape evolution modeling: concepts, complications and partial solutions. *Earth-Science Reviews* 185: 1088–1106, Fig.7 p.1098.

Space–time boundary conditions for soilscape models

Initial condition

The initial condition comprises values of all state variables, often stocks, presented in the soil(-scape) model at the beginning of the simulation period, both with depth and spatially. Models with high process detail usually require more state variable values at $T = 0$ than simpler (or empirical) models. Table 1 gives some examples of state variables and also names important variables related to transport of, for example, water, heat, soil mass that are based on the initial state variables.

The importance of correct values of initial state variables depends on various factors:

- Process dynamics: Highly dynamic processes will quickly render the initial value less important because a new steady state may be reached rapidly;
- Model sensitivity to changes in these state variables;

Table 1 Examples of state variables in soilscape and pedon models that must be initialized.

State variables input at $T = 0$	Pedon models	Soilscape models
<i>At pixel/site level</i>		
Elevation and derivatives (e.g., slope)	–	TTLEM ^a , HydroLORICA ^b
Soil depth	–	SSSPAM ^c , mARM3D ^d , TTLEM ^a , HydroLORICA ^b
Vegetation	SoilGen ^e , Witch-Aspects ^f	HydroLORICA ^b
<i>Per depth segment</i>		
Grain size distribution	SoilGen ^e , Witch-Aspects ^f	SSSPAM ^c , mARM3D ^d , TTLEM ^a , HydroLORICA ^b
Bulk density of soil (and rock) and segment thickness	SoilGen ^e , Witch-Aspects ^f	TTLEM ^a , HydroLORICA ^b
Vertical transport-related parameters	SoilGen ^e , HP1 ^g	HydroLORICA ^b
Lateral (surface) mass transport parameters	–	HydroLORICA ^b
Organic carbon content	SoilGen ^e , Witch-Aspects ^f	HydroLORICA ^b
Cations and anions in solution	SoilGen ^e	–
Cations at exchange complex and CEC	SoilGen ^e , Witch-Aspects ^f , HP1 ^g	–
Minerals (contents and stoichiometry)	SoilGen ^e , Witch-Aspects ^f	–
Volumetric water content	SoilGen ^e , Witch-Aspects ^f , HP1 ^g	–
Temperature	SoilGen ^e	–

^aCampforts et al., 2017.

^bVan der Meij et al., 2020.

^cWelivitiya et al., 2019.

^dCohen et al., 2010.

^eFinke et al., 2018.

^fGoddéris et al., 2006.

^gJacques et al., 2008.

- Simulation scenario: The dynamics at the upper soil boundary, duration of the simulation period and the occurrence (or not) of soil rejuvenation by erosion or deposition bringing fresh parent material close to the surface.

The initial condition represents a status in the past, therefore, it must be obtained by reconstruction if a real soil(-scape) is to be simulated, or by definition, in the case of a virtual situation. For soil status variables, inventories of the parent material are often the data source. These may comprise properties of the C-horizon, or that of shallow bedrock as derived from soil and geology maps and associated information systems. When the initial situation is an already developed soil and not simply the parent material, reconstruction of the properties is more complex, associated with uncertainty, and will use vertically and or laterally distributed soil data.

Similarly, the initial topography can be represented by landscapes resulting from a reconstructed geological history. Heimsath and coworkers (Minasny et al., 2015) and Campforts et al. (2016) used inventories of terrestrial cosmogenic nuclides to reconstruct soil production and redistribution spatially that could be used for landscape reconstruction. In other cases (e.g., Campforts et al., 2017) the current landscape was used as the initial topography.

Time-variant upper boundary condition

Changes at the upper boundary of the soil(-scape), i.e., the soil–atmosphere interface, influence matter or energy transfers into the modeled system. These inputs clearly relate to the factors of soil formation (Table 2). A notable difference between pedon and soilscape models is that erosion and deposition are inputs in the (1-D) pedon models, but they are the calculated results of surface mass redistribution processes in the soilscape models.

Slope and exposition are usually time-invariant in pedon models, but may change over time in soilscape models due to surface mass redistribution. In both cases they are not true boundary conditions but may still vary over time.

As a consequences of boundary inputs, rates at which soil processes take place are influenced. For instance, organic matter decomposition rates are influenced by water content and temperature, chemical reaction constants are modulated by temperature changes, and so on.

The importance of time-variant upper boundary conditions is that they allow the reconstruction of soil development to the present using reconstructed changes in soil forming factors, and also enable the calculation of how future global change will affect the soil(-scape). This is conditional on the processes covered by the used model (Minasny et al., 2015).

There is a rich variety of methods to obtain upper boundary inputs (Table 3). Three main cases can be identified:

1. Boundary conditions are reconstructed using proxies.

Proxies are measurable variables in soils or sediments that can be dated, e.g., via a depth–age model in the sampled material. The ages are obtained in most cases by ^{14}C dating, OSL (optically stimulated luminescence) dating or radioactive decay. Proxies can be botanical (pollen diagrams, diatoms, macrofossils, tree rings), zoological (ostracods, molluscs, chironomids) and chemical (magnetic susceptibility, stable isotopes), and are transformed to the target variable using knowledge on ecological niches or correlation to known values. Proxies can be combined to obtain better estimates of the target variable (e.g., Bos et al., 2017), and are often mapped by space–time interpolation during the estimation process as sampling sites and application sites are usually not the same (e.g., Davis et al., 2003).

Table 2 Time-variant upper boundary conditions in soil(-scape) models.

Upper boundary conditions	Processes influenced
Climate	Temperature Heat flow, correction of chemical equilibria, physical weathering Precipitation: water Water flow Precipitation: solutes Solute flow Evaporation Evapotranspiration
Organisms	Vegetation Biomass production, C-turnover, CO ₂ -diffusion, solute uptake Fauna Bioturbation Human influence: Fertilization Solute flow and -uptake Irrigation Water flow, biomass production Tillage Heat/water/solute flow modification Crop protection Biomass production
Relief	Slope Runoff Exposition Heat/water/solute flow modification
Mass redistribution	Erosion Removal of topsoil, texture grading Deposition: texture Chemical dissolution/precipitation, C-cycling, clay migration Mineralogy Mineral weathering and new formation Solute/exchange chemistry Chemical equilibria, cation exchange equilibria Change of boundary conditions All of the above
Time	

Markings  indicate influence by global change.

Table 3 Four examples of proxy- or model-reconstructed boundary conditions for soil models.

<i>From proxy data source</i>	<i>To</i>	<i>References</i>
Pollen diagrams, diatom assemblages	Vegetation; precipitation; potential evapotranspiration; summer- and winter temperature	Davis et al. (2003)
Magnetic susceptibility and remanescence	Precipitation	Maher and Thompson (1994)
<i>From earth system model</i>	<i>To</i>	<i>References</i>
HadGEM, IPSL-CM5A, MIROC-ESM-CH + Orchidee	Temperature; precipitation Net primary production	Keyvanshokouhi et al. (2019)
LOVECLIM	Natural vegetation; monthly average precipitation, potential evapotranspiration, temperature	Finke et al. (2018)

2. Boundary conditions are constructed using models of other earth system compartments.

Earth system models describe a larger part of this planet than the soil and therefore are useful to obtain inputs at the soil(-scape) upper boundary. In contrast to proxies, they can be used for both reconstruction (e.g., paleoclimatic models; Finke et al., 2018) and projection (e.g., Keyvanshokouhi et al., 2019) studies.

3. Boundary conditions are constructed using historical documents.

In some cases, and mostly for land use reconstructions, historical maps, observations and documents can be consulted to obtain boundary conditions but their temporal range is usually restricted to a few centuries or less.

Time-variant lower boundary conditions

While the upper boundary of a modeled soil system is the soil–atmosphere interface, the depth of the lower boundary of the system can be set in various ways, for instance (i) at the interface between unconsolidated (i.e., sediment, regolith) or consolidated material (i.e., bedrock), or (ii) at arbitrary, fixed depths. Over pedogenetic timescales, the interface between regolith and bedrock may move downwards due to soil production or sediment addition and it may move upwards due to erosion and denudation processes. This is a complication in terms of setting up a model, and therefore often a fixed depth is chosen that is large enough to absorb the effects of erosion and to include the A and B horizons minimally.

The most important boundary conditions that need to be specified are related to the hydrology and temperature of the modeled system.

Common bottom boundary conditions for water flow are: (i) the bottom flux is defined by a constant or variable water table, (ii) free drainage (deep or no groundwater table; the bottom flux is restricted by the hydraulic conductivity of the lowest compartment at unit hydraulic head gradient), (iii) zero drainage (impermeable layer).

In case (i), input of variation in the water table over time is required. This may be relevant when agricultural water management is implemented, e.g., by artificial drainage. Another way to input variation in the water table is using option (iii) where the layer at the average water table depth in a particular period is assigned a low permeability (e.g., Zwertvaegher et al., 2013). This is one way to couple a 3-D groundwater model to a 1-D-pedon model taking regional groundwater variations into account.

Most studies apply free drainage at the lower boundary though some examples exist where the bottom boundary condition switches between free drainage and zero drainage as a function of the aridity of the climate (e.g., Finke and Hutson, 2008).

Common bottom boundary conditions for heat flow are: constant temperature, a zero heat flux condition, or a heat reservoir. The heat reservoir is the most likely to be applied, as it dampens but does not exclude temperature change in the deep soil layers, though when modeling soil development in permafrost soils other options might be considered.

Heterogeneity

Heterogeneity can partly be accommodated by setting the initial conditions to reflect lateral and vertical variations in parent materials, but in some cases new heterogeneity can be caused by disturbances during a soil(-scape) simulation, and then affect the boundary conditions. Examples are:

- Tectonic and volcanic events. For tectonic events, at the top boundary, elevation and slope may respond, causing mass movements (Coulthard and van de Wiel, 2013). At the bottom boundary, mainly the water flow regime is affected. Volcanic events are largely unexplored with models, but river blockades and the formation of temporary lakes would have to be accounted for (Van Gorp et al., 2016).
- Interventions in the local and regional water management. At the top boundary, an irrigation boundary condition may become active. At the bottom boundary, mainly the water flow regime is affected when artificial drainage conditions are implemented.
- Permafrost thawing. At the bottom boundary, both the water flow regime and the heat flow regime may be affected once permafrost has disappeared, the active zone deepens or a talik is formed.

Identifying and minimizing uncertainty

Accuracy assessment

All models are associated with uncertainty, and therefore assessment of the accuracy of prediction of its major output variables is essential. Soil development models cover substantial temporal extents, varying between a few decades to millennia. For accuracy assessment, comparisons to field data are needed, which represent the current state of the soil system. Such assessment might not indicate whether all process constants are well-set, as various combinations of these constants might lead to similar accuracies. This problem of non-uniqueness may be partly avoided in two situations:

1. A chronosequence is present.

A soil chronosequence contains soils covering different periods in the climate and land use history. The overall effects of different histories of boundary conditions as well as the duration of soil development are then represented in both soil data and simulation results. Sauer and coworkers (Minasny et al., 2015) assessed model accuracy in a glacial rebound chronosequence in Norway for soil ages between 2100 and 11,050 years and concluded that deviations between various simulated and measured soil variables developed within 2100 years. As an alternative, if the study area includes a climatic gradient this may be used to evaluate model behavior under different boundary inputs. Buried paleosoils with boundary conditions derived from a paleoclimate model may also be used for accuracy assessment.

2. Monitoring data exist over a temporal extent comparable to the temporal autocorrelation length of the soil variable studied. This may be the case for soil organic carbon models that are compared to long-term field experiments.

Minasny et al. (2015) analyzed 29 soil(-scape) development model studies and found that pedon models are more often (11/14) compared to field data than soilscape models (5/15).

Soil and elevation data for calibration and accuracy assessment

Soil(-scape) evolution models can be compared to data on different spatial supports and extents, and can use both legacy data (with known age) and recent data. Soil inventories may be up to 75 years old, therefore, it may be useful to account for their age by selecting model outputs for the same period (e.g., Zwertvaegher et al., 2013) to avoid the possibility that recent land use change is not represented by the soil data.

- Landscape evolution models are frequently compared with digital topographic datasets to assess their validity. This works in two cases: when current landscapes are considered in equilibrium with their tectonic and climatic boundary conditions (in which case models run from a hypothetical starting landscape until they, too, reach equilibrium, Taylor Perron and Hamon, 2012), or when topography is well known for a point in the past as well as in the present (in which case models run from the past topography to the present—rarely the other way around; see work by Temme et al., in Minasny et al., 2015). In both cases, metrics of current day topography are extracted from digital topographic datasets and from model outputs (e.g., using TopoToolbox; Schwanghart and Scherler, 2014), and are then compared.
- Soilscape evolution models can similarly be evaluated with soil maps and digital topographic datasets. Either of the two conditions expressed above would have to be met on the topographic side. When assuming equilibrium, observed soil patterns would additionally have to reflect the developmental endpoint in a landscape in equilibrium with tectonics and climate, to be comparable to simulated soil patterns. When using a well-known starting landscape, soil patterns at that past point in time would also have to be known accurately and used as a boundary condition in the soil-landscape evolution model. A soil map can be used for quality testing of a simulated soilscape after conversion of the soilscape model output to the legend of the soil map. This could be done under the assumption that the soil map is without error by counting the fraction of the area with equal classification in both soil map and modeled soilscape (confusion matrix). However, no examples of such quality tests are yet known. The few known soilscape-soil map comparisons are based on datasets of pedons (e.g., Vanwalleghe et al., 2010; Zwertvaegher et al., 2013). The spatial coverage of such pedon datasets may be increased by one of the many methods under the umbrella of digital soil mapping (DSM), in which case uncertainty of the spatial soil data should be accounted for while testing model accuracy.
- Pedon models can be compared to field data obtained at the simulation locations. Minasny et al. (2015) reported various examples, where the most common soil variables to be tested were clay content, organic carbon content, soil pH, CEC and calcite content (Opolot, 2016).

Calibration

To improve model accuracy, sensitivity studies followed by model calibration can help. Sensitivity studies help to identify the model process constants that are most promising for calibration. There are many methods for sensitivity analysis (cf. Minasny et al., 2015) and the most feasible method largely depends on the computational demands by the soil(-scape) model. Exhaustive sensitivity analyzes, where all combinations of model parameters are explored, are usually beyond computational possibilities. A random sampling of the model parameter space has so far been used instead (e.g., Temme and Vanwalleghe, 2015).

Calibration can be done on a parameter by parameter basis, starting with the most sensitive process parameter and proceeding with increasingly less sensitive parameters. This approach is feasible when the computational demands of the model are high, which is often the case in soil(-scape) development models. More advanced methods (Minasny et al., 2015) that simultaneously vary several (sensitive) process parameters are only feasible with fast models.

Several studies have shown that the quality after calibration shows sensitivity to the boundary and initial inputs over the simulation period. This suggests that calibrated models cannot be transported to different settings. Keyvanshokouhi et al. (2016) showed that uncertainty in estimating the agricultural history of a site affects the quality after calibration and also demonstrated the effect of the uncertainty of the initial soil texture. Ongoing research suggests a strong relation between the (uncertain) dust deposition history and quality after calibration in the Chinese Loess Plateau.

To summarize, in soil development models the quality after calibration is conditional on the accuracy of various boundary and initial conditions. In synthetic scenario studies, these conditions are not reconstructed but defined. Therefore part of the uncertainty is reduced in these studies but at the cost of being relevant for a particular region. In particular, the computationally intensive soilscape evolution models may be restricted to this exploratory model use for the near future.

Outlook

Pedon and soilscape development models will be suitable instruments for evaluating scenarios of global change if they are hydrology-enabled, describe the soil-C cycle and accept land use management as upper boundary conditions. This stage has now been reached for some pedon and soilscape development models. Coupling with climate models has led to realistic scenarios for past (Finke et al., 2018) and future (Keyvanshokouhi et al., 2019) soil development. A necessary next step will be to consolidate this situation by increasing the set of models with this ability, which will increase the competition between such models via model comparison and lead to improvements. The work of the climate modeling community can be used as a template. Useful future activities are model testing on standardized data sets. Additionally, a larger suite of models will enable assessment of the effects of global change on soils in a similar way to that by the climate modeling community: by ensemble modeling. This stage has not yet been reached.

From the perspective of understanding soil–landscape relations, ideally, process-based models would be used to map how soils vary spatially in the landscape. This would remove the empiricity of DSM. Some cases already exist: Bonfatti et al. (2018) tested whether a simple landscape evolution model can predict soil thickness in a valley of the Rio Grande do Sul, Brazil. The results are promising but also have limitations due to erosion processes at short time scales and uncertainty involving soil formation of the area. Lozbennev et al. (2021) used a hydrological model to calculate runoff over the catchment. They found that the simulated runoff alone can explain the variation of the depth of secondary carbonates in soils and the soil taxa. Finke and coworkers (Minasny et al., 2015) used a distributed pedon model to find causes for “noisy” soil–landscape relations in a loess region in Belgium and identified treefalls as the likely source of the noise. The link between DSM and soilscape modeling is becoming closer, as process knowledge is being used increasingly in DSM (e.g., Angelini et al., 2016) and DSM can be used to identify major processes in a particular soilscape (e.g., Baltensweiler et al., 2020).

References

- Angelini ME, Heuvelink GBM, Kempen B, and Morrás HJM (2016) Mapping the soils of an Argentine Pampas region using structural equation modelling. *Geoderma* 281: 102–118.
- Baltensweiler A, Heuvelink GBM, Hanewinkel M, and Walthert L (2020) Microtopography shapes soil pH in flysch regions across Switzerland. *Geoderma* 380: 114663.
- Bonfatti BR, Hartemink AE, Vanwalleghe T, Minasny B, and Giasson E (2018) A mechanistic model to predict soil thickness in a valley area of Rio Grande do Sul, Brazil. *Geoderma* 309: 17–31.
- Bos JAA, De Smedt P, Demiddele H, et al. (2017) Multiple oscillations during the Lateglacial as recorded in a multi-proxy, high-resolution record of the Moervaart palaeolake (NW Belgium). *Quaternary Science Reviews* 162: 26–41.
- Campforts B, Vanacker V, Vanderborght J, Baken S, Smolders E, and Govers G (2016) Simulating the mobility of meteoric ¹⁰Be in the landscape through acoupled soil-hillslope model (Be2D). *Earth and Planetary Science Letters* 439: 143–157.
- Campforts B, Schwanghart W, and Govers G (2017) Accurate simulation of transient landscape evolution by eliminating numerical diffusion: The TTLEM 1.0 model. *Earth Surface Dynamics* 5: 47–66.
- Cohen S, Willgoose G, and Hancock G (2010) The mARM3D spatially distributed soil evolution model: Three-dimensional model framework and analysis of hillslope and landform responses. *Journal of Geophysical Research-Earth Surface* 115: F04013.
- Coulthard TJ and van de Wiel MJ (2013) Climate, tectonics or morphology: What signals can we see in drainage basin sediment yields? *Earth Surface Dynamics* 1: 13–27.
- Davis BAS, Brewer S, Stevensson AC, Guiot J, and Data Contributors (2003) The temperature of Europe during the Holocene reconstructed from pollen data. *Quaternary Science Reviews* 22: 1701–1716.
- Davy P and Lague D (2009) The erosion/transport equation of landscape evolution models revisited. *Journal of Geophysical Research, Earth Surface* 114(F3).
- De Vries W, Kros J, and Voogd JCH (1994) Assessment of critical loads and their exceedance on Dutch forests using a multilayer steady-state model. *Water, Air, & Soil Pollution* 76: 407–448.
- Finke PA (2012) Modeling the genesis of Luvisols as a function of topographic position in loess parent material. *Quaternary International* 265: 3–17.
- Finke PA and Hutson J (2008) Modeling soil genesis in calcareous löss. *Geoderma* 145: 462–479.
- Finke PA, Vanwalleghe T, Opolot E, Poesen J, and Deckers J (2013) Estimating the effect of tree uprooting on variation of soil horizon depth by confronting pedogenetic simulations to measurements in a Belgian loess area. *Journal of Geophysical Research - Earth Surface* 118(4): 2124–2139.
- Finke PA, Yin Q, Bernardini NJ, and Yu Y (2018) Climate-soil model reveals causes of differences between MIS 5e and MIS 13 paleosoils. *Geology* 46: 99–102.

- Goddéris Y, François LM, Probst A, et al. (2006) Modeling weathering processes at the catchment scale: The WITCH numerical model. *Geochimica Cosmochimica Acta* 70: 1128–1147.
- Hansen S, Jensen HE, Nielsen NE, and Svendsen H (1991) Simulation of nitrogen dynamics and biomass production in winter-wheat using the Danish simulation-model DAISY. *Fertilizer Research* 27: 245–259.
- Hjulstrom F (1935) *Studies of the Morphological Activity of Rivers as Illustrated by the River Fyris*. vol. 25, Bulletin Geological Institute Uppsala 221–527.
- Keyvanshokouhi S, Cornu A, Samouëlian A, and Finke P (2016) Evaluating SoilGen2 as a tool for projecting soil evolution induced by global change. *Science of the Total Environment* 571: 110–123.
- Jacques D, Simunek J, Mallants D, and van Genuchten MTh (2008) Modeling coupled water flow, solute transport and geochemical reactions affecting heavy metal migration in a podzol soil. *Geoderma* 145: 449–461.
- Keyvanshokouhi S, Cornu S, Lafolie F, et al. (2019) Effects of soil process formalisms and forcing factors on simulated organic carbon depth-distributions in soils. *Science of the Total Environment* 652: 523–537.
- Lozbenov N, Yurova A, Smirnova M, and Kozlov D (2021) Incorporating process-based modeling into digital soil mapping: A case study in the virgin steppe of the Central Russian Upland. *Geoderma* 383: , 114733.
- Maher BA and Thompson R (1994) Paleorainfall reconstructions from pedogenic magnetic susceptibility variations in the Chinese Loess and Paleosols. *Quaternary Research* 44: 383–391.
- Minasny B, Finke P, Stockmann U, Vanwallegghem T, and McBratney A (2015) Resolving the integral connection between pedogenesis and landscape evolution. *Earth-Science Reviews* 150: 102–120.
- Nishiizumi K, Kohl CP, Arnold JR, Klein J, Fink D, and Middleton R (1991) Cosmic ray produced ^{10}Be and ^{26}Al in Antarctic rocks: Exposure and erosion history. *Earth and Planetary Science Letters* 104(2–4): 440–454.
- Opolot E (2016) *Modeling Soil Evolution to Assess Soil System Behaviour Under Global Change*. PhD-thesis Ghent University.
- Schwanghart W and Scherler D (2014) TopoToolbox 2—MATLAB-based software for topographic analysis and modeling in Earth surface sciences. *Earth Surface Dynamics* 2: 1–7.
- Taylor Perron J and Hamon JL (2012) Equilibrium form of horizontally retreating, soil-mantled hillslopes: Model development and application to a groundwater sapping landscape. *Journal of Geophysical Research, Earth Surface* 117(F1), F01027.
- Temme AJAM and Vanwallegghem T (2015) LORICA – A new model for linking landscape and soil profile evolution: Development and sensitivity analysis. *Computers & Geosciences* 90.
- Tucker GE and Hancock GR (2010) Modeling landscape evolution. *Earth Surface Processes and Landforms* 35: 28–50.
- Van der Meij WM, Temme AJAM, Lin HS, Gerke HH, and Sommer M (2018) On the role of hydrologic processes in soil and landscape evolution modeling: concepts, complications and partial solutions. *Earth-Science Reviews* 185: 1088–1106.
- Van der Meij WM, Temme AJAM, Wallinga J, and Sommer M (2020) Modeling soil and landscape evolution—The effect of rainfall and land-use change on soil and landscape patterns. *The Soil* 6: 337–358.
- Van Gorp W, Schoorl JM, Temme AJAM, et al. (2016) Catchment response to lava damming: integrating field observation, geochronology and landscape evolution modeling. *Earth Surface Processes and Landforms* 41: 1629–1644.
- Vanwallegghem T, Poesen J, McBratney A, and Deckers J (2010) Spatial variability of soil horizon depth in natural loess-derived soils. *Geoderma* 157: 37–45.
- Vanwallegghem T, Stockmann U, Minasny B, and McBratney AB (2013) A quantitative model for integrating landscape evolution and soil formation. *Journal of Geophysical Research, Earth Surface* 118: 1–17.
- Wagenet RJ and Hutson JL (1986) Predicting the fate of nonvolatile pesticides in the unsaturated zone. *Journal of Environmental Quality* 15(4): 315–322.
- Welivitiya WDDP, Willgoose GR, and Hancock GR (2019) A coupled soilscape–landform evolution model: model formulation and initial results. *Earth Surface Dynamics* 7: 591–607.
- Whipple KX, Hancock GS, and Anderson RS (2000) River incision into bedrock: Mechanics and relative efficacy of plucking, abrasion and cavitation. *Geological Society of America Bulletin* 112: 490–503.
- Willgoose G (2018) *Principles of Soilscape and Landscape Evolution*. Cambridge: Cambridge University Press.
- Zwertvaegher A, Finke P, Smedt D, et al. (2013) Spatio-temporal modeling of soil characteristics for soilscape reconstruction. *Geoderma* 207–208: 166–179.

Climate and land surface models: Role of soil

Toby Richard Marthews^a, Holger Lange^b, Alberto Martínez-de la Torre^{a,c}, Richard J Ellis^a, Sarah E Chadburn^d, and Martin G De Kauwe^e, ^aUK Centre for Ecology & Hydrology (UKCEH), Wallingford, United Kingdom; ^bNorwegian Institute of Bioeconomy Research (NIBIO), Høgskoleveien, Ås, Norway; ^cMeteorological Surveillance and Forecasting Group, DT Catalonia, Agencia Estatal de Meteorología (AEMET), Barcelona, Spain; ^dCollege of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, United Kingdom; ^eSchool of Biological Sciences, University of Bristol, Bristol, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

This is an update of P. Smith, CLIMATE MODELS, ROLE OF SOIL, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 262–268, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00090-4>.

Introduction	753
Standard soil model components of land surface models	753
The role of the soil profile in land surface models – how deep is a soil profile?	753
Soil moisture model components (hydraulics and hydrology)	756
Soil biogeochemical cycling models (soil nutrient and carbon dynamics)	757
Deep soil	757
Soil heat balance and thermal dynamics models	758
Extreme environments	758
Drylands and dry forests	758
Tropical moist and wet forests	759
Cold regions	759
Peatlands and wetlands	759
New directions in land surface simulation	760
Hyperresolution	760
Coupled model environments	760
Conclusions	760
References	760

Abstract

The role of soil in current climate models is reviewed and discussed, with a focus on developments over the last two decades. Soil modeling may be divided into three major parts: simulation of soil hydrological dynamics, soil biogeochemistry and the soil thermal environment. Each of these three major parts is summarized with a brief description of current best practice and developments. Specific issues and modifications relevant to four extreme environments are highlighted: drylands, tropical moist and wet forests, cold regions, and peatlands and wetlands. Finally, current advances in the areas of hyperresolution and coupled model environments are discussed, which we see as the two leading edges of current soil model development.

Glossary

Baseflow Baseflow is here taken to be the lateral outflow from soil summed over all phreatic soil layers (i.e. at and below the water table). Note that *baseflow* here does not refer to the outflow at a particular ‘basal’ level, and also does not refer to the relatively constant minimal flow of a stream or river under low or absent rainfall.

Darcy–Buckingham Law The Darcy–Buckingham Law describes how hydraulic head controls the movement of water in both saturated and unsaturated soils (Henry Darcy published **Darcy’s Law** governing the saturated case in 1856, which Edgar Buckingham extended to the unsaturated case in 1907) (Hillel, 2004).

Groundwater Groundwater is the water held in the **soil phreatic zone**, i.e. the saturated soil zone between the water table and impermeable bedrock (including all water held in rivers, lakes and the ocean) (Fig. 2) (n.b. this is distinct from subsurface-water as defined below).

Inundation Inundation is any water flowing temporarily on the soil surface but not part of a lake or water course. This includes **detention storage**, i.e. water that is in transit as runoff toward a water body.

Peatlands Peatlands are areas with a naturally accumulated layer of dead organic material (peat) at the surface (if peat is still in the process of being formed, it is also termed a **mire**) (Parish et al., 2008).

Ponding Ponding is any stationary water that is temporarily on the soil surface but not part of a lake or water course. This includes **depression storage**, i.e. water occurring in small surface depressions.

Richardson–Richards Equation The Richardson–Richards Equation is the most widely-used equation governing the movement of water in unsaturated soils (Farthing and Ogden, 2017). Note that we follow Raats and Knight (2018) in referring to both authors in the name of this equation: Lewis Fry Richardson and Lorenzo Richards, who published effectively the same equation in 1922 and 1931, respectively.

Soil moisture/soil water Soil moisture / soil water is the water content of the soil profile. Note that we define the soil column to stretch down to impermeable bedrock, therefore this term includes all groundwater (as defined below).

Subsurface-flow Subsurface-flow is the lateral outflow from soil summed over all vadose soil layers. Note that, according to these definitions, the sum of all lateral outflow below the surface is (*Subsurface-flow* + *Baseflow*) (Fig. 3), i.e. $\text{Total Runoff} = (\text{Surface Runoff}) + (\text{Below Surface Runoff}) = \text{Surface-flow} + \text{Subsurface-flow} + \text{Baseflow}$.

Subsurface-water Subsurface-water is the water held in the soil vadose zone, i.e. the unsaturated soil zone between the surface and the water table (Fig. 2) (n.b. this is distinct from groundwater as defined above).

Surface-flow Surface-flow is flowing surface-water. Note that, although it perhaps seems counter-intuitive, for consistency we take the term *surface-water flooding* to exclude river and coastal flooding.

Surface-water Surface-water is the water held above or on the land surface (including overland flow, glaciers and snowpack, but excluding water held in rivers, lakes and the ocean¹).

Terra firme Terra firme (literally *solid ground*) is land that is neither seasonally nor permanently flooded (i.e. away from large wetlands, rivers or lakes). Accepted in English since the 1600s, this term is currently widely used in Brazil (e.g. *terra firme forest*), but was originally used to refer to the mainland territories of the Republic of Venice, Italy.

Water table/Phreatic surface The water table/phreatic surface is the surface below which all fractures, pores and other voids are saturated with water. Under some conditions, a capillary fringe or tension-saturated zone can form immediately above the water table where the soil is saturated with water, yet the water remains under tension.

Wetlands Wetlands are areas that are inundated or saturated by water at a frequency and for sufficient duration to support emergent plants adapted for life in saturated soil conditions (Parish et al., 2008) (n.b. all mires are wetlands, but not all wetlands are peatlands and not all peatlands are wetlands).

Key points

- Soil modeling may be divided into three major parts: simulation of soil hydrological dynamics, soil biogeochemistry and the soil thermal environment
- Each of these three major parts is reviewed and a summary provided of current best practice and developments
- Specific issues and modifications relevant to four extreme environments are highlighted: drylands, tropical moist and wet forests, cold regions and peatlands and wetlands
- Finally, current advances in the areas of hyperresolution and coupled model environments are discussed

Nomenclature

DEM	A Digital Elevation Model (DEM; aka a Digital Terrain Model, DTM) is an elevation map; a representation of the physical topography of a landscape. DEMs are generally derived from Digital Surface Models (DSMs), which are based on remote-sensing data collected by planes, drones and satellites. DSMs follow the highest level of all elements of the land surface (imagine gently laying a sheet over the Earth's surface). A DEM may be derived from a DSM by subtracting anthropogenic surface elements (e.g. buildings, bridges) as well as canopy height (in vegetated areas) and the depths of free water elements existing on the surface (e.g. glaciers, snow; q.v. Fig. 2). However, n.b. because of the relative unreliability of bathymetry data, especially for smaller rivers, and the depth of small water bodies is often not known and thus not subtracted for DEMs follow level L3 only on terra firme (q.v. Fig. 2), and actually follow the surfaces of rivers, glaciers, lakes and oceans elsewhere.
ESM	Earth system models (ESMs) are models that simulate the coupled interactions between the atmosphere, ocean, ice and land surface. Several ESMs are under development around the world and are usually built up in a modular fashion from pre-existing models that simulate individual components.
GEWEX and WCRP	The Global Energy and Water Cycle Exchanges Project (GEWEX) is the core project in the World Climate Research Programme (WCRP) concerned with studying the dynamics and thermodynamics of the atmosphere and interactions with the Earth's surface.

¹It may seem odd to categorize river, lake and ocean water as groundwater rather than surface-water, but this is a consequence of the definition of groundwater being water below the water table (Fig. 2).

ISMC	The International Soil Modeling Consortium (ISMC) aims to integrate and advance soil systems modeling, data gathering and observational capabilities at a global level.
LSM:	Land surface models (LSMs) are the part of Climate Models or ESMs that simulate processes happening at the Earth's surface. LSMs include standalone models as well as the component models of Climate and ESMs that are concerned with the dynamics of the land surface.
SOC and SOCD	Soil Organic Carbon (SOC) (% of dry soil by mass) is the amount of carbon (C) contained in SOM, where $SOC = c_{SOM} * SOM$ (approximately 50% by mass, i.e. $c_{SOM} \approx 0.50$, Pribyl (2010) , where $(1/c_{SOM})$ is the <i>van Bemmelen factor</i>). In a soil layer of depth D (m) it is most usual to quote either SOC or Soil Organic Carbon Density, SOCD (t C/ha), which is defined by $SOCD = SOC * DBD * D * 10000$, where DBD is the dry bulk density of the layer (g/cm^3 or t/m^3) (Rodeghiero et al., 2009).
SOM	Soil Organic Matter (SOM) (% of dry soil by mass) is composed of soil microbes including bacteria and fungi, decaying material from plants and animals, fecal material, and products formed from their decomposition (Campbell and Paustian, 2015).

Introduction

Soil plays a crucially important role in sustaining life on Earth. Many aspects of the complexity of soil and the dynamics of its constituents are well-understood ([Hillel, 2004](#)), but under climate change soil properties are beginning to change, fundamentally affecting their associated dynamics and the services soils provide to humans. In addition, climate change is bringing us into a new era of more frequent climate extremes ([IPCC, 2021](#)) and soils are being affected by large-scale processes that are generally not yet included in soil scientific models ([Kendon et al., 2021](#)). We provide here an overview of the current representation of soil processes in climate models and associated land surface models (LSMs). We identify the dominant soil-related issues that are of concern across the globe, and provide links to data sources and new research that summarize current innovations in this fast-moving field.

Standard soil model components of land surface models

A reliable characterization of the soil in terms of water balance, nutrient status and thermal regime is necessary for the simulation of soil in LSMs ([Fig. 1](#)). Global-scale LSMs calculate soil moisture, soil temperature and other soil variables continually during simulations, tracking also the feedback processes through which soil-mediated fluxes influence the lower layers of the atmosphere (land–atmosphere feedbacks).

The role of the soil profile in land surface models – how deep is a soil profile?

Soil profiles are theoretically taken to extend from the ground surface to impermeable bedrock (Level L1, [Fig. 2](#)). For a spatially-explicit soil simulation model, it is critically important to consider the extent of the model domain. Asking “How deep

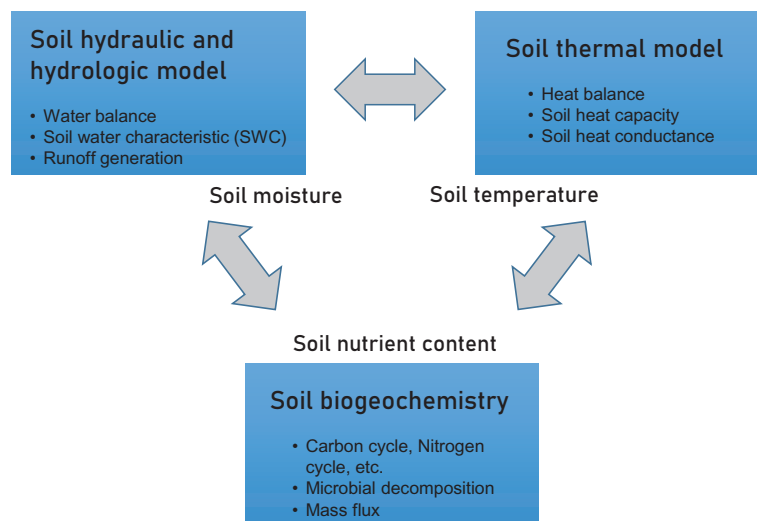


Fig. 1 A summary of soil processes of greatest relevance to LSMs and ESMs (Earth system models). Models exist that focus on all parts of this modeling sequence (see the ISMC [Model Portal](#) for a comprehensive list).

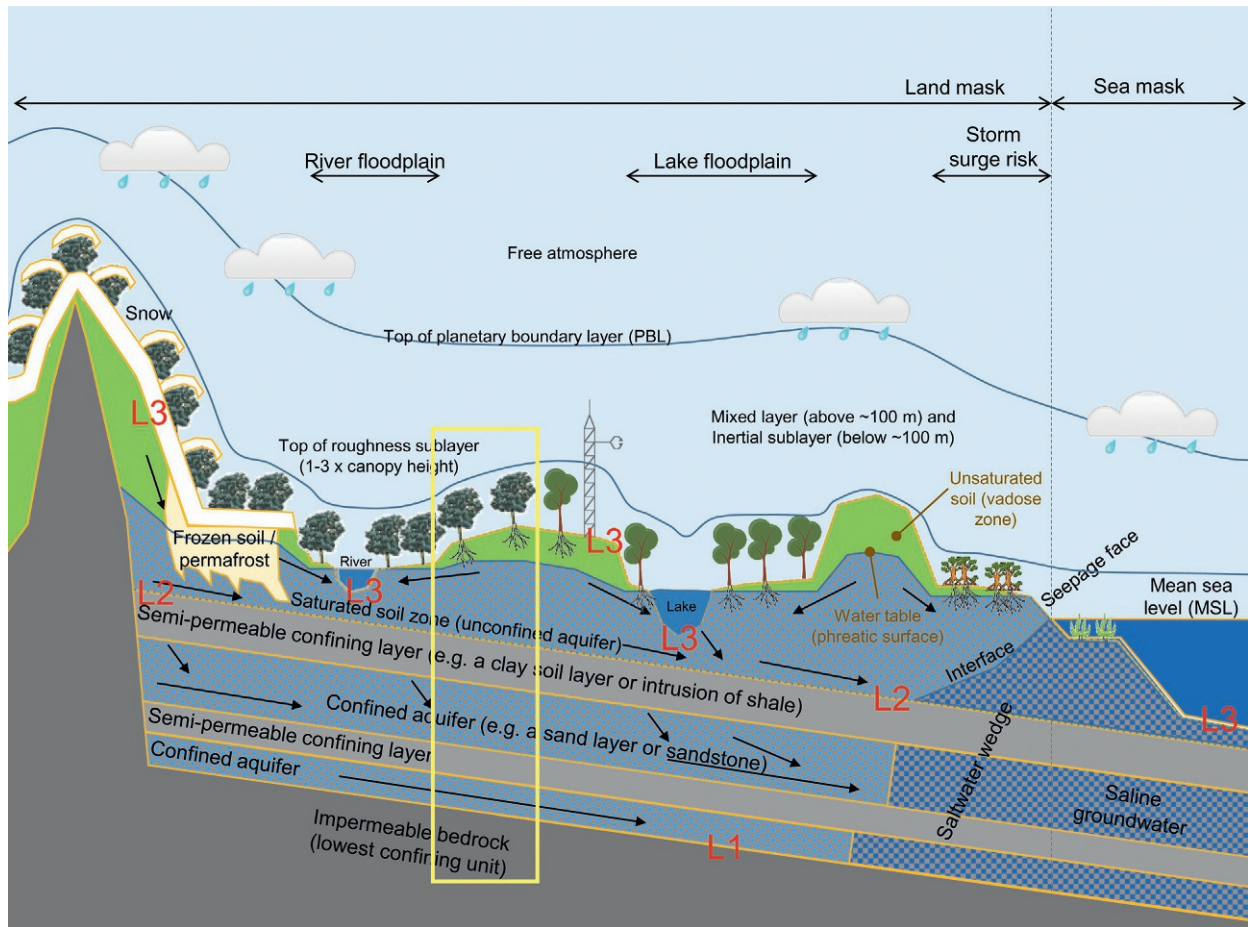


Fig. 2 A representation of standard conceptual levels in land surface models (LSMs): L1 is the bedrock surface, L2 is the surface of the shallowest confining layer (whether impermeable or semipermeable), L3 is the ground surface (underneath any free water present at the surface, i.e. glacial ice, solid snow, freshwater or seawater). Note that the water table runs below L3 over most of the land surface (fluctuating in response to rainfall), but is above L3 in rivers, lakes and the ocean because it follows the free water surface when that is above the soil. Level L1 will usually be less than 250 m below the surface, and only exceptionally down to approximately 700 m (Richter and Markewitz, 1995; Shangguan et al., 2017). The section outlined is shown in greater detail in Fig. 3.

is the soil?" is a deceptively simple question (Richter and Markewitz, 1995): the answer can vary greatly because of variation in the definition of "the soil" in different disciplines (Shangguan et al., 2017). Does the soil profile stretch all the way from the ground surface you stand on to the level of impermeable bedrock below? Is weathered rock included? What about an impermeable clay layer a few meters below the ground? If you are standing on leaf litter rather than bare soil, are those leaves part of the soil, too? Is a half-decomposed log nearby (coarse woody debris) also part of the soil? Here, we follow soil science norms and would answer "yes" to all five of these questions.

Most model simulations out of necessity make a practical division of the soil profile into two parts, the upper and lower soil profile:

- (1) *Upper soil profile – A and B horizons – Topsoil and subsoil* (Fig. 3): The upper part of the soil profile is considered to be the part from the ground surface (i.e. Level L3, Fig. 2) down to what is termed the soil simulation lower boundary (SLB). The SLB depth in LSMs is usually relatively shallow (<5 m), which means that at many global locations soil profiles in LSMs are simulated to be deeper than they are in reality (e.g. mountain regions) and in many other locations the direct simulation of soil involves only a small percentage of the material below Level L3. Many studies have shown that model simulations return improved soil flux gradients when finer soil layers, or a greater number of vertical soil layers, are used (e.g., Lekshmi and Arnepalli, 2017; Batelisi et al., 2020).
- (2) *Lower soil profile – C horizon* (Fig. 3): The lower part of the soil profile is considered to be the part from the SLB down to Level L1 (bedrock). In some models, the lower soil is simulated with heightened lateral in/outflow because of the application of a 3-D Richardson–Richards approach that would be too computationally intensive at shallow soil depths. Note, in most biomes the water table fluctuates continually between the upper and lower part of the soil profile, therefore if two different submodels handle the two domains the water table should be able to move realistically between the two within the simulation (Fan et al., 2007).

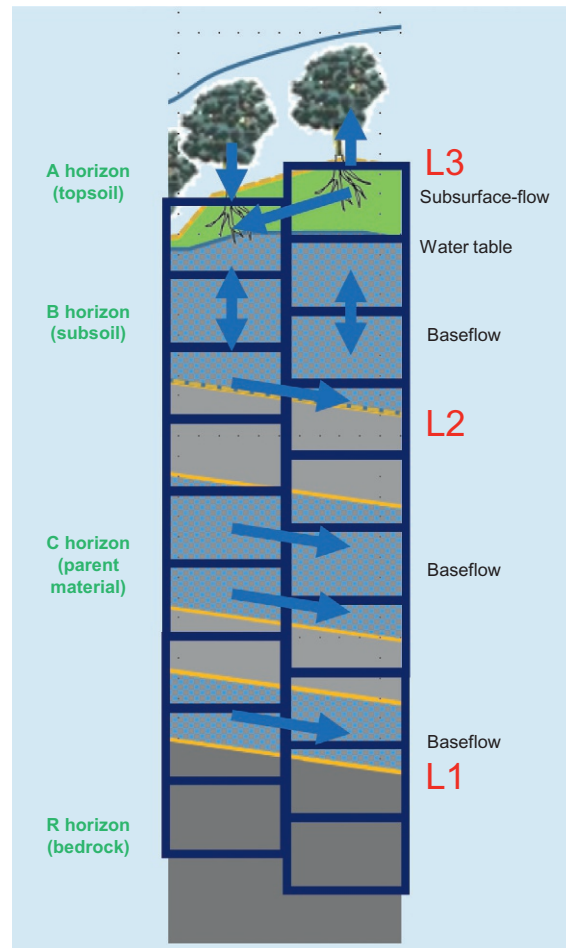


Fig. 3 An example 1-D soil profile in two neighboring simulated locations, where the soil simulation depth is at level L2 (Fig. 2). Note that if topographic flow directions at the surface are not parallel to flow directions at depth (as illustrated in this example location), then two sets of ancillary flow directions are required for the simulation of water flow in the landscape, with the former calculated from a DEM and the latter calculated from a DEM with soil depths subtracted (q.v. Pelletier et al., 2016; Shangguan et al., 2017). Although variation in soil surface elevation (microtopography) both spatially (i.e., within-gridcell, e.g. thermokarst landscapes and permafrost thaw, Aas et al. (2019)) and over time (e.g. peatland development, Baird et al., 2012) is important in many contexts, this is usually assumed not to affect subsurface flowpaths. The A, B, C and R horizons of a soil profile are closely related to the dynamics of water within the profile (Hillel, 2004), and in modeling frameworks the level L2 generally coincides with the lowest part of the B horizon (subsoil). Note that although some sources separate out an O horizon or humus layer at the upper end of the A horizon, here we include this in A. Similarly, some sources separate out an E Horizon or eluviation layer at the lower part of the A horizon, but we also include this in A. 'Regolith' is most widely used to refer to all unconsolidated rock (i.e. horizons A, B and C together) (although note Pelletier et al. (2016) used the term 'intact regolith' to refer to what we call here the soil C horizon).

Why is the SLB depth not simply fixed at Level L1 to capture the whole soil profile? One answer is historical: LSMs were developed originally as descriptions of the lower boundary layer of climate models, and a few meters of soil were considered sufficient for the most simple representation of the land surface. Second, global data on soil depths were lacking at the time, although this is no longer the case (Shangguan et al., 2017).

A third reason was, and remains, the challenge of simulating semi-permeable or impermeable layers at intermediate depths within the soil profile. The presence of such layers requires soil hydraulic properties to vary with depth in the soil profile simulation to allow the possibility of flow pathways at depth to differ from those at the surface (e.g. Fig. 3). Setting the depth SLB at a point above the surface of the first discontinuity (Level L2, Fig. 2) in theory allows the complications introduced by confining layers to be handled separately from surface soil simulation. Two difficulties arise with this approach, however:

- The effect on the atmosphere of not just the upper soil but also the lower soil profile is now known to be highly significant (Martinez-de la Torre and Miguez-Macho, 2019) and
- Many soils have semi-permeable or permeable layers (aquicludes) very close to the surface (e.g. lenses of ice or taliks in permafrost environments). The 'Australian Evergreen Paradox' provides a clear example of how this can have important implications: the marked dry season of the northern Australian monsoon tropics would lead a standard LSM to predict dominance by deciduous tree species, while in reality the presence of deep water aquifers within the reach of tree roots facilitates widespread dominance by evergreen species (Bowman and Prior, 2005).

Soil moisture model components (hydraulics and hydrology)

Soil moisture is a critically important variable both in the context of land management and scientific theory, and its status is determined by the interdependent processes of ‘soil hydrology’ (water movement within the soil) and ‘soil hydraulics’ (the mechanical behavior of water in the soil physical system), both at the surface and at depth. The soil at any particular location is characterized as a layered column (profile) where water can enter from above and can also enter or leave via a simulated lower boundary (SLB) (see Fig. 2), with profile moisture content determined by the water balance equation at the soil surface (Beven, 2012):

$$\begin{aligned} \text{(Total Water Input to Soil)} = & \text{(Precipitation)} + \text{(Irrigation input)} - \text{(Evaporation)} - \text{(Surface-flow)} \\ & + \text{(Below-surface lateral inflow}^*) - \text{(Below-surface lateral outflow}^*) \\ & + \text{(Capillary rise from below SLB)} - \text{(Drainage at level SLB of the soil)} \end{aligned}$$

*n.b. ‘Below-surface lateral in/outflow’ will correspond to (Subsurface-flow) if SLB is at the water table, or (Subsurface-flow) + (Baseflow) if SLB is at level L1 (Fig. 2).

The flow of water through individual soil layers and the vertical position of the water table are determined from these inputs by applying a similar water balance at each layer, modified to allow upward capillary flow through the application of a soil hydraulic model based on the Richardson–Richards Equation (Clark et al., 2015; Farthing and Ogden, 2017). Soil hydraulic models simulate the soil water characteristic (SWC), which is a relationship between soil matric suction ψ (Pa) and volumetric moisture content θ (cm^3/cm^3) parameterized for each soil in the simulation area (Table 1). These calculations result in an estimate of soil moisture status in each soil layer, water table depth and also surface and subsurface runoff. In organic soils (i.e. soils with a significant O horizon), especially, soil structure is important (Fatichi et al., 2020), therefore it is critical to apply these models carefully to capture reliably the fluxes close to the soil surface (e.g. when macropore or bypass flow is significant).

This water balance approach characterizes the state of the soil surface (including ponding) (Table 1). Runoff production models then produce estimates of surface and subsurface runoff, which may be used to predict the dynamics of flow velocity in local water courses and also any occurrence of inundation.

Current LSMs generally identify a reasonable soil SLB depth (above Level L2, Fig. 2, ideally close to the bottom of the soil’s B horizon, Fig. 3). Two different soil simulation models are applied for the upper and lower soil profiles (i.e. above and below the SLB). Upper soil profile simulators usually assume that lateral subsurface-flow and baseflow are zero so that a 1-D Richardson–Richards equation may be applied. Lower soil profile simulators, however, generally assume that the effects of vegetation (deep roots) are negligible and apply a Boussinesq equation (based on Darcy’s Law) that allows the 3-D dynamics of aquifers to be simulated.

Pedotransfer functions (PTFs) are used to source parameter estimates based on soil profile data (Weihermüller et al., 2021), e.g. from soil profile databases such as UNSODA or EU-HYDI, and are currently a focus of global theoretical development (see the current ISMC Working Group on Pedotransfer functions and land surface parameterization). Although the empirical nature of PTFs is theoretically undesirable and semi-physical PTFs do exist (Zhang et al., 2018), they remain the most practical simulation approach in the context of large-scale LSMs.

Table 1 Models concerned with water flow in or on the soil profile.

	Model
<i>(1) Water flow in the upper soil profile</i>	
Models of the soil water characteristic curve, based on assumptions about the soil pore-size distribution and soil tortuosity. Most LSMs follow either Brooks and Corey’s or van Genuchten’s models, combined with Mualem’s model of soil pore distributions (cf. Marthews et al., 2014; Van Looy et al., 2017).	Brooks & Corey–Mualem van Genuchten–Mualem
<i>(2) Water flow in the lower soil profile</i>	
Often called <i>groundwater models</i> (see e.g. Batelis et al. (2020), although note in this chapter we define groundwater in terms of water table position rather than in relation to a fixed soil depth (see Glossary)).	CABLE MODFLOW LEAFHYDRO ParFlow
<i>(3) Runoff production models</i>	
The water balance equation (see text) allows us to calculate runoff (Beven, 2012) under certain assumptions about the distribution of water within gridcells. Some runoff production approaches involve assuming soil moisture storage capacities follow a constant probability distribution (e.g. PDM, CABLE) or can be derived from the local topography (TOPMODEL).	PDM CABLE TOPMODEL

Usually, the calculations of water balance and water table position in the soil profile are divided into three submodels, although in the framework of Earth System Models (ESMs) they may be integrated more closely.

In LSMs, it is still unusual for both the upper and lower soil profiles to be simulated in detail, e.g. most crop models still assume a soil profile with a ‘freely draining’ lower boundary, which generally implies that the lower soil profile model has been reduced to a single drainage flux just to maintain water balance requirements. A few LSMs that simulate only upper soil profiles have been coupled with groundwater models (see Batelis et al., 2020), and more are attempting to improve the representation of lower soil profile dynamics in their simulations. This is because doing so can greatly improve predictions of water table depths and therefore surface soil wetness and rates of evapotranspiration (Martinez-de la Torre and Miguez-Macho, 2019). An implicit difficulty in using two separate models to simulate the upper and lower soil profile is to ensure that the water table can move realistically all the way from the surface (under wet conditions) down to the level of the first aquifer (under dry conditions), and for the water table surface within the soil (i.e. the belowground phreatic surface) to be realistic at the local scale (e.g. no discontinuities) (cf. peatland simulators that solve these difficulties, e.g., Baird et al., 2012).

Soil biogeochemical cycling models (soil nutrient and carbon dynamics)

Nutrient dynamics in soil (biogeochemistry) are critical to determining soil and plant function (health) both in terms of nutrients (e.g. nutrient addition from litter and loss via leaching) and physical structure (e.g. compacting or cracking). Nutrients enter the soil via rainfall, leaf litter and debris inputs from surface vegetation, and the erosion of minerals within the soil matrix itself. After a certain delay, depending on the processes involved, nutrients then leave the soil by soil respiration (notably carbon in the form of CO₂ (carbon dioxide) and CH₄ (methane) and soil erosion (e.g. during dust storms, landslide events or as leachate, dissolved or entrained in surface or subsurface runoff). These fluxes are simulated by ‘soil biogeochemistry models’, which include soil carbon dynamics and soil degradation models (Table 2).

Depending on the balance of input and output of plant nutrients, organic material may accumulate in the soil (e.g. peatland formation) or decrease (soil degradation), with a wide range of implications and impacts for both human and natural ecosystems that depend on those soils. Decomposition in soil is mediated by soil fauna and microbes and is also strongly affected by soil hydraulics; the position of the water table will dictate whether decomposition within the soil is aerobic or anaerobic. The dynamics of decomposition is currently a very active area of research, including whether the activities of soil microbes need to be explicit in the model, or can be approximated by decay kinetics (Table 2). Many recent projects have also focused on the role of soil moisture in the control of soil C dynamics (e.g. the Norwegian MOCABORS project).

Soil carbon (C) is the largest terrestrial C pool (IPCC, 2021). Many climate change solutions rely on the capability of soil to accumulate carbon and store it long-term or actively help to increase this capability (e.g., the FAO’s RECSOIL initiative). Knowing the residence time of carbon within the soil, and being able to predict how this may change with warming, is critical to estimating the effectiveness of these solutions. However, there is a limit to how much carbon can be stored in soil (SOC saturation; e.g. Campbell and Paustian (2015) estimate up to 9% SOC in mineral soils under agricultural use), therefore land management solutions intended to increase SOC storage have their limits, which must also be estimated. Models of carbon, water, nitrogen and even phosphorus dynamics are key to providing these estimates.

Deep soil

The importance of biogeochemical processes at depth in soil has been highlighted by many research projects, with attention drawn to our lack of both theory (Campbell and Paustian, 2015) and observations (Marthews et al., 2014) for the lower part of the soil profile. There has long been an assumption in soil science that at a certain depth in the soil biotic activity decreases to become negligible and below that the material is simply rock at varying stages of weathering (i.e. a C horizon, Fig. 3). Although soil organic matter (SOM) is highly concentrated at the surface because it derives mainly from decomposing surface litter (e.g. for carbon

Table 2 Overview of soil biogeochemistry models.

Model	K or M
<i>Top 5 models listed in Campbell and Paustian (2015):</i>	
CENTURY Soil Organic Matter Model	K
RothC Rothamsted Carbon model	K
DNDC (DeNitrification-DeComposition)	K
EPIC (Environmental Policy Integrated Climate)	K
DSSAT (Decision Support System for Agrotechnology Transfer)	K
<i>A selection of more recent models used with LSMs:</i>	
Yasso07	M
COMMISSION	M
MIMICS	M

Models are divided into those that simulate soil microbial decomposition via microbial and chemical kinetics (K) and those that attempt a more process-based representation of microbial dynamics (M) (cf., Chadburn et al., 2020).

accounting SOM it is assumed to be zero below 1-m depth (Rodeghiero et al., 2009), we cannot exclude the possibility of significant soil biogeochemical dynamics at depth. Soil sampling across the globe seldom extends below a few meters and information from microwave remote sensing is limited to the top few centimeters of the soil, therefore information on vertical distributions of SOM is currently quite limited.

Although tree roots below 10 m are rare in all biomes, microbial activity at depth can nevertheless be significant and does affect soil biogeochemical cycling (Richter and Markewitz, 1995; Campbell and Paustian, 2015; Lennon, 2020). Evidence is also accumulating that soil respiration measured at the soil surface does not represent total emission, as is commonly assumed (Rodeghiero et al., 2009), because an unknown quantity of CO₂ emitted from deeper soil layers is absorbed by water held in shallower soil layers before it reaches the surface (Richter and Markewitz, 1995; Gross and Harrison, 2019).

Tree roots may be rare below 10 m, but it does not mean that they are absent (Yang et al., 2016) and these taproots may be significant for vegetation persistence under drought. It should be expected that larger plant species will evolve root distributions that display a peak of coarse root density at the average annual water table depth (to absorb day-to-day water), a peak of fine root density close to the surface (to absorb nutrient input from surface litterfall) and a small number of taproots as deep as possible to be activated during drought or severe dry seasons. Global data on rooting depth are increasing (Yang et al., 2016), although it should be kept in mind that estimates of rooting depth aim to identify the level below which root activity becomes minimal. Therefore, they are arguably closer to an estimate of the depth of the base of the subsoil (B-C horizon boundary, Fig. 3) rather than a hard limit below which roots cannot occur (the activity of taproots may be episodic, e.g. in response to drought, and therefore challenging to capture from observations).

Soil heat balance and thermal dynamics models

The current ISMC Working Group on [Soil Thermal Properties](#) aims to evaluate the current use of empirical functions to estimate soil thermal properties (including heat conductivity and heat capacity). The flow of heat into and out of the soil determines soil temperature in each layer of the soil profile. Soil temperature is important because it influences the physical, chemical, and microbiological processes that take place in the soil and at the soil surface, including the types and rates of chemical reactions within the soil, and energy and mass exchange with the atmosphere (Hillel, 2004). Thermal conductivity is strongly influenced by hydrology, hence simulating soil moisture and groundwater (section “Soil moisture model components (hydraulics and hydrology)”) is an important prerequisite for modeling thermal dynamics (He et al., 2020). Models of soil thermal conductivity have a long history, but they are still under development (Table 3).

Recently, there has been progress in representing coupled heat and vapor flow at the soil surface, which is usually represented with relatively high uncertainty in LSMs (Lekshmi and Arnepalli, 2017). In addition, global warming will entail significant increases in soil temperatures and soil thermal dynamics. Currently, evidence is mixed on the dependence of land–atmosphere fluxes on soil temperature (e.g. soil respiration, Carey et al., 2016). Research is ongoing to identify these dependencies and linkages in order to estimate how much (or whether) rising soil temperatures will fuel significant increases in global temperatures (Carey et al., 2016).

Extreme environments

The standard approach to modeling of the soil environment outlined here (Fig. 1) should be considered a starting point only, and requires modification for more detailed work in particular biomes. Here, we focus on three example biomes where soil modeling encounters particular challenges. We consider these to be extreme environments in a global sense: for example, dryland forests cover only 7% of global land and would therefore be considered extreme by most of the global population, despite being the norm for the people for whom these areas are home.

Drylands and dry forests

There are many challenges to human life in global drylands that are caused by or aggravated by changes in the local soil environment. In these regions, soil erosion and loss can be overwhelming, and there is strong motivation for simulation models to be formulated that can address some of these challenges. In very dry soil, the liquid water content inside the soil matrix is no longer continuous so capillary forces become negligible and the strength of adsorptive forces controls water retention (i.e. soil moisture is dominantly hygroscopic). Standard soil hydraulic models arguably do not model these effects well, simply assuming

Table 3 Overview of soil thermal conductivity and thermal capacity models (refer to Rerak (2017)).

Model	Notes
Johansen model	Johansen (1977)
Johansen model with revised thermal conductivity	Chadburn et al. (2015) (recommended for organic soils)

that a certain amount θ_{res} of soil water is ‘residual’, in the sense of being unavailable to plants (Marthews et al., 2014). Improving the model representation of desorptive drying in very dry soil could greatly improve the reliability of LSMs when run for dryland environments.

Fire, soil compaction, desertification and large-scale erosion events are often considered infrequent events, but in drylands they are part of the normal seasonal cycle and therefore must be accounted for in simulations. Burning, for example, can greatly affect soil properties, depending on the extent, intensity and duration of the fire (Certini, 2005). Although heat in moist soil penetrates more deeply, the latent heat of vaporization prevents soil temperatures from exceeding 95 °C until water completely vaporizes. In completely dry soil, however, temperatures can typically rise to 200–700 °C depending on fuel availability. Below the surface, the temperature gradient is very steep: at 5 cm in the mineral soil temperatures rarely exceed 150 °C and often no heating occurs below 20–30 cm (Certini, 2005). The effects of fire on soil are numerous, however, including significant removal of organic matter, deterioration of both structure and porosity, loss of nutrients through volatilization, ash entrapment in smoke columns, leaching and erosion, and marked alteration of both quantity and specific composition of microbial and soil-dwelling invertebrate communities.

Tropical moist and wet forests

Moist and wet tropical forest environments present particular difficulties for soil models: their soils are usually not only macroaggregated, but also interlaced with a large root density and exist in areas of the globe where high precipitation over geological periods of time has produced unique soil types with different hydraulic and hydrological properties. In very wet soil, the soil is almost always saturated or close to saturation. Therefore, capillary forces become negligible compared to gravitational flow, and standard soil hydraulic models, which emphasize capillary effects, arguably do not model these effects well (Marthews et al., 2014).

Fire should also be mentioned in this context: although natural wildfires are generally restricted to drylands, prescribed burning can occur in any biome and is widely used in tropical forests as an aid to deforestation (e.g. the Amazonian Arc of Deforestation). In addition to the known effects on vegetation, fires greatly affect soil structure (see above) in ways that are not generally included in LSMs.

Cold regions

High latitude ecosystems have long been poorly represented in LSMs, broadly because of a lack of observational data (Chadburn et al., 2015). Boreal forests, tundra and other Arctic and Antarctic ecosystems have an important role in global budgets of CO₂ and CH₄ and influence much of the seasonal and annual climatology of the rest of the globe. The representation of cold soil processes within LSMs (e.g. snow, ice, frozen soil and permafrost dynamics) has recently come into sharp focus because the Arctic, in particular, is experiencing approximately twice the rate of warming as the rest of the globe (Arctic amplification, Richter-Menge et al., 2019).

In the boreal zone, the widespread thawing of permafrost is of great concern and may in the future lead to significant infrastructure damage in many regions. The [Global Terrestrial Network for Permafrost](#) (GTN-P) was developed in the 1990s to organize and manage a global network of permafrost observations for detecting, monitoring and predicting climate change. Many global models represent specific cold-region soil processes such as the phase change associated with freezing and thawing, and more recently have begun to include excess ice and thermokarst processes.

Peatlands and wetlands

Peatlands occur in all regions except the high polar regions: they represent some of the most productive ecosystems in the world and are of particular global focus today (e.g. the [International Peatland Congress 2021](#)). Peatlands, with their deep organic layers and large carbon content, tend to form in water-saturated soils where anoxic conditions suppress decomposition (Parish et al., 2008). Peat represents a globally-relevant and potentially unstable carbon stock, and often lies beneath some of the most biodiverse forest ecosystems on the planet. Progress has been made in representing peatland hydrological dynamics and carbon fluxes, but a reliable representation of the coupled interactions between carbon and hydrology that determines peatland stability and resilience has not yet been successfully formulated for large-scale LSMs (see discussion in Chadburn et al., 2020).

Wetland environments are characterized by seasonally or permanently inundated soils. In these areas, interaction of the flow regime and water table is crucial: if the flow regime is variable with frequent floods then fluvial input will be significant, but if the water table is very close to the surface then groundwater input will be significant. Wetlands are also key nodes in the global methane (CH₄) cycle; they are the largest natural source of methane that goes into the atmosphere (Chadburn et al., 2020). Methane emission is primarily a function of soil temperature and substrate C availability (Chadburn et al., 2020), but water flow is also important (with emission from flowing water often assumed to be negligible compared to that from stagnant water - cf. the division between depression storage and detention storage (see Glossary)).

Global peatlands demonstrate most clearly that soil is not just a passive substrate but a dynamic one that changes its volume and physical makeup over time, not only because of deposition or erosion of material by external processes (e.g. hydrological, eolian, including e.g. landslides) but also because of accumulation and loss through internal processes (Baird et al., 2012; Chadburn et al.,

2020). The need for LSMs to include a dynamic soil module that can simulate these processes is clear from the point of view of long-term peatland development.

New directions in land surface simulation

Hyperresolution

There are currently two leading edges of development in LSMs: First, the move to higher resolution (Beven et al., 2015), which is being driven predominantly by the movement in meteorology toward convection-permitting models that require sub-daily temporal resolution and spatial resolutions finer than 4–5 km (Kendon et al., 2021). At these spatial resolutions, many aspects of the land surface that were approximated as ‘subgrid’ processes previously must be given a more process-based treatment (Beven et al., 2015; Fisher and Koven, 2020).

Coupled model environments

A second leading edge is the move toward coupled model simulations, where there is two-way data exchange between simulation elements (e.g. land surface, ocean, atmosphere). Before the 2000s, climate models and LSMs were always separate models, but since then there has been a broad global effort to integrate simulated systems more closely and produce fully-coupled models of the whole global environment, i.e. Earth system models (ESMs) (Clark et al., 2015). The ESMs generally include a more integrated approach to the modeling of soil processes (Campbell and Paustian, 2015; Fisher and Koven, 2020). A great advantage of ESMs is that they allow the known influence of the soil on the atmosphere to be characterized much more reliably than either climate models or LSMs in isolation (e.g. the dynamics of the tropical rainbelt). Efforts to improve coupled modeling approaches are being coordinated through the WCRP’s Coupled Model Intercomparison Project Phase 6 (CMIP6), a worldwide modeling effort initiated in 2016 that is fostering improvements in climate models, LSMs and ESMs (IPCC, 2021).

Conclusions

Soil covers almost the entire land surface of the Earth: only at the highest reaches of mountain ranges or deep below permanent glaciers and ice sheets does the effect of soil on surface life become negligible. Soil influences or controls an enormous variety of aspects of the functioning of the land surface, ranging from the yield of crops in agricultural fields, to the dynamics of vegetation in natural ecosystems, to the velocity of floodwaters in river catchments. The shape of the soil surface around us defines many aspects of all of our lives.

Land Surface Models (LSMs) have been through several generations of development since they were first formulated in the 1960s, but have arguably not yet fulfilled their potential to solve key societal and scientific questions related to environmental resilience (Fisher and Koven, 2020). The representation of soil in LSMs and in elements of other climate models is undergoing rapid development, and the relevance and importance of soil to current efforts to understand and mitigate climate change is becoming increasingly recognized.

The United Nations Secretary General, António Guterres, said in his State of the Planet speech in December 2020: “Covid and climate have brought us to a threshold” and that “Making peace with nature is the defining task of the 21st century.” Although the effects of soil processes are only one part of climate change, they are a critically important part and it is therefore timely to identify the most important issues we face in each of the major global biomes.

A balance must be struck between model realism (which often requires complexity) and tractability (runtimes must not be impracticably long), the need for an improved representation of soil hydrological, biogeochemical and other processes is becoming clear. Looking ahead to the future, we anticipate a new generation of soil models with greater consideration in their make-up of the formation, development, organic and physical influences on the biosphere, and long-term accumulation, erosion and loss of soil profiles. We believe that to address the challenges of climate change, the multiple effects of soil processes will also need to be reflected in the context of all good climate and land surface models.

References

- Aas KS, Martin L, Nitzbon J, Langer M, Boike J, Lee H, Berntsen TK, and Westermann S (2019) Thaw processes in ice-rich permafrost landscapes represented with laterally coupled tiles in a land surface model. *The Cryosphere* 13: 591–609.
- Baird AJ, Morris PJ, and Belyea LR (2012) The DigiBog peatland development model 1: rationale, conceptual model, and hydrological basis. *Ecohydrology* 5: 242–255.
- Batellis SC, Rahman M, Kollet S, Woods R, and Rosolem R (2020) Towards the representation of groundwater in the Joint UK Land Environment Simulator. *Hydrological Processes* 34: 2843–2863.
- Beven K (2012) *Rainfall-Runoff Modelling: The Primer*, 2nd edn Chichester, UK: Wiley-Blackwell.
- Beven K, Cloke H, Pappenberger F, Lamb R, and Hunter N (2015) Hyperresolution information and hyperresolution ignorance in modelling the hydrology of the land surface. *Science China Earth Sciences* 58: 25–35.
- Bowman DMJS and Prior LD (2005) Why do evergreen trees dominate the Australian seasonal tropics? *Australian Journal of Botany* 53: 379–399.

- Campbell EE and Paustian K (2015) Current developments in soil organic matter modeling and the expansion of model applications: A review. *Environmental Research Letters* 10.
- Carey JC, Tang JW, Templer PH, Kroeger KD, Crowther TW, Burton AJ, Dukes JS, Emmett B, Frey SD, Heskell MA, Jiang L, Machmuller MB, Mohan J, Panetta AM, Reich PB, Reinsch S, Wang X, Allison SD, Bamminger C, Bridgham S, Collins SL, De Dato G, Eddy WC, Enquist BJ, Estiarte M, Harte J, Henderson A, Johnson BR, Larsen KS, Luo Y, Marhan S, Melillo JM, Peuelas J, Pfeifer-Meister L, Poll C, Rastetter E, Reinmann AB, Reynolds LL, Schmidt IK, Shaver GR, Strong AL, Suseela V, and Tietema A (2016) Temperature response of soil respiration largely unaltered with experimental warming. *Proceedings of the National Academy of Sciences of the United States of America* 113: 13797–13802.
- Certini G (2005) Effects of fire on properties of forest soils: A review. *Oecologia* 143: 1–10.
- Chadburn S, Burke E, Essery R, Boike J, Langer M, Heikenfeld M, Cox P, and Friedlingstein P (2015) An improved representation of physical permafrost dynamics in the JULES land-surface model. *Geoscientific Model Development* 8: 1493–1508.
- Chadburn SE, Aalto T, Aurela M, Baldocchi D, Biasi C, Boike J, Burke EJ, Comyn-Platt E, Dolman AJ, Duran-Rojas C, Fan YC, Friborg T, Gao Y, Gedney N, Gockede M, Hayman GD, Holl D, Hugelius G, Kutzbach L, Lee H, Lohila A, Parmentier FJW, Sachs T, Shurpali NJ, and Westermann S (2020) Modeled microbial dynamics explain the apparent temperature sensitivity of wetland methane emissions. *Global Biogeochemical Cycles* 34.
- Clark MP, Fan Y, Lawrence DM, Adam JC, Bolster D, Gochis DJ, Hooper RP, Kumar M, Leung LR, Mackay DS, Maxwell RM, Shen CP, Swenson SC, and Zeng XB (2015) Improving the representation of hydrologic processes in Earth System Models. *Water Resources Research* 51: 5929–5956.
- Fan Y, Miguez-Macho G, Weaver CP, Walko R, and Robock A (2007) Incorporating water table dynamics in climate modeling: 1. Water table observations and equilibrium water table simulations. *Journal of Geophysical Research-Atmospheres* 112.
- Farthing MW and Ogden FL (2017) Numerical solution of Richards' equation: A review of advances and challenges. *Soil Science Society of America Journal* 81: 1257–1269.
- Fatchi S, Or D, Walko R, Vereecken H, Young MH, Ghezzehei TA, Hengl T, Kollet S, Agam N, and Avissar R (2020) Soil structure is an important omission in Earth System Models. *Nature Communications* 11.
- Fisher RA and Koven CD (2020) Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *Journal of Advances in Modeling Earth Systems* 12.
- Gross CD and Harrison RB (2019) The case for digging deeper: Soil organic carbon storage, dynamics, and controls in our changing world. *Soil Systems* 3.
- He HL, Dyck M, and Lv JL (2020) A new model for predicting soil thermal conductivity from matric potential. *Journal of Hydrology* 589.
- Hillel D (2004) *Introduction to Environmental Soil Physics*. Amsterdam, Netherlands: Elsevier.
- IPCC (2021) *Climate Change 2021: The Physical Science Basis*. U.K.: Cambridge University Press.
- Johansen O (1977) *Thermal Conductivity of Soils (Varmeledningsevne av jordarter)*. Trondheim: Norwegian University of Science and Technology.
- Kendon EJ, Prein AF, Senior CA, and Stirling A (2021) Challenges and outlook for convection-permitting climate modelling. *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences* 379.
- Lekshmi KRA and Arnepalli DN (2017) A review on coupled heat and water vapour transport in unsaturated soils. *Geotechnical Frontiers 2017: Geotechnical Materials, Modeling, and Testing* 746–755.
- Lennon JT (2020) Microbial Life Deep Underfoot. *MBio* 11.
- Marthews TR, Quesada CA, Galbraith DR, Malhi Y, Mullins CE, Hodnett MG, and Dharssi I (2014) High-resolution hydraulic parameter maps for surface soils in tropical South America. *Geoscientific Model Development* 7: 711–723.
- Martinez-de la Torre A and Miguez-Macho G (2019) Groundwater influence on soil moisture memory and land-atmosphere fluxes in the Iberian Peninsula. *Hydrology and Earth System Sciences* 23: 4909–4932.
- Parish F, Sirin A, Charman D, Joosten H, Minayeva T, Silvius M, and Stringer L (eds.) (2008) *Assessment on Peatlands, Biodiversity and Climate Change: Main Report*. Wageningen: Global Environment Centre, Kuala Lumpur and Wetlands International.
- Pelletier JD, Broxton PD, Hazenberg P, Zeng X, Troch PA, Niu G, Williams ZC, Brunke MA, and Gochis D (2016) *Global 1-km Gridded Thickness of Soil, Regolith, and Sedimentary Deposit Layers*. Oak Ridge, Tennessee, USA: ORNL DAAC.
- Pribyl DW (2010) A critical review of the conventional SOC to SOM conversion factor. *Geoderma* 156: 75–83.
- Raats PAC and Knight JH (2018) The contributions of lewis fry Richardson to drainage theory, soil physics, and the soil-plant-atmosphere continuum. *Frontiers in Environmental Science* 6.
- Rerak M (2017) Selected soil thermal conductivity models. 4th Scientific and Technical Conference on Modern Technologies and Energy Systems (Wtue 2016) vol. 13.
- Richter DD and Markewitz D (1995) How deep is soil - soil, the zone of the earth's crust that is biologically-active, is much deeper than has been thought by many ecologists. *BioScience* 45: 600–609.
- Richter-Menge J, Druckenmiller ML, and Jeffries M (2019) *Arctic Report Card 2019*. NOAA.
- Rodeghiero M, Heinemeyer A, Schrumpp M, and Bellamy P (2009) Determination of soil carbon stocks and changes. In: Kutsch WL, Bahn M, and Heinemeyer A (eds.) *Soil Carbon Dynamics: An Integrated Methodology*, pp. 49–75. Cambridge, UK: CUP.
- Shangguan W, Hengl T, de Jesus JM, Yuan H, and Dai YJ (2017) Mapping the global depth to bedrock for land surface modeling. *Journal of Advances in Modeling Earth Systems* 9: 65–88.
- Van Looy K, Bouma J, Herbst M, Koestel J, Minasny B, Mishra U, Montzka C, Nemes A, Pachepsky YA, Padarian J, Schaap MG, Toth B, Verhoef A, Vanderborght J, van der Ploeg MJ, Weiermüller L, Zacharias S, Zhang YG, and Vereecken H (2017) Pedotransfer functions in earth system science: Challenges and perspectives. *Reviews of Geophysics* 55: 1199–1256.
- Weiermüller L, Lehmann P, Herbst M, Rahmati M, Verhoef A, Or D, Jacques D, and Vereecken H (2021) Choice of pedotransfer functions matters when simulating soil water balance fluxes. *Journal of Advances in Modeling Earth Systems* 13.
- Yang YT, Donohue RJ, and McVicar TR (2016) Global estimation of effective plant rooting depth: Implications for hydrological modeling. *Water Resources Research* 52: 8260–8276.
- Zhang YG, Schaap MG, and Zha YY (2018) A high-resolution global map of soil hydraulic properties produced by a hierarchical parameterization of a physically based water retention model. *Water Resources Research* 54: 9774–9790.

Encyclopedia of SOILS IN THE ENVIRONMENT

Volume Four: Pedology & Pedometric

EDITORS-IN-CHIEF

Michael J Goss

Professor Emeritus, School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G 2W1, CANADA

Margaret Oliver

Visiting professor, Soil Research Centre, Department of Geography and Environmental Science, Russell Building, University of Reading, Whiteknights, Reading, RG6 6DR, United Kingdom.

SECTION EDITORS

Dr Evgenia Blagodatskaya

Dept. of Soil Ecology, Helmholtz-Zentrum für Umweltforschung GmbH – UFZ, Theodor-Lieser Str. 4, D-06120 Halle (Saale), GERMANY

Professor Adrian Unc

Boreal Ecosystems and Agricultural Sciences, Grenfell Campus, Memorial University of Newfoundland, 20 University Drive, Corner Brook, Newfoundland A2H 5G4, CANADA

Dr. Siobhán Staunton

INRA, UMR Eco&Sols, Montpellier SupAgro, bâtiment 12, 2 Place Viala, 34060 Montpellier Cedex 2, Montpellier, FRANCE

Professor Baoshan Xing

University of Massachusetts – Amherst, College of Natural Sciences, Stockbridge School of Agriculture, Environmental & Soil Chemistry, Program, 410 Paige Laboratory, Amherst, MA 01003, USA

Dr Karin Mueller

New Zealand Institute for Plant & Food Research Ltd., Ruakura Research Centre, Private Bag 3123, Hamilton, New Zealand

Professor Christina Siebe

Instituto de Geología, Circuito de la Investigación Científica s/n, Ciudad Universitaria, Ciudad de México, CP 04510, México

Professor Rupert Baeumler

Institute of Geography, Friedrich-Alexander-University of Erlangen-Nuremberg, Wetterkreuz 15, 91058 Erlangen, Germany

Dr Uta Stockmann

Soil Processes and Function Organisational Unit, CSIRO Agriculture and Food, Black Mountain, GPO BOX 1700, Canberra ACT 2601, AUSTRALIA

Professor Paul Hallett

Room 2.23, Cruickshank Building, School of Biological Sciences, University of Aberdeen, Aberdeen, AB24 3UU, UNITED KINGDOM

Professor Dani Or

Department of Environmental Systems Science (D-USYS), Institute of Biogeochemistry and Pollutant Dynamics (IBP), Soil and Terrestrial Environmental Physics (STEP), CHN F 29.1 Universitaetstrasse 16, 8092 Zurich, SWITZERLAND



Set ISBN



Volume ISBN



Second Edition

Encyclopedia of SOILS IN THE ENVIRONMENT

Volume Five: Soil Physics



BIODIVERSITY,
ANTIBIOTICS, FOOD SECURITY
HUMAN AND ANIMAL HEALTH
PATHOGENS,
SOCIAL COHESION

Editors-in-Chief

Michael Goss and Margaret Oliver

Section Editors

Paul Hallett & Dani Or

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

SECOND EDITION

ENCYCLOPEDIA OF SOILS IN THE ENVIRONMENT

SECOND EDITION

EDITORS-IN-CHIEF

Professor Michael J Goss

*Professor Emeritus, School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G
2W1, CANADA*

Professor Margaret Oliver

*Visiting Professor, Soil Research Centre, Department of Geography and Environmental Science,
Russell Building, University of Reading, Whiteknights, Reading, RG6 6DR, United Kingdom*

VOLUME 5

SOIL PHYSICS

SECTION EDITORS

Paul Hallett

*Room 2.23, Cruickshank Building, School of Biological Sciences, University of Aberdeen, Aberdeen,
AB24 3UU, UNITED KINGDOM*

Dani Or

*Department of Environmental Systems Science (D-USYS), Institute of Biogeochemistry and Pollutant
Dynamics (IBP), Soil and Terrestrial Environmental Physics (STEP), CHN F 29.1 Universitaetstrasse
16, 8092 Zurich, SWITZERLAND*



ELSEVIER

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge MA 02139, United States

Copyright © 2023 Elsevier Ltd. All rights reserved

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

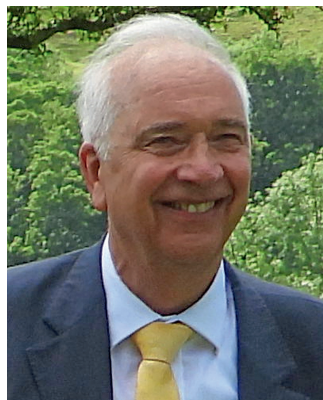
ISBN 978-0-12-822974-3

For information on all publications
visit our website at <http://store.elsevier.com>



Publisher: Oliver Walter
Acquisition Editor: Priscilla Braglia
Content Project Manager: Sona Chahal
Associate Content Project Manager: Nandhini Mahendran
Designer: Christian Bilbow

EDITORS BIOGRAPHY



Michael J Goss, PhD, MSc, BSc, is a professor emeritus in the School of Environmental Sciences at the University of Guelph. Before taking on the joint editor-in-chief responsibilities with Margaret Oliver, he held a similar position for *Soil Use and Management*, a journal of the British Society of Soil Science of which he is an honorary fellow. He was an inaugural member of the European Society for Agronomy. Michael started his career at the Letcombe Laboratory before moving to Rothamsted Experimental Station and the Macaulay Land Use Research Institute (now the James Hutton Research Institute). In 1990 he was appointed to the Research Chair in Land Stewardship at the University of Guelph and in 2005 became the director of its Regional Campus near Ottawa.

Michael's research has covered soil plant relations, soil management, and agriculture's footprint on the environment. His focus has been on the physical properties of soil and their impacts on root growth, which he showed to be influenced by the role of the root cap. He broadened his approach to include the effects of tillage and land drainage on root exploration. This work stimulated his interest in the transport of materials, such as plant nutrients, pathogenic microbes, antibiotics, and biologically active compounds through the soil and their contamination of water resources. Michael's knowledge in this field resulted in his involvement in the public inquiry into the death of seven people, including children, from a microbially contaminated drinking water supply. He contributed, both to the investigation into how the contamination occurred and if any fault could be attributed to local agricultural practices, and to how to prevent future events.

Michael also developed an interest in the interrelationships between soil microbes and root systems, particularly the tripartite symbiosis between roots of legumes, mycorrhizal fungi, and rhizobia. He has been an energetic member of collaborative research programs in France and Germany, and in Portugal, where he continues his interest in the role of mycorrhiza in sustainable land management.

Michael has authored more than 180 publications including two books, one in its 2nd edition, and edited two others.



Margaret Ann Oliver (BSc, PhD)

Margaret's interest in soil began in her penultimate year at school when learning about different types of soil and their formation. She chose the University of Bristol, United Kingdom for her first degree because the geography course offered soil science and she was also able to do a degree in geology alongside. Her love for both subjects remains undiminished. She did her PhD in soil spatial analysis at the University of Birmingham. Her early research focused on the multivariate and geostatistical analysis of soil data, which she continued to use in various research projects. Her research interests expanded to include the application of a wide range of numerical and geostatistical methods to soil and other data including pollen counts, forestry, radon emission, remotely sensed imagery, precision agriculture, and the incidence of childhood cancer. Her specific interests have been in sample design for spatial analysis and eventual mapping, risk analysis associated with prescribed thresholds in relation to soil pollutants or deficiencies of crop nutrients, and the relations between different scales of spatial variation. Her most

recent research was in the field of precision agriculture. At both the Universities of Birmingham and Reading, she taught topics in soil science, statistical methods, and spatial analysis. At Reading she had a research group working on various projects such as imagery, sampling and mapping, precision agriculture, and human health.

She has published more than 100 academic papers and has made several contributions to books. She is the co-author of three books: *Statistical Methods in Soil and Land Resource Survey*, Oxford University Press (Webster, R and Oliver, M. A. 1990); *Geostatistics for Environmental Scientists*, Second Edition, John Wiley & Sons (Webster, R and Oliver, M. A. 2007), and *How to do the Basic Steps in Geostatistics: The Variogram and Kriging*, Springer, Dordrecht (Oliver, M.A. & Webster, R.2015). She has edited two books: *Geostatistical Applications for Precision Agriculture* (Springer, Dordrecht, 2010) and *Precision Agriculture for Sustainability and Environmental Protection* (Earthscan from Routledge, 2013).

Margaret retired officially from Reading in 2004, but as yet has not managed to escape from academia. She has retained a link and a base at the University as a visiting professor since January 2005. After retirement, she developed and taught geostatistics courses to outside groups for several years. She then started on the long path of editorial work, as an associate editor of the *European Journal of Soil Science* and eventually the editor-in-chief, assistant editor of *Mathematical Geosciences*, co-editor of *Precision Agriculture* with John Stafford, and finally the editor-in-chief of the *Encyclopedia of Soils in the Environment* with Michael Goss.

Margaret was made an honorary member of the British Society of Soil Science in 2021 to reflect her commitment to soil science through fellow research, teaching, and editorial work. She was honored by the Pedometrics Commission of the International Union of Soil Science by having an award named after her in 2015. It is the Margaret Oliver award for young pedometricians.

EDITORIAL BOARD

Editors in Chief

Michael J Goss

Professor Emeritus, School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G 2W1, CANADA

Margaret Oliver

Visiting Professor, Soil Research Centre, Department of Geography and Environmental Science, Russell Building, University of Reading, Whiteknights, Reading, RG6 6DR, United Kingdom

Section Editors

Dr Evgenia Blagodatskaya

Dept. of Soil Ecology, Helmholtz-Zentrum für Umweltforschung GmbH – UFZ, Theodor-Lieser Str. 4, D-06120 Halle (Saale), GERMANY

Professor Adrian Unc

Boreal Ecosystems and Agricultural Sciences, Grenfell Campus, Memorial University of Newfoundland, 20 University Drive, Corner Brook, Newfoundland A2H 5G4, CANADA

Dr. Siobhán Staunton

INRA, UMR Eco&Sols, Montpellier SupAgro, bâtiment 12, 2 Place Viala, 34060 Montpellier Cedex 2, Montpellier, FRANCE

Professor Baoshan Xing

University of Massachusetts – Amherst, College of Natural Sciences, Stockbridge School of Agriculture, Environmental & Soil Chemistry, Program, 410 Paige Laboratory, Amherst, MA 01003, USA

Dr Karin Mueller

New Zealand Institute for Plant & Food Research Ltd., Ruakura Research Centre, Private Bag 3123, Hamilton, New Zealand

Professor Christina Siebe

Instituto de Geología, Circuito de la Investigación Científica s/n, Ciudad Universitaria, Ciudad de México, CP 04510, México

Professor Rupert Baeumler

Institute of Geography, Friedrich-Alexander-University of Erlangen-Nuremberg, Wetterkreuz 15, 91058 Erlangen, Germany

Dr Uta Stockmann

Soil Processes and Function Organisational Unit, CSIRO Agriculture and Food, Black Mountain, GPO BOX 1700 , Canberra ACT 2601, AUSTRALIA

Professor Paul Hallett

*Room 2.23, Cruickshank Building, School of Biological Sciences, University of Aberdeen, Aberdeen, AB24 3UU,
UNITED KINGDOM*

Professor Dani Or

*Department of Environmental Systems Science (D-USYS), Institute of Biogeochemistry and Pollutant Dynamics (IBP),
Soil and Terrestrial Environmental Physics (STEP), CHN F 29.1 Universitaetstrasse 16, 8092 Zurich, SWITZERLAND*

FOREWORD

Soil is the porous “skin” on the surface of the Earth that has formed over thousands of years of weathering, comprising minerals, organisms (both dead and alive), water, and air. Soil is more than just the sum of its parts. It serves several crucial ecological functions, from food and fiber production to water storage and purification to waste decomposition.

Increasingly we understand that everything is connected, and that soil is a “Critical Zone” that intersects with the hydrosphere (water), lithosphere (minerals), atmosphere (air), and biosphere (all living organisms). To quote Daniel Hillel, the 2012 World Food Prize laureate and editor of the First Edition of *The Encyclopedia of Soils in the Environment*, “[Soil] is the crucible of terrestrial life, within which biological productivity is generated and sustained...Our civilization depends on the soil...” He also promoted the importance of understanding soil to manage it effectively. This revised and reorganized second edition of the Encyclopedia aims to reflect the changing status of soil in the world. Within these five volumes, over 724 experts from many countries have provided updated, easy-to-understand chapters on soil, its properties, and its role in the environment to affect climate change, plant productivity, human health, and biodiversity.

Many technological advances have been made since the first edition of this encyclopedia that has allowed us to investigate soil physical, chemical, and especially biological properties more precisely, enabling us to identify and understand their roles in ecosystem services. The study of soils, especially in the environmental context, is critical to meeting the challenges of feeding, clothing, and cleaning up the world. The information presented in these volumes will be useful to many, including soil and environmental scientists, policymakers, land managers, educators, students, and anyone interested in learning more about soil.

Soil is under considerable threat from many sources—building and infrastructure, climate change, population pressure, and pollution, yet it is vital to agronomic productivity and continues to provide critical ecosystem services. Society needs to be engaged in how soil is managed and conserved. Most natural scientists are very dedicated and generous, including all of the authors in this Encyclopedia who have freely given their time to make these volumes so comprehensive. The study of soil, while often challenging, is fascinating—soil is a beautiful earthy material that can delight our senses with its variegated colors, gritty to smooth textures, and especially its characteristic aroma when freshly turned. Perhaps that is why gardeners and even some farmers tend to overwork their soil to its detriment—it *just feels good to get your hands in the dirt!*

Soil science is one of the most exciting, interesting, and rewarding disciplines, largely because it is multi-dimensional and interdisciplinary, basic and applied. It affects all of our lives. Thus, I encourage you to peruse these volumes and discover all kinds of interesting aspects of soil in the environment.

April Ulery

Former President, Soil Science Society of America
Professor, Soil & Environmental Sciences
New Mexico State University
Las Cruces, NM
The United States

PREFACE

This second edition of the *Encyclopedia of Soils in the Environment* follows in the footsteps of a highly rated first edition overseen by Professor Daniel Hillel as editor in chief, ably assisted by six editors. That edition provided a very valuable resource on the “state-of-the-art” in soil science in the context of the environment in the closing years of the twentieth century and the first three years of the current century. It set a standard that the editorial team for the second edition has aimed to match. The soil and environmental sciences have made huge strides in the almost 20 years since the first edition was published, and the new team has had to embrace these developments while retaining the basic components of these sciences. The project has taken 4 years to complete. It began with the appointment of the two editors-in-chief, professors Michael J. Goss and Margaret A. Oliver, followed by that of the nine subject editors and a pedology adviser.

The project was in its infancy (a few months only) when COVID-19 struck. The editors-in-chief discussed the format and structure of the second edition around the time that the first “lockdowns” around the world were being introduced. The first edition was alphabetically organized, whereas we decided to have a section-oriented approach in this edition, which we hope gives a more structured feel for the various topics embraced. The principal sections are soil biology, soil chemistry, the soil environment and its management, pedology (the origins and development of soil), pedometrics (the application of mathematics and statistics to soil), and soil physics. We were able to give our full attention to this development and to the selection of the scientists to join us on the road ahead. Furthermore, we were fortunate to choose a team that has maintained its enthusiasm despite the many tribulations we have faced along this path. The team comprises Jane Blagodatskaya and Adrian Unc (biology and microbiology), Siobhan Staunton and Baoshan Xing (chemistry), Karin Müller and Christina Siebe (environment and management), Rupert Bäumler (pedology), Uta Stockmann and Margaret Oliver (pedometrics), and Paul Hallett and Dani Or (physics). The editorial board has been mindful in their choice of subjects of the need for basic soil science and for state-of-the-art developments, with the aim that the audience for these volumes can be as large as possible.

During the gestation period of the Encyclopedia, methods of working for those in academia were turned on their heads. The change from “face-to-face” teaching to formulating and establishing online courses, initially for lockdown periods but then for much of the next 2 years, had huge consequences for academic staff worldwide. This meant that they had less time to devote to outside activities, such as writing chapters for an encyclopedia. As a result, some authors were unable to complete their chapters in time, leaving a few unfortunate gaps. The whole team has done as much as they can to limit any adverse effects, but we ask readers to appreciate the struggle that authors have had with illness (personal, within their families and work colleagues) and a great increase in their workload. Despite this, we are presenting a comprehensive document that should stand the test of the next 20 years of soils in the environment.

This century has seen an almost unprecedented interest in the soil and environmental sciences. It has been triggered not only by issues that the Earth is facing from global climate change, increasing population pressure, soil and land degradation, and food insecurity but also by the many new and sophisticated methods available, such as the DNA determination of microbial populations and their functional capabilities, proximal sensing, machine learning methods, and an ever-increasing array of technological tools. These developments have greatly increased our knowledge and appreciation of the complexity of soils, which has changed our understanding of some well-established aspects of the science. The soil and its environment are at the critical interface of the atmosphere, biosphere, hydrosphere, and lithosphere. They deserve to be taken seriously by the international community in the same way that air and water have been. Governments are now paying attention

to the soil and environment, and a legal and governance framework is emerging to look after these parts of our life support resources.

We have aimed, through our editing of chapters, to make the language of the chapters accessible to all, regardless of educational background. This does not mean that we have diminished the scientific content, but that readers should be able to gain the foothold they desire in these subjects to progress to greater knowledge and understanding of them. We have retained the overview of soil written by Daniel Hillel for the First Edition, first to honor his great contribution to soil science and second to demonstrate the significance of soil. It follows this Preface.

*Michael J. Goss
Margaret A. Oliver*

SOIL

The term “soil” refers to the weathered and fragmented outer layer of our planet’s land surfaces. Formed initially through the physical disintegration and chemical alteration of rocks and minerals by physical and biogeochemical processes, soil is influenced by the activity and accumulated residues of a myriad of diverse forms of life. As it occurs in different geologic and climatic domains, soil is an exceedingly varied body with a wide range of attributes.

Considering the height of the atmosphere, the thickness of the earth’s rock mantle, and the depth of the ocean, one observes that soil is an amazingly thin body—typically not much more than one meter thick and often less than that. Yet it is the fount of terrestrial life, within which biological productivity is generated and sustained. It acts like a composite living entity, a home to a community of innumerable microscopic and macroscopic plants and animals. A mere fistful of soil typically contains billions of microorganisms, which perform vital interactive biochemical functions. Another intrinsic attribute of the soil is its sponge-like porosity and its enormous internal surface area. That same fistful of soil may actually consist of several hectares of active surface, upon which physicochemical processes take place continuously.

Realizing humanity’s utter dependence on the soil, ancient peoples, who lived in greater closeness to nature than many of us today, actually revered the soil. It was not only their source of livelihood but also the material from which they built their homes and that they learned to shape, heat, and fuse into household vessels and writing tablets (ceramics, made of clayey soil, being the first synthetic material in the history of technology). In the Bible, the name assigned to the first human was Adam, derived from “adama,” meaning soil. The name given to the first earthling’s mate was Hava (Eve, in transliteration), meaning “living” or “life-giving.” Together, therefore, Adam and Eve signified quite literally “Soil and Life.”

The same powerful metaphor is echoed in the Latin name for human species—Homo, derived from humus, the material of the soil. Hence, the adjective “human” also implies “of the soil.” Other ancient cultures evoked equally powerful associations. To the Greeks, the earth was a manifestation of Gaea, the maternal goddess who, impregnated by Uranus (god of the sky), gave birth to all the gods of the Greek pantheon.

Our civilization depends on the soil more crucially than ever because our numbers have grown while available soil resources have diminished and deteriorated. Paradoxically, however, even as our dependence on the soil has increased, most of us have become physically and emotionally detached from it. Many of the people in the so-called developed countries spend their lives in the artificial environment of a city, insulated from direct exposure to nature, and some children may now assume as a matter of course that food originates in supermarkets.

Detachment has bred ignorance, and out of ignorance has come the delusion that our civilization has risen above nature and has set itself free of its constraints. Agriculture and food security, erosion and salination, degradation of natural ecosystems, depletion and pollution of surface waters and aquifers, and decimation of biodiversity—all these processes, which involve the soil directly or indirectly, have become abstractions to many people. The very language we use betrays disdain for that common material underfoot, often referred to as “dirt.” Some fastidious parents prohibit their children from playing in the mud and hurry to wash their “soiled” hands when the children nonetheless obey an innate instinct to do so. Thus, soil is devalued and treated as unclean though it is the terrestrial realm’s principal medium of purification, wherein wastes are decomposed and nature’s productivity is continually rejuvenated.

Scientists who observe soil closely see it in effect as a seething foundry in which matter and energy are in constant flux. Radiant energy from the sun streams onto the field and cascades through the soil and the plants growing in it. Heat is exchanged, water percolates through the soil’s intricate passages, and plant roots extract

water and transmit it to their leaves, which transpire it back to the atmosphere. Leaves absorb carbon dioxide from the air and synthesize it with soil-derived water to form the primary compounds of life. Oxygen emitted by the leaves makes the air breathable for animals, which consume and in turn fertilize plants.

Soil is thus a self-regulating bio-physio-chemical factory, processing its own materials, water, and solar energy. It also determines the fate of rainfall and snowfall reaching the ground surface—whether the water thus received will flow over the land as runoff, or seep downward to the subterranean reservoir called groundwater, which in turn maintains the steady flow of springs and streams. With its finite capacity to absorb and store moisture, and to release it gradually, the soil regulates all these phenomena. Without the soil as a buffer, rain falling over the continents would run off entirely, producing violent floods rather than sustained river flow.

Soil naturally acts as a living filter, in which pathogens and toxins that might otherwise accumulate to foul the terrestrial environment are rendered harmless. Since time immemorial, humans and other animals have been dying of all manner of diseases and have then been buried in the soil, yet no major human disease is transmitted by it. The term antibiotic was coined by soil microbiologists who, as a consequence of their studies of soil bacteria, specifically actinomycetes, discovered streptomycin (an important cure for tuberculosis and other infections). Ion exchange, a useful process of water purification, also was discovered by soil scientists studying the passage of solutes through beds of clay.

However unique in form and function, soil is not an isolated body. It is, rather, a central link in the larger chain of interconnected domains and processes comprising the terrestrial environment. The soil interacts both with the overlying atmosphere and the underlying strata, as well as with surface and underground bodies of water. Especially important is the interrelation between the soil and climate. In addition to its function of regulating the cycle of water, it also regulates energy exchange and surface temperature.

When land is cleared of vegetation and turned into a cultivated field, the native biomass above the ground is often burned and the organic matter within the soil tends to decompose. These processes release carbon dioxide into the atmosphere, thus contributing to the Earth's greenhouse effect and to global warming. On the other hand, the opposite act of afforestation and soil enrichment with organic matter, such as can be achieved through conservation management, may serve to absorb carbon dioxide from the atmosphere. To an extent, the soil's capacity to store carbon thus helps to mitigate the greenhouse effect.

Thousands of years are required for nature to create life-giving soil out of sterile bedrock. In only a few decades, however, unknowing or uncaring humans can destroy that wondrous work of nature. In various circumstances, mismanaged soils may be subject to erosion (the sediments of which tend to clog streambeds, estuaries, lakes, and coastal waters), to leaching of nutrients with attendant loss of fertility and eutrophication of water bodies, to waterlogging and impaired aeration, or to an excessive accumulation of salts that may cause a once-productive soil to become entirely sterile. Such processes of soil degradation, sometimes called "desertification," already affect large areas of land.

We cannot manage effectively and sustainably that which we do not know and thoroughly understand. That is why the tasks of developing and disseminating sound knowledge of the soil and its complex processes have assumed growing urgency and importance. The global environmental crisis has created a compelling need for a concentrated, concise, and definitive source of information—accessible to students, scientists, practitioners, and the general public—about the soil in all its manifestations in nature and in relation to the life of humans.

Daniel Hillel
May 2004

LIST OF CONTRIBUTORS FOR VOLUME 5

Andrey Abramov

Institute of Physicochemical and Biological Problems in Soil Science, Pushchino, Russia

Richard G Allen

Department of Soil and Water Systems, University of Idaho, Kimberly, ID, United States

Isabel Alves

Departamento de Ciências e Engenharia dos Biosistemas, Instituto Superior de Agronomia, Universidade de Lisboa, Lisboa, Portugal

Martha C Anderson

USDA-ARS, Hydrology and Remote Sensing Lab, Beltsville, MD, United States

Shmuel Assouline

Department of Environmental Physics and Irrigation, Agricultural Research Organization—Volcani Institute, Rishon LeZion, Israel

Jörg Bachmann

Institute of Soil Science, Leibniz Universität Hannover, Hannover, Germany

Steffen Beck-Broichsitter

Institute of Plant Nutrition and Soil Science, Christian-Albrechts-University Kiel, Kiel, Germany

Katrina E Bennett

Earth and Environmental Sciences, Los Alamos National Laboratory, Los Alamos, NM, United States

Markus Berli

Division of Hydrologic Sciences (DHS), Desert Research Institute, Las Vegas, NV, United States

Evgenia Blagodatskaya

Department of Soil Ecology, Helmholtz-Zentrum für Umweltforschung GmbH – UFZ, Halle (Saale), Germany

Guillaume Blanchy

Flemish Research Institute for Agriculture, Fisheries and Food (ILVO), Merelbeke, Belgium; Department of Environment, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium

Heye Reemt Bogen

Agrosphere Institute (IBG-3), Forschungszentrum Jülich GmbH, Jülich, Germany

Julia Boike

Alfred Wegener Institute Helmholtz Center for Polar and Marine Research (AWI), Potsdam, Germany; Department of Geography, Humboldt University Berlin, Berlin, Germany

Pascal Boivin

University of Applied Sciences of Western Switzerland, HES-SO, HEPIA-Agronomy, Geneva, Switzerland

Sara Bonetti

School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Naaran Brindt

Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, United States

Luca Brocca

Research Institute for Geo-Hydrological Protection, National Research Council, Perugia, Italy

Cosimo Brogi

Agrosphere Institute (IBG-3), Forschungszentrum Jülich GmbH, Jülich, Germany

Andrea Carminati

ETH Zürich, Zürich, Switzerland

David Caterina

Department of Urban and Environmental Engineering, University of Liège (ULiège), Liège, Belgium

Edwin Cey

University of Calgary, Calgary, AB, Canada

Brent Clothier

Plant and Food Research Ltd, Palmerston North, New Zealand

Simona Consoli

Università degli Studi di Catania, Catania, Italy

- Valentin Couvreur
Université catholique Louvain, Louvain-la-Neuve, Belgium
- Quirijn de Jong van Lier
Center for Nuclear Energy in Agriculture, University of São Paulo, Piracicaba, Brazil
- Philippe De Smedt
Department of Environment and Department of Archaeology, Ghent University, Ghent, Belgium
- Stéphane Declerck
Université catholique Louvain, Louvain-la-Neuve, Belgium
- TKK Chamindu Deepagoda
Department of Civil Engineering, Faculty of Engineering, University of Peradeniya, Peradeniya, Sri Lanka
- M Dentz
Spanish National Research Council (IDAEA-CSIC), Barcelona, Spain
- Ian C Dodd
Lancaster Environment Centre, Lancaster University, Lancaster, United Kingdom
- Jose Dörner
Instituto de Ingeniería Agraria y Suelos, Facultad de Ciencias Agrarias y Alimentarias, Universidad Austral de Chile, Valdivia, Chile
- Guido D'Urso
Universita' degli Studi di Napoli Federico II, Portici, Italy
- Wolfgang Durner
Institute of Geoecology, Technische Universität Braunschweig, Braunschweig, Germany
- Steven R Evett
Conservation and Production Research Laboratory, USDA-ARS, Bushland, TX, United States
- Robert Ewing
The Climate Corporation, Seattle, WA, United States
- Simone Fatichi
Department of Civil and Environmental Engineering, National University of Singapore, Singapore, Singapore
- Ty Ferre
Hydrology and Atmospheric Sciences, University of Arizona, Tucson, AZ, United States
- Heiner Fleige
Institute of Plant Nutrition and Soil Science, Christian-Albrechts-University Kiel, Kiel, Germany
- Shmulik P Friedman
Department of Environmental Physics and Irrigation, Institute of Soil, Water and Environmental Sciences, Agricultural Research Organization, The Volcani Center, Rishon LeZion, Israel
- Sarah Garré
Flemish Research Institute for Agriculture, Fisheries and Food (ILVO), Merelbeke, Belgium
- Kirill M Gerke
Schmidt Institute of Physics of the Earth of Russian Academy of Sciences, Moscow, Russia
- Behzad Ghanbarian
Kansas State University, Manhattan, KS, United States
- Teamrat A Ghezzehei
The University of California, Merced, Merced, CA, United States
- Marc-Oliver Goebel
Institute of Soil Science, Leibniz Universität Hannover, Hannover, Germany
- Sunny Eng Giap Goh
Universiti Malaysia Terengganu, Terengganu, Malaysia
- Michael J Goss
School of Environmental Sciences, University of Guelph, Guelph, ON, Canada
- Steve Green
Plant and Food Research Ltd, Palmerston North, New Zealand
- Pieter H Groenevelt
University of Guelph, Guelph, ON, Canada
- Jannis Groh
Forschungszentrum Jülich GmbH, Institute of Bio- and Geoscience IBG-3: Agrosphere, Jülich, Germany; Institute of Crop Science and Resource Conservation (INRES) – Soil Science and Soil Ecology, University of Bonn, Bonn, Germany; Leibniz Centre for Agricultural Landscape Research (ZALF), Landscape Functioning, Müncheberg, Germany
- Paulo Ivonir Gubiani
Federal University of Santa Maria, Santa Maria, Brazil
- Paul D Hallett
School of Biological Sciences, University of Aberdeen, Aberdeen, United Kingdom
- Shoichiro Hamamoto
Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan
- Jerry L Hatfield
National Laboratory for Agriculture and the Environment, Ames, IA, United States
- Masaki Hayashi
Department of Geoscience, University of Calgary, Calgary, AB, Canada

Joshua L Heitman
North Carolina State University, Raleigh, NC, United States

D Hillel
Formerly Columbia University, New York, NY, United States

Andrew D Hiron
University Centre Myerscough, Preston, United Kingdom

Jan W Hopmans
University of California Davis, Davis, CA, United States

Rainer Horn
Institute of Plant Nutrition and Soil Science, Christian-Albrechts-University Kiel, Kiel, Germany

Robert Horton
Iowa State University, Ames, IA, United States

Conrad Jackisch
Faculty of Geosciences, Geoengineering and Mining, Technische Universität Bergakademie Freiberg, Freiberg, Germany

Nicholas Jarvis
Department of Soil and Environment, Swedish University of Agricultural Sciences, Uppsala, Sweden

Mathieu Javaux
Earth and Life Institute, Agronomy, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

Yan Jin
University of Delaware, ISE Laboratories, Newark, DE, United States

Monika Joschko
Leibniz-Zentrum für Agrarlandschaftsforschung (ZALF) e. V., Müncheberg, Germany

Douglas L Karlen
Dlkarlen Consulting, LLC, Saint Paul, MN, United States

Marina V Karsanina
Schmidt Institute of Physics of the Earth of Russian Academy of Sciences, Moscow, Russia

Sheela Katuwal
Department of Poultry Science, University of Arkansas, Fayetteville, AR, United States

Thomas Keller
Department of Soil & Environment, Swedish University of Agricultural Sciences, Uppsala, Sweden; Agroscope, Department of Agroecology & Environment, Zürich, Switzerland

Ayse Kilic
University of Nebraska-Lincoln, Lincoln, NE, United States

John Koestel
Soil Fertility and Soil Protection, Agroscope - Standort Reckenholz, Zürich, Switzerland

Eva Kröner
Institute of Bio- and Geosciences, Agrosphere (IBG-3), Forschungszentrum Jülich, Jülich, Germany

William P Kustas
USDA-ARS, Hydrology and Remote Sensing Lab, Beltsville, MD, United States

Lars Kutzbach
Institute of Soil Science, Center for Earth System Research and Sustainability (CEN), Universität Hamburg, Hamburg, Germany

M Lamandé
Department of Agroecology, Aarhus University, Research Centre Foulum, Tjele, Denmark; Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway

Mats Larsbo
Department of Soil and Environment, Swedish University of Agricultural Sciences, Uppsala, Sweden

Volha Lazouskaya
University of Delaware, ISE Laboratories, Newark, DE, United States

Peter Lehmann
Physics of Soils and Terrestrial Ecosystems, Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland

Sally D Logsdon
Agricultural Research Service, USDA, Ames, IA, United States

N Lu
Colorado School of Mines, Golden, CO, United States

Maik Lucas
Department of Soil, Plant and Microbial Sciences, Michigan State University, East Lansing, MI, United States; DOE Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI, United States

Gabriele Manoli
School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Maria Marin
University of Aberdeen, Aberdeen, United Kingdom

- Wolde Mekuria
*International Water Management Institute (IWMI),
Addis Ababa, Ethiopia*
- Rajith Mukundan
*Bureau of Water Supply, New York City Department of
Environmental Protection, Kingston, NY, United States*
- Attila Nemes
*Faculty of Environmental Sciences and Natural Resource
Management, Norwegian University of Life Sciences, Ås,
Norway; Division of Environment and Natural Resources,
Norwegian Institute of Bioeconomy Research (NIBIO), Ås,
Norway*
- John R Nimmo
U.S. Geological Survey, Menlo Park, CA, United States
- John M Norman
*University of Wisconsin-Madison, Department of Soil
Science, Madison, WI, United States*
- Dani Or
*Institute of Biogeochemistry and Pollutant Dynamics, Soil
and Terrestrial Environmental Physics group, Swiss Federal
Institute of Technology (ETH Zurich), Zurich, Switzerland;
Division of Hydrologic Sciences, Desert Research Institute,
Reno, NV, United States*
- Christoforos Pappas
*Department of Civil Engineering, University of Patras,
Rio Patras, Greece*
- Paula Paredes
*Departamento de Ciências e Engenharia dos Biossistemas,
Instituto Superior de Agronomia, Universidade de Lisboa,
Lisboa, Portugal*
- J-Yves Parlange
*Department of Biological and Environmental Engineering,
Cornell University, Ithaca, NY, United States*
- Athanasios Paschalis
*Department of Civil and Environmental Engineering,
Imperial College London, South Kensington Campus,
London, United Kingdom*
- Luis S Pereira
*Departamento de Ciências e Engenharia dos Biossistemas,
Instituto Superior de Agronomia, Universidade de Lisboa,
Lisboa, Portugal*
- Everton Alves Rodrigues Pinheiro
Federal University of Tocantins, Gurupi, Brazil
- Dubravka Pokrajac
*School of Engineering, University of Aberdeen, Aberdeen,
United Kingdom*
- John H Prueger
*National Laboratory for Agriculture and the Environment,
Ames, IA, United States*
- Jaime Puértolas
*Department of Botany and Plant Ecology and Physiology,
University of La Laguna, San Cristóbal de La Laguna,
Spain*
- Thomas Pütz
*Forschungszentrum Jülich GmbH, Institute of Bio- and
Geoscience IBG-3: Agrosphere, Jülich, Germany*
- Alan W Rempel
*Department of Earth Sciences, University of Oregon,
Eugene, OR, United States*
- Dennis E Rolston
*Department of Land, Air and Water Resources, University
of California, Davis, CA, United States*
- Alejandro Romero-Ruiz
*Net Zero and Resilient Farming, Rothamsted Research,
Hertfordshire, United Kingdom*
- David Rudolph
University of Waterloo, Waterloo, ON, Canada
- Siul Ruiz
*Faculty of Engineering and Physical Sciences,
Bioengineering Group, University of Southampton,
Southampton, United Kingdom*
- David Russo
*Institute of Soil, Water and Environmental Sciences, The
Volcani Institute, Agricultural Research Organization,
Rishon LeZion, Israel*
- Thomas J Sauer
*National Laboratory for Agriculture and the Environment,
Ames, IA, United States*
- Steffen Schlüter
*Department of Soil System Science,
Helmholtz-Centre for Environmental Research - UFZ,
Halle, Germany*
- Andrea Schnepf
*Agrosphere Institute, IBG-3, Forschungszentrum Jülich
GmbH, Jülich, Germany*
- Stefan Schrader
Institute of Biodiversity, Braunschweig, Germany
- Chongyang Shen
*China Agricultural University, Haidian District, Beijing,
China*
- Nataline Simon
*Department of Urban and Environmental Engineering,
University of Liège (ULiège), Liège, Belgium*
- Ylva Sjöberg
*Department of Geosciences and Natural Resource
Management, University of Copenhagen, Copenhagen,
Denmark*

Tammo S Steenhuis

Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, United States; Faculty of Civil and Water Resources Engineering, Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar, Ethiopia

Alexia Stokes

University Montpellier, AMAP, INRAE, CIRAD, CNRS, IRD, Montpellier, France

Craig J Sturrock

Division of Agricultural and Environmental Sciences, School of Biosciences, University of Nottingham, Loughborough, United Kingdom

MC Sukop

Florida International University, Miami, FL, United States

Yukie Tanino

School of Engineering, University of Aberdeen, Aberdeen, United Kingdom

Christoph Tebbe

Institute of Biodiversity, Braunschweig, Germany

Desalegn Tegegne

International Water Management Institute (IWMI), Addis Ababa, Ethiopia

L ten Damme

Department of Agroecology, Aarhus University, Research Centre Foulum, Tjele, Denmark; Department of Soil & Environment, Swedish University of Agricultural Sciences, Uppsala, Sweden

Sally Thompson

Department of Civil, Environmental and Mining Engineering, University of Western Australia, Crawley, WA, Australia; Department of Civil and Environmental Engineering, University of California, Berkeley, CA, United States

Seifu A Tilahun

Faculty of Civil and Water Resources Engineering, Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar, Ethiopia; Department of Ecology and Conservation Biology, Texas A&M University, College Station, TX, United States

Kaitlin Togliatti

National Laboratory for Agriculture and the Environment, Ames, IA, United States

GC Topp

Agriculture and Agri-Food Canada, Ottawa, ON, Canada

Markus Tuller

Department of Environmental Science, The University of Arizona, Tucson, AZ, United States

Ariena HC van Bruggen

Department of Plant Pathology and the Emerging Pathogens Institute, University of Florida, Gainesville, FL, United States

JJH van den Akker

Wageningen University & Research, Wageningen Environmental Research, Wageningen, The Netherlands

Jan Vanderborght

Agrosphere Institute, IBG-3, Forschungszentrum Jülich GmbH, Jülich, Germany

Rodney T Venterea

U.S. Department of Agriculture, Agricultural Research Service, Soil and Water Management Research Unit, St. Paul, MN, United States

Harry Vereecken

Agrosphere Institute (IBG-3), Forschungszentrum Jülich GmbH, Jülich, Germany

Art Warrick

Environmental Science, University of Arizona, Tucson, AZ, United States

Jon M Wraith

College of Life Sciences and Agriculture, University of New Hampshire, Durham, NH, United States

Yuting Yang

Department of Hydraulic Engineering, Tsinghua University, Beijing, China

Anil Yildiz

Methods for Model-based Development in Computational Engineering, RWTH Aachen University, Aachen, Germany

Michael H Young

Bureau of Economic Geology, University of Texas at Austin, Austin, TX, United States

Mohsen Zarebanadkouki

Technical University of Munich, Freising, Germany

C Zhang

Hunan University, Changsha, Hunan, P.R. China

CONTENTS OF ALL VOLUMES

VOLUME 1: SOIL BIOLOGY

Overview

Edaphology	1
<i>RM Goss and AL Ulery</i>	

Biodiversity	10
<i>Monica A Farfan and Diana H Wall</i>	

Soil organisms

Bacteriophage (Viruses)	16
<i>Li Bi, Ji-Zheng He, Dan-Ting Yu, Li-Li Han, and Hang-Wei Hu</i>	

Archaea	25
<i>Mikhail V Semenov</i>	

Bacteria	31
<i>Mikhail V Semenov</i>	

Soil protists	39
<i>Stefan Geisen, Nathalie Amacker, and Alexandre Jousset</i>	

Fungi including mycorrhizal fungi	49
<i>F Buscot</i>	

Soil algae	61
<i>Karen Baumann and Patrick Jung</i>	

Soil arthropods: Underfoot and all around	70
<i>Ernest C Bernard</i>	

Nematodes	105
<i>Deborah A Neher and Thomas O Powers</i>	

Earthworms: Essential ecosystem engineers providing vital ecosystem services	112
<i>Maria JI Briones</i>	

Complex structures

Soil biofilms: Evolving concepts and ecological functions	120
<i>Yichao Wu, Marc Redmile-Gordon, Peng Cai, Chunhui Gao, and Qiaoyun Huang</i>	

Biocrusts	127
<i>Matthew A Bowker, Jasmine Anenberg, Hilina Yohannes Kebede, Anita Antoninka, and Lydia Bailey</i>	

Functioning

Studying soil microbial communities: Culture-dependent approaches	139
<i>Eduardo K Mitter, Dasiel Obregon, Micaela Tosi, Heather Slinn, and Kari Dunfield</i>	

Studying soil microbial communities: Culture-independent approaches	146
<i>Tolulope Mafa-Attoye, Heather Slinn, Micaela Tosi, Evan Mayer, Dasiel Obregon, and Kari Dunfield</i>	

Studying microbial function in soil: Culture-independent techniques <i>Micaela Tosi, Dasiel Obregon, Heather Slinn, and Kari Dunfield</i>	155
Kinetics of microbial processes: General principles <i>NS Panikov</i>	168
Kinetics of microbial processes: Applications to soil biology <i>NS Panikov</i>	186
Biogeochemical cycles	
Carbon cycle in soils: Dynamics and management <i>Charles W Rice, Carlos B Pires, and Marcos VM Sarto</i>	219
Carbon storage in soils <i>Isabelle Basile-Doelsch, Tiphaine Chevallier, Marie-France Dignac, and Amandine Erktan</i>	228
Soil organic nitrogen <i>EJ Daly and KA Congreves</i>	243
Nitrification: Process, organisms and management <i>Jeanette M Norton and Yang Ouyang</i>	253
Microbial denitrification <i>Michael J Goss, Sergey Blagodatsky, and Adrian Unc</i>	261
Symbiotic nitrogen fixation <i>Florence Mus and Hsin-Hua Wu</i>	270
Nitrogen: Non-Symbiotic Nitrogen fixation in soils <i>Vadakattu VSR Gupta and Sarah S Roley</i>	283
Phosphorus in soils—Biological interactions <i>Nataliya Bilyera and Iryna Loginova</i>	293
Biological transformations of sulfur in soil <i>Chaoqun Wang and Maxim Dorodnikov</i>	303
Plant uptake of mineral nutrients <i>Angela Hodge</i>	310
Higher plants and their interactions with soil organisms	
Soil Biology: Root form and function <i>Magda Garbowski, Grégoire Freschet, Louise Jackson, Cynthia Brown, and Louise Comas</i>	321
Root mucilage: Chemistry and functions in soil <i>Yutong Jiang, Lijun Hou, Fang Ding, and Joann K Whalen</i>	332
Root exudates and microorganisms <i>Aidan De Sena, Krisztina Mosdossy, Joann K Whalen, and Chandra A Madramootoo</i>	343
Root interactions with the microbiome from the rhizoplane to the bulk soil: An overview <i>Joann K Whalen and Shamim Gul</i>	357
Microorganisms & biological processes	
Soil respiration <i>Jeremiah D Vallotton, Sergey Blagodatsky, and Adrian Unc</i>	369
Anaerobic soil processes <i>PW Inglett, D Said-Pullicino, and K Ramesh Reddy</i>	379
Mineral-organic-microbial interactions <i>Balwant Singh, Sabina Yeasmin, and Donald L Sparks</i>	387

Microbial ecology & management

Habitat for soil organisms and their functions <i>Alexandra Kravchenko and Andrey Guber</i>	407
Plant growth promoters <i>Elisa Gamalero, Gustavo Santoyo, and Bernard R Glick</i>	418
Inoculation with microorganisms <i>Maureen O'Callaghan and David Wright</i>	428
Biocontrol of soil-borne plant diseases <i>Mehrdad Mohammadiani and Matthew Bakker</i>	439
Management of soil biology for crop protection from biotic and abiotic stresses <i>Michael J Goss, Isabel Brito, and Mário Carvalho</i>	447
Trees, hedges, agroforestry and microbial diversity <i>Josiah B Judson, Joseph Holden, Pippa Chapman, and Marcelo Valadares Galdos</i>	469
Urban soil microbiome: Activity, diversity and functioning <i>Maria Korneykova, Ekaterina Kozlova, Kristina Ivashchenko, Maria Vasileva, and Viacheslav Vasenev</i>	480
Microbiology of extreme soil environments <i>Patrice Dion</i>	494
Microbial mobility and transport in soils <i>Mohammad Reza Mosaddeghi, Adrian Unc, Ali Akbar Mahboubi, and Azadeh Safadoust</i>	512

Biology of soil amendments & contaminant degradation

Biology of compost <i>Heribert Insam, Thomas Klammersteiner, and María Gómez-Brandón</i>	522
Landfarming and its microbiology <i>Naga Raju Maddela, Balasubramanian Ramakrishnan, Kadiyala Venkateswarlu, and Mallavarapu Megharaj</i>	533
Biodegradation of organic contaminants <i>PB Hatzinger and JW Kelsey</i>	547
Biological aspects of manure management <i>Natasha Rayne and Bee Khim Chim</i>	558
Septic systems and municipal waste disposal on land <i>Charles P Gerba, Michael J Goss, and Adrian Unc</i>	567

Climate change

Climate change impacts on soil biology <i>Alberto Canarini, Lucia Fuchslueger, François-Xavier Joly, and Andreas Richter</i>	578
---	-----

VOLUME 2: CHEMISTRY**Basic Soil Chemistry**

Acidity and acidification <i>Siobhan Staunton</i>	1
pH <i>Siobhan Staunton</i>	8
Buffering capacity <i>Siobhan Staunton and Ravi Naidu</i>	15
Soil specific surface area <i>Emmanuel Arthur, Markus Tuller, and Kurt D Pennell</i>	22

Soil Analytical Chemistry

Cation exchange capacity and reactions <i>Donald S Ross and Ulf Skyllberg</i>	32
Testing for nutrients in the soil and the environment <i>Antonio P Mallarino</i>	43
Metal extraction and analysis <i>Guo Wang and Honghong Li</i>	50
Soil organic matter: Composition <i>Markus Kleber</i>	58
Physical fractionation techniques <i>Sam J Leuthold, Michelle L Haddix, Jocelyn Lavallee, and M Francesca Cotrufo</i>	68
The use of stable isotopes in soil science: Low atomic number elements <i>Chiara Pistocchi, Isabelle Bertrand, Kristýna Kantnerová, Astrid Oberson, Federica Tamburini, and Roland A Werner</i>	81
The use of radioactive isotopes to study soil pools and processes <i>Timothy I McLaren</i>	96
The use of stable isotopes in soil science: Metals <i>Cécile Quantin and Damien Guinoiseau</i>	107

Periodic Table

Alkali metals in soils <i>Guy J Levy and Asher Bar-Tal</i>	116
Alkaline earth metals in soil <i>Justin B Richardson</i>	126
Transition metals in soil <i>Siobhan Staunton</i>	135
Nitrogen and ammonia in soils <i>Nicholas T Girkin and Hannah V Cooper</i>	142
Phosphorus in soil <i>Luisella Celi, Maria Martin, and Elisabetta Barberis</i>	152
Aluminum in soils <i>Syuntaro Hiradate</i>	167
Iron in soils <i>Trolard Fabienne and Guilhem Bourrié</i>	176
Sulfur and selenium in soil <i>Dongli Liang, Fei Zhou, and Mingxing Qi</i>	192
Arsenic in soil <i>Md Imran Ullah Sarkar, Afsana Jahan, Md Tofail Hossain, Mohammad Mahmudur Rahman, and Ravi Naidu</i>	201
Non-transition elements: Metals and metalloids in soils <i>Natasha, Sana Khalid, Muhammad Shahid, Faiz Rabbani, and Camille Dumat</i>	214
Lanthanides, actinides and radon <i>Franco Ajmone Marsan and Mery Malandrino</i>	225
Halogens in soils <i>Svensson Teresia, Thiry Yves, Bueno Maïté, and Oelmann Yvonne</i>	234
Transformation of metals and metalloids by microorganisms <i>Susan M Glasauer, Euan P Burford, and Geoffrey Michael Gadd</i>	243
Dynamics of radionuclides <i>Siobhan Staunton</i>	256
Silicon and other non-metal elements from Group 14 <i>Bruno Delvaux and Zimin Li</i>	263

Soil Physical Chemistry

Chemical equilibria	274
<i>Arthur Paul Schwab</i>	
Dissolution and precipitation processes	284
<i>Zhenqing Shi, Shiwen Hu, Rong Li, and Yang Lu</i>	
Reactions, redox potential and kinetics	297
<i>Trolard Fabienne and Guilhem Bourrié</i>	
Sorption of organic chemicals in soil	315
<i>Joseph J Pignatello</i>	
Sorption—Oxyanions	327
<i>CP Schulthess, H Wijnja, W Yang, Anahita Khosravi, and Baoshan Xing</i>	
Sorption—Metals	336
<i>Daniel G Strawn and Donald L Sparks</i>	
Sorption–desorption kinetics	343
<i>Gautier Landrot and Donald L Sparks</i>	
Sorption-desorption hysteresis	355
<i>Wenhao Wu, Hongxia Zhu, and Kun Yang</i>	
Surface charge and zero-charge points	368
<i>J Chorover</i>	
Flocculation and dispersion	376
<i>Guy J Levy and Isaac Shainberg</i>	

Soil Carbon Chemistry

Soil organic matter dynamics	387
<i>Yu Luo and Jianming Xu</i>	
Soil proteins	396
<i>Siobhan Staunton</i>	
Biochar: Properties, mechanisms, and interactions in the soil	406
<i>Balwant Singh, Shamim Mia, and Marta Camps-Arbestain</i>	
Dissolved organic carbon	421
<i>Jason James and Rob Harrison</i>	
Black carbon, pyrogenic carbon	432
<i>Heike Knicker</i>	
Fire effects on soil	448
<i>Stefan H Doerr, Cristina Santín, and Jorge Mataix-Solera</i>	
Interfacial reactions of microorganisms with minerals and organic matter	458
<i>Rota Wagai, Puu-Tai Yang, and Klaus Kaiser</i>	
Pesticide fate in soils	470
<i>Pierre Benoit, Laure Mamy, Carole Bedos, and Enrique Barriuso</i>	
Soil organic matter in molecular simulations	483
<i>Martin H Gerzabek, Daniel Tunega, Edgar Galicia-Andrés, and Chris Oostenbrink</i>	

Soil Physical Analysis

Raman spectroscopy	497
<i>Anupam Das and Huiyuan Guo</i>	
Fourier transform infrared spectroscopic methods of soil analysis	510
<i>Derek Peak</i>	
Electron paramagnetic resonance spectroscopy: Part I Historic Perspectives	519
<i>Giorgio S Senesi and Nicola Senesi</i>	

Electron paramagnetic resonance spectroscopy: Part II the view forward <i>Robert L Cook, Emmanuel Ampiah, Phillip T Sprunger, and Randall W Hall</i>	533
Fluorescence Spectroscopy: Part I Principles <i>Yong Zhang, Shuai Guo, and Chaoxian Wei</i>	544
Fluorescence spectroscopy: Part II practical applications <i>Yong Zhang, Chaoxian Wei, and Shuai Guo</i>	552
Near and mid infrared soil spectroscopy <i>Wartini Ng, Brendan Malone, Budiman Minasny, and Sangho Jeon</i>	567
Electron microscopic methods (TEM, SEM and energy dispersal spectroscopy) <i>Jian Zhao and Xia Liu</i>	575
Application of atomic force microscopy (AFM) on environmental interfaces <i>Saikat Ghosh, Zhang Di, Bo Pan, and Baoshan Xing</i>	589
Applications of nuclear magnetic resonance for the study of soils <i>Daniel H Lysak, Myrna J Simpson, and Andre J Simpson</i>	605
Neutron-gamma analysis of soil elements <i>Aleksandr Kavetskiy, Galina Yakubova, Nikolay Sargsyan, Stephen A Prior, and H Allen Torbert</i>	625
X-Ray Diffraction <i>Chao Wang and Cheng Gu</i>	642
Soil analysis using laser-induced breakdown spectroscopy <i>Débora Marcondes Bastos Pereira Milori, Ladislau Martin-Neto, Hero T Gollany, Diego Victor de Babos, Luis Carlos Leva Borduchi, and Paulino Ribeiro Villas-Boas</i>	654
Calorimetric analysis <i>Xingmin Rong and Qiaoyun Huang</i>	662
Understanding of soil processes at the microscale—Use of NanoSIMS in soil science <i>Carsten W Mueller, Carmen Hoeschen, and Ingrid Koegel-Knabner</i>	670
Nanoparticles in soil <i>Di Lang, Quan Chen, Yang Liu, and Bo Pan</i>	681
Overview	
Soil chemistry and climate change <i>Jianming Xu, Yong Li, and Zhongmin Dai</i>	692
VOLUME 3: ENVIRONMENT AND MANAGEMENT	
Soil functions, potential uses of soil and ecosystem services	
Soil functions, soil potential uses and ecosystem services <i>Karl Stahr</i>	1
Farming systems to conserve land for nature <i>Rattan Lal</i>	12
Geoarchaeology and soils <i>Carlos E Cordova</i>	20
Civilizations and the role of soil, agriculture, and climate change <i>Jeffrey A Homburg, Eric C Brevik, and Jonathan A Sandor</i>	29
Ecosystem services provided by soils in highly anthropized areas (SUITMAs) <i>Jean Louis Morel, Geoffroy Séré, Viatcheslav Vasenev, and Thomas Nehls</i>	37
Structure and function of forested soils <i>Peter Clinton and Jen Owens</i>	56
Agroforestry <i>NV Thevathasan, AM Gordon, and PKR Nair</i>	68

Rangeland management <i>Alexandra Sandhage-Hofmann</i>	88
Effects on soil of grassland management for pasture, hay and silage <i>Carmen Segura, Laura Cardenas, Richard McDowell, Sarah Morgan, and Martin SA Blackwell</i>	102
Dryland farming: Technological and management options for sustainable agriculture and food systems <i>Ch Srinivasarao, G Ranjith Kumar, Ravula Manasa, Kiran Pilli, Samaresh Sahoo, S Rakesh, Sumanta Kundu, K C Nataraj, KV Rao, JVNS Prasad, and S Malleswari</i>	113
The effect of crop rotations on soil <i>Caio Fernandes Zani, Arlete Simões Barneze, Rogério Peres Soratto, and Charles Andrew Francis</i>	125
The contribution of organic farming systems to soil fertility—A systems perspective <i>Bernhard Freyer, Pierre Ellssel, Jürgen K Friedel, and Kurt Möller</i>	135
Soil functions to support natural plant communities and crops: Soil multifunctionality <i>Leidivan Almeida Frazão, Luiz Arnaldo Fernandes, Adriana Marcela Silva Olaya, and Igor Costa de Freitas</i>	146
Methane and nitrous oxide emissions from land <i>Joanna M Cloy and Keith A Smith</i>	153
Soil and land degradation	
Overview chapter on soil degradation <i>K Müller, Margaret A Oliver, and C Siebe</i>	165
Soil monitoring <i>Ben Marchant and Nicolas Saby</i>	172
Soil quality and health key indicators <i>Edoardo AC Costantini and Simone Priori</i>	181
Water induced soil erosion <i>R Jane Rickson</i>	193
Erosion by wind <i>John Tatarko</i>	208
Human exposure to aeroparticles emitted from unpaved roads due to traffic resuspension and wind erosion <i>Arón Jazcilevich and Ann G Wellens</i>	223
Soil salination processes and management <i>Uri Nachshon and Guy J Levy</i>	236
Degradation of soil structure and soil organic matter <i>Andrew S Gregory</i>	246
Drainage, soil water storage, buffering and filtering <i>Jeffrey Strock and Norm Fausey</i>	257
Environmental effects of irrigation <i>Sevilay Topcu</i>	266
Trade-offs of wastewater irrigation <i>Manzoor Qadir and Christopher A Scott</i>	277
Pollution and risks of trace elements in the soil environment <i>Peng Wang, Fang-Jie Zhao, and Steve P McGrath</i>	288
Pollutants—Persistent organic <i>Rai S Kookana and Divina A Navarro</i>	296
Pollution by microplastic in agricultural areas <i>Nicolas Beriot and Esperanza Huerta-Lwanga</i>	309

Municipal solid waste management: Production, management, and environmental effects <i>Bryant M Serre and Lynda H McCarthy</i>	321
Remediation of metals and organic contaminants in soil <i>Peter Sanderson, Md Mezbaul Bahar, Bhabananda Biswas, and Ravi Naidu</i>	333
Soil change under different scenarios	
Soil change under different scenarios <i>Joann K Whalen</i>	344
The effect of global change on the soil body <i>Sophie Cornu, David Montagne, and Antra Boca</i>	360
The effects of adequate and excessive application of mineral fertilizers on the soil <i>Bijay-Singh and Tek B Sapkota</i>	369
Foliar Application of Nutrients <i>Thomas Eichert and Victoria Fernández</i>	382
Mulches and cover crops part I: Types <i>KK Bandyopadhyay, CL Acharya, and KM Hati</i>	392
Mulches and cover crops part II: Role in soil health and climate resilient agriculture <i>KK Bandyopadhyay, CL Acharya, and KM Hati</i>	401
The potential agronomic and environmental applications of biochar: Prospects and challenges <i>Yu-Jun Wang, Ping Wu, Nanthi S Bolan, and Hai-Long Wang</i>	414
Soil management for carbon sequestration <i>Christopher Brock, Uwe Franko, and Martin Wiesmeier</i>	424
Soil tillage <i>Mathieu Lamandé, Lars Juhl Munkholm, and Trond Børresen</i>	433
Weed and soil management: A balancing act <i>Trevor Kenneth James and Charles Norman Merfield</i>	439
Terraces as traditional agricultural landscapes. Stability amidst change <i>Gerardo Bocco</i>	450
Soil quality protection policies and plans to ensure sustainability <i>Ram Swaroop Meena and Sandeep Kumar</i>	457
Construction of Technosols for green roofs and vertical green system <i>Eleonora Flores-Ramírez, Stefan Abel, Elizabeth Chávez-García, and Geoffroy Séré</i>	473
Management of park areas, sport fields, and school yards, including golf courses and public right of ways <i>Maxine J Levin and Anna Paltseva</i>	488
Soil-society relations	
Overview chapter on soil-society relations <i>C Siebe, Margaret A Oliver, and K Müller</i>	505
Soil art: Sensory and symbolic engagement with soils <i>Alexandra R Toland and Daniel Wolter</i>	509
Soil science in the criminal justice system <i>Lorna Anne Dawson</i>	521
Soil governance at the international, regional and national level <i>Harald Ginzky</i>	532
The economics of soils' contribution to human well-being <i>Bartosz Bartkowski and Julian R Massenber</i>	547

Soil and human health	555
<i>Margaret A Oliver and Eric C Brevik</i>	
Soil Science Education, Awareness and Outreach in the Modern World	572
<i>Laura Bertha Reyes-Sánchez and Eric C Brevik</i>	
Soils in landscape planning	580
<i>Seth Denizen</i>	

VOLUME 4: PEDOLOGY

Pedology

Pedology—Basic principles	1
<i>Rupert Bäumler, Uta Stockmann, and Margaret A Oliver</i>	

Factors of soil formation, pedogenic processes and mineralogy

Overview chapter on factors of soil formation	7
<i>Rupert Bäumler, Margaret A Oliver, and Uta Stockmann</i>	
Factors of soil formation: Climate	14
<i>Stefaan Dondeyne, Stephan Mantel, and Seppe Deckers</i>	
Factors of soil formation: Soil biota and organic matter as a factor for soil formation and weathering processes	25
<i>Scott Werts</i>	
Soil forming factors: Parent material	34
<i>Stephen Nortcliff</i>	
Parent materials: Loess	42
<i>Tobias Sprafke</i>	
Time	48
<i>Susan E Melzer, Caroline M Yonker, and Eugene F Kelly</i>	
Factors of soil formation—Human impacts	54
<i>Jonathan A Sandor, C Lee Burras, Michael Thompson, and Skye A Wills</i>	
Dynamic soils	69
<i>Markus Egli and Dennis Dahms</i>	
Interactions between soils and climate change	82
<i>Cornelia Rumpel</i>	
Primary minerals	91
<i>Balwant Singh</i>	
Mineral components of the soil clay fraction	109
<i>Darrell G Schulze</i>	
Short-range ordered aluminosilicates	121
<i>Tetsuhiro Watanabe, James B Harsh, and Rota Wagai</i>	
Metal oxides	135
<i>Andreas C Scheinost and Balwant Singh</i>	
Soil morphology	149
<i>Phillip R Owens and Zamir Libohova</i>	
Soil micromorphology	171
<i>Rosa M Poch and Georges Stoops</i>	
Soil Classification and distribution	
History and principles of soil classification	185
<i>Maria Gerasimova and Dmitry Konyushkov</i>	

Soil classification systems: World reference base and soil taxonomy	197
<i>David Badía-Villas</i>	
World reference base for soil resources (WRB)	206
<i>Stephan Mantel, Stefaan Dondeyne, and Seppe Deckers</i>	
World soil map based on soil taxonomy	218
<i>Eswaran Padmanabhan and Paul F Reich</i>	
Soils as archives of city history	232
<i>Hermine Huot and Jean Louis Morel</i>	
Classification of soils by ecozones	
Ecozones and soils – An introduction	246
<i>Rupert Bäumler</i>	
Cold Region Soils: Part I—Distribution, Properties and Processes	250
<i>Nicolas A Jelinski, Chien-Lu Ping, and John CF Tedrow</i>	
Cold region soils: Part II – Genesis and classification	263
<i>Nicolas A Jelinski, Chien-Lu Ping, and John CF Tedrow</i>	
Boreal forest soils	280
<i>Gabriele Broll</i>	
Soils of humid cool temperate regions	289
<i>Nicolas A Jelinski, Justin B Richardson, and Edward A Nater</i>	
Grassland soils in the cool–arid–temperate ecozone (Steppe, grassland)	299
<i>Mengyang You, Xiaobing Liu, Rupert Bäumler, and Margaret A Oliver</i>	
Soils from a warm-temperate semi-tropical ecozone (Mediterranean) with humid winter months	319
<i>J Torrent, Marco Pfeiffer, and Juan José Ibañez</i>	
Arid and semi-arid subtropical and tropical ecozone (II)	332
<i>Bernhard Lucke</i>	
Humid subtropical ecozone with emphasis on Acrisols and Alisols	343
<i>Nilton Curi, Sérgio Henrique Godinho Silva, and Alberto Vasconcellos Inda</i>	
Soils of the humid and sub-humid tropics	353
<i>Karl Stahr and Ludger Herrmann</i>	
Classification of soils by properties and environmental settings	
Sodic soils	368
<i>Guy J Levy and Uri Nachshon</i>	
Organic soils	383
<i>Heike Schimmel and Wulf Amelung</i>	
The pedology of anaerobic soils	398
<i>John Hollis and Allan Lilly</i>	
Alluvium and alluvial soils	417
<i>Janis L Boettinger</i>	
Hydric Soils	424
<i>MJ Vepraskas and JF Berkowitz</i>	
Wetland soils	436
<i>Cornelius Wilhelm van Huyssteen</i>	
Paddy soils: Characteristics, productivity and management	447
<i>Stephan M Haefele</i>	
Volcanic soils and non-volcanic andic soils	458
<i>Rupert Bäumler and G Uehara</i>	

PEDOMETRICS**Pedometrics**

Overview of Pedometrics	471
<i>Alexandre MJ-C Wadoux, Inakwu OA Odeh, and Alex B McBratney</i>	

Statistics and soil data analysis

Statistics in soil science	486
<i>Richard Webster</i>	

Multivariate analysis for soil science	499
<i>Si Yang Han, Patrick Filippi, Mercedes Román Dobarco, Januar Harianto, Mathew S Crowther, and Thomas FA Bishop</i>	

Machine learning and processing of large data	509
<i>Madlene Nussbaum</i>	

The issue of scale and change of support in the spatial analysis of environmental data	521
<i>A Castrignanò, A Belmonte, and N Romano</i>	

Soil measurements in space and time

Sampling for soil survey	535
<i>DJ Brus</i>	

Handling of soil samples in the field, in transport to the laboratory and in the laboratory	551
<i>Claudy Jolivet, Céline Ratié, Déborah Chavrit, Nicolas Soler-Dominguez, and Antonio Bispo</i>	

Digital Soil Morphometrics	568
<i>Yakun Zhang, Alfred E Hartemink, Jingyi Huang, and Budiman Minasny</i>	

Proximal soil sensing in the field	579
<i>Mikaela Tilse, Uta Stockmann, and Patrick Filippi</i>	

Proximal sensing of soil moisture	591
<i>Markus Tuller, Ebrahim Babaeian, and Scott B Jones</i>	

Proximal sensing of land surface temperature	600
<i>Ebrahim Babaeian and Markus Tuller</i>	

Proximal sensing of evapotranspiration	610
<i>Ebrahim Babaeian and Markus Tuller</i>	

Remote sensing of soil moisture	618
<i>Alexander Gruber and Jian Peng</i>	

Remote Sensing of Soil Organic Carbon	631
<i>Bas van Wesemael and Sabine Chabrilat</i>	

Pedotransfer functions and their application to soil water dynamics	642
<i>Marcel G Schaap, Yonggen Zhang, and Attila Nemes</i>	

Spatial and temporal soil variation

Geostatistics: Modelling spatial variation	655
<i>Richard Webster</i>	

Uncertainty assessment of spatial soil information	671
<i>Gerard BM Heuvelink and Richard Webster</i>	

Digital soil mapping: Evolution, current state and future directions of the science	684
<i>Brendan Malone, Dominique Arrouays, Laura Poggio, Budiman Minasny, and Alex B McBratney</i>	

Geographical information systems (GIS) and soils	696
<i>S Grunwald and J Böhner</i>	

Precision agriculture	710
<i>James A Taylor</i>	

Foundations, measurements and trends in pedodiversity <i>Juan José Ibañez and Marco Pfeiffer</i>	726
Soil landscape modelling	
Modeling soil development in a landscape context <i>Peter A Finke, Budiman Minasny, and Arnaud JAM Temme</i>	739
Climate and land surface models: Role of soil <i>Toby Richard Marthews, Holger Lange, Alberto Martínez-de la Torre, Richard J Ellis, Sarah E Chadburn, and Martin G De Kauwe</i>	751
VOLUME 5: SOIL PHYSICS	
Soil structure & mechanics	
Soil structure <i>Steffen Schlüter and John Koestel</i>	1
Aggregation <i>Teamrat A Ghezzehei</i>	8
Porosity and pore-size distribution <i>John R Nimmo, Sheela Katuwal, and Maik Lucas</i>	16
Pore-scale modelling of flow and transport phenomena in soils <i>Kirill M Gerke and Marina V Karsanina</i>	25
Soil texture: Its relevance, measurement and representation internationally <i>Attila Nemes</i>	35
Soil tilth <i>Thomas Keller, Douglas L Karlen, and Paul D Hallett</i>	48
Crusts and seals: Structural <i>Shmuel Assouline and Sally Thompson</i>	57
Soil mechanics <i>Markus Berli and Paul D Hallett</i>	69
Compaction <i>JJH van den Akker, L ten Damme, M Lamandé, and Thomas Keller</i>	85
Bioturbation—Physical processes <i>Siul Ruiz, Paul D Hallett, and Dani Or</i>	100
The reinforcement of soil by plant roots <i>Alexia Stokes and Anil Yildiz</i>	115
Physical effects of soil salinity <i>David Russo</i>	123
Shrink-Swell processes in soils <i>Pascal Boivin</i>	139
Water retention	
Water properties <i>N Lu, C Zhang, and D Hillel</i>	146
Water potential <i>Dani Or, Markus Tuller, and Jon M Wraith</i>	159
Water content and water potential measurement <i>Wolfgang Durner</i>	168
Soil water retention and characteristic curve <i>Markus Tuller and Dani Or</i>	187
Soil water repellency <i>Jörg Bachmann and Marc-Oliver Goebel</i>	203

Flow & transport processes

Aeration	216
<i>Rodney T Venterea and Dennis E Rolston</i>	
Capillarity	225
<i>Dani Or and Markus Tuller</i>	
Darcy's law	240
<i>Conrad Jackisch and Eva Kröner</i>	
Diffusion	247
<i>TKK Chamindu Deepagoda and Shoichiro Hamamoto</i>	
Hydrodynamics in soil	256
<i>Ty Ferre and Art Warrick</i>	
Thermodynamics of soil water	262
<i>Pieter H Groenevelt</i>	
Immiscible Fluids	271
<i>Yukie Tanino and Dubravka Pokrajac</i>	
Overland flow	287
<i>Tammo S Steenhuis, Rajith Mukundan, and Seifu A Tilahun</i>	
Direction-dependency of soil hydraulic and mechanical properties	294
<i>Steffen Beck-Brochsisitter, Heiner Fleige, Jose Dörner, and Rainer Horn</i>	
Macropores and macropore flow	306
<i>Nicholas Jarvis and Mats Larsbo</i>	
Unstable preferential flow in porous media	314
<i>Naaran Brindt, J-Yves Parlange, and Tammo S Steenhuis</i>	
Leaching processes	321
<i>Brent Clothier and Steve Green</i>	
Colloid transport	336
<i>Yan Jin, Chongyang Shen, and Volha Lazouskaya</i>	
Scaling of transport properties	353
<i>Robert Ewing and Behzad Ghanbarian</i>	
Soil influence on groundwater	364
<i>David Rudolph and Edwin Cey</i>	

Thermal properties and behaviour

Temperature effects of soil hydraulic properties	379
<i>Jan W Hopmans and Robert Horton</i>	
Heat and vapor transfer	385
<i>Sunny Eng Giap Goh, Joshua L Heitman, and Robert Horton</i>	
Radiation Balance	393
<i>Jerry L Hatfield, Kaitlin Togliatti, Thomas J Sauer, and John H Prueger</i>	
Freezing and thawing cycles	400
<i>Alan W Rempel</i>	
Arctic permafrost	410
<i>Julia Boike, Andrey Abramov, Katrina E Bennett, and Lars Kutzbach</i>	
Permafrost and cold region hydrology	419
<i>Masaki Hayashi and Ylva Sjöberg</i>	

Sensing Techniques

Soil water sensing by neutron scattering	427
<i>Steven R Evett</i>	

Time-domain reflectometry <i>GC Topp and Ty Ferre</i>	436
Geophysical methods for soil applications <i>Sarah Garré, Guillaume Blanchy, David Caterina, Philippe De Smedt, Alejandro Romero-Ruiz, and Nataline Simon</i>	444
Non-invasive imaging of soil processes <i>Craig J Sturrock</i>	459
Hydrology & water balance	
Soil processes in the hydrologic cycle <i>Cosimo Brogi, Harry Vereecken, Heye Reemt Bogena, and Luca Brocca</i>	469
Global water cycle from a soil perspective <i>Michael H Young and Dani Or</i>	482
Water evaporation from bare soil <i>Peter Lehmann</i>	490
Evapotranspiration <i>Yuting Yang</i>	501
Plant available water <i>Quirijn de Jong van Lier, Sally D Logsdon, Everton Alves Rodrigues Pinheiro, and Paulo Ivonir Gubiani</i>	509
Plant-water relations <i>Ian C Dodd, Andrew D Hiron, and Jaime Puértolas</i>	516
Water use efficiency: A review of spatial and temporal variability <i>Simone Fatichi, Athanasios Paschalis, Sara Bonetti, Gabriele Manoli, and Christoforos Pappas</i>	527
Root water uptake and myco-rhizosphere hydraulic properties <i>Andrea Carminati, Maria Marin, Valentin Couvreur, Mohsen Zarebanadkouki, Stéphane Declerck, and Mathieu Javaux</i>	543
Plant-soil-water relations <i>Jan Vanderborght, Quirijn de Jong van Lier, Andrea Schnepf, and Mathieu Javaux</i>	552
Crop and landscape water requirements <i>Luis S Pereira, Isabel Alves, and Paula Paredes</i>	571
Water harvesting <i>Wolde Mekuria and Desalegn Tegegne</i>	593
Irrigation methods <i>Shmulik P Friedman</i>	608
Soil–plant–atmosphere continuum <i>John M Norman, Martha C Anderson, and William P Kustas</i>	624
The Penman-Monteith equation <i>Richard G Allen, Ayse Kilic, Guido D'Urso, and Simona Consoli</i>	634
Infiltration <i>Ty Ferre and Art Warrick</i>	646
Solute transport <i>MC Sukop and M Dentz</i>	652
Lysimetry <i>Thomas Pütz and Jannis Groh</i>	667
Recognition of notable past soil scientists <i>Michael J Goss, Evgenia Blagodatskaya, Ariena HC van Bruggen, Monika Joschko, Stefan Schrader, and Christoph Tebbe</i>	680
Index	691

SOIL PHYSICS

Soil structure

Steffen Schlüter^a and John Koestel^b, ^aDepartment of Soil System Science, Helmholtz-Centre for Environmental Research - UFZ, Halle, Germany; ^bSoil Fertility and Soil Protection, Agroscope - Standort Reckenholz, Zürich, Switzerland

© 2023 Elsevier Ltd. All rights reserved.

This is an update of V.A. Snyder, M.A. Vázquez, STRUCTURE, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 54-68, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00533-6>.

Introduction	1
Soil structure and soil functions	2
Soil structure dynamics	3
Methods of structure analysis	5
Implications for modeling	6
Summary and outlook	7
References	7

Abstract

Soil structure is the spatial arrangement of particles, pores and organic components of soil. It influences many processes in soil and is at the same time shaped by them. Soil structural properties are therefore important indicators for soil functions. This chapter explains the links between soil structure and soil functions, describes important drivers of soil structure dynamics, provides an overview of methods to characterize soil structure at different spatial scales and discusses implications for modelling.

Key points

- Dichotomy of pore structure and solid structure.
- Constituents, functions and measurement techniques of soil structure are summarized with respect to spatial scale.
- Implications for models of water and matter cycles.

Introduction

Soil structure has been defined as the spatial heterogeneity of different components and properties of soil (Dexter, 1988). This comprises the spatial arrangement of both solid constituents and pores at various scales, e.g., from the way in which flocculated clay platelets assemble at the nm scale to earthworm burrows that stretch down the entire soil profile at m scale. There are two different approaches of characterizing soil structure: either by the size, shape and grade of soil aggregates as diagnostic features or by the volume fraction, size and connectivity of pores (Vogel et al., 2021). Both have merits in understanding soil functions as all processes in soil occur in pores and at pore surfaces, but it is the binding forces of particles that retain pore shape and connectivity. A comparison of scales is given in Fig. 1. Pores are typically classified according to their size and resulting functions. Macropores have a diameter larger than 10–50 μm (Blume et al., 2016), in which gravity exceeds capillarity. These pores drain quickly and are important for soil aeration. In mesopores, water is held against gravity by capillary forces that are weak enough to be overcome by plants to extract it. Mesopores not only constitute the reservoir of plant-available water, but also the habitable pore space for most microorganisms, many of which are restricted to aqueous environments. Micropores ($<0.2 \mu\text{m}$ according to German soil classification, Blume et al., 2016) are too small to accommodate most bacteria and water is too tightly bound in it to be extracted by plants (matric potential $<-1500 \text{ kPa}$). Different subcategories and size limits are used by other classification systems and disciplines, e.g., soil micromorphology or colloid and surface chemistry. Although most studies report macropores to be greater than 10–50 μm in size, some studies have used a cut-off as large 1000 μm (Luxmoore, 1981). Likewise, micropores are generally less than 0.2 μm , but the size range used as the cut-off can be as large as 10–30 μm (Luxmoore, 1981; Brewer, 1964). Pores are also classified into primary (or textural) pores, which are voids resulting from assembly of irregularly-shaped particles, and secondary (or structural) pores resulting from aggregation and structure-forming processes. These structural pores are the reason why a fine-textured, loamy soil can have an equal or even greater saturated hydraulic conductivity than a coarse-textured, sandy soil.

Soil structure also manifests itself as the spatial variability of the stability against mechanical stress. Different binding agents act at different scales and have different strengths, arising from very stable polyvalent cations and organo-mineral associations at the molecular scale, to roots and hyphae that enmesh particles at mm-cm scales. As a result, soils harbor a hierarchy of failure zones. When these are activated by mechanical stress, e.g., through tillage or sieving, soils crumble into fragments, which have traditionally been termed clods, peds or aggregates. Further information can be found in the articles on Aggregates and Tilth found in the encyclopedia.

Soil structure and soil functions

Many processes in soil are sensitive to soil structure and many soil functions depend on it (Figs. 1 and 2). The pore structure provides the pathways for water and matter fluxes, and the habitat for soil biota. In addition, the pores and pore-solid interfaces shape chemical reaction patterns in space, altogether giving rise to a variety of micro-environmental conditions encountered in soil. At the same time, the soil structure is shaped by many biotic and abiotic processes leading to various feedback mechanisms between soil structure and the processes occurring in it (Fig. 2). These feedbacks are the reason why soil structure can be used as an indicator of many essential ecosystem functions of soil like plant growth, maintenance of biodiversity, carbon storage, organic matter turnover or water retention (Rabot et al., 2018).

Due to the utility of soil structure in characterizing soil quality or health, scores of what constitutes “good” or “poor” soil structure have been developed for field assessment. These have been derived by regression analysis of the properties or interest, e.g., crop yield, infiltration capacity, with readily available proxies for soil structure such as bulk density or air capacity (Rabot et al., 2018). Bulk density is a very easy to obtain parameter defined by the mass of dry soil in a given volume. With compaction, more soil is present in a given volume at the expense of lost porosity, so density increases. As bulk density lumps all pores together and is influenced by soil texture, it can be a poor indicator of soil structure functions such as root growth (Kaufmann

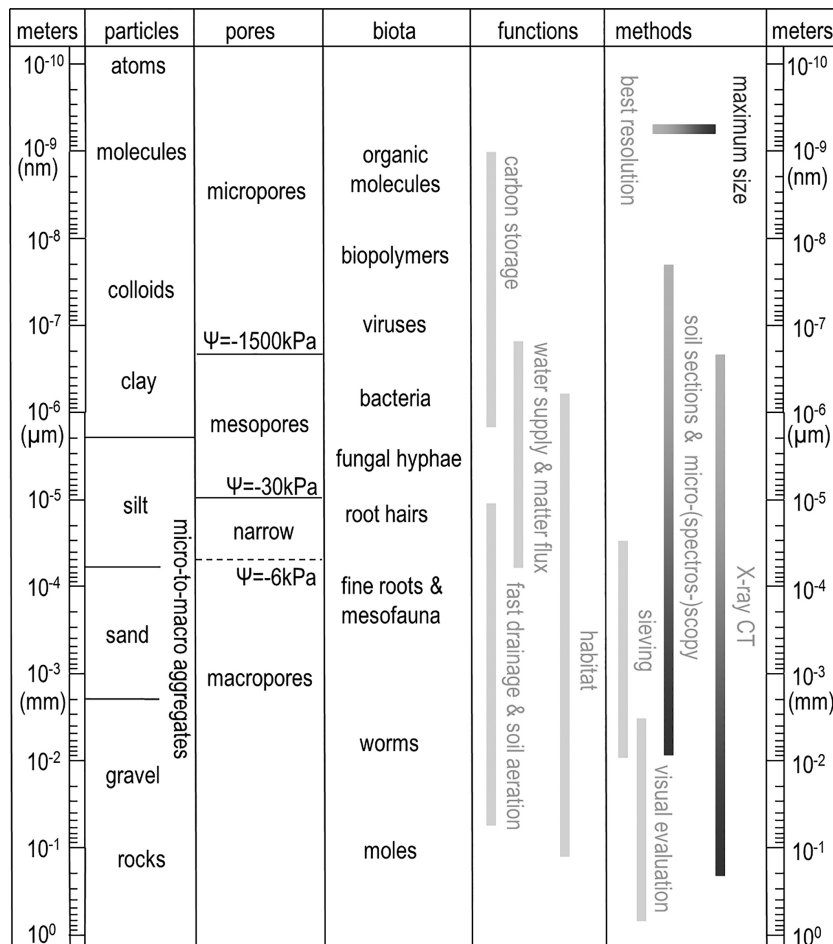


Fig. 1 Comparative scales in soil structure including abiotic and biotic constituents as well as a selection of associated soil functions. Size cut-offs for particles and pores according to German soil taxonomy (Blume et al., 2016).

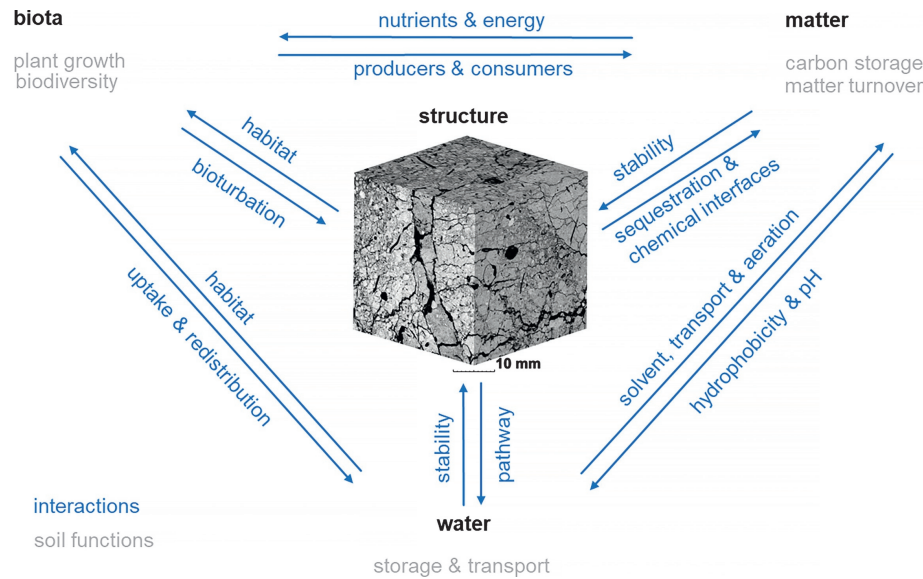


Fig. 2 Interactions between soil structure, soil biota, water and matter, as well as the soil functions that are maintained by these interactions.

et al., 2010). That is why other indicators like the degree of compactness that accounts for texture (Reichert et al., 2009), or more descriptive measurements of pore structure like the S-index (Dexter, 2004) have been advocated. Useful information on soil structure can also be estimated from water retention characteristics; particularly at critical cut-off points such as field capacity and wilting point. Mechanical properties of soil can also characterize how soil structure affects root growth and burrowing fauna. Therefore, combined mechanical and hydrological measurements, such as the least limiting water range (Da Silva et al., 1994), have been developed.

Soil structure dynamics

Soil structure is not static, but continuously modified by various processes compiled in Fig. 2. Weather has a large impact, primarily through the action of wetting and drying the soil that can cause structure to collapse or disaggregate under wetting, and crack under drying. In tilled agricultural soils, cultivation fragments the soil to produce tilth that can rapidly change depending on the stability of the soil. A common measurement of soil stability assesses the breakdown of aggregates through laboratory-imposed stresses of rapid or slow wetting, or mechanical perturbation.

Soil biota including plant roots, earthworms, termites, ants and others have the strength to reshape the pore space to explore the soil for resources and protection. This mixing process called bioturbation is a major driver for soil structure dynamics (compare Fig. 3a and b). Soils harbor an earthworm population of some 100 kg ha^{-1} , that can easily incorporate several tons of annual litter fall per ha (Schaefer and Schauer mann, 1990). In addition, plant roots and soil fauna have an indirect effect on soil structure by enhancing its stability through the release of organic compounds, gluing agents and physical enmeshment by roots, hyphae and plant residues. The soil carbon brought about by these organic amendments often correlates positively with soil structural stability, when land use and soil management have been stable over long periods of time, e.g., continuous forest, pasture, or tillage type. The soils have then evolved into a state in which structure stabilization by organo-mineral associations and organic matter stabilization by physical protection within the soil structure has reached an equilibrium. For a given type of soil management, correlation between organic matter and structural stability can be highly soil-dependent, with texture playing a major role. In many cases, a certain threshold level of organic matter is necessary before soil structure begins stabilizing (Fig. 4). The threshold value tends to increase with increasing soil clay content, suggesting that a critical amount of organic matter per unit mineral surface area must be exceeded before effective stabilization occurs.

Crack formation through shrinking and swelling of 2:1 clay minerals is an important abiotic structure-forming process (compare Fig. 3c). Once these desiccation cracks have been initiated in a drying event, they can be partially refilled by other particles, which is a dominant pedogenesis process in shrinking soils such as vertisols. In addition, clay platelets at the crack surfaces get reoriented by compression with the next swelling event, so that the failure zones persist and get reactivated upon the next drying events. Freezing and drying cycles can have a similar effect on soil structure, as water migrating to the freezing front dries out the soil elsewhere. In addition, when ice expansion leads to local pressure build up, it can lead to rupture and soil fragmentation.

In addition to naturally occurring biotic and abiotic structure forming processes, tillage and other soil management-induced processes can modify soil structure, sometimes unintentionally. The main goal of soil tillage, besides weed control and

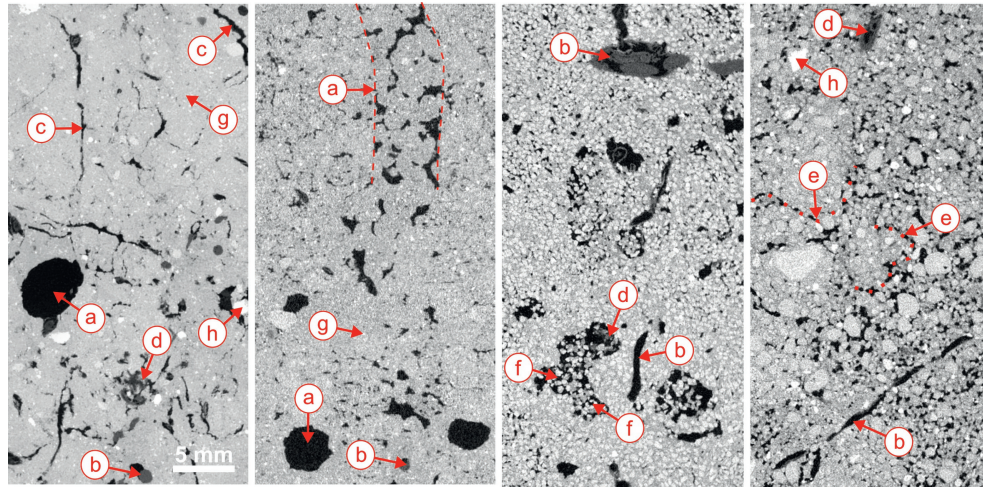


Fig. 3 Intact soil structure scanned with X-ray at a resolution of 20 μm . The samples were taken in (from left to right): Bad Lauchstädt, Germany (grassland, topsoil, 12% sand, 68% silt, 20% clay), Garzweiler, Germany (crop rotation, below plow layer, 5% sand, 81% silt, 14% clay), Hadera, Israel (citrus orchard, topsoil, 65% sand, 16% silt, 19% clay) and Kühnfeld, Halle, Germany (continuous maize, conventional tillage, topsoil, 63% sand, 25% silt, 12% clay). The annotated structure features comprise: (a) empty and re-filled earthworm burrows, (b) roots and root channels, (c) desiccation cracks, (d) plant residues, (e) hypothetical failure zones, (f) primary pores between particles, (g) soil matrix and (h) primary particles.

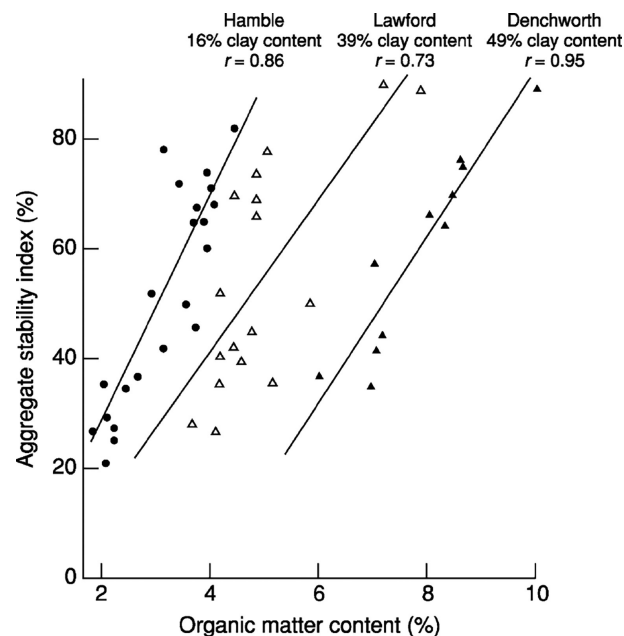


Fig. 4 Relation between structure stability and organic matter content in soil with varying texture. Structure stability was measured as aggregate stability to wet sieving (Douglas and Goss, 1982; Haynes and Beare, 1996). Reproduced with permission from Haynes RJ and Beare MH (1996) Aggregation and organic matter storage in Meso-thermal, humid soils. In: Carter MR and Stewart BA (eds) *Structure and Organic Matter Storage in Agricultural Soils*, p. 232. Boca Raton, FL: CRC Press. Redrawn from Douglas JT and Goss MJ (1982) Stability and organic matter content of surface soil aggregates under different methods of cultivation and in grassland. *Soil & Tillage Research* 2: 155–175. doi: 10.1016/0167-1987(82)90023-X.

incorporation of plant residues, is to loosen the soil and create a seedbed. Soil compaction may arise during land management. This can be intentional, e.g., rolling for soil reconsolidation after sowing, or unintended from machinery or livestock. Highly compacted soils have decreased pore space and connectivity, leading to poor aeration and greater mechanical resistance to root growth. Other land management practices may change soil structure and its stability, either as the primary goal or a secondary impact. This includes liming, which aims to flocculate clay and enhance cementation of particle contacts through cation bridges. Liming does this by replacing monovalent K^+ or Na^+ cations with polyvalent Ca^{2+} cations on reactive surfaces. Any management choice that increases carbon stocks or prevents the reduction of carbon stocks by protection against soil erosion, such as catch crops and manure application can have a similar, beneficial effect on the stability of soil structure.

Methods of structure analysis

Soil structure is an important parameter in soil survey and classification. Analyzing soil structure is inherently difficult due to the opaque nature of soil. Over the past decade, there has been a surge in the use of non-invasive methods like X-ray tomography that can visualize pore space, but this is limited to research laboratories and very expensive. Non-invasive approaches make it possible to measure the 3-D arrangement of solids and pores in soils directly, but indirect approaches provide a valuable estimation. Early investigations of soil structure therefore involved analyses of soil fragments, termed aggregates, as a proxy for the 3-D arrangement of particles and pores.

The simplest examination of soil structure can be done in the field from a visual inspection of soil aggregates that have been manually extracted from the soil by experts. The criteria used in this classification scheme are type or shape, class or size, and grade or distinctness of the aggregate. Different structure types are illustrated in Fig. 5. Visual soil structure characterizations normally take place in the field, often at soil profiles in freshly dug pits for soil surveying and classification.

Visual assessments of soil structure have been extended to assess the impacts of land management practices. These shovel-based approaches are rapid and use clear keys to minimize variability between operators inherent to qualitative assessments. One approach is the Visual Evaluation of Soil Structure (Guimarães et al., 2011) which extracts a section of soil to spade depth and assesses the breakup of the soil into fragments. The abundance, distribution, shape and size of visible pores and roots growing into the fragments, and evidence of mottling, are used to provide a soil structural quality score. Visual structure characterization has the advantage of simplicity and a very low demand in technical resources. It is however hindered by subjectivity, since both the extraction of the soil aggregates as well as their classification are based on the judgement of an expert.

Soil fragments can also be extracted and investigated in the laboratory under well-defined conditions, providing greater objectivity than rapid field assays. These can either use tilth, i.e., soil already fragmented into aggregates by tillage, or aggregates isolated from the soil through mechanical agitation and sorting by size through sieving (Kemper and Rosenau, 1986). The aggregates can then be investigated as individual entities, measuring properties such as aggregate density or strength, or used to assess soil aggregate stability. Soil aggregate stability tests mimic stresses experienced in the field that destabilize soils and may lead to erosion or crusting. The different stresses and their relevance are: (i) rapid wetting to simulate intensive rainfall on dry soil that induces slaking; (ii) capillary wetting similar to groundwater rise that induces slumping; and (iii) mechanical breakdown as would occur from raindrop impact or tillage (Le Bissonnais, 1996). Mechanical breakdown is typically imposed by stirring the soil, but the use of ultra-sonication has been advocated to impose more controlled energy inputs. Very high ultra-sonication energy coupled with oxidation of the organic material can be applied to break down larger aggregates to even smaller and more stable pieces. While aggregate-based approaches of soil structure characterization are still widely in use and of great utility for assessing soil structural quality and erosion resistance, their failure to capture the original 3-D arrangement of solids and pores of intact soil brings criticism (Vogel et al., 2021).

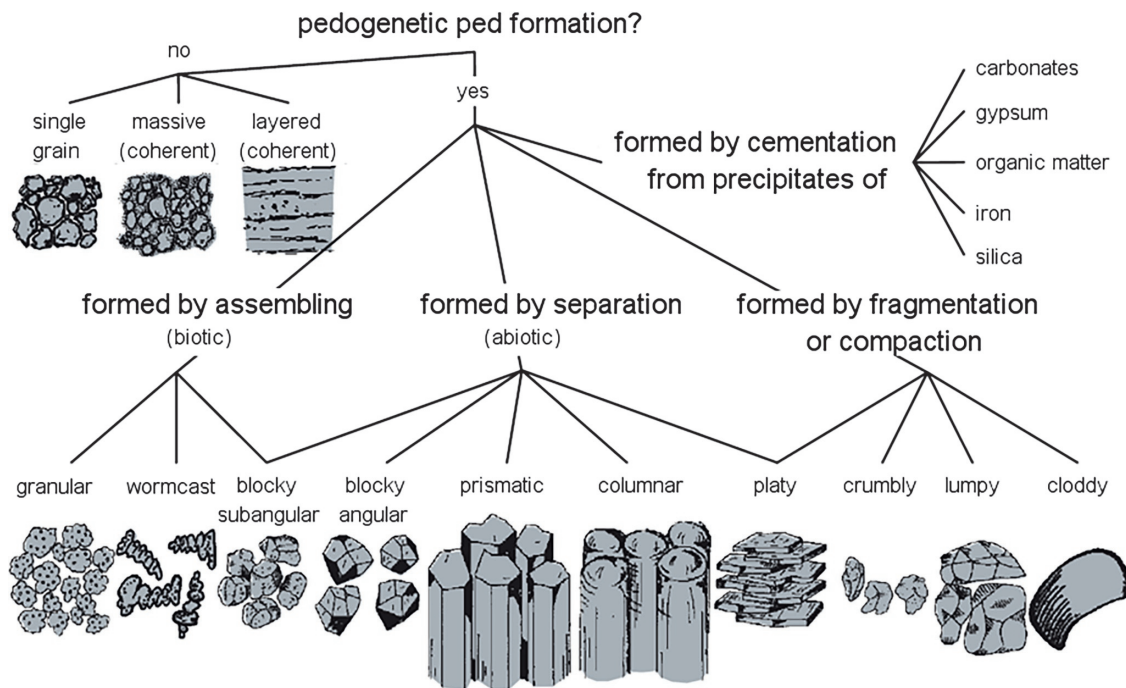


Fig. 5 Classification of soil structure by aggregate characterization based on visual appearance. Reproduced with permission from FAO (2006) *Guidelines for Soil Description*. 4th edn. FAO: Rome.

Stemming from early research on soil genesis and classification, 2-D thin sectioning approaches were developed to retain the intact structure of soil (Bullock et al., 1985). The sections are created by taking intact specimens that are impregnated with a resin and then attached to a glass slide before being cut and ground to 10–30 μm thickness. Micromorphological examination of thin sections allows for the analyses of the spatial arrangement of pores and solids without breaking up the original soil structure into aggregates. In addition to intact pore structure, soil thin sections have the added benefit of allowing for mapping biochemical heterogeneity and the spatial distribution of microbiota with various microscopy and micro-spectroscopy techniques. The method is however associated with a large workload and limited to 2-D structure characterization. 3-D X-ray imaging has emerged as a major tool to quantify structures in natural, undisturbed soil. Based on X-ray computed tomography, previously inaccessible structural information can be quantified, such as the connectedness of individual pores or the bottleneck in the pore connection from one side of the sample to the other.

The methods for structure analyses, described above, have characteristic scales at which they can be applied (Fig. 1). X-ray imaging can visualize pores at sub-micron resolution, but only for small samples a few mm in size. By rule of thumb, the respective image resolution is presently in the range of the sample diameter divided by 1000–2000. The image resolution is expected to improve in the coming years with further developments of the detector hardware. Thin-sections visualize pores down to the thickness of the section, so typically 10–30 μm diameter. The lateral resolution of micro-(spectro)scopic imaging applied to these soil sections varies with optics and spot sizes of different techniques. Aggregate analyses in the laboratory consider aggregate diameters between some micrometers and a few centimeters. Visual soil structure inspection is limited to structures larger than a millimeter but has the advantage, that it is applicable to larger scales, i.e., up to more than 1 m at a soil profile. Geophysical methods like electrical resistivity tomography (ERT) and georadar (GPR) have been tested for soil structure mapping at the plot and field scale (10 m and more). While these methods may offer useful insights, they suffer from large ambiguities in the measurement results (ERT) or are only applicable on dry soil with low clay contents (GPR).

Implications for modeling

The structure of the soil pore-network and its stability determine how gases are exchanged and water and associated solutes and colloids are flowing under a given hydraulic gradient. It therefore needs to be reflected in models that simulate gas and water and solute movement through the soil. The nature of the pore network makes this a non-trivial task. Soil structure may be characterized by an entropy, with complete disorder as in clean sand with only primary pores at the one extreme and complete order as with a bundle of capillaries at the other. In the majority of cases, soil structures exhibit intermediary entropies, which are inherently difficult to grasp within mathematical models. Early models conceptualized soil as one of the two extremes on the entropy scale, most prominently in the form of Richards' equation and the advection-dispersion equation, which fit best to the case of maximal entropy. Later, dual and multi domain models emerged to amend some of the obvious shortcomings of these simple models. Alternative approaches have aimed at capturing the pore-network hierarchy with fractal relationships. Albeit fractal dimensions can be derived from soil structure and similar power law scaling appears in some soil hydraulic functions, relationships with which pore-structures could be derived from hydraulic functions using only fractal properties (or vice versa) are yet not available despite decades of research. Yet another approach to implement the effects of soil structure are stochastic models that view soil as a spatially correlated random field of hydraulic properties. However, commonly assumed random fields match rather poorly to e.g., hydraulic conductivity fields obtained from X-ray image data of soils. Eventually, there is often no short-cut from measuring the pore structure in individual soil samples to answer more specific questions on water, solute and air movement in this soil, such as the amount of soil by-passed under preferential flow or the volume fraction of well-aerated soil. Köhne et al. (2009) give an overview of existing water flow and solute transport model approaches for structured soil and you can read more in the articles on Pore Scale Physical Processes, Scaling and Solute Transport found in this encyclopedia.

Carbon and nutrient cycles are driven predominantly by microbes and plant roots, which in turn require specific redox and moisture conditions. Carbon and nutrient cycles in soil are therefore also intimately coupled to the exchange of gases and movement and retention of water and dissolved organic compounds (See section "Soil structure and soil functions"). For example, root growth, and therefore the supply of fresh organic matter, is guided by macropore structures and water-movements in smaller pores. Microbial communities that mineralize carbon efficiently need environments with a good balance of water supply, as the solvent of carbon substrate and nutrients, and oxygen supply, as an energetically efficient electron donor for soil respiration. Respective model simulations need to be necessarily based upon the above-discussed hydraulic models, which makes predictive modelling of carbon mineralization and the associated biology a highly complex task (Moyano et al., 2013). The same holds for models of nutrient cycles (nitrogen, phosphorus, among others), for which similar links to soil structure exist.

The fact that soil structure is constantly evolving adds an extra layer of complexity to modelling soil systems and makes it a truly challenging field of research. Models for soil structure evolution are still at the very forefront of scientific research and have only recently been emerging.

Summary and outlook

Soil structure underpins flow and transport, microbial habitats, root growth channels and mechanical behavior. It is therefore an important property of soil that affects soil quality, with stresses from land mismanagement presenting a major challenge.

Soil is said to be one of the most complex materials known. The main ingredient for this complexity is the soil structure, which compartmentalizes soil into a large number of different regions with various degrees of connections between them. This renders understanding and modelling of physical, chemical, biological or ecological processes in soil systems a challenging endeavor, as they all strongly depend on individual soil structures. At the same time, this means that knowing the soil structure is key in appraising soil functioning. Soil structure derived indicators are therefore included when classifying soil quality.

Future developments in hardware will provide improved imaging methods with which not only the 3-D physical soil structure but also the associated chemical properties and embedded soil biology can be quantified better. Promising approaches involve correlative imaging, in which two or more complementary imaging methods are superimposed (Hapca et al., 2015; Schlüter et al., 2019). Improved computational power and modelling approaches will then enable to implement this improved information on the nature of the soil pore-network into 3-D models, which will be needed to further untangle the complexity of soil systems.

References

- Blume H-P, Brümmer GW, Horn R, Kandeler E, Kögel-Knabner I, Kretschmar R, Stahr K, Wilke B-M, Thiele-Bruhn S, and Welp G (2016) *Scheffer/Schachtschabel Soil Science*. Berlin: Heidelberg, Springer.
- Brewer R (1964) *Fabric and Mineral Analysis of Soils*. New York: Krieger.
- Bullock P, Fedoroff N, Jongerius A, Stoops G, and Tursina T (1985) *Handbook for Soil Thin Section Description*. Waine Research.
- Da Silva AP, Kay BD, and Perfect E (1994) Characterization of the least limiting water range of soils. *Soil Science Society of America Journal* 58: 1775–1781.
- Dexter AR (1988) Advances in characterization of soil structure. *Soil & Tillage Research* 11: 199–238.
- Dexter AR (2004) Soil physical quality: Part I. Theory, effects of soil texture, density, and organic matter, and effects on root growth. *Geoderma* 120: 201–214.
- Douglas JT and Goss MJ (1982) Stability and organic matter content of surface soil aggregates under different methods of cultivation and in grassland. *Soil & Tillage Research* 2: 155–175. [https://doi.org/10.1016/0167-1987\(82\)90023-X](https://doi.org/10.1016/0167-1987(82)90023-X).
- Guimarães RML, Ball BC, and Tormena CA (2011) Improvements in the visual evaluation of soil structure. *Soil Use and Management* 27: 395–403.
- Hapca S, Baveye PC, Wilson C, Lark RM, and Otten W (2015) Three-dimensional mapping of soil chemical characteristics at micrometric scale by combining 2D SEM-EDX data and 3D X-ray CT images. *PLoS One* 10: e0137205.
- Haynes RJ and Beare MH (1996) Aggregation and organic matter storage in Meso-thermal, humid soils. In: Carter MR and Stewart BA (eds.) *Structure and Organic Matter Storage in Agricultural Soils*, p. 232. Boca Raton, FL: CRC Press.
- Kaufmann M, Tobias S, and Schulin R (2010) Comparison of critical limits for crop plant growth based on different indicators for the state of soil compaction. *Journal of Plant Nutrition and Soil Science* 173: 573–583.
- Kemper WD and Rosenau RC (1986) *Aggregate stability and size distribution*. Methods of Soil Analysis. <https://doi.org/10.2136/sssabookser5.1.2ed.c17>.
- Köhne JM, Köhne S, and Simunek J (2009) A review of model applications for structured soils: A Water flow and tracer transport. *Journal of Contaminant Hydrology* 104: 4–35.
- Le Bissonnais Y (1996) Aggregate stability and assessment of soil crustability and erodibility 1. Theory and methodology. *European Journal of Soil Science* 47: 425–437.
- Luxmoore RJ (1981) Micro-, Meso-, and macroporosity of soil. *Soil Science Society of America Journal* 45: 671–672.
- Moyano FE, Manzoni S, and Chenu C (2013) Responses of soil heterotrophic respiration to moisture availability: An exploration of processes and models. *Soil Biology and Biochemistry* 59: 72–85.
- Rabot E, Wiesmeier M, Schlüter S, and Vogel HJ (2018) Soil structure as an indicator of soil functions: A review. *Geoderma* 314: 122–137.
- Reichert JM, Suzuki LEAS, Reinert DJ, Horn R, and Håkansson I (2009) Reference bulk density and critical degree-of-compactness for no-till crop production in subtropical highly weathered soils. *Soil and Tillage Research* 102: 242–254.
- Schaefer M and Schauermaier J (1990) The soil fauna of beech forests: Comparison between a mull and a moder soil. *Pedobiologia* 34: 299–314.
- Schlüter S, Eickhorst T, and Mueller CW (2019) Correlative imaging reveals holistic view of soil microenvironments. *Environmental Science & Technology* 53: 829–837.
- Vogel H-J, Balseiro-Romero M, Kravchenko A, Otten W, Pot V, Schlüter S, Weller U, and Baveye PC (2021) A holistic perspective on soil architecture is needed as a key to soil functions. *European Journal of Soil Science* 73: e13152.

Aggregation

Teamrat A Ghezzehei, The University of California, Merced, Merced, CA, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of J.R. Nimmo, AGGREGATION I Physical Aspects, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 28–35, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00532-4>.

Introduction	9
Mechanical aggregates (fragments)	10
Hierarchical aggregates	10
Life cycle of hierarchical aggregates	10
Sources of organic matter for aggregation	10
Organic matter forms and transformation relevant to aggregation	11
Biotic and abiotic feedback	12
Characterization	12
Dry aggregate size distribution	12
Water stable aggregation	12
Aggregate strength	13
Fractionation of organic matter in aggregated soils	13
Imaging	13
Functions and interpretations	13
Biogeochemical functions	14
Resistance to degradation	14
Flow and transport	14
Summary and outlook	14
References	14

Abstract

Aggregation is a vital characteristic of soil structure that affects its physical and biogeochemical properties. Aggregation results from the cohesion of primary minerals with organic or inorganic constituents. It depends on the dynamic balance between binding and fragmentation. There are two classes of aggregation. Mechanical aggregates are formed instantly by external forces and are often unstable. Hierarchical aggregates result from slow binding and are stable. Characterization of aggregation includes size, shape, stability, configuration, and their arrangement within soil. The nature of hierarchical aggregation is inferred from size separation of aggregates. Aggregate stability indicates their ability to persist under disruptive forces.

Glossary

Aggregate stability The ability of aggregates to persist under internal (e.g., rapid wetting or desiccation) and external (e.g., tillage or bioturbation) disruptive forces.

Hierarchical aggregation Binding of soil constituents by various abiotic and biotic binding factors spanning multiple scales.

Mechanical fragmentation Formation of soil aggregates by mechanical fragmentation of large clods.

Rhizodeposit Organic deposits added by living or dead roots, including short-lived root hairs, sloughed-off cells, exudates, and mucilage.

Rhizosheath An armor of soil particles that strongly binds to living roots and remains attached to the root system when pulled out of the soil and shaken vigorously.

Slaking The process of fragmentation of an aggregate by escaping entrapped air when rapidly submerged in water.

Key points

- Soil aggregates are essential features of soil structure, responsible for soil physical and biogeochemical characteristics relevant to agricultural and ecosystem functions.

- There are two broad classes of soil aggregates; aggregates formed by (a) mechanical fragmentation and (b) slow hierarchical binding of soil constituents. Mechanical fragments are formed instantly by external forces, while hierarchical aggregates are formed through the slow binding of soil constituents by various abiotic and biotic binding factors spanning multiple scales.
- The life cycle of soil hierarchical aggregation is closely associated with the biophysical dynamics of soil organic matter, including source, dynamics, and biotic-abiotic feedback. Higher-order hierarchical soil aggregation exists only when organic matter is the primary driver.
- Soil aggregation has three related characteristics: the size and shape of individual aggregates, the stability of aggregates, and the configuration and arrangement of aggregates within the undisturbed soil.
- Aggregate size separation and stability are determined through sieving and applying disruptive forces that reflect forces encountered in real field conditions.

Introduction

Soil aggregates are composite soil units formed by the cohesion of primary mineral with organic constituents that remain stable when subjected to disruptive forces. Aggregation is an essential feature of soil structure—the spatial arrangement of soil constituents and the resultant complex pore network, which spans a wide range of sizes and connectivity. The state of soil aggregation is a product of a dynamic balance of binding and fragmentation. It is characterized by the constant transfer of bound soil constituents between neighboring volume elements of aggregated soil, often bound together to form even larger and weaker aggregates. In addition, soil aggregates are more delicate than primary mineral constituents and readily respond to disturbances such as cultivation, fire, drought, flooding, and changes in vegetation. As a result, soil aggregation evolves faster than soil texture and mineralogy, which change at geologic time scales. Therefore, the evolution of soil aggregation is responsible for numerous alterations of soil physical and biogeochemical characteristics within the time frame relevant to agricultural and ecosystem functions.

Conversely, the processes responsible for aggregate dynamics, including biological activity, wetting-drying cycles, and dissolution and precipitation of soil constituents, are also directly influenced by the state of soil aggregation. There are two broad classes of soil aggregation: (a) mechanical fragmentation and (b) slow binding of hierarchical aggregates. Although mechanical and hierarchical aggregates exhibit similar morphology after separation, their functions are distinct (Fig. 1).

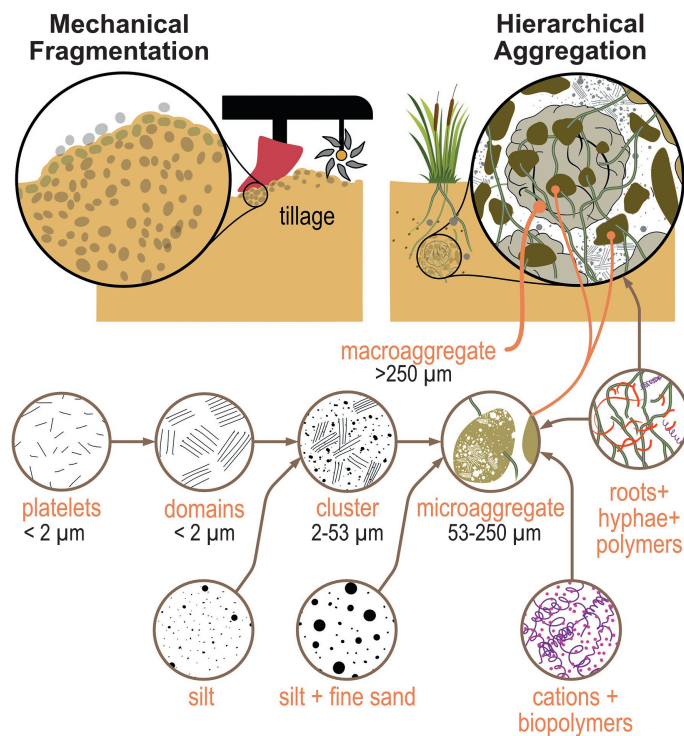


Fig. 1 Two main mechanisms of soil aggregation: (A) mechanical aggregation by breakdown of clods and (B) hierarchical aggregation by complex biotic and abiotic interactions.

Mechanical aggregates (fragments)

The fragmentation of large soil clods in cohesive soils, e.g., by tillage, freezing-thawing cycles, and burrowing animals, instantly creates an assemblage of differentiated *mechanical aggregates (fragments)* (Pires et al., 2017). These aggregates are readily observable in freshly prepared seedbeds (Hadas, 1997). The binding agents that hold these aggregates together vary with the soil composition and may include abiotic (clay, oxides, and carbonates) or biotic (organic matter). Their characteristic feature is that they form instantly by external forces. Mechanically derived aggregates are often unstable and revert to coarser blocks when neighboring aggregates coalesce by capillary attraction during wetting-drying cycles (Ghezzehei and Or, 2000).

Hierarchical aggregates

The slow binding of soil constituents by various abiotic and biotic binding factors spanning multiple scales results in *hierarchical aggregates* (Tisdall and Oades, 1982). Electrostatic bonding of clay aluminosilicate platelets (flocculation) forms $<2\ \mu\text{m}$ **clay domains**, the lowest element in the hierarchical order of soil aggregation. Flocculation is the balance of both repulsive and attractive forces. Repulsion arises due to similarly charged ions attached to clay surfaces. An attractive force emerges when the distance between clay platelets is small enough that opposing sides share ions. The distance is dependent on the hydrated size of ions. With their small hydration spheres, divalent cations such as Ca and Mg promote flocculation. On the other hand, the large, hydrated volume of monovalent ions such as Na and Li causes the dispersion of clays.

Further bonding of stable clay domains and silt particles results in the formation of 2–20 μm clusters. Colloidal aluminosilicates and Fe and Al oxides are the primary binding agents at this scale (Totsche et al., 2018). In addition, the desiccation of colloidal organic molecules derived from root exudation and microbial extracellular polymeric substances (EPS) can form hardened glue that binds clusters (Tisdall and Oades, 1982). Clusters are also referred to as medium-sized microaggregates.

The binding of clusters and fine and medium-sized sand grains results in the formation of large (20–250 μm) microaggregates. The nature and persistence of binding agents at this scale vary considerably with soil type and management but generally include processed organic macromolecules, polyvalent metal cation complexes, oxides, and highly disordered aluminosilicates (Totsche et al., 2018).

Macroaggregates ($>250\ \mu\text{m}$ but typically not bigger than several millimeters) constitute the largest class in the hierarchy. The bonds that hold macroaggregates together are usually weaker and more transient than those that bind microaggregates. These include enmeshing by hyphae and root hairs or binding by microbial EPS and root exudates (polysaccharides). Macroaggregates ($>250\ \mu\text{m}$) are formed during the rapid mineralization of fresh organic matter inputs. Microaggregates develop gradually within macroaggregates (Tisdall and Oades, 1982) as more of the organic matter is mineralized and forms a more stable association with the mineral surfaces.

Large soil clods ($>25\ \text{mm}$) are not considered part of the hierarchical aggregation concept. Such lumps need not contain lower-order aggregates and are typically formed by the desiccation of soils with a large fraction of clays, oxides, or carbonates. In agricultural soils, disruption of stable hierarchical aggregates by tillage and subsequent compaction by machinery can form dense clods (Dexter, 1988).

The study of soil aggregation and its function is usually preceded by the separation of (sieving) aggregates from bulk soil. Therefore, mechanical fragments are rarely distinguished from biologically derived hierarchical aggregates owing to their similarities in appearance and size. Moreover, sieving removes the spatial organizational context that aggregates occupy in soil structure. This ambiguity in our conceptual definition of soil aggregates conceals significant differences in the genesis and ecological functions of soil aggregates formed in natural and managed soils (Or et al., 2021, and references cited therein). This ambiguity is particularly crucial in managed soils subjected to reduced tillage, in which both forms of aggregate formation can coexist and interact.

Life cycle of hierarchical aggregates

Higher-order hierarchical soil aggregation exists only when organic matter is the primary driver (Tisdall and Oades, 1982). Therefore, *the hierarchical conceptual model plays a crucial role in explaining soil organic matter dynamics*. Conversely, the life cycle of soil aggregation is closely associated with the biophysical dynamics of soil organic matter (SOM), including source, dynamics, and biotic-abiotic feedback.

Sources of organic matter for aggregation

For most natural ecosystems, the predominant source of SOM is primary production by local vegetation. Other forms of organic matter inputs from outside sources (e.g., manure and compost inputs, deposition of eroded sediments, dissolved or suspended organic matter in surface or subsurface streams of water, and feces and remains of fauna) may play a role in specific conditions. Root-derived SOM inputs dominate overall stocks and aggregate-associated carbon (C) (Rasse et al., 2005). Roughly half of photosynthetically fixed plant C is allocated belowground as root tissues, rhizodeposits, and root respiration (Nguyen, 2003). A significant portion of the aboveground biomass is processed and lost before being incorporated into the underlying soil by mixing

(e.g., earthworms, rodents or agricultural practices) or leaching as dissolved organic matter (DOM). For example, the leaf residue of forest floors is degraded in the litter layer, with only a small fraction entering the underlying soil (Crow et al., 2009). The surface-applied residue is decomposed mainly in situ, with a small fraction getting incorporated in subsurface pools in the form of dissolved organic matter (DOM) that later gets incorporated into mineral-associated fractions (silt and clay) (Gale and Cambardella, 2000). Moreover, root-derived C persists for a much longer time in soils (Rasse et al., 2005). The physical form of root-derived organic matter and the environmental conditions it encounters in aggregated soils account for the majority of this preferential stability. The centrality of roots in soil aggregation is apparent when one gently pulls out or excavates the root system of most plants. Soils inhabited by dense and fine root systems, such as ryegrass, promote stable macroaggregates that withstand post-tillage collapse by wetting and drying (Cockroft and Olsson, 2000). Furthermore, the degree of aggregation is correlated with root density (Fig. 2).

Organic matter forms and transformation relevant to aggregation

A key feature of the organic matter storage that distinguishes well-aggregated soils is the relative quantity of particulate organic matter (POM) (Gale and Cambardella, 2000). The amount of soil incorporated in macroaggregates is correlated with macroaggregate-associated POM in undisturbed and restored prairie soils (Jastrow, 1996). Long-term no-till treatment or restoration to natural vegetation can restore the POM concentration to native levels over decades in some environments. Fresh POM input triggers the rapid formation of macroaggregates on a time scale of several weeks. The rapid decomposition of labile POM and the accompanying soil-binding effect of microbial EPS and hyphae are responsible for the formation of macroaggregates. Scanning Electron Microscopy (SEM) images of microaggregates developed in native soils reveal that many 100–200 μm microaggregates retain the elongated shape of root residues and release visible POM cores upon ultrasonic disruption (Tisdall and Oades, 1982). Smaller ($\sim 90 \mu\text{m}$) and presumably more processed microaggregates lose their elongated shape but retain some traces of the root POM residue or contain voids that once were occupied by roots. Moreover, high-resolution X-Ray $\mu\text{-CT}$ images of stable macroaggregates ($\sim 5 \text{ mm}$) reveal root-derived POM in their cores (Kravchenko et al., 2015). Macroaggregates are typically weak and can easily break if the soil is disturbed, e.g., by tillage or slaking. In macroaggregates that continue to mellow without disturbance, POM decomposition progresses slowly and may result in stable microaggregation within the volume previously occupied by larger macroaggregates. Therefore, *the cultivation of undisturbed or native soils preferentially degrades weak macroaggregates, resulting in the mineralization of the organic binding agents and rapid loss of organic matter content* (Tisdall and Oades, 1982).

Living roots also contribute to aggregation by depositing large amounts of rhizodeposits, including short-lived root hairs, sloughed-off cells, exudates, and mucilage. Mucilage and root hairs have soil-binding characteristics, often credited with soil-binding on root surfaces. This specialized form of aggregation, armoring of living roots by mineral-mineral and mineral-root attachments, is known as the *rhizosheath* (Read and Gregory, 1997). Rhizodeposits also fuel rhizosphere microbial activity and the production of soil-binding EPS and fungal hyphae. Roots are also responsible for frequent and pronounced wetting-drying cycles in the rhizosphere. The accompanying back-and-forth motion of water menisci can mobilize and reorient clay particles resulting in the tight coating around roots (Dorioz et al., 1993). In addition, the transport of soluble organic molecules and their subsequent preferential deposition at inter-particle contact points results in a strong gluing effect in the rhizosheath.

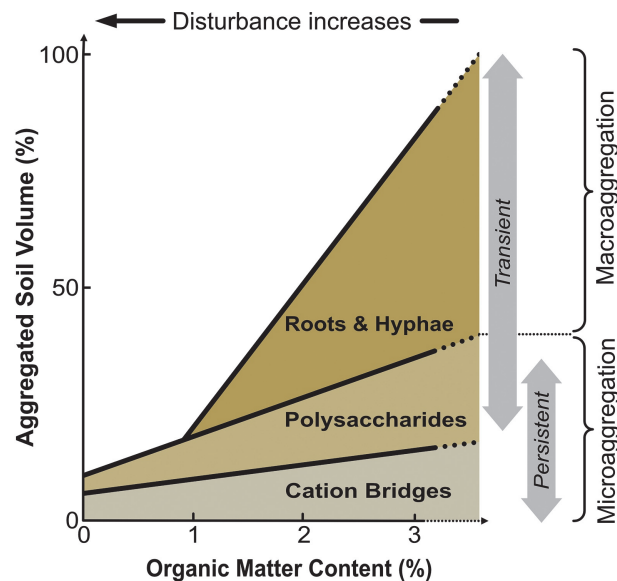


Fig. 2 Influence of organic matter content and type on formation and stability of microaggregation and macroaggregation (using data from Tisdall and Oades (1982)).

Biotic and abiotic feedback

The environmental conditions of soils are as crucial as the chemical nature of soil organic matter in regulating soil C dynamics and associated biogeochemical cycles (Schmidt et al., 2011). Soil organic matter is a continuum of materials of varying complexity, undergoing constant alteration. Within this framework, soil aggregates play an essential role in controlling the environmental conditions that in turn control organic matter transformations. Hierarchical soil aggregation creates a complex three-dimensional pore network with relatively porous inter-aggregate spacing and a tight interior. This arrangement leaves a relatively wet and oxygen-deprived core that limits the degradation of SOC. In addition, the organic molecules released by plant roots and microbes alter soil water retention characteristics. This aggregation-driven hydrodynamics results in spatial segregation of microbial communities and contrasting nutrient consumption patterns, with an overall slowdown of decomposition.

Characterization

The concept of soil aggregation implies three related characteristics (Martin et al., 1955): (a) the size and shape of individual aggregates, (b) the stability of aggregates, and (c) the configuration and arrangement of aggregates within the undisturbed soil. Aggregate size separation is used to infer the nature of the hierarchical aggregation, and aggregate stability indicates the ability of aggregates to persist under internal (e.g., rapid wetting or desiccation) and external (e.g., tillage or bioturbation) disruptive forces. Size separation of aggregates typically employs sieving in a dry or wet state and separates only those aggregates that withstand the sieving action. However, there are no straightforward methods for characterizing the morphology and size of undisturbed aggregates as they exist in their natural state (see “Imaging” section). A possible exception is the dry separation of mechanical aggregates, in which the aggregates are already pre-formed during fragmentation. Therefore, aggregate stability and aggregate size are practically considered together. Both have been adopted as quantitative indicators of soil fertility and health (Lehmann et al., 2020), and several standardized methods exist. The disruptive forces used to characterize aggregate sizes and stability are chosen to reflect forces encountered in real field conditions or to disrupt specific orders within the aggregate hierarchy (Dexter, 1988). The destruction of bonds that bind aggregates of a certain hierarchical order is accompanied by the destruction of all higher orders.

Dry aggregate size distribution

Dry aggregate stability mimics the resistance of tilled agricultural soils to wind erosion. In most cases, the airflow speed past soil aggregates is not strong enough to disintegrate aggregates (Nimmo and Perkins, 2002). However, the abrasive energy of wind-blown loose soil particles and smaller aggregates can damage and break down aggregates. Mechanical separation of aggregates using stacks of sieves or a rotary (tumbling) sieve simulates wind-driven abrasion. The resulting aggregate separates are typically smaller than the mechanical aggregates that exist on tilled soil beds. However, the fraction of soil that remains in intact aggregates after a prescribed mode of dry abrasion serves as a quantitative indicator of resistance against wind erosion. As aggregates break down to <0.8 mm in size, they are small enough to be moved by wind erosion by creeping across the surface or saltation, with further break down resulting in wind erosion of aggregates by suspension in air.

Water stable aggregation

Water stability refers to resistance to two distinct disruptive forces (Nimmo and Perkins, 2002). First, wetting of initially dry soil is prone to disintegration by the pressure of entrapped air that builds inside the engulfed pores and the swelling of clays (*slaking*). The resistance to slaking is an important indicator of soil stability during rapid wetting and resistance to raindrop splash. Second, the stability of aggregates against flowing water reflects the resistance of soils to sheet erosion and the stability of soil structure (pores and pore connectivity) during wetting and redistribution of soil water. Generally, water stability can be determined using a single sieve (e.g., 250 μm or 2000 μm) or a nest of sieves and subjecting soil aggregates to up-and-down motion while fully submerged underwater. When dry aggregates are subjected to wet sieving, the resulting stability reflects stability against both slaking and flowing water. When dealing with weak aggregates, the aggregates are prewetted under vacuum or by capillary action to eliminate the action of slaking. In addition, using a stack of sieves, one can fractionate the soil volume into silt-clay (0–53 μm), micro-aggregates (53–250 μm) and macroaggregates (>250 μm). However, the stack of sieves method does not distinguish between the separated particles of a certain order and similarly sized particles contained within stable particles of higher order.

When aggregate size classes are determined by either dry-sieving or wet-sieving, as described above, the distribution is often described in terms of arithmetic mean weight diameter (MWD) or geometric mean weight diameter (GMD). MWD implies normal distribution of soil aggregates, while GMD is associated with a log-normal distribution. The latter is consistent with most soils. They can be calculated by the equations:

$$MWD = \sum_{i=1}^n \omega_i \bar{x}_i, \text{ and} \quad (1)$$

$$GMD = \exp \left[\frac{\sum_{i=1}^n \varpi_i \log(x_i)}{\sum_{i=1}^n \varpi_i} \right], \quad (2)$$

to account for the number of sieve sizes used, n , the mean diameter of the i th sized sieve, x_i , and the portion of the initial sample weight retained on the i th sized sieve, ϖ_i after dry-sieving or wet-sieving.

Aggregate strength

Generally, aggregate strength decreases with size. Specifically, aggregates of higher hierarchical order (e.g., macroaggregates) are weaker than the lower-order aggregates contained within them. Therefore, the tensile strength of an aggregate of a certain order indicates the strength of the bonds between aggregates of lower hierarchical order (Dexter, 1988). Tensile strength is measured by crushing a single aggregate between two plates. The top plate is lowered at a constant speed while monitoring the applied force. This causes the soil aggregate to bulge outwards, creating a tensile stress at its center. The strength, TS , is calculated as a ratio of the force, F at failure to the cross-sectional area of the aggregate normal to the direction of the force. As soil aggregates have an irregular shape, the calculation of TS is not straightforward, but a generally accepted approximation is:

$$TS = 0.576 \frac{F}{\bar{d}^2}, \quad (3)$$

where $\bar{d}$ is the geometric mean diameter of the soil aggregate in the x-y, x-z and y-z directions (Dexter, 1988). Various equations have been proposed to calculate soil aggregate strength, including based on the energy of fragmentation. The statistical variability of soil aggregate strength and the change in strength with aggregate size can be used to assess the friability of soil.

Fractionation of organic matter in aggregated soils

Understanding the role of soil aggregation on organic matter persistence and turnover rate requires separating the bulk SOM into fractions that reflect its distribution within the hierarchical orders. There are several methods available (Poeplau et al., 2018). The purpose of these methods is to sequentially break the bulk soil into components that reflect different modes of organic matter storage within and between aggregates. Generally, the bulk soil is separated into different aggregate size classes by gentle shaking or sieving. Then, the free (particulate) organic matter is separated from the aggregates (soil particles) within each size by floatation. Variations of methods use water (density of 1 g/cm³) or other denser fluids. The dense aggregates separated in the previous step are further broken down by the application of ultrasonic energy or chemical dispersion using Na solution and separated into different size sub-classes. The material within each class in the previous step is further separated into free organic matter and heavy fraction by floatation. Ultimately, these methods yield free organic matter (usually assumed to be particulate) contained within each aggregate size class and organic matter associated with the mineral fraction in each size class.

Imaging

Scanning Electron Microscopy (SEM) is widely used to visualize the arrangement of soil constituents at multiple scales. SEM micrographs of aggregates acquired at appropriate magnifications are the foundations for the hierarchical soil aggregation concepts (Tisdall and Oades, 1982). However, SEM visualizes only the exposed surfaces of aggregates. In contrast, X-ray CT and thin-section microscopy provide access to the internal structure of individual aggregates. For example, CT images of large macroaggregates have revealed fragments of root residence that presumably fuel the aggregation process (Kravchenko et al., 2015). X-ray CT is the only tool with the potential to visualize aggregates and their configuration and arrangement within bulk soils. Visualizing mechanical aggregates within repacked soil columns has been successful (Aravena et al., 2011). However, visualizing hierarchical aggregates within undisturbed soils has not been successful yet (Koestel et al., 2021).

Functions and interpretations

The operational definition of soil aggregation varies with the functions of interest and does not usually coincide with the mode of formation (Or et al., 2021). For example, in agronomy, aggregation may refer to *tilth*—a favorable friable state of seedbed resulting from mechanical fragmentation by tillage (Hadas, 1997)—or the resilience of hierarchical aggregates in minimally disturbed agricultural soils (Six, 2004). In soil conservation, aggregation explains the resistance to degradation by erosion and nutrient losses. From a biogeochemical perspective, aggregates are hot spots of microbial activity and biodiversity, with the elevated nutrient cycling within the aggregate environment liberating nutrients locked up in the organic matter. However, aggregates also provide environmental protection for organic matter and associated nutrients against rapid mineralization by constraining oxygen availability and accessibility to microbial attack (Keiluweit et al., 2016). Because soil aggregation is governed by chemical, physical, and biological factors and influences multiple soil functions, it has emerged as one the key indicators of *soil health* (“the continued capacity of soil to function as a vital living ecosystem that sustains plants, animals, and humans”) (Lehmann et al., 2020).

Biogeochemical functions

Soil aggregation controls biogeochemical rates by influencing the size and connectivity of intra- and inter-aggregate pores and associated environmental variables, including wetness, aeration, and temperature (Schmidt et al., 2011). Experimental and modelling studies of soil aggregates revealed that the distribution of aerobic and anaerobic microbial activities, development of anoxic conditions, and microbial diversity are tightly coupled with the 3D flow and transport patterns within and between aggregates (Keiluweit et al., 2016).

The microaggregates formed within macroaggregates significantly reduce the turnover rate of their C by (a) physically isolating POM with extracellular polymeric substances (EPS) and clay coatings, (b) formation of anaerobic cores, (c) fostering tighter association with mineral fractions, and (d) alteration of C chemical structure (Rasse et al., 2005). Based on these observations, soil C can be conceptually partitioned into slow (micro-aggregate associated) and fast (free and macroaggregate associated) pools of C. Standardized procedures for the separation of macro- and micro-aggregates, particularly microaggregates within macroaggregates, are described in a previous section.

Therefore, depending on the unique biophysical conditions, aggregates can act as stable reservoirs of organic matter inputs (Jastrow, 1996) or biogeochemical hotspots responsible for increased greenhouse gas emissions.

Resistance to degradation

In addition to protecting soil organic matter, soil aggregation also provides resistance against the physical degradation of soils. The hierarchical model of soil aggregation also implies that smaller aggregates exclude large pores and are, therefore, denser. This porosity exclusion principle further suggests that larger aggregates are more porous and weaker because they contain large gaps between smaller and denser aggregates within them (Dexter, 1988). These ideas explain the resistance of aggregated soils to compaction and erosion.

Flow and transport

Soil aggregation has a profound effect on the size, distribution, and connectivity of soil pores. Therefore, aggregation also plays a direct role in water flow and the transport of gases, heat, and solutes. Moreover, alteration of the state of aggregation is a primary driver of alterations in flow and transport properties. Direct information about aggregate characteristics and the degree of aggregation are not directly represented in current mathematical and numerical models of flow and transport. However, the effects of aggregation are indirectly represented as bulk density, porosity, water retention characteristic, hydraulic conductivity, and aeration. For example, post-tillage fragmentation leads to the bimodality of pore size distribution, which is typically represented by dual-porosity and dual-permeability hydraulic function. In this case, the large inter-aggregate pores drain readily but act as fast conduits for infiltration and drainage when fully saturated. Inter-aggregate pores also serve as fast pathways for gas exchange. The intra-aggregate pores are usually smaller and retain moisture for more extended periods. In contrast, the effect of hierarchical aggregation of undisturbed soils on hydraulic functions is less pronounced (Koestel et al., 2021) because in situ formed aggregates do not produce large inter-aggregate pores.

Summary and outlook

Soil aggregation is a product of biological, chemical, and physical soil factors and processes that also exerts important feedback on these driving processes. Soil aggregation is recognized as a key indicator of soil health. However, several open questions remain regarding its definition and characterization. Current methods of characterization lack the ability to distinguish between aggregates formed by fragmentation from hierarchical aggregates formed primarily by bonding. These two processes are not independent of each other, with considerable interaction. Moreover, aggregates separated by sieving and other mechanical disturbance do not necessarily reflect the nature of intact aggregates as they exist in undisturbed soils. It is possible that the undisturbed soil contains soil constituents held by a continuum of binding energies. Thus, the aggregates that are the basis for much of our knowledge about soil aggregation are partially formed during the sieving process (Young et al., 2001). Recent advances in high-resolution 3D X-ray imaging hold promise in enabling in situ visualization and characterization of soil aggregation. Direct and quantitative links between the characteristics of soil aggregates and soil biological and physical functions are not fully developed yet. Although aggregation is recognized as a dynamic soil property, modelling the aggregation evolution and the resulting effects on soil functions and ecosystem models is still in its infancy.

References

- Aravena JE, et al. (2011) Effects of root-induced compaction on rhizosphere hydraulic properties - X-ray microtomography imaging and numerical simulations. *Environmental Science and Technology* 45(2): 425–431. Available at: <https://doi.org/10.1021/es102566j>.
- Cockroft B and Olsson KA (2000) Degradation of soil structure due to coalescence of aggregates in no-till, no-traffic beds in irrigated crops. *Australian Journal of Soil Research* 38(1): 61–70.

- Crow SE, et al. (2009) Earthworms, stand age, and species composition interact to influence particulate organic matter chemistry during forest succession. *Biogeochemistry* 92(1–2): 61–82. Available at: <https://doi.org/10.1007/s10533-008-9260-1>.
- Dexter AR (1988) Advances in characterization of soil structure. *Soil and Tillage Research* 11(3–4): 199–238.
- Dorioz JM, Robert M, and Chenu C (1993) The role of roots, fungi and bacteria on clay particle organization. An experimental approach. *Geoderma* 56(1–4): 179–194. Available at: [https://doi.org/10.1016/0016-7061\(93\)90109-X](https://doi.org/10.1016/0016-7061(93)90109-X).
- Gale WJ and Cambardella CA (2000) Carbon dynamics of surface residue-and root-derived organic matter under simulated no-till. *Soil Science Society of America Journal* 64(1): 190–195. Available at: <https://doi.org/10.2136/sssaj2000.641190x>.
- Ghezzehei TA and Or D (2000) Dynamics of soil aggregate coalescence governed by capillary and rheological processes. *Water Resources Research* 36(2): 367–379. Available at: <https://doi.org/10.1029/1999WR900316>.
- Hadas A (1997) Soil tillage - the desired soil structural state obtained through proper soil fragmentation and reorientation processes. *Soil and Tillage Research* 43(1–2): 7–40.
- Jastrow JD (1996) Soil aggregate formation and the accrual of particulate and mineral-associated organic matter. *Soil Biology and Biochemistry* 28(4–5): 665–676. Available at: [https://doi.org/10.1016/0038-0717\(95\)00159-X](https://doi.org/10.1016/0038-0717(95)00159-X).
- Keiluweit M, et al. (2016) Are oxygen limitations under recognized regulators of organic carbon turnover in upland soils? *Biogeochemistry* 127(2–3): 157–171. Available at: <https://doi.org/10.1007/s10533-015-0180-6>.
- Koestel J, et al. (2021) Approaches to delineate aggregates in intact soil using X-ray imaging. *Geoderma* 402: 115360 Available at: <https://doi.org/10.1016/j.geoderma.2021.115360>.
- Kravchenko AN, et al. (2015) Protection of soil carbon within macro-aggregates depends on intra-aggregate pore characteristics. *Scientific Reports* 5(October): 16261. Available at: <https://doi.org/10.1038/srep16261>.
- Lehmann J, et al. (2020) The concept and future prospects of soil health. *Nature Reviews Earth & Environment* 1(10): 544–553. Available at: <https://doi.org/10.1038/s43017-020-0080-8>.
- Martin JP, et al. (1955) Soil Aggregation. In: Norman AG (ed.) *Advances in Agronomy*, pp. 1–37. Academic Press. Available at [https://doi.org/10.1016/S0065-2113\(08\)60333-8](https://doi.org/10.1016/S0065-2113(08)60333-8).
- Nguyen C (2003) Rhizodeposition of organic C by plant: Mechanisms and controls. *Agronomie* 23: 375–396.
- Nimmo JR and Perkins KS (2002) Aggregate stability and size distribution. In: *Methods of Soil Analysis, Part 4 Physical Methods*, pp. 317–328. Madison, Wisc: Soil Science Society of America.
- Or D, Keller T, and Schlesinger WH (2021) Natural and managed soil structure: On the fragile scaffolding for soil functioning. *Soil and Tillage Research* 208: 104912 Available at: <https://doi.org/10.1016/j.still.2020.104912>.
- Pires LF, et al. (2017) Soil structure changes induced by tillage systems. *Soil and Tillage Research* 165: 66–79. Available at: <https://doi.org/10.1016/j.still.2016.07.010>.
- Poepplau C, et al. (2018) Isolating organic carbon fractions with varying turnover rates in temperate agricultural soils – A comprehensive method comparison. *Soil Biology and Biochemistry* 125: 10–26. Available at: <https://doi.org/10.1016/j.soilbio.2018.06.025>.
- Rasse DP, Rumpel C, and Dignac MF (2005) Is soil carbon mostly root carbon? Mechanisms for a specific stabilisation. *Plant and Soil* 269(1–2): 341–356. Available at: <https://doi.org/10.1007/s11104-004-0907-y>.
- Read DB and Gregory PJ (1997) Surface tension and viscosity of axenic maize and lupin root mucilages. *New Phytologist* 137(4): 623–628. Available at: <https://doi.org/10.1046/j.1469-8137.1997.00859.x>.
- Schmidt MWI, et al. (2011) Persistence of soil organic matter as an ecosystem property. *Nature* 10: 103.
- Six J (2004) A history of research on the link between (micro)aggregates, soil biota, and soil organic matter dynamics*1. *Soil and Tillage Research* 79(1): 7–31.
- Tisdall JM and Oades JM (1982) Organic-matter and water-stable aggregates in soils. *Journal of Soil Science* 33(2): 141–163.
- Totsche KU, et al. (2018) Microaggregates in soils. *Journal of Plant Nutrition and Soil Science* 181(1): 104–136. Available at: <https://doi.org/10.1002/jpln.201600451>.
- Young IM, Crawford JW, and Rappoldt C (2001) New methods and models for characterising structural heterogeneity of soil. *Soil and Tillage Research* 61(1): 33–45. Available at: [https://doi.org/10.1016/S0167-1987\(01\)00188-X](https://doi.org/10.1016/S0167-1987(01)00188-X).

Relevant websites

<https://casi.ucanr.edu>—Conservation Agriculture Systems Innovation.

<https://www.isqaper-is.eu>—Interactive Soil Quality Assessment.

<https://www.hydro.uni-jena.de/forschung/forschungsprojekte/mad-soil>—Microaggregates: Formation and turnover of the structural building blocks of soils.

<https://soilhealthnexus.org>—Soil Health Nexus.

<https://www.somfractionation.org>—Soil Organic Matter Fractionation.

<http://soilquality.org/home.html>—Soil Quality for Environmental Management.

<https://www.nrcs.usda.gov/conservation-basics/natural-resource-concerns/soils/soil-health/soil-health-assessment>—USDA NRCS Soil Health Assessment.

Porosity and pore-size distribution

John R Nimmo^a, Sheela Katuwal^b, and Maik Lucas^{c,d}, ^aU.S. Geological Survey, Menlo Park, CA, United States; ^bDepartment of Poultry Science, University of Arkansas, Fayetteville, AR, United States; ^cDepartment of Soil, Plant and Microbial Sciences, Michigan State University, East Lansing, MI, United States; ^dDOE Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of J.R. Nimmo, Porosity and Pore Size Distribution[☆], Reference Module in Earth Systems and Environmental Sciences, Elsevier, 2013, ISBN 9780124095489, <https://doi.org/10.1016/B978-0-12-409548-9.05265-9>.

Introduction	16
Porosity	17
Porosity in natural soils	17
Measurement of porosity	18
Pores and their spatial characterization	18
The nature of a pore	18
Measurement and representation of pore-size distribution	19
Based on measured geometry	19
Based on hydraulics	20
Comparison of methods	21
Typical features of a pore-size distribution	22
Relation to soil properties and processes	23
Effects of pore-space characteristics on other soil processes	23
Pore dynamics	23
Conclusions	24
References	24

Abstract

Porosity, the fraction of soil volume not occupied by solids, is relatively easy to conceptualize and measure. Pore-size distribution is a complex topic, in part from the lack of a clear and unique concept of a soil pore as a discrete object. Available tools for evaluating pore-size distribution involve traditional conventions and operational definitions applied to hydraulic property characterizations, as well as more recently developed imaging techniques. Application of these tools can provide effective characterizations of the pores and pore space that lead to an understanding of the relation between pore characteristics and soil functions.

Key points

- Porosity is a fundamental characteristic of soil that is straightforward to define and measure.
- The precise delineation of a pore and its characteristic size requires artificial, subjectively-established distinctions.
- Several methods that rely on different assumptions are available to measure the pore-size distribution of a soil.
- The quantified porosity and pore-size distribution are valuable for assessing diverse functions and properties of soil.

Introduction

The pore space is that portion of the soil's volume that is not occupied by or isolated by solid material. Its basic character affects and is affected by critical aspects of almost everything that occurs in the soil: movement of water, air, and other fluids; transport and reaction of chemicals; and residence of roots and other biota. By convention the definition of pore space excludes fluid pockets that are totally enclosed within solid material—vesicles or vugs, for example, that have no exchange with the pore space that has continuity to the boundaries of the medium. Thus we consider a single, contiguous pore space within the body of soil. In general, the pore space has fluid pathways that are tortuous, variably constricted, and highly connected. Fig. 1 shows example images of soil pore space at three different scales.

The pore space is often considered in terms of individual pores—an artificial concept that enables quantifications of its essential character. Traditionally in soil science and hydrology, pores have been conceptualized, measured, and applied with respect to the fluids that occupy and move within the pore space. Recent advances in imaging techniques have facilitated additional types of characterization in terms of observable geometry and topology.

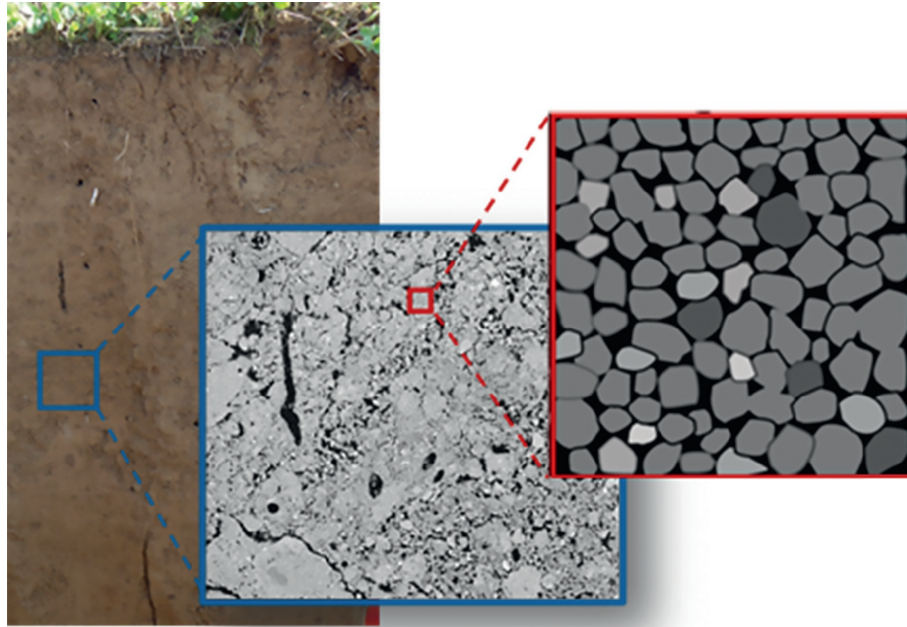


Fig. 1 View of a soil profile (loess soil) from an agricultural field at the Garzweiler mining area, Germany. The soil has pores of various sizes, shapes, and modes of origin. The image within blue rectangle is an X-ray computed tomography image of a slice of soil column obtained from the profile, in which the dark areas represent pores and the gray areas represent the soil matrix. At the resolution of $19\ \mu\text{m}$, only structural pores are visible in the image. Within the soil are also textural pores resulting from the packing of soil particles which is illustrated by the drawing of individual soil grains within the red rectangle. The gray areas represent soil particles and the dark area pores between the soil primary particles. Photo: *M. Lucas*.

Imaging techniques, however, cannot illustrate the total extent of the pore space, for example because the high resolution needed for the smallest pores can only be achieved for small samples, which do not contain large pores. Two-dimensional images have a further limitation in that the connectivity of the pores can only be incompletely shown.

Porosity

Porosity ϕ is the fraction of the total soil volume that is taken up by the pore space. Thus it is a single-value quantification of the amount of space available to fluid within a specific body of soil. Being simply a fraction of total volume, ϕ can range between 0 and 1, and typically falls between 0.3 and 0.7 for soils. With the assumption that soil can be considered as a continuum, adopted here as in much of soil science literature, porosity can be considered a function of position.

Porosity in natural soils

The porosity of a soil depends on several factors, including (1) degree of compaction, (2) breadth of the particle size distribution (polydisperse vs. monodisperse), (3) shape and size of particles, (4) cementing, and (5) aggregation. Mathematically considering an idealized medium of packed uniform spheres, ϕ must fall between 0.26 and 0.48, depending on the packing. Spheres randomly thrown together will have ϕ near the middle of this range, typically 0.30 to 0.35. A sand with grains nearly uniform in size (monodisperse) packs to about the same porosity as spheres. In a polydisperse sand, the fitting of small grains within the pores between large grains can reduce ϕ , conceivably below the 0.26 uniform-sphere minimum. The particular arrangement required to reduce ϕ to 0.26 or less is highly improbable, however, and ϕ usually exceeds 0.30 in natural soils. Particles more irregular in shape tend to have larger gaps between their nontouching surfaces, thus forming media of greater porosity. In porous rock such as sandstone, cementation or welding of particles not only creates pores that are different in shape from those of particulate media, but also reduces the porosity as the cementing material takes up space that would otherwise be pore space. Porosity in such a case can easily be less than 0.3, even approaching 0. Cementation can also have the opposite effect. In many soils, clay and organic substances cement particles together into compacted units surrounding large pores that are mechanically stable because of the rigidity of the cemented matrix. Pure matrix material might have a 0.35 porosity within it, but the medium as a whole has additional pore space such that ϕ of the bulk soil can be 0.5 or greater. Observed porosities can be as great as 0.8 to 0.9 in certain types of soils like peats (extremely high organic matter content), Andisols (volcanic ash soils), and ice-rich permafrost.

Porosity is often conceptually partitioned into two components, most commonly called textural and structural porosity. Fig. 1 shows examples of these. The textural component is the value the porosity would have if the arrangement of the particles were random, as described above for granular material without cementing. That is, the textural porosity might be about 0.3 in a granular medium. The structural component represents nonrandom structural influences, including macropores, and is arithmetically defined as the difference between the textural porosity and the total porosity.

The texture of the medium relates in a general but complex way to the pore-size distribution, as large particles may give rise to large pores between them, and small particles may give rise to small pores, or to additional large pores if their clay mineralogy causes aggregation. Additionally, the structure of the medium, especially the pervasiveness of features like aggregation, shrinkage cracks, and wormholes, substantially influences hydraulic properties.

Measurement of porosity

Porosity is the volume of pores divided by the total soil volume, which can be calculated using the difference between total sample volume and total particle volume. Usually this is not done directly but with measurements of bulk density ρ_b and particle density ρ_p . From their definitions as total particle mass divided by total volume (ρ_b) or by particle volume (ρ_p), the ratio ρ_b/ρ_p is the complement of ϕ , so that

$$\phi = 1 - \frac{\rho_b}{\rho_p} \quad (1)$$

For a soil of unknown ρ_p , a value of 2.65 g/cm³ (corresponding to quartz) is frequently used. Often the main source of error arises from uncertainty in the total soil volume used to determine ρ_b . If a minimally disturbed sample of soil of known shape (e.g. cylindrical) can be obtained, its volume can be calculated from its dimensions. Significant error can result from irregularities in the actual shape and from compaction during the sampling process. Alternatively, the measured volume can be that of the excavation from which a bulk soil sample has been extracted. Various methods such as the balloon and sand-fill methods, as well as by filling the lined hole with a known liquid such as water, can indicate the total volume.

Another way to determine porosity is from the volume of water contained in a saturated sample of known volume. The mass of saturated material less the oven-dry mass of the solids, divided by the density of water, gives the volume of water. Dividing this volume by the original sample volume gives the porosity. An analogous method is to determine the volume of gas in the pore space of a completely dry sample. Sampling and drying of the soil must be conducted in a way that does not compress the soil or otherwise alter its porosity. A pycnometer can measure the air volume in the pore space. It works with a gas-tight chamber enclosing the sample so that the internal gas-occupied volume can be perturbed by a known amount while the gas pressure is measured. Boyle's law indicates the total gas volume from the change in pressure resulting from the volume change. This total gas volume minus the volume of air-filled portions of the apparatus apart from the soil yields the total pore volume to be divided by the sample volume.

Tomographic imaging or the older technology of thin sections can produce a visualization from which porosity can be calculated. With a three-dimensional image the volume of pore space divided by total volume gives the volumetric porosity. With a two-dimensional image an analogous procedure can be followed using areas, giving the areal porosity over the imaged plane, or using segment lengths along a line through the sample to yield a linear porosity. If the medium is isotropic, these would numerically equal the volumetric porosity. Imaging tends to underestimate the total porosity, as the smallest pores cannot be resolved.

Pores and their spatial characterization

The nature of a pore

Because soil does not contain discrete objects with obvious boundaries that could be called individual pores, the precise delineation of a pore requires artificial, subjectively established distinctions. This contrasts with soil primary particles, which are easily defined, being discrete material objects with obvious boundaries. The criterion selected to partition pore space into individual pores is often not explicitly stated when pores or their sizes are discussed. Because of this inherent arbitrariness, some scientists argue that the concepts of pore and pore size should be avoided. Much valuable theory of the behavior of the soil-water-air system, however, has been built on these concepts, defined using criteria that are accepted widely, if not universally.

Pores can also be classified according to their origin, function, or other attributes. A textural/structural distinction is possible, as it is for porosity. Inter-granular pores are the major portion of soil textural porosity, as discussed above. Intra-granular or dead-end pores (if not entirely enclosed within solid) might empty or fill with water, but without contributing as directly to fluid movement through the medium. Intra-aggregate pores may be essentially equivalent to inter-granular pores within an aggregate. Inter-aggregate pores, including shrinkage cracks, are common types of macropores, in addition to biogenic pores, for example the channels left by decayed roots and tunnels made by burrowing animals.

Pore-size characterizations rely heavily on a conventional idealization of pores as being cylindrically symmetric, which is far from true, given the irregular shapes of grains and pores in real soils. For realistic pore shapes the effective width varies along the axis of flow. Relatively wide portions are often called pore *bodies* and the relatively narrow portions that separate the pore bodies called

pore *necks*. Other anatomical metaphors are sometimes used, the wide part of a pore being the “belly” or “waist”, and the constrictive part being the “throat”. In a medium dominated by textural pore space, like a sand, pore bodies are inter-granular spaces of dimensions typically slightly less than those of the adjacent particles. At another extreme, a long wormhole, especially if its diameter is essentially uniform along its length, might be considered a single pore. The boundaries of such an elongated pore are of three types: (1) interface with solid, (2) constriction—a plane through the locally narrowest portion of pore space, or (3) interface with another pore (e.g. a crack or wormhole) or a hydraulically distinct region of space (e.g. the land surface).

Pore sizes are often specified by an effective radius, r , of the pore body or neck. Alternative indicators of size include the cross-sectional area or the volume of a pore, and the hydraulic radius, defined as the ratio of the cross-sectional area to circumference, or of pore volume to specific surface.

Methods for measuring or estimating pore size fall mainly into two categories, those based on imaging and those based on hydraulic behavior. In the usual case where it is more important to characterize a medium than an individual pore, these are applied to measure many pores so as to obtain a pore-size distribution, as discussed further in the next section.

Measurement and representation of pore-size distribution

The pore-size distribution is the relative abundance of each pore size in a representative volume of soil. It can be represented with a function $f(r)$, which has a value proportional to the combined volume of all pores whose effective radius is within an infinitesimal range centered on r . Fig. 2 shows examples. Like porosity, $f(r)$ may be taken to comprise textural and structural components.

Based on measured geometry

The most obvious and straightforward measurements of pore size are with geometric analysis of images. X-ray computed tomography (CT) enables 3D imaging of the interior features within a soil. Access to CT scanners has increased dramatically since the 1990s (Vogel et al., 2022), and this 3D imaging technique has largely replaced the pore structural analysis of 2D thin sections.

From two- or three-dimensional images, bodies and necks can be measured by image analysis (Lindquist et al., 2000). In three dimensions, an often-used technique is to determine the maximum inscribed sphere in each pore, i.e. the diameter of the largest sphere that fits inside the pore (Hildebrand and Rügsegger, 1997). Image-based techniques, however, come with a strong trade-off between sample size and resolution; the small samples needed for high resolution do not capture large macropores. This trade-off also affects topological measures such as pore connectivity. The effect can be minimized by combining different sample sizes (Lucas et al., 2021; Vogel et al., 2010). In two-dimensional image analysis, the plane of the image does not always intersect the widest parts of the pore bodies and necks, so mathematical correction techniques are necessary to estimate unbiased pore body and neck sizes. In addition, information about topological features like pore connectivity is lost.

Three-dimensional analysis of geometry is also possible with impregnation techniques. In these, the soil pore space is filled with a resin or other liquid that solidifies. After solidification, the medium is broken up and individual blobs of solid resin, actually casts of the pores, are analyzed as particles. Incremental sectioning and imaging of the 3D impregnated block can also be used to reconstruct the 3D pore space, albeit with considerable effort.

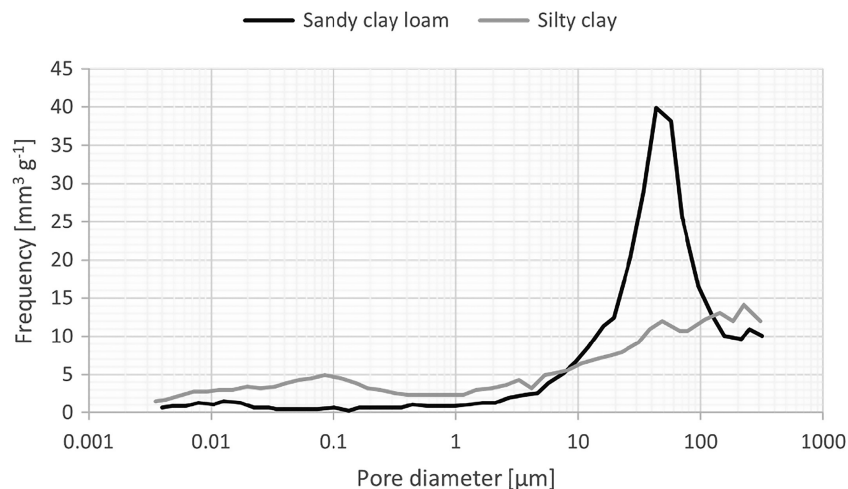


Fig. 2 Pore-size distributions of a sandy clay loam and a silty clay soil, measured by mercury intrusion. Adapted from Calderón MJ, Real M, Cabrera A, Koskinen WC, Cornejo J, and Hermosín MC (2015) Influence of olive oil mill waste amendment on fate of oxyfluorfen in southern Spain soils. *CLEAN – Soil, Air, Water* 43(7): 1107–1113. doi:10.1002/clen.201400560.

Based on hydraulics

More commonly used than imaging methods are those based on effective capillary size inferred from measured hydraulic properties. These indirectly measure pore sizes from data concerning fluid behavior in the unsaturated medium, usually the emptying or filling of pores during soil drying or wetting (Fig. 3). Most commonly this is done using a water retention curve $\theta(\psi)$, where θ is the volumetric water content and ψ is the matric pressure. The relation between the measured data and specific quantifications relies on a process-based interpretation of measurement techniques.

Because large pores fill or empty at ψ near 0, a medium that has many large pores will have a retention curve that drops rapidly to low θ as ψ decreases from 0. Conversely, one with very fine pores will retain much water even at highly negative ψ , thus having a retention curve with more gradual changes in slope. Quantified using capillary theory, the effective pore radius corresponds to the ψ at which a pore empties or fills according to

$$r = \frac{-2\sigma \cos \alpha}{\psi} \quad (2)$$

where σ is the surface tension and α is the contact angle. This formula can convert a measured $\theta(\psi)$ into an equivalent $\theta(r)$ curve. This curve is actually a cumulative pore-size distribution; the water content on a drying $\theta(r)$ indicates the combined volume of all pores with neck radius less than r . Applying the fundamental theorem of calculus, the pore-size distribution is simply the derivative. This function is the basic equivalent-capillary representation of pore-size distribution.

$$f(r) = \frac{d\theta}{dr} \quad (3)$$

This method has a serious limitation in media subject to pore blockage during emptying or filling (Mualem and Dagan, 1975). On a drying curve, for example, there can be pores that remain filled even when ψ is at a value that should have them empty according to Eq. (2), if they are completely surrounded by empty pores such that there is no path to accept water exiting from pores that are isolated within. The consequence is that the abundance of pores of larger size will be underestimated when inferred from the drying water retention curve. The reverse holds true for a wetting curve.

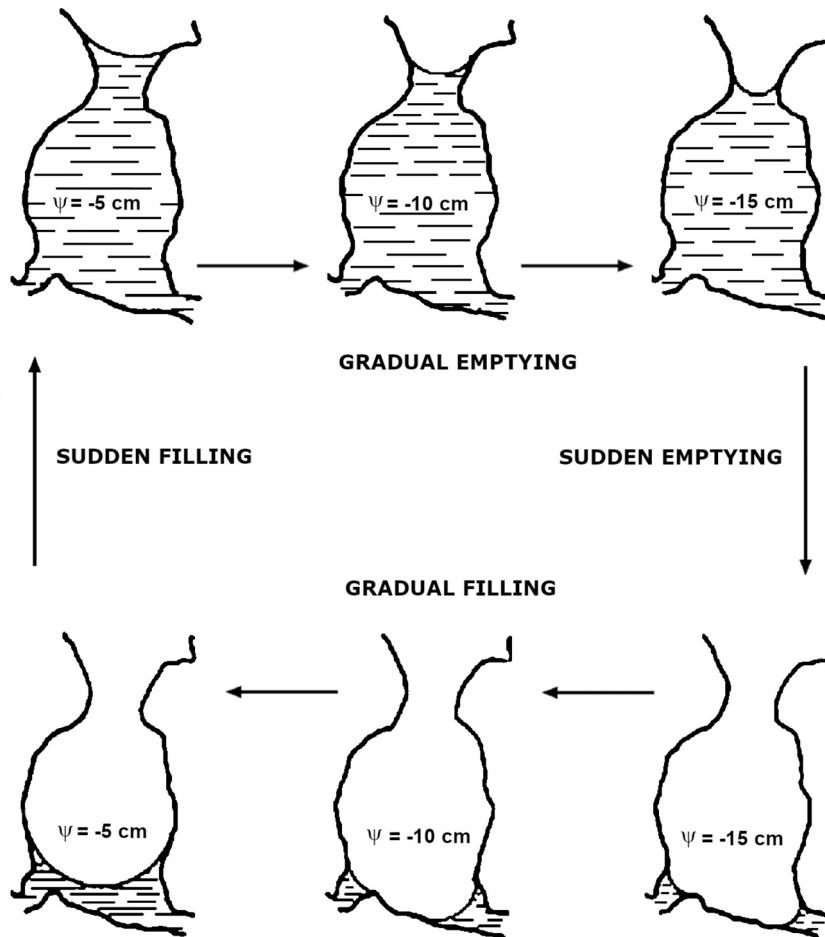


Fig. 3 Dynamics of a Haines jump. Adapted from Miller EE and Miller RD (1956) Physical theory for capillary flow phenomena. *Journal of Applied Physics* 27(4): 324–332.

Mercury porosimetry is analogous to the water-retention based method, but uses air instead of water as the wetting fluid and mercury instead of air as the nonwetting fluid. Mercury is forced into the pores of dry soil incrementally, and the relation between mercury content and mercury pressure recorded. Applying Eq. (2) with the appropriate values of surface tension and contact angle for mercury gives the estimated pore-size distribution.

The effective radius can be directly related to the radius of curvature of the air-water interface at the occurrence of Haines jumps, a basic phenomenon of capillary hysteresis, illustrated in Fig. 3. The pore necks, which control the value of ψ at which pores empty, are smaller than the pore bodies, which control the ψ at which pores fill. As the medium dries and ψ decreases, water retreats gradually as the air-water interface becomes more curved. At the narrowest part of the pore neck, this interface can no longer increase curvature by gradual amounts while satisfying the capillary relation (2), so it retreats suddenly to narrower channels elsewhere. An analogous phenomenon occurs during wetting, when the decreasing interface curvature cannot be supported by the radius of the pore at its maximum width. The volume that empties and fills in this way is essentially that of an individual pore, though not all pore space is subject to Haines jumps—water remains in crevices and in films (not shown in Fig. 3) coating solid surfaces. Various models and theories treat this space in different ways. By the definition above it is part of a pore in addition to the volume affected by the Haines jump.

The Haines jump concept can be used to give separate size distributions for pore necks and pore bodies. Applying Eq. (3) to data from a drying retention curve indicates neck radii; applied to wetting curve data, it indicates body radii.

Comparison of methods

The different measurement techniques of soil pore-size distribution do not give exactly the same results, as they are based on different concepts and biased by different characteristics of soil. Fig. 4 shows a diagram noting the ranges of validity of selected

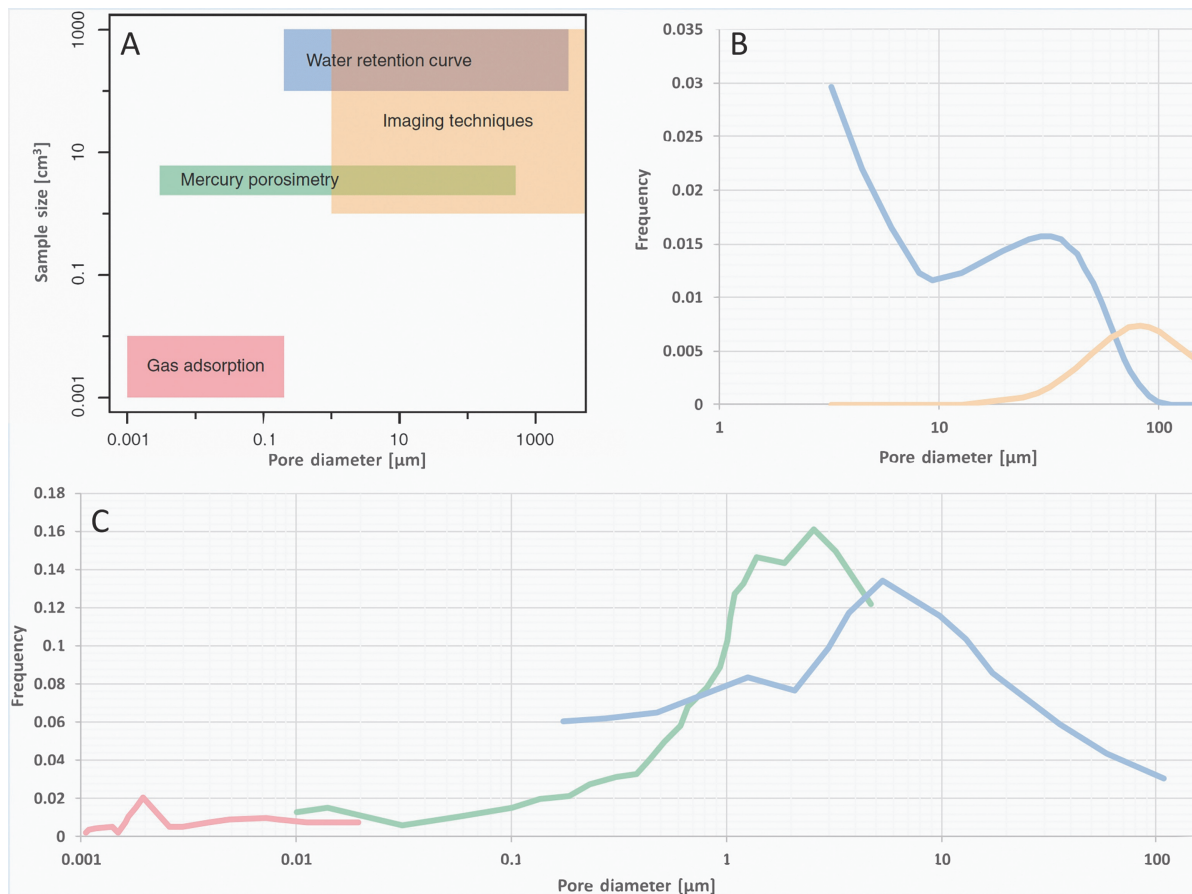


Fig. 4 Pore-size distributions obtained by different methods. (A) Comparison of the range of accessible sample size and pore size for the different methods. Both axes are represented with a logarithmic scale. (B) Pore-size distribution of a loamy sand soil derived from water retention (measured by the evaporation method) (blue) and computed microtomography (light brown). (C) Pore-size distribution of loess soil derived by water retention, mercury intrusion porosimetry, and gas adsorption. Curve colors in (B) and (C) correspond to those of the labeled rectangles in (A). (A) Adapted from Rabot E, Wiesmeier M, Schlüter S, and Vogel HJ (2018) Soil structure as an indicator of soil functions: A review. *Geoderma* 314: 122–137. doi:10.1016/j.geoderma.2017.11.009. (B) Adapted from Geistlinger H and Leuther F (2018) Evaporation study for real soils based on HYPROP hydraulic functions and micro-CT-measured pore-size distribution. *Vadose Zone Journal* 17(1): 180041, doi:10.2136/vzj2018.02.0041. (C) Adapted from Hajnos M, Lipiec J, Świeboda R, Sokołowska Z, and Witkowska-Walczak B (2006) Complete characterization of pore-size distribution of tilled and orchard soil using water retention curve, mercury porosimetry, nitrogen adsorption, and water desorption methods. *Geoderma* 135: 307–314. doi:10.1016/j.geoderma.2006.01.010.

methods, and examples of substantially different results from different methods. Results often differ even where the methods overlap in pore-size range. Different methods are affected differently by unintended influences. For methods based on retention or intrusion, emptying and filling depend on more than capillarity; contact angles, for example, are dynamic and are unlikely to equal handbook values, so they may deviate significantly for water and for mercury. The swelling of clays can be a major influence on a water retention curve, though negligible for retention measured with mercury. Imaging techniques are subject to some entirely different influences such as resolution and image quality. Another influence is the sample size typically used for the specific method (Fig. 4A). Macropores, for example, may be representatively captured in large soil samples used in imaging, but are not captured in the relatively small sample sizes used for gas adsorption. In Fig. 4B, the pore size distribution derived from a water retention curve (blue) indicates a greater abundance of smaller pores than that based on CT measurements (light brown), in part because the drying curve used for this purpose gives a measure of pore neck rather than body sizes, and also because it relies on dynamic accessibility of pores to the incoming fluid. In addition, for distributions measured by hydraulic means such as water retention, caution is needed in interpretation at the extremes, as the driest and wettest portions of a retention curve are determined by factors other than capillarity. In soil environmental applications, the water retention method is most often used, in large part because these data are more commonly available. One can also expect a water-based method to give better results for a water-based application such as predicting hydraulic conductivity.

Typical features of a pore-size distribution

Fig. 5 illustrates how $f(r)$ can be apportioned into different classifications of pore type. Corresponding graphs of the cumulative size distribution can be used equivalently.

Often a specific functional form or other representational model of pore-size distribution is useful. A normal or lognormal distribution can be fit to the data, for example. Self-similar or fractal forms of the distribution function also can be used, giving a power-law form of the cumulative pore-size distribution. If the actual distribution is unimodal, these functional forms can work well by themselves. Structural features, however, may give the soil a bimodal pore-size distribution, leading to several distinctive effects on water flow. Some soils may have three or more peaks. Multimodal distributions can be represented by superpositions of the normal or lognormal forms.

The minimum conceivable size of a soil pore, given that it must contain at least a few molecules, is about 1 nm. Fluid behavior in such small pores would be dominated by interaction with the solid material, and not likely describable by capillary laws or even standard thermodynamics. The smallest pores measured by water retention or mercury intrusion are typically 50 or 100 times larger than this, though with greater cost and effort, greater magnitudes of pressure and hence smaller measured r can be achieved. Because r relates inversely to ψ , the dry portion of the $\theta(r)$ retention curve, and correspondingly of $f(r)$, is confined to a small region near $r = 0$, where the apparent number of pores becomes large. In an extreme case in which θ is held to remain artificially finite as ψ goes to negative infinity, at $r = 0$ $f(r)$ approaches infinite height in the form of a delta function.

At either pore size extreme, where capillary phenomena lose dominance, Eq. (2) no longer applies, though it still can give r values corresponding to ψ . These values are purely artificial in character, having no relation to the actual radii of capillaries, though they may be useful for translating one property to another (discussed below). At the dry extreme, water is not likely to be in filled pores but rather in thin films that coat particles of soil.

Peaks in $f(r)$ are apparent if they fall within the r -measurement range of the method (see Fig. 4C). Such a peak corresponds to an inflection point in $\theta(r)$. The inverse relation of r to ψ means that $\theta(\psi)$ is more likely to show an inflection point than $\theta(r)$; if $\theta(r)$ has an inflection point, then $\theta(\psi)$ is mathematically required to have an inflection point, but the converse is not true. The case of a retention curve that shows a wide range of slopes (typically with an inflection point within the measured range and a distinct initial-Haines-jump or air-entry effect at a nonzero value of ψ) will have a defined peak in the pore-size distribution if the middle portion

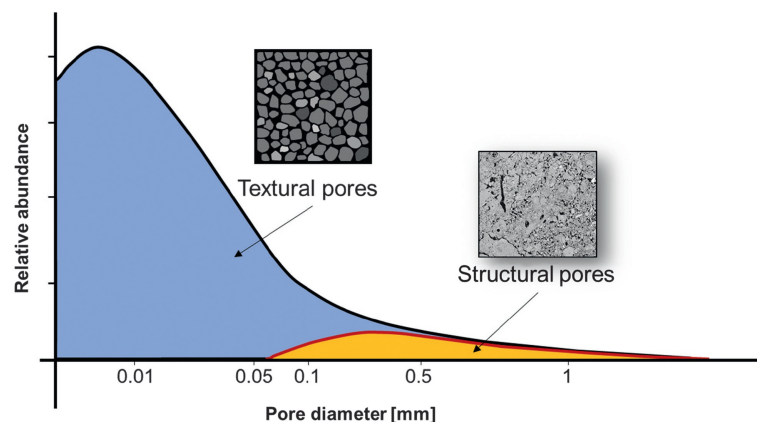


Fig. 5 Pore-size distribution with subregions distinguished for different types of pores on the basis of pore body and neck radii and drying/wetting history. Adapted from Lucas, M., Vetterlein, D., Vogel, H.-J., & Schlüter, S. (2021). Revealing pore connectivity across scales and resolutions with X-ray CT. *European Journal of Soil Science* 72(2): 546–560. doi:10.1111/ejss.12961.

of the retention curve has a slope that differs markedly from that of the two end segments. This is frequently true for monodisperse media and for repacked samples, in which pores larger than a certain r have been destroyed. For many soils, perhaps most, $f(r)$ does not have a peak except at $r = 0$, or so close to 0 that it cannot be measured. This is especially likely for media that are macroporous, polydisperse, or that show an air-entry effect with ψ very close to 0. Retention curves of this type may be best represented by a fractal or power-law model. Lognormal models or other representations that have a defined peak can often still be applied but with the peak outside the measurable range. Predictive models of hydraulic properties such as hysteresis and conductivity may be unaffected by the lack of a measurable peak even if they are applied using a pore-size distribution that resembles some form of normal distribution.

Relation to soil properties and processes

A major importance of a soil's porosity and pore-size distribution is that they relate to other soil properties in a complex and useful way; they can be interpreted to reveal important insights and used to predict other properties of soils.

Effects of pore-space characteristics on other soil processes

One obvious significance of porosity is that it is an upper limit of volumetric water content. It is similarly essential to the definition of degree of saturation, θ divided by porosity. Another significance is that within the pore space, the complement to θ is gas content. That is, volumetric gas content equals the difference between θ and porosity.

The magnitude of porosity roughly indicates complexity of structure, being greater for greater complexity, as noted above for aggregated soil. Additionally, the spatial variability of porosity is important, and also correlates with greater structural complexity. For pore-size distribution the most obvious connection is the relation to the water retention curve as elaborated above. In addition, if both wetting and drying water retention curves are known, their difference reveals the extent of hysteresis and thus gives information about both pore-neck and pore-body distributions. Conversely, if both neck and body sizes are known from image analysis, the magnitude of soil-water hysteresis can be estimated.

Because pores are fluid conduits, their size distribution is useful for predicting hydraulic conductivity K . The usual assumption applied is that the pores act hydraulically as if they are cylinders of the same effective radius, providing a simple geometry for applying hydrodynamic formulas (e.g. Childs and Collis-George, 1950). Popular models include those of Mualem (1976) and Burdine (1953), which have been analytically combined with widely used empirical formulas for retention curves. Gas and other types of fluid transport can be treated similarly.

For solute transport, the pore-size distribution affects solute convection similarly to hydraulic conductivity. Additionally, it affects solute dispersion, which is expected to be greater for a broader pore-size distribution. The pore-size distribution affects the sorption of solutes in a complex way: smaller pores are associated with longer residence times but most solutes may bypass them and pass quickly through the large pores with minimal opportunity to react. Interchange between fast- and slow-transporting portions of the pore space is a vital and much-investigated aspect of solute transport in soils. Smaller pores are associated with greater relative surface area, and thus may correlate with a greater capacity for sorption.

For particle transport, many aspects are essentially the same as for K and solute transport. Additionally, the phenomenon of straining depends critically on the proportion of pores smaller than a given particle size. This is the dominant factor in some particle-transport applications.

The large-radius portion of the pore-size distribution may be indicative of macropore space and thus relate to the processes of macropore flow and aeration. In addition, increased volume of macropores may be linked to their mode of origin (e.g., earthworm channels or root holes). It is at the same time an important habitat consideration for soil fauna, and may also be important for root growth in soil of high mechanical impedance. Similarly, at the other extreme, pores in the tens of microns size range may represent optimal microbial habitats, leading to high enzyme activity (Kravchenko et al., 2019).

The pore-size distribution is a critical aspect of the overall structure of a soil, the spatial arrangement of pores and solids. More about soil structure can be found in the soil structure article. To link pore structure to its functions, however, the pore space needs to be described by more than just its pore-size distribution and porosity. Additional important inputs include various topological characteristics of pore-space images, such as tortuosity or anisotropy. A particular concept of interest is the Minkowski functionals. In three dimensions these geometric measures include pore volume, surface area, and mean curvature, as well as connectivity described by the Euler number. The determination of the Minkowski functionals as a function of pore size can be used to describe the pore space fully by a small set of parameters and thus link pore structure to physical processes (Vogel et al., 2010).

Pore dynamics

Natural and human-induced soil processes create and destroy pores, and induce changes in their size and other attributes. Table 1 lists some of the common effects on pore-size distribution caused by routine processes. In general, if the soil is compacted by uniform stress, such as a weight imposed at the land surface, it normally loses large pores and gains small pores. Disturbance from irregular stresses, as during digging or repacking, has a variety of effects on pore size, often with the effect of a net decrease in the number of large pores and an increase in the number of small pores. Though small inter-granular pores are seldom closed

Table 1 Possible effects of routine soil processes on pore-size distribution.

Soil process	Effect	Further reference
Shrinkage	Enlarged macropores New macropores created Within an aggregate, intra-aggregate pores decrease in size, or increase in size if clay particles shrink	Swelling and shrinking Freezing and thawing processes at field scale
Swelling or frost-heaving	Decrease in the size of macropores Closed macropores Within an aggregate, intra-aggregate pores increase in size, or decrease in size if clay particles expand	Freezing and thawing cycles
Mechanical compression	Decrease in size of macropores Closed macropores Aggregates break up, reducing the number of intra-aggregate pores, thereby reducing the proportion of the smallest pores	Compaction
Disturbance from digging or plowing	Existing macropores destroyed Interclod macropores created Aggregates break up, reducing the number of intra-aggregate pores, also the fraction of the pore space represented by the smallest pores	Tilth
Biological activity	New macropores created, often with concurrent reduction of smaller pores, e.g., by compression of soil surrounding a growing root Enlarged macropores, as by ongoing traffic of ants or burrowing mammals Constricted or obstructed pores, e.g., through growth of microorganisms Increased aggregation, promoting the creation of inter-aggregate macropores, and possibly smaller inter-granular pores within aggregates	Root physical constraints Bioturbation; Siul Ruiz Reinforcement by roots
Chemical activity	Constricted or obstructed pores through formation of precipitates Enlarged pores through dissolution of precipitates Increased or decreased interparticle cohesion, with complex effects on pore size and structure	Salinity, Physical effects

completely, some processes can close macropores, in effect destroying them. Several types of processes can create pores. Roots, for example, can create a dense system of biopores whose tubular form can connect over long distances.

Conclusions

The characterization of pore space is a vital and fruitful aspect of soil investigation. Physical, chemical, and biological influences acting on the liquid, solid and gas constituents of the soil determine the form and development of pores, whose character in turn profoundly influences the nature and behavior of the soil. It relates to all major soil functions, including water storage and flow, carbon storage and organic matter turnover, as well as plant growth and biodiversity.

Soil porosity is fairly well standardized in definition and measurement techniques. Pore size, however, is not obvious how to define, and consequently has no unique standard of measurement. It is difficult to consider pores on an individual basis without talking about interactions with water and other materials. Yet pore-size concepts are central to topics like macropores, aggregation, fractures, soil matrix, and solute mobility. Pore size plays a key role in various proposed means of quantifying soil structure. It also has a major practical role in the prediction of hydraulic properties. New pore size concepts, measurement techniques, and relations to a great diversity of soil properties and processes are likely to remain a major emphasis in the study of soil.

References

- Burdine NT (1953) Relative permeability calculations from pore size distribution data. *Transactions of AIME* 198: 71–77.
- Childs EC and Collis-George N (1950) The permeability of porous materials. *Proceedings Royal Society London Ser. A* 201: 392–405.
- Hildebrand T and Rügsegger P (1997) A new method for the model-independent assessment of thickness in three-dimensional images. *Journal of Microscopy* 185(1): 67–75. <https://doi.org/10.1046/j.1365-2818.1997.1340694.x>.
- Kravchenko AN, Guber AK, Razavi BS, Koestel J, Blagodat'skaya EV, and Kuzyakov Y (2019) Spatial patterns of extracellular enzymes: Combining X-ray computed micro-tomography and 2D zymography. *Soil Biology and Biochemistry* 135: 411–419. <https://doi.org/10.1016/j.soilbio.2019.06.002>.
- Lindquist WB, Venkatarangan A, Dunsmuir J, and Wong T-F (2000) Pore and throat size distributions measured from synchrotron X-ray tomographic images of Fontainebleau sandstones. *Journal of Geophysical Research - Solid Earth* 105(B9): 21509–21527. <https://doi.org/10.1029/2000JB900208>.
- Lucas M, Vetterlein D, Vogel H-J, and Schlüter S (2021) Revealing pore connectivity across scales and resolutions with X-ray CT. *European Journal of Soil Science* 72(2): 546–560. <https://doi.org/10.1111/ejss.12961>.
- Mualem Y (1976) A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research* 12(3): 513–522.
- Mualem Y and Dagan G (1975) A dependent domain model of capillary hysteresis. *Water Resources Research* 11(3): 452–460. <https://doi.org/10.1029/WR011i003p00452>.
- Vogel HJ, Weller U, and Schlüter S (2010) Quantification of soil structure based on Minkowski functions. *Computers & Geosciences* 36(10): 1236–1245. <https://doi.org/10.1016/j.cageo.2010.03.007>.
- Vogel H-J, Balseiro-Romero M, Kravchenko A, Otten W, Pot V, Schlüter S, Weller U, and Bayevy PC (2022) A holistic perspective on soil architecture is needed as a key to soil functions. *European Journal of Soil Science* 73(1): e13152. <https://doi.org/10.1111/ejss.13152>.

Pore-scale modelling of flow and transport phenomena in soils

Kirill M Gerke and Marina V Karsanina, Schmidt Institute of Physics of the Earth of Russian Academy of Sciences, Moscow, Russia

© 2023 Elsevier Ltd. All rights reserved.

Introduction	25
Input data and pore space characterization	26
Numerical methods	28
Examples of the application of pore-scale modelling	30
Existing codes	30
Homogenization, upscaling and beyond	31
Summary and outlook	33
Acknowledgments	33
References	33

Abstract

Soil physical and other properties are usually studied in the laboratory or field to parameterize continuum-scale models. Such measurements may result in numerous inaccuracies or are based on false assumptions. With recent development in soil structure imaging and characterization techniques a detailed 3D information on the inner structure of a soil sample can be acquired. With such an input data pore-scale modelling approach is indispensable in simulating the processes taking part within the pore (or solid) space. Modelling approach allows to obtain such properties on any scale required and upscale them to continuum-scale. In this chapter we briefly describe theoretical and practical basics to use pore-scale modelling in soil science applications limiting our focus to flow and transport properties.

Key points

- Pore-scale modelling is an emerging technique that is capable of simulating soil processes on different scales where they are inaccessible by direct measurements
- With a variety of available free and commercial tools any soil scientists may incorporate pore-scale modelling into research
- Continuum-scale models will benefit from parameterization from or coupling with pore-scale models

Introduction

Soil physical properties such as hydraulic conductance (Ksat chapter reference), water retention curve and relative permeabilities are highly interconnected to all other soil functions, as they describe water and solute transport within soil. This is hard to overestimate the importance of these properties, as they affect moisture and nutrients availability to plants, (micro)organisms — all, in turn, determinate soil fertility. As such, physical soil properties are the fundamental building blocks of all major models aimed at describing soil processes, e. g., those including the Darcy and Richardson-Richards equations. What usually lacks the necessary attention is that all such properties are at the continuum-scale — this means they represent some “averaged” value attributed to a volume of soil. This averaging step, or homogenization (Fig. 1), is based on pore-scale information such as pressure and flow velocities fields. This is the information that can be obtained with the help of pore-scale modelling approaches.

Pore-scale modelling operates at one level below the continuum-scale and explicitly describes fluid interfaces and velocities. Unlike the continuum-scale, which does not know anything about the soil except for the homogenized property, pore-scale describes the so-called structure-property relationship — how fluids (solutes and air) interact with soil solids (e. g., mineral, clay and organic components and their mixtures). The interaction parameters, such as interfacial tensions, velocity slip or contact angles, are themselves homogenized values that can be measured either experimentally (but this is problematic for some properties) or modelled with the help of molecular dynamics (MD). The MD simulations describe the motion and interaction of individual molecules and, due to computational expense, allow only a limited modelling domain — e. g., a single nano-pore or fluid droplet on the substrate. A density-function theory (DFT), if properly parameterized and combined with MD (Vaganova et al., 2022), is another efficient way to assess pore-scale characteristics. Hence, pore-scale modelling is sandwiched in-between MD or DFT at the molecular scale and the classical continuum-scale and a part of the general multi-scale physical framework to describe flow and transport processes within porous media.

In soil physics and hydrology pore-scale simulations provide a novel perspective on understanding numerous aspects of the flow and allow the fixing of numerous flaws of the experimental techniques:

- 1) Unsaturated hydraulic conductances (two-phase flow soil experiments) are hard to measure experimentally. This has resulted in the total domination of unphysical statistical methods, such as pedotransfer functions;

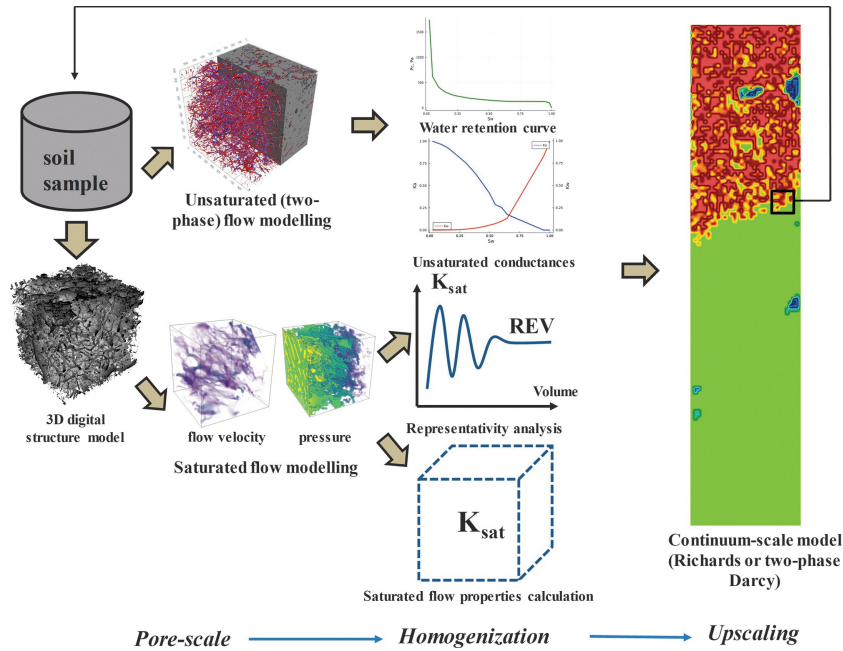


Fig. 1 Homogenization and continuum-scale flow modelling in soils based on the pore-scale simulations – a general scheme.

- 2) Measurement errors due to technological drawbacks (e. g., [Diamond, 2000](#)), such as tool inaccuracies or faulty physical assumptions;
- 3) Changes to the physical properties of samples under laboratory measurements by cracking, self-filtration and other processes;
- 4) Distortions by boundary effects due to coring or sampling;
- 5) Inability to account for dynamic structure influences on transport properties;
- 6) Failure to assess tensorial ([Guibert et al., 2016](#); [Gerke et al., 2019](#)) or even directional properties (e. g., soil cores provide quasi-1D flow properties);
- 7) Field or laboratory measurements provide no information if the studied soil is representative elemental volume (REV) ([Gerke and Karsanina, 2021](#)), as structural information (pore structure or flow field) is needed to assess representativeness.

While pore-scale modelling in soil science is still in early development, the potential of this approach to uncover important aspects of flow physics in soils is without doubt. Moreover, pore-scale flow and transport models can be coupled to other critical processes — mechanical ([Hallett and Newson, 2005](#); [Barbosa et al., 2022a](#)) or biological ([Schnepf et al., 2022](#)) models, to name just a few possible directions. The aim of this contribution is to provide a brief description of pore-scale modelling of flow and transport in general (with practical links to existing codebases) and to discuss why such techniques will inevitably become main-stream for solving various problems in soil physics and beyond.

Input data and pore space characterization

The critical input data for pore-scale models is the 3D geometry of the pore space of the studied soil sample (reference to soil structure chapter). Such information usually comes from X-ray tomography (XCT) imaging ([Abrosimov et al., 2021a](#)) (reference to X-ray tomography chapter) that received widespread adoption for characterizing soil structure. The fundamental problem of almost any imaging method is the ratio of the field-of-view to imaging resolution — the larger the soil sample, the lower the resolution. Consequently, covering hierarchical soil structural features at all interesting scales requires multi-scale imaging with sub-sampling ([Karsanina et al., 2018](#); [Lucas et al., 2021](#)). The leap in imaging techniques development and availability of the XCT devices is facilitating progress in pore-scale modelling.

Unfortunately, XCT imaging alone is either not enough or not applicable in each study situation. This problem is particularly pronounced for nano-scale imaging and owing to some organic matter invisibility on XCT — both can be mitigated with the help of focused ion beam — scanning electron microscopy (FIB-SEM) imaging ([Gerke and Karsanina, 2021](#)). In addition, thin-sections are not only useful to discern different solid phases, including organic matter and mineral crystal orientation, they also provide means of building 2D maps of microbial activity and mineral composition ([Raynaud and Nunan, 2014](#)), these later can be extended over 3D soil structural images.

The 2D nature of some imaging techniques fails to provide meaningful input data for pore-scale modelling methods due to their inherent need for 3D information. Stochastic reconstruction techniques allow re-creation of 3D structure based on limited 2D or 3D information. Different approaches, such as process-based methods (Barbosa et al., 2022b), multiple-point statistics and simulated annealing based algorithms and their hybrids, may provide an effective solution for different structures such as sandy soils, clayey soils or aggregates. The usage of stochastic reconstructions to produce soil images has been attempted in a limited number of studies (Zhang et al., 2005; Karsanina et al., 2015), but is expected to grow in the near future, especially because of the need to fuse multi-scale images into a single 3D digital structural model (Karsanina et al., 2018).

The majority of 3D images obtained by any approach, as mentioned above, will represent gray-scale image with voxel color representing X-ray attenuation or the intensity of electron back-scattering. To separate pore and solid voxels one needs to apply a segmentation procedure, or binarization. In our opinion, the soil science community still have not produced a superior specific segmentation procedure; moreover, automatic binarization techniques usually result in poor image processing (Abrosimov et al., 2021b). Among existing techniques, converging active contours (Sheppard et al., 2004), indicator kriging (Oh and Lindquist, 1999) and region growth (Hashemi et al., 2014) provide a good trade-off between quality and computational expense. If the original image is too noisy, an open-source non-local means filter (Bruns et al., 2017) can be a helpful pre-segmentation step. Note that due to the resolution limit, any segmentation of the XCT image by default introduces errors — see a discussion of this effect and a possible solution with the help of the deep supervised learning on physically modelled data in Lavrukhin et al. (2021).

After the 3D image of porous space is obtained, it can serve (together with any additional information) as input data for pore-scale modelling. However, one additional step in pore characterization is needed for so-called pore-network modelling approach. The idea is to simplify the geometry of the porous space by representing it with the help of pore and throat elements with a limited variety of shapes while conserving the topology. The classical approach to represent pore shapes is the circle-triangle-square or CTS approach (Valvatne and Blunt, 2004) was critiqued for its inability to represent non-convex shapes. It was later shown that one can remove the simplification of the shapes completely with the help of direct solutions and machine learning (Miao et al., 2017). The problem of topology conservation was recently solved (Zubov et al., 2022), as it was shown that all popular extraction methods fail in preserving pore-network's Euler numbers (Gerke et al., 2020). At the time of writing the dominant state-of-the-art extraction approaches include watershed-based (Rabbani et al., 2014; Gostick et al., 2016), median axis-based, maximal inscribed balls (Dong and Blunt, 2009) algorithms, and their hybrids (Gerke et al., 2020). In addition to the ability to effectively simulate single and multi-phase flow, the extraction of the pore-network models also allows morphological analysis such as pore and throat size distributions, pore shapes and related properties. The example of the 3D image of the soil pore space and of the extracted pore-network model is shown in Fig. 2 in the form of so-called ball-and-stick diagram (note that this representation is chosen due to visibility, pores are not spheres while throats are not cylinders).

For multi-flow simulation in addition to 3D pore geometry the parameters of the fluid-fluid and fluid-solid interactions are needed. These are usually supported in the form of the contact angles that are linked to interfacial tensions or free energy potentials. The value of static and dynamic (during fluids movements) contact angles depend on a number of parameters such as chemical constitution of the fluid and solids, roughness of solid walls and the history of the interaction between solids and fluids. Static contact angles are either measured in the laboratory on prepared surfaces (link to contact angle chapter) or can be measured in situ based on XCT tomography. Dynamic angles are obtained either through direct pore-scale modelling or recalculated from static counterparts using empirical models. Both static and dynamic contact angles of confined fluids in nanopores can be also modelled using MD or Hydrodynamical DFT approaches.

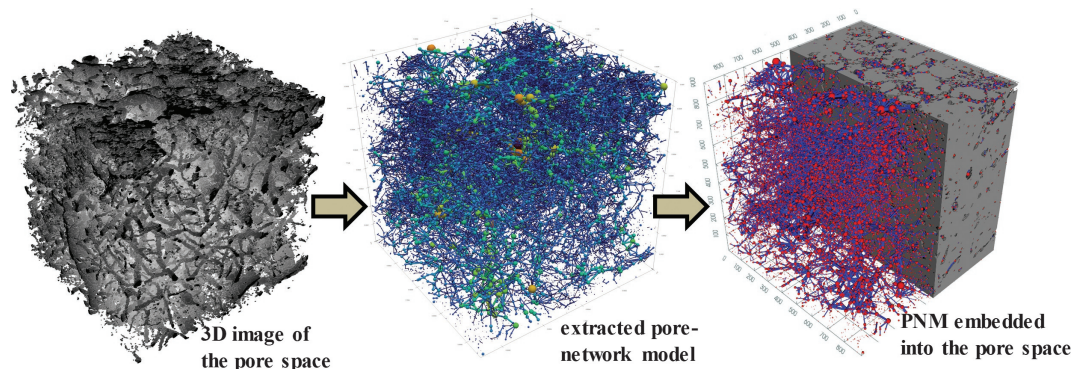


Fig. 2 The results of the pore-network extraction procedure — the PNM is extracted from the 3D pore space image (left) and is usually represented with the help of the ball-and-stick diagram (middle). Note that this does not mean that pores are spherical and throats are cylindrical. To visualize the correspondence of the PNM model to the original pore space geometry we also show PNM embedded into the solid phase of the soil sample studied (right).

Numerical methods

The flow of fluid(s) on the pore-scale is classically described by a Navier-Stokes equation:

$$\begin{cases} \frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla) \mathbf{v} - \frac{\mu}{\rho} \Delta \mathbf{v} + \frac{\nabla p}{\rho} = 0 \\ \operatorname{div} \mathbf{v} = 0 \end{cases}, \quad (1)$$

where $\mathbf{v} = (v_x, v_y, v_z)$ is velocity field, μ is fluid viscosity, ρ is fluid density, and p is pressure field. This equation can describe the flow in a wide variety of velocity magnitudes starting from the very slow to high-velocity turbulent regimes. It is generally assumed that porous media flow, including within soils, happens with very low velocities (as characterized by so-called Reynolds number, Re , that should be less than 1 for Darcy equation to hold after homogenization, $Re = \frac{\rho u l}{\mu}$, where u is the absolute flow velocity and l is the characteristic length of pore or channel). In the case of $Re < 1$, the flow happens in creeping or Stokes regime and can be described by a simplified version of Navier-Stokes equation (with eliminated inertia term) that is called Stokes equation:

$$\begin{cases} \mu \Delta \mathbf{v} - \nabla p = 0 \\ \operatorname{div} \mathbf{v} = 0 \end{cases}. \quad (2)$$

Unlike Eq. (1), Eq. (2) does not contain time — this form of the equation is used to show that in this case we are interested only in single-phase properties, one needs to solve a stationary problem and, thus, time can be excluded. The case where the transient problem is of interest, Stokes equation can include the time variable similarly to Eq. (1). Both Eqs. (1) and (2) can be solved numerically if some boundary conditions, e. g., pressure differences on the sides of the modelling domain, are known; and the spatial discretization of the modelling domain is performed. The latter is usually performed with cubical cells, as they are naturally arising from X-ray computed tomography images. However, depending on the numerical method utilized some other type of computation mesh such as tetrahedral or unstructured meshes could be preferred. This, however, add another complexity level on top of image processing, as one needs to create a mesh based on segmented pore space 3D image first.

Eqs. (1) and (2) are written for single-phase, or saturated flow. It is possible to modify these equations to include any number of fluids in case they are immiscible (by adding subscripts for all different fluids considered). However, to close the system we also need to describe the interaction between fluids and solid phases, as well as between the fluids themselves. Here, the choice of the model depends on the numerical method that is chosen and the boundary can be either exact or diffuse. In other words, we solve the same flow problem as for single-phase flow, but for each fluid involved plus interaction between fluids and with the solid phase. This results in a much more computational intensity needed for simulating multi-phase flow.

There are plenty of diverse numerical method to solve single- and multi-phase flow equations at the pore-scale: different classes of finite-difference (FDM), finite-element (FEM), finite-volume (FVM) methods, lattice-Boltzmann methods (LBM) and smooth-particle hydrodynamics (SPH). Multi-phase flows require the treatment of the phase boundary tracking, i.e. volume-of-fluid (VOF) approach or a phase field methods. They differ not only in terms of the solution of Eqs. (1) and (2), but also in terms of the description of the interactions between solid and fluid phases. For example, VOF method allows exact interface tracking, but this feature is computationally heavy and may produce artifacts for some specific pore geometries. The phase-field methods operate on the diffuse-boundary, e.g., as described by the Cahn-Hilliard potential model. Although in the case of the diffuse boundary there is no abrupt interface between the fluids, this approach allows more physical description of the interactions based on free energy (that can be utilized to parameterize contact angles). The choice of the method to use in each case is non-trivial and depends on the aims of the numerical study, as each approach may have its significant advantages and pitfalls. The authors tend to gravitate in the direction of phase-field models in case of direct multi-phase flow simulations within soils. The most popular LBM approach is easy to implement within the code and to parallelize, but the physical model behind, especially for multi-phase flow is questionable (despite a widespread delusion that LBM generalized to Navier-Stokes equations).

The computational efficiency of different methods is hard to be compared as it depends on numerous aspects of the numerical code, including the solution method, the architecture of computational resources (e. g., potential usage of hybrid CPU + GPU), the choice of the libraries or compilers and explicit or implicit numerical schemes. As a rule of thumb the more computationally expensive the method is, the more accurate the solution is. The majority of existing codes or software utilize explicit schemes due to the ease of implementation and parallelization (which is usually linear to the number of cores used). The design and coding of implicit numerical schemes is much more involved, but it is expected to start dominating direct pore-scale simulations in the future. The benefit of implicit schemes is that such schemes can solve stationary equations with a few iterations (Stokes flow) or allow significantly larger time steps that don't depend on the temporal discretization (unlike LBM or explicit methods). The drawback of such schemes is their complexity and a need to solve large scale systems of linear equations and complexity in their parallelization. However, with the usage of asymptotically optimal iterative linear solvers (i.e. multigrid methods) one can expect optimal performance that is proportional only to the size of the domain discretization.

Compared to direct pore-scale models which are based on Eqs. (1) and (2) solutions, pore-network models are orders of magnitude more computationally efficient. We argue that PNMs are actually the only available technology to perform practical pore-scale simulations of large volumes of soil material, as direct models require high-performance resources to describe rather small volumes (usually limited to 1000^3 voxels depending on the imaging resolution, porosity and number of phases among other factors). Pore-network simulators significantly reduce computation times by solving the flow on the graph with hydraulic

conductances for saturated and unsaturated flow that are already prescribed on each pore-throat-pore element based on simplified geometry (Tuller et al., 1999; Patzek and Kristensen, 2001) or based on neural network prediction (Miao et al., 2017). Therefore the solution for saturated hydraulic conductance takes a mere second, while two-phase (air-water) simulations may require from minutes to hours on a single CPU compared with days to weeks on a multi-CPU cluster for direct models. The significant variation for computational times for unsaturated flow is due to the complexity of pore-network used. Roughly speaking, PNM can be divided into two classes — quasi-static and dynamic models. The former assumes that viscous forces are negligible, i.e., flow occurs slowly and can be described within some capillary pressure steps — the displacement of one fluid with the other depends only on local menisci curvature. The latter takes viscous forces into account and describes some dynamic effects due to flow velocity (hence the name). The idea behind the first class of models is that all major unsaturated flow processes are rather slow and capillarity is the dominating force. Whether this is true for all soils during gravitational water flow is an open question, as soils have a wide variety of pore sizes. Some pressure boundary conditions, for example, water table above the soil, or larger pores may reduce the importance of capillary forces and make viscous forces dominant. In this case dynamic pore-network models are necessary to describe flow (Joekar-Niasar and Hassanizadeh, 2012). The ability to account for viscous forces does not come in for free — one needs to solve much more involved systems of equations and, thus, dynamic models are much more computational time consuming. But compared to quasi-static models they not only add viscous forces, but also have time resolution. The type of the model needs to be chosen based on soil type and flow conditions, depending on local flow properties it is possible to combine both models within the same pore-network solver.

Assuming low flow velocities the models described above deal with Darcian or linear flow regime — volume averaged velocities depend only on pressure gradient thus making permeability (or saturated hydraulic conductivity) independent of pressure differences. However, depending on pore sizes and flow conditions pre- and post-Darcian flow regimes can be potentially relevant to soil processes (Fig. 3). The term pre-Darcian is controversial — here we use it to refer to rarified or nano-confined conditions where Stokes' regime is not valid. Such deviations from the linear relationship result due to diffusion (for low Knudsen numbers — in case flowing molecules interact more often with the walls compared to each other), slip effects on the pore walls (for intermediate Knudsen numbers), and their combination (intermediate diffusion-slip flow regime). In soil processes pre-Darcian flow regime can potentially take place for air flow within sub-micron pores or for water flow within sub-micron pores with organic walls, but requires significant additional research to explore and describe them numerically (due to severe limitations in experimental studies with such pore systems in the laboratory). The post-Darcian regime leading to decreased relationship between flow and pressure gradient occurs due to vorticity originating within the pore space for flow velocities with Re above 1 (Muljadi et al., 2016). Note that such flow is usually not turbulent (in the statical turbulence theory sense) and can be considered to be a transition from Stokes to Navier-Stokes flow regime. Post-Darcian regime may influence the way we model flow within the macropores for the description of the preferential flow (reference to preferential flow chapter). The relevance of non-Darcian flow for soil processes description is an area of emerging research.

Non-Darcian flow processes can be explicitly modelled using MD approaches, but due to computational expenses solution for pores with diameter up to 50 nm can be achieved. For micron size pores some direct modelling approaches can be utilized, but due to their upscaled nature some parameters (e.g., tangential accommodation coefficient) may be necessary to be evaluated from MD or DFT simulations. If direct solution or an empirical model is available, a pore-network model can be parameterized and used for effective simulation on larger modelling domains (Mehmani et al., 2013).

In addition to single- and multi-phase fluid flow problems, pore-scale modelling approach allows simulation of numerous other transport coefficients and physical processes. One relevant example is the dispersion that can be easily simulated on top of velocity distributions resulting from solving Eqs. (1) or (2). To simulate mass transfer a particle tracking approach may be used — the

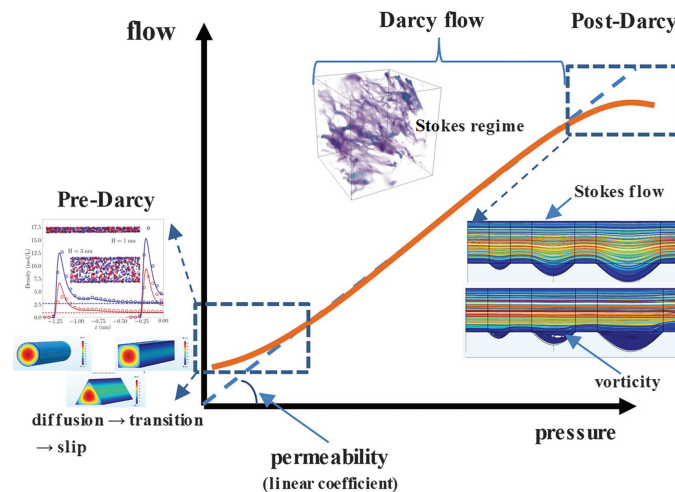


Fig. 3 Illustration of pre-Darcian, Darcian and post-Darcian regimes and their influence on the transport properties of soils.

“molecules” of the substance are placed within the fluid and are moved on each time step randomly due to diffusion (according to known diffusion coefficient) and along the local flow velocity vector due to advection. Parallel implementations for millions of particles can describe complex dispersion patterns within porous media (Khirevich and Patzek, 2019), including hierarchical structures. Direct solution can be again extended to pore-network models and utilized for computationally efficient mass transport models over huge soil volumes to obtain macroscopic dispersion coefficients. Other transport processes such as heat flow or mechanical properties can be simulated using aforementioned numerical methods, but with different systems of equations solved and, thus, are not considered in detail here.

Examples of the application of pore-scale modelling

To display the power of pore-scale modelling approaches here we shall simulate saturated and unsaturated flow for a single soil 3D image. As an object of the demonstration we chose a well-described soil sample Ah-1 (low porosity) from the work of (Gerke and Karsanina, 2021). The 900^3 voxels 3D geometry of this soil sample and the pore-network extracted with the approach of Gerke et al. (2020) has been already shown in Fig. 2. At first we solve the Stokes equation (Eq. 2) with the help of FDMSS Stokes solver that was exhaustively described by Gerke et al. (2018). The FDMSS solver is an open software and is optimized to perform on a regular machines such as a 3D image with the volume of 500^3 voxels can be processed on a modern laptops and up to 1000^3 voxels with the help of the desktop workstations. The results of the simulations along a single direction are shown in the form of the pressure and flow velocity fields are depicted on Fig. 4. After averaging of these fields one can compute the permeability (or saturated hydraulic conductivity which is essentially the same value normalized by the fluid viscosity and is temperature dependent for water). In addition to classical directional permeabilities it is also possible to compute a full permeability (or K_{sat}) tensor, but producing a physically relevant value for a REV volume soil would require geometrical periodization with the help of stochastic reconstructions (Gerke et al., 2019) or the knowledge of the exact pore geometry around the studied sample is needed otherwise.

For a two-phase flow case we chose a PNM modelling approach due to its computational efficiency for 900^3 voxels volume at hand. After the extraction from the 3D image the pore-network model was passed into a popular open PNM-solver by Valvatne and Blunt (2004). We have modelled the air-water (under room temperature conditions) drainage process with perfect water wetting conditions (0° contact angle) and under quasi-static assumption. The results of the simulations including resulting water retention curve and relative permeabilities for both air and water are shown in Fig. 5.

Here we limited our consideration to the computationally inexpensive methods that are freely available to the public in the form of open codes or free software and can be reproduced by any willing researcher; all visualization was performed with the Paraview software or vtk library.

Existing codes

One of the main messages we want to convey in this chapter is that any interested researcher can manage to start pore-scale simulations with the help of the existing software. Available resources include open codebases and commercial packages — Table 1 summarizes major existing resources and briefly discusses advantages and shortcomings of each potential solution. The information we provide here is not comprehensive, especially considering the vast amount of commercial solutions available, yet we believe that closed source commercial software does not provide a clear methodological way forward and, thus, are not a preferred way of doing

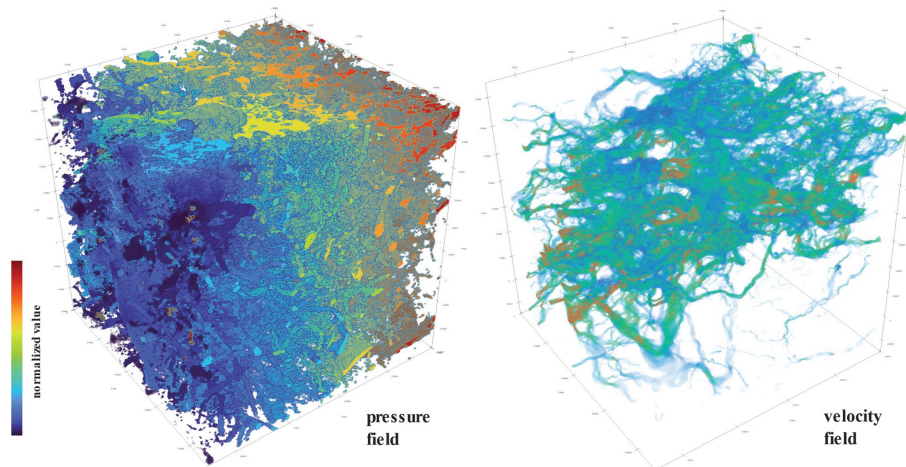


Fig. 4 Pressure and flow velocity field visualization after a single-phase direct flow simulations (using FDMSS free software). These fields are input data for homogenization procedures that allows permeability or saturated hydraulic conductivity (including their full tensorial representations) to be obtained.

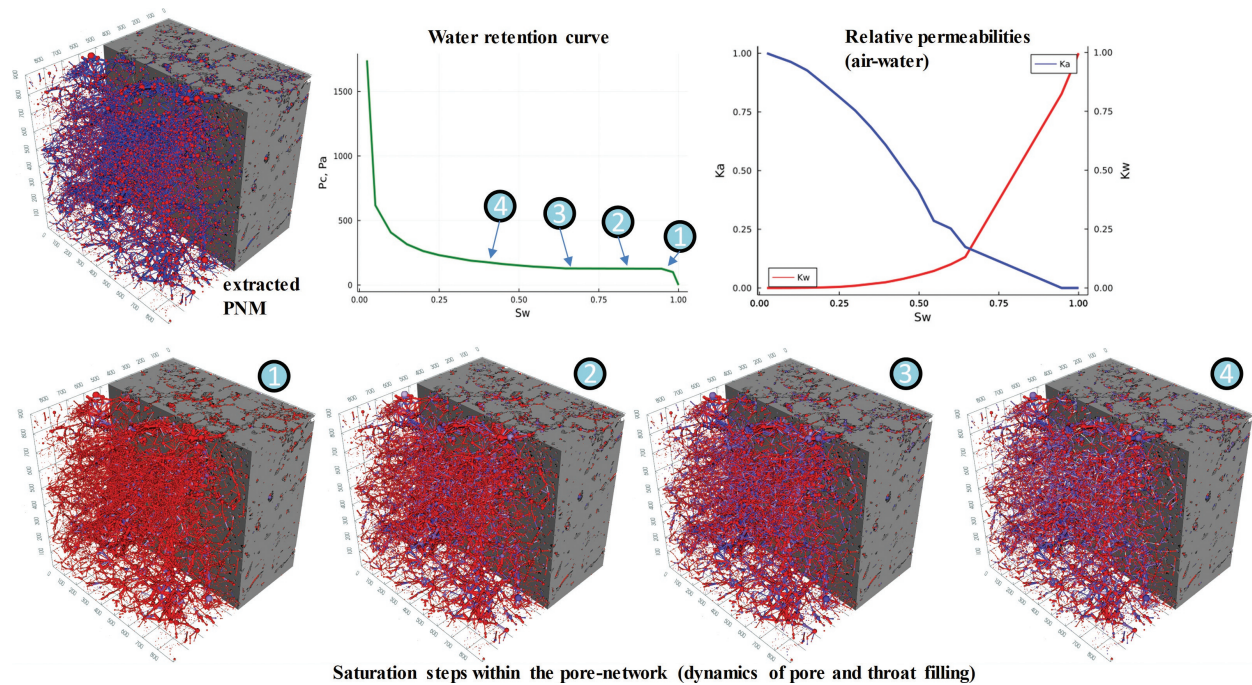


Fig. 5 The examples of two-phase (air displacing fully wetting water – drainage) flow simulations within a soil sample including visualizations of different pore filling stages that result in water retention curve and relative permeabilities (or unsaturated conductances).

science compared with codebases extensively described in research papers. We have to admit, however, that the quality of research codes is usually low compared to commercial products. Its usage may also require some programming skills to deal with or improve them to deal with ones current research needs.

Homogenization, upscaling and beyond

As mentioned in the Introduction, pore-scale modelling is able to mitigate numerous shortcomings of the field or laboratory methods to obtain soil physical properties, such as permeability (saturated hydraulic conductivity), water retention curve, relative permeabilities for air and water (unsaturated conductivities), dispersion coefficients. Among others benefits, the modelling approach allows to obtain unsaturated properties much faster than laboratory measurements and with known boundary conditions (e. g., pressure at the outlet), that are usually hard to manage in field measurements. Needless to say, once the 3D structure of the soil sample is known, the amount of modelling experiments that can be performed on this sample is virtually limitless — this is something that is hard to achieve in the laboratory. The ability of pore-scale modelling methods to determine the full tensorial physical properties is especially valuable for homogenization and upscaling (Fig. 1) in order to parameterize continuum-scale hydrological models.

Pore-scale modelling also allows us to homogenize soil volumes that do not exhibit REV properties — this is especially important considering the fact that majority of standard cores may lack representativity in case of structured soils (Gerke and Karsanina, 2021). This ability of pore-scale methods to operate on any volume of soil may also be invaluable in determination of coefficients of exchange between soil matrix and preferential flow paths, e. g., in the form of the macropores (Barbosa et al., 2022a). In addition, some mass transfer parameters, for example, mineral nutrition solution flow from nano-porous organic-mineral assemblages to the plant roots, may only be valid on the very small scale, well beyond the abilities of the laboratory measurements. In such applications, there is no alternative to pore-scale modelling.

Local knowledge of velocities or water-air menisci positions and phase connectivities may be needed to explain the living conditions of microorganisms and to simulate their habitat evolution, organic matter mineralization or CO₂ production processes. Another example could be the modelling of swell-shrink behavior of soils depending on the local availability of water, where water availability and air trapping can be, in turn, affected by the structural dynamics. Classical models for unsaturated parameters, such as the van Genuchten-Mualem model, can be modified to account for hysteresis, but such modifications are purely empirical and fail to account to all possible boundary conditions variations. Pore-scale modelling can describe any history of flow conditions and any number of hysteretic loops. In other words, pore-scale modelling approaches allow us to operate on any size of the modelling domain, at least from above it is limited only by the computational resources available — and this is the limit that requires us to upscale to the continuum-scale models, and to describe the parameters that continuum-scale needs to operate.

Table 1 A brief summary of existing pore-scale modelling resources.

<i>Code or software</i>	<i>Functionality</i>	<i>Advantages</i>	<i>Shortcomings</i>	<i>Reference to the source</i>
OpenFOAM	An open computational fluid dynamics codebase mining to solve numerous (coupled) problems, cross-platform (yet Linux is preferable)	Has a well documented solvers that can be programmed to solve virtually any system of equations	A very steep learning curve with programming skills required, some solvers are based on rather old technologies and are not always state-of-the-art, in general very computationally expensive methods, limited support of block matrices in linear systems	https://www.openfoam.com/ A porous media solver based on OpenFOAM core by Ali Raeini: https://github.com/ImperialCollegeLondon/poreFoam-singlePhase (single phase) https://github.com/ImperialCollegeLondon/poreFoam (two-phase as described by Raeini et al. (2012)) https://palabos.unige.ch/
Palabos	A cross-platform code for single and two-phase LBM flow simulations	A wide variety of boundary conditions, very well documented and robust package with a significant back up of scientific publications	Some programming skills are required to setup computations	
FDMSS	A finite-difference Stokes solver (single-phase flow) software	Very efficient compared to other examples in this table — one can simulate 500^3 voxels cubes on a modern laptop and up to 1000^3 voxels on desktop workstations, very easy to use	The accuracy of the solution is lower than for the LBM method	https://porenetwork.com/download/fdmss/ (Windows version, Linux version can be obtained from the authors on demand) The description and the manual are provided by Gerke et al. (2018)
OpenPNM	A open-source library for pore-network-based simulations — includes the PNM extractor and simulation parts	Easy to start and use python-based code-base	Somewhat rudimentary flow simulations as compared to other codes	https://openpnm.org/ (also includes other libraries such as porespy to characterize porous structures and more, Gostick et al., 2016)
Dong's code	A classical code for maximal inscribed-balls PNM extraction methodology, saves the results into the 'stat-oil' format	A long time leading open-source solution for PNM extraction	Has been shown to have some inaccuracies. The codebase is very hard to manage (even after code improvements performed by Ali Raeini)	https://github.com/ImperialCollegeLondon/pnextract (Dong and Blunt, 2009)
Valvatne's code	A classical code for single- and two-phase simulations in PNMs. Reads in 'stat-oil' format and works perfectly in pair with the Dong's extractor	A long time (and still is) leading open-source solution for PNM simulations	Produces unphysical results in some particular cases, the model within the code is sometimes different from the one described in the paper	https://github.com/ImperialCollegeLondon/pnflow (the original code as described by Valvatne and Blunt (2004) has been significantly refactored and improved by Ali Raeini)
Comsol	A versatile finite-element method commercial software for a variety of simulations, including pore-scale flow simulations (single, two- and three-phases) that can be coupled with other processes, e. g., mechanical deformations	Very well documented with huge learning resources and useful support services	Commercial software with closed code, lacks built-in mesher to convert 3D voxelized images to polygonal/unstructured meshes necessary to produce solutions with Comsol	https://www.comsol.com
Paraview	A current state-of-the-art visualization tool based on vtk library (was used to produce flow and PNM visualization in this chapter)	Very well documented with variety of formats to save data for visualization (easy to implement within ones code or export from major open codebases or commercial software)	Unstable and susceptible to occasional crashes (in case of problems try other version, Linux versions are preferable)	https://www.paraview.org/

The immediate question is “Can one combine the computational power of pore-scale and the efficiency of the continuum-scale models?” This can be achieved by coupling of these two scales into a single model that exchanges parameters — pore-scale provides upscaled physical flow properties, continuum-scale model updates all fluxes or processes within a large modelling domain and returns the boundary conditions (e.g., pressures, saturations) back to pore-scale to re-calculate the properties. As far as we know, Heiba et al. (1986) were the first to propose such a coupling. However, the implementation for such a coupling is non-trivial and only some simple 1-D preliminary studies are known (Sheng and Thompson, 2013). In addition to dynamic coupling, such models could include any number of scales in a Russian doll manner — continuum-scale links to pore-scale, which in turn is parameterized by MD/DFT models. Such linking can also be accelerated with the help of the machine learning. While still looking somewhat distant future, such coupled multi-scale multi-physics models are the most appealing approach to describe the multitude of soil processes on all necessary scales from nm to meters.

Summary and outlook

In this chapter we briefly described the current state of pore-scale models for soil processes with the focus on flow and transport problems. While there are tools that can solve flow governing equations (Navier-Stokes type equations or their simplifications), their application in soil physics is very limited, likely for historical reasons — pore-scale simulations received a lot of attention in petroleum engineering and related disciplines, while soil was for long considered to be a complex living substance that is hard to be described numerically. With the development of soil structure imaging methods on all scales (XCT imaging, thin-sections and FIB-SEM), a realization of a fully 3D digital model of soil structure that can account for dynamic changes and multi-phase soil constituents is a matter of time and effort (Karsanina et al., 2021). With such a model at hand, building a multi-scale pore-network model for large scale flow and transport simulations is only a practical challenge with all technological aspects already known.

In Table 1 we have summarized some major open or free and commercial tools for pore-scale simulations. With such tools any soil scientists can incorporate pore-scale modelling into her or his research. However, some care has to be taken while doing so — one needs to understand the underlying physics to make sure the correct model and parameters are chosen for the task.

In soil physics applications a lot of additional work is actually needed to verify pore-scale models and to understand which physical phenomena, for example, nano-scale transport, is important for which physical property or process. Such work is complicated by a gap between the volumes of soil operated upon in laboratory methods and pore-scale models; but this gap has been shrinking for the last 10 years and can be completely eliminated with the help of multi-scale pore-networks extracted from fused 3D images (Karsanina et al., 2018).

In summary, we still tend to think that pore-scale modelling is an evolving technology. By saying this we mean that to apply it to numerous soil problems “out of the box” a lot of addition research is necessary — one needs to make sure that all processes relevant to soils (as compared to hydrocarbon bearing rocks) are properly incorporated into the model. With the vast potential and ability to explore different scales and processes that pore-scale modelling provides, we are confident that such tuning is a matter of 5–10 year. For a variety of soil processes mentioned here, pore-scale modelling simply has no alternatives; moreover, it allows to study any volume (especially small areas or volumes dominated by sub-micron porosity) with any boundary conditions — something that is close to impossible to achieve in laboratory or field.

Acknowledgments

We thank our colleagues Dr. Nickolay Evstigneev and Dr. Aleksey Khlyupin for very fruitful flow modelling discussions, Rustem Sirazov, Andrey Ananov and Mikhail Gomorev for the help with figure creation.

References

- Abrosimov KN, Gerke KM, Fomin DS, Romanenko KA, and Korost DV (2021a) Tomography in soil science: From the first experiments to modern methods (a review). *Eurasian Soil Science* 54(9): 1385–1399.
- Abrosimov KN, Gerke KM, Semenov IN, and Korost DV (2021b) Otsu’s algorithm in the segmentation of pore space in soils based on tomographic data. *Eurasian Soil Science* 54(4): 560–571.
- Barbosa LA, Gerke KM, and Gerke HH (2022a) Modelling of soil mechanical stability and hydraulic permeability of the interface between coated biopore and matrix pore regions. *Geoderma* 410: 115673.
- Barbosa LA, Gerke KM, Munkholm LJ, Keller T, and Gerke HH (2022b) Discrete element modeling of aggregate shape and internal structure effects on Weibull distribution of tensile strength. *Soil and Tillage Research* 219: 105341.
- Bruns S, Stipp SL, and Sørensen HO (2017) Looking for the signal: A guide to iterative noise and artefact removal in X-ray tomographic reconstructions of porous geomaterials. *Advances in Water Resources* 105: 96–107.
- Diamond S (2000) Mercury porosimetry: an inappropriate method for the measurement of pore size distributions in cement-based materials. *Cement and Concrete Research* 30(10): 1517–1525.
- Dong H and Blunt MJ (2009) Pore-network extraction from micro-computerized-tomography images. *Physical Review E* 80(3), 036307.
- Gerke KM and Karsanina MV (2021) How pore structure non-stationarity compromises flow properties representativity (REV) for soil samples: Pore-scale modelling and stationarity analysis. *European Journal of Soil Science* 72(2): 527–545.

- Gerke KM, Vasilyev RV, Khirevich S, Collins D, Karsanina MV, Sizonenko TO, and Mallants D (2018) Finite-difference method Stokes solver (FDMSS) for 3D pore geometries: Software development, validation and case studies. *Computers & Geosciences* 114: 41–58.
- Gerke KM, Karsanina MV, and Katsman R (2019) Calculation of tensorial flow properties on pore level: Exploring the influence of boundary conditions on the permeability of three-dimensional stochastic reconstructions. *Physical Review E* 100(5): 053312.
- Gerke KM, Sizonenko TO, Karsanina MV, Lavrukhin EV, Abashkin VV, and Korost DV (2020) Improving watershed-based pore-network extraction method using maximum inscribed ball pore-body positioning. *Advances in Water Resources* 140: 10357.
- Gostick J, Aghighi M, Hinebaugh J, Tranter T, Hoeh MA, Day H, and Putz A (2016) OpenPNM: A pore network modeling package. *Computing in Science & Engineering* 18(4): 60–74.
- Guibert R, Horgue P, Debenest G, and Quintard M (2016) A comparison of various methods for the numerical evaluation of porous media permeability tensors from pore-scale geometry. *Mathematical Geosciences* 48(3): 329–347.
- Hallett PD and Newson TA (2005) Describing soil crack formation using elastic–plastic fracture mechanics. *European Journal of Soil Science* 56(1): 31–38.
- Hashemi MA, Khaddour G, François B, Massart TJ, and Salager S (2014) A tomographic imagery segmentation methodology for three-phase geomaterials based on simultaneous region growing. *Acta Geotechnica* 9(5): 831–846.
- Heiba AA, Jerauld GR, Davis HT, and Scriven LE (1986) Mechanism-Based Simulation of Oil Recovery Processes. In: *SPE Annual Technical Conference and Exhibition*. SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers.
- Joekar-Niasar V and Hassanizadeh SM (2012) Analysis of fundamentals of two-phase flow in porous media using dynamic pore-network models: A review. *Critical Reviews in Environmental Science and Technology* 42(18): 1895–1976.
- Karsanina MV, Gerke KM, Skvortsova EB, and Mallants D (2015) Universal spatial correlation functions for describing and reconstructing soil microstructure. *PLoS One* 10(5), e0126515.
- Karsanina MV, Gerke KM, Skvortsova EB, Ivanov AL, and Mallants D (2018) Enhancing image resolution of soils by stochastic multiscale image fusion. *Geoderma* 314: 138–145.
- Karsanina MV, Lavrukhin EV, Fomin DS, Yudina AV, Abrosimov KN, and Gerke KM (2021) Compressing soil structural information into parameterized correlation functions. *European Journal of Soil Science* 72(2): 561–577.
- Khirevich S and Patzek TW (2019) Three-dimensional simulation of tracer transport dynamics in formations with high-permeability channels or fractures: Estimation of oil saturation. *Physics of Fluids* 31(11), 113604.
- Lavrukhin EV, Gerke KM, Romanenko KA, Abrosimov KN, and Karsanina MV (2021) Assessing the fidelity of neural network-based segmentation of soil XCT images based on pore-scale modelling of saturated flow properties. *Soil and Tillage Research* 209: 104942.
- Lucas M, Vetterlein D, Vogel HJ, and Schlüter S (2021) Revealing pore connectivity across scales and resolutions with X-ray CT. *European Journal of Soil Science* 72(2): 546–560.
- Mehmani A, Prodanović M, and Javadpour F (2013) Multiscale, multiphysics network modeling of shale matrix gas flows. *Transport in Porous Media* 99(2): 377–390.
- Miao X, Gerke KM, and Sizonenko TO (2017) A new way to parameterize hydraulic conductances of pore elements: A step towards creating pore-networks without pore shape simplifications. *Advances in Water Resources* 105: 162–172.
- Muljadi BP, Blunt MJ, Raeini AQ, and Bijeljic B (2016) The impact of porous media heterogeneity on non-Darcy flow behaviour from pore-scale simulation. *Advances in Water Resources* 95: 329–340.
- Oh W and Lindquist B (1999) Image thresholding by indicator kriging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21(7): 590–602.
- Patzek TW and Kristensen JG (2001) Shape factor correlations of hydraulic conductance in noncircular capillaries: II. Two-phase creeping flow. *Journal of Colloid and Interface Science* 236(2): 305–317.
- Rabbani A, Jamshidi S, and Salehi S (2014) An automated simple algorithm for realistic pore network extraction from micro-tomography images. *Journal of Petroleum Science and Engineering* 123: 164–171.
- Raeini AQ, Blunt MJ, and Bijeljic B (2012) Modelling two-phase flow in porous media at the pore scale using the volume-of-fluid method. *Journal of Computational Physics* 231(17): 5653–5668.
- Raynaud X and Nunan N (2014) Spatial ecology of bacteria at the microscale in soil. *PLoS One* 9(1), e87217.
- Schnepf A, Carminati A, Ahmed MA, Ani M, Benard P, Bentz J, and Vetterlein D (2022) Linking rhizosphere processes across scales: Opinion. *Plant and Soil* 1–38.
- Sheng Q and Thompson K (2013) Dynamic coupling of pore-scale and reservoir-scale models for multiphase flow. *Water Resources Research* 49(9): 5973–5988.
- Sheppard AP, Sok RM, and Averdunk H (2004) Techniques for image enhancement and segmentation of tomographic images of porous materials. *Physica A: Statistical Mechanics and its Applications* 339(1–2): 145–151.
- Tuller M, Or D, and Dudley LM (1999) Adsorption and capillary condensation in porous media: Liquid retention and interfacial configurations in angular pores. *Water Resources Research* 35(7): 1949–1964.
- Vaganova M, Nesterova I, Kanygin Y, Kazennov A, and Khlyupin A (2022) Linking theoretical and simulation approaches to study fluids in nanoporous media: Molecular dynamics and classical density functional theory. *Chemical Engineering Science* 250: 117383.
- Valvatne PH and Blunt MJ (2004) Predictive pore-scale modeling of two-phase flow in mixed wet media. *Water Resources Research* 40(7): W07406.
- Zhang X, Deeks LK, Bengough AG, Crawford JW, and Young IM (2005) Determination of soil hydraulic conductivity with the lattice Boltzmann method and soil thin-section technique. *Journal of Hydrology* 306(1–4): 59–70.
- Zubov AS, Murygin DA, and Gerke KM (2022) Pore-network extraction using discrete Morse theory: Preserving the topology of the pore space. *Physical Review E* 106(5), 055304.

Soil texture: Its relevance, measurement and representation internationally

Attila Nemes, Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway;
Division of Environment and Natural Resources, Norwegian Institute of Bioeconomy Research (NIBIO), Ås, Norway

© 2023 Elsevier Ltd. All rights reserved.

This is an update of G.W. Gee, Texture, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, pp. 149–155, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00428-8>.

Introduction	36
Soil particle size measurements	36
Sample pre-treatment for fine earth measurements	37
Frequently used measurement methods	37
The pipette and hydrometer methods	37
The integral suspension pressure (ISP) method	37
Light scattering, laser diffractometry (LDM)	38
Compatibility of measurements using methods based on different measurement theories	38
Compatibility of measurements using methods of similar measurement theories	39
Advantages and disadvantages of different novel measurement techniques	40
Measurement effort and cost	40
The added value of continuous data—International compatibility	40
Particle size and textural classification schemes	40
Particle size schemes	40
Textural classification systems	41
Conversion of data using interpolation	41
Attempts at unification of soil particle-size measurements	43
Use of soil texture data to infer other soil properties	44
Inference to pore-size distribution	44
Use of texture data to estimate soil hydraulic properties	44
Limited success in estimating hydraulic conductivity	45
The tumble effect of misrepresenting soil texture	45
Other uses of soil texture	45
Summary	46
References	46

Abstract

Soil texture is one of the most fundamental and longest recorded soil properties. It refers to the balance between the various size fractions of the soil's solid constituents. This chapter summarizes commonly used and emerging methodologies to determine, and systems to represent soil particle-size distribution and to classify soil texture. Given the diversity in methodology as well as classification systems used internationally, we also address data compatibility issues and harmonization options. We close the chapter by presenting some of the classic uses of soil texture information.

Glossary

Bulk density The total mass of solids per unit soil volume.

Liquid limit The water content at which the soil changes from a plastic state to a liquid state.

Plasticity The soils ability to undergo deformation without cracking.

Plastic limit The moisture content at which soil can no longer be remolded without cracking.

Porosity Total volume of pores, whether filled by the liquid or gas phase, per total volume.

Specific surface area The total surface area of a material per unit of mass.

Void ratio The ratio of the space occupied by the voids to the space occupied by solids.

Particle-size distribution or grain size distribution or gradation Refers to the abundance of different sizes of individual soil particles.

Atterberg limits Proposed by Albert Atterberg (1846–1916), the Plastic Limit (PL) and Liquid Limit (LL) correspond to soil moisture levels at which the soil's consistency changes.

Brownian motion The random motion of particles suspended in a fluid medium.

Pedotransfer function A relationship (e.g., equation, table, decision tree) from which certain soil properties can be estimated from other soil properties.

The A-line It is an empirically chosen line on the plasticity chart that separates clays (soils that fall above A-line) and silts (soils that fall below the A-line). The A-line is given by the equation $PL = 0.73(LL - 20)$.

Key points

- Soil texture, deduced from a field test or soil particle-size analysis, is an essential soil property that is a product of soil formation and determines or influences key functional soil processes.
- Soil particle-size distribution is determined by a variety of conventional and new methods that are often based on different measurement theories and underlying assumptions and present substantial challenges of data compatibility.
- A uniform textural classification of soils world-wide is hindered by the diversity of measurement techniques and sample preparation practices being used, as well as variations in the fraction limits that are used to define the three basic textural fractions: sand, silt and clay content.
- Soil texture is used to infer a variety of other soil properties that are often more difficult to determine, and is frequently used to help delineate areas of similar hydrological behavior in large scale modeling applications.

Introduction

Soil consists of a collection of particles that differ widely in size and shape. Some particles are coarse enough to be seen easily with the naked eye, while others are so small that they can only be seen with powerful microscopes. The term ‘soil texture’ refers to the relative proportion of the various sizes of mineral soil particles. It is usually expressed with reference to the most dominant particle sizes found in the soil and has both qualitative and quantitative connotations.

Qualitatively, ‘soil texture’ refers to the ‘feel’ of the soil material. When rubbed by hand, the soil can give a feel that ranges from coarse and gritty to fine and smooth. Experienced soil classifiers can press soil into their hands and tell its texture, coarse or fine, or gradations in between. This test is often called ‘field texturing,’ ‘hand texturing’ or ‘feel analysis.’ A guide to performing a feel analysis is provided by USDA-NRCS (USDA-NRCS, website link). Gee (2005) warns that the expressions ‘light soil’ and ‘heavy soil’ are frequently used to characterize the general physical behavior of different soils. A coarse-grained, sandy soil tends to be loose, well-aerated, and easy to cultivate, therefore it is called a ‘light soil.’ A fine-textured soil tends to absorb water easily and becomes plastic and sticky when wet, and tight, compact, and cohesive when dry, therefore it is called a ‘heavy soil.’ These terms are misleading, since coarse-textured soils are generally denser, and have smaller total porosity than fine-textured soils, and thus are heavier rather than lighter per unit volume. Table 1 shows the range of densities and porosities that are typical for soils of various dominant textures. These values can vary, however, within a wide range, and it is noted that soil use and management can influence them substantially in both the short and long term.

Quantitatively, soil texture refers to the relative proportions of various sizes of soil particles, sometimes called ‘size separates,’ which is the main subject of this chapter.

Soil particle size measurements

The size of soil particles can vary over 6 orders of magnitude, ranging from large stones and rocks (greater than 250 mm in size) to tiny, submicron clays (less than 0.0001 mm). The shape of soil particles also varies widely. Particle shape is often an indication of the mineral assemblage and the degree of soil weathering that has taken place. For beach sand, composed of quartz, particles are uniformly rounded and almost spherical. In contrast, most clay materials are far from spherical, many being platelets, looking more like playing cards (e.g., kaolinite) or tubes (e.g., halloysite) than marbles. In spite of the variation in shape, soil particles are sized according to their ‘effective’ particle diameter. Because of the huge size range, no single method is optimal to measure all particle diameters. The analysis is most commonly referred to as analysis of particle-size or grain size distribution. Size measurements depend on effectively separating the soil into primary particles. Methods for soil analysis consist of disaggregation by washing, exposing the specimen to vibration (physical or by ultra-sound), or otherwise dispersing the soil into single units. Typically, any larger stones (cm size range) are measured directly by a tape, a ruler, or a caliper. The stony, gravelly and sandy fraction—consisting of particles in the mm size range down to approximately 0.05 mm are generally sieved, i.e., washed through a nest of screens over a given period of time, while the proportion of particles smaller than 0.05 mm is determined by sedimentation or techniques based on light-scattering.

Table 1 Typical bulk densities and associated pore system properties for soils of various textures (Gee, 2005).

Description	Bulk density (g cm^{-3})	Porosity (cm cm^{-3})	Void ratio (cm cm^{-3})
Clay	1.20	0.54	1.17
Loam	1.40	0.49	0.96
Sand	1.60	0.39	0.64

In soil science, by convention, only the fraction that consist of particles smaller than 2 mm are considered for the expression of soil texture. This is reflected in most, if not all textbooks and measurement protocols, by stating that the soil should be sieved to 2 mm prior to further analysis. The fraction coarser than 2 mm is then considered as a gravel or stone fraction, depending on the size of particles and the classification system used. It is noted that some historical data from some Eastern European countries adhered to a system in which the limit between sand and the coarse fraction was either 1 or 3 mm, but they have now adopted the 2 mm system.

Sample pre-treatment for fine earth measurements

In natural soil, particles are bound by both organic and inorganic agents into various sizes of soil aggregates. Prior to the particle-size analysis, those binding agents must be removed to release the elemental particles from the aggregates. For tropical soils or soils with large organic, iron, or aluminum contents, special treatments are deployed to remove these binding agents and release the primary particles. Soils of volcanic ash origin are notorious for aggregation, and the chosen pre-treatment by chemical or mechanical dispersion can significantly affect the amount of clay released. In one test, researchers have demonstrated that the clay content (i.e., the sum of particles smaller than 0.002 mm) of a volcanic ash soil varies from 1% to 56% depending on the amount of pre-treatment imposed on the sample (Gee, 2005). Some tropical soils consist largely of clay-size particles yet behave much like sands, in terms of their water-storage and related physical properties, because of the high degree of natural aggregation.

The classical approach to sample pre-treatment is that the sample is stirred in water containing a dispersing agent, such as sodium hexametaphosphate or sodium pyrophosphate, which swamps the clay surfaces with sodium, causing dispersion of the clay materials and accelerating disaggregation. Prior to this step, laboratories also often remove soil organic matter using hydrogen peroxide, sodium hypochlorite or disodium peroxodisulfate (Mikutta et al., 2005), and remove the primary inorganic binding agent, calcium carbonate, using hypochloric acid. The applicable ISO standard (International Organization for Standardization [ISO], 2020) leaves space for interpretation of these steps, and since they can be time and labor-intensive processes, in which experience may also be important, their application in different laboratories may differ. *At the time of writing of this chapter*, the FAO Global Soil Laboratory Network (FAO GLOSOLAN) is working on a better defined standard operating procedure (SOP) to be recommended for both the sample pre-treatment and the measurement of soil particle size distribution using the pipette and hydrometer methods, which is expected to help fill this gap.

Methods, such as laser diffractometry, use a much smaller amount of soil sample than sedimentation methods, and the measurement is based on a different background theory. For this analysis, laboratories still use sodium hexametaphosphate for sample pre-treatment, often in combination with physical dispersion using ultrasonic agitation.

Frequently used measurement methods

The pipette and hydrometer methods

Most particle-size measurements on the fine soil fraction (i.e., silt and clay fraction) rely on sedimenting a soil suspension (Gee and Or, 2002). Stokes' law for viscous drag on a settling spherical object is combined with buoyant and gravitational forces to obtain the settlement rate. Combining forces and solving for the particle velocity yields:

$$v = d^2 g (\rho_p - \rho_w) / (18\eta) \quad (1)$$

where d is the equivalent particle diameter, g is gravitational acceleration, ρ_p is particle density, ρ_w is the solution's density, and η is the solution's viscosity.

Knowing settlement depths allows for a straightforward calculation of the effective particle diameter. In this understanding, the pipette technique involves taking physical samples from given depths and after given settlement periods, while the hydrometer method relies on estimates of the effective density of the suspension as particles of various size settle.

Fig. 1 shows the pipette method and Fig. 2 shows a typical hydrometer used for soil-suspension density measurements. This equipment is relatively inexpensive and readily available. For routine analysis in soil, sedimentation, and engineering laboratories, the pipette and the hydrometer are the standard tools for fine-fraction analysis. Various laboratories typically determine 3 to 8 points on the particle size distribution curve at various pre-determined cut-off points that are then used to determine the texture class of the soil and can be fitted with a lognormal or other curve type.

The integral suspension pressure (ISP) method

There has been interest in developing a measurement method that is based on the same theory, Stokes' law, but is more automated and fine-tuned than the pipette or hydrometer methods. Initiated by Nemes et al. (2002), a research group in Hungary introduced a prototype (named ASTA, Kovács et al., 2004, 2006). Later, the method was further refined into the *integral suspension pressure* (ISP) method and successfully commercialized by Durner et al. (2017) and by Durner and Iden (2021) after further upgrades.

The development of sensory technology in the 2000s allowed the detection of very small changes in pressure at a given point in a suspension as particles of different sizes settle over time. Coupled with that technology, an integral calculation was programmed into the measurement software that interprets the pressure signal into a quasi-continuous soil particle-size distribution curve. This interpretation is limited by the same assumptions (e.g., particle sphericity, independent settling) that also apply to the pipette and

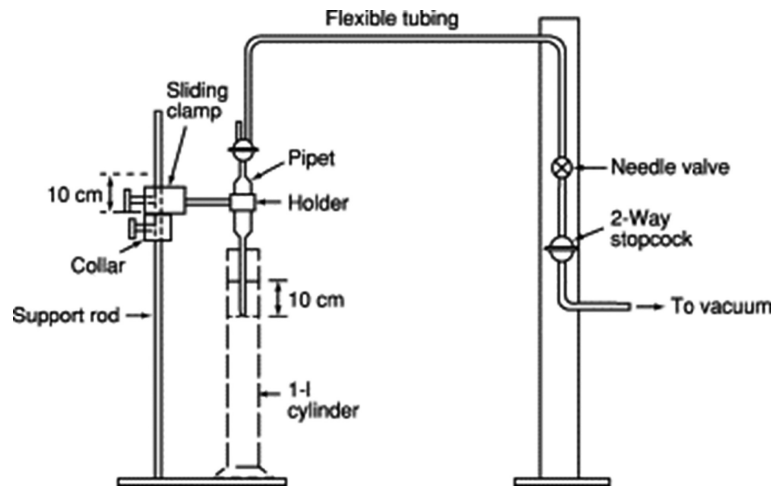


Fig. 1 Schematic of pipet apparatus and stand for sedimentation analysis of soil-particle size. Reproduced with permission from Syvitski JPM (Ed.) (1991) *Principles, Methods, and Application of Particle Size Analysis*. New York: Cambridge University Press.

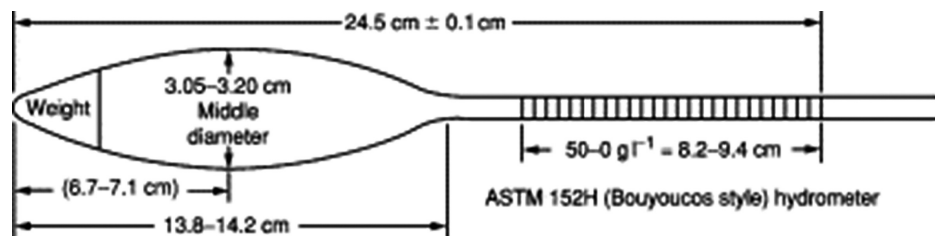


Fig. 2 Schematic of ASTM 152H-type hydrometer used for soil particle-size analysis. Reproduced with permission from Syvitski JPM (Ed.) (1991) *Principles, Methods, and Application of Particle Size Analysis*. New York: Cambridge University Press.

hydrometer methods. It is noted that for the full particle-size curve, the sand fraction still needs to be determined by sieving, since that fraction settles too quickly in a water-based suspension. When the measurement is initiated after shaking, sedimentation of the sand particles already starts in the first c.a. half a minute when the suspension is still in turbulent motion, and it is difficult to obtain a stable pressure signal. Nemes et al. (2020) tested and demonstrated some of the sensitivities of the method.

Light scattering, laser diffractometry (LDM)

In recent decades, the need for diagnostic tools in soil genesis and mineralogy have led to more sophisticated techniques being developed and tested. Several techniques have been investigated that are based on other theories, such as changes in an electric current or the scattering of a light beam by individual soil particles that are channeled through a sample cell in a dilute suspension.

Named after Wallace H. Coulter, the 'Coulter counter' technique is based on particles changing the electric current in the sample cell as they pass through. The technique has been used by numerous groups in soil and sediment research (e.g., Walker et al., 1974), but this technology has been more widely adopted in the medical field (e.g., hematology).

Laser-light scattering techniques, such as LDM, are gaining popularity in soil science to obtain size distributions of soil materials. Fig. 3 shows the primary components of a laser light-scattering device. Multiple laser diffractometer models are now on the market and soil laboratories worldwide make larger up-front investment costs by purchasing such devices that save on later measurement time and costs. Several technical aspects and limitations have been documented with respect to device settings, construction of dispersion units and the varied quality of homogenization of samples when different fractions are represented in different ratios, respectively, which are good to consult prior to working with the method (e.g., Ryżak and Bieganski, 2011; Sochan et al., 2012; Polakowski et al., 2015). Subsequently, Messing et al. (2021) and Svensson et al. (2022) highlighted the difficulties in comparing particle-size data obtained by different laser diffractometer models, hardware and software settings and the diversity in sample pre-treatments.

Compatibility of measurements using methods based on different measurement theories

Concerns have been raised about the compatibility of data obtained using different measurement techniques. Many studies have now reported that data obtained by laser diffractometry (LDM) and by sedimentation type methods (SM) do not compare well, and

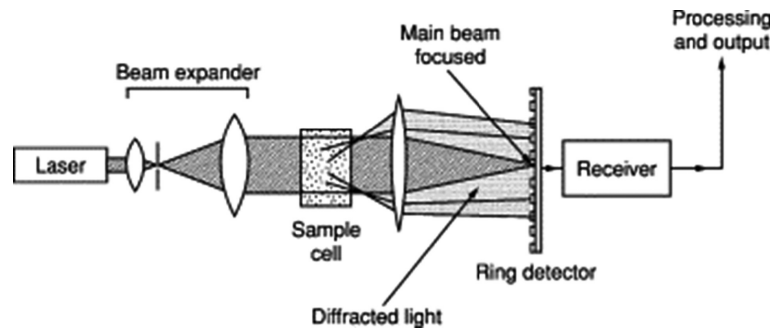


Fig. 3 Laser diffraction particle-size analyzer showing the primary components: light sources, sample, focusing lens, detector, and processing system. Reproduced with permission from Syvitski JPM (Ed.) (1991) *Principles, Methods, and Application of Particle Size Analysis*. New York: Cambridge University Press.

various conversion or adjustment options have been suggested (e.g., Makó et al., 2017; Yang et al., 2019; Svensson et al., 2022). Various reasons have been cited for such discrepancies, such as differences in the measurement theory, the representation of each particle fraction in dilute (LDM) or denser (SM) suspensions, that certain assumptions are invalid (e.g., sphericity of particles, independent motion of particles and the absence of Brownian-motion in the colloid phase). No agreement has yet been reached on the basis for data conversion.

The reason for the discrepancies in measurements lies in the combination of these cited factors. Soil particles are not spherical—often far from it—although the measurement methods assume that they are. This leads to uncertainties and biases in the subsequent calculations in different ways. In the sedimentation cylinder, non-spherical particles will be exposed to different drag on their sides and will tend to orient with the larger surface perpendicular to the direction of sedimentation, similar to how leaves fall from trees, and will not sediment steadily and according to their equivalent size. The same non-spherical particles provide different light-reflection patterns depending on their orientation at the time of passing the detection chamber driven by turbulent flow, and their sizes may therefore be misjudged. While the agreement in the sand and silt fraction appears to be rather linear, the discrepancies appear to be less well describable in the clay fraction, which suggests that additional factors, such as the presence of Brownian motion in the sedimenting suspension as well as potential differences in sample pre-treatment may play a role.

It is well documented that measurements by the LDM method and by the sedimentation type methods cannot be directly compared, so they need some kind of conversion or adjustment factor (Svensson et al., 2022). The same authors pointed out different approaches proposed for this in the literature. Typically, authors suggest an alternate reading point by LDM (e.g., a value in the 4–5 μm range) that should be understood as the reading that best approximates the clay content obtained by the conventional pipette method, but the recommended value varies study-by-study. These differences are most likely driven by differences in clay mineralogy of the soil samples (i.e., shape of the smallest particles), differences in hardware and their settings, and undoubtedly differences in sample pre-treatment. Finding an advanced conversion method will be necessary to avoid fragmentation in our methods and compare data obtained internationally. Without proper conversion, there is a risk of introducing bias to the obtained particle-size distribution curve (i.e., mostly underestimating clay content vs. the sedimentation type methods) that may further result in mis-classification of the soil in question, as demonstrated by, e.g., Makó et al. (2017), Yang et al. (2019) and Svensson et al. (2022).

It is technically possible to obtain data that appear to align well despite the use of a multitude of methods, such as sieving, sedimentation, and light scattering methods (e.g., Wu et al., 1993). It appears that with appropriate pre-treatments and method calibration, data compatibility can be achieved, but so far this remains a challenge in the fragmented practical laboratory work where different local or national standards, as well as cost- and work-time reduction efforts, frame some of the choices made when preparing samples or making measurements.

Compatibility of measurements using methods of similar measurement theories

Initial indications concerning the performance of the integral suspension method compared to the traditional sedimentation type methods are mixed. Acevedo et al. (2021) and Messing et al. (2021) reported sizeable biases between those methods, while Nemes et al. (2020) reported near-perfect alignment in the sand and silt fractions, but only poor correlations, although no substantial bias, in the clay fraction, despite the measurements having been performed in separate laboratories by separate staff. In the study of Nemes et al. (2020) the same sample pre-treatments were closely adhered to, whereas in the other two investigations, a variety of nationally recommended sample preparation methods were mixed with that recommended in ISP's documentation. This points toward the significance of sample pre-treatments when it comes to using ISP compared with other sedimentation type methods. While more supporting data should be collected, it is currently recommended that laboratories that change between sedimentation type methods keep following their regular sample pre-treatment steps to keep their new data *backward-compatible* with their historical data.

Advantages and disadvantages of different novel measurement techniques

When we compare data from ISP, LDM or any future technique, all such assessments and statistical evaluations typically assume that the data acquired using the pipette or hydrometer methods are correct and error-free. This is likely not the case, due, among other factors, to the assumed spherical shape rarely being met by natural soil particles. In addition, different laboratories often follow different sample preparation protocols that may result in different degrees of disaggregation, causing variation in the measurements. Nevertheless, *backward compatibility* with historical data is desirable, and hence for now, we make the practical assumption that those measurements are correct.

Measurement effort and cost

Beyond the compatibility of measurements, another key driver behind method development is the desire to increase the efficiency of data acquisition—i.e., the prospect of faster, easier, cheaper collection of data. To that effect, users have reported on the practical considerations needed to be taken into account when planning to operate one of the novel techniques.

LDM requires a substantial up-front investment, which is increasingly being made by laboratories in exchange for substantially cheaper operating costs—primarily in the form of saved work-time. Work-time is saved primarily because sample preparation is usually reduced by the highly automated nature of the technique. Moreover, using light-scattering techniques the solution does not need to sediment over time. However, users typically prefer to run replicate measurements. Because of its low running costs, the LDM technique is gaining popularity and its application may provide better *forward compatibility* with future data collections by many other users.

The ISP method is a computerized, semi-automated technique for PSD determination. It yields high resolution data with a distribution curve fitted to it. However, it still requires substantial human interaction and measurement time. Early indications are that available ISP versions appear to provide *backward compatible* data with conventional methods if the user applies the same sample pre-treatment, which is usually an elaborate process. Applying this technique requires the sedimentation of particles as conventional methods do, and it also requires the user to manually sieve the sand fraction and enter this data into the software, which are added work-time components. It is also noted that the ISP technique is sensitive to temperature or air-pressure fluctuations in the first minutes of the measurement (Nemes et al., 2020), that needs to be controlled by proper planning and potentially some minor investments to avoid a biased pressure signal propagating through the measurement. The technique, however, promises to provide the best option for both backward and international compatibility.

The added value of continuous data—International compatibility

A great advantage of both the ISP and the LDM methods is that the data obtained are nearly continuous and allows reading off the curve at any size limit without interpolations or post-processing required. If the only goal is to provide sand, silt and clay fractions according to a single classification system, a continuous PSD curve has minimal added value. However, with the growing internationalization of environmental and climate research, there is an ever increasing need to standardize the collection of new data, or at a minimum to make it easy to collect *internationally compatible* data. A continuous PSD curve promises to allow efficient direct interpretations according to different national and international systems simultaneously, without the need for elaborate interpolations that often hinder international collaborative efforts (e.g., Nemes et al., 1999). The complexity and some consequences of having to interpolate existing soil PSD curves is addressed below.

Particle size and textural classification schemes

Particle size schemes

For applications in agronomy and soil-science, the size-fractions of primary interest are those smaller than 2 mm due to their water retention characteristics. This texture range is usually represented by 3 main size separates, sand, silt and clay, where the sand and silt fractions are often further divided into sub-fractions (e.g., very fine, fine, medium, coarse and very coarse sand). However, different national and international systems exist for establishing the definitions and measurement. Fig. 4 presents an overview of several particle size measurement sequences in practical use in a number of European countries. The main system is recognized by FAO and the US Department of Agriculture (USDA). The system recognized by The International Union of Soil Science (IUSS) differs in the definition of the upper limit for the silt fraction (being defined at 0.02 mm by IUSS and 0.05 mm by FAO/USDA). Nevertheless, there are several other size cut-offs between silt and sand content in practical use world-wide, such as 0.06 mm in Scandinavia and Great Britain, or 0.063 mm being used dominantly in Germany. Note that the size separates have arbitrary size limits, and for the most part the differences between schemes appear to be relatively small.

Other systems have existed or still exist that make data comparison and conversion much harder if not impossible. As seen in Fig. 4, some countries defined soil clay content at a different limit (e.g., as <0.001 mm) to most others and international systems (i.e., <0.002 mm), but with proper interpolation—the topic being addressed below—an estimate of the 0.002 mm fraction can still be made. It is, however, a lot more challenging or close-to-impossible to give an estimate of any fraction according to the FAO-USDA system when the entire fine earth fraction was determined on sizes <1 mm, as was the case historically, e.g., in Poland (Ryzak and Bieganski, 2009) or Russia. Given that in such cases the 1–2 mm fraction was considered as ‘course fragments,’ it was discarded and therefore it would require extrapolation, rather than interpolation to achieve compatibility with more broadly used systems.

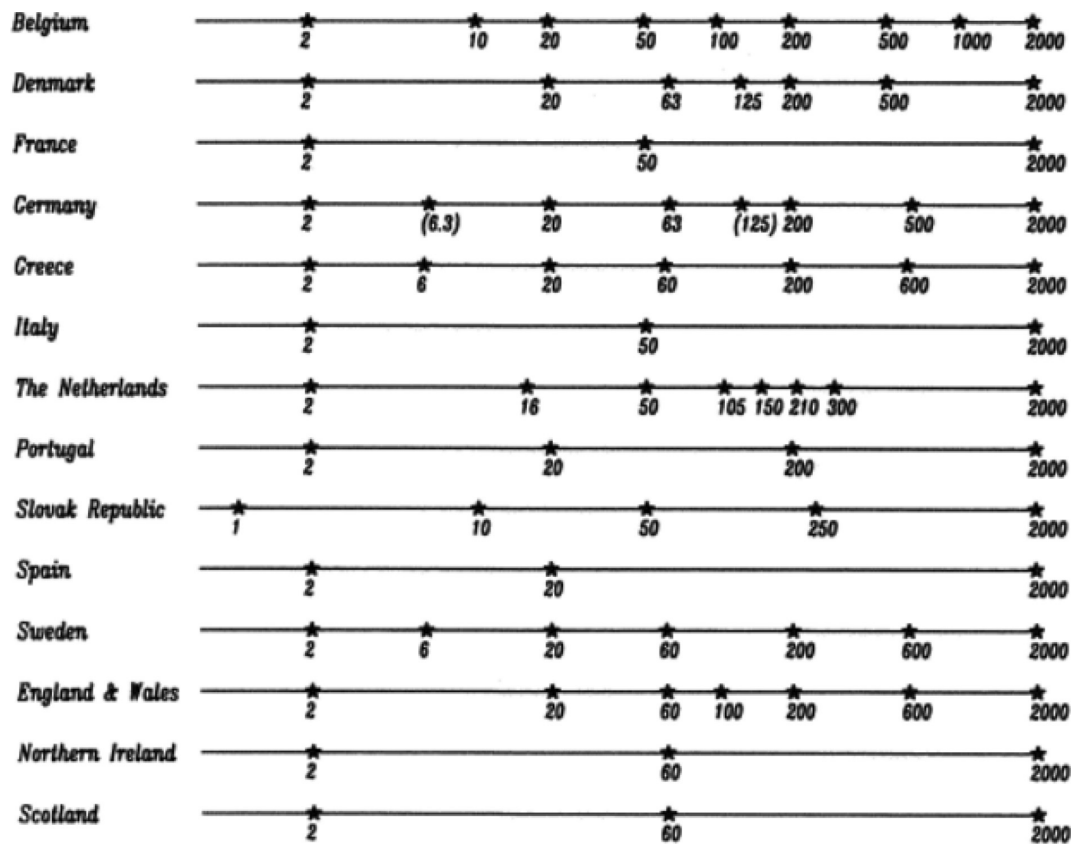


Fig. 4 Particle-size limits used by countries contributing data to the HYPRES database. Reproduced with permission from Nemes et al. (1999) *Geoderma* 90(3–4): 187–202.

Textural classification systems

Since soils rarely consist of a single size fraction, textural classification is based on different combinations of size separates. There have been dozens of different schemes proposed and used over the last ca. 150 years that are as diverse as the applications for which the textural data are used. Different disciplines often have their primary interest in a particular particle size range and present different practical or theoretical arguments to support their systems (Blott and Pye, 2012). Here we focus on the primary classification systems used to classify agricultural soils and later use these to demonstrate primary data conversion issues.

Fig. 5 illustrates four different soil texture triangles in current use in four data-rich countries. It is often not straightforward to convert the textural class of a soil from one system to another. For example, in the German system, silt content is defined as the fraction (weight percent) of particles in the size range of 0.002–0.063 mm, whereas in the other systems depicted in Fig. 5 it is defined as the fraction between 0.002 and 0.05 mm. A conversion of particle size data is necessary to convert between textural classification systems. Nemes and Rawls (2006) demonstrated that lack of conversion will increase the risk of introducing bias in any follow-up application, while an interpolation-based conversion introduces some random error but reduces the chances of bias in subsequent applications.

Conversion of data using interpolation

The internationalization of environmental research has presented a growing need to harmonize soil particle size data and present them according to a uniform textural classification system. Several researchers have tested various models (curve types) to fit part or the entire particle-size distribution curves (e.g., Hwang et al., 2002; Bayat et al., 2017) which could be used to make interpolations. In the work of Nemes et al. (1999) the authors tested several methods to achieve such a task for a large European database. The popular log-linear interpolation is simple to use but was proven unreliable when applied to a broad range of soils, especially when only the sand-silt-clay triplet was available. Among the tested methods, a free-form spline function and a k -nearest neighbor (k -NN) type pattern-recognition algorithm were better, more reliable alternatives (Fig. 6). The k -NN algorithm calculates the weighted average of the appropriate fraction from a limited number of similarly shaped curves in a large external database, and assigns it as the

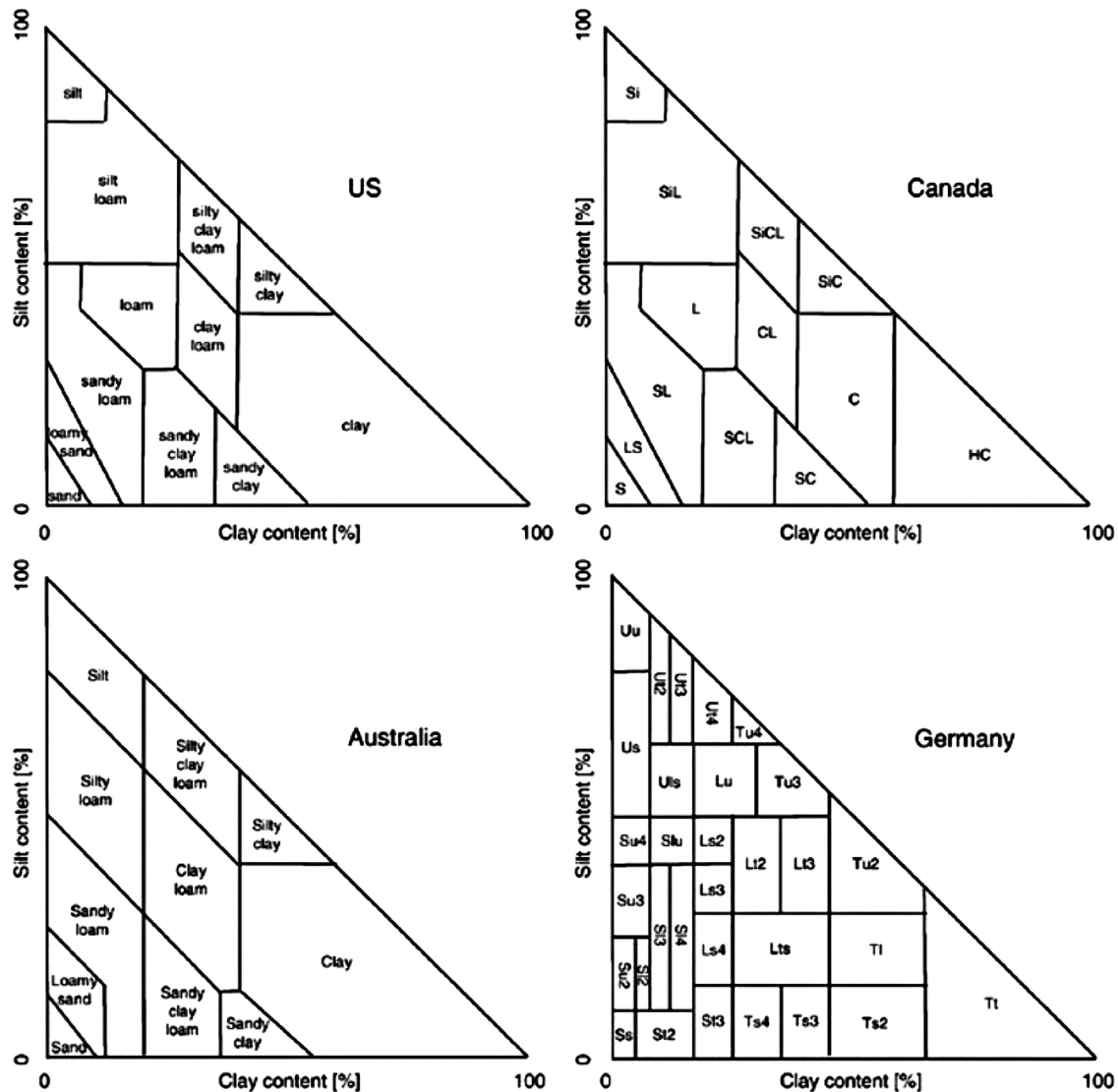


Fig. 5 Examples of different soil texture triangles used in four different countries. The US texture triangle is widely used in the World. Conversion between the US, Canadian, Australian and the FAO (not shown) systems requires only re-classification, while conversion to or from the German system requires data conversion or interpolation first. Moeys (2018—https://cran.r-project.org/web/packages/soiltexture/vignettes/soiltexture_vignette.pdf) provides R scripts to produce texture triangles based on different classification systems. Reproduced with permission from Bormann (2010) *Geoderma* 157: 142–153. DOI: <https://doi.org/10.1016/j.geoderma.2010.04.005>.

estimate. The approach has been successfully applied to harmonize particle-size distribution data in the European HYPRES (Wösten et al., 1999) and EU-HYDI (Weynants et al., 2013) databases, as well as in collaborative EU projects (e.g., H2020 OPTAIN—www.optain.eu). Blott and Pye (2001) and Moeys (2018—https://cran.r-project.org/web/packages/soiltexture/vignettes/soiltexture_vignette.pdf) offer conversion tools that can use various interpolation possibilities, however, along with Nemes et al. (1999), urge caution in the use of log-linear interpolation.

Irrespective of which method is used, the estimations become less reliable if the measured points are sparse. This adds uncertainty to the estimated size-fraction and as a consequence to the texture class that a sample falls into. However, the occurrence of random misclassification is seen to be less problematic than if textural classification is systematically shifted, due to not interpolating and converting data, to the appropriate system (Nemes and Rawls, 2006).

In some cases, the missing particle-size data point is at one of the ends of the distribution curve, in which case some type of extrapolation is required to fill the gap. Extrapolations beyond the measured data range should generally be avoided, however, in some cases that is the only option for integrating data into a larger data pool (e.g., that of a country into a region or continent). Should extrapolation be necessary, all efforts should be made toward testing the procedure and quantifying its uncertainty.

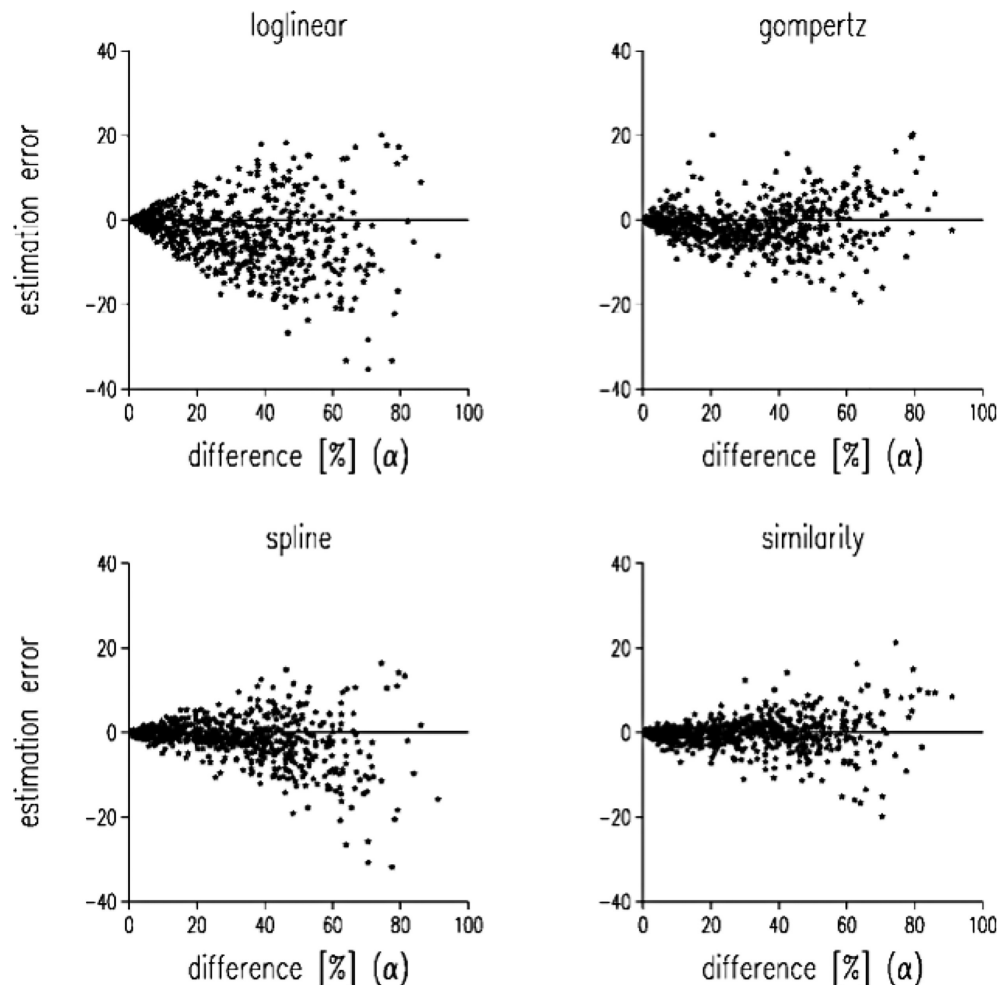


Fig. 6 Sensitivity of 4 interpolation procedures to data scarcity around the point to be interpolated. The tighter the points around the $Y = 0$ line, the more reliable a method is. The greater the gap between measured data points (X axis), the greater the interpolation uncertainty gets. The 'similarity' technique here and the 'k-nearest neighbor (k-NN) algorithm' referred to in the chapter text are identical. Reproduced from Nemes et al. (1999) *Geoderma* 90(3–4): 187–202.

Attempts at unification of soil particle-size measurements

It has been recognized that local solutions are necessary but not sufficient to resolve global environmental and climate issues. Research is therefore being internationalized, and the lack of a unified system to measure basic soil data has been identified as posing a challenge to international scale environmental research.

As summarized by [Blott and Pye \(2012\)](#), as early as 1962, a number of committees and working groups had been established in an attempt to reach an agreement on a common system of particle size classification. The Soil Science Society of America (SSSA) established a 'Particle Size Distribution in Soils' committee, while the Society of Economic Paleontologists and Mineralogists (SEPM, now the Society for Sedimentary Geology) appointed an 'Inter-Society Grain Size Study' committee. A series of meetings and workshops were held, with representatives from several societies, including SSSA, SEPM, ASTM, the Geological Society of America, the Highway Research Board, AASHTO and others. Despite all the efforts, with so many organizations involved it is perhaps not surprising that an agreement was not reached. Each organization had already too much invested in their respective classification systems by that time. SSSA and SEPM published their own position papers on the issue ([Committee on Particle Size and Distribution, 1967](#); [Tanner, 1969](#)). The latter, presenting the opinion of sedimentologists, outlined strong support for the phi scale, citing a number of major advantages: geometric basis, equal spacing, simple nomenclature, fine subdivisions, wide range of sizes and wide acceptance. Neither this direction, nor the use of the geometric mean particle diameter and its standard deviation gained traction in agricultural and environmental soil science beyond their occasional use ([Shiozawa and Campbell, 1991](#)). The latter approach assumes log-normality of the data and can work across measurement systems.

More than half-a-century later, the situation has not changed much: growth in the number of available measurement alternatives, sample pre-treatment and equipment setup options outpaces efforts to reach consensus over measurement standards and

protocols. Both the European SOPHIE initiative (<https://www.wur.nl/en/article/soil-program-on-hydro-physics-via-international-engagement-sophie.htm>) and FAO's Global Soil Partnership Network (FAO GLOSOLAN, <https://www.wur.nl/en/article/soil-program-on-hydro-physics-via-international-engagement-sophie.htm>) are engaged in developing common data acquisition protocol(s) for soil physical measurements, among them for obtaining soil particle-size distribution data.

The quest is challenging in multiple ways. As seen from the experience in the 1960s, investments made in their measurement standards and extensive data collections prove too important to each society or country to agree to changes, and risk losing *backward compatibility* to existing data. It is important to keep in mind, however, that some novel methods may prove to be sufficiently cost efficient, and may become the gold standard in the future—with or without providing compatible data today.

Use of soil texture data to infer other soil properties

Soil texture is one of the most fundamental properties of both the saturated and the unsaturated zones of the Earth's crust. It is one of the first properties that is measured and that has been mapped worldwide, despite the several different particle size scales and size class terminology schemes recognized in, e.g., sedimentology, geomorphology, soil science, ecology and civil engineering.

The significance of soil texture in determining other soil properties comes from the large size differences in the different fractions. For example, by definition, a sand particle is 100 s or 1000s of times greater in diameter than a clay particle, which makes it simple to depict that the size of the pore space between neighboring particles can differ by up to several orders of magnitude, and the surface area of the same mass of clay can be about 10,000 times more than that of the same mass of sand. Add to this that different clay mineral types have different shapes, specific surface area, and chemical reactivity to appreciate that the type and abundance of the size fractions will determine most of the soil's chemical properties and its response to changes in its environment.

Another important factor in the soil's functionality is its ability to store and exchange air and fluids (i.e., water, solutes, pollutants). The size distribution of particles has long been understood to influence the sizes of pores in the soil. Understanding physical laws such as Darcy's and Poiseuille's laws of flow or the Young-Laplace equation of capillary forces (see relevant other chapters) it should be easy to deduce that soils of different pore systems—both in terms of sizes and their complexities—will result in media that have different conductive and storage properties for both air and organic or inorganic fluids. It is not easy to make translations, such as how much clay it takes to equal a unit of silt or sand, or to determine to what extent different fractions have an effect on a particular soil property of interest. In addition, factors already mentioned like mineralogy, chemical and biological properties as well as the aggregation of particles complicate matters to the point that the use of particle size data to infer other soil properties often requires both (1) a good degree of empiricism and (2) the use of additional soil information. Therefore, it should not be too surprising that texture data alone do not generally relate directly to other fundamental physical properties.

Inference to pore-size distribution

It has been perceived for a long time that most soils exhibit log-normal PSD characteristics. Much has been published about the fractal nature of soil PSD and the possibility that the fractal dimension of PSD is related to an equivalent pore-size distribution that is then subsequently used to obtain estimates of hydraulic properties from those parameters (Gee, 2005). It has, however, been shown that many soils are actually multifractal, i.e., they do not present the same fractal dimension over the entire particle-size range, and therefore this direct linkage between soil PSD and its hydraulic properties has lost traction.

Nevertheless, the PSD curve still constitutes one of the basic soil characteristics and are useful toward diagnosing the origin of a soil, or in the empirical estimation of numerous physical, chemical and hydraulic soil properties. Since types of particle associations typically come with dominantly predictable pore-size distributions as well, the PSD curve is still useful as a proxy to describe pore system properties and predict fluid retention and conductivity properties of the soil—even if in a less direct sense.

Use of texture data to estimate soil hydraulic properties

Soil scientists have used soil texture information to make qualitative judgments about a number of physical properties for a long time, but to use texture for quantitative inferences became main-stream only decades later.

One of the main uses of soil texture information in agronomy and environmental soil science is through inference to soil hydraulic properties—primarily to soil water retention and hydraulic conductivity that are required to parameterize process-based simulation models concerned with, e.g., agro-hydrology, crop production, erosion processes, water and nutrient transport in the landscape, and land-surface processes at the regional to global scales. Since the late 1980s, the umbrella term *pedotransfer function* (PTF) is used to refer to such inferences. There is rich literature on this subject summarized in a compendium and extensive review by Pachepsky and Rawls (2004) and Van Looy et al. (2017) respectively.

There are many types of PTFs distinguished in the literature based on their inputs, outputs, underlying algorithms, etc. Here we highlight the two most basic PTF types, i.e., class and continuous PTFs. A class PTF is essentially an average of some soil property within a group of soils, typically a texture class, while a continuous PTF utilizes detailed particle-size data of each individual PSD curve separately. A class PTF returns the same estimate to each sample within the same class. It is noted that because the allocation of texture classes is rather arbitrary, there is no guarantee that such class boundaries are optimal for the functional grouping of soils. In fact, this approach has been challenged by Bormann (2010) and Twarakavi et al. (2010) who both presented alternate 'classes' that provide a more effective grouping by the hydrological functioning of soils. Continuous PTFs work with sand, silt and clay

contents of each sample and return different estimates for different soils, but it is easy to depict from the literature that more successful PTFs use not only PSD but additional inputs, such as soil bulk density or organic matter content, among other options.

Here we repeat that soil texture or PSD information is the first and most basic input to those estimation functions. It is noted that the notion behind pedotransfer functions is that they relate *available or easy to measure* soil physical properties to other, *difficult to measure*, soil properties, such as soil hydraulic properties. However, today the complexity and effort it takes to obtain reliable and comparable PSD data is underappreciated. When such estimation functions originated, mostly in the 1980s, there was a large pool of conventionally measured soil texture and particle size data available for exploration, and textbook knowledge was followed up by the successful derivation of estimation functions using soil texture or PSD as their input. Since then, limitations in the information content of soil texture and PSD have been explored, while their determination in practice has seen unprecedented diversification in methodology. Further progress in PTF-related international research is expectable only if we succeed to harmonize existing data and standardize measurement methodologies for future measurements to achieve both backward and forward data compatibility.

Limited success in estimating hydraulic conductivity

It has been increasingly recognized that PTFs have chronically failed to give reliable estimates of soil saturated hydraulic conductivity. This functional soil property is among the most variable properties of soil—if not the single most variable one—both spatially and temporally. It is mostly driven by the unique architecture of pores, especially macro and meso-pores, which are driven primarily by soil structural properties, rather than soil texture. Indeed, aggregation of the inorganic and organic constituents of soil drives the formation of the soil's 'expressed' pore architecture, which can then be modified on varied time scales by natural factors, such as rainfall, the activity of soil fauna as well as manifold human-induced interventions, including tillage, compaction, crop management, application of amendments. Yet the measurable PSD and soil texture does not change. The spatial arrangement of soil solids (soil structure) is the first thing that is destroyed during sample preparation for soil texture analysis, leaving the user with only limited information on the pore-size distribution of the soil. The fine, micron to sub-micron scale pore-system is rather predictable from the soil's clay and fine-silt content, but that information only explains part of the unsaturated functionality of the soil, but adds very little to its saturated hydraulic conductivity, which is driven by the structural (10s of microns to mm and cm scale) pore system's properties (e.g., its abundance, connectivity, tortuosity). Since the broad, quantitative representation of soil structure is still unresolved, the estimation of this property on the basis of soil particle-size data remains a challenge.

The tumble effect of misrepresenting soil texture

As noted in previous sections, the choice of PSD measurement method, choice of sample-preparation and measurement protocol, as well as the need for data interpolation all introduce certain degrees of uncertainty. Part of that uncertainty can be random, but very often a notable—often strong—systematic component (bias) is seen. The primary risk of systematic deviations in the representation of PSD is the misrepresentation or misclassification of soils by their texture. Examples have been shown by [Yang et al. \(2019\)](#), [Acevedo et al. \(2021\)](#) and [Svensson et al. \(2022\)](#) for cases when textural misclassification occurs as a result of differences in measurement theories or sample preparation choices, and by [Nemes and Rawls \(2006\)](#) who studied the case when misrepresentation is a result of lack of data interpolation and the resulting misuse of textural categorization.

If soil texture is systematically misrepresented, resulting estimations will also be misrepresented in a domino-like effect. This can be problematic both in small scale studies in which fine differences in flow processes are studied, as well as in large scale (catchment to global scale) studies that use class PTFs to represent soil hydraulic properties. In the latter case, systematic under or over-estimation of the properties for certain soil types may result in the misjudgment of their capacity to accept infiltrating water over large areas. It is therefore essential that textural biases are avoided, even if at the expense of introducing some random error through any data interpolation or conversion.

Other uses of soil texture

Climatic effects, such as seasonal freezing-thawing or frequent wetting and drying cycles, biological activity as well as frequent human influence, are rare in deeper soil layers that are of concern for some engineering purposes, or in layers that are already sealed. Therefore, inherent fluid retention and conductivity properties in deep, naturally textured or once-engineered soil layers will show a stronger correlation between their texture and hydraulic behavior. This is used when a soil's functionality is determined for engineering purposes, such as in the construction of earthen-dams, levees, landfills, or buildings. One of the primary soil properties of concern is the drainage capacity, whether its presence or absence is desirable. Large drainage capacity will help remove volumes of water from the surface and subsurface and contribute to good aeration, but will pose elevated pollution risks to the subsurface and groundwater resources. Soils with very little drainage capacity will serve well as 'impermeable' insulation layers, but aerobic biological processes (e.g., bio-remediation, organic decomposition processes) will be strongly hindered. One other important pool of soil properties of interest that are related to soil texture are its mechanical properties, such as precompression stress and shear-strength. Urbanization, and the general tendency of a growing amount of construction on reclaimed land is in conflict with the recent recognition of how vital our soil resources are for the future. Soil protection goals call for measures to be taken at construction sites, the extent of which depends on the knowledge of soil mechanical properties, which are primarily driven by its texture and actual moisture status.

Summary

Since particle-size distribution and soil texture are considered to be among the most fundamental soil properties, much effort has been invested since the early 1900s to measure and describe it and later to develop methods that relate it to a variety of other soil properties or characteristics like soil hydraulic properties. Yet, its representation is still problematic. I have reviewed the measurement methods predominantly in use, discussed the diversity in representing particle-size distribution as well as soil texture internationally, and argued the need to harmonize soil texture data internationally and reviewed the available options. The particle size distribution obtained from a laboratory analysis is almost always dependent on the method used, as well as the mechanical and chemical pretreatment that involves arbitrary decisions. As a result, we do not really know the absolute particle size distribution for a given soil. Nevertheless, sustained efforts are necessary to convert existing data between classification systems to avoid biases in the representation of soil texture and its subsequent applications. Despite known hindrances, the communities concerned with the representation of soil texture should work toward unifying data acquisition and subsequent classification standards. To this end, novel methods that can return quasi-continuous particle-size data may prove essential in the future.

References

- Acevedo SE, Contreras CP, Ávila CJ, and Bonilla CA (2021) Testing the integral suspension pressure method for soil particle size analysis across a range of soil organic matter contents. *International Agrophysics* 35(4): 357–363.
- Bayat H, Rastgou M, Nemes A, Mansourizadeh M, and Zamani P (2017) Mathematical models for soil particle-size distribution and their overall and fraction-wise fitting to measurements. *European Journal of Soil Science* 68(3): 345–364.
- Blott SJ and Pye K (2001) GRADISTAT: A grain size distribution and statistics package for the analysis of unconsolidated sediments. *Earth Surface Processes and Landforms* 26: 1237–1248.
- Blott SJ and Pye K (2012) Particle size scales and classification of sediment types based on particle size distributions: Review and recommended procedures. *Sedimentology* 59(7): 2071–2096.
- Bormann H (2010) Towards a hydrologically motivated soil texture classification. *Geoderma* 157(3–4): 142–153.
- Committee on Particle Size and Distribution (1967) Considerations relative to a common particle size scale of earthy materials. *Soil Science Society of America Proceedings* 31: 579–584.
- Durner W and Iden SC (2021) The improved integral suspension pressure method (ISP+) for precise particle size analysis of soil and sedimentary materials. *Soil and Tillage Research* 213: 105086.
- Durner W, Iden SC, and von Unold G (2017) The integral suspension pressure method (ISP) for precise particle-size analysis by gravitational sedimentation. *Water Resources Research* 53(1): 33–48.
- Gee GW (2005) *Texture (No. PNNL-SA-38359)*. Richland, WA (United States): Pacific Northwest National Lab.(PNNL).
- Gee GW and Or D (2002) 2.4 Particle-size analysis. *Methods of soil analysis. Part 4*, vol. 598, 255–293.
- Hwang SI, Lee KP, Lee DS, and Powers SE (2002) Models for estimating soil particle-size distributions. *Soil Science Society of America Journal* 66(4): 1143.1150. <https://doi.org/10.2136/sssaj20>.
- International Organization for Standardization (2020) Soil Quality—Determination of Particle Size Distribution in Mineral Soil Material—Method by Sieving and Sedimentation (ISO Standard no. 11277:2020). <https://www.iso.org/standard/69496.html>.
- Kovács B, Czinkota I, Tolner L, and Czinkota G (2004) The determination of particle size distribution (PSD) of clayey and silty formations using the hydrostatic method. *Acta Mineralogica-Petrographica* 45: 29–34.
- Kovács B, Czinkota I, Tolner L, and Czinkota G (2006) Fit method for calculating soil particle size distribution from particle density and settling time data. *Agrokémia és Talajtan* 55(1): 295–304.
- Makó A, Tóth G, Weynants M, Rajkai K, Hermann T, and Tóth B (2017) Pedotransfer functions for converting laser diffraction particle-size data to conventional values. *European Journal of Soil Science* 68(5): 769–782.
- Messing I, Soriano AM, Svensson DN, and Barron J (2021) *Soil Particle Size Distribution—Comparison Between Laser Diffraction, Integral Suspension Pressure and Sedimentation (Pipette) Methods*, EGU General Assembly 2021, online, 19–30 Apr 2021 EGU21–10869. <https://doi.org/10.5194/egusphere-egu21-10869>.
- Mikutta R, Kleber M, Kaiser K, and Jahn R (2005) Organic matter removal from soils using hydrogen peroxide, sodium hypochlorite, and disodium peroxodisulfate. *Soil Science Society of America Journal* 69(1): 120–135.
- Nemes A and Rawls WJ (2006) Evaluation of different representations of the particle-size distribution to predict soil water retention. *Geoderma* 132(1–2): 47–58.
- Nemes A, Wösten JHM, Lilly A, and Voshaar JO (1999) Evaluation of different procedures to interpolate particle-size distributions to achieve compatibility within soil databases. *Geoderma* 90(3–4): 187–202.
- Nemes A, Czinkota I, Czinkota G, Tolner L, and Kovács B (2002) Outline of an automated system for the quasi-continuous measurement of particle-size distribution. *Agrokémia és Talajtan* 51(1–2): 37–46.
- Nemes A, Angyal A, Makó A, Jacobsen JE, and Herczeg E (2020) *Measurement of Soil Particle-Size Distribution by the PARIO Measurement System: Lessons Learned and Comparison With Two Other Measurement Techniques*. EGU General Assembly 2020, Online, 4–8 May 2020, EGU2020-9832. <https://doi.org/10.5194/egusphere-egu2020-9832>.
- Pachepsky Y and Rawls WJ (eds.) (2004) *Development of Pedotransfer Functions in Soil Hydrology*. In: vol. 30. Elsevier.
- Polakowski C, Ryzak M, Bieganski A, Sochan A, Bartmiński P, Dębicki R, and Stelmach W (2015) The reasons for incorrect measurements of the mass fraction ratios of fine and coarse material by laser diffraction. *Soil Science Society of America Journal* 79(1): 30–36.
- Ryzak M and Bieganski A (2009) Metody wyznaczania rozkładu granulometrycznego gleb mineralnych. (Methods of determining the grain size distribution of mineral soils). In: *Acta Agrophysica* 175. Lublin, Poland: Institute of Agrophysics (IPAN). (In Polish).
- Ryzak M and Bieganski A (2011) Methodological aspects of determining soil particle-size distribution using the laser diffraction method. *Journal of Plant Nutrition and Soil Science* 174(4): 624–633.
- Shiozawa S and Campbell GS (1991) On the calculation of mean particle diameter and standard deviation from sand, silt, and clay fractions. *Soil Science* 152(6): 427–431.
- Sochan A, Bieganski A, Ryzak M, Dobrowolski R, and Bartmiński P (2012) Comparison of soil texture determined by two dispersion units of Mastersizer 2000. *International Agrophysics* 26(1).
- Svensson DN, Messing I, and Barron J (2022) An investigation in laser diffraction soil particle size distribution analysis to obtain compatible results with sieve and pipette method. *Soil and Tillage Research* 223: 105450. <https://doi.org/10.1016/j.still.2022.105450>.
- Tanner WF (1969) The particle size scale. *Journal of Sedimentary Petrology* 39: 809–812.
- Twarakavi NK, Šimůnek J, and Schaap MG (2010) Can texture-based classification optimally classify soils with respect to soil hydraulics? *Water Resources Research* 46(1).

- Van Looy K, Bouma J, Herbst M, Koestel J, Minasny B, Mishra U, Montzka C, Nemes A, Pachepsky YA, Padarian J, and Schaap MG (2017) Pedotransfer functions in Earth system science: Challenges and perspectives. *Reviews of Geophysics* 55(4): 1199–1256.
- Walker PH, Woodyer KD, and Hutka J (1974) Particle-size measurements by Coulter Counter of very small deposits and low suspended sediment concentrations in streams. *Journal of Sedimentary Research* 44(3): 673–679.
- Weynants M, Montanarella L, Tóth G, Arnoldussen A, Anaya Romero M, Bilas G, Børresen T, Cornelis W, Daroussin J, Conceição Gonçalves M, Haugen L-E, Hennings V, Houskova B, Iovino M, Javaux M, Keay CA, Kätterer T, Kværnø SH, Laktinova T, Lamorski K, Lilly A, Mako A, Matula S, Morari F, Nemes A, Patyka NV, Romano N, Schindler U, Shein EV, Sławiński C, Strauss P, Tóth B, and Wösten JHM (2013) *European Hydropedological Data Inventory (EU-HYDI)*. JRC Technical Reports EUR 26053 EN. 1831-9424. ISBN: 978-92-79-32355-3167. <https://doi.org/10.2788/5936>.
- Wösten JHM, Lilly A, Nemes A, and Le Bas C (1999) Development and use of a database of hydraulic properties of European soils. *Geoderma* 90: 169–185.
- Wu Q, Borkovec M, and Sticher H (1993) On particle size distributions in soils. *Soil Science Society of America Journal* 57: 883–890.
- Yang Y, Wang L, Wendroth O, Liu B, Cheng C, Huang T, and Shi Y (2019) Is the laser diffraction method reliable for soil particle size distribution analysis? *Soil Science Society of America Journal* 83(2): 276–287.

Relevant websites

- <https://www.fao.org/global-soil-partnership/glosolan/en/>—FAO Global Soil Partnership (GLOSOLAN): (Last Accessed 1 Oct 2022)
- https://cran.r-project.org/web/packages/soiltexture/vignettes/soiltexture_vignette.pdf—Moeys, J. 2018. The soil texture wizard: R functions for plotting, classifying, transforming and exploring soil texture data. (Last Accessed 1 Oct 2022)
- <https://www.wur.nl/en/article/soil-program-on-hydro-physics-via-international-engagement-sophie.htm>—Soil Program on Hydro-Physics via International Engagement (SOPHIE): (Last Accessed. 1 Oct 2022)
- https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/edu/?cid=nrcs142p2_054311—United States Department of Agriculture, Natural Resources Conservation Service (USDA-NRCS): Guide to Texture by Feel. (Last accessed: 1 Oct 2022).

Soil tilth

Thomas Keller^a, Douglas L. Karlen^b, and Paul D. Hallett^c, ^aAgroscope, Department of Agroecology & Environment, Zürich, Switzerland; ^bDlkarlen Consulting, LLC, Saint Paul, MN, United States; ^cSchool of Biological Sciences, University of Aberdeen, Aberdeen, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

This is an update of D.L. Karlen, TILTH, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 168–174, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00303-9>.

Introduction	48
History of tilth	49
Tilth of tilled and zero-tillage arable soils, and distinction from natural soil structure	50
Seasonal dynamics of soil structure in arable fields with tillage	51
Soil organic matter and tilth	52
Soil management effects on tilth	52
Tillage effects on soil tilth	53
Evaluating soil tilth in situ	53
Tilth indices	54
Summary	54
References	56

Abstract

Soil tilth describes the friability and physical quality of a soil for crop growth. Soil tilth is an old concept yet its direct quantification remains difficult. Tilth is characterized by measuring soil properties related to soil aggregates, pore structure, mechanical resistance and organic matter, or in the field using methods such as visual evaluation of soil structure. Soil tilth is positively influenced by management that increase soil organic matter content, and negatively influenced by soil compaction. Tillage to produce tilth can have positive or negative effects over time, the latter prompting shifts to reduced or no tillage.

Key points

- Soil tilth describes the structural state and physical conditions of a soil that support plant growth
- Soil management affects soil tilth through direct and indirect effects on soil structure
- Soil organic matter is a key factor for good soil tilth
- Soil tilth was traditionally associated with plowed soils, but zero-tillage has shown that it too can produce good tilth by biological processes and weathering
- Intensive tillage can degrade tilth if it depletes carbon and structural stability
- Soil tilth is characterized by measuring properties related to soil organic matter, aggregates, soil pore structure, and soil mechanical resistance
- Direct quantification of soil tilth remains challenging

Introduction

Soil tilth is a catch-all term used initially to describe tilled seedbeds and rooting zones, but applied more recently to reduced tillage systems, including zero-tillage that are not mechanically fragmented by farmers. From the look and feel of a soil, an agronomist in the field can gain a qualitative understanding of soil tilth from its ease of fragmentation, aggregated structure and the proliferation of roots. Signs of restricted root growth near the soil surface or shoot emergence would indicate poor soil tilth.

The scientific description of soil tilth is the physical condition of soil as related to its ease of tillage, fitness as a seedbed, and its impedance to seedling emergence and root penetration (Soil Science Society of America, 1996). Ease of tillage does not apply to zero-tillage soils, but the other properties do, so the approaches to measure tilth are interchangeable across soil management practices. To some extent, “soil tilth” could be used interchangeably with “soil physical quality” in agricultural soils, although measurements are typically limited to shallow depths where the impacts on seed germination, seedling emergence and early root growth are greatest. Soil tilth is an important qualitative indicator of soil physical condition, used to describe soil structure in terms such as mellowness or friability.

Many properties of a soil will affect its tilth. Soils with poor tilth are described as appearing lifeless, resembling brick or concrete, or being cloddy (Fig. 1). If soils with poor tilth have a high percentage of sand, they will often disaggregate or separate into their primary sand-, silt-, and clay-size particles, with little or no tendency to bind together. Soils with poor tilth can also be very massive



Fig. 1 Photos showing a deep gleyic Cambisol with loamy soil texture with good (left) and poor soil tilth (right). Photos: Bettina Koster, Agridea, Switzerland.

and difficult to till, seed, or work into a suitable seedbed. Soils with medium tilth are characterized as being somewhat cloddy with a tendency to “ball up” and be “rough pulling” (i.e., requiring greater energy to break up by cultivation) when working into a seedbed. Soils with good tilth are described as being crumb-like, easy to slice (like cutting butter), and spongy. Defining the descriptive terminology for soil tilth is a challenging task for those specializing in soil management. More specifically, tilth is difficult to quantify.

Whatever the difficulties over its definition, several qualitative approaches to the evaluation of the quality of tilth in the field have been developed using rapid spade assays, such as the Visual Evaluation of Soil Structure (Guimarães et al., 2011). Moreover, quantitative information on the structural state of soil can be obtained from measurements of soil mechanical, hydraulic and gas transport properties, typically involving sampling of intact soil cores (Groenevelt et al., 1984; Dexter, 1988; Kawamoto et al., 2006). With the widespread use of X-ray computed tomography imaging in soil science, it is now possible to visualize soil tilth in 3D and directly quantify soil pore structure (Peth, 2011).

History of tilth

The term ‘soil tilth’ appeared 1800 years ago in *Cato’s Roman Farm Management* as one of the most important aspects of agriculture “What is the first principle of good agriculture? To plow well. What is the second? To plow again; and the third is to manure.” However, even though the concept of soil tilth predating modern agriculture, a clear definition has never evolved because numerous objective and subjective perceptions have conveyed different meanings. For example, a 19th century childhood story explaining the concept of soil tilth, describes how in his dying moments an old man called his children to his side to tell them that if they would dig diligently in the garden, they would find a hidden treasure. The children did so, but alas, they found no treasure of silver or gold. The subsequent harvest from this much-worked garden, however, was so large that the father’s meaning gradually dawned upon his children and they realized the wisdom of his words. A 16th century description written to provide instructions for growing peas and beans tells the farmer how to determine if the soil is ready for planting. The 1523 book entitled “*Boke of Husbandry*” states that he or she should walk on the plowed ground and “if it synges or crye, or make any noise under thy fete then it is to[o] wete to sowe; and if it make no noyse, and wyll beare thy horses, thanne sowe in the name of God.” Good soil tilth was for a long time associated with (intensive) tillage, which explains some of the resistance toward conservation tillage and zero-tillage that still exists (Karlen et al., 1990). Although good tilth may be achieved by tillage, conservation agriculture has shown that it too can produce good tilth and that over time, more intensive tillage, such as plowing, can degrade tilth if it depletes carbon and structural stability.

The value of soil tilth was described and discussed in the journal *Science* about 100 years ago (Frierson, 1921; Alexander, 1922). By 1937, Yoder had already presented a comprehensive representation of soil tilth, and described key characteristics of favorable tilth including minimum mechanical resistance for roots, adequate water infiltration and aeration, promotion of microbial activity, as well as providing bearing capacity and traction for farm implements (Yoder, 1937). These aspects and trade-offs (e.g., “fluffy” soil vs. bearing capacity) describe well the state-of-the art of desired traits of managed soil structure (Or et al., 2021). The main challenge, then and now, is the quantification of tilth and the determination of target values, optimum ranges and thresholds of soil properties associated with favorable soil tilth. Moreover, Yoder (1937) also stated that soil tilth is dynamic – the quantification of soil structure dynamics still challenges scientists today (Or et al., 2021). Among the earliest scientific articles that mention “soil tilth” are Fenton (1930) who mentioned the benefit of “good soil tilth” for the formation of pasture, and Blair (1937) and Blair and Cashen (1938) who used the shape of soil compressibility curves as a measure to quantify soil tilth. In the 1930s, several authors began to describe how climatic conditions and various soil, crop, and animal management practices affected soil tilth. In the USA, much of this research was driven by the disastrous impacts of the ‘Dust Bowl’ in the 1930s. This early research sought to describe the structure and feel of the soil tilth that formed the seedbed, generally made up of fragments produced by the action of plowing. By sieving soil samples and separating the material into different sizes, they showed how freezing, drying winds, rainfall, tillage, animal compaction and other management factors affected tilth. Eventually, soil tilth became a “blanket” term describing all the soil conditions that determine the degree of fitness of a soil as an environment for the growth and development of a crop plant (Karlen et al., 1990; Hadas, 1997). Soil structure was identified as a key factor in determining soil tilth because it is a soil property that can be rapidly altered through tillage operations and changes in environmental factors such as rainfall or temperature.

Soil tilth therefore changes over a growing season due to the intensity and timing of tillage, and changes to soil physical structure by weather, plant roots and soil organisms. Over longer time periods, soil tilth and its seasonal dynamics may also change due to management practices. As virgin grassland or forest soils are cultivated, the organic matter concentrations, base (cation) saturation, and porosity decrease, but bulk density increases. These changes reduce the granulation or tendency of the soil to form stable aggregates and gradually deteriorate soil structure. Declining soil tilth is associated with increased runoff, erosion potential, need for tillage to prepare an adequate seedbed, and fertilizer to sustain reasonable and profitable crop yields.

Soil management research through the mid-20th century focused on identifying basic properties and processes that affect soil structure and tilth (i.e., flocculation, cementing by organic and inorganic colloids, wetting and drying cycles, freezing and thawing, organic matter cycling, biological activity, and tillage). There was a drive to optimize the environmental conditions of soil for seed germination and root growth. Although tillage to produce tilth was a central to such research, the role of organic matter, root growth and microorganisms was appreciated. Over time, tilth expanded beyond plowed seedbeds as intensive tillage was degrading vulnerable soils. In the 1970’s, Brazil combatted tillage induced soil degradation by pioneering zero-tillage and conservation agriculture approaches (de Freitas and Landers, 2014). Led by innovative farmers, government researchers (e.g., Brazilian Agricultural Research Corporation - EMBRAPA) and industry partners, the new practices developed not only improved the tilth of degraded soils, but also allowed agriculture to expand onto vulnerable soils in the Cerrado. This shows awareness of the importance of not only good soil tilth, but also the impacts of physically manipulating tilth on the sustainability of soil.

Significant progress has been made in understanding many of the scientific aspects of soil structure, organic matter cycling, activities of microorganisms, and effects of various soil and crop management practices. However, with respect to the adoption of improved soil management practices, economic pressures and political forces beyond the control of most farmers often contribute to declining soil tilth. Yoder (1937) posed the question whether using soils with poor tilth or regeneration of soils was more economical – a highly topical issue of today.

Tilth of tilled and zero-tillage arable soils, and distinction from natural soil structure

Although soil tilth is primarily associated with the soil structural state induced by tillage, the widespread use and increasing adoption of zero-tillage necessitates some considerations regarding the structural state of differently managed soils (e.g., plowed, reduced tillage, zero-tillage), and distinction between natural and managed soil structure (Or et al., 2021). In this context, we may use a broader definition of soil tilth as “the soil physical condition supporting plant growth.”

Soil structure of natural and managed soil are dominated by different processes (Fig. 2). The structural state and dynamics of soil in natural ecosystems is driven by feedbacks between biological (plant growth and decay, soil organisms) and abiotic processes (soil shrink-swell induced by drying and wetting, and freeze-thaw cycles), resulting in a soil characterized by biopores, desiccation cracks and soil aggregation. In contrast, tilth of managed, tilled soil is predominantly driven by mechanical fragmentation and subsequent consolidation and coalescence of fragments created by tillage. It is important to recognize the difference between aggregates formed by natural (e.g., biology, weather) processes, and tillage-induced “aggregates”, that are the result of a fragmentation process, and therefore these are better termed “fragments.” While aggregates and fragments could look similar in size and shape, their formation processes and ecological functions are different. The (annual) soil fragmentation by tillage still dominates the (annual) “life cycle” of managed soil tilth, and tillage-induced soil fragments characterize the structure and function of soil tilth in managed, tilled systems. Natural processes are not absent in managed soil, but play a less prominent role than in natural soil. Nevertheless, it has been shown that biopores created by earthworms and roots are more dominant in reduced tillage than in plowed systems. A second distinct difference between managed and natural soil is the vegetation and hence root abundance and distribution, as well as soil carbon inputs (roots, litter). Natural systems comprise of annual and perennial plants with more or less diverse above- and below-ground features, and vegetation with typically diverse root structures (plants with different root architectures and properties,

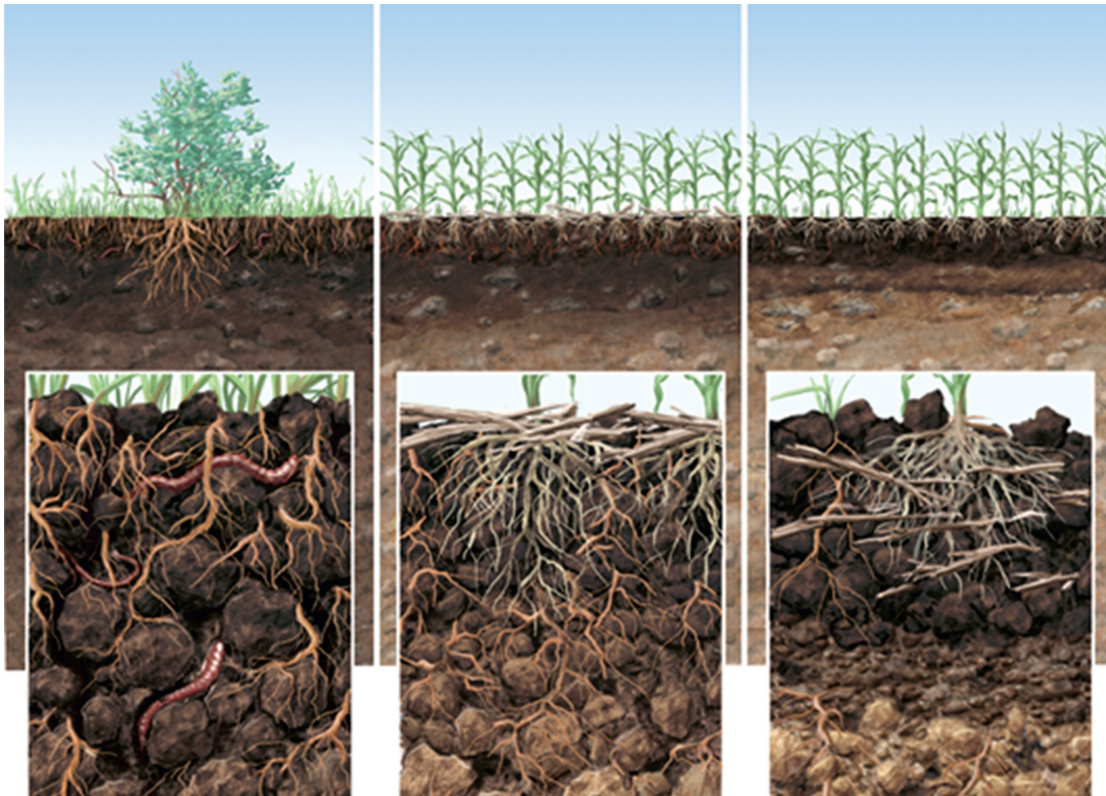


Fig. 2 Illustration of natural soil structure (left), and managed soil structure in a zero-tillage system (middle) and under conventional tillage (right). Natural soil structure is shaped by biopore and aggregate formation from soil fauna-vegetation-climate feedbacks. Tilled topsoil structure is largely governed by mechanical fragmentation and loosening, and subsequent consolidation. Note the distinct transition from the topsoil to the subsoil in conventionally tilled soil. Zero-tillage favors higher biological activity than conventional tillage due to the lack of mechanical disturbance but lacks the vegetation diversity of natural soil ecosystems. From Or D, Keller T, and Schlesinger WH (2021) Natural and managed soil structure: On the fragile scaffolding for soil functioning. *Soil and Tillage Research* 208: 104912.

and variable rooting depths). In contrast, plants on managed (arable) soils are typically annuals, which means: the growing season is limited (between sowing and harvest), parts of the biomass production and nutrients are removed from the soil (at harvest), and the diversity of plants is low (usually one main crop at a time, some weeds but these are usually not desired and repressed). Hence, even soil in zero-tillage systems that involve minimal mechanical soil disturbance to the depth of seeding (some disturbance in unavoidable due to coulters or disks on the seeder) develops a different tilth compared with natural soil, due to the differences in vegetation and removal of harvested plant parts in managed systems, irrespective of tillage system (Or et al., 2021). In addition, periodic mechanical stresses from farm vehicles may induce disturbance (i.e., compaction) in managed soil, irrespective of tillage system.

Seasonal dynamics of soil structure in arable fields with tillage

Annually tilled soil undergoes an annual cycle of fragmentation and loosening, and subsequent consolidation (Fig. 3). The topsoil (typically 0–0.2 m depth, but regional differences and traditions exist) is mechanically fragmented and loosened during primary tillage (there is a variety of tillage implements, including the widely used moldboard plow). The fragment size distribution produced by tillage is a complex function of implement properties, velocity, and soil conditions (i.e., texture, moisture and structure), and hence, prediction of how soil fragments during tillage is not straightforward. In conventional tillage systems, the top few centimeters (the depth depends on the crop to be sown) are further fragmented and fragments are sorted (e.g., larger fragments to the top) during seedbed preparation (secondary tillage, using a harrow or similar implement). Sowing involves again some mechanical disturbance of the soil, and frequently, the seedbed is compacted by rolling, also referred to as firming, after seeding. These three steps of what is typically referred to as conventional tillage (primary tillage, secondary tillage, sowing) can be done at three different occasions, or combined into one machinery pass – there is a variety of different tillage systems and implements. Moreover, certain crops require specific tillage operations, e.g., to create ridges (e.g., potatoes). The structure of tillage-loosened topsoil is unstable, and quickly consolidates, and the discrete fragments created by tillage coalesce over time (Sandin et al., 2018). The magnitude and rate of consolidation is dependent on the amount of precipitation, frequency and intensity of wetting-drying cycles, and stability of soil. Models to predict changes of certain features of soil structure (e.g., pore size distribution) following tillage exist (Or and Ghezzehei, 2002; Roger-Estrade et al., 2009; Chandrasekhar et al., 2019), but their parameterization remains challenging.

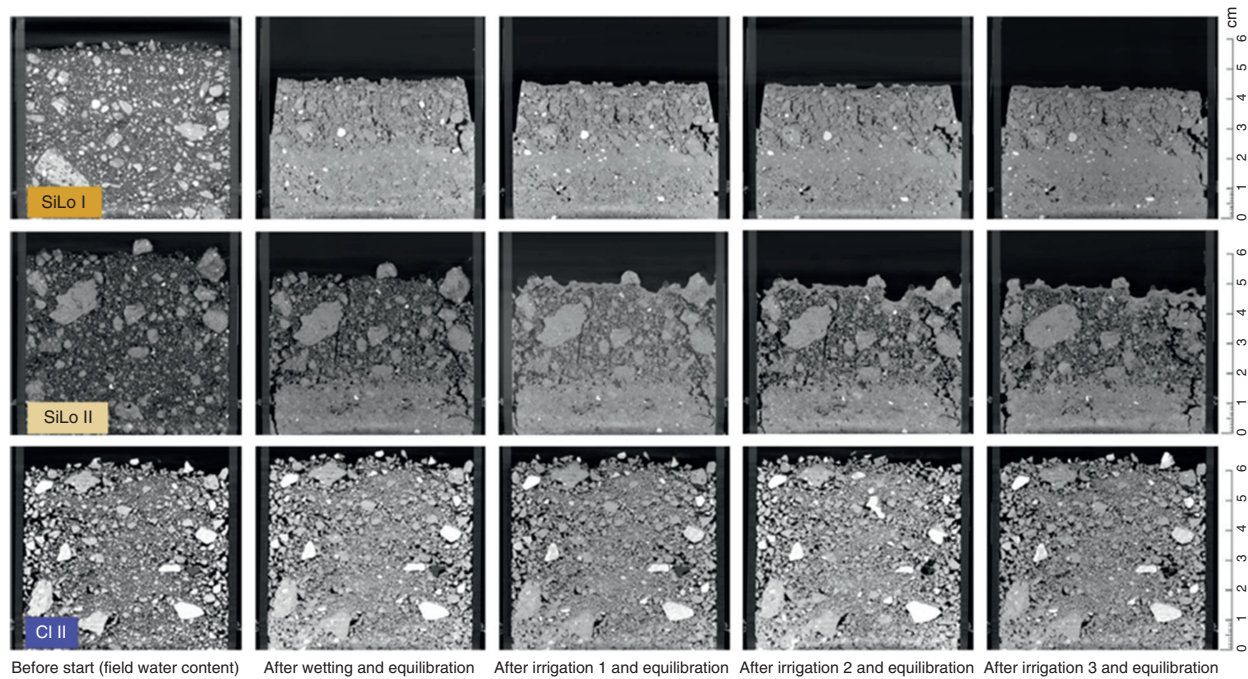


Fig. 3 Sequence of computed tomography images illustrating consolidation of tilled soil during wetting-drying cycles. In each cycle, the samples were wetted from above (rainfall simulation, 4 h at 5 mm h⁻¹) and then drained (i.e. equilibrated) to -30 hPa matric potential during at least 5 days. Soil textural classes are silt loam (22% clay; top images), silt loam (7% clay; center) and clay (54% clay; bottom). From Sandin M, Jarvis N, and Larsbo M (2018) Consolidation and surface sealing of nine harrowed Swedish soils. *Soil & Tillage Research* 181: 82–92.

Soil organic matter and tilth

Organic matter is one of the main factors affecting soil tilth because of its role in aggregation. Studies with disturbed and undisturbed soils show that as organic matter increases, bulk density decreases and there is more space for air and water exchange. The constant addition of organic matter is one reason that virgin grassland soils generally have excellent aggregation and structure. These soil characteristics occur because in humid regions grasses replace most of their roots and top growth each year. Also, when the above ground plant material dies, it falls on the soil surface where it is decomposed or mixed into the upper part of the soil by earthworms and other soil organisms. The dead roots and plant residue provide a food source for fungal and bacterial decomposition processes leading to the formation of soil aggregates, which modify effects of soil texture with regard to water and air relationships and root penetration. Increased water-stable aggregation, annual proliferation of plant roots, and the associated wetting and drying cycles gradually develop a soil structure that is stable and with good management generally resists degradation.

The regular addition of organic materials can improve tilth by creating larger and more stable aggregates that are not easily eroded by wind or water. Organic matter, both as residue on the soil surface and as a binding agent for aggregates near the surface, strongly influences water infiltration, retention, and the potential for particle detachment and erosion. Surface residues intercept rainfall or irrigation water and dissipate the energy before the drops hit the surface, where they can detach soil particles through the splash or begin to form a surface crust by filling the voids and cracks between the aggregates with fine particles. Organic materials on the soil surface also slow the water as it flows in surface runoff, thus increasing opportunities for the water to infiltrate. Another characteristic of soils with good tilth is the presence of biopores and pore space between the water-stable aggregates. These channels greatly enhance the ability of soil to conduct water from the surface into the subsoil, where because of good root proliferation (another characteristic of good tilth) plants can subsequently use the water to support growth and development.

These impacts of organic matter, biology and weather on soil tilth affect all soils, but they are especially important under zero-tillage. In addition to fuel and labor savings, zero-tillage can increase organic matter in the top few centimeters of soil, resulting in more water-stable aggregates, and a structure that accommodates rapid plant growth with adequate root proliferation. However, if compacted layers are not loosened by tillage root growth can be restricted, so biopores provide important pathways for roots to reach subsoils.

Soil management effects on tilth

Tilth, as an indicator of the physical condition of a soil, will be strongly influenced by any soil management practice that quantitatively affects soil structure, aggregation processes, or soil organic matter cycling. Incorporating perennial crops into

extended rotations, reducing tillage intensity and frequency, or including cover crops and optimizing nutrient and water management are among the most common practices that can be used to create the desired condition.

Surface mulch from crop residues, cover crops, leaves, or other organic amendments provides both physical protection from raindrop impact and a food source for earthworms and other soil fauna, and for soil microorganisms. The surface mulch also moderates soil temperature and moisture extremes at the soil surface. If soils are left bare and exposed, surface temperatures can be very high, causing the soil to become very dry. When this occurs, earthworms and other soil insects will move deeper into the soil leaving a surface zone that contains very few active organisms. Bacteria and fungi that live in thin films of water will die or become inactive, slowing the natural process of organic matter cycling and aggregation.

A range of management options exist to improve soil tilth, mainly by increasing the action of plant roots and surface mulches, through external inputs of organic matter such as manure or compost, diversified crop rotations, cover crops and incorporating temporary leys in the rotation, or by minimizing disturbance by tillage. Information on cover crops and soil tillage can be found in other articles in the Encyclopedia. Below the specific impacts of tillage to soil tilth are described.

Tillage effects on soil tilth

Although tillage improves tilth by loosening surface soil, disrupting crusts and creating a more favorable soil-water-air environment for plants, it can also be a primary cause for the long-term deterioration of tilth. The loosening process that often has positive short-term plant production benefits may create negative long-term effects because it may increase the rate of organic matter decomposition through chemical and microbial mineralization. This occurs because previously protected organic matter is exposed by the soil fragmentation induced during tillage and subsequently used as a food source by many microorganisms. Loss of organic matter decreases aggregate stability, increasing the potential for surface crusting, poor seedling emergence, runoff, erosion, and other indicators of poor soil tilth (e.g., decreased infiltration of rainfall or irrigation water, decreased soil water retention, and decreased nutrient cycling). Tillage itself or the wheel-traffic associated with management operations can further reduce tilth by increasing compaction. Wheel-traffic reduces total pore space and frequently increases water-filled pore space leading to reduced aeration, slower nutrient cycling, and a decreased volume of soil for plant root proliferation. These processes obviously occur in fields used for crop production, but they have similar negative effects on tilth at construction sites, in forests during logging operations, on campgrounds, athletic fields, or even desert areas where off-road recreational vehicles are allowed.

Tillage has been described as being addictive for soils because of the risk of a vicious cycle that can be established. Tillage provides improvements in tilth and nutrient cycling, but there is a risk that these periods of improvement become shorter and shorter as more of the organic matter is lost through oxidation. As a result of these interactions, monitoring soil tilth may be a useful indicator of the sustainability of an agricultural system. However, whether tillage results in soil degradation and loss of organic matter is dependent, among others, on timing of tillage operations and soil and climatic conditions. The change from a plowed to a reduced or zero-tillage system results in a stratification of soil organic carbon, with higher soil organic carbon concentrations near the surface in reduced or zero-tillage soil, but impacts on total carbon stocks are often small or negligible (Meurer et al., 2018).

In situations where nitrogen was the most limiting factor, increasing nitrogen application rates and therefore crop productivity has been shown to increase soil organic matter even with moldboard plowing. Development and adoption of reduced- or zero-tillage agricultural practices have the potential to improve soil tilth even more by slowing organic matter decomposition, maintaining aggregate stability, and preserving soil structure. However, impacts of reduced and zero-tillage on soil hydraulic properties and functions are not unambiguous, with a tendency toward less macroporosity but greater pore connectivity in zero-tillage compared with conventionally tilled soils (Wardak et al., 2022). Moreover, some studies found increased risk of nutrient leaching and higher nitrous oxide emissions in zero-tillage compared with conventional tillage systems (Abdalla et al., 2013), but other studies have found the opposite (Goss et al., 1993) or inconclusive trends. The critical management practices for these systems generally focus on crop residue management, crop rotation and cover crops. Maintaining crop residue on the soil surface, permanent plant cover, and the lack of loosening through tillage reduces dispersion of the surface aggregates by rainfall or runoff. By leaving crop stubble on untilled soil or having permanent plant cover, the potential effects of wind erosion are minimized. Reducing or eliminating tillage also diminishes tillage erosion and keeps soil from being moved downhill by tillage tools.

However, reducing or eliminating tillage (i.e., zero-tillage) is no silver bullet – it can result in crop yield declines, the lack of crop residue incorporation can lead to crop disease and weed problems, and the lack of mechanical loosening can result in increasing topsoil compaction and increase nitrous oxide emissions (Soane et al., 2012; Abdalla et al., 2013; Pittelkow et al., 2015; Wardak et al., 2022). Climate, soil type and crop management (e.g., crop rotation) are factors influencing these challenges (Pittelkow et al., 2015). Soil management involves trade-offs (e.g., between food and biomass production and regulating or habitat functions), and good management aims at maximizing synergies and minimizing trade-offs between soil functions.

Evaluating soil tilth in situ

Several factors can cause soil tilth to change. Some are tillage, compaction due to wheel traffic or grazing, reconsolidation due to rainfall or irrigation, and cultivation. Qualitative observations of surface crusts, cloddiness, poor seedling emergence, ponding,

restricted plant rooting, soil organism diversity and population (especially earthworms) and increased runoff, erosion, or compaction have traditionally been used to evaluate soil tilth. Visual soil evaluation methods, e.g., the “Visual Evaluation of Soil Structure (VESS)” method (Guimarães et al., 2011) that apply robust, hierarchical field-based tests to assign scores to tilth, can be used to assess the structural state of soil, yield semi-quantitative data, and could be used to evaluate the (long-term) temporal changes of soil tilth in relation to soil management (Guimarães et al., 2017). Some of the visual evaluation methods have been designed to result in specific advice for soil management, e.g., “SOILpak” (McKenzie, 1998).

The mechanical resistance experienced by a growing plant root or tillage implement can be measured from penetration resistance. Penetration resistance is measured most precisely with a penetrometer, although a stiff wire, spade, or tile finder can also be used for a simple assessment. For precise measurements, the force required to push a rod with a cone-shaped tip through the soil to a known depth is recorded. To minimize the variation and provide the most quantitative data possible, measurements should be taken 24–48 h after rainfall or irrigation so that the soil moisture content is relatively uniform throughout the measurement zone, ideally close to field capacity. In general, if the resistance exceeds 2 MPa, root elongation is reduced by 50% and more, but this is crop-species dependent (Dexter, 1987). If the soil has good tilth, it will be fluffy, granular and separate easily along the various structural planes of weakness, but if the soil structure were massive or compacted, the tilth by definition would be considered to be quite poor.

Measurements of infiltration rates with a rainfall simulator or infiltrometers are also useful for evaluating soil tilth. These measurements are based on the time required for a known volume of water to enter the soil. If a soil has good tilth, water entry will be moderately fast for the first 5–10 cm of water. If water continues to enter at a very high rate, it could indicate the soil has no ability to retain the water or that there are cracks, crevices, fractures, root channels, earthworm holes or other macropores carrying the water away. If water enters the soil at an extremely slow rate or not at all, tilth is generally in a less desirable condition perhaps due to compaction or surface crusting. Under those conditions, rainfall and irrigation water would probably run off, increasing the potential for soil erosion and further degradation of soil tilth.

Tilth indices

Efforts to develop a tilth index were initiated to improve soil management and provide guidance for custom or prescribed tillage. The need for a tilth index was envisioned because even though an experienced person may be able to determine if a soil has good tilth by sight and feel, there was no available method to measure or quantify it. Furthermore, if soil tilth can be quantified, it may be possible to measure seasonal variation and to understand the subtle effects of soil and crop management practices such as conservation tillage or extended rotations.

The initial effort assumed that tilth of a mineral soil could be characterized by examining bulk density, strength, aggregate characteristics, organic matter content, and consistency of a soil simultaneously. Bulk density, cone index (penetration resistance or strength), and an aggregate uniformity coefficient were chosen to reflect short-term manageable soil properties. Organic matter content was chosen to reflect long-term management effects. The plasticity index was used as a measure of consistency, which reflects the relative ease with which a soil can be deformed or ruptured. This index relates to soil water properties and is strongly influenced by inherent soil characteristics (predominantly clay type and amount). These five measurements or indicators were also chosen because they can be easily measured with a routine soil test. Singh et al. (1992) proposed a tilth index based on a multiplicative combination of tilth coefficients related to bulk density, cone index, organic matter content, aggregate uniformity, and plasticity index, and proposed non-limiting, critical, and limiting values for each of these soil properties (Table 1).

Target values and optimal ranges of properties associated with desired soil tilth are presented in Table 2. However, the derivation of such values is not trivial, because optima and limiting values depend on soil texture and soil type, crop species, and soil management systems. For example, zero-tillage soils may be more compact (higher bulk density, higher mean penetration resistance) but have a better connectivity of soil pores that allows roots to proliferate despite the higher soil compactness (Wardak et al., 2022). Visual soil evaluation methods can be used to obtain information on the structural state of soil and to derive quantitative soil structure scores, as described in the previous section. Direct quantification of soil structural characteristics can be obtained by X-ray computed tomography imaging, based on intact soil cores sampled in the field (e.g., Schlüter et al., 2018). Concurrently, and with an even greater emphasis on the biological processes that affect tilth, the concepts of soil quality and more recently soil health have emerged as a strategy to quantify the effects of soil management on physical, chemical and biological properties and processes in soil (Bünemann et al., 2018; Moebius-Clune et al., 2016), as reviewed in other articles in the Encyclopedia.

Summary

The concept of soil tilth may to some extent be easier for a farmer to recognize than for a scientist to quantify. Soil tilth is used to describe soil structure in terms such as mellowness and friability, and characterizes the soil physical quality for crop growth. Traditionally, tilth has been described in many ways usually focusing on the physical condition of a soil and how it responds to tillage, functions as a seedbed, or affects seedling emergence and plant root growth and development. Tilth can be quantified in the field using methods of visual evaluation of soil structure, and based on measurements of selected soil properties on intact soil cores

Table 1 Soil measurements and corresponding coefficients developed to compute a soil tilth index for Mollisols and Alfisols in the Midwestern U.S.A.

Indicator	Description	Tilth coefficient
Bulk density	Mass per unit volume of dry soil	1.0, for $BD \leq 1.3 \text{ g cm}^{-3}$ $-1.5 + 3.87 * BD - 1.5 * BD^2$, for $1.3 \leq BD \leq 2.1 \text{ g cm}^{-3}$ 0.0, for $BD \geq 2.1 \text{ g cm}^{-3}$
Penetration resistance (cone index)	A measure of soil strength indicating ease of root penetration, plant growth, and yield	1.0, for $CI \leq 1.0 \text{ MPa}$ $1.0 - 0.002 * CI - 0.01 * CI^2$, for $1.0 \leq CI \leq 10.0 \text{ MPa}$ 0.0, for $CI \geq 10.0 \text{ MPa}$
Soil organic matter content	The organic fraction of soil excluding non-decayed plant and animal residue	1.0 for $OM \geq 50 \text{ g kg}^{-1}$ $0.59 + 0.122 * OM - 0.008 * OM^2$, for $1\% \leq OM \leq 5\%$ 0.70, for $OM \leq 1\%$
Aggregate uniformity coefficient	Shape of the grain size (aggregate) distribution curve (all of equal size gives an AUC = 1.0)	1.0, for $AUC \geq 5$ $0.348 + 0.245 * AUC - 0.023 * AUC^2$, for $2 \leq AUC \leq 5$ 0.75, for $AUC \leq 2$
Plasticity index	The difference in water content between the liquid limit and the plastic limit of a soil (cohesiveness)	1.0 for $PI \leq 150 \text{ g kg}^{-1}$ $1.0 + 0.0009 * PI - 0.00016 * PI^2$, for $15\% \leq PI \leq 40\%$ 0.80, for $PI \geq 40\%$

$$\text{Tilth Index} = CF_{(BD)} * CF_{(CI)} * CF_{(UC)} * CF_{(OM)} * CF_{(PI)}$$

Where: BD = bulk density; CI = cone index; UC = uniformity coefficient;

OM = soil organic matter; PI = plasticity index

Tilth coefficients are between 0 and 1, a tilth coefficient of 1.0 reflects non-limiting values.

From Singh KK, Colvin TS, Erbach DC, and Mughal AQ (1992) Tilth index: An approach to quantifying soil tilth. *Transactions of ASAE* 35(6): 1777–1785.

Table 2 Soil properties of desired soil tilth: target values, optimum ranges and thresholds.

Soil structural property	Unit	Optimum range [†]	Estimated target values [¶]		
			Clay	Loam	Sandy loam
<i>Mechanical traits</i>					
Degree of compactness	Mg m ⁻³ /Mg m ⁻³	0.87			
Bulk density	Mg m ⁻³		1.23	1.50	1.54
Penetration resistance	MPa	≤2	1.6	1.5	1.4
<i>Soil aeration</i>					
Air-filled porosity	m ³ m ⁻³	≥0.1–0.14	0.05	0.09	0.12
Rel. gas diffusion coeff.	m ² s ⁻¹ /m ² s ⁻¹	≥0.0037–0.02	0.003	0.008	0.013
Air permeability	μm ²	≥20	16.4	28.2	36.1
<i>Preferential pathways for fluids and roots</i>					
Microporosity	m ³ m ⁻³	≥0.07			
<i>Water retention</i>					
Plant-available water	m ³ m ⁻³	≥0.15–0.2	0.17	0.19	0.17
Soil organic carbon cone.	kg ¹ kg ⁻¹	0.02			

From Or D, Keller T, and Schlesinger WH (2021) Natural and managed soil structure: On the fragile scaffolding for soil functioning. *Soil and Tillage Research* 208: 104912.

and aggregates sampled in the field. Soil tilth is positively influenced by soil organic matter and negatively influenced by soil compaction, whereas tillage can have positive or negative effects. Soil management, especially tillage practices, cover crops, and crop rotations, often affects soil tilth because of its direct or indirect effect on soil organic matter. Soil scientists, ecologists, and agricultural engineers are continuing to evaluate different biological, chemical, and physical measurements or indicators that can be used individually or combined into index values to quantitatively describe tilth and to use that information to improve soil management practices, such as conservation or zero-tillage agriculture, together with the use of cover crops or more diversified and extended crop rotations.

References

- Abdalla M, Osborne B, Lanigan G, Forristal D, Williams M, Smith P, and Jones MB (2013) Conservation tillage systems: A review of its consequences for greenhouse gas emissions. *Soil Use and Management* 29: 199–209.
- Alexander J (1922) The value of tilth in agriculture. *Science* 55: 156.
- Blair GWS (1937) Compressibility curves as a quantitative measure of soil tilth. *The Journal of Agricultural Sciences* 27: 541–556.
- Blair GWS and Cashen GH (1938) Compressibility curves as a quantitative measure of soil tilth, II. *The Journal of Agricultural Science* 28: 367–378.
- Bünemann EK, Bongiorno G, Bai Z, et al. (2018) Soil quality – A critical review. *Soil Biology and Biochemistry* 120: 105–125.
- Chandrasekhar P, Kreiselmeier J, Schwen A, Weninger T, Julich S, Feger K-H, and Schwärzel K (2019) Modeling the evolution of soil structural pore space in agricultural soils following tillage. *Geoderma* 353: 401–414.
- de Freitas PL and Landers JN (2014) The transformation of agriculture in Brazil through development and adoption of zero tillage conservation agriculture. *International Soil and Water Conservation Research* 2: 35–46.
- Dexter AR (1987) Mechanics of root growth. *Plant and Soil* 98: 303–312.
- Dexter AR (1988) Advances in characterization of soil structure. *Soil & Tillage Research* 11: 199–238.
- Fenton EW (1930) A botanical study of pasture plots. *Annals of Applied Biology* 17: 522–548.
- Frierson LS (1921) The value of tilth in agriculture. *Science* 54: 193–194.
- Goss MJ, Howse KR, Lane PW, Christian DG, and Harris GL (1993) Losses of nitrate-nitrogen in water draining from under autumn-sown crops established by direct drilling or mouldboard ploughing. *Journal of Soil Science* 44: 35–48.
- Groenevelt PH, Kay BD, and Grant CD (1984) Physical assessment of a soil with respect to rooting potential. *Geoderma* 34: 101–114.
- Guimarães RML, Ball BC, and Tormena CA (2011) Improvements in the visual evaluation of soil structure. *Soil Use and Management* 27: 395–403.
- Guimarães RML, Lamandé M, Munkholm LJ, Ball BC, and Keller T (2017) Opportunities and future directions for visual soil evaluation methods in soil structure research. *Soil & Tillage Research* 173: 104–113.
- Hadas A (1997) Soil tilth – The desired soil structural state obtained through proper soil fragmentation and reorientation processes. *Soil and Tillage Research* 43: 7–40.
- Karlen DL, Erbach DC, Kaspar TC, Colvin TS, Berry EC, and Timmons DR (1990) Soil tilth: A review of past perceptions and future needs. *Soil Science Society of America Journal* 54: 153–161.
- Kawamoto K, Moldrup P, Schjønning P, Iversen BV, Komatsu T, and Rolston DE (2006) Gas transport parameters in the vadose zone: Development and tests of power-law models for air permeability. *Vadose Zone* 5: 1205–1215.
- McKenzie D (1998) *SOILpak for cotton growers*, 3rd edn New South Wales, Australia: Orange.
- Meurer KHE, Haddaway NR, Bolinder MA, and Kätterer T (2018) Tillage intensity affects total SOC stocks in boreo-temperate regions only in the topsoil – A systematic review using an ESM approach. *Earth-Science Reviews* 177: 613–622.
- Moebius-Clune BN, Moebius-Clune DJ, Gugino BK, Idowu OJ, Schindellbeck RR, Ristow A, van Es HM, Thies JE, Shayler HA, McBride MB, Wolfe DW, and Abawi GS (2016) *Comprehensive Assessment of Soil Health – The Cornell Framework Manual, Edition 3.1*. Geneva, NY: Cornell University.
- Or D and Ghezzehei TA (2002) Modeling post-tillage soil structural dynamics: A review. *Soil & Tillage Research* 64: 41–59.
- Or D, Keller T, and Schlesinger WH (2021) Natural and managed soil structure: On the fragile scaffolding for soil functioning. *Soil and Tillage Research* 208: 104912.
- Peth S (2011) Noninvasive quantification of 3D pore space structures in soils. In: Glinski J, Horabik J, and Lipiec J (eds.) *Encyclopedia of Agrophysics*, pp. 516–519. Berlin: Springer.
- Pittelkow CM, Liang X, Linquist BA, van Groenigen KJ, Lee J, Lundy ME, van Gestel N, Six J, Venterea RT, and van Kessel C (2015) Productivity limits and potentials of the principles of conservation agriculture. *Nature* 517: 365–368.
- Roger-Estrade J, Richard G, Dexter AR, Boizard H, De Tourdonnet S, Bertrand M, and Caneill J (2009) Integration of soil structure variations with time and space into models for crop management. A review. *Agronomy for Sustainable Development* 29: 135–142.
- Sandin M, Jarvis N, and Larsbo M (2018) Consolidation and surface sealing of nine harrowed Swedish soils. *Soil & Tillage Research* 181: 82–92.
- Schlüter S, Grossmann C, Diel J, Wu G-M, Fischer S, Deubel S, and Rücknagel J (2018) Long-term effects of conventional and reduced tillage on soil structure, soil ecological and soil hydraulic properties. *Geoderma* 332: 10–19.
- Singh KK, Colvin TS, Erbach DC, and Mughal AQ (1992) Tilth index: An approach to quantifying soil tilth. *Transactions of ASAE* 35(6): 1777–1785.
- Soane BD, Ball BC, Arvidsson J, Basch G, Moreno F, and Roger-Estrade J (2012) No-till in northern, western and south-western Europe: A review of problems and opportunities for crop production and the environment. *Soil and Tillage Research* 118: 66–87.
- Soil Science Society of America (1996) *Glossary of Soil Science Terms*. Madison, WI: American Society of Agronomy 137.
- Wardak DLR, Padia FN, de Heer M, Sturrock CJ, and Mooney SJ (2022) Zero tillage has important consequences for soil pore architecture and hydraulic transport: A review. *Geoderma* 422: 115927.
- Yoder RE (1937) The significance of soil structure in relation to the tilth problem. *Soil Science Society of America Journal* 2: 21–33.

Crusts and seals: Structural

Shmuel Assouline^a and Sally Thompson^{b,c}, ^aDepartment of Environmental Physics and Irrigation, Agricultural Research Organization—Volcani Institute, Rishon LeZion, Israel; ^bDepartment of Civil, Environmental and Mining Engineering, University of Western Australia, Crawley, WA, Australia; ^cDepartment of Civil and Environmental Engineering, University of California, Berkeley, CA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	57
What is a soil seal?	58
The “standard” description	58
Why do seal layers form?	58
Consequences of the seal layer for rainfall-runoff-infiltration partitioning	59
Factors affecting seal formation	59
Modeling soil seals	60
Modeling the structure of the seal layer	60
Modeling seal formation	60
Modeling the seal hydraulic properties	61
Modeling water flow in a soil seal layer	62
Modeling case study—Can polymer application prevent seal layer formation	62
Implications	62
Infiltration and runoff	62
Vegetation, including agronomy and restoration	64
Ecological impact of crusts	65
Conclusions	66
References	67

Abstract

Crusts and seals form by the physical impact of raindrops on a soil surface. Raindrops rapidly wet and mechanically fragment aggregates on the soil surface, forming dense seal layers when fine sediment is washed into and clogs soil pores. When seal layers dry into hard crusts, long-lasting physical impacts restrict seedling emergence and alter the movement of water, energy and matter between the soil and the surface. Seal and crust layers affect important elements of landscape function, most notably by creating runoff-runon dynamics that support vegetation growth in drylands. This chapter introduces the physico-chemical basis of physical seal and crust formation, the history of conceptual and numerical models proposed to describe the properties and hydrological effects of seals and crusts, and the implications for agronomy and ecology.

Introduction

Key hydrological processes such as infiltration and evaporation are governed by the soil physical and hydraulic properties near the soil-atmosphere interface. Changes in the physical conditions of the soil surface form a compacted and hard layer—called a ‘seal’ when wet or a ‘crust’ when dry (Photo 1). Seals and crusts affect numerous hydrologic and biological processes. These changes occur frequently, dynamically during and between rainfall events, and can lead to large changes in soil physical properties. In this rapid change, physical crusts are distinct from the biological crusts found in arid or semiarid regions. Formed from microbial plant communities of cyanobacteria, green algae, lichens, mosses, and microfungi that colonize the near-surface soil, biological crusts require an establishment period of many years. In this chapter, we focus on structural seals and crusts. Natural and anthropogenic processes, such as fire or compaction, influence the structure of this upper soil layer. One of the most pervasive of these processes is the wetting of dry soils and the exposure of the soil surface to the impact of raindrops. This impact and rapid wetting can fragment large soil aggregates through a variety of physico-chemical mechanisms. The fragments produce a fine-textured and low permeability “seal” layer, which affects the flow processes in the soil profile.

Soil surface sealing can have severe agricultural, hydrological and environmental effects. Seals decrease the infiltration rate, reducing water movement to the root zone available to plants and diminishing the natural recharge of aquifers. The water that does not infiltrate forms surface runoff, increasing soil erosion. Reduced water supply and the mechanical strength of crusts inhibit seedling emergence and suppress plant growth, ultimately decreasing crop yields. In arid and semi-arid regions, soil surface sealing can induce a vicious cycle that interacts with drivers of desertification, such as overgrazing to intensify land degradation. Low vegetation cover—for example created by overgrazing—exposes the soil to rainfall, creating seals and crusts, which reduce plant vigor and growth, leaving even more of the soil area bare and exposed to rainfall. Conversely, the seal layer can be beneficial if lands are managed to maximize runoff, for example in water harvesting systems. Intriguingly, the function of many arid and semi-arid ecosystems may also depend on a mixture of sealed and unsealed soils being present in the landscape.

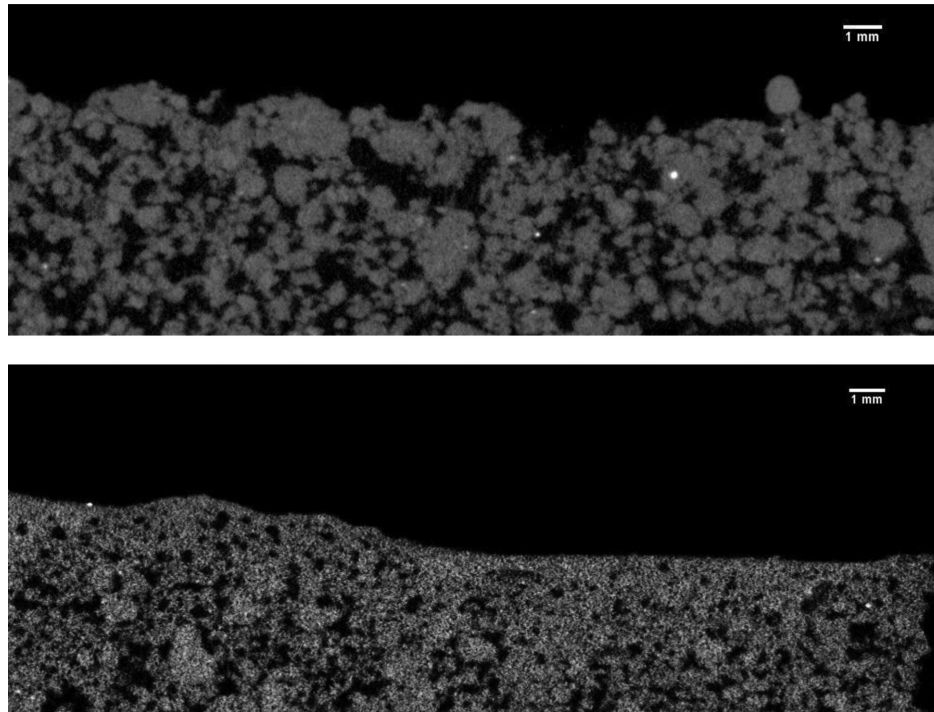


Photo 1 2D X-ray micro Computed Tomography images of the top soil surface after 5 min (top) and 14 min (bottom) of a constant rainfall of 60 mm h^{-1} intensity. The pictures are from a conference paper of Garbout A, Armenise E, Ahn S, Doerr S, Ritz K, Simmons R, Sturrock C, Mooney S (2014) Soil surface micro topology assessment using X-ray micro Computed Tomography. In: *4th International Conference on Surface Metrology, Hamburg, Germany*; <https://www.researchgate.net/publication/276026594>.

The formation of structural seals at the surface of bare soils exposed to the direct impact of raindrops is influenced by many factors involving soil properties, rainfall characteristics and flow conditions, as revealed by a large body of laboratory and field studies. This work has also produced several different conceptual, physical and empirical models of seal formation, seal properties, and seal effects on water and water vapor flow.

This section describes the morphology, phenomenology, and consequences of surface seal or crust layers, together with different conceptual and empirical modeling approaches to describe their properties and impacts.

What is a soil seal?

The “standard” description

Widely used descriptions state that seals or crusts “consist of two discrete and uniform layers, a 0.1 mm skin overlying a 2.0 mm washed-in layer” (McIntyre, 1958a, 1958b). In this description, the skin refers to the highly compacted, low porosity surface layer, and the washed-in layer refers to a deeper region whose properties are mostly impacted by the wash-in of fines to the pores. In practice, the thickness and properties of seal layers are highly variable—ranging from 0.1 mm up to 20.0 mm, and often the seal structure varies gradually with depth, rather than forming discrete layers (Photo 1). Light or electron microscopy, X-ray radiography or computed tomography of thin sections of the crust can all be used to measure and characterize these structural features of seals (Tackett and Pearson, 1965; Roth, 1997; Bresson et al., 2004; Garbout et al., 2014).

Why do seal layers form?

Structural seals form at the soil surface following the destruction of soil aggregates by rainfall. Two processes occur simultaneously when a raindrop collides with a bare soil surface: (1) a transfer of kinetic energy from the raindrop to the soil surface and (2) rapid imbibition of water by the soil. The transfer of energy causes mechanical changes to the soil surface, including compaction and particle detachment, while droplet splash can mobilize the fine material released by the rain impact. Simultaneously, rapid imbibition of water causes several changes to the soil. It enhances the collapse of unsaturated soil aggregates, facilitated by the sudden compression of entrapped air within the aggregates. This compression is particularly great when shrink-swell clays expand as they wet, when it can cause “explosive” destruction of aggregates, known as slaking. Wet soils also become more plastic and less cohesive, so that soil wetting increases the mechanical effects on the soil of raindrop impact.

During the early stage of seal formation, infiltration rates are often still large and there is little runoff. Under these conditions, loosened fine soil particles can be transported into the upper soil layer with infiltrating water. These fine particles fill the inter-aggregate voids in undisturbed soil, forming the seal layer (McIntyre, 1958a, 1958b). As infiltration continues, deeper aggregates are wetted, undergo further swelling and collapse, and the seal layer thickens. At the soil surface, raindrops continue to cause compaction (Bresson and Boiffin, 1990), increasing the bulk density and reducing the porosity of the seal layer.

Seal formation is a dynamic process that proceeds more slowly over time during rainfall. During the early stage of seal formation, the concentration of the fine particles in the percolating suspension is small and the open pore structure only slightly affected. The soil hydraulic conductivity allows large infiltration rates with considerable capacity to transport fine soil particles to depth. As the permeability of the surface declines, a decrease in the wash-in of fines and the rate of wetting of deeper aggregates prevents further growth of the seal layer. The saturated hydraulic conductivity of seal layers is often much smaller than the undisturbed soil. Ratios of the seal:undisturbed soil hydraulic conductivity range from 20% to as little as 0.01%. The reduction in conductivity scales with the thickness of the seal layer and the extent of compaction within it. For a given soil-rainfall system, the seal layer thickness (d_c), reaches its maximum shortly after the beginning of rainfall. It may take longer for maximum compaction of the seal layer to be reached, with some experimental studies finding that it was not attained until cumulative rainfall had reached approximately 50 mm.

Consequences of the seal layer for rainfall-runoff-infiltration partitioning

The main hydrological impact of the formation of seal layers is to reduce infiltration rates, so that rainfall is partitioned into runoff rather than infiltration (Morin and Benyamini, 1977; Romkens et al., 1986). This effect is illustrated in Fig. 1, which shows the evolution of infiltration rates over time for undisturbed soils (dotted line), for a soil on which a seal is dynamically forming (thin line), and a soil that was sealed prior to the rainfall event (thick line). Curves show a fitted infiltration model in each case.

Factors affecting seal formation

Rainfall intensity and kinetic energy play a major role in determining the seal properties and the rate of seal formation. The greater the input of kinetic energy into the soil surface via the rainfall, the more rapidly seal layers are formed and the more their properties differ from the undisturbed soil.

The formation and characteristics of the seal layer are also strongly affected by soil properties that relate to the cohesion of soil particles and stability of soil aggregates. These include (i) soil mineralogy and texture; (ii) aggregate size distribution; (iii) aggregate stability; (iv) initial bulk density; (v) initial water content distribution in the soil profile, (vi) application of phosphogypsum or polymers to the upper soil layer; (vii) slope, (viii) organic matter content; (ix) soil exchangeable sodium percentage (ESP); and (x) electrical conductivity (EC) of the applied water.

This multiplicity of soil and rainfall factors interact, so that the relationships between storm, soil and seal characteristics are complex and variable. This complexity means that direct comparison or extrapolation of experimental results is challenging. Therefore, soil scientists must rely on models to describe (i) the structure, (ii) the formation, (iii) the hydraulic properties and finally (iv) the flow of water within the seal layer.

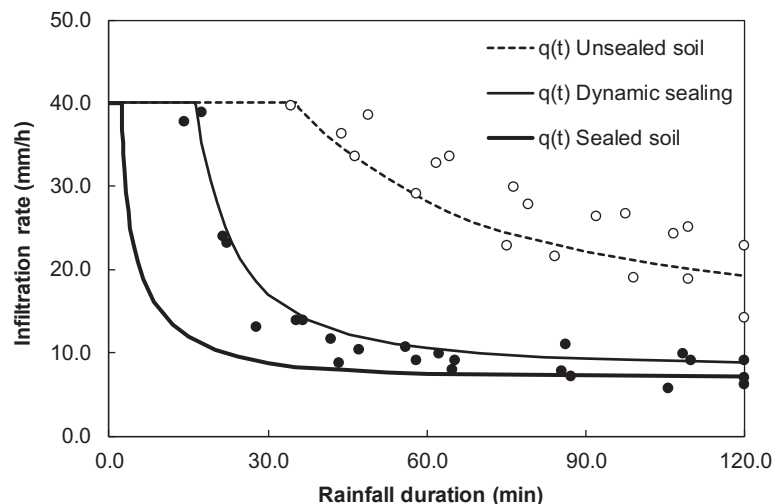


Fig. 1 Modeled infiltration through (i) a completely formed seal (bold line), (ii) a seal under formation (thin solid line), and (iii) an undisturbed silty clay loam soil (dashed line). The measured data (open and black dots) are from Baumhardt RL (1985) *The Effect of Rainstorm Characteristics on Soil Sealing and Infiltration*. Ph.D. thesis, Mississippi State University.

Modeling soil seals

Modeling the structure of the seal layer

Conceptual models that describe the structure of the seal layer attempt to relate physical properties of the soil within the seal layer to the depth of soil within it and the characteristics of the underlying soil. The simplest such models describe the seal layer as a soil layer of a given thickness and uniform properties (e.g., Hillel and Gardner, 1969, 1970).

Mualem and Assouline (1989) suggested instead that the seal bulk density, ρ_c , should be greatest at the surface where the soil is most exposed to raindrop impact, and should then decrease exponentially with depth, h , until it approaches that of the undisturbed soil, ρ below the seal layer:

$$\rho_c(h) = \rho + \Delta\rho_o e^{-\gamma h} \quad (1)$$

where h is the depth into the soil layer, $h = 0$ at the soil surface and increases in a downward direction, $\Delta\rho_o$ is the maximum change in bulk density between the undisturbed soil and the surface of the seal layer ($h = 0$), and γ is a characteristic parameter describing the soil-rainfall interaction. γ can be related to the thickness of the seal layer (d_c) by defining this thickness through truncating the distribution at the location where changes in soil bulk density with depth are insignificant, i.e., defining that:

$$\rho_c(d_c) - \rho \leq 10^{-3} \Delta\rho_o \quad (2)$$

Giving:

$$\gamma = -\ln(10^{-3})/d_c \quad (3)$$

This theoretical model was validated against accurate measurements of soil bulk density distribution with depth (Roth, 1997; Bresson et al., 2004), as shown in Fig. 2 for a loamy silt soil. Roth (1997) suggested that a sigmoidal function (dashed line in Fig. 2) could be a more appropriate model than an exponential one, recognizing that the maximally compacted soil condition could extend slightly below the soil surface. The exponential model can readily be modified to produce a sigmoidal form by raising h in Eq. (1) to a power n (Assouline, 2004). Practically, however, both forms are similar, with the exponential and sigmoidal functions diverging only in the upper 1 mm of the soil—a region so thin that no reliable bulk density data can be obtained.

Modeling seal formation

Seal layer formation, and therefore the infiltration and runoff processes that it mediates, are dynamic and change throughout rainfall events. This dynamism means that models of the seal layer formation are needed to understand the consequences of seal layers for the fate of rainfall. Assouline and Mualem (1997) developed a dynamic model to predict the evolution of the soil seal layer over time, given the properties of rainfall (or irrigation) and the undisturbed soil.

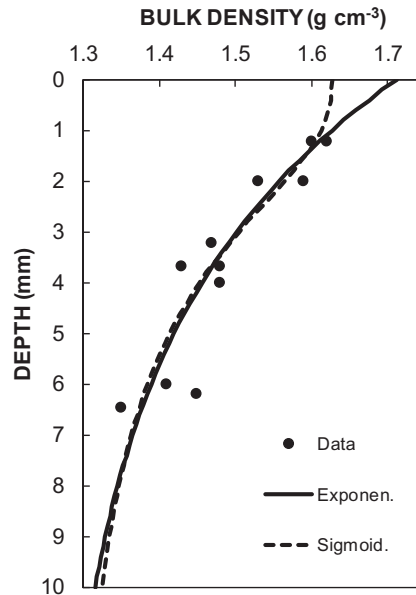


Fig. 2 The modeled distribution with depth of the bulk density within the completely formed seal layer according to the exponential model in Eq. (1) and the sigmoidal model of Roth (1997), compared with the experimental data of Roth (1997) for a loamy silt soil. Based on Assouline S (2004) Rainfall-induced soil surface sealing a critical review of observations, conceptual models, and solutions. *Vadose Zone Journal* 3(2): 570–591.

At the soil surface, the disturbance resulting from raindrop impact and imbibition is expressed in terms of the bulk density increase relative to the undisturbed soil, $\Delta\rho_o$, as a function of rainfall intensity, I , and time of exposure to rainfall, t :

$$\Delta\rho_o(I, t) = \Delta\rho_o(1 - e^{-\xi t}) \quad (4)$$

where $\Delta\rho_o$ retains its previous definition. The term ξ captures the time-dependence of seal layer formation based on rainfall and soil properties, as defined by:

$$\xi = \frac{6\omega k I}{D_{max}\Delta\rho_o\tau} \int_0^{D_{max}} D^2 f(D, I) dD \quad (5)$$

Here, $f(D, I)$ is the raindrop size distribution, D_{max} the maximum diameter of a drop, while the parameter k links the diameter, D , of drop to its velocity (and thus kinetic energy). The variable τ represents the shear strength of the undisturbed soil, and ω , is a fitting parameter.

The measured evolution of $\Delta\rho_o$ with rainfall duration is shown in Fig. 3, with the fitted model shown in black (Augeard et al., 2008).

Modeling the seal hydraulic properties

The thin and heterogeneous nature of the soil seal layer usually precludes direct measurement of its hydraulic properties. These properties are therefore usually inferred by comparing measurements of infiltration or flow through the seal layer to models of water flow. This requires representing the seal layer hydraulic properties within the model.

The simplest hydraulic model of a seal layer is to treat it as a saturated porous medium with fixed (low) hydraulic conductivity and thickness (Hillel and Gardner, 1970; Ahuja and Swartzendruber, 1973). The limitations of this approach are that (i) it does not describe the dynamic evolution of the seal layer and its hydraulic properties, and (ii) unsaturated properties of the seal, for example its water retention curve, must be addressed when considering the drying behavior of sealed (crusted) soils.

Seal hydraulic properties can therefore be modeled using similar approaches to describing the properties of undisturbed soil (Mualem and Assouline, 1989; Baumhardt et al., 1990). For example, retaining the concept of a uniform seal layer, Baumhardt et al. (1990) represented the seal water retention curve using the Brooks and Corey (1964) model, suggesting the following approach: (i) experimentally measuring seal conductance to evaluate the saturated hydraulic conductivity assuming a given or observed seal thickness; (ii) computing the saturated water content, θ_{sc} , of the compacted seal based on the estimated saturated hydraulic conductivity and applying the Kozeny-Carman relationship (iii) estimating the water entry value of the compacted seal, h_{wevs} , using the Poiseuille equation; and (iv) determining the Brooks and Corey (1964) exponent λ_c for the seal layer graphically, assuming there is a proportional increase in the initial water content of the seal, θ_{ic} , with bulk density. As an extension to this approach, the hydraulic properties of the seal layer can be related to the bulk density profile with depth (as per Fig. 1), leading to a depth-distributed representation of the seal layer hydraulics (Fig. 4).

Mualem and Assouline (1989) have presented a different approach based on the impact of bulk density on the parameters in the van Genuchten (1980) model of the water retention curve.

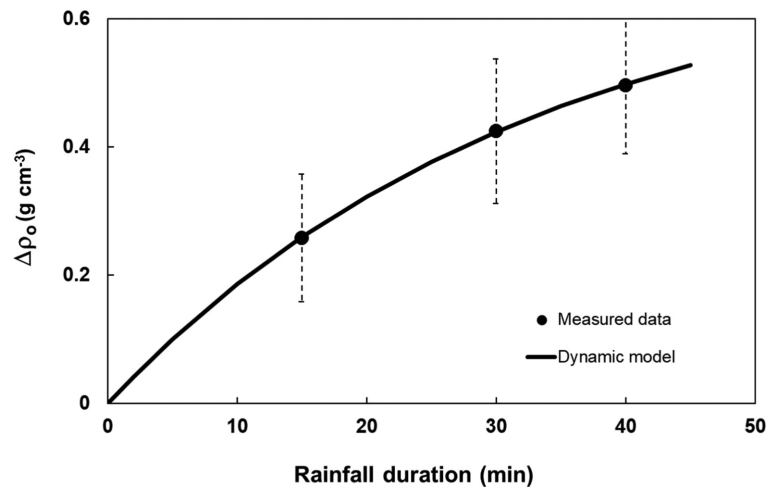


Fig. 3 Evolution of the change in measured bulk density at the soil surface $\Delta\rho_o$ as a function of the duration of a rainfall event with an intensity of 30.5-mm h⁻¹ (black dots). Vertical bars correspond to the standard error of the bulk density data. Eq. (2) is fitted on $\Delta\rho_o$ data to model its change in time. Based on Augeard B, Bresson LM, Assouline S, Gaskuel C, Kao C and Vauclin M (2008) Dynamics of soil surface bulk density: Role of water table elevation and rainfall duration. *Soil Science Society of America Journal* 72: 412–423.

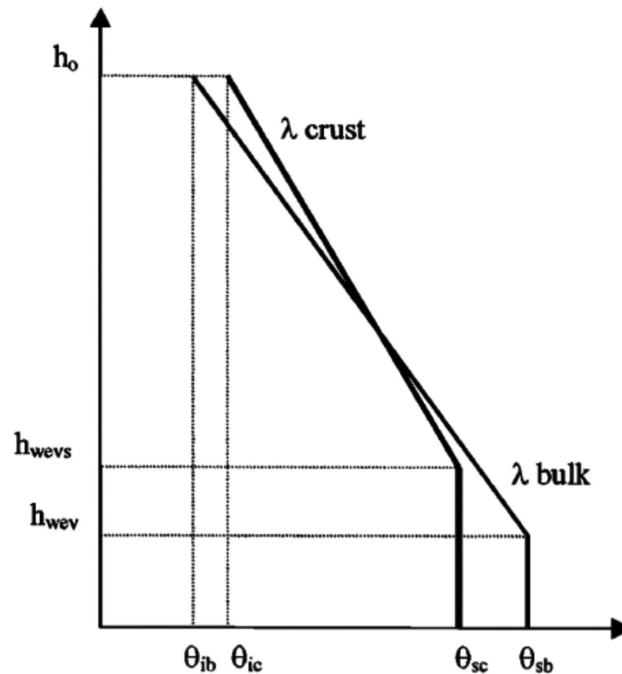


Fig. 4 The methodology suggested by Baumhardt (1985) for deriving the water retention curve of the seal layer. After Baumhardt RL, Romkens MJM, Whisler FD and Parlange JY (1990) Modeling infiltration into a sealing soil. *Water Resources Research* 26: 2497–2505.

Modeling water flow in a soil seal layer

Sealed soils are a special case of layered soils, and similar modeling approaches can be applied to both. The “special” aspects of the sealed soil are that seal physical and hydraulic properties are dynamic, related to those of the undisturbed soil on which the seal develops, and respond to the rainfall conditions prevailing during the seal layer formation. By linking a flow model with a seal development model such as the dynamic model of Assouline and Mualem (1997), a dynamic infiltration solution can be obtained, which allows a quantitative analysis of the soil and rainfall properties on seal formation (Assouline and Mualem, 2000).

Modeling case study—Can polymer application prevent seal layer formation

Many physical and chemical treatments are proposed to reduce the formation of soil seals, for example by manipulating the initial water content of the soil surface (Le Bissonnais and Singer, 1992), or the chemical conditions of the soil-water system (Shainberg, 1992; Levy and Rapp, 1998). Particularly promising treatments include applying small amounts of polymers to the soil surface, which can limit aggregate breakdown, reduce seal formation, increase infiltration rates and reduce runoff.

The effect of polymers on seal formation can be represented in Eq. (5) by modifying the soil shear strength parameter τ . Evaluating the sensitivity of modeled infiltration rates to τ therefore allows an evaluation of the potential outcomes of polymer treatment. Results comparing infiltration under sealed conditions and for a soil treated with 50 kg ha⁻¹ of the polymer *P-101*, which increases τ by a factor of 4, are presented in Fig. 5. As shown in Fig. 5, the modeled and observed infiltration rates agree well, demonstrating the sensitivity of seal formation, physical properties and hydrological outcomes to changes in soil shear strength. These results illustrate the potential value of simple models of soil sealing for prediction and inference, and highlight that a simple understanding of how soil amendments affects τ or K_s could be sufficient to project the hydrological effects of such treatments.

Implications

Infiltration and runoff

In most field settings, soil sealing does not occur uniformly across the landscape, but varies with heterogeneity in soil properties, soil surface cover, and local microtopographic elevation and slope. The effects of sealing on infiltration for homogeneous soils were shown in Fig. 1, and are also illustrated in the thin lines and data points (obtained from homogeneous soil column experiments of Baumhardt, 1985) in Fig. 6. This figure shows infiltration under a 120 min, 40 mm h⁻¹ storm. As expected, seal formation on homogeneous soils reduces the time of ponding (from ~40 min to ~20 min), and greatly reduces infiltration rates (Hopmans et al., 2007). The differences that arise between sealed and unsealed soils when a heterogeneous soil field is simulated, however, are less pronounced (Assouline and Mualem, 2002). In particular, (a) the time of ponding is almost identical between sealed and unsealed

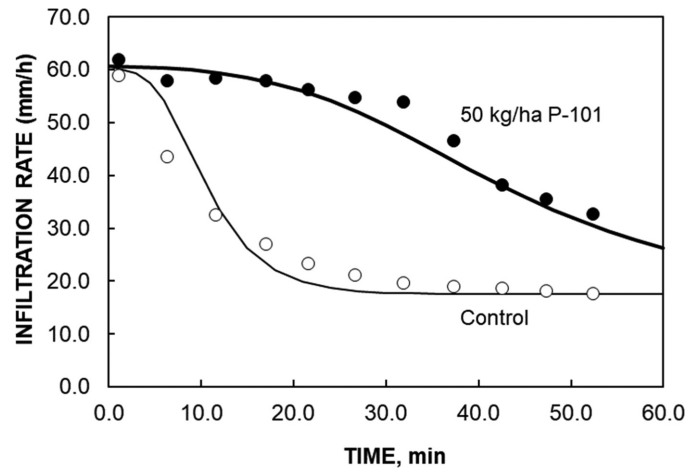


Fig. 5 Simulated effect of the application of polymer P-101 on infiltration during soil sealing using the model of Assouline and Mualem (1997) and the experimental data of Ben-Hur (2001) for a loess soil. Based on Assouline S (2004) Rainfall-induced soil surface sealing a critical review of observations, conceptual models, and solutions. *Vadose Zone Journal* 3(2): 570–591.

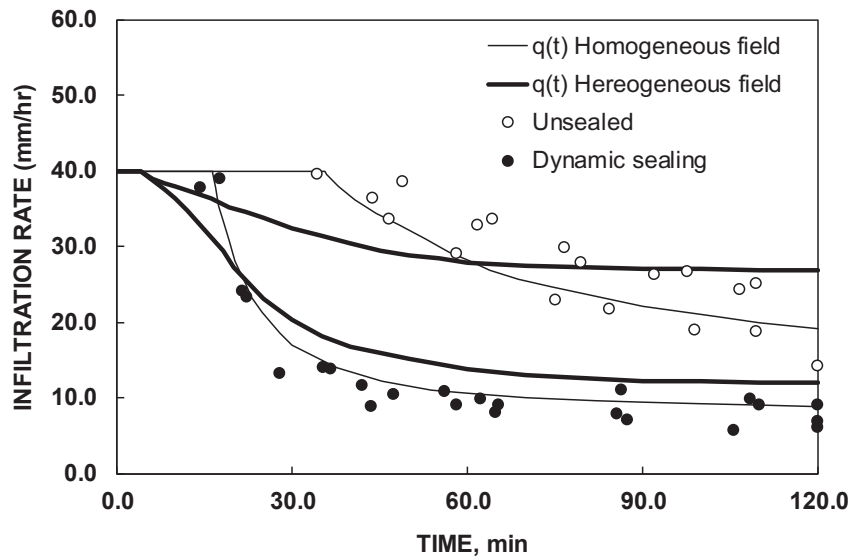


Fig. 6 The infiltration curves simulated for a uniform (thin line) and a heterogeneous field (thick line) for the unsealed case and during seal formation, along with the corresponding measured infiltration rates in packed homogeneous soil columns (open and black dots reported by Baumhardt, 1985). Based on Hopmans JW, Parlange JY and Assouline S (2007) Infiltration. In: *The Handbook of Groundwater Engineering*, JW Delleur (ed.), 2nd edn, CRC Press, Taylor and Francis Group: Boca Raton, ch. 7.

cases—reflecting the importance of low permeability soils in driving ponding before the seal layer forms; and (b) the steady state infiltration rate is greater for heterogeneous sealed soils than homogeneous sealed soils—likely reflecting a proportionally greater volume of infiltration moving through the most permeable of the heterogeneous soil properties. The modeled results assume random heterogeneity and model infiltration only—they do not therefore account for processes, such as runoff-runon, or depositional seal formation in microtopographic depressions, both of which further complicate the hydrology of sealed landscapes in field settings.

Scaling these results up further to consider runoff generation at field scales, synthetic experiments show that surface sealing—its presence and whether it is in place or forming dynamically—are the dominant factors determining runoff production from catchments where sealing could occur. As shown in Fig. 7, dynamic soil sealing increased runoff by a factor of 10, while a previously developed seal increased runoff by a factor of 20 (for a 40 mm h^{-1} , 45 min simulated rainfall event) (Assouline and Mualem, 2006). This high sensitivity suggests that (i) sealing should not be ignored in runoff assessments, and (ii) that methods to induce sealing could form part of rainwater harvesting designs—or conversely, treatments to reduce sealing could assist in mitigating flooding and erosion risks.

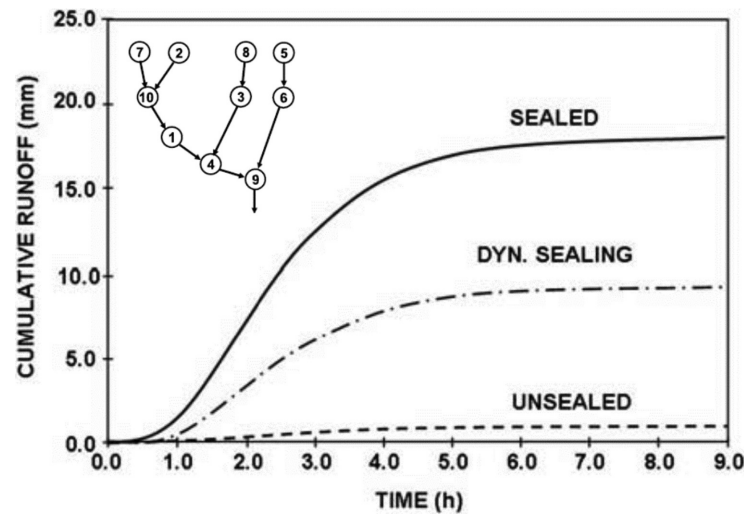


Fig. 7 Cumulative runoff versus time for three surface conditions in the case where 30 mm of rainfall at $I = 40 \text{ mm h}^{-1}$ are applied to a small semi-arid catchment (described in the inset) with different soil surface conditions. Based on Assouline S and Muallem Y (2006) Runoff from heterogeneous small bare catchments during soil surface sealing. *Water Resources Research* 42: W12405.

Vegetation, including agronomy and restoration

Seal and crust layers influence vegetation through altering the infiltration and evaporation of water, by presenting a layer with low hydraulic conductivity that reduces wetting and drying rates, and by predisposing the soil surface to erosion. These outcomes alter the water regime, energy regime and mechanical stresses experienced by growing vegetation, with generally negative effects on crops. For instance, in a comparison across multiple soil types and replicates, soil crusting reduced maize yields between 5% and 23% (average 15%), with the greatest yield reductions occurring under the most productive conditions (Nizami and Khan, 1991). Replanting costs associated with poorly controlled crusting in the United States in 1975 were estimated as \$US 10 million annually (\$50 million in today's values) (Edwards, 1977). Crusting is often particularly problematic for restoration practices in drylands prone to crust formation, and in high priority environments such as mine cover systems, where soil compaction is desirable to limit recharge into the covered area, but crusted soils inhibit plant establishment (Beckett et al., 2017).

Most research on crops in agronomic systems has focused on understanding the mechanical impact of crusting on seedling emergence, and on agronomic techniques for reducing crust formation. Cemented soil crusts form a barrier to emerging seedlings. Seedlings must either grow through a fracture or be able to exert enough force to break the crust. The mechanical strength of crusts, which varies with impact energy and rainfall chemistry—specifically the dispersivity of the rainwater, can vary from 0.5 MPa (low energy, low dispersivity) to 1.2 MPa (high energy, high dispersivity). The force exerted by emerging seedlings ranges from 0.15 N for alfalfa to over 4 N for maize, meaning that the pressures exerted by the seedlings are frequently too small to emerge through crusts, although results can be quite variable for a given crop type, as summarized in Table 1.

Several approaches are available to ameliorate the effects of crusting on seedling emergence. Mechanical disruption of crusts via tillage can create cracks allowing for emergence, but may also damage seedlings. Crusts can also be physically removed. Crusts can be watered to soften them and reduce their mechanical strength, although the resulting reductions in soil temperature can also impede seedling emergence. Alternatively, where crusts and cover cannot be altered, planting seeds together can assist in seedling

Table 1 Crust strength preventing emergence of several seed crops.

Crop type	Crust strength preventing emergence	References
Beans	0.03 MPa	Richards (1953)
Grain sorghum	0.2 MPa	Parker and Taylor (1965)
Pearl millet	0.02 MPa (1 seed per hole) 0.05 MPa (2 seeds per hole)	Sinha and Ghildyal (1979)
Cotton	0.03 MPa (1 seed per hole) 0.07 MPa (2 seeds per hole) 0.4 MPa	Sharma and Agarwal (1974) and Gerard (1980)
Cowpea/mungbeans	0.3 MPa	Chaudhry and Das (1980)
Maize	0.24 MPa (average pressure exerted by the seed) 0.7 MPa	Souty et al. (1992)

emergence from crusts—this approach is recommended in some forms of land restoration (Madsen et al., 2012) as well as agriculture. A variety of crop breeding efforts have also targeted physical and growth characteristics of seedlings that help them germinate in the presence of crusts (Anzoooman et al., 2018).

Increasingly, however, no-tillage agronomic practices are recommended to attempt to prevent the formation of the soil crust in the first place (Pareja- Sánchez et al., 2017). No-tillage residue management practices provide a canopy cover that intercepts raindrop impact and reduces evaporation, thus increasing plant-available soil water (Gabriel et al., 2021). Combining no-tillage practices that allow for integrating mulching and seeding into the untilled surface provide ways to further improve seedling germination (Xu et al., 2021). These practices protect the soil surface and limit soil drying, mitigating against crust formation. In the restoration context, new precision seeding technologies (e.g., Masarei et al., 2021) may provide similar benefits where opportunities to ameliorate the soil conditions themselves are limited.

Ecological impact of crusts

The modification of surface hydrology by soil crusts and seals influences the ecological functioning of landscapes. Crusted areas are integral to the “patchwork mosaic” concept of dryland ecosystem functioning (Aguiar and Sala, 1999), in particular the formation of “runoff-runon” units in which low infiltration capacity over crusted locations creates runoff which then infiltrates in the presence of vegetation within “runon” zones (Ludwig et al., 2005). This dynamic is shown in Fig. 8, which illustrated a situation where no infiltration occurs on crusted inter-patch areas, while within vegetated patches infiltration is enhanced relative to rainfall volumes at the upper edge of the patch (Crompton et al., 2019).

By supplying vegetated locations with water additional to that provided by precipitation alone, the runoff generated on sealed and crusted locations enables the persistence of vegetation in areas too arid to otherwise support it (Tongway and Ludwig, 1994). Infiltrating run-on also frequently contains additional nutrients, soil carbon and seeds which can fertilize and sustain the vegetation which traps them (Harman et al., 2014; Thompson et al., 2014). The recognition of the role that ‘runoff’ and ‘runon’ plays in the ‘Zai’ technique developed in Burkina Faso has led to a local farmer, Yacouba Sawadogo, to further develop it for both food crops and medicinal trees. In turn, this has contributed to countering desert encroachment and climate change (Ehlers and Goss, 2016; <https://www.youtube.com/watch?v=GYKOcq9cd1Q>; www.1080films.co.uk/yacoubamovie).

A consequence of the runoff-runon dynamic on sealed soils is that more vegetation cover is not necessarily “better” (Berghuis et al., 2020). Removing sealed areas can generate collapse of vegetation in drylands, as the plants cannot survive without the extra water. Vegetation cover that increases in response to land rehabilitation—e.g., removal of grazing pressures—may not reach 100%, but instead stabilize at a level where the water supplied by runoff provides sufficient resources to the plants. Thus, soil seals and crusts can be conceptualized as having a “dual role”—where too extensive they promote runoff which cannot be infiltrated and retained on the hillslope, creating a path to further degradation (Assouline et al., 2015). Where crusted areas are too small, by contrast, they will not supply enough water to enable full plant growth, effectively starving vegetation. This dynamic is shown in Fig. 9.

The importance of the seal layer specifically in generating these outcomes can be assessed by modeling sealed and unsealed soils on patchily vegetated hillslopes. Water content on hillslopes with and without seals, along with comparisons to field measurements, are shown in Fig. 10, which clearly illustrates the increased water content in sites where the hillslope soils were sealed (Sela et al., 2012).

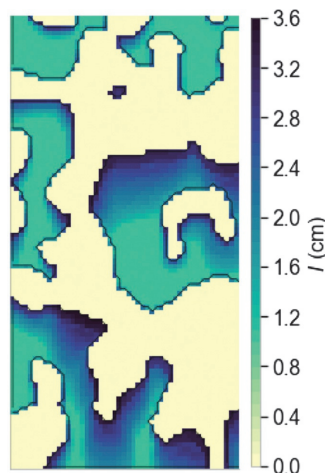


Fig. 8 Infiltration following a 5 cm/h storm on crusted soils with vegetated patches with a hydraulic conductivity of 8 cm/h as simulated with a coupled Saint Venant Equation—Richards Equation model. From Crompton O, Sytsma A and Thompson S (2019) Emulation of the Saint Venant equations enables rapid and accurate predictions of infiltration and overland flow velocity on spatially heterogeneous surfaces. *Water Resources Research* 55: 7108–7129.

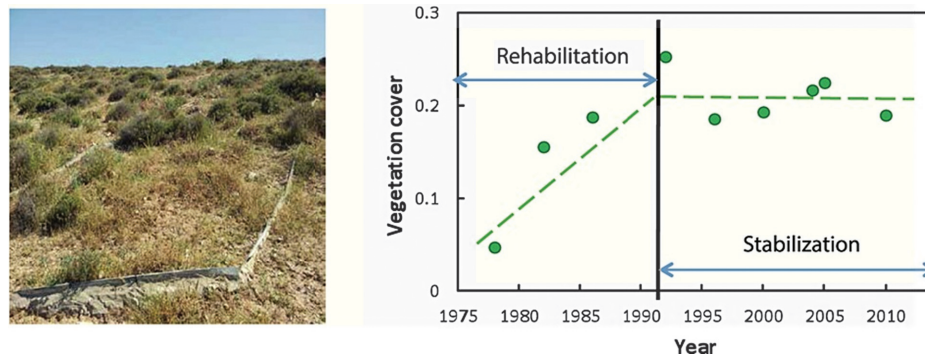


Fig. 9 Recovering runoff plots at the Lehavim Long Term Ecological Research site in the Negev Desert, Israel on the left. Over a period of 25 years recovery from overgrazing, the vegetation cover first increased before stabilizing at approximately 25%—allowing some 75% of the hillslope to act as a runoff “catchment” on the sealed soils there. After Assouline S, Thompson SE, Chen L, Svoray T, Sela S and Katul GG (2015) The dual role of soil crusts in desertification. *Journal of Geophysical Research – Biogeosciences* **120**: 2108–2119.

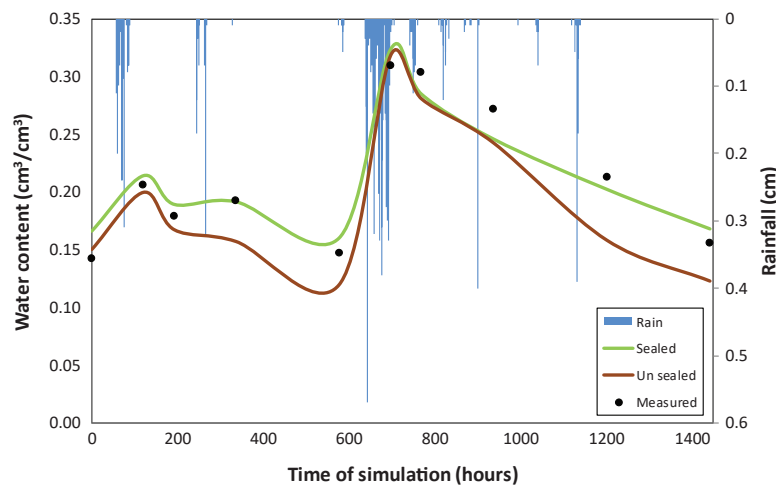


Fig. 10 Average water content for all measured samples (black dots) and the corresponding simulated values for sealed (green solid line) and unsealed (brown solid line) systems. Rainfall events are depicted by the blue columns. Based on Sela S, Svoray T and Assouline S (2012) Soil water content variability at the hillslope scale: Impact of surface sealing. *Water Resources Research* **48**(3): W03522.

At the hillslope scale, predicted cumulative infiltration is greatest where both seal layers and vegetation are present, and the runoff-runon dynamic is allowed to take place (Chen et al., 2013). This is shown in Fig. 11. When vegetation cover is significant, crusts can drive desertification, but this process is potentially self-limiting. For low vegetation cover, crusts mitigate against desertification by providing water subsidy to plant communities through a runoff-runon mechanism (Assouline et al., 2015; Svoray et al., 2021).

Conclusions

Soil seal layers and crusts introduce an additional dynamic into the infiltration process. In addition to a “typical” model in which infiltration rates are reduced as soils become more saturated, in soils with sealing, infiltration rates also decline as the structure of the soil surface itself is altered by the action of rainfall. Hydrologically seals reduce infiltration, promote runoff and can promote faster, more erosive flows on their usually smooth surface. Nearly 70 years of effort have produced high quality physical models to predict these effects at point scales.

Seal layers and crusts have confounding impacts on vegetation. In agricultural systems, their mechanical strength impedes the emergence of seedlings, reducing yields and requiring farmers mechanically, chemically or agronomically adapt to their presence. The proliferation of no till agriculture and the incorporation of no till, mulching and different targeted tillage approaches have all helped to mitigate the effects of crusts on agronomy. Similar challenges arise in other settings where seeding is required, including land and mine site remediation. Here, precision seeding and agricultural engineering techniques that can overcome soil compaction and crusting offer considerable scope to improve the economic efficiency of remediation activities.

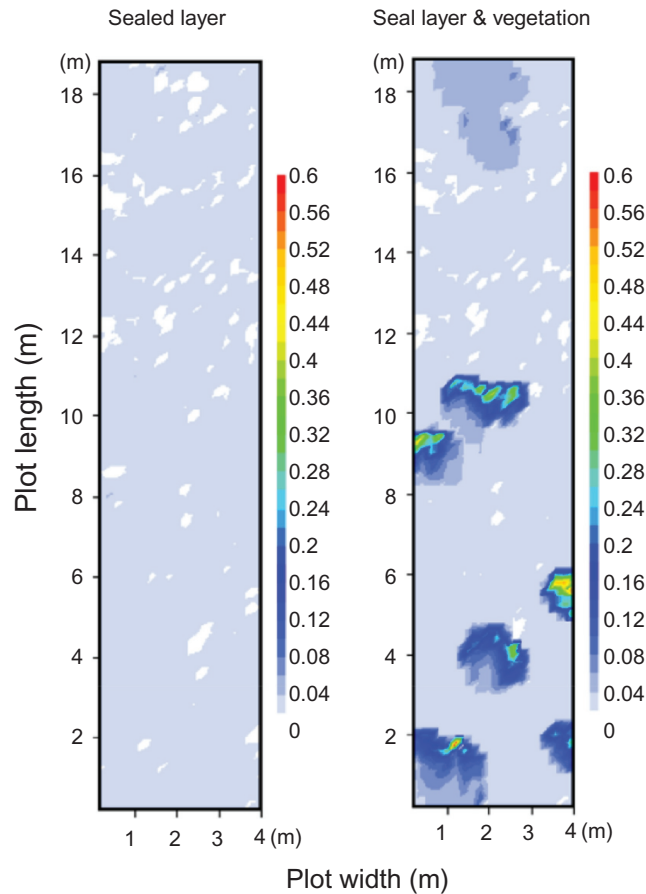


Fig. 11 Cumulative infiltration after a storm modeled on a hillslope with sealed soils (left) and sealed soils with vegetation patches present (right). The combination of sealing and vegetation leads to significantly greater cumulative infiltration than sealed bare soils especially in the vegetation patches. Based on Chen L, Sela S, Svoray T and Assouline S (2013) The role of soil-surface sealing, microtopography, and vegetation patches in rainfall-runoff processes in semiarid areas. *Water Resources Research* **49**: WR20360.

Ecologically, seal layers present a conundrum in that where too extensive they promote runoff of water, nutrients, seeds and other resources and their export from the landscape, contributing to degradation. However, landscapes without any sealed or impermeable areas force plants to depend on only the rain that falls above their roots for survival. In sufficiently dry landscapes, this can lead to seal layers being essential for the function and persistence of ecosystems. Thus, surface sealing contributes to the formation of local hydrological refugia (McLaughlin et al., 2017; Svoray et al., 2021) and form an integral physical component of functioning ecosystems.

References

- Aguilar MR and Sala OE (1999) Patch structure, dynamics and implications for the functioning of arid ecosystems. *Trends in Ecology & Evolution* 14(7): 273–277.
- Ahuja LR and Swartzendruber D (1973) Horizontal soil-water intake through a thin zone of reduced permeability. *Journal of Hydrology* 19: 71–89.
- Anzouman M, Christopher J, Mumford M, Dang YP, Menzies NW, and Kopitke PM (2018) Selection for rapid germination and emergence may improve wheat seedling establishment in the presence of soil surface crusts. *Plant and Soil* 426: 227–239.
- Assouline S (2004) Rainfall-induced soil surface sealing a critical review of observations, conceptual models, and solutions. *Vadose Zone Journal* 3(2): 570–591.
- Assouline S and Mualem Y (1997) Modeling the dynamics of seal formation and its effect on infiltration as related to soil and rainfall characteristics. *Water Resources Research* 33(7): 1527–1536.
- Assouline S and Mualem Y (2000) Modeling the dynamics of soil seal formation: Analysis of the effect of soil and rainfall properties. *Water Resources Research* 36: 2341–2349.
- Assouline S and Mualem Y (2002) Infiltration during soil sealing: The effect of areal heterogeneity of soil hydraulic properties. *Water Resources Research* 38: WR001168.
- Assouline S and Mualem Y (2006) Runoff from heterogeneous small bare catchments during soil surface sealing. *Water Resources Research* 42: W12405.
- Assouline S, Thompson SE, Chen L, Svoray T, Sela S, and Katul GG (2015) The dual role of soil crusts in desertification. *Journal of Geophysical Research – Biogeosciences* 120: 2108–2119.
- Augeard B, Bresson LM, Assouline S, Gaskuel C, Kao C, and Vauclin M (2008) Dynamics of soil surface bulk density: Role of water table elevation and rainfall duration. *Soil Science Society of America Journal* 72: 412–423.
- Baumhardt RL (1985) *The Effect of Rainstorm Characteristics on Soil Sealing and Infiltration*. Ph.D. thesis Mississippi State University.
- Baumhardt RL, Romkens MJM, Whisler FD, and Parlange JY (1990) Modeling infiltration into a sealing soil. *Water Resources Research* 26: 2497–2505.

- Beckett CTS, Glenn D, Bradley K, Guzzomi AL, Merritt D, and Fourie AB (2017) Compaction conditions greatly affect growth during early plant establishment. *Ecological Engineering* 106: 471–478.
- Ben-Hur M (2001) Soil conditioner effects on runoff and potato yield under sprinkler irrigation. *Agronomy Journal* 93: 1156–1163.
- Berghuis PMJ, Mayor AG, Rietkerk M, and Baudena M (2020) More is not necessarily better: The role of cover and spatial organization of resource sinks in the restoration of patchy drylands. *Journal of Arid Environments* 183: 104282.
- Bresson LM and Boiffin J (1990) Morphological characterisation of soil crust development stages on an experimental field. *Geoderma* 47: 301–325.
- Bresson LM, Moran CJ, and Assouline S (2004) Use of bulk density profiles from X-radiography to examine structural crust models. *Soil Science Society of America Journal* 68: 1169–1176.
- Brooks RH and Corey AT (1964) *Hydraulic Properties of Porous Media*. *Hydrology Papers*, vol. 3. Fort Collins: Colorado State University.
- Chaudhry KG and Das DK (1980) Seedling emergence of summer legumes through simulated soil crust. *Journal of the Indian Society of Soil Science* 28: 386–388.
- Chen L, Sela S, Svoray T, and Assouline S (2013) The role of soil-surface sealing, microtopography, and vegetation patches in rainfall-runoff processes in semiarid areas. *Water Resources Research* 49: WR20360.
- Crompton O, Sytma A, and Thompson S (2019) Emulation of the Saint Venant equations enables rapid and accurate predictions of infiltration and overland flow velocity on spatially heterogeneous surfaces. *Water Resources Research* 55: 7108–7129.
- Edwards WM (1977) *Soil Crusting*. vol. 57, North Central Region: US Agricultural Research Service, pp. 35–42.
- Ehlers W and Goss M (2016) *Water Dynamics in Plant Production*, 2nd edn. Wallingford: Cabi Publishing.
- Gabriel JL, García-González I, Quemada M, Martin-Lammerding D, Alonso-Ayuso M, and Hontoria C (2021) Cover crops reduce soil resistance to penetration by preserving soil surface water content. *Geoderma* 386: 114911.
- Garbout A, Armenise E, Ahn S, Doerr S, Ritz K, Simmons R, Sturrock C, and Mooney S (2014) Soil surface micro topology assessment using X-ray micro Computed Tomography. In: *4th International Conference on Surface Metrology, Hamburg, Germany*.
- Gerard CJ (1980) Emergence force by cotton seedlings. *Agronomy Journal* 72: 473–476.
- Harman CJ, Lohse KA, Troch PA, and Sivapalan M (2014) Spatial patterns of vegetation, soils, and microtopography from terrestrial laser scanning on two semiarid hillslopes of contrasting lithology. *Journal of Geophysical Research – Biogeosciences* 119: 163–180.
- Hillel D and Gardner WR (1969) Steady infiltration into crust topped profiles. *Soil Science* 108: 137–142.
- Hillel D and Gardner WR (1970) Transient infiltration into crust topped profiles. *Soil Science* 109: 69–76.
- Hopmans JW, Parlange J-Y, and Assouline S (2007) Infiltration. In: Delleur JW (ed.) *The Handbook of Groundwater Engineering*, 2nd edn. Boca Raton: CRC Press, Taylor and Francis Group. ch. 7.
- Le Bissonnais Y and Singer MJ (1992) Crusting, runoff, and erosion response to soil water content and successive rainfalls. *Soil Science Society of America Journal* 56: 1898–1903.
- Levy GJ and Rapp I (1998) Polymer effects on surface mechanical strength of a crusting loessial soil. *Australian Journal of Soil Research* 37: 91–101.
- Ludwig JA, Wilcox BP, Breshears DD, Tongway DJ, and Imeson AC (2005) Vegetation patches and runoff-erosion as interacting ecohydrological processes in semiarid landscapes. *Ecology* 86(2): 288–297.
- Madsen MD, Davies KW, Williams CJ, and Svejcar TJ (2012) Agglomerating seeds to enhance native seedling emergence and growth. *Journal of Applied Ecology* 49: 431–438.
- Masarei MI, Erickson TE, Merritt DJ, Hobbs RJ, and Guzzomi AL (2021) Engineering restoration for the future. *Ecological Engineering* 159: 106.
- McIntyre DS (1958a) Permeability measurements of soil crusts formed by raindrop impact. *Soil Science* 85: 261–266.
- McIntyre DS (1958b) Soil splash and the formation of surface crusts by raindrop impact. *Soil Science* 85: 185–189.
- McLaughlin BC, Ackerly DD, Klos PZ, Natali J, Dawson TE, and Thompson SE (2017) Hydrologic refugia, plants, and climate change. *Global Change Biology* 23: 2941–2961.
- Morin J and Benyamini Y (1977) Rainfall infiltration into bare soils. *Water Resources Research* 13: 813–817.
- Mualem Y and Assouline S (1989) Modeling soil seal as a non-uniform layer. *Water Resources Research* 25: 2101–2108.
- Nizami MI and Khan NA (1991) Soil crust effects on plant emergence and grain yield of maize. *Pakistan Journal of Agricultural Research* 12: 169–176.
- Pareja- Sánchez E, Ramos C, Lampurlanes J, Fuentes J, and Martínez C (2017) Long-term no-till as a means to maintain soil surface structure in an agroecosystem transformed into irrigation. *Soil and Tillage Research* 174: 221–230.
- Parker JJ and Taylor HM (1965) Soil strength and seedling emergence relations. 1. Soil type, moisture tension, temperature and planting depth effect. *Agronomy Journal* 57: 289–291.
- Richards LA (1953) Modulus of rupture as an index of crusting of soil. *Soil Science Society of America Proceedings* 17: 321–323.
- Romkens MJM, Baumhardt RL, Parlange MB, Whisler FD, Parlange JY, and Prasad SN (1986) Rain-induced surface seals: Their effect on ponding and infiltration. *Annals of Geophysics* 4: 417–424.
- Roth CH (1997) Bulk density of surface crusts: depth functions and relationships to texture. *Catena* 29: 223–237.
- Sela S, Svoray T, and Assouline S (2012) Soil water content variability at the hillslope scale: Impact of surface sealing. *Water Resources Research* 48(3), W03522.
- Shainberg I (1992) Chemical and mineralogical components of crusting. In: Sumner ME and Stewart BA (eds.) *Soil Crusting: Chemical and Physical Processes*, pp. 33–54. Boca Raton, FL: Lewis Publishers.
- Sharma DP and Agarwal RP (1974) Seedling emergence behaviour of bajra, cotton relationship with the modulus of rupture in alluvial soils. *Journal of the Indian Society of Soil Science* 28: 119–121.
- Sinha AK and Ghildyal BP (1979) Emergence force of crop seedlings. *Plant and Soil* 51: 153–156.
- Souty N, Stengel P, Rode C, and Tutubene R (1992) A mechanistic study of maize emergence through superficial crusts. *Soil and Tillage Research* 23: 125–140.
- Svoray T, Sela S, Chen L, and Assouline S (2021) Lateral flow and contributing area control vegetation cover in a semiarid environment. *Water Resources Research* 57: WR030998.
- Tackett JL and Pearson RW (1965) Some characteristics of soil crusts formed by simulated rainfall. *Soil Science* 99: 407–413.
- Thompson SE, Assouline S, Chen L, Trahtenbrot A, Svoray T, and Katul GG (2014) Secondary dispersal driven by overland flow in drylands: Review and mechanistic model development. *Movement Ecology* 2: 1–13.
- Tongway DJ and Ludwig JA (1994) Small-scale resource heterogeneity in semi-arid landscapes. *Pacific Conservation Biology* 1: 201–208.
- van Genuchten MT (1980) A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal* 44: 892–898.
- Xu C, Li R, Song W, Wu T, Sun S, Shen W, Hu S, Han T, and Wu C (2021) Integrating straw management and seeding to improve seed yield and reduce environmental impacts in soybean production. *Agronomy* 11: 1033.

Soil mechanics

Markus Berli^a and Paul D Hallett^b, ^aDivision of Hydrologic Sciences (DHS), Desert Research Institute, Las Vegas, NV, United States; ^bSchool of Biological Sciences, University of Aberdeen, Aberdeen, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

This is an update of S.K. Upadhyaya, STRESS–STRAIN AND SOIL STRENGTH, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 38–54, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00507-5>.

Introduction	71
A brief history of soil mechanics	71
The concept of stress	71
Total versus effective stress	73
The concept of strain	73
Constitutive relationships for soil materials	75
Linear-elastic soil behavior	77
Elastoplastic soil behavior	77
Elastoplastic soil behavior under compressive stresses	77
Soil failure - shear	78
Soil failure - fracture	80
Measuring soil mechanical behavior in the field	81
Cone penetrometer	81
Soil sinkage	82
Soil shear	82
Soil mechanics and land management	83
Outlook	84
References	84

Abstract

Soil mechanics describes the deformation of soils by mechanical stresses. Above critical stress levels, soils deform plastically so that the pore structure does not recover upon unloading. Soil moisture has a large impact on the mechanical behavior of soil. Dry soils are fairly elastic, moving toward elastoplastic behavior for moist soils, and then viscous flow for wet soils. Theories adopted from materials science and geotechnical engineering describe the response of soils to complex stresses under shear, compression or tension. However, they are insufficient to describe structurally complex, unsaturated soils typical near the surface.

Glossary

Effective stress Stress experienced in a soil due to the combined impacts of the applied mechanical stress, stress from water in the soil such as capillarity and air pressure.

Elastic A material that recovers strain completely upon removal of a mechanical stress.

Elasto-plastic A material that only partially recovers strain up the removal of a mechanical stress, with irrecoverable, plastic strain occurring beyond a critical stress level.

Mohr-Coulomb failure criterion Constitutive relationship describing the shear failure of a soil experiencing both shear and compressive stresses. It describes soil from its cohesion and internal friction between particles.

Strain Volumetric deformation of a material, often caused by an applied stress.

Stress The force acting per unit area, expressed as Pa. They can either be in shear, compression or tension.

Virgin Compression Curve Relationship between changes to void ratio or porosity with increasing compressive stress that has an initial elastic component (recompression line) with little deformation followed by greater plastic deformation (virgin compression line) above a precompression that defines the strength of the soil.

Viscous Time-dependent deformation of a material caused by its flow under a mechanical stress.

Key points

- Constitutive relationships describe the deformation (strain) of soil under mechanical stresses.
- The mechanical behavior of soil depends on soil moisture because of capillarity stresses that mechanically bond particles and the impact to plastic and viscous deformation.

- Soil pore structure is modified by mechanical deformation with both negative (e.g. compaction) and positive (fracture) impacts to the functioning of soil.
- Surface soils are structurally complex, making it difficult to apply soil mechanics theory originally developed for deeper, saturated and less structured soils.
- It is possible to modify the mechanical properties of soil by management, either by intentional fragmentation by tillage or through biological processes that bond and aggregate soil particles.

Nomenclature

ϕ	Angle of internal friction in Mohr-Coulomb failure criterion
χ	Parameter to relate matric potential (or mean hydrostatic stress) to effective stress
ν	Poisson's ratio
τ	Shear stress (kPa)
ϵ	Strain (m m^{-1})
σ	Stress (kPa)
σ_{1c}	Precompression stress (kPa)
ρ_b	Soil bulk density (kg m^{-3})
γ_{oct}	Deviatoric strain
τ_{oct}	Octahedral shear stress (kPa)
ρ_s	Soil particle density (kg m^{-3})
A	Area (m^2)
c	Cohesion parameter in Mohr-Coulomb failure criterion
E	Elastic or Young's modulus (kPa)
e	Void ratio ($\text{m}^3 \text{m}^{-3}$)
F	Force (N)
k	Sinkage parameter in the Bernstein equation of plate sinkage
K_{IC}	Fracture toughness ($\text{kPa m}^{1/2}$)
l	Length (m)
m	Constant in the Bernstein equation of plate sinkage
n	Porosity ($\text{m}^3 \text{m}^{-3}$)
p	Mean hydrostatic stress (kPa)
q	Deviatoric stress (kPa)
S_r	Degree of saturation ($\text{m}^3 \text{m}^{-3}$)
u	Deformation by stress (m)
Y	Geometry factor in fracture toughness equation

Subscripts

0	Initial condition
1, 2, 3	Subscripts for triaxial stress or strain. 1 is vertical; 2 and 3 are radial
a	Air
n	Normal direction (generally compression)
p	Pores
s	Soil
t	Shear direction
t	Solids
t	Tensile
v	Volumetric
w	Water

Introduction

Soil mechanics is a sub-discipline of soil science and geotechnical engineering that deals with the mechanical properties and processes of soils. Generally speaking, soil mechanics describes how the soil and its pore structure changes its shape or volume (or how it “deforms”) due to stresses acting on the soil. There are a wide range of soil mechanics applications in soil science, especially to predict and explain soil responses to the stresses exerted by compaction and tillage. The application of soil mechanics to soil science also extends to explore the physics of soil structure formation, shrinkage and swelling, erosion resistance, restrictions to root growth and bioturbation, and reinforcement of soil by plant roots.

Many soil mechanics concepts and methods used in soil science were initially developed to solve geotechnical engineering problems. Key issues for geotechnical engineering have some similarities to soil science. Building foundations, for instance, need to withstand the weight of buildings, whereas agricultural soils may need to withstand the weight of machinery or livestock. Evaluating geotechnical slope stability from its shear characteristics uses similar theories used by soil scientists to measure soil stability and wheel slip. Traditional geotechnical engineering, however, has focused on subsurface soils below the “dry crust.” These soils are mostly water-saturated and they are often assumed to be structureless. Describing these soils has been simplified to their component particle sizes (sand, silt, clay) and bulk density, and they tend to have mechanical properties that change little over time.

Soil scientists, however, are more interested in surface rather than subsurface soils. Surface soils are typically unsaturated, structured (i.e., have properties that are not sufficiently described in terms of particle sizes and bulk density only) and have much more dynamic properties than subsurface soils due to their greater exposure to climate and biota. This complicates understanding and a “unified theory” or “standard model” for the mechanics of unsaturated, structured surface remains elusive. Nevertheless, the basic concepts of soil mechanics such as stress, strain and deformation apply to unsaturated, structured surface soils.

This chapter provides an introduction to soil mechanics for a soil science audience. The chapter covers basic concepts taken from geotechnical engineering, and then describes methods adopted in soil science to solve practical problems involving unsaturated, structured surface soils.

A brief history of soil mechanics

Soil mechanics has been of interest for millennia, not so much as the research discipline we know today but as practical knowledge for a range of applications. Think first of early farmers who discovered the benefit of using a stick or later a plow to break up the soil for seed bed preparation, potters who took advantage of the moldability of clay, or early engineers in search of stable ground for constructing houses, roads and other structures. They were interested in either deforming and moving soil in a desired way (“soil dynamics”) or, the opposite, making sure the soil stayed in place (“soil statics”), the key problem for slope stability and foundation engineering. These early forms of soil mechanics were based on trial and error, but nevertheless effective and quite successful since the plow, pottery and principles of foundation engineering are still in use today.

It was not until the 18th century that soil mechanics was studied from a more theoretical angle, starting with the works by [Coulomb \(1776\)](#), later [Rankine \(1857\)](#) and ultimately [Terzaghi \(1925\)](#) and [Terzaghi and Peck \(1948\)](#), who is often considered the father of modern soil mechanics. Terzaghi applied the concepts of stress, strain and strength to soil as a deformable, water-saturated porous medium. Soils deform and eventually fail when exposed to increasing stress levels. The different strands of soil mechanics were brought together in what is known as Critical State Soil Mechanics (CSSM) ([Schofield and Wroth, 1968](#)), which provided a first framework that could describe the different forms of strain, deformation and failure that saturated soils undergo when exposed to external loads. Concurrently to the development of CSSM, efforts were made to describe the mechanics of soils that were not fully water-saturated, commonly termed a “unsaturated soil mechanics” ([Fredlund and Rahardjo, 1993](#); [Fredlund et al., 2012](#); [Lu and Likos, 2004](#)).

Soil scientists contributed to the development of geotechnical engineering principles, particularly on the behavior of unsaturated soils, resulting in the two disciplines starting to converge again (see e.g., [Lu and Likos, 2004](#)). Progress in material sciences and increasing computational capabilities allow for soil mechanics based on the principles of multi-physics and porous media theory, which couple mechanical, hydraulic and thermodynamic processes and properties of the soil ([Andrade and Mital, 2019](#); [de Boer, 2000](#); [Navarro et al., 2014](#)). Structured, unsaturated and biologically active surface soils can now be explored.

The concept of stress

In soil mechanics, stress is defined as a force, F , acting per unit area, A , on a body of soil. If F acts normal (i.e. perpendicular) to A , then one refers to that stress a “normal stress” σ defined as:

$$\sigma = F_n/A \quad (1)$$

If F acts parallel to A , then a body experiences a “shear stress”, τ defined as:

$$\tau = F_t/A \quad (2)$$

where F_n and F_t are the normal and shear forces, respectively. The stresses σ and τ have the units of a pressure (N/m^2 or Pa).

Because soil is a three-dimensional body, stresses can act with different magnitude and in different directions, dictated by how the applied stress is transmitted through structurally complex soils. The major stress directions experienced in soils are compression, shear and tension (Fig. 1). Compression stresses act normal to the center of the soil body and can be caused by the weight of machinery or livestock, or the weight of overburden soil. Soil compaction vulnerability can be quantified from the decrease in soil volume with increasing compression stress. Shear stresses slide a volume of soil against another volume, such as from the action of wheel slip, livestock trampling or tillage, leading to shape changes but not necessarily volume changes of the soil. Tension stresses pull the soil apart, either from an internal force such as shrinkage when the soil dries or through the action of an external force induced for example by tillage.

Tension, shear or compression stresses as depicted in Fig. 1 are ideal cases where stress only acts in one direction or within a two-dimensional plane. In practice, however, more complex stress situations occur in soils, such as the combined action of compression and shearing from wheels or livestock at the soil surface. In addition, since soils are three dimensional bodies, a combination of normal and shear stresses act in a three-dimensional space. Fig. 2 illustrates a stress situation where there is one compressive normal stress, termed σ_1 , acting vertically on a soil cylinder as well as two compressive normal stresses termed σ_2 and σ_3 acting radially on the soil cylinder. The stress state shown in Fig. 2 is called “triaxial” with σ_1 , σ_2 , and σ_3 the three principal stresses. In Fig. 2 a common assumption is made that $\sigma_2 = \sigma_3$, which further simplifies our stress considerations.

For a triaxial stress situation, there are two important stresses that define the stress state of the cylindrical body and that capture the interplay of the three principal stresses σ_1 , σ_2 , and σ_3 . The first important stress is the mean hydrostatic stress, denoted here as p , as:

$$p = \frac{\sigma_1 + \sigma_2 + \sigma_3}{3} = \frac{\sigma_1 + 2\sigma_3}{3}, \quad (3)$$

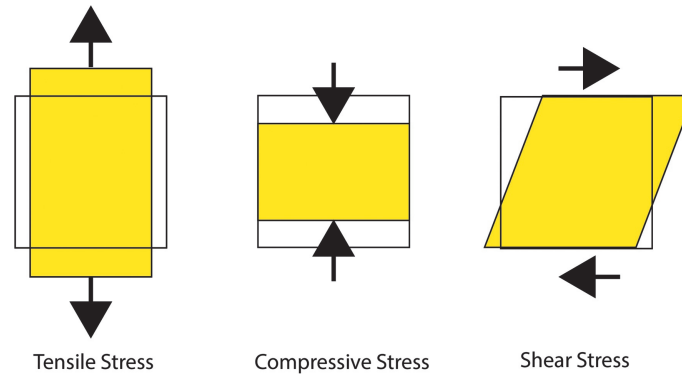


Fig. 1 Normal and shear stresses acting on a two-dimensional body of soil.

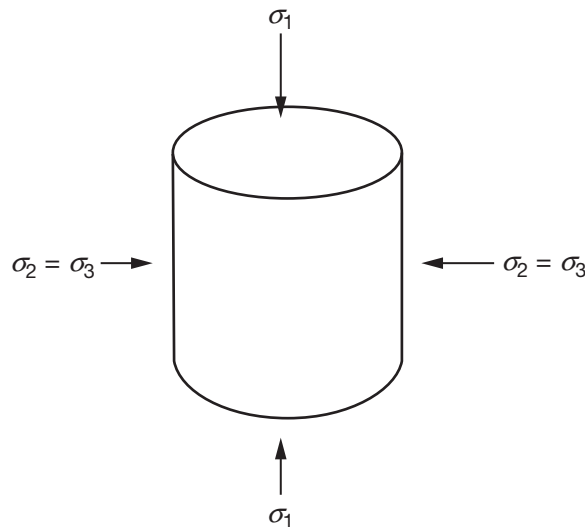


Fig. 2 Principal stresses σ_1 , σ_2 , and σ_3 acting on a cylindrical body of soil.

assuming $\sigma_2 = \sigma_3$ as shown in Fig. 2. The hydrostatic stress in a soil is similar to the pressure found inside of a column of a liquid such as water (therefore the name “hydrostatic” stress). The second important stress is the octahedral shear stress, denoted τ_{oct} :

$$\tau_{oct} = \frac{\sqrt{2}}{3}(\sigma_1 - \sigma_3) \quad (4)$$

Some soil mechanics models use the deviatoric stress, q , instead of the octahedral shear stress, τ_{oct} , with q defined as:

$$q = \sigma_1 - \sigma_3. \quad (5)$$

In simple terms, the triaxial stress state of a cylindrical body of soil can be described by the mean hydrostatic stress, p , and either the octahedral stress, τ_{oct} , or the deviatoric stress q . This is important because the mean hydrostatic stress controls volume changes (such as soil compaction) and the deviatoric (or octahedral) stress controls shape changes of the soil (such as breaking up soil by tillage, movement of an unstable slope or debris flow). In other words, shear deformation of the soil is caused by the difference between the principal stresses, whereas volume changes (such as compaction) are caused by the average of the three principal stresses.

Total versus effective stress

Soil is a three-phase system, consisting of solid particles, and pores filled with variable amounts of air and water depending on water content. This further complicates the mechanical characterization of the soil as stress can occur in each of these three phases. The total stress that the soil undergoes can be divided between solid, water and air stresses using the concept of effective stress. The mean hydrostatic stress, p , as given in Eq. (3), is an effective stress made up of the stress acting on the solids, p_t , pore water, p_w and pore air, p_a , simply defined by the sum of these stresses:

$$p = p_t + (p_w + p_a), \quad (6)$$

where p_a is often zero as it is irrelevant in saturated soil and rapidly equilibrates with atmospheric pressure in unsaturated soils. Stresses in the soil caused by external loads such as the weight of a building or a vehicle are total stresses, not effective stresses, and the effective stresses they cause depend on the water and air pressure. This means that part of the total stress, p_t , in the soil caused by a load applied on the soil surface can be offset by the pore water pressure, p_w . Hence, the effective stress, p , may be small or even zero and, therefore, little to no strain or deformation will occur even though the soil surface is loaded. If the load at the soil surface is applied long enough, then the pore water pressure may dissipate with time through drainage, causing the effective stress to increase and some slow but steady soil deformation under the load called “consolidation.” Consolidation is of great importance in geotechnical engineering due to foundations (e.g., Wesley, 2010).

For unsaturated soil, Eq. (6) still holds but needs to be extended. Under unsaturated conditions, p_w changes from positive to negative pressure or “tension.” Consequently, the effective stress acting on unsaturated soil is no longer smaller (or equal) to the total stress but can be greater than the total stress. One example of this phenomena is that moist, unsaturated sand is generally “firmer” than saturated, wet sand because of the “sandcastle effect.” A sandcastle built with wet, water saturated sand collapses, but moist, unsaturated sand holds together. If the sand dries more, water holding together the particles is lost and the sand gets weaker. The strength of moist sand is caused by capillary forces induced by the interface between water and air within the pores of the sand, affected by p_w , which is analogous to water potential.

In an unsaturated soil, p_w acts only in water-filled and not air-filled pores. This can be accommodated in Eq. (6) by a factor, χ that is related to the amount of water:

$$p = p_t - \chi p_w \quad (7)$$

known as Bishop’s effective stress equation (Bishop, 1959). Although χ depends on the saturation degree, S_r , or the moisture content, θ , of soil, it is an empirical parameter derived from mechanical tests. For fully water saturated soil, $\chi = 1$, simplifying Eq. (7) for unsaturated soil into Eq. (6) for saturated soil. For air-dry soil, χ is assumed to equal to 0, so the effective mean hydrostatic stress, p , is equal to the total mean hydrostatic stress, p_t . In between these two moisture extremes, $0 < \chi < 1$.

If an applied stress is much greater than the water potential of the soil, a total rather than an effective stress approach can be used as χp_w will have a small impact. To illustrate this, let’s assume the total stress in the soil under a heavy load is 200 kPa. Let’s also assume that soil below the load is unsaturated with a water content around field capacity, which is equivalent to a water pressure in the soil pores around –10 kPa. For these conditions, according to Eq. (7), the effective mean hydrostatic stress, p , will be less than 5% higher than the total mean hydrostatic stress, p_t , because we know that for unsaturated conditions $\chi < 1$, which means that the absolute value of χp_w is less than 10 kPa. Therefore, with the total stress approach, it can be assumed that the total stress in the soil is approximately equal to the effective stress. When using the total stress approach, it is important to consider that the material properties of the soil become moisture-dependent (Berli et al., 2015).

The concept of strain

When a body of soil is subjected to sufficiently large external forces, it deforms. The ratio of these deformations to the original dimensions of the body are known as ‘strain.’ As a simple one-dimensional example, let’s assume we have a slender rod of length

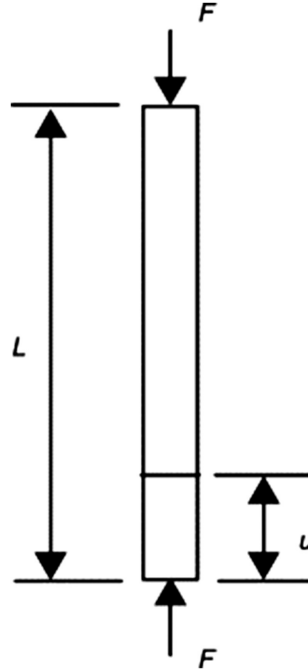


Fig. 3 A slender rod undergoing a compressive force, F , leading to compressive deformation, u , and compressive strain, ε .

l that is subjected to a uniaxial compressive force, F (Fig. 3). Due to the force, the length of the rod will decrease by a certain amount of compressive deformation, u . The compressive strain, ε , of the rod can be then defined as:

$$\varepsilon = u/l \quad (8)$$

This concept of one-dimensional strain can be extended to three dimensions for a cylindrical body of soil undergoing triaxial stresses as shown in Fig. 2. For each of the three principal stresses, σ_1 , σ_2 , and σ_3 , a corresponding principal strain ε_1 , ε_2 , and ε_3 , can be defined to represent the strains of the cylindrical body as a reaction to the corresponding principal stresses, σ_1 , σ_2 , and σ_3 .

Similar to the mean hydrostatic and deviatoric stresses, one can define a corresponding volumetric strain, ε_v , shear strain, ε_s , and deviatoric strain, γ_{oct} , respectively, given as:

$$\varepsilon_v = \varepsilon_1 + \varepsilon_2 + \varepsilon_3 \quad (9)$$

$$\varepsilon_s = \frac{2}{3}(\varepsilon_1 - \varepsilon_3) \quad (10)$$

$$\gamma_{oct} = \frac{\sqrt{2}}{3}(\varepsilon_1 - \varepsilon_3) \quad (11)$$

assuming $\sigma_2 = \sigma_3$ and, hence, $\varepsilon_2 = \varepsilon_3$. The volumetric strain, ε_v , is important because it is a measure of soil compaction. Changes in void ratio, Δe , of the soil can be used to calculate ε_v by:

$$\varepsilon_v = -\frac{\Delta e}{1 + e_0} \quad (12)$$

with e_0 the initial void ratio of the soil. The void ratio of the soil is defined as the ratio of total pore space of a soil, V_p , to the total volume of the solid space of the soil, V_s , according to:

$$e = \frac{V_p}{V_s} \quad (13)$$

Note the negative sign in Eq. (12) owing to the sign convention that treats compressive strains (and stresses) as positive. Void ratio is commonly used in soil mechanics because the denominator, V_s , does not change with strain. This is different from other areas of soil science where porosity or bulk density are used to describe the volumetric strain, ε_v . All these volumetric strain parameters are related. Void ratio, for example, can be easily calculated from other soil physical properties such as porosity, n , by:

$$e = \frac{n}{1 - n} \quad (14)$$

Or dry bulk density, ρ_b :

$$e = \frac{\rho_b}{\rho_s} - 1 \quad (15)$$

with ρ_s the particle density of the soil.

Eqs. (9) to (11) provide the foundation to calculate strain (and by extension deformation, if needed) of a body of soil undergoing triaxial stress conditions as given by Eqs. (3) to (5). What is still missing are the constitutive relationships to describe how stress and strain of the soil are related, one of the key tasks of soil mechanics. The constitutive relationships depend on the texture, moisture and structure of the soil.

Constitutive relationships for soil materials

Soil constitutive relationships describe how stress and strain are related, why different soils respond differently to the same level of stress or why the soil deforms differently depending on the level of applied stress. At low levels of stress, the corresponding strains are small and can disappear once the stress is removed. This type of stress-strain behavior is called “elastic”, like a metal spring that bounces back to its original size and shape once the stress is released. Above a yield stress, strain is irrecoverable after the stress has been removed. With increasing stress, more irrecoverable strain occurs. This type of stress-strain behavior is called “elasto-plastic.” When the stress level on the soil is further increased, the body of soil “fails.” Soil failure can be desired, such as in the case of plowing where the compacted surface soil is broken up into looser clods by shear failure of the soil at the plowshare. Soil failure, however, is often undesired, such as a failing foundation under a building, slope failure of a dam or a landslide, sometimes with devastating consequences for life and property.

Therefore, the strain response of soil to an applied stress can be elastic (recoverable), plastic (irrecoverable) or failure, depending on the mechanical properties of the soil and the amount of stress that is applied. Above a yield stress, plastic deformation occurs. Above the failure stress, failure occurs. Fig. 5 shows these various forms of strain responses of soil to applied stress, expressed in stress-strain diagrams. Linear-elastic stress-strain behavior shown in Fig. 5A is an ideal case where there is a linear increase in strain with increasing stress, with complete recovery of strain when the stress is removed. Soil rarely exhibits linear-elastic behavior (Fig. 5A), so a nonlinear-elastic constitutive relationship (Fig. 5B) describes soil loading behavior along a particular loading or unloading path.

Rigid-perfectly plastic (Fig. 5C) and elastic-perfectly plastic (Fig. 5D) models, respectively, are sometimes used in soil mechanics, particularly in limit equilibrium analysis to assess at what level of stress a soil fails. Fig. 5E represents the typical stress-strain behavior of soil under triaxial compression, described earlier in Figs. 2 and 4. Stress in Fig. 5E would correspond to the mean

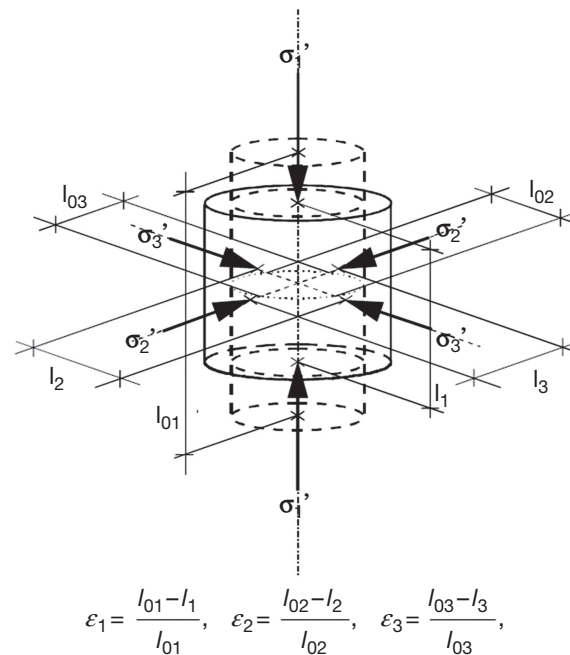


Fig. 4 Principle strains, ϵ_1 , ϵ_2 , and ϵ_3 of a cylindrical soil sample undergoing principle stresses σ_1 , σ_2 , and σ_3 . ϵ_1 , ϵ_2 , and ϵ_3 are defined in terms of the dimensional changes, l_1 , l_2 , l_3 of the cylindrical samples in the three principle directions relative to the original dimensions, l_{01} , l_{02} , l_{03} . Adapted from Berli M (2001) *Compaction of agricultural subsoils by tracked heavy construction machinery*. Institute of Terrestrial Ecology (ITOe). Swiss Federal Institute of Technology (ETH), Zurich.

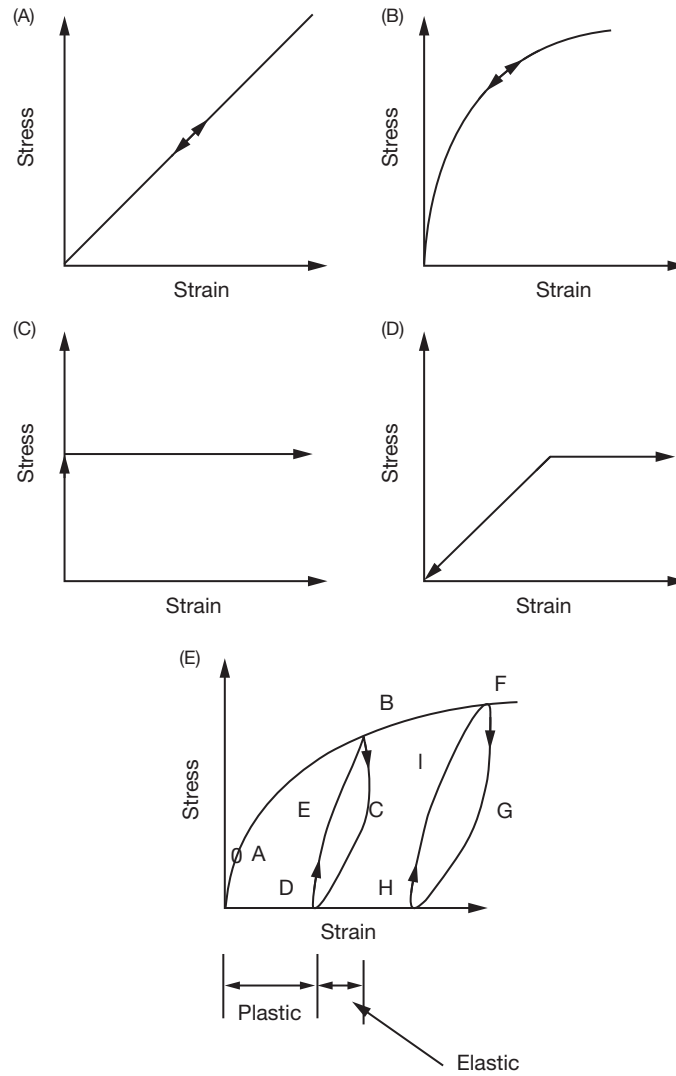


Fig. 5 Idealized stress–strain curves for solid materials: (A) linear-elastic; (B) nonlinear elastic; (C) rigid-perfectly plastic; (D) elastic-perfectly plastic; (E) and elastoplastic behavior.

hydrostatic stress given in Eq. (5) and strain to the volumetric strain provided by Eq. (9). In this example, a cylindrical soil specimen as shown in Figs. 2 and 4 exhibits elastic behavior until the applied stress reaches the yield stress (point A) on the soil stress–strain curve. The specimen will no longer deform elastically but will start to deform plastically if further stress is applied from point A to point B. The other letters indicate cycles of loading and unloading. These follow a different path due to plastic deformation, with some elastic recovery as seen from C–D or G–H in Fig. 5E. When loaded again the stronger, plastically deformed soil responds differently until the applied stress reaches the maximum stress applied in the previous loading cycle, where the stress–strain relationship returns to the same path as the initially unloaded specimen.

Beyond point F in Fig. 5E mechanical behavior depends on how the mean hydrostatic stress (Eq. 3) and the deviatoric stress (Eq. 5) evolve with increasing stress level. If primarily the mean hydrostatic stress, p , increases with increasing stress level while the deviatoric stress, q , remains low (what “low deviatoric stress” exactly means will be further elaborated in the section on failure criteria), then the density of the soil specimen will increase (or “harden”, as it is usually referred to in material science). The extreme case occurs if all three principal stresses are of the same magnitude and the deviatoric stress equals zero. Under these hydrostatic stress conditions, the soil specimen undergoes pure compaction, and the soil will become denser and denser with increasing applied stress. If, however, q in the soil specimen is high (again, “high deviatoric” will be further elaborated in the section on failure criteria), then the density of the soil specimen will decrease (or “loosen”) with increasing applied stress and the specimen will eventually “fail” when the deviatoric stress meets the failure criteria. In a nutshell, the fate of the cylindrical soil specimen very much depends on the relationship between p and q .

The stress-strain relationships shown in Fig. 5E are helpful to understand the different stages of stress-strain behavior a soil undergoes with increasing stress level. It is important to remember, however, that the curves shown in Fig. 5 are primarily conceptual models. When dealing with surface rather than subsurface soils, the mechanical response may be more complex due to the impacts of soil structure. Moreover, these curves assume that the relationship between stress and strain does not depend on time.

Many stresses acting at the soil surface, however, are very dynamic. These can be due to rapid imposed stresses, such as from a moving wheel or livestock, or slow processes such as the gradual slumping of a seedbed. Under a rapid dynamic stress, a delayed, viscous strain can occur, so the soil may deform very little even if exposed to stresses much higher than yield stress. The viscous response of the soil to stress also explains why multiple exposure of the soil to the same stress, such as by repeated wheeling by machinery, may result in increasing strain, even though the individual stress event is short. The viscous deformations that each pass causes are additive so after a sufficient amount of passes a sizeable amount of deformation can occur, even though the deformation from a single pass was small.

Viscous processes also have a large impact on the coalescence of seedbeds, and the gradual elongation of plant roots through soil. The time-dependency of soil stress-strain relationships has been largely neglected in soil science, only addressed in a few studies (Berli et al., 2006; Berli and Or, 2006; Ghezzehei and Or, 2001; Vyalov, 1986) despite its great importance for the mechanics of surface soils. Mechanical behavior can be described by “elasto-viscoplastic” stress-strain relationships where the elastic deformation is (almost) instantaneous, but the plastic deformation, at a sufficiently high stress level such as higher than yield stress, takes time to develop. For instance, pottery clay is an elasto-viscoplastic material, so it takes time to mold the lump of clay into a new shape even if a good amount of stress is applied to the lump.

Linear-elastic soil behavior

Even though linear elasticity is rarely observed in natural soils, many models and descriptions of soil mechanical processes assume linear-elastic behavior (Fig. 5A). Early theories of material mechanics assumed linear elasticity, and the simple stress-strain relationship makes it an ideal starting point for an introduction to soil mechanics.

Eqs. (1) and (6) define stress and strain for a one-dimensional (in soil mechanics also called “uniaxial”) compressive stress as shown in Fig. 1. For a linear elastic material, the following simple linear equation relates compressive stress, σ , with strain, ϵ :

$$\sigma = E\epsilon \quad (16)$$

with E the elastic or Young's modulus of the material. For a three-dimensional body, as shown in Figs. 2 and 4, undergoing a triaxial stress with $\sigma_1 > \sigma_2 = \sigma_3$, the axial compression of the specimen leads to a radial expansion of the specimen (Fig. 4). This means that the cylinder becomes “shorter” and “thicker” under triaxial stresses. The relationship between axial, ϵ_1 , and radial strain, ϵ_3 , assuming $\epsilon_2 = \epsilon_3$, is defined by the Poisson's ratio of the material, ν , given as:

$$\nu = -\frac{\epsilon_3}{\epsilon_1} \quad (17)$$

Note the minus sign in Eq. (15) indicating that if the compressive strain ϵ_1 is positive, according to the sign convention in soil mechanics, the “expansive strain” ϵ_3 has a negative value. With a uniform, linear-elastic body, two material properties, E and ν , fully describe the stress-strain behavior of the body. For a triaxial stress state, Eqs. (16) and (17) allow the relating of p and ϵ_v from Eqs. (3) and (12), as well as q and ϵ_s from Eqs. (5) and (10) according to:

$$p = \frac{E}{3(1-2\nu)}\epsilon_v \quad (18)$$

$$q = 3 \left[\frac{E}{2(1+\nu)} \right] \epsilon_s \quad (19)$$

So, under the assumption that soil can be described as a uniform, linear elastic body undergoing triaxial stress, all of soil mechanics can be described by Eqs. (18) and (19). More complex relationships are needed to describe elasto-plastic behavior, as introduced in the next section.

Elastoplastic soil behavior

This section outlines elastoplastic behavior of soil under compressive stresses, as illustrated in Fig. 5E. Mohr-Coulomb failure criterion is introduced to describe mechanical conditions when stresses are adequate to cause failure.

Elastoplastic soil behavior under compressive stresses

Testing of soil compressive stress can use a triaxial device, as illustrated in Fig. 4, or confined compression tests where the sample is restricted from moving axially, so $\epsilon_2 = \epsilon_3 = 0$ and only the axial strain, ϵ_1 , remains. Confined uniaxial compression tests are popular

to study the mechanics of surface soils because they are simpler to run than full triaxial tests and, for most practical purposes, provide sufficient information about the elasto-plastic behavior. Fig. 6 illustrates the idealized behavior of a soil sample under confined compression. Here the axes are flipped, with axial strain or its equivalence as void ratio (Eq. 12) on the y axis, and the logarithm of the applied compressive stress on the x axis. Plotting this way is the convention in soil mechanics. Two straight dashed lines approximate two stages of mechanical behavior. An initial recompression curve is the elastic component where little deformation occurs. On unloading the axial strain would recover and no compaction would occur, such as from light traffic on dry soils. Greater deformation occurs beyond a precompression stress, σ_{1c} , producing irrecoverable plastic deformation that follows the virgin compression line. The intersection between these two lines is approximately σ_{1c} .

The precompression stress is similar to yield stress. It depends on the stress history of the soil, so it increases when soils are damaged by compaction, but it can also partially recover over time through weathering or tillage. Many interpret precompression stress as the highest amount of axial stress that the specimen has experienced, so it is commonly used as an indicator of soil compaction history and damage. Fig. 6 illustrates the simple Casagrande approach to approximate σ_{1c} from the intersection between straight lines plotted as recompression and virgin compression curves, but other approaches have been proposed to derive σ_{1c} from the shape of the entire curve. At very high stresses, little pore space remains so axial deformation deviates from the virgin compression curve and eventually ceases. This is generally beyond the stresses that are typically applied to soils, so ignored in analysis. With real soil data, recompression and virgin compression curves may not be as obvious as Fig. 6 suggests.

Fig. 7 shows examples of measured and modeled axial strain (strain expressed in terms of void ratio, see Eq. 13) versus axial stress curves. These were carried out on undisturbed, unsaturated, structured silt-loam soil collected from three different depth of 12 cm 52 cm and 72 cm below the soil surface (Berli, 2001; Berli et al., 2015). Symbols represent measured and lines modeled stress-strain values using the approach by Casini (2012). The modeled compression curves are bi-linear and therefore similar to the idealized stress-strain curve shown in Fig. 6, but with a less distinct intersection between the recompression and virgin compression lines. The soils were initially equilibrated to different water suctions, which because of effective stress, shift behavior depending on the impacts to capillary cohesion and distribution water affecting χ (Eq. 7).

Soil failure - shear

Whereas a compressive stress (Fig. 1) in confined conditions can decrease pore space, depending on the directions of applied stresses and confining conditions from surrounding soil, a soil may deform and ultimately fail. Soil failure generally occurs from a combination of stresses applied to soils in different directions. This section introduces Mohr-Coulomb theory as an example of a simple failure criterion frequently used in soil mechanics. Mohr-Coulomb focuses on the internal friction within the soil mass as the main failure criterion by defining the maximum amount of shear stress, τ_{max} , a soil specimen can withstand before falling apart. For a triaxial stress situation with principal stresses σ_1 , σ_2 , σ_3 and $\sigma_2 = \sigma_3$ as defined in Fig. 2, the Mohr-Coulomb failure criterion can be written as follow:

$$\tau_{max} = c + \sigma_n \tan(\phi) \quad (20)$$

with

$$\sigma_n = \frac{\sigma_1 + \sigma_3}{2} \quad (21)$$

the normal stress on the failure plane, τ_{max} the maximum shear stress on the failure plane, c the cohesion and ϕ the angle of internal friction of the soil, respectively. Fig. 8 is a graphical representation of the Mohr-Coulomb failure line superimposed on a typical Mohr circle defined by the principal stresses σ_1 and σ_3 .

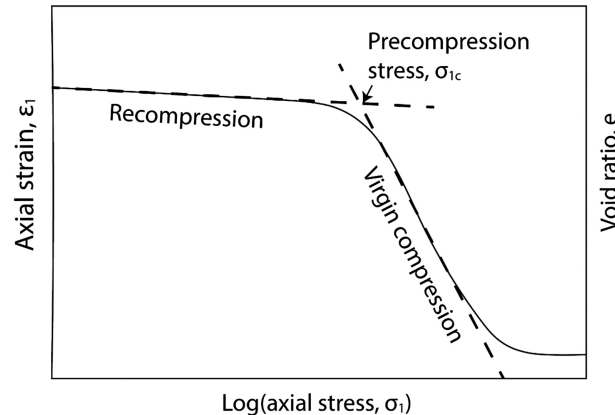


Fig. 6 Idealized stress-strain curve for an elastoplastic soil undergoing confined uniaxial compression. The transition from (elastic) recompression to (plastic) virgin compression is characterized by the precompression stress, σ_{1c} .

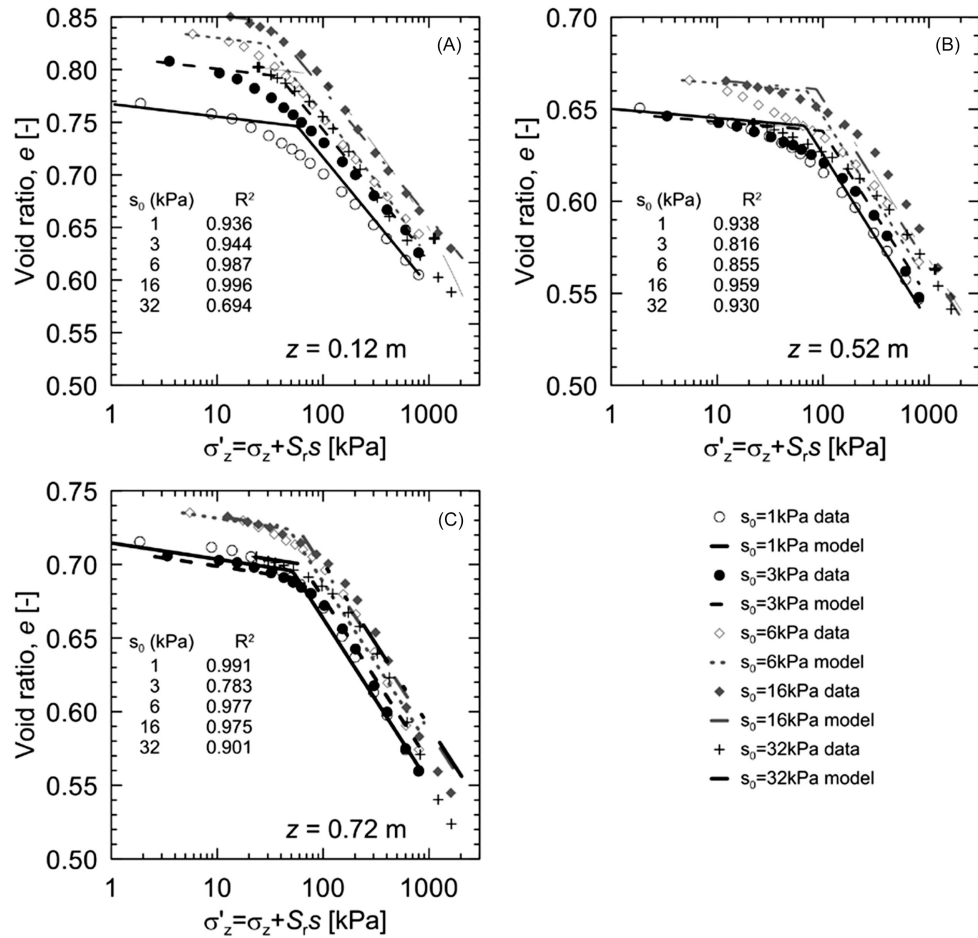


Fig. 7 Measured (symbols) and modeled (lines) of effective axial stress vs. void ratio curves employing the model by Casini (2012) for specimen of structured, unsaturated agricultural soil collected from depths of (A) 12 cm, (B) 52 cm, and (C) 72 cm. Each symbol represents the arithmetic mean of four to nine individual measurements (Berti et al., 2015 Fig. 6). s_0 indicating the initial suctions of the soil specimen.

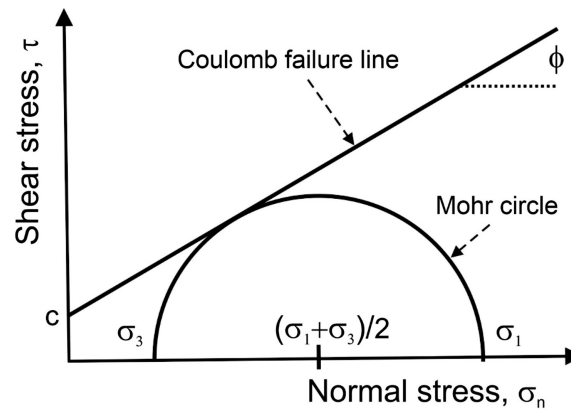


Fig. 8 Mohr-Coulomb failure criterion for soil showing the relationship between shear stress, τ and normal stress, σ_n , for a soil with cohesion, c , and angle of internal friction, ϕ .

Fig. 8 as well as Eqs. (20) and (21) imply that if the Mohr circle defined by the principal stresses σ_1 and σ_3 touches the Coulomb failure line, the soil specimen will fail. If the Mohr circle does not touch the Coulomb failure line, however, the principal stresses σ_1 and σ_3 do not cause the soil specimen to fail. The Mohr-Coulomb failure criterion tells us that there are three main parameters that define whether a soil fails or not. First there are the two principal stresses σ_1 and σ_3 . One can see from Eqn. (20) that with increasing normal (or compressive) stress, σ_n , also the maximum shear stress, τ_{max} , increases. i.e., the greater the normal stress, σ_n ,

the greater the shear stress the soil can withstand without failure. The most extreme case in this respect occurs if $\sigma_1 = \sigma_3$, i.e., all three principal stresses have the same magnitude and equal to the mean hydrostatic stress, p . For $\sigma_1 = \sigma_3 = p$, the Mohr circle shrinks to a single point and the Mohr-Coulomb failure criterion tells us that under these (hydrostatic) stress conditions no failure can occur.

The other extreme stress situation occurs if the axial stress, σ_1 , has a non-zero positive value but the radial stress component, σ_3 , becomes zero. This stress situation occurs in a uniaxial compression test. Fig. 9 shows the Mohr-Coulomb diagram with $\sigma_3 = 0$ for a soil with the same cohesion, c , and angle of internal friction, ϕ , as shown in Fig. 8. By comparing the stress values, σ_1 , from Figs. 8 and 9, for $\sigma_3 = 0$ the stress values σ_1 at failure are considerably smaller than for $\sigma_3 > 0$. The key parameter to assess failure is the difference between σ_1 and σ_3 , or what we defined in Eqn. (5) as the deviatoric stress, q . In the soil mechanics literature, other more advanced failure criteria are discussed, which are typically formulated in terms of the stresses of p and q (see e.g., Critical State Soil Mechanics; (Schofield and Wroth, 1968)). However, a comprehensive discussion of failure criterion is beyond the scope of this chapter.

Besides the stress state given by the principal stresses σ_1 and σ_3 , the two soil properties cohesion, c , and angle of internal friction, ϕ , play an important role in assessing soil failure. These depend on soil properties and from a practical standpoint they can be modified by soil management and amendments. Cohesion is a measure of how strongly the soil particles “stick” together, whereas the angle of internal friction is a measure of how hard it is to move soil particle relative to each other. Cohesion and angle of internal friction depend on the type of soil. Clays, for example, generally have high cohesion but a low angle of internal friction, whereas (dry) sand or gravel are examples of soils that exert high internal friction angles but have no cohesion. As illustrated in Figs. 8 and 9, as well as Eq. (20), greater values of c and ϕ result in greater maximum shear stress for a given normal stress. Note that c is independent of σ_n , meaning that a soil can have some cohesion without being stressed. To go back to the pottery example used earlier on, pottery clay sticks together (“has cohesion”) if it is not stressed.

Soil moisture, via effective stress, can also contribute to cohesion, even in sands that have little particle bonding when dry. Recall the sand castle effect where “at the right water content” the sand grains are held together by capillary forces, being responsible for what is sometime referred to as “apparent cohesion.” This apparent cohesion depends on the moisture content and disappears if the sand gets too wet or dries out. For soft surface soils, as found in many agricultural or forestry applications, cohesion plays an important role and primarily determines whether a soil fails. The angle of internal friction, however, plays less of a role, unless the applied stresses are considerably higher than the cohesion. Examples include under the wheel of a heavy vehicle or the hoof of an animal that is sinking into soft soil. This results in $\sigma_3 > 0$ but also $\sigma_1 \gg \sigma_3$ and, hence, according to Eq. (20) the angle of internal friction starts controlling the maximum shear stress.

Mohr-Coulomb characteristics can be measured by a triaxial test or a simpler to conduct direct shear test. In the direct shear test, a soil sample is placed in a split block or ring, with one half moved across the other to produce a defined shear failure zone. Weights are placed on the soil to control σ_n .

Soil failure - fracture

Depending on the direction of the applied stress and the plasticity of a soil, failure can be dominated by fracture. This occurs from the extension of pre-existing cracks in the soil. As applied stresses cannot pass across these cracks, the stresses are concentrated at the tips, greatly reducing the stress required to cause failure (Fig. 10). The weakest state of a soil is fracture from tensile stresses, which can occur from soil shrinkage upon drying. Fracture can also occur from shear and compressive stresses, as they can open up the tips of cracks, causing them to elongate under tension (Fig. 10).

Fracture by crack elongation in soils dominates in dry, brittle conditions. Often linear elastic behavior is assumed, with the failure stress under tension, σ_T calculated with fracture mechanics by:

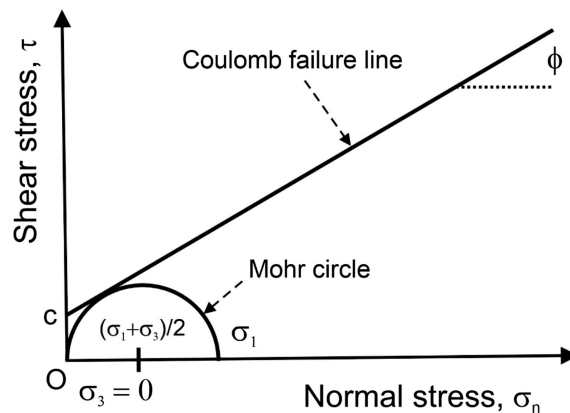


Fig. 9 Mohr-Coulomb failure criterion for $\sigma_3 = 0$ and a soil with same cohesion, c , and angle of internal friction, ϕ , as shown in Fig. 8.

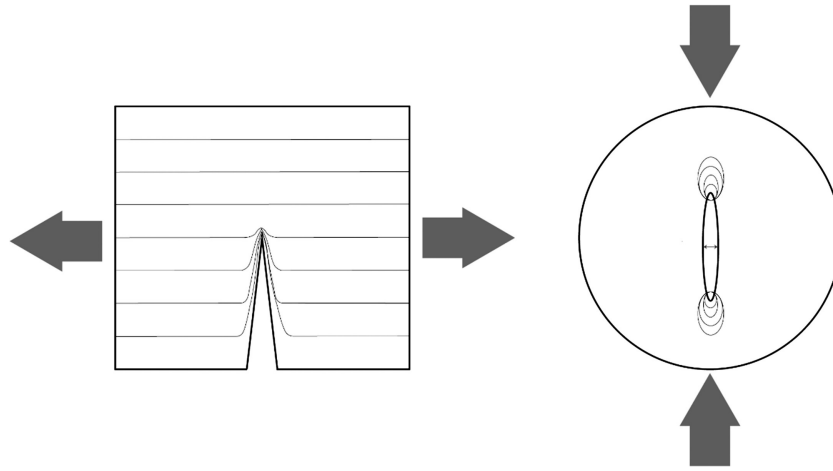


Fig. 10 Concentration of stress at the tip of cracks showing examples of direct tension of a half crack (left) and indirect tension caused by crack opening from a compressive stress on a disk (right). The gray arrows are the direction of the applied stress, with thinner lines showing stress transmission in the specimen.

$$\sigma_T = \frac{K_{IC}}{Y\sqrt{c}} \quad (22)$$

where K_{IC} is the fracture toughness, c is the length of the crack and Y is a geometry factor to account for the shapes of the sample and crack (Hallett and Newson, 2005). The parameter K_{IC} is independent of crack size, depending on the bond strength of soil particles at the crack tip.

Measuring K_{IC} of soil requires remolded specimens with controlled crack lengths and specific geometries to apply a tensile stress at the crack tip. In intact, structured soil it is impossible to measure. Here the transmission of stress and its concentration at crack tips will be affected by interacting, geometrically complex crack networks. The disk-shaped specimen illustrated on the right of Fig. 10 shows opening of a crack by a compressive stress, resulting in tensile strains developing at the crack tip. This is an idealized image of indirect tension stresses resulting at crack tips, which can occur regardless of the direction of the applied stress (Fig. 1).

As described previously, soils are not linear elastic materials, even when dry. Eq. (22) can be adapted to account for elasto-plastic behavior with nonlinear fracture mechanics, but this requires more complex analysis and testing.

Measuring soil mechanical behavior in the field

To use soil mechanics for practical applications, one needs to know the mechanical properties of a soil. Previous sections described uniaxial compression, triaxial compression and direct shear tests that are conducted in the laboratory. Cone penetrometers, shear vanes and sinkage plates are three examples of field testing equipment developed to measure the strength and compressibility of soils for tillage, traction, and soil-compaction studies. These devices provide soil-strength parameters that tend to depend on the geometry of the test device and the type of loading applied. Moreover, these soil parameters often do not represent any single soil property, but are usually functions of several fundamental engineering properties of soil. Therefore, soil parameters obtained using these devices (e.g., penetrometers, shear-vane devices, and shear graphs) are often called 'composite soil parameters.' There are also various techniques used to determine soil stickiness, shatter resistance, and cutting resistance. The cone penetrometer is well-known among these devices because it is a simple device that is very easy to use. Shear-vane devices and shear graphs have been used to obtain soil shear and sinkage parameters. They are useful in predicting the tractive ability of wheels and tracks using a semiempirical approach.

Cone penetrometer

Perhaps the most widely used device to measure soil strength in the field is the cone penetrometer. Although the cone penetrometer was developed to determine the mobility of off-road vehicles, it has been used to predict traction, draft requirements of tillage implements, and to quantify soil strength to indicate soil-compaction level and impedance to root growth. The most common form of this device consists of a polished steel cone, which is pushed against the soil and then the force of penetration is measured. The American Society of Agricultural Engineers (ASAE) standard (S313.3) describes the geometry of a standard cone penetrometer, whereas a second standard (EP542) outlines the proper procedure for using this device.

Since the force needed to penetrate the soil is related to the geometry of the device, a cone with a base diameter of 20.27 mm and an apex angle of 30° is selected as the standard (Fig. 11). For harder soil conditions, a smaller cone with a base diameter of 12.83 mm is used. A second key variable that influences the force of penetration is the penetration rate. ASAE Standard EP542

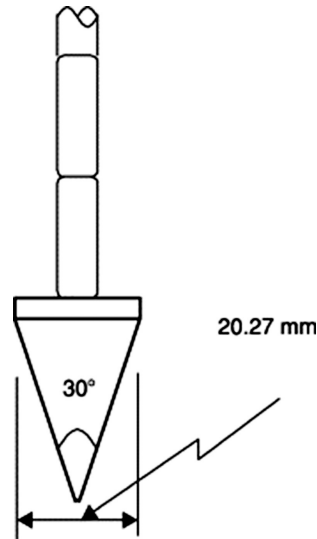


Fig. 11 Standard cone penetrometer of the American Society of Agricultural Engineers.

recommends a quasistatic rate of 1.83 m min^{-1} . Although it is difficult to control the insertion rate with handheld devices, hydraulically or electrically operated devices can be designed to operate at this standard speed. Force is usually measured using a load-sensing mechanism such as a load cell. Newer devices often include a depth-measuring sensor such as a potentiometer so that a soil penetration resistance profile can be obtained. The penetration resistance force is expressed as the cone index, which is the ratio of the force to the base area of the cone. The soil cone index value obtained using a soil cone penetrometer is a composite value that depends on soil texture, bulk density, and moisture content. In terms of engineering properties of soil, it depends on cohesion, soil internal angle of friction, soil metal friction, and adhesion.

One of the main concerns with the use of cone index to represent soil strength is its variability, especially in dry and cloddy conditions. ASAE standard EP542 recommends that in a given location at least 20 measurements be taken at a water content near to the field capacity of soil to obtain a representative measure of soil strength. With the advent of precision agriculture and the potential role of soil compaction in limiting water infiltration, drainage, and root growth, there is an increased interest in the cone penetrometer as a soil-strength mapping tool. Consequently, fully automated cone penetrometers with global positioning systems (GPS) to provide geographic position data are now commercially available.

Soil sinkage

Cone penetrometers are inadequate to measure soil characteristics relevant to tractive ability of wheels and tracks, so sinkage devices can be used. Soil sinkage devices consist of either circular or rectangular plates that are pushed against soil, and their load deformation characteristics are recorded (Fig. 12). The Bernstein equation is often used to relate applied pressure, p_s , to soil deformation, z as:

$$p_s = kzm, \quad (23)$$

where m is a constant and k is the sinkage parameter dependent on soil cohesion, angle of internal friction, and plate dimensions. Plate sinkage relationships have been used to model rolling resistances of wheels and tracks.

Soil shear

Shear characteristics of soil such as cohesion and angle of internal friction have been used to model the tractive ability of wheels and tracks, soil fragmentation by tillage tools, soil stability and the mechanical resistance to root growth. In soil science, a rotational shear vane is commonly used in the field (Fig. 13). The shear vanes come in different sizes, with larger vanes deployed in weaker soils to improve accuracy. A measurement of shear strength is calculated from the torque that needs to be applied to rotate the vane. Measurements can be made at different depths by pushing the vane deeper into the soil.

A range of other approaches can measure shear strength in the field, including a combined sinkage and shear device called a bevameter. Another approach is to insert a metal box into the top of the soil to a prescribed depth, excavate the surrounding soil to that depth, and then to winch the box horizontally, with the applied force recorded. Weights can also be placed above the soil for Mohr-Coulomb analysis (Eq. 20). This approach has been deployed to measure the reinforcement of soil by plant roots.

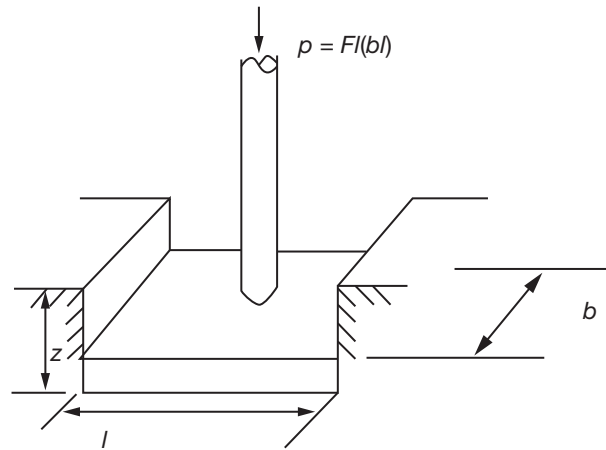


Fig. 12 A rectangular sinkage plate.

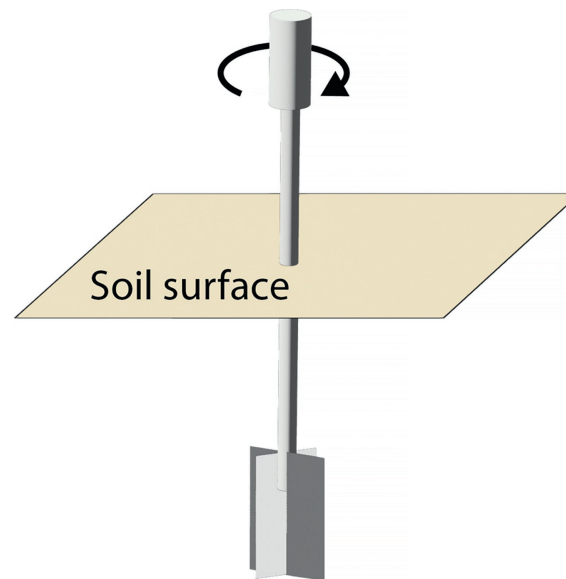


Fig. 13 A rotational shear vane. The torque required to rotate the vane in soil until it fails and the size of the shear vanes are used to calculate shear stress.

Soil mechanics and land management

Soils can be managed so that mechanical damage is avoided or mitigated, and to retain or improve mechanical properties to provide greater stability. Direct damage to soil arises from traffic deforming the soil by compaction, which is covered in detail in another article in the encyclopedia. Lighter machinery can decrease the stress and therefore strain experienced by the soil. Moreover, elastoplastic (and even viscous) deformation can be avoided by timing traffic on soils when they are drier and behave more as an elastic material that can withstand and recover from stress.

Tillage, predominantly by shear and fracture failure, breaks up compacted soils, resulting in drops in the precompression stress (Fig. 6), penetration resistance (Fig. 11) and shear strength (Fig. 13). Roots can therefore penetrate the soil more easily, but the weaker soil may also be more vulnerable to mechanical stresses from rainfall that leads to erosion. Poorer soil stability following tillage can increase viscous flow, so the tilth produced by tillage may slump over time.

Biology has a large impact on soil mechanics. Large-scale slope stability, for instance, can be improved by planting vegetation to provide reinforcement of soil by roots. Likewise, removing plant roots can make soils more vulnerable to damage by traffic, as occurs with overgrazing by livestock. At smaller-scale, biology binds soil particles, resulting in aggregation and the creation of a complex pore network. Microorganisms and plant roots excrete *exo*-polymeric substances that are surface active and act like glues. Fine plant roots and fungal hyphae provide stabilization at a slightly larger scale. These mechanical processes drive soil aggregation,

but they are often overlooked and rarely quantified in predominantly observational research on soil structure formation (Hallett et al., 2013).

Readers are referred to articles in this encyclopedia on Bioturbation, Soil Compaction, Soil Tilth, Root Reinforcement and Soil Aggregates where soil mechanics has been used to described processes.

Outlook

The mechanical behavior of soil is a fundamental physical property that is highly relevant to numerous soil functions, especially related to soil structure and stability. Two of the biggest threats facing soils, compaction and erosion, are driven by the resistance of soil to mechanical damage from traffic or rainfall, respectively. Mechanical deformation occurs when an imposed stress exceeds the strength of the soil. Beyond a critical stress, failure of the soil may occur either with negative consequences such as landslides or with positive impacts such as crack elongation to form soil structure. As surface soils are structurally complex and unsaturated, theories adopted from geotechnical engineering are insufficient to describe their mechanical behavior. However, advances in recent years and improved computer power have improved understanding substantially, including the use of X-Ray CT imaging to visualize localized deformation processes. The importance of soil pore structure is now widely appreciated, but moving forwards, the role of soil mechanics in the genesis and deformation of soil pores needs greater consideration.

Readers wishing to learn more about soil mechanics are guided to the following books:

Wesley (2010) introduces the fundamental of soil mechanics using at a range of geotechnical engineering applications as examples.

Lu and Mitchell (2019) provides up-to-date overview of soil mechanics with special focus on geotechnical engineering.

Lu and Likos (2004) as well as Fredlund et al. (2012) introduce unsaturated soil mechanics from a soil science as well as a geomechanics perspective.

De Boer (2000) provides an introduction to soil mechanics based on porous media theory as well as an excellent review of the history of soil mechanics.

References

- Andrade JE and Mital U (2019) Multiscale and multiphysics modeling of soils. In: Lu N and Mitchell JK (eds.) *Geotechnical Fundamentals for Addressing New World Challenges*. Cham, Switzerland: Springer.
- Berli M (2001) *Compaction of Agricultural Subsoils by Tracked Heavy Construction Machinery*. Institute of Terrestrial Ecology (ITOE). Swiss Federal Institute of Technology (ETH), Zurich.
- Berli M and Or D (2006) Deformation of pores in viscoplastic soil material. *International Journal of Geomechanics* 6: 108–118.
- Berli M, Accorsi ML, and Or D (2006) Size and shape evolution of pores in viscoplastic matrix under compression. *International Journal for Numerical and Analytical Methods in Geomechanics* 30: 1259–1281.
- Berli M, Casini F, Attinger W, Schulin R, Springman SM, and Kirby JM (2015) Compressibility of undisturbed silt loam soil – measurements and simulations from a process-based model. *Vadose Zone Journal* 14. <https://doi.org/10.2136/vzj2014.10.0153>.
- Bishop AW (1959) The principle of effective stress. *Teknisk Ukeblad* 106: 859–863.
- Casini F (2012) Deformation induced by wetting: A simple model. *Canadian Geotechnical Journal* 49: 954–960. <https://doi.org/10.1139/T2012-054>.
- Coulomb CA (1776) Essai sur une application des règles des maximis et minimis à quelques problèmes des statique relatifs à l'architecture. *Memoires de Mathematique et de Physique, présentés à l'Academie Royal de Science par divers Savans* 7: 343–382.
- de Boer R (2000) *Theory of Porous Media*. Berlin: Springer.
- Fredlund DG and Rahardjo H (1993) *Soil Mechanics for Unsaturated Soils*. New York: John Wiley & Sons.
- Fredlund DG, Rahardjo H, and Fredlund MD (2012) *Unsaturated Soil Mechanics in Engineering Practice*. New York: John Wiley & Sons.
- Ghezzehei TA and Or D (2001) Rheological properties of wet soils and clays under steady and oscillatory stresses. *Soil Science Society of America Journal* 65: 624–637.
- Hallett P and Newson T (2005) Describing soil crack formation using elastic-plastic fracture mechanics. *European Journal of Soil Science* 56: 31–38. <https://doi.org/10.1111/j.1365-2389.2004.00652.x>.
- Hallett PD, Karim KH, Glyn Bengough A, and Otten W (2013) Biophysics of the vadose zone: From reality to model systems and back again. *Vadose Zone Journal* 12: Available from: https://www.researchgate.net/publication/277390063_Biophysics_of_the_Vadose_Zone_From_Reality_to_Model_Systems_and_Back_Again. Accessed 29 June 2023.
- Lu N and Likos WJ (2004) *Unsaturated Soil Mechanics*. Hoboken, New Jersey: John Wiley & Sons.
- Lu N and Mitchell JK (2019) *Geotechnical Fundamentals for Addressing New World Challenges*. Cham, Switzerland: Springer.
- Navarro V, Asensio L, Alonso J, Yustres A, and Pintado X (2014) Multiphysics in implementation of advanced soil mechanics models. *Computers and Geotechnics* 60: 20–28. <https://doi.org/10.1016/j.compgeo.2014.03.012>.
- Rankine WJM (1857) On the stability of loose earth. *Philosophical Transactions of the Royal Society of London* 147: 9–27.
- Schofield AN and Wroth CP (1968) *Critical State Soil Mechanics*. London: McGraw-Hill.
- Terzaghi K (1925) *Erdbaumechanik auf bodenphysikalischer Grundlage*. Wien: Franz Deuticke.
- Terzaghi K and Peck RB (1948) *Soil Mechanics in Engineering Practice*. New York: John Wiley and Sons.
- Vyalov SS (1986) *Rheological Fundamentals of Soil Mechanics*. Amsterdam: Elsevier.
- Wesley LD (2010) *Fundamentals of Soil Mechanics for Sedimental and Residual Soils*. Hoboken, NJ: John Wiley & Sons.

Compaction

JJH van den Akker^a, L ten Damme^{b,c}, M Lamandé^{b,d}, and Thomas Keller^{c,e}, ^aWageningen University & Research, Wageningen Environmental Research, Wageningen, The Netherlands; ^bDepartment of Agroecology, Aarhus University, Research Centre Foulum, Tjele, Denmark; ^cDepartment of Soil & Environment, Swedish University of Agricultural Sciences, Uppsala, Sweden; ^dFaculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway; ^eAgroscope, Department of Agroecology & Environment, Zürich, Switzerland

© 2023 Elsevier Ltd. All rights reserved.

This is an update of J.J.H. van den Akker, B. Soane, *COMPACTION*, Editor(s): Daniel Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 285–293, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00248-4>.

Introduction	86
Factors and processes affecting the distribution and intensity of compaction	86
Compaction under running gear	87
Effect of loading characteristics on contact and soil stress	88
Soil compactability	89
Effects on soil physical and mechanical properties	91
Bulk density, porosity, and packing state	91
Hydraulic properties	91
Aeration characteristics	91
Soil strength characteristics	91
Compaction in crop production	92
Effects on germination and establishment	92
Effects on root growth and root distribution	92
Effects on plant growth, crop yield and soil biota	93
Interactive crop responses to compaction and fertilizer application	93
Crop responses to subsoil compaction	93
Modeling of crop responses to soil compaction	94
Effects on environmental components	95
Measurements of soil mechanical stresses	97
Strategies for the prevention of compaction	97
Compaction in forestry	98
Amelioration of compacted soils	98
Outlook	98
References	98
Further reading	99

Abstract

Soil compaction is a decrease in pore space driven by mechanical stresses. It is prevalent in agriculture, forestry and construction due to too high wheel loads by machinery, leading to adverse impacts to drainage, aeration and soil strength that affects plants and exchanges between soils and the wider environment. Compaction arises when a mechanical stress exerted on the soil exceeds its internal strength. The applied stress attenuates with depth, but mitigating subsoil compaction is more difficult than compaction of the upper soil layers.

Key points

- Soil compaction by the weight of machinery or animals causes a loss of porosity and pore connectivity.
- When the compaction stress exceeds a critical value, plastic deformation results in no recovery of soil pores when the stress is removed.
- Lighter machinery, low ground pressure tires and driving on fixed tramlines are approaches to decrease soil compaction risk.
- Compaction damage can be reversed in the topsoil by tillage or natural processes, but recovery may take a long time and subsoil damage by compaction is irreversible.

Introduction

Soil compaction occurs when an exerted mechanical stress on soil exceeds its internal strength. Pore space is lost with compaction, which can adversely impact soil functions including habitat for soil biota and roots, flood regulation, and climate regulation. The resulting environmental consequences extend beyond on-site impacts to the wider landscape. Compacted soils tend to have poorer capacity to produce plants, whether for agriculture, forestry or amenity purposes.

Compaction is a process of densification and it is typically associated with distortion, in which total and air-filled porosity and permeability are reduced, and soil strength is generally increased. Compaction typically also changes soil pore connectivity and continuity, hence modifying soil structure and associated properties. The term 'compaction' is used to identify a process and should be distinguished from the term 'compactness,' which indicates the state of packing of the solid soil constituents at a given time and position.

There are many causes of soil compaction, but here we focus on stresses exerted by machinery or grazed livestock. With machinery, the compaction process can be initiated by wheels, tracks or rollers of traffic such as tractors, depth-control wheels on cultivators and packer rollers. Soil densification can also be caused by heavy overburdens of ice and snow, by natural and management-induced hard-setting, and by illuviation of clay from the A-horizon into the B-horizon. These processes may take place slowly over long periods of time and result in the B-horizon having much greater bulk density and strength than the A-horizon. Densification also occurs during structural collapse near the soil surface due to the impact of rain, resulting in the formation of surface crusts.

In arable land with annual plowing, both topsoil and subsoil compaction should be considered. We define the subsoil as the soil beneath the regularly loosened layer (that is about 20–35 cm thick). This definition of the subsoil includes the pan layer as the upper part of the subsoil. This pan layer is, in many cases, less permeable for roots, water, and oxygen than the soil beneath it, and is the bottleneck for subsoil functioning. In contrast to the topsoil, the subsoil is not loosened annually but compaction is cumulative and, in the long run, a more or less field-wide compacted layer is created. The natural recovery potential of the subsoil is low, and therefore, subsoil compaction is nearly persistent. Evidence from field studies demonstrate that impairment of soil physical properties caused by subsoil compaction remains for decades (Berisso et al., 2012; Keller et al., 2017). An important cause of subsoil compaction is moldboard plowing in which the furrow-side tractor wheels apply appreciable loads directly to the upper surface of the subsoil, causing a plow pan.

Problems of compaction are widely distributed throughout the world, but tend to be most prevalent and increasing where heavy machinery is used in agriculture or forestry, in both temperate and tropical areas (Keller et al., 2019). Soils that are naturally fragile in structure, such as soils of humid tropical forests and light-textured soils in areas of low but erosive rainfall, are particularly prone to problems arising from compaction and subsequent high risks of erosion due to a reduction of permeability. Compaction is one of the major threats to soil resources. Estimates of land area affected by soil compaction are still uncertain, but studies in the 1990's state that around 35% of all arable land is affected by compaction in regions with heavy mechanization (Oldeman et al., 1991). Moreover, soil compaction was reported to be equal to or exceed 33 Mha of agricultural land in Europe and 18 Mha in Africa. A project on mapping of Soil and Terrain Vulnerability in Central and Eastern Europe (SOVEUR) by the United Nations Food and Agriculture Organization (FAO) and International Soil and Reference Information Centre (ISRIC) showed that compaction is the most widespread form of soil physical degradation on agricultural land in Central and Eastern Europe (Nachtergaele et al., 2002; Batjes et al., 2014). It has been estimated that about 25 Mha agricultural land was lightly and about 36 Mha moderately compacted in the year 2000. Soil compaction results in significant costs, both directly for the farmer in terms of, among others, yield loss and higher fuel use, and indirectly for society due to erosion, higher greenhouse gas emissions, flooding and landslides. Reliable estimates of the ecological and economic costs of soil compaction remain, however, difficult to assess.

Factors and processes affecting the distribution and intensity of compaction

When the stress applied to the surface soil exceeds a critical value, the soil compresses and does not recover when the stress is removed. Generally, the expression of soil strength used for soil compaction risk evaluation is the precompression stress, also called preconsolidation stress. It is defined from the sharp bend observed on the strain - log(stress) curve obtained from the uniaxial, confined compression of soil cores (Fig. 1). At stresses smaller than the precompression stress, the rate and amount of deformation is small and if the applied load is removed, the soil returns to its original structure as an elastic material. Once the precompression is exceeded, deformation is greater with increasing applied stress and the soil does not bounce back when the stress is removed. This expression of soil strength is somewhat idealized and originates in civil engineering soil mechanics (Casagrande, 1936) where it was introduced to characterize the reversible (elastic) and persistent (plastic) strain of homogenized clay during slow consolidation. Because the precompression stress represents the strength of a soil, we call it the precompression strength. In agricultural soil science with structured soils, exceeding of the precompression strength will result in plastic deformation with destruction of the structure including irreversible changes of properties and functions (Horn, 2003).

A Virgin Compression Curve based on a one-dimensional compression test on soil is shown in Fig. 1. It is usually measured using very slow tests that allow for soil drainage, following the approach developed for geotechnical engineering. The downward compression stress is referred to as the normal stress. Soils may also experience shear and tensile stresses that have elastic and plastic components and can occur concurrently with compression to create complex stress fields.

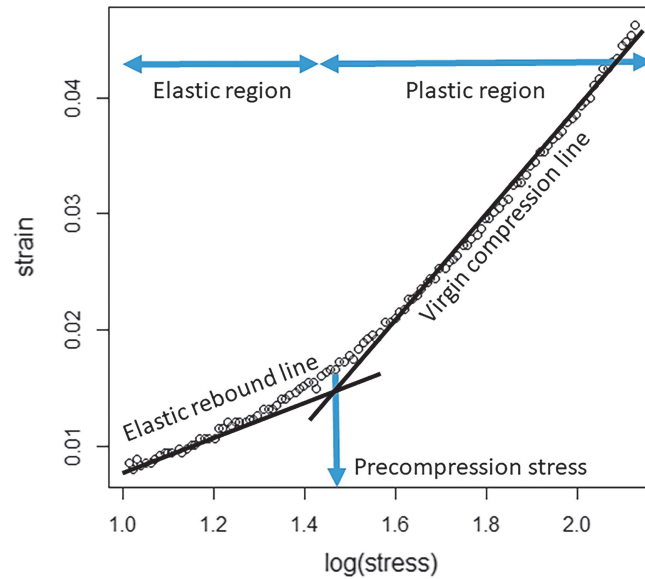


Fig. 1 Example of strain - $\log(\text{stress})$ experimental points obtained from a uniaxial, confined compression test. The precompression stress is defined as the stress at the sharp bend of the curve. Soil submitted to stress larger than the precompression stress or precompression strength will be compacted irreversibly following the virgin compression line.

Compaction under running gear

Compaction under wheels and tracks is a dynamic process in which a soil volume beneath the running gear is subject to normal and shear stresses that cause normal and shear strains. Fig. 2 illustrates how soil volume elements are affected by the passage of a wheel. During a wheel pass, a volume element at a certain depth endures not only volume change (i.e., compaction) but also distortion (i.e., shear deformation), as depicted by the volume elements 9–26. The soil displacement and deformation is also dynamic across the wheel track, due to differences in stress-state in the ground contact area, as can be seen in Fig. 3. The impact of a wheeling event on physical properties of the soil depends on the ability of the soil to resist the exerted stresses, i.e., on the strength of the soil. If the exerted soil stresses exceed the precompression strength or shear strength of the soil, then compaction or shear deformation will be accompanied by permanent structural deformations. This is what the wheel-load carrying capacity concept (Van den Akker, 2004; Van den Akker and Schjønning, 2004) is based on. Degradation of soil structure will be less severe if only the precompression strength is exceeded. In this case, the soil will mainly compress with limited distortion, and the soil pores will still retain a certain continuity.

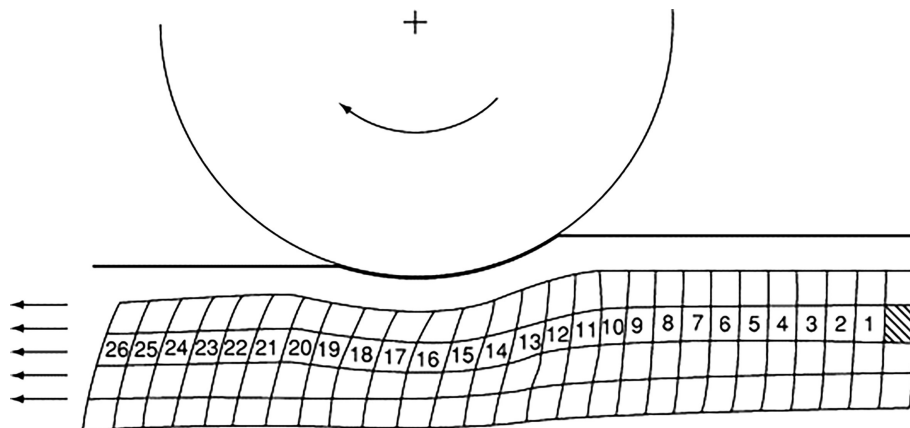


Fig. 2 Displacement and deformation in a vertical section of a soil volume element under a wheel. Reproduced with permission from Koolen AJ and Kuipers H (1983) Agricultural soil mechanics. *Advanced Series in Agricultural Sciences* 13, Heidelberg: Springer-Verlag.

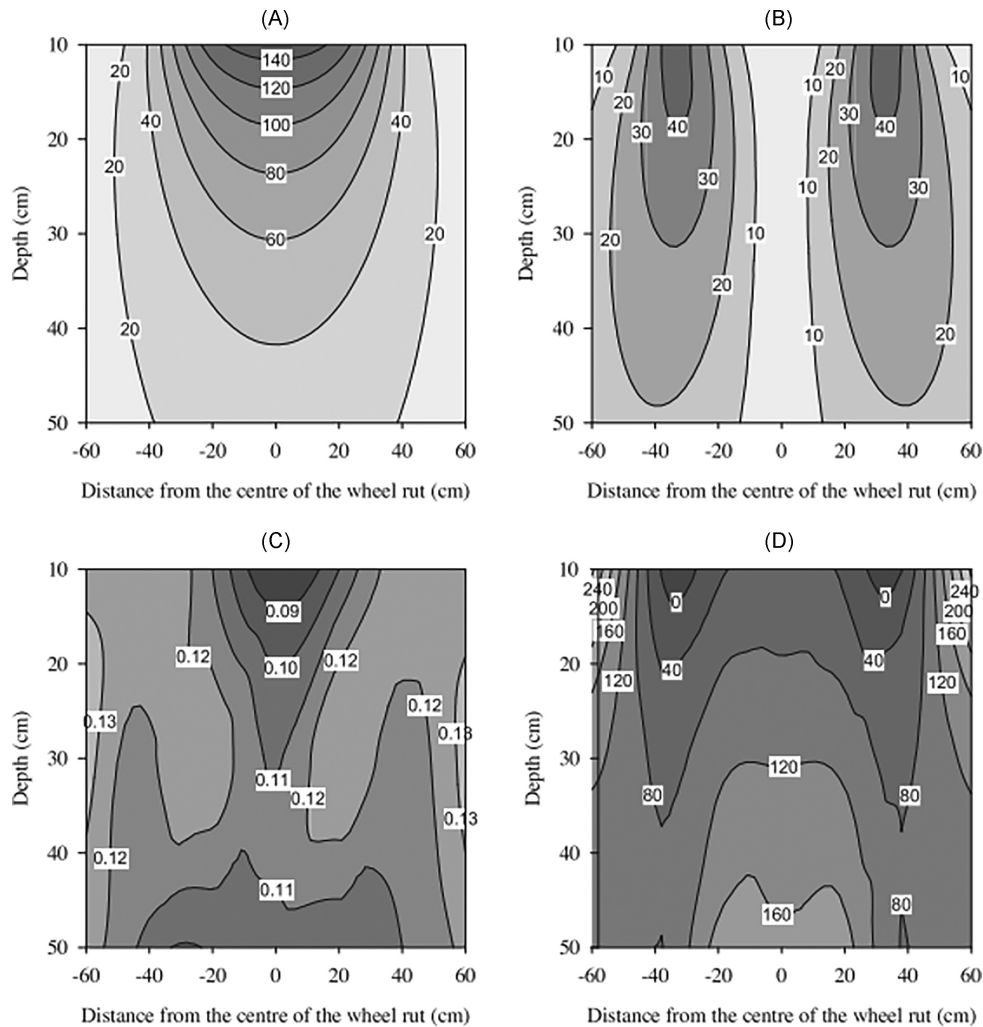


Fig. 3 Computed (A) mean normal stress (kPa) and (B) shear stress (kPa), and (C) interpolated measured air-filled porosity ($\text{m}^3 \text{m}^{-3}$) and (d) measured air permeability (μm^2) at a matric potential of -100 hPa , as a function of horizontal distance from the center of the wheel rut and depth in a clay loam soil; 4x repeated wheel load 61 kN; tire width 80 cm; inflation pressure 200 kPa. Figure from Berisso FE, Schjønning P, Lamandé M, Weisskopf, P, Stettler, M, and Keller T(2013) Effects of the stress field induced by a running tire on the soil pore system. *Soil & Tillage Research* 131: 36–46.

Effect of loading characteristics on contact and soil stress

Loading characteristics such as choice of tire or track, wheel load and inflation pressure are used to estimate the magnitude of stresses exerted on soil. Examples of loading situations for various vehicles and animals are shown in Table 1.

For practical purposes, the mean ground contact pressure P (kPa) is a useful parameter, defined as the wheel load divided by the ground contact area. The distribution of the load in the contact area is, however, hardly uniform under running gear on a deformable surface. Instead, the distribution is generally well described by a parabola, with a maximum stress at the centerline and zero stress at the edge of the ground contact area. The maximum stress in the ground contact area, that are typically much greater than P , determines the peak stress in the soil.

Peak stresses under caterpillar tracks are two to four times the mean ground contact pressure, and depend strongly on the firmness of the soil and design of the track system. Peak stresses under lugs can be two to five times higher than the mean ground contact pressure. The resulting soil stresses decrease, however, rapidly with depth, because the ground contact area of the lugs is small. If soil stresses at a depth of 0.2–0.3 m are considered, the peak stresses under the lugs can be neglected. Still, a parabolic stress distribution, with a maximum stress of one-and-a-half to two times the mean ground contact pressure generally provides a good estimate of the vertical stress distribution.

The stresses distributed in the ground contact area propagate into the soil profile. The larger the wheel load, the deeper into the soil the stresses propagate. Moreover, stresses propagate deeper when the soil is weaker. Track or tire dimensions and tire inflation pressure are the main drivers of stress in the upper depths of the soil (Arvidsson and Keller, 2004; Van den Akker, 2004), yielding smaller stresses for larger dimensions and lower pressures. Therefore, Low Ground Pressure tires can considerably reduce the risk of

Table 1 Examples of loads and ground contact pressures applied by agricultural and forest vehicles, and by animals.

Source	Total load (kN)	Load per (rear) wheel/track/hoof (kN)	Ground contact area (m ²)	Mean ground contact pressure (kPa)	Peak stress in ground contact area (kPa)
Tires					
270-kW tractor	100	40	0.65	65	170
65-kW tractor	40–50	10–15	0.2–0.3	40–75	90–200
Combine harvester (loaded)	250	85	0.75	110	220
Slurry tanker (loaded)	210	35–52	0.5–0.67	75–110	140–270
Forwarder	70–110	17–30	0.4–0.8	85–220	100–440
Tracks					
Tractor with tracks	140	70	2.6	30	80
Forwarder with tracks	100–140	35–70	2.0–2.4	55–75	100–150
Animal					
Horse	8	2–8		75–300	
Cow	4–6	1–6		120–480	

Note that ground contact area depends also on tire inflation pressure, tire flexibility and rut depth.

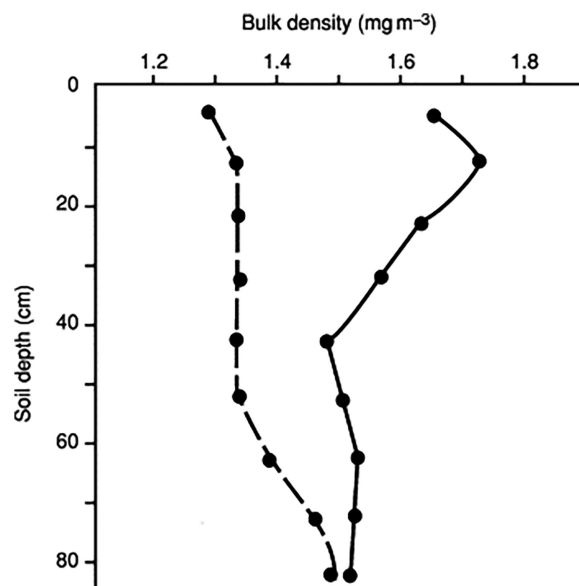


Fig. 4 Increase of bulk density to considerable depth when a bulldozer was used for forest clearing in Surinam (—no vehicle traffic; — bulldozer traffic). Reproduced with permission from Van der Weert R (1974) *Tropical Agriculture (Trinidad)* 51(2): 325–331.

soil compaction compared to standard tires (Ten Damme et al., 2020). The benefit gets smaller the deeper into the soil profile the stress propagates. Deeper in the soil profile, stress relates more closely to wheel load instead, which can be explained by the Söhne (1953) summation procedure: a pseudo-analytical continuum concept based on the Boussinesq (1885) solution for stress propagation in an isotropic half space medium. Compaction by heavy machinery extends then well into the subsoil (Fig. 4), despite the fact that heavier machinery generally has larger tires compared to lighter machinery.

However, many machinery-soil interactions that may influence the stress-state and distortion rate in soil remain poorly understood. Among these interactions are the effects of traction, repeated wheeling and loading time. More traction increases the risk of pore distortion due to changes in the stress state in soil (higher horizontal stress relative to vertical stress) and has been pointed out as a factor that explains greater subsoil deformation for smaller wheel loads (Pulido-Moncada et al., 2019; Ten Damme et al., 2021). With increasing loading time and number of wheeling events, soil deformation may gradually increase without substantial changes of the stress state.

Soil compactibility

Soils can vary in their ability to resist deformation, from being sufficiently strong to resist applied loads (low compactibility) to being so weak that they are compacted by even small loads (high compactibility). Well-structured soils combine good physical soil

properties with adequate strength compared to badly structured soil. Sandy soils with a single-grain structure and compacted massive soils can be very strong, but rootability and soil physical properties are then often poor. Roots have a reinforcing action and increase the elasticity and resistance of a soil to compaction.

Soil moisture content and soil water suction have dominant influences on soil compactibility. In soils where capillarity by water is the main cohesive force, strength increases with drying until it reaches a maximum, and then again decreases upon further dehydration because the cross-sectional area of the water menisci decreases so the impact of capillary suction is less. This cohesive force based on soil water suction is called apparent cohesion (Koolen and Kuipers, 1983). Drying also increases the true cohesion because cementing agents concentrate in the connecting points between the soil grains. Soil water suction increases the cohesion significantly and in this way the strength and resistance to compaction of a soil can vary considerably. In dry structured (aggregated) soils, soil water concentrates in the aggregates, and cohesion and soil strength in the aggregate increases considerably. Drying structured soils shrink and turn into an assembly of individual aggregates that fit together rather neatly. This assembly of aggregates has a very high interaggregate angle of internal friction and moderates interaggregate cohesion. This soil is strong and has a small compactibility. However, if such a soil is overloaded and compacted, the aggregates will be crushed and the interaggregate space will be lost, resulting in a dense compact soil.

Dry soils can easily resist loads. However, extremely dry sandy soils lose their apparent cohesion and can be deformed and compacted under small stresses. With increasing moisture content, soil compactibility increases until a condition known as the optimum moisture content for compaction is reached (Fig. 5A). At still wetter moisture contents, the soil becomes increasingly incompressible as the moisture fills ever more of the total porosity and further loss of air-filled porosity becomes impossible. Saturated soil is incompressible because water is incompressible. However, in reality, soils are hardly ever fully saturated. Although compaction may be minimal in field-saturated soils, the plastic flow of a wet soil results in a degradation of soil structure and macropores with accompanying diminishment of soil physical quality (Dawidowski and Lerink, 1990). Hence, dry and strong soils may be largely elastic, whereas at high moisture content and low strength, the soil's reaction to loading may be plastic flow.

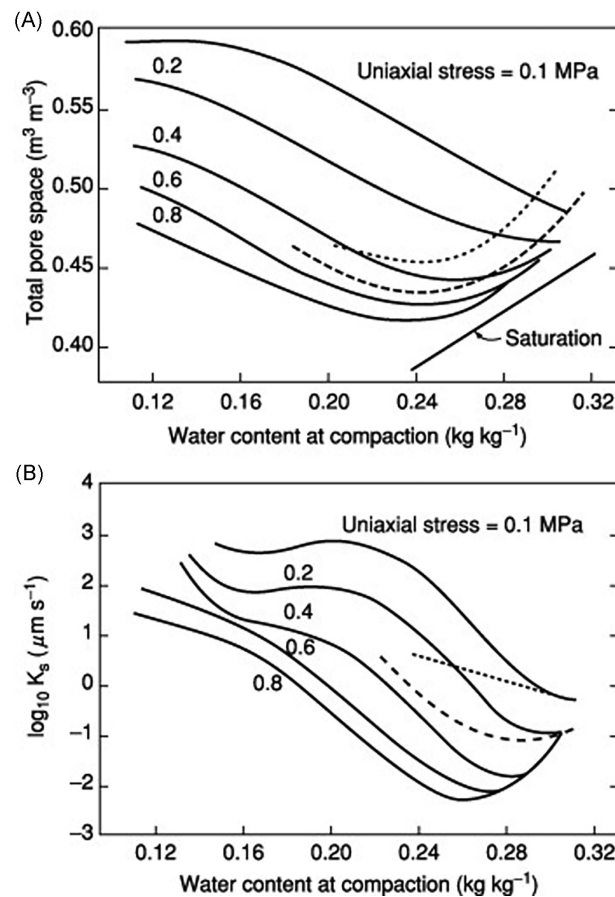


Fig. 5 Changes in (A) total porosity and (B) saturated hydraulic conductivity of a sandy clay loam subjected to field traffic during seedbed preparation with low ground pressure (marked as '.....', tire inflation pressure 40 kPa) or high ground pressure (marked as '- - -', tire inflation pressure 80 kPa) with the same wheel loads or subjected to uniaxial stresses of 0.1, 0.2, 0.4, 0.6, and 0.8 MPa at various water contents. Solid lines refer to compression of aggregated mixtures (after Dawidowski and Lerink, 1990). Reproduced with permission from Horton R, Ankeny MD, and Allmaras RR (1994) *Soil Compaction in Crop Production*. Amsterdam: Elsevier.

Increases in organic matter content tend to increase soil compactibility but also increase soil elasticity, the latter resulting in greater compaction resilience of pore structure. Peat soils dominated by organic matter, for instance, are quite resilient to compaction, despite initial large deformations. The value of the optimum moisture content for compaction increases as the organic matter content increases.

Effects on soil physical and mechanical properties

Bulk density, porosity, and packing state

Dry bulk density (the mass of soil solids per unit volume) is a direct and easy-to-measure indicator of changes in compactness. As it is calculated from the mass of material in a given volume, bulk density is affected by both porosity and particle density. At the same bulk density, a soil with a greater soil organic matter content will have a smaller porosity than a soil with less soil organic matter content. This impact can be removed by expressing changes in the packing state by total porosity or void ratio, described in the Stress-Strain and Soil Strength article. Total porosity is much easier to visualize than bulk density, as it is the total volume of pores in a given volume of soil. Void ratio is the volume of pores in relation to the volume of solids only.

Between soils of different textures, bulk density and porosity can change markedly, so it is difficult to interpret compaction occurrence from these measurements alone. Relative terms for packing state (e.g., relative density, degree of compactness) provide a ratio between a 'compacted' and 'uncompacted' state of the same soil. These enable the same threshold values to indicate over-compaction for soils with different texture (Håkansson et al., 1988).

Compaction primarily causes major reductions in macroporosity ($>50\ \mu\text{m}$), reduction of total porosity, and often an increase in microporosity. Major changes to hydraulic conductivity and gas diffusivity result, due to compaction impacts to pore architecture, i.e., pore size distribution, pore connectivity and continuity.

Hydraulic properties

Saturated hydraulic conductivity is very sensitive to the compaction process by wheels, especially if the deformation includes shearing and kneading of the soil. Fig. 5 shows that wet soils are easier to compact than dry soils, and that the effect on the saturated hydraulic conductivity is more than proportional to the compaction. In Fig. 5A the vertical stress in the topsoil beneath the Low Ground Pressure tire can be estimated at about 130 kPa (0.13 MPa) and beneath the High Ground Pressure tire on about 220 kPa (0.22 MPa). However, the compaction levels expressed in total pore space are comparable to uniaxial stresses in the range of 0.5–0.3 MPa. This can be explained by the dynamic stress field including shear stresses under wheels and resulting distortion of structure by shear failure resulting in the greater degradation of soil qualities than observed in uniaxial tests.

The decrease of macroporosity and increase of microporosity in the process of compaction results in larger water contents for a wide range of matric potentials in compacted versus uncompacted soil. The consequence is that compacted soils are generally characterized by a higher unsaturated hydraulic conductivity compared to their uncompacted state. However, near saturation, meso- and macropores are also water filled and contribute to the hydraulic conductivity. Thus, the saturated and near-saturated hydraulic conductivity of an uncompacted soil is greater than that of a compacted soil.

Aeration characteristics

A consequence of compaction is a reduction in the air-filled pore space. The air-filled pore space is an indication of the aeration status of the soil for plant growth and microbial activity. At air-filled porosity values less than 10%, oxygen deficiency is likely, especially in warm weather, whereas values less than 5% probably indicate incipient anaerobiosis. Analogous to saturated hydraulic conductivity, aeration status depends strongly on soil structure and continuous macropores. In a poorly structured compact soil, threshold values for air-filled porosity of 20% (oxygen deficiency) and 10% (anaerobiosis) are often used to indicate structural condition. In a well-structured soil, oxygen deficiency may start below 5% air-filled porosity and anaerobiosis below 2.5% air-filled porosity. Other indices of aeration status, such as oxygen diffusion rate, oxygen content of soil air, and air permeability, provide more precise indicators of the aeration level. Compaction that involves shearing disrupts pore continuity, which can decrease aeration with minimal changes to total porosity. Therefore, the values listed above only provide a rough estimate.

Soil strength characteristics

Soil strength resists the stresses causing compaction, and tends to increase with compaction. However, at the same bulk-density and moisture content, a well-structured soil resists stresses better than a poorly structured soil under moist conditions, Horn (2003). A wet soil is generally weaker than a dry soil. In periods with wet-weather conditions, compacted soils tend to become wetter than uncompacted soils due to limited infiltration capacity, smaller saturated hydraulic conductivity, and large microporosity. Consequently, the strength and trafficability of compacted soils become lower than that of well-structured soils that drain more readily. After a wet period, a compacted soil will stay wet longer with limited workability. In a dry period, the strength of a compacted soil will increase considerably and result in high trafficability. Soil tillage then requires more powerful equipment and a higher energy input to loosen the soil and break up clods of soil into fragments that form a good seedbed.

Compaction increases the penetration resistance for roots, resulting in limited rooting depth and reduced crop growth. At penetration resistances (measured with a cone with a diameter of 12.7 mm and top angle of 30 degrees) of 1.5 MPa and 3.0 MPa, root growth rates of many crops are reduced to approximately 50% and 0%, respectively. However, in well-structured soils, roots can make use of continuous macropores to penetrate deeply into the soil. In drying soils, strength and penetration resistance increase.

Compaction in crop production

Although compacted soils usually have unsatisfactory physical conditions for plant growth, the extent of loss of productivity depends on soil type, plant species, and weather conditions.

Effects on germination and establishment

Compact soils result in cloddy seedbeds after primary and secondary tillage, poor soil-seed contact, and reduced germination. Soils may develop a compact surface layer due to crusting after heavy rainfall, which has sufficient strength to restrict or even to inhibit seedling emergence, in particular for dicotyledonous species.

Effects on root growth and root distribution

Macropores ($>50\ \mu\text{m}$ diameter), through which roots can generally proliferate readily, are much reduced in compacted soil. As a result, root growth is restricted in development and depth or even inhibited. In Fig. 6, the limitations to root growth are

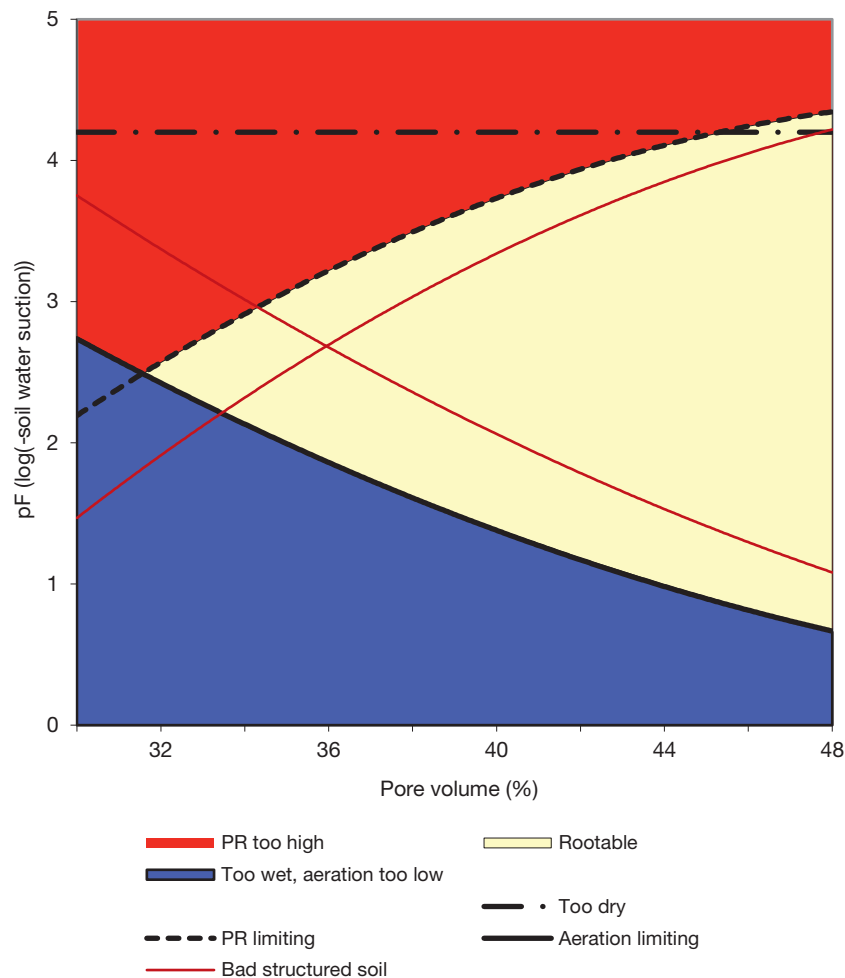


Fig. 6 A conceptual relationship between soil porosity and soil water potential in which soil aeration and mechanical resistance (PR, penetration resistance) meet specified root requirements. Root growth is insufficient in the red and blue shaded areas and impossible beyond a soil water potential of $-1500\ \text{kPa}$ (pF 4.2). Added to the figure are the same relationships if the structure is deteriorated resulting in a strong reduction of macropores (bad structured soil) (after Boone, 1988). Reproduced with permission from Lindstrom MJ and Voorhees MB (1994) *Soil Compaction in Crop Production*. Amsterdam: Elsevier.

conceptualized as relations between soil porosity and soil water potential at which soil aeration and mechanical resistance meet specified root requirements. In a soil with a certain pore volume, the soil water suction determines whether root growth is limited by too high mechanical resistance, by aeration problems, or is not limited. In Fig. 6, the two thin lines show the situation in the case of a poorly structured soil. A poorly structured soil with few continuous macropores needs more air-filled pores for aeration than a well-structured soil. Likewise, much smaller mechanical resistances are allowed in a poorly structured soil than in a well-structured soil because roots can follow the macropores in the well-structured soil.

Effects on plant growth, crop yield and soil biota

The level of compactness (often expressed in dry bulk density) influences the growth, yield, and quality of crops, although the impact depends on crop species, soil type, and weather conditions (Fig. 7). The optimum level of compactness for crop growth tends to be higher for sandy soils, in dry seasons, and for monocotyledonous species. At a compactness level less than optimum, crops suffer from reduced soil-root contact, reducing germination and nutrient transfer, while at a compactness level higher than optimum, root growth and aeration are restricted and denitrification can lead to N losses.

Soil biota and biological processes are influenced by soil compactness. This is partly because of the influence of pore size distribution and connectivity on the volume and connectivity of habitats for soil fauna, bacteria and fungi, and partly because compacted soils modify soil moisture and aeration conditions and may suffer from anaerobiosis, which in turn affects microbial metabolism markedly. The burrowing abilities of soil fauna, particularly earthworms, are reduced in compacted soils.

Interactive crop responses to compaction and fertilizer application

When farmers note that the growth of crops is adversely affected by compacted soils, a common action is to apply additional N fertilizer. However, the growth responses obtained may be less significant than those at lower levels of compactness (Fig. 8), and the additional nitrogen applications may be inefficient economically and detrimental to the environment. This is extremely important for soil sustainability. For instance, in Fig. 8, a compacted soil at 1.54 g cm^{-3} bulk density requires $3\times$ the applied nitrogen to obtain the same dry matter yield as an uncompacted soil at 1.37 g cm^{-3} bulk density. In compacted soils, roots are restricted from deep rooting, so if leaching occurs nitrogen is lost. Moreover, the poorer aeration status of compacted soils results in greater gaseous emissions of applied N, causing N-losses and producing N_2O .

Crop responses to subsoil compaction

Amelioration of subsoil compaction is much more expensive and less effective than loosening compacted topsoils. Avoidance of compaction in the subsoil is therefore a much greater need than in the topsoil. Seventeen years after a single full-field compaction action, crop and nitrogen yield losses can still be significant on compacted sites (Fig. 9). In this long-term experiment (Alakukku, 2000), only moderate axle loads were allowed on the fields after the initial compaction, to give the topsoil and subsoil the possibility to recover by natural processes. It should be noted that in practice, subsoils will endure heavy wheel loads from time to time, so most subsoils in agricultural fields are more or less partly compacted every year. This means that in practice, the recovery time of subsoils is limited, and the subsoil quality is never regained as much as in the long-term experiments. Note that the compaction effect was more pronounced in all years for harvested nitrogen than for grain dry matter. The strong decline of the effect in the first few years was mainly due to the recovery of the topsoil. The effect of compaction on grain and nitrogen yield was

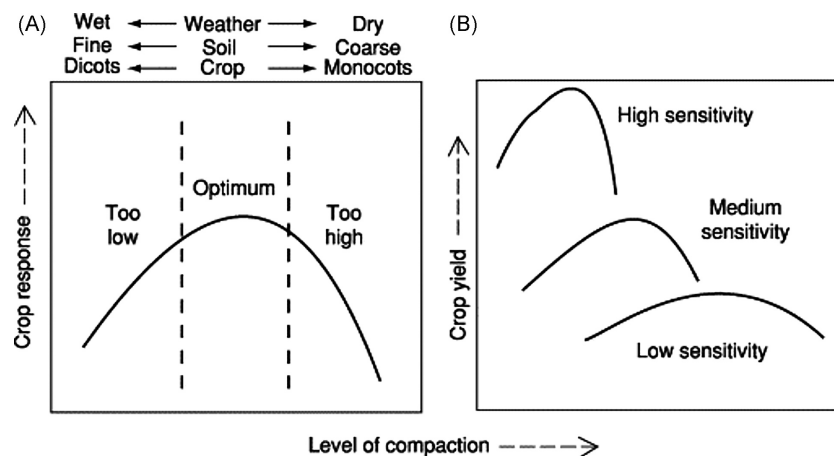


Fig. 7 Conceptual relationship of crop response to level of soil compactness in relation to weather, soil texture and (A) crop type and (B) crop sensitivity. Reproduced with permission from Lindstrom MJ and Voorhees WB (1994) *Soil Compaction in Crop Production*. Amsterdam: Elsevier.

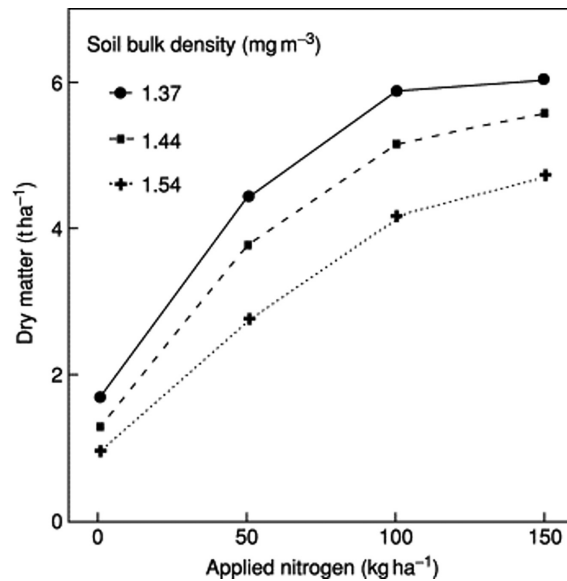


Fig. 8 Dry matter yields of ryegrass (first cut, mean of 3 years) in response to applied nitrogen fertilizer at three levels of bulk density (3–12 cm depth). Adapted from data in Douglas JT and Crawford CE (1993) *Grass Forage Science*. Oxford: Blackwell Science.

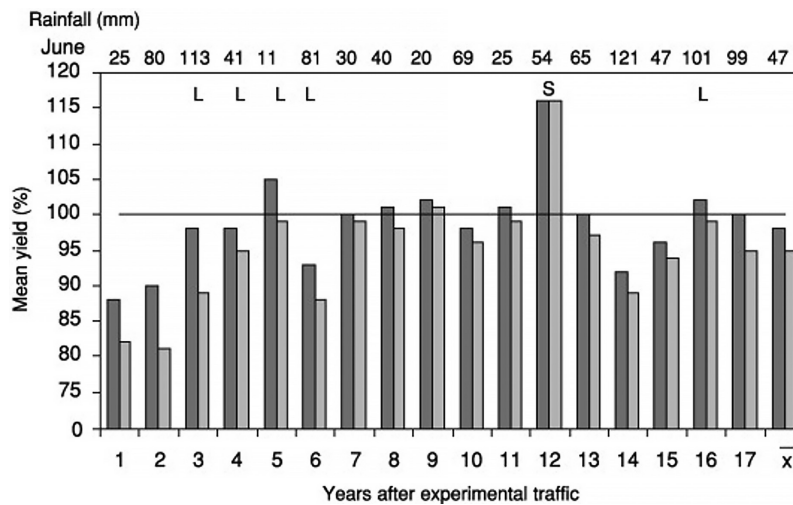


Fig. 9 Mean grain and nitrogen yields of annual crops in control treatment (=100%), and relative to the control in loading treatment of four passes with a 50 kN axle load in 1981 on clay soil, for 17 successive years after the loading. L, lodging; S, sprouting; relative grain yield (%); relative nitrogen yield (%). Reproduced with permission from Alakukku L (2000) *Advances in GeoEcology* 32. Reiskirchen: Catena Verlag.

influenced by rainfall during the subsequent growing seasons. In dry seasons, a moderately compacted subsoil sometimes may result in yield advantages due to reduced loss of soil water percolating below and out of the reach of the root zone. In wet seasons, however, the reduced permeability of the subsoil can lead to anaerobic conditions within the topsoil, resulting in direct damage to the crop and loss of nitrogen by denitrification. In dry seasons, compaction may prevent roots from growing into the subsoil due to high soil mechanical resistance, which limits the plants access to resources (particularly water and nitrogen) in the subsoil.

Modeling of crop responses to soil compaction

Crop growth does not reach its full potential when the uptake of water, oxygen, or nutrients is less than the demand of the crop. Potential crop growth is determined considering the prevailing weather conditions. Reduced crop growth may be caused by a reduction of the length of the growing period, low temperature, limited supply from the soil of water, oxygen, and nutrients to the root system, and a limited activity of the root system. Soil water plays a central role in these limiting factors, and effects of soil

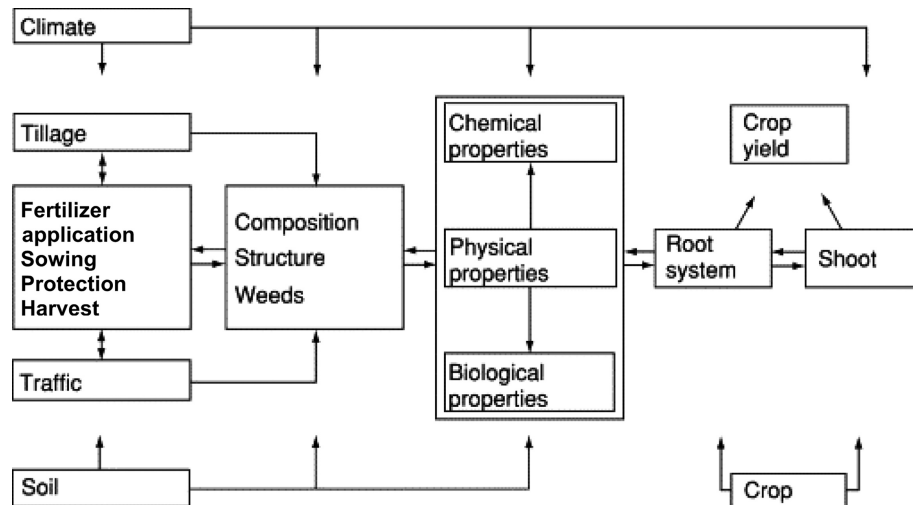


Fig. 10 Interrelationships among climate, soil management, soil properties, and crop growth. Reproduced with permission from Boone FR (1988) *Soil and Tillage Research*. Amsterdam: Elsevier.

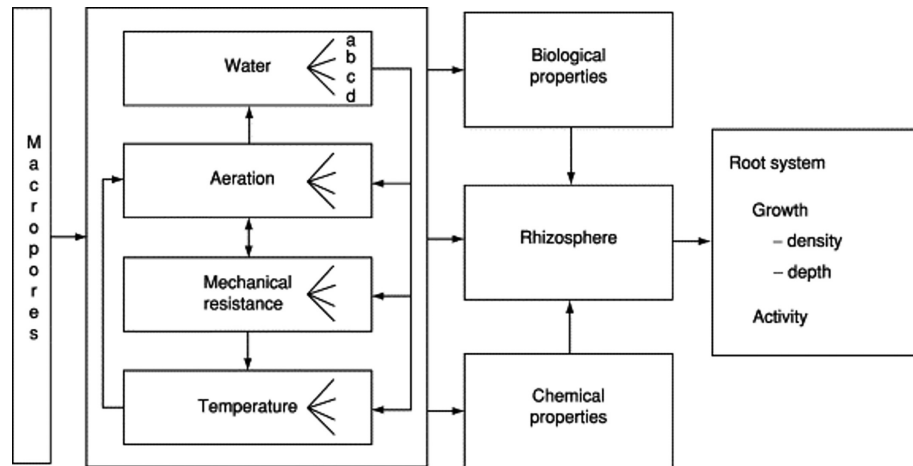


Fig. 11 Relationships among soil macropores, soil physical, chemical, and biological properties, rhizosphere and root system. a, surface boundary; b, storage; c, transport; d, sink or source aspects. Reproduced with permission from Boone FR (1988) *Soil and Tillage Research*. Amsterdam: Elsevier.

compaction on crop growth and biological functioning in relation to water are crucial. In Fig. 10, a scheme is presented of the interrelationships among soil tillage, field traffic, soil structure and soil physical, chemical, and biological properties. A part of this scheme is considered in more detail in Fig. 11. To simulate crop responses to soil compaction, all aspects presented in Figs. 10 and 11 should be included in a model. However, most crop models do not include all aspects, such as limitations of root growth due to increased soil mechanical resistance or lack of aeration, the role of macropores, reduced accessibility and availability of nutrients (e.g., loss of nitrogen by denitrification), and effects of reduced biological activity, and tend to underestimate the impact of compaction on crop growth.

Effects on environmental components

Soil compaction influences a number of environmental parameters, even at considerable distance from the original location at which the compaction occurred (Fig. 12). Compaction may change the fluxes of greenhouse gases from the soil to the atmosphere through mechanisms associated with effects on soil permeability, aeration, and crop development. Compaction increases CO₂ emissions because cultivation of compacted soils requires appreciably more energy than cultivation of uncompacted soils. Approximately 90% of the global N₂O emissions to the atmosphere come from soils. Compacted soils tend to be wetter than noncompacted soils, so denitrification is enhanced. The flux of N₂O increases rapidly as air-filled porosity declines (Fig. 13). Compaction from vehicle traffic prior to the establishment of cereal crops can cause marked increases in the N₂O fluxes during the early growth period in spring (Table 2).

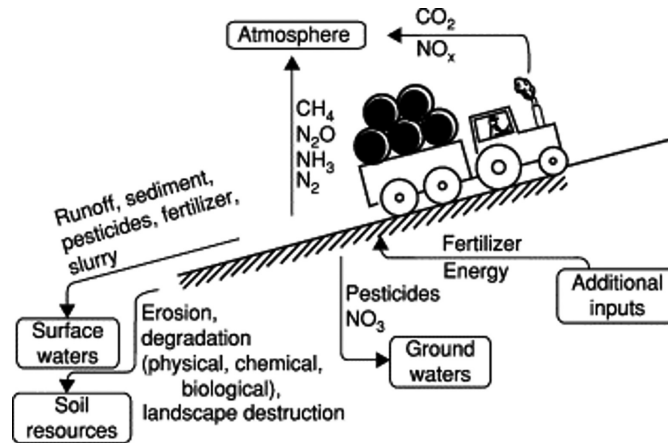


Fig. 12 A conceptual diagram showing the major pathways whereby compacted soil conditions may influence components of the environment. Reproduced with permission from Soane BD and Van Ouwerkerk (1995) *Soil and Tillage Research*. Amsterdam: Elsevier.

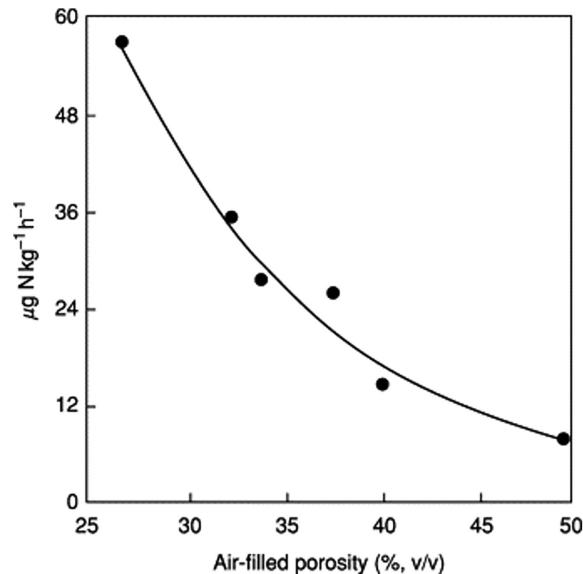


Fig. 13 Increased emission of nitrogen by denitrification as air-filled porosity decreases. Reproduced with permission from Sextone AJ, Parkin TB, and Tiedje JM (1988) *Soil Biology and Biochemistry*. Amsterdam: Elsevier.

Table 2 Influence of vehicle traffic (zero, light, heavy) prior to the establishment of wheat and spring barley on N_2O flux during the early spring growth period.

Crop	Period (dates and days)	Cumulative N_2O flux ($g\ N_2O-N\ ha^{-1}$)		
		Zero	Light	Heavy
Spring barley	16 May–14 July = 60 days	320	310	401
Winter wheat	8 March–8 May = 62 days	245	210	578

Adapted from data in Ball BC, Parker JP, and Scott A (1999) *Soil and Tillage Research*. Amsterdam: Elsevier.

Due to reduced permeability, compacted soils usually show greater runoff and hence greater risk of flooding and erosion than noncompacted soils. Surface rills and even gullies are sometimes directly associated with wheel tracks, particularly over seedbeds following periods of highly erosive rainfall. Surface waters may thus carry additional burdens of clay and silt, fertilizer, and pesticides when runoff occurs from areas of compacted soil.

Measurements of soil mechanical stresses

The stress state at a given point in the soil can be determined using strain gauge pressure transducers, more commonly called 'load cells' (Lamandé et al., 2015). Alternatively, one can use vibrating wire strain gages or strain gauge transducers, or fluid inclusion type probes. The strain-gauge pressure transducers can measure normal stresses and relate the stresses to a direction depending on the orientation of the transducer in the soil. The measurements may be made in a single or multiple directions simultaneously by a single load cell, depending on the number of faces of the transducer (Wiermann et al., 1999). Fluid inclusion probes may also be able to read stresses in specific directions, but usually they are used for measurements of mean normal stress.

The magnitude of stresses at a given point in the soil during wheeling is dynamic both in and across the driving direction. Several, similar sensors for measurements of a given stress across the full width of the wheel track provide the distribution of stresses during wheeling. However, to gain a complete picture of the stress-state under running gear, stresses need to be measured in multiple directions (the principal stresses, and mean normal and octahedral stresses). Stresses can be measured at different depths, which provides information on the propagation of stresses into the soil.

Strategies for the prevention of compaction

Options to manage soil compaction include minimizing compaction risks by adjusting induced stresses to soil strength, and restricting compaction to small areas such as tramlines as practiced in controlled traffic farming. The stresses induced by vehicles can be minimized by reductions in overall mass and reductions in wheel loads, and by increases in ground contact area and reduction of ground contact pressure. Low ground contact pressure will mainly reduce soil stresses in the topsoil and upper subsoil, as soil stresses in the deeper subsoil are primarily governed by the wheel load. Tires of greater width and diameter will increase ground contact area if combined with a reduction in inflation pressure (Ten Damme et al., 2020), and thereby reduce the ground contact pressure for a given wheel load. Limits in tire inflation pressure will also result in limiting the wheel load of that specific tire. The development of flexion technologies has played an important role here; inflation pressures for a 30 kN wheel load reduced from 140 kPa for a 16.9 R28 tire to 0.6 kPa for a 710/70 R 42 tire. Multiple axle systems (in particular tandem axles for trailers and tankers with low-pressure tires) reduce wheel loads and provide extra contact area, however, increase the number of times soil is wheeled if the axles are not positioned in an offset configuration. Dual wheels and tracks also allow for a larger ground contact area. However, if not implemented correctly, one risks compacting a larger soil volume. For tracks in particular, it is important to balance the machinery well to avoid high stresses under the front or rear of a vehicle, and to reduce peak stresses concentrated under the wheels and supporting rollers. In general, it is of primary importance to reduce wheel loads to reduce stresses in the deeper soil layers.

Tool-carriers (gantries, controlled traffic farming), up to 12 m width, have been found capable of providing traffic-free zones in which tillage requirement is reduced (because of the lack of compaction) and crop yield and crop quality may be improved. Vehicle traffic can also be reduced by combining different operations into a single-pass operation. However, an increase of wheel loads should be prevented.

Most importantly, the strength of the soil at the moment of field traffic should be considered. Trafficking and working soil when it is too wet should be avoided to preserve soil structural and mechanical stability. Tillage should be reduced as much as possible to optimize biological and physical processes that improve soil structure, in particular macroporosity.

Improving drainage will result in drier hence stronger soils. Increases in soil organic matter, either as a surface mulch or incorporated, will improve soil structure and increase soil elasticity and resilience and thereby reduce the risk on irreversible compaction.

Strategies and recommendations for prevention of soil compaction in agriculture often rely on simulation models. Such soil compaction models calculate stress propagation and soil failure in the soil profile for certain wheel or track loading characteristics and soil conditions (e.g., soil water content). Usually, these models use a stress propagation in the soil profile based on the theory of elasticity. A comparison between measured and simulated soil stresses highlighted the importance of accurate stress readings and realistic upper model boundary conditions, and suggested that the actual stress transmission could be predicted accurately according to the theory of elasticity for the conditions investigated (Keller et al., 2014). One approach in risk assessment using soil compaction models involves a quantitative comparison of stresses transmitted to the soil profile with soil strength, enabling evaluation of the risk of soil compaction for specific machinery and soil conditions. Often precompression stress is used as soil strength. Although there are some questions about the use of precompression strength (Keller et al., 2012) such tools can help farmers and advisors in planning and making decisions about specific traffic situations in the field. This approach can be used in decision support systems (e.g., www.terrano.dk), for example, in choosing the right equipment or tires. Comparisons of loading and soil conditions can also be used for mapping, for example, the wheel load carrying capacity (Schjønning et al., 2015) or for calculation of the number of trafficable days.

Simple rules of thumb to reduce soil compaction risk have been proposed. For example, at water contents around field capacity, traffic on agricultural soil should not exert vertical stresses in excess of 50 kPa at depths larger than 50 cm (Schjønning et al., 2012). Also, maximum values of mean ground contact pressures are recommended for different soil conditions to minimize topsoil and subsoil compaction, ranging from 80 kPa for wet soils in spring to 200 kPa for dry soils in summer (Tijink, 1998).

Compaction in forestry

Forestry operations often exert a large number of wheel passes with large wheel loads on very moist soils. Due to the rough forest environment with roots, branches, rootstocks, and often stones, and the highly dynamic wheel loads, tires of forest vehicles must be wear-resistant and with high inflation pressures. This causes a high risk of deep subsoil compaction, which will be very persistent (Riggert et al., 2019). On the other hand, the frequency of forestry management and the area involved is less than arable agriculture. However, the almost nonexistent and often only partial recovery of compacted subsoils by natural processes causes the degradation of soil quality by subsoil compaction to accumulate over time. In addition to a decrease in forest production, this results also in a greater risk of erosion and shallower rooting and vulnerability to dry or wet periods. Subsoil compaction can be limited to a relatively small part of the area by using controlled traffic on permanent skid trails with fixed GPS coordinates. Reduction of subsoil compaction is at least partly possible by using tracks and the use of slash (brush mats) dumped on the trail.

Amelioration of compacted soils

Natural weathering due to freezing and thawing is usually restricted to the top few centimeters of the surface soil and rarely penetrates to the subsoil. Swelling and shrinking arising from changes in water content can cause changes in soil structure that may lead to an increase in total porosity. Deep rooting is important to dry out deeper soil layers to induce shrinkage and improve structure. Roots of certain species can penetrate compacted layers, thereby creating new pores, but this does usually not involve an increase in total porosity because roots push the soil aside when growing through soil. Over time, however, roots interacting with soil organisms can promote aggregation and restructuring of soil, thereby decreasing compaction. Earthworms can decompact a soil by ingesting soil and releasing cast on the soil surface. Such improvements of macroporosity greatly improve permeability and gas diffusivity of soil. In this way, gradual improvements can be made in the soil physical quality of compacted layers. However, highly compacted parts, in particular layers that cannot be penetrated by roots, will recover very slowly and possibly not recover at all.

Compacted topsoils can be loosened by tillage. However, in many cases loosened compacted soils are cloddy and require more intensive secondary cultivation to achieve suitable seedbed tilth. Loosening subsoils requires high-draft special equipment and high traction, but it destroys the existing continuous macropores, and weakens soil structure and strength. Subsoiling should be done when the subsoil is dry with a minimum of loosening and with the aim to form cracks. Many loosened subsoils recompact within a few years, resulting in worsened soil physical properties and rootability. The result is that subsoils that have been loosened once must be loosened regularly every 4–5 years. Therefore, before subsoiling the subsoil should be inspected to determine rootability, aeration status, and drainability, to avoid unnecessary subsoiling.

Outlook

Compaction is one of the greatest forms of degradation facing soils. The impacts on biomass productivity affects food and fiber production, while the impact on the soil environment exacerbates the impacts of climate change. Compacted soils store less water, restrict root growth, erode more readily and emit more greenhouse gases. More erratic weather due to climate change results in fewer workable days, when soils are dry enough during spring and autumn operations to resist the stresses of machinery.

References

- Alakukku L (2000) Response of annual crops to subsoil compaction in a field experiment on clay soil lasting 17 years. *Advances in GeoEcology* 32: 205–208. Catena Verlag, Reiskirchen, Germany.
- Arvidsson J and Keller T (2004) Technical solutions to reduce the risk of subsoil compaction: Effects of dual wheels, tandem wheels and tyre inflation pressure on stress propagation in soil. *Soil & Tillage Research* 79: 191–205.
- Batjes NH, van Engelen VWP, and Dijkshoorn JA (2014) *Soil and Terrain Database for Central and Eastern Europe (SOVEUR), version 1.1*. ISRIC - World Soil Information. <http://www.fao.org/land-water/land/land-governance/land-resources-planning-toolbox/category/details/en/c/1032175/>.
- Berisso FE, Schjønning P, Keller T, Lamandé M, Etana A, de Jonge LW, Iversen BV, Arvidsson J, and Forkman J (2012) Persistent effects of subsoil compaction on pore size distribution and gas transport in a loamy soil. *Soil & Tillage Research* 122: 42–51.
- Boone FR (1988) Weather and other environmental factors influencing crop responses to tillage and traffic. *Soil & Tillage Research* 11: 283–324.
- Boussinesq J (1885) *Application des potentiels à l'étude de l'équilibre et des mouvement des solides élastiques*. Paris, France: Gauthier-Villars.
- Casagrande A (1936) The determination of the pre-consolidation load and its practical significance. In: Casagrande A (ed.) *Proceedings of the 1st International Soil Mechanics and Foundation Engineering Conference*, Cambridge, MA, 22–26 June 1936, vol. 3, pp. 60–64. Cambridge, MA: Graduate School of Engineering, Harvard University.
- Dawidowski JB and Lerink P (1990) Laboratory simulation of the effects of traffic during seedbed preparation on soil physical properties using a quick uni-axial compression test. *Soil & Tillage Research* 17: 31–45.
- Håkansson I, Voorhees WB, and Riley H (1988) Vehicle and wheel factors influencing soil compaction and crop response in different traffic regimes. *Soil & Tillage Research* 11: 239–282.
- Horn R (2003) Stress-strain effects in structured unsaturated soils on coupled mechanical and hydraulic processes. *Geoderma* 116: 77–88.
- Keller T, Arvidsson J, Schjønning P, Lamandé M, Stettler M, and Weisskopf P (2012) In situ subsoil stress-strain behavior in relation to soil precompression stress. *Soil Science* 177: 490–497.

- Keller T, Lamandé M, Arvidsson J, Berli M, Ruiz S, Schjønning P, and Selvadurai APS (2014) Transmission of vertical soil stress under agricultural tyres: Comparing measurements with simulations. *Soil & Tillage Research* 140: 106–117.
- Keller T, Colombi T, Ruiz S, Manalili MP, Rek J, Stadelmann V, Wunderli H, Breitenstein D, Reiser R, Oberholzer HR, Schymanski S, Romero-Ruiz A, Linde N, Weisskopf P, Walter A, and Or D (2017) Long-term soil structure observatory for monitoring post-compaction evolution of soil structure. *Vadose Zone Journal* 16. <https://doi.org/10.2136/vzj2016.11.0118>.
- Keller T, Sandina M, Colombia T, Horn R, and Ore D (2019) Historical increase in agricultural machinery weights enhanced soil stress levels and adversely affected soil functioning. *Soil & Tillage Research* 194(104): 293.
- Koolen AJ and Kuipers H (1983) *Agricultural Soil Mechanics. Advanced Series in Agricultural Sciences*, vol. 13.
- Lamandé M, Keller T, Berisso FE, Stettler M, and Schjønning P (2015) Accuracy of soil stress measurements as affected by transducer dimensions and shape. *Soil & Tillage Research* 145: 72–77.
- Nachtergaele FOF, van Linden GWJ, and Badjes NH (2002) Soil and terrain databases and their applications with special reference to physical soil degradation and soil vulnerability to pollution in central and eastern Europe. In: Pagliai M and RJA J (eds.) *Sustainable Land Management—Environmental Protection: A Soil Physical Approach*. Reiskirchen: Catena Verlag. *Advances in GeoEcology* 35: 45–55.
- Oldeman LR, Hakkeling RTA, and Sombroek WG (1991) *World Map of the Status of Human-induced Soil Degradation, an Explanatory Note*. Wageningen, Netherlands: ISRIC.
- Pulido-Moncada M, Munkholm LJ, and Schjønning P (2019) Wheel load, repeated wheeling, and traction effects on subsoil compaction in northern Europe. *Soil & Tillage Research* 186: 300–309.
- Riggert R, Fleige H, and Horn R (2019) An assessment scheme for soil degradation caused by forestry machinery on skid trails in Germany. *Soil Science Society of America Journal* 83: S1–S12. <https://doi.org/10.2136/sssaj2018.07.0255>.
- Schjønning P, Lamandé M, Keller T, Pedersen J, and Stettler M (2012) Rules of thumb for minimizing subsoil compaction. *Soil Use and Management* 28: 378–393.
- Schjønning P, van den Akker JJH, Keller T, Greve MH, Lamandé M, Simojoki A, Stettler M, Arvidsson J, and Breuning-Madsen H (2015) Driver-Pressure-State-Impact-Response (DPSIR) analysis and risk assessment for soil compaction—A European perspective. In: Sparks DL (ed.) *Advances in Agronomy*, pp. 183–237. Elsevier Inc.
- Söhne W (1953) *Druckverteilung im Boden und Bodenverformung unter Schlepperreifen. (Pressure distribution in the soil and soil deformation under tractor tyres)*. Grundlagen der Landtechnik 49–63.
- Ten Damme L, Stettler M, Pinet F, Vervaeke P, Keller T, Munkholm LJ, and Lamandé M (2020) Construction of modern wide, low-inflation pressure tyres per se does not affect soil stress. *Soil & Tillage Research* 204(104): 708.
- Ten Damme L, Schjønning P, Munkholm LJ, Green O, Nielsen SK, and Lamandé M (2021) Soil structure response to field traffic: Effects of traction and repeated wheeling. *Soil & Tillage Research* 213(105): 128.
- Tijink FGJ (1998) Mechanisation strategies to reduce traffic induced soil compaction. *Advances in Sugar Beet Research IIRB* 1: 57–66.
- Van den Akker JJH (2004) SOCOMO: A soil compaction model to calculate soil stresses and the subsoil carrying capacity. *Soil & Tillage Research* 79: 113–127.
- Van den Akker JJH and Schjønning P (2004) Subsoil compaction and ways to prevent it. Chapter 10. In: Van den Akker JJH, Schjønning P, Schjønning P, Elmholt M, and Christensen BT (eds.) *Managing Soil Quality: Challenges in Modern Agriculture*, pp. 163–184. Wallingford, Oxon, UK: CABI Publishing, CAB International. ISBN: 0 85199 671 X.
- Wiermann C, Way TR, Horn R, Bailey AC, and Burt EC (1999) Effect of various dynamic loads on stress and strain behaviour of a Norfolk sandy loam. *Soil & Tillage Research* 50: 127–135.

Further reading

- Batey T (2009) Soil compaction and soil management—A review. *Soil Use and Management* 25: 335–345. <https://doi.org/10.1111/j.1475-2743.2009.00236.x>.
- Håkansson I (ed.) (1994) Subsoil compaction by high axle load traffic. *Special Issue of Soil & Tillage Research* 29: 105–306.
- Horn R, van den Akker JJH, and Arvidsson J (eds.) (2000) Subsoil compaction: Distribution, processes and consequences. *Advances in GeoEcology* 32: 1–462.
- Larson WE, Blake G, Allmaras RR, Voorhees WB, and Gupta S (eds.) (1989) *Mechanics and Related Processes in Structured Agricultural Soils*, In: *NATO ASI Series E: Applied Sciences*, vol. 172.
- Monnier G and Goss MJ (eds.) (1987) *Soil Compaction and Regeneration*. Rotterdam, Netherlands: A. A. Balkema.
- Pagliai M and Jones RJA (eds.) (2002) Sustainable land management—environmental protection: A soil physical approach. *Advances in GeoEcology* 35: 1–588.
- Van den Akker JJH, Arvidsson J, and Horn R (eds.) (2003) Experiences with the impact and prevention of subsoil compaction in the European Community. *Special Issue of Soil Tillage Research* 73: 1–186.
- Van den Akker JJH and Soane B (2004) Compaction. In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*, vol. 4, 285–293.
- Van Ouwerkerk C (ed.) (1988) Tillage and traffic in crop production. *Special Issue of Soil Tillage Research* 11: 197–372.
- Van Ouwerkerk C and Soane BD (eds.) (1995) Soil compaction and the environment. *Special Issue of Soil Tillage Research* 35: 1–113.

Bioturbation—Physical processes

Siul Ruiz^a, Paul D Hallett^b, and Dani Or^{c,d}, ^aFaculty of Engineering and Physical Sciences, Bioengineering Group, University of Southampton, Southampton, United Kingdom; ^bSchool of Biological Sciences, University of Aberdeen, Aberdeen, United Kingdom; ^cInstitute of Biogeochemistry and Pollutant Dynamics, Soil and Terrestrial Environmental Physics group, Swiss Federal Institute of Technology (ETH Zurich), Zurich, Switzerland; ^dDivision of Hydrologic Sciences, Desert Research Institute, Reno, NV, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	100
Overview of bioturbation mechanical processes	101
Soil physical constraints to bioturbation	102
Bioturbation by penetrating plant roots	103
Bioturbation by earthworm burrowing	103
Models for soil bioturbation by growing plant roots and earthworms burrowing	104
Rodent excavation and burrowing activity	106
Habitat formation and nesting activity in soil by miner bees	107
Bioturbation by ants and termites in tropical and arid regions	108
Termite mounds physical functioning	109
Biophysical modeling of granular transport by ants and termites	112
Conclusions	113
References	113

Abstract

Bioturbation is the physical movement of soil by fauna or plant roots. The primary goal of bioturbation is for organisms to access resources in soil such as food, plant nutrients and water. Bioturbation includes soil translocation by burrowing and ingestion, soil deformation by organism expansion, and soil stabilization by compounds secreted by organisms and their tissue. These processes are extremely important to the formation of soil structure, with impacts to flow and transport properties. Damage to soils from intensive management, compaction or carbon depletion limits the capacity for bioturbation, but bioturbation also helps to restore degraded soils.

Key points

- Bioturbation is a key process in the formation and maintenance of soil structure.
- Bioturbation is caused by a range of organisms that live in soil, including plant roots, insects such as termites, and animals such as earthworms and burrowing mammals.
- *Penetrator* organisms push into soil whereas *excavator* organisms displace soil by carrying or shoveling it with limbs.
- Soil structure is altered by bioturbation through the translocation of soil by burrowing or ingestion, the mechanical deformation of soil by the forces from burrowing, and the secretion of mucilage and other compounds by soil-dwelling organisms.
- Changes to soil structure by bioturbation affect flow and transport properties of soil.
- Biopores are a product of bioturbation. These open channels formed by organisms are rapid transport pathways that also ease movement of subsequent organisms, especially root growth.
- Land mismanagement decreases the abundance and functional capacity of bioturbating organisms, with severe implications to soil.

Introduction

Bioturbation is defined as the alteration of soil structure induced by biological agents (i.e., plants or animals). More specifically, bioturbation comprises the biophysical processes that internally displace soil solids, induce the mixing of soil constituents, re-organize the physical pore spaces in soil, and redistribute soil organic compounds. There are many biophysical processes in bioturbation, which include direct impacts from crevice burrowing, soil ingestion and aggregate displacement, and indirect impacts from secretions by plant roots (rhizodeposition), earthworms (casts and mucus) or associated microorganisms. Bioturbation can enhance the transport of water and gasses, provide reusable channels for future root growth, and accelerate the rates of soil structure formation, thus mitigating adverse impacts of soil structure degradation and restore soil function.

Soil bioturbation in natural soils is central in forming and maintaining soil structure that provides essential ecosystem services for a variety of biomes. The resulting biopore networks provide preferential pathways for water flow and extend oxic conditions

deeper into soil profiles (Ruiz et al., 2021). Biopores act as hotspots for biological activity and create paths of least resistance for subsequent root growth (Colombi et al., 2017). For water, these structures improve deep recharge to groundwater (sometimes at the expense of enhanced risk of pollution) and improve overall soil water retention. Bioturbation has been shown to facilitate root growth, which can further enhance the amount of atmospheric carbon stored below ground in the form of root biomass and root exudates (Beer et al., 2010). Regions with rich plant-derived organic matter act to stimulate microbial activity (Kuzakov and Blagodatskaya, 2015) and provide nourishment for earthworms (Lavelle et al., 2007). Soil ingestion by earthworms can also augment microbial activity, stimulating the aggregation of soil particles, which also contribute to soil structure. The ensemble of effects attributed to soil bioturbation activity may enhance crop yields by up to 25% (Van Groenigen et al., 2014).

Bioturbation is central to carbon cycling through numerous processes. The most direct is through the growth and development of organisms for acquisition of nutrients and water; the secondary impact is the promotion of soil biological activity. Rhizodeposition and decaying roots temporarily sequester photosynthetically fixed atmospheric carbon in soils. The below-ground carbon investment can range from 20% to 85% of total plant biomass (Gabet et al., 2003). Fine roots are produced at annual rates of 0.3–1 kg per m² ground surface, depending on ecosystems (Ruiz et al., 2017), which are consistent with other estimates for primary productivity in the literature (Beer et al., 2007). These carbon investments become primary sources of soil organic matter (SOM), which provides energy for heterotrophic activities by many invertebrates.

Bioturbation can enhance soil structure formation rates by 1–2 orders of magnitude relative to abiotic processes. Invertebrates contribute considerably to the movement and mixing of soil, with earthworms reported to move 6.5–100 kg soil per m² surface annually, ants impacting up to ~3 kg soil per m² surface annually, and termites moving up to ~0.6–1 kg soil per m² surface annually (Gabet et al., 2003). In doing so, these organisms act to mix the organic constituents through upper soil horizons. Their activities may also help to reduce the runoff of nutrients such as nitrogen and phosphorus.

Organisms such as earthworms, miner bees, and rodents play an active role as ecosystem engineers, but physical constraints, nutrient availability and pH can prohibit their activity. A pedological example are the distinct horizons that form in temperate Podzols because low pH soils deter earthworm activity. Arid soils may be too dry or hard, and tropical soils may have very low pH (Ruiz et al., 2021), so ants, termites, and reptiles, adapted to these particular environmental niches, become the dominant bioturbators (De Bruyn and Conacher, 1990). This highlights the different strategies used by organisms to explore soil resources and create subterranean shelters across different biomes and climatic conditions. For example, termites are fascinating ecosystem engineers that build intricate galleries with specific architecture for regulating heat and moisture, superbly tuned to the functioning of symbiotic fungi they cultivate, and enabling their habitats to persist in warm tropical regions (Ocko et al., 2019).

When soils are intensively managed for agriculture, damage from compaction and tillage can impede burrowing. The current move towards less intensive tillage systems has an aim of restoring the capacity for biology, rather than mechanical cultivation, to form soil structure. This is especially important in regenerative agriculture, where soil bioturbation by plant roots and earthworms penetrating soil aid in restoring degraded soils.

Soil bioturbation processes and their impact on soil structure will be described in this chapter. Bioturbating organisms and their tunneling characteristics are listed, and split into those that generate tunnels through penetration processes (penetrators) and those that displace soil by carrying or shoveling it with limbs (excavators). The mechanics of bioturbation by growing plant roots and burrowing earthworms, both of which penetrate and expand soil cavities, will first be described. Then larger-scale bioturbation by mammals, and their resulting impact on soil functionality, will be shown. Granular transport by ants and termites, that dominate in drier and warmer regions, demonstrates climatic impacts. The chapter ends with a broader outlook of other bioturbation agents that affect soil physical properties.

Overview of bioturbation mechanical processes

Different organisms employ different strategies to accomplish resource exploration and niche formation via soil bioturbation. Organisms without limbs, such as earthworms and plant roots, generate biopores (largely) by axial penetration and radial expansion processes (Fig. 1a–d). Despite general similarity in the resulting cylindrical biopores produced by these two organisms, they differ in physical characteristics and form through slightly different mechanical processes. Plant roots impact mechanical stresses on the soil at the tissue level due to expansion of cells (Greacen and Oh, 1972), driven by the softening of cell walls by their polymers modifying in response to developmental cues and the environment (Fig. 1a). In contrast, organisms such as earthworms use their muscular system to displace (and sometimes ingest) soil under relatively wet conditions (Fig. 1c).

Limbed and toothed organisms can grab or shovel soil particles to form burrows and cavities (Fig. 1e–g) (Rosario Carotenuto et al., 2020). For clarity, we will designate these organisms as excavators. Examples of mammal excavators are rodents such as moles or badgers that use their claws and even teeth to break up and shovel soil to form burrows (Fig. 1e and f). Similarly, smaller organisms such as ants or termites make use of their mandibles to carry and displace soil particles (Fig. 1g).

Bioturbation processes are constrained by governing physical processes occurring across a range of scales, inhibiting penetrators vs excavators differently. A summary on their relative bioturbation activity can be found in Table 1. Bioturbation activity with respect to mass displaced per unit land area is compared to the amount of soil impacted by tillage activity, converted from soil volumes assuming a range of soil bulk densities from 1200 to 1800 kg m⁻³ (Or et al., 2021). While there are limited data on the absolute magnitude of annual mass impacted by soil bioturbation, estimates for the combined impact of earthworms and plant roots have been compiled and compared to that of tillage activity by humans (Table 1). Soil bioturbation impacts just over half the

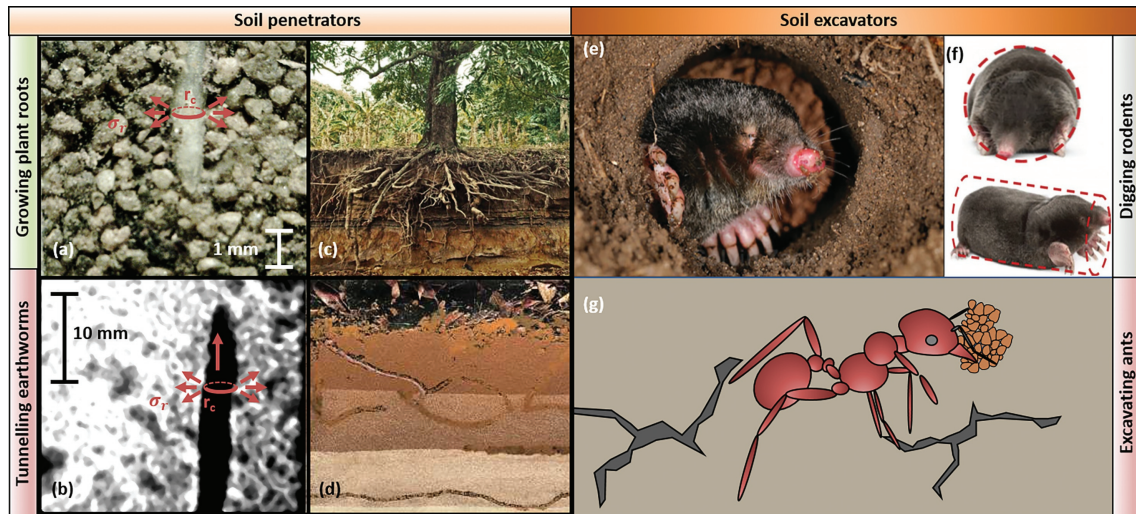


Fig. 1 Primary bioturbation processes. Left panels (a–d) highlight bioturbation through penetration processes such as plant root growth (a) or peristaltic motion of earthworms (b). Their penetration-cavity expansion processes result in channels that form biopores in soil (c and d). The right panels (e–g) highlight excavation processes through rodents digging (e and f) or ants carrying out material (g). Figure was adapted from Ruiz S, Schymanski S and Or D (2017) Mechanics and energetics of soil penetration by earthworms and plant roots—Higher burrowing rates cost more. *Vadose Zone Journal* doi: 10.2136/vzj2017.01.0021; Rosario Carotenuto A, Guarracino F, Šumbera R and Fraldi M (2020) Burrowing below ground: Interaction between soil mechanics and evolution of subterranean mammals. *Journal of the Royal Society Interface* 17: 20190521.

Table 1 Soil bioturbation by different organisms based on displaced soil mass per unit area annually (note affected soil depth vary widely across biomes and plant species).

	Tillage	Plant roots	Earthworms	Rodents	Ants	Termites
Annually displaced soil mass per land area [kg m^{-2}]	150–230	0.3–1.5	6–100	0.4–5.5	3	0.6–1
Annually displaced soil mass [kg]	$2.2\text{--}3.2 \times 10^{15}$	$1.2\text{--}1.7 \times 10^{15}$	–	–	–	$10.8\text{--}18 \times 10^6$

Taken from Or D, Keller T and Schlesinger WH (2021) Natural and managed soil structure: On the fragile scaffolding for soil functioning. *Soil and Tillage Research* **208**: 104912.

soil volumes affected by tillage. Considering that tillage is one of the largest geo-engineering activities on earth, the impact of soil bioturbation is likely underestimated and plays a significant role in shaping terrestrial ecosystems.

While extending these estimates to other soil bioturbation organisms is difficult, estimates considering global maps of invasive termite species can provide rough estimates of their displaced soil masses (Buczkowski and Bertelsmeier, 2017). Termite activity affects roughly $18 \times 10^6 \text{ m}^2$ of terrestrial land, displacing soil masses estimated to range from 10.8 to $18 \times 10^6 \text{ kg}$ annually. There are many organisms that may contribute to soil bioturbation at similar orders of magnitude.

The subsequent sections will expand on the physical processes that underpin different types of soil bioturbation and elucidate how these processes are constrained by varying soil conditions, ultimately highlighting how soil physical conditions can explain particular niches for differing soil bioturbation activity.

Soil physical constraints to bioturbation

Organisms are constrained in bioturbation by soil moisture and strength. As soils dry, capillary forces from water begin to increase the cohesion between particles, with strength increasing further through chemical bonding when soils are very dry. Organic matter and soil aggregation can decrease the strength of soil experienced by an organism, with feedback occurring when these organisms regenerate the structure of soil. When soils are waterlogged, bioturbation will be limited by oxygen availability. For plants, root growth rates will be constrained as soils approach the wilting point at -1500 kPa water potential.

To overcome these physical constraints from soil, organisms have developed growth and burrowing mechanisms described below. The ecology of burrowing organism is strongly dependent on soil physical conditions, with penetrators dominating in temperate and excavators dominating in arid regions. Whereas a temperate region may provide a soft, moist soil that is easy to push through, an arid region has dry soil that is lighter to carry.

Bioturbation by penetrating plant roots

Plant roots must overcome soil mechanical resistance, water-logging and periods of drought during growth (Ruiz et al., 2019). To a great extent, plant roots can overcome these stresses with their remarkable capacity to exert large pressures (>1 MPa), to stimulate biological activity and to modify surrounding soil (the rhizosphere). Root growth deforms and displaces the soil matrix to form biopores that persist long after roots decompose. During their active growth, plant roots slough off cells and exude polymeric substances, presumably for reducing friction and improving hydraulic contact with the surrounding soil, which results in lining and reinforcement of biopores. Exuded polymeric substances also improve water retention, enhance soil aggregation and provide a conductive and living environment around the roots often seen as a rhizosheath (Basirat et al., 2019). Furthermore, root water uptake to meet atmospheric demand (transpiration flux) extends the depths and extent of soil drying. This, in turn, may enhance abiotic effects of wetting-drying cycles that can form cracks and assist soil aggregation.

The ability of roots to penetrate soil relies on multi-scale growth processes (Pritchard, 1994). At the molecular level, roots are largely constructed by parenchyma cells that maintain thin yet rigid primary walls with a remarkably high mechanical yield strength, estimated at around $0.5\text{--}2 \times 10^6$ Pa (Gibson, 2012). Extension of these cells occurs when their internal pressures exceed both the yield of the cell wall and the external mechanical impedance exerted by the surrounding soil. Cellular extension is complemented by an enzymatic loosening processes, causing periodic reorientation of cellulose fibrils in the cell wall (Pritchard, 1994). This enables tip wise extensional growth to reduce frictional effects near the root interface with the soil, and ultimately it allows roots to locally impart pressures upwards of $1\text{--}2 \times 10^6$ Pa; magnitudes great enough to fracture chalk rock (Misra et al., 1986). Under more mechanically stressed conditions, such as compaction, plant roots display a radial thickening in the zone of elongation behind the root cap (Hettiaratchi et al., 1990). The radial growth, primarily associated with the outer cortical cells, together with the enhanced production of root hairs by the epidermal cells, helps anchor the root behind the zone of cell extension. The internal arrangement of tissues limits the ability of roots to grow through smaller pores, as they lack mechanisms to become thinner. However, some studies have also suggested that radial thickening allows roots to exploit pre-existing fractures to propagate cracks. Alternatively, radial thickening may be the means of redistributing local stresses on the root. The multiscale growth process that enables plant roots to exert such high pressures also limits the actuation speeds of plants. Growth rates are constrained by poroelastic limitations, which are the fastest water driven actuation processes within a plant. As a result, root growth rates are in the order of $1\text{--}2 \times 10^{-7}$ m s⁻¹ (Dumais and Forterre, 2012) (Fig. 1a and b). A more rigorous mechanical description of this will be described later in the text.

The cumulative impact of a plants root system on soil is highly dynamic, as laterals growing out of primary roots are in a constant flux of growth and death. This results in an evolving network of tunnels. Furthermore, the accumulation of rhizodeposits in the rooting zone act to reinforce the tunneling structures. As such, these networks are more persistent than a pore produced through only mechanical means. The persistence of root biopores acts as a reliable avenue for subsequent preferential root growth, as they provide paths of least mechanical resistance. This is very important in particularly hard soils.

Beyond the biopores formed by root growth, bioturbation can result from root-soil entanglement, particularly for trees (Fig. 2). Root-soil entanglement acts to enhance bulk soil mechanical properties, reducing hillslope susceptibility to shallow landslides (Fig. 2b). However, the same entanglement by a root network can lead to a phenomena is known as tree throw (Fig. 2c) (Gabet et al., 2003). Strong forces associated with wind, snow overload, or even sufficient root decay can lead to the toppling of trees. As the tree falls, the soil mass entangled with the roots can be pulled out with the tree, resulting in a pit where the tree was originally rooted. Subsequent to the uprooting, the roots begin to decay, thus the soil is released, creating a mound that forms under the decaying tree. This process occurs over a period of 5–10 years (Gabet et al., 2003).

Bioturbation by earthworm burrowing

Earthworm bioturbation plays an important role in soil formation, stimulation of biological activity and in organic carbon transformations, thereby promoting soil health in temperate regions. As mentioned above, earthworm burrows improve physical conditions for soil ecological functioning (i.e., water flow, retention and aeration), and provide preferred pathways that can be reused by growing plant roots (Ruiz et al., 2021). Earthworms line their burrows with casts and mucus excretion, which incorporate organic constituents, making these hotspots for biological activity (Lavelle et al., 2007). Certain earthworm ecotypes (i.e., permanent deep burrowing anecic earthworms) are also known to pull leaf litter from the surface deep into their burrows (Palm et al., 2013). This activity enhances the pool of soil organic matter and augments the introduction of carbon to the profile, adding to contributions from root growth and decay. The mechanics of burrow formation follow two main processes; the primary is simply soil displacement by muscle actuation and hydroskeleton (Dorgan, 2010) action, and the other is by soil ingestion. Limited experimental evidence suggests that while earthworms ingest large quantities of soil, active burrowing is often done by mechanical means while in search of carbon rich regions in the soil (Ruiz et al., 2017).

The primary mechanism of earthworm burrowing is penetration-cavity expansion. This uses a system of circumferential and longitudinal muscles (i.e., their hydroskeleton) for mobility. Earthworms can alternate between contracting and stretching these muscle fibers in order to generate peristalsis (Fig. 3a–c) (Purves et al., 2003). This cyclic forward propulsion and subsequent expansion allows earthworms to penetrate and anchor into soil, enabling their ability to crevice burrow. From the action of their muscle fibers earthworms can produce actuation speeds on the order of $1\text{--}5 \times 10^{-4}$ m s⁻¹, which is three orders of magnitude faster

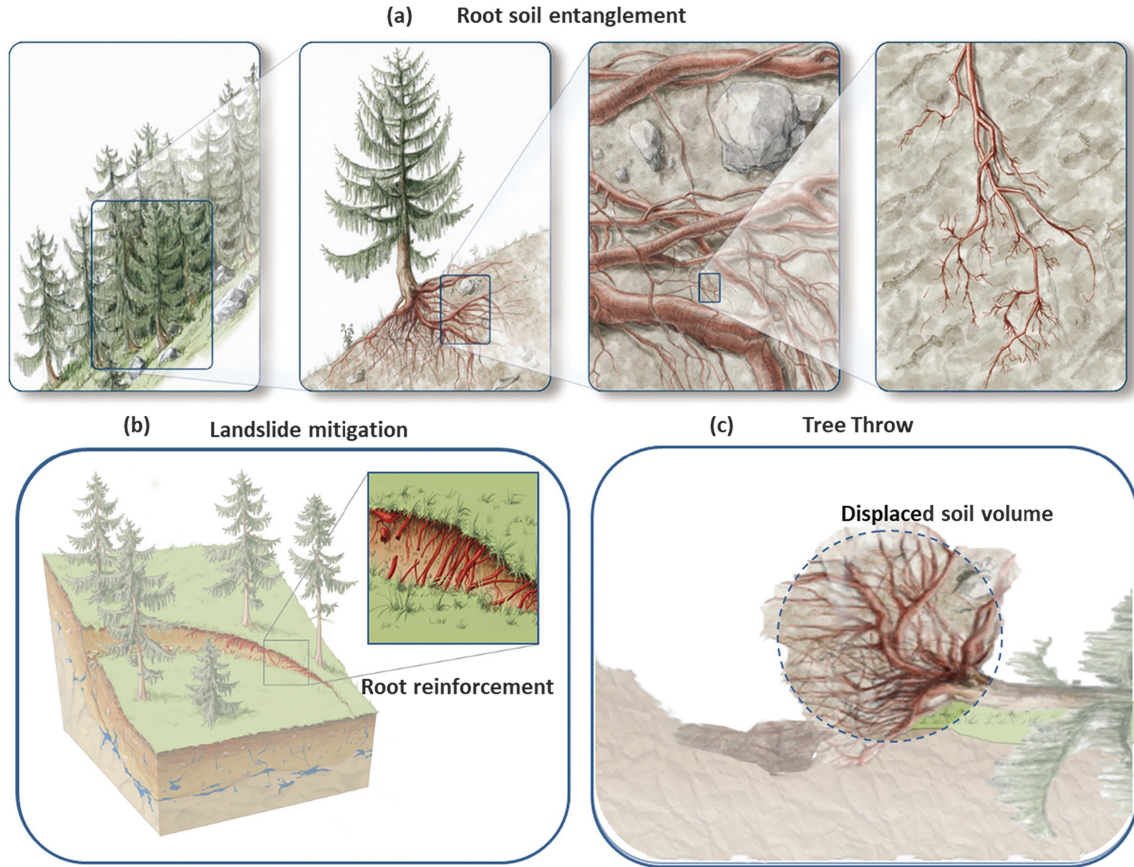


Fig. 2 Influence and scale of root-soil entanglement. The root-soil entanglement spans multiple scales from fine roots to rooting networks of an entire plant and even include multiple plants. (b) The presence of roots can act to mitigate landslides by acting to mechanically reinforce the soil. (c) For aging roots or dead vegetation, large winds and storms can act to uproot the tree; a processes known as tree throw. Tree throw can act to displace large volumes of soil bound to the root network.

than the rates that plant roots grow. As a trade-off to quicker actuation speeds, the absence of a rigid cell wall greatly reduces the maximum pressures that earthworms can exert, which are only on the order of $0.6\text{--}2 \times 10^5$ Pa (Ruiz and Or, 2018). This ultimately restricts the soil mechanical conditions that are suitable for earthworm activity.

Models for soil bioturbation by growing plant roots and earthworms burrowing

Both the penetration-cavity expansion burrowing of earthworms and tip extension of plant roots into soil can be described as local penetration processes, where soil is displaced and deformed radially about these biological surfaces. Recent models quantified the pressures required by earthworms and plant roots to form biopores in wet soils based on the mechanics of radial cavity expansion in an elasto-viscoplastic medium (Ruiz et al., 2017). More specifically, the model defines the radial stresses induced by either an earthworm's hydroskeleton or expanding root tissue at a cavity wall as σ_r (Ruiz et al., 2017). The minimum mechanical stress (pressure) required to expand a radial cavity in wet soil is given as:

$$\sigma_r(R_p) = P_L - 2s_u \ln\left(\frac{R_p}{r_c}\right) = s_u \quad (1)$$

where r_c [m] is the cavity radius, P_L [Pa] is the pressure at the cavity interface, R_p [m] is the radius of the elasto-viscoplastic interface (i.e., remote elastic effect), and s_u [Pa] is the soil shear strength. The ratio between the cavity zone and the inelastic zone converges to the ratio between the shear modulus of rigidity and the soil strength:

$$\left(\frac{R_p}{r_c}\right)^2 \rightarrow \frac{G}{s_u} \quad (2)$$

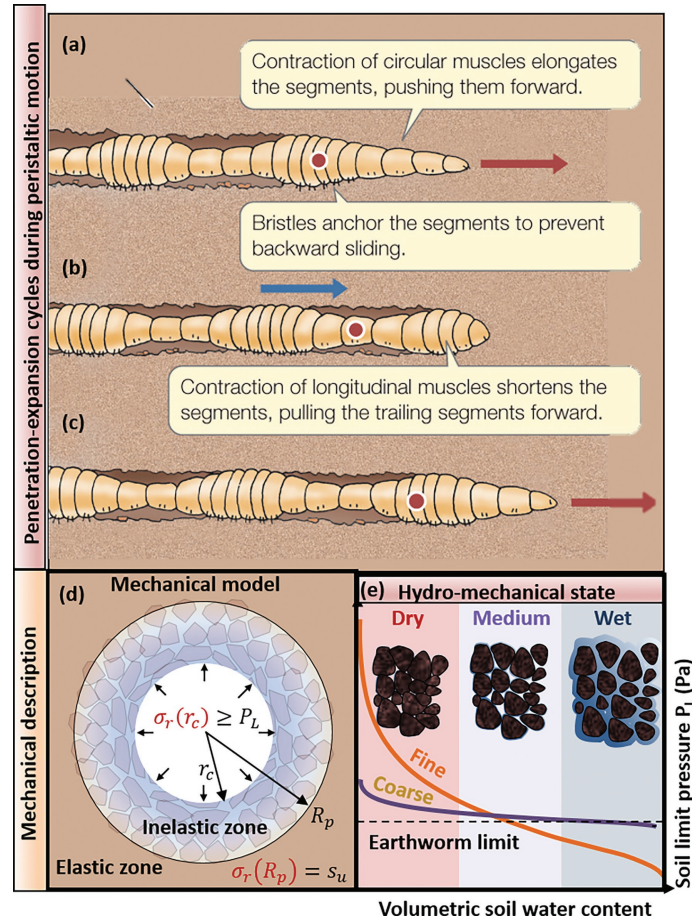


Fig. 3 Penetration-cavity expansion model for soil bioturbation (Ruiz et al., 2017) and (Dorgan, 2015). Example illustrates earthworm peristaltic motion that translates into penetration-expansion processes (a–c). Physical characteristics can be mechanically described by cavity expansion in wet soils (c), where the soil mechanical properties are linked to soil saturation state and texture class (d). Figure was adapted from Purves WK, Sadava D, Orians GH and Heller HC (2003) *Life: The Science of Biology: Volume III: Plants and Animals*. Macmillan.; Ruiz SA, Bickel S and Or D (2021) Global earthworm distribution and activity windows based on soil hydromechanical constraints. *Communications Biology* 4: 1–9.

where G [Pa] is the shear modulus of the soil. Substituting in the ratio in Eq. (2) into Eq. (1) determines the lower limit pressure required for cavity expansion:

$$P_L = s_u \left(1 + 2 \ln \left(\frac{R_p}{r_c} \right) \right) = s_u \left(1 + \ln \left(\frac{G}{s_u} \right) \right) \quad (3)$$

The resulting expression provides the minimum amount of pressure an earthworm or plant root would have to exert to expand a cavity radially in soil, based on the soil mechanical properties (Fig. 3d). Note that G and s_u are dependent on soil texture class and soil moisture status (Ruiz et al., 2017).

To consider the distinction between earthworms and plant roots, viscosity should be considered, as it takes into account the differences in their respective penetration rates ($1 \times 10^{-4} \text{ m s}^{-1}$ vs. $1 \times 10^{-7} \text{ m s}^{-1}$ respectively). Further distinctions can be found based on their respective pressure limitations. An earthworm's hydroskeleton is mechanically limited to a maximum pressure of $P_w = 2 \times 10^5 \text{ Pa}$ (Ruiz and Or, 2018). Thus, earthworms are mechanically impeded by soil conditions when $P_L \geq P_w$. Plant roots, may continue to grow in mechanically restrictive conditions by contrast, as $P_r = 2 \times 10^6 \text{ Pa}$. These constraining soil mechanical conditions are linked to the hydration status and texture of soil (Ruiz et al., 2017).

The resulting mechanical constraints on soil bioturbation have been used to outline global regions that permit earthworm bioturbation (Fig. 4). Furthermore, this mechanistic description also enables predictions for earthworm activity windows, which were found to be consistent with annual seasonal cycles. It is worth noting that there are auxiliary constraints that may also impede their biophysical activity such as water availability, soil pH, and temperature.

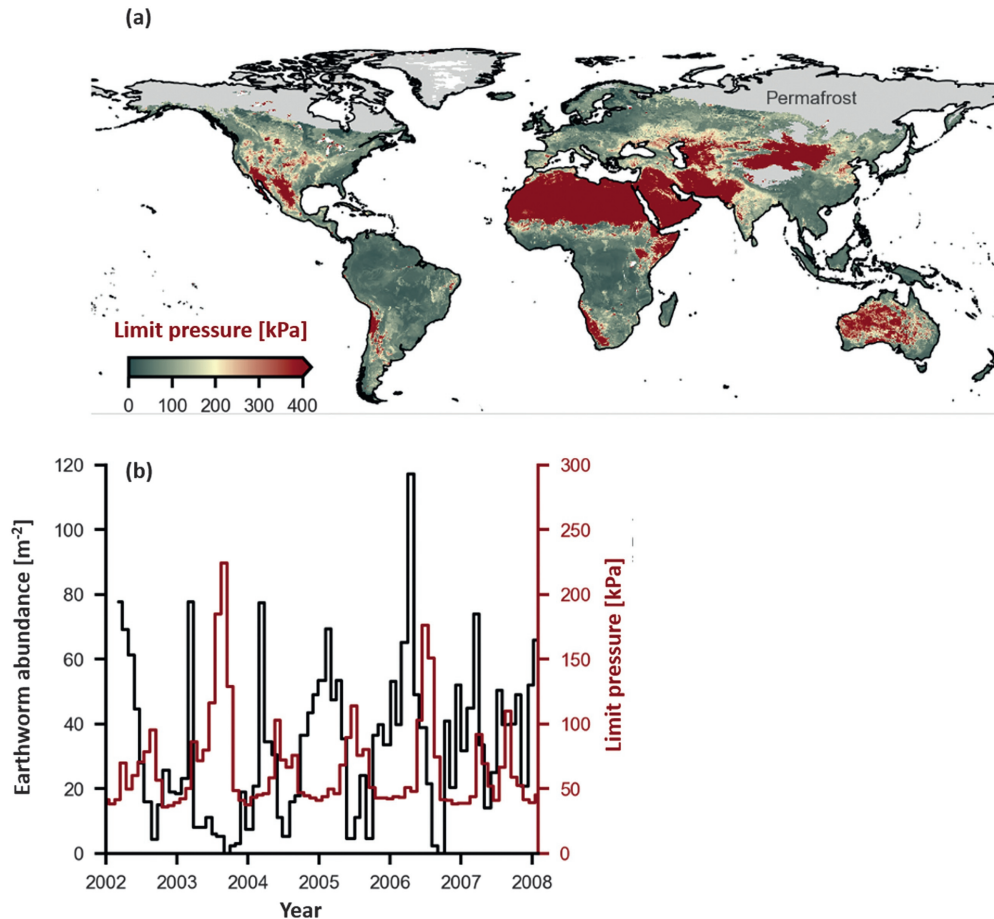


Fig. 4 Mechanically hospitable regions for earthworm activity (a) and activity time windows (b). The map in (a) illustrates minimum pressure required to expand a cavity in soil (limit pressure) associated with physiological pressure limitations of earthworms, highlighting regions that would impede earthworm activity in yellow to red. Activity windows in (b) illustrate the inverse correlation between limit pressure and monitored earthworm abundance in the United Kingdom. Figure adapted from Ruiz SA, Bickel S and Or D (2021) Global earthworm distribution and activity windows based on soil hydromechanical constraints. *Communications Biology* 4: 1–9.

Rodent excavation and burrowing activity

Burrowing activity by rodents, such as gophers and badgers, is another form of soil bioturbation. It is estimated that rodent populations alone dislodge 4–55 t of soil per ha annually (Table 1). Such displacement can produce 3–15 m³ of burrows per ha annually, depending on the relative abundance of different rodents. Under favorable conditions, gopher abundance can be as high as 125 per ha. These conditions appear to be proportionate to vegetated ground cover. Thus, rodent burrowing will have their most dominant impact in natural grasslands, then forests, and lastly in agricultural settings (Bocek, 1986).

Rodent burrow structures consist of openings to the soil surface, nesting regions, food storage sections, and feeding tunnels (often parallel to the surface). Tunnel depths are associated with several soil physical characteristics such as texture and water table height, however, they are also influenced by root penetration depth as plant roots are a primary source of food. Individual gophers burrow tunnels exceeding 150 m cumulative length, with up to one third of the tunnels being horizontal. Most of the burrowing length is dedicated to nesting compartments and tunnels to the surface (often in sections less than 1 m long). Nesting and storage chambers are often placed at greater depths (0.5–1 m depth), see Fig. 4. However, certain gopher species construct nesting mounds on the surface (Bocek, 1986; Hickman, 1977) (Fig. 5).

Rodent burrows result in tunnel sizes much larger than earthworm galleries, with gopher tunnels having mean radii of about 0.03 m (Vleck, 1981). The influence that they have on soil fertility is not immediately evident. The larger tunnel sizes are often associated with visible mounds on the surface. Initial tunneling activities may adversely impact vegetation by cutting roots and even main stems, causing an initial reduction in vegetation by as much as 20% (Grant et al., 1980). Despite these initial effects, evidence suggests that the soil re-worked by this activity becomes broken up and loosened, and decaying plant biomass becomes mixed with the impacted soil. These soils ultimately become more friable and richer in soil organic matter, which facilitates subsequent soil

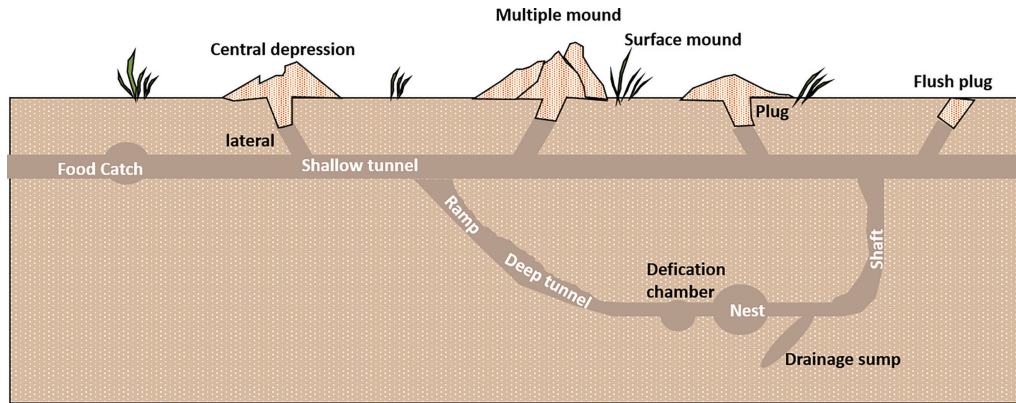


Fig. 5 Diagram of gopher burrowing system. Highlighted are shallow and deep tunnels consisting of food compartments, nests, and drainage systems. Figure adapted from Hickman GC (1977) Burrow system structure of *Pappogeomys castanops* (Geomysidae) in Lubbock County, Texas. *American Midland Naturalist* 97: 50–58.

biological and physical functionality. Abandoned tunnels eventually collapse under gravity and rain fall, and gradually fill with material from the surface (Bocek, 1986). Consequently, the soil becomes more fertile through this loosening, as it permits greater air and water transport through the pore space. Lastly, the loosened soil enables better plant root soil penetration during root tip growth.

Depending on the behavioral group, rodents tunnel through soil using either their frontal claws (i.e., scratch-digging) or their strong incisor teeth (i.e., chisel-tooth digging) to loosen soil (Fig. 6) (Samuels and Valkenburgh, 2009). Loosened soil will then be pressed radially along the tunnel or displaced into regions of the system that are no longer in use (Grinnell, 2020). Limited studies have tried to mechanically quantify these activities by assessing the mechanics associated with the penetration of rodent incisors into soil (Ma et al., 2020). While previous studies have provided certain insights into some of the underlying mechanics associated with the burrowing activity (Ma et al., 2020), investigations with more rigorous experimental and modeling methodologies are needed.

Habitat formation and nesting activity in soil by miner bees

Miner bees are solitary bees that construct nests in the soil, preferring well drained conditions with soils comprised of moderately high silt and clay contents (Lybrand et al., 2020). As they feed on flower nectar and act as pollinators, their preferred habitats are in areas with abundant floral diversity (Rau, 1929). Despite being somewhat solitary, they maintain their tunnels neighboring other miner bees. These crowded villages may consist of 100–1000 burrows that the bees themselves have difficulty distinguishing. The tunnels accessing these burrows end in large “mud chimneys” or turrets that are visible above ground (Fig. 7a). While the exact purpose of the turrets is not clear, certain speculation suggests that these chimneys serve as land marks for returning bees (Rau, 1929). Turret sizes are about 0.005–0.01 m in diameter and vary in length (upwards of 0.06 m). Below ground, the tunnels are hollow, tortuous and hard to follow due to the closeness to other nests. They are not sealed by bees. Both tunnels and turrets contain greater concentrations of clay, which make them hard. Clay is also used deep in the tunnels by bees to form sealed compartments

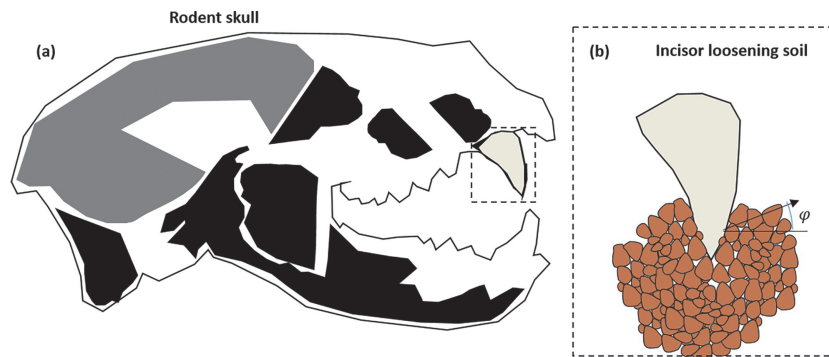


Fig. 6 Mechanical description of badger teeth loosening soil. (a) simplified illustration of a badger skull with the incisor colored in tan. Skull image is adapted from Zagrai et al. (2019). (b) penetration of incisors in soil is reliant on soil internal friction. Note that some rodents also use their claws to scratch and dig through soil (Samuels and Valkenburgh, 2009). Figure is adapted from Ma Y, Wang H, Zhuang J, Qi H and Yu J (2020) Effects of bionic curves on penetration force under difference soils. *Applied Sciences* 10: 529.

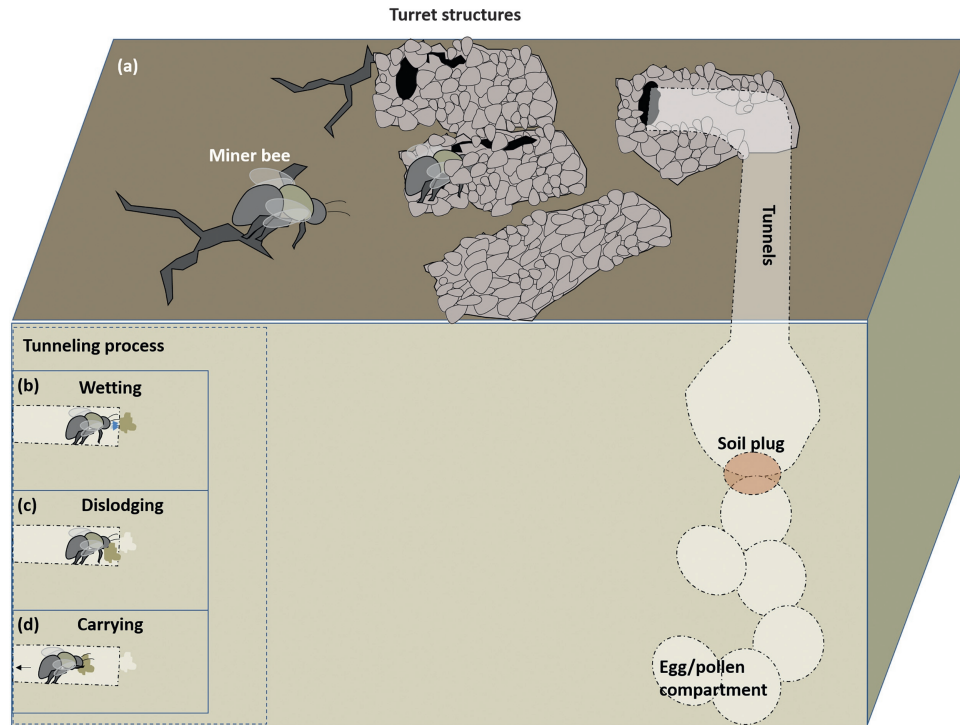


Fig. 7 Miner bee tunnel characteristics and process. (a) Miner bee tunnels have chimney like turret structures that extend above their tunnels. Their below ground tunnels consist of compartments for eggs and pollen that can be sealed off by a dense soil plug. The process of creating tunnels requires bees to wet the soil (b) and subsequently dislodge moistened soil with their mouths (c). They will then carry dislodged soil out with their front legs (d).

that store pollen and eggs (Rau, 1929). This allows the bees to keep their tunnels open without worrying about intruders or harsh weather conditions.

The formation of below ground nests is an intricate excavation process that requires the bees to work through dense clay. Bees will transport water in their gullet extracted from mud puddles up to 5 m away from their tunnels, with a round trip journey taking 15–30s. Water is used to moisten hard clays in order to reduce their strength and rigidity (Fig. 7b). Once sufficiently soft, bees can then remove a mouthful of soil (Fig. 7c). They will subsequently carry removed soil with their front legs and back out of their nest to the opening with their hind legs (Fig. 7d). The soil is used to build and reinforce their chimney structures (Rau, 1929). Few studies have focused on the interplay between bee habitats and soil structure (Lybrand et al., 2020). It is likely that the nesting tunnels provide preferential flow pathways for water and extend oxygenated zones in soil. Chemical analysis found soils impacted by miner bee communities were rich in lipid like compounds, which may act to stabilize the tunnels. This would also provide organic matter to be cycled into the soil.

Current studies have yet to propose mechanistic models for soil bioturbation by miner bees, in particular soil mechanical limits for the dislodging process that occurs during the moistening and removal of soil mass. The transport of soil particles may have certain similarities to other soil dwelling insects. Certain elements of these processes will be elaborated next on the section for bioturbation by ants and termites.

Bioturbation by ants and termites in tropical and arid regions

While earthworms and plant roots play a prominent role in soil structure formation in temperate regions, other organisms dominate bioturbation in tropical and arid regions. Richness maps indicate that these regions are dominated by ant and termite bioturbation resulting from their tunneling activities during nest or gallery formation (Fig. 8) (Tuma et al., 2020). Sampling biases in these regions limit the amount of data on their activity, which makes it difficult to ascertain the direct impact and importance of ants and termites. However, there are a few studies that have reported the impact that these soil bioturbation agents have on soil structure.

The influence of ants and termites on soil structure and associated properties is somewhat similar to earthworm burrows and root channels. Ant and termite galleries and nests enhance water infiltration and soil aeration. Furthermore, these galleries also provide paths of least resistance for root growth. The spatial extent that a single gallery may influence may range from 13 to 50 m in

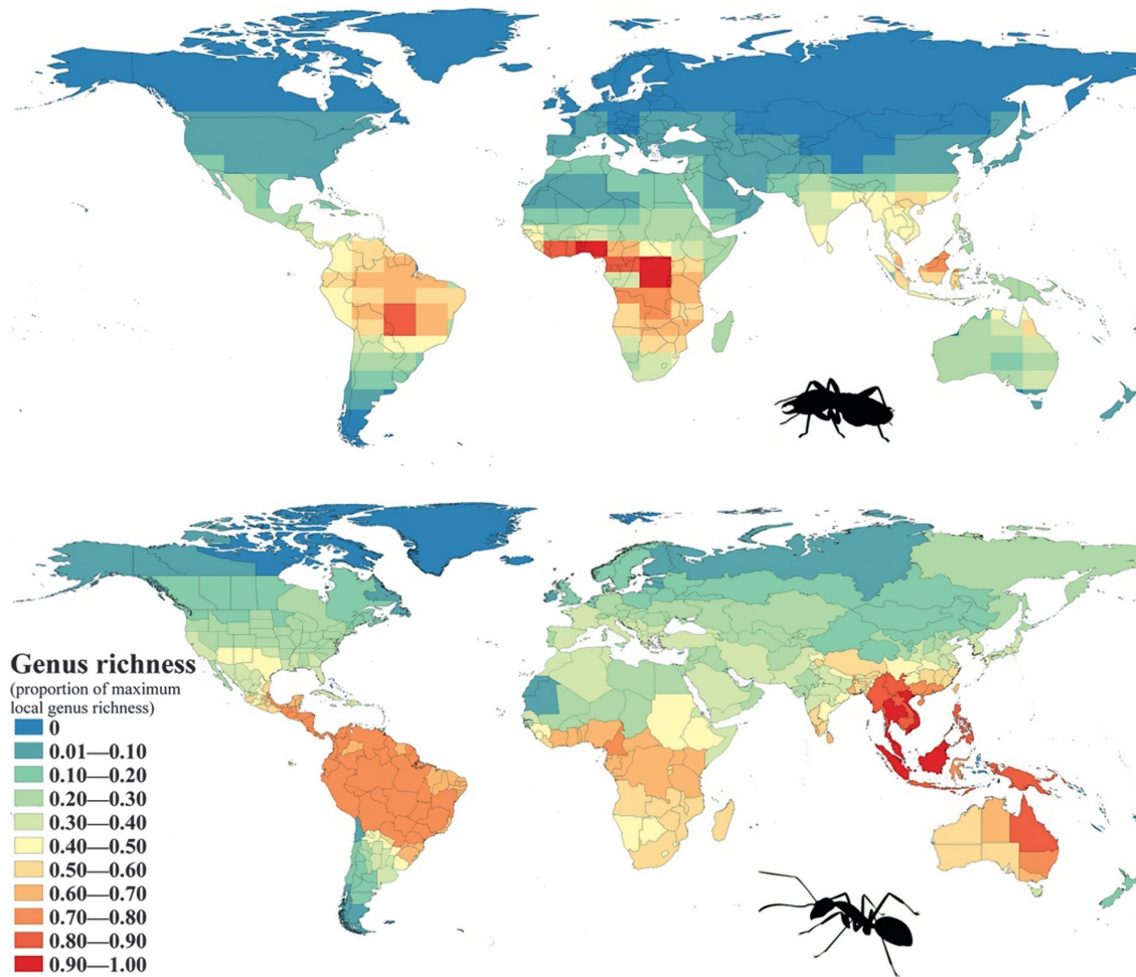


Fig. 8 Global richness maps of termites (top) and ants (bottom) replicated from Tuma et al. (2020).

radius from the location of a mound (De Bruyn and Conacher, 1990). While the formation of galleries may be similar, termites will incorporate organic constituents in their gallery formation. Mounds formed by termites act as a biological crust that protect the soil against water and wind erosion. Furthermore, this enhances soil aggregation by termites, which does not occur for ants (Jouquet et al., 2021). In the following, we outline models for termite mound physical functionality and the physical process involved in the construction of nests and galleries highlighting constraints to nest formation and rates of tunnel formation based on their excavation process.

Termite mounds physical functioning

Termites maintain favorable habitats in otherwise harsh warm conditions in arid and tropical regions by constructing large porous mound structures for controlling internal temperatures and humidity. Variations in termite mound morphologies can be attributed to heat and mass transport, and termites modify the mound shape in response to local odor concentrations (e.g., metabolic gases, pheromones, etc.) reflecting diffusion and airflow patterns. Temperatures in the mound are modulated by diurnal fluctuations and advection through the mounds caused by airflow. Models have been developed to understand these modulations based on physical principles. One model specifically considers three physical phenomena: temperature evolution based on the heat equation, with air flow regulated by Darcy's law, and odor transport based on diffusion-advection (Ocko et al., 2019). Steady state mound morphologies can be predicted based on physical non-dimensional parameters such as Peclet number, Biot number and relative mound thickness. Considering an average odor concentration at the mound walls of ϕ_c , a mound radius of R_m , and a mound height of h_m , then the surface area of the mound must be enough to balance the volumetric flux of odor generated inside the mound J (Fig. 9).

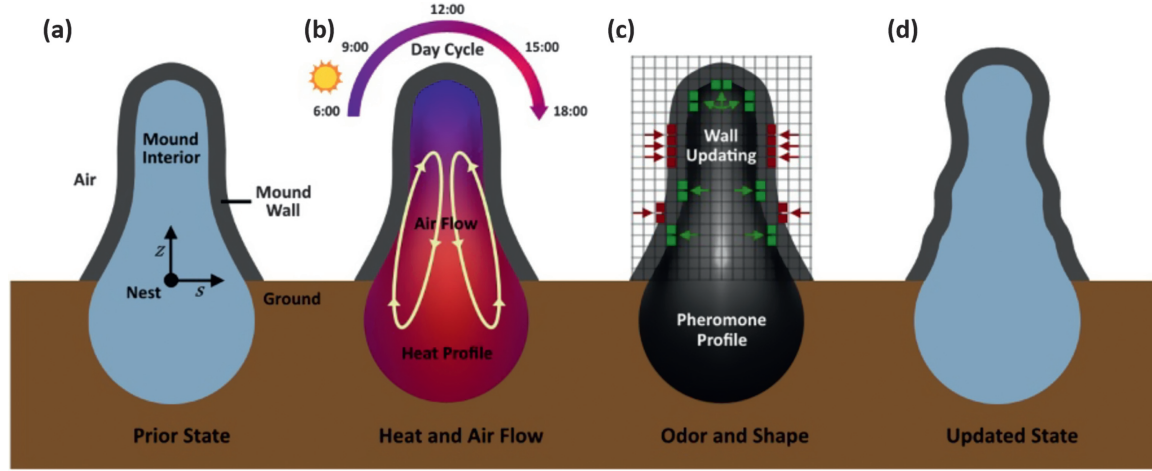


Fig. 9 Physical processes that drive morphological evolution of termite mounds as adapted from Ocko SA, Heyde A and Mahadevan L (2019) Morphogenesis of termite mounds. *Proceedings of the National Academy of Sciences* 116: 3379–3384. (a) Initial mound state and associated disjointed regions. (b) Response of heat profile and airflow under diurnal temperature fluctuations. (c) Odor profile which subsequently drives extension of the mound. (d) The updated mound morphology.

The flux can be given by

$$J = R_m^2 \phi_c D_{diff} / h_m, \quad (4)$$

where $D_{diff} [\text{m}^2 \text{s}^{-1}]$ is the diffusivity of the volatile odor. Thus, a mature mound at steady-state should have a radius described by:

$$R_c = \sqrt{\frac{J h_m}{D_{diff} \phi_c}} \quad (5)$$

The Peclet number (i.e., the ratio between advective and diffusive fluxes) is defined as

$$Pe = \frac{R_c (\kappa \alpha \rho g \Delta T)}{D_{diff} \eta} \quad (6)$$

where $\Delta T [\text{K}]$ is the diurnal temperature change, $\kappa [\text{m}^2]$ is the mound permeability, $\eta [\text{Pa s}]$ is the air viscosity, and $\alpha \rho g [\text{Pa m}^{-1} \text{K}^{-1}]$ is the buoyancy parameter. Similarly, the Biot number (i.e., ratio between thermal conduction and convection) is defined as

$$Bi = \frac{\sqrt{\tau D_T}}{R_c} \quad (7)$$

where $\tau [\text{s}]$ is the diurnal period and $D_T [\text{m}^2 \text{s}^{-1}]$ is the thermal diffusivity. Lastly, the relative thickness is given by

$$Th = \frac{h_m}{R_c} \quad (8)$$

Termites will adapt their mound's morphological trends and features such as sphericity and volume based on these non-dimensional values, thus presenting a physical rationale for mound designs (Fig. 10). While these mounds can explain how termites modulate the conditions in their habitats, other mechanisms associated with soil physical conditions may help to explain their global distributions.

While temperature and airflow play an important role in the termite nest structure, regulating moisture conditions is also critical for their habitats. A cross sectional view of a termite nest (Fig. 11a) elucidates several conduit structures and galleries that play a role in guiding soil moisture and vapor through the nest, regulating the nest humidity (Turner, 2006). These structures give rise to the moisture fluxes in and out of the nest (Fig. 11b) (Turner, 2006). Fluid is drawn inward from the adjacent soil based on differences in matric potential described by

$$J_{n,s} = K_h (\Psi_s - \Psi_n) \quad (9)$$

where $K_h [\text{m s}^{-1} \text{Pa}^{-1}]$ is the hydraulic conductance of the soil, $\Psi_s [\text{Pa}]$ is the soil water potential, and $\Psi_n [\text{Pa}]$ is the water potential of the nest. Evaporation of the nest is based on the difference between the nest and atmospheric humidity (based on partial pressures) given by

$$J_{n,a} = K_v (\Delta p H_2 O) \quad (10)$$

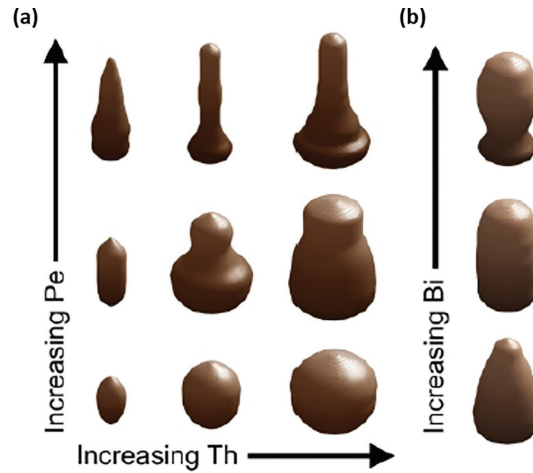


Fig. 10 Termite mound morphologies according to physical principles as adapted from Ocko SA, Heyde A and Mahadevan L (2019) Morphogenesis of termite mounds. *Proceedings of the National Academy of Sciences* 116: 3379–3384. (a) Termite mound absolute size increases with relative thickness (Th), but its sphericity reduces with increasing Peclet number (Pe). (b) Mounds are more bottom heavy for low Biot (Bi) numbers, cylindrical for medium Biot numbers and top heavy for high Biot numbers.

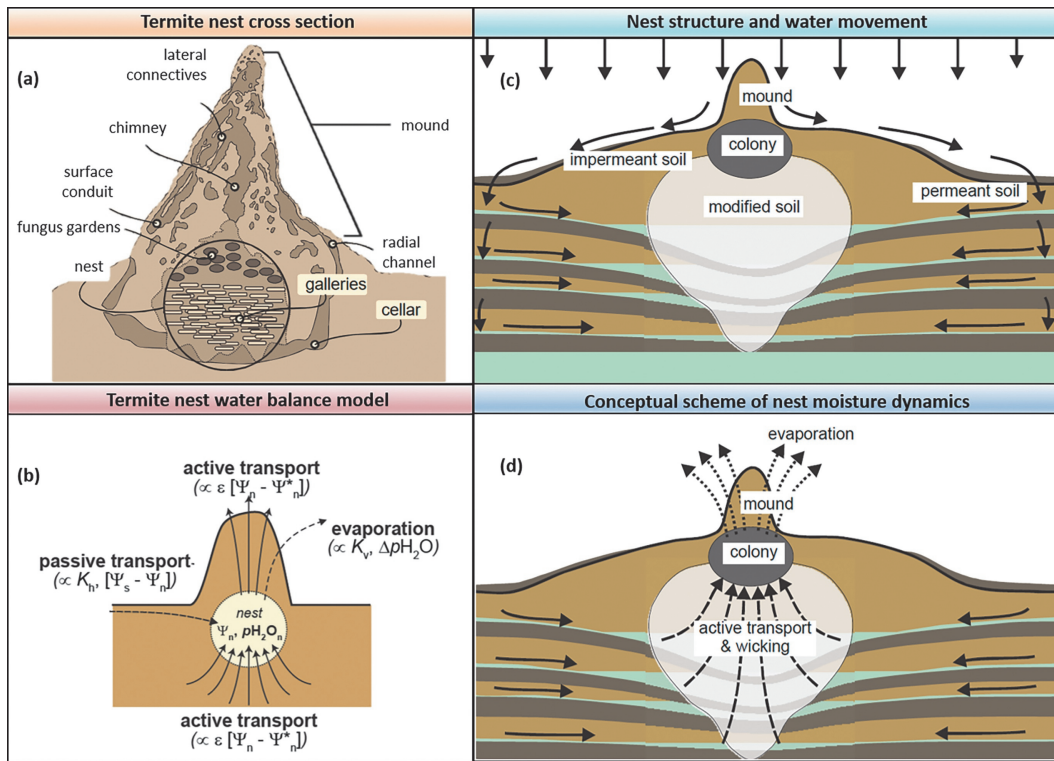


Fig. 11 Termite mound architecture and its role in regulating moisture as adapted from Turner (2006). (a) Termite nest and mounds consist of several conduit, chimney and channel structures that conduct and transport water in the form of liquid moisture in the soil pore space as well as water vapor. (b) These conduits serve as the basis for a model describing the water flux balance in the nest, considering the transport of water into the nest from adjacent soil, the transport of water out of the nest via evaporation, and the active transport of moisture in and out of the nest by targeted amendment of nest properties. (c) Conceptual functioning of the model considering precipitation above the mound and water permeating along adjacent permeable routes and (d) movement of moisture through the nest.

where K_v [$\text{m s}^{-1} \text{Pa}^{-1}$] is the mound's vapor conductance and Δp_{H_2O} [Pa] is the difference between the partial pressure of vapor in the nest and the atmosphere. Water can be further conveyed into the nest from the deep soil or onto the surface of the nest by the transport of damp soil, which helps achieve a target nest moisture content, Ψ_n^* . These fluxes are described and regulated by the expression:

$$J_{n,n^*} = \varepsilon(\Psi_n - \Psi_n^*) \quad (11)$$

where ε [$\text{m s}^{-1} \text{Pa}^{-1}$] is a regulating error signal that reflects the difference between the moisture of the nest Ψ_n [Pa] and the “target” nest moisture Ψ_n^* [Pa]. A conceptual description of the water movement around and subsequently through the nest can be seen in Fig. 11c and d, which illustrate the regulation process of soil moisture through the nest.

The construction of the fluid conduits in and around the mounds play a role in augmenting soil around the termite mounds themselves. This impacts the water cycle and local hydrology near the mound locations. Ultimately, the construction of the termite nests acts to enhance soil water retention.

Biophysical modeling of granular transport by ants and termites

Bioturbation by ants and termites displaces soil by carrying away discrete soil aggregates or sand grains of size D with their mandibles of size M (Fig. 12a) (Espinoza and Santamarina, 2010). Models describing how harvester ants manually transport soil based on a particle removal model showed that ants used different approaches to remove particles depending on grain size relative to their mandible width and soil water content (Fig. 12b). For fine textured soils, ants remove soil and form loose aggregates at low moisture conditions. When soil moisture is higher, ants are forced to pull fine particles in several successive grabs until they can form a stable aggregate. When aggregates reach a diameter of roughly 3 mm, ants are able to transport them (Espinoza and Santamarina, 2010).

In contrast to finer soil particles, dry coarse particles such as sand grains can be handled by ants in single grabs. However, the number of dry coarse soil particles that can be transported in a single grab is limited by friction and by capillarity forces. Modeling aggregation of particles is based primarily on consideration of capillarity and van der Waals forces. The weight of a transportable soil aggregate is affected by capillary forces between spherical particles with diameter d [m], expressed as:

$$F_c = \frac{\pi}{2} \gamma d \left(2 - \left(\frac{8}{9} G \theta_m \right)^{\frac{1}{4}} \right) \quad (12)$$

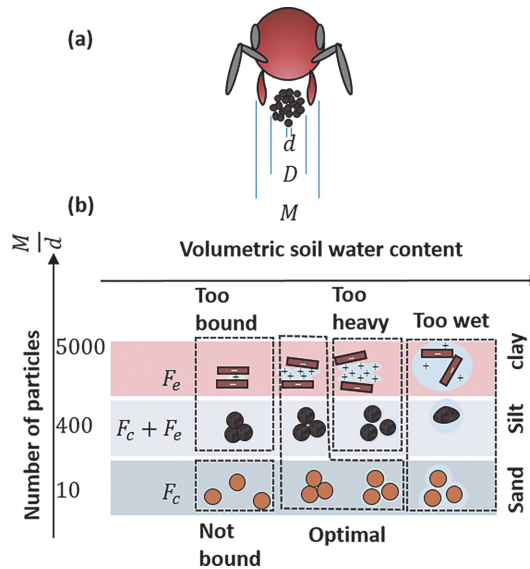


Fig. 12 Granular transport process of ants and termites adapted from Espinoza DN and Santamarina JC (2010) Ant tunneling—A granular media perspective. *Granular Matter* 12: 607–616. (a) Ants are constrained by a soil aggregate of a size D , as this cannot exceed their mandible size M . Aggregate sizes are made up of smaller particle sizes d , which are aggregated depending on their particle size. (b) Soil physical constraints depend on texture class, moisture content, electrical forces (F_e) and capillary forces (F_c). Overly dry soils rich in silt and clay may be too strongly bound for removal due to electrical forces. Dry sand may be removed, but it may require several passes to move multiple particles. Mildly wetted soils may have sufficient capillarity for optimal movement. Overly wet silt and clay may be too heavy to carry. If soil is too wet, then there might not be sufficient binding.

where γ [N m^{-1}] is the surface tension, G [$\text{kg}_{\text{grain}} \text{m}^3_{\text{water}} \text{kg}^{-1} \text{m}^{-3}_{\text{grain}}$] is the specific gravity of a mineral grain, and θ_m [$\text{kg}_{\text{water}} \text{kg}^{-1}_{\text{grain}}$] is the gravimetric water content (Espinoza and Santamarina, 2010). For small particles, van der Waals forces may become dominant, which are estimated as:

$$F_E = \frac{A_h}{24h^2}d \quad (13)$$

where A_h ($=0.83 \times 10^{-20}$ J) is the Hamaker constant for quartz-water-quartz, and h [m] is the inter-particle spacing between the two particles of size d [m] (Espinoza and Santamarina, 2010). Besides the mandible size constraint, ants are constrained by the weight of a particle that they can carry. The weight is given as:

$$F_w = \frac{\pi}{6}d^3Gg\rho_w \quad (14)$$

where g [$=9.81 \text{ m s}^{-2}$] is the gravitational acceleration, and ρ_w [$=1000 \text{ kg}_{\text{water}} \text{m}^{-3}_{\text{water}}$] is the density of water (Espinoza and Santamarina, 2010). The maximum weight an ant can lift is on the order of $3\text{--}6 \times 10^{-3}$ N, thus there is a balance between particle size and weight that is dependent on aggregate moisture content and particle sizes. This affects the tunneling rates of ants.

Considering the processes above, ants can construct biopores more easily under drier soil conditions (prevailing in arid regions), which is in stark contrast with conditions favored by earthworm bioturbation in wet soil. These details are consistent with findings that ants and termites are dominant bioturbation agents in semi-arid and arid regions, not only because earthworms are mechanically prohibited from these regions, but the physical process that ants and termites rely on is facilitated by drier conditions with coarse grained soil.

Conclusions

Bioturbation of soil arises from plant roots, insects and mammals. The underpinning penetration-cavity expansion bioturbation model associated with growing plant roots and tunneling earthworms (penetrators) provides a mechanistic model to describe the formation of biopore networks under different climatic conditions and soil types. This model can also be useful in outlining limitations for below ground activity by earthworms, explaining their reduced abundance in arid and tropical regions. Excavation of soil by ants and termites contributes to soil structure maintenance in arid and tropical regions where earthworm activity is limited. The environmental selection of ants and termites can be explained by biophysical models for their granular soil dislodging and transport activities, which are easier in dry conditions and coarse soils. Future studies may be able to extend these models and formulate similar constraint maps as those that exist for global earthworm distributions (Ruiz et al., 2021).

Excavation activity by larger rodents is complex in terms of short term (adverse) impacts on existing vegetation and long-term contributions to soil health. The available mechanistic and ecological models remain sketchy, and more work is needed to assess the effects of rodent bioturbation on ecosystem functioning. Understanding mechanistic biophysical principles associated with bioturbation processes can help quantify conditions that promote and restrict subterranean activity by different organisms. Empirical evidence suggests a significant impact of soil bioturbation on mechanical and transport properties of soil, with effects on ecosystem functioning and soil health. Therefore bioturbation should be incorporated in modern Earth system models of near surface processes. Recent studies highlight the importance of bioturbation in landscape formation processes (e.g., tree throw (Gabet et al., 2003)) and in soil alteration and production, but the modeling lags behind.

References

- Basirat M, Mousavi SM, Abbaszadeh S, Ebrahimi M, and Zarebanadkouki M (2019) The rhizosheath: A potential root trait helping plants to tolerate drought stress. *Plant and Soil* 445: 565–575.
- Beer C, Reichstein M, Ciais P, Farquhar G, and Papale D (2007) Mean annual GPP of Europe derived from its water balance. *Geophysical Research Letters* 34. <https://doi.org/10.1029/2006GL029006>.
- Beer C, Reichstein M, Tomelleri E, Ciais P, Jung M, Carvalhais N, Christian R, Arain MA, Baldocchi D, Bonan GB, et al. (2010) Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate. *Science* 329: 834–838.
- Bocek B (1986) Rodent ecology and burrowing behavior: Predicted effects on archaeological site formation. *American Antiquity* 51: 589–603.
- Buczkowski G and Bertelsmeier C (2017) Invasive termites in a changing climate: A global perspective. *Ecology and Evolution* 7: 974–985.
- Colombi T, Braun S, Keller T, and Walter A (2017) Artificial macropores attract crop roots and enhance plant productivity on compacted soils. *Science of the Total Environment* 574: 1283–1293.
- De Bruyn LL and Conacher AJ (1990) The role of termites and ants in soil modification—A review. *Soil Research* 28: 55–93.
- Dorgan KM (2010) Environmental constraints on the mechanics of crawling and burrowing using hydrostatic skeletons. *Experimental Mechanics* 50: 1373–1381.
- Dorgan KM (2015) The biomechanics of burrowing and boring. *The Journal of Experimental Biology* 218: 176–183.
- Dumais J and Forterre YEL (2012) “Vegetable dynamics”: The role of water in plant movements. *Annual Review of Fluid Mechanics* 44: 453–478.
- Espinoza DN and Santamarina JC (2010) Ant tunneling—A granular media perspective. *Granular Matter* 12: 607–616.
- Gabet EJ, Reichman O, and Seabloom EW (2003) The effects of bioturbation on soil processes and sediment transport. *Annual Review of Earth and Planetary Sciences* 31: 249–273.
- Gibson LJ (2012) The hierarchical structure and mechanics of plant materials. *Journal of the Royal Society Interface* 9: 2749–2766.
- Grant W, French N, Folse JR, and L. (1980) Effects of pocket gopher mounds on plant production in shortgrass prairie ecosystems. *The Southwestern Naturalist* 25: 215–224.
- Greacen E and Oh J (1972) Physics of root growth. *Nature* 235: 24–25.
- Grinnell J (2020) *The Burrowing Rodents of California as Agents in Soil Formation*. University of California Press.

- Hettiaratchi D, Goss M, Harris J, Nye P, and Smith K (1990) Soil Compaction and Plant Root Growth [and Discussion]. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 329: 343–355.
- Hickman GC (1977) Burrow system structure of *Pappogeomys castanops* (Geomysidae) in Lubbock County, Texas. *American Midland Naturalist* 97: 50–58.
- Jouquet P, Henry-des-Tureaux T, Bouet C, Labiadh M, Caquineau S, Boukbida HA, Ibarra FG, Hervé V, Bultelle A, and Podwojewski P (2021) Bioturbation and soil resistance to wind erosion in Southern Tunisia. *Geoderma* 403: 115198.
- Kuzyakov Y and Blagodatskaya E (2015) Microbial hotspots and hot moments in soil: Concept & review. *Soil Biology and Biochemistry* 83: 184–199.
- Lavelle P, Barot S, Blouin M, Decaëns T, Jimenez JJ, and Jouquet P (2007) Earthworms as key actors in self-organized soil systems. *Theoretical Ecology Series* 4: 77–106.
- Lybrand RA, Fedenko J, Tfaily M, and Rao S (2020) Soil properties and biochemical composition of ground-dwelling bee nests in agricultural settings. *Soil Science Society of America Journal* 84: 1139–1152.
- Ma Y, Wang H, Zhuang J, Qi H, and Yu J (2020) Effects of bionic curves on penetration force under difference soils. *Applied Sciences* 10: 529.
- Misra R, Dexter A, and Alston A (1986) Maximum axial and radial growth pressures of plant roots. *Plant and Soil* 95: 315–326.
- Ocko SA, Heyde A, and Mahadevan L (2019) Morphogenesis of termite mounds. *Proceedings of the National Academy of Sciences* 116: 3379–3384.
- Or D, Keller T, and Schlesinger WH (2021) Natural and managed soil structure: On the fragile scaffolding for soil functioning. *Soil and Tillage Research* 208: 104912.
- Palm J, Loes N, Van Schaik MB, and Schröder B (2013) Modelling distribution patterns of anecic, epigeic and endogeic earthworms at catchment-scale in agro-ecosystems. *Pedobiologia* 56: 23–31.
- Pritchard J (1994) The control of cell expansion in roots. *New Phytologist* 127: 3–26.
- Purves WK, Sadava D, Orians GH, and Heller HC (2003) *Life: The Science of Biology: Volume III: Plants and Animals*. Macmillan.
- Rau P (1929) The biology and behavior of mining bees, *Anthophora abrupta* and *Entechnia taurea*. *Psyche* 36: 155–181.
- Rosario Carotenuto A, Guarracino F, Šumbera R, and Fraldi M (2020) Burrowing below ground: Interaction between soil mechanics and evolution of subterranean mammals. *Journal of the Royal Society Interface* 17: 20190521.
- Ruiz SA and Or D (2018) Biomechanical limits to soil penetration by earthworms: Direct measurements of hydroskeletal pressures and peristaltic motions. *Journal of the Royal Society Interface* 15: 20180127.
- Ruiz S, Schymanski S, and Or D (2017) Mechanics and energetics of soil penetration by earthworms and plant roots—Higher burrowing rates cost more. *Vadose Zone Journal*. <https://doi.org/10.2136/vzj2017.01.0021>.
- Ruiz S, Koebernick N, Duncan S, Fletcher DM, Scotson C, Boghi A, Marin M, Bengough AG, George T, and Brown L (2019) Significance of root hairs at the field scale—modelling root water and phosphorus uptake under different field conditions. *Plant and Soil* 447(8): 1–24.
- Ruiz SA, Bickel S, and Or D (2021) Global earthworm distribution and activity windows based on soil hydromechanical constraints. *Communications Biology* 4: 1–9.
- Samuels JX and Valkenburgh BV (2009) Craniodental adaptations for digging in extinct burrowing beavers. *Journal of Vertebrate Paleontology* 29: 254–268.
- Tuma J, Eggleton P, and Fayle TM (2020) Ant-termite interactions: An important but under-explored ecological linkage. *Biological Reviews* 95: 555–572.
- Turner JS (2006) Termites as mediators of the water economy of arid savanna ecosystems. In: *Dryland Ecohydrology*. Springer.
- Van Groenigen JW, Lubbers IM, Vos HM, Brown GG, De Deyn GB, and Van Groenigen KJ (2014) Earthworms increase plant production: A meta-analysis. *Scientific Reports* 4: 6365.
- Vleck D (1981) Burrow structure and foraging costs in the fossorial rodent, *Thomomys bottae*. *Oecologia* 49: 391–396.
- Zagrai G, Șeicaru A, Lixandru D, and Zagrai Măierean A (2019) Morphology of the skulls in badger (*Meles Meles*) and otter (*Lutra Lutra*)—Comparative aspects. *Lucrari Stiintifice-Universitatea de Stiinte Agricole a Banatului Timisoara, Medicina Veterinara* 52: 125–130.

The reinforcement of soil by plant roots

Alexia Stokes^a and Anil Yildiz^b, ^aUniversity Montpellier, AMAP, INRAE, CIRAD, CNRS, IRD, Montpellier, France; ^bMethods for Model-based Development in Computational Engineering, RWTH Aachen University, Aachen, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	115
Soil shear strength	115
Direct shear tests	116
Triaxial testing	117
How vegetation impacts the hydraulic properties of soil	118
How plant roots reinforce soil mechanically	118
Outlook	121
References	122

Abstract

The biophysical interactions between soil and vegetation are discussed from a mechanical and hydrological perspective. How these interactions impact the stability of soil on slopes and protect against shallow landslides is presented, along with a brief outline of models used to calculate the reinforcing effects of roots on soil, and how these models have been derived in the last half-century. Methods are discussed for measuring root-permeated soil shear strength in the field and laboratory. Future challenges for research are outlined, and include the difficulties encountered when integrating large roots and the architecture of complex root systems into mechanical models, as well as scaling up to the landscape level.

Key points

- Plants with numerous, strong roots increase soil shear strength
- Direct soil-root shear tests quantify reinforcement, but need a small normal force
- Plant species mixtures enhance both soil mechanical and hydraulic properties.
- Root system architecture should be better integrated in mechanical models

Introduction

Plant roots reinforce soil against mass movement of substrate through a mixture of physical and chemical processes. These processes are controlled by the functioning of plants and microbes that are impacted by diverse root traits, as well as by environmental factors, including soil type, soil water content, soil cohesion and pH. Soil reinforcement mechanisms have been extensively reviewed in the last two decades, mostly from the geotechnical perspective of improving slope stability against shallow landslides (Giadrossich et al., 2017; Stokes et al., 2009; Fig. 1), but also with the aim of understanding fundamental processes related to ecosystem functioning (Freschet et al., 2021).

Most studies on plant root systems and soil reinforcement have focused on measurements of soil shear strength at near-saturation levels, root mechanical properties and root area ratio (RAR), the fraction of a plane of soil occupied by roots. These data are then used in models that calculate the contribution of vegetation to slope stability. Nearly all studies find that plant species, which have roots that are strong in tension and a large RAR, increase soil shear strength (Freschet et al., 2021). This knowledge can be used in defining planting and management schemes applicable to natural and artificially constructed slopes, a practice that originated 4000 years ago for the management of flood waters in China, and is now common practice in many countries worldwide.

In this chapter, we focus on the fundamental mechanisms driving soil reinforcement and how vegetation shapes these processes. We explain the factors that govern soil shear strength and how plant root systems physically interact with soil to alter shear strength. We review the most common methods for measuring soil and root physical properties related to soil reinforcement, and conclude with an update on the most recent models used to estimate the role of vegetation in stabilizing slopes.

Soil shear strength

Soil shear strength (τ) is defined as the resistance to deformation by the action of tangential (shear) stress. Soil shear strength is made up of cohesion between particles and resistance to particles sliding over each other due to friction or interlocking. Cohesion is composed of true cohesion and apparent cohesion. True cohesion is a function of soil mineralogy and results from chemical bonds between particles. Apparent cohesion, however, is determined by water tension within the soil and is strongly influenced by water



Fig. 1 A shallow landslide on a vegetated slope. Photograph is kind courtesy of Dr. Frank Graf.

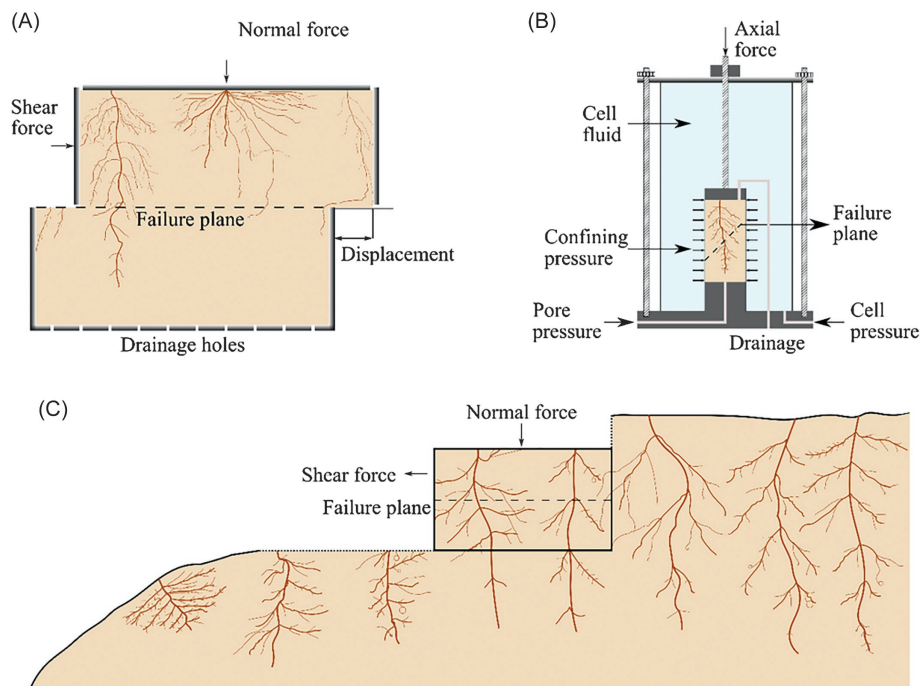


Fig. 2 Methods to determine the shear strength of root-permeated soils: (A) direct shear test, (B) triaxial tests, and (C) in situ direct shear tests. Illustrations are not drawn to scale.

content (Flerchinger et al., 2005). Additional cohesion from plant roots can also enhance soil shear strength as roots are stronger in tension than soil and add reinforcement as soil is deformed during shear. Soil shear strength and the contribution of roots to shear strength are usually measured via direct shear tests (Fig. 2A) and triaxial tests (Fig. 2B).

Direct shear tests

Direct shear tests are commonly employed in geotechnical engineering practice and research to determine the shear strength of a soil. The test is generally performed in a container consisting of two halves, where one half is moved relative to the other (Fig. 2A). The

interface between two boxes acts as a failure plane, where the sample is sheared - not at its weakest plane. A normal force is applied at the surface of the sample to induce a normal stress (σ_n) at the failure plane, and a horizontal force is applied to one half of the box.

A direct shear test can be performed as either displacement-controlled (i.e., the container is pushed at a constant displacement rate while measuring the force required) or stress-controlled (i.e., the force is increased at constant increments and displacement is recorded). ISO 17892-10 (International Organization for Standardization, 2018) or ASTM D 3080 (ASTM International, 2012) standards provide guidelines for conventional direct shear testing. Since the pioneering studies on soil reinforcement by Wu et al. (1979) or Waldron and Dakessian (1982), direct shear testing is one of the most frequently used methods to measure the shear strength of root-permeated soils. Tests can be performed either in situ (or in the laboratory (Mickovski and van Beek, 2009; Ghestem et al., 2014; Yildiz et al., 2019)). In situ tests are the most realistic in terms of field conditions (Fig. 2C): roots are fully embedded in undisturbed soil and anchored below the potential shear surface, and soil can be saturated before the test begins. However, in situ tests are time consuming and are usually performed on a flat ground surface because of complications with testing on a slope. Additionally, as soil and root system characteristics vary spatially and temporally, a large number of measurements must be performed across a site and at different points in time, to assess which areas are the most susceptible to shallow landslides. To overcome these difficulties, Meijer et al. (2019a) adapted a garden corkscrew weeder so that peak soil shear strength could be measured rapidly in the field. This method measures the resistance against pull-out extraction of a cylinder of (rooted) soil caught within the screw and can be used on steep slopes.

The aforementioned standards define limitations on maximum particle size of a soil that can be tested in direct shear testing. ASTM D-3080 limits the maximum particle size to one tenth of the shear box width and one sixth of the specimen height. A standard commercially available direct shear apparatus (mostly capable of testing square samples of 60×60 mm or circular samples of 60 mm in diameter) cannot accommodate soils with larger-size particles for testing. As vegetated soils are generally procured from field sites, they can contain gravel-sized particles. Furthermore, the root diameters to be tested in a direct shear apparatus can also exceed the size limitations of the standards. In laboratory studies, due to these size limitations and the importance of monitoring hydrological aspects during testing, many studies investigating the shear strength of root-permeated soils are conducted with non-conventional testing devices (Yildiz et al., 2019).

Root-permeated samples to be tested in a laboratory direct shear apparatus are generally prepared in two ways. The first approach is to prepare a reconstituted specimen that has segments of roots taken from a living plant inserted into the shear box. This approach enables the control and measurement of parameters related to root morphology, such as diameter, angle, length and tortuosity. However, it is clearly not representative of the natural growth state of roots in undisturbed soil. The second technique is to grow plants in a shear box either under controlled, e.g., in a growth chamber or glass house (Yildiz et al., 2019), or monitored conditions, e.g., in the field with continuous monitoring of meteorological parameters (Veylon et al., 2015), (see Table 4 in Freschet et al. (2021) for a review of root system characteristics and shear strength of root-permeated soil, in comparative studies with several plant species per study). Much more realistic conditions prevail, especially with regard to root anchorage in soil and the root-soil interface; however there is minimal information about the number of roots, root area ratio, or any other root characteristics before the test commences.

Direct shear tests are performed under the application of a load perpendicular to the shearing plane. Tests are generally performed for at least three levels of normal load to be able to calculate shear strength parameters. Considering the typical depth of a shallow landslide, i.e., 1.0–2.0 m, (Norris et al., 2008), the normal load applied during direct shear testing of root-permeated soils should be comparable to the overburden pressure that the soil experiences at shallow depths. Therefore, in comparison to conventional shear tests applied in geotechnical engineering, generally much smaller pressures should be employed, taking into account only the weight of soil above the potential shear plane and the weight of vegetation (Veylon et al., 2015).

Triaxial testing

An alternative method to determine the shear strength of soil is via triaxial testing (Fig. 2C). These tests are performed on cylindrical specimens, usually with a height to diameter ratio of 2:1. The specimen is contained in a cell, which is filled generally with water, and is subjected to an all-around radial pressure, i.e., cell pressure. An axial load is applied on the specimen until failure is observed, while the deformations are measured (Terzaghi et al., 1996). Either consolidated (i.e., excess pore water pressure in the sample generated due to loading is dissipated) or unconsolidated (i.e., sample is tested before the dissipation of the excess pore water pressure) samples can be tested under drained or undrained conditions. Four main stress paths, namely axial compression, axial extension, lateral compression, and lateral extension can be applied on a sample. Triaxial testing has various advantages over direct shear testing, which can be summarized briefly as follows:

- Sample fails at the weakest plane,
- Easier to monitor deformations, and calculate strains,
- More realistic representations of stress state under field conditions can be generated,
- Samples can be tested under various stress paths (e.g., extension or compression) and conditions (e.g., drained and undrained).

Triaxial testing on root-permeated soils is much rarer than direct shear tests mainly due to the complexity of the procedure. However, it is sometimes inevitable as the nature of the problem and the conditions identified in the field dictate the use of triaxial testing on root-permeated soils. For example, Foresta et al. (2020) investigated the volume collapse of pyroclastic soils in Southern

Table 1 Details of studies using triaxial testing on root-permeated soils.

Test	Plant species	Soil Sample Size: Height, Diameter [mm]	Confining pressure [kPa]	References
Consolidated Undrained	<i>Alnus incana</i>	140.0, 70.0	50, 75, 100	Zhang et al. (2010)
Consolidated Drained	<i>Robinia pseudoacacia</i>	80.0, 39.1	100, 200, 300, 400	
Unconsolidated Undrained	<i>Atriplex canescens</i> , <i>Caragana korshinskii</i> , <i>Zygophyllum xanthoxylon</i> , <i>Nitraria tangutorum</i> , <i>Lycium chinense</i>	125.0, 61.8	10, 20, 30, 40	Hu et al. (2013)
Consolidated Drained and Consolidated Undrained	Poaceae	79.3, 36.4	10, 30, 50	Foresta et al. (2020)

Italy under fully saturated undrained conditions, and so performed consolidated drained and undrained triaxial tests on samples planted with grasses, to assess the effects of roots on the occurrence of static liquefaction. Only a few similar studies exist that have used triaxial testing (see Table 1), and they follow the 2:1 aspect ratio of cylinder height to diameter suggested by various standards. A minimal number of studies tested soils at confining pressures comparable to the stress state of a soil element at a shallow depth.

How vegetation impacts the hydraulic properties of soil

In a soil with a large water content, the pore water pressure can increase and reduce soil shear strength. Pore water pressure is the pressure (isotropic normal force per unit area) exerted by the fluid phase, where pores are filled or partially filled with water and is averaged over a representative elementary volume of soil containing many pores, rather than an individual pore. If vegetation is present and actively growing (i.e., not in seasonal dormancy), water can be removed via evapotranspiration, i.e., the combination of evaporation from the soil and plant transpiration. The forces between air, water, and soil particles lead to changes in matric suction. Matric suction, or negative pore water pressure, is the free energy change in a unit volume of water, when isothermally transferred from the soil water state to the free water state and is defined at the soil-water-air representative elementary volume. Matric suction is therefore the pressure dry soil exerts on the surrounding material to equalize the moisture content in the overall block of soil and it is estimated as the difference between pore air pressure (u_a) and pore water pressure (u_w). Negative pore water pressures produce matric suction and greater shearing resistance. This process is termed hydrological or hydric cohesion.

When soil is saturated on a vegetated site, positive pore-water pressures reduce soil shear strength and can also produce uplift pressure (u_z). Hydrostatic uplift (also known as uplift pressure), is an upward pressure applied to a structure (such as vegetation), that has the potential to raise it relative to its surroundings. Hydrostatic uplift can be problematic if it is greater than the weight of the structure (e.g., plant mass) exerted downwards. The downward weight of the plant is therefore reduced, and mechanical integrity of the soil and vegetation can be compromised (Liu et al., 2021).

Vegetation also causes changes in soil hydrological conditions through the interception of precipitation by the canopy, and channelling of rainfall down stems and along roots. Furthermore, as roots grow through soil, they create channels - also termed biotic macropores (Ghestem et al., 2011) - along which water can flow (Fig. 3). In addition to forming biotic macropores, plant roots are major drivers of soil structure formation through aggregation at the root-soil interface. This phenomenon occurs due to hydrological stresses from root-water uptake, mechanical deformation from root growth and the secretion of organic compounds that provide substrate for microorganisms and actively bind soil particles. Mycorrhizal fungi are particularly important soil organisms that are associated with roots, soil aggregation and stability. Additionally, plant roots can promote slope drainage by creating macropores and functioning as hillslope-scale preferential flow paths that by draining subsurface water away from potentially unstable sites reduces pore-water pressure. However, when root channels converge or when subsurface flow abruptly terminates in the slope (e.g., if roots cannot penetrate bedrock and growth accumulates over the bedrock surface), water pressure may concentrate in critical zones, thus promoting localized zones of instability (Ghestem et al., 2011). The role of coarse roots (with a diameter > 2 mm) on preferential flow is therefore highly variable, whereas finer roots (with a diameter < 2 mm) have a more consistent effect on infiltration processes via matrix flow, i.e., uniform flow through fine pores. However, many fine roots can also block pores and greatly reduce hydraulic flow. It is therefore advisable to promote species mixtures on a vegetated slope, so that a range of root branching densities (Freschet et al., 2021) and pore sizes throughout the soil profile will enhance infiltration and drainage processes (Nespoulous et al., 2019).

How plant roots reinforce soil mechanically

Mohr-Coulomb theory is a mathematical model used to define the shear strength (τ) of soils at different effective stresses. The generalized form of the Mohr-Coulomb failure criterion using effective stress parameters can be expressed as follows (Lu and Likos, 2006):

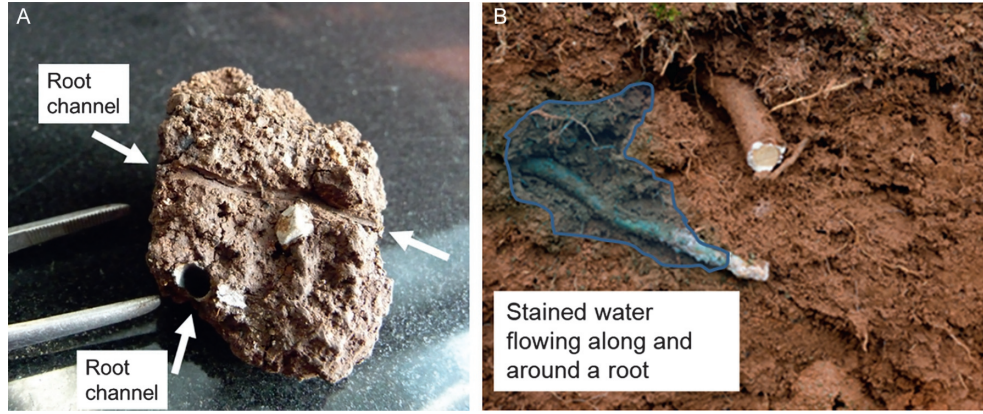


Fig. 3 (A) Macropores (channels formed after roots have died and decayed) in a silty loam and (B) water stained with Brilliant Blue FCF dye flowing along and around a live root, but next to a root where no stained water was observed, demonstrating the highly variable nature of preferential flow processes around coarse roots on vegetated slopes. (A) Reproduced with kind permission from Ghestem M, Sidle RC, and Stokes A (2011) The influence of plant root systems on subsurface flow: Implications for slope stability. *Bioscience* 61: 869–879; (B) reproduced with kind permission from Nespoulous J, Merino-Martin L, Monnier Y, Bouchet DC, Ramel M, Dombey R, Viennois G, Mao Z, Zhang J-L, Cao K-F, le Bissonnais Y, Sidle RC, and Stokes A (2019) Tropical forest structure and understorey determine subsurface flow through biopores formed by plant roots. *Catena* 181: 104061.

$$\tau = c' + (\sigma - u_a) \tan \phi' + (u_a - u_w) \tan \phi^b \quad (1)$$

where c' is cohesion, σ is the normal stress, u_a is the pore air pressure, u_w is the pore water pressure, ϕ' is the internal friction angle, and ϕ^b is an additional friction angle to express the contribution of matric suction to shear strength in the partially saturated state. When soil is saturated, Eq. (1) simplifies to:

$$\tau_s = c' + (\sigma - u_w) \tan \phi' \quad (2)$$

In the 1970s, pioneering modelling approaches incorporated the action of plant roots into Mohr-Coulomb theory (Waldron, 1977; Wu et al., 1979), and they were based on the equilibrium of forces acting on a soil element permeated with roots. One of the key assumptions of such models is the full mobilization of root tensile strength as the root-soil is sheared and deformed (Fig. 4A).

As the soil is sheared, roots are stretched and pulled out of the soil, and subjected to a tensile force that can be divided into two components, acting in the tangential and normal directions. Long, well-anchored roots will increase pull-out resistance up to a critical root length, above which roots will break in tension instead of slipping out of the soil. Roots can also slip out of soil if the tensile stress in the root is greater than the bonding resistance between the soil and the root (Fan, 2012). However, the root failure type (breakage or slippage) was not considered in these pioneering models. The shear strength of the root-permeated soil (τ_r), can be expressed as follows:

$$\tau_r = c' + (\sigma' + t_r \cos \theta) \phi' + t_r \sin \theta \quad (3)$$

where θ is the angle between the vertical axis and the root after deformation, and t_r is the tensile stress on the root (Eq. 6). By substituting the shear strength of the soil at the saturated state, τ_s , into Eq. (3), we can define τ_r as:

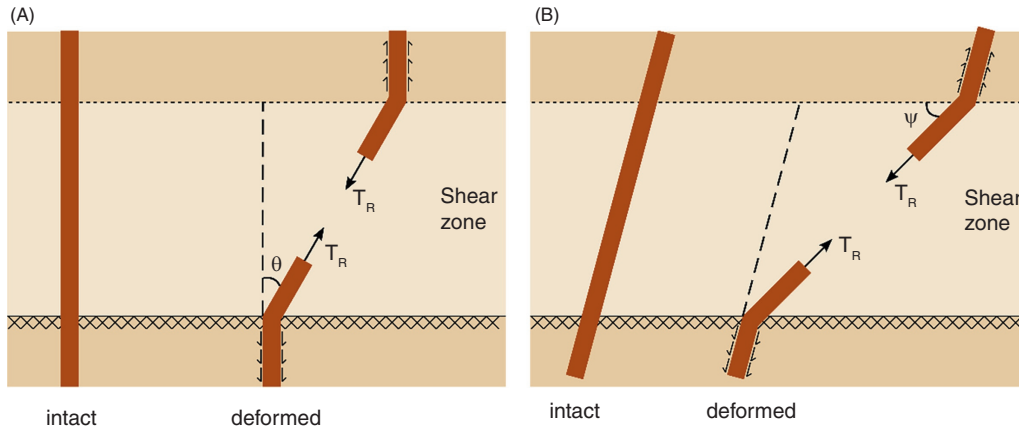


Fig. 4 Intact and deformed shape of roots crossing a shear plane during deformation. an initially (A) perpendicular and (B) inclined root segment. Modified from Gray DH and Ohashi H (1983) Mechanics of fiber reinforcement in sand. *Journal of Geotechnical Engineering* 109: 335–353.

$$\tau_r = \tau_s + t_r(\sin\theta + \cos\theta\phi') \quad (4)$$

If an initially inclined root is deformed until an angle ψ is reached (See Fig. 4B), Eq. (5) becomes (Gray and Ohashi, 1983):

$$\tau_r = \tau_s + t_r[\sin(90 - \psi) + \cos(90 - \psi)\phi'] \quad (5)$$

Considering multiple perpendicular roots of various diameters passing through a potential failure plane, t_r can be calculated as:

$$t_r = \sum_{i=1}^n T_r^i \frac{A_r^i}{A} \quad (6)$$

where T_r^i is the tensile strength of the roots of diameter class i , A_r^i is the cross-sectional area of the roots of diameter class i , and A is the area of the shear plane. Root tensile strength is calculated as the force required per cross-sectional area of the root to cause failure, either through breakage or non-reversible deformation (see Giadrossich et al. (2017) for methods to measure tensile strength). Meijer (2021) proposed the use of a cumulative and continuous root diameter distribution instead of diameter classes, as information is lost concerning their real diameters, if roots are sorted into separate diameter classes. Wu et al. (1979) argued that the additional strength due to roots has the “characteristics of cohesion,” and the term “additional cohesion due to roots” (c_r) was coined, which can be generalized in Eq. (7).

$$c_r = k' T_r RAR \quad (7)$$

where the ratio A_r/A in Eq. (6) is denoted as RAR (root area ratio). Wu et al. (1979) suggested that the term $(\sin\theta + \cos\theta\phi')$ in Eq. (5) is insensitive to a range of values for root angle (θ , Fig. 4A) and effective internal friction angle of the soil (ϕ'). Also, the effects of roots on soil internal friction angle are variable and poorly understood (Veylon et al., 2015). Therefore, Wu et al. (1979) provided an average value of 1.2 as k' to substitute in Eq. (7). However, subsequent studies have shown that substituting 1.2 for k' in Eq. (7) generally overestimates the additional cohesion from roots, and alternatives to k' have since been suggested (see Mao (2022) for a review). One source of this overestimation is the mobilization of tensile strength of the roots and shear strength of soil. Fannin et al. (2005) demonstrated one isolated in situ direct shear test displaying two peaks in the shear stress - strain behavior and attributed the second peak to the presence of roots. Pollen et al. (2004) also noted that peak shear stress of soil and tensile stress of roots occurred at different displacements, and root strength may not be fully mobilized before failure. Therefore, Cohen et al. (2011) suggested that a simple addition of the strength of both components of the root-soil matrix would result in an overestimation.

As a result of these investigations, a new type of model was developed that accounts for variability in root diameter and tensile strength of a bundle of roots crossing a shear plane, where root failure is more likely to be progressive than simultaneous and roots can either break or slip out of soil during failure (Waldron and Dakessian, 1982). To simulate this progressive root failure, Pollen et al. (2004) examined an algorithm used in materials science called the Fiber Bundle Model (FBM). FBMs consider the sharing of load or energy over a bundle of roots that differ in size and material properties and a range of FBM models are now available for estimating c_r (Fig. 5 (Cohen et al., 2009), see Mao (2022) for a review). The order of progressive root failure can now be modelled by determining the critical displacement at which roots slip or break (Cohen et al., 2011), thereby simulating more effectively the failure of the soil-root matrix in shear tests (Fannin et al., 2005).

To correct for the over estimation of c_r in the model by Wu et al. (1979), a factor, k'' was suggested by Preti (2006), and is the ratio of c_r estimated by a FBM to that estimated in the Wu et al. (1979) model, based on the same root bundle. Several studies used a constant for k'' (or as a product of k'' and k') whereas others quantified k'' based on field data (Bischetti et al., 2009). Meijer (2021) simulated the variation of k'' as a function of load-sharing laws, root tensile strength and diameter variation, therefore enabling FBM users to quickly select a value for k'' depending on their needs. Typical values of c_r vary from 1 to 25 kPa depending on the type of soil and vegetation, and assumptions made in model development (see Table 5-2 in Norris et al., 2008, for typical values of c_r). Although values >100 kPa can be found in the literature, they are often derived from measurements close to the soil surface, where it is unlikely that soil failure will occur in shear.

Although much research has focused on the contribution of plant roots to soil cohesion, studies have usually focused on the inclusion of thin roots (< 10 mm in diameter) that are mostly subjected to tensile loading and so can be incorporated into the modelling framework described above. However, the impact of larger and stiffer roots that act in bending and compression has rarely been considered, although studies are now beginning to stress the importance of introducing these elements into slope stability models (Giadrossich et al., 2019). More detailed investigations into the effect of large, structural roots on slope stability, and the way root systems are designed in terms of architecture (Ghestem et al., 2014; Meijer et al., 2019b), are needed to close this knowledge gap. Also, the geometrical design of a root system, for example, branching intensity, will impact soil reinforcement as the presence of branches counteracts potential slippage between the soil and roots, allowing the mobilization of larger axial forces within the root (Meijer et al., 2019b). One way forward is through the use of scaled-down models of 3D root system architecture in geotechnical centrifuges (Zhang et al., 2020). Geotechnical centrifuges allow the testing of materials such as soil and rock, that have non-linear mechanical properties that depend on the effective confining stress and stress history. The centrifuge increases g-forces on the physical model, to produce identical self-weight stresses in the model and prototype. Zhang et al. (2020) successfully applied this technique to scaled down (1:20) physical models of tree root architecture (using real data and obtained via 3D printing). Through a series of tests in normal and elevated gravity conditions, Zhang et al. (2020) were able to determine the impact of soil moisture and root system architecture on the resistance of models to being pushed over laterally (considered as a proxy of loading

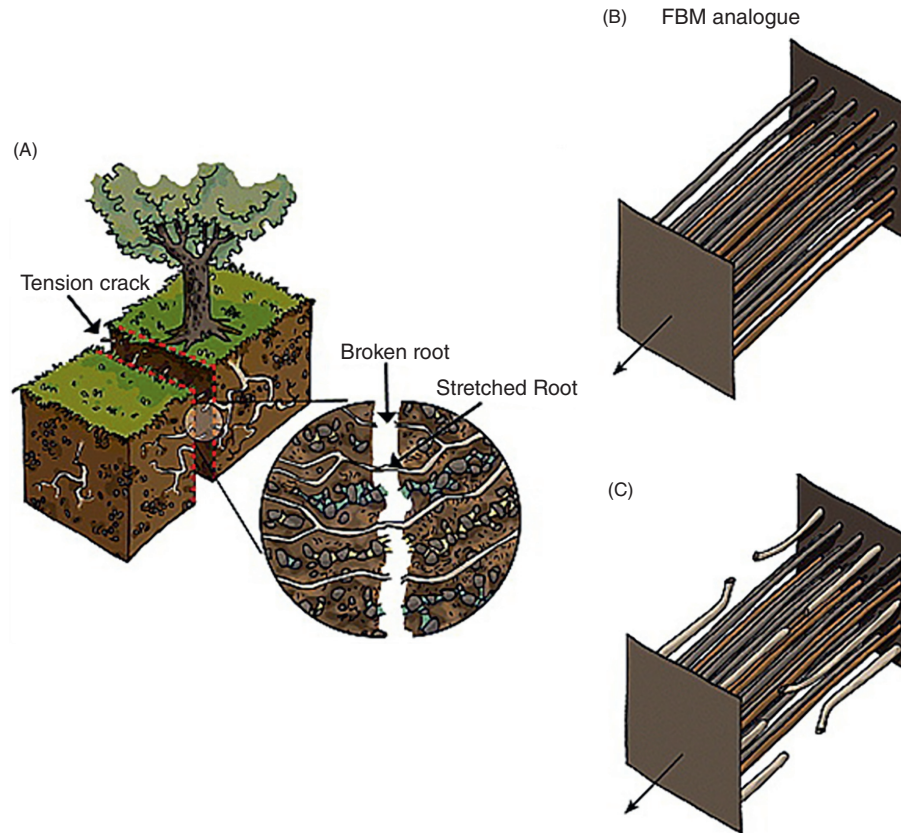


Fig. 5 During (A) a shallow landslide or soil failure in tension, plant roots are stretched, broken or can slip out of soil. In (B) Fiber Bundle Models (FBM), roots are modelled as a cluster of elements subjected to tensile forces. FBMs consider the sharing of load or energy over the modelled bundle of roots, that differ in size and material properties. (C) The modelled failure of roots is not simultaneous, reflecting field conditions. Image reproduced with kind permission from Cohen D, Lehmann P, and Or D (2009) Fiber bundle model for multiscale modeling of hydromechanical triggering of shallow landslides. *Water Resources Research* 45.

due to debris flow or wind). Although complex and costly, centrifugal technology will help our understanding of plant anchorage and soil failure processes, as well as validating the integrity of numerical soil reinforcement models.

Different types of models have been formulated over the years to predict landslide hazard at the slope level (see Murgia et al. (2022) for a review). Usually, an estimation of the Factor of Safety (FoS, i.e., the ratio of resisting forces to driving forces), based on physical data of soil and vegetation properties is calculated (Norris et al., 2008). A slope is considered to be stable if the FoS is > 1.0 . Engineered slopes are generally designed to have a FoS between 1.2 and 1.4 (Norris et al., 2008). Natural slopes have highly variable FoS depending on their geological and engineering properties. Studies that estimate FoS help explain why a slope has failed locally, but upscaling slope stability models to the landscape level is a challenge, partially because of a lack of suitable data to evaluate models, the challenge of selecting realistic input data in a heterogeneous environment, and also because of a lack of understanding of biophysical processes at different scales. There have been few attempts to model the influence of vegetation on slope stability over a landscape. However, the recent availability of accessible databases and techniques such as airborne LiDAR (Light Detection and Ranging) for assessing vegetation, makes it easier to obtain reliable vegetation, soil and hydrological input data. Similarly, the use of cutting-edge methods such as sediment source fingerprinting, that differentiates the properties of soil from different sources, (using e.g., signatures such as elemental geochemistry, radionuclides or environmental DNA in soil (Frankl et al., 2022)), would allow the investigation of complex and often poorly understood relationships between vegetation cover, substrate mass movement and geomorphological response at the catchment scale. Understanding, modelling and predicting mass movement processes over time and space in a vegetated landscape is one of the next major challenges for soil and plant biophysicists.

Outlook

Plant roots have a large impact on soil stability, operating at scales from individual aggregates up to the landscape. Changes to land management away from natural systems with perennial root systems, to areas temporally devoid of vegetation due to agriculture, forestry, construction or land degradation, have resulted in soils that are much more mechanically weak, and hence susceptible to landslides or other forms of instability. There is great scope to use vegetation to improve the physical stability of soil by the action of

plant roots, as has been observed in major restoration projects such as on the Loess Plateau in China. Geotechnical engineering projects can also replace nails, piles or concrete facings with vegetation to stabilize slopes, providing a green solution that is also cost-effective. The uptake of such approaches, however, has been limited by the trust in available predictive models and uncertainties caused by the interaction between soil, plants and weather.

References

- ASTM International (2012) *ASTM D3080-04. Standard Test Method for Direct Shear Test of Soils Under Consolidated Drained Conditions*. <https://doi.org/10.1520/D3080-04>.
- Bischetti GB, Chiaradia EA, Epis T, and Morlotti E (2009) Root cohesion of forest species in the Italian Alps. *Plant and Soil* 324: 71–89.
- Cohen D, Lehmann P, and Or D (2009) Fiber bundle model for multiscale modeling of hydromechanical triggering of shallow landslides. *Water Resources Research* 45.
- Cohen D, Schwarz M, and Or D (2011) An analytical fiber bundle model for pullout mechanics of root bundles. *Journal of Geophysical Research* 116: F03010.
- Fan C-C (2012) A displacement-based model for estimating the shear resistance of root-permeated soils. *Plant and Soil* 355: 103–119.
- Fannin RJ, Eliadorani A, and Wilkinson JMT (2005) Shear strength of cohesionless soils at low stress. *Géotechnique* 55: 467–478.
- Flerchinger GN, Lehrs GA, and McCool DK (2005) Freezing and thawing - processes. In: *Encyclopedia of Soils in the Environment*, pp. 104–110. Elsevier.
- Forsta V, Capobianco V, and Cascini L (2020) Influence of grass roots on shear strength of pyroclastic soils. *Canadian Geotechnical Journal* 57: 1320–1334.
- Frankl A, Evrard O, Cammeraat E, Tytgat B, Verleyen E, and Stokes A (2022) Tracing hotspots of soil erosion in high mountain environments: How forensic science based on plant eDNA can lead the way. An opinion. *Plant and Soil* 476: 729–742.
- Freschet GT, Roumet C, Comas LH, Weemstra M, Bengough AG, Rewald B, Bardgett RD, de Deyn GB, Johnson D, Klimešová J, Lukac M, McCormack ML, Meier IC, Pagès L, Poorter H, Prieto I, Wurzbürger N, Zadworny M, Bagniewska-Zadworna A, Blancaflor EB, Brunner I, Gessler A, Hobbie SE, Iversen CM, Mommer L, Picon-Cochard C, Postma JA, Rose L, Ryser P, Scherer-Lorenzen M, Soudzilovskaia NA, Sun T, Valverde-Barrantes OJ, Weigelt A, York LM, and Stokes A (2021) Root traits as drivers of plant and ecosystem functioning: Current understanding, pitfalls and future research needs. *New Phytologist* 232: 1123–1158.
- Ghestem M, Sidle RC, and Stokes A (2011) The influence of plant root systems on subsurface flow: Implications for slope stability. *Bioscience* 61: 869–879.
- Ghestem M, Veylon G, Bernard A, Vanel Q, and Stokes A (2014) Influence of plant root system morphology and architectural traits on soil shear resistance. *Plant and Soil* 377: 43–61.
- Giadrossich F, Schwarz M, Cohen D, Cislighi A, Vergani C, Hubble T, Phillips C, and Stokes A (2017) Methods to measure the mechanical behaviour of tree roots: A review. *Ecological Engineering* 109: 256–271.
- Giadrossich F, Cohen D, Schwarz M, Ganga A, Marrosu R, Pirastu M, and Capra GF (2019) Large roots dominate the contribution of trees to slope stability. *Earth Surface Processes and Landforms* 44: 1602–1609.
- Gray DH and Ohashi H (1983) Mechanics of fiber reinforcement in sand. *Journal of Geotechnical Engineering* 109: 335–353.
- Hu X, Brierley G, Zhu H, Li G, Fu J, Mao X, Yu Q, and Qiao N (2013) An exploratory analysis of vegetation strategies to reduce shallow landslide activity on loess hillslopes, Northeast Qinghai-Tibet Plateau, China. *Journal of Mountain Science* 10: 668–686.
- International Organization for Standardization (2018) *Geotechnical investigation and testing - Laboratory testing of soil - Part 10: Direct shear tests*.
- Liu H, Mao Z, Wang Y, Kim JH, Bourrier F, Mohamed A, and Stokes A (2021) Slow recovery from soil disturbance increases susceptibility of high elevation forests to landslides. *Forest Ecology and Management* 485: 118891.
- Lu N and Likos WJ (2006) Suction stress characteristic curve for unsaturated soil. *Journal of Geotechnical and Geoenvironmental Engineering* 132: 131–142.
- Mao Z (2022) Root reinforcement models: classification, criticism and perspectives. *Plant and Soil* 472: 17–28.
- Meijer GJ (2021) A generic form of fibre bundle models for root reinforcement of soil. *Plant and Soil* 468: 45–65.
- Meijer G, Bengough G, Knappett J, Loades K, and Nicoll B (2019a) Measuring the Strength of Root-Reinforced Soil on Steep Natural Slopes Using the Corkscrew Extraction Method. *Forests* 10: 1135.
- Meijer G, Muir Wood D, Knappett JA, Bengough AG, and Liang T (2019b) Root branching affects the mobilisation of root-reinforcement in direct shear. In: *E3S Web of Conferences* 92 12010.
- Mickovski SB and van Beek LPH (2009) Root morphology and effects on soil reinforcement and slope stability of young vetiver (*Vetiveria zizanioides*) plants grown in semi-arid climate. *Plant and Soil* 324: 43–56.
- Murgia I, Giadrossich F, Mao Z, Cohen D, Capra G, and Schwarz M (2022) Modeling shallow landslides and root reinforcement: A review. *Ecological Engineering* 181. <https://doi.org/10.1016/j.ecoleng.2022.106671>.
- Nespoulous J, Merino-Martín L, Monnier Y, Bouchet DC, Ramel M, Dombey R, Viennois G, Mao Z, Zhang J-L, Cao K-F, le Bissonnais Y, Sidle RC, and Stokes A (2019) Tropical forest structure and understorey determine subsurface flow through biopores formed by plant roots. *Catena* 181: 104061.
- Norris JE, Stokes A, Mickovski SB, Cammeraat E, Van Beek R, Nicoll BC, and Achim, A. (Eds.). (2008) *Slope Stability and Erosion Control: Ecotechnological Solutions*. Springer Science & Business Media.
- Pollen N, Simon A, and Collison A (2004) Advances in assessing the mechanical and hydrologic effects of riparian vegetation on streambank stability. In: *Riparian Vegetation and Fluvial Geomorphology*, pp. 125–139. American Geophysical Union.
- Preti F (2006) Stabilità dei versanti vegetati. In: Sauli G, Cornellini P, and Preti F (eds.) *Manuale 3 Ingegneria Naturalistica Sistemazione dei versanti. Regione Lazio (in Italian)*, 137–168. Cap. 10.
- Stokes A, Atger C, Bengough AG, Fourcaud T, and Sidle RC (2009) Desirable plant root traits for protecting natural and engineered slopes against landslides. *Plant and Soil* 324: 1–30.
- Terzaghi K, Peck RB, and Mesri G (1996) *Soil Mechanics in Engineering Practice*, 3rd edn New York, NY: Wiley-Interscience Publication. John Wiley & Sons, Inc.
- Veylon G, Ghestem M, Stokes A, and Bernard A (2015) Quantification of mechanical and hydric components of soil reinforcement by plant roots. *Canadian Geotechnical Journal* 52: 1839–1849.
- Waldron LJ (1977) The shear resistance of root-permeated homogeneous and stratified soil. *Soil Science Society of America Journal* 41: 843–849.
- Waldron LJ and Dakessian S (1982) Effect of grass, legume, and tree roots on soil shearing resistance. *Soil Science Society of America Journal* 46: 894–899.
- Wu TH, McKinnell WP III, and Swanston DN (1979) Strength of tree roots and landslides on Prince of Wales Island, Alaska. *Canadian Geotechnical Journal* 16: 19–33.
- Yildiz A, Graf F, Rickli C, and Springman SM (2019) Assessment of plant-induced suction and its effects on the shear strength of rooted soils. *Proceedings of the Institution of Civil Engineers: Geotechnical Engineering* 172: 507–519.
- Zhang C-B, Chen L-H, Liu Y-P, Ji X-D, and Liu X-P (2010) Triaxial compression test of soil-root composites to evaluate influence of roots on soil shear strength. *Ecological Engineering* 36: 19–26.
- Zhang X, Knappett JA, Leung AK, Ciantia MO, Liang T, and Danjon F (2020) Small-scale modelling of root-soil interaction of trees under lateral loads. *Plant and Soil* 456: 289–305.

Physical effects of soil salinity

David Russo, Institute of Soil, Water and Environmental Sciences, The Volcani Institute, Agricultural Research Organization, Rishon LeZion, Israel

© 2023 Elsevier Ltd. All rights reserved.

This is an update of D. Russo, Salinity I Physical Effects, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, pp. 442–453, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00529-4>.

Introduction	124
Water retention and swelling in salt-affected soils	125
Soil hydraulic conductivity in salt-affected soils	126
Modelling hydraulic conductivity and water retention in salt-affected soils	128
Calculations of the $b(h, R, C_0)$ and the $N(h, R, C_0)$ functions	128
Calculations of the $\theta(h, R, C_0)$ and the $K(h, R, C_0)$ functions	129
General implications for flow and transport	131
Flow and transport in a realistic salt-affected soil	133
The coupling-interaction parameters on the field scale	133
The equations governing flow and transport of interacting solutes in spatially heterogeneous, variably saturated soils	134
A case study	135
The physical domain and the simulated flow scenario	135
Results and discussion	135
Concluding remarks	137
References	138

Abstract

Salinity has become an increasing threat to soil physical properties due to drought and water shortages. There are impacts of salinity on plants due to osmotic impacts on soil pore water, and to soil water flow and transport from soil solution-soil matrix interactions. As the soil solution sodium adsorption ratio increases, the effect of salinity-induced, soil solution-soil matrix interactions on flow and transport worsens, as both soil solution concentration and capillary pressure head decrease. In finer textured soil, particularly those with smectite clay, pore structure changes exacerbate the impacts from salinity due to swelling, when sodium salts are present. This article provides background theory and a case study for soils irrigated with low-quality water.

Key points

- A shortage of rain and water resources in semi-arid and arid zones necessitates the use of low-quality water (LQW) such as saline water and treated sewage water (TSW) for irrigation
- Use of LQW for irrigation may accelerate contamination or salinization of soils and aquifers underneath irrigated regions
- Quantitative field-scale descriptions of chemical transport in the unsaturated zone are essential for basic understanding of LQW transport in near-surface geological environments, and for assessing reliably the threat posed by using LQW for irrigation to these environments
- Quantitative description of LQW transport requires the coupling of the flow to the transport through the dependence of the soil hydraulic properties on the solute concentrations.

Units and Nomenclature

a_i ($i = 1$ to 4)	Coefficients in Eqs. (3), (5) and (6) [dimensionless]
A_s	Specific surface area [$\text{cm}^2 \text{g}^{-1}$]
A_d	Specific surface area for adsorption [$\text{cm}^2 \text{g}^{-1}$]
A_e	Specific surface area for exclusion [$\text{cm}^2 \text{g}^{-1}$]
B	Half distance between two adjacent tactoids [cm]
$B(\text{Eq. 2})$	$kLTC_0/10^3(R + 1)$ [erg cm^{-3}] ($1 \text{ erg} = 1 \text{ g cm}^2 \text{ s}^{-2}$)
$C_{(.)}$	Concentration of the ion (.) [meql^{-1}]
C	$\text{can} + C_{\text{Ca}}$ [meql^{-1}]
$c_{(.)}$	Molar concentration of the ion (.) [mol l^{-1}]
C_0	$c_{\text{Na}} + c_{\text{Ca}}$ [mol l^{-1}]

d	Position of the mid-plane between two adjacent tactoids [cm]
D	The dielectric constant of water [dimensionless]
E_f	Elution factor [dimensionless]
F	Faraday constant [2.893×10^{11} esu meq ⁻¹] (1 esu = 1 g ^{1/2} cm ^{3/2} s ⁻¹)
G	$(8\pi C_0 Le^2 / 10^3 DkT)^{1/2}$ [molecules electron ⁻¹ cm ⁻¹]
h	Capillary pressure head [cm]
h_b	Air entry value of h [cm]
i, j	Indices [dimensionless]
k	Boltzmann's constant [1.38×10^{-16} erg deg. ⁻¹ molecule ⁻¹]
K	Hydraulic conductivity [cm s ⁻¹]
K^r	Relative hydraulic conductivity [dimensionless]
ℓ	Average number of pores having a unit length [cm]
L	Avogadro number [6.0225×10^{23} molecules mol ⁻¹]
P	Swelling pressure [erg cm ⁻³]
P_w , P_a and $P_c = P_w - P_a$	Water, air and capillary pressures, respectively [erg cm ³]
r	Equivalent radius of water-filled pore [cm]
R	c_{Na}/c_{Ca} [dimensionless]
R_f	Retardation factor [dimensionless]
s	Position of the plane of the charged surface [cm]
SAR	Sodium adsorption ratio ($=C_{Na}/\sqrt{C_{Ca}/2}$) [(meql ⁻¹) ^{1/2}]
T	Absolute temperature [Kelvin degrees]
V	Volume [cm ³]
v_i	Sub-volume ($\sum v_i = V$) [cm ³]
x	Distance [cm]
Y	Scaled electrical potential [molecule electron ⁻¹]
Y_s	Value of Y at the charged surface [molecule electron ⁻¹]
Y_d	Value of Y at the mid-plane between two adjacent tactoids [molecule electron ⁻¹]
$Z_{(.)}$	Valence of the ion (.) [electron molecule ⁻¹]
α	$8000\pi/LDkT$ (Eq. 7) [erg mol ⁻¹]
β	Pore size distribution index [dimensionless]
γ	Specific gravity of water [g cm ⁻² s ⁻²]
Γ_s	Surface charge density [esu cm ⁻²]
Γ_{Cl}	Quantity of the excluded chloride per unit area of the charged surface [meq cm ⁻²]
Γ_{Na}	Quantity of sodium ionic charges per unit area of the charged surface [esu cm ⁻²]
δ	Space charge density [esu cm ⁻³]
ϵ	Electronic charge [4.803×10^{-10} esu electron ⁻¹]
η	Dynamic viscosity of water [g cm ⁻¹ s ⁻¹]
θ	Volumetric water content [cm ³ cm ⁻³]
θ_s , θ_r	Saturated and residual values of θ [cm ³ cm ⁻³]
Θ	Effective water saturation [dimensionless]
κ	Permeability [cm ²]
ρ_b	Bulk density [g cm ⁻³]
σ	Surface tension at the air-water interface [erg cm ⁻²]
v	Volume of water retained by clay particles in a unit volume of soil [cm ³ cm ⁻³]
φ	Porosity [cm ³ cm ⁻³]
ψ	Electrical potential [erg esu ⁻¹]

Introduction

Soil salinity arises from dissolved salts in soil solution, introduced by capillary rise and evaporation from deeper soil or from irrigation with water containing salts. A shortage of rain and water resources exists in semi-arid and arid zones, necessitating the use of low-quality water (LQW) such as saline water and treated sewage water (TSW) for irrigation, thereby preserving scarce fresh water resources to meet the increasing demand of the urban sector (e.g., [Hamilton et al., 2007](#) and [Qadir et al., 2007](#)). Inasmuch as LQW may contain relatively large quantities of soluble salts, its use for irrigation may accelerate the contamination or the salinization of soils and aquifers underneath the irrigated regions (e.g., [Assouline et al., 2015](#)).

Quantitative field-scale descriptions of chemical transport in the vadose (unsaturated) zone, therefore, are essential for improving the basic understanding of the transport of LQW in near-surface geological environments. They provide predictive tools, that, in turn, can be used to predict the future spread of LQW in these environments and to assess, reliably, the threat to soils and the underlying water supplies posed by using LQW irrigation.

Modelling irrigation with LQW, and, particularly, with TSW, requires at least three sub-systems to be taken into account in the effort: (i) the N-C-O sub-system describing the cycling of nitrogen and carbon compounds in the unsaturated zone; (ii) the major ions sub-system (Cl , SO_4 , HCO_3 , Ca , Mg , Na , and K) contributing to salinity; and (iii) phosphate, boron, and trace metals. The focus of this chapter is on the effect of soil salinity on the soil hydraulic properties, through soil solution-soil matrix (SS-SM) interactions. Consequently and for simplification, the practical soil-water system containing only the major ions in the irrigation water is considered here.

The SS-SM interactions include phenomena such as cation exchange, anion exclusion, precipitation and dissolution, swelling and reorganization of the soil colloid structures and, concurrently, rearrangement of the soil pore size distribution; the latter may affect the soil hydraulic conductivity and soil water retention (e.g., McNeal, 1968; Russo and Bresler, 1977a,b). These interactions, therefore, may affect water flow and solute transport markedly. The magnitude of the SS-SM interactions depends upon the types and amounts of inorganic and organic soil colloids. Among soil mineral colloids, the most reactive constituents are the smectite minerals (e.g., montmorillonite, beidellite, montronite), characterized by a dioctahedral structure. The presence of these minerals imparts a considerable cation exchange capacity or specific surface area to soils in arid and semi-arid regions. Furthermore, the smectite minerals are capable of considerable expansion at moderate to high exchangeable sodium levels in the presence of relatively low-salt soil solution concentrations. For comprehensive reviews, see Bresler et al. (1982) and Shainberg (1984). Hence, they can impart substantial salinity-dependent soil water retention and soil hydraulic conductivity changes, and, concurrently, may considerably affect water flow and solute transport (e.g., Russo, 1988, 2010, 2013).

The present chapter concentrates on the effect of salinity on soil physical properties relevant to flow and transport, in particular soil water retention and hydraulic conductivity. Experimental evidence on the effect of soil salinity on soil water retention and hydraulic conductivity, together with an approach to model these entities in salt-affected soils are presented and discussed. Implications of the results of these analyses for flow and transport in salt-affected soils are demonstrated and assessed.

Water retention and swelling in salt-affected soils

The amount of water retained by soil depends largely upon the soil's mineral and organic colloid contents, although the structural arrangement of soil pores may also be of considerable importance. In particular, the presence of substantial quantities of soil organic matter or smectite minerals results in a considerable water retention capacity. Soils with high smectite content may also swell considerably more in the presence of high sodium or high salt concentrations. Soils high in organic colloids are much more stable and do not tend to swell excessively under adverse chemical conditions. They may disperse, however, at high pH levels or in the presence of low salt concentrations (Bresler et al., 1982).

Following Russo and Bresler (1977a), the effects of concentration and composition of the soil solution on the water retention on the local (Darcy) scale, in a soil containing 20% clay (dominated by montmorillonite) are depicted in Fig. 1, for a simple but practical soil water system containing Cl , Na and Ca ions only.

The functional relationships (under wetting conditions) in Fig. 1 are given in terms of the capillary pressure head, h , (considered here as a positive quantity), volumetric soil-water content, θ , and equilibrium solution concentration, $C = C_{\text{Na}} + C_{\text{Ca}}$, where C is concentration in meq/l , and composition (expressed in terms of sodium adsorption ratio, $\text{SAR} = C_{\text{Na}}/\sqrt{(C_{\text{Ca}}/2)}$). The points in Fig. 1 represent actual measurements for various values of SAR and C . The solid lines represent best visual fits between measured points. Note that the ordinate scale of Fig. 1 is a shift upward by one order of magnitude, and the data accordingly translated along this axis, to distinguish between different sets of concentrations. In general, each of the water retention relationships in Fig. 1 can be represented by a single continuous curve. When solution concentration exceeds 50 meq/l , a single-valued function is obtained for any value of SAR. The same is true for $\text{SAR} = 0$ and $2 \text{ meq/l} \leq C \leq 5 \text{ meq/l}$, as can be seen from the lowest curve of Fig. 1. Within the range $2 \text{ meq/l} \leq C \leq 10 \text{ meq/l}$ and $0 \leq \text{SAR} \leq 50$, water content, θ depends not only on h , but also on both concentration and composition of the soil solution. For a given h , θ increases as C decreases and SAR increases.

The experimental findings presented in Fig. 1 can be qualitatively explained in terms of electrostatic properties and the porous nature of the soil. It is assumed that the soil pore space can be divided into structural and textural porosity. The latter is regarded as an intrinsic property of the soil, arising from the size and arrangement of basic particles within the soil mass. In these pores the $h(\theta)$ function can be determined quantitatively from the pressure that develops between the clay platelets, within the clay mass between the pores as the soil solution composition and concentration vary.

The classical view of the osmotic pressure of soil colloids is that cations in the diffuse layer between clay particles reduce the free energy of the water relative to that in the external equilibrium solution. This causes water molecules to diffuse into the space between the clay particles, since the cations are electro-statically constrained from leaving the interlayer region. Hence, the particles move apart to accommodate the extra water. If, on the other hand, the clay particles are confined, hydraulic pressure develops between them. For a Na/Ca-Cl soil-water system, this pressure may be determined, based on the mixed-ion diffuse double-layer theory (Bresler, 1972), see Eq. (2) below.

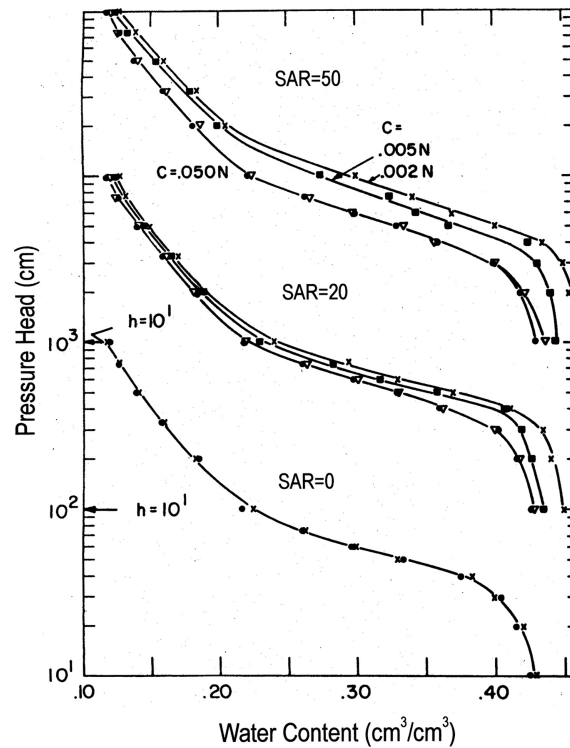


Fig. 1 Capillary pressure head, h , as a function of water content, θ , for selected values of the concentration, C , and the sodium adsorption ratio, SAR, of the equilibrium solution. Arrows indicate that the point $h = 10$ cm is shifted and the data are translated along the h -axis. Reproduced from Russo D and Bresler E (1977a) Effect of mixed Na-Ca solutions on the hydraulic properties of unsaturated soils. *Soil Science Society of America Journal* 41: 714–717.

Eq. (2) provides a reasonably good first-order approximation of the water retention forces acting within the textural porosity. It predicts that for a given hydrostatic pressure, P , the volume of water retained by the clay mass (which is proportional to the distance between the clay particles) increases as the equilibrium solution becomes more diluted (C decreases) and less dominated by the divalent ion (SAR increases). In the structural pores, which are sensitive to external forces arising from interactions between clay particles within the textural porosity, the $h(\theta)$ function can be determined from the pressure within the water-filled pores relative to that within the air-filled pores. In the former pores, at equilibrium, the pressure increases as the pore radius decreases, or for a given pressure, higher values of θ are associated with a larger number of pores having a given corresponding radius. Thus, the functional relationships between h and θ in the structural porosity depend on the total porosity and the pore size distribution. For a given pressure, the space between the clay particles increases as SAR increases and as C decreases. This leads to an increasing amount of water retained by the clay mass, and concurrently to the swelling of the clay mass so that it occupies a larger volume. For a constant-volume system, changes in the volume of the clay mass are at the expense of the structural porosity and its pore size distribution. Thus, as can be observed in Fig. 1, for a given capillary pressure head, the amount of water retained by the soil increases with decreasing C and increasing SAR.

Soil hydraulic conductivity in salt-affected soils

The retention curves in Fig. 1 represent the soil's pore size distribution to some extent. The data in Fig. 1, therefore, suggests that swelling of the clay fraction within the soil, induced by relatively low C and high SAR, may alter the geometry of the soil pores and thus affect the intrinsic soil permeability, $\kappa(\theta)$. Investigations (e.g., McNeal, 1968) have confirmed that soil permeability or hydraulic conductivity, $K(\theta) = \kappa(\theta)\gamma/\eta$, where γ and η are the specific gravity and the dynamic viscosity of water, respectively, is highly affected by the soil solution concentration, C , and composition, expressed in terms of SAR. It has been suggested that the swelling of clay particles in a confined soil system causes the size of large soil pores to decrease, and that breakdown of domains of clay platelets (tactoids) (Blackmore and Miller, 1961; Shainberg and Kaiserman, 1969) and the subsequent movement and dispersion of the clay platelets may further block soil pores. Low permeability can result from such geometric restrictions. Investigators found a positive correlation between relative soil permeability and the swelling of extracted soil clays, which, in turn, was affected by the soil solution concentration and composition. Inasmuch as soil solution composition may affect the breakdown of tactoids (and the subsequent movement and dispersion of clay platelets) on the one hand, and soil swelling on the

other hand, in much the same way, the relative importance of swelling and dispersion in affecting soil permeability remains essentially unresolved (Shainberg and Kaiserman, 1969). There are situations (e.g., relatively coarse-textured soils associated with relatively high influx rates), in which the de-mixing of adsorbed ions, which concentrates Na on external particle surfaces, may lead to the breakdown of tactoids and the subsequent movement and dispersion of clay platelets.

Following Russo and Bresler (1977a), the effect of concentration and composition of the soil solution on the soil hydraulic conductivity function $K(\theta)$ at the local (Darcy) scale, for the same soil as in Fig. 1, is depicted in Fig. 2.

The functional relationships presented in Fig. 2 are given in terms of water content, θ , and equilibrium solution concentration, C and composition, SAR. The points in Fig. 2 represent actual measurements for various values of SAR and C . The solid lines represent best visual fits between measured points. Note that the abscissa scale in Fig. 2 has been shifted to the right and the data are accordingly translated along the θ -axis to distinguish among different sets of concentrations.

In general, each of the $K(\theta)$ functions in Fig. 2 can be represented by a single continuous curve. Furthermore, for the calcium-saturated system ($SAR = 0$ and $2 \text{ meq/l} \leq C \leq 50 \text{ meq/l}$), a single-valued $K(\theta)$ is obtained. In the mixed Na-Ca system, however, values of $K(\theta)$ decrease with both θ and C for a given SAR. The differences among the various functions increase as θ becomes larger and as C becomes smaller. These differences are pronounced when SAR and θ are greater than 15 and 0.25, respectively, and when C is equal to or less than 10 meq/l . In other words, relatively low values of soil water content compensate for the negative effects of high SAR and low C on the hydraulic conductivity.

The experimental data depicted in Fig. 2 can be explained qualitatively in terms of soil-water-clay interactions and the porous nature of the soil. The unsaturated hydraulic conductivity function, $K(\theta)$, depends upon the size distribution of water-filled pores and the total water-filled porosity. As mentioned in the previous section, the diffuse double-layer theory for mixed-electrolyte systems predicts that for a given pressure, the space between clay particles, and, concurrently, the amount of water retained by the clay mass, increases as SAR increases and as C decreases. For a constant-volume system, changes in the volume of the clay mass are at the expense of the quantity and distribution of the soil pores. Thus, for a given h , the amount of water retained by the soil increases (Fig. 1) and the hydraulic conductivity decreases (Fig. 2) as SAR becomes larger and solution concentration, C decreases. Since the swelling of clay decreases as h increases, the increase in the water volume retained and the decrease in hydraulic conductivity become smaller as h increases. In addition, both water content and hydraulic conductivity, and concurrently water velocity, may considerably decrease with increasing h . Consequently, increasing h may reduce the possibility that movement and dispersion of clay platelets will lead to additional blocking of the pore space available for water flow.

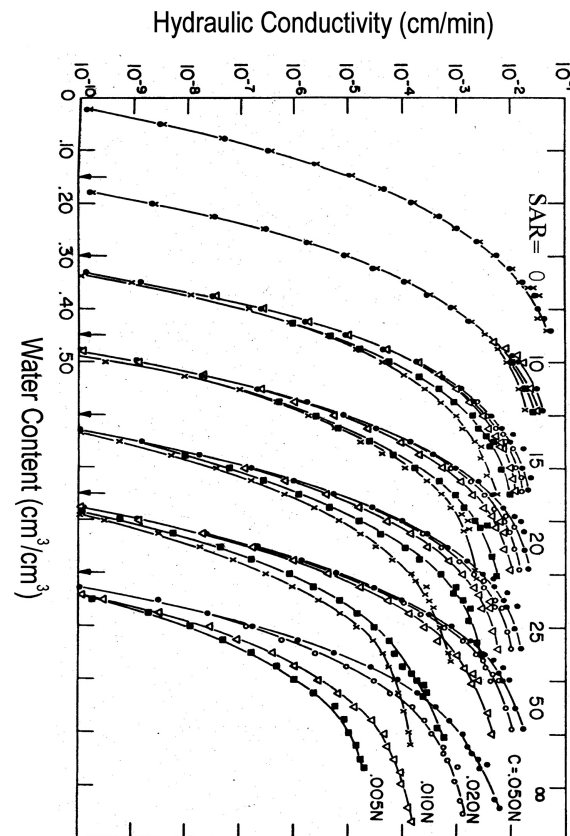


Fig. 2 Hydraulic conductivity, K , as a function of water content, θ , for selected values of the concentration, C , and the sodium adsorption ratio, SAR, of the equilibrium solution. Arrows indicate that the point $\theta = 0$ is shifted and the data are translated along the θ -axis. Reproduced from Russo D and Bresler E (1977a) Effect of mixed Na-Ca solutions on the hydraulic properties of unsaturated soils. *Soil Science Society of America Journal* 41: 714–717.

Modelling hydraulic conductivity and water retention in salt-affected soils

In the preceding sections it was shown that in salt-affected soils, both the local-scale soil hydraulic conductivity, $K(h)$, and the soil water retention, $\theta(h)$, may depend on soil solution concentration and composition. Ion precipitation or dissolution will have impacts, but following Russo and Bresler (1977b) and Russo (1988), a simple but credible soil-water system can contain just Na, Ca and Cl ions. The equilibrium solution is quantified in terms of the sum of the molar concentrations of the cations, $C_0 = c_{Na} + c_{Ca}$, and the ratio between the molar concentration of mono- and di-valent cations, $R = c_{Na}/c_{Ca}$.

For the modelling of the soil hydraulic conductivity, $K(h, R, C_0)$, and the soil water retention, $\theta(h, R, C_0)$, functions at the local (Darcy) scale, processes and entities relevant to the micro (pore) scale should be taken into account. It is assumed that: (i) the soil clay fraction is mostly montmorillonite (having a constant effective surface charge density, Γ_s), which may exist in packets of platelets or domains (tactoids); (ii) the ions on the external surfaces of the tactoids (the "exchange phase") are distributed according to the planar mixed-ion diffuse double-layer theory; (iii) the number of clay platelets per tactoid depends on the pressure acting on the soil-water system, and on the exchangeable sodium percentage, ESP; and (iv) due to the continuity of the pressure in the soil solution, the pressure difference between two adjacent clay particles and the bulk solution (given by the swelling pressure, P (Eq. (2) below) is equal to the pressure difference across the air-water interface within the soil pores. In other words, the difference between the pore water pressure, P_w , and the atmospheric pressure, P_a , as given by the capillary pressure, $P_c = P_w - P_a = -\gamma h$.

Calculations of the $b(h, R, C_0)$ and the $N(h, R, C_0)$ functions

For the Na/Ca-Cl soil-system considered here, the general equation of the interacting double layer is obtained from the first integration of the one-dimensional Poisson-Boltzmann equation given by:

$$\frac{dY}{dx} = \frac{G}{\sqrt{R+1}} [R(e^{Z_{Na}Y} - e^{Z_{Na}Y_d}) + (e^{Z_{Ca}Y} - e^{Z_{Ca}Y_d}) + (R+2)(e^{Z_{Cl}Y} - e^{Z_{Cl}Y_d})]^{1/2} \quad (1)$$

where $Y = e\psi/kT$ is the scaled electrical potential, $Z_{()}$ is the valence of the ion ($\cdot$); ψ is the electrical potential; e is the charge on the electron; k is Boltzmann's constant; T is the absolute temperature; x is the distance to the charged surface (where $x = s$ and $Y = Y_s$); Y_d is Y at the midplane between two adjacent tactoids (at the plane $x = d$, where $dY/dx = 0$), $G = (8\pi C_0 L e^2 / 10^3 D k T)^{1/2}$; L is Avogadro's number, and D is the dielectric constant of the bulk solution.

For dilute solutions, the swelling pressure, P , i.e., the pressure difference between the equilibrium solution (where $Y = 0$) and the mid-plane between two adjacent tactoids (at the plane $x = d$, where $Y = Y_d$), is obtained by substituting the Boltzmann distribution law into the Van't Hoff relation. For the Na/Ca-Cl soil system considered here, with tactoids which are at a distance of $2b$ apart (where $b = d - s$), the swelling pressure, P , is given by:

$$P = B [R(e^{Z_{Na}Y_d} - 1) + (e^{Z_{Ca}Y_d} - 1) + (R+2)(e^{Z_{Cl}Y_d} - 1)] \quad (2)$$

where $B = kLTC_0/10^3(R+1)$. Substituting $P = \gamma h$ in (2), one can obtain after rearrangement

$$a_1 e^{Z_{Na}Y_d} + a_2 e^{Z_{Ca}Y_d} + a_3 e^{Z_{Cl}Y_d} + a_4 = 0 \quad (3)$$

Thus, the solution of the nonlinear equation (3) yields Y_d as a function of h , R , and C_0 . Estimation of $b(h, R, C_0)$ also requires the knowledge of the scaled electrical potential at the charged surface, Y_s , which can be estimated from the integration of Eq. (1) and from the condition of electro-neutrality requiring that the surface charge density, Γ_s , be equal to the total space charge density, $\delta(x)$, within the range $s \leq x \leq d$, thus

$$\left(\frac{dY}{dx}\right)_{x=s} = \frac{4\pi e \Gamma_s}{DkT} \quad (4)$$

Evaluating (1) at the plane $x = s$ (where $Y = Y_s$), using (4), one obtains after rearrangement

$$a_1 e^{Z_{Na}Y_s} + a_2 e^{Z_{Ca}Y_s} + a_3 e^{Z_{Cl}Y_s} + a_4 = 0 \quad (5)$$

where $a_4 = -\{2\pi 10^3(R+1)\Gamma_s^2/LDkTC_0\} + a_1 e^{Z_{Na}Y_d} + a_2 e^{Z_{Ca}Y_d} + a_3 e^{Z_{Cl}Y_d}$, and a_i , $i = 1, 2, 3$ as before.

The solution of the nonlinear equation (5) yields Y_s as a function of Γ_s , R , C_0 , and $Y_d(h, R, C_0)$. From the values of Y_d and Y_s , the distance from the midplane $x = d$ (where $dY/dx = 0$ and $Y = Y_d$) to the charged surface (where $x = s$ and $Y = Y_s$), $b = d - s$, is obtained by direct integration of (1). Rearranging (1) yields

$$\frac{G}{\sqrt{R+1}} \int_d^s dx = \int_{Y_d}^{Y_s} (a_1 e^{Z_{Na}Y} + a_2 e^{Z_{Ca}Y} + a_3 e^{Z_{Cl}Y} + a_4)^{-1/2} dY \quad (6)$$

where $a_4 = -[a_1 e^{Z_{Na}Y_d} + a_2 e^{Z_{Ca}Y_d} + a_3 e^{Z_{Cl}Y_d}]$ and a_i , $i = 1, 2, 3$ as before. Thus, a numerical integration of (6) yields $b = d - s$ as a function of h , R , and C_0 .

Once $Y_d(h, R, C_0)$ has been estimated from Eq. (3), the fractional excess of Na on the exchange phase, Γ_{Na}/Γ_s is estimated from

$$\frac{\Gamma_{Na}}{\Gamma_s} = \frac{R\sqrt{C_0}}{\Gamma_s\sqrt{\alpha(R+1)}} \sinh^{-1} \left(\frac{\Gamma_s}{\sqrt{C_0}} \frac{\sqrt{\alpha(R+1)}}{(R+4 \cosh Y_d)} \right) \quad (7)$$

where Γ_{Na} is the quantity of sodium ionic charges per unit area of the charged surface, and $\alpha = 8000\pi/L\epsilon\kappa T$. Note that, inasmuch as $Y_d = Y_d(h, R, C_0)$, the exchangeable sodium percentage, $ESP = 100\Gamma_{Na}/\Gamma_s$ is also a function of h , R and C_0 . For given pressure head, h and $ESP(h, R, C_0)$, the number of clay platelets per tactoid, $N = N[h, ESP(h, R, C_0)]$ is calculated by using experimental data on the dependence of N on h in a calcium system, and by experimental data on the dependence of N on ESP in clay suspensions (Shainberg and Kaiserman, 1969).

Calculations of the $\theta(h, R, C_0)$ and the $K(h, R, C_0)$ functions

Following Russo and Bresler (1977b) and Russo (1988), to estimate the $K(h, R, C_0)$ and the $\theta(h, R, C_0)$ functions at the local scale, a confined volume element, V , of a homogeneous and isotropic porous medium with uniform porosity, ϕ^0 , is considered. This soil volume, which is large compared with the typical pore size of the porous medium, is subdivided into n equal sub-volumes, v_i , $i = 1$ to n . Each sub-volume, v_i , is associated with single values of the average pressure head, h_i , the equivalent radius of the water-filled pores, $r_i^0 = 2\sigma/\gamma h_i$, (with σ being the surface tension at the air-water interface), and the average number of pores having a unit length, $\ell_i = \phi^0 v_i / \pi (r_i^0)^2$. Making use of the functions $b(h_i, R, C_0)$ and $N(h_i, R, C_0)$, the water content, θ_i , in each v_i is calculated from:

$$\theta_i(h_i, R, C_0) = \theta_i^0(h_i) + [v_i(h_i, R, C_0) - v_i^0(h_i)] \quad (8)$$

where θ_i^0 is the water content in a reference state (that is, a soil in equilibrium with a concentrated calcium chloride solution), at pressure head, h_i . v_i^0 and v_i are the volumes of water retained by the clay particles in a unit volume of the soil, in the reference state, and in equilibrium with a soil solution of a given R and C_0 , given by:

$$v_i^0(h_i) = \frac{A_s}{2N_i^0(h_i)} \{ 2b_i^0(h_i) + 2b_0 [N_i^0(h_i) - 1] \} \rho_b \quad (9)$$

$$v_i(h_i, R, C_0) = \frac{A_s}{2N_i^0(h_i)} \{ 2b_i(h_i, R, C_0) [1 + N_i^0(h_i) - N_i(h_i, R, C_0)] + 2b_0 [N_i(h_i, R, C_0) - 1] \} \rho_b \quad (10)$$

respectively, where A_s and ρ_b are the specific surface area and the bulk density, respectively, of the soil at the reference state, the superscript zero refers to the reference state, and $2b_0$ is the thickness of the water film between two adjacent clay platelets within a tactoid ($2b_0 = 9 \times 10^{-8}$ cm, regardless of h , R , or C_0). Similarly, since in the confined soil element the increment in the water retained by the clay fraction is at the expense of the pore space in which water flow takes place, the effective porosity, ϕ_i , in each sub-volume, v_i , is given by

$$\phi_i(h_i, R, C_0) = \theta_i^0(h_i) - [v_i(h_i, R, C_0) - v_i^0(h_i)] \quad (11)$$

and the mean equivalent radius of the pores, r_i , in each sub-volume, v_i is given by

$$r_i(h_i, R, C_0) = \left[\frac{\phi_i(h_i, R, C_0) v_i}{\pi \ell_i} \right]^{1/2} \quad (12)$$

Given the values of $\phi_i(h_i, R, C_0)$ and $r_i(h_i, R, C_0)$, the relative hydraulic conductivity is calculated on the basis of the capillary bundle theory (e.g., Childs and Collis-George, 1950; Brutsaert, 1967). Using a modified Series-Parallel model of the porous medium, the hydraulic conductivity relative to the reference state, is given (Marshall, 1958) by

$$K_i^r(h_i, R, C_0) = \frac{[\phi_i(h_i, R, C_0)]^2 \sum_{j=1}^n [2j-1] [r_j(h_j, R, C_0)]^2}{[\phi_i^0(h_i)]^2 \sum_{j=1}^n [2j-1] [r_j^0(h_j)]^2} \quad (13)$$

where ϕ_i^0 is the soil effective porosity in sub-volume, v_i , in the "stable" reference state, $h_i < h_{i+1}$, and $i = 1$ to n . Thus, the soil hydraulic conductivity, K_i , given by

$$K_i(h_i, R, C_0) = K_i^0(h_i) K_i^r(h_i, R, C_0) \quad (14)$$

where $K^0(h)$ is the hydraulic conductivity function of the soil at the "stable" reference state.

Following Russo and Bresler (1977b), a comparison between the calculated and measured water retention curves in the soil presented in the previous figures, for different combinations of $C = 10^{-3} C_0 [1 + (R+1)^{-1}]$ and $SAR = R[10^3 C_0 (R+1)^{-1}]^{1/2}$, is depicted in Fig. 3.

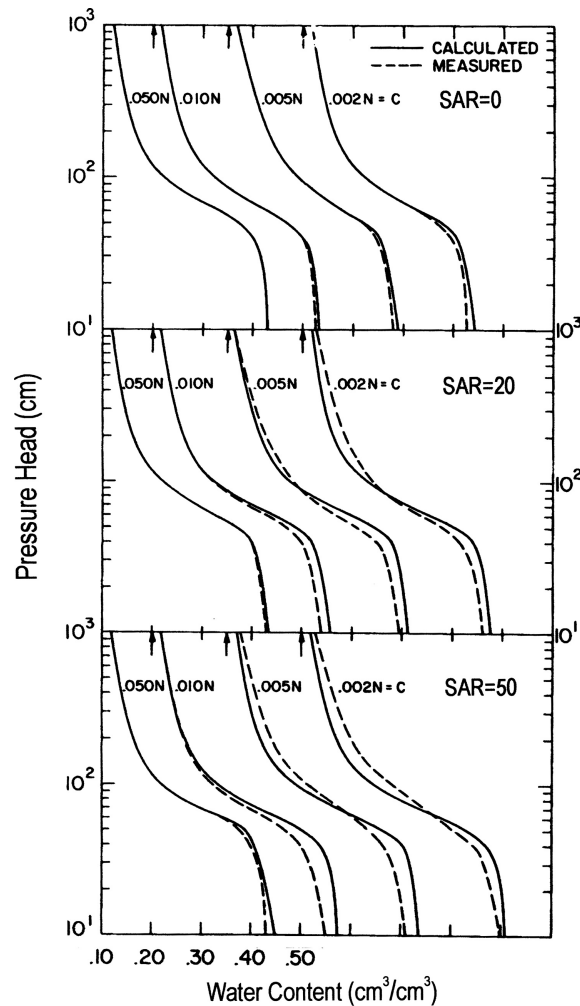


Fig. 3 Capillary pressure head, h , as a function of water content, θ , for selected values of the concentration, C , and the sodium adsorption ratio, SAR, of the equilibrium solution. Computed results (solid lines) are compared with measured data (dashed lines). Arrows indicate that the point $\theta = 0.10$ is shifted and the data are translated along the θ -axis. Reproduced from Russo (1976) *M.Sc. thesis (in Hebrew)*, Faculty of Agriculture in Rehovot, The Hebrew University, Jerusalem.

The abscissa scale in Fig. 3 is shifted to the right, and the data are accordingly translated along the θ axis to distinguish between the calculated and the measured curves for different sets of concentrations. For calcium-saturated soil ($SAR = 0$), the experimental retention curves are independent of the solution concentration, whereas the calculated curves are slightly dependent on it. The small discrepancy between the calculated and the measured curves can serve as a measure of the error in the assumptions and of the ability of the calculations to predict soil-water retention. For the mixed systems, there is reasonably good agreement between measured and calculated retention curves. The agreement improves for lower sodic systems (SAR decreases) and as the system becomes more concentrated (C increases).

A comparison between calculated and measured relative hydraulic conductivity, $K^r = K(\Theta, SAR, C)/K^0(\Theta)$, where $\Theta = (\theta - \theta_r)/(\theta_s - \theta_r)$ is the effective water saturation, and θ_s and θ_r are the saturated and residual values of θ , respectively, for the same soil as in the previous figures, for various combinations of Θ , C and SAR , is presented in Fig. 4.

The abscissa scale in Fig. 4 is shifted to the right and the data accordingly translated along the SAR axis to distinguish between the calculated and the measured curves for different sets of concentrations. It can be seen (Fig. 4) that there is relatively good agreement between calculated and measured data in most cases. For the combinations of Θ , SAR and C depicted in Fig. 4, the calculated K^r deviates by less than twice the value of the measured K^r .

The comparison between calculated and experimental results indicates that a reasonable first-order approximation to the soil-water retention and soil hydraulic conductivity functions can be obtained for relatively wide ranges of soil-solution concentrations and compositions. This can be achieved by taking into account the porous nature of the soil, the interactions between the charged solid phase and the ions in the liquid phase within the soil, and the effects of the water pressure and the soil solution composition on the organization and structure of the clay platelets.

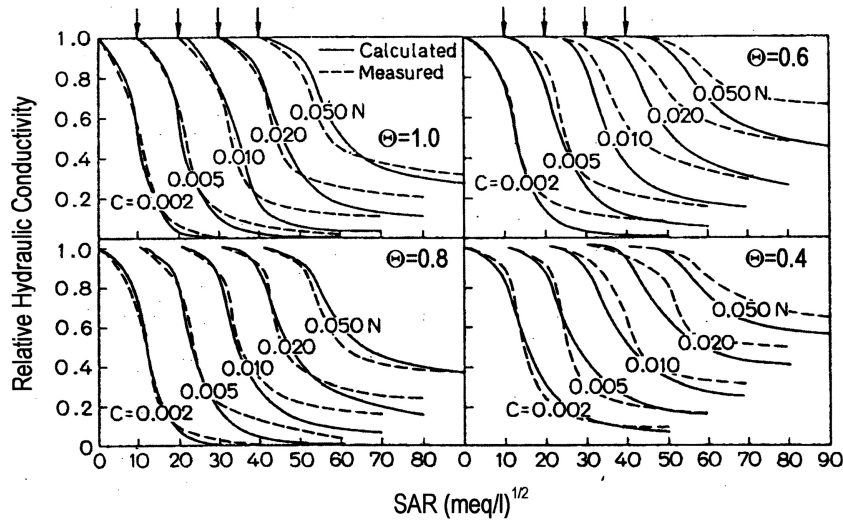


Fig. 4 Relative hydraulic conductivity, K_r , as a function of the sodium adsorption ratio, SAR, of the equilibrium solution, for selected values of the solution concentration, C , and degree of water saturation, Θ . Computed results (solid lines) are compared with measured data (dashed lines). Arrows indicate that the point $SAR = 0$ is shifted and the data are translated along the SAR-axis. Reproduced from Russo D and Bresler E (1977b) Analysis of the saturated-unsaturated hydraulic conductivity in a mixed Na-Ca soil system. *Soil Science Society of America Journal* 41: 706–710.

Soils of differing textures, containing clay fractions that are mostly montmorillonite, have essentially the same surface charge density ($\Gamma_s \approx 4.5 \times 10^4$ esu/cm²); consequently, they are characterized by essentially the same $\Gamma_{Na}(h, R, C_0)/\Gamma_s$ (Eq. 7) and $\Gamma_{Cl}(h, R, C_0)$ (see Eq. (16) below) relationships. On the other hand, the $\theta(h, R, C_0)$ and the $K^f(h, R, C_0)$ functions (Fig. 5), and, concurrently, the $K(h, R, C_0) = K^0(h)K^f(h, R, C_0)$ functions, associated with soils of different texture (see Table 1) may be quite different (Russo, 1988).

The functions depicted in Fig. 5 suggest that the SS-SM interactions affect the differences among the various soil types, with respect to their hydraulic properties. In the case of the water retention curves, the differences between the values of θ associated with the various soils generally increase as h and SAR increase and as C decreases. As for the hydraulic conductivity function, the differences between the values of K^f , and concurrently, between values of $K = K^f K^0$, associated with the various soils increase as SAR increases and as h and C decrease. Fig. 5 suggests that because of SS-SM interactions, the capability of the coarse-textured soil to transmit water and solutes faster than the finer-textured soils is extended to a larger pressure head (smaller degree of saturation) as the soil solution SAR increases and as C decreases.

Both the experimental and the theoretical results presented in the previous sections suggest that for a given soil solution concentration and composition, the effect of the SS-SM interactions on the hydraulic conductivity decreases with decreasing water saturation. Following Russo (1988), this is demonstrated in Fig. 6, in which combinations of soil solution concentration (C) and composition (SAR) required to maintain $K^f \geq 0.75$ (that is, no more than a 25% reduction in K relative to the “stable” reference state), are given for various degrees of water saturation, and for soils of different textures (Table 1). Note that the curves in Fig. 6 generalize the concept of the “threshold concentration” in terms of soil properties and soil water status. Consequently, they may be used for water quality classification in relation to soils of different textures, and for irrigation management. For a given SAR , the solution concentration required to maintain $K^f \geq 0.75$ (that is, the “threshold concentration”) decreases as the soil water content decreases and as the soil’s texture becomes coarser. Fig. 6 suggests that a decrease in the water application rate, and, concurrently, in soil water saturation, may improve the efficiency of the reclamation of sodic soils, especially, fine-textured ones.

General implications for flow and transport

Generally, for a given soil and given boundary and initial conditions, water flow is governed by the soil hydraulic properties. Thus, the effect of the SS-SM interactions is to reduce the water flow relative to that of a “stable” reference soil. For a given soil, the retardation of the water movement by the SS-SM interactions, will increase as either the initial soil solution SAR or the SAR of the applied water increases, as either the initial soil solution concentration or the solution concentration of the applied water decreases, and as the surface water application rate increases.

Solute transport, in turn, is governed by water flow and by cation exchange and anion exclusion. In the case of transport, therefore, the SS-SM interactions may act in two opposing directions: the retardation of the solute movement because of reductions in the soil hydraulic conductivity and because of cation exchange, and the acceleration of the solute movement because of anion exclusion. The retardation effect caused by the reductions in the soil hydraulic conductivity, and, concurrently, in water velocity, increases with increasing SAR and decreasing C and h , especially in relatively fine-textured soils. The retardation effect because of Na/Ca exchange is given by the retardation factor, R_f

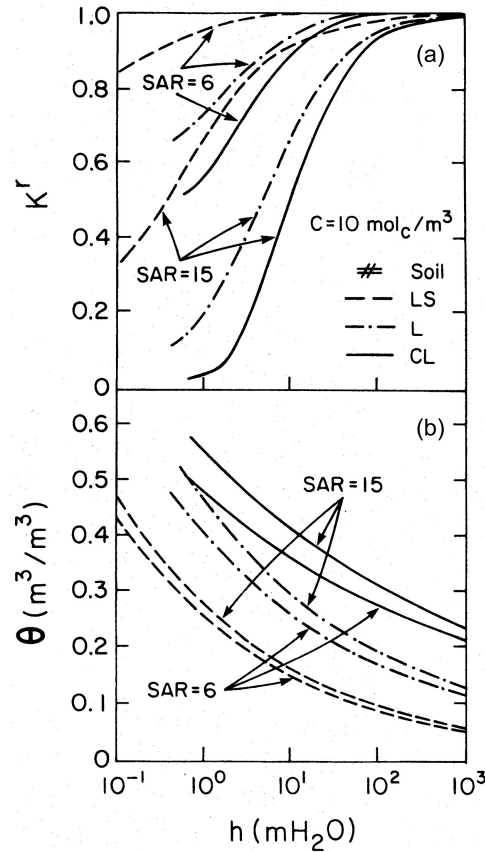


Fig. 5 Relative hydraulic conductivity, K_r , (a) and water content, θ , (b) as a function of the capillary pressure head, h for selected values of the concentration, C , and the sodium adsorption ratio, SAR, of the equilibrium solution, for three soils of different texture, loamy sand (LS), loam (L), and clay loam (CL). Reproduced from Russo D (1988) Numerical analysis of the non-steady transport of interacting solutes through unsaturated soil. 1. Homogeneous systems. *Water Resources Research* 24: 271–284.

Table 1 Representative values for relevant soil physical properties.

Soil texture	Clay fraction	A_s cm ² /g	CEC meq/g	K_s cm/h	h_b cm	B	θ_s cm ³ /cm ³	θ_r cm ³ /cm ³	ρ_b g/cm ³
Loamy sand	0.06	3.6×10^5	0.056	5.56	9.1	0.30	0.410	0.015	1.53
Loam	0.19	9.5×10^5	0.145	2.50	48.0	0.23	0.451	0.042	1.43
Clay Loam	0.34	1.7×10^6	0.248	0.77	63.4	0.15	0.476	0.068	1.33

A_s , CEC and ρ_b are the soil specific surface area, cation exchange capacity and bulk density, respectively;

K_s , θ_s and θ_r are the soil saturated hydraulic conductivity, and saturated and residual water contents, respectively;

h_b and β are the soil bubbling pressure head and pore size distribution index, respectively.

Reproduced from Russo D (1988) Numerical analysis of the non-steady transport of interacting solutes through unsaturated soil. 1. Homogeneous systems. *Water Resources Research* 24: 271–284.

$$R_f(h, R, C_0) = 1 + \frac{[R + 1] \Gamma_{Na}(h, R, C_0) F A_{ad} \rho_b}{Z_{Na} R C_0 \theta(h)} \quad (15)$$

where F is the Faraday constant, A_{ad} is the specific adsorption surface area of the soil, and Γ_{Na} is the quantity of sodium ionic charges per unit area of the charged surface (Eq. 7). The retardation caused by Na/Ca exchange (expressed in terms of R_f) increases with increasing SAR and decreasing C and h , especially in relatively fine-textured soils.

Anion exclusion can be calculated from the distribution of the scaled electrical potential, $Y(x)$ (Eq. 1) within the domain $s \leq x \leq d$, which, in turn, for given values of $b(h, R, C_0)$ and $Y_s(h, R, C_0)$, is obtained from the solution of Eq. (1), using the Runge-Kutta method. Applying the Boltzmann distribution law for the Na/Ca-Cl system considered here, the amount of excluded chloride per unit surface area, is calculated from

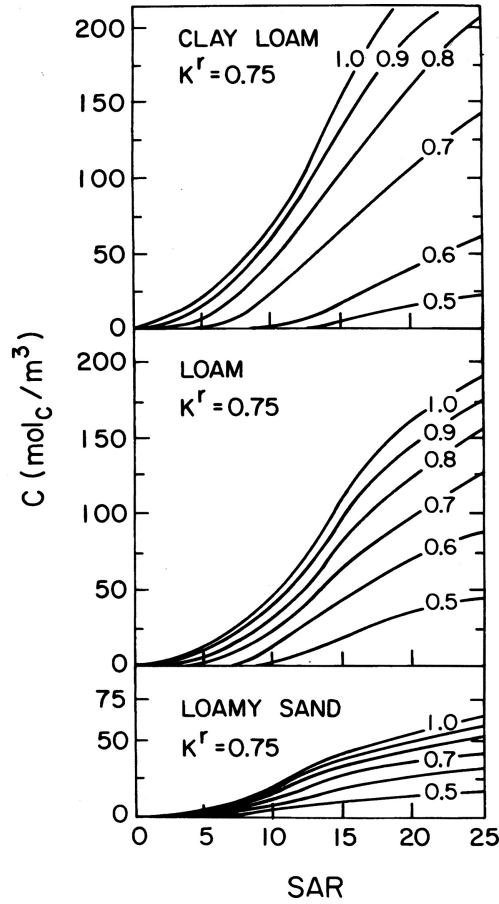


Fig. 6 Combinations of the concentration, C , and the sodium adsorption ratio, SAR, of the equilibrium solution, and degree of water saturation, Θ (the numbers labeling the curves) at which 25% reduction in hydraulic conductivity occurs, for three soils of different texture, loamy sand (LS), loam (L), and clay loam (CL). Reproduced from Russo D (1988) Numerical analysis of the non-steady transport of interacting solutes through unsaturated soil. 1. Homogeneous systems. *Water Resources Research* 24: 271–284.

$$\Gamma_{Cl}(h, R, C_0) = \frac{|Z_{Cl}| C_0 (R + 2)}{(R + 1)} \cdot \int_s^d \{1 - \exp[-Y(x; h, R, C_0)]\} dx \quad (16)$$

which, in turn, is used to calculate the elution factor, E_f , given by

$$E_f(h, R, C_0) = 1 - \frac{[R + 1] \Gamma_{Cl}(h, R, C_0) A_{ex} \rho_b}{|Z_{Cl}| C_0 [R + 2] \theta(h)} \quad (17)$$

where A_{ex} is the specific exclusion surface area.

The acceleration effect due to the chloride exclusion, which is determined by the second term on the right-hand side of Eq. (17), increases with increasing SAR and decreasing C and h , especially in relatively fine-textured soils. The net effect of the SS-SM interactions on the solute transport is determined, therefore, by the relative magnitudes of all of these processes, which, in turn, depend on the soil type and the boundary and initial conditions imposed on the flow system.

Flow and transport in a realistic salt-affected soil

The coupling-interaction parameters on the field scale

For given soil, crop, or weather, or quantity and quality of the irrigation water, solute transport couples with water flow because of its dependence on the Darcy velocities. In the case of irrigation with interacting mixed Na/Ca salts, water flow itself is coupled to the transport through the dependence of the hydraulic properties on solute concentrations. In this case, the relevant soil properties (referred to hereafter as the coupling-interaction parameters, CIPs), include the soil hydraulic conductivity, K , the soil water content,

θ , the retardation factors for Na and Ca, R_f^{Na} and R_f^{Ca} , respectively, and the elution factor for Cl, E_f , all as functions of the pressure head and the composition of the soil solution.

The theoretical framework described earlier in this chapter evaluates the CIPs on the macroscopic (Darcy) scale, pertinent to flow and transport processes at laboratory scale. Evaluation of the CIPs on a larger scale, more amenable to flow and transport processes in realistic, field-scale, spatially heterogeneous soils, requires coupling of the theoretical approach with measured spatial distributions of the relevant soil properties. The latter include the hydraulic conductivity and the water retention functions at an inert reference state, the soil cation exchange capacity, CEC, the soil specific surface area, A_s , and the soil bulk density, ρ_b . Following Russo (2013), the resultant spatial distributions of the CIPs, in turn, are quantified in terms of their probability density functions (PDFs), expressed as functions of the soil solution concentration and composition and the pressure head (Fig. 7).

The equations governing flow and transport of interacting solutes in spatially heterogeneous, variably saturated soils

It is assumed here that water flow is described locally by the 3-D Richardson-Richards equation, the physical parameters of which are viewed as realizations of 3-D, second-order, stationary random space functions, characterized completely by a constant mean and a two-point, spatial covariance. It is further assumed that the transport of each of the ions of the mixed Na/Ca salt is described locally by the classical, one-region, 3-D advection-dispersion equation (ADE). When the ions in the soil solution interact with the soil matrix, the properties of the spatially heterogeneous soil that affect flow and transport are expressed as functions of the spatial coordinate vector, $\underline{x}$, and flow-controlled attributes. The latter include the pressure head, ψ , the sum of the molar concentration of the cations, $C_0 = c_{\text{Na}} + c_{\text{Ca}}$ [mol ℓ^{-1}], and the ratio between the molar concentration of the mono- and the divalent cations, $R = c_{\text{Na}}/c_{\text{Ca}}$, in the soil solution.

In a Cartesian coordinate system (x_1, x_2, x_3), where x_1 is directed vertically downwards, assuming local isotropy, considering water uptake by the plant roots and neglecting local anisotropy and the effect of osmotic gradients on the flow, the Richardson-Richards equation governing flow in a non-rigid, 3-D variably saturated, spatially heterogeneous soil (Russo et al., 2004) is:

$$\frac{\partial(\theta^r \theta)}{\partial t} = \sum_{i=1}^3 \frac{\partial}{\partial x_i} \left[K^r K \frac{\partial \psi}{\partial x_i} \right] - \frac{\partial K^r K}{\partial x_1} - S_w \quad (18)$$

where t is time; $\psi = \psi(\underline{x}, t)$ is the pressure head; $\theta = \theta(\psi, \underline{x})$ and the scalar $K = K(\psi, \underline{x})$, are the volumetric water content and the hydraulic conductivity, respectively, of the soil in a rigid “stable” reference state, in which the changes in the soil PSD due to SS-SM interactions are negligibly small (i.e., when $R \rightarrow 0$, and $C_0 \rightarrow \infty$); $\theta^r = \theta^r(C_0, R, \psi, \underline{x})$ and $K^r = K^r(C_0, R, \psi, \underline{x})$ are relative water content, and relative hydraulic conductivity, respectively, resulting from the effect of the SS-SM interactions on the soil PSD, and $S_w = S_w(\underline{x}, t)$ is a sink term representing water uptake by the plant’s roots.

Similarly, assuming that the entire water-filled pore space is mobile, neglecting ion precipitation or dissolution, and neglecting ion uptake by the plant’s roots, the ADE governing transport of the mixed Na/Ca salt (Russo et al., 2004) is:

$$\frac{\partial(\theta_m^* c_m)}{\partial t} = \sum_{i=1}^3 \sum_{j=1}^3 \frac{\partial}{\partial x_i} \left\{ \theta D_{ij} \frac{\partial c_m}{\partial x_j} \right\} - \sum_{i=1}^3 \frac{\partial(u_i \theta c_m)}{\partial x_i} \quad (19)$$

where $m = 1$ to 3 indicates the chloride, sodium and calcium ions, respectively; c_m [mol ℓ^{-1}] is the resident solute concentration of the m -th ion, expressed as mass per unit volume of soil solution; θ_m^* is the “effective” volumetric water content, given by $\theta_m^* = \theta + \theta_{\text{exm}}$ and $\theta_m^* = \theta + \theta_{\text{adm}}$, for $m = 1$ and $m = 2, 3$, respectively, where $\theta_{\text{exm}}(\psi, \underline{x}, R, C_0, c_m)$ ($m = 1$) and $\theta_{\text{adm}}(\psi, \underline{x}, R, C_0, c_m)$ ($m = 2, 3$) are the exclusion and the adsorption volume fractions, respectively, calculated from the planar, mixed-ion diffuse-

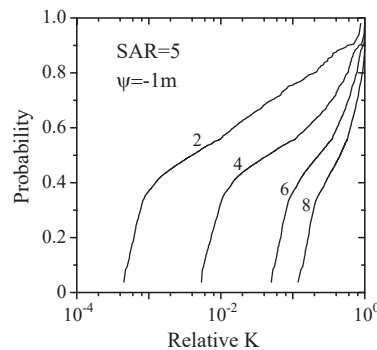


Fig. 7 Cumulative distribution function of relative hydraulic conductivity, K_r , for different chloride concentrations (denoted by the numbers labeling the curves, in $\text{meq} \ell^{-1}$). Reproduced from Russo D (2013) Consequences of salinity-induced-time-dependent soil hydraulic properties on flow and transport in salt-affected soils. *Procedia Environmental Sciences* 19: 623–632, <https://doi.org/10.1016/j.proenv.2013.06.071>.

double-layer theory; u_i ($i = 1, 2, 3$) are the components of the Eulerian water velocity vector, and D_{ij} ($i, j = 1, 2, 3$) are the components of the pore-scale dispersion tensor. When molecular diffusion is small enough to be excluded from the analysis, D_{ij} given (Bear, 1972) as:

$$D_{ij} = \lambda_T |\underline{u}| \delta_{ij} + (\lambda_L - \lambda_T) u_i u_j / |\underline{u}| \quad (20)$$

where λ_L and λ_T are the longitudinal and the transverse components of the pore-scale dispersivity tensor; δ_{ij} is the Kronecker delta, and $|\underline{u}| = (u_1^2 + u_2^2 + u_3^2)^{1/2}$.

Details of the numerical methods used to solve the governing equations (18, 19), subject to the appropriate boundary and initial conditions are described elsewhere (Russo et al., 2004). The “mixed” form of Richardson-Richards equation governing flow in a three-dimensional domain (Eq. 18) is approximated by a fully implicit Euler scheme with a truncation error of $O(\Delta t, \Delta x_1^2, \Delta x_2^2, \Delta x_3^2)$. This scheme is convergent and unconditionally stable for the linear diffusion equation and has been found appropriate for flow problems in highly heterogeneous porous media, in which both steep gradients and saturated regions can develop.

The resulting system of nonlinear algebraic equations with respect to the pressure head, ψ , is solved iteratively, by applying the so-called modified Picard method. Picard (external) iterations are applied for both capillary and gravity terms and the resulting system of linear algebraic equations is solved by the polynomial preconditioned conjugate gradient (PPCG) method. The hydraulic conductivity between the numerical cells, the so-called inter-block conductivity employed in the solution, is estimated by means of a modified version of an asymptotic weighting scheme. The modified scheme avoids the smearing of steep wetting fronts and does not create oscillations for gravity-dominated flow, and is therefore appropriate for flow problems in highly heterogeneous flow systems associated with steep gradients or saturated regions.

The numerical scheme employed for the solution of Eq. (19) is designed to minimize numerical dispersion for advection-dominated transport problems. The scheme can be utilized even in the purely advective case ($Pe \rightarrow \infty$), where Pe is the grid Péclet number, introducing some limited numerical dispersion that smears a sharp concentration front (e.g., $c = 1$ to $c = 0$ step) over ~ 3 numerical cells in the 1-D case. The iterative scheme used to couple the $m = 3$ sets of Eq. (19), and the iterative solution method needed because of the coupling between the flow equation and the transport equations through the dependence of $K(\psi)$ and $\theta(\psi)$ on C_0 and R , are described elsewhere (Russo et al., 2004).

A case study

The physical domain and the simulated flow scenario

The case study considered here involved a citrus orchard planted on a structured clay soil in a region with typical Mediterranean climate characterized by a distinct rainy period during the winter and a long dry season over the rest of the year that necessitates multiple irrigation events. Realistic features of the flow domain, taken into account in the simulation, include the irrigation water quantity and quality, the spatial pattern of the irrigation system at the soil surface, the spatial variability of the relevant soil properties and the flow-controlled attributes, the spatial pattern of the trees' roots, and the time-dependent atmospheric forcing conditions imposed on the soil surface.

The flow scenario started at the beginning of the irrigation season (April 1st) and proceeded for 1 year. The irrigation water has moderate values of chloride concentration, $C = 10$ meq/l, and sodium adsorption ratio, $SAR = 5.7(\text{meq/l})^{1/2}$, while the rainwater has small values, i.e., $C = 1.4$ meq/l, and $SAR = 0.5(\text{meq/l})^{1/2}$. A pulse ($t_0 = 0.05\text{d}$) of a tracer solute was applied during the first irrigation event.

Results and discussion

Following Russo (2013), profiles of the horizontally-averaged mean and standard deviation (SD) of the chloride concentration, C , and the soil ESP = $f(h, SAR, C)$, for three elapsed times, for both the interacting and the inert cases, are depicted in Fig. 8a and b. These times correspond to the middle of the irrigation season, the transition period between the irrigation season and the rainy season, and the end of the rainy season, respectively. Fig. 8a suggests that the differences between the sets of both the mean and the SD profiles of C are minimal during the transition period. Most significant differences between the two sets occur during the rainy season. At the end of this season, in the inert reference case, the mean C profiles exhibit the symmetrical Fickian distribution with concentration at the soil surface that approaches its counterpart under rainwater. In contrast, for the interacting case, the mean profiles, and, particularly, the SD profiles, exhibit shallower concentration peaks, and, significant secondary concentration peaks close to the soil surface.

Fig. 8b demonstrates that the Na/Ca exchange is a very slow process compared to solute movement (Fig. 8a). This stems from the nonlinearity of the Na/Ca exchange isotherm, which, in turn, increases with increasing SAR and decreasing C . Consequently, both the increase in the soil ESP during the irrigation season, and the decrease in the soil ESP during the rainy season are confined to the upper part (≈ 1 m) of the soil profile. Fig. 8b suggests that the differences between the two sets of the profiles of the mean and the SD of the soil ESP are most substantial during the rainy season. In the inert case, both the mean and the SD of the soil ESP close to the

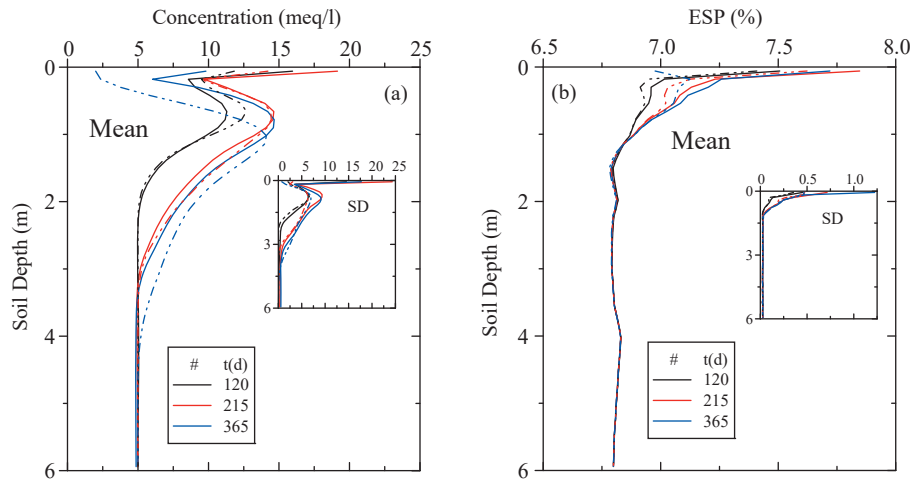


Fig. 8 Profiles of the mean values and the standard deviations (SD) of the soil solution chloride concentration, C , (a) and the soil exchangeable sodium percentage, ESP (b), for different elapsed times. Reproduced from Russo D (2013) Consequences of salinity-induced-time-dependent soil hydraulic properties on flow and transport in salt-affected soils. *Procedia Environmental Sciences* 19: 623–632, <https://doi.org/10.1016/j.proenv.2013.06.071>.

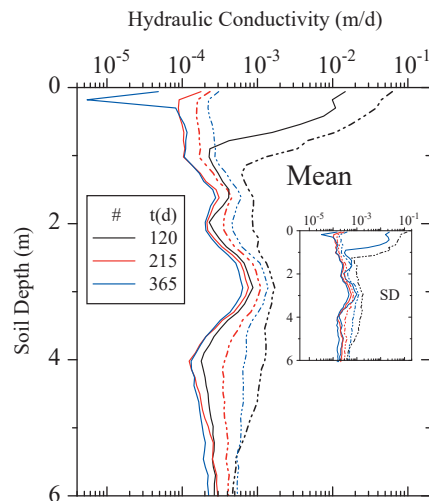


Fig. 9 Profiles of the mean values and the standard deviations (SD) of the soil hydraulic conductivity, K , for different elapsed times. Reproduced from Russo D (2013) Consequences of salinity-induced-time-dependent soil hydraulic properties on flow and transport in salt-affected soils. *Procedia Environmental Sciences* 19: 623–632, <https://doi.org/10.1016/j.proenv.2013.06.071>.

soil surface decrease substantially during the rainy season, while their counterparts associated with the interacting case only slightly decrease.

Profiles of the horizontally-averaged mean and SD of the soil hydraulic conductivity, K , at three elapsed times (as in the previous figures) are depicted in Fig. 9; profiles of the mean and the SD of K associated with the concentration-dependent flow system (solid lines) are compared with their inert, reference counterparts (dashed-dotted lines). Fig. 9 demonstrates the substantial decrease in the mean K with increasing time, particularly in the upper part of the soil profile. Because the effect of the SS-SM interactions on the soil PSD is enhanced by a combination of low-concentration and high-SAR soil solution, the decrease in the mean K is particularly substantial during the rainy season, when the soil solution is diluted of salts, and less during the irrigation seasons when salts in the soil solution are more concentrated.

The movement and spread of a tracer solute in the spatially heterogeneous soil may serve as a measure of the capability of the soil to transfer water and solutes. Important characteristics of the tracer transport are the time-dependent, spatial moments of its concentration field, which, in turn, provide measures of the mass, location and spread of the tracer plume. The first normalized spatial moment about the origin, defines the coordinate location of the centroid of the tracer solute, $R_i(t)$, ($i = 1, 2, 3$), and is used to define the effective solute velocity vector, $\underline{V}_i$, as the time rate of change of the displacement of the centroid of the tracer

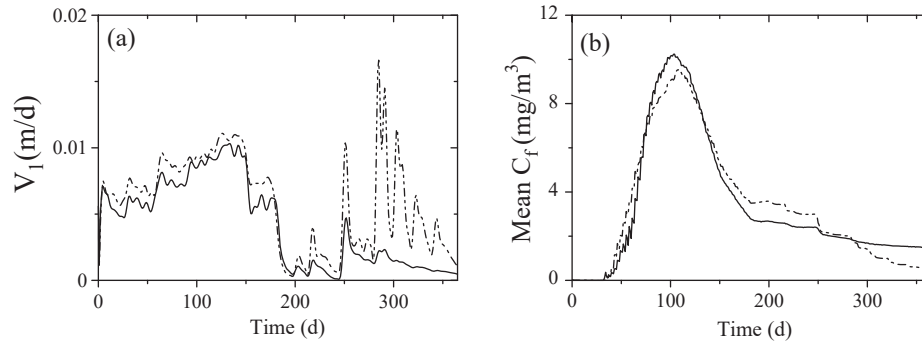


Fig. 10 Longitudinal component of the tracer solute effective velocity, V_1 , (a), and mean solute tracer flux concentration, C_f , at a horizontal CP located at vertical distance, L_1 , from the injection zone at the soil surface (b), as functions of time. Reproduced from Russo D (2013) Consequences of salinity-induced-time-dependent soil hydraulic properties on flow and transport in salt-affected soils. *Procedia Environmental Sciences* 19: 623–632, <https://doi.org/10.1016/j.proenv.2013.06.071>.

solute, i.e., $\underline{V}(t) = [dR_1/dt, dR_2/dt, dR_3/dt]^T$. The longitudinal component of $\underline{V}$, V_1 , is depicted as a function of time in Fig. 10a for both the inert and the interacting cases. The decrease in $V_1(t)$ associated with the interacting case as compared with the inert case, is attributed to the reduction in the hydraulic conductivity caused by the rearrangement of the soil's PSD; clearly the effective solute velocity substantially decreases after the first few rainstorms during the rainy season.

Another characteristic tracer transport relationship is the solute mean breakthrough curve (BTC). Consider the case of an instantaneous pulse of unit-mass, uniformly injected at the soil surface. The tracer solute mean BTC at a given horizontal control plane (CP) located at distance, L , from the soil surface, represents the travel time, τ , probability density function (PDF), $f(\tau; L)$, which, in turn, is a global descriptor of the response of the flow system to a contamination or salinization event. Mean solute tracer BTC as a function of time is depicted in Fig. 10b for both the inert and the interacting cases. This figure clearly demonstrates that changes in the transport properties of the soil related to soil sodification coupled with diluted soil solution, lead to a highly skewed mean solute tracer BTC, which exhibits an exceedingly large tail.

It should be emphasized that the effect of the SS-SM interactions on the soil hydraulic properties is essentially reversible. The relatively concentrated soil solution during the irrigation season may compensate in part for the negative effects of the SAR, which, in turn, increases slowly. As a result, the soil hydraulic conductivity may recover during the irrigation seasons. This reversible, cyclic behavior, however, may persist as long as the soil ESP does not exceed a soil-dependent, threshold value, ESPc. When the soil ESP exceeds ESPc, a breakdown of the clay particle domains, and the subsequent relocation of the clay particles, may lead to a continuous decrease in the soil hydraulic conductivity. Eventually, this may lead to an unrecoverable situation in which most of the soil surface is essentially sealed.

Concluding remarks

Salinity has many physical impacts on soil, particularly when the clay fraction is dominated by smectite minerals (e.g., montmorillonite), with solutions containing Na, Ca and Cl ions. These conditions are used in this article, with the analysis based on several simplifying assumptions, considering processes and entities at three different length-scales: (i) the micro (pore) scale, i.e., the physicochemical interactions between the soil solution and the soil matrix, the structure of the clay particles, and the soil PSD; (ii) the local (Darcy) scale, i.e., water flow and solute transport and the constitutive relationships for unsaturated flow; (iii) the field scale, i.e., the spatial structure of the heterogeneous soil and the statistics of relevant soil properties and flow-controlled attributes.

The results of the analysis presented in this chapter suggest that SS-SM interactions may alter the geometry of the soil pores, and, thus may affect the soil hydraulic conductivity, $K(h)$, and soil water retention, $\theta(h)$, functions. The effect of these interactions on $K(h)$ and $\theta(h)$, the retardation factor, R_f , and the elution factor, E_f , and, concurrently, on water flow and solute transport may be significant. This effect generally increases as the soil solution's sodium adsorption ratio (SAR) increases; as both the soil solution concentration and the capillary pressure head decrease; as the clay fraction of the soil increases; and as the soil-texture becomes finer.

The case study and the results of the simulations presented in this chapter are relevant to fine-textured soils whose clay fraction is dominated by montmorillonite. The results of the simulations suggest that enhanced SS-SM interactions, induced by increasing soil exchangeable sodium during the irrigation season and dilution of the soil solution during the rainy season, may considerably reduce soil hydraulic conductivity during the rainy season, particularly, close to the soil surface, and, concurrently, may reduce water velocity. Consequently, enhanced SS-SM interactions, may greatly slow down the solute movement, may substantially increase the tailing of the mean solute breakthrough at a given horizontal control plane (CP), and may change markedly the pattern of the solute concentration profiles.

Inasmuch as the effect of the SS-SM interactions on the soil hydraulic properties is essentially reversible, the relatively concentrated soil solution during the irrigation seasons may compensate in part for the negative effects of the SAR, which, in turn, increases slowly; consequently, during the irrigation seasons, the soil hydraulic conductivity may recover. This reversible, cyclic

behaviour, however, may persist as long as the soil ESP does not exceed a soil-dependent, threshold value, ESP_c . Additional simulations (not shown here) suggest that the value of ESP_c was not reached during a period of 7 years, allowing the aforementioned cyclic behaviour to persist. This is expected in view of the values of C and SAR employed in the simulations, which, in turn, are typical of treated-wastewaters used for irrigation.

References

- Assouline S, Russo D, Silber A, and Or D (2015) Balancing water scarcity and quality for sustainable irrigated agriculture. *Water Resources Research* 51. <https://doi.org/10.1002/2015WR017071>.
- Bear J (1972) *Dynamics of Fluids in Porous Media*, p. 764. New York: Elsevier.
- Blackmore AV and Miller RD (1961) Tactoid size and osmotic swelling in Ca- montmorillonite. *Soil Science Society of America Proceedings* 25: 169–173.
- Bresler E (1972) Interacting diffuse layers in mixed mono-divalent ionic systems. *Soil Science Society of America Proceedings* 36: 891–896.
- Bresler E, McNeal BL, and Carter DL (1982) *Saline and Sodic Soils: Principles-Dynamics-Modeling*. Berlin Heidelberg and New York: Springer. 236 pp.
- Brutsaert W (1967) Some methods of calculated unsaturated permeability. *Transactions of ASAE* 10: 400–404.
- Childs EC and Collis-George N (1950) The permeability of porous materials. *Proceedings of the Royal Society of London, Series A* 201: 392–405.
- Hamilton AJ, Stagnitti F, Xiong X, Kreidl SL, Benke KK, and Maher P (2007) Wastewater irrigation: The state of play. *Vadose Zone Journal* 6: 823–840. <https://doi.org/10.2136/vzj2007.0026>.
- Marshall TJ (1958) A relation between permeability and size distribution of pores. *Journal of Soil Science* 9: 1–8.
- McNeal BL (1968) Prediction of the effect of mixed salt solutions on soil hydraulic conductivity. *Soil Science Society of America Proceedings* 32: 190–193.
- Qadir M, Wichnells D, Raschid-Sally L, Singh Minhas P, Drechsel P, Bahri A, and McCormick P (2007) Agricultural use of marginal-quality water and challenges. In: Molden D (ed.) *Water for Food; Water for Life. A Comprehensive Assessment of Water Management in Agriculture*, p. 425. London, U. K: Earthscan.
- Russo D (1988) Numerical analysis of the non-steady transport of interacting solutes through unsaturated soil. 1. Homogeneous systems. *Water Resources Research* 24: 271–284.
- Russo D (2010) Analysis of transport of mixed Na/Ca salts in a three-dimensional heterogeneous variably saturated soil. In: Levy GJ, Bar-Tal A, and Fine P (eds.) *Treated Wastewater in Agriculture: Use and impacts on the soil environments and crops* 14, pp. 418–436. New Jersey: Willey-Blackwell Publishing Ltd.
- Russo D (2013) Consequences of salinity-induced-time-dependent soil hydraulic properties on flow and transport in salt-affected soils. *Procedia Environmental Sciences* 19: 623–632. <https://doi.org/10.1016/j.proenv.2013.06.071>.
- Russo D and Bresler E (1977a) Effect of mixed Na-Ca solutions on the hydraulic properties of unsaturated soils. *Soil Science Society of America Journal* 41: 714–717.
- Russo D and Bresler E (1977b) Analysis of the saturated-unsaturated hydraulic conductivity in a mixed Na-Ca soil system. *Soil Science Society of America Journal* 41: 706–710.
- Russo D, Zaidel J, and Laufer A (2004) Numerical analysis of transport of interacting solutes in a three-dimensional unsaturated heterogeneous soil. *Vadose Zone Journal* 3: 1286–1299.
- Shainberg I (1984) The effects of electrolyte concentration on the hydraulic properties of sodic soils. In: Shainberg I and Shalhevet J (eds.) *Soil Salinity Under Irrigation: Processes and Management*, pp. 49–64. New York: Springer-Verlag.
- Shainberg I and Kaiserman A (1969) Kinetics of the formation and breakdown of Ca-montmorillonite tactoids. *Soil Science Society of America Proceedings* 33: 547–551.

Shrink-Swell processes in soils

Pascal Boivin, University of Applied Sciences of Western Switzerland, HES-SO, HEPIA-Agronomy, Geneva, Switzerland

© 2023 Elsevier Ltd. All rights reserved.

This is an update of D. Smiles, P.A.C. Raats, Swelling and shrinking, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 115–123, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00425-2>.

Introduction	139
Origin of soil shrinkage	140
Basics of soil shrinkage	141
Shrinkage measurement	142
Shrinkage modeling	142
Properties and processes affecting shrinkage curves	143
Significance	143
Perspectives	143
References	144
Further reading	145

Abstract

Soil shrinking and swelling with water, leading to crack formation, is the primary abiotic soil structure forming process. Shrinking is mostly due to the soil colloids forming the soil *plasma*, which produces a characteristic shrinkage curve (SC) on drying. Soil shrinkage has physical, geotechnical and pedological applications with relevance to (i) its impact on soil hydrodynamics (preferential flow in cracks and changes in the hydraulic characteristics) and (ii) improved impact diagnosis and upscaling of physical properties by deterministic SC modeling. In natural soils, organic matter and aggregation affect shrinkage and therefore the development of soil structure by drying and wetting.

Glossary

COLE Coefficient of linear extension.

Plasma That part of the soil solid material that is easily deformed, reorganized or concentrated by the processes of soil formation. It includes all the material, mineral or organic, of colloidal size and relatively soluble material that is not contained in the skeleton grains.

Structural porosity Pore system made of the biopores, the cracks and the packing voids.

Key points

- All soils containing colloids show a shrinkage curve
- The characteristic shrinkage curve (SC) on drying should be considered in the hydrodynamics of non-rigid soils
- SC deterministic modeling allows characterization of the plasma and structural pore systems
- SC analysis allows improvement in the soil physical characterization of the soil
- Applications in hydrodynamics: non-rigid soils, by-pass flow
- Applications in structure characterization, compaction diagnosis
- Spatialization and dynamics of soil physical properties

Introduction

Most soils show shrink-swell behavior, induced by the forces resulting from changes in water content and their impact on pore volumes. The extent of soil shrinkage is generally defined as the bulk density or specific volume change of soil relative to its water content (Haines, 1923). It is, therefore, a physical property of undisturbed soil structure that has a large impact on a range of soil processes, especially pore structure dynamics and transport. The volume change with water content is due to the shrinking and swelling of soil constituents and their structural arrangement.

Soil shrinkage has been studied for more than a century and was first considered as an indicator of soil structural stability. It is evident in some soils by the creation of cracks, which is considered as the primary and major abiotic process of soil structure

formation (see Soil Structure section). Vertisols, for instance, shrink extensively during drying, which provides one of the key diagnostic features of their classification. During the 20th century, the scope of interest in soil shrinkage widely increased in soil science and civil engineering. In engineering applications, the question is mostly related to heavy swelling-clay soils whose shrinking properties may severely damage buildings. In soil science, the investigations covered a wide range of topics, ranging from a fundamental understanding of pedogenic processes to applications in understanding water flow, structural stability or pore structure dynamics. Soil scientists have explored the role of soil constituents in soil shrinkage, observing a characteristic shrinkage curve (SC) that relates specific volume (or bulk density) of the soil to its water content. Using the shrinkage curve in modeling and analysis has applications in soil physics and soil physical characterization from the aggregate to field scale. There is capacity with soil shrinkage modeling to bridge the knowledge on soil fabric, soil constituents and soil hydrodynamics. In shrinking soils, hydrodynamics is affected by both matric potential draining soil pores and changes to the size of pores from shrinkage. Although the former process is studied extensively in soil physics, the second process is often overlooked, so soil shrinkage modeling plays an important role in overcoming a major limitation faced by soil physics.

Origin of soil shrinkage

If we revisit the example of Vertisols, their large macroscopic shrink-swell behavior is closely related to their large content of swelling clays (see Pedology section in this Encyclopedia). Seasonally dependent crack networks develop in these soils with cycles of wetting and drying. Smaller concentrations of swelling clays and other colloids observed in other soils results in less shrinkage, leading to smaller size cracks whose network may only be observable on thin sections. Mineral soils are formed of voids, skeleton grains, and colloids, namely clay minerals bound to soil organic carbon (SOC) and oxides forming the soil *plasma* (*"Glossary of Soil Science Terms | Soil Science Society of America,"*, 2020). The plasma consists primarily of domains of silt- and clay-size particles coated with SOC and oxides. It comprises small (nm to μm) interparticle pores. Because clay particles are major constituents of the plasma, it is also referred to as the clay matrix, clay phase, or organo-mineral complex. Clay minerals and SOC shrink and swell with changes in water content. The shrinkage behavior of clay pastes has been well described and depends on the clay mineral properties (e.g., Tessier, 1990). The shrinkage of the soil, therefore, depends on the clay content, clay type, SOC and oxides acting as binding elements, thus limiting the shrinkage of the plasma (Boivin et al., 2009, 2004; Goutal-Pousse et al., 2016).

As a structured medium, the soil has a porosity. Various classifications of soil pores based on size classes have been proposed (see Soil Structure section). Size classes are user, tool, and scale dependent. Therefore, they do not reflect intrinsic properties of the structured soil and its materials. Soil shrinkage occurs at the expense of soil porosity and is, therefore, related to the nature and behavior of the soil pores. Soil pore size distribution, shape, and connectivity have been investigated for their implications in transport properties or as habitat of soil biota. At its simplest, a rigid, homogeneous, and uniform pore network has been assumed in water transport modeling, but structural complexity of the pore network has become increasingly appreciated and feasible to model with higher speed computers. The deformation of the structure, especially its changes due to shrinkage, is often overlooked.

The solid material (either minerals or organic matter) and pore system together constitute the soil fabric. The physical structure of the soil fabric was long ago described by micromorphologists (e.g., Brewer, 1964), which allowed identification of two categories of soil pore systems, based on their origin, shape and function; namely the structural pores and the plasma pores. Plasma pores are equivalent to textural porosity, which refers to the clay pores and the lacunar voids between skeleton grains and the clay matrix. Taking this approach, the structural porosity includes cracks, biopores, packing voids and vughs. Because of their origin, plasma pores are sub- μm , while the volume of structural pores is mostly represented by pores larger than $5\ \mu\text{m}$.

The structural and plasma pores show distinct shrink-swell behavior upon changes in soil water content. From saturation to the *Air Entry* (AE) point, no air entry occurs in plasma pores (Tessier, 1990). Therefore, the plasma pores remain saturated along the corresponding water content range, and the clay paste volume follows a 1:1 line with the water content (Fig. 1), until AE is reached,

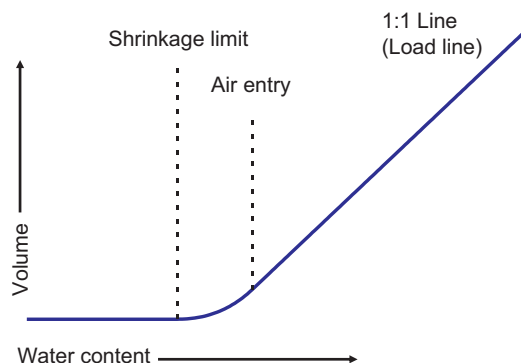


Fig. 1 Schematic shrinkage curve of a clay paste.

which may account for a large soil shrinkage in swelling clay soils. In contrast, air enters into the structural pores upon soil drying, and water menisci exert a shrinkage pressure on the soil by Jurin's law of capillarity and approximated by the Laplace equation.

Basics of soil shrinkage

Soil shrinkage was studied early in the early 20th century for its impact on soil structure, including disputed roles of colloidal coatings (Haines, 1923; Hardy, 1923). Later, soil shrinkage was investigated for geotechnical problems like foundation integrity or in soil science as an indicator of structure stability (McGarry and Daniells, 1987). The concept of linear extensibility was introduced to characterize the one-dimension change of soil with water content, which is usually quantified with the Coefficient of Linear Extensibility (COLE) (Grossman et al., 1968):

$$COLE = \frac{L_m}{L_d} - 1 \quad (1)$$

where L_m is the length of the moist soil unit and L_d is the length of the dry soil unit. COLE indicates the potential of the soil to shrink and swell with water. As the extent of shrinkage can result in soil layer cracking and soil subsidence, there is a great interest in measuring COLE for engineering or preferential flow modeling (Coppola et al., 2015). COLE is used in soil classification to categorize shrinking soils such as Vertisols, which require at least 0.06 mm m^{-1} linear shrinkage in their characterization. Testing of COLE is simple. It involves filling a trough of soil with a remolded soil paste and then measuring the length change after the soil has been dried. However, shrinkage is a volumetric, so the volume change is related to the COLE by:

$$\frac{dV}{V} = \left(\frac{dL}{L}\right)^n \quad (2)$$

with V , the soil specific volume ($\text{cm}^3 \text{g}^{-1}$), L the soil length, and n the geometric factor (Towner, 1986) with $n = 3$ in case of isotropic shrinkage.

Using a remolded paste, however, destroys soil structure. Because of the strong relationship between soil structure and soil SC, COLE is much more descriptive of realistic field conditions if it is measured on undisturbed soil, contrary to what is sometimes practiced in geotechnical engineering for clayey materials.

COLE measures the extremes of shrinkage between maximum wetness and dryness. The availability of quasi continuous SC measurement techniques over a range of water contents on undisturbed soil cores allows for the general shape of the soil SC to be split into different linear domains with curvilinear transitions (Braudeau et al., 2004) (Fig. 2). These main domains, from water saturation to air dry, are the structural, basic and residual shrinkage domains. In some soils, however, an additional swelling domain is observed close to water saturation and was interpreted as a property of poorly aggregated soil. Another sharp shrinkage domain has been observed near to the air-dry state in Swiss and Canadian soils with large SOC content (unpublished). Despite these possible additional domains at the two ends of the SC, most of the water content change is observed from structural to residual

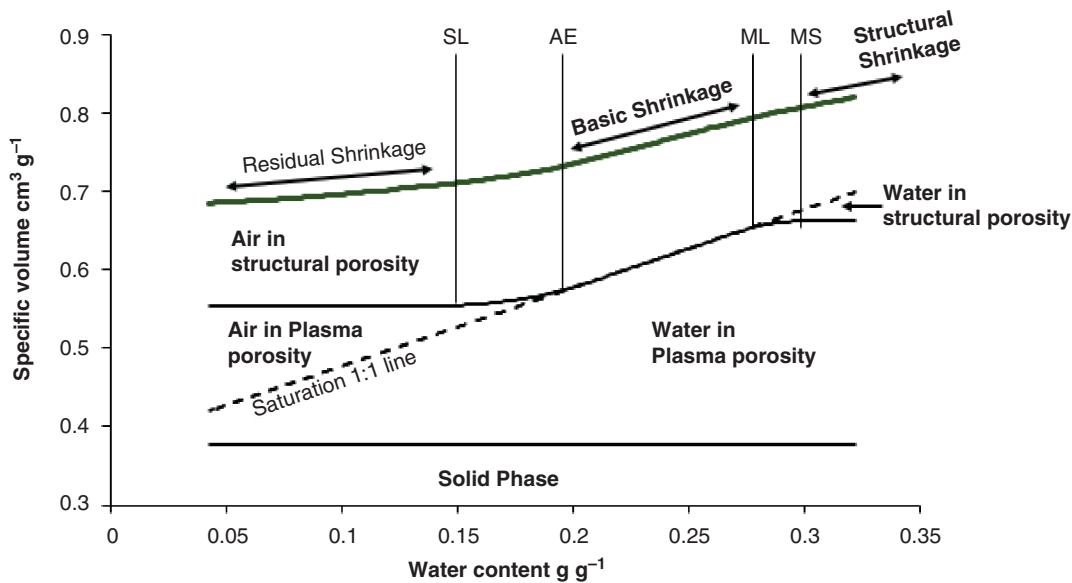


Fig. 2 Cambisol shrinkage curve showing the typical S shape, soil shrinkage domains, water saturation line and distribution of air and water in the pore systems.

shrinkage. The magnitude of these domains depends on soil type and colloid content, though the S-shape is characteristic of many soils, e.g., in Vertisols, Calcisols, Cambisols, Luvisols, Ferralsols etc.

The structural shrinkage occurs primarily due to the drainage of structural pores. Its slope has been related to soil structural stability since it depicts the capacity of the structure to withstand the menisci forces. The basic shrinkage (Fig. 2), also referred to as proportional shrinkage (Mitchell, 1992), is related to the proportional shrinkage of the clay. For clay, this slope is equal to 1 (Fig. 1), but for structured soil it usually ranges from less than 0.1 to more than 1.5, depending of the tendency of the structure to collapse or to remain rigid upon plasma shrinking. Therefore, for the same plasma shrinkage, the lower the soil shrinkage slope, the denser the crack network in the soil and the greater its volume.

Shrinkage measurement

Soil shrinkage can be measured in situ using displacement transducers that record the movement of gauges inserted at different depths (Coquet et al., 1998), though this is technically demanding and difficult to interpret. However, shrinkage is mostly measured on extracted undisturbed cores or clods. Once soil is removed from its environment, however, the overburden pressures of the overlying soil layers is not accounted for. Different devices have been proposed in the last decades to measure the shrinkage of clods, allowing high accuracy and quasi continuously measurement of SC. Examples include laser barriers (Braudeau et al., 1999), displacement transducers (Fig. 3) (Boivin et al., 2004; Schindler et al., 2015) and X-Ray CT (Peth et al., 2010). Linear or circumference changes are converted into volume change using Eq. (2), with estimation of the geometric factor when the soil wet and dry volumes are measured. The geometric factor of soil cores has been generally reported to be close to 3 on average, though its value changes along the SC (Boivin, 2007).

Shrinkage modeling

Shrinkage modeling was performed early in soil science, with the aim of quantifying the structural stability of the soil. In past decades models were proposed that took benefit from the high quality SCs obtained with quasi-continuous SC measurement devices. Two categories can be distinguished. One category of equations aims to reproduce quantitatively the SC, for instance to account for its influence on hydraulic conductivity (Horn et al., 2014). Due to the characteristic S-shaped curve of the SC, one approach modified the Van Genuchten equation. Another category is deterministic models based on considerations of the nature and behavior of soil pores, accounting for the properties of plasma and structural pores with decreasing water content, as presented above. In the first category of equations, the number of parameters conditions the ability of the model to reproduce the observation, which opposes the objective of using a minimum number parameter.

Deterministic models have been proposed from particle scale to layer scale, with special emphasis on clod scale. The most developed models (Braudeau et al., 2004, 1999; Chertkov, 2012) are based on common concepts, namely soil SC accounting for the combined effect of shrinkage of plasma (that can be referred to as clay matrix or microporosity) and structural pores. The shrinkage of plasma is considered, as depicted in Fig. 1, and that of the structural porosity, including cracks, drains with soil drying with limited deformation, being accounted for by the slope of the structural shrinkage (Fig. 2). To our best knowledge, model comparisons have rarely been performed on high resolution SCs (Boivin et al., 2006), therefore, we will not discuss this aspect further.

In Boivin (2007), the clay pastes sub-model of Chertkov (2012) and Braudeau et al. (1999) were found to be extremely close, thus implying a similar modeling of the complementary structural porosity volume for successful fitting. However, only the XP model (Braudeau et al., 1999) has been successfully applied to a large number of undisturbed sample series from different soil types. The thermodynamic formalism of the soil shrinkage proposed by Braudeau et al. (2014) is based on previous research starting with Sposito (1972). Based on the two-pore systems description, it yields results close to those obtained with XP. Note that overburden is not taken into account in this model, thus applicable to isolated soil structure units.

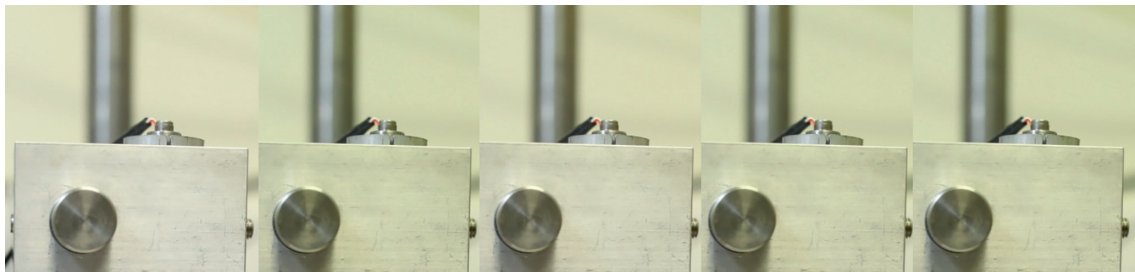


Fig. 3 Undisturbed cambisol sample shrinking from -10 hPa (left) to air-dry (right) in a shrinkage apparatus (Schäffer et al., 2008). Note the displacement transducer on the top of the sample (picture C. Deluz).

Properties and processes affecting shrinkage curves

Soil shrinkage is affected by clay type, clay content, iron and aluminum oxides, SOC content (see above) and carbonate content (Zolfaghari et al., 2016). These factors show linear relationships with the shrinkage parameters as determined with the XP model, or modified Van Genuchten equation of shrinkage. For instance, shrinkage increases with clay content and decreases with SOC and iron oxides contents, whereas the structural pore volume and the readily available water increase with SOC content. In that respect, the spatial variability of soil colloids allows one to account for the spatial variability of soil physical properties. Moreover, the coefficients of variation (CV) of the SC parameters (XP model) are very small (from 7% to 14%, Boivin, 2007; Boivin et al., 2006) compared with CVs for most soil physical parameters reported, thus opening the door to effective mapping of the soil shrinkage properties and derived pore systems parameters.

Soil shrinkage is also affected by external factors such as cropping practices, trafficking or biological activity, and shrinkage modeling has been used to determine structure degradation indicators (Johannes et al., 2019), to characterize tillage effects (Mallory et al., 2011), soil compaction (Peng et al., 2009) and structure recovery (Fell et al., 2018; Goutal-Pousse et al., 2016), or the effect of soil biota on soil structure (Milleret et al., 2009). Fitting XP on soil SCs allowed a distinction to be made between the effect of soil constituents, their variability, and factors such as trafficking on biological activity, thus allowing for improved field diagnosis.

Significance

Many developments in soil physics require assumptions about the pore network, such as homogeneity and rigidity, to simplify the numerical solution of the equations. These assumptions oppose the known behavior of structured soils containing colloids, i.e., non homogeneous and uniform pore network, and geometry changes with water content. This has raised many issues in application, particularly in parameter estimation and mapping. The physical properties, as determined according to mechanics theories, exhibit large variability, unstable variances with measurement device size, and poor correlation with the distribution of soil constituents. The developments of pedotransfer functions (Bouma and van Lanen, 1987) are aimed at bridging the resulting gap between soil physical properties and easy to map soil properties, in particular colloid content, by establishing local correlations. To some extent this can be considered doomed to failure owing to the fundamental contradiction between soil mechanics assumptions and soil properties. With that respect, one may consider the development of shrinkage modeling as a promising way to bridge this gap.

Fluid transfer modeling equations including the shrink-swell properties of the soil were proposed (Garnier et al., 1997). However, the soil volume change with water remains seldom considered in fluid transfer modeling despite its strong impact (Horn et al., 2014). The introduction of shrink-swell properties in water transfer models allows the dynamic accounting of preferential flow at layer scale (Coppola et al., 2015). A full modeling method including (i) overburden (Smiles and Raats, 2005), (ii) preferential flow in cracks and (iii) the dual porosity (accounting for transfers between plasma aggregates and structural pores) is still to be formulated.

Significant progress in soil structure characterization has been gained with shrinkage analysis. In the context of climate change, and SOC depletion leading to poor quality soils, it is of key importance to improve our capacity to monitor and model changes in soil physical properties. For instance, shrinkage analysis allowed to emphasize the role of SOC in soil structure and soil porosity development (Boivin et al., 2009), to propose improved soil compaction diagnosis (Goutal-Pousse et al., 2016; Schäffer et al., 2008) and effective structure degradation indicators (Johannes et al., 2019).

Perspectives

In re-designing water transfer models at profile scale, taking into account soil shrinkage is essential. Addressing this challenge requires growing interest in developing deterministic shrinkage models, which include the two-soil pore-systems (structural and plasmic) and their response to changes in soil constituents (particularly SOC content). This opens the door to models that are capable of dynamic adaptation of soil physical properties to, e.g., soil management, and to sharply improve model parameter specialization.

To do so, research is needed that can fill the known and suspected knowledge-gaps remaining. For example, the deterministic models of shrinkage include a transfer of water from plasma porosity to structural porosity, which needs to be further investigated (Barbosa et al., 2022), and overburden should be included in the description of water potential and confining stresses from the surrounding soils that would be found in the field.

Another perspective is to improve our understanding of soil complex processes. The role of structural porosity in microorganism activity and the related carbon or nitrogen cycles has started to be deciphered, in particular with imaging technology (Kravchenko et al., 2019). Changes to soil pore structure with shrinkage is likely to have a large impact on microbial habitats, activity and physical protection. Such a fundamental knowledge is key to our understanding of soil ecosystem services management, particularly in cropland, but it is hard to study in situ and at large scale. On the other hand, shrinkage analysis does not allow the capture of fine structural pores architecture, but it does allow precise, inexpensive quantification of fine structural pore volume, size distribution,

and their relationships with constituents and practices at large scale. Therefore, research on the links between fine structural porosity architecture and the shrinkage signature of soil structure should be worked out. The impacts extend beyond physical processes, including the capacity of soil pores to support biological activity and diversity, but shrinkage impacts have been overlooked.

Finally, though deterministic modeling of soil SC seems more promising, the models are based on assumptions that have been only partially validated, at best. If the above perspectives stand, it is of critical importance to progress in deterministic SC modeling associated with a fully validated physical formalism of the underlying processes.

References

- Barbosa LAP, Gerke KM, and Gerke HH (2022) Modelling of soil mechanical stability and hydraulic permeability of the interface between coated biopore and matrix pore regions. *Geoderma* 410: 115673. <https://doi.org/10.1016/j.geoderma.2021.115673>.
- Boivin P (2007) Anisotropy, cracking, and shrinkage of vertisol samples. Experimental study and shrinkage modeling. *Geoderma* 138: 25–38.
- Boivin P, Garnier P, and Tessier D (2004) Relationship between clay content, clay type and shrinkage properties of soil samples. *Soil Science Society of America Journal* 68: 1145–1153.
- Boivin P, Garnier P, and Vauclin M (2006) Modeling the soil shrinkage and water retention curves with the same equations. *Soil Science Society of America Journal* 70: 1082–1093.
- Boivin P, Schaeffer B, and Sturmy W (2009) Quantifying the relationship between soil organic carbon and soil physical properties using shrinkage modelling. *European Journal of Soil Science* 60: 265–275.
- Bouma J and van Lanen HAJ (1987) Transfer functions and threshold values: From soil characteristics to land qualities. In: *Quantified Land Evaluation Procedures, Proc. International Workshop on Quantified Land Evaluation Procedures*, 106–111. 27 April - 2 May 1986. Washington, DC.
- Braudeau E, Costantini JM, Bellier G, and Colleuille H (1999) New device and method for soil shrinkage curve measurement and characterization. *Soil Science Society of America Journal* 63: 525–535.
- Braudeau E, Frangi J-P, and Mohtar RH (2004) Characterizing nonrigid aggregated soil–water medium using its shrinkage curve. *Soil Science Society of America Journal* 68: 359–370.
- Braudeau E, Assi AT, Boukcim H, and Mohtar R (2014) Physics of the soil medium organization part 1: Thermodynamic formulation of the pedostructure water retention and shrinkage curves. *Frontiers in Environmental Science* 2. <https://doi.org/10.3389/fenvs.2014.00004>.
- Brewer R (1964) *Fabric and Mineral Analysis of Soils*. New York: John Wiley and Sons.
- Chertkov VY (2012) An integrated approach to soil structure, shrinkage, and cracking in samples and layers. *Geoderma* 173–174: 258–273. <https://doi.org/10.1016/j.geoderma.2012.01.010>.
- Coppola A, Comegna A, Dragonetti G, Gerke HH, and Basile A (2015) Simulated preferential water flow and solute transport in shrinking soils. *Vadose Zone Journal* 14. <https://doi.org/10.2136/vzj2015.02.0021>.
- Coquet Y, Touna J, and Boivin P (1998) Comparison of soil linear shrinkage curve from extracted cores and in situ. *Australian Journal of Soil Research* 36: 765–781.
- Fell V, Matter A, Keller T, and Boivin P (2018) Patterns and factors of soil structure recovery as revealed from a tillage and cover-crop experiment in a compacted orchard. *Frontiers in Environmental Science* 6. <https://doi.org/10.3389/fenvs.2018.00134>.
- Garnier P, Perrier E, Jaramillo RA, and Baveye P (1997) *Numerical model of 3-dimensional anisotropic deformation and 1-dimensional water flow in swelling soils*. <https://doi.org/10.1097/00010694-199706000-00003>.
- Glossary of Soil Science Terms (2020) <https://www.soils.org/publications/soils-glossary>. Soil Science Society of America [WWW Document], URL (accessed 5.25.20).
- Goutal-Pousse N, Lamy F, Ranger J, and Boivin P (2016) Structural damage and recovery determined by the colloidal constituents in two forest soils compacted by heavy traffic. *European Journal of Soil Science* 67: 160–172. <https://doi.org/10.1111/ejss.12323>.
- Grossman RB, Brasher BR, Franzmeier DP, and Walker JL (1968) Linear extensibility as calculated from natural-clod bulk density measurements. *Soil Science Society of America Journal* 32: 570–573. <https://doi.org/10.2136/sssaj1968.03615995003200040041x>.
- Haines WB (1923) The volume changes associated with variations of water content in soil. *The Journal of Agricultural Science* 13: 296–311.
- Hardy F (1923) The physical significance of the shrinkage coefficient of clays and soils. *The Journal of Agricultural Science* 13: 243–264. <https://doi.org/10.1017/S0021859600003567>.
- Horn R, Xinhua P, Heiner F, and Jose D (2014) Pore rigidity in structured soils—only a theoretical boundary condition for hydraulic properties? *Soil Science and Plant Nutrition* 60: 3–14. <https://doi.org/10.1080/00380768.2014.886159>.
- Johannes A, Weiskopf P, Schuln R, and Boivin P (2019) Soil structure quality indicators and their limit values. *Ecological Indicators* 104: 686–694. <https://doi.org/10.1016/j.ecolind.2019.05.040>.
- Kravchenko AN, Guber AK, Razavi BS, Koestel J, Quigley MY, Robertson GP, and Kuzyakov Y (2019) Microbial spatial footprint as a driver of soil carbon stabilization. *Nature Communications* 10: 3121. <https://doi.org/10.1038/s41467-019-11057-4>.
- Mallory JJ, Mohtar RH, Heathman GC, Schulze DG, and Braudeau E (2011) Evaluating the effect of tillage on soil structural properties using the pedostructure concept. *Geoderma* 163: 141–149. <https://doi.org/10.1016/j.geoderma.2011.01.018>.
- McGarry D and Daniells IG (1987) Shrinkage curve indices to quantify cultivation effects on soil structure of a vertisol. *Soil Science Society of America Journal* 51: 1575–1580.
- Milleret R, Le Bayon C, Lamy F, Gobat JM, and Boivin P (2009) Impact of root, mycorrhiza and earthworm on soil physical properties as assessed by shrinkage analysis. *Journal of Hydrology* 373: 499–507.
- Mitchell AR (1992) Shrinkage terminology: Escape from "Normalcy." *Soil Science Society of America Journal* 56: 993–994.
- Peng X, Dörner J, Zhao Y, and Horn R (2009) Shrinkage behaviour of transiently- and constantly-loaded soils and its consequences for soil moisture release. *European Journal of Soil Science* 60: 681–694. <https://doi.org/10.1111/j.1365-2389.2009.01147.x>.
- Peth S, Nellesen J, Fischer G, and Horn R (2010) Non-invasive 3D analysis of local soil deformation under mechanical and hydraulic stresses by μ CT and digital image correlation. *Soil and Tillage Research* 111: 3–18. <https://doi.org/10.1016/j.still.2010.02.007>. IZMIR conference (ISTRO 2009).
- Schäffer B, Schuln R, and Boivin P (2008) Changes in shrinkage of restored soil caused by compaction beneath heavy agricultural machinery. *European Journal of Soil Science* 59: 771–783.
- Schindler U, Doerner J, and Mueller L (2015) Simplified method for quantifying the hydraulic properties of shrinking soils. *Journal of Plant Nutrition and Soil Science* 178: 136–145. <https://doi.org/10.1002/jpln.201300556>.
- Smiles D and Raats PAC (2005) Swelling and shrinking. In: *Encyclopedia of Soils in the Environment*, pp. 115–123. Oxford, UK; Boston: Academic Press.
- Sposito G (1972) Thermodynamics of swelling clay-water systems. *Soil Science* 114: 243–249.
- Tessier D (1990) Behaviour and microstructure of clay minerals. In: De Boett M, Hayes M, and Herbillon A (eds.) *Soil Colloids and Their Association in Aggregates*, pp. 387–415. New York: Plenum Press.
- Towner GD (1986) Anisotropic shrinkage of clay cores, and the interpretation of field observations of vertical soil Movement. *Journal of Soil Science* 37: 363–371.
- Zolfaghari Z, Mosaddeghi MR, and Ayoubi S (2016) Relationships of soil shrinkage parameters and indices with intrinsic soil properties and environmental variables in calcareous soils. *Geoderma* 277: 23–34. <https://doi.org/10.1016/j.geoderma.2016.04.022>.

Further reading

- Braudeau E and Bruand A (1993) Determination of the clay shrinkage curve using the shrinkage curve of the undisturbed soil sample—Application to a soil sequence in Ivory-Coast. *Comptes Rendus de l'Academie des Sciences Serie II* 316: 685–692.
- Sisson JB and Wierenga PJ (1981) Spatial variability of steady-state infiltration rates as a stochastic process. *Soil Science Society of America Journal* 45: 699–704.
- Stirk GB (1954) Some aspects of soil shrinkage and the effect of cracking upon air entry into the soil. *Australian Journal of Soil Research* 5: 279–290.

Water properties

N Lu^a, C Zhang^b, and D Hillel^{c,†}, ^aColorado School of Mines, Golden, CO, United States; ^bHunan University, Changsha, Hunan, P.R. China; ^cFormerly Columbia University, New York, NY, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of D Hillel, WATER, PROPERTIES, Editor (s): D Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 290–300, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00442-2>

Introduction	146
Molecular structure	147
Hydrogen bonding	148
States of water	149
Water-ion interaction	151
Ionization and pH	151
Solvent properties of water	152
Osmotic pressure	153
Solubility of gasses	154
Vapor pressure	154
Adsorption of water on soil surface and interlamellar	155
Surface tension	156
Density and compressibility	156
Density of free water	156
Density of soil water	157
Dynamic and kinematic viscosity	157
References	158

Abstract

The ubiquity of water in our living world is briefly surveyed. Fundamental properties of water are described from atomic to bulk scales, including water molecular structure; unique role of hydrogen bond in water molecular structure; thermodynamic conditions defining water phases; stable and metastable states, and phase transitions; interaction of water and dissolved ion; dissolutions of gasses in liquid water; dependence of vapor pressure on temperature in the presence of liquid water or ice; soil water retention mechanisms of adsorption and capillarity; and water and soil water's macroscopic physical properties of density, compressibility, and dynamic and kinematic viscosity.

Key points

- The ubiquitousness of water in our living world is briefly surveyed.
- Fundamental properties of water are described from atomic to bulk scales, including water molecular structure; unique role of hydrogen bond in water molecular structure; thermodynamic conditions defining water phases; stable and metastable states, and phase transitions; interaction of water and dissolved ion; dissolutions of gases in liquid water; dependence of vapor pressure on temperature in the presence of liquid water or ice; soil water retention mechanisms of adsorption and capillarity; and water and soil water's macroscopic physical properties of density, compressibility, and dynamic and kinematic viscosity.

Introduction

Our Earth is the planet of life, primarily because it is blessed with the precise ranges of temperature and pressure that make possible the existence in a liquid state of a singular substance called water. So ubiquitous is water on our globe, covering nearly three-quarters of its surface, that the entire planet really should be called 'Water' rather than 'Earth.' However, as Coleridge's Ancient Mariner complained, most of the water everywhere is unfit to drink. Less than 2.5% of the water on earth is 'fresh' (i.e., non-saline) water, and that amount is spatially and temporally unevenly distributed. Humid regions are endowed with an abundance of it, even with a surfeit, so that often the problem is how to dispose of excess water. Arid and semiarid regions, on the other hand, are afflicted with a chronic shortage.

[†]Deceased

Life as we know it most likely began in an aquatic medium and water is still the principal constituent of living organisms. Water is, literally, the essence of life. As Vladimir Vernadsky (1863–1945) wrote a century ago: “Life is animated water.” Though we appear to be solid, we are really liquid bodies, similar to gelatin, which also seems solid but is in fact largely water, made consistent by the presence of organic material. The analogous material in the cells of our bodies is protoplasm. The water content of a newborn infant is nearly 75% water by mass, and even in adults it is over 55%. Actively growing herbaceous plants typically contain over 90% water. Far from being a bland, inert liquid, water is a basic reactant substance, a powerful solvent and an effective transporter of numerous substances.

The importance of water was recognized early in history, yet little was known about its real nature. In the Middle Ages, people believed that fresh water emanated magically from the bowels of the earth. They could not imagine that all the water flowing in innumerable springs and mighty rivers (such as the Nile, which appeared to the ancient Egyptians to come out of the driest desert!) could possibly result from so seemingly feeble a source as rain and snow. The first to conjecture this was Leonardo da Vinci, but only in the latter part of the 17th century did the English astronomer Edmond Halley and, separately, the Frenchman Claude Perrault prove the principle by calculation and measurement. Water was long thought to be a single element, until early in the 18th century, when it was found to consist of hydrogen and oxygen in combination.

Notwithstanding its ubiquity, water remains something of an enigma, possessing unusual and anomalous attributes. Perhaps the first anomaly is that, being a compound of two elements and having relatively low molecular weight, water co-exists as liquid and vapor at room temperature. (Its sister compound, hydrogen sulfide, H_2S , has a boiling-point temperature of -60.7°C .) Compared with other common liquids, water has unusually high melting and boiling points, heats of fusion and vaporization, specific heat, dielectric constant, viscosity, and surface tension.

Molecular structure

One cubic meter of liquid water at 20°C and 1 atm pressure (101.3 kPa) contains about 3.7×10^{28} (37 billion billion billion) molecules, the diameter of each is about 2.75×10^{-10} m (2.75×10^{-1} nm, or about 2.75 Angstrom units). The chemical formula of water is H_2O , which signifies that each molecule consists of two atoms of hydrogen and one of oxygen. There are three isotopes of hydrogen (^1H , ^2H , ^3H) and three of oxygen (^{16}O , ^{17}O , ^{18}O), which can form 18 combinations. However, all isotopes but ^1H and ^{16}O are quite rare.

The hydrogen atom consists of a positively charged proton and a negatively charged electron. The oxygen atom consists of a nucleus having a positive charge of eight protons, surrounded by eight electrons, of which six are in the outer shell. Since the outer electron shell of hydrogen lacks one electron and that of oxygen lacks two electrons to satisfy quantum stability, one atom of oxygen can combine with two atoms of hydrogen in an electron-sharing molecule.

The strong intermolecular forces in liquid water are caused by the electrical polarity of the water molecule, which in turn is a consequence of the arrangement of electrons in its oxygen and hydrogen atoms (Fig. 1). The oxygen atom shares a pair of electrons with each of the two hydrogen atoms, through overlap of the 1 s orbitals of the hydrogen atoms with two hybridized sp^3 orbitals of

Box 1

A Farewell to Teardrops.

For centuries, conventional wisdom held that larger stones fall faster than small ones, that the Earth is flat yet the sun revolves around it, and that the sun and moon are of equal size. We would like to believe that in our time all baseless notions have been replaced by sound science. But have they?

Consider the shape of a raindrop. The conventional standard is a teardrop, rounded at the bottom and tapering to a point at the top (Fig. B1). So prevalent is that image that it is printed in textbooks and used as a logo by irrigation companies and even by conferences sponsored by the United Nations. Alas, the vertically elongated, top-pointed raindrop is a physical impossibility.

A drop forming at the tip of a spout assumes that shape only at the very instant of detachment. Because of surface tension, any drop suspended in air ‘balls up’ spontaneously into a sphere. In a cloud, spherical droplets tend to grow by condensation or coalescence until reaching a critical weight equal to the buoyant force provided by the surrounding air, over which they begin to fall.

Air bypassing a falling drop acquires greater velocity around the curved ‘waist’ of the drop than near its top or bottom. By Bernoulli’s law ($p + \rho v^2/2 = \text{constant}$), the pressure of the air alongside the drop is lowered relative to that of the air above and below the drop. Consequently, the drop compresses vertically and comes to resemble an ellipsoid. Such is indeed the shape of small drops (<2 mm).

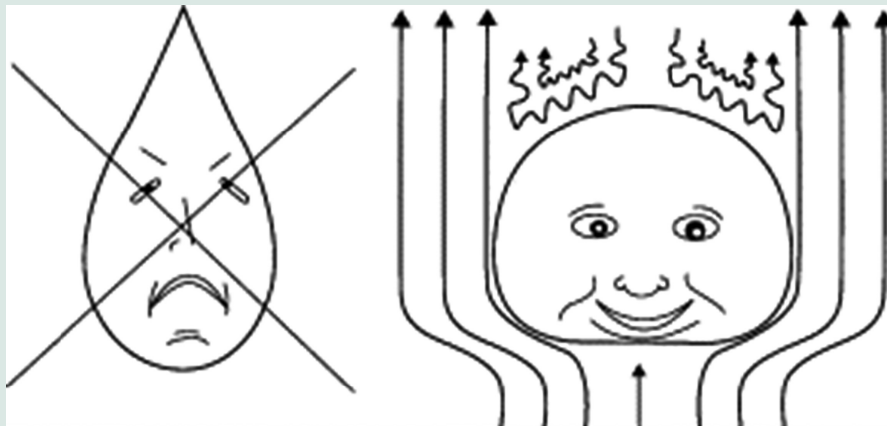
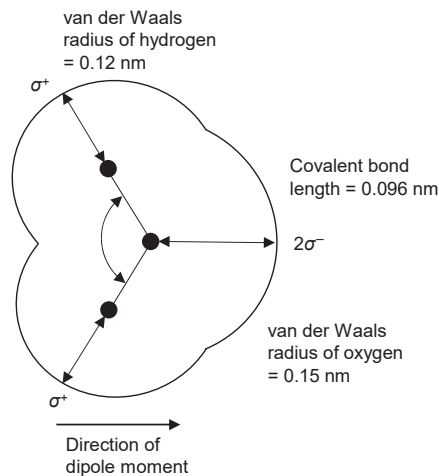
In the case of larger drops, the laminar streams of air flowing past the drop may not converge smoothly above the top, but may leave a turbulent wake there. In such a wake, the pressure is lower than it would be in an ideally laminar flow regime. The reduced air pressure at the top causes the drop to bulge a bit there, thus acquiring the appetizing shape of a miniature hamburger bun (described by the appropriately named McDonald as early as 1954).

If a drop were to continue accelerating, it might eventually spread out to form a pancake. But Stokes’ law intervenes, decreeing that the drop’s acceleration be countered by the increasing viscous resistance of the air. When air resistance equals the gravitational force, acceleration ceases and the drop continues to fall at a constant ‘terminal’ velocity and with a more or less constant shape. It finally slaps the ground with its flattened face going ‘plop’.

Some of us may think it unfair that such an exquisitely sculpted natural body should have no more glorious a fate than to splatter down on some dry bit of earth. But that is where—having at once lost its distinctive shape as it enters the labyrinthine interstices of the soil—our drop might well give life to a thirsty plant, perhaps even to a sunflower. Anyhow, we bid farewell to the sad countenance of the teardrop, a singularly inappropriate symbol for happy rain.

(Continued)

(Continued)

**Fig. B1** Conventional and real raindrops (Hillel, 1998).**Fig. 1** Model of a liquid water molecule (Hillel, 1998). The curved lines represent the borders at which van der Waals attractions are counterbalanced by repulsive forces.

the oxygen atom. The H—O—H bond in water is not linear but bent  at an angle of 104.5° . That angle deviates slightly from the perfectly tetrahedral arrangement of the oxygen atom's four possible sp^3 orbitals, which would have an angle of 109.5° . The mean H—O interatomic distance is 9.57×10^{-2} nm. The arrangement of electrons in the molecule gives it electrical asymmetry or polarity. The electronegative oxygen atom tends to attract the single electrons of the hydrogen atoms, leaving the hydrogen nuclei bare. Hence, each of the two hydrogen atoms has a local partial positive charge. The oxygen atom, in turn, has a partial negative charge, located in the zone of the unshared orbitals. Thus, though the water molecule has no net charge, it forms an electrical dipole.

Hydrogen bonding

Every hydrogen proton, while attached primarily to a particular molecule, is also attracted to the oxygen of the neighboring molecule, with which it forms a secondary link known as a hydrogen bond. Though the intermolecular link resulting from dipole attraction is not as strong as the primary link of the hydrogen to the oxygen of its own molecule, water can be regarded as a polymer of hydrogen-bonded molecules. This structure is most complete in ice crystals, in which each molecule is linked to four neighbors via four hydrogen bonds, thus forming a hexagonal lattice with a rather open structure (Fig. 2). When the ice melts, this rigid

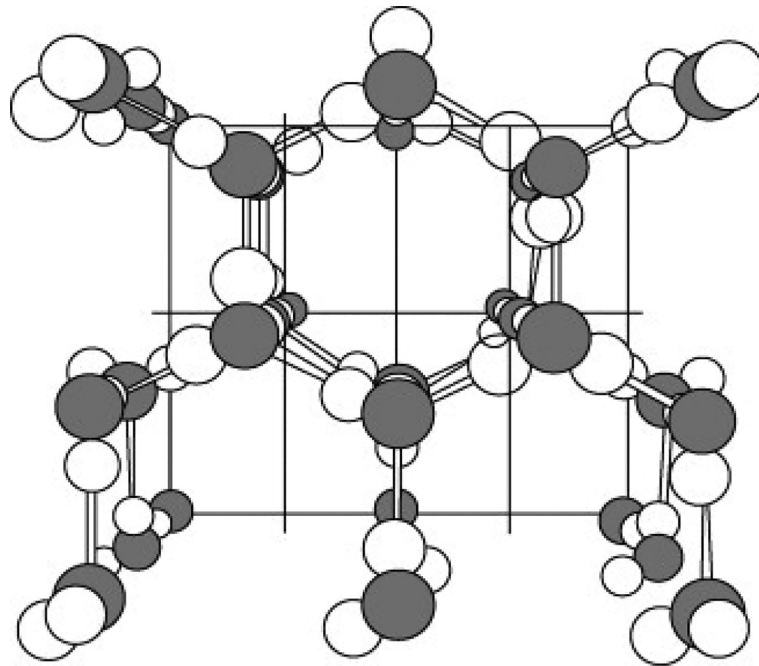


Fig. 2 Schematic structure of an ice crystal (Hillel, 1998). The oxygen atoms are shown in gray and the hydrogen atoms in white. The pegs linking adjacent molecules represent hydrogen bonds.

structure collapses partially, so additional molecules can enter the intermolecular spaces and each molecule can have more than four near neighbors. For this reason, liquid water can be denser than ice at the same temperature, and thus the ice sheet formed on the surface of lakes and ponds in winter remains on the top rather than sinks to the bottom as it would if ice were denser than liquid water.

States of water

In the vapor or gaseous state, water molecules are largely independent of one another and occur mostly as monomers signified as $(\text{H}_2\text{O})_1$. Occasionally, colliding molecules may fuse to form dimers $(\text{H}_2\text{O})_2$ or even trimers, $(\text{H}_2\text{O})_3$, but such combinations are rare. However, in the solid state a rigidly structured lattice forms with a tetrahedral configuration (Fig. 2) that can be schematically depicted as sheets of puckered hexagonal rings (Fig. 3). As many as nine alternative ice forms can occur when water freezes,

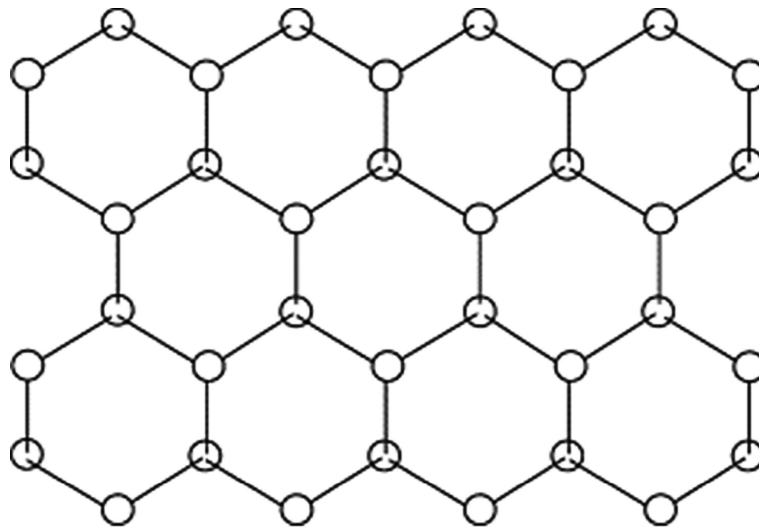


Fig. 3 The crystalline structure of ice I (Hillel, 1998).

depending on prevailing temperature and pressure conditions. Fig. 3 pertains to ice I, the familiar form, which occurs and is stable at ordinary atmospheric pressure.

The orderly structure of ice does not totally disappear in the liquid state. The polarity and hydrogen bonds continue to bind water molecules together, though the structural forms that develop in the liquid state are much more flexible and transient than in the rigidly structured solid state. Hydrogen bonds in liquid water form an extensive three-dimensional network, the detailed features of which appear to be short-lived. According to the 'flickering cluster' model discovered by Frank and Wen and modified by Erland, the molecules of liquid water associate and dissociate repeatedly in transitory or flickering polymer groups, designated $(\text{H}_2\text{O})_n$, having a quasicrystalline internal structure. These microcrystals, as they were, form and melt so rapidly on the scale of picoseconds and randomly that, on a macroscopic scale, water appears to behave as a homogeneous liquid (Fig. 4).

What determine water phases are pressure and temperature (Fig. 5). In transition from ice to liquid (fusion) and from liquid to vapor (vaporization), hydrogen bonds must be broken (while in freezing and condensation they are re-established). Hence relatively high temperatures and energies are required to achieve these water phase transitions. To thaw 1 kg of ice, $3.35 \times 10^5 \text{ J}$ (80 cal g^{-1}) must be supplied. Conversely, the same energy (the latent heat of fusion) is released in freezing. At the boiling point (100°C , at atmospheric pressure), water passes from the liquid to the gaseous state and in so doing it absorbs $2.26 \times 10^6 \text{ J kg}^{-1}$.

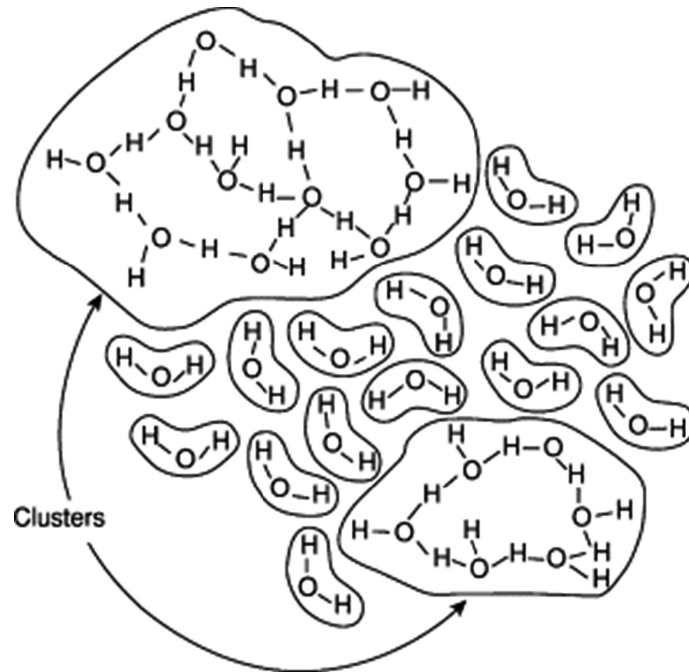


Fig. 4 Schematic illustration of 'flickering clusters,' showing polymeric associations and monomeric molecules in liquid water (Hillel, 1998).

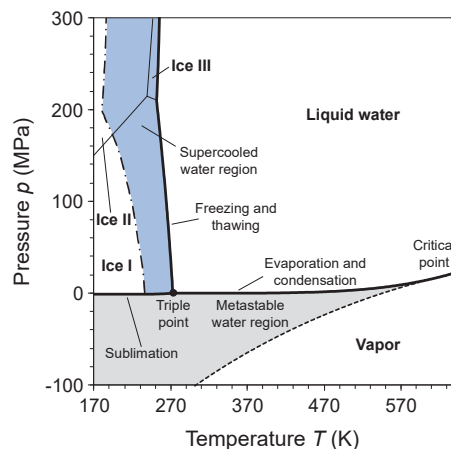


Fig. 5 Schematic illustration of water phase diagram in temperature and pressure (Debenedetti, 1996). Note: the pressure limits of cavitation are determined from the van der Waals equation of state.

(540 cal g⁻¹). This amount of heat is known as the latent heat of vaporization. Water can be vaporized at temperatures below 100 °C, but such vaporization requires greater heat or lower atmospheric pressure. At 30 °C, the latent heat is about 2.43×10^6 J kg⁻¹ (580 cal g⁻¹). Sublimation is the direct transition from ice to vapor, and the heat absorbed by this process is equal to the sum of the latent heats of fusion and vaporization.

Under certain pressure and temperature conditions, water is at the metastable state (Fig. 5), i.e., liquid water is cooled below the freezing point without crystallizing (supercooling) and liquid water is stretched below the saturated vapor pressure without gasifying (cavitation). Impurities like dissolved gasses and soil particles can greatly control the transition from liquid to vapor (cavitation) or conversely from vapor to liquid (condensation). Rough surfaces of soil particles often provide sites to entrap gasses that promote cavitation or to retain ice nuclei that trigger freezing. Small environmental perturbations in pressure and temperature can break the metastable states and yield spontaneous cavitation, condensation, or freezing.

Because of the existence of electromagnetic field near soil particle surface and the polarity of water molecules, water is attracted to soil particle surface and into clay lamellae, where much elevated inter-water molecular pressure is produced. Such elevated pressure could be as high in magnitude as gigapascal (GPa) for the formation of the first adsorbed layer of water molecules, leading to a depression of water freezing or supercooling temperature in soil to as low as -22 °C (i.e., 251 K in Ice III structure regime) (Fig. 5). The phenomenon of supercooling of soil water widely occurs in regions of thawing and freezing.

Water-ion interaction

Ionization and pH

Because of its small atomic mass and the tightness with which its single electron is bound to the oxygen atom, the nucleus of the hydrogen atom in the water molecule exhibits a finite tendency to dissociate from the oxygen with which it is covalently associated and to 'jump' to the adjacent water molecule, with which it is hydrogen-bonded. Such an event produces two ions: the hydronium cation (H₃O⁺) and the hydroxyl ion (OH⁻). The reaction described is reversible, and should be written as:



However, by convention it is written simply as:



and one speaks of 'hydrogen ions' rather than of 'hydronium ions.'

Although the self-ionization of water is small, its consequences are extremely important. The ionization is reversible, and it tends to an equilibrium state in which the rate of dissociation into ions equals the rate of ion reassociation to form molecules once again. For such a system in equilibrium (at which the concentration of each of the species H₂O, H⁺, and OH⁻ remains constant), the law of mass action applies; i.e., the ratio of concentrations of the products and the reactants must be constant. Using brackets to denote concentration, we can write this in the following way:

$$K_{\text{equal}} = [\text{H}^+][\text{OH}^-]/[\text{H}_2\text{O}] \quad (1)$$

Since the number of water molecules undergoing dissociation at any given time is very small relative to the total number of water molecules present, [H₂O] can be considered constant. Assuming this concentration to be 55.5 mol L⁻¹ (the number of grams per liter divided by the gram molecular weight: 1000/18 = 55.5 mol L⁻¹), we can simplify the equilibrium constant expression as follows:

$$55.5K_{\text{equal}} = [\text{H}^+][\text{OH}^-], \text{ or } K_w = [\text{H}^+][\text{OH}^-] \quad (2)$$

in which K_w is a composite constant called the ion product of water. In fact, the concentrations of H⁺ and OH⁻ ions in pure water at 25 °C are 10⁻⁷ mol L⁻¹, an extremely small value when compared to the overall concentrations of (largely undissociated) water, namely 55.5 mol L⁻¹. Thus, K_w at 25 °C is 10⁻¹⁴. If the hydroxyl ion concentration [OH⁻] is changed, the hydrogen ion concentration [H⁺] changes automatically to maintain the constancy of the product, and vice versa. An excess concentration of hydrogen ions over the concentration of hydroxyl ions imparts to the aqueous medium the property of acidity, whereas a predominance of hydroxyl ions produces the opposite property of alkalinity or basicity. A condition in which the concentrations of H⁺ and OH⁻ are equal is called neutrality.

The ion product of water, K_w , is the basis for the pH scale, a measure of the concentration of H⁺ (and of OH⁻ as well) in any aqueous solution in the range of concentration between 1.0 mol L⁻¹ H⁺ and 1.0 mol L⁻¹ OH⁻. The pH is defined as:

$$\text{pH} = \frac{\log_{10} 1}{[\text{H}^+]} = -\log_{10} [\text{H}^+] \quad (3)$$

As already stated, in a precisely neutral solution at 25 °C, the pH of such a solution is:

$$[\text{H}^+] = 1.0 \times 10^{-7} \text{ mol l}^{-1}$$

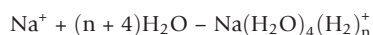
The pH of such a solution is:

$$\text{pH} = \log_{10} \left[\frac{1}{1 \times 10^{-7}} \right] = 7.0$$

The pH value of 7.0 for a neutral solution is thus not arbitrary, but derives from the value of the ion product of water at 25 °C. The pH scale is logarithmic, not arithmetic. If two solutions differ in pH by one unit, then one solution has 10 times the hydrogen ion concentration of the other. Thus, a pH of 6 implies a hydrogen ion concentration of 10^{-6} and a hydroxyl ion concentration of 10^{-8} . A pH of 5 indicates $[\text{H}^+] = 10^{-5}$ (acidity 10 times that of the above) and $[\text{OH}^+] = 10^{-9}$.

Solvent properties of water

Water dissolves or disperses many substances because of its polar nature. Hence it has been called the universal solvent. All chemical substances have finite solubilities in water, but these solubilities range widely. Many crystalline salts and other ionic compounds readily dissolve in water but are nearly insoluble in nonpolar liquids such as chloroform or benzene. Since the crystal lattice of salts, such as sodium chloride, is held together by very strong electrostatic attractions between alternating positive and negative ions, considerable energy is required to pull these ions away from one another. However, water dissolves sodium chloride because the strong electrostatic attraction between water dipoles and the Na^+ and Cl^- ions, forming stable hydrated Na^+ and Cl^- ions, exceeds the attraction of these ions to each other. In the case of Na^+ , hydration is represented by the process:



illustrated in Fig. 6. In addition to hydration, there is also the hydrolysis of metal species, a reaction in which the metal ion displaces one of the protons (hydrogen) of water to form basic hydroxides.

Ion solvation is also aided by the tendency of the solvent to oppose the electrostatic attraction between ions of opposite charges. This is characterized by the dielectric constant D , which is defined by the relationship:

$$F = \kappa e_1 e_2 / Dr^2$$

Here F is the attractive force between two ions of opposite charge, κ is Coulomb's constant, e_1 and e_2 are charges on the ions, and r is the distance between them. Water has an extremely high dielectric constant (compared to 1.0 of air), as shown in Table 1. For instance, the attractive force between Na^+ and Cl^- ions at a given distance in water is less than one-third that in ethanol and only one-40th that in benzene. This fact greatly facilitates hydration of ions and dissolution of the crystal lattices of salts in water.

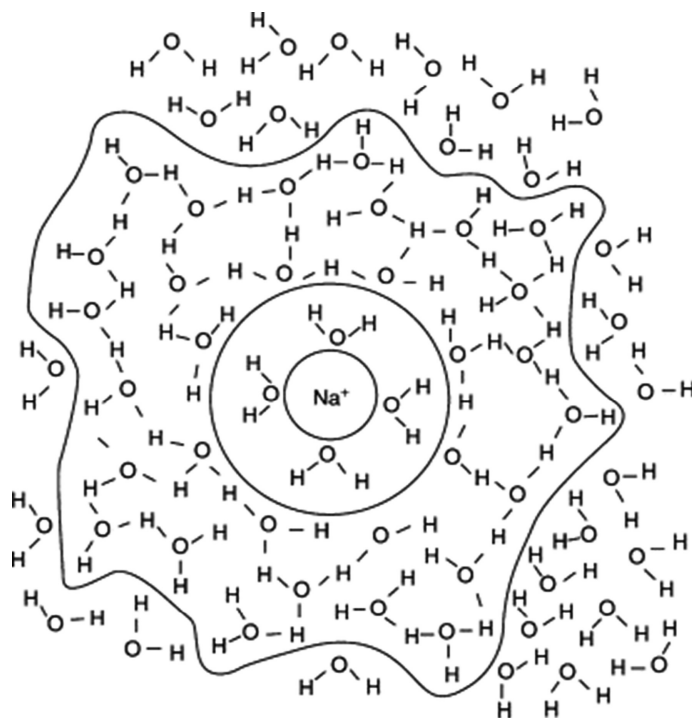


Fig. 6 A model of the hydration 'atmosphere' of a sodium ion (Hillel, 1998). An inner shell of more or less rigidly structured water is surrounded by a cluster of looser but still structure-enhanced water, the whole floating in a sea of 'free' water.

Table 1 Dielectric constants of some liquids (20 °C) (Hillel, 1998).

Water	80.4	Acetone	21.4
Methanol	33.0	Benzene	2.3
Ethanol	25.3	Hexane	1.9

The effect of a solute on the solvent is manifest in a set of properties, namely, the colligative properties of solutions, which depend on the number of solute particles per unit volume of solvent. Solutes produce such characteristic effects in the solvent as depression of the freezing point and of the vapor pressure, and elevation of the boiling point. They also endow a solution with the property of osmotic pressure. Theoretically, 1 mol of an ideal solute dissolved in 1 kg of water at a pressure of 760 mm of mercury (0.1 MPa, or 1 bar) depresses the freezing point by 1.86 °C and elevates the boiling point by 0.543 °C. Such a solution also yields an osmotic pressure of 2.24 MPa (22.4 atm) in an appropriate apparatus. However, aqueous solutions usually deviate considerably from ideal behavior, and the deviations are greater the higher the concentrations. The quantitative relationships given above are exact only at infinite dilution.

Osmotic pressure

Owing to the constant thermal motion of all molecules in a fluid (above a temperature of absolute zero), solute species spread throughout the solution in a spontaneous tendency toward a state of equal concentration throughout. This migration of solutes in response to spatial differences in concentration is called diffusion that is a consequence of the second law of thermodynamics: matters move from high energy place to low energy place.

If a physical barrier is interposed between two regions, across the path of diffusion, and if that barrier is permeable to molecules of the solvent but not to those of the solute, the former will diffuse through the barrier in a process called osmosis (from the Greek $\omega\sigma\mu\sigma$, meaning 'push'). As in the case of unhindered diffusion, this process tends toward a state of uniform concentration even across the barrier. Barriers permeable to one substance in a solution but not to another are called selective or semipermeable membranes. Membranes surrounding cells in living organisms, for example, exhibit selective permeability to water while restricting the diffusion of solutes between the cells' interior and their exterior environment. Water molecules cross the membrane in both directions, but the net flow of water is from the more dilute solution to the more concentrated, following the second law of thermodynamics.

Fig. 7 is a schematic representation of a pure solvent separated from a solution by a semipermeable membrane. Solvent will pass through the membrane and enter the solution compartment, driving the solution level up the left-hand tube until the hydrostatic pressure of the column of dilute solution on the left is sufficient to counter the diffusion pressure of the solvent molecules drawn into the solution through the membrane. This process will cease when the chemical potentials of the solutions in both sides are equal or reach equilibrium. The hydrostatic pressure at equilibrium, when solvent molecules are crossing the membranes in both directions at equal rates, is the osmotic pressure of the solution.

In dilute solutions, the osmotic pressure is proportional to the concentration of the solution and to its temperature according to the following equation:

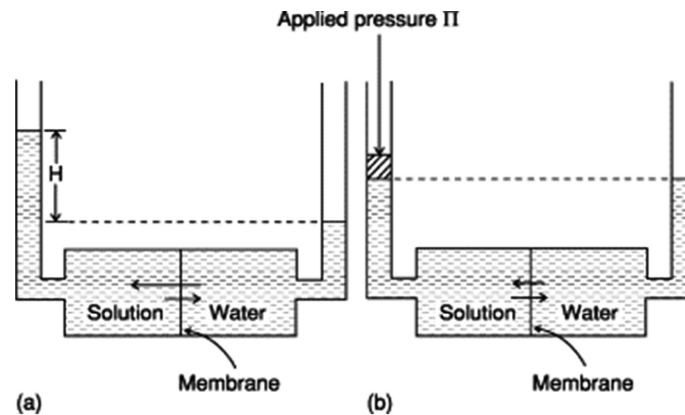


Fig. 7 Osmosis and osmotic pressure (Hillel, 1998): (a) in osmosis, the flow of water molecules through the membrane into the solutions is at first greater than the reverse flow from the solution into the water compartment due to a lower chemical potential in the solution compartment. The hydrostatic pressure due to the column of expanded solution increases the rate of water flow from the solution to the water compartment until, at equilibrium, the opposite flows are equal and the chemical potentials in both compartments are the same. (b) The osmotic pressure of the solution is equal to the hydrostatic pressure Π , which must be applied to the solution to equalize the rate of flow to and from the solution and produce a net flow of zero.

$$\Pi = MRT \quad (5)$$

Here Π is the osmotic pressure in atmospheres (to be multiplied by 0.101 to obtain megapascal units), M the total molar concentration of the solute (whether molecules or dissociated ions), T the temperature in degrees Kelvin, and R the gas constant (0.08205 liter atm deg.⁻¹ mole). The osmotic pressure increase with temperature is associated with the corresponding increase of the molecular diffusivity (self-diffusion coefficient) of water, D_w . According to the Einstein-Stokes equation,

$$D_w = kT/6\pi r\eta$$

where $k = R/N$, the Boltzmann constant (1.38×10^{-23} J K⁻¹); r is the rotation radius of the molecule ($\sim 1.5 \times 10^{-4}$) and η is the viscosity.

In soil, the concentration of solution is usually non-uniform near particle surface due to the formation of electrical double layer. Thereby, osmotic pressure in soil varies spatially near particle surface (Lu and Zhang, 2019). This phenomenon stems from the crystalline structure of clay minerals, where particle surfaces are usually negatively charged and balanced by counterions. This additional electrical field emanates from the particle surfaces, attracting cations, and repelling anions. The resulted cation concentration profile exhibits a maximum near the particle surface and decays to a constant value of the bulk solution concentration at the far field.

Solubility of gasses

The concentration of dissolved gasses in water in equilibrium with a gaseous phase generally increases with pressure and decreases with temperature. According to Henry's law, the mass concentration of a dissolved gas c_m is proportional to the partial pressure of the gas p_i :

$$c_m = s_o p_i / p_o \quad (7)$$

where s_o is the solubility coefficient of the particular gas in water and p_o is the total pressure of the atmosphere. The volume concentration is similarly proportional:

$$c_v = s_v p_i / p_o$$

where s_v is the solubility coefficient expressed in terms of volume ratios (i.e., c_v is the volume of dissolved gas relative to the volume of water). The values of s_o and s_v are determined experimentally. If the gas does not react chemically with the liquid, these properties should remain constant over a range of pressures, especially at low partial pressures of the dissolved gasses. Solubility is, however, strongly influenced by temperature. Table 2 gives the s_v values of several atmospheric gasses at various temperatures.

Vapor pressure

According to the kinetic theory, molecules in a liquid are in constant motion, which is an expression of their thermal energy. These molecules collide frequently, and occasionally one or another at the surface absorbs sufficient momentum to leap out of the liquid and into the atmosphere above it. Such a molecule, by virtue of its kinetic energy, thus changes from the liquid to the gaseous phase. This kinetic energy is then lost in overcoming the potential energy of intermolecular attraction while escaping from the liquid. At the same time, some of the randomly moving molecules in the gaseous phase may strike the surface of the liquid and be absorbed in it.

The relative rates of these two directions of movement depend upon the concentration of vapor in the atmosphere relative to its concentration at a state of equilibrium (i.e., when the movement in both directions is equal). An atmosphere that is at equilibrium with a body of pure water at standard atmospheric pressure is considered to be saturated with water vapor, and the partial pressure of the vapor in such an atmosphere is called the saturation (or equilibrium) vapor pressure. The vapor pressure at equilibrium with any body of water depends on the physical condition of the water (pressure and temperature) and its chemical condition (solutes) but is independent of the absolute or relative quantity of liquid or gas in the system.

Table 2 Solubility coefficients of gasses in water (Lu and Likos, 2004).

Temperature (°C)	Nitrogen (N ₂)	Oxygen (O ₂)	Carbon dioxide (CO ₂)	Air (without CO ₂)
0	0.0235	0.0489	1.713	0.0292
10	0.0186	0.0380	1.194	0.0228
20	0.0154	0.0310	0.878	0.0187
30	0.0134	0.0261	0.665	0.0156
40	0.0118	0.0231	0.530	0.0141

The solubilities of various gasses (particularly oxygen) in varying conditions strongly influence such vital soil processes as oxidation and reduction, and respiration by roots and microorganisms.

Table 3 Physical properties of water vapor (Hillel, 1998).

Temperature (°C)	Saturation vapor pressure (torr)		Vapor density in saturated air (kg m ⁻³)		Diffusion coefficient (m ² s ⁻¹) (× 10 ⁻⁴)
	Over liquid	Over ice	Over liquid (× 10 ⁻³)	Over ice (× 10 ⁻³)	
-10	2.15	1.95	2.36	2.14	0.211
-5	3.16	3.01	3.41	3.25	—
0	4.58	4.58	4.85	4.85	0.226
5	6.53	—	6.80	—	—
10	9.20	—	9.40	—	0.241
15	12.78	—	12.85	—	—
20	17.52	—	17.30	—	0.257
25	23.75	—	23.05	—	—
30	31.82	—	30.38	—	0.273
35	42.20	—	39.63	—	—
40	55.30	—	51.1	—	0.289
45	71.90	—	65.6	—	—
50	92.50	—	83.2	—	—

The saturation vapor pressure rises with temperature. As the kinetic energy of the molecules in the liquid increases, so does the evaporation rate. Consequently, a higher concentration of vapor in the atmosphere is required for the rate of return to the liquid to match the rate of escape from it. A liquid arrives at its boiling point when the vapor pressure becomes equal to the atmospheric pressure. If the temperature range is not too wide, the dependence of saturation vapor pressure on temperature is expressible by the equation (Table 3):

$$\ln p_o = a - b/T \quad (8)$$

where $\ln p_o$ is the logarithm to the base e of the saturation vapor pressure p_o , T is the absolute temperature, and a and b are constants.

As mentioned, the vapor pressure also depends on the hydrostatic pressure of the liquid water. At equilibrium with drops of water (which have a hydrostatic pressure greater than atmospheric), the vapor pressure is oversaturated and greater than in a state of equilibrium with free water (which has a flat interface with the atmosphere). On the other hand, in equilibrium with adsorbed or capillary water under a potential lower than free water, the vapor pressure is smaller than in equilibrium with free water. This phenomenon is called vapor pressure lowering. The curvature of water drops is considered to be positive, as these drops are convex toward the atmosphere, whereas the curvature of capillary water menisci is considered negative, as they are concave toward the atmosphere.

For water in capillaries, in which the air-water interface is concave, the Kelvin equation applies:

$$-(\mu_1 - \mu_1^o) = RT \ln (p_1^o/p_1) = 2\gamma v_1 \cos \alpha / r_c$$

in which $(\mu_1 - \mu_1^o)$ is the change in potential of the water due to the curvature of the air-water interface, γ is the surface tension of water, α is the contact angle, v_1 is the partial molar volume of water, and r_c is the radius of the capillary.

Water present in the soil invariably contains solutes, mainly electrolytic salts, in variable concentrations. Thus, soil water should properly be called the soil solution. The composition and concentration of the soil solution affect soil behavior. While in humid regions the soil solution may have a concentration of but a few parts per million, in arid regions the concentration may become as high as several percent. The ions commonly present are H^+ , Ca^{2+} , Mg^{2+} , Na^+ , K^+ , NH_4^+ , OH^- , Cl^- , HCO_3^- , NO_3^- , SO_4^{2-} , and CO_3^{2-} . The vapor pressure of electrolytic solutions is less than that of pure water. The equation is:

$$v_1 \Pi_o = RT \ln (p_1^o/p_1) = \mu_1 - \mu_1^o$$

wherein Π_o is the osmotic pressure of a nonvolatile solute, μ_1^o and p_1^o are the chemical potential and vapor pressure of the liquid in its pure state, and μ_1 and p_1 are the same for the solution. Thus, the soil solution has a lower vapor pressure than pure water, even when the soil is saturated. In unsaturated soil the capillary and adsorptive effects further lower the potential and hence also the vapor pressure.

Adsorption of water on soil surface and interlamellar

Adsorption is an interfacial phenomenon resulting from the differential forces of attraction or repulsion occurring among molecules or ions of different phases at their exposed surfaces. As a result of cohesive and adhesive forces coming into play, the zones of contact among phases may exhibit a concentration or a density of material different from and higher than that inside the phases themselves.

As different phases come in contact, various types of adsorption can occur: adsorption of gasses on solids, of gasses on liquid surfaces, and of liquids on solids.

The interfacial forces of attraction or repulsion may themselves be of different types, including electrostatic or ionic (Coulombic) forces, intermolecular forces such as van der Waals and London forces, and short-range repulsive (Born) forces. The adsorption of water upon solid surfaces is generally of an electrostatic nature. The polar water molecules attach to the charged faces of the solids and to the ions adsorbed on them. This adsorption of water is the mechanism causing the strong retention of water by clay at high suctions, resulting higher pore water pressure and density near soil particle surface and between clay interlamellar.

The interaction of the charges of the solid with the polar water molecules may impart to the adsorbed water a distinct structure in which the water dipoles assume an orientation dictated by the charge sites on the solids. The elevated pore water pressure in this adsorbed 'phase' could be as large as GPa in order of magnitude, and may induce abnormally elevated water density, highly suppressed fusion temperature, and significantly increased viscosity compared to ordinary liquid water at the same temperature. The adsorption of water on clay surfaces is an exothermic process, resulting in the liberation of an amount of heat known as the heat of wetting.

Surface tension

Surface tension is a phenomenon occurring typically, but not exclusively, at the interface of a liquid and a gas. The liquid behaves as if it were covered by an elastic membrane in a constant state of tension that tends to cause the surface to contract. To be sure, no such membrane exists but a thin layer with highly variable inter-water molecular stress, yet the analogy is a useful one if not taken too literally. If we draw an arbitrary line of length L on a liquid surface, there will be a force F pulling the surface to the right of the line and an equal force pulling the surface leftwards. The ratio F/L is the surface tension and its dimensions are those of force per unit length. The same phenomenon can also be described in terms of energy. Increasing the surface area of a liquid requires work, which remains stored as internal energy in the enlarged surface, just as energy can be stored in a stretched spring. That internal energy can perform work if the enlarged surface is allowed to contract again. Energy per unit area has the same dimensions as force per unit length.

An explanation for occurrence of surface tension is given in Fig. 8. A molecule inside the liquid is attracted in all directions equally by the cohesive forces of neighboring molecules, while a molecule at the surface of the liquid is attracted into the relatively dense liquid phase by a net force greater than that attracting it toward the rarified gaseous phase. This unbalanced force draws the surface molecules inward into the liquid and results in the tendency for the surface to contract. This is why drops of a liquid in air as well as bubbles of air in a liquid assume the shape of a sphere, which is a body of minimal surface exposure relative to its volume. The curvature of such sphere gives rise to a pressure difference across the liquid-air interface, viz. capillary pressure. A detailed explanation for this phenomenon can be found in the Chapter on Capillarity.

Different liquids exhibit different surface tension values with air, as in the following list at 20 °C:

Water, $7.27 \times 10^{-2} \text{ N m}^{-1}$ (72.7 dyn cm^{-1});
 Ethyl ether, $1.7 \times 10^{-2} \text{ N m}^{-1}$ (17 dyn cm^{-1});
 Ethyl alcohol, $2.2 \times 10^{-2} \text{ N m}^{-1}$ (22 dyn cm^{-1});
 Benzene, $2.9 \times 10^{-2} \text{ N m}^{-1}$ (29 dyn cm^{-1});
 Mercury, 0.43 N m^{-1} (430 dyn cm^{-1}).

Density and compressibility

Density of free water

The open packing of water molecules in ice and liquid water accounts for their relatively low densities. Unlike most substances, water exhibits a point of maximum density (at 4 °C), below which the substance expands due to the formation of a hexagonal lattice structure, and above which the expansion is due to the increasing thermal motion of the molecules. The coefficient of thermal

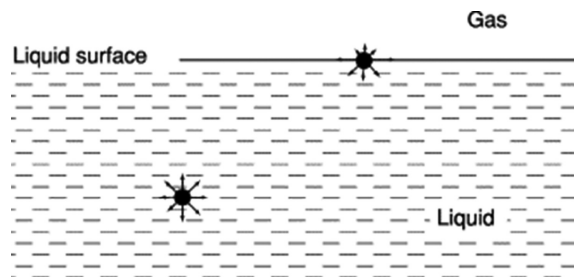


Fig. 8 Cohesive forces acting on a molecule inside the liquid and at its surface.

expansion of water is rather small, and in the normal temperature range of, say, 4–50 °C, the density diminishes only slightly, from 1000 to 988 kg m⁻³.

The compressibility of water, C_w , can be defined as the relative change in density with change in pressure:

$$C_w = (1/\rho_w)(d\rho_w/dP) \quad (9)$$

At 20 °C and at atmospheric pressure, the compressibility of pure water is about $4.6 \times 10^{-10} \text{ m}^2 \text{ N}^{-1}$. In the normal situations encountered on the surface of the Earth, water can usually be considered incompressible. The compression of water cannot be ignored, however, in the case of deep confined aquifers, which may be subject to a pressure of, say, 10 MPa or more.

Density of soil water

Water in soil can exhibit a density up to 1.872 g/cm³ (Zhang and Lu, 2018), depending on soil types mineral and water content (Fig. 9). Such a high-density value stems from the highly compressive pressure at the order of GPa near the particle surface induced by adsorption (Zhang and Lu, 2019). The magnitude of adsorption usually decays with the distance from particle surface or interlamellar surface (Lu and Zhang, 2019). Accordingly, the elevated soil water density decreases. As a result, the average water density in soil generally decays with water content because higher water content suggests more water lies further away from particle surface. The decay rate depends on the characteristics of adsorption strength of soil therefore varies with soil types, i.e., more prevalence in clayey soils with higher adsorption strength (Zhang and Lu, 2020).

Dynamic and kinematic viscosity

When a fluid is moved in shear (i.e., when adjacent layers of fluid are made to slide over each other), the force required is proportional to the velocity of shear. The proportionality factor is called the viscosity. As such, it is the property of the fluid to resist the rate of shearing and this can be visualized as an internal friction. The coefficient of viscosity η is defined as the force per unit area necessary to maintain a velocity difference of 1 m s⁻¹ between two parallel layers of fluid which are 1 m apart. The viscosity equation is:

$$\tau = \frac{F_s}{A} = \eta \, du/dx \quad (10)$$

wherein τ is the shearing stress, consisting of a force F_s acting on an area A ; η (dimensions: mass/(length × time)) is the coefficient of dynamic viscosity; and du/dx is the velocity gradient perpendicular to the stressed area A .

The ratio of the dynamic viscosity of a fluid to its density is called the kinematic viscosity, designated ν . It expresses the shearing-rate resistance of a fluid independently of the density. Thus, while the dynamic viscosity of water exceeds that of air by a factor of about 50 (at room temperature), its kinematic viscosity is actually lower.

Fluids of lower viscosity flow more readily and are said to possess greater fluidity (which is the reciprocal of viscosity). As shown in Table 4, the viscosity of water diminishes by over 2% per 1 °C rise in temperature, and thus decreases by more than half as the temperature increases from 5 °C to 35 °C. The viscosity is also affected by the type and concentration of solutes present.

Water is the origin of all the known forms of life. That is why the search for life usually begins with the search for water, e.g., the theme of NASA's Mars Exploration Program, "Seek Signs of life", begins with an exploration strategy known as "Follow the Water". Water occupies about half volume of cells and controls various biological processes by acting as a solvent, transporter, reactant, messenger, or catalyst (e.g., Brini et al., 2017). These indispensable roles come from water's unique polar molecular structure, for example, (1) the high polarity emanates electrostatic fields to attract other polar substances or ionic materials, facilitating the

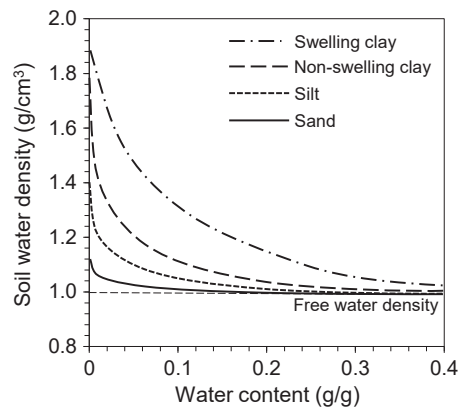


Fig. 9 Water density in soils as a function of water content and soil types (Zhang and Lu, 2018).

Table 4 Physical properties of liquid water (Hillel, 1998).

Temperature (°C)	Density (kg m ⁻³) (× 10 ³)	Specific heat (J kg ⁻¹ deg C) (× 10 ³)	Latent heat (vaporization) (J kg ⁻¹) (× 10 ⁶)	Surface tension (kg s ⁻²) (× 10 ⁻²)	Thermal conductivity (J m ⁻⁶ s deg C)	Dynamic viscosity (kg m ⁻¹ s) (× 10 ⁻²)	Kinematic viscosity (m ² s ⁻¹) (× 10 ⁻⁶)
-10	0.99794	4.27	2.53	—	—	—	—
-5	0.99918	4.23	2.51	7.64	—	—	—
0	0.99987	4.22	2.50	7.56	0.561	0.1787	1.79
4	1.00000	4.21	2.49	7.5	0.570	0.1567	1.57
5	0.99999	4.207	2.49	7.48	0.574	0.1519	1.52
10	0.99973	4.194	2.48	7.42	0.587	0.1307	1.31
15	0.99913	4.19	2.47	7.34	0.595	0.1139	1.14
20	0.99823	4.186	2.46	7.27	0.603	0.1002	1.007
25	0.99708	4.18	2.44	7.19	0.612	0.0890	0.897
30	0.99568	4.18	2.43	7.11	0.620	0.0798	0.804
35	0.99406	4.18	2.42	7.03	0.629	0.0719	0.733
40	0.99225	4.18	2.41	6.95	0.633	0.0633	0.661
45	0.99024	4.18	2.40	6.87	0.641	0.0596	0.609
50	0.98807	4.186	2.38	6.79	0.645	0.0547	0.556

dissolution and transportation of biomolecules, such as proteins and nucleic acids (e.g., Riveros-Perez and Riveros, 2018); (2) for non-polar molecules, the stronger hydrogen bonds among polar water molecules give rise to the hydrophobic effect, which is responsible for protein folding, formation of cell membranes, and adsorption of proteins into lipid layers (e.g., Ball, 2017; Riveros-Perez and Riveros, 2018); and (3) water molecule's polarity is the origin of surface tension and capillarity, which drives water rise in xylem of plants, and can be implemented in pharmacologic formulations and technologies (e.g., Brini et al., 2017; Riveros-Perez and Riveros, 2018). Additionally, water is the most common reactant for biological actions, e.g., cellular respiration and photosynthesis in plants. Summarizing, water involves in all the biochemical reactions directly or indirectly, and its unique properties play a vital role in sustaining the entire biosphere and living system.

References

- Ball P (2017) Water is an active matrix of life for cell and molecular biology. *Proceedings of the National Academy of Sciences* 114(51): 13327–13335.
- Brini E, Fennell CJ, Fernandez-Serra M, Hribar-Lee B, Lukšič M, and Dill KA (2017) How water's properties are encoded in its molecular structure and energies. *Chemical Reviews* 117(19): 12385–12414.
- Debenedetti PG (1996) *Metastable Liquids: Concepts and Principles*. Princeton, NJ: Princeton University Press.
- Hillel D (1998) *Environmental Soil Physics*. Boston: Academic Press.
- Lu N and Likos WJ (2004) *Unsaturated Soil Mechanics*. New Jersey: John Wiley and Sons.
- Lu N and Zhang C (2019) Soil sorptive potential: Concept, theory, and verification. *Journal of Geotechnical and Geoenvironmental Engineering* 145(4): 04019006.
- Riveros-Perez E and Riveros R (2018) Water in the human body: An anesthesiologist's perspective on the connection between physicochemical properties of water and physiologic relevance. *Annals of Medicine and Surgery* 26: 1–8.
- Zhang C and Lu N (2018) What is the range of soil water density? Critical reviews with a unified model. *Reviews of Geophysics* 56(3): 532–562.
- Zhang C and Lu N (2019) Unitary definition of matric suction. *Journal of Geotechnical and Geoenvironmental Engineering* 145(2): 02818004.
- Zhang C and Lu N (2020) Soil sorptive potential: Its determination and predicting soil water density. *Journal of Geotechnical and Geoenvironmental Engineering* 146(1): 04019118.

Water potential

Dani Or^a, Markus Tuller^b, and Jon M Wraith^c, ^aDivision of Hydrologic Sciences, Desert Research Institute, Reno, NV, United States; ^bDepartment of Environmental Science, The University of Arizona, Tucson, AZ, United States; ^cCollege of Life Sciences and Agriculture, University of New Hampshire, Durham, NH, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	159
Historical perspective	159
Definition of soil water potential	160
Components of the water potential	161
Measurement of soil water potential components	163
Total water potential	163
Pressure potential	164
Matric potential	164
Osmotic potential	166
Summary	167
References	167

Abstract

The concept of soil water potential is central to the characterization of water flow in unsaturated soils, to water retention, to plant available soil water and to soil strength. The concept is rooted in thermodynamic theory and reflects the attraction of soil and water, equilibrium between the liquid and vapor phases of soil water, and how gravity, pressure and soluble solutes affect soil water energy state and its exchanges with the atmosphere and plants. We review the historical roots of the water potential concept and offer practical definitions of the various components that contribute to water potential. Advances in measurement methods offer new insights into the relative influences of water potential components on water flow and exchanges with plants and the atmosphere.

Key points

- The historical origin of the soil water potential concept is reviewed.
- Various components that contribute to soil water potential are defined.
- Contemporary state-of-the-art measurement methods for all soil water potential components are presented.

Introduction

Water status in soils is characterized by both the amount of water present and its energy state. Soil water is subjected to forces of variable origin and intensity, thereby acquiring different quantities and forms of energy. The two primary forms of energy of interest here are kinetic and potential. Kinetic energy is acquired by virtue of motion and is proportional to velocity squared. However, because the movement of water in soils is relatively slow (usually less than 0.1 m h^{-1}), its kinetic energy is negligible. Potential energy, which is defined by the position of soil water within a soil body and by internal conditions, is largely responsible for determining soil water status under isothermal conditions. Like all other matter, soil water tends to move from where the potential energy is higher to where it is lower, in pursuit of equilibrium with its surroundings. The magnitude of the driving force behind such spontaneous motion is a difference in potential energy across a distance between two points of interest. At a macroscopic scale, we can define potential energy relative to a reference state. The standard state for soil water is defined as pure and free water (no solutes and no external forces other than gravity) at a reference pressure, temperature, and elevation, and is arbitrarily given the value of zero.

Historical perspective

Buckingham (1907) introduced the concept of *capillary potential* to quantify the energy state soil water as affected by the attraction between soil and water. In his 1907 monograph *Studies on the movement of soil moisture*, Buckingham wrote: “Then we may define ψ as the work required per centigram to pull water away from the mass of soil. We shall call ψ the capillary potential of the soil.” Buckingham’s definition of capillary potential has been formulated as a purely mechanical variable (no thermal effects) and for certain applications it remains an operational definition for the driving forces for flow in partially saturated porous media. We surmise that the reason that Buckingham did not pursue a more general definition of the water potential based on Gibbs thermodynamic

formalism was the complexity of gradients in capillary potential relative to gradients of linear transport laws such as Ohm's and Fourier's (for electric and temperature gradients). In reference to capillary flow, Buckingham wrote: "Furthermore, the other factor in the equation, namely, the gradient S , is not a simple and directly measurable quantity like a head of water, an electrical potential, or a temperature. It is a gradient of a quantity ψ , the attraction of the soil for water; and ψ depends in some as yet unknown way, differing from soil to soil, on water content of the soil, which can itself be measured only by tedious and not very accurate methods."

Buckingham had been exposed to Gibbs thermodynamic theory during his PhD studies in Leipzig (he graduated in 1893) under Wilhelm Ostwald. During that time, Ostwald translated Gibbs' work to German and lectured extensively on his interpretation of Gibbs' theory. After his graduation, Buckingham (1900) wrote a textbook on thermodynamics while teaching at Bryn Mawr College, a women's liberal arts college in Pennsylvania. In the preface to the book "An Outline of the Theory of Thermodynamics" that was published in 1900 Buckingham acknowledged the strong influence that Gibbs and Duhem, the French thermodynamicist, had on his writing. This solid background in thermodynamics and recognition of the importance of potential as a driving force for flow and attainment of equilibrium were the foundations for the pioneering work he later published during his tenure with the Bureau of Soils where he wrote his 1907 monograph and introduced the concept of potential to soil physics. Despite the limited scope of the mechanical potential for addressing capillary flows in soil, Buckingham added that "it is obvious that a rigorous treatment of the subject, with no restrictions imposed on either the water content or the soluble salt content of the soils, would have to use thermodynamic reasoning. But the simple conception of a mechanical potential will suffice for present purposes, though it is not impossible that with more comprehensive experimental data available we should have to use thermodynamic potential or the free energy."

It took nearly 4 decades for a more comprehensive theory of water potential to be developed using a broader thermodynamic basis. In their monograph on "The Thermodynamics of Soil Moisture" Edlefsen and Anderson (1943) applied Gibbs' approach to soil and plant physiology. They recognized the limitations of the mechanical potential (proposed by Buckingham) and wrote "Potential, for example, takes no explicit account of the effect of temperature on the total energy change of a system. Fortunately, another function, called 'free energy' was invented many years ago in the field of thermodynamics. This function combines all the criteria and characteristics of both potential and entropy that are most useful in the study of the thermodynamics of soil moisture and its use by plants."

Definition of soil water potential

Soil water is subject to several force fields, the combined effects of which result in a deviation in potential energy relative to the reference state, called the "total soil water potential" (ψ_T) defined as: "The amount of work that an infinitesimal unit quantity of water at equilibrium is capable of doing when it moves (isothermally and reversibly) to a pool of water at similar standard (reference) state, i.e., similar pressure, elevation, temperature and chemical composition."

A more formal definition of water potential requires the use of thermodynamic quantities such as Gibbs free energy G (or molar Gibbs free energy, g). The relations are based on the fundamental equation of internal energy (U) of an open system (i.e., mass of different substances can be exchanged reversibly with the surrounding while T and P vary mildly across the boundaries):

$$dU = TdS - PdV + \sum_{i=1}^j \frac{\partial U}{\partial n_i} dn_i \quad (1)$$

where T is the temperature (K), S the entropy ($J K^{-1}$), P the pressure (Pa) and V is the volume (m^3). The last term is a summation over changes in the amounts of substances, each representing the contribution to the chemical potential μ ($J mol^{-1}$) of the system:

$$\mu_i = \left(\frac{\partial U}{\partial n_i} \right)_{S,V,n} \quad (2)$$

The Gibbs free energy is defined as: $G = U + PV - TS$, and it offers certain advantages over the use of the internal energy (U) for soil water applications. Taking the total differential dG and replacing dU with Eq. (1) yields:

$$dG = VdP - SdT + \sum_{i=1}^j \mu_i dn_i \quad (3)$$

We can normalize the Gibbs free energy, entropy, and volume (the extensive variables) per mol, i.e., g ($J mol^{-1}$), S ($J mol^{-1} K^{-1}$), V ($m^3 mol^{-1}$) to find for a single pure substance (water) that: $d\mu = dg = VdP - SdT$ (Bejan, 1997). In other words, the molar Gibbs free energy is equal to the chemical potential. The foregoing analysis is not meant to offer a rigorous thermodynamic derivation of water potential in soil but to provide a few thermodynamic links with definitions below.

Based on thermodynamic terms, a formal definition of water potential ψ ($J m^{-3}$) that parallels the verbal definition above is given as (Slatyer and Taylor, 1960):

$$\psi = \frac{(\mu_w - \mu_{w,0})}{\bar{v}_{w,0}} \quad (4)$$

with μ_w ($J mol^{-1}$) the chemical potential of water in a soil sample, $\mu_{w,0}$ ($J mol^{-1}$) the chemical potential for pure water at reference state (originally referred to the same temperature as the sample) where the definition is extended to a similar

reference state (temperature, elevation, pressure, and chemical composition). Similarly, the partial molar volume at reference state $\bar{v}_{w,0}$ ($\sim 18 \times 10^{-6} \text{ m}^3 \text{ mol}^{-1} \sim 0.001 \text{ m}^3 \text{ kg}^{-1}$) converts the molar units (J mol^{-1}) to J m^{-3} or units of pressure (Pa).

Components of the water potential

The primary forces acting on soil water held within a rigid soil matrix under isothermal conditions can be conveniently grouped as: (1) matric forces resulting from interactions of the solid phase with the liquid and gaseous phases; (2) osmotic forces owing to differences in chemical composition of the soil solution; and (3) body forces induced by gravitational and other inertial force fields (e.g., centrifugal). The thermodynamic approach, whereby potential energy rather than forces is used, is particularly useful for equilibrium and flow considerations. Equilibrium would require the vector sum of these different forces acting on a body of water in different directions to be zero—this is an extremely difficult criterion to deal with in soils. On the other hand, potential energy, mathematically defined as the negative integral of the force over the path taken by an infinitesimal amount of water when it moves from a reference location to the point under consideration, is a scalar and not a vector. Subsequently, we can express the total potential as the algebraic sum of the component potentials corresponding to the different fields acting on soil water:

$$\Psi_T = \Psi_m + \Psi_s + \Psi_p + \Psi_z \quad (5)$$

where the component potentials ψ_i are discussed below.

ψ_m is the matric potential resulting from the combined effects of capillarity and adsorptive forces within the soil matrix. The primary mechanisms for these effects include: (1) capillarity caused by liquid-gas interfaces forming and interacting within the irregular soil-pore geometry; (2) adhesion of water molecules to solid surfaces due to short-range London-van der Waals forces and extension of these effects by cohesion through hydrogen bonds formed in the liquid; and (3) ion hydration and water participating in diffuse double layers—particularly near clay surfaces. There is some disagreement regarding the practical definition of this component of the total potential. Some consider all contributions other than gravity and solute interactions at a reference atmospheric pressure. Others use a tensiometer to measure and provide a practical definition of the matric potential in a soil volume of interest that is in contact with a tensiometer's porous cup. The value of ψ_m ranges from zero when the soil is saturated to increasingly negative values as the soil becomes drier (note that $\psi_m = 0 \text{ mm}$ is greater than $\psi_m = -1000 \text{ mm}$; in analogy, a temperature of 0°C is greater than -10°C).

The representation of flow and transport in unsaturated soils, particularly at low water contents, commonly lumps capillary and adsorptive forces without distinguishing their individual contributions to the matric potential (Tuller et al., 1999). However, as emphasized by Nitao and Bear (1996), the distinctive consideration of adsorptive surface forces and liquid films is essential for advanced flow and transport theories. Hence, we seek to expand the “capillary-centered” standard treatment by incorporating the contribution of adsorptive forces to the matric potential component of soil water. Following Philip (1977), we thus express the liquid-vapor interface as a surface of constant, partial specific Gibbs free-energy (or matric potential) made up of an adsorptive component (A) and a capillary component (C) (more rigorous representations of these contributions are found in Radke (2015) and Kaptay (2018)):

$$\Psi_m = A(h) + C(\kappa) \quad (6)$$

with κ as the mean curvature of the liquid-vapor interface, and h as the distance from the solid to the liquid-vapor interface taken normal to the solid surface (i.e., thickness of the adsorbed film). The capillary component C is given by the classic Young–Laplace equation:

$$C(\kappa) = \frac{-2\sigma\kappa}{\rho} \quad (7)$$

where κ is positive for an interface concave outward from the liquid, σ is the surface tension at the interface, and ρ is the density of the liquid.

The adsorptive component in Eq. (6) is attributed to two types of surface forces (Derjaguin et al., 1987). The first kind includes longer-range ($>500 \text{ \AA}$) electrostatic forces (e.g., diffuse double layer, DDL), and shorter-range ($<100 \text{ \AA}$) van der Waals and hydration forces, responsible for molecular interactions and structural changes in water molecules near the solid surface. The second kind of surface forces is comprised of longer-range forces due to the overlapping of two interfacial regions (e.g., mutual attraction between two clay platelets across a slit-shaped pore space). The combined effect of such interfacial interactions results in a difference in chemical potentials between the liquid in the adsorbed film and the bulk liquid phase. This difference in chemical potentials may be expressed as an equivalent interfacial force per unit area of the interface, termed the “disjoining pressure” (Π). The disjoining pressure is a function of liquid film thickness (h), and it can also be viewed as the difference between a normal component of liquid film pressure, P_N (in equilibrium with the gaseous phase $P_N = P_G$), and the pressure in the bulk liquid phase, P_L :

$$\Pi(h) = P_N(h) - P_L = P_G - P_L \quad (8)$$

The disjoining pressure is related to more conventional thermodynamic quantities such as Gibbs free energy. Gibbs free energy (G) per unit area of the interface may be defined on the basis of $\Pi(h)$ isotherms for constant pressure P_L , temperature T , and chemical (μ) and electric potentials of the liquid-gaseous and the liquid–solid interfaces as:

$$G(h) = - \int_{\infty}^h \Pi(h) dh \quad (9)$$

The value of $G(h)$ is equal to the work of thinning the film in a reversible isobaric-isothermal process from ∞ to a finite thickness h , with $\Pi(h) = -(\partial G/\partial h)_{T, P_L, \mu, \psi_g, \psi_s}$. The use of $\Pi(h)$ as the basic thermodynamic property is not a mere change of notation, as $\Pi(h)$ has advantages in cases where Gibbs thermodynamic theory is difficult to define, such as when interfacial zones overlap to the extent that the film does not retain the intensive properties of the bulk phase. The use of the disjoining pressure is advantageous from an experimental point of view because of the relative ease in accounting for different contributions (e.g., electrostatic effects).

The disjoining pressure is a sum of several components, similar to the concept of total soil water potential discussed above. The primary components of $\Pi(h)$ in porous media are molecular, $\Pi_m(h)$, electrostatic, $\Pi_e(h)$, structural, $\Pi_s(h)$, and adsorptive $\Pi_a(h)$:

$$\Pi(h) = \Pi_m(h) + \Pi_e(h) + \Pi_s(h) + \Pi_a(h) \quad (10)$$

$\Pi_m(h)$: The molecular component originates from van der Waals interaction between macro-objects (e.g., parallel clay plates). Various expressions, with $\Pi_m(h)$ often proportional to h^{-3} , were derived by [Paunov et al. \(1996\)](#) and [Iwamatsu and Horii \(1996\)](#).

$\Pi_e(h)$: The electrostatic component of the disjoining pressure is calculated from the solution of the Poisson-Boltzmann equation for the DDL with appropriate boundary conditions. Approximate solutions are adequate for many applications, often with $\Pi_e(h) \propto h^{-2}$.

$\Pi_s(h)$: Some controversy exists regarding the origin of the structural component; some attribute it to changes in the structure (density) of water adjacent to solid surfaces and deformation of hydrated shells, while others attribute this force to the presence of a layer with a lower dielectric constant near the surface. Regardless of its exact origin, this component is responsible for the so-called hydration repulsion which stabilizes dispersion and prevents coagulation of some colloidal particles, even at high electrolyte concentrations: $\Pi_s(h) \propto h^{-1}$.

$\Pi_a(h)$: The adsorptive component of the disjoining pressure results from nonuniform concentrations in the water film due to unequal interaction energies of solute and solvent with interfaces in nonionic solutions. This is different than the nonuniform distribution of charged ions. This component of the disjoining pressure is likely to become very important for interactions between nonpolar molecules (e.g., NAPLs) which give rise to repulsive forces in the liquid film.

The form of the disjoining pressure isotherm $\Pi(h)$ is determined by the nature of surface forces. While the molecular component $\Pi_m(h)$ is always present, the influence of other components depends on surface properties, liquid polarity and its composition, and adsorption of dissolved components. The range of the electrostatic forces in dilute solutions of a 1:1 electrolyte (10^{-6} – 10^{-7} mol L⁻¹) is in the range of 0.3–1.0 μ m. Consequently, relatively thick films of water and aqueous electrolyte solutions ($h > 50$ nm) are stable mainly through the $\Pi_e(h)$ component of disjoining pressure. The magnitude and contribution of $\Pi_e(h)$ primarily depend on the charges of the film and substrate surfaces. Dispersion forces become appreciable in the range $h < 50$ nm, and their influence is enhanced by large differences between the permittivity of the liquid and the solid. The forces of structural repulsion may come to play when the film thickness is less than 10 nm.

ψ_s is the solute or osmotic potential determined by the presence of solutes in soil water, which lower its potential energy and its vapor pressure. The effects of ψ_s are important when: (1) there are appreciable amounts of solutes in the soil; and (2) in the presence of a selectively permeable membrane or a diffusion barrier which transmits water more readily than salts. The effects of ψ_s are otherwise generally negligible when only liquid water flow is considered, and no diffusion barrier exists. The two most important diffusion barriers in the soil are: (1) soil-plant root interfaces (cell membranes are selectively permeable); and (2) air–water interfaces, thus, when water evaporates, salts are left behind. In dilute solutions the solute potential, also called the osmotic pressure, is proportional to the concentration and temperature according to:

$$\Psi_s = -RTC_s \quad (11)$$

where ψ_s is in kPa, R is the universal gas constant (0.008314 kPa m³mol⁻¹K⁻¹), T is absolute temperature (K), and C_s is solute concentration (mol m⁻³). A useful approximation which may be used to estimate ψ_s in kPa from the electrical conductivity of the soil solution at saturation (EC_s) in deciSiemens per meter (dS m⁻¹) is:

$$\Psi_s \approx -36EC_s \quad (12)$$

ψ_p is the pressure potential, defined as the hydrostatic pressure exerted by unsupported water saturating the soil above a point of interest. Using units of energy per unit weight provides a simple and practical definition of ψ_p as the vertical distance from the point of interest to the free water surface (unconfined water table elevation). The convention used here is that ψ_p is always positive below a water table, or zero if the point of interest is at or above the water table. In this sense, nonzero magnitudes of ψ_p and ψ_m are mutually exclusive: either ψ_p is positive and ψ_m is zero (saturated conditions), or ψ_m is negative and ψ_p is zero (unsaturated conditions), or $\psi_p = \psi_m = 0$ at the free water table elevation. Note that some prefer to combine the pressure and matric components into a single term, which assumes positive values under saturated conditions and negative values under unsaturated conditions. Based on operational and explanatory considerations, we prefer to adopt the more commonly used separate components protocol.

ψ_z is the **gravitational potential**, which is determined solely by the elevation of a point relative to an arbitrary reference point and is equal to the work needed to raise a body against the Earth's gravitational pull from a reference level to its present position. When expressed as energy per unit weight, the gravitational potential is simply the vertical distance from a reference level to the point of interest. The numerical value of ψ_z itself is thus not important (it is defined with respect to an arbitrary reference level)—what is important is the difference (or gradient) in ψ_z between any two points of interest. This value is invariant of the reference level location.

Soil water is at equilibrium when the net force on an infinitesimal body of water equals zero everywhere, or when the total potential is constant in the system. Though the last statement is a logical consequence of the definitions above, it is not strictly true. Constant total potential is a necessary but not a sufficient condition, and, for thermodynamic equilibrium to prevail, three conditions must be met simultaneously: (1) thermal equilibrium, or uniform temperature; (2) mechanical equilibrium, meaning no net convection-producing force; and (3) chemical equilibrium, meaning no net diffusional transport or chemical reaction. In most practical applications, however, the macroscopic definition of the total potential and equilibrium conditions based on it are completely adequate.

The difference in chemical and mechanical potentials between soil water and pure water at the same temperature is known as the soil water potential ψ_w :

$$\Psi_w = \Psi_m + \Psi_s + \Psi_p \quad (13)$$

Note that the gravitational component (ψ_z) is absent in this definition. Soil water potential is thus the result of inherent properties of soil water itself, and of its physical and chemical interactions with its surroundings, whereas the total potential ψ_T includes the effects of gravity (an “external” and ubiquitous force field).

Total soil water potential and its components may be expressed in several ways depending on the definition of a “unit quantity of water.” Potential may be expressed as (1) energy per unit of mass; (2) energy per unit of volume; or (3) energy per unit of weight. A summary of the resulting dimensions, common symbols, and units are presented in Table 1.

Only μ has actual units of potential; ψ has units of pressure, and h of head of water. However, the above terminology (i.e., potential energy expressions rather than units of potential per se) is widely used in a generic sense in the soil and plant sciences. The various expressions of soil water energy status are equivalent, with:

$$\mu = \frac{\Psi}{\rho_w} = gh \quad (14)$$

where ρ_w is density of water (1000 kg m^{-3} at 20°C) and g is gravitational acceleration (9.81 m s^{-2}).

Measurement of soil water potential components

Total water potential

Psychrometric methods are commonly used for the measurement of the water potential (ψ_w) in soils (also in plants and food). The potential of the soil solution is in thermodynamic equilibrium with its ambient water vapor. Taking the vapor pressure above pure water at reference state ($\psi_w = 0$) as e_0 , the vapor pressure (e) over a salt solution or water held in soil pores by matric forces is depressed relative to the reference state, i.e., $e < e_0$. A convenient measure obtained by the psychrometer is the relative vapor pressure of the ambient soil atmosphere, which is related to the water potential (ψ_w) of soil water through the well-known Kelvin equation:

$$RH = \frac{e}{e_0} = e^{\left(\frac{M_w \psi_w}{\rho_w RT}\right)} \quad (15)$$

where e is water vapor pressure (kPa), e_0 is saturated vapor pressure at the same temperature, M_w is the molecular weight of water ($0.018 \text{ kg mol}^{-1}$), R is the ideal gas constant ($8.31 \text{ J K}^{-1} \text{ mol}^{-1}$ or $0.008314 \text{ kPa m}^3 \text{ mol}^{-1} \text{ K}^{-1}$), T is absolute temperature (Kelvin), and ρ_w is the density of water (kg m^{-3}). Rearranging and taking a log-transformation of Eq. (15) yields an expression for water potential ψ_w :

$$\Psi_w = \frac{RT\rho_w}{M_w} \ln \left(\frac{e}{e_0} \right) \quad (16)$$

Table 1 Expressions for the total soil water potential and its components.

Units	Symbol	Name	Dimensions	SI units	CGS units
Energy/mass	μ	Chemical potential	$L^2 t^{-2}$	J kg^{-1}	erg g^{-1}
Energy/volume	ψ	Soil water potential, suction, or tension	$ML^{-1} t^{-2}$	Nm^{-2} (Pa)	erg cm^{-3}
Energy/weight	h	Pressure head	L	m	cm

L , length, M , mass; t , time

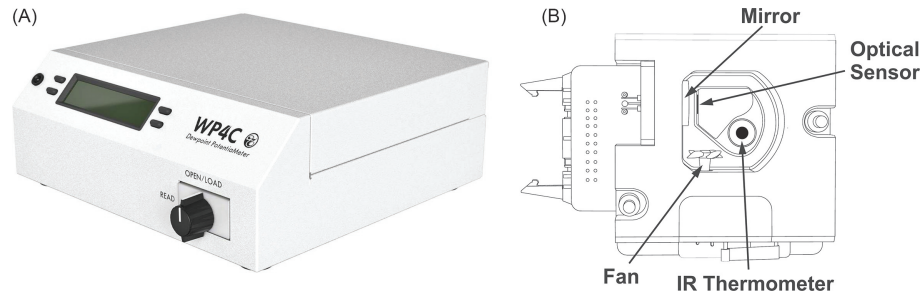


Fig. 1 (A) A WP4C laboratory psychrometer (METER Group, Pullman, WA, USA). (B) Sample chamber of the WP4C psychrometer.

The water potential in drier soils is lower, such that fewer water molecules “escape” into the ambient atmosphere, resulting in lower relative humidity (lower relative vapor pressure). Concentrated soil solutions having lower osmotic potentials have similar effect on reducing vapor pressure, as more water molecules are associated with hydrated salt molecules and are less free to “escape” the liquid state. The inability to distinguish between matric and osmotic effects limits psychrometric measurements to soil water potential only. In some cases where the osmotic potential is negligible, psychrometric measurements are used to infer the matric potential.

Fig. 1A depicts a WP4C laboratory psychrometer (METER Group, Pullman, WA, USA). The stainless steel cup that contains the soil sample is sealed against a sensor block (Fig. 1B). The soil sample equilibrates with the vapor in the headspace. A fan controls the boundary layer conductance of the dew point mirror and shortens equilibration time. An infrared (IR) thermometer measures the sample temperature, and a dew point mirror measures the dew point temperature of the air in the headspace, which allows calculation of the vapor pressure of the air in the headspace as the saturation vapor pressure at dew point temperature. When the water potential of the sample and the headspace air are in equilibrium, the measurement of the headspace vapor pressure and sample temperature (from which saturation vapor pressure is calculated) gives the water potential of the sample according to Eq. (16).

Temperature plays a crucial role for accurate ψ_w measurements. Most critical is the measurement of the difference between sample and dew point temperature. To achieve a ψ_w measurement resolution of 0.05 MPa, temperature differences need to be measured at a resolution of 0.006 °C.

Psychrometers measure the combined matric (ψ_m) and solute (ψ_s) potentials. The solute potential can be approximated based on the electrical conductivity (EC) of the soil solution (Eq. (12)). The matric potential may be calculated by subtracting ψ_s from ψ_w . The measurement range of the WP4C Dewpoint PotentiaMeter is from –100 to –300,000 kPa.

Pressure potential

Piezometers are commonly applied to measure ψ_p . A piezometer (Fig. 2A) is a tube that is inserted into the soil to depths below the water table and that extends to the soil surface and is open to the atmosphere. The bottom of the piezometer is perforated to allow soil water under positive hydrostatic pressure to enter the tube. Water enters the tube and rises to a height equal to that of the unconfined water table. The elevation of the free water table is measured relative to the soil surface using a steel tape with a water level sounder or other electro-optic devices that indicate when the tip of the steel tape touches the water table. The value of pressure potential expressed as energy per weight is simply the vertical distance from a point of interest to the surface of the free water table. Pressure potentials above the water table surface are always zero (nonzero pressure and matric potentials are mutually exclusive). For automated continuous water level measurements modern loggers that rely on pressure transducers for the measurement of the hydrostatic and barometric pressures are commonly employed (Fig. 2B).

Matric potential

Tensiometers or heat-dissipation sensors are commonly applied to measure soil matric potential. A tensiometer consists of a porous cup, usually made of ceramic and having very fine pores, connected to a vacuum gauge through a water-filled tube (Fig. 3). The porous cup is placed in intimate contact with the bulk soil at the depth of measurement. When the matric potential of the soil is lower (more negative) than inside the tensiometer, water moves from the tensiometer along a potential energy gradient to the soil through the saturated porous cup, thereby creating suction sensed by the gauge. Water flow into the soil continues until equilibrium is reached and the suction inside the tensiometer equals the soil matric potential. When the soil is wetted, flow may occur in the reverse direction, i.e., soil water enters the tensiometer until a new equilibrium is attained. The tensiometer equation is given as:

$$\Psi_m = \Psi_{gauge} + (z_{gauge} - z_{cup}) \quad (17)$$

with ψ_{gauge} the reading at the vacuum gauge location and z indicating depth. The vertical distance from the gauge plane to the cup, expressed as a negative quantity when the cup is lower than the gauge as is usually the case, must be added to the matric potential

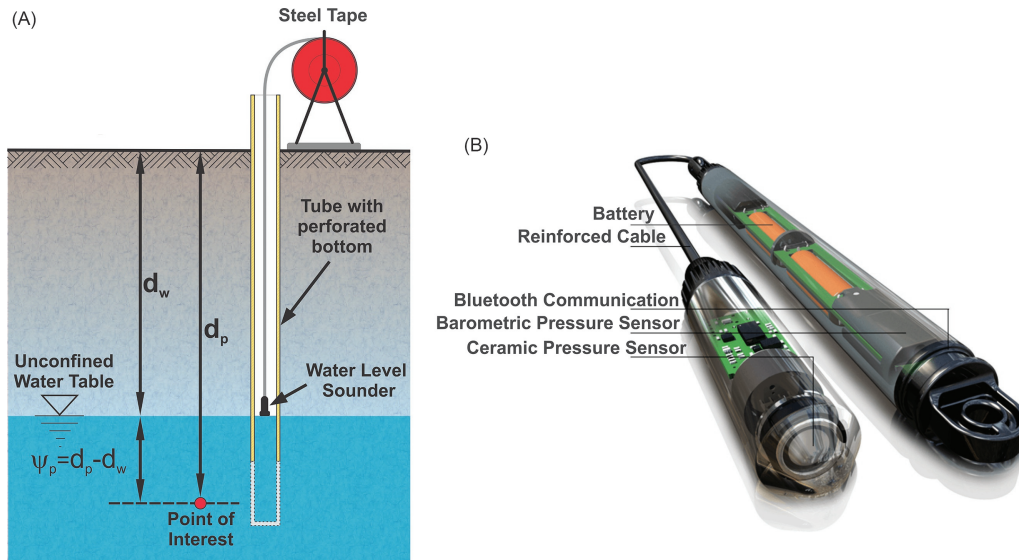


Fig. 2 (A) The concept of piezometer measurements. (B) HOB0[®] MX2001 water level logger (Copyright 1996–2012 Onset Computer Corporation, 470 MacArthur Blvd, Bourne, MA 02532 U.S.A., all rights reserved).

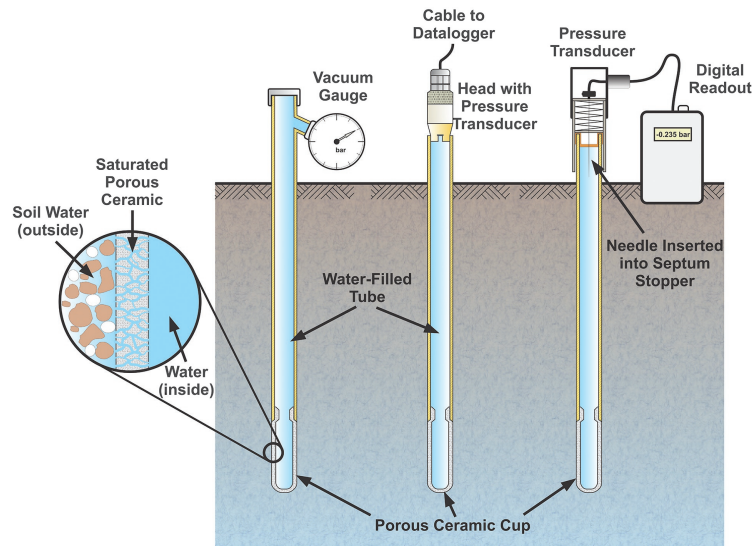


Fig. 3 Tensiometers for matric potential measurement using vacuum gauges and electronic pressure transducers.

measured by the gauge (ψ_{gauge}) in order to obtain the matric potential at the depth of the cup. This accounts for the positive head exerted by the overlying tensiometer water column at the depth of the ceramic cup. Note that using the difference in vertical elevation is appropriate only when potentials are expressed per unit of weight. Electronic sensors called pressure transducers often replace mechanical vacuum gauges. The transducers convert mechanical pressure into an electric signal which can be more easily and more precisely measured. In practice, pressure transducers can provide more accurate readings than other gauges, and in combination with data-logging equipment are able to supply continuous measurements of matric potential.

The tensiometer range is limited to suctions (absolute values of the matric potential) of less than 100 kPa, i.e., 1 bar, 10 m head of water, or 1 atm. This is because the air entry value of the porous cup is equivalent to 100 kPa. Therefore, other means are needed for matric potential measurement under drier conditions.

Heat-dissipation sensors may be applied for a matric potential range from -1 to -1000 MPa. The rate of heat dissipation in a porous medium is dependent on the medium's specific heat capacity, thermal conductivity, and density. The heat capacity and thermal conductivity of a porous matrix are affected by its water content. Heat dissipation sensors contain heating elements in line-

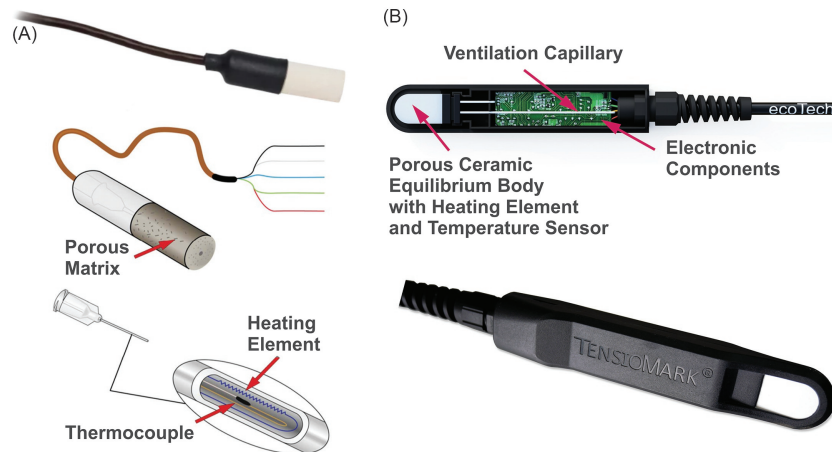


Fig. 4 (A) 229 heat dissipation matrix potential sensor (Campbell Scientific, Inc., Logan, UT, USA). (B) Tensiomark heat dissipation matrix potential sensor (ecoTech Umwelt-Meßsysteme GmbH, Bonn, Germany).

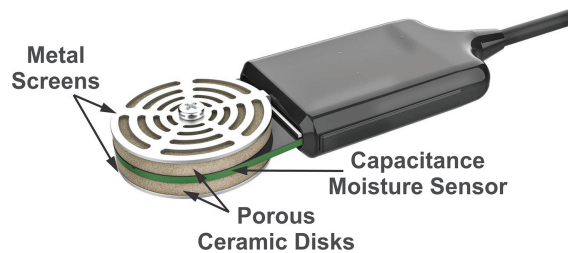


Fig. 5 TEROS 21 Water Potential Sensor (METER Group, Pullman, WA, USA).

or point-source configurations embedded in a rigid porous matrix with fixed pore space. The measurement is based on application of a heat pulse by applying a constant current through the heating element for specified time period, and analysis of the temperature response measured by a thermocouple placed at a fixed distance from the heating source. With the heat dissipation sensor buried in the soil, changes in soil water matric potential result in a gradient between the soil and the porous ceramic matrix, inducing water flow between the two materials until a new equilibrium is established. The water flow changes the water content of the ceramic matrix which, in turn, changes the thermal conductivity and heat capacity of the sensor, and hence the measured temperature response to the applied heat pulse.

Fig. 4A depicts the 229 heat dissipation matrix potential sensor (Campbell Scientific, Inc., Logan, UT, USA) that consists of a heating element and thermocouple placed in epoxy in a hypodermic needle, which is encased in a porous ceramic matrix. The sensor's measurement range is from -10 to -2500 kPa. It requires soil specific calibration and an excitation current of 50 mA. A Tensiomark heat dissipation sensor (ecoTech Umwelt-Meßsysteme GmbH, Bonn, Germany) with a measurement range from -1 to -1000 MPa is shown in Fig. 4B.

Additional means for in-situ measurements of ψ_m , are sensors such as the TEROS 21 (METER Group, Pullman, WA, USA) (Fig. 5). This sensor consists of a capacitance soil moisture sensor that is sandwiched between porous ceramic disks with known water characteristic. Metal screens attached to the ceramic disks prevent fringing of the electromagnetic field into the surrounding soil. The ceramic disks are brought in close hydraulic contact with the surrounding soil, which allows the ψ_m of the soil and porous ceramic substrate to equilibrate. The ψ_m is inferred from the measured volumetric water content of the ceramic substrate and its known water characteristic. The measurement range of the TEROS 21 sensor is from -5 to $-100,000$ kPa.

Osmotic potential

Soil water solutions contain varied quantities and compositions of dissolved salts. The relationships between the salt concentration and ψ_s , and the possibility for estimating ψ_s from the electrical conductivity (EC) of the soil solution, were discussed above. Conventional measurement of soil solution EC involves solution extraction from saturated soil samples and measuring the EC using an electrical conductivity meter (Fig. 6A).

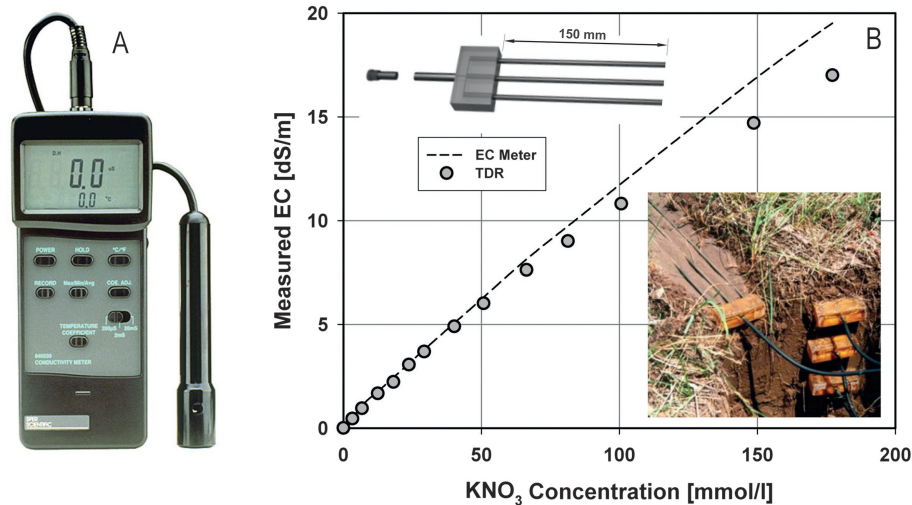


Fig. 6 (A) Handheld electrical conductivity (EC) meter; (B) Time domain reflectometry (TDR) probes and solution EC vs concentration measured with TDR and EC meter. Reproduced from Mmolawa, K.B. and Or, D. (2000) Root zone solute dynamics under drip irrigation: a review. *Plant and Soil* 222: 163–190.

Electrical conductivity meters rely on Ohm's law:

$$E = I R \quad (18)$$

with E the electromotive force (volts), I the current flow (amperes), and R the resistance (ohms). For constant voltage, the current flowing through the solution is inversely proportional to the electrical resistance, or directly proportional to the electrical conductance. The solution EC is thus determined from known voltage and electrode geometry and measurement of the electric current. More recently, a variety of in situ methods, such as time-domain reflectometry (TDR), have been used to deduce soil bulk EC from electromagnetic signal attenuation (Mmolawa and Or, 2000), hence enabling simultaneous measurements of water content and soil EC in the same undisturbed soil volume (Fig. 6B). Concurrent knowledge of θ and EC can be used to infer the soil solution EC, and hence to estimate ψ_s .

Summary

The soil water potential framework is central to the quantification of the energy state of soil water and features prominently in the driving forces for water and vapor flow in soil. Soil water potential is also important for exchanges with the atmosphere and for soil-plant interactions including plant response to changes in water availability (a similar concept applies to flow and transport within plants). We provided a historical context for the origins of the soil water potential and presented the various components that contribute to soil water potential. We reviewed and discussed contemporary measurement methods used for soil water potential determination in laboratory and field conditions.

References

- Bejan A (1997) *Advanced Engineering Thermodynamics*, 2nd ed. New York: John Wiley & Sons.
- Buckingham E (1900) *An Outline of the Theory of Thermodynamics*. New York: The MacMillan Company.
- Buckingham E (1907) *Studies on the Movement of Soil Moisture*. Washington D.C.: US Department of Agriculture, Bureau of Soils, 61.
- Derjaguin BV, Churaev NV, and Muller VM (1987) *Surface Forces*. New York: Plenum.
- Edlefsen NE and Anderson ABC (1943) *Thermodynamics of Soil Moisture*. vol. 15, Hilgardia, 31–298.
- Iwamatsu MI and Horii K (1996) Capillary condensation and adhesion of two wetter surfaces. *Journal of Colloid Interface Science* 182: 400–406.
- Kaptay G (2018) The chemical (not mechanical) paradigm of thermodynamics of colloid and interface science. *Advances in Colloid and Interface Science* 256: 163–192.
- Mmolawa KB and Or D (2000) Root zone solute dynamics under drip irrigation: a review. *Plant and Soil* 222: 163–190.
- Nitao JJ and Bear J (1996) Potentials and their role in transport in porous media. *Water Resources Research* 32(2): 225–250.
- Paunov VN, Dimova RI, Kralchevsky PA, Broze G, and Mehreteab A (1996) The hydration repulsion between charged surfaces as an interplay of volume exclusion and dielectric saturation effects. *Journal of Colloid Interface Science* 182: 239–248.
- Philip JR (1977) Unitary approach to capillary condensation and adsorption. *Journal of Chemical Physics* 66(11): 5069–5075.
- Radke CJ (2015) Gibbs adsorption equation for planar fluid–fluid interfaces: Invariant formalism. *Advances in Colloid and Interface Science* 222: 600–614.
- Slatyer RO and Taylor SA (1960) Terminology in plant- and soil-water relations. *Nature* 187: 922–924.
- Tuller M, Or D, and Dudley LM (1999) Adsorption and capillary condensation in porous media: Liquid retention and interfacial configurations in angular pores. *Water Resources Research* 35(7): 1949–1964.

Water content and water potential measurement

Wolfgang Durner, Institute of Geocology, Technische Universität Braunschweig, Braunschweig, Germany

© 2023 Elsevier Ltd. All rights reserved.

This is an update of GS Campbell, CS Campbell, Water Content and Potential, Measurement[☆], *Reference Module in Earth Systems and Environmental Sciences*, Elsevier, 2013, ISBN 9780124095489, <https://doi.org/10.1016/B978-0-12-409548-9.05333-1>.

Introduction	169
Terminology and units of soil water content and soil water potential	169
Measurement of water content	170
Reference method: Gravimetric measurements	170
Reference method: Volumetric measurements	171
Sensor technology for dielectric properties	171
Time domain reflectometry (TDR)	171
Capacitance probes	173
Frequency domain reflectometry (FDR)	173
Sensor technology based on electric conductivity	173
Sensor technology for thermal properties	173
Sensor technology for nuclear methods	174
Noninvasive techniques	174
Measurement of water potential	174
Tensiometers—direct potential measurement in the wet range	175
Solid-Phase Sensors (SPS): Potential measurement in the medium suction range	177
The dry range: Determining soil water potential from vapor pressure	178
Measurement characteristics of water potential sensors	180
Concluding remarks	181
Appendix: Market overview of commercial sensors with Relevant Websites	181
References	186

Abstract

In this article, I briefly review techniques for measuring the soil water content and soil water potential. Whereas a large number of methods have been developed to measure these properties in the past, only a few techniques are of interest nowadays, are commercially available, and used as standard sensors. Water content is nowadays primarily measured with sensors that measure the electric permittivity of soils. For soil water potential, no single technique is suited to cover the whole range of suctions. Depending on the wetness, tensiometers, equivalent porous media sensors, or psychrometric methods are suited.

Key points

- The water content and matric potential of the soil are measured in situ by placing electronic sensors in the soil.
- Today, the measurement of water content in natural soils is mostly based on the measurement of permittivity.
- Water content measurements using sensors embedded in the soil are highly influenced by local variability of soil properties and sensor placement conditions. Furthermore, they encompass only small sample volumes.
- Matric potentials in soils can be determined very accurately under wet conditions using tensiometers, but these instruments fail under dry conditions.
- Commercially available matric potential sensors for the medium moisture range are based on the principle of equilibrating an equivalent porous medium. In practice such sensors cover water potential ranges between -10 and -5000 kPa. They are widely available, but have rarely been tested against each other under real field conditions.
- Measurement of total water potential under dry conditions is possible but rarely applied in the field. It can be measured using psychrometric techniques, but the instruments are affected considerably by temperature conditions.
- The link between large-scale soil moisture monitoring, e.g., by cosmic ray probes or satellite microwave monitoring and local soil moisture monitoring is an ongoing research topic.

Nomenclature

Gravimetric water content w [kg kg ⁻¹]	Mass of water per mass of dry soil.
Gravitational potential ψ_g [kPa]	Energy per volume of water caused by an elevation difference to a reference plain. Positive above the reference plain, negative below it.
Hydraulic potential Φ [kPa]	Sum of matric and gravitational potential, i.e., $\Phi = \psi_m + \psi_g$ resp. $H = h_m + h_g$.
Hydrostatic potential ψ_h [kPa]	Energy per volume of water exerted by positive fluid pressure due to a surcharge of water at water saturation, positive.
Matric potential ψ_m [kPa]	Energy per volume of water exerted on the water phase by capillary forces and water-solid interactions, negative in unsaturated soils.
Osmotic potential ψ_o [kPa]	Energy per volume of water caused by the presence of solutes in the water phase, always negative.
pF	Logarithmic unit for suction, defined as $\text{pF} = \log_{10}(h \text{ [cm]})$.
Pressure head h [m]	Energy per weight of water. All water potential components can be expressed in units of pressure heads.
Suction	Absolute value of the matric potential, either in units of pressure [kPa] or in units of pressure head [m]
Tension	Synonym to suction.
Total potential ψ_t or h_t	Sum of all potential components, i.e., $\psi_t = \psi_m + \psi_g + \psi_o + \psi_h$ in pressure units, of $h_t = h_m + h_g + h_o + h_h$ in units of pressure head.
Volumetric water content θ [m ³ m ⁻³]	Volume of water per volume of soil.

Introduction

Soil moisture influences the exchange of energy, water, solutes, and carbon within the soil-vegetation-atmosphere system (e.g., Vereecken et al., 2008). Two variables define soil moisture, namely water content and water potential. Both, soil water content and soil water potential, fundamentally affect a wide variety of biophysical processes, such as seed germination, plant nutrition, and plant wilting. Soil water content controls to a great extent soil mechanical, thermal, chemical and biological properties. Soil water potential, in contrast, is the key variable that determines dissipative water movement in soil and the availability of water to plants. Hydrological processes, such as infiltration, redistribution, evaporation, transpiration, deep percolation and groundwater renewal, are caused by gradients in soil water potential (Bittelli, 2010). Knowledge of the relationship between water content and matric potential, which is the most important component of the total water potential in unsaturated soils, is required to model the soil water balance. Such modeling provides an indispensable tool for many applications in agricultural and horticultural management, such as managing irrigation, minimizing fertilizer use and loss, and optimizing agricultural production. It is also key for assessing contaminant transport in soil and water exchange between the soil-atmosphere interface, which relates to weather forecast and local climate. In this article, we briefly review techniques for measuring the soil water content and soil water potential. Although a large number of experimental methods have been developed to measure these quantities over the last century, there are now only a few techniques that are used widely, commercially available and used as standard sensors.

Terminology and units of soil water content and soil water potential

Soil water content is a measure of the amount of water (volume or mass) contained in a unit volume or mass of soil. If the measure is the volume of water per unit volume of soil, the water content is called the ‘volume wetness’ or ‘volumetric water content’, θ [m³ m⁻³]. If it is the mass of water per unit mass of dry soil it is called the ‘mass wetness’ or ‘gravimetric water content’, w [kg kg⁻¹]. The gravimetric water content is given by the mass of water per unit mass of dry soil. The two water content measures are related by:

$$\theta = \frac{\rho_b}{\rho_w} w \quad (1)$$

where ρ_b [kg m⁻³] is the bulk density of the soil (mass of dry soil per unit volume) and ρ_w [kg m⁻³] is the density of water ($\approx 10^3$ kg m⁻³). The volumetric water content can range from 0 to a maximum value that is equal to the porosity of a soil. In typical soils, porosity is between 0.35 and 0.7, but in specific media (e.g. peat) the porosity can reach values of >95% and accordingly volumetric water contents also can be this large. Very often, water contents are listed without units (volume per volume or mass per mass cancel each other out) and are expressed as percent (θ in the range 0 ... 100%). However, this is poor practice and expressing units minimizes interpretation errors.

Soil water potential reflects the energy state of water in porous media and incorporates capillarity, adsorption, gravity, osmosis and positive pressure under saturated conditions. This energy state is given relative to water in a standard state, in which it is free of solutes, free from external forces except gravity, and at a reference pressure, reference temperature and reference elevation. In particular, it is free from any forces that can be exerted to the water phase by nearness to the soil matrix, such as solid-liquid interphase forces (capillary forces). These forces provide a key component of soil water potential, and are summarized as matric potential, ψ_m (Luo et al., 2022). The elevation of a water-filled point in soil with respect to a reference elevation (often the soil surface) determines the gravitational potential (ψ_g). Presence of solutes in the soil solution decreases the potential energy of water with respect to solute free water and is expressed as osmotic potential (ψ_o). Under saturated conditions, water overlying over a point of interest exerts a pressure, which is termed soil hydrostatic potential (ψ_h) (also called pressure potential). When quantifying the water potential, at least these components have to be taken into account, and the total water potential ψ_t is thus given by

$$\psi_t = \psi_m + \psi_g + \psi_o + \psi_h \quad (2)$$

Note that in Eq. (2) only ψ_m or ψ_h can be different from zero, since under water-saturated conditions ψ_m is zero by definition, and in unsaturated conditions there is no pressure from overlying water, i.e., $\psi_h = 0$. Thus, ψ_h is a positive pressure component compared with the matric potential, which is negative. Other forces that act on the water in a porous medium and contribute to total potential are the air pressure in pores, expressed as pneumatic potential, and the force that acts on water in the case of a load of an unconsolidated matrix, which is termed overburden potential. However, the latter two pressures do not prevail in typical soils, since air pressure is assumed to be equal in a continuous air phase, and particle load is in most soils held by particle-particle contacts. Thus, they can normally be neglected. Finally, the osmotic pressure is in most cases neglected and assumed to show significant gradients in soils only in the case of strong salt content gradients, which rarely occurs under humid climates. However, when considering water uptake by plant roots in drying soils, the osmotic component may become important. In any situation, the matric potential and the gravitational potential must be taken into account. These two components are combined to form the 'hydraulic potential' which is a key variable in any simulation of water transport in unsaturated porous media.

To quantify the energy state of soil water, it must be related to an amount of water. It can be expressed as energy per mass of water (J kg^{-1}), as energy per volume of water (J m^{-3}), which is equivalent to pressure (Pa), or as energy per unit weight of water ($\text{J (9810 kg m s}^{-2})^{-1}$). The latter unit appears not straightforward on first sight, but since $1 \text{ J} = 1 \text{ kg m}^2 \text{ s}^{-2}$, it leads to the expression of the water potential as *pressure head*, in units of water column length (m or cm). This is convenient, and thus it is commonly used in soil physics and vadose zone hydrology (Durner and Or, 2005). The relation between pressure unit and length unit is given by $\psi = \rho_w gh$ where $g = 9.81 \text{ m s}^{-2}$ is the gravitational acceleration. In soil physics practice, often used units for pressure head are cm and hPa, and we can set $h = 1 \text{ cm}$ approximately equal to $\psi = 1 \text{ hPa}$.

Soil water potential in unsaturated conditions is a negative number, because it requires energy to transfer soil water from an unsaturated soil to the reference state. To avoid the negative sign and to represent the matric potential as positive number, which also allows to express the potential on a logarithmic scale, the term "suction" and "tension" are used to express pressures or pressure heads as positive variables. Dependent on the state of soil dryness, the matric potential and thus the energy state of soil water typically ranges over several orders of magnitude. It is thus convenient to express the potential on a log scale, which intrinsically requires to express the potential as positive number, i.e. as suction. Specifically in mainland Europe, the use of the "pF" value has become established and is frequently used:

$$pF = \log_{10}(|h|) \text{ with } h \text{ in [cm]} \quad (3)$$

where h is the matric potential expressed as pressure head in the unit cm. Although the pF unit may have a somewhat clumsy definition, its usage is widely understood and practical.

The water suction defines a cascade of key moisture states, such as field capacity, permanent wilting point, air dry soil and oven dry state. Table 1 gives a synoptic overview to these states. The uppermost line (pF 0) indicates the almost fully saturated state, where the matric pores are all filled with water and only pores with diameters $>300 \mu\text{m}$, i.e., so-called macropores, are drained. At the other end of the moisture scale are the air dry and oven dry state. These are somewhat affected by the humidity of the ambient air, but the listed values (pF 6.0 and pF 6.9) can be seen as typical for normal environments (e.g., $\sim 50\%$ relative humidity at 20°C).

Measurement of water content

Reference method: Gravimetric measurements

The reference measurement for water content is obtained by determining the mass loss by evaporation of water from a soil sample during oven drying. Samples are typically dried for at least 24 h at 105°C , or longer if mass loss has not ceased. Note that at this condition not all water molecules in a soil are removed from the system. Rather, it is an operational difference of water content between an actual state and a dryness that is defined by the oven dry reference state, as indicated in Table 1. Also, one should be aware that air-dry soil, which is dried at room or outside temperature rather than in an oven, does not have a water content of zero. The difference between air-dry and oven-dry water content can be small for coarse sands but large for fine-textured soils.

Table 1 Water potential units.

	ψ [kPa]	h m H_2O	pF	Pore diameter [μm]	Relative humidity
Quasi saturated	0.1	0.01	0.0	3000	1.00000
	1.0	0.10	1.0	300	0.99999
FC (near GW)	6.0	0.61	1.8	50	0.99996
	10	1.0	2.0	30	0.99993
FC (far from GW)	33	3.4	2.5	9	0.99976
	100	10	3.0	3	0.99926
	1000	102	4.0	0.3	0.99264
PWP	1500	153	4.2	0.2	0.98898
	10^4	1'020	5.0	0.03	0.92877
Air dry	$\sim 10^5$	$\sim 10'000$	~ 6		0.47763
Oven dry	$\sim 8 \cdot 10^5$	$\sim 80'000$	~ 6.9		0.00062

GW: groundwater. FC, field capacity, a typical value for the drained upper limit of water potential in soil; value of pF 1.8 is used in conditions close to groundwater; value of pF 2.5 is used traditionally for conditions far from groundwater. PWP, a typical value for the lower limit of plant-available water in soil. Value for "Air dry" is typical but varies with atmospheric humidity. Value for "Oven dry" is typical but varies with oven humidity and oven temperature. "Pore diameter" is diameter of the largest water-filled cylindrical pore at the indicated potential. Relative humidity is computed assuming a temperature of 20 °C.

Apart from errors in weighing soils before and after oven drying, the two most serious errors come from loss of mass of nonaqueous volatiles in the sample and uncertainty of the point at which the sample is really "dry". Increasing or decreasing the drying temperature changes the mass loss, and this change can be substantial in some soils. Variation in oven humidity can also lead to some uncertainty. Typical errors are on the order of 0.005 kg kg^{-1} . There is a number of faster drying methods which give quicker results at the expense of some accuracy. According to [Campbell and Campbell \(2005\)](#), drying of samples in a microwave oven can decrease the time required for a measurement to approximately 20 min, while still giving acceptable accuracy for many purposes.

Reference method: Volumetric measurements

Mass wetness is easily determined in laboratory studies, but has little relevance in the field, where one is mainly interested in the ability of the soil to store water for later transpiration and evaporation. Here volume wetness is the relevant measurement. The direct method for measuring volume wetness is to determine the water mass through oven-drying, computing from it the water volume, and relating this to a sample of known volume. Soil cores taken by pressing sharpened steel cylinders of 100 cm^3 , 250 cm^3 or even larger are used to obtain samples of known volume. In collecting the sample, utmost care must be taken not to compress the soil core by the sampling process. From the known volume and the soil dry mass, the bulk density is calculated. The volume wetness is then obtained using Eq. (1). Since this is a tedious and time-consuming measurement, which furthermore is subject to large uncertainty owing to the large spatial variation found in the field, volume wetness is often inferred in situ from measurements of other soil properties.

Sensor technology for dielectric properties

Destructive sampling to measure soil water content is time-consuming, can interfere with plots and only measures one point in time. Sensors connected to data loggers are therefore commonly used to measure water content. Dielectric sensors determine the relative electric permittivity, ϵ [–] (formerly: dielectric constant) of the soil. The permittivity is then converted to a soil moisture content measurement using a calibration equation, which can be either empirical ([Topp and Davis, 1985](#)) or based on theoretical considerations ([Roth et al., 1990](#)). The relative dielectric permittivity of water is approximately 80 at microwave frequencies, whereas it is between 3 and 5 for soil minerals, organic matter, and ice. A measure of soil permittivity is therefore strongly influenced by the amount of liquid water present in the soil. Since ice has a dielectric permittivity near that of soil minerals, dielectric methods measure unfrozen water content in frozen soil.

Dielectric sensors are commonly used for soil moisture measurement in research and agricultural applications because they are (as such) relatively accurate, durable, and can provide continuous measurements over time. Thus, sensors of this type have become dominant for the past 30 years or so for soil moisture monitoring purposes. Three methods are typically used to measure the permittivity of the soil: time domain reflectometry (TDR), frequency domain reflectometry (FDR), and capacitance. [Fig. 1](#) gives an overview of water content sensors that were used in a sensor comparison study by [Jackisch et al. \(2020\)](#). A variety commercially available sensors is listed in [Table A1](#) in the appendix.

Time domain reflectometry (TDR)

The TDR sensor works by sending a high-frequency electromagnetic signal down a probe inserted into the soil. Parallel conductors are placed in the soil forming a transmission line, and one the time required for an electromagnetic pulse to traverse the

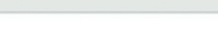
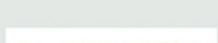
	System name	Manu- facturer	Measurement principle ^{*1}	Probe length, integral volume ^{*2}	Range	T ^{*3}	Image
Soil water content	Trase TDR Trase	TRASE	TDR (tangent intersection of pulse in 10 ps)	20 cm, ≈ 1000 cm ³	0 – 1 m ³ m ⁻³		
	Trime Pico32 Pico32	IMKO	TDR (time sampling of 1GHz TDR pulse in 3 ps)	11 cm, ≈ 250 cm ³	0 – 1 m ³ m ⁻³	•	
	Trime Pico64 Pico64	IMKO		16 cm, ≈ 1250 cm ³	0 – 1 m ³ m ⁻³	•	
	Trime T3P T3P	IMKO	(11 cm guides on tube probe)	11 cm, ≈ 1000 cm ³	0 – 1 m ³ m ⁻³		
	ThetaProbe ML2x	Delta-T	I (impedance of 100 MHz signal)	6 cm, ≈ 75 cm ³	0 – 0.5 m ³ m ⁻³	•	
	HydraProbe Hydra	Stevens	I (impedance of 50 MHz signal)	4,5 cm, ≈ 40 cm ³	0 – 1 m ³ m ⁻³	•	
	10HS	METER (Decagon)	I (impedance of 70 MHz signal)	10 cm, ≈ 1300 cm ³	0 – 0.57 m ³ m ⁻³		
	5TM	METER (Decagon)		5 cm, ≈ 715 cm ³	0 – 1 m ³ m ⁻³	•	
	EC5	METER (Decagon)		5 cm, ≈ 250 cm ³	0 – 1 m ³ m ⁻³		
	WET 2 WET	Delta-T	I (20 MHz signal on central rod)	6.8 cm, ≈ 500 cm ³	0 – 1 m ³ m ⁻³	•	
	Profile Probe PR2/6 PR2	Delta-T	I (100 MHz signal on pairs of steel rings)	6x along tube probe 5 cm, ≈ 3100 cm ³	0 – 1 m ³ m ⁻³		
	EnviroSCAN JKI	Sentek	I (100-500 MHz signal on pairs of steel rings)	≈ 6 cm, ≈ 350 cm ³	0 – 0.65 m ³ m ⁻³		
	TMS3 buriable TMS3	Tomst	TDT (no. of pulses received of 2.5 GHz transmis- sion frequency)	≈ 10 cm	0 – 1 m ³ m ⁻³	•	
	TMS Tomst	Tomst		≈ 10 cm	0 – 1 m ³ m ⁻³	•	
	SMT100 SMT	Truebner	Freq. of TDT ring oscillator	≈ 10 cm	0 – 0.6 m ³ m ⁻³	•	

Fig. 1 Selection of water content sensors, as used in the comparative study of Jackisch et al. (2020). From Jackisch C, Germer K, Graeff T, Andrä I, Schulz K, Schiedung M et al. (2020). Soil moisture and matric potential—an open field comparison of sensor systems. *Earth System Science Data* 12(1): 683–697. Reproduced with courtesy of Copernicus.

transmission line, reflect from the distal end, and return along the same line is measured. The time delay between the transmitted and reflected signals is used to determine the speed of propagation of the electromagnetic disturbance. This speed is controlled by the dielectric permittivity of the surrounding medium. It is computed from

$$\varepsilon = \left(\frac{ct}{2L} \right)^2 \quad (4)$$

where c is the speed of light ($3 \times 10^8 \text{ m s}^{-1}$), t [s] is the time to traverse the transmission line, and L [m] is the length of the transmission line. TDR measurements obviously require very fast response circuitry. The time resolution must be on the order of 100 ps. Equipment with this kind of time resolution is necessarily relative expensive. TDR, due to its higher-frequency excitation, is less influenced by electrical conductivity of the soil than the capacitance-based probes. See also Chapter entitled Time-Domain Reflectometry.

Capacitance probes

Capacitance probes provide a cheaper alternative since slower circuitry is connected to the transmission line. A voltage pulse is applied through a resistor of value R and the transmission line acts like a capacitor, with the soil as the dielectric. The time constant for charging the capacitor (time to charge to 63% of its final voltage) is RC , where C is the capacitance of the probe. The capacitance is related to geometric factors and is directly proportional to the dielectric permittivity of the medium surrounding the probe. The charging time is therefore directly related to the soil permittivity. Capacitance probes are relatively inexpensive and easy to operate. Furthermore, the sensor geometry is very adaptable, facilitating the development of a variety of configurations and making them particularly usable for use in solid phase (SPS) matric potential sensors (see section “Measurement of water potential”). However, capacitance probes are influenced by soil type and require calibration. Also, there is concern about the influence of soil salinity and soil temperature on capacitance sensors (Kelleners et al., 2004).

Frequency domain reflectometry (FDR)

FDR probes are also used to measure the capacitance, based on the change in frequency of a reflected radio wave or resonant frequency (Kelleners et al., 2004). In this measurement method, an oscillator is used to pass an electromagnetic signal through a metal waveguide. FDR probes are considered accurate, but must be calibrated for the type of soil in which they are buried. Compared to TDR probes, they offer a faster response time. However, the term FDR is often misused because most frequency sensors use a single frequency rather than a frequency range.

Sensor technology based on electric conductivity

When accuracy must be weighed against affordability and a cheap sensor is needed, a Raspberry Pi moisture sensor is an option. The physical measurement principle involves measuring the electrical conductivity or resistance of the soil to determine its moisture content. The physical measurement is usually implemented using simple circuits and analog-to-digital converters (ADCs) to interface with the Raspberry Pi. The analog readings from the sensors are converted into digital values that can be processed and interpreted by the Raspberry Pi for further analysis. The method provides a general indication of soil moisture. It is not as accurate as more advanced commercial sensors, but can still be useful for simple projects or experiments in various fields such as agriculture, gardening, or environmental monitoring. Adjustments and calibrations are required depending on the application and environmental conditions.

Sensor technology for thermal properties

Both thermal conductivity and heat capacity of soil vary with water content. The relationship with heat capacity is an easy one to use for determining water content. The water content can be computed from the volumetric heat capacity of soil:

$$\theta = \frac{C - C_s}{C_w} \frac{\rho_b}{\rho_s} \quad (5)$$

Here C , C_s , and C_w [$\text{J m}^{-3} \text{ K}^{-1}$] are the volumetric heat capacities of the bulk soil, the soil minerals, and water; and ρ_b [kg m^{-3}] and ρ_s [kg m^{-3}] are the bulk and particle densities. Since C_w is typically about twice as large as C_s , changes in water content have a fairly large effect on C . If the bulk density of the soil is known and C is measured, all values on the righthand side of Eq. (5) are known, so water content can be computed. Typically, C is measured using a dual-needle thermal properties sensor. In the recent Tensiomark™ sensor, the sensor element is imbedded between two porous ceramic plates. The sensor head is glued into a approx. 125 mm long plastic case, in which the electronics are situated. Fig. 1 includes this and other sensors of the heat dissipation type.

As indicated above, an impressive variety of dielectric and thermal sensors is commercially available (Table A1). A common problem to all such sensors is the rather limited support volume. Since soil-embedded sensors measure the properties of the dielectric within the electromagnetic field of the probe or transmission line, it is important that a representative sample of the soil lies within the electromagnetic field of the sensor. Unfortunately, this field lies close to the surface of the sensor, so any air gaps or compaction around the sensor may adversely affect the measurement. This problem is less given for sensors based on gamma ray attenuation or thermalization of fast neutrons, described next.

Sensor technology for nuclear methods

Gamma rays and neutrons both interact strongly with soil water. The methods used most widely for water-content measurement involve attenuation of transmitted gamma radiation and backscattering of neutrons, though neutron transmission and gamma backscatter have also been used. Gamma attenuation is typically a laboratory method. A soil column, 5–20 cm thick, is irradiated by a collimated beam of gamma radiation. Sources with appropriate energy levels are ^{137}Cs and ^{241}Am . The volume wetness is computed from:

$$\theta = \frac{\ln\left(\frac{I_d}{I_w}\right)}{\mu_w S} \quad (6)$$

where I_d and I_w are the numbers of gamma photons passing through the dry and wet soil in unit time, respectively, μ_w is the mass attenuation coefficient for water, and S is the soil-column thickness. It is possible to determine the bulk density and the water content of a sample without measuring the gamma attenuation in the dry soil column by measuring the attenuation of two gamma beams having different energies.

Neutron scattering is a method for measuring volume wetness in the field. In a typical application, a neutron source such as $^{241}\text{Am}/\text{Be}$ is lowered into an access tube in the soil. The neutrons from the source initially have high energy, but as they collide with low-atomic-mass nuclei (mainly hydrogen) in the soil, they slow and become ‘thermal’ neutrons, with energies typical of atoms at room temperature. These slow neutrons can be counted with a relatively simple detector. Since the main source of hydrogen nuclei in the soil is water, and since the number of scattered neutrons is proportional to the number of collisions with these hydrogen nuclei, there is a relationship between neutron count and water content:

$$\theta = a + b \frac{I}{I_{std}} \quad (7)$$

where a and b are soil-specific empirical constants, I is the count in soil, and I_{std} is the count in a standard scattering medium. Neutron scattering is much less affected by air gaps around the sensor than are dielectric methods. The scattering volume (i.e., the signal support volume) for the neutrons varies with soil-water content from approximately from a sphere of 15 cm diameter in saturated soil to, e.g., 70 cm in dry soil.

Noninvasive techniques

The spatial characteristics of soil moisture data can be characterized using the “scale triplet” proposed by Blöschl and Sivapalan (1995), encompassing support, spacing, and extent. The support refers to the integration volume of the measurement methods, the spacing to the distance between single measurements, and the extent to the area over which measurements are available (e.g., area of measurement network). Embedded soil moisture sensors have very small support, and spacing and extent depends on the number of sensors that can be placed and managed. To obtain moisture patterns at the km scale relevant to land surface and atmospheric models, they are of limited use.

Intensive progress is thus currently on-going with respect to techniques that allow continuous noninvasive and contactless measurements of soil moisture dynamics at the field to basin scale. According to Bogaen et al. (2015) these are (1) cosmic-ray neutron probes, (2) Global Navigation Satellite System reflectometry, (3) ground-based microwave radiometry, (4) gamma-ray monitoring, (5) terrestrial gravimetry, and (6) low-frequency electromagnetic surface waves. The treatment of these techniques is beyond the scope of this Chapter. However, we must be aware that any of these techniques produces proxies of soil moisture with different characteristics. They all depend on ground truth measurements for calibration, and they all are also affected more or less by vegetation, snow, atmospheric conditions, land surface properties, and other factors.

As a single example, cosmic ray neutron sensing receives currently considerable attention in the scientific community. It is used to measure the average soil moisture at the horizontal scale of 10^2 m and depth of a few decimeters. A stationary probe measures neutrons produced by cosmic rays and moderated by hydrogen atoms, which are mainly found in soil water, and emitted into the atmosphere. There they are mixed on a scale of hundreds of meters and inversely correlated in density with soil moisture. A neutron detector captures all neutrons of a specific medium energy range that cross its path. To get meaningful readings, the instruments capture the surrounding neutrons over a period of hours. Since their number is very small, natural fluctuations introduce large errors into the measurements. In addition, water that is bound in plants or even snow falsify the measured values. Even cattle that curiously approach the sensors can be a source of error.

Measurement of water potential

Water potential in soils ranges from a pressure head $h = 0$ at full saturation to air dry ($\sim \text{pF } 6$) or even oven dry ($\sim \text{pF } 6.9$) conditions, where water content is zero. The available methods for soil water potential measurement differ with respect to the suction range in which they are applicable. No single instrument or method exists that is able to capture soil water potential measurement across the full range, from saturation to oven dryness. In the wet and moist range, the pore water pressure of the soil is measured directly with tensiometers. In the medium range of moisture content, porous reference media (solid phase sensors, SPS) in which the degree of water saturation changes with water potential, are suitable. Gypsum blocks, granular matrix sensors, heat dissipation matric

potential sensors and filter paper are examples. In relatively dry soils, water potential can be derived from vapor pressure measurements by psychrometry. All methods for measuring soil water potential require hydraulic equilibrium between soil water in the domain in contact with a sensor and the measurement device itself. The contact requires connection of the water phase from the soil to the inside of the sensor, except for psychrometers, where contact is provided by the vapor phase. In this section, we describe the principles of these measurement techniques, with an emphasis on tensiometer measurements.

Tensiometers—direct potential measurement in the wet range

Tensiometers for matric potential determination are historically the earliest used measurement technologies for soil water potential in the field. The fundamental principles go back to the seminal work of Willard Gardner, while Lorenzo A. Richards is credited with developing the first durable tensiometer design suitable for field use in the 1920s. According to Or (2001), historical evidence suggests that the initial design was actually proposed even earlier, by Burton E. Livingston as far back as 1908. Tensiometers are particularly suited for measurement of soil water energy status at high water contents where transport processes are relatively rapid (e.g., solute transport from the vadose zone down to groundwater). However, due to their limited measurement range, they cannot be used in drier conditions, where soil water becomes limiting for plant growth. Under relatively moist soil conditions, tensiometric measurements are robust, reliable, and accurate, with a considerable body of theoretical and practical knowledge gained by soil physicists, hydrologists, manufacturers, and other practitioners. However, tensiometers require maintenance and their use at frozen conditions is problematic. In the following, we discuss principles of tensiometric measurements, tensiometer design and measurement range, and show recent developments. Reviews on all aspects of the practical applicability of tensiometers are available for further information, e.g., by Young and Sisson (2002), Durner and Or (2005), and Whalley et al. (2013).

A tensiometer generally consists of a water-filled reservoir connected to a pressure sensor (manometer or electronic pressure transducer), which is brought into contact with the soil via a rigid porous membrane that is permeable to water and impermeable to air. In most cases, the membrane is in the form of a porous ceramic cup with size of a few cm, and pure water is used as the internal pressure transfer fluid. The water phase outside the cup and the liquid in the interior of the tensiometer are bridged through water-filled fine pores of the cup. Water flows between the surrounding soil and the water-filled reservoir until the suction inside the tensiometer is equilibrated to the soil's matric potential. In modern tensiometers, the interior is almost perfectly rigid. If no air bubbles are present, the change in water volume in response to matric potential changes is negligible, resulting in almost instant equilibrium and a quick response time of the instrument. Following pressure equilibration, the liquid pressure in the tensiometer is directly measured. To function properly, the pores of the membrane must remain impermeable to the non-wetting fluid (air), i.e., they must remain at all operation states completely saturated with water. Tensiometers measure the matric potential component of the water potential. The pore openings of the membranes are in the μm range, which is still about 6000 times larger than the size of a hydrated Na molecule (Young and Sisson, 2002), making them permeable to dissolved ions. Therefore, the pressure measured inside the porous cup is not affected by the osmotic potential component. Still, the pores are small enough to withstand an air entry up to some bar pressure difference.

Installation of tensiometers is straightforward. The porous cups are usually connected to a shaft of specific length that allows them to be inserted at a desired measurement depth into pre-drilled access holes. The diameter of the access hole at the target depth is slightly smaller than that of the porous cup to ensure adequate contact. In principle, a partial contact between the cup and the surrounding material would be sufficient, since the contact area does not influence the equilibrium pressure, but it will affect the response time. Also, a change of the material properties near the cup, for example by compression of the surrounding soil, has minimal influence on the measured value of the soil water potential, because the energy status is continuous for different porous materials in hydraulic equilibrium. In stony soils, where the contact area of the soil's water phase to the tensiometer cup can be very small, it may be advantageous to facilitate water pressure transfer by inserting a slurry of silica flour into the borehole. In contrast to the water content, the water potential in soils does not change abruptly at the material boundaries. Despite this spatial continuity of the state variable, however, experience shows that, especially in the presence of roots, the local variability of the matric potential can become very high as soon as the soils begin to dry out.

In general, it is not the absolute pressure of the soil water that is of interest, but the capillary pressure, which is defined as the pressure difference between the water and the gas phase in the soil. Therefore, the pressure sensors used for tensiometers are usually differential sensors that measure the difference between the pressure in the tensiometer cell and the ambient air pressure. This atmospheric reference air pressure is in most cases equal to the gas pressure at the location of the measurement. However, there are cases where rainfall traps a layer of air between a sharp wetting front and the groundwater, significantly increasing the atmospheric pressure. Furthermore, in soils that are very wet and where the continuity of the gas phase is limited, the local pressure of the trapped air is not identical to the ambient air pressure, adding a pneumatic component to the water potential. When tensiometric pressure is interpreted as matric potential in such cases, it is a misunderstanding and contributes to a class of phenomena called "dynamic effects" (Durner et al., 2014).

In the past, U-manometer tensiometers with pressure measurement by mercury, mechanical manometer tensiometers and septum tensiometers were used, but they have lost their importance, not least because of poor sensitivity (see section "Measurement characteristics of water potential sensors", below). Modern tensiometers measure the water pressure in the reservoir with electronic pressure transducers. Features of such tensiometers are integrated amplifier, digital signal conditioning for the electrical output, temperature compensation, filling level indication, and the possibility for external refilling in situ. Fig. 2 shows the setup of a commercially available modern tensiometer for field use with a cup diameter of 20 mm (left), and a tensiometer for laboratory use

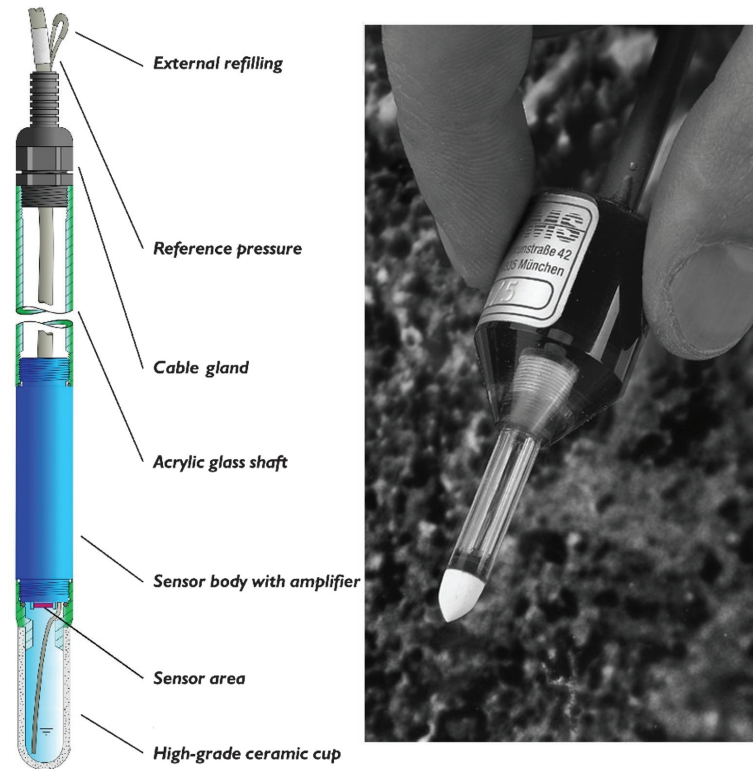


Fig. 2 Advanced pressure transducer tensiometers. Left: Pressure transducer tensiometer T8 by UMS. Right: Mini tensiometer T5 by UMS. Reproduced by permission of METER Group AG, Munich, Germany.

with a cup diameter of 5 mm (right). Comparative studies have shown that the size of the tensiometer cup is irrelevant for measuring matric potential, and apart from the problem of mechanical robustness, small tensiometers can perform excellently, even in field use (Jackisch et al., 2020).

The main disadvantage of tensiometers is their limited measuring range. Under thermodynamically stable conditions, the requirement that the water inside the tensiometer must be able to transfer the suction, requires an absolute liquid pressure which is above the vapor pressure at ambient temperature. When the absolute pressure in the tensiometer goes below the water vapor pressure, cavitation of the water phase occurs followed by spontaneous boiling of water. This limits the measurable suction to the pressure difference between the actual ambient atmospheric pressure and the internal vapor pressure. Once a bubble is there, it expands and water is sucked through the porous membrane into the dry surrounding soil, and the tensiometer acts essentially as a micro-irrigation device. In practice, measurements are often reliable only up to ~ 80 kPa ($\sim pF$ 2.9). However, thermodynamic equilibration can be delayed, expressed as a boiling delay (deferescence). This metastable state can prevail for extended periods of time and allows an extension of the tensiometric measurement range. The theoretical limit of suction is given by the energy required for cavitation in a pure water phase, which requires a calculated tensile strength of 140 MPa (Zheng et al., 1991). Physically reachable maximum suctions are far from this limit, but Temperley and Chambers (1946) reported that thoroughly degassed liquids in smooth-walled containers can have a tensile strength of up to 10% of the theoretical value. Ridley and Burland (1994), at Imperial College London, developed a miniature pressure transducer in which the volume of water between the pressure transducer and the porous stone was minimized and reported measured water potentials down to -1500 kPa with a response time of a few minutes. However, it appears that reliable tensiometers of this type are not commercially available. Practical limits with the best commercial laboratory tensiometers filled with completely gas-free water are in the range of -100 to -400 kPa (Schindler et al., 2010).

In an effort to extend the measurement range of tensiometers, the introduction of so-called osmotic tensiometers has been repeatedly reported. In principle, a tensiometer cup with much narrower pores, which acts as a semipermeable membrane, is filled with a solution with a high osmotic potential, such as polyethylene glycol. In contact with free water, a very high overpressure builds up in the tensiometer interior, e.g., of 1500 kPa. In contact with soil, the soil water suction reduces this positive pressure, theoretically until cavitation occurs inside of the osmotic tensiometer. Earliest efforts date back to the Peck and Rabbidge (1969) who used polyethylene glycol with a molecular weight of 20,000 to extend the measurement range to a water potential of -1500 kPa. Experience has shown that the practicability of the osmotic tensiometer is limited and many difficulties stand in the way of a routine application, despite recent attempts with improved membrane materials and sensor configurations. Key problems

are temperature sensitivity, sensor drift and leakages of the osmotic solutions. Recent developments and an overview about the state of the technique are reported by Van Der Ploeg et al. (2009) and Rahardjo et al. (2021).

Solid-Phase Sensors (SPS): Potential measurement in the medium suction range

Measurement of soil matric potential in the intermediate pressure range (-10 kPa to -5000 kPa) poses considerable experimental challenge. Solid phase sensors (SPS) consist of equivalent porous media (gypsum blocks, granular matrix material, felt, filter paper) embedded in the soil. The soil water pressure of the surrounding soil equilibrates with the pressure inside the porous media by drawing water out of the SPS (when drying) or releasing it into the SPS (when wetting). Internal water content is measured and the corresponding suction in the sensor is calculated from an internal calibration function (actually the water retention curve of the SPS material). In essence, the problem of suction measurement is replaced by a water content measurement in a sensor. The principles for measuring internal water content in the SPS can vary, and basically any principle discussed for direct water content sensors can be used for measurement in SPS. Common sensors use either electrical resistance, heat dissipation, or dielectric properties of the sensor material.

Classic equivalent porous media materials are filter papers, blocks of gypsum, fiberglass or nylon. Also, granular matrix sensors have become available which consist of gypsum wafers embedded in granular matrix (Scanlon et al., 2002). SPS measure only the matric potential as they are permeable for the gas phase in their operation range and neither pneumatic potential nor osmotic potential contribute to the signal. Since SPS require a certain amount of water to flow from the sensor into the soil (drying) or from the soil into the sensor (in a wetting period) there are system-related disadvantages in terms of response time. In addition, hysteresis is unavoidable, so that in phases of transient moisture changes of the surrounding soil the sensor's accuracy may be limited.

The pore-size distribution of the SPS material can be customized to provide a high sensitivity in a desired matric potential range. Modern SPS are suited to measure potentials in the range -10 to -1000 kPa. Promises by manufacturers that their sensors will cover larger ranges should be treated with skepticism. For any SPS, there is a certain suction threshold, below which the porous medium of the sensor is fully saturated and cannot register changes in matric potential. On the other side of the suction range, above a certain suction, the sensor's porous material is rather dry and again further changes in suction will no longer lead to a significant change of water content. Or and Wraith (1999) and Noborio et al. (1999) identified a need for porous materials having a wide range of pore sizes. Or and Wraith (1999) reported a tradeoff between the sensor's matric potential range and its sensitivity to changes in the surrounding soil. Wraith and Or (2001) proposed that equivalent porous media of sensors should have similar particle and/or pore-size distributions as the surrounding soil under investigation.

In the following, we shortly characterize traditional Gypsum blocks, dielectric matric potential sensors, head dissipation sensors, and for historic reasons the filter paper method. Dielectric and heat-dissipation undergo an intense commercial development. A current list is provided in Table A2, and a couple of sensors used in a field comparison study by Jackisch et al. (2020) is shown in Fig. 3.

Gypsum blocks with embedded electrodes to measure the electric resistance have been used for more than 80 years in agricultural applications (Buoyoucos and Mick, 1940). The electrical conductivity in these blocks is buffered by the ion strength of the saturated gypsum solution (~ 0.2 S/m). This is contrary to nylon and fiberglass blocks, where the electrical conductivity is dependent also on the ionic strength of the soil solution. The problems of gypsum block sensors are a limited temporal stability, a long response time, appreciable hysteresis, and the need for individual calibration. Furthermore, the measurements are temperature dependent and



Fig. 3 Example of a capacitance sensor embedded into a porous medium. Reproduced by permission of METER Group Inc., Pullman, WA.

must be corrected accordingly. Handling and disadvantages of these probes are well known and documented (Scanlon et al., 2002), and it appears that not much development has taken place over recent decades.

Dielectric matric potential sensors are made up of a porous equivalent medium in which the water content is measured by TDR, FDR or capacitance technology. Early attempts were recorded by Or and Wraith (1999) and Noborio et al. (1999) who worked with embedded TDR. Nowadays, a wide range of sensors are commercially available in which the internal water content is determined using the capacitance method, and great care is taken to calibrate the sensors with respect to the matric potential. An example of such a sensor is shown in Fig. 4. One difficulty with this type of sensor is placement in undisturbed soil. Unlike tensiometers, it is not sufficient to simply drill a narrow hole from the soil surface to the desired depth; a trench is required from which the sensor is inserted into the soil to accommodate the cabling and sensor geometry.

Although the material properties in the immediate vicinity of the sensor do not directly affect the transmission of water suction from the soil to the sensor, the practical inconvenience of creating a significant disturbance in order to place the sensor should still be taken into account. Basically, this problem is similar to heat dissipation sensors described next.

Heat-dissipation sensors (HDS) were probably first described by Phene et al. (1971) and were critically evaluated by Reece (1996). Here, the matric potential is indirectly inferred by measuring the water content of a porous media based on its thermal properties. The temperature dependence of the heat conductivity makes the heat dissipation sensors very sensitive to the ambient temperature, although corrections of the effect have been developed (Flint et al., 2002). HDS have an outer matrix made of ceramic or felt (e.g., Matile et al., 2013) that encases a stainless-steel needle. Inside the needle is an electrical resistor that runs the length of the probe (typically few cm) and a temperature sensor, either a thermocouple or thermistor. To measure thermal conductivity, a known amount of current is passed through the resistor for a preset amount of time and data are collected on the temperature change over time. The thermal conductivity can be calculated from the slope of the temperature versus the logarithm of time. Because the sensor's matrix properties remain fixed, the thermal conductivity is a measure of the water content of the matrix, which, in turn, is related to the water potential of the soil. These sensors must be calibrated and, at midrange water potentials, have a strong temperature dependence.

Finally, the *filter-paper method* is mentioned here for historical reasons. Contrary to all SPS methods above, it is not suitable for monitoring, but only for a single measurement of a matric potential in a destructive context. A single piece of Whatman No. 42 filter paper is brought in contact with a soil sample until equilibrium, which generally takes 2–3 days. The filter paper is then weighed and dried. The water content of the paper is used to calculate the matric potential using an empirical relationship between the water content of the filter paper and its water potential (Campbell and Campbell, 2005).

The dry range: Determining soil water potential from vapor pressure

If soil water is in temperature and vapor equilibrium with a vapor phase, the potential in the vapor phase becomes the same as that of the liquid phase. The potential in the vapor phase can be determined by measuring its relative humidity. The relation is expressed by the Kelvin equation (Or and Wraith, 1999)

$$RH = e/e_0 = \exp(M_w g h / RT) \quad (8)$$

where RH [–] is the relative humidity, e [Pa] is water vapor pressure, e_0 [Pa] is saturated vapor pressure at the same temperature, M_w is the molecular weight of water ($0.0018 \text{ kg mol}^{-1}$), g is the gravitational acceleration (9.81 m s^{-2}), R is the ideal gas constant ($8.31 \text{ J K}^{-1} \text{ mol}^{-1}$), and T is the absolute temperature [K]. The relative humidity of the air can be determined either from the wet-bulb depression or from the dew point temperature, using a chilled mirror. Since air is an effective diffusion barrier for most solutes, the corresponding water potential includes the osmotic and the matric potential, which is contrary to water tensiometers and SPS sensors.

Psychrometers measure the difference between a dry bulb and wet bulb temperature. The dry bulb is at the temperature of the surrounding soils, the wet bulb at the temperature of an evaporating surface. The lower the humidity, the higher will be the rate of evaporation from the wet bulb, and thus the temperature depression below ambient. Rearranging (Eq. 8) and taking a log-transformation leads to

$$h = \frac{RT}{M_w g} \ln(e/e_0) \quad (9)$$

For the range of e/e_0 near 1, which is the usual state in soils in humid climates, Eq. (9) can be simplified to

$$h = \frac{RT}{M_w g} \left(\frac{e}{e_0} - 1 \right) = 0.471 \cdot 10^6 \text{ T} \left(\frac{e}{e_0} - 1 \right) \quad (10)$$

with h in m (Rawlins and Campbell, 1986).

To determine h , it is feasible to measure the wet-bulb depression or dew-point depression of the air in the headspace. However, for soil that is moist enough to support plant growth, the relative humidity is typically above 0.99 (Table 1). This means that the psychrometer or dew-point meter used must be exceptionally precise and sensitive with respect to wet vs. dry temperature measurements. To obtain precise psychrometric water potential measurements of around -10^5 kPa , the accuracy required for temperature difference measurements is 0.005°C .

	System name	Manufacturer	Measurement principle ^{*1}	Probe length, integral volume ^{*2}	Range	T ^{*3}	Image
Matric potential	T4	METER (UMS)	Direct tension of water at pressure transducer	6 cm	0 – 850 hPa		
	T5	METER (UMS)		0.6 cm	0 – >1000 hPa		
	T8	METER (UMS)		6 cm	0 – 850 hPa	•	
	TS1 TS	METER (UMS)		6 cm	0 – 850 hPa	•	
	SIS	METER (UMS)	Electric resistance in equivalent porous medium	6 cm	0 – 2000 hPa	•	
	WATER-MARK Gypsum	Irrrometer		8.2 cm	0 – 2000 hPa		
	MPS-1	METER (Decagon)	I (impedance of 70 MHz signal in equivalent porous medium)	4.5 cm	100 – 5000 hPa	•	
	MPS-2	METER (Decagon)		4.5 cm	90 – ∞ hPa	•	
	MPS-6	METER (Decagon)		4.5 cm	90 – ∞ hPa	•	
	TensioMark TM	ecoTech	Heat pulse dissipation in porous membrane	1 cm	1 – 6500 hPa	•	
	Heat Dissipation HeatD	bambach		1 cm	1 – 6500 hPa	•	
	pFMeter	ecoTech	Heat pulse dissipation in equivalent porous medium	4 cm	1 – ∞ hPa	•	
	Polymer Tensiometer POT	Wageningen University	Direct tension of hydrophile polymer	3 cm	0 – 1.6 MPa		

*1) TDR = time-domain reflectometry, I = impedance measurement of capacitance, TDT = time-domain transmission.

*2) Est. assuming a cylindrical volume. *3) Probe also measures temperature. Image copyrights by the manufacturers

Fig. 4 Selection of water potential sensors, as used in the comparative study of Jackisch et al. (2020). From Jackisch C, Germer K, Graeff T, Andrä I, Schulz K, Schiedung M et al. (2020). Soil moisture and matric potential—an open field comparison of sensor systems. *Earth System Science Data* 12(1): 683–697. Reproduced with courtesy of Copernicus.

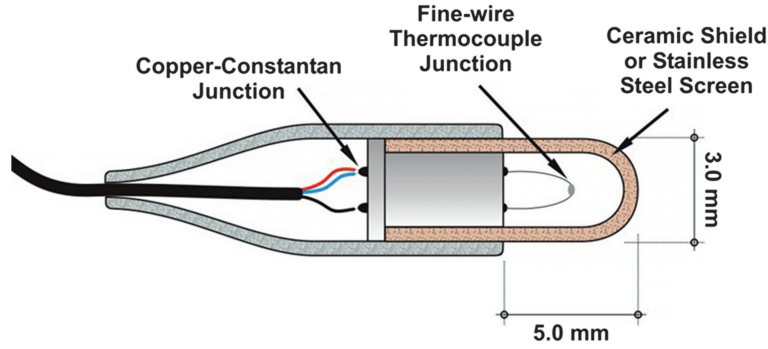


Fig. 5 Field thermocouple psychrometer. From: Muñoz-Carpena R (2004) *Field devices for monitoring soil water content: BUL343/AE266*, 7/2004. Edis 2004(8).

A soil psychrometer for use in the field consists of one junction of the thermocouple suspended in a thin-walled porous ceramic or stainless screen cup buried in the soil (Fig. 5). Another thermocouple is embedded in an insulated plug to measure the ambient temperature at the same location. The *RH* in the air chamber is calculated from the “wet bulb” vs a reference “dry bulb” temperature difference. By an electrical current, the suspended thermocouple is cooled below the dew point until water condenses on the junction. The cooling current then stops, and as water evaporates, it draws heat from the junction, depressing it below the temperature of the surrounding air until it attains wet bulb temperature. The difference in temperatures between the wet and dry bulb is related to the relative humidity by the psychrometer equation

$$\frac{e}{e_0} = 1 - \left(\frac{s + \gamma'}{e_0} \right) \Delta T \quad (11)$$

where s is the slope of the saturation water vapor pressure curve ($s = de_0/dT$), γ is the psychrometric constant ($\sim 0.067 \text{ kPa K}^{-1}$ at 20°C), and ΔT is the temperature difference (K) between the dry and wet bulb.

Due to their sensitivity to thermal gradient effects, psychrometers are not suitable for use at shallow soil depths. Additionally, the requirement for equilibrium among different phases causes psychrometers to have a relatively slow response time.

Theoretically, it is not necessary to calibrate the calculation of the total water potential for specific soils since it is based on fundamental physical principles. However, in practice thermocouple psychrometers are calibrated empirically using solutions of known water potential. The measurement range for water potential toward the wet end is approx. -30 to 200-kPa , whereas in the dry range it reaches -300 MPa . For further details, the reader is referred to Andraski and Scanlon (2002).

In lab instruments, water potential is measured by equilibrating the liquid phase water of a sample with the vapor phase water in the headspace of a closed chamber. The dew point depression technique has proven to be a fast and accurate method for determining water potential. A widely used instrument of that type is shown in Fig. 6 (e.g. Gubiani et al., 2013; Schelle et al., 2013). It uses a chilled mirror dew point technique to measure the water potential. A thermoelectric (Peltier) cooler controls the mirror temperature. A beam of infrared light is directed onto the mirror and reflected back to a photodetector, which detects the change in reflectance when condensation occurs on the mirror. A thermocouple attached to the mirror accurately measures the dew-point temperature. Since both dewpoint and sample surface temperatures are simultaneously measured, the need for complete thermal equilibrium is eliminated, and measurement time is reduced to less than 5 min. Readings are reported directly in MPa, with accuracy for the water potential range of 0 to $-10 \text{ MPa} \pm 0.1 \text{ MPa}$ and from -10 MPa to $-60 \text{ MPa} \pm 1\%$. This reflects exceptional accuracy of differential temperature measurements, as an accuracy of $\pm 0.1 \text{ MPa}$ requires a dew point temperature accuracy of $\pm 0.001 \text{ K}$.

Measurement characteristics of water potential sensors

All of the water potential measuring instruments discussed in this chapter affect the signal to be measured. The measurement characteristic of an instrument is determined by the relation between the sensitivity and the conductance of the instrument. The sensitivity expresses the ratio between the change of the signal (pressure change Δp) with respect to the amount of water ΔV that is required to flow from the surrounding soil to (or from) the instrument (Richards, 1949),

$$S = \Delta p / \Delta V \quad (12)$$

The conductance is dependent on the area of exchange between the instrument and the soil, and on the flow resistance of the instrument. In the case of tensiometers, this resistance is determined by the thickness and the saturated conductivity of the porous membrane. The conductance of SPS instruments is established by the (un)saturated conductivity function of the equivalent porous medium (EPM). Additionally, the capability of the adjacent soil to deliver the necessary water flow needs to be considered.

For modern advanced tensiometers that are perfectly filled, the measurement characteristic is very good for the whole measurement range, due to the high sensitivity of the pressure transducers and the fact the ΔV is close to zero. In the case of SPS with equivalence porous media, it is dependent on the volume and porosity of the sensor, its unsaturated conductivity, but also on



Fig. 6 WP4C Dewpoint PotentialMeter. Reproduced with courtesy of METER Group Inc., Pullman, WA.

the unsaturated conductivity of the surrounding soil. In principle, the sensitivity of SPS is lower, since the water content of the sensor must change. In an ideal SPS, the porosity characteristic of the sensor would be equal to the soil itself (Wraith and Or, 2001). For psychrometric methods, the measurement characteristic might be limited by the required equilibration of the gas phase vapor pressure and temperature.

Concluding remarks

This chapter highlights techniques for measuring soil water content and energy state. All described techniques are limited to small volumes of measurement. In the future, there is potential for larger scale estimation of water content in areas and volumes, particularly near the earth-atmosphere interface, through the use of proxies such as ground penetrating radar, electrical tomography, cosmic ray sensing, or passive microwave radiation mounted on satellites or drones. Despite this, local in situ measurements will remain necessary to obtain accurate ground truth and understand vertical flow and transport processes in soils. To this end, there has been ongoing development of commercially available water content sensors using dielectric methods (Table A1), including sensors with wireless communication and profile probes. Nonetheless, the primary challenge in the field remains soil heterogeneity in conjunction with small measurement volumes and the practical issue of correctly placing the sensors in the soil undisturbed (Stevenson et al., 2021). Consequently, interpreting absolute water content from individual sensors remains uncertain.

The challenge of measuring water potential is particularly difficult. The problem of accurately measuring suction in the field over a significant distance or area remains unsolved, compounded by the fact that no single matric potential sensor can cover the entire suction range from saturation to air dryness. Tensiometers provide measurements in suction ranges of about 0–10² kPa, SPS sensors cover (in an optimistic view) the range 10¹–10⁴ kPa, and thermocouple psychrometers can achieve measurements within a range of 10^{2.5}–10⁵ kPa. The accuracy of modern tensiometers is excellent, but their practical use is restricted, as they need to be refilled when they run dry, and cannot function until the soil moisture conditions improve. To overcome these limitations, technical development focuses on SPS sensors despite their principal drawbacks (Table A2). However, there is still a lack of rigorous scientific proof of their long-term reliability in field measurements, and in some cases, the results of studies appear to contradict some manufacturers' claims (Jackisch et al., 2020).

Appendix: Market overview of commercial sensors with Relevant Websites

See Tables A1 and A2.

Table A1 Market overview of water content sensors.^a

No.	Company	Sensor designation	VWC Accuracy	Measuring range	EC Accuracy	Measuring range	T Accuracy	T Measuring range	Measuring method ^b	Transmission	Cost ^c	Spec sheet
1	Soil Scout Oy Ltd.	Soil Scout	±2%	n.s.	±0.2 dS/m	0–20 dS/m	±0.1 °C	–40 to +80 °C	Capacitance/FDR	Wireless	\$6099 (incl. 6 × sensors, 1 × base station, 1 × echo repeater)	https://campaigns.soilscout.com/wireless-soil-sensor-get-real-time-soil-data?gclid=CjwKCAjwoMSWBhAdEiwAVJ2ndvNv5t8MYIFy-4u_EH6vaGz7FuINH1v6ZlwaAE2h3FsqAxJlf2MU2RoCTggQAvD_BwE#lp-pom-block-2
2	Spectrum Technologies	SMEC-300	3% @ EC <8 mS/cm	0% to approx. 65% vol. H ₂ O (with calibration for substrates)	2%	0–10 mS/cm	0.6 °C	–50 to +85 °C	Capacitance/FDR	Cable 1.8 m, can be extended up to 15 m	€249.00	https://www.specmeters.com/weather-monitoring/sensors-and-accessories/sensors/soil-moisture-sensors/smec300/
3	Spectrum Technologies	SM-100	3% at EC <8 mS/cm	0% to approx. 65% vol. Water content (with calibration for substrates)	–	–	–	–	Capacitance/FDR	Cable 1.8 m, can be extended up to 15 m	€105–115	https://www.specmeters.com/weather-monitoring/sensors-and-accessories/sensors/soil-moisture-sensors/sm100/
4	METER	TEROS 11	±0.03 m ³ /m ³ (±3.00% VWC) typical in mineral soils that have solution EC <8 dS/m Medium specific calibration ±0.01–0.02 m ³ /m ³ (±1–2% VWC) in any porous medium	Mineral soil calibration: 0.00–0.70 m ³ /m ³ Soilless media calibration: 0.0–1.0 m ³ /m ³	–	–	±0.5 °C from –40 to 0 °C ±0.3 °C from 0 to +60 °C	–40 to +60 °C	Capacitance/FDR	Cable, 5–20 m	\$198–239	https://www.metergroup.com/en/meter-environment/products/teros-11-soil-moisture-sensor
5	METER	TEROS 12	Generic calibration: ±0.03 m ³ /m ³ (±3.00% VWC) typical in mineral soils that have solution EC <8 dS/m	0.00–0.70 m ³ /m ³	±5% +0.01 dS/m from 0 to 10 dS/m ±8% from 10 to 20 dS/m	0–20 dS/m (bulk)	±0.5 °C from –40 to 0 °C ±0.3 °C from 0 to +60 °C	–40 to +60 °C	Capacitance/FDR	Cable, 5–20 m	\$234–276	https://www.metergroup.com/en/meter-environment/products/teros-12/teros-12-tech-specs
6	Sentek Technologies	Drill and Drop	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	Capacitance/FDR	Wireless	\$250–350 per sensor \$500–2500 per data logger	https://sentektechnologies.com/products/soil-data-probes/drill-drop/
7	Sentek Technologies	Diviner 2000	n.s.	Oven dry to saturation	–	–	–	–	Capacitance/FDR	Portable system incl a data display unit and a portable probe	\$250–350 per sensor \$500–2500 per data logger	https://sentektechnologies.com/wp-content/uploads/2022/02/SEN5878_A4_Diviner_4LR.pdf
8	Acclima Inc.	TDR-310 W	–2 to +2%	0–100%	(0–1000 uS/cm): –25 +25 uS/cm (1000–2500 uS/cm): –2.5 +2.5% (2500–6000 uS/cm): –5 +5%	0–6000 uS/cm	(+5 to +35 °C): –0.25 +0.25 °C (–15 to +50 °C) –0.5 +0.5 °C	–40 to +60 °C	TDR	Cable	\$250–350 per sensor \$1000–3500 per data logger	https://acclima.com/wp-content/uploads/2022/02/SEN5878_A4_Diviner_4LR.pdf
9	Spectrum Technologies	FieldScout TDR 350	±3.0%	Up to 98.3% depending on soil type	±0.1 mS/cm	0–5 mS/cm	n.s.	–30 to +60 °C	TDR	Handheld device	€1/779.05	https://www.specmeters.com/soil-and-water/soil-moisture/fieldscout-tdr-meters/fieldscout-tdr-350-soil-moisture-meter-with-case/
10	Campbell Scientific	CS655	±1% (with soil-specific calibration) where solution EC <3 dS/m ±3% (typical with factory VWC model) where solution EC <10 dS/m	0–100%	±5% of reading +0.05 dS/m	0–8 dS/m	±0.1 °C (for typical soil temperatures [0–40 °C] when probe body is buried in soil) ±0.5 °C (for full temperature range)	–50° to +70 °C	TDR	Cable	\$250–350 per sensor \$1000–3500 per data logger	https://www.campbellsci.cc/cs655

11	Campbell Scientific	SoilVUE™10	±1.5% typical with most soils Soils with high organic matter (>12% soil organic carbon) or high clay content (>45% clay) may need a soil-specific calibration due to the dispersive nature of these materials	0–100%	±2% (0–2.5 dS/m) ±5% (full range)	0–10 dS/m	±0.15 °C (between –30° and +40 °C)	n.s.	TDR	Cable	\$250–350 per sensor \$1000–3500 per data logger	https://www.campbellsci.cc/soilvue10
12	Campbell Scientific	CS650	±1% (with soil-specific calibration) ±3% (typical with factory VWC model) where solution EC <3 dS/m	0–100%	±5% of reading +0.05 dS/m	0–3 dS/m	±0.1 °C (for typical soil temperatures [0–40 °C] when probe body is buried in soil) 0.5 °C (for full temperature range)	–50° to +70 °C	TDR	Cable	\$250–350 per sensor \$1000–3500 per data logger	https://www.campbellsci.cc/cs650
3	Soilmoisture Equipment Corp (SEC)	TRASE System 1	±2% FS	0–100%	–	–	–	–	TDR	Cable	n.s.	https://ictinternational.com/products/trase-system-1/tr-trase-system-1/
14	Soilmoisture Equipment Corp (SEC)	HandiTrase handheld TDR	±2% FS	0–100%	–	–	–	–	TDR	Handheld device	n.s.	https://www.soilmoisture.com/HT2/
15	IMKO	TRIME-PICO 64/32	±2% (0–40%) ±3% (40–70%)	0–100%	±2% (0–6 dS/m) ±3% (6–20 dS/m)	0–6 mS/cm	±0.2 °C	–40 to +70 °C	TDR	Cable	n.s.	https://www.lombardemarozzini.com/sites/default/files/pico_flyera4_english_web_0.pdf
16	Grodan	GroSens	5% V/V	0–100%	±0.5 mS/cm	0–10 mS/cm	±1 °C	0–+50 °C	Multi-sensor measuring system	Wireless	€2666.67 for handheld meter €1100 for sensor €933.33 for receiver €555.00	https://www.grodan101.com/siteassets/grodan-101/downloads/grosens_brochureup_final_2019.pdf
17	Delta-T	ML3 ThetaProbe	±0.01 m ³ m ^{–3} (1%) with soil-specific calibration	0–1.0 m ³ m ^{–3}	Salinity errors <0.035 m ³ m ^{–3} from 0.05 to 0.4 m ³ m ^{–3}	50–500 mS m ^{–1}	±0.5 °C (0–+40 °C for temp sensor) ±0.75 °C (–20 to +60 °C for temp sensor)	–20 to +60 °C	Impedance	Handheld device		https://delta-t.co.uk/product/ml3/#overview
18	Stevens Water Monitoring Systems	HydraProbe	±0.01 WFV for most soils ±0.03 max for fine textured soils	0–100%	±2.0% or 0.02 S/m	0–1.5 S/m	±0.3 °C	–10 °C to +60 °C	Impedance	Cable	€695.00	https://stevenswater.com/products/hydraprobe/
19	METER	10HS	With standard calibration equation, 0.03 m ³ /m ³ (3% VWC) typical in mineral soils that have solution electrical conductivity <10 dS/m	0–57% (mineral soil calibration) 0–69% (soilless media calibration)	–	–	–	–	Impedance	Cable	€133	https://www.metergroup.com/en/meter-environment/products/ech20-10hs-soil-moisture-sensor
20	METER	EC5	±0.03 m ³ /m ³ typical in mineral soils that have solution EC <8 dS/m ±0.02 m ³ /m ³ in any porous medium (±2%)	0–100%	–	–	–	–	Impedance	Cable	€125	https://www.metergroup.com/en/meter-environment/products/ech20-ec-5-soil-moisture-sensor
21	Delta-T	SM150T	±0.03 m ³ m ^{–3} (3%)	0–1.0 m ³ m ^{–3}	–	–	±0.5 °C	0–40 °C	Impedance	Handheld device	€270.00	https://delta-t.co.uk/wp-content/uploads/2017/01/SM150T-user-manual-version-1.0.pdf
22	Delta-T	WET 2	±0.03 m ³ m ^{–3} (3%)	0–1.0 m ³ m ^{–3}	±10 mS m ^{–1}	0–300 mS m ^{–1}	±1.5 °C	–5 to 50 °C	Impedance measurement of capacitance	Handheld device	€1'940.00	https://pdf.agriexpo.online/pdf/delta-t-devices/soil-moisture-measurement/176237-13717.html#open64148

(Continued)

Table A1 (Continued)

No.	Company	Sensor designation	VWC Accuracy	Measuring range	EC Accuracy	Measuring range	T Accuracy	T Measuring range	Measuring method	Transmission	Cost	Spec sheet
23	Delta-T	Profile Probe PR2	$\pm 0.04 \text{ m}^3 \text{ m}^{-3}$ (4%) with soil specific calibration	$0\text{--}1.0 \text{ m}^3 \text{ m}^{-3}$	—	—	—	—	Impedance measurement of capacitance	Handheld device	€1'340.00	https://pdf.agriexpo.online/pdf/delta-t-devices/soil-moisture-measurement/176237-13717.html#open64148
24	Tomst	TMS-4	$\pm 1\%$	$0\text{--}100\%$	—	—	$\pm 0.5 \text{ }^\circ\text{C}$	$-40 \text{ to } +60 \text{ }^\circ\text{C}$	TDT	wireless	€88/sensor	https://www.datenlogger-store.de/mwdownloads/download/link/id/1956
25	Truebner	SMT100	Using factory calibration up to $\pm 3\%$ (VWC) in mineral soils with moderate salinity from 0 to 50% VWC Using medium specific calibration up to $\pm 1\%$ (VWC)	$0\text{--}60\%$ volumetric water content (up to 100% volumetric water content with limited accuracy)	—	—	$\pm 0.2 \text{ }^\circ\text{C}$, max. $\pm 0.4 \text{ }^\circ\text{C}$ over full range	$-40 \text{ to } +80 \text{ }^\circ\text{C}$ (analog version $-40 \text{ to } +60 \text{ }^\circ\text{C}$)	TDT	Cable	€149/sensor	https://www.truebner.de/en/smt100.php
26	Truebner	SMT50	Using factory calibration up to $\pm 3\%$ (VWC) in mineral soils with moderate salinity from 0 to 50% VWC Using medium specific calibration up to $\pm 1\%$ (VWC)	$0\text{--}60\%$ volumetric water content (up to 100% volumetric water content with limited accuracy)					Capacitance/FDR	digital/analog	57.98 €	https://dvs-beregnung.de/kapazitiver-bodenfeuchte-sensor-inkl-temperaturmessung-truebner-smt50-arduino?gclid=CjwKCAjwrNmWBhA4EiwAHbjEQLcdeyrL40mUehYDWwrJKQD51eBDFw8NZHZS0yi0e4j9qyeXk0ZLYBoCWY0QAvD_BwE
27	Baywa	SWM 5000	$0 \dots 70\%$				$0 \text{ }^\circ\text{C} \dots 60 \text{ }^\circ\text{C}$		FDR/NTC		886-€	
28	Conrad	M5 Stack									17-€	https://www.conrad.de/de/p/m5-stack-u101-bodenfeuchtesensor-1-st-2380033.html?hk=SEM&WT.mc_id=google_pla&gclid=CjwKCAjwrNmWBhA4EiwAHbjEQL_wwwmM018X_j8uQLAxidlLBA3j8zFITJSIDtbhAdu2CUmM_EPUXhoCMAQAQAvD_BwE

^aThis overview makes no claim to completeness. No liability for the correctness of the data. Source is the information of the providers accessible on the Internet.

^bPrices are approx. Estimates from websites. Costs for loggers or handheld readout devices must be added.

Table A2 Market overview of matric potential sensors.^a

No.	Company	Sensor designation	Claimed Accuracy	Claimed suction range [kPa]	Measuring method ^a	Approx. Costs per sensor ^b	Spec sheet
1	METER	TEROS 21 ^b	±10% +2 kPa	9 ... 200	SPS Capacitance	€ 270	https://www.metergroup.com/en/meter-environment/products/teros-21-soil-water-potential-sensor
2	Watermark/ Mosler	SS200/MMM	n.s.	0 ... 200	SPS: constant ion concentration	€ 57	
3	Irrrometer	Watermark	n.s.	0 ... 240	SPS Electrical resistance in granular matrix	\$ 50	https://www.irrometer.com/pdf/sensors/403%20WATERMARK%20Sensor-WEB.pdf
4	ecoTech	Tensiomark	±5% +3 kPa	1 ... 10 ⁶	SPS Heat capacitance	€ 830	https://ictinternational.com/content/uploads/2020/10/ecoTech-Manual-Tensiomark.pdf
5	UGT	Full Range Tensiometer FRT 15D ^b	±0.5% FS	−100 ... 1500	Osmotic Tensiometer	n.s.	https://ugt-online.de/loesungen/full-range-tensiometer-frt-15d/
6	Irrrometer	“SR” Irrrometer	±3-2-3% of span ASME B40.1 Grade B	0 ... 100	Tensiometer (mechanical manometer)	\$86 \$140–\$155 for transducer	https://www.irrometer.com/pdf/irrometers/101%20IRRROMETER-Family-brochure.pdf
7	Stelzner	Tensiometer Classic	±10% ±2 kPa	0 ... 70	Tensiometer (mechanical manometer)	€ 50	https://pronova.de/en/products/agricultural-measuring/moisture-analysis/1116/classic-plug-in-tensiometer
8	Fiedler Elektronika	TMS11A	±0.3 kPa	0 ... 65	Tensiometer	€ 250	https://www.fiedler.company/en/products/meteorological-stations-and-measuring-sensors/soil-moisture-sensors/tms11a-soil-tensiometer
9	METER	TEROS 32 ^b	±0.15 kPa	−85 ... 100	Tensiometer	€ 400	https://www.metergroup.com/en/meter-environment/products/teros-32-soil-tensiometer
10	UGT	Tensio 153e	±0.2 kPa	−30 ... 100	Tensiometer	€ 560	https://www.ugt-online.de/produkte/bodenkunde/tensiometer/tensio-153e/
11	UIT	Tensiometer	0.5% FS	−100 ... 100	Tensiometer	n.s.	https://www.uit-gmbh.de/de/produktetails__762/?getProdInfos=2312&

^aThis overview makes no claim to completeness. No liability for the correctness of the data. Source is the information of the providers accessible on the Internet. Stated ranges are in some cases not trustworthy.

^bPrices are approx. Estimates from websites. Costs for loggers or handheld readout devices must be added.

References

- Andraski BJ and Scanlon BR (2002) 3.2.3 Thermocouple Psychrometry. *Methods of Soil Analysis: Part 4 Physical Methods*, vol. 5, 609–642.
- Bittelli M (2010) Measuring soil water potential for water management in agriculture: A review. *Sustainability* 2(5): 1226–1251.
- Blöschl G and Sivapalan M (1995) Scale issues in hydrological modelling: A review. *Hydrological Processes* 9(3–4): 251–290.
- Bogena HR, Huisman JA, Güntner A, Hübner C, Kusche J, Jonard F, et al. (2015) Emerging methods for noninvasive sensing of soil moisture dynamics from field to catchment scale: A review. *Wiley Interdisciplinary Reviews Water* 2(6): 635–647.
- Buoyoucos GJ and Mick AH (1940) *An Electrical Resistance Method for the Continuous Measurement of Soil Moisture Under Field Conditions*. Technical Bulletin, vol. 172. East Lansing, MI: Michigan Agricultural Experiment Station.
- Campbell GS and Campbell CS (2005) Water content and potential, measurement. In: *Encyclopedia of Soils in the Environment*. Amsterdam: Elsevier.
- Durner W and Or D (2005) Chapter 73: Soil water potential measurement. In: Anderson MG and McDonnell JJ (eds.) *Encyclopedia of Hydrological Sciences*, pp. 1089–1102. John Wiley & Sons, Ltd. <https://doi.org/10.1002/0470848944.hsa077a>.
- Durner W, Diamantopoulos E, Iden SC, and Scharnagl B (2014) Chapter 17: Hydraulic properties and non-equilibrium water flow in soils. In: Teixeira WG, Ceddia MB, Ottoni MV, and Donnagema GK (eds.) *Application of Soil Physics in Environmental Analyses: Measuring, Modelling and Data Integration*. Progress in Soil Science Series, vol. 2014, pp. 403–434. Dordrecht: Springer. ISBN: 978-3-319-06012-5.
- Flint AL, Campbell GS, Ellett KM, and Calissendorff C (2002) Calibration and temperature correction of heat dissipation matric potential sensors. *Soil Science Society of America Journal* 66(5): 1439–1445.
- Gubiani PI, Reichert JM, Campbell C, Reinert DJ, and Gelain NS (2013) Assessing errors and accuracy in dew-point potentiometer and pressure plate extractor measurements. *Soil Science Society of America Journal* 77(1): 19–24.
- Jackisch C, Germer K, Graeff T, et al. (2020) Soil moisture and matric potential—An open field comparison of sensor systems. *Earth System Science Data* 12(1): 683–697.
- Kelleners TJ, Soppe RWO, Robinson DA, Schaap MG, Ayars JE, and Skaggs TH (2004) Calibration of capacitance probe sensors using electric circuit theory. *Soil Science Society of America Journal* 68(2): 430–439.
- Luo S, Lu N, Zhang C, and Likos W (2022) Soil water potential: A historical perspective and recent breakthroughs. *Vadose Zone Journal* 21(4), e20203.
- Matile L, Berger R, Wächter D, and Krebs R (2013) Characterization of a new heat dissipation matric potential sensor. *Sensors* 13(1): 1137–1145.
- Noborio K, Horton R, and Tan CS (1999) Time domain reflectometry probe for simultaneous measurement of soil matric potential and water content. *Soil Science Society of America Journal* 63: 1500–1505.
- Or D (2001) Who invented the tensiometer? *Soil Science Society of America Journal* 65(1): 1–3.
- Or D and Wraith JM (1999) A new TDR-based soil matric potential sensor. *Water Resources Research* 35: 3399–3407.
- Peck AJ and Rabbidge RM (1969) Design and performance of an osmotic tensiometer for measuring capillary potential. *Soil Science Society of America Journal* 33(2): 196–202.
- Phene CJ, Hoffman GJ, and Rawlins SL (1971) Measuring soil matric potential in situ by sensing heat dissipation within a porous body: I. Theory and sensor construction. *Soil Science Society of America Journal* 35(1): 27–33.
- Rahardjo H, Shen Y, Tsen-Tieng DL, et al. (2021) New osmotic tensiometer development. *Geotechnical Testing Journal* 44(3): 722–740.
- Reebs CF (1996) Evaluation of a line heat dissipation sensor for measuring soil matric potential. *Soil Science Society of America Journal* 60(4): 1022–1028.
- Richards LA (1949) Methods of measuring soil moisture tension. *Soil Science* 68(1): 95.
- Ridley AM and Burland JB (1994) A new instrument for the measurement of soil moisture suction. *Géotechnique* 44(3): 551–556.
- Roth K, Schulin R, Flüher H, and Attinger W (1990) Calibration of time domain reflectometry for water content measurement using a composite dielectric approach. *Water Resources Research* 26(10): 2267–2273.
- Scanlon BR, Andraski BJ, and Bilskie J (2002) Miscellaneous methods for measuring matric or water potential. In: Dane JH and Topp GC (eds.) *Methods of Soil Analysis, Part 4, Physical Methods*. SSSA Book Series No. 5, pp. 643–670. Madison, WI: Soil Science Society of America.
- Schelle H, Heise L, Jänicke K, and Durner W (2013) Water retention characteristics of soils over the whole moisture range: A comparison of laboratory methods. *European Journal of Soil Science* 64(6): 814–821.
- Schindler U, Durner W, von Unold G, and Müller L (2010) Evaporation method for measuring unsaturated hydraulic properties of soils: Extending the measurement range. *Soil Science Society of America Journal* 74(4): 1071–1083.
- Stevenson ME, Kumpan M, Feichtinger F, Scheidl A, Eder A, Durner W, et al. (2021) Innovative method for installing soil moisture probes in a large-scale undisturbed gravel lysimeter. *Vadose Zone Journal* 20(1), e20106.
- Temperley HNV and Chambers G (1946) The behaviour of water under hydrostatic tension. *Proceedings of the Physical Society of London* 53: 420–436.
- Topp GC and Davis JL (1985) Measurement of soil water content using time-domain reflectometry (TDR): A field evaluation. *Soil Science Society of America Journal* 49(1): 19–24.
- Van Der Ploeg MJ, Gooren HPA, Bakker, et al. (2009) Polymer tensiometers with ceramic cones: Direct observations of matric pressures in drying soils. *Hydrology and Earth System Sciences* 14(10): 1787–1799.
- Vereecken H, Huisman JA, Bogena H, et al. (2008) On the value of soil moisture measurements in vadose zone hydrology: A review. *Water Resources Research* 44(4).
- Whalley WR, Ober ES, and Jenkins M (2013) Measurement of the matric potential of soil water in the rhizosphere. *Journal of Experimental Botany* 64(13): 3951–3963.
- Young MH and Sisson JB (2002) 3.2.2 Tensiometry. In: Dane JH and Topp GC (eds.) *Methods of Soil Analysis, Part 4, Physical Methods*. SSSA Book Series No. 5pp. 575–608. Madison, WI: Soil Science Society of America.
- Zheng Q, Durben DJ, Wolf GH, and Angell CA (1991) Liquids at large negative pressures: Water at the homogeneous nucleation limit. *Science* 254(5033): 829–832.

Soil water retention and characteristic curve

Markus Tuller^a and Dani Or^b, ^aDepartment of Environmental Science, The University of Arizona, Tucson, AZ, United States; ^bDivision of Hydrologic Sciences, Desert Research Institute, Reno, NV, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of M. Tuller, D. Or, Water Retention and Characteristic Curve, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 278–289, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00376-3>.

Introduction	187
The matric potential	188
The bundle of cylindrical capillaries model	188
Liquid retention and pore shape	189
Representing the SWC	190
Parametric SWC models	190
SWC parameters	190
Global SWC datasets – from PTFs to CoGTFs	191
Physical constraints for SWC parameter estimation	192
Fractal representation of the soil pore space and the SWC	193
Physically based SWC models	193
Water adsorption and dry-end SWC relations	195
Pore scale models for realistic pore spaces	195
Pore network models and the SWC	197
Hysteresis in the SWC	197
Measurement of the SWC	198
Pressure plate apparatus and pressure flow cells	198
The evaporation method	199
Dewpoint hygrometer method	200
Paired sensors for field SWC measurement	200
References	201

Abstract

The relations between the amount of water retained in unsaturated soils and the retaining capillary and adsorptive forces are described by the soil water characteristic (SWC) curve. This function is soil specific and plays an important role in many agricultural, environmental, and engineering applications. A soil specific SWC is used for crop water management and determining plant available soil water, for representing water flow and contaminant transport, and for assessing natural disasters such as flooding or landslides. This chapter presents fundamental and novel concepts and techniques for representation and quantification of the SWC curve and estimation of its parameters at local and global scales. Standard measurement methods such as the Richards pressure plate apparatus and Tempe cells are presented along with contemporary techniques based on chilled-mirror dewpoint psychrometry, dielectric measurements and dynamic evaporation. We review different pore space representation models, the concept of hysteresis, and discuss adsorptive forces that dominate the dry end of SWC.

Key points

- Novel and established concepts for representation and modeling of the soil water characteristic curve are presented
- Standard soil water characteristic curve measurement methods are described along with contemporary state-of-the-art techniques
- Various pore space representation models, the concept of hysteresis, and the contribution of adsorptive forces to soil water retention are discussed

Introduction

The soil water characteristic (SWC) curve describes the amount of water retained in the soil pores (expressed as mass or volume water content, θ_m or θ_v) under equilibrium at a given matric potential. The SWC is an important hydraulic property related to the size and connectedness of pore spaces, hence, is strongly affected by soil texture and structure, and by other constituents, including organic matter and clay minerals. Modeling the water distribution and flow in partially saturated soils requires knowledge of the

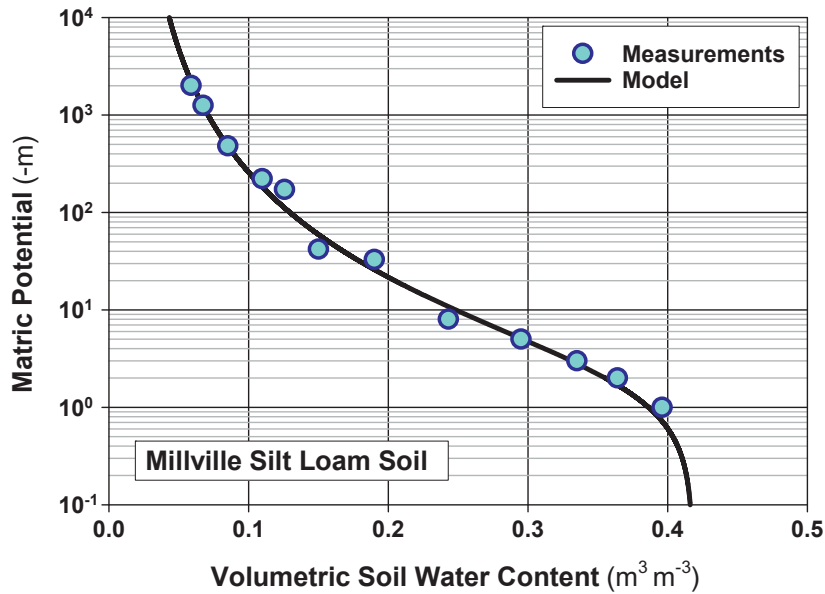


Fig. 1 Measured SWC data pairs (symbols represent a matric potential value and a corresponding water content) for Millville silt loam soil shown with a continuous SWC parametric model (to be discussed below).

SWC, therefore it plays a critical role in water management and in the simulation of solute and contaminant transport in the environment. Typically, a SWC is highly nonlinear and relatively difficult to measure accurately. Because the matric potential (ψ_m) extends over several orders of magnitude for the range of water contents commonly encountered in practical applications, the matric potential is often plotted on a logarithmic scale. Fig. 1 depicts SWC measurements together with a fitted continuous SWC model for Millville silt loam soil.

The matric potential

The matric potential represents the change in energy state of soil water relative to a reference, and it is the component of water potential attributed to the effects of capillary and adsorptive forces acting between the liquid, gas, and solid phases. Capillarity in unsaturated soils originates from the formation of curved air-water interfaces that, due to the resulting air-water surface tension and the contact angle, cause a lowering of the pressure in the liquid phase relative to the gas phase. Hence, capillarity in soil emerges under partially saturated conditions (i.e., in the presence of the nonwetting air phase), where based on simple radial geometry, a meniscus curvature radius (R) is a function of the pressure difference between air and water (termed capillary pressure, P_c) expressed by the Young–Laplace equation:

$$P_0 - P_c = \Delta P = \frac{2\sigma}{R} \quad (1)$$

where P_0 is the atmospheric pressure (conventionally referenced as zero), P_c is the pressure of the liquid side of the meniscus (in soil water), and σ is the surface tension of the air-water interface. If soil pores were to behave as a bundle of capillary tubes, the relationships between matric potential and soil pore radii would be described by these relations alone. However, in real soils, forces other than capillarity are at play and give rise to adsorption of water to soil surfaces to form thin liquid films. Generally speaking, the amount of water associated with water films scales with the soil specific surface area, hence, adsorptive forces are most important in clayey and organic soils with large specific surface areas. Additionally, water films are affected by electrostatic forces most prominent in soils with appreciable active clay content (e.g., montmorillonite). In sandy soils, adsorption is small (because the surface area is low) and capillarity dominates the SWC. Generally, the soil matric potential reflects the combined effects of capillarity and surface adsorption, which are often difficult to separate (see Water Potential and Capillarity chapters).

The bundle of cylindrical capillaries model

Conceptual models for the SWC and liquid distribution in partially saturated porous media have evolved based on the “bundle of cylindrical capillaries” (BCC) representation of pore-space geometry. The BCC representation postulates that at a given matric potential a portion of interconnected cylindrical pores is completely liquid-filled, whereas larger pores are completely empty (Fig. 2). This convenient idealization of soil pore space enables a linkage between the soil pore size distribution and the SWC based on the capillary-rise equation. (see Capillarity and Macropores and Macropore Flow chapters). However, such representation

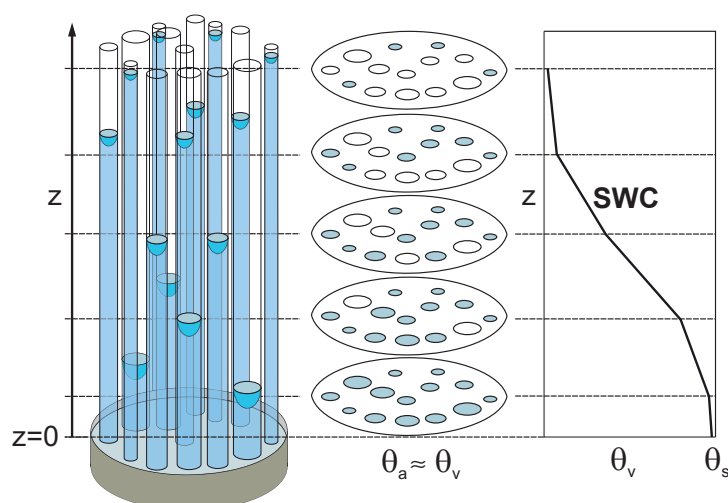


Fig. 2 Relationship between the pore space, as represented by the bundle of cylindrical capillaries model, and the SWC. Greater capillary rise against the pull of gravity occurs in smaller pores, which have smaller radii of meniscus curvatures. Note that z indicates the elevation above free water (representing the gravitational force), θ_a stands for pore volume, θ_v for volumetric water content, and θ_s for volumetric water content at saturation.

imposes serious limitations on the practical application of the BCC model to natural porous media as discussed in the following section, including the unrealistic underlying geometry that precludes dual water–air occupancy within the same pores and lack of consideration of adsorbed water films.

Liquid retention and pore shape

Liquid retention in the porous matrix of a soil is highly dependent on the shape and angularity of individual pores. The inspection of a thin section of Devonian sandstone (Scholle, 1979) (Fig. 3A) and a scanning electron micrograph of calcium-saturated montmorillonite (Fig. 3B) reveals that natural pore spaces do not resemble cylindrical capillaries, as assumed for the BCC model. When angular pores are drained, a fraction of the wetting phase remains in the pore corners (Fig. 3C). This aspect of “dual occupancy” of the invaded portion of the tube, not possible in cylindrical tubes, more realistically represents liquid configurations and mechanisms for maintaining hydraulic continuity in porous media. Liquid-filled corners and crevices play an important role in displacement rates of oil and in other transport processes in partially saturated porous media. The relationships between liquid retention and pore angularity as well as dynamic wetting phase displacement processes (i.e., pore snap-off and snap-on) are known and well documented (Tuller et al., 1999).

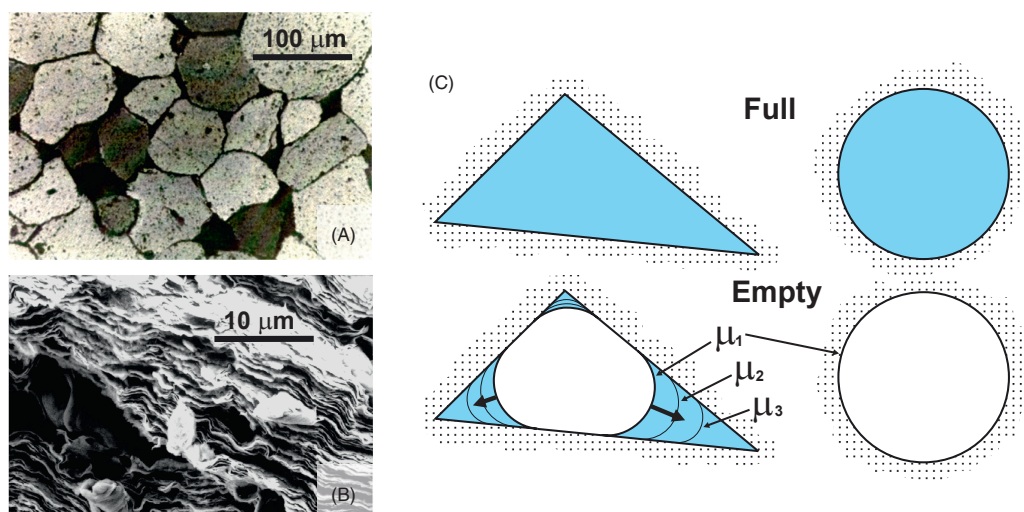


Fig. 3 Angular pore space in natural porous media: (A) thin section of Devonian sandstone; (B) scanning electron micrograph (SEM) of calcium-saturated montmorillonite; (C) liquid retention in triangular and cylindrical pores (μ is the chemical potential). (A) Reproduced from Scholle PA (1979) *A Color Illustrated Guide to Constituents, Textures, Cements, and Porosities of Sandstones and Associated Rocks*. American Association of Petroleum Geologists Memorandum 28. Tulsa, OK: American Association of Petroleum Geologists with permission of the AAPG, whose permission is required for further use.

Representing the SWC

Measuring the SWC is laborious and time-consuming. Measured ψ_m - θ pairs are often fragmentary and usually constitute relatively few measurements over the wetness range of interest. For modeling and analysis purposes, and for characterization and comparison between different soils and scenarios, it is therefore beneficial to represent the SWC in a mathematically continuous form. Several approaches, ranging from empirical parametric expressions to physically based models with parameters derived from measurable medium properties to pore network or lattice Boltzmann simulations can be employed to represent the continuous SWC.

Parametric SWC models

Various expressions have been proposed to represent the shape of SWC using continuous and differentiable functions. These are often empirical relations requiring a few parameters. The commonly used functions have been selected for their parsimony (as few parameters as possible), simple and consistent parameter estimation, and close fit to measured θ - ψ_m values including the wet and dry limits (wet and dry ends). Among the many models we present here are the widely used parametric SWC models proposed by Brooks and Corey (1964) and by van Genuchten (1980). The van Genuchten (VG) model relates the relative saturation (Θ) to the matric potential by the following expression:

$$\Theta = \frac{\theta - \theta_r}{\theta_s - \theta_r} = \left[\frac{1}{1 + (\alpha \psi_m)^n} \right]^m \quad (2)$$

where θ_r and θ_s are the residual and saturated water contents, respectively, ψ_m is the matric potential, and α , n , and m are parameters directly dependent on the shape of the $\theta(\psi_m)$ curve. Because of the strong correlation between n and m , a common simplification is to assume that $m = 1 - 1/n$. Thus, the parameters required for estimation of the model are θ_r , θ_s , α , and n . θ_s is relatively easy to measure or can be deduced from soil bulk density, leaving only the three unknown parameters θ_r , α , and n to be estimated from experimental data. Note that θ_r is sometimes taken as θ at -1.5 MPa, $\theta_{air\ dry}$, or a similar low matric potential value.

Another well-established parametric model is denoted as BC (Brooks and Corey, 1964):

$$\begin{aligned} \Theta &= \frac{\theta - \theta_r}{\theta_s - \theta_r} = \left(\frac{\psi_b}{\psi_m} \right)^\lambda & \psi_m > \psi_b \\ \Theta &= 1 & \psi_m \leq \psi_b \end{aligned} \quad (3)$$

where ψ_b is a parameter related to the soil matric potential at air entry (b represents “bubbling pressure”) and λ is related to the soil pore-size distribution. Matric potentials are expressed as positive quantities (i.e., in absolute values) in both the VG and BC parametric expressions. Campbell (1974) proposed a function similar to BC to express degree of saturation ($S = \theta/\theta_s$) in terms of the air entry matric potential ψ_b and an exponent b that can be related to soil texture:

$$\frac{\theta}{\theta_s} = \left(\frac{\psi_b}{\psi_m} \right)^{\frac{1}{b}} \quad (4)$$

ψ_b and b are derived as functions of the geometric mean diameter and geometric standard deviation of the particle-size distribution. Eq. (4) is often used in relation to the fractal idealization of the soil porous system as discussed below.

Estimation of VG or BC parameters from experimental data requires sufficient data points to characterize the shape of the SWC, and a program to perform nonlinear regression. Many computer spreadsheet software packages provide relatively simple and effective mechanisms to perform nonlinear regression. Fig. 4 depicts the parametric VG and BC models fitted to silt loam $\theta(\psi_m)$ data. The resulting best-fit parameters for the VG model are: $\alpha = 0.358 \text{ m}^{-1}$; $n = 1.42$; $\theta_s = 0.418 \text{ m}^3 \text{ m}^{-3}$; and $\theta_r = 0.04 \text{ m}^3 \text{ m}^{-3}$ (with $r^2 = 0.99$). For the BC model, the best-fit parameters are: $\lambda = 0.24$; $\psi_b = 0.79 \text{ m}$; $\theta_s = 0.418 \text{ m}^3 \text{ m}^{-3}$; and $\theta_r = 0.01 \text{ m}^3 \text{ m}^{-3}$ (with $r^2 = 0.98$). Note that the most striking difference between the VG and the BC models is the discontinuity at $\psi = \psi_b$ in the BC model.

The parametric models introduced thus far rely on curve-fitting using parameters related to the specific shape of the employed mathematical function (see Fig. 4) but are not directly derived from the physical characteristics of the soil or other porous media. Certain models have described the relationships between the SWC and the soil pore-size distribution using statistical functions, such as the log-normal distribution used by Kosugi (1994). Conceptually, they represent soil pores as a BCC, whose radii are log-normal distributed and simply apply the capillary-rise equation to establish a relationship between the matric potential and effective medium saturation (see Capillarity chapter). The applicability of these conceptual models is often limited to coarse-textured soils where capillary forces dominate the matric potential and adsorptive forces are negligible (see Water Potential chapter).

SWC parameters

The parameters derived from SWC measurements across different soil types for the VG and BC models are available from a variety of sources. The UNSaturated SOil hydraulic DATABASE (UNSODA) compiled by the US Salinity Laboratory (Nemes et al., 2001)

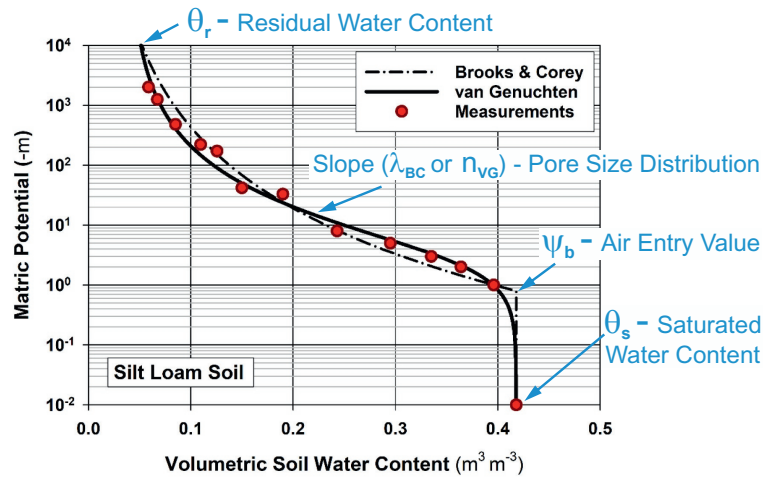


Fig. 4 The van Genuchten and Brooks and Corey parametric models fitted to measured silt loam SWC data. Some of the most prominent features of the SWC are associated with parameters of these two common models as directly highlighted on the figure.

Table 1 Typical van Genuchten model parameters (α , n) including residual (θ_r) and saturated (θ_s) water contents compiled from the UNSODA database.

Textural class	N	θ_r ($\text{cm}^3\text{cm}^{-3}$)	θ_s ($\text{cm}^3\text{cm}^{-3}$)	α (cm^{-1})	n
Sand	123	0.058	0.37	0.035	3.19
Loamy sand	51	0.074	0.39	0.035	2.39
Sandy loam	78	0.067	0.37	0.021	1.61
Loam	61	0.083	0.46	0.025	1.31
Silt	3	0.123	0.48	0.006	1.53
Silt loam	101	0.061	0.43	0.012	1.39
Sandy clay loam	37	0.086	0.40	0.033	1.49
Clay loam	23	0.129	0.47	0.030	1.37
Silty clay loam	20	0.098	0.55	0.027	1.41
Silty clay	12	0.163	0.47	0.023	1.39
Clay	25	0.102	0.51	0.021	1.20

N is the number of soils or samples of a given textural class from which the mean values are compiled.

contains an exhaustive collection of soil-water retention and unsaturated hydraulic conductivity information for soils of different textures from around the world. While the authors have attempted to provide some indices concerning quality or reliability of the compiled data, users are advised to use their own experience and discretion when adapting such data to their own applications. Table 1 lists values for the VG parameters α and n , and the residual and saturated water contents for various soil textural classes compiled from UNSODA. Note the large variations in SWC relationships for given soil textural classes. Gutmann and Small (2007) have shown that these variations may exceed differences across soil textures. The origins of these large variations are attributed to lack of consideration of soil structure that may alter SWC parameters for the same soil texture. See Or (2019) for a recent discussion.

Global SWC datasets – from PTFs to CoGTFs

For spatially extensive modeling of hydrologic processes, agroecosystems and land-surface modules of global climate models, soil hydraulic parameterization is an essential input. The present parameterization of soil hydraulic properties is based on detailed maps of soil type and supported by other remotely sensed attributes, using the formalism of pedotransfer functions (PTFs), which primarily use soil information, or covariate-supported geotransfer functions (CoGTFs) that harness numerous other geophysical covariates (vegetation, climate, geomorphic features). The term PTF was coined by Bouma (1989) with the aim of “translating data we have to parameters we need.” Practically, we seek to use a few measured SWC datasets to derive functional relationships with easy-to-measure attributes that are spatially available. These correlations are then used to map the inferred parameters at the desired spatial resolution. The history of regressing easy-to-measure soil attributes for inference of SWC-related parameters dates back the work of Briggs and Lane (1907) and later Veihmeyer and Hendrickson (1927) in the context of deriving plant available water from soil attributes.

We distinguish SWC datasets such as UNSODA (Nemes et al., 2001), HYBRAS (Ottoni et al., 2018), WOSIS (Batjes et al., 2017), AFSPDB (Leenaars et al., 2014) from global maps of SWC parameters. The latter involves the PTF mapping mentioned earlier. Modern SWC parameter maps such as generated with Rosetta 3 (Zhang and Schaap, 2017), HiHydroSoil (de Boer, 2016), CoGTF (Gupta et al., 2021), and the Dai et al. (2019) dataset, provide essential inputs for land surface models. However, these maps suffer from a chronic lack of SWC data, are limited by regional calibration, provide uneven geographical coverage, and yield poorly constrained parameter values. Gutmann and Small (2007) have shown that soil textural classes explain only 5% of the variance of soil hydraulic properties (SHP), suggesting that estimated SHP variations exceed expected differences due to soil texture classification.

A recently compiled Global Soil Hydraulic Properties (GSHP) database (Gupta et al., 2021) encompasses nearly 16,000 well-curated SWC sets spanning 2750 locations globally (as compared with the widely used UNSODA database that contains 218 SWCs primarily from the US and Europe). The CoGTF approach broadens aspects of classical PTFs by using remotely sensed covariates, machine learning, and spatial cross-validation tools to improve training and constraining estimated soil hydraulic parameter values. Table 2 lists mean values for the VG parameters α and n , and the residual and saturated water contents together with their standard deviations for various soil textural classes compiled from the GSHP database.

Physical constraints for SWC parameter estimation

Parameter estimation for the highly nonlinear SWC may result in numerous combinations of parameter sets with similarly good fits to measured data. This nonuniqueness (related to a phenomenon known as sloppy parameter space) may result in combinations of SWC parameters that are not physically consistent. These are identified via SWC-related applications that use the same parameters for example to define the soil unsaturated hydraulic conductivity function, characteristic capillary lengths for different soil types (Lehmann et al., 2008), infiltration behavior at steady state and other physical phenomena that rely on SWC information and its parameters. A potential avenue for identifying unphysical parameter combinations is to use these auxiliary physical phenomena as constraints for SWC parameter estimation (striking a balance between goodness of fit to SWC measurements and ensuring physical consistency) as demonstrated in a study by Lehmann et al. (2020) using a SWC-derived soil trait known as characteristic evaporation length L_c . This is the depth over which capillary flow is maintained and supplies soil water to the soil surface during stage 1 evaporation (when this length is exceeded by a receding drying front, the evaporation plane migrates below the surface giving rise to vapor-transport dominated stage 2 evaporation). The scarce direct measurements of L_c for different soil types (texture classes) have been compiled and are represented (color bars) with the theoretical L_c (black lines) in Fig. 5A from Lehmann et al. (2020). Note that mean theoretical L_c values vary with the underlying parameters deduced from SWC databases (Rosetta 3 or Carsel and Parrish, 1988). As illustrated in Fig. 5B the constrained VG parameter values (red) honor the L_c constraint with minor deviations in terms of goodness of fit.

Tests of the constrained SWC parameters show improved hydrologic performance at small and larger scales. While it is extremely difficult to assess the usefulness of such constrained parameter estimation for the performance of complex models, the approach adds robustness to the estimated parameters at relatively low cost in terms of SWC fit. The concept of injecting physical constraints to SWC parameter estimation can be extended to include other physical constraints for improving PTF estimated SWCs, for example, by consideration of effects on infiltration from vegetation cover and soil structure.

Table 2 Globally-averaged van Genuchten SWC model parameters compiled from the *Global Soil Hydraulic Properties* (GSHP) database.

Textural class	N	θ_r ($m^3 m^{-3}$)		θ_s ($m^3 m^{-3}$)		α (m^{-1})		n (–)	
		$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
Sand	443	0.05	0.03	0.36	0.09	3.83	4.09	2.59	1.06
Loamy sand	397	0.04	0.03	0.39	0.1	8.11	9.83	1.74	0.57
Sandy loam	997	0.05	0.05	0.41	0.1	6.04	8.33	1.56	0.37
Loam	622	0.08	0.08	0.46	0.11	7.31	8.59	1.39	0.31
Silt	23	0.03	0.02	0.5	0.07	2.02	1.83	1.49	0.38
Silt loam	620	0.05	0.06	0.48	0.1	4.46	6.67	1.41	0.32
Sandy clay loam	399	0.09	0.07	0.43	0.08	11.86	11.29	1.41	0.31
Clay loam	286	0.09	0.1	0.52	0.13	11.53	11.31	1.31	0.28
Silty clay loam	338	0.09	0.09	0.53	0.1	9.07	10.78	1.37	0.35
Silty clay	213	0.13	0.11	0.54	0.09	15.85	13.14	1.35	0.38
Sandy clay	113	0.13	0.09	0.46	0.09	16.07	11.87	1.4	0.33
Clay	901	0.18	0.12	0.56	0.09	13.78	12.06	1.4	0.36

N is the number of SWCs of a given soil textural class from which parameter means ($\bar{x}$) and standard deviations (*s*) were computed.

Reproduced from Gupta S, Papritz A, Lehmann P, Hengl T, Bonetti S, and Or D (2022) Global Soil Hydraulic Properties dataset based on legacy site observations and robust parameterization. *Scientific Data* 9:444, doi:10.1038/s41597-022-01481-5.

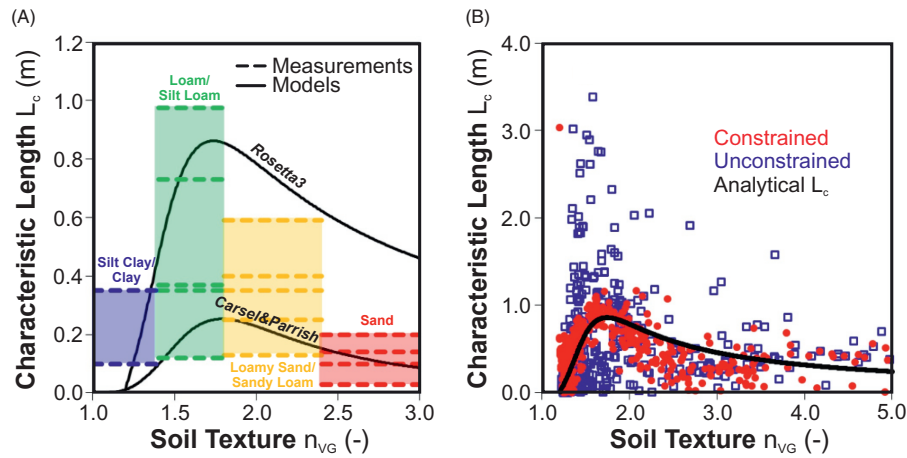


Fig. 5 (A) Measurement ranges and theoretical relations between L_c and soil texture (represented via typical values of the VG parameter n_{VG}) for different databases (Rosetta 3 and CARSel and Parrish, 1988); (B) Comparison of constrained (red) and original unconstrained (blue) van Genuchten n_{VG} parameters. Constrained parameters were obtained by requiring simultaneous fit to the SWC and L_c data (the latter represented as black line by an analytical expression). Reproduced from Lehmann P, Bickel S, Wei Z, and Or D (2020) Physical constraints for improved soil hydraulic parameter estimation by pedotransfer functions. *Water Resources Research* 56:e2019WR025963 doi:10.1029/2019WR025963.

Fractal representation of the soil pore space and the SWC

Fractals are hierarchical, often highly complex, spatial, or temporal systems that are generated with iterative algorithms obeying simple scaling rules (Mandelbrot, 2010). Patterns within such systems repeat themselves over a defined range of scales (self-similarity). This enables the reproduction of statistical properties of a particular pattern at other length or time scales. Fractal geometry can be applied to quantitatively describe irregularity and shape of natural objects by estimating their fractal dimension. Several theoretical models have been proposed to derive the SWC from the fractal representations of the soil porous system (Gimenez et al., 1997). There are two general approaches, based on either surface or mass fractals. Surface fractal models assume that water is only present in the form of adsorbed liquid films on pore surfaces, whereas mass fractal models assume that only capillary water obeying the capillary-rise equation is present within the fractal system. As with the BCC approach, fractal models for the SWC are based on derivation of the pore-size distribution from the fractal structure under consideration, neglecting pore connectivity and topology. Crawford (1994) presented the following relationship between the mass-fractal dimension (D_m) and the degree of saturation (S):

$$s = \left(\frac{\psi_m}{\psi_b} \right)^{(D_m - d_e)} \quad (5)$$

where ψ_m is the matric potential under consideration, ψ_b is the matric potential at the air entry point, and d_e is the embedding dimension. The embedding dimension equals 2 in two-dimensional systems and 3 in three-dimensional space. Note that the mass-fractal dimension D_m is always less than d_e . Because of the identical functional form, Eq. (5) is commonly equated with Campbell's version of the BC SWC function (Eq. 4) to derive Campbell's b -factor from the fractal dimension ($b = -1/(D_m - d_e)$). Fractal approaches are limited due to the assumption that fractal scaling is valid over an infinite range of matric potentials. In reality natural porous media have lower and upper scaling limits related to their minimum and maximum pore sizes.

Physically based SWC models

In contrast to parametric models based on BCC and fractal approaches of soil pores, recent physically based models have adapted a more nuanced representation of soil pores considering angular pores and specific surface area (absent in BCC representation) to simultaneously accommodate capillary and adsorptive phenomena. Despite the unavoidable complexity, the angular pore space model offers several advantages over the BCC approach, including a more realistic representation of natural pores with dual occupancy of wetting and nonwetting phases in the same pores, and accommodation of adsorptive forces for soils with high specific surface areas, which improves SWC representation also under dry conditions.

Tuller et al. (1999) applied the augmented Young-Laplace equation that considers capillary and adsorptive contributions to the matric potential to calculate water-air interfaces within a cross-section of angular pores. This building block and abrupt water displacement mechanisms (snap-off or snap-on) enabled assembly of SWC functions for imbibition and drainage building from pore scale representation. Fig. 6 depicts a typical transition of a unit pore element from dry to wet (imbibition) that includes spontaneous liquid displacement (snap-off) in slits (Fig. 6B) and in the central pore (Fig. 6D). A statistical upscaling scheme using a gamma density function for the central pore length L (Or and Tuller, 1999) enabled transition to the SWC of soil samples.

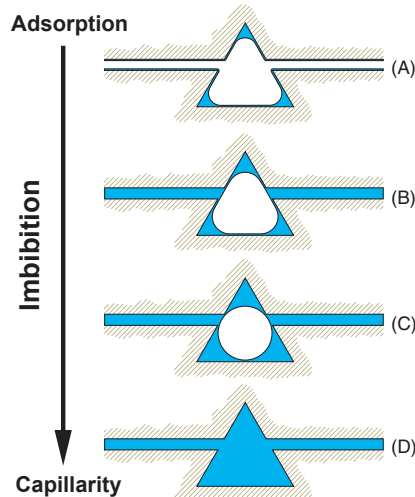


Fig. 6 Unit element filling stages during imbibition. (A) Liquid films adsorbed on pore and slit walls and liquid held in corners due to capillary forces at low matric potentials; (B) spontaneous slit fill-up (capillary condensation); (C) pore snap-off; (D) full unit element.

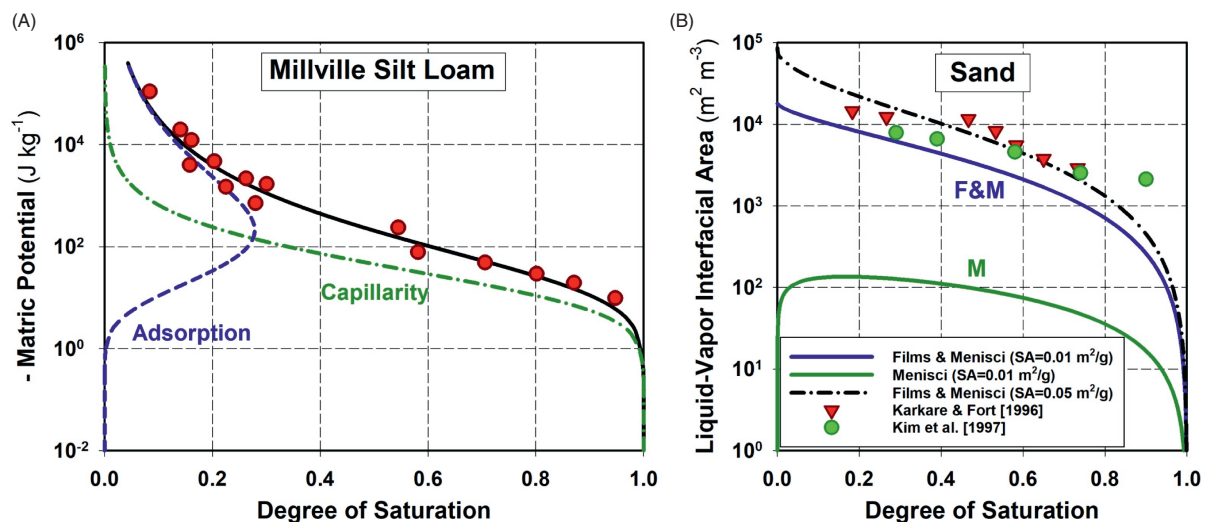


Fig. 7 (A) Application of the angular pore-space model for Millville silt loam soil. Note the capillary and absorptive contributions to the soil-water characteristic (SWC); (B) comparison of model-calculated capillary (menisci) and adsorptive (films) contributions to liquid-vapor interfacial area as a function of saturation in an artificial sand mixture and interfacial area measurements obtained with interfacial tracers (Karkare and Fort, 1996; Kim et al., 1997). SA, surface area; M, menisci; F&M, films and menisci. Reproduced from Or D and Tuller M (1999) Liquid retention and interfacial area in variably saturated porous media: Upscaling from single-pore to sample-scale model. *Water Resources Research* 35(12): 3591–3606.

The gamma distribution represents observed soil pore-size distributions and enables analytical expressions for the sample scale SWC. The upscaling procedure represents the degree of saturation (S) as a function of matric potential expressed as the sum of various pore-filling mechanisms as depicted in Fig. 6. Fig. 7A depicts a sample application for Millville silt loam soil. Because the model calculates the configuration of liquid-vapor interfaces within pores and films, adsorptive and capillary contributions to the SWC can be separated (Fig. 7A). Such separation is not possible with any of the empirical or semiempirical approaches presented above. Another important feature of this model is the ability to estimate liquid-vapor interfacial area as shown in Fig. 7B, comparing model estimates for sand with known surface area of $0.01\text{--}0.05\text{ m}^2\text{g}^{-1}$. The magnitude and changes in liquid vapor interfacial area play an important role in multiphase flow processes, bioremediation of contaminated soils, gas-exchange phenomena, and virus and colloid transport.

The complexity of the physical SWC model of Tuller and Or (TO) impeded its practical application. Several simplification have been proposed such as in the work of Lebeau and Konrad (2010), who represented soil pore spaces by combining parallel capillary tubes (similar to the BCC representation) with flat and smooth particle surfaces to facilitate partitioning the SWC into capillary and adsorptive components. Another simplified representation of the TO angular pore model was proposed by Diamantopoulos and Durner (2015). They considered triangular pores for modeling water retention and the effect of the contact angle on water flow to derive closed form expressions for sample-scale liquid retention, hydraulic conductivity, and liquid-vapor interfacial area as a function of the matric potential.

Water adsorption and dry-end SWC relations

Various approaches have been proposed to estimate the dry end of the SWC (e.g., Rossi and Nimmo, 1994; Jensen et al., 2015). These have been based on arbitrary assignment of limiting matric potential values and extrapolation of the water content at these low matric potentials. Here we show that a benefit of the physical representation of the SWC based on the TO model is the simple attribution of adsorptive and capillary components to the matric potential as seen in Fig. 7A. The adsorptive surface forces dominate at the dry end of a SWC (Fig. 7A) where the remaining soil water is held (primarily) in the form of thin liquid films, which gives rise to strong relationship between the dry end soil water content (often referred to as residual water content, θ_r) and soil specific surface area (SA). The dominance of adsorption over capillarity at the SWC dry end greatly simplifies the representation of SWC as demonstrated by Tuller and Or (2005). Considering soil surface-water interactions dominated by van der Waals adsorptive forces, gives rise to a nearly universal scaling relationship for SWC curves at low water contents (Fig. 8A). The relationship can be applied to estimate θ_r [$\text{m}^3 \text{m}^{-3}$] for matric potentials $\psi < -10 \text{ MPa}$ as:

$$\theta_r \cong h \times SA \times \rho_b \quad (6)$$

where SA [$\text{m}^2 \text{kg}^{-1}$] is the soil specific surface area, ρ_b [kg m^{-3}] is the dry soil bulk density, and h [m] is the film thickness as a function of matric potential:

$$h = \sqrt[3]{\frac{A_{svl}}{6\pi\rho_w g \psi}} \quad (7)$$

where A_{svl} [J] is the Hamaker constant for solid-vapor interactions through the intervening liquid, ψ [m] is the matric potential, ρ_w [kg m^{-3}] is the density of the liquid, and g [m s^{-2}] is acceleration due to gravity.

Conversely, the same strong relationship can be applied to deduce soil specific area as illustrated in Tuller and Or (2005) and in several follow up studies (Arthur et al., 2013; Leão and Tuller, 2014). Eq. (8) that relates gravimetric water content θ_m [kg kg^{-1}] to water film thickness and soil specific surface area may be fitted by nonlinear regression to adsorption isotherms measured with the dewpoint hygrometer method (see measurement section below) with SA as a free fitting parameter (Fig. 8B).

$$\theta_m = h \times SA \times \rho_w \quad (8)$$

Table 3 compares the SA obtained by fitting Eq. (8) to water adsorption isotherms measured with the dewpoint hygrometer method and SA measured with the Ethylene Glycol Monoethyl Ether (EGME) adsorption method for a wide range of soil textures and clay mineralogies (Tuller and Or, 2005).

While the SWC dry end is also affected by water held in nano-roughness by capillary forces (that support nanoscale interfacial curvatures), evidence suggests that the contribution to the dry end water content is relatively small and can be combined with surface-area dominated adsorption for many practical applications. However, this does not hold for coarse textured and low surface area soils, where roughness is important.

Pore scale models for realistic pore spaces

With the rapid advancement of computational resources (i.e., high performance computing), several sophisticated simulation techniques including lattice Boltzmann models, smoothed particle hydrodynamics, or volume of fluid methods gained popularity

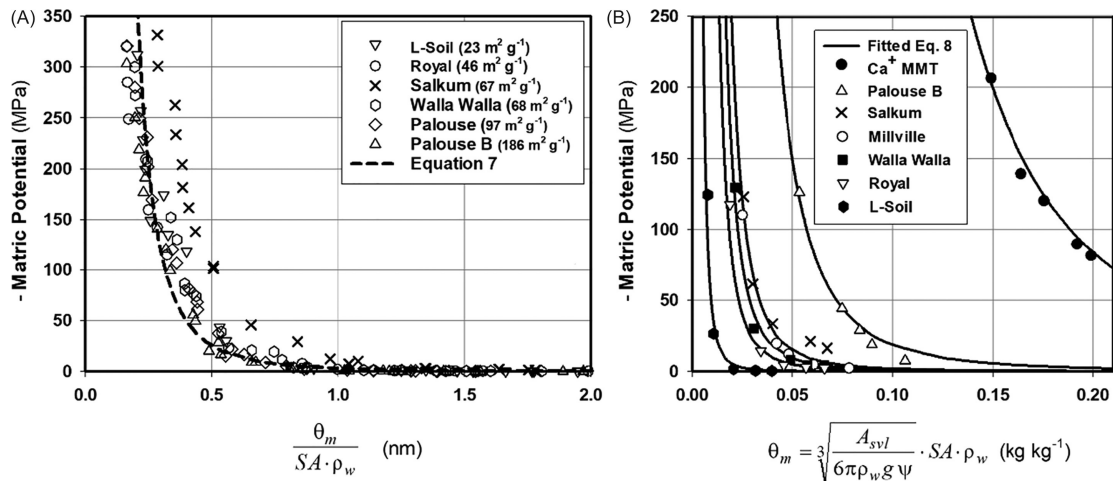


Fig. 8 (A) SWC curves scaled by soil specific surface area (data are from Campbell and Shiozawa, 1992). The dashed line depicts the adsorption isotherm given in Eq. (7). (B) Eq. (8) fitted to measured data obtained with the WP4C laboratory hygrometer (METER Group, Inc., Pullman, WA, USA). Reproduced from Or D and Tuller M (2005) Water films and scaling of soil characteristic curves at low water contents. *Water Resources Research* 41: W09403. doi:10.1029/2005WR004142.

Table 3 Comparison of SA obtained by fitting Eq. (8) to water adsorption isotherms and SA measured with the EGME adsorption method.

Soil series	Eq. (8) (m^2g^{-1})	EGME (m^2g^{-1})
L-Soil	24	25
Royal	58	45
Walla Walla	71	70
Millville	72	73
Salkum	84	51
Palouse B	181	203
Ca^{2+} Montmorillonite	597	760

Reproduced from Tuller M and Or D (2005) Water films and scaling of soil characteristic curves at low water contents. *Water Resources Research* 41: W09403, doi:10.1029/2005WR004142.

in recent years. Accurate three-dimensional discretization of the soil porous system obtained with X-ray computed micro tomography (μCT), magnetic resonance imaging (MRI), or focused ion beams scanning electron microscopy (FIB-SEM) are used as boundaries for fluid flow simulations (Blunt et al., 2013).

Lattice Boltzmann models (LBM) are descendants of lattice gas cellular automata, which follow the motions of individual particles and were first presented as a viable means of solving Navier–Stokes equations. In the most simplistic sense, LBMs work with distributions of particles at each lattice point rather than with individual particles. LBMs simulate interactions of hypothetical particles confined to a regular lattice, which greatly enhances solvability of the Boltzmann equation. This has advantages for certain types of simulations in that averaging is not required to obtain smooth velocity fields. The lattice Boltzmann method has emerged as a powerful tool for simulation of multiphase fluid systems, including water and water vapor. The LBM incorporates complex details of pore shape that characterize realistic porous media, fluid factors, and solid–fluid interactions in a physically sound way, leading to realistic simulation of interfaces between different fluids or between a liquid and its vapor and adsorption of wetting films. Interfaces arise, deform, and migrate in virtually any pore geometry rather naturally without the need for complex interface tracking algorithms. An example of LBM application to compute water retention in complex porous media is shown in Fig. 9, which is based on two-dimensional (2-D) images of a real soil (Sukop and Or, 2003). Vapor and liquid boundaries of equal pressure were applied to the top and bottom of the domain, respectively, in steps corresponding to equal increments of $\log(-\text{matric potential})$. The matric potential is expressed in mass units per time step squared (equivalent to energy per unit volume in a 3-D system). Each potential step was terminated when the relative change in fluid mass in the domain was negligibly small. The simulated liquid behavior and SWC curve show appreciable differences in fluid configurations during wetting and drying that resulted from hysteresis.

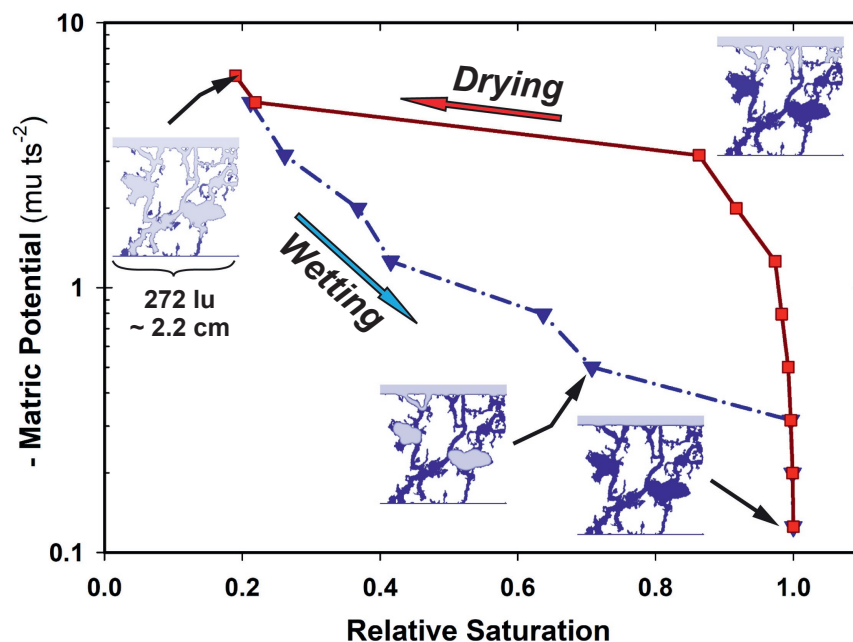


Fig. 9 Lattice Boltzmann model simulation of liquid distribution in a complex pore system and computation of the soil-water characteristic curve. Reproduced from Sukop M and Or D (2003) Lattice Boltzmann method for modeling liquid–vapor interface configurations in porous media. *Water Resources Research* 40(1): W01509.

Smoothed particle hydrodynamics (SPH) was first applied in astrophysics to simulate gas flow dynamics in the course of the fission of protostars. A continuous density field of fluid particles is characterized by a superposition of bell-shaped smoothing functions centered on a set of points. The gradient of the field is represented by the same superposition of the gradients of the smoothing function. A set of properties, such as the particle mass is associated with each particle. The mass can be thought of as being smoothed or smeared by the smoothing function so that the contribution of an individual particle to the fluid density field can be specified and the density field defined (Meakin and Tartakovsky, 2009). The SPH equation for the flow of a perfect fluid is based on the equation of motion $d\bar{v}/dt = -\nabla P/\rho$, where $\bar{v}$ is the local fluid velocity, ∇P the pressure gradient, and ρ the fluid density. The effects of viscosity on fluid flow can be included in SPH simulations by considering the viscous dissipation term in the Navier-Stokes equation. Major advantages of SPH include conservation of mass without the need for further computation (i.e., the particles already represent mass), subtraction of pressure from weighted contributions of neighboring particles rather than by solving linear systems of equations. Furthermore, in contrast to grid-based techniques (e.g., volume of fluid method), which track fluid boundaries, SPH directly creates a free surface for two-phase interacting fluids (i.e., particles represent the denser fluid (water) and empty space represents the lighter fluid (air)). These provisions allow SPH fluid flow simulations in real time.

The volume of fluid method (VOF) provides a powerful, yet efficient means for tracking fluid interfaces (Golparvar et al., 2018). It accurately defines the location and shape of the free surface and simulates its movement during each time step via solution of the Navier-Stokes equation. The VOF method is based on an indicator function (or a set of $n - 1$ functions in the case of n distinct fluid phases), which is advected with the flow. The indicator function is the set of volume fractions within each element of the computational grid (Huang et al., 2005). For air-water interfaces (i.e., in partially saturated soils), the volume fractions for a single phase are sufficient. The effects of surface tension are expressed via a body force that acts on the fluid in cells near (curved) interfaces in which the gradient of the indicator function is non-zero. For multiphase fluid flow in a porous medium, a fluid-fluid-solid contact angle model must be considered in VOF simulations. In summary, while the techniques above are well suited for insights into pore-scale processes, their applicability to sample or profile scales processes is limited.

Pore network models and the SWC

In contrast to the LBM, SPH, and VOF methods, network models consider simple and regular pore geometries (Berkowitz and Ewing, 1998). They were developed based on the idea that pore space may be represented as an interconnected network of capillary tubes whose radii represent the dimensions of the pores within a porous medium. For a given matric potential, liquid-vapor interfacial configurations within the network can be exactly determined, based on pore-scale capillary and dynamic displacement considerations. The macroscopic SWC is then determined based on geometric volume averaging of the spatially distributed liquid within the network. A primary advantage of pore network models is the explicit consideration of pore connectivity and topology in a simplified and mathematically tractable framework. Some of the limitations involve oversimplification of pore-scale physics (e.g., neglect of adsorptive pore-scale processes), incomplete understanding of interface migration and routing, inadequate technologies for inference of network parameters from real samples, and significant computational burden for detailed 3-D networks even at a core scale (greater than 100 mm).

Hysteresis in the SWC

Water content and the potential energy of soil water are not uniquely related, because the amount of water present at a given matric potential is dependent on the pore-size distribution and the properties of air–water–solid interfaces. An SWC relationship may be obtained by: (1) taking an initially saturated sample and applying suction or pressure to desaturate it (desorption), or (2) gradually wetting an initially dry soil (sorption). These two pathways produce curves that in most cases are not identical; the water content of the “drying” curve is higher for a given matric potential than that of the “wetting” branch (Fig. 10A). This phenomenon is called hysteresis, defined as “the phenomenon exhibited by a system in which the reaction of the system to changes is dependent upon its past reactions to change.” The hysteresis in the SWC can be related to several phenomena, including: (1) the “ink bottle” effect, resulting from nonuniformity in shape and sizes of interconnected pores; drainage is governed by the smaller pore radius r , whereas wetting is dependent on the large radius R (Fig. 10B); (2) different liquid–solid contact angles for advancing and receding water menisci (Fig. 10C); (3) entrapped air in a newly wetted soil (e.g., pore doublet); and (4) swelling and shrinking of the soil during wetting and drying. The role of individual factors remains unclear and is subject to ongoing research. Part of the hysteresis phenomena may be attributed to measurement artifacts, for example, due to differences between tension- and pressure-induced desaturation. A potentially important aspect of desorption methods under tension is the possibility of liquid displacement (drainage) even in the absence of a continuous gaseous phase due to cavitation initiated by encapsulated gas bubbles or the liquid’s own vapor pressure. Surface heterogeneity and impurities in soil and rock water are conducive to lowering the cavitation threshold. Also see chapter on Hysteresis.

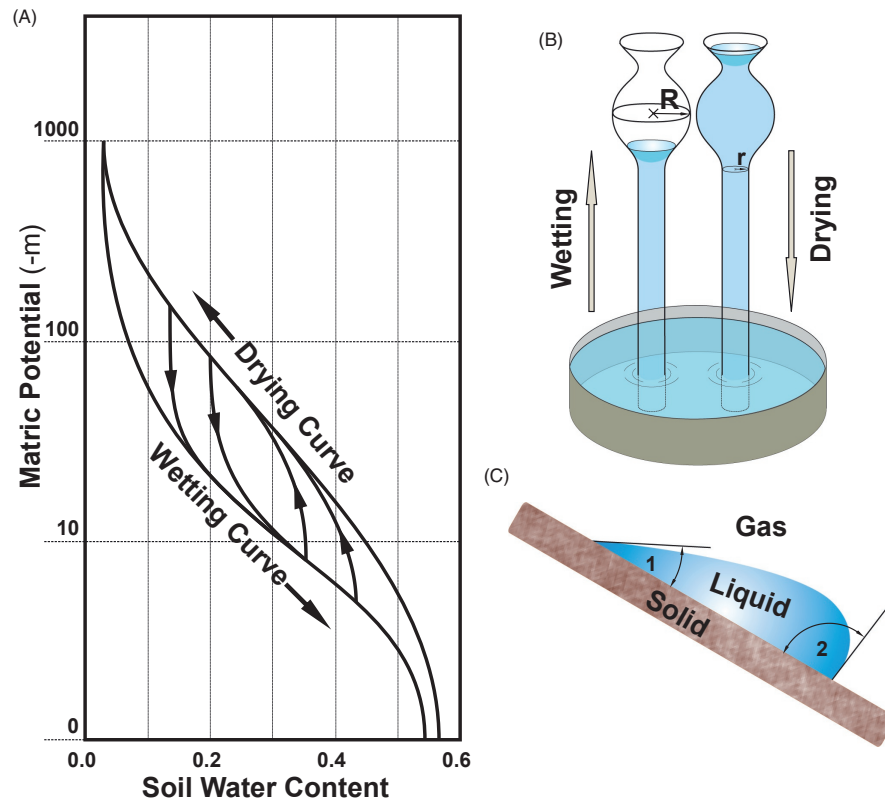


Fig. 10 (A) Hysteresis of the soil water characteristic. (B) The ink bottle effect and (C) the contact angle effect as potential mechanisms for hysteresis.

Measurement of the SWC

A variety of methods may be used to obtain requisite ψ_m and θ values to estimate the SWC. Potential experimental problems include: the limited functional range of the tensiometer, which is often used for in situ measurements; inaccurate θ measurements in some cases; the difficulty in obtaining undisturbed samples for laboratory determinations; and a slow rate of equilibrium under low matric potential (i.e., dry soils). In situ methods are preferred in defining SWCs, as is measuring over a wide range of ψ_m and θ values. An effective method to obtain simultaneous measurements of ψ_m and θ_v is by installing time-domain reflectometry (TDR) probes in close proximity to transducer tensiometers or more advanced ψ_m sensors (e.g., TEROS 21 sensor, Meter Group, Inc.; Tensiometer sensor, ecoTech Umwelt Meßsysteme) with the attributes monitored during variable soil wetness. Large changes in ψ_m and θ_v are expected under highly evaporative conditions near the soil surface, or in the presence of active plant roots.

Pressure plate apparatus and pressure flow cells

The pressure plate apparatus comprises a pressure chamber that encloses a water-saturated porous plate, which allows water but not air to flow through its pores (Fig. 11A). The porous plate is at atmospheric pressure at the bottom, while the top surface is at the applied pressure of the chamber. Soil samples, usually sieved to less than 2 mm, are placed in retaining rings in contact with the porous plate and allowed to saturate by immersion in water. The porous plate with saturated soil samples is then placed in the chamber and a known N_2 or air gas pressure is applied to force water out of the soil through the plate. Water flows out of the soil until equilibrium between the force exerted by the air pressure and the force by which soil water is being held by the soil (ψ_m) is attained. Soil water retention at the wet end (less than 1 bar) is strongly influenced by the soil structure and its natural pore-size distribution. Hence, “undisturbed” intact soil samples (cores) are preferred over repacked samples for this portion of the SWC. The pressure flow cell (also known as Tempe cell) can hold intact soil samples encased in metal rings (Fig. 11B). The operation of the Tempe cell follows that of the pressure plate, except its pressure range is usually limited to 0–1 bar or 0.1 MPa. The porous ceramic material used in pressure plates and flow cells must be completely water-saturated prior to use. Following equilibrium between soil matric potential ψ_m and the applied air pressure, the soil samples are removed from the apparatus, weighed wet, then oven-dried to determine the mass water content gravimetrically. These may be converted to volume water contents through knowledge of the sample bulk densities. The water content of repacked soils at a given matric potential should not be used to infer θ of intact soils at the same ψ_m , due to modified pore sizes and pore geometry. Multiple pressure steps may be applied to the same soil sample when

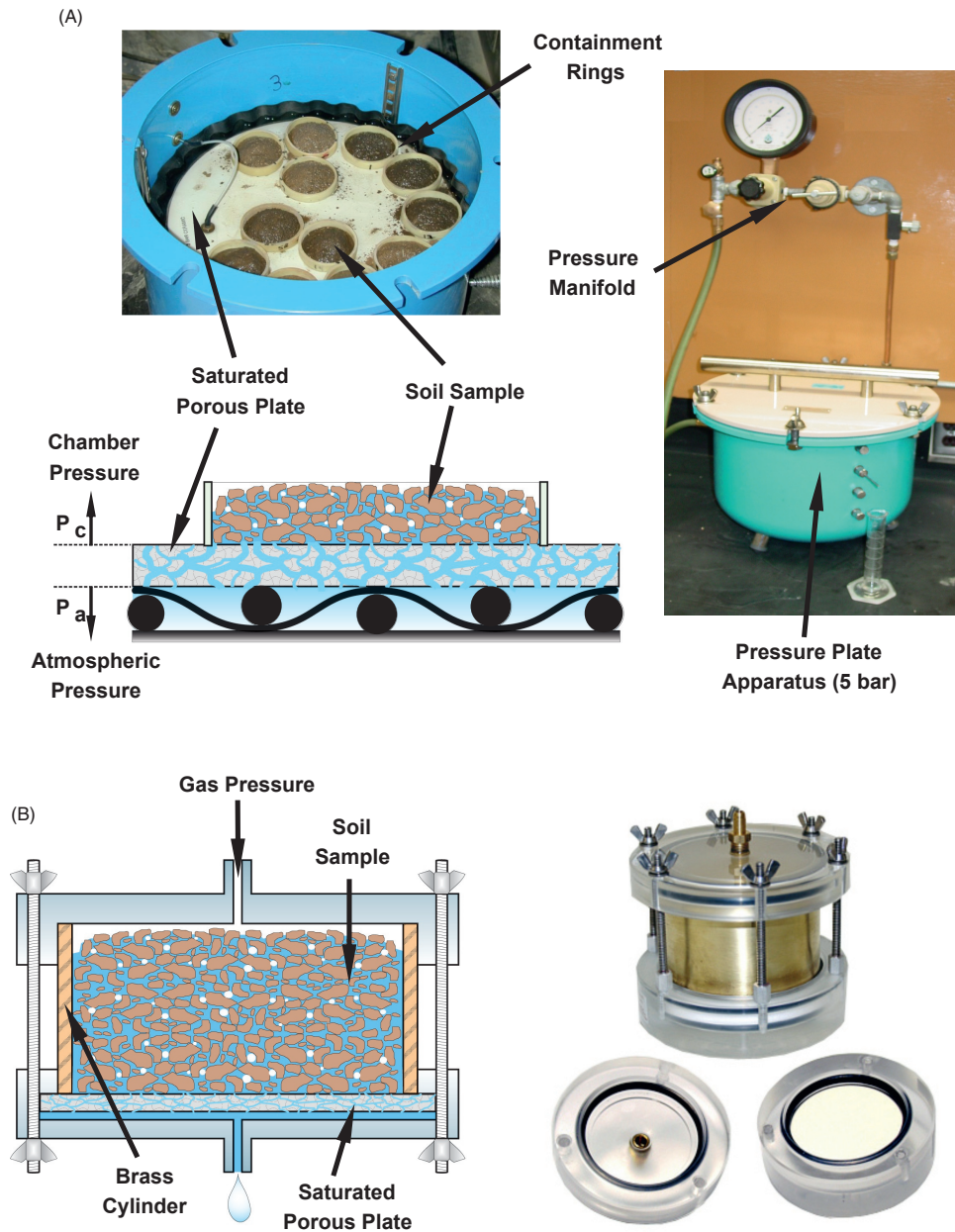


Fig. 11 (A) Pressure plate apparatus; and (B) Tempe pressure flow cell used to desaturate soil samples to specified matric potentials.

using Tempe cells. The cells may be sequentially disconnected from the pressure source, weighed to determine the change in water content from the previous step, then reconnected and the next pressure step applied. Outflow of water from the cells may be collected to calculate or confirm changes in sample water contents.

The evaporation method

Another laboratory technique for measurement of the SWC is the so-called evaporation method. It involves evaporation of water from the surface of an initially saturated soil column while concurrently measuring the change in matric potential and the water mass loss. Advanced benchtop evaporation systems such as the HYPROP (METER Group, Inc., Pullman, WA, USA) consist of two miniature transducer tensiometers for matric potential measurements, one installed close to the evaporating surface and the second close to the bottom of the column, and a highly accurate laboratory scale (Fig. 12) that continuously measures the water mass loss. The HYPROP system is fully automated and interfaced with a computer for monitoring the experiment and recording the measurements. From the readings of the 2 tensiometers (kPa), the average matric potential and the hydraulic gradient are calculated.



Fig. 12 State-of-the-art HYPROP benchtop evaporation system (METER Group, Inc., Pullman, WA, USA).

The mass difference, measured with the scale, is used to calculate the volumetric water content and the water flow rate. The experiment is terminated when either one of the tensiometers runs dry or the mass changes become marginal. Then, the remaining moisture content is determined by oven-drying the sample at 105 °C for 24 h. The analytic procedure assumes that the matric potential and water content distribute linearly through the column and that matric potential changes are linear between two evaluation points. The initial water content is determined from the total loss of water (i.e., evaporation and water loss by oven drying). Because HYPROP measurements are limited to the wet end of the SWC due to the functional range of the tensiometers, obtained data are commonly combined with WP4C Dewpoint Hygrometer data for dry conditions (see below). A specifically developed software package allows combining wet- and dry-end measurements and provides means to parameterize the [van Genuchten \(1980\)](#) and [Brooks and Corey \(1964\)](#) SWC models.

Dewpoint hygrometer method

Accurate dry-region SWC data may be measured with hygrometers that rely on the chilled-mirror dew point technique. These instruments measure the relative humidity (RH) in soil air that is considered to be in equilibrium with the soil water phase and convert RH into matric potential with the well-known Kelvin equation assuming negligible osmotic potential. State-of-the-art laboratory hygrometers include the WP4C Dewpoint PotentiaMeter and the fully automated Vapor Sorption Analyzer (VSA) (METER Group, Inc., USA) ([Arthur et al., 2014](#)) (Fig. 13). The WP4C instrument is capable of measuring matric potentials between 1 MPa and 300 MPa. The soil samples need to be brought to different predetermined moisture levels (e.g., 0.01, 0.02, 0.03, 0.04, and 0.05 kg kg⁻¹) and stored in a refrigerator for about 7 days to ensure equilibrium. Once equilibrium is attained, the samples are transferred into WP4C stainless steel sampling cups and soil water potential (matric + osmotic) is measured. Immediately after the measurements the mass of the wet samples is measured. After oven-drying, the exact gravimetric water content of each sample that corresponds to the measured water potential is determined. The VSA is a fully-automated instrument that provides two different isotherm generation modes – the Dynamic Dewpoint Isotherm (DDI) and the Dynamic Vapor Sorption (DVS) methods to measure water vapor adsorption as well as desorption isotherms (+ hysteresis). The VSA dries or wets an initially air-dry (or oven-dry) soil sample and automatically measures the water potential with a chilled-mirror dewpoint technique while simultaneously recording the sample mass throughout the drying or wetting process with a high-precision magnetic balance. Further details regarding measurement technology and procedures and the theoretical basis for subsequent calculations are presented in [Arthur et al. \(2014\)](#).

Paired sensors for field SWC measurement

In spite of the importance of measuring SWCs in situ, few suitable methods are available for field or in situ application. Paired sensors may be used to measure θ and ψ_m in the same or closely adjacent soil volumes, over a range of soil wetness. Examples include the use of neutron moisture meters or TDR sensors, together with tensiometers or more advanced ψ_m sensors (e.g., TEROS 21 sensor, Meter Group, Inc.; Tensiometer sensor, ecoTech Umwelt Meßsysteme). Application of paired sensors is often constrained by differences in soil volumes sampled by the respective sensors, different time constants required (e.g., many θ sensors obtain instantaneous measurements, while ψ_m methods require equilibrium), and the fact that few matric-potential techniques function over the entire range of interest for wetness. This commonly provides limited overlap in soil water retention measured using different techniques, and measurements obtained using different methods may not be consistent in the ranges of overlap.

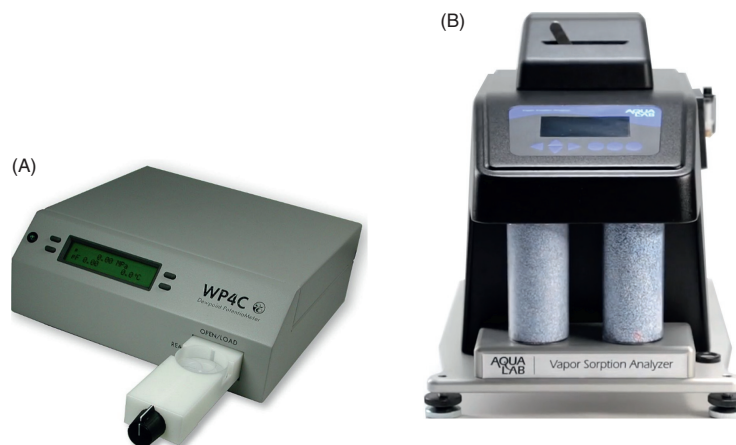


Fig. 13 WP4C and VSA laboratory hygrometers (METER Group, Inc., Pullman, WA, USA).

References

- Arthur E, Tuller M, Moldrup P, and de Jonge LW (2014) Evaluation of a fully automated analyzer for rapid measurement of water vapor sorption isotherms for applications in soil science. *Soil Science Society of America Journal* 78: 754–760. <https://doi.org/10.2136/sssaj2013.11.0481n>.
- Arthur E, Tuller M, Moldrup P, Resurrection AC, Meding MS, Kawamoto K, Komatsu T, and de Jonge LW (2013) Soil specific surface area and non-singularity of soil-water retention at low saturations. *Soil Science Society of America Journal* 77: 43–53. <https://doi.org/10.2136/sssaj2012.0262>.
- Batjes NH, Ribeiro E, van Oostrum A, Leenaars J, Hengl T, and Mendes de Jesus J (2017) WoSIS: Providing standardised soil profile data for the world. *Earth System Science Data* 9: 1–14. <https://doi.org/10.5194/essd-9-1-2017>.
- Berkowitz B and Ewing RP (1998) Percolation theory and network modeling applications in soil physics. *Surveys in Geophysics* 19: 23–72.
- Blunt MJ, Bijeljic B, Dong H, Gharbi O, Iglauer S, Mostaghimi P, Paluszny A, and Pentland C (2013) Pore-scale imaging and modelling. *Advances in Water Resources* 51: 197–216. <https://doi.org/10.1016/j.advwatres.2012.03.003>.
- Bouma J (1989) Using soil survey data for quantitative land evaluation. *Advances in Soil Science* 9: 177–213.
- Briggs LJ and Lane JWM (1907) *The moisture equivalents of soils*. (U.S. Department of Agriculture Bureau of Soils. Bulletin) U.S. Government Printing Office.23. Retrieved from <https://books.google.com/books?id=oWp2rgEACAAJ>.
- Brooks RH and Corey AT (1964) *Hydraulic Properties of Porous Media*. *Hydrology Papers, No. 3*. Fort Collins, CO: Colorado State University Press.
- Campbell GS and Shiozawa S (1992) Prediction of hydraulic properties of soils using particle-size distribution and bulk density data. In: van Genuchten MT, et al. (ed.) *Proceedings of the International Workshop on Indirect Methods for Estimating Hydraulic Properties of Unsaturated Soils*, pp. 317–328. Riverside: Univ. of Calif.
- Campbell GS (1974) Simple method for determining unsaturated conductivity from moisture retention data. *Soil Science* 117: 311–314.
- Carsel RF and Parrish RS (1988) Developing joint probability distributions of soil water retention characteristics. *Water Resources Research* 24(5): 755–769.
- Crawford JW (1994) The relationship between structure and the hydraulic conductivity of soil. *European Journal of Soil Science* 45: 493–502.
- Dai Y, Xin Q, Wei N, Zhang Y, Shangguan W, Yuan H, et al. (2019) A global high-resolution data set of soil hydraulic and thermal properties for land surface modeling. *Journal of Advances in Modeling Earth Systems* 11(9): 2996–3023. <https://doi.org/10.1029/2019MS001784>.
- de Boer F (2016) *HiHydroSoil: A High Resolution Soil Map of Hydraulic Properties (Version 1.2)*, Report Future Water. 134.
- Diamantopoulos E and Durner W (2015) Closed-form model for hydraulic properties based on angular pores with lognormal size distribution. *Vadose Zone Journal* 14(2). <https://doi.org/10.2136/vzj2014.07.0096>.
- Gimenez D, Perfect E, Rawls WJ, and Pachepsky YA (1997) Fractal models for predicting soil hydraulic properties: A review. *Engineering Geology* 48: 161–183.
- Golparvar A, Zhou Y, Wu K, Ma J, and Yu Z (2018) A comprehensive review of pore scale modeling methodologies for multiphase flow in porous media. *Advances in Geo-Energy Research* 2(4): 418–440. <https://doi.org/10.26804/ager.2018.04.07>.
- Gutmann E and Small E (2007) A comparison of land surface model soil hydraulic properties estimated by inverse modeling and pedotransfer functions. *Water Resources Research* 43: W05418. <https://doi.org/10.1029/2006WR005135>.
- Gupta S, Lehmann P, Bonetti S, Papritz A, and Or D (2021) Global prediction of soil saturated hydraulic conductivity using random forest in a Covariate-based GeoTransfer Function (CoGTF) framework. *Journal of Advances in Modeling Earth Systems* 13: e2020MS002242. <https://doi.org/10.1029/2020MS002242>.
- Huang H, Meakin P, and Liu M (2005) Computer simulation of two-phase immiscible fluid motion in unsaturated complex fractures using a volume of fluid method. *Water Resources Research* 41: W12413. <https://doi.org/10.1029/2005WR004204>.
- Jensen DK, Tuller M, de Jonge LW, Arthur E, and Moldrup P (2015) A New Two-Stage Approach to predicting the soil water characteristic from saturation to oven-dryness. *Journal of Hydrology* 521: 498–507.
- Karkare MV and Fort T (1996) Determination of the air-water interfacial area in wet “unsaturated” porous media. *Langmuir* 12: 2041–2044.
- Kim H, Rao PSC, and Annable MD (1997) Determination of effective air-water interfacial area in partially saturated porous media using surfactant adsorption. *Water Resources Research* 33(12): 2705–2711.
- Kosugi K (1994) Three-parameter lognormal distribution model for soil water retention. *Water Resources Research* 30(4): 891–901.
- Leão TP and Tuller M (2014) Relating soil specific surface area, water film thickness, and water vapor adsorption. *Water Resources Research* 50: 7873–7885. <https://doi.org/10.1002/2013WR014941>.
- Lebeau M and Konrad J-M (2010) A new capillary and thin film flow model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research* 46: W12554. <https://doi.org/10.1029/2010WR009092>.
- Leenaars J, Kempen B, van Oostrum A, and Batjes N (2014) In: Arrouays, et al. (ed.) *Africa Soil Profiles Database: A compilation of georeferenced and standardized legacy soil profile data for Sub-Saharan Africa*, 51. 2014.
- Lehmann P, Bickel S, Wei Z, and Or D (2020) Physical constraints for improved soil hydraulic parameter estimation by pedotransfer functions. *Water Resources Research* 56: e2019WR025963. <https://doi.org/10.1029/2019WR025963>.
- Lehmann P, Assouline S, and Or D (2008) Characteristic lengths affecting evaporative drying of porous media. *Physical Review E* 77(5), 056309. <https://doi.org/10.1103/PhysRevE.77.056309>.

- Mandelbrot BB (2010) *Les Objets Fractals: Forme, Hasard et Dimension*. Paris, France: Flammarion.
- Meakin P and Tartakovsky AM (2009) Modeling and simulation of pore-scale multiphase fluid flow and reactive transport in fractured and porous media. *Reviews of Geophysics* 47: RG3002. <https://doi.org/10.1029/2008RG000263>.
- Nemes A, Schaap MG, Leij FJ, and Wösten JHM (2001) Description of the unsaturated soil hydraulic database UNSODA version 2.0. *Journal of Hydrology* 251(3–4): 151–162.
- Or D (2019) The tyranny of small scales - On representing soil processes in global land surface models. *Water Resources Research* 55. <https://doi.org/10.1029/2019WR024846>.
- Or D and Tuller M (1999) Liquid retention and interfacial area in variably saturated porous media: Upscaling from single-pore to sample-scale model. *Water Resources Research* 35(12): 3591–3606.
- Ottoni MV, Ottoni Filho TB, Schaap MG, Lopes-Assad MLR, and Rotunno Filho OC (2018) Hydrophysical database for Brazilian soils (HYBRAS) and pedotransfer functions for water retention. *Vadose Zone Journal* 17(1–17): 2018.
- Rossi C and Nimmo JR (1994) Modeling of soil water retention from saturation to oven dryness. *Water Resources Research* 30(3): 701–708.
- Scholle PA (1979) *A Color Illustrated Guide to Constituents, Textures, Cements, and Porosities of Sandstones and Associated Rocks*. American Association of Petroleum Geologists Memorandum 28. Tulsa, OK: American Association of Petroleum Geologists.
- Sukop M and Or D (2003) Lattice Boltzmann method for modeling liquid–vapor interface configurations in porous media. *Water Resources Research* 40(1): W01509.
- Tuller M and Or D (2005) Water films and scaling of soil characteristic curves at low water contents. *Water Resources Research* 41: W09403. <https://doi.org/10.1029/2005WR004142>.
- Tuller M, Or D, and Dudley LM (1999) Adsorption and capillary condensation in porous media-liquid retention and interfacial configurations in angular pores. *Water Resources Research* 35(7): 1949–1964.
- van Genuchten MT (1980) A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal* 44: 892–898.
- Veihmeyer FJ and Hendrickson AH (1927) The relation of soil moisture to cultivation and plant growth. Proc. 1st intern. *Congress of Soil Science* 3: 498–513.
- Zhang Y and Schaap MG (2017) Weighted recalibration of the Rosetta pedotransfer model with improved estimates of hydraulic parameter distributions and summary statistics (Rosetta3). *Journal of Hydrology* 547: 39–53. <https://doi.org/10.1016/j.jhydrol.2017.01.004>.

Soil water repellency

Jörg Bachmann and Marc-Oliver Goebel, Institute of Soil Science, Leibniz Universität Hannover, Hannover, Germany

© 2023 Elsevier Ltd. All rights reserved.

This is an update of J. Letey, WATER-REPELLENT SOILS, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 301-306, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00519-1>.

Soil water repellency: How relevant?	203
Wettability and surface energy	204
What makes a soil water repellent?	206
Quantification of soil water repellency	207
Methods based on drop infiltration	207
Capillary rise method (CRM) and water entry pressure	207
Repellency index (RI)	208
Sessile drop method (SDM)	208
Wilhelmy plate method (WPM)	208
Surface free energy components	209
Soil processes affected by water repellency	209
Infiltration and preferential flow	209
Microscale water configuration	210
Evaporation	211
Indirect effects of soil water repellency	212
Aggregate stability	212
Erosion	212
Microbial activity and carbon sequestration	212
Plant growth	212
Moisture-affected dynamics of soil water repellency	212
Modelling	213
Dealing with soil water repellency	214
Conclusions and outlook	214
References	214

Abstract

Hydrophobicity ("fear of water") and hydrophilicity ("affinity to water") are qualitative terms that describe the extent of interaction of solid surfaces with water, normally observed in the 3-phase system solid-liquid-air. Hydrophobic dry soil does not imbibe water spontaneously, but rather causes it to ball up and remain at the surface. The water may then penetrate preferentially at other locations, resulting in fast and local movement into deeper layers while leaving dry spots in close proximity. Soil water repellency occurs worldwide and is the cause of many important phenomena like poor water infiltration, heterogenous and fragmented soil moisture patterns, increased overland flow, erosion, reduced irrigation water efficiency and poor plant growth. Due to expected climate change effects, the impact of soil water repellency can be expected to become even more important in the future.

Key points

- Soil water repellency is a decrease in the wetting rate of soil due to its surface properties.
- Extremely water repellent soil has a soil-water contact angle >90 degrees and does not wet. If soils have a contact angle <90 degrees but >0 degree, they are subcritically water repellent.
- Preferential flow paths and dry regions are typical of water repellent soils.
- Water repellency affects erosion, aggregate stability and carbon cycling.

Soil water repellency: How relevant?

Soil water repellency (SWR) was first reported more than 100 years ago, mainly as a rare curiosity associated with specific environments (DeBano, 2000). The early research identified the importance of soil organic matter (SOM) and microorganisms to SWR development, prompting research on how these interactions affect physical, chemical and biological soil processes. As SWR

research increased, its prevalence globally became evident, even in unexpected wet and cool maritime environments. Beyond visually obvious SWR, where water beads on soil surfaces with no infiltration, water infiltration into most soils has been observed to be somewhat influenced by SWR. Soil management and irrigation practices were found to be large drivers of SWR development.

An increase in SWR may cause a substantial reduction in soil water availability, often resulting in poor plant growth and poor irrigation efficiency (Olorunfemi, 2014). Large areas of land can be affected by SWR, such as 2 million ha of agricultural land in South-Western Australia (Blackwell, 2000) alone, where SWR presents a major challenge to agricultural productivity. Nowadays, in conjunction with a still growing world population and increasing demand for food, research to increase soil productivity is important. Worldwide, new trends in agriculture including no-till systems and intensified use of (treated) waste water for irrigation, and the amendment of soil with biochar, manure and other organic compounds have been observed to increase the level of SWR. Recently, further challenges that may exacerbate SWR have emerged. Of particular importance is the recognition of climate change effects, with expected more frequent heavy rainfall and extended water stress periods. In short, SWR may be considered as a feature of many soils that presents a serious threat to a range of ecosystem services.

Examples of processes and phenomena, either initiated or enhanced by SWR, are poor water infiltration, increased overland flow, enhanced wind and water erosion, heterogeneous and fragmented soil moisture patterns, rapid water and solute transport caused by preferential water flow pathways, reduced irrigation water efficiency, poor plant growth and soil carbon sequestration as schematically depicted in Fig. 1.

Few of these SWR-affected processes are represented adequately in recent soil physics textbooks, where it is generally assumed that soils are perfectly wettable. Established and accepted theories based on “classical” capillary concepts like the Richardson-Richards equation for water transport are, strictly speaking, only valid for hydrophilic soils (Or and Tuller, 2022).

This chapter is a continuation of the Chapter “Water Repellent Soils” by Letey (2005), included in the previous edition of the Encyclopedia of Soils in the Environment. The text also refers to various aspects of Or and Tuller’s chapter on “Capillarity” in the present edition. We finally refer to the excellent historical review on more than 100 years of SWR research by DeBano (2000) and the most recent review by Smettem et al. (2021).

Wettability and surface energy

Soil water repellency controls the water dynamics of dry or partly saturated soil (Fig. 2). Instead of immediately infiltrating within seconds, water balls up on the water repellent soil surface. The time it takes for a drop of water to penetrate the soil is a measure of the persistence of SWR, often termed and measured as “water drop penetration time”, WDPT. Additional water at the surface will flow as surface flux to local depressions and may penetrate preferentially along coarse pores or macropores, resulting in fast and preferential flow into deeper layers, while leaving permanently dry spots in close proximity. Extensive infiltration through the entire cross-section does not occur until the water is in contact with the soil for a time equal to or greater than the WDPT, or the water head

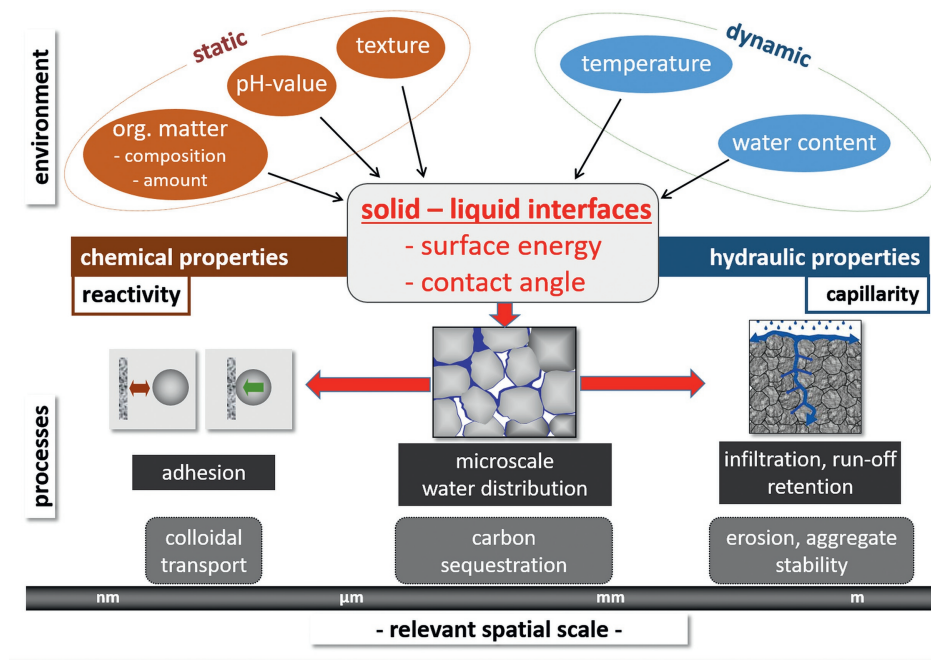


Fig. 1 Chemical and physical factors that influence soil water repellency and selected scale-dependent soil processes affected by soil water repellency.



Fig. 2 Water droplets on dry peat. Water balls up and remains at the surface. Photo by S.K. Woche.

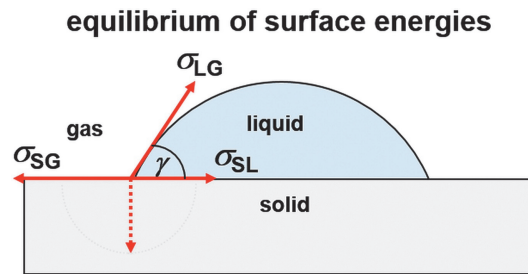


Fig. 3 Subcritical water repellent solid surface. The contact angle γ is the angle between the liquid wetting front and the solid surface at the 3-phase contact line and results from equilibrium of the three surface energies: solid-gas σ_{SG} , liquid-gas σ_{LG} , and solid-liquid σ_{SL} . The dashed arrow shows the strain field component caused by the forces at the 3-phase contact line (Good, 1992). Modified from Good RJ (1992) Contact angle, wetting, and adhesion: a critical review. *Journal of Adhesion Science and Technology* 6: 1269–1302.

at the surface becomes greater than the water-entry pressure. Over time, contact between water and soil mineral particles may decrease SWR, with the time taken defined as the persistence of SWR.

The solid-gas surface energy, σ_{SG} , is the fundamental physicochemical property of a solid that governs wettability and capillarity. The expression which connects the surface energies of the solid-liquid-air system with the visible behavior of water, as displayed by the contact angle (CA) visible at the 3-phase boundary line of a water droplet (Fig. 3), was established by Young (1805) and is known as Young's equation.

$$\cos \gamma = \frac{\sigma_{SG} - \sigma_{SL}}{\sigma_{LG}} \quad (1)$$

In Eq. (1) γ is the (intrinsic) CA [°], σ is the surface energy [mJ m⁻²], and the indices S, L, and G denote the solid, liquid, and gas phase, respectively. From Eq. (1) it is evident that wetting of a solid surface by a liquid is controlled by competition between the different surface energies (Fig. 3). For details on the impact of surface roughness and chemical heterogeneity on the CA, we refer to Or and Tuller (2022).

For soil particles, the CA usually varies between 0 degree (completely wettable) and around 140 degrees (extremely water repellent). Soils showing CA >90degrees are usually classified as hydrophobic, whereas soils characterized by CA <90degrees are often referred to as subcritical water repellent (Fig. 4). Zisman (1964) proposed the so-called critical surface energy σ_c of a solid, which equals the σ_{LG} of a liquid showing a zero-degree CA with this solid. The value of σ_c can be determined by measuring CA using a series of liquids with different surface energy. Low solid surface energy σ_{SG} imposes weak attraction between particles and the liquid phase. Generally, pure soil minerals are completely wettable or exhibit small CA due to high-energy surfaces (Lewin et al., 2005). However, soil particles under natural conditions are usually covered by organic substances or mixed with particulate organic matter (Doerr et al., 2000).

Several studies show that SWR is positively correlated with the soil organic carbon content (e.g., Varela et al., 2005). Organic matter usually reduces the particle surface energy, where particularly nonpolar hydrophobic functional groups like —CH₃ likely decrease the wettability and cause hydrophobicity (Tschapek, 1984). This may result in a large number of non-polar sites on the particle surfaces, thick enough to mask the mineral surface properties. The chemical drivers are the amount of hydrophobic —CH groups relative to that of hydrophilic C=O, C—OH, and C—NH₂ groups at the outermost solid-liquid interface of a few Ångström thickness that determine the wettability of SOM (Ferguson and Whitesides, 1992).

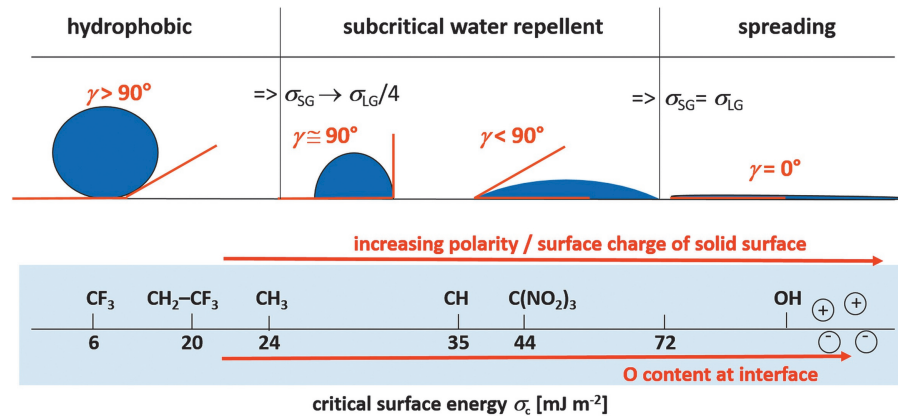


Fig. 4 Different wetting regimes are characterized by the relationship between solid-air σ_{SG} and liquid-air σ_{LG} surface energies. The critical surface energy σ_c increased (i) with increasing atomic oxygen (O) concentration and lower carbon (C) concentration and (ii) higher concentration of surface charges.

What makes a soil water repellent?

There are several factors affecting soil wettability (see Fig. 1), the most important of which are briefly discussed below. Soils with high SOM content are usually more likely to develop SWR. The origin of SOM is primarily plant derived organic matter like roots and their exudates, plant tissues and in situ formed pedogenic organic compounds like humic substances. The composition of SOM was found to be more relevant to explain the extent of SWR than the geologic substrate that forms soil minerals. Certain trees, shrubs and grass species are associated with the occurrence of SWR because of the SOM they deposit in soils. Most prevalent are evergreen trees (e.g., *Eucalyptus* spp., *Pinus* spp.) due to the considerable amount of resins, waxes or essential oils they produce in their stems and leaves.

In the soil, various other biological drivers contribute to SWR development. Microbial biofilms and fungal activity create hydrophobic tissues and *exo*-polymeric substances that coat soil minerals and affect soil water characteristics (Doerr et al., 2000). The roots of some plant species exude organic compounds that enhance water affinity and storage capacity above a threshold water content, but switch to water repellency at a lower water content. Hysteretic effects, i.e., differences between the sorption and desorption water retention curves, increase significantly with increasing root exudate concentration (Moradi et al., 2012). Soil moisture dynamics and transport processes at the root-soil interface have become a subject of intensive research (Benard et al., 2018). Numerous studies show a complex modification of physical properties in the rhizosphere compared to surrounding bulk soil.

Due to the wide spectrum and sources of compounds, SWR has been observed for many soils, but it is most prevalent in sandy soils due to the low particle volume to surface ratio and normally uncharged surfaces. In contrast, pedogenic mineral compounds like iron oxides or clay minerals have much greater surface area and usually exhibit charged surfaces or surfaces with functional groups that support wettability, thus reducing SWR. In addition, SWR is often found to be negatively correlated with pH, which is explained by a decreasing polarity of SOM functional groups due to protonation at low pH (de Jonge et al., 1999).

Another factor potentially affecting SWR is the occurrence of wildfires, which are frequently shown to either create or enhance the severity of SWR (Varela et al., 2005), e.g., by reinforcing the bonding of hydrophobic organic molecules to soil particles and chemically intensifying their hydrophobicity by pyrolysis. However, fire-induced reduction and even destruction of SWR in topsoils have also been reported depending on the extent of burning and SOM properties.

However, the modification of water repellency due to environmental processes is a behavior of soil which is still not well understood. To explain changes in wettability as a response to changing physical environmental conditions, Ferguson and Whitesides (1992) proposed a conceptual model of the outermost organic interface, combining theoretical considerations with different experimental methods. They analyzed the wettability of functionalized polyethylene surfaces with several interface-sensitive techniques (CA measurement, X-ray photoelectron spectroscopy, XPS, and attenuated total reflectance Fourier-transform infrared spectroscopy, ATR-FTIR). From this information “interphases” were defined according to the physical depth specifically probed by the different techniques (Fig. 5). The “CA interphase” refers to the outer 0.5–1.0 nm layer whose chemical compounds determines surface wettability. The “XPS interphase” refers to approximately the outer 1–3 nm, the analysis depth of XPS, while the “ATR-FTIR interphase” as the deepest interphase of interest refers to the outer 1 μ m or deeper.

Ferguson and Whitesides (1992) reported that functional groups may still be present close to the CA interphase, but located too deep to influence wetting. They are nonetheless accessible to reagents in solution and able to migrate into the CA interphase, and Ferguson and Whitesides (1992) termed the portion of the interface where this type of interaction is possible as “sub-CA interphase”. The depth of the sub-CA interphase is between that of the XPS and the ATR-FTIR interphase. It is determined by permeability, pore structure, and solid–liquid interactions and is generally less well defined than the depths associated with the respective analytical techniques. This model also explains moisture and temperature affected modification of soil wettability.

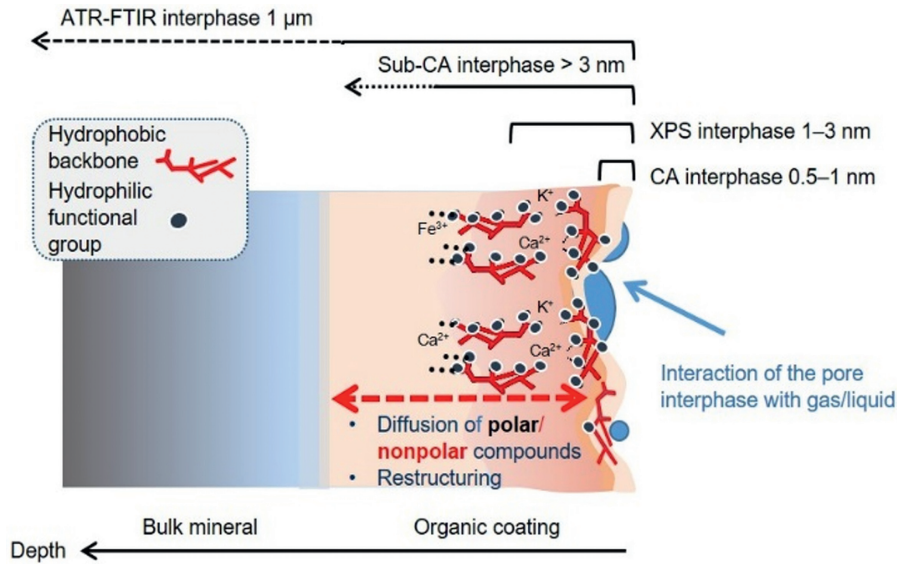


Fig. 5 Schematic representation of the different analytical penetration depths at the solid interphase. Figure redrawn after Ferguson GS and Whitesides GM (1992) Thermal reconstruction of the functionalized interface of polyethylene carboxylic acid and its derivatives. In: Schrader ME, Loeb GI (eds.), *Modern Approaches to Wettability: Theory and Applications*. Springer US: Boston, MA, pp. 143–177; Ellerbrock RH, Gerke HH, Bachmann J, Goebel MO (2005) Composition of organic matter fractions for explaining wettability of three forest soils. *Soil Science Society of America Journal* **69**: 57–66.

Quantification of soil water repellency

Methods based on drop infiltration

It is impossible to measure directly the surface energy of irregularly shaped particles such as soil grains. Therefore, indirect and simple methods based on the penetration behavior of liquids into porous medium of interest were developed (Wallis and Horne, 1992; Letey, 2005). This group of methods includes the water drop penetration time (WDPT) test, the molarity of a water-ethanol droplet (MED) test and numerous subvariants based on the same principle. The WDPT test is the most common method and applied in the laboratory as well as in the field. This test places a drop of water of a controlled volume onto the soil surface and the time required to completely penetrate into the soil is determined. Usually, 5 main WDPT classes are distinguished: non repellent (0–5 s), slightly repellent (5–60 s), strongly repellent (60–600 s), severely repellent (600–3600 s) and extremely repellent (>3600 s) (Dekker and Ritsema, 1994). This test only provides qualitative information, but it is useful to characterize soils with a high degree of water repellency and for in situ investigations under ambient moisture conditions. Over time, the drop of water may transform hydrophobic organic compounds to hydrophilic, so WDPT provides a measurement of SWR persistence.

The MED test is a variant of the WDPT test. It uses water-ethanol mixtures to reduce the surface energy σ_{LG} (i.e., the surface tension) of the solution with increasing ethanol concentration. The 90 degrees-surface tension (σ_{nd}) is the surface tension of the solution where transition from sitting on the surface to penetration occurs. The target value is the ethanol concentration at which the droplet penetrates within 5 s. This concept assumes that a liquid can only penetrate completely and quickly into a porous medium for $\gamma < 90$ degrees.

Theoretical relationships between various surface tensions of liquid and solid and CA have been combined to obtain the following two equations (Letey, 2005):

$$\sigma_{SG} = \sigma_{nd}/4 \quad (2)$$

$$\cos \gamma = (\sigma_{nd}/\sigma_{LG})^{1/2} - 1 \quad (3)$$

Accordingly, the measurement of σ_{nd} can be used to calculate the value of σ_{SG} and γ . It should be noted that the MED test is somewhat unsatisfactory, because of its empirical nature and the lack of a physical foundation (Letey, 2005).

Capillary rise method (CRM) and water entry pressure

The capillary rise method (CRM) can be applied to determine the CA of granular material such as soil particles. Air-dry and sieved (usually <2 mm) material is packed into a tube, which is closed at the lower end with filter paper. The base of the upright tube is brought into contact with a test liquid in a reservoir, and the mass gain of the tube can be recorded with an electronic balance. The maximum height h_m that is reached by the wetting front can be expressed as (Washburn, 1921):

$$h_m = \frac{2\sigma_{LG} \cos \gamma_{st}}{\rho g R_{st}} \quad (4)$$

where h_m is the height of the wetting front above the base of the tube [m], ρ is the density of the liquid [Mg m^{-3}], g is the acceleration due to gravity [m s^{-2}], R_{st} is the mean static radius of pores for an idealized bundle of parallel capillaries [m], and γ_{st} is the static CA [$^\circ$] (Wang and Wallach, 2021). However, reaching equilibrium takes time, often weeks even for small columns. The rate of the rise of the wetting front in the powder-filled glass tube (dh/dt) by means of Poiseuille's law is:

$$w^2 = c \frac{\rho^2 \sigma_{LG} \cos \gamma}{\eta} \quad (5)$$

Eq. (5) suggests a linear relationship between the liquid uptake into the column w^2 and t , with the slope m . This equation cannot be used directly to determine γ , because h is a function of the two unknown parameters R_{st} and γ . This applies to a liquid that partly wets, such as water, as well as to a low surface tension liquid that completely wets the particles, such as *n*-hexane, yielding $\cos \gamma = 1$. The effective pore radius and the CA are not known a priori, therefore *n*-hexane is used as a reference liquid to evaluate first the soil-specific factor c in Eq. (5). Parameter c can be derived from the slope of the function $w^2 = mt$ and the physical properties of *n*-hexane. Once c is known, the capillary rise experiment is carried out with water, from which the CA for water is then determined (Siebold et al., 1997).

A disadvantage of the CRM is that the testing liquid (water or soil solution) does not penetrate for $\gamma > 90$ degrees, which excludes hydrophobic soil from measurement. In this case, a positive pressure, h_p , is needed for penetration and a water head equal to or greater than h_p must exist for water entry. Note that the sign in Eq. (4) changes for $\gamma > 90$ degrees. Measurement of h_p requires an apparatus that allows the height of a water column above the soil to be gradually increased and the ability to determine when the water begins to penetrate the soil. It is interesting to note that the infiltration rate into wettable soils is only slightly affected by the depth of ponded water, whereas infiltration into water-repellent materials is drastically affected by the depth of ponded water h_p .

Repellency index (RI)

An index of SWR can also be determined from the sorptivities of water and ethanol, with the latter being considered as a completely wetting liquid due to its small surface tension. Hexane, as used for CRM, can also be used for these measurements. The water repellency index (RI) compares the infiltration rates of both liquids from a tension infiltrometer, either in the field or the laboratory. Due to ethanol being used as a wetting liquid, MiniDisk infiltrometers (Meter, Pulman, USA) are increasingly used for field studies because they require less liquid than conventional infiltrometers. In the laboratory, small-scale infiltrometers attached to logging balances can obtain measurements of RI at millimeter resolution, such as on the surface of soil aggregates. Common to tension infiltrometers is an ability to control and adjust the tension head of the infiltrating liquid, thereby minimizing macropore flow and allowing for hydraulic conductivity to be measured using multiple tensions. An advantage of the RI approach with infiltrometers is the measurement of intact soil samples, as abrasion from sieving affects SWR. However, the approach does not provide a direct measurement of CA and the shape of the wetting front in hydrophobic soils may affect measurements. Disadvantages described above for CRM also apply. For further details see Olorunfemi (2014).

Sessile drop method (SDM)

The sessile drop method (SDM) measures the CA of a sessile drop on soil minerals directly using a goniometer fitted microscope. Advanced devices use video recording and computer-aided CA evaluation of the drop contour as shown schematically in Fig. 3. The measurement is static, taken immediately after placing the drop onto the sample, but it can also be made dynamically (i.e., advancing and receding CA mode) by adding or removing liquid from the drop during the measurement or by tilting the sample. Soil samples are usually prepared by pressing a one-grain layer of dry soil particles onto a glass slide covered by double-sided adhesive tape (Bachmann et al., 2000). The initial CA is evaluated immediately after placing a small drop of water (e.g. 1 μL). An advantage of the SDM is that different liquids may be used to evaluate both wettability and surface energy components of soil particles (see Eq. 8). It should be noted that the sessile drop CA, in the strict sense, does not allow for the derivation of exact thermodynamically valid data because soil particle surfaces are rough, heterogeneous, anisotropic, elastic, deformable, and may react with the wetting liquid (Drehlich et al., 2019). Nevertheless, the derived values are useful approximations, which can be applied for many research questions.

Wilhelmy plate method (WPM)

This method is based on a procedure originally proposed by Wilhelmy (1863). A solid body that is partly immersed into a liquid is subject to at least three different forces: the gravity force which acts vertically downward, the buoyancy force which acts vertically upward and the interfacial menisci forces at the 3-phase contact line around the body. In the original setup, a platinum plate is connected with a balance and then lowered into a reservoir filled with a test liquid, providing a measurement of surface tension, assuming a zero CA of the plate. The principle has been inverted for soil studies, by using a liquid with a known σ_{LG} (e.g., water) to evaluate the unknown CA (Bachmann et al., 2003). A small glass slide, with adhesive tape and a one-grain layer of soil particles on

all sides can be used. The buoyancy force F_b increases linearly with the plate's depth of immersion. At any stage, while the plate is partly immersed, the total force F_t that is acting on the plate can be given as:

$$F_t = W - F_b - F_w = W - V\rho g + l_w\sigma_{LG}\cos\gamma \quad (6)$$

The symbol W [kg] in Eq. (6) denotes the mass of the sample, V [m³] is the volume of the immersed part of the sample, and l_w [m] is the wetted length of the sample. The CA can be calculated by:

$$\cos\gamma = \frac{(F_t + V\rho g)}{l_w\sigma_{LG}} \quad (7)$$

One advantage of the WPM is that it covers the entire CA range between 0 and 180 degrees and, depending on the immersion dynamics and direction (into or out of the liquid reservoir), allows for the measurement of dynamic advancing and receding CA, similar to SDM. Due to the physical principle applied, advancing CA measured with WPM are larger and receding CA smaller than corresponding static CA measured with SDM.

Surface free energy components

Based on CA data it is possible to determine polar and non-polar surface free energy components. These components are, for instance, important to predict interactions between colloidal particles (e.g., clay minerals and bacterial cells) and soil pore surfaces in aqueous media (Bos et al., 1999). The relevant surface energy components of colloids and adsorbing surfaces can be determined experimentally using the van Oss-Chaudhury-Good Eq. (8) (van Oss et al., 1988):

$$(1 + \cos\gamma)\sigma_{LG} = 2\left(\sqrt{\sigma_{SG}^{LW}\sigma_{LG}^{LW}} + \sqrt{\sigma_{SG}^{+}\sigma_{LG}^{-}} + \sqrt{\sigma_{SG}^{-}\sigma_{LG}^{+}}\right) \quad (8)$$

The superscripts 'LW', '+', and '-' denote the non-polar Lifshitz-van der Waals component, the electron-acceptor (acid) component, and the electron-donor (base) component, respectively. To determine the three unknown variables, σ_{SG}^{LW} , σ_{SG}^{+} , σ_{SG}^{-} , in Eq. (8), the solution of a system of three independent linear equations is needed. This can be achieved by measuring CA using three different liquids (with two of them being polar) with known surface energy components, e.g. deionized water, ethylene glycol, and α -bromonaphthalene.

Soil processes affected by water repellency

Water repellency is an important controlling factor of soil hydraulic processes and affects potentially all processes in the soil where water is involved. SWR can have significant effects on water infiltration, water storage, solute transport in the unsaturated zone, soil water balance, surface runoff and soil erosion, soil fertility, irrigation efficiency, aggregate stability and the protection of SOM against microbial degradation. Not all of these SWR impacts are negative. Small levels of SWR, previously described as subcritical, may enhance aggregate stability and the physical protection of SOM. It is evident that SWR has important repercussions on different spatial scales, ranging from the landscape to the pore scale.

Infiltration and preferential flow

If one imagines a strong precipitation event occurring after a long drought period with precipitation falling on a dry water repellent soil surface, the immediate effects of SWR on soil hydrology become intuitively conceivable. At the onset of such an event, infiltration will be considerably reduced and precipitation may pond or run-off the soil surface, with a variable portion eventually infiltrating via preferential flow paths into the soil. The initial effect of SWR leads to a different partitioning of precipitation, decreasing the amount infiltrating water and increasing the amount of run-off water. The impact of SWR can be very pronounced, as demonstrated in a study by Lemmnitz et al. (2008), where the fraction of surface run-off for a subcritically water repellent plot (CA $\approx$ 45 degrees) was found to be 96.4% as compared to 11.4% for a completely wettable plot that received the same amount of precipitation.

Infiltration into a water repellent bulk soil matrix will usually not occur until water has been in contact with the soil surface for a time equal to or greater than the WDPT, corresponding to a CA < 90 degrees. The degree of repellency usually reduces with time after contact with water (Wang and Wallach, 2021), which explains the typically observed increase in infiltration rate with ongoing precipitation. In flat or gently sloping terrain, where run-off flow is typically small and water may pond on the water repellent soil surface, infiltration can also be initiated when the depth of the water layer at the surface is greater than the water-entry pressure. The special features of water repellent soils are clearly reflected in their infiltration dynamics. While a wettable soil is typically characterized by a fast initial infiltration rate, which decreases and eventually becomes relatively constant with time when sufficient water is available at the surface, the rate of infiltration into a water repellent soil is slow during the initial phase of infiltration and increases with time (Fig. 6).

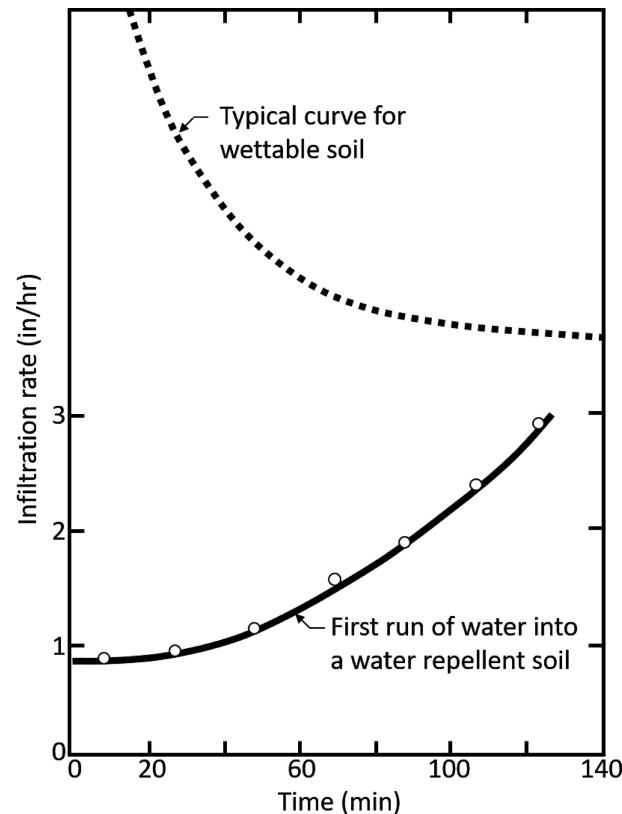


Fig. 6 Typical infiltration curves of a water repellent and a wettable soil. Figure redrawn from Letey J, Osborn J, Pelishek RE (1962) The influence of the water–solid contact angle on water movement in soil. *International Association of Scientific Hydrology Bulletin* 7: 75–81.

If the water infiltrates through preferential flow paths into the subsoil it will by-pass large domains of the topsoil, which can lead to dry surface soil and greater moisture content in the subsoil. In general, preferential flow effects triggered by SWR can cause partitioning of infiltrating water, potentially leading to the formation of wet and dry domains in close proximity, where differences in volumetric water content can reach more than 35% (e.g. Dekker and Ritsema, 1995). These domains can be stable for months. Water repellent properties can induce very high levels of small (mm)-scale variability in soil water movement (Hallett et al., 2004), but the exclusion of water from water repellent soil domains can also be effective at a larger (m)-scale (Oostindie et al., 2008).

As SWR is strongly dependent on SOM content, it is usually mostly pronounced in topsoils. The effects of SWR-induced preferential flow therefore typically decrease with depth, frequently leading to a more homogeneous moisture distribution in the deeper subsoil. But SWR has also been found to increase with depth, exerting a strongly significant impact on soil moisture distribution and water movement in the deeper subsoil. Even if a subsoil is completely wettable, the preferential flow phenomena, once initiated in the topsoil or in the organic layer, may also be effective at deeper depths. For instance, in a sloped forest soil Gerke et al. (2015) observed surface run-off underneath an organic topsoil (biomat) layer caused by a water repellent A horizon. From the analysis of the staining pattern after irrigation (Fig. 7), they concluded that the biomat layer and mineral soil surface topography generated locally preferential flow pathways in deeper mineral soil layers instead of homogeneous infiltration into the wettable mineral soil.

However, these phenomena are not restricted to hillslopes, but can also be found in flat terrain, as visualized by the preferential infiltration pattern found for a loamy sandy beech forest site, reaching depths of more than 130 cm in the wettable subsoil (Fig. 8). For a more quantitative discussion on the mechanisms causing preferential flow phenomena we refer to Letey (2005) and Hendrickx et al. (1993).

Microscale water configuration

SWR may also have important repercussion on the microscale (pore or particle scale) distribution and continuity of the liquid phase in the soil matrix. Microscopic water distribution in the wetted pore system of a water-repellent soil can differ strongly from a wettable soil at the same soil water content (Muehl et al., 2012). Environmental scanning electron micrographs have revealed pronounced differences in the microscale water configuration on subcritically water repellent ($CA \approx 50$ degrees) and strongly

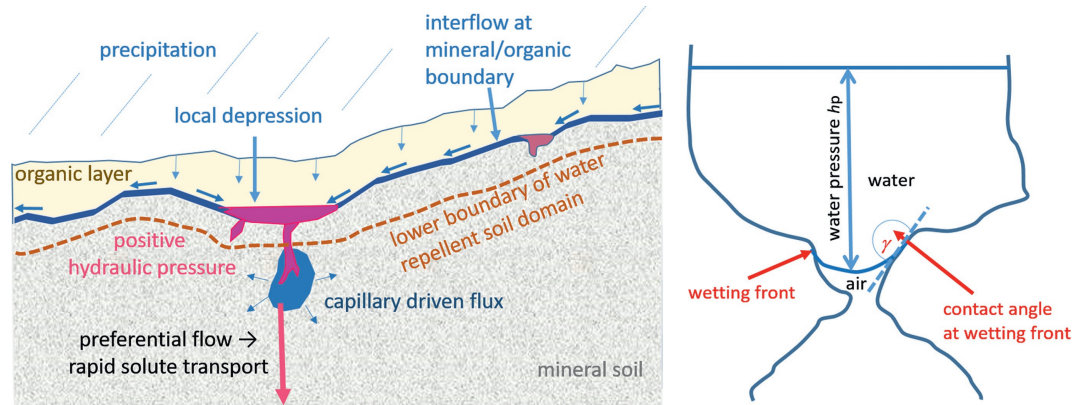


Fig. 7 Schematic representation of water meniscus entering an air-wet pore throat. Negative values are for water infiltration with $CA < 90$ degrees, positive values for $CA > 90$ degrees, as indicated by the height of the water level. Consequences are surface run-off generation and local infiltration into the hydrophobic mineral topsoil layer. Figures redrawn after Gerke KM, Sidle RC, Mallants D (2015) Preferential flow mechanisms identified from staining experiments in forested hillslopes. *Hydrological Processes* **29**, 4562–4578.



Fig. 8 Moisture distribution 24 h after infiltration with “Brilliant Blue” (irrigation rate 22 mm/h, total amount 200 mm). The testing site is a 100 year beech stand on loamy sand. Organic top layer and topsoil were rated as hydrophobic ($CA > 90$ degrees). Zones of preferential water flow are clearly indicated by dyed areas. Formation of corresponding flow pathways accelerate water and solute transport into ground water and reduces plant available water.

hydrophobic ($CA \approx 160$ degrees) particles (Goebel et al., 2011). While slightly water repellent particles were wetted uniformly, promoting good water connectivity between the particles, water formed small droplets on the hydrophobic particles, resulting in greatly reduced water connectivity between individual particles, confined by the diameter of the water drops formed between them. In addition, water film thickness on particle surfaces can be strongly affected by particle wettability. It has been shown that water film thickness on slightly water repellent quartz particles ($CA = 10$ degrees) is reduced by a factor of 5 as compared to completely wettable quartz particles (Chruaev, 2000), showing that already small differences in wettability can have a strong impact on water film thickness.

Evaporation

The development of a water repellent surface layer can also reduce subsequent drying of deeper soil layers by preventing evaporation and upward capillary movement of water. The effectiveness of SWR to reduce evaporation seems to be more affected by the location (depth) of the water repellent layer within the soil profile than by its thickness (Shokri et al., 2008). The reduction in evaporation in the presence of a water repellent layer is usually explained by the interruption of capillary flow, suggesting that the flux across the water repellent layer is purely diffusive. However, investigations with mixtures of hydrophilic and hydrophobic sand grains revealed gradually decreasing evaporation with an increasing fraction of hydrophobic grains (Shokri et al., 2009). This suggests that the primary mechanism for the reduction in evaporation is not necessarily a complete interruption of capillary flow but rather a reduction in capillary driving forces.

Indirect effects of soil water repellency

Aggregate stability

Several studies have shown that increasing SWR favors the stability of soil aggregates. The protection of aggregates due to SWR is mainly related to the reduction of the initial wetting rate, which diminishes the build-up of air-pressure in soil pores, thus reducing the slaking stress. A decrease in aggregate stability is often accompanied by a decrease in hydrophobic organic matter compounds, such as lipids. The important role of hydrophobic organic compounds was demonstrated by [Piccolo and Mbagwu \(1999\)](#), showing that aggregate stability markedly increased after the application of hydrophobic substances. The particular role of SOM in stabilizing soil aggregates is not only caused by its effect on SWR but also due to mechanical stabilization mechanisms (e.g. by enmeshment of mineral particles), which often makes it difficult to clearly differentiate the effects of SWR on aggregate stability. However, [Mataix-Solera and Doerr \(2004\)](#), for instance, could demonstrate for soils with similar organic carbon contents that aggregate stability is clearly enhanced with increasing SWR. [Goebel et al. \(2005\)](#) showed that a CA difference of only 13 degrees can significantly reduce water uptake rates and enhance aggregate stability of air-dry aggregates immersed in water, concluding that wettability is a better predictor of the initial aggregate breakdown dynamics than the total amount of organic carbon, particularly at very low organic carbon contents.

Erosion

Because of their reduced infiltration capacity, water repellent soils have often been associated with the occurrence of soil erosion, and several authors found a positive relationship between the degree of SWR and the amount of soil mass displaced by erosion ([Shakesby et al., 2000](#)). Particularly, in coarse-textured soils or soils with low organic matter content, SWR can strongly enhance erosional processes. For instance, in a study on a loamy sand soil, [Bodí et al. \(2012\)](#) found a more than 20-fold increase in sediment yield when hydrophobic (WDPT >3600 s) compared to the wettable counterpart (WDPT = 0 s). For soil erosion it is often difficult to differentiate the effects of SWR from other co-varying factors. In particular, the existence or lack of a vegetation and litter cover strongly affects whether SWR will have an effect on soil erosion. In this context, the spatial distribution of SWR can also be important for its potential to induce soil erosion, meaning that although SWR can induce high runoff locally, the effects at larger scales can be strongly attenuated ([Coelho et al., 2004](#)). Likewise, an enhanced aggregate stability often found in water repellent soil can increase the resistance to erosion.

Microbial activity and carbon sequestration

In contrast to wettable soils, water repellent soils are often characterized by less available water, which in turn is often more heterogeneously distributed in the soil matrix, leading to the formation of persisting dry domains. Soil microbial activity is strongly governed by the availability of water, so it can be reduced or even completely absent in such dry soil domains, meaning that residing SOM may be effectively protected. Beyond these effects, which are generally relevant at the profile scale, the previously described impact of SWR on water film thickness and continuity (heterogeneity) can have a strong impact on microbial activity. As shown by [Or et al. \(2007\)](#), diffusion rates of extracellular enzymes to SOM and of dissolved substrates back to the microbial cells are proportional to the thickness of the water film surrounding soil particles. The reduction in diffusion can result in a physical separation of microorganisms from substrates and nutrients and may cause dormancy and long-term starvation ([Kieft et al., 1993](#)). In summary, SWR can contribute to the formation of biologically non-preferred soil domains at different spatial scales, where microbial decomposition processes are reduced or even temporarily interrupted.

Plant growth

SWR-induced reduction in water infiltration and formation of preferential flow paths can significantly affect soil water recharge and the amount of plant-available water in deeper soil layers ([Lozano-Parra et al., 2016](#)), with important effects on irrigation efficiency ([Olorunfemi, 2014](#)). A study by [Ursino and Rulli \(2010\)](#), investigating the effect of water availability on vegetation patterns, suggested that post-fire vegetation patterns can be significantly affected by SWR. In this context, it is important to note that a change in vegetation structure may have substantial consequences for SWR itself, because increasing soil dryness can favor drought resistant vegetation, whose litter is typically rich in oils and waxes ([Doerr et al., 2000](#)). Over the long-term, this can further enhance SWR.

Moisture-affected dynamics of soil water repellency

Wetting of water repellent soil is governed by at least two mechanisms. First, increasing moisture content decreases the CA. This arises because already wetted surfaces (water menisci or water-film coated soil particles) exhibit a zero-CA to the contact area of a water drop placed onto the moist surface. Hence, the observed CA is a mixed signal resulting from the fractional area of either water repellent (grain) or wettable (water) surfaces, according to the Cassie-Baxter composite surface model ([Or and Tuller, 2022](#)). Secondly, beside the “pure” water content effect, a reorganization of the molecular structure at the solid interfaces in contact with water takes place (see [Fig. 4](#)).

During rainfall the wetting resistance of the soil surface can be reduced when the volumetric water content θ reaches a threshold, called the critical water content θ_{crit} ([Dekker and Ritsema, 1994](#)). Values of θ_{crit} can range from 2% by volume for a dune sand to about 40% by volume for a SOM-rich silt loam soil. [Fig. 9](#) shows soil wettability in terms of CA as a function of θ and capillary

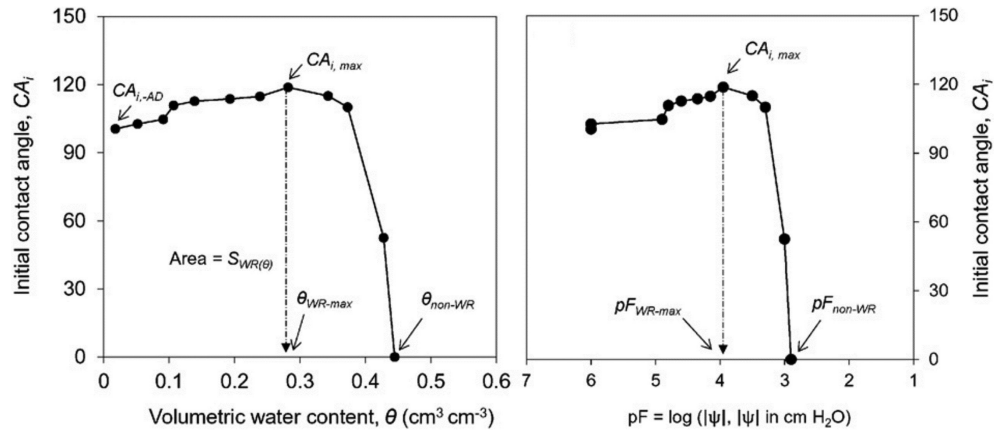


Fig. 9 Schematic diagram showing the initial SDM CA as a function of volumetric water content θ (left) and the pF-value (right). Area under the curve ($S_{WR(\theta)}$), the θ at the maximum SWR (θ_{WR-max}), the θ where SWR ceased (θ_{non-WR}), the maximum CA ($CA_{i,max}$), the pF at the maximum SWR (pF_{WR-max}) and the pF where SWR ceased (pF_{non-WR}). Figure taken from Wijewardana NS, Müller K, Moldrup P, Clothier B, Komatsu T, Hiradate S, de Jonge LW, Kawamoto K (2016) Soil-water repellency characteristic curves for soil profiles with organic carbon gradients. *Geoderma* **264**: 150–159.

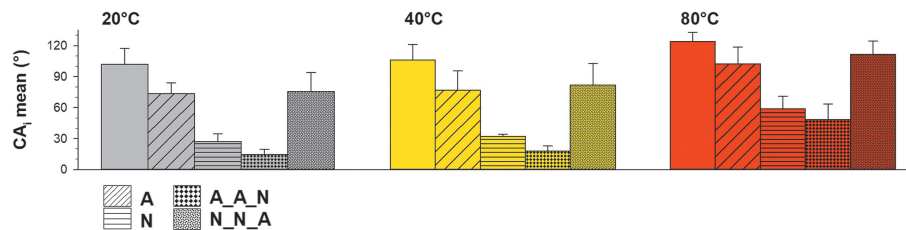


Fig. 10 Initial sessile drop CA for Ah horizons of sandy beach forest soils initially treated at 20, 40, and 80 °C. Letter 'A' stands for normal air-drying after water addition, 'N' means shock-freezing of wet soil material with subsequent freeze-drying. Repeated letters mean a series of treatments, e.g., A_A_N denotes repeated air-drying, rewetting and finally shock-freezing. Figure by S.K. Woche.

pressure (pF values) over a wide moisture range. It shows a decrease in CA at pF values larger than permanent wilting point ($>pF\ 4.2$) and a CA maximum at pF 4.2, which might be explained by reduced surface energy due to adsorbed water vapor as reported by Goebel et al. (2004). The CA-decrease at higher moisture contents could be caused by the formation of water films and water menisci as discussed above.

Structural reorganization of the multi-component interface, either caused by contact with water or by temperature, is shown in Fig. 10. Both treatments affect the wetting properties in the CA-sensitive interphase (see Fig. 5), which is a reversible process.

Displayed for each variant are the mean CA of five forest soils ($n = 50$) after the first wetting–drying cycle (A = air-drying after wetting, N = shock-freezing in liquid N₂ and freeze-drying after wetting) and after three repeated wetting–drying cycles, with finally a reversed drying procedure (A_A_N). Initial CA were similar for the 20 °C and 40 °C pretreatment before wetting–drying cycles followed, while samples pretreated at 80 °C had a higher CA level for the initial treatment and the following cycles. Air-drying always led to larger CA than shock-freezing/freeze-drying. Dynamic CA fluctuations reflect temperature and wetting effects in accordance with the interface model shown in Fig. 5. Recently, Woche et al. (2017) confirmed a close connection between the molecular setup in the outermost interface of soil particles by comparing XPS data to determine surface chemical composition with sessile drop CA to quantify the degree of SWR. Finally, Bachmann et al. (2021) concluded that shock-freezing/freeze-drying allows a direct access to the ambient wetting properties of the solid–water interface which might support the next steps in hydraulic model development.

Modelling

Hydraulic models of SWR are being developed from phenomenological approaches (Dekker and Ritsema, 1994), complex models based on the Richardson-Richards equation (Deurer and Bachmann, 2007) or pore-scale modelling in the rhizosphere (Benard et al., 2018). Model calculations based on the Lattice-Boltzmann simulation technique have shown a good agreement between pore-scale simulations of a sessile drop CA (Kang et al., 2018) and common approaches to account for the effect of roughness (Wenzel effect) and chemically composite surfaces (Cassie-Baxter effect; Bachmann and McHale, 2009). It is essential to unravel the question of when SWR will occur or disappear and the governing factors for dynamically changing wettability. Mao et al. (2019) discussed in their review paper possible mechanisms for the occurrence and persistence of SWR from nano to ecosystem scale. They

concluded that understanding the molecular setup at soil particle surfaces is essential to understand SWR as the basis for various model concepts able to distinguish clearly between processes occurring in wettable and water repellent soil.

Dealing with soil water repellency

Different techniques have been developed to mitigate the effects of SWR. Wetting agents are the most effective method to reduce or eliminate SWR, at least for some time. Generally, wetting agents are differentiated into three classes of surfactants, i.e. nonionic surfactants, anionic surfactants and block polymer nonionic surfactants, whereas nonionic surfactants are most widely used in agricultural production (Olorunfemi, 2014). Several mechanisms are associated with surfactants. The surfactant lowers the surface tension of the water, which reduces the solid-liquid CA, i.e., the soil becomes more wettable (Letey, 2005). In turn, reducing the surface tension of water may lower the capillary forces, which potentially decrease the infiltration rate when the soil is wettable or only slightly water repellent. Further, surfactant molecules are adsorbed by the soil as the solution moves downward.

Surfactants generally have a molecular structure consisting of a polar (hydrophilic) end and a non-polar (hydrophobic) end. In the soil, the hydrophobic end adsorbs to water repellent coatings via hydrophobic adsorption mechanisms, whereas the hydrophilic part of the molecule remains at the coating surface. Adsorption of the surfactant molecule converts a water-repellent soil to a wettable soil upon drying. However, as the surfactant is adsorbed from the solution, the surface tension of the solution increases again and wettability decreases. Therefore, surfactants are most effective in treating SWR when it is associated with the surface layer of the soil (Letey, 2005). Converting the uppermost soil layer to a wettable one, e.g. by surfactants, acts as if the level of water ponding is increasing with the depth of the wettable layer, which improves infiltration into deeper, still repellent layers. Although wetting agents are effective in improving water infiltration, they are expensive and only temporarily effective. Therefore, surfactants are most commonly applied in highly managed golf course soils, but their use in agriculture is increasing, particularly for higher value horticulture.

Mitigating SWR includes watering practices to keep the water content above the critical value and to reduce water loss due to evaporation, i.e., via mulch layers on the soil surface. Furthermore, tillage may lead to abrasion of hydrophobic organic coatings and might improve wettability, as can adding hydrophilic clay into the surface layer of a water-repellent sand. Conversely, shifts to reduced tillage may decrease severe soil water repellency because of the impacts to soil biology and organic matter decomposition dynamics. An extended overview on this topic is provided by Ritsema and Dekker (2003).

Conclusions and outlook

Soil water repellency has been described for decades. It can be considered as a phenomenon that occurs worldwide and may affect specifically different ecosystem service functions, depending on the respective land use system and climatic conditions. A challenge of current SWR research is to finally develop a mechanistic understanding of bio-physical interactions, including models and concepts that are consistent on all spatial scales of interest, from catchment hydrology down to particle interfaces. With climate change and the increased use of greywater for irrigation, the prevalence of SWR and its impacts, can be expected to worsen over time.

References

- Bachmann J and McHale G (2009) Superhydrophobic surfaces: A model approach to predict contact angle and surface energy of soil particles. *European Journal of Soil Science* 60: 420–430.
- Bachmann J, Horton R, van Der Ploeg RR, and Woche S (2000) Modified sessile drop method for assessing initial soil-water contact angle of sandy soil. *Soil Science Society of America Journal* 64: 564–567.
- Bachmann J, Woche SK, Goebel MO, Kirkham MB, and Horton R (2003) Extended methodology for determining wetting properties of porous media. *Water Resources Research* 39.
- Bachmann J, Söffker S, Sepelmeier N, Goebel M-O, and Woche SK (2021) The effect of temperature and wetting–drying cycles on soil wettability: Dynamic molecular restructuring processes at the solid–water–air interface. *European Journal of Soil Science* 72: 2180–2198.
- Benard P, Zarebanadkouki M, and Carminati A (2018) Impact of pore-scale wettability on rhizosphere rewetting. *Frontiers in Environmental Science* 6: 16.
- Blackwell PS (2000) Management of water repellency in Australia, and risks associated with preferential flow, pesticide concentration and leaching. *Journal of Hydrology* 231–232: 384–395.
- Bodi MB, Doerr SH, Cerdà A, and Mataix-Solera J (2012) Hydrological effects of a layer of vegetation ash on underlying wettable and water repellent soil. *Geoderma* 191: 14–23.
- Bos R, van der Mei HC, and Busscher HJ (1999) Physico-chemistry of initial microbial adhesive interactions—Its mechanisms and methods for study. *FEMS Microbiology Reviews* 23: 179–230.
- Churaev NV (2000) *Liquid and Vapour Flows in Porous Bodies: Surface Phenomena*. Routledge & CRC Press.
- Coelho C, Ferreira AJD, Boulet AK, and Keizer JJ (2004) Overland flow generation processes, erosion yields and solute loss following different intensity fires. *Quarterly Journal of Engineering Geology and Hydrogeology* 37: 233–240.
- de Jonge LW, Jacobsen OH, and Moldrup P (1999) Soil water repellency: Effects of water content, temperature, and particle size. *Soil Science Society of America Journal* 63: 437–442.
- DeBano LF (2000) Water repellency in soils: A historical overview. *Journal of Hydrology* 231: 4–32.
- Dekker LW and Ritsema CJ (1994) How water moves in a water repellent sandy soil. 1. Potential and actual water repellency. *Water Resources Research* 30: 2507–2517.
- Dekker LW and Ritsema CJ (1995) Finger-like wetting patterns in 2 water-repellent loam soils. *Journal of Environmental Quality* 24: 324–333.
- Deurer M and Bachmann J (2007) Modeling water movement in heterogeneous water-repellent soil: 2. A conceptual numerical simulation. *Vadose Zone Journal* 6: 446–457.
- Doerr SH, Shakesby RA, and Walsh RPD (2000) Soil water repellency: Its causes, characteristics and hydro-geomorphological significance. *Earth Science Reviews* 51: 33–65.

- Drehlich JW, Boinovich L, Chibowski E, Della Volpe C, Holysz L, Marmur A, and Siboni S (2019) Contact angles: History of over 200 years of open questions. *Surface Innovations* 8: 3–27.
- Ferguson GS and Whitesides GM (1992) Thermal Reconstruction of the Functionalized Interface of Polyethylene Carboxylic Acid and Its Derivatives. In: Schrader ME and Loeb GI (eds.) *Modern Approaches to Wettability: Theory and Applications*, pp. 143–177. US, Boston, MA: Springer.
- Gerke KM, Sidle RC, and Mallants D (2015) Preferential flow mechanisms identified from staining experiments in forested hillslopes. *Hydrological Processes* 29: 4562–4578.
- Goebel MO, Bachmann J, Woche SK, Fischer WR, and Horton R (2004) Water potential and aggregate size effects on contact angle and surface energy. *Soil Science Society of America Journal* 68: 383–393.
- Goebel MO, Bachmann J, Woche SK, and Fischer WR (2005) Soil wettability, aggregate stability, and the decomposition of soil organic matter. *Geoderma* 128: 80–93.
- Goebel M-O, Bachmann J, Reichstein M, Janssens IA, and Guggenberger G (2011) Soil water repellency and its implications for organic matter decomposition—Is there a link to extreme climatic events? *Global Change Biology* 17: 2640–2656.
- Good RJ (1992) Contact angle, wetting, and adhesion: A critical review. *Journal of Adhesion Science and Technology* 6: 1269–1302.
- Hallett PD, Nunan N, Douglas JT, and Young IM (2004) Millimeter-scale spatial variability in soil water sorptivity: Scale, surface elevation, and subcritical repellency effects. *Soil Science Society of America Journal* 68: 352–358.
- Hendrickx JMH, Dekker LW, and Boersma OH (1993) Unstable wetting fronts in water-repellent field soils. *Journal of Environmental Quality* 22: 109–118.
- Kang H, Lourenço SDN, and Yan WM (2018) Lattice Boltzmann simulation of droplet dynamics on granular surfaces with variable wettability. *Physical Review E* 98: 012902.
- Kieft TL, Amy PS, Brockman FJ, Fredrickson JK, Bjornstad BN, and Rosacker LL (1993) Microbial abundance and activities in relation to water potential in the Vadose zones of arid and semiarid sites. *Microbial Ecology* 26: 59–78.
- Lemmitz C, Kuhnert M, Bens O, Güntner A, Merz B, and Hüttl RF (2008) Spatial and temporal variations of actual soil water repellency and their influence on surface runoff. *Hydrological Processes* 22: 1976–1984.
- Letey J (2005) Water-Repellent Soils. In: Hillel D, Hatfield JH, Powlson DS, Rosenzweig C, Scow KM, Singer MJ, and Sparks DL (eds.) *Encyclopedia of Soils in the Environment*. Elsevier/Academic Press.
- Lewin M, Mey-Marom A, and Frank R (2005) Surface free energies of polymeric materials, additives and minerals. *Polymers for Advanced Technologies* 16: 429–441.
- Lozano-Parra J, van Schaik NLMB, Schnabel S, and Gomez-Gutierrez A (2016) Soil moisture dynamics at high temporal resolution in a semiarid Mediterranean watershed with scattered tree cover. *Hydrological Processes* 30: 1155–1170.
- Mao J, Nierop KGJ, Dekker SC, Dekker LW, and Chen B (2019) Understanding the mechanisms of soil water repellency from nanoscale to ecosystem scale: A review. *Journal of Soils and Sediments* 19: 171–185.
- Mataix-Solera J and Doerr SH (2004) Hydrophobicity and aggregate stability in calcareous topsoils from fire-affected pine forests in southeastern Spain. *Geoderma* 118: 77–88.
- Moradi AB, Carminati A, Lamparter A, Woche SK, Bachmann J, Vetterlein D, Vogel H-J, and Oswald SE (2012) Is the rhizosphere temporarily water repellent? *Vadose Zone Journal* 11.
- Muehl GJH, Ruehlmann J, Goebel M-O, and Bachmann J (2012) Application of confocal laser scanning microscopy (CLSM) to visualize the effect of porous media wettability on unsaturated pore water configuration. *Journal of Soils and Sediments* 12: 75–85.
- Olorunfemi I (2014) Soil hydrophobicity: An overview. *Journal of Scientific Research and Reports* 3: 1003–1037.
- Oostindie K, Dekker LW, Wesseling JG, and Ritsema CJ (2008) Soil surfactant stops water repellency and preferential flow paths. *Soil Use and Management* 24: 409–415.
- Or D and Tuller 2022 (This edition of the Encyclopedia of Soils in the Environment)
- Or D, Smets BF, Wraith JM, Dechesne A, and Friedman SP (2007) Physical constraints affecting bacterial habitats and activity in unsaturated porous media—A review. *Advances in Water Resources* 30: 1505–1527.
- Piccolo A and Mbagwu JSC (1999) Role of hydrophobic components of soil organic matter in soil aggregate stability. *Soil Science Society of America Journal* 63: 1801–1810.
- Ritsema CJ and Dekker LW (eds.) (2003) *Soil Water Repellency; Occurrence, Consequences, and Amelioration*, p. 352. Elsevier.
- Shakesby RA, Doerr SH, and Walsh RPD (2000) The erosional impact of soil hydrophobicity: Current problems and future research directions. *Journal of Hydrology* 231–232: 178–191.
- Shokri N, Lehmann P, and Or D (2008) Effects of hydrophobic layers on evaporation from porous media. *Geophysical Research Letters* 35: L19407.
- Shokri N, Lehmann P, and Or D (2009) Characteristics of evaporation from partially wettable porous media: Evaporation from wettable porous media. *Water Resources Research* 45.
- Siebold A, Walliser A, Nardin M, Oppliger M, and Schultz J (1997) Capillary rise for thermodynamic characterization of solid particle surface. *Journal of Colloid and Interface Science* 186: 60–70.
- Smettem KRJ, Rye C, Henry DJ, Sochacki SJ, and Harper RJ (2021) Soil water repellency and the five spheres of influence: A review of mechanisms, measurement and ecological implications. *Science of the Total Environment* 787.
- Tschapek M (1984) Criteria for determining the hydrophilicity-hydrophobicity of soils. *Zeitschrift für Pflanzenernährung und Bodenkunde* 147: 137–149.
- Ursino N and Rulli MC (2010) Combined effect of fire and water scarcity on vegetation patterns in arid lands. *Ecological Modelling* 221: 2353–2362.
- van Oss CJ, Chaudhury MK, and Good RJ (1988) Interfacial Lifshitz-van der Waals and polar interactions in macroscopic systems. *Chemical Reviews* 88: 927–941.
- Varela ME, Benito E, and de Blas E (2005) Impact of wildfires on surface water repellency in soils of northwest Spain. *Hydrological Processes* 19: 3649–3657.
- Wallis MG and Horne DJ (1992) Soil Water Repellency. In: Stewart BA (ed.) *Advances in Soil Science. Advances in Soil Science*, vol. 20, pp. 91–146. New York, NY: Springer.
- Wang Z and Wallach R (2021) Water infiltration into subcritical water-repellent soils with time-dependent contact angle. *Journal of Hydrology* 595: 126044.
- Washburn EW (1921) The Dynamics of capillary flow. *Physics Review* 17: 273–283.
- Wilhelmy L (1863) Über die Abhängigkeit der Capillaritäts-Constanten des Alkohols von Substanz und Gestalt des benetzten festen Körpers. *Annals of Physical Chemistry* 195: 177–217.
- Woche SK, Goebel M-O, Mikutta R, Schurig C, Kaestner M, Guggenberger G, and Bachmann J (2017) Soil wettability can be explained by the chemical composition of particle interfaces—An XPS study. *Scientific Reports* 7: 42877. EP.
- Young T (1805) An essay on the cohesion of fluids. *Philosophical Transactions of the Royal Society* 95: 65–87.
- Zisman WA (1964) Relation of the equilibrium contact angle to liquid and solid constitution. In: *Contact Angle, Wettability, and Adhesion, Advances in Chemistry*, pp. 1–51. American Chemical Society.

Aeration

Rodney T Venterea^a and Dennis E Rolston^b, ^aU.S. Department of Agriculture, Agricultural Research Service, Soil and Water Management Research Unit, St. Paul, MN, United States; ^bDepartment of Land, Air and Water Resources, University of California, Davis, CA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	216
Soil air content and composition	217
The dissolved 'atmosphere'	218
Soil gas-phase dynamics	219
Respiration and oxygen uptake	219
Heterogeneity and variability	219
Other gas production processes	220
Gas exchange mechanisms	220
Diffusion	220
Convection	221
Aeration requirements	221
Plants	221
Remediation of contaminated soils	222
Measurement of soil gas chemical composition	222
Summary	223
References	223

Abstract

Soil aeration affects plant growth, nutrient cycling, greenhouse gas emissions, soil remediation, and soil morphology. In a general sense, soil aeration is the interchange of gases between the atmosphere and the earth, but often refers more specifically to oxygen status which is largely affected by respiration of plant roots and soil microbes. Soil-atmosphere exchange of gases also results from other biogeochemical reactions that consume or produce gases and physical processes of diffusion and convection which are influenced by soil texture, structure, water content, and temperature. This chapter provides an overview of the key processes affecting and affected by soil aeration.

Key points

- Soil aeration influences soil fertility and plant growth and also affects numerous biochemical processes that have diverse ecological impacts.
- 'Aeration' may refer to a wide range of gaseous species that can be present in soil or may be used more specifically to refer to oxygen status.
- Soil-atmosphere gas exchange results from biogeochemical reactions that consume or produce gas concentration gradients that drive diffusion.
- Gas transport processes are influenced by soil physical properties including texture, structure, water content and temperature.
- Dissolved gases play a critical role in regulating soil microbial and chemical activities.
- This chapter provides an overview of the key processes affecting and affected by soil aeration.

Introduction

Soil aeration status is a critical factor influencing soil fertility and plant growth (Ben-Noah and Friedman, 2018) and affecting numerous soil biochemical processes. Many of the soil processes affected by aeration status have important, and diverse, ecological impacts including, but not limited to, greenhouse gas emissions (Wang et al., 2021), tropospheric ozone (O₃) production (Lu et al., 2021), stratospheric O₃ depletion (Ravishankara et al., 2009), and mercury speciation (Xu et al., 2019). Aeration is also an important factor in remediation of contaminated soil (Fragkou et al., 2021). Over decadal or longer time scales, aeration also impacts soil genesis and morphology which can impact shorter-term processes such as nutrient cycling (Gasparatos et al., 2019).

In a general sense, aeration is the interchange of gases between the atmosphere and the earth, and the term 'aeration' may be used in referring to a wide range of gaseous species that can be present, and move in or out, of soil. Gases expected to be present in soil air include not only those that predominate in earth's lower atmosphere, i.e., dinitrogen (N₂), oxygen (O₂) and argon, but a range of

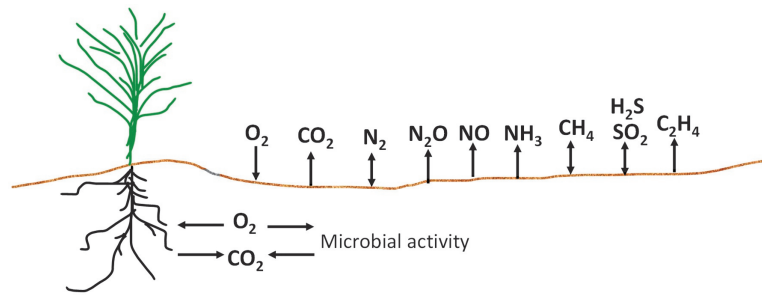


Fig. 1 Schematic diagram indicating the general flow directions of important gases within soil.

other 'biogenic' gases produced by microbial and plant processes within soil which are present in much greater concentrations than in the atmosphere. Thus, one way to characterize a soil's aeration status is by comparing the composition and concentrations of the gases in soil pores to that of the above-ground atmosphere (Fig. 1).

In soil, we are often concerned with O_2 status because it has a critical impact on a range of biological processes, both plant and microbial. In general, O_2 moves from the atmosphere into soil and is consumed by respiration which in turn results in the production of carbon dioxide (CO_2). Thus, soil O_2 and CO_2 concentrations often trend in opposite directions relative to atmospheric levels. In addition to respiration, other microbiological processes including decomposition, nitrification, and denitrification, as well as chemical reduction/oxidation reactions can also affect, and be affected by, soil aeration status. When O_2 becomes depleted, this can alter some of predominant biochemical and chemical processes and pathways and this may in turn alter the production of other gases. For this reason, O_2 can be considered as a 'master regulator' of soil biochemistry and therefore the term 'aeration' is often used specifically to refer to a soil's O_2 status.

The production or consumption of O_2 , CO_2 and other gases in soil creates concentration gradients that drive gas movement via diffusion within the soil and flux across the soil-atmosphere interface. Pressure gradients arising from natural barometric variations or from incorporated materials may also drive gas movement via convection. The physical processes of diffusion and convection are influenced by soil physical properties including texture, structure, water content and temperature. Human activities can result in the accidental or intentional introduction of gases in the soil profile such as fumigants, anhydrous ammonia, pesticides, and various organic chemical contaminants, for example, gasoline constituents with high vapor pressures such as benzene, toluene and xylene (Rolston, 2002), which can also create concentration and pressure gradients that drive soil gas exchange.

This chapter provides an overview of the key processes affecting and affected by soil aeration. Soil aeration has been reviewed previously with a focus on aeration as it relates to plants (Glinski and Stepniewski, 1985) and more recently by Ben-Noah and Friedman (2018). The early 20th century assessment of soil aeration by Russell and Appleyard (1915) also provides valuable historical context for the role of aeration in relation to soil biological processes. These references and others cited below are recommended as additional references.

Soil air content and composition

Soil is a multi-phase system comprised of varying proportions of gas, liquid and solid materials. Therefore, the relative proportions of each of these phases are important properties representing the state of the soil system at a given time and location. The term 'soil air' is often used to refer to the soil gas phase, although the composition of soil gas is often much different than what is normally referred to as atmospheric 'air.' The relative amount of air contained in a soil on a volume basis, or the soil-air content (ϵ , $m^3 \text{ air } m^{-3} \text{ soil}$), is directly related to its bulk density (ρ , $Mg \text{ soil } m^{-3} \text{ soil}$) and volumetric water content (θ , $m^3 \text{ H}_2\text{O } m^{-3} \text{ soil}$). To estimate ϵ under a given set of conditions, the soil's total porosity (ϕ), which may be occupied by water and air in different proportions, is first determined from $\phi = 1 - \frac{\rho}{\rho_p}$, where ρ_p is the particle density generally assumed to be $2.65 \text{ Mg } m^{-3}$ for mineral soils (Jury and Horton, 2004). This allows the determination of ϵ from $\epsilon = \phi - \theta$. Since ρ of natural soil varies from approximately $1.0 \text{ Mg } m^{-3}$ to $1.8 \text{ Mg } m^{-3}$, ϕ varies between 0.3 and 0.6, meaning that the total pore space represents 30–60% of the total soil volume. Thus, under dry soil conditions, ϵ can vary over this same range, but will decrease below this range as soil-water content increases. Another index that is often used as an indicator of soil-water content, and as a proxy for soil aeration status, is water-filled pore space (WFPS) which represents the fraction, or percentage, of the total pore space occupied by water, calculated from $\frac{\theta}{\phi}$ (Linn and Doran, 1984). This calculation can also be convenient for expressing soil-air content where the percentage of the total pore space occupied by air, or air-filled pore space (%AFPS) is given by $100 - \%WFPS$ or $100 (1 - \frac{\theta}{\phi})$. These relationships can be used to determine, for example, in a soil with $\rho = 1.2 \text{ Mg } m^{-3}$, the pore space would be equally occupied by water and air ($WFPS = AFPS = 50\%$) when $\theta = 0.274$.

The composition of soil air depends on the relative magnitude of the sources and sinks of the various gas components and on the rate of interchange between soil air and atmospheric air. If a soil were completely 'aerated' the concentrations of gases in soil air would be similar to that in the atmosphere. Under most conditions, O_2 concentrations in soil air will be somewhat below that in the ambient atmosphere ($\sim 20\%$ by volume) while CO_2 concentrations will be relatively elevated.

Under some conditions, O_2 concentrations can be completely depleted. Conditions where O_2 is absent are referred to as 'anoxic' and processes taking place in the absence of oxygen are commonly referred to as 'anaerobic' processes. Under many conditions, soil

profiles are neither fully aerated nor fully anoxic but are somewhere between these extremes (Scott, 2000). In addition to being needed to estimate the soil-air content, different indices of soil-water content (e.g., WPFS) are often used as proxies for O₂ availability because it is often difficult to directly measure soil O₂ concentrations.

The major constituent of the atmosphere (N₂) does have both a biological sink (biological N fixation) and source (denitrification) in soil. However, due to its high background concentration in the atmosphere (~78%), N₂ concentrations in soil air are particularly difficult to measure. Methods to determine soil air N₂ concentrations or rates of soil-to-atmosphere N₂ flux driven by denitrification have been developed (Yang and Silver, 2012; Butterbach-Bahl et al., 2002; Burgin and Groffman, 2012). Earlier studies have detected substantial variation in N₂ concentrations that appeared to be affected by rainfall events that may have trapped gases rising from below, dissolved or displaced N₂ in the soil air (Hinkle, 1994).

Among the biogenic trace gases, CO₂ concentrations in soil air can be as high as 10–40 times that in the atmosphere (Chen et al., 2005) which is currently 412 ppm or 0.0412% by volume. Depending on the particular soil system and its land use, the soil air is also expected to contain varying amounts of other gases including ammonia (NH₃), nitric oxide (NO), nitrous oxide (N₂O), nitrous acid (HONO), methane (CH₄), hydrogen sulfide (H₂S), and ethylene (C₂H₄) all of which may be present in amounts that far exceed their concentrations in the atmosphere. These gases result from numerous microbial or plant processes that cannot be fully reviewed here but are covered to some degree in later sections.

Another important difference between the above-ground atmosphere and soil air is that relative humidity in soil air rarely drops below 90% due to the presence of strong capillary and adhesive forces between soil surfaces and water molecules (Or et al., 2004), whereas in the atmosphere, relative humidity varies over a much broader range. One consequence is that water vapor concentrations in soil air are commonly in the range of 1–5% by volume, which is generally greater than or similar to concentrations of biogenic gases such as CO₂, CH₄ and N₂O. This can complicate the quantification and characterization of soil air chemical composition, depending on whether or not water vapor concentrations are taken into account. Water vapor transport in soil and across the soil-atmosphere interface can also contribute to convective movement of soil air in complex ways due to its coupling with energy and liquid water fluxes (Bittelli et al., 2008).

The dissolved ‘atmosphere’

In their treatise *The atmosphere of the soil: Its composition and the causes of variation*, Russell and Appleyard (1915) stated “Thus it appears that there are two atmospheres in the soil: one present as free gas filling the pores, and practically as rich in oxygen as ordinary air, the other dissolved in the surface films of water and other substances, almost devoid of oxygen and consisting mainly of carbon dioxide with some nitrogen”. Using Russell and Appleyard (1915) terminology, we can refer to compounds including not only O₂, CO₂ and N₂, but other biogenic or soluble gases such as N₂O, NO, NH₃ and CH₄, as occupying two distinct soil ‘atmospheres.’ While the soil atmosphere comprised of gas-filled pores is of critical importance in regulating the transport of gases within the soil and in exchange with the above-ground atmosphere, the atmosphere comprised of gases dissolved in soil water is of critical importance because it is the medium for biological and chemical activity. Although our understanding of the dissolved atmosphere has been expanded considerably since 1915, the most conspicuous feature is the same as in the gas-phase, i.e., O₂ tends to be depleted and CO₂ increased in soil solution relative to water in equilibrium with the above-ground atmosphere. This effect may be further enhanced because the rate of diffusion in liquid is substantially less than in the gas phase, with diffusion coefficients generally differing by a factor of 10,000 (Table 1). Another important property affecting dissolved gas concentrations is the Henry’s

Table 1 Henry’s Law constants and diffusion coefficients for major atmospheric and trace gases in soils.

Gas	Henry’s Law constant at 25 °C (K_H^{25}) ^a $\frac{\text{mol}}{\text{L} \cdot \text{atm}}$	Henry’s Law temperature dependency constant (TDC) ^b K	Diffusion coefficient in water (D_{water}) ^c $\frac{\text{cm}^2}{\text{s}}$	Diffusion coefficient in air (D_{air}) ^c $\frac{\text{cm}^2}{\text{s}}$
Low Henry’s Law gases (least liquid phase partitioning)				
N ₂	0.0006	1300	0.000020	0.195
O ₂	0.0013	1500	0.000031	0.221
CH ₄	0.0014	1600	0.000021	0.211
Intermediate Henry’s Law gases				
NO	0.019	1400	0.000023	0.203
N ₂ O	0.025	2600	0.000018	0.160
CO ₂	0.035	2600	0.000021	0.161
High Henry’s Law gases (greatest liquid phase partitioning)				
HONO	50	4200	0.000018	0.160
NH ₃	60	4900	0.000022	0.231

^aHenry’s law constant for solubility in water at 25 °C (K_H^{25} , $\frac{\text{mol}}{\text{L} \cdot \text{atm}}$). From NIST Chemistry WebBook, SRD 69, <https://webbook.nist.gov/chemistry/>.

^bTemperature dependence constant (TDC) to determine Henry’s Law constant at any temperature (T, K) using: $K_H^T = K_H^{25} \exp(TDC * (T^{-1} - 298.15^{-1}))$. From NIST Chemistry WebBook, SRD 69, <https://webbook.nist.gov/chemistry/>.

^cDiffusion coefficients in water (D_{water} , cm² s⁻¹) and air (D_{air} , cm² s⁻¹) determined using the U.S. EPA On-line Tools for Site Assessment Calculation, <https://www3.epa.gov/ceampub/learn2model/part-two/onsite/estdiffusion-ext.html>.

Law constant (K_H) which describes the extent to which a compound will partition between gas and liquid phases under equilibrium conditions. For example, the K_H values in Table 1 can be used to estimate the dissolved concentration (C_d , mol L⁻¹) based on the gas-phase partial pressure (P_g , atm) using

$$C_d = P_g K_H \quad (1)$$

The main gases present in soil fall into three categories with respect to their tendency to partition to the liquid phase, with N₂, O₂ and CH₄ having K_H values that are approximately 10 times less than N₂O, NO and CO₂, which are in turn approximately 1000 times less than HONO and NH₃ (Table 1).

Soil gas-phase dynamics

Respiration and oxygen uptake

Respiration carried out by plant roots and soil microbes are commonly referred to as autotrophic and heterotrophic respiration, respectively, though it should be noted that some soil microbes including nitrifying bacteria are considered autotrophic because they utilize inorganic compounds as their source of carbon for biosynthesis. Rates of O₂ consumption and CO₂ production are directly related to the rate of plant and microbial growth, which in turn is related to several environmental factors including air and soil temperature, substrate availability, and soil moisture (Ben-Noah and Friedman, 2018; Glinski and Stepniewski, 1985). Most notably, autotrophic and heterotrophic respiration both tend to increase with increasing temperature.

Under aerobic conditions and when carbohydrates are the predominant carbon source, the amounts of CO₂ produced and O₂ consumed on a molar basis sometimes tend to be about equal, and therefore the ratio between them (CO₂ consumed per O₂ produced), which is commonly referred to as the respiratory quotient (RQ), tends to be approximately 1.0. An estimate of the RQ can be used, for example, to estimate O₂ consumption when only CO₂ respiration data are available, which is often the case because the latter is usually more readily measured. However, deviation of RQ away from 1.0 will tend to occur when portions of the soil volume become anoxic, when carbon sources other than carbohydrates are being metabolized, or when evolved CO₂ becomes converted to soluble carbonates under alkaline conditions (Angert et al., 2015; Glinski and Stepniewski, 1985; Ben-Noah and Friedman, 2018). Under these conditions, estimating O₂ based on an assumed RQ value of 1.0 can lead to substantial errors.

Heterogeneity and variability

Because soils are multi-phasic and physically, chemically and biologically heterogeneous, the composition of soil air can differ depending on the spatial scale at which it is characterized. On the scale of the soil profile (cm to m), concentrations of O₂ in soil layers near the surface (soil-atmosphere interface) will be closer to ambient levels and generally tend to decrease with depth. However, substantial spatial variation may also occur laterally and at small (<1 mm) spatial scales due to the presence of high biological activity and rapid O₂ consumption occurring close to incorporated fertilizers or manure (Nguyen et al., 2017), naturally occurring organic materials undergoing decomposition (Parkin et al., 1987), or due to active root O₂ uptake. Soil-water content, texture and structure affect the ability of gases to move through soil pores and can further limit the rates of O₂ diffusion into such regions of high O₂ consumption. Further restriction to gas transport can be presented by the presence of mucilage layers around plant roots (Ben-Noah and Friedman, 2018). These factors can result in the interior of large soil aggregates becoming anoxic even though O₂ levels in the larger inter-aggregate pore spaces may be relatively high (Renault and Stengel, 1994; Ebrahimi and Or, 2015; Højberg et al., 1994). The spatial distributions described above for O₂ tend to be reversed for biogenic gases, i.e., gas concentrations tend to be smaller close to the soil-atmosphere interface and to increase with depth and within soil aggregates. Soil aggregate size will interact with hydration cycles and carbon availability to determine O₂ availability and therefore net production of biogenic gases such as N₂O and CH₄ inside of soil aggregates (Ebrahimi and Or, 2018).

In addition to small-scale spatial variations, substantial temporal variations in overall soil-gas composition and in the concentrations of individual gases can occur, for example as a result of diel and seasonal changes in radiation and temperature which affect microbial and/or plant activities. At a given location, soil O₂ uptake rates, responding to changes in autotrophic or heterotrophic respiration, tend to change (i) at hourly time scales over the course of the day in response to diel changes in temperature, (ii) over seasonal time scales in response to changes in temperature and shifts in vegetation activity, and (iii) may also vary from year to year due to inter-annual weather variation (Glinski and Stepniewski, 1985; Ben-Noah and Friedman, 2018). Soil concentrations of O₂ and CO₂ also depend upon the rate that the soil is able to exchange these gases with the atmosphere via diffusion and convection. The diurnal and annual variabilities in soil gas concentrations are generally much greater for clayey soils than for sandy soils due to the ability of sandy soils to transmit gases at a higher rate (larger soil-gas diffusion coefficients and air permeabilities) and maintain more constant concentrations.

Large variations in soil respiration have been measured across different ecosystems and climatic regions which are expected to affect the soil aeration status (Bond-Lamberty and Thomson, 2010; Jian et al., 2021). Assessing 245 plant species from across the globe, Han and Zhu (2021) found that species-specific root respiration showed 60-fold variation and was positively correlated with root N concentration and negatively correlated with root tissue density, with deciduous having significantly greater root respiration than evergreen species. In managed agricultural systems, activities such as tillage (Venterea et al., 2006) and fertilizer amendment (Nguyen et al., 2017) can substantially affect soil respiration rates and O₂ availability.

Because both heterotrophic and autotrophic respiration tend to increase with temperature, there is a concern that increasing temperatures associated with greenhouse gas-driven climate change will force increased soil respiration and concomitant losses of soil carbon to the atmosphere. However, soil respiration is considered to be the least well-constrained component of the terrestrial carbon cycle due to large uncertainties which complicate the modeling of soil respiration and its response to temperature and other factors across ecosystems and climatic regions (Jian et al., 2021). For example, studies have found that plant respiration responses to temperature were less pronounced when plants are allowed to acclimate to increasing temperatures (Bryla et al., 2001) and that root and rhizosphere respiration were more temperature-sensitive than microbial respiration alone, especially in colder regions (Li et al., 2020). Other related concerns include effects of altered snowfall in temperate regions (Aanderud et al., 2013) and soil warming in high-latitude regions (Elberling and Brandt, 2003) on release of CO₂ due to altered soil respiration and/or physically trapped CO₂.

Other gas production processes

The majority of gas production in soil results from oxidation-reduction (or redox) reactions (Glinski and Stepniewski, 1985; Hillel, 2004). These processes are generally connected with the transfer of electrons from soil organic matter (or organic contaminants) to oxidized inorganic compounds catalyzed by enzymes produced by soil microorganisms. Some soil microbes, such as nitrifying bacteria, are lithotrophs that obtain energy by transferring electrons among different inorganic compounds. For well-aerated conditions (aerobic), O₂ is the electron acceptor. When O₂ becomes limiting (anaerobic), other substances will accept electrons or be reduced. Examples of compounds that can be reduced under anaerobic conditions are nitrate (denitrification), manganic manganese, ferric iron, sulfate, and perchlorate (a natural and anthropogenic contaminant). In waterlogged or very wet soils and sediments that are unable to transmit O₂ sufficiently fast through the profile, O₂ is the first to disappear followed by nitrate and then sequentially by the reduction of manganese, iron, and sulfate. Perchlorate is reduced at about the same point as nitrate. Several toxic organic chemicals also undergo redox reactions that greatly affect the kind and toxicity of the reaction products.

Soil O₂ availability can be a critical variable with regard to the production of other gases in soil. As O₂ levels decrease and other compounds serve as electron acceptors, the types of gases generated from the predominant redox reactions also tend to change. This can result in important shifts in the composition of gases emitted to the atmosphere and the net greenhouse gas impacts of those emissions. These effects are particularly evident with respect to soil carbon and nitrogen cycling processes because they affect not only CO₂ but more potent greenhouse gases including methane (CH₄) and nitrous oxide (N₂O). As O₂ levels decrease, processes that consume CH₄ tend to decrease, while processes that produce CH₄ are promoted. The amounts of N₂O produced from both nitrification and denitrification also tend to increase as O₂ levels decrease (Venterea, 2007). Soil O₂ levels can also affect the production of gases that are not greenhouse gases but have other ecological impacts, including nitric oxide (NO) produced during nitrification or denitrification, and hydrogen sulfide produced when sulfate is reduced. Some gases produced in soils are not necessarily associated with redox reactions. When fertilizer materials such as urea and ammonium salts are applied to the soil, reactions can occur, particularly in soils with a pH greater than about 7.5 or 8, that produces ammonia gas (NH₃). Fairly large emission of NH₃ can occur, for instance, if urea is applied to the soil surface since urea hydrolysis results in a large increase in pH near the fertilizer granules, with NH₃ being emitted. The deeper the material is placed, the less NH₃ will be emitted. Ammonia gas may also be produced in soil from the incorporation of animal waste.

In addition to O₂, gases including hydrocarbons, N₂O, NO, CH₄, and some sulfur gases may move from the atmosphere into the soil and be consumed by biological processes. Thus, the soil may act as a significant sink for atmospheric gases under some circumstances.

Some biogenic gases, including CO₂, HONO and NH₃, participate in acid/base reactions that alter their molecular form in a manner that affects their phase stability. For both CO₂, which exists in equilibrium with water as carbonic acid (H₂CO₃), and for HONO, at higher pH, as the acidic species tend to become deprotonated, their ionic forms are increasingly soluble and therefore there is decreased partitioning of the protonated forms to the gas phase. In contrast, partitioning of NH₃ to the gas phase increases at increased pH due to decreased protonation of NH₄⁺ to favor more NH₃ in solution. In this case, the amount of NH₃ gas emitted depends not only on acid/base (NH₃/NH₄⁺) reactions, and therefore on pH, but also depends on soil surface reactions, i.e., sorption or cation exchange of NH₄⁺. Thus, for example, in soils with similar pH and similar total NH₃ + NH₄⁺ levels, soils having greater NH₄⁺ sorption capacity will have a smaller proportion of the dissolved N in the form of NH₃ (Venterea et al., 2015) and therefore less likelihood of emitting gaseous NH₃.

Gas exchange mechanisms

Diffusion

Diffusion is considered to be the principal mechanism driving the exchange of gases between the soil and atmosphere. The diffusion velocities of gas mixtures in soil are related to each other in a complex manner dependent upon the mole fraction of each gas, the molar fluxes of each gas, and the binary diffusion coefficient of each gas pair. General equations for steady transport of a multicomponent mixture of gases have been developed based on gas kinetic theory (Curtiss and Hirschfelder, 1949). If gravity effects are ignored or diffusion occurs only horizontally, the well-known Stefan-Maxwell equations provide the theoretical framework for diffusion of gases in soils. Fick's law is generally applicable for only a few special cases (Jaynes and Rogowski, 1983). One of these cases is for the diffusion of a trace gas in a binary mixture, meaning that the mole fraction of the trace gas is small. Since the binary diffusion coefficients of N₂-air and O₂-air are very similar and CO₂ may be considered to exist in trace

amounts, the diffusion of the two major gases associated with aeration may come under this case. Diffusion of some of the other gases existing in soil may not meet this criterion, however (Rolston and Moldrup, 2002).

Assuming that the special case conditions are met, Fick's law is given by:

$$\frac{M_g}{At} = f_{g,d} = -D_p \frac{dC_g}{dx} \quad (2)$$

where M_g is the amount of gas diffusing (kg gas), A is the cross-sectional area of the soil (m^2 soil), t is time (s), $f_{g,d}$ is the gas flux density ($kg \text{ gas } m^{-2} \text{ soil } s^{-1}$) due to molecular diffusion, C_g is concentration in the gaseous phase ($kg \text{ gas } m^{-3} \text{ soil air}$), x is distance (m soil), and D_p is the soil-gas diffusion coefficient ($m^3 \text{ soil air } m^{-1} \text{ soil } s^{-1}$).

The soil-gas diffusion coefficient is the main variable controlling the degree of soil aeration. It is largest when the soil is dry and approaches zero as the soil becomes very wet or saturated. The D_p has been related both empirically and theoretically to the soil-air content by a number of authors (Moldrup et al., 2013; Ben-Noah and Friedman, 2018). Since discrepancies between measured soil-gas diffusion coefficients stemming from Eq. (1) and those calculated from the many empirical equations occur, it is often desirable to measure the soil-gas diffusion coefficient for particular situations. Laboratory methods using soil cores and field methods for measuring soil-gas diffusion coefficients have been developed (Rolston and Moldrup, 2002).

Convection

In addition to molecular diffusion processes, soil gases may also exchange with the atmosphere through convection (advection). Convective flow of gases means that the whole air parcel is moving through soil pores due to a pressure difference (gradient) between the soil and the atmosphere. Pressure gradients may develop due to barometric pressure changes in the atmosphere; wind blowing across the soil surface or against a hill or other landscape feature; infiltration of rainfall or irrigation water into the soil, soil-water redistribution, and evaporation; temperature differences across the upper part of the soil profile; and density differences due to high concentrations of gases that have densities much different than air (Scott, 2000; Jury and Horton, 2004).

Convective gas flow is described by a form of Darcy's Law

$$f_{g,c} = -\frac{K_a}{\mu} \left[\frac{dP}{dx} \right] \quad (3)$$

where $f_{g,c}$ is the flux density of gas due to convection ($m^3 \text{ gas } m^{-2} \text{ soil } s^{-1}$), K_a is the air permeability ($m^3 \text{ gas } m^{-1} \text{ soil}$), μ is the viscosity of the gas mixture (Pa s), P is pressure (Pa), and x is distance (m soil).

Air permeability is strongly influenced by soil-water content. Air permeability attains a maximal value for dry soil and decreases as the soil becomes wet until it reaches zero at saturation. This is caused by progressive blockage of the soil pores by water. Several field and laboratory methods for measuring K_a have been developed involving either steady-state or non-steady-state flow, though steady-state measurements of gas permeability are preferred. Most field methods are based on the same principles as the laboratory methods, i.e., they involve measuring the flow rate of air through a soil column under known pressure differences across the column.

Convective processes occur rapidly and sporadically. Thus, it is very difficult to observe, measure, and predict the gas exchange that occurs by convection. During infiltration into soil, there is ample evidence that air pressure increases ahead of the water wetting front, and convection of gases occurs. It is often assumed that diffusion is the predominant gas exchange process because it is operating continuously whereas convection occurs episodically (Ben-Noah and Friedman, 2018). Of the convective processes, rainfall and irrigation may contribute the most to aeration of soils, with estimates that rainfall may account for 7–9% of total aeration. Adequately quantifying these processes and being able to predict and model convective flow processes is a goal yet to be fully attained.

Aeration requirements

Plants

The plant response to inadequate aeration or lack of O_2 is highly dependent upon the plant species, stage of growth, and upon several soil and environmental conditions such as temperature, water relations, and occurrence of toxic byproducts of anaerobic conditions. The response of plants to inadequate aeration may be due to either direct or indirect effects. The direct effect is because of the lack of oxygen for root respiration. The indirect effect is due to changes in redox conditions that affect nutrient and water availability, soil pH, buildup of toxic concentrations of metabolites and metals, and the viability of pests and diseases (Glinski and Stepniewski, 1985; Ben-Noah and Friedman, 2018).

The direct effect of lack of O_2 or slow diffusion of O_2 to plant roots has been characterized by a measurement called the Oxygen Diffusion Rate (ODR). The ODR is measured by placing a cylindrical platinum electrode into the soil (Stolzy and Letey, 1964). Oxygen is reduced at the electrode surface and creates a current that can be measured. Diffusion of O_2 to the water covered electrode is meant to mimic the diffusion through the water film around roots. Many studies have attempted to relate the ODR to plant response (Glinski and Stepniewski, 1985; Ben-Noah and Friedman, 2018). ODR values smaller than about $0.2 \mu g \text{ cm}^{-2} \text{ min}^{-1}$ indicate potentially poor aeration for many plant species. Values larger than about $0.4 \mu g \text{ cm}^{-2} \text{ min}^{-1}$ are indicative of relatively good aeration for growth of most plants.

The plant symptoms of poor aeration are poor root growth or root death, negative geotropism of roots, depressed shoot growth, wilting, leaf senescence, abortion of flowers, and termination of shoot apex growth (Glinski and Stepniewski, 1985). Poor aeration may also result in decreased transpiration; accumulation of ethylene, ethanol, and other metabolites; and decreased uptake of nutrients, most notably nitrogen, potassium, and phosphorus. Substantial variation in the sensitivity of nutrient deficiency, and other negative impacts, to O₂ availability has been observed among different plant species (Ben-Noah and Friedman, 2018; Glinski and Stepniewski, 1985). Reduced conditions in soil may also result in increased solubility of some chemicals that are toxic to plants.

Remediation of contaminated soils

With the large use of chemicals in modern society, both inorganic and organic chemicals, either intentionally or accidentally, end up in soil as contaminants with varying degrees of toxicity (Gerstl et al., 1989). Large amounts of petroleum products end up contaminating soil, both as point and non-point sources of pollution. The hydrocarbons in crude petroleum include alkanes, cycloalkanes, aromatics, polycyclic aromatics, asphaltenes, and resins. In addition, chlorinated solvents, pesticides, detergents, metals, and other kinds of chemicals may pollute soil. Microbial processes in soil can degrade many of the organic compounds, but the biodegradability of various compounds is greatly influenced by their physical state and toxicity as well as soil environmental conditions including soil-water content, soil pH, temperature, levels of inorganic nutrients, levels of electron acceptors, and aeration (Baker and Herson, 1994). Organic contaminants provide a source of carbon to microorganisms, or they provide electrons, which the organisms can use to obtain energy. Metal contaminants may undergo redox reactions affecting their solubility and toxicity which also is dependent upon the aeration status of the soil.

In most cases, biodegradation of organic contaminants depends on the activities of aerobic organisms. Thus, the presence of an adequate supply of O₂ in the soil is essential for biodegradation or bioremediation to occur. In general, it takes 2–3 kg of O₂ to degrade 1 kg of petroleum hydrocarbon. On the other hand, some compounds like the highly chlorinated chemicals are not easily degraded and may be broken down more effectively under anaerobic conditions. Some chemicals, perchlorate for instance, are only broken down under anaerobic conditions. Table 2 gives a list of some chemicals indicating whether they are degradable by aerobic or anaerobic processes.

Measurement of soil gas chemical composition

Researchers have been measuring the composition of soil air for more than 100 years, initially focusing on O₂ and CO₂ (Russell and Appleyard, 1915). Early methods combined the installation of metal soil probes to varying depths that allowed for extraction of soil air samples that were then analyzed using the Haldane apparatus. More recently, soil probes have been developed that allow gas sample collection from soil depth intervals of 0.05–0.50 m while minimizing physical disturbance and accounting for dilution effects (Petersen, 2014). Vertical profiles of gas concentrations over smaller depth intervals (e.g., 20 mm) can be obtained in undisturbed soil cores using horizontally inserted needles as probes (Venterea and Rolston, 2002). Samples collected via probes can be analyzed by gas chromatography or chemiluminescence to detect a wide range of soil gas species including O₂, CO₂, NO, N₂O and CH₄ or can be equipped with in situ sensors that can provide semi-continuous measurement of CO₂ and O₂ without the need for sample extraction (Chen et al., 2005; Owens et al., 2016; Jarecke et al., 2016). Planar optode technology allows for time lapse visualizations of soil O₂ concentration dynamics in two dimensions (Nguyen et al., 2017). Microelectrodes for in situ measurement of O₂ concentrations at scales of <50 mm have been in use for decades (Lemon and Erickson, 1952) and improvements in the technology are still underway including for measurement of other gases including N₂O (Xing et al., 2021; Revsbech, 2021). Due to the specialized methods and equipment required to directly quantify soil-gas O₂ concentrations, soil-water content or other indices such as WFPS or relative diffusivity (Owens et al., 2016) are often used to infer soil O₂ status, although such inference does not account for other factors such as O₂ consumption or transport through water-filled pores which may vary independent of water content.

Table 2 Organic chemicals and their biodegradability.

<i>Chemical class</i>	<i>Examples</i>	<i>Biodegradability</i>
Aromatic hydrocarbons	Benzene, toluene	Aerobic and anaerobic
Ketones and esters	Acetone, methyl ethyl ketone	Aerobic and anaerobic
Petroleum hydrocarbons	Fuel oil	Aerobic
Chlorinated solvents	Trichloroethylene, perchloroethylene	Aerobic (methanotrophs), Anaerobic (reductive dechlorination)
Polyaromatic hydrocarbons	Anthracene, beno[a]pyrene, creosote	Aerobic
Polychlorinated biphenyls	Arochlors	Unclear
Organic cyanides		Aerobic

Adapted from Baker KH and Herson DS (1994) *Bioremediation*. New York: McGraw-Hill.

Summary

Aeration is the interchange of various gases between the atmosphere and soil and the various reactions that either consume or produce gases in the soil. The composition of soil air depends on the relative magnitude of both the sources and sinks of the various gas components, the interchange between soil air and atmospheric air, and the partitioning of the gases between the gaseous, liquid, and solid (mineral and organic matter) phases of the soil. The two major gases associated with aeration are oxygen (O_2) and carbon dioxide (CO_2) where O_2 moves from the atmosphere to soil and is consumed by plant roots and microorganisms and CO_2 moves from the soil, where it is produced by plant and microbial respiration, to the atmosphere. In addition to root and microbial respiration, O_2 may also be consumed by reaction with metals and other compounds. Soil O_2 availability can be a critical variable with regard to the production of other gases in soil including the potent greenhouse gases methane and nitrous oxide and gases that have important ecological impacts like nitric oxide and ammonia. The two transport mechanisms that result in aeration of soils are molecular diffusion and convection. Although convection can result in significant transport under certain situations, such as infiltration of rainfall or irrigation water into soil, diffusion is considered to be the dominant mechanism of exchange over the long term. The plant symptoms of poor aeration are poor root growth or death, negative geotropism of roots, depressed shoot growth, wilting, leaf senescence, abortion of flowers, and termination of shoot apex growth. Aeration is also important for the soil's ability to degrade pollutants. In most cases, biodegradation of organic contaminants depends on the activities of aerobic organisms. Thus, the presence of an adequate supply of O_2 in the soil is essential for biodegradation or bioremediation to occur. On the other hand, some chemicals will only degrade under anaerobic conditions. Knowledge of a chemical's biodegradability and whether the chemical will degrade under aerobic or anaerobic conditions plays a major role in design of effective bioremediation schemes.

References

- Aanderud ZT, Jones SE, Schoolmaster DR, Fierer N, and Lennon JT (2013) Sensitivity of soil respiration and microbial communities to altered snowfall. *Soil Biology and Biochemistry* 57: 217–227.
- Angert A, Yakir D, Rodeghiero M, Preisler Y, Davidson EA, and Weiner T (2015) Using O_2 to study the relationships between soil CO_2 efflux and soil respiration. *Biogeosciences* 12(7): 2089–2099.
- Baker KH and Herson DS (1994) *Bioremediation*. New York: McGraw-Hill.
- Ben-Noah I and Friedman SP (2018) Review and evaluation of root respiration and of natural and agricultural processes of soil aeration. *Vadose Zone Journal* 17(1): 170119.
- Bittelli M, Ventura F, Campbell GS, Snyder RL, Gallegati F, and Pisa PR (2008) Coupling of heat, water vapor, and liquid water fluxes to compute evaporation in bare soils. *Journal of Hydrology* 362(3): 191–205.
- Bond-Lamberty B and Thomson A (2010) A global database of soil respiration data. *Biogeosciences* 7(6): 1915–1926.
- Bryla DR, Bourma TJ, Hartmond U, and Eissenstat DM (2001) Influence of temperature and soil drying on respiration of individual roots in citrus: Integrating greenhouse observations into a predictive model for the field. *Plant, Cell & Environment* 24(8): 781–790.
- Burgin AJ and Groffman PM (2012) Soil O_2 controls denitrification rates and N_2O yield in a riparian wetland. *Journal of Geophysical Research – Biogeosciences* 117: G01010.
- Butterbach-Bahl K, Willibald G, and Papen H (2002) Soil core method for direct simultaneous determination of N_2 and N_2O emissions from forest soils. *Plant and Soil* 240(1): 105–116.
- Chen D, Molina JAE, Clapp CE, Venterea RT, and Palazzo AJ (2005) Corn root influence on automated measurement of soil carbon dioxide concentrations. *Soil Science* 170(10): 779–787.
- Curtiss CF and Hirschfelder JO (1949) Transport properties of multicomponent gas mixtures. *The Journal of Chemical Physics* 17(6): 550–555.
- Ebrahimi A and Or D (2015) Hydration and diffusion processes shape microbial community organization and function in model soil aggregates. *Water Resources Research* 51(12): 9804–9827.
- Ebrahimi A and Or D (2018) Dynamics of soil biogeochemical gas emissions shaped by remolded aggregate sizes and carbon configurations under hydration cycles. *Global Change Biology* 24(1): e378–e392.
- Eberling B and Brandt KK (2003) Uncoupling of microbial CO_2 production and release in frozen soil and its implications for field studies of arctic C cycling. *Soil Biology and Biochemistry* 35(2): 263–272.
- Fragkou E, Antoniou E, Daliakopoulos I, Manios T, Theodorakopoulou M, and Kalogerakis N (2021) In situ aerobic bioremediation of sediments polluted with petroleum hydrocarbons: A critical review. *Journal of Marine Science and Engineering* 9(9): 1003.
- Gasparatos D, Massas I, and Godelitsas A (2019) Fe-Mn concretions and nodules formation in redoximorphic soils and their role on soil phosphorus dynamics: Current knowledge and gaps. *Catena* 182: 104106.
- Gerstl Z, Chen Y, Mingelgrin U, and Yaron B (1989) *Toxic Organic Chemicals in Porous Media*. Berlin: Springer-Verlag.
- Glinski J and Stepniewski W (1985) *Soil Aeration and Its Role for Plants*. New York: CRC Press.
- Han M and Zhu B (2021) Linking root respiration to chemistry and morphology across species. *Global Change Biology* 27(1): 190–201.
- Hillel D (2004) *Introduction to Environmental Soil Physics*. New York: Elsevier Academic Press.
- Hinkle ME (1994) Environmental conditions affecting concentrations of He , CO_2 , O_2 and N_2 in soil gases. *Applied Geochemistry* 9(1): 53–63.
- Højberg O, Revsbech NP, and Tiedje JM (1994) Denitrification in soil aggregates analyzed with microsenors for nitrous oxide and oxygen. *Soil Science Society of America Journal* 58(6): 1691–1698.
- Jarecke KM, Loecke TD, and Burgin AJ (2016) Coupled soil oxygen and greenhouse gas dynamics under variable hydrology. *Soil Biology and Biochemistry* 95: 164–172.
- Jaynes DB and Rogowski AS (1983) Applicability of Fick's Law to Gas Diffusion. *Soil Science Society of America Journal* 47(3): 425–430.
- Jian J, Vargas R, Anderson-Teixeira KJ, Stell E, Herrmann V, Horn M, Kholod N, Manzoni J, Marchesi R, Paredes D, and Bond-Lamberty BP (2021) *A Global Database of Soil Respiration Data, Version 5.0*. ORNL Distributed Active Archive Center.
- Jury WA and Horton R (2004) *Soil Physics*. New York: John Wiley & Sons, Inc.
- Lemon E and Erickson A (1952) The measurement of oxygen diffusion in the soil with a platinum microelectrode. *Soil Science Society of America Journal* 16(2): 160–163.
- Li J, Pendall E, Dijkstra FA, and Nie M (2020) Root effects on the temperature sensitivity of soil respiration depend on climatic condition and ecosystem type. *Soil and Tillage Research* 199: 104574.
- Linn DM and Doran JW (1984) Effect of water-filled pore space on carbon dioxide and nitrous oxide production in tilled and nontilled soils. *Soil Science Society of America Journal* 48(6): 1267–1272.
- Lu X, Ye X, Zhou M, Zhao Y, Weng H, Kong H, Li K, Gao M, Zheng B, Lin J, Zhou F, Zhang Q, Wu D, Zhang L, and Zhang Y (2021) The underappreciated role of agricultural soil nitrogen oxide emissions in ozone pollution regulation in North China. *Nature Communications* 12(1): 5021.

- Moldrup P, Chamindu Deepagoda TKK, Hamamoto S, Komatsu T, Kawamoto K, Rolston DE, and de Jonge LW (2013) Structure-dependent water-induced linear reduction model for predicting gas diffusivity and tortuosity in repacked and intact soil. *Vadose Zone Journal* 12(3): pp. vzj2013.01.0026.
- Nguyen QV, Wu D, Kong X, Bol R, Petersen SO, Jensen LS, Liu S, Brüggemann N, Glud RN, Larsen M, and Bruun S (2017) Effects of cattle slurry and nitrification inhibitor application on spatial soil O₂ dynamics and N₂O production pathways. *Soil Biology and Biochemistry* 114: 200–209.
- Or D, Tuller M, and Wraith JM (2004) Soil water potential. In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*. New York: Academic Press.
- Owens J, Clough TJ, Laubach J, Hunt JE, Venterea RT, and Phillips RL (2016) Nitrous oxide fluxes, soil oxygen, and denitrification potential of urine- and non-urine-treated soil under different irrigation frequencies. *Journal of Environmental Quality* 45(4): 1169–1177.
- Parkin TB, Starr JL, and Meisinger JJ (1987) Influence of sample size on measurement of soil denitrification. *Soil Science Society of America Journal* 51(6): 1492–1501.
- Petersen SO (2014) Diffusion probe for gas sampling in undisturbed soil. *European Journal of Soil Science* 65(5): 663–671.
- Ravishankara AR, Daniel JS, and Portmann RW (2009) Nitrous Oxide (N₂O): The dominant ozone-depleting substance emitted in the 21st century. *Science* 326(5949): 123–125.
- Renault P and Stengel P (1994) Modeling oxygen diffusion in aggregated soils: I. Anaerobiosis inside the aggregates. *Soil Science Society of America Journal* 58(4): 1017–1023.
- Revsbech NP (2021) Simple sensors that work in diverse natural environments: The micro-Clark sensor and biosensor family. *Sensors and Actuators B: Chemical* 329: 129168.
- Rolston DE (2002) The soil gas phase: 4.1. Introduction. In: Dane JH and Topp GC (eds.) *Methods of Soil Analysis. Part 4. Physical Methods*. Madison, WI: Soil Science Society of America.
- Rolston DE and Moldrup P (2002) The soil gas phase: 4.3. Gas diffusivity. In: Dane JH and Topp GC (eds.) *Methods of Soil Analysis. Part 4. Physical Methods*. Madison, WI: Soil Science Society of America.
- Russell EJ and Appleyard A (1915) The atmosphere of the soil: Its composition and the causes of variation. *The Journal of Agricultural Science* 7(1): 1–48.
- Scott HD (2000) *Soil Physics: Agricultural and Environmental Applications*. Ames: Iowa State University Press.
- Stolzy LH and Letey J (1964) Characterizing soil oxygen conditions with a platinum microelectrode. In: Norman AG (ed.) *Advances in Agronomy*. Academic Press.
- Venterea RT (2007) Nitrite-driven nitrous oxide production under aerobic soil conditions: Kinetics and biochemical controls. *Global Change Biology* 13(8): 1798–1809.
- Venterea RT and Rolston DE (2002) Nitrogen oxide trace gas transport and transformation: 1. Evaluation of data from intact soil cores. *Soil Science* 167(1): 35–48.
- Venterea RT, Baker JM, Dolan MS, and Spokas KA (2006) Carbon and nitrogen storage are greater under biennial tillage in a Minnesota corn-soybean rotation. *Soil Science Society of America Journal* 70(5): 1752–1762.
- Venterea RT, Clough TJ, Coulter JA, Breuillin-Sessoms F, Wang P, and Sadowsky MJ (2015) Ammonium sorption and ammonia inhibition of nitrite-oxidizing bacteria explain contrasting soil N₂O production. *Scientific Reports* 5: 12153.
- Wang C, Ma X, Wang G, Li G, and Zhu K (2021) Implication of O₂ dynamics for both N₂O and CH₄ emissions from soil during biological soil disinfestation. *Scientific Reports* 11(1): 6590.
- Xing L, Qin W, Manevski K, Zhang Y, Hu C, Zhang L, Dong W, Wang Y, Li X, Gaudel G, and Qin S (2021) An improved microelectrode method reveals significant emission of nitrous oxide from the rhizosphere of a long-term fertilized soil in the North China Plain. *Science of the Total Environment* 783: 147011.
- Xu J, Buck M, Eklöf K, Ahmed OO, Schaefer JK, Bishop K, Skjölberg U, Björn E, Bertilsson S, and Bravo AG (2019) Mercury methylating microbial communities of boreal forest soils. *Scientific Reports* 9(1): 518.
- Yang WH and Silver WL (2012) Application of the N₂/Ar technique to measuring soil-atmosphere N₂ fluxes. *Rapid Communications in Mass Spectrometry* 26(4): 449–459.

Relevant websites

- https://en.wikipedia.org/wiki/Soil_gas—Wikipedia entry: Soil Gas.
- <https://webbook.nist.gov/chemistry/> - NIST Chemistry WebBook.
- <https://www3.epa.gov/ceampubl/learn2model/part-two/onsite/estdiffusion-ext.html>—EPA On-line Tools for Site Assessment Calculation.
- https://daac.ornl.gov/SOILS/guides/SRDB_V5.html—NASA Open Access Earth Science Data: A Global Database of Soil Respiration Data, Version 5.0.
- <https://www.youtube.com/watch?v=i6RldRQHlfi>—Lecture: Evaluation of Root Respiration and of Natural and Agricultural Processes of Soil Aeration by Ilan Ben-noah and Shmulik Friedman.

Capillarity

Dani Or^a and Markus Tuller^b, ^aDivision of Hydrologic Sciences, Desert Research Institute, Reno, NV, United States; ^bDepartment of Environmental Science, The University of Arizona, Tucson, AZ, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of D. Or, M. Tuller, CAPILLARITY, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 155-164, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00339-8>.

Introduction	225
A brief historical perspective on capillarity	226
The origins of capillarity in porous media	226
Surface tension	226
Contact angle formed between liquid and solid surfaces	228
Curved surfaces and capillarity	229
The capillary rise model	230
Capillarity in soils	230
Capillarity in angular pores	231
Capillary rise in angular tubes	232
Dynamic aspects of capillarity	233
Capillary rise dynamics	233
Dynamic contact angle	235
Heterogeneous surfaces and microscale hysteresis	236
Contact angle on chemically heterogeneous and rough surfaces	236
Capillarity and soil cohesion	237
Hysteresis	238
Summary	239
References	239

Abstract

Capillarity emerges from interactions between air-water interfaces and soil surface and pore spaces. The tendency of water to adhere and spread over wettable surfaces while minimizing air-water interfacial area gives rise to spontaneous capillary rise in soil pores and crevices and to water retention against forces of gravity. Capillarity affects soil aqueous phase configurations that determine hydraulic pathways for water flow in unsaturated soil towards plant roots and surface evaporation, and sustain gas and nutrient diffusion pathways for biological activity. Capillary effects vary with soil type (surface and pore space properties) and contribute to the soil matric potential (along with film adsorption).

Key points

- The historical context of the capillarity concepts is provided.
- The ingredients of capillarity and their liquid-solid interactions are reviewed.
- Equilibrium capillary rise in cylindrical and angular capillaries is quantified.
- Dynamic aspects of capillarity are presented.

Introduction

The coexistence of gaseous, liquid, and solid phases in soil pores gives rise to various interfacial phenomena that, for example, lead to spreading of liquid droplets on solid surfaces, liquid spontaneously rising in narrow soil pores, interfacial mechanical stresses that hold water-wet sand grains together, or the entrapment and retention of liquid in crevices. These processes are attributed to the phenomenon of capillarity that controls the amount and configuration of water retained in soil following wetting events. Capillarity also controls aqueous phase connectivity critical for diffusion of solutes and for biological activity in soils (e.g., bacterial cell dispersion and microbial accessibility to nutrients). The control over water flow and solute transport rates, aspects of biological activity and water availability for plants highlight the central role of capillarity in agriculture, ecology, environmental and climatic processes. Focusing on the soil aqueous phase, we will review the interfacial properties of water and interactions with the soil solids to develop a quantitative representation of capillarity in soil.

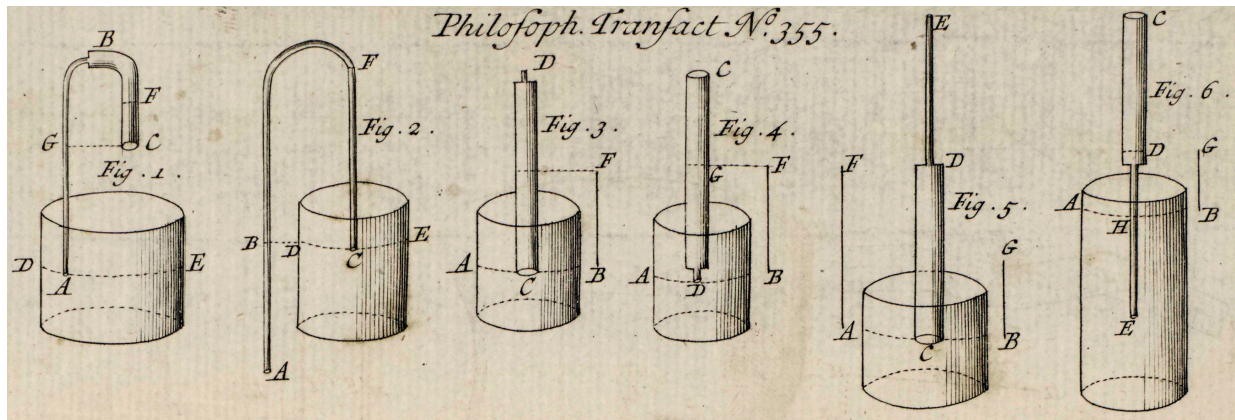


Fig. 1 Historical figures from the paper of Jurin, (1718). Note the perpetual capillary paradox described on the left panel (original Fig. 1). Reproduced from Jurin J (1718) II. An account of some experiments shown before the Royal Society; with an enquiry into the cause of the ascent and suspension of water in capillary tubes. *Philosophical Transactions of the Royal Society of London*, **30** (355): 739–747.

A brief historical perspective on capillarity

One of the earliest accounts of capillarity in a quantitative and now familiar way has been reported by the English scientist Jurin, (1718). His paper starts with posing an apparent paradox of perpetual motion in a capillary siphon tube pulling water from a reservoir and dripping it back to the same reservoir (Fig. 1). In his own words Jurin writes “Some days ago, a method was proposed to me by an ingenious friend, for making a perpetual motion, which seemed so plausible, and indeed so easily demonstrable from an observation by the late Mr. Hawksbee, said to be grounded upon experiment that, thou I am far from having any opinion of attempts of this nature, yet, I confess, I could not see why it should not succeed. “But as searches after things impossible in themselves are frequently observed to produce other discoveries unexpected by the inventor”. Jurin, (1718) correctly attributed the spontaneous rise of water into capillaries to “the attraction of periphery, or section of the surface of the tube, to which the upper surface of the water is contiguous and coheres”. About 100 years later, in 1805, two renowned scientists Simon Laplace in France and Thomas Young in England published independent studies on the phenomenon of a pressure jump across curved liquid-gas interfaces that are presented below as the Young-Laplace equation. In his essay on the cohesion of fluids, Young (1805) not only described the relations between curved interfaces and the onset of pressure in the liquid phase, but he also laid the foundation for the concepts of contact angle and wettability that we will explore later. In 1907, Edgar Buckingham introduced the concept of capillary potential for quantifying the attractions between water and soil solids and pores. Buckingham (1907, p. 29) wrote: “Then we may define ψ as the work required per centigram to pull water away from the mass of soil. We shall call ψ the capillary potential of the soil”. This definition of capillary potential remains a keystone in the quantitative representation of driving forces for flow in partially saturated porous media.

The origins of capillarity in porous media

The phenomenon of capillarity in porous media results from two primary opposing forces: liquid adhesion to solid surfaces, which tends to spread the liquid; and cohesive surface tension forces of liquids, which act to reduce liquid-gas interfacial area per unit volume of liquid. The resulting liquid-gas interface configuration under equilibrium reflects a balance between these forces. The phenomenon of capillarity is thus dependent on solid and liquid interfacial properties such as surface tension, contact angle, and solid surface roughness, and importantly, on the geometry of pore spaces as discussed below.

Surface tension

At the interface between water and other fluids (e.g., air) or soil solids, water molecules are exposed to different forces than are molecules within the bulk liquid. For example, water molecules in the bulk liquid are subjected to uniform cohesive forces due to formation of flickering (short-lived) hydrogen bonds with neighboring molecules on all sides. Water molecules at the air-water interface experience a net attraction into the liquid because of lower density of water molecules on the air side of the interface, with more hydrogen bonds being formed at the liquid side (Fig. 2). On average, water molecules at the surface form only three hydrogen bonds compared with the four formed in bulk water (Henn, 2003). The result is a membrane-like water surface that tends to contract and reduce the amount of its excess surface energy. Evidence suggests that water molecules are oriented with hydrogen atoms pointing into the vapor phase (Henn, 2003).

The surface tension reflects the amount of interfacial energy per unit area, or the energy required to bring molecules from the bulk liquid to increase the surface. Although the air-water surface may appear quiescent, it is a region experiencing violent collisions and heavy traffic from molecules condensing from the vapor phase or breaking loose (evaporating) from the surface with an average

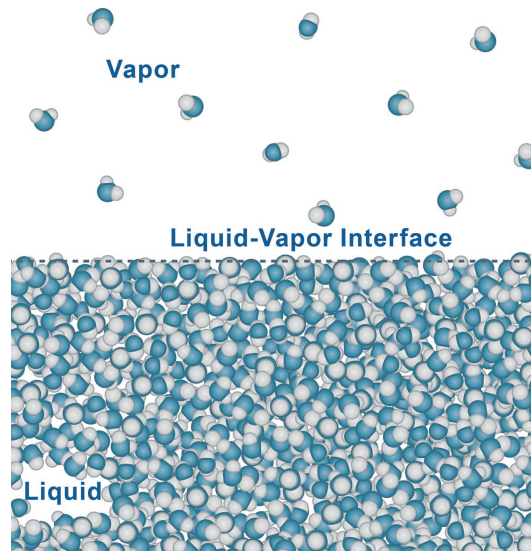


Fig. 2 The air-water interface where water molecules at the surface experience different forces than in bulk with net attraction inwards (this creates the analogy of tension). The excess free energy at the surface implies that a force (equal to the surface tension) is required to bring water molecules from the bulk to increase surface area.

residence time of a molecule at the surface of less than 0.1 ms (Adamson, 1990). Water molecules are also exchanged with the bulk water at high rates (defined by the temperature and self-diffusion).

Liquids vary in their surface tension σ (Table 1) based on intermolecular forces. Surface tension typically decreases (linearly) with increasing temperature. Increasing temperatures and thermal expansion reduce the density of the liquid and cohesive forces (e.g., fraction of intact hydrogen bonds) with the thermal changes affecting the vapor pressure above the surface and solubility.

Table 1 Liquid-vapor interfacial tensions for various liquids.

<i>Liquid</i>	<i>Temp.</i> [°C]	<i>Surface Tension</i> [mN m ⁻¹]
Water	20	72.94
	25	72.13
Methylene iodide	20	67.00
Glycerin	24	62.60
Ethylene glycol	25	47.30
Dimethyl sulfoxide	20	43.54
Propylene carbonate	20	41.10
1-Methyl naphthalene	20	38.70
Dimethyl aniline	20	36.56
Benzene	20	28.88
Toluene	20	28.52
Chloroform	25	26.67
Propionic acid	20	26.69
Butyric acid	20	26.51
Carbon tetrachloride	25	26.43
Butyl acetate	20	25.09
Diethylene glycol	20	30.90
Nonane	20	22.85
Methanol	20	22.50
Ethanol	20	22.39
Octane	20	21.62
Heptane	20	20.14
Ether	25	20.14
Perfluoromethylcyclohexane	20	15.70
Perfluoroheptane	20	13.19
Hydrogen sulfide	20	12.30
Perfluoropentane	20	9.89

Reproduced from Adamson AW (1990) *Physical Chemistry of Surfaces*, 5th edn. New York: John Wiley.

Soluble substances can increase or decrease surface tension. If the affinity of the solute molecules or ions to water molecules is greater than the affinity of the water molecules to one another, then the solute tends to be drawn into the solution and to cause an increase in surface tension (e.g., electrolytic solutes). For example, a 1% NaCl concentration increases the surface tension of an aqueous solution by 0.17 mN m^{-1} at 20°C . If, on the other hand, the attraction between water molecules is greater than their attraction to the solute molecules, then the latter tend to be relegated towards the surface, reducing its tension. That is the effect of many organic solutes, particularly detergents.

Contact angle formed between liquid and solid surfaces

When a liquid drop is placed on a solid surface, the angle formed between the solid-liquid (SL) interface and the liquid-gas (LG) interface (Fig. 3) is referred to as the equilibrium (or static) contact angle (γ). Two equivalent approaches have been used to describe the equilibrium contact angle on smooth and chemically homogeneous planar surfaces: (1) a force balance approach, and (2) an interfacial free-energy minimization. The force balance formulation considers interfacial tensions (σ_{ij}) as forces per unit length; hence the force balance at the contact line of a drop resting on a solid surface under equilibrium requires the vector sum of the forces acting to spread the drop (outward) to be equal to opposing surface tension and cohesion forces. The free-energy minimization approach regards interfacial tension as energy per unit area, and calculates changes in surface free energy (ΔF) due to infinitesimal displacement (ΔA):

$$\Delta F = \Delta A(\sigma_{SL} - \sigma_{GS}) + \Delta A \cos \gamma \sigma_{LG} \quad (1)$$

The result is identical whether considering the minimization of free energy, with $\Delta F/\Delta A = 0$, or taking a balance of forces tangential to the solid surface; both cases yield the Young equation:

$$\sigma_{LG} \cos \gamma + \sigma_{SL} - \sigma_{GS} = 0 \quad (2)$$

with L , G , and S indicating liquid, gas, and solid, respectively, and σ_{ij} the respective interfacial surface tensions. The equilibrium contact angle is therefore:

$$\cos \gamma = \frac{\sigma_{GS} - \sigma_{SL}}{\sigma_{LG}} \quad (3)$$

Liquids that are attracted to solid surfaces (adhesion) more strongly than to other liquid molecules exhibit a small contact angle, and the solid is said to be 'wetable' by the liquid (Fig. 3A). Conversely, when the attractive forces within the liquid are larger than the adhesive force with the solid surface, the liquid 'repels' the solid and γ is large (Fig. 3B). Fig. 4 illustrates differences in wettability of a silt soil (Bachmann et al., 2000). In Fig. 4B a water droplet is resting on a soil surface that was treated to become water-repellent ($\gamma = 70^\circ$). In contrast, Fig. 4A depicts a wettable soil surface. In general, the contact angle of water on clean glass, and presumably on most soil minerals, is small, and for mathematical convenience is often taken as $\gamma = 0^\circ$. Note that for most natural soil surfaces $\gamma > 0^\circ$ (for clean quartz $\gamma \approx 20^\circ$).

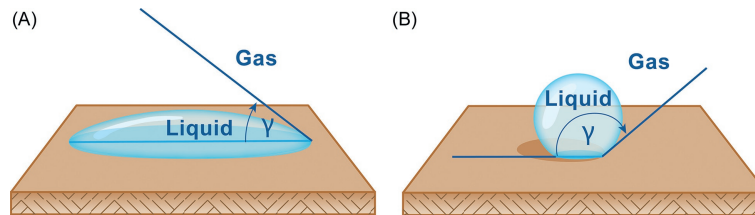


Fig. 3 Liquid-solid-gas contact angles: (A) hydrophilic surface ($\gamma < 90^\circ$) where liquid wets the surface; (B) hydrophobic surface ($\gamma > 90^\circ$) where liquid 'repels' the surface.

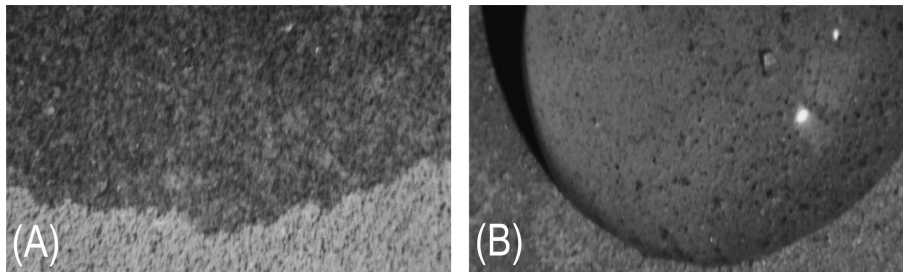


Fig. 4 (A) Wettable silt soil surface ($\gamma \sim 0^\circ$); (B) treated water-repellent silt soil surface ($\gamma = 70^\circ$). Reproduced from Bachmann J, Ellies A, and Hartge KH (2000) Development and application of a new sessile drop contact angle method to assess soil water repellency. *Journal of Hydrology*, 231–232: 66–75.

Curved surfaces and capillarity

When the forces that spread the liquid (adhesion and spreading on solids, or gas pressure within a bubble) are in balance with surface tension that tends to minimize interfacial area, the resulting liquid-gas interface is curved. In porous media, the liquid-gas interface shape reflects the interplay of forming a specific contact angle with solids on the one hand, and energetic constraints that minimize interfacial area within a pore. As shown independently by Laplace (1806) and Young (1805), a pressure difference forms across the curved interface, where the pressure at the concave side of the interface is greater by an amount that is dependent on the radius of the interface curvature and the surface tension of the liquid. For a hemispherical liquid-gas interface having radius of curvature R , the pressure difference is given by the Young-Laplace equation:

$$\Delta P = \frac{2\sigma}{R} \quad (4)$$

where $\Delta P = P_L - P_G$ when the interface curves into the gas (e.g., water droplet in air); or $\Delta P = P_G - P_L$ when the interface curves into the liquid (e.g., air bubble in water, water in a small glass tube). In many instances a bubble may not be spherical, or an element of liquid may be confined by irregular solid surfaces, resulting in two different radii of curvature such as water held in pendular rings between two spherical solid particles (Fig. 5). The Young-Laplace equation for this case is given by:

$$\Delta P = \sigma \left(\frac{1}{R_1} + \frac{1}{R_2} \right) \quad (5)$$

Note that this equation reduces to Eq. (4) for spherical geometry with $R_1 = R_2$, and the sign of R is negative for convex interfaces ($R_2 < 0$) and positive for concave interfaces ($R_1 > 0$). For an interface forming in a linear crevice or within a fracture, $R_2 \rightarrow \infty$, hence Eq. (5) reduces to: $\Delta P = \sigma/R_1$, where R_1 equals half the fracture aperture.

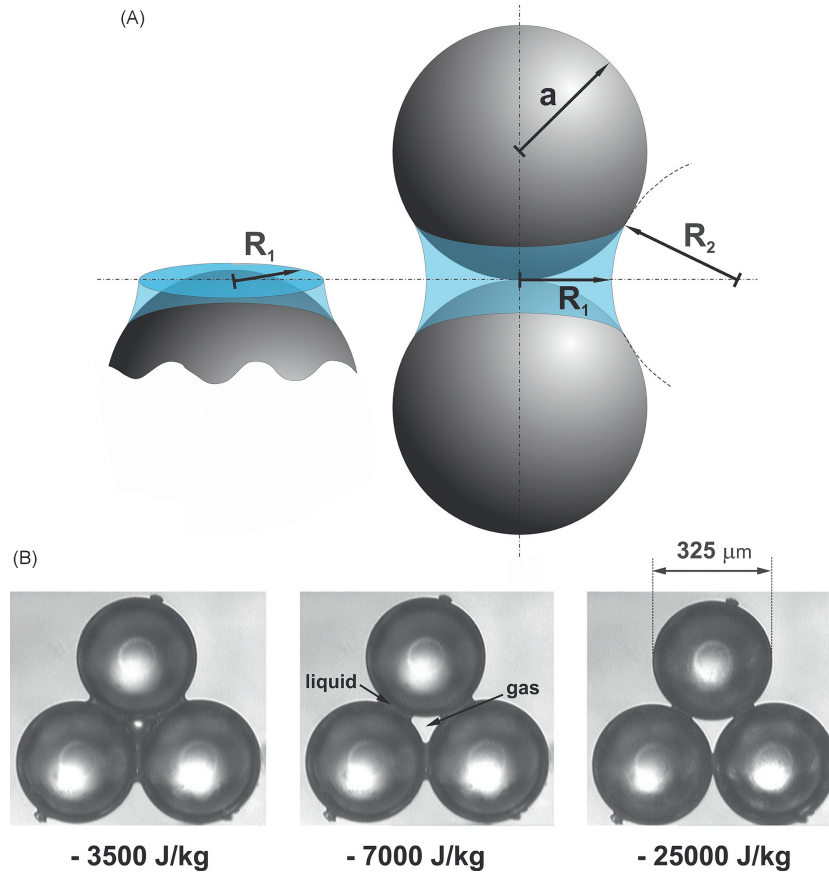


Fig. 5 (A) Radii of curvature and shape of water held in pendular space between two spherical grains. Note that for two equal spheres with radius a , the relationship between R_2 and R_1 is given as: $R_2 = R_1^2 [2(a - R_1)]$. (B) Water menisci held between three spherical glass beads at different capillary pressures.

The capillary rise model

When a cylindrical glass tube of small diameter (capillary) is dipped into free water, a meniscus forms in the tube owing to the contact angle between water and the tube walls, and minimum surface energy requirements. The smaller the tube radius, the larger the degree of curvature and the pressure difference across the air-water interface (Fig. 6). The pressure at the water side (P_w) is lower than atmospheric pressure (P_0). This pressure difference causes water to rise into the capillary until the upward capillary force is balanced by the weight of the water column. In a cylindrical tube, the radius of meniscus curvature (R) is related to the tube radius r by $R = r / \cos \gamma$; consequently, the equilibrium height of capillary rise in a cylindrical tube with contact angle γ is:

$$h = \frac{2\sigma \cos \gamma}{\rho_w g r} \quad (6)$$

where g is the acceleration due to gravity, and ρ_w is the liquid density. For water at 20°C in a glass capillary with $\gamma = 0^\circ$, the capillary rise equation simplifies to: $h(\text{mm}) = 15/r(\text{mm})$. It is instructive to introduce here the concept of capillary length λ_C , given as:

$$\lambda_C = \sqrt{\frac{\sigma}{\rho_w g}} \approx 3 \text{ mm} \quad (7)$$

This length is used as a yardstick to gauge the relative importance of capillarity relative to gravity for different liquids. For example, the capillary length for water defines a limit for rain drop sizes (that seldom exceed this diameter, Villermaux and Bossa, 2009). For capillary rise applications, a factor of 2 is added to the term under the radical ($\lambda_C \approx 4.2 \text{ mm}$ for water). The ratio of pore size or capillary radius r to the capillary length λ_C squared defines a dimensionless number known as the Bond number (B_O) (Or, 2008):

$$B_O = \left(\frac{r}{\lambda_C} \right)^2 \quad (8)$$

For example, $B_O = 1$ marks the condition where capillarity and gravity forces are equal (for a combination of capillary size and liquid). For water in soil pores with radii larger than 4.2 mm ($B_O > 1$) gravity overcomes capillary forces and we may observe gravity dominated flow (preferential flow).

Capillarity in soils

The complex geometry of soil pore space creates numerous combinations of interfaces, capillaries, and wedges in which water is retained, and results in a variety of air-water and solid-water configurations. Water is drawn into or held by these interstices in proportion to the resulting capillary forces. Here again, we may define a macroscopic capillary length that gauges the importance of capillarity and gravity via the use of the soil sorptivity function. In addition, water is adsorbed on to solid surfaces with considerable force at close distances. Due to practical limitations of present measurement methods, no distinction is made between the various mechanisms affecting water in porous matrices (i.e., capillarity and surface adsorption). Common conceptual models for water retention in porous media and matric potential rely on a simplified picture of soil pore space as a 'bundle of cylindrical capillaries'.

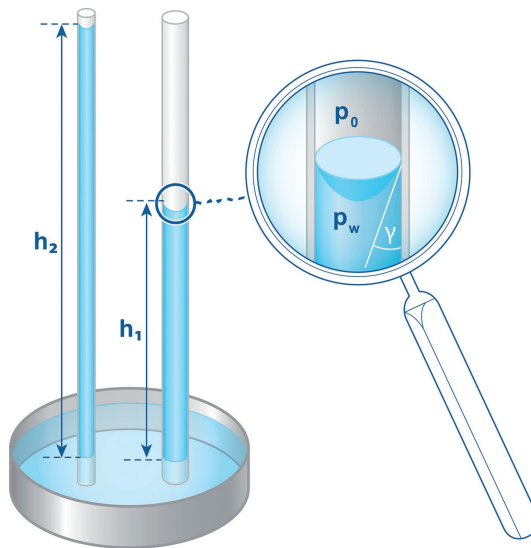


Fig. 6 Capillary rise in cylindrical tubes with different radii.

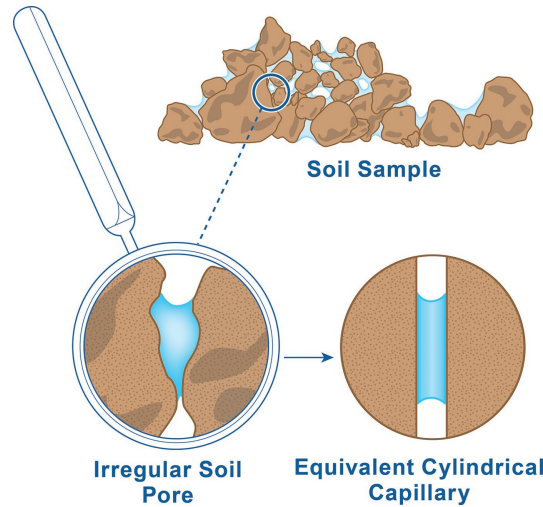


Fig. 7 Idealization of natural soil pore spaces as equivalent cylindrical capillaries.

The primary conceptual steps made in such models are illustrated in Fig. 7. The representation of soil pores as equivalent cylindrical capillaries greatly simplifies modeling and parameterization of soil pore space and relies heavily on the capillary rise equation (Eq. 6).

Capillarity in angular pores

Cursory inspection of scanning electron micrographs of soils and other natural porous media (Fig. 8) shows that pore spaces formed by aggregation of primary particles and mineral surfaces tend to be angular and slit-shaped, rarely resembling cylindrical tubes. Such observations and other shortcomings of the 'cylindrical capillary' model have led to the development of new models for capillarity in angular and slit-shaped pores. Capillarity in angular pores is quite different from the behavior in cylindrical pores with equivalent cross-sectional area. For example, when angular pores are drained, a fraction of the wetting phase (water) remains in the pore corners (Fig. 9A). This aspect of 'dual occupancy' of wetting and nonwetting phases, not possible in cylindrical tubes, more realistically represents liquid configurations and the mechanisms for maintaining aqueous phase continuity in porous media. Liquid-filled corners and crevices play an important role in displacement and transport processes in partially saturated porous media. For all (regular and irregular) polygons with n corners, the total water filled area (Aw_t) at a given matric potential is simply the sum of the water-filled areas in each corner (Fig. 9A). This sum is given by the simple equation (Tuller et al., 1999):

$$Aw_t = r(\mu)^2 F(\alpha) \quad (9)$$

With

$$F(\alpha) = \sum_{n=1}^i \left(\frac{1}{\tan(\frac{\alpha_i}{2})} - \frac{\pi(180 - \alpha_i)}{360} \right) \quad (10)$$

where μ is the matric potential and $F(\alpha)$ is a shape factor dependent on pore angularity (corner angles α_i) only.

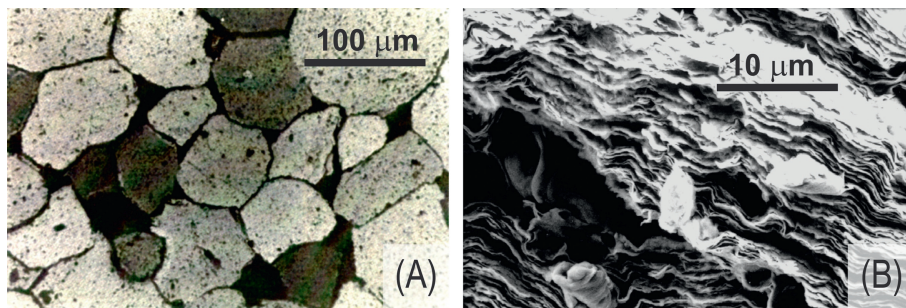


Fig. 8 (A) Thin section of Devonian sandstone, revealing angular pore space. (B) Scanning electron micrograph of calcium-saturated montmorillonite clay. (A) Reproduced from Roberts JN, Lawrence M, and Schwartz LM (1985) Grain consolidation and electrical conductivity in porous media. *Physical Review B* 31: 5990–5997.

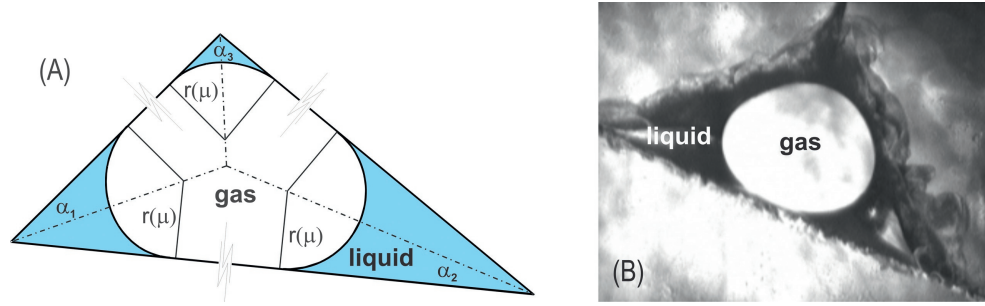


Fig. 9 (A) Dual-occupancy of wetting and nonwetting phases in triangular pores; (B) liquid-vapor interfacial configuration in a triangular glass pore (~ 2 mm).

In contrast to a piston-like filling or emptying of circular capillaries, angular pores undergo different filling stages and spontaneous displacement in the transition from dry to wet soil conditions (or vice versa). As the soil dries, the remaining water is retained in crevices and pore corners by capillary forces. When the water potential increases, the radii of curvatures in the corners increase until they make contact and form an inscribed circle (in the pore cross section). Subsequently, liquid spontaneously fills up the pore in a process termed pore snap-on. When an angular pore is drained, air invades the central pore region first and gradually displaces liquid to the corners. For completeness, one must also consider the role of liquid films due to adsorption to solid surfaces.

Capillary rise in angular tubes

Capillary rise in angular tubes is a simple extension of the capillary rise in cylindrical tubes presented above (Eq. 6). Following [Bico and Querre \(2002\)](#), the capillary force (F) exerted over the tube's perimeter P is balanced by the liquid weight (W) in a tube of cross-sectional area A (under equilibrium conditions), and this force balance is expressed as:

$$F = P\sigma \cos \gamma = W = \rho_w g h A \quad (11)$$

where g is the acceleration due to gravity, ρ_w is the liquid density, γ is the liquid-solid-gas contact angle, and h is the height of capillary rise.

For a rigorous account of the total weight of the liquid in the examples depicted in [Fig. 10](#) (square cross section with sides a , or equilateral triangular tube with sides a), we must account for the weight of liquid fingers over the main meniscus at h . The total weight of the liquid for the square tube is calculated as:

$$W_{\text{square}} = \rho_w g h a^2 + 4\rho_w g \int_h^\infty \left(1 - \frac{\pi}{4}\right) r^2(z) dz \quad (12)$$

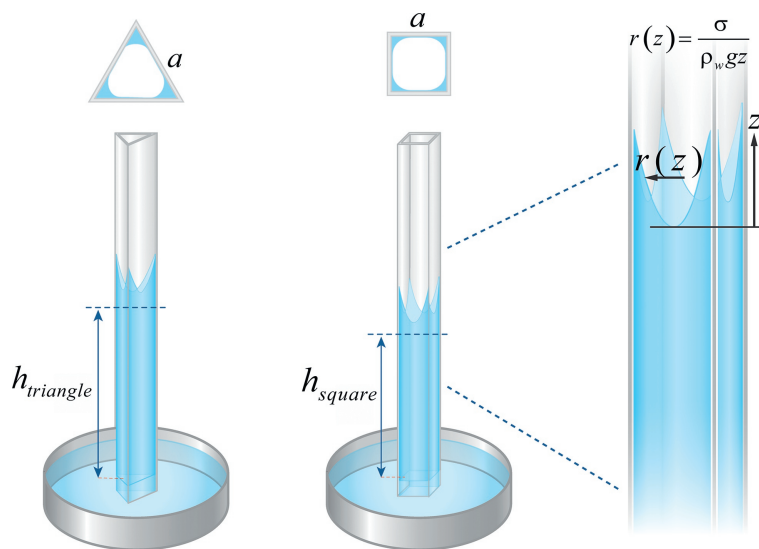


Fig. 10 Equilibrium capillary rise in angular tubes.

In equilibrium, the force balance (Eq. 11) yields the capillary rise height, h :

$$h_{\text{square}} = \frac{(2 + \sqrt{\pi}) \sigma \cos \gamma}{\rho_w g a} = \frac{3.77 \sigma \cos \gamma}{\rho_w g a} \quad (13)$$

For all practical purposes, however, the counterweight of liquid fingers over the main meniscus can be ignored (lowering the capillary height estimate by less than 6%) in favor of a much simpler approximation considering the first term on the RHS of Eq. (13) only. This simplification yields:

$$h_{\text{square}} \approx \frac{4 \sigma \cos \gamma}{\rho_w g a} \quad (14)$$

Similarly, for an equilateral triangular tube the capillary rise is approximately:

$$h_{\text{triangle}} \approx \frac{12 \sigma \cos \gamma}{\sqrt{3} \rho_w g a} \quad (15)$$

Dynamic aspects of capillarity

Capillary rise dynamics

The equilibrium height of fluid rise in a capillary (Eq. 6) does not contain any information regarding the rate of rise and the time scales for attaining equilibrium, which may be of significant importance for industrial and natural processes. Under certain conditions, we may apply a simple force balance between a driving capillary force F_σ :

$$F_\sigma = 2\pi R \sigma \cos \gamma \quad (16)$$

and a retarding viscous force F_η (assuming Poiseuille flow in a tube):

$$F_\eta = 8\pi\eta x \frac{dx}{dt} \quad (17)$$

to represent the rate of capillary flow into a horizontal cylindrical capillary. Additionally, we may introduce inertial forces according to:

$$m \frac{d^2 x}{dt^2} = F_\gamma - F_\mu \quad (18)$$

where m is the mass of the liquid in the capillary, x is distance, and t is time. Substitution of the forces (Eqs. 16 and 17) into Eq. (18) and integration (neglecting higher-order terms) yields the so-called Lucas-Washburn-Rideal (LWR) equation (Lucas, 1918; Washburn, 1921; Rideal, 1921):

$$x = \sqrt{\left(\frac{R \sigma \cos \gamma}{\eta} \right) t} \quad (19)$$

which describes the rate of liquid penetration into a horizontal capillary with the dependency of x on $\sqrt{t}$. It is interesting to note that Washburn's neglect of inertial effects and Rideal's truncation of higher-order terms (r^{-n} , $n > 2$) in his series solution yield the same solution (Eq. 19). Note that a similar dependency of x on $\sqrt{t}$ is observed for early stages of water imbibition into soil (during early infiltration where gravity is unimportant) and these relations are used to define soil sorptivity.

The analytical representation of dynamic capillary rise with gravity is mathematically challenging. Several simplified analytical solutions for the rate of capillary rise in vertical capillaries have been proposed, such as the following implicit solution:

$$\frac{\rho g R^2}{8\eta} t = z(t) - z_e \ln \left[1 - \frac{z(t)}{z_e} \right] \quad (20)$$

The solution diverges as $z(t)$ approaches the equilibrium capillary rise z_e (Eq. (6)). Another approximate solution has been proposed, based on the introduction of a retardation coefficient β :

$$z(t) = z_e \left(1 - \exp \left[-\frac{\sigma \cos \gamma}{\beta z_e} t \right] \right) \quad (21)$$

The solution converges to the equilibrium capillary rise z_e (Eq. 6) for long periods of time. Fig. 11A depicts comparison of Eqs. (20) and (21) with capillary-rise measurements of silicon oil (PDMS 10) in a glass capillary, with $r = 0.315 \text{ mm}$ ($\sigma = 20.1 \text{ mN m}^{-1}$; $\eta = 10 \text{ mPa}$; and $\rho = 0.935 \text{ kg m}^{-3}$) (Hamraoui and Nylander, 2002). The nondimensional retardation coefficient for water in glass capillaries ranges from $\beta = 0.5$ for large radii ($r > r_c$), representing friction dissipation due to contact line motion and contact angle adjustment, $\beta = 0.7$ for small radii ($r < r_c$) representing primarily viscous dissipation. The critical radius r_c is related to an interesting

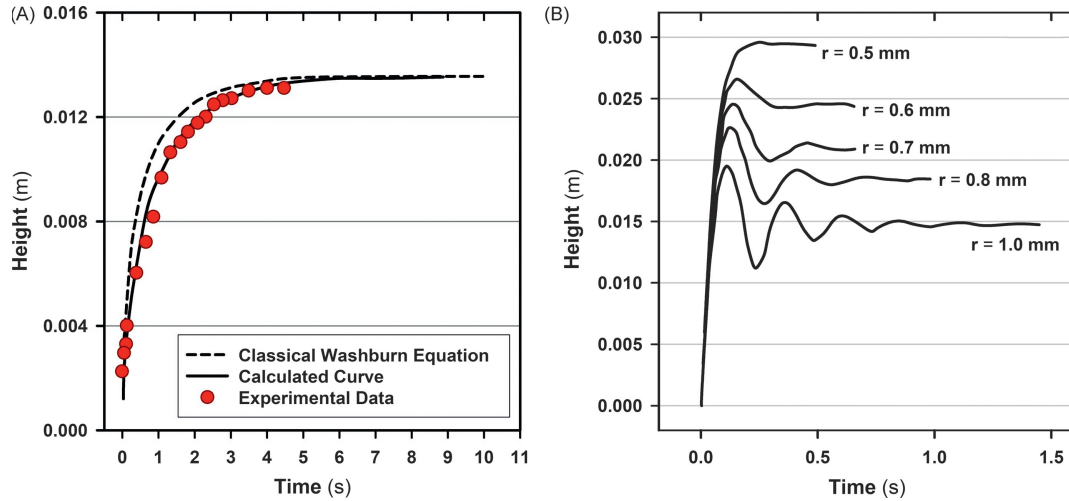


Fig. 11 (A) Comparison of measurements and theoretical models for capillary rise dynamics of silicon oil (PDMS 10) in glass capillary with $r = 0.315 \text{ mm}$ (calculated curve from Eq. (21); classic Washburn equation from Eq. (20)); (B) inertia-induced oscillations during capillary rise of water in different glass capillary sizes (numerical simulations). Note that inertial oscillations vanish for capillaries smaller than $r = 0.474 \text{ mm}$ according to Eq. (22). Adapted from Hamraoui A and Nylander T (2002) Analytical approach for the Lucas–Washburn equation. *Journal of Colloid and Interface Science* **250**: 415–421.

feature of capillary rise in the presence of gravity, namely inertia-induced oscillations in large capillaries (Quere et al., 1999), as depicted in Fig. 11B. The inertial oscillations disappear in capillaries of radii smaller than r_c :

$$r_c = \frac{2(\sigma \cos \theta \eta^2 \rho^2 g^3)^{1/5}}{\rho g} \quad (22)$$

A comprehensive overview of the transition from inertial to viscous limited capillary rise is given by Fries and Dreyer (2008) and summarized in Fig. 12. Initially, capillary rise is primarily inertial (there is not enough liquid to be retarded by viscous/friction forces) and the capillary rise is proportional to t (elapsed time). Gradually, viscous forces become dominant, and the capillary rise regime expressed by the LWR solution emerges with capillary rise rate proportional to $\sqrt{t}$. Interestingly, the meniscus velocity during

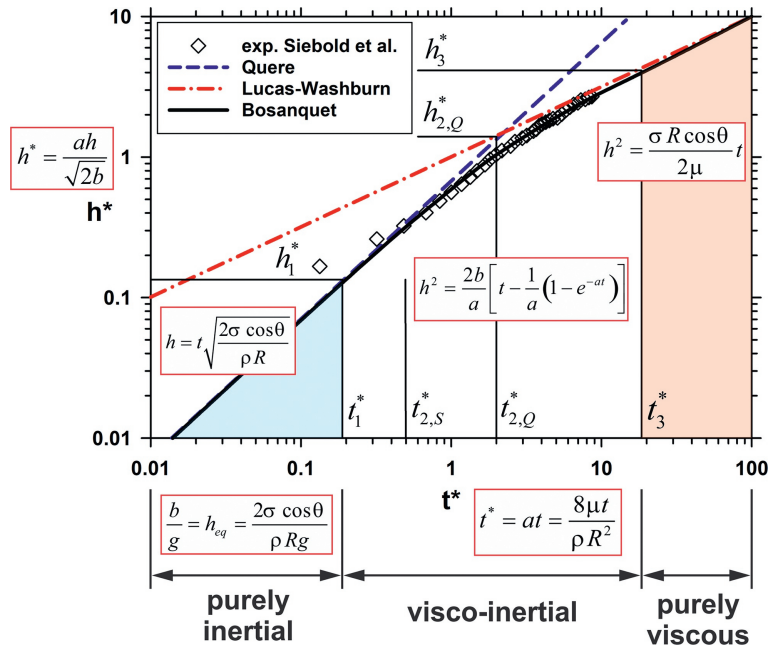


Fig. 12 Dimensionless representation of dynamic capillary rise regimes transitions from inertial to viscous capillary rise rates (and ultimately to equilibrium height, h_{eq}). The driving force for all dynamic regimes is capillarity (surface tension and curved interfaces). The proportionality of the capillary rise rate $h(t)$ as a function of time for different regimes given by different solutions. Modified from Fries N and Dreyer M (2008) The transition from inertial to viscous flow in capillary rise. *Journal of Colloid and Interface Science*, **327**: 125–128.

the early inertial regime may vary from 0.38 m s^{-1} to 3.8 m s^{-1} as the capillary radius varies from $1000 \text{ }\mu\text{m}$ to $10 \text{ }\mu\text{m}$. These high interfacial velocities are particularly important during motion of infiltration and drainage fronts and these rapid velocities are associated with the Haines jumps (Haines, 1930) as shown by Moebius and Or (2014).

Dynamic contact angle

The contact angle formed between a flowing liquid front (advancing or receding) and a solid surface is not constant but reflects the interplay between capillary and viscous forces. The relative importance of these forces is often expressed by the so-called capillary number, $Ca = \eta v / \sigma$, with η the liquid dynamic viscosity, and v the contact line velocity. The dependency of the dynamic contact angle γ_D on the velocity of the contact line during complete wetting can be described by a nearly universal behavior according to the so-called Tanner law:

$$\gamma_D^3 = A Ca \quad (23)$$

where A is a constant (~ 94 for γ_D in radians). Eq. (23) fits the data of Hoffman for $Ca < 0.1$ and $\gamma_D < 130^\circ$ (Fig. 13). The complete range of Hoffman's data fitted to the empirical expression:

$$\gamma_D(\text{rad}) = \cos^{-1} \left\{ 1 - 2 \tanh \left[5.16 \left(\frac{Ca}{1 + 1.31 Ca^{0.99}} \right)^{0.706} \right] \right\} \quad (24)$$

as depicted by a continuous line in Fig. 13.

For conditions of partial wetting ($\gamma_s > 0$), the relationships between contact angle and Ca are less universal. It has been postulated that at low Ca the apparent dynamic contact angle remains close to the static angle but rapidly deviates when Ca exceeds the value for γ_s (Fig. 13). This postulate is formalized by the following expression:

$$\gamma_D^3 - \gamma_s^3 = A Ca \quad (25)$$

Additional examples of advancing and receding contact angle dependency on capillary number are shown in Fig. 14 (Hirasaki and Yang, 2002). Note that for receding contact angle there is a critical Ca above which the contact angle vanishes. The theoretical basis for the postulate in Eq. (25) was first derived by Voinov (1976), using hydrodynamic approximations near the moving contact line, resulting in:

$$\gamma_D^3 - \gamma_s^3 = 9Ca \ln(Y/Y_m) \quad (26)$$

where Y/Y_m is a ratio of macroscopic length over which the contact angle is defined ($\sim \text{mm}$) to molecular length where continuum theories fail ($\sim \text{nm}$). Application of Eq. (26) with $Y/Y_m = 10^5$ to the data of Hoffman is depicted in Fig. 13. A key shortcoming of such hydrodynamic models for a dynamic contact angle is the lack of consideration of the effects and interactions with solid surface properties. Most natural surfaces are seldom flat and dry, and propagation of contact lines invariably involves motion over precursory films that cover surface roughness and overcoming rough geometry during contact line motion.

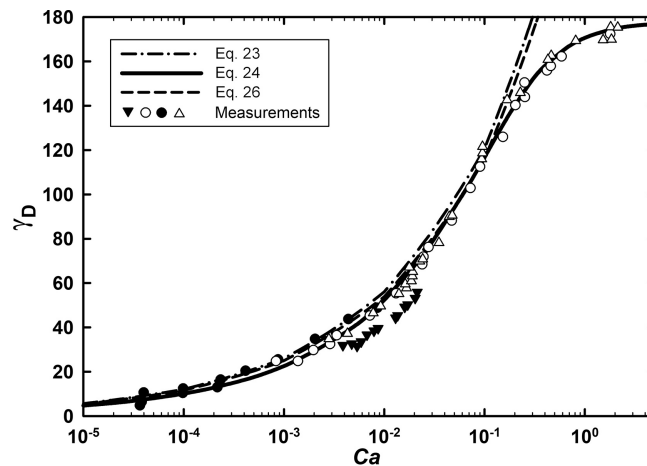


Fig. 13 Experimental results of Hoffmann fitted with Eq. (24) (Hoffman, 1975), and approximations given by Eqs. (23) and (26). Note that, for water flow in soils, the capillary number Ca rarely exceeds the range of values between 10^{-6} and 10^{-4} (for $v = 1 \text{ mm s}^{-1}$, $Ca = 1 \times 10^{-5}$, Friedman, 1999). Note that for natural soil conditions, the contact angle for water is seldom close to 0° (as shown here), and the intercept for low Ca under most practical conditions will rarely fall below 20° .

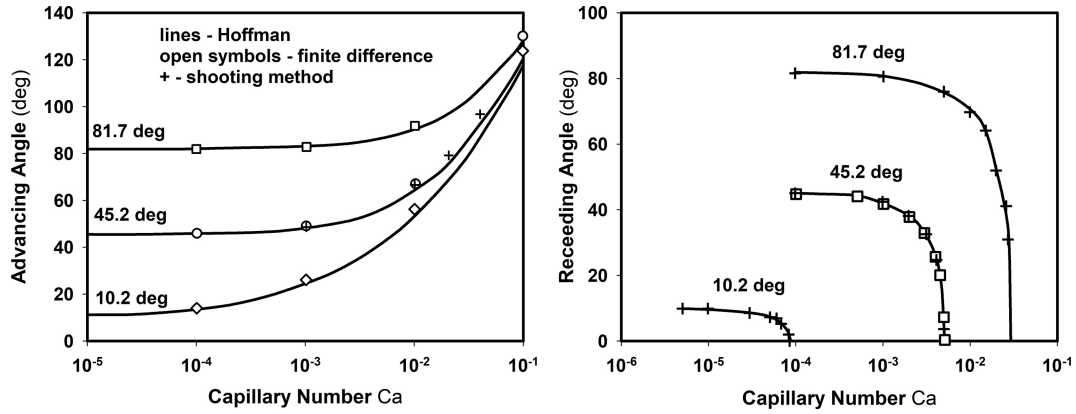


Fig. 14 Finite difference computation versus Eq. (25) and parameters from Kistler (1993) for advancing (left) and receding (right) contact angle as a function of Ca . Reproduced from Hirasaki GJ and Yang SY (2002) Dynamic contact line with disjoining pressure, large capillary numbers, large angles and pre-wetted, precursor, or entrained films. In: Mittal KL (ed.) *Contact Angle, Wettability and Adhesion*, Vol. 2, pp. 1–30. VSP, Zeist, The Netherlands.

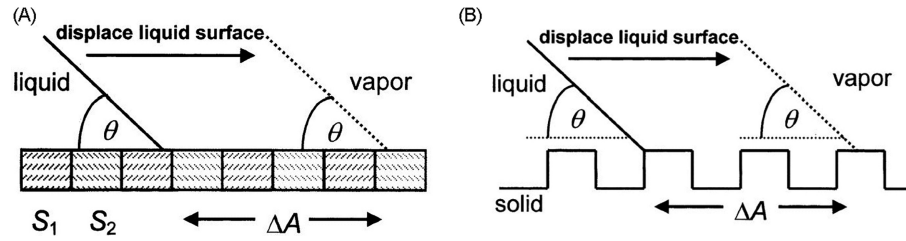


Fig. 15 Definition sketch for contact angle formation on (A) chemically heterogeneous surface and (B) rough surface with $Ro = \Delta A / \Delta A_0$, where A_0 is the projected area over a smooth surface. Reproduced from McHale G and Newton MI (2002) Frenkel's method and the dynamic wetting of heterogeneous planar surfaces. *Colloids and Surfaces, A: Physicochemical and Engineering Aspects*, **206**: 193–201.

Heterogeneous surfaces and microscale hysteresis

Contact angle on chemically heterogeneous and rough surfaces

Consider a chemically heterogeneous surface made up of patches of solids (or grains) with two different equilibrium contact angles γ_a and γ_b , and with the fraction of the area occupied by a solid given as f (Fig. 15). The apparent equilibrium contact angle (γ_e) for the composite surface is given by the semiempirical Cassie equation:

$$\cos \gamma_e = f \cos \gamma_a + (1 - f) \cos \gamma_b \quad (27)$$

An example of the Cassie law for contact angle of water on a sand surface with increasing amounts of hydrophobic grains is shown in Fig. 16. The Cassie law (Eq. 27) is in remarkable agreement with experimental data for sand (Fig. 16) and silt surfaces. An interesting extension of the Cassie law for porous surfaces (soil, fabric, etc.) predicts that the apparent contact angle (γ_e) should be proportional to surface porosity (n):

$$\cos \gamma_e = (1 - n) \cos \gamma_a - n \quad (28)$$

The negative sign associated with porosity is due to the nonwetting properties of empty pores (i.e., air with $\cos \gamma_{air} = -1$).

These concepts of mixed wettability can be incorporated into the capillary rise model (Eq. 6) where capillary rise takes place in slits formed between two walls of different wettability. The same study applies the Cassie law to liquid retention in porous media and demonstrates these effects on the hydraulic properties of unsaturated porous media with varying surface wettability. In addition to surface chemical heterogeneity, the roughness of a surface is known to alter its wettability properties by increasing the wettable surface area per unit projected area, and by enabling a complex interplay between macroscopic contact angle and microscale geometry, leading to gas entrapment and a patchwork of micro-interfaces underneath the wetting fluid. A spectacular demonstration of surface roughness-induced super hydrophobicity with a water drop resting on a fractal hydrophobic surface and forming a contact angle of about 170° is shown in Fig. 17A. Such enhanced hydrophobicity is not only important for a variety of engineering and industrial treatments aimed at waterproofing of surfaces and fabrics, but it may also be important for explaining wettability properties of natural soil surfaces. Assuming that surface roughness defined by the ratio of the real surface area relative to flat surface ($Ro = \Delta A / \Delta A_0$) affects only the solid-liquid and solid-vapor interfacial areas, minimization of surface free energy results in the so-called Wenzel equation:

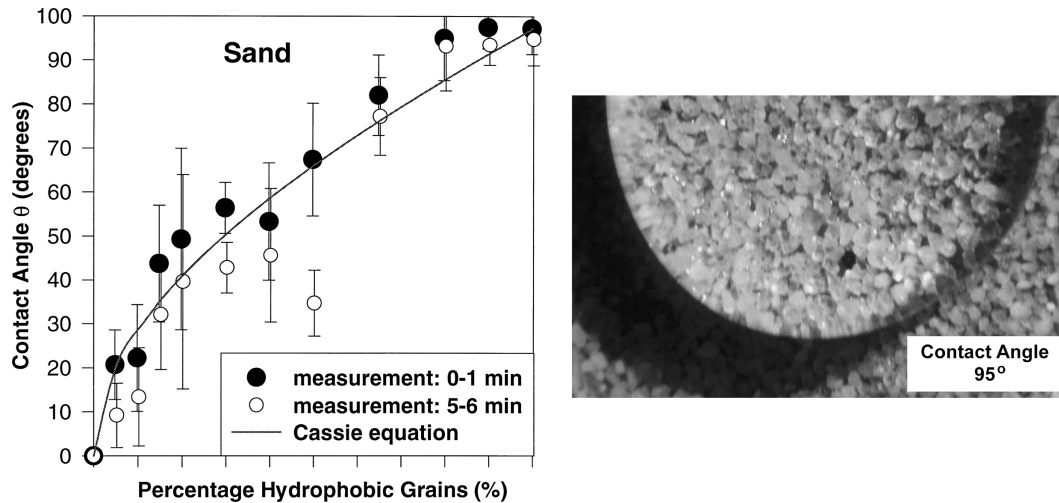


Fig. 16 Application of the Cassie law to (A) experimental results of contact angle with sand surfaces containing different proportions of hydrophobic (treated) sand grains; and (B) an image of a water droplet on nonwetting sand forming a contact angle of 95° . Adapted from Bachmann J, Ellies A, and Hartge KH (2000) Development and application of a new sessile drop contact angle method to assess soil water repellency. *Journal of Hydrology*, **231–232**: 66–75.

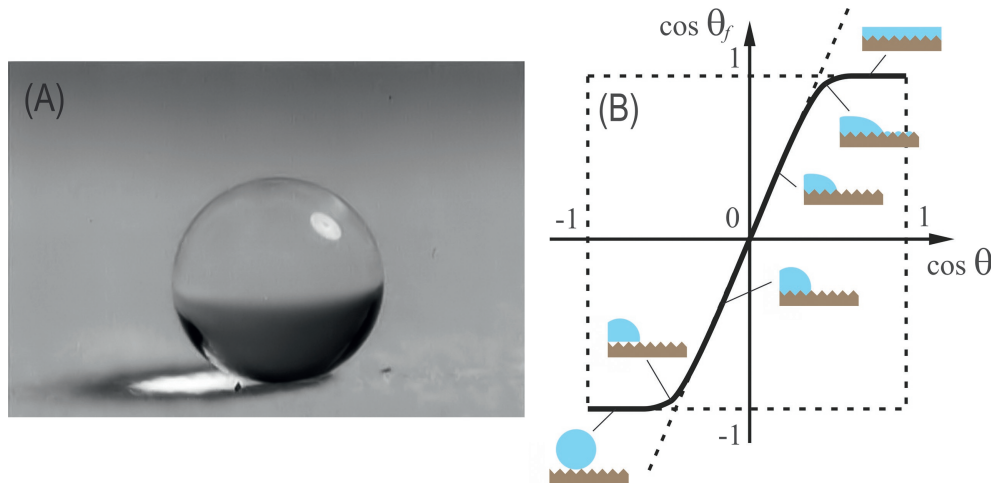


Fig. 17 (A) Water drop ($r = 1\text{ mm}$) resting on fractal rough surface with $R_O = 4.4$ (Eq. 29); and (B) apparent contact angles as a function of surface microroughness for a range of surfaces with different wettability. Reproduced from Onda T, Shibuichi S, Satoh N, and Tsujii K (1996) Super-water-repellent fractal surfaces. *Langmuir*, **12**: 2125–2127.

$$\cos \gamma_e = R_O \cos \gamma \quad (29)$$

where γ is the static contact angle for a smooth surface of similar chemical composition (see scheme in Fig. 15B).

The scope of surface influence is more complicated than predicted by simple expressions such as the Cassie and Wenzel equations. Other factors such as details of roughness geometry, interfacial pinning, and air trapping conspire to accentuate surface wetting properties as shown in Fig. 17B. The scheme depicted in Fig. 17B is based on experimental results showing the apparent contact angle on a rough surface plotted against the static contact angle on a smooth surface with similar chemical composition (to isolate the influence of surface roughness). Subsequent studies have shown a range of behaviors and asymmetry between the hydrophobic ($\cos \gamma < 0$) and hydrophilic ($\cos \gamma > 0$) sides of Fig. 17B. It is interesting to note that certain roughness patterns induce formation of air patches trapped underneath the liquid (similar to water drops resting on surfaces of some plant leaves).

Capillarity and soil cohesion

An important consequence of the Young-Laplace equation (Eq. 5) is the onset of a tensile force that pulls soil particles together across pendular water held at grain contacts, which is a familiar phenomenon to anyone attempting to build a sandcastle on the beach. Keen (1924) and Haines (1925) developed basic equations that relate capillary-induced cohesion and particle size and water

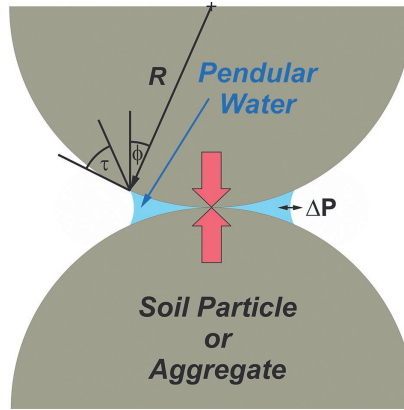


Fig. 18 Sketch illustrating capillary cohesion between soil particles.

content of idealized soils (i.e., spherical grains of known radius). For equal-sized spherical grains, the tensile force F_{Cap} pulling the grains together is calculated as:

$$F_{Cap} = 2\pi R \sigma_{LV} \sin \phi \sin (\phi + \tau) - \Delta P \pi R^2 \sin^2 \phi \quad (30)$$

where R is the particle radius, ΔP is the Young-Laplace pressure (Eq. 5) with high negative values (tensile forces) for very small particles, ϕ and τ are the filling angles related to the amount of pendular water between grains (Fig. 18), and σ_{LV} is the surface tension of the liquid-vapor interface. These concepts have been embedded in calculations relevant to soil mechanics and provide a bridge between soil water retention and mechanical properties under certain simplifying assumptions. For example, [Ghezzehei and Or \(2000\)](#) have used these concepts to quantify rates of soil aggregate coalescence due to tensile (capillary) forces using a slightly modified version of Eq. (30) that considers flattening of contacts between aggregates.

Hysteresis

The amount of liquid retained in a porous medium is not uniquely defined by the value of the matric potential but is also dependent on the 'history' of wetting and drying. This phenomenon, known as hysteresis, is closely related to various aspects of pore geometry, capillarity, and surface wettability. The macroscopic manifestation of hysteresis in soil water retention (or soil water characteristic) is rooted in several microscale mechanisms, including: (1) differences in liquid-solid-gas contact angles for advancing and receding water menisci (Fig. 19A), which is accentuated during drainage and wetting at different rates; (2) the 'ink bottle' effect resulting from nonuniformity in shape and sizes of interconnected pores, as illustrated in Fig. 19B, whereby drainage of the irregular pores is governed by the smaller pore radius r , and imbibition is dependent on the larger radius R . Additional effects stem from pore angularity; (3) differences in air-entrapment mechanisms; and (4) swelling and shrinking of the soil under wetting and drying, respectively. From early observations to the present, the role of individual factors remains unclear, and hysteresis is a subject of ongoing research.

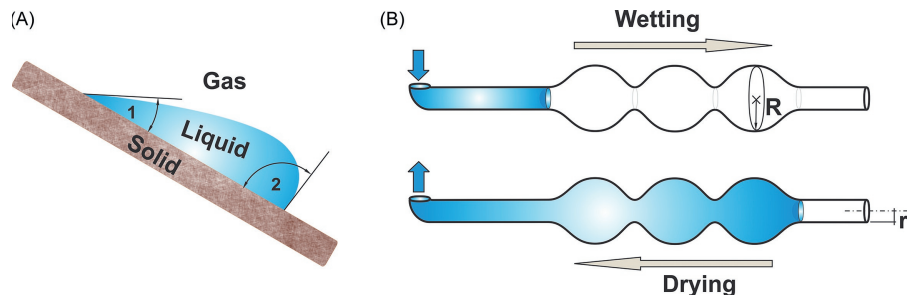


Fig. 19 Two microscale mechanisms for hysteresis in capillary behavior: (A) differences between advancing and receding contact angle; and (B) the 'ink bottle' effect depicting two different amounts of liquid retained in identical pores under the same matric potential. Note that we keep the capillaries horizontal to avoid gravity effects on imbibition and drainage processes.

Summary

Capillarity is central to the understanding of water behavior in shallow and partially saturated soils, it represents the roles of air-water interfaces and formation of contact lines that envelop water in soil. Capillarity reflects the forces that hold water in unsaturated soil against the action of gravity and give rise to spontaneous flow towards evaporating surfaces or plant roots. The configuration of remaining water in unsaturated soil is defined by capillarity and with it the hydraulic continuity required for mass flow. We reviewed the components of capillary rise in soil pores and dynamics of capillary processes. Capillarity imparts internal mechanical strength to partially saturated soil and granular media due to the action of tensile forces that pull grain contacts together.

References

- Adamson AW (1990) *Physical Chemistry of Surfaces*, 5th edn. New York: John Wiley.
- Bachmann J, Ellies A, and Hartge KH (2000) Development and application of a new sessile drop contact angle method to assess soil water repellency. *Journal of Hydrology* 231–232: 66–75.
- Bico J and Querre D (2002) Rise of liquids and bubbles in angular capillary tubes. *Journal of Colloid and Interface Science* 247: 162–166.
- Buckingham E (1907) *Studies on the Movement of Soil Moisture. Bulletin (United States. Bureau of Soils), No. 38*, Washington, DC: USDA.
- Friedman SP (1999) Dynamic contact angle explanation of flow rate-dependent saturation-pressure relationships during transient liquid flow in unsaturated porous media. *Journal of Adhesion Science and Technology* 13: 1495–1518.
- Fries N and Dreyer M (2008) The transition from inertial to viscous flow in capillary rise. *Journal of Colloid and Interface Science* 327: 125–128.
- Ghezzehei TA and Or D (2000) Dynamics of soil aggregate coalescence governed by capillary and rheological processes. *Water Resources Research* 36(2): 367–379.
- Haines WB (1925) Studies in the physical properties of soil. II. A note on the cohesion developed by capillary forces in an ideal soil. *Journal of Agricultural Science* 15(4): 529–535.
- Haines WB (1930) Studies in the physical properties of soil. V. The hysteresis effect in capillary properties, and the modes of moisture distribution associated therewith. *Journal of Agricultural Science* 20: 97–116.
- Hamraoui A and Nylander T (2002) Analytical approach for the Lucas–Washburn equation. *Journal of Colloid and Interface Science* 250: 415–421.
- Henn AR (2003) The surface tension of water calculated from a random network model. *Biophysical Chemistry* 105(2–3): 533–543.
- Hirasaki GJ and Yang SY (2002) Dynamic contact line with disjoining pressure, large capillary numbers, large angles and pre-wetted, precursor, or entrained films. In: Mittal KL (ed.) *Contact Angle, Wettability and Adhesion*. vol. 2, pp. 1–30. Zeist, The Netherlands: VSP.
- Hoffman RL (1975) A study of advancing interface. *Journal of Colloid and Interface Science* 50: 228–241.
- Jurin J (1718) II. An account of some experiments shown before the Royal Society; with an enquiry into the cause of the ascent and suspension of water in capillary tubes. *Philosophical Transactions of the Royal Society of London* 30(355): 739–747. <https://doi.org/10.1098/rstl.1717.0026>.
- Keen BA (1924) On the moisture relationships in an ideal soil. *The Journal of Agricultural Science* 14: 170.
- Kistler SF (1993) Hydrodynamics of wetting. In: Berg JC (ed.) *Wettability*, pp. 311–429. New York: Marcel Dekker.
- Laplace PS (1806) *Mecanique Celeste, Supplement to the*, 10th edn Paris: Courier.
- Lucas R (1918) Über das Zeitgesetz des kapillaren Aufstiegs von Flüssigkeiten. *Kolloid Zeitschrift* 23: 15–22.
- Moebius F and Or D (2014) Inertial forces affect fluid front displacement dynamics in a pore-throat network model. *Physical Review E* 90: 023019.
- Or D (2008) Scaling of capillary, gravity and viscous forces affecting flow morphology in unsaturated porous media. *Advances in Water Resources* 31: 1129–1136.
- Quere D, Raphael E, and Ollitrault J-Y (1999) Rebounds in a capillary tube. *Langmuir* 15: 3679–3682.
- Rideal EK (1921) On the flow of liquids under capillary pressure. *Philosophical Magazine* 44: 1152–1159.
- Tuller M, Or D, and Dudley LM (1999) Adsorption and capillary condensation in porous media: Liquid retention and interfacial configurations in angular pores. *Water Resources Research* 35(7): 1949–1964.
- Villermaux E and Bossa B (2009) Single-drop fragmentation determines size distribution of raindrops. *Nature Physics* 5(9): 697–702.
- Voinov OV (1976) Hydrodynamics of wetting. *Fluid Dynamics* 11: 714.
- Washburn EW (1921) The dynamics of capillary flow. *Physical Review* 17: 273–283.
- Young T (1805) III. An essay on the cohesion of fluids. *Philosophical Transactions. Royal Society of London* 95: 65–87.

Darcy's law

Conrad Jackisch^a and Eva Kröner^b, ^aFaculty of Geosciences, Geoengineering and Mining, Technische Universität Bergakademie Freiberg, Freiberg, Germany; ^bInstitute of Bio- and Geosciences, Agrosphere (IBG-3), Forschungszentrum Jülich, Jülich, Germany

© 2023 Elsevier Ltd. All rights reserved.

This is an update of D. Swartzendruber, Darcy's Law, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 363-369, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00344-1>.

Definition and introduction	240
The Darcy model	240
Linear flow models	241
The Darcy experiment	242
Validity and limits of the model	242
Application of the Darcy equation to soil water and the extension to unsaturated cases	244
Isotropy and scale (in)variance	244
Darcy-scale and representative elementary volume	244
Transition to non-Darcian flow and non-uniform flow in soils	245
Summary	245
References	245

Abstract

The famous Darcy model relates the pressure gradient and the hydraulic conductivity of a porous filter layer to the flow rate of water through the filter. This macroscale linear flow model for saturated water flow through porous media is central for a whole class of environmental and technical processes. We present the foundations, assumptions and limitations of the model.

Key points

- Linear flow model for water through porous media
- Flow rate depends on pressure gradient and bulk material conductivity
- Linear approximation only valid for laminar flow cases dominated by viscous forces

Definition and introduction

Darcy's law is a linear flow model for water flow through a saturated porous medium. In 1856 the French engineer Henry P.G. Darcy presented his famous equation (Darcy, 1856) stating that flow rate (q) depends on the pressure gradient (∇H) and on the hydraulic conductivity (K) (Eq. 1):

$$q = K \nabla H \quad (1)$$

Based on a series of experiments in Dijon, Darcy discovered that the flow rate through a saturated sand column was proportional to the hydraulic gradient (pressure difference) in the system. The coefficient of proportionality is termed the saturated hydraulic conductivity.

The flow and redistribution of water through soils is crucial for the hydrologic cycle and most terrestrial ecohydrological processes (Rodríguez-Iturbe, 2000). Although the flow of a liquid in a porous medium is a complex multi-phase process and soils have to be addressed as structured and dynamic systems (Rabot et al., 2018), the water flow at saturated and stationary conditions is a central reference for the function of different soils. Notwithstanding the limitations of his linear flow model, Darcy provided the first analytical solution to unsteady, saturated flow. Its relevance to a whole class of environmental and technical processes is undisputed.

The Darcy model

The first publication of the Darcy model is referred to as appendix D to his book about the public water infrastructure in the city of Dijon, France (Darcy, 1856). When addressing filtration through sand layers he conducted experiments to determine the laws of water flowing through the filters. Darcy demonstrated with sand column experiments (Darcy, 1856; see Fig. 3) that *the volume of*

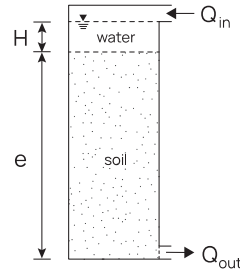


Fig. 1 A water-saturated soil column to illustrate Darcy's law for saturated filter flow. Steady state when $Q_{in} = Q_{out}$.

water passing through a layer of sand of a given nature is proportional to the pressure and in inverse proportion to the thickness of the layers crossed. This led to his original equation for volume flux of filtered water at atmospheric pressure under the filter (Eq. 2):

$$Q = \frac{ks}{e}(H + e) \quad (2)$$

where Q is the volume of filtered water, k is a coefficient depending on the nature of the sand, s is the surface area of a filter, e is the thickness of the layer of sand, H is the pressure as height of the water on the filter (Fig. 1).

In later forms of the same equation, the nature of his equation as linear flow model becomes more clearly accessible as stated in Eq. (1): A flux (q [LT^{-1}]) through a porous medium is defined by the saturated hydraulic conductivity (K [LT^{-1}]) as the invert of the hydraulic resistivity of the porous medium and the pressure gradient (∇H [L^{-1}]) as pressure difference per filter length ($\nabla H = \Delta H/e$).

Linear flow models

The Darcy equation is one form of the general linear flow model defining a flux as conductivity times a gradient. Further examples for these models are given in Table 1. As such it is a macroscopic description of the flow process with fundamental significance – and with a series of assumptions and limitations.

In analogy to other linear flow models (Table 1) one can directly deduce the implications of the Darcy equation: Flow through a porous medium (i) requires an initial gradient to be depleted, (ii) is governed by a specific conductivity, and (iii) both terms are assumed to act linearly. The gradient is a difference of the driving potential (here water head e.g., as analog to the electrical potential in Ohm's law) acting on a cross section over a given depletion length (here, length of the filter). The specific conductivity is treated as material constant and can be understood as the inverse of the resistivity of the material against the gradient depletion (here hydraulic conductivity e.g. as analog to the electrical resistivity).

The strength of the Darcy equation resides in the general clarification that an observed flow through a filter layer is proportional to the gradient. Because hydraulic conductivity is used in topographic indexes like the soil-topographic wetness index (Walter et al., 2002) one should note that the actual flow can deviate substantially from the saturated hydraulic conductivity depending on this gradient. Likewise, measuring flow alone is insufficient to derive hydraulic conductivity as material property. The driving gradient has to be considered alongside.

In addition to the famous experiments with constant gradients, Darcy appears to have been the first to apply the linear flow equation for an unsteady flow through a filter with falling head at the upper side of the filter and constant atmospheric pressure at

Table 1 Overview about general linear flow models.

Equation	Quantity transported across a flow region	transport coefficient of flow region	difference in driving potential across system
Ohm's law of electric currents			
$I = \sigma \frac{\Delta V A}{L} = \frac{1}{R} V$	$I =$ electric charge per time (Cs^{-1}, A)	$\sigma =$ specific electrical conductivity (Scm^{-1}) or invert of resistance ($m\Omega^{-1}$)	$\Delta V =$ difference in electrical potential (V)
Darcy's law of fluid flow through a porous medium			
$Q = k \frac{\Delta h A}{L}$	$Q =$ volume per time [$L^3 T^{-1}$]	$k =$ hydraulic conductivity [LT^{-1}]	$\Delta h =$ difference in head of water column [L]
Fick's law of diffusion flux			
$J = D \frac{\Delta C A}{L}$	$J =$ mass per area per time [$ML^{-2} T^{-1}$]	$D =$ Diffusion coefficient [$L^2 T^{-1}$]	$\Delta C =$ difference in concentration [ML^{-3}]
Fourier's law of thermal conduction			
$q = K \frac{\Delta T A}{L}$	$q =$ heat flux density [MT^{-3}]	$K =$ Thermal conductivity [$MLT^{-3}\Omega^{-1}$] ($WK^{-1}m^{-1}$)	$\Delta T =$ difference in temperature [Ω]

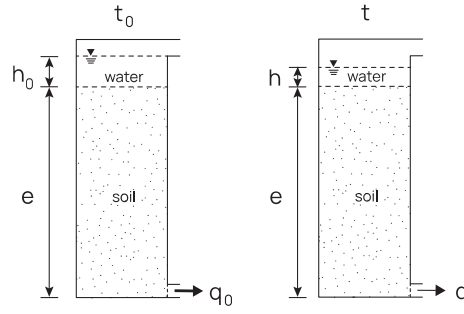


Fig. 2 A water-saturated soil column to illustrate Darcy's law for saturated filter flow with falling head (h).

the lower outlet (Darcy, 1856, appendix D and E, Fig. 2). This was done by combining Eq. (1) with continuity of the flux (q) through a filter surface (s) resulting in changes of the head (dh) over time (dt):

$$\frac{q}{s} = \frac{k}{e}(h + e) = - \frac{dh}{dt} \quad (3)$$

Eq. (3) can be resolved to:

$$\ln(h + e) = \ln(h_0 + e) - \frac{k}{e}[t - t_0] \quad (4)$$

and thus

$$\ln q = \ln q_0 - \frac{k}{e}[t - t_0] \quad (5)$$

which allows for the determination of the filter properties based on falling head observations. The notation corresponds to the original and Eq. (2) with h as water head above the filter, e as filter length, t as time and q as drained volume. The subscript 0 refers to the initial variable values.

The Darcy experiment

The properties and background of the Darcy equation becomes clearer, when one revisits the original study (Darcy, 1856). The main focus resides on clearing water in different arrangements of sandy filter layers. Obviously, the recognition of the gradient-dependent filter flow is paramount for such a study as it allows the unravelling of the properties of the porous media and filter length from observed pressure differences and fluxes.

The central experiment was conducted with a sand column apparatus (Fig. 3 as reported in Brown, 2002). The sands from the banks of the Saône river were packed into long steel tubes with manometers at both ends and a tap at the bottom. Apparently, the water supply was directly connected to the laboratory's water system. Repeated experiments lead to the recognition of the linear relationship between gradient and flow scaled by the "material coefficient" (saturated hydraulic conductivity).

One should note the scale of the apparatus with a considerable filter length of about 3 m (the exact values differ between the figure and the manuscript description) and a tube diameter of 0.35 m.

Validity and limits of the model

In accordance to his experiments and application cases, Darcy himself was well aware about limits to his model. He noted that filter fluxes with moderate gradients through sandy layers should allow for the linear simplification. In the following, different dimensions and implications of inherent simplifications in Darcy's law will be explained. A summary of the validity limits is given in Fig. 4.

The conceptualization of the macroscale model is also referred to as "Darcy-scale." Actual microscopic processes including friction at the liquid-solid interfaces, non-linear scaling of the flow with the pore radius and dynamic viscosity are all subsumed into the macroscale constant for hydraulic conductivity (see Bernoulli's Principle (1738) and Hagen-Poiseuille equation (Hagen, 1839; Poiseuille, 1840). This reduction further implies cases to be limited to saturated conditions in essentially homogeneous and isotropic pore spaces.

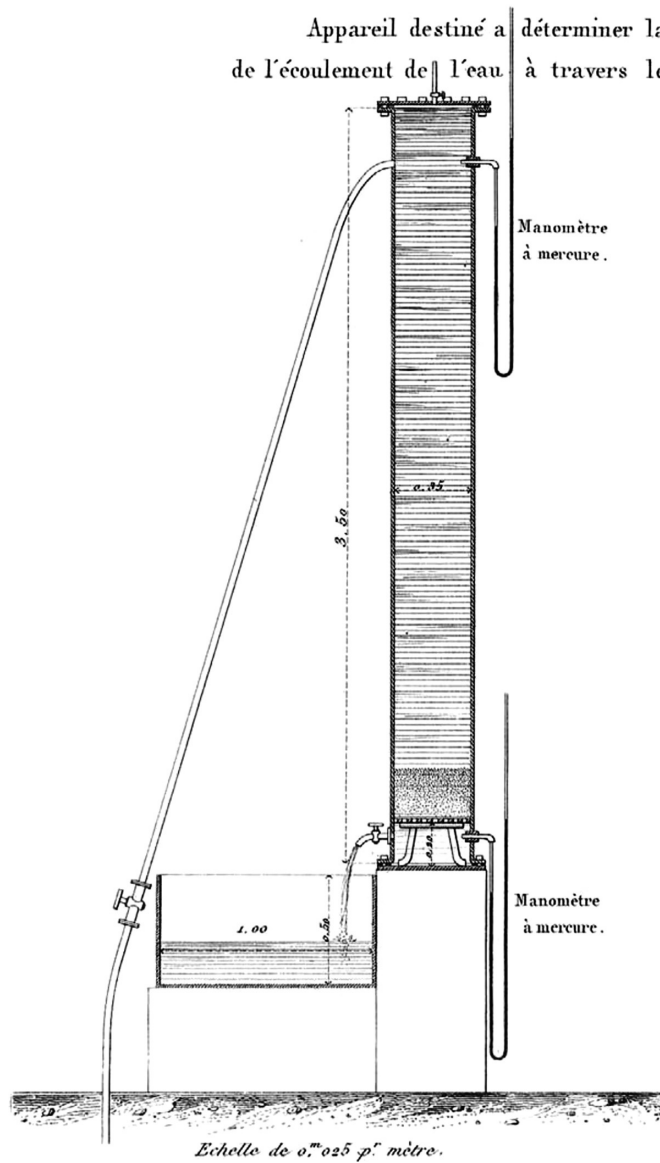


Fig. 3 Sand column apparatus of Darcy filtration experiments (Darcy, 1856, Fig. 3, Plate 24 in Brown, 2002).

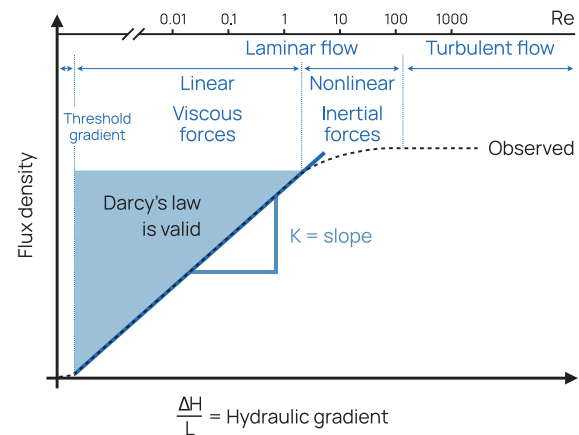


Fig. 4 Validity of Darcy flow in linear range (courtesy of D. Or). Re is Reynolds number, K hydraulic conductivity, ΔH pressure difference, L length of the filter.

Application of the Darcy equation to soil water and the extension to unsaturated cases

Classical soil physical measurement procedures strongly adhere to analyses in soil cores and packed physical models. As such, they remain within the general realm of soils as filter layers in the sense of Darcy. Obviously, Darcy's law required to be extended to allow for unsaturated (but stationary) conditions by defining hydraulic conductivity as a function of soil water content ($K(\theta)$) known as Darcy-Buckingham equation (Nimmo and Landa, 2005):

$$q = K(\theta) \nabla H \quad (6)$$

The conceptual idea behind this is that if the pore space is not fully occupied by water, water films along the solid particles of the soil exhibit longer pathways and thinner cross sections for the flow. One can therefore expect (i) K to decrease with decreasing water content and (ii) $K(\theta)$ to be a highly non-linear exponential function. Given that soil water content (θ) is not a complete state variable, its counterpart matric potential (the soil capillary tension sucking water into air-filled pore space, ψ) is also used to determine unsaturated hydraulic conductivity.

To overcome the limitation to stationary conditions, the water passage through different calculation nodes was described by the continuity equation of the mass balance:

$$\frac{\partial \theta}{\partial t} = - \frac{\partial q}{\partial z} \quad (7)$$

The change of soil water content ($\partial \theta$) in a calculation node over time (∂t) equals the change in drained volume ($-\partial q$) in space (∂z). Eq. (6) can be directly set into Eq. (7) yielding the well-known Darcy-Richards equation, which thus still adheres to the general linear foundation of Darcy's law.

Unsurprisingly, it has subsequently become a core topic of soil physics to define and resolve the limits of the well-known mismatch of heterogeneous pore space reality and multi-phase flow in soils with the assumption of macroscopic homogeneity, isotropy and linearity and laminarity. Practitioners should note that even modern analytical tools to measure saturated hydraulic conductivity and soil water retention characteristics in undisturbed soil cores still rely on the conceptual foundation of Darcy's law including the referring Darcy-scale. Although experimental evidence for the limitations of the Darcy equation has existed at least since Schneebeli (1955), technical capacities to resolve soil water dynamics at the actual pore level emerged only recently (Lucas et al., 2020).

Isotropy and scale (in)variance

Although Darcy referred to different layers of filter sands, a strong contrast of hydraulic properties, for example, between soil layers or under contrasting antecedent conditions, is omitted or subsumed in the model for an overall filter property. In contrast to the original experiments, with packed columns several meters in length, or sandy river banks, in most natural soils such isotropic conditions are rather rarely found (Gerke, 2006).

Among others, Miller and Miller (1956) derived Darcy's law from the Stokes equation by averaging sub-scale variance and boundary effects to the macroscopic Darcy-scale. To do so, they introduced four assumptions for their approximation:

1. The geometric properties of the solid part of the porous medium is homogeneous, isotropic and permanent so that these properties are independent of position, orientation and time.
2. The liquid is uniform in its surface tension, contact angle, density and viscosity throughout the flow system.
3. The properties of the gaseous phase are constant over the flow system.
4. All liquid and all gaseous proportions are connected so that neither isolated droplets nor bubbles exist.

Sposito et al. (1979) and others have presented further theoretical elaborations deriving the macroscopic description from fluid mechanics, Markov processes and statistical thermodynamics. Their studies highlighted many of the implications when reducing the complex multi-phase flow in soils to the linear flow model. They came to similar conclusions as Darcy himself, that *the theory would be expected to work best with coarse materials in the range of sands, and perhaps silts, at moderate liquid contents.*

Darcy-scale and representative elementary volume

Following the notion that models are generally valid at a specific scale, the definition of the representative elementary volume (REV) is deeply connected to the Darcy-scale, too. Given the four central assumptions by Miller and Miller (1956) and acknowledging the state of science in laboratory and field soil physical analytical procedures, the REV can hardly be unraveled from the Darcy-scale, when most measurements closely relate to the linear flow model.

Roth (2008) revisits the question of why scaling of water flow through porous media is so complicated. Darcy's law subsumes the many non-linear effects of the fluid (i.e., temperature, pressure and solute depending viscosity), of the pore space geometry (e.g., non-monotonous, non-circular capillaries) and of the multi-phase surfaces into one parameter in a linear model. The upscaling of

these microscale effects to the macroscale description is limited to cases and the extent to which the pore-geometry itself can be upscaled. The corresponding characteristic length of the REV must be large enough to fulfill the stationarity assumption (Roth, 2008), yet small enough to resolve driving processes (Vogel and Ippisch, 2008). To fulfill both requirements has been described as closure problem (Whitaker, 1986).

Transition to non-Darcian flow and non-uniform flow in soils

Despite uniform pore space characteristics, the observed flow has been found to deviate from the linear relation in the Darcy equation (Fancher and Lewis, 1933; Hansbo, 2001). Irmay (1958) interprets the Darcy equation as the first term in the more general and non-linear Forchheimer equation. At higher Reynolds numbers (as the ratio of inertial forces to viscous forces used to discern laminar and turbulent flow conditions), the non-linear contribution of inertia to the transport of momentum in the flow field become prevalent (Yazdchi et al., 2010, 2011) (see upper validity limit in Fig. 4). This transition is found to happen already for Reynolds numbers between 0.4 and 3 (Tachibana et al., 2017). Above the transition range a sharp decrease of the permeability coefficient was found in the same study, manifesting the shift toward turbulent fluxes at high Reynolds numbers.

The disagreement of observed soil water dynamics with the assumptions in Darcy's law have been debated also with respect to macropores (Beven and Germann, 1982, 2013) and fractures (Berkowitz, 1989). Measurements of non-uniform infiltration (Flury et al., 1994; Nimmo, 2021) question the applicability of Darcy's findings to natural and rarely saturated soils. The heterogeneous pore networks are particularly difficult to upscale to a unique macroscopic description. Moreover, dispute about the assumption of stationarity of the flow field as laminar flow is inevitable as the heterogeneity also hinders the inherently assumed perfect mixing (Jackisch and Zehe, 2018; Berkowitz and Zehe, 2020). For infiltration, Germann (2020) challenges the applicability of the Darcy-concept of filter flow even further by explaining water dynamics based on Newton's viscous shear flow.

After Hagen-Poiseuille (1841) the radius of a pore quadratically scales the mean velocity and enters the viscosity-based form to the power of 4. With non-uniformly distributed pore sizes, consequently the critical Reynolds number for laminar, creeping flow can be locally surpassed in pores with larger radius. Whitaker (1986) pointed to spatial deviations of the pressure and velocity within an assumed REV. In soils the non-uniform boundary conditions and variances in the flow field can become self-reinforcing when the pore structures become reconfigured (Zehe et al., 2021).

At the other end of the pore size spectrum, flow in clay soils with a very low hydraulic conductivity was also found to deviate from Darcy's simplification with pressure gradients building up before flow starts (Miller and Low, 1963) (see lower validity limit in Fig. 4). Unlike with Darcian flow with a continuous hydraulic head in the filter, pore pressure diffusion around adsorbed water at the large specific surface of the mineral-water interface was found to retard the flow reaction (Tang et al., 2016).

Summary

Darcy's linear flow model is fundamental for many environmental and technical processes as it clarifies that a pressure gradient needs to act on a porous media with a specific bulk resistance or conductivity to generate flow, depleting this gradient. As with most linear model approximations, special care has to be taken about the embedded assumptions and the limits for any application.

References

- Berkowitz B (1989) Boundary conditions along permeable fracture walls: Influence on flow and conductivity. *Water Resources Research* 25: 1919–1922. <https://doi.org/10.1029/wr025i008p01919>.
- Berkowitz B and Zehe E (2020) Surface water and groundwater: Unifying conceptualization and quantification of the two “water worlds” *Hydrology and Earth System Sciences* 24: 1831–1858. <https://doi.org/10.5194/hess-24-1831-2020>.
- Bernoulli D (1738) *Hydrodynamica, Sive de viribus et motibus fluidorum commentarii*. Sumptibus Johannis Reinholdi Dulseckeri, Argentorati (i.e. Strasbourg).
- Beven K and Germann P (1982) Macropores and water flow in soils. *Water Resources Research* 18: 1311–1325. <https://doi.org/10.1029/wr018i005p01311>.
- Beven K and Germann P (2013) Macropores and water flow in soils revisited. *Water Resources Research* 49: 3071–3092. <https://doi.org/10.1002/wrcr.20156>.
- Brown GO (2002) Henry Darcy and the making of a law. *Water Resources Research* 38. <https://doi.org/10.1029/2001wr000727>. 11-1–11-12.
- Darcy H (1856) In: Dalmont V (ed.) *Les fontaines publiques de la ville de Dijon*. Paris: Librairie des corps impériaux des ponts et chaussées et des mines.
- Fancher GH and Lewis JA (1933) Flow of Simple Fluids through Porous Materials. *Industrial and Engineering Chemistry* 25: 1139–1147. <https://doi.org/10.1021/ie50286a020>.
- Flury M, Flühler H, Jury WA, and Leuenberger J (1994) Susceptibility of soils to preferential flow of water: A field study. *Water Resources Research* 30: 1945–1954. <https://doi.org/10.1029/94wr00871>.
- Gerke HH (2006) Preferential flow descriptions for structured soils. *Journal of Plant Nutrition and Soil Science* 169: 382–400. <https://doi.org/10.1002/jpln.200521955>.
- Germann P (2020) Viscosity Controls Rapid Infiltration and Drainage. *Not the Macropores, Water* 12: 337. <https://doi.org/10.3390/w12020337>.
- Hagen GHL (1839) III. Über die Bewegung des Wassers in engen cylindrischen Röhren. *Poggendorfs Annalen der Physik und Chemie* 46(2): 423–442.
- Hansbo S (2001) Consolidation equation valid for both Darcian and non-Darcian flow. *Géotechnique* 51: 51–54. <https://doi.org/10.1680/geot.2001.51.1.51>.
- Irmay S (1958) On the theoretical derivation of Darcy and Forchheimer formulas. *Eos, Transactions of the American Geophysical Union* 39: 702–707. <https://doi.org/10.1029/tr039i004p00702>.
- Jackisch C and Zehe E (2018) Ecohydrological particle model based on representative domains. *Hydrology and Earth System Sciences* 22: 3639–3662. <https://doi.org/10.5194/hess-22-3639-2018>.
- Lucas M, Vetterlein D, Vogel H-J, and Schlüter S (2020) Revealing pore connectivity across scales and resolutions with X-ray CT. *European Journal of Soil Science* 130: 305–315. <https://doi.org/10.1111/ejss.12961>.

- Miller RJ and Low PF (1963) Threshold Gradient for Water Flow in Clay Systems. *Soil Science Society of America Journal* 27: 605–609. <https://doi.org/10.2136/sssaj1963.03615995002700060013x>.
- Miller EE and Miller RD (1956) Physical Theory for Capillary Flow Phenomena. *Journal of Applied Physics* 27: 324–332. <https://doi.org/10.1063/1.1722370>.
- Nimmo JR (2021) The processes of preferential flow in the unsaturated zone. *Soil Science Society of America Journal* 85: 1–27. <https://doi.org/10.1002/saj2.20143>.
- Nimmo JR and Landa ER (2005) The Soil Physics Contributions of Edgar Buckingham. *Soil Science Society of America Journal* 69: 328–342. <https://doi.org/10.2136/sssaj2005.0328>.
- Poiseuille JLM (1840) Physiques - Recherches experimentales sur le mouvement des liquides dans les tubes de tres petits diametres. *Academie des Sciences, Comptes Rendus* 11: 961–967.
- Rabot E, Wiesmeier M, Schlüter S, and Vogel H-J (2018) Soil structure as an indicator of soil functions: A review. *Geoderma* 314: 122–137. <https://doi.org/10.1016/j.geoderma.2017.11.009>.
- Rodriguez-Iturbe I (2000) Ecohydrology: A hydrologic perspective of climate-soil-vegetation dynamics. *Water Resources Research* 36: 3–9. <https://doi.org/10.1029/1999wr900210>.
- Roth K (2008) Scaling of water flow through porous media and soils. *European Journal of Soil Science* 59: 125–130. <https://doi.org/10.1111/j.1365-2389.2007.00986.x>.
- Schneebeli G (1955) Expériences sur la Limite de Validité de la Loi de Darcy et L'Apparition de la Turbulence Dans un Écoulement de Filtration. *La Houille Blanche* 41: 141–149. <https://doi.org/10.1051/lhb/1955030>.
- Sposito G, Gupta VK, and Bhattacharya RN (1979) Foundation theories of solute transport in porous media: A critical review. *Advances in Water Resources* 2: 59–68. [https://doi.org/10.1016/0309-1708\(79\)90012-5](https://doi.org/10.1016/0309-1708(79)90012-5).
- Tachibana I, Moriguchi S, Takase S, Terada K, Aoki T, Kamiya K, and Kodaka T (2017) Characterization of transition from Darcy to non-Darcy flow with 3D pore-level simulations. *Soils and Foundations* 57: 707–719. <https://doi.org/10.1016/j.sandf.2017.08.003>.
- Tang L, Chen H, and Song J (2016) Process of pore pressure diffusion in saturated clay soil and impact of adsorbed water. *Geosciences Journal* 20: 649–665. <https://doi.org/10.1007/s12303-016-0002-4>.
- Vogel H-J and Ippisch O (2008) Estimation of a critical spatial discretization limit for solving Richards' equation at large scales. *Vadose Zone Journal* 7: 112–114. <https://doi.org/10.2136/vzj2006.0182>.
- Walter MT, Steenhuis TS, Mehta VK, Thongs D, Zion M, and Schneiderman E (2002) Refined conceptualization of TOPMODEL for shallow subsurface flows. *Hydrological Processes* 16: 2041–2046. <https://doi.org/10.1002/hyp.5030>.
- Whitaker S (1986) Flow in porous media I: A theoretical derivation of Darcy's law. *Transport in Porous Media* 1: 3–25. <https://doi.org/10.1007/bf01036523>.
- Yazdchi K, Srivastava S, and Luding S (2010) On the transition from creeping to inertial flow in arrays of cylinders. *Mechanics of Solids, Structure and Fluids* 9: 767–772. <https://doi.org/10.1115/imece2010-37689>.
- Yazdchi K, Srivastava S, and Luding S (2011) Microstructural effects on the permeability of periodic fibrous porous media. *International Journal of Multiphase Flow* 37: 956–966. <https://doi.org/10.1016/j.ijmultiphaseflow.2011.05.003>.
- Zehe E, Loritz R, Edery Y, and Berkowitz B (2021) Preferential pathways for fluid and solutes in heterogeneous groundwater systems: Self-organization, entropy, work. *Hydrology and Earth System Sciences* 25: 5337–5353. <https://doi.org/10.5194/hess-25-5337-2021>.

Diffusion

TKK Chamindu Deepagoda^a and Shoichiro Hamamoto^b, ^aDepartment of Civil Engineering, Faculty of Engineering, University of Peradeniya, Peradeniya, Sri Lanka; ^bGraduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan

© 2023 Elsevier Ltd. All rights reserved.

This is an update of T.M. Addiscott, P. Leeds-Harrison, DIFFUSION, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 389–394, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00346-5>.

Introduction	247
How does diffusion occur?	248
Fick's laws of diffusion	248
Maxwell-Stefan diffusion equation	249
Importance of diffusion in soil	249
Different diffusion regimes in soil	250
Diffusion coefficients in soil-liquid and soil-gaseous phases	250
Solute diffusion coefficient in soil	250
Modeling solute diffusion	250
Soil-gas diffusion coefficient	251
Modeling soil-gas diffusion	252
Unified modeling approach for solute and gas diffusion in soil	253
Concluding remarks	254
References	255

Abstract

Diffusion is the movement of a chemical substance in the gas or liquid phase along a chemical gradient and constitutes a primary mechanism of mass transfer in natural and engineered soil ecosystems. Many ecosystems services and functions in soils are linked to diffusion-controlled mass transfer processes in soil-aqueous and soil-gaseous phases. The mathematical framework describing diffusion in soil is based on Fick's laws of diffusion, with extended versions available for applications in multi-component mixtures. Diffusion coefficients in soil-aqueous and soil-gaseous phases are determined from measurements of chemical concentration over time using specific apparatus under controlled boundary conditions. Predictive-descriptive diffusivity models offer a fast and convenient alternative to estimate diffusion coefficients in both aqueous and gaseous phases using easily measurable soil properties.

Key points

- Diffusion-controlled processes constitute a primary mechanism of mass transfer in soil.
- The basic mathematical framework describing diffusion in soil is founded on Fick's laws.
- Half-cell and one-chamber methods are generally adapted to measure solute and gas diffusion coefficients, respectively.
- Predictive-descriptive models are used to estimate diffusion coefficients in soil from basic soil properties

Introduction

Diffusion constitutes a key mechanism by which mass transfer takes place in soils and other porous domains. The word 'diffusion' comes from the Latin *diffundere*, meaning *spreading out*. Technically, diffusion can be described as the net movement of chemical substances in a gas or a liquid in response to a chemical gradient, until a state of equilibrium prevails within the system with respect to the chemical species. Diffusion can also occur in solid state, but it is of little interest for applications in soil in comparison to important implications in liquid and gaseous states.

Diffusion is often referred to as a process that occurs along a concentration gradient, although the true energy gradient for movement is provided by the partial pressure (in the gas phase) or differential chemical potential (in the liquid phase) of the particular gas or chemical in the medium. A simple example is an airtight chamber that is partitioned into two compartments by a moveable middle plate in such a way that each unit is filled with two different gases, but at the same total pressure. Since the partial pressures of the two gases in the two units are different, removing the middle plate instantly makes a partial pressure gradient, causing the movement of two gases from high partial pressure to low partial pressure until the partial pressure gradients, and hence the net movement across any section, cease. A similar process occurs in a liquid solution for chemicals (or solutes) along a chemical potential gradient. Essentially, diffusion is non-directional and may occur in any direction where the particular gradient exists.

How does diffusion occur?

Diffusion originates from *Brownian motion*, the random thermal movement of ions or molecules at the microscopic scale, named after Robert Brown (1773–1858), the first to study the phenomenon (Brown, 1828). Given a chemical potential gradient, the probability is that molecules will make a net movement from high potential regimes to low potential regimes, the phenomenon referred to as diffusion. Diffusion can thus be interpreted as the macroscopic manifestation of Brownian motion occurring at the microscopic level. Because gradients in partial pressure or concentration are gradually lessened by diffusion but never eliminated, partial pressures (of gases) and concentrations (of solutes) become uniform only at an infinite time (although in practice many diffusive flows become too small to measure after a relatively short time).

In the absence of an advective flow, which occurs due to an induced pressure gradient, diffusion becomes the prominent mechanism of mass transport in soils (Bear, 1972). In the presence of advective flows, however, the magnitude of advective transport dictates the relative importance of the diffusive transport. Moreover, advective flow triggers *dispersion*, the enhanced mixing observed at microscopic scale due to the velocity fluctuation around soil particles: this should not be confused with diffusion. Though the two mechanisms are implicitly analogous, two distinct differences exist between diffusion and dispersion: dispersion is velocity-dependent whereas diffusion is not, and dispersion is controlled by the direction of flow but diffusion is essentially concentration-dependent irrespective of the flow direction (Fig. 1).

Fick's laws of diffusion

Adolf Eugen Fick (1829–1901) developed the pioneering mathematical framework to describe the phenomena of diffusion. The work has almost exclusively been used as the constitutive law to predict diffusive fluxes in continuum mechanistic models

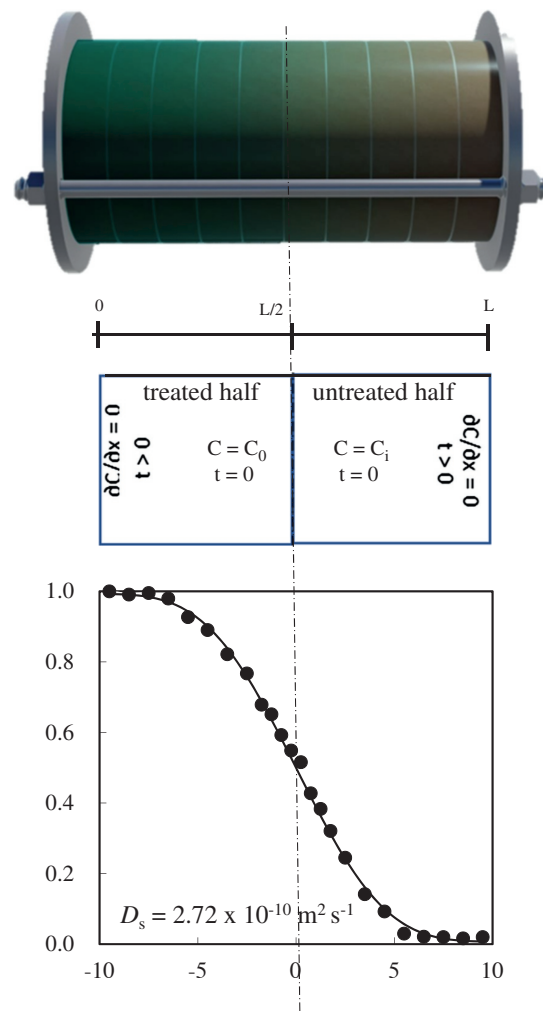


Fig. 1 Schematic of the half-cell apparatus for measuring soil diffusion coefficient, D_s (top), initial and boundary conditions (middle) and a typical solute concentration profile across the cell length (bottom). Unpublished data from Hamamoto et al. (2009) on Nishi-Tokyo Andisol (bulk density = 0.7 g cm^{-3} ; saturation = 0.95).

(Fick, 1855). Fick's mathematical expressions are principally analogous to other classical flux equations in which flux is expressed proportional to a gradient in a property of interest in the system, with a system-specific coefficient as the proportionality constant. Fick's diffusion expressions came under two fundamental laws. Fick's first law of diffusion relates the diffusive flux to the concentration gradient as follows:

$$J = -D \frac{\partial c}{\partial x} \quad (1)$$

where J is the diffusive flux ($\text{mol m}^{-2} \text{s}^{-1}$), D is the diffusion coefficient ($\text{m}^2 \text{s}^{-1}$), c is the concentration (mol m^{-3}), and x (m) is the spatial dimension. The negative sign indicates that the diffusive fluxes are directed against the concentration gradient.

Combining Fick's first law of diffusion with the continuity equation (Eq. 2)

$$\frac{\partial c}{\partial t} = -\frac{\partial J}{\partial x} \quad (2)$$

invokes Fick's second law of diffusion (Eq. 3), which describes the one-dimensional spatial-temporal relation of diffusion as follows:

$$\frac{\partial c}{\partial t} = D \frac{\partial^2 c}{\partial x^2} \quad (3)$$

where t is time (s). Though Fick's laws dominate in classical diffusion-controlled mass transfer applications due to their simplicity and reduced computational expense, the laws are applicable only for restricted cases as described below.

Maxwell-Stefan diffusion equation

Since the classical Fick's laws implicitly assume the presence of no other components affecting the flux of a specific chemical component, they are ideally applied for chemicals constituting a binary system with molecules in diluted concentrations. In a multicomponent system with individual components having comparable molar fractions, Fick's laws tend to yield inaccurate estimations. Therefore, the Maxwell-Stefan (MS) diffusion equation, an extended version of the Fick's equations, is generally applied to estimate diffusive fluxes in multicomponent mixtures. The MS constructive theory, developed independently by James Maxwell (for dilute gases) and Joseph Stefan (for liquids), can be presented as (Bird et al., 2007):

$$\nabla x_a = -\sum_{b=1}^N \frac{x_b N_a - x_a N_b}{D_{ab}} \quad (4)$$

where ∇ is vector differential operator (m^{-1}), x_a and x_b are the molar fractions of species a and b (dimensionless), respectively ($1 < a, b < N$; $a \neq b$), N_a and N_b are molar fluxes of species a and b ($\text{mol m}^{-2} \text{s}^{-1}$), respectively, D_{ab} is the binary diffusion coefficient ($\text{m}^2 \text{s}^{-1}$) for the gas pair a and b in a multicomponent system. Accordingly, for an N -component system, $N(N-1)/2$ number of binary diffusion coefficients are required (Bird et al., 2007).

Importance of diffusion in soil

Diffusion of chemical species in soil occurs in both aqueous and gaseous phases and hence plays a key role in natural and engineered soil ecosystems. Many useful regulatory and provisioning ecosystem services, ranging from nutrient cycling to pollutant transport, can be potentially linked to diffusion-controlled mass transfer processes.

In the vadose zone, diffusion becomes the dominant transport mechanism in water for dissolved chemicals when sufficient time has elapsed following infiltration and redistribution of water. Diffusion ensures an adequate supply of nutrients and dissolved ions to the depleted root zones caused by root uptake of essential plant nutrients. Dissolved oxygen, consumed during rhizosphere respiration and microbial metabolism, is continually replenished by diffusion of oxygen from distant soil-water systems. Deeper in the soil and at larger scale, diffusion offers a primary mode of communication between a stagnated aquifer and a coexisting flow domain, leading to the exchange and spreading of chemicals and contaminants. Diffusion predominates in engineered soil systems, which are hydraulically isolated, causing them to contain chemicals or waste materials. The efficiency of clean-up technologies for contaminated sites with dissolved contaminants largely depends on how well diffusive contaminant migration in the soil is understood.

Soil-gas diffusion, on the other hand, occurs predominantly in the gaseous phase, as gas diffusion in liquid is nearly a four-order-of-magnitude slower process than that in the air (Cussler, 2009). Soil-gas diffusion facilitates soil aeration by the continuous supply of oxygen to plant roots and prompt removal of root-zone carbon dioxide, thereby maintaining a healthy soil atmosphere for plant growth. Mitigation of diffusion-dominated emission of greenhouse gases from agroecosystems is essential to regulate future climate change. Volatile organic compounds may diffuse in soil from a leaky underground storage facility toward the ground surface and enter into overlying infrastructure through foundation cracks, thereby affecting indoor air quality (Unnithan et al., 2021). Landfills constitute another complex and unorthodox ecosystem, where diffusion and off-site emission of landfill gases (predominantly methane and carbon dioxide) may potentially lead to deadly catastrophes such as explosions or fire hazards.

Different diffusion regimes in soil

In porous media with a wide spectrum of differently-sized pores, diffusion becomes fundamentally pore size-dependent; the ratio of pore size to the size of the diffusing molecules dictates the governing diffusion mechanism. If pores have sizes smaller than the size of diffusing molecules, they are inaccessible and hence lead to a complete *impasse* (with no diffusion). When molecules are of comparable dimensions with pore size, other diffusion mechanisms may occur, initiated with *configurational diffusion* (Siddiqui et al., 2021). *Knudsen diffusion* is associated with pore sizes that are larger than that of molecules but are smaller than the mean-free distance of the molecules (Clifford and Hillel, 1986). In pore domains where pores are larger than the mean-free distance of molecules, *Fickian diffusion* predominates (Jaynes and Rogowski, 1983), and is very apparent in natural soils. A combination of different diffusion regimes is most likely to occur in soils with broadly ranging pore sizes, whereas the predominant pore domain size dictates the governing diffusion regime for specific molecular constituents. While different constitutive theories explain the aforementioned diffusion regimes, the Fickian-based mathematical framework is widely adopted to describe diffusion in applied soil science.

Diffusion coefficients in soil-liquid and soil-gaseous phases

Solute diffusion coefficient in soil

Diffusion of dissolved chemical constituents (referred hereafter as solutes) in the soil-liquid phase is typically characterized by the solute diffusion coefficient in soil, (D_s , $\text{m}^2 \text{s}^{-1}$), which can be conveniently derived from Fick's laws. Since different Fickian-based formulations appear in the literature to calculate solute diffusion coefficients, care must be exercised when comparing the coefficients calculated from different theoretical derivations.

Amongst different methods for the practical measurement of solute diffusion coefficients, the so-called half-cell method is the most popular. In brief, this comprises two halves of a cell, with one given a shot of the solutes, with the other being free of them. They are then brought into an instant and complete hydraulic contact, enabling the solutes to migrate from the treated cell to the untreated one, solely by diffusion.

The solute diffusion coefficient can be determined from the observed concentration vs. time data along the length of the half-cell, solving the diffusion equation under specified boundary conditions (Flury and Gimmi, 2002).

Governing equation

$$\frac{\partial c}{\partial t} = \frac{D_s}{R\theta} \frac{\partial^2 C}{\partial x^2} \quad (5)$$

where R is the retardation factor (dimensionless), and θ is the soil moisture content ($\text{m}^3 \text{m}^{-3}$).

Initial and Boundary conditions

$$\begin{aligned} \frac{\partial C(x=0, t > 0)}{\partial x} &= 0; \\ C\left(0 \leq x \leq \frac{L}{2}, t=0\right) &= C_0; \quad C\left(\frac{L}{2} < x \leq L, t=0\right) = C_i \\ \frac{\partial C(x=L, t > 0)}{\partial x} &= 0 \end{aligned}$$

Analytical solution

$$\frac{C - C_i}{C_0 - C_i} = 0.5 + \frac{2}{\pi} \sum_{m=1}^{\infty} \frac{\exp(-D_s m^2 \pi^2 t / RL^2 \theta)}{m} \cos\left(\frac{m\pi x}{L}\right) \sin\left(\frac{m\pi}{2}\right) \quad (6)$$

where C_0 is the initial liquid-phase concentration of the treated half-cell (g cm^{-3}), C_i is the initial liquid-phase concentration in the untreated half-cell (g cm^{-3}), L is the total column length (cm), and t is time (s).

Modeling solute diffusion

Measurement of the solute diffusion coefficient in soil is experimentally time-consuming and instrumentally challenging. Many practical constraints surround the half-cell method, including the difficulty in maintaining a continuously replenished reservoir of solutes in the treated cell, and the challenge to maintain a perfect hydraulic contact between the two half cells, particularly in low-saturation conditions (Flury and Gimmi, 2002). Therefore, it is a convenient practice to estimate D_s from easily-measurable properties such as soil moisture content (θ , $\text{m}^3 \text{m}^{-3}$) and total porosity (Φ , $\text{m}^3 \text{m}^{-3}$). Several predictive models are available in the literature including the classical Millington and Quirk (1961) model:

$$\frac{D_s}{D_{o,l}} = \frac{\theta^{10}}{\Phi^2} \quad (7)$$

where $D_{o,l}$ ($\text{m}^2 \text{s}^{-1}$) is the solute diffusion coefficient in free water. Olesen et al. (1996) found that the soil type-independent Millington and Quirk (1961) model often overpredicts the solute diffusion in soil and proposed the following empirical formulation:

$$\frac{D_s}{D_{o,l}} = 0.45\theta \left(\frac{\theta}{\Phi} \right)^{0.3b} \quad (8)$$

where b (dimensionless) is the Campbell pore size distribution parameter which makes the model (Eq. 8) soil type-dependent. Later studies, for example Olesen et al. (2000) and Moldrup et al. (2007), emphasized the presence of a 'percolation threshold' (θ_{th} , $\text{m}^3 \text{m}^{-3}$), a moisture content below which the diffusion ceases due to the presence of disconnected moisture regimes. A more generalized model incorporated with percolation threshold can be written as

$$\frac{D_s}{D_{o,l}} = H\theta(\theta - \theta_{th}) \quad \theta > \theta_{th} \quad (9a)$$

$$\frac{D_s}{D_{o,l}} = 0 \quad \theta \leq \theta_{th} \quad (9b)$$

where H is an empirically determined parameter by fitting the model to the measured data. Olesen et al. (2000) proposed a constant H ($=1.1$) regardless of soil type, whereas Moldrup et al. (2007) correlated H to soil type via clay content of the soil. Along the same line, later studies accounted for additional soil complexities in the model, for example Hamamoto et al. (2009) who empirically linked soil density (ρ_b) in functions for H ($=3.7\text{--}4.0\rho_b$) as well as θ_{th} ($=0.42\text{--}0.27\rho_b$).

A study from Hunt et al. (2014) discussed a theoretical framework for percolation threshold by combining the percolation theory and effective medium approach. The theoretical model includes a critical moisture content (θ_c , $\text{m}^3 \text{m}^{-3}$) which is conceptually analogous to the percolation threshold, and a moisture content at 'crossover point', (θ_x , $\text{m}^3 \text{m}^{-3}$), where transition from percolation to effective medium scaling takes place. The model takes the form of:

$$\frac{D_s}{D_{o,l}} = \frac{1 - \theta}{\theta_x - \theta_c} \left(\frac{\theta - \theta_c}{1 - \theta_c} \right)^2 \quad \theta_c < \theta < \theta_x \quad (10a)$$

$$\frac{D_s}{D_{o,l}} = \frac{\theta - \theta_c}{1 - \theta_c} \quad \theta_x < \theta < 1 \quad (10b)$$

Hunt et al. (2014) emphasized the exponent 2 (in Eq. 10a) as a universal scaling factor in the light of theoretical exposition of conduction properties, while a priori information is required to estimate the critical moisture contents.

Soil-gas diffusion coefficient

Founded on Fick's laws, the soil-gas diffusion coefficient (D_p , $\text{m}^2 \text{s}^{-1}$) characterizes the diffusion of gas in the functional soil gaseous phase. As observed with the solute diffusion, it is generally normalized with the gas diffusion coefficient in free air ($D_{o,g}$, $\text{m}^2 \text{s}^{-1}$) and expressed as soil-gas diffusivity ($D_p/D_{o,g}$), a parameter that can be related only to the porous media characteristics, mainly the air content (ϵ , $\text{m}^3 \text{m}^{-3}$) and tortuosity of the gaseous phase (τ , dimensionless). The tortuosity can be geometrically defined as the ratio between the true length a gas molecule will traverse between two points in space in a given time to the shortest (Euclidean) distance between the two points.

Chamber-based methods are generally used to measure the soil-gas diffusion coefficient with one-chamber or two-chamber approaches. A brief account of the one-chamber approach is mentioned here while descriptive information on both methods can be found elsewhere (Rolston and Moldrup, 2002).

The one-chamber apparatus typically consists of an air-tight chamber, provisioned at one end to mount the soil sample, and a mechanism to isolate the sample from the chamber often by means of a movable plate (Fig. 2). N_2 and O_2 are used as the experimental gas pair, which makes up 99% of the atmospheric composition, and hence constitutes a near-binary system to which Fick's laws can be applicable.

The chamber, isolated from the sample mounted atop with the plate in the closed position (Fig. 2), is initially flushed with N_2 to make the air inside O_2 -free. The movable plate is then removed enabling the atmospheric O_2 to move into the chamber, while N_2 moves out to the atmosphere following their opposite concentration gradients. Using an oxygen sensor mounted on the chamber wall, the change in O_2 concentration is continuously monitored. An analytical solution is used to calculate the soil-gas diffusion coefficient under the predetermined boundary conditions.

Governing equation

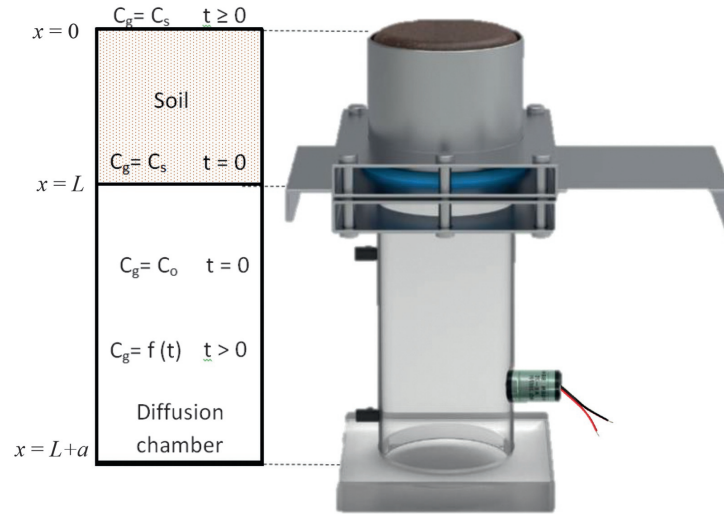


Fig. 2 Schematic of the one-chamber apparatus for measuring soil-gas diffusion coefficient (right) and initial and boundary conditions (left).

Assuming the soil is non-reactive (physically, chemically, and biologically) within the experimental timeframe (i.e., the retardation factor = 1);

$$\frac{\partial c}{\partial t} = \frac{D_p}{\varepsilon} \frac{\partial^2 c}{\partial x^2} \quad (11)$$

Initial and boundary conditions

$$C(x = 0, t \geq 0) = C_s$$

$$C(0 < x < L, t = 0) = C_s$$

$$C(L < x < L + a, t = 0) = C_s$$

$$C(L < x < L + a, t > 0) = f(t)$$

Analytical solution

$$\frac{C - C_s}{C_o - C_s} = \sum_{n=1}^{\infty} \frac{2h \exp(-D_p \alpha_n^2 t / \varepsilon)}{L(\alpha_n^2 + h^2) + h} \quad (12)$$

At some time greater than zero, terms with $n \geq 2$ become negligible with respect to the first term, and the equation reduces to,

$$\frac{C - C_s}{C_o - C_s} = \frac{2h \exp(-D_p \alpha_1^2 t / \varepsilon)}{L(\alpha_1^2 + h^2) + h} \quad (13)$$

where ε the air content ($\text{m}^3 \text{m}^{-3}$), L is the sample height (m), and h is the chamber height (m). Herein, α_1 is the positive root of α in the equation;

$$\alpha L \tan \alpha L = hL$$

The gradient of the C vs. t plot becomes linear in sufficiently large time, and the D_p can be calculated from the gradient $(-D_p \alpha_1^2 / \varepsilon)$. Detailed information can be found in [Rolston and Moldrup \(2002\)](#).

Modeling soil-gas diffusion

Similar to solute diffusion, laboratory measurement of the soil-gas diffusion coefficient is difficult and is often challenged by gas leaks through joints and valves, fine cracks in chamber materials, or misalignments in the sliding plate, just to name a few. Models often offer a quick and convenient alternative, provided that an adequate accuracy can be achieved. They are generally based on easily measurable soil physical properties such as air-filled porosity (ε) and total porosity (Φ).

The pioneering work on modeling soil-gas diffusivity dates back to 1904 when Edgar Buckingham developed a simple yet powerful model (Eq. 14; Buckingham, 1904) which stands even today amid an extensive series of diffusivity models developed over the last century.

$$\frac{D_p}{D_{o,g}} = \varepsilon^2 \quad (14)$$

The power exponent 2 in the Buckingham model makes it principally related to the analogous transport models, and hence provides a universal expression for soil-gas diffusivity ($D_p/D_{o,g}$). A series of power-law models subsequently appeared with different power exponents, e.g., 3/2 (Marshall, 1959), 4/3 (Millington, 1959). The Millington and Quirk (1961) model is presently the most widely used model predicting soil-gas diffusivity, despite not being originally derived for gas diffusion modeling.

$$\frac{D_p}{D_{o,g}} = \left(\frac{\varepsilon^{10}}{\Phi^2} \right) \quad (15)$$

Notably, this model performs better in structureless sands compared to silty and clayey soils. Several conceptual predictive models appeared subsequently (e.g., Troeh et al., 1982; Moldrup et al., 2000). The Troeh et al. (1982) model takes the form of

$$\frac{D_p}{D_{o,g}} = \left(\frac{\varepsilon - u}{1 - u} \right)^v \quad u < \varepsilon < 1 \quad (16)$$

where u and v are fitting coefficients in Eq. (16). The parameter u represents the air content below which gas diffusion ceases due to the disconnected air phase, caused by moisture-induced or compaction-induced pore discontinuity. Eq. (16) is comparable with Eq. (10a) discussed above in relation to solute diffusion. The parameter u represents a gas percolation threshold that can be described within the framework of percolation theory.

Across a broader array of soil textures, the water-induced linear reduction (WLR)-Marshall model (Eq. 17; Moldrup et al., 2000) was found to perform well for sieved-and-repacked soils.

$$\frac{D_p}{D_{o,g}} = \varepsilon^{1.5} \left(\frac{\varepsilon}{\Phi} \right) \quad (17)$$

However, since the conceptual development of the model (Eq. 17) does not account for the properties inherent to undisturbed soils (e.g., soil heterogeneity, layering, soil density variations), the model tends to overpredict the measurements in structurally undisturbed soils. To account for density across a broad range of undisturbed soils, Chamindu Deepagoda et al. (2011) introduced the so-called density-corrected gas diffusivity model:

$$\frac{D_p}{D_{o,g}} = 0.1 \left(2 \left(\frac{\varepsilon}{\Phi} \right)^3 + 0.04 \frac{\varepsilon}{\Phi} \right) \quad (18)$$

On a data-set of 280 intact soil samples, this empirically-based model performed better than existing models, but was still affected by complex soil structure. To improve predictions further, a media complexity factor, C_m was introduced to the WLR-Marshall model (Eq. 17) in the structure-dependent water-induced linear reduction (SWLR) model (Moldrup et al., 2013) (Eq. 19).

$$\frac{D_p}{D_{o,g}} = \varepsilon^{(1 + C_m \Phi)} \left(\frac{\varepsilon}{\Phi} \right) \quad (19)$$

The model provides excellent predictions due to its flexibility to represent different structure status by assigning different C_m values for repacked ($C_m = 1.0$) and undisturbed ($C_m = 2.1$) soils.

A scatterplot comparison of the performance of selected predictive models against undisturbed Danish soil data is illustrated in Fig. 3.

Unified modeling approach for solute and gas diffusion in soil

Considering the strong analogy of diffusive transport processes in soil-aqueous and soil-gaseous phases, both can be described in a unified mathematical framework. Diffusive transport in aqueous and gaseous phases is controlled by the fluid content and the tortuosity of the associated pore network. The analogy between gas and solute diffusion processes for structureless porous media (e.g. sand) has been demonstrated by Hamamoto et al. (2010). However, the gas phase tortuosity was mainly controlled by the fluid-phase (air) content and total porosity (i.e., by soil moisture conditions and soil compaction), while soil type (texture) was an additional major control of the liquid-phase tortuosity. Therefore, for silty and clayey soils, higher soil surface area attributed to higher fines contents reduced water layer coating soil particles, forming more tortuous water-filled pore networks. Unified predictive models combining soil-gas and solute diffusion coefficients have been proposed in past studies (Hamamoto et al., 2010, 2012). Hamamoto et al. (2012) introduced the Liquid and Gas diffusivity (LIGA) model;

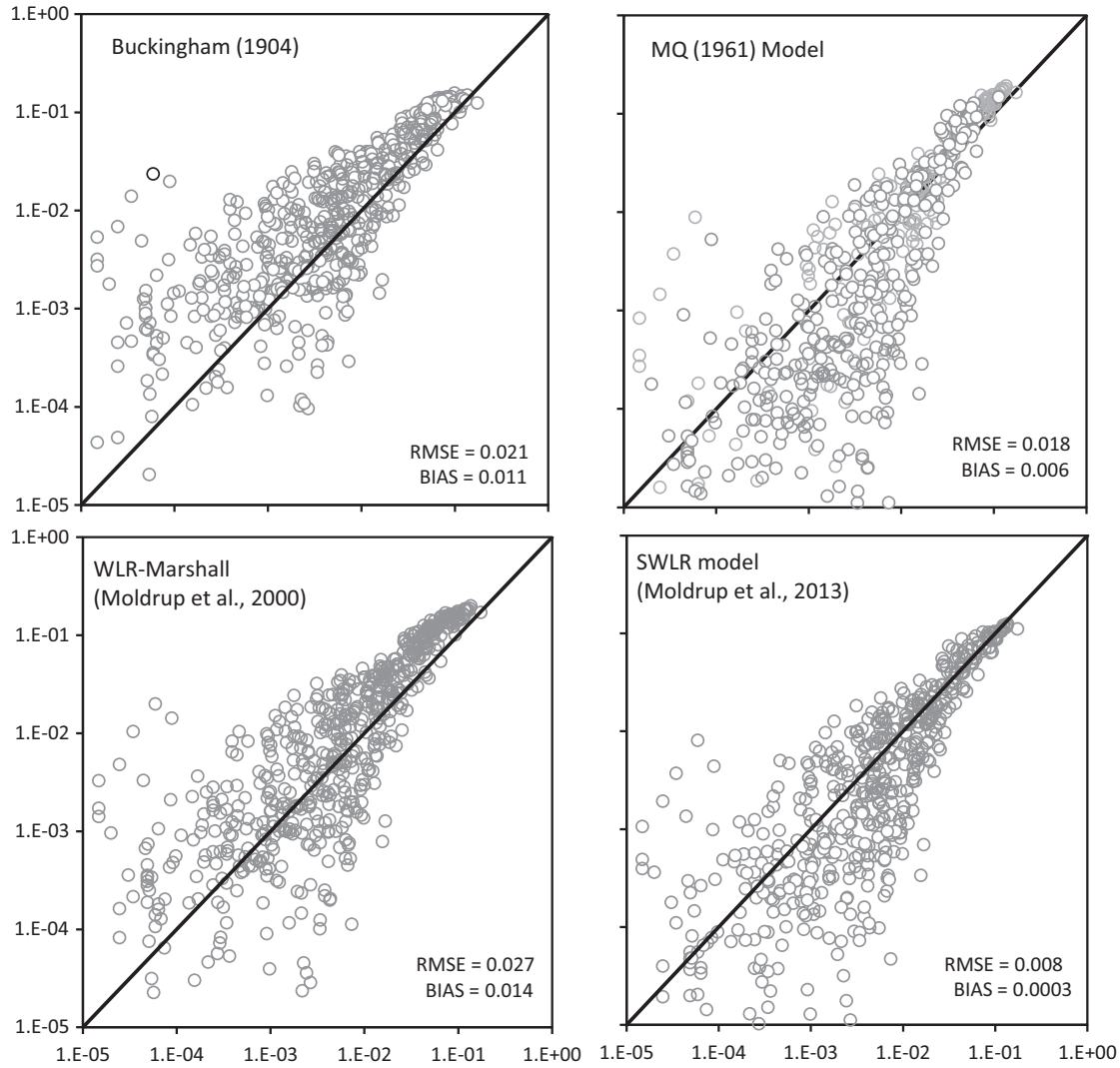


Fig. 3 Scatterplot comparison of four predictive gas diffusivity models tested against 150 undisturbed soil data. MQ, Millington and Quirk; WLR, Water-induced linear reduction; SWLR, Structure-dependent water-induced linear reduction. Modified from Chamindu Deepagoda TTK, Moldrup P, and Schjønning P et al. (2011) Density-corrected models for gas diffusivity and air permeability in unsaturated soil. *Vadose Zone Journal* 10: 226–238.

$$\frac{D_i}{D_{o,i}} = \left\{ \frac{(\kappa - \kappa_{th})}{2 - (\kappa - \kappa_{th})} \right\} \left(\frac{\kappa - \kappa_{th}}{\Phi - \kappa_{th}} \right) \quad (20)$$

where subscript i indicates s for solute diffusion coefficient or g for gas diffusion coefficient. κ is the fluid content defined as a soil–air content (ϵ , $\text{m}^3 \text{m}^{-3}$) for $D_{s,g}$ and soil–water content (θ , $\text{m}^3 \text{m}^{-3}$) for $D_{s,l}$, respectively. κ_{th} is the percolation threshold ($\text{m}^3 \text{m}^{-3}$) defined as;

$$\begin{aligned} \kappa_{th} &= 0.11\Phi^4 \quad \text{for } D_g \\ \kappa_{th} &= P \times \text{FINES}_{vol} \quad \text{for } D_g \end{aligned}$$

where P is a constant, suggested as 0.6 in Hamamoto et al. (2012). FINES_{vol} is defined as the volumetric content of fines (including clay and organic matter). Nevertheless, such attempts to develop unified diffusivity models for soil are still at the preliminary stages and require broader experimental and numerical involvements in the future.

Concluding remarks

Accurate characterization of diffusion-controlled processes in soil–aqueous and soil–gaseous phases is essential to understand the important roles they play in regulating overall soil ecosystem health. It further provides useful implications to develop best

engineering practices to control related environmental problems, for example cleanup of contaminated sites, management of engineered landfills, etc. Further developments in experimental and computational capabilities will be highly beneficial to better characterize diffusion processes in soils and other environmental domains. Future computational attempts may put further emphasis on integrating both solute and gas diffusion modeling into a unified mathematic framework by means of universal scaling parameters.

References

- Bear J (1972) *Dynamics of Fluids in Porous Media*. New York: Dover Publications.
- Bird RB, Stewart WE, and Lightfoot EN (2007) *Transport Phenomena*, 2 edn, United States of America: John Wiley & Son.
- Brown R (1828) A brief account of microscopical observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *The Philosophical Magazine* 21: 161–173.
- Buckingham E (1904) *Contributions to Our Knowledge of the Aeration of Soils*. Washington, D. C: U.S. Department of Agriculture, Bureau of Soils.
- Chamindu Deepagoda TKK, Moldrup P, Schjønning P, et al. (2011) Density-corrected models for gas diffusivity and air permeability in unsaturated soil. *Vadose Zone Journal* 10: 226–238.
- Clifford SM and Hillel D (1986) Knudsen diffusion: The effect of small pore size and low gas pressure on gaseous transport in soil. *Soil Science* 141: 289–297.
- Cussler EL (2009) *Diffusion: Mass Transfer in Fluid Systems*. Cambridge University Press.
- Fick A (1855) V. On liquid diffusion. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 63: 30–39.
- Flury M and Gimmi T (2002) Solute diffusion. In: Dane JH and Topp GC (eds.) *Methods of Soil Analysis: Part 4 Physical Methods*. SSSA Book Series, vol. 5, pp. 1323–1351. Madison, WI: SSSA.
- Hamamoto S, Moldrup P, Kawamoto K, and Komatsu T (2010) Excluded-volume expansion of Archie's law for gas and solute diffusivities and electrical and thermal conductivities in variably saturated porous media. *Water Resources Research* 46: W06514.
- Hamamoto S, Moldrup P, Kawamoto K, and Komatsu T (2012) Maxwell's law based models for liquid and gas phase diffusivities in variably-saturated soil. *Soil Science Society of America Journal* 76: 1509–1517.
- Hamamoto S, Perera MSA, Resurreccion A, et al. (2009) The solute diffusion coefficient in variably-compacted, unsaturated volcanic ash soils. *Vadose Zone Journal* 8: 942–952.
- Hunt AG, Ghanbarian B, and Ewing RP (2014) Saturation dependence of solute diffusion in porous media: Universal scaling compared with experiments. *Vadose Zone Journal* 13(4). <https://doi.org/10.2136/vzj2013.08.0204>.
- Jaynes DB and Rogowski AS (1983) Applicability of Pick's law to gas diffusion. *Soil Science Society of America Journal* 47: 425–430.
- Marshall TJ (1959) The diffusion of gases through porous media. *Journal of Soil Science* 10: 79–82.
- Millington RJ (1959) Gas diffusion in porous media. *Science* 130: 100–102.
- Millington RJ and Quirk JM (1961) Permeability of porous solids. *Journal of the Chemical Society, Faraday Transactions* 57: 1200–1207.
- Moldrup P, Chamindu Deepagoda TKK, and Hamamoto S (2013) Structure-dependent water-induced linear reduction model for predicting gas diffusivity and tortuosity in repacked and intact soil. *Vadose Zone Journal* 12(3). <https://doi.org/10.2136/vzj2013.01.0026>.
- Moldrup P, Olesen T, Blendstrup H, et al. (2007) Predictive–descriptive models for gas and solute diffusion coefficients in variably saturated porous media coupled to pore-size distribution: IV. Solute diffusivity and the liquid phase impedance factor. *Soil Science* 172: 741–750.
- Moldrup P, Olesen T, Gamst J, et al. (2000) Predicting the gas diffusion coefficient in repacked soil: Water induced linear reduction model. *Soil Science Society of America Journal* 64: 1588–1594.
- Olesen T, Moldrup P, Henriksen K, and Petersen LW (1996) Modeling diffusion and reaction in soils: 1V. New models for predicting ion diffusivity. *Soil Science* 161: 633–645.
- Olesen T, Moldrup P, Yamaguchi T, and Rolston DE (2000) Constant slope impedance factor model for predicting the solute diffusion coefficient in unsaturated soil. *Soil Science* 166: 89–96.
- Rolston DE and Moldrup P (2002) Gas diffusivity. In: Dane JH and Topp GC (eds.) *Methods of Soil Analysis: Part 4*. SSSA Book Series 5, pp. 1113–1139. Madison, WI: SSSA.
- Siddiqui MAQ, Salvemini F, Ramandi HL, Fitzgerald P, and Roshan H (2021) Configurational diffusion transport of water and oil in dual continuum shales. *Scientific Reports* 11: 2152. <https://doi.org/10.1038/s41598-021-81004-1>.
- Troeh FR, Jabro JD, and Kirkham D (1982) Gaseous diffusion equations for porous materials. *Geoderma* 27: 239–253.
- Unnithan A, Bekele DN, Chadalavada S, and Naidu R (2021) Insights into vapour intrusion phenomena: Current outlook and preferential pathway scenario. *Science of the Total Environment* 796: 148885.

Hydrodynamics in soil

Ty Ferre^a and Art Warrick^b, ^aHydrology and Atmospheric Sciences, University of Arizona, Tucson, AZ, United States; ^bEnvironmental Science, University of Arizona, Tucson, AZ, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of T.P.A. Ferré, A.W. Warrick, HYDRODYNAMICS IN SOILS, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 227-230, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00377-5>.

Introduction	257
Steady-state flow through a saturated medium	257
Transient flow through a saturated medium	258
Steady-state flow through an unsaturated medium	260
Transient flow through an unsaturated medium	261
Conclusions	261
Further reading	261

Abstract

The movement of water in the subsurface, “Hydrodynamics,” touches many areas of the natural and human-influenced world. Ecosystems are shaped, contaminants spread, valleys flood, and slopes fail, all under the influence of hydrodynamic processes. These processes can be divided into four main subtopics based on whether flow is constant or changing in time and on whether the pores are or are not entirely filled with water. While these conditions share underlying concepts, there are specific considerations that make each unique.

Glossary

Richardson-Richards’ Equation A simplification of the coupled flow equations describing the movement of air and water through the same medium based on an assumption that air is infinitely mobile. This is commonly applied for soil physics and hydrology problems.

Saturated A condition under which all of the pores are filled with water. In some cases, immobile, trapped air is allowed and the condition is deemed “effective saturation.”

Steady-state A condition under which there are no changes in the state of the system with time; this includes both hydrostatic and non-zero flow conditions.

Storage The volume of water present in a given volume of porous medium.

Transient A condition under which there is a change in state of the system at any location; this only includes flowing conditions, but it may be limited to internal redistribution.

Unsaturated A condition under which the pores are filled with more than one fluid, often air and water.

Dynamic viscosity Dynamic viscosity is the resistance of internal shear of a fluid.

Flux The flow of a liquid per unit cross sectional area perpendicular to flow, [m s^{-1}].

Hydraulic conductivity The ease with which water can move through a porous medium under a given gradient, which depends on the pore structure, the fluid properties, and the volumetric water content, [$\text{m}^3 \text{m}^{-3}$].

Permeability The ease with which any fluid can move through a porous medium under a given gradient, which depends on the pore structure and the volumetric content of the fluid, [$\text{m}^3 \text{m}^{-3}$].

Pressure head The gage (relative to atmospheric) pressure of water divided by the weight of water, [m].

Specific storage A property of the medium and the fluid that describes the ability of a unit volume of medium to change the volume of stored water, while remaining fully saturated, per unit change in pressure head.

Specific yield A property of the medium and the fluid that describes the ability of a unit area (plan view) of medium to change the volume of stored water by drainage per unit change in pressure head at every depth.

Key points

- Flow occurs in response to an applied energy gradient in space;
- Steady-state flow through a homogeneous, saturated medium leads to a linear trend in hydraulic head with distance;
- The primary storage mechanism in a saturated medium, medium compaction, provides little storage capacity, so transient changes propagate relatively quickly;
- Steady-state flow through an unsaturated porous medium may give rise to changes in water content with distance, which leads to spatial variation in hydraulic conductivity, resulting in nonlinear trends in hydraulic head with distance; and

- Propagation of applied changes via transient flow through a variably saturated medium can be significantly more delayed than through a saturated medium because of the larger storage changes associated with changes in water content with time.

Nomenclature

K	Hydraulic conductivity [m s^{-1}]
k	Permeability [m^2]
q	Flux
Ss	Specific storage
Sy	Specific yield
μ	Dynamic viscosity ($\text{kg m}^{-1} \text{s}^{-1}$)
ψ	Pressure head [m]

Introduction

Hydrodynamics is, literally, the movement of water. Whereas hydrostatics can describe the distribution of water in a medium when it is not flowing, hydrodynamics describes how water flows and how that flow changes in time. As such, an understanding of hydrodynamics is fundamental to topics ranging from plant growth, to soil conservation, to construction, to weather prediction. Fluid flow always occurs due to a spatial energy gradient. However, the specific conditions give rise to very different descriptions of flow. As a result, soil-water dynamics can be subdivided into flow through a water-saturated or unsaturated medium, and under steady-state (constant in time) or transient conditions. Groundwater flow problems typically consider only saturated flow. Flow through the vadose zone requires study of flow through variably saturated media.

Steady-state flow through a saturated medium

The Darcy equation describes the response of a fluid in a porous medium to an imposed energy gradient. This simple formula, equivalent to Ohm's Law for electrical flow, forms the foundation of discussions of hydrodynamics. The one-dimensional form of the Darcy equation can be written as:

$$Q = -k \frac{\rho g}{\mu} A \frac{d(\psi + z)}{dL} \quad (1)$$

where Q is the flow ($\text{m}^3 \text{s}^{-1}$), A is the cross-sectional area perpendicular to flow (m^2), k is the intrinsic permeability (m^2), ρ is the fluid density (kg m^{-3}), g is the acceleration due to gravity (ms^{-2}), μ is the dynamic viscosity of the fluid ($\text{kg m}^{-1} \text{s}^{-1}$), ψ is the pressure head (m), z is the elevation (m), and L is the distance in the direction of flow (m) (note that the "head formulation" is used commonly to describe water flow. This is defined as the potential energy per unit weight of fluid. It is used widely because its units, length, are convenient for both laboratory and field applications. In the formulation, the influence of gravity is referred to as the elevation head and it acts only in the vertical direction.) The negative sign is necessary due to the definition of a gradient to define flow as occurring from a region of higher energy to lower energy. This can be rewritten in terms of the flux, $q(\text{ms}^{-1})$ as:

$$q = \frac{Q}{A} = -k \frac{\rho g}{\mu} \frac{d(\psi + z)}{dL} \quad (2)$$

Flow occurs in the direction of decreasing energy of the fluid, and the rate of flow depends directly upon the ability of the medium to permit flow (permeability), the fluid density, and the acceleration due to gravity, and inversely upon the resistance of the fluid to internal deformation (dynamic viscosity). Introducing the hydraulic conductivity, $K (\text{ms}^{-1})$ allows simplification to:

$$Q = -KA \frac{d(\psi + z)}{dL} \quad (3)$$

Considering flow in the horizontal, x -direction, the flux density is:

$$q_x = -K \frac{d(\psi + z)}{dx} = -K \frac{d\psi}{dx} \quad (4)$$

This indicates that horizontal flow is driven only by pressure head gradients. The flux density for flow in the vertical, z -direction, is:

$$q_z = -K \frac{d(\psi + z)}{dz} = -K \left(\frac{d\psi}{dz} + 1 \right) \quad (5)$$

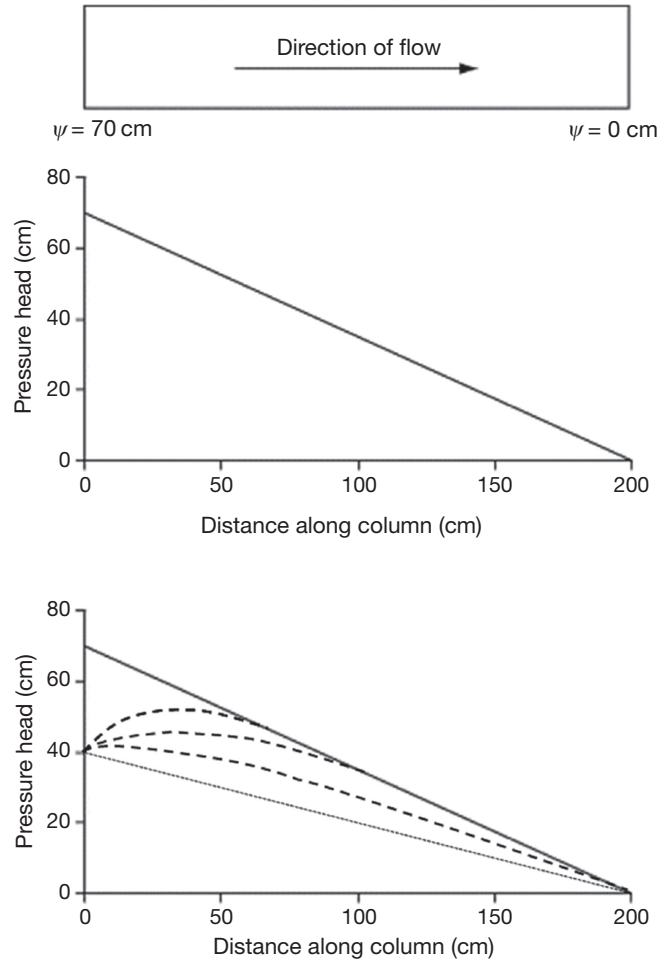


Fig. 1 Pressure-head profiles along a horizontal column during steady-state saturated flow. Note that the hydraulic head is equivalent to the pressure head if we assume that the center height of the column is located at $z = 0$.

This shows that vertical flow is driven by both gravity and pressure gradients. Note further that the hydraulic conductivity may be different in the horizontal and vertical directions, for example due to layering caused by depositional processes (known as anisotropic conditions).

As a baseline, it is useful to know how the pressure head, elevation head, and hydraulic head are distributed within homogeneous systems (K does not vary with location) undergoing steady state flow. Fig. 1 shows a horizontal column of uniform cross-sectional area. The elevation head is constant everywhere. The pressure and hydraulic head gradients must be constant with x location to achieve steady state conditions. If this were not the case, then there would be more flow into some regions than out of those regions, which would cause the amount of water in storage to change with time, violating the steady state assumption. This is shown for a 200-cm-long, homogeneous horizontal column for which the left boundary is maintained at a constant pressure head of 70 cm and the right boundary is held at a constant pressure head of 0 cm (top panel, Fig. 1).

The two-dimensional form of the Darcy equation follows directly from the one-dimensional form. For instance, two-dimensional flow in the x - z plane is described as:

$$q_i = -K_{ij} \frac{d(\psi + z)}{dx_j} \quad (6)$$

This equation includes the term K_{ij} , which can be read as “the conductivity in the i -direction due to a gradient in the j -direction.” The derivative term describes the head gradient in the j -direction.

Transient flow through a saturated medium

It can be confusing to learn that a medium can be saturated while also experiencing a change in water stored with time. This occurs through two transient processes in response to a change in the water pressure: a change in the porosity, and a change in water density. The former occurs because water that resides in a medium under a positive pressure head acts to expand the matrix of the

medium, essentially pushing the soil grains apart. If the water pressure is reduced, the effective stress on the medium will increase and the medium will compact. By compacting, the porosity of the medium is reduced as water is released from storage while maintaining full water saturation. Simultaneously, the decreased pressure of the water allows the water itself to expand, thereby occupying more volume per unit mass of water. However, this expansion of water is negligibly small for most water flow problems. These two processes combined are referred to as the specific storage of a medium. This lumped property is critical to understanding the transient response of a saturated porous medium.

To describe transient flow in a water saturated medium, we begin with Eq. (6), which describes flow in response to a gradient. Consider the three-dimensional region shown in Fig. 2. Each edge has a constant length, ΔL . Flow is two-dimensional in the x - z plane. At steady state the sum of q_x in and q_z in will be equal and opposite to the sum of q_x out and q_z out, and there won't be any change in water stored with time. However, under transient flow conditions, these sums will not be equal. The net difference in the inflow and the outflow will equal the change in the volume of water stored, ΔV_w , per unit time elapsed, Δt , per unit area of the region. This is written as a mass balance equation:

$$dQ_x + dQ_z = \frac{dV_w}{dt} \quad (7)$$

For simplicity, we assume that all sides of the box have the same size, $dx = dy = dz$. Dividing by the volume of the box gives:

$$\frac{dq_x}{dx} + \frac{dq_z}{dz} = \frac{dV_w}{(dxdydz)dt} \quad (8)$$

We can then describe the change in volume of water per volume of the box as the change in porosity as a function of the change in pressure head, which is approximately equal to the specific storage, S_s [L^{-1}]. This leads to the flow equation:

$$\int \frac{dq_i}{dx_i} dx_i = S_s \int \frac{d\psi}{dt} dt \quad (9)$$

In words, this states that the divergence of the flow is equal to the change in storage with time.

Given the flow equation, consider what would happen in the column shown in Fig. 1 if the pressure head on the left boundary was increased instantaneously. The gradient would increase at the left boundary, causing more water to flow into the column. The increase in pressure would increase the water storage, so some of this increased flow would be added to storage and less of it would remain to flow farther to the right. As a result, the storage mechanism acts to slow the propagation of the change in pressure head imposed at the right boundary. Eventually, a new steady state condition would be established with a steeper left-to-right gradient and more water stored within the column. In reality, the storage capacity of a saturated medium is very small, so the transient period would be relatively short. If these processes are scaled up to field-relevant distances, however, the delays due to transient flow can last for days, months, or even centuries!

To visualize transient, saturated flow, consider a transition between two steady-state conditions (Fig. 1, bottom panel). The solid line shows the steady state conditions discussed earlier. The lowest, dotted line relates to steady state flow with a lower head on the left boundary. The dashed lines show the transition between these two steady state conditions. Note that the change occurs at the left boundary first (the topmost dashed line) and then propagates farther to the right with increasing time. At early time, there is no change in the outflow from the right boundary, even though the left boundary condition has changed. The transient profiles are nonlinear, even for saturated flow, indicating that flow into and out of any volume of porous medium are not equal. This leads to a change in storage, and therefore a change in the head, with time.

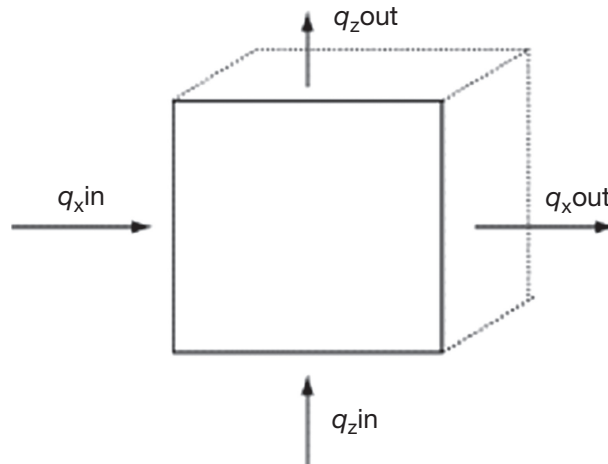


Fig. 2 Three-dimensional unit cell for consideration of conservation of mass. The flux density into (q in) and out of (q out) the cell is shown in the x and z directions only for clarity.

Steady-state flow through an unsaturated medium

The Darcy equation is equally applicable to steady-state flow through both saturated and unsaturated media. However, there is a critical difference when this relationship is applied to unsaturated flow. Specifically, the hydraulic conductivity (or intrinsic permeability) of the medium varies with the pressure head of the water in the medium. More accurately, the hydraulic conductivity depends on the water content (the fraction of the pores that are filled with water), because water only flows through water-filled pores under steady state conditions. As the water pressure decreases below atmospheric pressure, air can invade soil pores. Due to capillarity, the largest pores allow air to enter at higher water pressures. As a result, these pores, which contribute the most to the overall hydraulic conductivity, drain at higher (less negative) pressure. This leads to a very rapid decrease in hydraulic conductivity as a medium transitions from full to partial saturation. The equation describing unsaturated flow is deceptively simple:

$$q_i = -K_{ij}(\theta) \frac{d(\psi + z)}{dx_j} \quad (10)$$

where θ is the volumetric water content $[-]$, defined as the volume of water in a given total volume of medium.

The ramifications of the dependence of the hydraulic conductivity on the volumetric water content can be quite dramatic. For example, consider the case of steady-state, one-dimensional, horizontal flow through a homogeneous medium, such as shown for saturated conditions in Fig. 1. The pressure-head profile is shown in Fig. 3 for a 200-cm-long, sand-filled column with the left boundary held at a constant pressure head of 0 cm and the right boundary held at a constant pressure head of -70 cm. Notice that, unlike the case of steady-state one-dimensional flow through a saturated medium, the pressure head varies nonlinearly in the direction of flow. This occurs because the hydraulic conductivity decreases steadily with decreasing pressure head from the saturated value at the left boundary to a much lower value at the right boundary. To maintain steady-state flow throughout the column, the magnitude of the gradient must increase to compensate for this decreasing hydraulic conductivity. Therein lies the difficulty of calculating the pressure-head distribution during unsaturated flow. Even under the apparently simple conditions of steady-state, one-dimensional horizontal flow, the pressure-head distribution depends on the hydraulic conductivity distribution, which depends on the pressure head distribution. This “circular” dependence describes a nonlinear condition.

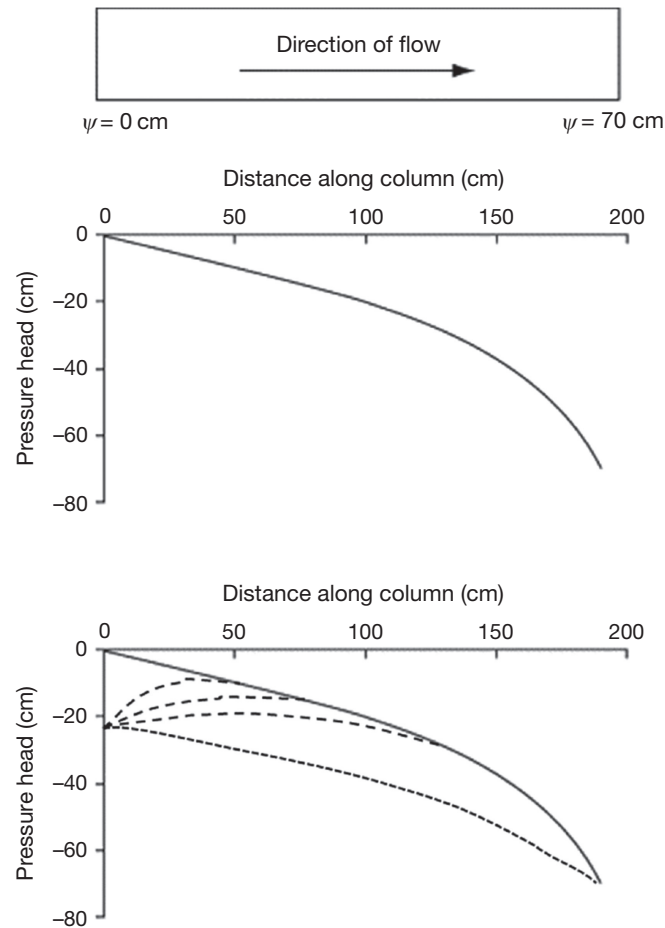


Fig. 3 Pressure-head profiles along a horizontal column during steady-state unsaturated flow.

Transient flow through an unsaturated medium

Transient unsaturated flow is fundamentally different than saturated flow and steady-state unsaturated flow. During transient unsaturated flow, water enters pores that were previously occupied by air (or air enters previously water-filled pores). If we assume that the air can move easily, such that it does not impede the movement of water, the flow of water can be defined using only one equation—the Richardson-Richards' equation (Eq. 10). In addition to the nonlinear nature of unsaturated flow, as shown in Fig. 3, the mechanism of storage results in far greater storage capacity. That is, you can change the volumetric water content by much more than you can change the porosity by changing the water pressure. Remembering that changes in storage act to slow the propagation of a change applied to a system (e.g., the change in the pressure at the left boundary discussed above), we can expect that transient conditions have much more impact on flow under unsaturated conditions.

Transient, unsaturated flow appears qualitatively similar to transient saturated flow (Fig. 3, bottom panel). Both steady state profiles are nonlinear, as discussed above. But, as for saturated conditions, the change occurs at the left boundary first (the topmost dashed line) and then propagates farther to the right with increasing time. There is a similar period during which there is no change in the outflow from the right boundary, even though the left boundary condition has changed. Because of the storage mechanism, filling and draining of the pores, leads to much larger changes in storage than the corresponding mechanisms under saturated conditions, transient conditions propagate more slowly through unsaturated soils. This increased time is magnified by lower hydraulic conductivity of the partially saturated soils.

Conclusions

The flow of water through porous media is critical to many fields of study in natural science and engineering. A basic understanding of the underlying processes, such as is presented here, is relatively accessible. But, understanding how water moves through a complex, transient system requires experience and the help of computer models. Specific models are constructed to examine how multiple fluids move and interact at the scale of a few soil pores. Other models track water movement from the atmosphere to the deep subsurface at the scale of the entire Earth. Still other models are built to predict the impacts of changes in irrigation practices, water extraction for municipal use, or land cover change at every intermediate scale. Each of these applications of hydrodynamics requires supporting data and nuanced understanding of water movement and storage to fully capture the many roles of water in the environment.

Further reading

- Cohen A and Cherry J (2020) *Conceptual and Visual Understanding of Hydraulic Head and Groundwater Flow*. The Groundwater Project. <https://books.gw-project.org/conceptual-and-visual-understanding-of-hydraulic-head-and-groundwater-flow/>.
- Darcy H (1856) *Les Fontaines Publique de la Ville de Dijon*. Paris, France: Dalmont.
- Gardner WR (1958) Some steady state solutions of the unsaturated moisture flow equation with application to evaporation from a water table. *Soil Science* 85(4): 228–232.
- Pinder GF and Celia MA (2006) *Subsurface Hydrology*. John Wiley and Sons <https://doi.org/10.1002/0470044209>.
- Richards LA (1931) Capillary conduction of liquid through porous media. *Physics* 1(1931): 318–333.

Thermodynamics of soil water

Pieter H Groenevelt[†], University of Guelph, Guelph, ON, Canada

© 2023 Elsevier Ltd. All rights reserved.

This is an update of P.H. Groenevelt, Thermodynamics of Soil Water, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 163–168, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00429-X>.

Introduction	262
Classical (equilibrium) thermodynamics of soil water	263
The tensiometer (and the pressure membrane apparatus)	264
The Maxwell relations and their use in the hydrostatics of swelling soils	264
Non-equilibrium thermodynamics of soil water	265
L_{TV} (thermo-filtration) and L_{VT} (thermo-osmosis)	266
L_{TD} and L_{DT} (thermo-diffusive processes)	267
L_{TE} and L_{ET} (thermo-electric processes)	267
L_{VD} (osmosis, also called capillary osmosis) and L_{DV} (reverse osmosis or salt sieving)	267
L_{VE} (electro-osmosis) and L_{EV} (streaming current)	267
L_{DE} (electro-phoresis) and L_{ED} (diffusion current)	267
The “linearity” argument	267
The thermodynamic versus the mechanistic theory	268
Modern development and outlook	268
References	268

Abstract

The fundamentals of thermodynamics are applied to soil water. The first part deals with equilibrium (classical) thermodynamics. The essential point is the mathematical expression for the chemical potential of soil water. The second part deals with non-equilibrium (irreversible) thermodynamics. The key element is the development of a mathematical expression for the total energy dissipation (entropy production), taking into account all possible transport processes (entropy producing phenomena) occurring in the chosen “system.” This expression is then used to construct transport equations for all those transport processes. Each transport equation contains a flux as a linear function of all possible driving forces. In this scheme all possible Onsager Reciprocal Relations are exposed in the symmetric matrix of transport coefficients.

Key points

- Thermodynamics applied to soil physical processes of heat, solute and mass transfer is described.
- The article builds from the origins of thermodynamics as a discipline to its application to soil science.
- Processes described for soil occur not far from equilibrium, such as Darcy’s Law, but some catastrophic processes in soil may benefit from recent developments in Non-linear Non-equilibrium Thermodynamics.

Introduction

The word thermodynamics is derived from the ancient Greek words for heat, $\theta\epsilon\rho\mu\omicron\varsigma$, and force or power, $\delta\upsilon\nu\alpha\mu\iota\varsigma$. It presents the science of all forms of energy and mass, including entropic (“waste”) heat contained in the mass at ambient temperature. The origin of the word suggests that this branch of science deals with both the statics (equilibrium) and the dynamics (non-equilibrium) of energy and mass. Statics (“Classical Thermodynamics”) deals with the state of the system in which no heat or mass transfer occurs (this state is extremely rare in natural soils). Dynamics (“Non-equilibrium Thermodynamics”) deals with transport processes of mass and heat. “Soil water” is often used interchangeably with the term “soil moisture.” Here we distinguish between the two terms:

[†] Author deceased.

“soil water” indicating the chemical component H_2O in the soil and “soil moisture” meaning the soil solution. Thermodynamics distinguishes the different chemical components in the system. It deals with the interaction between and transfer of these components and heat.

Classical (equilibrium) thermodynamics of soil water

The birth of thermodynamics is rather confusing. The founders tried to deal with such things as the effectiveness of steam engines (S. Carnot). R Meyer, who was the first (1842) to publish the equivalence of work and heat, based his conclusion on his observations of the degree of “redness” of human blood. No wonder that thermodynamics had a shaky start. The first mathematical formulation of the “First Law of Thermodynamics” (the law of conservation of energy) came from Helmholtz and Joule in 1847. All these scientists were dealing with systems, which were “on the move.” Yet the system had to be at equilibrium and the “changes” had to be infinitesimally small and “reversible” (meaning that the changes can be reversed without any loss of useful energy at the ambient temperature, or entropy production). An excellent overview of these matters can be found in Pippard’s book (Pippard, 1964).

One would expect the First Law of thermodynamics to equate integral values of the different forms of energy. Yet the equations always come in differential form first. At first only “closed” systems were considered, implying that no mass could move into or out of the “system.” It was not until J. Willard Gibbs, 1876, 1878 published his two great treatises “The Equilibrium of Heterogeneous Substances” (Gibbs, 1876, 1878), that the “system” was opened up and many of the mysteries of the previous 50 years were clarified. The Transactions of the Connecticut Academy were not widely read in Europe and it was not until G.N. Lewis published his famous book in 1923, 20 years after Gibbs’ death, that his works became widely known. In the year of his retirement, Gibbs remarked that during the 30 years of his teaching at Yale University, he estimated that only half a dozen of his students had benefited from his lectures (Gibbs, 1961), which were published some 58 years after his death.

Physicists, like Maxwell and Einstein, immediately recognized the intrinsic value of Gibbs’s work. Chemists and, in particular, scientists in applied disciplines were much slower to accept and understand Gibbs’s work. In the field of Soil Science, the pioneers were Edlefsen and Anderson (1943). Closer to home, it was Cary (1963, 1964) who brought Irreversible Thermodynamics into Soil Physics. Cary concentrated his work on the coupled transport process called thermo-osmosis.

We start with the integrated Gibbs Equation:

$$E = U + \Psi = TS - PV + \sum_i \mu_i m_i + \sum_i \psi m_i \quad (1)$$

where E is the energy of the system, U is the “internal” energy, and Ψ is the sum of the “free” energy contained in the components of the system and the energy derived from external force fields, e.g. the gravity force ($\psi = gh$).

U consists of three “blobs” of different forms of energy, the first one (TS) being the entropic energy (heat in the system at the ambient temperature, T), the second one ($-PV$) being the energy the system has lost in order to create a volume, V , for itself to exist at the ambient pressure, P , the third one being the sum of the chemical energy of each component (i) in the system, with m_i being the mass of component i and μ_i being the chemical potential of component i .

Ψ is derived from external force fields, such as gravity or an electrical or magnetic field (in the gravity field $\psi = gh$, g being the gravitational acceleration and h the height above the reference level). If other forms of energy (such as kinetic energy or viscous stress energy) play a role in the transformations of energy, they should be added to Eq. (1). Although kinetic and viscous stress energy play a dominant role in energy transformations in the soil solution on the Navier-Stokes scale, on the Darcy scale they are almost always negligible, and will therefore be ignored here.

“Gibbs” energy, G , is defined by

$$G \equiv U - TS + PV = \sum_i \mu_i m_i \quad (2)$$

Thus, μ_i is the specific (per unit mass) Gibbs free energy of constituent i .

The first expressions of the First Law were all in differential form, even after Gibbs “opened up” the system:

$$dU = TdS - PdV + \sum_i \mu_i dm_i \quad (3)$$

When Eq. (3) is combined with the total (mathematical) derivation of Eq. (2), the result is:

$$SdT - VdP + \sum_i m_i d\mu_i = 0 \quad (4)$$

This “left-over” is known as the Gibbs-Duhem equation.

From Eq. (4) it follows that

$$d\mu_w = -Sm_w^{-1}dT + Vm_w^{-1}dP - \sum_{(k \neq w)} m_k m_w^{-1} d\mu_k \quad (5)$$

The last term of Eq. (5) represents the concentration dependent part of $d\mu_w$.

This is the central equation in equilibrium thermodynamics of soil water (Groenevelt and Bolt, 1969). The pressure term contains all surface tension effects. This expression is used to formulate the driving force in Darcy’s law and subsequently the non-conjugated driving forces in the thermodynamic transport equation for the soil solution.

The guiding principle is now, that when the system is in equilibrium, $d\mu_w = 0$, or, the specific Gibbs free energy of water is the same everywhere, or, as far as water is concerned, there is no transport and no entropy production. In short: the water is at rest. The inverse statement: “when $d\mu_w = 0$, the water is at rest” is not necessarily true and is not a fundamental principle of thermodynamics. It is possible that the last two terms of the RHS of Eq. (5) (temporarily) balance each other, making $d\mu_w = 0$, but both water and the solute are on the move, and entropy is being produced. This is possible when “leaky” semi-permeable membranes are present. Heavy clay soils are excellent examples of “leaky” semi-permeable membranes.

Therefore, the concept of the “total” potential of the soil water (the sum of all the component potentials) is not very useful. When the whole system is at rest (complete equilibrium), then the total potential of the water is equal everywhere in the system.

But, when the total potential of the water is equal everywhere, this does not necessarily mean that the water is at rest. The gradient of the total potential is not the appropriate driving force in the flux equation of water. This will become clearer in the section on non-equilibrium thermodynamics.

The apprentice in thermodynamics often asks him/herself: “What is the practical use of all this?” Thermodynamics is still not a coherent, comprehensive science. Developments are still happening haphazardly. Some scientists, like the American mathematician Clifford Truesdell, complained bitterly about the inconsistencies and the lack of mathematical rigor, using words like “thermodynamics was approached through detours in the fog of word-play.”

Here, only two of the practically useful results will be discussed:

The tensiometer (and the pressure membrane apparatus)

Early scientists, working on the thermodynamics of soil moisture, such as L. A. Richards, G. H. Bolt and M. J. Frissel (1960), immediately realized that water in the soil is present in the soil solution and that the soil solution is never pure water. The soil solution contains dissolved salts and ions. In addition, it contains counter-ions, that is a surplus of cations, necessary to counter-balance the surplus of electrons present in the lattices of clay particles. The electrical double layer acts as an osmometer and the counter-ions cause a volume element of the solution, close to a clay surface, to carry a volume charge. This gives the soil solution an additional opportunity to possess energy (in the presence of an electrical field). In addition to the above, there are micro-force fields acting on the solution that is in close proximity of solid surfaces (London-Van der Waals forces). All this implies that Eq. (5) should contain a few more terms if the thermodynamic system were chosen on the Navier-Stokes scale.

The first requirement in Soil Physics is to formulate and assess the hydrostatics of soil water, i.e., establish the characteristic relationships when the water is at rest. There are no “membranes” in nature or in the laboratory that are semi-permeable to heat. Thus, when a system is brought to equilibrium the temperature will be the same everywhere. The energy level in the external force fields, such as gravity, can easily be established separately. To measure the energy level of soil water derived from internal force fields (menisci, electrical double layers, London-Van der Waals forces), it is now advantageous to bring the soil solution in equilibrium with either a compartment containing a solution that reflects the composition of the soil solution or one that contains pure water. There are membranes that are (“perfectly”) semi-permeable to all dissolved materials in the soil solution. In the latter case (using a perfect osmometer and measuring the hydrostatic pressure in the osmometer with respect to atmospheric pressure), one measures the energy level of the soil water at the ambient temperature and at the particular position in the external force field, resulting from all internal force fields plus all dissolved materials (the total osmotic pressure of the soil solution). In the former case, using a perfectly “leaky” membrane, such as a sintered ceramic cup (tensiometer cup) or a Visking membrane in a pressure plate apparatus, the solution inside the compartment (tensiometer cup) will reflect the composition of the “free salts” in the soil solution and will show the “equilibrium” solution. By measuring the hydrostatic pressure in a tensiometer with respect to atmospheric pressure, one measures the energy level of all internal force fields (including electrical double layers), but excluding the free salt. This energy level is called the matric potential. The counter-ions cannot equilibrate in the tensiometer cup. They act as internal mini-osmometers and can be considered part of the matrix (the solid phase of the soil). Because most soils are pretty “leaky” as far as the exclusion of dissolved materials is concerned, the water in the soil moves largely as a solution. It is therefore the gradient of the matric potential that is the dominant driving force on the soil water (in addition to the force of the gravitational field). The imperfect semi-permeability of a clay soil, causing osmotic pressure differences to drive the water, will be dealt with in the section “Non-equilibrium thermodynamics” (see below). The action of osmotic pressure gradients will then be formulated as “capillary” osmosis and the imperfect semi-permeability will be evaluated by the “reflection coefficient.”

The Maxwell relations and their use in the hydrostatics of swelling soils

Before entering the field of non-equilibrium states (entropy production), it is useful to pay attention to the Maxwell relations. Here, one is crossing the border area between classical and irreversible thermodynamics. In equilibrium thermodynamics, the temperature and pressure were allowed to change but only in very small steps, δT and δP . Entropy production was not allowed. It began to look like differentiation, but differentiating with respect to what? Time or space? Once the independent variable is chosen (i.e., space), the driving forces for the transport processes can be formulated. Applying cross-differentiation in the matrix of driving forces, one obtains the Maxwell relations.

As a prime example of the use of a Maxwell relation we present here the following. For many decades soil scientists have searched for an expression for the differential of the matric potential (tensiometer pressure), p , with respect to the load (or overburden pressure), P , at constant ratio of mass of water and mass of solids. Here, both the dependent and the independent variable are

intensive variables. The appropriate Maxwell relation shows that the above differential is equal to the differential relationship between the void ratio (volume of voids per volume of solids), e , and the moisture ratio (volume of water per volume of solids) at constant load pressure, P . Here, both the dependent and the independent variables are extensive variables (Groenevelt and Bolt, 1972).

Thus:

$$dp(dP)^{-1} \text{ at constant } m_w(m_{\text{solids}})^{-1} = de(d\vartheta)^{-1} \text{ at constant } P \quad (6)$$

where p = the tensiometer pressure, P = the load pressure, e = the void ratio and ϑ = the moisture ratio (volume of water per volume of solids) at constant load pressure, P .

This means that one can predict the behavior of the tensiometer reading upon loading from the slopes of the shrinkage lines for different values of P . Because the shrinkage lines for different values of P are nearly parallel, one only needs to measure the unloaded shrinkage line to predict how the tensiometer reading changes when moving deeper into the soil profile or upon loading by machinery.

Non-equilibrium thermodynamics of soil water

The thermodynamic “system” is now chosen to be a unit volume element in the soil solution. All the extensive variables are expressed per unit volume. Subsequently, each of the terms of the differential form of the Gibbs equation (Eq. 3) are written as differentials with respect to time.

The mathematical sloppiness of classical thermodynamics disappears once the terms of the differential form of the Gibbs equations are transformed into proper time differentials and one enters the realm of non-equilibrium thermodynamics.

Next, for each of the terms separately, an appropriate conservation (continuity) equation is constructed. The general form of such a conservation equation shows that the change of the content of the entity of concern, with time, is equal to the negative of the divergence of the flux of that entity (inflow minus outflow), plus or minus one or more source or sink terms. Thus, for the entropic energy term one writes:

$$d(SV^{-1}) dt^{-1} = -\text{div } j_s + \sigma \quad (7)$$

where SV^{-1} is the entropy per unit volume of solution; j_s is the flux of entropy and σ is the entropy production term. This latter term plays a central role in non-equilibrium thermodynamics. If all the processes and forms of energy are accounted for, and nothing is counted twice, and all algebra is carried out correctly, then the entropy production term is non-negative. As the absolute temperature is always positive, the energy dissipation term, $T\sigma$, must also be non-negative. When σ , and thus $T\sigma$, is zero, the system is at equilibrium and nothing moves. If anything moves, then entropy is produced, and σ and $T\sigma$ are positive. Energy is being dissipated. That is, energy is transformed from useful energy to waste energy (= heat at the ambient temperature). This is the “Second Law of Thermodynamics.”

As the first term on the RHS of Eq. (3) is already multiplied by T , one finds, after replacing the time differentials of the different terms in Eq. (3) by the RHS of their appropriate conservation equations, the product $T\sigma$ as the source term of the entropy conservation equation. After all the appropriate substitutions are made, the term $T\sigma$ is singled out (written explicitly) and the equation is then called “the (energy) dissipation function.”

On the RHS of the dissipation function one usually finds the sum of a number of products of fluxes and forces. One may reshuffle these forces and the fluxes, under the strict principles that the rules of algebra are carried out correctly and that nothing is forgotten or counted twice.

The resulting complete expression for the energy dissipation is then the sum of products of fluxes and conjugated forces.

$$T\sigma = -j_q \text{ grad } T - j^V \text{ grad } H - j^D \text{ grad } \pi - j^E \text{ grad } E \quad (8)$$

where j_q is the caloric (Fourier) heat flux; j^V is the (Darcy) flux of the soil solution; j^D is the (Fick) diffusion flux and j^E is the electric current (usually indicated by capital I).

$\text{grad } \pi$ is the gradient of the osmotic pressure and $\text{grad } E$ is the gradient of the electrostatic potential.

One recognizes the primary forms of dissipation, viz. the dissipation of heat energy, pressure energy, mixing or unmixing energy, and electrical energy.

If one now constructs simple linear homogeneous transport equations relating a flux to its conjugated flux (the one that is occurring in the same product), one gets, respectively, the transport equations of Fourier (1822), Darcy (1856), Fick (1855), and Ohm (1827).

Physics and Chemistry have long recognized that “coupled” transport can occur. Examples of such coupled transport processes are osmosis, electro-osmosis, thermo-osmosis and the Peltier effect. A coupled transport process takes place when a flux arises due to a non-conjugated force (that is a force that occurs in a product other than the one in which the flux of concern occurs).

Non-equilibrium thermodynamics now postulates that each flux, occurring in the energy dissipation function is a linear, homogeneous (no intercept) function of all forces occurring in the same equation for the total energy dissipation. Thus:

$$j_q = -L_{TT} \text{grad } T - L_{TV} \text{grad } H - L_{TD} \text{grad } \pi - L_{TE} \text{grad } E \quad (9)$$

$$j^V = -L_{VT} \text{grad } T - L_{VV} \text{grad } H - L_{VD} \text{grad } \pi - L_{VE} \text{grad } E \quad (10)$$

$$j^D = -L_{DT} \text{grad } T - L_{DV} \text{grad } H - L_{DD} \text{grad } \pi - L_{DE} \text{grad } E \quad (11)$$

$$j^E = -L_{ET} \text{grad } T - L_{EV} \text{grad } H - L_{ED} \text{grad } \pi - L_{EE} \text{grad } E \quad (12)$$

The diagonal (backward slash) terms of the right-hand sides of Eqs. (9)–(12), with the L_{KK} coefficients, represent the “straight”, well-known transport processes, and are known as the “laws” of Fourier (1822), Darcy (1856), Fick (1855), and Ohm (1827), respectively. All other (off-diagonal) terms represent coupling processes, and their coefficients, L_{KM} , are known as coupling coefficients. They come in pairs of twins, and are located symmetrically across the diagonal. They represent all possible coupling phenomena that can occur, in this case in the presence of the four driving forces occurring in this scheme.

If the energy dissipation equation is complete and all algebra has been carried out correctly, then the two twin coefficients are equal (Groenevelt, 1971):

$$L_{KM} = L_{MK}. \quad (13)$$

Thus the matrix of coefficients in Eqs. (9)–(12) is symmetrical. In the above matrix there are six of these pairs of equal twins. These equalities are known as the Onsager Reciprocal Relations (ORRs).

After the publication of his, now famous, papers, Onsager’s (1931a, 1931b) work received little attention, until, Miller (1960), working at the Lawrence Radiation Laboratory at the University of California in Livermore, Ca., under the auspices of the U.S. Atomic Energy Commission, published the results of a vast array of experiments trying to verify the ORRs. His conclusions were: “For thermoelectricity, electro-kinetics, isothermal diffusion, and anisotropic heat conduction, the experimental checks are sufficiently good that the validity of these relations is practically unquestionable,” and “In view of the above, the author concludes that the experimental evidence is overwhelmingly in favor of the validity of the Onsager reciprocal relations.”

Miller’s work has remained unknown in the field of Soil Physics for a long time and Onsager’s work has remained in doubt and sometimes ridiculed even after he was awarded the Nobel Prize (see the cartoon in the Saturday Review (1969, p.88) as reproduced in Groenevelt (1971)).

It was Elrick’s group of soil physicists that undertook the first thorough experiment concerning one of the coupled transport processes in the laboratory (Elrick et al., 1976).

To keep everything under complete control, they chose as “porous medium” homogeneous, homo-ionic montmorillonite, a fine clay, instead of real soil. The results, involving capillary osmosis and reverse osmosis, as well as electro-osmosis and streaming current, are presented in Groenevelt et al. (1978).

The fundamental value of the procedure outlined above is that all possible forms of energy dissipation are accounted for, even if they have never been observed.

For Soil Physics they are extremely important and useful (Groenevelt, 1971).

After the merciless denigration of the Onsager Reciprocal Relations by Truesdell (1968), soil physicists should rise up and accept the validity and the great practical usefulness of ORRs. The framework of the matrix of coefficients in Eqs. (9)–(12), is comparable, in nature but not quite in stature, to the periodic system as proposed by Mendeleev. For his systematic framework Onsager received the Nobel Prize in Chemistry (1968), which Mendeleev never received. The first such prize in chemistry was awarded to Jacobus van’t Hoff, even though Mendeleev was still alive. The refusal by the Nobel Committee to award the prize to Mendeleev continued until his death in 1907.

The framework makes it possible, in case the measurement of a certain coefficient is difficult or expensive, to measure its twin, which often appears to be easier or cheaper to measure.

When building models for transport processes in clay soils, e.g., based on electrical double layer theory, one can use these ORRs to verify the correctness of the model: if the cross-coefficients are not equal, the researcher can be assured that somewhere she or he made a mistake.

The actual occurrence of a coupled transport process always relies on some kind of a selection mechanism, such as a mechanism that can select between molecules of different chemical nature, e.g., water and salts, or a mechanism that can select between “hot” and “cold” molecules, e.g., a liquid-gas interface (Groenevelt and Elrick, 1976; Groenevelt and Kay, 1974). It should be noted that convection (advection) is never a selection mechanism.

The coupled transport processes are discussed here in pairs of twin phenomena, and indicated by their (equal) coefficients:

L_{TV} (thermo-filtration) and L_{VT} (thermo-osmosis)

Both these phenomena are very common in soils. The transport of heat due to a water potential gradient in the absence of a temperature gradient and the transport of water due to a temperature gradient in the absence of a water potential gradient are occurring constantly in the soil. In unsaturated soil they are due to evaporation and condensation [the selection mechanism here is the heat of vaporization or condensation]. In saturated soils the magnitude of these phenomena is very small [the selection mechanism is now the heat of wetting]. In frozen soils they are caused by freezing and melting [here, the selection mechanism is the heat of freezing or melting and solidification or sublimation].

These phenomena and the values of the two coefficients are discussed in the literature by Kay and Groenevelt (1974).

L_{TD} and L_{DT} (thermo-diffusive processes)

In the literature these phenomena are known as, respectively, the Dufour effect and the Soret effect (Mortimer and Eyring, 1980). The magnitude of these effects in soils is small, but their occurrence is quite probable.

L_{TE} and L_{ET} (thermo-electric processes)

In the literature these phenomena are known as, respectively, the Peltier (1834) effect and the Seebeck (1822) effect. The magnitude of these phenomena in soils is small, but their occurrence is quite likely.

L_{VD} (osmosis, also called capillary osmosis) and L_{DV} (reverse osmosis or salt sieving)

These phenomena are common in soils. The magnitude is directly related to the clay content of the soil. [The selection mechanism lies with the electrical double layer]. Clay particles expel negative ions and therefore, they expel dissociated salts (negative adsorption). The longstanding conflict as to whether the osmotic pressure (potential) should be added to the hydraulic potential of water to produce the “total” potential of water, the gradient of which then serves as the driving force on the water, is here resolved. As the value of L_{VD} is almost always smaller, than the value of L_{VV} (except for a perfectly semi-permeable membrane), the concept of the “total” potential is useless. The ratio L_{VD} over L_{VV} is known as the **reflection coefficient**.

L_{VE} (electro-osmosis) and L_{EV} (streaming current)

These electrokinetic phenomena also find their cause in the existence of electrical double layers. The most commonly observed result of the effects is the streaming potential in clay soils.

L_{DE} (electro-phoresis) and L_{ED} (diffusion current)

These phenomena, together with osmosis and reverse osmosis, electro-osmosis and streaming current are extensively discussed in the literature by Elrick et al. (1976) and Groenevelt and Elrick (1976). They calculated the magnitude of all six coefficients on the basis of electrical double layer theory.

The “linearity” argument

In his original theory, Onsager (1931a,b) stipulated that “the system” should be “close to equilibrium.” He was trying to wrestle the branch of science, Thermodynamics, out of the 19th century constraint that the system should be “closed” and at equilibrium. Ever since its publication up to the awarding of the Nobel Prize to Onsager in 1968, and long after that, the question has been raised: “How close is close to equilibrium?” Here, an argument to resolve this conundrum is as follows:

Onsager’s theory deals with the dissipation of energy (entropy production). There are many processes that dissipate energy. In Physics and Chemistry, the processes often carry the name of the scientist who formulated the process by means of an equation. Names, such as Fourier, Fick, Ohm, and Darcy, are examples of transport equations that are accompanied by energy dissipation. These equations contain a driving force which results in a flux. These two components are linearly and homogeneously related by a transport coefficient. When the driving force is zero, there is no flux and no energy dissipation: the system is at equilibrium. The driving force and the resulting flux are “conjugated.” The theory of Onsager deals with the energy dissipation during processes where the driving force and the flux are not “conjugated.” These processes are called “coupled transport processes” and carry names such as osmosis, thermo-osmosis, and electro-osmosis, or they carry the name of the first observer, such as “Dufour effect,” “Seebeck effect.” Onsager reasoned that each flux is also a linear homogeneous function of each non conjugated force that acts upon the system. In general, these non-conjugated relations are of a secondary intensity as the two related primary (conjugated) mechanisms of energy dissipation. By setting each flux as a homogeneous, linear function of each possible driving force, the set of transport coefficients appears as a matrix. Onsager’s “discovery” is now that this matrix is symmetrical, i.e., that each non-diagonal coefficient is equal to its counterpart on the other side of the diagonal:

$$L_{KM} = L_{MK} \quad (14)$$

The proposed solution for the conundrum now reads:

“As long as the primary (diagonal) process-behavior is homogeneous and linear, the secondary (non-diagonal) process must also be homogeneous and linear. So, as long as Darcy’s law holds (no turbulence in the soil water), one can be assured that the non-linear dissipation process is also linear.”

The general transport equations now formulate the different fluxes as the sum of a set of products of a driving force and a transport coefficient. These transport coefficients form a square matrix of which the diagonal elements represent the conjugated transport processes and the off-diagonal coefficients represent the non-conjugated (couple) transport processes.

The “discovery” by Onsager is now that the matrix of coefficients is symmetrical. Obviously, the requirement that the system should not be “too far from equilibrium” is linked to the linearity of the coupled transport processes, which, in turn, is subjugated by the linearity of the conjugated transport processes.

So, as long as the conjugated laws hold, e.g., Darcy’s law holds (no turbulence in the soil), the laws for non-conjugated transport processes (capillary-osmosis, thermos-osmosis and electro-osmosis) also must hold (Kay and Groenevelt, 1974).

Note: When Onsager received his Nobel prize for Chemistry in 1968, 37 years after he formulated the “Reciprocal Relations”:

$$L_{KM} = L_{MK} \quad (15)$$

the Nobel Committee obviously ignored Nobel’s stipulation that the prize should be awarded for recent work, a stipulation that prevented Mendeleev from receiving the prize because the work was “too old.” Onsager’s work is now 90 years old. A greater recognition in the sciences that deal with porous media such as soils and membranes is overdue (Katchalsky and Curran, 1965).

The thermodynamic versus the mechanistic theory

Apart from the thermodynamic theory for transport processes and their coupling, there is also a mechanistic theory (Hillel, 2005). In the mechanistic theory there is no accounting for the energy dissipation (entropy production) accompanying the transport process. The terms entropy and entropy production do not occur in the mechanistic theory. Therefore, the symmetry of the matrix of transport coefficients is not guaranteed. For the most important coupled transport processes in soils, water and heat transport, the so-called Philip and De Vries theory (Philip and de Vries, 1957), the relation between the two coupling coefficients is not discussed.

Modern development and outlook

The branch of science called “Non-equilibrium Thermodynamics” is now split into two sub-branches. The first one, dealt with above, is now called “Linear Non-equilibrium Thermodynamics.” Indeed, all transport equations presented above are linear.

This sub-branch of science deals with processes that are “not far from equilibrium.”

The second sub-branch is now called “Non-linear Non-equilibrium Thermodynamics” (Glansdorff and Prigogine, 1971).

Its great proponent was Ilya Prigogine, who received the Nobel Prize for Chemistry for his work in 1977. This sub-branch of science deals with “violent” processes (“far away from equilibrium”), such as Liesegang rings and atom bombs. The natural transport processes in the soil are never “violent”, and, as long as Darcy’s law (one of the primary energy dissipating processes) holds, all coupling processes fall in the park of “not far from equilibrium.” However, if there are cases where Darcy’s law breaks down and turns into the non-linear Forchheimer equation (that is when the water in the soil becomes turbulent), soil physicists may have to turn their attention to this latest branch of thermodynamics.

References

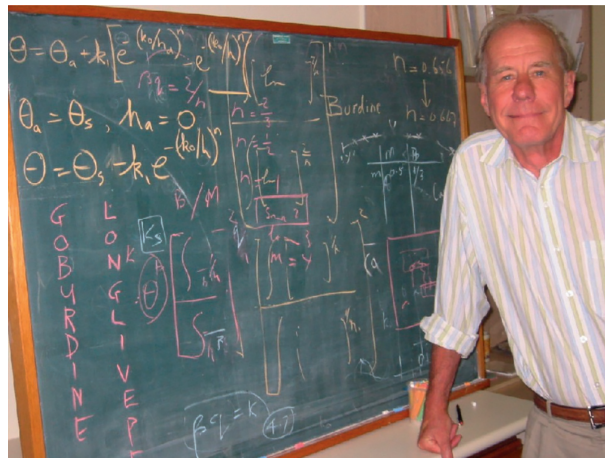
- Bolt GH and Frissel MJ (1960) Thermodynamics of soil moisture. *Netherlands Journal of Agricultural Science* 8: 57–78.
- Cary JW (1963) Onsager’s relation and the non-isothermal diffusion of water vapor. *The Journal of Physical Chemistry* 67: 126–129.
- Cary JW (1964) An evaporation experiment and its irreversible thermodynamics. *International Journal of Heat and Mass Transfer* 7: 531–538.
- Darcy H (1856) *Les fontaines publiques de la ville de Dijon*. Paris: Dalmont.
- Edlefsen NE and Anderson ABC (1943) Thermodynamics of soil moisture. *Hilgardia* 15: 31–160.
- Erick DE, Smiles DE, Baumgartner N, and Groenevelt PH (1976) Coupling phenomena in saturated homo-ionic montmorillonite: I. Experimental. *Soil Science Society of America Journal* 40: 490–491.
- Gibbs JW (1961) The scientific papers of J. Willard Gibbs. *Thermodynamics*. Vol. I, p. 434. Dover.
- Fick A (1855) Ueber diffusion. *Annalen der Physik* 170(1): 59–86. <https://doi.org/10.1002/andp.18551700105>.
- Fourier J (1822) *Théorie Analytique de la Chaleur*. Didot: Firmin.
- Gibbs JW (1876) The equilibrium of heterogeneous substances. *Transactions of the Connecticut Academy of Arts and Sciences*.
- Gibbs JW (1878) The equilibrium of heterogeneous substances. *Transactions of the Connecticut Academy of Arts and Sciences*.
- Glansdorff P and Prigogine I (1971) *Thermodynamic Theory of Structure, Stability and Fluctuations*. London: Wiley.
- Groenevelt PH (1971) Onsager’s reciprocal relations. *Search* 2: 264–267.
- Groenevelt PH and Bolt GH (1969) Non-equilibrium thermodynamics of the soil-water system. *Journal of Hydrology* 7: 358–388.
- Groenevelt PH and Bolt GH (1972) Water retention in soil. *Soil Science* 113: 238–245.
- Groenevelt PH and Erick DE (1976) Coupling phenomena in saturated homo-ionic montmorillonite; II Theoretical. *Soil Science Society of America Journal* 40: 820–823.
- Groenevelt PH, Erick DE, and Blom TJM (1978) Coupling phenomena in saturated homo-ionic montmorillonite; III. Analysis. *Soil Science Society of America Journal* 42: 671–674.
- Groenevelt PH and Kay BD (1974) On the interaction of water and heat transport in frozen and unfrozen soils: II The liquid phase. *Soil Science Society of America Proceedings* 38: 400–404.
- Hillel D (2005) Thermal properties and processes. In: Hillel D (ed.) *The Encyclopedia of Soils in the Environment*. Elsevier.
- Katchalsky A and Curran PF (1965) *Non-Equilibrium Thermodynamics in Biophysics*. Cambridge: Harvard University Press.

- Kay BD and Groenevelt PH (1974) On the interaction between water and heat transport in frozen and unfrozen soil: I Basic theory: The vapor phase. *Soil Science Society of America Proceedings* 38: 395–400.
- Miller DG (1960) Thermodynamics of irreversible processes. The experimental verification of the Onsager Reciprocal Relations. *Chemical Reviews* 60: 15–37.
- Mortimer RG and Eyring H (1980) Elementary transition state theory of the Soret and Dufour effects. *Proceedings of the National Academy of Science* 77(4): 1728–1731. <https://doi.org/10.1073/pnas.77.4.1728>.
- Ohm GS (1827) *Die galvanische Kette, mathematisch bearbeitet*. Berlin: T. H. Riemann.
- Onsager L (1931a) Reciprocal relations in irreversible processes. I. *Physics Review* 37: 405.
- Onsager L (1931b) Reciprocal Relations in Irreversible Processes. II. *Physics Review* 38: 2265.
- Peltier JCA (1834) *Annales de Chimie et de Physique* 56: 371–386.
- Phillip JR and de Vries DA (1957) Moisture movement in porous materials under temperature gradients. *Transactions of the American Geophysical Union* 38: 222–228.
- Pippard AB (1964) *The Elements of Classical Thermodynamics*. Cambridge: Cambridge University Press.
- Seebeck TJ (1822) Magnetische Polarization der Metalle und Erze durch Temperature-Differenz. *Abhandlungen der Königl. Akademie der Wissenschaften zu Berlin* 265–373.
- Truesdell C (1968) In: Parkus H and Sedov LI (eds.) *Irreversible Aspects of Continuum mechanics and Transfer of Physical Characteristics in Moving Fluids*, p. 373. Vienna: Springer.

Addendum

Pieter Hendrik Groenevelt: August 8, 1937–February 10, 2023. [Groenevelt and Bolt \(1969, 1972\)](#) (Photograph provided by Cameron Grant).

The unexpected death of Pieter happened when he was in the middle of polishing this chapter. The editorial team felt it would be appropriate to pay tribute to him as part of what turned out to be his last solo paper.



Pieter was born on the island of Java in what was then the Dutch East Indies, to where his father – an agricultural engineering consultant – and mother had emigrated from the Netherlands. Following the Japanese invasion of the island early in 1942, Europeans were interned, with Pieter, his mother and three sisters being locked up in what Pieter described as a series of concentration camps. His father, who was captured after the fall of Java, was shipped to the island of Kyushu, Japan as a prisoner-of-war. There he died from pneumonia less than 2 years later as a result of working in the horrendous conditions in a coal mine. Pieter was dying from starvation and tropical diseases when the war ended but managed to survive until the gates of the camp were opened and the internees got access to the abundant food outside. Pieter wrote that his experiences at that period in his life were as bad as it gets and he would be able to deal with whatever was thrown at him later.

His mother returned to the Netherlands with Pieter and his sisters in 1946. Pieter worked hard to catch up with his missed schooling and eventually gained a place at the University of Wageningen, gaining a Bachelor of Science in Tropical Land and Water Management, followed by a Master of Science in Irrigation and Drainage. He then went on to study for a doctorate in Soil Physics guided by G. ‘Gerry’ H. Bolt focused on non-equilibrium thermodynamics. Pieter’s thesis, entitled ‘Coupling phenomena in soil transport processes’ resulted in a 30-page review article on non-equilibrium thermodynamics related to the soil-water system (Groenevelt and Bolt, 1969). After a brief period working on the measurement of soil density and water content by gamma-ray attenuation, Pieter was awarded a CSIRO Postdoctoral Fellowship to work in the CSIRO Division of Environmental Mechanics, Canberra, Australia. His tenure of that position was abruptly and acrimoniously curtailed following a disagreement with the Director, John Philip, forcing Pieter, his wife and children to seek a new position at very short notice. Happily, this was solved by an invitation to join the Department of Land Resource Science at the University of Guelph, Canada, which remained his academic home.

The move provided new environments for Pieter to think about, starting with the arctic which piqued his interest in the functioning of freezing and frozen soils. He produced a series of papers dealing with transport of heat and water in frozen and unfrozen soils, tackling the issues of water potential and structural changes resulting from freezing and thawing. Pieter continued to work on thermodynamics but he also developed his analytical skills, which he applied to practical problems, particularly those associated with soil management. He considered the effects of tillage and cropping systems on soil stability and structure,

particularly in relation to soil erosion. The application of liquids, such as those from landfills, either in circulating systems or spray irrigation, was considered both from the viewpoint of the breakdown of surface structure but also from the changes in contact angle between the liquid and the soil fabric. The latter aspect was also investigated with respect to liquid manure and the impact on evaporation from soil. The different approaches adopted to evaluate the effects on soil conditions following the various waste materials caused Pieter to develop new ways of assessing changes.

New approaches were developed for using time domain reflectometry to measure bulk soil electrical conductivity (σ) and the dielectric constant. To correct for influences of the measuring system on σ , reflection measurements were compared with a reference measurement performed in air. To avoid using empirical equations to deal with the effects of changing water content on electrical conductivity, the theory of dispersive behavior was used to develop theoretical model.

To measure the saturated hydraulic conductivity and of slowly permeable clay liners, using single-ring or well permeameters (e.g., Guelph permeameter), Pieter and his colleagues produced an analytical solution describing the transient, capillary-dominated, early-time estimate of these flow parameters. Here, an analytical solution was utilized for early-time, one-dimensional infiltration at an initial constant head followed by a falling-head phase.

Pieter's childhood experiences resulted in him being incredibly supportive of students. He was a very patient and kind mentor but was also able to encourage them to achieve more than they thought themselves capable. There was always a strong element of freedom and fun in his approach but it was always backed up by enthusiasm and inexhaustible hard work. As a colleague he was also there to help if asked. He was fiercely analytical but his criticisms were always balanced.

Immiscible Fluids

Yukie Tanino and Dubravka Pokrajac, School of Engineering, University of Aberdeen, Aberdeen, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

This is an update of RJ Lenhard, JH Dane, M Oostrom, Immiscible fluids, Editor(s): D Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 239–247, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00571-3>.

Introduction	273
Fluid distribution in the pore spaces at equilibrium	274
Extension to air-NAPL-water-soil systems	277
Movement of immiscible fluids in soils	279
Overview of multiphase numerical models used for field-scale simulations	280
Relative permeability	280
Three-phase relative permeability under strongly water-wetting conditions	281
Partitioning of NAPL components among the fluid and solid phases	283
Distribution of NAPL components at local equilibrium	284
Challenges	285
Summary/outlook	285
Acknowledgment	285
References	286

Abstract

Immiscible fluids are ubiquitous in the subsurface, occurring naturally (air and water in the vadose zone) or because of human activity [non-aqueous phase liquids (NAPLs)]. The chapter gives an overview of the dominant physical and chemical processes that govern the fate and transport of air, water, NAPL, and NAPL components, in soil. The vertical distribution of air, water, and NAPL at equilibrium is explained. Laboratory measurements of relative permeabilities of water-wet sandstones from two studies are presented, and salient features identified. The simplest models used to describe the mass transfer of NAPL components in soil between aqueous, gaseous, and solid phases are described.

Key points

- Immiscible fluids are ubiquitous in the subsurface, occurring naturally or because of human activity.
- Field-scale simulators use empirical relationships between relative permeability and capillary pressure and phase saturations, which result from complex processes and are therefore difficult to estimate accurately.
- Even for the simplest air-water-NAPL fluid combinations at equilibrium, the relative permeability to NAPL depends on the saturation of all three fluids *and* their history.
- A number of empirical models for three-fluid phase relative permeability have been proposed, but their applicability is poorly understood.
- Mass transfer of NAPL components within strongly water-wet soil involves three main processes: (a) *volatilization*, (b) *dissolution*, and (c) *sorption*.
- Emerging areas of research include new closure equations based on detailed information of pore structure extracted from X-ray micro-computed tomography data.

Nomenclature

Roman letters

a_{wo} (m^{-1})	specific interfacial area between the aqueous phase and NAPL mixture per unit bulk volume
C_s^β ($kg\ kg^{-1}$)	mass of the sorbed NAPL component β per unit dry mass of soil
C_v^β ($mol\ m^{-3}$)	mass of dissolved NAPL component β per unit volume of the aqueous phase
C_{w*}^β ($mol\ m^{-3}$)	aqueous solubility of pure component β
C_g^β ($mol\ m^{-3}$)	concentration of NAPL component β in the gas phase
$D(\Theta_w)$ ($m^2\ s^{-1}$)	soil water diffusivity

f_{oc} (kg kg ⁻¹)	fraction of organic carbon of soils
g (m s ⁻²)	gravitational acceleration
h (m)	matric head
H (m)	height of NAPL column (cf. Fig. 5)
$J(S_w)$ (.)	Leverett J-function
k (m ²)	intrinsic permeability
k_α (m ²)	effective permeability to phase α
$k_{r\alpha}$ (.)	relative permeability to phase α ($= k_\alpha/k$)
k_{wo}^β (m s ⁻¹)	mass transfer coefficient between the aqueous phase and NAPL mixture
$K(\Theta_w)$ (m s ⁻¹)	hydraulic conductivity
K_d^β (m ³ kg ⁻¹)	soil/water partitioning or soil/water distribution coefficient of NAPL component β
K_{oc}^β (m ³ kg ⁻¹)	organic carbon partitioning coefficient
K_H^β (Pa m ³ mol ⁻¹)	Henry's Law constant
p^β (Pa)	partial pressure of NAPL component
p_*^β (Pa)	equilibrium vapor pressure of pure component β
Δp (Pa)	difference in pressures of two fluids across the interface
p_c (Pa)	capillary pressure ($= p_n - p_w$)
p_n (Pa)	pressure of the non-aqueous phase
p_w (Pa)	pressure of the aqueous phase
P_n (Pa)	macroscopic pressure in the non-wetting phase
P_w (Pa)	macroscopic pressure in the wetting phase
P_α (Pa)	macroscopic pressure of fluid α
P_{wn} (Pa)	macroscopic capillary pressure
P_{ow} (Pa)	NAPL-water macroscopic capillary pressure
P_{aw} (Pa)	air-water macroscopic capillary pressure
P_{ao} (Pa)	air-NAPL macroscopic capillary pressure
q_α (m s ⁻¹)	volumetric flow rate of fluid α per unit bulk cross-sectional area
r (m)	pore radius
R (J K ⁻¹ mol ⁻¹)	universal gas constant ($= 8.314$ J K ⁻¹ mol ⁻¹)
R_1, R_2 (m)	principal radii of curvature of the interface
S_α (.)	saturation of fluid phase α
S_{gr} (.)	residual gas saturation
S_{wi} (.)	irreducible water saturation
V_α (m ³)	volume of fluid α within REV
V_v (m ³)	volume of void space within REV
V_t (m ³)	bulk volume of REV
z (m)	vertical coordinate
z_c (m)	critical height

Greek letters

θ (rad)	contact angle
θ_{aw} (rad)	air-water contact angle
θ_{ow} (rad)	NAPL-water contact angle
Θ_w (.)	water content ($= \phi S_w$)
μ_α (Pa s)	dynamic viscosity of fluid α
ξ (.)	parameter defining NAPL (oil) extent in vertical equilibrium profile
ρ_α (kg m ⁻³)	density of fluid α
σ (N m ⁻¹)	interfacial tension
σ_{ow} (N m ⁻¹)	equilibrium NAPL-water interfacial tension
σ_{aw} (N m ⁻¹)	equilibrium air-water interfacial tension
σ_{ao} (N m ⁻¹)	equilibrium air-NAPL interfacial tension
ϕ (.)	porosity
χ_o^β (.)	mole fraction of NAPL component β in the NAPL

Subscripts and superscripts

i	initial (non-equilibrium)
o	NAPL (oil) phase
w	aqueous phase
a	air phase
g	gaseous (air) phase
n	non-wetting phase
r	relative (permeability) or residual (saturation)
α	fluid phase; $\alpha = (w, a, g, o)$ denote water, air, gas, oil, respectively
β	NAPL component

Abbreviations

REV (m^3)	representative elementary volume
NAPL	non-aqueous phase liquid
μCT	micro-computed tomography

Introduction

Soils are located near the surfaces of land on Earth. As such, they contain variable amounts of air and water, which are immiscible, i.e., they cannot mix to form a single phase in all ratios. Soils are also subject to contamination by human activities. Many of the contaminants commonly found in soils are liquids that are immiscible with air and water. These immiscible contaminants are principally fuel hydrocarbons (petroleum liquids) and industrial solvents, and are referred to as non-aqueous phase liquids (NAPLs). Fuel hydrocarbons tend to be NAPLs that are less dense than water, or *light* NAPLs (LNAPLs). Industrial solvents, especially chlorinated hydrocarbons, tend to be denser than water, or *dense* NAPLs (DNAPLs). Examples of common NAPLs are listed in Table 1.

The increased interest in better understanding NAPL behavior in soils is largely driven by our concern for water quality. As water moves downward through soils, dissolved NAPL components can migrate to potable water supplies. Furthermore, volatile components of NAPLs may partition to the air occupying the pores of the soil and move through the soil. These volatilized components can subsequently partition into pore water and enter potable water supplies. The migration of NAPL components and, ultimately, their contamination of groundwater are affected by the distribution of NAPL in soils.

The behavior of water, air, and NAPL in soils can be explored using simulation models. These simulators solve fundamental equations governing the movement of fluid phases to predict their movement over a selected area of interest, i.e., the simulation domain. To produce reliable predictions, models need to be given the soil characteristics, the initial distribution of relevant fluid phases across the simulation domain, and relevant multiphase flow properties, e.g., relative permeability to the three fluids. Fig. 1 presents a sample of a simulation model output. It shows the consequences of a spillage of two NAPLs on the soil surface, a LNAPL (top row) and a DNAPL (bottom row) (Kikumoto and Nakamura, 2017). During the initial stages (the left two columns) both NAPLs follow a similar pattern—they move downwards until they reach a layer of low permeability. The low permeability zone acts as a barrier, causing them to change their main direction of flow to horizontal and to spread over its surface until they can move downwards again toward the next barrier, where the movement becomes horizontal again. The differences between the migration of the LNAPL and the DNAPL are subtle. The former moves somewhat slower: in the second column its front has just reached the bottom of the lower barrier, whereas the DNAPL front has moved well below it. As the migration continues, the difference between the LNAPL and the DNAPL becomes more and more apparent, with the former moving slower and creating a mound above the groundwater table (blue line) and the latter penetrating the groundwater. Beyond the time instances shown in the figure this distinctly different behavior of LNAPL and DNAPL would continue, with the former moving along the groundwater surface and the latter along the impermeable aquifer bottom. Simulations of this kind offer valuable insights into behavior of fluids in soils at field (m- to km-) scales. However, at present, the reliability of these simulations is strongly restricted by the uncertainty in some of the input parameters, most notably the permeability and relative permeability and their distribution in space.

The movement of air, water, and NAPLs in soils is driven by differences in density between the fluids (i.e., gravity), differences in fluid pressures between the fluid phases, and the gradient of pressure within each fluid phase. Pressure distribution within all fluid phases depends on the size and shape of the pores containing the fluids and the configuration of fluids inside these pores, which may even depend on whether the volume occupied by a particular fluid phase is increasing or decreasing in the soil. Once the net forces acting on the fluids become zero, the system has reached a mechanical equilibrium and the movement stops. The corresponding distribution of fluids depends on all parameters affecting fluid motion, including the displacement history.

Table 1 Density, viscosity, aqueous solubility, and interfacial tension of selected NAPLs at selected temperatures.

	density ρ (kg m ⁻³)	dynamic viscosity μ (mPa s)	interfacial tension with water σ (mN m ⁻¹)	solubility in water (mg L ⁻¹)
aviation gasoline	695.3 ^a , A80 at 15 °C 709 ^{b,c} at 15 °C 715.1 ^a , A100 at 15 °C 705 ^{b,c} at 20 °C	1 ^a at 15 °C 0.443 ^{b,c} at 15 °C 0.419 ^{b,c} at 20 °C	42.2 ^a , A100 at 15 °C 31.7 ^a , A80 at 15 °C 49 – 51 ^{b,c} at 20 °C	100 ^a , A100 at room temperature
leaded gasoline	729.0 ^a at 15 °C 735 ^{b,d} at 15 °C 731 ^{b,d} at 20 °C	0.6 ^a at 15 °C 0.451 ^{b,d} at 20 °C	18.0 ^a at 15 °C 49 – 51 ^{b,d} at 20 °C	169 ^a at 20 °C 240 ^e at 20 °C
home heating oil	864.1 ^a at 15 °C 866.0 ^a at 15 °C 840 ^{a,e} at 20 °C 876.85 ^b at 10 – 29 °C	2.15 ^b at 15 °C 2.00 ^b at 20 °C 4 ^a at 20 °C 4.04 ^e at 20 °C	14.7 ^a at 15 °C 50 ^b at 20 °C	3 ^a , 56 ^a at 20 °C 3.12 ^e at 22° ± 2 °C
benzene	884 ^b at 15 °C 876.5 ^f at 20 °C 878.5 ^f at 20 °C 877 ^b at 21 °C	0.70 ^b at 15 °C 0.6468 ^g at 20 °C 0.647 ^f at 20 °C 0.65 ^b at 20 °C	35.0 ^{b,h} at 20 °C 28.7 ⁱ at 25.0 °C	1780 ^f at 20 °C 820 ^j at 22 °C 1780 ± 45 ^k at 25° ± 1.5 °C 1795 ^{b,g} at 25 °C
dichloromethane (DCM, methylene chloride)	1334 ^b at 15 °C 1325 ^b at 20 °C	0.449 ^j at 15 °C	28.3 ^h at 20 °C	13780 ^b at 20 °C 13200 ^m at 25 °C
trichloroethylene (TCE)	1467 ^b at 15 °C 1460 ^b at 20 °C	0.613 ^b at 15 °C 0.583 ^b at 20 °C	34.5 ^b at 24 °C	1100 ^m at 20 °C 1100 ^b at 25 °C
tetrachloroethylene (PCE, PERC)	1635 ^b at 15 °C 1625.0 ^f at 20 °C 1626 ^b at 20 °C	0.935 ^b at 15 °C 0.884 ^b at 20 °C 0.890 ^f at 20 °C	47.5 ^h at 20 °C 44.4 ^b at 25 °C	150 ^{m,f} at 20 °C 400 ^j at 25 °C

Data were taken directly from the sources cited; where the cited reference is not the original source of the data, the original work was not reviewed.

A80, Avgas 80; A100, Avgas 100.

^aOil Properties Database (Environment Canada Environmental Science and Technology Centre, 2006).

^bChemical Hazard Response Information System (USCG (US Coast Guard), 1999). Source reports temperature and density values in °F and lb. ft.⁻³; where data permit, viscosity and density data have been linearly interpolated to 15 °C and 20 °C.

^c< 4.86 g lead gal⁻¹.

^dautomotive gasoline <4.23 g-lead gal⁻¹.

^eShiu et al., 1990.

^fInstallation Restoration Program Toxicology Guide (Lu, 1989).

^gHoward and Durkin, 1974.

^hGirifalco and Good, 1957.

ⁱMurphy et al., 1957.

^jChiou et al., 1977.

^kMcAuliffe, 1966.

^lMercer and Cohen, 1990.

^mPearson and McConnell, 1975.

Fluid distribution in the pore spaces at equilibrium

At equilibrium the distribution of air, water, and NAPL in soil is dictated by their pressures and the molecular forces of attraction to each other and to the surfaces of the soil grains. 'Wettability' describes the propensity of a fluid to maintain contact with a solid surface in the presence of a second fluid, which in turn depends on a number of physical and chemical properties of the fluids and the solid. One measure of wettability is the contact angle, θ , that the fluid makes with the solid surface: the smaller the θ , the greater its wettability. A 'wetting fluid' ($\theta < 90^\circ$), will spread over the solid surface and a 'nonwetting fluid' ($\theta > 90^\circ$) tends to avoid contact with the solid. The wettability of a fluid determines whether it will occupy larger or smaller pores in the soil, and whether it breaks up into isolated clusters ('capillary trapped') or forms wetting layers along the grain surface at low saturations.

In soils, air is the nonwetting fluid, and water is generally the wetting fluid. Therefore, water typically occupies the smallest pore spaces in soils and air occupies the largest pore spaces. It is generally accepted that NAPLs, if present, spread over water surfaces and form a layer between the water and air. NAPLs occupy pore spaces intermediate to those occupied by water and air. However, there are several exceptions to these conditions. One exception occurs under mixed wettability, in which there is pore-scale and sub-pore scale heterogeneity in the hydrophilicity of the grain surface. Mixed wettability may arise if organic compounds render hydrophobic the surface of grains in direct contact with NAPLs, resulting in NAPLs wetting parts of the grains instead of water. Another exception is where the NAPL has a complex structure in which one end of the compound is hydrophilic and the opposite end is hydrophobic.

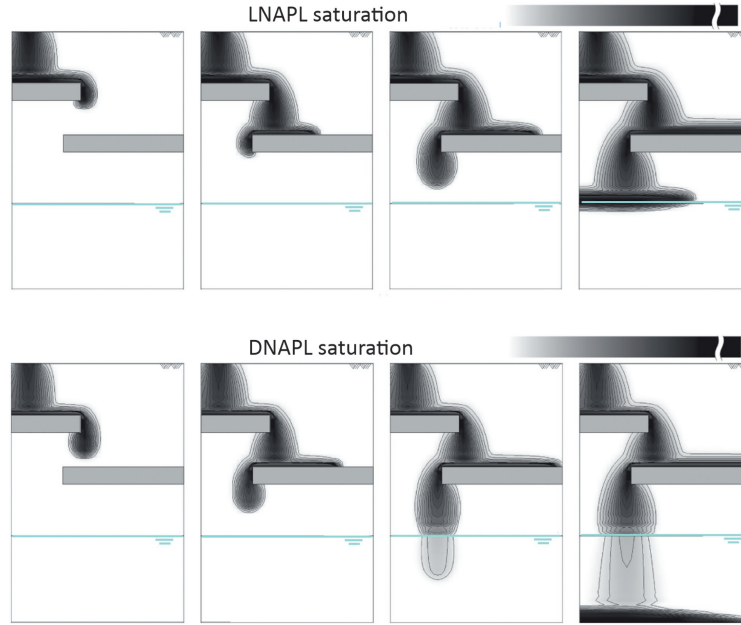


Fig. 1 Migration of a LNAPL (top row) and a DNAPL (bottom row) in an idealized soil profile containing regions of low permeability (grey rectangles) located above the water table (blue horizontal lines). NAPL concentration is shown as a shade of grey. Vertical columns correspond to four instances after the start of spillage, with time increasing from left to right. Reproduced with permission and modified from Kikumoto M and Nakamura K (2017) Simulation of seepage flow of non-aqueous phase liquid in vadose zone. *Environmental Geotechnics* 4(3): 171–183. doi: 10.1680/jenge.15.00011.

Contact angle is a function of not only the interfacial energy of each fluid with the solid surface but also the interfacial tension between the fluid pairs. Interfacial tension is the net force of attraction between molecules at the interface of two immiscible fluids per unit length. It can also be defined as energy per unit interfacial area required to increase that area by one unit. In soils, the distribution of immiscible fluids is different for identical pore geometries and fluid pressures if the interfacial tensions are different. Further details about interfacial tension, including tabulated values for selected liquids and their vapor, can be found in the *Capillarity* chapter.

A basic equation of capillarity is the Young-Laplace equation, which considers a single interface separating two immiscible fluids at mechanical equilibrium, such as water and air in a single pore. It relates the difference in fluid pressures across the interface to the interfacial tension and the curvature of the interface at mechanical equilibrium. The Young-Laplace equation is:

$$\Delta p = \sigma \left(\frac{1}{R_1} + \frac{1}{R_2} \right) \quad (1)$$

in which Δp is the difference in pressures of the fluids across the interface, σ is the interfacial tension between the fluid pair, and R_1 and R_2 are the principal radii of curvature of the interface in orthogonal directions. Δp is closely related to the *capillary pressure*, which is conventionally defined in petroleum engineering and in soil physics as:

$$p_c = p_n - p_w \quad (2)$$

where p_n and p_w are the pressures of the non-aqueous phase and the aqueous phase, respectively. The interface separating the nonwetting and wetting fluids is curved convex toward the nonwetting fluid, which means that the pressure of the nonwetting fluid is greater than the wetting fluid. Consequently, the capillary pressure is always positive using the definition given by Eq. (2). In a cylindrical pore the radii of curvature of the fluid-fluid interface in Eq. (1) are equal ($R_1 = R_2 = R$), i.e., the interface has a shape of a sphere. From the geometry of the interface and the pore, the interface radius can be converted to the pore radius, r , by $\cos\theta = r/R$. Eq. (1) then reduces to:

$$\Delta p = \frac{2\sigma \cos\theta}{r}. \quad (3)$$

The pores within the soils have complex geometry and fluid phases within them create numerous interfaces. Fig. 2 shows the distribution of air in a packed column of spherical particles initially saturated with water, then allowed to drain by gravity, and finally imbibed with water again (Tanino et al., 2020). Each air cluster has multiple interfaces, some of them with the solid and some with water. However, for predicting behavior of fluids at the field scale or even the Darcy scale it is impractical—and often unnecessary—to simulate the distribution of fluid phases at such level of detail. Spatial averaging is therefore applied to all relevant properties and fundamental laws to switch from the scale of individual pores to Darcy scale, which is referred to in this chapter as the macroscopic scale. Averaging is performed over a representative elementary volume (REV) which is sufficiently large to yield statistically representative average properties but sufficiently small to exclude macroscopic heterogeneity (Bear, 1960).

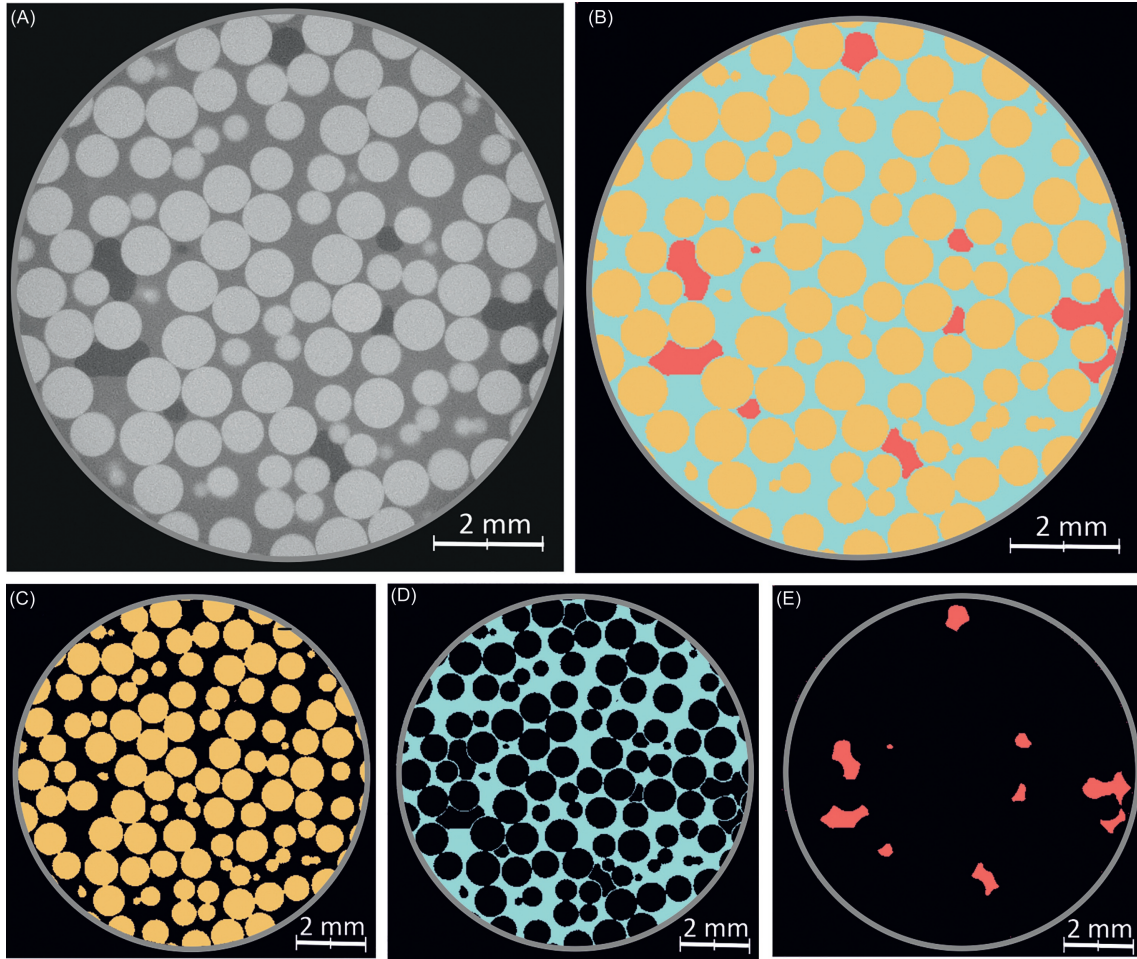


Fig. 2 A cross-section of a 38 mm-long, 21 mm-diameter packed column of glass spheres after secondary imbibition visualized by X-ray micro-computed tomography (μ CT). Each pixel in the tomogram (A) was labelled as being occupied by the solid (brown), water (blue), or air (red) to generate the segmented image (B). Configurations of the solid, water, and air are shown in (C), (D), and (E), respectively. Raw image from Tanino Y, Ibekwe A and Pokrajac D (2020) Impact of grain roughness on residual nonwetting phase cluster size distribution in packed columns of uniform spheres. *Physical Review E* **102**(1): 013109. doi: 10.1103/PhysRevE.102.013109 (A); Post-processing and segmentation by Hamza H (2021) *Impact of Grain Roughness on Capillary Trapping in Porous Media*. MSc thesis, University of Aberdeen: United Kingdom. (B–E).

The size of REV can be based on various quantities such as porosity and saturation. Porosity can be expressed as:

$$\phi = \frac{V_v}{V_t} \quad (4)$$

where V_v is the volume of the voids within the REV and V_t is the bulk volume of the REV which includes both voids and the solid phase. In Fig. 2B, an air-water-solid system, V_v is the total volume occupied by water (blue) and air (red), and V_t contains also solid phase (brown). The fractional volume of the void space occupied by fluid phase α , is termed *saturation*, and is defined as:

$$S_\alpha = \frac{V_\alpha}{V_v} = \frac{V_\alpha}{\phi V_t} \quad (5)$$

where $\alpha = (w, a, o)$ for water, air, and NAPL, respectively, and V_α is the volume of the fluid phase α within the REV. The sum of the saturations of all fluid phases occupying the pores within a REV is equal to unity.

Fig. 3 illustrates the process of establishing the size of the REV for porosity and the residual air saturation. It shows how both porosity and the air saturation vary with the size of the volume from which they have been determined until a certain point when they become stable. This point defines the lower limit of the REV, i.e., the minimum size of the averaging volume which can produce physically meaningful results. It can also be noticed that the minimum REV for porosity and air saturation are different. This means that the REV for porosity should not be extrapolated to other quantities. In practical terms the REV determines the size of a laboratory sample which is sufficiently large to produce physically meaningful results.

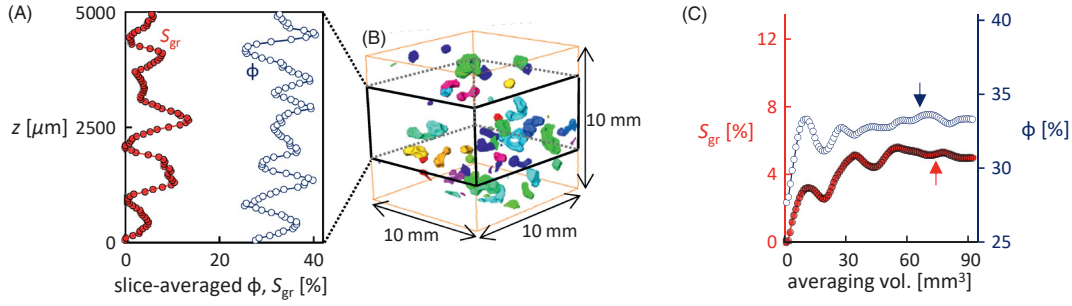


Fig. 3 Distribution of porosity and residual air saturation after secondary imbibition in the same packed column of glass spheres as that in Fig. 2. (A) Vertical distribution of S_{gr} and ϕ averaged over 4.9 mm-diameter $\times$ 5.0 μm -thick tomograms. (B) 3D rendering of the residual air in the middle of the packed column; water and grains are not shown. (C) S_{gr} (red) and ϕ (blue) averaged over cylindrical volumes of constant (4.9 mm-diameter) cross-section and increasing number of tomograms, at 50 μm intervals, in the direction of increasing z . Arrows in (C) depict REV for ϕ (blue) and S_{gr} (red) and are intended as a guide to the eye only. Data in (A, C) from Pentury JG (2021) *Impact of Grain Roughness on Capillary Trapping in Porous Media*. MSc Thesis, University of Aberdeen: United Kingdom. (B) Modified from Tanino Y, Ibekwe A and Pokrajac D (2020) Impact of grain roughness on residual nonwetting phase cluster size distribution in packed columns of uniform spheres. *Physical Review E* **102**(1): 013109. doi: 10.1103/PhysRevE.102.013109.

Upscaling from a single pore to the REV needs to be done for all relevant quantities, including capillary pressure. Macroscopic capillary pressure is usually defined as:

$$P_{wn} = P_n - P_w \quad (6)$$

where P_n , P_w are the *macroscopic* pressures in the non-wetting and wetting fluid, respectively. P_n , P_w are average pressures over the entire fluid volume within the REV. It is possible to derive an alternative definition by rigorous averaging of capillary pressure across individual interfaces within the REV (Starnoni and Pokrajac, 2020), but this definition is yet to be applied to practical simulation models. It should be noted that the two definitions do not in general yield identical values even when the fluid phases are at equilibrium.

Fluid content in soil is largely determined by the macroscopic capillary pressure. For a two-phase system the relationship $P_{wn}(S_w)$ can be expressed by using

$$J(S_w) = P_{wn} \frac{\sqrt{k/\phi}}{\sigma \cos \theta} \quad (7)$$

where $J(S_w)$ is the Leverett J-function, commonly used to scale the results of $P_{wn}(S_w)$ measurements, and k is intrinsic (single phase, liquid) permeability. The RHS of Eq. (7) is very similar to a re-arranged version of Eq. (3) in which the pore radius is replaced with an estimate of the representative radius of pores based on permeability and porosity.

Several models have been proposed to describe the relationship between capillary pressure and fluid saturations mathematically. The most common models for describing S_{α} - P_{α} relations have originally been proposed for two-phase (air-water) soil systems. Brooks and Corey (1964) model provides a relationship between the air-water capillary pressure, P_{aw} , and the dimensionless effective saturation with water:

$$\hat{S}_w = \frac{S_w - S_{wi}}{1 - S_{wi}} \quad (8)$$

where S_{wi} is the irreducible (i.e., irreducible by gravity or realistic pressure gradients) water saturation. The relationship contains two parameters: the so-called air capillary entry pressure and a fitting parameter related to the pore-size distribution. van Genuchten (1980) model relates S_w to P_{aw} , and has two fitting parameters; widely used empirical models for S_{α} - P_{α} are presented in the *Water Retention and Characteristic Curve* chapter.

Extension to air-NAPL-water-soil systems

In air-NAPL-water systems in soils, water is usually the most wetting fluid, air is the non-wetting fluid, and the NAPL has intermediate wettability. The spatial distribution of the three fluid phases depends on the relative magnitudes of their respective interfacial tensions. When a drop of LNAPL is dispensed onto a flat water surface in contact with air, it may behave in the following ways (Hirasaki, 1993):

- If $\sigma_{aw}^i < \sigma_{ow}^i + \sigma_{ao}^i$, where superscript i denotes the initial values of the interfacial tensions between the phases, three phases will meet at a point and the drop will remain stationary.
- If $\sigma_{aw}^i > \sigma_{ow}^i + \sigma_{ao}^i$ (Fig. 4A) it is not possible to balance the three interfacial tensions and NAPL spreads as a film. The film thickness depends on the relative strength of the interfacial tensions at equilibrium. If equilibrium air-water interfacial tension,

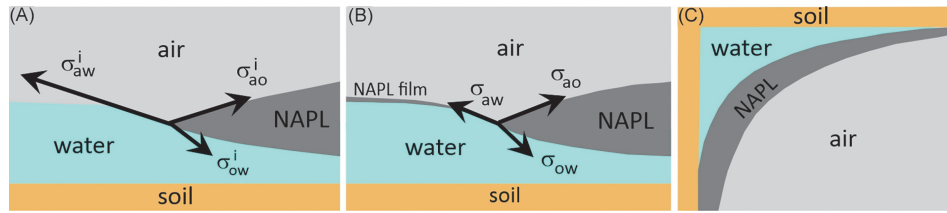


Fig. 4 Three fluids in contact. If $\sigma_{aw}^i > \sigma_{ow}^i + \sigma_{ao}^i$ (A) the NAPL will spread between the air and water and form a thin film. The system will be in equilibrium with a NAPL lens if equilibrium air-water interfacial tension is smaller than the sum of the other two (B). Similar spreading phenomena occur in porous media, where NAPL layers often separate water and air in the corners of pores (C). Based on Figs. 1 and 2 in Zhou D and Blunt MJ (1997) Effect of spreading coefficient on the distribution of light non-aqueous phase liquid in the subsurface. *Journal of Contaminant Hydrology* 25: 1–19. doi: 10.1016/S0169-7722(96)00025-3.

σ_{aw} , is smaller than the sum of the other two interfacial tensions, the resulting film is very thin and in equilibrium with the remaining NAPL (Fig. 4B). In contrast, if σ_{aw} still dominates, the NAPL film may swell infinitely. The latter case is referred to as a spreading system. Note that $\sigma \leq \sigma^i$ for any pair of phases.

The same spreading phenomena can occur in the pore space of soil. Water, as most wetting fluid phase, occupies the narrowest portions of the pore space and the corners of larger pore bodies; air is non-wetting and hence occupies the middle region of the larger pores; NAPL may form a layer between them (Fig. 4C). For such a configuration, NAPL and water combined may be treated as a single, pseudo-wetting phase, i.e., the air-NAPL capillary pressure can be represented as a function of total liquid saturation $S_w + S_o$, as first suggested by Leverett (1941) and subsequently confirmed by Parker et al. (1987). This approximation is only valid for a spreading system.

Zhou and Blunt (1997) use the above approximation and Leverett J-function to express capillary pressures in a water-wet air-NAPL-water system that may be established in soil by gravity drainage. Fig. 5A is a sketch of the vertical arrangement of the three fluid phases in a soil, in the vicinity of the groundwater table: starting from the deeper layers of soil where it is fully saturated with water, $z = 0$ is the level at which NAPL is first mobile, or connected, and $z = H$ is the height at which air is first continuous. This means that below $z = 0$ the only fluid present is water, between $z = 0$ and $z = H$ two fluids—water and NAPL—are present, and above $z = H$ all three fluids are present. The maximum elevation for the region that contains NAPL is determined by the following parameter:

$$\xi = \frac{\sigma_{ow} \cos \theta_{ow} (\rho_o - \rho_a)}{\sigma_{ao} \cos \theta_{ao} (\rho_w - \rho_o)} \quad (9)$$

If $\xi \leq 1$ oil spreads across the entire soil column. However most fluid systems have $\xi > 1$, with oil saturation reducing to zero at a level z_c given by

$$z_c = \frac{\xi H}{\xi - 1} \quad (10)$$

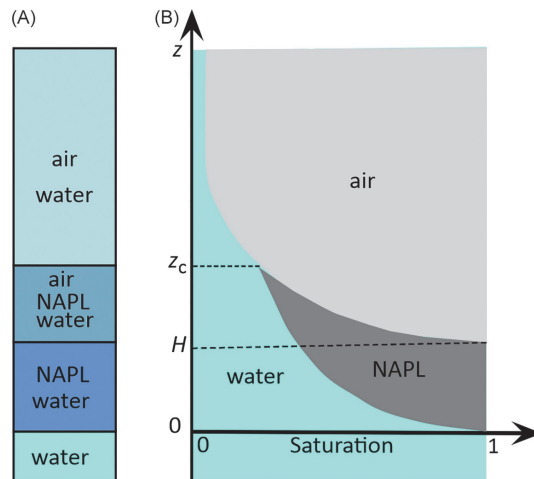


Fig. 5 Vertical equilibrium in a soil with $\xi > 1$ (Eq. 9) close to the water table. Typical arrangement of air, water, and NAPL (A) and their relative volumes (B) as a function of depth. Modified from Zhou D and Blunt MJ (1997) Effect of spreading coefficient on the distribution of light non-aqueous phase liquid in the subsurface. *Journal of Contaminant Hydrology* 25: 1–19. doi: 10.1016/S0169-7722(96)00025-3.

The expressions for oil-water, air-oil, and air-water capillary pressures are, respectively:

$$\begin{aligned} P_{ow} &= (\rho_w - \rho_o) g z = \sqrt{\frac{\phi}{k}} \sigma_{ow} \cos \theta_{ow} J(S_w), 0 \leq z \leq z_c \\ P_{ao} &= (\rho_o - \rho_a) g (z - H) = \sqrt{\frac{\phi}{k}} \sigma_{ao} \cos \theta_{ao} J(S_w + S_o), \quad H \leq z \\ P_{aw} &= (\rho_w - \rho_a) g z - (\rho_o - \rho_a) g H = \sqrt{\frac{\phi}{k}} \sigma_{aw} \cos \theta_{aw} J(S_w), \quad z_c \leq z \end{aligned} \quad (11)$$

The first equality sign follows from the hydrostatic pressure distribution in water, NAPL, and air. The second equality sign relates capillary pressures to fluid saturations using the Leverett J-function (see Eq. 7). For known $J(S_w)$, solving above equations for S_w , S_o , S_a yields the distribution of the three fluids as functions of the level above the water-oil interface ($z = 0$). The typical shape of these distributions is shown in Fig. 5B.

Based on the above equations, Zhou and Blunt (1997) have also developed a method for using measurements of $S_w(z)$ in an air-water system in capillary-gravity equilibrium to predict the distribution of NAPL, water, and air in a three-phase experiment. They assumed a strongly water-wet system with $\cos \theta_{aw} = \cos \theta_{ow} = 1$. The measured and predicted three-phase distributions are in good agreement, confirming that the approximate model of the three-phase behavior was able to capture its main features in the analyzed strongly water-wet gravity-drained systems.

Another, more general extension of the two-phase S_α - P_α relations to air-NAPL-water systems was proposed by Parker et al. (1987). It involves using a scaling parameter to multiply the capillary pressure between each pair of fluid phases and relating these scaled capillary pressures to the corresponding effective liquid saturations. The authors have shown that it is possible to find a set of scaling parameters such that relationship between the scaled pressures and the corresponding liquid saturations is universal, i.e., it is identical for all possible pairs of fluids in both two-phase and three-phase systems. The scaling factors can be determined from S_α - P_α data in two-phase or three-phase systems or predicted from ratios of interfacial tensions.

The relationship between fluid content and capillary pressure is very complex and often not unique, i.e., it depends on the history of fluid movements which causes hysteresis in S_α - P_α relationship. There are three major reasons for hysteresis:

- *Fluid entrapment.* When displaced by a more wetting fluid, the nonwetting fluid can become trapped and isolated in the soil pores. This happens only when wetting fluid is displacing the non-wetting fluid, and causes the wetting fluid to occupy larger pores on imbibition saturation paths than it would on drainage saturation paths for the same wetting-fluid saturation. In larger pores the radii of curvature of the interfaces will also be larger (see Eq. 3), which means the capillary pressure will be smaller (see Eq. 1).
- *Pore-geometry effects,* which include the so-called ‘ink-bottle effect’: an imbibing wetting fluid will stop in larger pores whereas a draining wetting fluid will stop at the narrower pores.
- *Contact angle changes with direction of fluid movement* (draining versus imbibing). Advancing and retreating contact angles for the same fluid are usually different. A different contact angle for the same fluid content will yield a different capillary pressure. However, the change in contact angles in soils with direction of fluid movement is very complex, because the grain surfaces of soil are irregular at the microscopic scale.

Further details and a discussion of hysteresis can be found in the chapter *Hysteresis*.

In air-water systems, there are only two possible combinations of fluid movement: water is either imbibing (and displacing air out of a sample) or draining (letting air in). In air-NAPL-water systems, there are six possible combinations of fluid movement, which makes predicting, or modelling, three-fluid S_α - P_α behaviour more difficult. Despite difficulties it is possible to model hysteretic air-NAPL-water S_α - P_α relations in which water is the wetting fluid. Parker and Lenhard (1987) have proposed a model which accounts for fluid entrapment by using apparent saturations and phase saturation-path history. The authors have shown that the model can successfully reproduce several cycles of imbibition and drainage in an air-NAPL-water system.

Movement of immiscible fluids in soils

The description of the movement of a fluid in unsaturated soils is based on a modification of Darcy’s law for saturated water flow, which was developed in 1856. In 1907, Buckingham modified Darcy’s law to describe unsaturated water flow by replacing the saturated hydraulic conductivity with an unsaturated hydraulic conductivity that is a function of water content or capillary pressure. Over the years, the modification made by Buckingham and the recognition that the hydraulic conductivity consists of fluid-dependent and -independent elements has yielded what is commonly referred to as a generalized form of Darcy’s law:

$$\mathbf{q}_\alpha = -\frac{k_\alpha(S_\alpha)}{\mu_\alpha} (\nabla P_\alpha - \rho_\alpha \mathbf{g}) \quad (12)$$

for phase α , in which $\mathbf{q}_\alpha$ is the volumetric flow rate of α per unit bulk cross-sectional area of the porous medium (also known as the Darcy velocity), $k_\alpha(S_\alpha)$ is the permeability of that medium to α at a particular saturation S_α , μ_α is the viscosity of α , P_α is the pressure

of α , ρ_α is the density of α , and g is the gravitational acceleration. The first term in brackets in Eq. (12) is the gradient of pressure, $\nabla P_\alpha = \partial P_\alpha / \partial x_i$, and x_i , $i = 1, 2, 3$ is a symbol representing the Cartesian coordinates x , y , and z .

Darcy's law is combined with the mass balance equation for fluid α , which can be expressed as:

$$-\nabla \rho_\alpha q_\alpha = \frac{\partial \rho_\alpha \phi S_\alpha}{\partial t} \quad (13)$$

yielding the flow equation

$$\nabla \left[\frac{\rho_\alpha k_\alpha(S_\alpha)}{\mu_\alpha} (\nabla P_\alpha - \rho_\alpha g) \right] = \frac{\partial \rho_\alpha \phi S_\alpha}{\partial t}. \quad (14)$$

A flow equation can be written for each fluid phase α present in soil. The sum of the saturations of all these phases is equal to unity.

In unsaturated soil, it is commonly assumed that air in the soil is not moving and is at atmospheric pressure so that only the flow equation for water is required, and it is more practical to use water content, $\Theta_w = \phi S_w$, instead of saturation. Water is assumed incompressible ($\rho_w = \text{const}$), and its pressure is expressed via the matric head, h , defined as:

$$h = \frac{P_w}{\rho_w g} \quad (15)$$

Expressing Eq. (14) in terms of the matric head, and assuming that water is moving only in the vertical direction yields Richard's equation:

$$\frac{\partial}{\partial z} \left[K(\Theta_w) \left(\frac{\partial h}{\partial z} + 1 \right) \right] = \frac{\partial \Theta_w}{\partial t} \quad (16)$$

where z is the vertical coordinate pointing upwards and $K(\Theta_w) = \rho_w g k_w / \mu_w$ is the hydraulic conductivity. The unity in the brackets expresses the effect of gravity. If water is also moving in a horizontal direction an analogous equation applies, but without the unity.

Overview of multiphase numerical models used for field-scale simulations

Simulation of multiphase flow in soils is based on the flow equations applied to field-scale problems. The process starts with defining a conceptual model, i.e., the description of the domain of interest and the relevant processes. This governs the selection of the flow equations, which are solved numerically using a computer code. The input data usually required for simulation of multiphase flow include soil and fluid characteristics such as ϕ , ρ_α , $k_\alpha(S_\alpha)$, and $P_c(S_\alpha)$, the initial distribution of the fluid phases α across the simulation domain, boundary conditions, and computational parameters such as spatial and temporal steps. Based on these inputs the code calculates how the distribution of pressure and saturation of each fluid phase across the simulation domain changes in time. If physically possible, after sufficiently long simulation time all fluid phases will stop moving because equilibrium pressure and saturation distributions have been reached.

The general procedure described above can be applied to solve the Richards equation for either water content or pressure. In the former case the known model input includes water density and viscosity and the following relationships: $K(\Theta_w)$ and the soil water diffusivity $D(\Theta_w) = K(\Theta_w) dh/d\Theta_w$. These relationships can be expressed using the empirical formulas such as [van Genuchten \(1980\)](#) or [Brooks and Corey \(1964\)](#) models, or even as look-up tables based on experimental results. Starting from the initial distribution of water within the soil column, simulation produces a series of water distributions at different times. The corresponding water pressure can be tracked via $D(\Theta_w)$.

To simulate movement of both water and air, and possibly an additional fluid such as NAPL, the simulation model needs to solve either two-(fluid) phase or three-phase flow equations. In the former case there is a flow Eq. (14) for each phase, and another equation stating that the sum of fluid saturations equals unity. These are combined with an equation that relates capillary pressure to the pressures in the two phases (Eq. 6). The input data include density, viscosity, and $k_\alpha(S_\alpha)$ for each phase, and the relationship between P_c and saturation of one of the phases. The simulation software calculates the distribution of both pressure and saturation for each phase as they evolve in time.

The process for three-phase simulations is analogous, with input data including density, viscosity, and $k_\alpha(S_\alpha)$ for each phase, and relationships between capillary pressure and a saturation for each combination of two phases. Results show the evolution of saturation and pressure distribution for each phase.

The accuracy of simulation models crucially depends on the accuracy of all input data, some of which are very difficult to estimate. The remaining part of this section is focused on the key parameter which determines the motion and ultimately the distribution of fluids in a soil—relative permeability, $k_{r\alpha} (= k_\alpha/k)$. For this reason, the accuracy of relative permeability is critical for obtaining good multiphase flow simulation results.

Relative permeability

Where the pore space of a porous medium is occupied by more than one fluid phase, the permeability of that medium to one phase is affected by the presence of the other phases, and $k_\alpha = k_\alpha(S_\alpha)$. Because the cross-sectional area available to each phase is restricted

by the presence of the other phases, it is widely assumed in numerical simulators that $k_{rx} \leq 1$. However, laboratory experiments in water-NAPL-rock systems show that the permeability to NAPL is enhanced by the presence of water such that $k_{ro} > 1$ if S_o is sufficiently high and the water is stationary (Berg et al., 2008 and references therein). More recently, Christensen and Tanino (2017) showed that this enhanced permeability occurs not only in water-wet rock but in mixed-wet rock also.

Even in two-fluid systems relative permeability is notoriously difficult to measure under capillary-dominated conditions representative of the subsurface. $k_x(S_x)$ for a given system is a function not only of S_x but also saturation history and wettability. In mixed-wet media such as NAPL-contaminated soils and oil reservoir rock, capillary end effects demand the use of Darcy-scale simulators to interpret displacement experiments and to obtain $k_x(S_x)$ and residual NAPL or gas saturation (e.g., Tanino and Christensen, 2019). Where three fluids are present in the pore space, $k_x(S_x)$ is a function not only of its own saturation but that of the other two fluids. Furthermore, saturation history now comprises the co-evolution of three parameters (S_x for $\alpha = o, w, g$) which is not necessarily monotonic.

Because of the above challenges, three-phase relative permeabilities are rarely measured in petroleum engineering or in soil sciences. Instead, relative permeabilities that are needed to predict fluid flow in air-NAPL-water-soil systems are normally estimated. The general approach is to extrapolate from two-phase relative permeabilities, which in turn may be measured directly on relevant samples or estimated from empirical correlations. Two commonly used empirical models in petroleum engineering to describe two-phase relative permeability are the Corey (1954)-type power laws and LET correlations (Lomeland et al., 2005). These functions are presented in the chapter *Water Retention and Characteristic Curve*.

Three-phase relative permeability under strongly water-wetting conditions

Investigators began to study relative permeability in air-NAPL-water systems in the mid-1950s, largely focusing on strongly water-wet conditions. Water and air (and gaseous phases in general) are the wetting and non-wetting phases, respectively, under strongly water-wet conditions irrespective of the presence of NAPL. In contrast, NAPL will preferentially wet the grains relative to gas, but will be repelled relative to water. Consequently, the relative permeabilities to water and air depend mainly on their respective saturations and saturation histories but the relative permeability to NAPL depends on the saturation of all three fluids and their history. We provide selected examples of each below.

Air relative permeability. Hysteresis exists in $k_x(S_x)$ relations primarily because of the entrapment of fluids with lesser wettability by fluids with greater wettability (see *Hysteresis* chapter for other causes). Because air is the non-wetting phase in an air-water, air-NAPL, or an air-NAPL-water system, it will be trapped by capillary forces when its saturation is decreasing, i.e., when pore space occupied by air is invaded by water or NAPL. As expected, laboratory measurements of $k_{rg}(S_g)$ clearly display hysteresis, with $k_{rg}(S_g)$ smaller in the direction of decreasing S_g below a critical S_g (e.g., blue, Fig. 6). The critical S_g presumably corresponds to the saturation at which some of the gas becomes discontinuous (cf. Figs. 2E and 3B). Note that two-phase and three-phase $k_{rg}(S_g)$ are approximately the same at the same S_g and direction of S_g change. Accordingly, $k_{rg}(S_g)$ in air-water-NAPL systems is conventionally approximated by the two-phase $k_{rg}(S_g)$.

Water relative permeability. In strongly water-wetting systems, physical reasoning suggests that water will not be trapped and, accordingly, $k_{rw}(S_w)$ is expected to display only a weak dependence on saturation history. Indeed, laboratory measurements of $k_{rw}(S_w)$ of strongly water-wetting sandstones in three-phase systems have been found to depend largely on the value of S_w and not whether S_w is increasing or decreasing (e.g., markers, Fig. 7). In contrast, in some two-phase systems at low S_w (≤ 0.6), $k_{rw}(S_w)$ at a given S_w is larger in the direction of increasing S_w (blue) than decreasing S_w (red) (lines, Fig. 7). While $k_{rw}(S_w)$ in Berea sandstone

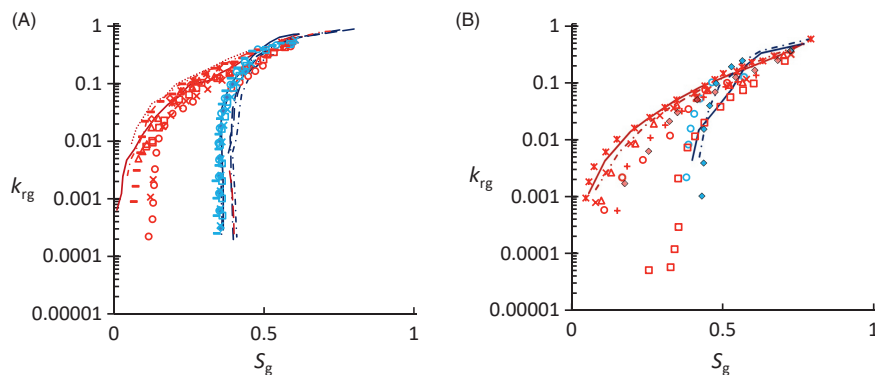


Fig. 6 Relative permeability of a Berea sandstone core (A) and Bentheimer sandstone cores (B) to gas with two (solid lines: water-gas; all other lines: oil-gas) or three fluids (markers) present. Red and blue denote increasing and decreasing S_g , respectively. Data from Oak MJ (1990) Three-phase relative permeability of water-wet Berea. In: *Proceedings, SPE/DOE Enhanced Oil Recovery Symposium, Tulsa, Oklahoma, April*. doi: 10.2118/20183-MS; Oak MJ, Baker LE and Thomas DC (1990) Three-phase relative permeability of Berea sandstone. *Journal of Petroleum Technology* 42(8): 1054–1061. doi: 10.2118/17370-PA (A) and Alizadeh AH and Piri M (2014) The effect of saturation history on three-phase relative permeability: An experimental study. *Water Resources Research* 50: 1636–1664. doi: 10.1002/2013WR014914 (B).

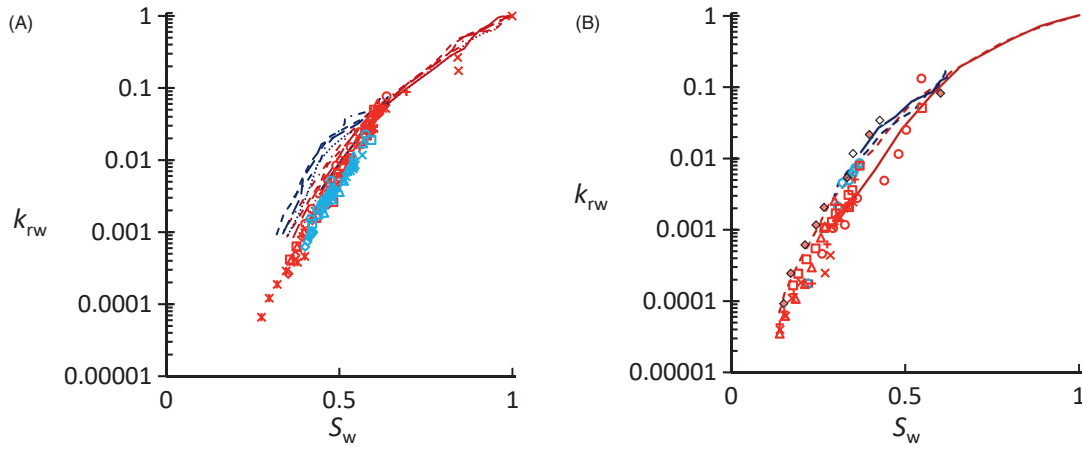


Fig. 7 Relative permeability of a Berea sandstone core (A) and Bentheimer sandstone cores (B) to water with two (solid lines: water-gas; all other lines: water-oil) or three fluids (markers) present as reported in Oak, 1990 and Oak et al., 1990 (A) and Alizadeh and Piri, 2014 (B). Red and blue denote decreasing and increasing S_w , respectively. Data from: (A) Oak MJ, Baker LE and Thomas DC (1990) Three-phase relative permeability of Berea sandstone. *Journal of Petroleum Technology* 42(8): 1054–1061. doi: 10.2118/17370-PA; Oak MJ (1990) Three-phase relative permeability of water-wet Berea. In: *Proceedings, SPE/DOE Enhanced Oil Recovery Symposium, Tulsa, Oklahoma, April*. doi: 10.2118/20183-MS.; (B) Alizadeh AH and Piri M (2014) The effect of saturation history on three-phase relative permeability: An experimental study. *Water Resources Research* 50: 1636–1664. doi: 10.1002/2013WR014914.

appears to be slightly lower for a three-fluid system than the corresponding two-fluid (i.e., gas-water or NAPL-water) system at the same S_w for increasing S_w (blue lines and markers, Fig. 7A), the difference is generally interpreted to be negligible, and $k_{rw}(S_w)$ in air-water-NAPL-soil systems is conventionally approximated by the two-phase $k_{rw}(S_w)$.

NAPL relative permeability. NAPL is the wetting phase relative to air but the non-wetting phase relative to water. It is thus possible for gas to mobilize and displace oil that has been trapped by water. Consequently, and in contrast to k_{rw} , NAPL relative permeability is a function not only of its own saturation but also of the relative amounts of the other two fluids. Thus three-phase $k_{ro}(S_o)$ cannot be approximated by two-phase $k_{ro}(S_o)$ (e.g., Fig. 8, top row), and there can be large differences in NAPL relative permeability because of hysteresis.

Nonetheless, several salient features have been identified. When NAPL displacement is dominated by layer drainage and $S_w \ll S_o$, theory predicts that $k_{ro} \sim S_o^2$ (Fenwick and Blunt, 1998). $k_{ro}(S_o)$ of some gas-water-NAPL fluid systems indeed display a quadratic dependence at low S_o (e.g., Fig. 8A; see also Fenwick and Blunt, 1998); at higher S_o , the exponent is larger than 2. A $k_{ro} \sim S_o^2$ regime is not observed in other gas-water-NAPL combinations (e.g., Fig. 8B), although this may be due to limitations on the range of S_o considered.

Several empirical models have been proposed to predict relative permeability where all three fluids are flowing, with two most popular models in petroleum engineering being those of Baker (1988) and Stone (1970). Both models involve the extrapolation of two-flowing phase relative permeability data to three-flowing phase conditions. The model by Baker (1988) involves saturation-weighted interpolation between NAPL relative permeability in NAPL-water systems, k_{ro}^{ow} , and NAPL relative permeability in three fluid (i.e., water-NAPL-gas) systems at irreducible water saturation, $k_{ro}(S_{wi})$, at the three-phase S_o :

$$k_{ro}(S_o) = \frac{(S_w - S_{wi}) k_{ro}^{ow}(S_o) + (S_g - S_{gr}) k_{ro}(S_{wi}, S_o)}{(S_w - S_{wi}) + (S_g - S_{gr})} \quad (17)$$

where S_{gr} is residual gas saturation in the three-flow phase system. The model by Stone (1970) is given by:

$$k_{ro}(S_o) = S_{oe} \frac{k_{ro}^{ow}(S_w)}{1 - S_{we}} \frac{k_{ro}^{og}(S_g)}{1 - S_{ge}} \quad (18)$$

where

$$\begin{aligned} S_{oe} &= \frac{S_o - S_{or, \min}}{1 - S_{wi} - S_{or, \min}} \\ S_{we} &= \frac{S_w - S_{wi}}{1 - S_{wi} - S_{or, \min}} \\ S_{ge} &= \frac{S_g}{1 - S_{wi} - S_{or, \min}} \end{aligned} \quad (19)$$

and $S_{or, \min}$ is the minimum value of S_{or} under three-phase flow conditions; $S_{or, \min}$ may be predicted from S_{or} in water-oil and oil-gas systems by, e.g., saturation-weighted interpolation, or used as a fitting parameter (Kianinejad and DiCarlo, 2016). Note that it is not always possible to match the saturation history of the three-phase flow with those of $k_{ro}^{ow}(S_w)$ and $k_{ro}^{og}(S_g)$. If S_g is increasing

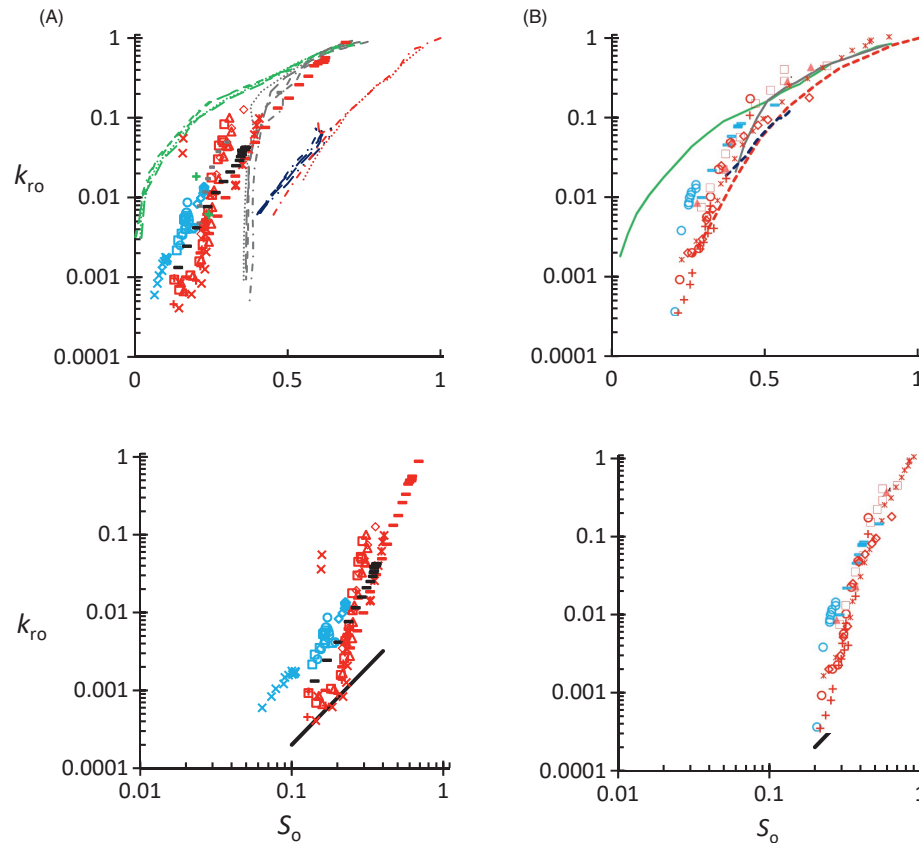


Fig. 8 Relative permeability of a Berea sandstone core (A) and Bentheimer sandstone cores (B) to NAPL with two (lines) or three fluids (markers) present. Lines in top row: water-NAPL drainage (green) and imbibition (grey); NAPL-gas drainage (red) and imbibition (blue). Markers: red and blue denote decreasing and increasing S_o , respectively. Bottom row only shows three-phase data; solid black line is $k_{ro} \sim S_o^2$. Data from (A) Oak MJ (1990) Three-phase relative permeability of water-wet Berea. In: *Proceedings, SPE/DOE Enhanced Oil Recovery Symposium, Tulsa, Oklahoma, April*. doi: 10.2118/20183-MS and Oak MJ, Baker LE and Thomas DC (1990) Three-phase relative permeability of Berea sandstone. *Journal of Petroleum Technology* **42**(8): 1054–1061. doi: 10.2118/17370-PA and (B) Alizadeh AH and Piri M (2014) The effect of saturation history on three-phase relative permeability: An experimental study. *Water Resources Research* **50**: 1636–1664. doi: 10.1002/2013WR014914.

and S_o and S_w are decreasing in the three-phase flow, one would choose $k_{ro}^{og}(S_g)$ measured in the same direction (i.e., increasing S_g and decreasing S_o). However, the correct choice for $k_{ro}^{ow}(S_w)$ is not clear, since S_w and S_o cannot change in the same direction at the same time.

Partitioning of NAPL components among the fluid and solid phases

NAPL components in soils are present in various forms. They can be dissolved in pore water or in a NAPL phase, or adsorbed on the grain surfaces. In addition, volatile components may be found in the gaseous phase. To predict the fate of NAPL components in the subsurface, the partitioning of the components between liquid, gaseous, and solid phases must be considered.

Fig. 9 illustrates the partitioning pathways in a soil-water-air-NAPL system (Newell et al., 1995). Mass transfer of NAPL components within strongly water-wet soil involves three main processes: (a) *volatilization* from the aqueous phase and NAPL into the gaseous phase; (b) *dissolution* from the gaseous phase and NAPL into the aqueous phase; and (c) *sorption* from the aqueous phase onto and into the solid phase. Because water coats the grains in water-wet media, gas-to-grain and NAPL-to-grain mass transfer are believed to be rare. Instead, NAPL components in the gas phase and the NAPL must first partition into the aqueous phase before adsorbing on grain surfaces.

Several different models have been used to describe the mass transfer of compounds between aqueous, gaseous, and solid phases in soil. Because many mass-transfer processes are assumed to occur much faster than the convective flow of the fluids, mass transfer is commonly modelled as being instantaneous and local equilibrium is assumed.

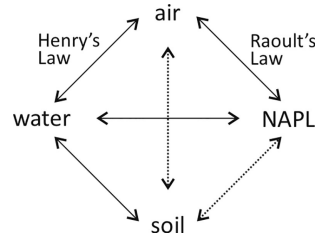


Fig. 9 Partitioning pathways between the four phases that may be found in the unsaturated zone. Solid arrows depict main pathways; dashed arrows depict less likely pathways. Modified from Newell CJ, Acre SD, Ross RR and Huling SG (1995) *Light Nonaqueous Phase Liquids. Ground Water Issue*. US EPA; Report No. EPA/540/S-95/500.

Distribution of NAPL components at local equilibrium

Soil-water partitioning (sorption) Sorption is the result of two mechanisms: *adsorption* onto the surface of grains and *absorption* into the internal volume (e.g., microporosity) of the grains. Under idealized conditions—where microporosity is limited, absorption is negligible, and conditions are steady—sorption equilibrium may be achieved relatively quickly. Equilibrium sorption of many organic, nonpolar chemicals can be described by a linear sorption isotherm of the form

$$C_s^\beta = K_d^\beta C_w^\beta \quad (20)$$

where C_s^β is the equilibrium sorbed concentration of NAPL component β [mass of sorbed chemical β per unit dry mass of soil], C_w^β is the equilibrium aqueous concentration of β [mass dissolved per unit volume of the aqueous phase], and the constant of proportionality K_d^β is the equilibrium soil/water partitioning (or distribution) coefficient for β [$L^3 M^{-1}$].

Where the mass fraction of organic carbon in the soil—a routinely measured property in the laboratory—is greater than $f_{oc} = O(0.1-1)$ wt%, sorption of NAPL components is controlled by the organic content of the soil and K_d^β can be estimated by:

$$K_d^\beta = f_{oc} K_{oc}^\beta \quad (21)$$

in which K_{oc}^β is the organic carbon partitioning coefficient of NAPL component β , which can be estimated from solubility data and molecular structure (see *Solute Transport* chapter for a table of K_{oc} values). In low f_{oc} soils, sorption onto inorganic surfaces becomes significant and $K_d^\beta > f_{oc} K_{oc}^\beta$ (Fitts, 2013 and references therein). A linear sorption isotherm assumes K_d^β is independent of concentration, which is not valid at high concentrations. More complex models and further discussion can be found in the *Organic Chemicals* chapter.

Water-air partitioning (volatilization) The partitioning of NAPL components between the gaseous phase and the aqueous phase is described by Henry's Law,

$$p^\beta = K_H^\beta C_w^\beta \quad (22)$$

where p^β is the partial pressure of NAPL component β in the gaseous mixture at equilibrium with the aqueous phase, C_w^β is the concentration of β in the aqueous phase [moles of β per unit volume of aqueous phase] and K_H^β [$atm m^3 mol^{-1}$] is the constant of proportionality known as Henry's Law constant. The partial pressures can be converted into concentrations of the NAPL components in the gaseous phase using, e.g., the ideal gas law, which yields:

$$C_g^\beta = \frac{p^\beta}{R T} = \frac{K_H^\beta}{R T} C_w^\beta \quad (23)$$

where C_g^β is the number of moles of component β volatilized into a unit volume of the gaseous phase, R is the ideal gas constant, and T is the temperature.

NAPL-air partitioning (volatilization) The partitioning of NAPL components between the gaseous phase and the NAPL is often described by Raoult's Law,

$$p^\beta = p_*^\beta x_o^\beta \quad (24)$$

where p^β is now the equilibrium partial pressure of NAPL component α in the gaseous mixture over the NAPL mixture, p_*^β is the equilibrium vapor pressure of the pure component β at the same temperature, and x_o^β is the mole fraction of component β in the NAPL mixture [moles of β /moles of all molecules in the NAPL phase]. The larger the value of p_*^β , the more volatile the component. As in Eq. (22) above, the ideal gas law can be applied to convert partial pressure to concentration, and Eq. (24) becomes:

$$C_g^\beta = \frac{p^\beta}{R T} = \frac{p_*^\beta}{R T} x_o^\beta \quad (25)$$

Since $x_o^\beta \leq 1$, Eq. (25) implies that the equilibrium concentration of each component β in the gaseous phase will be lower than if the NAPL was 100% β .

NAPL-water partitioning (dissolution) If the liquid phases both behave as ideal mixtures, an analogue of Raoult's Law may be used to describe the partitioning of NAPL components between a multicomponent NAPL mixture and the aqueous phase at equilibrium:

$$C_w^\beta = C_{w*}^\beta x_o^\beta \quad (26)$$

where C_{w*}^β is the aqueous solubility of pure component β (e.g., benzene), i.e., the concentration of β in an aqueous phase that is in contact and in equilibrium with pure liquid β [moles L^{-1}]. The more water-soluble components (i.e., one with larger C_{w*}^β) will partition into groundwater more readily than the less soluble components (e.g., larger molecular weight hydrocarbons) in the NAPL phase, which in turn implies that the NAPL phase will become enriched in the less soluble components over time. The range of values of C_w^β in 31 gasoline samples from Florida, United States can be found in Table II, [Cline et al., 1991](#).

Challenges

In a real system, several factors may give rise to deviations from Eq. (24). For example, the presence of highly water-soluble constituents (e.g., alcohols), or cosolvents, is known to enhance the aqueous solubility of NAPL compounds. Similarly, if mass transfer cannot be assumed to be instantaneous, rate-limited mass transfer needs to be used to describe changes in NAPL component partitioning. Rate-limited interphase mass transfer is often approximated by first-order kinetic relationships representing the combined effect of all physical nonequilibrium processes affecting the migration between phases. The driving force is the concentration gradient across the interfaces separating the phases. The interphase exchange rate per unit volume [$M T^{-1} L^{-3}$] for first-order dissolution of NAPL constituents into the aqueous phase is usually given by:

$$\frac{\partial}{\partial t} (\phi S_w C_w^\beta) = k_{wo}^\beta a_{wo} [C_{w*}^\beta x_o^\beta - C_w^\beta(t)] \quad (27)$$

where k_{wo}^β [$L T^{-1}$] is the mass transfer coefficient between the aqueous phase and NAPL mixture and a_{wo} [L^{-1}] is the specific interfacial area between the two phases per unit bulk volume. As shown in Eq. (27), NAPL dissolution rates are dependent on the interfacial contact area. The relation between rate-limited NAPL dissolution and interfacial area has been the focus of several research projects.

At first glance, Henry's Law and Raoult's Law appear to be the same. However, there is a difference. Ideal Henry's Law behavior is observed for dilute solutions (i.e., low concentrations of component β), while ideal Raoult's Law applies to a mixture that is nearly pure (high concentrations). The two are thermodynamically consistent; however, significant deviations are frequently observed at intermediate concentration (mole fraction) ranges.

More broadly, the presence of NAPL in the soil is believed to render soils mixed-wet, with the grain surface in the middle of NAPL-invaded pores rendered less hydrophilic or even hydrophobic relative to surfaces in pores that remain water or air-filled. Under such conditions, direct mass transfer from NAPL to the solid phase may no longer be negligible. Similarly, near the top of the unsaturated zone where liquid saturation is sufficiently low, sorption from the vapor phase may be significant (see, e.g., [Unger et al., 1996](#) for a discussion). This is a topic for future work, involving high resolution imaging and sophisticated post-processing techniques that can resolve NAPL layers in mixed-wet systems.

Summary/outlook

Immiscible fluids are ubiquitous in the subsurface, occurring naturally (vadose zone) or because of human activity (NAPL contamination). Field-scale simulation models have been developed to predict the distribution and migration of immiscible fluids in the subsurface and changes to their composition. These simulators use empirical relationships between relative permeability and capillary pressure and phase saturations, which result from complex processes and are therefore difficult to estimate accurately, particularly where the pore space is occupied by three distinct phases. Many empirical correlations have been proposed, but they have limited physical bases and our ability to select the appropriate correlation for a given site is unsatisfactory. Indeed, for non-idealized wetting conditions expected in real soils, we lack both laboratory observation and predictive models. Furthermore, challenges remain in upscaling to field scale and simulating the flow field in heterogeneous porous media which is necessary for quantifying the fate and transport of immiscible fluids at system scales.

Emerging areas of research to address these gaps in knowledge include new closure equations based on detailed information of pore structure extracted from μ CT imaging. Recently the wider availability of μ CT has enabled detailed studies of immiscible fluids at the micron-scale. These relatively new data sets provide direct insight into pore-scale processes and can be used to improve the relationships used in field-scale numerical models. The theoretical framework exists, but experimental data to validate it and to guide its revision are still lacking.

Acknowledgment

The authors thank Hussein Hamza and Josua G. Pentury for providing unpublished details from their MSc theses ([Hamza, 2021](#); [Pentury, 2021](#)).

References

- Alizadeh AH and Piri M (2014) The effect of saturation history on three-phase relative permeability: An experimental study. *Water Resources Research* 50: 1636–1664. <https://doi.org/10.1002/2013WR014914>.
- Baker LE (1988) Three-phase relative permeability correlations. In: *Proceedings, SPE/DOE Enhanced Oil Recovery Symposium, Tulsa, Oklahoma, April 17–20*, SPE-17369-MS. <https://doi.org/10.2118/17369-MS>.
- Bear J (1960) *Dynamics of Fluids in Porous Media*. American Elsevier Publishing Company, Inc.
- Berg S, Cense AW, Hofman JP, and Smits RMM (2008) Two-phase flow in porous media with slip boundary condition. *Transport in Porous Media* 74(3): 275–292.
- Brooks R and Corey T (1964) Hydraulic properties of porous media. In: *Hydrology Papers 3*. Fort Collins, Colorado: Colorado State University.
- Chiou CT, Freed VH, Schmedding DW, and Kohnert RL (1977) Partition coefficient and bioaccumulation of selected organic chemicals. *Environmental Science & Technology* 11(5): 475–478. <https://doi.org/10.1021/es60128a001>.
- Christensen M and Tanino Y (2017) Enhanced permeability due to apparent oil/brine slippage in limestone and its dependence on wettability. *Geophysical Research Letters* 44(12). <https://doi.org/10.1002/2017GL073603>.
- Cline PV, Delfino JJ, and Rao PSC (1991) Partitioning of aromatic constituents into water from gasoline and other complex solvent mixtures. *Environmental Science & Technology* 25(5): 914–920. <https://doi.org/10.1021/es00017a013>.
- Corey AT (1954) The interrelation between gas and oil relative permeabilities. *Producers Monthly* 19: 38–41.
- Environment Canada Environmental Science and Technology Centre (2006) *Oil Properties Database*. https://etc-cte.ec.gc.ca/databases/OilProperties/oil_prop_e.html. (Accessed 23 January, 2022).
- Fenwick DH and Blunt MJ (1998) Network modeling of three-phase flow in porous media. *SPE Journal* 3: 86–96. SPE-38881-PA. <https://doi.org/10.2118/38881-PA>.
- Fitts CR (2013) 10: Groundwater chemistry. In: Fitts CR (ed.) *Groundwater Science*, 2nd edn, pp. 421–497. Academic Press. <https://doi.org/10.1016/B978-0-12-384705-8.00010-8>.
- Girifalco LA and Good RJ (1957) A theory for the estimation of surface and interfacial energies. I. Derivation and application to interfacial tension. *Journal of Physical Chemistry* 61(7): 904–909. <https://doi.org/10.1021/j150553a013>.
- Hamza H (2021) *Impact of Grain Roughness on Capillary Trapping in Porous Media*. MSc thesis United Kingdom: University of Aberdeen.
- Hirasaki GJ (1993) Structural interactions in the wetting and spreading of van der Waals fluids. *Journal of Adhesion Science and Technology* 7(3): 285–322. <https://doi.org/10.1163/156856193X00718>.
- Howard PH and Durkin PR (1974) Benzene. In: *Environmental Sources of Contamination, Ambient Levels, and Fate*, Final Report. Report No. PB-244139.
- Kianinejad A and DiCarlo DA (2016) Three-phase oil relative permeability in water-wet media: A comprehensive study. *Transport in Porous Media* 112: 665–687. <https://doi.org/10.1007/s11242-016-0669-z>.
- Kikumoto M and Nakamura K (2017) Simulation of seepage flow of non-aqueous phase liquid in vadose zone. *Environmental Geotechnics* 4(3): 171–183. <https://doi.org/10.1680/jenge.15.00011>.
- Leverett MC (1941) Capillary behaviour in porous solids. *Transactions of AIME* 142: 152–169. <https://doi.org/10.2118/941152-G>.
- Lomeland F, Ebeltoft E, and Thomas WH (2005) A new versatile relative permeability correlation. In: *Proceedings of the International Symposium of the Society of Core Analysts, 21–25 August, Toronto, Canada, SCA2005-32*.
- Lu P-Y (1989) *The Installation Restoration Program Toxicology Guide*. vol. 2. <https://doi.org/10.2172/5440488>. ORNL/M-975-Vol.2. United States.
- McAuliffe C (1966) Solubility in water of paraffin, cycloparaffin, olefin, acetylene, cycloolefin, and aromatic hydrocarbons. *Journal of Physical Chemistry* 70(4): 1267–1275. <https://doi.org/10.1021/j100876a049>.
- Mercer JW and Cohen RM (1990) A review of immiscible fluids in the subsurface: Properties, models, characterization and remediation. *Journal of Contaminant Hydrology* 6(2): 107–163.
- Murphy NF, Lastovica JE, and Fallis JG (1957) Correlation of interfacial tension of two-phase three-component systems. *Industrial and Engineering Chemistry* 49(6): 1035–1042.
- Newell CJ, Acre SD, Ross RR, and Huling SG (1995) Light Nonaqueous Phase Liquids. In: *Ground Water Issue*. US EPA. Report No. EPA/540/S-95/500.
- Oak MJ (1990) Three-phase relative permeability of water-wet Berea. In: *Proceedings, SPE/DOE Enhanced Oil Recovery Symposium, Tulsa, Oklahoma, April*. <https://doi.org/10.2118/20183-MS>.
- Oak MJ, Baker LE, and Thomas DC (1990) Three-phase relative permeability of Berea sandstone. *Journal of Petroleum Technology* 42(8): 1054–1061. <https://doi.org/10.2118/17370-PA>.
- Parker JC and Lenhard RJ (1987) A model for hysteretic constitutive relations governing multiphase flow. 1. Saturation–pressure relations. *Water Resources Research* 23: 2187–2196.
- Parker JC, Lenhard RJ, and Koppusamy T (1987) A parametric model for constitutive properties governing multiphase flow in porous media. *Water Resources Research* 23: 618–624.
- Pearson CR and McConnell G (1975) Chlorinated C₁ and C₂ hydrocarbons in the marine environment. *Proceedings of the Royal Society of London B* 189: 305–332. <https://doi.org/10.1098/rspb.1975.0059>.
- Pentury JG (2021) *Impact of Grain Roughness on Capillary Trapping in Porous Media*. MSc Thesis United Kingdom: University of Aberdeen.
- Shiu WY, Bobra M, Bobra AM, Majanen A, Suntuo L, and Mackay D (1990) The water solubility of crude oils and petroleum products. *Oil and Chemical Pollution* 7: 57–84.
- Staronni M and Pokrajac D (2020) On the macroscopic momentum balance equation for the fluid-fluid interfaces in two-phase porous media flows. *Advances in Water Resources* 135: 103487. <https://doi.org/10.1016/j.advwatres.2019.103487>.
- Stone HL (1970) Probability model for estimating three-phase relative permeability. *Journal of Petroleum Technology* 22(2): 214–218. SPE-2116-PA. <https://doi.org/10.2118/2116-PA>.
- Tanino Y and Christensen M (2019) Imbibition capillary pressure and relative permeability of mixed-wet microporous rock: New insights from history matching. *Transport in Porous Media* 129: 121–148. <https://doi.org/10.1007/s11242-019-01280-4>.
- Tanino Y, Ibekwe A, and Pokrajac D (2020) Impact of grain roughness on residual nonwetting phase cluster size distribution in packed columns of uniform spheres. *Physical Review E* 102(1): 013109. <https://doi.org/10.1103/PhysRevE.102.013109>.
- Unger DR, Lam TT, Schaefer CE, and Kosson DS (1996) Predicting the effect of moisture on vapor-phase sorption of volatile organic compounds to soils. *Environmental Science & Technology* 30(4): 1081–1091. <https://doi.org/10.1021/es950065f>.
- USCG (US Coast Guard) (1999) *Chemical Hazard Response Information System (CHRIS) Hazardous Chemical Data*. Washington, DC: U.S. Government Printing Office.
- van Genuchten MT (1980) A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal* 44: 892–898. <https://doi.org/10.2136/sssaj1980.03615995004400050002x>.
- Zhou D and Blunt MJ (1997) Effect of spreading coefficient on the distribution of light non-aqueous phase liquid in the subsurface. *Journal of Contaminant Hydrology* 25: 1–19. [https://doi.org/10.1016/S0169-7722\(96\)00025-3](https://doi.org/10.1016/S0169-7722(96)00025-3).

Overland flow

Tammo S Steenhuis^{a,b}, Rajith Mukundan^c, and Seifu A Tilahun^{b,d}, ^aDepartment of Biological and Environmental Engineering, Cornell University, Ithaca, NY, United States; ^bFaculty of Civil and Water Resources Engineering, Bahir Dar Institute of Technology, Bahir Dar University, Bahir Dar, Ethiopia; ^cBureau of Water Supply, New York City Department of Environmental Protection, Kingston, NY, United States; ^dDepartment of Ecology and Conservation Biology, Texas A&M University, College Station, TX, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of TS Steenhuis, L Agnew, P Gérard-Marchant, MT Walter, OVERLAND FLOW, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 130–133, ISBN: 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00568-3>

Introduction	287
Early work	287
Experimental—Overland flow and watershed storage considerations	289
Overland flow models	291
Modeling overland flow with the SCS-CN	291
Saturation excess runoff models	291
Water quality impacts of saturation excess overland flow	292
Conclusion	293
References	293

Abstract

The movement of water across the soil surface is termed overland flow. Too much overland flow can cause major flooding, potentially causing damage to property and loss of life. While it is generally taught that runoff events occur when rainfall intensity exceeds the soil infiltration capacity, the most significant overland flow occurs after a large amount of rain over several days. The soil profile becomes saturated, and the rainfall cannot infiltrate and runs off. Therefore, the water stored in the watershed is a better predictor of overland flow than infiltration capacity. Best management practices that consider the runoff areas near streams are more effective than those implemented on the hillsides.

Key points

- Overland flow can cause major flooding, damage to property and loss of life.
- Most large floods are caused by saturation excess overland flow, where the soil becomes saturated, and the precipitation cannot infiltrate anymore and runs off as overland flow.
- Potential watershed storage before the rainstorm is a better indicator of the amount of overland flow than infiltration capacity.
- Infiltration capacity in well vegetated areas is usually greater than most rainfall intensities.
- Best management practices placed at runoff source areas near streams are more effective than those installed on the hillsides.

Introduction

Overland flow describes the runoff of rain before entering a stream channel without interacting with the groundwater. Commonly two mechanisms are identified for generating overland flow: infiltration excess and saturation excess. Infiltration excess runoff is defined as the process where runoff is generated when the rainfall intensity exceeds the soil infiltration capacity. Horton, (1933, 1939) measured it while performing experiments in 3.6 m long runoff plots. Hence it is referred to as ‘Hortonian runoff’. Hortonian flow is common during storms with very high rainfall intensities in arid and semi-arid areas (where significant soil crusting and surface sealing occurs), during rain events, irrigated fields, and in urban areas (Dunne and Leopold, 1978).

Early work

The Hortonian runoff concept does not meaningfully explain storm runoff in humid and sub-humid regions with well-structured soils where the infiltration capacity is greater than the average rainfall intensity (Hewlett and Hibbert, 1967; Dunne and Black, 1970). Runoff is generated in areas that are or become saturated during a rainfall event. This process is termed ‘saturation-excess

overland flow'. Unlike Hortonian flow, where overland flow can be clearly defined as water running over the surface, overland flow in saturated areas is less distinct. Overland flow may partially flow laterally through topsoil (hence the term throughflow coined by [Kitkby and Chorley, \(1967\)](#), cited by [Dunne and Black \(1970\)](#) in areas where the soils are saturated. In undisturbed forests ([Hewlett and Hibbert, 1967](#); [Sidle et al., 2000](#)) and abandoned grasslands ([Lyon et al., 2006](#)), all the flow is throughflow and overland flow is only visible over a short distance of the total flow path. [Hewlett and Hibbert, 1967](#)) wrote, 'Above the zone of saturation, we may regard such movement (i.e., lateral flow) as due to thickening of the water films surrounding soil particles and a resulting pulse in water flux as the saturated zone is approached'. [McDonnell \(2009\)](#) notes that Hewlett and Hibbert lacked any discussion of preferential flow paths. It should not be a surprise, since the notion of preferential macropore flow was only introduced in the late 1970s and early 1980s ([Bouma and Dekker, 1978](#); [Beven and Germann, 1981](#)). Thus the "thickening of the films" of Hewlett and Hibbert means that the water fills the macropores and pipes, greatly enhancing the conductivity of the soil. Consequently, throughflow occurs only when the soil becomes saturated due to a high groundwater table or a (shallow) layer restricting the downward percolation and forms a perched water table. When the soil is dryer, only the smaller pores hold the water. All early publications agree that contributions to overland flow from hillslopes without a perched water table are small during a rainstorm. On theoretical grounds, [Freeze and Witherspoon \(1968\)](#) supported this argument that water involved in the runoff event is only generated near and in the saturated area. Only over longer periods is the groundwater table affected by flow from the hillslopes and the amount of overland flow originating from the saturated area during rainfall events. Thus, connectivity in the landscape between areas, which are sources of runoff, depends on whether the soil profile is sufficiently wet so that water can be transported either overland or through large pores and pipes and along the soil-bedrock interface ([Torres et al., 1998](#)).

Saturated areas occur in the landscape where the incoming flow (both rainfall and lateral flow from upslope) is greater than water flowing out as lateral flow due to a reduction in slope, soil depth or hydraulic conductivity ([Beven and Kirkby, 1979](#); [Beven, 2001](#)). Under these conditions, the water table will eventually reach the surface given sufficient rain. Once saturated, any additional rainfall (irrespective of intensity) becomes overland flow. Areas prone to saturation occur at the bottom of slopes, usually in topographically converging regions with a high groundwater table or on hillsides with a hard pan (fragipan) at shallow depth ([Fig. 1](#)).

Overland flow from these saturated areas consists of direct runoff and return flow. Direct runoff originates from rain falling on the saturated area and running off as overland and lateral flow ([Hewlett and Hibbert, 1967](#)). Return flow occurs after the rain stops when the soils drain from saturation to a quasi-equilibrium moisture content, when all large pores are empty and the hydraulic gradient in the vertical direction is constant ([Brindt et al., 2022](#)).

The quasi-equilibrium moisture content is defined, similarly to field capacity of well-drained soils, as the moisture content at which the direct drainage is negligible. While for well-drained soils, field capacity is reached after saturating a soil when the unsaturated hydraulic conductivity becomes limiting, for soils with a high water table or fragipan at shallow depth, the vertical flow stops when the hydraulic gradient becomes constant with depth. The latter occurs when the matric potential decreases linearly with height above the water table, with the same quantity as the gravity potential increases with height. Therefore, the moisture content at equilibrium conditions with height can be found through the soil moisture characteristic curve ([Brindt et al., 2022](#)). If the soil is sloping, lateral flow occurs until the large pores are empty (defined as pores greater than 0.3–0.8 mm in diameter), when the conductivity of the soil matrix is one or two orders smaller than saturated hydraulic conductivity.

During periods of rainfall, the soil wets up and reaches saturation. As the rain continues, the area of saturated soil expands, and the overland flow per unit of rainfall will increase. Conversely, dry periods will decrease the extent of the saturated area, and the soil might dry out, depending on its location in the landscape. Hence the term 'variable source area' (VSA) comes with the expanded saturated area ([Dunne and Leopold, 1978](#)). The extent of the saturated area is illustrated in [Fig. 2](#) for the Catskills in upstate New York, where potential evaporation exceeds average rainfall during the summer and potential evaporation is less than precipitation during the winter ([Walter et al., 2000](#)).

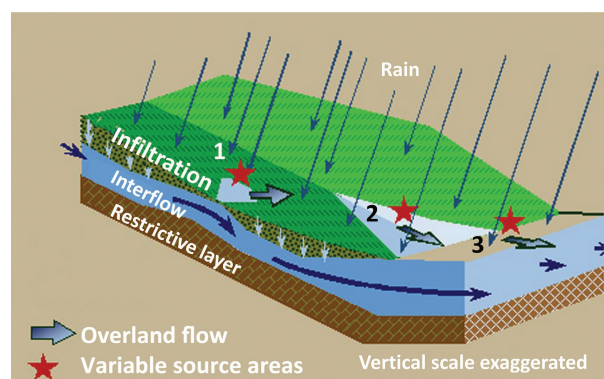


Fig. 1 Incidence of saturation-excess hydrology: 1, shallow soil; 2, convergence area; and 3, decreasing downhill slope. VSA, variable source area. http://soilandwater.bee.cornell.edu/research/VSA/processes/processes_sat.html

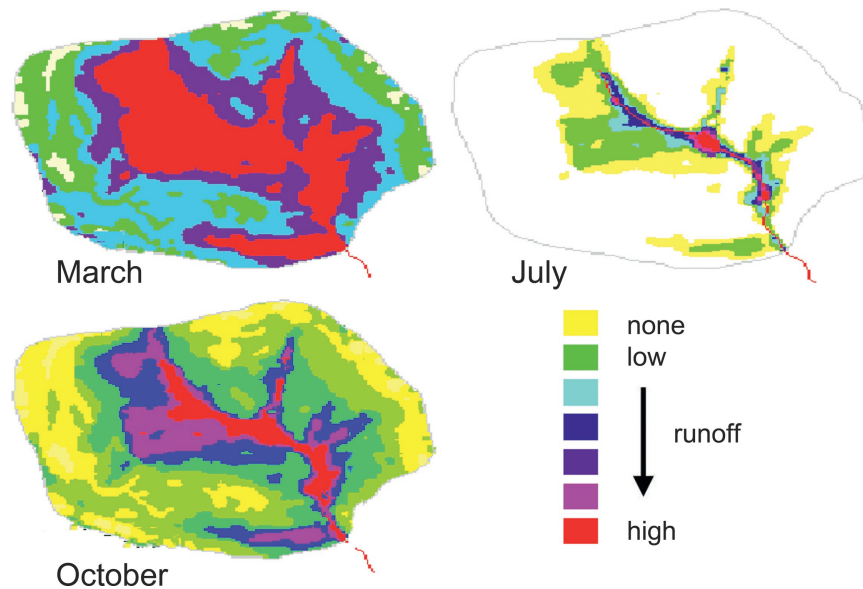


Fig. 2 Seasonal changes of saturated areas (variable source areas) in a watershed in New York state. Color denotes areas more or less prone to saturation (Walter et al., 2000).

Experimental—Overland flow and watershed storage considerations

An interesting case study on saturation excess overland flow is the 95 ha Debre Mawi watershed near Lake Tana in Ethiopia. In this watershed, discharge was observed for 9 years from 2010 to 2018 at the outlet and in 4 small watersheds nested within. In the second and third years, the government installed, with the help of the farmers, soil and water conservation structures consisting of 60 cm deep and 50 cm wide infiltration furrows covering 70% of the watershed area (Mhired et al., 2020). These furrows were filled with soil within 2–3 years and were ineffective during the last 4 years of the case study. The area of Eucalyptus trees increased 10-fold to 5% of the watershed area in 2018 from 0.5 ha of small eucalyptus trees in 2010.

Yearly rainfall in the Debre Mawi watershed is around 1200 mm/year, of which 80% falls during the rain phase from the beginning of June to the middle of September. Over the study period, annual rainfall increased on average by 23 mm/year. Once the topsoil dries out during the dry phase, actual evaporation is negligible. Eucalyptus, unlike native trees, only transpire during the dry phase. Because Eucalyptus was planted in small woodlots, they transpire approximately twice the potential evaporation rate (about 10 mm/day) during the dry monsoon phase (Enku et al., 2020).

In the Debre Mawi watershed, saturation excess overland flow is generated from the valley bottom once sufficient rain has fallen. A hillside in the Debre Mawi watershed is shown in the photograph of Fig. 3. The grass is in the foreground, and cropland is shown further upslope. Two piezometers, P9 and P8, were installed in the grassland and one piezometer P7 in the cropland in 2018. The graph in Fig. 3 shows the water table depths after the piezometers were installed on July 15. In the grassland, the water table was at the surface from the first week of August until the middle of September. The water table in the cropland (piezometer P7) only increased slightly in the middle of July after a 50 mm rainfall. A volcanic dike perpendicular to the slope caused the high water table in the grassland that forced the water to the surface. This area was too wet to grow crops. Only higher up the slope, could crops be grown, where the soils were sufficiently drained, as shown by the groundwater depth readings in P7.

The runoff coefficients for the overland flow for all storms with rainfall above 8 mm are shown in the graph in Fig. 3. Storms on consecutive days were combined. The overland flow was minimal before July 15, when the water table rose and decreased rapidly after September 15 with the falling water tables in the grassland. This demonstrated clearly that overland flow was dependent on the degree of saturation in the watershed. Degraded soils that have a slowly permeable layer at 20–80 cm depth also generate overland flow from the hillside.

The effect of the increasing land area of eucalyptus trees on annual runoff at the outlet of the watershed was investigated over 9 years. We plotted the annual runoff coefficient instead of runoff to remove some of the annual rainfall variability (Fig. 4). The coefficient is defined as the yearly overland flow divided by the rainfall during the rain phase. In addition, we plotted in Fig. 5 the cumulative rainfall for 3 mm of runoff to occur. Annual runoff coefficients trended downward with the increased acreage of mature eucalyptus trees (Fig. 4).

The reduction in yearly runoff is caused by an increase in cumulative rainfall before significant runoff starts (Fig. 5). The explanation is as follows. Since runoff is generated by saturation excess runoff, the watershed needs to “fill up” before any significant overland flow can occur (Tilahun et al., 2016). The storage is depleted during the dry phase. The increased land area of eucalyptus trees decreases the soil storage to a greater degree because deep roots can reach a greater portion of the water. Consequently, more

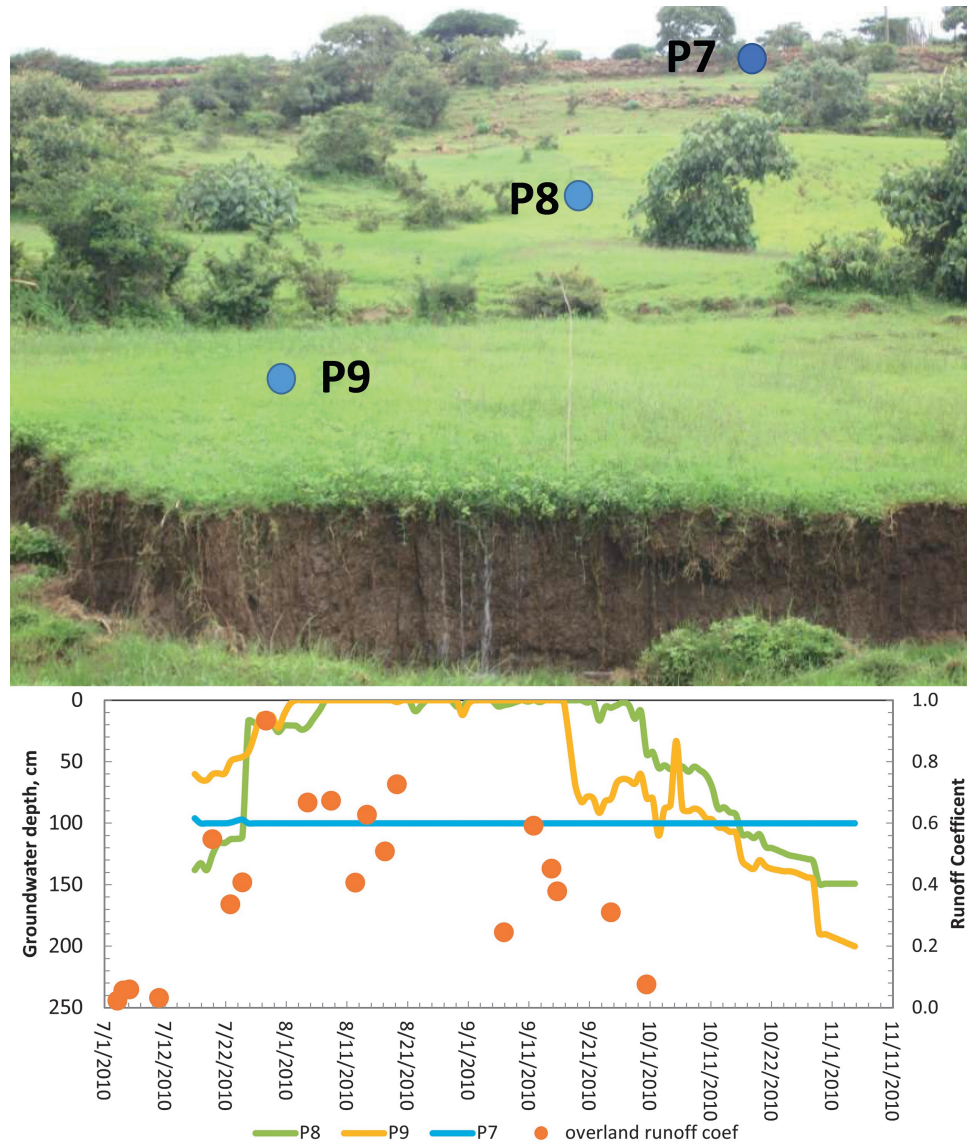


Fig. 3 Photograph of a hillslope in the Debre Mawi from the top watershed divide with cropland to the bottom with grassland where it drains. The cropland looks small, but the length is about four times that of the grassland; The graph shows the groundwater table depths for the piezometers in the grassland (P8, P9) and the cropland (P7) and runoff coefficients at the watershed outlet.

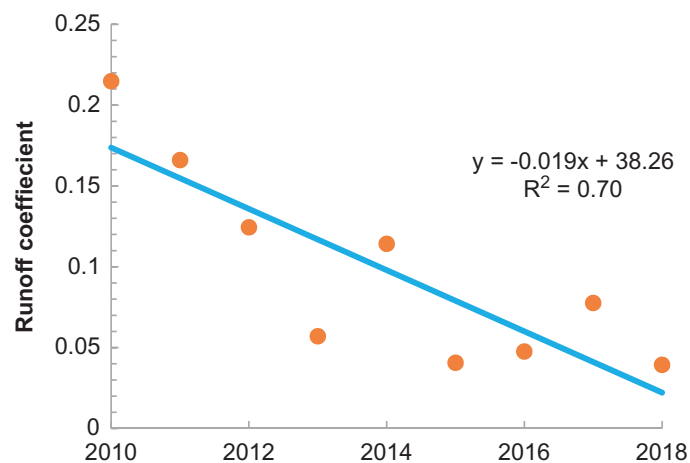


Fig. 4 Annual runoff coefficient (i.e., quotient of annual runoff and annual rainfall in the monsoon rain phase) for the Debre Mawi watershed from 2010 to 2018.

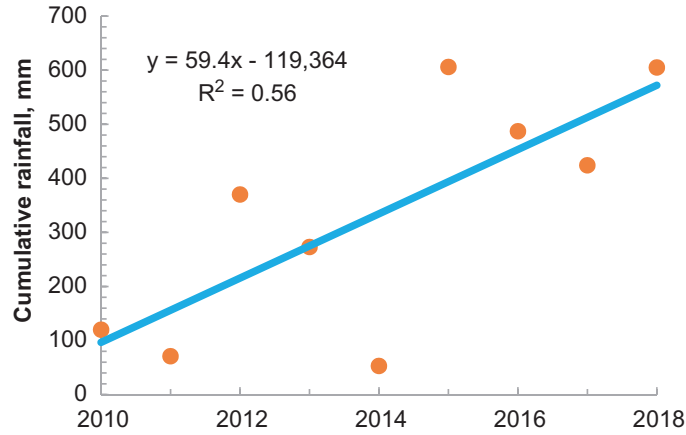


Fig. 5 Amount of precipitation from the beginning of the rain phase to cause the first 3 mm of cumulative runoff at the outlet for the 95 ha Debre Mawi watershed.

precipitation is needed to fill up the depleted storage. The implication is that overland flow is decreasing by the amount that evaporation is increasing. Thus, a fixed amount of rainfall is needed each year to fill up the storage. It means that if the rain decreases, the runoff will be less by the same amount and not by a fraction, as would be the case if infiltration excess is the primary runoff mechanism. It also explains some of the variability around the trendline in Fig. 5. For example, 2015 had the least rainfall of the 9 years and the lowest runoff coefficient.

Overland flow models

Modeling overland flow with the SCS-CN

Infiltration excess type models predict runoff when the rainfall intensity exceeds the infiltration capacity of the soil or, more often, use the Soil Conservation Service Curve Number method (SCS-CN) as interpreted in the SCS handbooks. In this approach, curve numbers are assigned based on soil type and plant characteristics that determine the infiltration rate in the soil.

Although the SCS-CN was construed as an infiltration excess model, it is more likely to describe saturation excess runoff. This is because Victor Mockus developed the SCS-CN equation in the 1950s and 1960s to produce rainfall-runoff curves of a type found in natural watersheds. The watersheds referred to by Mockus are in the Appalachian mountains where saturation excess dominates.

It can be shown theoretically that the SCS-CN method is based on saturation excess by taking the derivative of the discharge with respect to the precipitation. The derivative is equal to the extent of the saturated portion A_s of the watershed when one assumes that rainfall runs off from the saturated area and infiltrates from unsaturated soil (Steenhuis et al., 1995). It can be expressed as

$$A_s = 1 - \frac{\bar{S}^2}{[P + \bar{S}]^2} \quad (1)$$

where P is the effective rainfall (total precipitation minus the amount that brings the soil to the equilibrium moisture content) and $\bar{S}$ is the average storage in the watershed, which is related to the curve number.

Based on the amount of rain required to saturate the soil, the available storage before runoff occurs, σ , can be expressed as a function of the area that becomes subsequently saturated (Schneiderman et al., 2007)

$$\sigma = \bar{S} \left[\sqrt{\frac{1}{1 - A_s}} - 1 \right] \quad (2)$$

where σ is the available storage in the soil (or the amount of water needed to bring the soil to saturation from field capacity) as a function of the area that becomes subsequently saturated, A_s . Hence $A_s = 0$ for the wettest area with the highest topographic index and $A_s = 1$ for the area with the lowest topographic index.

The propensity to saturate can be found by ranking the landscape according to the topographic index defined (Beven and Kirkby, 1979) as the volume of water delivered per unit contour length divided by the ability of the soil profile to transmit the flow (Beven and Kirkby, 1979). Thus, the extent of the saturated area can be found from the total effective rainfall and average storage in the watershed. The latter can be linked with a curve number.

Saturation excess runoff models

The SWAT (Soil and Water Assessment Tool) model is likely the most widely used worldwide because it is open source and has an extensive literature database that makes it easy to use. However, because it predicts runoff according to infiltration excess, the runoff

source areas are not the same as in a watershed with saturation excess runoff. Changing the location of the runoff areas can be easily accomplished by redefining the soils according to the topographic index in wetness classes. By ranking the wetness classes according to the topographic index, the available storage can be calculated, and runoff can be determined as the amount of rain that cannot be stored in the soil. Models such as SWAT-VSA (Easton et al., 2008) and SWAT-wil (Steenhuis et al., 2019) are based on this principle. SWAT-HS (Hoang et al., 2017) introduces a surface interflow reservoir to accommodate flow between wetness classes.

Another model to predict the saturated-excess overland flow is the TOPMODEL. It is based on calculating a groundwater table height. When the groundwater intersects with the soil surface, runoff occurs. In the TOPMODEL, the assumption is made that a groundwater table underlies the watershed, which is unrealistic for many mountainous areas. However, since most of the saturated areas occur near streams, this limitation is not severe as a groundwater table is present in these areas. An alternative method for simulating saturated areas in a sloping watershed with a hardpan is the DHSVM (Wigmosta et al., 1994) and SMR (Frankenberger et al., 1999) models. In these models, water is routed by interflow from one cell to the next.

Water quality impacts of saturation excess overland flow

Water quality of overland flow depends on the soil chemical concentration at the location where it is generated. Therefore, best management practices to improve water quality should consider the location of surface runoff. Many practices aimed at reducing contaminant transport, such as terraces, soil bunds and infiltration furrows implemented on the hillsides, are based on Hortonian concepts. However, the effectiveness of these practices is diminished where Hortonian processes are not governing runoff generation.

Wherever VSA hydrology is a dominant process, there will be regions within a watershed that are more susceptible to producing runoff and delivering it to surface water bodies than other regions. These areas can be considered hydrologically sensitive areas (HSA). Recognizing the existence of HSAs allows watershed-scale water quality efforts to be focused on those areas where HSAs coincide with land uses that potentially contribute pollutants (Walter et al., 2000; Rittenburg et al., 2015). The intersection of an HSA and a pollutant loading area is called the 'critical management zone.' The best management practice for this area would be limiting or prohibiting polluting activity from these HSAs. A contrasting approach is finding methods and means of eliminating the hydrological sensitivity of the critical zones, although recent research suggests attempts to remove hydrologic sensitivity through drainage practices may reroute pollutants from overland flow to subsurface flow, with little reduction in concentration at a downstream target location. In either case, the HSAs warrant primary attention when preserving or improving water quality.

One example in upstate New York, where HSAs play a role in management practices, is whether dairy operators should spread manure on steep slopes or in flat areas. The past dogma was to avoid manure spreading on steep slopes and instead maximize spreading in flat areas. Under Hortonian flow, steep areas might arguably produce the most rapidly moving runoff and, therefore, the greatest potential for erosion and transport of manure. However, saturation excess overland flow is identified as the dominant process, where all rainwater infiltrates on the sloping lands and drains very rapidly, resulting in almost no substantial runoff (Hoang et al., 2019). Conversely, the flat areas, especially at the base of hillslopes, are especially prone to saturation and, thus, prone to runoff generation (Walter et al., 2000; Rittenburg et al., 2015). Fig. 6 shows manure deposition in the wrong place from an HSA standpoint. A manure-spreading policy more consistent with the recognized hydrology would promote spreading manure higher up in the watershed and minimize spreading on low, flat areas susceptible to runoff generation (Hoang et al., 2019).



Fig. 6 Manure deposition in the wrong place from a hydrologically sensitive area standpoint (courtesy of Glenn Warner). Courtesy of Glenn Warner. Professor Emeritus. Natural Resources & the Environment University of Connecticut, Storrs, Connecticut

Conclusion

While infiltration excess runoff is generally considered the main cause of runoff, saturation excess on vegetated surfaces is the cause of many large floods. Especially during large storms, rainwater fills up the soils, and any additional rain becomes runoff. It is therefore essential to consider both runoff processes in the design of flood control structures and best management practices to improve water quality.

References

- Beven KJ (2001) *Rainfall-Runoff Modelling: The Primer*. Chichester, UK: Wiley.
- Beven K and Germann P (1981) Water-Flow in soil macropores. 2. A combined flow model. *Journal of Soil Science* 32: 15–29.
- Beven KJ and Kirkby MJ (1979) A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological Sciences Journal* 24: 43–69.
- Bouma J and Dekker CW (1978) A case study on infiltration into dry clay soil. I. Morphological observations. *Geoderma* 20: 27–40.
- Brindt N, Pacenka S, Richards BK, et al. (2022) Self-organizing hydrological processes in a runoff source area. *Catena*. <https://doi.org/10.1016/j.catena.2021.105955>.
- Dunne T and Black RD (1970) An experimental investigation of runoff production in permeable soils. *Water Resources Research* 6: 478–490.
- Dunne T and Leopold LB (1978) *Water in Environmental Planning*. New York: W.H. Freeman.
- Easton ZM, Fuka DR, Walter MT, et al. (2008) Re-Conceptualizing the soil and water assessment (SWAT) model to predict runoff from variable source areas. *Journal of Hydrology* 348: 279–291.
- Enku T, Melesse AM, Ayana EK, et al. (2020) Groundwater use of a small eucalyptus patch during the dry monsoon phase. *Biologia* 75: 853–864.
- Frankenberger JR, Brooks ES, Walter MT, Walter MF, and Steenhuis TS (1999) A GIS-based variable source area model. *Hydrological Procedure* 13: 804–822.
- Freeze RA and Witherspoon PA (1968) Theoretical analysis of regional ground water flow: 3. Quantitative interpretations. *Water Resources Research* 4: 581–590.
- Hewlett JD and Hibbert AR (1967) Factors affecting the response of small watersheds to precipitation in humid regions. In: Sopper WE and Lull HW (eds.) *Forest Hydrology*, pp. 275–290. Oxford, UK: Pergamon Press.
- Hoang L, Schneiderman EM, Moore KEB, et al. (2017) Predicting saturation-excess runoff distribution with a lumped hillslope model: SWAT-HS. *Hydrological Procedure* 31: 2226–2243.
- Hoang L, Mukundan R, Moore KEB, et al. (2019) Phosphorus reduction in the New York City water supply system: A water-quality success story confirmed with data and modeling. *Ecological Engineering* 135: 75–88.
- Horton RE (1933) The role of infiltration in the hydrologic cycle. *Eos, Transactions of the American Geophysical Union* 14: 446–460.
- Horton RE (1939) Analysis of runoff-plat experiments with varying infiltration-capacity. *Eos, Transactions of the American Geophysical Union* 20: 693–711.
- Kitkby MJ and Chorley RJ (1967) Overland flow, throughflow and erosion. *International Association of Scientific Hydrology. Bulletin* 12(2): 5–21.
- Lyon SW, Seibert J, Lembo AJ, Walter MT, and Steenhuis TS (2006) Geostatistical investigation into the temporal evolution of spatial structure in a shallow water table. *Hydrology and Earth System Sciences* 10: 113–125.
- McDonnell J (2009) Hewlett, J.D. and Hibbert, A.R. 1967: Factors affecting the response of small watersheds to precipitation in humid areas. In Sopper, W.E. and Lull, H.W., editors, *Forest hydrology. Progress in Physical Geography* 33: 288–293.
- Mhiret DA, Dagnew DC, Guzman CD, et al. (2020) A nine-year study on the benefits and risks of soil and water conservation practices in the humid highlands of Ethiopia: The Debre Mawi watershed. *Journal of Environmental Management* 270: 110885.
- Rittenburg RA, Squires AL, Boll, et al. (2015) Agricultural BMP effectiveness and dominant hydrological flow paths: Concepts and a review. *Journal of the American Water Resources Association (JAWRA)* 305–329.
- Schneiderman EM, Steenhuis TS, Thongs DJ, et al. (2007) Incorporating variable source area hydrology into the curve number based Generalized Watershed Loading Function model. *Hydrological Procedure* 21: 3420–3430.
- Sidele R, Tsuboyama Y, Noguchi S, et al. (2000) Stormflow generation in steep forested headwaters: A linked hydrogeomorphic paradigm. *Hydrological Procedure* 14: 369–385.
- Steenhuis TS, Winchell M, Rossing J, Zollweg JA, and Walter MF (1995) SCS runoff equation revisited for variable-source runoff areas. *Journal of Irrigation & Drainage Engineering* 121: 234–238.
- Steenhuis TS, Schneiderman EM, Mukundan R, et al. (2019) Revisiting SWAT as a saturation-excess runoff model. *Watermark* 11: 1427.
- Tilahun SA, Ayana EK, Guzman CD, et al. (2016) Revisiting storm runoff processes in the upper Blue Nile basin: The Debre Mawi watershed. *Catena* 143: 47–56.
- Torres R, Dietrich WE, Montgomery D, Anderson S, and Loague K (1998) Unsaturated zone processes and the hydrologic response of a steep, unchanneled catchment. *Water Resources Research* 34: 1865–1879.
- Walter MT, Walter MF, Brooks ES, et al. (2000) Hydrologically sensitive areas: Variable source area hydrology implications for water quality risk assessment. *Journal of Soil and Water Conservation* 55: 277–284.
- Wigmosta MA, Vail LW, and Lettenmaier DP (1994) A distributed hydrology-vegetation model for complex terrain. *Water Resources Research* 30: 1665–1679.

Direction-dependency of soil hydraulic and mechanical properties

Steffen Beck-Broichsitter^a, Heiner Fleige^a, Jose Dörner^b, and Rainer Horn^a, ^aInstitute of Plant Nutrition and Soil Science, Christian-Albrechts-University Kiel, Kiel, Germany; ^bInstituto de Ingeniería Agraria y Suelos, Facultad de Ciencias Agrarias y Alimentarias, Universidad Austral de Chile, Valdivia, Chile

© 2023 Elsevier Ltd. All rights reserved.

Introduction	294
Anisotropy of air and water flow in the vadose zone	294
Anisotropy of gas diffusivity, air-permeability, and air-filled porosity	295
Anisotropy of saturated and unsaturated hydraulic conductivity and permeability	297
Anisotropy of heterogenous structured materials analyzed by X-ray CT image analysis	301
(An-)Isotropy of mechanical properties and functions in structured soils	301
Stress strain processes in structured unsaturated soils	301
Quantification of structure dependent changes by tensorial stress strain and shear stress components	302
Consequences of the strength anisotropy for land sliding processes	303
Conclusion	304
References	304

Abstract

Soil structure induces vertical or horizontal direction dependency in water flow and mechanical behavior. These effects are exacerbated by anthropogenic induced plate structure formation, where soil deformation due to mechanical loads can lead to a distinct horizontal anisotropy of the soil pore system. This promotes surface runoff and thus increased soil erosion and may reduce biomass productivity. The direction-dependent differences in intensity parameter (i.e., hydraulic conductivity, air-permeability) due to soil deformation can be effectively linked to the pore size distribution and continuity.

Key points

- Direction-dependent differences in soil mechanical or transport properties are strongly related to the pore size distribution and continuity
- The preferential direction of air or water transport in soil is horizontal under compaction and vertical under cracking.
- Prismatic soil structure indicates slightly greater shear resistance in vertical than horizontal directions
- Platy soil structure indicates slightly greater shear resistance in horizontal than vertical directions

Introduction

Soil physical properties may vary from one location to another (heterogeneity), as well as from one direction to another at the same location (anisotropy). While the directional behavior (isotropy or anisotropy) of soil hydraulic conductivity is of primary importance in analyzing rates and directions of water flow and solute transport in the vadose zone, in terms of mechanical properties it is of great relevance in evaluating soil deformation processes since each mechanical stress applied to the soil surface results in a 3-dimensional stress propagation.

The directional behavior of soil physical properties governs not only hydrologic processes (e.g., infiltration, recharge, evaporation, and runoff) but also soil aeration and root growth. It has been documented that the induced mechanical load can result in a platy soil structure when the induced mechanical stresses exceed the stability of the soil structure. The latter results in direction-dependent behavior of e.g., air permeability and hydraulic conductivity due to a predominantly higher pore continuity and connectivity in a horizontal than in a vertical direction. At the same time, since the mechanical strength also has direction-dependency, root growth can be impeded vertically, affecting potential soil exploration. The following chapter examines both the direction dependency of hydraulic and mechanical properties in soils, considering theoretical aspects, results of different studies, and new methodologies used to evaluate direction dependent processes in soils.

Anisotropy of air and water flow in the vadose zone

Soils generally consist of several and differently structured horizons with direction-dependent properties due to the textural composition and soil aggregate formation (Dörner and Horn, 2006). Even in the same layer, the heterogeneous arrangements of

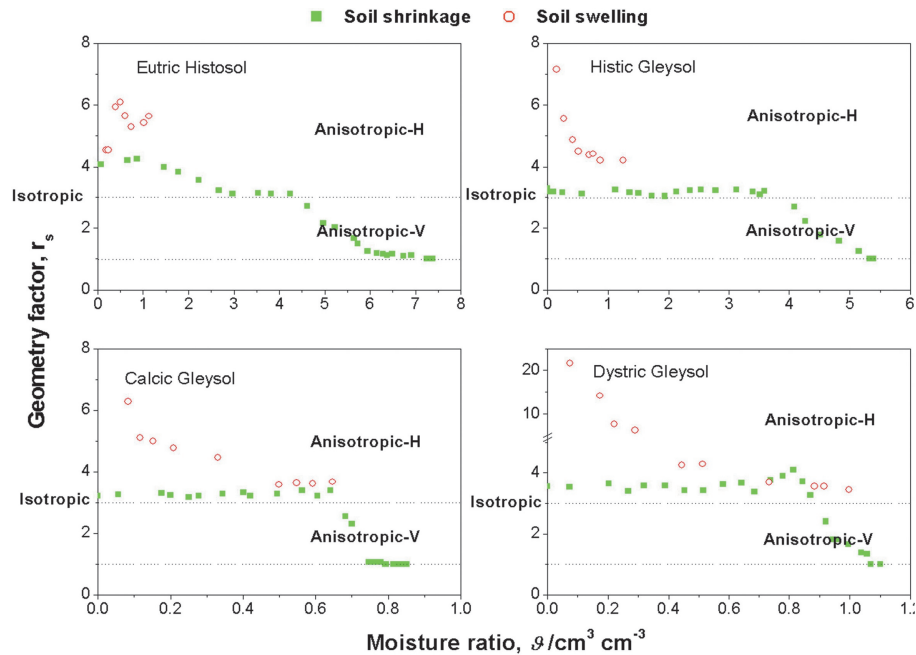


Fig. 1 Effect of swell shrinkage processes on the geometry factor, r_s , as an expression of the isotropy or anisotropic volume changes (Peng and Horn, 2007).

soil particles and aggregates due to anthropogenic and pedogenetic effects result in a discrepancy of pore continuity and tortuosity between the different directions (Beck-Broichsitter et al., 2020c). The resulting anisotropy is generally influenced by soil structure (e.g., platy, columnar, etc.), therefore exhibiting a pattern of pores with a distinct directional bias. Such heterogeneous soil structure has been found to influence soil deformation (Peng and Horn, 2007), soil strength and transport processes (Dörner and Horn, 2006).

The anisotropy of air and water flow through unsaturated porous media can vary depending on time, land use management, degree of saturation, and annually varying wetting and drying cycles, biological activity, or anthropogenic impacts (Reszkowska et al., 2011). Air and water fluxes in soils are mostly multidimensional and their direction-dependency is strongly reliant on aggregate formation and orientation, as well as pore size distribution (Horn and Smucker, 2005; Dörner and Horn, 2006). In addition, soil compaction can increase the probability of preferential flow through existing or emerging macropores produced by pore system restructuring (i.e., crumbly to platy). For instance, topsoil compaction may cause a platy structure to form, but over time it may change to crumbly through swelling-shrinkage or freezing-thawing cycles (Jarvis, 2007) or deep-rooting plants (Beck-Broichsitter et al., 2020c) (Fig. 1).

Swelling and shrinkage can produce anisotropic processes (Fig. 1). Starting from a completely homogenized saturated soil, the initial vertical anisotropic shrinkage during drying is followed by a primarily horizontal anisotropic swelling (Hartge and Horn, 2016). Repeated shrinkage and swelling patterns pass through a complete isotropic volume change, but anisotropy can occur from strengthening that induces primarily horizontal volume changes. Dörner and Horn (2009) also demonstrated these processes for various soils and geological origin with the help of the geometry factor, r_s to measure anisotropy.

A platy soil structure is typical for engineered barriers (Beck-Broichsitter et al., 2020b) and mechanical-induced topsoil compaction of arable fields (Schlüter et al., 2018). Another cause of direction dependent fluxes in arable soils is from subsoil compaction. Fig. 2 presents the results of water flow modelling in a structured Stagnic Luvisols in Northern Germany under arable management (Dörner and Horn, 2006). Anisotropic conditions of the saturated hydraulic conductivity occurred due to plow pan formation that led to a deviation of the water flow vectors, generating preferential flows parallel to the slope at 15–25 cm depth. In contrast, in well-aggregated soil, isotropic conditions generally occur, where no statistically significant differences in the hydraulic conductivities between the vertically and horizontally oriented samples are found (Reszkowska et al., 2011).

Anisotropy of gas diffusivity, air-permeability, and air-filled porosity

Soil aeration depends on gas diffusivity, D_p , and air-permeability, k_a , which is largely affected by the air-filled porosity, θ_a . Whereas D_p and k_a describe the intensity of transport, θ_a describes the capacity of transmissive pores. They are therefore referred to as intensity and capacity parameters, respectively. All these parameters are a function of initial soil structure and stability (Beck-Broichsitter et al., 2020a), pore continuity, connectivity, and tortuosity (Schjønning and Koppelgaard, 2017), degree of compaction, volumetric water content, and pore size distribution (Chief et al., 2008).

Gas diffusivity and air-permeability are tensors that can indicate isotropic or anisotropic behavior due to direction-dependent D_p - and k_a -values in three spatial dimensions (Dörner and Horn, 2006). The one-dimensional advective steady-state air flow in soils can be described by Darcy's law as,

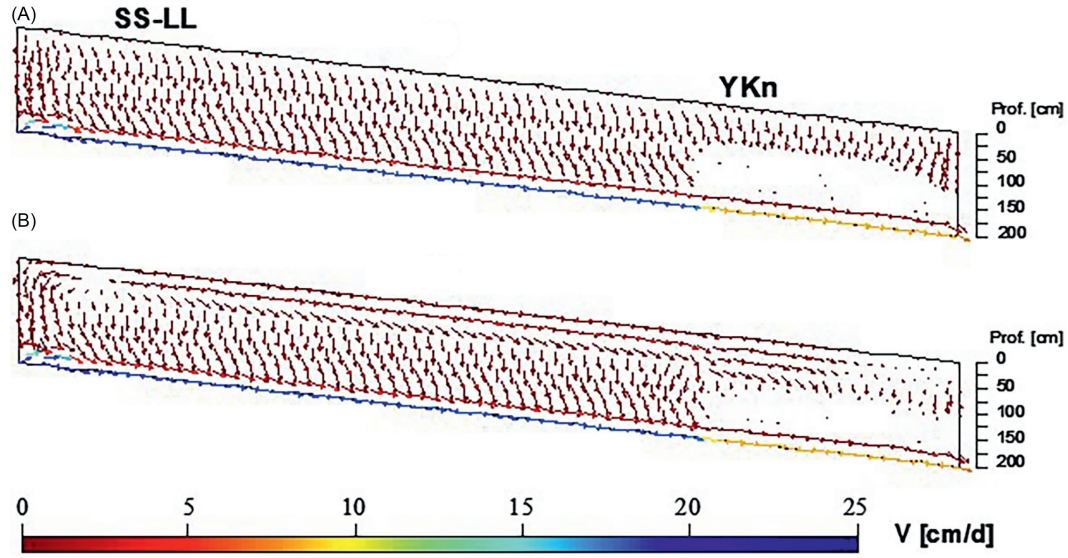


Fig. 2 Water flow vectors (v) with (A) isotropic conditions of saturated hydraulic conductivity and (B) anisotropic saturated hydraulic conductivity at the plow pan depth (15–25 cm). SS-LL: Stagnic Luvisol, YKn: Anthrosol derived from colluvic material (Dörner and Horn, 2006). The soil profile can be seen in Fig. 5.

$$\nabla^2 \left(\frac{k_a}{\eta} p_e \right) = 0 \quad (1)$$

where η is the air viscosity (Pa s^{-1}), k_a is the air permeability (m^2), p_e (Pa) is the differential pressure from atmospheric pressure, p_a (Pa), and at steady-state ($p_e < p_a$).

The three-dimensional air flow in soils can be numerically solved as (i.e., Chief et al., 2008),

$$\frac{k_{ax}}{\eta} \left(\frac{\partial^2 p_e}{\partial x^2} \right) + \frac{k_{ay}}{\eta} \left(\frac{\partial^2 p_e}{\partial y^2} \right) + \frac{k_{az}}{\eta} \left(\frac{\partial^2 p_e}{\partial z^2} \right) = 0 \quad (2)$$

The anisotropy ratios, AR, can be expressed in a direction-dependent form (Table 1), where θ_{axi} , k_{axi} , θ_{aBxi} are air-filled porosity, air permeability, and blocked-air-filled porosity in x direction, and θ_{azi} , k_{azi} , θ_{aBzi} are the corresponding parameter in z direction, while assuming x direction = y direction (Ball et al., 1988, 2001).

For sandy, coarse-loamy, and fine-silty arable soils and in a bulk density range between 1.3 g cm^{-3} and 1.5 g cm^{-3} , the k_a -derived anisotropy ratios can vary between 0.06 and 15.8 (Chief et al., 2008). In Fig. 3, for loamy sand with crumbly structure in a bulk density range between 1.15 g cm^{-3} and 1.25 g cm^{-3} , the AR_{ka} values can vary between 0.5 and 1.4, while a bulk density range between 1.75 g cm^{-3} and 1.85 g cm^{-3} and a platy structure result in AR_{ka} values between 1.5 and 2.2 (Beck-Broichsitter et al., 2020a). Furthermore, anisotropy conditions might introduce significant errors of up to two orders of magnitude in on-site intrinsic permeability measurements (Chief et al., 2008).

The blocked air-filled porosity in both directions can reach up to $1/4$ of a soil's total porosity (Dörner and Horn, 2006; Beck-Broichsitter et al., 2020a). In aggregated soils, lower θ_{aB} values can be expected along the dominant axis presenting an anisotropic behavior. The blocked-air-filled porosity is the pore space that does not take part in convective air transport (Ball et al., 1988).

In arable used soils, the $AR_{\theta_{aB}}$ values can vary between 0.3 and 0.9 in coherent, angular-blocky, and subangular-blocky topsoil layer ($\leq 35 \text{ cm}$ depth) and up to 12.0 in platy-structured plow pans, while subsoil layers ($> 35 \text{ cm}$ depth) with subangular-blocky

Table 1 Standard formulas calculating the air-filled porosity, θ_a , the air permeability (μm^2), k_a , and the blocked air-filled porosity, θ_{aB} ($\text{cm}^3 \text{ cm}^{-3}$), at each measurement step (i.e., saturation, water content, pore size distribution), I, respectively in x and z direction, while assuming $x = y$ (Ball et al., 2001; Dörner and Horn, 2006).

Standard formulas	Anisotropy ratios
$\theta_a = \left(1 - \frac{\rho_b}{\rho_s} \right) - \theta_v$	$AR_{\theta_a} = \sum_{i=1}^n \frac{\Delta \theta_{axi}}{\Delta \theta_{axi}}$
$k_a = M \cdot \theta_a^N$	$AR_{k_a} = \sum_{i=1}^n \frac{\Delta k_{axi}}{\Delta k_{axi}}$
$\theta_{aB} = 10^{\frac{(-\log(M))}{N}}$	$AR_{\theta_{aB}} = \sum_{i=1}^n \frac{\Delta \theta_{aBxi}}{\Delta \theta_{aBxi}}$

N and M are empirical parameter (Ball et al., 1988), ρ_b (g cm^{-3}) is the dry bulk density, ρ_s (g cm^{-3}), is particle density, θ_v the volumetric water content ($\text{cm}^3 \text{ cm}^{-3}$), and AR_{θ_a} , AR_{k_a} , and $AR_{\theta_{aB}}$ the corresponding anisotropy ratios.

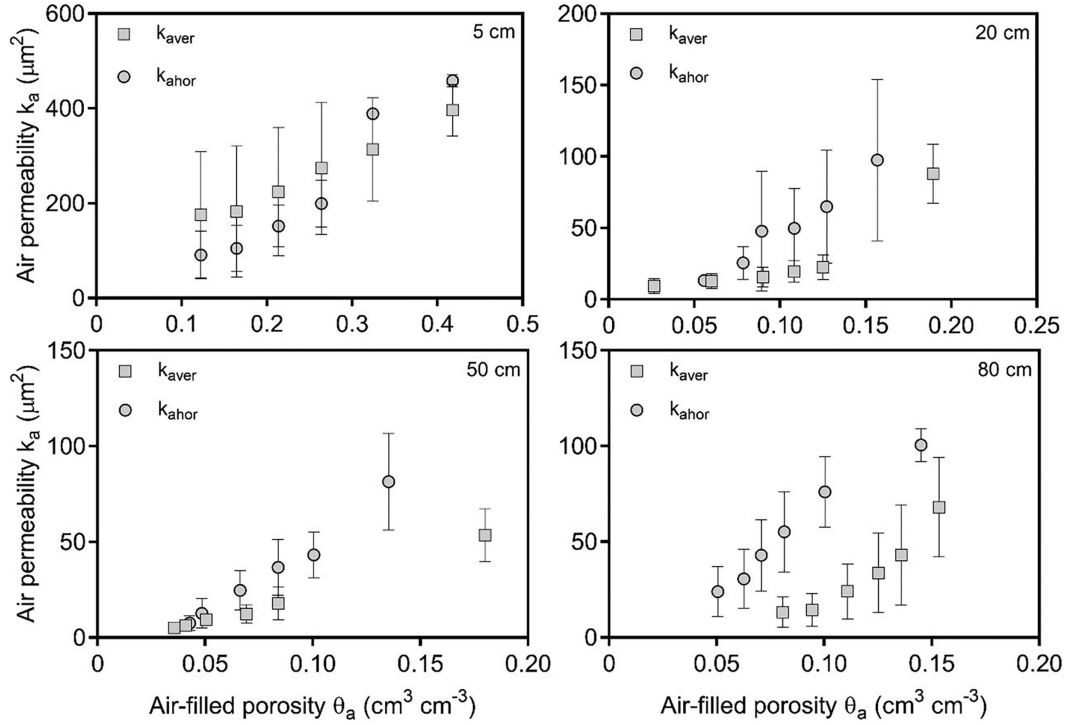


Fig. 3 Air permeability, k_a (μm^2), versus air-filled porosity, θ_a ($\text{cm}^3 \text{cm}^{-3}$), of a loamy sand with ρ_b between 1.15 g cm^{-3} and 1.25 g cm^{-3} in 5 cm depth, and ρ_b between 1.75 g cm^{-3} and 1.85 g cm^{-3} in 20 cm, 50 cm, and 80 cm depths, samples taken in vertical (ver) and horizontal (hor) direction (Beck-Broichsitter et al., 2020a). The dashed lines indicate the confidence interval CI: 95%.

structure tend to $AR_{\theta_{aB}}$ values between 0.6 and 5.8 (Dörner and Horn, 2006). A similar tendency is shown by Beck-Broichsitter et al. (2020b) with $AR_{\theta_{aB}}$ values < 1.0 for crumbly structured topsoils and $AR_{\theta_{aB}}$ values up to 6.8 for platy structured subsoils.

The one-dimensional diffusive gas flux in soils is controlled by the soil-gas diffusion coefficient, D_p , outlined as soil-gas diffusivity D_p/D_0 and can be expressed with a predictive model (i.e., Moldrup et al., 2013) as,

$$\frac{D_p}{D_0} = p\theta_a^\chi \left[\frac{\theta_a}{\theta_T} \right]^{T_a} \quad (3)$$

where θ_T is the total porosity ($\text{m}^3 \text{m}^{-3}$), p , χ , and T_a are model parameters.

The typical fitting parameters for pure sand and sand size fractions are $p = 1$ and $\chi = 4/3$, and $p = 1$, while for soils with total porosities up to $0.5 \text{ cm}^3 \text{cm}^{-3}$ for fractured, clayey soils $\chi = 2$ (Moldrup et al., 2013).

The three-dimensional soil gas diffusion assuming quasi-steady state behavior (Maier et al., 2020) can be numerically solved by Fick's law as,

$$\frac{D_{px}}{D_0} \left(\frac{\partial^2 C}{\partial x^2} \right) + \frac{D_{py}}{D_0} \left(\frac{\partial^2 C}{\partial y^2} \right) + \frac{D_{pz}}{D_0} \left(\frac{\partial^2 C}{\partial z^2} \right) = 0 \quad (4)$$

where D_{px}/D_0 is the soil-gas diffusivity in x direction, D_{py}/D_0 is soil-gas diffusivity in y direction. Assuming $D_{px}/D_0 = D_{py}/D_0$, the anisotropy ratio can be calculated as,

$$AR_{\frac{D_p}{D_0}} = \sum_{i=1}^n \frac{\frac{D_{px}}{D_0}}{\frac{D_{pz}}{D_0}} \quad (5)$$

It must be remembered that soil gas diffusivity in multidimensional porous media depends on the rigidity of soil structure (Horn et al., 2019) and therefore the tortuosity and pore connectivity is affected by swelling and shrinkage or freezing and thawing cycles (Villagra et al., 2020), even in subangular blocky and platy structured soils.

Anisotropy of saturated and unsaturated hydraulic conductivity and permeability

The possibility to predict the soil hydraulic properties on various scales from small aggregates of few μm with an intra-aggregate pore system up to the hillslope scale of square kilometers (Dusek and Vogel, 2018) depends on the structure of the pore system. Of major importance is the presence and distribution of macropores (MaP) with $d > 0.3 \text{ mm}$, structural pores or inter-aggregate pores

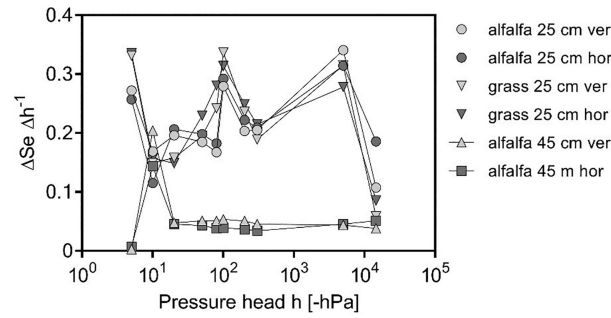


Fig. 4 Pore size distribution as function of the slope of the water retention curve, $\Delta Se \Delta h^{-1}$, of a Haplic Luvisol profile (Ap/Ah, E and E/Bt, Btg, CBkg, and C horizons) in Fig. 1 with an alfalfa and grass plot in 25 cm and 45 cm depth in vertical (ver) and horizontal (hor) directions (Beck-Broichsitter et al., 2020c). The h-value of -10 hPa (about 9.81 cm) was assumed as boundary between the macropores MaP ($h_1 \geq -10$ hPa), and wide coarse pores wCP ($-10 > h_2 \geq -60$ hPa).

including wide coarse pores (wCP) with $0.3 \geq d > 0.05$ mm, and textural pores known as matrix, intra-aggregate, and inter-particle pores including narrow coarse pores (nCP) with $0.05 \geq d > 0.01$ mm, medium pores (MP) with $0.01 \geq d > 0.0002$ mm, and fine pores (FP) with $d \leq 0.0002$ mm (Horn and Smucker, 2005).

The pore size distributions in Fig. 4 are characterized by direction-dependent differences in the structural peaks between water potentials of -10 hPa and -60 hPa, and the matrix peaks at around -100 hPa and at around -5000 hPa. Pedogenesis of a soil profile results in greater vertical than horizontal cracks, producing anisotropy in pore size distribution between vertical and horizontal directions (Fig. 4).

An anisotropy ratio (AR) can be defined as the ratio of Ks-values obtained in vertical (ver) and horizontal (hor) directions (Beck-Broichsitter et al., 2020a, b) as,

$$AR_{Ks} = \frac{K_{s_{hor}}}{K_{s_{ver}}} \quad (6)$$

Eq. 6 describes not only these flux differences in vertical and horizontal direction, but it is also important for determining the unsaturated hydraulic conductivity (Hartge and Horn, 2016).

In Fig. 5, the saturated hydraulic conductivity, Ks, shows structure-dependent differences but the AR_{Ks} values also strongly depend on land use and the clay content. For grassland sites, AR_{Ks} values >1 are most obvious for the topsoil, while the subsoil horizons of arable sites show comparatively more pronounced AR_{Ks} values, also depending on the clay content (Horn et al., 2019).

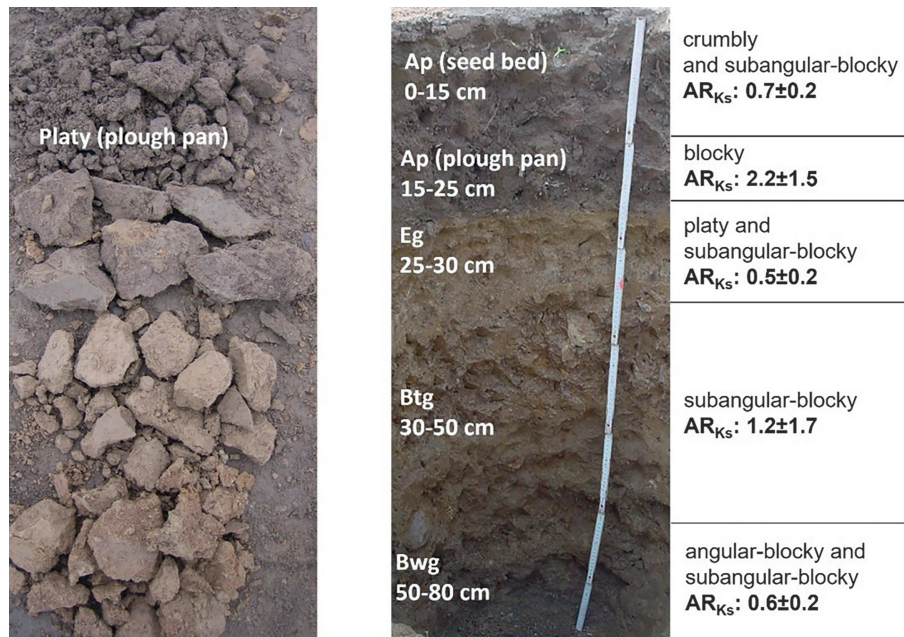


Fig. 5 Conventional tilled Stagnic Luvisol derived from glacial Weichselian till in Northern Germany with different structure types. Anisotropy ratios (arithmetic mean $\pm$ standard deviation) of the saturated hydraulic conductivity, AR_{Ks} , in five different soil horizons (Horn et al., 2019).

Platy structures show mostly dominating horizontal fluxes caused by trampling, grass harvesting, and slurry application with heavy machinery through stress induced particle rearrangement.

In summary, crumbly formed or subangular blocky structured soil horizons mostly indicate isotropic fluxes (Hartge and Horn, 2016), while prismatic structured soils exhibit a predominantly vertical flow caused by the development of secondary pores (shrinkage cracks) or biopores (Uteau et al., 2013) in contrast to a platy structure with predominantly horizontal flow.

In Fig. 6, anisotropy of Ks values, especially in the subsoils under arable land use for a range of soils can be explained by the development of a platy soil structure because of the management (tillage). Sandy soils (Podzols) show only minor anisotropy effects. Groundwater-affected soils have greater Ks in the subsoil due to their prismatic structure.

The unsaturated hydraulic conductivity, $K(Se)$ cm d^{-1} , can be determined as a function of the effective saturation, Se (-), with the combined Mualem-van Genuchten type analytical function and the bimodal retention function (Priesack and Durner, 2006),

$$K(Se) = K_s \left(\sum_{i=1}^k w_i Se_i \right)^\tau \left(\frac{\sum_{i=1}^k w_i \alpha_i [1 - (1 - Se_i^{1/m_i})^{m_i}]}{\sum_{i=1}^k w_i \alpha_i} \right)^2 \quad (7)$$

where Se is the effective saturation (-) with $(\theta - \theta_r)/(\theta_s - \theta_r)$, θ_s is saturated and θ_r is residual water content, respectively, α (cm^{-1}), n (-) and m (-), with $m = 1 - 1/n$, are empirical van Genuchten fitting parameters for both pore domains (index i), integer k denotes the

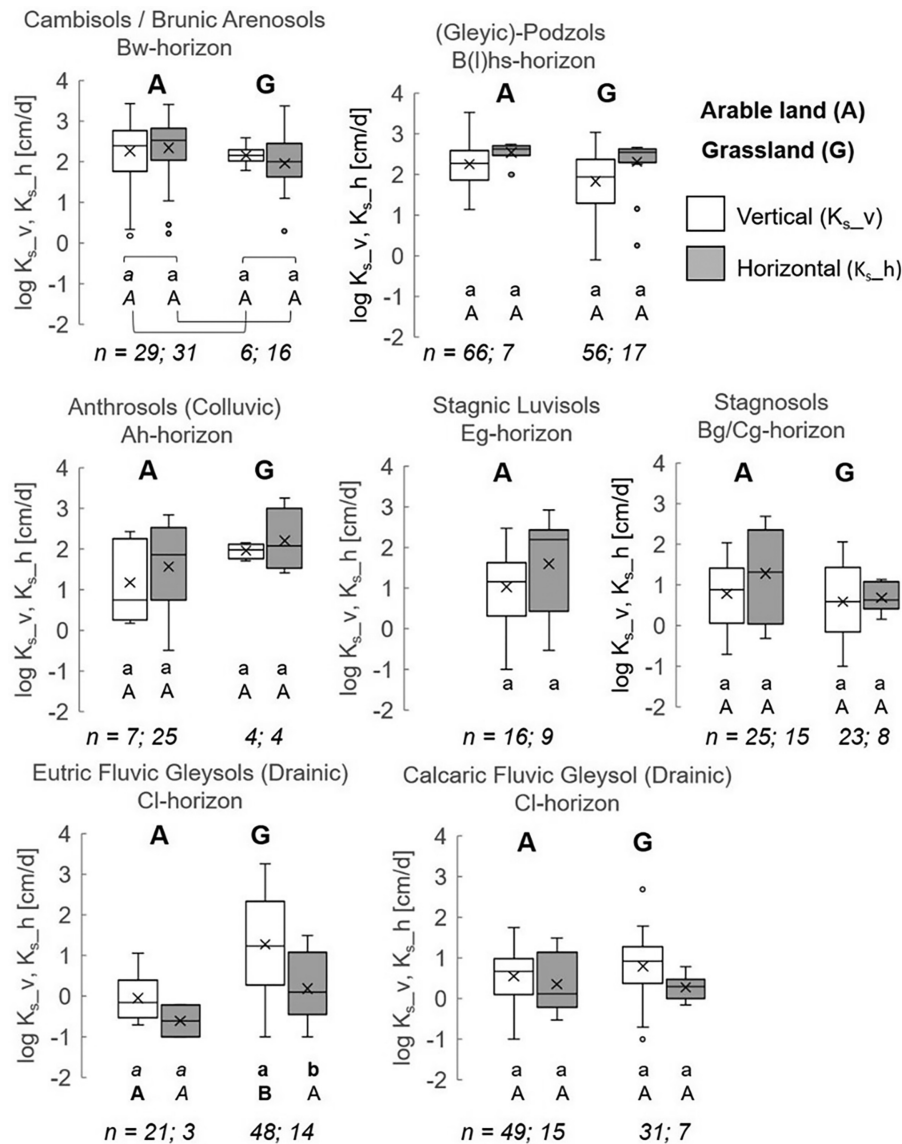


Fig. 6 Saturated hydraulic conductivity in vertical, K_{s_v} , and horizontal, K_{s_h} , direction of subsoil horizons (≤ 60 -cm depth) from representative soil types in Northern Germany under arable use (A) and grassland use (G); number of samples. Significant differences between K_{s_v} and K_{s_h} values are denoted as lowercase letters (ab), those between land use types by uppercase letters (AB), with a significance level of 0.05 (Horn et al., 2019).

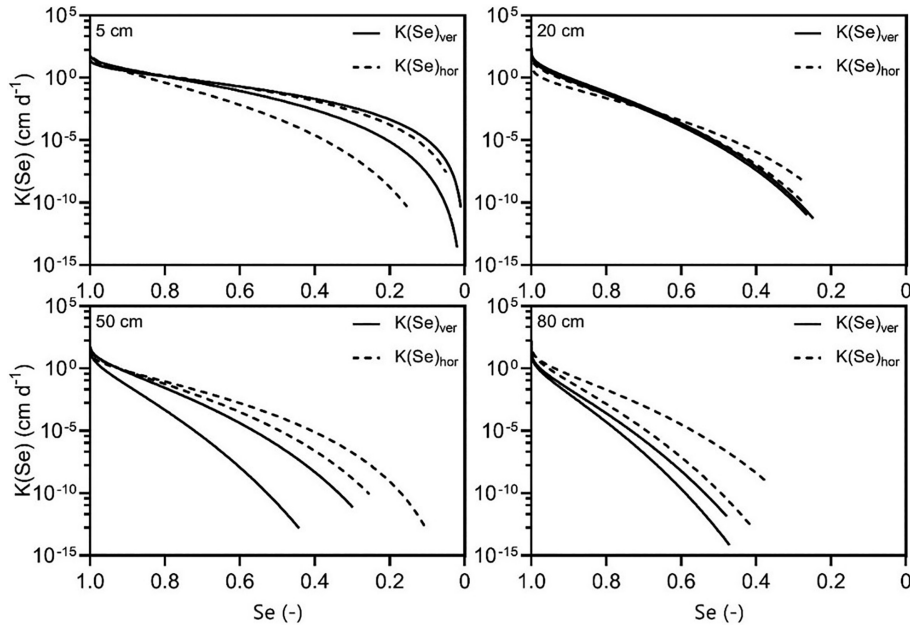


Fig. 7 Unsaturated hydraulic conductivity, $K(Se)$ (cm d^{-1}), of a loamy sand with ρ_b between 1.15 g cm^{-3} and 1.25 g cm^{-3} at 5 cm depth, and ρ_b between 1.75 g cm^{-3} and 1.85 g cm^{-3} at 20 cm, 50 cm, and 80 cm depths, samples taken in vertical (ver) and horizontal (hor) directions (Beck-Broichsitter et al., 2020b).

modality of the model (i.e., $k = 2$ for bimodal), w_i is a weighting factor for the sub-curves of each pore domain ($0 < w_i < 1$ and $\sum w_i = 1$), and $\tau(-)$ is the tortuosity.

In Fig. 7, the results indicate direction-dependent differences in the $K(Se)$ values with depth, including the crumbly-structured soil at 5 cm depth and in the artificially compacted and therefore more platy structured soil at 20 cm, 50 cm, and 80 cm depths (Beck-Broichsitter et al., 2020c). The $K(Se)$ values were mostly related to the K_s values and predominantly to the existence of macropores (Beck-Broichsitter et al., 2020b), and the degree of saturation. Thus, the lower the saturation, the smaller is the share and area of conductive contact points per flow cross section resulting in a strongly decreasing unsaturated soil, while limiting the water flow in the soil matrix (Reszkowska et al., 2011).

Values of AR ranging between 4.5 and 1 in the nearly saturated range (Se ranging from 1.0 and 0.8) suggest an increased potential for lateral flow above a less permeable bedrock (Fig. 7). The $K(Se)$ values were also related to the mean K_s values with 554 cm d^{-1} in horizontal and 123 cm d^{-1} in vertical directions, and macropore volumes of 7.4% in vertical and 13.1% in horizontal directions. The predominantly horizontal-oriented pore network was created by plant roots and root fragments (Dohnal et al., 2012), and differently sized stones and rock fragments (Fig. 8).

The soil water diffusivity, $D(\theta)$ ($\text{cm}^2 \text{ d}^{-1}$), is the product of $K(\theta)$ and the differential water capacity, $dh/d\theta$ (i.e., slope of the water retention function), and the effective $D(\theta)$ values for the i -th pore size classes and can be calculated as,

$$D_i(\theta_i) = \int_{\theta_i^-}^{\theta_i^+} K(\theta_i) \frac{dh}{d\theta_i} d\theta \quad (8)$$

where $\theta^+(h^+)$ is the upper and $\theta^-(h^-)$ is the lower ranging between the i -th pore size classes: MaP ($h_1 \geq -10 \text{ hPa}$), wCP ($-10 > h_2 \geq -60 \text{ hPa}$), nCP ($-60 > h_3 \geq -300 \text{ hPa}$), MP ($-300 > h_4 \geq -15,000 \text{ hPa}$), and FP ($h_5 < -15,000 \text{ hPa}$).

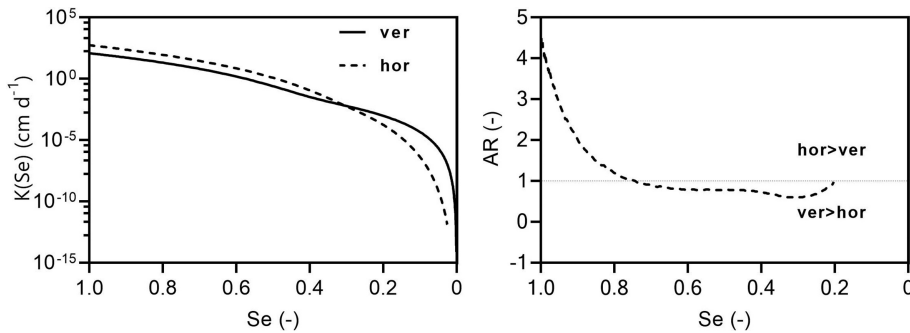


Fig. 8 Unsaturated hydraulic conductivity, $K(Se)$ (cm d^{-1}), (left) and anisotropy ratio, AR (-), (right) of a sandy loamy BwC horizon of a Dystric Cambisol (AE/Bs/Bsw/BwC horizons) under forest use in 40 cm depth with ρ_b of 1.25 g cm^{-3} , samples taken in vertical (ver) and horizontal (hor) direction.

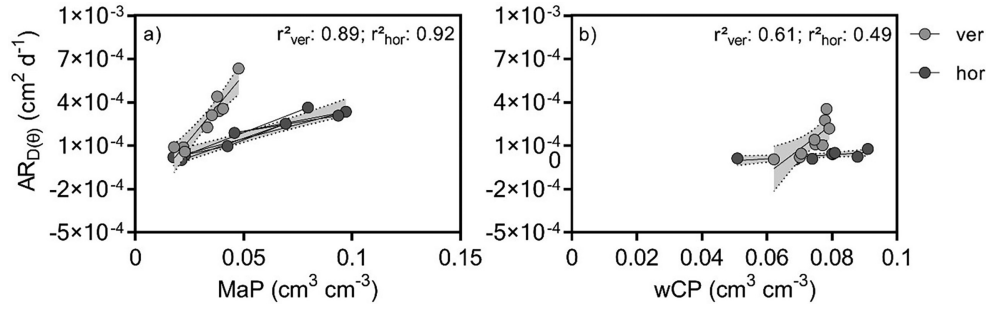


Fig. 9 D(θ)-weighted average anisotropy ratio, $AR_D(\theta)$, as a function of the volume fraction of macropores, MaP ($h_1 \geq -10$ hPa), and wide coarse pores, wCP ($-10 > h_2 \geq -60$ hPa), of a sandy loamy BwC horizon of a Dystric Cambisol (AE/Bs/Bsw/BwC horizons) under forest use in 40 cm depth with ρ_b of 1.25 g cm^{-3} , samples taken in vertical (ver) and horizontal (hor). The dashed lines indicate the confidence intervals of 95% of the relations between $AR_D(\theta)$ and the volume fraction of MaP and wCP, respectively.

A diffusivity-dependent effective weighted anisotropy ratio, $AR_{D(\theta)}$, per pore size class, i , can be calculated as,

$$AR_{D(\theta)} = D_i(\theta_i) \int_{\theta^+}^{\theta^-} \frac{1}{AR_i} d\theta \quad (9)$$

In Fig. 9, the D(θ)-weighted anisotropy ratios, $AR_D(\theta)$, indicate a positive linear relationship for the MaP (r^2 : 0.89–0.92) and wCP (r^2 : 0.49–0.61). The $AR_{D(\theta)}$ values strongly increase with the volume fraction of macropores and wide coarse pores, respectively (Beck-Broichsitter et al., 2020b, c).

The intensity-capacity concept, proposed by Horn et al. (2014), explores whether it is possible to predict intensity values (i.e., hydraulic conductivity, air permeability) with capacity parameters (i.e., bulk density, pore size distribution).

An intensity parameter like soil water diffusivity, $D(\theta)$, can be derived from the $K(\theta)$ function but it is less related to capacity parameters such as porosity and pore size distribution. This results because intensity parameters are also functions of pore continuity and connectivity and not directly of capacities derived from overall volumes (Horn et al., 2014).

Anisotropy of heterogenous structured materials analyzed by X-ray CT image analysis

Gas transport parameters such as air permeability and gas diffusivity can be effectively combined with porosity measured by X-ray computed tomography (CT) for multi-scale analyses (Piccoli et al., 2019). There is a direct link between soil physical properties and structure as obtained by imaging that can be used as the realistic boundary conditions for simulating processes and properties (Schlüter et al., 2018). Conventional soil mechanical methods can also be combined with X-ray CT to analyze the direction-dependent functional recovery of pores in the compacted subsoil (Beck-Broichsitter et al., 2020c).

Tensorial structural characteristics are needed to determine the degree of anisotropy and the preferred orientation (Doubé et al., 2010). The mean intercept length (MIL) tensor is the most used technique (Klatt et al., 2017) and the degree of the CT-based anisotropy (A_{CT}) can be calculated as,

$$A_{CT} = 1 - \frac{\text{Long axis eigenvalue}}{\text{Short axis eigenvalue}} \quad (10)$$

In Fig. 10, the sandy loamy BwC horizon of a Dystric Cambisol (AE/Bs/Bsw/BwC horizons) under forest use and formed by plant roots and rock fragments indicates a CT-based macroporosity (MaPCT) of $0.028 \text{ cm}^3 \text{ cm}^{-3}$ for vertical and $0.047 \text{ cm}^3 \text{ cm}^{-3}$ for horizontal oriented soil cores. The A_{CT} values are larger for the horizontal than for the vertical sampled soil core, while the predominant orientation of macropores is x-axis for the vertical and z-axis for the horizontal oriented soil core and therefore parallel to the sampling direction (90°).

(An-)Isotropy of mechanical properties and functions in structured soils

Stress strain processes in structured unsaturated soils

Mechanical stresses applied to a soil surface result in a 3-dimensional stress propagation and a coinciding induced strain. The extent to which these stresses are transmitted in soils depends on the internal soil strength and become more pronounced as soils become less aggregated, drier due to more negative matric potential, or abundant in organic carbon content. Furthermore, stress transmission depends on texture, clay mineralogy, and composition of organic acids.

Generally, mechanical processes across 3 dimensions in soils are described by the stress-strain equation (i.e., Hartge and Horn, 2016) as,

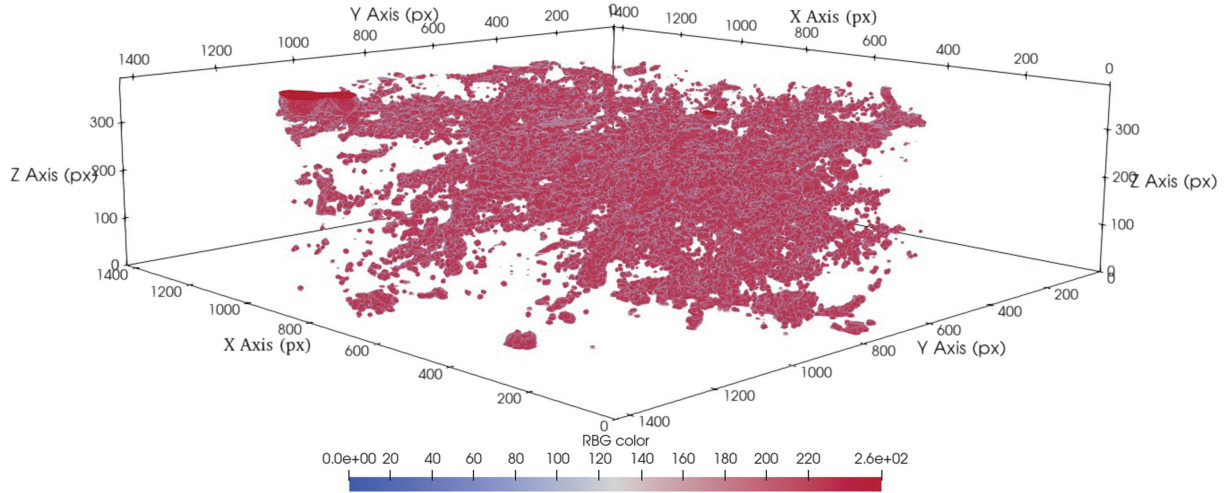


Fig. 10 X-ray CT image of the macropore volume ($\text{MaP}_{\text{CT}}: d > 1.134 \text{ pixel}$) of a sampled soil core using Fiji/ImageJ (Koestel, 2018) as ratio between phase volume (voxel) and bulk volume (voxel) of the region of interest, ROI.

$$\begin{bmatrix} \sigma_x(\vec{x}, t) & \tau_{xy}(\vec{x}, t) & \tau_{xz}(\vec{x}, t) \\ \tau_{xy}(\vec{x}, t) & \sigma_y(\vec{x}, t) & \tau_{yz}(\vec{x}, t) \\ \tau_{xz}(\vec{x}, t) & \tau_{yz}(\vec{x}, t) & \sigma_z(\vec{x}, t) \end{bmatrix} = f \begin{bmatrix} \varepsilon_x(\vec{x}, t) & \varepsilon_{xy}(\vec{x}, t) & \varepsilon_{xz}(\vec{x}, t) \\ \varepsilon_{xy}(\vec{x}, t) & \varepsilon_y(\vec{x}, t) & \varepsilon_{yz}(\vec{x}, t) \\ \varepsilon_{xz}(\vec{x}, t) & \varepsilon_{yz}(\vec{x}, t) & \varepsilon_z(\vec{x}, t) \end{bmatrix} \quad (11)$$

where f defines the soil specific material properties. In the case of a linear-elastic isotropic material, the tensor notation reads as,

$$\begin{bmatrix} \sigma_x & \tau_{xy} & \tau_{xz} \\ \tau_{xy} & \sigma_y & \tau_{yz} \\ \tau_{xz} & \tau_{yz} & \sigma_z \end{bmatrix} = \frac{E}{1 + \nu} \begin{bmatrix} \varepsilon_x - \varepsilon_m & \varepsilon_{xy} & \varepsilon_{xz} \\ \varepsilon_{xy} & \varepsilon_y - \varepsilon_m & \varepsilon_{yz} \\ \varepsilon_{xz} & \varepsilon_{yz} & \varepsilon_z - \varepsilon_m \end{bmatrix} + \frac{E}{3 \cdot (1 - 2 \cdot \nu)} \cdot \begin{bmatrix} \varepsilon_m & 0 & 0 \\ 0 & \varepsilon_m & 0 \\ 0 & 0 & \varepsilon_m \end{bmatrix} \quad (12)$$

where E is $\Delta\sigma/\Delta\varepsilon$ (Young's modulus or modulus of elasticity) and ν is Poisson's ratio.

The first tensor, ε_m , describes the volume constant shear deformation and the second tensor the volumetric strain with the mean volume strain ε_m defined as (Hartge and Horn, 2016),

$$\varepsilon_m = \frac{1}{3} \cdot (\varepsilon_x + \varepsilon_y + \varepsilon_z) = \frac{1}{3} \cdot \varepsilon_{\text{vol}} \quad (13)$$

For example, Riggert et al. (2016) documented the variation in stress transmission in structured unsaturated soils under conventional, conservation, and forestry management by determination of the stress state components as well as the tensorial varying volumetric displacements as a function of the applied stresses. The proportions of the major and minor stresses as well as the mean normal stress (MNS) or octahedral shear stress (OCTSS) depend on soil aggregation like coherent, prismatic, polyhedral, subangular blocky, crumbly, or platy structures.

The internal soil strength (precompression stress or soil rigidity) quantifies the consequences of mechanical, hydraulic, chemical, or biological stresses on the tensorial components of the stress strain ratio (Horn and Blum, 2020). While stress transmission in incompressible water is completely isotropic, soils behavior ranges from mostly vertically anisotropic if they are very weak (i.e., moist, with mostly coherent structure and fine textured) to an increased proportion of horizontal deformation in stronger aggregated soils with subangular blocky, crumbly structure or soils with platy structure. In this case the total affected stressed soil volume is reduced. The vertical anisotropy fades off and finally reaches isotropy. Coinciding with this, the deviator stress at rest k ($\sigma_1^* k = \sigma_3$) varies between 0.2 and > 1 depending on the internal soil strength.

Fröhlich (1934) defined these strength dependent changes with the concentration factor, which for substrates ranged between 3 for very hard (isotropic deformation behavior) up to 9 for weak and/or wet material conditions. However, because of limited soil rigidity, a precompression stress dependency needs to be incorporated which affects the volumetric stress strain, too (Horn and Blum, 2020). Exceeding the internal soil strength causes an increase in the concentration factor, a more intense vertical soil deformation and a coinciding increased vertical anisotropy.

Quantification of structure dependent changes by tensorial stress strain and shear stress components

Based on the vectorial composition of the strength properties and functions, the direction dependent changes in the shear resistance parameters (angle of internal friction, ϕ , and cohesion, c) and precompression stress (P_c), can also have isotropy or anisotropy patterns. For instance, stress induced breakage of a platy structure is highly direction dependent, with the intensity of the stress application beyond the rigidity limits differing in vertical and horizontal directions (Fazekas and Horn, 2005). From shear

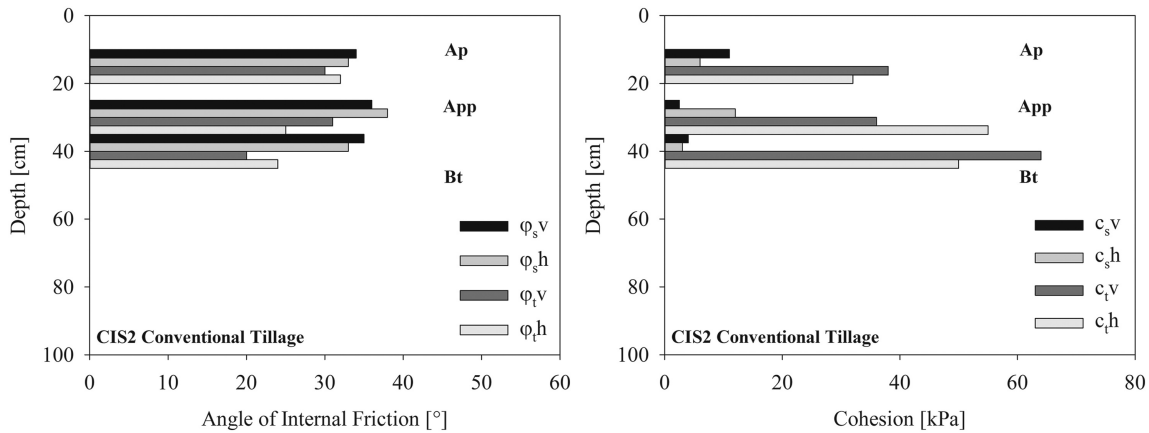


Fig. 11 Angle of internal friction, ϕ , and cohesion, c , for a Haplic Luvisol derived from glacial Weichselian till (site CIS2) as a function of stress applied and sampling direction. The matric potential was -60 hPa, ϕ_{sv} and ϕ_{sh} denote: structure-related angle of the friction of the samples taken vertically (v) and horizontally (h), while c_{sv} and c_{sh} define the corresponding cohesion values. The factors s and t define the stress ranges applied: s is stresses smaller than the precompression stress, t is stress within the virgin compression stress (Dörner and Horn, 2009).

measurements and stress strain analyses the directional dependence can be quantified, with the applied stresses possibly coinciding with a new anisotropic platy structure formation in deeper soil depths.

Furthermore, P_c values strongly depend on soil aggregation and show very distinct differences for soil samples collected in vertical vs. horizontal directions. For a platy structure, the ratios of horizontal to vertical P_c values (AR_{P_c}) are greater than 1 due to their formation by the complex stresses of wheeling. The latter was observed in a platy structure plow pan in a loess derived Luvisol down to 60 cm depth (Fazekas and Horn, 2005). On the hand, less anthropogenically affected soils with a crumblier structure show AR_{P_c} values <1 . Likewise, the ratio of the horizontal to vertical shear parameters for c (A_c) and ϕ (A_ϕ) follow similar trends in disturbed compared with well-managed soils, being >1 for a platy structure but changing to <1 for sites under conservation tillage that improved soil structure (i.e., Fazekas and Horn, 2005). These changes underline the tensorial effects and reactions on soil strength due to its limited rigidity, which can be derived from the stress dependent changes in these ratios. For a coherent structure, ϕ and c values are mostly identical, while a prismatic structure indicates slightly greater ϕ and c values in vertical than horizontal direction, while platy structured soils show an opposite trend (i.e., Dörner and Horn, 2009).

In Fig. 11, the subangular blocky-structured Ap horizon of a Luvisol, derived from glacial till, shows values for the shear resistance lower than the P_c , while the sampling direction indicates no differences in ϕ values, but greater c for samples taken vertically (Dörner and Horn, 2009).

Consequences of the strength anisotropy for land sliding processes

In Fig. 12, the direction dependent shear parameters in aggregated soils also modify the bearing capacity induced volumetric strain behavior as it depends directly on the ϕ and the c (Hartge and Horn, 2016). The prediction of the tensile failure pattern (extension, Rankine's active pressure distribution) requires the application of those values determined for vertically sampled soils, but the

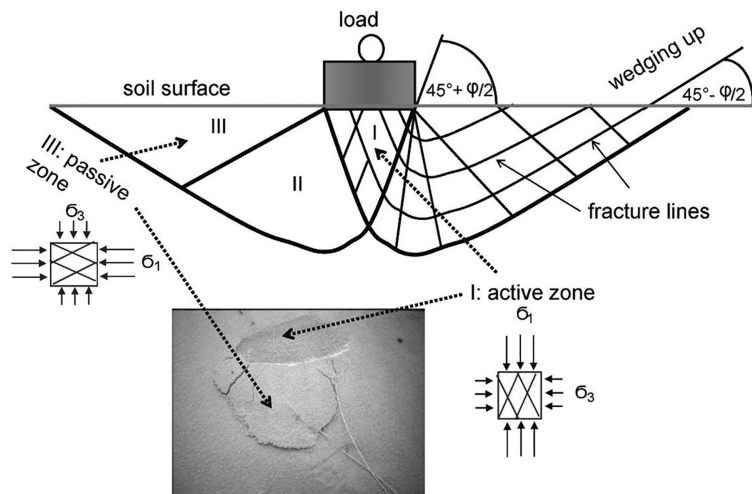


Fig. 12 Schematic description of the bearing capacity induced volumetric strain behavior (Hartge and Horn, 2016).

opposite is true for the Rankine's passive pressure distribution. Horizontally sampled soil parameters are required to define the deformation and failure patterns as the exit angle, which reads as $45^\circ - \varphi/2$. Well-structured soils have a flatter exit angle of the laterally moved soil volume (passive movement) compared with weaker aggregated soils (Hartge and Horn, 2016).

Conclusion

Soil structure not only has a significant influence on soil stability (P_c) and the shear parameters φ and c but also on the vertical or horizontal flow of water. Soil deformation ($P_c < \text{stress}$) due to mechanical loads can lead to a distinct horizontal anisotropy of the soil pore system because of anthropogenic induced plate structure formation. This promotes surface runoff and thus increased soil erosion and may reduce biomass productivity. Shear parameters are thus tensors which depend on soil aggregation and other internal properties.

Direction-dependent differences in the soil water diffusivity as a link between intensity and capacity parameters are strongly related to the pore size distribution. The greatest impacts are from macropores to narrow coarse pores. Thus, by weighting the anisotropy ratio, $AR_{D(\theta)}$, with the soil water diffusivity, both the capacity and the intensity parameters were combined. Small-scale direction-dependent differences in the soil hydraulic capacity and intensity parameter will be used for model-based upscaling for better understanding of water flow at the catchment scale.

However, additional research is needed to extend these results to further soil profiles information under various tillage systems and soil properties, and to finally document these interactions more completely. In conclusion, mechanical and hydrological soil functions are tensors that depend on the soil structure and the stresses applied in comparison with soil mechanical stability.

References

- Ball BC, O'Sullivan MF, and Hunter R (1988) Gas diffusion, fluid flow and derived pore continuity indices in relation to vehicle traffic and tillage. *Journal of Soil Science* 39: 327–339. <https://doi.org/10.1111/J.1365-2389.1988.TB01219.X>.
- Ball BC, Smith KA, and Mullins CE (2001) Gas movement and air-filled porosity. In: Smith KA and Mullins CE (eds.) *Soil and Environmental Analysis: Physical Methods, Revised and Expanded*, pp. 499–539. New York, NY, USA: Marcel Dekker.
- Beck-Broichsitter S, Fleige H, Gerke HH, and Horn R (2020a) Effect of artificial soil compaction in landfill capping systems on anisotropy of air-permeability. *Journal of Plant Nutrition and Soil Science* 1–11. <https://doi.org/10.1002/jpln.201900281>.
- Beck-Broichsitter S, Fleige H, Dusek J, and Gerke HH (2020b) Anisotropy of unsaturated hydraulic properties of compacted mineral capping systems seven years after construction. *Soil and Tillage Research* 204. <https://doi.org/10.1016/j.still.2020.104702>.
- Beck-Broichsitter S, Gerke HH, Leue M, von Jeetze PJ, and Horn R (2020c) Anisotropy of unsaturated soil hydraulic properties of eroded Luvisol after conversion to hayfield comparing alfalfa and grass plots. *Soil and Tillage Research* 198: 104553. <https://doi.org/10.1016/j.still.2019.104553>.
- Chief K, Ty P, Ferré A, and Hinnell AC (2008) The effects of anisotropy on in situ air permeability measurements. *Vadose Zone Journal* 7: 941–947. <https://doi.org/10.2136/vzj2007.0164>.
- Dohnal M, Vogel T, Sanda M, and Jelinkova V (2012) Uncertainty analysis of a dual-continuum model used to simulate subsurface hillslope runoff involving oxygen-18 as natural tracer. *Journal of Hydrology and Hydromechanics* 60(3): 194–205. <https://doi.org/10.2478/v10098-012-0017-0>.
- Dörner J and Horn R (2006) Anisotropy of pore functions in structured Stagnic Luvisols in the Weichselian moraine region in N Germany. *Journal of Plant Nutrition and Soil Science* 169: 213–220. <https://doi.org/10.1002/jpln.200521844>.
- Dörner J and Horn R (2009) Direction-dependent behaviour of hydraulic and mechanical properties in structured soils under conventional and conservation tillage. *Soil and Tillage Research* 102: 225–232. <https://doi.org/10.1016/j.still.2008.07.004>.
- Doobe M, Klosowski MM, Arganda-Carreras I, Cordelieres FP, Dougherty RP, Jackson JS, Schmid B, Hutchinson JR, and Shefelbine SJ (2010) BoneJ: free and extensible bone image analysis in ImageJ. *Bone* 47: 1076–1079. <https://doi.org/10.1016/j.bone.2010.08.023>.
- Dusek J and Vogel T (2018) Hillslope hydrograph separation: The effects of variable isotopic signatures and hydrodynamic mixing in macroporous soil. *Journal of Hydrology* 563: 446–459. <https://doi.org/10.1016/j.jhydrol.2018.05.054>.
- Fazekas O and Horn R (2005) Interaction between mechanically and hydraulically affected soil strength depending on time of loading. *Journal of Plant Nutrition and Soil Science* 168(1): 60–67. <https://doi.org/10.1002/jpln.200421381>.
- Fröhlich OK (1934) *Druckverteilung im Baugrund*. Wien, NY: Springer.
- Hartge KH and Horn R (2016) In: Horton R, Horn R, Bachmann J, and Peth S (eds.) *Essential Soil Physics, An Introduction to Soil Processes, Structure, and Mechanics*, p. 391. Stuttgart, Germany: Schweizerbart Science Publishers.
- Horn R and Blum W (2020) Effect of land-use management systems on coupled physical and mechanical, chemical, and biological soil processes: how can we maintain and predict soil properties and functions? *Frontiers of Agricultural Science and Engineering* 7(3), 104709. <https://doi.org/10.15302/J-FASE-2020334>.
- Horn R, Peng X, Fleige H, and Dörner J (2014) Pore rigidity in structured soils – only a theoretical boundary condition for hydraulic properties? *Soil Science and Plant Nutrition* 60: 3–14. <https://doi.org/10.1080/00380768.2014.886159>.
- Horn R and Smucker A (2005) Structure formation and its consequences for gas and water transport in unsaturated arable and forest soils. *Soil and Tillage Research* 82: 5–14. <https://doi.org/10.1016/j.still.2005.01.002>.
- Horn R, Mordhorst A, Fleige H, Zimmermann I, Burbaum B, Filipinski M, and Cordsen E (2019) Soil Type and land use effects on tensorial properties of saturated hydraulic conductivity in Northern Germany. *European Journal of Soil Science* 71(2): 179–189. <https://doi.org/10.1111/ejss.12864>.
- Jarvis NJ (2007) A review of non-equilibrium water flow and solute transport in soil macropores: Principles, controlling factors and consequences for water quality. *European Journal of Soil Science* 58: 523–546. <https://doi.org/10.1111/j.1365-2389.2007.00915.x>.
- Klatt MA, Schröder-Turk GE, and Mecke K (2017) Mean-intercept anisotropy analysis of porous media. II. Conceptual shortcomings of the MIL tensor definition and Minkowski tensors as an alternative. *Medical Physics* 44(7): 3663–3675. <https://doi.org/10.1002/mp.12280>.
- Koestel J (2018) SoilJ – An ImageJ plugin for the semi-automatic processing of 3-D X-ray images of soils. *Vadose Zone Journal* 17: 170062. <https://doi.org/10.2136/vzj2017.03.0062>.
- Maier M, Gartiser V, Schengel A, and Lang V (2020) Long term soil gas monitoring as tool to understand soil processes. *Applied Sciences* 10: 8653. <https://doi.org/10.3390/app10238653>.

- Moldrup P, Chamindu Deepagoda TTK, Hamamoto S, Komatsu T, Kawamoto K, Rolston DE, and Wollesen de Jonge L (2013) Structure-dependent water-induced linear reduction model for predicting gas diffusivity and tortuosity in repacked and intact soil. *Vadose Zone Journal* 12: 1–11. <https://doi.org/10.2136/vzj2013.01.0026>.
- Peng X and Horn R (2007) Anisotropic shrinkage and swelling of some organic and inorganic soils. *European Journal of Soil Science* 58: 98–107. <https://doi.org/10.1111/j.1365-2389.2006.00808.x>.
- Piccoli I, Schjønning P, Lemandé M, Zanini F, and Morari F (2019) Coupling gas transport measurements and X-ray tomography scans for multiscale analysis in silty soils. *Geoderma* 338: 576–584. <https://doi.org/10.1016/j.geoderma.2018.09.029>.
- Priesack E and Durner W (2006) Closed-form expression for the multi-modal unsaturated hydraulic conductivity. *Vadose Zone Journal* 5: 121–124. <https://doi.org/10.2136/vzj2005.0066>.
- Reszkowska A, Krümmelbein J, Peth S, Horn R, Zhao Y, and Gan L (2011) Influence of grazing on hydraulic and mechanical properties of semiarid steppe soils under different vegetation type in Inner Mongolia, China. *Plant and Soil* 340: 59–72. <https://doi.org/10.1007/s11104-010-0405-3>.
- Riggert R, Fleige H, Kietz B, Gaertig T, and Horn R (2016) Stress distribution under forestry machinery and consequences for soil stability. *Soil Science Society of America Journal* 80(1): 38–47. <https://doi.org/10.2136/sssaj2015.03.0126>.
- Schjønning P and Koppelgaard M (2017) The Forchheimer approach for soil air permeability measurement. *Soil Science Society of America Journal* 81: 1045–1053. <https://doi.org/10.2136/sssaj2017.02.0056>.
- Schlüter S, Großmann C, Diel J, Wu G-M, Tischer S, Deubel A, and Rücknagel J (2018) Long-term effects of conventional and reduced tillage on soil structure, soil ecological and soil hydraulic properties. *Geoderma* 332: 10–19. <https://doi.org/10.1016/j.geoderma.2018.07.001>.
- Uteau D, Pagenkemper SK, Peth S, and Horn R (2013) Root and time dependent soil structure formation and its influence on gas transport in the subsoil. *Soil and Tillage Research* 132: 69–76. <https://doi.org/10.1016/j.still.2013.05.001>.
- Villagra K, Carvajal-Vanegas D, Beck-Broichsitter S, and Horn R (2020) A simulated effect of wetting and drying periods for two textural soils amended with biochar at a catchment scale. *Agricultural Engineering International: The CIGR e-journal*. 22: 9–21.

Macropores and macropore flow

Nicholas Jarvis and Mats Larsbo, Department of Soil and Environment, Swedish University of Agricultural Sciences, Uppsala, Sweden

© 2023 Elsevier Ltd. All rights reserved.

Introduction	306
What are macropores?	307
What are macropore and preferential flows?	307
Why is macropore flow important?	308
Macropore network characteristics and macropore flow	308
Modeling macropore flow and transport	311
Conclusions and outlook	312
References	313

Abstract

Macropores are large pores in soil created by both physical processes (e.g., swell-shrink) and biological agents (e.g., plant roots and soil fauna). Macropores have profound impacts on various critical soil ecosystem services/disservices (e.g., flood prevention, contaminant leaching). In this article, we first discuss the nature of macropores and explain why they are important. We then discuss how the “architecture” of macropore networks affects water flow, before outlining some approaches that are commonly used to model water flow through soils containing macropores. We conclude with some suggestions for future research that may prove fruitful.

Glossary

Macropores Large diameter structural pores in soil created either by physical (i.e., swell/shrink) or biological (root turnover, faunal activity) processes.

Macropore flow The preferential channeling of water flowing through soils in macropores due to an extreme contrast in permeability with surrounding matrix pores.

Non-equilibrium flow Refers to the fact that single representative values of pressure potential and/or solute concentration cannot be identified for a soil volume through which water is flowing. This is a diagnostic feature of all preferential flow processes, including macropore flow.

Percolating pore space That part of the pore space which is connected to both the upper and lower surfaces of a considered soil volume.

Percolation threshold A critical (minimum) value of soil porosity needed for the pore space to become long-range connected. Above the threshold, the pore space is said to percolate.

Preferential flow Flowing water is channeled through a limited part of the soil pore space.

Key points

- Macropores are large diameter structural pores in soil created by either physical (e.g., swell-shrink) or biological (e.g., faunal activity) processes.
- Macropores allow fast “preferential” channeling of water through soils due to the extreme contrast in permeability with surrounding matrix pores.
- Macropores have both detrimental environmental impacts (e.g., enhancement of contaminant leaching risks) but also significant beneficial ecosystem services (e.g., reduction of runoff and flooding risks).

Introduction

Soil macropores and the occurrence of macropore flow were recognized as long ago as the late 19th century. From measurements of drainage flow and the leaching of agrochemicals at Rothamsted in the U.K., Lawes et al. (1882) observed that . . . “in a heavy soil, channel drainage will in most cases precede general drainage, a portion of the water escaping by the open channels before the body of the soil has become saturated; this will especially be the case if the rain fell rapidly, and water accumulates on the surface.” They further concluded that “... the drainage water from a soil may thus be of two kinds: it may consist (1) of rain-water which has passed, with but little alteration in composition, down the open channels of the soil; or (2) of the water discharged from the pores of a saturated soil.” However, this accurate, yet qualitative,

description of rapid non-equilibrium water flow through soils containing macropores was largely ignored during the ensuing century. Instead, quantitative approaches introduced by Edgar Buckingham and Lorenzo Richards in the early 1900's laid much of the foundation of soil water physics in the decades that followed. This classical theory of soil water flow is based on an underlying assumption of homogeneity, that unique values of water potential, water content and hydraulic conductivity can adequately characterize a soil volume (i.e., that a so-called representative elementary volume—REV—exists for these properties). Solute transport theory developed along conceptually similar lines, with the advection-dispersion equation (ADE) becoming extensively used from the middle of the 20th century. The significance of macropores was re-discovered in the 1970s and early 1980s, as new experimental observations of their effects on soil water flow and solute leaching began to challenge these established paradigms (Thomas and Phillips, 1979; Beven and Germann, 1982). Research into all aspects of macropore flow has dramatically intensified during the last 40 years (Jarvis et al., 2016; Nimmo, 2020).

What are macropores?

Macropores are comparatively large pores in soil that develop as a consequence of various mechanisms of soil structure formation, including physical processes such as wetting and drying and freeze-thaw cycles, which cause soil cracking, and the actions of biological agents, especially the growth and decay of plant roots and the activity of soil fauna (Meurer et al., 2020). In most soils, macroporosity will tend to decrease with depth in parallel with a decline in the intensity of these structure-forming processes. Below the plant root zone, macroporosity derived from biological activity and soil wetting and drying cycles will usually be minimal, although fissures of geological origin may be present in the underlying soil parent material. Macropores tend to be biological “hot-spots,” especially in undisturbed soils (e.g., untilled soils or subsoils in arable land), supporting a significantly greater biomass, activity and functional diversity of the microbial community. This is because surface-vented macropores efficiently supply nutrients and oxygen to roots and microorganisms and because living plant roots often grow preferentially in or close to macropores, which provides a supply of carbon from root exudation and turnover. Compared with the bulk soil, the interfaces between subsoil macropores and bulk soil are often characterized by coatings and linings that are enriched in either clay or organic carbon, or both. These internal interfaces in soil may be more impermeable and/or hydrophobic than the bulk soil and also more chemically reactive (see references in Jarvis, 2007).

What are macropore and preferential flows?

The presence of macropores can result in extreme contrasts in permeability at the pore-scale, which can lead to large spatial variations in pore water velocities and solute displacement as water infiltrates the soil during precipitation events. Water flow in macropores is one of three processes which can result in so-called preferential flow in soil, such that flowing water is channeled through only a limited fraction of the total pore space. A convergence of flow toward a limited number of flow paths leads to non-equilibrium in pressure potential transverse to the flow direction, which is a diagnostic feature of preferential flow (Beven and Germann, 1982). These non-equilibrium hydraulic conditions can be exemplified by non-monotonic wetting patterns recorded in plot experiments in the field. In these situations, soil water monitoring devices (e.g., tensiometers or time-domain reflectometry probes) located deep in the soil in contact with water-conducting macropores show a faster response to precipitation than probes installed at shallower depths, but not in contact with macropores (see studies cited in Jarvis et al., 2016). It is not straightforward to identify a specific minimum pore diameter that enables non-equilibrium flow conditions to develop, not least because pore continuity (or connectivity) also plays an important role (see “Macropore network characteristics and macropore flow”). Nevertheless, evidence from the use of a variety of experimental methods suggests that pores larger than ca. 0.3–0.6 mm in diameter will generally allow preferential flow (Jarvis, 2007).

Preferential flow can also occur in the form of unstable “finger flow” in homogeneous coarse-textured or water repellent soils. The third preferential flow process, often termed funnel flow (or simply heterogeneous flow) occurs as a consequence of macroscopic spatial variations in hydraulic properties in heterogeneous soils. We will not further discuss heterogeneous flow in this chapter as, in principle, it can be dealt with by extending applications of the Richardson-Richards and ADE equations to two or three spatial dimensions. In contrast, macropore flow is a pore-scale phenomenon (Jarvis, 2007; Jarvis et al., 2016) that invalidates the use of continuum model approaches like the Richardson-Richards' equation for water flow and the ADE for solute transport, as a REV cannot be identified at a given depth in the soil for the key state variables (i.e., water content, matric potential and solute concentration).

It should be noted that the three types of preferential flow may occur simultaneously in some soil types (Jarvis et al., 2016). However, in general, macropore flow is probably the most important of the three processes, especially in finer-textured soils where they provide a sufficient contrast in permeability with the surrounding (bulk or matrix) pore space. Furthermore, macropores are ubiquitous in all soils, they are often continuous along large parts of the soil profile and pore water flow velocities in macropores can be very large. Data collated from the literature suggests that macropore flow velocities in soils are typically between 0.1 and 10 m d⁻¹, with a median value of 3.5 m d⁻¹ (Nimmo, 2007; Nimmo et al., 2021). A word of caution is appropriate here, as many of these experiments were conducted with precipitation intensities and amounts that are uncommon under natural conditions in most

climates (see [Jarvis, 2007](#)). Flow velocities in macropores are sensitive to water input rates at the soil surface, so that values toward the lower end of this reported range would be more likely for rainfall events of more typical intensities.

Soil physics theory suggests that local pressures close to atmospheric pressure are required to generate water flow into macropores (see references in [Jarvis, 2007](#)). These near-saturation conditions will occur more frequently in soil characterized by a limited matrix permeability, for example, in clay soils or soils compacted by vehicle traffic. Water repellency, which often develops during the drying of uncultivated soils under natural vegetation, can also strongly enhance the generation of macropore flow ([Jarvis, 2007](#); [Jarvis et al., 2016](#)) as it reduces the infiltration capacity of the soil matrix and the required water potentials can then be quickly reached, even for rainfall events of low intensity. Near-saturated conditions need only occur locally to generate macropore flow. As noted above, preferential flow is a non-equilibrium process, so that once it has been initiated, water flowing in macropores may rapidly “by-pass” the surrounding dry soil matrix, providing the supply rate to the macropores is larger than the imbibition rate into the matrix. In this respect, the distinct physico-chemical properties of macropore/matrix interfaces may drastically reduce uptake rates of water and solutes into the matrix, which enhances the strength of macropore flow and thus also the risk of leaching of contaminants (see studies cited by [Jarvis, 2007](#)).

Irrespective of size, water flow in all soil pores is governed by the same fundamental momentum balance between driving forces (gravity and matric potential gradients), inertial forces, and the resistances to flow arising from friction between water and the solid surfaces and within the water itself. Although these same forces operate on water in all pores, the balance between them differs for water flowing rapidly in macropores compared with slower flowing water in much smaller matrix pores. Firstly, gravity dominates the driving force for water flow in macropores as significant matric potential gradients are highly unlikely to develop. Furthermore, the pore water velocities in macropores may become so large that the acceleration terms in the momentum balance may not be negligible and the flow is no longer laminar (see studies cited in [Jarvis, 2007](#)). Such flow conditions violate assumptions implicit in the Richardson-Richards equation describing creep flow.

Why is macropore flow important?

Water flow in macropores is associated with potentially significant ecosystem disservices. For example, on strongly sloping land, it is implicated in a loss of slope stability, landslides and gully erosion (e.g., [Hencher, 2010](#); [Wilson et al., 2013](#)). Fast non-equilibrium water flow through macropores can also seriously impact the quality of groundwater by short-circuiting the chemical and biological capacity of the soil in the unsaturated zone to retain and degrade potential contaminants (e.g., [Jarvis, 2007](#); [Clothier et al., 2008](#)). In this respect, macropore flow is an especially critical process for otherwise relatively immobile or quickly degrading solutes that are foreign to the soil and whose effects on aquatic ecosystems or human health are significant at small leached fractions (e.g., pesticides; [Jarvis, 2007](#)). In contrast, the leaching of relatively mobile solutes that are also indigenous to the soil (e.g., nitrate) will be less affected by macropore flow. In a similar way, macropore flow can also impact surface water quality via discharges from field drainage systems. However, in this context, it should be remembered that an absence of macropores would severely impair drainage, leading to surface runoff, which in most cases would result in significantly greater adverse impacts on surface water quality.

Macropores also provide significant beneficial ecosystem services related to soil hydrology. In particular, macropores open to the soil surface enhance infiltration capacity and drainage rates ([Beven and Germann, 1982](#)) thereby minimizing the risks of severe waterlogging, surface runoff and soil erosion. In terms of impacts on the soil water balance, macropores that function as non-equilibrium flow pathways through the unsaturated zone will increase recharge to groundwater, while correspondingly reducing soil water contents and evapotranspiration. However, although they play a critical role in partitioning precipitation at the soil surface between infiltration and surface runoff, the impact of macropores on the re-distribution of infiltrated water in soil and recharge to groundwater is likely to be quite limited in most cases ([Or, 2019](#)), certainly in comparison with their impact on the risk of contaminant leaching ([Jarvis, 1998](#)).

Macropore network characteristics and macropore flow

[Allaire et al. \(2009\)](#) and [Jarvis et al. \(2016\)](#) comprehensively reviewed the wide variety of experimental methods that have been used to study the properties of macropores and macropore flow in both the laboratory and in the field, so only a brief summary is provided here. Common approaches to studying macropore flow in the field include the quantification of dye staining patterns following infiltration as well as the use of routine soil physical techniques to monitor the time-course of responses of either soil water potentials or water contents to precipitation events. In the laboratory, the extent of preferential flow is commonly assessed by analyzing breakthrough curves for non-reactive tracer compounds that represent travel-time distributions under steady water flow. Macropore flow has also been monitored at intermediate scales on large outdoor monoliths or lysimeters under more field-like (i.e., transient) conditions.

The morphological properties of the macropores themselves are clearly also of great interest. Although they are difficult to characterize in the field, some methods have been developed based on soil excavation following the infiltration of hard-setting materials ([Allaire et al., 2009](#)). However, these methods are not widely employed as they are laborious and because they only reveal very large macropores. In the laboratory, early studies used staining approaches in combination with thin-sectioning techniques to

quantify the properties of soil macropores. In more recent years, this approach has been largely supplanted by X-ray scanning technologies as these enable the quantification of a range of geometrical and topological properties of soil macropores in 3D and for larger soil samples (see Fig. 1). The use of X-ray scanning has, for example, revealed that soil macroporosity is generally comprised of a number of individual networks or “clusters” separated from one another by smaller matrix pores, with each network being characterized by multiple pathways of variable thickness, including bottleneck constrictions and “dead-ends.”

Recent studies making use of X-ray tomography have confirmed expectations that some of the key concepts in a mathematical theory of connectivity termed “percolation theory” should be applicable to soil macropore networks (Jarvis et al., 2017a; Lucas et al., 2020; Koestel et al., 2020). In particular, these studies show that soils appear to follow one of the underlying premises of percolation theory, in that the long-range connectivity of macroporosity is achieved at a critical value termed the percolation threshold. For soils with macroporositities above this threshold value, individual clusters merge into one “giant” long-range connected cluster. The dependence of pore connectivity on porosity means that more sparsely distributed larger diameter macropores will generally be less well connected than smaller diameter macropores or mesopores that are usually characterized by larger pore volumes. In this respect, Koestel et al. (2020) showed that dense networks of pores between 80 and 250 μm in diameter were especially important in linking together much larger macropores that would otherwise be disconnected.

Experimental studies that combined imaging methods to quantify the properties of macropore networks with experiments to quantify water flow and solute transport on the same samples have led to new insights into how the properties of macropore networks influence macropore flow. For example, Larsbo et al. (2014) related macropore network characteristics quantified by X-ray to the strength of macropore flow determined from steady-state tracer breakthrough experiments under near-saturated flow conditions for 18 topsoil columns sampled from four arable sites, all but two of which had percolating macroporosity. They found that for soils with small macroporosity, the macropore saturation and the diameter of the largest water-filled pore was larger, which led to stronger preferential flow (Fig. 2). Similar experimental results were reported by Katuwal et al. (2015) and by Larsbo et al. (2016). At their field site, Larsbo et al. (2016) reported that preferential flow was significantly weaker for soils rich in organic carbon that were characterized by dense well-connected networks of pores 200–600 μm in diameter.

These findings can be explained by a conceptual model based on percolation concepts (Larsbo et al., 2014), such that the degree of preferential flow will increase as macroporosity decreases, providing that macroporositities are larger than the percolation threshold where long-range connectivity breaks down (see schematic illustration in Fig. 3). Below this threshold, the macropore clusters are disconnected and many will not be open to the water supply at the soil surface. Water flowing into disconnected macropore clusters will accumulate at dead-ends or “bottlenecks,” resulting in saturation and the re-distribution of water to the surrounding matrix pores, a process termed internal catchment. The development of local near-saturated conditions in the soil matrix adjacent to another macropore cluster would then be required to trigger further macropore flow. Thus, both experimental

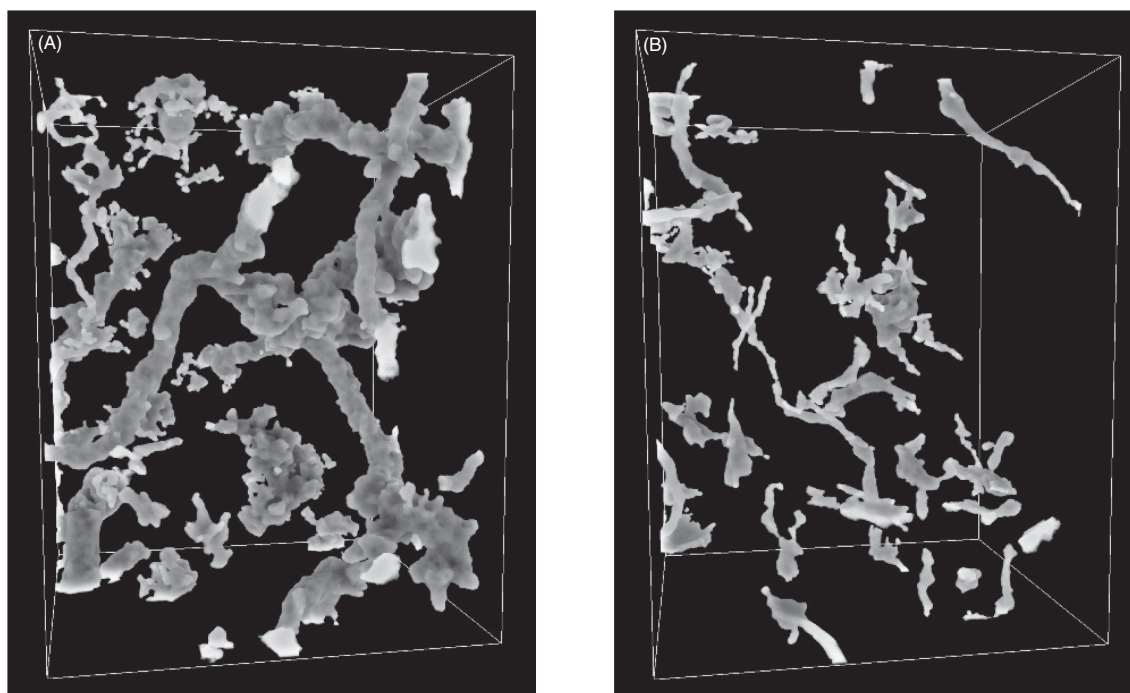


Fig. 1 X-ray images of macropore networks (pore space $>500 \mu\text{m}$ in diameter) in two of the samples included in the study by Koestel et al. (2020). Cylindrical aluminum cores (65 mm in diameter, 60 mm in height) were taken from a soil pit in an agricultural field at Ås, $\sim 30 \text{ km}$ south of Oslo, Norway, at depths of 4–10 cm in the topsoil (a) and 90–96 cm in the subsoil (b). The imaged volumes correspond to a cross-section of $4 \times 4 \text{ cm}$ and a height of 5.2 cm. The original image resolution was $80 \mu\text{m}$.

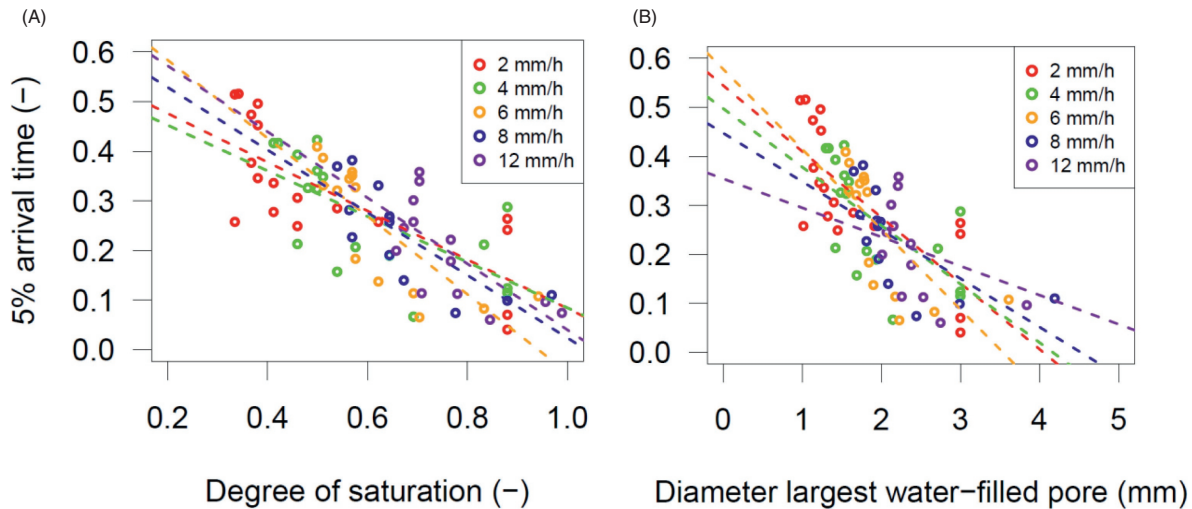


Fig. 2 Relationships between the 5% tracer arrival time and (a) the degree of saturation in the X-ray imaged pore space and (b) the diameter of the largest water-filled pore, for four soils at five different steady water flow rates varying between 2 and 12 mm h⁻¹. Dashed lines represent linear regression lines for each flow rate. All regression slopes were significantly (p value < 0.05) different from zero. The 5% tracer arrival time is a non-parametric measure of the strength of preferential flow, which is defined as the (normalized) time at which 5% of an applied tracer pulse has leached from the bottom of the core. Figure re-drawn from Larsbo, M, Koestel, J, Jarvis, N (2014) Relations between macropore network characteristics and the degree of preferential solute transport *Hydrology and Earth System Sciences* 18: 5255–5269.

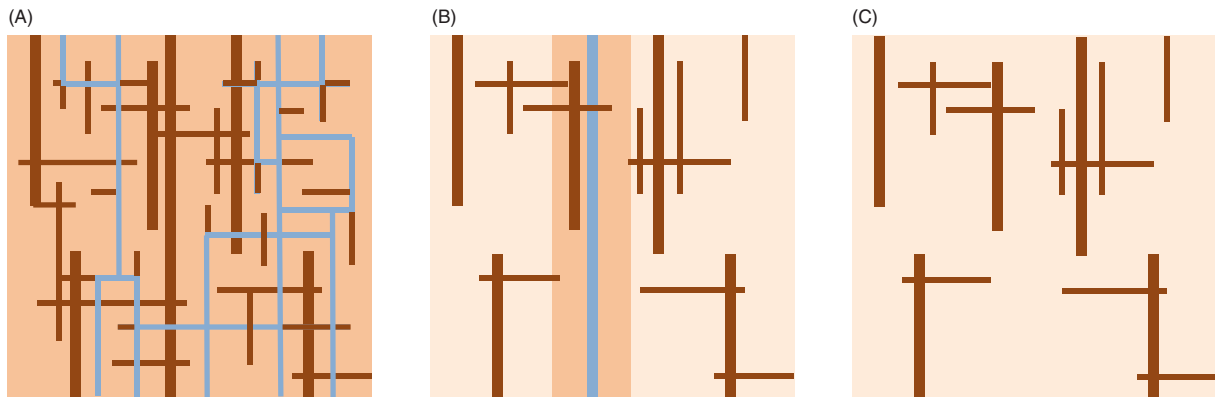


Fig. 3 Schematic illustration of the relationship between pore connectivity, water-filled flow pathways, macroscopic flow capacity (saturated hydraulic conductivity) and the strength of preferential transport. Two classes of soil pore are represented in the figures with contrasting porosity and diameter: a dense network of mesopores and six larger diameter macropores. Conducting water-filled (backbone) pores are indicated in light-blue, empty or dead-end pores in dark brown (a) a dense well-connected percolating pore network (a single pore cluster) with a uniformly high (macroscopic) saturated hydraulic conductivity denoted by the dark shading. The connected mesopore network has a sufficiently large flow capacity to conduct all the incoming water, so the degree of preferential flow is consequently low (b) this network was created from the first network by removing some of the mesopores to create six isolated pore clusters. The flow capacity of the mesopore network is now severely impaired, because it is poorly connected at the sample scale. The water flow is therefore now diverted to a cluster containing a single macropore that is continuous through the sample. The sample now exhibits a high degree of preferential flow. At the macroscopic scale, this is reflected in a continuous high conductivity pathway through the sample (dark shading) surrounded by a low conductivity matrix (light shading) (c) after removing the single continuous macropore, sample-scale continuity breaks down, which reduces the strength of preferential flow. Saturated hydraulic conductivity is also now uniformly low. Figure adapted from Larsbo, M, Koestel, J, Jarvis, N (2014) Relations between macropore network characteristics and the degree of preferential solute transport. *Hydrology and Earth System Sciences* 18: 5255–5269.

and theoretical studies suggest that preferential flow may occur through soil containing disconnected macropore networks, although it does so to a more limited extent (see studies cited in Jarvis et al., 2016). A lack of connected macroporosity can also lead to soil waterlogging, surface runoff and erosion during especially heavy rainstorms.

Precipitation rates will usually be much smaller than the soil saturated hydraulic conductivity in soils with percolating macropore networks. In such cases, the water flow in macropores will be unsaturated and its geometrical configuration will vary with pore network characteristics, input rates and the degree of saturation (e.g., Jarvis, 2007; Jarvis et al., 2016, 2017b; Nimmo, 2020). At low

macropore saturations, “film” or “rivulet” flow may take place along macropore surfaces. As water inputs and the degree of saturation increase, capillary “bridging” across the narrowest sections of macropores and “pulse” flow may occur (Jarvis, 2007). Bottlenecks along the flow pathways, together with air entrapment, mean that macropores usually remain unsaturated during water flow even under extremely high input rates or even saturation of the soil surface (Beven and Germann, 1982; Jarvis, 2007; Nimmo, 2020). The contrasting flow configurations that develop in response to variations in input rates at the surface have consequences for pore water velocities in macropores and the strength of preferential flow. This is briefly discussed in section “Modeling macropore flow and transport” in relation to the widely-used kinematic wave model, where the kinematic exponent should reflect differences in flow configurations in soil macropores.

Modeling macropore flow and transport

A wide variety of modeling approaches have been developed that can account for macropore flow. We do not attempt to provide a comprehensive review here, referring the reader instead to Šimůnek et al. (2003), Gerke (2006) and Jarvis et al. (2016). Instead, we focus on a few of the more widely-used concepts, highlighting their strengths and limitations. Ideally, water flow in soil macropores would be modeled directly at the pore-scale using the fundamental Navier-Stokes equations of flow, or approximations of them. In this way, the impacts of macropore flow at larger scales (e.g., soil profiles) would emerge as a consequence of the characteristics of the macropore network and the imposed boundary conditions. Recent advances in X-ray imaging technologies and computational methods have enabled some studies of this kind (e.g., Scheibe et al., 2015). However, numerical solutions of the pore-scale flow equations are computationally extremely demanding and the sample sizes used for X-ray imaging are limited by the size and resolution of the equipment. As a result, this kind of approach has so far been limited to simulation domains that are too small to be of practical interest and almost certainly much smaller than would constitute a REV for macropore flow in structured soils (Koestel et al., 2020).

At the opposite extreme of model complexity, steady-state pore water flow velocities and travel times in macropores have been estimated from various analytical solutions of the Navier-Stokes flow equations that can be derived for highly idealized macropore geometries and flow configurations, such as free water films of constant width or fully saturated flow through cylindrical pores (i.e., the Hagen-Poiseuille equation). These models can all be encapsulated in a single generalized equation by making use of the concept of an effective soil hydraulic radius, R_h (m), which is defined as the ratio between the macropore water content θ_{mac} ($\text{m}^3 \text{m}^{-3}$) and the wetted specific surface area in soil ($\text{m}^2 \text{m}^{-3}$). In this approach, the rate of gravity-driven laminar water flow, q_{mac} (m s^{-1}), through a macroscopically homogeneous macropore “domain” with only a limited range of macropore diameters, can be expressed as:

$$q_{mac} = \left(\frac{g}{G \mu \tau} \right) \theta_{mac} R_h^2 \quad (1)$$

where g is the acceleration due to gravity ($= 9.8 \text{ m s}^{-2}$), G is a flow geometry factor (dimensionless) varying between 2 for cylindrical geometry and 3 for planar geometry, μ is the kinematic viscosity ($= 10^{-6} \text{ m}^2 \text{s}^{-1}$) and τ (dimensionless) is the flow path tortuosity. Putting appropriate values for R_h in Eq. (1) leads to the special cases of this general model that are applicable to idealized flow geometries. For example, by definition, R_h takes values of $d/4$, $d/2$, and d for saturated cylindrical channels, saturated slits and free planar films, where d is the channel diameter, slit width, or film width. However, great care should be exercised in using simple models based on Eq. (1), as the geometrical configuration of water flowing through macropores in natural soils may not reflect the assumptions underlying the chosen form of the model. For example, the Hagen-Poiseuille equation, which is derived for saturated water flow through a cylindrical pore, is not appropriate to describe macropore flow in natural soils characterized by complex and irregular inter-connected pore networks, through which the flow is usually unsaturated (Jarvis et al., 2016, 2017a; see section “Macropore network characteristics and macropore flow”).

For natural soils that are subjected to highly intermittent precipitation regimes in the field, both macropore water contents and flow rates inevitably vary with time (i.e., transient conditions prevail). The so-called kinematic wave equation has been proposed as a flexible and relatively parsimonious approach to modeling variably-saturated water flow in macropores (Germann, 1985). Jarvis et al. (2017b) showed that an easy-to-use form of the kinematic equation can be derived from Eq. (1) by (i.) assuming a power law relationship between the wetted surface area and water content in the macropores, and (ii.) replacing the difficult-to-estimate parameters that explicitly characterize the geometry of flow by the saturated hydraulic conductivity of the macropores, $K_{s(mac)}$, which can be more easily measured. This leads to (Jarvis et al., 2017b):

$$q_{mac} = K_{s(mac)} \left(\frac{\theta_{mac}}{\phi_{mac}} \right)^\alpha \quad (2)$$

where ϕ_{mac} is the conducting macroporosity and α is the kinematic exponent, which reflects the geometrical configuration of water flowing through the macropores (Jarvis et al., 2017b). The model is completed by combining Eq. (2) with a continuity equation:

$$\frac{\partial \theta_{mac}}{\partial t} = -\frac{\partial q_{mac}}{\partial z} \quad (3)$$

where t is time and z is height. Eqs. (2) and (3) can be solved either analytically for relatively simple initial and boundary conditions (Germann, 1985) or numerically for complex, irregular conditions (Larsbo et al., 2005). Making use of concepts from percolation theory, the conducting macroporosity ϕ_{mac} in Eq. (2) can be identified with the “backbone” of the percolating clusters (i.e., the macropore network excluding dead-end macropores). To maintain strict physical realism, the kinematic exponent should lie between 1 and 3, as this ensures that pore water velocities in macropores increase with increases in flow rate, saturation and wetted surface area (Jarvis et al., 2017b). A value for the kinematic exponent of 1 implies that the velocity of water flow in macropores will be at a maximum independent of input rate and macropore water content (Nimmo, 2007; Jarvis et al., 2017b). At the other extreme (i.e., $\alpha = 3$), flow velocities are at a minimum and wetted surface areas are constant irrespective of θ_{mac} (e.g., Jarvis et al., 2017b).

Combining a macropore flow model with descriptions of flow and transport in the soil matrix significantly enhances the potential range of applications of the model to help resolve practical questions in management, policy and decision-making. Dual-permeability models (Šimůnek et al., 2003; Gerke, 2006) are commonly used for such purposes. These models treat the soil macroporosity and matrix as two interacting flow regions without any explicit consideration of the geometry and topology of flow paths in either domain. Classical theory (i.e., the Richardson-Richards equation and ADE) is generally adopted to describe flow and transport in the soil matrix. Different approaches have been used to calculate water flow in the macropore domain. In some models, the Richardson-Richards equation is used to calculate water flow in both flow domains. One potential problem with this is that some of the underlying assumptions (e.g., that the viscous resistances to flow are dominant and flow is laminar) may not be appropriate for the flow of water in macropores (Jarvis, 2007). A parsimonious alternative approach is to employ the kinematic wave equation to model flow in the macropore region (Larsbo et al., 2005). As with all models, dual-permeability approaches suffer from some theoretical weaknesses and limitations (Vogel, 2019). One important limitation is that any macroporosity present is implicitly assumed to be fully connected, which means that internal catchment due to dead-end macropores can only be simulated at one depth in the soil profile. Another potentially key question that has not been resolved, and which is common to all continuum model approaches, is whether a REV can be identified for macropores (Beven and Germann, 1982; Koestel et al., 2020). At present, the experimental evidence is supportive for topsoil characterized by relatively dense macropore networks (Fig. 2; Larsbo et al., 2014), whereas this condition may not be met for sparsely distributed macropores in subsoil (Fig. 1; Koestel et al., 2020). The approximate first-order treatment of mass transfer between the two distinct pore regions also introduces model error in terms of an apparent time-dependence of the mass transfer coefficient. Another important issue in applying these models is that some of the parameters, especially those controlling exchange of water and solutes between the two flow regions, are difficult or even impossible to measure, and must therefore be derived by model calibration. This emphasizes the importance of adopting suitable methods to deal with model parameter and prediction uncertainties. Nevertheless, despite both theoretical and practical limitations, repeated calibration and testing of dual-permeability models against data from laboratory columns, lysimeters and field plots (Köhne et al., 2009a,b) has generated sufficient confidence in their reliability that they are now commonly used as tools by practitioners and stakeholders for various management purposes.

Conclusions and outlook

Macropore flow has been extensively studied since the early 1980s, but its regulation by pore system characteristics remained poorly understood for many years. This is because until recently the soil has been a “black box,” with the nature of its pore system unseen and therefore impossible to directly quantify. In more recent years, advances in non-destructive X-ray scanning techniques and image analysis tools have allowed researchers to directly observe and quantify the complex multi-scale geometry of the soil pore system, and the water flowing within it, at a steadily improving resolution. This has led to a better understanding of how pore network properties control non-equilibrium macropore flow at the pore to core scale. In this respect, concepts borrowed from percolation theory have also proven useful to interpret experimental results, in particular to highlight the role and significance of network connectivity. However, in the context of macropores and macropore flow, percolation theory has so far mostly remained an explanatory rather than a predictive model. There should therefore be considerable scope to make more use of concepts from percolation theory in macropore flow models. We also expect to see an increasing use of non-invasive imaging techniques to directly test and parameterize macropore flow models. This will lead to an improved understanding of the process itself and more reliable model predictions in practical applications with less need to resort to calibration.

For some management purposes (e.g., as decision-support tools for pesticide registration), model simulations at the soil profile scale are sufficient, which means that important questions related to parameterization and prediction uncertainty are manageable. However, macropore flow also has important consequences for hydrological processes and water quality at much larger scales (e.g., catchments, regions) and in such cases model end-users need to “bridge” across this scale gap for management and decision-making (Vogel, 2019; Nimmo et al., 2021). Hydropedological approaches based on the mapped spatial distribution of representative soil-hydrological types in combination with pedotransfer functions and rules may offer a pragmatic and reliable way to employ soil profile-scale models at the landscape scale (Lin, 2003; Or, 2019; Vogel, 2019).

Macropore networks are dynamic, with changes induced by climate and land use and mediated by physical and biological processes operating at time-scales ranging from seconds to decades (e.g., Meurer et al., 2020). Some models consider the effects of swelling and shrinking and the impacts of tillage and subsequent consolidation at annual time-scales, but so far the temporal dynamics of macroporosity at longer time-scales arising as a consequence of the activity of biological agents (e.g., plant root

production and faunal bioturbation) have been largely neglected. The temporal dynamics of soil macropore networks are especially relevant to the pressing issue of soil degradation, so implementing such processes into models would be desirable.

References

- Allaire S, Roulier S, and Cessna AJ (2009) Quantifying preferential flow in soils: A review of different techniques. *Journal of Hydrology* 378: 179–204.
- Beven K and Germann P (1982) Macropores and water flow in soils. *Water Resources Research* 18: 1311–1325.
- Clothier B, Green SR, and Deurer M (2008) Preferential flow and transport in soil: Progress and prognosis. *European Journal of Soil Science* 59: 2–13.
- Gerke HH (2006) Preferential flow description for structured soils. *Journal of Plant Nutrition and Soil Science* 169: 382–400.
- Germann PF (1985) Kinematic wave approach to infiltration and drainage into and from soil macropores. *Transactions of the American Society of Agricultural Engineers* 28: 745–749.
- Hencher SR (2010) Preferential flow paths through soil and rock and their association with landslides. *Hydrological Processes* 24: 1610–1630.
- Jarvis N (1998) Modelling the impact of preferential flow on non-point source pollution. In: Selim HH and Ma L (eds.) *Physical Non-Equilibrium in Soils: Modeling and Application*, pp. 195–221. Ann Arbor Press.
- Jarvis N (2007) Review of non-equilibrium water flow and solute transport in soil macropores: Principles, controlling factors and consequences for water quality. *European Journal of Soil Science* 58: 523–546.
- Jarvis N, Koestel J, and Larsbo M (2016) Understanding preferential flow in the vadose zone: Recent advances and future prospects. *Vadose Zone Journal*. <https://doi.org/10.2136/vzj2016>.
- Jarvis N, Larsbo M, and Koestel J (2017a) Connectivity and percolation of structural pore networks in a cultivated silt loam soil quantified by X-ray tomography. *Geoderma* 287: 71–79.
- Jarvis N, Koestel J, and Larsbo M (2017b) Reply to ‘Comment on “Understanding preferential flow in the vadose zone: Recent advances and future prospects” by N. Jarvis et al. *Vadose Zone Journal* 16. <https://doi.org/10.2136/vzj2017.01.0034r.09.0075>.
- Katuwal S, Norgaard T, Moldrup P, Lamandé M, Wildenschild D, and de Jonge L (2015) Linking air and water transport in intact soils to macropore characteristics inferred from X-ray computed tomography. *Geoderma* 237–238: 9–20.
- Koestel J, Larsbo M, and Jarvis N (2020) Scale and REV analyses for porosity and pore connectivity measures in undisturbed soil. *Geoderma* 366: 114206.
- Köhne JM, Köhne S, and Šimůnek J (2009a) A review of model applications for structured soils: (A) Water flow and tracer transport. *Journal of Contaminant Hydrology* 104: 4–35.
- Köhne JM, Köhne S, and Šimůnek J (2009b) A review of model applications for structured soils: (B) Pesticide transport. *Journal of Contaminant Hydrology* 104: 36–60.
- Larsbo M, Roulier S, Stenemo F, Kasteel R, and Jarvis N (2005) An improved dual-permeability model of water flow and solute transport in the vadose zone. *Vadose Zone Journal* 4: 398–406.
- Larsbo M, Koestel J, and Jarvis N (2014) Relations between macropore network characteristics and the degree of preferential solute transport. *Hydrology and Earth System Sciences* 18: 5255–5269.
- Larsbo M, Koestel J, Kätterer T, and Jarvis NJ (2016) Preferential transport in macropores is reduced by soil organic carbon. *Vadose Zone Journal*. <https://doi.org/10.2136/vzj2016.03.0021>.
- Lawes JB, Gilbert JH, and Warington R (1882) *On the Amount and Composition of the Rain and Drainage Waters Collected at Rothamsted*. London: William Clowes & Sons Ltd.
- Lin H (2003) Hydropedology: Bridging disciplines, scales, and data. *Vadose Zone Journal* 2: 1–11.
- Lucas M, Vetterlein D, Vogel H-J, and Schlüter S (2020) Revealing pore connectivity across scales and resolutions using X-ray CT. *European Journal of Soil Science* 72: 546–560.
- Meurer K, Barron J, Chenu C, et al. (2020) A framework for modelling soil structure dynamics induced by biological activity. *Global Change Biology* 26: 5382–5403.
- Nimmo J (2007) Simple predictions of maximum transport rate in unsaturated soil and rock. *Water Resources Research* 43: W05426. <https://doi.org/10.1029/2006WR005372>.
- Nimmo J (2020) The processes of preferential flow in the unsaturated zone. *Soil Science Society of America Journal* 85: 1–27.
- Nimmo JR, Perkins KS, Plampin MR, et al. (2021) Rapid response unsaturated zone hydrology: Small-scale data, small-scale theory, big problems. *Frontiers in Earth Science* 9: 613564.
- Or D (2019) The tyranny of small-scales—On representing soil processes in global land surface models. *Water Resources Research* 55. <https://doi.org/10.1029/2019WR024846>.
- Scheibe TD, Perkins WA, Richmond MC, et al. (2015) Pore-scale and multiscale numerical simulation of flow and transport in a laboratory-scale column. *Water Resources Research* 51: 1023–1035.
- Šimůnek J, Jarvis NJ, van Genuchten MT, and Gärdenäs A (2003) Review and comparison of models for describing nonequilibrium and preferential flow and transport in the vadose zone. *Journal of Hydrology* 272: 14–35.
- Thomas GW and Phillips RE (1979) Consequences of water movement in macropores. *Journal of Environmental Quality* 8: 149–156.
- Vogel H-J (2019) Scale issues in soil hydrology. *Vadose Zone Journal* 18: 190001.
- Wilson GV, Nieber JL, Sidle RC, and Fox GA (2013) Internal erosion during soil pipe flow: State of the science for experimental and numerical analysis. *Transactions of the American Society of Agricultural and Biological Engineers* 56: 465–478.

Unstable preferential flow in porous media

Naaran Brindt, J-Yves Parlange, and Tammo S Steenhuis, Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of TS Steenhuis, JY Parlange, YJ Kim, DA DiCarlo, JS Selker, PA Nektarios, DA Barry, F Stagnitti, UNSTABLE FLOW, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 197–201, ISBN: 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00433-1>

Introduction	315
Early work	315
Observations	315
Conditions for wetting front instability	316
Checking stability criteria using experimental evidence	317
Dynamic contact angle	317
Modeling gravity-driven fingers fingered flow	318
Continuum models	319
Semi-continuum modeling	319
Pore network models	319
Summarizing remarks	319
References	320

Abstract

Unstable flow of water through preferential flow paths can take place in porous media as columns, also called fingers. To occur, the matric potential at the fingertip must be greater than at the surface, with a discontinuous matric potential across the wetting front. We describe the past research on unstable flow, including the criteria that cause the instability, and present a theory to explain how the contact angle changes dynamically at the front. This theory leads to more realistic models for unstable flow.

Glossary

Contact angle The angle formed between the surface and the line tangent to the edge meniscus near the soil surface. The contact angle depends on the physical properties of the soil and the liquid but also on the velocity of the wetting front.

Dynamic contact angle The contact angle that changes as a function of the velocity of the water flow. When the contact angle increases, the matric potential will increase as well and is the reason the fingertip is the wettest part of the finger.

Fingers In soils consist of wetted columns in otherwise dryer homogeneous soil.

Matric potential or capillary pressure Is the potential energy of water per unit volume relative to pure water under atmospheric pressure. It is always negative and is related to how strongly the water is being held in the soil films surrounding the grains in the soil. Water films are not present in dry soil, and the matric potential does not exist. Hence it can be considered discontinuous at the wetting front.

Preferential flow Refers to the uneven and often rapid movement of water and solutes through porous media, typically soil, characterized by regions of enhanced flux such that only a small fraction of media participates in most of the flow, allowing much faster transport of a range of contaminants, including pesticides, nutrients, trace metals, and manurial pathogens.

Unstable Flow The flow is considered unstable when disturbances are amplified and maintained in a regular pattern, i.e., a wavy infiltration front develops in column-like structures called fingers

Vadose zone Is the variably saturated zone between the ground surface and the permanent groundwater

Key points

- Preferential flow is the rule rather than the exception in soils.
- There are still gaps in the understanding of how water and pollutants move preferentially in the unsaturated zone.
- Preferential flow can occur in structured soils in regions with greater conductivity and as unstable flow in homogeneous and layered coarse-textured or water repellent soils.
- This chapter discusses unstable preferential flow.

Introduction

The importance of understanding preferential flow processes cannot be overstated, considering its relevance to groundwater recharge, crop irrigation, and vadose-zone transport of nutrients and contaminants. According to the late Louis Dekker, preferential flow is the rule rather than the exception (personal communication to T. S. Steenhuis, Summer 2008). Preferential flow affects the residence time of infiltrating water in the vadose zone and impacts the degradation and adsorption of pollutants and nutrients. Preferential flow can occur as macropore flow through locally high conductive paths or as column-type flow in a homogeneous soil due to unstable flow conditions. Both types of preferential flow are similar in that the gravity force dominates sorptive processes. In this section, we are only concerned with unstable finger flow formation driven by gravity that, according to the Science Citation Index, is not well recognized. The key terms “infiltration” and “soil” resulted in over 20,000 publications in June 2022, while only 270 of these contained the word “unstable” in addition.

Generally, it is assumed in the theory describing the water fluxes in the vadose zone that small perturbations at the wetting front during infiltration will dissipate over time. However, research has shown that this assumption is not always valid. Small perturbations can grow over time and combine to form preferential paths in homogeneous soils, commonly called fingers in the literature. However, naming them columns such as proposed by Parlange et al. (2002) in their review of earlier studies on unstable flow before 2002 would have been more appropriate (Figs. 1 and 2). The sharp boundary between the wet path and the dry soil differs from the classic soil physical infiltration front with monotonously increasing moisture content. This results from a discontinuity in matric potential at the wetting front with the dry soil, retarding the movement and resulting in moisture buildup. For gravity-driven, column-like features having a sharp boundary with the dry soil, moisture content decreases, and pressure becomes more negative monotonically in the opposite direction of the flux (Figs. 1 and 2).

Early work

Observations

In this research field, Jamison (1946) and Bond (1964) were the first to notice the presence of narrow wet “channels” or “tongues” in otherwise dry sandy soil. In the experiments of Hill and Parlange (1972), the column persistence was controlled, given sufficient time for vapor transport to establish an equilibrium by hysteresis between the soil water in the finger (on a wetting branch) and the outside (on a drying branch, Liu et al., 1995; DiCarlo et al., 1999). Parlange et al. (2002) showed that the width of the two-dimensional columns, d , in slab chambers is usually well approximated by:

$$d = \frac{\pi S^2}{K_s(\theta_s - \theta_i)} \frac{1}{1 - \frac{Q}{K_i}} \quad (1a)$$

Which can be simplified to

$$d = \frac{2\pi}{\alpha} \quad (1b)$$

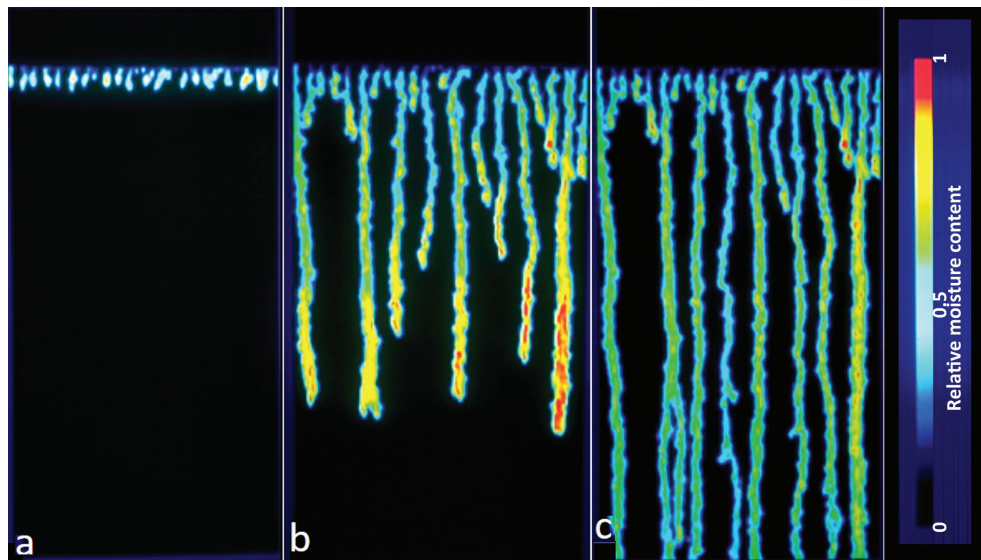


Fig. 1 Fingers formation and development in coarse sand below a fine sand layer in a flow slab experiment after: (a) 2 min, (b) 5 min, and (c) 15 min of water infiltration. The fine sand layer where water was uniformly distributed is not shown. The color indicates the moisture contents, with red the closest to saturation. Then the colors for decreasing moisture contents are yellow, light green, light blue, dark blue, and black. Reproduced from Glass RJ, Steenhuis TS and Parlange JY (1989) Mechanism for finger persistence in homogeneous, unsaturated, porous media: Theory and verification. *Soil Science* **148**: 60–70.

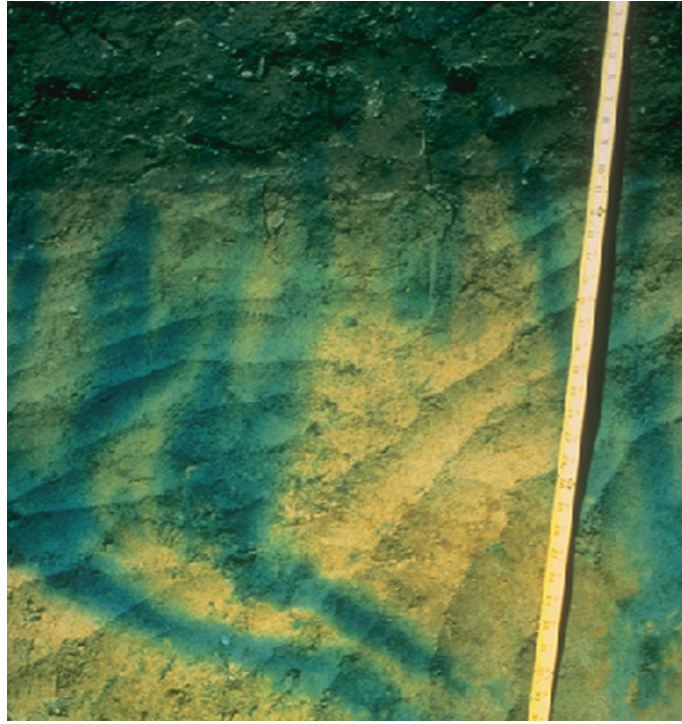


Fig. 2 Finger flow pathways (dyed blue with the infiltrating water) in a soil profile in coastal sediments in Delaware. Water moves down like fingers and then encounters a coarser layer over which it flows sideways (Picture made by Jan Boll).

where S is the sorptivity, Q is the imposed water flux, and K_s is the saturated conductivity of the coarse sand layer, θ_s is the near-saturated moisture content at the fingertip and θ_i is the initial moisture content, α is a constant in the exponential unsaturated hydraulic conductivity equation and is approximately equal to the reciprocal of the height of the capillary fringe. For the three-dimensional fingers, the π should be replaced by 4.8.

Saffman and Taylor (1958) were the first to show instability in a Hele-Shaw cell (mimicking a porous medium). Their experiment with nearly 4000 citations in Google Scholar profoundly impacted our understanding of instability phenomena in porous media. There are, however, several significant differences between this ideal study and unstable flow in soils. In Hele-Shaw cells, hysteresis is irrelevant; consequently, fingers widen behind the wetting front and merge in Hele-Shaw cells. Hence, the length of the fingers remains approximately the same when moving downwards. In soils, downward moving fingers have the same width and remain distinct. Another complication was the use of oil by Saffman and Taylor, leading to the expression “viscous fingering”. Since similar patterns can be obtained with water infiltrating in unsaturated soil with entrapped air, viscosity is not a very important characteristic for the instability to form.

Conditions for wetting front instability

For gravity-driven column-like wetted volumes, the stability criteria of Saffman and Taylor (1958) can be written for the case where water moves down under gravity and the viscosity of air is negligible (Parlange and Hill, 1976; Parlange et al., 2002),

$$K_s > Q \quad (2)$$

Raats (1973) postulated that flow is unstable when the pressure at the wetting front, h_L is greater than the pressure at the surface h_0 (i.e., is more negative at the surface than at the wetting front):

$$h_0 - h_L < 0 \quad (3)$$

Raats (1973) derived the condition in Eq. (3) by assuming that velocity of a homogeneous, stable front, v_w increased with depth for the perturbation to grow in time and become unstable:

$$\frac{dv_w}{dL} > 0 \quad (4)$$

where L is the wetting front depth. Raats (1973) also noted that the increased contact angle for dry water repellent soils would also meet the condition in Eq. (3).

Checking stability criteria using experimental evidence

Several researchers have confirmed the finding of Hill and Parlange (1972). In the laboratory, unstable fingered flow occurs when a fine layer restricts the flow to dry, coarse sand and under field conditions in layered soils, confirming Eqs. (2) and (4). The fine layer restricts the flow (Eq. 2), and the wetting front velocity is smaller in the fine soil than in the coarse soil below (Eq. 4).

Another method to reduce the flow rate is to apply water at less than the saturated conductivity (Selker et al., 1992a). They and others have shown that this technique will produce unstable fingered flow in dry, wettable sandy soils. In this case, small amounts of initial moisture strongly inhibited instability pattern development (Yao and Hendrickx, 1996; Selker et al., 1992a) also measured that the capillary pressure at the wetting front was less than above the fingertip and confirmed the validity of Eq. (3). However, the downward velocity of the finger was constant; thus, the condition in Eq. (4) that the velocity increased with depth was not met.

Finally, following the instability criteria in Eq. (2), experimentally increasing the flow rate to near-saturated hydraulic conductivity produced monotonic wetting profiles in wettable soil (Geiger and Durnford, 2000). However, unlike the stability criteria in Eqs. (2) and (4), extremely small flow rates of less than 0.12 mm/h did not result in fingered flow (Hendrickx and Yao, 1996).

As postulated by Raats (1973), Dekker and coworkers in the early 1990s, Dekker and Ritsema (1994) and Ritsema and Dekker (1994) showed that soil water repellency in the field caused unstable flow through permanent paths. Since these early experiments, other researchers have found that unstable fingered flow occurred in water repellent soils of various textures, confirming the validity of the condition stated in Eq. (3).

In addition to these field experiments, slab chambers experiments filled with sand of different water repellency degrees were conducted in the late 1990s by Bauters et al. (2000) and Wang et al. (1998). Soils can have different degrees of water repellency depending on the contact angle. It is often called subcritical water repellency when the contact angle is less than 90 degrees. As few as three water repellent sand grains in a hundred caused the soil to become subcritical water repellent and resulted in wetting front instability (Bauters et al., 2000). In the Bauters experiments, water ponded on a strong water repellent soil surface before it infiltrated suddenly. Thus, the unstable flow in water repellent sands does not satisfy the instability criteria outlined in 4. Eq. (2) is immaterial when water ponds at the surface. Condition 3 of greater pressure at the top than at the wetting front was met (Bauters et al., 2000).

Entrapped air also caused finger formation. The air is entrapped when a soil surface is wetted uniformly, and the wetting front is saturated in an otherwise confined domain. In these cases, the downward movement of the wetting front compresses the air underneath, as shown experimentally by Wang et al. (1998). These fingers are therefore not caused by instability criteria set in Eqs. (2)–(4) since they do not show the high moisture content at the front.

Dynamic contact angle

In all cases where gravity-driven fingers form under atmospheric pressure, the moisture content and pressure—when measured—were greater at the fingertip than some distance above it, as postulated by Raats (1973) and is a characteristic of gravity-driven fingers. Criteria in Eqs. (2) and (4) were not always met in fingered flow experiments.

The implicit assumption for establishing the instability criteria in the 1980s was that Darcy's law and the mass conservation could describe the unstable phenomena at the wetting front. Some studies (e.g., Geiger and Durnford, 2000) inferred that the capillary pressure of the edge of the finger (including the fingertip) is discontinuous. For discontinuous pressure, the capillary pressure below the wetting front is independent of that at the wetting front. The capillary pressure, h (m), under static conditions at the discontinuous wetting front capillary pressure can be expressed as (Raats, 1973):

$$h = -\frac{2\sigma \cos \theta}{\rho g r} \quad (5)$$

where σ is the surface tension (N/m), ρ is the density of the fluid (kg/m³), g is the gravitational constant (m/s²), r is the pore diameter and is the θ contact angle. For a moving wetting front, the capillary pressure can be expressed as a function of the dynamic contact angle, θ_d

$$h = -\frac{2\sigma \cos \theta_d}{\rho g r} \quad (6)$$

The dynamic contact angle is dependent on the static contact angle, surface tension, viscosity, and the velocity of water in the pore (Hoffman, 1975) and can be expressed based on experimental evidence (Baver et al., 2014) as

$$\theta_d = \theta_s + \left(1 - \frac{\theta_s}{\pi}\right) \arccos [1 - 2 \tanh (4.96 Ca^{.702})] \quad (7)$$

where θ_s is the static contact angle (radians) of the water with silica grains, and Ca is the dimensionless capillary number defined as:

$$Ca = \frac{\mu V}{\sigma} \quad (8)$$

where μ is the viscosity of the liquid (Pa s), V is the contact line velocity (m/s), and σ is the surface tension (N/m) between air and water.

To find the velocity of the water in the pores, experimental observations with a high-speed camera by Moebius and Or (2012) have shown that water moves as bursts at the wetting front with velocities in pores that are more than 50 times faster than the average pore velocity. The pore velocity of the water at the wetting front can be obtained for small fluxes by assuming that water is flowing through one pore at a time. From a mass balance point of view, when water goes down one pore at a time, the minimum pore velocity is equal to the flux (Q , m^3/s) in the finger divided by the area of the pore neck. It might be greater when there is a pause between the breakthroughs (Steenhuis et al., 2013). By calculating the velocity this way, we can use Eqs. (7) and (8) to find the dynamic contact angle, θ_d . Substituting θ_d in Eq. (6), the capillary pressure at the wetting front can be found. The moisture content can be calculated using the soil characteristic curve.

Thus, the assumption that the velocity in the pore affects the dynamic contact angle leads directly to the instability criteria postulated by Raats (Eq. 3): The dynamic contact angle increases the contact angle at the wetting front and thereby lowers the capillary pressure at the wetting front below that of the capillary pressure in the finger itself.

The theory of the dynamic contact angle was tested by Steenhuis et al. (2013). They calculated the capillary pressures at the wetting front with Eqs. (6)–(8) for the extensive experiments in 1.3 cm columns carried out by DiCarlo (2007). The observed and predicted capillary pressures are given in Fig. 3 for the initial moisture contents ranging from 0 to $0.14 \text{ cm}^3 \text{ cm}^{-3}$. The pore neck diameter was fitted for one experiment. The static contact angle was 50° for dry soil and decreased linearly to 20° for moisture contents of 9% and greater. The observed and predicted data followed the same trend.

The theory also explains previous experimental findings. One of the empirical findings was that finger formation is suppressed in wettable soils with residual moisture (Hendrickx and Yao, 1996). These small amounts of water form films around the grain, making the pressure across the wetting front continuous and resulting in monotonic infiltration profiles with increasing moisture content and capillary pressure towards the soil surface.

Glass et al. (1989) and others showed that wettable soils with uneven moisture distribution generated by previous fingered flow events would infiltrate again through these fingers. Water enters the finger under lower matric potential than the dry soil because of the films around the grains in the finger than are not present in the dry sand. Only when the soil dries out can the path of the finger be at a different location.

Although both wettable and water repellent soils experimentally produce gravity-driven fingered flow, the behavior at the wetting front is different. In water repellent soils, the contact angle is greater in dry soil than in wet soil. Therefore, the instability criterion in Eq. (3) is met independently of a dynamic contact angle tied to the pore velocity as postulated already by Raats (1973). Only for subcritical water repellent soils, the increase in contact angle due to the pore velocity might be significant.

Modeling gravity-driven fingers fingered flow

Since Hill and Parlange (1972) published their observations of gravity-driven fingers in well-sorted sands, researchers have developed models to simulate this unstable phenomenon with the decreasing water content and matric potential opposite the flow direction. The models and analytical solutions (called models for short) can be distinguished into roughly three categories: (1) continuum models where pressure across the wetting front is continuous, and the water flow is simulated for the complete

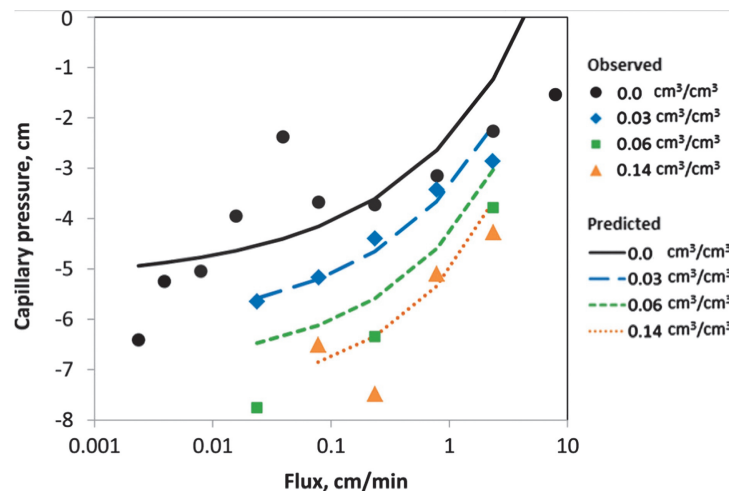


Fig. 3 Observed and predicted capillary pressures in the fingertip as a function of moisture content and imposed fluxes. Observed data (symbols) are from infiltration experiments by DiCarlo (2007) in small 1.3 cm wide columns filled with 20/30 sand and are averages over all replicates as each flux. Predicted lines are based on Eqs. (6)–(8).

profile. (2) semi-continuous models that take the matric potential discontinuity at the wetting front as a given and use soil physics theory to predict the water movement behind and in front of the wetting front, and (3) pore network models that adapt the contact angle at the finger edge.

Continuum models

When the pressure in the soil is continuous, the Richardson-Richards equation applies. However, the equation is unconditionally stable and can only produce a monotonic temporal change in saturation, i.e., the moisture at each point in the soil profile can either increase or decrease. In contrast, as discussed in the previous section, gravity-driven fingering is characterized by an abrupt change in saturation at the wetting front and nonmonotonic saturation temporal change within the fingers (DiCarlo, 2013). As a result, solution obtained using Richardson-Richards' equation cannot produce a gravity-driven unstable flow. Several approaches tackled this limitation by adding a term to the Richardson-Richards equation. A drawback of the continuum models is that they do not reproduce the sharp wetting front.

Semi-continuum modeling

Semi continuum models treat the wetted finger and the surrounding soil as discrete entities. Selker et al. (1992b) published the first type of this model. They predicted the decreasing water content behind the wetting front with the Richardson-Richards equation by using a moving coordinate system, constant pressure at the wetting front, and a finger width predicted by using the equation of Parlange and Hill (1976). DiCarlo (2013) extended the approach of Selker and coworkers by adding pore-scale physics to the continuum flow equation at the wetting front. Kmec et al. (2021) used the DiCarlo approach and combined ideas from continuum mechanics and invasion percolation models to divide the modeling domain into finite subdomains, each represented by average capillary pressure values. Brindt and Wallach (2020) adapted the Stefan heat-flow moving boundary problem boundary approach for characterizing the wetting front advancement of gravity-driven unstable flow by controlling the contact angle at the wetting front. The main disadvantage of the semi continuum models is that matric potential at the front cannot be determined independently and must be obtained from experimental results.

Pore network models

Pore network models have pore and throat elements that can only have discrete saturation states. This modeling approach produced finger-like wetting patterns with drying behind the tip for all fluxes, contrary to the experimental results. Other pore network models incorporated rules that allowed partial wetting of pores (Hughes and Blunt, 2000).

Common to all the above mentioned models is that the wetting front is retarded. The retardation in continuum models is achieved by ad-hoc measures of which the physical validity remains in doubt (DiCarlo, 2013). Pore-based network models have been relatively successful in simulating the transition between monotonic and unstable profiles as a function of the flux. Still, it requires a more robust description of the interface of the two fluids (DiCarlo, 2013). The dynamic contact angle theory presented can be used as a framework for modeling the pressures at the interphase of the two fluids for better performance of pore network models.

Summarizing remarks

Water tends to flow in preferential paths since it is thermodynamically more favorable. Unstable gravity-driven fingered flow is one of the consequences. Unstable flow occurs when the matric potential at the wetting front is greater than at the surface, made possible by a matric potential discontinuity across the wetting front due to the lack of water films in the dry soil. In monotonic moisture content profiles, water films aid the advance of the water. When water is strongly adsorbed to the grains in dry soil (and the distinct water films still need to be formed), water movement through the pore necks is retarded. Since the resistance to flow is different among pore necks, water will flow through one pore neck at a time in a short time. The rapidly flowing water will cause a flattening of the meniscus compared to the pores behind the front lowering the matric potential. Since the lower matric potential is in equilibrium with greater water content, the tip of the finger is wetter than the remaining part.

Since the quantity of water has become limited in many parts of the world, water conservation practices such as wastewater irrigation will become more prevalent. Organics in the wastewater can make soils more water repellent. Thus, we expect that areas irrigated with wastewater will show preferential flow independent of the soil texture, thereby reducing irrigation efficiency, as noted by Jamison almost a century ago. Therefore, understanding unstable flow processes may help drain less water from soil profiles, saving water and keeping the groundwater free of additional pollutants.

References

- Bauters TWJ, Steenhuis TS, DiCarlo DA, et al. (2000) Physics of water repellent soils. *Journal of Hydrology* 231–232: 233–243.
- Baver CE, Parlange J-Y, Stooft CR, et al. (2014) Capillary pressure overshoot for unstable wetting fronts is explained by Hoffman's velocity-dependent contact-angle relationship. *Water Resources Research* 50: 5290–5297.
- Bond RRD (1964) The influence of the microflora on the physical properties of soils. II. Field studies on water repellent sands. *Australian Journal of Soil Research* 2: 123.
- Brindt N and Wallach R (2020) The moving-boundary approach for modeling 2-D gravity-driven stable and unstable flow in partially wettable soils. *Water Resources Research* 56: e2019WR025772.
- Dekker LW and Ritsema CJ (1994) How water moves in a water repellent sandy soil: 1. Potential and actual water repellency. *Water Resources Research* 30: 2507–2517.
- DiCarlo DA (2007) Capillary pressure overshoot as a function of imbibition flux and initial water content. *Water Resources Research* 43: W08402.
- DiCarlo DA (2013) Stability of gravity-driven multiphase flow in porous media: 40 Years of advancements. *Water Resources Research* 49: 4531–4544.
- DiCarlo DA, Bauters TWJ, Darnault CJG, Steenhuis TS, and Parlange J-Y (1999) Lateral expansion of preferential flow paths in sands. *Water Resources Research* 35(2): 427–434.
- Geiger SL and Durnford DS (2000) Infiltration in homogeneous sands and a mechanistic model of unstable flow. *Soil Science Society of America Journal* 64: 460–469.
- Glass RJ, Steenhuis TS, and Parlange J-Y (1989) Mechanism for finger persistence in homogeneous, unsaturated, porous media: Theory and verification. *Soil Science* 148: 60–70.
- Hendrickx JMH and Yao T (1996) Prediction of wetting front stability in dry field soils using soil and precipitation data. *Geoderma* 70: 265–280.
- Hill DE and Parlange J-Y (1972) Wetting front instability in layered soils. *Soil Science Society of America Journal* 36: 697–702.
- Hoffman RL (1975) A study of the advancing interface I. Interface shape in liquid-gas systems. *Journal of Colloid and Interface Science* 50: 228–241.
- Hughes RG and Blunt MJ (2000) Pore scale modeling of rate effects in imbibition. *Transport in Porous Media* 40: 295–322.
- Jamison VC (1946) The penetration of irrigation and rain water into sandy soils of central Florida. *Soil Science Society of America Journal* 10: 25–29.
- Kmec J, Fürst T, Vodák R, and Šir M (2021) A two-dimensional semi-continuum model to explain wetting front instability in porous media. *Scientific Reports* 11: 3223.
- Liu Y, Parlange J-Y, Steenhuis TS, and Haverkamp R (1995) A soil water hysteresis model for fingered flow data. *Water Resources Research* 31: 2263–2266.
- Moebius F and Or D (2012) Interfacial jumps and pressure bursts during fluid displacement in interacting irregular capillaries. *Journal of Colloid and Interface Science* 377: 406–415.
- Parlange J-Y and Hill DE (1976) Theoretical analysis of wetting front instability in soils. *Soil Science* 122: 236–239.
- Parlange J-Y, Steenhuis TS, Li L, Barry DA, and Stagnitti F (2002) Column flow in stratified soils and fingers in Hele-Shaw cells: A review. In: Raats PAC, Smiles D, and Warrick AW (eds.) *Geophysical Monograph Series 129: Environmental Mechanics: Water, Mass and Energy Transfer in the Biosphere*, pp. 79–85. Washington, DC: American Geophysical Union.
- Raats PAC (1973) Unstable wetting fronts in uniform and nonuniform soils. *Soil Science Society of America Journal* 37: 681–685.
- Ritsema CJ and Dekker LW (1994) How water moves in a water repellent sandy soil: 2. Dynamics of fingered flow. *Water Resources Research* 30(9): 2519–2531.
- Saffman PG and Taylor G (1958) The penetration of a fluid into a porous medium or Hele-Shaw cell containing a more viscous liquid. *Proceedings of the Royal Society A Mathematical Physical and Engineering Sciences* 245: 312–329.
- Selker J, Leclercq P, Parlange J-Y, and Steenhuis T (1992a) Fingered flow in two dimensions: 1. Measurement of matric potential. *Water Resources Research* 28: 2513–2521.
- Selker J, Parlange J-Y, and Steenhuis T (1992b) Fingered flow in two dimensions: 2. Predicting finger moisture profile. *Water Resources Research* 28: 2523–2528.
- Steenhuis TS, Baver CE, Hasanpour B, et al. (2013) Pore scale consideration in unstable gravity driven finger flow. *Water Resources Research* 49: 7815–7819.
- Wang Z, Feyen J, and Elrick D (1998) Prediction of fingering in porous media. *Water Resources Research* 34: 2183–2190.
- Yao T and Hendrickx JMH (1996) Stability of wetting fronts in dry homogeneous soils under low infiltration rates. *Soil Science Society of America Journal* 60: 20–28.

Leaching processes

Brent Clothier and Steve Green, Plant and Food Research Ltd, Palmerston North, New Zealand

© 2023 Elsevier Ltd. All rights reserved.

This is an update of B.E. Clothier, S. Green, LEACHING PROCESSES, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 336–343, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00388-X>.

Introduction	321
Drainage quantity	323
Leachate quality	324
Preferential leaching processes	324
Field measurement devices	327
Nitrogen leaching	327
Salt leaching	328
Modelling and risk assessment	330
Pesticide leaching	332
Upscaling	333
Summary	334
References	334

Abstract

Soil is the thin and biologically active skin of our Green Planet. Soil, by virtue of its intrinsic fertility, plays a key role in sustaining our primary-production systems that seek to feed our growing populations. To enhance the productivity of these agricultural systems, irrigation water, fertilizers, salts, and pesticides are often applied to crops growing on soil. As well, soils are sometimes used for the disposal, either intentional or accidental, of certain waste products. Also the soil surface is the depository of atmospheric compounds that might either be natural, such as salt, or compounds from human activities, such as sulfur. Since soil lies astride the main pathway along which rainwater leaches to either groundwater or surface water bodies, the water quality of these receiving water reserves is strongly influenced by land management, and the myriad of interlinked water-transport processes and chemical-exchange mechanisms that operate in the soil domain. Development of sustainable land-management strategies, and the design of environmental-protection protocols, demand that leaching process be well understood.

Key points

- Globally some 720 mm of rains falls on soil every year. Of this, some 310 mm on average either drains through the soil, or runs off into receiving water bodies.
- The flows of agrichemicals, nutrients and salts have their origin in the surface soil, and their transmission to receiving waters is by leaching with the drainage water.
- Drainage flows and leaching pathways are chaotic in the structured soils that typify the earth's surface, and this leads to rapid and far-reaching leaching.
- Novel measurement devices and modern modelling approaches are providing new information and improved understanding of drainage and leaching.
- Better land management practices are needed to maintain aquifer recharge and protect groundwater quantity, and to prevent degradation of ground and surface waters through the leaching of nutrients, agrichemicals, and salts.

Introduction

The world is facing a wicked trilemma. We need to feed and sustain a rapidly growing population, and we must adapt to the exigencies of climate change, all the while functioning within the planet's safe operating space. In 2009 the Stockholm Resilience Institute published a framework to assess how, across seven control variables, the world was operating relative to the planetary boundaries for a safe operating space (Rockström et al., 2009). They updated this in 2015 (Steffen et al., 2015), and this assessment is presented here in Fig. 1. The greatest exceedance of the planetary boundary layer is for biosphere integrity, closely followed by the biogeochemical flows of nitrogen and phosphorus. Rockström et al. (2009) commented that "... modern agriculture is a major cause of environmental pollution, including large-scale nitrogen- and phosphorus-induced environmental change." They added that "... the additional amounts of nitrogen and phosphorus activated by humans are now so large that they significantly perturb the global cycles of these two important elements."

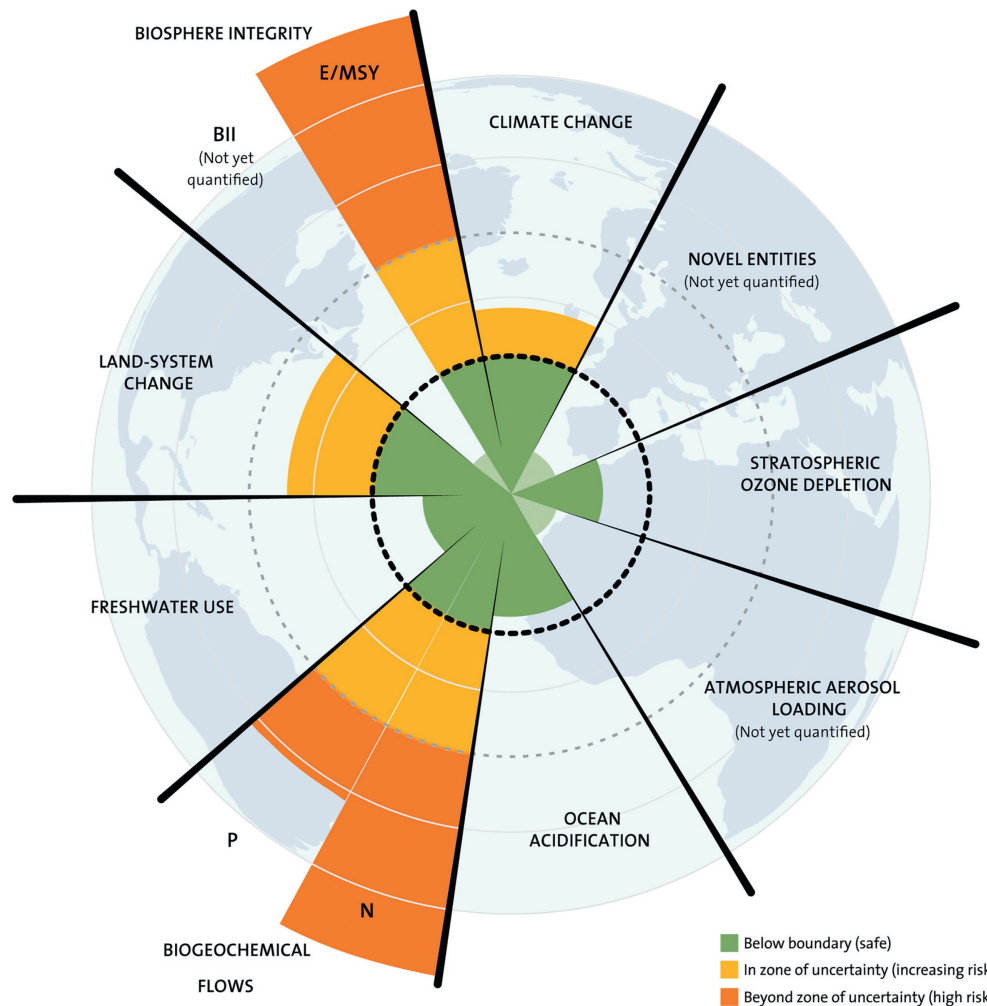


Fig. 1 The current status of seven control variables relative to the planetary boundaries for Earth's safe operating-space. The green sectors are within the safe boundary, with orange being increasing risk, and red is high risk. Here P is phosphorus, N is nitrogen, E/MSY is extinctions per millions species per year, and BII is the Biodiversity Intactness Index. Based on Steffen W, Richardson K, Rockström J, Cornell SE, Fetzer I, Bennett EM, Biggs R, Carpenter SR, de Vries W, de Wit CA, Folke C, Gerten D, Heinke J, Mace GM, Persson LM, Ramanathan V, Reyers B, and Sörlin S (2015) Planetary boundaries: Guiding human development on a changing planet. *Science* 347:6219. doi:10.1126/science.1259855.

The biogeochemical flows of nitrogen and phosphorus, as well as other agrichemicals have their origin in the critical zone of the top few meters of our soils, and their transmission to receiving water bodies is by leaching processes through soil. In addition, the leaching of salts resulting from excessive irrigation using saline waters, especially in arid and semi-arid regions, is resulting in soil salinity and shallow groundwater problems that will require effective management of water and salt in order to avert global food-security issues (Wichelns, 2015).

Soil is the thin and biologically active skin of our Green Planet. Soil, by virtue of its intrinsic fertility, plays a key role in sustaining our primary-production systems that seek to feed our growing populations. To enhance the productivity of these agricultural systems, irrigation water, fertilizers, and pesticides are often applied to crops growing on soil. As well, soils are sometimes used for the disposal, either intentional or accidental, of certain waste products. Also, the soil surface is the depository of atmospheric compounds that might either be natural, such as salt, or compounds from human activities, such as sulfur. Since soil lies astride the main pathway along which rainwater leaches to either groundwater or surface water bodies, the water quality of these receiving water reserves is strongly influenced by land management, and the myriad of interlinked water-transport processes and chemical-exchange mechanisms that operate in the soil domain. Development of sustainable land-management strategies, and the design of environmental-protection protocols, demand that leaching process be well understood.

Leaching encompasses the gamut of biophysical mechanisms and chemical exchange processes that can result in altered chemical concentrations in any pore-water that is percolating through soil. There are two aspects to leaching; the quantity of the water that is draining through the soil and destined for subterranean, or surface water bodies; and the chemical concentration, or quality, of that pore water. We use the term leaching to encompass these two meanings; the deep seepage of water, more commonly called drainage, plus the leaching export of chemicals carried by that drainage.

In this chapter, we first explore the controls on the quantity of drainage water quitting the base of the rootzone (D , mm). We then describe the chemical processes that lead to changes in the quality of that drainage water. For risk assessment purposes it is necessary to know both the concentration of a chemical in the leachate (C , mol m⁻³), plus the total chemical load (per unit ground area) that is moving to receiving water bodies (M , mol m⁻²; or more commonly in kg ha⁻¹). The total leaching of chemical, M (load), is the product of D (quantity) and C (quality). Thus, in our discussion of leaching processes, we need to consider both the transport of water, along with the chemical processes operating in soil. We outline, with examples, the mathematical description of transport processes, and the chemical exchange mechanisms that determine the leachate load.

Physically, soil is a porous medium comprising a complex arrangement of the three phases of solid, liquid and air. Much to the chagrin of scientists, convective and diffusive flow through soil does not behave in the uniform and isotropic way assumed by their simple, theoretical descriptions. Rather, aggregates, cracks and macropores serve to create an apparently chaotic flow regime that leads to rapid and far-reaching transport of chemicals. Modelling methods that account for this preferential flow are described. The application of such models for risk assessment purposes is also discussed. We conclude with a modelling assessment of the risks posed to groundwater by the leaching of the pesticides that are applied to grapes growing on soil above an unconfined aquifer.

Drainage quantity

Globally, an average of some 720 mm of rainfall (RF , mm) falls on soil each year. The many nooks and crannies in the soil provide it with a capillary attractiveness so that there is some tendency for the infiltrating rainwater to be absorbed by the soil, near the surface. The ever-present force of gravity tends to draw complementary water deeper into the profile beyond the grasp of the roots. On balance, about 410 mm of the global average rainfall is lost back as vapor to the atmosphere, either by direct free-water evaporation from the soil and directly off plant surfaces (E , mm), or by transpiration through plants (T , mm). Any transpired water must first have been captured from the soil by the plant's roots. The complement of evaporation, namely drainage, D , and runoff, R , averages therefore about 310 mm per year, since on an annual timescale the change in soil-water storage is negligible. Drainage water acts a vehicle, carrying with it passenger solutes that can leach the soil of chemicals that may compromise the quality of receiving waters. For productive systems, rainfall might be augmented by irrigation (I). Indeed, irrigation comprises about 80–85% of the world's consumption of fresh water. And unfortunately some 20% of global gross irrigation water-usage comprises nonrenewable groundwater extraction (Wada et al., 2012), and this nonrenewable contribution more than tripled between 1960 and 2000.

Both for production purposes and environmental protection, it is critical to determine whether the partitioning of rootzone water, and its dissolved solutes, is either into the plants by uptake, or dispatch by drainage toward groundwater, D . Along with the surface evaporation of water, roots are the arbiters of this partitioning, and they largely determine water and solute fate (Fig. 2). In simple terms, roots penetrate the soil to some effective depth, z_R (mm). From within this rootzone they extract liquid water, and some dissolved solutes. Thus rhizosphere processes critically affect both the quantity and quality of leachates leaving the rootzone.

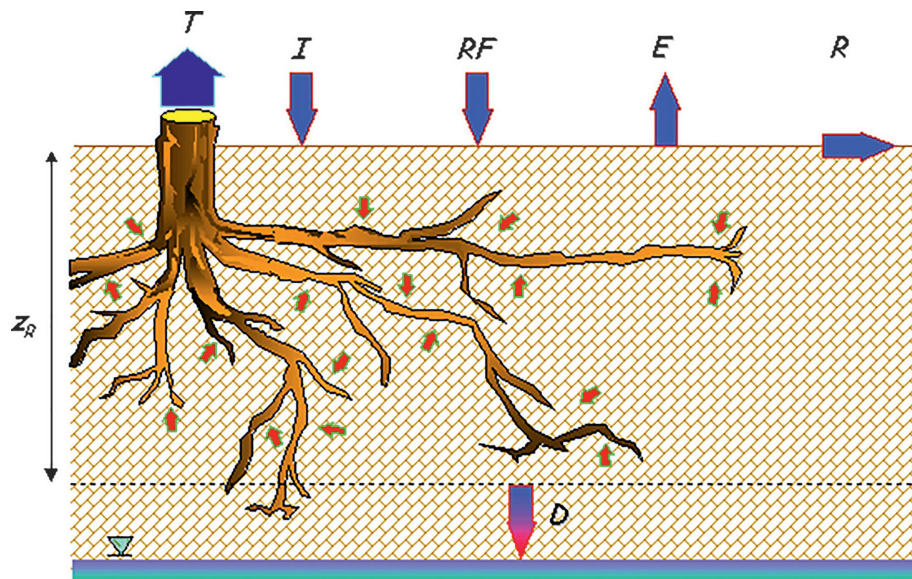


Fig. 2 The fluxes of water in the rootzone. Here the effective depth of rooting is z_R . Inputs into the rootzone include rainfall (RF), as well as irrigation (I) in some cases. Not all of the applied water might infiltrate, as there could be runoff (R). Evaporative losses are either by direct evaporation of water from the soil surface (E), or by transpiration from plants (T). Drainage of water from the base of the rootzone toward groundwater, D , carries with it leaching solutes.

The water balance of the rootzone, over a given time interval Δt , can be written to find the drainage at depth z_R , as

$$D = RF + I - R - E - T - \Delta S, \quad (1)$$

where ΔS is the change, over Δt , in the storage of soil water in the rootzone (mm). The variation in the storage of water can be found from the change in the soil-water content profile, $\theta(z)$ ($\text{m}^3 \text{m}^{-3}$), over Δt , integrated from the surface down to the base of the rootzone,

$$\Delta S = \int_0^{z_R} \Delta \theta(z) dz. \quad (2)$$

Thus the quantity of drainage, D , is implicitly linked to the weather (RF , R , E , and T), as well as to the nature of the soil and the vegetation cover, plus the effective depth and distribution of plant roots.

At a mechanistic level, this interplay between soil-water diffusion and convection, and root-water extraction, can be described by the Richardson-Richards equation,

$$\frac{\partial \theta}{\partial t} = \frac{d\theta}{dh} \frac{\partial h}{\partial t} = \frac{\partial}{\partial z} \left(K(h) \frac{\partial h}{\partial z} \right) - \frac{\partial K(h)}{\partial z} - U(h(z), t) \quad (3)$$

where h (mm) is the soil-water pressure head, which is a function of the soil's water content, and K (mm s^{-1}) is the pressure-head dependent, soil-water conductivity function. The first term on the right-hand side describes the diffusive nature of the capillary-induced movement of water through soil. The second term accounts for direct convection in response to gravity. The last term, U (mm s^{-1}), is the depth-dependent extraction of water by plant roots. Many forms have been proposed for U (Feddes et al., 1976), and all rely on the theoretical insight provided by Gardner's (1960) seminal work on the uptake of water by roots.

The Richardson-Richards equation is the theoretical scheme used to model and predict the partitioning of soil-water fate into either plant uptake (T), or drainage (D) (Clothier and Green, 1997).

Leachate quality

The chemical concentration of the soil's pore-water solution, C , depends on a complex of linked interactions between transient transport processes, chemical exchange mechanisms, as well as production, extraction and degradation reactions. For any chemical species, the convection-dispersion equation describes the resident concentration at any time and location in the soil as

$$\frac{\partial \theta C}{\partial t} + \frac{\partial \rho S}{\partial t} = \frac{\partial}{\partial z} \left(\theta D_s \frac{\partial C}{\partial z} \right) - \frac{\partial q_w C}{\partial z} + \sum \Gamma_i \quad (4)$$

where S is the concentration of chemical that is retained by the soil matrix (mol kg^{-1}), ρ is the soil's bulk density (kg m^{-3}), D_s is the solute's diffusion-dispersion coefficient ($\text{mm}^2 \text{s}^{-1}$), q_w is the water flux density (mm s^{-1}), and Γ_i is the source/sink term for the i reactions that can produce, extract or degrade the chemical species concerned ($\text{mol m}^{-3} \text{s}^{-1}$). The structure of this formula (Eq. 4) highlights the inextricable link between water movement (q_w) (Eq. 3), and chemical-transport processes.

As well, Eq. (4) accounts for the link between the chemical concentration in the soil solution (C), and the amount of chemical that is adsorbed to the surfaces of the soil's matrix (S). This relationship is called the adsorption isotherm. One commonly used form that describes this exchange process is the Langmuir isotherm;

$$S = S_{\max} \frac{mC}{1 + mC} \quad (5)$$

where $S_{\max}$ is the maximum adsorption (mol kg^{-1}), and m is an empirical constant ($\text{m}^3 \text{mol}^{-1}$). Often, however, for simplicity, or in an absence of better information, a linear isotherm is assumed,

$$S = K_D C \quad (6)$$

where K_D is called the distribution coefficient ($\text{m}^3 \text{kg}^{-1}$).

Eqs. (3) and (4) provide us with a mathematical framework that is commonly used to interpret and predict the leaching processes that determine the quantity of leachate (D), and the leachate load of chemical (M).

Preferential leaching processes

The macroscopic theory, on which Eqs. (3) and (4) is based, requires that the soil be a uniform isotropic medium. Unfortunately, soil is a highly structured medium, such that the three phases of solid, liquid, and air are anisotropic. And there is often a high degree of connectedness between these phases. Soil structure is typically characterized by aggregates, fractures, and biogenic macropores (Fig. 3).

The spatial topology of the macroporous networks that typify structured-soil determines the speed and extent of preferential flow processes. As a result of local non-equilibrium processes, there can be more rapid and deeper-than-expected leaching of surface-applied chemicals, relative to that predicted by Eqs. (3) and (4).

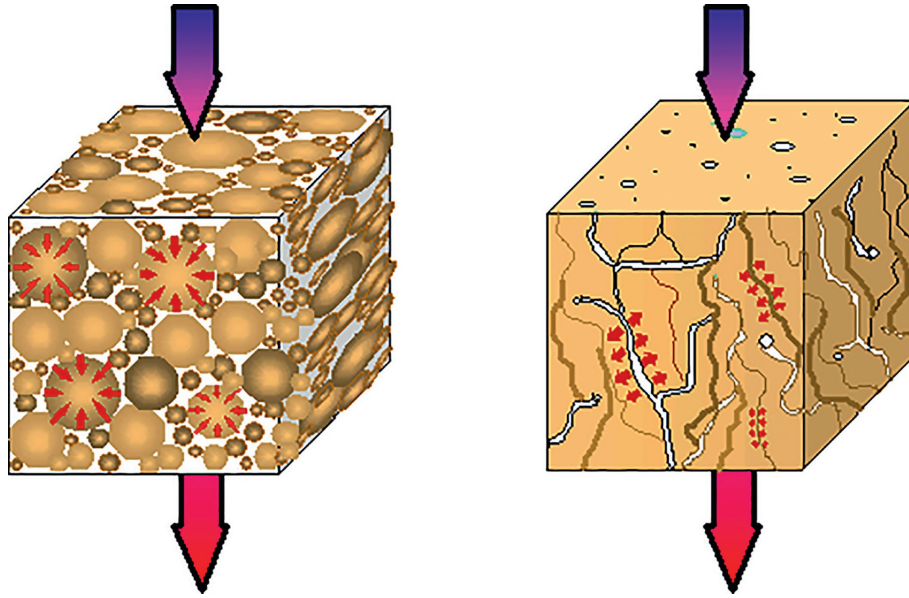


Fig. 3 Soil structure is characterized either by aggregates (left), or macropores and fractures (right). Such structure results in rapid and far-reaching transport processes that lead to preferential leaching of applied chemicals. In such structured soils, there is meanwhile likely to be reduced leaching of chemicals produced from within the soil's matrix.

Not all of the soil's pore water would therefore seem to be actively involved in water transport and chemical exchange. Aggregates and macropores can create a wide spectrum of pore-water velocities associated with a solution that is either invading the soil, or leaching from it. In an aggregated soil, the more-rapid water-flow paths will be those that are connected throughout the inter-aggregate domain. So inter-aggregate water will preferentially wend its way around the aggregates, which tend to possess smaller, yet albeit connected intra-aggregate pores. These would not seem to be as effective at leaching chemicals from rootzone soil, because the flow velocities are lower and the connected path-lengths shorter.

Alternative non-mechanistic transfer-function schemes have been developed for modelling the preferential leaching of chemicals through structured soil. Process-based transport schemes are difficult to parameterize, and so transfer-function models have been considered because they are less data-intensive, and they can support either a deterministic or stochastic framework. The text of [Jury and Roth \(1990\)](#) provides a comprehensive treatment of transfer functions and solute movement through soil, and describes how the solute travel-time probability density function can be taken as a fundamental property of the soil system.

Pedogenic and biogenic processes can impart upon soil a macroporous structure characterized by a longitudinal connectedness. The enhanced mobility of flow down these variously sized macropores, as predicted by Poiseuille's Law of flow through a pipe, is abetted by macropore connectivity to produce preferential transport of the invading solution, or enhanced leaching of some part of the resident solution.

In an illustrative study of preferential leaching in the laboratory with undisturbed cores of soil growing pasture, [Scotter and Kanchanasut \(1981\)](#) used two dye tracers to mimic leaching. Methylene blue dye, applied in free-water ponded on the soil's surface ($h_o = 0$), was found to stain only a small volumetric fraction of the soil's volume ([Fig. 4](#)). When water containing rhodamine-B dye was applied to the soil in water at the slight unsaturated surface potential of -20 mm, the pattern of leaching was found to involve a larger volumetric fraction of the soil's porosity. Obviously, leaching in structured soils can, depending on hydraulic conditions, involve differing volumes of the soil's wetted porosity.

One of the most-popular mechanistic models currently used to model the preferential leaching that results from either the presence of aggregates or macropores, is that presented by [van Genuchten and Wierenga \(1976\)](#). Occam's razor is simply used to cut the soil's volumetric water content θ ($\text{m}^3 \text{m}^{-3}$) in two. This binary assumption does not account for changes in the leaching processes as a function of the soil's degree of unsaturation (cf. [Fig. 4](#)), but it does better represent the preferential leaching processes that prevail in field soils. The great simplicity of this model is that one fraction of the soil's water, θ_{im} ($\text{m}^3 \text{m}^{-3}$), is considered, for leaching purposes, to be effectively stagnant. The complementary portion of the soil's water, θ_m ($\text{m}^3 \text{m}^{-3}$), is deemed to be mobile, as might be imagined of that water which is between the aggregates, or flowing in the macropores.

A two-domain approximation for the leaching of a conservative solute under steady flow at pore-water velocity v_m , (mm s^{-1}), was given by [van Genuchten and Wierenga \(1976\)](#) as.

$$\theta_m \frac{\partial C_m}{\partial t} + \theta_{im} \frac{\partial C_{im}}{\partial t} = \theta_m D_s \frac{\partial^2 C_m}{\partial z^2} - v_m \theta_m \frac{\partial C_m}{\partial z}, \quad (7)$$

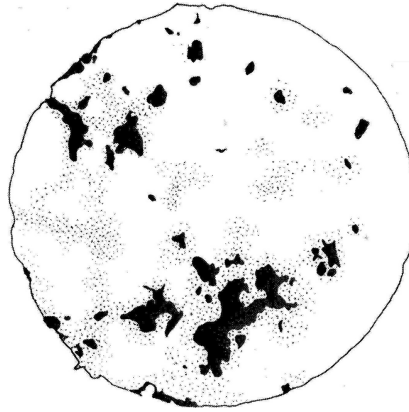


Fig. 4 Contrasting patterns of the leaching of two dyes through a 120 mm diameter core of undisturbed soil. The dark area was stained by a dye tracer contained in saturated flow from a free-water pond at the surface ($h_0 = 0$), and the stippled area was from dye leached by unsaturated water applied at the slight negative head of $h_0 = -20$ mm. From Scotter DR and Kanchanasut P (1981) Anion movement in a soil under pasture. *Australian Journal of Soil Research* 19:299–307.

where C_m and C_{im} are the concentrations of inert chemical in the mobile and immobile domains, and $v_m = q_w/\theta_m$. It is assumed that molecular diffusion in the direction of leaching in the mobile domain is negligible, and that the only mechanism to transfer solute between domains is, for analytical convenience, limited to a diffusion-like process that can be described by.

$$\theta_{im} \frac{\partial C_{im}}{\partial t} = \alpha (C_m - C_{im}) \quad (8)$$

where α (s^{-1}) is a first-order mass transfer coefficient.

To measure θ_{im} independently, Clothier et al. (1992) proposed using a tracer-filled tension disk-permeameter to allow leaching into the soil of a solution at some concentration C_m . Ignoring any effects that might accrue from interdomain diffusion (Eq. 8), the flux concentration in the mobile domain in the soil just under the disk, would be C_m . Thus by removing soil samples from under the disk, and determining the average resident concentration of solute C^* , the mobile fraction could be found using the partitioning formula

$$\theta C^* = \theta_m C_m + \theta_{im} C_{im}. \quad (9)$$

If a tracer not initially present in the soil were used, and if diffusive exchange (Eq. 8) were considered sufficiently weak, then C_{im} should remain essentially zero over the time scale of the experiment. So,

$$\theta_m = \theta \left(\frac{C^*}{C_m} \right), \quad (10)$$

via which θ_m can be obtained independently from measurements immediately under the disk of just C^* and θ . Their results for a silt loam suggested that for leaching, θ_m/θ is about 0.7–0.75 at a pressure head of $h_o = -40$ mm. Only about three-quarters of the soil's water-filled porosity appeared to be actively involved in leaching. This qualitatively corroborates the cross-sectional staining areas in Fig. 4, where it could be presumed that about 50% of the area were pore-space, and of that only about 75% was stained by dye.

Subsequently, Jaynes and Horton (1997) added a novel aspect to this technique. Through the use of the sequential invasion of i inert tracers into the same soil under the permeameter, they showed how differences in the respective resident concentrations of C_i^* would reflect the rate of diffusion, α , into the immobile domain. The resident concentration in the soil of any tracer i would reflect the diffusion that had taken place during the time t_i it had been present in the mobile domain. From a regression of the measurements using

$$\ln \left[1 - \frac{C_i^*}{C_{m,i}} \right] = \frac{-\alpha}{\theta_{im}} t_i + \ln \left[\frac{\theta_{im}}{\theta} \right], \quad (11)$$

they could deduce both θ_m and α . From many experiments on various soils, they found α to range between about 5×10^{-7} and $5 \times 10^{-6} s^{-1}$.

For aggregated soils, Jarvis (1994) presented a numerical model, the aptly named MACRO, which allows convection between the inter-aggregate mobile zone and the intra-aggregate pores. This convection is considered gravity-free, and is simply driven by water content differences between the respective domains. Solute transfer between these domains is allowed both by diffusion, as well as by direct convection with the transferring water. By simulation, Jarvis (1994) concluded that "... at short times, macropore flow increases leaching of non-reactive solute. At longer times, leaching is reduced by macropore flow [and it is] critical to account for the effects of macropore flow in clay soils on the leaching of sorbing, and degrading compounds."

Field measurement devices

Better management is needed as to how we use nitrogen and phosphorus, as well as manage salt, to grow the crops required to feed the world's burgeoning population within the safe operating-space for planet Earth (Fig. 1). The management thinker Peter Drucker stated that "you can't manage what you can't measure." Therefore there is an imperative to measure leaching processes so that we can improve land-management practices in relation to the biochemical flows of nitrogen and phosphorus (Steffen et al., 2015), salts, and pesticides.

Leaching, in terms of kg of nutrients, active ingredients, or salt per hectare of land per unit time, is the product of the drainage of water, in m of depth equivalent ($\text{m}^3 \text{m}^{-2}$), times the concentration, in mg L^{-1} , of the nutrients, pesticides, or salt in the drainage water.

By encapsulating a representative volume of soil in a sealed container it is possible to monitor drainage, and to measure the concentration in the leachate leaving the base of the enclosed soil. This set-up is known as a lysimeter. Lysimeters enable quantification of the complete mass balance. Lysimeters come in a variety of forms: repacked soil, or undisturbed cores; free-draining base, or controlled tension drainage; weighable, or non-weighing; shallow ones for grasses, or deeper ones for trees. A Special Section of the Vadose Zone Journal (Pütz et al., 2018) highlighted the roles that lysimeters can play in quantifying evaporation and evapotranspiration, chemical leaching and recharge, plus modelling and process evaluation. Pütz et al. (2018) concluded that lysimeters bridge the gap between laboratory studies and field measurements. Lysimeters are expensive in terms of capital costs and time requirements, and so they are rarely used for monitoring of leaching across landscapes.

A cheaper and simpler way to infer leaching losses is to sample the soil solution through using suction cups. Suction cups are constructed out of a porous medium such as ceramic, or a fritted material, that when inserted into the soil allows ingress of the soil solution into a reservoir when a suction is applied to the cup. The chemical concentration in the aliquot withdrawn from the soil solution is presumed to be the flux concentration of the solute that would be leaching. Furthermore, to calculate the areal loading of the leaching of solutes, there needs to be a contemporaneous measure of the drainage rate. This can be done via Darcy's Law, or using a water balance estimate, or predictions through transport modelling. Skaggs et al. (2002) noted that the "... solute concentrations measured with porous cup solution samplers are known to be affected by the magnitude, method and duration of the applied suction." Indeed, the chemical concentration within the suction cup is likely to be an admixture of the mobile flux-concentration, and the immobile resident-concentration of solute in the soil solution. There can also be other practical issues. Duwig et al. (2000) used suction cups to infer nitrate leaching through an uplifted coral atoll in the South Pacific. The soil there is very permeable, and this created measurement difficulties for there was a narrow time-window during the drainage cycle when the suction cup could actually withdraw any aliquots from the soil solution. In a study of nitrate leaching under irrigated sugarbeet in Montana, USA, Jabro et al. (2016) observed high variability between suction-cup replicates in the measured nitrate-N concentrations. They concluded that suction-cup samplers are not recommended in heterogeneous soils that exhibit preferential macroporous flow-patterns.

A cheaper alternative to enclosed lysimeters are the so-called pan lysimeters where a collection receptacle is inserted into the soil profile and water is collected into a reservoir through gravity (Richards, 1950). These devices are also known as 'zero tension' lysimeters, because water enters the collection container as free drainage-water. This free-water sampling of the draining soil solution by pan lysimeters means that measured flux-concentration of chemicals in the reservoir might not be the true flux-concentration of solutes moving through the soil under both gravitational and capillary forces.

Earlier, Ivie and Richards (1937) had conceived of the 'equilibrium tension' lysimeter for field use, whereby the drainage water could be drawn through a porous plate or membrane under a controlled vacuum. This has led to the more-recent development of the water flux-meter, a passive capillary-wick lysimeter, which Gee et al. (2009) described as "... a compromise between the complications and expense of equilibrium tension lysimeters and the simplicity of the less accurate pan lysimeters." Gee et al. (2009) provided a comprehensive and critical review of the efficacy and performance of the various forms of the passive-wick drainage fluxmeter (DFM).

Green et al. (2010) described their version of a DFM (Fig. 5). The DFM comprises a convergence ring to align the flow vertically, under which is a funnel containing a mix of sand and diatomaceous earth within which the untangled threads of a fiberglass wick are distributed. The wick length is 600 mm and this feeds into a tipping spoon to record the drainage hydrograph, before the soil solution is then stored underneath in a reservoir, from which it can be pumped to the surface using a vacuum pump. The leachate concentration in the drainage solution can then be measured.

Nitrogen leaching

Green et al. (2010) described the drainage and nitrogen (N) leaching results they obtained from an array of 45 manually sampled DFMs installed on a dairy farm near Lake Taupo in the center of the North Island of New Zealand. The soil is the free-draining Taupo sand, an Immature Orthic Pumice Soil. The pasture was rotationally grazed at about 3 cows per hectare, and received 4 split applications of urea each at $45 \text{ kg-N ha}^{-1} \text{y}^{-1}$. As expected for this winter-wet climate, most of the drainage occurred during the middle of the year (Fig. 6A). The average leaching loss on nitrate was around $40 \text{ kg ha}^{-1} \text{y}^{-1}$ over the 2 years (Fig. 6B). The variability in the individual nitrate measurements at any sampling time was large, as would be expected for pasture grazed by cows due

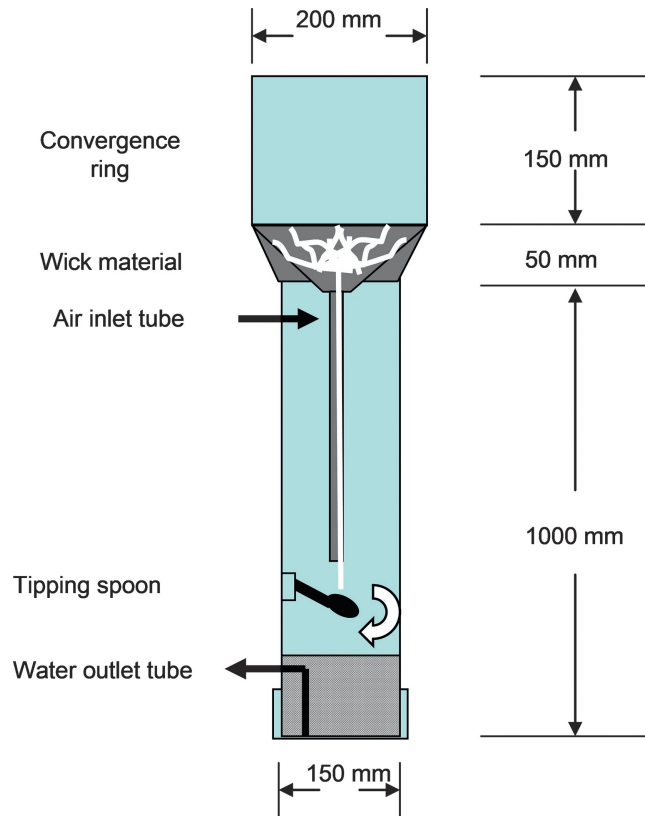


Fig. 5 A schematic of a passive tension fluxmeter. The fluxmeter is buried at the base of the rootzone, and drainage water enters the convergence rings by gravity and in response to capillary tension as a result of the 60 cm long fiberglass wick. The water then drips out via a tipping spoon that records the flow hydrograph, and it is then stored in the reservoir, awaiting removal of the leachate by a vacuum pump through the outlet tube that reaches up to the surface. Adapted from Green S, Deurer M, Clothier B, Andrews S, Cauldwell K, Roberts A, Wellwood M, and Thomas P (2010) Water and nitrogen movement under agricultural and horticultural land. Proceedings of the Workshop “Farming’s Future: Minimising footprints and maximising margins”. Fertiliser & Lime Research Centre, Massey University, 10–11 February 2010. [Farmed Landscapes Research Centre Paperlist10. (massey.ac.nz)].

the complex spatial patterns and timings of urinations and the depositions of dung. This highlights the value of low-cost devices like tension DFMs that mimic leaching processes at the local scale, yet enable sufficient replications to obtain representative data for the larger field scale.

Green et al. (2010) used the SPASMO (Soil Plant Atmosphere System Model) of Green et al. (2008) to simulate both the drainage and nitrate leaching through this free-draining pumice soil under this dairy pasture. These predictions are shown in Fig. 6 and well represent the N leaching processes.

Salt leaching

About one third of world’s land surface, some 50 million km², is either arid or semi-arid. These lands receive less than 400 mm of rain. Twenty percent of the world’s population, about 1.2 billion people, live in such arid, or semi-arid regions. Agricultural production in these regions is affected by soil salinity as the irrigation water used to grow the crops generally contains a mixture of naturally occurring salts (Ayers and Westcot, 1994). Inevitably then salts are added to the soil with each irrigation, whereas the plant roots only extract fresh water by osmosis. The extent to which salts build up in the soil depends on irrigation water salinity (EC_w , dS m⁻¹), irrigation management, and the adequacy of drainage to provide the leaching of salts (Wichelns, 2015). Leaching of salts can be achieved by applying an excess of irrigation water so that a fraction of the applied water drains through the rootzone carrying with it a portion of the accumulated salts. This excess water aliquot is known as the leaching fraction, LF , and is defined as

$$LF = \frac{\text{Depth of water leached below the rootzone}}{\text{Depth of irrigation water applied}}. \quad (12)$$

The goal of using a sufficient LF is to ensure that the soil-water salinity in the rootzone (EC_{sw}) does not reach a level above which losses in crop yield will occur. In the hyper-arid United Arab Emirates (UAE), Al Muaini et al. (2019) used an LF of 0.33 to irrigate date palms with water having an EC_w of 5 dS m⁻¹. Their measurements of EC_{sw} in the rootzone showed the diurnal cycling in

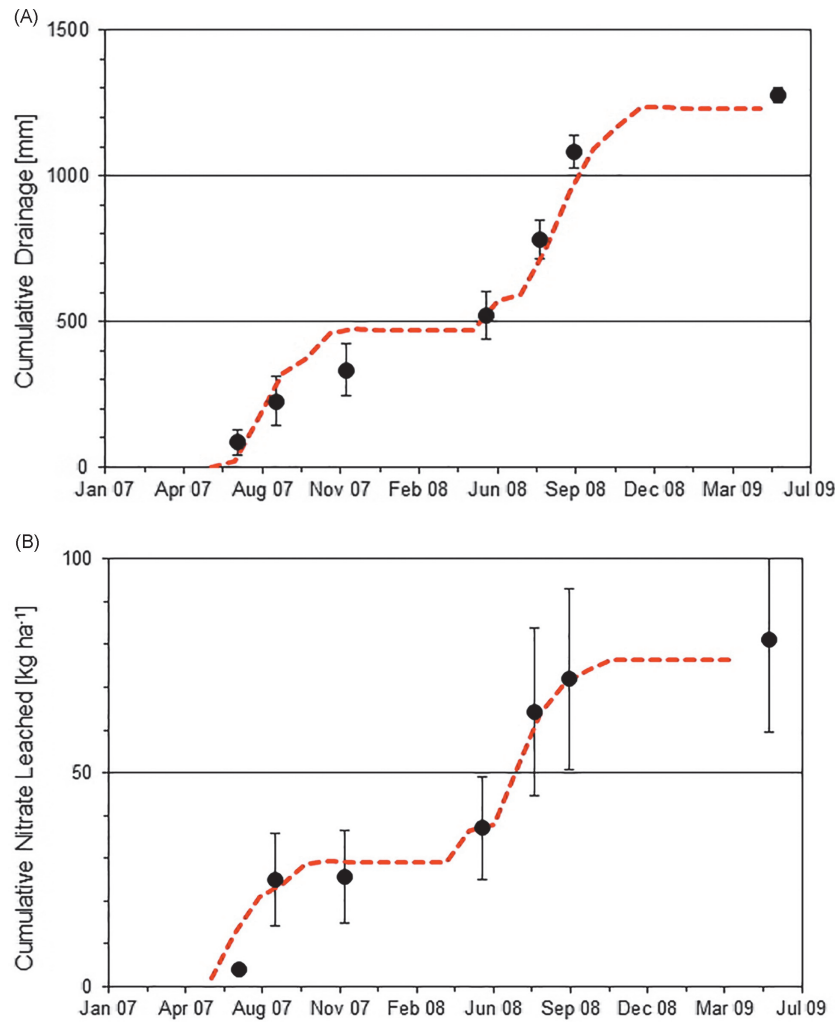


Fig. 6 Top (A): The time series of drainage from an array of 45 tension drainage fluxmeters (DFM) under rotationally grazed pasture on a dairy farm near Lake Taupo, New Zealand. Bottom (B): The time series of cumulative nitrate leaching as measured by the array of 45 DFMs. The red lines are modelled results using the Soil Plant Atmosphere System Model (SPASMO) of Green *et al.* (2008) taking into account the patches of dung and urine from the grazing animals. Adapted from Green S, Deurer M, Clothier B, Andrews S, Cauldwell K, Roberts A, Wellwood M, and Thomas P (2010) Water and nitrogen movement under agricultural and horticultural land. Proceedings of the Workshop “Farming’s Future: Minimising footprints and maximising margins”. Fertiliser & Lime Research Centre, Massey University, 10–11 February 2010. [Farmed Landscapes Research Centre Paperlist10. (massey.ac.nz)].

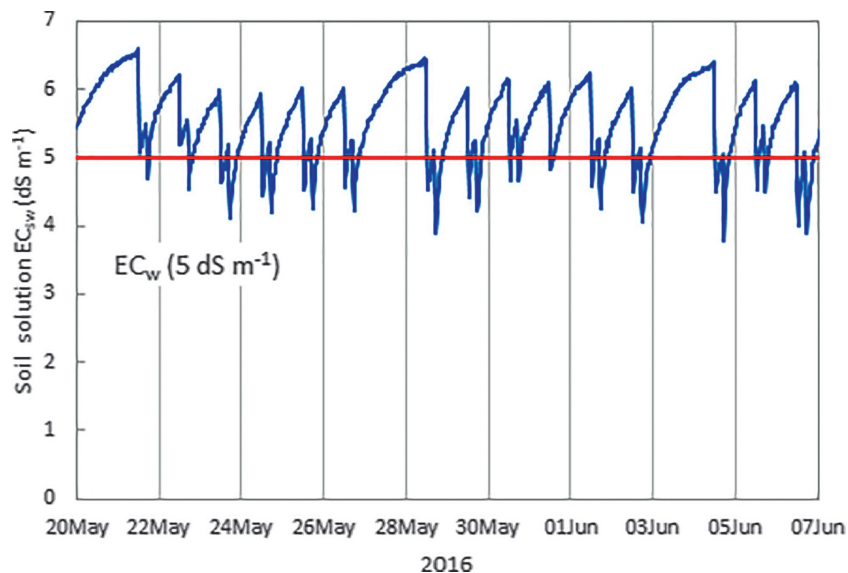


Fig. 7 The soil-water electrical conductivity EC_{sw} inferred from the measured water content (θ) and bulk soil electrical conductivity EC_b under a date palm irrigated with $5\ dS\ m^{-1}$ water (horizontal line). The irrigation regime, I , was chosen to provide a leaching fraction LF of 0.33. Two daily irrigations were carried out, one in the morning, and the other in the afternoon. There were no irrigations on Fridays (21 May, 28 May, and 4 June). Modified from Al Muaini A, Sallam O, Green SR, Kennedy L, Kemp P and Clothier BE (2019) The blue and grey water footprints of date production in the saline and hyper-arid deserts of the United Arab Emirates. *Irrigation Science* doi:10.1007/s00271-019-00642-6.

soil-water salinity with the build-up in EC_{sw} following the twice daily irrigations, and especially on Fridays when there were no irrigation (Fig. 7). This LF of one third was successful in leaching out the excess salts left after root-water extraction by the date palm roots, as there was no net build-up over time in the rootzone soil-water salinity.

Whereas this LF of one third was successful in maintaining a good rootzone EC_{sw} , the leaching of the excess salts in the drainage LF could pose a risk to the underlying aquifer. Ayers and Westcot (1994) showed that a simple mass-balance calculation estimates the salinity of the drainage water, EC_{dw} , to be

$$EC_{dw} = \frac{EC_w}{LF} \quad (13)$$

Thus with an LF at one third, the leachate salt-salinity in this case would be threefold that of the irrigation water, namely 15 dS m^{-1} . Over time, this higher salinity drainage water would likely find its way to the aquifer, and thereby increase the salinity of the underlying groundwater reserves.

Small, autonomous desalination plants are being increasingly used by farmers in arid regions to 'freshen-up' saline groundwater for use in irrigating high-value crops. This would seem to obviate the LF concerns raised above. Al Muaini et al. (2019) found the benefit cost ratio to be 1.4 if a small desalination unit was used to turn 15 dS m^{-1} groundwater into 5 dS m^{-1} water for irrigating date palms. This reduces the salt loading on the groundwater under crops now irrigated with fresher groundwater. However, there is then the problem of disposing of the reject brine from the desalination unit. Lyra et al. (2019) explored the feasibility, benefits, and risks of using the reject brine from desalination plants in an integrated land-based aquaculture system in conjunction with irrigation of a halophytic crop such as *Salicornia*. Current work seeks to understand better the long-term environment impact of salt leaching from such an Integrated Aquaculture Agriculture System (IAAS), and to develop improvements in food security and to enhance economic sustainability.

Modelling and risk assessment

Jarvis (1994) showed that despite the complicated mathematical forms of the coupled flow and transport Eqs. (3) and (4), and even (7) with (8), modern computers and efficient software schemes are easily able to solve them to predict the impact of land management on leaching processes. The results from models can then be used in risk-assessment frameworks to help us better understand the functioning of environmental systems, and to guide decisions for sustainable land management.

Initially, we provide some results from simplified modelling of the drainage processes under short pasture that is grazed, and a pine forest. A simple, lumped-parameter model (Green et al., 1999; Clothier et al., 2000) provides us with insight into the processes of leaching in terms of the controls on drainage quantity (D), and the contrast between the chemical concentration in the leachate (C), and total leachate load (M).

The simulations considered pasture and pines growing on soil typical of the Nelson region of New Zealand. Actual weather data from 1972 to 1998 were used. The maximum rooting depth, z_R , of 0.5 m was assumed for pasture, whereas the pine trees were considered able to root effectively to any prescribed depth. Simulations using a lumped parameter model were carried out for $z_R = 0.5, 1, 1.5, 2, 2.5$, and 3 m (Fig. 8).

When pasture and pines are grown on soils of a similar shallow depth of 0.5 m, the ET for the pines is lower due to sparseness of the canopy in the early stages of tree growth, and again following the thinnings. Nonetheless, due to interception losses by the trees reducing the effective rainfall, there is still a lower drainage quantity (D) under pines for such shallow soils. As the soil depth increases, the rise in ET restricts the drainage quantity (D) under the pines. For a shallow soil of $z_R = 0.5 \text{ m}$, the annual leachate load under pines is about 195 mm, or 20% of RF . When the soil is deep, say $z_R = 3 \text{ m}$, $D = 23 \text{ mm}$, or just 2.5% of RF . Thus the amount of water draining through soil is critically dependent on the nature of the vegetation cover and the soil's properties, and in particular the depth of effective rooting z_R .

Pasture production and pine tree growth were modelled through a nitrogen-balance scheme. Consequently, we could therefore also use this model to assess the risk posed to groundwater quality by the leaching of nitrate from the rootzone of these crops. We choose to present here, the average-annual nitrate-leaching concentrations, C , and the total annual loadings, M , for the pasture ($z_R = 0.5 \text{ m}$), and the pines with $z_R = 2 \text{ m}$ (Fig. 9).

There are two distinctive characteristics – the difference between the two land uses, and the skewed and non-Gaussian nature of the CPF's. The average concentration (probability 0.5) is much lower for pines (2.3 mg L^{-1}), than for pasture (12.1 mg L^{-1}), since a lot more of the nitrogen ends up stored in the plant tissue. Through the repeated grazings, and the additions of dung and urine, there are greater inputs of N into the pastoral system. The CPF for the trees is the most non-Gaussian, and this is due to the infrequent high loadings of nitrate that occur when, occasionally, a wet-year follows a dry year, and there is a surge in nitrate leaching. The simple, lumped-parameter model provides us with the amount of drainage (D) so that this can be multiplied by the nitrate concentration (C) to obtain a loading of nitrate of the receiving waters (M). The model results in Fig. 9 show that the loading on groundwater from grazed pasture is predicted to be $41 \text{ kg-N ha}^{-1} \text{ y}^{-1}$. However for the pine trees growing on this 2 m deep soil, the low concentrations of nitrate combined with the low amount of drainage, mean that there is simulated to be a very low leachate loading of just $0.6 \text{ kg-N ha}^{-1} \text{ y}^{-1}$. So whereas pine trees are predicted to have a negative impact on the quantity of drainage (D) reaching receiving water bodies, they are considered, by virtue of a lower total loading (M), to have a beneficial role in maintaining the quality of

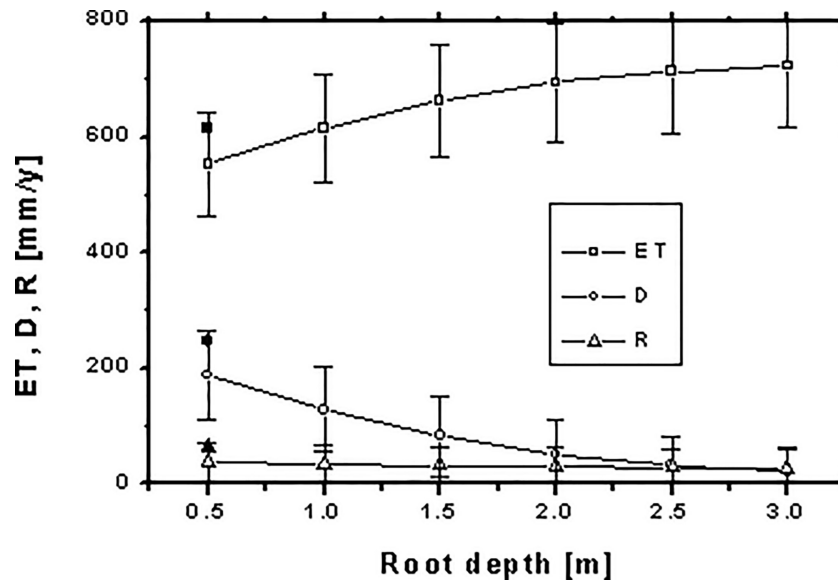


Fig. 8 The influence of the depth of the rootzone on the amount of drainage (D) 'lost' by pine trees (open symbols) in comparison to pasture (closed symbols). Pasture is assumed to root to a maximum depth of z_R of 0.5 m, whereas the pines could extend their roots beyond 3.0 m, if soil depth permits. The mean annual rainfall over the 27 years was 974 mm. Reproduced with permission from Clothier BE et al. *Modelling Rootzone Processes: A Tool for Risk Assessment*. Water 2000. NZWWA.

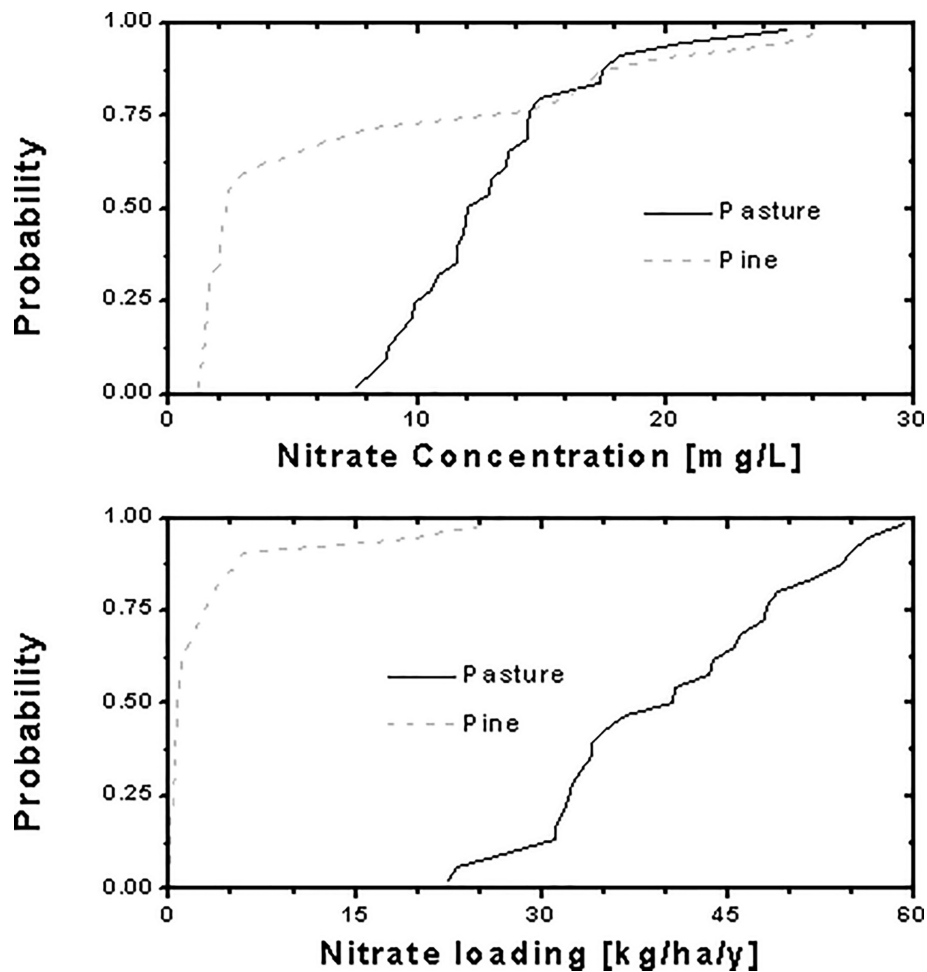


Fig. 9 The predicted cumulative probability functions (CPF) of: (A) the concentration of nitrate in the leachate under pasture, and that under pine trees (soil depth 2 m, but the rooting depth for pasture $z_R = 0.5$ m, and pines $z_R = 2.0$ m). This CPF relates to the annualized value that was derived by running the model in daily time steps over 27 years, and summing the results to form annual totals, and (B) the total loading of nitrate on the receiving water bodies. Reproduced with permission from Clothier BE et al. *Modelling Rootzone Processes: A Tool for Risk Assessment*. Water 2000. NZWWA.

leachates moving to groundwaters. The interplay between land management and soil characteristics determine the quantity of drainage and the water quality of leaching.

Pesticide leaching

We continue with the use of the more-sophisticated modelling framework of SPASMO (Green et al., 2002) to predict the risks posed by the leaching of the active ingredient (AI) of three pesticides applied, in a simulation, to a grape vineyard on a Twyford silt loam located above an unconfined aquifer in Hawke's Bay, New Zealand. This framework can account, in a simple way, for the mobile-immobile partitioning of fast and slow-flow pathways of leaching through soil. The SPASMO simulations considered actively growing grape vines, and the grassed inter-row. The calculations were run on daily time-steps using a 27-year sequence of weather data (1972–1999) recorded from the local airport at Napier. The soil's physical and hydraulic properties were taken from the New Zealand Soils Database. Three herbicides, simazine, linuron and amitrole, were 'applied' at the label rates on the 15th September each year. This modelling of leaching predicts if, and how often, the soil-solution concentration at 3 m might exceed the Maximum Allowable Value (MAV) of 2 ppb of the AI.

Simazine, is quite mobile because of its low exchange with the soil's organic carbon (Eq. 6). Also it is persistent with a long degradation half-life of $t_{1/2} = 50$ days. Linuron is less mobile because of greater exchange with the soil's organic matter, and it is less persistent with $t_{1/2} = 33$ days. Amitrole is also less mobile because of low exchange, but it is much less persistent with $t_{1/2} = 3$ days. Therefore, concerning the likelihood of leaching to groundwater, we would expect the risks to be ranked simazine > linuron > amitrole. This expectation is borne out in the simulated leaching results, presented for a depth z of 3 m (Fig. 10). The average simazine concentration in the soil solution leaching through Twyford silt loam, over the 30-year period, is predicted to be just $0.03 \mu\text{g l}^{-1}$, although 10% of the time it will exceed $0.12 \mu\text{g l}^{-1}$, or 6% of MAV. Simazine should be used with greater caution, for it poses a greater leaching risk than either of the other two pesticides.

The risk of pesticide leaching also depends on the depth to groundwater. Fig. 11 shows the model results for linuron, as a function of depth. Here the data are presented as a cumulative probability function of exceedance. During the time taken to leach deeper into the soil profile, there will be degradation of the pesticide. For linuron, this means that there is predicted to be, on average (50% exceedance), about an order of magnitude drop in the concentration of active ingredient with every meter of soil depth. Even in the worst case of $z = 1$ m, the concentration of linuron will still be about an order of magnitude less than the MAV. Nonetheless, this analysis shows that shallow unconfined groundwaters are at the greatest risk of contamination from the leaching of surface-applied chemicals.

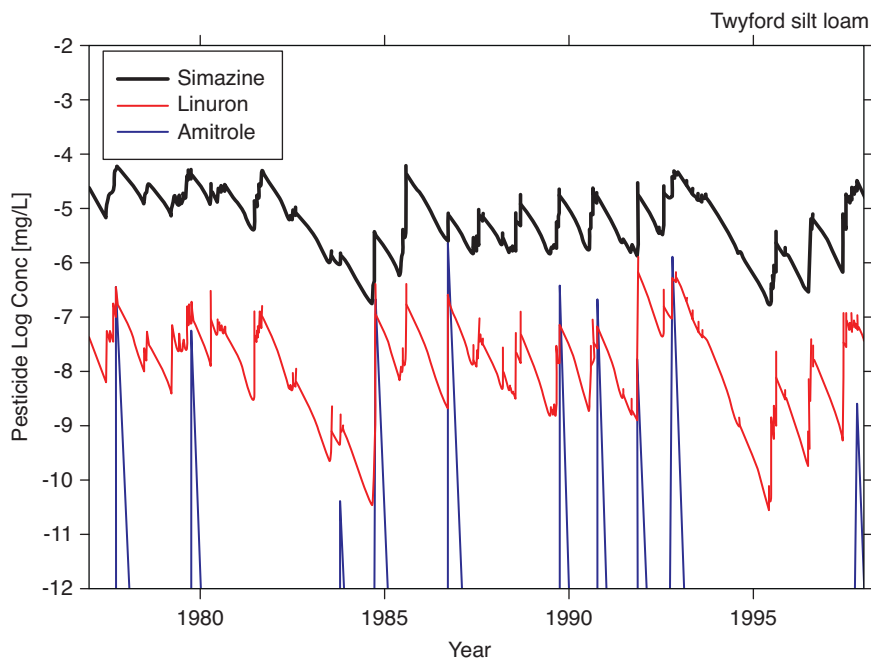


Fig. 10 Simulated herbicide concentrations in the soil solution at 3 m depth under grapevines growing on a Twyford silt loam. Each year, over 30 years, the herbicides were considered to be applied at the label rate, to 100% of the soil surface, in mid-September.

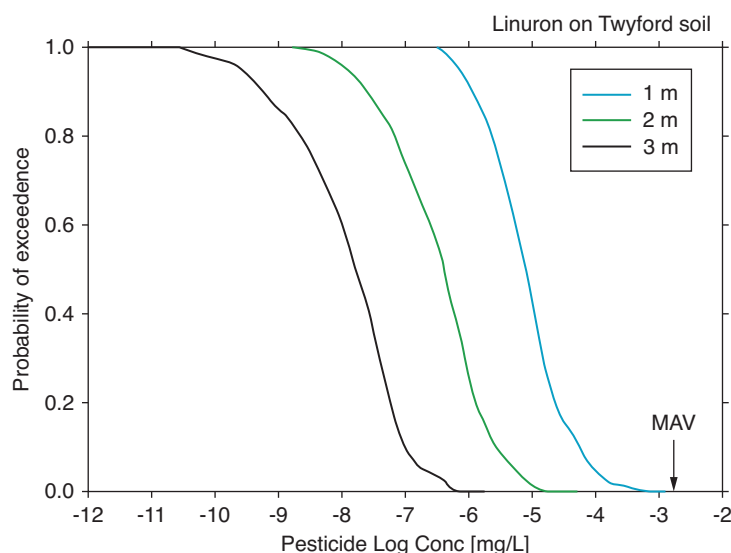


Fig. 11 The predicted probability distribution for the concentration of linuron in leachate at 3 depths under grapevines growing on a Twyford silt loam. Linuron was considered to be applied once each year, in mid-September, at the label rate of 1.0 kg ha^{-1} of active ingredient.

Upscaling

Whereas we are obtaining better measurements and improved modelling of local scale leaching processes, there is a challenge to upscale these to the field, farm and landscape scale. The inability to monitor or model at the larger spatial scale poses a problem for resource managers and regulators to develop practices and policies to protect the quality of both surface and subterranean water resources across whole catchments.

The first upscaling jump is to link rootzone leaching processes to those throughout the rest of the critical zone of the underlying vadose column of soil (Glaser et al., 2019), and the upper parts of the saturated region of the deeper aquifers (Banwart and Sparks, 2017). Gurevich et al. (2021) investigated the relationships between nitrogen measurements in the soil column, the groundwater monitoring-well scale (GW), plus the field and farm scale. They monitored in the rootzone (RZ) down to 3 m, the deep vadose zone (DVZ) down to 7 m, as well as the upper saturated zone (SZ) down to 14 m under a commercial almond orchard. Within the coarse-textured parts of the orchard, Gurevich et al. (2021) found that the measured RZ nitrate concentrations closely correlated with the GW concentrations underneath, indicating the significance of far-reaching and preferential flow paths. Overall, Gurevich et al. (2021) concluded that traditional soil-science RZ monitoring, through aggregated samples, could be extended to the DVZ, SZ and GW. The rider was, however, that about a dozen monitoring sites across the 56 ha orchard would be needed to assess GW nitrate impacts from land management practices on the surface.

Further spatial upscaling from the field and farm-scale to the whole catchment, with its complex mixing between surface and groundwaters, remains a major challenge, both to monitor and to model. The magnitude of this challenge is because, as Alexander et al. (2002) noted, there is a need to account "... for the variety of sources and removal processes and the nature of the interactions." They added that "... large quantities of nutrients are removed on the landscape and in the waterways, but the rates vary widely in response to such factors as climate, topography, soil, vegetation, and the physical and hydraulic properties of streams and reservoirs." Rivett et al. (2008) found that denitrification losses of nitrogen were more probable in confined aquifers and in the near-river environments of the riparian strip and the hyporheic zones. They listed eight areas of research where improved understanding would be required to enable mechanistic upscaling to the catchment scale.

Such complexity mitigates against the current usage of complex process-based or mechanistic models. Alexander et al. (2002) applied the SPARROW (Spatially Referenced Regression on Watershed Attributes) model of surface water-quality to estimate empirically the rates of nutrient delivery from both point and non-point sources to streams, rivers, and lakes. The catchment they studied was the 13,900 km² Waikato River Basin in Central New Zealand. There are 4 connected sub-catchments within this basin: Taupo, Waipa, Upper Waikato, and the Lower Waikato from which the Waikato River flows into the sea. The difference between the landscape yield of nutrients, and the watershed yield, is the attenuation. This represents the nutrients removed in streams and reservoirs. Across the four sub-catchments, Alexander et al. (2002) found nitrogen losses of 76% (Taupo), 42% (Upper Waikato), Waipa (39%) and 45% (Lower Waikato). The respective attenuations for phosphorus were found to be 89%, 59%, 53% and 51%. The Taupo sub-catchment is unusual because it is the least developed with forested land and Lake Taupo itself comprising over 50% of the catchment's area. These results highlight the complexity and the difficulty of upscaling leaching processes from the local scale, to the farm scale, right through to the catchment level.

Summary

The local leaching of chemicals through soil by percolating pore-water is the result of a complex of interconnected biophysical and biochemical processes. Firstly, biophysical mechanisms determine the quantity of water that quits the rootzone, destined for either groundwater or surface receiving waters. The prevailing weather, the nature of the vegetative cover, and the properties of the soil conspire to determine the amount of leachate. Secondly, the biochemical processes of exchange, production, and degradation control the concentration of chemical in the leaching solution.

It is imperative that better land-management practices be adopted to reduce the biochemical flows of nutrients, salts, and pesticides into receiving water bodies. Measurements and modelling will be critical in developing these new land-use practices. And soil and water monitoring across the landscape will be important for tracking progress to this goal of improving water-resource quality. At the local-scale, new measurement devices and mechanistic modelling are providing useful knowledge about leaching processes. However, there are challenges as to how these local measurements and predictions are upscaled to provide understanding at the field, farm, and catchment scale. New monitoring protocols and empirical models are helping in this quest.

References

- Al Muaini A, Sallam O, Green SR, Kennedy L, Kemp P, and Clothier BE (2019) The blue and grey water footprints of date production in the saline and hyper-arid deserts of the United Arab Emirates. *Irrigation Science*. <https://doi.org/10.1007/s00271-019-00642-6>.
- Alexander RB, Elliott AH, Shankar U, and McBride GB (2002) Estimating the sources and sinks of nutrients in the Waikato Basin. *Water Resources Research* 38(12): 1268–1280.
- Ayers RS and Westcot DW (1994) *Water Quality for Agriculture, Irrigation and Drainage Paper 29, rev. 1*. Rome: Food and Agriculture Organization of the United Nations.
- Banwart SA and Sparks DL (2017) Quantifying and managing soil functions in Earth's critical zone combining experimentation and mathematical modelling. *Advances in Agronomy* 142: 2–423. ISBN: 978-0-12-812222-8.
- Clothier BE and Green SR (1997) Roots: The big movers of water and chemical in soil. *Soil Science* 162: 534–543.
- Clothier BE, Kirkham MB, and MacLean JE (1992) *In situ* measurement of the effective transport volume for solute moving through soil. *Soil Science Society of America Journal* 56: 733–736.
- Clothier BE, Green SR, Roygard JFK, Vogeler I, and Duwig C (2000) Modelling rootzone processes: A tool for environmental risk assessment. In: *Proceedings Water 2000 Conference, NZ Water and Wastes Association, Auckland, 19–23 March 2000*.
- Duwig C, Becquer T, Vogeler I, Vauclin M, and Clothier BE (2000) Water dynamics and nutrient leaching through a cropped Ferralsol in the Loyalty Islands (New Caledonia). *Journal of Environmental Quality* 29: 1010–1018.
- Feddes RA, Kowalik PJ, Malinka KK, and Zradny H (1976) Simulation of field water uptake by plants using a soil water dependent root extraction function. *Journal of Hydrology* 31: 13–26.
- Gardner WR (1960) Dynamic aspects of water availability to plants. *Soil Science* 89: 63–73.
- Gee GW, Newman BD, Green SR, Meissner R, Rupp H, Zhang ZF, Keller JM, Waugh WJ, van der Velde M, and Salazar J (2009) Passive wick fluxmeters: Design considerations and field applications. *Water Resources Research* 45: W04420. <https://doi.org/10.1029/2008WR007088>.
- Glaser B, Jackisch C, Hopp L, and Klaus J (2019) How meaningful are plot-scale observations and simulations of preferential flow for catchment models. *Vadose Zone Journal* 18: 1–18. <https://doi.org/10.2136/vzj2018.08.0146>.
- Green SR, Clothier TMM, and Millar A (1999) Risk assessment of the irrigation requirements of field crops in a maritime climate. *Journal of Crop Production* 2: 353–377.
- Green SR, Clothier BE, Caspari HW, and Neal SM (2002) Rootzone processes, tree water use and the equitable allocation of water to olives. In: Smiles DE, et al. (ed.) *Heat and Mass Transfer in the Natural Environment, American Geophysical Union Monograph Monograph 129* 337–345.
- Green SR, Clothier BE, van den Dijssel C, Deurer M, and Davidson P (2008) Measurement and modelling the stress response of grapevines to soil-water deficits in their rootzones. Chapter 12. In: Ahuja L, et al. (ed.) *Modeling the Response of Crops to Limited Water: Recent Advances in Understanding and Modeling Water Stress Effects on Plant Growth Processes. Soil Science Society America Monograph* 357–386. Madison, WI: Soil Science Society of America.
- Green S, Deurer M, Clothier B, Andrews S, Cauldwell K, Roberts A, Wellwood M, and Thomas P (2010) Water and nitrogen movement under agricultural and horticultural land. In: *Proceedings of the Workshop "Farming's Future: Minimising footprints and maximising margins". Fertiliser & Lime Research Centre, Massey University, 10–11 February 2010*. Farmed Landscapes Research Centre Paperlist10. (massey.ac.nz).
- Gurevich H, Baram S, and Harter T (2021) Measuring nitrate leaching across the critical zone at the field to farm scale. *Vadose Zone Journal*, e20094. <https://doi.org/10.1002/vzj2.20094>.
- Ivie JO and Richards LA (1937) A meter for recording slow liquid flow. *Review Scientific Instrumentation* 8: 86–89. <https://doi.org/10.1063/1.1752247>.
- Jabro JD, Stevens WB, Iversen WM, Allen BL, and Sainju UM (2016) Suction cup samplers for estimating nitrate-nitrogen in soil water in irrigated Sugarbeet production. *Journal of Environmental Protection* 7: 1342–1354. <https://doi.org/10.4236/jep.2016.710117>.
- Jarvis NJ (1994) The MACRO Model (Version 3.1) - Technical description and sample simulations. In: Reports & Dissertations 19. Uppsala: Swedish University of Agricultural Science. 51 pp.
- Jaynes DB and Horton R (1997) Field parameterization of the mobile/immobile domain model. In: Selim M and Ma L (eds.) *Physical Nonequilibrium in Soils: Modeling and Application*. Chelsea, MI: Ann Arbor Press.
- Jury WA and Roth K (1990) *Transfer functions and solute movement through soil: Theory and applications*. Basel: Birkhauser, p. 226.
- Lyra DA, Al-Shihhi RMS, Nuqui R, Robertson SM, Christiansen A, Ramachandran S, Ismail S, and Al-Zaabi AM (2019) Multidisciplinary studies on a pilot coastal desert modular farm growing *Salicornia bigelovii* in United Arab Emirates. Chapter 16. In: Hasanuzzaman M, et al. (ed.) *Ecophysiology, Abiotic Stress Responses and Utilization of Halophytes*. Springer. https://doi.org/10.1007/978-981-13-3762-8_16.
- Pütz T, Fank J, and Flury M (2018) Lysimeters for vadose zone research. *Vadose Zone Journal* 17: 180035. <https://doi.org/10.2136/vzj2018.02.0035>.
- Richards LA (1950) Laws of soil moisture. *Eos, Transactions of the American Geophysical Union* 31(5): 750.
- Rivett MO, Buss SR, Morgan P, Smith JW, and Bemment CD (2008) Nitrate attenuation in groundwater: A review of biogeochemical controlling processes. *Water Research* 42(16): 4215–4232. <https://doi.org/10.1016/j.watres.2008.07.020>.
- Rockström J, Steffen W, Noone K, Persson Å, Chapin FS III, Lambin EF, Lenton TM, Scheffer M, Folke C, Schellnhuber HJ, Nykvist B, de Wit CA, Hughes T, van der Leeuw S, Rodhe H, Sörlin S, Snyder PK, Costanza R, Svedin U, Falkenmark M, Karlberg L, Corell RW, Fabry VJ, Hansen J, Walker B, Liverman D, Richardson K, Crutzen P, and Foley JA (2009) A safe operating space for humanity. *Nature* 461: 472–475. <https://doi.org/10.1038/461472a>.
- Scotter DR and Kanchanasut P (1981) Anion movement in a soil under pasture. *Australian Journal of Soil Research* 19: 299–307.
- Skaggs TH, Wilson GV, Shouse PJ, and Leij FJ (2002) 6.4 Solute transport: Experimental methods. In: Dane JH and Topp GC (eds.) *Methods of Soil Analysis*, <https://doi.org/10.2136/sssabookser5.4.c57>.

- Steffen W, Richardson K, Rockström J, Cornell SE, Fetzer I, Bennett EM, Biggs R, Carpenter SR, de Vries W, de Wit CA, Folke C, Gerten D, Heinke J, Mace GM, Persson LM, Ramanathan V, Reyers B, and Sörlin S (2015) Planetary boundaries: Guiding human development on a changing planet. *Science* 347: 6219. <https://doi.org/10.1126/science.1259855>.
- Van Genuchten MT and Wierenga PJ (1976) Mass transfer studies in sorbing porous media. I. Analytical solutions. *Soil Science Society of America Journal* 40: 473–480.
- Wada Y, van Beek LPH, and Bierkens MFP (2012) Nonsustainable groundwater sustaining irrigation: A global assessment. *Water Resources Research* 48: W00L06. <https://doi.org/10.1029/2011WR010562>.
- Wichelns D and Qadir M (2015) Achieving sustainable irrigation requires effective management of salt, soil salinity, and shallow groundwater. *Agricultural Water Management* 157: 31–38. <https://doi.org/10.1016/j.agwat.2014.08.016>.

Colloid transport

Yan Jin^a, Chongyang Shen^b, and Volha Lazouskaya^a, ^aUniversity of Delaware, ISE Laboratories, Newark, DE, United States;
^bChina Agricultural University, Haidian District, Beijing, China

© 2023 Elsevier Ltd. All rights reserved.

Introduction	337
Colloidal forces and interactions	337
DLVO forces	337
Van der Waals force/energy	337
Electrostatic or double layer force/energy	338
Non-DLVO forces	339
The DA and SEI technique	340
DLVO interaction energy/force profiles and maps	340
Processes governing colloid transport	343
Deposition	344
Release	344
Straining	345
Retention at AWI, contact line and film straining	345
Experimental approaches	346
Factors influencing colloid transport	348
Surface roughness and charge heterogeneity	348
Colloid shape	349
Modeling colloid transport	349
Equilibrium adsorption and kinetic deposition	349
Blocking and ripening	349
Transport models	350
Upscaling	350
Conclusion	350
References	351

Abstract

Colloid transport in porous media plays important roles in many agricultural, environmental, and engineering processes. Colloids vary in their source, composition, shape and structure but they all possess large specific surface areas and have high reactivity. Deposition or attachment, release or detachment, and straining are the processes controlling colloid transport in saturated media; attachment to air-water interfaces and the contact line, and film straining are additional processes affecting colloid transport through unsaturated media. The interactions of colloids with interfaces can be described by classical Derjaguin-Landau-Verwey-Overbeek (DLVO) theory, which considers van der Waals and electrical double layer forces and by extended DLVO, which includes non-DLVO forces (e.g., hydration, steric and hydrophobic). Current approaches to modeling colloid transport build on the classical colloid filtration theory. More complete mechanistic understanding of colloid retention and release processes and experimental validation will improve model development and our ability to manage both the risks and beneficial applications of environmental colloids.

Key points

- Colloid transport in porous media is relevant to multiple agricultural, environmental and engineering processes
- Deposition, release, and straining are the main processes controlling colloid transport in saturated porous media; retention at the air-water interface and contact line and film straining are additional processes affecting colloid transport in unsaturated media
- The Derjaguin-Landau-Verwey-Overbeek (DLVO) theory and extended DLVO theory are used for describing colloidal interactions
- Capillary force governs the interaction of colloids at the air-water interface and contact line
- Colloid filtration theory (CFT) is the theoretical basis for quantifying colloid transport in porous media under favorable conditions
- Surface roughness and charge heterogeneity play a critical role in colloid deposition and release under unfavorable conditions
- Mechanistic and continuum models are used to simulate colloid transport in porous media
- Upscaling from the pore-scale mechanistic calculations to predict colloid transport at continuum scales in natural media remains a challenge

Introduction

Colloids are particles in the size range of 1 nm to 10 μm . They are ubiquitous in subsurface porous media and can be generated in situ or introduced from external sources. Colloidal particles found in porous media like soil include inorganic (e.g., soil clay particles, oxyhydroxides of Fe and Al, carbonates, silica), organic (e.g., soil organic materials), engineered (e.g., manufactured nanoparticles), biological (e.g., bacteria and viruses) (McCarthy and McKay, 2004), and micro- and nanoplastics (Yu and Flury, 2021). Despite the differences in their source, composition and structure, colloids have two important properties in common: (1) they have very small sizes that give rise to large specific surface areas and high reactivity and (2) they can remain in suspension for long periods of time and therefore can potentially travel long distances in subsurface media. These properties are the key parameters controlling colloids' fate and transport in porous media (DeNovio et al., 2004).

Colloid transport in porous media has attracted intense research attention due to its relevance to agricultural and environmental concerns and important engineering processes such as membrane filtration and petroleum reservoir production. Early interests in colloid mobilization and transport in soil dates back to the 1930s, due to the important roles that colloids play in soil genesis and erosion and their influence on hydraulic conductivity in relation to irrigation management (Kretzschmar and Schafer, 2005). Classic colloid filtration theory (CFT) was developed in the 1970s for predicting the rate of particle removal in deep-bed filtration water treatment systems (Yao et al., 1971), which inspired more than four decades of active research on colloid retention and transport in porous media. Colloid-facilitated contaminant transport represents another intensely focused area of colloid research since the late 1980s as mobile colloids can serve as carriers of strongly sorbing substances (e.g., heavy metals, radionuclides, hydrocarbons, pesticides, nutrients) (McCarthy and Zachara, 1989). Research on the transport of biocolloids (e.g., bacteria and viruses), which is of significant relevance to various environmental and public health issues, including protection of drinking water supplies, in situ bioremediation, and riverbank filtration (Jin and Flury, 2002), attracted intense interests in the 1990s. Study of nanoparticles (NP), having one or more external dimensions in the range of 1–100 nm, represents a more recent focus in colloid research. Although NPs occur naturally and are ubiquitous in soil and water, engineered nanoparticles (ENPs) have been the center of most research to date because of their wide and varied use in many industrial and environmental applications and their potential impact on human health and ecological systems. There is an emerging interest in the role that soil colloids play in complexing and stabilizing soil organic carbon (OC) as well as the fate and transport of colloidal OC, especially in the context of global carbon cycling (Yan et al., 2018; Afsar et al., 2020).

Three major processes control the extent of colloid transport in porous media: deposition or attachment, release or detachment, and straining. Deposition involves transport from bulk fluid to the vicinity of a collector (e.g., soil mineral grain) surface and subsequent attachment on the surface. Straining describes the process of colloids being trapped in down-gradient pore throats that are too narrow to permit passage. Attachment and detachment are controlled by interaction forces (or energies) between colloids and collector surfaces (Ryan and Elimelech, 1996; Shen et al., 2020). The interaction forces/energies are quantitatively described by the standard and extended Derjaguin-Landau-Verwey-Overbeek (DLVO) theory, which are described in detail below.

Colloidal forces and interactions

DLVO forces

DLVO interaction force/energy is the sum of van der Waals (vdW) and electrostatic forces/energies and is expressed as a function of separation distance between interacting surfaces.

Van der Waals force/energy

The vdW force originates from quantum mechanics, acts between atoms, molecules and surfaces, and is composed of induction force, orientation force, and dispersion force. The Hamaker technique is the most commonly used to determine the vdW force/energy between two surfaces (Hamaker, 1937). This approach calculates the vdW energy (U_{vdW}) as sum of the interactions between colloids 1 and 2 in a vacuum as

$$U_{\text{vdW}} = \int_{V_1} \int_{V_2} u(h) q_1 q_2 dV_1 dV_2 \quad (1)$$

where q_1 and q_2 are molecules per unit volume, V_1 and V_2 are volumes of colloids 1 and 2, respectively, and h is the separation distance between the two particles. When the interaction bodies (e.g., a colloidal particle and a collector) are immersed in water, U_{vdW} between the colloid and collector surface is obtained using the following equation developed by Hamaker (1937):

$$U_{\text{vdW}} = U_{\text{PC}} + U_{\text{WW}} - U_{\text{PW}} - U_{\text{WC}} \quad (2)$$

where U_{PC} is U_{vdW} between the colloid and collector in vacuum, U_{WW} is U_{vdW} between the water colloid and the water collector in vacuum, U_{PW} is the U_{vdW} between the colloid and the water collector in vacuum, U_{WC} is the U_{vdW} between the water colloid and the collector in vacuum. The expressions used to calculate U_{PC} , U_{WW} , U_{PW} , and U_{WC} all have the form of $-A_{\text{PC}}f(h)$, $-A_{\text{WW}}f(h)$, $-A_{\text{PW}}f(h)$, and $-A_{\text{WC}}f(h)$, respectively, where $f(h)$ is a function determined by the geometrical configuration for an interaction (e.g., colloid size

and shape, separation distance). A_{PC} , A_{WW} , A_{PW} , A_{WC} are Hamaker constants for colloid-collector, water-water, colloid-water, and water-collector systems, respectively. Therefore, Eq. (2) can be rewritten as

$$U_{vdW} = -(A_{PC} + A_{WW} - A_{PW} - A_{WC})f(h) = -A_H f(h) \quad (3)$$

where A_H is the Hamaker constant for a colloid-water-collector system ($A_H = A_{PC} + A_{WW} - A_{PW} - A_{WC}$). The expression for calculating vdW interaction energy between a spherical colloid and flat surface is given by

$$U_{vdW} = -\frac{A_H}{6} \left[\frac{R}{h} + \frac{a}{h+2R} + \ln \left(\frac{h}{h+2R} \right) \right] \quad (4)$$

where R is the radius of a spherical colloid.

For more complex scenarios, such as interactions between hollow particles, between a colloid and a bubble, and between two air bubbles, modifications to Eqs. (1) and (2) can be made for calculating U_{vdW} for these interaction configurations (see Table 1). The expression for calculating U_{vdW} between a spherical particle and a flat air-water interface (AWI) is

$$U_{vdW} = -\frac{A_{WW} - A_{PW}}{6} \left[\frac{R}{h} + \frac{a}{h+2R} + \ln \left(\frac{h}{h+2R} \right) \right] \quad (5)$$

As the value of A_{WW} is generally smaller than A_{PW} , the Hamaker constant for the colloid-AWI interaction is negative and thus the vdW force is repulsive. The expression for calculating the vdW energy between a spherical particle and a spherical bubble is

$$U_{vdW} = -\frac{A_{WW} - A_{PW}}{12} \left[\frac{\gamma}{x^2 + xy + x} + \frac{\gamma}{x^2 + xy + x + y} + 2 \ln \left(\frac{x^2 + xy + x}{x^2 + xy + x + y} \right) \right] \quad (6)$$

where $x = \frac{H}{(R_1 + R_2)}$ and $\gamma = \frac{R_1}{R_2}$ (R_1 and R_2 are radii of the particle and bubble, respectively). The expression for calculating U_{vdW} between two spherical bubbles is given by

$$U_{vdW} = -\frac{A_{WW}}{12} \left[\frac{\gamma}{x^2 + xy + x} + \frac{\gamma}{x^2 + xy + x + y} + 2 \ln \left(\frac{x^2 + xy + x}{x^2 + xy + x + y} \right) \right] \quad (7)$$

The Hamaker approach assumes pairwise summation of vdW force/energy, which may be violated in condensed matter when many-body interactions and retardation effects are significant. An alternative approach to calculate U_{vdW} , the Lifshitz approach, avoids the pairwise summation assumption by ignoring the atomic structure and calculates the Hamaker constant by treating the interacting surfaces as continuous media with well-defined bulk properties such as their dielectric constants and refractive indices. However, this approach involves complex mathematical equations, so simplifications are necessary even for calculating simple interaction configurations (e.g., sphere-flat surface interaction). The simplified Lifshitz approach has the same expressions as the Hamaker approach for calculating vdW force/energy for simple interaction configurations and the only difference between the two approaches is the way to calculate the Hamaker constant.

Electrostatic or double layer force/energy

When a charged particle or collector surface is immersed in an electrolyte solution, an ionic cloud, also called an electrostatic double layer (DL), is formed around each surface. Overlap of the DLs results in DL energy/force when the particle approaches the collector surface. The Poisson-Boltzmann (PB) equation is used to determine the electrostatic potential distribution between a particle and a surface, and DL energy can then be obtained by analyzing the change in free energy of the DL system (Hogg et al., 1966). If the electrolyte is symmetric (i.e., when the cations and anions have the same valence), the PB equation is written as

$$\nabla^2 \Psi = \kappa^2 \sinh(\Psi) \quad (8)$$

where ∇^2 is the Laplacian operator, $\Psi (=ve\psi/kT)$ is reduced surface potential, v is charge number, e is electronic charge, ψ is electrostatic potential, k is the Boltzmann constant, T is absolute temperature, and κ is the inverse Debye screening length. The PB equation is nonlinear thus an analytical solution of electrical potential distribution does not exist even for a spherical particle. Accordingly, the PB equation has been frequently linearized (valid for small surface potentials, e.g., ≤ 25 mV) and given by

Table 1 Expressions used to calculate vdW energies between two interaction bodies in water.

Inter action configurations	Expressions
Colloid-collector	$U_{vdW} = U_{PC} + U_{WW} - U_{PW} - U_{WC}$
Colloid-AWI	$U_{vdW} = U_{WW} - U_{PW}$
Colloid-colloid	$U_{vdW} = U_{PP} + U_{WW} - 2U_{PW}$
Hollow colloid filled with water-collector	$U_{vdW}^{out} - U_{vdW}^n$
Hollow colloid filled with air-collector	$U_{vdW}^{out} - U_{vdW}^n + U_{WW}^n - U_{WC}^n$
Bubble-collector or colloid-AWI	$U_{WW} - U_{WC}$
Bubble-bubble	U_{WW}

$$\nabla^2 \Psi = \kappa^2 \Psi \quad (9)$$

Eq. (9) can be solved to obtain electrostatic potential at any point between the two interacting surfaces. Calculated electrostatic potential distribution can be adopted to calculate DL energy using the expression developed by Verwey and Overbeek (1948) and assuming appropriate boundary conditions, which include two extreme cases: constant surface potential and constant surface charge. However, the true condition lies somewhere between the two extremes. In addition to the two extreme conditions, linear superposition approximation (LSA) has also been used as a boundary condition (Shen et al., 2020). Analytical equations can be obtained, using the above three boundary conditions, to calculate DL energy per unit area between two parallel infinite planar surfaces (Table 2). Various analytical and numerical methods such as the Derjaguin approach (DA) and surface element integration (SEI) technique commence with these analytical equations to calculate DL energies between a colloid and a surface (Hogg et al., 1966; Bhattacharjee and Elimelech, 1997; Lin and Wiesner, 2010).

H is the separation distance, ϵ_0 is the dielectric permittivity in vacuum, ϵ is the relative dielectric permittivity of solvent, κ is the inverse Debye screening length, ν is the charge number, e is the electronic charge, ψ_1 and ψ_2 are the zeta potentials of two interacting surfaces, κ is Boltzmann constant, and T is absolute temperature, R is colloid radius.

Non-DLVO forces

In addition to the DLVO forces, non-DLVO forces are considered under certain circumstances. Hydration, steric repulsion, and hydrophobic interactions are three short-range repulsive forces that colloids commonly experience in saturated porous media. Hydration force F_{HR} , or hydration repulsion, occurs because it requires energy to dehydrate interacting surfaces that contain ionic or polar species, which is extremely significant and can inhibit colloids from approaching a collector surface at very short separation distances (i.e., < 0.5 – 0.6 nm, twice the thickness of a water molecule). However, hydration repulsion decays rapidly with separation distance in an exponential manner and is influenced by various factors including hydrophilicity of the interacting surfaces and solution ionic strength. Quantitative estimation of hydration energy remains difficult; thus only empirical models are available for determining F_{HR} . An example of the empirical models is

$$F_{HR} = \frac{E_0}{\lambda} \exp(-h/\lambda) \quad (10)$$

where λ is characteristic decay length, E_0 is the maximum repulsive energy per unit area at the closest possible separation distance, equal to $N_A C_h C_e \lambda$ (N_A is the Avogadro number, C_h is the proportionality constant defined as “hydration constant,” and C_e is the electrolyte concentration).

Steric repulsion results from interactions between surfaces with adsorbed chains and/or chain elements (e.g., surfactants, polymers, and organic matter) at short separation distances (i.e., less than twice the thickness of the adsorbed layer) (Ryan and Elimelech, 1996). At such short distances, the polymer chains on the two surfaces interdigitate and steric repulsion arises. In addition to interdigitation, the brushes are simultaneously compressed when the opposite polymer brushes contact each other. Therefore, both entropic and elastic effects contribute to steric repulsion. However, if some of the polymer chains on one

Table 2 Expressions for calculating DL energy for the interaction between two parallel infinite planar surfaces (E) and for the interaction between a spherical colloid and a planar surface using the DA technique (U_{DA}) or method of Lin and Wiesner (2010) (U_{LW}).

Boundary condition	Expressions
CSP	$E^{CSP}(H) = \frac{\epsilon_0 \epsilon \kappa}{2} \left[(\psi_1^2 + \psi_2^2) (1 - \coth(\kappa H)) + 2\psi_1 \psi_2 \operatorname{csch}(\kappa H) \right]$ $U_{DA}^{CSP}(H) = \pi \epsilon \epsilon_0 R \left[(\psi_1^2 + \psi_2^2) \ln(1 - \exp(-2\kappa H)) + 2\psi_1 \psi_2 \ln\left(\frac{1 + \exp(-\kappa H)}{1 - \exp(-\kappa H)}\right) \right]$ $U_{LW}^{CSP}(H) = [U_{DA}^{CSP}(H) + U_{DA}^{CSP}(H + 2R)] + [\beta^{CSP}(H) - \beta^{CSP}(H + 2R)]$ $\beta^{CSP}(x) = \frac{\pi}{\kappa} \epsilon \epsilon_0 \left[\frac{1}{2} (\psi_1^2 + \psi_2^2) \operatorname{Li}_2(\exp(-2\kappa x)) + 2\psi_1 \psi_2 [\operatorname{Li}_2(-\exp(-\kappa x)) - \operatorname{Li}_2(\exp(-\kappa x))] \right]$
CSC	$E^{CSC}(H) = 32 \epsilon \epsilon_0 \kappa \tanh\left(\frac{v\psi_1}{4kT}\right) \tanh\left(\frac{v\psi_2}{4kT}\right) \frac{(kT}{v\epsilon})^2 \exp(-\kappa H)$ $U_{DA}^{CSC}(H) = \pi \epsilon \epsilon_0 R \left[-(\psi_1^2 + \psi_2^2) \ln(1 - \exp(-2\kappa H)) + 2\psi_1 \psi_2 \ln\left(\frac{1 + \exp(-\kappa H)}{1 - \exp(-\kappa H)}\right) \right]$ $U_{LW}^{CSC}(H) = [U_{DA}^{CSC}(H) + U_{DA}^{CSC}(H + 2R)] + [\beta^{CSC}(H) - \beta^{CSC}(H + 2R)]$ $\beta^{CSC}(x) = \frac{\pi}{\kappa} \epsilon \epsilon_0 \left[-\frac{1}{2} (\psi_1^2 + \psi_2^2) \operatorname{Li}_2(\exp(-2\kappa x)) + 2\psi_1 \psi_2 [\operatorname{Li}_2(-\exp(-\kappa x)) - \operatorname{Li}_2(\exp(-\kappa x))] \right]$
LSA	$E^{LSA}(H) = \frac{\epsilon_0 \epsilon \kappa}{2} \left[-(\psi_1^2 + \psi_2^2) (1 - \coth(\kappa H)) + 2\psi_1 \psi_2 \operatorname{csch}(\kappa H) \right]$ $U_{DA}^{LSA}(H) = 64 \pi \epsilon \epsilon_0 R \left(\frac{kT}{v\epsilon} \right)^2 \tanh\left(\frac{v\psi_1}{4kT}\right) \tanh\left(\frac{v\psi_2}{4kT}\right) \exp(-\kappa H)$ $U_{LW}^{LSA}(H) = [U_{DA}^{LSA}(H) + U_{DA}^{LSA}(H + 2R)] + \frac{64 \pi \epsilon \epsilon_0}{\kappa} \left(\frac{kT}{v\epsilon} \right)^2 \tanh\left(\frac{v\psi_1}{4kT}\right) \tanh\left(\frac{v\psi_2}{4kT}\right) [-\exp(-\kappa H) + \exp(-\kappa(H + 2R))]$

CSP, constant surface potential; CSC, constant surface charge; LSA, linear superposition approximation.

surface are adsorbed on the other interacting surface, the polymers can draw the surfaces closer to each other, causing attractive bridging forces. In this case, the repulsive steric and attracting bridging forces can exist simultaneously and have coupled but non-additive effects on the interaction of the surfaces.

Hydrophobic interaction involves the configurational rearrangement of water molecules as two hydrophobic surfaces come together, particularly common when AWI, which is superhydrophobic, is involved. Although hydrophobic interaction has long been believed to be due to a "hydrophobic bond," this interaction has later been recognized to act at a longer range than a typical covalent bond. Israelachvili and Pashley (1982) were the first to directly measure the forces between two hydrophobic surfaces in aqueous solution. They found that hydrophobic force decayed exponentially with separation distance with a characteristic length of ~ 1 nm. The origin of hydrophobic interaction is not completely known but has been suggested to include bridging of surface nano-/micro-bubbles, bubble cavitation during surface approaching or separation, or entropy elevation during the disruption and change of hydrogen bonding network at the water-hydrophobic interface.

Under unsaturated conditions, colloids can interact with AWI, and such interactions largely depend on colloid charge and hydrophobicity. If interactions are attractive, the thin water film between colloids and AWI breaks, and colloids experience capillary forces when they penetrate the AWI. The capillary force arises due to change in the shape of the AWI upon contacts with colloids and corresponding actions of surface tension force and Laplace pressure (Flury and Aramrak, 2017). The magnitude of capillary forces is typically much larger than the DLVO interaction forces (Gao et al., 2008), so they can play a significant role in the aggregation, retention and release of colloids in unsaturated media. Calculations of the capillary force require solving the Laplace equation of capillarity and determining the meniscus profile around the colloids with different configurations (Flury and Aramrak, 2017).

In a porous medium with low water content, a capillary meniscus is formed due to capillary force, which leads to pinning of a colloid to the stationary solid surface (Fig. 1A). In this case, the capillary force (F_{cap}) normal to the AWI can be calculated by

$$F_{cap} = \pi R \gamma [2 \sin \phi \sin(\theta + \phi) + \cos \theta (1 + \cos \phi)^2] - \sin \phi \quad (11)$$

where γ is surface tension of water, θ is contact angle, and ϕ is filling angle. As water content in the porous medium increases, the adsorbed water film expands, resulting in a free water surface (Fig. 1B). The capillary force can be determined

$$F_{cap} = \pi R^2 \rho g \left\{ \frac{Hc + Rc[1 + \cos(\pi - \phi)]}{Rc} + \frac{2}{(Rc)^2} \sin(\phi + \theta) \sin \phi \right\} \quad (12)$$

where g is gravitational acceleration, H is thickness of the free water film, $c = \sqrt{(\Delta \rho g)/\gamma}$ is the capillary number and $\Delta \rho$ is the difference between water and air density. The thickness of H is related to the filling angle ϕ , and their relationship is described by the Laplace equation. When the water film thickness H is larger than the colloid diameter by further increase of water content (Fig. 1C), the direction of the capillary force changes and become opposite to the adhesive force between the colloid and solid surface. The maximum capillary force in the opposite direction (F_{det}) can be estimated

$$F_{det} = 2\pi R \gamma \sin^2 \left(\frac{\theta}{2} \right) \quad (13)$$

The DA and SEI technique

The DA and SEI techniques are the two most commonly used methods for interaction energy calculations, which are based on analytical equations of differential interaction energies between two planar surfaces per unit area or between an area element and an infinite planar surface, as described previously. The DA technique calculates sphere-sphere interaction energy by discretizing sphere surfaces into infinitesimally ring-shaped flat plates and then sums up differential interaction energies between each pair of ring plates (e.g., dS_1 and dS_2 in Fig. 2A). The differential interaction energies between the ring plates are calculated using the expressions shown in Table 2. Calculations using the DA technique are accurate only for large colloids and when $\kappa R > 5$ (κ is the inverse Debye screening length). The SEI technique calculates the interaction energy for a sphere-flat interaction (Fig. 2B), developed by Bhattacharjee and Elimelech (1997). In this case, the sphere is discretized into small elements, and the total interaction energy is the sum of differential interaction energies between each area element and the planar surface. The SEI technique has overcome the limitations of DA and can accurately calculate interaction energies even for nano-sized particles and at all possible separation distances.

DLVO interaction energy/force profiles and maps

DLVO interaction energy profile is used for interpreting attachment and detachment of colloids on/from surfaces. A standard DLVO interaction energy profile is constructed by plotting total interaction energy (e.g., sum of vdW and DL energy) as a function of separation distance between a colloid and a collector surface. When a colloid and surface are like-charged, the electrical DL energy is repulsive and the combination of the repulsive DL energy with the attractive vdW energy results in an infinitely deep primary energy well, called the primary minimum, at a small separation distance, a maximum energy barrier, and a shallow attractive energy well, called the secondary minimum, at a larger distance. Therefore, the classic DLVO theory predicts that colloids are irreversibly attached at the deep primary minima with infinite depths.

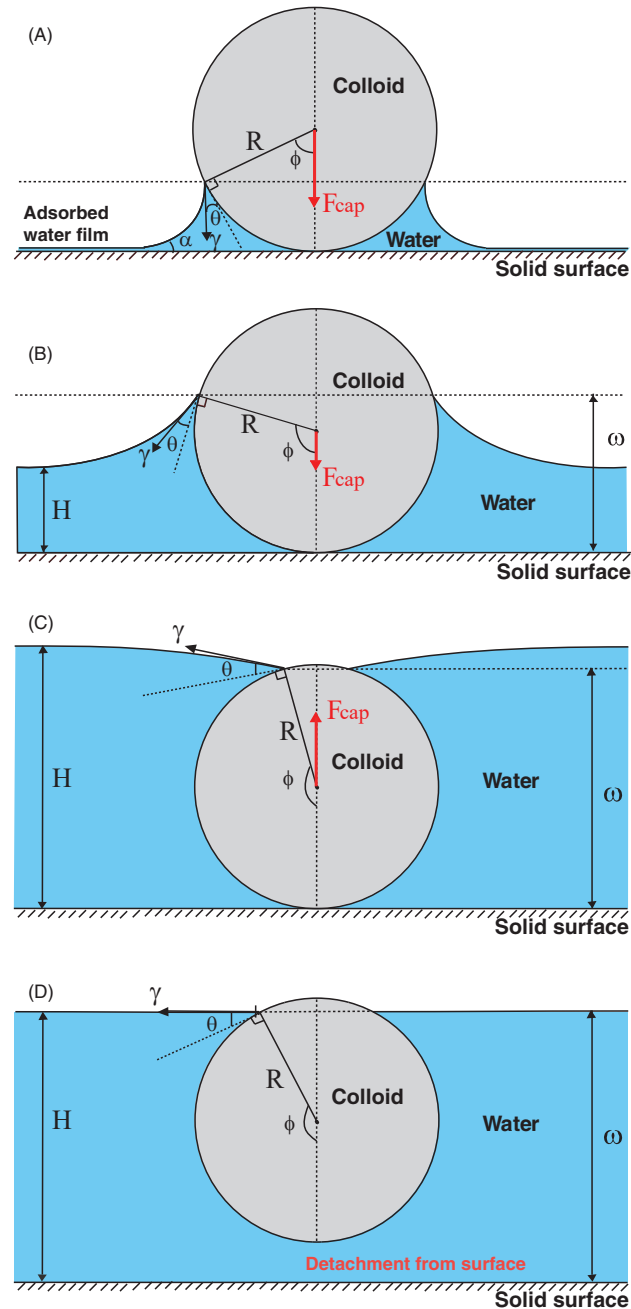


Fig. 1 Capillary forces acting on a colloid for different configurations of the AWI. (A) capillary meniscus for a dry porous medium with adsorbed water film, (B) free water film with thickness H smaller than the colloid diameter $2R$ ($H < 2R$), (C) water film thickness expanded to $H > 2R$, and (D) further expansion of water film thickness leads to detachment of colloid from solid surface. Figure and caption from Flury M and Aramrak S (2017) Role of air-water interfaces in colloid transport in porous media: A review. *Water Resources Research* 53: 5247–5257, <https://doi.org/10.1002/2017WR020597>, with permission.

Predictions by the classic DLVO theory is often found to be in contrast to observations from many column experiments, which show that colloids attached at primary minima can be released by reducing solution ionic strength and by increasing hydrodynamic shear (Ryan and Elimelech, 1996; Shen et al., 2020). Surface force measurements in microscopic experiments unambiguously showed the existence of very strong short-range repulsion at separation distances less than a few nanometers. This is likely due to the existence of the non-DLVO forces such as hydration and hydrophobic interactions, as previously discussed. To date, only empirical expressions are available to estimate these short-range interaction forces. Alternatively, the effect of short-range repulsion is included by designating a minimum separation distance corresponding to shear plane distance and the thickness of hydration layers between surfaces and by calculating Born potential energy (Ryan and Elimelech, 1996). The depth of primary minimum becomes limited when the aforementioned short-range repulsions are involved (Fig. 3). Colloids can overcome the energy barriers to attach at the

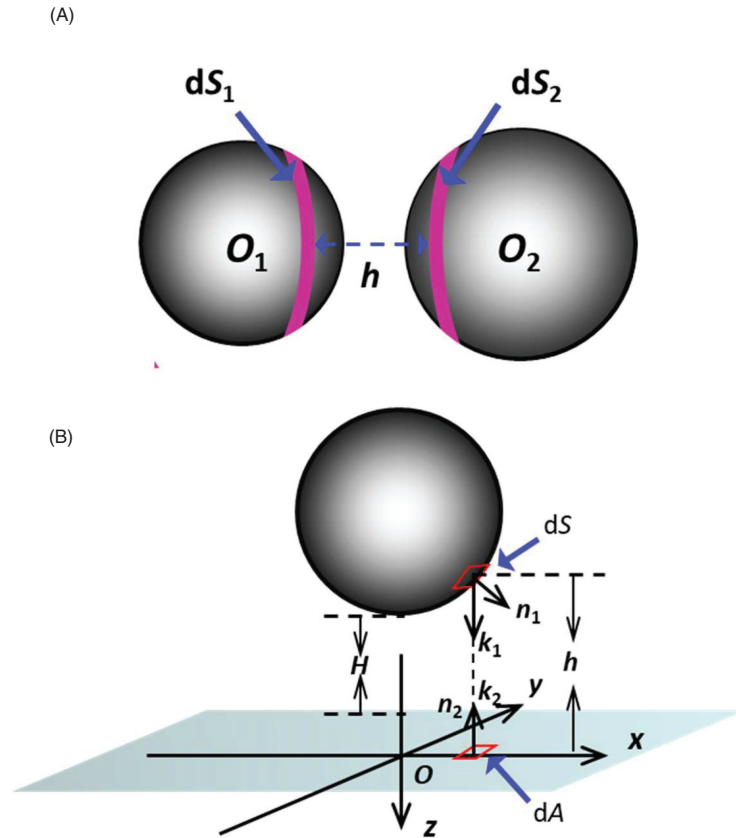


Fig. 2 Schematic of interaction between two spherical particles or between a spherical particle and a flat plate. dS_1 and dA are differential area elements on colloid and collector surfaces, respectively, h is local distance between the elements dS_1 and dS_2 or between dS and dA , n_1 and n_2 are unit outwards normal to the particle and collector surfaces, respectively, k_1 and k_2 are unit vectors directed towards the positive and negative z axis, respectively, H is separation distance. Modified from Shen C, Jin Y, Zhuang J, Li T, and Xing B (2020) Role and importance of surface heterogeneities in transport of particles in saturated porous media. *Critical Reviews in Environmental Science and Technology* 50: 244–329, with permission.

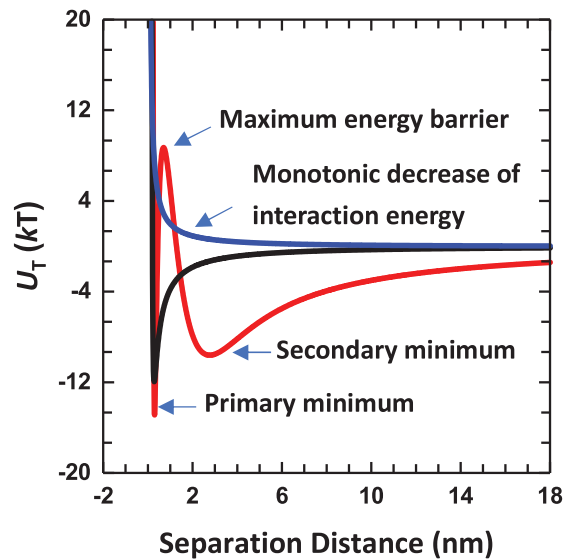


Fig. 3 DLVO interaction energy U_T as a function of separation distance. For the red curve, a primary minimum, maximum energy barrier, and a secondary minimum is present; In black curve, only a primary minimum is present; In blue curve, the interaction energy decreases monotonically with increasing separation distance. Modified from Shen C, Jin Y, Zhuang J, Li T, and Xing B (2020) Role and importance of surface heterogeneities in transport of particles in saturated porous media. *Critical Reviews in Environmental Science and Technology* 50: 244–329, with permission.

primary minima if the energy barriers are small and the adhesive torques acting on the colloids at the primary minima are larger than applied torques (e.g., hydrodynamic torque). Colloids may also attach at the secondary minima if the secondary energy wells are deep enough to prevent the colloids from being released by hydrodynamic shear. Colloid attachment in the presence of repulsive energy barriers is termed “unfavorable”, which is the most prevalent condition in natural porous media where colloids and collector surfaces are like charged. When colloid and collector surfaces are oppositely charged, both the DL and $\mathcal{W}$ energies are attractive and only the primary minimum exists in the interaction energy profile. In this case, colloids nearby a collector surface will attach at the primary minima as the attraction and the attachment is “favorable”.

A DLVO interaction energy profile/curve is sufficient to characterize the interactions between a particle and a collector surface in the absence of surface heterogeneity (Fig. 4A). However, if physical and chemical heterogeneities exist on the collector surface, the interaction energy profile changes with the horizontal location of the colloid (Fig. 4B). Shen et al. (2020) showed that the interaction energy can decrease monotonically with separation distance at low ionic strength values when nanoscale protruding asperities exist on the collector surface (Fig. 3). This happens when colloids experience repulsion at all separation distances, thus there are no energy wells available for attachment. Notably, to describe the variation of interaction energy with the horizontal position due to surface heterogeneities, DLVO interaction energy maps can be constructed by computing interaction energies at a given separation distance and over a range of locations (x, y). Shen et al. (2012) presented DLVO interaction energy and force maps in the xz or yz plane by changing the vertical positions (x, z) or (y, z) in a rasterized manner (Fig. 5). The DLVO interaction energy and force maps in xz or yz plane have the advantage of clearly showing variations of the primary minimum, energy barrier, and secondary minimum due to the influence of surface heterogeneity.

Processes governing colloid transport

Fig. 6 provides a schematic of major locations and mechanisms for colloid retention in saturated (Fig. 6A) and unsaturated (Fig. 6B) porous media. Colloid transport in saturated porous media is controlled by deposition at the solid-water interface (SWI) and straining (Bradford et al., 2006; Shen et al., 2020). While colloid retention at SWI and straining occur in both saturated and unsaturated media, there are two major complications in the unsaturated porous media: the presence of an air phase and transient

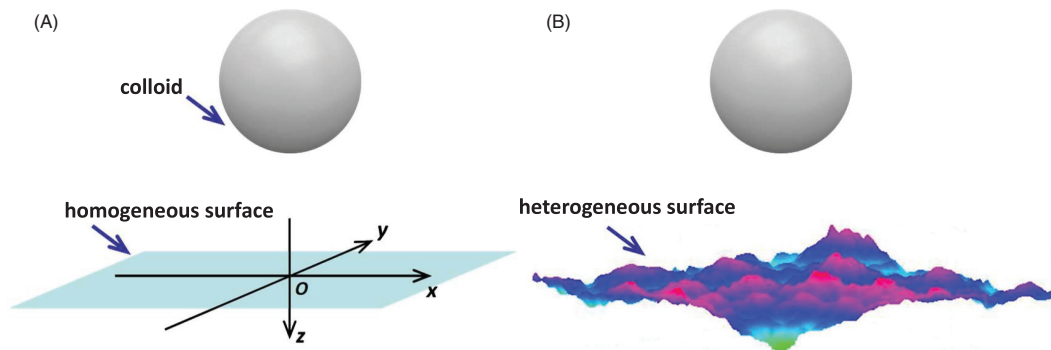


Fig. 4 Illustration of a colloid interacting with a (A) homogeneous or (B) heterogeneous surface.

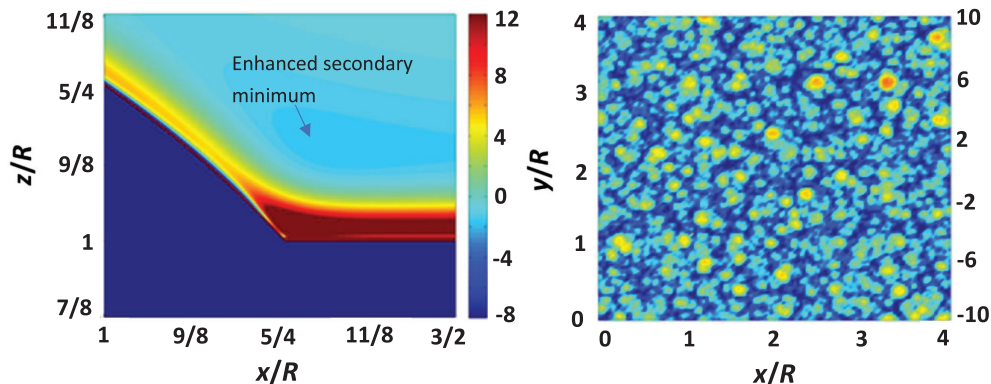


Fig. 5 DLVO interaction energy maps in xz or xy plane. R is colloid radius. The scale bar between the plots represents the interaction energy in units of kT (k is Boltzmann constant and T is absolute temperature). Modified from Shen C, Wang F, Li B, Jin Y, Wang L-P, and Huang Y (2012) Application of DLVO energy map to evaluate interactions between spherical colloids and rough surfaces. *Langmuir* 28: 14681–14692, with permission.

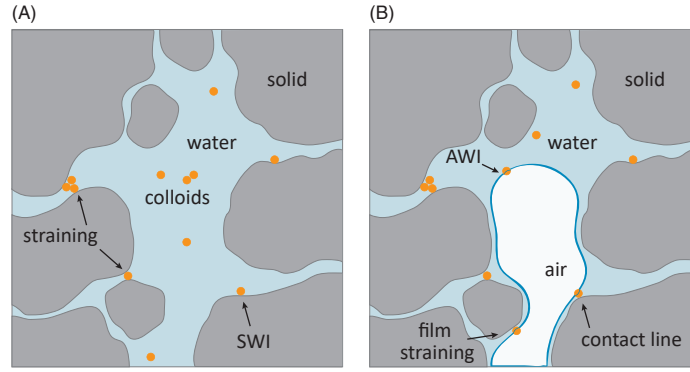


Fig. 6 Schematic of colloid retention mechanisms in (A) saturated and (B) unsaturated porous media. (A) Major colloid retention mechanisms in saturated porous media are retention at the solid-water interface (SWI) and straining. (B) Additional colloid retention mechanisms in unsaturated porous media include film straining and retention at the air-water interface (AWI) and contact line.

flow events (McCarthy and McKay, 2004). The presence of air phase affects colloid partitioning and distribution between the phases and results in additional mechanisms of colloid retention including retention at the air-water-solid interface (contact line) and the AWI, and by film straining (e.g., Bradford et al., 2006).

Deposition

Deposition is a process that involves collision of a colloid in bulk solution with a collector surface and subsequent immobilization at the surface. According to the CFT, deposition is controlled by three individual retention mechanisms: diffusion, interception and sedimentation (Yao et al., 1971). Deposition is quantitatively represented by the single collector removal efficiency η in the CFT, which is defined as the ratio of the overall colloid deposition rate onto a collector to convective transport of upstream colloids toward a collector (Ryan and Elimelech, 1996). The single collector removal efficiency can be obtained by solving the convective-diffusion equation, taking into account the aforementioned three deposition mechanisms or by quantifying the trajectory of particles, using the Lagrangian method in both cases.

Existing theoretical calculations of deposition rate or single collector removal efficiency can be orders of magnitude smaller than experimental observations under unfavorable conditions, thus they cannot accurately predict colloid transport in porous media (Molnar et al., 2015; Shen et al., 2020). Consequently, an additional factor, the attachment efficiency α , is used in the CFT to account for the discrepancies. The single collector removal efficiency η is decoupled as the product of single collector contact efficiency η_0 , which quantifies the transport step from bulk solution to the vicinity of a collector surface, and the attachment efficiency α , which quantifies the attachment step. Various correlation equations have been developed by regressing dimensionless parameters governing colloid deposition against the theoretical single collector contact efficiency value over a broad range of parameter values (Yao et al., 1971; Rajagopalan and Tien, 1976; Tufenkji and Elimelech, 2004), and a summary of the correlation equations can be found in Molnar et al. (2015) and Shen et al. (2020).

The attachment efficiency α represents the fraction of collisions with the collector surface that results in attachment, which is determined by conducting column experiments and calculated using the following expression

$$\alpha = -\frac{2}{3} \frac{d_c}{(1-f)L\eta} \ln(C/C_0) \quad (14)$$

where d_c is collector diameter, f is porosity of a porous medium, L is column length, C and C_0 are colloid concentration of influent and effluent suspensions, respectively. Numerous theoretical expressions have also been developed to calculate the attachment efficiency. For example, the Boltzmann model is used to calculate the attachment efficiency at primary minima

$$\alpha = \exp(-U_{\max}) \quad (15)$$

where $U_{\max}$ is the maximum energy barrier.

Attachment efficiency at secondary minima can be calculated using the Maxwell equation (Hahn et al., 2004) and a number of empirical correlations (Shen et al., 2020). Although deposition at secondary minima has been a subject under debate since the development of DLVO theory, experimental evidence using techniques, such as total internal reflection microscopy, has provided unambiguous support for its occurrence (Shen et al., 2020).

Release

Deposition is reversible from the primary minimum under certain conditions, including when short-range repulsion reduces the depth of the primary minimum and when hydrodynamic torques exceed the adhesive torques at high flow velocities. Torque balance analysis can be used to determine the release from primary minima; rolling has been found to be the governing mechanism

causing the release of colloids from primary minima compared to lifting and sliding (Bergendahl and Grasso, 2000). Additional mechanisms that can cause colloid release from primary minima include: presence of nanoscale surface protruding asperities, which can significantly reduce the primary minimum depth at low ionic strength, and increased electrostatic repulsion and steric repulsion by increasing pH or introducing surfactants (e.g., polymers and natural organic matter). Colloids can even spontaneously detach from primary minima by Brownian diffusion if the primary minimum depths are reduced to be comparable to the average kinetic energy of a colloid ($1.5 kT$, k is Boltzmann constant and T is absolute temperature).

Colloids that are attached via secondary minimum association are more readily released compared to those attached at primary minima. They can be released by decreasing solution ionic strength because this decreases the depths of secondary minima (Hahn et al., 2004). These colloids may also translate and rotate along the collector surfaces due to hydrodynamic shear when the collector surfaces are relatively smooth. However, deposition in deep secondary minima is stable when hydrodynamic shear is absent (e.g., at stagnation point regions), or when tangential attractive forces exist (e.g., presence of nanoscale protruding asperities).

Straining

Straining refers to the retention of colloids in down-gradient pore throats, formed due to intersection of at least two interfaces that are too small for the colloids to pass through. Straining mainly occurs at the interaction of only two solid-water interfaces, termed grain-grain contact or wedging. Straining is controlled by the size of colloids and pore size distribution of a porous medium and can be influenced by chemical factors such as solution ionic strength and surface tension (Bradford et al., 2006). In contrast to deposition that causes exponential retention profiles under favorable conditions, colloids retained via straining are mainly distributed at the inlet of a porous medium, resulting in hyperexponential retention profiles.

Retention at AWI, contact line and film straining

AWI plays a critical role in colloid transport in porous media, and can both reduce and enhance transport depending on various factors (Flury and Aramrak, 2017). The processes and mechanisms controlling colloid retention at the AWI and contact line depend on the configuration of air phase (e.g., as a continuous phase or trapped air bubbles) and corresponding hydrodynamic regime. Two hydrodynamic regimes are typically considered: steady-state and transient, under which the roles that the AWI plays in colloid transport can be remarkably different.

Steady-state regime represents flow under constant saturation (water content) and flow rate. Under this regime, air is trapped in porous media, AWI boundaries remain reasonably constant, and therefore the AWI and contact line can be considered stationary. Under steady-state conditions, there is more colloid retention in unsaturated porous media compared to saturated ones at the same colloid and solution chemistries (DeNovio et al., 2004), which increases with decreasing water content (e.g., Wan and Tokunaga, 1997). The increased retention has been attributed to retention at the AWI and on the contact line and to film straining.

Interaction of a colloidal particle with AWI is governed by forces including the repulsive vdW force, electrostatic interaction, and hydrophobic force. Capillary force, which has been described previously, acts on colloids only after they penetrate and attach to the AWI. The majority of the colloids found in the environment are hydrophilic and negatively charged, thus the net interaction force between colloids and AWI is either repulsive or weakly attractive (due to hydrophobic force). Under such conditions, AWI does not serve as a major retention site for colloid attachment from the bulk solution. The net interaction force may be different for positively charged colloids so the role of AWI needs to be considered for individual cases.

Under steady-state conditions, contact line is stationary and colloid motion towards the contact line is determined by the flow field in the bulk solution. When colloid motion is parallel to contact line, colloids can be retained in the contact line regions at locations with low or zero flow velocities. Colloids can also be carried to the contact line by convective flows, Brownian motion, or evaporation and be retained via so-called film straining (Wan and Tokunaga, 1997). In unsaturated porous media where water is present in the form of pendular rings around grain-to-grain contacts and thin films on grain surfaces, retention can occur both within disconnected pendular rings and in thin films for colloids with sizes comparable to or smaller than the thickness of the thin films. These colloids can be re-mobilized back into bulk solution upon a transient event (infiltration). In general, colloids can be retained on the contact line with or without penetrating AWI (Flury and Aramrak, 2017). Fig. 7A illustrates colloid retained both at the AWI and the contact line and the role of capillary force. Alternatively, colloids can be trapped within the liquid wedge on the contact line (Fig. 7B). In this case, the thin film around the particle doesn't break; the theoretical treatment of such retention was proposed by Gao et al. (2008).

Transient flow regime in unsaturated porous media includes such events as imbibition (infiltration), drainage, and evaporation. Transient flow has been recognized as the main process governing colloid mobilization in soil, which is driven by multiple mechanisms including thin-film expansion, mobilization with moving AWI or contact line, and hydrodynamic shear (Aramrak et al., 2011; DeNovio et al., 2004). During transient flow, the moving AWI and contact line can mobilize colloids and transport them with the moving AWI. There are a number of forces acting on an attached colloid upon its interaction with a moving contact line, and the force balance determines whether a colloid is mobilized or remains on the surface (Fig. 8). These forces include adhesion (DLVO) force between the attached colloid and solid collector surface, hydrodynamic force, and capillary force. Capillary force is the governing force of colloid mobilization, and its magnitude and direction determine whether a colloid is mobilized or pinned against the collector surface (Lazouskaya et al., 2013). Therefore, colloid and collector contact angles and surface tension are the parameters that affect the capillary force value and colloid mobilization. Once removed from the collector surface by the capillary force, colloids remain attached at AWI and can be further transported with the moving AWI.

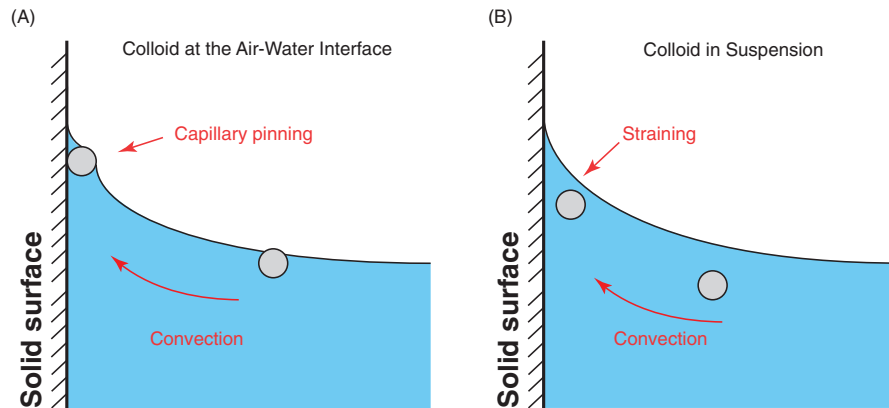


Fig. 7 Colloid retention on the contact line at steady-state conditions: (A) pinning capillary force on the colloid, and (B) colloid trapping from the solution in the wedge. Modified from Flury M, Aramrak S (2017) Role of air-water interfaces in colloid transport in porous media: A review. *Water Resources Research* 53: 5247–5257, doi:10.1002/2017WR020597, with permission.

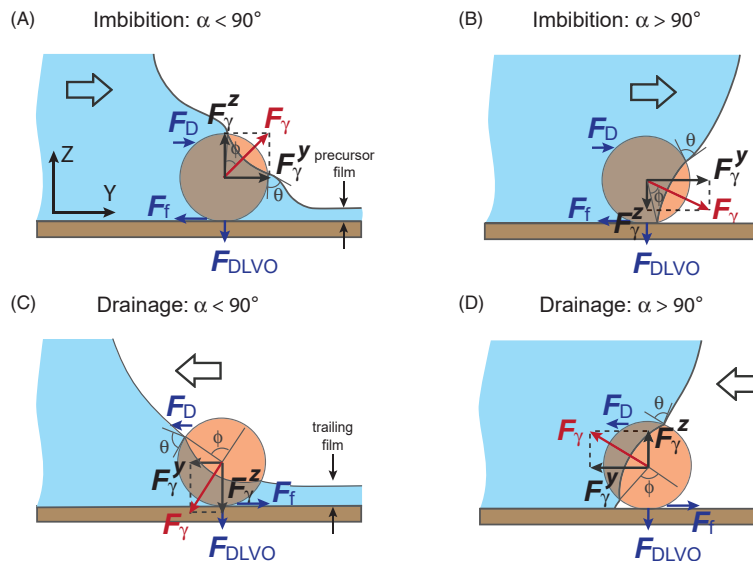


Fig. 8 Schematic of force balance acting on an attached colloid during imbibition (A, B) and drainage (C, D) for hydrophilic ($\alpha < 90^\circ$) and hydrophobic ($\alpha > 90^\circ$) collectors where α is collector contact angle. F_γ is surface tension component of capillary force, F_{DLVO} is DLVO force, F_D is hydrodynamic force, and F_f is friction force. Large arrows indicate the flow direction. Modified from Lazouskaya V, Wang LP, Or D, Wang G, Caplan JL, and Jin Y (2013) Colloid mobilization by fluid displacement fronts in channels. *Journal of Colloid and Interface Science* 406: 44–50, doi: 10.1016/j.jcis.2013.05.078, with permission.

Colloid mobilization mechanisms of model spherical colloids during imbibition and drainage in porous media are relatively well understood (Lazouskaya et al., 2013). However, mobilization of natural colloids in soil is more complex and can vary depending on soil properties (composition, organic matter content) and other parameters such as moisture content, duration and history of AWI transients, and evaporation. Overall, transient conditions (either chemical or physical such as ionic strength or flow rate) are favorable for colloid mobilization. Additionally, water loss with drainage and evaporation induces capillary stresses and weakens soil capillary walls, which increases the number of mobilized colloids with subsequent infiltration events, and drying duration accentuates these effects (e.g., Michel et al., 2010). Freezing/thawing cycles mobilize even more colloids than wetting/drying events due to the additional mechanical stresses on soil structural units (Flury and Aramrak, 2017).

Experimental approaches

Colloid transport is determined by the relative magnitudes of concurrent processes of retention (deposition, straining) and release or mobilization (Sen and Khilar, 2006). Laboratory methods to study and quantify colloid retention and release in natural and

model porous media include batch and column experiments, pore-scale studies using microscopy, X-ray tomography and other advanced tools. These methods focus on different processes and they are often used to complement each other.

Similar to studying chemical sorption, batch experiments are conducted to measure attachment and detachment of colloids to and from solids (e.g., sand or soil). In typical batch experiments, a solution containing a known concentration of colloids is added to a known quantity of sand/soil at a predetermined solid-to-liquid mass ratio. The mixture is then shaken on a rotary table or a shaker at a specified rotation speed for a specified amount of time to reach (apparent) equilibrium. The final or equilibrium concentration of colloids in solution is measured and the amount of attached colloids is calculated based on the mass balance, which is then used to construct equilibrium isotherms using linear, Langmuir or Freundlich models. The colloid attachment process is typically considered as linear and reversible and a distribution coefficient (K_D) is obtained and used to calculate a retardation factor that can be used in transport models (Schijven and Hassanizadeh, 2000). In this manner, important parameters such as colloid attachment and detachment can be measured under various chemical conditions (e.g., solution pH and ionic strength). The main advantages of batch experiments are their simplicity and being less labor-intensive compared to column and field studies.

Laboratory column experiments under saturated or unsaturated porous media dominate colloid transport research. The experimental setup includes packing a column with porous media at desired water saturation, followed by introducing influent colloid suspension. The effluent is collected and analyzed for colloid concentration, and the results are used to generate breakthrough curves. Column experiments can be conducted at steady state or transient state (DeNovio et al., 2004). Although results from batch and column experiments often cannot be compared directly due to different solid-to-solution ratios and other interaction conditions, parameters from batch experiments (e.g., attachment and detachment coefficients) are sometimes used to aid interpretation of results from column experiments (Bradford et al., 2006). Although column experiments provide valuable data in colloid transport, mechanistic understanding is hindered by the combined effects of multiple processes that occur simultaneously in columns.

With the progress in microscopy, it became possible to distinguish colloid retention mechanisms and their contributions to colloid transport by conducting pore-scale and meso-scale visualization experiments (Ochiai et al., 2006). These experiments complement column experiments and generate important insights into colloid retention mechanisms especially in unsaturated systems. Typical pore-scale experiments include a micromodel device (e.g., etched glass and silicon pore networks, media-packed flow cells, and capillary channels) connected to a syringe pump and visualized with a microscope (Lazouskaya et al., 2006, 2011, 2013). Various micromodel designs represent different pore and AWI configurations. AWI can be introduced as entrapped air bubbles, as a free surface in open capillaries and packed media, and as phase boundary in two-phase flows. Similar to column experiments, pore-scale experiments can be carried out to study colloid behavior during both steady-state and transient flow processes. Analysis of these configurations and corresponding results clarified how the roles of AWI and contact line in colloid retention and mobilization differ in steady-state systems with stationary AWI and contact line and in the systems with moving AWI and contact line (e.g., two-phase flow) (Flury and Aramrak, 2017). The common limitations of pore-scale experiments include the small representative area of study and the two-dimensional nature of such systems. Fig. 9 illustrates fluorescent colloids and a moving AWI during imbibition and drainage in a microfluidic channel with a trapezoidal cross section. Because the channel depth was only 20 μm , colloid retention at AWI and on the contact line could be clearly distinguished.

Imaging studies at the intermediate meso-scale provide the means to investigate the pore structure and topography of a porous medium, configurations of water films and entrapped air, and colloid retention in three-dimensional domains. These studies employ advanced imaging techniques such as magnetic resonance imaging, X-ray and neutron tomography, and light-based

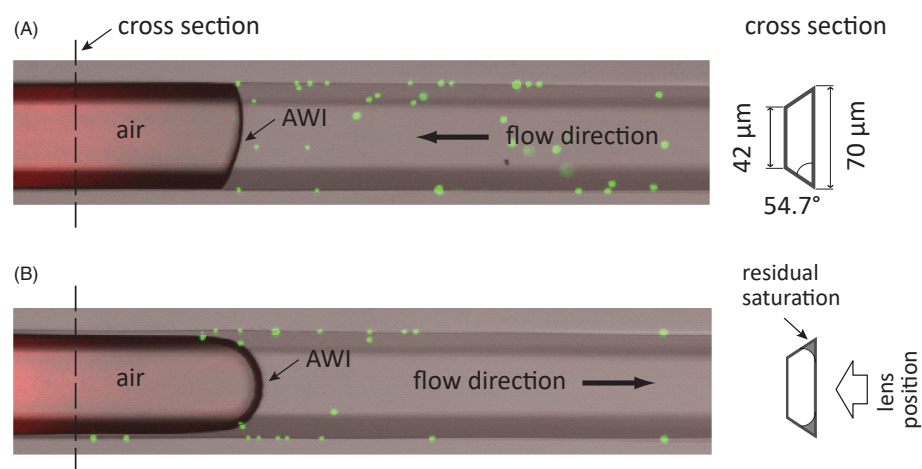


Fig. 9 Images of microfluidic channel acquired with the confocal microscope and the schematic cross-sections in two regimes: (A) imbibition and (B) drainage. The dashed lines mark the approximate location of the schematic cross-sections. Modified from Lazouskaya V, Wang L, Gao H, Shi X, Czymbek K, and Jin Y (2011) Pore-scale investigation of colloid retention and mobilization in the presence of a moving air–water interface. *Vadose Zone Journal* 10: 1250–1260, doi:10.2136/vzj2011.0003, with permission.

microscopy methods (Ochiai et al., 2006). Compared to pore-scale experiments, the drawback of these methods is the lower resolution. Consequently, at this scale the studies involved either large particles (e.g., $\sim 5\ \mu\text{m}$, $36\ \mu\text{m}$) or colloidal aggregates (Ochiai et al., 2006).

Field studies of colloid transport are scarce, due to the difficulty in conducting well-controlled experiments in undisturbed soil and potential artifacts related to experimental procedures. Experiments at the field scale typically focus on colloid mobilization or breakthrough measurements in soil lysimeters. Several field studies have attributed the peaks in measured colloid concentration to passage of moving AWIs during infiltration and drainage events (DeNovio et al., 2004; Flury and Aramrak, 2017). Other parameters that affect colloid mobilization in field experiments include changes in solution chemistry (e.g., ionic strength and pH) and hydrodynamic shear stress. In addition to colloid mobilization, field studies on colloid-facilitated transport of contaminants have also been conducted (Sen and Khilar, 2006). However, more field-scale studies are needed in all areas of colloid transport including colloid, biocolloid, micro/nanoplastics, and colloid-facilitated transport.

Factors influencing colloid transport

All factors that govern the processes of colloid deposition and release affect colloid transport in porous media. They include properties of colloids and porous media, solution chemistry (e.g., composition, pH and ionic strength), and hydrodynamics (water content and flow velocity). The effects of these parameters have been studied and documented extensively (Ryan and Elimelech, 1996; Flury and Aramrak, 2017; Shen et al., 2020). Thus, the sections below only cover the factors that have been the focus in more recent studies.

Surface roughness and charge heterogeneity

Colloids of environmental relevance have rough surfaces, thus their behavior in natural environments deviates from predictions by standard theories developed based on colloids with ideal smooth surfaces. Nanoscale protruding asperities act to reduce local repulsive energy barriers thus increase colloid deposition in primary minima under unfavorable conditions. On the other hand, the protruding asperities reduce the depths of primary minima, and accordingly colloids deposited atop the protruding asperities are more readily released by reducing solution ionic strength or increasing flow velocity, leading to reduced colloid deposition in primary minima under favorable conditions (Bizmark et al., 2020). In addition, colloids are favorably deposited at secondary minima when solution ionic strength is low and the secondary minimum wells are enhanced at concave asperities (Shen et al., 2020). Colloids can also deposit at the primary minima in the area close to the vertex of the concave surfaces where the repulsive energy barriers are reduced or even disappear (Fig. 10). Colloid deposition at primary minima at concave locations is irreversible when solution ionic strength is reduced because the primary minimum depth is deep at all ionic strength values (Li et al., 2017).

Surface chemical heterogeneity can also reduce repulsive energy barriers and increase colloid deposition in primary minima under unfavorable conditions. In addition, chemical heterogeneity increases primary minimum depth and thus adhesion, which makes colloid deposition in primary minima less reversible (Shen et al., 2020). For macroscale chemical heterogeneity, the

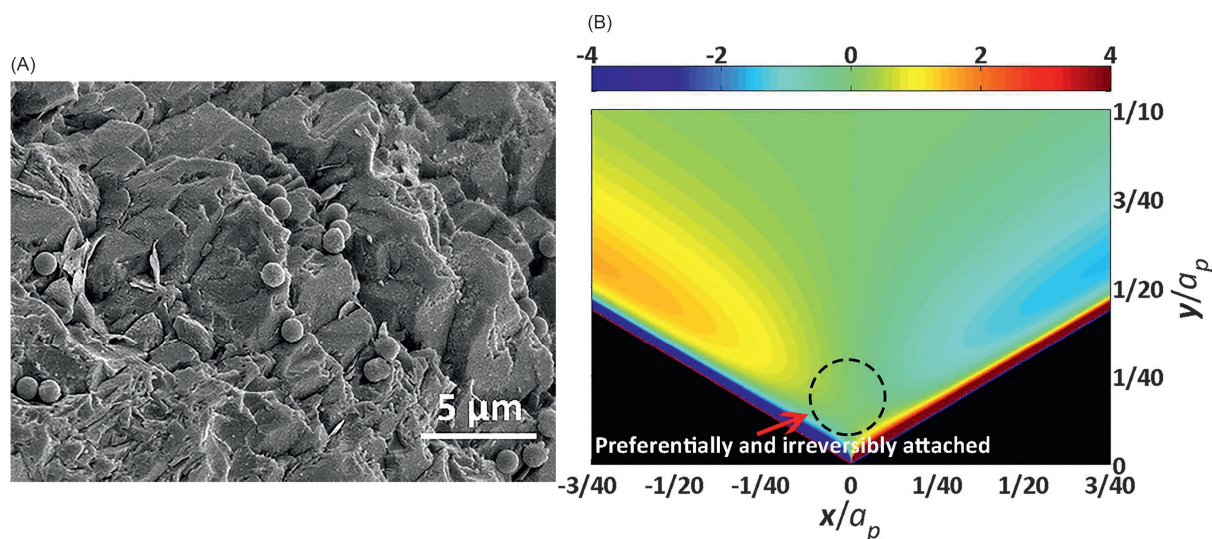


Fig. 10 (A) A scanning electron microscope (SEM) image illustrating the irreversible deposition of colloids at concave sand surface; (B) A DLVO interaction force map showing the absence of energy barrier at the vertex of a concave surface due to intersecting of two half planar surfaces. Figures from Li T, Jin Y, Huang Y, Li B, and Shen C (2017) Observed dependence of colloid detachment on the concentration of initially attached colloids and collector surface heterogeneity in porous media. *Environmental Science & Technology* 51: 2811–2820, with permission.

attachment efficiency is proportional to the fraction of heterogeneity and can be quantified by the patchwise model (Johnson et al., 1996). However, when both micro- and nano-sized chemical heterogeneities are present, the model by Pazmino et al. (2014), which uses trajectory analysis and takes into account torque balance, can be used.

Colloid shape

Colloids are highly variable in shape (e.g., plate-like clays, rod-like goethite and bacteria, thin-wire like carbon nanotubes). The current knowledge on colloid transport in porous media is primarily based on studies that used spherical colloids as models, whereas only few studies have looked at the effect of shape exclusively (Liu et al., 2010; Aramrak et al., 2013). Results from several studies conducted under saturated conditions are not consistent, ranging from enhanced, reduced, and similar retention of rod-shaped colloids compared to spherical colloids (e.g., Ma et al., 2020), which are likely due to the coupled effects of colloid shape, hydrodynamics, and physicochemical properties of colloid and collector surfaces. Non-spherical colloids have different translational and rotational trajectories due to flow field and Brownian motion compared to spherical colloids. Under unsaturated conditions, the presence of AWI and contact line can further alter transport behavior of non-spherical colloids. For example, capillary forces are stronger for particles with edges compared to smooth spherical particles (Flury and Aramrak, 2017).

Modeling colloid transport

Equilibrium adsorption and kinetic deposition

The CFT considers colloid deposition as a first-order kinetic and irreversible process but does not consider deposition at secondary minima (Ryan and Elimelech, 1996). This approach is inadequate for quantifying colloid transport in natural porous media where interactions of colloids and medium surfaces occur under mostly unfavorable conditions where deposition at secondary minima is an important mechanism. New advances include combining equilibrium adsorption equations (e.g., linear isotherm and Langmuir model) and kinetic interaction into the convection-diffusion equation to simulate colloid transport in porous media. In some cases, using an equilibrium adsorption model is warranted for simulating colloid deposition. For example, the presence of nanoscale protruding asperities can cause shallow primary energy wells and thermodynamic equilibrium may establish between the tips of nanoscale protruding asperities and bulk solutions.

Blocking and ripening

Blocking refers to the phenomenon that colloids, previously attached on a collector surface, inhibit subsequent attachment of approaching colloids in their vicinity. Blocking occurs when interactions between the attached and approaching colloids are unfavorable. Blocking can be quantified by conducting column experiments to measure the value of C/C_0 at the plateau of breakthrough curves. The Langmuir approach is most frequently used to describe colloid blocking as

$$\omega = \frac{S_{max} - S}{S_{max}} \quad (16)$$

where ω is fraction of porous medium available for deposition, S_{max} is colloid retention capacity, S is colloid concentration on the solid phase of a porous media. The use of Langmuir function to simulate the blocking process has been debated because it displays a linear dependence on fractional surface coverage while microscopic examinations have shown that nonlinear relationship exists when surface coverage is high (Shen et al., 2020). Alternatively, the random sequential adsorption (RSA) model has been adopted to account for the nonlinear blocking effects. The classic RSA model delineates irreversible deposition of hard spherical colloids on a homogeneous surface in a monolayer, which is written as

$$\omega = 1 + a_1 \left(\frac{S}{S_{max}} \right) + a_2 \left(\frac{S}{S_{max}} \right)^2 + a_3 \left(\frac{S}{S_{max}} \right)^3 + \dots \quad (17)$$

where a_1 , a_2 , and a_3 are -4 , $6\sqrt{3}/\pi$, and 1.404 , respectively. The RSA model has been modified later by considering colloid shape, Brownian diffusion, and colloidal forces (e.g., attractive colloid-surface DL interaction, repulsive colloid-colloid DL interaction, and gravitational force).

Ripening is opposite to blocking, referring to the process of deposited colloids on a collector surface favoring subsequent colloid deposition. Ripening occurs when the colloid-colloid interaction is more favorable compared to colloid-surface interactions. Ripening results in an increase of deposition rate with time and is denoted by a descending plateau in breakthrough curves from column experiments. To date, ripening can only be quantified using empirical models such as the second-order kinetics model. The mass balance equation for colloid deposition via ripening as a second-order process is expressed as

$$\frac{\rho_b}{f} \frac{\partial S}{\partial t} = k_r SC \quad (18)$$

where ρ_b is bulk density and k_r is colloid interaction rate coefficient.

Transport models

Transport of colloids in porous media has been simulated by abstract, mechanistic, and continuum scale models (Babakhani et al., 2017). Material flow analysis and multi-media models are abstract models developed based on the mass balance principle covering scales from global down to local levels. Most abstract models use simple algebraic equations, and only a few attempts have been made to address the influence of various factors on colloid transport.

Different from solutes that exhibit random motion, colloids have limited diffusive motion and possess largely deterministic trajectories. Therefore, the likelihood of colloids reaching collector surfaces can be mechanistically predicted. Lagrangian and Eulerian methods are the two most common mechanistic approaches for calculating the single collector removal efficiency η at the pore scale. Lagrangian methods describe trajectories of colloids governed using Newton's second law by considering the forces and torques acting on the colloids. Eulerian methods track colloid concentration distribution in time and space. For colloid transport and deposition in fluids, the distribution of colloid concentration is described by the convective-diffusion equation. The calculated single collector removal efficiency at pore scale is related to the deposition rate at continuum scale via upscaling, shown in detail below.

In continuum-scale models, microscopic properties are averaged over a representative elementary volume (REV) and the average property such as porosity and colloid concentration is treated as a value that represents the property throughout the REV (Molnar et al., 2015). Based on this assumption, the mass balance equations are solved at the macroscopic scale and partial differential equations that describe spatial and temporal rate of change of the key macroscopic properties are obtained. For example, the advection-dispersion equation with incorporation of terms (e.g., partition, deposition, and straining) has been frequently used to simulate the transport. Particularly, dual-deposition rate models and dual-region or multiporosity models have been developed. The values of parameters in these terms can be obtained by using the models to simulate data from experiments such as batch and column examinations.

Upscaling

The pore scale parameter η_p quantifies the transport of colloids from bulk solution to the vicinity of a collector surface and subsequent immobilization on a collector surface. The pore scale parameter η_p determined via the mechanistic models can be related to deposition rate k_d at continuum scale via the following expression

$$k_d = \frac{3}{2} \frac{(1-f)}{d_c f} U \eta_p \quad (19)$$

where d_c is collector diameter. Eq. (19) assumes that the control volume is depicted by a collection of identical pore-scale collectors consisting of the same volume and possessing the same porosity as the control volume (Molnar et al., 2015). This assumption has been validated by experimental observations. However, upscaling using Eq. (19) faces challenges under unfavorable conditions because of the accumulation of colloids in the near surface domain (e.g., via secondary-minimum association). The accumulation in the near surface domain can cause extended tailing in colloid breakthrough curves during elution in column transport experiments through: (1) long residence times of colloids in the near-surface fluid domain and (2) transfer of colloids in the near surface domain from upstream to downstream collectors (Johnson and Hilpert, 2013). The upscaling is further complicated at field scale due to the high heterogeneity of subsurface environments and the interaction of various transport processes (e.g., aggregation, deposition, release, straining, and blocking).

Conclusion

Colloid transport, the movement of particles of sizes from 1 nm to 10 μm , in soils and groundwater aquifers, plays crucial roles in many agricultural, environmental and engineering processes. These processes include transport of chemical pollutants, pesticides, fertilizers, microbial pathogens, manufactured nanoparticles and micro- and nanoplastics, as well as beneficial management practices such as water purification and filtration, groundwater remediation using microbes and nanoparticles, and petroleum reservoir production. Research on colloid transport through porous media in the last ~ 35 years has significantly improved our overall understanding. New advances are particularly notable in the following areas: (1) identification of colloid interactions in secondary minima and straining as important mechanisms of colloid retention in addition to deposition in primary minima under unfavorable conditions; (2) revealing mechanisms that control short-range forces and incorporation of the short-range forces into the classical DLVO theory to account for discrepancies between predictions from traditional theories and experimental observations; (3) development of theoretical frameworks to account for physical and chemical heterogeneity of colloids and collector surfaces; (4) initiation and making significant progress on research of colloid retention and transport in unsaturated media by identifying the roles that the AWI and the contact line play; (5) employment of advanced instruments and pore-scale investigations to significantly improve mechanistic understanding on the interactions and processes involved in colloid transport; (6) expanding the scope of colloid transport research from soil colloid mobilization, colloid-facilitated contaminant transport to include transport of microbial pathogens, manufactured NPs and micro- and nanoplastics.

Despite the significant advances, challenges and knowledge gaps remain. Models that are used to quantify colloid transport at realistic scales are largely empirical due to our incomplete understanding of the mechanisms involved, our inability to experimentally validate many of the assumptions used in these models, as well as the extremely complex nature of the subsurface porous media. These will be topics under continued investigation. Moreover, new areas of interest related to colloid transport in porous media will undoubtedly emerge. For example, the stability of soil organic carbon and its transport has been found to be closely related to its ability to complex with soil minerals, suggesting a potentially very important role that soil colloids may play in global carbon cycling. Another relevant topic that has attracted intense research attention is the fate and transport of micro- and nanoplastics in soil due to their widespread pollution in terrestrial systems. Therefore, colloid transport in porous media and its many applications will remain an active research topic in the foreseeable future.

References

- Afsar MZ, Goodwin C, Beebe TP Jr., Jaisi DP, and Jin Y (2020) Quantification and molecular characterization of organo-mineral associations as influenced by redox oscillations. *Science of the Total Environment* 704: 1354549. <https://doi.org/10.1016/j.scitotenv.2019.135454>.
- Aramrak S, Flury M, and Harsh JB (2011) Detachment of deposited colloids by advancing and receding air-water interfaces. *Langmuir* 27: 9985–9993.
- Aramrak S, Flury M, Harsh JB, Zollars RL, and Davis HP (2013) Does colloid shape affect detachment of colloids by a moving air–water interface? *Langmuir* 29(19): 5770–5780. <https://doi.org/10.1021/la400252q>.
- Babakhani P, Bridge J, Doong R, and Phenrat T (2017) Continuum-based models and concepts for the transport of nanoparticles in saturated porous media: A state-of-the-science review. *Advances in Colloid and Interface Science* 246: 75–104.
- Bergendahl J and Grasso D (2000) Prediction of colloid detachment in a model porous media: Hydrodynamics. *Chemical Engineering Science* 55: 1523–1532.
- Bhattacharjee S and Elimelech M (1997) Surface element integration: A novel technique for evaluation of DLVO interaction between a particle and a flat plate. *Journal of Colloid and Interface Science* 193: 273–285.
- Bizmark N, Schneider J, Priestley RD, and Datta SS (2020) Multiscale dynamics of colloidal deposition and erosion in porous media. *Science Advances* 6: eabc2530.
- Bradford SA, Simunek J, Bettahar M, van Genuchten MT, and Yates SR (2006) Significance of straining in colloid deposition: Evidence and implications. *Water Resources Research* 42: W12S15. <https://doi.org/10.1029/2005WR004791>.
- DeNovio NM, Saiers JE, and Ryan JN (2004) Colloid movement in unsaturated porous media: Recent advances and future directions. *Vadose Zone Journal* 3: 338–351. <https://doi.org/10.2136/vzj2004.0338>.
- Flury M and Aramrak S (2017) Role of air-water interfaces in colloid transport in porous media: A review. *Water Resources Research* 53: 5247–5257. <https://doi.org/10.1002/2017WR020597>.
- Gao B, Steenhuis TS, Zevi Y, Morales VL, Nieber JL, Richards BK, McCarthy JF, and Parlange J-Y (2008) Capillary retention of colloids in unsaturated porous media. *Water Resources Research* 44: W04504. <https://doi.org/10.1029/2006WR005332>.
- Hahn MW, Abadiz D, and O'Melia CR (2004) Aquasols: On the role of secondary minima. *Environmental Science & Technology* 38: 5915–5924.
- Hamaker HC (1937) The London-van der Waals attraction between spherical particles. *Physica* 4: 1058–1072.
- Hogg R, Healy TW, and Fuerstenau DW (1966) Mutual coagulation of colloidal dispersions. *Transactions of the Faraday Society* 62: 1638–1651.
- Israelachvili J and Pashley R (1982) The hydrophobic interaction is long range, decaying exponentially with distance. *Nature* 300: 341–342.
- Jin Y and Flury M (2002) Fate and transport of viruses in porous media. *Advances in Agronomy* 77: 39–102.
- Johnson WP and Hilpert M (2013) Upscaling colloid transport and retention under unfavorable conditions: Linking mass transfer to pore and grain topology. *Water Resources Research* 49: 5328–5341.
- Johnson PR, Sun N, and Elimelech M (1996) Colloid transport in geochemically heterogeneous porous media: Modeling and measurements. *Environmental Science & Technology* 30: 3284–3293.
- Kretzschmar R and Schafer T (2005) Metal retention and transport on colloidal particles in the environment. *Elements* 1: 205–210.
- Lazouskaya V, Jin Y, and Or D (2006) Interfacial interactions and colloid retention under steady flows in a capillary channel. *Journal of Colloid and Interface Science* 303: 171–184.
- Lazouskaya V, Wang L, Gao H, Shi X, Czymmek K, and Jin Y (2011) Pore-scale investigation of colloid retention and mobilization in the presence of a moving air–water interface. *Vadose Zone Journal* 10: 1250–1260. <https://doi.org/10.2136/vzj2011.0003>.
- Lazouskaya V, Wang LP, Or D, Wang G, Caplan JL, and Jin Y (2013) Colloid mobilization by fluid displacement fronts in channels. *Journal of Colloid and Interface Science* 406: 44–50. <https://doi.org/10.1016/j.jcis.2013.05.078>.
- Li T, Jin Y, Huang Y, Li B, and Shen C (2017) Observed dependence of colloid detachment on the concentration of initially attached colloids and collector surface heterogeneity in porous media. *Environmental Science & Technology* 51: 2811–2820.
- Lin S and Wiesner MR (2010) Exact analytical expressions for the potential of electrical double layer interactions for a sphere-plate system. *Langmuir* 26: 16638–16641.
- Liu Q, Lazouskaya V, He Q, and Jin Y (2010) Effect of particle shape on colloid retention and release in saturated porous media. *Journal of Environmental Quality* 39: 500–508.
- Ma H, Bolster C, Johnson WP, Li K, Pazmino E, Camacho KM, Anselmo AC, and Mitragotri S (2020) Coupled influences of particle shape, surface property and flow hydrodynamics on rod-shaped colloid transport in porous media. *Journal of Colloid and Interface Science* 577: 471–480. <https://doi.org/10.1016/j.jcis.2020.05.022>.
- McCarthy JF and McKay LD (2004) Colloid transport in the subsurface: Past, present, and future challenges. *Vadose Zone Journal* 3: 326–337. <https://doi.org/10.2136/vzj2004.0326>.
- McCarthy JF and Zachara JM (1989) Subsurface transport of contaminants. *Environmental Science & Technology* 23: 496–502.
- Michel E, Majdalani S, and Di-Pietro L (2010) How differential capillary stresses promote particle mobilization in macroporous soils: a novel conceptual model. *Vadose Zone Journal* 9: 307–316. <https://doi.org/10.2136/vzj2009.0084>.
- Molnar IL, Johnson WP, Gerhardt JL, Willson CS, and O'Carroll DM (2015) Predicting colloid transport through saturated porous media: A critical review. *Water Resources Research* 51: 6804–6845.
- Ochiai N, Kraft EL, and Selker JS (2006) Methods for colloid transport visualization in pore networks. *Water Resources Research* 42: W12S06. <https://doi.org/10.1029/2006WR004961>.
- Pazmino E, Trauscht J, Dame B, and Johnson WP (2014) Power law size-distributed heterogeneity explains colloid retention on soda lime glass in the presence of energy barriers. *Langmuir* 30: 5412–5421.
- Rajagopalan R and Tien C (1976) Trajectory analysis of deep-bed filtration with sphere-in-cell porous-media model. *AIChE Journal* 22: 523–533.
- Ryan JN and Elimelech M (1996) Colloid mobilization and transport in groundwater. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* 107: 1–56.
- Schijven JF and Hassanizadeh SM (2000) Removal of viruses by soil passage: Overview of modeling, processes, and parameters. *Critical Reviews in Environmental Science and Technology* 30: 49–127. <https://doi.org/10.1080/10643380091184174>.
- Sen TK and Khilar KC (2006) Review on subsurface colloids and colloid-associated contaminant transport in saturated porous media. *Advances in Colloid and Interface Science* 119(2–3): 71–96. <https://doi.org/10.1016/j.cis.2005.09.001>.

- Shen C, Wang F, Li B, Jin Y, Wang L-P, and Huang Y (2012) Application of DLVO energy map to evaluate interactions between spherical colloids and rough surfaces. *Langmuir* 28: 14681–14692.
- Shen C, Jin Y, Zhuang J, Li T, and Xing B (2020) Role and importance of surface heterogeneities in transport of particles in saturated porous media. *Critical Reviews in Environmental Science and Technology* 50: 244–329.
- Tufenkji N and Elimelech M (2004) Correlation equation for predicting single-collector efficiency in physicochemical filtration in saturated porous media. *Environmental Science & Technology* 38: 529–536.
- Verwey EJW and Overbeek JTG (1948) *Theory of the Stability of Lyophobic Colloids*. Amsterdam: Elsevier.
- Wan J and Tokunaga TK (1997) Film straining of colloids in unsaturated porous media: Conceptual model and experimental testing. *Environmental Science & Technology* 31(8): 2413–2420. <https://doi.org/10.1021/es970017q>.
- Yan J, Maneski R, Vasilas B, and Jin Y (2018) Mobile colloids and colloidal organic carbon: An underestimated carbon pool in global carbon cycles? *Frontiers in Environmental Science* 6: 148. <https://doi.org/10.3389/fenvs.2018.00148>.
- Yao K-M, Habibian MT, and O'Melia CR (1971) Water and waste water filtration: Concepts and applications. *Environmental Science & Technology* 5: 1105–1112.
- Yu Y and Flury M (2021) Current understanding of micro- and nanoplastics in soil. *Vadose Zone Journal* 20: e20108. <https://doi.org/10.1002/vzj2.20108>.

Scaling of transport properties

Robert Ewing^a and Behzad Ghanbarian^b, ^aThe Climate Corporation, Seattle, WA, United States; ^bKansas State University, Manhattan, KS, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of R.P. Ewing, SCALING I Transport Processes, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 477-485, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00570-1>.

Introduction	354
Scale	354
Gaussian solute transport	355
Spreading of solute in a tube	356
The advection-dispersion equation (ADE)	357
Alternatives to the ADE	357
Mobile-immobile (MIM)	358
Transfer functions	359
Fractional ADE (FADE)	359
Continuous-time random walk (CTRW)	360
Percolation theory	361
Observations and mechanisms	361
Acknowledgments	362
References	362

Abstract

Solute transport is generally assumed to be as described by the advection-dispersion equation (ADE). But a wide range of observations leads to the conclusion that dispersion increases with scale (distance or time). That is, fitting the ADE to early-time dispersion data and extrapolating to later time will almost always underestimate the data that are (eventually) collected at that late time. This inconvenient finding has motivated extensive research and model development. Nonetheless, predictions of future dispersion frequently get the scaling wrong.

Glossary

ADE Advection-dispersion equation, a model of dispersive transport in which the solute travels at the mean fluid velocity, and disperses symmetrically as if it were diffusing.

BTC Breakthrough curve.

CLT Convective lognormal transfer, a specific kind of streamtube model, in which the distribution of streamtube velocities is given a lognormal form.

CTRW Continuous time random walk, a model of dispersive transport in which solute jumps between streamlines, with jump probability and timing governed by ancillary functions.

Dispersivity Dispersion coefficient divided by velocity.

FADE Fractional advection-dispersion equation, a generalization of the ADE using fractional calculus.

MIM Mobile-immobile, a model of dispersive transport in which transport takes place in the mobile region, with solute diffusing between the mobile and immobile regions.

MRMT Multi-rate mass transfer, a generalization of the MIM with many immobile regions, each exchanging solute with the mobile region at different rates.

PT Percolation theory, a conceptual model of porous materials and the pathways through them.

Streamtube model A transport model in which solute molecules flow along streamlines or streamtubes, and do not jump from one to another.

Transfer function See streamtube model.

Key points

1. Solute transport is commonly assumed to be as described by the advection-dispersion equation (ADE).
2. Dispersion is often observed to increase with scale (distance or time).
3. Many explanations and models have been proposed for this scale-dependence of dispersion.
4. We generally do not know in advance what model will best describe how dispersion in a given situation will evolve with time and/or distance.

Introduction

'Scaling of transport properties' refers to the common and often troublesome observation that the model parameters describing transport at one scale are frequently incorrect at another scale. The word 'scale' denotes extent, usually in space or time. Thus, to say that 'transport *scales*' is to say that different results are obtained from measurements made at different spatial or temporal scales. The most common situation is that one measures dispersion (using whatever metric one chooses) at a small scale and predicts it at a larger scale, only to find that the actual dispersion as measured later at the larger scale is greater than predicted. Dispersion coefficients compiled from many separate experiments tend to increase with distance. How a model parameter (such as the dispersion coefficient) changes with scale need not be linear: the exponent governing the coefficient's change with time or distance need not equal unity. Is this scaling intrinsic to the transport process itself, a function of the media, or both?

The concept of scaling is historically grounded in widespread use of the advection-dispersion equation (ADE), which uses a diffusion analogy to describe dispersive transport. To say that 'transport scales' often simply means that it does not conform to a simple diffusion-like description. Scaling of dispersion was initially thought to be an exceptional case, sometimes called 'anomalous dispersion,' implying that dispersion which conforms to the ADE is the norm. But so-called anomalous transport is increasingly understood as typical, with dispersion that conforms to the ADE ('Gaussian dispersion' or 'Fickian transport') now being seen as the special case.

This article discusses the simple case of one-dimensional longitudinal transport of a nonreactive, neutral, conservative solute during steady isothermal saturated flow, ignoring any density and viscosity differences between the native fluid (for example, the original soil pore water) and the 'new' fluid (say, soil pore water with a tracer or pollutant added). This simple case is already sufficiently difficult, and yet the reader is cautioned that real-world situations also include reactive, degrading, sorbing, or charged solutes moving non-isothermally and transiently under multiple driving and dispersing forces.

Scale

Soil science and hydrology are concerned with processes taking place from the micrometer scale to the kilometer scale and beyond, and from the essentially instantaneous to a geologic timescale. It is common to speak of scales in a colloquial manner—'pore-scale,' 'lab-scale,' 'overnight'—and the mental images that these terms invoke can be useful, although they are not quantitative. Less useful are terms such as 'microscale,' which carry widely different meanings between (and even within!) disciplines. Because soil science is inherently interdisciplinary, such words often have different meanings even to different soil scientists. If quantitative terms are needed, then using the size itself—for example, 'decimeter-scale'—can reduce ambiguity.

Some properties are specific to a defined range of scales. To illustrate, suppose one wants to measure the porosity of a sand using computed tomography (CT). If the CT unit measures with a resolution of 0.01 mm, then most voxels (volume elements from the CT scan) will represent either the inside of a solid particle or the center of a pore. In neither case will it be useful to speak of the porosity of an individual voxel: porosity is a property of a volume and becomes meaningless as that volume shrinks to a point. If we coarsen the scale slightly, each individual voxel will sample just a few grains and pores, and so individual measurements will be highly variable. But when porosity is measured over a sufficiently large number of grains and pores (at least 30 in each direction is common guideline), the measured value is stable. This illustrates the concept of the representative elementary volume (REV): the smallest volume over which a given property can be measured with reasonable confidence. Notice that the REV must be larger than the characteristic scale of variability of the medium (if one exists).

We must also distinguish between scale and level. Scale is strictly about physical extent, while level involves the emergence of new properties. The two are easily confused, because changes in physical size often entail changes in level also. As a trivial example, a soil core 10 cm deep may come from a single horizon, but 'scaling up' to a 20-cm core may yield a sample with two distinct horizons in it—a difference in level as well as scale. Likewise, scaling up from a 10-cm core to a pedon may introduce fractures or wormholes into the system, and so, rather than studying a soil matrix, one now faces a macropore-matrix or dual-permeability system. The distinction between scale and level may be quite subtle. For example, buoyancy effects may not be evident during fast flow of a weak salt solution through a 1-cm-diameter soil column, but they may be evident during slower flow through a larger-diameter soil column. In the second case, the increase in size (scale) allows the emergence of a previously negligible mechanism (level), so the second system under study is not equivalent to the first. It can be useful to compare relevant dimensionless groups to help discern

whether changing the scale also allows new processes to come into play. Relevant dimensionless groups may include the Péclet number (ratio of advective to diffusive forces), the capillary number (ratio of viscous to capillary forces), the Bond number (ratio of buoyancy to capillary forces), the Reynolds number (ratio of inertial to viscous forces), and the Schmidt number (ratio of kinematic viscosity to molecular diffusivity).

A further difficulty is that many parameters of interest can each be best measured within a fairly narrow range of scales. For example, dispersion is generally estimated from solute concentration, based upon samples at the scale of a liter or less. Measuring solute concentration over a large area is generally accomplished by collecting many small samples, because collecting the entire outflow of a large area is impractical if not impossible. So studies of dispersion generally operate at many scales: the scale of the individual observations (where different observations may themselves have different scales), the scale of the current composite (for example, the inferred pollutant plume based upon some 100 individual core samples), the scale of the desired prediction, and the characteristic scale, if any, of the soil's heterogeneity. While we are mainly concerned here with the disparity between the middle two (scale of the current composite and the scale of the desired prediction), any difference in scale may contribute to scaling.

Gaussian solute transport

Hydrodynamic dispersion (for convenience here called simply 'dispersion') emerges from the interplay of two mechanisms: mechanical dispersion and molecular diffusion. In simplest terms, mechanical dispersion is caused by the medium, while diffusion is inherent to the fluid. Mechanical dispersion results from differences in travel time along different streamlines. Streamlines differ in velocity due to heterogeneity of the medium, such as variations in pore size and connectivity, and the presence of structure across multiple scales from pore-scale correlations to larger-scale layering and fractures. When flow is rapid, mechanical dispersion contributes most of the total dispersion, while at slow flow rates diffusion is also evident.

Diffusion as a macroscopic phenomenon results from the random motion of solvent and solute molecules. The diffusion equation describes net solute flux as being proportional to, and opposite in direction from, the concentration gradient. In one dimension and in bulk fluid (no solid phase):

$$\frac{\partial C}{\partial t} = -D_{\text{aq}} \frac{\partial^2 C}{\partial x^2} \quad (1)$$

where C is solute concentration (M L^{-3}), t is time (T), D_{aq} is the diffusion coefficient ($\text{L}^2 \text{T}^{-1}$) of the solute in water, and x is distance (L). In the presence of a solid phase, the molecular diffusion coefficient can be adjusted by the porosity ϕ of the medium and the tortuosity τ of the diffusive path length, giving an effective diffusion coefficient: $D_{\text{eff}} = D_{\text{aq}} \phi / \tau$.

Dispersion is often estimated using data from an experiment in which a tracer is introduced into steadily flowing water. That source may be an instantaneous spike or a step input; for this illustration we assume a step. At a point downstream from where tracer was introduced, a steady increase in the solute's flux concentration is seen as the native water is replaced by water with the tracer (If instead the tracer's resident concentration were measured at multiple points inside the sample at a single instant, the concentration would be seen to decrease with distance downstream from its peak at the point of introduction.). The output of a flux concentration dispersion experiment can be summarized graphically using a break-through curve (BTC) (Fig. 1). For a laboratory experiment, the x-axis is typically presented in units of pore volumes, the amount of water needed to replace the original water in the sample precisely once. The y-axis is expressed as the ratio of the input concentration to the peak concentration, C/C_0 . This normalization of the two axes allows curves from different experiments (e.g., using different flow rates or column lengths) to be compared.

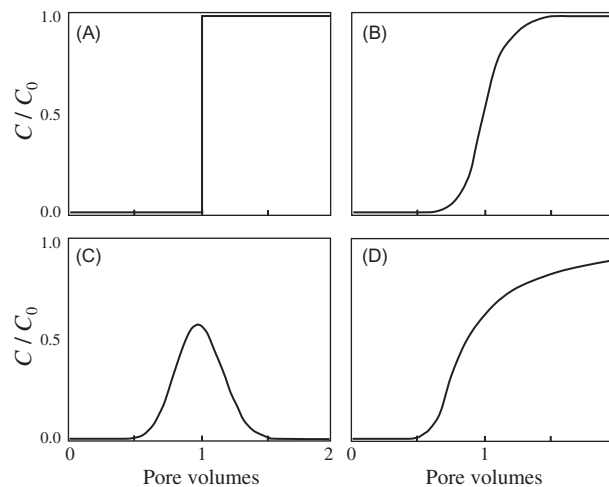


Fig. 1 Sample flux concentration breakthrough curves for: (A) a step input with piston flow; (B) a step input with Gaussian dispersion; (C) a Dirac (pulse) input with Gaussian dispersion; and (D) non-Gaussian dispersion of a step input. See text for explanation.

Spreading of solute in a tube

Griffiths (1910) showed that a drop of dye, placed into water flowing through a glass tube, moves at the mean water velocity, with its dispersion along the tube being symmetrical. Taylor's (1953) subsequent analysis showed that radial diffusion (normal to the direction of flow) explained the movement of solute molecules between the fast-flowing tube center and the negligible velocities near the wall. While diffusion is only one component of dispersion, Taylor showed that dispersion of a solute in fluid flowing steadily through a tube can—under certain conditions—be precisely described by superimposing the mean fluid velocity on a diffusion-like dispersion term. The main condition is simply that flow be sufficiently slow that solute molecules have time to diffuse freely between fast (tube center) and slow (tube wall) regions, a condition that is also well approximated, at least at the pore scale, under natural conditions in many soils and aquifers.

Taylor's papers on longitudinal dispersion of solutes in a tube, combined with further analysis by Aris (1956), shows that the dispersion behaves like diffusion, in the sense that it yields a Gaussian concentration profile the width of which grows in proportion to the square root of time (Fig. 2). Aris's theoretical expression for the dispersion coefficient D ,

$$D = D_{\text{aq}} + \frac{a^2 v_0^2}{192 D_{\text{aq}}} \quad (2)$$

where a is the tube radius (L) and v_0 the velocity (L T^{-1}) in the tube's center, gives the dispersion coefficient as the sum of diffusive and mechanical dispersive forces. For this equation to hold, $a^2/4D_{\text{aq}} t$ must be small: there must be sufficient time for radial diffusion to move molecules a distance equal to the radius of the tube. In terms of the Péclet number, this condition may be expressed as $Pe \ll 4 l/a$, where l is the distance from the dye injection point at which the measurement is made. When the elapsed time is too short, radial diffusion is insignificant and dispersion can be approximated via the parabolic velocity profile. Because for any given velocity and tube radius the time condition does not hold for some short time or distance, short times and/or distances are called 'preasymptotic.'

The conditions for Taylor's analysis can be shown by plotting the ratio of total dispersion to diffusion as a function of the Péclet number (Fig. 3). The solid line represents the Aris equation (Eq. 3), which shows different behavior in each of three regions; the dashed lines show the behavior that would be seen if the time condition were violated. In the first region, diffusion is the main contributor to total dispersion. In the second region, the slope of the line is approximately 2. The Péclet number above which the time condition fails, $Pe > 4 l/a$, is shown as a gray region rather than a single line because it changes for different l . In this third, preasymptotic region, dispersion is linearly related to velocity.

Taylor's and Aris's analyses established much of how both engineers and soil scientists now think about dispersion. The analogy thereby established between dispersion and diffusion has been broadly applied to solute movement in porous media. The analogy is particularly attractive because the diffusion equation has been widely studied, and operating within this analogy gives access to preexisting mathematical solutions. Concepts such as a particular aquifer test having preasymptotic flow conditions also derive from this history.

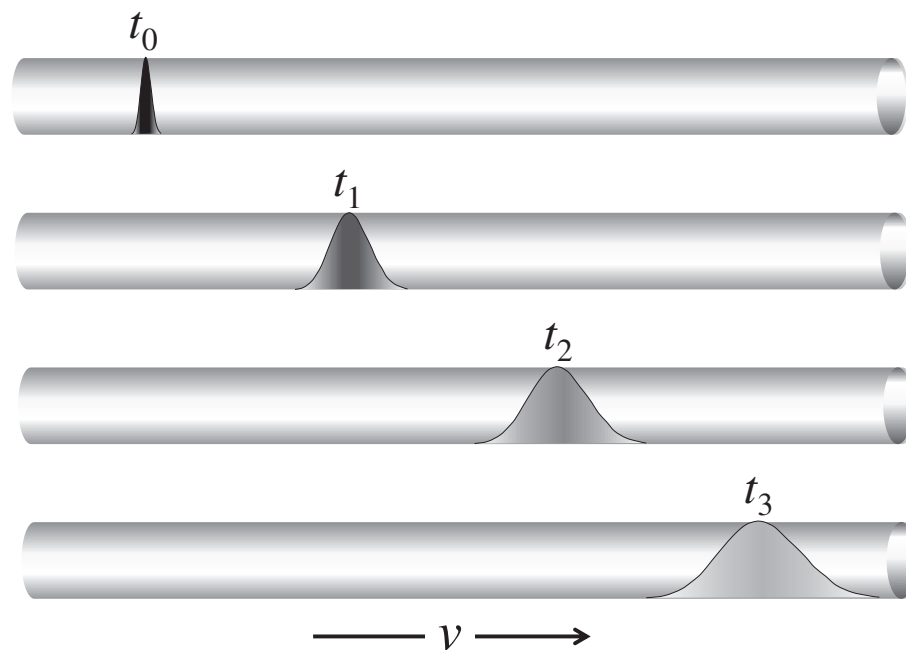


Fig. 2 Schematic of solute flowing and dispersing with distance in a tube into which it was introduced at a single instant.

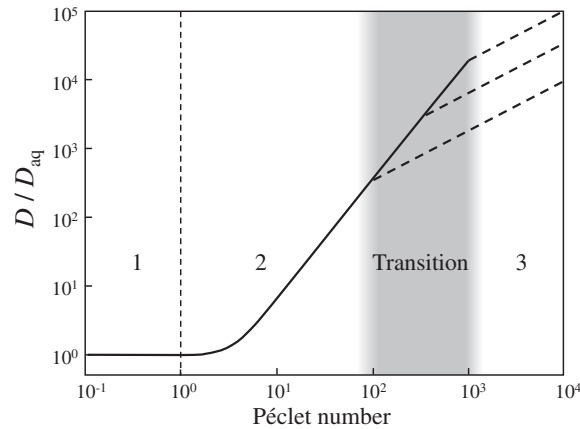


Fig. 3 Mechanical dispersion for flow through a tube, divided by diffusion, as a function of the Péclet number. Above some Péclet number whose exact value varies with length, dispersion transitions to being linear with time; below that value, dispersion scales with the square of time. See text for explanation.

The advection-dispersion equation (ADE)

Superimposing a mean velocity on the diffusion equation yields the ADE, also called the convection–dispersion equation (CDE). A one-dimensional form of the ADE for nonreactive solutes describes the change in concentration as resulting from the sum of dispersive and advective terms:

$$\frac{\partial C}{\partial t} = D \frac{\partial^2 C}{\partial x^2} - v \frac{\partial C}{\partial x} \quad (3)$$

where v is the average linear water velocity (L T^{-1}). An additional pre-factor R (retardation) for the lefthand side accounts for solute moving slower or faster than the water, generally caused by sorption. Analytical solutions to this equation have been derived for a wide variety of initial and boundary conditions, and several computer programs are also available. In a simple case, a point injection of a solute will move at the mean water velocity, while spreading as if it were diffusing with a diffusion coefficient equal in magnitude to the dispersion coefficient D . Even in soils and aquifers where dispersion is well described by the ADE, non-Fickian (or non-Gaussian) behavior is often seen at early times or short distances, then increasingly Fickian behavior will be seen with increasing distance or time. Accordingly, non-Fickian behavior is sometimes called ‘preasymptotic,’ implying that all transport eventually becomes Fickian. However, non-Fickian output is sometimes observed to persist across many length scales.

Dispersion is often observed to increase approximately linearly with fluid velocity. The ADE can in this case be generalized with respect to velocity by defining a dispersivity $\alpha = D/v$ (with units of L), and replacing the dispersion coefficient D in the ADE with α . This dependence on velocity may seem counterintuitive. If dispersion is the sum of molecular diffusion and mechanical dispersion, then mechanical dispersion (which results from the medium) should be independent of velocity, whereas slower flow (having more time) should result in greater diffusion-based dispersion. But as Taylor showed, D increases linearly with v within a specific range of Péclet numbers, so dispersion increasing with velocity is reasonable.

The assumption that dispersion can be treated mathematically as if it were diffusion often leads to a poor match between model and data. This is frequently because one or more of the equation’s assumptions have been violated: probably either the medium is not uniform (let alone tubular!) or there is insufficient exchange between solute molecules traveling on different streamlines. Interestingly, the ADE is unable to account correctly for dispersion resulting from vertical flow through layered soils (conceptually equivalent to different tubes in series, as shown in Fig. 4B): it predicts identical results regardless of which soil comes first, while experimentally the order of the layers affects the final value of the dispersion coefficient.

Because diffusion and flow are not strictly additive with respect to dispersion except under specific conditions, the basic concept of the ADE is not sufficiently general for many situations found in the real world. Much recent research on solute dispersion reveals struggles with the poor predictions produced by the ADE, frequently focusing on developing alternative conceptual models. In addition, the classical advection-dispersion (ADE) equation predicts that the dispersivity should be independent of length scale, while both lab and field experimental measurements imply the opposite (Molz et al., 1983; Neuman, 1990; Xu and Eckstein, 1995).

Alternatives to the ADE

Solute transport is often poorly described by the ADE, particularly when a porous medium is strongly heterogeneous. In heterogeneous formations, due to spatial heterogeneity, solute migration is characterized by features that do not conform to Fickian transport (Neuman and Zhang, 1990; Berkowitz et al., 2006). But even in homogeneous media such as sand packs, there is

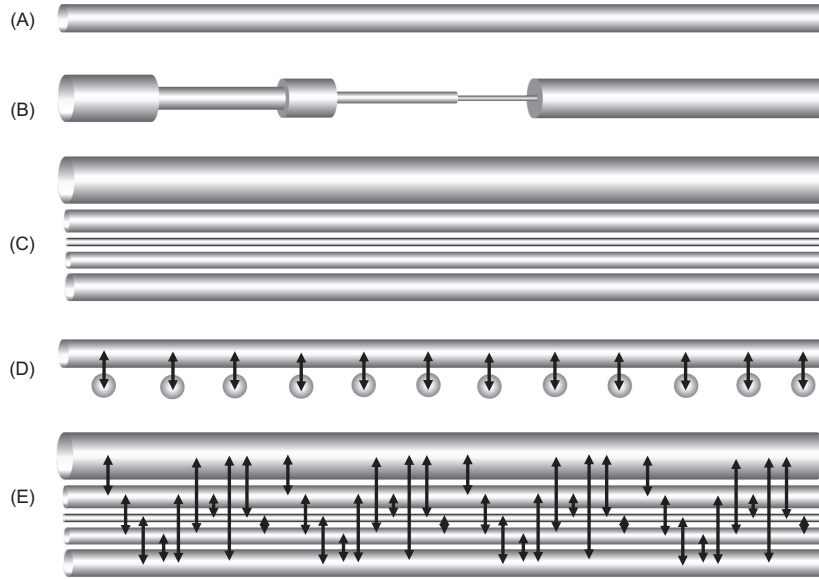


Fig. 4 (A) A single tube with constant cross section, analogous to the ADE; (B) a tube with varying cross section, analogous to flow through a layered soil; (C) a bundle of tubes of different but constant cross sections, analogous to the CLT; (D) a single tube with dead-end 'pores', analogous to the MIM; and (E) a bundle of tubes of different cross sections with solute exchange between tubes, analogous to the CTRW.

evidence (Scheidegger, 1959; Lévy and Berkowitz, 2003; Bromly and Hinz, 2004) that solute transport displays non-Fickian behavior.

Alternatives to the ADE fall into two distinct groups: those that are built explicitly on the foundation of the ADE (dispersivity, MIM and MRMT, FADE), and those of independent provenance (CTRW, PT). All of these models other than PT are mathematically related. For instance, Dentz and Berkowitz (2003) demonstrated that there exists an analytical relationship between CTRW and MRMT. Metzler and Klafter (2000) showed that the time- and space-fractional ADEs are limit forms of CTRW. Furthermore, the fractional mobile-immobile model (FMIM) can be derived from the MRMT with a special memory function (Schumer et al., 2003). Finally, Sun et al. (2020) clarified the mathematical connections between CTRW, multi-rate mass transfer (MRMT), and fractional advection dispersion equations (FADE).

Mobile-immobile (MIM)

The mobile-immobile (MIM) model was first introduced in the petroleum engineering literature to explain dispersion in the presence of dead-end pores. It was adopted by van Genuchten and Wierenga (1976) in the soil physics community to describe dispersion in structured soils. The concept is that the soil is composed of mobile regions (such as interaggregate pores) through which most or all of the flow actually occurs, which are interconnected with immobile regions (e.g., the interior of aggregates) that exchange mass with mobile regions by diffusion only (Fig. 4D). The model can be considered to describe chemical sorption that is instantaneous at some sites and slow at others. The MIM model divides the domain into high- and low-permeability zones with appropriate coupling terms to account for the exchange of water and/or contaminants.

$$f_m \theta \frac{\partial C_m}{\partial t} + (1 - f_m) \theta \frac{\partial C_{im}}{\partial t} = f_m \theta D_m \frac{\partial^2 C_m}{\partial x^2} - v_m f_m \theta \frac{\partial C_m}{\partial x} \quad (4a)$$

$$(1 - f_m) \theta \frac{\partial C_{im}}{\partial t} = \mu (C_m - C_{im}) \quad (4b)$$

where θ is the volumetric water content, f is a water fraction (either mobile or immobile), the subscripts m and im refer respectively to the mobile and immobile zones, and μ is a first-order mass transfer coefficient ($L^2 T^{-1}$).

The MIM model given in (Eqs. 4a and 4b) is based on several assumptions, chief of which are (i) advective transport is neglected in the low-permeability zone, and (ii) solute exchange between the mobile and immobile zones is described via a quasi-empirical first-order rate term proportional to the difference between the average concentrations of the mobile and immobile zones.

Because diffusive exchange between regions is generally slow relative to advection, a given layer of soil is generally out of equilibrium with the solute flowing through it for some time, so this model is said to describe nonequilibrium flow. When the immobile region has zero volume, or when the diffusive exchange is instantaneous, the MIM reduces to the ADE; this extreme can be thought of as describing a soil of low heterogeneity. As might be expected from its close ties with the ADE, the MIM predicts identical scaling behaviors (Table 1). On a practical level, the MIM often suffers from nonunique solutions; that is, a given BTC may be equally well fitted by more than one single set of parameters. This non-uniqueness implies that the equation has too many fitting

Table 1 Scaling with time (T) of solute mass mean and standard deviation according to four transport models.

Model	Mean	Standard deviation	Standard deviation/mean
ADE	t	$t^{1/2}$	$t^{-1/2}$
MIM	t	$t^{1/2}$	$t^{-1/2}$
CLT	t	t	Constant
CTRW ($0 < \beta < 1$)	t^β	t^β	Constant
CTRW ($1 < \beta < 2$)	t	$t^{(3-\beta)/2}$	$t^{(1-\beta)/2}$

ADE, advection–dispersion equation; MIM, mobile–immobile model; CLT, convective lognormal transfer function model; CTRW, continuous-time random walk model.

Adapted from: Jury WA, Gardner WR and Gardner WH (1991) *Soil Physics*, 5th edn. New York: John Wiley; Skaggs TH and Leij FJ (2002) Solute transport: Theoretical background. In: Dane JH and Topp GC (eds.) *Methods of Soil Analysis*, Part 4, *Physical Methods*. Madison, WI: Soil Science Society of America, ch. 6.3; Berkowitz B, Kosakowski G, Margolin G and Scher H (2001) Application of continuous time random walk theory to tracer test measurements in fractured and heterogeneous porous media. *Ground Water* **39**(4): 593–604.

parameters (three for the MIM versus one for the basic ADE), and casts some doubt on the physical interpretation of those parameters.

The MIM can be generalized by positing multiple components of the porous medium, each interacting with the fluid flow through its own rates and mass transfer coefficients; this is the Multiple-Rate Mass Transfer (MRMT) model of Haggerty and Gorelick (1995).

The MIM model is single-rate and predicts an exponential trend toward an asymptotic state where the mobile mass is a constant positive fraction of the injected mass. However, tracer measurements observed at the MADE site exhibited a continuous power-law decrease in mass recovery (Harvey and Gorelick, 2000). Such a power-law trend may be explained by a fractal distribution of rate coefficients. Schumer et al. (2003) generalized the traditional MIM model and proposed fractional MIM (FMIM) equations with power-law waiting times in the immobile zone. The FMIM model is equivalent to previous models of MIM transport with power law memory functions, and is the limiting case governing continuous time random walks with heavy tailed random waiting times.

Transfer functions

Another common conceptual model is the so-called streamtube model or transfer function model, in which solute transport is considered to take place through multiple parallel but noninteracting streamlines (Fig. 4C). Dispersion results from the differences in velocity among the various streamtubes. Because in principle it is possible to reproduce any possible breakthrough curve by generating the correct streamtube distribution, this method is very flexible. This model is essentially a multiple-tube equivalent of the dispersion that would result from flow in a tube in the absence of diffusion, and so, because their dispersion results from insufficient mixing, streamtube models may be considered to be preasymptotic. Streamtube models are most applicable to media in which flow really does take place through parallel but noninteracting pathways, such as horizontal flow through an aquifer with homogeneous layers of contrasting properties. Some variants of this model allow dispersion within each streamtube, while others assume no local dispersion (i.e., piston flow). One common version, Jury's convective lognormal transfer function model (CLT), specifies the distribution of the solute travel times (or, by analogy, the tube radii) and thus the form of the final BTC.

The streamtube models assume zero interaction between streamlines, in contrast to the ADE, which assumes that streamlines do interact. These two models therefore represent extremes along a continuum of streamline interaction. In principle, the streamtube model should better represent early or preasymptotic times, before streamlines have had time to intermix adequately, and the ADE later times.

While the ADE predicts that solute-spreading scales with the square root of time, the CLT predicts solute-spreading that scales linearly with time (Table 1). Solute movement that is consistent with a streamtube-like mechanism will therefore appear to scale if analyzed using the ADE; likewise, solute transport that is well described by the ADE will scale incorrectly if analyzed using the CLT.

Fractional ADE (FADE)

The classical ADE can be generalized through the use of fractional calculus. Meerschaert et al. (1999) were among the first to propose this by applying Lévy's family of stable distribution. They developed the fractional advection-dispersion equation:

$$\frac{\partial C}{\partial t} = -v \frac{\partial C}{\partial x} + D \frac{1 + \omega}{2} \frac{\partial^2 C}{\partial x^2} + D \frac{1 - \omega}{2} \frac{\partial^2 C}{\partial (-x)^2} \quad (5)$$

where $1 < \zeta \leq 2$ is the stable distribution index and $-1 \leq \omega \leq 1$ is the skewness. Eq. (5) with $\zeta = 2$ and $\omega = 0$ reduces to the traditional ADE model.

Initial models were space-fractional (fractional in space), meaning that the order of space in the partial differential equation is fractional, not necessarily an integer. Variants proposed later include the time-fractional dispersion equation with drift (Liu et al.,

Table 2 Summary of different models developed by means of fractional calculus (Kelly et al., 2017).

Model	Governing equation	Remarks
Space-fractional ADE (sFADE)	$\frac{\partial C}{\partial t} = -V \frac{\partial C}{\partial x} + D \frac{1+\omega}{2} \frac{\partial^2 C}{\partial x^2} + D \frac{1-\omega}{2} \frac{\partial^2 C}{\partial (-x)^2}$	ζ Stable index, ω skewness
Time-fractional ADE (tFADE)	$\frac{\partial^{\gamma} C}{\partial t^{\gamma}} = -V \frac{\partial C}{\partial x} + D \frac{\partial^2 C}{\partial x^2}$	γ Time-fractional exponent
Fractional mobile-immobile (FMIM)	$\frac{\partial C}{\partial t} = -V \frac{\partial C}{\partial x} + D \frac{\partial^2 C}{\partial x^2} - \kappa \frac{\partial^{\gamma} C}{\partial t^{\gamma}}$	γ Time-fractional exponent, κ capacity coefficient
Temporally tempered Lévy motion (TTLM)	$\frac{\partial C}{\partial t} = -V \frac{\partial C}{\partial x} + D \frac{\partial^2 C}{\partial x^2} - \kappa \frac{\partial^{\gamma+\lambda} C}{\partial t^{\gamma+\lambda}}$	γ Time-fractional exponent, κ capacity coefficient, λ tempering rate

2003), the fractional mobile-immobile equation (Schumer et al., 2003), and the temporally tempered Lévy motion model (Meerschaert et al., 2008) (Table 2).

Kelly et al. (2017) stated that “While sFADE [space-fractional ADE] provides acceptable fits for the BTCs under consideration, sFADE does admit nonphysical behavior (negative dispersion) that may manifest itself at other measurement locations or times.”

Continuous-time random walk (CTRW)

Continuous-time random walk (CTRW) theory has also been used to study contaminant migration. It is based on the generalization of the “random walk” method where the particle transition times, as well as jump lengths, are accounted for. The CTRW has been successfully developed by Berkowitz, Scher, and co-workers to quantify non-Fickian transport in a broad variety of porous and fractured media. The basis of the CTRW framework is the portrayal of transport as a sequence of transition rates and the incorporation of the full spectrum of these rates into the transport equations.

The CTRW is derived from the Master equation of statistical mechanics using appropriate memory functions to represent non-Fickian behavior. Within the CTRW framework, the transport equation in one dimension and Laplace space is

$$u\tilde{C}(x, u) - C_0(x) = -\tilde{M}(u) \left[v_t \frac{\partial}{\partial x} \tilde{C}(x, u) - D \frac{\partial^2}{\partial x^2} \tilde{C}(x, u) \right] \quad (6)$$

where $\tilde{C}(x, u)$ is the Laplace transform of concentration at location x , u is the Laplace variable and $\sim$ indicates Laplace space, $C_0(x)$ is the initial concentration at location x , and v_t is the transport (solute) velocity.

In Eq. (6), $\tilde{M}(u)$ is the memory function that accounts for the unknown heterogeneities at the pore scale and is given by (Berkowitz and Scher, 2009)

$$\tilde{M}(u) = t_1 u \frac{\tilde{\psi}(u)}{1 - \tilde{\psi}(u)} \quad (7)$$

where t_1 is a median time for transitions between pore bodies, and $\tilde{\psi}(u)$ is the transition rate probability and the Laplace transform of $\psi(t)$, the waiting (or transition) time distribution. In the literature, various choices of $\tilde{\psi}(u)$, such as decoupled and coupled joint distributions have been made to determine $\tilde{M}(u)$ in Eq. (7). For example, in the earlier framework of the CTRW, a simple power law such as $\psi(t) \sim t^{-1-\beta}$ was applied (Scher et al., 1991). Geometry and velocity distributions, in conjunction with particle mixing rules, were also used to map the particle movement between fracture intersections onto a joint probability density $\psi(s, t)$, the probability per time for a transition between fracture intersections separated by s with a difference of arrival times of t (Berkowitz and Scher, 1998).

At sufficiently large time, the transport in a real system will become Fickian at length scales sufficiently greater than the largest heterogeneity scale meaning that the medium acts as a more homogeneous system. This behavior can be suitably addressed using the following truncated power-law (also known as TPL) version of $\psi(t)$, which enables detailed investigation of the transition from non-Fickian to Fickian transport:

$$\psi(t) = \left[\left(\frac{t_1}{t_2} \right)^{\beta} \exp \left(\frac{t_1}{t_2} \right) \Gamma \left(-\beta, \frac{t_1}{t_2} \right) \right]^{-1} \frac{\exp(-t/t_2)}{t_1 (1 + t/t_1)^{1+\beta}} \propto \exp(-t/t_2) (1 + t/t_1)^{-1-\beta} \quad (8)$$

where t_1 and t_2 are the limits of the power-law spectrum, $\Gamma(a, x)$ is the incomplete Gamma function, and β is the exponent of the waiting (or transition) time distribution. β typically ranges between 0 and 2. The value of β is non-universal, meaning that it depends on the heterogeneity and morphology of the porous medium.

In Eq. (8), $\psi(t) \sim (t/t_1)^{-1-\beta}$ for $t_1 \leq t \leq t_2$, and $\psi(t)$ decreases exponentially for $t \geq t_2$. The interplay between β and t_2 has a substantial impact on the entire shape of a migrating solute plume (Berkowitz and Scher, 2009). By fitting the CTRW solution to either experimental data or numerical simulations, different values of β e.g., between 0 and 1 (Cortis and Berkowitz, 2004), 1 and 2 (Berkowitz and Scher, 2009), and > 2 have been reported in the literature. Fickian-like transport (Gaussian behavior) arises when $\beta > 2$, in agreement with the central limit theorem. The $0 < \beta < 2$ range corresponds to highly heterogeneous porous media. An attractive aspect of the CTRW is that fitting a breakthrough curve gives an estimate of β , which also indicates how the transport will scale.

Percolation theory

Percolation theory can be applied to study porosity- and saturation-dependent treatment of dispersion in porous media using the concepts from critical path analysis, cluster statistics of percolation, and fractal scaling of percolation clusters (Hunt and Skinner, 2008; Ghanbarian-Alavijeh et al., 2012). Within the percolation theory framework, one can calculate spatial solute distributions at an instant in time and arrival time distributions as a function of system size and properties of the medium. Unlike all the previous models, percolation theory is not derived from or related to the ADE.

One first models the pore space of the medium by assuming that the pore size distribution follows a specific type of probability density function, for example a power law. The pore size distribution leads to the probability distribution of the hydraulic conductance, g , for a given water content θ and distance from the source x as follows (Hunt et al., 2014)

$$W(g|x; \theta) \propto Ei \left[\left| 1 - \left(\frac{g}{g_c} \right)^{\frac{3-D_p}{3}} \right| \left[\left(\frac{x}{L} \right)^{\frac{\xi - \phi + \theta - \theta_c}{\xi - \theta_c}} \right]^{\frac{2}{v}} \right] \quad (9)$$

where Ei is the exponential integral, g is the hydraulic conductance, g_c is the critical hydraulic conductance, D_p is the pore space fractal dimension, v is the universal exponent of the correlation length (equal to 0.88 in three dimensions), ξ is the parameter of the generalized power-law distribution (Ghanbarian-Alavijeh et al., 2012), and θ_c is the critical water content for percolation.

The probability distribution function for solute arrival times in a system of length x , $W(t|x; \theta)$ is found using the identity $W(t|x; \theta)dt = gW(g|x; \theta)dg$. To apply such an identity, one must first have an expression for the typical transit time, $t(g|x; \theta)$ of a cluster of length x and controlling conductance g . The typical time for transport through such a cluster is generated by isolating the effects of cluster topology and the conductance distribution. If one solves for the saturation dependence of $t(g|x; \theta)$ for the power-law pore-size distribution model, the result is (Ghanbarian-Alavijeh et al., 2012)

$$t(g|x; \theta) = \left(\frac{x}{L} \right)^{D_b} \frac{t_0 D_p}{3 - D_p} \frac{\left[\left(\frac{g_c}{g} \right)^{1 - \frac{D_p}{3}} \left(\frac{\xi - \phi + \theta}{\xi - \phi + \theta - \theta_c} \right) - 1 \right]}{\left[1 - \left(\frac{g}{g_c} \right)^{\frac{D_p}{3}} \left(\frac{\xi - \phi + \theta - \theta_c}{\xi - \phi + \theta} \right)^{\frac{D_p}{3-D_p}} \right]} \left[\frac{1}{(\xi - \phi + \theta - \theta_c)^{(D_b-1)v}} \right] \left[\frac{1}{1 - \left(\frac{g}{g_c} \right)^{1 - \frac{D_p}{3}}} \right]^{(D_b-1)v} \quad (10)$$

where D_b is the backbone fractal dimension described below.

All the parameters in Eq. (10) have a physical meaning. Recall that the exponent for the correlation length, v , is universal and equal to 0.88 in three dimensions (Hunt et al., 2014). The exponent for the fractal dimensionality of the backbone is known to have only a limited variability (in particular with flow dimensionality and whether the percolation problem is random or invasion). However, it could also have a wide range of values if the pore space is correlated depending on the level of correlation. The critical volume fraction for percolation θ_c will not typically be known, but a good estimation can be made from soil water retention data, where it can be identified with the minimum water content measured (residual water content). The fractal dimensionality, D_p , and the parameter ξ can also be extracted from the soil water retention curve.

Ghanbarian-Alavijeh et al. (2012) applied the percolation theoretic approach, Eqs. (9) and (10), to determine the scale dependence of dispersivity, compared with more than 2200 experiments measured over 10 orders of magnitude of length scale, and found good agreement between the theory and experiments. Specific comparisons of predictions for arrival time distribution with experiments on a single column over a range of saturations also demonstrated that their model is accurate and predictive.

Sahimi (2012) applied power-law scaling laws from percolation theory and linked the exponent β in the CTRW model to scaling exponents characterizing power-law behavior of electrical and hydraulic conductivities, as well as correlation length, percolation probability, and accessible fraction of pores on the sample spanning cluster. Sahimi (2012) argued that nonuniversality of β could be due to the nonuniversality in continuum systems or in percolation models with long-range correlations.

Observations and mechanisms

There is a strong tendency for dispersivity to be greater at larger scales, as documented in several meta-analyses (e.g., Neuman, 1990; Schulze-Makuch, 2005). Unfortunately, the scatter in the data is great enough that no universal scaling behavior can unambiguously be inferred. While the cause of this trend is not clear, the most common explanation is that, as a solute travels farther, it encounters increasing scales of heterogeneity. This means that, at larger scales, solute molecules will on average sample a larger range of velocities than they did at a smaller scale. The travel time for each molecule is the sum of many different times at many different velocities; increasing the range of velocities will therefore (assuming random and independent sampling by each molecule) result in a wider distribution of travel times and thereby a larger dispersivity. A fractal medium, with its ever-increasing scale of heterogeneity, results mathematically in dispersivity increasing with distance, i.e., scaling.

But increasing scales of heterogeneity are apparently not necessary for scaling to occur. For example, one laboratory experiment used a two-dimensional flow cell packed with two sands: 16% of the volume consisted of blocks of fine sand randomly placed within a coarse sand matrix, with the hydraulic conductivities differing by a factor of 20. Within approximately 1 m, the fitted dispersivity increased from 0.073 to 0.127 m. The skew of the data was fairly constant, suggesting that a change from preasymptotic

to asymptotic behavior was not occurring. Similarly, the Borden aquifer tests saw dispersivity increase by a factor of 8 over a distance of some 10 m in what was considered a homogeneous sand. And of course dispersion that occurs primarily through a streamtube-like mechanism, wherein regions that differ in velocity are isolated from each other, will also show scaling of ADE-calculated dispersivity.

Discussion of the 'scale of heterogeneity' also raises the question of the importance of heterogeneity at smaller scales. If much of dispersion is driven by heterogeneity, and that heterogeneity exists at all scales, then presumably the nature of the heterogeneity below the scale of heterogeneity is also important. But because both models and measurements are blind to processes below their respective finite-size scales, there are always smaller-scale processes that they cannot take into account, but which still influence the large-scale outcome. An exclusive focus on the largest-scale heterogeneities implicitly assumes that the smaller-scale heterogeneities are either negligible or diffusive in nature, but neither is necessarily the case. The smaller-scale influences can be significant and non-diffusive; for example, the presence of undetected cracks or wormholes in a soil profile may lead to arrival of solute at the water table well in advance of prediction.

Soils and other geologic porous media are neither purely random nor purely structured; instead they vary in their degree of randomness across most or all scales. But while understanding of soil properties has historically concentrated on the measurements themselves, and only relatively recently on the spatial structure of the properties, generally the topological structure is still ignored. As a consequence, researchers in the theory of contaminant transport have long considered (for example) what proportion of an aquifer is sand or clay, and even how thick the average sand layer might be, but not the question of whether there is a continuous sand pathway through the system. Continuity and interconnectedness are critical to diffusive and advective exchange across streamlines, and to the range of velocities encountered within a given streamline. They therefore make the difference between the ADE and streamtube models, and so are critical determinants of the scaling behavior of a given soil or aquifer.

In most situations, solute transport in soils or other geologic porous media is subject to all of the mechanisms mentioned so far: diffusive exchange between neighboring streamlines, parallel pathways with little exchange between them, and diffusive exchange between more and less mobile regions of the medium. There are also other important dispersive processes, such as advective (and even turbulent) exchange between streamlines, streamline-specific sorption based upon mineralogy, biodegradation based upon differences in microhabitat, and streamlines with nonconstant velocity, that do not show up in these models. Certainly the extent to which each mechanism contributes to dispersion is difficult to determine in advance, nor can one know in advance whether the length scale at which dispersion is to be predicted is greater than or less than the scale of heterogeneity (if any) of the medium.

In summary, predicting the scaling of transport is quite uncertain. New mathematical models such as FADE and CTRW give great flexibility in fitting, and open the possibility of gaining hints about the scaling. This is a topic that continues to advance. Nonetheless, natural systems are complex, and the practitioner should not assume that answers will be easy, nor early answers correct.

Acknowledgments

B. Ghanbarian is grateful to Brian Berkowitz, Weizmann Institute of Science, for fruitful discussion of the CTRW model, and to Kansas State University for support through faculty start-up funds.

References

- Aris R (1956) On the dispersion of a solute in a fluid flowing through a tube. *Proceedings of the Royal Society of London* 235(1200): 67–77.
- Berkowitz B and Scher H (1998) Theory of anomalous chemical transport in random fracture networks. *Physical Review E* 57(5): 5858–5869.
- Berkowitz B and Scher H (2009) Exploring the nature of non-Fickian transport in laboratory experiments. *Advances in Water Resources* 32(5): 750–755.
- Berkowitz B, Cortis A, Dentz M, and Scher H (2006) Modeling non-Fickian transport in geological formations as a continuous time random walk. *Reviews of Geophysics* 44: RG2003.
- Bromly M and Hinz C (2004) Non-Fickian transport in homogeneous unsaturated repacked sand. *Water Resources Research* 40(7): 5–9.
- Cortis A and Berkowitz B (2004) Anomalous transport in "classical" soil and sand columns. *Soil Science Society of America Journal* 68(5): 1539–1548.
- Dentz M and Berkowitz B (2003) Transport behavior of a passive solute in continuous time random walks and multirate mass transfer. *Water Resources Research* 39(5): 1111.
- Ghanbarian-Alavijeh B, Skinner TE, and Hunt AG (2012) Saturation dependence of dispersion in porous media. *Physical Review E* 86(6): 066316.
- Griffiths A (1910) On the movement of a coloured index along a capillary tube, and its application to the measurement of the circulation of water in a closed circuit. *Proceedings of the Physical Society of London* 23: 190–197.
- Haggerty R and Gorelick SM (1995) Multiple-rate mass transfer for modeling diffusion and surface reactions in media with pore-scale heterogeneity. *Water Resources Research* 31(10): 2383–2400.
- Harvey C and Gorelick SM (2000) Rate-limited mass transfer or macrodispersion: Which dominates plume evolution at the Macrodispersion Experiment (MADE) site? *Water Resources Research* 36(3): 637–650.
- Hunt AG and Skinner TE (2008) Longitudinal dispersion of solutes in porous media solely by advection. *Philosophical Magazine* 88(22): 2921–2944.
- Hunt AG, Ewing RP, and Ghanbarian B (2014) *Percolation Theory for Flow in Porous Media*. vol. 880. Springer.
- Kelly JF, Bolster D, Meerschaert MM, Drummond JD, and Packman AI (2017) FracFit: A robust parameter estimation tool for fractional calculus models. *Water Resources Research* 53(3): 2559–2567.
- Lévy M and Berkowitz B (2003) Measurement and analysis of non-Fickian dispersion in heterogeneous porous media. *Journal of Contaminant Hydrology* 64(3–4): 203–226.
- Liu F, Anh VV, Turner I, and Zhuang P (2003) Time fractional advection-dispersion equation. *Journal of Applied Mathematics and Computing* 13(1): 233–245.
- Meerschaert MM, Benson DA, and Bäumer B (1999) Multidimensional advection and fractional dispersion. *Physical Review E* 59(5): 5026–5028.
- Meerschaert MM, Zhang Y, and Baeumer B (2008) Tempered anomalous diffusion in heterogeneous systems. *Geophysical Research Letters* 35: L17403. <https://doi.org/10.1029/2008GL034899>.

- Metzler R and Klafter J (2000) The random walk's guide to anomalous diffusion: A fractional dynamics approach. *Physics Reports* 339(1): 1–77.
- Molz FJ, Güven O, and Melville JG (1983) An examination of scale-dependent dispersion coefficients. *Ground Water* 21(6): 715–725.
- Neuman S (1990) Universal scaling of hydraulic conductivities and dispersivities in geologic media. *Water Resources Research* 26(8): 1749–1758.
- Neuman SP and Zhang YK (1990) A quasi-linear theory of non-Fickian and Fickian subsurface dispersion: 1. Theoretical analysis with application to isotropic media. *Water Resources Research* 26(5): 887–902.
- Sahimi M (2012) Dispersion in porous media, continuous-time random walks, and percolation. *Physical Review E* 85(1): 016316.
- Scheidegger AE (1959) An evaluation of the accuracy of the diffusivity equation for describing miscible displacement in porous media. In: *Proceedings of the 2nd Theory of Fluid Flow in Porous Media Conference* 101–116.
- Scher H, Shlesinger M, and Bendler J (1991) Time-scale invariance in transport and relaxation. *Physics Today* 26–34.
- Schulze-Makuch D (2005) Longitudinal dispersivity data and implications for scaling behavior. *Ground Water* 43(3): 443–456.
- Schumer R, Benson DA, Meerschaert MM, and Baeumer B (2003) Fractal mobile/immobile solute transport. *Water Resources Research* 39(10): 1296. <https://doi.org/10.1029/2003WR002141>.
- Sun L, Qiu H, Niu J, Hu BX, Kelly JF, et al. (2020) Comparison of negative skewed space fractional models with time nonlocal approaches for stream solute transport modeling. *Journal of Hydrology* 582: 124504.
- Taylor GI (1953) Dispersion of soluble matter in solvent flowing slowly through a tube. *Proceedings of the Royal Society of London, Series A* 219: 189–203.
- van Genuchten MT and Wierenga PJ (1976) Mass transfer studies in sorbing porous media I. Analytical solutions. *Soil Science Society of America Journal* 40(4): 473–480.
- Xu M and Eckstein Y (1995) Use of weighted least-squares method in evaluation of the relationship between dispersivity and field scale. *Ground Water* 33(6): 905–908.

Soil influence on groundwater

David Rudolph^a and Edwin Cey^b, ^aUniversity of Waterloo, Waterloo, ON, Canada; ^bUniversity of Calgary, Calgary, AB, Canada

© 2023 Elsevier Ltd. All rights reserved.

This is an update of Y. Bachmat, *GROUNDWATER AND AQUIFERS*, Editor(s): Daniel Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 153–168, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00369-6>.

Introduction	364
Groundwater hydrology	365
Groundwater occurrence	366
Groundwater flow	366
Geochemistry of groundwater	367
Groundwater contamination	368
Groundwater as a natural and anthropogenic resource	369
Impact of soils on groundwater	370
Soil influence on groundwater recharge and flow	370
Soil influence on groundwater quality	371
Land management influence	373
Rural soil environments	373
Urban soil environments	374
Groundwater impact on soils	374
Conclusions	375
References	377

Abstract

Soils play an important role in controlling both the hydrologic and biochemical interaction between the land surface and the underlying groundwater system. A dynamic synergy exists between how soil processes influence groundwater quantity and quality and how the nature of groundwater flow influences soil water content and composition, and surface water ecology. The soil zone functions as both a reactive filter for infiltrating water and a region of infiltration above the water table facilitating the replenishment of groundwater resources. Rapid anthropogenic development combined with the uncertainties of a warming climate are leading to observable depletion of the groundwater resource, along with poorly understood, yet potentially significant changes in the complex interaction between soils and groundwater.

Key points

- Importance of groundwater in the global water balance
- Fundamentals of groundwater flow systems
- Influence of soils on groundwater quality and quantity
- Biogeochemical characteristics and interactions between soils and underlying groundwater resources
- Conjunctive interaction between soils and groundwater and impacts on aquatic ecology and groundwater contamination

Introduction

Within the global inventory of freshwater, groundwater accounts for approximately 30% of the overall total, yet it is estimated to represent close to 95% of the accessible, unfrozen freshwater resource (UNESCO, 2022). Groundwater is relied upon as a municipal, industrial and agricultural water source worldwide, providing more than half of the irrigation water demand, and representing the primary water source for over two billion people. Groundwater also plays a significant role in sustaining aquatic ecosystems, ranging from wetlands to streams and lakes, and functions as a strategic reserve in times of drought. As surface water resources are approaching or exceeding their allocation limits worldwide, groundwater is being relied upon progressively to support a wide variety of human needs. Warming climatic trends have begun to influence the replenishment, storage and temperature of groundwater, although the long-term impacts on the resource remain uncertain (Earman and Dettinger, 2011).

Both the quantity and quality of groundwater resources are facing unprecedented challenges from anthropogenic activities and changing climatic drivers. Overexploitation has resulted in more than half of the world's main aquifer systems being depleted at a rate greater than the natural replenishment. This has resulted in the loss of critical baseflow required to sustain terrestrial water bodies and has led to excessive land subsidence in some areas. Overpumping has resulted in the progressive capture of deep, lower

quality groundwater, causing the chronic degradation of regional groundwater quality by geogenic contaminants (Levy et al., 2021). The legacy of agricultural land management and municipal and industrial growth worldwide has resulted in the extensive contamination of vital groundwater supply systems, leading to adverse human health and ecosystem impacts. Despite the strategic value of groundwater, it remains poorly understood, quantified and managed at a global scale.

Groundwater is conventionally defined as the water that fills the pores, fractures and open cavities within subsurface materials beneath the phreatic level or water table, which is generally referred to as the saturated zone. Below the water table fluid pressure is considered positive, exceeding atmospheric pressure. The region overlying the water table and extending to ground surface is referred to as the vadose zone, where the residing pore water is held in tension within the pore network and subsurface materials are predominantly variably saturated (Fig. 1). Pore water tension (matric potential) may be highly variable both spatially and temporally within the vadose zone with greater tensions leading to smaller degrees of saturation as illustrated in Fig. 1. Near the water table, pore water tensions are generally low and the geologic materials in this region may remain saturated as water is held by capillarity of the pores. This is referred to as the zone of tension saturation or the capillary fringe (Fig. 1). The upper section of the vadose zone contains the soil water zone, which represents the interface between the ground surface and the variably saturated materials overlying the groundwater system (Fig. 1). The soil zone can be highly variable in thickness, mineral composition and in its physical, chemical and microbiological nature.

The majority of the rooting zone for terrestrial vegetation is contained within the unsaturated soil zone and it is normally characterized by its large component of organic carbon. Surface vegetation, the soil zone, and shallow groundwater are closely interlinked, often described and studied together as one unit referred to as the Critical Zone. The soil zone represents a vital component of the Critical Zone (Lin, 2010) (Fig. 1) controlling infiltration of terrestrial water and possessing a large capacity for geochemical and biological reactivity, which can significantly influence the soil water quality. As such, the Earth's surficial soils can fundamentally influence both the quality and quantity of groundwater at the local and regional scale. The nature of this influence depends on the physical, chemical and biological characteristics of the soil, which are in turn affected by local climatic conditions. The intrinsic interaction between soils and groundwater is the primary focus of this chapter.

Groundwater hydrology

Water contained within the terrestrial subsurface is conventionally compartmentalized as soil water and groundwater. Soil water exists within the variably-saturated vadose zone normally under fluid pressures lower than ambient atmospheric pressure (or tension). Groundwater is defined as the vast reserve of subsurface water that exists below the water table under fully saturated conditions at positive pressures at or above atmospheric pressure (Fig. 1). In soil physics, the term matric potential refers to the tension, with negative values (tension) in the vadose zone, whereas the positive pressure in the groundwater shows the pressure potential. The interaction between soil water and groundwater has wide ranging implications for the quantity and quality of water throughout all components of the global water balance.

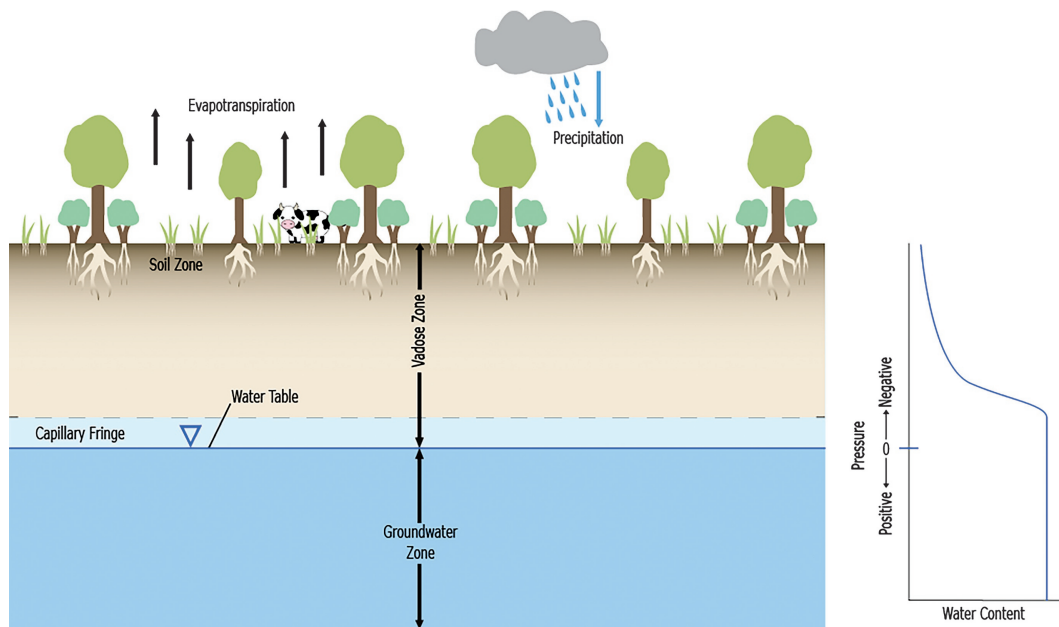


Fig. 1 Conceptual diagram of the shallow subsurface showing the transition from ground surface through the soil and vadose zones to groundwater. The relationship between pore water content and pressure head is provided in the illustration on the right.

Groundwater occurrence

At a global scale, an estimated 68% of freshwater is trapped in glacial ice with, remarkably, only 1%–2% in surface water systems including all of Earth's lakes and rivers. The balance of the total freshwater exists as groundwater. The amount of freshwater circulating through the soil zone is an extremely small fraction of the overall global water balance (Shiklomanov, 1993). Groundwater provides drinking water to over 50% of the world's population and accounts for about 25% of all anthropogenic freshwater needs on Earth (UNESCO, 2022). Relative to surface water, groundwater exists as a highly accessible resource in most areas of the world, making it available to a wide range of users, including communities and industries. As such, it has tremendous societal and economic value, yet it remains poorly understood and undervalued.

The occurrence and movement of soil water and groundwater varies both spatially and temporally due to the complexity of the physical conditions within the subsurface and the changing hydrologic conditions throughout the annual cycle. Groundwater exists within lithified or consolidated geologic strata (rock) and non-lithified or unconsolidated strata (loose sediments) below the water table. The characteristics of the sediments influence the mobility, quality and accessibility of the groundwater resource. Of critical importance in understanding the occurrence and movement of groundwater are the physical parameters that define the permeability and porosity of the geologic material within which groundwater occurs. These parameters influence both the ease with which groundwater can flow under a driving force (permeability) and the storage capacity (porosity). Key controlling parameters include:

- **Porosity:** volume of voids per unit volume of the porous medium.
- **Effective porosity:** volume of interconnected voids per unit elementary volume of the porous medium. While porosity is a local measure of the total water storage capacity, effective porosity is a local measure of the well-connected pore space through which water actually moves.
- **Specific storage:** volume of water released from (or taken into) storage per unit volume of porous medium, per unit decline (or rise) of the piezometric (hydraulic) head.
- **Storativity:** equates to the specific storage multiplied by the saturated thickness of the geologic unit at the given point.
- **Specific yield:** the volume of water that is released from storage in an unconfined aquifer (explained below) per unit drop in the water table. The specific yield of an unconfined aquifer is larger than the specific storage of a comparable confined aquifer due to the process of actual dewatering and draining of the soil pores which occurs in the unconfined setting.
- **Permeability (or intrinsic permeability) and hydraulic conductivity:** permeability is a characteristic of the void space which affects its ability to transmit a fluid and is related to the size and interconnectivity of the pores within the sediment or rock. Hydraulic conductivity is directly related to the permeability of the porous medium but also incorporates properties of the fluid, such that it is directly proportional to fluid density and inversely proportional to fluid viscosity.
- **Transmissivity:** the discharge of groundwater through a unit width of the saturated zone as a result of a unit hydraulic gradient. Transmissivity, at a given point in the horizontal plane, equals the hydraulic conductivity, multiplied by the saturated thickness of the aquifer at the given point.
- **Hydraulic head and hydraulic gradient:** hydraulic head in the subsurface refers to a measure of the fluid potential of the water at the point of measurement and is a combination of the fluid pressure and the gravitational potential energy related to the elevation of the water above a chosen datum point, which is commonly taken as sea level. The hydraulic gradient is the rate of change in the hydraulic head over a specified distance in the direction of flow and is often referred to as the driving force of subsurface water flow.

Natural geologic strata containing groundwater are often categorized as either aquifers or aquitards. Aquifers consist of relatively permeable materials and are capable of providing useful or economically-feasible volumes of water when pumped for anthropogenic uses or which supply significant replenishment through upward seepage to surface water bodies such as lakes, streams and wetlands. Examples of typical consolidated aquifer materials include sandstones, porous or karstic limestones and highly fracture rocks, and some examples of unconsolidated aquifer materials include glacially-deposited sands and gravels, fluvial and alluvial sediments and coarse eolian deposits (Fetter and Kremer, 2021) (Fig. 2). Geologic units with low permeability, where groundwater flow is reduced, are referred to as aquitards and are often characterized by finer grained materials. Rock types like shale, granite and unfractured carbonate rocks are examples of consolidated aquitards and clay, glacial till and fine silts are examples of unconsolidated aquitards. Groundwater strata are further defined as being *unconfined* if bounded from above by the phreatic surface or the water table or *confined* when an aquifer unit is bounded from above by an aquitard of lower permeability such that the hydraulic head within the aquifer unit exists at levels higher than the elevation of the interface between the aquifer and the aquitard (Fig. 2). Aquifer and aquitard units are frequently encountered in layered sequences and when the aquitard units are permeable enough to permit a reasonable transmission of water vertically between the aquifer units, the aquifers are referred to as being *semiconfined*. The regional movement of groundwater through the complex subsurface environment is influenced by the distribution and nature of the annual replenishment of meteoric water recharging the groundwater flow system through the surficial soil zone as described below.

Groundwater flow

Groundwater is normally in a state of permanent motion from areas of natural and artificial recharge to points of natural and artificial discharge. Renewable groundwater is part of the hydrologic cycle, as seasonal meteoric and surface waters percolate

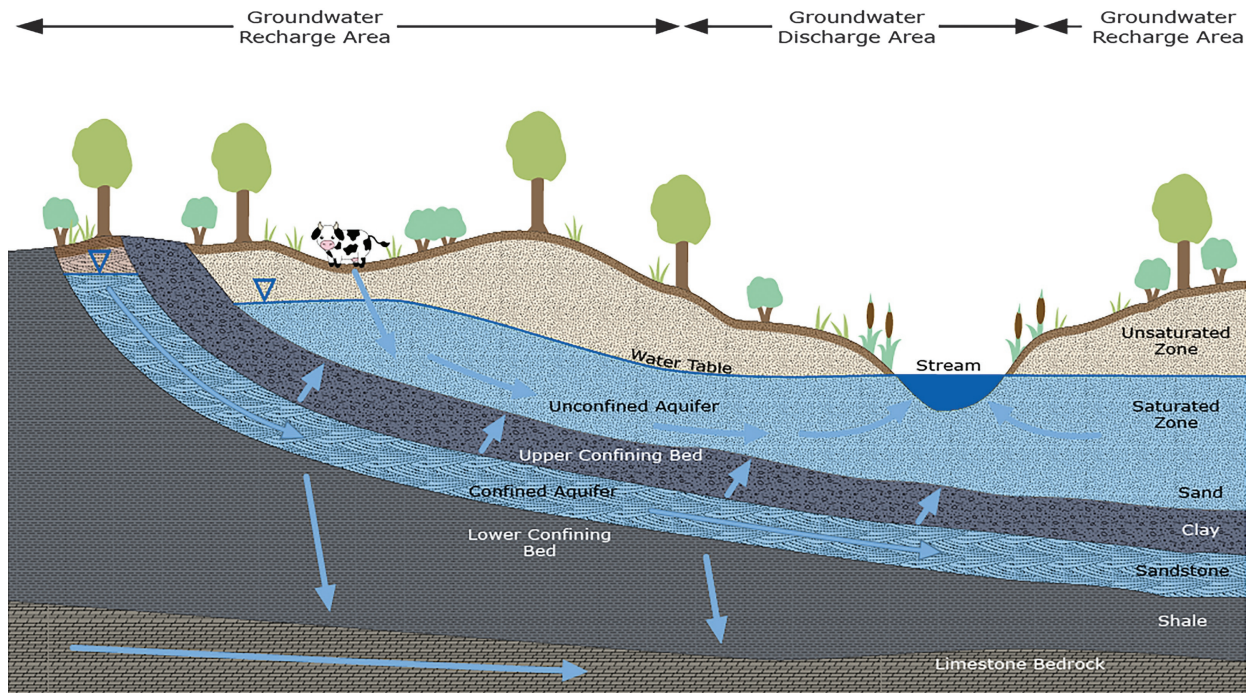


Fig. 2 Regional groundwater flow system from recharge to discharge areas through natural geologic units, including confined and unconfined aquifers and low permeability aquitards.

downward through the soil and unsaturated zone to the saturated portion of the subsurface, driving groundwater flow from areas of recharge to natural discharge outlets such as springs, wetlands, streams, lakes, and oceans (Fig. 2). The vadose zone, including the soil horizons, functions as a highly dynamic and reactive filtering agent overlying the groundwater flow system (Figs. 1 and 2). The permeability, composition and thickness of the vadose zone controls seepage rates and the chemical or biological reactivity above the water table, influencing the infiltration timing, distribution and water quality in recharge areas (Fig. 2). The characteristics of the shallow soil zone are highly variable spatially and transient, fundamentally influencing groundwater both locally and regionally. This is addressed in a subsequent section. The movement of groundwater within the saturated zone is usually slow and variable, ranging from millimeters per year up to meters or tens of meters per day. As a result, the residence time or age of water in the saturated zone (i.e., the time it takes to travel from its point of entry into the saturated zone to the point of discharge) may range from several weeks to thousands of years. The length of time groundwater remains within the subsurface influences its chemical nature due to the water-rock or sediment interaction that occurs along the flow path, which will also be discussed in more detail in a subsequent section.

The rate or flux of groundwater in the subsurface is described by a variable, termed specific discharge. It is defined at any point as the volume of water which passes per unit time through a unit area of porous medium perpendicular to the direction of flow at that point. Because flow takes place through the void spaces only, the average groundwater velocity is estimated by dividing the specific discharge by a factor equal to the effective porosity, defined earlier. The quantification of groundwater flow conventionally relies on Darcy's Law, which defines the specific discharge as the product of the hydraulic gradient and the hydraulic conductivity. Groundwater moves from regions of high hydraulic head to regions of low hydraulic head. By measuring hydraulic head distribution within the subsurface with a network of piezometers or monitoring wells and combining these data with laboratory (e.g., grain size analysis, permeameter) or in situ (e.g., single well response tests, pumping tests) estimates of hydraulic conductivity, groundwater flow direction and rate can be determined at local or regional scales (Fetter and Kremer, 2021). Due to the inherent geologic heterogeneity within the subsurface (hydraulic conductivity and porosity) and the transient nature of the hydraulic factors driving the flow system (recharge, discharge and hydraulic gradient), groundwater flow is highly variable both spatially and temporally.

Geochemistry of groundwater

The dissolved chemical load within natural groundwater is influenced by the chemical composition of infiltrating surficial waters, interactions within the vadose zone, specifically within the soil horizons and the nature of the geologic materials that groundwater encounters along its flow path from recharge regions to discharge points. Infiltrating precipitation will enter the shallow soil environment with an initial aqueous geochemical composition that reflects the local atmospheric conditions. In most cases the

water will have a relatively neutral pH, low total dissolved solids (TDS) and high dissolved oxygen (DO) concentration. Once within the shallow vadose zone, the infiltrating water enters a chemically and biologically reactive environment associated with the soil horizons where the chemical composition may change due to a wide range of reactions. In addition, chemical constituent concentrations may increase as a result of the removal of soil water through evapotranspiration (Poeter et al., 2020). As a result of the combined effect of open access to atmospheric oxygen and the production of carbon dioxide (CO_2) through root respiration and organic decay, the shallow vadose zone can be highly reactive where the development of acidic conditions within the infiltrating porewater results in the dissolution of soil minerals and subsequent increases in TDS. In addition to an increase in ionic composition of the infiltrating porewater, other soluble species including forms of carbon, nitrogen and chemicals derived from surface sources of contamination may also migrate through the shallow vadose zone. The fate and transport of these different species will be influenced by biogeochemical reactions within the soil horizons and the underlying unsaturated zone as discussed in more detail below.

Once the infiltrating water passes through the vadose zone and enters the groundwater flow system, the geochemical composition will continue to evolve. This evolution depends primarily on the chemistry of the recharge water from the vadose zone, the physical and mineralogical makeup of the geologic materials the groundwater will flow through, and the residence time within the subsurface. As the groundwater encounters a diverse assemblage of minerals along the flow path, dissolution will cause concentrations of major, minor and trace constituents to increase depending on the solubility of the constituent, dissolution rates and the length of time the water is exposed to specific mineral types. This is often referred to as water-rock interaction. In the case of a short residence time (e.g., <1 year), the geochemical composition of the groundwater may be very similar to that of the recharge water. However, older groundwater that has remained in the subsurface for much longer time periods (perhaps millennia), are frequently characterized by much greater TDS values, in some cases close to that of seawater (Poeter et al., 2020). Although the chemical composition of groundwater can vary widely, the ionic species that account for the majority of the mass of dissolved solids (>95%) include the cations calcium (Ca^{2+}), sodium (Na^+), and magnesium (Mg^{2+}) and the anions bicarbonate (HCO_3^-), chloride (Cl^-) and sulfate (SO_4^{2-}). Typical minor constituents include iron (Fe^{2+}), potassium (K^+), nitrate (NO_3^-), manganese (Mn^{2+}), silicon (Si^{4+}), fluoride (F^-) and dissolved organic carbon (DOC), but these tend to occur at much lower concentrations (Poeter et al., 2020).

Groundwater contamination

Both natural processes and anthropogenic activities can result in the contamination of groundwater at local and regional scales. As outlined in the previous section, groundwater with extended subsurface residence times can contain high levels of major and minor dissolved constituents that may exceed human or animal recommended drinking water limits. This can include high levels of salts in salinized groundwater that can result in soil salinity over very large areal extents (Scanlon et al., 2007; Poeter et al., 2020; Hassani et al., 2020). These are examples of natural processes leading to groundwater contamination resulting in degradation of water quality and potential impacts on soil characteristics. Anthropogenic sources of groundwater contamination are conventionally grouped in two categories: point and nonpoint or distributed sources (Fig. 3). Examples of point sources are landfills, septic weeping beds, above and below ground storage tanks for hazardous liquids, fuel stations and accidental spills. Point sources of contamination tend to be local in nature although over long periods of time, can result in spatially extensive regions of contaminated groundwater. Examples of nonpoint (distributed) sources include land application of agrichemicals, irrigation with poor-quality waters, atmospheric deposition of pollutants and urban runoff including road de-icers (Fig. 3). These produce large-scale effects in both time and space, which are often overlapping and difficult to monitor and relate to each source separately.

Numerous cases of groundwater contamination, particularly from landfills, mine-tailings, agricultural fertilizers and industrial operations have been documented worldwide. It is worth noting that, except for particular cases of direct contact between local sources and outlets of water and/or contaminants, the effects from most of the above-mentioned sources of groundwater contamination are not immediate. Due to flow of contaminants through soil to groundwater, there can be a long time-lag between the release of a contaminant to the subsurface and the detection of its impact on a groundwater receptor such as a public supply well or surface water body. The fate and transport of a specific pollutant will depend on the thickness and reactive nature of the soil and vadose zone, the flow path and residence time within the saturated zone, and the natural degradation and transformation capacity of the subsurface materials along the flow path. Processes such as adsorption, ion exchange, dilution, and chemical and biochemical reactions can all influence the nature and persistence of a wide range of groundwater contaminants. Once groundwater becomes contaminated, however, it will usually require a very expensive and time-consuming effort to achieve a desired level of clean up or remediation, if it is even technically feasible to achieve. Groundwater remediation strategies and technologies are extremely active areas of research and engineering practice (Fetter et al., 2018).

A specific groundwater contaminant category of global concern relates to water borne pathogens such as some strains of *Escherichia coli* (*E. coli*) bacteria and viruses, which can represent a significant and immediate threat to human health. Bacteria are generally not mobile in sand aquifers because they are large enough to be filtered out by the sand, but can be mobile in gravel aquifers, porous or karstic limestone and fractured rock aquifers. Viruses, which are much smaller, are more mobile than bacteria in aquifers. Harmful viruses can remain active for as long as 3 years and may travel many kilometers in high permeability aquifers (Poeter et al., 2020; Taylor et al., 2004). Microbiological water quality testing typically relies on fecal indicator bacteria, such as *E. coli*, that are easy to detect in water samples and their presence is an indicator of the potential presence of human pathogens

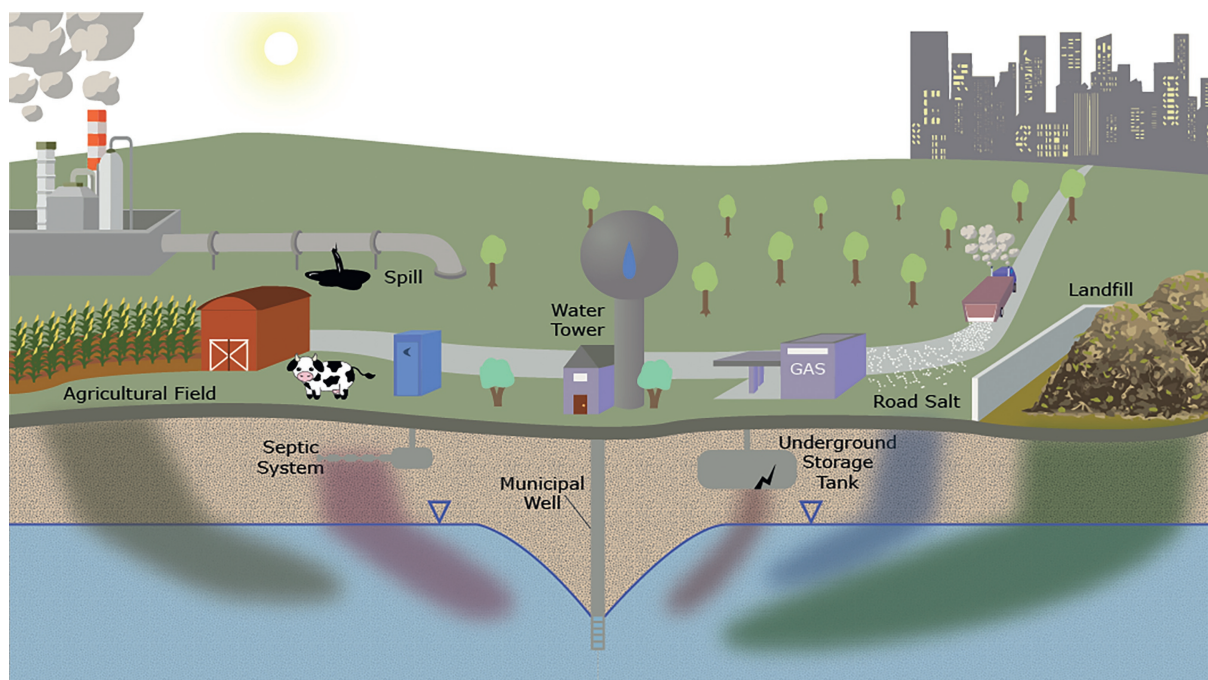


Fig. 3 Examples of point and diffuse sources of contamination resulting in zones of soil and groundwater contamination in the vicinity of a public water supply well.

including viruses. Although there are now effective methods for sampling and analysis of viruses in well water, such testing is rarely done, so that the frequency of illness attributable to viruses in groundwater is poorly known.

Groundwater as a natural and anthropogenic resource

Groundwater, as a natural resource, plays a fundamental role in supporting aquatic ecosystems, surface water flows, reservoir levels and diverse anthropogenic demands. While groundwater accounts for over 95% of the planet's liquid fresh water, it also provides approximately 50% of the total drinking water supply worldwide and 43% of all water used for irrigation (UNESCO, 2015, 2022). As surface water allocations reach maximum sustainable limits, groundwater will be required to meet increasing water demands from all sectors. Under projected warming climatic scenarios, the availability of surface water resources may become less reliable in some regions (Erler et al., 2018). Because groundwater tends to be more protected from variable climatic conditions due to the immense storage capacity within the subsurface reservoirs and protection from losses through evapotranspiration, it represents a relatively resilient and secure water source, particularly under conditions of extended drought.

Although groundwater is considered a renewable resource, being replenished through the annual precipitation cycle, the sustainable extraction capacity is finite. Overexploitation of aquifer systems worldwide has resulted in a continual lowering of groundwater levels, land subsidence and depletion of aquifer storage volume (Scanlon et al., 2012). The implications of lowering groundwater levels are perhaps most acutely observed in relation to the effect on surface water systems and terrestrial aquatic ecosystems. Groundwater supplies a relatively continuous form of replenishment to surface water systems including streams and lakes in the form of baseflow (de Graaf et al., 2019). The release of relatively warm groundwater during frozen winter months in cold climatic regions, sustains surface water flow beneath ice cover, which is essential for aquatic wildlife. Discharging groundwater also carries both essential nutrients and other geochemical constituents vital to sustaining aquatic ecosystems, but at the same time can also carry potential contaminants present within the groundwater system (Lewandowski et al., 2020).

One of the most essential, yet poorly understood roles of regional groundwater flow is the critical support it provides in sustaining Earth's fragile wetlands. Wetlands exist and are sustained in areas of continuous shallow water table conditions where the roots of phreatophytes, common in wetland environments, can access the required water supply from groundwater (Fig. 4). Lowering of water tables due to groundwater extraction activity or changing recharge conditions can result in serious and permanent damage to wetland features, in some cases causing their complete disappearance (Poeter et al., 2020; de Graaf et al., 2019).

Long-term depletion of global groundwater resources is also threatening the sustainability of irrigation practices worldwide (e.g., Scanlon et al., 2012). Attempts to mitigate these threats involve the intentional enhancement of groundwater recharge through a variety of approaches, including water harvesting. Redirection of runoff waters from impermeable surfaces or other drainage features to constructed infiltration basins facilitates soil water replenishment and groundwater recharge, augmenting subsurface water required to support the agricultural demand.

The progressive degradation of groundwater quality as a result of both natural and man-made contamination processes, as summarized above, represents another major threat to the long-term utility of this strategically-significant resource. For example, it

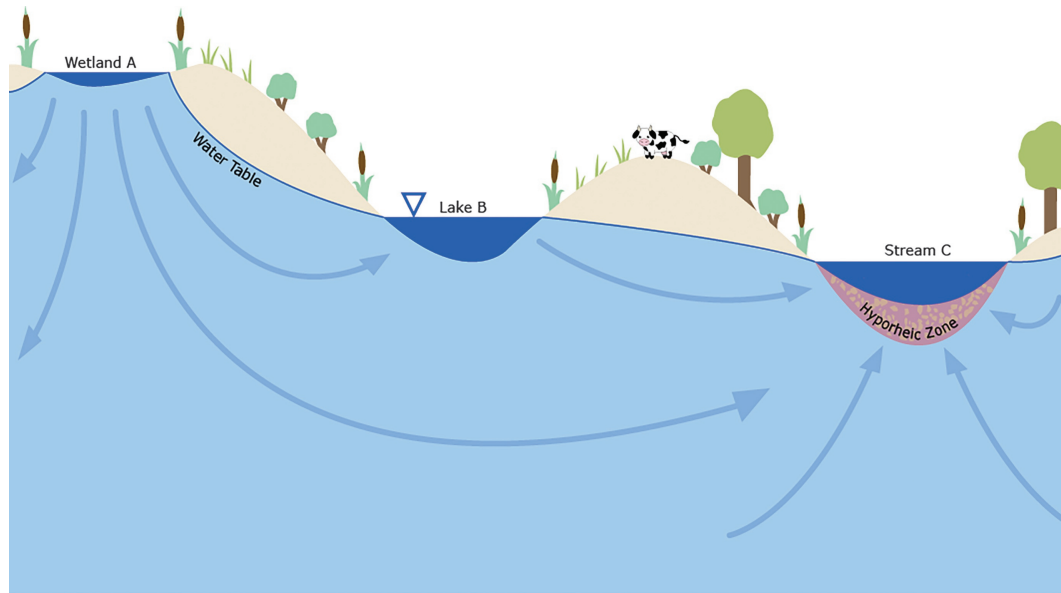


Fig. 4 Groundwater-surface interactions within a regional groundwater flow system showing surface water bodies under losing (Wetland A), flow-through (Lake B), and gaining conditions (Stream C). The hyporheic zone, a region of intense surface-groundwater mixing, is shown beneath Stream C.

is estimated that approximately 25% of all public supply wells in the United States produce groundwater that has concentrations of at least one contaminant of human health concern that is above the recommended drinking water limit (Toccalino and Hopple, 2010). As noted above, discharging groundwater with degraded water quality can also have serious impacts on the health of surface waters, wetlands, soils and associated ecological environments. The interaction between groundwater and surface water systems, under either recharging or discharging conditions, occurs at the terrestrial surface within the soil zone. The important role this shallow component of the Critical Zone plays in controlling the exchange of water and geochemical constituents between groundwater and surface water is explored in more detail in the next section.

Impact of soils on groundwater

Soil influence on groundwater recharge and flow

Water that percolates below the root zone and replenishes groundwater (i.e., reaches the water table) is termed groundwater recharge. A simple water budget expression for the soil zone can be used to estimate groundwater recharge as follows:

$$R = P + Q_{in} - ET - Q_{out} - \Delta S \quad (1)$$

where R is groundwater recharge, P is precipitation, ET is evapotranspiration, Q_{in} and Q_{out} are lateral flows of water into or out of the region of interest, respectively, in the form of surface or subsurface flows, and ΔS is the change in soil water storage. All components are expressed as rates, typically normalized to the area of interest (e.g., mm d^{-1}) but they can also be expressed as volumetric rates (e.g., $\text{m}^3 \text{d}^{-1}$).

The timing, magnitude and distribution of groundwater recharge varies considerably based on soil conditions, topography, vegetation, and climate, and the corresponding complex interplay between these different factors (Keese et al., 2005; Scanlon et al., 2006). Soil zone properties, including texture, structure and chemical or mineralogical composition, determine how water is partitioned between surface (runoff) and subsurface (infiltration) during precipitation events, affecting the amount of water in the soil zone available for potential groundwater recharge. A significant structural feature in most soil systems is macro-porosity, which refers to large continuous pores within the soil matrix normally formed by vegetation, soil fauna, desiccation or weathering after initial soil genesis. Preferential flow along soil macropore networks allows water to be rapidly transported vertically or laterally in the subsurface, bypassing the soil matrix, increasing the risk of groundwater contamination (Beven and Germann, 2013). Coarse-textured soils typically have greater infiltration capacity, making them more amenable to groundwater recharge, but preferential flow through macro-porosity in fine-grained soils can also lead to significant groundwater recharge. In agricultural settings, tillage practices and soil compaction from vehicular traffic can alter soil structure and macro-porosity, thereby affecting soil infiltration and drainage characteristics. The composition, structure and permeability of the native geologic materials underlying the soil zone will also impact groundwater recharge. Layers with little permeability in the vadose zone can result in perched water table conditions (isolated zones of water saturation above the regional water table), affecting subsurface flow patterns and rates of groundwater recharge.

Vegetation and land management conditions influence recharge potential in two main ways: (1) by altering soil infiltration characteristics, and (2) perhaps more importantly, they are a major control on ET losses from the soil zone which can limit groundwater recharge rates. Different land covers (e.g., vegetated vs. non-vegetated) or vegetation types will change ET rates, altering soil moisture conditions and the amount of water available for recharge (Keese et al., 2005). Vegetation is often closely linked with climate, which is another key control on groundwater recharge. Climate determines the amount, timing and type of precipitation inputs as well as the evapotranspirative demand, with recharge rates generally greater in humid climates than in arid or semi-arid regions. In many agricultural regions, irrigation inputs can further contribute to groundwater recharge. Due to interactions between various factors, the combination of climate and geologic conditions give rise to regions with similar ecological communities (termed ecoregions), physiography, or soil pedology that often share common recharge dynamics.

Topography drives groundwater flow systems, with groundwater recharge occurring in topographically higher areas and groundwater discharge in lower lying areas as described by Toth (1963), leading to the adage that the water table is considered a “subdued replica of topography” in many humid regions. However, groundwater flow patterns are complicated by undulations in topography and heterogeneous subsurface geologic formations, resulting in more complex local and regional flow systems (Fig. 5). Furthermore, climatic conditions can alter groundwater recharge patterns, particularly in more arid regions with seasonal precipitation or snow-melt dominated settings. In these regions, surface water runoff during rainy seasons or spring melt is shaped by topography and contributes to localized groundwater recharge beneath topographic lows (e.g., wadis, playas, wetlands). A classic example is depression-focused recharge, where snowmelt runoff from seasonally-frozen soils collects in topographic depressions and, following soil thaw in spring, results in groundwater recharge that varies both spatially and seasonally (Hayashi et al., 1998; van der Kamp and Hayashi, 1998).

Soil influence on groundwater quality

Water that migrates through the soil zone and recharges groundwater is subject to natural and anthropogenic influences that strongly influence groundwater quality. The soil zone is often referred to as a natural water filter and plays an important role in altering water quality through physical, geochemical, and biological processes. The physical filtration capacity of soil is determined by a range of soil characteristics, including porosity, pore geometry (e.g., size, shape, connectivity), structure, clay content, organic matter content, and soil particle physicochemical properties. Physical filtration is generally considered to be the removal of suspended colloidal or insoluble particles, including clay particles and microbes (e.g., viruses, bacteria, protozoa), during infiltration or percolation of water through the soil. Other mechanisms of filtration in partially saturated soils include pore straining, film straining, and surface deposition or adsorption. While these processes are generally effective at removing a large fraction of colloids (often >99%) in many soils, the importance of preferential flow in structured soils is increasingly being recognized as a critical pathway for colloidal (especially microbial) transport that can result in rapid movement of microbes to groundwater systems (Taylor et al., 2004).

Geochemical reactions in soil are an important attenuation mechanism for many dissolved groundwater contaminants of concern. Soils have charged particle surfaces that can transfer dissolved ionic species from solution in the process of sorption (attachment to the surface) and desorption (detachment from surfaces). Since most soils contain negatively charged surfaces, species such as Na^+ , Ca^{2+} , Mg^{2+} undergo cation exchange reactions that alter both soil properties and porewater solution compositions. Other solute species such as phosphate and many heavy metals are commonly attenuated by soil sorption, although the process can be reversible meaning that contaminant migration is slowed (termed retardation) but the contaminants are not completely removed and may be remobilized under certain conditions. Organic contaminants, including pesticides and petroleum hydrocarbons, are also subject to sorption and desorption processes that are largely controlled by the amount of organic matter in the soil profile. Reduction-oxidation (or redox) reactions in soil and groundwater are important in dictating the chemical composition of groundwater along a flow path. The natural exposure of soils to atmospheric oxygen means that soil porewater is initially oxidized (oxic). Recharge below the water table decreases or eliminates exposure to atmospheric oxygen, causing the water to eventually evolve to more reducing (or anoxic) conditions along the flow path. The persistence of dissolved oxygen (aerobic conditions), which is related to microbial activity, carbon availability and other biogeochemical factors, will affect groundwater redox conditions and, in turn, strongly influence other key redox sensitive elements like nitrogen, iron, manganese, and sulfur. The role of soil organic matter and denitrification in controlling groundwater nitrate concentrations is an excellent example of the important role that redox reactions play in the soil profile. Nitrate leaching from the soil zone is a major risk to groundwater in many agricultural landscapes, but denitrification of nitrate to N_2 gas can help to reduce nitrate loads and mitigate these risks. Thermodynamically, denitrification first requires the depletion of dissolved oxygen, which is often achieved through oxidation of dissolved organic carbon from the soil zone or other electron donors where present (e.g., sulfide oxidation).

The third main type of soil “filtering” process is biological, with organisms (predominantly micro-organisms) that proliferate in the soil zone responsible for the transformation or degradation of potential contaminants. Soil microbial communities are incredibly diverse and abundant, with global soil microbial biomass (expressed in terms of carbon mass) on the order of tens of billions of tonnes, equating to roughly 10^5 – 10^9 organisms cm^{-3} (Serna-Chavez et al., 2013; Magnabosco et al., 2018). The majority of microorganisms are present in the top meter or two of the soil profile, but they are adapted to survive in environments up to kilometers below the Earth’s surface (Magnabosco et al., 2018). Degradation of pollutants by microbes is termed biodegradation, and is important for the removal of a wide range of organic compounds, notably pesticides and petroleum hydrocarbons, from soil and groundwater systems. Potential contaminants are typically broken down into less toxic or non-toxic products by native

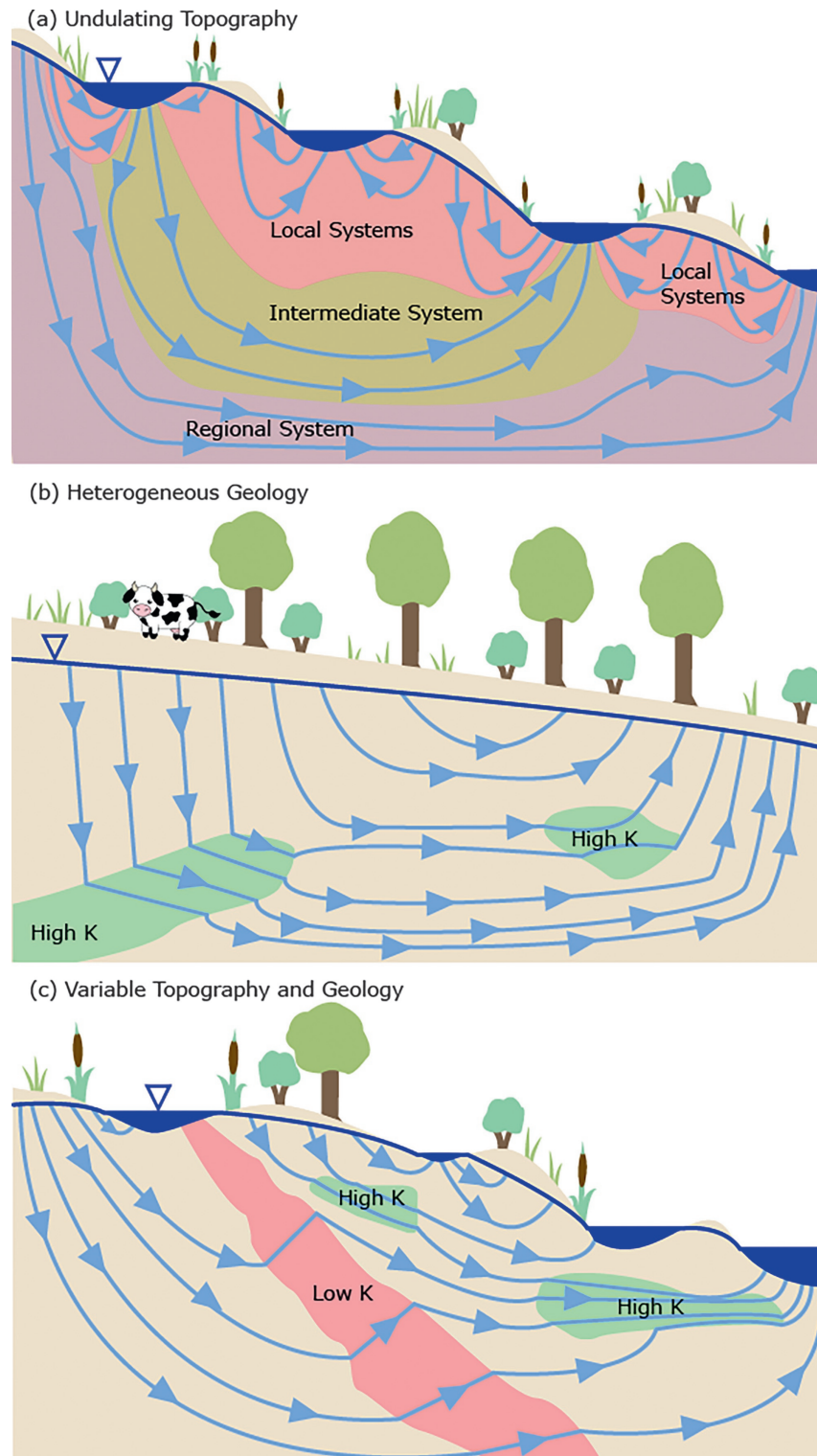


Fig. 5 Cross-sectional examples of groundwater flow systems demonstrating the important roles of (a) topography, (b) geologic heterogeneity, and (c) complex interaction between topography and geology in controlling subsurface groundwater flow patterns. K is the hydraulic conductivity. Delineation of characteristic local, intermediate, and regional groundwater flow systems are shown in the top diagram.

microbial communities that rely on the reactions as an energy source. Micro-organisms are also involved in transformation reactions of inorganic compounds and metals, with many geochemical reactions being microbially mediated. Key redox reactions, such as denitrification noted above, depend on natural microbial communities present in the subsurface to proceed. As such, the natural attenuation of potential contaminants offered by soils serves to protect underlying groundwater quality from degradation (Chamy and Rosenkranz, 2013).

Biogeochemical processes can also result in the introduction of pollutants to soil pore water and groundwater. Dissolution of chemical compounds naturally present in soil or underlying geologic materials will result in groundwater contamination if concentrations exceed an acceptable concentration limit (often related to drinking water or some other intended water use). A classic example is the extensive groundwater arsenic contamination that has been identified in regions such as Bangladesh, India, Mongolia, China, Argentina, United States and Mexico (Podgorski and Berg, 2020), but dissolution of natural contaminants such as nitrate, fluoride, asbestos and chromium has also been reported around the world. Redox and acid-base reactions in soil can exacerbate groundwater contamination. Carbon cycling in soils alters pore water chemistry via dissolution or degassing of atmospheric and plant-produced carbon dioxide, typically resulting in carbonic acid production that promotes weathering and dissolution. In soils or geologic materials with abundant sulfide minerals, sulfide oxidation can further acidify the soils resulting in the dissolution and mobilization of heavy metals to groundwater. This process is particularly noteworthy at mining sites, where it has attracted the term acid mine drainage (Nordstrom, 2011).

Sources of groundwater contamination typically originate from a wide range of natural and anthropogenic sources at the surface, which vary in spatial extent (e.g., point, non-point or diffuse, linear) and duration (e.g., one-time spill, continuous input). Diffuse sources with continuous loading can result in significant accumulation of contaminants stored in the vadose zone at a regional scale, which are then carried into groundwater during recharge events. Vadose zone and groundwater flow velocities are much slower than surface water flows, resulting in long residence times that can range from days to decades (or longer), depending on groundwater recharge rates. Consequently, long-term contaminant loadings, such as the buildup of soil nitrogen in agricultural watersheds, creates a legacy of contamination for a long time into the future. Given the long lag times, changes in land or water management practices implemented today could take long times (decades or longer) to achieve improvements in groundwater quality in the absence of other natural or engineered attenuation mechanisms (Rudolph et al., 2015).

Land management influence

Land management decisions shape the hydrological and geochemical function of soils in different settings, in turn affecting groundwater conditions. The starkest differentiation is between urban and rural soil environments, with noteworthy influences outlined further below.

Rural soil environments

Globally, agriculture land use comprises about 50% of the habitable land area (Ritchie and Roser, 2019). In these rural areas, agricultural and land management practices can dramatically alter the soil environment and the local water balance, ultimately influencing rates and timing of groundwater recharge. A global meta-study by Mohan et al. (2018) has shown that land use and vegetation (land cover) conditions, together with climatic factors (precipitation and ET) were the most important controls on groundwater recharge. Conversion of land to agricultural production in many regions of the world involves replacing deep-rooted perennial native vegetation (including forest or shrub lands) with shallow-rooted annual crops, reducing ET and consequently increasing groundwater recharge rates, in some instances by up to one or two orders of magnitude (Han et al., 2017; Scanlon et al., 2007). Similarly, changing cultivation practices, such as the conversion from perennial grasses to annual row crops or the adoption of certain tillage practices, can alter soil structure and affect soil water partitioning (infiltration-runoff-ET) during and after precipitation events, effectively increasing deep drainage and groundwater recharge potential. As noted earlier, soil compaction from agricultural vehicles can also significantly influence shallow soil characteristics and reduce infiltration capacity.

The combination of soil and climate conditions in a region often results in a pronounced seasonality to groundwater recharge. Using stable water isotope data from 54 locations around the world, Jasechko et al. (2014) showed a greater proportion of groundwater recharge over winter seasons in arid and temperate regions, and more recharge in wet seasons in tropical regions. The greater relative contribution of winter precipitation to groundwater recharge in arid and temperate regions can be attributed in part to reduced ET potential over winter. In addition, for snowmelt driven watersheds in higher latitude regions the proportion of winter recharge can increase further due to the rapid input of water during spring freshet when soils are still frozen, often resulting in depression-focused recharge described earlier. As one might intuitively expect, groundwater recharge in the tropics is driven largely by high-intensity rainfall events, occasionally following a similar depression-focused recharge behavior, and resulting in the largest proportion of annual recharge during the wet season.

Irrigation and drainage of agricultural lands both have important consequences for groundwater systems. Irrigation, which is essential for agricultural production in many arid and semi-arid regions, will affect underlying groundwater resources in different ways depending on the source of water used for irrigation. Where irrigation water is derived from local or imported surface water sources, enhanced groundwater recharge occurs. This is due to increased water inputs that result in drainage below the root zone, termed irrigation return flows, as well as water losses from irrigation canals and conveyance systems. If irrigation water is sourced from local groundwater systems, groundwater pumping will result in a decline in groundwater levels, which may or may not be offset by increased groundwater recharge rates associated with irrigation return flow. Normally in these situations, only a fraction of

the water pumped for irrigation is returned to the aquifer in the form of groundwater recharge, hence the net result of irrigation is a decline in groundwater levels. In more humid regions or agricultural areas with poorly draining soils, artificial subsurface drainage systems including mole drains and tile drains may be used to artificially lower the water table and promote more timely and effective access to a field for agricultural operations. Such artificial drainage of the landscape drastically alters both the surface and subsurface hydrologic flow patterns in a catchment, with important consequences for transport of any potential contaminants (Golmohammadi et al., 2017). In essence, shallow groundwater flows into the shallow drains and is rerouted horizontally via the interconnected drainage network to nearby surface water drainage pathways. In areas of natural groundwater recharge, this would theoretically reduce the rates of recharge and any potential groundwater contaminant loads. However, even though shallow enhanced drainage systems clearly alter the near-surface soil water balance, the corresponding influence on groundwater recharge rates and water quality is not fully understood and is still a topic of active research.

Agricultural land use in rural watersheds is often associated with the input of nutrients and agrichemicals to improve agricultural productivity, but these compounds can be leached through the soil profile and degrade groundwater quality. Organic (e.g., manure, compost) and inorganic fertilizer inputs in areas of more intensive agriculture result in the export of nitrogen (typically as nitrate), phosphate, chloride and other chemical constituents (e.g., Ca, Mg) to underlying groundwater systems. While it can be difficult to attribute groundwater nitrate contamination to specific sources owing to the complexities of soil N cycling, multiple N sources (e.g., manure, inorganic fertilizer, organic matter mineralization, atmospheric N₂ fixation), and subsurface attenuation mechanisms (e.g., denitrification), there is clear evidence that nitrate contamination at watershed scales is strongly correlated to agricultural land use (Böhlke, 2002). In contrast, pesticides (and their metabolites) are often detected in shallow groundwater, but concentrations tend to be low due to sorption and relatively rapid degradation in the soil profile. Enhanced groundwater recharge, whether due to land use alteration or irrigation, has also been shown to have important consequences for soil and groundwater salinization, which is described in more detail in the following section.

Urban soil environments

Paving and the proliferation of impervious surfaces in urban environments reduces the soil infiltration potential and increases stormwater runoff, suggesting a reduction in groundwater recharge potential. However, leakage from water distribution and sewer infrastructure networks provides additional sources of groundwater recharge and ET decreases due to the reduction in bare soil and vegetated surfaces. As a result, several studies have noted increased rates of groundwater recharge in urban settings (Lerner, 2002; Tubau et al., 2017) and corresponding changes in the spatial distribution of recharge are expected but less well documented.

Many municipalities are moving toward green stormwater infrastructure (GSI) and low-impact development technologies to better manage stormwater flows in the face of increasing urbanization and changing climate patterns. GSI essentially involves stormwater management approaches that rely on soil processes, notably infiltration, soil water storage and ET, to conserve or restore natural hydrological and ecological function in urban environments (Chini et al., 2017). Examples of GSI technologies include green roofs, bioswales and bioretention systems, rain gardens, and permeable pavements. Most GSI systems involve vegetated “green” spaces that store or infiltrate stormwater runoff collected from nearby impervious surfaces (e.g., roofs, parking lots, roadways), thereby reducing stormwater runoff volumes and improving water quality for downstream receptors. While GSI has many benefits, it is not yet clear what influence they would have on urban groundwater systems. Enhanced localized infiltration beneath GSI systems has the potential to increase groundwater recharge or induce lateral subsurface flow that could adversely impact other urban infrastructure (e.g., water inflows to basements), but there is little research to date on these potential concerns.

Urbanization brings with it the storage and utilization of a wide range of chemicals that can threaten groundwater supplies. Fertilizers and pesticides are routinely applied in urban landscapes, creating the same issues that arise in agricultural regions (see above). Commercial and industrial operations in urban centers have historically stored, utilized, or disposed of a wide range of potential contaminants, including underground fuel tanks, septic systems (nutrients, microbial contaminants), chlorinated and volatile organics, polyaromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs), and metals. Municipal wastes are frequently disposed of in landfills that become localized and concentrated sources of these same contaminants. In cold regions, the application of road salts to improve winter driving and pedestrian conditions leads to the accumulation of salts in soil and underlying groundwater. Similarly, PAHs originating from asphalt paved roadways or burning of hydrocarbon fuels can leach to groundwater. Water leakage from urban water distribution and sewer systems can become distributed linear sources of contamination, releasing water treatment byproducts (e.g., chloroform, nutrients, detergents and surfactants, pathogens) or other emerging contaminants (e.g., pharmaceuticals, microplastics, nanomaterials and per- and polyfluorinated substances (PFAS)) to the subsurface. Little is currently known about the capacity of soils to retain or remove many emerging contaminants, and when combined with typically long vadose zone and groundwater storage times described earlier, there is the potential for a legacy of future water quality concerns in these environments.

Groundwater impact on soils

Just as soils play an important role in influencing groundwater systems, groundwater conditions will similarly impact the soil zone through various interconnections and feedback mechanisms. One prime example of these feedback mechanisms is water table depth, which tends to be maintained within a certain range by interactions between the soil and groundwater zones. Increasing rates of groundwater recharge will tend to cause the water table to rise, which in turn can increase soil moisture due to capillary rise

above the water table. The increasing soil moisture will tend to result in greater rates of ET, thereby reducing the groundwater recharge potential and stabilizing the water table within a certain range below the ground surface. In the other direction, declining water tables can lower soil moisture conditions, reducing ET (and soil hydraulic conductivity), leading to a reduction in vertical water movement and limits on water table decline. In cold regions where soils freeze seasonally, subzero temperatures cause the formation of pore ice and a corresponding decrease in soil water potential (or increase in matric suction), in a process called cryosuction. The result is a vertical migration of soil moisture toward the freezing front from unfrozen soil below, which can create ice lenses, generate frost heaving, and cause freezing-induced declines in the water table.

Groundwater-based interactions between the water table, soil capillarity, and vegetation (i.e., ET demand) can also have important consequences for soil quality and agricultural productivity, primarily due to salinization and waterlogging of soils. As noted earlier, increased groundwater recharge due to land alteration or irrigation can raise the water table to near the surface, where capillarity and evapotranspiration combine to concentrate soluble salts in the soil zone. The increased soil salinity and poor drainage characteristics (i.e., waterlogging) associated with rising groundwater levels, which tend to be exacerbated in arid climates and low-lying areas, have degraded soil quality and agricultural productivity in large tracts of land across the globe (Han et al., 2017; Scanlon et al., 2007). In a similar way, natural groundwater discharge can alter soil conditions and result in the accumulation of salts and other minerals at Earth's surface. Groundwater discharge carries with it many dissolved compounds that can either (a) be discharged to surface water bodies, or (b) if ET exceeds the rate of groundwater discharge, be deposited by chemical precipitation in surficial soils. Soil pore water in many groundwater discharge areas is therefore often characterized by elevated levels of dissolved solids, resulting in localized zones of soil salinity, sodicity, and enrichment of minerals from geologic weathering (e.g., sulfur, iron, lithium) that can be both beneficial (e.g., mining, animal salt licks) and harmful (e.g., soil hardpans, saline soil degradation).

As part of the natural hydrologic cycle, groundwater discharge to the Earth's surface is important to many surficial systems. Groundwater discharge via seeps and springs provides baseflow that sustains flow in streams and rivers and water levels in lakes and wetlands. The shallow zone immediately around stream beds, where groundwater-surface interactions are more prevalent, is termed the hyporheic zone (Fig. 4) and is responsible for enhanced storage and exchange of water, thermal energy, nutrients, contaminants, and dissolved gases (oxygen) that are vitally important for riparian vegetation and aquatic organisms.

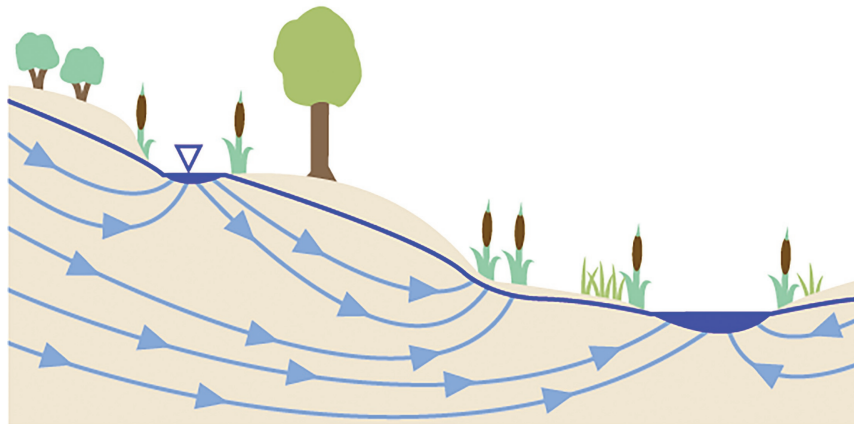
In recent years, there has been increasing recognition that groundwater plays a critical role in supporting aquatic and terrestrial ecosystems, providing valuable ecosystem services in a wide range of landscapes (Griebler and Avramov, 2015). The function of these groundwater-dependent ecosystems is often not fully understood, with groundwater contributions varying spatially (direct or indirect discharge, large or small area) and temporally (perpetual, seasonal, intermittent), and ecosystems influenced by both water quantity and quality. Groundwater dependent ecosystems and environmental flows in rivers and streams are at risk from excessive groundwater pumping, with global estimates indicating that approximately 42%–79% of watersheds with groundwater pumping are at risk of reaching environmentally sustainable stream flow limits by 2050 (de Graaf et al., 2019).

Subsurface geology, particularly the distribution of rock and sediment permeability, is one of the most important controls on groundwater flow, thus influencing the distribution, rate, and timing of groundwater discharge. Wetlands with groundwater dependent vegetation and hydric soils are typically situated in discharge areas at the end of subsurface flow paths that are largely determined by surface topography and subsurface geologic heterogeneity (USGS, 1996). For example, groundwater fed wetlands are often situated near topographic depressions or breaks in slope, where more permeable strata outcrop at surface, or above the downgradient end of laterally discontinuous aquifer units (Fig. 6). Flow pathways and landscape position will also affect the wetland hydroperiod, or the duration and portion of the year when inundated with water, thereby regulating soil formation and vegetation characteristics.

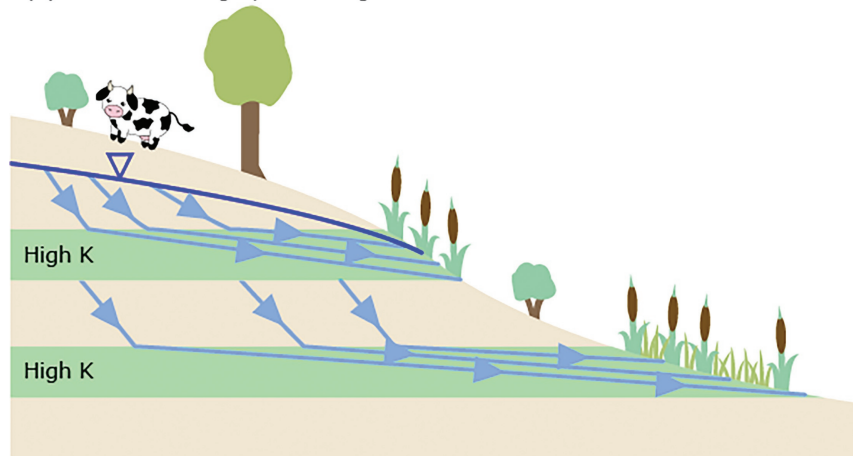
Conclusions

Soils play a critical and complex role in controlling the hydrologic, chemical, and biological interaction between the land surface and the underlying groundwater system, a region often referred to as the shallow Critical Zone. Many aspects of this interaction are poorly understood and require additional investigation and research. This has become particularly evident in light of the potential influence of climate warming. There is a dynamic synergy between how soil processes influence groundwater quantity and quality, and how shallow groundwater flow systems control soil water content, soil mineral and pore water composition, and soil ecologic capacity. The combined influence of rapid and extensive anthropogenic change and evolution within the hydrologic cycle as weather patterns change is leading to noticeable, yet highly unpredictable impacts within both the near-surface soil environment and the deeper subsurface groundwater regime.

(a) Depressions and Slope Breaks



(b) Areas of Stratigraphic Change



(c) Subsurface Geology

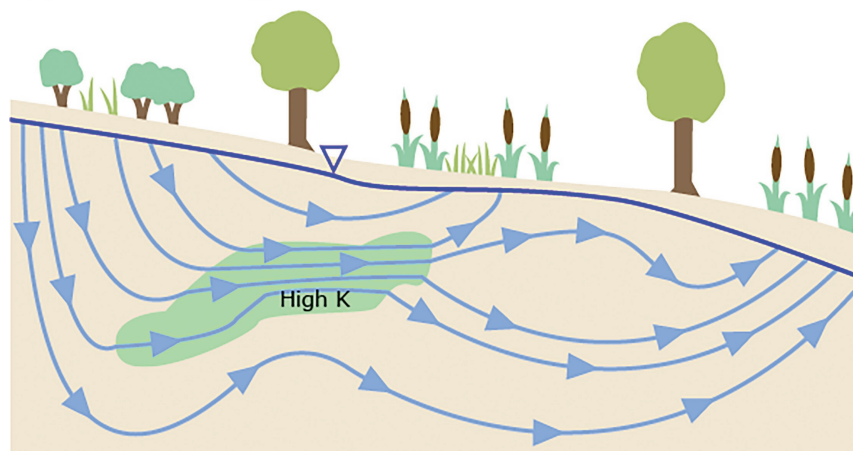


Fig. 6 Groundwater flow cross-sections illustrating common locations of ground-fed wetlands (a) at topographic depressions and breaks in slope, (b) at changes in geologic strata, and (c) within ground discharge areas controlled by subsurface geology (note the wetland situated above the high permeability unit outlined in green). Modified from USGS (1996) *National Water Summary on Wetlands Resources*. United States Geological Survey. Water Supply Paper 2425. <https://doi.org/10.3133/wsp2425>.

As urbanization continues, groundwater extraction increases, and agricultural land management expands, groundwater recharge and the critical soil processes that sustain it are changing. In some areas, this is resulting in reductions in groundwater replenishment and deterioration of groundwater and soil quality. The capacity for the soil system to function as a natural filter, protecting subsurface water from contaminant migration, is beginning to decline as the reactive capacity of the soils progressively decrease through the long-term exposure to various chemical species. This is already having profound effects on groundwater quality as infiltrating contaminant mass builds up in the soil zone. The chronic concentration of minerals in the shallow soil system due to the combined effects of mineralized groundwater discharge and evapotranspiration is leading to soil degradation including extensive soil salinity. The potential interactive processes within the soil profile related to emerging contaminants such as microplastics and nanoparticles, however, remain largely unknown.

The storage capacity and ability of the soil profile and vadose zone to control infiltration rates and vertical pore water mobility through processes related to partially saturated flow conditions and capillarity are crucial in preserving water within the subsurface environment. These processes are also fundamentally important in providing required chemical and biological reaction times necessary to naturally attenuate chemicals and pathogens before they reach groundwater. Anthropogenic modifications to the water cycle through the introduction of artificial subsurface drainage systems (e.g., tiles and mole drains) and irrigation can have profound impacts on these processes by altering soil characteristics, pore water content and quality, groundwater recharge and the nature of the interaction between shallow groundwater and surface water bodies. As agricultural water demands increase and water management strategies designed to maximize food production become more necessary, soil and groundwater conditions within the rural setting are likely to continue to degrade globally.

The incorporation of relevant soil processes and soil-groundwater interaction into studies of groundwater resource management and sustainability will be essential when considering either anthropogenic or ecological applications. This will be particularly true when employing models designed to predict soil, hydrologic, and ecosystem responses under rapidly changing climatic conditions. In many cases, these processes are still poorly understood and rapidly evolving and represent an area of critical future research.

References

- Beven K and Germann P (2013) Macropores and water flow in soils revisited. *Water Resources Research* 49(6): 3071–3092.
- Böhle JK (2002) Groundwater recharge and agricultural contamination. *Hydrogeology Journal* 10(1): 153–179.
- Chamy R and Rosenkranz F (2013) *Biodegradation: Life of Science*. London: IntechOpen. <https://doi.org/10.5772/52777>.
- Chini CM, Canning JF, Schreiber KL, Peschel JM, and Stillwell AS (2017) The green experiment: Cities, green stormwater infrastructure, and sustainability. *Sustainability* 9(1): 105.
- de Graaf IE, Gleeson T, Van Beek LPH, Sutanudjaja EH, and Bierkens MF (2019) Environmental flow limits to global groundwater pumping. *Nature* 574(7776): 90–94. <https://doi.org/10.1038/s41586-019-1594-4>.
- Earman S and Dettinger MD (2011) Potential impacts of climate change on groundwater resources – A global review. *Journal of Water and Climate Change* 2(4): 213–229. <https://doi.org/10.2166/wcc.2011.034>.
- Erlar AR, Frey SK, Khader O, et al. (2018) Simulating climate change impacts on surface water resources within a lake-affected region using regional climate projections. *Water Resources Research* 55(1): 130–155. <https://doi.org/10.1029/2018WR024381>.
- Fetter CW and Kremer D (2021) *Applied Hydrogeology*, 5th edn. Long Grove, IL: Waveland Press.
- Fetter CW, Boving T, and Kremer D (2018) *Contaminant Hydrogeology*, 3rd edn. Long Grove, IL: Waveland Press.
- Golmohammadi G, Rudra R, Prasher S, et al. (2017) Impact of tile drainage on water budget and spatial distribution of sediment generating areas in an agricultural watershed. *Agricultural Water Management* 184: 124–134.
- Griebler C and Avramov M (2015) Groundwater ecosystem services: A review. *Freshwater Science* 34(1): 355–367.
- Han D, Currell MJ, Cao G, and Hall B (2017) Alterations to groundwater recharge due to anthropogenic landscape change. *Journal of Hydrology* 554: 545–557.
- Hassani A, Azapagic A, and Shokri N (2020) Predicting long-term dynamics of soil salinity and sodicity on a global scale. *Proceedings of the National Academy of Sciences* 117(52): 33017–33027. <https://doi.org/10.1073/pnas.2013771117>.
- Hayashi M, van der Kamp G, and Rudolph DL (1998) Water and solute transfer between a prairie wetland and adjacent uplands, 1. Water balance. *Journal of Hydrology* 207(1–2): 42–55.
- Jasechko S, Birks SJ, Gleeson T, et al. (2014) The pronounced seasonality of global groundwater recharge. *Water Resources Research* 50(11): 8845–8867.
- Keese KE, Scanlon BR, and Reedy RC (2005) Assessing controls on diffuse groundwater recharge using unsaturated flow modeling. *Water Resources Research* 41(6): W06010.
- Lerner DN (2002) Identifying and quantifying urban recharge: A review. *Hydrogeology Journal* 10(1): 143–152.
- Levy ZF, Jurgens BC, Burow KR, et al. (2021) Critical aquifer overdraft accelerates degradation of groundwater quality in California's Central Valley during drought. *Geophysical Research Letters* 48(17), e2021GL094398. <https://doi.org/10.1029/2021GL094398>.
- Lewandowski J, Meinikmann K, and Krause S (2020) Groundwater–surface water interactions: Recent advances and interdisciplinary challenges. *Water* 12(1): 296. <https://doi.org/10.3390/w12010296>.
- Lin H (2010) Earth's Critical Zone and hydrogeology: Concepts, characteristics, and advances. *Hydrology and Earth System Sciences* 14(1): 25–45.
- Magnabosco C, Lin LH, Dong H, et al. (2018) The biomass and biodiversity of the continental subsurface. *Nature Geoscience* 11(10): 707–717.
- Mohan C, Western AW, Wei Y, and Saft M (2018) Predicting groundwater recharge for varying land cover and climate conditions—A global meta-study. *Hydrology and Earth System Sciences* 22(5): 2689–2703.
- Nordstrom DK (2011) Hydrogeochemical processes governing the origin, transport and fate of major and trace elements from mine wastes and mineralized rock to surface waters. *Applied Geochemistry* 26(11): 1777–1791.
- Podgorski J and Berg M (2020) Global threat of arsenic in groundwater. *Science* 368(6493): 845–850.
- Poeter E, Fan Y, Cherry JA, Wood W, and Mackay D (2020) *Groundwater in Our Water Cycle – Getting to know Earth's most important fresh water source*. The Groundwater Project, Guelph, Ontario, Canada <https://doi.org/10.21083/978-1-7770541-1-3>.
- Ritchie H and Roser M (2019) *Land Use*. Published online at OurWorldInData.org. Retrieved from: <https://ourworldindata.org/land-use>.
- Rudolph DL, Devlin JF, and Bekeris L (2015) Challenges and a strategy for agricultural BMP monitoring and remediation of nitrate contamination in unconsolidated aquifers. *Groundwater Monitoring and Remediation* 35(1): 97–109.
- Scanlon BR, Keese KE, Flint AL, et al. (2006) Global synthesis of groundwater recharge in semiarid and arid regions. *Hydrological Processes* 20(15): 3335–3370.

- Scanlon BR, Jolly I, Sophocleous M, and Zhang L (2007) Global impacts of conversions from natural to agricultural ecosystems on water resources: Quantity versus quality. *Water Resources Research* 43(3): W03437.
- Scanlon BR, Faunt CC, Longuevergne L, and McMahon PB (2012) Groundwater depletion and sustainability of irrigation in the US High Plains and Central Valley. *Proceedings of the National Academy of Sciences* 109(24): 9320–9325. <https://doi.org/10.1073/pnas.1200311109>.
- Serna-Chavez HM, Fierer N, and Van Bodegom PM (2013) Global drivers and patterns of microbial abundance in soil. *Global Ecology and Biogeography* 22(10): 1162–1172.
- Shiklomanov LA (1993) World freshwater resources. In: Gleick PH (ed.) *Water in Crisis: A Guide to World's Freshwater Resources*, pp. 13–24. New York: Oxford University Press.
- Taylor R, Cronin A, Pedley S, Barker J, and Atkinson T (2004) The implications of groundwater velocity variations on microbial transport and wellhead protection—review of field evidence. *FEMS Microbiology Ecology* 49(1): 17–26.
- Toccalino PL and Hopple JA (2010) *The quality of our Nation's waters: Quality of water from public-supply wells in the United States, 1993–2007: Overview of major findings*. U. S. Geological Survey Circular 1346, <https://doi.org/10.3133/cir1346>.
- Toth J (1963) A theoretical analysis of groundwater flow in small drainage basins. *Journal of Geophysical Research* 68(16): 4795–4812.
- Tubau I, Vázquez-Suñé E, Carrera J, Valhondo C, and Criollo R (2017) Quantification of groundwater recharge in urban environments. *Science of the Total Environment* 592: 391–402.
- UNESCO (2015) *The United Nations World Water Development Report 2015: Water for a Sustainable World*. Paris: France. ISBN: 978-92-3-100071-3. 122 pp.
- UNESCO (2022) *The United Nations World Water Development Report 2022: Groundwater: Making the Invisible Visible*. Paris: France. ISBN: 978-92-3-100507-7. 225 pp.
- USGS (1996) *National Water Summary on Wetlands Resources*. United States Geological Survey. <https://doi.org/10.3133/wsp2425>. Water Supply Paper 2425.
- van der Kamp G and Hayashi M (1998) The groundwater recharge function of small wetlands in the semi-arid northern prairies. *Great Plains Research* 8(1): 39–56.

Temperature effects of soil hydraulic properties

Jan W Hopmans^a and Robert Horton^b, ^aUniversity of California Davis, Davis, CA, United States; ^bIowa State University, Ames, IA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	380
Soil water retention	380
Unsaturated hydraulic conductivity	381
Unsaturated soil water flow modeling	382
Long-term implications	383
Summary	384
References	384

Abstract

Temperature dependence of soil hydraulic properties is typically accounted for by the so-called surface tension-viscous flow (STVF) theory, considering changing soil water properties only. However, there remains much uncertainty on temperature effects, as most reported research has been through laboratory experimentation on small soil cores using different methodologies for a limited range of soil types. Though effects of soil water properties can be measured accurately to verify the STVF theory, significant deviations are often reported, especially to quantify temperature dependence of the soil unsaturated hydraulic conductivity. In part, this is explained by soil temperature changing soil particle-soil solution interactions, as well as the soil's entrapped air and water volumes affecting pore connectivity and water retention.

Key points

- Effect of temperature on soil water properties is largely determined by changing water viscosity and surface tension (STVF theory).
- Full understanding of the temperature effects on soil hydraulic properties is incomplete because research has been limited to controlled laboratory studies, rather than in-situ field investigations.
- Temperature effects on field soil hydrology are especially relevant because of global warming implications.

Nomenclature

$\alpha_h(T)$	Temperature scaling factor of soil water retention curve (–)
$\alpha_K(T)$	Temperature scaling factor of hydraulic conductivity (–)
β	Conductivity parameter (–)
θ	Volumetric water content ($\text{m}^3 \text{m}^{-3}$)
θ_s	Saturated volumetric water content ($\text{m}^3 \text{m}^{-3}$)
θ_r	Residual volumetric water content ($\text{m}^3 \text{m}^{-3}$)
σ	Surface tension of water (Nm^{-1})
γ	Contact angle (degree)
ρ	Is the density of water (kg m^{-3})
l	Tortuosity parameter (–)
η	Connectivity parameter (–)
μ	Dynamic viscosity of water (N s m^{-2})
ν	Kinematic viscosity (m^2/s)
h_m	Soil water matric potential or matric pressure head, (m)
k	Intrinsic soil permeability (m^2)
g	Acceleration due to gravity (9.8 m s^{-2}).
C	Soil water capacity (m^{-1})
$D(\theta)$	Soil water diffusivity (m^2/s)
K	Soil unsaturated hydraulic conductivity (m s^{-1})
K_s	Saturated hydraulic conductivity (m s^{-1})

q_w	Darcy flux density (m s^{-1}),
S_e	Effective soil saturation (–)
T_r	Reference soil temperature ($^{\circ}\text{C}$)
z	Vertical soil position ($z > 0$, upwards, m)

Introduction

Soil temperature has a range of impacts on soil hydraulic behavior, driven primarily by the temperature dependence of the surface tension and viscosity of pore water. Improved understanding of temperature effects on soils will become increasingly relevant as climatic changes are likely affecting soil hydrologic processes with temperature increasing globally thereby altering both natural and agricultural ecosystems in ways yet to be determined.

Diurnal fluxes in soil water with depth are an example of temperature effects, driven by surface temperature changes that cause water to move from warmer to colder regions. Vapor phase transport is also influenced by the impact of temperature on the vapor pressure of water, described in the Heat and Water Transfer article.

Using physical principles as outlined below, unsaturated soil water flow is largely controlled by the physical arrangement of soil particles in relation to the water and air phases within the soil's pore space, as determined by pore size distribution and water-filled porosity or volumetric water content, θ (m^3 water/ m^3 bulk soil). In addition to θ , the volume of water is sometimes defined by degree of saturation, $S = \theta/\theta_s$, or effective saturation (S_e) as defined by

$$S_e = \frac{\theta - \theta_r}{\theta_s - \theta_r} \quad (1)$$

normalizing the mobile water between values of zero and one, while defining a residual water content, θ_r , for water that is considered immobile.

When considering temperature effects on the soil hydraulic properties, we will distinguish between soil water solution properties and soil-water interactions with other mechanisms that affect soil structure, pore geometry and pore connectivity. Temperature dependence of soil water properties only is determined by the so-called surface tension-viscous flow (STVF) theory.

Soil water retention

The soil-water retention function describes the relationship between the volumetric water content retained in soil, θ , and the soil capillary and adsorptive forces. It is also known as the soil-water release or soil-water characteristic function. These two water-retaining forces in unsaturated soils combined are defined as the matric forces and are sometimes also called suction forces. These suction forces increase as the size of the water-filled pores decreases, such as occurs by drainage, root water uptake or soil evaporation. When expressed relative to the reference potential of free water, the water potential in unsaturated soils is negative (the soil-water potential is less than the water potential of free water at atmospheric pressure). It is often referred to as the soil water matric potential. Hence, the matric potential decreases or becomes more negative as the soil water content decreases. In soils, water potential is usually expressed with respect to a unit volume of water, thereby attaining units of pressure (Pa) or per unit weight of water, so that the potential represents the equivalent height of a column of water (m). The pressure head equivalent of the combined adsorptive and capillary forces is defined as the matric pressure head, h_m . Since the matric forces are controlled by pore size distribution, specific surface area, and type of physico-chemical interactions at the solid-liquid interfaces, the soil water retention curve is soil-specific.

By way of the unique relationship between capillary water pressure and the radius of curvature of the air-water interface, and using the analogy between capillary tubes and the irregular pores in porous media, a relationship can be derived between h_m and its corresponding effective pore radius, r_e , or

$$|h_m| = \frac{2\sigma\cos\gamma}{\rho g r_e} \quad (2)$$

where σ and γ are defined as the surface tension and wetting angle (of wetting fluid with solid surface), respectively, ρ is the density of water, and g is the acceleration due to gravity (9.8 m s^{-2}). This capillary equation simplifies to $|h_m| = 0.15/r_e$, when both h_m and r_e are expressed in cm, for fully hydrophilic soils, as the surface tension is assumed to apply for pure water at room temperature. Moreover, in most studies of soil water flow and transport, isothermal soil conditions are assumed, ignoring the temperature dependency of the water parameters in Eq. (2).

However, in field conditions soil temperature is a dynamic variable that changes with both time and soil depth, so that soil and water parameters entering in Eq. (2) should be corrected for temperature changes. Whereas temperature effects on water properties such as surface tension and density are well defined, experimental confirmation of theoretical temperature effects on the soil water

retention are limited. Philip and de Vries (1957) were among the first to account for temperature effects on water properties, which was adapted by Hopmans and Dane (1986a) in their experimental work by relating the matric head at temperature T ($^{\circ}\text{C}$) to its corresponding value at a reference temperature, T_r , by a scaling factor $\alpha_h(T)$, or

$$h_m(T) = \alpha_h(T) \cdot h_m(T_r) \quad (3a)$$

where,

$$\alpha_h(T) = \frac{\left(\frac{\sigma}{\rho}\right)_T}{\left(\frac{\sigma}{\rho}\right)_{T_r}}, \quad (3b)$$

with $\alpha_h(T)$ decreasing as $T > T_r$. In addition to this work, others followed to validate such a theoretical correction, confirming that such correction for water properties is largely inadequate to account for temperature effects on the soil water retention curve. To explain the observed temperature effect discrepancies, Hopmans and Dane (1986b) investigated contributions of changing entrapped air volumes with temperature, whereas Nimmo and Miller (1986) suggested that solute concentrations of main organic substances decreased the surface tension of soil solution as compared to pure water, thereby affecting its temperature dependence differently. Later, Liu and Dane (1993) accounted for additional temperature effects by considering entrapped water pockets that transition to continuous soil water as temperature is increased. Experiments by Bachmann et al. (2001) were able to largely explain deviations because of significant changes in contact angle with temperature, and that therefore temperature effects were soil-specific. The hypothesis was confirmed by Joshi et al. (2019) from experiments using three different-textured soils.

Defining the soil water capacity as slope of the soil water retention curve by $C(h_m) = d\theta/dh_m$, one can derive that

$$C_T(h_m) = \frac{C(h_m, T_r)}{\alpha_h(T)}. \quad (4)$$

Unsaturated hydraulic conductivity

The relationship between the soil's unsaturated hydraulic conductivity, K , and volumetric water content, θ , is the other essential fundamental soil hydraulic property needed to describe soil water movement. It is also a function of the water and soil matrix properties, and determines water infiltration, evaporation and drainage rates, and is strongly affected by soil water content. It is defined by Darcy's equation, which relates the soil water flux to the total driving force for flow. Applying Darcy's law in the vertical dimension only, the magnitude of flow can be computed from the steady state flow equation (Darcy's law):

$$q_w = -K(\theta) \left(\frac{\partial h_m}{\partial z} + 1 \right), \quad (5)$$

where q_w is the Darcy volume flux density (m s^{-1}), z is vertical position ($z > 0$, upwards, m), and $K(\theta)$ denotes the unsaturated hydraulic conductivity (m s^{-1}). In this expression, the unsaturated hydraulic conductivity is related to the intrinsic soil permeability, k (m^2), by

$$K = \frac{\rho g k}{\mu} = \frac{g k}{\nu}, \quad (6)$$

where μ denotes the dynamic viscosity of water (N s m^{-2}), ν is the kinematic viscosity (m^2/s), and ρ and g are as defined before. The usage of permeability instead of conductivity allows application of the flow equation to fluids other than water with different density and viscosity values as k is defined as being a soils property only, independent of the transmitting fluid.

Soil hydraulic conductivity is typically assumed to be independent of temperature, yet from Eq. (6), it is now clear that ν is very much a temperature dependent parameter. Theoretical predications of temperature-dependent K have only accounted for temperature effects on ν , e.g., by defining a scaling factor $\alpha_K(T)$, or

$$K(T) = \alpha_K(T) \cdot K(T_r), \quad (7)$$

where $\alpha_K(T) = \frac{\nu_r}{\nu}$ and is increasing as $T > T_r$. This arises because the temperature sensitivity for μ is much larger than for ρ for liquid fluids. The relative magnitudes of $\alpha_K(T)$ versus $\alpha_h(T)$ using the STVF theory are presented in Fig. 1, and clearly show that predicted soil water retention effects of temperature are relatively small compared with the hydraulic conductivity effects, the latter being several times larger.

Most experimental $K(\theta, T)$ measurements typically find that observed effects are two to five times larger than those predicted by the STVF theory (Hopmans and Dane, 1986a; Constantz, 1982; Joshi et al., 2019). In part, this could be explained because of changing soil structure and pore characteristics by temperature, affecting the intrinsic permeability k in Eq. (6), as suggested by Gao and Shao (2015) and Kelishadi et al. (2018). Other possible reasons for the increased temperature effect could be explained by (a) changes in entrapped air volume, (b) a temperature coefficient of viscosity that is affected by pore water solutes, and (3) by vapor transport and enhanced liquid water flow, driven by imposed temperature gradients at relatively small soil water content values.

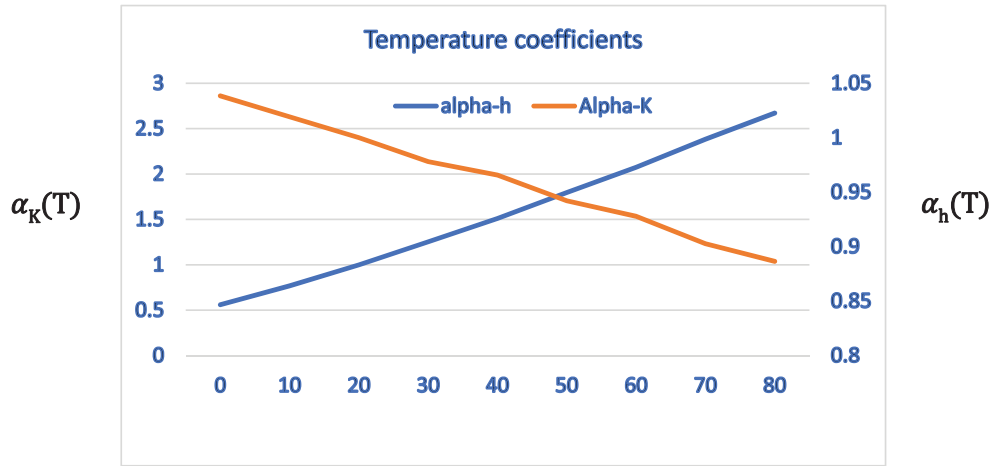


Fig. 1 Effect of temperature on soil water properties, as defined by temperature scaling factors $\alpha_h(T)$ and $\alpha_K(T)$, using the SFVF theory, with $T_r = 20^\circ\text{C}$ (Eqs. (3b) and (7), respectively).

Alternatively, when considering the temperature effects on $K(h)$, one finds these to be limited. This is so because the temperature effects on surface tension and viscosity are counteracting (Hopmans and Dane, 1986a) and offsetting each other.

Functional models for unsaturated hydraulic conductivity are based on pore size distribution, pore geometry and connectivity, and require integration of the soil water retention functions to obtain analytical expressions for the unsaturated hydraulic conductivity (Mualem, 1976). The resulting expressions relate the relative hydraulic conductivity, K_r , defined as the ratio of the unsaturated hydraulic conductivity, K , and the saturated hydraulic conductivity, K_s , to the effective saturation, and can be written in the following generalized form

$$K_r(S_e) = S_e^l \left[\frac{\int_0^{S_e} |h_m|^{-\eta} dS_e}{\int_0^1 |h_m|^{-\eta} dS_e} \right]^\beta \quad (8)$$

where l and η are parameters related to the tortuosity and connectivity of the soil pores, and the value of the parameter β is determined by the method of evaluating the effective pore radii. The moisture-dependency is highly nonlinear, with a decrease in K of 4–5 or more orders of magnitude within field-representative changes in water content. Model fitting techniques assume a prescribed form for the soil water retention curve, to be substituted in Eq. (8).

The ratio of the unsaturated K with the water capacity (C) is defined as the soil water diffusivity, $D(\theta)$, with dimension of m^2/s , or $D(\theta) = K(\theta)dh_m/d\theta = K(\theta)/C(\theta)$. The concept of diffusivity arises from a combination of Darcy law with the mass balance equation, when expressed in water content rather than soil water matric potential. Using Eqs. (3a) and (7), one can derive that the temperature effect on $D(\theta)$ can be computed from:

$$D_T(\theta) = K_T(\theta)/C_T(\theta, h_m) = \alpha_h(T_r) \alpha_K(T_r) * D(\theta, T_r), \quad (9)$$

indicating that the temperature effects are partly offset. However, since the temperature effect on $\alpha_K(T)$ is mostly dominating (Fig. 1), soil water diffusivity typically increases with temperature and may be dominated by temperature-induced changes in soil pore characteristics (Gao and Shao, 2015), thereby changing the intrinsic soil permeability in Eq. (6).

Unsaturated soil water flow modeling

Since the Darcy equation is strictly defined for steady state water flow conditions, the mass conservation principle is applied and combined with Eq. (5) to yield the so-called Richardson-Richards equation to solve for temporal changes in h_m or θ , at any depth z and time t :

$$\frac{\partial \theta}{\partial t} = C(h_m) \frac{\partial h_m}{\partial t} = \frac{\partial}{\partial z} \left[K(h_m) \left(\frac{\partial h_m}{\partial z} + 1 \right) \right], \quad (10)$$

and is valid for isotropic and isothermal conditions. With dynamic soil temperatures, the soil hydraulic parameters must be adjusted for temperature through the temperature corrections above, such as demonstrated by Hopmans and Dane (1985). However, in addition Philip and de Vries (1957) showed that additional water flow occurs through (a) liquid water flow by

temperature gradients and (b) vapor water transport occurring by both matric potential and temperature gradients. This requires that both the water and heat flow must be considered through their coupled transport equations, with transport coefficients defined by a total of four diffusivity coefficients, accounting for both liquid and vapor transport by matric potential gradients as well as by soil temperature gradients (see Heat and Water Transfer article to learn more about vapor transport). The complete expressions to define the coupled water and heat transport are extensive and include the need to quantify temperature effects on soil thermal properties (e.g., Hopmans and Dane, 1986c). For that reason, we refer to both Philip and de Vries (1957) and more recent work by Nassar and Horton (1992).

Long-term implications

Implications of diurnal soil temperature fluctuations and long-term temperature changes are extensive. As an example of diurnal variations, we demonstrate the effect of diurnal soil temperature variations on soil water infiltration with time in Fig. 2a, as presented by Jaynes (1990). The synchronized response is likely a direct consequence of changes in surface tension and viscosity with temperature (STVF theory). The modeling study by Jungqvist et al. (2014) showed that air temperature could increase more rapidly than soil temperature across the climatic gradient of Sweden (Fig. 2b), and indicated possible long-term impacts could be a shift in biomes from middle-boreal to southern-boreal in the northern most region of the country.

Assessment of temperature effects on soil properties has been done mostly in laboratory settings, allowing for testing of the STVF theory. Although such experiments allow for correction of unsaturated flow processes such as infiltration, drainage and consequences on soil water storage and plant available water on the short term (diurnal and seasonal), increasingly more relevant implications could be ascertained resulting from rising air and soil temperatures over the long-term by climate change. This includes soil water evaporation, soil water storage, plant-available water, and groundwater recharge, while other implications such as soil carbon stock, GHG emissions and soil thawing are intricately related to soil hydraulic properties. Therefore, the remaining discussion will focus only on consequences of increasing soil temperatures.

Detailed analysis of increased temperature on soil water retention shows that soil water storage and plant-available is reduced. This can be readily shown from the STVF theory, because of lowered surface tension and viscosity of water that decreases the soil's field capacity. Though relatively minor, together with the increased soil water evaporation, it is projected that most climatic regions will experience increasing risks of plant water stress and drought. Soil water evaporation will be enhanced by increasing air temperatures and vapor pressure deficits, coupled with changing soil water and soil vapor diffusivity values. In concert, as suggested by Green et al. (2019), the soil carbon stock is likely to decrease as climate warming continues, and much of the affected land surface may transition from a carbon sink to a carbon source in the future, thereby accelerating climate change.

Further implications of global warming come from the thawing of permafrost. Permafrost soils stay frozen for two or more years, often below a shallow layer of surface soil that freezes and thaws every year. Often the permafrost contains large amounts of organic matter that decomposes when thawed thereby emitting CO₂ in the atmosphere. Many permafrost soils are found in boreal forests that are potentially very productive and contain many wetlands because of the seasonal freezing and thawing. Understanding the implications of warming of these permafrost areas requires knowledge of their hydrologic behavior, largely controlled by soil hydraulic properties during and between freezing/thawing cycles. There is limited information available as their measurement in such conditions is extremely difficult, as complications arise because of soil structural changes by expansion of soils because of soil freezing, largely impacting soil water retention and hydraulic conductivity.

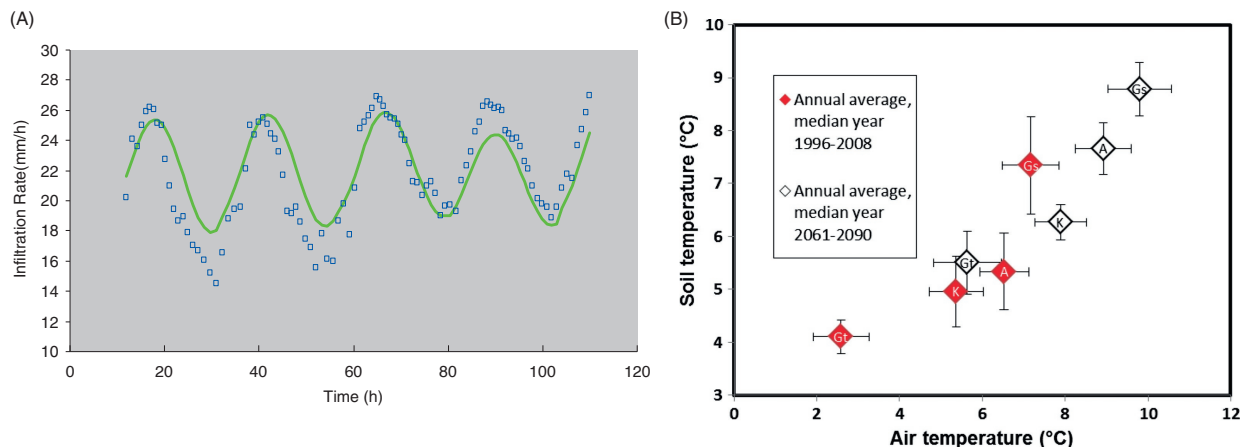


Fig. 2 (a) Effect of diurnal changes of soil temperature on soil infiltration (symbols are measured values and the solid curve is derived from temperature influence on water viscosity, from Jaynes, 1990), and (b) projected change in long-term annual air and soil temperature (20 cm depth) for various biomes across Sweden (Fig. 7 of Jungqvist et al.).

Summary

In brief, there remains much uncertainty on the temperature dependence of soil hydraulic properties. Much of the uncertainty arises because reported research has been through laboratory experimentation on small soil cores using different methodologies for a limited range of soil types. Though effects of soil water properties can be measured accurately to verify the STVF theory, significant deviations are often reported, especially to quantify temperature dependence of the soil unsaturated hydraulic conductivity. In part, this is explained by soil temperature changing soil particle-soil solution interactions because of soil solution chemistry as well as by changing soil pore connectivity and water retention as entrapped air and water volumes contract and expand (or change).

Increasingly, both water scarcity and climatic changes are making large fractions of continental populations vulnerable. Both are global issues that require attention to scientific agendas and policymakers, and for which mitigating soil and water management practices at local scales are crucial, to offset impacts of increasing temperatures globally. Therefore, future research should focus on the monitoring and analysis of in situ field-based experiments, especially with increasing scientific interest on the effects of global warming on soil water availability, drought and water security, and the need to include temperature-dependent soil physical characteristics in land surface simulation models that couple water and energy transfer between vegetated soils and the atmosphere above.

References

- Bachmann J, Horton R, Grant SA, and van der Ploeg RR (2001) Isothermal and nonisothermal evaporation from four sandy soils of different water repellency. *Soil Science Society of America Journal* 65: 1599–1607.
- Constandant J (1982) Temperature dependence of unsaturated hydraulic conductivity of two soils. *Soil Science Society of America Journal* 46: 466–470.
- Gao H and Shao M (2015) Effects of temperature changes on soil hydraulic properties. *Soil and Tillage Research* 153: 145–154.
- Green JK, et al. (2019) Large influence of soil moisture on long-term terrestrial carbon uptake. *Nature* 565: 476–479. <https://www.nature.com/articles/s41586-018-0848-x>.
- Hopmans JW and Dane JH (1985) Effect of temperature-dependent hydraulic properties on soil water movement. *Soil Science Society of America Journal* 49: 51–59.
- Hopmans JW and Dane JH (1986a) Temperature dependence of soil hydraulic properties. *Soil Science Society of America Journal* 50: 4–9.
- Hopmans JW and Dane JH (1986b) Temperature dependence of soil water retention curves. *Soil Science Society of America Journal* 50: 562–567.
- Hopmans JW and Dane JH (1986c) Thermal conductivity of two porous media as a function of water content, temperature, and density. *Soil Science* 142: 187–195.
- Jaynes DB (1990) Temperature variations effect on field-measured infiltration. *Soil Science Society of America Journal* 54: 305–312.
- Joshi DC, Iden S, Peters A, Das BS, and Durner W (2019) Temperature dependence of soil hydraulic properties: Transient measurements and modeling. *Soil Science Society of America Journal* 83: 1628–1636.
- Jungqvist G, Oni SK, Teutschbein C, and Futter MN (2014) Effects of climate change on soil temperature in Swedish Boreal forests. *PLoS One* 9(4): 1–10.
- Kelishadi H, Mosaddeghi MR, Ayoubi S, and Mamedov AI (2018) Effect of temperature on soil structural stability as characterized by high energy moisture characteristic method. *Catena* 170: 290–304.
- Liu HH and Dane JH (1993) Reconciliation between measured and theoretical temperature effects on soil water retention curves. *Soil Science Society of America Journal* 57: 1202–1207.
- Mualem Y (1976) New model for predicting hydraulic conductivity of unsaturated porous-media. *Water Resources Research* 12: 513–522.
- Nassar IN and Horton R (1992) Simultaneous transfer of heat, water, and solute in porous media. I. Theoretical development. *Soil Science Society of America Journal* 56: 1350–1356.
- Nimmo JR and Miller EE (1986) The temperature dependence of isothermal moisture vs potential characteristics of soils. *Soil Science Society of America Journal* 50: 1105–1113.
- Philip JR and de Vries DA (1957) Moisture movement in porous materials under temperature gradients. *Transactions of the American Geophysical Union* 38(2): 222–232.

Heat and vapor transfer

Sunny Eng Giap Goh^a, Joshua L Heitman^b, and Robert Horton^c, ^aUniversiti Malaysia Terengganu, Terengganu, Malaysia; ^bNorth Carolina State University, Raleigh, NC, United States; ^cIowa State University, Ames, IA, United States

© 2023 Elsevier Ltd. All rights reserved.

Introduction	386
Mechanisms for transport of liquid and water vapor in soil	386
The governing equation for liquid and water vapor	387
Mechanisms driving heat transfer	388
The governing equation for heat transfer	388
Pore-scale model	389
Applications of heat and water movement	390
Heat and water transfer continuum model limitations	390
Additional applications of coupled heat and mass transfer models	391
Summary	391
References	391

Abstract

Heat and water transfer in soils are governed by equations that correspond to heat and mass conservation. These soil physics equations deal with fluxes and storage of heat and mass (liquid water, water vapor) within soil. The theory has many useful applications, for example, soil temperature variations that affect crop production, water irrigation management, water vapor release to the atmosphere, ground heat exchange systems in commercial and housing areas, land surface modeling of surface temperature, fate and transfer of volatile organic compounds, and modeling of greenhouse gas emissions. The macroscopic scale mechanisms driving heat and mass fluxes described in this Chapter indicate the current knowledge used to interpret observations, and also expose the limitations of existing models. A pore-scale model is also briefly described which represents the fundamental mechanisms (microscopic scale) more precisely than the currently used representative elementary volume continuum models (macroscopic scale).

Key points

- Describe the mechanisms driving liquid, water vapor and heat flux in the soil
- Present the heat and mass transfer governing equations that can describe soil moisture content and temperature distributions in the soil
- Describe applications of the equations and possible future improvements

Nomenclature

q_L	Liquid water flux ($kg \cdot m^{-2} \cdot s^{-1}$)
ψ_m	Soil water matric potential (m)
K	Hydraulic conductivity ($m \cdot s^{-1}$)
z	Vertical axis (m) and it is oriented positively upward
ρ_L	Liquid water density ($kg \cdot m^{-3}$)
i	A unit vector with value that could be either +1 or – 1 for vertically upward or downward, respectively.
D_{TL}	Thermal liquid diffusivity ($K \frac{\partial \psi_m}{\partial T}$, $m^2 \cdot s^{-1} \cdot K^{-1}$)
T	Soil temperature (K)
q_v	Water vapor flux ($kg \cdot m^{-2} \cdot s^{-1}$)
D	Water vapor diffusion ($m^2 \cdot s^{-1}$)
Ω	Tortuosity factor (dimensionless)
θ_a	Volumetric air content ($m^3 \cdot m^{-3}$)
ρ_v	Water vapor density ($\rho_s h_r = \rho_s h_m$, $kg \cdot m^{-3}$). The ρ_s is saturated water vapor density ($kg \cdot m^{-3}$), h_r is relative humidity (dimensionless), and h_m is matric potential effect on relative humidity (dimensionless)
D_{mv}	Isothermal vapor diffusivity ($\frac{D\Omega\theta_a\rho_s}{\rho_L} \frac{\partial h_m}{\partial \psi_m}$, $m \cdot s^{-1}$)
D_{Tv}	Thermal vapor diffusivity ($\eta \frac{D\Omega\theta_a}{\rho_L} \left(h_m \frac{\partial \rho_s}{\partial T} + \rho_s \frac{\partial h_m}{\partial T} \right)$, $m^2 \cdot s^{-1} \cdot K^{-1}$), where η is vapor enhancement factor (dimensionless)
θ	Total water content ($m^3 \cdot m^{-3}$)
t	Time in second

θ_L	Volumetric water content ($m^3 \cdot m^{-3}$)
q_h	Energy heat flux ($J \cdot m^{-2} \cdot s^{-1}$)
λ_*	Thermal conductivity of wet porous medium ($W \cdot m^{-1} \cdot K^{-1}$)
L_o	Latent heat of vaporization ($J \cdot kg^{-1}$) at T_o
c_v	Specific heat of water vapor ($J \cdot kg^{-1} \cdot K^{-1}$)
T_o	An arbitrary reference temperature (K)
c_L	Specific heat of liquid water ($J \cdot kg^{-1} \cdot K^{-1}$)
S_h	Total heat content ($J \cdot m^{-3}$)
C_d	Volumetric heat capacity of dry soil solid matrix including the organic matter ($\rho_b(c_o\phi_o + c_m\phi_m)$, $J \cdot m^{-3} \cdot K^{-1}$). The ρ_b is soil bulk density ($kg \cdot m^{-3}$), c_o is specific heat of organic matter ($J \cdot kg^{-1} \cdot K^{-1}$), ϕ_o is organic mass fraction (dimensionless), c_m is specific heat of soil minerals ($J \cdot kg^{-1} \cdot K^{-1}$), and ϕ_m is mineral mass fraction (dimensionless)
W	Differential heat of wetting ($J \cdot kg^{-1}$)
$L(T)$	Latent heat of vaporization ($J \cdot kg^{-1}$) at T
$Q_b(h)$	Bond invasion potential ($N \cdot m^{-2}$)
γ	Surface tension ($N \cdot m^{-1}$)
P_g	Gas pressure ($N \cdot m^{-2}$)
$P_l(h)$	Liquid pressure ($N \cdot m^{-2}$) at elevation h oriented vertically positively downward
$\Delta\rho$	Density difference between two fluids ($kg \cdot m^{-3}$)
r_b	Throat radius (m)
g	Gravity acceleration ($m \cdot s^{-2}$)
L	Vertical height of model medium (m)
B	Bond number (dimensionless)

Introduction

Soil water is vital for sustaining life in terrestrial environments. Soil water may come from sources such as precipitation and upward movement of groundwater, or by human interventions with inputs such as irrigation. A consistent soil water supply to support plant root adsorption and transpiration is necessary to sustain crop productivity. In both cropped and natural systems, soil water is transferred back to the atmosphere through a combination of evaporation and transpiration. These processes tightly connect the terrestrial hydrologic cycle (mass transfer) to the land-surface energy balance (heat transfer). Soil water is in a cycle of continuous exchange with its surroundings according to gradients in water potential and temperature. Through these exchanges, soil water storage can vary between extremely dry and saturation. Since soil water dynamics are governed by both mass and heat mechanisms, accounting for these mechanisms makes possible forecasting and scenario testing, and enhances decision-making strategies for allocating water resources.

Because soil water is a key variable that exerts control on a range of environmental processes, understanding mechanisms that drive its redistribution is broadly important to a range of problems in agriculture, engineering, and environmental sciences. Some problems can be addressed without considering the full complexity of these mechanisms. Mass transfer of soil water modeled by the Richardson-Richards equation often excludes vapor transfer. As soil temperature gradients become pronounced, water vapor mechanisms are included in the mass transfer model. Heat transfer within soil is often simplified to conduction and modeled by Fourier's law. When phase transition and fluid mass transfer cannot be neglected, latent heat of vaporization, and sensible heat transfer mechanisms of vapor and liquid water are included in the heat transfer equation. An effort at simplification may include decoupling and solving the heat and mass transfer equations separately. Alternatively, continued growth in our understanding of the complexity in water and heat transfer mechanisms can provide useful insights to address new and emerging challenges, such as global climate change and uncertainty. Our capacity to address new problems is supported by our ability to extrapolate relationships from basic mechanisms. In the following sections, we present macroscopic, continuum-based theory for mechanisms driving water movement and heat transfer, we briefly introduce a pore-scale model, and we discuss applications and limitations to existing theory.

Mechanisms for transport of liquid and water vapor in soil

Water flux in unsaturated soil is described by the Buckingham-Darcy equation,

$$q_L = -\rho_L K \frac{d\Psi_m}{dz} - \rho_L Ki \quad (1)$$

The left term is taken as the mass of water moving through a unit cross-sectional area in a unit of time. Eq. (1) assumes that water density does not change with space (z), and its Cartesian coordinate is aligned vertically upward positive. The first term on the right

is the water flux rate due to the matric potential gradient, which represents water flow from higher to lower potential regions. The negative sign on the term is to cancel the negative value in the denominator (dz). The second term on the right represents the gravity water flux, and the negative sign indicates gravity water flux is oriented upward.

Under the influence of temperature, Eq. (1) can be extended to account for a temperature gradient-driven flux (Milly, 1982),

$$q_L = -\rho_L K \frac{d\Psi_m}{dz} - \rho_L D_{TL} \frac{dT}{dz} - \rho_L Ki \quad (2)$$

The second term on the right is the water flux due to a temperature gradient, which is water transfer from higher to lower temperature regions. The term can be neglected when there is a homogenous temperature, i.e., spatially uniform temperature. Hydraulic conductivity is also influenced by soil temperature (Hopmans and Horton, 2022).

In addition, there is a water vapor flux in unsaturated soil. The vapor flux is often described by Fick's law,

$$q_v = -D\Omega\theta_a \frac{d\rho_v}{dz} \quad (3)$$

The equation is taken as the mass of vapor transfer in a unit cross-sectional area in a unit time. The equation describes water vapor movement from high to low vapor density regions. The water vapor transport occurs only in the pore space corresponding to the volumetric air content. The travel distance of the water vapor in the soil increases as it moves around and over the soil particles, liquid films, and pore water as given by a tortuosity factor. These two parameters reduce the water vapor flux in soil compared to that in open air space.

The fundamental independent variables driving water vapor flux in the soil are temperature and matric potential. Hence, Eq. (3) can be expanded as below,

$$q_v = -\rho_L D_{mv} \frac{d\Psi_m}{dz} - \rho_L D_{Tv} \frac{dT}{dz} \quad (4)$$

The first term on the right side of the equation represents the matric potential gradient driven water vapor flux from higher to lower pressure regions, and the second term represents the temperature gradient driven water vapor flux from higher to lower temperature regions.

The combined liquid water and water vapor fluxes give a total mass flux in the soil, as

$$\frac{q_L + q_v}{\rho_L} = -(K + D_{mv}) \frac{d\Psi_m}{dz} - (D_{TL} + D_{Tv}) \frac{dT}{dz} - Ki \quad (5)$$

The governing equation for liquid and water vapor

The mass conservation of liquid water and water vapor in the soil in the vertical spatial dimension is given by,

$$\frac{\partial(\rho_L \theta)}{\partial t} = -\frac{\partial(q_L + q_v)}{\partial z} \quad (6)$$

The variable $\rho_L \theta$ in the storage term is given by the mass of liquid water and water vapor in a unit volume of the soil. By placing Eq. (5) and $\rho_L \theta = \rho_L \theta_L + \rho_v \theta_a$ into Eq. (6) and assuming water density does not change with time, Eq. (6) can be rewritten as,

$$\frac{\partial\left(\theta_L + \frac{\rho_v}{\rho_L} \theta_a\right)}{\partial t} = -\frac{1}{\rho_L} \frac{\partial\left(-\rho_L(K + D_{mv}) \frac{\partial\Psi_m}{\partial z} - \rho_L(D_{TL} + D_{Tv}) \frac{\partial T}{\partial z} - \rho_L Ki\right)}{\partial z} \quad (7)$$

Also, assuming that liquid water density does not change with space, the equation becomes,

$$\frac{\partial\theta_L}{\partial t} + \frac{1}{\rho_L} \frac{\partial}{\partial t}(\rho_v \theta_a) = \frac{\partial}{\partial z} \left[(K + D_{mv}) \frac{\partial\Psi_m}{\partial z} + (D_{TL} + D_{Tv}) \frac{\partial T}{\partial z} + Ki \right] \quad (8)$$

The water vapor density is a function of matric potential and temperature, and the volumetric air content is a function of total porosity and volumetric water content. The matric potential is related to the soil moisture content by a water retention function. The model further assumes that the total porosity does not change with time. Using the chain rule method, the second term on the left side of Eq. (8) can be expanded into (Heitman et al., 2008a),

$$\left(1 + \frac{\theta_a}{\rho_L} \frac{\partial\rho_v}{\partial\Psi_m} \frac{\partial\Psi_m}{\partial\theta_L} - \frac{\rho_v}{\rho_L}\right) \frac{\partial\theta_L}{\partial t} + \frac{\theta_a}{\rho_L} \frac{\partial\rho_v}{\partial T} \frac{\partial T}{\partial t} = \frac{\partial}{\partial z} \left[(K + D_{mv}) \frac{\partial\Psi_m}{\partial z} + (D_{TL} + D_{Tv}) \frac{\partial T}{\partial z} + Ki \right] \quad (9)$$

The right side of Eq. (9) indicates the mass flux resulting from the matric potential gradients, the temperature gradients, and gravity. Both the matric potential and temperature gradients are imposed on the mass transport of liquid water and water vapor, whereas gravity is only applied to liquid water. The mass fluxes result in the time-varied soil moisture content and soil temperature, as indicated in the left side of Eq. (9).

Mechanisms driving heat transfer

Heat flux in unsaturated soil is given by (de Vries, 1958),

$$q_h = -\lambda_* \frac{dT}{dz} + L_o q_v + c_v(T - T_o)q_v + c_L(T - T_o)q_L \quad (10)$$

The left side of the equation is taken as the heat energy that moves across a unit cross-sectional area in a unit of time. The first term on the right is heat conduction in soil. The second term is the heat transfer by water vapor movement from the latent heat of vaporization. The last two terms are heat transfers by sensible heat movement due to water vapor and liquid water fluxes, respectively. Eq. (10) does not include the sensible heat movement by water transfer in plant roots.

The governing equation for heat transfer

Heat energy conservation is given by,

$$\frac{\partial S_h}{\partial t} = -\frac{\partial q_h}{\partial z} \quad (11)$$

The equation indicates the flux of heat energy in the vertical spatial dimension on the equation right side, which results in the change in the heat storage with time on the equation left side. If the incoming heat flux is greater than the outgoing heat flux, the heat storage will increase, and vice versa.

Heat storage in the soil is given by (de Vries, 1958),

$$S_h = C_d(T - T_o) + c_v \rho_v \theta_a(T - T_o) + c_L \rho_L \theta_L(T - T_o) + L_o \rho_v \theta_a - \rho_L \int_{t_o}^t W \frac{\partial \theta_L}{\partial t} dt \quad (12)$$

The first term on the right is the heat storage in the soil solid phase, which is volumetric heat capacity of the soil solids. The second and third terms are heat storage in the water vapor and liquid water, respectively. The latent heat of vaporization is in the fourth term. The fifth term includes heat involved with differential wetting. The negative sign on the differential wetting implies that an increase of soil moisture content would reduce heat storage.

Inserting Eq. (12) into the left side term of Eq. (11) gives (heat storage varied with time),

$$\begin{aligned} \frac{\partial S_h}{\partial t} = & \left\{ (T - T_o)(c_L \rho_L - c_v \rho_v) - L_o \rho_v - \rho_L W + \theta_a[(T - T_o)c_v + L_o] \frac{\partial \rho_v}{\partial \Psi_m} \frac{\partial \Psi_m}{\partial \theta_L} \right\} \frac{\partial \theta_L}{\partial t} \\ & + \left\{ C_d + c_v \rho_v \theta_a + c_L \rho_L \theta_L + \theta_a[(T - T_o)c_v + L_o] \frac{\partial \rho_v}{\partial T} \right\} \frac{\partial T}{\partial t} \end{aligned} \quad (13)$$

The spatial heat flux given in the right side of Eq. (11) combined with heat flux over a cross-sectional area as in Eq. (10) is combined to give,

$$-\frac{\partial q_h}{\partial z} = \frac{\partial}{\partial z} \left[(\lambda_* + L_o \rho_L D_{Tv}) \frac{\partial T}{\partial z} + L_o \rho_L D_{mv} \frac{\partial \Psi_m}{\partial z} - (T - T_o)(c_v q_v + c_L q_L) \right] \quad (14)$$

Taking the latent heat of vaporization at reference soil temperature as $L(T) + (c_L - c_v)(T - T_o)$, the terms become,

$$-\frac{\partial q_h}{\partial z} = \frac{\partial}{\partial z} \left[(\lambda_* + L_o \rho_L D_{Tv}) \frac{\partial T}{\partial z} + \rho_L [L(T) + (c_L - c_v)(T - T_o)] D_{mv} \frac{\partial \Psi_m}{\partial z} - (T - T_o)(c_v q_v + c_L q_L) \right] \quad (15)$$

Placing both Eqs. (13) and (15) into Eq. (11) gives (Heitman et al., 2008a),

$$\begin{aligned} & \left\{ (T - T_o)(c_L \rho_L - c_v \rho_v) - L_o \rho_v - \rho_L W + \theta_a[(T - T_o)c_v + L_o] \frac{\partial \rho_v}{\partial \Psi_m} \frac{\partial \Psi_m}{\partial \theta_L} \right\} \frac{\partial \theta_L}{\partial t} \\ & + \left\{ C_d + c_v \rho_v \theta_a + c_L \rho_L \theta_L + \theta_a[(T - T_o)c_v + L_o] \frac{\partial \rho_v}{\partial T} \right\} \frac{\partial T}{\partial t} \\ & = \frac{\partial}{\partial z} \left[(\lambda_* + L_o \rho_L D_{Tv}) \frac{\partial T}{\partial z} + \rho_L [L(T) + (c_L - c_v)(T - T_o)] D_{mv} \frac{\partial \Psi_m}{\partial z} - (T - T_o)(c_v q_v + c_L q_L) \right] \end{aligned} \quad (16)$$

The right side of Eq. (16) gives the heat fluxes by conduction, latent heat of vaporization, water vapor sensible heat, and liquid water sensible heat. Heat fluxes by soil temperature gradient and soil matric potential gradient result in time-varied moisture content and temperature in the soil as on the left side of the equation. When modeling soil moisture content and temperature distributions, Eqs. (9) and (16) are solved simultaneously.

Fig. 1 shows the apparent thermal conductivity (λ_{app}) values for selected soils at various moisture content values. The λ_{app} value represents the thermal conductivity as a summation of all transfer mechanisms included in Eq. (10) (Goh and Noborio, 2016). In conditions where liquid and vapor fluxes become insignificant, the equation reduces to the pure thermal conductivity, that is the first term on the right side of Eq. (10), excluding the temperature gradient.

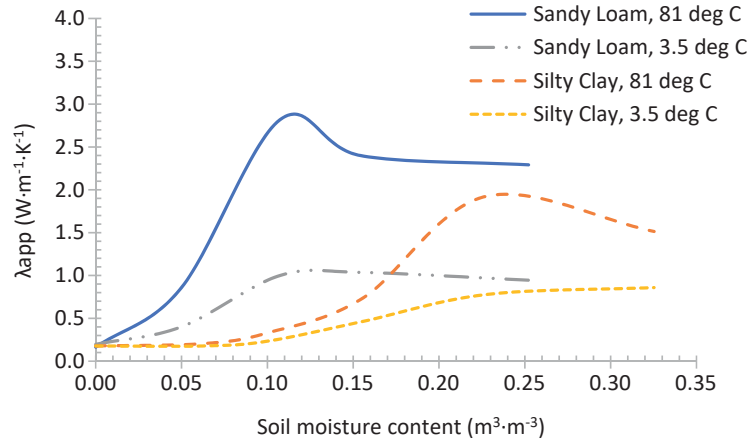


Fig. 1 Apparent soil thermal conductivity with soil moisture content for sandy loam and silty clay soils at 3.5 and 81 °C. Note the effects of both temperature and texture over the range of soil moisture contents. Source: Lu S, Ren T, and Horton R (2020) Estimating the components of apparent thermal conductivity of soils at various water contents and temperatures. *Geoderma*, 376, 114530.

Pore-scale model

A pore-scale model represents phenomena at the pore level, which can be considered microscopic in relative comparison to the continuum approach phenomena at the macroscopic level (i.e., representative elementary volume), for example, Darcy's law. For this reason, a pore-scale model can be viewed as a more fundamental approach than the macroscopic level model. Pore-scale models consist of the volume of fluids method, lattice Boltzmann, stochastic rotation dynamics, pore-network model, level set method, and phase-field. The pore-network model can be used to model water distribution in porous media. The model is represented by random sizes of network pores and throats that correspond to the empty spaces between the particles (e.g., sand particles) and the linkages for mass transfer between the empty spaces. Throats have different widths or radii for cylinder throats.

The Prat (1993) model uses step rules (or algorithms) to guide the mechanisms of standard invasion percolation into an initially saturated two-dimensional porous medium (permeable at the top boundary) by neglecting evaporation. These steps are: (1) at each step, non-wetting fluid invades the widest throat (and adjacent pore) that has the lowest capillary pressure (or suction), (2) regions of isolated wetting fluid, which are isolated clusters of pores and throats filled with wetting fluid, are trapped and cannot be invaded due to wetting fluid incompressibility [where the isolated clusters are parts of the central single wetting fluid cluster before it was separated by invading non-wetting fluid], and (3) the invasion end when the non-wetting fluid reaches a pore at the exit surface at the bottom boundary.

Prat (1993) then introduced a drying algorithm to replace the invasion percolation algorithm. Steps are: (1) identify all clusters in the network, (2) identify the throat with the lowest capillary pressure for each cluster, (3) calculate the evaporation for each boundary cluster, (4) the calculated evaporation flux in step 3 is assigned to the throat in step 2, for each cluster, (5) the throat with the adjacent pore in step 2 is drained, and (6) update the fluid distribution in the network. Including the effect of gravity, the jump potential is estimated for each throat,

$$Q_b(h) = -\frac{2\gamma}{r_b} + (P_g - P_l(h)) = -\frac{2\gamma}{r_b} \left(1 - \frac{\Delta\rho g L r_b}{2\gamma} \frac{(L-h)}{L} \right) = -\frac{2\gamma}{r_b} \left(1 - B \frac{(L-h)}{L} \right) \quad (17)$$

The throat invasion potential is given by the difference between the surface tension resulted pressure ($-2\gamma/r_b$) and gas-liquid hydrostatic pressure ($P_g - P_l(h)$). Equation (17) is used to estimate and identify the throat with the lowest capillary pressure in step 2. Including the effect of viscosity, Darcy's law equation, the Hagen-Poiseuille capillary equation, and the mass balance equation are used (Prat, 2002). Similarly, the equations for water vapor diffusion between pores through the throat, evaporation rate above water meniscus at the adjacent throat, and mass balance in the vapor phase are also reported in Prat (2002). Refer to Nowicki et al. (1992) for further information on these equations. The effects of a temperature gradient on the surface tension gradient and the saturation vapor partial pressure gradient are included. In addition, a more detailed pore-scale approach has been used to study vapor transport in porous medium, for instance, the vapor enhancement mechanism caused by the connected capillary pathways (or the liquid island effect) (Shahraeeni and Or, 2012).

In addition to the investigation of fundamental mechanisms at the microscopic level, parameters at the macroscopic level can be estimated, for example, liquid relative permeability, capillary pressure, fluid effective diffusion, the pressure of capillary (Nowicki et al., 1992), and the effective heat transfer coefficient.

Applications of heat and water movement

The soil moisture and heat transfer equations developed by Philip and de Vries (1957) and de Vries (1958) have been used in several soil studies. The heat and moisture model has been extended to include other mechanisms. For instance, Nassar and Horton (1997) extended the equation to include solute transport mechanisms. Liang et al. (2017) used the soil heat and moisture movement model coupled with crop growth to model the soil moisture content, temperature, leaf area index, dry matter, and crop yield in rice fields. Bai et al. (2020) improved the soil water and heat equations to include soil deformation. Although the advantage of having a fully coupled heat and moisture simulation model is evident, some researchers have also provided meaningful estimates with a simplified decoupled version of the equations. For instance, Ni et al. (2019) decoupled the soil heat and moisture mechanisms into a simplified model for estimating the mass transport using the Richards equation and for estimation of soil temperature using an empirical equation. Heitman et al. (2008b) estimated soil water evaporation with depth and time.

The soil heat and moisture equation have been used to model soil temperature and water content distributions. Modeling of soil heat and moisture has also been possible under slightly more extreme conditions, as noted in the study of Kelleners et al. (2016) on water infiltration of snowmelt. Massman (2015) likewise used soil heat and moisture equations to study high-temperature sand-water-steam systems, for wildfires and prescribed burns.

Moreover, soil heat and moisture modeling has been reported in specific engineering applications. Chalhoub et al. (2017) studied ground heat exchangers using simplified heat and moisture equation corresponding to Fourier's law thermal conductivity driven heat flux and simple water budget equation based on soil water content, rainfall rate, and evaporation rate.

Soil heat and moisture equations have also been applied in studies of specific mechanisms. For example, Vanderborght et al. (2017) summarized the application of heat and water transport to justify the need for a 3-dimensional representation of processes to study vapor transport in soil.

Heat and water transfer continuum model limitations

The heat and water transfer equations have some limitations (Heitman et al., 2008a). The limitations are based on considered, not considered, and yet discovered, understood, or quantified phenomena. First of all, the governing equations described in this review are limited to mass and heat transfer in a vertical direction. When mass and heat transfer take place in two- or three-dimensional spaces, the one-dimensional governing equation should be expanded to account for those dimensions. The mass transport is driven by soil matric potential and temperature gradients and gravitational force, while heat transfer is based on conduction, latent heat convection, and sensible heat convection. The governing equations can be extended to account for solute transport and osmotic effects (Nassar and Horton, 1997). For structured soil, the pore domain could be fractioned into dual-permeability, that is, separate permeability for the soil matrix and the macropores (fractures).

The equation assumes rigid porous media. Under non-rigid media conditions, the effect of bulk density can be included to describe water and temperature distributions (Kojima et al., 2018). Also, hysteresis in water retention curves may or may not be accounted for in the model, depending on the constitutive equation used. Milly (1982) included hysteresis. In addition, local thermodynamic equilibrium and equilibrium liquid-vapor phase usually are assumed. A non-equilibrium liquid-vapor change model has been proposed, but it has limited precise predictions of soil moisture distributions (Massman, 2015).

Furthermore, uncertainty in parameter calibration on unsaturated hydraulic conductivity and vapor flux has been a common concern (Heitman et al., 2008a). Additional uncertainty is encountered when the model soil system is exposed to open air with atmospheric pressure fluctuations due to wind gusts, possibly adding more uncertainty to model predictions (Scanlon and Milly, 1994). Water evaporation from the soil surface and the associated drying front depth (length), which is the result of the force balance between capillarity, gravity, and viscous dissipation (Lehmann et al., 2008), may also affect the application of heat and mass transfer theory.

Mechanisms leading to the excessive vapor fluxes sometimes observed under temperature gradients are also not fully understood (Philip and de Vries, 1957). Cass et al. (1984) quantified the vapor enhancement factor at various temperatures, soil moisture contents and air pressures. Hiraiwa and Kasubuchi (2000) simplified the experimental method by excluding pressure control and simplified the mathematical equation for estimation. Air volume expansion and contraction by diurnal temperature variation has also been investigated in the field (Parlange et al., 1998). A pore-scale model which incorporates the liquid island effect confirms the Philip and de Vries's vapor enhancement postulation at low soil moisture content values (Shahraeeni and Or, 2012). Other than air flux, a liquid to vapor phase change which results in vapor volume expansion advection (inferred from the universal gas law) has also been suggested (Goh and Noborio, 2016). Again, the Lu et al. (2020) measurements identified excessive vapor flux which reaffirms the need for including new mechanisms that could be considered as either air volume expansion advection, vapor volume expansion advection, or a combination of both. Another mechanism, vapor buoyancy, has not been investigated even within pore-scale models (Shahraeeni and Or, 2012).

Additional applications of coupled heat and mass transfer models

Coupled heat and mass transfer models have useful applications to several earth and environmental system topics, including fate and transport of volatile organic materials such as BTEX compounds, pesticides, and greenhouse gases (GHG). Dynamic heat and water transfer connects to soil temperature and water content distributions. Temperature and water content impact microbial activity, which influences chemical sinks and production rates, and they also influence gas diffusion rates. Therefore, coupled heat and water transfer is not only important to estimate dynamic soil temperature and water content as connected to surface heat balance and surface water balance, but the dynamic soil temperature and water content distributions are necessary to estimate organic chemical transport and greenhouse gas fluxes.

Summary

Work to-date reveals the mechanisms driving fluxes of heat and mass in an unsaturated porous medium. The heat and mass transfer equations can be solved simultaneously to describe the temperature and water distributions in soil. Model estimated temperature and moisture distributions in soil have many important applications in agriculture and engineering, and the diversity of applications continues to expand. The model can be used as the basis for expansions to include additional mechanisms such as solute transport, greenhouse gas emission, soil deformation, dual-porosity, and dual-permeability, or for investigations of unclear mechanisms like the vapor enhancement factor. The pore-scale model provides promising opportunities to study heat and mass transfer mechanisms in undisturbed soils, especially the effects of soil particle arrangement and the resulting distributions of pore and throat sizes. However, similar to the macroscopic scale theory, microscopic model mechanisms have to be calibrated and validated.

References

- Bai R, Lai Y, Zhang M, and Ren J (2020) Study on the coupled heat-water-vapor-mechanics process of unsaturated soils. *Journal of Hydrology* 585: 124784. <https://doi.org/10.1016/j.jhydrol.2020.124784>.
- Cass A, Campbell GS, and Jones TL (1984) Enhancement of thermal water vapor diffusion in soil. *Soil Science Society of America Journal* 48(1): 25–32. <https://doi.org/10.2136/sssaj1984.03615995004800010005x>.
- Chalhoub M, Bernier M, Coquet Y, and Philippe M (2017) A simple heat and moisture transfer model to predict ground temperature for shallow ground heat exchangers. *Renewable Energy* 103: 295–307. <https://doi.org/10.1016/j.renene.2016.11.027>.
- de Vries DA (1958) Simultaneous transfer of heat and moisture in porous media. *Eos, Transactions American Geophysical Union* 39(5): 909–916. <https://doi.org/10.1029/TR039i005p00909>.
- Goh EG and Noborio K (2016) An improved heat flux theory and mathematical equation to estimate water vapor advection as an alternative to mechanistic enhancement factor. *Transport in Porous Media* 111(2): 331–346. <https://doi.org/10.1007/s11242-015-0596-4>.
- Heitman JL, Horton R, Ren T, Nassar IN, and Davis DD (2008a) A test of coupled soil heat and water transfer prediction under transient boundary temperatures. *Soil Science Society of America Journal* 72(5): 1197–1207. <https://doi.org/10.2136/sssaj2007.0234>.
- Heitman JL, Xiao X, Horton R, and Sauer TJ (2008b) Sensible heat measurements indicating depth and magnitude of subsurface soil water evaporation. *Water Resources Research* 44: W00D05. <https://doi.org/10.1029/2008WR006961>.
- Hiraiwa Y and Kasubuchi T (2000) Temperature dependence of thermal conductivity of soil over a wide range of temperature (5–75°C). *European Journal of Soil Science* 51(2): 211–218. <https://doi.org/10.1046/j.1365-2389.2000.00301.x>.
- Hopmans JW and Horton R (2022) *Temperature effects of soil hydraulic properties*, 2nd edn. Encyclopedia of soils in the environment. ch. 00206.
- Kelleners TJ, Koonce J, Shillito R, Dijkema J, Berli M, Young MH, Frank JM, and Massman WJ (2016) Numerical modeling of coupled water flow and heat transport in soil and snow. *Soil Science Society of America Journal* 80(2): 247–263. <https://doi.org/10.2136/sssaj2015.07.0279>.
- Kojima Y, Heitman JL, Sakai M, Kato C, and Horton R (2018) Bulk density effects on soil hydrologic and thermal characteristics: A numerical investigation. *Hydrological Processes* 32: 2203–2216. <https://doi.org/10.1002/hyp.13152>.
- Lehmann P, Assouline S, and Or D (2008) Characteristic lengths affecting evaporative drying of porous media. *Physical Review E* 77(5): 056309. <https://doi.org/10.1103/PhysRevE.77.056309>.
- Liang H, Hu K, Qin W, Zuo Q, and Zhang Y (2017) Modelling the effect of mulching on soil heat transfer, water movement and crop growth for ground cover rice production system. *Field Crops Research* 201: 97–107. <https://doi.org/10.1016/j.fcr.2016.11.003>.
- Lu S, Ren T, and Horton R (2020) Estimating the components of apparent thermal conductivity of soils at various water contents and temperatures. *Geoderma* 376: 114530.
- Massman WJ (2015) A non-equilibrium model for soil heating and moisture transport during extreme surface heating: The soil (heat–moisture–vapor) HMV-Model Version 1. *Geoscientific Model Development* 8(11): 3659–3680. <https://doi.org/10.5194/gmd-8-3659-2015>.
- Milly PCD (1982) Moisture and heat transport in hysteretic, inhomogeneous porous media: A matrix head-based formulation and a numerical model. *Water Resources Research* 18(3): 489–498. <https://doi.org/10.1029/WR018i003p00489>.
- Nassar IN and Horton R (1997) Heat, water, and solution transfer in unsaturated porous media: I – theory development and transport coefficient evaluation. *Transport in Porous Media* 27(1): 17–38. <https://doi.org/10.1023/a:1006583918576>.
- Ni J, Cheng Y, Wang Q, Ng CWW, and Garg A (2019) Effects of vegetation on soil temperature and water content: Field monitoring and numerical modelling. *Journal of Hydrology* 571: 494–502. <https://doi.org/10.1016/j.jhydrol.2019.02.009>.
- Nowicki SC, Davis HT, and Scriven LE (1992) Microscopic determination of transport parameters in drying porous media. *Drying Technology* 10(4): 925–946. <https://doi.org/10.1080/07373939208916488>.
- Parlange MB, Cahill AT, Nielsen DR, Hopmans JW, and Wendroth O (1998) Review of heat and water movement in field soils. *Soil and Tillage Research* 47(1–2): 5–10. [https://doi.org/10.1016/S0167-1987\(98\)00066-X](https://doi.org/10.1016/S0167-1987(98)00066-X).
- Philip JR and de Vries DA (1957) Moisture movement in porous materials under temperature gradients. *Eos, Transactions American Geophysical Union* 38(2): 222–232. <https://doi.org/10.1029/TR038i002p00222>.

- Prat M (1993) Percolation model of drying under isothermal conditions in porous media. *International Journal of Multiphase Flow* 19(4): 691–704. [https://doi.org/10.1016/0301-9322\(93\)90096-D](https://doi.org/10.1016/0301-9322(93)90096-D).
- Prat M (2002) Recent advances in pore-scale models for drying of porous media. Drying 2000: Selected Papers from the 12th International Symposium IDS2000 *Chemical Engineering Journal* 86(1): 153–164. [https://doi.org/10.1016/S1385-8947\(01\)00283-2](https://doi.org/10.1016/S1385-8947(01)00283-2).
- Scanlon BR and Milly PCD (1994) Water and heat fluxes in desert soils: 2. Numerical simulations. *Water Resources Research* 30(3): 721–733. <https://doi.org/10.1029/93wr03252>.
- Shahraeeni E and Or D (2012) Pore scale mechanisms for enhanced vapor transport through partially saturated porous media. *Water Resources Research* 48(5): 1–16. <https://doi.org/10.1029/2011WR011036>.
- Vanderborght J, Fetzer T, Mosthaf K, Smits KM, and Helmig R (2017) Heat and water transport in soils and across the soil-atmosphere interface: 1. Theory and different model concepts. *Water Resources Research* 53(2): 1057–1079. <https://doi.org/10.1002/2016WR019982>.

Radiation Balance

Jerry L Hatfield*, Kaitlin Togliatti, Thomas J Sauer, and John H Prueger, National Laboratory for Agriculture and the Environment, Ames, IA, United States

© 2023 Elsevier Ltd. All rights reserved.

Radiation balance	393
Incoming solar radiation	395
Albedo	395
Incoming longwave radiation	397
Terrestrial longwave radiation	397
Radiation balance	398
Implications	399
References	399

Abstract

Radiation balance represents the difference between the incoming and reflected solar radiation and the incoming and emitted longwave radiation at the earth's surface. A positive energy balance indicates that the energy impinging on the earth's surface has the capacity to evaporate water, heat the air, heat the ground, or drive photosynthesis while a negative energy balance shows that energy is being lost from the earth back to the atmosphere. It is this balance of energy that determines our ability to sustain life on earth.

Key points

- Radiation balance determines the amount of energy available at the earth's surface to heat the air or ground, evaporate water, or drive the photosynthetic process.
- Radiation balance is comprised of the short-wave (sunlight) and long-wave (derived from temperature of objects) components at the earth's surface.
- Radiation balance varies with time of day, time of year, location on the earth, and surface properties of ground.
- Radiation balance is affected by the climate and affects the climate and understanding these effects will enhance our ability to ensure food supplies, stable ecosystems, and environmental quality.

Radiation balance

Life on Earth depends upon the flow of energy. Living objects balance their energy inputs with outputs to create a comfortable environment. Humans have an advantage because they create clothing and shelter to help modify their environment. Plants and soil also have inputs of energy, but have limited controls on the dissipation of energy and even less control on the energy inputs into the system. When we begin to think about the balance of energy in the soil or plant system, we can begin to understand some of the remarkable balances that exist within the natural environment. Understanding the surface radiation and energy balances provides an opportunity to explore the role of different components (e.g., crop residue, tillage, cover crops), on soils and the environment (Monteith and Unsworth, 2013; Rosenberg et al., 1983).

The balance of radiation is similar to a checkbook or savings account. How much money is in it at any one time depends upon how much has been put into the account and how much has been taken out. This balance can be expressed very simply in word form. Net radiation is the difference between the gains (incoming solar radiation and absorbed long-wave radiation) and losses (reflected and transmitted solar radiation and emitted longwave radiation). This 'budget' can be expressed mathematically as:

$$R_n = (1 - \alpha)S_t + L_d - \varepsilon\sigma T_s^4 \quad (1)$$

where R_n is the net radiation at the surface expressed as the amount of energy (joules) per unit area (meters squared) per unit time (seconds) or watts per meter squared. All of the radiation terms (S_t , L_d , and $\varepsilon\sigma T_s^4$) in Eq. (1) must have the same units to avoid problems. As we expand these terms, S_t is the solar radiation input, α the albedo or reflectance of the surface, L_d the input of longwave radiation, $\varepsilon\sigma T_s^4$ is the energy lost as longwave radiation from the surface. Shortwave radiation represents the energy in the 0.3- to 3- μ m range, while longwave radiation represents wavelengths of more than 3 μ m. Another way to consider this difference is that shortwave radiation originates from the sun, with a surface temperature near 6000 K, while longwave radiation is emitted from objects with temperatures near 300 K. Current technology with radiometers allows for the direct measurement of these components

*Retired

and is shown in Fig. 1 as a 24 h sequence of incoming shortwave, reflected shortwave, incoming longwave, and emitted longwave for 4 days over a maize field in central Iowa. The convention that is used for the radiation balance components is that positive values show energy directed toward the earth's surface and negative energy values are directed toward the atmosphere. In this sequence of days, there are days in which the solar radiation is affected by cloud cover and show a large variation in the values, which affects how much solar radiation is reflected back to the sky. The longwave components do not exhibit a large change when it is cloudy (Fig. 1). Only the longwave components are detectable at night because there is no energy from the sun after sunset and before sunrise (Fig. 1). On a cloudy day, there is a marked reduction in solar radiation and longwave components become nearly equal. The radiation balance for a surface is the combination of these components and are shown for these 4 days in Fig. 2. Radiation balance values throughout the day do not always produce a smooth curve because of the presence of clouds. Nighttime values are zero or below because the Earth's surface tends to lose more energy than it receives. Energy available as net radiation is further divided into other components. We express the energy balance as:

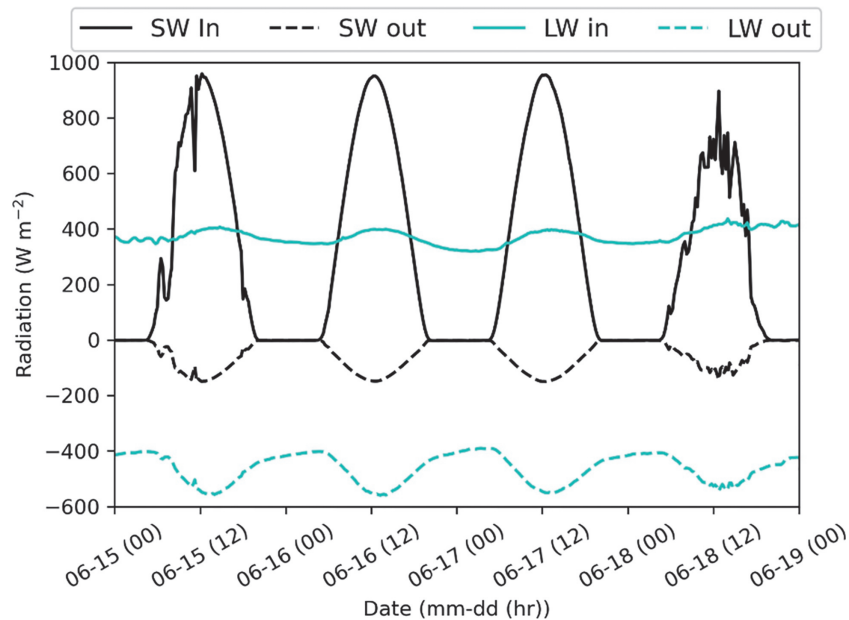


Fig. 1 Components of net radiation for 4 days (DOY 167-170) in June over a maize canopy in central Iowa showing shortwave (SW) and longwave (LW) inputs from the sun (in) and emitted from the back from the land surface (out).

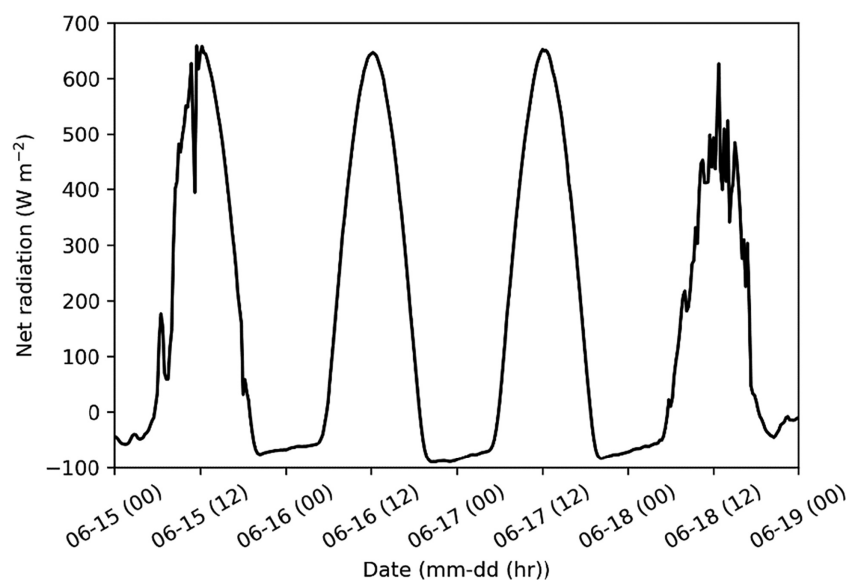


Fig. 2 Net radiation balance for 4 days (DOY 167-170) during the midsummer growing season in central Iowa based on the components shown in Fig. 1.

$$R_n = LE + H + G \quad (2)$$

where LE represents the energy used to evaporate water, H the energy to heat the air, and G the amount stored in the soil volume. The more energy input to the surface the greater the potential magnitude of these components (see section “energy balance”).

Incoming solar radiation

Solar radiation originates at the Sun and travels through space and the Earth’s atmosphere before it reaches the Earth’s surface. Solar radiation consists of two components, the direct component (S_b) and the diffuse (S_d) component. Diffuse radiation comes from all angles, while the direct beam follows the path of the Sun throughout the day. Diffuse radiation is formed by the scattering of the direct beam by particles in the atmosphere. Physically, we can observe the effect of direct and diffuse radiation by sharpness of shadows. On a clear day, the outlines of the shadows are clearly defined, while on a cloudy day there are no shadows yet we still can see. Mathematically, we represent incoming solar radiation as:

$$S_t = S_b \sin \beta + S_d \quad (3)$$

The angle of the Sun’s beam above a horizontal surface, β , distributes the energy in the beam over a different area than if the surface were perpendicular to the beam. The magnitude of direct and diffuse radiation depends upon the solar elevation (the angle of the Sun’s beam above the horizon), the amount of scattering and absorption by atmospheric gases and aerosols, and the amount of scattering and absorption by water droplets and ice particles in the clouds. Particles in the atmosphere have a large impact because they scatter radiation in all directions, some back to space and some toward the Earth’s surface. As the number of particles in the atmosphere increases the portion of diffuse radiation increases.

There is large variation in the incident solar radiation at many locations throughout the year because of the changing angle of inclination of the Earth with respect to the Sun. At the equator there is little difference; however, at the midlatitudes there are large differences between summer and winter. At 42°N latitude, clear-sky solar radiation incident on the Earth at midday is approximately 26 J m⁻² per day during summer, but approximately 9 J m⁻² per day during winter. Clouds affect these totals and need to be considered. There are tables of clear-sky solar radiation incident on the Earth’s surface for any location: these tables have been assembled through measurements and empirical calculations.

The slope of the surface is a critical factor in the radiation balance. If there is a south-facing slope in the northern hemisphere then more energy is going to be absorbed than if there is a north-facing slope. More subtle differences exist if the slope of the soil surface is changed with tillage implements. The creation of a furrow with a tillage tool that runs east to west across the field creates a surface in which one part will receive more direct solar radiation than the other. This will cause the soil surface and soil temperatures to be warmer on the south side and cooler on the north side in the northern hemisphere.

Temperatures for bare soil at the surface under these conditions may be 10–20 °C different at midday. This same furrow shape placed in a north-south direction will have differences in surface temperatures from morning to afternoon. The geometry of the surface plays a critical role in the radiation balance through the effect of direct-beam solar radiation. Manipulation of the soil surface through tillage implements can create a specific microclimate at the soil surface that influences temperature-dependent processes within the soil.

Albedo

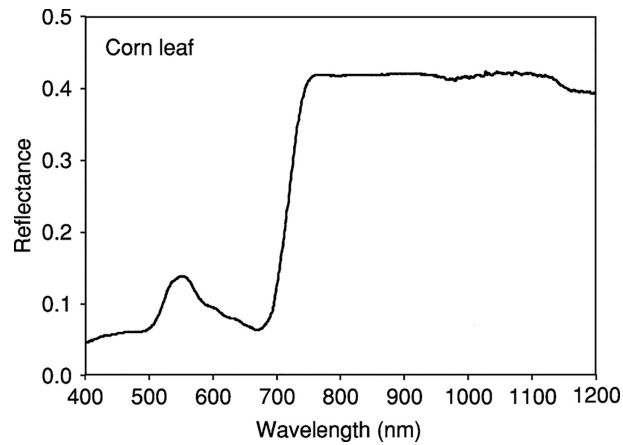
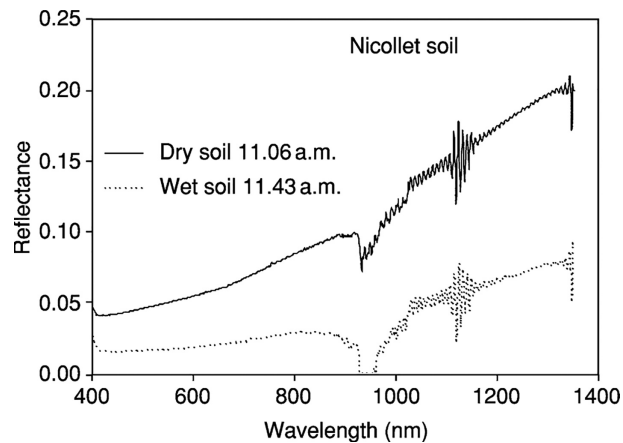
Albedo of a surface is the ratio of amount of solar radiation reflected back to the atmosphere to the amount of incoming solar radiation (S_r/S_i). We see a highly reflective surface as white, and a low-reflectance surface as black. A dark surface will absorb more radiation than a light-colored surface. The range of albedo values for different surfaces is shown in Table 1.

Albedo is more complex than assuming a single constant value for an agricultural soil surface. Albedo values for a soil surface change with soil water content, type of soil, tillage practice, amount and type of crop-residue cover, and age of residue. Agricultural soils are modified by tillage, residue placement, and the growth of a crop and throughout the year there is a constant change in the albedo of a soil surface. Starting with a smooth soil surface with an albedo of 0.20, the addition of water to a soil surface decreases the albedo to 0.15, with a slow return to the original value as the surface dries. Water is an effective absorber of short and longwave radiation and adding water decreases the albedo of any surface. If the same surface is roughened by tillage, there is a further decrease in the albedo of a dry surface to near 0.15. This decrease is due to the change in the soil surface that traps more direct- and diffuse-beam solar radiation because of the change from a smooth to a rough surface. Wetting this tilled surface further decreases the albedo. Adding a fresh crop residue increases the albedo because of the highly reflective nature of the crop residue; however, as the residue ages, the albedo decreases and may be only slightly different from the soil surface in the spring. Soil surfaces are continually undergoing changes that affect their albedo. This will impact the radiation balance and ultimately the evaporation of water from the surface, heating the air, and warming the soil.

The albedo values shown in Table 1 represent an integrated value over the shortwave wavelengths and are determined by measured the incoming solar radiation with a pyranometer and reflected radiation with an inverted pyranometer of the same type. Within the solar radiation spectrum, there are differences in reflectance, as shown in Fig. 3 for a corn leaf. There is a dramatic change

Table 1 Albedo values for different natural surfaces.

Surface	Albedo
Bare soil, dry	0.20
Bare soil, wet	0.15
Bare soil, fresh corn residue	0.35
Bare soil, weathered corn residue	0.22
Corn canopy	0.22
Wheat canopy	0.26
Rice canopy	0.24
Deciduous forest with leaves	0.18
Coniferous forest	0.16
Tundra	0.20
Fresh snow	0.80

**Fig. 3** Reflectance spectra for a corn leaf in the visible and near-infrared wavelengths.**Fig. 4** Reflectance of solar radiation across visible and near-infrared wavelengths for a smooth, tilled, Nicollet soil when wet and dry.

in the reflectance between the visible and near-infrared wavelengths from $0.6 \mu\text{m}$ to $0.7 \mu\text{m}$. The difference in reflectance spectra across the wavelengths in plant material shows low reflectance values in the blue and red portions of the spectrum, an increase in the green wavelengths, and then a sharp increase in the near-infrared wavelengths. Reflectance values for a bare soil across the solar radiation wavelengths show a different pattern from a leaf and tend to increase linearly with wavelength, not showing the large differences among portions of the spectrum (Fig. 4). Tilling the soil to change the roughness and increases in soil water content will affect soil reflectance. There is a decrease in the reflectance values for all wavelengths (Fig. 4). Adding crop residue to the soil

surface also changes the reflectance of the surface and changes the energy-balance components. Melting of snow greatly decreases albedo, resulting in soil warming.

When solar radiation impinges on a leaf or any other object, the radiation may be reflected, absorbed, or transmitted. If we imagine ourselves underneath a thin layer of white cloth, from the upper side we would see the white color owing to reflection of all wavelengths; however, underneath the cloth we would still see because of the light being passed through (transmitted). Some of the light would be absorbed by the cloth. Under a 10-cm layer of soil, there would be no transmitted light and there would be no shortwave radiation, therefore no visibility. This same principle applies when layers of material are added to the soil surface. A layer with a high light transmittance allows solar radiation to reach the soil surface and warm the soil, evaporate water, or warm the air next to the soil surface. When a highly reflective and low-transmittance material, such as snow, is placed on the soil surface, there is little change in the temperature of the soil surface with varying solar radiation affecting the radiation balance. Transmittance values for different materials vary widely, as shown in Table 2 for both the shortwave and longwave regions of the spectrum. Transmittance of solar radiation through materials, especially plant material, has a large effect on the radiation balance at the soil surface. Very dense canopies transmit some light through all of the layers of leaves to the soil surface. Humans are able to see in this region; however, a leaf receives little energy. Plants grown under other canopy layers often have very large leaves with less chlorophyll (more yellow) than the same plant grown in full sunlight. The radiation balance of these plants and the soil under dense canopies is much different from the environment without this canopy.

Incoming longwave radiation

All objects with an absolute temperature of more than 0 K emit energy. Objects in our atmosphere such as gases, particles, ice crystals, or water droplets have temperatures above absolute zero and emit energy. The energy emitted is equivalent to the 4th power of their temperature as described by the Stefan-Boltzmann law $E = \epsilon \sigma T^4$. The Stefan-Boltzmann constant, σ , is $5.67 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-1}$. The amount of energy emitted by an object greatly increases as the temperature increases. The emissivity term, ϵ , can be considered an efficiency factor. The higher the emissivity of an object the more effective the emission of longwave radiation away from the object. Longwave radiation is measured with radiometers designed to capture only these wavelengths; however, there are relatively few of those instruments located around the world. Incoming longwave radiation is generally estimated from mathematical equations in which the emissivity of the atmosphere, ϵ_a , is estimated through a combination of water vapor or temperature measurements to modify the energy relationship to $\epsilon_a \sigma T_a^4$. In one of the original relationships for this estimation, it was proposed that $\epsilon_a = 0.51 + 0.66e_0^{1/2}$, where e_0 is the water vapor pressure of the atmosphere. There have been several attempts to refine this relationship, for example, where ϵ_a is estimated by $0.575e_0^{1/7}$ and ϵ_a is best represented by $0.70 + 5.95 \times 10^{-5} e_0 \exp. (1500/T_0)$, with T_0 the air temperature in degrees Kelvin (Hatfield et al., 1983). Both e_0 and T_0 are obtained from meteorological station data at a height of 2 m above the surface and are available every hour throughout the day. A comparison of all of these different methods with data collected across the United States has found that there is little difference among the methods and all predicted longwave radiation to within 5% of the measured values. Emissivity values for clear skies can range from 0.6 to 0.75, and these equations apply to the clear-sky condition. When there are clouds, the ϵ_a term can be considered to be near 1. Consider what happens to air temperature on a clear versus a cloudy night. On the clear night there is a rapid drop in air temperature because the atmosphere is contributing little energy to the Earth's surface. On a cloudy night or an extremely humid night there is little change in air temperature because the longwave radiation lost from the Earth is captured by the clouds or water vapor and reemitted back to Earth's surface.

Terrestrial longwave radiation

Objects on the Earth's surface have a temperature and emit energy in relationship to their temperature as discussed for longwave energy emitted from the atmosphere. The same principle applies to the soil surface or objects on the Earth's surface. The last term in Eq. (1), $\epsilon_a \sigma T_s^4$, represents the loss of energy away from the surface into the atmosphere. Emissivity values for objects on Earth are

Table 2 Transmittance values for different natural surfaces.

Surface	Transmittance
Corn leaves, visible	0.08
Corn canopy, visible	0.05
Wheat leaves, visible	0.07
Wheat canopy, visible	0.06
Fresh corn residue, visible	0.05
Weathered corn residue, visible	0.04
White plastic, visible	0.95
Black plastic, visible	0.0

Table 3 Emissivity values for different natural surfaces.

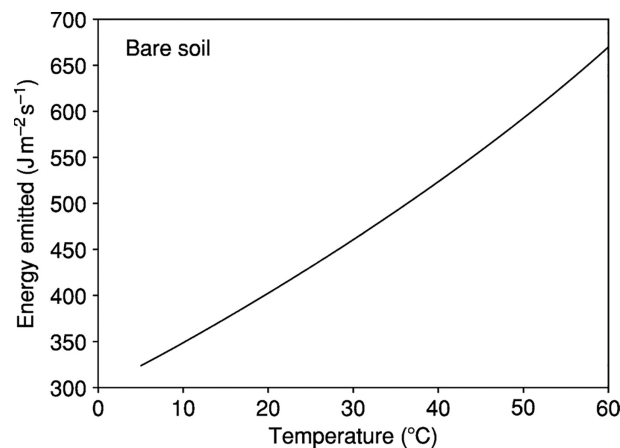
Surface	Emissivity
Bare soil, dry	0.97
Bare soil, wet	0.99
Bare soil, fresh corn residue	0.97
Bare soil, weathered corn residue	0.98
Corn canopy	0.94
Wheat canopy	0.96
Deciduous forest, foliage present	0.99
Coniferous forest	0.99

close to 1; however, there is variation, as shown in Table 3. The variation is small and often the assumption is that 0.95 or 0.97 are adequate mean values that can be used. The T_s term in the relationship represents the surface temperature of the object. This would be the temperature at the outer surface of an object or at the very top layer of the soil surface. There are large gradients of temperature in the soil, and for a dry soil the surface temperature at midday may exceed 50 °C while at 10 mm below the surface the temperature is less than 40 °C. In the estimation of the longwave energy lost from the surface, a longwave radiometer can be inverted over the surface to measure the amount of energy escaping or estimated using measurements of surface temperature obtained from infrared thermometers combined with an estimate of emissivity for the surface. The latter approach works quite well for smooth or uniform surfaces; however, for rough or undulating surfaces or surfaces comprised of a combination of plants and soil, there is a large variation in surface temperature and measurements would have to be made that would represent the average value for the entire surface area. The magnitude of the energy lost from surfaces with different temperatures is shown by Fig. 5. As the surface temperature increases, the energy emitted increases rapidly, because emission is a function of the 4th power of surface temperature.

Radiation balance

Each of the terms in Eq. (1) has been discussed separately to develop an appreciation for the variation within each component and the factors that affect the magnitude of the energy within each term. As these components are reassembled into the radiation balance for a surface it is possible to see how soil-management factors might affect the radiation balance. There are changes throughout the day as shown in Fig. 2. If the radiation balance is expanded into the components for each hour throughout the day, the differences that each component adds to the balance can be seen. The largest factor is the solar radiation term, and the difference between the 2 days is a direct result of the incoming solar radiation. The least amount of variation among days is introduced by the incoming longwave radiation, L_d , since this factor is primarily affected by air temperature. Longwave loss is actually highest during the day when the surface temperatures are highest. As these relationships are expanded into annual radiation balances, surface temperature and air temperature variations among days across the seasons have a large impact. If the annual radiation balance over the year is computed, the sum of the inputs and losses is nearly zero.

On a daily interval, changing the soil surface through tillage or crop-residue management has a large impact on the radiation balance. If a smooth, dry soil is tilled, there are several changes in the radiation balance components. Incoming solar and longwave radiation do not change. The albedo decreases so there is a greater adsorption of solar radiation leading to a greater amount of energy available at the soil surface. However, evaporation of soil water from the wet, rough surface decreases the surface

**Fig. 5** Energy emitted from a soil surface under different surface temperatures.

temperature, creating a situation in which the surface emits less energy. The end result is a surface that has an increased amount of energy available than did the surface before tillage. This effect lasts only as long as there is soil water to evaporate, but there is some effect of the rough surface on reducing the albedo of a dry soil; however, this effect is small compared with the initial period with a wet soil surface.

The final example is the soil surface to which crop residue has been added. In the case of fresh crop residue there is an increase in the albedo of the surface and a reduction in net radiation. Weathered residue decreases the albedo and allows the surface to have an increased net radiation loading. Wetting and drying of the crop residue has some effect on the radiation balance because of the effect of water on both the albedo and emissivity of the surface. Soil-management practices have a large impact on the radiation balance, so knowledge of which factor is affected by management will help provide an improved understanding of the consequences of management decisions.

Implications

The radiation balance of a surface represents the balance of shortwave gains and losses and longwave gains and losses. The radiation balance changes throughout the day and year owing to incoming solar radiation, which has the largest effect on the radiation balance, followed in importance by the cloud cover. The next major factor is the slope of the surface, which affects the amount of solar radiation absorbed. The slope factor can be both a large-scale (slope of hills and mountains) and microscale factor (creation of furrows and ridges with tillage implements). Humans have a large impact on the radiation balance and can manipulate the energy balance through all of the components in the radiation balance.

The radiation balance determines our ability to effectively produce food required for food security; however, it is not the only factor in the complex set of processes linked in plant growth. Understanding the basic source of energy at the earth's surface provides the context for how ecosystems function and thrive.

References

- Hatfield JL, Reginato RJ, and Idso SB (1983) Comparison of long-wave radiation calculation methods over the United States. *Water Resources Research* 19: 285–288.
- Monteith JL and Unsworth MH (2013) *Principles of Environmental Physics*, 4th edn. New York: Elsevier.
- Rosenberg NJ, Blad BL, and Verma SB (1983) *Microclimate: The Biological Environment*, 2nd edn. New York: John Wiley.

Freezing and thawing cycles

Alan W Rempel, Department of Earth Sciences, University of Oregon, Eugene, OR, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of B. Sharratt, FREEZING AND THAWING I Cycles, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 98–103, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00528-2>.

Introduction	400
Dominant controls	401
Thermal conditions	401
Liquid supply	403
Constitutive behavior	403
External loads	405
Solutes	406
Climatic dependence of freeze–thaw behavior	406
Climate change	407
Mitigation strategies	407
Outlook	408
References	409

Abstract

Pore-scale interactions cause ice in sediments to melt over a range of subzero (centigrade) temperatures; this is a consequence of the phenomenon known as premelting. Liquid potential gradients in ice-infiltrated sediments drive flow from warmer to colder temperatures and from fine-grained to coarse-grained regions, supplying ice growth that can produce frost damage as a result of ice–sediment stresses that strengthen with cooling. On thawing, elevated liquid contents can reduce sediment strength dramatically while drainage is inhibited by residual pore ice. Appreciable enhancements to solute concentration during freezing and dilution during thawing also have important biological and mechanical consequences.

Key points

- surface energy and wetting interactions produce temperature-dependent variations in pore space ice and liquid saturation levels, with notable reductions in permeability under colder ice-clogged conditions
- stresses, transmitted between wetted ice and sediment surfaces, strengthen and promote deformation in colder regions, where ice growth is supplied by residual liquid driven down potential gradients that parallel the conductive heat flux
- liquid supply down large temperature gradients during transient freezing promotes deformation at cold temperatures and thawing reduces sediment strength while drainage is impeded by nearby ice
- freeze–thaw impacts are minimal in temperate climates, increase in severity in colder climates, and diminish in severity within climates that are cold enough for near-surface unfrozen conditions to be rare
- climate change is expected to reduce freeze-thaw damage in some temperate climates that warm enough to reduce the occurrence and severity of freezing; freeze-thaw damage is expected to increase in cold climates that are warming rapidly and experiencing more freeze-thaw cycles

Introduction

Freeze-thaw cycles can have dramatic consequences, including frost damage to infrastructure, heave of the ground surface, and enhanced erosion of water-logged soils. The modest $\sim 10\%$ density contrast between ice and liquid water is incidental to most of this behavior. Instead, in porous materials like soils, the mechanical disruption associated with freeze–thaw cycles can be traced to two primary underlying factors: (1) the intermolecular forces that transmit stresses between ice and sediment surfaces across intervening liquid films that thin to enable higher stress transmission at colder temperatures; and (2) the progressive clogging by ice of potential flow pathways that restricts the mobility of residual water (referred to collectively as premelted liquid), which is contained both in the films and in high-curvature regions where ice–liquid surface energy enables equilibrium coexistence. Transient heat flow during freeze–thaw cycling results in temperature gradients, which produce gradients in ice–sediment stresses (factor 1) that are balanced by liquid pressure gradients to satisfy force equilibrium. These pressure gradients commonly drive liquid flow toward colder subzero (centigrade) temperatures and draw water from liquid-saturated fine-grained aggregates into larger macropores where ice is present. Permeability reductions cause the liquid flux to diminish along increasingly constricted flow paths

(factor 2), with mass balance satisfied by ice deposition. The accommodation of this additional ice often entails significant deformation (tied to factor 1), sometimes unloading sediment particle contacts to produce segregated ice and driving frost damage at a variety of scales. Upon thawing, the liberation of liquid quantities that exceed close-packed pore volumes causes local strength reductions and enhanced erosion, particularly when drainage pathways are restricted by elevated ice concentrations in nearby regions that remain frozen (tied to factor 2).

Dominant controls

From this basic understanding of how freeze-thaw cycles disrupt soil systems, several factors emerge as dominant controls on the severity of consequences in different environments. These can broadly be categorized as pertaining to (i) local time-evolving thermal conditions, (ii) the distribution and supply of moisture, (iii) soil constitutive behavior, and the presence of additional (iv) mechanical and (v) chemical loads. Interrelationships between these factors lead to the rich variety of observed responses to freeze-thaw cycling, rendering the summary here more illustrative than comprehensive.

Thermal conditions

Frost damage tends to be greatest when the temperature at the ground surface cycles between cold and warm conditions, with thermal properties and the frequency of freeze-thaw cycling determining the depth range over which damped thermal signals propagate. The modern understanding for why such temperature cycling produces damage is rooted in the modified phase behavior of ice and liquid water that is associated with their confinement in small (i.e. pore) spaces (e.g. see Fig. 1C).

Thermodynamic arguments first elucidated by Gibbs (1878) enable the absolute temperature $T = T_{eq}$ at an ice-liquid interface with curvature K to be written in terms of its offset from a reference temperature T_0 in response to changes in the local liquid pressure P_l and the force per unit area P_T (referred to in the literature as the disjoining pressure or the thermomolecular pressure) that is exerted on the ice by a foreign substrate, so that

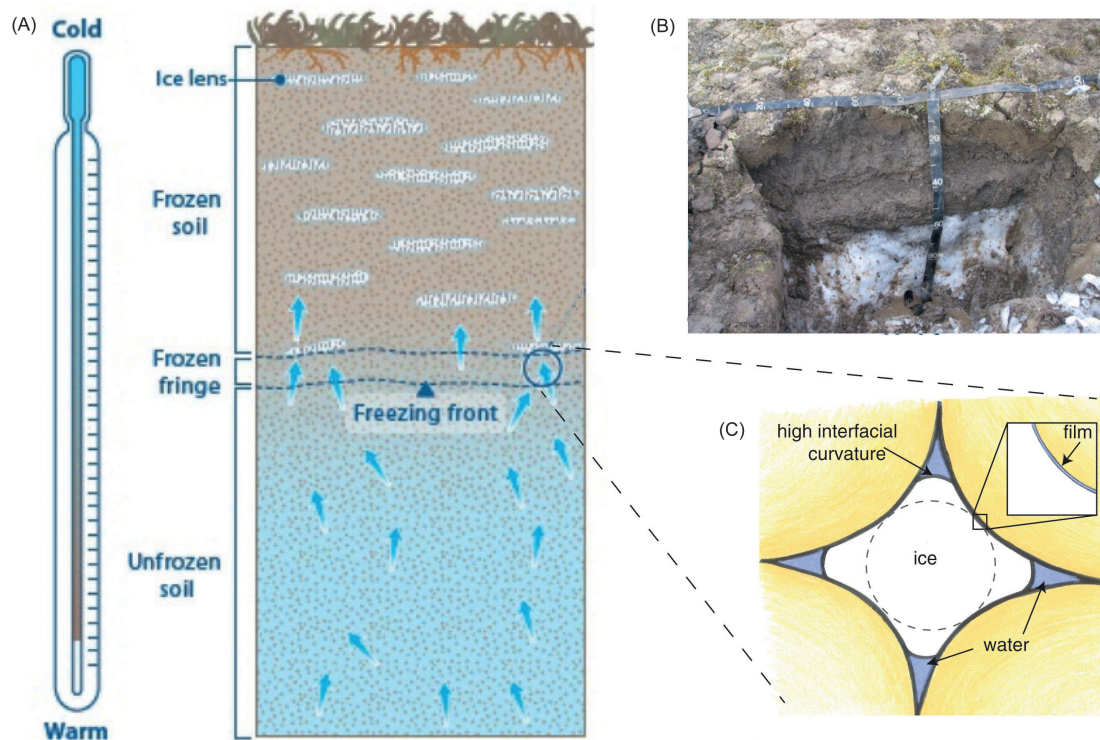


Fig. 1 (A) Schematic diagram of a soil column, with segregated ice lenses in the cold frozen soil near the ground surface and unfrozen soil at the warmer temperatures prevailing at depth. Immediately beneath the lowest ice lens is a frozen fringe in which the pore space is characterized by gradients in the relative proportions of ice and the liquid water, which is drawn up from below to supply lens growth. (B) Photograph of ice-infiltrated sediments with visible segregated ice in an excavation to a depth of several decimeters. Figure components provided by Dani Or. (C) Close-up view of the ice–water configuration in an idealized pore above the freezing front at which ice first appears. Larger residual liquid pockets found near particle contacts are bounded by surfaces with high ice–water interfacial curvature K and joined by thin liquid films that separate ice from particle surfaces while enabling the transmission of significant disjoining pressures P_T .

$$T_{eq} \approx T_0 - \frac{T_0}{\rho_l \mathcal{L}} (P_T + \gamma_{il} \mathcal{K}) - \frac{\rho_l - \rho_i}{\rho_l} \frac{T_0}{\rho_l \mathcal{L}} (P_l - P_0). \quad (1)$$

Here, ρ_i is ice density, ρ_l is liquid density, $\mathcal{L}$ is the latent heat of fusion, γ_{il} is the ice–liquid surface energy, and the reference pressure P_0 is defined so that $T_{eq} = T_0$ when $P_l = P_0$ under the bulk equilibrium conditions that pertain when the compressive normal stress in the ice σ_n is equal to the liquid pressure P_l , or equivalently $\sigma_n - P_l = P_T + \gamma_{il} \mathcal{K} = 0$. The approximation in Eq. (1) holds under typical environmental conditions since deviations from bulk equilibrium tend to be sufficiently small that the finite compressibility of the liquid can be neglected and the ice can also be treated as perfectly rigid; it should be noted that any changes in liquid solute content c will complicate matters by causing commensurate changes to the reference temperature so that $T_0 = T_0(c, P_0)$. For the important limiting case of pure water (i.e. with $c = 0$) and values of P_0 commonly encountered in near-surface environments (i.e. $P_0 \approx 10^5 - 10^6$ Pa), the ratio $T_0/\rho_l \mathcal{L}$ is approximately 9×10^{-7} K Pa⁻¹, so that a temperature depression, or undercooling, of $T_0 - T_{eq} = 0.1$ K along a flat ice–liquid interface (i.e. with $\mathcal{K} = 0$) entails a thermomolecular, or disjoining pressure of $P_T \approx 10^5$ Pa in a well-drained system for which P_l is maintained at P_0 . This indicates that along an interface where the normal stress on the ice and the pressure in adjacent liquid differ, even a modest undercooling can lead to stress transmission between the ice and pore walls that is large enough to support an appreciable load and often induce significant compaction. Moreover, since the density ratio that multiplies the final term $(\rho_l - \rho_i)/\rho_l \approx 0.08$ is much less than unity, changes in the local temperature tend to be accommodated primarily by changes in P_T and $\mathcal{K}$ rather than by changes in the liquid pressure on its own, leading to significant stress gradients even when liquid transport is restricted so that P_l deviates from P_0 .

As described by Eq. (1), the strength of the net force per unit area P_T (i.e. factor 1 discussed in the “Introduction” Section) exerted between ice and soil particles (or other foreign substrates such as rock, pipes or concrete) across intervening premelted liquid films increases at colder temperatures that are offset further below bulk melting – that is 0 °C at atmospheric pressure for pure water and the liquidus temperature when solutes are present (e.g. Dash et al., 2006). In addition to the local temperature, the temperature gradient is important as well, both because of its influence on the potential gradients that drive liquid supply through a process that is often referred to as cryosuction, and because of its connection to changes in relative permeability that restrict liquid transport (i.e. factor 2 discussed in the “Introduction” Section) (e.g. O’Neill and Miller, 1985; Rempel et al., 2016; Walder and Hallet, 1986). A predictive understanding of the effects of freeze–thaw cycling is hence intimately tied to assessments of how the local temperature in ice-bound porous media varies through time and space (e.g. a common wintertime spatial pattern is illustrated in Fig. 1A).

Because the depth range over which significant temperature changes occur near the ground surface is typically much shallower than the horizontal length scales upon which variations of similar magnitude take place, thermal conditions in the subsurface are often approximated well by idealized one dimensional energy balance considerations. The time t and depth z varying conductive heat flux perpendicular to the ground surface can be written in terms of the effective thermal conductivity K_e as

$$Q(z, t) = -K_e \frac{\partial T}{\partial z}. \quad (2)$$

Crucially, the surface heat flux $Q_s = Q(z = 0, t)$ that drives temperature changes below is sensitive not only to local meteorological conditions, but also to the insulating effects of snow-cover and crop residues, variations in radiative exchange that are associated with differing slope aspect, albedo, and canopy cover, and factors controlling convective heat exchange, such as surface roughness and the presence or absence of windbreaks (e.g. Sharratt, 1993; Sharratt et al., 1998; Henry, 2007; Santamouris et al., 1999). Beneath the surface, heat is transported primarily by conduction, with advective contributions from heat carried by moving water or air, typically (though not always) negligible, but in a narrow range of temperatures slightly colder than T_0 (controlled by the soil “Constitutive behavior” discussed below) the latent heat associated with changes in ice content can significantly affect the thermal profile (for example in Fig. 1A such conditions prevail in the labeled frozen fringe).

Setting $T = T_{eq}$ and differentiating Eq. (1) with respect to depth, while making a substitution to express the middle term as proportional to the difference between the normal compressive stress on the ice σ_n and the liquid pressure P_l yields the liquid potential gradient as

$$\frac{\partial P_l}{\partial z} - \rho_l g_z \approx \frac{\rho_l \mathcal{L}}{T_0} \frac{\partial T}{\partial z} + \frac{\rho_l}{\rho_i} \left(\frac{\partial \sigma_n}{\partial z} - \rho_i g_z \right) \quad (3)$$

where T_0 has been treated as constant (valid when solute concentration is uniform, e.g. $c = 0$), and g_z is the component of gravitational acceleration perpendicular to the ground surface. In the idealized case where the stress state in the ice can be approximated as isotropic with a depth dependence that is hydrostatic (i.e. $\partial \sigma_n / \partial z = \rho_i g_z$), the term in parentheses on the right disappears to reveal a simple direct relationship between temperature gradients (or the conductive heat flux represented by Eq. 2) and liquid potential gradients. Since water flow is driven by liquid potential gradients, enhanced water transport is favored when the conductive heat flux Q is high, but the “Constitutive behavior” discussed below entails large reductions in permeability at colder temperatures that impede water transport. As a result, at small sub-zero temperatures, water transport is only modestly restrained, while the capacity for deformation is limited by relatively weak stress transmission between ice and sediment surfaces (i.e. P_T is small). At cold temperatures, large stresses are generated between ice and sediment surfaces, but deformation rates are limited by the restrictions in water supply that accompany the formation of ice lenses and the clogging of pores with high concentrations of ice. As temperatures change through time, these competing factors enable deformation to be most rapid when local conditions are cold enough for ice–sediment stresses to support external loads, yet warm enough for liquid flow to supply ice growth, with thermal

gradients sufficiently steep, both to promote the potential gradients that drive liquid transport and to conduct away latent heat (e.g. Rempel et al., 2016). Freeze–thaw cycling is particularly effective at providing the thermal conditions needed to drive the ice accumulation in sediments that produces frost damage.

Liquid supply

Mechanical disruption from freeze–thaw cycling is most intense when the pore space is completely saturated with ice and liquid water so that gradients in liquid flux can be accommodated only by ice deposition that drives associated deformation, rather than simply by compressing or displacing pockets of residual air and water vapor (e.g. Walder and Hallet, 1986). Substituting the liquid potential gradient from Eq. (3) into the idealized one-dimensional version of Darcy’s law, the discharge velocity q_z of the residual liquid in ice-infiltrated soils becomes

$$q_z \approx -\frac{k}{\mu} \left(\frac{\rho_l \mathcal{L}}{T_0} \frac{\partial T}{\partial z} + \frac{\rho_l}{\rho_i} \frac{\partial \sigma_n}{\partial z} - \rho_l g_z \right), \quad (4)$$

where μ is the liquid viscosity and k is the permeability. (Crucially, k depends on the spatially and temporally varying ice content, as described in the section on “Constitutive behavior”.) The first term in parentheses describes how the liquid water that supplies ice growth is often drawn from warmer temperatures where the liquid potential is elevated relative to colder zones where stronger ice–sediment interactions require that liquid pressures drop to satisfy force balance constraints. To make the connection between liquid transport and heat transport more explicit, the expression for Q from Eq. (2) can be substituted for the temperature gradient to write

$$q_z = \frac{k \rho_l \mathcal{L}}{\mu T_0 K_e} Q - \frac{k}{\mu} \left(\frac{\rho_l}{\rho_i} \frac{\partial \sigma_n}{\partial z} - \rho_l g_z \right), \quad (5)$$

which highlights the fact that in the many common situations for which the first term on the right dominates (i.e. when $\partial \sigma_n / \partial z \approx \rho_l g_z$), the discharge velocity is directly proportional to the conductive heat flux; this motivates the popular term for liquid flow in ice-infiltrated sediments: cryosuction.

Consider, for example, the thermal scenario illustrated schematically in Fig. 1A. When temperature increases monotonically with depth, the lower reaches of partially frozen sediments are supplied by liquid drawn from warmer unfrozen zones below. Here, the liquid flux is directly tied to the (negative, or upwards directed) conductive heat flux through Eq. (5), while the liquid pressure in the unfrozen zone is typically determined by regional groundwater flow considerations. In contrast, during the warm phases of freeze–thaw cycling, the temperature profile is more complicated, with elevated temperatures near the surface giving way to freezing conditions at depth (i.e. Q is positive near the ground surface), typically reaching a minimum (where $Q = 0$) before warming as the depth increases further still (i.e. Q is negative at depth). Such a thermal profile enables continued ice accumulation and associated deformation in the portion of the sediments that overlies colder, ice-clogged layers, as liquid is induced to flow down from above. A special case of this behavior that is perhaps more intuitive, occurs following periods of near-surface thaw in which the moisture that is liberated by melting elevates the water table, helping to cause liquid to seep downwards and refreeze to increase ice concentrations at greater depths.

The situation is more complicated when heterogeneities in the distribution of pore sizes are significant. For example, in temperature regimes that allow fine-grained aggregates to remain fully saturated with liquid water while adjacent macropores contain ice (see “Constitutive behavior” Section), liquid can be drawn out to freeze and drive aggregate compaction, leading to appreciable disruption of soil structure. More generally, however, even such small-scale persistent liquid reservoirs are not required for gradual ice redistribution, as depressed liquid pressures in cold regions can pull in moisture from warmer regions, supplied by partial melting in ice-bound sediments or surface snow. Indeed, even when temperature gradients are completely absent, the difference between the final two terms in the parentheses of Eqs. (4) and (5) describes the potential available to drive liquid transport whenever gradients in the normal stress exerted on the ice surface deviate from hydrostatic so that $\partial \sigma_n / \partial z \neq \rho_l g_z$.

Constitutive behavior

Along a contorted ice–liquid interface within the sediment pore space (e.g. see Fig. 1C), Eq. (1) describes how phase equilibrium is perturbed by the influence of surface energy wherever the interfacial curvature $\mathcal{K}$ is nonzero and by the net disjoining pressure P_T that results from intermolecular forces involving both condensed water phases (i.e. liquid and ice) as well as the molecules of the pore walls (e.g. Dash et al., 2006). The extension of ice–liquid equilibrium from a single bulk melting temperature (i.e. 0°C at atmospheric pressure for pure water) to a lower temperature range results from increasing differences between the ice normal stress and liquid pressure as the temperature drops; this general feature of equilibrium solid–liquid phase behavior, known as premelting, is observed and expected for many different materials besides water. Importantly, the combined effects of $P_T + \gamma_{il}\mathcal{K}$ can be equated with the difference between the compressive ice normal stress and the liquid pressure $\sigma_n - P_l$ irrespective of the detailed nature of the microphysical interactions that generate P_T . However, at a given value of P_T the thickness d of the liquid film that separates the ice

from sediment particles or other foreign substrates depends on how this net repulsive force per unit area is generated. Commonly, in the limit that $\mathcal{K} \rightarrow 0$, $P_T(d)$ is approximated using a power law relationship between thickness d and undercooling $T_0 - T_{eq}$; a logarithmic dependence can be more appropriate when the ice and substrate have a finite surface charge (e.g. Hansen-Goos and Wettlaufer, 2010), but at moderate pH values (i.e. 6–9) predictions suggest that charge densities on ice surfaces are sufficiently weak that this logarithmic regime may be of limited relevance (Daigle, 2021). A rich variety of other possibilities besides simple power-law and logarithmic dependencies exist, all of which share the general feature that thinner films enable stronger net interaction strengths so that $\partial P_T / \partial d < 0$.

The soil freezing curve that describes how the volume fraction of ice within the pore space S_i increases as the liquid water content is diminished at colder temperatures is important for altering the conductive heat flux as a result of latent heat effects and for modulating liquid transport through its influence on permeability variations. Soil freezing curves can be approximated numerically with reference to detailed pore-scale geometrical considerations and film wetting behavior, and measured empirically with a range of techniques, including nuclear magnetic resonance, time domain reflectometry, and calorimetry (e.g. Chen et al., 2020; Hansen-Goos and Wettlaufer, 2010; Watanabe and Wake, 2009). Representative experimental data for the liquid saturation ($1 - S_i$) as a function of the depression beneath the bulk melting temperature, or undercooling ($T_0 - T_{eq}$), is shown for several sediments in Fig. 2A (taken from Horiguchi and Miller, 1983; Watanabe and Osada, 2016). In mathematical models that describe freeze–thaw processes, the change in liquid saturation with undercooling is often approximated by a simple power law (i.e. a straight line on a log–log plot), though it should be recognized that such parameterizations fit the data well over a limited saturation range and can deviate considerably from measured values both at higher and lower saturation levels. Moreover, differences in the pore-scale partitioning between ice and liquid at the same undercooling depending on whether the sample

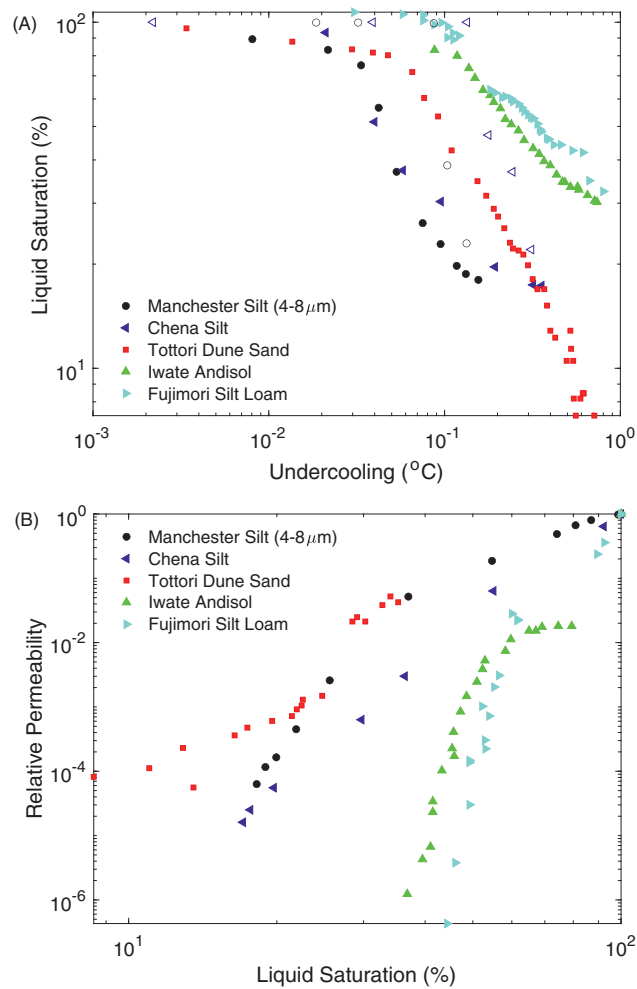


Fig. 2 Constitutive data for different sediments saturated with a mixture of liquid water and ice. (A) Measured unfrozen liquid saturation as a function of undercooling $T_0 - T_{eq}$. (B) Relative permeability as a function of liquid saturation. Data for sieved Manchester Silt (4–8 μm) and Chena Silt taken from Horiguchi and Miller (1983), data for Tottori Dune Sand, Iwate Andisol, and Fujimori Silt Loam taken from Watanabe and Osada (2016, all measurements shown here were made at the same Darcy velocity of 0.53 m/day). In (A) closed symbols were measured along a melting pathway, whereas open symbols for the Manchester Silt and Chena Silt represent measurements made as the temperature was reduced, showing considerable hysteresis. In (B) all measurements were taken along a melting pathway.

achieved that thermal state during warming (thawing) or during cooling (freezing) can produce considerable hysteresis in measured saturation levels, as shown for Manchester Silt and Chena Silt in Fig. 2A (Horiguchi and Miller, 1983). Because solid ice cannot normally be superheated whereas liquid water is readily supercooled, measurements made along a warming pathway are expected to better represent a sequence of equilibrium states.

Liquid water tends to preferentially wet both ice surfaces and pore walls, causing the micro-scale liquid configuration to resemble that which characterizes vapor–liquid equilibrium in the vadose zone (see Fig. 1C). At warmer sub-zero temperatures, the first ice crystals to reach equilibrium with liquid water occupy the centers of the largest pores (illustrated by the dashed circle in Fig. 1C), which enable the lowest possible values of interfacial curvature $\mathcal{K}$. For example, when the characteristic radius of the largest pores is r_p , recognizing that $\mathcal{K} \approx 2/r_p$ for the largest spherical ice crystals that fit in the pore space, Eq. (1) indicates that these crystals first reach equilibrium at temperature $T_f \approx T_0 [1 - 2\gamma_{il}/(\rho_i \mathcal{L} r_p)]$, which is approximately -0.005°C when $r_p = 10^{-5}$ m and -0.05°C when $r_p = 10^{-6}$ m. The implication is that fine-grained sediments can remain fully liquid-saturated at temperatures that are cold enough for ice to be in equilibrium with residual liquid in coarse-grained sediments. For example, when a soil structure is characterized by larger macropores that separate fine-grained aggregates, residual liquid can remain within the micron-scale pores of the aggregates long after the liquid content of the macropores has frozen almost entirely. Cooling further facilitates the progressive invasion of ice into smaller pores and the ice–liquid interface is distorted into the crevices between adjacent sediment grains (i.e. the larger liquid pockets in Fig. 1C), with the surface-energy penalty associated with increased interfacial curvature compensated by the free energy reduction that accompanies crystallization of residual liquid. The liquid films, which coat pore walls, thin as temperature is reduced, enabling P_T to increase, yet the liquid stores in high-curvature regions diminishes more quickly so that the residual fraction of liquid held in films dominates volumetrically at very cold temperatures. At warmer sub-zero temperatures, curvature effects exert the dominant control on residual liquid volume, enabling the liquid saturation to be determined with reasonable accuracy from the pore geometry and surface curvature considerations alone, without precise knowledge of the functional dependence of $P_T(d)$ (e.g. Chen et al., 2020). It is in this curvature-dominated regime that changes in ice saturation with temperature are most dramatic, enabling latent heat effects to have their greatest contributions to the local energy balance and causing temperature gradients to be diminished. Broadly speaking, sediments that are less well sorted are characterized by a correspondingly larger range of pore sizes that enables the liquid saturation to change more gradually with temperature than in more well-sorted sediments. Interestingly, although the change in liquid saturation level with undercooling in ice-infiltrated sediments is broadly comparable to the change in liquid saturation level with matric potential for the same soil placed within the vadose zone, several studies have highlighted important quantitative differences between freezing–thawing curves and drying–wetting curves that are not fully understood, but may be at least partially sourced to differing degrees of fine-scale particle rearrangements (e.g. Watanabe and Osada, 2016; Ren and Vanapalli, 2019).

The geometrical controls on the temperature-dependent ice saturation $S_i(T_0 - T_{eq})$ also govern changes in the flow paths that determine how the permeability k decreases with increased ice content. It is technically challenging to measure permeability changes with ice saturation, but several reliable datasets have been reported (e.g. Horiguchi and Miller, 1983; Watanabe and Osada, 2016), and successful strategies have been developed to enable numerical predictions for k , particularly in the regime for which S_i is dominated by curvature effects (e.g. Watanabe and Flury, 2008; Chen et al., 2021). Fig. 2B shows measurements of the relative permeability scaled by its liquid-saturated value as a function of liquid saturation level in the same five sediments represented in Fig. 2A. The relationship between relative permeability and liquid saturation is often approximated by a power-law in mathematical models of freeze–thaw processes. Potential liquid transport pathways with the largest cross-sectional areas tend to become ice-clogged first, when the liquid saturation level is still relatively high; because these broader flow paths represent the lowest resistance to water transport under liquid-saturated conditions, their blockage by ice can cause the relative permeability to decrease dramatically with small changes in the liquid saturation level. Considering how changes take place with temperature, significant ice volume fractions can develop and impede residual liquid transport through the pore space of coarse-grained sediments under warmer conditions with S_i increasing dramatically over narrower temperature ranges than in fine-grained sediments. However, despite the relative abundance of liquid water at cold temperatures in the most fine-grained, clay-rich sediments, their low saturated permeabilities limit water transport and hence the ability of freeze–thaw cycles to affect change. As noted above, the dramatic heterogeneities in pore size that often characterize healthy soil structure, for example with clay and organic-rich aggregates separated by macropores, can lead to the development of isolated fine-grained, liquid-saturated pockets embedded in a frozen ice matrix that effectively prohibits the large-scale material transport that is enhanced in these structures under unfrozen or partially vapor-saturated conditions. Local transport from fine grained regions toward macropores can nevertheless take place, accompanied by compaction in regions from which liquid is withdrawn and by the expansion and fracture propagation elsewhere that is needed to accommodate ice growth.

External loads

Net ice–sediment thermomolecular pressures P_T increase by approximately $1.1 \text{ MPa } ^\circ\text{C}^{-1}$ as the temperature is depressed below the normal bulk melting point (i.e. 0°C when effects of solutes are negligible, colder and concentration dependent when they are not). Roads, pipelines, buildings and other common infrastructure elements can be considered as external loads that must be supported by a combination of liquid pressures and sediment-contact stresses under unfrozen conditions, whereas thermomolecular pressures contribute as well when ice is present in the pore space. Segregated ice growth in through-going lenses that span many pore

dimensions (shown in Fig. 1A and B) is particularly important for unloading sediment contacts, requiring that thermomolecular pressures be sufficient to both balance external loads entirely and to support the reduced liquid potentials that are needed to draw additional supply. Mathematical descriptions of segregated ice development invoke two viable mechanisms, either with the gradual unloading of sediment contacts progressing as a consequence of load transfer via connected pore-spanning ice surfaces in zones referred to as frozen fringes (e.g. O'Neill and Miller, 1985; Rempel et al., 2004), or through ice-supported fracture propagation processes emanating from anomalously large pores in heterogeneous media (e.g. Style et al., 2011). When external loads are distributed unevenly, ice accumulation can be mechanically inhibited where larger stresses must be generated to produce displacement, favoring ice growth instead in locations where volumes are expanded more easily. Stones, fence posts, and other objects that penetrate conditions spanning from frozen to thawed are subject to more rapid temperature changes than can take place in adjacent water-logged soils from which latent heat must be removed as ice forms. This can distort isotherms, thereby affecting the timing, rates and consequences of ice growth and sometimes enabling frost action to produce self-organized patterns (e.g. Kessler and Werner, 2003; Peterson and Krantz, 2003).

Solutes

The colligative effects of solutes depress the local melting temperature, enabling liquid to persist to colder temperatures and suppressing ice growth. Effectively, this can be framed as making the reference temperature T_0 in Eq. (1) a function of the solute concentration c . When ice does form, solute molecules are rejected, enhancing their concentration in the residual liquid. Sediment surface chemistry can be affected, causing changes in cohesion and also altering the temperature dependence of melt film thickness. Even starting from relatively dilute liquid solutions in unfrozen conditions, the dramatic reductions in liquid volume during freezing can concentrate solutes enough to affect microbial ecology and cause damage to plants (e.g. Henry, 2007). Thawing liberates abundant quantities of fresh water that cause dilution with appreciable consequences as well.

Climatic dependence of freeze–thaw behavior

Fig. 3 shows the average annual number of days with freezing conditions at the ground surface (Kim et al., 2017). The impacts of freeze–thaw cycles depend on the severity, frequency and duration of freezing conditions. In many temperate regions, for example, winter ground surface temperatures are often frozen at night and thawed during the subsequent day, so that many tens of freeze–thaw cycles accumulate each year (i.e. most of the red and some of the orange regions in Fig. 3). Under favorable conditions, where near-surface water contents are high and soils are sufficiently fine-grained for curvature effects to prevent immediate pore ice formation, yet sufficiently coarse-grained to facilitate permeable liquid transport or situated on hillslopes that act as seepage faces, high night-time heat fluxes that are often dominated by radiation to clear and cold night skies support the growth of needle ice at the ground surface (also known as pipkrake, and as frost flowers or false fungus when hosted on dead wood), while the associated latent heat release prevents freezing conditions from penetrating any deeper. The diurnal needle-ice cycle can loosen a thin veneer of surficial sediments, in part aided by the temporary concentration of excess liquid that is liberated during daily thaw periods, but no substantial external loads can be displaced. In more coarse-grained, or less fully saturated soils under the same environmental

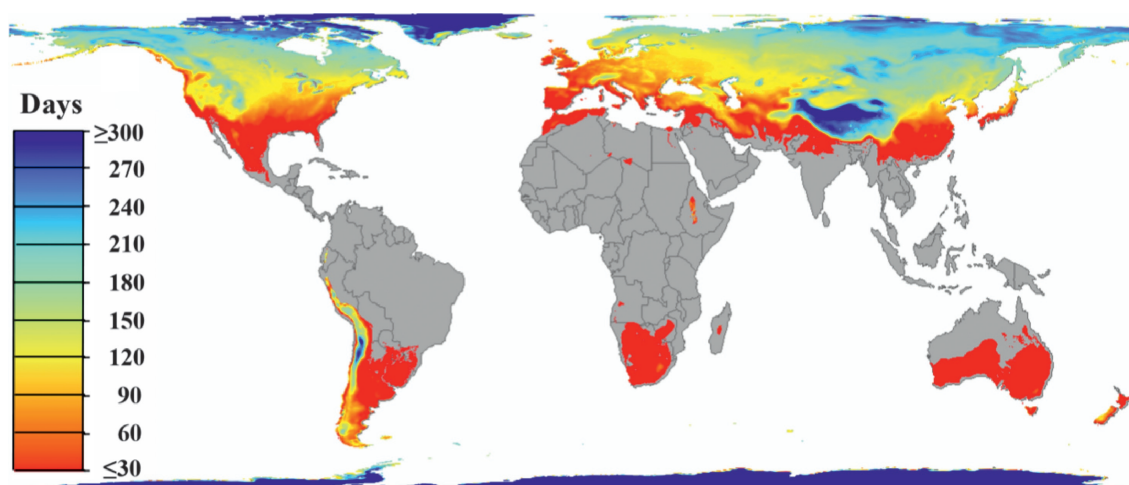


Fig. 3 Average annual number of days with freezing conditions at the ground surface. Note that this map includes glaciated regions, some of which are unfrozen at their beds. Data obtained using satellite microwave brightness temperatures over the period from 1979 to 2014 by Kim et al. (2017).

forcing, frozen conditions can penetrate more deeply (i.e. several centimeters), allowing ice to effectively bind particles together and increase cohesive strength, but the effects are ephemeral and rarely involve substantial water redistribution.

More dramatic freeze–thaw impacts take place when daily mean temperatures are sufficiently low for frozen conditions to penetrate further (i.e. decimeters, as shown in Fig. 1B) below the ground surface and be maintained for several days, weeks, or months (i.e. most of the orange and yellow and some of the lighter green regions in Fig. 3). Diurnal cycles of warmer and colder surface temperatures, notably once again modulated strongly by site-specific radiative forcing, as well as factors affecting convective heat transfer (i.e. wind) and the presence or absence of insulating materials (i.e. snow), can produce oscillations in the heat flux direction and thereby drive liquid transport cycles that supply segregated ice growth, with the potential for significant disruption to infrastructure (e.g. frost heave, frost cracking) and modification of soil structure. This can include both expansion and partial rejuvenation of compacted soils by ice-induced fracture processes, and compaction of fine-grained aggregates as a result of water transport toward ice-bound macropores. The largest ice concentrations are supplied by the most active liquid transport, occurring where temperatures are just sufficiently cold for thermomolecular pressures to support external loads and displace pore walls to accommodate segregated ice growth, while remaining sufficiently warm that the relative permeability is not too reduced, with steeper temperature gradients that accompany diurnal temperature swings of higher amplitude favoring more rapid liquid supply both by generating higher liquid potential gradients and by reducing transport distances from unfrozen liquid reservoirs. With spring thaw, or the onset of “mud season”, previously ice-bound soils lose their integrity while underlying drainage pathways remain frozen and effectively impermeable, often causing further infrastructure damage (e.g. potholes, pipe failure) and soil structure modification (e.g. Ferrick and Gatto, 2005).

In colder climates still, with extended winter conditions that keep surface temperatures depressed well below 0 °C, frozen conditions prevail through most of the year and freeze–thaw cycles can be greatly diminished both in number and effect (i.e. most of the green and some blue regions in Fig. 3). Changes in the direction of heat flow that are driven by surface warming and cooling do produce gradients in liquid potential that enable small changes in ice distribution, but dramatic reductions in relative permeability at cold ambient temperatures limit their magnitude so that winter disruption to infrastructure is often minor and the most notable mechanical effect is the dramatic increase to near-surface cohesive strength (e.g. Qi et al., 2006). Strong, dry winds over cold, bare soils comprised of frozen aggregates and air-filled macropores can encourage significant vapor transport that lowers total near-surface water content. Under the warmer conditions encountered during shoulder seasons, freeze–thaw cycles accumulate and are often accompanied by active near-surface segregated ice growth with attendant deformation and structure modification. Persistent positive surface temperatures that melt near-surface ice accumulations over deeply frozen soils can lead to appreciable seasonal surface water-logging that renders formerly sound wintertime travel routes impassable.

Climate change

Climate change is altering freeze–thaw patterns, producing enhanced effects in some locations while diminishing effects in others. In many cases, ongoing changes to seasonal temperatures may have a relatively straightforward impact on freeze–thaw cycles, typically causing them to decrease in number (e.g. in parts of North America by as many as one cycle fewer per year on average over each of the past 40 years, Kim et al., 2017, see Fig. 4) and often reducing the maximum depth of frost penetration. As a result, some regions with milder climates are becoming entirely frost-free, regions with more severe climates are experiencing less disruption associated with freeze–thaw behavior, and the most frigid locations, many of which are among the most rapidly warming places on Earth, are becoming less deeply frozen and as a result may become subject to enhanced freeze–thaw effects. Simple projections based solely on changes in mean temperature are clearly inadequate, however, as they miss potential local impacts of climate-driven changes in (i) seasonally dependent amplitudes of diurnal surface temperature forcing, (ii) the quantity, seasonality and phase (i.e. snow vs. rain) of precipitation, (iii) the persistence and timing of cloud cover and its effects on radiative forcing, (iv) the type, development, and density of vegetation, (v) the dynamics of microbial communities and nutrient cycling, and (vi) land use and infrastructure development. Modifications to these and other factors that affect the dominant controls on freeze–thaw behavior can both compound and mitigate the effects of baseline warming.

Mitigation strategies

Common strategies aimed at decreasing the damaging effects of freeze–thaw cycles most often take aim at the thermal and hydrological controls, and also sometimes involve more dramatic interventions such as the replacement of fine-grained soils with coarser, less frost-susceptible alternatives. Encouraging insulation by measures (e.g. involving crop residue management) that result in thicker snow-cover can be effective at dampening surface temperature variations and so limiting the potential gradients that drive liquid transport, but snow also provides an abundant water source that can be mobilized to supply segregated ice growth that restricts spring drainage and exacerbates soil erosion. Infrastructure investments can be protected by measures to reduce near-surface moisture content and also to encourage in situ pore freezing, as is more common in coarser materials, rather than segregated ice growth habits. The simple strategy of extending excavations below the frost penetration depth is extremely effective for protecting structures in moderately cold climates, whereas employing refrigeration or elevating structures on pilings to protect infrastructure by ensuring that frozen ground doesn’t thaw can be successful strategies in more severe environments. Tillage practices

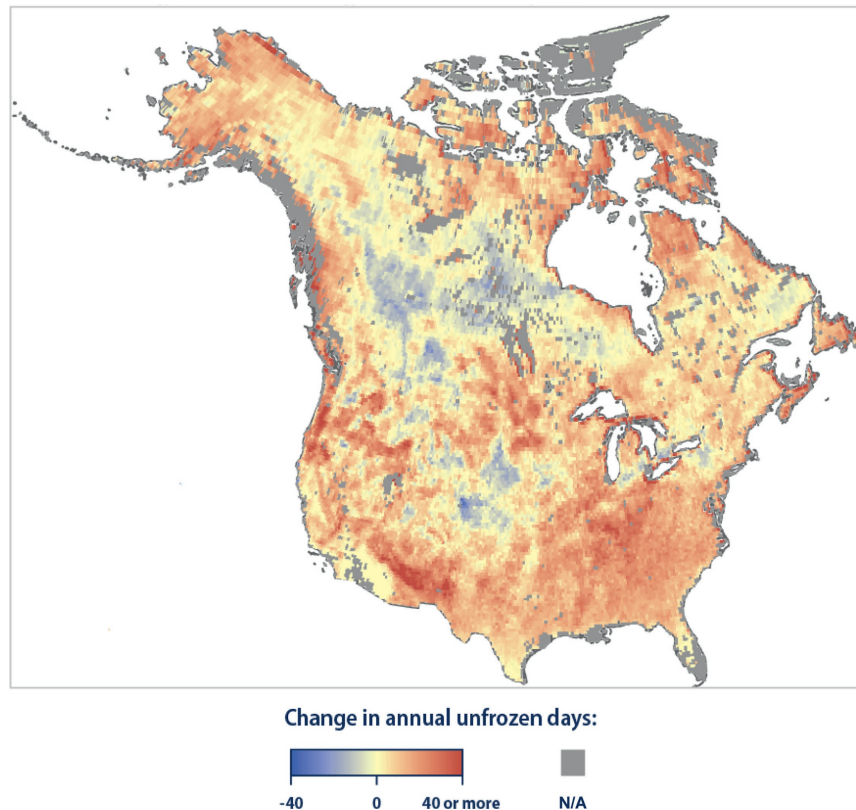


Fig. 4 Change in the annual number of frozen days in North America over the 40-year period from 1979 to 2019. Dark gray areas represent permanently frozen surfaces like glaciers and land that freezes only very rarely. Data derives from a 2020 update by Kim and Kimball from data originally published by Kim et al. (2017), made available by the United States Environmental Protection Agency as one of the Climate Change Indicators at epa.gov/climate-indicators.

that loosen and rejuvenate soil structure can also enhance fall drainage and lower the effective thermal conductivity further by mixing in crop residues so that frost penetration depths are reduced, which can be particularly effective at limiting freeze–thaw disturbance when subzero temperatures are encountered only above the water table. Where snow cover is minimal, the influence of albedo on radiative heat transfer can be employed to encourage earlier spring thaw by darkening the ground surface, for example with autumn residue burning.

Outlook

The basic processes involved in freeze–thaw cycling are well-understood, yet their combinations and interactions can be complex and remain challenging to describe in predictive quantitative simulations. Uncertainties in soil constitutive behavior are confounded by heterogeneity, with constraints generally much better for soil freezing curves than for relative permeability treatments, which are inherently much more difficult to validate. The evolution of geotechnical properties and soil structure with repeated freeze–thaw cycling requires further study. Most model treatments of freeze–thaw behavior are idealized to consider fully water and ice saturated porous media, despite the importance and ubiquity of partial air-saturation, particularly in agricultural applications. Efforts that account for solute redistribution, including the effects of changing pore-water composition on soil freezing curves and relative permeability functions, are also poorly represented in the literature. Despite considerable progress in unravelling the mechanics of segregated ice growth, the relative importance of ice-lens nucleation by sediment unloading across discrete temperature horizons in comparison with ice-lens growth by fracture propagation initiated within anomalously large pores remains unresolved. Further attention is also needed to formulate predictive descriptions of gradual ice redistribution within fully ice-bound soils, and to account for poroelastic effects, particularly in regard to the modification of soil structure. Concern over the effects of global climate change and the need to protect infrastructure investments have motivated renewed interest in these and related topics, some of which will undoubtedly be addressed more comprehensively in the coming years.

References

- Chen J, Mei S, Irizarry JT, and Rempel AW (2020) A Monte Carlo approach to approximating the effects of pore geometry on the phase behavior of soil freezing. *Journal of Advances in Modeling Earth Systems* 12(10). e2020MS002117.
- Chen J, Mei S, and Rempel AW (2021) Estimating permeability of partially frozen soil using floating random walks. *Water Resources Research* 57(12). e2021WR030598.
- Daigle H (2021) Structure of the electrical double layer at the ice–water interface. *The Journal of Chemical Physics* 154(21), 214703.
- Dash JG, Rempel AW, and Wettlaufer JS (2006) The physics of premelted ice and its geophysical consequences. *Reviews of Modern Physics* 78(3): 695.
- Ferrick M and Gatto LW (2005) Quantifying the effect of a freeze–thaw cycle on soil erosion: Laboratory experiments, Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research. *Group* 30(10): 1305–1326.
- Gibbs JW (1878) On the equilibrium of heterogeneous substances (part ii). *Transactions of the Connecticut Academy of Arts and Sciences* 3: 343–524.
- Hansen-Goos H and Wettlaufer J (2010) Theory of ice premelting in porous media. *Physical Review E* 81(3), 031604.
- Henry HA (2007) Soil freeze–thaw cycle experiments: Trends, methodological weaknesses and suggested improvements. *Soil Biology and Biochemistry* 39(5): 977–986.
- Horiguchi K and Miller RD (1983) Hydraulic conductivity functions of frozen materials. In: *Proceedings of 4th International Permafrost Conference* 504–508.
- Kessler M and Werner B (2003) Self-organization of sorted patterned ground. *Science* 299(5605): 380–383.
- Kim Y, Kimball JS, Glassy J, and Du J (2017) An extended global earth system data record on daily landscape freeze–thaw status determined from satellite passive microwave remote sensing. *Earth System Science Data* 9(1): 133–147.
- O'Neill K and Miller RD (1985) Exploration of a rigid ice model of frost heave. *Water Resources Research* 21(3): 281–296.
- Peterson RA and Krantz WB (2003) A mechanism for differential frost heave and its implications for patterned-ground formation. *Journal of Glaciology* 49(164): 69–80.
- Qi J, Vermeer PA, and Cheng G (2006) A review of the influence of freeze–thaw cycles on soil geotechnical properties. *Permafrost and Periglacial Processes* 17(3): 245–252.
- Rempel AW, Wettlaufer J, and Worster MG (2004) Premelting dynamics in a continuum model of frost heave. *Journal of Fluid Mechanics* 498: 227–244.
- Rempel AW, Marshall JA, and Roering JJ (2016) Modeling relative frost weathering rates at geomorphic scales. *Earth and Planetary Science Letters* 453: 87–95.
- Ren J and Vanapalli SK (2019) Comparison of soil-freezing and soil-water characteristic curves of two Canadian soils. *Vadose Zone Journal* 18(1): 1–14.
- Santamouris M, Mihalakakou G, Psiloglou B, Eftaxias G, and Asimakopoulos D (1999) Modeling the global solar radiation on the Earth's surface using atmospheric deterministic and intelligent data-driven techniques. *Journal of Climate* 12(10): 3105–3116.
- Sharratt B (1993) Freeze–thaw and winter temperature of agricultural soils in interior Alaska. *Cold Regions Science and Technology* 22(1): 105–111.
- Sharratt B, Benoit G, and Voorhees W (1998) Winter soil microclimate altered by corn residue management in the northern Corn Belt of the USA. *Soil and Tillage Research* 49(3): 243–248.
- Style RW, Peppin SS, Cocks AC, and Wettlaufer JS (2011) Ice-lens formation and geometrical supercooling in soils and other colloidal materials. *Physical Review E* 84(4), 041402.
- Walder JS and Hallet B (1986) The physical basis of frost weathering: Toward a more fundamental and unified perspective. *Arctic and Alpine Research* 18(1): 27–32.
- Watanabe K and Flury M (2008) Capillary bundle model of hydraulic conductivity for frozen soil. *Water Resources Research* 44(12).
- Watanabe K and Osada Y (2016) Comparison of hydraulic conductivity in frozen saturated and unfrozen unsaturated soils. *Vadose Zone Journal* 15(5).
- Watanabe K and Wake T (2009) Measurement of unfrozen water content and relative permittivity of frozen unsaturated soil using NMR and TDR. *Cold Regions Science and Technology* 59(1): 34–41.

Arctic permafrost

Julia Boike^{a,b}, Andrey Abramov^c, Katrina E Bennett^d, and Lars Kutzbach^e, ^aAlfred Wegener Institute Helmholtz Center for Polar and Marine Research (AWI), Potsdam, Germany; ^bDepartment of Geography, Humboldt University Berlin, Berlin, Germany; ^cInstitute of Physicochemical and Biological Problems in Soil Science, Pushchino, Russia; ^dEarth and Environmental Sciences, Los Alamos National Laboratory, Los Alamos, NM, United States; ^eInstitute of Soil Science, Center for Earth System Research and Sustainability (CEN), Universität Hamburg, Hamburg, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	410
Characteristics of permafrost-affected soils	411
Distribution of Northern Hemisphere Permafrost	411
Physical characteristics of permafrost-affected soils	411
Cryo-pedogenetic processes	414
Hydrology of permafrost-affected soils	414
Energy and water balances	414
Snow-vegetation feedbacks	415
Coarse-scale feedbacks on changing permafrost-affected soils to Arctic hydrology	415
Gas transport	415
Subsidence, degradation, irreversible processes, and the fate of permafrost	416
Summary	416
References	417

Abstract

Permafrost is the term used to describe ground (soil or rock) that remains at or below 0 °C throughout the year for at least two consecutive years. In this chapter, we focus on distribution and physical characteristics of Arctic terrestrial permafrost. We discuss the constituents of permafrost (minerals, organic matter, water, ice, and gas) as well as the presence and importance of ice, leading to patterned ground formation. We explain cryo-pedogenetic processes, hydrology, energy, and water balances, snow-vegetation feedbacks, and gas transport in permafrost regions. Finally, we discuss permafrost degradation and its effects on the energy, water, and carbon balances of the Arctic.

Key Points

- Permafrost is spatially distributed across the globe and exerts important controls on land surface energy, water, and carbon balances and the global climate system.
- Permafrost underlies 11% of the total surface of the Earth's exposed land surface.
- Permafrost strongly affects soil genesis due to a unique set of cryo-pedogenetic processes.
- The hydrology of permafrost-affected soils is complex due to seasonal freezing and thawing of the active layer.
- Snow cover and vegetation exert important controls on energy and water balances of permafrost-affected soils and fluxes.
- Thawing of ground-ice-rich permafrost leads to subsidence, with often irreversible changes of landscape topography, hydrology, and carbon cycling.
- The warming Arctic climate is expected to increase the production and release of CO₂ and CH₄ in permafrost-affected soils.

Introduction

Permafrost is an important component of the Earth's cryosphere. It plays a key role in the global carbon cycle, as well as in the ecosystems and infrastructure in Arctic and sub-Arctic climate zones. Permafrost-affected soils have unique energy, water, and carbon dynamics due to periodic freeze-thaw cycles and the underlying continuously frozen ground. Borehole measurements show that Arctic permafrost has been warming during the early 21st century (Biskaborn et al., 2019). Earth system model simulations demonstrate that the carbon release from degrading permafrost due to climate warming represents an important global feedback on climate change, which affects the plausibility of achieving the temperature targets of the United Nations Framework Convention on Climate Change (de Vrese and Brovkin, 2021).

By definition, permafrost encompasses ground (soil, sediments, rocks) that remains at or below 0 °C for at least two consecutive years (French, 2017; van Everdingen, 2005). The permafrost table is the upper boundary of the permafrost, and the ground above the permafrost table is called the suprapermafrost layer. Permafrost layers may be just meters deep in the southernmost permafrost zones, or it may be more than a kilometer deep further north. Permafrost can contain layers of solid ice throughout, which can be up

to several meters thick. The deepest permafrost is located in northern Yakutia where it reaches 1500 m (Melnikov, 1966). Deeper down in the soil is a perennially unfrozen layer where the geothermal heat flux prevents temperatures from falling below the freezing point.

We focus on processes of Arctic terrestrial permafrost, thus excluding formerly terrestrial, now flooded permafrost (submarine permafrost) which underlies large proportions of the Arctic shelf areas. We define “Arctic” to be the region of the common base map provided by the definition of vegetation and associated characteristics (Circumpolar Arctic Vegetation Map, CAVM Team, 2003).

Permafrost layers may not actually be frozen (part or all of the pore water turned into ice), because the freezing point of included water can be depressed several degrees below 0 °C. Water, in either liquid or solid phase, may or may not be present in permafrost. To differentiate between the temperature and water phase of permafrost, the terms “cryotic”/“non-cryotic” and “frozen”/“unfrozen” have been proposed. The former terms refer solely to the temperature of the material, whereas the latter terms refer to the presence of ice in the ground (van Everdingen, 2005; French, 2017). Therefore, perennially cryotic ground is synonymous with permafrost, and permafrost may be “unfrozen,” “partially frozen” or “frozen,” depending upon the state of the ice and water content. The active layer is seasonally frozen; the top layer of the ground freezes in the winter and thaws during the summer. Permafrost-affected soils are closely related to the active layer and upper sections of permafrost.

The thermal and hydraulic dynamics of permafrost-affected soils control the fluxes of energy, water, and carbon between the land surface and the atmosphere. This is of particular interest with regards to current climate warming, which is already pronounced in the Arctic and is anticipated to increase dramatically in the coming decades (Meredith et al., 2019). The warming of the climate leads to degradation of the permafrost and its landforms (Karjalainen et al., 2020), thereby mobilizing large quantities of organic carbon in thawing permafrost soils (Nitzbon et al., 2020). Decomposition and mineralization of organic carbon are expected to contribute to a further increase of atmospheric CO₂ concentration, fueling a positive feedback mechanism to the Earth’s climate (Previdi et al., 2021). Depending on the development of Arctic hydrology, increases in methane production and emission from degraded, wet Arctic soils can lead to additional increases of atmospheric radiative forcing. Permafrost-affected soils can be free of vegetation or covered by treeless tundra vegetation or coniferous taiga forests. While some tree species in the taiga biome can grow on permafrost-affected soils, tree growth is generally prevented by cold summers and short growing seasons in the tundra biome. Changes in vegetation, such as movements of shrub- and tree-lines, affect energy, water, and carbon fluxes and modulate the response to climate change of permafrost-affected soils.

Characteristics of permafrost-affected soils

Distribution of Northern Hemisphere Permafrost

Permafrost underlies an area of $14\text{--}16 \times 10^6 \text{ km}^2$ —around 15% of the exposed Northern Hemisphere or 11% of the total Earth’s exposed land surface (Obu, 2021). Permafrost zones are defined as continuous (90–100% of area), discontinuous (50–90%), sporadic (10–50%), and isolated (10% or less). The mean annual ground temperature (MAGT, °C) at the top of the permafrost is shown in Fig. 1; these model results were validated against ground observations of temperature from borehole measurements (Obu et al., 2019). MAGT ranges from approximately 0 °C to –25 °C in the permafrost. The 0 °C isotherm shows great variation latitudinally, reaching much further south in Siberia compared with North America, for example. Sometimes, permafrost layers are found in areas where they cannot form under present climatic conditions, which reflects past conditions when the climate was colder. These deposits are called “relict permafrost” and are often covered by dozens of meters of unfrozen soil.

Physical characteristics of permafrost-affected soils

The material forming permafrost can be bedrock, sediment, or soil, consisting of mineral and organic solid compounds, ice, liquid water, and gas. Ice existing in the pores of frozen soil is named pore ice. When the absolute ice content exceeds the pore space (porosity) under unfrozen conditions, it is referred to as excess ice. Fig. 2 gives an example of the ice content of a frozen permafrost core section from the Yedoma complex of Pleistocene origin (Lena River Delta, Siberia). The laboratory and computer tomography results show a high ice content, comprising a large volumetric fraction (between 39% and 48%) of excess ice as well as a complex ice distribution (Fig. 2) (Nitzbon et al., 2021, 2022). The ice content and soil texture of permafrost in the Arctic are largely unknown and depend on factors such as history (duration of formation), soil type, soil texture, soil water chemistry, and temperature (Fischer et al., 2020). Thermal and hydrological processes are highly sensitive to the presence of ice and water in the ground. Simulation of permafrost requires implementation of these physical processes, as well as the parameterization of the thermal and hydrologic properties.

In permafrost areas, downward thawing from the surface in the warm season forms the active layer, several centimeters to a couple of meters deep, depending on climatic conditions and soil properties. The annual temperature range is the highest at the soil surface, ranging approximately between –30 °C and +30 °C, and is dampened with depth. In ground with changing volumetric liquid water content due to freezing and thawing processes, the internal energy can change through sensible heat transfer, expressed as a change in temperature, as well as through latent heat transfer, expressed in a change in liquid water or ice content. The freezing and thawing of wet soil is affected by latent heat release due to phase change, which slows the soil warming and cooling.

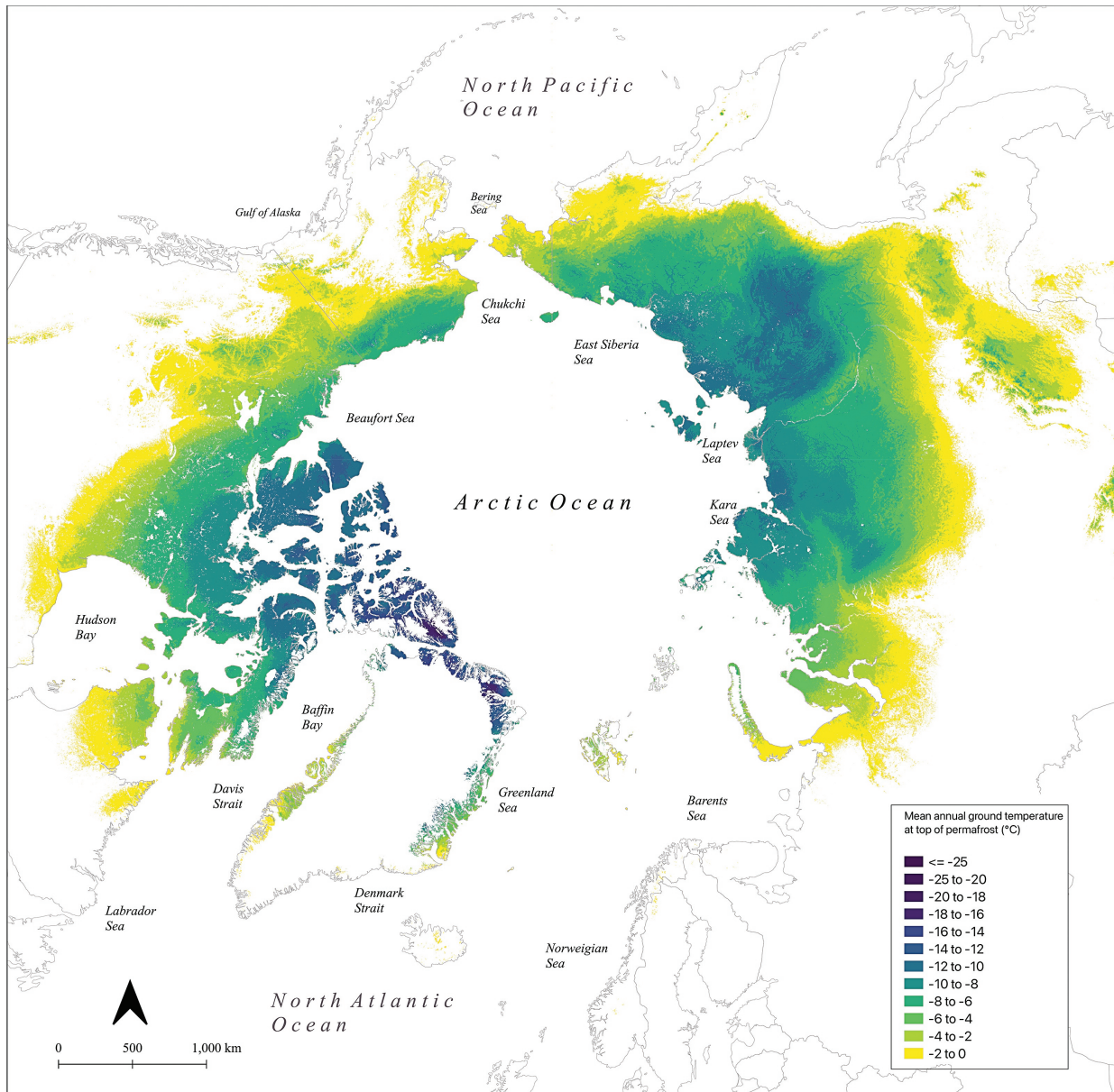


Fig. 1 The mean annual ground temperature (MAGT, °C) as the modeled mean annual ground temperatures at the top of the permafrost for the Northern Hemisphere at 1 km spatial resolution. The 0°C isotherm is indicated by the boundary between colored and noncolored areas. Adapted from Obu J, Westermann S, Bartsch A et al. (2019) Northern hemisphere permafrost map based on TTOP modelling for 2000–2016 at 1 km² scale. *Earth-Science Reviews* **193**, 299–316. doi: 10.1594/PANGAEA.888600.

Soil thermal properties, i.e., volumetric heat capacity and thermal conductivity, depend on the constituents of the soil, i.e., the volumetric fractions of mineral, organic, liquid water, ice, and air, and their distribution (Fig. 2). The composition of the solid soil constituents changes with depth as determined by the soil stratigraphy. The contents of liquid water and air are variable with both depth and in time due to both thermal and hydrological processes. While the volumetric heat capacity is obtained as the weighted sum of all constituents, the thermal conductivity cannot be calculated easily from the corresponding values of its constituents because their spatial arrangement has a dominant influence.

Permafrost-affected soils are often characterized by a surface organic layer that consists of mostly organic matter. This surface organic layer appears to be one of the most important factors controlling the thermal regime. The thickness of this organic layer can range from a few centimeters to several decimeters, depending on the rates of accumulation and decomposition of plant litter. Often, the soil surface is covered by moss layers. These organic surface layers have a large effect on the thermal and hydrologic processes of the active layer. This is due to the fact that the thermal diffusivity of organic matter is an order of magnitude

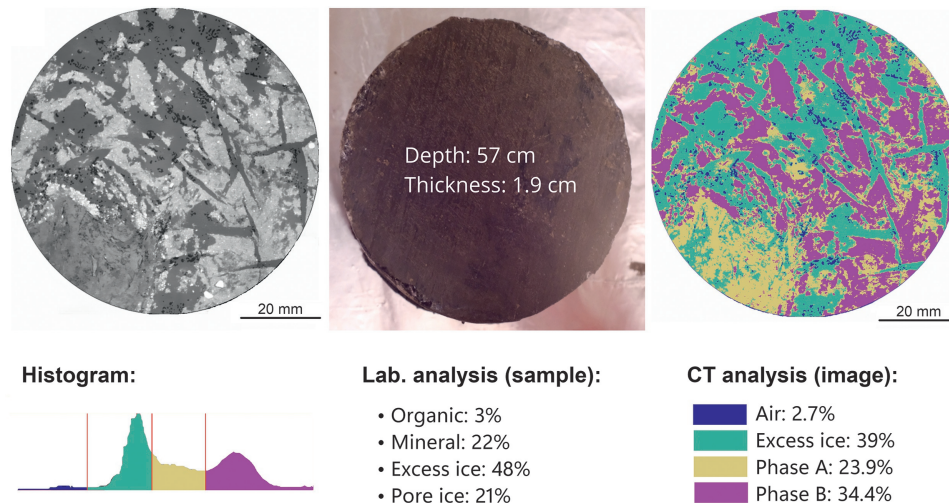


Fig. 2 Results of X-ray computed tomography image and laboratory analysis of the same permafrost core, drilled in 2017 in the soil of the Yedoma complex, Lena River Delta, Siberia. Left: Raw image of an X-ray computed tomography of one sample slice (0.005 cm) using computer tomography (1 cm slice = 200 images). The images obtained show different colors based on density differences of the material. Lighter grays indicate higher densities (for example, mineral grains in white), dark gray low densities (for example, gas in black), and light gray medium density (excess ice). Middle: Photograph of same permafrost core at 57 cm depth (below land surface). The composition (organic, mineral, excess, and pores) is the result of standard soil laboratory analysis (thickness of slice: 1.9 cm). Right: Classification of the imaged sample slice (left image) using manual thresholding segmentation of the histogram. This image-based classification cannot resolve the composition of the material but shows a complex spatial arrangement of all the constituents at high resolution. Varying constituents of pore ice and organic and mineral material are thus designated as mixed sediment phases A and B. Figure reproduced from Nitzbon J, Gadylyaev D, Schlüter S, Köhne JM, Grosse G and Boike J (2022) Brief communication: Unravelling the composition and microstructure of a permafrost core using X-ray computed tomography. *The Cryosphere* 16: 3507–3515. doi: 10.5194/tc-16-3507-2022. See further explanation and online video 3D material; Nitzbon J, Gadylyaev D Boike J (2021) *Material Composition in Permafrost Core Through Computer Tomographic Imaging*. Copernicus Publications. doi: 10.5446/56864; <https://av.tib.eu/media/56864>.

smaller than that of mineral soil constituents. In addition, when pores are filled with air, this decreases the soil thermal conductivity. The low thermal conductivity of the surface organic layer creates an effective insulator, decreasing the heat exchange between permafrost and the atmosphere, cooling the soil in the summer (Porada et al., 2016). On the other hand, when saturated, or especially when icy, the thermal conductivity is orders of magnitude higher. From a hydrological perspective, a higher organic matter content increases the soil porosity, affecting the flow and storage of water. The hydraulic conductivity (of peat) ranges across several orders of magnitude between dry and saturated conditions. Including a surface organic layer in earth system models has largely improved their representations of permafrost thermal dynamics (Chadburn et al., 2015).

Permafrost was often regarded as being impermeable to water flow (“aquitard”), but actually water does migrate in frozen soils under thermal and pressure gradients. In soils colder than 0 °C, a certain fraction of liquid water remains in the soil pores because of capillary and adsorption forces, surface effects, and an electrical double layer. As the temperature drops below 0 °C in a porous medium such as soil, the liquid water in the largest voids will freeze. As a consequence, the liquid-ice interface will move into pores with smaller radii, leaving some water in liquid form. For ice and liquid water in thermodynamic equilibrium, a characteristic relationship between soil temperature and liquid water content exists. This relationship depends on the total amount of water present in the material, on the texture of the material (i.e., fine vs. coarse), and on the purity of the water. Thus, freezing of ground and phase change of water occurs gradually over a certain temperature range below the freezing point, in which liquid water and ice coexist. This relationship of the amount of liquid water present at a given subfreezing temperature (soil freezing characteristic curve; SFCC) depends mainly on the share of different constituents in the soil and its texture. To describe the SFCC mathematically, a common assumption is that the freezing of soil can be described in the same way as the drying of soil, with the latter expressed as the soil water retention curve (SWRC). Both empirical formulations, SFCC and SWRC, are implemented in models for the calculation of hydraulic and thermal properties of seasonally freezing soils.

When permafrost degrades, or ground ice thaws, the ground surface will subside due to the volume reduction of ice to water or lateral export of melt water. The opposite process, heave, is the upward displacement of material due to freezing. Soil characteristics are important in terms of the amount of heave and subsidence, such as soil type, grain size distribution, and the availability of water.

Heave and subsidence of the ground surface can offer insight into processes of heat and mass transfer in freezing and thawing soils. The typical seasonal amplitudes are within the range of 20–120 mm (Gruber, 2020). Subsidence can be measured directly, but recent techniques have focused on measurement of permafrost-affected soil displacements from electrical resistivity measurements, differential interferometric synthetic aperture radar (Antonova et al., 2018), and remote sensing. The existence of ice-rich permafrost has substantial implications for thermal and hydrological processes, as well as permafrost stability. Melting of massive ground ice

creates topographic and hydrologic changes, with long-lasting effects and feedbacks on the landscapes and carbon turnover (Nitzbon et al., 2020).

Cryo-pedogenetic processes

The thermal regime of permafrost-affected soils and particularly the cycles of freeze-thaw lead to special cryogenic soil formation processes. These cryo-pedogenetic processes include ice segregation, frost heave and subsidence, up-freezing of coarse fragments, frost sorting, thermal contraction and expansion, and gelifluction. One important effect of these processes—on their own or in combination—is cryoturbation. This term refers to all differential soil movements induced by deep frost penetration and frost action (alternate freezing and thawing) (van Everdingen, 2005). An extensive description of the occurrence and processes of cryoturbated soils is provided by Ping et al., (2015).

Ice segregation is an important frost action process behind many cryo-pedogenetic processes and cryoturbation phenomena. The coexistence of solid and liquid phases of water in freezing soils allows water to flow to the freezing front, where it freezes and can form ice lenses (segregation ice or excess ice). Frost heave of soil occurs when soil water freezes due to the volume expansion during ice nucleation (about 9% increase) and is greatly enhanced where ice lens formation takes place. Ice segregation thus causes differential frost heave, leading to increased heterogeneity of microtopography and soil properties. Over longer time periods, this can also lead to the formation of complex patterned ground. Small-scale thermo-mechanical soil models have been successfully used to simulate the formation of nonsorted circles (Nicolisky et al., 2008). The formation of non-sorted circles depends primarily on water-logged conditions and small-scale vegetation heterogeneities. Furthermore, ice segregation drives movement of coarse fragments (frost-pull or frost-push hypotheses, French, 2017), which in turn can lead to frost sorting and patterned ground formation (Kessler and Werner, 2003). Intensity of ice segregation depends on the frost susceptibility of soil materials, which is a function of the soil's pore size distribution, and the rate of temperature drop during freezing.

Another important process leading to cryoturbation is thermal contraction of frozen ground during rapid cooling in winter. This process leads to the formation of vertical cracks several meters deep, often arranged in polygonal networks. When meltwater runs into these cracks and refreezes, several-millimeter-wide ice veins form. If this process is repeated over many years, massive ice wedges build up. In summer, as the frozen ground in the center of polygons warms and expands, the presence of new ice veins prevents it from returning to its former position. This creates strong compression forces that deform the soil and cause soil material to be thrust over the ice veins/wedges at the polygon rims, as described by Lachenbruch, 1962. The soils of the elevated polygon rims typically show cryoturbation, whereas the soils in the depressed centers do not. If ice bodies contained in permafrost-affected soils thaw, the collapse of previously overlying or adjacent soil materials leads to further forms of cryoturbation.

Cryo-pedogenic processes strongly influence the carbon turnover and sequestration of permafrost-affected soils (Ping et al., 2015). However, small-scale soil heterogeneity affected by cryoturbation is not yet represented in most earth system models, but can be shown to have effects on larger-scale energy, water, and carbon balances (Beer, 2016; Shu et al., 2020).

Hydrology of permafrost-affected soils

At fine spatial scales (i.e., meters), water flow in pathways into and through the active layer and along frozen soil layers can influence soil properties in a number of important ways. In the Arctic, there are unique features of permafrost-affected soils that have variable hydrologic and drainage properties. In particular, cryo-pedogenetic processes that form patterned ground have unique soil-hydrologic processes, such as the seasonal increase of water storage potential along with active layer thaw, which are not present outside the permafrost regions. Water moves through the polygonal tundra landscapes in distinctive ways; for example, water flow occurs along the ring-like shape of the polygon center near the polygon rim (Harp et al., 2021). Polygonal tundra geometries affect water movement, drainage timing and patterns, and the trajectory of formation of taliks, which are unfrozen “windows” through the permafrost. Landscape and climate factors influencing polygonal tundra environments and thermokarst pool development are discussed in more detail in Abolt et al. (2020). At scales beyond the soil column (i.e., roughly tens of meters), hillslope processes become dominant. Recent research findings for hydrology of one-dimensional soil column energy and water balances, albedo, and snow-vegetation interactions are discussed below.

Energy and water balances

Energy and water balances of Arctic permafrost-affected soils are highly seasonal. Water abundance plays a large role for the surface energy balance because of the latent heat flux (Q_E) that is driven by evaporation of water, either by direct evaporation from wet surfaces or by plant transpiration. When Arctic land surface and soil conditions are dry, the sensible heat flux (Q_H) and Q_E are similar, or Q_E might even be larger in arid locations with low relative humidity. In wet tundra environments with open water (e.g., lakes, wet Arctic plain environments) and locations with greater relative humidity, Q_E will be the largest in summer. Soil properties, such as thermal conductivity and porosity, have been found to be the main drivers of ground thaw in these environments. Finally, liquid water is warmer than the frozen soil and it transfers energy with it, leading to the development of conduits of flow, preferential pathways through which water, nutrients, and contaminants may move along and through the soil active layer.

Albedo is an important driver of energy balance in Arctic soils, largely because the amount of light reflected has a strong influence on radiation balances, with different effects for summer and winter. It is one of the primary feedback mechanisms of snow and vegetation in Arctic permafrost-affected soils. Albedo can vary significantly with respect to summer vegetation cover, resulting in more radiation either reflected or absorbed at the ground surface. In the winter, the white and bright snow cover reflects energy, leading to reduced warming of the surface. Snow albedo decreases have been linked to black carbon on snow in Arctic environments (Pu et al., 2021). Fires and changes in surface water extents can interact with albedo signatures during the snow season and result in lowered albedo in the boreal Arctic (Webb et al., 2021).

Snow-vegetation feedbacks

In Arctic environments, water in all its forms interacts closely with vegetation, resulting in impacts to the permafrost-affected soils (e.g., Grünberg et al., 2020). Vegetation, particularly shrubs, causes snow to pile up or accumulate along the margins and within shrub patches in Arctic environments, owing to a number of important factors including topography, vegetation composition (density, height, and presence or absence), and wind. Deeper snow protects the ground surface from penetrating cold temperatures, which can insulate the soil and lead to changes in permafrost condition, such as the development of taliks forming conduits for water, nutrient, and gas fluxes. Shallower snow, on the other hand, leads to reduced insulation of the soil and enhanced downward penetration of cold into the permafrost.

Both snow and vegetation are shifting in the Arctic under climate change impacts, with variable responses to permafrost-affected soils which are, in part, dependent on vegetation type and interactions with snow (Bennett et al., 2022). Shrubs are known to be increasing in the Arctic, resulting in changes to snow albedo in the autumn, leading to snow melt that can form a negative feedback, thereby driving or stabilizing permafrost development (Belke-Brea et al., 2020). Similarly, boreal forest canopies alter the surface energy balance of underlying permafrost by inhibiting more than 90% of the solar radiation and suppressing turbulent heat fluxes (Stünzi et al., 2021a). The loss of boreal forests can lead to an increase in the ground surface temperatures and an increase in the active layer thickness (Stünzi et al., 2021b). On the other hand, both shrub and boreal forests can lead to larger below-canopy snow accumulations and a delay of snow melt, which thermally protects permafrost from cold winter air temperatures (Stünzi et al., 2021a). Overall, reductions in snow cover extent across the Arctic can lead to early snow melt, resulting in shorter snow seasons. Recent findings on the importance of rain-driven permafrost thaw play a substantial role in deepening the active layer via both thermal heat transfer and changing soil thermal conductivity, mechanisms that have been overlooked in past research. Summer precipitation and fires can also interact to variably influence upland and lowland shrub development in the Arctic tundra environments (Chen et al., 2021), leading to enhanced permafrost thaw in disturbed and wetland areas.

Coarse-scale feedbacks on changing permafrost-affected soils to Arctic hydrology

At very coarse scales, the hydrology of Arctic permafrost-affected soils can impact global climate through the release of carbon dioxide, methane, and other greenhouse gasses owing to increased thawing of permafrost and deepening of active layers. Enhanced atmospheric moisture in the Arctic can cause increased runoff from thawing permafrost-affected soils, leading to increased runoff to the Arctic Ocean. However, at these coarse scales, it is difficult to understand impacts because the tools that we utilize to study them (i.e., earth system models) may not include the physical processes, such as thermodynamics of the permafrost, vegetation dynamics, biogeochemistry, and subgrid scale snow processes. Additionally, some of these processes are still not well understood, for example, the effects of warming-induced vegetation changes on soil organic carbon turnover. Dynamic vegetation model simulations by Zhang et al. (2013) showed that positive feedbacks on climate warming, for example, reduced snow-season albedo through tree-line advance and shrub expansion may dominate over negative feedbacks, such as increased carbon sequestration in plants and soils. Newer earth system models incorporate these dynamics for Arctic systems, which are in development, testing, and validation phases (<https://ngee-arctic.ornl.gov>; <https://arcticinterface.org>; <https://e3sm.org>).

Gas transport

Permafrost-affected soils and sediments are often rich in organic matter, and biogenic mineralization of this organic matter leads to the release of greenhouse gasses, CO₂ and CH₄, into the soil pore space. Microbial production of CO₂ and CH₄ increases with temperature and is greater in thawed than in frozen soils, yet production continues at low rates even in frozen soils at temperatures as low as −16 °C (Panikov and Dedysh, 2000; Rivkina et al., 2004). Soil gasses exist in the soil pore space either as free gas or dissolved in the thawed or frozen pore water solution. CH₄ can also occur in permafrost as clathrate hydrates. The theoretical stability depth of methane clathrate hydrate in permafrost regions is about 200 m; however, meta-stable methane hydrates have been found in ice-rich permafrost at considerably lower depths (Rivkina et al., 2001; Chuvilin et al., 2018; Dallimore and Collett, 1995).

Gasses are transported within the soils by diffusion or advection. Diffusion of gasses is more efficient in air than in liquid water (order of magnitude of diffusivity ratio: 10⁴); gas diffusion in ice is even less efficient than in liquid water (order of magnitude of diffusivity ratio: 10²). Advective gas transport can be driven by free gas flow, one important mechanism being the upward migration of gas bubbles (ebullition), or by the flow of pore water solution containing dissolved gasses. Large contents of ice as well as hydrates restrict both diffusive and advective gas migration. Therefore, gasses produced in situ in permafrost-affected soils by organic matter mineralization or transported from deeper gas reservoirs can accumulate below near-to-impermeable ice-bearing

permafrost. Thawing of ground ice can lead to sudden blowouts from gas-filled cavities, which can form pockmarks or even deep craters (Bogoyavlensky et al., 2021).

The active layer of permafrost-affected soils, which thaws in the warm season and allows plant growth, is most dynamic in terms of soil gas production, consumption, and transport. In addition to heterotrophic respiration, autotrophic respiration by roots of vascular plants releases large quantities of CO_2 into the soil pore space. Furthermore, root exudation enhances soil organic matter mineralization both to CO_2 under oxic conditions and to CH_4 and CO_2 under anoxic conditions. The active production of CO_2 and CH_4 in the active layer leads to a seasonal buildup of these gasses, but due to their small distance to the soil surface, and because the soil is unfrozen, gas export from the soil to the atmosphere is efficient. Due to the fast turbulent mixing of emitted gasses in the atmospheric boundary layer, concentration gradients from soil to the atmosphere are steep, fostering diffusive transport. Atmospheric turbulence can also create pressure pumping at the soil surface, leading to advective gas transport out of the soil. Furthermore, pressure pumping as well as shaking of plant structures by wind gusts can trigger the release of soil gas bubbles (Wille et al., 2008). During the refreezing period in the fall, cryostatic pressure buildup in the active layer can lead to squeezing-out and sudden outbursts of CH_4 and CO_2 (Mastepanov et al., 2008).

Vascular plants are essential for the transport of CH_4 from the soil to the atmosphere, in addition to diffusion and ebullition through the soil pore space. Many wetland plants have a unique feature called aerenchyma, which are air-filled tissues found in their leaves, culms, rhizomes, and roots. Aerenchyma allows efficient transport of oxygen from the leaves to the oxygen-deficient soil zones around the roots, while also promoting the transport of CH_4 from the rhizosphere to the atmosphere. This plant-mediated transport can be due merely to enhanced diffusion, but advective transport mechanisms have also been observed; these are driven by pressure gradients within the plant structures.

Arctic warming is expected to increase land-atmosphere fluxes of CO_2 and CH_4 . Thawing will make permafrost more permeable to soil gasses stored in or below the permafrost. Moreover, thawing will mobilize previously frozen organic matter, making it more easily available for the active microbial communities in the expanding active layer soils. But probably the dominant effect of Arctic warming on the greenhouse gas fluxes in permafrost regions will be through the effects on biogenic organic matter decomposition via changes in soil temperature and moisture regimes in active layer soils and in plant productivity and composition (e.g., Olefeldt et al., 2013).

Subsidence, degradation, irreversible processes, and the fate of permafrost

Gradual deepening of the active layer, i.e., “top-down” thawing of permafrost, is the typical way in which thawing proceeds in much of the permafrost region. However, thawing develops differently in terrain that hosts ice-rich permafrost deposits. When ice-rich ground thaws or massive ground ice melts, permafrost loses its stability, and the solid ground material subsides as a consequence. This process is referred to as thermokarst, and it involves the formation of characteristic landforms, such as thermokarst lakes, thermokarst depressions, or thermokarst mounds. Compared to gradual, top-down thawing, thermokarst can induce positive feedbacks, such as water impoundment, causing more rapid thawing of the ground.

In contrast to gradual thaw, this type of thaw is typically not captured by large-scale climate models, as these types of models do not specifically represent permafrost deposits containing excess ice (Koven et al., 2013). One of the few permafrost model studies that considers abrupt thaw processes, including thermokarst lake formation and drainage, shows that these abrupt thaw processes can mobilize large amounts of previously frozen organic carbon in deep permafrost layers during the 21st century (Nitzbon et al., 2020). One reason why thermokarst is not represented in most earth system models is that the process operates at fine spatial scales (meters to hundreds of meters), resulting in land cover spatial heterogeneities which are far below the lateral resolution of coarse-scale models (tens to hundreds of kilometers). Characteristic thermokarst landforms and other patterned ground features in permafrost regions are linked to heterogeneities in many surface and subsurface properties, such as vegetation, microtopography, soil texture, and water and ice content. These fine-scale heterogeneities in turn affect the transport of energy and water over the landscape (i.e., snow redistribution or water fluxes over a polygonal terrain). Numerical modeling using a tiling approach showed that lateral processes and the distribution of excess ice play key roles in the pathways of landscape evolution and permafrost degradation in ice-rich terrain (Nitzbon et al., 2020).

Summary

Degradation (warming and thawing) of Arctic permafrost is an ongoing global phenomenon and this process may be irreversible on shorter than millennial timescales. As permafrost thaws, increasing and deepening heat and water transfer through the active layer and taliks occurs as a result and will mobilize nutrients, carbon, contaminants, and greenhouse gasses. Due to the highly complex combination of permafrost state, hydrology, soil, and vegetation properties, the effects of Arctic warming on energy, water, and carbon fluxes in permafrost landscapes are difficult to predict and remain highly uncertain.

Thermokarst-related processes cause changes in the energy and water balances of ice-rich terrain, which result in substantially larger amounts of projected permafrost degradation and thaw-affected carbon under a warming climate, compared to projections for ice-poor terrain. The role of hydrology (wetting, drying) and the presence of excess ice and thermokarst will likely determine the fate of permafrost.

References

- Abolt CJ, Young MH, Atchley AL, et al. (2020) Feedbacks between surface deformation and permafrost degradation in ice wedge polygons, Arctic Coastal Plain, Alaska. *JGR Earth Surface* 125(3), e2019JF005349. <https://doi.org/10.1029/2019JF005349>.
- Antonova S, Sudhaus H, Strozzi T, et al. (2018) Thaw subsidence of a Yedoma Landscape in Northern Siberia measured in situ and estimated from TerraSAR-X interferometry. *Remote Sensing* 10: 494. <https://doi.org/10.3390/rs10040494>.
- Beer C (2016) Permafrost sub-grid heterogeneity of soil properties key for 3-D soil processes and future climate projections. *Frontiers in Earth Science* 4: 81. <https://doi.org/10.3389/feart.2016.00081>.
- Belke-Brea M, Domine F, Barrere M, et al. (2020) Impact of shrubs on winter surface albedo and snow specific surface area at a low Arctic site: In situ measurements and simulations. *Journal of Climate* 33(2): 597–609. <https://doi.org/10.1175/JCLI-D-19-0318.1>.
- Bennett KE, Miller G, Busey R, Chen M, Lathrop ER, Dann JB, Nutt M, Crumley R, Wilson CJ, Iversen CM, Wulfschlegler SD, et al. (2022) Spatial patterns of snow distribution in the sub-Arctic. *The Cryosphere* 16: 3269–3293. <https://doi.org/10.5194/tc-16-3269-2022>.
- Biskaborn BK, Smith SL, Noetzi J, et al. (2019) Permafrost is warming at a global scale. *Nature Communications* 10: 264. <https://doi.org/10.1038/s41467-018-08240-4>.
- Bogoyavlensky V, Bogoyavlensky I, Nikonov R, et al. (2021) New catastrophic gas blowout and giant crater on the Yamal Peninsula in 2020: Results of the expedition and data processing. *Geosciences* 11(2): 71. <https://doi.org/10.3390/geosciences11020071>.
- CAVM Team (2003) *Circumpolar Arctic Vegetation Map. (1:7,500,000 scale), Conservation of Arctic Flora and Fauna (CAFF) Map No. 1*. Anchorage, Alaska: U.S. Fish and Wildlife Service.
- Chadburn S, Burke E, Essery R, et al. (2015) An improved representation of physical permafrost dynamics in the JULES land-surface model. *Geoscientific Model Development* 8: 1493–1508. <https://doi.org/10.5194/gmd-8-1493-2015>.
- Chen Y, Hu FS, and Lara MJ (2021) Divergent shrub-cover responses driven by climate, wildfire, and permafrost interactions in Arctic tundra ecosystems. *Global Change Biology* 27(3): 652–663. <https://doi.org/10.1111/gcb.15451>.
- Chuvilin E, Bukhanov B, Davletshina D, Grebenkin S, and Istomin V (2018) Dissociation and self-preservation of gas hydrates in permafrost. *Geosciences* 8: 431. <https://doi.org/10.3390/geosciences8120431>.
- Dallimore SR and Collett TS (1995) Intrapermafrost gas hydrates from a deep core hole in the Mackenzie Delta, Northwest Territories. *Canada Geology* 23(6): 527–530. [https://doi.org/10.1130/0091-7613\(1995\)023%3C0527:IGHFAD%3E2.3.CO;2](https://doi.org/10.1130/0091-7613(1995)023%3C0527:IGHFAD%3E2.3.CO;2).
- de Vrese P and Brovkin V (2021) Timescales of the permafrost carbon cycle and legacy effects of temperature overshoot scenarios. *Nature Communications* 12: 2688. <https://doi.org/10.1038/s41467-021-23010-5>.
- Fischer DA, Lacelle D, and Pollard W (2020) A model of unfrozen water content and its transport in icy permafrost soils: Effects on ground ice content and permafrost stability. *Permafrost and Periglacial Processes* 31(1): 184–199. <https://doi.org/10.1002/ppp.2031>.
- French HM (2017) *The Periglacial Environment*, 4th edn. John Wiley & Sons.
- Gruber S (2020) Ground subsidence and heave over permafrost: hourly time series reveal interannual, seasonal and shorter-term movement caused by freezing, thawing and water movement. *The Cryosphere* 14(4): 1437–1447. <https://doi.org/10.5194/tc-14-1437-2020>.
- Grünberg I, Wilcox EJ, Zwieback S, Marsh P, and Boike J (2020) Linking tundra vegetation, snow, soil temperature, and permafrost. *Biogeosciences* 17: 4261–4279. <https://doi.org/10.5194/bg-17-4261-2020>.
- Harp DR, Zlotnik V, Abolt CJ, et al. (2021) New insights into the drainage of inundated ice-wedge polygons using fundamental hydrologic principles. *The Cryosphere* 15(8): 4005–4029. <https://doi.org/10.5194/tc-15-4005-2021>.
- Karjalainen O, Luoto M, Aalto J, et al. (2020) High potential for loss of permafrost landforms in a changing climate. *Environmental Research Letters* 15: 104065. <https://doi.org/10.1088/1748-9326/abafds>.
- Kessler MA and Werner BT (2003) Self-organization of sorted patterned ground. *Science* 299(5605): 380–383. <https://doi.org/10.1126/science.1077309>.
- Koven CD, Riley WJ, and Stern A (2013) Analysis of permafrost thermal dynamics and response to climate change in the CMIP5 earth system models. *Journal of Climate* 26(6): 1877–1900. <https://doi.org/10.1175/JCLI-D-12-00228.1>.
- Lachenbruch AH (1962) *Mechanics of Thermal Contraction Cracks and Ice-Wedge Polygons in Permafrost*. Special GSA Papers, No. 70 New York: Geological Society of America.
- Mastepanov M, Sigsgaard C, Dlugokencky E, et al. (2008) Large tundra methane burst during onset of freezing. *Nature* 456: 628–630. <https://doi.org/10.1038/nature07464>.
- Melnikov PI (1966) *The depth of upper earth's crust cooling at Yakutsk autonomous region of USSR/Geothermal research and usage of Earth's deep heat*. Moscow: Science, pp. 110–111.
- Meredith M, Sommerkorn M, Cassotta S, Derksen C, Ekaykin A, Hollowed A, Kofinas G, Mackintosh A, Melbourne-Thomas J, Muelbert MMC, Ottersen G, Pritchard H, and Schuur EAG (2019) Polar Regions. In: Pörtner H-O, Roberts DC, Masson-Delmotte V, Zhai P, Tignor M, Poloczanska E, ... Weyer NM (eds.) *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*, pp. 203–320. Cambridge/New York, NY: Cambridge University Press. <https://doi.org/10.1017/9781009157964.005>.
- Nicolson DJ, Romanovsky VE, Tipenko GS, and Walker DA (2008) Modeling biogeophysical interactions in non-sorted circles in the Low Arctic. *Journal of Geophysical Research – Biogeosciences* 113(G3). <https://doi.org/10.1029/2007JG000565>.
- Nitzbon J, Westermann S, Langer M, et al. (2020) Fast response of cold ice-rich permafrost in northeast Siberia to a warming climate. *Nature Communications* 11: 2201. <https://doi.org/10.1038/s41467-020-15725-8>.
- Nitzbon J, Gadylyayev D, Schlüter S, Köhne JM, Grosse G, and Boike J (2022) Brief communication: Unravelling the composition and microstructure of a permafrost core using X-ray computed tomography. *The Cryosphere* 16: 3507–3515. <https://doi.org/10.5194/tc-16-3507-2022>.
- Nitzbon J, Gadylyayev D, and Boike J (2021) Material composition in permafrost core through computer tomographic imaging. In: *Visualization of a permafrost core composition using computer tomographic imaging*, p. 2021. Copernicus Publications. <https://doi.org/10.5446/56865>.
- Obu J (2021) How much of the earth's surface is underlain by permafrost? *Journal of Geophysical Research - Earth Surface* 126(5): e2021JF006123. <https://doi.org/10.1029/2021JF006123>.
- Obu J, Westermann S, Bartsch A, et al. (2019) Northern hemisphere permafrost map based on TTOP modelling for 2000–2016 at 1 km² scale. *Earth-Science Reviews* 193: 299–316. <https://doi.org/10.1016/j.earscirev.2019.04.023>.
- Olefelt D, Devito KJ, and Turetsky MR (2013) Sources and fate of terrestrial dissolved organic carbon in lakes of a Boreal Plains region recently affected by wildfire. *Biogeosciences* 10(10): 6247–6265. <https://doi.org/10.5194/bg-10-6247-2013>.
- Panikov NS and Dedysh SN (2000) Cold season CH₄ and CO₂ emission from boreal peat bogs (West Siberia): Winter fluxes and thaw activation dynamics. *Global Biogeochemical Cycles* 14(4): 1071–1080. <https://doi.org/10.1029/1999GB900097>.
- Ping CL, Jastrow JD, Jorgenson MT, Michaelson GJ, and Shur YL (2015) Permafrost soils and carbon cycling. *The Soil* 1(1): 147–171. <https://doi.org/10.5194/soil-1-147-2015>.
- Porada P, Ekici A, and Beer C (2016) Effects of bryophyte and lichen cover on permafrost soil temperature at large scale. *The Cryosphere* 10(5): 2291–2315. <https://doi.org/10.5194/tc-10-2291-2016>.
- Previdi M, Smith KL, and Polvani LM (2021) Arctic amplification of climate change: A review of underlying mechanisms. *Environmental Research Letters* 16: 093003. <https://doi.org/10.1088/1748-9326/ac1c29>.
- Pu W, Shi T, Cui J, Chen Y, Zhou Y, and Wang X (2021) Enhancement of snow albedo reduction and radiative forcing due to coated black carbon in snow. *The Cryosphere* 15(5): 2255–2272. <https://doi.org/10.5194/tc-15-2255-2021>.

- Rivkina E, Gilichinsky DA, McKay C, and Dallimore S (2001) Methane distribution in permafrost: evidence for an inter pore pressure methane hydrate. In: Permafrost Response on Economic Development, Environmental Security and Natural Resources, pp. 487–496. Dordrecht: Springer. https://doi.org/10.1007/978-94-010-0684-2_33.
- Rivkina E, Laurinavichius K, McGrath J, et al. (2004) Microbial life in permafrost. *Advances in Space Research* 33(8): 1215–1221. <https://doi.org/10.1016/j.asr.2003.06.024>.
- Shu S, Jain AK, Koven CD, and Mishra U (2020) Estimation of permafrost SOC stock and turnover time using a land surface model with vertical heterogeneity of permafrost soils. *Global Biogeochemical Cycles* 34(11). <https://doi.org/10.1029/2020GB006585>. e2020GB006585.
- Stünzi SM, Boike J, Cable W, et al. (2021a) Variability of the surface energy balance in permafrost-underlain boreal forest. *Biogeosciences* 18: 343–365. <https://doi.org/10.5194/bg-18-343-2021>.
- Stünzi SM, Boike J, Gädeke A, Herzsich U, Kruse S, Pestryakova LA, Westermann S, and Langer M (2021b) Sensitivity of ecosystem-protected permafrost under changing boreal forest structures. *Environmental Research Letters* 16(8). <https://doi.org/10.1088/1748-9326/ac153d>.
- van Everdingen RO (ed.) (2005) *Multi-Language Glossary of Permafrost and Related Ground-Ice Terms*. Boulder, Colorado: World Data Centre for Glaciology. Available at: <http://nsidc.org/fgdc/glossary>.
- Webb EE, Loranty MM, and Lichstein JW (2021) Surface water, vegetation, and fire as drivers of the terrestrial Arctic-boreal albedo feedback. *Environmental Research Letters* 16(8): 084046. <https://doi.org/10.1088/1748-9326/ac14ea>.
- Wille C, Kutzbach L, Sachs T, Wagner D, and Pfeiffer E-M (2008) Methane emission from Siberian arctic polygonal tundra: Eddy covariance measurements and modeling. *Global Change Biology* 14: 1395–1408. <https://doi.org/10.1111/j.1365-2486.2008.01586.x>.
- Zhang W, Miller PA, Smith B, et al. (2013) Tundra shrubification and tree-line advance amplify arctic climate warming: results from an individual-based dynamic vegetation model. *Environmental Research Letters* 8(3): 034023. <https://doi.org/10.1088/1748-9326/8/3/034023>.

Relevant websites

<https://www.permafrost.org/>

Feedback mechanism

<https://www.youtube.com/watch?v=3qyT43pbUus>—Arctic feedbacks, University of Colorado Boulder

Energy and water balances

https://link.springer.com/chapter/10.1007/978-3-030-31379-1_2#Bib1—The Energy Balance of Permafrost Soils and Ecosystems.

Albedo

<https://www.sciencedirect.com/topics/earth-and-planetary-sciences/albedo>.

High resolution videos of the composition and microstructure of a permafrost core

<https://www.awi.de/en/science/geosciences/permafrost-research/research-focus/permafrost-outreach/permafrost-resources.html>.

https://www.youtube.com/playlist?list=PLnngXB12Nji-oNXZKS8rAlwQrZQQvc2_X.

<https://av.tib.eu/media/56864>.

Permafrost and cold region hydrology

Masaki Hayashi^a and Ylva Sjöberg^b, ^aDepartment of Geoscience, University of Calgary, Calgary, AB, Canada; ^bDepartment of Geosciences and Natural Resource Management, University of Copenhagen, Copenhagen, Denmark

© 2023 Elsevier Ltd. All rights reserved.

Introduction	419
Hydrological processes in seasonally frozen soil	420
Hydraulic conductivity of frozen soil	420
Role of macropores in frozen-soil infiltration	420
Land-use effects on frozen soil infiltrability	422
Effects of snowcover on infiltration and runoff	422
Spatial scale of runoff generation	422
Hydrological processes in permafrost regions	422
Impact of permafrost on the distribution of water	422
Taliks, springs, and frost mounds	424
Seasonality, travel times, and flow paths	424
Catchment-scale effects of permafrost	424
Effects of climate change on soil hydrology in cold regions	425
Climate-change effects on the hydrology of seasonally frozen soil	425
Climate-change effects on permafrost hydrology	425
Concluding remarks	426
References	426

Abstract

Permafrost and seasonally frozen soil are a dominant feature of cold regions around the world. The ice forming in frozen soil influences the hydrology of cold regions by: (1) reducing soil infiltrability and increasing surface runoff over frozen soil surfaces, (2) forming an impermeable barrier at the top of ice-rich soil and controlling the lateral flow of groundwater, and (3) providing a source of melt water during warm seasons.

Key points

- Soil freeze-thaw processes have a strong influence on soil hydraulic conductivity and infiltrability, and the partitioning of snowmelt into infiltration and runoff.
- Snowmelt infiltration in frozen soil is controlled by macropore networks that are open to flow even when the frozen soil matrix allows little infiltration. The hydrology of cold region is sensitive to agricultural practices that alter the development of macropore networks.
- Permafrost influences the partitioning of surface and subsurface water by limiting storage and flow of water in the ground and promoting storage and flow near and at the ground surface. Near surface flow paths are generally associated with fast travel times and water chemistry reflecting flow through organic rich soil layers.
- Climate-change effects on frozen soil has a large degree of uncertainty due to the complex feedback between soil, vegetation, and snowcover that provides insulation from cold air. A warmer winter (and shorter duration of snowcover) does not necessarily mean shallower soil frost.

Introduction

Permafrost and seasonally frozen soil cover approximately 50% of exposed land surfaces in the Northern Hemisphere (Zhang et al., 2003). In this article ‘cold region’ refers to the region in which snow and frozen soil have substantial effects on hydrological processes. The soil in these regions freezes and thaws in response to cooling and warming of the soil surfaces. Repeated freezing and thawing can influence soil physical, chemical, and biological processes (Hayashi, 2013) and alter soil structure and properties.

The ice forming in pore spaces influences the hydrology of cold regions by: (1) reducing soil infiltrability and increasing surface runoff over frozen soil surfaces, (2) forming an impermeable barrier at the top of ice-rich soil and controlling the lateral flow of groundwater, and (3) providing a source of melt water during warm seasons. The warming of climate will alter the depth and seasonality of soil frost and the thickness and spatial extent of permafrost. This will also affect the hydrology through changes in the three processes mentioned above.

Hydrological processes in seasonally frozen soil

Hydraulic conductivity of frozen soil

The blockage of flow by ice in pore spaces is the primary reason why hydrological processes are sensitive to soil freezing and thawing. Since soil hydraulic conductivity is determined by pore-size distribution and pore connectivity, it is sensitive to which pores are blocked and how. Soil matric potential is negative in the vadose zone meaning that the thermodynamic potential of liquid soil water is lower than that of free water. This results in freezing point depression, whereby soil water remains in the liquid phase at temperatures substantially below 0 °C. Given that the water suspended in smaller soil pores has a lower matric potential, the degree of freezing point depression is also greater. Therefore, ice formation in water-saturated soils starts in the largest water-filled pores. Since the permeability of individual soil pores is proportional to the square of pore diameter, the blockage of largest pores greatly reduces soil hydraulic conductivity.

The relationship between soil temperature, T (°K), and matric potential, μ_m (Pa) is represented by the Clausius-Clapeyron equation (Williams and Smith, 1989, p.190):

$$\frac{d\mu_m}{dT} = \frac{\rho_w L_f}{T} \quad (1)$$

where L_f (J kg⁻¹) is the latent heat of fusion, and ρ_w (kg m⁻³) is the density of liquid water. For example, the soil matric potential corresponding to -0.1 °C is roughly equal to -120 kPa. Therefore, the water in unsaturated soil at $\mu_m = -120$ kPa (or matric potential head of -12 m) does not freeze until the temperature drops below -0.1 °C (see Fig. 1 for an example). A theoretical pore diameter corresponding to -120 kPa is approximately 2 μm, implying that pores larger than a few microns are filled with air (Fig. 2a) (Mohammed et al., 2018).

These large, air-filled pores can transmit the infiltrating water sourced by rain or snowmelt. However, if the soil temperature is sufficiently low, the infiltrating water will freeze in pores, and the ice forming in these pores will block the flow of water and prevent further infiltration (Fig. 2c). Therefore, the history of repeated freezing and thawing and the timing of infiltration are important for the partitioning of snowmelt water into infiltration and surface runoff (Fig. 3b).

In addition to downward infiltration, liquid water in soil pores can move upward from a deeper zone to the freezing front (Fig. 3a), which is the boundary between partially frozen soil above and unfrozen soil below, driven by the matric potential gradient generated by the low potential of frozen soil (see above). When the unsaturated hydraulic conductivity of unfrozen soil is negligible, due to the small water content, water vapor plays a substantial role in the upward transfer of water from a deeper, warmer soil to the freezing front driven by the vapor pressure gradient. The upward flow of soil water to the freezing front from the water table (Fig. 3a) can cause the latter to recede during the frozen period in shallow unconfined aquifers (Ireson et al., 2013).

Role of macropores in frozen-soil infiltration

The theoretical framework presented above is useful for understanding the flow through frozen soil matrix. However, most soils in vegetated areas have macropores formed by decaying plant roots, animal burrows, and other processes. These macropores remain dry during freeze-thaw cycles and provide open conduits for infiltration. Relatively fast movement of liquid water in macropores can cause a thermal disequilibrium, whereby the temperature of water flowing through macropores can be substantially higher than the temperature of soil matrix. Depending on the rate of water inputs from the surface (e.g., snowmelt rate) and flow velocity, macropore networks in frozen soil can remain open and allow unrestricted infiltration (Fig. 2c). However, macropore networks

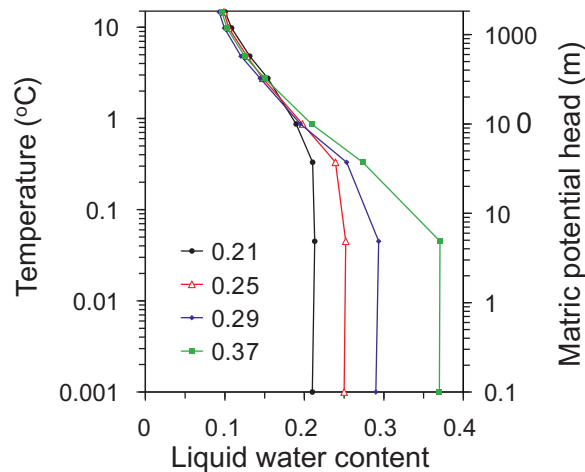


Fig. 1 Relation between liquid water content and temperature for silt loam soil for different total water contents (WC). The soil is saturated (i.e. no air in pore spaces) for the curve of total WC = 0.37. Other curves show the cases in which the soil is unsaturated prior to freezing (Watanabe and Wake (2009)).

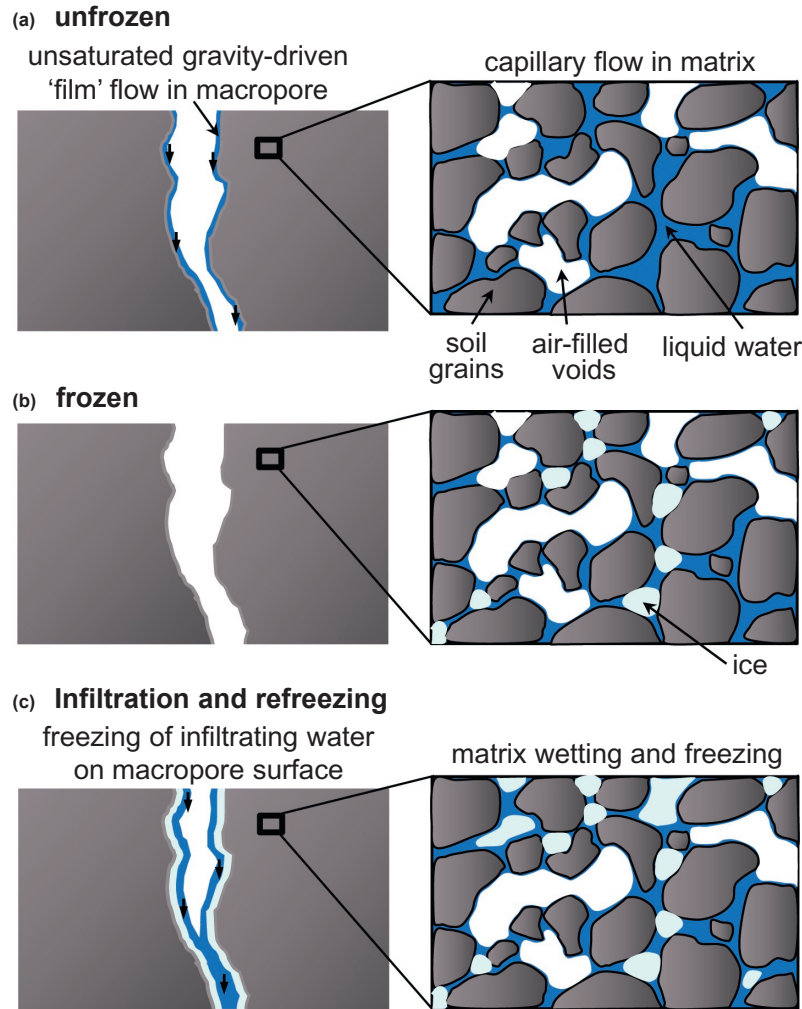


Fig. 2 Schematic diagram of unsaturated soil depicting the distribution of air, liquid water, and ice in the soil matrix and macropores. (a) Infiltration in unfrozen soil. (b) Frozen soil in winter with no infiltration. (c) Infiltration and refreezing of water during spring melt or midwinter melt events. Adapted from Mohammed AA, Kurylyk BL, Cey EE and Hayashi M (2018) Snowmelt infiltration and macropore flow in frozen soils: overview, knowledge gaps, and a conceptual framework. *Vadose Zone Journal* 17. doi:10.2136/vzj2018.04.0084.

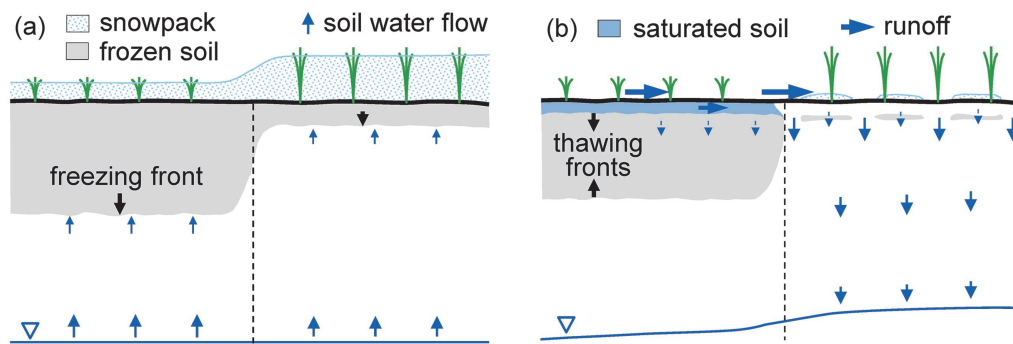


Fig. 3 Schematic cross sections demonstrating hydrological fluxes affected by frozen soil. (a) Midwinter condition showing the influence of vegetation height on snowcover thickness and soil frost depth. Upward arrows indicate the freezing-induced soil water flow. (b) Spring condition showing the surface and subsurface runoff generated by frozen soil having limited infiltrability (left half). The runoff is transferred to the area with no or thin soil frost, where it infiltrates and recharges groundwater.

may have local constrictions, which can slow the water movement and provide sufficient time for freezing and the blockage of macropore networks (Watanabe and Kugisaki, 2017). Therefore, the presence of macropore network does not necessarily cause sustained infiltration at a high rate in soils subjected to large temperature swings due to diurnal warming/cooling or a mid-winter warm event (e.g., foehn) followed by a cold period.

Land-use effects on frozen soil infiltrability

Given that macropore networks have a strong influence on the infiltrability of frozen soil, it is important to understand the factors controlling the formation and destruction of soil macropores. In general inversion tillage can improve the soil structure over a short time frame (e.g., one growing season) but it destroys macropores and prevent the development of well-connected macropore networks. As a result, cropped fields under conventional tillage typically have smaller frozen soil infiltrability compared with undisturbed grasslands having the same soil texture (Hayashi, 2013). Therefore, conventionally tilled fields tend to generate greater amounts of snowmelt runoff over frozen soil than undisturbed grasslands.

Zero-tillage and continuous cropping practices allow the soil to preserve macropore networks, thereby reducing the amount of snowmelt runoff. On the other hand, grasslands subjected to animal grazing may be affected by compaction by animals and may have less developed macropore networks compared to natural grasslands. Therefore, the contrast in snowmelt runoff generation between cropped fields under zero-till and grazed grasslands is smaller than between conventionally tilled fields and undisturbed grasslands.

Effects of snowcover on infiltration and runoff

Snowcover has a strong influence on frozen-soil processes by providing thermal insulation of soil surfaces and releasing snowmelt water. Soil freezing and thawing is controlled by energy fluxes and storage in a complex system consisting of snowpack, plant canopy and crop residue, and soil. The thickness of snowpack and the height and density of plant canopy are spatially heterogeneous, which can result in a substantial spatial variability in soil freezing and thawing processes (Fig. 3a). Snow accumulation and melt processes are influenced by topography and land-use practices such as crop management, grazing, and tree harvesting in forests. Therefore, the variability in topography and land use can have a strong influence on soil freezing/thawing and snowmelt infiltration and runoff processes at various scales ranging from field plots (e.g., 1 m) to a landscape (e.g., 10^2 – 10^3 m).

Spatial scale of runoff generation

Heterogeneous distribution of snowpack and frozen soil results in a high spatial variability of snowmelt runoff generation within a catchment. Therefore, some areas of a catchment may have frozen soils generating runoff and other areas may be unfrozen and accommodate the infiltration of snowmelt runoff transferred from frozen areas (Fig. 3b). Depending on the sizes and connectivity of runoff-generating areas within the catchment, frozen soil may not have a measurable influence on catchment-scale runoff. Therefore, it is important to consider spatial scales for understanding the effects of frozen soil on hydrological processes.

For example, the Northern Great Plains of North America is subjected to seasonal soil frost up to 1–1.5 m in depth and has a high density of topographically closed depressions ranging in size from 10^2 to 10^4 m². Snowmelt runoff generated over frozen soils in the catchment of these depressions are retained until it starts to spill over the edge of the depression. A large fraction of snowmelt water collected in the depression infiltrates and recharges local groundwater (i.e. depression-focused groundwater recharge). As a result, catchment-scale runoff measured in stream gauging stations is much smaller than local-scale runoff generated over frozen grounds. Similar scale-dependent snowmelt runoff has been observed in forested slopes, where the canopy density controls snowpack thickness, which in turn controls the extent of frozen soil and snowmelt runoff (Stähli, 2005).

Hydrological processes in permafrost regions

Impact of permafrost on the distribution of water

In permafrost regions, which cover approximately 25% of the northern hemisphere landmass, at least some of the ground material is perennially frozen, although unfrozen water may also exist within permafrost due to freezing point depression caused by soil matric potential (Eq. 1) or by high solute concentration. The active layer is the layer of ground above permafrost which thaws and refreezes seasonally as temperature fluctuates between annual maximum and minimum (Fig. 4a). Permafrost is defined as a perennially cryotic zone (Fig. 4a), meaning temperature $\leq 0^\circ\text{C}$ for two or more consecutive years, which is different from a perennially frozen zone (French, 2018, p.72). Therefore, the boundary between the active layer and permafrost may be ambiguous in some cases (gray zone in Fig. 4a). Under a warming climate, temperature profiles can deviate substantially from the equilibrium profile resulting in a perennially non-cryotic ($T > 0^\circ\text{C}$) zone above permafrost, referred to as talik (Fig. 4b). Fig. 4c and d demonstrate the difference in seasonal progression of soil freezing and thawing in cases with and without talik. The active layer typically reaches a maximum thickness ranging from decimeters to a few meters depending on local climate and ground thermal properties. In regions with continuous permafrost, where at least 90% of the ground surface is underlain by permafrost, the active

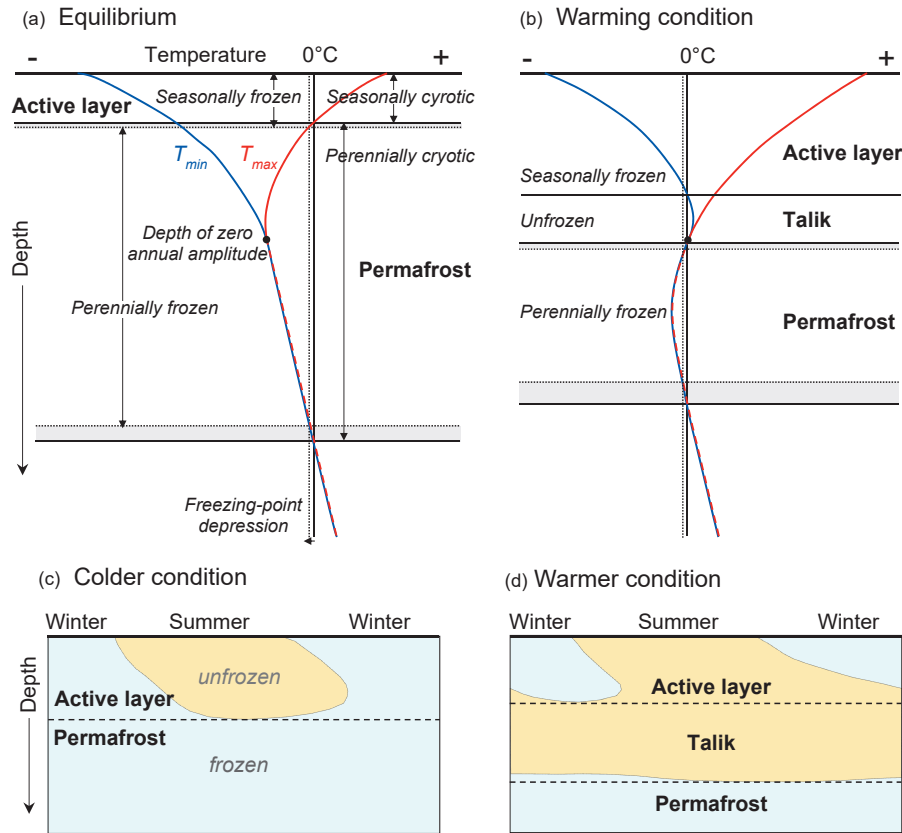


Fig. 4 (a) Schematic temperature-depth profile through permafrost under thermal equilibrium showing annual maximum (T_{max}), minimum (T_{min}), and the depth of zero annual amplitude. (b) Temperature-depth profile under a warming condition depicting the development of talik between the active layer and permafrost. (c) Schematic diagram showing seasonal progression of freezing and thawing of the active layer directly overlying permafrost. (d) Seasonal progression of freezing and thawing under a warmer condition depicting the presence of talik between the active layer and permafrost.

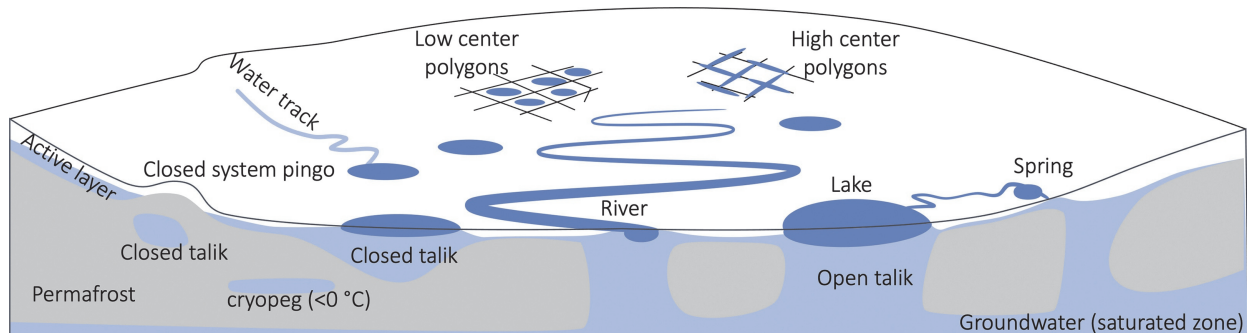


Fig. 5 Hydrological features and landforms in a landscape with continuous permafrost (left side of river) and discontinuous permafrost (right side of river).

layer constitutes the main aquifer of consideration for most hydrological concerns. In the active layer the water table is perched on top of the frost table which continuously thaws deeper over the summer season (Fig. 5).

Percolation from the active layer into deeper aquifers below the permafrost is restricted by low permeability frozen ground, which leads to an abundance of wetlands and lakes in many lowland permafrost areas. Polygonal terrain is common in lowland tundra areas and forms by repeated infiltration of water into thermal contraction cracks over hundreds to thousands of years, creating a polygonal pattern of ice wedges below the tundra surface (Fig. 5) (Liljedahl et al., 2016). Due to the massive ice content in ice wedges, the ground above them is raised above the surrounding tundra and water is often ponded in the center of so-called low centered polygons. When ice wedges degrade, the raised polygon rims turn into troughs leading to redistribution of water into the rims of so-called high center polygons. In high-center polygonal terrain, the troughs provide connected drainage paths which may lead to less ponding of surface water and more runoff, compared to low-center polygonal terrain (Liljedahl et al., 2016). The perched water table and abundance of wetlands, such as ice wedge polygons, can contribute to relatively high evapotranspiration

rates in some permafrost areas, as much water is stored near or at the ground surface (Sjöberg et al., 2021). Evaporation tends to dominate over transpiration, although few published studies have been focused on the partitioning of evapotranspiration in permafrost areas (Young-Robertson et al., 2018). With decreasing cover of permafrost into the discontinuous and sporadic permafrost regions the number of connections between aquifers increase and the regional partitioning of surface water and groundwater is less affected by the presence of permafrost.

Taliks, springs, and frost mounds

Taliks are important storages and conduits for (liquid) groundwater (Fig. 5). Taliks can be non-cryotic ($> 0^{\circ}\text{C}$) or cryotic ($\leq 0^{\circ}\text{C}$). Non-cryotic taliks are often found below water bodies, such as lakes, streams, and wetlands, or on relatively warmer south-facing slopes (in the northern hemisphere) or in locations with thick insulating snow cover which. Cryotic taliks are kept unfrozen due to a lowering of the freezing point of water. Cryopegs are perhaps the most common type of cryotic taliks and are defined as layers of ground that remain unfrozen at temperatures below 0°C due to large concentrations of dissolved solids. They are often found in coastal permafrost areas. Open taliks penetrate the permafrost completely, while closed taliks are contained above or within the permafrost (Woo, 2012). Open taliks allow for discharge of groundwater from below the permafrost to the ground surface and thereby the formation of springs. During winter, the freezing of discharging groundwater can lead to build-up of icings (German *aufeis*, Russian *naled*), for example in stream beds and near springs. Icings typically range in size from a few square meters to tens of square kilometers depending on the supply of water (Ensom et al., 2020).

Where ice blocks the discharge of groundwater through the ground surface, mounds can form from the build-up of ice from freezing groundwater below the surface (Woo, 2012). Pingos are perennial frost mounds that can reach heights above 100 m and are classified into open system or closed system types (Fig. 5) (Yoshikawa and Kane, 2021). Open system pingos receive water through underlying taliks connecting them to the subpermafrost aquifer. Closed system pingoes are found on top of closed taliks in the continuous permafrost zone, typically in drained lake beds. As permafrost aggrades after lake drainage, water in the closed talik is pressurized and expelled into the pingo ice core. Distinct types of frost mounds, namely palsas and peat plateaus, are found in peatlands. These landforms, which are typically less than a few meters high, are formed by ice segregation and often create locally drained islands in otherwise waterlogged peatlands. While closed system pingos are typically found in continuous permafrost areas, open system pingos are common in the discontinuous permafrost areas, and palsas and peat plateaus are common features near the climatic margin of permafrost in sporadic and isolated permafrost zones.

Seasonality, travel times, and flow paths

The seasonal development of the active layer contributes to high seasonal variability in water flow paths and travel times. The seasonal deepening of the frost table in the active layer can lead to a pronounced transmissivity feedback in permafrost areas, as flow paths in the early summer go through near surface material with high hydraulic conductivity (Carey and Woo, 2001). Such shallow flow paths are associated with short travel times, and as the summer progresses and the frost table thaws deeper, flow paths may deepen, and travel times may grow longer. The seasonal development of flow paths also results in seasonally variable water chemistry in streams draining permafrost areas. In the active layer, groundwater flow may move predominantly in the organic topsoil in early summer but move into deeper mineral soils with thawing over the summer and the water chemistry tends to reflect such changes (O'Donnell et al., 2021). Water in subpermafrost aquifers in continuous permafrost areas tends to move slowly and can be several thousand years old and often has large concentrations of dissolved solids.

Catchment-scale effects of permafrost

Streams that drain permafrost regions are commonly characterized as flashy. The presence of permafrost generally leads to very low streamflow during the winter season due to limited groundwater inflow and solid precipitation (snow). The peak flow during spring freshet is in contrast very high, reflecting the shallow and rapid flow paths during this season. Near surface water is also drained through water tracks which form stream-like drainage networks in many continuous permafrost regions. While there is no formal definition of water tracks, they are commonly referred to as local to regional linear features with high soil moisture, sometimes including overland flow, and a thick active layer (Trochim et al., 2016). By increasing the landscape hydrologic connectivity, they have important impact on the drainage response of permafrost catchments. They are easily distinguishable in tundra landscapes by relatively lush vegetation. Just like seasonal progression of thaw can be reflected in water chemistry, the catchment-scale extent of permafrost may also impact the chemistry of streams. In a catchment with continuous permafrost much water moves through organic-rich near surface flow paths while with less permafrost more water moves through mineral soil layers. For example, high concentrations of dissolved organic carbon (DOC) in stream water can be seen as an indicator of permafrost whereas higher levels of weathering products in stream water may indicate less permafrost in the catchment.

Effects of climate change on soil hydrology in cold regions

Climate-change effects on the hydrology of seasonally frozen soil

A rise in winter temperature will generally increase the rain-to-snow ratio of winter precipitation and shorten the period of snowcover. While there is a fair degree of confidence in this prediction over the global scale, the effects of warming on soil frost will likely have substantial regional- and local-scale variabilities. This is because the complex interplay between temperature and precipitation affects snow accumulation and melt processes, which in turn affects soil freezing and thawing through thermal insulation, or lack thereof, by snowcover (Iwata et al., 2010).

Field experiments in cold regions around the world have unequivocally shown that the removal of snowcover lowers soil temperature and increases frost depths. In contrast, a thickening snowcover results in warmer soil temperatures but a shallower frost depth (Fig. 3a). Therefore, if the warming causes frequent snowcover depletion during warm spells followed by cold spells, it will increase the frost depth instead of decreasing it. Such a scenario may increase or decrease snowmelt runoff, depending on the amount, intensity, and timing of snowmelt; and the infiltrability of near-surface soil (see below).

The energy input for midwinter snowmelt is primarily provided by sensible heat flux from the warm air over the melting snowpack because the radiation input is relatively small in midwinter. The sensible heat flux gain during a warm spell is partially offset by the latent heat flux loss associated with evaporation or sublimation. As such, the rate of midwinter melt is much less than that of spring melt driven by radiation, even under a warmer climate. Therefore, the midwinter release of meltwater from the snowpack is also smaller unless it is enhanced by rain on snow. As a result, the meltwater release rate may be small enough to be absorbed by the unsaturated frozen soil. In such case, the meltwater will wet the surface soil and refreeze during the subsequent cold spell, and repeated melt-refreeze cycles will reduce the soil infiltrability. If the soil is still frozen and the snowcover is left on the ground in the spring, the high snowmelt rate driven by both radiation and sensible heat flux can generate a large volume of runoff.

The climate-change effects on the hydrology of seasonally frozen catchments depend on the complex interaction between changes in air temperature and precipitation. Therefore, the prediction of climate-change effects requires careful consideration of energy and water transfer processes in the atmosphere-snow-soil system and its feedback with vegetation covers and land management practices.

Climate-change effects on permafrost hydrology

Warming and subsequent thawing of permafrost may impact hydrology in several ways (Walvoord and Kurylyk, 2016). A thickening of the active layer may result in a lowering of the water table in the active layer and thereby a deepening of active layer water flow paths (Fig. 6). Warming may lead to higher potential evapotranspiration, while a deeper water table and increased connectivity may limit the water available for actual evapotranspiration (Sjöberg et al., 2021; Andresen et al., 2020). The net effect of warming and permafrost thaw on evapotranspiration is so far poorly constrained.

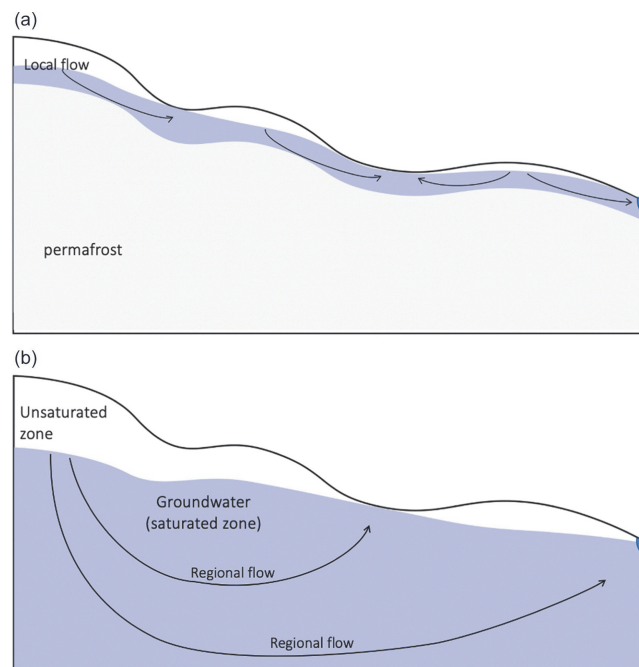


Fig. 6 Impact of permafrost on flow paths and water table (conceptual). Panel (a) illustrates local flow paths with a high groundwater table perched by permafrost, while panel (b) shows how without permafrost longer flow paths are present.

The formation of a supra-permafrost talik directly below the seasonal frost may allow for groundwater flow during the cold season (Walvoord et al., 2019). In general, thawing permafrost is expected to lead to increased connectivity and storage of (liquid) groundwater. Such catchment-scale changes may be detectable in historical streamflow records. Increases in winter streamflow have been observed in rivers across the Arctic and have been interpreted as signs of reduction in catchment-scale permafrost (St. Jacques and Sauchyn, 2009). The storage-discharge relationship of a catchment has also been leveraged to infer changes in permafrost from analyzing changes in the shape of the hydrograph during recession flow periods (Lyon and Destouni, 2010). Changes in stream water chemistry, such as decreasing DOC concentration or rising concentrations of weathering products have been interpreted as signs of deeper water flow paths due to thawing permafrost (O'Donnell et al., 2021).

Concluding remarks

The permanent or seasonal presence of ice is a distinct feature of soil hydrology in cold regions. While the thermodynamic principles governing soil freezing and thawing are reasonably well established in laboratory settings, our understanding of field-scale processes has many challenges. The infiltrability of frozen soil has a strong influence on the partitioning of snowmelt water into infiltration and runoff, which in turn affects the magnitude and timing of water inputs to river systems. The frozen soil infiltrability is strongly dependent on the complex interaction between soil matrix and macropore networks, as well as the history of repeated freeze-thaw and midwinter infiltration events. To advance knowledge on the effects of frozen soil on hydrological processes, an important challenge is to improve our capability to predict field-scale soil infiltrability in various environments.

The presence of permafrost has profound implications on the partitioning of surface and subsurface water storage and flow paths. Water tables perched near the ground surface by permafrost forces flow through soil layers with typically high hydraulic conductivity and organic content. Thawing of permafrost may lead to increases in groundwater connectivity and storage, deeper and slower groundwater flow paths, and shifts in water chemistry. At the catchment scale, such changes in storage and connectivity are challenging to observe and current understanding is to a large extent built on theoretical assumptions.

References

- Andresen CG, Lawrence DM, Wilson CJ, McGuire AD, Koven C, Schaefer K, Jafarov E, Peng S, Chen X, Gouttevin I, Burke E, Chadburn S, Ji D, Chen G, Hayes D, and Zhang W (2020) Soil moisture and hydrology projections of the permafrost region—A model intercomparison. *The Cryosphere* 14: 445–459.
- Carey SK and Woo M-K (2001) Slope runoff processes and flow generation in a subarctic, subalpine environment. *Journal of Hydrology* 253: 110–129.
- Ensom T, Makarieva O, Morse P, Kane D, Alekseev V, and Marsh P (2020) The distribution and dynamics of aufeis in permafrost regions. *Permafrost and Periglacial Processes* 31: 383–395.
- French HM (2018) *The Periglacial Environment*, 4th edn., p. 515. John Wiley & Sons.
- Hayashi M (2013) The cold vadose zone: Hydrological and ecological significance of frozen-soil processes. *Vadose Zone Journal* 12. <https://doi.org/10.2136/vzj2013.03.0064>.
- Ireson AM, van der Kamp G, Ferguson G, Nachshon U, and Wheeler HS (2013) Hydrogeological processes in seasonally frozen northern latitudes: understanding, gaps and challenges. *Hydrogeology Journal* 21: 53–66.
- Iwata Y, Hayashi M, Suzuki S, Hirota T, and Hasegawa S (2010) Effects of snow cover on soil freezing, water movement, and snowmelt infiltration: A paired plot experiment. *Water Resources Research* 46: W09504. <https://doi.org/10.1029/2009WR008070>.
- Liljedahl A, Boike J, Daanen R, et al. (2016) Pan-Arctic ice-wedge degradation in warming permafrost and its influence on tundra hydrology. *Nature Geoscience* 9: 312–318. <https://doi.org/10.1038/ngeo2674>.
- Lyon SW and Destouni G (2010) Changes in catchment-scale recession flow properties in response to permafrost thawing in the Yukon River basin. *International Journal of Climatology* 30: 2138–2145.
- Mohammed AA, Kurylyk BL, Cey EE, and Hayashi M (2018) Snowmelt infiltration and macropore flow in frozen soils: overview, knowledge gaps, and a conceptual framework. *Vadose Zone Journal* 17. <https://doi.org/10.2136/vzj2018.04.0084>.
- O'Donnell J, Douglas T, Barker A, and Guo L (2021) Changing biogeochemical cycles of organic carbon, nitrogen, phosphorus, and trace elements in Arctic rivers. In: Yang D and Kane DL (eds.) *Arctic Hydrology, Permafrost and Ecosystems*. Cham: Springer. https://doi.org/10.1007/978-3-030-50930-9_11.
- Sjöberg Y, Jan A, Painter SL, Coon ET, Carey MP, O'Donnell JA, and Koch JC (2021) Permafrost promotes shallow groundwater flow and warmer headwater streams. *Water Resources Research* 57: e2020WR027463.
- St. Jacques JM and Sauchyn DJ (2009) Increasing winter baseflow and mean annual streamflow from possible permafrost thawing in the Northwest Territories, Canada. *Geophysical Research Letters* 36: L01401. <https://doi.org/10.1029/2008GL035822>.
- Stähli M (2005) Freezing and thawing phenomena in soils. In: Anderson MG (ed.) *Encyclopedia of the Hydrological Sciences*, pp. 1069–1076. New York: John Wiley & Sons.
- Trochim ED, Jorgenson MT, Prakash A, and Kane DL (2016) Geomorphic and biophysical factors affecting water tracks in northern Alaska. *Earth and Space Science* 3: 123–141.
- Walvoord MA and Kurylyk BL (2016) Hydrologic impacts of thawing permafrost—A review. *Vadose Zone Journal* 15. <https://doi.org/10.2136/vzj2016.01.0010>.
- Walvoord MA, Voss CI, Ebel BA, and Minsley BJ (2019) Development of perennial thaw zones in boreal hillslopes enhances potential mobilization of permafrost carbon. *Environmental Research Letters* 14: 015003.
- Watanabe K and Kugisaki Y (2017) Effect of macropores on soil freezing and thawing with infiltration. *Hydrological Processes* 31: 270–278.
- Watanabe K and Wake T (2009) Measurement of unfrozen water content and relative permittivity of frozen unsaturated soil using NMR and TDR. *Cold Regions Science and Technology* 59: 34–41.
- Williams PJ and Smith WM (1989) *The Frozen Earth: Fundamentals of Geocryology*. Cambridge: Cambridge University Press.
- Woo M-K (2012) *Permafrost Hydrology*. Berlin, Germany: Springer-Verlag. <https://doi.org/10.1007/978-3-642-23462-0>.
- Yoshikawa K and Kane DL (2021) Permafrost features and talik geometry in hydrologic system. In: Yang D and Kane DL (eds.) *Arctic Hydrology, Permafrost and Ecosystems*. Cham: Springer. https://doi.org/10.1007/978-3-030-50930-9_14.
- Young-Robertson JM, Raz-Yaseef N, Cohen LR, Newman B, Rahn T, Sloan V, et al. (2018) Evaporation dominates evapotranspiration on Alaska's Arctic Coastal Plain. *Arctic Antarctic and Alpine Research* 50: e1435931.
- Zhang T, Barry RG, Knowles K, Ling F, and Armstrong RL (2003) Distribution of seasonally and perennially frozen ground in the Northern Hemisphere. In: Phillips M, et al. (ed.) *Proceedings of the 8th International Conference on Permafrost*, Zürich, Switzerland. 21–25 July 2003, pp. 1289–1294. Rotterdam, The Netherlands: A.A. Balkema.

Soil water sensing by neutron scattering

Steven R Evett, Conservation and Production Research Laboratory, USDA-ARS, Bushland, TX, United States

© 2023, Published by Elsevier Ltd.

Introduction	427
How it works	428
Interferences	428
Neutron moisture meters	428
Access tubes and depth control	429
NMM statistics	430
NMM calibration, accuracy, and representativity	431
NMM safety and use considerations	431
Cosmic ray neutron scattering	432
Outlook	434
References	434

Abstract

Soil water content can be sensed by counting thermal (energy at room temperature) neutrons that are scattered from collisions of more energetic (fast) neutrons with the nuclei of atoms. Repeated collisions reduce the energy of fast neutrons until they are at thermal energy, a process called moderation. Hydrogen atoms, mainly found in soil water, are more effective moderators than are most elements found in soils, so the count of thermal neutrons is a good indicator of the amount of water in soil. Fast neutron sources can be man-made or natural (cosmic ray collisions with atoms). Neutron scattering approaches can be used for access tube measurements or larger-scale near-surface measurements using the cosmic ray neutron scattering method.

Key points

- The neutron moisture meter (NMM, also called neutron probe) is an accurate ($\text{RMSE} < 0.01 \text{ m}^3 \text{ m}^{-3}$) method for assessing soil and vadose zone profile water content and changes in storage in profiles that may extend to $> 100 \text{ m}$ depth.
- The NMM senses a much larger soil volume ($14,000\text{--}65,000 \text{ cm}^3$) than do in situ soil water sensors that rely on electromagnetic measurements ($60\text{--}600 \text{ cm}^3$).
- The NMM uses a man-made radioactive source of fast neutrons that prohibits its unattended use.
- Hydrogen in the soil, mostly in soil water, preferentially slows fast neutrons until they reach a thermal energy state.
- The source and detector of thermal neutrons are placed together in a probe that is lowered into the soil through an access tube.
- The count of thermal neutrons increases with the soil water content.
- The cosmic ray neutron scattering (CRNS) method employs at least two detectors that are placed at a height of approximately 1 m above the soil surface, one to sense thermal neutrons and one to sense fast neutrons.
- The CRNS method estimates the soil water content in a much larger (radius of $130\text{--}240 \text{ m}$) but also much more ill-defined volume of soil.
- Depth of CRNS sensing ranges from 15 cm to 70 cm but most of the sensed volume is at less than 20-cm depth.

Introduction

The method of sensing soil water content by counting slow neutrons scattered back to a detector was conceived of due to understanding of neutron interactions with matter developed in the 1940s (Belcher et al., 1950; Pieper, 1949), and an effective neutron moisture meter (NMM) using a man-made radioactive source of energetic neutrons was demonstrated by Grant (1975). Seven decades later, the neutron scattering method used in access tubes remains a competitive method for repeated determination of changes in soil profile water content (Evett et al., 2008; IAEA, 2000) because it is nondestructive, has a large sensing volume that reduces variability compared to other methods, can be field calibrated for accurate sensing, works successfully to depths not easily attained with other methods (e.g., $> 100 \text{ m}$), and works well in stony soils and cracking clays in which other methods work poorly. The large volume sensed means that a given precision can be obtained with fewer replicates than are required for other methods such as the capacitance and time domain methods, that soil disturbance during access tube installation has minimal effect on results (unlike electronic sensor methods), and that field calibration is successful because volumetric soil samples can be obtained from within the volume measured by the probe at each depth. The technology is mature with a wide literature base describing

applications and problems. The Cosmic Ray Neutron Scattering (CRNS) method, which relies on naturally occurring neutrons, was conceived by Desilets and Zreda (2003) and developed into a useful method more recently (Desilets and Zreda, 2013; Franz et al., 2012a, 2012b; Rosolem et al., 2013). The two methods have much in common but also are distinctly different in what they can reliably sense and in how they may be used.

How it works

The neutron thermalization method relies on a source of fast neutrons (mean energy of 5 MeV) and a detector of slow neutrons (~ 0.025 eV or 300 K). Fast neutron sources are either man-made as in the case of the neutron moisture meter (NMM) or natural as in the case of the CRNS method. Energetic neutrons emitted from a NMM source ($\sim 10^9$ /s) are either slowed through repeated collisions with the nuclei of atoms in the soil (scattering and thermalization) or are absorbed by those nuclei, the latter being one source of interference with the method. A small fraction of scattered neutrons will enter the detector. Of these, an even smaller fraction ($\sim 10^3$ /s) will be slowed to thermal (room temperature) energy levels and will be detected. Soil density and chemical composition affect the concentration of thermalized neutrons around the detector. The most common atoms in soil (aluminum and silicon) scatter neutrons with little energy loss because they have much greater mass than a neutron. However, the energy of a neutron is halved if it hits a hydrogen nucleus, because the mass of the hydrogen nucleus is the same as that of the neutron. On average, 19 collisions with hydrogen are required to thermalize a neutron, a process called moderation. Other relatively efficient neutron moderators are carbon, nitrogen, and oxygen (about 120, 140 and 150 collisions, respectively). On the time scales of common interest in water and environmental management, changes in soil carbon and nitrogen content are minor and have little effect on changes in the concentration of thermal neutrons, although absolute quantities of these elements and of oxygen in the soil may affect the calibration of the neutron method as is the case with soil with large concentrations of CaCO_3 . Also, on these time scales, changes in soil hydrogen and oxygen content occur mainly due to changes in soil water content. The detection of thermal neutrons is thus most affected by changes in water content; and volumetric water content can be accurately and precisely related to the count of thermal neutrons through calibration of the NMM. For the NMM the count of slow neutrons is positively related to the soil water content. For the CRNS method, the source of fast neutrons is natural collisions of cosmic rays with atomic nuclei in the atmosphere near the ground. The CRNS detector is positioned above ground and detects thermal neutrons scattered back into the atmosphere from the soil, vegetation, and from water in the air. The back scattered slow neutron count is reduced as water content of the soil, vegetation and atmosphere increase, resulting in an inverse relationship between soil water content and thermal neutron count for the CRNS method.

Interferences

The organic matter content of soil affects the NMM calibration because hydrogen and carbon effectively thermalize neutrons. Also, organic matter and most clays contain important amounts of hydrogen, some not in the form of water, that may not be driven off by heating to 105 °C (the standard temperature for drying soil samples). Separate calibrations thus are often required for soil layers that differ in organic matter or clay content from layers above or below. In arid or semi-arid zones, many soils have layers rich in CaCO_3 and CaSO_4 that require separate calibration (Van Bavel et al., 1956). Boron, cadmium, chlorine, iron, fluorine, lithium and potassium are atoms that absorb neutrons, and soils or soil horizons that contain large or fluctuating amounts of such elements will require separate calibrations or adjustments in data interpretation. For example, separate calibration may be required for soils with appreciable concentrations of boron, chloride salts or iron, such as Oxisols or soils rich in magnetite. These interferences with the neutron scattering method have been rather thoroughly discussed in the literature for the NMM but are less well understood for the CRNS method. Neither method is much affected by soil salinity within the range survivable by plants.

Neutron moisture meters

Typical forms of NMM equipment include: (1) a profiling meter with a source—detector pair assembled into a cylindrical probe that is lowered to user-chosen depths into a hole in the soil, and (2) a flat-based meter that is placed on the soil surface with the source and detector fixed at separate locations inside the base of the meter. The surface meter senses a volume that is roughly hemispherical and extends into the soil for a distance that decreases as soil water content and soil density increase, and which varies from ~ 0.15 m in wet soil to ~ 0.3 m in dry soil (Van Bavel et al., 1956). The precision is less than can be attained with a profiling NMM meter; and it suffers even more when soil moisture changes greatly with depth near the surface (Van Bavel et al., 1961), a common occurrence. Nakayama and Allen (1990) reported that stringent conditions are required for good precision including: (1) flattening the surface to fit the meter bottom with no air gaps, (2) marking the measurement site so that the meter can be repeatedly placed in identical position, and (3) using a neutron absorber shield made of cadmium around the meter (except for the bottom) to reduce effects of surrounding vegetation. Even with these conditions, however, the strong depth dependency of calibration coefficients and the inability to accurately estimate the depth of reading led to great uncertainty as to the accuracy of measurements.

The profiling NMM (Fig. 1B) is more commonly used. It is operated at user-chosen depths in the soil. The hole is lined with a cylindrical access tube in order to protect the probe and ensure a constant hole diameter. A cable connects the probe to a counter,

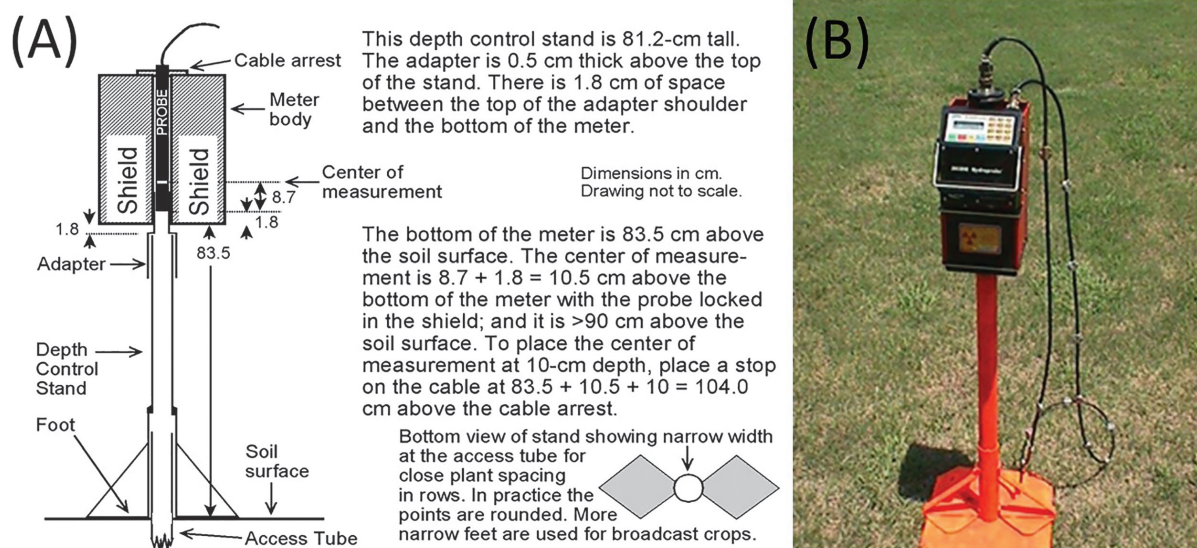


Fig. 1 (A) Cross-sectional schematic of a profiling neutron moisture meter in place on top of a depth control stand (Evelt et al., 2003). The stand is in place over an access tube that has been inserted into the soil. The probe is locked in the meter shield and standard counts may be taken in this position so long as vegetation is absent for a distance of at least 3 m. For the dimensions given here, to measure at 10-cm depth, the probe must be lowered 104 cm through the stand and into the access tube. Clamps placed on the cable ensure that the lowering distance is maintained for that a subsequent reading depths. Dimensions vary for meters of different manufacture. (B) Photograph of a Campbell Pacific Nuclear NMM resting on a depth control stand and with the probe locked in the shield. The stand is itself resting on a base plate that has a length of access tube welded normal to its center to prevent the stand from toppling. Cable clamps are visible.

data storage and display module. The cable allows the probe to be lowered into the tube and stopped at intervals to measure the thermal neutron concentration. Common probe diameters are 38 mm and 51 mm. The readout and control unit are typically attached to a block of high-density polyethylene that acts as the instrument shield. The probe is locked into this shield when not in use. Probes from different manufacturers have different physical arrangements of the source of fast neutrons and the detector, and this arrangement affects the calibration (McCauley and Stone, 1972). The source may be directly beneath the detector or is centered around or on one side of it. The source-detector geometry in modern NMMs has little effect on the attainable accuracy and precision (Allen et al., 1993; Dickey et al., 1993; Evelt and Steiner, 1995). The source in modern meters is a mixture of beryllium and americium-241 with an activity ranging from 0.4 GBq to 1.9 GBq. The ^{241}Am decays, emitting an alpha particle, which then is absorbed by a Be atom, which in turn produces ^{12}C and a fast neutron in the nuclear reaction: $^9\text{Be}(\alpha, n)^{12}\text{C}$. The sensed volume is approximately a sphere with radius, R (cm), that decreases with increases in volumetric water content (θ_v , $\text{m}^3 \text{m}^{-3}$), according to

$$R = 15(\theta_v)^{-1/3} \quad (1)$$

About 95% of the measured slow neutrons are from this volume (IAEA, 1970).

Access tubes and depth control

Successfully employed access tubing materials include stainless steel, mild steel, galvanized steel, polyvinyl chloride (PVC), polycarbonate, and polyethylene plastics, and aluminum. Calibration is affected by all access tube materials but more so by the hydrogen in all plastics and the neutron absorber chlorine in PVC tubes. The neutron absorber iron affects calibration in steel tubes, while aluminum is nearly transparent to neutrons. Thus, it is imperative that a NMM be calibrated in the same tubing as will be used in the field. Although calibration precision decreases slightly if plastic tubes are used (Allen et al., 1993), precision and accuracy are much more dependent on the tube installation and calibration methods employed than on tube material. Recommendations for installation of access tubes are given in Hignett and Evelt (2002) and Evelt et al. (2008) give recommendations for choice of access tube material and installation methods.

Access tubes are commonly installed vertically with a few cm of tubing protruding above the soil surface, mostly to prevent soil and water from entering the tube, which is further ensured by use of a tube covering or plug. The common practice of placing the NMM on top of the access tube near the soil surface before lowering the probe for readings is not recommended for two reasons. First, when the NMM is placed near the soil surface, the shield in the meter body may influence near-surface counts to a degree that depends strongly on the height of the meter above the soil (Stone et al., 1993). Second, in field use, the height of access tubes above the soil is likely to change with tillage, rainfall induced settling, erosion or deposition, or other factors, resulting in an equivalent change in the depth of probe placement. For readings above 0.3-m depth, centimeter scale differences in the depth of the probe will

strongly influence the neutron count and the calibration equation due to loss of neutrons to the atmosphere (Grant, 1975; Van Bavel et al., 1961).

Evetts et al. (2003) showed how these problems are addressed using a depth control stand. The stand comprises a length of metal tubing, the same size as the access tube, fixed to a 0.2-m length of slightly larger tubing that is in turn supported by a foot resting directly on the soil surface (Fig. 1A). The larger diameter of the lower length of tubing allows it to be slipped over the top of an access tube. Cable stops are arranged to achieve the desired depths of placement of the probe, thus maintaining the reading depths at exact distances relative to the soil surface. Standard counts taken with the probe locked in the shield are routinely taken during NMM calibration and field use. Standards taken with the meter too close to the soil surface will vary with the water content of the soil (Allen and Segura, 1990; Dickey, 1990). The stand described is tall enough (>80 cm) to be suitable for taking standard counts without interference from soil water content, with the NMM mounted on the stand and the probe locked in the meter shield. Using these tools, accurate calibration and subsequent readings may be achieved at depths as shallow as 10 cm (Table 1) (Evetts and Steiner, 1995).

NMM statistics

The Poisson probability distribution describes the random process of neutron emission from the source. For processes described by the Poisson distribution, a series of counts over equal time periods will have a standard deviation that is equal to the square root of the mean value. One result of this fact is that the coefficient of variation of counts can be reduced by increasing the counting time. The sample mean, m , is computed as

$$m = \frac{1}{N} \sum_{i=1}^N x_i \quad (2)$$

where x_i is the value of a single count taken over a time period and N is the number of counts (all taken with the probe in one position and for the same length of time). The sample standard deviation, s , is computed as

$$s = \left[\frac{1}{N-1} \sum_{i=1}^N (x_i - m)^2 \right]^{1/2} \quad (3)$$

For a properly operating meter with the probe in a constant environment, the mean value of the ratio $s/(m)^{1/2}$, called the Chi ratio, should be close to unity. The χ^2 statistic can be used to describe the probability distribution of the ratio:

$$\frac{s}{m^{1/2}} = \left[\frac{\chi^2}{N-1} \right]^{1/2} \quad (4)$$

Values of χ^2 (Chi-squared) for a given probability level (P) are given in statistical tables for different values of $(N-1)$. Writing the right-hand-side of Eq. (4) for the upper and lower limits of χ^2 , we can obtain upper and lower values of the Chi ratio for the chosen probability level and number of samples. For example, for 32 samples and a 95% probability level, the values of χ^2 are 17.5 for $P = .975$ and 48.1 for $P = .025$; and the Chi ratio from Eq. (4) should be between 0.75 and 1.25 about 95 times in every hundred. Some meters divide the count by a fixed number in order to reduce the displayed count to a reasonably small value. In computing Chi ratios for such meters, the user should first multiply the recorded counts by the factor that the meter used to reduce them. Modern NMMs have a standard count function that takes many counts and computes the mean standard count and Chi ratio, making it easy for the user to obtain those values for use in calibration, quality control, and everyday computation of water contents from a calibration equation. Chi ratio values that are repeatedly more than unity or less than unity are cause for concern and likely will require repair of the NMM, as will large and consistent deviations of the standard count from the mean of previous standard counts. A consequence of the Chi ratio tending to unity is that the standard deviation decreases quickly as count time and the value of m increases (Fig. 2).

Table 1 Calibrations of water content (θ_v , $\text{m}^3 \text{m}^{-3}$) vs. count ratio (C_R) for the Amarillo fine sandy loam using the methods in Evetts and Steiner (1995), including a depth control stand.

Depth (cm)	Equation	RMSE	r^2	N
10	$\theta_v = 0.014 + 0.2172C_R$	0.004	0.997	6
30–190	$\theta_v = -0.063 + 0.2371C_R$	0.007	0.988	44
30–90	$\theta_v = -0.066 + 0.2421C_R$	0.008	0.988	24
110–190	$\theta_v = -0.057 + 0.2299C_R$	0.006	0.992	20

RMSE is root mean squared error, N is the number of samples, and r^2 is the coefficient of determination for the regression analysis. From Evetts SR and Steiner JL (1995) Precision of neutron scattering and capacitance type moisture gages based on field calibration. *Soil Science Society of America Journal* 59: 961–968. doi: 10.2136/sssaj1995.03615995005900040001x.

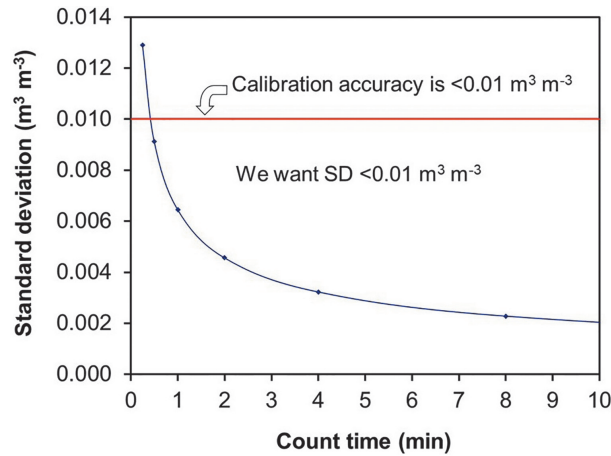


Fig. 2 Decrease in the standard deviation of the mean count as the count time increases for a particular model of neutron moisture meter. Accuracy of field calibration with this meter is typically $<0.01 \text{ m}^3 \text{ m}^{-3}$. One can use a count time of 1 min to maintain equivalent precision.

NMM calibration, accuracy, and representativity

Manufacturers provide NMM calibration equations based on measurements taken in standard samples, but these are seldom useful for soil water content determination (e.g., Dickey, 1990). Calibration of NMMs involves taking neutron count measurements, x , with the probe at different depths below the soil surface and standard counts, x_s , with the probe locked in the meter shield, and then computing the count ratio, C_R , defined as

$$C_R = x/x_s \quad (5)$$

for each depth. Correlating the measured count ratio values with independently determined volumetric water content, θ_v ($\text{m}^3 \text{ m}^{-3}$), values results in a calibration equation. For modern meters and the normal range of values of soil water content, the calibration is linear and of the form

$$\theta_v = b_0 + b_1 C_R \quad (6)$$

where b_0 and b_1 are the calibration coefficients as determined by linear regression, Count ratio values are used because the source activity and thus counts will decline over time, and because the detector efficiency is somewhat temperature dependent (Evet, 2000a). Recommendations for taking standard counts are given in Hignett and Evett (2002) and Evett et al. (2008), as are recommendations for field calibration using the wet site—dry site method of Evett and Steiner (1995). Careful field calibrations done using the wet site—dry site method and the depth control stand should attain root mean squared errors $<0.01 \text{ m}^3 \text{ m}^{-3}$ and r^2 values greater than 0.9, even for depths near the surface (e.g., 10 cm in Table 1).

Although the NMM has been proven to be an accurate means of assessing soil profile water storage, changes in storage, and crop water use (using the soil water balance equation), the individual depth-wise water content values should be well understood as being representative of a large volume, not a point. Fig. 3 illustrates this with a comparison of depth-wise water content data from two NMM access tubes compared with the depth-wise water content values from a system of three profiles of TDR sensors (see Time Domain Reflectometry article). Both the NMM and TDR sensors were calibrated for the soil in question. The top five TDR data are from depths of 2, 6, 10, 20, and 30 cm while the top two NMM values are from depths of 10 cm and 30 cm. As can be seen, the 10-cm NMM value provides close to an average of TDR water contents from the surface to 30 cm, and the 30-cm NMM values provide close to an average of TDR water contents from the 20 cm to 50 cm depths. Despite the differences in depth-wise values, profile water content calculated from the NMM and TDR methods over multiple dates was strongly correlated ($r^2 = 0.95$). Due to the large sensing volume of the NMM relative to that of electromagnetic sensing technologies, such as TDR, the number of neutron probe access tubes needed to represent the soil profile water content and change in storage for a small (tens of meters) experimental plot may be as small as one (Evet et al., 2009). For larger fields, several neutron probe access tubes may be needed to represent the field mean water content change in storage for accurate calculation of field mean evapotranspiration (Evet et al., 2012).

NMM safety and use considerations

Safety concerns include radiation safety and back and knee strains incurred during repeated bending and kneeling to operate meters placed on access tubes. The depth control stand described above has virtually eliminated physical injuries where it is used because it allows users to work standing up. Due to the small levels of radioactivity involved, the ALARA (As Low As Reasonably Achievable) principle of reducing exposure guides most radiation safety rules. Radiation received by users may be reduced by increasing distance

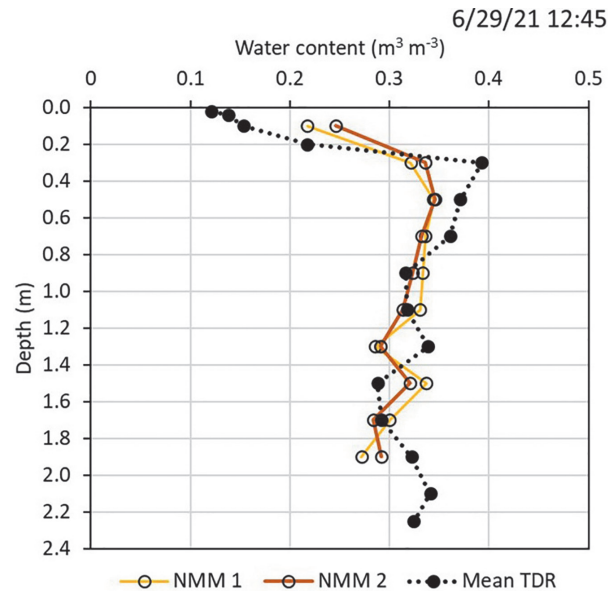


Fig. 3 Profiles of depth-wise water content values from a NMM used in two access tubes and of the mean of three nearby TDR sensors at each depth. All were in a monolithic weighing lysimeter. Both NMM and TDR sensors had soil-specific calibrations.

from the meter, decreasing time spent near the meter, and increasing shielding. The probe should always be locked into the shield except when it is lowered into an access tube. Users should be made aware that the source emits radiation at all times, even when the meter is turned off and batteries removed. Guidelines for ALARA use of the NMM are found in [Evetts \(2000b\)](#). The International Atomic Energy Agency maintains an Internet site of useful information on radiation safety and hazardous materials transport (HYPERLINK "<http://www.iaea.org/worldatom/>" <http://www.iaea.org/worldatom/>).

Due to regulation, the NMM method is not usable for automatic measurements. Due to its large measurement volume, the method is inappropriate where detailed vertical definition is required. This can be particularly important near the surface where water content often changes rapidly with depth. In such cases, the NMM can be used for deeper measurements in conjunction with time domain reflectometry (TDR) measurement of the near-surface soil water content ([Evetts et al., 1993](#)). The time and effort required to install access tubes and calibrate for each soil type is nontrivial. There is also a substantial cost for the equipment and for necessary training and licenses to handle and transport radioactive materials.

Radiation safety concerns are not pertinent to the CRNS method because the radiation involved is naturally occurring. The cost of CRNS equipment is larger than that of NMM equipment but CRNS equipment is easily transported for mobile applications without radiation safety concerns.

Cosmic ray neutron scattering

The COSMOS (Cosmic Ray Soil Moisture Observing System) cosmic ray-induced neutron detection system was introduced in 2008–2010 ([Desilets et al., 2010](#); [Shuttleworth et al., 2010](#); [Zreda et al., 2008](#)). Many later publications have referred to the method used in COSMOS as the Cosmic Ray Neutron Scattering (CRNS) method. The method has been used to map surface soil water variations using a tomographic approach ([Franz et al., 2015](#)), and is promising for testing of satellite remote sensing-based soil water products ([Crow et al., 2012](#)). The flux rate of fast neutrons produced by collisions of cosmic rays with atoms in the atmosphere near the ground is known to be affected by altitude and natural fluctuations in the cosmic ray flux rate, so the method depends on continuous knowledge of the cosmic ray flux rate at the specific geographic location and altitude where the method is deployed. This knowledge is provided by operators of CRNS networks. CRNS equipment typically consists of, at minimum, two thermal neutron detectors placed approximately 1 m above the ground surface, one of which is unshielded and acts to count thermalized neutrons that pass into it. The other detector is shielded with high density polyethylene plastic that prevents thermal neutrons from passing through to the detector. Fast neutrons are scattered in the shield such that some of them pass into the detector, which thus acts to count the fast neutron flux. The fast neutron flux is reduced as soil water content increases because fast neutrons are thermalized through interaction with hydrogen in the soil water. Thus, the fast neutron count is inversely proportional to soil water content (Eq. 7). Neutron detections are counted by a high-speed electronic counter, the data are temporarily stored in an embedded computer, and then transmitted, typically by satellite transmission, to the network base station where data are stored, adjusted for cosmic ray intensity at the CRNS unit location, and made available to the user. The CRNS method may be deployed as a static installation (e.g., [Fig. 4](#)) or as a mobile, even backpack, deployment.

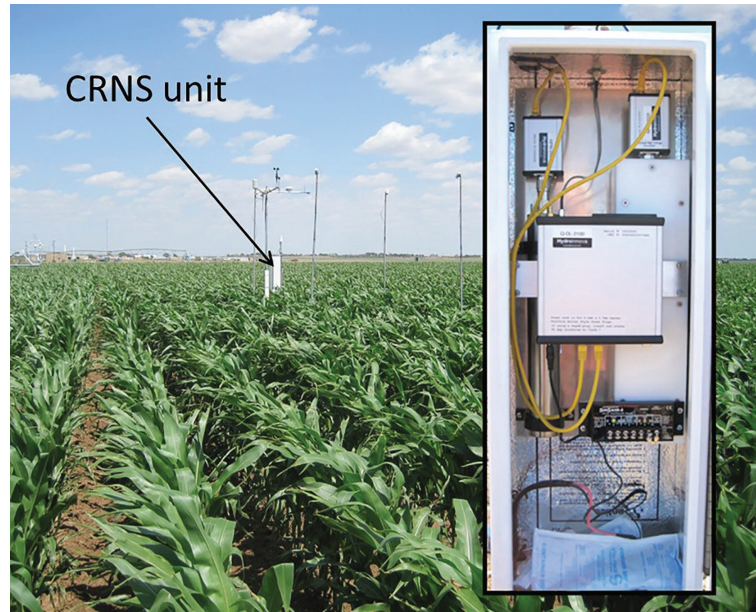


Fig. 4 CRNS equipment installed at a large weighing lysimeter at the center of a 220-m square corn field. The inset photo shows a COSMOS CRNS unit with thermal (left) and fast neutron (right) detectors, counters, and embedded computer. In the field photograph the COSMOS unit is the white vertical rectangular object to left of center.

The CRNS method responds to the soil water content near the surface and to the near-surface atmospheric humidity and vegetation over a large area (100 s of meters). Assuming that atmospheric humidity effects are negligible, soil water content, θ_v , at time t may be inferred using

$$\theta_v(t) = \rho_{bd} \{a_o [F(t) \times N/N_0(t)] - a_1\}^{-1} - a_2\} - \theta_{LW} - \theta_B - \theta_{SOC} \quad (7)$$

where ρ_{bd} is the mean dry soil bulk density, N is the time dependent fast neutron count rate, N_0 is the time dependent fast neutron count over dry soil, θ_{LW} is the soil lattice water content, θ_B is the biomass water content equivalent, θ_{SOC} is the soil organic carbon water content equivalent, and a_o , a_1 and a_2 are fitting parameters with values of 0.0869, 0.3720, and 0.1236, respectively. The correction factor, $F(t)$, accounts for variable absolute humidity, solar activity, and absolute humidity of the air (IAEA, 2017). Rosolem et al. (2013) described a correction for absolute humidity of the air.

The depth of soil sensed is specific to the soil and is inversely related to its water content. Shuttleworth et al. (2010), using Monte Carlo analysis, determined an effective depth of sensing to be <0.15 m for wet soil and >0.7 m for dry soil; and they determined that an effective sensing diameter was between 650 m and 700 m and relatively independent of soil water content. Later studies have shown considerable smaller radii and depth of sensing (Franz et al., 2012a; Köhli et al., 2015). Except quite near the installation site, the CRNS method does not respond to soil water changes deeper than ~20 cm. Köhli et al. (2015) found that the sensing radius was in the range of 130–240 m depending on soil water content, vegetation, and humidity of the air; and that the depth of sensing ranged from 15 cm to 83 cm, decreasing quickly and exponentially with distance from the sensor. For example, in experiments at Bushland, Texas, the CRNS did not respond to subsurface drip irrigation amounts ranging from 12 mm to 50 mm where the drip tapes were installed at depths of 32–34 cm. Recent work has attempted to reduce the spatial weighting of CRNS data, which is caused by the radially decreasing sensing depth (Scheffele et al., 2020), thereby turning it into an unweighted integral soil water content over the support volume that would be used more easily in other applications. The reduction of spatial weighting did, however, require the installation of at least one and preferably three or more accurate soil water sensing profiles at distances <100 m from the CRNS and with three or more sensors at depths from just below the surface (e.g., 5 cm) down to at least 50-cm depth. Several studies have focused on the influence of above-ground biomass on the CRNS signal (Hornbuckle et al., 2012; Franz et al., 2013; Bogena et al., 2013) and found that the hydrogen in biomass decreases the thermal neutron intensity and the accuracy of the CRNS. Subsequently, research investigated the use of empirical corrections (Baatz et al., 2015) and the ratio of thermal to fast neutrons to correct for biomass (Tian et al., 2016; Jakobi et al., 2018; Vather et al., 2020). The precision of N is described by Eq. (3), so precision may be improved by increasing either the counting time or using additional or improved detectors, both of which increase the total count and so reduce the standard deviation. Mobile CRNS equipment typically includes multiple or improved detectors for this reason. For static installations, precision may be improved by integrating counts over 12-h or longer periods, for example by using a running average as was done in COSMOS.

Despite the impediments discussed here, the CRNS method is rapidly maturing and is recognized as an important approach to bridging the gap between satellite derived soil water content data products and the point-scale or nearly point-scale soil water data provided by networks of in-situ electromagnetic sensors and the neutron probe. Integration of data from networks of CRNS units

and networks of in situ sensors may be a powerful tool for better understanding of soil-plant-atmosphere dynamics and watershed hydrology (Fersch et al., 2020).

Outlook

Advances in neutron detection continue and will have a strong impact on the neutron scattering method for soil water monitoring for multiple reasons, including appreciable reduction of the cost relative to commercially available ^3He and BF_3 based detectors. Boron lined proportional counters have proven effective (Weimar et al., 2020). This detector comprises a layer of $^{10}\text{B}_4\text{C}$ deposited on a copper foil that is used to line the inside of an Al tube. Multiple tubes may be used to increase neutron count rate and decrease SD. The detector is not yet commercially available. Flynn et al. (2021) described a ^6Li foil detector that is appreciably less expensive than available ^3He and BF_3 based detectors and is commercially available. Field comparison with a commercially available ^3He based CRNS showed similar performance. Another detector design based on a plastic scintillometer wrapped with thermal neutron detectors has been demonstrated by Stevanato et al. (2019). All three emerging neutron detection systems can provide additional spectral information about neutron energy that is postulated to be useful in discriminating between vegetative, atmospheric, and soil water, which could improve the usefulness of the CRNS for soil water monitoring.

References

- Allen RG and Segura D (1990) Access tube characteristics and neutron meter calibration. In: Harris SR (ed.) *Irrigation and Drainage, Proceedings of the 1990 National Conference, Durango, CO, 11–13 July 1990, New York, NY, USA: American Society of Civil Engineers*, pp. 21–31.
- Allen RG, Dickey G, and Wright JL (1993) Effect of moisture and bulk density sampling on neutron moisture gauge calibration. In: Allen RG and Neale CMU (eds.) *Management of Irrigation and Drainage Systems, Integrated Perspectives, Proceedings of the 1993 ASCE National Conference on Irrigation and Drainage Engineering, Park City, UT, July 21–23, 1993, New York, NY, USA, American Society of Civil Engineers*, pp. 1145–1152.
- Baatz R, Bogen HR, Franssen JH, Huisman JA, Montzka C, and Vereecken H (2015) An empirical vegetation correction for soil water content quantification using cosmic ray probes. *Water Resources Research* 51: 2030–2046. <https://doi.org/10.1002/2014WR016443>.
- Belcher DJ, Cuykendall TR, and Sack HS (1950) *The Measurement of Soil Moisture and Density by Neutron and Gamma-Ray Scattering. Tech. Development Report No. 127(1950)*. Washington, DC: U.S. Civil Aeronautics Administration.
- Bogen HR, Huisman JA, Baatz R, Hendricks Franssen HJ, and Vereecken H (2013) Accuracy of the cosmic-ray soil water content probe in humid forest ecosystems: The worst case scenario. *Water Resources Research* 49: 5778–5791. <https://doi.org/10.1002/wrcr.20463>.
- Crow WT, Berg AA, Cosh MH, Loew A, Mohanty BP, Panciera R, de Rosnay P, Ryu D, and Walker JP (2012) Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products. *Reviews of Geophysics* 50: RG2002. <https://doi.org/10.1029/2011RG000372>.
- Desilets D and Zreda M (2003) Spatial and temporal distribution of secondary cosmic-ray nucleon intensities and applications to in-situ cosmogenic dating. *Earth and Planetary Science Letters* 206: 21–42. [https://doi.org/10.1016/S0012-821X\(02\)01088-9](https://doi.org/10.1016/S0012-821X(02)01088-9).
- Desilets D and Zreda M (2013) Footprint diameter for a cosmic-ray soil moisture probe: Theory and Monte Carlo simulations. *Water Resources Research* 49(6): 3566–3575.
- Desilets D, Zreda M, and Ferré TPA (2010) Nature's neutron probe: Land surface hydrology at an elusive scale with cosmic rays. *Water Resources Research* 46: W11505. <https://doi.org/10.1029/2009WR008726>.
- Dickey GL (1990) Factors affecting neutron gauge calibration. In: Harris SR (ed.) *Irrigation and Drainage, Proceedings of the 1990 National Conference, Durango, CO, 11–13 July 1990, New York, NY, USA: American Society of Civil Engineers*, pp. 9–20.
- Dickey G, Allen RG, Wright JL, Murray NR, Stone JF, and Hunsaker DJ (1993) Soil bulk density sampling for neutron gauge calibration. In: Allen RG and Neale CMU (eds.) *Management of Irrigation and Drainage Systems, Integrated Perspectives, Proceedings of the 1993 ASCE National Conference on Irrigation and Drainage Engineering, Park City, UT, July 21–23, 1993, New York, NY, USA: American Society of Civil Engineers*, pp. 1103–1111.
- Evelt SR (2000a) Some Aspects of Time Domain Reflectometry (TDR), Neutron Scattering, and Capacitance Methods of Soil Water Content Measurement. In: *Comparison of Soil Water Measurement Using the Neutron Scattering, Time Domain Reflectometry and Capacitance Methods. IAEA-TECDOC-1137*, pp. 5–49. Vienna: International Atomic Energy Agency. <https://www.iaea.org/publications/5932/comparison-of-soil-water-measurement-using-the-neutron-scattering-time-domain-reflectometry-and-capacitance-methods>.
- Evelt SR (2000b) *Nuclear Gauge Module I, Design, Theory, and Operation*. Beltsville, MA: Nuclear Gauge Train-the-Trainer Course, USDA-Radiation Safety Staff.
- Evelt SR and Steiner JL (1995) Precision of neutron scattering and capacitance type moisture gages based on field calibration. *Soil Science Society of America Journal* 59: 961–968. <https://doi.org/10.2136/sssaj1995.03615995005900040001x>.
- Evelt SR, Howell TA, Steiner JL, and Cresap JL (1993) Evapotranspiration by soil water balance using TDR and neutron scattering. In: Allen RG and Neale CMU (eds.) *Management of Irrigation and Drainage Systems, Integrated Perspectives, Proceedings of the 1993 ASCE National Conference on Irrigation and Drainage Engineering, Park City, UT, July 21–23, 1993, New York, NY, USA, American Society of Civil Engineers*, pp. 914–921.
- Evelt SR, Tolk JA, and Howell TA (2003) A depth control stand for improved accuracy with the neutron probe. *Vadose Zone Journal* 2(4): 642–649. <https://doi.org/10.2136/vzj2003.6420>.
- Evelt SR, Heng LK, Moutonnet P, and Nguyen ML (eds.) (2008) *Field Estimation of Soil Water Content: A Practical Guide to Methods, Instrumentation, and Sensor Technology*. Vienna, Austria: International Atomic Energy Agency IAEA-TCS-30. http://www-pub.iaea.org/MTCD/Publications/PDF/TCS-30_web.pdf.
- Evelt SR, Schwartz RC, Tolk JA, and Howell TA (2009) Soil profile water content determination: Spatiotemporal variability of electromagnetic and neutron probe sensors in access tubes. *Vadose Zone Journal* 8(4): 926–941. <https://doi.org/10.2136/vzj2008.0146>.
- Evelt SR, Schwartz RC, Howell TA, Baumhardt RL, and Copeland KS (2012) Can weighing lysimeter ET represent surrounding field ET well enough to test flux station measurements of daily and sub-daily ET? *Advances in Water Resources* 50: 79–90. <https://doi.org/10.1016/j.advwatres.2012.07.023>.
- Fersch B, Francke T, Heistermann M, Schrön M, Döpfer V, Jakobi J, Baroni G, Blume T, Bogen HR, Budach C, Gränzig T, Förster M, Güntner A, Hendricks Franssen H-J, Kasner M, Köhli M, Kleinschmit B, Kunstmann H, Patil A, Rasche D, Scheffele L, Schmidt U, Szulc-Seyfried S, Weimar J, Zacharias S, Zreda M, Heber B, Kiese R, Mares V, Mollenhauer H, Völkisch I, and Oswald S (2020) A dense network of cosmic-ray neutron sensors for soil moisture observation in a highly instrumented pre-Alpine headwater catchment in Germany. *Earth System Science Data* 12: 2289–2309. <https://doi.org/10.5194/essd-12-2289-2020>.
- Flynn KD, Wyatt BM, and McInnes KJ (2021) Novel cosmic ray neutron sensor accurately captures field-scale soil moisture trends under heterogeneous soil textures. *Water* 13: 3038. <https://doi.org/10.3390/w13213038>.
- Franz TE, Zreda M, Ferré TPA, Rosolem R, Zweck C, Stillman S, Zeng X, et al. (2012a) Measurement depth of the cosmic ray soil moisture probe affected by hydrogen from various sources. *Water Resources Research* 48(8): W08515.

- Franz TE, Zreda M, Rosolem R, and Ferré TPA (2012b) Field validation of a cosmic-ray neutron sensor using a distributed sensor network. *Vadose Zone Journal* 11(4). <https://www.soils.org/publications/vzj/pdfs/11/4/vzj2012.0046>.
- Franz TE, Zreda M, Rosolem R, Hornbuckle BK, Irvin SL, Adams H, Kolb TE, Zweck C, and Shuttleworth WJ (2013) Ecosystem-scale measurements of biomass water using cosmic ray neutrons. *Geophysical Research Letters* 40: 3929–3933. <https://doi.org/10.1002/grl.50791>.
- Franz TE, Wang T, Avery W, Finkenbinder C, and Brocca L (2015) Combined analysis of soil moisture measurements from roving and fixed cosmic ray neutron probes for multiscale real-time monitoring. *Geophysical Research Letters* 42: 3389–3396. <https://doi.org/10.1002/2015GL063963>.
- Grant DR (1975) Measurement of soil moisture near the surface using a moisture meter. *Journal of Soil Science* 26: 124–129.
- Hignett C and Evett SR (2002) Neutron thermalization. In: Dane JH and Topp GC (eds.) *Methods of Soil Analysis, Part 4 – Physical Methods*, pp. 501–521. Madison, WI: Soil Science Society of America. Section 3.1.3.10.
- Hornbuckle B, Irvin S, Franz T, Rosolem R, and Zweck C (2012) The potential of the COSMOS network to be a source of new soil moisture information for SMOS and SMAP. In: *Proceedings of the 2012 IEEE International Geoscience and Remote Sensing Symposium, Munich, Germany, 22–27 July 2012*.
- IAEA (1970) *Neutron Moisture Gauges*. Tech. Rep. Ser. No. 112 Vienna, Austria: International Atomic Energy Agency.
- IAEA (2000) *Comparison of Soil Water Measurement Using the Neutron Scattering, Time Domain Reflectometry and Capacitance Methods*. IAEA-TECDOC-1137 Vienna, Austria: International Atomic Energy Agency. <https://www.iaea.org/publications/5932/comparison-of-soil-water-measurement-using-the-neutron-scattering-time-domain-reflectometry-and-capacitance-methods>.
- IAEA (2017) *Cosmic Ray Neutron Sensing: Use, Calibration and Validation for Soil Moisture Estimation*. IAEA-TECDOC-1809 Vienna, Austria: International Atomic Energy Agency. https://www-pub.iaea.org/MTCD/Publications/PDF/TE-1809_web.pdf.
- Jakobi J, Huisman JA, Vereecken H, Diekkruiger B, and Bogaen HR (2018) Cosmic ray neutron sensing for simultaneous soil water content and biomass quantification in drought conditions. *Water Resources Research* 54: 7383–7402. <https://doi.org/10.1029/2018WR022692>.
- Köhli M, Schrön M, Zreda M, Schmidt U, Dietrich P, and Zacharias S (2015) Footprint characteristics revised for field-scale soil moisture monitoring with cosmic-ray neutrons. *Water Resources Research* 51: 5772–5790. <https://doi.org/10.1002/2015WR017169>.
- McCauley GN and Stone JF (1972) Source-detector geometry effect on neutron probe calibration. *Soil Science Society of America Proceedings* 36: 246–250.
- Nakayama FS and Allen SG (1990) Application of neutron soil surface water monitoring for plant establishment. In: Harris SR (ed.) *Irrigation and Drainage, Proceedings of the 1990 National Conference, Durango, CO, 11–13 July 1990, New York, NY, USA, American Society of Civil Engineers*, pp. 210–217.
- Pieper GF (1949) *The Measurement of Moisture Content of Soil by the Slowing of Neutrons*. Master's thesis. Ithaca, NY: Cornell University.
- Rosolem R, Shuttleworth WJ, Zreda M, Franz TE, Zeng X, and Kurc SA (2013) The effect of atmospheric water vapor on neutron count in the cosmic-ray soil moisture observing system. *Journal of Hydrometeorology* 14(5): 1659–1671. <https://doi.org/10.1175/JHM-D-12-0120.1>.
- Scheffele LM, Baroni G, Franz TE, Jakobi J, and Oswald SE (2020) A profile shape correction to reduce the vertical sensitivity of cosmic-ray neutron sensing of soil moisture. *Vadose Zone Journal* 19: e20083. <https://doi.org/10.1002/vzj2.20083>.
- Shuttleworth WJ, Zreda M, Zeng Z, Zweck C, and Ferré TPA (2010) The COsmic-ray Soil Moisture Observing System (COSMOS): A non-invasive, intermediate scale soil moisture measurement network. In: *BHS Third International Symposium, Managing Consequences of a Changing Global Environment, Newcastle 2010, British Hydrological Society*.
- Stevanato L, Baroni G, Cohen Y, Fontana CL, Gatto S, Lunardon M, Marinello F, Moetto S, and Morselli L (2019) A novel cosmic-ray neutron sensor for soil moisture estimation over large areas. *Agriculture* 9: 202. <https://doi.org/10.3390/agriculture9090202>.
- Stone JF, Allen RG, Gray HR, Dickey GL, and Nakayama FS (1993) Performance factors of neutron moisture probes related to position of the source on the detector. In: Allen RG and Neale CMU (eds.) *Management of Irrigation and Drainage Systems, Integrated Perspectives, Proceedings of the 1993 ASCE National Conference on Irrigation and Drainage Engineering, Park City, UT, July 21–23, 1993, New York, NY, USA: American Society of Civil Engineers*, pp. 1128–1135.
- Tian Z, Li Z, Liu G, Li B, and Ren T (2016) Soil water content determination with cosmic-ray neutron sensor: Correcting aboveground hydrogen effects with thermal/fast neutron ratio. *Journal of Hydrology* 540: 923–933. <https://doi.org/10.1016/j.jhydrol.2016.07.004>.
- Van Bavel CHM, Underwood N, and Swanson RW (1956) Soil moisture measurement by neutron moderation. *Soil Science* 82: 29–41.
- Van Bavel CHM, Nielsen DR, and Davidson JM (1961) Calibration and characteristics of two neutron moisture probes. *Soil Science Society of America Proceedings* 25(5): 329–334.
- Vather T, Everson CS, and Franz TE (2020) The applicability of the cosmic ray neutron sensor to simultaneously monitor soil water content and biomass in an *Acacia mearnsii* forest. *Hydrology* 7: 48. <https://doi.org/10.3390/hydrology7030048>.
- Weimar J, Köhli M, Budach C, and Schmidt U (2020) Large-scale boron-lined neutron detection systems as a ^3He alternative for Cosmic Ray Neutron Sensing. *Frontiers in Water* 2: 16. <https://doi.org/10.3389/frwa.2020.00016>.
- Zreda M, Desilets D, Ferre TPA, and Scott R (2008) Measuring soil moisture content non-invasively at intermediate spatial scale using cosmic-ray neutrons. *Geophysical Research Letters* 35(21). <https://doi.org/10.1029/2008GL035655>.

Time-domain reflectometry

GC Topp^a and Ty Ferré^b, ^aAgriculture and Agri-Food Canada, Ottawa, ON, Canada; ^bHydrology and Atmospheric Sciences, University of Arizona, Tucson, AZ, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of G.C. Topp, T.P.A. Ferré, TIME-DOMAIN REFLECTOMETRY, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 174–181, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00508-7>.

Introduction	437
Measurement of soil water content using TDR	437
Instrumentation and wave-guides	440
Impact of salinity on water content measurement	441
Measurement of bulk electrical conductivity using TDR	442
Impact of water content on pore-water salinity measurement	442
Conclusion	442
References	443

Abstract

Time-domain reflectometry (TDR) is an electromagnetic (EM) technique using radar principles at radio frequency to estimate water content and electrical conductivity of soil. Buried parallel rods are inserted into soil, with the velocity of the TDR wave affected by the soil water content. In addition, the magnitude of the wave is decreased by electrical losses associated with the bulk electrical conductivity of the soil. From the dielectric properties of the soil, equations allow for the TDR waveform to be converted to water content and electrical conductivity. TDR has been combined with other measurements such as with tensiometers and the cone penetrometer.

Key points

- TDR is a sensor approach to estimate water content and electrical conductivity of soil.
- As a TDR wave propagates through soil, guided by the probe rods, its velocity is decreased by the presence of water.
- Equations to convert from TDR waveform to water content and electrical conductivity are described.
- TDR has become one of the most widely used soil water sensor approaches.

Nomenclature

c	Velocity of light (m s^{-1})
L	Length (m)
t	Time(s)
v	Propagation velocity of electromagnetic waves (m s^{-1})
V	Voltage (volt)
Z_0	Characteristic impedance of probe (ohms)
Z_i	Output impedance of instrument (ohms)
α	Attenuation coefficient
ϵ''_r	Imaginary component of relative dielectric permittivity (relative dielectric loss)
ϵ'_r	Real component of relative dielectric permittivity (dielectric constant)
ϵ_0	Dielectric permittivity of free space (F m^{-1})
ϵ_r	Relative dielectric permittivity (dielectric constant)
ϵ_{ra}	Apparent relative dielectric permittivity (dielectric constant)
θ_v	Volumetric soil water content ($\text{m}^3 \text{m}^{-3}$)
P	Reflection coefficient
σ_0	Electrical conductivity (S m^{-1})
ω	Frequency of signal ($2\pi \text{s}^{-1}$)

Introduction

The vital role of water in maintaining the life of the landscape requires the development of techniques to monitor and sustain water supply and its quality (Clothier and Green, 1997). Time-domain reflectometry (TDR) is an established technology for use in soil and landscape processes, which has become widely used in highly diversified applications (Lekshmi et al., 2014; Binley et al., 2015). TDR offers rapid, in situ, nondestructive measurement of soil water content and ionic solute concentration. This is an electromagnetic (EM) technique using radar principles at radio frequency where estimates of water content and electrical conductivity of soil can be made separately from the same wave (Dowding, 2001). The effect of the soil and water on the propagation velocity of the EM wave is analyzed to provide a reliable measure of the water content. The bulk electrical conductivity (EC) is determined from an analysis of the rate of decrease in amplitude during the propagation of the wave in soil (Topp and Reynolds, 1998). The principles of EM wave propagation are also used for the design of soil probes and to indicate the precision of measurements. The strong interaction between water content and electrical conductivity usually hinders the independent and separate determination of these properties in soil. This difficulty is overcome with TDR where both measurements are from the same sampled region and with the same EM wave.

Water molecules have unique electrical properties that determine the electrical properties of soil (Topp et al., 1980). Because of the strong dependence of soil electrical properties on the amount of water in a soil and on the quality of that water, electrical and electromagnetic measurements can be very useful for characterizing the soil water content. One method with which these properties can be measured is TDR. TDR can be thought of as one-dimensional radar using radio waves that propagate along a wave-guide. This wave-guide is constructed from parallel metal rods that are inserted in the soil. As radio waves travel through any material, the characteristics of the wave are continuously changed by the medium through which the wave is traveling. These changes in wave properties are measured by TDR and a wave equation analysis is used to estimate two electrical parameters of the soil: dielectric permittivity (dielectric constant) and electrical conductivity. These properties are then used to estimate the water content and the electrical conductivity of the soil, respectively (Ferré and Topp, 2000).

As the name implies, time-domain reflectometry relies on the measurement of the travel time of reflected waves. In this way, it is similar to other radar applications. In TDR, the wave velocity is measured by recording the time difference between the entry of a wave into the soil and the return of the reflected wave from the end of the TDR probe. Currently available TDR instruments operate in the range of radio frequencies (10–1000 MHz). In this frequency range, the dependence of the wave velocity on the water content is not strongly influenced by the electrical conductivity of the pore water or the electrical properties of the soil solids. This is supported by direct measurements made since the mid-1970s that show that, while other soil factors such as density, texture, temperature, and soluble salts affect the TDR wave velocity, their effects are less pronounced than are the effects of water content (Lekshmi et al., 2014). That is, the soil water content is the most important soil factor affecting the travel time measured by TDR. This property of TDR makes the method very effective and efficient for the measurement of soil water content.

Soil salinity is an issue of great importance in relation to irrigation and agricultural production in arid and semiarid regions. It has long been realized that the electrical conductivity of pore water can be related to pore-water salinity. However, the bulk soil electrical conductivity is not solely dependent upon the concentration and composition of salts in solution. This property is also highly dependent on the soil texture, the temperature, and the water content. TDR can be used to infer the bulk electrical conductivity of a soil. The energy loss of the transmitted signal is related to the bulk electrical conductivity of the soil. This energy loss is determined from the relative amplitudes of the signal source and that of the wave reflected from the end of the probe. TDR has the advantage of simultaneously measuring both the water content and the bulk electrical conductivity in approximately the same volume of porous medium. Under some conditions, this information can be used to infer the bulk electrical conductivity, and thereby the salinity, of the pore water (Wraith, 2002).

In the 40-plus years since TDR was first applied to soil measurements (Topp et al., 1980), the method has become a de facto standard for soil water content measurement and monitoring (Lekshmi et al., 2014). The popularity of the method for environmental monitoring and research arises from a combination of its accuracy in a wide range of soils and its relative ease of use compared with many other available techniques (Ferré and Topp, 2002). TDR provides real-time, in-situ soil water content measurements. Multiple soil probes can be incorporated into a switching network connected to a single instrument and data logger allowing for remote automated monitoring. For most soils, the accuracy of measurements of volumetric water content change is within $\pm 0.02 \text{ m}^3 \text{ m}^{-3}$ without the need for soil-specific calibration, and similar absolute water contents can be achieved with calibration. There is considerable flexibility in the design and placement of TDR probes, allowing users to modify water content measurement networks to conform to the requirements of any specific study. Finally, because TDR measures the volumetric water content, the data are directly applicable to hydrologic water balance analyses with no need for the measurement of supporting soil parameters such as bulk density.

Measurement of soil water content using TDR

Measuring the volumetric water content of a soil is difficult. As a result, methods have been developed that infer the volumetric water content based on the water potential, bulk electrical conductivity, or dielectric permittivity. The water potential–volumetric water content relationship is nonlinear and hysteretic. The bulk electrical conductivity depends on many factors in addition to the

water content. In contrast, the bulk dielectric permittivity of a soil is highly correlated with the volumetric water content. For these reasons, recent advances in the development of indirect methods of measuring the volumetric water content have focused on methods that rely on the measurement of dielectric permittivity. TDR has long been used as a laboratory method to measure the dielectric Permittivity of liquids. In an unrelated field, TDR has been widely applied to the location of faults in buried cables. However, it was not until it was demonstrated that a single correlation existed between the bulk dielectric permittivity and the volumetric water content for a wide range of porous materials that these two distinct uses came together to give rise to the TDR method for soil water content measurement. Since the 1970s, TDR has been applied at many scales and under a broad range of conditions spanning agricultural and forestry environments and seasonal conditions from summer to winter. In addition, the method has been improved and modified.

Most TDR instruments launch a fast-rise voltage step (rise time < 200 ps) along a transmission line that is connected to a probe that is inserted in the soil or medium of interest (O'Connor and Dowding, 1999). The voltage pulse propagates as a planar EM wave, traveling in the soil and guided by the conductors that form the probe. The properties of the soil that govern the propagation of the TDR pulse are described collectively by the propagation constant of the soil, from which velocity of propagation, v (m s^{-1}), and attenuation coefficient (α) are derived. The TDR instrument records and displays the voltage returning to the instrument as a function of elapsed time following the launch of the voltage pulse. This gives rise to a characteristic plot known as the waveform.

The waveform is a record of the voltage that arrives back at the instrument after being reflected from the end of the probe. Commonly, the returning voltage, V , is displayed as a normalized voltage, known as the reflection coefficient (ρ), as a function of time (t) (Fig. 1, where $V/V_0 = 1 - \rho$), where V_0 is the voltage of the transmitted pulse. From the TDR waveform it is possible to determine both v and α . As discussed above, v is used to infer the volumetric water content while α is used to determine the electrical conductivity. A portion of the TDR step pulse travels a distance, L_{cable} (m), along the connecting cables to the probe and then reflects back to the instrument. A portion of the energy that enters the probe travels an additional distance, L (m), along the probe, and then travels the total distance back to the instrument. Taking the difference in these travel times to be t (s), the velocity of the wave can be determined from the known probe length, L , as:

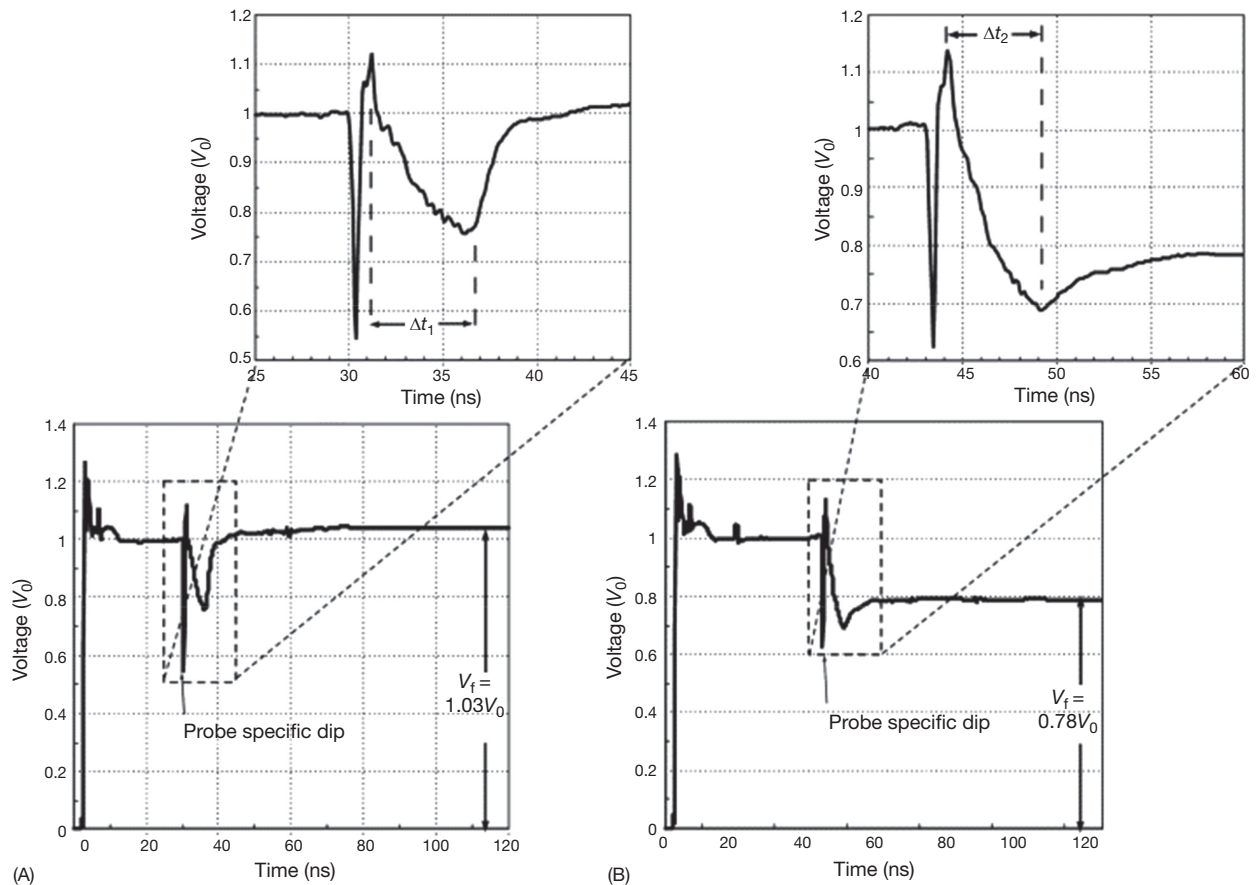


Fig. 1 Two time-domain reflectometry curves from 20-cm probes in silty clay loam soil. The soils are at similar water content but the soil solution is more conductive in (B), giving a smaller return reflection and resulting lower V_f . In (A) $\theta_v = 0.304 \text{ m}^3 \text{ m}^{-3}$ and $\sigma_0 = 57 \text{ mS m}^{-1}$ and in (B) $\theta_v = 0.271 \text{ m}^3 \text{ m}^{-3}$ and $\sigma_0 = 95 \text{ mS m}^{-1}$.

$$v = \frac{2L}{t} \quad (1)$$

From electromagnetic wave propagation theory, v is given in terms of relative dielectric permittivity and electrical conductivity of soil, which can be expressed as:

$$v = \frac{c}{\sqrt{\epsilon_{ra}}} \quad (2)$$

Where c ($= 3 \times 10^8 \text{ m s}^{-1}$) is the velocity of light and other EM waves in vacuum, and ϵ_{ra} is the apparent relative dielectric permittivity or apparent dielectric constant that controls the rate of propagation of the step pulse. Combining Eqs. (1) and (2) gives:

$$\epsilon_{ra} = \left(\frac{ct}{2L} \right)^2 \quad (3)$$

More complete analyses of the EM wave equations show that ϵ_{ra} is primarily the result of the relative dielectric permittivity (real component) or dielectric constant of the soil, ϵ_r' . Also, ϵ_{ra} is affected to some extent by the electrical conductivity, σ_0 (S m^{-1}), the frequency of the signal, ω ($2\pi \text{ s}^{-1}$), and the relative dielectric loss or imaginary component of relative dielectric permittivity, ϵ_r'' . The effect of these factors on EM waves is expressed as:

$$\epsilon_{ra} = \frac{\epsilon_r'}{2} \left[1 + \sqrt{1 + \left(\frac{\epsilon_r'' + \frac{\sigma_0}{\omega\epsilon_0}}{\epsilon_r'} \right)^2} \right] \quad (4)$$

where ϵ_0 is the dielectric permittivity of free space ($= 8.85 \times 10^{-12} \text{ F m}^{-1}$). Most TDR applications in soil have assumed that $(\epsilon_r'' + \sigma_0/\omega\epsilon_0)/\epsilon_r' \ll 1$, which leads to $\epsilon_{ra} \approx \epsilon_r'$. This assumption allows for the determination of the volumetric water content without prior knowledge of the bulk electrical conductivity.

The high relative dielectric permittivity of water imparts to wet soil a very strong dependence of ϵ_{ra} on the water content. The early laboratory work in TDR adopted an empirical approach, making no assumption about the 'state' of water in porous materials. Instead, it related the TDR-measured ϵ_{ra} to the volumetric soil water content, θ_v ($\text{m}^3 \text{ m}^{-3}$), based on oven drying to 105°C . This early work gave a consistent empirical relationship between relative dielectric permittivity and volumetric water content for a wide range of soils and became a model for calibration of the TDR as a method for measurement of water content. Furthermore, it was found that the relationship is essentially independent of soil bulk density, texture, ambient temperature, and salt content, within the range of these parameters most often encountered. The equation now used widely for calibration is:

$$\theta_v = -5.3 \times 10^{-2} + 2.92 \times 10^{-2} \epsilon_{ra} - 5.5 \times 10^{-4} \epsilon_{ra}^2 \div 4.3 \times 10^{-6} \epsilon_{ra}^3 \quad (5)$$

This relationship has extremely broad applicability, providing good descriptions for the water content of soils of widely varying properties, such as iron-rich volcanic soils, the unfrozen water content in frozen soils, soils with up to 50% gravel, oil shale waste, and crushed concrete. The widely obtained agreement between observations and Eq. (5) has not been true for observations made in low-density soils, mineral soils with high organic or clay contents and artificial soils, such as glass beads. The widespread use of TDR has resulted in a number of applications in which Eq. (5) cannot be used, and some effort has been devoted to finding alternative relationships between ϵ_{ra} and θ_v . However, no relationship has emerged that shows more general applicability without the requirement for other supporting measurements (e.g., clay content, temperature). Various approaches have been presented over the years (e.g. Whalley, 1993; Kojima et al., 2023; Wang et al., 2023).

Alternatives to empirically derived calibrations are often based on dielectric mixing models. The application and use of mixing models require varying levels of prior knowledge about the soil, such as porosity, density, and the relative dielectric permittivity of soil components. The application of mixing models for water content determinations has shown that 'square root' mixing models are most common. Such models are equivalent to a linear relationship between $\sqrt{\epsilon_{ra}}$ and the TDR travel time, t in Eq. (3), which leads to a convenient linear calibration for water content. The linear relationship between θ_v and is supported by calibration data from numerous sources and is very similar to the polynomial expression in Eq. (5) (Fig. 2). Fitting a linear relationship where possible presents a significant improvement over fitting a polynomial calibration curve because there are only two parameters to define, making the fitting procedure much simpler. A linear regression of $\sqrt{\epsilon_{ra}}$ on θ_v gives rise to (Fig. 2):

$$\theta_v = 0.115\sqrt{\epsilon_{ra}} - 0.176 \quad (6)$$

The constant term includes the electrical properties of the dry solids, primarily the bulk density and the relative permittivity of the constituent material. The slope coefficient in the first term on the right-hand side embodies the effects of dissolved solutes and clay surfaces on the electrical properties of the water phase as seen by TDR. This slope changes relatively slowly as salt and clay contents increase. Clay and salt content do introduce some curvature to the relationship between $\sqrt{\epsilon_{ra}}$ and θ_v , indicating the importance of calibration checks to assure an appropriate calibration for high-resolution measurements and for unique soil conditions. Precise calibration equations become more important where absolute measures of water content are required as contrasted with differences or changes in water where more general calibrations usually suffice.

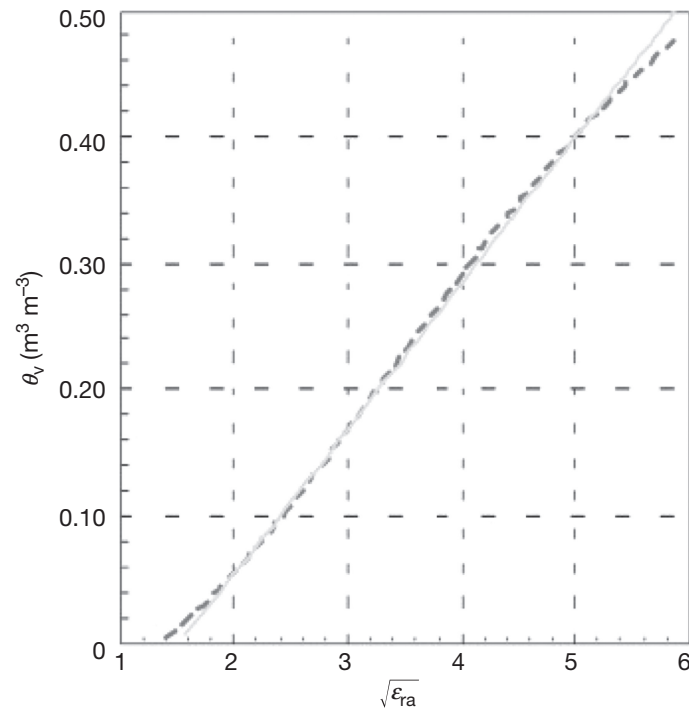


Fig. 2 The similarity between Eq. (5) (dashed line) and Eq. (6) (solid line) over the water content range ($0.05\text{--}0.45\text{ m}^3\text{ m}^{-3}$) applicable for most soils. Note the x-axis is $\sqrt{\epsilon_{ra}}$ and shows the linearity of Eq. (6) (White et al., 1994).

Instrumentation and wave-guides

The original and still widely used TDR instrument for soil was the portable cable-tester. This instrument displays the TDR waveform on a screen, allowing analyses to be performed and recorded manually (Lekshmi et al., 2014). Alternatively, the data can be recorded and analyzed digitally on a personal computer (PC). Currently, a number of companies offer TDR instrumentation aimed directly at soil and environment studies. These offer automated analysis of the TDR waveform as a part of the basic instrument with network switching for multiple probes as an option. Most of these instruments make use of general calibration relationships that closely approximate Eqs. (5) and (6). In addition, they allow for the calibration relationships to be customized for the desired soil characteristics.

One of the most important components of a TDR system is the soil probe or transmission line. The basic elements are the conductive components, often parallel metallic rods, which act as wave-guides. Currently the most common soil probes are of the parallel rod type, consisting of two or three parallel rods. These may vary in length from $0.1\text{--}1.0\text{ m}$ and with probe separations from $0.01\text{--}0.1\text{ m}$. Probes that use bare metallic rods measure total water content along the path of the TDR pulse regardless of the distribution of water along this path. However, the distribution of electromagnetic energy is not uniform in the directions perpendicular to the rods. Furthermore, the distribution of energy depends on the number of rods, their diameters, and their separations. A rod separation to rod diameter ratio less than 10 is a practical rule of thumb for probe design. Further, the rod diameter should be large in relation to the dominant pore size, e.g., 10 times the dominant particle or aggregate size. We have found that 6-mm rods spaced at 50 mm have worked well in a variety of studies in tilled and untilled agricultural soil.

Many other probe configurations have come into current use, including profiling probes with the conductors mounted on opposite sides of a dielectric central core. The parallel wires mounted on a central core have been diversified to include a helical-wrapped pair on a single shaft probe and serpentine pathways on a plastic surface giving a surface probe. In all cases the energy density of the propagating wave decreases rapidly with distance from the conductors. Spatial weighting of the measurements in the cross-section perpendicular to the direction of wave propagation is a very important consideration, especially for probes that have nonmetallic components near the metal rods. Numerical analyses have been applied to the wave-equation response from parallel rod probes embedded in a porous medium. One aspect of this work was the definition of sample areas in the transverse plane that are not limited to predefined shapes, allowing for a more realistic description of the lateral sensitivity of different probe configurations. In general, TDR probes should be installed in a manner that minimizes variability of the water content within their sampling volume, especially in the direction that is transverse to the rods.

An important application of TDR relates to the measurement of shallow soil water content profiles. Three general approaches to water content profiling are depicted in Fig. 3 using vertical and horizontal probes. In general, vertical rods (Fig. 3A) give the most

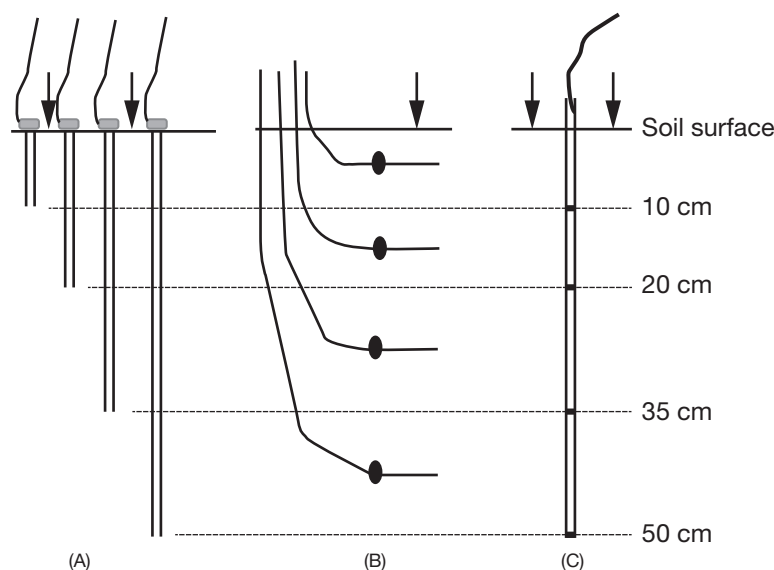


Fig. 3 Three options for measurement of water content profiles. (A) Adjacent vertical; (B) horizontal; (C) segmented vertical. Probe lengths and intervals can be adjusted to the monitoring requirements. Installing probes in (A) and (C) off vertical results in a compromise between vertical and horizontal installation but optimizes the advantages from each (Ferré and Topp, 2002).

accurate measurement of cumulative water from the surface to the ends of the rods, regardless of the water content distribution with depth. Horizontal rods (Fig. 3B) give higher resolution of the water content profile, with less accurate measurements of the total water stored to a given depth. It is possible to place probes depicted in Fig. 3A and C at an angle off the vertical. Angled rods at say 45 degrees give an optimal compromise between horizontal and vertical configurations. Another consideration is the potential for causing cracks in the soil or gaps around the rods. These features can have a significant impact on TDR measurements and on water movement, and vertical emplacements are most susceptible to these influences. In addition, water content profiles that are determined by the differencing of adjacent measurements (e.g., Fig. 3A) introduce the error from both measurements, which can lead to a cumulative error as large as $\pm 0.03 \text{ m}^3 \text{ m}^{-3}$. On the other hand, horizontal rods require excavation for insertion at depth, which can also affect water movement. In addition, because the head of a horizontal probe is placed beneath the ground surface, it is important that it be watertight to avoid the influence of water entry. Practically, vertical rods are the simplest to install. However, their tendency to move vertically out of the soil during winter by the processes of frost-heave can be countered somewhat by the use of angled rods.

An alternative to the continuous-rod probes is the segmented probe depicted in Fig. 3C. This probe configuration makes use of diode-shortening to achieve probe segmentation electronically and is available commercially. This probe is essentially a series of vertical probes placed in the same hole. As a result of its design, these probes provide both the water content profile and the total profile stored water from a single installation. Installation of this type of probe is of critical importance as any disturbance of the soil adjacent to the probe is in the most sensitive part of the measured region. Similarly, air gaps or preferential water flow along the probe will also affect measurements.

Impact of salinity on water content measurement

The explicit effect of the bulk electrical conductivity on TDR water content determinations has not been fully evaluated. Research is continuing with varied EM techniques to define conditions when the direct effect of bulk electrical conductivity (EC) results in unacceptable error in water content estimates. It appears, however, that the influence is very small for mineral soils having a clay content less than 50% and pore-water electrical conductivities less than 5 dS m^{-1} . For example, measurements in a soil of 55% clay gave an overestimate of water content of $0.014 \text{ m}^3 \text{ m}^{-3}$ due to the assumption that $\epsilon_{ra} = \epsilon_r'$. Typically, high EC conditions lead to difficulty in measuring the water content due to total extinction of the waveform through energy loss. This places a practical limit on the maximum length of TDR rods. However, to maintain acceptable water content accuracy, probe length should be at least 7.5 cm. If the EC is too high to use probes at least 7.5 cm in length, nonconductive rod coatings can be used to minimize the effects of signal loss. These coatings change the water content calibration and they should only be used when absolutely necessary.

Some of the limitations of TDR analysis and application may be addressed by a new approach to TDR waveform analysis. Standard analyses involve interpretation of the time of arrival of reflections. However, in addition to this time-based analysis, a TDR waveform can be analyzed in the frequency domain using Fourier analysis. Conversion of TDR data into the frequency domain provides additional frequency-dependent information about the electrical properties of the soil and water. Frequency domain analyses of measurements made with 3-cm length probes may provide high-precision measurement of relative dielectric permittivity in saline soils where the time-domain analysis is not feasible. Further developments of this approach may extend the range of applicability of TDR.

Measurement of bulk electrical conductivity using TDR

As the TDR wave propagates through soil guided by the probe rods, its velocity is decreased by the presence of water. In addition, the magnitude of the wave is decreased by electrical losses associated with the bulk electrical conductivity of the soil, σ_0 . The decrease in magnitude of the wave is expressed in the attenuation coefficient, α . It is possible to express σ_0 in terms of the values obtained from the TDR waveform (Fig. 1):

$$\sigma_0 = \left(\frac{1}{120\pi L} \right) \left(\frac{Z_0}{Z_i} \right) \left(\frac{2V_0}{V_f} - 1 \right) \quad (7)$$

where Z_0 (ohms) is the characteristic impedance of the TDR probe (a factor of the probe geometry), Z_i is the output impedance of the TDR instrument, V_0 (volts) is the initial step voltage output and V_f is the final voltage remaining after all multiple reflections from the probe, respectively. In practice, the quantity $(Z_0/120\pi L Z_i)$ is often treated as a calibration constant for the instrument, cable, and probe combination, which can be obtained from measurements in solutions of known EC. After the probe and TDR instrument have been characterized, the bulk EC is calculated from the value of V_f as the only independent variable at each measurement site or measurement time. Although the measurement is straightforward, it is important that V_f be determined well after all multiple reflections have returned and the TDR waveform has reached a final constant value.

The bulk EC measured by TDR has applications in research into solute transfer processes, particularly ionic solute breakthrough experiments. At a heterogeneous field site, using vertically installed TDR rods, solute mass flux measured with TDR was in good agreement with that obtained from solution samplers. The TDR-measured EC has been used to monitor nitrate levels in the field with a view to improving crop nutrient management and minimizing leaching to groundwater.

Impact of water content on pore-water salinity measurement

The EC of the soil solution is often desired for soil and environment studies, particularly in relation to soil salinity assessment or monitoring migration of polluting chemicals. The relationship between bulk EC and solution EC is strongly dependent on both the soil water content and the clay content of the soil. TDR offers the distinct advantage of measuring both water content and EC in the same sample and under identical conditions. Thus, TDR can be very useful for establishing both the water content and the pore-water EC simultaneously. TDR has been used in both direct and indirect approaches and to elucidate the relationship among water content, bulk EC, and solute EC. However, it is not always possible to determine the pore water EC if the water content varies along the TDR probes, as this will give a gradient of pore water EC along the probe.

Conclusion

Since the early applications in soils, TDR has revolutionized the field measurement of soil water content. In 1991, a survey of soil physicists in four countries reported TDR was used one-third as often as gravimetric sampling and one-half as often as neutron moderation. It is most likely that TDR use had surpassed both these methods by the end of the 1990s, making it more ubiquitous than any other technique. The capability to log TDR data and the relatively nondestructive nature of the installations have allowed in situ monitoring, contributing to the much wider use of TDR. For water balance and water-use efficiency studies, TDR is particularly applicable because the probe designs and orientations can be chosen to meet a variety of conditions, such as agricultural row crops, drip irrigation patterns, and forested landscapes. Through the application of TDR in soil much has been learned about the soil's electromagnetic properties in the radio frequency range. This knowledge has assisted in the development of other electromagnetic techniques, such as ground-penetrating radar and remote sensing using either active or passive microwave radars. TDR is specifically valuable for 'ground-truthing' radar applications because both are interpreting measured EM properties of soil. As a consequence, TDR can be applied to defining the sampling regions of radar applications by appropriate choice of probe geometries for the TDR. Water content measurement by capacitance techniques has been accelerated recently, in part because appropriate frequencies and probe geometries were identified through TDR studies. Although capacitance devices are less expensive, they have not yet offered the ability to measure EC as can be done by TDR.

TDR is prominent among several emerging technologies being used to enable the long-neglected investigation of roots as the major agents of water and chemical transfer in soil. TDR water content measurements also improve the value and significance of other environmental parameters. For example, frequent measurements of temperature, water content by TDR, and oxygen concentrations in the rooting zone of maize can help to demonstrate when tillage and trafficking limit both O_2 supply and crop growth. Prior to TDR, measurements of O_2 supply for roots and other biota in the soil had not been possible. Time-domain transmissiometry (TDT), which is similar to TDR, has been combined with a soil penetrometer to measure both soil strength (compaction) and water content. The combination of sensors greatly enhances the value of the soil penetrometer as a diagnostic tool for soil compaction.

Regulations now play a leading role in the need to monitor soil contamination and harmful infiltration of contaminated solutions into groundwater. Cities in the USA such as Los Angeles, Tucson, Las Vegas, and Glendale are using TDR for such

monitoring. Agencies from around the world have reported using TDR in environmental monitoring applications as diverse as nonaqueous liquid-phase concentrations (Sweden), subsurface pollutant detection (United Arab Emirates), changes in groundwater level and crude oil thickness (USA), characterization of diesel-contaminated soil (Canada), and measuring water-table elevations within sidewalk systems (USA).

The application of TDR in soil is continuing to develop and diversify in ways that will enhance the value of TDR for soil and environmental monitoring. The analyses of spatial sensitivity within probes have facilitated designs for specific uses, often allowing combinations of TDR with other measurements such as with tensiometers and the cone penetrometer.

References

- Binley A, Hubbard SS, Huisman JA, Revil A, Robinson DA, Singha K, and Slater LD (2015) The emergence of hydrogeophysics for improved understanding of subsurface processes over multiple scales. *Water Resources Research* 51: 3837–3866.
- Clothier BE and Green SR (1997) Roots: the big movers of water and chemical in soil. *Soil Science* 162: 534–543.
- Dowding CH (ed.) (2001) *Second International Symposium and Workshop on Time Domain Reflectometry for Innovative Geotechnical Applications*. IL, USA: Infrastructure Technology Institute at Northwestern University Evanston.
- Ferré PA and Topp GC (2000) Time-domain reflectometry sensor techniques for soil water content measurements and electrical conductivity measurements. In: Baltes H, Gopel W, and Hesse J (eds.) *Sensors Update*. vol. 7, pp. 277–300. Weinheim: Wiley-VCH.
- Ferré PA and Topp GC (2002) Time domain reflectometry. In: Dane J and Topp GC (eds.) *Methods of Soil Analysis, Part 4*, 3rd edn, pp. 434–446. Madison, WI: Soil Science Society of America.
- Kojima Y, Okumura K, Aoki S, Noborio K, Kamiya K, and Horton R (2023) A four-parameter-based thermo-TDR approach to estimate water and NAPL contents of soil liquid. *Geoderma* 429: .
- Lekshmi SUS, Singh DN, and Baghini MS (2014) A critical review of soil moisture measurement. *Measurement* 54: 92–105.
- O'Connor KM and Dowding CH (1999) *GeoMeasurements by Pulsing TDR Cables and Probes*. Boca Raton: CRC Press 416.
- Topp GC and Reynolds WD (1998) Time domain reflectometry: A seminal technique for measuring mass and energy in soil. *Soil & Tillage Research* 47: 125–132.
- Topp GC, Davis JL, and Annan AP (1980) Electromagnetic determination of soil-water content - measurements in coaxial transmission-lines. *Water Resources Research* 16: 574–582.
- Wang ZJ, Hua S, Timlin D, Kojima Y, Lu ST, Sun WG, Fleisher D, Horton R, Reddy VR, and Tully K (2023) Time domain reflectometry waveform interpretation with convolutional neural networks. *Water Resources Research* 59: .
- Whalley WR (1993) Considerations on the use of time-domain reflectometry (TDR) for measuring soil-water content. *Journal of Soil Science* 44: 1–9.
- White I, Knight JH, Zegelin SJ, and Topp GC (1994) Comments on "Consideration of the use of time-domain reflectometry (TDR) for measuring soil water content" by Whalley WR. *European Journal of Soil Science* 45: 503–508.
- Wraith JM (2002) Time domain reflectometry. In: Dane J and Topp GC (eds.) *Methods of Soil Analysis, Part 4*, 3rd edn, pp. 1289–1297. Madison, WI: Soil Science Society of America.

Geophysical methods for soil applications

Sarah Garré^a , Guillaume Blanchy^{a,b} , David Caterina^c , Philippe De Smedt^d ,

Alejandro Romero-Ruiz^e , and Nataline Simon^c, ^aFlemish Research Institute for Agriculture, Fisheries and Food (ILVO), Merelbeke, Belgium; ^bDepartment of Environment, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium; ^cDepartment of Urban and Environmental Engineering, University of Liège (ULiège), Liège, Belgium; ^dDepartment of Environment and Department of Archaeology, Ghent University, Ghent, Belgium; ^eNet Zero and Resilient Farming, Rothamsted Research, Hertfordshire, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction	445
Geophysical variables and pedo- or petrophysical models	445
Applications of geophysical methods in soil science	445
Electrical resistivity tomography (ERT)	446
Challenges and limitations	447
Electromagnetic induction (EMI)	448
Challenges and limitations	449
Magnetometry and magnetic soil mapping	449
Challenges and limitations	450
Ground penetrating radar (GPR)	450
Challenges and limitations	451
Induced polarization (IP)	451
Challenges and limitations	452
Nuclear magnetic resonance (NMR)	452
Challenges and limitations	452
Seismic methods	452
Challenges and limitations	454
Distributed temperature sensing (DTS)	454
Challenges and limitations	454
Forward modelling and inversion	455
Data standardization and open science	456
Conclusion	456
References	457

Abstract

Traditional methods to assess soil properties or monitor soil processes typically rely on point data. They often fail to represent the heterogeneity of soil. Airborne remote sensing techniques allowed us to measure proxies for soil properties or variables at regional scale, but often covering only the first few centimeters of soil. Ground-based geophysical methods can fill in the gap between these two scales. They primarily observe variations in the thermal, electrical, magnetic, electromagnetic or seismic properties of the subsurface to map its characteristics or to monitor dynamic changes and can yield 2- or 3-D quantitative information on soil properties or processes. Geophysical methods have been applied in various soil-related research fields: from soil mapping and precision agriculture, to geotechnical engineering and soil remediation and to archeology and forensic investigation. This chapter briefly introduces the ground-based geophysical techniques appropriate for soil applications.

Key points

- Geophysical methods are key to bridge the gap between point measurements and remote sensing techniques for soil applications.
- This chapter introduces the ground-based geophysical techniques relevant for soil applications, commonly used equipment and limitations of each technique
- The use of geophysical techniques for soil applications ranges from soil mapping and precision agriculture, through geotechnical engineering and soil remediation, to archeology and forensic investigation.
- We discuss critical data processing steps such as inversion and pedophysical relationships.

Introduction

The soil plays a crucial role in the redistribution of energy, water and solutes in terrestrial systems. Relevant scales range from the single pore and the soil-root interface, through the field scale, to encompass the entire globe. The multi-scale heterogeneity of soils presents a challenge to the representation of processes over different scales and to the parameterization of corresponding models. Traditional observation methods for soil processes often rely on (a network of) point data from manual sampling or sensors that capture states and attributes of soil system functioning. These point data often fail to capture the heterogeneity of soil properties at larger scale. Over recent decades, airborne remote sensing techniques (from drones to satellites) generated proxies for near-surface soil properties at the regional scale (e.g. SMOS, SMAP, Sentinel-1, Sentinel-2 campaigns). Nevertheless, those methods only interrogate the first few centimeters of the soil. In addition, the time-intervals between measurements and spatial resolution are rather coarse (days and 10–100 s of meters, respectively). Ground-based geophysical methods can fill in the gap between local point measurements and remote sensing products. Geophysical methods typically observe variations in the thermal, electrical, magnetic, electromagnetic or seismic properties of the subsurface to map its characteristics or to monitor dynamic changes. They can yield 2- or 3-D information on soil properties (e.g., texture or organic matter content) or state variables of interest (e.g., temperature, soil moisture, salinity) at centimeter or decimeter resolution.

This chapter briefly introduces key ground-based geophysical techniques relevant for soil applications, commonly used equipment and limitations of each technique. The use of unmanned aerial vehicles (UAVs) for geophysical surveys (e.g. EMI, GPR, magnetometry) or sensors mounted on towers (e.g. L-band radiometry, cosmic-ray neutron sensing) are not covered here, but is certainly also relevant for soil applications. Airborne GPR and EM are increasingly being used on UAVs.

Geophysical variables and pedo- or petrophysical models

For a geophysical measurement, we register the response of the soil to a signal which is naturally present (e.g. gravity) or which is actively introduced by the observer (e.g. an electrical current). For soil applications, electromagnetic energy is most used, alongside acoustic waves and gravitational fields. In all cases, the observed change in voltage in the measurement device is related to the geophysical response. The most straightforward example of this process is the direct current (DC) resistivity method, whereby a current I is introduced into the soil, after which the resultant potential difference (or voltage V) between pairs of electrodes is recorded with a voltmeter. The resistance of the recorded volume is then obtained through Ohm's law. To translate the collection of signals (e.g. resistances) to a distribution of the geophysical property (e.g. the electrical resistivity distribution of the bulk soil), we then have to incorporate information on the measurement geometry, the physical processes taking place during the measurement and the nature of the soil in a process called 'inversion.'

Typically, the inferred geophysical properties are not the primary property of interest, but they are used as a proxy for a soil characteristic or state variable. Therefore, relationships are established between these properties of interest and the geophysical properties. Such relationships are called petrophysical or, in the field of soil science, pedophysical relationships. A pedophysical relationship expresses the influence of different soil properties and state variables on the geophysical property. Soil electrical resistivity is for example affected by, among others, pore water salinity, soil texture, porosity, temperature and soil moisture. Therefore, several sources of independent data or calibration data are necessary to develop pedophysical relationships. We give an overview of the geophysical methods, the properties measured and examples of the derived properties and states in [Table 1](#).

Applications of geophysical methods in soil science

During the past decade geophysical measurement and data processing techniques became more affordable and easier to deploy. As such, these instruments offered the potential to explore and monitor the spatio-temporal dynamics of soil. In their review on hydrogeophysics, [Binley et al. \(2015\)](#) pointed out the expanding range of applications: from ecosystem science and biogeochemistry to agronomy and soil mapping. Since then, the possibilities have only expanded. Civil and geotechnical engineering problems are increasingly addressed using geophysics, especially when the application is water-related. For example, ERT, GPR, and self-potential (SP), combined with high-precision leveling, are used to assess sinkhole risks and several authors have used ERT to monitor landslide activity. Another common application is soil pollution and remediation monitoring in polluted sites. The efficiency of geophysical techniques for soil remediation monitoring greatly depends on the nature of the pollution and how it results in specific electrical or magnetic signatures. Geophysical methods have become ubiquitous in archaeological practice and heritage management. The first applications started in the 1950s and 1960s. Magnetometry instruments are the workhorse for archaeological prospecting, complemented with ground penetrating radar, direct-current resistivity and electromagnetic induction surveys. Archeological applications have also driven the deployment of geophysical instrument (arrays) geared at high sensitivity and sampling density. Geophysics is also used for forensic investigations. For example, [Doro et al. \(2022\)](#) used ERT to locate clandestine graves and monitor human decay within the subsurface.

Precision agriculture is one of the fastest growing soil-centered areas for adopting geophysical methods. Where EMI has been used for years as a qualitative soil mapping tool or for more detailed experimental design, multi-geometry devices now allow for

Table 1 Geophysical methods, their properties measured and examples of derived properties and states.

<i>Geophysical Method</i>	<i>Common abbreviations</i>	<i>Geophysical Properties</i>	<i>Examples of Derived Properties and States</i>
Electrical resistivity tomography	ERT, ERI	Electrical conductivity	Water content, clay content, pore water conductivity
Induced polarization	IP, EIT	Electrical conductivity, chargeability	Water content, clay content, pore water conductivity, surface area, permeability
Spectral induced polarization	SIP	As above but with frequency dependence	Water content, clay content, pore water conductivity, surface area, permeability, geochemical transformations
Self-potential	SP	Electrical sources, electrical conductivity	Water flux, permeability
Electromagnetic induction	EMI, FDEM, TDEM	Electrical conductivity	Water content, clay content, salinity
Magnetometry	MAG	Magnetic susceptibility	Ferromagnetic iron oxide content, redox conditions
		Magnetic susceptibility, Remanent magnetization	Ferromagnetic iron oxide content, redox conditions.
Ground penetrating radar	GPR	Permittivity	Water content, porosity, stratigraphy
Seismic		Elastic moduli and bulk density	Lithology, ice content, cementation state, pore fluid substitution
Seismoelectrics		Electrical current density	Water content, permeability
Nuclear magnetic resonance	NMR	Proton density	Water content, permeability
Gravity		Bulk density	Water content, porosity
Distributed temperature sensing	DTS	Temperature	Water content, thermal properties (conductivity and diffusivity)

Adapted from Binley A., Hubbard SS, Huisman JA, Revil A, Robinson DA, Singha K, and Slater LD (2015) The emergence of hydrogeophysics for improved understanding of subsurface processes over multiple scales. *Water Resources Research* 51(6):3837–3866. ISSN 0043-1397. <https://doi.org/10.1002/2015wr017016>.

inversion of the apparent electrical conductivity, unlocking information of vertical heterogeneity in addition to horizontal patterns. Combination with machine learning techniques has resulted in new possibilities. RomeroRuiz et al. (2021) used seismic to locate areas suffering from chronic soil compaction. GPR has been used at several sites to locate tile drainage components. Also, the monitoring potential of EMI and ERT and its relationship with soil moisture has been extensively used. Several researchers use geoelectrical methods to study the root water uptake of trees and crops and estimate the efficiency of applied irrigation patterns. This has led to the application of EMI, and also ERT, in plant phenotyping through the assessment of soil moisture depletion.

Previously, the phenotyping community struggled to incorporate environmental effects into its studies in field environments, especially for the below-ground part of the plant. Geophysics offers a window into that ‘hidden half.’ Fig. 1 gives an overview of the field setup of different geophysical techniques covered in the chapter.

Electrical resistivity tomography (ERT)

Electrical Resistivity Tomography (ERT) is a technique, which infers the bulk electrical resistivity of the soil between electrodes. The bulk electrical resistivity corresponds to the combined resistivity of soil particles, pore water and air. A basic measurement system consists of four electrodes (A, B, M, N), a transmitting electrode pair and a receiving electrode pair. The transmitter electrode pair applies a quasi-direct current (I) to the ground using rectangular DC current pulses or an AC current at low frequency to avoid polarization of injection electrodes A and B. The receiver measures the voltage (V) between the potential electrodes M and N (see Fig. 2). Resistivity meters then register a resistance (R) for each combination of four electrodes, based on Ohm’s law $R = V/I$. For a tomography, many different combinations of current and potential electrodes are used for injection and measurement. The electrodes can be inserted into the soil surface or at several depths as borehole electrodes or a combination of both.

From the current, the voltage and a geometric configuration factor (k), the apparent electrical resistivity (ρ_a) is calculated. The geometric factor depends on the spatial arrangement of the four electrodes and the geometry of the measurement area (e.g. presence of topography). Resistance is an extrinsic property (i.e., depends on the way it is measured), whereas resistivity is an intrinsic property (i.e., depends only on the material in which the measurement is performed). We call the obtained resistivity “apparent”, because it represents the resistivity of a hypothetical, homogeneous medium, which would give the same resistance value for the applied electrode arrangement. The measured, apparent resistivity is a complex average of the resistivities of the various materials that the current encounters.

A soil-specific pedophysical relationship is necessary to derive any soil state variable or property from electrical resistivity measurements. A well-known electrical resistivity model is the empirical Waxman and Smits (1968) model that extends Archie’s Law to account for surface conduction. Apparent resistivity is also affected by temperature. Ma et al. (2010) gives an overview of temperature correction models.

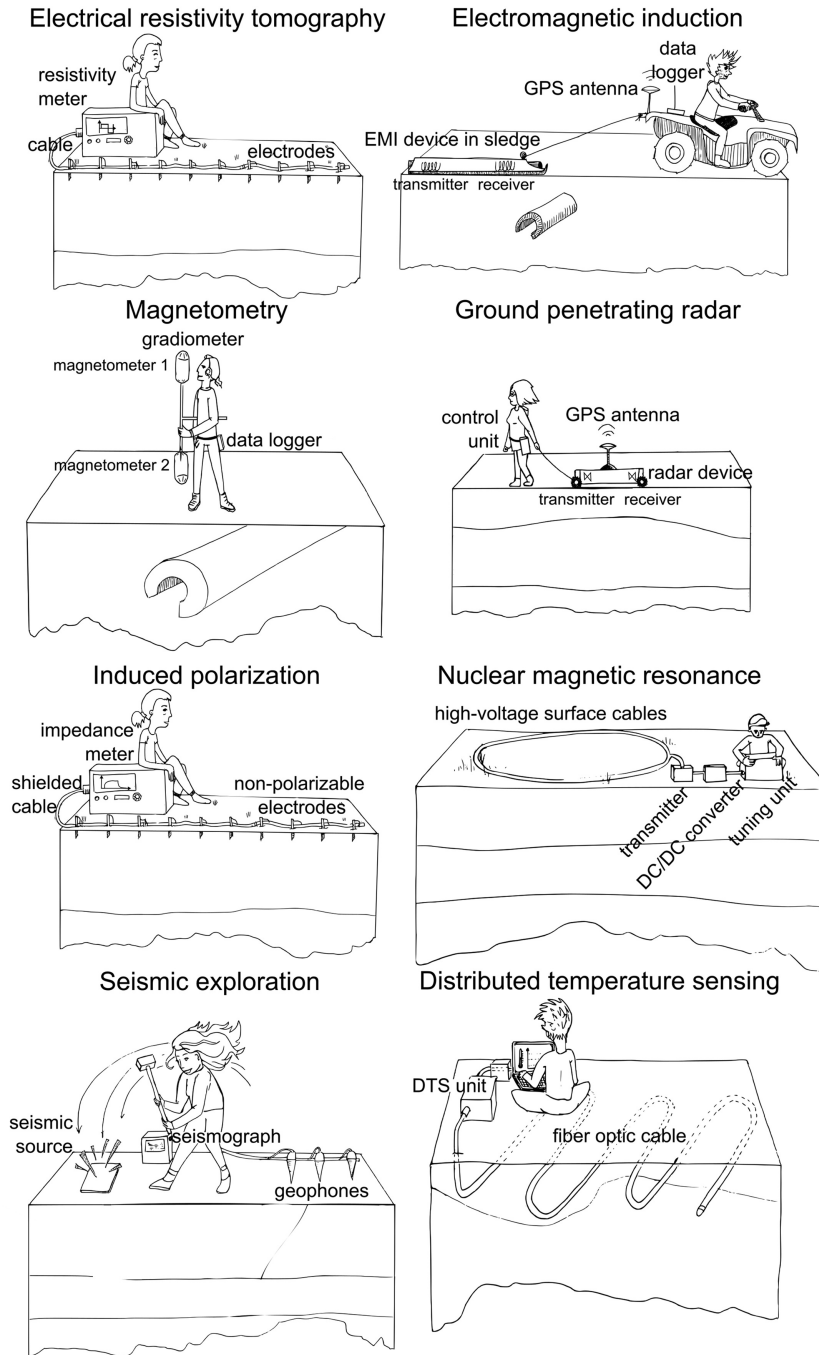


Fig. 1 Overview of the field setup for the different geophysical techniques for soil applications covered in this chapter.

Challenges and limitations

At the core of geoelectrical methods for application to soil is the assumption that the electrical conductivity map obtained of the subsurface can be translated into soil properties or state-variables via site-specific pedo-electrical relations calibrated with laboratory or field data. Common practice assumes spatially homogeneous relations, which neglects the impact of soil heterogeneity on those relations. This can cause significant errors in translating the electrical measurements to the hydraulic state variables. Standardized approaches to establish and upscale pedophysical relationships are, therefore, crucial to increase the accuracy of the method. Furthermore, the inversion remains a source of uncertainty, since an infinite number of electrical conductivity models fit the noisy geoelectrical data equally well.

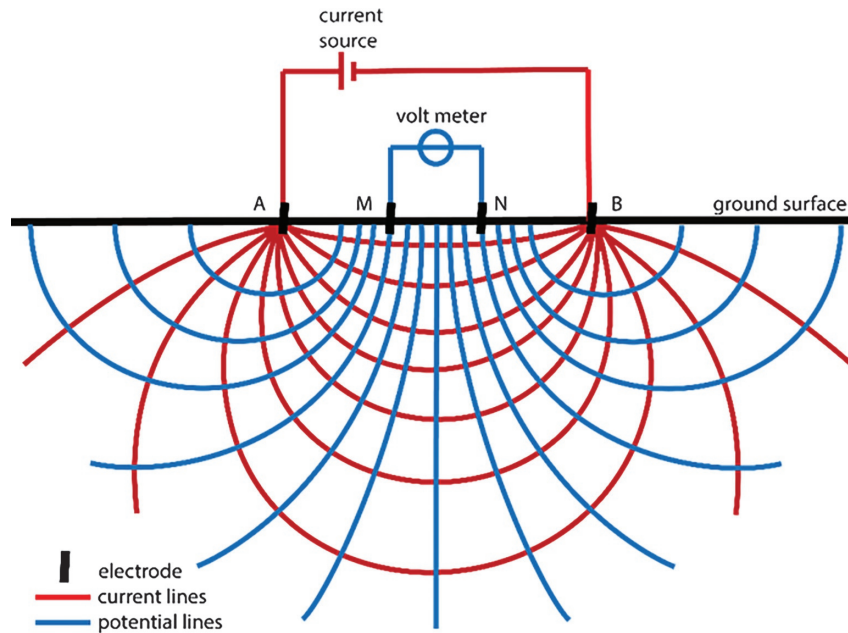


Fig. 2 Basic concept of electrical resistivity measurement of the subsurface.

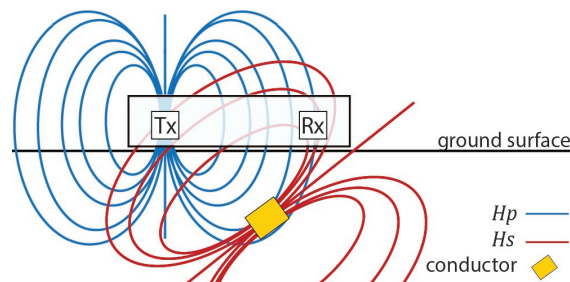


Fig. 3 Diagram of the FDEM working principle, indicating the primary (H_p) and secondary (H_s) magnetic fields.

Electromagnetic induction (EMI)

EMI uses low-frequency electromagnetic fields to investigate electrical and magnetic properties of the subsurface. For most soil applications, EMI measurements are performed in the frequency domain and described as frequency domain electromagnetics (FDEM). This means that the analysis of the signal is done with respect to frequency, rather than time. A magnetic field oscillating at a frequency between 1 and 100 kHz is generated by passing an alternating current through a transmitter coil (Tx, see Fig. 3) that consists of a small wire loop. This primary field (H_p) induces eddy currents¹ in conductive parts of the soil, which in turn generate a secondary magnetic field (H_s) that induces an oscillating current in the receiver coil (Rx). The size and shape of the subsurface volume that influences the observed current depends on the separation and respective orientation of transmitter versus receiver coil. For soil applications, coils are generally horizontally or vertically coplanar, or in a perpendicular orientation. One or more transmitter-receiver pairs are fixed in a rigid boom (so-called fixed-loop systems) with coil separations ranging from 0.20 m to 6.00 m. Although multi-frequency instruments are available, fixed-loop instruments used for soil applications operate at a single operating frequency, but integrate multiple transmitter-receiver pairs to enable geometrical depth sounding.

Evaluating the ratio of the secondary (H_s) to the primary magnetic field (H_p) at the receiver when both fields are out-of-phase or in-phase allows electromagnetic properties of the investigated volume to be derived. In low-conductive environments, the out-of-phase response is quasi-linearly related to the subsurface apparent conductivity. This approximate, but straightforward, relationship, is only valid under certain conditions, requiring a sufficiently small Tx-Rx distance, operating frequency and subsurface conductivity (Callegary et al., 2007). It is typically referred to as the low induction number approximation (LIN). Some instruments use multiple frequencies for different Tx-Rx distances to remain under this condition. When this condition is not met, e.g., in saline

¹Eddy currents (also called Foucault's currents) are loops of electrical current induced within conductors by a changing magnetic field in the conductor according to Faraday's law of induction.

environments, the relationship becomes more complex and a full evaluation of Maxwell's equations is needed to obtain reliable apparent conductivity values (Hanssens et al., 2019). Commercially available instruments developed for soil applications usually output the out-of-phase response as apparent electrical conductivity (σ_a) using the LIN approximation. Hence, they underestimate subsurface apparent electrical conductivity at higher conductivities and do not always offer access to the raw data.

In low-conductive environments (ca. <100 mS/m) the in-phase response informs on changes in soil magnetic susceptibility (κ). Although not widely used, current instrumentation allows reliable large-area mapping of soil susceptibility alongside apparent conductivity. Relating the in-phase response to quantitative changes in subsurface magnetic susceptibility can be ambiguous due to complex spatial sensitivities and the influence of conductivity on this component. Nevertheless, recent advances show the potential of soil magnetic susceptibility modelling (Klose et al., 2018; Delefortrie et al., 2018).

Challenges and limitations

Although current FDEM instruments are increasingly well-calibrated, issues of absolute and relative signal calibration burden data interpretation. Offsets from absolute zero, i.e., resulting in a non-zero response when making air readings, mainly exacerbate quantitative interpretation of the in-phase response, while, particularly under the influence of temperature variations, drift with time introduces errors in all signal components. Accounting for these issues is required to allow reliable quantitative data interpretation. To this end, adaptive calibration (Thiesson et al., 2014) and drift correction protocols (Delefortrie et al., 2014) are available. Resolving these issues can aid exploiting the full potential of inversion methods geared toward 3D models of soil conductivity and magnetic susceptibility.

Magnetometry and magnetic soil mapping

Magnetometry is a passive method that measures ambient magnetic fields. The interplay between Earth's magnetic field and magnetic materials in the subsurface lies at the heart of magnetometric prospecting. Magnetic materials can be considered as magnetic dipoles that are magnetized by the influence of Earth's magnetic field (see Fig. 4). In addition, those materials can possess a residual magnetization (remenance), which persists after an external magnetic field is removed and remains independent of weak magnetic fields. This is common in anthropogenic materials, such as iron objects or clay bricks, that obtain remanence through the production process. At any location, the resulting total magnetic field can be measured and described by its intensity, and its directional components. Crucial in magnetometer surveying is accounting for the influence of diurnal variations in Earth's magnetic field intensity, which are in the same order of magnitude of soil magnetic variations as natural and anthropogenic processes. To remediate this, stationary magnetometer measurements are conducted alongside roving measurements and subtracted from the survey data. Alternatively, magnetometers can be used in gradiometer mode, whereby two magnetometers are positioned at a fixed distance along the same axis. In such a configuration, the difference between the observations from both sensors is taken to remove diurnal variations, and provide more detailed insight into soil variations.

Two main types of magnetometers are in use: (i) total field magnetometers measuring the total magnetic field intensity and providing a scalar measurement (e.g., alkali-vapor (mainly cesium-vapor) magnetometers), and (ii) magnetometers recording a vector component (e.g., fluxgate instruments). The latter type is often configured in gradiometer mode as standard, i.e., as fluxgate gradiometers. More detailed discussions of the principles of magnetometry can be found in Everett (2013). Apart from dominating the use of geophysical methods in archaeology, and use in unexploded ordnance detection, magnetometry is rarely deployed in soil applications. Pilot studies, however, have shown that even the weakly magnetic signal rendered by iron oxides in soils can be used to detect influence of soil tillage (Mathé and Lévêque, 2003), whereas clay brick drainage pipes are straightforward magnetic targets due to their strong magnetic remanence (Rogers et al., 2006).

Alongside magnetometry, other methods enable magnetic soil mapping by targeting the induced magnetization, or magnetic susceptibility, directly. Complementing the potential of FDEM instruments, dedicated susceptibility meters (e.g., the Bartington MS2) are used to map the magnetic signal of topsoils in pollution (e.g., Declercq et al. (2019)) or ecological studies (e.g., Grimley

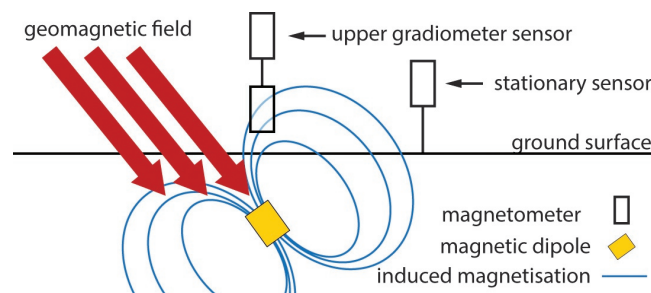


Fig. 4 Diagram of the working principle of a magnetometer showing various instrument configurations (vertical gradiometer, whereby observations of the upper sensor are subtracted from those from the bottom sensor) and inclusion of a stationary base station for recording diurnal geomagnetic variations).

et al. (2008)). Such instruments, consisting of inductor-capacitor circuits operating through principles of electromagnetic induction, only offer small depths of investigation that are limited to the upper centimeters below the surface. Downhole logging configurations, however, can be deployed to gain better insight into vertical magnetic variations.

Challenges and limitations

The inherent relationship between the soil magnetic signal and its mineralogical composition, as well as its independence from short-term moisture dynamics, presents good opportunities for soil applications. Broadly described, the soil magnetic signal is influenced by pedogenic processes (such as weathering, reduction and oxidation of iron-bearing minerals) and anthropogenic processes (e.g., the heating of soil or the input of detrital materials such as magnetic particulate matter), which both primarily occur in the upper soil layer, as well as the lithology of the parent material (e.g., weathering of igneous bedrock). A comprehensive overview discussion on pathways to soil magnetism is provided by Jordanova (2017).

Magnetometer surveys have potential to study soil erosion (Jordanova et al., 2014), redox variations (Grimley et al., 2008), or as an indicator for the spatial distribution of soil nutrients (Miroshnychenko et al., 2021). Here, combined use of adaptive instrumentation, for instance FDEM instruments operating at different frequencies, or the joint deployment and inversion of FDEM and magnetometry instrumentation can provide a more robust basis for field applications of magnetic soil mapping. Nevertheless, more work is needed to consolidate the fundamental framework for relating magnetic observations to investigated aspects of soil functioning across different soil types (Jordanova, 2017).

Ground penetrating radar (GPR)

Ground penetrating radar (GPR) uses electromagnetic fields in the mega-to-gigahertz (MHz - GHz) range to characterize the soil. Compared to EMI, these higher frequencies put GPR in the propagation regime, whereby the emitted fields behave as waves that propagate, reflect and refract in the media they travel through. Acquisition of GPR data can be done very rapidly (see Fig. 1) and, due to its wave-based nature, this method provides a much higher vertical (depending on soil properties and the source frequency) and horizontal resolution than other geophysical methods (e.g., ERT or EMI). For this reason, the GPR method is one of the most widely used geophysical methods in soil sciences and near surface surveying. GPR instruments are composed of a transmitting (Tx) and a receiving (Rx) antenna (see Fig. 5). The former sends an electromagnetic wave into ground, after which the receiver records the reflections of the emitted wave as it travels through the soil. The velocity of the wave is determined by the soil dielectric permittivity (ϵ). When the electromagnetic wave encounters a change in permittivity, part of the energy of the wave is reflected on the interface, while the remaining energy is transmitted or refracted. Recording these reflections at the receiver provides information on the structure of the subsurface (Fig. 5). These reflections are recorded in function of time, resulting in a so-called GPR-trace at each measurement location, traces can be aggregated in transects (radargram) or 3D volume. Several processing steps are needed to allow reliable interpretation of GPR data: the removal of interfering frequencies (e.g., background removal and 'dewow'), converting wave traveltimes to depth (velocity analysis and migration) and adjusting signal gain (to compensate, for instance, for energy loss with depth). An overview of standard GPR processing steps is given by Cassidy (2009).

The dielectric permittivity can be linked to other soil characteristics such as soil moisture content. This relationship has been widely used for Time-Domain Reflectometry (TDR) soil moisture sensors. Similarly, GPR has been used to map soil moisture by measuring the direct (not reflected) wave (Fig. 5) that travels along the ground surface or using air-launched methods. However, these methods usually only assess the soil moisture of the upper layer (few centimeters). Transmission GPR can also be used to measure larger volume by using the wave speeds between boreholes. Huisman et al. (2003) provided a review of the use of GPR for soil moisture estimation. GPR can also be used to locate the depth of soil horizons with different soil texture and associated water retention properties (thus producing abrupt vertical changes in water content) (André et al., 2012). In water-saturated conditions, GPR methods may help in locating zones with different porosities (i.e., different water content) or fractures, both associated with zones of increased permeability. Due to its simplicity, the empirical Topp's equation is typically used to relate the dielectric permittivity with water content and vice-versa. Yet, other physics-based models between the soil bulk ϵ and the dielectric properties

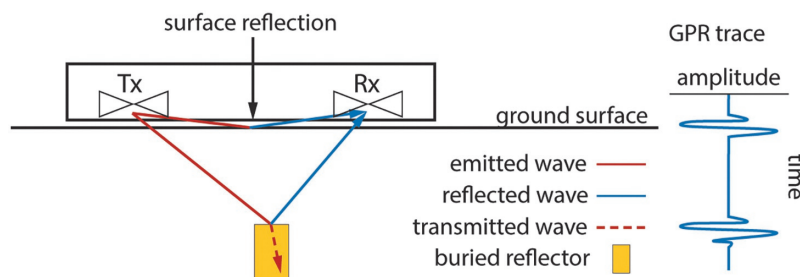


Fig. 5 Diagram of the GPR working principle. On the right, a GPR trace is shown as would be obtained from the received reflections shown in the diagram.

of soil individual constituents exist and are often based on volume averaging (Roth et al., 1990). Such models are successful in providing information about soil texture, soil structure or even help monitoring ice content in the soil.

The primary use of GPR for soil applications, however, is mapping out the individual traces along profiles and volumes recorded with a fixed transmitter-receiver distance. In these so-called reflection surveys, the location and extent of interfaces between materials with different permittivities are identified by visualizing their recorded reflections in GPR traces across surveyed areas. Hereby, one key advantage of GPR is the potential for the 3D discrimination of small subsurface permittivity discontinuities. Alongside use for environmental applications (Knight, 2001), this potential is particularly exploited in archaeology where the ability for identifying anthropogenic (e.g. Verdonck et al. (2020)) and palaeoenvironmental (e.g. Chapman et al. (2009)) phenomena of GPR is unparalleled.

Challenges and limitations

Despite its unique advantages in subsurface exploration, a key limitation in GPR surveying is energy loss of the travelling radar wave, leading to decreasing signal-to-noise ratios. This energy loss increases with travel distance in the propagation direction, attenuation and the amount of reflected and scattered energy (Annan, 2009). With higher conductivity (for instance due to high clay content or salinity), attenuation will rise, resulting in a limited depth of investigation or impeding use of GPR altogether. Attenuation has a stronger influence on high frequency waves. While lower frequency waves are more resilient to such dampening, these offer only limited spatial resolution (Annan, 2009). The choice of a GPR survey's central frequency therefore implies a trade-off between resolution and depth of investigation that takes into account the expected conductivity of the study area. In soil with high stone content, the energy loss due to signal scattering can also decrease the depth of investigation. Generally, GPR methods perform better in partially water-saturated soils with textures containing low or moderate clay contents. Despite these limitations, the GPR methods will continue to be one of the most widely used geophysical methods in soil science due to the unmatched sensitivity of ϵ to water content changes, their versatility and the relatively well-developed set of tools for interpretation of GPR signals.

Induced polarization (IP)

Polarization represents the ability of a soil to store and release electrical charges. Induced Polarization (IP) measurements can be conducted in the time-domain (Time Domain Induced Polarization or TDIP) or in the frequency-domain (Frequency Domain Induced Polarization or FDIP). When deployed in the field, the layout used is the same as for conventional electrical resistivity measurements. In the time-domain, the voltage at the receiver electrodes is not only measured during direct current injection, but also after the current has been switched off. The voltage decay is not instantaneous, but rather exhibits a gradual decrease that reflects polarization mechanisms. A common way to quantify the polarization is to integrate the voltage decay as shown in Fig. 6A. This

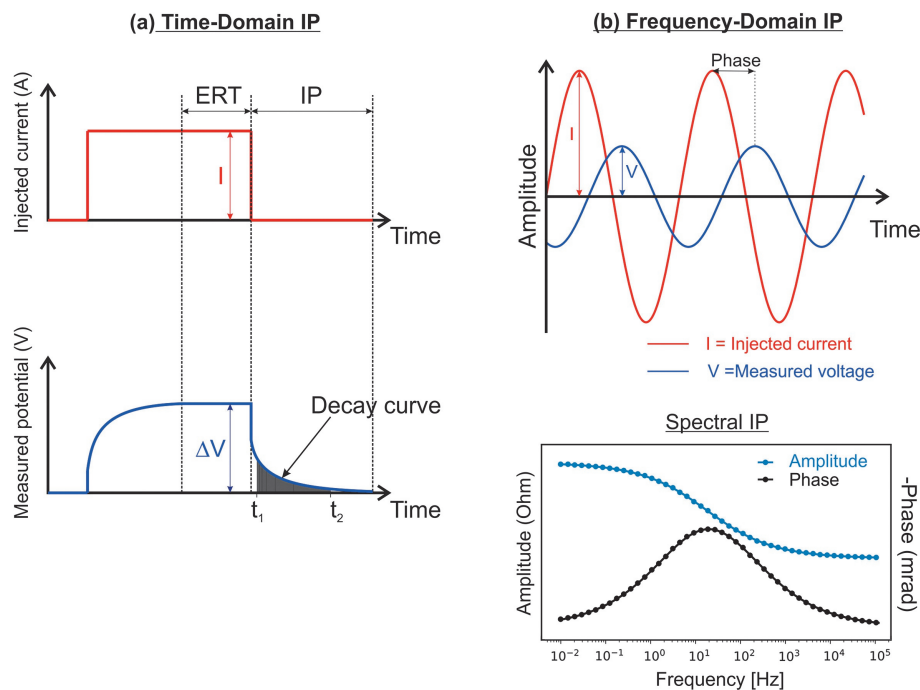


Fig. 6 Induced polarization measurements in the time-domain and in the frequency-domain.

results in an apparent chargeability usually expressed in mV/V. In the frequency-domain, an alternating current is applied and the relative phase-shifted voltage is measured. Data collected consists of a magnitude (i.e., transfer resistances, R) and a phase shift between the applied current and the measured voltage (Fig. 6B). When phase and magnitude are recorded over a frequency range from 10^{-3} Hz to several kHz, this is generally referred to as Spectral Induced Polarization (SIP). More information about IP methods is available in Kemna et al. (2012).

In the low frequency range (below 1 MHz), four main mechanisms explain the polarization occurring in porous media: (i) electric double layer polarization occurring in the layers surrounding mineral grains and depending on the nature of the grains; (ii) membrane polarization varying with the grain or pore size distribution, (iii) electrode polarization occurring in the presence of conductive minerals and (iv) the Maxwell-Wagner polarization (generally observed for frequencies >1 kHz) that occurs in multiphase systems when changes in dielectric permittivity and conductivity exist between the different phases. In contrast to ERT, IP allows electrolytic conduction to be distinguished from surface conduction. This is of great interest as soils often have significant surface conductivity.

Challenges and limitations

Polarization depends on numerous parameters such as the type and amount of conductive minerals, the presence of clay minerals, the pore fluid composition, the pore saturation, the presence of insoluble hydrocarbons, the size and shape of the pore space, the size and shape of the grains or the temperature, just like resistivity. While this can be seen as a strength of the IP method, it can also complicate the interpretation of data as these parameters can vary simultaneously. Compared with resistivity, however, the link between the physico-chemical mechanisms described above and the IP response observed in the field or in the laboratory is not yet fully understood. Finally, collecting reliable IP data requires more caution and takes more time than resistivity measurements. Nevertheless, IP reduces ambiguity compared with a DC resistivity measurement.

Nuclear magnetic resonance (NMR)

Nuclear Magnetic Resonance (NMR) measurements quantify molecular properties of matter by exposing atomic nuclei in a magnetic field to electromagnetic radiation. NMR is particularly sensitive to the presence of protons ^1H and is therefore well suited to investigate systems containing water. In soil science, NMR is used to infer permeability, water content, total porosity, pore size distribution and to quantify bound and free fluid fractions. Under an external magnetic field, protons behave like small bar magnets that orient themselves along the direction of the magnetic field. In doing so, they create a macroscopic magnetization (step 0 in Fig. 7) whose magnitude is directly proportional to the proton density and, by extension, to the water content. NMR measurements are classically achieved by producing an oscillating magnetic pulse with a transmitting current coil at a particular frequency known as the Larmor or resonance frequency which tips the magnetization away from the external magnetic field (step 1 in Fig. 7). The resulting transverse magnetization creates in turn a small signal that can be measured by a receiving current coil. When the excitation pulse is switched off, the amplitude of the measured current decreases rapidly while the system progressively returns to equilibrium (steps 2 and 3 in Fig. 7). The amplitude of the signal just after the excitation pulse is directly related to the water content. The relaxation time is influenced by multiple parameters including the pore size distribution. For more information about the NMR principles and applications for near-surface characterization, the interested reader is referred to Behroozmand et al. (2014).

Challenges and limitations

While borehole NMR has long been successfully used to characterize gas and oil reservoirs (Freedman, 2006) and more recently for near-surface hydrogeological and environmental studies (Kirkland and Codd, 2018), surface NMR still remains relatively little used outside the research community, despite its great potential, mainly because of its low signal to noise ratio. To overcome this, further development of the technique is needed. Possible areas of improvements include electromagnetic noise reduction (Kremer et al., 2022) and more efficient instruments and measurements techniques to increase production rate (Behroozmand et al., 2014). Interesting instruments for soil science are, for example the DART system, a slimline borehole NMR, or the single-sided DISCUS equipment.

Seismic methods

Seismic methods involve the measurement of wave-fields produced by artificially or naturally-generated seismic waves that propagate in the subsurface or near the soil surface. An earthquake is a common form of a seismic wave. Typically, seismic measuring arrays involve a transect of seismic sensors (geophones) deployed on the soil surface or along boreholes, and a seismic source is successively activated at different spatial locations. In soil applications, the source often is a hammer impact on a plate (see Fig. 8). The geophones measure the ground movement resulting from body (including pressure and shear) waves propagating through the soil, scattering at soil interfaces and from surface waves propagating at the soil surface. Pressure waves are the fastest seismic waves, produce soil displacement parallel to the direction of propagation and are the smallest in amplitude.

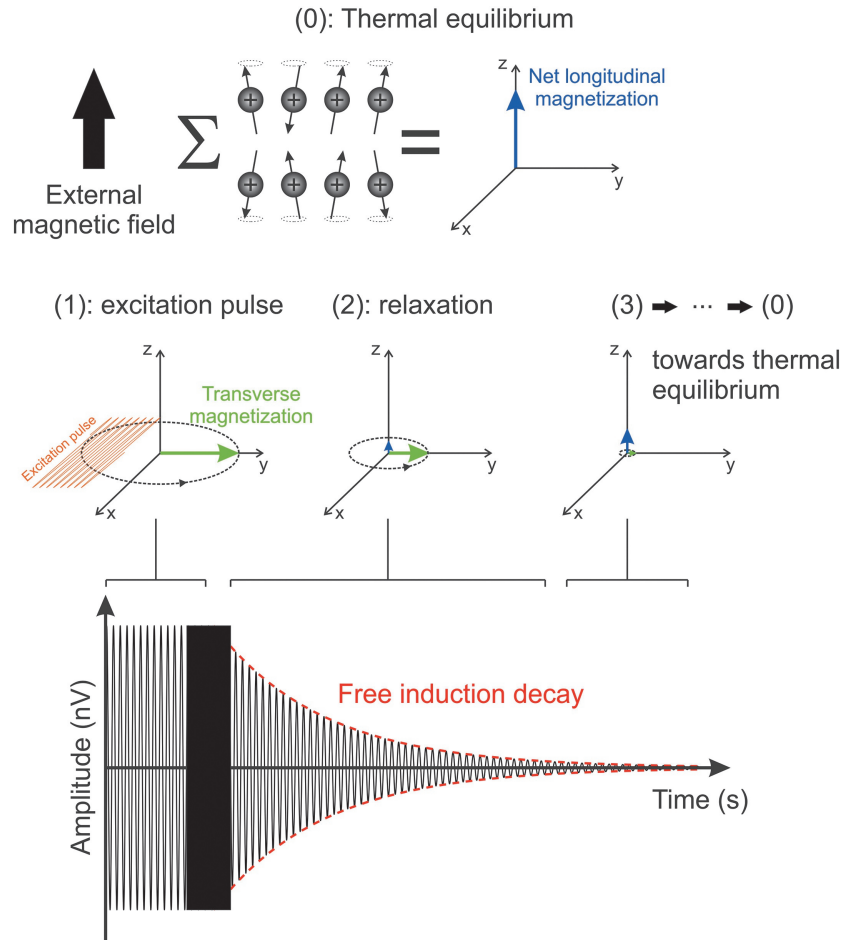


Fig. 7 Protons behave like small bar magnets called spins. Under the effect of an external magnetic field (e.g., the earth's magnetic field), they tend to orient themselves in the direction of the field creating a macroscopic (so-called longitudinal) magnetization (blue arrow in step 0). The amplitude of this magnetization is directly proportional the proton density and therefore to the water content. However, the longitudinal magnetization cannot be quantified as long as it is oriented along the direction of the external field. Therefore, it needs to be tipped away from it. This is achieved by applying an excitation pulse at a particular frequency called the resonance or Larmor frequency (step 1). The resulting oscillating transverse magnetization (green arrow in steps 1 to 3) generates a small signal which in turn can be detected by a receiver coil. At the end of the excitation pulse, the spins progressively realign along the external magnetic field (blue arrow pushes back) and a loss of spins synchronization progressively occurs (green arrow decays)(steps 2 and 3). This results in a decrease of the received signal.

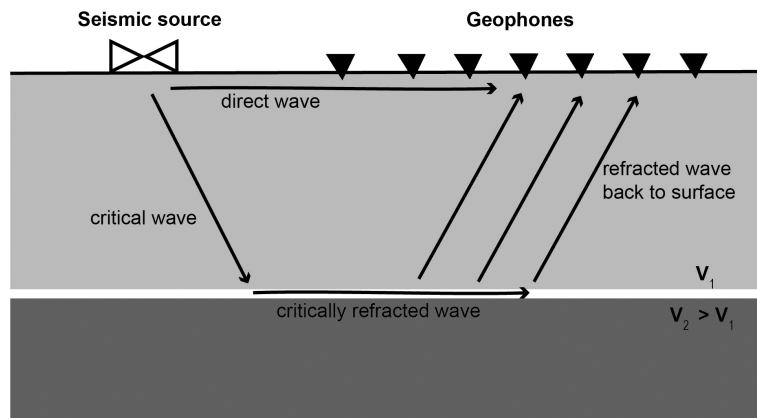


Fig. 8 Basic principles of seismic measurements.

Relative to pressure waves, shear waves and surface waves are slower (both travelling at a similar velocity) but produce higher amplitude soil displacement that is perpendicular to the direction of propagation. Seismic velocities depend on properties of the bulk soil such as resistance to compression and to shear stress, but also on soil moisture and capillary pressure (Romero-Ruiz et al., 2018). For this reason, they provide a direct link to soil mechanical properties (e.g., penetration resistance).

Seismic characterization and imaging of soils can be achieved using a wide variety of methods addressing wave phenomena including refraction and reflection of pressure waves and dispersion of surface waves (Steeple, 2005). High amplitude surface waves dominate the seismograms and there is an increased interest in their application for characterizing soils (e.g., soil compaction and permafrost monitoring (Dou and Ajo-Franklin, 2014)). Nevertheless, pressure waves are the first arriving waves in the seismograms, are easy to detect and, thus, methods that target them are simpler and the most used. For example, seismic refraction tomography, is used to infer a subsurface map of pressure wave velocities (Sheehan et al., 2005). It measures the travel-time of pressure waves that propagate through the soil and are refracted back to the surface after interacting at soil interfaces (e.g., layers). The travel-times are extracted from seismograms and used to infer a pressure wave velocity subsurface map that explains such observations.

Challenges and limitations

Seismic methods are underused in soil applications mainly due to limitations in survey design and interpretation tools. Geophones need to be buried to sense ground movement, and located close to the seismic source (decimeters or a few meters) to achieve the necessary vertical resolution and avoid attenuation; which limits acquisition at large scales. Unlike other geophysical methods, automatic seismic acquisition has seldom been developed nor tested. These restrictions further limit the capacity to monitor active seismic signals over time. In addition, for such small scales, simultaneously arriving pressure, shear and surface waves overlap in the recorder seismograms and are often difficult to disentangle for individual interpretation. Despite the limitations identified, the direct dependence of seismic waves on the soil mechanical properties makes seismic methods highly relevant for soil studies, given that they can resolve mechanical states that are not directly observable by other more commonly used geophysical methods (e.g., ERT, EMI and GPR).

Distributed temperature sensing (DTS)

Distributed Temperature Sensing (DTS) provides temperature measurements at high spatial and temporal resolutions (each 5 to 60 s at 0.1–2 m spacing) along a fiber optic (FO) cable at scales of meters to tens of kilometers. Fig. 9A shows the measurement principle of DTS. The measurement unit emits short laser pulses into the fiber optic cable. As the light pulse travels, a portion of the photons interact with the crystalline structure of the optical fiber, resulting in a backscattered signal returning back toward the DTS unit. As detailed in Fig. 9B, the backscattered signal spectrum has Rayleigh band, Brillouin bands, and Raman bands. Most DTS systems estimate temperature from the Raman backscattering signal, by measuring the returning signal at two distinct wavelengths: the Stokes (-Raman) component and to the anti-Stokes (-Raman) component. Contrary to the Raman Stokes component, the intensity of the Raman Anti-Stokes component depends on the temperature of the surrounding soil. The temperature is estimated from the ratio of the amplitudes of the Stokes to the Anti-Stokes signals. The time of arrival of the backscattered light signal depends on the location of the measurement along the fiber, thus leading to the exact measurement location.

Both passive- or active-DTS surveys are possible (or both methods combined). Passive-DTS continuously monitors natural temperature variations all along the fiber optic cable, while active-DTS surveys monitor induced temperature changes by injecting an artificial heat source in the soil or cable. Soil temperature is particularly variable in time and space. Passive-DTS is a great tool to monitor those spatio-temporal temperature dynamics at high spatial and temporal resolutions over long distances. In addition, both passive- and active-DTS experiments can be used to estimate soil thermal properties. As soil thermal properties depend on moisture content, monitoring temperature changes through fiber optic DTS not only provide estimates of thermal conductivity (Abesser et al., 2020), and thermal diffusivity but also of soil moisture content (Ciocca et al., 2012).

Challenges and limitations

While active-DTS measurements have been used effectively to estimate soil thermal conductivity, their use to estimate water contents and soil heat capacity or diffusivity remains challenging. One of the reasons is the structure of fiber optic cables, which influences the temperature increase for early periods in time of heating, and affects the analysis of the thermal response curve, used to estimate thermal properties and water contents. This issue can be addressed by applying longer heating and cooling periods. Unfortunately, prolonged-time heat injection can significantly disturb the soil water content field and then reduce the accuracy of the soil water content estimate (Wang et al., 2020).

One of the main limitations to using this method to study soil thermal properties is that the deployment of fiber optic cables in soils may significantly disturb their structures and lead to preferential pathways for infiltration. Hence, FO-cable installation is a critical point for field soil applications of DTS methods. Since heat transfers are different in soils and in water or air, the contact between the soil and the cable must be perfect to avoid air and water pockets, that highly affect temperature response. Likewise, accurate estimates of the depth installation of the cable are required, especially to estimate soil properties from passive-DTS measurements (Steele-Dunne et al., 2010). To do so, plows can be used to ensure the proper installation of the cable and control its burial depth (Simon et al., 2022).

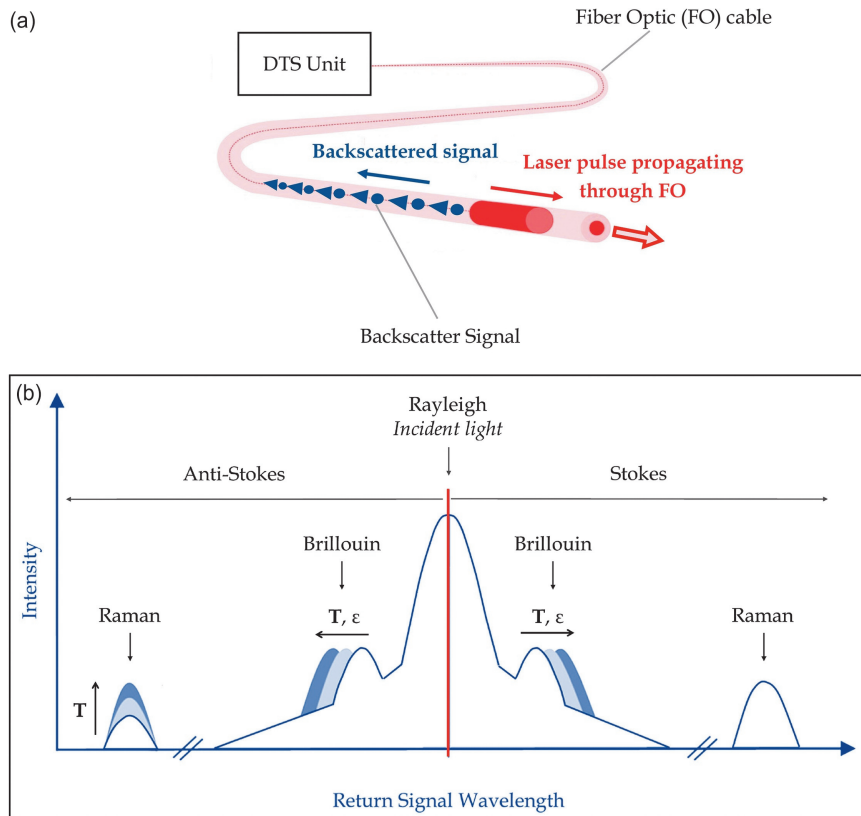


Fig. 9 (A) Principle of FO-DTS measurements. (B) Spectrum of the backscattered light returning back toward the DTS unit. Apart from the Rayleigh band which has the same wavelength as the incident laser pulse, the backscattered signal is shifted toward higher and lower wavelengths, called Stokes and Anti-Stokes components respectively. Stokes and Anti-Stokes signals result from Brillouin and Raman scatterings. Brillouin Stokes and Anti-Stokes components are both sensitive to strain (ϵ) and temperature (T) variations whereas, for Raman bands, only the amplitude of the Anti-Stokes component is temperature-dependent.

Forward modelling and inversion

Most of the measured geophysical properties are 'apparent', i.e. represent a volume-average of heterogeneous soil properties for a given arrangement of emitters and receivers. Those apparent properties require a data inversion process (see Fig. 10) to obtain an estimation of their 'real' distribution in the soil. In such a process, we use a 'forward model' which simulates the measurements one would measure for a given distribution of subsurface properties. Once we have those simulated data, we compare them with the real data. This comparison is typically done with an objective function, which we minimize iteratively by running the forward model again for different properties of the subsurface (obtained, for instance, by using the gradient of the objective function with respect to the model properties) until the difference between the simulated and real data is minimal. In many cases, several distributions will fit the data equally well, so most inversion algorithms have to add additional constraints (e.g., a smoothness constraint) in the objective function to select the final estimated distribution.

In addition, the sensitivity, resolution and scales of the different geophysical methods can vary significantly. A combination of techniques is, therefore, often necessary to extract variations of one specific soil property or state variable. Instead of treating every data source independently, translating them to soil properties or variables and using that for interpretation, they can be combined in a joint interpretation scheme. Hinnell et al. (2010) discussed several approaches in which geophysical data and models and hydrological models are more or less coupled, so that the physics of the soil system can constrain the interpretation of the geophysical data.

Attempts for joint inversion of geophysical methods only started to be reported in the last 15–20 years. Yaramanci et al. (2002) used the optimization method of simulated annealing to jointly invert SNMR and vertical electrical sounding data for both synthetic and real cases, eventually inferring mobile and immobile water contents. Doetsch et al. (2010) used time-lapse borehole ERT and GPR to track an injected saline plume. They looked at the coupling between the two methods through constraints in the inversion scheme. Jardani et al. (2013) combined ERT and SP in a stochastic framework, i.e., looking for property statistics rather than its actual deterministic distribution. In their case, the inversion was of the hydrological problem (flow and transport model, looking primarily at the hydraulic conductivity).

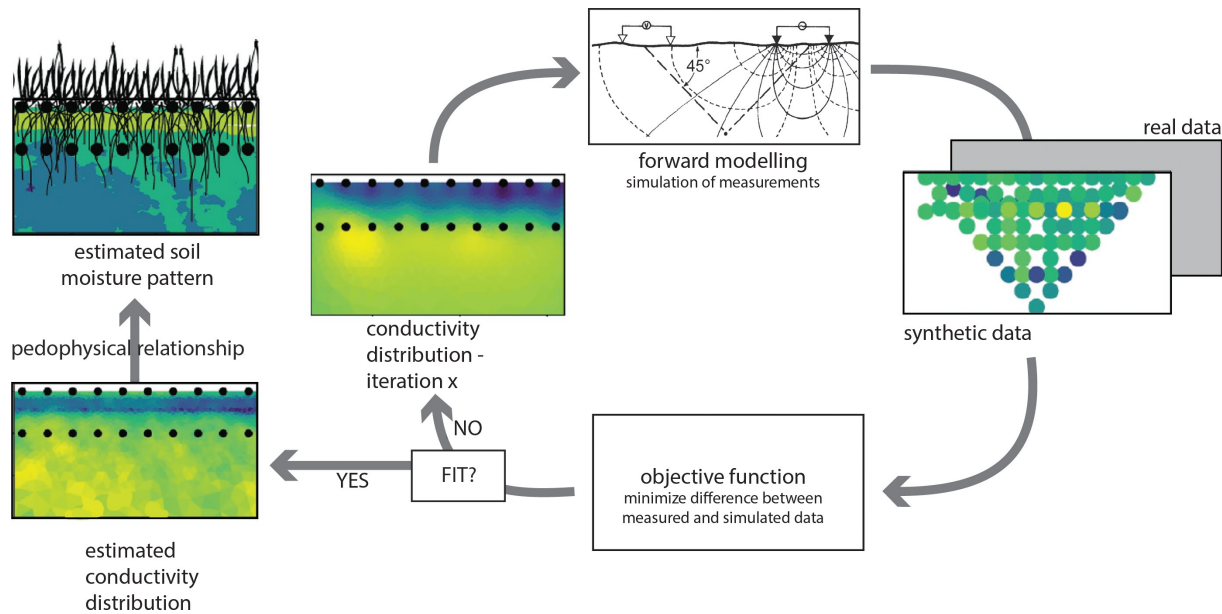


Fig. 10 Schematic overview of a forward modelling and inversion pipeline in geophysics.

Data standardization and open science

Several geophysical methods originated from the commercial oil and mineral exploration industry. As such, instruments are mostly closed source and often lack standardization, manufacturers tending to produce instrument-specific data formats. On the hardware side, the black box model of commercial instruments can sometimes be a barrier. For instance, EMI instruments often need to be corrected for drift and calibration on top of the manufacturer calibration and without knowing about the hardware-born causes of the drift. Another example is from GPR systems that typically provide no information about the emitting source. More transparency on the internal specifications of commercial instruments and open hardware instruments (e.g. OhmPi resistivity meter) built on top of open-source platforms (e.g., Raspberry Pi, Arduino) could alleviate these issues.

Geophysical data is increasingly processed using open-source software. These software packages can be bound to a geophysical method, such as the open-source FDEM1DFWD or FEMIC codes that propose forward models for electromagnetic induction. Alternatively, multi-method geophysical frameworks such as SimPEG or pyGIMLI enable processing of different types of geophysical methods together (coupled or joint inversion). Other open-source projects such as ResIPy, EMagpy, GPRpy propose a graphical user-interface to process resistivity, EMI and GPR data respectively. This makes them more accessible for non-specialists and in an educational context. The Software Underground (<https://softwareunderground.org/>) project summarizes several initiatives about open-source software for geophysics.

One of the main challenges of geophysics for soil applications remains the use of pedophysical relationships to link geophysical variables to soil properties. These relationships are often site-specific, can present hysteresis and are subjected to heterogeneity. Building worldwide databases of pedophysical relationships from lab and field measurements can help in building more robust relationships but also to develop standards (e.g., Catalog of AgroGeophysical Studies, <https://agrogeophy.github.io/catalog/>).

Conclusion

Geophysical techniques map a wide range of proxies for soil properties and state variables. Through this capacity, geophysical data complement precise (invasive) measurements with 2- or 3-D continuous information, while equally bridging the gap between point observations and coarse satellite data. Their non-invasive character allows for timelapse surveys for monitoring programs.

Despite these advantages, the multiscale heterogeneity of soils burdens the interpretation of geophysical data, exacerbating, for instance, the use of pedophysical relationships to quantify specific soil properties. To advance the application of soil geophysics, more data and measurement standards are needed, alongside benchmarking of processing algorithms and open science initiatives.

Nevertheless, as the sole means for non-invasive soil investigation over a large depth range, it is unthinkable to imagine geophysics apart from soil science. As an integral part of field research, especially combined with complementary data types, the use of geophysics in soil applications will grow more relevant with time.

References

- Abesser C, Ciocca F, Findlay J, Hannah D, Blaen P, Chalari A, Mondanos M, and Krause S (2020) A distributed heat pulse sensor network for thermo-hydraulic monitoring of the soil subsurface. *Quarterly Journal of Engineering Geology and Hydrogeology* 53(3): 352–365. <https://doi.org/10.1144/qjegh2018-147>. ISSN 1470-9236.
- André F, van Leeuwen C, Saussez S, Van Durmen R, Bogaert P, Moghadas D, de Ress'eguié L, Delvaux B, Vereecken H, and Lambot S (2012) High-resolution imaging of a vineyard in south of France using ground-penetrating radar, electromagnetic induction and electrical resistivity tomography. *Journal of Applied Geophysics* 78: 113–122. <https://doi.org/10.1016/j.jappgeo.2011.08.002>. ISSN 0926-9851.
- Annan A (2009) *Electromagnetic Principles of Ground Penetrating Radar*. Elsevier. pp. 1–40. <https://doi.org/10.1016/b978-0-444-53348-7.00001-6>.
- Behroozmand AA, Keating K, and Auken E (2014) A review of the principles and applications of the NMR technique for near-surface characterization. *Surveys in Geophysics* 36(1): 27–85. <https://doi.org/10.1007/s10712-014-9304-0>. ISSN 0169-3298.
- Binley A, Hubbard SS, Huisman JA, Revil A, Robinson DA, Singha K, and Slater LD (2015) The emergence of hydrogeophysics for improved understanding of subsurface processes over multiple scales. *Water Resources Research* 51(6): 3837–3866. <https://doi.org/10.1002/2015wr017016>. ISSN 0043-1397.
- Callegary JB, Ferr'e TPA, and Groom RW (2007) Vertical spatial sensitivity and exploration depth of low-induction-number electromagnetic-induction instruments. *Vadose Zone Journal* 6(1): 158–167. <https://doi.org/10.2136/vzj2006.0120>. ISSN 1539-1663.
- Cassidy NJ (2009) *Ground Penetrating Radar Data Processing, Modelling and Analysis*. Elsevier, pp. 141–176. <https://doi.org/10.1016/b978-0-444-53348-7.00005-3>.
- Chapman H, Adcock J, and Gater J (2009) An approach to mapping buried prehistoric palaeosols of the Atlantic seaboard in Northwest Europe using GPR, geoarchaeology and GIS and the implications for heritage management. *Journal of Archaeological Science* 36(10): 2308–2313. <https://doi.org/10.1016/j.jas.2009.06.015>. ISSN 0305-4403.
- Ciocca F, Lunati I, Van de Giesen N, and Parlange MB (2012) Heated optical fiber for distributed soil-moisture measurements: A lysimeter experiment. *Vadose Zone Journal* 11(4). <https://doi.org/10.2136/vzj2011.0199>. ISSN 1539-1663.
- Declercq Y, Samson R, Castanheiro A, Spassov S, Tack FM, Van De Vijver E, and De Smedt P (2019) Evaluating the potential of topsoil magnetic pollution mapping across different land use classes. *Science of the Total Environment* 685: 345–356. <https://doi.org/10.1016/j.scitotenv.2019.05.379>. ISSN 0048-9697.
- Delefortrie S, De Smedt P, Saey T, Van De Vijver E, and Van Meirvenne M (2014) An efficient calibration procedure for correction of drift in EMI survey data. *Journal of Applied Geophysics* 110: 115–125. <https://doi.org/10.1016/j.jappgeo.2014.09.004>. ISSN 0926-9851.
- Delefortrie S, Hanssens D, and De Smedt P (2018) Low signal-to-noise FDEM in-phase data: Practical potential for magnetic susceptibility modelling. *Journal of Applied Geophysics* 152: 17–25. <https://doi.org/10.1016/j.jappgeo.2018.03.003>. ISSN 0926-9851.
- Doetsch J, Linde N, and Binley A (2010) Structural joint inversion of time-lapse crosshole ERT and GPR traveltime data. *Geophysical Research Letters* 37(24). <https://doi.org/10.1029/2010gl045482>. ISSN 0094-8276.
- Doro KO, Emmanuel ED, Adebayo MB, Bank C-G, Wescott DJ, and Mickleburgh HL (2022) Time-lapse electrical resistivity tomography imaging of buried human remains in simulated mass and individual graves. *Frontiers in Environmental Science* 10. <https://doi.org/10.3389/fenvs.2022.882496>. ISSN 2296-665X.
- Dou S and Ajo-Franklin JB (2014) Full-wavefield inversion of surface waves for mapping embedded low-velocity zones in permafrost. *Geophysics* 79(6): EN107–EN124. <https://doi.org/10.1190/geo2013-0427.1>. ISSN 0016-8033.
- Everett ME (2013) *Near-Surface Applied Geophysics*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139088435>.
- Freedman R (2006) Advances in NMR logging. *Journal of Petroleum Technology* 58(01): 60–66. <https://doi.org/10.2118/89177-jpt>. ISSN 0149-2136.
- Grimley DA, Wang J-S, Liebert DA, and Dawson JO (2008) Soil magnetic susceptibility: A quantitative proxy of soil drainage for use in ecological restoration. *Restoration Ecology* 16(4): 657–667. <https://doi.org/10.1111/j.1526-100x.2008.00479.x>. ISSN 1061-2971.
- Hanssens D, Delefortrie S, Bobe C, Hermans T, and De Smedt P (2019) Improving the reliability of soil EC-mapping: Robust apparent electrical conductivity (rECa) estimation in ground-based frequency domain electromagnetics. *Geoderma* 337: 1155–1163. <https://doi.org/10.1016/j.geoderma.2018.11.030>. ISSN 0016-7061.
- Hinnell AC, Ferr'e TPA, Vrugt JA, Huisman JA, Moysey S, Rings J, and Kowalsky MB (2010) Improved extraction of hydrologic information from geophysical data through coupled hydrogeophysical inversion. *Water Resources Research* 46(4). <https://doi.org/10.1029/2008wr007060>. ISSN 0043-1397.
- Huisman JA, Hubbard SS, Redman JD, and Annan AP (2003) Measuring soil water content with ground penetrating radar: A review. *Vadose Zone Journal* 2(4): 476–491. <https://doi.org/10.2136/vzj2003.4760>. ISSN 1539-1663.
- Jardani A, Revil A, and Dupont J (2013) Stochastic joint inversion of hydrogeophysical data for salt tracer test monitoring and hydraulic conductivity imaging. *Advances in Water Resources* 52: 62–77. <https://doi.org/10.1016/j.advwatres.2012.08.005>. ISSN 0309-1708.
- Jordanova N (2017) *Soil Magnetism: Applications in Pedology, Environmental Science and Agriculture*. London, UK: Academic Press.
- Jordanova D, Jordanova N, and Petrov P (2014) Pattern of cumulative soil erosion and redistribution pin-pointed through magnetic signature of Chernozem soils. *Catena* 120: 46–56. <https://doi.org/10.1016/j.catena.2014.03.020>. ISSN 0341-8162.
- Kemna A, Binley A, Cassiani G, Niederleithinger E, Revil A, Slater L, Williams KH, Orozco AF, Haegel F-H, Horst A, Kruschwitz S, Leroux V, Titov K, and Zimmermann E (2012) An overview of the spectral induced polarization method for near-surface applications. *Near Surface Geophysics* 10(6): 453–468. <https://doi.org/10.3997/1873-0604.2012027>. ISSN 1569-4445.
- Kirkland CM and Codd SL (2018) Lowfield borehole NMR applications in the near surface environment. *Vadose Zone Journal* 17(1): 1–11. <https://doi.org/10.2136/vzj2017.01.0007>. ISSN 1539-1663.
- Klose T, Guilleminot J, Simon F-X, and Tronicke J (2018) Toward subsurface magnetic permeability imaging with electromagnetic induction sensors: Sensitivity computation and reconstruction of measured data. *Geophysics* 83(5): E335–E345. <https://doi.org/10.1190/geo2017-0827.1>. ISSN 0016-8033.
- Knight R (2001) Ground penetrating radar for environmental applications. *Annual Review of Earth and Planetary Sciences* 29(1): 229–255. <https://doi.org/10.1146/annurev.earth.29.1.229>. ISSN 0084-6597.
- Kremer T, Irons T, Müller-Petke M, and Larsen JJ (2022) Review of Acquisition and Signal Processing Methods for Electromagnetic Noise Reduction and Retrieval of Surface Nuclear Magnetic Resonance Parameters. *Surveys in Geophysics*. <https://doi.org/10.1007/s10712-022-09695-3>. ISSN 0169-3298.
- Ma R, McBratney A, Whelan B, Minasny B, and Short M (2010) Comparing temperature correction models for soil electrical conductivity measurement. *Precision Agriculture* 12(1): 55–66. <https://doi.org/10.1007/s11119-009-9156-7>. ISSN 1385-2256.
- Mathé V and Lévêque F (2003) High resolution magnetic survey for soil monitoring: detection of drainage and soil tillage effects. *Earth and Planetary Science Letters* 212(1–2): 241–251. [https://doi.org/10.1016/s0012-821x\(03\)00241-3](https://doi.org/10.1016/s0012-821x(03)00241-3). ISSN 0012-821X.
- Miroshnichenko M, Kruglov O, Nazarov P, and Kovalenko S (2021) *Identification of the Structure of Soil Cover by Magnetic Susceptibility*. Springer International Publishing 57–68. https://doi.org/10.1007/978-3-030-68394-8_6.
- Rogers MB, Baham JE, and Dragila MI (2006) Soil iron content effects on the ability of magnetometer surveying to locate buried agricultural drainage pipes. *Applied Engineering in Agriculture* 22(5): 701–704. <https://doi.org/10.13031/2013.22001>. ISSN 1943-7838. [Online; accessed 2022-08-31].
- RomeroRuiz A, Linde N, Baron L, Solazzi SG, Keller T, and Or D (2021) Seismic signatures reveal persistence of soil compaction. *Vadose Zone Journal* 20(4). <https://doi.org/10.1002/vzj2.20140>. ISSN 1539-1663.
- Romero-Ruiz A, Linde N, Keller T, and Or D (2018) A Review of Geophysical Methods for Soil Structure Characterization. *Reviews of Geophysics* 56(4): 672–697. <https://doi.org/10.1029/2018rg000611>. ISSN 8755-1209.
- Roth K, Schulín R, Flüher H, and Attinger W (1990) Calibration of time domain reflectometry for water content measurement using a composite dielectric approach. *Water Resources Research* 26(10): 2267–2273. <https://doi.org/10.1029/wr026i010p02267>. ISSN 0043-1397.

- Sheehan JR, Doll WE, and Mandell WA (2005) An evaluation of methods and available software for seismic refraction tomography analysis. *Journal of Environmental and Engineering Geophysics* 10(1): 21–34. <https://doi.org/10.2113/jee10.1.21>. ISSN 1083-1363.
- Simon N, Bour O, Faucheux M, Lavenant N, Le Lay H, Fovet O, Thomas Z, and Longuevergne L (2022) Combining passive and active distributed temperature sensing measurements to locate and quantify groundwater discharge variability into a headwater stream. *Hydrology and Earth System Sciences* 26(5): 1459–1479. <https://doi.org/10.5194/hess-26-1459-2022>. ISSN 1607-7938.
- Steele-Dunne SC, Rutten MM, Krzeminska DM, Hausner M, Tyler SW, Selker J, Bogaard TA, and van de Giesen NC (2010) Feasibility of soil moisture estimation using passive distributed temperature sensing. *Water Resources Research* 46(3). <https://doi.org/10.1029/2009wr008272>. ISSN 0043-1397.
- Steeple DW (2005) *Shallow Seismic Methods*. Dordrecht: Springer Netherlands 215–251. https://doi.org/10.1007/1-4020-3102-5_8.
- Thiesson J, Kessouri P, Schamper C, and Tabbagh A (2014) Calibration of frequencydomain electromagnetic devices used in nearsurface surveying. *Near Surface Geophysics* 12(4): 481–491. <https://doi.org/10.3997/1873-0604.2014012>. ISSN 1569-4445.
- Verdonck L, Launaro A, Vermeulen F, and Millett M (2020) Ground-penetrating radar survey at Falerii Novi: A new approach to the study of Roman cities. *Antiquity* 94(375): 705–723. <https://doi.org/10.15184/aqy.2020.82>. ISSN 0003-598X.
- Wang M, Li X, Chen L, Hou S, Wu G, and Deng Z (2020) A modified soil water content measurement technique using actively heated fiber optic sensor. *Journal of Rock Mechanics and Geotechnical Engineering* 12(3): 608–619. <https://doi.org/10.1016/j.jrmge.2019.11.003>. ISSN 1674-7755.
- Waxman M and Smits L (1968) Electrical conductivities in oil-bearing shaly sands. *Society of Petroleum Engineers Journal* 8(02): 107–122. <https://doi.org/10.2118/1863-a>. ISSN 0197-7520.
- Yaramanci U, Lange G, and Hertrich M (2002) Aquifer characterisation using Surface NMR jointly with other geophysical techniques at the Nauen/Berlin test site. *Journal of Applied Geophysics* 50(1–2): 47–65. [https://doi.org/10.1016/S0926-9851\(02\)00129-5](https://doi.org/10.1016/S0926-9851(02)00129-5). ISSN 0926-9851.

Non-invasive imaging of soil processes

Craig J Sturrock, Division of Agricultural and Environmental Sciences, School of Biosciences, University of Nottingham, Loughborough, United Kingdom

© 2023 Elsevier Ltd. All rights reserved.

Introduction	459
Computed tomography	460
X-ray computed tomography	460
Key principles	460
Types of X-ray CT system	460
X-ray source	461
Manipulator stage	462
Detector	462
Image analysis	462
Gamma-ray tomography	464
Neutron tomography	464
Magnetic resonance imaging	464
Positron emission tomography	466
Applications	467
Summary and outlook	467
Appendix: Supplementary material	467
References	467

Abstract

Non-invasive imaging techniques provide a unique ability to investigate soil processes with minimal disturbance to its natural state. This chapter provides an overview of the key non-invasive techniques, their operating principals and their areas of application to quantify soil processes. e.g. structural development of soil pore networks, plant root system architecture, rhizosphere interactions, water and chemical movement. The digital nature of the data integrates well with structural modelling approaches of soil processes from both laboratory and field-based experiments.

Key points

- Non-invasive imaging allows for direct quantification of soil structure and associated measures of processes in three dimensions
- Collection of 3D digital data of soils either field structured or repacked facilitates modelling of soil process
- Combination of different modalities at multiple scales, using correlative imaging approaches, offer exciting prospects for understanding the physical, chemical and biological behavior of soils in the future

Introduction

The study of soil pore structure and its impact on many soil processes has historically relied on the use of indirect measurements (e.g., soil water retention characteristic, nitrogen sorptivity and mercury porosimetry), primarily due to the inherent opacity of the soil matrix. Such methods tend to simplify the system, assuming that pores are cylindrical in shape and therefore overlook the complex heterogenous nature of the soil ecosystem. Direct imaging approaches to quantify soil pore micromorphology based on the use of soil thin sections have been widely used since the 1940s. However, the preparation of soil thin sections is highly laborious, destructive and generates often discontinuous 2D images, producing limited information on the connectivity of pore networks. Over the last few decades, there has been significant development and application of non-invasive imaging techniques, which have proved to be invaluable in gaining new insights and understanding of processes in soil. For example, the role of soil structure, operating across spatial and temporal scales, has a direct consequence on processes such as gas and water flow, plant root development, root water and nutrient uptake, carbon and nutrient cycling and biological activity (Young et al., 2001). The multiple interactions between these processes mean that soil structure is dynamic with the connectivity, tortuosity and heterogeneity of the 3D soil pore network changing over time. The study of such processes therefore requires methods that capture the 3D nature of the soil system. There are several different non-invasive techniques, each with their own advantages, for studying different parts of the soil system. In this chapter, the main non-invasive imaging techniques to assess soil processes in 3D will be introduced and an overview of their key principles, strengths, and main applications to soils will be provided. These techniques are Computed

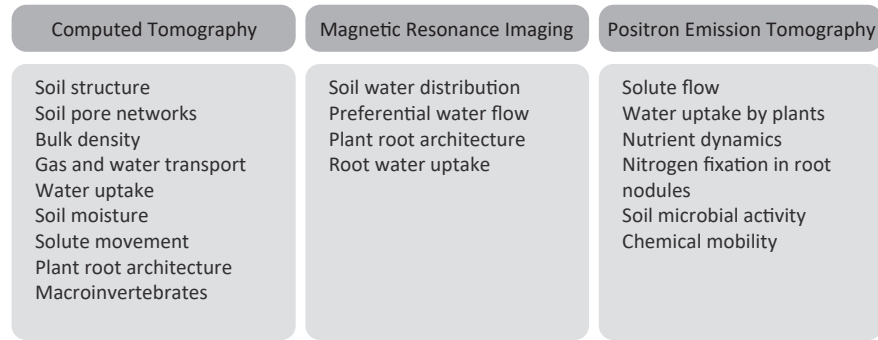


Fig. 1 Applications of 3D non-invasive imaging techniques to soil processes.

Tomography (CT), Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET). Each of these three methods differ in the way they generate images of the material under investigation as well as their potential applications. Fig. 1 provides an overview of some of the key areas of application of each of these techniques. Geoelectrical methods, such as Electrical Resistivity Imaging (EIT), which by proxy can non-invasively measure the spatial and temporal variability of soil processes especially in the field will not be covered in this chapter. Please see Loke et al. (2013) and Cimpoiu et al. (2020) for reviews of EIT.

Computed tomography

Computed Tomography (CT) is a non-invasive imaging technique that allows the measurement of the interior and exterior structure of an object in three dimensions. The technique is based on calculating the attenuation of a radiation source as it passes through a sample. A series of projection images at incremental angles through the sample are collected and then mathematic algorithms are used to compute multiple cross-sectional images that map the attenuation of the incident radiation. Post scanning, computer workstations are used to run the algorithms in a process known as reconstruction. The name tomography is derived from the Greek words tomos, meaning 'to slice or section' and graphein, which means 'to write.' There are different forms of CT defined by the type of radiation that is used to collect the images e.g. X-ray, gamma or neutron tomography. As each type of radiation has different interaction properties with the constituent materials of the sample, each type of CT offers a range of application purposes in the study of soil processes.

X-ray computed tomography

X-ray Computed Tomography (XCT) is the most commonly used non-invasive imaging technique in soil science. It was originally developed as a medical diagnostic tool in the late 1960's by Sir Godfrey Hounsfield who designed the first commercial whole-body medical CT scanner (EMI Mk1) in 1972.

Key principles

X-ray Computed Tomography allows measurement of the X-ray attenuation coefficient of a material (e.g. a soil core) by calculating the reduction in intensity of an incident X-ray beam (I_0) to that which has been transmitted through the material under investigation (I). This relationship is described by Lambert-Beer's law:

$$I = I_0 \exp(-\mu D) \quad (1)$$

where μ is the linear X-ray attenuation coefficient (L^{-1}) and D is the thickness of the material (L^{-1}). The linear attenuation coefficient (μ) is a function of the energy of the incident X-ray beam and both the physical density and atomic number of the constituent components of the material. As soils fundamentally contain multiple phases in nature, Eq. (1) can be extended to account for the differences in the linear attenuation coefficients for air, water and solid components of the soil:

$$I = I_0 \exp \left(- \left[(1 - \theta_p) \mu_s \rho_s D + \theta_p S_w \mu_w \rho_w D \right] \right). \quad (2)$$

ρ_s and ρ_w are the densities and μ_s and μ_w are the linear attenuation coefficients of solid and water, respectively. θ_p denotes the total porosity and S_w the water saturation. As air has a very low attenuation and therefore a negligible contribution to the overall linear attenuation coefficient it is omitted from the equation.

Types of X-ray CT system

There are three main categories of X-ray CT system; medical, industrial and synchrotron based (Fig. 2). Each of these categories have a slightly different configurations but share four common components (i) X-ray source (either conventional X-ray tube or synchrotron light), (ii) a sample manipulator stage or gantry, (iii) an X-ray detector, and (iv) a computer for acquisition,



Fig. 2 Examples of different types of X-ray-based Computed Tomography system. (A) Medical CT scanner at Queens Medical Centre, Nottingham, (B) Industrial X-ray CT system at Hounsfield Facility, University of Nottingham, (C) Internal components of industrial CT system ([right to left] X-ray tube, soil core in rotating stage and detector panel) and (D) Synchrotron Light Source Facility where tomography is one of the many techniques used for investigation of samples. (A) Courtesy Sacha Mooney, University of Nottingham. (B–D) www.diamond.ac.uk © Diamond Light Source Ltd. 2016.

reconstruction, and downstream analysis of the image data. The configuration of these components has important consequences for the size of object that can be scanned and the spatial resolution of the images. Imaging resolution is dictated by the size of the X-ray beam focal spot, the size and number of pixel elements of the detector, and the distances of the detector and the sample from the X-ray focal spot. As medical CT systems are designed around imaging a human body (meter to decimeter size range), the source-object-detector configuration differs from that of industrial and synchrotron-based CT, where the X-ray source and detector rotate around the sample (patient), whereas it is the sample that rotates on a manipulator stage with source and detector remaining static in the latter two (Fig. 2C). Medical CT resolution is generally limited to 1–2 mm, so most applications of CT to soil science in the last 10 years use industrial or synchrotron-based systems. These generally offer a higher spatial and contrast resolution, which permit discrimination of individual soil phases, enabling the resolution of microscopic pore characteristics (often with feature recognition $<2\text{--}3\text{ }\mu\text{m}$). However, with increasing levels of spatial resolution the field of view of the scan is generally reduced. For example, typical scanning resolution of a 5 cm diameter soil core would be approximately $20\text{--}30\text{ }\mu\text{m}$ on an industrial CT system. However, to image at higher resolution (e.g. $<5\text{ }\mu\text{m}$) a much smaller, ‘soil aggregate scale’ sample would be needed. A wide range of industrial CT systems are now available in the marketplace that offer versatility in sample size, resolution and X-ray power. Furthermore, with synchrotron light systems and some industrial systems (e.g. 3D X-ray microscopes), optical magnification is possible to enhance image sharpness and resolution. In some cases, advanced scanning modes are available in these systems which allow higher magnification scans of a larger object (e.g. region-of-interest scanning). For industrial and synchrotron systems, the X-ray source, stage, and detector are housed within an appropriately shielded lead cabinet or ‘hutch’ to prevent accidental exposure of the operator of the system to ionizing radiation. A further description of the key parts of an X-ray CT system is provided below.

X-ray source

X-rays are generated by either an X-ray tube or a synchrotron light source. X-ray tubes used in medical or industrial CT are mostly based on the Coolidge tube, also known as a hot cathode tube, where the thermionic effect is used for electron production. The evacuated chamber of the tube contains a cathode and anode. The cathode is a tungsten wire filament that is heated by the supply of a low voltage from a generator to release electrons. By application of a high voltage between the cathode and anode the electrons are accelerated to strike the anode (target) at the end of the tube. As the electrons enter the electric field of the anode material atoms, they are rapidly slowed down resulting in a loss of their kinetic energy and release of X-ray radiation and heat. This radiation is called *bremsstrahlung* (German for braking radiation). Secondly, characteristic radiation is also produced if an accelerated electron directly strikes and ejects a bound electron from the inner atomic shell of the anode material. X-rays are produced when an electron, from

the outer atomic shell, falls back to the inner shell due to its reduction in energy. X-ray tubes can be supplied with an anode made from different materials (e.g. tungsten, molybdenum, copper) which will produce characteristic radiation with a different energy spectrum. This is useful depending on the composition of the sample to be investigated as molybdenum or copper targets can be used to image materials with lower densities such as plant tissue. The combination of the two processes of X-ray production result in an X-ray beam that is polychromatic, thus displaying a range of wavelengths (and therefore energies) depending on the maximum acceleration voltage and current of the electron beam and the anode material.

X-ray tube type in industrial CT systems are defined by whether the electron beam is transmitted through or reflected off the target. Transmission tubes offer a greater level of control and precision to focus of the electron beam to achieve sub-micron focal spots size, providing benefits in image sharpness for imaging at high resolutions (0.5–2 μm). However, they are limited in power and therefore not as well suited for larger samples where higher X-ray energies and flux are required to fully penetrate the sample. Micro-focus (e.g. focal spot typically $>3 \mu\text{m}$) CT systems typically have a reflection target X-ray tube which is positioned at an approximate 45° angle inside the cabinet and the electron beam is reflected off the surface of the X-ray target (Fig. 2C). Generally, higher power X-rays are required to penetrate larger samples or samples consisting of more electron dense materials. For example, using an industrial system, although the X-ray tube energy may be in the 70–120 kV range for a small soil aggregate (2–5 cm diameter), much higher energies (approximately 150–200 kV range) are required to achieve sufficient X-ray penetration for larger soil cores (e.g. 5–10 cm diameter). Many industrial CT manufacturers now offer to supply their system with dual tube configurations, thereby extending the operational functionality of the system and permitting both nanofocus and microfocus imaging capability.

Synchrotron light sources are a type of electron accelerator that produce very high brilliance X-rays. The generated X-ray beam is parallel (unlike cone or fan beams) and can be tuned to a specific wavelength (monochromatic) to enhance the image quality beyond what is capable on a lab-based industrial CT system. There are >50 synchrotron light facilities globally which range in size (main storage ring can be 0.5 to >2 km in circumference) and energy level. For example, Diamond Light Source, UK (<http://www.diamond.ac.uk/>), is a medium energy level source (3 GeV) with a main storage ring circumference of 562 m, and 33 beamlines (see Fig. 2D). Each beamline is allocated to a different type of experimental technique which broadly falls into three categories (diffraction, imaging or spectroscopy). In addition to tomographic imaging of soil structure, synchrotron beamlines can be used to map the chemical soil environment using absorption, fluorescence and diffraction spectroscopy techniques often with ultra-high elemental sensitivity and resolution (nanoscale). Detailed reviews of soil, plant and environmental science applications using synchrotron light are provided by Burke et al. (2015) and Lombi and Susini (2009).

Manipulator stage

The manipulator is used to securely hold and position the sample during the scan. It consists of a mechanical multi-axis CNC stage. The stage allows rotation of the sample to enable collection of a series of projection images at different oblique angles to the X-ray source. In industrial systems magnification is geometric so moving sample closer to the X-ray source will result in a higher magnification. Motion in the vertical axis facilitates imaging of longer samples such as soil cores.

Detector

Measurement of X-ray attenuation by the sample is achieved using an X-ray detector. The detector essentially captures X-rays and converts their intensity into an electrical signal. There have been significant advances in detector technology, especially in recent years. In modern industrial X-ray CT systems, flat-panel scintillation detectors are most common. These detectors offer higher sensitivity, improved signal-to-noise and quantum efficiency, larger dynamic range and increased number and size of the individual pixels compared to previous generation of scanner. All these factors contribute to improved spatial and contrast resolution, enhanced feature detection capabilities and ever faster scan times.

Image analysis

Irrespective of imaging modality, a key part of all non-invasive imaging techniques is the discrimination of discrete phases in the images. Broadly in soils, these phases are air (pores), water and solid material, however the solid material may consist of a wide range of materials from primary soil particles (sand, silt and clay) to plant roots, seeds, residues, particulate organic material and macroinvertebrates. With CT, the reconstructed images which map the X-ray attenuation co-efficient of these phases, are typically represented in either 8, 16 or 32-bit grayscale image format, with low attenuating materials (e.g. air) corresponding to low gray values and highly attenuating (e.g. mineral grains with high atomic numbers) corresponding to high gray values (e.g. for an 8-bit grayscale CT image low attenuation [0 = black] to high attenuation [255 = white] (See Fig. 3). The number of individual images depends on the number of pixels of the detector and generally range between 1000 and 3000. The spatial structure of the soil pore network and other components can be visualized and quantified using 3D image analysis techniques (Fig. 4). This typically follows a series of stepwise processes involving image contrast enhancement, noise reduction, material segmentation or thresholding (via pixel classification), quantification and visualization (See Fig. 5). There are numerous different methods of noise reduction and threshold algorithms available and the final numerical output of the image analysis workflow is heavily dependent on each of these steps, especially the choice of thresholding algorithm (e.g. see Baveye et al. (2010)). Therefore, caution must be taken to ensure the image analysis workflow is appropriate for your own images and specific application as differences in soil properties (geological background, texture, moisture, and organic matter content) and scanner settings can impact on accuracy of the measurements. In the last 5 years, the development of artificial intelligence methods based on convolutional neural networks have become increasingly

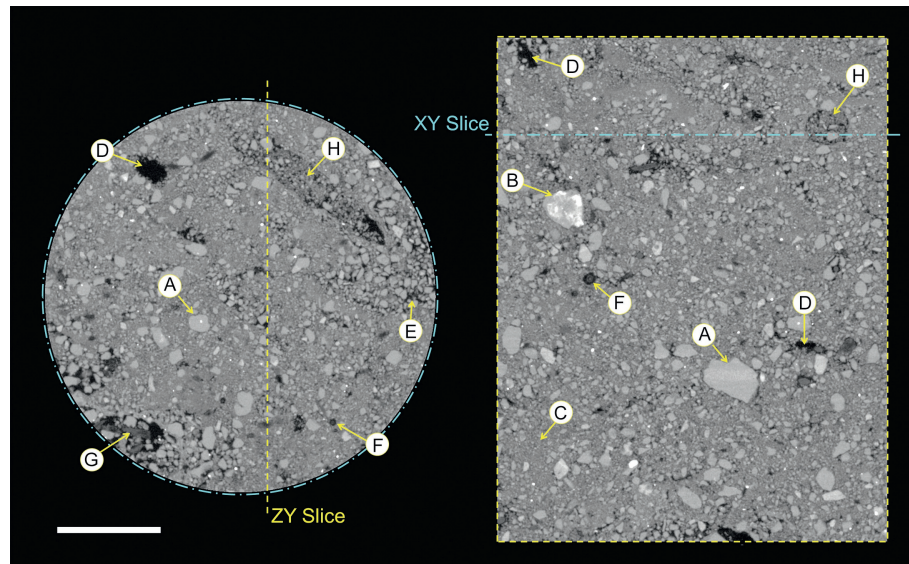


Fig. 3 Example X-ray CT slice images (XY slice (left) and ZY slice image (right)) showing soil structure of a loamy sand soil core collected from WITS Rural Facility, South Africa (40 mm diameter x 50 mm depth). Low image gray levels (dark) relate to lower density materials (e.g. air, organic matter) whereas higher image gray levels relate to higher density materials (primary particles). The annotations highlight some of the key features in the images (A) primary particles (sand grain), (B) primary particle with higher density inclusions, (C) bulk soil matrix, (D) air-filled biopore, (E) inter-particle porosity, (F) plant roots, (G) organic residues (H) in-filled biopore. Scale bar 12 mm. Original image courtesy Professor Kate Parr, University of Liverpool.

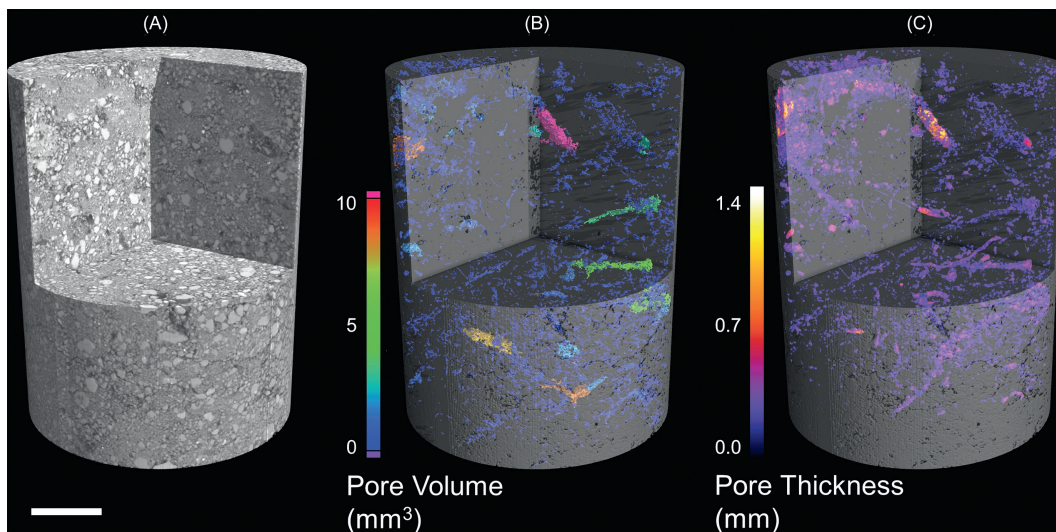


Fig. 4 Three-dimensional visualization and quantification of soil structure of a loamy sand soil core collected from WITS Rural Facility, South Africa (40 mm diameter x 50 mm depth) using X-ray CT. (A) Gray scale 3D rendered image of solid material, (B) quantification of soil pore volume calculated using VGStudioMAX defect analysis module (Volume Graphics GmbH, Germany) and (C) soil pore thickness calculated using *BoneJ* plugin for *Fiji* image analysis software. Scale bar 10 mm.

popular and offer advantages in fully automated segmentation of image data of soils and plant roots (Han et al., 2019; Soltaninejad et al., 2020). To extract the wealth of quantitative information about the tortuosity, connectivity, and topology of the features of interest, a number of 3D software tools and approaches are available. Commercial analysis packages such as *Avizo* (ThermoFisher Scientific, USA), *VGStudioMax* (Volume Graphics GmbH, Germany) or *Dragonfly* (Object Research Systems, Canada) offer versatile solutions for advanced processing of CT data, however their high cost can be a limitation for some users. Opensource free software such as *ImageJ/Fiji* (Schindelin et al., 2012), which has collections of community developed plugins specifically designed for analysis of 3D porous networks (e.g. *BoneJ* (Doube et al., 2010)), soils (e.g. *SoilJ* (Koestel, 2018)) and plant roots (e.g. *Rootline* (Phalempin et al., 2021)) provide useful alternatives to commercial packages. As the image data are digital, the information lends itself well to translation for modelling approaches, where structural models are required.

For further reading, Schlüter et al. (2014) provide an excellent in-depth review of imaging processing of multiphase CT images appropriate to soil processes.

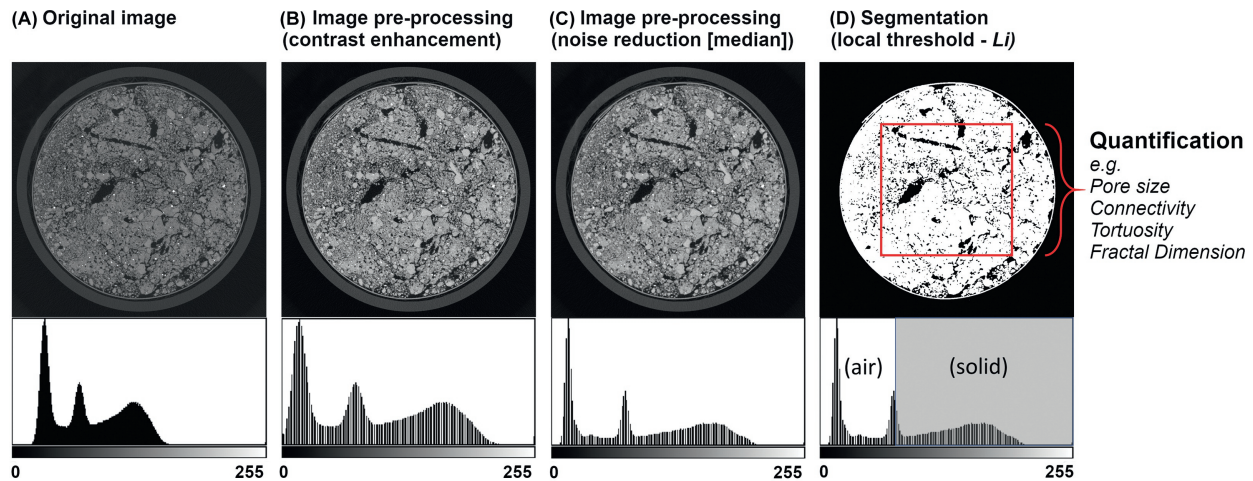


Fig. 5 Basic image processing steps to classify X-ray CT image pixels as air filled pores before geometrical quantification of soil pore structures in 3D.

(A) Original CT image of a sandy loam soil (B) Image pre-processing: contrast enhancement (note stretching of histogram) (C) Image pre-processing: noise reduction using a median filter (1.5 radius), (D) Segmentation of solid from pores by the classification of pixels based on their gray level to a binary image. The *Li* local threshold algorithm in FIJI Software was used in this example. Histograms of image gray levels are presented below each image.

Gamma-ray tomography

Another type of tomography used in the soil sciences is Gamma-ray tomography. It is similar in fundamental principles to X-ray CT, but as the name suggests, measures the attenuation of gamma-rays as they are transmitted through the sample of interest. Gamma-rays are emitted by the nucleus of certain radionuclides following radioactive decay. Commonly used sources are ^{241}Am , ^{133}Ba , ^{137}Cs , and ^{60}Co , with each radionuclide displaying different gamma-ray emission energy, activity, and half-life properties. Gamma-ray tomography has not been used as extensively as X-ray CT in soil science. Gamma-ray CT systems are often smaller, simpler, more portable, less expensive, have greater penetrating power and less scatter in comparison X-ray CT systems that typically require large electrical generators to power their X-ray tubes. However, the acquisition radiograph images are generally of lower quality due to limitations on the intensity of the gamma-rays from the radioactive source. Overall scan times can be therefore longer due to the need for longer detector exposure times. For further reading, [Pires et al. \(2011\)](#) provide an excellent review of the use of Gamma-ray tomography in soil science and [Johansen \(2015\)](#) give a detailed overview of the technology and some general applications.

Neutron tomography

Neutron tomography is a non-invasive imaging technique that uses neutrons to map the internal structure of a material of interest in 3D via their attenuation of neutrons. The acquisition principles are similar to those used for X-ray and Gamma-ray CT. However, as neutrons have a neutral charge, they generally have a very weak interaction with some electron dense materials. In comparison to X-rays, which are mainly attenuated by the electron shell of atoms, the attenuation of neutrons is dominated by interaction with the atomic nuclei, so lighter, nucleon dense elements such as hydrogen and boron have a much stronger interaction with neutrons compared to heavier elements such as silicon, aluminum or lead. Neutron imaging has several advantages as compared to other imaging techniques. These advantages include the ease of differentiation of roots as they contain water, making them distinct in neutron radiographs. Neutron tomography can also enable the tracking of water movement and uptake as it can differentiate between isotopes of hydrogen (protium and deuterium) which can be used to track water movement in real time imaging ([Zarebanadkouki et al., 2013](#); [Tötze et al., 2021](#)). One of the major disadvantages with neutron tomography is the limited penetrating power of neutrons, which limit the thickness of specimens that can be successfully imaged (limited approximately 20 mm in soils) ([Mawodza et al., 2020](#)). Exposing study specimens to neutrons also makes them radioactive soon after imaging, thus introducing handling hazard for up to a few hours. Another limitation that may reduce the widespread use of neutron imaging is the relative scarcity of imaging facilities. Most neutron imaging facilities are based at neutron research reactors or spallation sources such as Paul Scherrer Institut in Switzerland (www.psi.ch/en/num) or the ISIS Neutron and Muon Source at the Rutherford Appleton Laboratory in UK (www.isis.stfc.ac.uk).

Magnetic resonance imaging

Magnetic resonance imaging (MRI) is a valuable non-invasive tomographic technique used for the investigation of soil processes involving the distribution or movement of water. It can also be applied to image plant tissues, as they are predominately water

based. In comparison to the electromagnetic waves and high-energy particle transmission techniques described above, which measure the absorption of ionizing radiation by the sample, MRI is emission-based and uses the effect of nuclear magnetic resonance (NMR) of certain atomic nuclei (mainly hydrogen and therefore water) to generate 2- or 3-D images. MRI scanners use a strong magnetic field to align the spin (angular momentum) of the hydrogen nuclei (protons), creating weak magnetization of the sample. This magnetism can be manipulated with application of time-dependent radio frequency pulses. A detector coil around the sample is used to measure the change in orientation of the rotational axis of the spinning nuclei (called precession) through induction of a weak voltage. Fourier analysis is used to extract spatial information on the distribution and quantity of protons from the detector coil signal. For an accessible introduction to the principles of NMR we refer the reader to Scheenen et al. (2000) and to Callaghan (1993) who provide a thorough description of the theory of the technique.

The application of MRI to characterize moisture content in soils was first demonstrated in the 1980s, both in the field and the laboratory, identifying its potential as for studies of preferential flow pathways. The use of tracers with fluorinated compounds or paramagnetic ions like Cu^{2+} , Ni^{2+} or Mn^{2+} can be useful to improve the contrast of water. Fig. 6A shows an example of the temporal mapping of water flow in a soil core using MRI (Haber-Pohlmeier et al., 2021).

Soils with strong magnetic properties (ferromagnetic particles) can present a challenge to MRI as they suppress the signal of the soil water. However, Pflugfelder et al. (2017) screened six different agricultural soils with a range of textures for use in a MRI study to image barley roots. Only one of the soils proved unsuitable for use with MRI, highlighting the potential of the technology for studying plant root growth (see Fig. 6B for an example of root imaging) in a range of natural soil types (van Dusschoten et al., 2016). Despite this, there are still relatively few studies using MRI to look at root-soil interactions, especially compared to X-ray CT. The method gives detailed information on root traits such as mass, length, diameter, tip number, branching angles, spatial distribution, and growth rate. Detection of roots with diameters down to 200 μm is possible and the technique also allows simultaneous measurement of moisture content in the soil surrounding roots. Recently, low magnetic field MRI has been developed to further

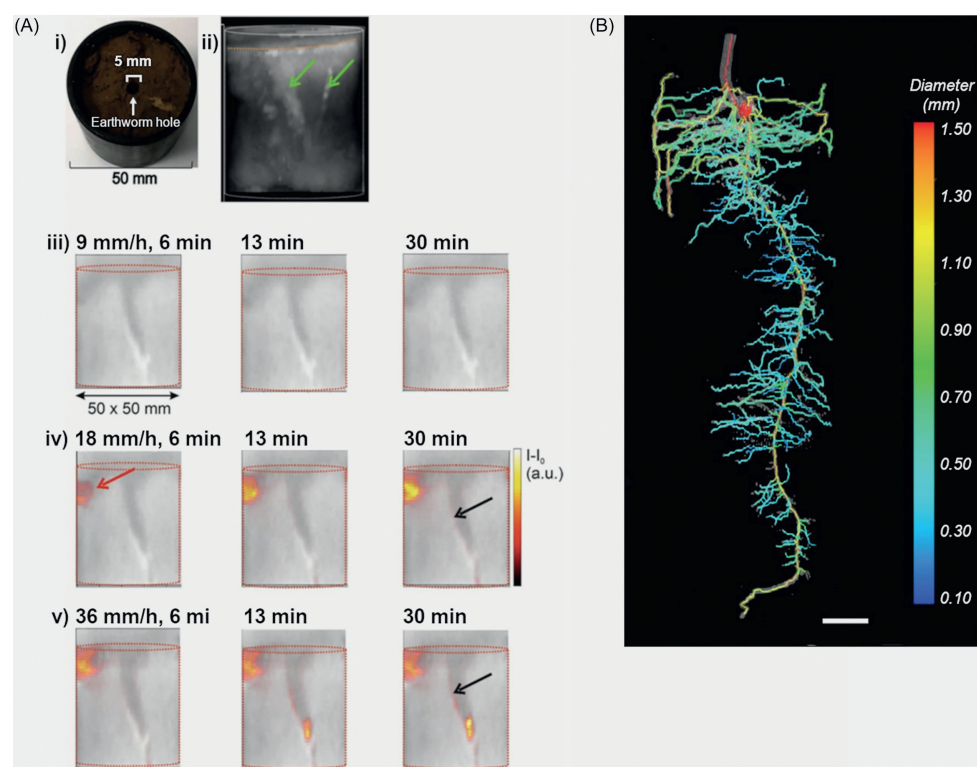


Fig. 6 Examples of MRI imaging of soil and plant processes. (A) Imaging flow in a natural soil core during irrigation. (i) Setup. (ii) ZTE image of the natural core (maximum intensity projection) with no irrigation. Matrix size 128^3 , FOV: $70 \times 70 \times 90$ mm. The dotted orange line indicates the soil surface, the white arrows point to the large wormhole and a smaller macropore, probably caused by a degraded root. (iii–v) Central vertical slice with different irrigation rates (top to bottom) and time-points after start (left to right). Shown are difference images at the given time-points minus $t = 0$. The red arrow indicates ponding water near the left edge. The black arrows indicate film flow at the walls of the large wormhole. (B) Magnetic resonance image of a maize root system (15 DAS) co-registered with a diameter estimation of each root segment based on the local MRI intensity and drawn in false colors. Scale bar 20 mm. (A) Figure modified from Haber-Pohlmeier S, Caterina D, Blümich B, and Pohlmeier A (2021) Magnetic resonance imaging of water content and flow processes in natural soils by pulse sequences with ultrashort detection. *Molecules* 26: 5130. under creative commons license (<http://creativecommons.org/licenses/by/4.0/>). (B) Figure modified from van Dusschoten D, Metzner R, Kochs J, Postma JA, Pflugfelder D, Bühler J, Schurr U, and Jahnke S (2016) Quantitative 3D analysis of plant roots growing in soil using magnetic resonance imaging. *Plant Physiology*, 170, 1176–1188. under creative commons license (<http://creativecommons.org/licenses/by/4.0/>).

mitigate the effects of paramagnetic materials in soil on image quality further extending the utility and wider use of technique for the study of root-soil interactions (Bagnall et al., 2020).

Positron emission tomography

Although less widely used compared to CT, Positron Emission Tomography (PET) allows unique insights at high sensitivity of dynamic processes in soils. The technique involves the introduction of a positron-emitting nuclide tracer to the soil or plant system and then maps its spatial and temporal distribution. Tracer location is detected by the emission of paired high-energy photons (as gamma-rays) which are produced following an annihilation event due to the interaction of a positron from the tracer and an electron of the sample material. The emitted gamma rays are detected by arrays of scintillator crystal detector blocks arranged in a ring or stack of rings in the scanner. Although the spatial resolution of most PET scanners is in the 1–5 mm range, due to the physics of the emission, and the detection of the coincident photons, PET imaging has very high sensitivity and accurate estimation of radiotracer concentration. Several positron-emitting isotopes can be created which have potential applications in studies of soil and plant processes. The most commonly used are ^{11}C , ^{13}N , ^{15}O , ^{18}F and ^{64}Cu . For example, both ^{18}F labelled fluoride (F^-) and fluorodeoxyglucose (FDG) have been used to image water transport in soils and sediments, however, each of these compounds have different physiochemical properties compared to water so caution must be taken on interpretation the results. Solute transport (Richter and Grundig, 2000) and herbicide migration (Kulenkampff et al., 2018) in soils has also been successfully visualized using ^{64}Cu labelled compounds. PET imaging has also allowed in situ quantification of biological activity within soils. Kinsella et al. (2012) used ^{18}F labelled FDG to characterize the metabolic activity of a subsurface strain of *Rahnella* (sp. Y9602) that has been shown to promote uranium sequestration in groundwater and soils. The tracer is metabolized by the bacteria and can be spatially resolved in the experimental system, which consisted of a column of fine sand with two bands of slit loam soil. The results showed a higher activity in the finer textured silt loam (see Fig. 7A). PET has also showed promise for quantification of microbial activity in soil sediments and crusts. In plant science, PET has been used extensively to image dynamics of macronutrients, trace elements, and signaling compounds (see Fig. 7B). Furthermore, the combination of PET with X-ray CT (Garbout et al., 2012) and MRI

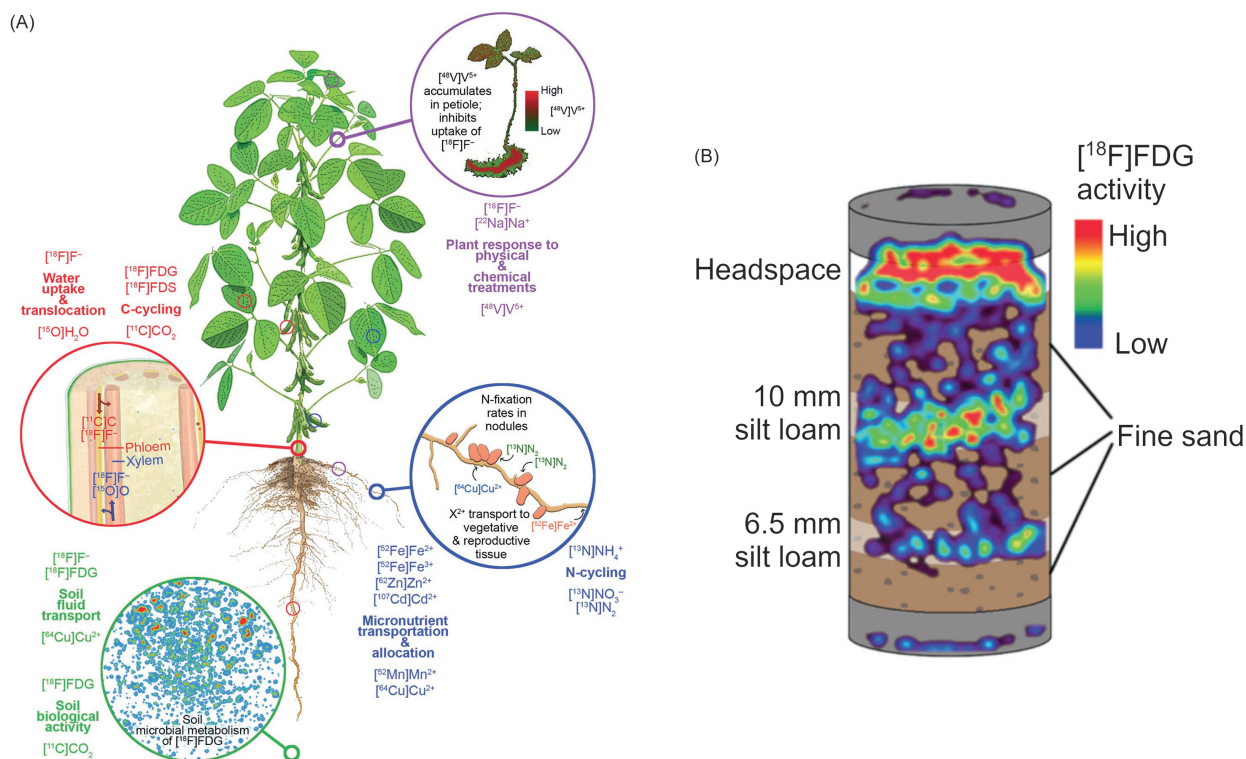


Fig. 7 Examples of Positron Emission Tomography imaging used to study soil processes. (A) An overview of positron imaging in plants. Examples of key areas of research are outlined in the inserts. (B) Example of Positron emission tomography (PET) imaging of *Rahnella* sp. Y9602 activity using $^{18}\text{F}\text{FDG}$ applied to silt loam layers interspersed with sterile fine sand layers in a synthetic soil column. (A) From Schmidt MP, Mamet SD, Ferrieri RA, Peak D, and Siciliano SD (2020) From the outside in: An overview of positron imaging of plant and soil processes. *Molecular Imaging* 19: 1536012120966405. under creative commons license (<http://creativecommons.org/licenses/by/4.0/>). (B) Figure from Schmidt MP, Mamet SD, Ferrieri RA, Peak D, and Siciliano SD (2020) From the outside in: An overview of positron imaging of plant and soil processes. *Molecular Imaging* 19: 1536012120966405. [Originally modified from Kinsella K, Schlyer DJ, Fowler JS, Martinez RJ, and Sobecky PA (2012) Evaluation of positron emission tomography as a method to visualize subsurface microbial processes. *Journal of Hazardous Materials* 213: 498–501] under creative commons license (<http://creativecommons.org/licenses/by/4.0/>).

(Jahnke et al., 2009) based imaging, information of the porous structure of soil can be generated in combination with root structure and the movement of the radio tracers in the plant, opening exciting new opportunities to aid understanding of the rhizosphere. For further reading on the use of PET imaging in the soil and plant sciences see (Schmidt et al., 2020).

Applications

Both X-ray and Gamma-ray CT have been used to investigate; fluid transport in relation to the physical structure of a soil (e.g. porosity, pore size distribution, surface area); generate data for simulation of fluid flow into a soil and other porous media and to measure the impact of pore microstructure on permeability in addition to measurement of soil quality indicators bulk density and total porosity. By far the most widely applied use of CT is in the 3D quantification of soil structure, where soil scientists have investigated the effects of land and crop management, impact of plant roots, microbes, macroinvertebrates on structural genesis (e.g. see Helliwell et al. (2013) for further examples). X-ray CT has also been used extensively for the quantification of the 3D structure of plant roots and their development over time (Mooney et al., 2012). Although, X-ray CT offers the potential of imaging plant root structures at higher spatial resolutions (e.g. $<5\ \mu\text{m}$ resolution imaging of root hairs using synchrotron based X-ray CT (Koebernick et al., 2017)), differentiation of roots in the soil is still a major challenge due to common X-ray attenuation values with other material in the soil such as water filled pore space and organic based residues. The use of Neutron, Magnetic Resonance and Positron Emission imaging techniques offer alternative approaches with greater sensitivity for plant root imaging although the imaging resolution is coarser.

Summary and outlook

This chapter has outlined the use of three of the main non-invasive imaging techniques currently available to investigate soil processes in 3D: Computed Tomography, Magnetic Resonance Imaging and Positron Emission Tomography. Although there have been significant advances in the development of each of these techniques over the last 30 years, there are still a number of gaps in understanding of soil processes that need to be addressed. In terms of CT, the wider availability of industrial CT systems, due to their lower cost and easier intuitive operational protocols, bring the technology to a greater number of soil scientist where equipment may become a standard tool in the soil science laboratory. Advances in X-ray detectors have facilitated improved image quality and ever faster acquisition times, giving the soil scientist the ability to increase sample replication numbers to give greater confidence in capturing the heterogeneity displayed by 3D pore networks. Advanced imaging capabilities at national synchrotron light facilities also offer combined use of different modalities. For example, X-ray CT and time-of-flight neutron imaging was recently shown to offer greater discrimination and localization of structural phases in soils (soil minerals and soil organic matter) than if the techniques were used independently (Koestel et al., 2022). Furthermore, emerging new methodologies that combine structural and chemical imaging also offer exciting new avenues to a deeper understanding of soil systems. These correlative imaging methods use image registration techniques to combine 3D tomographic data sets with biochemical microscopic data of different modalities and scales (e.g. light, fluorescence and electron microscopy and secondary ion mass spectrometry) (Schlüter et al., 2019). This unique capability to simultaneously quantify 3D heterogenous habitat space in soil and map microbial communities and organic matter distributions are set to revolutionize our understanding of biogeochemical and physical processes in natural soil systems.

Appendix: Supplementary material

Supplementary material related to this chapter can be found on the accompanying CD or online at <https://doi.org/10.1016/B978-0-12-822974-3.00151-8>.

References

- Bagnall GC, Koonjoo N, Altobelli SA, Conradi MS, Fukushima E, Kuethe DO, Mullett JE, Neely H, Rooney WL, Stupic KF, Weers B, Zhu B, Rosen MS, and Morgan CLS (2020) Low-field magnetic resonance imaging of roots in intact clayey and silty soils. *Geoderma* 370: 114356.
- Baveye PC, Laba M, Otten W, Bouckaert L, Dello Sterpaio P, Goswami RR, Grinev D, Houston A, Hu Y, Liu J, Mooney S, Pajor R, Sleutel S, Tarquis A, Wang W, Wei Q, and Sezgin M (2010) Observer-dependent variability of the thresholding step in the quantitative analysis of soil images and X-ray microtomography data. *Geoderma* 157: 51–63.
- Burke IT, Mosselmans JFW, Shaw S, Peacock CL, Benning LG, and Coker VS (2015) Impact of the Diamond Light Source on research in Earth and environmental sciences: Current work and future perspectives. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 373: 20130151.
- Callaghan PT (1993) *Principles of Nuclear Magnetic Resonance Microscopy*. Oxford University Press on Demand.
- Cimpoeaşu MO, Kuras O, Pridmore T, and Mooney SJ (2020) Potential of geoelectrical methods to monitor root zone processes and structure: A review. *Geoderma* 365: 114232.
- Doube M, Kłosowski MM, Arganda-Carreras I, Cordelières FP, Dougherty RP, Jackson JS, Schmid B, Hutchinson JR, and Shefelbine SJ (2010) BoneJ: Free and extensible bone image analysis in ImageJ. *Bone* 47: 1076–1079.
- Garbout A, Munkholm LJ, Hansen SB, Petersen BM, Munk OL, and Pajor R (2012) The use of PET/CT scanning technique for 3D visualization and quantification of real-time soil/plant interactions. *Plant and Soil* 352: 113–127.

- Haber-Pohlmeier S, Caterina D, Blümich B, and Pohlmeier A (2021) Magnetic resonance imaging of water content and flow processes in natural soils by pulse sequences with ultrashort detection. *Molecules* 26: 5130.
- Han Q, Zhao Y, Liu L, Chen Y, and Zhao Y (2019) A simplified convolutional network for soil pore identification based on computed tomography imagery. *Soil Science Society of America Journal* 83: 1309–1318.
- Helliwell JR, Sturrock CJ, Grayling KM, Tracy SR, Flavel RJ, Young IM, Whalley WR, and Mooney SJ (2013) Applications of X-ray computed tomography for examining biophysical interactions and structural development in soil systems: a review. *European Journal of Soil Science* 64: 279–297.
- Jahnke S, Menzel MI, van Dusschoten D, Roeb GW, Buhler J, Minwuyelet S, Blumler P, Temperton VM, Hombach T, Streun M, Beer S, Khodaverdi M, Ziemons K, Coenen HH, and Schurr U (2009) Combined MRI-PET dissects dynamic changes in plant structures and functions. *Plant Journal* 59: 634–644.
- Johansen GA (2015) 7 - Gamma-ray tomography. In: Wang M (ed.) *Industrial Tomography*. Woodhead Publishing.
- Kinsella K, Schlyer DJ, Fowler JS, Martinez RJ, and Sobecky PA (2012) Evaluation of positron emission tomography as a method to visualize subsurface microbial processes. *Journal of Hazardous Materials* 213: 498–501.
- Koebnick N, Daly KR, Keyes SD, George TS, Brown LK, Raffan A, Cooper LJ, Naveed M, Bengough AG, Sinclair I, Hallett PD, and Roose T (2017) High-resolution synchrotron imaging shows that root hairs influence rhizosphere soil structure formation. *New Phytologist* 216: 124–135.
- Koestel J (2018) SoilJ: An ImageJ plugin for the semiautomatic processing of three-dimensional X-ray images of soils. *Vadose Zone Journal* 17: 170062.
- Koestel J, Fukumasu J, Larsbo M, Herrmann AM, Ariyathilaka P, Magdysyuk OV, and Burca G (2022) Potential of combined neutron and X-ray imaging to quantify local carbon contents in soil. *European Journal of Soil Science* 73: e13178.
- Kulenkampff J, Stoll M, Grundig M, Mansel A, Lippmann-Pipke J, and Kersten M (2018) Time-lapse 3D imaging by positron emission tomography of Cu mobilized in a soil column by the herbicide MCPA. *Scientific Reports* 8.
- Loke MH, Chambers JE, Rucker DF, Kuras O, and Wilkinson PB (2013) Recent developments in the direct-current geoelectrical imaging method. *Journal of Applied Geophysics* 95: 135–156.
- Lombi E and Susini J (2009) Synchrotron-based techniques for plant and soil science: opportunities, challenges and future perspectives. *Plant and Soil* 320: 1–35.
- Mawodza T, Burca G, Casson S, and Menon M (2020) Wheat root system architecture and soil moisture distribution in an aggregated soil using neutron computed tomography. *Geoderma* 359: 113988.
- Mooney SJ, Pridmore TP, Helliwell J, and Bennett MJ (2012) Developing X-ray Computed Tomography to non-invasively image 3-D root systems architecture in soil. *Plant and Soil* 352: 1–22.
- Pflugfelder D, Metzner R, van Dusschoten D, Reichel R, Jahnke S, and Koller R (2017) Non-invasive imaging of plant roots in different soils using magnetic resonance imaging (MRI). *Plant Methods* 13: 102.
- Phalempin M, Lippold E, Vetterlein D, and Schlüter S (2021) An improved method for the segmentation of roots from X-ray computed tomography 3D images: Rootline v.2. *Plant Methods* 17: 39.
- Pires L, Cassaro F, Bacchi O, and Reichardt K (2011) *Gamma-Ray Computed Tomography in Soil Science: Some Applications*.
- Richter M and Grundig M (2000) Positron emission tomography for studies of water flow in soil columns. In: *6th International Congress on Applied Mineralogy (ICAM 2000)*, Jul 17-19 2000 Gottingen, Germany 653–655.
- Scheenen TWJ, van Dusschoten D, de Jager PA, and van As H (2000) Quantification of water transport in plants with NMR imaging. *Journal of Experimental Botany* 51: 1751–1759.
- Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, Tinevez J-Y, White DJ, Hartenstein V, Eliceiri K, Tomancak P, and Cardona A (2012) Fiji: an open-source platform for biological-image analysis. *Nature Methods* 9: 676–682.
- Schlüter S, Sheppard A, Brown K, and Wildenschild D (2014) Image processing of multiphase images obtained via X-ray microtomography: A review. *Water Resources Research* 50: 3615–3639.
- Schlüter S, Eickhorst T, and Mueller CW (2019) Correlative imaging reveals holistic view of soil microenvironments. *Environmental Science & Technology* 53: 829–837.
- Schmidt MP, Mamet SD, Ferrieri RA, Peak D, and Siciliano SD (2020) From the outside in: An overview of positron imaging of plant and soil processes. *Molecular Imaging* 19: 1536012120966405.
- Soltaninejad M, Sturrock CJ, Griffiths M, Pridmore TP, and Pound MP (2020) Three dimensional root CT segmentation using multi-resolution encoder-decoder networks. *IEEE Transactions on Image Processing* 29: 6667–6679.
- Tötze C, Kardjilov N, Hilger A, Rudolph-Mohr N, Manke I, and Oswald SE (2021) Three-dimensional in vivo analysis of water uptake and translocation in maize roots by fast neutron tomography. *Scientific Reports* 11: 10578.
- van Dusschoten D, Metzner R, Kochs J, Postma JA, Pflugfelder D, Bühler J, Schurr U, and Jahnke S (2016) Quantitative 3D analysis of plant roots growing in soil using magnetic resonance imaging. *Plant Physiology* 170: 1176–1188.
- Young IM, Crawford JW, and Rappoldt C (2001) New methods and models for characterising structural heterogeneity of soil. *Soil and Tillage Research* 61: 33–45.
- Zarebanadkouki M, Kim YX, and Carminati A (2013) Where do roots take up water? Neutron radiography of water flow into the roots of transpiring plants growing in soil. *The New Phytologist* 199: 1034–1044.

Soil processes in the hydrologic cycle

Cosimo Brogi^a, Harry Vereecken^a, Heye Reemt Bogena^a, and Luca Brocca^b, ^aAgrosphere Institute (IBG-3), Forschungszentrum Jülich GmbH, Jülich, Germany; ^bResearch Institute for Geo-Hydrological Protection, National Research Council, Perugia, Italy

© 2023 Elsevier Ltd. All rights reserved.

Introduction	469
Soil water in the global water cycle	470
Key regulatory features of water in soil	470
Energy related processes	470
Hydrologic processes	471
Biogeochemical processes	472
Soil water in the context of watershed management	472
Impact of soil water on ecosystem services	473
Representing and monitoring water in soils	474
Mass balance and energy transfer methods	474
Measurement techniques for soil water content sensing	475
Soil water and the issue of scales	476
Examples of the combination of soil water measurements and modelling	476
The value of soil water content information	476
The importance of soil water content and its spatial variability in respect to crop productivity and carbon fluxes	478
On the use of soil water content data at different scales	478
Summary and outlook	480
References	480

Abstract

Water in soils is only 0.0012% of Earth's water and yet it is paramount in regional scale processes, in the climate system, and for life on Earth. The storage and retention of water in soils affects energy partitioning processes at the land surface, plants transpiration, hydrological processes, and water management practices. Finally, it controls soil's ability to act as a sink or source for greenhouse gases with an indirect yet strong control on the global climate system.

Key points

- Water in soils is a small yet key fraction of the global Earth's water.
- Water in soils strongly affects hydrological, energy related, and biogeochemical processes.
- Water in soils has a large impact on human activities and on ecosystem services.
- The monitoring of soil water content provides large benefits and multiple techniques are available, each with positive and negative aspects.

Introduction

Water stored in soils is only 0.0012% of Earth's hydrosphere. Yet, the storage and retention of water in soils, its quality, and its high temporal and spatial variability is paramount in the climate system and for life on Earth in general. Such water affects regional scale processes such as the energy partitioning between latent and sensible heat fluxes at the Earth's surface, and thus air temperature, and sometimes precipitation, drought, and heat waves. The lack of soil water also constrains the ability of plants to perform photosynthesis by limiting transpiration processes and thus reducing biomass production. It directly impacts major hydrological processes such as infiltration, runoff, and groundwater recharge, as well as water management practices, such as irrigation and drainage. Finally, at longer time scales, water controls the ability of soil to act as a sink for greenhouse gases (e.g., carbon sequestration) or source (soil respiration, N emissions) thus applying an indirect yet strong control on the global climate system.

In this chapter, we first introduce water in soils within the global water cycle and then discuss regulatory features of water in soils from energy related processes to hydrology and biogeochemistry. A discussion on soil water in the context of watershed management and its impacts on ecosystem services follows. Methods for monitoring soil water content are introduced with a discussion on the issue of the spatial scale of water content in soil. Finally, several examples are described in which soil water content sensing and monitoring were combined with modelling environments.

Soil water in the global water cycle

The global amount of freshwater is generally constant and represents 3% of Earth's hydrosphere with the remaining being saline water that is mostly stored in oceans and seas, which occupy 70% of Earth's surface, or as groundwater (Aeschbach-Hertig and Gleeson, 2012; Bogen et al., 2004; Dorigo et al., 2021). However, this small percentage of freshwater is the only portion that can potentially be utilized by most terrestrial animals and by the vast majority of plants. Water contained in the unsaturated zone, which is itself a small portion of Earth's freshwater, is generally defined as water in soils.

Freshwater stored in soil as soil moisture is one of the storage pools of the water cycle, which is a conceptual model that describes and separates the processes and pathways through which the ~ 1.4 billion m^3 of Earth's water is transferred and stored. Freshwater can be found in glaciers and icecaps in its solid state, in clouds and air humidity in its gaseous state, and in lakes, streams, rivers, soils and groundwater reservoirs in its liquid state. A variety of processes such as precipitation, evaporation, transpiration, condensation, infiltration, runoff, melting, and groundwater flow can transfer freshwater from one of these storage pools to another. Water in soils represents about 0.05% of the total stock of freshwater (0.0012% of the hydrosphere) and is thus one of the smaller pools. However, especially when compared to groundwater, soil water contributes to or controls large land surface fluxes (Fig. 1). Thus, it strongly impacts terrestrial energy cycles through its influence on the energy partitioning at the surface and affects hydrological processes that are coupled to energy cycles through evapotranspiration (Bogen et al., 2004; Seneviratne et al., 2010). Finally, soil water is linked to the biogeochemical cycles of carbon and nutrients through the coupling between plant photosynthetic activity and transpiration.

Key regulatory features of water in soil

Energy related processes

Every year, approximately 5500 TW of energy, mostly from solar radiation, are used to return to the atmosphere $\sim 70,000 \text{ km}^3$ of water that is stored in Earth's soils. It is especially the water stored in the topsoil that strongly affects the energy fluxes at the land surface and thus the related meteorological and atmospheric inputs and vegetation states. Here, the net radiation that reaches the Earth's surface must be balanced by (a) fluxes of latent and sensible heat directed towards the atmosphere, (b) ground heat fluxes

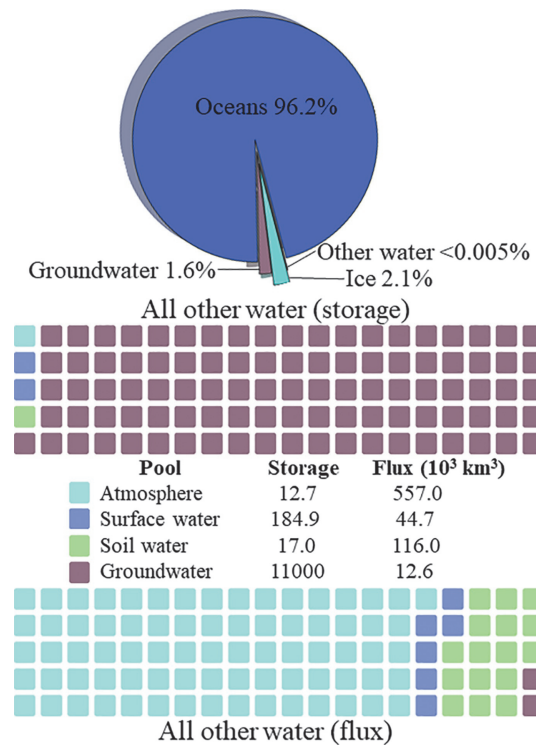


Fig. 1 The global water storage pools of oceans, groundwater, ice, and other water. Of the other water pools, atmosphere, surface water, and soil water are shown together with groundwater in terms of storage and fluxes, both with values (in 10^3 km^3) and with graphical representation of their magnitude. As indicated with text citations, data from: Aeschbach-Hertig W and Gleeson T (2012) Regional strategies for the accelerating global problem of groundwater depletion. *Nature Geoscience* 5(12): 853–861; Bogen H, Hake JF and Vereecken H (2004) *Water and Sustainable Development*. Forschungszentrum Jülich; Dorigo W, Dietrich S, Aires F, Brocca L, Carter S, Cretaux JF, Dunkerley D, Enomoto H, Forsberg R and Güntner A (2021) Closing the water cycle from observations across scales: Where do we stand? *Bulletin of the American Meteorological Society* 102(10): E1897–E1935.

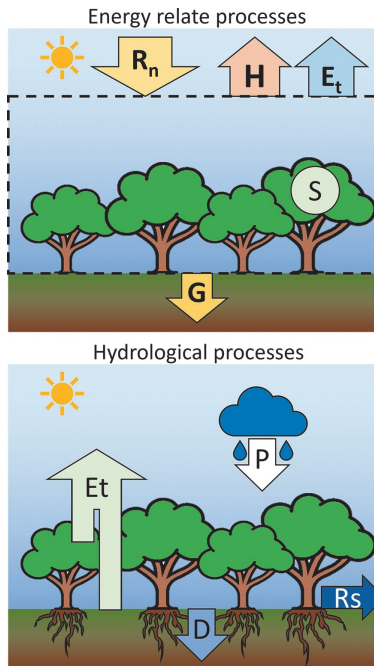


Fig. 2 The upper panel shows the energy related fluxes where R_n is net radiation, H is sensible heat, E_t is latent heat, G is ground heat, and S is heat stored in vegetation. The lower panel shows the hydrological processes where P is precipitation, E_t is evapotranspiration, D is infiltration, and R_s is surface runoff.

directed towards the subsurface, and (c) heat stored in the vegetation (Fig. 2). Variations in soil water content can modulate feedbacks between the land surface and the atmosphere by changing the partitioning between latent heat, sensible heat, and ground heat fluxes. Soil water availability is a limiting factor for evapotranspiration and thus latent heat, which also tightly couples the energy related and the hydrological processes (Jung et al., 2010). For example, when soil dries out, sensible heat increases, thus reducing latent heat because actual evapotranspiration falls below potential evapotranspiration. Also, the soil heat transport, and thus ground heat flux, is influenced by water movement in soil as this can redistribute heat vertically. Variations in soil water content also affect the turbulent and highly rotational air flows called eddies, the currents that transport water vapor, heat, and gaseous substances across the Earth's surface and in between surface and atmosphere. In parallel, plants covering the soil can alter heat fluxes, but such cover is strictly related to soil water availability as it will be discussed later.

Hydrologic processes

When rainfall reaches the land surface, it has a chance to infiltrate the soil. However, the infiltration rate will depend on multiple factors (Fig. 2). First, vegetation can intercept rainfall before it reaches the soil and part of this water will evaporate before doing so. The amount of interception depends on plant structure, branching, leaf geometry, and rainfall intensity, all aspects that vary considerably in time and space. When rainfall reaches the soil surface, the amount and rate at which it can infiltrate depends on rainfall intensity, vegetation type, degree of surface sealing and soil management practices (e.g., frequency and type of tillage for agricultural land). Furthermore, local factors such as the soil's hydraulic properties, initial soil water content, and the roughness of the soil surface as well as non-local terrain properties such as slope, aspect, and landscape type and location play an important role. All these vary greatly in space as different soils can be found in different areas and at different depths. For example, higher infiltration rates can be the consequence of soil heterogeneity and, especially in forest soils, of the presence of biopores and macropores. Moreover, antecedent soil saturation status can favor or reduce infiltration. A large soil water content, on one hand, can increase hydraulic conductivity, since a greater portion of pores are transmitting water, whereas on the other hand it can decrease infiltration due to the reduced suction gradient (Schumann, 1998). When precipitation rate exceeds soil infiltration capacity, water ponds on the soil surface. The ponded water can either infiltrate at a later stage, return to the atmosphere by evaporation, or generate surface runoff. Depending on the volume and energy of rainfall and runoff, erosion and transport of solutes and solids can occur. This runoff water will move towards lower elevations and, if no further infiltration or evaporation occurs, reach freshwater bodies through streams and river networks, and be eventually discharged into seas or oceans.

The water stored in soil can leave the soil profile either by the process of evapotranspiration or by drainage and leaching to groundwater or to surface waters. Evapotranspiration returns to the atmosphere about 63% of the total precipitation that reaches the land surface and consists of two major components. The first component is soil evaporation in which soil water evaporates from the soil surface. This occurs when the water vapor pressure at the evaporating surface is higher than that of the atmosphere. The water returned to the atmosphere through evaporation is purified from residuals or pollutants that remain in the soil. The second

component is the process of plant transpiration, whereby the water stored in the root zone is taken up by plants through their roots and released back into the atmosphere via the stomata in the leaves. When water availability and root extension are not a limiting factor, transpiration tends to be controlled by the available solar radiation and can be increased with the presence of wind. However, water limitation is generally a constraint to transpiration rates, especially in agriculture. In addition, plants generally cannot access all the water that is present in the rootzone, and their uptake abilities are strongly influenced by plant and soil characteristics. A paramount factor for transpiration is thus soil matric potential which represents the force that retains water in soils. In the unsaturated zone, the soil matric potential is negative and is mostly controlled by capillary forces. Plants can take up water when the soil matric potential is between the soil matric potential at field capacity and wilting point. These two values vary between vegetation types and plants varieties. When the amount of water in soil is such that the soil matric potential is larger than the soil matric potential at field capacity, soil cannot hold this water against gravity and water drains downwards and eventually reaches the groundwater.

Biogeochemical processes

Soil water affects many biogeochemical processes in soils that are related to important matter cycles such as carbon, nitrogen, phosphate, and other nutrients. It controls the reaction rates of processes, such as decomposition of organic matter and the related mineralization of organically bound N, the formation of greenhouse gases, nitrification, denitrification as well as sorption, desorption, and chemical binding of inorganic and organic compounds.

The reservoir of carbon in soils, stored both in organic and inorganic forms, is larger than that of atmosphere and terrestrial vegetation combined and is largely determined by the soil water content. Boreal soils and waterlogged peat soils in particular have great storage potential because low temperatures, oxygen levels, and pH values inhibit the decomposition of carbon in the soil. Although peatlands make only 3% of the terrestrial land surface, they store about one fifth of the soil carbon stocks. In tropical forests, high soil water content can lower the production and release into the atmosphere of carbon dioxide from the soil. In these environments, soil respiration is strongly affected by short-term soil water content dynamics and by seasonal patterns in temperature, which has a positive control on soil respiration if soil water content is not a limiting factor (Wood et al., 2013). Changes in soil carbon stock depend on input rates relative to losses and are generally slow processes. Carbon turnover is a relatively rapid process in the topsoil whereas deeper in the soil profile carbon residence times are longer. Most of the carbon inputs are plant based and as such are directly or indirectly dependent on soil water such as above- and below ground litter, root exudates, and C dissolved in the subsoil. The same applies to carbon losses through mineralization, leaching, and carbonate weathering. Furthermore, carbon can be lost via erosion and gained through deposition, which are again mediated by overland water flows.

Similar considerations can be made for nutrients such as nitrogen, phosphorus, potassium, and trace elements that have pivotal role in the support of plant growth and are a key limiting factor in most soil-vegetation ecosystems. For most of these nutrients, plant uptake is a relatively inefficient process and losses, especially agricultural products removed from the system, are difficult to balance with inputs. This is the reason why agricultural productivity is often boosted by fertilizer application practices. Water transports nutrients out of the rooting zone by either releasing them into the atmosphere via gas effluxes (e.g., NH_4 , N_2O , and N_2) or by transport through erosion, leaching, and subsurface flow that directs such nutrients into the groundwater or surface water bodies. Finally, soil water content regimes can affect soil biota which are either soil animals or micro-organisms, such as fungi, algae, and bacteria. Microbial activity is generally enhanced by high soil water availability, especially after precipitations, and bacteria in particular are influenced as their diversity is promoted by intermediate moisture conditions (Bickel and Or, 2020; Huang et al., 2021).

Water dynamics in soil, mostly due to soil heterogeneity, control the storage and transformation of anthropogenic substances that can be temporary restrained from reaching the atmosphere, surface water, or groundwater. Pollutants may originate from point-scale sources, such as industry and landfills, or from non-point sources, such as agricultural land. The recycling of such polluting materials and substances contributes to the protection of human health and is strictly dependent on the regulation of the water flow in soils where such elements are found. Thus, specific covers or barriers can be designed to limit the water flow through a waste-laden zone, with an interesting example being the artificial freezing of soil. Such measures generally provide a stable physical and chemical environment with limited releases to the surrounding biosphere though dilution, dispersion, retardation, and decay.

Soil water in the context of watershed management

Water in soils, together with other morphological and climatic aspects, such as temperature and the occurrence of seasonal soil freezing, strongly affects infiltration, groundwater recharge, baseflow, runoff, and erosion. Such variables are part of watershed management practices, which aim at managing natural resources and human needs. For example, soil water becomes key when there is a need to mitigate the adverse impacts of extremely wet or dry conditions. The amount of water that is already present in soils when precipitations occur, combined with other local and non-local variables, can lead to extreme runoff and flooding which affect human activities and natural environments. Flooding can cause soil erosion and redeposition, thus removing fertile topsoil layers on one hand and reducing water quality downstream on the other. It can also affect vegetation and reduce its contribution to interception, evapotranspiration, and seasonal patterns in water use that affect runoff and infiltration. In the prolonged absence of precipitation, drought is defined by the low availability of water in soils. Low soil water content can contribute to overgrazing or to

the spread of wildfire, thus exposing soils to erosion, especially when precipitation is renewed. On one hand, such erosion contributes to the mobilization of large stocks of organic carbon and thus to atmospheric carbon dioxide emissions. On the other hand, it diminishes soil water storage capacity and is responsible for the loss of agricultural land leading to enhanced risk of food insecurity.

Various human activities (e.g., agriculture, industry, and tourism) critically rely on the support of soil water and of freshwater in general. This has led to a conceptual subdivision between two kinds of freshwater resources, referred to as blue and green water (Falkenmark and Rockström, 2006). Blue water is stored in surface and groundwater bodies, which can be accessed to support human activities, whereas green water originates from precipitation and contributes to soil water. Green water can be either productive when it contributes to transpiration or non-productive (i.e., soil evaporation and interception). Strong interactions occur between these two kinds of water, and green water has a pivotal role in the production of biomass for human use and consumption (Fig. 3). Such importance can be exacerbated by the geographical and climatic characteristics of a given region, by human intervention, and by future climate change scenarios. When green water is not sufficient, blue water can be harvested from natural or artificial water bodies and reservoirs or can be moved from the groundwater to the surface by water-well extraction. Such water is then returned to the soil, for example in the case of irrigation. Here, an accurate and timely characterization of soil water dynamics in the root zone can enable more efficient and thus more sustainable irrigation practices. Nonetheless, only part of irrigated water will be used by crops while the remaining quantity will contribute to further evaporation, transport of solutes in the groundwater, and runoff of poor water quality. Consequently, irrigation strongly impacts soil water and soil quality. One example is soil salinization which can take place not only in irrigated land but also in areas affected by saltwater intrusion to the groundwater due to the exploitation of coastal aquifers. Another impact of irrigation practices is the increased transport of substances due to deep percolation and leaching of excess nutrients. Nonetheless, irrigation is a key solution to insufficient availability of green water and to the consequent burden for local food production, which is why some countries use more than 35% of their renewable water resources for irrigation (Faurès et al., 2002). These figures are expected to increase in the upcoming years. As such, it is paramount to balance the advantages and disadvantages of irrigation practices and increase crop water use efficiency as excessive irrigation has the potential to stress downstream ecosystems and communities by modifying the natural water cycle.

Impact of soil water on ecosystem services

As a consequence of the topics discussed in the previous sections, most of the processes related to water in soils are tightly interlinked and affect ecosystem services that are themselves related to environmental, societal, and population issues (Smith et al., 2015). This is especially true in the case of provisioning services, which are involved in the creation of physical products, and on regulating services, which are benefits obtained by the regulation of ecosystem processes (Vereecken et al., 2016).

Ecosystem provisioning services rely on soil water that regulates the production of food from plants and animals, raw materials (e.g., timber, pulp, fibers, fodder, and fertilizers), genetic resources, freshwater, energy (e.g., biofuel and wood), and ornamental resources. To a large extent, this is due to the tight relationship between water availability, plant transpiration, and photosynthesis but also to the availability of nutrients and to soil quality which are both impacted by the water in soils. Excessive water inputs can transport nutrients through erosion, leaching, and subsurface flow. Such nutrients are not only made unavailable to plants but negatively impact the water quality of ground and surface freshwater resources. Water has an impact on the ability of soils to provide

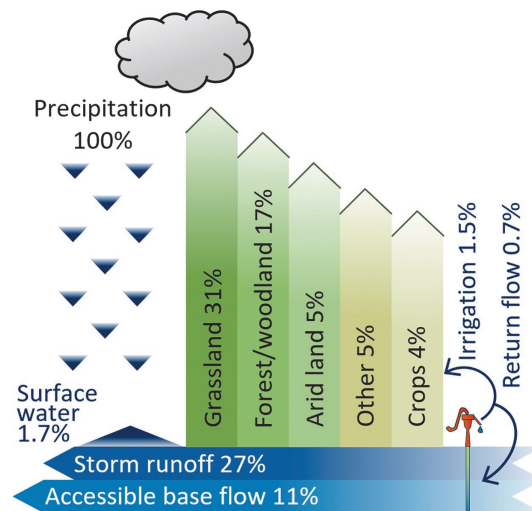


Fig. 3 Green and blue water global flows with fluxes obtained from Falkenmark et al. (2004). As indicated with text citations, data from: Falkenmark M, Rockström J and Rockström J (2004) *Balancing Water for Humans and Nature: The New Approach in Ecohydrology*. Earthscan.

additional provisioning services. For example, it impacts soil consolidation and strength and thus its ability to provide physical support and viable biological habitats.

In the context of regulatory services, water in soils indirectly affects climate through the regulation of energy related processes, plant communities, and their impact on surface albedo. This can, for example, affect the occurrence of rainfall events, drought, and extreme heatwaves. Over longer time scales, by changing levels of carbon sequestration, soil serves as a key sink of greenhouse gases. Here, changes in the amount of water in soils can significantly affect carbon stocks, having strong impacts on and feedbacks to the Earth's climate system. For example, by regulating soil temperature, soil water variations can increase or decrease the rate of microbial decomposition of organic matter. In addition, large portion of land can become a high-rate source of greenhouse gases, as is the case in the rapid thawing of permafrost soils. In addition, strong precipitation events lead to the transport of dissolved carbon or of particulate organic matter. Another indirect effect on carbon stocks is the occurrence of dry and saturated conditions that affect plant activity and thus the amount and composition of litter inputs and rhizodeposition. Within this framework, the promotion of carbon stocks in soils not only has the potential to offsets greenhouse emissions but also to maintain or improve soil hydrological functions.

Representing and monitoring water in soils

Mass balance and energy transfer methods

Mass balance methods are mathematical models used to describe changes in the water that is stored in a control volume (i.e., a soil profile). These methods make use of the principle of conservation of mass, where the water inputs are equal to the outputs. One of the simplest approaches is the water balance equation:

$$\partial S = P - ET - R$$

where ∂S is the change in soil water content, P and ET are precipitation and evapotranspiration respectively, and R is runoff which can be divided in surface and groundwater runoff. Models such as the water balance generally assume that water flow is driven by gravity whereas more complex models can consider both gravity and matric potential. Such models are often coupled to a crop module, since plants can exert a strong control over water flow. The most common approach that includes matric potential to describe water movement in soil is the Richardson-Richards equation (Vereecken et al., 2016):

$$\frac{\partial \theta_w}{\partial t} = \frac{\partial}{\partial z} \left[K(h) \left(\frac{\partial h}{\partial z} - 1 \right) \right] - Q$$

Where θ_w is volumetric water content, t is time, z is the vertical coordinate, and K is the hydraulic conductivity of the soil as a function of soil matric potential h . Such function is highly non-linear and can only be approximated using numerical methods. In unsaturated conditions, both θ_w and K are related to h and can be estimated from soil hydraulic properties (e.g., soil textural information, bulk density, and organic carbon content) by using pedotransfer functions (Van Looy et al., 2017). Plant water uptake is taken into account by the Q term which represents water that is lost due to plant transpiration and is equal to (Vanderborght et al., 2021):

$$Q = \varphi(h) Q_p(z)$$

where the potential root water uptake Q_p is multiplied by a stress factor φ accounting for soil matric potential conditions. This factor φ varies between 0 and 1 with the latter representing optimal soil matric potential (and thus water availability) conditions for plant transpiration. The φ factor is calculated using threshold values of h that can vary between different plants. Generally, extreme values of 0 for φ are assumed in saturated soils (anaerobic stress) and in soils below wilting point. The Richardson-Richards equation has a long history of successful applications but is commonly restricted to one-dimensional models. In part, this is a consequence of its high computational demand, but its application also needs precise and spatially distributed soil information (e.g., soil characteristics and hydraulic properties) that can be challenging to collect. Depending on the application, a one-dimensional modelling environment that does not consider lateral flow can neglect important components of the analyzed system, one example being the study of groundwater pumping and its effects on the surrounding aquifer. Thus, catchment-scale studies and other large-scale applications often make use of relatively simpler methods that do not fully characterize θ_w and K .

Energy transfer methods, similarly to mass balance methods, also require a control volume to be defined but in this case energy is conserved. Here the following surface energy balance equation can be applied:

$$R_n = H + E_L + G + \frac{dS}{dt}$$

where the net radiation R_n is balanced by fluxes of energy out of the control volume and by changes of energy stored within the control volume. The former fluxes are represented by sensible heat (H) and latent heat (E_L) fluxes towards the atmosphere and by ground heat flux (G) towards the soil. The latter change in energy is represented by dS/dt which is the change in the heat that is stored in the system (e.g., in the vegetation). Here, the change in soil water content at the surface is due to actual evapotranspiration ET , which is related to latent heat through the latent heat of vaporization ($ET = E_L/(\lambda \rho_w)$) resulting in:

$$ET = \left(R_n - H - G - \frac{dS}{dt} \right) / (\lambda_v \rho_w)$$

where ρ_w is the density of water. Alternatively, if temperature and vapor pressure deficit at the surface (T_s and e_s) and at a certain height (T_a and e_a) are known, the actual ET can be obtained through the Bowen ratio $B = H/E_L$. Here, B can be calculated using:

$$B \approx \frac{T_s - T_a}{e_s - e_a} \quad \text{and} \quad E_L = \frac{R_n - G}{1 + B}$$

if it is assumed that the change in heat stored by the vegetation is small enough and can thus be neglected.

Measurement techniques for soil water content sensing

A wide range of measurement techniques that determine the temporal and spatial variability of soil water content have been developed in the past 50 years, as such information is essential for a large number of studies. At the point-scale, both Electromagnetic (EM) methods and tools based on thermal soil properties (heat pulse sensors) are viable measurement techniques, with the former being the most common pick. EM methods rely on the dependency of the soil dielectric permittivity on soil water content. Since the dielectric permittivity of liquid water is much higher than that of other soil components, soil water is the primary factor that governs EM wave propagation in the soil. Such methods comprise the use of time domain reflectometry (TDR), time domain transmission (TDT), and capacitance and impedance sensors. TDR and TDT sensors measure the propagation velocity of EM waves along open and closed transmission lines, respectively, whereas capacitance sensors typically determine soil water content by measuring the charge time of a capacitor (i.e., the soil-probe system) for a given voltage. Typically, capacitance sensors operate at a measurement frequency between 50 MHz and 150 MHz while TDR and TDT operate at higher frequencies. High frequencies are favorable since they result in soil water content measurements that are less influenced by the electrical conductivity and by the imaginary dielectric permittivity of the soil. The need for automated measurement systems have largely increased the number of low-cost sensors, such as those for irrigation purposes. More recently, wireless sensor network has appeared as a promising approach for soil water content monitoring over large areas and with a high temporal resolution, which is particularly useful for the observation of ecohydrological processes, for the spatial characterization of soil properties, and for the validation of remotely sensed soil water content products (Bogena et al., 2010).

At an intermediate scale, several new hydrogeophysical measurement technologies have emerged that attempt to address the need for field scale soil water content measurements (Bogena et al., 2015). Some of these are addressed in more details in other chapters that are focused on proximal and remote sensing of soil water. Cosmic-ray neutron sensing (CRNS), for example, counts secondary fast neutrons near the soil surface that are created by primary cosmic-ray particles. These fast neutrons are closely related to the root zone soil water content within a horizontal footprint of about 200 m. Alternatively, the Global Navigation Satellite System Reflectometry (GNSS-R) signal can be used to infer soil water content since it is affected by the soil permittivity, and thus soil water, at the land surface within a radius of 50 m. Another promising approach is the use of permanently installed gamma-ray intensity monitoring stations. Here, the attenuation of gamma radiation originated from the decay of radioactive isotopes in the soil can be sensed within a horizontal footprint of 25 m approximately. Finally, non-invasive geophysical methods such as Electromagnetic Induction (EMI), Ground penetrating Radar (GPR) and Electrical Resistivity (ER) can be used to detect and study soil water content dynamics alongside water salinity, soil textural properties, and porosity among other properties. EMI induces a secondary magnetic field in the ground through the emission of a primary magnetic field and uses it to measure the apparent electrical conductivity of the soil. In GPR, an EM wave is propagated in the subsurface and its reflections are sensed at the soil surface to measure soil permittivity or electrical conductivity. ER provides information on the resistivity distribution of the subsurface which can be used to monitor soil water dynamic processes over time and space (e.g., through ER tomography).

At the regional to global scale, satellite-based sensors such as the Soil Moisture and Ocean Salinity (SMOS), Soil Moisture Active and Passive (SMAP), and Advanced Scatterometer (ASCAT) are used to monitor soil water content with a temporal resolution of a few days. However, such products have a rather coarse resolution of tens of kilometers. Higher resolution data can be obtained by synthetic aperture radars (SAR) such as the European Space Agency's Sentinel-1 and Japan Aerospace Exploration Agency's ALOS-2, alas with lower temporal resolution. In the next decade, several new perspectives are envisaged with the upcoming SAR missions NISAR (NASA ISRO Synthetic Aperture Radar), ROSE-L (Radar Observing System for Europe at L-band), and the BIOMASS mission operating at longer wavelengths. Additionally, an emerging technique for measuring soil water content from space is based on GNSS-R technology onboard small satellites (e.g., Cyclone GNSS) that provides good spatial and temporal resolution (<5 km, sub-daily) at low cost.

Alongside soil water content, another fundamental variable is soil water potential, which drives water dynamics by reflecting the energy state of water in the porous soil. Despite its paramount importance, measurements of soil water potential has not experienced the same fast-paced development found in soil water content measurements (Durner and Or, 2006). The main reasons being the need for hydraulic contact between the measuring instrument and the surrounding soil, and the lack of a measurement technique that can measure across a wide soil water content range. The selection of a particular instrumentation will thus depend on the range of interest of each given application. The most common instruments are tensiometers, which are effective between $\sim 10^{-1}$ and $\sim 10^2$ kPa (high soil water content) and are thus suited for the investigations of relatively rapid processes such as solutes transport to the groundwater. The concept behind tensiometers is long-known and several improvements were made to increase the

match between the true soil state and the tensiometer reading (advanced tensiometers) or to reduce the necessity for continuous maintenance (self-filling tensiometers). However, tensiometers generally require skilled operators, are strongly affected by soil temperature in ways that are challenging to correct, and have a limited measurement range. For example, they are generally unsuited for agricultural studies in which low soil water content limits plant growth. In the intermediate range, reference wettable porous media such as gypsum, fiberglass, and nylon are used to measure between $\sim 10^1$ and 10^5 kPa. In this range, the impossibility of obtaining direct measurements is resolved by measuring soil water content or water-content related properties (e.g., through heat dissipation or electromagnetic properties) in a reference material that is in hydraulic equilibrium with the surrounding soil. Generally, factory calibration allows the retrieval of soil water potential from soil water content. Such instruments consist of porous materials that have a wide range of pore sizes. Blocks of gypsum with embedded electrodes have been used for more than 80 years and, more recently, sensors made of gypsum wafers embedded in a granular matrix have become available. Finally, thermocouple psychrometers measurements are used in the range between $\sim 10^5$ and 10^8 kPa (very low soil water content). But their measurement range and their high susceptibility to thermal gradients makes them unsuited for the use at shallow soil depths.

Soil water and the issue of scales

Soil water content measurements cover a broad range of scales due to the numerous available sensing techniques and their fundamental differences. There is often an additional difference between measurements and modelling, which requires a change of scale between measurements and predictions. To simplify this issue, a scale triplet that applies to both measurements and models was proposed by Western and Blöschl (1999). Here, three components: spacing, extent, and support, are needed to uniquely specify the dimensions in space of a measurement or a model. Spacing is the distance between samples or elements of a model, extent is the spatial coverage (e.g., length of a transect or length of a model domain), and support is the volume of influence of a measurement or the size of a model element. Fig. 4 shows the support (black lines) and extent (colored area) of a set of measurement techniques. It is apparent how overlaps and gaps between measurements often occur, which is why the simultaneous use of different techniques is both promising and challenging. It is thus imperative to establish a framework where information from measurements and modelling predictions are optimally combined at various spatial scales and within different disciplines (Vereecken et al., 2008).

An appropriate combination of multi-scale measurements and modelling of soil water can provide large benefits to the rapidly expanding establishment of terrestrial observatories. These are research instruments composed of geographically distributed infrastructures from different fields of research that are networked through real-time, state-of-the-art communication technology (Bogena et al., 2006). The focus is the observation of terrestrial systems, which consist of subsurface, land surface (including biosphere), lower atmosphere, and anthroposphere. As shown in Fig. 4, systems are hierarchically organized along evolving scales and structures with hydrological units representing the basic scaling units. Even though terrestrial systems are extremely complex, most process-based climate and biosphere models generally provide a simplified representation of the terrestrial component. Terrestrial observatories could help remedy such deficiency and contribute to the validation and development of novel model concepts as well as methods for upscaling parameters, fluxes, and state variables. They could also provide information on interactions between the soil-vegetation and atmosphere, on the impact of land use and climate change, and on the effects of human interventions. Finally, terrestrial observatories represent a key added value to the validation and development of integrated approaches that can measure and model soil water across different scales, thus bridging key gaps in hydrological and terrestrial sciences.

Examples of the combination of soil water measurements and modelling

The value of soil water content information

The increasing use of nondestructive and remote sensing techniques to measure and monitor soil water content has resulted in an unprecedented and still growing data availability across a range of scales. An improved spatial and temporal characterization of soil water content brings a number of scientific benefits in hydrology for which Vereecken et al. (2008) offered a thorough review. Overall, the authors conclude that a unified concept for multiscale observation and interpretation of soil water content data is needed but the hydrologic community somehow lacks behind other disciplines in this regard. For this, the implementation of multiscale monitoring sites such as terrestrial observatories and appropriate databases are needed to validate upscaling theories and catchment scale models.

What benefits result from the acquisition of soil water content data? Data assimilation is for example used to improve model predictions of water fluxes and to estimate vadose zone properties. Key methods are direct insertion, statistical correction assimilation, Newtonian nudging, inverse modeling, variational approaches, linear regression equations, and sequential data assimilation methods, also known as Kalman filtering (KF) techniques. Of the latter, standard KF is the most prominent and general estimator for systems where linear dynamics are the dominant governing factor. Unfortunately, the dynamics of interest are in this case not linear, which motivates the use of extended KF and, more recently, ensemble KF which avoids many of the problems of the former filter by implementing a Monte Carlo method. The relative ease of implementation and computational efficiency make ensemble KF a very attractive option in vadose zone hydrology.

Near surface soil water content measurements have also been used to improve the retrieval of the spatiotemporal variability of water storage in the root zone and within the soil profile. In doing this, several issues should be considered such as (a) the

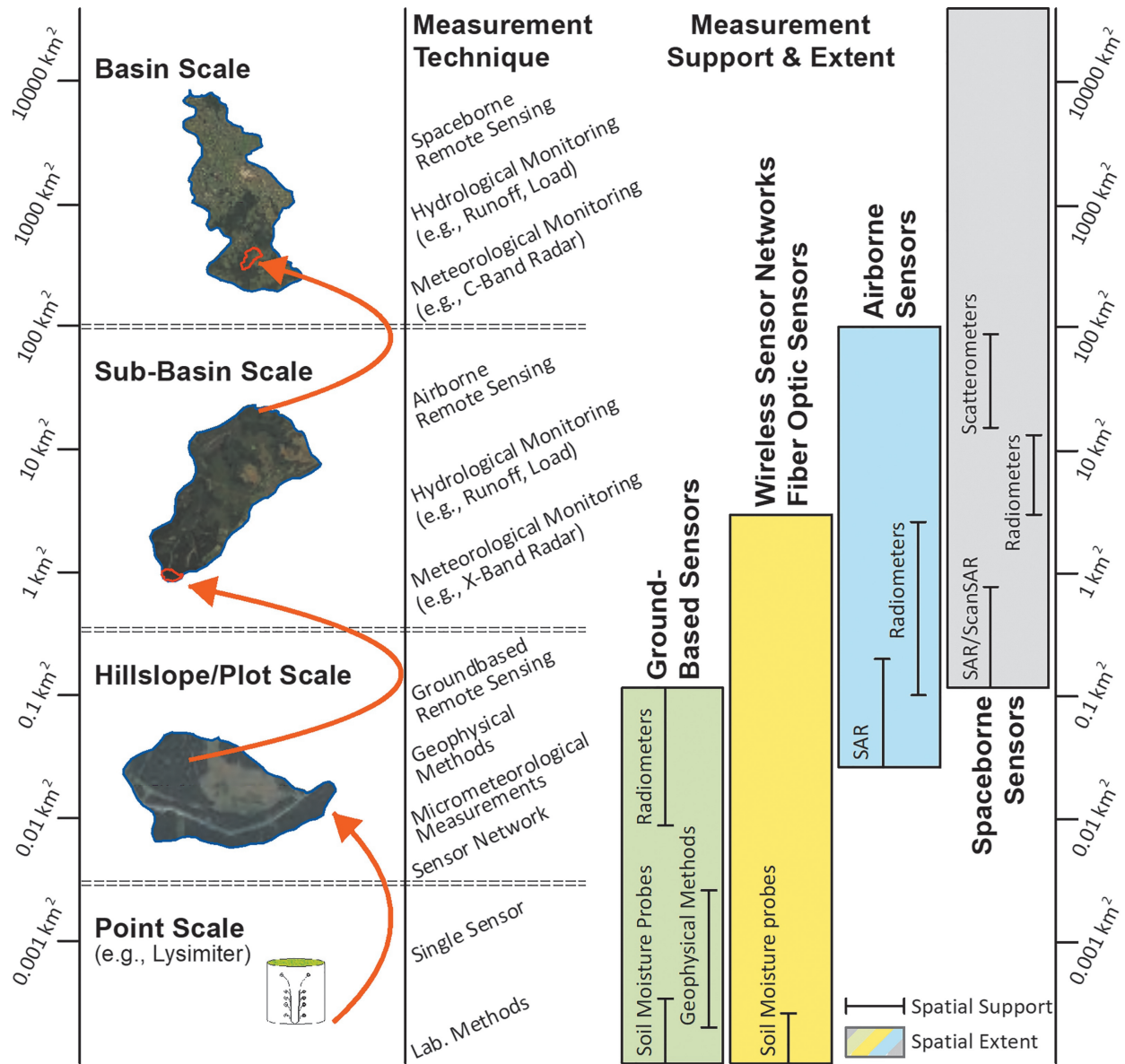


Fig. 4 On the left side, hierarchy of evolving scales from the point scale to the regional basin scale and relative measurement techniques and upscaling from Bogaen et al. (2004). On the right side, support and extent scales of soil moisture measurements and observations from Vereecken et al. (2008). Here, ground-based sensors are soil moisture probes (e.g., capacitance probes and TDR), geophysical methods (GPR and EMI), and radiometers (e.g., ELBARA); wireless sensor networks are soil moisture probes (e.g., capacitance probes); airborne sensors are SAR (e.g., E-SAR) and radiometers (e.g., ESTAR, PBMR, PALS); and spaceborne sensors are SAR/ScanSAR (e.g., ERS1-2, RADARSAT, ENVISAT, JERS, and ALOS), radiometers (e.g., SMMR, AMSR-E, and SMOS), and scatterometer (e.g., ERS1-2 and METOP). Was modified from data and figures From: (i) Bogaen H, Hake JF, and Vereecken H (2004) Water and Sustainable Development. Forschungszentrum Jülich. and (ii) Vereecken H, Huisman J, Bogaen H, Vanderborght J, Vrugt J and Hopmans J (2008) On the value of soil moisture measurements in vadose zone hydrology: A review. Water Resources Research 44(4).

requirement of a mathematical model that incorporates root water uptake, (b) an a priori knowledge of initial soil water content, and (c) a correct specification of hydraulic parameters. A priori information on soil hydraulic properties can be obtained using pedotransfer functions as direct measurements are rarely available. Additional state variables (e.g., temperature, leaf area index (LAI), water table depth) are needed to improve estimations. A further limitation is the eventual lack of correlation between the soil water content at the surface and in the subsurface, which can be overcome by coupling surface and subsurface soil water content. Finally, the accuracy of the input data affects the accuracy of the output with field-based measurements techniques showing strong accuracy and thus a useful addition to satellite-based measurements.

Soil hydraulic properties at the column scale are often estimated from flux measurements (e.g., evaporation and outflow) within an inverse modelling framework. In contrast, applications at the field-scale are scarce and showed that additional information (e.g., soil matric potential, soil structural characteristics, and measured hydraulic properties) and homogeneous soil assumptions are

required to constrain parameter estimation. Recently, an increase in computational power, novel optimization methods, the availability of remotely sensed soil water content and heat fluxes, and geophysical measurements has generated new opportunities for the estimation of hydraulic properties at the field to catchment scale.

Hydrological fluxes can be estimated from the spatiotemporal characterization of soil water content by using a hydrological model that is as complex as the system being studied. For example, there is a long tradition in estimating bare soil evaporation from surface soil water content. Furthermore, remotely sensed data are used to characterize soil water fluxes although this approach is generally limited to the shallow topsoil. A promising combination of remote sensing and ground-based methods (e.g., geophysical measurements) might bridge this gap in the upcoming years and provide additional benefits, for example in the prediction of the groundwater recharge. Applications to further fluxes such as runoff are still under debate, although it is known that soil water content patterns can strongly influence runoff generation. Similarly, the integration of a mechanistic based approach of root water uptake in the Richardson-Richards equation is challenging as it requires information about the soil root architecture, parameterization, and its development in time. For surface energy fluxes, it is known that the quality of regional-scale climate and weather simulations might be diminished by neglecting small-scale soil water content variability. However, not all studies corroborated such scenario, especially when soil water content is rather homogeneous in space due to, for example, soil drying.

The importance of soil water content and its spatial variability in respect to crop productivity and carbon fluxes

In the following studies, measurements of soil water content and related parameters were combined with the agroecosystem model AgroC to study water content dynamics and their effects on local water scarcity, crop performance, and carbon fluxes at the Selhausen test site near the Jülich Research Centre (Germany). The following examples are focused on the year 2016, which had hydroclimatic conditions close to the 30-year average. Thus, the importance of such research could be enhanced in years with more extreme drought conditions. The AgroC model couples (a) the SOILCO2 module for vertical water, heat, and CO₂ fluxes, (b) the RothC module for turnover of organic carbon, and (c) the SUCROS module for simulating crop growth and organic matter accumulation. Here, SOILCO2 solves the one-dimensional Richardson-Richards equation while, in SUCROS, the potential root water uptake is reduced by a stress factor φ that depends on fixed dry or wet soil conditions.

A series of 81 spatially distributed AgroC simulations were performed over an area of $\sim 1 \times 1$ km comprising five crops: sugar beet, silage maize, potato, winter barely, and winter wheat (Brogi et al., 2020). Information on soil hydraulic properties for 18 soil types were retrieved from a geophysics-based soil mapping product (Brogi et al., 2019). Simulations of soil water content dynamics were compared with measurements to corroborate the robustness of the soil description. For this, atmospheric forcing was provided by a climate station located within the study area. The simulations allowed the reconstruction of spatial and temporal patterns of leaf area index (LAI) for the five crops across the studied area, which were successfully compared with RapidEye satellite images captured throughout the growing season. The reduction of LAI and associated yield loss differed across the study area. Lower LAI and yields were found on soil-crop combinations where available water was limited by unfavorable soil characteristics, high demand from the crops, and low precipitation (Fig. 5A and B). By using this method, precise maps of simulated yield dependence on water availability were produced, which outperformed the use of commonly-available soil representations (Brogi et al., 2021). Such maps are relevant in the support of agricultural decision-making for crop type and variety selection, precision irrigation and fertilizer application, seeding date selection, and crop rotation.

A 2.8 ha sugar beet field located within the same 1×1 km study area was investigated (Herbst et al., 2021). Measurements of soil water content, net ecosystem exchange (NEE) from both point-scale chamber and field-scale eddy covariance (EC) methods, soil respiration, soil temperature, and crop physiology were carried out. These were combined with the AgroC model to simulate and understand the within-field patterns of CO₂ fluxes linked to soil water availability, soil characteristics, and state variables. During dry periods with low soil water content, the NEE measured at 18 locations with the chamber method was generally in agreement with the EC measurements but showed strong within-field variability. These 18 points were thus divided in two clusters based on the underlying soil characteristics that impacted soil water availability, and which were inferred from geophysical measurements and direct soil sampling. Simulations and measurements showed that limited soil water and the resulting reduction in root water uptake were the main driver of the spatiotemporal variability in crop growth and CO₂ fluxes. Here, the cluster with the stronger water limitation had consistently smaller biomass growth, greater soil respiration, and lower magnitude of NEE (Fig. 5C and D). This study showed how combined EC and point scale measurements can characterize complex within-field patterns in NEE related to soil water availability and how agroecosystem modelling can be used for the gap-filling of NEE measurements during drought periods. It also showed that patterns in carbon fluxes related to soil water content dynamics can be successfully measured and simulated at the field and sub-field scale, which could be of interest for future assessments of the carbon footprint of agricultural practices.

On the use of soil water content data at different scales

A variety of examples, where information on soil water content was used across scales in model evaluation, data assimilation, or refinement of coarse products can be found in literature. At a relatively small hillslope-scale, Gebler et al. (2019) combined land surface-subsurface modelling with data assimilation of soil water content on a 38 ha grassland test site in Rollesbroich (Germany). High-resolution soil water content was obtained from a dense sensor network with 61 locations and three depths (5, 20, and 50 cm). The ensemble KF was used to assimilate such measurements within the Terrestrial System Modelling Platform (TerrSysMP)

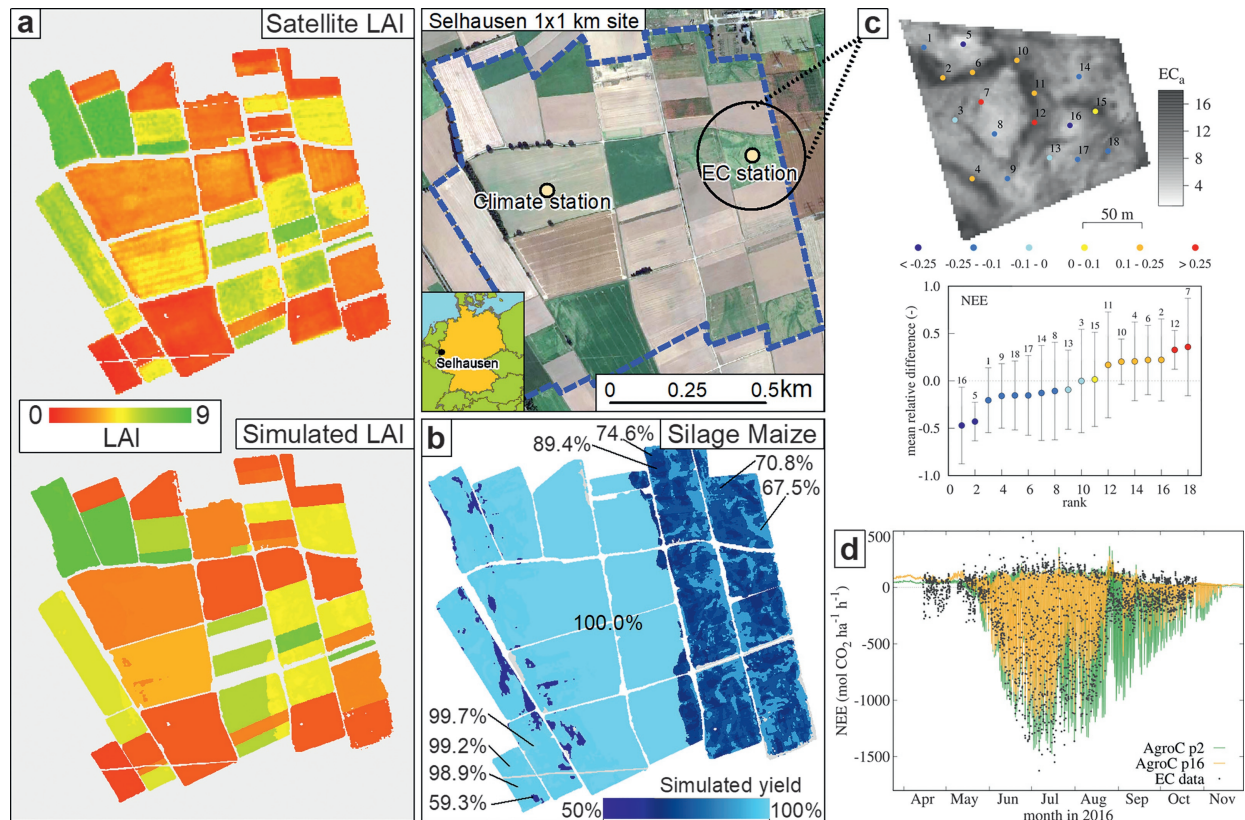


Fig. 5 Location and outline of the 1×1 km Selhausen test site (North Rhine-Westphalia, Germany) with (A) satellite-derived leaf area index (LAI) and LAI simulated with AgroC at the 9th of June 2016, (B) yield map for silage maize for the 2016 growing season, (C) apparent electrical conductivity (ECa) map of a selected 2.8 ha field with the locations of 18 point-scale net ecosystem exchange (NEE) measurements and their ranked mean relative differences, and (D) hourly eddy covariance (EC) measurements of NEE and AgroC simulations of points p2 (average soil moisture) and p16 (low soil moisture). Was modified from data and figures from: (i) Brogi C, Huisman J, Herbst M, Weihermüller L, Klosterhalfen A, Montzka C, Reichenau T and Vereecken H (2020) Simulation of spatial variability in crop leaf area index and yield using agroecosystem modeling and geophysics-based quantitative soil information. *Vadose Zone Journal* 19: e20009 and (ii) Herbst M, Pohlig P, Graf A, Weihermüller L, Schmidt M, Vanderborght J and Vereecken H (2021) Quantification of water stress induced within-field variability of carbon dioxide fluxes in a sugar beet stand. *Agricultural and Forest Meteorology* 297: 108242.

coupled to the Parallel Data Assimilation Framework (PDAF) on both a real-world example and a synthetic illustration (mimicking the real-world circumstances). This study showed that combining data assimilation and high-resolution physically based models can lead to meaningful improvements in soil water content and discharge estimation at the hillslope scale. Although better results were obtained for the synthetic illustration, especially in the simulation of discharge, reduced root mean square error of soil water content predictions was obtained for the real-world example as well. The relatively poorer improvements in the real-world example were partially attributed to model structural errors, biases in prior parameterization, and site-specific effects, which highlighted the importance of the representation of local small-scale processes and of a precise geostatistical parameterization. Furthermore, it suggested that multiple data types (e.g., evapotranspiration) should be assimilated to improve model predictions.

At the intermediate catchment scale, Baatz et al. (2017) showed the benefits of assimilating soil water content data in the Community Land Model CLM4.5. Such data was provided by nine cosmic ray neutron sensors (CRNS) located within the Rur catchment (Germany). Since a CRNS device can only provide local information due to the ~ 240 m maximum footprint, the local ensemble Kalman filter (LETKF) was used to update soil water content at unmonitored locations. Estimates of soil water content were improved by joint state-parameter estimates both during the assimilation (2011–2012) and during the evaluation period (2013). When a biased soil map was used as input for the simulations, RMSE and bias improved considerably and approached those obtained by using an actual regional soil map. The benefits of the joint state-parameter updates were confirmed with additional jackknife experiments. At one level, this study shows the potential of CRNS to improve subsurface parameterization in regional land surface models, especially in areas of the globe where prior information on soil properties is limited due to the scarcity of detailed soil maps. More broadly, the advantages provided by CRNS networks highlight the potential benefits of further extending CRNS networks at the European scale, which will require further harmonization as later proposed by Bogaen et al. (2022).

At the continental scale, Naz et al. (2019) assimilated coarse resolution soil water content maps derived from the European Space Agency (ESA) Climate Change Initiative (CCI) for the period 2000–2006 with the Community Land Model (CLM3.5) integrated with the Parallel Data Assimilation Framework (PDAF). This was carried out for Europe at a high spatial resolution of

~3 km. The assimilation of satellite data provided meaningful improvements in estimates of soil water content, which improved over all the investigated regions, except Scandinavia, and provided better predictions for summer and autumn seasons compared with winter and spring seasons. This discrepancy was attributed to data gaps, and to the presence of snow and frozen soils that limited the assimilation of soil water content in the selected modelling environment. Overall, total runoff showed only marginal improvements, especially in central Europe. This study proved (a) the potential of satellite sourced soil water content data assimilation to produce improved high-resolution simulations at the continental scale and (b) the added value of long-term observation of soil water content for hydrologic simulations in climate change studies. With the recent availability of COSMO-REA6, the period was extended until 2015 using the proposed methodology and thus obtaining a reliable high-resolution long-term soil water content product (Naz et al., 2020).

Summary and outlook

The small amount of Earth's water stored in soils is key in hydrological, biogeochemical, and energy related processes as it feeds back to vegetation growth, soil and air temperature, infiltration, and carbon storage among other processes. Thus, knowledge of the temporal and spatial variability of soil water is a paramount component of multidisciplinary research as it deeply affects natural and human environments. This is especially true for regions where extreme flooding is due to rainfall exceeding the infiltration capacity of the soils or where drought events are becoming more frequent and severe. Multiple methods and techniques are available to measure, monitor, represent, and simulate water in soils and its dynamics. Such methods vary in scale and in field of application, which sometimes results in powerful combinations of multiple techniques and models but also adds several complexities. Thus, the high spatial and temporal variation of soil water content and the fact that multiple scales (from the small field-scale to global) must be considered, should direct the hydrologic community towards more unified concepts for multiscale observations, predictions, and interpretation of water in soils.

References

- Aeschbach-Hertig W and Gleeson T (2012) Regional strategies for the accelerating global problem of groundwater depletion. *Nature Geoscience* 5(12): 853–861.
- Baatz R, Franssen H-JH, Han X, Hoar T, Bogena H, and Vereecken H (2017) Evaluating the value of a network of cosmic-ray probes for improving land surface modeling. *Hydrology and Earth System Sciences* 21: 2509–2530.
- Bickel S and Or D (2020) Soil bacterial diversity mediated by microscale aqueous-phase processes across biomes. *Nature Communications* 11(1): 1–9.
- Bogena H, Hake J-F, and Vereecken H (2004) *Water and Sustainable Development*. Forschungszentrum Jülich.
- Bogena H, Schulz K, and Vereecken H (2006) Towards a network of observatories in terrestrial environmental research. *Advances in Geosciences* 9: 109–114.
- Bogena H, Herbst M, Huisman J, Rosenbaum U, Weuthen A, and Vereecken H (2010) Potential of wireless sensor networks for measuring soil water content variability. *Vadose Zone Journal* 9(4): 1002–1013.
- Bogena HR, Huisman JA, Güntner A, Hübner C, Kusche J, Jonard F, Vey S, and Vereecken H (2015) Emerging methods for noninvasive sensing of soil moisture dynamics from field to catchment scale: A review. *Wiley Interdisciplinary Reviews Water* 2(6): 635–647.
- Bogena H, Schrön M, Jakobi J, Ney P, Zacharias S, Andreasen M, Baatz R, Boorman D, Duygu BM, and Eguibar-Galán MA (2022) COSMOS-Europe: A European Network of Cosmic-Ray Neutron Soil Moisture Sensors. *Earth System Science Data Discussions* 14(3): 1125.
- Broggi C, Huisman JA, Pätzold S, von Hebel C, Weihermüller L, Kaufmann MS, van der Kruk J, and Vereecken H (2019) Large-scale soil mapping using multi-configuration EMI and supervised image classification. *Geoderma* 335: 133–148. <https://doi.org/10.1016/j.geoderma.2018.08.001>.
- Broggi C, Huisman J, Herbst M, Weihermüller L, Klosterhalfen A, Montzka C, Reichenau T, and Vereecken H (2020) Simulation of spatial variability in crop leaf area index and yield using agroecosystem modeling and geophysics-based quantitative soil information. *Vadose Zone Journal* 19: e20009.
- Broggi C, Huisman JA, Weihermüller L, Herbst M, and Vereecken H (2021) Added value of geophysics-based soil mapping in agro-ecosystem simulations. *The Soil* 7(1): 125–143.
- Dorigo W, Dietrich S, Aires F, Brocca L, Carter S, Cretaux J-F, Dunkerley D, Enomoto H, Forsberg R, and Güntner A (2021) Closing the water cycle from observations across scales: Where do we stand? *Bulletin of the American Meteorological Society* 102(10): E1897–E1935.
- Durner W and Or D (2006) Soil water potential measurement. In: *Encyclopedia of Hydrological Sciences*. Wiley.
- Falkenmark M and Rockström J (2006) The new blue and green water paradigm: Breaking new ground for water resources planning and management. *Journal of Water Resources Planning and Management* 132(3): 129–132.
- Falkenmark M, Rockström J, and Rockström J (2004) *Balancing Water for Humans and Nature: The New Approach in Ecohydrology*. Earthscan.
- Faurès J-M, Hoogeveen J, and Bruinsma J (2002) *The FAO Irrigated Area Forecast for 2030*. Rome, Italy: FAO1–14.
- Gebler S, Kurtz W, Pauwels V, Kollet S, Vereecken H, and Hendricks Franssen HJ (2019) Assimilation of high-resolution soil moisture data into an integrated terrestrial model for a small-scale head-water catchment. *Water Resources Research* 55(12): 10358–10385.
- Herbst M, Pohl P, Graf A, Weihermüller L, Schmidt M, Vanderborght J, and Vereecken H (2021) Quantification of water stress induced within-field variability of carbon dioxide fluxes in a sugar beet stand. *Agricultural and Forest Meteorology* 297: 108242.
- Huang H, Calabrese S, and Rodríguez-Iturbe I (2021) Variability of ecosystem carbon source from microbial respiration is controlled by rainfall dynamics. *Proceedings of the National Academy of Sciences* 118(52): e2115283118.
- Jung M, Reichstein M, Ciais P, Seneviratne SI, Sheffield J, Goulden ML, Bonan G, Cescatti A, Chen J, and De Jeu R (2010) Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature* 467(7318): 951–954.
- Naz BS, Kurtz W, Montzka C, Sharples W, Goergen K, Keune J, Gao H, Springer A, Hendricks Franssen H-J, and Kollet S (2019) Improving soil moisture and runoff simulations at 3 km over Europe using land surface data assimilation. *Hydrology and Earth System Sciences* 23(1): 277–301.
- Naz BS, Kollet S, Franssen H-JH, Montzka C, and Kurtz W (2020) A 3 km spatially and temporally consistent European daily soil moisture reanalysis from 2000 to 2015. *Scientific Data* 7(1): 1–14.
- Schumann AH (1998) Infiltration: Introduction. In: *Hydrology and Lakes*, p. 418. Netherlands: Springer. https://doi.org/10.1007/1-4020-4513-1_129.
- Seneviratne SI, Corti T, Davin EL, Hirschi M, Jaeger EB, Lehner I, Orlowsky B, and Teuling AJ (2010) Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews* 99(3–4): 125–161.

- Smith P, Cotrufo M, Rumpel C, Paustian K, Kuikman P, Elliott J, McDowell R, Griffiths R, Asakawa S, and Bustamante M (2015) Biogeochemical cycles and biodiversity as key drivers of ecosystem services provided by soils. *The Soil* 1(2): 665–685.
- Van Looy K, Bouma J, Herbst M, Koestel J, Minasny B, Mishra U, Montzka C, Nemes A, Pachepsky YA, and Padian J (2017) Pedotransfer functions in Earth system science: Challenges and perspectives. *Reviews of Geophysics* 55(4): 1199–1256.
- Vanderborght J, Couvreur V, Meunier F, Schnepf A, Vereecken H, Bouda M, and Javaux M (2021) From hydraulic root architecture models to macroscopic representations of root hydraulics in soil water flow and land surface models. *Hydrology and Earth System Sciences* 25(9): 4835–4860.
- Vereecken H, Huisman J, Boga H, Vanderborght J, Vrugt J, and Hopmans J (2008) On the value of soil moisture measurements in vadose zone hydrology: A review. *Water Resources Research* 44(4).
- Vereecken H, Schnepf A, Hopmans JW, Javaux M, Or D, Roose T, Vanderborght J, Young M, Amelung W, and Aitkenhead M (2016) Modeling soil processes: Review, key challenges, and new perspectives. *Vadose Zone Journal* 15(5): 1–57.
- Western AW and Blöschl G (1999) On the spatial scaling of soil moisture. *Journal of Hydrology* 217(3-4): 203–224.
- Wood TE, Detto M, and Silver WL (2013) Sensitivity of soil respiration to variability in soil moisture and temperature in a humid tropical forest. *PLoS One* 8(12): e80965.

Global water cycle from a soil perspective

Michael H Young^a and Dani Or^b, ^aBureau of Economic Geology, University of Texas at Austin, Austin, TX, United States; ^bDivision of Hydrologic Sciences, Desert Research Institute, Reno, NV, United States

© 2023 Elsevier Ltd. All rights reserved.

Global and terrestrial water cycle	482
Soil mediated fluxes	484
Anthropogenic effects on the global water cycle—Blue and Green water	485
Scales of measurement and modeling of soil mediated hydrologic processes	485
Key soil properties	486
Measurement and modeling of soil water at continental to global scale	486
Conclusions and outlook	487
References	488

Abstract

Important aspects of the global water cycle are influenced by soil properties and processes. Vegetation-promoted soil structure and associated biopores affect rainfall partitioning into infiltration and runoff at the catchment to continental scales; soil affects water storage and its availability for plant transpiration; resistance to soil surface evaporation varies with soil properties; and rates of groundwater recharge are modified by soil properties. The influences of soil properties and processes on regional water cycle, particularly for infiltration and surface runoff, are accentuated in humid regions and decrease for drier climate. Additionally, changes in land cover and properties from disturbances (wildfire) alter the role of soils in the hydrologic cycle. Similar effects are observed following changes to land use (e.g., clear cutting, tillage). New technologies are allowing the scientific community to more precisely quantify the global water balance, and at finer spatial and temporal scales.

Key points

- Global water balance is highly dynamic, and being altered by anthropogenic disturbances.
- Water balance is influenced by soils in nuanced ways, and depending on the regional climate.
- Satellite remote sensing is significantly improving our ability to measure global water cycle components.

Global and terrestrial water cycle

Most of Earth's diversity of life resides and is supported within a thin layer of soil atop Earth's terrestrial surface. Soil represents the primary growth medium for plants for food and ecosystem functioning, a sink for—and source of—atmospheric CO₂, a home for the most biodiverse component of life, a moderator of Earth's energy balance, and finally a significant influencer of the terrestrial water cycle, at scales from soil profile to continents. Because all life forms rely on water to function, soil water fluxes and stores affect not only the pace and partitioning of hydrologic cycle components and surface energy balance, but also sustain most of Earth's biomass (Bar-On et al., 2018). Remarkably, despite the significantly smaller area of habitable terrestrial surfaces (less than half the ocean surface area), the global biomass is overwhelmingly terrestrial (Fig. 1) and is directly supported by soil water storage (and transpiration), while the total net primary productivity on land (56.4 Gt C/yr) is similar to oceans (48.5 Gt C/yr). More significant is the nearly 80 times more biomass on land than in oceans (see Fig. 1b).

Since the previous version of this encyclopedia, a greater understanding of the many roles of soil in modulating global water cycle and climatic variables has emerged, supported by ever increasing information, tools and maps for integrating spatial and temporal observations at unprecedented resolutions. The fusion of satellite remote sensing, globally harmonized soil datasets, and georeferenced analyses offer exciting new opportunities for injecting key soil processes that affect local and global water cycle. This chapter will provide a summary of current understanding and a peek into potential future research that will further sharpen our knowledge base.

Clearly, the importance of the global water cycle can be viewed through many lenses. This chapter, however, is specifically related to the influences of soils on the water cycle at different scales—but focused more on the continental and global scales—and influence of the water cycle on soil properties and processes. These relationships are being researched vigorously today by the scientific community, especially with the growth in satellite platforms being designed and launched at faster rates, and the further development of machine learning and artificial intelligence approaches that help us parameterize point-scale or meter-scale processes using ground-based and space-based measurements. Scholarly articles are being published continuously on findings related to these topics.

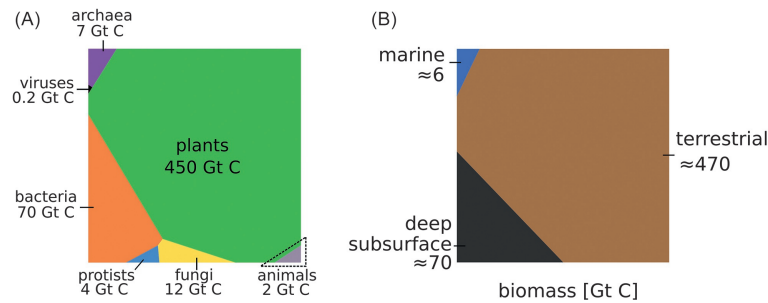


Fig. 1 (a) Global distribution of biomass, (b) distribution of biomass within different compartments of the biosphere. Nearly 82% of the global biomass is associated with plants; 13% bacteria; all other organisms account for the remaining 5%. (b): After Bar-On YN, Philips R, Milo R (2018) The biomass distribution on Earth. *Proceedings of the National Academy of Sciences* **115** (25): 6506–6511. doi: 10.1073/pnas.1711842115.

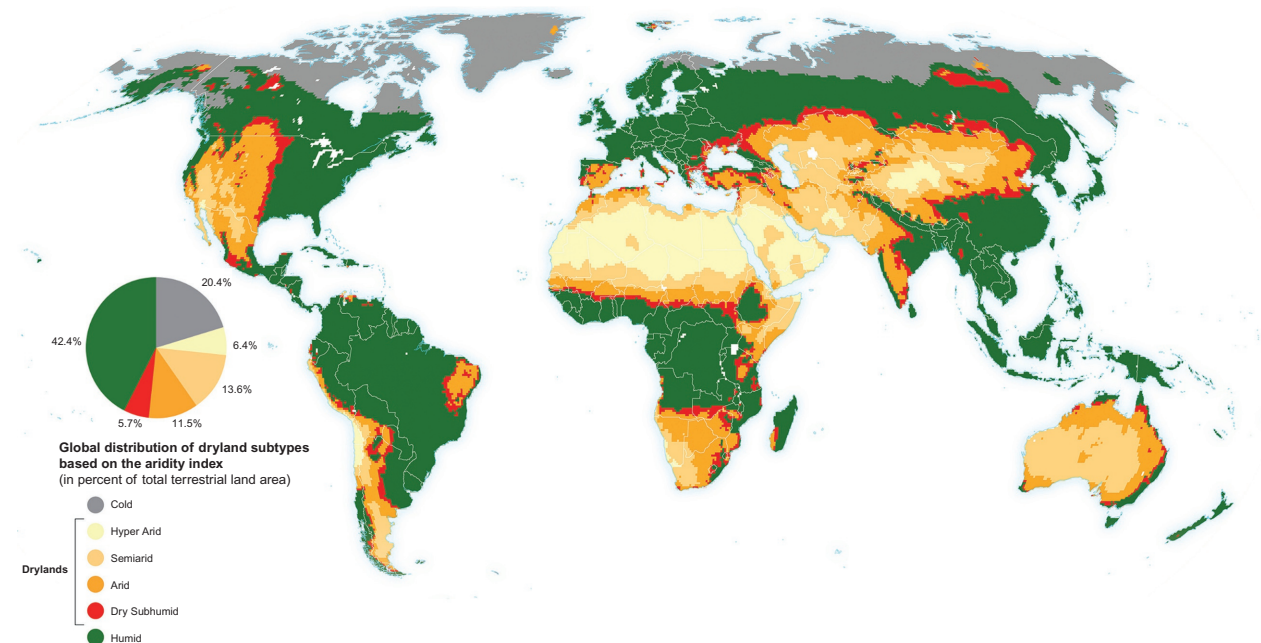


Fig. 2 Global distribution of lands of various aridity levels, from arid to humid and their relative area. Modified from World Atlas of Desertification (WAD) (2022) Joint Research Center, European Commission. <https://wad.jrc.ec.europa.eu/patternsaridity> (Accessed May 2022).

The global water cycle is complex, the amounts and partitioning of precipitation to different water balance components vary greatly across geographic regions. In particular, the ratio of precipitation to potential evapotranspiration that defines the aridity index (Fig. 2) (World Atlas of Desertification (WAD), 2022) determines water availability for vegetation and humans and is used often to distinguish humid from arid biomes. The amount of precipitation and its partitioning to evaporation and transpiration greatly affect surface energy balance as well. It is estimated that 20–22% of global precipitation and 14% of global evaporation is attributed to the land surface, the rest occurring over oceans. Recent advances in Earth observing platforms and computational power to fuse satellite and land measurements can now provide maps of precipitation at spatial resolution of 1 km² and temporal resolution of an hour (e.g., the CHelsea project) (Karger et al., 2021); the Global Precipitation Measurement program (National Aeronautical and Atmospheric Administration (NASA), 2022a, 2022b); ERA5-Land program (European Centre for Medium-Range Weather Forecasts (CMWRF), 2022). Following rainfall, evapotranspiration (ET) is the largest component of the annual terrestrial water cycle that is strongly modulated by soil processes (Fig. 3) (Oki and Kanae, 2006). Nearly 70% of terrestrial precipitation is sent back to the atmosphere as ET, with transpiration (T) through plants accounting for 2/3 of ET and evaporation (E) from land surfaces (including interception) accounting for the remaining 1/3 of ET. The convergence of high resolution climatic and soil information (e.g., OpenGeoHub, 2022) at the km scale, globally, opens unprecedented opportunities for improved understanding of how soil properties and processes affect rainfall partitioning and water storage and ultimately vegetation response and surface energy balance.

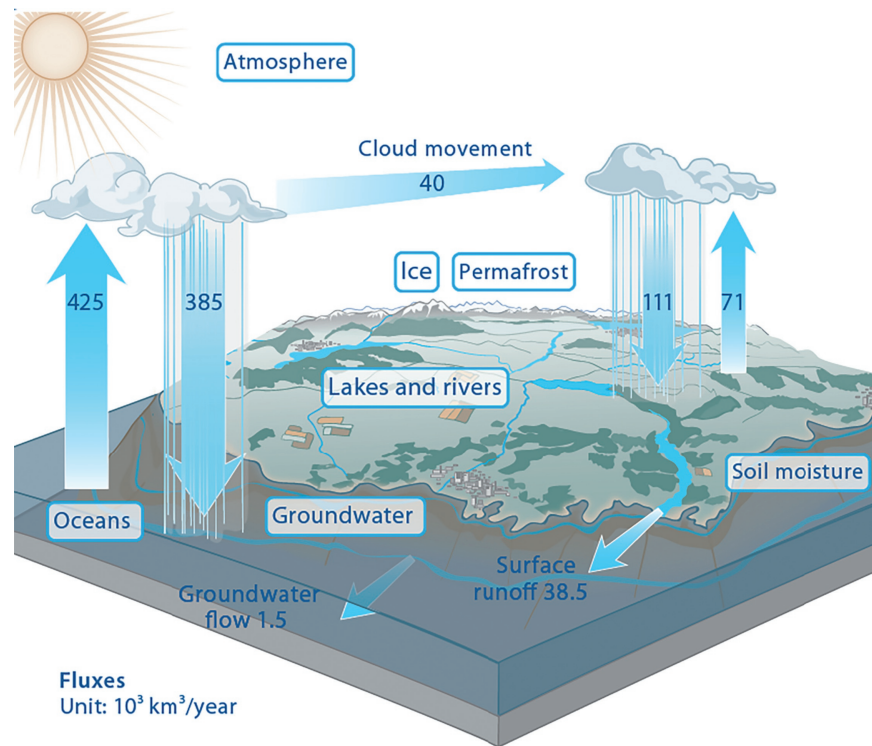


Fig. 3 Primary annual fluxes (numbers in arrows are times $1000 \text{ km}^3/\text{yr}$) and freshwater stores (shown in boxes) connected with the global water cycle. Modified from Oki T and Kanae S (2006) Global hydrological cycles and world water resources. *Science* **313**: 5790. doi: 10.1126/science.1128845.

The residence time of water varies depending on the “pool” in which the water enters. For example, a molecule of water remains in the atmosphere for less than a week, in the ocean for several thousand years, in soil for a year at most (unless it travels to groundwater), or in groundwater from a matter of weeks/months to millions of years (e.g., Kralik, 2015; Wang, 2020). These are the average turnover times in various compartments of the hydrosphere.

Soil mediated fluxes

Fluxes of water across the land-atmosphere interface, whether pointed upward into the atmosphere or downward into the soil, are controlled or moderated by soil properties and processes, chief among them are the soil infiltration capacity and the unsaturated hydraulic conductivity function. Both, the infiltration rate and ability of unsaturated soil to transmit water depend on soil antecedent wetness. The relations between soil wetness and these important water transport functions has been studied for decades, often parameterized and expressed using closed-form expressions (e.g., Brooks and Corey, 1964; van Genuchten, 1980) and infiltration parameters. Much like infiltration rates and downward movement of water determined by soil wetness, also surface evaporation and plant water uptake are critically dependent on the ability of soil to support capillary flow to surface evaporation, or supply the transpiration needs of plants via root water uptake. This nonlinear dependency also highlights the intrinsic ability of soil to limit water losses to the atmosphere, when hydraulic conductivity becomes limiting. This is particularly important in arid regions in which the atmospheric demand is very high (low relative humidity and high temperature and solar radiation); in effect, soil limits evaporative losses, sheltering soil water that can support sparse vegetation (Or et al., 2013).

The control exerted by soil surfaces on infiltration affect groundwater recharge and runoff generation, and subsequent lateral water flow toward streams or other surface water bodies (Maxwell and Kollet, 2008; Taylor et al., 2012). As the spatial scale increases toward global, soil properties and soil structure, particularly the presence of macropores, still exerts influence on the partitioning of precipitation into surface runoff and infiltration in Earth system models (Fatichi et al., 2020), one of the few practical ways to simulate water balance at the global scale.

Notwithstanding the importance of soil water fluxes for life itself, changes in soil water storage play a similarly important role in climate regulation. The ingredients and manifestations of the “soil moisture memory” in the climate literature (Koster and Suarez, 2001; Seneviratne et al., 2006; Seneviratne and Koster, 2012) are rooted in soil water holding capacity, and determination of plant available soil water, with specific considerations of soil type and depth and vegetation type and its rooting depths. The soil water holding capacity extends the effects of rainfall events and modulates temperature extremes. These are largely due to plant root water uptake and transpiration at plant controlled rates during intervals between rainfall events. Soil water storage and response of plants from different biomes (e.g., forests vs. grasslands) have been shown to play an important role in the occurrence (or suppression) of

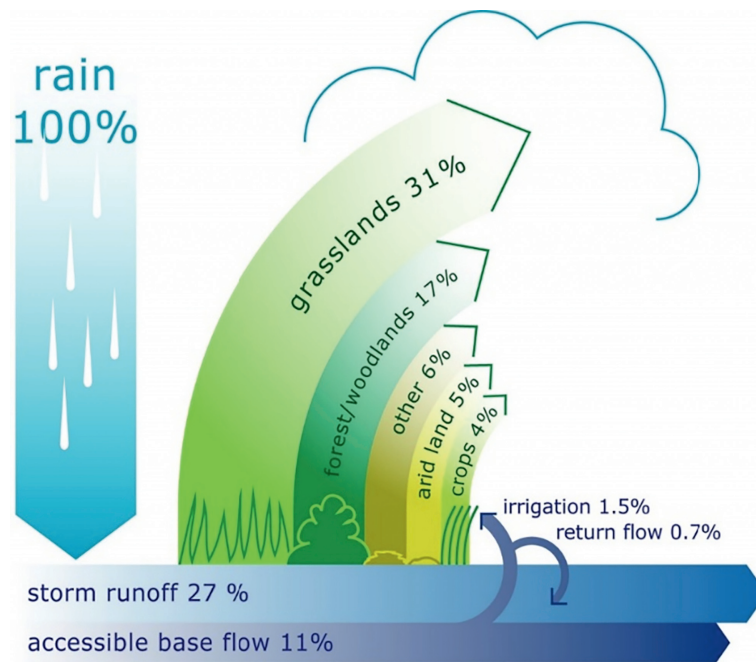


Fig. 4 The partitioning of terrestrial precipitation into different components of “blue” and “green” water based on land cover and use. The irrigation and extraction of “blue” water is from groundwater only, and estimated at around 1000 km³/yr (Margat and van der Gun, 2013). Image modified from International Soil Resources Information Center (ISRIC) (2022) <https://www.isric.org/utilise/global-issues/water> (Accessed April 2022).

heat waves (Miralles et al., 2019; Aminzadeh et al., 2021). Some have advocated mitigating the onset of heatwave by expanding the use of irrigation (Thiery et al., 2020), although economic considerations and availability of water sources for such proposed climate engineering remain debatable.

Anthropogenic effects on the global water cycle—Blue and Green water

The combined effects of growing global population with its water needs and changes in climate patterns (rainfall, extreme events, and evaporation) have altered components of the natural water balance. These changes are exacerbated by massive diversion of rivers and withdrawal of surface and groundwater resources. For example, it is estimated that human “blue” water withdrawals for agricultural irrigation alone are about 2500 km³/yr globally, representing around 70% of human “blue” water use, with 50–70% of this water consumed through irrigated crop ET (Rost et al., 2008). However, not all agriculture depends on “blue” water (irrigation); nearly 80% of global croplands are rainfed, using so-called “green” water for production of 60–70% of the world’s food (5000 km³/yr). Together with the “green” water required to support pastures and grazing (20,000 km³/yr; Rockström et al., 1999), nearly 27,500 km³/yr of water support food production for humans (combining “blue” and “green” water). This amount represents 25–30% of the terrestrial annual rainfall and highlights the magnitude of human involvement in the global water cycle (Fig. 4) (International Soil Resources Information Center (ISRIC), 2022). It also emphasizes the central role of soil in these processes, as all “green” water (by definition) is extracted from stored soil water. The unequal spatial distribution of “green” and “blue” water extraction patterns reflects climatic (rainfall for storage or irrigation primarily in arid and Mediterranean or seasonal climates) and socio-economic conditions (population density, infrastructure for “blue” water extraction and irrigation systems).

Scales of measurement and modeling of soil mediated hydrologic processes

As implied above, the scales of measurement and representation of soil hydrologic processes are intimately linked with the scales of the questions being addressed. An ongoing challenge for the scientific community is matching measurements at a certain scale with theoretical frameworks and their parameters often conceived and formulated at other scales (Dooge, 1997; Or, 2019). Studies for the water balance, for example, can be applied to a hierarchy of scales ranging from a single soil profile to a field, a watershed (a few hectares to 1000s of km²), a climatic region (from 100,000s to millions of km²) or to a global scale, assuming that the experiments are appropriately conceived and implemented. These scales were defined in part because they are tactile and comprehensible to the general public and to scientists who study them, not because the physical processes are bounded by them (even though they might be).

However, other means of assessing hydrologic responses across different spatial scales are emerging with advances in Earth observing and modeling capabilities. [Fatichi et al. \(2020\)](#) recently conducted numerical experiments on the influence of soil structure on water balance at global and ecosystem scales (e.g., grasslands, shrublands, mixed forests, etc.), rather than remaining tied to strict spatial scales listed above. The motivation is that biological processes (vegetation type, root growth, creation of biopores) that influence soil structure tend to group according to climatic environmental conditions (i.e., biomes); hence, considering the processes for each biome, and by classifying land cover accordingly (using, for example, remote sensing), researchers can apply knowledge gained at point- or watershed-scales, to larger land areas with similar ecosystem characteristics, or potentially at smaller scales when different approaches are combined (e.g., [Fang et al., 2021](#)).

The combination of Earth observation platforms and advanced analytics, like machine learning and soil pedotransfer functions, can reduce errors when parameter values are measured at one scale and applied to another scale, with the goal of predicting water balance components and closing mass balance. Conceptual and numerical models that are used to represent or simulate hydrologic processes at these different scales are becoming vital tools for researchers and decision makers who are seeking to extrapolate knowledge either up or down in scale, and into the future. Changes in water balance can then be predicted as a function of changes to climate or land use, or both, which will become increasingly common.

Key soil properties

Aspects of global water budgets that are moderated by soils can be attributed to a few key physical properties, including soil texture, soil bulk density and organic matter content. Soil texture refers to the relative percentages of sand, silt and clay fractions of mineral soil, as well as those fractions larger than 2 mm in diameters, often referred to as gravel or stones. The textural components generally control the amount of water held in soils against gravity; here finer-textured soils retain more water than coarser-textured soils, especially as the soils become drier. The means that water in soils with higher clay and silt content is more available to plant roots for transpiration and potentially for soil evaporation as well. In coarse-textured, sand-rich soils, water drains more readily (downward toward groundwater or laterally toward springs) and may no longer be available to exchange back into the atmosphere through ET processes. Soil structure refers to the arrangement of soil particles into aggregates (or peds) and the formation of soil horizons after pedogenic processes lead to concentrations of salts or clay in discrete layers, or both. Soil structures can influence the water balance in soils by formation of larger pores, through which more water will flow (and at faster rates), or potentially formation of lower conductivity layers where flow rates are lower, and where water can temporarily pond, creating localized, saturated zones. These phenomena are scale-dependent, are more likely observed at local- or ecosystem-scales, but less at global scales ([Fatichi et al., 2020](#)).

Some researchers have reported that soil organic matter, with its small size and high surface area, can significantly increase the water holding capacity of soil (e.g., [Bouyoucos, 1939](#); [Ankenbauer and Loheide, 2017](#)), improving water access for plants, increasing soil evaporation, and reducing deep recharge. Other studies have indicated that water holding capacity is influenced to a limited extent ([Minasny and McBratney, 2018](#)). These studies, while somewhat conflicting, point to a need to better understand how soil carbon manifests in the water budget and then, importantly, how to apply that knowledge at scales needed to address questions. Recently, [Bagnall et al. \(2022\)](#) highlighted the benefit of incorporating organic carbon levels into pedotransfer functions that could yield soil hydraulic properties (e.g., soil water retention and hydraulic conductivity functions) which govern soil water budgets. Another approach for incorporating emergent properties (e.g., soil structure is the so-called geo-transfer functions (GTFs)) ([Gupta et al., 2022a, 2022b](#)). GTFs expand the range of covariates considered for estimating soil hydraulic properties to include climatic, land cover, and geomorphic features, which all play critical roles in soil formation, and hence near surface water balance. The expansion of covariates beyond soil-dependent variables used for classical PTF have proven transformative in capturing nuances such as the type of clay in tropical regions ([Lehmann et al., 2021](#)) and aspects of soil structure that are correlated with precipitation, temperature and vegetation cover.

Measurement and modeling of soil water at continental to global scale

Measurements used to assess water balance are more often applied at the smaller end of the spatial scales, generally from point to regional scale, with most settling on watershed scales to capture the irregular boundaries of watersheds that contain surface water within it. At this scale, even for larger watersheds, uncertainty and error in surface water measurements are reduced, in part because the headwaters are fully contained in the region of interest and outfall is a point in space. Starting around 2015, various space agencies have designed, launched and operated satellite systems capable of estimating soil water content, from which this water pool can be quantified. Sensors on these satellite systems use radar and the dielectric response of ground surface as the surrogate indicator for soil water content. Two examples of systems used by the global Earth science community include the Soil Moisture Active Passive (SMAP) system ([Entekhabi et al., 2010](#)), operated by the Jet Propulsion Laboratory at NASA, and the Soil Moisture and Ocean Salinity (SMOS) system ([Kerr et al., 2010](#)), operated by the European Space Agency. Other systems producing soil water data products include ASCAT ([Bartalis et al., 2007](#)) and AMSR-2 ([Kim et al., 2015](#)), and when combined, several systems are able to provide water content at higher-than-designed spatiotemporal resolution, thereby coming closer to matching capabilities of these systems with the needs of the scientific communities ([Peng et al., 2021](#)).

In addition to new technology developed for interrogating surface properties at the global scale, the derivative benefits from digital innovation are equally important. For example, with the combination of more point-scale measurement systems being deployed globally, these disparate datasets can be harmonized and integrated into global or continental scale databases of water content (e.g., Dorigo et al., 2011, 2013), and soil hydraulic properties [e.g., POLARIS (Chaney et al., 2019); SoilGrids250 (Hengl et al., 2017)]. The advent of high performance computing and advanced analytics is providing the opportunity for the scientific community to estimate pools of soil water, surface water and ocean water at increasingly higher spatiotemporal resolutions, now approaching 1 km and lower (e.g., Fang et al., 2021). These new products potentially overcome technical issues in forecasting soil water at both the localized, single-grid and the regional river basin scales.

The increased spatial resolution of rainfall and soil information permits derivation of new soil-water related maps of plant available soil water or the average moisture conditions experienced by soil biota over climatic time scales. For example, a map of the so-called “climatic water content” (Bickel and Or, 2020) (Fig. 5) has been instrumental in explaining drivers for biome-scale soil bacterial diversity (Bickel and Or, 2020), enabled delineation of global earthworm distribution (whose activity is critically dependent on soil hydromechanical conditions) (Ruiz et al., 2021; see also chapter on Bioturbation). Other global climatic soil moisture data bases are now available, such as reported in Sugmin and Orth (2021) and Muñoz-Sabater et al. (2021) (ERA5-Land). The synthesis of spatially (and temporally) resolved rainfall record with soil information and hydraulic parameters provides new insights into how much plant available soil water can be stored in different soils. While the median is about 160 mm water per 1000 mm soil profile (about 22,000 km³ globally), this quantity varies widely with soil type (primarily texture and structure). Moreover, under certain climatic conditions the profile is rarely replenished to its potential, especially in hyper-arid regions, whereas in other, more humid climates the soil profile may remain close to field capacity the entire year.

Conclusions and outlook

Soil processes exert a dominant influence on the water cycle, outside of precipitation itself. While the influence is recognized at the local and watershed scales, new research is highlighting that soil processes also influence the water cycle at continental and global scales as well, though the scientific community has only begun to study the ramifications fully. Fatichi et al. (2020) showed through models that soil structure in the form of biopores (or other large preferential flow pathways) can influence surface runoff and deep percolation, hence reducing the role of ET, primarily in humid and vegetated regions. Following Fatichi’s work, Bonetti et al. (2021) proposed a framework for incorporating these structures into soil pedotransfer functions that can more easily be used in models at any chosen scale. Tashie et al. (2021) used machine learning and streamflow recession curves to constrain soil hydraulic properties used in models at the catchment or larger scales, with a goal of eventually reaching continental scale. Cosh et al. (2021), recognizing the value of point-scale monitoring, but at the continental scale, advocate for creating a coordinated network of networks, from which the scientific community and public could gain greater value. These efforts are showing the need to understand how water is partitioned into components of infiltration and surface runoff, and then again into groundwater recharge, plant transpiration and soil evaporation, all of which effect groundwater sustainability and ecosystem functions. Over the last 3–4 decades, the development, deployment and harmonizing of advanced remote sensing technology, point-scale sensing equipment, and the computational strength to process this information have significantly increased our overall understanding of the water cycle.

The rapid advances in technology and the changing community perspective toward asking globally-relevant questions provides new opportunities for placing the influence of soils on the water cycle in proper context. For example, several overriding questions that relate to the water cycle and the influences of soil are top-of-mind in the water and soil community. These include: How can we

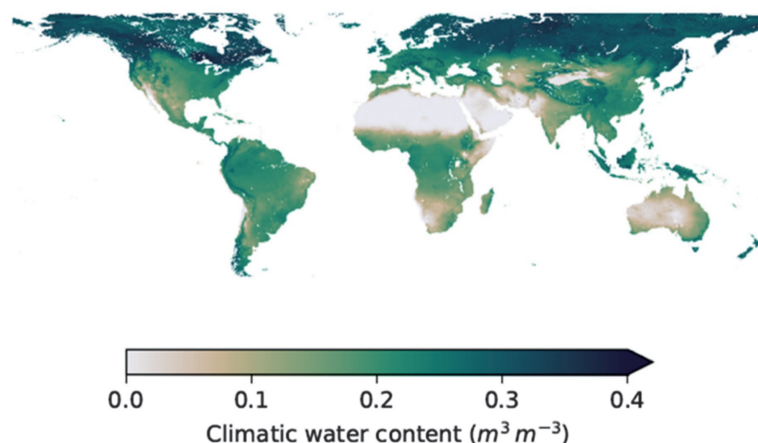


Fig. 5 Climatic water content representing the estimated soil water content averaged over 36 years of satellite and reanalyzed precipitation record, considering soil type (field capacity) and rainfall amounts and interval (internal drainage and evaporation). See Bickel S and Or D (2020) Soil bacterial diversity mediated by microscale aqueous-phase processes across biomes. *Nature Communications* 11: 1–9 for detail.

correctly understand ET rates and processes and the influence of soil properties and processes on them at spatio-temporal scales of interest, especially at the continental and global scales? To what extent can we track the plant-available soil water as a means of understanding impacts of climate change on global vegetation? And, how will ecosystem functions and services change in different types of land covers in very wet and dry conditions (e.g., after floods or droughts) that are predicted from weather extremes. Answers to these questions require ongoing efforts and research.

Given the availability of monitoring systems for groundwater levels (but not always for withdrawal); surface water levels in rivers, lakes and oceans; and precipitation, it should be clear that uncertainty in, and availability of, soil water content and storage at the continental and global scales are still significant areas requiring improvement. So, what areas of research and understanding are needed now? Two areas are probably being pursued the most. First, remote sensing methods have continued to advance in sophistication, providing data at higher and higher resolution, with more frequent return periods, and shorter latencies between overpass and data availability. Fang et al. (2021), for example, report on a workflow that provides water content data across the continental U.S. at a spatial resolution of 400 m, though it does require combining output from several different sources (satellites and point-scale data) to validate. Future technological development of newer satellites (e.g., the NASA-ISRO SAR mission, satellite is scheduled to launch in 2023–2024) (National Aeronautical and Atmospheric Administration (NASA), 2022a, 2022b) would monitor Earth's surface at around 200 m resolution. Coupled with harmonized soil water content data, collected with accurate point-scale sensors at regional scales, machine learning approaches can be more effectively used to quickly and more accurately generate gridded water content data. Future technological development coupled with advanced analytical and computational approaches can improve our understanding of how soil processes affect the water balance at global scales. Second, the development of new maps of soil and landscape properties, developed using different types of transfer functions with or without machine learning, are needed. For example, if climate change leads to persistent drought conditions and reduced availability of soil water for plants, wholesale changes in ecosystem characteristics and services could occur, but our understanding of the volume of plant available water in soil is difficult—in part because of the large scale of measurement and in part because the soil characteristics are not parameterized in a way that facilitates analyses. Gupta et al. (2022a, 2022b) recently reported on a framework for mapping plant available soil water at the global scale, making better use of available datasets of soil parameters, and showed that they could quantify soil water residing between field capacity and wilting point. From this information, and with knowledge of local and regional climate, more accurate estimates of ET rates can be ascertained (Bickel and Or, 2020, 2021).

References

- Aminzadeh M, Roderick ML, and Or D (2021) Using the complementary relationship between actual and potential evaporation to diagnose the onset of heatwaves. *Water Resources Research* 57: e2020WR029156.
- Ankenbauer KJ and Loheide SP (2017) The effects of soil organic matter on soil water retention and plant water use in a meadow of the Sierra Nevada, CA. *Hydrological Processes* 31: 891–901. <https://doi.org/10.1002/hyp.11070>.
- Bagnall DK, Morgan CLS, Cope M, Bean GM, Cappellazzi S, Greub K, Liptzin D, Norris CL, Rieke E, Tracy P, Aberle E, Ashworth A, Bañuelos Tavarez O, Bary A, Baumhardt RL, Borbón Gracia A, Brainard D, Brennan J, Briones Reyes D, et al. (2022) Carbon-sensitive pedotransfer functions for plant available water. *Soil Science Society of America Journal* 1–18. <https://doi.org/10.1002/saj2.20395>.
- Bar-On YN, Phillips R, and Milo R (2018) The biomass distribution on Earth. *Proceedings of the National Academy of Sciences* 115(25): 6506–6511. <https://doi.org/10.1073/pnas.1711842115>.
- Bartalis Z, Wagner W, Naeimi V, Hasenauer S, Scipal K, Bonekamp H, Figa J, and Anderson C (2007) Initial soil moisture retrievals from the METOP-A Advanced Scatterometer (ASCAT). *Geophysical Research Letters* 34: L20401.
- Bickel S and Or D (2020) Soil bacterial diversity mediated by microscale aqueous-phase processes across biomes. *Nature Communications* 11: 1–9.
- Bickel S and Or D (2021) Dataset of global climatic soil water contents and consecutive dry days. *Zenodo*. <https://doi.org/10.5281/zenodo.5078258>.
- Bonetti S, Wei Z, and Or D (2021) A framework for quantifying hydrologic effects of soil structure across scales. *Communications Earth and Environment*. <https://doi.org/10.1038/s43247-021-00180-0>.
- Bouyoucos GJ (1939) Effect of organic matter on the water holding capacity and the wilting points of mineral soils. *Soil Science Journal* 47: 377–383. <https://doi.org/10.1097/00010694-193905000-00005>.
- Brooks RH and Corey AT (1964) *Hydraulic Properties of Porous Media*. Hydrology Paper No 3 Fort Collins, CO: Civil Engineering Department, Colorado State University.
- Chaney NW, Minasny B, Herman JD, Nauman TW, Brungard C, Morgan CLS, et al. (2019) POLARIS soil properties: 30-m probabilistic maps of soil properties over the contiguous United States. *Water Resources Research* 55: 2916–2938. <https://doi.org/10.1029/2018WR022797>.
- Cosh MH, Caldwell TG, Baker CB, Bolten JD, Edwards N, Goble P, Hofman H, Ochner TE, Quiring S, Schalk C, Skumanich M, Svoboda M, and Woloszyn ME (2021) Developing a strategy for the National Coordinated Soil Moisture Monitoring Network. *Vadose Zone Journal* 20: e20139. <https://doi.org/10.1002/vzj2.20139>.
- Dooge JC (1997) Scale problems in hydrology. In: Buras N (ed.) *Reflections on Hydrology: Science and Practice*. Wiley. <https://doi.org/10.1029/SP048p0084>.
- Dorigo WA, Wagner W, Hohensinn R, Hahn S, et al. (2011) The international soil moisture network: A data hosting facility for global in situ soil moisture measurements. *Hydrology and Earth System Sciences*. <https://doi.org/10.5194/hess-15-1675-2011>.
- Dorigo WA, Xaver A, Vreugdenhil M, Gruber A, et al. (2013) Global automated quality control of in situ soil moisture data from the international soil moisture network. *Vadose Zone Journal*. <https://doi.org/10.2136/vzj2012.0097>.
- Entekhabi D, Njoku EG, O'Neill PE, Kellogg KH, Crow WT, Edelstein WN, Entin JK, Goodman SD, Jackson TJ, and Johnson J (2010) The soil moisture active passive (SMAP) mission. *Proceedings of the IEEE* 98: 704–716.
- European Centre for Medium-Range Weather Forecasts (CMWRF) (2022) <https://www.ecmwf.int/> (Accessed April, 2022).
- Fang B, Lakshmi V, Cosh M, and Hain C (2021) Very high spatial resolution downscaled SMAP radiometer soil moisture in the CONUS using VIIRS/MODIS data. *IEEE Journal Selected Topics in Applied Earth Observations and Remote Sensing*. <https://doi.org/10.1109/JSTARS.2021.3076026>.
- Faticchi S, Or D, Walko R, Vereecken H, Young MH, Ghezzehei T, Hengl T, Kollet S, Agam N, and Avissar R (2020) Soil structure—An important omission in earth system models. *Nature Communications* 11: 522. <https://doi.org/10.1038/s41467-020-14411-z>.
- Gupta S, Papritz A, Lehman P, Hengl T, Bonetti S, and Or D (2022a) Global mapping of soil water characteristics parameters—Fusing curated data with machine learning and environmental covariates. *Remote Sensing*. <https://doi.org/10.3390/rs14081947>.

- Gupta S, Lehmann P, Bickel S, Bonetti S, and Or D (2022b) Global mapping of plant-available soil water (PASW)—Revisiting field capacity and local climate effects. *Journal of Advances in Modeling Earth Systems*. Submitted.
- Hengl T, Mendes de Jesus J, Heuvelink GBM, Ruiperez GM, Kilibarda M, Blagotić A, et al. (2017) SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One* 12(2), e0169748. <https://doi.org/10.1371/journal.pone.0169748>.
- International Soil Resources Information Center (ISRIC) (2022) <https://www.isric.org/utilise/global-issues/water> (Accessed April 2022).
- Karger DN, et al. (2021) Global daily 1 km land surface precipitation based on cloud cover-informed downscaling. *Scientific Data*. <https://doi.org/10.1038/s41597-021-01084-6>.
- Kerr YH, Waldteufel P, Wigneron J-P, Delwart S, Cabot FO, Boutin J, Escorihuela M-J, Font J, Reul N, and Gruhier C (2010) The SMOS mission: New tool for monitoring key elements of the global water cycle. *Proceedings of the IEEE* 98: 666–687.
- Kim S, Liu YY, Johnson FM, Parinussa RM, and Sharma A (2015) A global comparison of alternate AMSR2 soil moisture products: Why do they differ? *Remote Sensing of Environment* 161: 43–62.
- Koster RD and Suarez MJ (2001) Soil moisture memory in climate models. *Journal of Hydrometeorology* 2: 558–570.
- Kralik M (2015) How to estimate mean residence times of groundwater. *Procedia Earth and Planetary Sciences*. <https://doi.org/10.1016/j.proeps.2015.07.070>.
- Lehmann P, Leshchinsky B, Gupta S, Mirus B, Bickel S, Lu N, and Or D (2021) Clays are not created equal: How clay mineral type affects soil parameterization. *Geophysical Research Letters*. <https://doi.org/10.1029/2021GL095311>.
- Margat J and van der Gun J (2013) *Groundwater Around the World: A Geographic Synopsis*. Boca Raton, FL: CRC Press.
- Maxwell RM and Kollet SJ (2008) Interdependence of groundwater dynamics and land-energy feedbacks under climate change. *Nature Geoscience*. <https://doi.org/10.1038/ngeo315>.
- Minasny B and McBratney AB (2018) Limited effect of organic matter on soil available water capacity. *European Journal of Soil Science* 69: 39–47. <https://doi.org/10.1111/ejss.12475>.
- Miralles DG, Gentile P, Seneviratne SI, and Teuling AJ (2019) Land-atmospheric feedbacks during droughts and heatwaves: State of the science and current challenges. *Annals of the New York Academy of Sciences* 1436(1): 19.
- Muñoz-Sabater J, Dutra E, Agustí-Panareda A, Albergel C, Arduini G, Balsamo G, Boussetta S, Choulga M, Harrigan S, Hersbach H, Martens B, Miralles DG, Piles M, Rodríguez-Fernández NJ, Zsoter E, Buontempo C, and Thépaut J-N (2021) ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data* 13: 4349–4383. <https://doi.org/10.5194/essd-13-4349-2021>.
- National Aeronautical and Atmospheric Administration (NASA) (2022a) https://www.nasa.gov/mission_pages/GPM/main/index.html (Accessed April 2022).
- National Aeronautical and Atmospheric Administration (NASA) (2022b) https://nisar.jpl.nasa.gov/system/documents/files/15_NISARApplications_SoilMoisture1.pdf (Accessed May 2022).
- Oki T and Kanae S (2006) Global hydrological cycles and world water resources. *Science* 313: 5790. <https://doi.org/10.1126/science.1128845>.
- OpenGeoHub (2022) <https://opengeohub.org/about-openlandmap/> (Accessed April 2022).
- Or D (2019) The tyranny of small scales—On representing soil processes in global land surface models. *Water Resources Research* 55. <https://doi.org/10.1029/2019WR024846>.
- Or D, Lehmann P, Shahraneen E, and Shokri N (2013) Advances in soil evaporation physics—A review. *Vadose Zone Journal*. <https://doi.org/10.2136/vzj2012.0163>.
- Peng J, Albergel C, Balenzano A, Brocca L, et al. (2021) A roadmap for high-resolution satellite soil moisture applications—confronting product characteristics with user requirements. *Remote Sensing of Environment*. <https://doi.org/10.1016/j.rse.2020.112162>.
- Rockström J, Gordon JL, Folke C, Falkenmark M, and Engwall M (1999) Linkages among water vapor flows, food production, and terrestrial ecosystem services. *Conservation Ecology* 3(2).
- Rost S, Gerten D, Bondeau A, Lucht W, Rohwer J, and Schaphoff S (2008) Agricultural green and blue water consumption and its influence on the global water system. *Water Resources Research* 44: W09405. <https://doi.org/10.1029/2007WR006331>.
- Ruiz SA, Bickel S, and Or D (2021) Global earthworm distribution and activity windows based on soil hydromechanical constraints. *Communications Biology* 4: 612. <https://doi.org/10.1038/s42003-021-02139-5>.
- Seneviratne SI and Koster RD (2012) A revised framework for analyzing soil moisture memory in climate data: Derivation and interpretation. *Journal of Hydrometeorology* 13(1): 404–412.
- Seneviratne SI, Koster RD, Guo Z, Dirmeyer PA, Kowalczyk E, Lawrence D, Liu P, Lu C-H, Mocko D, Oleson KW, and Verseghy D (2006) Soil moisture memory in AGCM simulations: Analysis Impact of the soil hydrology of global land-atmosphere coupling experiment (GLACE) data. *Journal of Hydrometeorology* 7: 1090–1112.
- Sugmin O and Orth R (2021) Global soil moisture data derived through machine learning trained with in-situ measurements. *Scientific Data* 8(1): 170. <https://doi.org/10.1038/s41597-021-00964-1>.
- Tashie A, Pavelsky T, Band L, and Topp S (2021) Watershed-scale effective hydraulic properties of the continental United States. *Journal of Advances in Modeling Earth Systems* 13. <https://doi.org/10.1029/2020MS002440>. e2020MS002440.
- Taylor RG, Scanlon BR, Doll P, Rodell M, et al. (2012) Ground water and climate change. *Nature Climate Change*. <https://doi.org/10.1038/nclimate1744>.
- Thiery W, Visser AJ, Fischer EM, Hauser M, Hirsch AL, Lawrence DM, Lejeune Q, Davin EL, and Seneviratne SI (2020) Warming of hot extremes alleviated by expanding irrigation. *Nature Communications* 11(1): 1–7.
- van Genuchten MT (1980) A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal* 44(5): 892–898.
- Wang HF (2020) *Groundwater Storage in Confined Aquifers*. Guelph, Canada: The Groundwater Project. ISBN: 978-1-7770541-7-5.
- World Atlas of Desertification (WAD) (2022) *Joint Research Center, European Commission*. <https://wad.jrc.ec.europa.eu/patternsaridity> (Accessed May 2022).

Water evaporation from bare soil

Peter Lehmann, Physics of Soils and Terrestrial Ecosystems, Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland

© 2023 Elsevier Ltd. All rights reserved.

Introduction	490
Evaporation process in the hydrologic cycle and surface energy balance	491
Evaporation rate dynamics	492
First and second stage of evaporation	492
Evaporation planes and drying fronts	493
Mechanisms sustaining stage-1 evaporation	493
The characteristic length for stage-1 evaporation	493
Continuity of capillary flow paths	493
Capillary forces vs. gravity	494
Capillary vs. viscous flow	494
Characteristic lengths of evaporation from shallow water tables	495
Characteristic length and stage-2 evaporation	495
Stage-2 evaporation and diffusion across a dry soil layer	495
Effects of structure formation and land use on evaporation stages	496
Quantification of the evaporation process at different scales	497
Pore scale considerations	497
Imaging at the column scale	497
Processes at the field scale	497
From catchments to global analyses	497
Summary	498
References	499

Abstract

Evaporation of water emitted from soil surfaces is a key process of the hydrologic cycle and the surface energy balance. It can be limited by the solar energy to vaporize water, the amount of precipitation, the soil hydraulic properties and surface characteristics (albedo, roughness, and presence of residues). Evaporation rates are large for wet soil surfaces with high atmospheric demand (controlled by energy, vapor pressure deficit, and wind velocity) but become very small when the vapor is emitted below the surface and must diffuse through a dry soil layer.

Key points

- Evaporation links the hydrologic cycle with the surface energy balance
- Evaporation stages can be classified by comparison with potential evaporation rate and by the position of the evaporation plane
- In stage-1 evaporation, capillary flow pathways connect the drying soil surface with the wet subsurface
- The length of capillary flow paths depends on soil hydraulic properties and atmospheric conditions
- Soil structures and heterogeneities have large effects on capillary flow paths and evaporation rate
- The quantification of evaporation rate can be measured with various techniques at different scales

Introduction

Evaporation of water from bare soil is an important pathway in the hydrologic cycle to move water in its liquid state back to the atmosphere as vapor. Because the corresponding phase change from liquid to gas requires a large amount of energy, the evaporation process is a key factor in both the hydrologic cycle and Earth surface energy balance. Evaporation is a relatively slow process (compared to soil water flow during heavy rainfall and ground water fluctuations) but it is very dynamic, with evaporation rates changing with time as controlled by soil properties and atmospheric conditions (see examples in Fig. 1). The key to understand evaporation rate dynamics is the quantification of water distribution at the surface and down the soil profile. In terrestrial systems, vapor is emitted from soil (soil evaporation) and from plant leaves (transpiration). After rainfall events, vapor can also evaporate from water intercepted by plant cover. The combination of these processes is denoted as evapotranspiration. In this chapter, the focus is on evaporation from the soil surface.

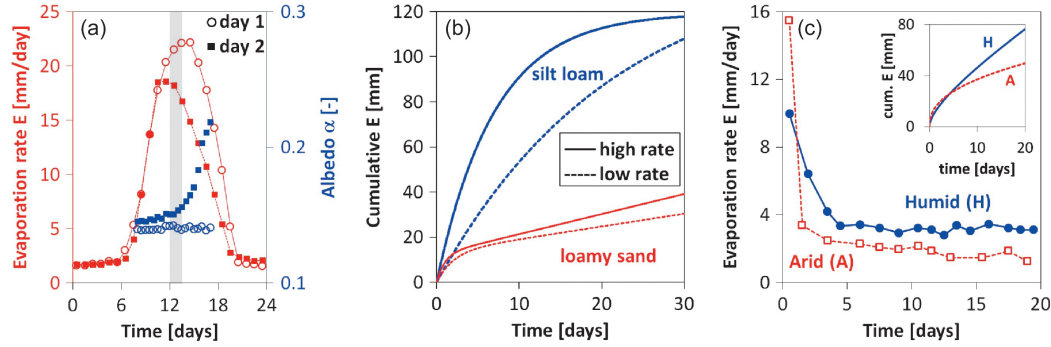


Fig. 1 Examples of evaporation dynamics from natural soils in field (a) and lab experiments (b and c). The evaporation rate depends on the atmospheric conditions and the soil properties. (a) The evaporation rate E (red) from a clay loam (Idso et al., 1974) is reduced during the cool night and for dry surfaces, indicated by increasing surface albedo α (shown in blue); the gray bar marks the concurrent change of E and α during the second day. (b) The response of evaporation losses on hot and windy atmospheric conditions (large rate) depends on the soil type (Jalota and Prihar, 1986) with greater fluxes for fine textured silt loam soil. (c) As already shown by Buckingham (1907), evaporation losses do not always increase with the atmospheric demand (the soil surface and air flow were warmer under ‘arid’ compared to ‘humid’ conditions): for conditions with a shallow water table, the unsaturated flow is limiting in case of high demands and a dry surface layer evolves that greatly resists vapor transport, reducing the evaporation rate.

Evaporation process in the hydrologic cycle and surface energy balance

Evaporation is relevant to water transport from the soil to the atmosphere and the consumption of solar energy. It is an important process of the hydrologic cycle, linking soil systems with the atmosphere. Evaporation E (or evapotranspiration ET in case of plant cover) is a component of the water balance, typically formulated for a soil layer of depth L . The water balance describes the change of water storage ΔS over time interval Δt as balance of fluxes as stated in Eq. (1):

$$\Delta S = P - R - E - D \quad (1)$$

with precipitation P , surface runoff R , evaporation E and drainage to deeper soil layers D . The fluxes are expressed as water volume per soil cross sectional area and time and are typically expressed as millimeters per day for a daily water balance. A change of storage ΔS corresponds to a change of water content $\Delta \theta$ equal to $\Delta S/L$. For example, the evaporation of 10 mm water (corresponding to 10 L of water evaporated from one square meter cross section) during a sunny day from a soil layer of 0.5 m thickness corresponds to a decrease of water content of 2% (when all other fluxes in Eq. (1) are zero). At global scale, evaporation from soil accounts annually for 180 mm of 940 mm precipitation (about 20%, Oki et al., 2005).

The amount of energy required to evaporate 1 kg of water is denoted as latent heat of evaporation ℓ and is about $2.45 \cdot 10^6 \text{ J kg}^{-1}$. Applied to the previous example of 10 mm evaporation per day (equal to 10 kg m^{-2}), energy of about 30 W (Joule per second) is required per square meter. This energy comes from the sun with 325 W m^{-2} arriving at the atmosphere and 136 W m^{-2} absorbed at the surface (averaged over day and night and geographic latitudes of land cover; Wild et al., 2015). For the given example, 10% of the solar energy is used for evaporation (i.e., $30/325$). This amount agrees with global averages for terrestrial surfaces (38 W m^{-2} for evapotranspiration). Formally, the surface energy balance equals:

$$R_N = R_S(1 - \alpha) + R_L^{\downarrow} - R_L^{\uparrow} = \ell \cdot E + H + G \quad (2)$$

with net radiation R_N , short wave radiation R_S , albedo α defining the amount of reflected shortwave radiation (depending, for example, on color and surface wetness of the soil, see Fig. 1a), and incoming and outgoing longwave radiation R_L . The net radiation R_N is partitioned into the latent heat flux of evaporation $\ell \cdot E$, the sensible heat flux H to the atmosphere, and the heat flux to the soil G . Eq. (1) and (2) show that the evaporation process links the hydrologic cycle with the surface energy balance (term E is in both equations). The evaporation rate E does thus depend on the rainfall amount P and the incoming energy R_N . Accordingly, the evaporation of a system or a climatic region can be characterized with respect to the climatic factor limiting the evaporation (P in water limited or R_N in energy limited systems).

The strong coupling between the hydrologic cycle and surface energy balance makes soil evaporation sensitive to climate change and change in land use. Higher temperatures increase the evapotranspiration rate. A change in land use, such as forest clearing, affects both soil properties and energy balance (change in surface reflectance and shading). Based on analysis of satellite data for the period between 1980 and 2017, Liu et al. (2021) found a significant increase of evapotranspiration in 42% of the land surface area (and significant decrease in less than 5% of area). The main driver for this increase was a change in precipitation and potential evapotranspiration. Projections of evaporation to the future with land surface models predict an increase of evapotranspiration from 550 to 580 mm per year (5.6% increase) or 560 mm per year (2.4%) depending on the emission scenario (Pan et al., 2015). For certain vegetation covers, the predicted increase could be much greater (between 9% and 24% for different scenarios). However, the response of soil evaporation to a change in atmospheric conditions is highly non-linear, as already indicated in Fig. 1 and the predicted values depend on the level of process details implemented in the land surface models. In this chapter, these details of the evaporation process and the mechanisms controlling its dynamics are presented.

Evaporation rate dynamics

The evaporation rate cannot only be limited by atmospheric conditions but by the amount of water in the soil and at the soil surface (e.g., limitation occurs when water in deeper soil layers is not transported to the surface). To distinguish between limitations by soil properties and by atmospheric conditions, the actual evaporation rate (amount of vapor emitted from soil per surface area and time) must be compared to the potential evaporation rate (amount of water emitted from a free water surface under the same atmospheric conditions). A first evaporation stage occurs when the amount of soil water and the surface water content is not limiting, such that the actual evaporation rate is very close to the potential evaporation rate and depends on temperature (an indicator for available energy) and the vapor pressure deficit in the air. More specifically, the actual evaporation flux is determined by the vapor pressure difference (between the evaporating surface and the atmosphere) and a transport coefficient (gaseous diffusion coefficient across a thin viscous boundary layer or a turbulent eddy-diffusion coefficient at larger scale). The potential evaporation rate increases with temperature, vapor pressure difference between the surface and the atmosphere, and wind velocity. For conditions with no or only moderate wind, the evaporation rates from a free water surface or a wet soil surface are very similar (actual evaporation rate close to the potential evaporation rate). As long as the water molecules evaporating at the surface are replaced by liquid water flow from the soil interior, the surface remains wet, and the evaporation rate remains constant. Eventually, when the water supply from the soil to the surface becomes too slow, the evaporated water is not replaced, and the surface becomes very dry. The presence of a dry soil surface layer indicates that evaporation (phase change) occurs within the porous medium and the vapor molecules travel through the dry soil before they are emitted to the atmosphere. The vapor transport through a dry soil layer is small and decreases with increasing thickness of the dry soil layer. The evolution of the actual evaporation rate can thus often be classified into a first evaporation stage (stage-1) with rates close to the potential evaporation rates and a second evaporation stage with the evaporation rate decreasing with time. The classification into first and second evaporation stage is not always straightforward because the evaporation rate can drop below the potential rates for different reasons as discussed next.

First and second stage of evaporation

In contrast to a completely water covered surface, the water content of a wet soil cannot be larger than its porosity. The reduced water content compared to a free water surface (that is typically used to determine the potential evaporation rate) can affect the actual evaporation rate. At the pore scale, the vapor emitted from water filled pores is not just transported vertically across the boundary layer between soil and atmosphere but diffuses as well radially across the dry surface. This lateral spreading across a hemispherical region around the water filled pores can be described as additional transport resistance compared to the one-dimensional transport across a subviscous boundary layer (Schlünder, 1988). However, under most conditions, the distances between water filled pores are small compared to the thickness of the boundary layer (that depends on atmospheric conditions) and the additional resistance of the vapor shell is also small, compared to the one-dimensional transport across the boundary layer. The sizes of the vapor shells depend on the water content and the pore size distribution. For conditions with small vapor shells compared to the boundary layer, the actual evaporation rate can be as high as the potential rate. But for thin boundary layers (a function of wind velocity) and for large pores and large spacings between water filled pores, the vapor shell resistance becomes larger and reduces the evaporation rate considerably. For evaporation from coarse textured soils under conditions with high wind velocities, the evaporation rate drops from the very beginning and no stage with constant evaporation rate can be observed (see Fig. 2).

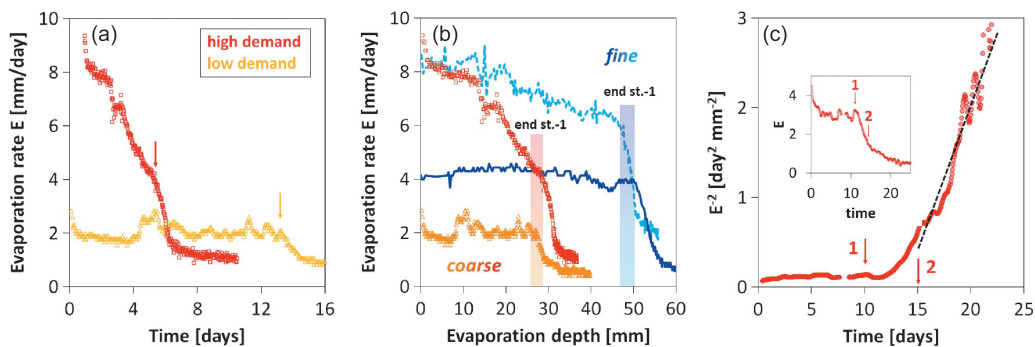


Fig. 2 Dynamics of evaporation rates and definition of evaporation stages. (a) The evaporation rate dynamics from a coarse sand (0.3–0.9 mm particles) changes with atmospheric conditions (high and low atmospheric demand). The arrows mark the end of stage-1 and the detachment of the evaporation plane from the surface. (b) Evaporation rates expressed as function of evaporation depth (cumulative mass loss per cross-section) for coarse and fine sand (0.1–0.6 mm). The mass loss at end of stage-1 evaporation (shaded rectangle) depends on soil texture but not on atmospheric conditions. (c) After end of stage-1, water recedes from the surface until a new evaporation front is formed at onset of stage-2 evaporation (the number 1 marks the end of stage-1, 2 the start of stage-2 evaporation). The start of stage-2 evaporation can be determined by plotting E^2 (evaporation rate E) as function of time (Brutsaert and Chen, 1995), here shown for a sand with 0.1–0.3 mm particles.

In the case of thin boundary layers or large pore sizes, the absence of a period of constant evaporation rate is caused by transport processes in the boundary layer. For fine textured media the evaporation rate can drop by a different process related to the limited water flow by capillarity from the wet subsoil to replace the water evaporated from the surface. In these fine textured soils, the pores are so small (a few microns in radius or less), that the water flow is smaller than the potential evaporation rate and the actual evaporation rate may never reach the potential evaporation rate.

There are thus three causes for decreasing evaporation rates and the lack of evaporation stages with an actual evaporation rate that equals the potential rate: (i) vapor diffusion across vapor shells in the boundary layer, (ii) liquid flow through small pores and (iii) vapor transport through a dry soil layer (see previous section). The classification of evaporation stages based on the evaporation rate alone (i.e., equal or smaller than the potential rate) does not differentiate between hydration stages of the surface (i.e., how much water is at the surface) that are important for ecohydrological processes. An alternative classification of the evaporation stages can thus be based on the hydration stage of the surface and the related question of where the phase change occurs (directly at the surface or within the porous medium). In the following, stage-1 is assigned to conditions when evaporation (phase change) occurs at the surface and stage-2 for conditions with completely dry surface layers and evaporation occurring exclusively inside the porous medium.

Evaporation planes and drying fronts

Using evaporation stage classification based on the presence of a dry soil surface layer (stage-2 evaporation); stage-1 evaporation contains all situations with water evaporating directly from the surface. To keep liquid water at the surface, evaporated water must be replaced by water flow. In the special case of a shallow ground water table, the water can be supplied sustainably from the water table without changing the water content in the profile over time. For other conditions without water supply from a reservoir, the water content in the soil layer decreases during evaporation. For a wet soil layer (water content close to saturation), the water evaporated from the surface is replaced by water from the soil below, reducing the water content in the profile. Extracting water from the soil profile corresponds to air invasion and thus increasing air content, forming an interface between the wet and the partially dry soil denoted as the drying front. During stage-1 evaporation, the drying front recedes deeper into the porous medium (replacing water evaporated at the surface) and the water content in the partially dry region above the front decreases.

In the second stage of evaporation, a second interface (in addition to the drying front) between the very dry soil surface layer (Heller, 1968) and the partially wet soil is formed. At this interface, the evaporation rate is greatest (compared to the levels below) and is denoted as secondary drying front, evaporation plane, or vaporization plane. During the second stage evaporation process, the evaporation plane recedes deeper into the soil, increasing the diffusive pathways from the evaporation plane to the surface. The increasing length that must be overcome by diffusion causes a steady decrease of the actual evaporation rate (Fig. 2c).

Mechanisms sustaining stage-1 evaporation

Sustained high evaporation rates related to the presence of liquid water at the surface requires a water supply from the subsurface to replace that evaporated from the surface. The water supply is driven by a capillary pressure difference between the wet subsurface and the drying surface. At the pore scale, the air-water interface (meniscus) is positioned at the soil surface in small pores but below the surface in larger pores that drained during the process. The capillary pressure difference between the receding water meniscus in large pores and the meniscus pinned at the surface in small pores drives water flow to the surface (Scherer, 1990). The flow capacity of the capillary pathways pumping water to the surface, defines how long the stage-1 evaporation can be sustained. This capacity depends on pore size distribution (capillary pressure difference between large and small pores), the initial water content (defining capillary pressure in largest water filled pores), and the influence of concurrent drainage (and water extraction by roots). To explain the mechanisms controlling the duration of stage-1 evaporation, a simplified situation of evaporation from a saturated homogeneous soil column without concurrent drainage is considered before the effects of more complex conditions in presence of soil structures are discussed at the end of this section. In such a simplified case, air invades the large pores and forms the drying front, separating the water saturated region from one that is partially air-filled. The capillary pressure at the drying front equals the air entry value of the porous medium. During evaporation with very negative water potential in the atmosphere, smaller pores are drained sequentially and make the surface capillary pressure more negative. The decrease of capillary pressure at the surface invokes a water flow from the drying front (where pressure stays constant at the air-entry value). This water flow can be limited by two conditions: the capillary flow paths become topologically disconnected at a critical water content or the hydraulic conductivity of the flow becomes too low to sustain the potential evaporation rate. In the next subsection, the assessment of the maximum length of these capillary pathways will be presented.

The characteristic length for stage-1 evaporation

Continuity of capillary flow paths

The length of the capillary pathways that can sustain the potential evaporation rate defines the maximum depth of the soil layer than can be drained by stage-1 evaporation. On one hand, this depends on a critical water content θ_{crit} and matric potential head h_{crit} when the capillary flow paths disconnect. These critical values can be deduced from measurements of the soil water characteristics curve (relationship between water content θ and matric potential head h). At the dry end when the water content

converges to its residual value, the water extraction by applied matric potential diminishes because it becomes controlled by the flow along adsorbed water films. As a simple estimate for the convergence toward the residual water content, the soil water retention curve can be linearized as shown in Fig. 3 for two different models of soil water characteristics with the critical pressure h_{crit} that equals to the pressure value when the linearized water saturation equals 0. The corresponding water content of the original soil water retention curve at this pressure h_{crit} is used as estimate for the critical water content θ_{crit} when capillary flow paths break-up. Note that this rule to assign a minimum matric potential value to sustain capillary flow differs from other approaches that assign much higher absolute values (150–1500 m or ∞). Conceptually, this criterion is related to the pressure and water content range that limits applicability of Richards equation with the underlying flow equation (Buckingham-Darcy) and hydraulic functions (capillary bundles with Hagen-Poiseuille flow).

Capillary forces vs. gravity

The pressure to attain the residual water content is a function of soil texture as manifested in the soil water characteristics curve. It is moderately negative (small absolute values) for coarse and more negative for fine textured media. Also, the capillary pressure at the drying front (interface between wet and partially dry soil) from where water is pumped by capillarity to the surface, can be texture dependent. The capillary pressure at the drying front equals the initial capillary pressure h_0 at the onset of evaporation or the air-entry value h_{air} for initially water saturated conditions. In the second case, it is a function of soil texture (h_{air} is typically more negative in fine textured soils). In both cases, it defines the matric potential head at the deepest level where water to be pumped to the evaporation plane at the surface is replaced by air. The maximum driving pressure difference to supply water to the surface is thus $h_{air} - h_{crit}$ (or $|h_{crit}| - |h_{air}|$ using absolute values of matric potential). This capillary pressure difference must overcome the height difference (gravitational potential) between the drying front and the surface. The maximum height difference, and thus the maximum length of the capillary flow paths, is just $h_{air} - h_{crit}$ and can be denoted as ‘gravity length’ L_G . Because the texture effect on h_{crit} is much stronger (several meters) than on the air-entry value h_{air} , the gravity length depends on soil texture and is longer in fine textured soils. Due to the stronger capillary forces, a deeper soil layer can be drained by stage-1 evaporation for finer textures (as long as hydraulic conductivity is not limiting—see next paragraph) as shown by larger stage-1 mass losses in Figs. 1b and 2b.

Capillary vs. viscous flow

The capillary pressure difference pumping water to the surface must overcome (in addition to height differences) the resistance to move the viscous liquid (water). Accordingly, there is an additional criterion to sustain stage-1 evaporation related to the rate of water supply along capillary flow paths and the corresponding viscous losses. The capillary pressure difference $h_{air} - h_{crit}$ is counteracted by both gravity and viscous flow and the maximum length of capillary flow is reduced compared to the gravity length. There are different possible approaches to account for viscous flow against gravity driven by a capillary pressure difference (see next paragraph). A simple approach to account for viscous effects can be to assign an effective hydraulic conductivity K_{eff} for the flow region between the drying front and the surface (Lehmann et al., 2008). The maximum length of capillary flow paths, that is denoted as characteristic length of evaporation L_C , equals:

$$L_C = \frac{|h_{crit}| - |h_{air}|}{1 + E_0/K_{eff}} \quad (3)$$

with stage-1 evaporation rate E_0 . The smaller the hydraulic conductivity K_{eff} , the larger the effect of viscous flow resistance and the more pronounced the length reduction compared to gravity length $L_G = |h_{crit}| - |h_{air}|$. Both the nominator and denominator of Eq. (3) depend on soil texture, with increasing values for fine textured soils (large absolute values of h_{crit} but small K_{eff} for fine

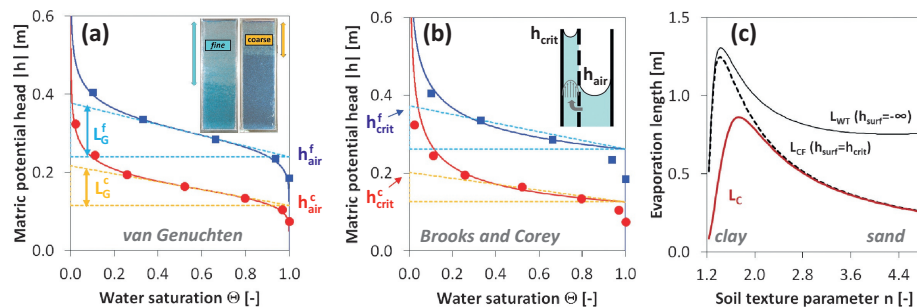


Fig. 3 Characteristic lengths of evaporation determined by (i) capillary pressure differences pumping water to the evaporating surface and (ii) the hydraulic conductivity of capillary flow paths. The range of capillary pressure differences L_G can be deduced from the soil water characteristic curve (SWC) between air-entry value h_{air} and a critical capillary pressure h_{crit} close to residual water content assigned to capillary flow cessation. The values can be determined by linearization of the SWC (van Genuchten parameterization in (a) and Brooks and Corey in (b)) with superscripts ‘c’ and ‘f’ for coarse and fine sand. The inset in (a) shows the dye distribution at end stage-1 with the drying front at depth L_G . The two capillaries in (b) represent the conceptual model of water flow from the receding meniscus in the large pore to the interface retained at the surface in the small pore. (c) Three different characteristic lengths of capillary flow paths from water table L_{WT} (Eq. 4), from the capillary fringe L_{CF} (Eq. 5) and from the drying front L_C (Eq. 3) are expressed as a function of soil texture (using the shape parameter n of the soil water characteristics curve). The three formulations differ with respect to the capillary pressure at the surface h_{surf} and the unsaturated flow across the profile.

textures). The maximum length of capillary pull to sustain evaporation rate E_0 is thus short for coarse textures because of small L_C and for fine textured media due to high values E_0/K_{eff} (see Fig. 3b). The largest characteristic length L_C can thus be found for intermediate textures (loam) due to moderately large hydraulic conductivity and capillary pressure differences.

Characteristic lengths of evaporation from shallow water tables

In systems with a shallow water table at depth L_{WT} , the evaporation rate E_0 can be sustained as long as the length of capillary paths reach the surface and the water table is replenished (Shokri and Salvucci, 2011). The existence of a relationship between maximum water table depth and sustainable evaporation rate was analyzed by Gardner (1958). Using a steady state solution and hydraulic conductivity $K(h)$ dropping exponentially with decreasing matric potential head h , the water flow can be integrated analytically along the profile and the maximum water table depth L_{WT} to sustain evaporation rate E_0 equals:

$$L_{WT} = \frac{1}{\beta} \ln \left[1 + \frac{K_s}{E_0} \right] \quad (4)$$

with saturated hydraulic conductivity K_s and the specific matric potential head $|1/\beta|$ where $K(1/\beta) = K_s/e$ (Euler number e). As shown in Fig. 3c, the length L_{WT} is larger than L_C because (i) it is considered in the derivation of Eq. (4) that the matric potential at the surface (driving the capillary flow) can go to $-\infty$ and (ii) the capillary pathways start at the free water table and not at the drying front. The mathematical procedure (and assumptions) used to develop Eq. (4) can be applied as well for the water flow from the drying front at capillary pressure h_{air} to the surface with matric potential h_{crit} . The expression for this length L_{CF} equals:

$$L_{CF} = \frac{1}{\beta} \ln \left[\frac{E_0 + K_s \exp(-\beta|h_{air}|)}{E_0 + K_s \exp(-\beta|h_{crit}|)} \right]. \quad (5)$$

Note that the thickness of the saturated region above a water table is denoted as the capillary fringe. Like the drying front for evaporation of saturated soils, the top of the capillary fringe is an interface between saturated and unsaturated conditions with matric potential head h_b . Eq. (5) thus corresponds to the length of flow paths from the capillary fringe (CF) to the surface as denoted by index CF. As shown in Fig. 3c, the lengths L_{CF} and L_C are very similar for coarse textured soils but differ for finer textures due to the different description of the unsaturated flow across the profile.

Characteristic length and stage-2 evaporation

The characteristic length of evaporation (defining the thickness of soil layer that can be drained by stage-1 evaporation) is relevant as well for stage-2 when the evaporation plane is below the surface. Two situations with respect to the evaporation plane below the surface must be distinguished. For evaporation from a shallow water table, the maximum length L_{WT} increases with decreasing evaporation rate E_0 . For an evaporation plane below the surface at depth z , the vapor must be transported across a dry soil layer by gas diffusion at a rate $E_z < E_0$. If the capillary flow is fast enough to sustain E_z (but too small to sustain E_0), the evaporation plane stays at a depth z below the surface. Without a shallow water table, the evaporation plane recedes deeper into the soil during stage-2 evaporation with a rate controlled by vapor diffusion (see next paragraph). Because the evaporation rate from the receding interface is reduced by diffusion through the dry layer, the viscous losses are reduced as well and the length of capillary flow paths from the wet subsurface to the receding evaporation plane, as determined by Eq. (3), becomes longer compared to the value computed with E_0 .

Stage-2 evaporation and diffusion across a dry soil layer

The presence of a dry surface layer reduces the evaporation rate very effectively by slow vapor diffusion from the receding evaporation plane to the surface across the tortuous diffusive pathways. The evaporation rate is determined by Fick's law of diffusion and depends on the diffusion coefficient in the dry soil D_s , the pressure difference between the vapor pressure at the evaporation plane (C_{sat} , considering saturated vapor pressure) and in the atmosphere C_{atm} , and the length of the diffusive pathways, i.e., the thickness of the dry soil layer. Because the Fickian flux is inversely proportional to the thickness of the dry layer, the evaporation rate drops continuously with the retreat of the evaporation plane. But the change from relatively constant stage-1 evaporation rates to smaller and decreasing stage-2 evaporation rates (following the Fickian law) is not instantaneous but occurs during a transition period. Experimental results based on dye deposition patterns (the deposited dye marks the position of the evaporation front) suggest that the evaporation plane during this transition period does not move gradually from the surface but 'jumps' to a depth ξ (Shokri and Or, 2011). Once the liquid is detached from the surface, a new evaporation front is formed at a depth ξ that is remarkably independent of soil texture and is about 10 mm. After the formation of the new evaporation front, it recedes then more gradually into the soil as function of time given by:

$$E(t) = D_s \cdot \frac{(C_{sat} - C_{atm})/\rho}{\sqrt{\xi^2 + t \cdot \frac{2D_s(C_{sat} - C_{atm})/\rho}{\Delta\theta}}} \quad (6)$$

with water density ρ and the water content difference $\Delta\theta$ between the wet region below the drying front and the dry layer above the vaporization plane (Or et al., 2013).

Effects of structure formation and land use on evaporation stages

The mechanisms limiting stage-1 evaporation that were explained in the previous subsections for homogeneous soils (no structures) can also be used to assess the effects of soil structure (and structure destruction) on evaporation losses. In natural soils with vegetation or large organic matter content, a surface layer with plant litter or large aggregates can be formed with small capillary forces and accordingly a short characteristic length L_c . Due to the weak capillary forces, the capillary pressure difference pumping water to the surface is small and stage-1 evaporation rate can only be sustained for a short period of time; as a result, the evaporation front recedes quickly from the surface after rainfall or irrigation. The addition of a layer with weak capillary forces can thus reduce stage-1 evaporation and evaporation losses in general as is shown in Fig. 4 for various types of surface structures. In addition to large aggregates or plant litter, crop residues can cover the surface as mulching layers (Fig. 4a). Also casts of earthworms, coarse sand layers, salt crusts (Fig. 4b) and layers of hydrophobic material (Fig. 4c) have the same suppressing effect on evaporation losses.

For the surface layers presented in Fig. 4, the evaporation losses were reduced by layering. However, depending on the capillary forces and thickness of the individual layers, layering can result in prolongation of the stage-1 evaporation process. This occurs when a fine textured material is on top of a coarser one and the pore sizes of the two materials intersect. This can formally be expressed as $h_{crit}^c < h_{air}^f$, with more negative critical capillary pressure in the coarse material h_{crit}^c than the air entry value in the fine material h_{air}^f (with superscripts c and f for coarse and fine material, respectively). In that case, an enhanced capillary pressure difference of $h_{air}^c - h_{crit}^f$ (compared to $h_{air}^c - h_{crit}^c$ in the case of homogeneous coarse material) drives the water supply to the surface. This situation corresponds to the compaction of the surface layer, increasing the capillary pull to the evaporation plane at the surface. Note that for layering with increasing pore sizes from top to bottom, the typical water content profile (water content increasing with depth) can be inversed, with the wetter layer on top of the drying layer (Goldberg et al., 2021). Structure formation and cultivation of land can also generate other structures than aggregates and layers. Freezing and thawing, burrowing activities of animals, wind erosion and compaction along wheel tracks can result in a vertical arrangement of materials with contrasting properties. The parallel arrangement of fine (strong capillary forces) and coarse textured materials (weak capillary forces) results in preferential air invasion and drying front formation in coarse material, pumping water to the fine material that can remain saturated. Later in the process, the evaporation plane in the coarse textured region starts to recede with the evaporation plane in the fine material still pinned at the surface. The duration of stage-1 evaporation (presence of hydraulically connected water at the surface) can be extended compared to homogeneous material due to the enhanced capillary pressure difference between the air-entry value h_{air}^c of the coarse material and the critical matric potential value in the fine h_{crit}^f . An example of enhanced water losses from a laboratory experiment is shown in Fig. 4c.

In addition to heterogeneities within the soil, the roughness of the surface can affect the evaporation rate as well. Surface roughness increases the surface area (compared to a flat surface), and wind and radiation induce a complex pattern of the boundary layer thickness and vapor concentrations along the surface (controlling the local evaporation rates and the viscous losses to transport water to the drying surface). Depending on the details of roughness geometry, atmospheric conditions, and hydraulic soil properties, the evaporation ratio from rough to smooth surfaces can vary considerably (Vanderborght et al., 2017).

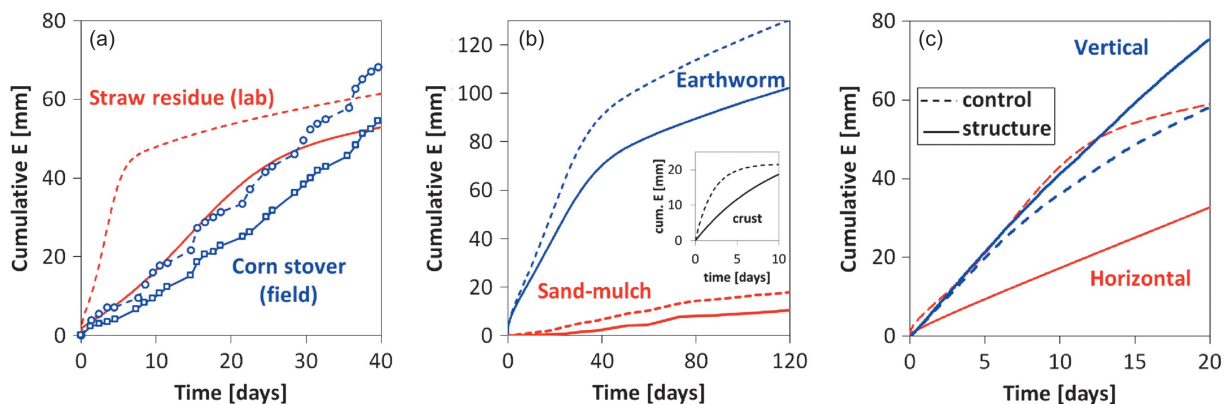


Fig. 4 Experimental results on the role of surface layers and structure formation on evaporation losses. The evaporation rate E is cumulated with time to compare the evaporation loss from a heterogeneous soil ('structure'; solid line) to a homogeneous soil ('control'; dashed line). (a) Covering the surface with straw (red; Bond and Willis, 1969) or corn stover (blue; Klocke et al., 2009) affects surface energy balance and capillary flow to the surface. (b) Adding surface layers with low capillary forces stop water supply to the surface: casts of earthworms on silt loam (blue; Ma et al., 2020), a coarse sand layer on top of sandy loam (red; Chen et al., 2020), and a salt crust of a sandy soil (inset; Zhang et al., 2013). Note that the red line shows a field experiment in cold season in a desert with low evaporation rate. (c) Lab experiments with coarse sand with a hydrophobic layer fixing evaporation plane below the surface in red (Shokri et al., 2008) and the vertical arrangement of sand layers increasing the capillary pull in blue.

Quantification of the evaporation process at different scales

Pore scale considerations

The evaporation stage transition as described in the previous sections depends on pore sizes and liquid phase continuity that can be quantified using three-dimensional imaging and simulations with pore networks (Nasir et al., 2022). In contrast to the conceptual picture of capillary bundles (as implemented in the hydraulic functions to model evaporation at larger scale), measurements and modeling reveal that water does not only flow in saturated small ‘pipes’ but along crevices and particle contacts (Tuller and Or, 2001). The dynamics of the liquid configuration in corners and its effect on evaporation rate can be studied in singular capillaries with angular cross-section, or in microfluid networks with well-defined geometries obtained with various etching techniques (Ding and Geistlinger, 2021). Despite the idealized geometry used in microfluid networks, such systems show the dominant role of liquid pinned in corners to sustain stage-1 evaporation rates. As long as continuous structures of liquid pinned in corners exist, water will flow to the surface along a capillary pressure difference and sustain stage-1 evaporation rate.

Imaging at the column scale

The combination of standard laboratory experiments (monitoring mass balance and point measurements of water content and matric potential head) with imaging methods offers a possibility to quantify evaporation rates and water content (and temperature) profiles simultaneously. Measuring the mass loss by soil evaporation and comparing it with the measured potential evaporation rates (e.g., from a water-filled column) gives information about the evaporation stage (constant or decreasing). The simultaneous imaging of dye deposition that is added in small concentrations to the water (for example, Brilliant Blue) enables the monitoring of the first and second drying front. The transition to stage-2 evaporation is marked by the onset of dye deposition below the surface (starting at a depth ζ around 10 mm as explained previously). More advanced imaging techniques (using X-ray, nuclear magnetic resonance, or neutron imaging) can delineate the evolution of the soil moisture profile during the process. Evaporation stages can also be characterized by thermal imagery of the surface, linking the evaporation rate with temperature patterns (temperature is cooler in wet regions with high evaporation rates; Nachshon et al., 2011).

Note that the evaporation process at the column scale can also be used to measure the unsaturated hydraulic conductivity and the soil water characteristics curve by simultaneous measurements of water content and matric potential gradients induced by evaporation (Schindler et al., 2010).

Processes at the field scale

In contrast to experiments under laboratory conditions, the boundary conditions in the field are more irregular with diurnal and seasonal cycles (see Fig. 1a). In addition to the temporal variations at the field scale, the spatial heterogeneities can be much larger and more irregular compared to lab experiments. The mass balance (and thus the evaporation rate) cannot be deduced from point sensors due to such heterogeneities (i.e., point measurements are not representative). To measure the water balance (and evaporation rate) under field conditions, weighable (micro-)lysimeters can be used. In contrast to micro-lysimeters that can be filled non-destructively (size in the order of 0.1 m), larger lysimeters with diameters at meter scale are typically re-packed and installed at a lysimeter station to ensure proper maintenance and instrumentation (an exception is the lysimeter network SOILCan with large soil monoliths taken non-destructively in the field and transported to the lysimeter station; Pütz et al., 2016). Evaporation from lysimeters is determined by measurements of the other terms of the mass balance equation (drainage, change in water storage, precipitation, and runoff; see Eq. (1)). Complementary to the actual evaporation rate, the potential evaporation rate can be measured from the water level in an evaporation pan (for example, Class A evaporation pan). Lysimeter networks (TERENO-SOILCan (Pütz et al., 2016) and SOPHOS (Chief et al., 2009)) allow observation of changes in water and energy balance during climate change and with different covers (e.g., bare soil or different types of plants). An alternative approach to the quantification of evaporation by measuring weight change of lysimeters, is the measurement of vapor fluxes in the atmosphere using the eddy-covariance method. The method requires the simultaneous measurement of wind speed (using a sonic anemometer) and vapor concentration (infrared gas analyzer) to determine the upwards directed evaporation fluxes. Depending on the height of the eddy covariance measurement (instruments can be installed on a high tower), the dimensions of the footprint of the measurements (i.e., the area that contributes to measured fluxes) can be several 100 m (Chu et al., 2021). This method is thus appropriate to measure evaporation at larger scales as discussed next.

From catchments to global analyses

The evaporation losses of catchments can be deduced from long-term averages, such that the water storage term (ΔS) in Eq. (1) becomes small and evaporation can be estimated as difference between precipitation P and river discharge $R + D$ (Graf et al., 2014). For high eddy covariance towers (see previous paragraph), the measurements average the evaporation rate over large areas and can be used as well to assess the water balance of slopes and small catchments. A global network of Eddy-covariance measurement sites (<https://fluxnet.org/>) provides data to quantify the mechanisms controlling evaporation rates in different climatic regions. From a global point of view, FLUXNET data provide a sparse point pattern of evaporation data and the gaps between must be filled with satellite-based approaches and modeling. Remote sensing from satellites enables mapping of key factors controlling the evaporation

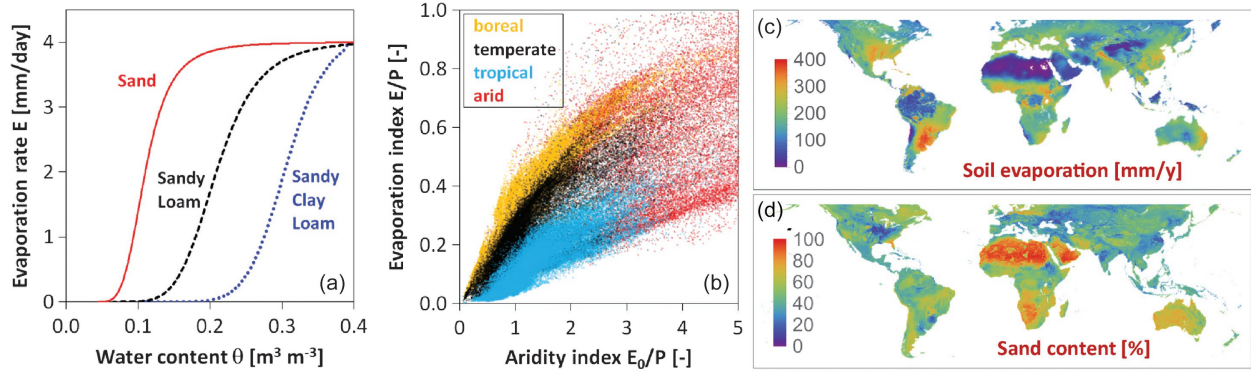


Fig. 5 Characteristics of bare soil evaporation at global scale. (a) Evaporation rate as function of water content for three different soil textures (Eq. 7). The evaporation rate is limited by the transport resistance through the unsaturated soil and the atmospheric boundary layer. (b) Simulated actual evaporation E and potential evaporation rate E_0 scaled by precipitation amount P for different climatic regions. A fraction of precipitated water is lost by leakage to deeper soil layers. The presence of plant cover in non-arid regions reduces the soil evaporation by transpiration, water interception and shading. (c) Global map of soil evaporation (without evaporation of water intercepted from plant cover) averaged over 36 years period computed with daily time steps and spatial resolution of $1/4^\circ$. (d) Soil texture information is used to (i) estimate the thickness of the soil layer contributing to stage-1 information and (ii) to compute the texture dependent evaporation rate shown in (a).

fluxes. Measurements in the infrared range of the spectrum provide information on temperature and in microwave range on air humidity (Holmes, 2019). Typically, the potential evaporation rate is deduced from temperature and relative humidity measurements and the actual evaporation rate is then expressed as a function of the potential rate using modeling concepts of various complexities. Some estimates of actual evaporation do not rely on soil information but estimate the actual rate as a function of temperature dependent vapor pressure deficit and relative humidity (Mu et al., 2011). More complex models estimate the actual evaporation rate as a function of modeled water content (Miralles et al., 2011). The relationship between evaporation rate and water content depends on soil texture (see Fig. 5a) and is related to the flow resistance through the unsaturated soil surface layer (Merlin et al., 2016; Lehmann et al., 2018). The evaporation rate from the soil surface depends on both (i) transport resistance of viscous flow in the soil and (ii) gas diffusion across the atmospheric boundary layer, and can formally be expressed as function of water content θ :

$$E(\theta) = \frac{E_0 \cdot K_{eff}(\theta)}{E_0 + K_{eff}(\theta)} \quad (7)$$

with potential evaporation rate E_0 and the effective hydraulic conductivity K_{eff} of the soil layer. The effective hydraulic conductivity can be estimated as $K_{eff} \sim 4K(\theta_{surf})$ with the water content at the surface θ_{surf} (Haghighi et al., 2013). Another effect of soil texture on evaporation rate dynamics is the texture-dependent thickness of the soil layer that contributes to stage-1 evaporation (see Fig. 3c). To model these textural effects on evaporation at the global scale, maps of soil texture of unprecedented resolution can be used (www.isric.org/explore/soilgrids, www.openlandmap.org). The water content controlling the actual evaporation rate is calculated for soil profiles consisting of a few soil layers using bucket models or solving the Richards equation. In the example shown in Fig. 5, the evaporation was calculated at global scale for a soil texture dependent layer thickness as defined by the characteristic length L_C in Eq. (3), defining the soil layer contributing actively to stage-1 evaporation (Or and Lehmann, 2019).

To compare evaporation dynamics from different catchments and regions, the evaporation rate E is plotted in Fig. 5b against the potential evaporation rate E_0 , both values scaled by the precipitation P . The figure shows the separation of regions with different climate and different soil properties. The amount of evaporation can be limited by energy (potential rate E_0 is smaller than precipitation P), water (potential rate E_0 is larger than precipitation P) but also depends on the soil hydraulic properties and the length of capillary flow paths pulling water to the evaporating surface. Note that the comparison of actual vs. potential fluxes scaled by the precipitation P is a common diagnostic tool for the analysis of evapotranspiration, i.e., the sum of transpiration, soil evaporation and evaporation of water intercepted by plant leaves (Budyko analysis; Liu et al., 2021).

Summary

The vapor transport from terrestrial systems to the atmosphere consumes globally about 60% and 25% of incoming water and energy fluxes, respectively. Notwithstanding that about 70% of global evapotranspiration can be assigned to transpiration (Paschalis et al., 2018), the evaporation from soils discussed in this chapter is a key process of the hydrologic cycle and the surface energy balance. The evaporation rate changes in time with (i) atmospheric conditions (available energy, vapor pressure deficit in the

atmosphere, wind and surface properties) and (ii) the available water in the soil. With respect to the amount of soil water, not the mere amount in the profile but the capacity to bring soil water to the evaporating surface determines if the actual evaporation rate can meet the atmospheric demand. This capacity is defined by the hydraulic conductivity and the capillary pressure difference between the wet subsoil and the drying soil surface. It depends on the length of capillary flow paths sustaining the surface evaporation rate imposed by the atmospheric conditions. This length changes with soil texture and imposed rate. The disconnection of the capillary flow paths results in the retreat of the evaporation plane from the surface and the formation of a dry soil layer. The presence of a dry layer reduces the evaporation rate that is now controlled by vapor diffusion across the dry soil layer from the receding evaporation plane.

References

- Bond JJ and Willis WO (1969) Soil water evaporation: Surface residue rate and placement effects. *Soil Science Society of America Journal* 33(3): 445–448.
- Brutsaert W and Chen D (1995) Desorption and the two stages of drying of natural tallgrass prairie. *Water Resources Research* 31(5): 1305–1313.
- Buckingham E (1907) Studies on the movement of soil moisture. In: *Bureau of soils*. Washington: Department of agriculture.
- Chen J, Xie X, Zheng X, Xue J, Miao C, Du Q, and Xu Y (2020) Effects of sand-mulch thickness on soil evaporation during the freeze–thaw period. *Hydrological Processes* 34(13): 2830–2842.
- Chief K, Young MH, Lyles BF, Healey J, Koonce J, Knight E, and Dana G (2009) *Scaling environmental processes in heterogeneous arid soils: Construction of large weighing lysimeter facility*. Las Vegas, NV: Desert Res. Inst..
- Chu H, Luo X, Ouyang Z, Chan WS, Dengel S, Biraud SC, and Zona D (2021) Representativeness of Eddy-Covariance flux footprints for areas surrounding AmeriFlux sites. *Agricultural and Forest Meteorology* 301: 108350.
- Ding Y and Geistlinger H (2021) Impact of surface roughness on evaporation in 2D micromodels. *Water Resources Research* 57(11), e2021WR029861.
- Gardner WR (1958) Some steady-state solutions of the unsaturated moisture flow equation with application to evaporation from a water table. *Soil Science* 85(4): 228–232.
- Goldberg N, Assouline S, Mau Y, and Nachshon U (2021) Compaction effects on evaporation and salt precipitation in drying porous media. *Hydrology and Earth System Sciences Discussions* 1–40.
- Graf A, Bogen HR, Drüe C, Hardelauf H, Pütz T, Heinemann G, and Vereecken H (2014) Spatiotemporal relations between water budget components and soil water content in a forested tributary catchment. *Water Resources Research* 50(6): 4837–4857.
- Haghighi E, Shahraeeni E, Lehmann P, and Or D (2013) Evaporation rates across a convective air boundary layer are dominated by diffusion. *Water Resources Research* 49(3): 1602–1610.
- Heller JP (1968) The drying through the top surface of a vertical porous column. *Soil Science Society of America Journal* 32(6): 778–786.
- Holmes TR (2019) Remote sensing techniques for estimating evaporation. In: Maggioni V and Massari C (eds.) *Extreme hydroclimatic events and multivariate hazards in a changing environment*, pp. 129–143. Elsevier.
- Idso SB, Reginato RJ, Jackson RD, Kimball BA, and Nakayama FS (1974) The three stages of drying of a field soil. *Soil Science Society of America Journal* 38(5): 831–837.
- Jalota SK and Prihar SS (1986) Effects of atmospheric evaporativity, soil type and redistribution time on evaporation from bare soil. *Soil Research* 24(3): 357–366.
- Klocke NL, Currie RS, and Aiken RM (2009) Soil water evaporation and crop residues. *Transactions of the ASABE* 52(1): 103–110.
- Lehmann P, Assouline S, and Or D (2008) Characteristic lengths affecting evaporative drying of porous media. *Physical Review E* 77(5), 056309.
- Lehmann P, Merlin O, Gentile P, and Or D (2018) Soil texture effects on surface resistance to bare-soil evaporation. *Geophysical Research Letters* 45(19): 10–398.
- Liu J, You Y, Li J, Sitch S, Gu X, Nabel JE, and Kong D (2021) Response of global land evapotranspiration to climate change, elevated CO₂, and land use change. *Agricultural and Forest Meteorology* 311: 108663.
- Ma L, Shao MA, and Li T (2020) Characteristics of soil moisture and evaporation under the activities of earthworms in typical anthrosols in China. *Sustainability* 12(16): 6603.
- Merlin O, Stefan VG, Amazirh A, Chanzy A, Ceschia E, Er-Raki S, and Khabba S (2016) Modeling soil evaporation efficiency in a range of soil and atmospheric conditions using a meta-analysis approach. *Water Resources Research* 52(5): 3663–3684.
- Miralles DG, Holmes TRH, De Jeu RAM, Gash JH, Meesters AGCA, and Dolman AJ (2011) Global land-surface evaporation estimated from satellite-based observations. *Hydrology and Earth System Sciences* 15(2): 453–469.
- Mu Q, Zhao M, and Running SW (2011) Improvements to a MODIS global terrestrial evapotranspiration algorithm. *Remote Sensing of Environment* 115(8): 1781–1800.
- Nachshon U, Shahraeeni E, Or D, Dragila M, and Weisbrod N (2011) Infrared thermography of evaporative fluxes and dynamics of salt deposition on heterogeneous porous surfaces. *Water Resources Research* 47(12).
- Nasir M, Kaito K, Patmonoaji A, Mahardika MA, She Y, Matsushita S, and Suekane T (2022) Three-dimensional pore-scale observation of drying process of porous media. *International Journal of Heat and Mass Transfer* 196: 123299.
- Oki T, Hanasaki N, Shen Y, Kanae S, Masuda K, and Dirmeyer P (2005) Global water balance estimated by land surface models participated in the GSWP-2. In: *Preprints, 19th Conf. on Hydrology, San Diego, CA, Amer. Meteor. Soc., 6.5*, <https://ams.confex.com/ams/pdfpapers/85226.pdf>.
- Or D and Lehmann P (2019) Surface evaporative capacitance: How soil type and rainfall characteristics affect global-scale surface evaporation. *Water Resources Research* 55(1): 519–539.
- Or D, Lehmann P, Shahraeeni E, and Shokri N (2013) Advances in soil evaporation physics—A review. *Vadose Zone Journal* 12(4): 1–16.
- Pan S, Tian H, Dangel SR, Yang Q, Yang J, Lu C, and Ouyang Z (2015) Responses of global terrestrial evapotranspiration to climate change and increasing atmospheric CO₂ in the 21st century. *Earth's Future* 3(1): 15–35.
- Paschalis A, Faticchi S, Pappas C, and Or D (2018) Covariation of vegetation and climate constrains present and future T/ET variability. *Environmental Research Letters* 13(10): 104012.
- Pütz T, Kiese R, Wollschläger U, Groh J, Rupp H, Zacharias S, and Vereecken H (2016) TERENO-SOILCan: A lysimeter-network in Germany observing soil processes and plant diversity influenced by climate change. *Environmental Earth Sciences* 75(18): 1–14.
- Scherer GW (1990) Theory of drying. *Journal of the American Ceramic Society* 73(1): 3–14.
- Schindler U, Durner W, von Unold G, and Müller L (2010) Evaporation method for measuring unsaturated hydraulic properties of soils: Extending the measurement range. *Soil Science Society of America Journal* 74(4): 1071–1083.
- Schlünder EU (1988) On the mechanism of the constant drying rate period and its relevance to diffusion controlled catalytic gas phase reactions. *Chemical Engineering Science* 43(10): 2685–2688.
- Shokri N and Or D (2011) What determines drying rates at the onset of diffusion controlled stage-2 evaporation from porous media? *Water Resources Research* 47(9).
- Shokri N and Salvucci GD (2011) Evaporation from porous media in the presence of a water table. *Vadose Zone Journal* 10(4): 1309–1318.
- Shokri N, Lehmann P, and Or D (2008) Effects of hydrophobic layers on evaporation from porous media. *Geophysical Research Letters* 35(19).

- Tuller M and Or D (2001) Hydraulic conductivity of variably saturated porous media: Film and corner flow in angular pore space. *Water Resources Research* 37(5): 1257–1276.
- Vanderborght J, Fetzner T, Mosthaf K, Smits KM, and Helmig R (2017) Heat and water transport in soils and across the soil-atmosphere interface: 1. Theory and different model concepts. *Water Resources Research* 53(2): 1057–1079.
- Wild M, Folini D, Hakuba MZ, Schär C, Seneviratne SI, Kato S, and König-Langlo G (2015) The energy balance over land and oceans: An assessment based on direct observations and CMIP5 climate models. *Climate Dynamics* 44(11–12): 3393–3429.
- Zhang JG, Xu XW, Lei JQ, Li SY, Hill RL, and Zhao Y (2013) The effects of soil salt crusts on soil evaporation and chemical changes in different ages of Taklimakan Desert Shelterbelts. *Journal of Soil Science and Plant Nutrition* 13(4): 1019–1028.

Evapotranspiration

Yuting Yang, Department of Hydraulic Engineering, Tsinghua University, Beijing, China

© 2023 Elsevier Ltd. All rights reserved.

This is an update of G. Stanhill, Evapotranspiration, Editor(s): Daniel Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 502–506, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00359-3>.

Introduction	501
Historical development of the concept	502
Measurement of evapotranspiration	502
Models for evapotranspiration	502
Magnitude of evapotranspiration and its components	504
Evapotranspiration in a changing environment	506
Evapotranspiration for water management	506
Summary	507
References	507

Abstract

Evapotranspiration is the second largest hydrological flux over land, which is also the linkage between the energy, water and carbon cycles. Here I review the definition, concept development and quantification of evapotranspiration. Merits and limitations of mathematical evapotranspiration models are discussed. The magnitude of evapotranspiration and its components are summarized, the responses of evapotranspiration to global environmental changes and the use of evapotranspiration information for water management are also briefly discussed.

Key points

- Evapotranspiration is a fundamental linkage between Earth's water, carbon and energy cycles.
- State-of-the-art evapotranspiration models often fail to represent the feedback between internal processes of evapotranspiration.
- Evapotranspiration estimates agree reasonably well among existing studies while quantifications of evapotranspiration components still contain large uncertainties.
- Responses of evapotranspiration to climate and land use/cover changes and the consequent impacts on the climate system are important future research directions.

Introduction

Evapotranspiration represents the process of water phase transition from liquid to gas form and then transfer into the atmosphere. Evapotranspiration may occur from the Earth's surface (e.g., the soil or a water body), through plant leaves (termed transpiration) or from rainfall intercepted by objects above ground (primarily vegetation in natural environments). The term evapotranspiration (ET) combines these three components, though interception is not explicitly mentioned in the compound term. Among different processes within the terrestrial hydrological cycle, land surface evapotranspiration plays an important role in determining the global water balance and linking the global water, energy and carbon cycles. On the one hand, ET moves a vast amount of water vapor upwards into the atmosphere and supports land surface precipitation. On the other hand, ET controls land-atmosphere feedback via modulation of the surface energy budget. From a water balance perspective, land surface evapotranspiration is the second largest water flux (after precipitation) within the terrestrial hydrological cycle, which accounts for about 65% of land surface rainfall (Oki and Kanae, 2006). This ratio can be even exceed 90% in arid regions (Budyko, 1974). Like other hydrological fluxes, evapotranspiration is commonly expressed in terms of depth per time, and common units of mm day^{-1} . When spatially integrated over an area (m^2 , such as a field, catchment, basin or country), the dimensions of evapotranspiration become volume per time, and common units are $\text{m}^3 \text{day}^{-1}$ or GI day^{-1} . In addition, because the phase change of water from liquid to gas is associated with the increase in internal energy, the latent heat of vaporization (2257 J g^{-1}) is usually used to convert the units of evapotranspiration into its energy equivalent (i.e., latent heat flux, usually in W m^{-2}).

Historical development of the concept

The formal appearance of the term “evapotranspiration” in the scientific literature was firstly made by an American geographer and climatologist, Charles Thornthwaite, who used the term “potential evapotranspiration” for climate classification (Thornthwaite, 1948). In contrast to actual evapotranspiration that occurs under real surface conditions, potential evapotranspiration defines the rate of evapotranspiration that would occur over surfaces with an (often hypothetical) unlimited supply of water.

Regardless of evapotranspiration or potential evapotranspiration, the term “evapotranspiration” itself has been refuted by a number of scientists in the field, e.g., Howard Penman, John Monteith, Charles Priestley and more recently by Miralles et al. (2020), Monteith (1986) and Penman (1948), on basis that it was clumpy and unnecessary, because the three components of “evapotranspiration” essentially share the same physical processes and the term “evaporation” was equally valid for the total water loss via vaporization and its components. Historically, the term “evaporation” was first documented in the sixteenth century (and later by Dalton, 1802) and also covered the water loss from plants.

Despite all the objections, however, the use of “evapotranspiration” has been embraced and become increasingly popular over recent decades. One possible reason is perhaps to emphasize the physiological controls on transpiration that is indeed different from a purely physical process of water loss from water/soil/wet canopy surfaces and to highlight the role of vegetation in controlling the terrestrial hydrological cycle that was overlooked at one time. The term “evapotranspiration” has also been commonly adopted by ecologists and agricultural scientists to stress the simultaneous control of plant stoma on transpiration and photosynthesis processes.

Measurement of evapotranspiration

Among all components within the hydrological cycle, ET is invisible and arguably the most difficult one to be accurately quantified. Field measurements of ET are generally classified into two groups: mass balance methods and micrometeorological methods. The mass balance methods use evaporation pans and weighing lysimeters to directly measure mass loss attributed to ET. An evaporation pan is used to determine the evaporation rate from open water surfaces under the prevailing climatic conditions. With proper implementation and calibration, pan evaporation is often considered a standard measure of potential ET. However, as evaporation pan is often too small to exert influence on the above atmosphere, the rate of evaporation recorded by evaporation pans are not necessarily equivalent to that would be measured over extensive wet surfaces as required in the definition of potential ET, if the surrounding surfaces are water-stressed. For such cases, Brutsaert (2015) introduced a term called “apparent potential evaporation” to reflect the difference.

Weighing lysimeter is another mass balance-based equipment for direct ET measurement and is often considered to be the most reliable approach of quantifying ET with the highest accuracy, yet still subject to local calibrations. Various kinds of weighing lysimeters are developed for specific purposes, ranging from cup-size micro-lysimeters, which measure soil evaporation beneath vegetation canopies or between plants, to large-scale lysimeters that ensure sufficient area and depth to support mature forest and trees growing in them. Other field methods instrument a soil profile with water content sensors to resolve other components of the water balance other than ET (similar to a lysimeter under field conditions).

Measurement of ET using micrometeorological methods includes the energy balance Bowen ratio system and the eddy-covariance system. The energy balance Bowen ratio system quantifies ET from observations of net radiation, ground heat fluxes and the vertical gradients of temperature and humidity at two fixed heights, with an assumption that the transfer coefficients of heat and water vapor are the same within the near-surface boundary layer. The eddy-covariance system measures sensible and latent heat flux through statistical covariance between the heat and moisture variation and vertical velocity using rapid responses sensors at frequencies typically equal to or greater than 10 Hz. With specific gas analyzers, the eddy-covariance system can also be used to measure exchanges of other tracer gases (e.g., CO₂, CH₄) between the land and the atmosphere. Due to its sound physical basis and fast technology advancements over the past three decades, the eddy-covariance system is now widely recognized as the best method to directly measure ET (and other tracer gases) at the ecosystem scale and is widely accepted in many important boundary layer experiments, in particular FLUXNET (Baldocchi et al., 2001). However, one unsolved issue with the eddy-covariance system is it often suffers from the energy balance non-closure problem, i.e., the sum of sensible and latent heat fluxes measured with eddy-covariance is often less than net radiation minus ground heat flux measured concurrently. There also remain difficulties in applying the method over tall, heterogeneous canopies, sloping surfaces, and when atmospheric conditions change rapidly (Monteith and Unsworth, 2013).

There are also methods for directly measuring the three ET components. Besides micro-lysimeter that measures evaporation from soil surfaces as mentioned above, the sap flow technique is designed to measure whole-plant transpiration using heat as a tracer of sap flow movement within a plant. The difficulty is then to upscale transpiration from a single tree to the entire ecosystem of interest. Canopy interception water loss can be quantified as the difference between rainfall recorded above and below vegetation canopies.

Models for evapotranspiration

Although field-based measurements are generally more reliable and accurate in quantifying ET, they are often costly and require substantial efforts in their installations and maintenances. Additionally, these field-based measurements can only provide ET

measures at the site level and often over limited periods. Consequently, mathematical models are important tools for estimating ET at varying spatial and temporal scales.

Numerous factors can affect the ET processes, including the meteorological conditions, surface moisture status, vegetation coverage, plant physiological characteristics and soil hydraulic attributes (Stanhill, 2005; Lehmann et al., 2018), so existing ET models are highly diverse in their structures and complexities depending on whether and how these factors are represented in the models. Earlier developments of ET models are made in two parallel pathways: the aerodynamic approach (Dalton, 1802; Sutton, 1934; Taylor, 1922) and the energy balance approach (Bowen, 1926; Brunt, 1939; Cummings and Richardson, 1927). The aerodynamic approach (also known as the bulk formula) estimates ET from vapor pressure gradients and a transfer coefficient determined from the wind profile, whereas the energy balance approach calculates ET using estimates of both net radiation (R_n) and the Bowen ratio (the ratio of sensible to latent heat flux). Although both methods are physically-based and theoretically sound, they had been rarely used in practical ET estimations in the early years, mainly due to the lack of surface temperature (T_s) measurements. In aerodynamic approaches, the surface temperature was needed to quantify the saturated vapor pressure at the evaporating surface, while in energy balance studies T_s was needed to estimate the net radiation and the Bowen ratio.

In 1948, Howard Penman reconciled the radiative and aerodynamic aspects of evaporation and provided a so-called combination approach for evaporation prediction (Penman, 1948). The major innovation of Penman was to eliminate surface temperature T_s via the assumption of saturated vapor pressure at the evaporating surface and linear interpolation of the air temperature from a reference height above the evaporating surface (e.g. 2 m above the surface), at which air temperature and vapor pressure were conventionally measured. As only standard meteorological data were needed, the Penman model was soon adopted by hydrologists and irrigation engineers and the approach is still widely used in the fields of hydrology, agriculture and micrometeorology.

However, although T_s is not explicitly needed in the Penman model, the Penman model still suffers from two issues regarding T_s . First, the temperature for assessing the slope of the saturated vapor pressure-temperature relationship (the Clausius-Clapeyron relations) is unknown. In practical applications, this slope is usually assessed using near-surface air temperature, but such treatment is only strictly valid if there were no temperature difference between the evaporating surface and the height at which air temperature is measured (Monteith, 1981). McColl (2020) offers a simple solution to this issue. A more difficult challenge emerges from the use of R_n as an independent forcing in the Penman model. However, R_n is always a function of T_s because T_s directly determines outgoing long-wave irradiance (via the Stefan-Boltzmann relation). One way to resolve the R_n representation problem is to assume the surface is at the air temperature and calculate the so-called isothermal net radiation. The fact that the surface is not at air temperature can then be accounted for by introducing an additional (radiative) resistance (Monteith and Unsworth, 2013). Nevertheless, these modifications to the Penman model are rarely used in current applications. At a more fundamental level, T_s is itself not independent of evaporation; instead, T_s can be seen as the consequence of a balance between surface radiative forcing and energy partitioning. In other words, neither R_n nor T_s can be considered to be truly independent of evaporation. This follows that the Penman model may be used for evaporation estimation but not for process understanding or evaporation attribution. Similar issues also exist in the model of Priestley and Taylor (1972) and others that require R_n as an input. Recently, by explicitly acknowledging the inter-dependence between radiation, surface temperature and evaporation, Yang and Roderick (2019) developed a maximum evaporation theory that circumvents the needs of R_n and T_s in estimating evaporation in a purely physically-based framework, which is a major advancement in the field and can be the basis of a new approach for evaporation estimation.

Originally, the Penman and Priestley-Taylor models were developed for evaporation from open water surfaces, hence not suitable for water-stressed conditions. In addition, stomata of plant leaves represent an additional resistance (by design) to vapor transfer between canopies and the atmosphere, which is accounted for in neither the Penman nor the Priestley-Taylor model. To account for these effects, Monteith (1965) introduced another resistance term called the surface resistance into the Penman equation and yielded a new model for ET estimation that is now widely known as the Penman-Monteith model. The difficulty of applying the Penman-Monteith model lies in the reliable determination of the surface resistance, which describes the resistance of vapor transfer from a combined soil and vegetation surface and cannot be determined independently from the first principles. It was until a large amount of ET observations became available (the value of surface resistance can then be determined by inverting the Penman-Monteith model with observed ET), which motivated the development of empirical surface resistance models (e.g., the Jarvis type models; (Jarvis, 1976)), the Penman-Monteith model had become increasingly popular in practical ET estimations. Similar approaches can also be applied to the Priestley-Taylor model by empirically reducing the Priestley-Taylor coefficient under potential conditions to reflect the environmental limitations on ET under non-potential conditions (e.g., Martens et al., 2017).

In addition to surface resistance, the Penman-Monteith type equations also require another resistance term (i.e., aerodynamic resistance) that quantifies the heat and water vapor transfer from the evaporating surface into the atmosphere, which is a function of wind speed, surface roughness and atmospheric stability. For a neutral atmosphere and a wind speed of 5 m s^{-1} (measured at 2 m height), typical values of aerodynamic resistance are 68 s m^{-1} , 45 s m^{-1} , 18 s m^{-1} and 6.5 s m^{-1} for water surfaces, grass (vegetation height = 0.1 m), agricultural crops (vegetation height = 1 m) and forest (vegetation height = 10 m), respectively (Shuttleworth, 1993).

Quantification of ET at regional or even larger scales has been rapidly advancing since satellite remote sensing data became readily available in the 1970s. In general, ET models based on remote sensing can be grouped into three types: surface energy balance (SEB) models, vegetation index (VI)—land surface temperature (T_s) space models, and Penman-Monteith-based models. An SEB model explicitly deals with sensible heat flux (based on concurrent observations of land surface temperature and near-surface air temperature) and calculates latent heat flux (or ET) as a residual in the surface energy budget equation (Yang et al., 2015). A VI— T_s space model determines the partitioning of surface available energy into sensible and latent heat fluxes based on the interpretation of the image (pixel) distribution in VI— T_s space (Carlson, 2007). Different from the SEB and VI— T_s models that rely on remotely sensed T_s , the Penman-Monteith based models do not require T_s as model input. Instead, they calculate ET

directly from the Penman-Monteith type equations in which the surface resistance is estimated based on remotely sensed VI and concurrent ground observations/satellite retrievals of meteorological conditions (Mu et al., 2007; Zhang et al., 2016). It should be noted that traditional satellites (e.g., Landsat, MODIS, AVHRR) only have moderate-to-low temporal coverages, which requires an additional temporal scaling method to upscale instantaneous ET retrievals at the satellite-overpassing time to longer time scales. This upscaling procedure often introduces additional uncertainties in the final ET estimates. Recent developments of geostationary satellites, which provide temporally continuous observations of land surface variables, represent an effective and promising way to overcome uncertainties raised by traditional upscaling methods and should be considered as a key focus in future studies.

Magnitude of evapotranspiration and its components

Considerable efforts have been made to estimate global terrestrial ET. At the global level, these estimates based on different methodologies are typically around 500 mm per year (ranging between 450 and 550 mm yr⁻¹) (Mianabadi et al., 2019; Zhang et al., 2016). However, the magnitude of ET shows a great spatial variability across global terrestrial environments (Table 1). The rate of ET is generally higher in tropical regions and gradually decreases towards both poles (Fig. 1A). At the continental scale, South America

Table 1 Evapotranspiration and its components in different Geographic regions.

Geographic region	Total ET (mm yr ⁻¹)	E_c (mm yr ⁻¹)	E_c/ET (%)	E_s (mm yr ⁻¹)	E_s/ET (%)	E_i (mm yr ⁻¹)	E_i/ET (%)
Global	487	366	75	49	10	72	15
Northern Hemisphere	392	265	68	76	19	51	13
Southern Hemisphere	741	568	77	60	8	113	15
Asia	361	254	70	58	16	49	14
Africa	481	383	80	49	10	49	10
North America	362	265	73	48	13	49	14
South America	949	726	77	32	3	191	20
Oceania	372	276	74	82	22	14	4
Europe	396	293	74	24	6	79	20

E_c : plant transpiration; E_s : soil evaporation (including snow sublimation); E_i : canopy interception water loss.

Data is taken from the GLEAM-v3.5a global ET product from 1982 to 2012 Martens B, Miralles DG, Lievens H, van der Schalie R, de Jeu RAM, Fernández-Prieto D, Beck HE, Dorigo WA and Verhoest NEC (2017) GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development* 10(5): 1903–1925.

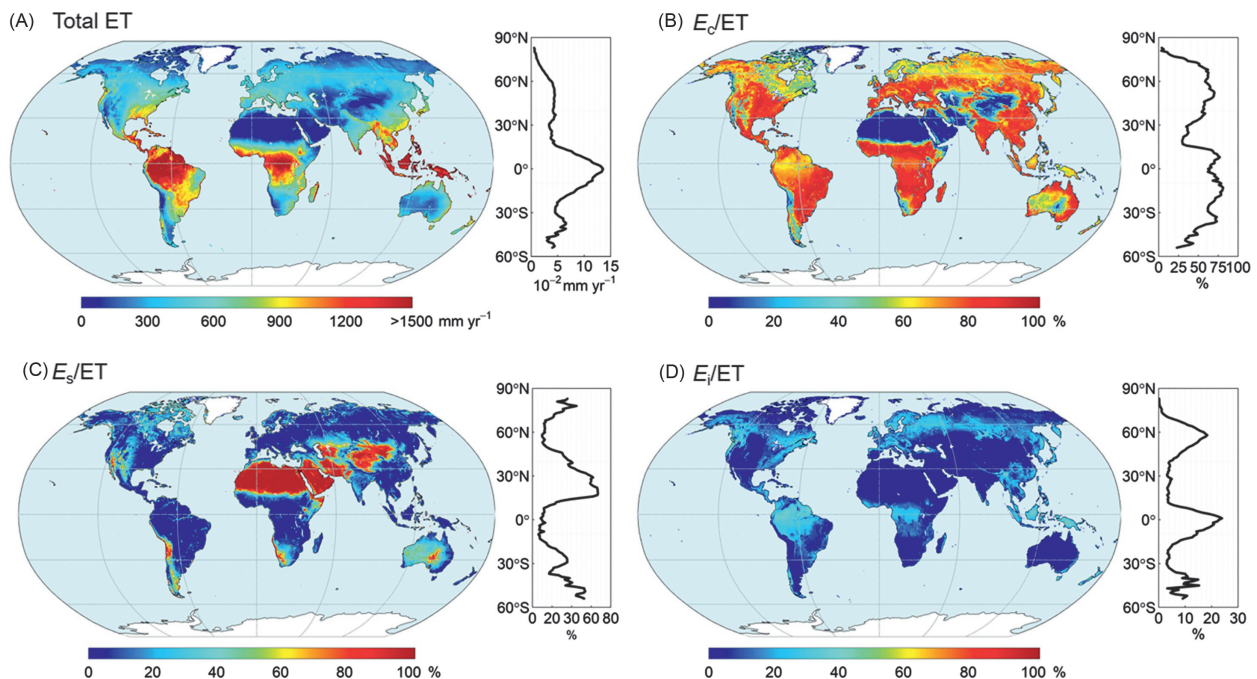


Fig. 1 Global pattern of mean annual ET and its components. (A) Total ET, (B) E_c/ET , (C) E_s/ET and (D) E_i/ET . E_c : plant transpiration; E_s : the sum of evaporation from soil/open water surface and sublimation from snow surfaces; E_i : canopy interception water loss. Data is taken from the GLEAM-v 3.5a global ET product from 1982 to 2012 Martens B, Miralles DG, Lievens H, van der Schalie R, de Jeu RAM, Fernández-Prieto D, Beck HE, Dorigo WA and Verhoest NEC (2017) GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development* 10(5): 1903–1925.

shows a way larger mean annual ET than other continents, due to a high ET rate over Amazonia. Africa has the second largest mean annual ET that is close to the global mean value, primarily driven by a high ET rate from the Congo basin despite a low ET in the Sahara Desert. For the remaining continents, their mean annual ET are all lower than (or close to) 400 mm yr^{-1} . The magnitude of ET also shows an evident distinction between climatic aridity zones (Table 2). In global humid regions where ET is primarily controlled by energy availability, the mean annual ET is around 570 mm yr^{-1} . By contrast, in arid regions where water is the most limiting factor of ET, the mean annual ET is lower than 130 mm yr^{-1} . At the biome level, forests generally show a higher ET rate than other biome types, with the highest ET rates found in evergreen broadleaf forests (primarily tropical rainforests) and the lowest ET rates occurred in deserts (Table 3).

While the estimates of total ET in different studies show a certain degree of agreement, quantification of ET components is more difficult and their estimates diverge substantially among studies. Plant transpiration (E_c) is commonly acknowledged as the largest component of ET at the global scale, yet existing estimates of the ratio of transpiration to total ET vary greatly from less than 25% to as large as 90% (Fatichi and Pappas, 2017; Wei et al., 2017). The ratio of E_c /ET estimated from land surface models (22–58%) are generally lower than those from other methods, whereas isotope-based approaches often give the highest E_c /ET estimates (65–90%; e.g., Good et al., 2015; Jasechko et al., 2013; Wei et al., 2017). The E_c /ET ratio estimated from methods based on remote sensing is typically between 50% and 75%, which are broadly consistent with measured values across (limited) global ecosystems (50–65%) (Schlesinger and Jasechko, 2014; Wei et al., 2017). Using field observations to constrain climate model predictions, a recent study reported a global E_c /ET ratio of $62\% \pm 6\%$, which seems to be a reasonable estimate (Lian et al., 2018). Compared with E_c , the other two ET components account for much smaller proportions of total ET. Typical values of soil evaporation (E_s) estimates are about 10–25% of total ET and those of canopy interception (E_i) are 10–20% of total ET (Miralles et al., 2016; Zhang et al., 2016).

Spatially, places with a larger E_c /ET usually also show a greater E_i /ET, which are primarily distributed in forest ecosystems as well as in humid regions. An interesting phenomenon is that the humid region does not necessarily have the highest E_c /ET value, due to more interception water loss in the region, compared with sub-humid and semi-arid regions (Table 2). Arid regions have the smallest E_c /ET but the greatest E_s /ET value due to sparse vegetation coverage in the region and thus ET consists primarily of non-biological water loss. For different biomes, the greatest E_c /ET ratio is found in cropland, possibly due to artificial measures that intend to minimize non-biological water loss and thereby maximize water use efficiency in these locations. In contrast, the smallest E_c /ET occurs in deserts, where vegetation cover is sparse and the total ET is dominated by soil evaporation.

Table 2 Evapotranspiration and its components in different climatic regions.

Climatic region	Total ET (mm yr^{-1})	E_c (mm yr^{-1})	E_c /ET (%)	E_s (mm yr^{-1})	E_s /ET (%)	E_i (mm yr^{-1})	E_i /ET (%)
Arid	129	56	43	72	58	1	1
Semi-arid	386	304	79	71	18	11	3
Sub-humid	419	340	81	39	9	40	10
Humid	570	428	75	36	6	106	19

E_c : plant transpiration; E_s : soil evaporation (including snow sublimation); E_i : canopy interception water loss.

Data is taken from the GLEAM-v3.5a global ET product from 1982 to 2012 Martens B, Miralles DG, Lievens H, van der Schalie R, de Jeu RAM, Fernández-Prieto D, Beck HE, Dorigo WA and Verhoest NEC (2017) GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development* **10**(5): 1903–1925.

Table 3 Evapotranspiration and its components in different biome types.

Biome type ^a	Total ET (mm yr^{-1})	E_c (mm yr^{-1})	E_c /ET (%)	E_s (mm yr^{-1})	E_s /ET (%)	E_i (mm yr^{-1})	E_i /ET (%)
ENF	495	313	63	36	7	146	29
EBF	1335	943	71	22	2	370	28
DNF	364	253	70	50	14	61	17
DBF	680	528	78	22	3	130	19
MF	519	352	68	39	8	128	25
Shrubland	288	214	74	62	22	12	4
Savanna	587	461	79	36	6	90	15
Grassland	376	293	78	66	18	17	5
Cropland	502	430	86	47	9	25	5
Desert	72	13	18	59	82	0.4	0.6

E_c : plant transpiration; E_s : soil evaporation (including snow sublimation); E_i : canopy interception water loss.

^aThe biome type classification is taken from the MODIS land use classification product (MCD12Q1, 2012), where ENF, EBF, DNF, DBF and MF stand for Evergreen Needleleaf Forest, Evergreen Broadleaf Forest, Deciduous Needleleaf Forest, Deciduous Broadleaf Forest and Mixed Forest, respectively.

Data is taken from the GLEAM-v3.5a global ET product from 1982 to 2012 Martens B, Miralles DG, Lievens H, van der Schalie R, de Jeu RAM, Fernández-Prieto D, Beck HE, Dorigo WA and Verhoest NEC (2017) GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development* **10**(5): 1903–1925.

Evapotranspiration in a changing environment

Due to the intrinsic nature of ET in linking the water, carbon and energy cycles, changes in ET have been widely studied to assess the impacts of global change on the Earth system. According to the IPCC AR6 assessment report, global mean surface temperature has increased by about 1 °C in the last decade compared to the pre-industrial level (Gulev et al., 2021). Meanwhile, observations and climate model simulations report a global hydrologic sensitivity of about 2% K⁻¹ (Allen and Ingram, 2002; Boer, 1993), suggesting that the global ET has increased by ~2% since the industrial revolution. This ET increase is so much smaller than the increase in globally averaged atmospheric specific humidity that is close to the Clausius-Clapeyron value of around 7% K⁻¹. This is mainly because the increased surface temperature directly yields an increased outgoing longwave radiation so that changes in net radiation (and thus the total available energy for ET) are small (Roderick et al., 2014). The relatively low hydrologic sensitivity suggests a somewhat muted response of ET to climate warming. Nevertheless, this relatively small ET increase only applies to the global scale, at which ET mostly consists of evaporation from ocean surfaces. On land, trends in ET can be much larger yet with high spatial heterogeneity.

Elevation in atmospheric CO₂ concentration (CO₂) is the ultimate driver of anthropogenic climate warming. Besides its direct greenhouse gas effect, elevated (CO₂) also affects vegetation water use via the so-called fertilization effect. On one hand, plant stomata partially close in a higher (CO₂) environment, which leads to a higher stomatal (and thus surface) resistance and consequently a lower transpiration rate at the leaf level. Field observations reported a mean increase in leaf-level stomatal resistance of 22% for an increase in (CO₂) of 201 ppm (from 366 to 567 ppm; Ainsworth and Rogers, 2007), which is equivalent to a stomatal resistance sensitivity to changes in (CO₂) of 0.11% ppm⁻¹. This is in good agreement with climate model projections that show a mean surface resistance sensitivity to changes in (CO₂) of 0.09% ppm⁻¹ (Yang et al., 2019). Studies further showed that the reduction in ET caused by (CO₂)-induced partial stomatal closure is of a similar magnitude to the increased ET caused by warming-induced vapor pressure deficit increases in global non-water-limited regions, leaving the ET changes strictly follow that of net irradiance (Milly and Dunne, 2016; Yang et al., 2019). On the other hand, the fertilization effect of (CO₂) on vegetation increases foliage cover and plant rooting depth (Nie et al., 2013; Zhu et al., 2016). With all else held constant, increased foliage cover and plant rooting depth would result in higher water consumption (i.e., transpiration) by vegetation. Due to the opposing responses of vegetation water use to elevated (CO₂) at the leaf level and the canopy level, the sign and magnitude of the net (CO₂) impact on transpiration (and ET) still remain contentious. Field observations based on ecosystem-scale Free Air CO₂ Enrichment experiments do not show a consistent ET change (increase or decrease) under elevated (CO₂) either (Donohue et al., 2017).

Land use and land cover changes also affect terrestrial ET. Of particular interest, revegetation, as one of the most effective measures for ecosystem remediation and climate mitigation, has been carried out across the globe over the past few decades, resulting in a significant greening trend, especially in China and India (Chen et al., 2019). These extra plants consume more water, resulting in an enhanced ET in revegetated areas. Studies have been arguing that the enhanced ET caused by revegetation may lead to local water resource imbalance, as evidenced by declines in soil moisture and streamflow in revegetation regions (e.g., Feng et al., 2016). On the other hand, enhancement in ET increases local moisture supply. This, together with revegetation-induced alternations in local energy partitioning that changes the atmospheric stability within the lower boundary layer, potentially enhance local moisture recycling and hence promote the formation of more precipitation that sustains the continuity of ET increases.

Evapotranspiration for water management

Knowledge of ET is important for efficient water resources management, especially for irrigation scheduling. Prior to ET information being available, the need for irrigation was determined in several ways, for example, using visual indicators such as plant color or leaf wilting (Doorenbos and Pruitt, 1977). However, these indicators appear too late to prevent yield reduction or decline in quality. In comparison, ET directly indicates the crop water requirement, which, in combination with soil characteristics, can be used to determine not only when to irrigate but also the amount of water needed (Goyal, 2014). Consequently, adopting ET-based irrigation scheduling is also an important way of water-saving in irrigation regions by avoiding over-irrigation (e.g., flood irrigation with fixed amount) and the likely consequent soil salinity (Chhabra, 1996; Yang et al., 2012). Because of all these benefits, many methods have been proposed to determine ET for different crops and agricultural regions to guide irrigation scheduling over years, among which the “crop coefficient—reference crop ET” method recommended by FAO may be arguably the most widely acknowledged and used method in today’s irrigation scheduling practice all over the world (Allen et al., 1998). In addition to agricultural water management, information of ET has also been increasingly used in the Integrated Water and Environment Management (IWEM) at the basin scale (Mendicino and Senatore, 2012; Wu et al., 2014), since ET not only quantifies the total basin water consumption but also represents the central linkage between the water cycle and other ecosystem or environment services (e.g., agricultural production and carbon sequestration). The difference between water input (mainly precipitation) and consumption (ET) also determines the amount of environmental flow that fundamentally supports the aquatic ecosystems within the basin. Nevertheless, despite the great progress in remote sensing ET techniques over the past few decades, the precision of ET information over large spatial extents and long temporal periods still remains a key challenge to successful basin-wide IWEM practice.

Summary

As the second largest hydrological fluxes on land, evapotranspiration fundamentally links the Earth's water, carbon and energy cycles and serves as a key indicator of global change. However, the quantification of evapotranspiration, in particular its components, has long been a challenge. Field methods are costly and can only quantify evapotranspiration at a point scale and for relatively short periods, whereas mathematical models often fail in presenting the feedback between internal processes of evapotranspiration. Averaged over global terrestrial environments, plant transpiration is the largest component of total evapotranspiration, which somehow justifies the use of the combined term "evapotranspiration" such that to highlight the role of terrestrial vegetation in controlling evapotranspiration. With the ongoing climate change and intensified human activities that considerably alter land surface characteristics, the quantification of responses in evapotranspiration to these global change signals and the consequent feedbacks to the climate system are key scientific questions to be answered in future studies.

References

- Ainsworth AE and Rogers A (2007) The response of photosynthesis and stomatal conductance to rising $[CO_2]$: Mechanisms and environmental interactions. *Plant, Cell & Environment* 30(3): 258–270.
- Allen MR and Ingram WJ (2002) Constraints on future changes in climate and the hydrologic cycle. *Nature* 419(6903): 228–232.
- Allen GR, Pereira SL, Raes D, and Smith M (1998) *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements*. Rome: Food and Agricultural Organization of the United Nations (FAO)300. Publication No. 56.
- Baldocchi D, et al. (2001) FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society* 82(11): 2415–2434.
- Boer GJ (1993) Climate change and the regulation of the surface moisture and energy budgets. *Climate Dynamics* 8(5): 225–239.
- Bowen IS (1926) The Ratio of Heat Losses by Conduction and by Evaporation from any Water Surface. *Physical Review* 27(6): 779–787.
- Brunt D (1939) *Physical and Dynamical Meteorology*. Cambridge: Cambridge University Press, p. 428.
- Brutsaert W (2015) *Hydrology: An Introduction*. Cambridge, UK: Cambridge University Press, p. 605.
- Budyko MI (1974) *Climate and Life*. San Diego, California: Academic, p. 508.
- Carlson T (2007) An Overview of the "Triangle Method" for Estimating Surface Evapotranspiration and Soil Moisture from Satellite Imagery. *Sensors* 7(8): 1612–1629.
- Chen C, et al. (2019) China and India lead in greening of the world through land-use management. *Nature Sustainability* 2(2): 122–129.
- Chhabra R (1996) Irrigation and salinity control. In: Chhabra R (ed.) *Soil Salinity and Water Quality*. London: Routledge.
- Cummings NW and Richardson B (1927) Evaporation from Lakes. *Physical Review* 30(4): 527–534.
- Dalton J (1802) Experimental essays on the constitution of mixed gas; on the force of steam or vapor from water and other liquids in different temperatures, both in a Torricellian vacuum and in air; on evaporation and on the expansion of gases by heat. *Manchester Literary and Philosophical Society* 5: 535–602.
- Donohue RJ, Roderick ML, McVicar TR, and Yang Y (2017) A simple hypothesis of how leaf and canopy-level transpiration and assimilation respond to elevated CO_2 reveals distinct response patterns between disturbed and undisturbed vegetation. *Journal of Geophysical Research – Biogeosciences* 122(1): 168–184.
- Doorenbos J and Pruitt WO (1977) *Crop Water Requirements*. FAO Irrigation and Drainage Paper 24 Rome: Food and Agriculture Organization of the United Nations.
- Fatchi S and Pappas C (2017) Constrained variability of modeled T:ET ratio across biomes. *Geophysical Research Letters* 44: 6795–6803.
- Feng X, et al. (2016) Revegetation in China's Loess Plateau is approaching sustainable water resource limits. *Nature Climate Change* 6(11): 1019–1022.
- Good SP, Noone D, and Bowen G (2015) Hydrologic connectivity constrains partitioning of global terrestrial water fluxes. *Science* 349(6244): 175–177.
- Goyal MR (2014) Water management for agronomic crops in Trinidad. In: Goyal MR and Harmsen EW (eds.) *Evapotranspiration Principles and Applications for Water Management*. Toronto: Apple Academic Press.
- Gulev SK et al. (2021) Changing State of the Climate System Rep.
- Jarvis PG (1976) The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field. *Philosophical Transactions of the Royal Society B* 273: 593–610.
- Jasechko S, Sharp ZD, Gibson JJ, Birks SJ, Yi Y, and Fawcett PJ (2013) Terrestrial water fluxes dominated by transpiration. *Nature* 496(7445): 347–350.
- Lehmann P, Merlin O, Gentile P, and Or D (2018) Soil texture effects on surface resistance to bare-soil evaporation. *Geophysical Research Letters* 45(19): 10398–10405.
- Lian X, et al. (2018) Partitioning global land evapotranspiration using CMIP5 models constrained by observations. *Nature Climate Change* 8(7): 640–646.
- Martens B, Miralles DG, Lievens H, van der Schalie R, de Jeu RAM, Fernández-Prieto D, Beck HE, Dorigo WA, and Verhoest NEC (2017) GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development* 10(5): 1903–1925.
- McColl KA (2020) Practical and theoretical benefits of an alternative to the penman-monteith evapotranspiration equation. *Water Resources Research* 56(6): e2020WR027106.
- Mendicino G and Senatore A (2012) The role of evapotranspiration in the framework of water resource management and planning under shortage conditions. In: Irmak A (ed.) *Evapotranspiration Remote Sensing and Modeling*. IntechOpen.
- Mianabadi A, Coenders-Gerrits M, Shirazi P, Ghahraman B, and Alizadeh A (2019) A global Budyko model to partition evaporation into interception and transpiration. *Hydrology and Earth System Sciences* 23(12): 4983–5000.
- Milly PCD and Dunne KA (2016) Potential evapotranspiration and continental drying. *Nature Climate Change* 6: 946.
- Miralles DG, et al. (2016) The WACMOS-ET project—Part 2: Evaluation of global terrestrial evaporation data sets. *Hydrology and Earth System Sciences* 20(2): 823–842.
- Miralles DG, Brutsaert W, Dolman AJ, and Gash JH (2020) On the use of the term "Evapotranspiration" *Water Resources Research* 56(11): e2020WR028055. <https://doi.org/10.1029/2020WR028055>.
- Monteith J (1965) Evaporation and environment. *Symposium of the Society of Experimental Biology* 19: 205–224.
- Monteith JL (1981) Evaporation and surface temperature. *Quarterly Journal of the Royal Meteorological Society* 107(451): 1–27.
- Monteith JL (1986) Howard latimer penman, 10 April 1909—13 October 1984. *Biographical Memoirs of Fellows of the Royal Society* 32: 377–404.
- Monteith JL and Unsworth MH (2013) *Principles of Environmental Physics*, 4th edn. Oxford, UK: Academic Press, p. 401.
- Mu Q, Heinsch FA, Zhao M, and Running SW (2007) Development of a global evapotranspiration algorithm based on MODIS and global meteorology data. *Remote Sensing of Environment* 111(4): 519–536.
- Nie M, Lu M, Bell J, Raut S, and Pendall E (2013) Altered root traits due to elevated CO_2 : A meta-analysis. *Global Ecology and Biogeography* 22(10): 1095–1105.
- Oki T and Kanae S (2006) Global hydrological cycles and world water resources. *Science* 313(5790): 1068–1072.
- Penman HL (1948) Natural evaporation from open water, bare soil and grass. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences* 193(1032): 120–145.
- Priestley CHB and Taylor RJ (1972) On the assessment of surface heat flux and evaporation using large-scale parameters. *Monthly Weather Review* 100: 81–92.

- Roderick ML, Sun F, Lim WH, and Farquhar GD (2014) A general framework for understanding the response of the water cycle to global warming over land and ocean. *Hydrology and Earth System Sciences* 18(5): 1575–1589.
- Schlesinger WH and Jasechko S (2014) Transpiration in the global water cycle. *Agricultural and Forest Meteorology* 189–190: 115–117.
- Shuttleworth WJ (1993) Evaporation. In: Maidment DR (ed.) *Handbook of Hydrology*. New York: McGraw-Hill Education.
- Stanhill G (2005) Evapotranspiration. In: Hillel D (ed.) *Encyclopedia of Soils in the Environment*. Academic Press.
- Sutton OG (1934) Wind structure and evaporation in a turbulent atmosphere. *Proceedings of the Royal Society of London, Series A* 146(858): 701–722.
- Taylor GI (1922) Diffusion by Continuous Movements. *Proceedings of the London Mathematical Society* s2–20(1): 196–212.
- Thornthwaite CW (1948) An Approach toward a Rational Classification of Climate. *Geographical Review* 38(1): 55–94.
- Wei Z, Yoshimura K, Wang L, Miralles DG, Jasechko S, and Lee X (2017) Revisiting the contribution of transpiration to global terrestrial evapotranspiration. *Geophysical Research Letters* 44(6): 2792–2801.
- Wu B, Jiang L, Yan N, Perry C, and Zeng H (2014) Basin-wide evapotranspiration management: Concept and practical application in Hai Basin, China. *Agricultural Water Management* 145: 145–153.
- Yang YT and Roderick ML (2019) Radiation, surface temperature and evaporation over wet surfaces. *Quarterly Journal of the Royal Meteorological Society* 145: 1118–1129.
- Yang YT, Shang S, and Jiang L (2012) Remote Sensing temporal and spatial patterns of evapotranspiration and the responses to water management in a large irrigation district of North China. *Agricultural and Forest Meteorology* 164: 112–122.
- Yang YT, Long D, Guan HD, Liang W, Simmons C, and Batelaan O (2015) Comparison of three dual-source remote sensing evapotranspiration models during the MUSOEXE-12 campaign: Revisit of model physics. *Water Resources Research* 51(5): 3145–3165.
- Yang YT, Roderick ML, Zhang SL, McVicar TR, and Donohue RJ (2019) Hydrologic implications of vegetation response to elevated CO₂ in climate projections. *Nature Climate Change* 9(1): 44–48.
- Zhang Y, et al. (2016) Multi-decadal trends in global terrestrial evapotranspiration and its components. *Scientific Reports* 6: 19124. <https://doi.org/10.1038/srep19124>.
- Zhu Z, et al. (2016) Greening of the Earth and its drivers. *Nature Climate Change* 6(8): 791–795. <https://doi.org/10.1038/nclimate3004>.

Plant available water

Quirijn de Jong van Lier^a, Sally D Logsdon^b, Everton Alves Rodrigues Pinheiro^c, and Paulo Ivonir Gubiani^d, ^aCenter for Nuclear Energy in Agriculture, University of São Paulo, Piracicaba, Brazil; ^bAgricultural Research Service, USDA, Ames, IA, United States; ^cFederal University of Tocantins, Gurupi, Brazil; ^dFederal University of Santa Maria, Santa Maria, Brazil

© 2023 Elsevier Ltd. All rights reserved.

Introduction	509
The upper limit of plant available water: Field Capacity	510
The lower limits of plant available water	511
Permanent wilting point	511
Limiting hydraulic conditions	511
Other limitations to plant water availability	512
Plant available water: The role of soil texture and structure	513
Plant available water and soil texture	513
Plant available water and soil structure: The role of soil management	513
The use of plant available water	513
Use of TAW and RAW in irrigation management	513
Use of TAW and RAW in hydrological models	514
Use of TAW and RAW as ecological indicators	514
Concluding remarks	514
References	514

Abstract

Soil water is only partially available to plants. A wet soil quickly drains to field capacity (FC), and during further drying through evapotranspiration, water potential and hydraulic conductivity decrease, and root water uptake becomes more difficult. At some water content, the limiting point (LP), drought stress leads to transpiration reduction. When reaching the permanent wilting point (WP), root water uptake is too little for a plant to survive. The difference between FC and WP is called total available water (TAW), whereas FC minus LP represents the readily available water (RAW). As both texture and structure affect soil hydraulic properties, they can be linked to TAW and RAW. Poor soil management may negatively affect plant available water.

Introduction

Not all water in the soil can be used equally by plants. The maximum amount of water a soil can hold and make available to plants has been the subject of research of many disciplines including agronomy, soil science, climatology, hydrology, and ecology.

The availability of soil water for root water uptake is determined by rooting characteristics and soil hydraulic properties. There may be limitations caused by macroscopic or microscopic porosity. The macroscopic limitation refers to very wet conditions, in which the soil water will quickly drain through macropore flow and move below the reach of the rooting system. The microscopic limitations determine the drought stress on the dry side. Water is held by capillarity stresses within the rooting zone but can no longer be extracted by plant roots due to its low potential energy and reduced soil hydraulic conductivity.

The plant available water fraction, also referred to as *total available water* (TAW), is commonly defined as the difference between the volumetric water contents at field capacity (FC) and the permanent wilting point (WP):

$$TAW = FC - WP \quad (1)$$

The values of FC and WP are typically obtained from soil water retention characteristics, with FC ranging from -50 to -300 hPa water potential and WP corresponding to $-15,000$ hPa water potential.

Another term with a similar meaning, but with a slightly different purpose, is the *readily available water* (RAW). The lower limit of RAW is the limiting soil water content (LP). Rather than being based on water potential like WP, LP is based on the limiting hydraulic conductivity causing the onset of plant drought stress:

$$RAW = FC - LP \quad (2)$$

RAW can also be interpreted as the fraction p of TAW that can be depleted before the onset of drought stress:

$$RAW = p \cdot TAW \quad (3)$$

where p is the depletion factor, ranging between 0 and 1, and defined as

$$p = \frac{FC - LP}{FC - WP} \quad (4)$$

For a healthy plant, when the soil water content is within the RAW range, the transpiration is at its potential or sink limited level, depending exclusively on the atmospheric demand. On the other hand, below LP transpiration falls below its potential level and becomes source limited (soil resistance to the water flow is increased), meaning that the water availability is insufficient to sustain the atmospheric water demand.

The amount of available water depends not only on FC, LP, and WP, but also on rooting depth Z_r . A plant with a deeper rooting system exploits a larger volume of soil and has more water at its disposal. TAW and RAW can therefore be multiplied by Z_r to find the respective total and readily available storage capacity in length units, e.g., mm. As root depth depends greatly on plant species and phenological stage, the storage capacity is always a variable quantity. For instance, the root depth of most annual crops is less than 1 m, while deep-rooted natural plants in a forest may reach several meters. Furthermore, the role of active roots and root length density stratification in the soil profile is not included in the root depth parameter. For local scales with homogeneous cropped fields, determination of root depth may not be a critical issue. However, for catchment and regional scales, the diversity of plant species and phenological stages make root depth determination extremely challenging.

The upper limit of plant available water: Field Capacity

Field capacity (FC) is a concept originating from irrigation management and associated with the soil water content after major drainage has ceased. Although it refers to a discontinuity of the water films in soil pores, there are various interpretations based on a specified matric potential, time after thorough wetting, or minimal flow rate at the profile bottom. Usually, field capacity is determined in the laboratory after desorbing an undisturbed soil sample to a specified matric potential (often -300 or -100 hPa). Sometimes field capacity is determined in the field one or 2 days after thorough wetting of a soil, then taking soil samples for volumetric water content. Other field approaches examine matric potential using a tensiometer and measure soil water content with sensors.

Simulated water movement in soil has been used to determine when drainage is small and field capacity is reached (Meyer and Gee, 1999; Reynolds, 2018). Sometimes the modeled approaches result in unrealistically long predicted times to cessation of drainage, up to 100 or even 1000 days for soils with high clay contents. Part of the problem is ignoring the effect of soil structure and biopores or cracks, which means that soils with high clay contents can have large pores that drain fast. The main problem with modeling approaches is that they are usually determined for soil areas without plants. Field approaches are more realistic than modeled flux-based output (Chandler et al., 2017). Fig. 1 gives an example using the field procedure of Chandler et al. (2017) to determine field capacity from soil water content data. Often field-determined field capacity is wetter than laboratory-determined field capacity (De Jong Van Lier, 2017).

An important issue in interpreting FC as the upper limit for PAW is the time to reach FC associated with a negligible bottom flux. Laboratory-based approaches ignore the plant root—soil interaction in water uptake. Many field studies have shown that plants often use water from the soil wetter than the drained upper limit. Roots can extract soil water to meet the atmospheric demand even as drainage continues (De Jong Van Lier, 2017; Logsdon, 2019). Roots do not wait 2 days or more before taking up soil water unless there is aeration stress which may lower the upper limit of soil water content for plant water uptake (Letey, 1985; Da Silva and Kay, 2004).

The depth of root growth also affects plant available water provided by soils. As upper soil layers dry, plants might extend their root systems to explore deeper soil, contravening a fixed root zone. Wetter soil below the root zone results in a gradient to the drier root zone, replenishing some of what the roots have already extracted and supplying additional soil water that can be extracted by

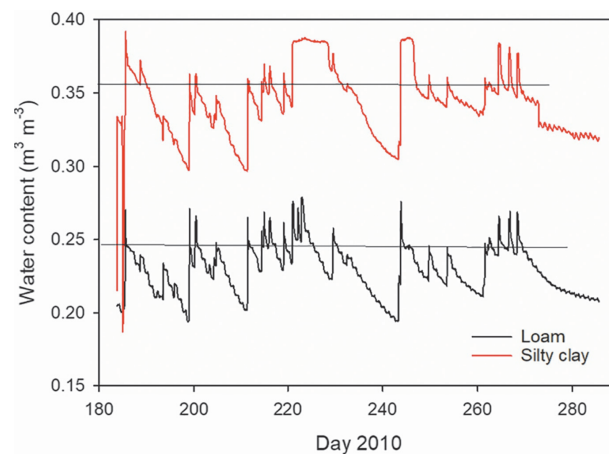


Fig. 1 Upper soil water cutoff using the procedure of Chandler et al. (2017) for a loam and a silty clay soil. Unpublished data is from the study by Logsdon (2015) for water content 0–15 cm depth.

the roots (Logsdon, 2019). This replenishment through the soil may occur on a diurnal basis: root water extraction at day, and upward water replenishment at night (Cordeiro et al., 2015). In addition, water can move from deeper roots to shallow roots, which then release the water for later uptake, which is called hydraulic lift (Caldwell et al., 1998).

The lower limits of plant available water

Permanent wilting point

The lower limit of plant available water, the permanent wilting point (WP) was introduced at the beginning of the 20th century by Briggs and Shantz (1912). From experimental data with wheat plants, they found that the *wilting coefficient* could be accurately estimated by dividing the “moisture equivalent” (water content retained by the soil in opposition to a centrifugal force of 100 g) by the factor 1.84. Nowadays, the most common “definition” of WP is in terms of a matric potential of $-15,000$ hPa. This value is almost without exception used in models and allows accurate predictions. WP corresponds to a condition at which the soil water potential and the soil hydraulic conductivity are so low that not even a minimal amount of water can be absorbed by the root system. Although a dependence of WP on crop characteristics, soil hydraulic properties, and atmospheric conditions is to be expected, the main reason for using a fixed matric potential for WP is the experimental difficulty of its measurement, as well as the relative irrelevance of its exact determination, at least for agronomic applications (Savage et al., 1996).

An indirect method to determine wilting point is to see where soil water content levels off at the lower end, leaving residual water held in disconnected pore spaces. This requires automated soil water sensor data over wet and dry times of the season.

WP measurements directly from plants are rarely performed, but when performed usually follow the Briggs and Shantz (1912) and Taylor and Ashcroft (1972) protocol. The method consists of growing plants in small pots, with evaporation from the soil surface prevented by a seal of wax over the soil surface. After establishment of the plants, the pots are kept immersed in a water bath to avoid condensation of the soil moisture on the pot walls. The plant is allowed to grow until it wilts and then it is allowed to recover in a dark, humid chamber. If it recovers, the procedure is repeated. If the plant does not recover in 24 h, the corresponding soil water content is considered to be the WP.

Limiting hydraulic conditions

Whereas WP defines a water content that is so low that plant growth and survival is no longer possible, transpiration will be initially limited when soils are wetter. This limiting water content is sometimes referred to as “trigger point” in irrigation management, or as limiting or critical water content (LP) in soil physics. It is the lower limit of readily available water (RAW), corresponding to the water content at which the overall hydraulic resistance of the pathway between soil and leaf becomes limiting to the transpiration rate. Consequently, at LP crop transpiration turns from its sink-limited value in the constant rate phase (determined by atmospheric properties like vapor pressure deficit, wind speed, temperature, and incoming radiation) into a source (=soil) limited value in the falling rate phase. The source-limited rate is primarily determined by soil, root, and xylem hydraulic properties. In particular, the soil hydraulic conductivity is highly dependent on soil water content and matric potential.

The correct determination of LP is important to agronomic (irrigation) management, as well as for hydrological modeling. It is more important than WP because soil water content reaches values below the limiting water content much more frequently than wilting point. A process-based determination of LP requires a full hydraulic parameterization of the soil-plant-atmosphere pathway (Cowan, 1965), and relies on numerical solutions requiring detailed information on hydraulic retentivity and conductivity, root system geometry, and root and xylem hydraulics (Javaux et al., 2013; De Jong Van Lier et al., 2013).

For applications requiring a less detailed approach, simple (piecewise linear) models have been proposed to describe the relative transpiration T_a/T_p as a function of water content, normally including FC, LP, and WP as threshold values. A commonly applied model of this type is the transpiration reduction function used in the model Aquacrop (Fig. 2).

In the FAO methodology, LP is defined as a function of FC and WP and a depletion factor p according to

$$LP = WP + (1 - p)(FC - WP) \quad (5)$$

Values for p are given by Doorenbos et al. (1986) for four crop groups with decreasing sensitivity to drought. Values of p vary between 0.125 and 0.875 as a function of the atmospheric demand expressed as potential transpiration.

Unlike water content, matric potential and its gradient can be linked directly to soil water movement and root water extraction. Therefore, it makes more sense to express threshold values of soil water availability (FC, LP, and WP) in terms of water potential. The widely used reduction function proposed by Feddes et al. (1978) is included in important hydrological models like Hydrus (Šimůnek et al., 2016) and SWAP (Kroes et al., 2017). It describes the root water uptake (RWU) reduction factor α as a piecewise linear function of matric potential (Fig. 3). Factor α is defined as

$$\alpha = \frac{S_a}{S_p} \Leftrightarrow S_a = \alpha S_p \quad (6)$$

where S_p ($\text{m}^3 \text{m}^{-3} \text{d}^{-1}$) and S_a ($\text{m}^3 \text{m}^{-3} \text{d}^{-1}$) are the potential and actual root water uptake rate, respectively. The resulting transpiration rate is obtained by integrating S_a over the rooting zone. This integration can be performed using weighting factors,

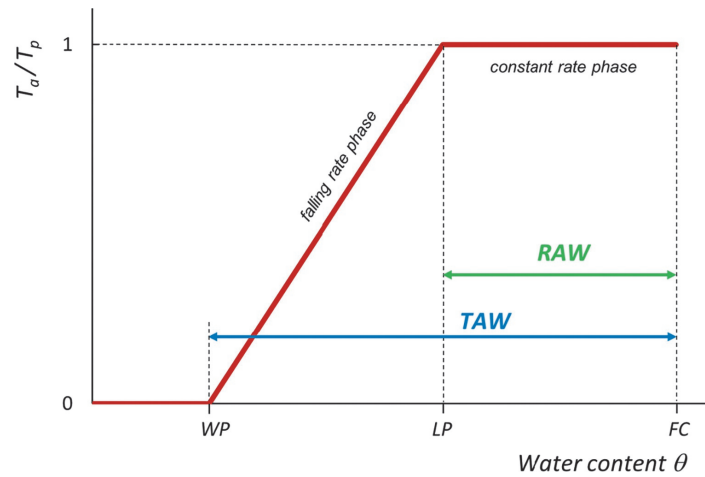


Fig. 2 The transpiration reduction function used in Aquacrop / FAO, defining relative transpiration T_a/T_p as a function of water contents FC, LP and WP, showing Total Available Water (TAW) and Readily Available Water (RAW).

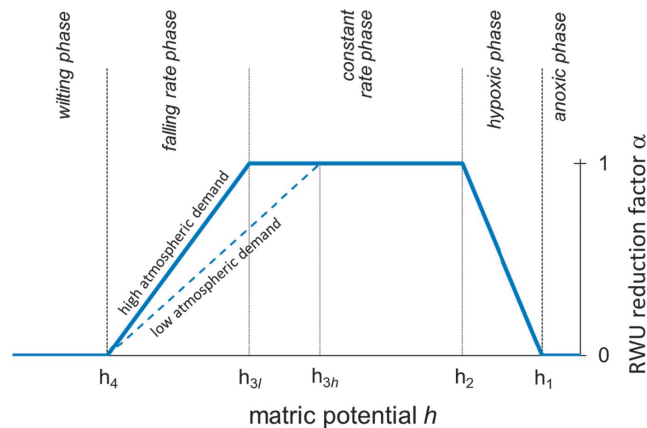


Fig. 3 The Feddes reduction function, defining the RWU reduction factor α as a function of matric potential, showing the five phases of RWU.

usually the root length density. Matric potential h_3 delimits the constant and falling rate phase and corresponds to LP, with values between -100 and -1000 hPa depending on crop species, according to [Taylor and Ashcroft \(1972\)](#).

The measurement of LP or h_3 is based on the determination of the water content at the onset of relative transpiration $T_a/T_p < 1$ (Figs. 2 and 3). By evaluating different plant species, phenological stages, soil types, and atmospheric conditions, we may determine how LP depends on these factors. Usually, plants are allowed to establish in pots in greenhouse experiments until a predefined development stage. At this point, the pots are split into two groups, one continuing to be watered, the other one without a water supply. Evaporation is minimized by sealing the soil surface with wax or a layer of coarse sand, and pots are weighed daily to determine the soil water content and the average transpiration rate. The transpiration rate of the irrigated plants is taken as potential transpiration T_p , while the transpiration rate of the non-irrigated plants is the actual transpiration T_a . A daily evaluation of the relative transpiration T_a/T_p allows finding when $T_a/T_p < 1$, the transition from the constant to the falling phase (Figs. 2 and 3), corresponding to LP. The accuracy of the determination depends on the homogeneity of soil characteristics (physical, chemical, and biological), root proliferation, and leaf area in both treatments (irrigated and non-irrigated). As LP depends on atmospheric conditions, similar atmospheric condition around the onset of $T_a/T_p < 1$ also makes the determination of LP more accurate.

Other limitations to plant water availability

In addition to the limits of plant available water related to soil hydraulic properties, as discussed above, other physical limitations to the access of plants to soil water may be present. In particular, those limitations caused by reduced aeration in wet soil or by too high a penetration resistance for plant roots in the dry soil have been considered. An attempt to integrate these limiting factors of different nature into one indicator was proposed by [Letey \(1985\)](#). The resulting Least Limiting Water Range (LLWR) is constrained at the wet side by drainage (FC) or aeration requirements, whereas the dry side may be controlled by hydraulic properties (WP) or mechanical limitations to root growth.

Plant available water: The role of soil texture and structure

Plant available water and soil texture

Plant available water is defined by FC and WP, and both can be assessed by a matric potential. The combination with existing pedotransfer functions based on soil texture (sand, silt, and clay content) allows predictions of plant available water for different texture classes. Fig. 4 shows an example of the relation between texture class and FC, WP, and available water. The vertical distance between FC and WP equals the plant available water. In this example, it is largest for (silt) loam texture. It should be remembered that soil hydraulic properties like FC and WP are not only determined by texture, but also by structure. Therefore, Fig. 4 may give a rough indication but cannot be used to assess plant available water in any soil and under any circumstances.

Plant available water and soil structure: The role of soil management

Together with soil texture, soil structure determines soil hydraulic properties. In its turn, soil structure is determined by pedogenetic factors in the long term, and by soil management in the short term. Soil management may affect soil structure in many ways. The most important processes regarding plant available water are compaction and loosening.

Compaction is caused by heavy machinery and results in the reduction of total soil porosity, but mainly macroporosity. It usually affects the surface layer of a soil, impacting the infiltrability, rootability, aeration, and water redistribution. Compaction has a negative impact on soil quality and plant growth, but its direct impact on plant available water is less straightforward.

The loosening and mixing of surface soil is one objective of tillage. Although tillage effects on soil structure can be of short duration, the mixing and aeration of carbon-rich topsoil will lead to enhanced oxidation of organic matter, causing a reduction of soil carbon content. Humic substances are important in stabilizing soil structure; hence the loss of organic C may lead to a deterioration of soil structure, with consequences for plant available water too. The mismanagement of organic C may thus lead to compaction and a decrease of plant available water.

The use of plant available water

Use of TAW and RAW in irrigation management

The limiting soil water content LP defines the onset of drought stress, and as such it is essential for the proper optimization of irrigation decisions based on the soil water content status. As mentioned, for irrigation purposes the depletion threshold p is an empirical factor that ranges between 0 and 1, and it is listed for crop groups and specific atmospheric water demands, e.g., in the FAO irrigation guidelines. According to these guidelines, p together with FC and WP determines LP (Eq. 5). However, LP represents the limiting soil hydraulic conditions, therefore, its proper estimation should consider soil hydraulic properties (retention and conductivity), root density distribution, and the atmospheric water demand. The correct determination of upper and lower limits of TAW and RAW is important, and mismatches may lead to erroneous irrigation scheduling, and consequently, reduced crop yields due to under or over-irrigation.

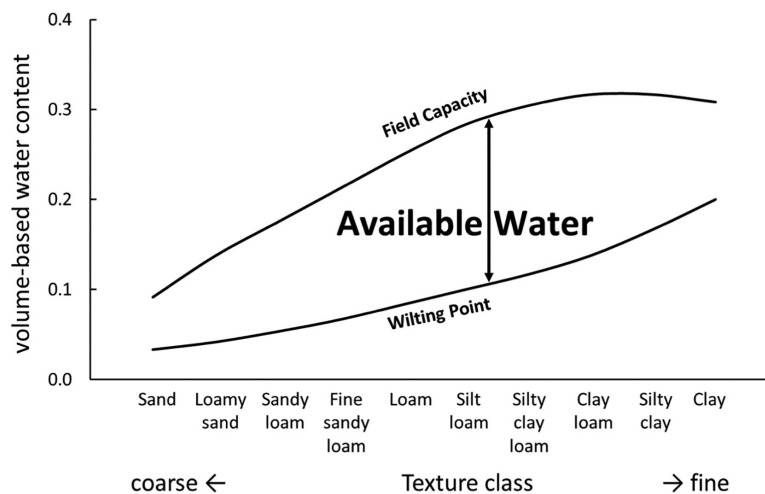


Fig. 4 Field capacity, wilting point and plant available water as a function of texture class. Adapted from Ohio State University (2005) *Ohio Agronomy*. 14th edn. Ohio State University.

Use of TAW and RAW in hydrological models

The goals of most agro-hydrological models are similar, i.e., to predict the soil water balance and crop yield. However, they may differ considerably in the way they describe and parameterize the system. Process-based models solve the water flow with the Richardson-Richards equation coupled with a sink term to account for root water uptake. Although these models represent the state-of-the-art in vadose zone hydrology models, they require detailed soil hydraulic parameterization (water retention and hydraulic conductivity). On the other hand, the “bucket” or compartment models use much simpler water transfer calculations. These models, such as Aquacrop (FAO), implement TAW in their water balance computing routine, treating the soil as a reservoir (bucket analog) with upper (FC) and lower (WP) limits of available water. These water balance “bucket” models are extensively used to describe land surface processes at large scales, especially when coupled with General Circulation Models (GCMs). Applications of TAW to evaluate soil water balance at a regional scale requires resetting its input parameters, moving from localized measurement to mapping, given that the larger the scale the greater the level of uncertainty.

Use of TAW and RAW as ecological indicators

Given the simplified description of the soil water balance provided by TAW and RAW, these indicators are commonly integrated with lumped models to describe the hydrological and ecological functioning at ecosystem scales. Estimating the maximum amount of water a soil can hold, aligned to high-temporal resolution data on soil water content status retrieved from remote sensing, is an important strategy to identify “pools” of potential transpirable water across ecosystems and how they shift over the temporal scale. This information is required to investigate the interactions of soil water and climate, especially the feedback between land surface evapotranspiration and precipitation, with consequent impacts on net primary productivity of terrestrial ecosystems and biogeochemical cycles (e.g., nitrification and CO₂). The main natural factors controlling soil water storage gradients across ecosystems are those related to pedological processes (e.g., parent material, weathering agents and local climate), resulting in soil groups with different thickness and texture, which in part control the vegetation dynamic and its feedback exerted on soil water fluxes.

Concluding remarks

Indicators of plant available water commonly consider upper and lower bounds based on soil hydraulic behavior alone. However, plant water use and development are also affected by other factors, most importantly by mechanical and aeration constraints. Soil water availability and gas diffusion rate (e.g., O₂ and CO₂) within the soil profile are dynamic processes that vary greatly in space and time according to climate variables, plant characteristics, soil hydraulic properties, and soil water status. Single values of water content or water potential alone cannot portray the adequate or critical conditions to which plants are exposed.

All the static limits (FC, WP, LP) are actually dynamic and depend on soil hydraulic properties, root distribution and extent, depth to the water table, as well as soil compaction and aeration status of the soil. Accounting for these factors will improve our prediction of plant water use and irrigation timing, even if just involving empirical adjustments to the simplified linear approaches.

References

- Briggs LJ and Shantz HL (1912) The wilting coefficient and its indirect determination. *Botanical Gazette* 53(1): 20–37. <https://doi.org/10.1086/330708>.
- Caldwell MM, Dawson TE, and Richards JH (1998) Hydraulic lift: Consequences of water efflux from the roots of plants. *Oecologia* 113: 151–161.
- Chandler DG, Seyfried MS, McNamara JP, and Hwang K (2017) Inference of soil hydrologic parameters from electronic soil moisture records. *Frontiers in Earth Science* 5: 25. <https://doi.org/10.3389/feart.2017.00025>.
- Cordeiro MRC, Krahn V, Ranjan RS, and Sager S (2015) Water table contribution and diurnal water redistribution within the corn root zone. *CBE Journal* 57. <https://doi.org/10.7451/CBE.2015.57.1.39>.
- Cowan IR (1965) Transport of water in the soil-plant-atmosphere system. *Journal of Applied Ecology* 2(1): 221–239. <https://doi.org/10.2307/2401706>.
- Da Silva AP and Kay BD (2004) Linking process capability analysis and least limiting water range for assessing soil physical quality. *Soil and Tillage Research* 79: 167–174.
- De Jong Van Lier Q (2017) Field capacity, a valid upper limit of crop available water? *Agricultural Water Management* 193: 214–220. <https://doi.org/10.1016/j.agwat.2017.08.017>.
- De Jong Van Lier Q, Van Dam JC, Durigon A, Santos MAD, and Metselaar K (2013) Modeling water potentials and flows in the soil–plant system comparing hydraulic resistances and transpiration reduction functions. *Vadose Zone Journal* 12(3). <https://doi.org/10.2136/vzj2013.02.0039>.
- Doorenbos J, Kassam AH, Branscheid V and Bentvelsen CLM (1986) *Yield Response to Water*. 33. <http://agris.fao.org/agris-search/search.do?recordID=SO2005100014> (Accessed 24 April, 2020).
- Feddes RA, Kowalik PJ, and Zaradny H (1978) Simulation of Field Water Use and Crop Yield. <https://www.cabdirect.org/cabdirect/abstract/19800705674>. (Accessed 23 April, 2020).
- Javaux M, Couvreur V, Vanderborght J, and Vereecken H (2013) Root water uptake: From three-dimensional biophysical processes to macroscopic modeling approaches. *Vadose Zone Journal* 12(4). <https://doi.org/10.2136/vzj2013.02.0042>.
- Kroes JG, Van Dam JC, Bartholomeus RP, Groenendijk P, Heinen M, et al. (2017) SWAP Version 4; Theory Description and User Manual. <https://edepot.wur.nl/416321>.
- Letej J (1985) Relationship between soil physical properties and crop production. In: Stewart BA (ed.) *Advances in Soil Science*, pp. 277–294. New York, NY: Springer.
- Logsdon SD (2015) Event- and site-specific soil wetting and seasonal change in amount of soil water. *Soil Science Society of America Journal* 79(3): 730–741. <https://doi.org/10.2136/sssaj2014.08.0327>.
- Logsdon S (2019) Should upper limit of available water be based on field capacity? *Agrosystems, Geosciences & Environment* 2(1): 190066. <https://doi.org/10.2134/age2019.08.0066>.
- Meyer PD and Gee GW (1999) Flux-based estimation of field capacity. *Journal of Geotechnical and Geoenvironmental Engineering* 125(7): 595–599. [https://doi.org/10.1061/\(ASCE\)1090-0241\(1999\)125\(7\)595](https://doi.org/10.1061/(ASCE)1090-0241(1999)125(7)595).

- Reynolds WD (2018) An analytic description of field capacity and its application in crop production. *Geoderma* 326: 56–67. <https://doi.org/10.1016/j.geoderma.2018.04.007>.
- Savage MJ, Ritchie JT, Bland WL, and Dugas WA (1996) Lower limit of soil water availability. *Agronomy Journal* 88(4): 644–651. <https://doi.org/10.2134/agronj1996.00021962008800040024x>.
- Šimůnek J, van Genuchten MT, and Šejna M (2016) Recent developments and applications of the Hydrus computer software packages. *Vadose Zone Journal* 15(7): vzj2016.04.0033. <https://doi.org/10.2136/vzj2016.04.0033>.
- Taylor SA and Ashcroft GL (1972) *Physical Edaphology; The Physics of Irrigated and Nonirrigated Soils*. San Francisco: W.H. Freeman.

Plant-water relations

Ian C Dodd^a, Andrew D Hiron^b, and Jaime Puértolas^c, ^aLancaster Environment Centre, Lancaster University, Lancaster, United Kingdom; ^bUniversity Centre Myerscough, Preston, United Kingdom; ^cDepartment of Botany and Plant Ecology and Physiology, University of La Laguna, San Cristóbal de La Laguna, Spain

© 2023 Elsevier Ltd. All rights reserved.

This is an update of C Gimenez, M Gallardo, RB Thompson, Plant-Water Relations, Editor(s): D Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 231–238, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00459-8>.

Introduction	516
Water in plants	517
Plant water status indicators	518
Direct indicators of plant water status	518
Relative water content	518
Water potential and its components	518
Measuring plant water potential	518
Indirect indicators of plant water status	519
Soil water status	519
Stomatal conductance	519
Sap flow	519
Stem diameter variations	520
Leaf or canopy temperature and spectral reflectance techniques	520
Applying plant-water relations in irrigation scheduling	521
Physiological responses to plant water deficits	522
Expansive growth	523
Stomatal responses	524
Root signals and their application in agriculture	524
Future opportunities	525
References	525

Abstract

Herbaceous plants can lose many times their own weight in water daily. Such transpirational losses may lead to plant water deficit, which constrains leaf expansion and photosynthesis. We review water transport in the soil-plant-atmosphere continuum, considering driving forces and constraints. Measurements that directly or indirectly indicate plant water status are introduced, to underpin their use in crop irrigation scheduling. In planta mechanisms (hydraulic and chemical signaling) regulating plant physiological responses provide opportunities to alter plant genetics or management (including irrigation) to increase plant water use efficiency, improve species selection in managed systems, and better understand vegetation dynamics of natural ecosystems.

Key points

- Describe the driving forces for water movement in the soil-plant-atmosphere continuum.
- Recognize how to measure plant water status using various techniques.
- Appreciate how plant water status measurements can be used in irrigation scheduling.
- Understand how hydraulic and chemical signals regulate plant responses to water deficit.

Introduction

Water constitutes the largest single chemical component of plants, yet daily water losses via plant transpiration can greatly exceed the volume of water within herbaceous plants. When uptake of soil water by plant roots is less than transpiration requirements (determined by the evaporative demand of the atmosphere), plant water deficit ultimately occurs. Various measurements of plant water status are useful for different disciplines, from irrigation scheduling to minimize or regulate crop water deficit to fundamental studies assessing the impact of climate change on natural ecosystems. Decreased plant water status affects numerous plant processes including cell expansion (which is the most sensitive to water deficit), stomatal closure and photosynthesis. Soil drying decreases leaf water potential and alters root-to-shoot chemical signaling, which regulates these processes. An enhanced understanding of these processes provides opportunities for crop management to minimize environmental impacts of agriculture while maintaining food production. Moreover, such knowledge helps refine models forecasting the impact of climate change on natural plant communities, and supports prediction of the success of tree-planting efforts to mitigate the impacts of climate change.

Water in plants

Water normally constitutes more than 90% of fresh weight in herbaceous plants, and more than 50% in woody plants. Most of this water (60–90%) is located within the cells, with the remainder in the cell walls and intercellular spaces. The water in cell walls and intercellular spaces forms a continuum (the apoplast) throughout the plant.

The total volume of water within a plant is very small in relation to the total volume of water transpired during its lifetime. Even on a daily basis, the volume of water within plants is insufficient to account for daily transpiration requirements on a warm, sunny day. Stomatal opening to allow CO₂ absorption for photosynthesis (A) allows water loss as transpiration (E) that can be considered as the 'cost' that plants incur, with their ratio (A/E) defining instantaneous water use efficiency. To maximize photosynthesis and provide carbohydrates for sustaining vegetative and reproductive growth, stomata must remain open for as long as possible during daylight periods, but this also maximizes transpiration.

Water has many functions within plants:

- maintaining cell and tissue turgor to allow cell enlargement (growth) and structural integrity.
- transport of mineral and organic solutes within (xylem and phloem) cells throughout the plant.
- participation in metabolic activities including CO₂ reduction in photosynthesis.
- transpirational cooling to avoid excessive daytime heating from incoming solar radiation.

Plant water status determines its growth and development, crop productivity in agricultural systems, and plant survival in natural systems. Plant water status either directly or indirectly affects almost every plant physiological process, although the sensitivity of various processes varies considerably. It is determined by various atmospheric, plant, and soil factors:

- Soil water availability
- Root capacity to absorb water
- Tissue hydraulic conductance (the ability to transport absorbed water)
- Stomatal responses that regulate transpiration (water loss to the atmosphere)
- Atmospheric evaporative demand (determined by radiation, humidity, temperature, wind)

Plant water status is commonly characterized by its water potential (Ψ), which is a measure of its free energy status that determines the direction of water movement within the soil-plant-atmosphere system. Ψ represents the work involved in moving 1 mole of pure water from a selected point within the plant (or soil) to a reference point at the same temperature and at atmospheric pressure. Ψ varies from zero at the reference point to negative values within the plant and soil. It is normally measured in units of pressure, with megapascals (MPa) being conventional.

Leaf water potential (Ψ_{leaf}) is determined by water loss to the atmosphere by transpiration (E) and water absorption from the soil, which depends on soil water potential (Ψ_{soil}) and the combined resistance to water movement within the roots and shoots (r). Eq. (1) describes these relationships:

$$\Psi_{\text{leaf}} = \Psi_{\text{soil}} - E/r \quad (1)$$

In saturated soils, with all the pore spaces filled with water, Ψ_{soil} approaches 0 MPa (the Ψ of pure water with no stresses). However, solutes (such as salts) in the soil solution decrease its osmotic potential and thus lower Ψ_{soil} . As water drains from the pore space due to gravity, Ψ_{soil} decreases further. In well-drained sandy soils with larger pore spaces, weak matric forces holding water molecules to the soil particles decrease Ψ_{soil} (–0.01 MPa is typical) (see Capillarity). In clay soils with smaller pore spaces, stronger matric forces result in lower Ψ_{soil} after drainage (typically –0.03 MPa). Evaporation from soil surfaces also depletes water, which decreases Ψ_{soil} . However, in vegetated soils, root water absorption to support transpiration has the greatest effect in lowering Ψ_{soil} . Together, evaporation and transpiration comprise evapotranspiration (ET). Along with soil texture and pore structure, ET determines changes in soil water content and hence, the rate of Ψ_{soil} decline in drying soils.

For transpiration to occur, Ψ_{leaf} must be lower (more negative) than Ψ_{soil} . When transpiration ceases (for instance, during night-time), Ψ_{leaf} tends to equilibrate with Ψ_{soil} . In the absence of significant night-time transpiration, leaf water potential by dawn (or after sufficient time to equilibrate with null transpiration) is considered a proxy Ψ_{soil} . Ψ_{leaf} tends to equilibrate with that of the wettest part of the rootzone, such that a few roots in moist deeper soil layers can take up sufficient water to sustain leaf water status.

Increasing evaporative demand (quantified by vapor pressure deficit, VPD) increases transpiration and hence (if Ψ_{soil} remains constant) decreases Ψ_{leaf} . Some plant species close their stomata to limit transpiration at greater VPDs to prevent damaging leaf water deficits. Greater evaporative demands during the night coupled with incomplete stomatal closure prevent Ψ_{leaf} completely equilibrating with Ψ_{soil} and cause some nocturnal transpiration (typically <10% of day-time transpiration). Inter- and intra-specific variation in the ability to limit transpiration at high day-time VPDs and at night have been proposed as plant breeding targets for water-limited environments (Coupel-Ledru et al., 2016).

Since stomata close as Ψ_{leaf} decreases, thereby limiting CO₂ uptake, both soil and atmospheric dryness restrict plant carbon gain. Global warming increases air temperatures and evaporative demand, and consequently ET. As both soil drying and greater ET decrease Ψ_{leaf} and, in turn, stomatal aperture, models predict that global temperature increases should decrease plant carbon fixation (Grossiord et al., 2020).

During daylight periods, water absorption lags behind transpiration due to the high resistance to water flow from soil into root xylem tissue. As atmospheric evaporative demand increases during the morning, transpiration increases, which lowers the water potential of cells from which water is evaporating. Within the plant, water moves from non-evaporating parenchyma cells inside the leaf (with a higher Ψ), toward the evaporating cells (with a lower Ψ), establishing a Ψ gradient. This gradient is transmitted throughout the plant-soil system, enabling continuous water movement. Lower evaporative demand in the late afternoon decreases transpiration. However, water uptake by roots continues until parenchyma cells fully rehydrate. When Ψ_{leaf} equals Ψ_{soil} , which usually occurs during the night, plant and soil water status are in equilibrium, and root water uptake ceases.

Plant water status indicators

Direct indicators of plant water status

Relative water content

Plant water content is the fraction of the total fresh tissue weight that is water, determined on either a fresh- or dry-weight basis. Since this varies considerably between different plant tissues and organs (that vary in tissue density), it is usually normalized according to maximal water content. This relative water content (RWC) is the water content of plant tissue or a plant organ relative to its maximum water content when it is fully hydrated (i.e., saturated). RWC measures the degree of tissue hydration, and is expressed as:

$$\text{RWC} = (\text{FW} - \text{DW}) / (\text{SFW} - \text{DW}) \quad (2)$$

where FW is the fresh weight (grams), DW the dry weight (grams), and SFW is the saturated (or turgid) fresh weight (grams).

Water potential and its components

Plant water status is most commonly measured as water potential (Ψ). Total water potential (Ψ) comprises the osmotic (solute) potential (Ψ_s), pressure potential (Ψ_p), matric potential (Ψ_m), and gravitational potential (Ψ_g) according to:

$$\Psi = \Psi_s + \Psi_p + \Psi_m + \Psi_g \quad (3)$$

Osmotic potential (Ψ_s): results from dissolved solutes in cell sap and is proportional to solute concentration and inversely proportional to cell water volume. In plants, Ψ_s is always negative and decreases as solutes concentrate as the plant dehydrates. Pressure potential (Ψ_p) is a measure of tissue turgor produced by the diffusion of water into the protoplast (mainly the vacuoles) of cells enclosed by largely inelastic cell walls. This produces an outward force against the cell walls that is opposite to the inward movement of water. In living cells, Ψ_p is typically positive (except at plasmolysis when $\Psi_p = 0$). However, in xylem conduits where transpiration maintains tension within the water column, Ψ_p is negative.

Matric potential (Ψ_m): arises from the action, on water, of electrostatic forces of attraction associated with cell wall and colloidal surfaces, and of capillary forces associated with narrow transport conduits. In plants it is considered to be negligible but is the dominant component of soil water potential. For the same volume of soil water, Ψ_m is typically lower in soils with smaller pores (clay soils). Thus, root water extraction from these soils requires even lower plant water potentials.

Gravitational potential (Ψ_g): results from gravitational forces acting on the water within plants and is only significant in very tall trees as its gradient is 0.01 MPa m^{-1} .

In most situations, total plant water potential is considered as the sum of the pressure potential (Ψ_p) and osmotic potential (Ψ_s). As both these components depend on tissue water content, RWC is related to Ψ_s and Ψ_p , and thus Ψ (Fig. 1). In a fully hydrated plant, RWC equals 1, Ψ_p is positive, Ψ_s is negative, and Ψ (the sum of Ψ_p and Ψ_s) equals zero. As RWC progressively declines, both Ψ_p and Ψ_s decline, and consequently Ψ becomes more negative. When RWC has declined sufficiently for Ψ_p to equal zero, the plant loses turgor and wilts.

Measuring plant water potential

Several different methods are available to measure plant water potential. The pressure-chamber method can rapidly measure Ψ of whole leaves or stems in the field. Traditional thermocouple psychrometry and hygrometry can accurately measure small tissue samples (e.g., a leaf disc), but require very stable environmental conditions and are unsuitable for field use. However, stem and leaf psychrometers are able to overcome these problems and are a promising tool for continuous in situ measurements of water potential (Dixon and Tyree, 1984). Pressure potential (Ψ_p) can only be measured directly by the pressure microprobe, which is inserted into cell protoplasm, but this technique is generally restricted to laboratory measurements. However, the development of pressure potential probes coupled to magnetic devices allows continuous measurements of leaf turgor in several species (Zimmermann et al., 2013), with potential uses for irrigation scheduling.

In practice, both Ψ_{leaf} and Ψ_s are routinely measured, with Ψ_p calculated as the difference between Ψ_{leaf} and Ψ_s . Osmotic potential (Ψ_s) is commonly measured by either thermocouple psychrometry or hygrometry after freezing and thawing the sample to break cell membranes and reduce Ψ_p to zero. The pressure chamber can also be used to construct a pressure–volume curve, using the relationships illustrated in Fig. 1 to determine Ψ_s from coupled measurements of Ψ_{leaf} and relative water content during leaf dehydration. These pressure–volume curves also yield other parameters relevant to plant cell water relations such as RWC, Ψ_{leaf} at

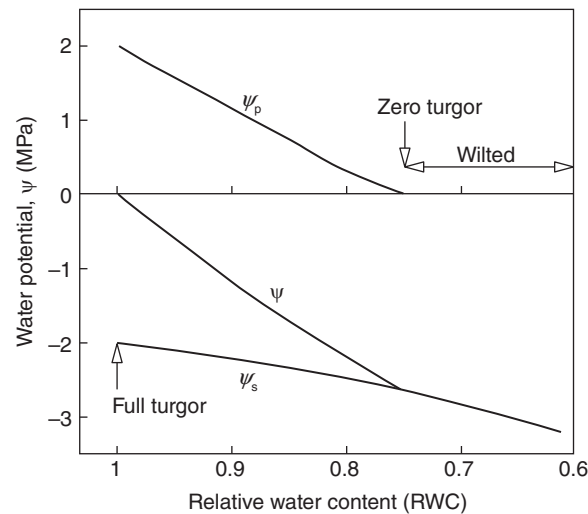


Fig. 1 Höfler-Thoday diagram illustrating the relationships between total water potential (Ψ), pressure potential (Ψ_p), osmotic potential (Ψ_s), and relative water content (RWC) as a cell or tissue loses water from a fully turgid state. Full turgor, and wilting are indicated by arrows. Adapted from Jones HG (2014) *Plants and Microclimate*. Cambridge: Cambridge University Press, with permission.

turgor loss point (i.e., when $\Psi_p = 0$), Ψ_s at full turgor, or bulk elasticity modulus (ϵ) or the rate of change of Ψ_p with change in cell volume (V) in turgid cells ($\epsilon = V \cdot \Delta\Psi_p / \Delta V$).

Indirect indicators of plant water status

Soil water status

Soil drying is easily quantified using a range of sensors that determine either volumetric soil water content (e.g., time domain reflectometry (TDR) probes that measure soil dielectric permittivity, and apply an empirical equation to convert it to water content) or soil water potential, Ψ_{soil} (e.g., tensiometers). The relationship between these variables is known as the soil moisture release curve, with soil texture, pore structure and organic matter content determining how much water the soil can hold at a given water potential. Although variation in Ψ_{soil} determines plant physiological responses, it is often more convenient to measure these responses directly, as Ψ_{soil} varies spatially and it can be difficult to know (especially in large trees) where in the soil the plant is extracting water from, and thus where sensors should be placed. This problem is compounded in large orchards, where it is impossible to instrument every tree.

Stomatal conductance

Stomatal conductance (g_s) depends on the number of stomata per unit leaf area (stomatal density) and the degree of stomatal opening. In well-watered plants, larger g_s should increase transpiration thereby decreasing Ψ_{leaf} . Stomatal closure can prevent further decreases in Ψ_{leaf} by reducing transpiration. Nevertheless, decreased Ψ_{leaf} can induce stomatal closure as the soil dries, thereby decreasing g_s . Thus, g_s and Ψ_{leaf} are intimately connected, as discussed more fully in the section Stomatal Responses. Both dynamic and steady-state diffusion porometers, and infra-red gas analyzers (IRGAs) can rapidly measure stomatal conductance, although this variable changes rapidly throughout the day as environmental conditions (e.g., evaporative demand) fluctuate.

Sap flow

Sap-flow sensors can continuously measure transpiration flow in stems, trunks and branches as the ascent of sap within xylem tissue (Vandegehuchte and Steppe, 2013). Sap flow rates are usually directly related to stomatal opening, increasing during the morning to maximum values and then decreasing with stomatal closure in the late afternoon (Fig. 2a). However, they can be out of phase with transpiration when water is stored in stems or branches (capacitance effects). All sap flow sensors measure changes of temperature along the stem after applying heat at a certain point, based on the principle that the applied heat is transported by the sap moving through the sapwood and the velocity of this transport is positively related with sap flow rate.

The various techniques to measure sap flow rate can be grouped in two approaches. The stem (or trunk) sector heat-balance method continuously electrically heats a section of the entire stem circumference and measures axial and radial heat-loss. The mass flow rate of sap is calculated as a function of the heat dissipated by the ascending sap. Alternatively, the heat-pulse method places the heater centrally and temperature sensor probes (either upstream and downstream sides) inside the trunk and calculates sap velocity as a function of the time required by the flowing sap to transport heat to a particular location.

Sap flow can be measured as sap flow rate (i.e., the rate of water mass change per unit time, usually g h^{-1}) or sap flux density (the rate of water volume per unit sapwood volume and time). While sap flow measurements are all based on the heat-balance

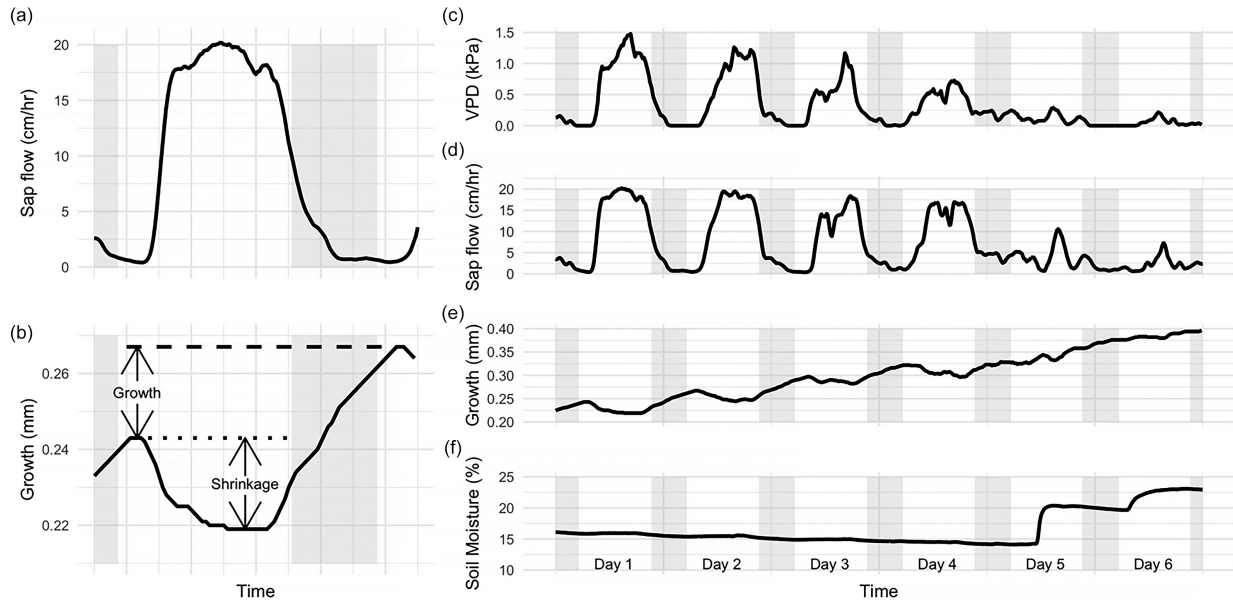


Fig. 2 Changes in environmental conditions and physiological responses of a silver birch (*Betula pendula*) tree. Panels (a) and (b) focus on sap flow and radial stem growth on Day 1. As sap flow increases during the day, the trunk supplements water from the root system with internal water reserves, principally from the phloem parenchyma; causing the stem to shrink. Stem growth occurs during the night when sap flow decreases. Atmospheric evaporative demand or vapor pressure deficit (VPD - c) is the driving force for sap flow (d). Diurnal fluctuations in stem diameter that were prominent on Days 1 and 2 (e) diminish after rain events rehydrate the soil on Days 5 and 6 (f). Deploying sensors that measure these variables can aid irrigation scheduling, particularly if the data can be visualized in real-time via a digital dashboard. Shaded rectangles indicate period between sunset and sunrise. A.D. Hirons, unpublished.

approach, techniques that measure sap flux density are diverse and can use either of the two approaches. Most of these techniques cannot distinguish between low and zero flows and cannot detect negative (i.e., descending) flows, but the most advanced techniques based on the heat-pulse method, such as the heat ratio technique, can detect very small or even reverse flows.

Stem diameter variations

Stem or trunk diameter fluctuates diurnally in transpiring plants, according to their water status (De Swaef et al., 2015). Stem diameter decreases during the morning (Fig. 2b), as rapid transpiration rates are partly supported by some water from stem (mostly phloem) tissue being incorporated into the transpiration stream. Rehydration of this tissue during the afternoon (as transpiration slows with decreased evaporative demand) increases stem diameter. Rehydration continues overnight with continued water uptake from the soil. Commonly, stem diameter has a maximum daily value just before dawn (coinciding with the maximum daily plant water potential), and a minimum daily value in the early afternoon (as plant reaches its minimum daily water potential). The magnitude of the daily stem contraction depends on atmospheric evaporative demand and plant water status, and increases with increasing plant water deficit. Consequently, these variations in stem diameter can very sensitively measure plant water status.

Stem growth is indicated by increased daily maximum stem-diameter values. Decreased growth rates, due to soil drying or other environmental stresses, are apparent as a change in the slope of daily maximum stem-diameter values. In mature plants, daily maximum stem-diameter values tend to be more constant. Since plant age and developmental stage affect the sensitivity of stem diameter fluctuations to evaporative demand, monitoring plant water status in this way usually requires reference baseline monitoring of well-watered controls grown in the same conditions. Sensors incorporating sensitive pressure transducers, known as linearly variable differential transducers (LVDT), connected to data loggers allow stem diameter to be continuously monitored.

Leaf or canopy temperature and spectral reflectance techniques

Transpiration cools the leaf, thus canopy temperature (i.e., average temperature of total leaf surface) is often appreciably less than the air temperature. Partial stomatal closure, in response to increasing evaporative demand or soil drying, increases canopy temperature as less heat is dissipated through the evaporation of water. Since absolute canopy temperature depends on prevailing air temperatures, it is conventional to calculate a Crop Water Stress Index (CWSI):

$$\text{CWSI} = ((T_{\text{canopy}} - T_{\text{air}}) - (T_{\text{wet}} - T_{\text{air}})) / ((T_{\text{dry}} - T_{\text{air}}) - (T_{\text{wet}} - T_{\text{air}})) \quad (4)$$

where T_{air} is air temperature, T_{canopy} is canopy temperature, T_{dry} is the temperature of a dry reference surface and T_{wet} is the temperature of a wet reference surface. Thus, the decrease in canopy temperature caused by transpirational cooling is expressed as a function of the maximum cooling possible in that environment. Typical reference surfaces are leaves sprayed with water (wet reference) or covered in vaseline to prevent transpiration (dry reference), or surrogate leaves such as pieces of card (Fig. 3). Canopy

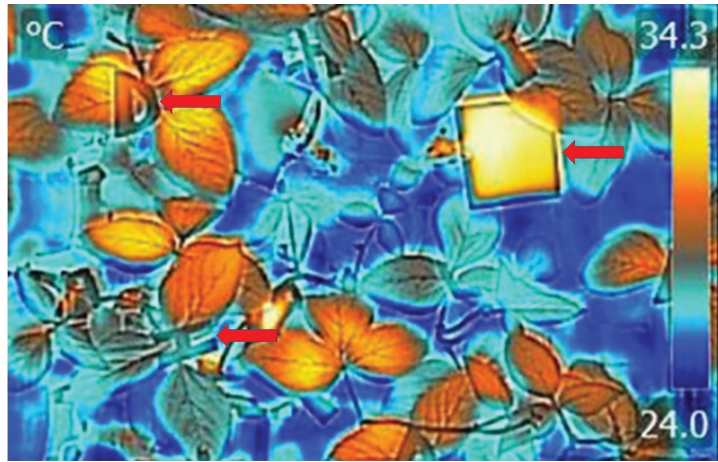


Fig. 3 False color thermal image superimposed to visual image of a soybean (*Glycine max*) canopy comprising individual well-watered plants (light blue-grey colors) and droughted plants (orange-brown colors) grown in pots in a greenhouse, with the bench indicated by dark-blue colors. Surface temperature is provided by the scale at right. The dry reference surface is the arrowed yellow rectangle, while individual well-watered plants and droughted plants are identified by placing red masking tape on the leaves in the letters H and D respectively (also arrowed). J. Puértolas, unpublished.

temperature is most commonly measured with infrared thermometers or thermal infrared cameras, with thermal images easily able to discriminate well-watered or droughted plants.

Reflectance indices are based on statistical techniques to determine the most physiologically relevant wavelengths that are correlated with plant and soil water content (Gerhards et al., 2019), including the

- Green-to-red ratio index (GRI) based on the vegetation reflectance (R) ratio at 550 nm and 670 nm (R_{550}/R_{670}).
- Water band index (WBI) based on the vegetation reflectance (R) ratio at 900 and 970 nm (R_{900}/R_{970}).
- Normalized Difference Vegetation Index (NDVI) based on vegetation reflectance (R) in the near infrared (NIR) and visible red (VR) regions thus:
 (1) $NDVI = (R_{NIR} - R_{VR}) / (R_{NIR} + R_{VR})$
- Water balance index (WABI) based on independent changes in leaf water content (1500 nm) and nonphotochemical quenching photoprotection (531 nm).

Whereas GRI relies on cheaper (multispectral or visible light) sensors, the others require more expensive hyperspectral sensing. As reflectance in specific wavelengths is determined by plant metabolite concentrations (e.g., photosynthetic pigments), ratiometric techniques are needed to discriminate water deficit responses from those influenced by other stresses (e.g., nutrient stress, pest or disease attack). As with all indirect indicators of plant water status, the sensitivity of these indices to water deficit will vary between species and thus determine their utility in irrigation scheduling in managed systems or vegetation monitoring in natural systems.

Applying plant-water relations in irrigation scheduling

These indicators of plant water status may be applied to assist growers in irrigation scheduling or to explore physiological stresses of plants in natural systems. Plant-based measurements offer an advantage over soil-moisture monitoring or estimating crop evapotranspiration (ET) requirements in that the actual unit of production (the plant) is being assessed. Nevertheless, soil or micrometeorology-based calculations determine the irrigation volume to apply, whereas plant water status indicators can indicate when to irrigate, but not the amount. Thus plant-based measurements can supplement other sources of data to inform grower decision-making.

Interpreting plant water status indicators requires comparison with previous data usually attained with fully irrigated plants. Such comparisons establish threshold values (e.g., a maximum allowable daily stem-diameter contraction; a minimum stomatal conductance), with irrigation needed when these values are exceeded. To distinguish variation in plant water status due to insufficient soil water from variation due to atmospheric demand, plant water status data are often normalized according to evaporative demand.

Irrigation timing can be determined by predawn or midday measurements of leaf water potential, based on prior relationships between Ψ_{leaf} and growth in the species of interest (Améglio et al., 1999). The aim is to keep Ψ_{leaf} above threshold values, at which growth is limited, by irrigation. Since Ψ_{leaf} can rapidly vary due to fluctuations in evaporative demand, stem water potential (Ψ_{stem}) is sometimes measured by enclosing leaves (to prevent transpiration) an hour before the measurement. While Ψ_{leaf} and Ψ_{stem} are sensitive measures of plant water status (in species that do not maintain Ψ_{leaf} as the soil dries), they are destructive, time-consuming

and require a skilled operator. These discrete manual measurements can attain only a limited amount of data in comparison to measurements that are readily automated.

Recent developments in sensor and data technology allow plant water status to be continuously monitored, using displacement transducers that measure stem diameter variations (LVDT) and sap-flow sensors (Steppe et al., 2008). Using LVDT sensors, irrigation scheduling can occur automatically when the selected indicator of plant water status reaches a defined limit. Trunk-diameter measurements with LVDT sensors are closely correlated with variations in plant water status in many tree species. Daily stem shrinkage, the difference between daily maximum values and daily minimum values (Fig. 2b), can detect water stress with irrigation protocols involving: (1) selecting the stem shrinkage most suitable for an individual species and particular growth stage, and (2) relating stem shrinkage to reference values of well-watered crops and normalizing them for VPD (3) to determine empirically the relationships between stem shrinkage and irrigation volume required. Placing several LVDT sensors within a crop can account for plant-to-plant variability to ensure average crop response is measured.

Since LVDT sensors do not directly determine the irrigation volume required, sap-flow sensors can be used to determine the amount of water consumed by the plant. Irrigation scheduling can then replenish this water at specified times or time intervals. Nevertheless, the cost of these sensors means that relatively few can be deployed within a crop, thus variability between plants needs to be considered when scaling from single plants to crops. Since identifying individual plants that represent the “average response” of the crop is essential for determining sensor placement, and thus adequate management of sensor-based irrigation, there is increased interest in remote sensing of the entire cropping area to assist irrigation scheduling decisions.

Remote sensing of plants offers advantages that (1) the measurement itself does not perturb plant physiology (e.g., leaves are removed to determine Ψ_{leaf} , inserting sap flow sensors into tree trunks can induce wound responses) and (2) the entire canopy evapotranspiration can be assessed. Multiple platforms (e.g., drones, aircraft, satellites) can image crop canopies, with biomass determinations, evapotranspiration estimates and spectral reflectance indices (indicating crop performance) all providing useful information to the irrigator.

Possibly the most useful technique involves thermal imaging of the crop to calculate a Crop Water Stress Index. Thermal imaging is especially useful across a large cropping area, to identify individual plants (or groups of plants) that may be water-stressed due to local variation in soil properties (e.g., limited soil volume) or failures in irrigation hardware (e.g., blocked drippers). Any irrigation scheduling decisions based on remote sensing require efficient image processing to minimize time between image capture, data processing and providing recommendations to the grower (Khanal et al., 2017).

Physiological responses to plant water deficits

Plant water deficit occurs when plant water status is reduced sufficiently to affect normal plant functioning (e.g., plant growth, stomatal conductance, photosynthetic rate). Absolute values of these variables are poor indicators of water stress, as they are affected by complex interactions of atmospheric, plant (phenological stage and water deficit history) and soil factors, that can modify plant response at a given water potential. Commonly, these values are expressed in relation to those of similar well-irrigated plants.

Plant water deficits are caused by limited root water absorption (e.g., because of dry, salty, cool, or poorly aerated soil) or high atmospheric evaporative demand. Plant water deficits induced solely due to large evaporative demand occur because water potential in transpiring leaves declines as leaf tissues lose water. These deficits depend on atmospheric conditions and are typically greatest at midday and on hot, dry, windy, sunny days. Plant water deficits induced by decreased soil water potential become progressively more intense, until alleviated by rain or irrigation. Since dry atmospheres increase evapotranspiration and therefore the rate of soil drying, they accelerate soil water deficits.

Temporal changes in Ψ_{stem} and sap flow as the soil dries are indicated (Fig. 4). Prolonged soil drying decreases Ψ_{stem} over time until the plant is unable to absorb sufficient water to rehydrate at night. Further declines in Ψ_{stem} promote stomatal closure and sap flow is reduced to conserve water. When plants are unable to recover cell and tissue turgor, which occurs at a threshold soil-water potential value (approximately -1.5 MPa in many herbaceous species, but typically lower, between -2 MPa and -3 MPa, in woody species), the plant permanently wilts.

Severe water deficits can even occur in plants growing in moist soil or hydroponic nutrient solutions, when high evaporative demand causes large transpiration rates that exceed the rate of water uptake by roots. These situations inhibit day-time growth, but can be compensated by night-time growth if night temperatures are not limiting. Since recurrent atmospheric water deficits can reduce yield in intensive greenhouse crop-production systems, fogging systems can be used to prevent them.

Different plant processes show variable sensitivity to water deficits (Table 1). Processes such as cell enlargement, cell wall synthesis, and protein synthesis are very sensitive, being affected by relatively small reductions in Ψ (Hsiao, 1973). Other processes, such as stomatal opening and CO_2 assimilation, generally require larger reductions in Ψ . As plant water deficits intensify, an increasing number of plant processes are affected (Table 1), and there can be complex interactions between them. Two of these processes, expansive growth and stomatal responses, are discussed more fully below.

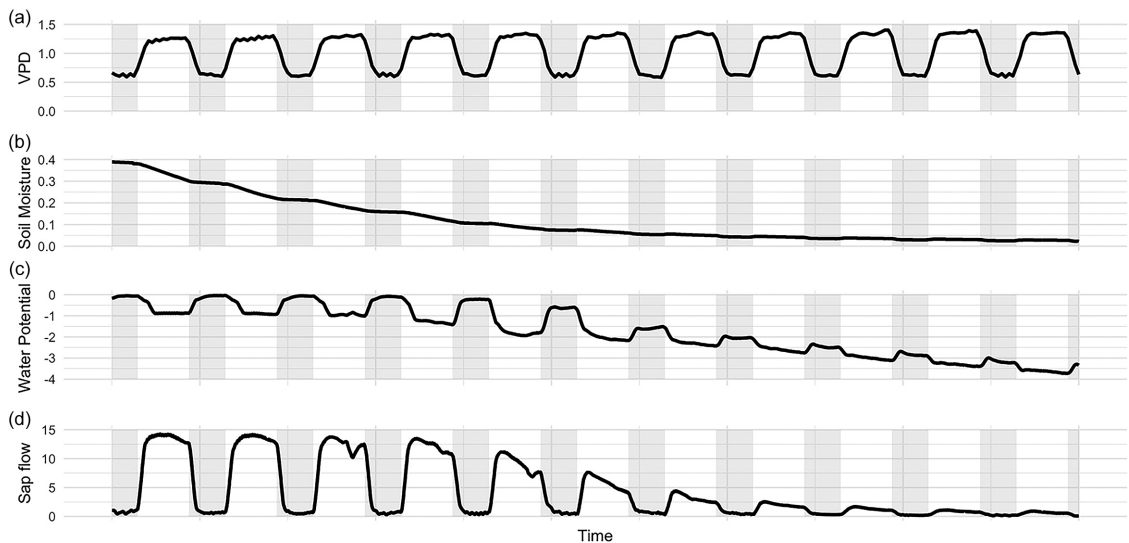


Fig. 4 Changes in environmental conditions and physiological responses of a holm oak (*Quercus ilex*) experiencing a drying cycle within a controlled environment room. Diurnal fluctuations in vapor pressure deficit (VPD—kPa) (a) remain stable during this period. As soil moisture (g g^{-1}) (b) declines, stem water potential (MPa) (c) does not return to well-watered values of approximately 0 MPa overnight. Further declines in water potential (c) promote stomatal closure and the decline in sap flow (d). Shaded rectangles represent the night period. A.D. Hirons, unpublished.

Table 1 Plant process affected by water stress and their sensitivity.

Process affected	Sensitivity to stress			Remarks
	Very sensitive			
	Relatively insensitive			
	Reduction in tissue Ψ required to affect process ^a			
	0 MPa	1 MPa	2 MPa	
Cell enlargement				Fast-growing tissue
Wall synthesis				
Protein synthesis				Fast-growing tissue
ABA accumulation				Depends on species
Stomatal opening				
CO ₂ assimilation				Depends on species
Respiration				
Sugar accumulation				

Length of the horizontal lines represent the range of stress levels within which a process becomes first affected. Dashed lines signify deductions based on more tenuous data.
^aWith Ψ of well-watered plants under mild evaporative demand as the reference point.
Hsiao TC (1973) Plant responses to water stress. *Annual Review of Plant Physiology* **24**: 519–570.

Expansive growth

Since cell enlargement is largely driven by turgor pressure exerted on cell walls, it is especially sensitive to water deficit, which is very important as it reduces leaf area expansion (the photosynthetic surface). Cellular turgor pressure (Ψ_p) is directly related to cell expansion according to the Lockhart equation:

$$dV/dt = m/\Psi_p - \Psi_{th} \tag{5}$$

where the relative change in cell volume (V), with time (t), depends on Ψ_p above a threshold value (Ψ_{th}) below which no cell expansion occurs and cell wall extensibility (m). Increases in cell wall extension during growth are irreversible. Typically, Ψ_{th} is normally only slightly less than Ψ_p values, thus small reductions in Ψ_p can inhibit growth. Both m and Ψ_{th} can vary with previous exposure to water stress and growing conditions. To maintain growth at lower Ψ_p in dry soil stress, plants can adapt by either reducing Ψ_{th} and/or increasing m . Although Eq. (5) provides an elegant conceptual theory, and is now used in functional-structural plant growth models (Coussement et al., 2018), most recent work emphasizes that cell wall biochemistry regulates expansive growth. While diurnal fluctuations in leaf turgor (e.g., caused by fluctuating evaporative demand over the course of a day) seem important in determining leaf elongation rate over minutes to hours, over days turgor seems less important with plant hormones regulating cell wall properties (Munns et al., 2000).

Stomatal responses

In well-watered plants growing at relatively low evaporative demands, light conditions and the low CO₂ partial pressure of substomatal cavities primarily determine stomatal opening. Under these conditions, the stomata are fully open during daylight periods, maximizing photosynthetic rates.

With soil or air drying, as leaf water potential declines, stomatal guard cells lose turgor and their aperture decreases to reduce plant water losses, thus compromising photosynthesis as access to CO₂ is reduced. Nevertheless, stomatal closure can occur even when leaves are fully turgid (Gollan et al., 1986), especially in some (isohydric) species in which partial stomatal closure acts to maintain Ψ_{leaf} homeostasis. In anisohydric species, stomata are less sensitive to Ψ_{leaf} which typically declines. While these stomatal responses are relatively rapid, stomata respond relatively slowly in plants that are recovering from water stress. Although re-watering can rapidly (within minutes) increase Ψ_{leaf} to well-watered values, stomatal conductance can take several hours, or even days, to recover.

In angiosperms, where guard cells are smaller than surrounding epidermal cells, guard cell turgor is usually decoupled from the turgor of surrounding epidermal cells to avoid the mechanical advantage of the large epidermal cells over the smaller guard cells. Regulated solute movement (especially K⁺ ions) to and from the guard cells changes their osmotic potential, increasing or decreasing their turgor compared with epidermal cells at the same water potential. Plant water deficits promote the efflux of K⁺ ions from guard cells, causing loss of cell turgor, which induces stomatal closure (Buckley, 2019). Re-watering induces synchronized changes in K⁺ concentration and guard cell water content prior to stomatal opening. These fluxes of ions require ion transport activation that are mainly regulated by the phytohormone abscisic acid, which is ubiquitously synthesized by plant cells in response to water stress.

Root signals and their application in agriculture

Traditional theory holds that a reduction in soil water potential decreases plant pressure potential, thereby decreasing cell expansion and promoting stomatal closure, a paradigm termed “hydraulic signaling” (Christmann et al., 2013). However, leaf growth inhibition and partial stomatal closure can occur even without changes in pressure potential of leaf cells, with these processes related more strongly to changes in soil water status than to leaf water status. Alternatively, the “chemical signaling” paradigm holds that plants can ‘sense’ drying soil in the root zone and communicate this information to the leaves by a means other than reduced leaf water status. Much work on chemical signaling has focused on the potent anti-transpirant hormone abscisic acid, ABA (Wilkinson and Davies, 2002). Considerable debate has concerned whether hydraulic or chemical signaling dominates *in planta*. These signaling systems are not mutually exclusive, as low plant water status stimulates ABA biosynthesis and enhances stomatal sensitivity to ABA, while increased ABA levels decrease stomatal conductance (thereby maintaining Ψ_{leaf}) and leaf hydraulic conductance while increasing root hydraulic conductance.

While there is broad consensus that ABA is important in regulating stomatal behavior through the mechanisms mentioned in the previous section (since ABA-deficient mutants have high stomatal conductance and low Ψ_{leaf}), two theories have emerged regarding its action *in planta*. “Root-to-shoot ABA signaling” holds that ABA is produced in roots (especially the root tips) in drying soil and is transported to leaves in the xylem sap, where it induces stomatal closure before Ψ_{leaf} decreases. Alternatively, “shoot-to-root ABA signaling” holds that the rate of shoot ABA biosynthesis increases in response to decreased Ψ_{leaf} , or that ABA is redistributed in the leaves toward the guard cells, and that much of this ABA also moves to the roots to increase root ABA accumulation. Given the importance of ABA in preventing plant desiccation, it seems plausible that both signaling systems operate (Li et al., 2018), with their relative magnitude determined by whether the roots or shoots are the first to perceive soil drying or increased evaporative demand respectively.

ABA can induce stomatal closure in isolated epidermis at nanomolar concentrations (typically found in the xylem sap of well-watered plants). Quantitative variation in ABA responses between different plant species might be explained by variation in the ABA concentrations required to induce stomatal closure. For example, while the ABA concentrations in maize xylem sap are sufficient to cause stomatal closure, those in wheat are not. The distribution of xylem-supplied ABA within the leaf seems critical in determining whether the stomata close. In well-watered plants, xylem sap of many species is relatively acidic, causing much of the ABA to be partitioned into mesophyll cells away from the guard cells. As the soil dries, the xylem sap becomes more alkaline, which results in more of the ABA reaching the guard cells. However, soil drying does not alter xylem sap pH in some species yet the stomata close (Sharp and Davies, 2009), suggesting the existence of other chemical root signals.

Other plant hormones such as jasmonates and strigolactones have demonstrable anti-transpirant action (Huntenburg et al., 2022), whereas others such as auxins and cytokinins can antagonize ABA-induced stomatal closure. Soil drying can alter the concentrations of all of these hormones, which are also involved in regulating other processes such as responses to biotic stress (jasmonates), shoot branching or tillering (strigolactones), cell expansion (auxin) and cell division (cytokinins). Thus, soil drying causes multiple changes in shoot phenotype, with the importance of root-to-shoot and shoot-to-root signaling varying between these different hormones.

Another hormone that seems important in regulating plant responses to soil drying and other stresses such as mechanical impedance is the volatile hormone ethylene (Pandey et al., 2021). In compact soil, ethylene accumulates around the root system thereby inhibiting root growth, which may compromise water uptake. Soil drying increases root concentrations of the ethylene precursor 1-amino-cyclopropane carboxylic acid (ACC), and some of this ACC can be transported to the shoot to enhance ethylene synthesis. However, soil drying decreases shoot ethylene production of many species, which may help the stomata to close since ACC (and ethylene) antagonize ABA-induced stomatal closure. Ethylene also promotes leaf senescence which decreases the

duration of leaf photosynthesis. Thus, genotypes that are less able to produce ethylene may produce higher yields when grown in drying soil.

Recognition of bi-directional hormonal signaling may stimulate commercial application of specific agronomic techniques to enhance crop drought resilience. Although the surgical technique of grafting has long been used in horticultural crops to ensure genetically distinct rootstocks and scions with desirable traits (e.g., disease resistant rootstocks, high quality fruit), modern genome editing procedures can increase or decrease plant hormone signaling. Since plant hormones can have antagonistic effects on root and shoot function (e.g., cytokinins restrict root elongation, but enhance shoot growth), grafting could allow different hormone production in roots and shoots. Attenuating root cytokinin production should increase root elongation in drying soils, while augmenting shoot cytokinin production should sustain vegetative growth as the soil dries. Grafting can increase water use efficiency of commercial crops (Cantero-Navarro et al., 2016), linked to changes in root-to-shoot chemical signaling. Increased understanding of the importance of root-shoot communication may provide opportunities to reconcile crop stress resilience and productivity.

Alternatively, recognition of root-to-shoot signaling has stimulated the development of an irrigation technique that deliberately withholds water from part of the rootzone, by watering alternate furrows or just using one irrigation line of two placed either side of the crop row. Such partial rootzone drying (Dodd et al., 2015) aims to create distinct dry and wet parts of the rootzone, with the former stimulating the production of root signals that restrict vegetative growth and stomatal opening (thereby decreasing water loss) and the latter supplying sufficient water to the shoots to maintain turgor. Such root signaling is dynamic, and is promoted by moderate soil drying when the roots in dry soil still contribute to the transpiration stream, but attenuated by prolonged drying when limited sap flow from the roots in dry soil restricts transmission of the signal. Carefully managing the irrigation volumes to the wet and dry rootzones theoretically provides opportunities to enhance stomatal closure to increase crop water use efficiency, or minimize stomatal closure to maintain photosynthesis. Measuring some of the plant water status indicators described above may allow such precision in irrigation management.

Future opportunities

While agriculture has traditionally emphasized maximizing water use (keeping the stomata open to maintain photosynthesis), this strategy is viable only when sufficient water is available in the soil profile, or supplemental irrigation is possible. Since climate change exacerbates existing tensions between agricultural and domestic, ecological and industrial water use, farmers of the future will need to manage limited water supplies. Adopting deficit irrigation techniques such as partial root zone drying (that supply less water than crop evapotranspiration requirements) provide opportunities to maximize biomass allocation to the harvested portion to increase crop quality. Partially decoupling economic productivity from irrigation water requirements offers opportunities to reconcile some of the competing demands for water.

Understanding plant-water relations is not only useful to develop techniques that improve water use efficiency in agriculture. Climate change impacts on natural plant ecosystems and managed forest systems are mostly linked to higher evaporative demand that decreases soil water availability. Predicting their future capacity to fix carbon in new climatic scenarios and making decisions on adaptive forest management and plant species conservation will depend on understanding relationships between water use and carbon fixation. While knowledge of plant water relations has significantly advanced at the individual plant level, approaches applying the remote sensing techniques described in this chapter (Jones, 2018) will be necessary to respond to this new challenge. As ecosystem complexity increases (e.g., from traditional monoculture cropping to regularly spaced agroecological systems to natural plant systems with vegetation of varying height and structure), understanding how species interactions determine plant water relations will become increasingly important.

References

- Améglio T, Archer P, Cohen M, et al. (1999) Significance and limits in the use of predawn leaf water potential for tree irrigation. *Plant and Soil* 207: 155–167.
- Buckley TM (2019) How do stomata respond to water status? *New Phytologist* 224: 21–36.
- Cantero-Navarro E, Romero-Aranda R, Fernandez-Munoz R, et al. (2016) Improving agronomic water use efficiency in tomato by rootstock-mediated hormonal regulation of leaf biomass. *Plant Science* 251: 90–100.
- Christmann A, Grill E, and Huang J (2013) Hydraulic signals in long-distance signaling. *Current Opinion in Plant Biology* 16: 293–300.
- Coupeledru A, Lebon E, Christophe A, et al. (2016) Reduced nighttime transpiration is a relevant breeding target for high water-use efficiency in grapevine. *Proceedings of the National Academy of Sciences USA* 113: 8963–8968.
- Coussement JR, De Swaef T, Lootens P, Roldan-Ruiz I, and Steppe K (2018) Introducing turgor-driven growth dynamics into structural-functional plant models. *Annals of Botany* 121: 849–861.
- De Swaef T, De Schepper V, Vandegehuchte MW, and Steppe K (2015) Stem diameter variations as a versatile research tool in ecophysiology. *Tree Physiology* 35: 1047–1061.
- Dixon MA and Tyree MT (1984) A new stem hygrometer, corrected for temperature gradients and calibrated against the pressure bomb. *Plant, Cell & Environment* 7: 693–697.
- Dodd IC, Puértolas J, Huber K, et al. (2015) The importance of soil drying and re-wetting in crop phytohormonal and nutritional responses to deficit irrigation. *Journal of Experimental Botany* 66: 2239–2252.
- Gerhards M, Schlerf M, Mallick K, and Udelhoven T (2019) Challenges and future perspectives of multi-/hyperspectral thermal infrared remote sensing for crop water-stress detection: A review. *Remote Sensing* 11: 10.
- Gollan T, Munns R, and Passioura JB (1986) Soil water status affects stomatal conductance in fully turgid wheat and sunflower leaves. *Australian Journal of Plant Physiology* 13: 459–464.
- Grossiord C, Buckley TN, Cernusak LA, et al. (2020) Plant responses to rising vapor pressure deficit. *New Phytologist* 226: 1550–1566.

- Hsiao TC (1973) Plant responses to water stress. *Annual Review of Plant Physiology* 24: 519–570.
- Huntenburg K, Puértolas J, de Ollas C, and Dodd IC (2022) Bi-directional, long-distance hormonal signalling of soil water availability. *Physiologia Plantarum* 174: e13697.
- Jones HG (2018) Thermal imaging and infrared sensing in plant ecophysiology. In: Sánchez-Moreiras A and Reigosa M (eds.) *Advances in Plant Ecophysiology Techniques*. Cham: Springer https://doi.org/10.1007/978-3-319-93233-0_8.
- Khanal S, Fulton J, and Shearer S (2017) An overview of current and potential applications of thermal remote sensing in precision agriculture. *Computers and Electronics in Agriculture* 139: 22–32.
- Li W, de Ollas C, and Dodd IC (2018) Long-distance ABA transport can mediate distal tissue responses by affecting local ABA concentrations. *Journal of Integrative Plant Biology* 60: 16–33.
- Munns R, Passioura JB, Guo J, Chazen O, and Cramer GR (2000) Water relations and leaf expansion: Importance of time scale. *Journal of Experimental Botany* 51: 1495–1504.
- Pandey BK, Huang GQ, Bhosale R, et al. (2021) Plant roots sense soil compaction through restricted ethylene diffusion. *Science* 371: 276–280.
- Sharp R and Davies WJ (2009) Variability among species in the apoplastic pH signalling response to drying soils. *Journal of Experimental Botany* 60: 4361–4370.
- Steppe K, De Pauw DJ, and Lemeur R (2008) A step towards new irrigation scheduling strategies using plant-based measurements and mathematical modelling. *Irrigation Science* 26: 505–517.
- Vandegehuchte MW and Steppe K (2013) Sap-flux density measurement methods: working principles and applicability. *Functional Plant Biology* 40: 213–223.
- Wilkinson S and Davies WJ (2002) ABA-based chemical signalling: The co-ordination of responses to stress in plants. *Plant Cell and Environment* 25: 195–210.
- Zimmermann U, Bitter R, Marchiori PER, et al. (2013) A non-invasive plant-based probe for continuous monitoring of water stress in real time: A new tool for irrigation scheduling and deeper insight into drought and salinity stress physiology. *Theoretical and Experimental Plant Physiology* 25: 2–11.

Water use efficiency: A review of spatial and temporal variability

Simone Fatichi^a, Athanasios Paschalis^b, Sara Bonetti^c, Gabriele Manoli^c, and Christoforos Pappas^d, ^aDepartment of Civil and Environmental Engineering, National University of Singapore, Singapore, Singapore; ^bDepartment of Civil and Environmental Engineering, Imperial College London, South Kensington Campus, London, United Kingdom; ^cSchool of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; ^dDepartment of Civil Engineering, University of Patras, Rio Patras, Greece

© 2023 Elsevier Ltd. All rights reserved.

This is an update of M.B. Kirkham, Water-Use Efficiency, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, pp. 315–322, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00441-0>.

Introduction and definitions	528
Relevance of WUE for linking water and carbon cycles	530
WUE observations and simulations	530
Crop water productivity	531
WUE variability and environmental dependencies	533
Stomatal conductance modeling and WUE	537
Trends in water use efficiency	537
Summary	539
References	540
Further reading	541

Abstract

Water Use Efficiency (WUE) expresses the trade-off between carbon assimilation (or subsequent carbon storage) and water release, two concurrent gas fluxes essential for plant functioning. Here, we review metrics that have been introduced to quantify WUE across spatial and temporal scales and corresponding observational methods. The physiological and environmental dependencies of WUE for crops and unmanaged ecosystems are discussed with a quantitative examination of observed and simulated WUE values. A constrained range of WUE across ecosystems and climates is highlighted. However, WUE increased considerably in the past century in response to growing atmospheric CO₂ concentration and it will likely continue along the same trajectory in the near future.

Glossary

WUE Generic water use efficiency [various units].
WUE_l Leaf scale water use efficiency [$\mu\text{mol CO}_2 \text{ mmol}^{-1} \text{ H}_2\text{O}$] or [$\text{gC kg}^{-1} \text{ H}_2\text{O}$].
iWUE Leaf scale intrinsic water use efficiency [$\mu\text{mol CO}_2 \text{ mol}^{-1} \text{ H}_2\text{O}$].
iWUE^{eco} Ecosystem scale intrinsic water use efficiency [various units].
IWUE Inherent water use efficiency [$\text{gC kPa m}^{-2} \text{ mm}^{-1}$].
WUE^{eco} Ecosystem scale water use efficiency computed with ET [$\text{gC m}^{-2} \text{ mm}^{-1}$].
WUE_T^{eco} Ecosystem scale water use efficiency computed with transpiration [$\text{gC m}^{-2} \text{ mm}^{-1}$].
A_n Net carbon assimilation [$\mu\text{mol CO}_2 \text{ m}^{-2} \text{ leaf s}^{-1}$].
g_s Stomatal conductance [$\mu\text{mol CO}_2 \text{ m}^{-2} \text{ leaf s}^{-1}$].
T Transpiration [$\text{mmol H}_2\text{O m}^{-2} \text{ leaf s}^{-1}$] or [mm h^{-1}] or [mm day^{-1}] or [mm yr^{-1}].
ET Evapotranspiration [mm day^{-1}] or [mm yr^{-1}].
GPP Gross primary production [$\text{gC m}^{-2} \text{ yr}^{-1}$].
NPP Net primary production [$\text{gC m}^{-2} \text{ yr}^{-1}$].
VPD Vapor pressure deficit [Pa].
C_a Atmospheric CO₂ concentrations [ppm] or [Pa].
C_i Leaf internal (or intercellular) CO₂ concentrations [ppm] or [Pa].
WP_{B/Tr} Water productivity as the ratio of biomass to transpiration [$\text{gDM m}^{-2} \text{ mm}^{-1}$] or [kgDM m^{-3}].
WP_{B/ET} Water productivity as the ratio of biomass to evapotranspiration [$\text{gDM m}^{-2} \text{ mm}^{-1}$].
WP_{Y/Tr} Water productivity as the ratio of yield to transpiration [$\text{gDM m}^{-2} \text{ mm}^{-1}$].
WP_{Y/ET} Water productivity as the ratio of yield to evapotranspiration [$\text{gDM m}^{-2} \text{ mm}^{-1}$].
WP* Water productivity normalized per potential ET and CO₂ concentration [$\text{gDM m}^{-2} \text{ day}^{-1}$].
B Produced crop biomass [$\text{gDM m}^{-2} \text{ day}^{-1}$].
HI Harvest Index [–].
Y Agricultural yield [$\text{gDM m}^{-2} \text{ day}^{-1}$].
Rubisco Ribulose-1,5-bisphosphate carboxylase-oxygenase, the key enzyme involved in the first major step of carbon fixation in photosynthesis.

Key points

- Definitions of Water Use Efficiency (WUE) are reviewed and WUE values across scales and ecosystems are provided.
- Water productivity for a given crop is relatively constant for a given CO₂ concentration and potential evapotranspiration.
- WUE dependence on biome type and precipitation is weak.
- Current remote sensing vegetation products are unlikely to properly characterize WUE.
- WUE has been increasing in response to growing atmospheric CO₂ concentration.

Introduction and definitions

Various definitions of Water Use Efficiency (WUE) have been proposed depending on the spatial scale of interest (leaf, plant, or ecosystem level) and productivity aspects of crops and natural vegetation. In its most general form, WUE represents the ratio between the carbon assimilated by, or stored in, vegetation (e.g., Gross and Net Primary Production (GPP, NPP), biomass, yield) and the flux of water released to the atmosphere (e.g., transpiration, evapotranspiration) over the same unit of space and time.

At the leaf scale, the water use efficiency (WUE_l) [$\mu\text{mol CO}_2 \text{ mmol}^{-1} \text{ H}_2\text{O}$] is expressed as the ratio between the instantaneous net carbon assimilation A_n [$\mu\text{mol CO}_2 \text{ m}^{-2} \text{ leaf s}^{-1}$] and transpiration T [$\text{mmol H}_2\text{O m}^{-2} \text{ leaf s}^{-1}$] (Nobel, 2009, p. 404):

$$WUE_l = \frac{A_n}{T}. \quad (1)$$

As WUE_l is known to depend on atmospheric conditions and directly on atmospheric Vapor Pressure Deficit (VPD), a metric that discounts for the role of VPD is the leaf-scale intrinsic water use efficiency (iWUE) [$\mu\text{mol CO}_2 \mu\text{mol}^{-1} \text{ H}_2\text{O}$] or [$\mu\text{mol CO}_2 \text{ mol}^{-1} \text{ H}_2\text{O}$] (Medlyn et al., 2017):

$$iWUE = \frac{A_n}{T} VPD^* = \frac{A_n}{1.6g_s} \quad (2)$$

where VPD* [-] is the VPD computed between the leaf interior and the air at the leaf surface and normalized by atmospheric pressure, g_s [$\mu\text{mol CO}_2 \text{ m}^{-2} \text{ leaf s}^{-1}$] is the stomatal conductance for CO₂, and 1.6 is the ratio between water vapor and CO₂ diffusivity in air.

Considering Fick's law of diffusion, the net assimilation can be expressed as the product of stomatal conductance and the gradient of CO₂ concentration, $A_n = g_s(C_a - C_i)$, and iWUE can be re-written as:

$$iWUE = \frac{g_s(C_a - C_i)}{1.6g_s} = \frac{C_a}{1.6} \left(1 - \frac{C_i}{C_a} \right) \quad (3)$$

where C_a and C_i [-] are the CO₂ concentrations at the leaf surface (often approximated with the atmospheric CO₂ concentration) and the leaf CO₂ internal concentration in the sub-stomatal cavity, respectively. Eq. (3) has three major implications. First, iWUE is solely dependent on C_a and C_i/C_a . Therefore, all plant physiological and environmental controls (except C_a) are summarized in the C_i/C_a value. Note also that a factor of three on the range of C_i/C_a between 0.3 and 0.9 would lead to a factor of seven in iWUE change ($42\text{--}294 \mu\text{mol CO}_2 \text{ mol H}_2\text{O}^{-1}$) at current C_a levels of 420 ppm (Fig. 1), which represents a reasonable expectation for iWUE variability. Second, for constant C_i/C_a , the iWUE scales linearly with C_a , which explains why iWUE has been extensively used in studies looking at CO₂ effects on WUE (e.g., Frank et al., 2015). Third, at least in some classic literature, there has been a consensus that C_i/C_a is a relatively conserved quantity, with values around 0.7 for C₃ plants and 0.4–0.5 for C₄ plants (e.g., Morison and

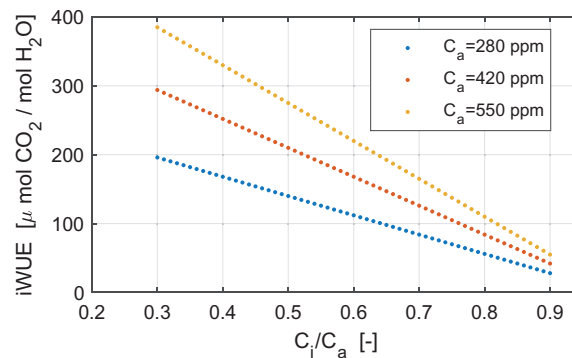


Fig. 1 Values of iWUE as a function of the ratio, C_i/C_a , between leaf internal (sub-stomatal cavity) and atmospheric CO₂ concentration for three different C_a values (280, 420, and 550 ppm).

Gifford, 1983; Long et al., 2004; Wong et al., 1985). If C_i/C_a is indeed constant, this would automatically imply that leaf-scale iWUE is solely a function of C_a for C_3 and C_4 plants (e.g., $iWUE = 0.1875 \cdot C_a$ for C_3), regardless of differences in plant physiological attributes or environmental conditions. In other words, for a given level of C_a , only two values of iWUE would exist (one for C_3 and one for C_4 plants). The constancy of C_i/C_a across different conditions has been used to explain the conservative behavior of the water used by crops to produce biomass (Steduto et al., 2007). However, C_i/C_a is not as constant across species and environmental conditions as often reported. The C_i/C_a ratio was shown to be affected by VPD (Morison and Gifford, 1983) and recently compiled datasets challenge the idea of a very conservative C_i/C_a value (Lin et al., 2015; Adams et al., 2020). Note also that a decrease of C_i/C_a from 0.8 to 0.7 (i.e., a -12% change) leads to a change in iWUE of +50%. Therefore, minor differences in C_i/C_a can have very important implications for WUE.

As the analysis moves from leaf scale to canopy and ecosystem scale, there are other processes that can influence WUE. First, a plant must compete for water with physical processes such as soil evaporation and drainage. Furthermore, canopy architecture, phenological cycle of vegetation, and, especially, the temporal dynamics of Leaf Area Index (LAI) can have an important role on water use efficiency, independently of leaf-scale physiology. Additional environmental factors such as nutrient availability and water stress can also affect the plant or ecosystem capacity of using water. In analogy with the leaf scale, the ecosystem scale intrinsic water use efficiency ($iWUE^{eco}$) is defined as (Beer et al., 2009):

$$iWUE^{eco} = \frac{GPP}{G_{s,can}} \quad (4)$$

where GPP [$gC\ m^{-2}\ h^{-1}$] is the Gross Primary Production and $G_{s,can}$ [$mol\ m^{-2}\ s^{-1}$] is the ecosystem scale canopy conductance (Lavergne et al., 2019). However, a more widely used and easier to measure metric at the ecosystem scale is the so-called inherent water use efficiency iWUE [$gC\ kPa\ m^{-2}\ mm^{-1}$]:

$$IWUE = \frac{GPP}{ET} VPD \quad (5)$$

where VPD [kPa] is computed using air temperature, and ET [$kg\ H_2O\ m^{-2}\ h^{-1}$] (or equivalently [$mm\ h^{-1}$]) is the ecosystem scale evapotranspiration, which sums the biotic transpiration and the abiotic evaporation from ground and from vegetation interception. More recently, it has been suggested that to better incorporate the nonlinear effect of VPD on ecosystem water use, an underlying water use efficiency, expressed as $uWUE = GPP \cdot VPD^{0.5}/ET$, should be used (Zhou et al., 2014). Alternatively, the basic ecosystem scale water use efficiency, WUE^{eco} [$gC\ m^{-2}\ mm^{-1}$], can be computed as:

$$WUE^{eco} = \frac{GPP}{ET} \quad (6)$$

Eq. (6) is simply the ratio of the gross amount of assimilated carbon and the total evapotranspiration at the ecosystem scale over a given period. A more plant-centered metric is: $WUE_T^{eco} = GPP/T$, where total ET is substituted with only the amount of transpired water T , which depends on plant behavior and its response to environmental conditions. However, WUE_T^{eco} is extremely difficult to measure and WUE^{eco} is generally reported in most studies. The ecosystem scale metrics $IWUE$ and WUE^{eco} are typically computed at the monthly or annual scale since GPP and ET have considerable variability at hourly and daily time scales.

Additional definitions of water use efficiency have been used in agronomy where the focus is on the amount of water used for generating crop biomass, crop yield, or marketable products such as fruits and oil, rather than GPP itself. For these agronomic applications, water use efficiency (often referred to as water productivity) can be measured as: $WP_{B/ET}$ [kg dry or fresh biomass $m^{-2}\ mm^{-1}$ transpiration], $WP_{B/ET}$ [kg dry or fresh biomass $m^{-2}\ mm^{-1}$ evapotranspiration], $WP_{Y/Tr}$ [kg dry matter yield $m^{-2}\ mm^{-1}$ transpiration], or $WP_{Y/ET}$ [kg dry matter yield $m^{-2}\ mm^{-1}$ evapotranspiration] (Steduto et al., 2007, 2012). Often these quantities are expressed in $m^{-3}\ H_2O$ rather than m^{-2} ground $mm^{-1}\ H_2O$. These values are reported for comparing water productivity levels across crops (Table 1) or for a given crop in different climates or agricultural management settings (Zwart and Bastiaanssen, 2004; Katerji et al., 2008; Vico and Porporato, 2015). For non-grain crops, such as sugarcane, cotton, or vegetable oils the water use efficiency is often expressed as the mass of the marketable products divided by the water use, e.g., as kg of sucrose, kg of lint, or kg of oil per $m^2\ mm$ or m^3 of evapotranspiration, respectively.

Table 1 Water productivity $WP_{Y/ET}$ computed as the ratio of yield (total yield—not grain yield) to evapotranspiration [$kgDM\ m^{-3}$] (or equivalently [$gDM\ m^{-2}\ mm^{-1}$]) for major crops.

Crop	$WP_{Y/ET}$ [$gDM\ m^{-2}\ mm^{-1}$]
Wheat	1.1 (0.6–1.5)
Rice	1.1 (0.4–1.6)
Maize	1.4 (0.8–2.4)
Soybean	0.47 (0.3–0.65)
Sorghum	1.2 (0.7–1.7)
Cotton Lint	0.23 (0.14–0.34)

Reported central values and range of variability (in brackets) are synthesized from four review articles (Zwart and Bastiaanssen, 2004; Katerji et al., 2008; Vico and Porporato, 2015; Kikal and Irmak 2017) and the FAO book on crop yield response to water availability (Steduto et al., 2012).

Relevance of WUE for linking water and carbon cycles

Given this mosaic of metrics used to express water use efficiency, it is extremely important to always present units and definitions explicitly. However, regardless of the selected metric, WUE represents a fundamental descriptor of the connection between the water and carbon cycles. Several reviews and synthesis articles have analyzed WUE in the past (e.g., Steduto, 1996; Howell, 2001; Katerji et al., 2008; Medlyn et al., 2017; Hatfield and Dold, 2019), even though syntheses of its magnitude and variability collected from different data sources and across ecosystems and environmental conditions are less common. WUE expresses the tradeoff between the plant capability of assimilating carbon and producing biomass and the water cost of such an operation. This tradeoff is essential for food production anytime water represents a limited resource for crops. Indeed, increasing WUE has motivated research toward improving agronomical techniques (Hsiao et al., 2007), especially irrigation (Howell, 2001), and it has also motivated crop breeding efforts (Condon et al., 2002, 2004; Franks et al., 2015). Any amelioration of WUE would result in the production of more or equal amounts of agricultural yield or any final product using a smaller amount of water. Concurrently, WUE is a very important trait of natural vegetation and ecosystems, because a higher WUE could potentially provide a competitive advantage in periods of water scarcity (Peters et al., 2018; Hatfield and Dold, 2019; Adams et al., 2021). Furthermore, studying future changes in WUE can be of high relevance to better understand how vegetation mediates water availability in a given region and how water availability modifies carbon budgets. For instance, if variations in climatic conditions lead to an increase (or decrease) in WUE, it would mean that vegetation productivity will increase (decrease) for unmodified water consumption or that water consumption will decrease (increase) for unmodified productivity or any combination of these possibilities. If the magnitude of the change is relevant, implications for terrestrial carbon budget (Keenan et al., 2013; Frank et al., 2015) and water resource availability (Knauer et al., 2017) could be significant. Therefore, knowledge of environmental controls and physiological and biophysical mechanisms affecting WUE under current and future climatic conditions is essential for Earth System Sciences including applications in agronomy, hydrology, and carbon cycle.

WUE observations and simulations

At the leaf scale, water use efficiency (i.e., WUE_l and $iWUE$) can be measured directly by means of leaf gas exchange measurements where A_n and T are inferred by changes in CO_2 and water vapor concentrations inside a leaf chamber (e.g., Medrano et al., 2015). Another way for obtaining WUE_l is related to the isotopic composition of the leaf carbon (C). Following Eq. (3), the sole knowledge of the ratio C_i/C_a is sufficient to compute $iWUE$ as C_a is known for a given historical period. During diffusion of CO_2 through stomata and photosynthesis processes there is a quantifiable discrimination ($\Delta^{13}C$) between the lighter $^{12}CO_2$ compared to the heavier $^{13}CO_2$ stable molecules. Based only on the knowledge of plant material $\delta^{13}C_p$, i.e., the ratio of ^{13}C and ^{12}C compared to an international standard, it is possible to compute plant $\Delta^{13}C$ (Laverne et al., 2019). The quantity $\Delta^{13}C$ has been found to be related to C_i/C_a in the first products of photosynthesis (Farquhar et al., 1982; Farquhar et al., 1989) Laverne et al., 2019) and, even though methodological concerns related to the role of mesophyll conductance (Seibt et al., 2008) and forest stand dynamics (Marchand et al., 2020) could be important, in its simplest form the $\Delta^{13}C - C_i/C_a$ relation can be expressed as:

$$\Delta^{13}C = a + (b - a) \left(\frac{C_i}{C_a} \right) \quad (7)$$

where $a = 4.4\text{‰}$ is the fractioning due to CO_2 diffusion and $b = 27\text{‰}$ is the apparent fractioning by Rubisco during carboxylation (Laverne et al., 2019). Strictly speaking, Eq. (7) can be used once leaf level $\delta^{13}C$ is known. However, with some adjustment to account for post-photosynthetic fractioning, it is typically used to infer C_i/C_a , and thus $iWUE$ (Eq. 3), based on the value of $\delta^{13}C_p$ observed in the cellulose of tree-rings and represents an indirect way to reconstruct average $iWUE$ in given periods (or years) for the whole plant (Peñuelas et al., 2011; Frank et al., 2015; Adams et al., 2020).

At the ecosystem scale, water use efficiency (i.e., WUE^{eco} and $EWUE$) can be computed based on observed latent heat (ET in energy units), VPD, and inferred GPP values from eddy-covariance flux measurements above vegetation (e.g., Beer et al., 2009). To compute $EWUE^{eco}$, an estimation of canopy conductance $G_{s,can}$ is required, which is typically derived by inverting the Penman-Monteith equation (Laverne et al., 2019). Using latent heat as a measure of total water use provides lower values of WUE^{eco} compared to WUE_T^{eco} as transpiration is only a fraction of total ET. To avoid this problem, several studies filter for days and hours where ground evaporation and evaporation from interception are expected to be negligible (e.g., Keenan et al., 2013; Mastrotheodoros et al., 2017). However, a physiological interpretation of WUE^{eco} and $EWUE$ remains complex due to the uncertainty embedded in flux tower ET and GPP observations (Knauer et al., 2018). Estimates of WUE^{eco} can be also obtained from global maps of GPP and ET derived from flux-tower data by means of machine learning techniques such as FLUXCOM (Jung et al., 2020).

Spaceborne remote sensing products offer the capability to estimate WUE^{eco} from field to global scales (e.g., Xue et al., 2015). However, most remote sensing products provide very indirect estimates of GPP (or NPP) and ET, since these quantities are derived from observations of vegetation radiation reflections in different wavebands, and of outgoing thermal radiation (e.g., Zhang et al., 2019b), and effects of other environmental factors are modeled. Therefore, the remotely sensed WUE^{eco} estimates are subjected to considerable uncertainties and are largely the product of simple models applied to the original observations.

Finally, estimates of WUE^{eco} can be obtained using the results of Terrestrial Biosphere Models (e.g., [Huang et al., 2015](#)). These models mechanistically solve energy, water, and carbon budgets and thus provide all necessary quantities to compute $iWUE^{eco}$, WUE^{eco} and WUE_T^{eco} . However, these values depend on the model capability to correctly reproduce water and carbon fluxes, which are influenced by the model internal structure, with non-negligible inter-model differences in the simulations of WUE^{eco} ([Paschalis et al., 2020](#)).

Crop water productivity

Given the economic and societal importance of increasing agricultural yield and diminishing water use for irrigation, a large attention has been dedicated to improving crop water use efficiency both at leaf ($iWUE$) and field (WUE^{eco}) scale, and consequently improving $WP_{B/ET}$ and $WP_{Y/ET}$. The catchy slogan “more crop per drop” ([Kijne et al., 2003](#)) summarized this effort for improving $WP_{Y/T}$ especially through selective breeding. However, the early enthusiasm for breeding for higher water use efficiency has been questioned ([Blum, 2009](#)) for reasons detailed below. Agricultural yield (Y [$gDM\ m^{-2}\ day^{-1}$]) can be decomposed in various terms to explicitly account for $WP_{B/ET}$ or $WP_{B/T}$ ([Passioura, 1996](#)):

$$Y = B \cdot HI = T \cdot WP_{B/T} \cdot HI = ET \cdot \frac{T}{ET} \cdot WP_{B/T} \cdot HI = ET \cdot WP_{B/ET} \cdot HI \quad (8)$$

where B [$gDM\ m^{-2}\ day^{-1}$] is the produced crop biomass, HI [–] is the Harvest Index expressed as the yield divided the biomass; and by definition: $WP_{B/T} = B/T$ [$gDM\ m^{-2}\ mm^{-1}$] and $WP_{B/ET} = B/ET$ [$gDM\ m^{-2}\ mm^{-1}$]. Eq. (8) suggests that for a given amount of evapotranspired water “ET,” the yield can be improved (i) by increasing the amount of water that is used for transpiration rather than for other ET pathways, i.e., the T/ET ratio; (ii) by increasing the carbon allocation to the harvested products (e.g., grains, fruits), i.e., the HI value; (iii) by enhancing the amount of carbon acquired for unit of transpired water, i.e., $WP_{B/T}$ itself. These are the three major mechanisms also described in [Condon et al. \(2004\)](#) to improve WUE by breeding of crop varieties. A strong linearity between biomass production (and yield) in crops and transpiration has been historically observed ([Fig. 2a–c](#)) since the work of [de Wit \(1958\)](#), who reassessed biomass yield and T observations carried out in containers at the beginning of the 20th century ([Briggs and Shantz, 1913, 1914](#); [Shantz and Piemeisel, 1927](#); [Dillman, 1931](#)). These linear relationships, even though with lower correlations and significant site to site variability, have been widely confirmed in the literature (e.g., [Hanks, 1983](#)) using field measurements of yield and ET ([Fig. 2d](#)) or, better, ET divided by potential evapotranspiration (E_{POT}) ([Fig. 2e](#)). ET is used rather than T , as separating the two in the field is extremely challenging. However, assuming that a large part of the variability is due to uncertainty in field observations and lack of knowledge of T/ET , once transpiration is re-scaled with potential evapotranspiration (E_{POT}), and CO_2 effects are accounted for by rescaling C_a with a reference value C_a^* , biomass production can be expressed as ([Steduto et al., 2007](#)):

$$B = WP^* \frac{T}{E_{POT}} \left(\frac{C_a}{C_a^*} \right) \quad (9)$$

The coefficient of proportionality WP^* [$gDM\ m^{-2}\ day^{-1}$] can be assumed to be nearly constant across a range of environmental conditions for each crop type and, for a given level of mineral nutrition (e.g., [Hanks, 1983](#); [Sinclair et al., 1984](#); [Steduto et al., 2007](#)). This assumption is especially true for a given location, while variability in the B (or Y) vs T/E_{POT} relation across locations can be considerable, but still lower than the variability among different crops ([Fig. 2d](#) and [e](#)). Typically, WP^* values range from 10 to 33 $gDM\ m^{-2}\ day^{-1}$ ([Table 2](#)). For a summer potential evaporation of 5 $mm\ day^{-1}$ and Carbon Use Efficiency (e.g., NPP to GPP ratio) of 0.5, these values lead to a WUE_i of 2.0–6.6 $gC\ kg\ H_2O^{-1}$, values just slightly larger than the WUE_i obtained from gas exchange observations (see next Section). As only a fraction of GPP represents yield and not all ET is transpiration, average $WP_{Y/ET}$ for major crops are smaller (0.25–0.7 $gC\ kg\ H_2O^{-1}$) with larger values for C_4 crops ([Table 1](#)). WP^* may be reduced by nutrient limitations, particularly lack of nitrogen. For some species with high rates of metabolic respiration for oil producing seeds or expensive fruits, pre-anthesis and post-anthesis values of WP^* might differ ([Table 2](#)) with lower values occurring toward the end of the growing season ([Steduto et al., 2007](#)). From Eq. (9) it follows that for $C_a = C_a^*$:

$$WP_{B/T} = \frac{WP^*}{E_{POT}} \quad (10)$$

which also means that, for a given climate, $WP_{B/T}$ must be nearly constant, with very limited room to modify yields by acting on $WP_{B/T}$. Eqs. (8) and (10) can be also linked to the relation between yield and evapotranspiration that has been the foundation of yield response to water availability in the FAO books ([Doorenbos and Kassam, 1979](#); [Steduto et al., 2012](#)):

$$\left(1 - \frac{Y}{Y_X} \right) = K_y \left(1 - \frac{ET}{E_X} \right) \quad (11)$$

where Y_X is the maximum yield corresponding to the maximum ET, $E_X \approx E_{POT}$. Eq. (11) is a water production function and can be applied to all agricultural crops (i.e., herbaceous, trees, and vines) and describes a linear proportionality between actual ET and Y . The only parameter is a yield response factor (K_y) that captures the complex relation between crop productivity and water use by a crop, and thus summarizes many biological, physical, and chemical processes ([Steduto et al., 2012](#)). If WP^* is constant, and T/ET

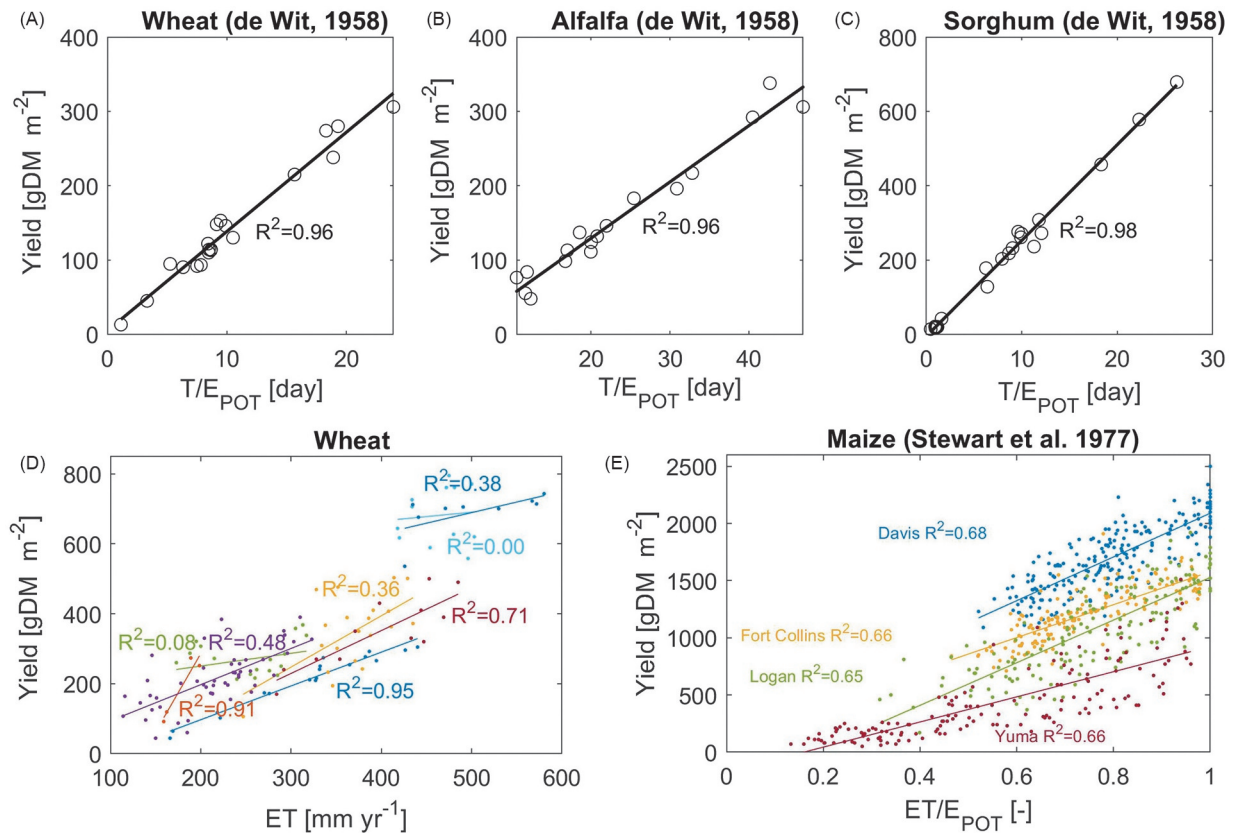


Fig. 2 Agricultural yields (Y , dry matter biomass in yield) as a function of the ratio between transpiration (T) and potential evapotranspiration (E_{POT}), (a) for Wheat ($n = 20$) (b) Alfalfa ($n = 17$) and (c) Sorghum ($n = 21$). Data were summarized by [de Wit \(1958\)](#) based on the original publications of [Briggs and Shantz \(1913, 1914\)](#), [Shantz and Piemeisel \(1927\)](#) and [Dillman \(1931\)](#), where crops were grown in closed containers and the surface of the soil was sealed to prevent direct evaporation. Experiments were carried out in four different locations of the USA Great plains: Delhart (Texas), Akron (Colorado), Newell (South Dakota) and Mandan (North Dakota) in the period 1911–1922. For Sorghum data were also from Bombay (India) ([Kanitkar et al., 1943](#)). (d) Wheat yield as a function of evapotranspiration (ET) (lack of information about potential evapotranspiration during the wheat growing season precludes the possibility to use ET/E_{POT}), data are derived from eight individual studies summarized in [Vico and Porporato \(2015\)](#). (e) Maize yield as a function of the ratio between evapotranspiration (ET) and potential evapotranspiration (E_{POT}). Crops were grown under field conditions in 1974 and 1975 and they underwent various treatments modifying irrigation amount, irrigation frequency, and salinity ([Stewart et al., 1977](#)). Results are four different locations in the USA (Davis ($n = 216$, blue), Fort Collins ($n = 161$, yellow), Logan ($n = 157$, green), and Yuma ($n = 154$, red)). In all plots, we report the linear fitting and corresponding coefficient of determination R^2 .

Table 2 The water productivity WP^* [gDM m⁻² day⁻¹] coefficient for different crop types as reported in AquaCrop reference manual—Annex ([Raes et al., 2018](#)).

Crop	WP^* [gDM m ⁻² day ⁻¹]	Fractional reduction (during yield formation)
Wheat	15.0	1.0
Rice	19.0	1.0
Maize	33.7	1.0
Soybean	15.0	0.6
Barley	15.0	1.0
Sorghum	33.7	1.0
Cotton	15.0	0.7
Sunflower	18.0	0.6
Sugarcane	30.0	1.0
Potato	18–20	1.0
Tomato	18	1.0
Sugar beet	17.0	1.0
Quinoa	10.5	0.9
Tef	14.0	1.0
Dry Bean	14.0	0.9

The reference value and the fractional reduction during yield formation for each crop are reported.

and HI are the same in potential and actual conditions, then $K_y = 1.0$. Observations suggest indeed a relatively limited range of K_y values around 0.8 to 1.35 for various crops (Doorenbos and Kassam, 1979).

The fact that WP^* has been historically observed to be a conserved quantity for a given crop (Steduto et al., 2007), it implies that once differences in evaporative demand and C_a are discounted, $WP_{B/ET}$ and $WP_{Y/ET}$ are also conserved. This makes it particularly challenging to improve $WP_{B/ET}$ beyond the values already achieved with best agricultural practices. However, breeding of crop varieties that are more efficient in their water use has been shown to be possible (Condon et al., 2004). While it is relatively easy to select for genotypes with higher iWUE at the leaf scale, initial expectations of substantial water savings were not maintained at the overall plant and field scale (Medrano et al., 2015), where genotypes with higher iWUE generally had lower productivity and yields (Condon et al., 2004). This is because the iWUE increase is often obtained by lowering g_s , rather than increasing A_n (Blum, 2009), therefore growth and yield are not stimulated at the plant level. Evidence suggests that genetically modified plants may not be better able to cope with drought than selection-bred cultivars as smaller leaf area and lower g_s may decrease productivity in the field and genetic modifications may delay stress onset rather than conferring drought resistance or avoidance (Lawlor, 2013).

Additionally, other important factors such as the length of the growing season or the biomass production in the early growing phase before anthesis have been shown to affect yield more significantly than iWUE itself, complicating the approaches for improving WUE^{eco} without compromising productivity (Condon et al., 2004). Therefore, alternative strategies aiming at reducing ground evaporation and increasing photosynthesis through higher LAI could be more beneficial (Condon et al., 2004). For instance, growing crops in a less water demanding atmosphere (lower VPD) can improve WUE^{eco} . An increase in WUE^{eco} could be obtained indirectly by fertilizer application in nutrient poor soils. When nutrients are limiting, fertilizers increase GPP and yield, often without increasing ET proportionally especially in wet years or ecosystems. Soil texture has also been shown to be able to affect $WP_{Y/ET}$ with lower values in clay soils where crop root water uptake may be hindered (Katerji and Mastrorilli, 2009). More generally, improvements in WUE can be obtained by acting on agricultural practices and management (Hatfield et al., 2001). Agricultural practices, such as crop residue management, tillage, mulching, weed control, use of intercropping, are shown in certain experiments to increase water available in the root zone, WUE^{eco} and $WP_{B/ET}$ (Hatfield et al., 2001; Turner, 2004; Hatfield and Dold, 2019). However, experiments with various agricultural practices have also shown a decrease or lack of response in $WP_{B/ET}$ (Hatfield et al., 2001). Differences among studies and difficulty in reproducing certain positive responses suggest that results are likely climate and crop specific and are also affected by the specific conditions of the experiment. Other practices such as crop row spacing and use of appropriate irrigation practices have also shown improvements in $WP_{B/ET}$. Specifically, plant spacing modifies the plant radiation environment and photosynthetically active radiation absorption, whereas irrigation might remove water stress, improving overall $WP_{B/ET}$. Techniques such as drip irrigation can minimize ground evaporation and evaporation from interception, therefore increasing the T/ET ratio and $WP_{B/ET}$ even further. In summary, management techniques acting on nutrient availability, irrigation application, water retention in the root zone, root water uptake, soil evaporation, and the biomass allocation toward the yield part of the crop are the best leverages to improve $WP_{Y/ET}$ (Hsiao et al., 2007) as they mostly act upon T/ET and HI.

WUE variability and environmental dependencies

It is often textbook knowledge that, at the leaf scale, plants in dry environments have higher WUE to cope with limited water availability (e.g., Chapin et al., 2002, p. 114). Concurrently, C_4 plants have lower C_i/C_a and higher WUE_i than C_3 plants because of their capability of avoiding photorespiration losses and attaining higher photosynthetic rates for a given C_a . Observations from recently compiled global databases of leaf-level gas exchange (Lin et al., 2015; Medlyn et al., 2017) and tree-ring isotope data (Adams et al., 2020) do not support a considerable climate dependence of iWUE. iWUE observations clearly suggesting that the C_4 photosynthetic pathway confers a considerably higher iWUE, while C_3 grasses and crops have the lowest iWUE values. iWUE values range from 21 to 151 $\mu\text{mol CO}_2 \text{ mol}^{-1} \text{ H}_2\text{O}$ (5–95 quantile) or 21–102 if we exclude C_4 plants (Fig. 3a), indicating that iWUE and C_i/C_a differ among C_3 and C_4 species. A somewhat smaller range of iWUE, 33–80 $\mu\text{mol CO}_2 \text{ mol}^{-1} \text{ H}_2\text{O}$ (5–95 quantile), is obtained from tree-ring isotope analysis, consequent on the database containing more values but only C_3 species, as it is limited to trees (Fig. 3b). Leaf-level iWUE correlation with WUE_i is small ($r^2 = 0.12$) (Fig. 3c), demonstrating an important role for the VPD conditions at which measurements are carried out. Leaf-level WUE_i range is 1.8–6.4 $\mu\text{mol CO}_2 \text{ mmol}^{-1} \text{ H}_2\text{O}$ (5–95 quantile) [corresponding to 1.2–4.2 $\text{gC kg H}_2\text{O}^{-1}$], with strong overlaps among plant functional types, where only evergreen needleleaf species show substantially higher WUE_i values (Fig. 3d). It must be noted, however, that Crassulacean Acid Metabolism (CAM) plants, that keep their stomata open at night when VPD is lowest and not during the daytime, can reach WUE_i values of 4–20 $\mu\text{mol CO}_2 \text{ mmol}^{-1} \text{ H}_2\text{O}$, much higher than for other species (Lambers and Oliveira, 2019, p. 243). These are not represented in Fig. 3. The most striking observation in Fig. 3 is, however, the large overlap of iWUE and WUE_i values across different plant functional types, showing that different iWUE values might co-exist in the same biome and, concurrently, that different biomes might have very similar average iWUE. Leaf-scale and tree-ring iWUE show a very weak negative dependence ($r^2 = 0.032$ and $r^2 = 0.076$ respectively) with precipitation, with fitting curves (C_3 species only) largely overlapping for precipitation lower than 2000 mm yr^{-1} (Fig. 3b), thus not supporting the hypothesis that on average plants in arid and semiarid biomes use water considerably more efficiently, than in other climates.

Overall, iWUE variability among C_3 species spans less than a factor of five and, in total, a factor of seven, which likely represents the leaf physiological and morphological flexibility in controlling iWUE, similar to the factor of seven derived from theoretical expectations (Eq. 3). This is also reflected in the values of the estimated C_i/C_a ratio, considerably lower for C_4 species (0.29–0.73)

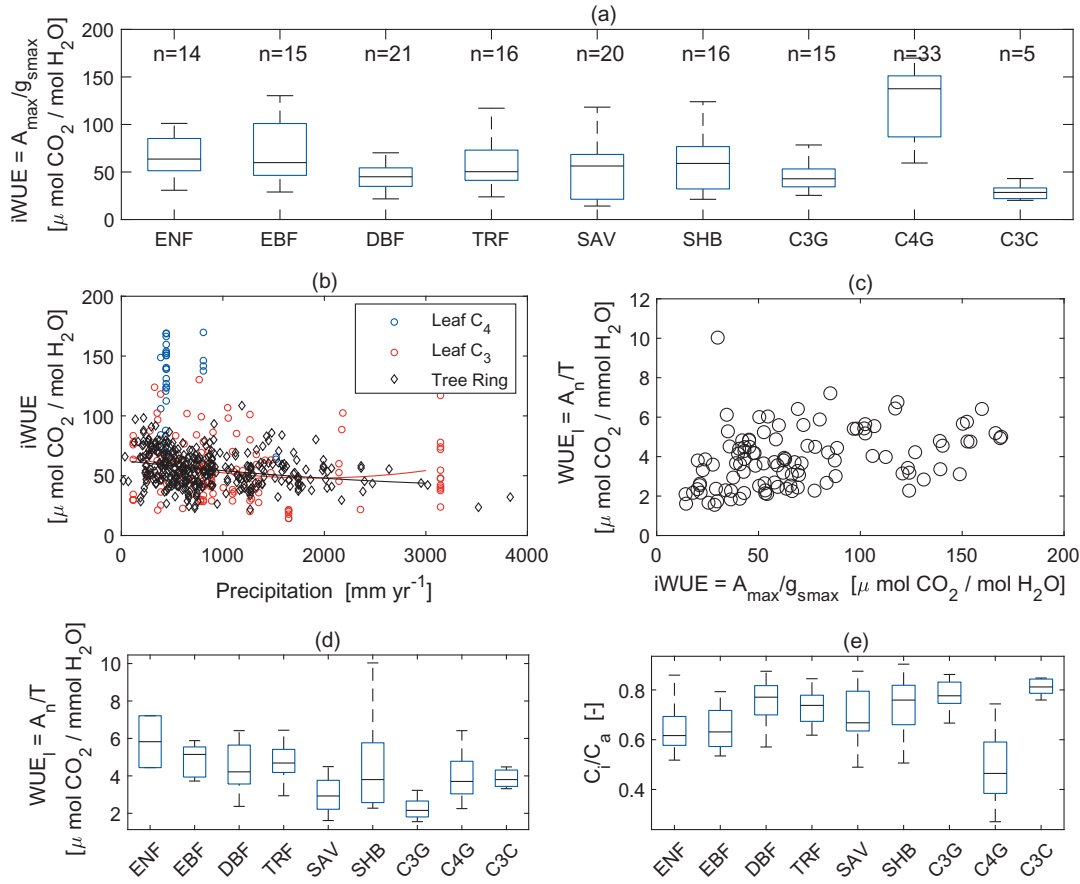


Fig. 3 (a) Boxplot of iWUE computed from the maximum values of net assimilation A_{max} and stomatal conductance $g_{s,max}$ of a given species for different biomes (ENF, evergreen needleleaf forest; EBF, evergreen broadleaf forest; DBF, deciduous broadleaf forest; TRF, tropical rainforest; SAV, savannah; SHB, shrub; C3G, C3 grass; C4G, C4 grass; C3C, C3 crops). Data are from the Lin et al., 2015 database (53 sites –155 species-sites, as species-sites with less than 4 observations have been excluded). (b) Scatter plot between iWUE and local average annual precipitation in the Lin et al., 2015 database (53 sites –155 species-sites) subdivided in C₃ and C₄ species and in the Adams et al., 2020 database (181 sites –362 species-sites, iWUE values are averaged of the time series). Precipitation is extracted from the GPCC version 2020 database (<https://www.dwd.de/EN/ourservices/gpcc/gpcc.html>) based on site latitude and longitude. Second order polynomials are used to fit the data for C₃ species. (c) Scatter plot between WUE_i and iWUE computed from A_{max} and $g_{s,max}$. (d) Boxplot of WUE_i for different biomes computed as the mean of net assimilation divided transpiration for each species-site. (e) Boxplot of $C_i/C_a for different biomes computed as the mean for each species-site starting from iWUE values.$

than for C₃ (0.55–0.86) (0.53–0.83 using data from Adams et al., 2020) but showing a substantial overlap among different biomes and slightly lower values for evergreen broadleaf and needleleaf species (Fig. 3e). Variations in average iWUE across different biomes are indeed substantially smaller than across species and reasonable differences are more within a factor of two rather than five or seven (Fig. 3a).

The magnitude of multi-annual WUE^{eco} at the ecosystem scale can be inferred from flux tower observations (Fluxnet 2015 synthesis¹) and remote sensing products (MODIS ET and GPP²). Flux tower derived WUE^{eco} range is $0.63\text{--}4.50\text{ gC m}^{-2}\text{ mm}^{-1}$ (5–95 quantile), providing a slightly larger magnitude of variability in WUE^{eco} than the leaf-level WUE_i and confirming that biomes are not an important determinant of WUE^{eco} (Medlyn et al., 2017). The weak correlation between WUE^{eco} and precipitation rather suggests that water limited sites have lower and not higher WUE^{eco} (Fig. 4a). However, this does not necessarily imply a decrease in leaf-level WUE_i as ET rather than T is used to compute WUE^{eco} , and a decreasing T/ET ratio in low precipitation regions would lead to a decrease in WUE^{eco} even with unchanged WUE_i .

The range of WUE^{eco} derived from MODIS remote sensing products is quite similar but with lower maxima, $0.62\text{--}3.96\text{ gC m}^{-2}\text{ mm}^{-1}$ (5–95 quantile for areas with precipitation larger than 50 mm yr^{-1}), than the one obtained from flux towers (Fig. 4c), but it also indicates the weak dependence on precipitation. The global maps of WUE^{eco} derived from FLUXCOM product (obtained through a machine learning extrapolation of flux tower data using various covariates, Jung et al., 2020) and MODIS tend to agree on the patterns of WUE^{eco} over temperate regions of North America, Europe, Asia, and over equatorial tropical

¹<https://fluxnet.org/data/fluxnet2015-dataset/>

²<https://modis.gsfc.nasa.gov/data/>

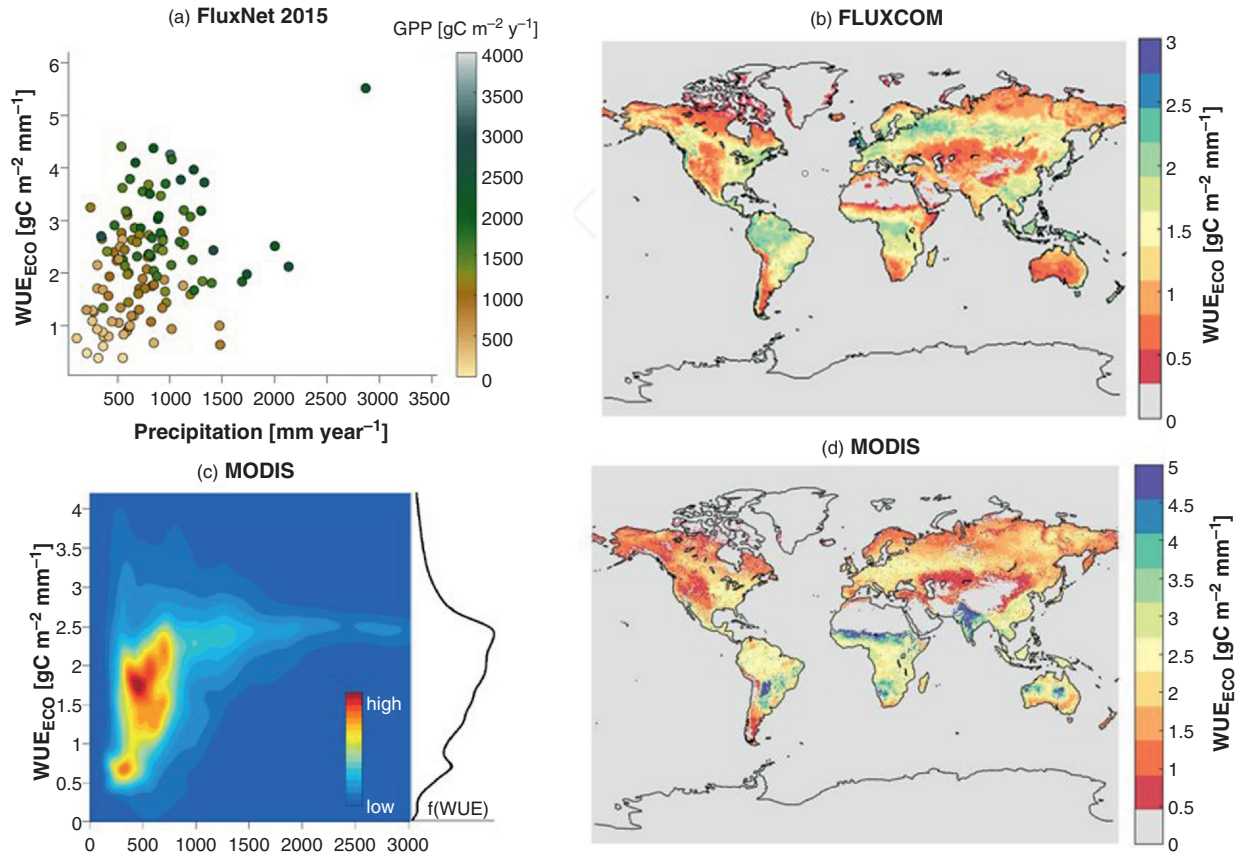


Fig. 4 (a) Scatter plot between WUE^{eco} and precipitation computed from the Fluxnet 2015 database (123 Sites) for different levels of Gross Primary Production (GPP). (b) Global map of WUE^{eco} computed from GPP and ET obtained from the FLUXCOM product, period 1980–2013. Reported values are the average WUE using the ANN, MARS and RF algorithms for GPP and the MLM algorithm for ET. All products used the RS + METEO data sets developed using meteorological forcing from CRUNCEP. (c) Density plot of WUE^{eco} as a function of precipitation computed from MODIS GPP and ET products. (d) Global map of WUE^{eco} computed from the MODIS GPP and ET products, period 2000–2016.

regions, but they show contrasting patterns in various semiarid regions such as Sahel, central Australia, East India and Pakistan (Fig. 4b and d). In these regions, MODIS WUE^{eco} is at the maximum, while FLUXCOM WUE^{eco} has very low values (see also Sun et al., 2016). Local flux-tower observations and process-based models (Sun et al., 2016) support low values of WUE^{eco} in semi-arid and arid regions. This evidence suggests that MODIS algorithms might be particularly problematic in approximating ET and/or GPP in regions with sparse vegetation. While the pattern of FLUXCOM derived WUE^{eco} appears more realistic, the maximum values rarely exceed 3.0 gC m⁻² mm⁻¹ (0.25–2.05, the 5–95 quantile) in contrast with local flux-tower observations and process-based model results (Sun et al., 2016), which report values above 4.0 gC m⁻² mm⁻¹. The FLUXCOM product also provides very gradual spatial variability in WUE^{eco}. These patterns are likely the result of a smoothing, induced by the machine learning algorithm and of the 0.5° spatial scale of the FLUXCOM product (Fig. 4b). Overall, FLUXCOM information suggests that peaks of WUE^{eco} are obtained in moist tropical forests and in forests at the border between temperate and boreal climates in Russia and Canada. In these regions water stress is likely minimal and atmospheric water demand moderate, with both factors allowing plants to take full advantage of the available water to assimilate carbon, and compounding effects are related to dense canopies with a high LAI. Increasing IWUE with higher soil moisture and larger LAI have been observed across flux tower sites (Beer et al., 2009), which supports this assessment.

Regions which are either strongly temperature-limited, such as arctic and sub-arctic regions, or strongly water-limited, such as semi-arid and arid regions and characterized by deserts and sparse vegetation (e.g., south-west USA, central Asia, south Africa, central Australia) have a considerably lower WUE^{eco}. In these regions, vegetation is not able to use water to assimilate carbon at the highest capacity, either because other factors (e.g., temperature, light) are limiting or because soil and atmospheric water limitations themselves decrease GPP. The vegetation in these regions cannot attain a dense cover with high LAI. These patterns have been confirmed by process-based models (Sun et al., 2016) and by studies estimating global WUE^{eco} with alternative methodologies (Cheng et al., 2017).

Simulations with the mechanistic terrestrial biosphere model T&C (Fatichi et al., 2016; 2021) for 119 sites characterized by different climates and biomes support the weak dependence of multi-annual WUE^{eco} on precipitation, showing lower WUE^{eco} values in semiarid regions, even though with a large scatter, in agreement with the FLUXCOM product (Fig. 5a). The WUE^{eco} range

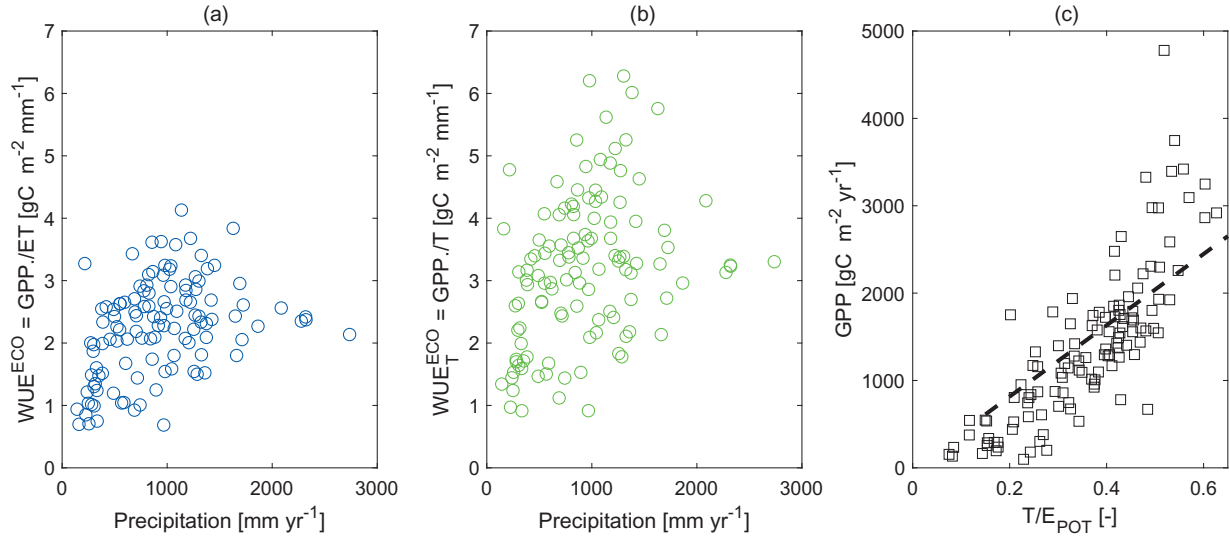


Fig. 5 Scatter plot of ecosystem scale (a) WUE^{eco} and (b) WUE_T^{eco} as a function of annual precipitation computed with the terrestrial biosphere model T&C for 119 sites characterized by different climates and biomes. (c) Scatter plot between GPP and transpiration divided potential evapotranspiration (T/E_{POT}) for the same 119 sites. A fitting linear function, without intercept, $GPP = 4080 (T/E_{POT})$ [$gC\ m^{-2}\ year^{-1}$] ($r = 0.77$) is also illustrated in panel c (dashed line).

of $0.9\text{--}3.5\ gC\ m^{-2}\ mm^{-1}$ (5–95% percentile) is also very similar to the ones obtained from flux-towers and other process-based models (Sun et al., 2016) with a median value of $2.33\ gC\ m^{-2}\ mm^{-1}$ demonstrating that ecosystem scale WUE^{eco} variability (a factor of 4) is not too dissimilar from variability at the leaf scale. Mechanistic terrestrial biosphere models allow analysis of WUE_T^{eco} , which is a quantity that cannot be easily derived from observations. Model simulations with T&C shows that the patterns of WUE^{eco} are largely preserved using WUE_T^{eco} even though WUE_T^{eco} values are larger ($1.3\text{--}5.4\ gC\ m^{-2}\ mm^{-1}$), as expected (Fig. 5b). These values compare well with the observed instantaneous WUE_i of $1.2\text{--}4.2\ gC\ m^{-2}\ mm^{-1}$, showing that canopy and ecosystem upscaling and long-term averaging are not generating any major departure from the magnitude of the fundamental leaf physiological process. This also suggests that the variability of WUE_i observations across species and environmental forcing is sufficient to cover a quite large range of natural conditions, representative of various canopies, ecosystems, climates, and soils.

Simulations also confirm a considerable scatter of WUE_T^{eco} values, increasing with precipitation ($r = 0.48$, $P < 0.001$) and decreasing with temperature ($r = -0.24$, $P = 0.007$, not shown), while the negative correlations with radiation ($r = -0.64$, $P < 0.001$) and VPD ($r = -0.59$, $P < 0.001$) are much stronger (not shown), reinforcing the idea of WUE^{eco} and WUE_T^{eco} mostly decreasing with increasing evaporative demand, and increasing with higher atmospheric CO_2 (e.g., Eq. 9). Note, however, that these are spatial patterns in WUE^{eco} observed across sites and might differ from temporal environmental dependencies of WUE^{eco} at a given site. In this regard, studies across the US found that WUE^{eco} is sensitive to VPD (Novick et al., 2016) and, for a given site, flux-tower observations have shown that $iWUE^{eco}$ increases with increasing VPD for different levels of soil moisture, at least as far as VPD does not exceed $2\text{--}3\ kPa$ (Zhang et al., 2019a). A similar pattern is also observed in $iWUE$ from leaf-scale observations with $iWUE$ increasing with VPD up to $3.5\ kPa$ and decreasing thereafter (not shown). A normalization of transpiration with potential evapotranspiration, similar to what is commonly done for crops, does lead to a much stronger correlation with GPP ($r = 0.77$, $P < 0.001$) in T&C simulations (Fig. 5c). The derived multi-site $WP^* = 11\ gDM\ m^{-2}\ day^{-1}$ (assuming a biomass conversion efficiency of 50%) is on the lower end of values published for crops.

In semi-arid regions WUE_T^{eco} tends to decrease following WUE^{eco} , as the T/ET ratio in the simulations does not depend on precipitation or aridity (Paschalis et al., 2018), confirming that vegetation adapted to arid and semi-arid regions is not able to use water more efficiently than other biomes at the ecosystem scale and over annual and multi-annual scales, contrary to common expectations. However, an increase of $iWUE$ with aridity has been shown, even though with very weak correlation, using $iWUE$ inferred from tree rings (Adams et al., 2021). Large values of WUE^{eco} ($2.15\ gC\ m^{-2}\ mm^{-1}$) for the corresponding aridity level (only $280\ mm$ of precipitation) have been also reported for the Yatir forest in Israel (Helman et al., 2017). This might suggest that WUE of trees growing in arid regions might behave differently from grassland and shrubland ecosystems.

Looking in time rather than space, $iWUE$ has been inferred to increase in response to extreme droughts at the regional scale based on inversion of atmospheric isotopic composition (Peters et al., 2018). However, this can be an indirect effect of increasing VPD values. Another study showed contrasting patterns of WUE^{eco} response to drought conditions, with WUE^{eco} increasing with drought in arid ecosystems and decreasing in semi-arid/sub-humid ones (Yang et al., 2016). An important point is that WUE^{eco} and $iWUE$ can exhibit seasonal variations linked to climate seasonality and leaf phenology (e.g., Zhu et al., 2014). Additionally, simply referring to their definitions, WUE^{eco} and $iWUE$ are expected to exhibit diverging responses to increasing VPD associated with periods of soil or atmospheric water stress. $iWUE$ increases during atmospheric droughts, as transpiration response to VPD is sub-linear, and the opposite should occur for WUE^{eco} as the rate of change of ET is expected to be less than that for GPP.

(Zhang et al., 2019a). Responses to decreasing soil moisture are likely more complex and non-monotonic, as iWUE might increase for moderate water limitations, where transpiration is more sensitive than A_n to low levels of soil moisture. Then iWUE can decrease to nearly zero or even negative values ($A_n < 0$ in Eq. 2) when soil moisture completely limits photosynthesis, but some residual transpiration still occurs. However, given the species-to-species variability and the multiple factors controlling ecosystem scale ecohydrological dynamics, a full characterization of WUE responses to atmospheric water limitations and droughts remains an open question.

Other studies have investigated the interactions between WUE and meteorological forcing, such as wind (Schymanski and Or, 2016) and increasing aerosol concentrations (Gao et al., 2018), with WUE decreasing when direct radiation is stronger due to a cleaner atmosphere, and increasing with higher wind speed at the leaf scale.

Stomatal conductance modeling and WUE

The simulated WUE^{eco} in terrestrial biosphere models is a result of upscaling processes from the leaf to the canopy and to the entire ecosystem (Medrano et al., 2015; Medlyn et al., 2017). Yet, the parameterization of leaf-level stomatal conductance represents an essential component of the simulated iWUE and WUE^{eco} in these models. Mechanistic descriptions of leaf-level stomatal conductance and transpiration have been proposed (Buckley, 2019), however, they require many parameters difficult to determine for ecosystem scale applications. Therefore, in terrestrial biosphere models, parameterizations of stomatal conductance (g_s) as a function of A_n are often based on semi-empirical expressions such as Ball-Berry—Eq. (12) (Ball et al., 1987) or Leuning—Eq. (13) (Leuning, 1995) or expressions derived from stomatal conductance optimization approaches, including Medlyn—Eq. (14) (Medlyn et al., 2011). Each equation is formulated slightly differently, but they all contain a key parameter (m [–], a_1 [–] and g_1 [kPa^{0.5}], respectively) that controls iWUE:

$$g_s = g_0 + m \cdot A_n \frac{R_H}{C_s} \quad (12)$$

$$g_s = g_0 + a_1 \cdot \frac{A_n}{(C_{s/i} - \Gamma) \left(1 + \frac{VPD}{D_0}\right)} \quad (13)$$

$$g_s = g_0 + \left(1 + \frac{g_1}{\sqrt{VPD}}\right) \frac{A_n}{C_a} \quad (14)$$

where g_0 [$\mu\text{mol CO}_2 \text{ m}^{-2} \text{ leaf s}^{-1}$] represents the residual stomatal conductance at zero net assimilation due to cuticular conductance and imperfect stomatal closure (Duursma et al., 2019). C_a , C_s , C_i [ppm] are the CO_2 concentrations in the atmosphere, leaf surface, and leaf interior, R_H [–] is the relative humidity, Γ [ppm] is the CO_2 compensation point, VPD [kPa] is theoretically computed from leaf temperature (but more often using air temperature) and assuming vapor saturated conditions inside the leaf, and D_0 [kPa] is an empirical parameter modulating VPD response in Eq. (13).

Parameters m , a_1 , and g_1 are inversely proportional to iWUE. In early models, only two (or three if conifers were separated) values of m ($m = 9$ for C_3 , and $m = 4$ for C_4 plants) were assumed (Sellers et al., 1996) as a corollary of the assumption that C_i/C_a should have been constant within C_3 and C_4 plants. However, now we know that their values change across species and biomes (Lin et al., 2015; Medlyn et al., 2017; Miner et al., 2017) and potentially also over time (Lammertsma et al., 2011; Mastrotheodoros et al., 2017), with important implications for modeling iWUE and WUE^{eco} . An estimate of the values of a_1 and g_1 is obtained from the global leaf scale gas exchange database (Lin et al., 2015) and it shows the strong inverse correlation between those parameters and iWUE ($r^2 = 0.65$ and $r^2 = 0.67$, respectively for a power-law fitting) (Fig. 6a and b) and the very tight correlation ($r^2 = 0.96$) between the two parameters (Fig. 6c). This suggests that both Eq. (13) and Eq. (14) might provide very similar responses in terms of WUE once properly parameterized. Using the dataset compiled by Lin et al. (2015), the estimated range of g_1 is 1.1 to 7.6 (5–95 percentile), and with $D_0 = 1$ kPa, a_1 ranges from 2.6 to 15.0, providing a variability in the order of 5–7 times, in line with the variability in WUE_l and iWUE themselves. The ranges provided here are relevant for realistic parameterizations of terrestrial biosphere models.

Trends in water use efficiency

According to Eq. (3), iWUE should increase linearly with C_a if the C_i/C_a ratio remains constant with varying C_a , as indicated by physiological literature (e.g., Morison, 1985; Drake et al., 1997) and process-based modeling (Fig. 7a). Stomatal conductance, indeed, generally decreases with C_a (Fig. 7b, Medlyn et al., 2001). The increase in iWUE at the leaf and plant scale with C_a has been widely documented, (e.g., Wullschlegel et al., 2002) but the magnitude of reported WUE increments with C_a has varied significantly in the literature (Walker et al., 2021). Part of this spread is related to methodological differences in various studies, but part is also due to different analyzed periods as C_a changed through time. Once we normalized published results on a per unit of CO_2 change (ppm^{-1}), the observed increase in iWUE seems to be constrained between 0.2% and 0.5% $\text{ppm}^{-1} \text{ CO}_2$, which is aligned with or larger than theoretical expectations of a 0.26% $\text{ppm}^{-1} \text{ CO}_2$ increase in the last 20 years (Fig. 7c). Although iWUE has increased more

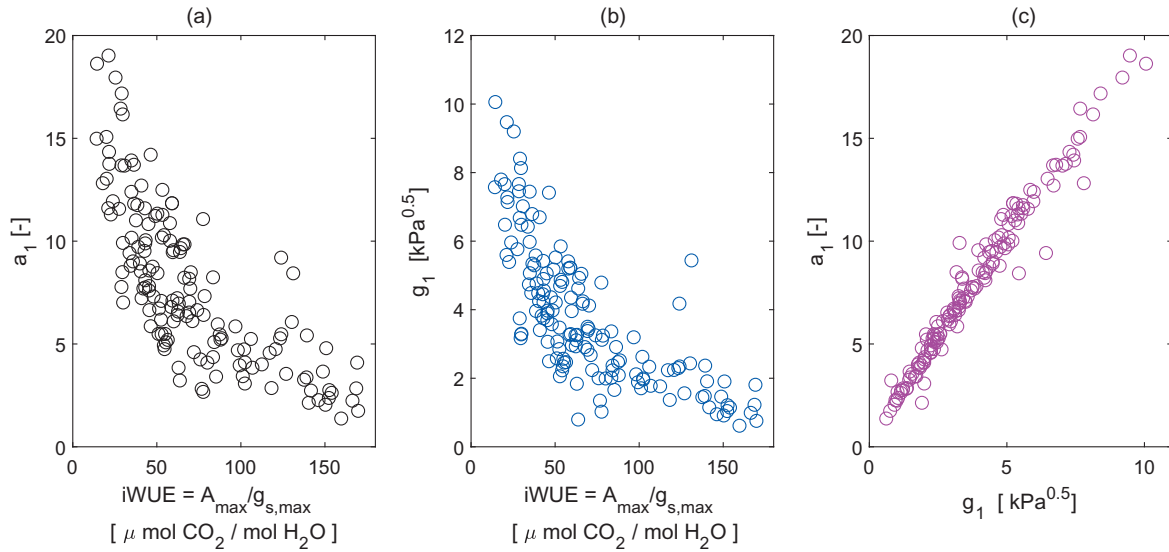


Fig. 6 Scatter plot between $iWUE$, computed from the maximum values of net assimilation $A_{\max}$ and stomatal conductance $g_{s,\max}$ for each species using (a) the parameter a_1 of the Leuning model of stomatal conductance and (b) the parameter g_1 of the Medlyn model of stomatal conductance. The parameters are fitted to the single leaf-gas exchange observations in the database of Lin et al. (2015) and then averaged for species-sites (53 sites and 155 species/sites). (c) Scatter plot between the a_1 and g_1 parameters.

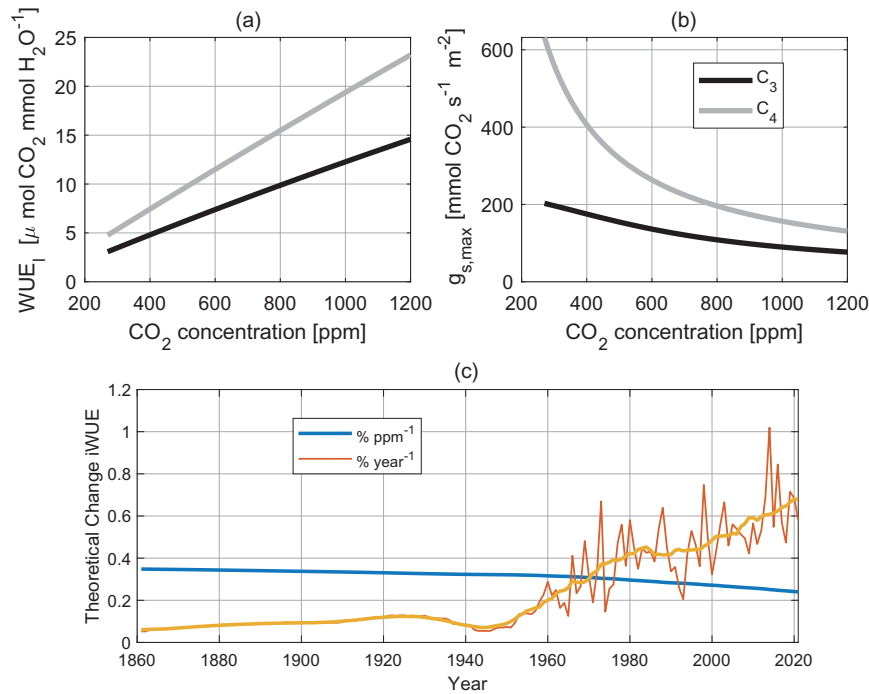


Fig. 7 (a) Leaf scale WUE_i , and (b) stomatal conductance g_s , as a function of atmospheric CO_2 concentration (C_a). Results are obtained with the T&C model for optimal meteorological conditions and realistic parameters for a C_3 and C_4 plants. (c) Expected changes in $iWUE$ [$\% \text{ ppm}^{-1} \text{ CO}_2$] or [$\% \text{ year}^{-1}$] in the period 1860–2020, assuming a constant ratio C_i/C_a , changes in [$\% \text{ year}^{-1}$] are smoothed with a 15-year moving average (orange line).

in $\% \text{ year}^{-1}$ in recent decades as C_a increased with time, the opposite is true for $iWUE$ sensitivity to CO_2 [$\% \text{ ppm}^{-1} \text{ CO}_2$], which decreases as C_a grows (Fig. 7c). More specifically, estimates based on the isotopic content of tree rings suggest that $iWUE$ increased by $0.20\% \text{ ppm}^{-1} \text{ CO}_2$ between 1850 and 2000 in 53 sites worldwide (Leonardi et al., 2012) and $0.27\text{--}0.31\% \text{ ppm}^{-1} \text{ CO}_2$ in 1100 tropical trees (van der Sleen et al., 2015). Other studies found larger trends of $0.46\% \text{ ppm}^{-1} \text{ CO}_2$ (Saurer et al., 2004), $0.44\% \text{ ppm}^{-1} \text{ CO}_2$ (Saurer et al., 2014), $0.41\% \text{ ppm}^{-1} \text{ CO}_2$ (Peñuelas et al., 2011), $0.38\% \text{ ppm}^{-1} \text{ CO}_2$ (Mathias and Thomas 2021), $0.22\% \text{ ppm}^{-1} \text{ CO}_2$ for broadleaf and $0.35\% \text{ ppm}^{-1} \text{ CO}_2$ for conifer trees in Europe (Frank et al., 2015), and a median $0.37\% \text{ ppm}^{-1} \text{ CO}_2$ was

reported during the last 40 years (Silva and Anand, 2013). Analysis of atmospheric CO₂ isotope data also supports an increase in the ¹³C/¹²C isotopic discrimination of land photosynthesis, which suggests a global increase in iWUE during the 20th century of 0.28% ppm⁻¹ CO₂ (Keeling et al., 2017), proportional to the C_a increase. A study using tree ring isotopes show that iWUE sensitivity to CO₂ has considerably declined with iWUE trends decreasing from 0.75% ppm⁻¹ CO₂ to 0.30% ppm⁻¹ CO₂ throughout the 21st century, pointing to a necessary positive trend in C_i/C_a to explain the result (Adams et al., 2020). The computed sensitivity in the early century appears very large in comparison to theoretical expectations and other studies and have been indeed revised to lower values based on new knowledge of mesophyll conductance, which points to minor changes in C_i/C_a (Gong et al., 2022).

CO₂ enrichment experiments such as the Duke and ORNL Free-Air Carbon dioxide Enrichment experiments (FACE) also show a clear increase in ecosystem-scale WUE measured as net primary production over transpiration with changes of 0.15% ppm⁻¹ CO₂ and 0.20% ppm⁻¹ CO₂ for Duke and ORNL, respectively (De Kauwe et al., 2013). These values are on the low-side of tree-ring derived increase in iWUE, but are obtained from forest scale observations with abrupt CO₂ enrichments.

Trends in IWUE can also be estimated using eddy covariance observations of carbon, water, and energy fluxes between the land surface and the atmosphere, although with noticeable uncertainties (Knauer et al., 2018). Using eddy covariance observations, Keenan et al. (2013) detected a large 1.23% ppm⁻¹ CO₂ increase in IWUE across forest sites in the Northern Hemisphere. This increase was surprising as it is more than five times larger than the theoretically expected trend and much larger than trends observed in tree rings and FACE experiments (Medlyn and De Kauwe, 2013). The result prompted other authors to conclude that the observed changes in IWUE should be explained by variables other than C_a and that trends of this magnitude are not a large-scale phenomenon (Knauer et al., 2017). A subsequent study with longer and updated flux tower data reassessed this trend at 0.58% ppm⁻¹ CO₂ (Mastrotheodoros et al., 2017), and model attribution showed that roughly 1/3 of this trend was due to air temperature and VPD effects and the rest explained by changes in C_a and potentially plant trait plasticity—driven by environmental changes and reflected in trends in plant functional traits including, but not limited to, the sensitivity of reference stomatal conductance to CO₂ (e.g., the a₁, g₁, or m parameter).

Terrestrial biosphere model analyses at global scale also report IWUE increases to the order of 0.18–0.21% ppm⁻¹ CO₂ for the 21st century (Ito and Inatomi, 2012; Huang et al., 2015). An analytical WUE model for global terrestrial ecosystems developed by directly upscaling WUE_i globally provided a 0.41% ppm⁻¹ CO₂ trend in WUE^{eco} in the period 1982–2011 (Cheng et al., 2017) with the increase in WUE^{eco} positively related to rising CO₂ concentration and increased canopy LAI, and negatively influenced by increased VPD. More recently, Ueyama et al. (2020) used a global network of flux tower observations combined with a modeling approach, to quantify the sensitivity of GPP and ET to rising C_a during 1990–2014 and estimated an iWUE^{eco} trend of 0.20% ppm⁻¹ CO₂ a little smaller than previous estimates. A study combining different data streams, i.e., tree rings, eddy covariance, and atmospheric observations reported an overall increase in WUE^{eco} of 0.52% ppm⁻¹ CO₂ between 1900 and 2010 (Dekker et al., 2016).

Remote sensing observations also provided spatial patterns of WUE^{eco} trends. Global trends vary in the order of –0.13 to +0.073% ppm⁻¹ CO₂ over the 2000–2013 period (Tang et al., 2014; Xue et al., 2015), which are however at odds with all the other estimates, and they are highlighting the problem of assessing WUE temporal and spatial variability by using remote sensing of vegetation products.

Summary

The capacity of a plant to acquire or store carbon for a given amount of water transpired can be quantified with several metrics that express a water use efficiency (or water productivity, as it is often called for crops). These metrics are fundamental to understand and properly quantify the linkage between the water and carbon cycles and for agronomic purposes. Theoretical considerations suggest that the intrinsic water use efficiency (iWUE) should be constant if C_i/C_a does not vary across species, or more realistically to have a variability in the order of five to seven times when C_i/C_a varies. This half-order of magnitude variability is confirmed by leaf-level gas exchanges, tree ring isotope, and ecosystem scale observations. It also represents a plausible range of variability for essential WUE parameters in the representation of stomatal conductance in terrestrial biosphere models. However, for a given species, long-term iWUE is rather constrained and repeated evidence from agronomy suggests a conservative value of water productivity WP_{B/T} for a given crop once CO₂ and potential evapotranspiration effects are considered. Such a conservative water productivity poses challenges for enhancing agricultural yields by breeding for higher iWUE and it rather suggests that management and improvements of T/ET and HI should lead to higher yields for unit of water used rather than changes in iWUE. In natural vegetation, eddy covariance flux observations and mechanistic modeling reproduce similar patterns of WUE^{eco} with a relatively weak positive dependence of WUE^{eco} on precipitation. These data especially disprove the intuitive thinking that water limited ecosystems should use water more efficiently. Values of WUE^{eco} are indeed generally smaller in arid and semi-arid regions and mechanistic modeling suggests that WUE_T^{eco} is likely lower too. Across sites, WUE^{eco} is negatively correlated with air temperature and solar radiation or, more generally, with potential evapotranspiration as known to occur for crops. At leaf-level and single sites, however, iWUE and IWUE have been shown to be positively correlated with VPD, at least for non-extreme VPD levels. Remote sensing derived WUE values are instead providing different WUE^{eco} patterns in semiarid regions questioning the robustness of their GPP or ET estimates in these sparsely vegetated ecosystems. Finally, multiple literature sources demonstrate that iWUE and IWUE increase with C_a at rates generally larger than the 0.26% ppm⁻¹ CO₂ derived from theoretical expectations. This suggests that other environmental controls (e.g., VPD, air temperature), ecosystem dynamics or plasticity in plant traits might have contributed to the larger than

expected increase of IWUE over the last few decades. While a C_a driven trend in WUE^{eco} will very likely persist in the future, the continuation of the positive feedbacks on the WUE^{eco} trend is less certain with iWUE sensitivity to C_a , which might already be decreasing with time. The continuation of such a strong positive trend in WUE^{eco} , however, might be essential as it affects the trade-offs between carbon uptake and water availability from local to global scales.

References

- Adams MA, Buckley TN, and Turnbull TL (2020) Diminishing CO_2 -driven gains in water-use efficiency of global forests. *Nature Climate Change* 10(5): 466–471.
- Ball J, Woodrow I, and Berry J (1987) A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions. In: *Progress in Photosynthesis Research: Volume 4 Proceedings of the Vllth International Congress on Photosynthesis Providence, Springer Netherlands*, 221–224.
- Beer C, Ciais P, Reichstein M, et al. (2009) Temporal and among-site variability of inherent water use efficiency at the ecosystem level. *Global Biogeochemical Cycles* 23(2): GB2018.
- Blum A (2009) Effective use of water (EUW) and not water-use efficiency (WUE) is the target of crop yield improvement under drought stress. *Field Crops Research* 112: 119–123.
- Cheng L, Zhang L, Wang Y-P, et al. (2017) Recent increases in terrestrial carbon uptake at little cost to the water cycle. *Nature Communications* 8: 110.
- Condon AG, Richards RA, Rebetzke GJ, and Farquhar GD (2002) Improving intrinsic water-use efficiency and crop yield. *Crop Science* 42: 122–131.
- Condon AG, Richards RA, Rebetzke GJ, and Farquhar GD (2004) Breeding for high water-use efficiency. *Journal of Experimental Botany* 55: 2447–2460.
- De Kauwe MG, Medlyn BE, Zaehle S, et al. (2013) Forest water use and water use efficiency at elevated CO_2 : A model-data intercomparison at two contrasting temperate forest FACE sites. *Global Change Biology* 19: 1759–1779.
- de Wit CT (1958) Transpiration and crop yields. In: *Versl. Landbouwk. Onderz. 64.6 Inst. Biol. Chem. Res. On Field Crops and Herbage, Wageningen, The Netherlands*.
- Dekker SC, Groenendijk M, Booth BBB, Huntingford C, and Cox PM (2016) Spatial and temporal variations in plant water-use efficiency inferred from tree-ring, eddy covariance and atmospheric observations. *Earth System Dynamics* 7(2): 525–533.
- Doorenbos J and Kassam AH (1979) *Yield Response to Water*. FAO Irrigation and Drainage Paper No. 33 FAO: Rome.
- Farquhar GD, O'Leary MH, and Berry JA (1982) On the relationship between carbon isotope discrimination and the intercellular carbon dioxide concentration in leaves. *Australian Journal of Plant Physiology* 9: 121–137.
- Faticchi S, Peleg N, Mastrotheodoros T, Pappas C, and Manoli G (2021) An ecohydrological journey of 4500 years reveals a stable but threatened precipitation-groundwater recharge relation around Jerusalem. *Science Advances* 7(37): eabe6303. <https://doi.org/10.1126/sciadv.abe6303>.
- Frank D, Poulter B, Saurer M, et al. (2015) Water-use efficiency and transpiration across European forests during the Anthropocene. *Nature Climate Change* 5: 579–583.
- Hanks JR (1983) Yield and water-use relationships: an overview. In: Taylor HM, Jordan WA, and Sinclair TR (eds.) *Limitations to Efficient Water Use in Crop Production*, pp. 393–411. Madison: ASA.
- Hatfield JL and Dold C (2019) Water-use efficiency: Advances and challenges in a changing climate. *Frontiers in Plant Science* 10: 103. <https://doi.org/10.3389/fpls.2019.00103>.
- Hsiao TC, Steduto P, and Fereres E (2007) A systematic and quantitative approach to improve water use efficiency in agriculture. *Irrigation Science* 25: 209–231.
- Huang M, Piao S, Sun Y, et al. (2015) Change in terrestrial ecosystem water-use efficiency over the last three decades. *Global Change Biology* 21(6): 2366–2378. <https://doi.org/10.1111/gcb.12873>.
- Katerji N, Mastrorilli M, and Rana G (2008) Water use efficiency of crops cultivated in the Mediterranean region: Review and analysis. *European Journal of Agronomy* 28(4): 493–507.
- Keenan TF, Hollinger DY, Bohrer G, et al. (2013) Increase in forest water-use efficiency as atmospheric carbon dioxide concentrations rise. *Nature* 499(7458): 324–327.
- Kijne JW, Barker R, and Molden DJ (eds.) (2003) *Water Productivity in Agriculture: Limits and Opportunities for Improvement*, p. 332. UK: CABI.
- Knaauer J, Zaehle S, Reichstein M, et al. (2017) The response of ecosystem water-use efficiency to rising atmospheric CO_2 concentrations: Sensitivity and large-scale biogeochemical implications. *The New Phytologist* 213(4): 1654–1666. <https://doi.org/10.1111/nph.14288>.
- Lavergne A, Graven H, De Kauwe MG, et al. (2019) Observed and modelled historical trends in the water-use efficiency of plants and ecosystems. *Global Change Biology* 25: 2242–2257.
- Lawlor DW (2013) Genetic engineering to improve plant performance under drought: Physiological evaluation of achievements, limitations, and possibilities. *Journal of Experimental Botany* 64(1): 83–108.
- Leuning R (1995) A critical appraisal of a combined stomatal-photosynthesis model for C_3 plants. *Plant, Cell & Environment* 18(4): 339–355.
- Lin Y-S, Medlyn BE, Duursma RA, et al. (2015) Optimal stomatal behaviour around the world. *Nature Climate Change* 5(5): 459–464. <https://doi.org/10.1038/nclimate2550>.
- Mastrotheodoros T, Pappas C, Molnar P, et al. (2017) Linking plant functional trait plasticity and the large increase in forest water use efficiency. *Journal of Geophysical Research—Biogeosciences* 122(9): 2393–2408. <https://doi.org/10.1002/2017JG003890>.
- Medlyn BE, Barton CVM, Broadmeadow MSJ, et al. (2001) Stomatal conductance of forest species after long-term exposure to elevated CO_2 concentration: A synthesis. *The New Phytologist* 149(2): 247–264.
- Medlyn BE, De Kauwe MG, Lin Y-S, et al. (2017) How do leaf and ecosystem measures of water-use efficiency compare? *New Phytologist* 216: 758–770.
- Medlyn BE, Duursma RA, Eamus D, et al. (2011) Reconciling the optimal and empirical approaches to modelling stomatal conductance. *Global Change Biology* 17(6): 2134–2144.
- Medrano H, Tomás M, Martorell S, et al. (2015) From leaf to whole-plant water use efficiency (WUE) in complex canopies: Limitations of leaf WUE as a selection target. *The Crop Journal* 3: 220–228.
- Morison JLL (1985) Sensitivity of stomata and water use efficiency to high CO_2 . *Plant, Cell & Environment* 8(6): 467–474.
- Paschalis A, Faticchi S, Pappas C, and Or D (2018) Covariation of vegetation and climate constrains present and future T/ET variability. *Environmental Research Letters* 13: 104012.
- Sinclair TR, Tanner CB, and Bennett JM (1984) Water-use efficiency in crop production. *Bioscience* 34: 36–40.
- Steduto P (1996) Water use efficiency. In: Pereira LS, Feddes RA, Gilley JR, and Lesaffre B (eds.) *Sustainability of Irrigated Agriculture. NATO ASI Series (Series E: Applied Sciences)*, vol. 312. Dordrecht: Springer. https://doi.org/10.1007/978-94-015-8700-6_12.
- Steduto P, Hsiao TC, and Fereres E (2007) On the conservative behavior of biomass water productivity. *Irrigation Science* 25: 189–207. <https://doi.org/10.1007/s00271-007-0064-1>.
- Steduto P, Hsiao TC, Raes D, and Fereres E (2012) *Crop Yield Response to Water*. Rome: Food and Agriculture Organization of the United Nations Italy.
- Vico G and Porporato A (2015) Ecohydrology of agroecosystems: Quantitative approaches towards sustainable irrigation. *Bulletin of Mathematical Biology* 77: 298–318. <https://doi.org/10.1007/s11538-014-9988-9>.
- Yang Y, Guan H, Batelaan O, et al. (2016) Contrasting responses of water use efficiency to drought across global terrestrial ecosystems. *Scientific Reports* 6: 23284. <https://doi.org/10.1038/srep23284>.
- Zhang Q, Ficklin D, Manzoni S, Wang L, Way D, Phillips RP, and Novick KA (2019a) Response of ecosystem intrinsic water use efficiency and gross primary productivity to rising vapor pressure deficit. *Environmental Research Letters* 14: 074023.

Further reading

For further reading, articles referenced in the Chapter that could not be listed in the reference list are reported here.

- Adams MA, Buckley TN, Binkley D, Neumann M, and Turnbull TL (2021) CO₂, nitrogen deposition and a discontinuous climate response drive water use efficiency in global forests. *Nature Communications* 12(1): 1–9.
- Briggs LJ and Shantz HL (1913) *The Water Requirement of Plants. I. Investigations in the Great Plains in 1910 and 1911*. US Department of Agriculture, Bureau of Plant Industry Bulletin No. 284. Washington, DC: US Department of Agriculture.
- Briggs LJ and Shantz HL (1914) Relative water requirements of plants. *Journal of Agricultural Research* 3: 1–65.
- Buckley TN (2019) How do stomata respond to water status? *The New Phytologist* 224: 21–36.
- Chapin FS III, Matson PA, and Mooneyk HA (2002) *Principles of Terrestrial Ecosystem Ecology*. New York: Springer.
- Dillman AC (1931) Water requirement of certain crop plants and weeds in the Northern Great Plains. *Journal of Agricultural Research* 42: 187–238.
- Drake BG, Gonzalez-Meler MA, and Long SP (1997) More efficient plants: A consequence of rising atmospheric CO₂? *Annual Review of Plant Physiology and Plant Molecular Biology* 48: 609–639.
- Duursma RA, Blackman CJ, López R, et al. (2019) On the minimum leaf conductance: Its role in models of plant water use, and ecological and environmental controls. *The New Phytologist* 221: 693–705.
- Farquhar GD, Ehleringer JR, and Hubick KT (1989) Carbon isotope discrimination and photosynthesis. *Annual Review of Plant Physiology and Plant Molecular Biology* 40: 503–537.
- Faticchi S, Leuzinger S, Paschalis A, et al. (2016) Partitioning direct and indirect effects reveals the response of water limited ecosystems to elevated CO₂. *Proceedings of the National Academy of Sciences USA* 113(45): 12757–12762. <https://doi.org/10.1073/pnas.1605036113>.
- Franks PJ, Doheny-Adams TW, Britton-Harper ZJ, and Gray JE (2015) Increasing water-use efficiency directly through genetic manipulation of stomatal density. *New Phytologist* 207(1): 188–195.
- Gao X, Gu F, Mei X, et al. (2018) Light and water use efficiency as influenced by clouds and/or aerosols in a rainfed spring maize cropland on the loess plateau. *Crop Science* 58: 853–862.
- Gong XY, Ma WT, Yu YZ, Fang K, Yang Y, et al. (2022) Overestimated gains in water-use efficiency by global forests. *Global Change Biology* 28: 4923–4934.
- Hatfield JL, Sauer TJ, and Prueger JH (2001) Managing soils to achieve greater water use efficiency: A review. *Agronomy Journal* 93: 271–280. <https://doi.org/10.2134/agronj2001.932271x>.
- Helman D, Osem Y, Yakir D, and Lensky IM (2017) Relationships between climate, topography, water use and productivity in two key Mediterranean forest types with different water-use strategies. *Agricultural and Forest Meteorology* 232: 319–330.
- Howell TA (2001) Enhancing water use efficiency in irrigated agriculture. *Agronomy Journal* 93: 281–289.
- Ito A and Inatomi M (2012) Water-use efficiency of the terrestrial biosphere: A model analysis focusing on interactions between the global carbon and water cycles. *Journal of Hydrometeorology* 13(2): 681–694.
- Jung M, Schwalm C, Migliavacca M, et al. (2020) Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the FLUXCOM approach. *Biogeosciences* 17: 1343–1365. 10.5194/bg-17-1343-2020, 2020.
- Kanitkar NV, Gode RB, and Gokhale DH (1943) Water requirement of Rabi Jowar in the scarcity tracts of the Bombay Deccan. *Indian Journal of Agricultural Sciences* 13(3): 235–251.
- Katerji N and Mastroianni M (2009) The effect of soil texture on the water use efficiency of irrigated crops: Results of a multi-year experiment carried out in the Mediterranean region. *European Journal of Agronomy* 30(2): 95–100.
- Keeling RF, Graven HD, Welp LR, et al. (2017) Atmospheric evidence for a global secular increase in carbon isotopic discrimination of land photosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* 114: 10361–10366.
- Knauer J, Zaehle S, Medlyn BE, et al. (2018) Towards physiologically meaningful water-use efficiency estimates from eddy covariance data. *Global Change Biology* 24: 694–710.
- Kukul MS and Irmak S (2017) Spatial and temporal changes in maize and soybean grain yield, precipitation use efficiency, and crop water productivity in the US Great Plains. *Transactions of the ASABE* 60(4): 1189–1208.
- Lambers H and Oliveira RS (2019) Plant water relations. In: *Plant Physiological Ecology*. Cham: Springer.
- Lammertsma EI, de Boer HJ, Dekker SC, Dilcher DL, Lotter AF, and Wagner-Cremer F (2011) Global CO₂ rise leads to reduced maximum stomatal conductance in Florida vegetation. *Proceedings, National Academy of Sciences, United States of America* 108(10): 4035–4040. <https://doi.org/10.1073/pnas.1100371108>.
- Leonardi S, Gentilesca T, Guerrieri R, et al. (2012) Assessing the effects of nitrogen deposition and climate on carbon isotope discrimination and intrinsic water-use efficiency of angiosperm and conifer trees under rising CO₂ conditions. *Global Change Biology* 18(9): 2925–2944.
- Long SP, Ainsworth EA, Rogers A, and Ort DR (2004) Rising atmospheric carbon dioxide: plants FACE the future. *Annu. Rev. Plant Biol.* 55: 591–628.
- Marchand W, Girardin MP, Hartmann H, et al. (2020) Strong overestimation of water-use efficiency responses to rising CO₂ in tree-ring studies. *Global Change Biology* 26: 4538–4558.
- Mathias JM and Thomas RB (2021) Global tree intrinsic water use efficiency is enhanced by increased atmospheric CO₂ and modulated by climate and plant functional types. *PNAS* 118(7): e2014286118. <https://doi.org/10.1073/pnas.2014286118>.
- Medlyn B and De Kauwe M (2013) Biogeochemistry: Carbon dioxide and water use in forests. *Nature* 499(7458): 287–289. <https://doi.org/10.1038/nature12411>.
- Miner GL, Bauerle WL, and Baldocchi DD (2017) Estimating the sensitivity of stomatal conductance to photosynthesis: A review. *Plant, Cell and Environment* 40: 1214–1238.
- Morison JLL and Gifford RM (1983) Stomatal sensitivity to carbon dioxide and humidity: A comparison of two C₃ and two C₄ grass species. *Plant Physiology* 71: 789–796.
- Nobel PS (2009) *Physicochemical and Environmental Plant Physiology*. Oxford, UK: Elsevier Academic Press.
- Novick KA, Ficklin DL, Stoy PC, et al. (2016) The increasing importance of atmospheric demand for ecosystem water and carbon fluxes. *Nature Climate Change* 6: 1023.
- Paschalis A, Faticchi S, Zscheischler J, et al. (2020) Rainfall-manipulation experiments as simulated by terrestrial biosphere models: where do we stand? *Global Change Biology* 26(6): 3336–3355. <https://doi.org/10.1111/gcb.15024>.
- Passioura JB (1996) Drought and drought tolerance. *Plant Growth Regul.* 20: 79–83.
- Peñuelas J, Canadell JG, and Ogaya R (2011) Increased water-use efficiency during the 20th century did not translate into enhanced tree growth. *Global Ecology and Biogeography* 20: 597–608.
- Peters W, van der Velde IR, van Schaik E, et al. (2018) Increased water-use efficiency and reduced CO₂ uptake by plants during droughts at a continental scale. *Nature Geoscience* 11: 744–748.
- Raes D, Steduto P, Hsiao TC, and Fereres E (2018) *AquaCrop: Version 6.0-6.1: Reference Manual: Annexes*. Rome: FAO. <http://www.fao.org/documents/card/en/c/BR244E/>. Accessed on: Aug. 23 2021.
- Saurer M, Siegwolf RTW, and Schweingruber FH (2004) Carbon isotope discrimination indicates improving water-use efficiency of trees in northern Eurasia over the last 100 years. *Global Change Biology* 10(12): 2109–2120. <https://doi.org/10.1111/j.1365-2486.2004.00869.x>.
- Saurer M, Spahni R, Frank DC, et al. (2014) Spatial variability and temporal trends in water-use efficiency of European forests. *Global Change Biology* 20(12): 3700–3712. <https://doi.org/10.1111/gcb.12717>.
- Schymanski SJ and Or D (2016) Wind increases leaf water use efficiency. *Plant, Cell & Environment* 39: 1448–1459. <https://doi.org/10.1111/pce.12700>.
- Seibt U, Rajabi A, Griffiths H, and Berry JA (2008) Carbon isotopes and water use efficiency: Sense and sensitivity. *Oecologia* 155: 441–454.
- Sellers PJ, Randall DA, Collatz GJ, et al. (1996) A revised land surface parameterization (SiB2) for atmospheric GCMs. 1. Model formulation. *Journal of Climate* 9: 674–705.
- Shantz HL and Piemeisel LN (1927) The water requirement of plants at Akron, Colo. *Journal of Agricultural Research* 34: 1093–1190.

- Silva LCR and Anand M (2013) Probing for the influence of atmospheric CO₂ and climate change on forest ecosystems across biomes. *Global Ecology and Biogeography* 22(1): 83–92.
- Stewart JL, Danielson RE, Hanks RJ, et al. (1977) *Optimizing Crop Production Through Control of Water and Salinity Levels in the Soil*. Logan: Utah Water Lab. Pub. no. PRWG 151-1191.
- Sun Y, Piao S, Huang M, et al. (2016) Global patterns of ecosystem WUE. *Global Ecology and Biogeography* 25: 311–323.
- Tang X, Li H, Desai AR, et al. (2014) How is water-use efficiency of terrestrial ecosystems distributed and changing on Earth? *Scientific Reports* 4: 7483. <https://doi.org/10.1038/srep07483>.
- Turner NC (2004) Agronomic options for improving rainfall-use efficiency of crops in dryland farming systems. *Journal of Experimental Botany* 55(407): 2413–2425.
- Ueyama M, Ichii K, Kobayashi H, et al. (2020) Inferring CO₂ fertilization effect based on global monitoring land—Atmosphere exchange with a theoretical model. *Environmental Research Letters* 15: 084009. <https://doi.org/10.1088/1748-9326/ab79e5>.
- van der Sleen P, Groenendijk P, Vlam M, et al. (2015) No growth stimulation of tropical trees by 150 years of CO₂ fertilization but water-use efficiency increased. *Nature Geoscience* 8: 24–28.
- Walker AP, De Kauwe MG, Bastos A, et al. (2021) Integrating the evidence for a terrestrial carbon sink caused by increasing atmospheric CO₂. *New Phytologist* 229(5): 2413–2445. <https://doi.org/10.1111/nph.16866>.
- Wong S-C, Cowan IR, and Farquhar GD (1985) Leaf conductance in relation to rate of CO₂ assimilation. *Plant Physiology* 78: 821–825.
- Wullschlegel SD, Tschaplinski TJ, and Norby RJ (2002) Plant water relations at elevated CO₂—Implications for water-limited environments. *Plant, Cell & Environment* 25(2): 319–331.
- Xue B-L, Guo Q, Otto A, et al. (2015) Global patterns, trends, and drivers of water use efficiency from 2000 to 2013. *Ecosphere* 6(10): 174. <https://doi.org/10.1890/ES14-00416.1>.
- Zhang Y, Kong D, Gan R, et al. (2019b) Coupled estimation of 500 m and 8-day resolution global evapotranspiration and gross primary production in 2002–2017. *Remote Sensing of Environment* 222: 165–182.
- Zhou S, Yu B, Huang Y, and Wang G (2014) The effect of vapor pressure deficit on water use efficiency at the subdaily time scale. *Geophysical Research Letters* 41: 5005–5013.
- Zhu X, Yu G, Wang Q, Hu Z, Han S, Yan J, et al. (2014) Seasonal dynamics of water use efficiency of typical forest and grassland ecosystems in China. *Journal of Forest Research* 19(1): 70–76.
- Zwart SJ and Bastiaanssen WG (2004) Review of measured crop water productivity values for irrigated wheat, rice, cotton, and maize. *Agricultural Water Management* 69(2): 115–133.

Root water uptake and myco-rhizosphere hydraulic properties

Andrea Carminati^a, Maria Marin^b, Valentin Couvreur^c, Mohsen Zarebanadkouki^d, Stéphane Declerck^c, and Mathieu Javaux^c,
^aETH Zürich, Zürich, Switzerland; ^bUniversity of Aberdeen, Aberdeen, United Kingdom; ^cEarth and Life Institute, Agronomy, Université Catholique de Louvain, Louvain-la-Neuve, Belgium; ^dTechnical University of Munich, Freising, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	543
Principles of water flow from the soil into the roots	544
Soil water flow	544
Water flow in the root	546
Rhizosphere processes and their impact on root water uptake	546
Root hairs	546
Mucilage	547
Mycorrhizal fungi	548
Rhizosphere hydraulic properties	549
Root water uptake and rhizosphere properties in their ecological and environmental context	550
References	550

Abstract

The ability of plant roots to extract water from the soil at the rate required to sustain the transpiration rate is constrained by the soil hydraulic conductivity. As soils dry, their hydraulic conductivity decreases by several orders of magnitude and the root water potential required to drive soil water flow rapidly drops. Eventually, the soil can no longer sustain the transpiration rate. Plants have multiple strategies to cope with soil hydraulic limitations: besides stomatal regulation, plants alter the hydraulic properties of the rhizosphere by secreting viscous compounds, growing hairs and hosting mycorrhiza.

Key points

- Water loss from leaves creates a potential gradient in the plant that drives water uptake from soil.
- Stomata regulate the rate of water loss from leaves.
- At high rates of water loss from leaves, the rate of water movement to roots through soil can be the limiting factor of transpiration.
- Roots can increase the rate of water capture from soil by improving root-soil contact, secreting compounds or by associations with mycorrhizal fungi.

Introduction

The uptake of soil water by roots is driven by water losses from the leaves into the atmosphere. Plants are big movers of water and understanding the factors controlling root water uptake is necessary for properly modelling and predicting water exchange between the soil and the atmosphere. Root water uptake is driven by a gradient in water potential across the soil-plant-atmospheric continuum. Water evaporates from the leaves into the atmosphere across stomata. The transpiration rate depends on the difference in vapor pressure between stomata and the atmosphere (also referred to as vapor pressure deficit or VPD) and on the canopy conductance. The latter depends on stomatal conductance (which plants can regulate at short time scale) and boundary layers.

The water loss from the leaves generates a decrease in leaf water potential that passively drives the water from the soil, into the roots, along the xylem, and finally to the stomata. The gradients in water potential that drive water along this pathway depend on the hydraulic conductivities of key elements: the soil, the rhizosphere (soil in vicinity of and affected by roots), the root-soil interface, the roots (including its radial and axial components and their arrangement along the root architecture), and the xylem up to the leaves. All these conductivities are variable and are affected, to a different extent, by water potential.

In this chapter, we focus on the conductivity of the soil, of the root, and of the mycorrhizosphere (Fig. 1). Particular attention is dedicated to the water flow across the rhizosphere, which is defined as the soil in the vicinity of and affected by roots (i.e., a few mm around the roots). Large gradients in water potential occur across the rhizosphere in dry soil with low conductivity and eventually trigger stomatal closure. Plants have mechanisms to mediate these gradients by growing root hairs, symbiosis with mycorrhizae or by altering the rhizosphere hydraulic properties, such as mucilage secretion. After a short review of water flow in soil and roots, we discuss physical and biological mechanisms shaping water flow across the rhizosphere and the impact of root hairs, mycorrhiza and mucilage on root water uptake.

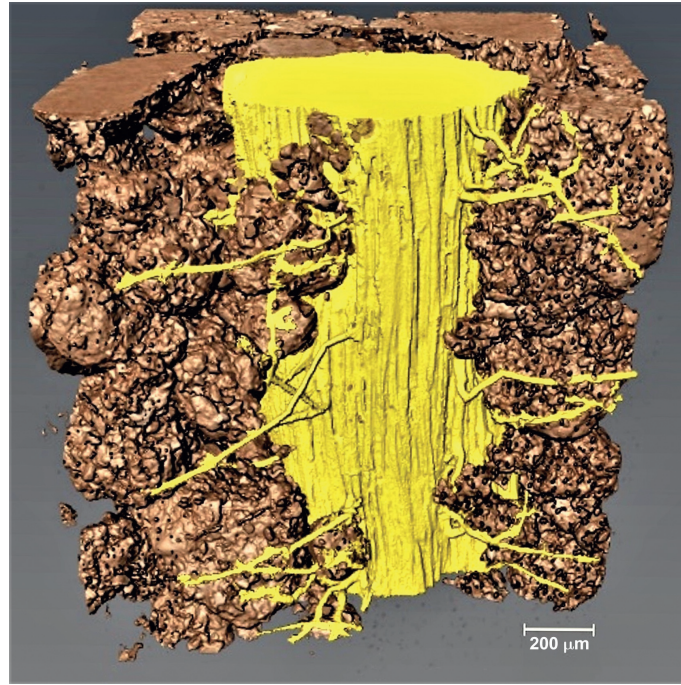


Fig. 1 Rendering of a synchrotron-based X-ray CT of a maize root in a loamy soil with root hairs protruding across the rhizosphere. Image courtesy by Patrick Duddek and Mutez Ahmed. Details on the image are in Duddek P, Carminati A, Koebernick N, Ohmann L, Lovric G, Delzon S, Rodriguez-Dominguez CM, King A, and Ahmed MA (2022) The impact of drought-induced root and root hair shrinkage on root–soil contact. *Plant Physiology*.

Principles of water flow from the soil into the roots

Soil water flow

Soil water fluxes are driven by gradients in soil water potential. In soils, the water potential [hPa] is made of three components: gravimetric (ψ_g), matric (ψ_m) and osmotic (ψ_o) potentials, with the latter not affecting water flow in soils. Volumetric soil water content θ [$\text{m}^3 \text{m}^{-3}$] and matric potential ψ_m are related through a function called the soil water retention curve. The amount of plant available water in a soil depends on typical points of this retention function (field capacity and wilting points) and on the plant root depth. Soil water holding capacity provides a buffer for plants to endure dry spells and is an important bulk soil characteristic in the long term (days-weeks). Typical values for soil water holding capacity range between 4% and 30%.

Besides its ability to store water, soil is also characterized by its ability to conduct water, called soil hydraulic conductivity $k(\theta)$ [m s^{-1}]. This is a highly nonlinear function of water content because not only do the number of wet pores decrease with water content, but also their diameter, and thereby the resistance that water has to overcome to flow. In addition, smaller pores are often more tortuous and less connected. The soil conductivity curve controls soil water fluxes within the soil profile and toward roots. In the rhizosphere, the role of the soil hydraulic conductivity is crucial as it generates an important resistance when the soil dries out.

The soil water flux is a function of the soil water potential gradient, which is described by the Buckingham-Darcy equation:

$$q = -k(\theta) \left(\frac{\partial h}{\partial x} + \frac{\partial z}{\partial x} \right) \quad (1)$$

where $\mathbf{x}$ is the position vector ($\mathbf{x} = [x, y, z]$) [m], the matric head is $h = \frac{\psi_m}{\rho g}$ [m], ρ is the water density [kg m^{-3}] and g is the gravitational acceleration [m s^{-2}].

The transfer of water within a soil profile is governed by the Richardson-Richards equation, which in three dimensions is expressed as:

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial \mathbf{x}} \left[k(\theta) \left(\frac{\partial h}{\partial \mathbf{x}} + \frac{\partial z}{\partial \mathbf{x}} \right) \right] - S \quad (2)$$

where t is time [s] and S is the sink term corresponding to the volumetric root water uptake ($S > 0$) or release ($S < 0$) [$\text{m}^3 \text{m}^{-3} \text{s}^{-1}$]. The sink term in Eq. (2) represents the net uptake per volume of soil, V [m^3]. In the literature, two different approaches have been proposed for estimating S . Some authors define S as a function of macroscopic volumetric properties like root length density or soil

water status like matric potential or matric flux potential (van Lier et al., 2006). On the other hand, microscopic approaches compute S from an explicit simulation of the segment root water uptake:

$$S = \frac{\sum_i j_{r,i} A_i}{V} \quad (3)$$

where $j_{r,i}$ [m sec^{-1}] is the root water uptake per root surface (A_i) of a root segment i ($i = 1:n$) located within the soil volume V . In such a case, j_r must be explicitly calculated for each root segment between bulk soil and root xylem and will be a function of the hydraulic conductance in the rhizosphere K_{rh} [$\text{m s}^{-1} \text{MPa}^{-1}$] and in the root radial pathway K_r [$\text{m s}^{-1} \text{MPa}^{-1}$]:

$$j_r = \frac{1}{\frac{1}{K_{rh}} + \frac{1}{K_r}} (\psi_b - \psi_x) \quad (4)$$

where ψ_b and ψ_x are the water potential (including all the three components: gravimetric, matric/hydraulic and osmotic) in the bulk soil and the root xylem, respectively [MPa].

Calculating K_{rh} can be extremely challenging given the complex nature of the rhizosphere and the different possible pathways for water between the bulk soil and the root surface. Different pathways have been identified: the matrix pathway, i.e., through the soil pores, the root hair pathway, and the mycorrhiza pathway. The two latter pathways will be further developed in the last sections.

The flow through the soil matrix is obtained by solving the Richardson-Richards equation axisymmetrically, by considering the concentric flux of water from the bulk soil to the root surface (and neglecting gravity):

$$\frac{\partial \theta}{\partial t} = \frac{1}{r} \frac{\partial}{\partial r} \left[rk(\theta) \left(\frac{\partial h}{\partial x} \right) \right] \quad (5)$$

where r is the radial coordinate [m]. Numerous authors have provided semi-analytical solutions to this equation under steady-state ($\partial \theta / \partial t = 0$) or steady rate ($\partial \theta / \partial t = \text{constant}$) using the flux matric potential approach (van Lier et al., 2006). One important aspect of these solutions is that the soil resistance to water flux becomes a function of the water flux itself. An increase of transpiration demand generates a lower water potential at the root interface, which decreases the hydraulic conductivity in the soil. Solving Eq. (5) using steady rate conditions allows calculating an effective rhizosphere conductance K_{rh} for varying j_r (Fig. 2).

In addition to the conductivity drop, rhizosphere matric conductance also depends on the contact between the root surface and the liquid phase. If the contact between root and the liquid phase decreases, water uptake concentrates on a narrow part of the root-soil interface and further increases the conductivity drop. When the soil dries, soil and roots may shrink and adversely affect the partial contact between soil and roots. For example, Faiz and Weatherley (1982) found that the roots of sunflowers shrank by 25% when the leaf water potential dropped to around -1.5 MPa. These observations suggest that under soil drying conditions, the contact between roots and their surrounding soil is reduced due to root and soil shrinkage, decreasing the hydraulic conductance of the rhizosphere and limiting water flow between roots and soil.

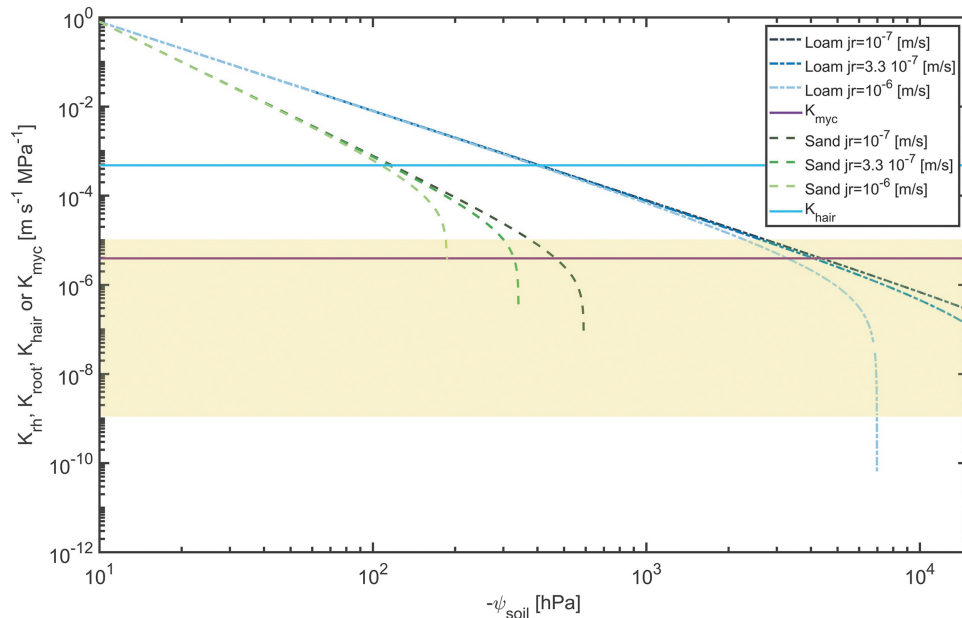


Fig. 2 Hydraulic conductance of the rhizosphere K_{rh} [$\text{m s}^{-1} \text{MPa}^{-1}$] of a sandy and a loamy soil and for three water fluxes at the root-soil interfaces j_r [m s^{-1}]. The large yellow bad is the range of root radial conductance K_r [$\text{m s}^{-1} \text{MPa}^{-1}$], according to Draye et al. (2010). K_{hair} and K_{myc} are the hydraulic conductances of the root hair zone and mycorrhiza estimated using the Hagen-Poiseuille law (Eq. 8).

Water flow in the root

Down to the subcellular scale, primary cell walls are “corridors” for water between protoplasts that contain the cytosol and organelles. Just like in soil, in these fully hydrated corridors whose porosity is about 70%, water flow is driven by gradients in water potential, which is given by the hydrostatic (Ψ_m) (similar to the matric potential defined in soil) and gravitational (Ψ_g) potential. However, pore size in cell walls is extremely small – about $5 \cdot 10^{-9}$ – relative to the dominant porosity class in soils with the finest texture – the clay “ultra-microporosity” spanning the 10^{-7} to $5 \cdot 10^{-6}$ m spectrum. As a consequence, despite large uncertainties, the hydraulic conductivity of cell walls is thought to be orders of magnitude lower than that of heavy clay soils at water saturation. The respective orders of magnitude are: 10^{-10} to 10^{-7} versus 10^{-6} to 10^{-4} m² MPa⁻¹ s⁻¹ as reviewed by [Couvreur et al. \(2018\)](#).

The boundary between the cell wall and the cytosol is the plasma membrane. The inside of this phospholipid bilayer is hydrophobic, which drastically limits the transport of water. However, “gates” to the cytosol exist under the form of transmembrane water channels called aquaporins. Because they decouple the transport of water and solutes, membranes allow the formation of regions with distinct compositions in solutes and different osmolalities. Membranes thereby allow the process of osmosis, which explains how water molecules tend to spontaneously flow toward the side with highest osmolality as a result of their high affinity for solutes. To capture it mathematically, a new component is added to the potential of water: the osmotic potential (Ψ_o , more negative with increasing osmolality). Applying the second law of thermodynamics to water transport, water passively flows down differences of summed Ψ_m , Ψ_g and Ψ_o across selective membranes (the sum of these components is also called total water potential Ψ_{tot}). A membrane failing to decouple water and solute transport is considered as “leaky.” A direct consequence of the leakiness is that the osmotic potential difference across the membrane is not as effective at driving water flow, and Ψ_o is then corrected by a “reflection coefficient” σ equal to zero if the membrane is fully permeable and up to one when fully selective. These principles translate mathematically as follows:

$$j_m = Lp (\Psi_{m,w} - \Psi_{m,c} + \sigma(\Psi_{o,w} - \Psi_{o,c})) \quad (6)$$

where j_m [m s⁻¹] is the water flow rate per unit membrane surface, Lp [m MPa⁻¹ s⁻¹] is the membrane hydraulic conductance per unit surface, $\Psi_{m,w}$ and $\Psi_{m,c}$ [MPa] are the pressure potentials on the sides of the cell wall and cytosol, respectively, and $\Psi_{o,w}$ and $\Psi_{o,c}$ [MPa] the (negative) osmotic potentials on the sides of the cell wall and cytosol, respectively. Note that the gravitational potential is usually neglected at this scale. Lp is defined as conductance per root surface.

The composite nature of the root subcellular structures arranged in parallel and in series allows multiple pathways for water from the root surface to xylem vessels. The apoplastic, transmembrane and symplastic pathways (the last two frequently referred to as the combined cell-to-cell pathway) rely dominantly on water transport through cell walls, plasma membranes and plasmodesmata, respectively. While they have been conceptualized as pathways in parallel with their respective conductivities and reflection coefficients, recent unified modelling and experimental approaches suggest strong alterations of water pathways along the root radius ([Couvreur et al., 2018](#)) and the absence of a purely apoplastic pathway. Hydrophobic root diffusion barriers in the endodermis (and occasionally exodermis) play a key role in both of these observations. The lignin based Casparian strip forms a scaffold in radial cell walls and the suberin lamellae is deposited between all membranes and primary cell walls, drastically limiting permeation across these cell walls and membranes, respectively, in the concerned cell layers. Thereby, these structures locally force water through transmembrane or symplastic pathways, while functionally creating a filter for nutrients responding to the direct environment of roots. There appears to be a high level of coordination between the deposition of diffusion barriers and the expression of specific aquaporins in their direct vicinity so that the relative importance of these multiple factors has been hard to decipher.

The mainstream model for radial water flow across vascular plant roots has remained untouched for multiple decades and is a common building block in Functional-Structural Plant Models. It simply assumes that root tissues are analogous to a “big-membrane,” hence its formulation very close to Eq. (6):

$$j_r = k_r (\Psi_{m,s} - \Psi_{m,x} + \sigma_r(\Psi_{o,s} - \Psi_{o,x})) \quad (7)$$

where j_r [m s⁻¹] is the radial water flux at the root surface, k_r is the root radial conductance per root surface as defined in Eq. (4), σ_r is the root reflection coefficient, $\Psi_{m,s}$ and $\Psi_{m,x}$ [MPa] are the water pressure potentials at root surface and in xylem, respectively, and $\Psi_{o,s}$ and $\Psi_{o,x}$ [MPa] the (negative) osmotic potentials at root surface and in xylem, respectively.

Measured k_r values range between 10^{-9} and 10^{-5} [m s⁻¹ MPa⁻¹] according to [Draye et al. \(2010\)](#) (yellow region in Fig. 2).

Rhizosphere processes and their impact on root water uptake

Root hairs

Root hairs represent about 2% of the root mass but they greatly increase the root surface area, which can be up to 3-fold larger in the presence of root hairs. The volume of soil influenced by roots in the presence of root hairs is further increased by their ability to access finer pores than the main root axis.

Controlled laboratory studies have demonstrated that root hairs facilitate water uptake. [Segal et al. \(2008\)](#) reported a greater depletion of soil water in the root hair zone of the barley (*Hordeum vulgare*) wild type compared with the hairless mutant. Similarly, [Tanaka et al. \(2014\)](#) showed a 47% reduction in water absorption for the hairless mutant of *Arabidopsis thaliana* relative to the wild type. In terms of plant water status, under low evaporative demand in controlled environmental conditions, [Dodd and Diatloff](#)

(2016) showed no difference in the transpiration rate of the no root hairs mutant of barley and the wild-type plants. In contrast, under higher evaporative demands, the ability of root hairs to take up water from drying soil was reported by Carminati et al. (2017b). As soil dried, the leaf water potential of the 'no root hairs' mutant decreased more rapidly than that of the wild type, implying a reduced capacity to take up water at high transpiration rates. Root hairs facilitate water uptake by bridging the gaps between roots and soil, either by directly taking up water or by creating a capillary bridge for water to flow across the gap (Carminati et al., 2017b) and thereby reducing the decline in water potential at the root-soil interface in rapidly transpiring plants.

Field evidence of the significance of root hairs for water uptake by barley under drought was reported by Marin et al. (2021). The field experiment was carried out in Scotland during a typically wet growing season (2017) and during the driest growing season ever recorded at the site (2018), under contrasting soil textures (clay loam and sandy loam). The study showed that root hairs did not confer a notable advantage to barley under optimal conditions, but under soil water deficit, root hairs enhanced plant water status and stress tolerance. This resulted in less negative leaf water potential and lower leaf abscisic acid concentration in the presence of root hairs for plants grown in clay loam, while no significant effect of root hairs was found in sandy loam soil (Marin et al., 2021).

The role of root hairs in plant water uptake has been elucidated mainly in barley. Cai et al. (2021) evaluated the effects of maize root hairs on plant-soil water relations in sand and loam and found no clear effect of root hairs in either soil textures. This might be related to the early shrinkage of root hairs in maize (Duddek et al., 2022). Root hairs show intra- and interspecific variations in length and density. In angiosperms, root hairs vary in length from zero (i.e., species with no root hairs) to 1.5 mm. If comparing barley and maize for instance, barley produces significantly longer root hairs (2-fold) and at a greater root hair length density (6-fold) than maize. Therefore, the role of root hairs in plant water uptake is equivocal, apparently being species and soil specific.

To give an approximate estimation of the capacity of root hairs to transport water, we calculate the effective conductance of root hairs, treating them as capillaries arranged in parallel and using the Hagen-Poiseuille law:

$$j_r = \frac{N \pi r^4 \Delta\psi}{8 \mu L} \quad (8)$$

where j_r is the water flux at the root soil interface [m s^{-1}], N is the number of root hairs per root surface [m^{-3}], r is the hair radius, $\Delta\psi$ is the difference in water potential, μ is viscosity [Pa s] and L is the hair length [m]. K_{hair} is calculated as:

$$K_{\text{hair}} = \frac{N \pi r^4}{8 \mu L} \quad (9)$$

Using $N = 9.5$ (assuming 30 hairs per root length and a root radius of 0.5 mm), hair radius $r = 6 \mu\text{m}$, and setting $L = 10 \text{ mm}$ (to compare it to other domains that were set to 10 mm), we obtain $4.8 \times 10^{-3} \text{ m s}^{-1} \text{ MPa}^{-1}$ (Fig. 2). This is a pretty large value and it might be an overestimate by assuming that viscosity is the same as water and by not adding the permeability of the root membrane. Water has to cross at least one membrane to be taken up by roots. However, it shows how the number of hairs and their radius is potentially sufficient to constitute a significant parallel pathway for water uptake.

Mucilage

Plants contribute a large portion of their photosynthetic products to the soil via their roots in a process called rhizodeposition. Rhizodeposition comprises the release of water-soluble components, the secretion of insoluble components, such as mucilage, the sloughing off of root border cells, the senescence of root epidermis and root hairs, and the release of gases such as CO_2 and ethylene. Mucilage is a polymeric gel that is mainly secreted by the Golgi apparatus of the root cap, but it can also be secreted by the epidermal cells of mature roots (McCully and Boyer, 1997). Its formation is readily visible in air-born roots of maize (Fig. 3). The root tip continuously secretes mucilage in the soil, and as the root elongates, mucilage is translocated along roots. Major components of root mucilage are galactose (~39–42%), fucose (~22–30%), mannose (~11–14%), arabinose (~8–11%), xylose (~1–4%), glucose (~1–4%) and glucuronic acid (~3–5%).

The chemical composition and quantity of mucilage varies across species, developmental stages of plants, and environmental conditions. Despite these variations, mucilage shares common intrinsic properties. Mucilage increases viscosity and lowers the surface tension and water potential of the soil solution. Read and Gregory (1997) showed that the viscosity of maize mucilage is twice that of water already at concentrations of 0.7 mg g^{-1} (weight of dry mucilage per unit weight of the liquid solution), and it increases non-linearly as the concentration of mucilage increases (1000 times higher than pure water at ca. 10 mg g^{-1}). The higher viscosity of mucilage compared to water causes a reduction in the hydraulic conductivity of the rhizosphere. At low concentrations, mucilage increases the saturated hydraulic conductivity of soils with small mean particle size (fine-textured soils), probably due to its swelling capacity and the consequent increase in soil porosity. However, as mucilage concentration increases, the saturated hydraulic conductivity decreases dramatically. Mucilage has a smaller surface tension than pure water, probably due to the presence of phospholipid components in the root exudates. Read and Gregory (1997) found that the surface tension rapidly declined with increasing mucilage concentration, decreasing from $72 \text{ [mN m}^{-1}]$ down to $45\text{--}50 \text{ mN m}^{-1}$ at concentrations of 1 mg g^{-1} in water. McCully and Boyer (1997) found that when fully saturated mucilage exuded by the nodal roots of the maize plant can hold an amount of water 300 times its dry weight, but a large fraction of this water could not be held against negative water potentials as the soil dries. These authors concluded that mucilage should not play a significant role in soil-plant water relations. However, when mucilage is embedded into a solid matrix, it increases the water holding capacity of the soil at any given water potential (Carminati



Fig. 3 Mucilage secreted by nodal roots of the maize plant. The photo was taken a day after a rain event in the field.

et al., 2017a). This evidence suggests that the effect of mucilage on soil water retention must be amplified in soils by additional forces that emerge from the interaction between polymers and soil matrix.

Mucilage of most plants contains a small fraction of phospholipids and anionic polysaccharides. These components are amphiphilic substances that contain both hydrophilic and hydrophobic functional regions. Therefore, when mucilage dries, the hydrophilic regions of these components may bind to each other or to the soil and expose the hydrophobic parts toward the air-filled phase of the pore space, causing water repellency in the rhizosphere (Hallett et al., 2003).

Several hypothetical functions have been attributed to mucilage such as impacting the dynamics of nutrients in the rhizosphere, keeping the soil surrounding the root tip wet and hydrated (McCully and Boyer, 1997) and therefore protecting root tips from dehydration in drying soil, acting as a lubricant at the root surface, thereby improving root penetration in soil, and improving the contact between roots and soil particles. Another putative function attributed to mucilage is the facilitation of water and nutrient uptake by roots. This beneficial function is related to an improved contact between roots and the soil, improving the soil's water-holding capacity, and improved connectivity of the liquid phase within soil pores by altering the liquid configuration. Carminati et al. (2017a) conceptualized the function of mucilage in the retention and flow of water in drying soil. They showed that the combination of increased viscosity and decreased surface tension of the liquid phase could alter the spatial configuration of the liquid phase within the soil pores, enhancing the connectivity of the liquid phase during soil drying.

Mycorrhizal fungi

The vast majority of land plants form associations with mycorrhizal fungi. Four main types of mycorrhizas are recognized based on morphological characteristics of root tissues and host plants: arbuscular mycorrhizas, ectomycorrhizas, ericoid mycorrhizas and orchid mycorrhizas, that form associations with 72%, 2%, 1.5% and 10% of vascular plants, respectively. Arbuscular mycorrhizal fungi (AMF) are therefore, by far, the more widespread, distributed in the majority of terrestrial ecosystems and which concern the vast majority of agricultural and horticultural crops. We will thus focus our attention on this main guild of mycorrhizal fungi.

Arbuscular mycorrhizal fungi are obligate root symbionts that have co-evolved with land plants for more than 400 million years. Their roles in plant nutrition and growth as well as alleviation of biotic and abiotic stresses have been repeatedly demonstrated, making these soil microorganisms key players in crop production (Smith and Read, 2008) and plant ecology in natural systems. AMF colonize the cortical cells of the root system, developing intraradical mycelium (IRM) and arbuscules, and extend tens of cm away from the roots into the soil as extraradical mycelium (ERM), colonizing soil pores with hyphae having diameters up to 10 times finer than that of root hairs. They produce up to 100 m of hyphae per gram of soil in pastures and prairies (Miller et al., 1995). ERM can thus be considered as an extension of the root system, increasing the uptake of minerals (e.g., phosphorus, nitrogen) and water, meeting plant nutritional needs, in exchange for organic compounds issued from the photosynthesis.

The role of AMF in the water supply of plants has been reported since the early 1970s. Safir et al. (1971) were the first to suggest that the symbiosis of AMF with soybean affects the water relations, probably via improved phosphorus (P) nutrition. This

hypothesis was supported up to the 1980s until a couple of studies demonstrated that AMF could also affect water relations by other mechanisms than P nutrition (Harris et al., 1985). Since then, several hypotheses have been proposed and are still being explored.

Fungal ERM allow AMF-colonized plants to explore more soil, accessing pores unreachable to roots and root hairs, taking up more water than non-colonized plants and transporting water via the hyphae from far distances to the roots. The tridimensional ERM structures maintain water continuity from soils to plant by by-passing air gaps. The hyphal length per unit soil mass (m g^{-1}) reported in the field varies from 68 to 101 in prairie (Miller et al., 1995), 45 to 74 in ungrazed pasture (Miller et al., 1995) and 10 to 35 in maize cropping (Kabir and Koide, 2002). The diameter of hyphae in the soil generally varies from 1.2 to 18 μm with the smallest ones involved in uptake being approximately 2 μm . Their growth is estimated between 0.3 and 3.3 mm a day with values reported for *Rhizophagus fasciculatum* symbiotic with *Trifolium subterraneum* of 10 m g^{-1} after four weeks and 20 m g^{-1} after 5 weeks (Scheltema et al., 1985).

AMF strongly affects the expression of aquaporins (i.e., water channels) involved in the transport of water. Aquaporins have been reported in the hyphae developing within the soil (favoring the entry of water within the hyphae) as well as in the peri-arbuscular membrane, which is the membrane that surrounds an arbuscule and is contiguous to the plasma membrane of the cortical cell (it plays a central role in nutrient exchange between the symbionts). Kikuchi et al. (2016) proposed a model in which plant transpiration creates a water potential gradient, allowing water to move within the extraradical hyphae through the cytosol toward the cortical cells. A mycorrhiza-inducible plant aquaporin on the periarbuscular membrane and an aquaglyceroporin located on the fungal plasma membrane (responsible for water transport across the plasma membrane) further mediate the water flow from fungal cell to root cell. Thus, root colonization impacts gene expression of aquaporins, which is correlated with changes in root hydraulic properties that increase water transport capacity of mycorrhizal roots. In the model proposed by Kikuchi et al. (2016), water may also move through the lumen of tubular vacuoles containing aquaporins. Interestingly, inorganic phosphorus accumulates within these tubular vacuoles and is translocated toward the roots by the water flow. Host transpiration and the fungal aquaporin thus play key roles in P translocation.

Stimulation of root hair growth and protection of host plants against the oxidative damage under abiotic, including osmotic stresses, have also been reported. Gas exchange capacities were also increased in AMF-colonized plants, probably associated with a decrease in stomatal resistance and increase in transpiration as reported by Zhu et al. (2011) in maize. Finally, AMF have been shown to impact plant water uptake capacity indirectly by the production of exudates, which alter soil pore structure through aggregation and can also influence hydrophobicity.

We repeated the calculations described above for root hairs but using $r = 1 \mu\text{m}$, $L = 1 \text{ cm}$ and $N = 100$ (estimated assuming 20 m of ERM per gram of soil, a root length density of 2 cm cm^{-3}) appropriate for AMF hyphae. Using Eq. (8), we obtain $K_{\text{myc}} = 3.9 \cdot 10^{-6} [\text{m s}^{-1} \text{ MPa}^{-1}]$ (Fig. 2). This value is much smaller than that calculated for root hairs, due to the smaller radius of ERM hyphae, however it is larger than the conductance of a sandy soil for soil matric potentials more negative than -200 hPa to -500 hPa , while it needs 10 times more negative water potential in a loam for mycorrhiza to be as conductive as the soil matrix. This simple calculation suggests that a significant flow across mycorrhizae can take place, particularly in coarse textured soils.

Rhizosphere hydraulic properties

Plant water uptake from the soil is influenced by the hydraulic properties of the rhizosphere, which differ from those of bulk soil as a result of biophysical processes occurring at the root-soil interface as discussed above. Processes affecting rhizosphere hydraulic properties include soil compaction by root growth, air-filled gaps caused by roots and soil shrinking and swelling during drying and wetting cycles, root hair intrusion, and mucilage production. Imaging methods have provided insight into the dynamics of water content and soil structure around the roots. X-ray CT has been particularly useful for resolving the soil structure dynamics around the roots. Aravena et al. (2011) showed that root-induced compaction increases contact areas between aggregates, leading to an increase in unsaturated hydraulic conductivity of the rhizosphere soil. Koebemick et al. (2017) investigated the effect of root hairs on soil porosity and pore connectivity, reporting increased soil macroporosity in the rhizosphere in the presence of root hairs. As roots shrink, they create gaps between soil and roots, reducing water transfers and therefore helping plants to prevent water loss when the soil dries.

Neutron radiography and tomography have been useful to reveal unexpected gradients in water content around the roots, with increasing water contents toward the root surface during the drying phase (Carminati et al., 2010). These contradictory observations were explained as the effect of mucilage exuded by roots on water holding capacity and rhizosphere hydrophobicity of the rhizosphere. Hallett et al. (2003) directly measured the hydraulic characteristics of the rhizosphere using a miniaturized infiltrometer device and found reduced water sorptivity in moist rhizosphere soil of barley compared with bulk soil, due to increased water repellency in the rhizosphere.

The increased hydrophobicity induced by mucilage in the rhizosphere also results in a high contact angle between liquid phase and soil particles (Moradi et al., 2012). Zarebanadkouki et al. (2016) found that during a drying and wetting cycle in a sandy soil, the hydraulic conductivity of the root-soil interface creates a temporary limit to root water uptake as a result of rhizosphere hydrophobicity. They showed that water repellency of the rhizosphere reduced root water uptake by a factor of 6.5 times during the first 2–3 h after irrigation. Despite the occurrence and mechanisms of water repellency in the rhizosphere being partly clarified, their consequences on emerging water relations between soil and plants remain unclear. The reason is that rhizosphere hydrophobicity is often observed as a local phenomenon along the root system and it is not clear if this local limitation of root water uptake has any

impact on the overall soil-plant water relationship. Rhizosphere hydrophobicity might even be a positive rhizosphere trait that hydraulically disconnects the root segments from the driest soil layers, therefore reducing leakage of water from root system to the dry soil. Rhizosphere hydrophobicity could locally and temporarily disconnect the root segment located in the driest soil layers, therefore reducing the risk of hydraulic failure.

Rhizosphere hydraulic properties also differ from those of bulk soil also as a result of other biofilms present in the soil, such as extracellular polymeric substances (EPS) secreted by bacteria and exudates from mycorrhizal fungi. Similar to mucilage, these materials connecting soil particles do not break up upon drying, due to their high viscosity and low surface tension, enhancing the soil water retention and connectivity of the liquid phase.

Root water uptake and rhizosphere properties in their ecological and environmental context

The physics of soil water flow indicates that soil drying and the consequent drop in soil hydraulic conductivity limit the availability of water to plants. The drop in soil hydraulic conductivity around the roots is exacerbated by soil and root shrinkage and the consequent reduction in root-soil contact, which causes a further loss of conductivity.

Plants can respond to such limitations in soil water availability with multiple strategies. One strategy is to reduce the transpiration demand by closing stomata, with the cost of reduced growth and carbon assimilation. A second strategy is to increase the surface of roots active in water uptake. A combination of both strategies is the allocation of more carbon to roots compared to shoot, which reduces the ratio between transpiration and root surface active in water uptake. An additional strategy is the modification of rhizosphere properties.

The three previous sections reviewed the several mechanisms that plants have to engineer the physical properties of their rhizosphere. Growth of root hairs and symbiosis with mycorrhiza increase the capacity of roots to extract water from dry and low conductive soils and thus increase the plant water potential needed to sustain transpiration during drought. However, their effectiveness is soil and plant specific. The additional strategy for plant roots is to modify their rhizosphere by the secretion of mucilage. Mucilage secretion leads to a wetter and more hydraulically connected rhizosphere compared to the bulk soil during drying, and to a less severe decrease in unsaturated hydraulic conductivity and diffusion coefficient. On the other hand, mucilage induces a temporary water repellency in the rhizosphere, particularly in sandy soils, which might be functional to maintain the region around the roots aerated upon rewetting and regulate water efflux from roots to dry soils layers. Despite its potential to maintain the connectivity between soil and roots, the relevance of mucilage on root water uptake has not yet been demonstrated due to challenges in quantifying the amount and properties of mucilage in-situ and to the lack of appropriate mucilage deficient genotypes. EPS produced by microorganisms have similar properties as mucilage and, one can speculate, have a similar beneficial role on root water uptake, indicating the important link between plants and soil microbiota for soil water resources.

Root strategies to improve the acquisition of water from the soil can be summarized in two groups: (1) root development and exploration of larger volumes of soil; and (2) improvement of the hydraulic conductivity of the rhizosphere via root hairs, mycorrhiza and the production of biofilms. Both strategies require a carbon input: the first strategy requires carbon input for plant growth; the second strategy consists in the investment of carbon into the rhizosphere. The two strategies are not alternatives and are likely to be coordinated in space and time. Hairs and mucilage are functional for some time, then hairs shrink (Duddek et al., 2022) and mucilage is degraded. Therefore, their production must match with the root development in such a way that these traits are active in the root segments that are hydraulically active (radially and axially).

In summary, we have shown the physics of soil water flow to the root and how, during soil drying, the local drop in conductivity around the roots becomes a limiting factor for transpiration. Plants have multiple mechanisms to cope with soil hydraulic limitation. Prominent examples are root hairs, mycorrhiza and mucilage exudation. Understanding plant and soil interactions in the rhizosphere is a key challenge in soil and plant research and it has important implications for the ability of plants to cope with drought. Finally, it should be noted that to fully comprehend the implications of rhizosphere processes, one should include these processes into a root system perspective, where roots dynamically grow in a variable (in space and time) soil.

References

- Aravena JE, Berli M, Ghezzehei TA, and Tyler SW (2011) Effects of root-induced compaction on rhizosphere hydraulic properties - X-ray microtomography imaging and numerical simulations. *Environmental Science and Technology* 45: 425–431.
- Cai G, Carminati A, Abdalla M, and Ahmed MA (2021) Soil textures rather than root hairs dominate water uptake and soil-plant hydraulics under drought. *Plant Physiology* 187: 858–872.
- Carminati A, Moradi AB, Vetterlein D, Vontobel P, Lehmann E, Weller U, Vogel HJ, and Oswald SE (2010) Dynamics of soil water content in the rhizosphere. *Plant and Soil* 332: 163–176.
- Carminati A, Benard P, Ahmed MA, and Zarebanadkouki M (2017a) Liquid bridges at the root-soil interface. *Plant and Soil* 417.
- Carminati A, Passioura JB, Zarebanadkouki M, Ahmed MA, Ryan PR, Watt M, and Delhaize E (2017b) Root hairs enable high transpiration rates in drying soils. *New Phytologist* 216(3): 771–781.
- Couvreux V, Faget M, Lobet G, Javaux M, Chaumont F, and Draye X (2018) Going with the flow: Multiscale insights into the composite nature of water transport in roots. *Plant Physiology* 178: 1689–1703.
- Dodd IC and Diatloff E (2016) Enhanced root growth of the brb (bald root barley) mutant in drying soil allows similar shoot physiological responses to soil water deficit as wild-type plants. *Functional Plant Biology* 43: 199–206.

- Draye X, Kim Y, Lobet G, and Javaux M (2010) Model-assisted integration of physiological and environmental constraints affecting the dynamic and spatial patterns of root water uptake from soils. *Journal of Experimental Botany* 61: 2145–2155.
- Duddek P, Carminati A, Koebernick N, Ohmann L, Lovric G, Delzon S, Rodriguez-Dominguez CM, King A, and Ahmed MA (2022) The impact of drought-induced root and root hair shrinkage on root–soil contact. *Plant Physiology* 189(3): 1232–1236.
- Faiz SMA and Weatherley PE (1982) Root contraction in transpiring plants. *New Phytologist* 92: 333–343.
- Hallett PD, Gordon DC, and Bengough AG (2003) Plant influence on rhizosphere hydraulic properties: Direct measurements using a miniaturized infiltrometer. *New Phytologist* 157: 597–603.
- Harris D, Pacovsky RS, and Paul EA (1985) Carbon economy of soybean-Rhizobium-Glomus associations. *New Phytologist* 101: 427–440.
- Kabir Z and Koide RT (2002) Effect of autumn and winter mycorrhizal cover crops on soil properties, nutrient uptake and yield of sweet corn in Pennsylvania, USA. *Plant and Soil* 238: 205–215.
- Kikuchi Y, Hijikata N, Ohtomo R, Handa Y, Kawaguchi M, Saito K, Masuta C, and Ezawa T (2016) Aquaporin-mediated long-distance polyphosphate translocation directed towards the host in arbuscular mycorrhizal symbiosis: Application of virus-induced gene silencing. *The New Phytologist* 211: 1202–1208.
- Koebernick N, Daly KR, Keyes SD, George TS, Brown LK, Raffan A, Cooper LJ, Naveed M, Bengough AG, Sinclair I, et al. (2017) High-resolution synchrotron imaging shows that root hairs influence rhizosphere soil structure formation. *New Phytologist* 216: 124–135.
- Marin M, Feeney DS, Brown LK, Naveed M, Ruiz S, Koebernick N, Bengough AG, Hallett PD, Roose T, Puértolas J, et al. (2021) Significance of root hairs for plant performance under contrasting field conditions and water deficit. *Annals of Botany* 128: 1–16.
- McCully ME and Boyer JS (1997) The expansion of maize root-cap mucilage during hydration. 3. Changes in water potential and water content. *Physiologia Plantarum* 99: 169–177.
- Miller RM, Jastrow JD, and Reinhardt DR (1995) External hyphal production of vesicular-arbuscular mycorrhizal fungi in pasture and tallgrass prairie communities. *Oecologia* 103: 17–23.
- Moradi AB, Carminati A, Lamparter A, Woche SK, Bachmann J, Vetterlein D, Vogel H-J, and Oswald SE (2012) Is the rhizosphere temporarily water repellent? *Vadose Zone Journal* 11: vzj2011.0120.
- Read DB and Gregory PJ (1997) Surface tension and viscosity of axenic maize and lupin root mucilages. *New Phytologist* 137: 623–628.
- Safir GR, Boyer JS, and Gerdemann JW (1971) Mycorrhizal enhancement of water transport in soybean. *Science (New York, N.Y.)* 172: 581–583.
- Scheltema MA, Abbott LK, Robson AD, and Death G (1985) The spread of glomus fasciculatum through roots of trifolium subterraneum and lolium rigidum. *New Phytologist* 100: 105–114.
- Segal E, Kushnir T, Muallem Y, and Shani U (2008) Water uptake and hydraulics of the root hair rhizosphere. *Vadose Zone Journal* 7: 1027–1034.
- Smith SE and Read D (2008) Mineral nutrition, toxic element accumulation and water relations of arbuscular mycorrhizal plants. *Mycorrhizal Symbiosis*. 145–VI.
- Tanaka N, Kato M, Tomioka R, Kurata R, Fukao Y, Aoyama T, and Maeshima M (2014) Characteristics of a root hair-less line of Arabidopsis thaliana under physiological stresses. *Journal of Experimental Botany* 65: 1497–1512.
- van Lier Q, de Metselaar JK, and van Dam JC (2006) Root water extraction and limiting soil hydraulic conditions estimated by numerical simulation. *Vadose Zone Journal* 5: 1264–1277.
- Zarebanadkouki M, Ahmed MA, and Carminati A (2016) Hydraulic conductivity of the root-soil interface of lupin in sandy soil after drying and rewetting. *Plant and Soil* 398: 267–280.
- Zhu XC, Bin SF, Liu SQ, and Liu TD (2011) Effects of arbuscular mycorrhizal fungus on photosynthesis and water status of maize under high temperature stress. *Plant and Soil* 346: 189–199.

Plant-soil-water relations

Jan Vanderborght^a, Quirijn de Jong van Lier^b, Andrea Schnepf^a, and Mathieu Javaux^c, ^aAgrosphere Institute, IBG-3, Forschungszentrum Jülich GmbH, Jülich, Germany; ^bCenter for Nuclear Energy in Agriculture, University of São Paulo, Piracicaba, Brazil; ^cEarth and Life Institute, Agronomy, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

© 2023 Elsevier Ltd. All rights reserved.

This is an update of RA Feddes, JC van Dam, Plant-soil-water relations, Editor(s): D Hillel, *Encyclopedia of Soils in the Environment*, Elsevier, 2005, Pages 222–230, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00520-8>.

Introduction	552
Mechanisms and characteristics of water flow in the soil-plant system	553
Description of uptake using models	558
Uptake based on root hydraulics (non-saline soils)	558
Uptake including soil hydraulics	561
Upscaling and reducing model complexity	564
Uptake in saline soils	567
Conclusions and outlook	569
References	569

Abstract

Water flows in the soil-plant-atmosphere system from higher to lower water potentials. The rate that water flows depends on potential differences and conductances along the different parts of the water flow pathways from soil to root, shoot, leaves, and atmosphere. Soil-plant-atmosphere models simulate water uptake from the soil by plant roots based on this principle, but differ in their representation of flow paths in the system which ranges from detailed representations of the 3D root architecture to a single flow path from a lumped root zone. By upscaling, detailed information about soil and root properties can be implemented in lumped models.

Key points

- Models of root water uptake, depending on both plant and soil properties are reviewed.
- With root age, root hydraulics change, with young tips taking up water at the most rapid rate.
- Roots redistribute water based on water potential gradients in the soil and plant tissue.
- Understanding root water uptake in detail aids the selection of ideotypes in crop breeding.

Introduction

Evaporation from the land surface is an important component of the terrestrial water cycle and on a global scale, it equates to about 60–70% of the total amount of continental precipitation. There are two main fluxes of evaporation: a flux of water that evaporates from the surface of mesophyll cells within plant leaves and exits the plant through the stomata (little openings in the leaf), often called the transpiration flux (T), and a flux of water that evaporates from surfaces (E), such as the soil surface, leaf surfaces that have intercepted rainwater, and open water surfaces of rivers and lakes. The sum of both fluxes is evapotranspiration (ET). Since the evaporation of water consumes considerable energy, evapotranspiration is a key component of the land surface energy balance. When the amount of water that is transported away from the surface can be supplied by new water that flows to the evaporating surfaces (e.g., the internal cavities below the open stomata, or the soil surface) the air at the evaporating surface remains saturated with water and the evapotranspiration rate can be derived from solving a surface energy balance that is coupled with a description of vapor and sensible heat transfers in the atmospheric boundary layer. Under these conditions, evapotranspiration is ‘energy limited’ and occurs at a ‘potential’ rate. This potential evapotranspiration, ET_{pot} , corresponds to the evaporation from a ‘well-watered’ surface and it ranges from 1 mm d⁻¹ for cool and humid weather conditions up to 10 mm d⁻¹ for hot and dry conditions (1 mm of water evaporated corresponds to 1 L of water that evaporates from a 1 m² surface). A simple approach to splitting up the potential evapotranspiration of a crop, ET_{pot} , into evaporation from the soil surface, E_{pot} , and transpiration, T_{pot} , uses Beer’s Law and splits up total energy that is absorbed by the surface into a part that is absorbed by the canopy and is used for transpiration, and a part that is used for soil water evaporation (Ritchie, 1972):

$$T_{pot} = ET_{pot}(1 - \exp(-\kappa LAI)) \quad (1)$$

where LAI is the leaf area index or the surface area of the leaves that covers a unit of soil surface [$\text{m}^2 \text{m}^{-2}$], and κ $[-]$ is an extinction coefficient that ranges between 0.4 and 0.75 for different crops and vegetation types.

The supply of water for evapotranspiration is provided by water flowing through the soil-plant system towards the evaporation surfaces. The water stored in vegetation ranges roughly from 5 to 10 kg m^{-2} for crops and grassland, to 30–40 kg m^{-2} for forests (roughly calculated from dry matter estimates (Whittaker and Likens, 1973) and a plant moisture content of 500–1000% (kg of water kg^{-1} dry weight) for crops and grasses and 100% for living wood) which is not much larger than the typical daily transpiration of 5 kg m^{-2} . In consequence, the transpiration rate is almost instantaneously translated into water uptake from the soil by roots. Water stored in trees can be depleted and used for transpiration, which leads to a phase shift between diurnal transpiration and root water uptake cycles. The water stored in trees, however, is insufficient to supply the transpiration demand over several days. When the water supply is insufficient to supply the evapotranspiration demand, the evapotranspiration becomes supply limited and is reduced compared with ET_{pot} . Since the extraction of water from the soil profile by evaporation and transpiration occurs over different ranges of depths, the dynamics of these processes in response to a drought spell differ considerably. Evaporation from the soil surface can dry out only a few cm of the soil profile, 0.1–0.15 m according to Allen et al. (1998), so that the evaporation becomes supply-limited relatively rapidly. Boesten and Stroosnijder (1986) found that in temperate humid climates, evaporation can become supply limited in as little as 2 days after the last rainfall. Plants can extract water over the extent of their whole root system and dry out the soil much deeper than evaporation from the soil surface. Maximal root depths vary strongly depending on the vegetation type, climate, soil texture, and groundwater table depth. A compilation by Fan et al. (2017) of maximal rooting depths reported in several studies showed a large variation ranging from 1 m to tens of meters, with a maximal observed root depth of 68 m of Shepherd's tree (*Boscia albitrunca*) in the Kalahari desert. Teuling et al. (2006) observed an e-folding decay rate of several weeks for transpiration (i.e., the period over which the transpiration shrinks by 63%) by deep rooted vegetation. This illustrates that plants may tap a large stock of deep soil water that helps them keep up transpiration during long dry spells.

Roots are not uniformly distributed in soil profiles and root densities decline strongly with depth. On average, 50% of the root mass is found in the upper 0.3 m of the soil profile whereas 95% of the root mass can be found in the upper 2 m (Schenk and Jackson, 2002), but there are large variations between vegetation, climate, and soil. Estimating how much water can be extracted by the root system depends on how water and roots are distributed in the soil profile.

In this chapter, we link the detailed description of water uptake by single root segments and how it is influenced by the soil properties in the rhizosphere, the small soil volume surrounded and influenced by roots, to the description of the plant available water in the entire root zone. We describe which soil and root properties define the relation between soil water content and its distribution in the soil profile and plant water uptake. This relationship is important to predict the impact on vegetation stress, growth, and carbon assimilation of prolonged drought spells, which are forecast by climate models for several regions of the world. It also plays a crucial role in irrigation scheduling to control the crop's root zone soil moisture.

Mechanisms and characteristics of water flow in the soil-plant system

Water is absorbed in the liquid state by the roots from the soil. The general flow pattern is shown in Fig. 1: soil $\rightarrow$ root cells $\rightarrow$ endodermis $\rightarrow$ xylem $\rightarrow$ leaf cells $\rightarrow$ stomata $\rightarrow$ air (the latter as vapor flow). Most of the transport ($>80\%$) through the root and leaf cells takes place along the microcapillaries situated in the cell wall. However, the roots contain endodermis cells in which the continuous system of microcapillaries in the cell walls is interrupted by the so-called Casparian strips. Consequently, most of the water moving through a plant has to pass through the cells itself. This means that the permeability of the endodermis cell and the

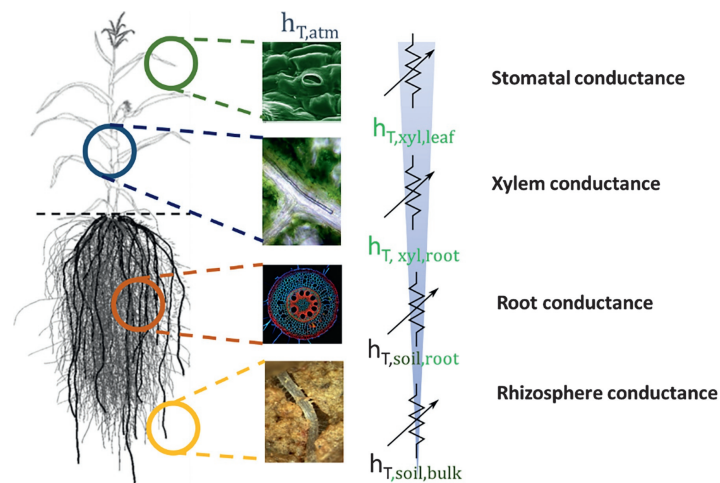


Fig. 1 Water flow path in the soil-plant-atmosphere continuum (SPAC) and water potentials and conductances in different parts of the continuum.

semipermeable membrane (plasmalemma) can be an important determinant in controlling the water flow. Hence cold soil temperature or poor soil aeration will induce a low level of metabolic activity in the endodermis cells, which may cause a large resistance across the semipermeable membrane against water uptake.

Water in this system flows from locations with a high total water potential to locations with a lower potential (generally expressed as negative values, being under suction). The total water potential is defined as the energy required to bring a unit amount of water from a reference state, normally defined as free, pure liquid water at atmospheric pressure and at a reference height, to a certain state. In soil hydrology, the unit amount of water is defined in terms of a mass of water multiplied by the gravitational acceleration, with the units of the water potential corresponding with length units. The total water potential expressed in length units, h_T , represents the energy required to move water along an equivalent water column with height h_T . In plant sciences, the unit amount of water is defined as a unit volume of water and the units of the total water potential are pressure units, ψ_T . h_T and ψ_T are related as:

$$h_T = \frac{\psi_T}{\rho_w g} \quad (2)$$

where ρ_w is the mass density of water and g the gravitational acceleration rate. For $g \approx 10 \text{ m s}^{-2}$, $h_T [\text{m}] \approx 10^2 \psi_T [\text{MPa}]$. We will use length units to express total water potentials. Total water potentials can be split into partial potentials and for soil-plant systems, the following partial potentials are of relevance:

$$h_T = h_g(z) + h_o(c) + h_p \quad (3)$$

where $h_g(z)$ is the gravitational potential, $h_o(c)$, the osmotic potential, and h_p the pressure potential. $h_g(z)$ represents the energy that is required to move water against the gravitational force and simply corresponds with the elevation, z , above the reference level. $h_o(c)$ is the osmotic potential and represents the energy that is required to move pure water into a solution across a membrane that is permeable for water molecules but not for ions (e.g., plant cell membranes). The osmotic potential h_o is related to the salt concentration in the soil pores or in the plant cells, c , and is negative since energy is released when pure water is mixed with a solution:

$$h_o = - \frac{RT \sum_i c_i}{\rho_w g} \quad (4)$$

where R is the universal gas constant ($8.314 \text{ J K}^{-1} \text{ mol}^{-1}$), T the absolute temperature [K], c_i the concentration of ion i in the soil solution [mol m^{-3}]. h_p is the pressure potential and corresponds with the difference in the water pressure and the reference water pressure, i.e., the atmospheric pressure. h_p in unsaturated soils is typically negative, i.e., lower than the atmospheric pressure, since attractive forces of pore walls generate curved air-water interfaces across which a pressure drop results from the air-water surface tension, and it decreases with decreasing soil moisture content. In living plant cells that are not supported by strong cell walls, e.g., leaf cells, h_p must be positive to maintain cell turgor. This implies that osmotic potential in these cells must be negative to sustain water flow from the soil to the leaves, or to avoid leaves losing water to the soil. At very negative total water potentials, the osmotic potential cannot be kept sufficiently low, so that the turgor of the cells is lost and plants wilt. To avoid leaf water potentials dropping to too low values when the soil dries out or when the transpiration stream is large, plants reduce transpiration by stomatal closure which results from turgor loss in stomatal guard cells. During the night, when transpiration is close to zero, the plant water potentials equilibrate again with the soil water potentials and plants regain turgor. Examples of potentials and partial potentials in different parts of the soil-plant system during daytime when plants transpire are given in Table 1. In xylem vessels, which are made up of dead cells without cell membranes, very negative water pressure potentials far below -10 m may occur so that water is under tension and in a metastable state. The strong molecular cohesion forces of water in combination with the anatomical properties of the xylem can sustain this metastable state of water. Strong cell walls and small holes or pits in the order of tens of nanometers connecting vessels, avoid the collapse of xylem vessels and avoid the formation of air bubbles or embolisms that can spread through the vessels and block the water flow. The cohesion-tension theory, introduced already at the end of the 19th century (Böhm, 1893; Dixon and Joly, 1895), states that water is pulled by negative pressures in the plant from the soil into the plant. The existence of such low and negative pressures or tensions of water in the xylem is debated

Table 1 Illustration of total water potentials (h_T) and partial potentials (all in m) in a tree-soil system: gravitational potential (h_g ; reference level is the soil surface), osmotic potentials (h_o), and pressure potentials (h_p) in the soil, root xylem, stem xylem, and leaves for wet and dry soil conditions.

Compartment	Wet soil				Dry soil			
	h_T	h_g	h_o	h_p	h_T	h_g	h_o	h_p
Soil	-1.5	-0.5	0	-1	-150	-0.5	0	-149.5
Root xylem	-30.5	-0.5	-10	-20	-170	-0.5	-10	-159.5
Stem xylem	-70	10	-10	-70	-250	10	-10	-250
Leaves	-100	10	-160	50	-300	10	-310	0

Partially after Rodríguez-Iturbe and Porporato (2004).

and other mechanisms that lead to low water potentials, e.g., attractive forces to cell walls and network structures of gels generating a matric potential, have been proposed.

For water fluxes in the soil plant system, the total water potentials are of importance and in general, the following relation between T and the total water potentials in the bulk soil of the root zone, $h_{T,soil}$ and in the plant, $h_{T,plant}$ can be defined:

$$T = K_{rs} (h_{T,soil} - h_{T,plant}) \quad (5)$$

where K_{rs} is the soil-plant conductance that represents the lumped conductance of the flow path from the soil to the leaves. Its units depend on whether the transpiration is determined as the volume of water taken up per plant and per time [$\text{m}^3 \text{d}^{-1}$] or as the volume of water taken up by the vegetation per unit of soil surface area per time [d^{-1}]. Based on observations that even in wet soils (i.e., $h_{T,soil} \approx 0$) crops may wilt during midday (i.e., $h_{T,leaf} \approx -150 \text{ m}$) on days when the transpiration rate is large (for daily averaged transpiration rate between 8 and 9 mm d^{-1} , the transpiration rate at midday is around 30 mm d^{-1}), a rough estimate of the plant part of K_{rs} yields $K_{rs} = 2 \times 10^{-4} \text{ d}^{-1}$ (Cowan, 1965). Of course, this value depends strongly on the crop or vegetation type and only gives an order of magnitude. Simplifying this flow path to a single path that crosses different compartments, as illustrated in Fig. 1, K_{rs} may be split up using Ohm's Law in partial conductances as:

$$K_{rs} = \left(K_{soil-root}^{-1} + K_{cortex}^{-1} + K_{endodermis}^{-1} + K_{xylem}^{-1} + K_{leafcells}^{-1} \right)^{-1} \quad (6)$$

These conductances are extensive properties and their values include the length of the flow path through the compartment and the cross-sectional area of the flow path. Multiplying these conductances by the path length and dividing them by the cross-sectional area, conductivities, which are intrinsic properties, are obtained. The xylem conductivity is in general eight orders of magnitude larger than the root radial conductivity (lumped cortex and endodermis conductivity). But, as the xylem constitutes all but 1 mm of the total flow path length in the plant, that means 99.9% of the flow path length in a 1 m high maize plant and 99.99% of the flow path length in a 10 m high tree. Considering that the cross-sectional area of flow through the xylem is much smaller than the absorbance surface of roots, the xylem conductance and non-xylem conductance (lumped cortex, endodermis, and leaf cell conductances) are of similar magnitude in plants (Sperry et al., 2003). The soil and xylem conductances decrease when the water pressures become more negative. The pressure potential at which the xylem conductivity drops to 50% of the fully saturated conductivity is generally lower than -100 m and lower than the pressure potential at which stomata start closing (Carminati and Javaux, 2020). This suggests that plants try to avoid failure of the xylem by closing stomata and controlling the plant water potential. It also implies that the plant conductance may be considered to be independent of the plant water potential when the plant is not extremely stressed.

For wet soil conditions, the hydraulic conductivity of the soil is considerably higher than that of root tissues and potential drops along the soil-root-leaf pathways occur mainly within the plant (Fig. 2) (Draye et al., 2010). For drier soil conditions, i.e., for

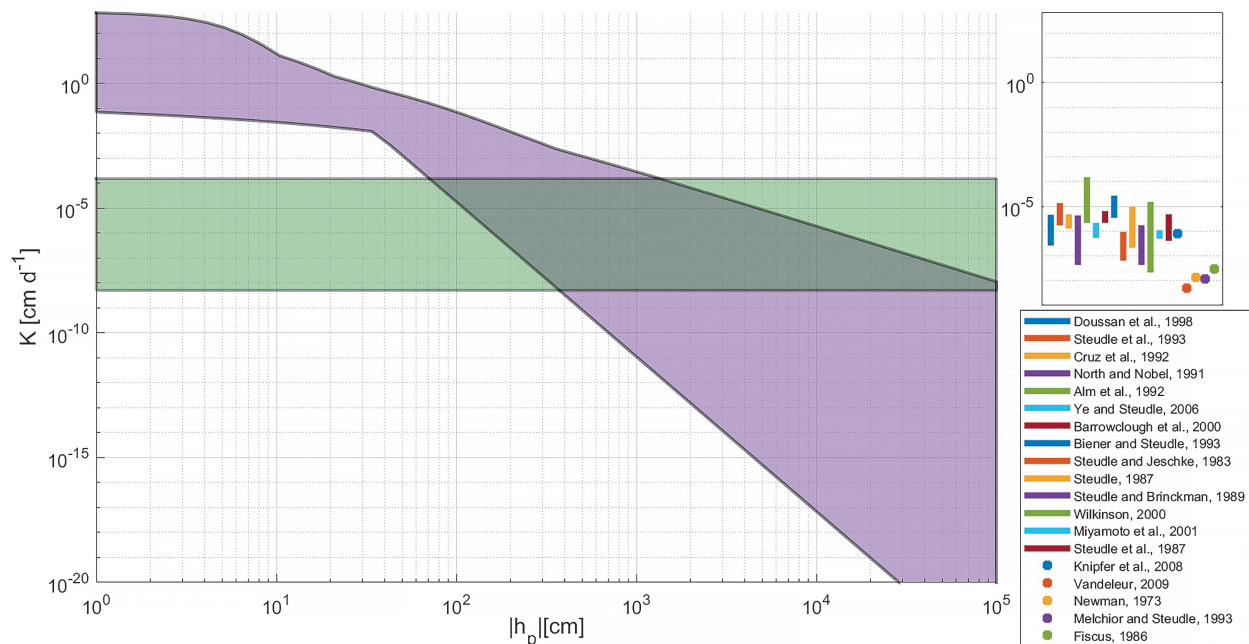


Fig. 2 Envelopes of typical soil conductivity curves (purple area) and apparent root conductivity values assuming an averaged radius of 250 μm (green area) redrawn from the literature. The upper right plot represents root conductivity values from 19 studies. Adapted from Draye X, Kim Y, Lobet G and Javaux M (2010) Model-assisted integration of physiological and environmental constraints affecting the dynamic and spatial patterns of root water uptake from soils. *Journal of Experimental Botany* 61: 2145–2155, where the original reference sources can be found.

pressure potentials smaller than -1 m in sand and -10 m in loam down to -150 m in clay, the conductivity of the soil becomes limiting and large pressure drops between the bulk soil and the soil-root interface emerge and Eq. (5) becomes nonlinear. Since the water capacity of the soil, i.e., how the soil water content changes per unit change of soil water pressure, is very low when the soil-root conductance becomes limiting, soil and plant water potentials drop rapidly when the soil dries out further. To avoid a drop in plant water potentials that disrupts the flow path in the xylem within the reaction time of stomata, stomata should close and reduce the transpiration losses at plant water potentials that are far above damaging pressures in the xylem. The critical leaf water potentials at which stomata start closing should therefore be far above the damaging pressures in the xylem and should adapt to the soil properties and the transpiration rates (Carminati and Javaux, 2020). The mechanism that closes stomata is turgor loss of the guard cells when water potentials in these cells become too low. Since guard cells are hydraulically connected to the other cells in the leaf, the leaf water potential is the direct signal for stomatal closure. However, many different signals play a role in stomatal regulation, which affect the dynamics and the magnitude of the stomatal response to a decrease in water potential (Lawson and Matthews, 2020). An additional signal that measures the change in transpiration rate per unit change in leaf water potential and to which stomata respond has been proposed (Carminati and Javaux, 2020). This would allow plants to sense non-linearities in the soil-plant flow path and control transpiration and plant pressure losses by timely closing of stomata. Stomatal regulation strategy is sometimes split between isohydric and anisohydric behavior. While plants with isohydric behavior have a strict stomatal regulation to keep leaf water potentials above a limiting threshold, plants with anisohydric behavior have a smoother stomatal regulation that generates greater variation of leaf water potentials with transpiration rate. The simplest way to represent the effect of drought on transpiration and root water uptake for an isohydric plant is to assume that when a critical leaf water potential is reached, stomata (partially) close and reduce the transpiration rate so that the leaf water potentials do not drop below this critical leaf water potential. The critical leaf water potentials that plants try to keep constant by reducing the transpiration depend on the plant species, but as discussed above, may also depend on the soil type and transpiration demand.

Eqs. (5) and (6) are useful to obtain a qualitative overview of the flow process in the soil-plant system and the water potential and of pressure drops along the flow path in this system. However, for a quantitative prediction, more detailed descriptions are required. The main reason is the spatial variation of soil water potentials, soil moisture contents, and root distributions in the root zone. The largest variations are with depth, but lateral variations may also be significant. The interplay between non-uniform water potentials and non-uniform root distributions in the root zone in combination with non-linear conductances in the system (i.e., conductances that depend on water potentials), needs to be understood to determine an average soil water potential in the root zone and the soil-plant conductance. This is needed to calculate the plant water potential and transpiration. How these root zone averaged potentials, conductances, and water fluxes change with time depends on the distribution of the water uptake from the soil profile and how water redistribution in the soil-plant system can compensate for spatial variations in root water uptake. Vertical and lateral variations in soil water potential result in, but are also caused by, non-uniform water uptake due to non-uniform root distributions in the root zone. Starting from a uniformly wet soil profile or a soil profile with a uniform soil water potential, water uptake occurs mainly in the upper soil layers where the root density is largest (largest absorbing root surface) and from where the flow path distance through the xylem to the transpiring leaf surfaces is the smallest. This results in faster drying by roots of the upper than the deeper soil layers and a stronger decrease in soil water potentials in the upper than in the deeper layers. To maintain the transpiration rate, the plant water potential will decrease and the difference between soil and plant water potential will become larger in the deeper soil layers than in the upper layer. As a consequence, the uptake from the drier upper soil layers will decrease and be compensated by increased uptake from deeper layers. Fig. 3 shows an example of this phenomenon that was observed in a lysimeter during the growing season of a spring barley crop. With time, the upper soil layers dry out and the root system progressively starts depleting deeper soil layers. This “moving sink” phenomenon is called root water uptake compensation.

During the night, when transpiration from leaves is close to zero, the water potential in the plant tends to equilibrate with the average soil water potential. The difference in water potential between the upper and the deeper soil then induces a water flow from the deeper soil layers to the upper soil via the root system that bypasses the water flow through the soil pores. This phenomenon of water uptake by deeper roots and root exudation in shallower layers is called hydraulic lift. Roots of deep rooting trees provide direct connections and transfer water over tens of meters between deeper soil layers, where they may tap from groundwater, and surface soil layers. The magnitude of these water transfers ranges from 0.04 to 1.3 mm d^{-1} in the empirical literature, and from 0.1 to 3.23 mm d^{-1} in the modeling literature (Neumann and Cardon, 2012). Hydraulic lift continues the uptake process during the night and provides a water stock that is easily accessible during the day. It may also support shallow rooting vegetation with water and keep the upper soil layers moist so that microbial activity, which mineralizes organic material and releases nutrients for plants; and nutrient mobility in the upper layers are maintained. The redistribution of water in the soil profile via roots may also occur in the downward direction when the upper soil layers are rewetted by precipitation and the lower soil layers are dry. This downward transfer via roots avoids water loss by evaporation from the soil surface so it can be used later by the vegetation. Hydraulic redistribution via the root system and root water uptake compensation play an important role in the water supply of seasonally dry ecosystems (Amenu and Kumar, 2008) (Fig. 4). The impact of hydraulic lift and root redistribution on water uptake depends on the interplay between the root distribution, the root, and soil hydraulic properties. Using a simulation study with root systems that have a high or a low root radial conductivity and an infinite xylem conductivity, which lead to, respectively a large and small hydraulic redistribution via the root system, it was found that hydraulic redistribution had the largest impact in root systems with a non-uniform root distribution. The effect of different root conductivities and hydraulic redistribution on the water uptake was found to be similar in different soil types.

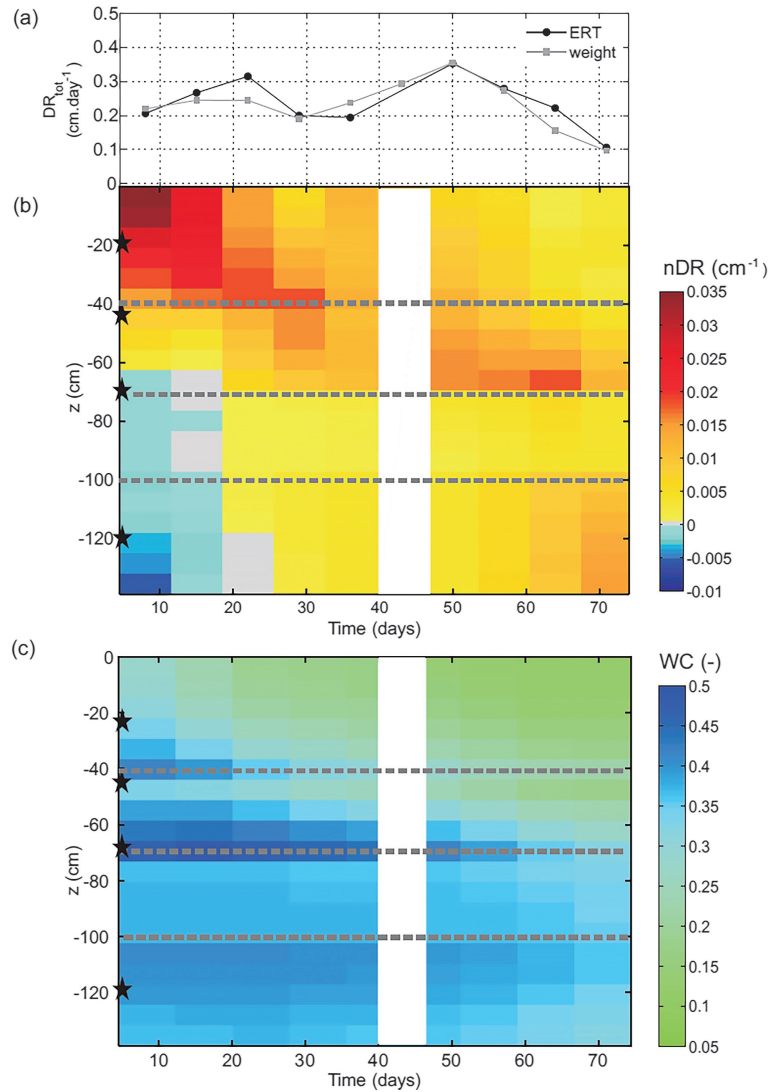


Fig. 3 (a) Total weekly water depletion rate (DR_{tot}) estimated by electrical resistivity tomography (ERT) and weight measurements during the growing season of a spring barley crop; (b) normalized local water depletion rates (nDR) as a function of depth and time; (c) volumetric water content (WC) measured with ERT as a function of depth and time. The gray dashed line indicates horizon interfaces. From Garre S, Javaux M, Vanderborght J, Pages L and Vereecken H (2011) Three-dimensional electrical resistivity tomography to monitor root zone water dynamics. *Vadose Zone Journal* **10**: 412–424.

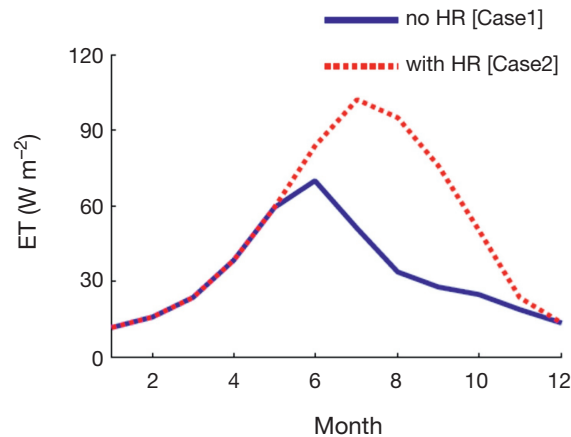


Fig. 4 Effect of hydraulic redistribution and root water uptake compensation (HR) on the transpiration in a seasonally dry ecosystem (Sierra Nevada, CA, United States) with a long dry summer and a wet winter season. Solid blue line and dotted red line are average annual transpiration cycles simulated by a model that does not (blue line) and does (red line) consider redistribution and compensation. From Amenu GG and Kumar P (2008) A model for hydraulic redistribution incorporating coupled soil-root moisture transport. *Hydrology and Earth System Sciences* **12**: 55–74.

Description of uptake using models

Uptake based on root hydraulics (non-saline soils)

The single flow path in the soil-root-plant system shown in Fig. 1 can be split up in several flow paths from the soil to the absorbing surfaces of single root segments and within the connected root segments to the root collar and further to the leaves. When neglecting the water storage in root segments, the root system can be represented by a network of leaky pipes and the flow in a single root segment is described by:

$$\frac{d}{d\ell} k_x \frac{dh_{T,xylem}}{d\ell} = -2\pi r_{root} k_r (h_{T,soil-root} - h_{T,xylem}) \quad (7)$$

where ℓ is the axial coordinate of the root, k_x [$\text{m}^3 \text{s}^{-1}$] and k_r [s^{-1}] are the intrinsic axial and radial root segment conductances, r_{root} is the root segment radius, and $h_{T,xylem}$ [m] and $h_{T,soil-root}$ [m] are the water potentials in, respectively, the xylem and at the soil-root interface, which include both the pressure potential and the elevation potential (osmotic potentials are not included here). k_r lumps the cortex and endodermis conductances. Where two, or more segments come together in a root branch, the xylem pressures at the connection point are the same for all segments and the flow towards that connection point equals the flow away from it. When $h_{T,soil-root}$ is known along all root segments, the uptake along each root segment and the distribution of $h_{T,xylem}$ in the entire root system can be calculated. The intrinsic conductances depend on the plant species, but large variations of these properties may also be observed between different plant genotypes of the same species. Furthermore, they can also be ‘plastic’ root properties that adapt to the soil and climate conditions. k_r and k_x vary over time or with distance along a root from its distal (youngest) to its proximal (oldest) ends. With the maturation of the xylem, k_x increases to a maximal value whereas the development of Casparian strips leads to a reduction of k_r with root age (Fig. 5a and b).

Considering a single root with constant k_x and k_r , the uptake along the root decreases with distance from the proximal end of the root since the pressure in the xylem increases. An analytical solution can be derived that expresses the uptake and the xylem pressures along the root as a function of the following dimensionless parameter α (Landsberg and Fowkes, 1978):

$$\alpha = l_{root} \sqrt{\frac{2\pi r k_r}{k_x}} \quad (8)$$

For larger α values, the uptake declines more from the proximal to the distal end of the root, whereas it is more homogeneous for small α values. The ratio of radial to axial root conductance and α therefore define a maximal root length along which water can be taken up effectively from a soil profile with a homogeneous soil water potential. Considering ranges of observed conductances, these analytical solutions indicate that root water uptake by a root segment depends on the distance of the segment to the root collar and that xylem conductances can be limiting for root water absorbance. The water uptake along a single root is also strongly affected by the distribution of conductances along the root. The lower k_r near the proximal (older) and higher k_r near the distal end of the root and the opposite for k_x lead to more uptake near the root tips. This is shown for a lateral root of maize in Fig. 5. A growing root with a non-uniform distribution of k_r and k_x leading to a higher uptake near root tips can better explore and access water deeper in the soil profile than a root with constant k_x and k_r values, which will focus uptake near the proximal end.

For real root architectures and for non-homogenous soil water potential distributions, the flow into and within the root network can be calculated numerically when the soil water potentials at the soil root interfaces, $h_{T,soil-root}$ are known (Fig. 6). To simulate soil

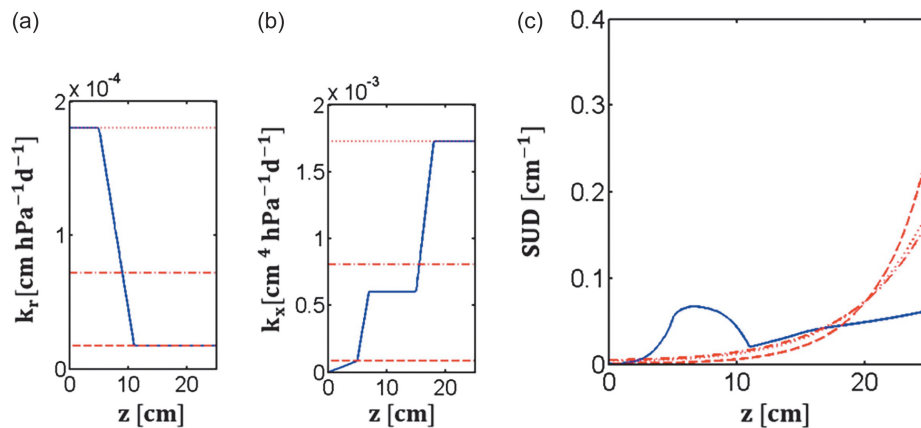


Fig. 5 Distribution along the length of a lateral maize root, z (with $z = 0$ at the root tip), of: radial root conductivity k_r (a), axial root conductivity k_x (b), and ‘standard uptake density’ (SUD, fraction of uptake per unit root length for a uniform soil water potential). Blue lines represent a root with a non-uniform distribution of k_r and k_x and red lines represent roots with uniform k_x and k_r . 1 hPa is equivalent to a pressure head of approximately 1 cm. From Meunier F, Couvreur V, Draye X, Zarebanadkouki M, Vanderborght J and Javaux M (2017) Water movement through plant roots—Exact solutions of the water flow equation in roots with linear or exponential piecewise hydraulic properties. *Hydrology and Earth System Sciences* 21: 6519–6540.

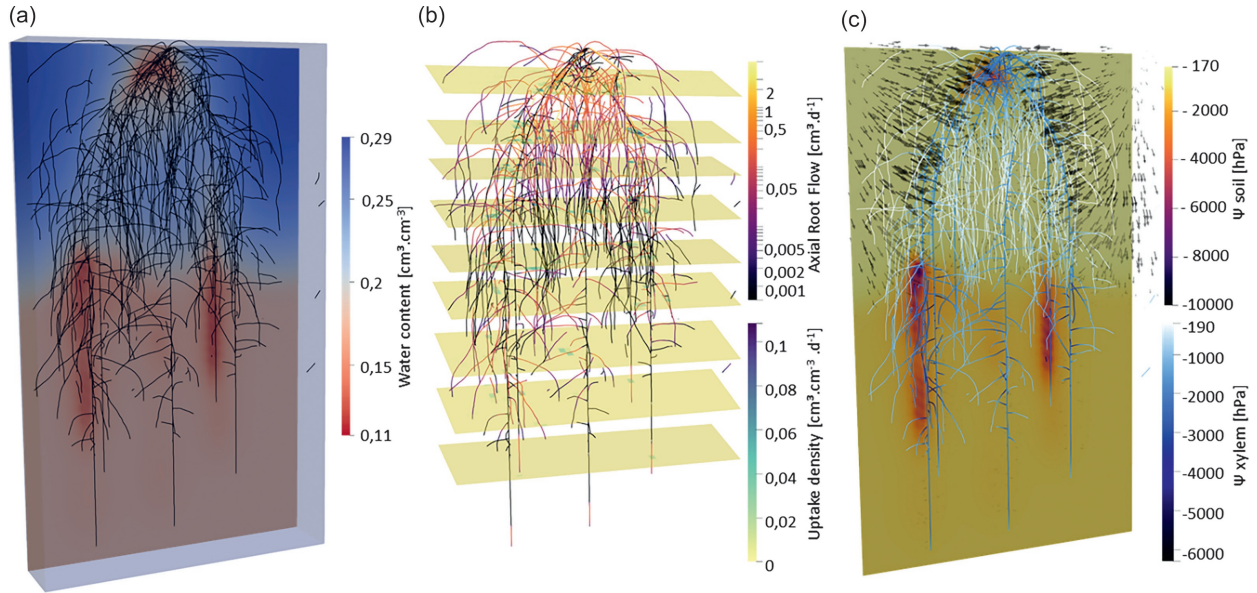


Fig. 6 Water fluxes in a 3D root system and in the soil for a soil profile with a wet top layer over a dry layer. (a): Soil water content; (b): uptake density (sink term) in horizontal planes and axial root flow; (c): xylem water potential, soil water potential and flux in the soil (black arrows).

water potential changes at the soil root interface that result from water uptake and water flow in the soil profile, the flow in the root system must be coupled to the flow equation in the soil (Javaux et al., 2008):

$$\frac{\partial \theta}{\partial t} = \nabla \cdot \mathbf{K} \nabla h_{T,soil} - S(\mathbf{x}) \quad (9)$$

where θ is the volumetric water content, $\mathbf{K}$ is the soil hydraulic conductivity tensor [m s^{-1}] and $S(\mathbf{x})$ is the sink term [$\text{m}^3 \text{m}^{-3} \text{s}^{-1}$] that represents the uptake of water per bulk volume of soil by all root segments in an infinitesimal soil volume $d\Omega$ around $\mathbf{x}$ (positive for uptake, negative for water efflux). When the soil water potentials do not vary horizontally and soil water flow is only in the vertical direction, the flow equation can be written in a one-dimensional form as:

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z} \left[K \left[\frac{\partial h_{p,soil}}{\partial z} + 1 \right] \right] - S(z) \quad (10)$$

where z is the vertical co-ordinate, $h_{p,soil}$ is the soil water pressure potential, and $S(z)$ represents the averaged uptake in a soil layer with infinitesimal thickness dz at depth z . To solve the equations, functions that relate θ and K with $h_{p,soil}$, i.e., the soil hydraulic functions, are required. The averaged sink term distributions show a similar sensitivity to the ratio of the radial to axial root segment conductance as the single root uptake. When soil water potentials are uniform and when the ratio of radial to axial conductivity is sufficiently small, the root uptake profile corresponds with the root length density distribution. For a large axial conductance, and consequently a large root system conductance, small decreases in soil water potential near the soil surface where uptake is initially larger due to the greater root density are compensated by a larger uptake deeper in the soil profile so that the root water uptake profile becomes more uniform with depth (Fig. 7a). For smaller axial conductances, the water uptake is initially concentrated near the soil surface (Fig. 7c) and then shifts downwards when the upper soil layer dries out. This shows that root hydraulic properties, root architecture, and the soil water potential profile determine the root water uptake distribution with depth and that this distribution differs from the root density profile.

The uptake integrated over all layers in the root zone corresponds with the plant transpiration rate (assuming that there is no water storage capacity in the plant):

$$T_{act} = \int_{L_{root}}^{z_{surf}} S(z) dz \quad (11)$$

where L_{root} is the depth of the bottom of the rootzone and z_{surf} is the elevation of the soil surface (m). When the soil is sufficiently wet and the soil hydraulic conductivity is considerably larger than the root radial conductivity, the water potential at the soil-root interfaces can be approximated by the bulk soil water potential and $S(z)$ can be described as (Couvreur et al., 2012):

$$S(z) = K_{rs}SUF(z)[h_{T,soil,eff} - h_{T,plant}] + K_{comp}SUF(z)[h_{T,soil}(z) - h_{T,soil,eff}] \quad (12)$$

where SUF is the 'standard uptake fraction' [m^{-1}] and $h_{T,soil,eff}$ is the effective soil water potential in the root zone, and K_{comp} a compensatory uptake conductance [s^{-1}]. $SUF(z)dz$ represents the fraction of the total root water uptake that is taken from a soil layer

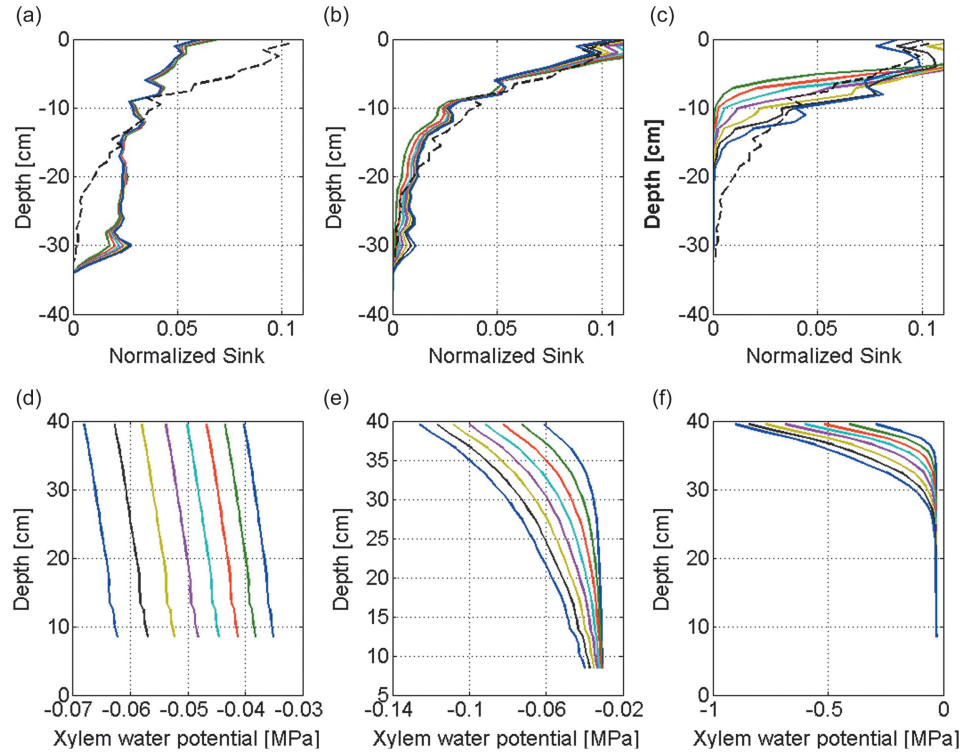


Fig. 7 Vertical profiles of the horizontally averaged sink term (top row a–c) and xylem water potentials during a dry spell bottom row (d–f) that are simulated for a 3D root system with three different sets of root hydraulic properties: left column (a and d) high axial conductivity ($k_x = 4.32 \text{ cm}^3 \text{ d}^{-1}$), middle column (b and e) medium axial conductivity ($k_x = 4.32 \times 10^{-2} \text{ cm}^3 \text{ d}^{-1}$), and right column (c and f) low axial conductivity ($k_x = 4.32 \times 10^{-4} \text{ cm}^3 \text{ d}^{-1}$). The radial conductivity ($k_r = 1.73 \times 10^{-4} \text{ d}^{-1}$). Colors represent successive days and the dashed black line is the normalized root length density. From Draye X, Kim Y, Lobet G and Javaux M (2010) Model-assisted integration of physiological and environmental constraints affecting the dynamic and spatial patterns of root water uptake from soils. *Journal of Experimental Botany* **61**: 2145–2155.

with thickness dz at depth z when the soil water potential is uniform in the entire root zone. $h_{T\text{soil},\text{eff}}$ is a weighted average of the soil water potential in the root zone and the weights correspond with SUF :

$$h_{T\text{soil},\text{eff}} = \int_{L\text{root}}^{z\text{surf}} SUF(z) h_{T\text{soil}}(z) dz \quad (13)$$

Combining Eqs. (11), (12), and (13) and considering that the integral of SUF over the root zone equals one, $h_{T\text{soil},\text{eff}}$ can be used as the average soil water potential in the root zone in Eq. (5) to calculate the total root water uptake for a given plant water potential. Or, when the transpiration rate is known, e.g., the potential transpiration rate, the plant water potential can be calculated. It was shown by theoretical analysis of the flow equations in the root system, that the definition of $h_{T\text{soil},\text{eff}}$ as in Eq. (13) gives for all soil water potential distributions in the rootzone a correct estimation of the plant water potentials and transpiration rate, and that K_{comp} can be approximated by K_{rs} (Vanderborght et al., 2021).

The first term of Eq. (12) represents the uptake by the root system when the soil water potential is uniform in the root zone and equal to the effective soil water potential. The second term represents the uptake compensation by the root system for a heterogeneous distribution of soil water potentials. At depths where the soil is dry and $h_{T\text{soil}}(z) < h_{T\text{soil},\text{eff}}$, this term is negative and the uptake is reduced and vice versa for depths where the soil is wetter. At night when there is no uptake and the first term in Eq. (12) is zero, there is a negative uptake or an exudation of water by the root system, i.e., hydraulic root redistribution, in drier regions of the soil, and this redistribution is independent of the transpiration rate.

Because the root segment conductivities are assumed to be independent of the root water potentials and the water potential at the soil root interface is assumed to be equal to the bulk soil water potential, the relation between the transpiration rate and the difference between the soil and plant water potentials in Eq. (5) is linear. When transpiration is plotted versus $h_{T\text{soil},\text{eff}}$ and when stomata regulate the plant water potential so that it does not drop below a critical value, $h_{T\text{plant},\text{crit}}$ (i.e., perfect isohydric plant) then the maximal transpiration rate that is possible for a certain $h_{T\text{soil},\text{eff}}$ is given by a straight line with a slope K_{rs} (Fig. 8). The ratio of actual transpiration, i.e., the transpiration rate when the critical plant water potential is reached, to the potential transpiration rate is represented by a linear function with a steeper slope for lower potential transpiration rates. This implies that for a lower transpiration demand by the atmosphere, the plant reduces the transpiration rate at drier soil conditions than at higher potential transpiration rates. Interestingly, this corresponds to the stress functions proposed by Feddes et al. (1978), which describe the reduction in root water uptake as a function of soil water potential and potential transpiration.

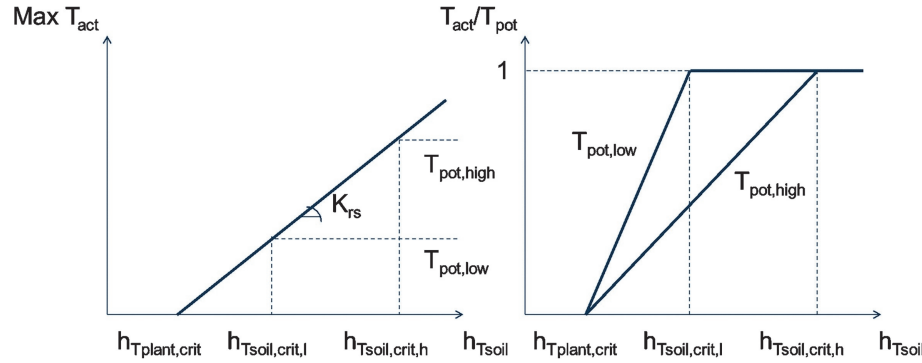


Fig. 8 (Left) Maximal transpiration for a critical water potential in the plant $h_{T,plant,crit}$ and a given root system conductance, K_{rs} , versus the soil water potential, $h_{T,soil}$. When $h_{T,soil}$ reaches $h_{T,plant,crit}$, transpiration goes to zero. $h_{T,soil,crit,l}$ and $h_{T,soil,crit,h}$ represent the minimal soil water potentials for which the maximal plant transpiration can supply a, respectively, low and high potential transpiration or transpiration demand. (Right) Ratio of actual to potential transpiration versus soil water potential for a high and a low T_{pot} .

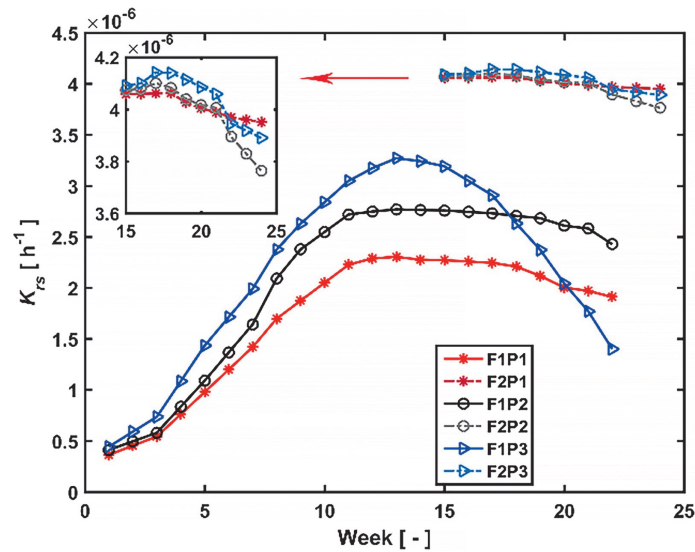


Fig. 9 Evolution of the root system conductance (K_{rs}) of a wheat crop from Feb. 11 until July, 24 in two soils (stony F1 and silty F2) and three water treatments (P1 dryout shelter, P2 rain fed and P3 irrigated). From Cai GC, Vanderborght J, Langensiepen M, Schnepf A, Huing H and Vereecken H (2018) Root growth, water uptake, and sap flow of winter wheat in response to different soil water conditions. *Hydrology and Earth System Sciences* 22: 2449–2470.

According to Eq. (12) and approximating K_{comp} by K_{rs} , K_{rs} and $SUF(z)$ are ‘sufficient’ properties of the root hydraulic system to describe the uptake from a soil profile. They are calculated directly from the root hydraulic architecture and the root segment hydraulic properties and translate the properties of the root system into its water uptake function. They are dynamic since they are related to root growth, root system development, and maturation. This growth depends on the soil properties and on the reaction of the plant to stress conditions. K_{rs} increases when the root system grows (Cai et al., 2018) (Fig. 9) and $SUF(z)$ is related to the root density in a certain layer but may differ from it, especially when the resistance to flow in the xylem is important (see Fig. 7).

Uptake including soil hydraulics

In dry soil, the low soil hydraulic conductivity leads to a drop in soil water potential along the radial water flow path from the bulk soil to the soil-root interface. Therefore, the bulk soil water potentials that are used in the soil water flow equation Eq. (9) should be scaled down to the water potentials at the soil root interface to calculate the flow into and within the root system (Eq. 7). This downscaling implies that the sink term in Eq. (9) depends not only on the root hydraulic properties but also on the soil. Downscaling can be done using ‘brute force’ by explicitly representing the 3D geometry of the roots in the 3D coupled soil-root model and treating the root system as an internal boundary in the soil domain instead of a distributed sink term. However, the computational costs to resolve the flow and water potential fields around 0.1 mm diameter roots are immense. Assuming that near the root segments, the flow in the soil is axisymmetric and perpendicular to the root and not influenced by gravity or water potential

gradients that drive the flow at a larger scale, a cylindrical rhizosphere can be constructed around the root and the flow in this rhizosphere can be described using a 1D axial symmetric flow equation:

$$\frac{\partial \theta_{rhizo}}{\partial t} = \frac{1}{r} \frac{\partial}{\partial r} \left[rK(h_{p,rhizo}) \frac{\partial h_{p,rhizo}}{\partial r} \right] + \nabla \cdot K \nabla h_{T,soil} \quad (14)$$

where θ_{rhizo} and $h_{p,rhizo}$ are the water contents and pressure heads in the rhizosphere. The term $\nabla \cdot K \nabla h_{T,soil}$ represents the change in water content in the rhizosphere per unit time due to divergence of flow at a larger scale. The inner boundary condition is defined at the root radius where the fluxes in the soil and root segments are matched, whereas the outer boundary is a no-flow boundary condition at half the average distance between root segments, which can be derived from the root length (RLD) density as:

$$r_{rhizo} = \sqrt{\frac{1}{\pi RLD}} \quad (15)$$

By averaging the rhizosphere model over the volume of the rhizosphere, the macroscale model is obtained, and the macroscale sink term is equal to the flow to the root divided by the volume to the rhizosphere:

$$S(\mathbf{x}) = \frac{2r_{root}}{r_{rhizo}^2 - r_{root}^2} K(h_{p,soil-root}) \frac{dh_p}{dr} \Big|_{r_{root}} \quad (16)$$

The rhizosphere part in Eq. (14) can either be solved numerically or analytically. Analytical solutions require some assumptions, such as the steady rate assumption, which means that the change in water content per time does not differ with the radial distance (i.e., $\frac{\partial \theta}{\partial t} = \text{constant}$). Then, the radial water flux at the soil root surface, q_r [m s^{-1}] and the sink term $S(\mathbf{x})$, can be expressed as:

$$q_r = -\frac{K_{soil-root} B}{r_{root}} (h_{p,soil} - h_{p,soil-root}) S(\mathbf{x}) = \frac{2K_{soil-root} B}{r_{rhizo}^2 - r_{root}^2} (h_{p,soil} - h_{p,soil-root}) \quad (17)$$

where $K_{soil-root}$ is an average soil conductivity [m s^{-1}] of the rhizosphere, B is a scaling factor, and $h_{p,soil}$ is the bulk soil water pressure head that corresponds with the averaged water content in the rhizosphere. $K_{soil-root}$ and B are obtained as in de Jong van Lie et al. (2008):

$$K_{soil-root} = \frac{\int_{-\infty}^{h_{p,soil}} K(h) dh - \int_{-\infty}^{h_{p,soil-root}} K(h) dh}{h_{p,soil} - h_{p,soil-root}} \quad (18)$$

$$B = \frac{2(\rho^2 - 1)}{1 - (0.53\rho)^2 + 2\rho^2(\ln \rho + \ln(0.53))}; \rho = \frac{r_{rhizo}}{r_{root}} \quad (19)$$

To calculate the sink term using Eq. (16) or Eq. (17), the soil water pressure potential at the soil-root interface, $h_{p,soil-root}$, must be known, which depends on the water flow in the root system. This implies that the flow equations in the root system and in the rhizosphere need to be coupled. By combining Eq. (17) with Eq. (7), the flow from the soil into the root xylem can be described as:

$$\frac{d}{dt} k_x \frac{dh_{T,xylem}}{dt} = -2\pi r_{root} \left(k_r^{-1} + \left(\frac{K_{soil-root} B}{r_{root}} \right)^{-1} \right)^{-1} (h_{T,soil} - h_{T,xylem}) \quad (20)$$

In contrast to Eq. (7), Eq. (20) is non-linear since $K_{soil-root}$ depends on $h_{T,soil}$ and $h_{T,xylem}$ ($h_{T,xylem}$ defines also $h_{T,soil-root}$ which is needed to calculate $K_{soil-root}$). Eq. (20) implies that the relation between plant water potentials, transpiration, root water uptake distribution, and bulk soil water potentials depends on both soil and root hydraulic properties. As an example, Fig. 10a and b shows a maize root system and the axial and radial root hydraulic properties as a function of the root segment ages. The maximal uptake that can be supplied by the root system, when the water potential in the root collar is $-16,000$ cm and the bulk soil water pressure head, $h_{p,soil}$ is uniform in the root zone, is plotted in Fig. 10c versus $h_{p,soil}$ for two different soils and for the case that the pressure at the soil-root interface is equal to $h_{p,soil}$. The slope of this curve corresponds to the soil-plant conductance, K_{rs} , defined in Eq. (5). When only the conductances in the plant are considered (which are assumed to be independent of the water potentials), this relation is linear. However, the non-linear soil hydraulic conductivity introduces a non-linearity in this relation, which is soil type dependent. For the silt loam soil, the soil plant conductance is constant with $h_{p,soil}$ and independent of the soil hydraulic conductivity for pressure heads larger than -2000 cm and decreases with decreasing $h_{p,soil}$. For the loamy sand soil, this decrease already starts at $h_{p,soil} = -100$ cm. Fig. 10d illustrates how these soil properties influence the ratio of the supply to demand, T_{act}/T_{pot} function for two T_{pot} values representing a large and a small demand.

The effect of root length density is illustrated in Fig. 10e and f. The root density was varied by considering more plants per surface area and a case with four times more plants per surface area led to a four times larger root density being simulated. A greater root density resulted in a larger K_{rs} and shifted the critical $h_{p,soil}$ at which the transpiration demand cannot be supplied by the soil-root system to more negative values.

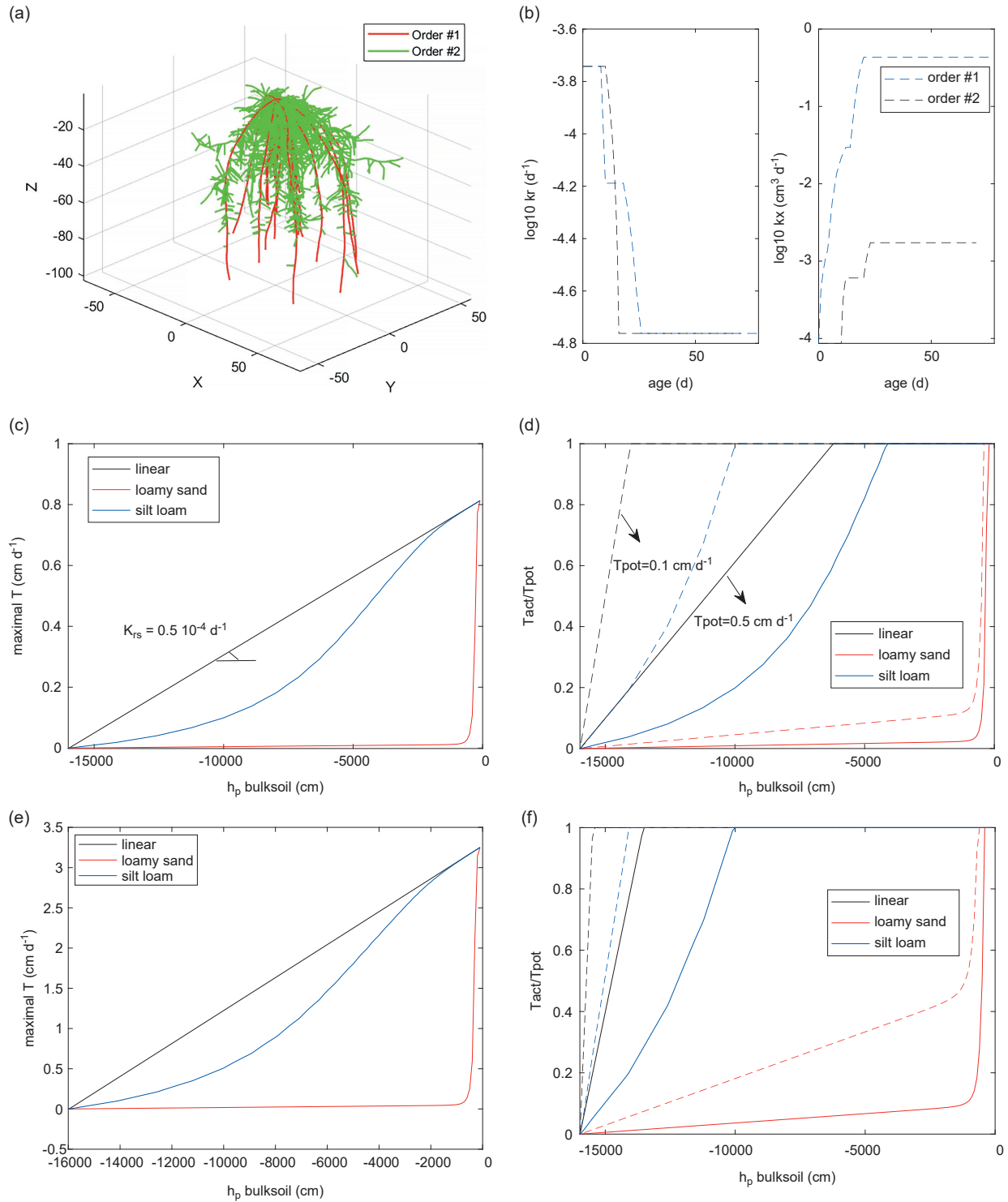


Fig. 10 Maize root system (a), and the radial, k_r , and axial conductances, k_x , of primary and secondary root segments as a function of the root segment ages; maximal supply rates or uptake rates by the soil-root system as a function of the bulk soil water pressure, $h_p \text{ bulksoil}$, for a critical plant water potential of -160 m and in three different soils (different colored lines) (a) and (e), and supply-demand ratios (T_{act}/T_{pot}) as a function of $h_p \text{ bulksoil}$ for two demand rates (dashed vs. full lines) and for a normal planting density of maize (0.75 m and 0.15 plant distance between and within rows) (c) and (d) and a four times higher plant density (e) and (f).

Upscaling and reducing model complexity

The relationships between transpiration and soil water potential in Fig. 10 illustrate the impact of the hydraulic soil properties for a homogeneous distribution $h_{p,soil}$ in the root zone on the root water uptake. When parts of the root zone become too dry and $h_{p,soil-root}$ differs from $h_{p,soil}$, the upscaled linear flow equations in the root system (Eq. 12) cannot be used to calculate the uptake and the plant water potentials directly from the non-uniform distribution of $h_{p,soil}$. Upscaling of the sink term $S(x)$ associated with a single root segment and its rhizosphere (Eqs. 16 or 17) to the sink term of a horizontal soil layer $S(z)$ requires a few simplifying assumptions. The different formulations of $S(z)$ that are used in models come from different approximations and simplifications that were used in this upscaling. An assumption that is common to all models, and which is the basis of Eq. (15), is that root segments are uniformly distributed and parallel to each other and that the soil water potentials and xylem water potentials are uniform in the soil volume over which the sink term is upscaled. Under this assumption, the water flow from the soil to the root xylem in a certain layer can be calculated directly from the average soil and xylem potentials in the layer and the radial root and rhizosphere conductances (Eq. 20). The upscaling of the radial conductances, which are all in parallel for a certain soil volume/layer, is trivial and equal to taking the sum of the conductances of all segments in a layer. The approaches that are used to relate leaf water potentials (or root collar potentials), to the average xylem and soil water potentials and uptake from a soil volume or layer can be classified into two groups: the parallel and big root approaches. Selecting between these depends on how the root segments in a certain soil volume are assumed to be connected to the root collar and the plant (Fig. 11).

The parallel root system is representative of a fibrous root system in which roots take up water near the root tips. The big root system could be representative of a tap root system with horizontal laterals. The upscaled axial conductances are obviously related to the assumed architecture of the upscaled network and linking the axial conductances of root segments to upscaled conductances remains a challenge. The upscaled root hydraulic model in Eq. (12) corresponds with a parallel root system, but the parameters of this model, $SUF(z)$ and K_{rs} , can be calculated directly from the true 3D root architecture without making assumptions. The impact of the axial conductances is therefore included in these parameters and the model may represent quite accurately 3D root hydraulic architectures (Vanderborght et al., 2021). Knowing the upscaled radial conductances, an upscaled effective axial conductance can be derived directly for each depth z from $SUF(z)$ and K_{rs} using Eq. (12). The obtained upscaled radial and axial conductances for each soil layer represent the properties of the upscaled parallel root model which is coupled to the rhizosphere model to solve for water potentials at the soil-root interface and the sink term profile in the 1D soil water flow equation (Eq. 10). Often models are further simplified by neglecting the axial resistance to flow in the root system, so that the xylem water potential is everywhere the same and the difference between parallel and big root models disappears (de Jong van Lier et al., 2013). These upscaled models consider both the root and soil hydraulic properties to calculate root uptake. Because of the non-linearity of the rhizosphere equation, the sink term needs to be solved iteratively, which increases the computational effort. Another simplification neglects resistance to flow in the rhizosphere and uses Eq. (12) to calculate the uptake from the soil water potentials directly without iterations. This approach only considers the root hydraulic properties and is appropriate when the soil is sufficiently wet but may overestimate the capacity of the root system to dry out the soil. It also leads to an overestimation of hydraulic redistribution and exchange into dry soil layers via the root system. As the approaches mentioned above consider water flow in the root system, they reproduce phenomena such as

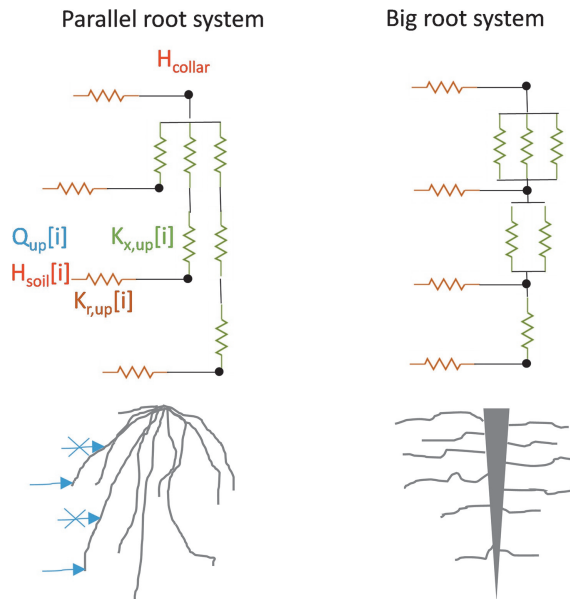


Fig. 11 Resistance networks considered in the parallel and big root systems and corresponding root system architectures. The parallel root system corresponds with a fibrous root system where water uptake takes place only near root tips, whereas a big root system corresponds with a taproot system with horizontal laterals.

root water uptake compensation and hydraulic lift. The local water uptake at a certain depth does not depend solely on soil water potentials at that depth but is also influenced by the water potential distribution in the entire root zone.

A further set of models do not simulate water flow in the root system and describe water uptake at a certain depth as a function of the atmospheric demand and the soil water potential at that depth. Using a similar formulation as Eq. (12), the sink term is described as:

$$S(z) = \alpha(h_{p,soil}(z), T_{pot})SUF(z)T_{pot} \quad (21)$$

where $SUF(z)$ is the normalized uptake distribution when the soil is sufficiently wet and α is a stress factor that varies between 0 and 1, which depends on the local soil water pressure head and the potential transpiration rate. $SUF(z)$ is mostly derived from the normalized root length density and α is a piecewise linear function with threshold soil water pressures, $h_{p,soil,crit}$, that depend on T_{pot} (Feddes et al., 1978).

$$\alpha(h_{p,soil}) = \frac{h_{p,soil} - h_{T,plant,crit}}{h_{p,soil,crit}(T_{pot}) - h_{T,plant,crit}} \text{ for } h_{T,plant,crit} < h_{p,soil} < h_{p,soil,crit}(T_{pot}) \quad (22)$$

The functional form of the stress factor α corresponds with supply functions that are derived using linear root hydraulics (see Fig. 8b). However, α is expressed in terms of soil water pressure heads whereas root hydraulics use total soil water potentials, which include elevation. The α function has been derived from observations of crop transpiration and root zone soil water potentials to obtain crop specific parameters. The α -function lumps the impact of the root system and soil hydraulic properties into one single function. Variations in hydraulic properties of root systems and of the relation between stomatal closure and leaf water potentials of different crops are depicted in different critical pressure heads for different crops (Table 2).

Root systems of a certain crop vary considerably with crop development and in reaction to soil and environmental conditions. Considering only root hydraulics, the critical soil water pressure head at which the uptake starts declining is obtained from:

$$h_{p,soil,crit} = \frac{T_{pot}}{K_{rs}} + h_{T,plant,crit} \quad (23)$$

For the example of the crop with $K_{rs} = 2 \times 10^{-4} \text{ d}^{-1}$ and $h_{T,plant,crit} = -150 \text{ m}$, $h_{p,soil,crit}$ would be -50 m for a daily averaged transpiration rate of 5 mm d^{-1} (which corresponds to a peak transpiration at midday of 20 mm d^{-1}), which is much more negative than the values listed in Table 2. The relatively high threshold pressure heads in Table 2, which are considerably larger than thresholds that are derived from root hydraulics (see also Fig. 10), reflect the impact of the soil hydraulic resistances on the stress functions. But, parameterizations of the α -stress function do not consider soil hydraulic properties.

The RWU modeled by Eq. (21) at a certain depth is not influenced by soil water pressures at other depths and does not increase when root water uptake decreases at other depths where the soil dries out. Redistribution of RWU and RWU compensation are hence not considered. To account for these processes without explicitly simulating flow in the root system, an additional stress index ω that quantifies the overall stress in the root zone was introduced (Jarvis, 1989):

$$\omega = \int_{L_r}^0 \alpha(h_{p,soil}(z), T_{pot})SUF(z)dz \quad (24)$$

When this stress index is larger than a critical index, ω_c , local reduction in water uptake due to local water stress can be fully compensated by redistributing the local uptake so that the supply, T_{act} , satisfies the demand, T_{pot} . The total uptake is described by the following equation (see Fig. 12):

$$T_{act} = \omega \frac{T_{pot}}{\max[\omega, \omega_c]} \quad (25)$$

For $\omega_c = 1$, the root system does not compensate whereas for $\omega_c \rightarrow 0$, the root system extracts all soil water at the potential transpiration rate until the water content in the root zone reaches the wilting point. When root water uptake compensation is considered, the model simulates that plants extract more water at a potential rate and dry out the soil more before transpiration

Table 2 Critical water potentials of the α -stress function (Eq. 22) (Wesseling, 1991).

Crop	$h_{p,soil,crit}$ (large T_{pot}) (m)	$h_{p,soil,crit}$ (small T_{pot}) (m)	$h_{T,plant,crit}$ (m)
Potatoes	-3.2	-6.0	-160
Sugar beet	-3.2	-6.0	-160
Wheat	-5.0	-9.0	-160
Pasture	-2.0	-8.0	-80
Corn	-3.25	-6.0	-80

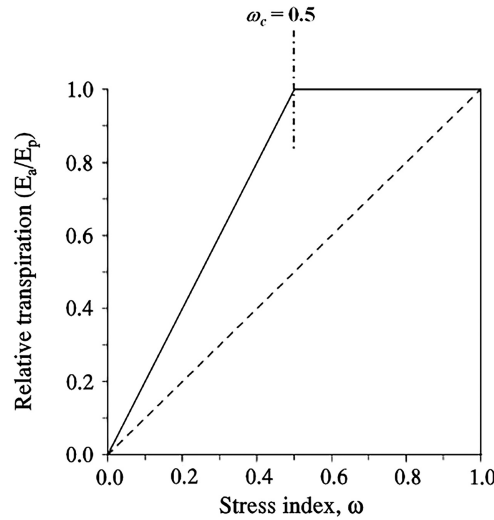


Fig. 12 Relative transpiration (T_{ac}/T_{pot}) versus stress index ω (Eq. 24) predicted by a model that considers root water uptake compensation: Eq. (25). The solid line represents T_{ac}/T_{pot} for a critical ω_c of 0.5 and dashed line represents T_{ac}/T_{pot} when there is no compensation, i.e., $\omega_c = 1$. From Jarvis NJ (2011) Simple physics-based models of compensatory plant water uptake: Concepts and eco-hydrological consequences. *Hydrology and Earth System Sciences* 15: 3431–3446.

is reduced. For a given stress level $\omega < \omega_c$, the simulated uptake is a factor ω_c^{-1} larger when compensation is considered. The local root water uptake is described by either:

$$S(z) = \alpha(h_{p,soil}(z), T_p)SUF(z) \frac{T_{pot}}{\max[\omega, \omega_c]} \quad (26)$$

or

$$S(z) = \left(\frac{\alpha(h_{p,soil}(z), T_{pot})}{\omega_c} - \max\left[\frac{\omega}{\omega_c}, 1\right] + 1 \right) SUF(z)T_{pot} \quad (27)$$

Both formulations predict the same uptake distribution when transpiration is lower than T_{pot} ($\omega < \omega_c$) but deviate from each other when the transpiration demand can be supplied ($\omega > \omega_c$). Eq. (27) was derived using a mechanistic description of water flow in the rhizosphere to root segments assuming that the root hydraulic conductance was large compared to the soil hydraulic conductance (Jarvis, 2011). In contrast to Eq. (26), which predicts $S \geq 0$ for all conditions, Eq. (27), predicts hydraulic lift and water redistribution via the root system, i.e., $S < 0$, at depths where $\alpha + \omega_c - \omega < 0$. This occurs only when plants are not stressed, i.e., $\omega > \omega_c$, and at locations where the soil is considerably drier than the mean soil water content, $\alpha < \omega$. Through the mechanistic derivation of Eq. (27), α , ω , and SUF are related to the soil and root system hydraulic properties and to the transpiration demand, T_{pot} , and can be derived from, in theory, measurable properties. Jarvis (2011) derived ω_c from resistances to flow in the rhizosphere while neglecting the flow resistance in the root system, whereas for relatively wet soil conditions, the latter is much larger than the former (see Fig. 2). Consequently, root water uptake redistribution under wet soil conditions was probably overestimated. Simply replacing the soil by root conductivities in the relation derived by Jarvis (2011), which would be more appropriate for wet soil conditions but less for dry conditions, gives the following expression for ω_c :

$$\omega_c = - \frac{T_{pot}}{K_{rs}h_{T,plant,crit}} \quad (28)$$

The critical stress index can be interpreted as the ratio of the potential transpiration to the maximal water supply to the canopy when the soil is saturated. When root density is considered a proxy for SUF and the soil profile is split into a wet or saturated domain and a dry domain, ω_c represents the minimal fraction of roots in saturated soil that is required to supply the potential transpiration. The smaller this ratio or fraction, the larger is the potential of the root system to redistribute water uptake and supply the water demand. Considering the crop with $K_{rs} = 2 \times 10^{-4} \text{ d}^{-1}$ and $h_{T,plant,crit} = -150 \text{ m}$, ω_c drops from 1 (no compensation) at $T_{pot} = 30 \text{ mm d}^{-1}$ to $\omega_c = 0.33$ for $T_{pot} = 10 \text{ mm d}^{-1}$. Mostly, ω_c is treated as a fixed parameter and is obtained from model fitting. When the dependence of ω_c on climate, soil, and root system dynamics is not considered, the fitted parameter should not be used to make predictions outside the range of conditions that were represented in the fitted dataset. Root hydraulic considerations show that ω_c and threshold pressures of the α -functions are related to each other since they depend on the same root hydraulic parameters. This implies that ω_c and α -function parameters should be derived together and not sequentially whereby, for example, parameters of the α function were derived first from fitting model simulations that did not consider compensation, i.e., assuming $\omega_c = 1$, to observations.

When ω_c is small, the root system can redistribute the water uptake within the root zone so that the supply rate does not depend strongly on the distribution of water in the root zone. In such a case, the reduction of water uptake compared to non-stressed

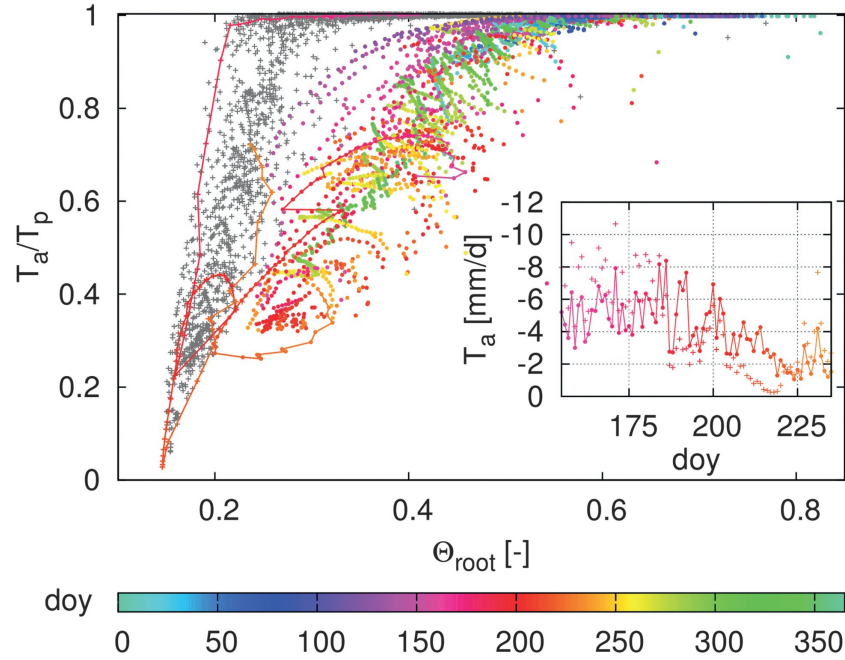


Fig. 13 T_{act}/T_{pot} simulated by a model that considers root water uptake compensation (gray dots) and by a model that does not consider root water uptake compensation (colored dots, color refers to the day of year, doy) and plotted versus the mean water saturation in the root zone $\langle \Theta \rangle_{root}$. The trajectory depicts a specific time series from June to the end of August and the inset shows the simulated T_{act} and infiltration during this period. From Schlüter S, Vogel HJ, Ippisch O and Vanderborght J (2013) Combined Impact of soil heterogeneity and vegetation type on the annual water balance at the field scale. *Vadose Zone Journal* 12: 1–17.

conditions can be described as a function of the average water content in the root zone. Fig. 13 shows simulated transpiration rate reductions in grassland that are plotted versus the average soil moisture content in the root zone and that are obtained using a model that does or does not ($\omega_c = 1$) consider root water uptake compensation. When compensation is considered, the reduction in transpiration occurs at lower water contents and the relation between the reduction factor and average soil moisture is less scattered.

This approach is used in 0D (zero-dimensional) soil water balance models that are used for irrigation planning (Doorenbos and Kassam, 1979), or in ecohydrology (Rodríguez-Iturbe and Porporato, 2004). It can describe the change in total water stored in the root zone over time as a function of different in- and outflows, which themselves are functions of the stored water content. The reduction functions consider a threshold fraction of the total plant-available water in the root zone (or its opposite, a threshold depletion fraction). For water contents above this threshold, the transpiration demand can be met by the supply. For water contents below this limit, the reduction factor is described as a decreasing function, either linear or convex, with decreasing water content (or increasing depletion factor). The threshold water content depends on the crop or vegetation type, transpiration demand, and soil type. In line with the process based models, greater water content (lower depletion factor) thresholds are used for larger transpiration rates, for vegetables than for grain crops (although no explicit link is made to the root system but increased thresholds would correspond to crops having a smaller root system conductance, shallow rooting depths and small densities) and for fine textured rather than coarse textured soils (Fig. 14). Although threshold soil water potentials are more negative for fine than for coarse soils (see Fig. 10d and f), the water contents are greater at these thresholds in fine-textured soils than in soils of coarse texture.

Uptake in saline soils

In saline soils, the osmotic potential is an important component of the total soil water potential. Since water flowing from the soil-root interface to the xylem must cross a cell membrane and since cell membranes show selectivity for transport of water compared to transport of ions, differences in osmotic potential between the xylem and the soil root interface drive water fluxes between the soil and the xylem. In spring before leaf burst and before the establishment of the transpiration stream, osmotic potentials in the xylem of tree roots are reduced to absorb water from the soil and induce a sap flow under a positive turgor towards the leaf buds. In saline soils, the low osmotic potentials in the soil water reduce the flow from the soil to the xylem. The leaky-pipe flow equation for a single root segment considering differences in osmotic potentials between the soil and the root xylem is given by:

$$\frac{d}{dt} k_x \frac{d(h_{p,xylem} + z)}{dl} = -2\pi r_{root} k_r (h_{p,soil-root} - h_{p,xylem} + \sigma(h_{o,soil-root} - h_{o,xylem})) \quad (29)$$

where σ is a ‘membrane reflection’ coefficient that varies between 0 (the membrane is not selective or is leaking and ions can pass like water) and 1 (membrane is perfectly selective and ions cannot pass) (Hopmans and Bristow, 2002). Gradients in osmotic

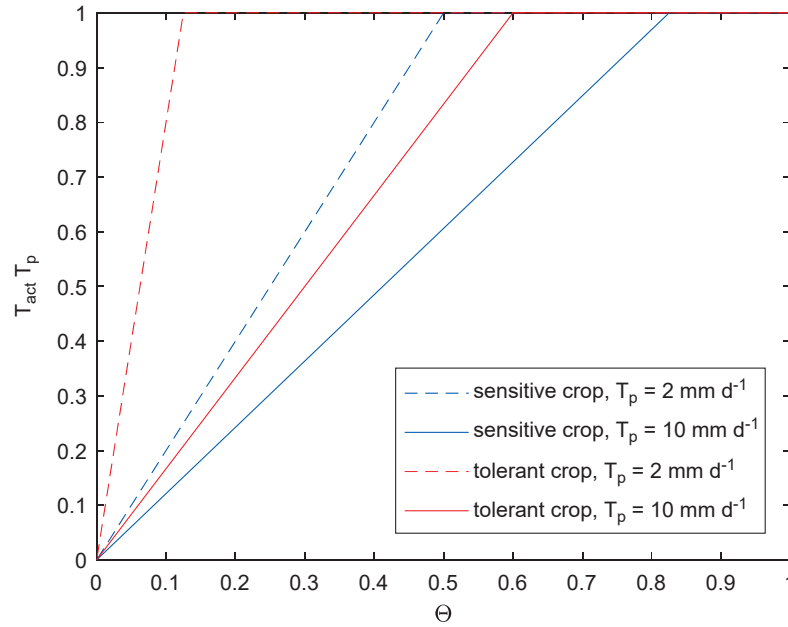


Fig. 14 Transpiration reduction function based on the available soil moisture relative to the maximal plant available soil moisture in the root zone, Θ , which is the ratio of $(\theta - \theta_w)/(\theta_{FC} - \theta_w)$ where θ_w is the water content at the permanent wilting point ($h_{p,soil} = -150$ m) θ_{FC} is the water content at ‘field capacity’. After Doorenbos J and Kassam AH (1979) *Yield Response to Water*. FAO: Rome.

potential drive a water flow when the mobility of ions is restricted compared to the mobility of water molecules. This is generally not the case within the soil or within the xylem, so water flow in the soil or in the xylem does not depend on gradients in osmotic potential. That is why the gradient in osmotic potential is not included in the left-hand side of Eq. (29). To describe the osmotic potentials in the soil and in the xylem, salt concentrations need to be predicted (Eq. 4) and transport equations that describe the transport of solutes need to be solved. As the transport of salts is driven by water fluxes and the water fluxes, in turn, depend on the salt concentrations, flow and transport in the soil-plant system are two-way coupled. Schröder et al. (2013) developed and used a model that simulates 3D salt transport in a coupled soil-root architecture model and accounts for the impact of osmotic potentials on root water uptake. They found that upscaled root hydraulic models that use total soil water potentials, including osmotic potential at soil root interfaces, can be applied directly to calculate root water uptake from saline soils. Upscaled soil models use bulk soil salinities or osmotic potentials to assess the impact of salinity on water uptake and use α -type stress functions similar to those used to evaluate the impact of low soil water pressure heads on root water uptake, e.g., Eq. (22) with $h_{p,soil}$ replaced by $h_{o,soil}$, (Maas and Hoffman, 1977) or using an S-shaped function (van Genuchten and Hoffman, 1984):

$$\alpha(h_{o,soil}) = \frac{1}{1 + \left(\frac{h_{o,soil}}{h_{o,soil,50}}\right)^p} \quad (30)$$

where p is a shape factor which is often taken to be 3 and $h_{o,soil,50}$ is the osmotic potential at which the potential transpiration is reduced by 50%. Because of salt accumulation, the osmotic potentials at soil-root interfaces are more negative than in the bulk soil and this difference increases with increasing advection flux towards the interface. This indicates that the salinity stress function should also show a sensitivity to the water uptake per root length with stress at less negative $h_{o,soil}$ for larger T_p and, or lower root length densities (de Jong van Lie et al., 2009). Another problem is mixed salinity and drought stresses. When water potentials at the soil-root interface are considered, the stress functions can be expressed in terms of the sum of $h_{p,soil-root}$ and $h_{o,soil-root}$. As the relations between bulk soil and interface potentials are different for osmotic and water pressure potentials, in theory a single stress function that depends on the bulk soil total water potential cannot be defined. This is in agreement with observations that the transpiration response shows different sensitivities to changes in $h_{p,soil}$ than to changes in $h_{o,soil}$. The transpiration response also shows interaction effects with a larger sensitivity to different salinity levels under wetter soil conditions than under drier conditions when the effect of $h_{p,soil}$ dominates. The following combined stress functions and variants of it have been proposed (Dudley and Shani, 2003; Homaei et al., 2002):

$$\alpha(h_{p,soil}, h_{o,soil}) = \frac{1}{1 + \left(\frac{h_{o,soil}}{h_{o,soil,50}} + \frac{h_{p,soil}}{h_{p,soil,50}}\right)^p} \quad (31)$$

or,

$$\alpha(h_{p,soil}, h_{o,soil}) = \alpha_p(h_{p,soil})\alpha_o(h_{o,soil}) \quad (32)$$

Different normalization factors for osmotic and pressure potential in Eq. (31) give different weight to the osmotic and drought stresses whereas Eq. (32) uses a multiplicative stress function. Because of the empirical nature of the stress functions, their parameters cannot be derived from direct measurements but must be derived from fitting functions of the soil pressure heads and osmotic potentials in the root zone to the observed plant responses. Therefore, it is difficult to define which functions and parameterizations are more appropriate. A mechanistic process model that combines water flow and solute transport in the soil root system may provide an alternative to obtain insight into the form of stress functions and how their parameters are related to soil and plant properties and to transpiration.

Conclusions and outlook

Water flow in the soil-plant system can be described according to the same physical laws and principles as the ones that apply to laminar flow in a tube or to flow in a porous medium. How much water can be taken up by plant roots from the soil profile without critical leaf water potentials being exceeded depends on the root and water distribution in the soil profile, and on the hydraulic properties of the soil and root segments. In root water uptake models that do not consider the detailed 3D water flow paths in the soil towards the root system, these dependencies are implicitly represented as a function of averaged soil water potentials or water contents in the root zone. The parameterization of these functions should therefore depend on both soil and root properties. A proper representation of the impact of root properties on these water uptake functions is important to identify root ideotypes based on model simulations. The concept of ideotypes is used in plant breeding and refers to a plant model which behaves or performs in a particular manner in a certain environment. Because plants grow, root systems and root distributions are, especially for annual plants, dynamic. This growth depends on environmental conditions so that plants and root systems show phenotypic plasticity, i.e., the same genotype develops differently in contrasting soil and climate conditions. Therefore, parameters of root water uptake functions should be coupled to the development and growth of a plant.

References

- Allen RG, Pereira LS, Raes D, and Smith M (1998) *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements*. Rome: FAO.
- Amenu GG and Kumar P (2008) A model for hydraulic redistribution incorporating coupled soil-root moisture transport. *Hydrology and Earth System Sciences* 12: 55–74.
- Boesten JTTI and Stroosnijder L (1986) Simple-model for daily evaporation from fallow tilled soil under spring conditions in a temperate climate. *Netherlands Journal of Agricultural Science* 34: 75–90.
- Böhm J (1893) Capillarität und saftsteigen. *Berichte der Deutschen Botanischen Gesellschaft* 11: 203–212.
- Cai GC, Vanderborght J, Langensiepen M, Schnepf A, Hugging H, and Vereecken H (2018) Root growth, water uptake, and sap flow of winter wheat in response to different soil water conditions. *Hydrology and Earth System Sciences* 22: 2449–2470.
- Carminati A and Javaux M (2020) Soil rather than xylem vulnerability controls stomatal response to drought. *Trends in Plant Science* 25: 868–880.
- Couvreux V, Vanderborght J, and Javaux M (2012) A simple three-dimensional macroscopic root water uptake model based on the hydraulic architecture approach. *Hydrology and Earth System Sciences* 16: 2957–2971.
- Cowan IR (1965) Transport of water in the soil-plant-atmosphere system. *Journal of Applied Ecology* 2: 221–239.
- de Jong van Lier QD, van Dam JC, Metselaar K, de Jong R, and Duijnsveld WHM (2008) Macroscopic root water uptake distribution using a matrix flux potential approach. *Vadose Zone Journal* 7: 1065–1078.
- de Jong van Lier QD, van Dam JC, and Metselaar K (2009) Root water extraction under combined water and osmotic stress. *Soil Science Society of America Journal* 73: 862–875.
- de Jong van Lier Q, van Dam JC, Durigon A, dos Santos MA, and Metselaar K (2013) Modeling water potentials and flows in the soil–plant system comparing hydraulic resistances and transpiration reduction functions. *Vadose Zone Journal* 12: 1–20.
- Dixon HH and Joly J (1895) XII. On the ascent of sap. *Philosophical Transactions of the Royal Society of London* 186(B): 563–576.
- Doorenbos J and Kassam AH (1979) *Yield Response to Water*. Rome: FAO.
- Draye X, Kim Y, Lobet G, and Javaux M (2010) Model-assisted integration of physiological and environmental constraints affecting the dynamic and spatial patterns of root water uptake from soils. *Journal of Experimental Botany* 61: 2145–2155.
- Dudley LM and Shani U (2003) Modeling plant response to drought and salt stress: Reformulation of the root-sink term. *Vadose Zone Journal* 2: 751–758.
- Fan Y, Miguez-Macho G, Jobbagy EG, Jackson RB, and Otero-Casal C (2017) Hydrologic regulation of plant rooting depth. *Proceedings of the National Academy of Sciences of the United States of America* 114: 10572–10577.
- Feddes RA, Kowalik PJ, and Zaradny H (1978) *Simulation of Field Water Use and Crop Yield*. Wageningen: Wiley.
- Homaee M, Feddes RA, and Dirksen C (2002) Simulation of root water uptake III. Non-uniform transient combined salinity and water stress. *Agricultural Water Management* 57: 127–144.
- Hopmans JW and Bristow KL (2002) Current capabilities and future needs of root water and nutrient uptake modeling. *Advances in Agronomy*. vol. 77, pp. 103–183. Elsevier.
- Jarvis NJ (1989) A simple empirical-model of root water-uptake. *Journal of Hydrology* 107: 57–72.
- Jarvis NJ (2011) Simple physics-based models of compensatory plant water uptake: Concepts and eco-hydrological consequences. *Hydrology and Earth System Sciences* 15: 3431–3446.
- Javaux M, Schröder T, Vanderborght J, and Vereecken H (2008) Use of a three-dimensional detailed modeling approach for predicting root water uptake. *Vadose Zone Journal* 7: 1079–1088.
- Landsberg JJ and Fowkes ND (1978) Water-movement through plant roots. *Annals of Botany* 42: 493–508.
- Lawson T and Matthews J (2020) Guard cell metabolism and stomatal function. *Annual Review of Plant Biology* 71: 273–302.
- Maas EV and Hoffman GJ (1977) Crop salt tolerance—Current assessment. *Journal of Irrigation and Drainage Division* 103: 115–134.
- Neumann RB and Cardon ZG (2012) The magnitude of hydraulic redistribution by plant roots: A review and synthesis of empirical and modeling studies. *New Phytologist* 194: 337–352.
- Ritchie JT (1972) Model for predicting evaporation from a row crop with incomplete cover. *Water Resources Research* 8: 1204.
- Rodríguez-Iturbe I and Porporato A (2004) *Ecohydrology of Water-Controlled Ecosystems*. Cambridge: Cambridge University Press.
- Schenk HJ and Jackson RB (2002) The global biogeography of roots. *Ecological Monographs* 72: 311–328.

- Schröder N, Lazarovitch N, Vanderborght J, Vereecken H, and Javaux M (2013) Linking transpiration reduction to rhizosphere salinity using a 3D coupled soil-plant model. *Plant and Soil* 377: 277–293.
- Sperry JS, Stiller V, and Hacke UG (2003) Xylem hydraulics and the soil–plant–atmosphere continuum: Opportunities and unresolved issues. *Agronomy Journal* 95: 1362–1370.
- Teuling AJ, Seneviratne SI, Williams C, and Troch PA (2006) Observed timescales of evapotranspiration response to soil moisture. *Geophysical Research Letters* 33.
- van Genuchten MT and Hoffman GJ (1984) Analysis of crop salt tolerance data. In: Shainberg I and Shalhevet J (eds.) *Soil Salinity Under Irrigation, Processes and Management*. vol. 51, pp. 258–271. New York: Springer Verlag.
- Vanderborght J, Couvreur V, Meunier F, Schnepf A, Vereecken H, Bouda M, and Javaux M (2021) From hydraulic root architecture models to macroscopic representations of root hydraulics in soil water flow and land surface models. *Hydrology and Earth System Sciences* 25: 4835–4860.
- Wesseling JG (1991) *Meerjarige Simulaties Van Grondwateronttrekking Voor Verschillende Bodemprofielen, Grondwatertrappen En Gewassen Met Het Model SWATRE*. Wageningen: Wageningen University & Research.
- Whittaker RH and Likens GE (1973) Primary production: The biosphere and man. *Human Ecology* 1: 357–369.

Crop and landscape water requirements

Luis S Pereira, Isabel Alves, and Paula Paredes, Departamento de Ciências e Engenharia dos Biosistemas, Instituto Superior de Agronomia, Universidade de Lisboa, Lisboa, Portugal

© 2023 Elsevier Ltd. All rights reserved.

Introduction and main concepts	571
Crop evapotranspiration	572
Reference evapotranspiration	573
Crop coefficients	574
The single crop coefficient approach	575
K_c adjustment for climate	580
K_c for other vegetation and adjustment to non-pristine conditions	581
The dual crop coefficient approach	581
K_{cb} for unknown conditions and partial-cover crops	582
Landscape water requirements	584
Irrigation water requirements	590
Limitations	591
Conclusions	592
References	592

Abstract

This article summarizes the essential concepts and methodologies for estimating plant water and irrigation requirements. Reference evapotranspiration is defined relative to a reference crop with constant canopy characteristics, allowing to parameterize the surface and aerodynamic resistances to heat and vapor fluxes. Crop evapotranspiration is obtained using crop coefficients, which reflect the canopy differences relative to the reference crop. Both time-averaged and dual-crop coefficients are explained, together with corrections for climate and plant density. An adaptation aiming at estimating landscape water requirements is made by introducing the concept of the landscape coefficient. Finally, essential information on the soil water balance to estimate crop and landscape irrigation requirements is provided.

Key points

- Reference evapotranspiration (ET_o) represents the climatic demand for evaporation;
- The crop coefficient (K_c) accounts for the specific crop characteristics that deviate from the reference crop surface (height, canopy structure and density, and albedo);
- Standard K_c values can be transferred between locations after correction for climate;
- Crop ET is reduced when crops are under water stress or when plant density during the mid-phase is below full ground cover;
- The dual K_c refers to the sum of K_{cb} relative to the crop and K_e representing soil evaporation, with K_{cb} calculable from the fraction of ground cover and plant height;
- Landscape water requirements can be estimated with a similar procedure, where plant type, density and management water stress are taken into account, as well as local microclimate as influenced by plot size and surrounding areas;
- Irrigation water requirements aim at satisfying plant water requirements and can be determined from the soil water balance.

Introduction and main concepts

Crop water requirements (CWR) are defined as the depth of water [mm] needed to meet the water consumed through evapotranspiration (ET_c) by a disease-free crop, growing in large fields under non-restricting soil conditions including soil water and fertility, and achieving full production potential under the given growing environment. If *crop evapotranspiration* (ET_c) is defined as the rate of evapotranspiration [mm d^{-1}] of a given crop as influenced by its growth stages, environmental conditions, and crop management to achieve the potential crop production, then the CWR is the sum of ET_c for the entire crop growth period. When management or environmental conditions deviate from the optimal, then that rate of evapotranspiration requires adjustment to the prevailing conditions and is called *adjusted* or *actual crop evapotranspiration* ($ET_{c \text{ act}}$). Both CWR and ET_c concepts apply to either irrigated or rainfed crops.

For irrigated crops, the concept of *CWR* has to be complemented by that of irrigation water requirement (*IWR*), which is the net depth of water [mm] that is required to be applied to a crop to complement the natural water input to fully satisfy its specific crop water requirement. The *IWR* is the fraction of *CWR* not satisfied by rainfall, soil water storage and groundwater contribution. When it is necessary to add a leaching fraction to assure appropriate leaching of salts in the soil profile, this depth of water is also included in *IWR*. In practice, *IWR* has to be converted into gross irrigation requirements (*GIR*) to take into consideration the beneficial water use fraction (*BWUF*) of the irrigation systems utilized.

Crop evapotranspiration

The rate of evapotranspiration ET [mm d^{-1}] can be computed with the Penman-Monteith (PM) equation:

$$ET = \frac{1}{\lambda} \frac{\Delta (R_n - G) + \rho c_p (e_s - e_a)/r_a}{\Delta + \gamma (1 + r_s/r_a)} \quad (1)$$

where λ is the latent heat of vaporization [kg m^{-3}], $R_n - G$ is the net balance of energy available at the surface [$\text{MJ m}^{-2} \text{d}^{-1}$], $(e_s - e_a)$ represents the vapor pressure deficit (*VPD*) of air at the reference (weather measurement) height [kPa], ρ represents mean air density [kg m^{-3}], c_p represents specific heat of air at constant pressure [$\text{kJ kg}^{-1} \text{°C}^{-1}$], Δ represents the slope of the saturation vapor pressure-temperature relationship at mean air temperature [kPa °C^{-1}], γ is the psychrometric constant [kPa °C^{-1}], r_s is the bulk surface resistance [s m^{-1}], and r_a is the aerodynamic resistance [s m^{-1}].

The PM Eq. (1) can theoretically be utilized for the direct calculation of crop evapotranspiration when the crop specific surface and aerodynamic resistances are available. However, data for these crop characteristics are scarce for most crops.

The transfer of heat and vapor from the evaporative surface into the air in the turbulent layer above a canopy is determined by the aerodynamic resistance r_a [s m^{-1}] between the surface and the reference level above the canopy:

$$r_a = \frac{\ln \left(\frac{z_m - d}{z_{om}} \right) \ln \left(\frac{z_h - d}{z_{oh}} \right)}{k^2 u_z} \quad (2)$$

where z_m is the height of wind velocity measurements [m], z_h is the height of air temperature and humidity measurements [m], d is the zero plane displacement height [m], z_{om} is the roughness length relative to momentum transfer [m], z_{oh} is the roughness length relative to heat and vapor transfer [m], u_z is the wind velocity at height z_m [m/s] and k is the von Karman constant ($= 0.41$).

Eq. (2) assumes that the evaporative surface may be represented as a “big leaf” inside the canopy. However, exchanges in the top layer of the canopy, between heights $d + z_{om}$ and the crop height h [m] are important as sources of vapor fluxes, namely on complete cover crops with a wet soil surface. Adopting $d + z_{om}$ as the level of the evaporative surface can lead to overestimation of r_a and underestimation of r_s . Thus, as an alternative, r_a can be computed from the top of the canopy, h :

$$r_a = \frac{\ln \left(\frac{z_m - d}{z_{om}} \right) \ln \left(\frac{z_h - d}{h - d} \right)}{k^2 u_z} \quad (3)$$

Both parameters d and z_{om} depend upon the crop height, h , and canopy architecture. Information exists relating d and z_{om} to h . Most of these relationships are crop specific. More general functions also consider the leaf area index, *LAI*, or the plant area index (Table 1).

The height z_{oh} is estimated as a fraction of z_{om} , commonly $z_{oh} = 0.1 z_{om}$ for short and fully developed canopies. The factor 0.2 is often preferred for tall and partial-cover crops. However, there is relatively small impact on *ET* calculations from selecting a z_{oh}/z_{om} ratio between 0.1 and 0.2.

The surface resistance, r_s [s m^{-1}], for full-cover canopies is often expressed by

$$r_s = r_l / LAI_{eff} \quad (4)$$

where r_l is the bulk stomatal resistance of a well-illuminated leaf [s m^{-1}], and LAI_{eff} is the effective leaf area index [dimensionless], usually taken as 0.5 *LAI* for a full cover crop. r_l usually increases as a crop matures and begins to ripen. Typical values for r_l and r_s are listed in Table 2. The use of these equations for prediction of crop water requirements is difficult due to differences among varieties and crop management practices. Information on stomatal conductance or stomatal resistance available in the literature is mainly oriented to physiological or eco-physiological studies, rather than practical agricultural management. Information on bulk stomatal resistances and how they vary with climate or water availability is still scarce. Eq. (4) has the limitation of considering surface

Table 1 Relationships of d and z_{om} with crop height, h , and leaf area index, *LAI*.

d	z_{om}	Comment
$a_1 h$ (a_1 commonly ranges between 0.6 and 1)	$b_1 h$ (b_1 commonly ranges between 0.03 and 0.12)	Most studies on field crops
0.7 h	0.1 h	Complete cover crops
0.67 h	0.12 h	Dense vegetation canopies ($LAI > 2$)
$h \left[1 - \frac{2}{\alpha LAI} (1 - e^{-\alpha LAI/2}) \right]$	$h e^{\alpha LAI/2} (1 - e^{-\alpha LAI/2})$	Perrier (1982)

α is an adjustment factor for *LAI* with $\alpha = 2$ for $f \geq 0.5$ and $\alpha = [2(1 - f)]^{-1}$ for $f < 0.5$, with f being the proportion of *LAI* lying above $h/2$.

Table 2 Typical values of the leaf resistance per unit leaf area (r_l) and bulk stomatal resistance (r_s) for several canopy types.

Canopy type	r_l ($s\ m^{-1}$)	$r_{l\ min}$ ($s\ m^{-1}$)	r_s ($s\ m^{-1}$)	$r_{s\ min}$ ($s\ m^{-1}$)
Tropical forest			125–150	50
Deciduous forest		45–150	70–160	50–60
Aspen	400			
Eucalyptus	200–400			
Maple	400–700	250		
Coniferous			70–150	30–60
Crops, general	50–320		30–130	20–150
Grain sorghum	200	150–200	100–140 (LAI = 2)	
Snapbeans		130		
Soybeans	120			50
Maize	160	70	80	25–40
Barley	150–250		45–70	40
Wheat		50		30
Alfalfa	80	50	40	
Sugar beets	100	50		
Clipped grass (0.15 m)	100–150		80–120	70
Clipped and irrigated grass (0.10–0.12 m)	75	40	40–60	25–30

Parameters $r_{l\ min}$ and $r_{s\ min}$ are minimum daytime values when all environmental variables are optimum.

From Allen RG, Pruitt WO, Businger JA, Fritschen LJ, Jensen ME, Quinn FH (1996) Evaporation and transpiration. In: Wootton TP, Cecilio CB, Fowler LC, Hui SL, Heggen RJ (Eds.) ASCE Handbook of Hydrology. New York: ASCE, pp. 125–252.

resistance to be a pure physiological parameter while in reality it also includes an aerodynamic component that regulates the transport of water vapor inside the canopy.

In general, r_s varies according to:

$$r_s = r_a \left(\frac{\Delta}{\gamma} \beta - 1 \right) + (1 + \beta) \frac{\rho c_p VPD}{\gamma (R_n - G)} \quad (5)$$

where β is the Bowen ratio (the ratio between the sensible and latent heat fluxes). In this equation β plays the role of a water-stress indicator. This equation illustrates that weather variables interact and their influences are interdependent, thus adding to the difficulties in appropriately selecting r_s .

Current research work is focused on improving our ability to apply the PM equation or multi-layer ET models to specific agricultural crops. Meanwhile, the main use of the PM equation is to compute the reference evapotranspiration, with ET_c being determined using crop coefficients.

Reference evapotranspiration

Crop evapotranspiration, ET_c [mm], can be calculated by multiplying the reference evapotranspiration, ET_o [mm], by a dimensionless crop coefficient, K_c :

$$ET_c = K_c ET_o \quad (6)$$

Reference evapotranspiration aims at representing the climatic demand for evaporation, and is defined as the ET from an extensive surface of reference vegetation having a uniform, standardized height, that is actively growing, free of diseases and parasites and not short of water that completely covers the soil surface. Two crops have been used as reference: cool-season clipped grass and alfalfa.

The reference grass crop is a hypothetical crop with an assumed height of 0.12 m having a fixed surface resistance and an albedo of 0.23, closely resembling an extense surface of green grass of uniform height, actively growing and adequately watered. The reference evapotranspiration ET_o [mm d⁻¹] can then be easily computed with the PM Eq. (1) since the aerodynamic and surface resistance terms can be parameterized, resulting in the FAO-Penman-Monteith (FAO- ET_o or PM- ET_o) equation:

$$ET_o = \frac{0.408 \Delta (R_n - G) + \gamma C_n u_2 (e_s - e_a)}{\Delta + \gamma (1 + C_d u_2)} \quad (7)$$

where, in addition to variables defined for Eq. (1), T is mean daily air temperature [°C] and u_2 is wind speed [m s⁻¹], both at 2 m height. C_n and C_d are coefficients that differ with calculation time step. For daily computations, surface resistance is considered to be 70 s m⁻¹, which leads to a C_n value of 900 and $C_d = 0.34$. When using an hourly time-step computation, $C_n = 37$, while $C_d = 0.24$ during daytime ($R_n > 0$), assuming a surface resistance of 50 s m⁻¹; during nighttime, $C_d = 0.96$, given a r_s of 200 s m⁻¹.

The alfalfa reference ET (ET_r) is defined as the ET from full cover, well-watered, actively growing alfalfa with a standardized height of 50 cm. In this case C_n and C_d in Eq. (7) are 1600 and 0.38 respectively for daily calculations, and 66 and 0.25 (daytime) or 1.7 (nighttime) for hourly calculations. Due to the lower surface resistance and higher aerodynamic roughness of alfalfa, ET_r values are higher than ET_o and approximate near maximum rates of ET expected from well-watered vegetation.

The computation of the PM- ET_o equation parameters should follow the procedures described by Allen et al. (1998). However, in many parts of the world data on some weather variables are often missing from datasets or are of low or questionable quality, namely solar radiation, air humidity and wind speed. They can then be estimated using the procedures proposed by Allen et al. (1998), with this being the approach recommended for most situations. The Hargreaves equation is an alternative that is based on values of T_{max} and T_{min} only and may give acceptable ET_o estimates, namely in hyper-arid and arid conditions (Raziei and Pereira, 2013) particularly when properly parameterized (Paredes et al., 2020).

Recently the use of remotely sensed data, such as those retrieved from geostationary satellites, to estimate ET_o has become popular. Reanalysis products are another important source of atmospheric variables, such as shortwave radiation and temperature, which are made available at various spatial resolutions and allow high temporal sampling. Reanalysis has the advantage of providing consistent information that span along several decades (Paredes et al., 2021).

Crop coefficients

While ET_o represents the climatic demand on evaporation, the K_c represents the influence of the specific crop characteristics on crop evapotranspiration. Being little influenced by climate, standard values for K_c can be transferred between locations and climates. Values of K_c based on ET_o are generally 1.2–1.4 greater than those based on the alfalfa reference ET (ET_r) given the differences between the two reference crops. When using tabulated values the user must be careful in selecting the appropriate K_c values according to the reference ET equation used. Grass reference ET_o is considered throughout this article.

K_c represents an integration of the effects of three primary characteristics that distinguish the crop from the reference: crop height, h (affecting roughness and aerodynamic resistance); crop-soil surface resistance, r_s (related to leaf area, fraction of ground covered by vegetation, leaf age and condition, degree of stomatal control, and soil surface wetness); and albedo, α , of the crop-soil surface (influenced by the fraction of ground covered by vegetation and soil surface wetness).

Two K_c approaches are considered. The first uses a time-averaged K_c to include multi-day effects of evaporation from the soil. The second concerns the basal crop coefficient, to obtain crop transpiration, and another coefficient representing the evaporation from the soil.

The crop coefficient curve represents the changes in K_c over the length of the growing season (Fig. 1). Its shape relates to changes in the vegetation and ground cover during plant development and maturation that affect the ratio ET_c/ET_o . Shortly after planting of annuals, or the initiation of new leaves for perennials, the value for K_c is often small. The K_c increases from that initial value, $K_{c\text{ ini}}$, at the beginning of rapid plant development and reaches a maximum, $K_{c\text{ mid}}$, at the time of maximum or near maximum plant development, the mid-season period. During the late season period, as leaves begin to senesce, the K_c begins to decrease until it reaches a lower value, $K_{c\text{ end}}$, at the end of the growing period.

The form for the equation used in the dual K_c approach is:

$$K_c = K_s K_{cb} + K_e \quad (8)$$

where K_s is the stress reduction coefficient [0–1], K_{cb} is the basal crop coefficient [0–~1.4], and K_e is the soil water evaporation coefficient [0–~1.4]. K_{cb} represents the ratio ET_c/ET_o when the soil surface layer is dry but the average soil water content of the root zone is adequate to sustain full plant transpiration, thus representing the base-line potential K_c in the absence of evaporation from the soil (Fig. 2). K_s reduces the value of K_{cb} , when the soil water content is not sufficient. Eq. (8) requires the calculation of a daily soil water balance for the root zone.

A time-averaged K_c can be calculated as:

$$K_c = \overline{K_{cb} + K_e} \quad (9)$$

where $\overline{K_{cb} + K_e}$ represents the time averaged sum of the basal K_{cb} and the soil evaporation coefficient, K_e , given actual values or possible scenarios of soil wetting frequencies. Typical shapes for the K_{cb} , K_e and $\overline{K_{cb} + K_e}$ curves are shown in Fig. 2. When summed,

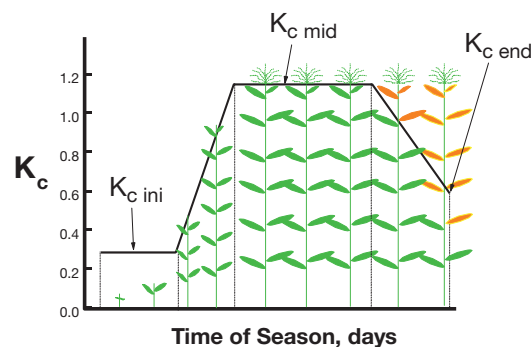


Fig. 1 Generalized crop coefficient curve from planting or initiation to harvesting or dormancy. From Allen RG, Pereira LS, Raes D and Smith M (1998) Crop evapotranspiration. In: *Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56. Rome: FAO.

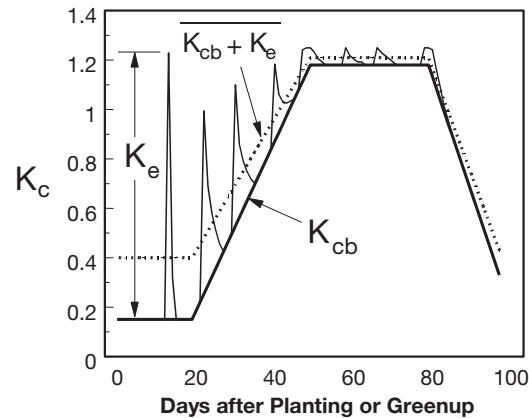


Fig. 2 Crop coefficient definitions showing the basal crop coefficient K_{cb} , which approximates canopy transpiration, the soil evaporation coefficient, K_e , and time-averaged $K_{cb} + K_e$ values. From Allen RG, Pereira LS, Raes D and Smith M (1998) Crop evapotranspiration. In: *Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56. Rome: FAO.

the values for K_{cb} and for K_e represent the total crop coefficient, K_c . The time-averaged K_c is useful for planning, irrigation system design, and typical irrigation management or for large scale water balance studies. The dual K_c is best where effects of day-to-day variation in soil surface wetness are important to estimate the resulting impacts on daily ET_c , soil moisture profile, and deep percolation.

The single crop coefficient approach

A simple procedure may be used to construct the K_c curve (Fig. 3):

1. Divide the growing period into four general growth stages that describe crop phenology or development, and determine the lengths [days] of these stages. The four crop growth periods are:
 - i) *Initial*: for annual crops, duration of this phase is from planting date to approximately 10% ground cover. For perennials, the planting date is replaced by the "greenup" date, when initiation of new leaves occurs.
 - ii) *Crop Development*: from 10% ground cover to effective full cover, which often occurs at the initiation of flowering or when LAI reaches 3.
 - iii) *Mid-Season*: from effective cover to start of maturity, which is often indicated by the beginning of the ageing, yellowing or senescence of leaves, leaf drop, or the browning of fruit.
 - iv) *Late Season*: from start of maturity to harvest or full senescence. For some perennial vegetation in frost-free climates, crops may grow year-round so that the date of termination may be taken as the same as the date of "planting".
2. Identify the three K_c values that correspond to $K_{c\text{ ini}}$, $K_{c\text{ mid}}$ and $K_{c\text{ end}}$.
3. Connect straight line segments through each of the four growth stage periods

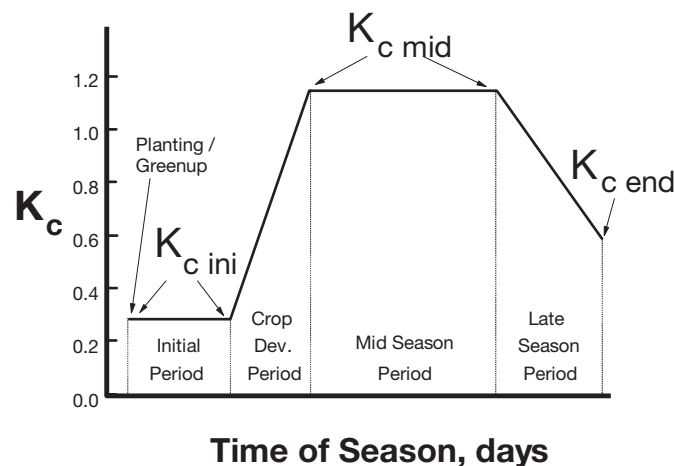


Fig. 3 Crop coefficient (K_c) curve and crop development stages definitions. From Allen RG, Pereira LS, Raes D and Smith M (1998) Crop evapotranspiration. In: *Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56. Rome: FAO.

The length of crop growth stages are crop-specific and change duration with crop variety, planting date, cultivation practices and weather conditions, mainly air temperature. The length of crop growth stages may be predicted using cumulative degree-based equations or plant growth models.

The lengths of the initial and development periods may be relatively short for deciduous trees and shrubs that develop new leaves in the spring at relatively fast rates. The $K_{c\text{ ini}}$ should then reflect the ground condition prior to leaf initiation, including the amount of grass or weed cover, soil wetness, tree density, and mulch density. The length of the late-season period may be relatively short for crops harvested before senescence or for vegetation killed by frost. The value for $K_{c\text{ end}}$ should reflect the soil surface condition and that of the vegetation following plant death or harvest. Indicative lengths of growth stages are given in Allen et al. (1998). However, tabulated lengths may deviate from the actual ones, so local observations or information should be used to incorporate the effects of plant variety, climate, cultural practices and water and salinity stresses.

Values for $K_{c\text{ ini}}$, $K_{c\text{ mid}}$ and $K_{c\text{ end}}$ are listed in Tables 3 and 4 for various agricultural crops. Usually there is close similarity in K_c within the same crop group, since the plant height, leaf area, ground coverage and water management are generally similar.

The $K_{c\text{ ini}}$ values in Tables 3 and 4 are only approximate because they vary widely with soil wetting conditions and because ET during the initial stage for annual crops is predominantly evaporation from the soil. Therefore, estimates for $K_{c\text{ ini}}$ must consider the frequency of irrigation and rainfall that wet the soil surface as well as the presence of crop residue or plastic or organic mulches.

Evaporation from bare soil, E_s [mm d^{-1}], can be characterized as occurring in two stages (Fig. 4). During stage 1, termed the “energy-limited” stage and having a duration t_1 [days], moisture is transported to the soil surface at a rate sufficient to supply the potential rate of evaporation, E_{so} [mm d^{-1}], which is governed by energy availability at the soil surface. E_{so} can be estimated as:

$$E_{so} = 1.15 ET_o \quad (10)$$

where ET_o is averaged for the initial period [mm d^{-1}].

Stage 2 is termed the “soil water-limited” stage, where hydraulic transport of subsurface water to the soil surface is smaller, thus making $E_s < E_{so}$. A portion of the evaporation occurs from below the soil surface, and energy is supplied by transport of heat into the soil profile. E_s decreases as soil moisture decreases and can be assumed to be linearly proportional to the depth of water remaining in the evaporation layer.

When the time interval [days] between two successive wettings, t_w , is greater than t_1 , $K_{c\text{ ini}}$ can be determined from:

$$K_{c\text{ ini}} = \frac{TEW - (TEW - REW) \exp\left(\frac{-(t_w - t_1) E_{so} \left[1 + \frac{REW}{TEW - REW}\right]}{TEW}\right)}{t_w ET_o} \quad (11)$$

where REW is the readily evaporable water, corresponding to the depth of evaporation when stage 1 drying is complete [mm], TEW is the total evaporable water, i.e. the maximum evaporation depth when soil evaporation effectively ceases [mm]. From the concept of stage 1 drying results $t_1 = REW/E_{so}$.

When $t_w < t_1$, the entire process resides within stage 1, then

$$K_{c\text{ ini}} = E_{so}/ET_o \quad (12)$$

Where furrow or trickle irrigation is practiced, and only a portion of the soil surface is wetted, $K_{c\text{ ini}}$ in Eqs. (11) and (12) should be reduced in proportion to the average fraction of wetted soil surface, f_w (ranging from 0, when no rain or irrigation occur, to 1). Indicative values for f_w are shown in Table 5. The infiltration depth from irrigation, I_w [mm], should also be adjusted:

$$I_w = I/f_w \quad (13)$$

where I is the total irrigation depth [mm].

REW is greater for medium-textured soils and is smaller for coarse soils. Maximum values for REW (REW_{max}) may be predicted according to soil texture:

$$\begin{aligned} REW_{\text{max}} &= 20 - 0.15 (\text{Sa}) \quad \text{for Sa} > 80\% \\ REW_{\text{max}} &= 11 - 0.06 (\text{Cl}) \quad \text{for Cl} > 50\% \\ REW_{\text{max}} &= 8 + 0.08 (\text{Cl}) \quad \text{for Sa} < 80\%, \text{ Cl} > 50\% \end{aligned} \quad (14)$$

where Sa and Cl are the fractions of sand and clay in the soil [%].

The TEW value is governed by the depth of soil contributing to evaporation, z_e [100–150 mm]. The soil water-holding properties within this evaporative layer, the presence of a hydraulically limiting layer beneath it, the unsaturated hydraulic conductivity, the conduction of sensible heat into the soil, and any root extraction of water from the evaporative layer all influence TEW . An approximation to TEW_{max} is:

$$TEW_{\text{max}} = z_e (\theta_{FC} - 0.5 \theta_{WP}) \quad (ET_o \geq 5 \text{ mm d}^{-1}) \quad (15)$$

$$TEW_{\text{max}} = z_e (\theta_{FC} - 0.5 \theta_{WP}) \sqrt{\frac{ET_o}{5}} \quad (ET_o < 5 \text{ mm d}^{-1}) \quad (16)$$

where θ_{FC} and θ_{WP} are the soil water content at field capacity and wilting point [mm mm^{-1}]. Typical values for θ_{FC} and θ_{WP} are given in Table 6.

Table 3 Single (time-averaged) crop coefficients, K_c , basal crop coefficient, K_{cb} , for non-stressed well-managed vegetable crops in subhumid climates ($RH_{min} \approx 45\%$, $u_2 \approx 2$ m/s) for use with the FAO Penman-Monteith ET_o , and indicative mean maximum plant heights, h , maximum root depths, $z_{r\max}$, and depletion fractions for no stress, p .

Crop	$K_{c\ ini}^a$	$K_{c\ mid}$	$K_{c\ end}$	$K_{cb\ ini}$	$K_{cb\ mid}$	$K_{cb\ end}$	Maximum Crop Height (h , m)	Maximum Root Depth ^b ($z_{r\max}$, m)	Depletion Fraction ^c (p)
a. Roots, tubers, bulbs and stem vegetable crops									
Asparagus	0.50	0.95	0.30	0.15	0.90	0.20	0.20–0.80	1.20–1.80	0.45
Carrots	0.70	1.05	0.95	0.15	0.95	0.85	0.30	0.30–0.50	0.30
Cassava: Year 1	0.30	1.00	0.60	0.15	0.90	0.45	1.00	0.50–0.80	0.50
Year 2	0.30	1.10	0.60	0.15	1.00	0.45	1.50	0.70–1.00	0.50
Celery	0.70	1.05	1.00	0.15	0.95	0.90	0.60	0.30–0.50	0.20
Garlic	0.70	1.05	0.70	0.15	0.95	0.60	0.50	0.30–0.50	0.30
Onions: Dry	0.70	1.05	0.65	0.15	0.95	0.55	0.45	0.30–0.60	0.30
Green	0.70	1.05	0.90	0.15	0.95	0.80	0.45	0.30–0.60	0.30
Seed	0.70	1.05	0.80	0.15	0.95	0.70	0.85	0.30–0.60	0.30
Parsnip	0.50	1.05	1.00	0.15	0.95	0.90	0.40	0.50–1.00	0.40
Potato: Long season	0.50	1.15	0.40	0.15	1.10	0.35	0.60	0.40–0.60	0.40
Short season	0.50	1.15	0.75	0.15	1.10	0.65	0.60	0.40–0.60	0.40
Radish	0.50	0.90	0.85	0.15	0.85	0.75	0.30	0.30–0.50	0.30
Rutabaga	0.50	1.10	1.00	0.15	1.00	0.90	0.60	0.50–1.00	0.50
Sweet potato	0.50	1.10	0.60	0.15	1.05	0.50	0.50	1.00–1.20	0.40
Table beets	0.50	1.05	1.00	0.15	0.95	0.90	0.40	0.60–1.00	0.45
Taro	0.50	1.10	1.05	0.15	1.05	1.00	1.20	0.30–0.40	0.40
Turnip	0.50	1.10	1.00	0.15	1.00	0.90	0.60	0.50–1.00	0.50
b. Leaves and flowers vegetable crops									
Artichoke	0.50	1.00	0.95	0.15	0.95	0.90	0.80	0.60–0.90	0.45
Broccoli	0.70	1.10	1.10	0.15	1.00	1.00	0.60	0.40–0.60	0.40
Brussel sprouts	0.70	1.05	1.05	0.15	0.95	0.95	0.75	0.40–0.60	0.40
Cabbage	0.70	1.05	1.00	0.15	0.95	0.90	0.40	0.30–0.50	0.40
Cauliflower	0.70	1.05	1.00	0.15	0.95	0.90	0.40	0.40–0.60	0.40
Lettuce	0.70	1.00	1.00	0.15	0.95	0.95	0.35	0.30–0.50	0.30
Spinach	0.70	1.00	1.00	0.15	0.90	0.90	0.35	0.30–0.50	0.20
c. Fruit and pod vegetable crops									
Bell pepper	0.60	1.10	1.05	0.15	1.05	1.00	0.70	0.50–1.00	0.30
Chili pepper	0.60	1.10	0.80	0.15	1.00	0.75	0.75	0.50–1.00	0.30
Chili, tabasco pepper: First harvest	0.60	1.10	0.70	0.15	1.05	0.65	0.80	0.50–1.00	0.40
Second harvest	0.70	1.05	0.60	0.15	1.00	0.55	0.80	0.50–1.00	0.40
Cucumber: For processing	0.50	1.05	0.90	0.15	1.00	0.80	0.30	0.60–1.20	0.40
Fresh market	0.60	1.05	0.80	0.15	1.00	0.70	0.30	0.60–1.20	0.40
Eggplant	0.60	1.05	1.00	0.15	1.00	0.90	0.80	0.40–1.00	0.45
Melon	0.50	1.05	0.85	0.15	0.95	0.75	0.30	0.60–1.20	0.40
Melon, cantaloupe	0.50	1.00	0.80	0.15	0.95	0.70	0.30	0.60–1.20	0.40
Okra	0.50	0.95	0.90	0.15	0.85	0.80	0.90	0.60–0.80	0.40
Pumpkin, winter squash	0.50	1.00	0.80	0.15	0.95	0.70	0.40	0.90–1.50	0.40
Squash, Zucchini	0.50	1.00	0.75	0.15	0.95	0.65	0.50	0.60–1.10	0.50
Strawberries	0.40	0.80	0.75	0.30	0.75	0.65	0.20	0.30–0.50	0.25
Tomato: For processing	0.70	1.10	0.90	0.15	1.05	0.85	0.70	0.60–1.20	0.40
Fresh market	0.70	1.10	1.00	0.15	1.05	0.90	0.70–1.50	0.60–1.20	0.40
Watermelon	0.50	1.05	0.70	0.15	0.95	0.60	0.40	0.80–1.50	0.40
d. Herbs, spices and special vegetable crops									
Basil: Single harvest	0.60	1.00	0.95	0.15	0.95	0.90	0.40	0.30–0.60	0.30
Multiple harvest	0.60	1.00	0.50	0.15	0.95	0.45	0.40	0.30–0.60	0.30
Black cumin	0.70	1.05	0.60	0.15	1.00	0.50	0.40	0.50–0.80	0.40
Coriander	0.60	1.10	0.75	0.15	1.05	0.70	0.30	0.30–0.50	0.30
Lemon balm: Single harvest	0.60	1.05	1.00	0.15	1.00	0.95	0.50	0.30–0.50	0.40
Multiple harvest	0.60	0.90	0.65	0.15	0.85	0.60	0.50	0.30–0.50	0.40
Mint	0.60	1.15	1.10	0.15	1.10	1.05	0.40	0.40–0.80	0.40
Parsley	0.60	1.00	0.90	0.15	0.95	0.85	0.30	0.30–0.50	0.30
Pineapple	0.50	0.90	0.90	0.15	0.80	0.80	1.20	0.30–0.60	0.50
Saffron	0.70	0.95	0.35	0.15	0.85	0.30	0.20	0.50–0.80	0.30

^aThe $K_{c\ ini}$ in the Table are only indicative. The procedure relative to Eqs. (11) and (12) should be used.

^bThe larger values for $z_{r\max}$ correspond to rainfed and/or water stressed conditions.

^cThe tabulated value for p must be corrected when $ET_c \neq 5$ mm d⁻¹ using $p = p_{table} + 0.04 (5 - ET_c)$.

From Allen, R.G., Pereira, L.S., Raes, D. and Smith, M. (1998). Crop evapotranspiration. In: Guidelines for Computing Crop Water Requirements. FAO Irrigation and Drainage Paper 56. Rome: FAO.; Pereira, L.S., Paredes, P., Lopez-Urrea, R., Hunsaker, D.J., Mota, M., Mohammadi Shad, Z. (2021a). Standard single and basal crop coefficients for vegetable crops, an update of FAO56 crop water requirements approach. Agricultural Water Management 243, 106196.

Table 4 Single (time-averaged) crop coefficients, K_c , basal crop coefficients, K_{cb} , for non-stressed well-managed field crops and forages in subhumid climates ($RH_{min} \approx 45\%$, $u_2 \approx 2$ m/s) for use with the FAO Penman-Monteith ET_o , and indicative mean maximum plant heights, h , maximum root depths, $z_{r\ max}$, and depletion fractions for no stress, p .

Crop	$K_{c\ ini}^a$	$K_{c\ mid}$	$K_{c\ end}$	$K_{cb\ ini}$	$K_{cb\ mid}$	$K_{cb\ end}$	Maximum Crop Height (h, m)	Maximum Root Depth ^b ($z_{r\ max}$, m)	Depletion Fraction ^c (p)
a. Grain legumes									
Bean: Green	0.50	1.05	0.95	0.15	1.00	0.85	0.50–0.70	0.50–0.90	0.45
Dry	0.50	1.05	0.40	0.15	1.00	0.30	0.50–0.70	0.50–0.90	0.45
Black and green gram:									
Black gram (dry)	0.40	1.15	0.35	0.15	1.10	0.30	0.50–0.70	0.60–1.00	0.45
Green gram	0.40	1.15	0.65	0.15	1.10	0.55	0.60–0.90	0.40–1.00	0.45
Chickpea (garbanzo)	0.40	1.10	0.35	0.15	1.05	0.25	0.50–0.70	0.70–1.00	0.45
Cowpea:									
Green	0.40	1.10	0.60	0.15	1.05	0.50	0.60–0.80	0.60–1.30	0.45
Dry	0.40	1.10	0.45	0.15	1.05	0.35	0.60–0.80	0.60–1.30	0.45
Faba bean:									
Fresh	0.50	1.10	1.05	0.15	1.05	0.95	0.80	0.50–0.70	0.45
Dry	0.50	1.10	0.40	0.15	1.05	0.30	0.80	0.50–0.70	0.45
Groundnut (peanut)	0.50	1.10	0.60	0.15	1.05	0.50	0.50	0.50–1.00	0.50
Lentil	0.50	1.05	0.30	0.15	1.00	0.20	0.50	0.60–0.80	0.50
Pea:									
Fresh	0.50	1.15	1.10	0.15	1.10	1.05	0.60	0.60–1.00	0.35
Dry	0.50	1.10	0.30	0.15	1.05	0.25	0.80	0.60–1.00	0.45
Soybean	0.50	1.15	0.35	0.15	1.10	0.25	0.80	0.60–1.40	0.50
b. Fiber crops									
Cotton	0.40	1.10	0.50	0.15	1.05	0.40	1.20	1.00–1.70	0.60
Flax (Linseed)	0.40	1.05	0.25	0.15	1.00	0.20	1.00	1.00–1.50	0.50
Sisal	0.30	0.55	0.55	0.15	0.50	0.50	1.50	0.50–1.00	0.80
c. Oil crops									
Camelina	0.40	1.10	0.45	0.15	1.05	0.40	0.80	1.00–1.50	0.60
Canola	0.50	1.10	0.35	0.15	1.05	0.25	1.00–1.50	0.80–1.30	0.50
Castorbean	0.35	1.10	0.55	0.15	1.05	0.45	1.00–1.50	0.80–1.30	0.60
Linseed (Flax)	0.35	0.95	0.25	0.15	0.90	0.20	0.90	1.00–1.50	0.60
Mustard	0.35	1.15	0.40	0.15	1.10	0.35	1.50–2.00	0.50–1.10	0.55
Safflower	0.35	1.10	0.25	0.15	1.05	0.20	1.10	1.00–2.00	0.60
Sesame	0.35	1.05	0.25	0.15	1.00	0.20	1.30	1.00–1.50	0.60
Sunflower	0.40	1.20	0.30	0.15	1.15	0.25	2.00	0.80–2.00	0.45
d. Sugar crops									
Sugar beet	0.50	1.05	0.75	0.15	1.00	0.65	0.50	0.70–1.20	0.55
Sugar cane	0.40	1.20	0.80	0.15	1.15	0.70	3.00–4.00	1.00–1.50	0.60
e. Cereals and grains									
Amaranth grain	0.30	1.10	0.25		1.05	0.20	2.00	0.50–1.50	0.55
Barley	0.30	1.05	0.25		1.00	0.20	0.70–0.90	0.60–1.20	0.55
Oats	0.30	1.05	0.25		1.00	0.20	0.80–1.10	1.00–1.50	0.55
Pearl Millet	0.30	1.10	0.35		1.05	0.25	2.00	1.00–2.00	0.55
Quinoa	0.30	1.10	0.50		1.05	0.45	1.00–1.20	0.60–1.20	0.55
Rye	0.30	1.00	0.35		0.95	0.30	0.90	0.60–1.20	0.55
Teff	0.30	1.05	0.25		1.00	0.20	1.10	0.60–1.20	0.55
Wheat, common:									
Winter, low grain moisture	0.60 ^d	1.15	0.25		1.10	0.20	0.70–1.10	1.00–1.50	0.55
Winter, high grain moisture	0.60 ^d	1.15	0.55		1.10	0.45	0.70–1.10	1.00–1.50	0.55
Spring, low grain moisture	0.60	1.15	0.25		1.10	0.20	0.70–1.10	1.00–1.50	0.55
Spring, high grain moisture	0.60	1.15	0.45		1.10	0.40	0.70–1.10	1.00–1.50	0.55
Wheat, durum:									
Winter, low grain moisture	0.60 ^d	1.05	0.25		1.00	0.20	0.70–1.10	1.00–1.50	0.55
Winter, high grain moisture	0.60 ^d	1.05	0.55		1.00	0.45	0.70–1.10	1.00–1.50	0.55
Maize:									
Grain, low moisture	0.40	1.20	0.30		1.15	0.25	2.50–3.50	0.60–1.50	0.50
Grain, high moisture	0.40	1.20	0.65		1.15	0.60			
Silage	0.40	1.20	0.95		1.15	0.85	2.50–3.20	0.60–1.50	0.50
Sweet	0.40	1.15	1.05		1.10	1.00	1.50–2.50	0.60–1.50	0.50
Sorghum:									
Grain	0.40	1.10	0.45	0.15	1.05	0.35	1.50–2.00	1.00–1.50	0.55
Silage	0.40	1.10	0.90	0.15	1.05	0.85	2.00–3.00	1.00–1.50	0.55
Sweet	0.40	1.15	0.95	0.15	1.10	0.90	3.00–4.00	0.60–1.50	0.55

Table 4 (Continued)

Crop	$K_{c \text{ ini}}$	$K_{c \text{ mid}}$	$K_{c \text{ end}}$	$K_{cb \text{ ini}}$	$K_{cb \text{ mid}}$	$K_{cb \text{ end}}$	Maximum Crop Height (h, m)	Maximum Root Depth ($z_{r \text{ max}}$, m)	Depletion Fraction (p)
Rice:									
Flooded	1.05	1.20	1.05	0.15	1.15	0.90	1.00	0.50	0.20 θ_{sat}
Flooded, dry seeding	0.85	1.20	1.05	0.15	1.15	0.90	1.00	0.50	0.20 θ_{sat}
Flooded, anticipated cut-off	1.05	1.20	0.80	0.15	1.15	0.70	1.00	0.50	0.20 θ_{sat}
Intermittent	0.95	1.20	1.00	0.15	1.15	0.85	1.00	0.70	0.20 θ_{sat}
Aerobic, sprinkler irrigation	0.90	1.10	0.95	0.15	1.05	0.85	0.80	1.00	0.35 ASW
Aerobic, surface irrigation	0.90	1.10	0.95	0.15	1.05	0.85	0.80	1.00	0.35 ASW
Rainfed	0.80	1.10	0.80	0.15	1.05	0.70	0.80	1.20	–
g. Forages									
Alfalfa Hay:									
averaged cutting effects	0.40	0.95	0.90				0.70	1.00–2.00	0.55
individual cutting periods	0.40 ^e	1.20 ^e	1.15 ^e	0.30 ^e	1.15 ^e	1.10 ^e	0.70	1.00–2.00	0.55
Grazing Pasture:									
rotated grazing	0.40	0.85–1.05	0.85	0.30	0.80–1.00	0.80	0.15–0.30	0.50–1.50	0.60
extensive grazing	0.30	0.75	0.75	0.30	0.70	0.70	0.10	0.50–1.50	0.60
Turf grass:									
cool season	0.90	0.95	0.95	0.85	0.90	0.90	0.10	0.50–1.00	0.40
warm season	0.90	0.95	0.95	0.85	0.90	0.90	0.10	0.50–1.00	0.50

^aThe $K_{c \text{ ini}}$ in the Table are only indicative. The procedure relative to Eqs. (11) and (12) should be used.

^bThe larger values for $z_{r \text{ max}}$ correspond to rainfed and/or water stressed conditions.

^cThe tabulated value for p must be corrected when $ET_c \neq 5 \text{ mm d}^{-1}$ using $p = p_{\text{table}} + 0.04 (5 - ET_c)$.

^d0.60 for non-frozen soils; in case of frozen soils $K_{c \text{ ini}} \sim 0.40$.

^eThese K_c coefficients for hay crops represent immediately following cutting; at full cover and immediately before cutting, respectively. The growing season is described as a series of individual cutting periods.

From Allen, R.G., Pereira, L.S., Raes, D. and Smith, M. (1998). Crop evapotranspiration. In: Guidelines for Computing Crop Water Requirements. FAO Irrigation and Drainage Paper 56. Rome: FAO.; Pereira, L.S., Paredes, P., Hunsaker, D.J., López-Urrea, R., Mohammadi Shad, Z. (2021b). Standard single and basal crop coefficients for field crops. Updates and advances to the FAO56 crop water requirements method. Agricultural Water Management 243, 106466.

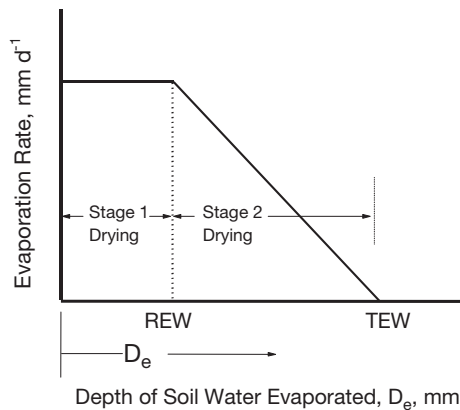


Fig. 4 Two stages model for soil evaporation. *TEW* and *REW* stand for total and readily evaporable water, respectively. From Allen RG, Pereira LS, Raes D and Smith M (1998) Crop evapotranspiration. In: *Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56. Rome: FAO.

Table 5 Indicative values of the average fraction of wetted soil surface, f_w .

Irrigation method	f_w
Precipitation	1.0
Sprinkler, basin and border irrigation	1.0
Furrow irrigation, every furrow, narrow bed	0.6–1.0
Furrow irrigation, every furrow, wide bed	0.4–0.6
Furrow irrigation with alternate furrows	0.3–0.5
Trickle irrigation	0.3–0.4

Allen et al. (1998).

Table 6 Typical values of soil water content at field capacity, θ_{FC} , soil water content at wilting point, θ_{WP} , readily evaporable water, REW, and total evaporable water, TEW, for different soil types.

Soil type (USA Soil Texture Classification)	Soil water characteristics			Amount of water that can be depleted by evaporation for $Z_e = 0.10$ m	
	θ_{FC} $m^3 m^{-3}$	θ_{WP} $m^3 m^{-3}$	$(\theta_{FC} - \theta_{WP})$ $m^3 m^{-3}$	Stage 1 REW mm	Stages 1 and 2 TEW ^a mm
Sand	0.07–0.17	0.02–0.07	0.05–0.11	2–7	6–12
Loamy sand	0.11–0.19	0.03–0.10	0.06–0.12	4–8	9–14
Sandy loam	0.18–0.28	0.06–0.16	0.11–0.15	6–10	15–20
Loam	0.20–0.30	0.07–0.17	0.13–0.18	8–10	16–22
Silt loam	0.22–0.36	0.09–0.21	0.13–0.19	8–11	18–25
Silt	0.28–0.36	0.12–0.22	0.16–0.20	8–11	22–26
Silt clay loam	0.30–0.37	0.17–0.24	0.13–0.18	8–11	22–27
Silty clay	0.30–0.42	0.17–0.29	0.13–0.19	8–12	22–28
Clay	0.32–0.40	0.20–0.24	0.12–0.20	8–12	22–29

Allen et al. (1998).

^aTEW = $(\theta_{FC} - 0.5 \theta_{WP}) Z_e$.

The average total water available for evaporation, W_a [mm], during each drying cycle is computed from the average depth [mm] added to the evaporative layer at each wetting:

$$W_a = P_{mean} + W_{ini}/n_w \quad (17)$$

where W_{ini} is the available soil water [mm] in the evaporation layer at the time of planting, n_w is the number of wetting events and P_{mean} is the average depth (mm) of water added to the evaporating layer at each wetting event. P_{mean} can be obtained with:

$$P_{mean} = \left(\sum P_n + \sum I_w \right) / n_w \quad (18)$$

but where each value of P_n and I_w must be limited to $P_n \leq TEW_{max}$ and $I_w \leq TEW_{max}$. The values for TEW and REW in Eq. (11) are calculated from TEW_{max} and REW_{max} as:

$$TEW = \min(TEW_{max}, W_a) \quad (19)$$

and

$$REW = REW_{max} \left[\min \left(\frac{W_a}{TEW_{max}}, 1 \right) \right] \quad (20)$$

When Eq. (14) to Eq. (16) are used, appropriate checking of results is required.

K_c adjustment for climate

The $K_{c \text{ mid}}$ and $K_{c \text{ end}}$ values in Tables 3 and 4 are valid for a standard subhumid climate where $RH_{min} = 45\%$ and $u_2 = 2 \text{ m s}^{-1}$. Under humid and calm conditions, the K_c for full-cover agricultural crops generally do not exceed 1.0 by more than about 0.05, because full-cover agricultural crops and the reference crop behave similarly regarding absorption of short-wave radiation, the primary energy source for evaporation under humid and calm conditions. Because the VPD is small under humid conditions, differences in ET caused by differences in r_a between the agricultural and the reference crop are also small, especially with low-to-moderate wind speeds. The values of K_c are thus less dependent on differences between the aerodynamic components of ET_c and ET_o . In contrast, under arid conditions, the effect of differences in r_a between the agricultural and the reference crop on ET_c become more pronounced because the VPD is often large. Hence, K_c will be larger under arid conditions, mainly for tall crops. Because the $K_{c \text{ mid}}$ and $K_{c \text{ end}}$ in Tables 3 and 4 represent conditions where $RH_{min} \approx 45\%$ and $u_2 \approx 2 \text{ m s}^{-1}$, when climatic conditions deviate from these values, the tabulated values need to be adjusted to the local climate:

$$K_c = K_{c \text{ (Table)}} + [0.04(u_2 - 2) - 0.004(RH_{min} - 45)] \left(\frac{h}{3} \right)^{0.3} \quad (21)$$

where $K_{c \text{ (Table)}}$ represents the $K_{c \text{ mid}}$ or $K_{c \text{ end}}$ taken from Tables 3 or 4, u_2 is the average daily wind speed at 2 m height [m s^{-1}], RH_{min} is the average daily minimum relative humidity [%], and h is the average plant height [m], with all averages referring to the midseason or the late-season period. Despite indicative values for h being presented in Tables 3 and 4, it is better to determine h from field observations. When crops are allowed to senesce and dry in the field ($K_{c \text{ end}} < 0.45$), no adjustment is necessary for $K_{c \text{ end}}$.

K_c for other vegetation and adjustment to non-pristine conditions

For large stands (larger than about 500–2000 m²) of vegetation not listed in the tables, for which the one-dimensional flux equations such as the PM Eq. (1) can be applied, a maximum upper limit on ET exists due to energy constraints. It is therefore possible to establish a maximum K_c for full cover vegetation, using:

$$K_{c \max} = \max \left\{ \left\{ 1.2 + [0.04 (u_2 - 2) - 0.004 (RH_{\min} - 45)] \left(\frac{h}{3} \right)^{0.3} \right\}, K_{cb} + 0.05 \right\} \quad (22)$$

where u_2 , $RH_{\min}$ and h are average values for the period considered, which may refer to the entire growth stage under consideration or, when more detailed computations are applied, be averaged for shorter periods (e.g., 5 days). $K_{c \max}$ ranges from about 1.05 up to 1.30.

The K_c values in Tables 3 and 4 represent potential water use by healthy, disease-free, and densely planted stands of vegetation, with adequate levels of soil water. When conditions are less than optimal, and crop height, plant density or leaf area index fall below full ground cover (that is, when $f_c < 0.90$ or $LAI < 3$), a reduction of the $K_{c \text{ mid}}$ and/or $K_{c \text{ end}}$ should be made, using:

$$K_c = K_{\text{soil}} + K_d \left[\max \left(K_{c \text{ full}} - K_{\text{soil}}, \frac{K_{c \text{ full}} - K_{\text{soil}}}{2} \right) \right] \quad (23)$$

where K_{soil} is the average K_c for bare soil, that should reflect the soil type and frequency of wetting as well as ET_o , and $K_{c \text{ full}}$ is the K_c for a full cover stand. K_d is a density coefficient, that will be further developed in a later section.

Other procedures, including remote sensing, may be used to estimate the K_c values for non-pristine conditions (Anderson et al., 2012; Pôças et al., 2020).

The dual crop coefficient approach

The basal crop coefficient, K_{cb} (Eq. 8) represents primarily the transpiration component of crop ET . Its use provides for separate adjustment for wet soil evaporation immediately following rain or irrigation events. This results in more accurate estimates of ET_c when computed on a daily basis. The K_{cb} curve has the same shape as the K_c curve (Fig. 1) and three values for K_{cb} are also required to draw the K_{cb} curve: $K_{cb \text{ ini}}$, $K_{cb \text{ mid}}$, and $K_{cb \text{ end}}$ respectively at the initial, mid-season and end-season. Recommended values for K_{cb} can be found in Tables 3 and 4, which must be adjusted for climate using a similar equation to Eq. (21), with the form:

$$K_{cb} = K_{cb(\text{Table})} + [0.04(u_2 - 2) - 0.004(RH_{\min} - 45)] \left(\frac{h}{3} \right)^{0.3} \quad (24)$$

The computation of the soil water evaporation coefficient K_e is based on the fact that evaporation from the soil is governed by the amount of energy available at the soil surface, which depends, in turn, on the portion of total energy that has been consumed by plant transpiration. Whenever the soil is wet and exposed to radiation, evaporation from the soil is predicted to occur at a maximum rate, thus K_e is also at a maximum. However, the sum $K_c = K_{cb} + K_e$ is limited by a maximum value of K_c ($K_{c \max}$, Eq. 22) following rain or irrigation. K_e decays after a wetting event depending on the cumulative amount of water evaporated from the surface soil layer. K_e can be calculated from

$$K_e = K_r (K_{c \max} - K_{cb}) \quad (25)$$

where K_r is the evaporation reduction coefficient [0–1]. Understandably, K_e is limited by the fraction of wetted soil exposed to sunlight, f_{ew} [0.01–1], so:

$$K_e \leq f_{ew} K_{c \max} \quad (26)$$

The method used to estimate evaporation from soil is similar to the one used to compute $K_{c \text{ ini}}$ where the evaporation rate is at the maximum rate until the depth of water depleted by evaporation, D_e [mm], equals REW (Fig. 4). When $D_e > REW$, the evaporation process is in stage 2, and its rate decreases in proportion to the remaining water. Therefore, K_r (Eq. 25) may be calculated as:

$$K_r = 1 \quad \text{for } D_e \leq REW \quad (27a)$$

$$K_r = \frac{TEW - D_e}{TEW - REW} \quad \text{for } D_e > REW \quad (27b)$$

D_e , the current depth of water depleted from the fraction of wetted soil exposed to sunlight, is computed from the daily water balance of the upper 100–200 mm of the soil:

$$D_{e \ i} = D_{e \ i-1} - (P_i - RO_i) - \frac{I_i}{f_w} + \left(\frac{K_e ET_o}{f_{ew}} \right)_i + T_{s \ i} \quad [0 \leq D_{e \ i} \leq TEW] \quad (28)$$

where P is the precipitation [mm], RO is runoff [mm] [$0 \leq RO \leq P$], I is the irrigation depth [mm], $(K_e ET_o/f_{ew})$ is the evaporation from the f_{ew} fraction of the exposed soil surface [mm], T_s is the transpiration from the f_w fraction of the evaporating soil layer [mm], and the subscript i refers to the day of estimation. To initiate the water balance, $D_e = 0$ immediately following a heavy rain or irrigation, or $D_e = TEW$ if a long time has passed since the last wetting. Precipitation (P) less than $0.2 ET_o$ may be ignored. For most applications $RO = 0$ and for most crops, except for very shallow-rooted crops, T_s can be neglected.

When the complete soil surface is fully wetted (e.g., by precipitation or sprinkler irrigation), $f_{ew} = 1 - f_c$, where f_c is the average fraction of ground covered by vegetation [0–0.99]. For irrigation systems where only a fraction of the soil surface is wetted, f_{ew} may be calculated as:

$$f_{ew} = \min(1 - f_c, f_w) \quad (29)$$

f_c should be observed in the field. When not observed, f_c can be estimated daily from:

$$f_c = \left(\frac{K_{cb} - K_{c \min}}{K_{c \max} - K_{c \min}} \right)^{1 + 0.5h} \quad (30)$$

$K_{c \min}$ is the minimum K_c for dry, bare soil [0.00–0.20]. The exponent " $1 + 0.5h$ " represents the effect of plant height on shading the soil and in increasing the K_{cb} . When programming, it is useful to set $(K_{cb} - K_{c \min}) \geq 0.01$ for numerical stability.

K_{cb} values are reduced when the soil water content in the root zone is too low to sustain transpiration at potential levels. The reduction is made through the water stress coefficient, K_s [0–1], that can be computed as:

$$K_s = \frac{\theta - \theta_{WP}}{\theta_p - \theta_{WP}} \quad (\theta < \theta_p) \quad (31)$$

where θ is actual average soil water content in the root zone [mm mm^{-1}] and θ_p is the threshold θ below which transpiration is decreased due to water stress [mm mm^{-1}]. By definition, $K_s = 1.0$ for $\theta > \theta_p$. The threshold θ_p is:

$$\theta_p = (1 - p) (\theta_{FC} - \theta_{WP}) \quad (32)$$

where p is the depletion fraction for no stress [0–1]. Indicative values can be found in Tables 3 and 4. The determination of K_s requires a daily balance of soil water content.

K_{cb} for unknown conditions and partial-cover crops

For vegetation with partial ground cover ($LAI < 3$) where the K_{cb} is not known, $K_{cb \text{ mid}}$ can be approximated as:

$$K_{cb \text{ mid}} = K_{c \min} + K_d (K_{cb \text{ full}} - K_{c \min}) \quad (33)$$

where K_d is a density coefficient, $K_{c \min}$ is the minimal basal K_c for bare soil (0.15 under typical agricultural conditions, but reaching values close to 0 when large periods without rainfall or irrigation occur) and $K_{cb \text{ full}}$ is the basal K_{cb} anticipated for the crop under full cover conditions ($LAI > 3$) under full water supply and corrected for climate, with:

$$K_{cb \text{ full}} = F_r \left(\left\{ \min[(1.0 + k_h h) 1.2] \right\} + 0.04(u_2 - 2) - 0.004(RH_{\min} - 45) \left(\frac{h}{3} \right)^{0.3} \right) \quad (34)$$

with $k_h = 0.1$ for tree and vine crops and tall field crops, and 0.2 for short crops and vegetables (Pereira et al., 2021c). The parameter F_r can be approximated by:

$$F_r \approx \frac{\Delta + \gamma(1 + 0.34u_2)}{\Delta + \gamma \left(1 + 0.34u_2 \frac{r_l}{r_{typ}} \right)} \quad (35)$$

where r_l and r_{typ} are the mean and the typical leaf resistance. r_{typ} is normally considered to be 100 s m^{-1} (Allen et al., 1998), with Mobe et al. (2020) using a value corresponding to the minimum unstressed canopy resistance when estimating K_{cb} values for apple orchards. F_r introduces an empirical adjustment for vegetation that exhibits more stomatal control than is typical of most agricultural crops, and plays an important role in the case of trees and vines. The standard value for F_r is 1 for annuals before leaf senescence, and decreases during late season due to leaf senescence. For perennials F_r also decreases during the late season but is also smaller than 1.0 when pruning and training control plant vigor. Pereira et al. (2021c) provide tables of F_r values for a wide variety of annuals and perennial crops.

For small, isolated stands, the internal boundary layer above the vegetation may not be in equilibrium with the underlying surface, particularly with well ventilated, tall trees. Small expanses of tall vegetation surrounded by shorter cover can result in a "clothesline effect" where the interchange between air and vegetation is much more efficient than over large expanses of homogeneous vegetation. In these cases, ET from isolated vegetation stands on a per unit area basis, may be significantly greater (1.6–2 times) than the corresponding ET_o computed for a grass reference (Allen et al., 1992), and $K_{cb \text{ full}}$ has to be increased beyond the value given by the equations above. On the other hand, the elevated ET from tall, narrow stands of vegetation can reduce ET

from adjacent shorter vegetation due to shading, blocking of wind and also by cooling and humidifying the air, so an integrated approach has to be considered where there is a mixture of vegetation types and heights, as developed for landscapes hereafter.

The density coefficient K_d describes the reduction in K_{cb} due to a reduction in the amount of vegetation. When LAI is known K_d can be approximated as

$$K_d = 1 - e^{-0.7 LAI} \quad (36)$$

where 0.7 is a common value for the light extinction coefficient inside the canopy. A different value can be used if more appropriate information regarding the specific crop is available.

When estimates of ground surface covered by vegetation are available, K_d may be estimated as:

$$K_d = \min \left(1, M_L f_{c\text{eff}}, f_{c\text{eff}}^{1/(1+h)} \right) \quad (37)$$

where $f_{c\text{eff}}$ is the effective fraction of soil covered or shaded by vegetation near solar noon [0.01–1], and M_L describes the effect of canopy density on shading and on maximum relative ET per fraction of ground shaded [1.5–2.0 during mid-season].

The calculation of $f_{c\text{eff}}$ depends on the plant spacing pattern. For tree crops and randomly planted vegetation, $f_{c\text{eff}}$ depends on the effect of the angle of the sun above the horizon (β) on the shaded area during the period of maximum ET (generally between 11:00 h and 15:00 h) (Fig. 5), with:

$$f_{c\text{eff}} = \frac{f_c}{\sin \beta} \leq 1 \quad (38)$$

For row crops, information is necessary to first compute the height to width ratio (HWR) of the crop row, namely the angle of plant row from east-west direction (Γ), the crop height (h) and the crop width as viewed from an east-west direction (Fig. 6), so that:

$$HWR = \frac{h \cos \Gamma}{\text{width}} \quad (39)$$

with $f_{c\text{eff}}$ being then calculated as:

$$f_{c\text{eff}} = f_c \left[\frac{HWR}{\tan \beta} \right] \quad (40)$$

The M_L multiplier imposes an upper limit on the relative magnitude of transpiration per unit area as represented by $f_{c\text{eff}}$ and is an attempt to simulate the physical limits on water flux through the plant (root, stem, leaves). It mostly ranges from 1.5 to 2.0 during the mid-season but as it depends on the canopy density and thickness it can be modified to fit the specific vegetation and the respective crop stage. Crops with more vertical production training systems, or with less leaf density or with leaves of very small dimension have lower M_L values. M_L values impact K_d mostly when $h > 3.0$ and $f_c < 0.60$, so careful parameterization has to be performed when dealing with high trees or sparse crops (Pereira et al., 2021c).

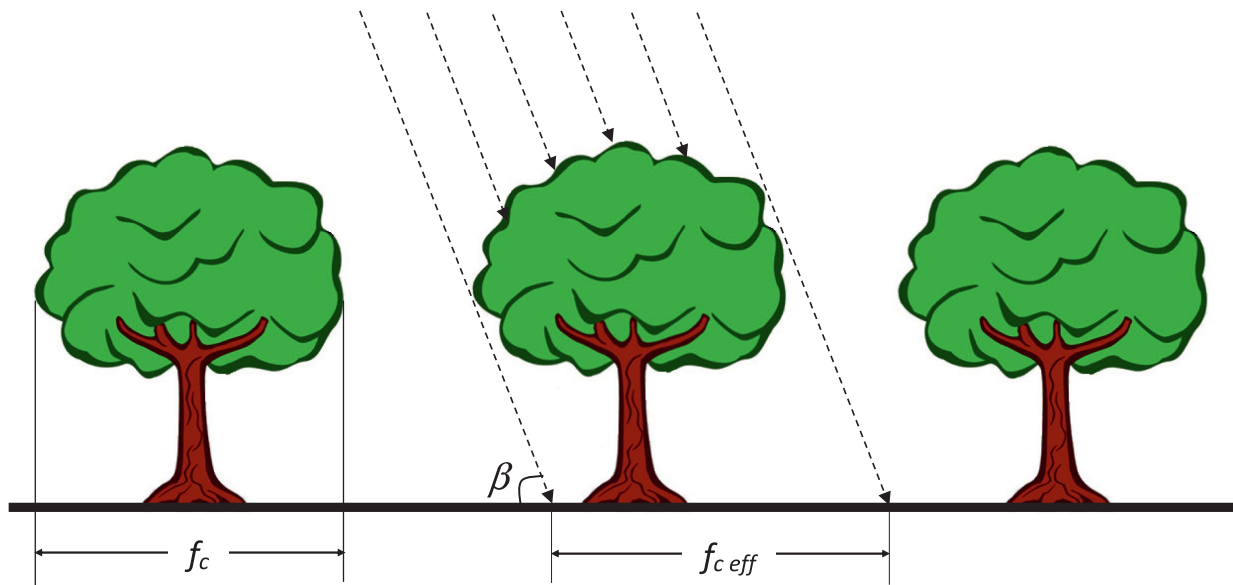


Fig. 5 Schematic illustration of the calculation of $f_{c\text{eff}}$ from f_c and β (Allen and Pereira, 2009).

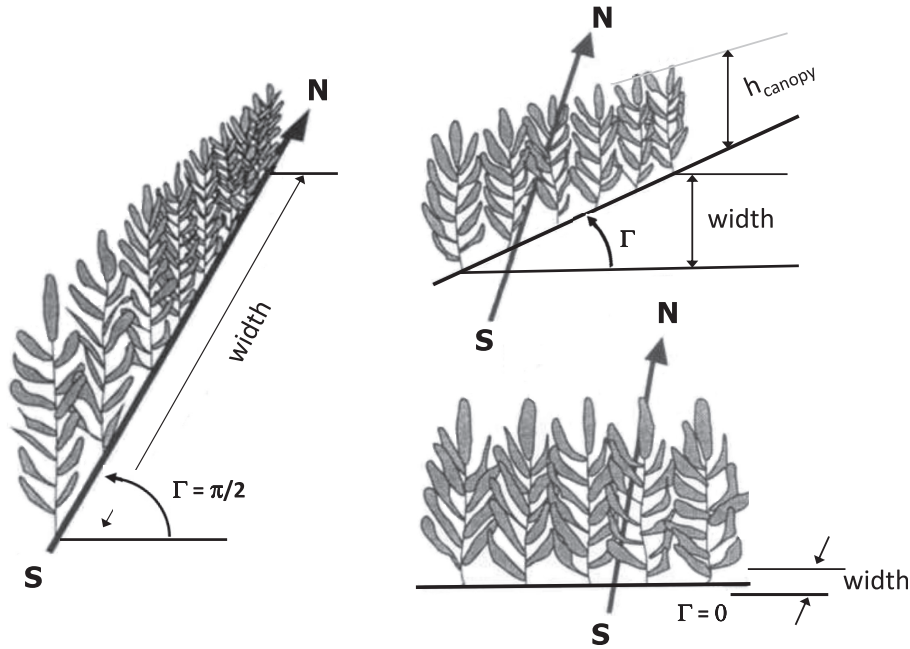


Fig. 6 Schematic representation of a row crop showing main canopy characteristics—width, height and angle Γ of plant row from an east-west direction that are used for the calculation of the height to width ratio (HWR) and effective ground cover $f_{c\ eff}$ for row crops. Adapted from Allen RG, Pereira LS, Raes D and Smith M (1998) Crop evapotranspiration. In: *Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56. Rome: FAO.

In orchards it is common to have natural vegetation or grass covering the ground with the purpose of enhancing infiltration of rainfall and reducing soil erosion. To consider the contribution of the ground cover to the total evapotranspiration, K_{cb} can be estimated as:

$$K_{cb} = K_{cb\ cover} + K_d \left(\max \left[K_{cb\ full} - K_{cb\ cover}, \frac{K_{cb\ full} - K_{cb\ cover}}{2} \right] \right) \quad (41)$$

where $K_{cb\ cover}$ is the K_{cb} of the ground cover in the absence of tree cover. The second term of the $\max$ function sets the estimated K_{cb} to half the difference between $K_{cb\ full}$ and $K_{cb\ cover}$ when this difference is negative, and accounts for the effects of the shading of the active ground cover by the taller trees. The value of $K_{cb\ cover}$ should be chosen adequately so as to reflect the density and vigor of the surface cover on sunlit areas.

Updated values for $K_{cb\ ini}$, $K_{cb\ mid}$, $K_{cb\ end}$, $K_{c\ ini}$, $K_{c\ mid}$ and $K_{c\ end}$ are listed in Tables 7–10 for various tree crops and plant densities. Usually there is close similarity in K_c within the same crop group. Full information regarding the crops and the tabulated values can be found in Pereira et al. (2020, 2021a, 2021b, 2021c) and Rallo et al. (2021).

Landscape water requirements

Landscapes are big water users in urban and residential areas, creating the need for more rational and efficient water use as competition for limited water resources increases (Pereira et al., 2009). Procedures to estimate water use and set objectives regarding irrigation of these surfaces have recently been adapted from those already developed for agricultural crops. These procedures take into account the fact that landscapes deviate from agricultural crops in several aspects:

- landscapes most often do not have the dimensions needed for the internal boundary layer above the vegetation to be in equilibrium with the underlying surface, as is implied in the Penman-Monteith equation. Therefore, compensating adjustments are necessary to account for effects of local surroundings on plant water use;
- landscapes often comprise mixtures of types of vegetation, with different heights, structure and resistance to water stress, and have non-uniform spatial configurations;
- contrary to agricultural crops, where irrigation aims at maximizing biomass production /marketable yield, landscape irrigation aims at maintaining an appealing appearance. Many plants species used in landscaping can tolerate some degree of water stress without impacts on health, longevity and appearance and thus deficit irrigation is often adopted, contrary to agricultural crops.

Table 7 Single (time-averaged) crop coefficients, K_c , basal crop coefficient, K_{cb} , for non-stressed well-managed vine, fruit crops, berries and hop in subhumid climates ($RH_{min} \approx 45\%$, $u_2 \approx 2$ m/s) for use with the FAO Penman-Monteith ET_o , and indicative mean maximum plant heights, h , maximum root depths, $z_{r\max}$, and depletion fractions for no stress, p .

Crop	$K_{c\ ini}^a$	$K_{c\ mid}$	$K_{c\ end}$	$K_{cb\ ini}$	$K_{cb\ mid}$	$K_{cb\ end}$	Maximum Crop Height (h, m)	Maximum Root Depth ^b ($z_{r\max}$, m)	Depletion Fraction ^c (p)
Grapes: table or raisin	0.30			0.20				1.0–2.0	0.35
Young ($f_c = 0.20$ –0.35)		0.60	0.55		0.55	0.45	1.0–1.8		
Overhead trellis ($f_c = 0.90$ –0.95)		0.95	0.70		0.90	0.65	2.0		
Horizontal trellis ($f_c = 0.60$ –0.70)		0.90	0.80		0.85	0.70	1.7–1.8		
T trellis ($f_c = 0.40$ –0.50)		0.80	0.60		0.75	0.55	2.0		
Y shape ($f_c = 0.60$ –0.90)		0.70	0.55		0.65	0.50	2.2		
Grapes: wine	0.30			0.20				1.0–2.0	0.45
Young ($f_c = 0.15$ –0.30)		0.45	0.40		0.40	0.30 ³	1.0–1.5		
Pergola ($f_c = 0.50$ –0.60)		0.65	0.50		0.60	0.45	2.0		
VSP ($f_c = 0.25$ –0.45)		0.70	0.45		0.465	0.40	1.5–2.0		
Guyot ($f_c = 0.15$ –0.50)		0.50	0.40		0.45	0.35	1.5–2.0		
Kiwi ($f_c = 0.80$ –0.90)	0.30	0.95	0.90	0.20	0.90	0.80	2.0	1.0–2.0	0.45
Highbush blueberries	0.30			0.20				0.4–0.7	0.50
Young ($f_c = 0.20$ –0.30)		0.40	0.30		0.35	0.20	<1.0		
Low ($f_c = 0.30$ –0.50)		0.50	0.35		0.45	0.25	1.5–1.8		
Medium ($f_c = 0.50$ –0.70)		0.80	0.35		0.75	0.30	1.5–1.8		
High ($f_c > 0.70$)		0.95	0.40		0.90	0.35	1.8–2.0		
Hops ($f_c = 0.10$ –0.20)	0.30	1.00	0.85	0.15	0.95	0.80	5–6	0.80–1.00	0.50

Allen et al. (1998); Allen and Pereira (2009); Rallo et al. (2021); Pereira et al. (2021c).

^aThe $K_{c\ ini}$ in the Table are only indicative. The procedure relative to Eq. (11) and (12) should be used.

^bThe larger values for $z_{r\max}$ correspond to rainfed conditions.

^cThe tabulated value for p must be corrected when $ET_c \neq 5$ mm d⁻¹ using $p = p_{table} + 0.04 (5 - ET_c)$.

Table 8 Single (time-averaged) crop coefficients, K_c , basal crop coefficient, K_{cb} , for non-stressed well-managed evergreen fruit tree crops in subhumid climates ($RH_{min} \approx 45\%$, $u_2 \approx 2$ m/s) for use with the FAO Penman-Monteith ET_o , and indicative mean maximum plant heights, h , maximum root depths, $z_{r\max}$, and depletion fractions for no stress, p .

Crop	$K_{c\ ini}^a$	$K_{c\ mid}$	$K_{c\ end}$	$K_{cb\ ini}$	$K_{cb\ mid}$	$K_{cb\ end}$	Maximum Crop Height (h, m)	Maximum Root Depth ^b ($z_{r\max}$, m)	Depletion Fraction ^c (p)
Citrus									0.50
Young ($f_c < 0.25$)	0.55	0.40	0.60	0.45	0.35	0.35	1.0–1.5	1.2–1.5	
Low density ($f_c = 0.25$ –0.40)	0.70	0.55	0.65	0.60	0.50	0.50	2.3–4.5	1.1–1.5	
Medium density, short ($f_c = 0.40$ –0.65)	0.80	0.60	0.70	0.60	0.55	0.55	<3.5	0.8–1.1	
Medium density, tall ($f_c = 0.40$ –0.65)	0.95	0.65	0.75	0.85	0.60	0.60	>3.5	0.8–1.1	
High density, short ($f_c \geq 0.65$)	0.95	0.70	0.80	0.85	0.65	0.65	2.5–3.5	0.8–1.1	
High density, tall ($f_c > 0.65$)	0.95	0.90	0.95	0.85	0.85	0.85	4.0–4.5	0.8–1.1	
Olives Young ($f_c = 0.15$ –0.30)	0.80	0.25	0.60	0.15	0.20	0.15	1.5–2.0		0.40
Traditional, low density ($f_c = 0.15$ –0.20)	0.80	0.30	0.60	0.30	0.25	0.20	2.5–3.0		0.65
Traditional, medium density ($f_c = 0.20$ –0.30)	0.85	0.45	0.65	0.30	0.40	0.30	3.0–3.5		0.65
Intensive (hedge prune ($f_c = 0.30$ –0.40))	0.80	0.45	0.70	0.30	0.40	0.35	2.5–3.5		0.45
Super intensive, medium density ($f_c = 0.25$ –0.35)	0.85	0.50	0.70	0.30	0.45	0.30	3.0–3.5		0.40
Super intensive high density ($f_c = 0.35$ –0.45)	0.85	0.55	0.75	0.30	0.50	0.40	3.0–4.0		0.40

Allen et al. (1998); Allen and Pereira (2009); Rallo et al. (2021); Pereira et al. (2021c).

^aThe $K_{c\ ini}$ in the Table are only indicative. The procedure relative to Eq. (11) and (12) should be used.

^bThe larger values for $z_{r\max}$ correspond to rainfed conditions.

^cThe tabulated value for p must be corrected when $ET_c \neq 5$ mm d⁻¹ using $p = p_{table} + 0.04 (5 - ET_c)$.

Table 9 Single (time-averaged) crop coefficients, K_c , basal crop coefficient, K_{cb} , for non-stressed well-managed temperate climate deciduous fruit tree crops in subhumid climates ($RH_{\min} \approx 45\%$, $u_2 \approx 2$ m/s) for use with the FAO Penman-Monteith ET_o , and indicative mean maximum plant heights, h , maximum root depths, $z_{r \max}$, and depletion fractions for no stress, p .

Crop	K _c <i>ini</i> ^a	K _c <i>mid</i>	K _c <i>end</i>	K _{cb} <i>ini</i>	K _{cb} <i>mid</i>	K _{cb} <i>end</i>	Maximum Crop Height (h, m)	Maximum Root Depth ^b (z _{r max} , m)	Depletion Fraction ^c (p)
a. Stone fruit trees									
<i>Apricots, cherry, plum</i>									
Young (<i>f_c</i> = 0.15–0.30)	0.35	0.55	0.40	0.15	0.50	0.30	1.5–2.0	1.0–2.0	0.50
Low density (<i>f_c</i> = 0.30–0.40)	0.40	0.60	0.45	0.20	0.55	0.35	2.0–3.0		
Med. Density (<i>f_c</i> = 0.40–0.50)	0.45	0.80	0.55	0.25	0.75	0.50	2.5–3.5		
High density (<i>f_c</i> = 0.50–0.60)	0.50	0.90	0.60	0.30	0.85	0.55	2.5–4.0		
Very high density (<i>f_c</i> > 0.60)	0.50	1.05	0.70	0.30	1.00	0.65	2.5–5.0		
<i>Peach</i>									
Young (<i>f_c</i> = 0.15–0.30)	0.40	0.55	0.45	0.20	0.50	0.35	1.5–2.5	1.0–2.0	0.50
Low density (<i>f_c</i> = 0.25–0.40)	0.40	0.65	0.50	0.20	0.60	0.40	2.5–3.0		
Med. Density (<i>f_c</i> = 0.40–0.60)	0.45	0.90	0.65	0.25	0.85	0.60	2.5–3.5		
High density (<i>f_c</i> > 0.60)	0.50	1.05	0.70	0.30	1.00	0.65	3.0–5.0		
b. Pome fruit trees									
<i>Apples</i>									
Young (<i>f_c</i> < 0.25)	0.40	0.40	0.35	0.25	0.35	0.25	1.0–1.8	1.0–2.0	0.50
Low density (<i>f_c</i> = 0.25–0.30)	0.50	0.50	0.40	0.30	0.45	0.30	3.0–3.6		
Med density (<i>f_c</i> = 0.30–0.40)	0.50	0.70	0.55	0.30	0.65	0.45	3.0–3.6		
High density (<i>f_c</i> = 0.40–0.50)	0.50	0.90	0.65	0.30	0.85	0.45	3.0–3.6		
Very high density (<i>f_c</i> > 0.50)	0.50	1.00	0.65	0.30	0.95	0.69	3.0–4.5		
<i>Pears</i>									
Young (<i>f_c</i> < 0.25)	0.40	0.40	0.30	0.25	0.35	0.20	1.0–1.5	1.0–2.0	0.50
Low density (<i>f_c</i> = 0.25–0.35)	0.50	0.55	0.45	0.30	0.50	0.35	2.0–3.0		
Med density (<i>f_c</i> = 0.35–0.40)	0.50	0.85	0.65	0.30	0.80	0.60	2.0–3.0		
High density (<i>f_c</i> = 0.40–0.50)	0.50	1.00	0.75	0.30	0.95	0.70	2.0–3.0		
Very high density (<i>f_c</i> > 0.50)	0.50	1.10	0.85	0.30	1.05	0.80	3.3–3.6		
c. Nut trees									
<i>Almonds</i>									
Young (<i>f_c</i> = 0.15–0.30)	0.35	0.40	0.35	0.15	0.20	0.25	1.0–2.0	1.0–2.0	0.40
Low density (<i>f_c</i> = 0.30–0.40)	0.40	0.45	0.40	0.20	0.40	0.30	2.0–4.0		
Medium density (<i>f_c</i> = 0.40–0.50)	0.45	0.65	0.50	0.20	0.60	0.40	3.0–5.0		
High density (<i>f_c</i> = 0.50–0.60)	0.50	0.90	0.65	0.20	0.85	0.60	4.0–5.0		
Very high density (<i>f_c</i> > 0.60)	0.50	1.05	0.75	0.20	1.00	0.70	4.0–5.5		
<i>Hazelnut</i>									
Young (<i>f_c</i> = 0.15–0.30)	0.30	0.45	0.40	0.20	0.40	0.20	1.0–2.0	1.0–2.0	0.40
Low density (<i>f_c</i> = 0.30–0.40)	0.35	0.75	0.55	0.25	0.70	0.45	4.0–5.0		
High density (<i>f_c</i> = 0.45–0.60)	0.40	0.95	0.65	0.30	0.90	0.55	2.5–3.0		

Table 9 (Continued)

<i>Crop</i>	K_c <i>ini</i>	K_c <i>mid</i>	K_c <i>end</i>	K_{cb} <i>ini</i>	K_{cb} <i>mid</i>	K_{cb} <i>end</i>	<i>Maximum Crop Height</i> (h, m)	<i>Maximum Root Depth</i> ($z_{r\ max}$, m)	<i>Depletion Fraction</i> (p)
Pecan								1.0–2.5	0.40
Young ($f_c = 0.25$ – 0.30)	0.35	0.55	0.50	0.25	0.50	0.35	1.0–2.0		
Low density ($f_c = 0.30$ – 0.60)	0.45	0.75	0.55	0.35	0.70	0.45	2.0–4.0		
Medium density ($f_c = 0.60$ – 0.70)	0.50	0.85	0.60	0.40	0.80	0.55	4.0–7.0		
High density ($f_c > 0.70$)	0.50	1.00	0.65	0.40	0.95	0.60	7.0–10.0		
Pistachios								1.0–1.5	0.40
Young, low density ($f_c = 0.15$ – 0.30)	0.30	0.45	0.35	0.20	0.40	0.25	1.5–2.0		
Med. density ($f_c = 0.30$ – 0.5)	0.35	0.80	0.55	0.25	0.75	0.50	2.0–3.0		
High density ($f_c > 0.50$)	0.40	0.95	0.65	0.30	0.90	0.60	3.0–4.0		
Walnut orchard								1.7–2.4	0.50
Young/low density ($f_c < 0.30$)	0.35	0.50	0.45	0.25	0.45	0.25	1.0–2.0		
Med. density ($f_c = 0.45$ – 0.75)	0.45	0.85	0.50	0.35	0.80	0.40	1.5–3.5		
High density ($f_c = 0.75$ – 0.85)	0.50	0.95	0.55	0.40	0.90	0.50	3.5–7.0		
Very high density ($f_c > 0.85$)	0.50	1.05	0.60	0.40	1.00	0.55	5.0–8.0		
d. Other fruit trees									
Fig tree								0.60–1.0	0.50
Young, low ($f_c < 0.30$)	0.40	0.45	0.35	0.15	0.40	0.25	1.0–1.5		
Medium ($f_c = 0.30$ – 0.50)	0.45	0.65	0.45	0.20	0.60	0.40	2.0–3.0		
Persimmon								0.60–1.2	0.50
Young, low ($f_c = 0.15$ – 0.25)	0.40	0.50	0.40	0.15	0.45	0.30	1.0–2.0		
Medium ($f_c = 0.30$ – 0.50)	0.45	0.95	0.75	0.20	0.90	0.70	2.5–3.5		
Pomegranate								0.50–1.5	0.40
Young ($f_c = 0.15$ – 0.25)	0.35	0.35	0.30	0.15	0.30	0.20	1.0–2.0		
Low ($f_c = 0.25$ – 0.35)	0.35	0.50	0.40	0.20	0.45	0.30	2.5–3.5		
Medium ($f_c = 0.35$ – 0.45)	0.40	0.60	0.45	0.20	0.55	0.40	2.5–3.5		
High ($f_c > 0.45$)	0.40	0.85	0.60	0.20	0.80	0.55	2.5–3.5		

(Allen et al., 1998; Allen and Pereira, 2009; Rallo et al., 2021; Pereira et al., 2021c).

^{a1} The K_c *ini* in the Table are only indicative and valid for typical irrigation management and soil wetting frequency.

^bThe larger values for $z_{r\ max}$ correspond to rainfed conditions.

^cThe tabulated value for p must be corrected when $ET_c \neq 5\text{ mm d}^{-1}$ using $p = p_{\text{table}} + 0.04 (5 - ET_c)$.

For landscapes, distinction should be made between target ET (ET_L) and actual ET ($ET_{L\ act}$) values. ET_L is based on minimum ET levels, relative to climate, necessary to sustain a healthy, attractive landscape while $ET_{L\ act}$ is based on landscape type and on actual water availability (Allen et al., 2007). When the landscape receives more water than that required by the target that includes intentional stress, then actual ET values may exceed target ET values. Under these conditions, landscape vegetation is able to exploit the additional available water. In contrast, when actual water stress levels are higher than the targeted ones the actual ET is lower than target ET values.

The water use from a landscape (ET_L), using an analogy with agricultural crops, can be estimated by

$$ET_L = K_L ET_o \quad (42)$$

where K_L is the landscape coefficient. An innovative approach to K_L was first developed by Costello et al. (2000), labeled as the WUCOLS (Water Use Classification of Landscape Species) method, and later modified by Romeo and Dukes (2010) for the LIMP (Landscape Irrigation Management Program), where the K_L is decoupled into four other coefficients that represent the factors that impact ET from landscapes—vegetation type, density, local climate, and soil water management. A similar approach to K_L was adopted by the Irrigation Association (IA) (2011). The multi-component decoupling method facilitates its application to the very different types of vegetation and environments that can be found in landscapes.

In this approach, the K_L has the form.

$$K_L = K_v K_d K_{mc} K_{sm} \quad (43)$$

Table 10 Single (time-averaged) crop coefficients, K_c , basal crop coefficient, K_{cb} , for non-stressed well-managed tropical and sub-tropical fruit tree crops in subhumid climates ($RH_{\min} \approx 45\%$, $u_2 \approx 2$ m/s) for use with the FAO Penman-Monteith ET_o , and indicative mean maximum plant heights, h , maximum root depths, $z_{r \max}$, and depletion fractions for no stress, p .^a

Crop	$K_{c \text{ ini}}^a$	$K_{c \text{ mid}}$	$K_{c \text{ end}}$	$K_{cb \text{ ini}}$	$K_{cb \text{ mid}}$	$K_{cb \text{ end}}$	Maximum Crop Height (h , m)	Maximum Root Depth ^b ($z_{r \max}$, m)	Depletion Fraction ^c p
Avocado								0.5–1.0	0.50
Young, low ($f_c = 0.10\text{--}0.20$)	0.40	0.65	0.60	0.25	0.55	0.45	1.0–2.0		
Medium ($f_c = 0.30\text{--}0.50$)	0.50	0.80	0.70	0.30	0.75	0.60	2.0–3.0		
High ($f_c > 0.50$)	0.50	1.00	0.85	0.30	0.95	0.80	3.0–4.0		
Banana: 1st year ($f_c = 0.80\text{--}0.95$)	0.50	1.05	0.95	0.15	1.00	0.90	2.0–2.5	0.5–0.9	0.35
2nd year ($f_c = 0.75\text{--}0.85$)	1.00	1.15	1.10	0.60	1.05	1.00	2.0–2.5	0.5–0.9	0.35
Cherimoya ($f_c = 0.50\text{--}0.65$)	0.50	0.65	0.55	0.30	0.60	0.45	2.5–4.0	1.0–2.0	0.50
Coconut ($f_c = 0.70\text{--}0.80$)	0.90	0.95	0.90	0.85	0.90	0.86	7.0–15.0	0.7–1.1	0.65
Coffee								0.9–1.5	0.40
Young, low density ($f_c = 0.15\text{--}0.25$)	0.70	0.75	0.70	0.60	0.70	0.65	1.0–2.0		
High density ($f_c = 0.40\text{--}0.60$)	0.90	0.95	0.95	0.80	0.90	0.90	2.5–3.5		
Very high density ($f_c > 0.60$)	0.90	1.00	1.00	0.80	0.95	0.95	2.0–3.0		
Date palm								1.5–2.5	0.50
Young, low density ($f_c = 0.15\text{--}0.25$)	0.50	0.50	0.55	0.45	0.45	0.45	8.0–10.0		
High density ($f_c = 0.30\text{--}0.60$)	0.80	0.70	0.65	0.70	0.65	0.60	8.0–10.0		
Very high density ($f_c > 0.60$)	0.90	0.85	0.75	0.85	0.80	0.70	10.0–15.0		
Guava ($f_c > 0.70$)	0.50	0.80	0.60	0.30	0.75	0.55	1.8–3.0	0.60–1.0	0.40
Mango								1.1–5.0	0.40
Young, low ($f_c = 0.10\text{--}0.20$)	0.30	0.40	0.35	0.20	0.35	0.25	1.0–2.0		
Medium ($f_c = 0.30\text{--}0.70$)	0.35	0.75	0.60	0.25	0.70	0.55	2.5–3.0		
High ($f_c > 0.70$)	0.35	0.90	0.70	0.25	0.85	0.65	5.0–5.5		
Papaya ($f_c > 0.70$)	0.45	1.00	0.75	0.25	0.95	0.70	1.5–3.0	0.5–2.0	0.40

^aThe $K_{c \text{ ini}}$ in the Table are only indicative and valid for typical irrigation management and soil wetting frequency.

^bThe larger values for $z_{r \max}$ correspond to rainfed conditions.

^cThe tabulated value for p must be corrected when $ET_c \neq 5$ mm d⁻¹ using $p = p_{\text{table}} + 0.04 (5 - ET_c)$.

From: Allen RG, Pereira LS, Raes D and Smith M (1998) Crop evapotranspiration. In: *Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56. Rome: FAO; Rallo G, Paço TA, Puig-Sirera A, Paredes P, Massai R, Provenzano G, Pereira LS (2021) Updated single and dual crop coefficients for tree and vine fruit crops. *Agricultural Water Management* **250**: 106645; Pereira LS, Paredes P, Melton F, Johnson L, Mota M and Wang T (2021c) Prediction of crop coefficients from fraction of ground cover and height: Practical application to vegetable, field and fruit crops with focus on parameterization. *Agricultural Water Management* **252**: 106663.

where K_v is a vegetation species factor (0.7–1.2), K_d is a vegetation density factor (0–1.0), K_{mc} is a microclimate factor (0.5–1.5), and K_{sm} is a managed water stress factor (0–1.0). Despite the apparent complexity of this approach, it can be easily implemented as each coefficient can be estimated separate from the others based on visual observation of the landscape, coupled with management objectives. Upper bounds of 1.00 for the modifying coefficients K_d and K_{sm} simplify their estimation and reduce uncertainties.

Though the method has been developed for landscapes, it can also be used to estimate ET of riparian strips and land corridors, or of small farms, undulating terrain, and multiple microclimates within an agricultural area.

The vegetation coefficient K_v is the maximum K_L expected for a specific single or mixture of plant species under full or nearly full ground cover, full soil water supply and where no microclimate adjustments are required and is estimated as:

$$K_v = \frac{ET_v}{ET_{o \text{ mc}}} \quad (44)$$

where $ET_{o \text{ mc}}$ is a local reference evapotranspiration if it were possible to measure weather data over a well-watered grass surface to determine ET_o in the local climate, and ET_v is the ET of the specific single or mixture of plant species with full ground cover (>70–80% of the ground covered or shaded by vegetation), where there are no microclimate adjustments required and under full soil water supply. The K_v is thus the equivalent of the K_c for agricultural crops, and Tables 3, 4, 7–10 can assist in choosing an adequate value given the specific plant structure and height. Also, many types of landscape vegetation tend to exhibit similar values for K_v due to similarities in height, total leaf area, stomatal response, and energy absorption. Therefore, condensed tables of typical values for general species types have been developed to provide general estimates for K_v (Table 11). Turfgrass coefficients are generally less than 1.0, contrary to what could be expected, since these surfaces are normally kept shorter than the 12 cm of the reference grass surface used to estimate ET_o .

The vegetation coefficient is not an equivalent to the plant or species coefficient (K_{sp}) considered by the WUCOLS model, as the K_{sp} is based on the water needs to maintain health and appearance, thus including a stress tolerance factor. Therefore, when using tabulated values, it is important to confirm the model they are to be used with.

Table 11 General vegetation factors (K_v) for general plant types for coverage (shading) of the ground >70–80% and full water supply.

Type of vegetation	K_v
Trees	1.15
Shrubs, desert species	0.70
Shrubs, non-desert species	0.80
Groundcover	1.00
Annuals (flowers)	0.90
Mixture of trees, shrubs, and groundcover	1.20
Turfgrass (cool season and warm season)	0.90

Allen et al. (2020).

Landscapes can vary considerably in terms of vegetation density, as there can be great variation in plant spacing, size, and maturity. The *density coefficient*, K_d , already described above (Eq. 37), can be used to correct the density impacts on ET based on fraction of ground cover, height and stomatal control of the specific vegetation.

Reference ET_o , which represents weather-based effects on ET , constitutes a regional estimate of reference evapotranspiration based on measured weather data from an open, vegetated area. ET from small sized patches of vegetation ($ET_{o\ mc}$) is influenced by the local microclimate, created by surrounding dry areas, buildings or pavements. The *microclimate coefficient* (K_{mc}) is defined as

$$K_{mc} = \frac{ET_{o\ mc}}{ET_o} \quad (45)$$

and accounts for impacts on ET by exposure to sun (e.g., north oriented versus south oriented areas), external shading by buildings, hot and cool areas, reflected and emitted radiation from structures, wind blockage by buildings and other vegetation, and transfer of sensible heat from low ET surroundings, such as structures and paved areas, that make $ET_{o\ mc}$ depart from the ET_o that is observed in large, well-watered, exposed surfaces.

Values for K_{mc} are listed in Table 12 for general classes of vegetation. The “high” category ($K_{mc} > 1.00$) reflects microclimate conditions that increase evapotranspiration, such as planting in direct sunlight near paved or other non-vegetated surfaces, near reflective or heat-absorbing surfaces such as windows or buildings, or in exposed, windy conditions. The “low” category ($K_{mc} < 1.00$) represents environments where the plantings are shaded, away from dry, hot surfaces and buildings or shielded from wind. The average or medium category ($K_{mc} = 1.00$) represents environments similar to those that characterize reference conditions for ET_o measurement, where buildings, pavements, shading and reflective surfaces do not influence the ET by the landscape (e.g., parks). Values for K_{mc} can be interpolated between the high, average, and low classes and should be selected for each sector of a landscape. It is advisable to perform local measurements to confirm or to derive local K_{mc} values.

Typically, the objective of landscape irrigation is to promote appearance rather than biomass production, unlike agricultural crops. Therefore, irrigation of landscapes is normally managed in order to apply less water than what is needed for full ET_v , thus reducing ET and introducing an intentional water stress factor, K_{sm} :

$$K_{sm} = \frac{ET}{ET_v} \quad (46)$$

where ET is actual evapotranspiration by the vegetated surface and $K_{sm} < 1$. The magnitude of the stress factor depends on the physiological requirements and morphological characteristics of the plants and the desired or minimum acceptable appearance (Table 13).

Table 13 helps making the distinction between water needs and stress tolerance. While all types of turfgrass have the same water needs, as noted by the same K_v (Table 11), given their similarity in height and structure, stress tolerance of cool season turfgrass is lower than that of warm season turfgrass. As a consequence, the latter can be maintained with less irrigation water and is thus recommended in areas with higher ET_o and/or limited water availability.

Computation of K_{sm} can be made according to Eqs. (31) and (32).

Besides savings of water, deficit irrigation can bring savings in mowing or pruning needs due to the reduced growth caused by water stress.

Table 12 Microclimate factor (K_{mc}) for landscape plant types.

Vegetation	K_{cm}		
	High (harsh environment)	Average (reference condition)	Low (protected environment)
Trees	1.4	1.0	0.5
Shrubs	1.3	1.0	0.5
Groundcover, flowers	1.2	1.0	0.5
Mixture of trees, shrubs, and groundcover	1.4	1.0	0.5
Turfgrass	1.2	1.0	0.8

Allen et al. (2020).

Table 13 Managed stress factor (K_{sm}) for landscape plant types and soil water depletion fraction (p) for no transpiration-reducing stress.

Vegetation category	K_{sm}			
	High stress	Average managed stress	Low stress	Depletion fraction (p)
Trees	0.4	0.6	0.8	0.6
Shrubs, desert species	0.3	0.4	0.6	0.6
Shrubs, non-desert species	0.4	0.6	0.8	0.6
Groundcover	0.3	0.5	0.8	0.5
Annuals (flowers)	0.5	0.7	0.8	0.4
Mixture of trees, shrubs, and groundcover	0.4	0.6	0.8	0.6
Cool-season turfgrass	0.7	0.8	0.9	0.4
Warm-season turfgrass	0.6	0.7	0.8	0.5

Allen et al. (2020).

Irrigation water requirements

The soil water balance is calculated for the effective rooting depth as:

$$\theta_i = \theta_{i-1} + \frac{(P_i - RO_i) + I_{wi} - ET_{ci} - DP_i + GW_i}{1000 z_{ri}} \quad (47)$$

where, in addition to the symbols used before, DP_i represents deep percolation [mm], GW_i is groundwater contribution [mm], and z_{ri} is the rooting depth [m], all referred to day i . DP is often estimated as $DP_i = 0$ when $\theta_i \leq \theta_{FC}$ and $DP_i = 1000 (\theta_i - \theta_{FC}) z_{ri}$ otherwise. GW is estimated from soil hydraulic properties and the water table depth. z_{ri} can be predicted assuming a linear variation from planting to maximum rooting. Maximum root depths for most common crops are presented in Tables 3, 4, 7–10.

The latest date for scheduling irrigation to avoid water stress is when $\theta_i = \theta_p$ (Eq. 32). However, irrigation is often scheduled when the “management-allowed depletion,” MAD , is attained. Generally, $MAD < p$ when there is risk aversion or uncertainty, and $MAD > p$ when plant water stress is intentional. Then

$$\theta_i = \theta_{MAD} = (1 - MAD) (\theta_{FC} - \theta_{WP}) + \theta_{WP} \quad (48)$$

In the case of landscapes where deficit irrigation is to be implemented ($K_{sm} < 1.0$), Table 14 can assist in the selection of a target MAD prior to irrigation that produces the desired K_{sm} on average.

The net irrigation depth to be applied will be

$$I_{wi} = 1000 z_{ri} (\theta_{FC} - \theta_i), \quad (49)$$

which summed for the entire season leads to the irrigation water requirement (IWR):

$$IWR = \frac{ET_c - P_e - GW - \Delta S}{1 - LR} \quad (50)$$

where P_e is the effective precipitation (gross precipitation less all runoff and deep percolation), GW is groundwater contribution or capillary rise, ΔS is the change in soil water storage in the root zone between planting and harvesting, and LR is the leaching requirement (the percentage of irrigation water that must pass through the root zone to keep the salinity of the soil below a specified value). The soil water balance is currently computed through crop-water simulation models, which allow the selection of best irrigation scheduling alternatives, e.g., the SIMDualKc model that adopts the dual crop coefficient approach (Rosa et al., 2012).

Table 14 Management allowed depletion (*MAD*) fraction to produce the stated managed stress factor (K_{sm}) given the depletion fraction for no *ET*-reducing stress (p) and assuming complete refilling of the root zone with each irrigation (*MAD* is expressed as a decimal).

Average K_{sm}	$p = 0.3$	$p = 0.4$	$p = 0.5$	$p = 0.6$	$p = 0.7$
1.00	0.30	0.40	0.50	0.60	0.70
0.95	0.47	0.57	0.66	0.75	0.86
0.90	0.55	0.65	0.73	0.81	0.88
0.85	0.62	0.71	0.79	0.86	—
0.80	0.68	0.76	0.83	0.89	—
0.75	0.74	0.80	0.87	—	—
0.70	0.78	0.84	0.89	—	—
0.65	0.82	0.88	—	—	—
0.55	0.90	—	—	—	—
0.50	—	—	—	—	—

—, indicates that the *MAD* value approaches or exceeds 1, so the soil water content approaches or exceeds the permanent wilting point and the vegetation is in danger of dormancy or death.

Allen et al. (2020).

Table 15 Indicative values of beneficial consumption fraction (*BWUF*).

System	BWUF (%)
• Surface irrigation, precision levelling	
- Furrow	65–85
- Border	70–85
- Basin	70–90
• Surface, traditional	
- Furrow	40–70
- Border	45–70
- Basin	45–70
- Basin, rice fields	25–50
• Sprinkler	
- Solid set	65–85
- Hand-move lateral	65–80
- Side-roll wheel move	65–80
- Traveler sprinkler	55–70
- Lateral move systems, center pivot	65–85
• Micro-irrigation	
- Trickle, ≥ 3 emitters per plant	85–95
- Trickle, < 3 emitters per plant	80–90
- Bubblers and sprayers	85–95
- Line source emitters	70–90

Pereira and Trout (1999).

The gross irrigation water requirement is computed as

$$GIWR = \frac{IWR}{BWUF} \quad (51)$$

where *BWUF* is the beneficial water use fraction, which includes ET_c and *LR*, and is total water applied. Indicative values of *BWUF* are presented in Table 15.

Limitations

Reliable estimates of crop or landscape water needs are necessary in many situations, such as hydrologic studies, watershed management, and irrigation planning and management. In recent decades, the approach proposed by FAO (Allen et al., 1998) to estimate crop water needs has been consolidated, given the sound physical basis and appropriate parameterization of the Penman-Monteith equation for the calculation of the reference evapotranspiration, together with the availability of tabulated, transferable values of single time averaged or dual crop coefficients for a wide variety of crops.

Despite the robustness of the method, there are still many situations where it is not properly used. A common error is using other equations rather than the $PM-ET_o$ equation to calculate reference evapotranspiration. This obviously leads to or requires different crop coefficients from the correspondent tabulated values, when conducting studies to determine crop coefficients or calculate crop water needs. Studies using a different ET equation should therefore consider the ratio between ET_o equations to allow the conversion of the derived K_c values using those equations or to be able to use the tabulated K_c values that refer to the PM reference ET_o .

FAO 56 introduced the distinction between actual and standard crop coefficients, with the latter referring to crops cultivated in pristine conditions. However, cultivation practices are often far from pristine due to water stress, salinity stress or stresses produced by various insufficiencies of agronomic nature, which are often not recognized or even accepted by many researchers or users, and lead to an incorrect use of the tabulated K_c values.

Conclusions

The population increase and the current politics towards a “greener” agricultural sector demands greater food production on less farmed land. In addition, the climate change challenges, in particular the decrease in water availability and increased competition with other users, enhances the importance of minimizing water use in agriculture and promoting irrigation water savings while maintaining or increasing crop production. This may be attained with an adequate irrigation scheduling and adoption of irrigation systems with higher efficiency (Pereira et al., 2009).

Further research is required regarding the relationships between soil evaporation, transpiration, fraction of soil wetted or fraction of ground covered by the crop, namely in fruit crops and other partial-cover crops. The impacts of mulches on K_c and K_{cb} values constitute another area where further needs of research have been identified. Finally, more research is welcomed on real-time irrigation decision support systems with particular focus on the use of soil water balance models that allow the integration of information based on remote sensing and sensors data.

References

- Allen RG and Pereira LS (2009) Estimating crop coefficients from fraction of ground cover and height. *Irrigation Science* 28: 17–34.
- Allen RG, Prueger JH, and Hill RW (1992) Evapotranspiration from isolated stands of hydrophytes: Cattail and bulrush. *Transactions of ASAE* 35(4): 1191–1198.
- Allen RG, Pereira LS, Raes D, and Smith M (1998) Crop evapotranspiration. In: *Guidelines for Computing Crop Water Requirements*. Rome: FAO. FAO Irrigation and Drainage Paper 56.
- Allen RG, Wright JL, Pruitt WO, Pereira LS, and Jensen ME (2007) Water Requirements. In: Hoffman GF, Evans RG, Jensen ME, Martin DL, and Elliot RL (eds.) *Design and Operation of Farm Irrigation Systems*, 2nd edn., pp. 208–288. St. Joseph, MI: ASABE.
- Allen RG, Dukes MD, Snyder RL, Kjølsgren R, and Kilic A (2020) A review of landscape water requirements using a multicomponent landscape coefficient. *Transactions of the ASABE* 63(6): 2039–2058.
- Anderson MC, Allen RG, Morse A, and Kustas WP (2012) Use of Landsat thermal imagery in monitoring evapotranspiration and managing water resources. *Remote Sensing of Environment* 122: 50–65.
- Costello LR, Matheny NP, and Clark JR (2000) A guide to estimating irrigation water needs of landscape plantings in California. In: *The Landscape Coefficient Method and WUCOLS III*. University of California Cooperative Extension, California Department of Water Resources.
- Irrigation Association (IA) (2011) *Irrigation*, 6th edn Falls Church, VA: Irrigation Association.
- Mobe NT, Dziki S, Zirebwa SF, Midgley SJE, von Loeper W, Mazvimavi D, Ntshidi Z, and Jovanovic NZ (2020) Estimating crop coefficients for apple orchards with varying canopy cover using measured data from twelve orchards in the Western Cape Province, South Africa. *Agricultural Water Management* 233: 106103.
- Paredes P, Pereira LS, Almorox J, and Darouich H (2020) Reference grass evapotranspiration with reduced data sets: Parameterization of the FAO Penman-Monteith temperature approach and the Hargreaves-Samani equation using local climatic variables. *Agricultural Water Management* 240: 106210.
- Paredes P, Trigo I, de Bruin H, Simões N, and Pereira LS (2021) Daily grass reference evapotranspiration with Meteosat Second Generation shortwave radiation and reference ET products. *Agricultural Water Management* 248: 106543.
- Pereira LS and Trout TJ (1999) Irrigation methods. In: van Lier HN, Pereira LS, and Steiner FR (eds.) *CIGR Handbook of Agricultural Engineering. Land and Water Engineering*, vol. I, pp. 297–379. St. Joseph, MI: ASAE.
- Pereira LS, Cordery I, and Iacovides I (2009) Coping with water scarcity. In: *Addressing the Challenges*. Dordrecht: Springer.
- Pereira LS, Paredes P, Melton F, Johnson L, Wang T, López-Urrea R, Cancela JJ, and Allen R (2020) Prediction of crop coefficients from fraction of ground cover and height. Background and validation using ground and remote sensing data. *Agricultural Water Management* 241: 106197.
- Pereira LS, Paredes P, Lopez-Urrea R, Hunsaker DJ, Mota M, and Mohammadi Shad Z (2021a) Standard single and basal crop coefficients for vegetable crops, an update of FAO56 crop water requirements approach. *Agricultural Water Management* 243: 106196.
- Pereira LS, Paredes P, Hunsaker DJ, López-Urrea R, and Mohammadi Shad Z (2021b) Standard single and basal crop coefficients for field crops. Updates and advances to the FAO56 crop water requirements method. *Agricultural Water Management* 243: 106466.
- Pereira LS, Paredes P, Melton F, Johnson L, Mota M, and Wang T (2021c) Prediction of crop coefficients from fraction of ground cover and height: Practical application to vegetable, field and fruit crops with focus on parameterization. *Agricultural Water Management* 252: 106663.
- Perrier A (1982) Land surface processes: vegetation. In: Eagleson PS (ed.) *Land Surface Processes in Atmospheric General Circulation Models*, pp. 395–448. Cambridge: Cambridge University Press.
- Pôças I, Calera A, Campos I, and Cunha M (2020) Remote sensing for estimating and mapping single and basal crop coefficients: A review on spectral vegetation indices approaches. *Agricultural Water Management* 233: 106081.
- Rallo G, Paço TA, Puig-Sirera A, Paredes P, Massai R, Provenzano G, and Pereira LS (2021) Updated single and dual crop coefficients for tree and vine fruit crops. *Agricultural Water Management* 250: 106645.
- Raziei T and Pereira LS (2013) Estimation of ET_o with Hargreaves-Samani and FAO-PM temperature methods for a wide range of climates in Iran. *Agricultural Water Management* 121: 1–18.
- Romeo CC and Dukes MD (2010) *Residential Benchmarks for Minimal Landscape Water Use*. Gainesville: University of Florida Water Institute.
- Rosa RD, Paredes P, Rodrigues GC, Alves I, Fernando RM, Pereira LS, and Allen RG (2012) Implementing the dual crop coefficient approach in interactive software. 1. Background and computational strategy. *Agricultural Water Management* 103: 8–24.

Water harvesting

Wolde Mekuria and Desalegn Tegegne, International Water Management Institute (IWMI), Addis Ababa, Ethiopia

© 2023 Elsevier Ltd. All rights reserved.

Definition	593
Introduction	594
History of water harvesting	594
Principles, concepts and components	595
Water harvesting and hydrological processes	595
Categories of water harvesting	597
Micro-catchment systems	598
External/Macro-catchment system	600
Floodwater harvesting	601
Rooftop harvesting	602
Groundwater harvesting	603
Environmental and socio-economic benefits of water harvesting and limitations	604
Conclusion	605
References	606

Abstract

Water harvesting is the intentional collection and concentration of rainwater and runoff to offset irrigation demands. Secondary benefits include decreased flood and erosion risk. Water harvesting techniques include micro- and macro-catchment systems, floodwater harvesting, and rooftop and groundwater harvesting. The techniques vary with catchment type and size, and the method of water storage. Micro-catchment water harvesting, for example, requires the development of small structures and targets increased water delivery and storage to the root zone whereas macro-catchment systems collect runoff water from large areas. The sustainability of water harvesting techniques at the local level are usually constrained by several factors such as labor, construction costs, loss of productive land, and maintenance, suggesting that multiple solutions are required to sustain the benefits of water harvesting techniques.

Key points

- Water harvesting is important in rainfed agroecosystems to offset moisture deficit, particularly in semiarid environments.
- Water harvesting can also decrease run-off on slopes and help mitigate flood risk.
- Water harvesting represents a good example of sustainable management of water resources that contribute to water and food security.
- Water harvesting involves diverse methods of collecting and managing floodwaters and surface runoff.
- About 3 million small reservoirs are in operation in semi-arid regions, supporting 15% of the world population.
- Water harvesting focuses on the efficient use of available water.
- Water harvesting technologies reduce complete crop failure by 7–20%.
- Water harvesting technologies can bridge up to 40% of the yield gap of attributable to water deficit.
- Both socio-economic and biophysical drivers constrain water harvesting.
- Multiple solutions are required to implement and sustain water harvesting techniques.

Definition

Water harvesting has been defined and classified in several ways. A general definition of water harvesting is the process of concentrating precipitation through runoff and storing it for beneficial use (Oweis et al., 2012). However, a broader definition is the collection and concentration of rainwater and runoff and its productive use for irrigation of annual crops, pastures, and trees, for domestic and livestock consumption and for ground water recharge (Prinz, 2011). Although there are slight variations in definitions, the term ‘water harvesting’ generally refers to the collection of rainstorm-generated runoff from an area (a catchment) to provide water for human, animal, or crop use. A secondary benefit of water harvesting on slopes is decreased runoff induced erosion. The water thus collected can either be utilized immediately, as for irrigation, or be stored in aboveground ponds or in subsurface reservoirs, such as cisterns or shallow aquifers, for subsequent utilization. As such, water harvesting is an ancient practice that has enabled some societies to survive in semiarid and arid areas where other sources of fresh water (e.g., rivers, lakes, or aquifers) are scant or unavailable.

Introduction

Water scarcity and quality issues caused by climate change and human actions are increasingly threatening food security and nutrition through their impacts on food systems (Young et al., 2021). This mainly presents a higher risk to the livelihoods of rural people by reducing crop and livestock yields (Payus et al., 2020). The FAO's State of Food and Agriculture report (FAO, 2020) indicated that 3.2 billion people live in agricultural areas with high to very high-water shortages or scarcity, of whom 1.2 billion people—roughly one-sixth of the world's population—live in severely water-constrained agricultural areas. This same report further elaborated that water scarcity is a particularly serious issue in Northern Africa and Western Asia, where per capita freshwater has declined by more than 30% and where the average annual volume of water per person barely reaches 1000 m³, which is conventionally considered the threshold for severe water scarcity. In terms of agricultural land affected, 128 million hectares of rainfed cropland (11%) and 656 million hectares of pastureland (14%) face frequent droughts, while 171 million hectares of irrigated cropland (60%) are subject to high or very high-water stress.

Either a change in demand and user practices, or alternate water resources, are required to sustain the agricultural sector. In this line, improved water management strategies, which includes a range of options is a crucial component to reduce water risks and attain potential yields in agriculture for improved food security and nutrition (FAO, 2020). The inherent challenges are to manage and adapt to weather variability, and to use water from rainfall more productively. According to Duveskog et al. (2003), there are two broad strategies for increasing yields in rainfed agriculture: (a) collecting or harvesting more water and infiltrating it into the root-zone; and (b) conserving water by reducing root-zone evaporation and drainage losses.

Water harvesting technology is especially relevant to the semi-arid and arid areas where the problems of environmental degradation, drought and population pressures are most evident (Piemontese et al., 2020). It is an important component of the package of remedies for these problem zones, and interest in water harvesting has been renewed in recent decades due to the growing demand for water for agriculture and urban development caused by higher population pressure and climate change. This interest has also led to increases in the understanding, implementation and management of water harvesting systems (Oweis et al., 2012). Other applications of water harvesting include the collection of run-off water to mitigate erosion (LaFevor and Ramos-Scharron, 2021) or flood risk (Ward et al., 2021). This chapter presents the history; principles, concepts, and components of water harvesting; the relationship between water harvesting and the hydrological processes including soil erosion and flood prevention; categories of water harvesting; and benefits and constraints of water harvesting.

History of water harvesting

Water harvesting has been employed for thousands of years for multiple purposes (Hofman-Caris et al., 2019). The art of inducing, collecting, and utilizing runoff has been practiced by desert-dwelling communities since antiquity, across disparate regions covering southwestern Asia, northern Africa, and southwestern and North America. Remnants of extensive water-harvesting systems are found in these regions. One frequently referenced example from the ancient world is the works of the ancient Nabateans, who inhabited the Negev Desert of southern Israel some 2000 years ago. To sustain their population in the desert environment, the Nabateans built extensive terraces and channels, and hewed out numerous cisterns. The Nabateans were also skilled at choosing appropriate sites for their cisterns and at ensuring that they could be filled with water annually by harnessing rainfall and runoff.

Historically there were two types of water-harvesting techniques, referred to as runoff farming and water spreading. Both techniques enable farming in desert areas of annual and perennial crops. Runoff farming is defined as the application of water-harvesting techniques to enable farming in desert areas (that is to say, collecting runoff from sloping areas and using it to irrigate plots or tracts of flat land, where crops could be grown). In ancient times, stone dikes were constructed across tributary streambeds, thus forming a series of terraced fields. These fields were watered by flows arriving from the upper watershed of the stream, as well as from the adjacent hillslopes. The slopes were often divided into sections, the runoff from each of which was led by means of constructed channels to specific fields (Fig. 1). The water retained and infiltrated in each field was then utilized by annual crops (e.g., wheat or barley) or by perennial crops (e.g., grapevines, olive trees, or other fruit trees).

Water spreading, another form of water harvesting, consists of diverting flash floods from intermittent streams (known as wadis in the Middle East, arroyos in the American Southwest, and dongas in parts of Africa) on to adjacent tracts of land. It is a simple form of flood irrigation, controlled by dikes, check dams, or channels designed to direct and spread the expected flow. It is generally used to irrigate pastures and rangelands, but in places also to sustain groves of trees in arid areas. Water spreading is typically practiced on small watersheds of a few square kilometers. The approach does not work well in larger watersheds because extreme rainfall may pool to such an extent that simple diversion structures are breached. More complex systems can alleviate this risk, such as the construction of dams designed not to retain the floods but merely to detain and regulate them. Farm units located downstream from the dam then get supplied with controlled flows of water. Such dams, called 'detention dams,' are built across a stream to temporarily impound the flood. A large-diameter open pipe and/or leaky barriers is laid through the dam to permit downstream flow at a predetermined rate, as through the drain of a bathtub. Thus, a flash flood that would normally last just a few hours is made to flow through the pipe and on to a series of fields for perhaps several days. The field dikes (made of compacted earth or stone) can then be built economically and safely to withstand floods of a known maximal intensity, so farming operations can be planned accordingly (Fig. 2).

A growing awareness of the potential of water harvesting for improved crop production and sustainable intensification of agriculture arose in the 1970s and 1980s (Piemontese et al., 2020). Water harvesting has been widely applied in different

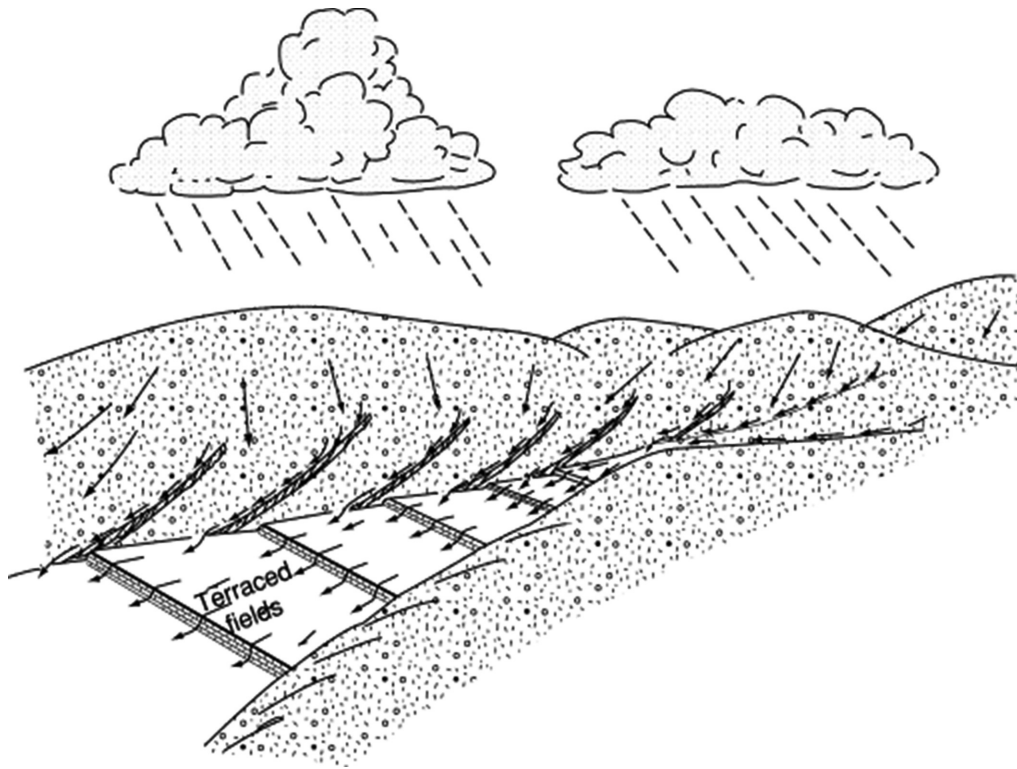


Fig. 1 Ancient runoff farming in the Negev: water trickling off barren slopes was directed to terraced fields in the wadi bed.

socio-ecological contexts with hotspots in large portions of East Africa and Southeast Asia (Piemontese et al., 2020; Fig. 3). Following such development, several water harvesting technologies have been set up in different regions of the world since the 1980s to restore degraded landscapes, conserve soil and water and improve agricultural productivity. For example, globally, about 3 million small reservoirs are in operation in semi-arid regions, supporting 15% of the world's population in semi-arid regions (Mady et al., 2020).

Principles, concepts and components

The basic principles of water harvesting are related to crop-water relationships and are focused on the efficient use of available water (Duveskog et al., 2003). They are about capturing precipitation falling in one area to increase the amount of water available for crop production and decrease runoff risk (Mekdaschi and Liniger, 2013). The principles are summarized in Table 1.

The basic principles of water harvesting summarized in Table 1 indicate that water harvesting is part of sustainable land management or integrated water resource management strategies (Mekdaschi and Liniger, 2013). The principles cover diverse water management strategies including the management of excess water from rainfall and seasonal flooding; increasing rainwater capture and availability, making use of surface runoff; improving direct water infiltration and reducing evaporation; soil water conservation practices that prevent surface runoff and keep rainwater in place; and increasing water use efficiency. Such coverage of diverse water management strategies suggest that a combination of strategies is required to improve the productivity of water.

For the water collection systems to be successful, the main determining factors are the water source, the type of water lifting or pumping system, and the water application mechanisms. Water sources are mainly rainwater and runoff, but harvested water may refill the other water sources such as river/lakes, ponds/springs, and shallow groundwater for later use. As a result, water harvesting usually has four basic components: catchment or collection area, conveyance system, storage component and application area or target (Mekdaschi and Liniger, 2013). The components could be adjacent to each other or connected by a conveyance system or in some cases the storage and application areas may also be the same, typically where water is concentrated in the soil for direct use by plants (Fig. 4).

Water harvesting and hydrological processes

The hydrological cycle of the earth is the sum of all processes in which water moves from the land and ocean surface to the atmosphere and back in the form of precipitation. Of the many processes involved in the water cycle, rainfall and runoff are the most important process related to water harvesting. Whenever the rate at which rainwater is applied to the soil surface exceeds the rate of infiltration into the soil, the excess tends to accumulate over the surface. Where the surface is not perfectly flat and smooth, the excess

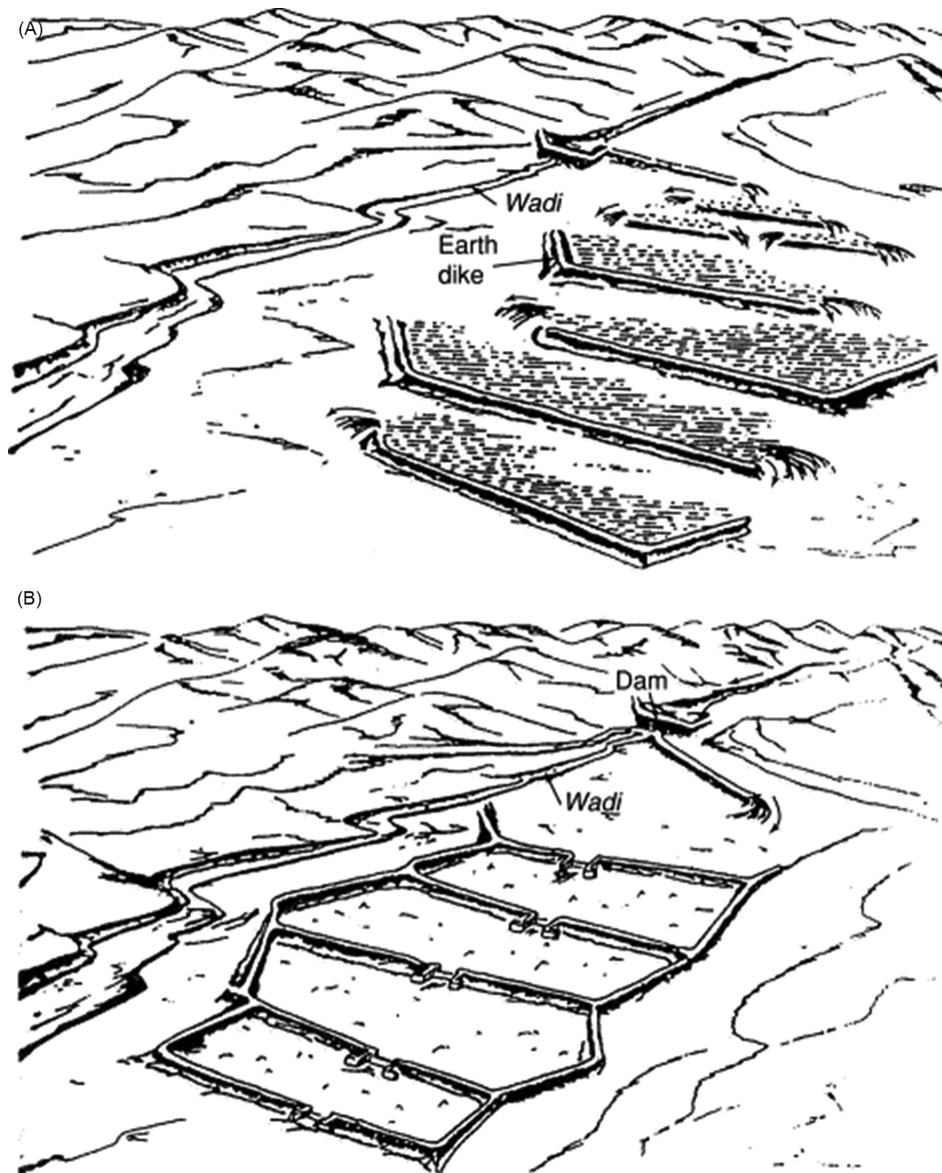


Fig. 2 Two general methods of water spreading by diversion from an intermittent stream (a wadi) on to adjacent land: (a) diversion of an uncontrolled flow over unlevelled land by means of a zigzag series of earthen dikes; and (b) diversion of a controlled flow from a detention dam to a series of level basins with concrete- or stone-lined spillways.

water collects in depressions, forming puddles. The total volume of water thus held, per unit area, is called ‘surface-storage capacity.’ It depends on the geometric irregularities (roughness) of the surface as well as on the overall slope of the land. Only when the surface storage is filled, and the puddles begin to overflow can actual runoff begin. The term ‘surface runoff’ thus represents the portion of the water supply to the surface that is neither absorbed by the soil nor accumulates on its surface, but that runs downslope.

Surface runoff typically begins as sheet flow but, as it accelerates and gains in erosive power, it eventually scours the soil surface to create channels causing soil erosion. Water harvesting technologies have the potential to reduce the peak discharge and heavy rainfall to cause soil erosion, as the systems were designed by diverting the precipitation runoff to the area where it can be used or stored (Kim et al., 2021). One of the primary benefits of water harvesting technologies at field level is to reduce soil erosion and to improve rainfall infiltration and conservation in the soil profile (Bossio et al., 2007). In this line, studies reported the benefit of water harvesting technologies such as tied ridges, mulching, conservation tillage, and furrow systems in improving on-site water availability for plant growth and reducing runoff and soil erosion from agricultural fields (e.g., Tolossa et al., 2020). Water harvesting technologies can also improve soil-moisture content in the soil layer due to increased infiltration of ponded water (e.g., Araya and Stroosnijder, 2010).

Also, hydrological processes at catchment scale are controlled by many physical and climatic factors. Hydrological processes such as runoff, infiltration, groundwater recharge, etc. are also likely to change following the treatment of catchments with different water harvesting technologies (Garg et al., 2021). Several studies conducted across the world demonstrated significant reductions in

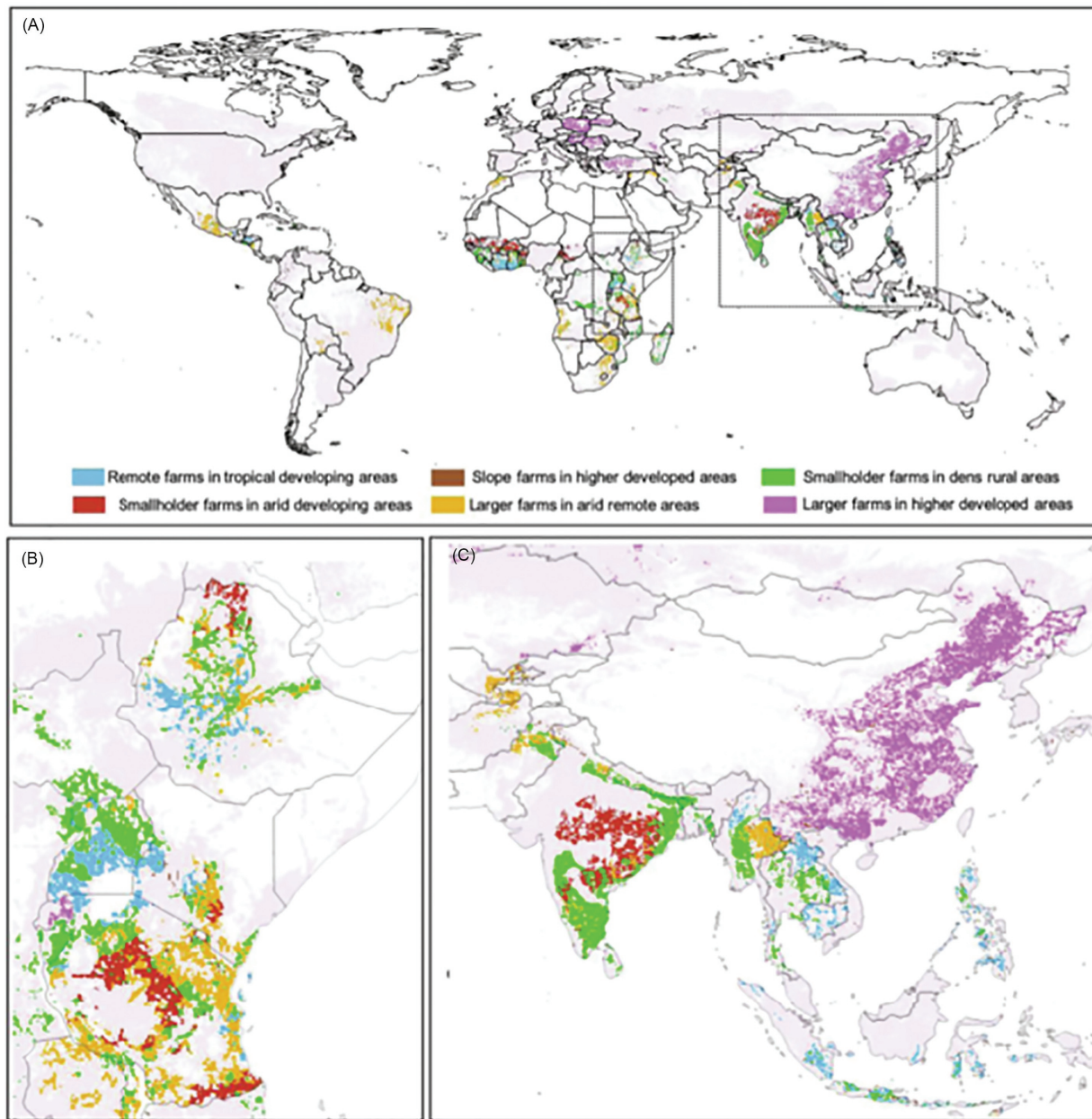


Fig. 3 (a) Global map of archetypes of successful water harvesting, and regional snapshots in (b) East Africa and (c) South-Eastern Asia. When multiple overlapping archetypes, the archetype with the smallest spatial extent is shown. From Piemontese L, Castelli G, Fetzer I, Barron J, Liniger H, Harari N, Bresci E and Jaramillo F (2020) Estimating the global potential of water harvesting from successful case studies. *Global Environmental Change* **63**: 102121.

runoff and sediment yield at the catchment outlets after the construction of diverse water harvesting technologies such as earth bunds with trenches, stone bunds, and check-dams (e.g., Dagnew et al., 2017; Garg et al., 2021). For example, a study conducted in South Asia demonstrated a reduction in surface runoff and soil erosion by more than 60% and 75%, respectively following the implementation of water harvesting technologies (Garg et al., 2021). A study in Ethiopia showed that soil bunds with 50-cm-deep furrows resulted in a reduction of sediment loss by half compared to untreated catchments (Dagnew et al., 2017). Similarly, a study in Southeast Asia demonstrated that water harvesting technologies are efficient in reducing runoff and total sediment yield at the catchment scale.

Categories of water harvesting

Catchment type and size, and the method of water storage are the two most frequently used criteria to classify water harvesting systems. Based on these criteria, five groups of water harvesting systems are identified: (i) micro-catchment systems, (ii) external/macro-catchment systems, (iii) floodwater harvesting, (iv) rooftop/courtyard harvesting and (v) groundwater harvesting.

Table 1 Principles of water harvesting.

Principles	Description
Use rainwater effectively.	Addresses uneven and erratic rainfall conditions through implementing storage techniques to increase the availability of water for crops.
Make effective use of soil water reserves.	Effective use of stored water in the soil profile through crop choice and tillage.
Take measures to avoid runoff.	Physical soil and water conservation and agronomic measures to reduce generation of runoff and soil erosion.
Avoid wasting water through evaporation.	Agronomic measures to reduce water that evaporates directly from bare soil, including the use of modern irrigation practices and controlled timing of water application.
Reduce water losses through drainage.	The combination of agronomic, agroforestry and soil-based measures that increase ground cover, improve soil structure and soil water holding capacity.
Planned irrigation	The use of appropriate irrigation methods, time of application and consideration of crop-water requirements.

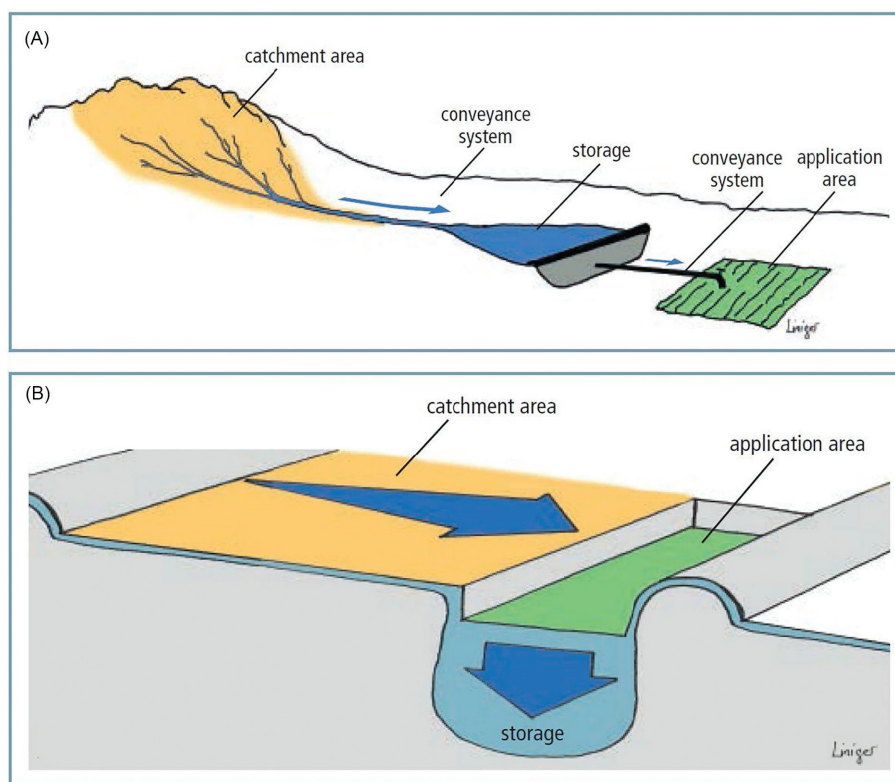


Fig. 4 Basic components of two water harvesting systems: (a) catchment area, storage and application area are clearly separated and connected by conveyance systems; (b) catchment area is bordering application area. Storage is in the soil or ground below the application area. Image Credit Mekdaschi SR and Liniger H (2013) *Water Harvesting: Guidelines to Good Practice*. Centre for Development and Environment (CDE): Bern; Rainwater Harvesting Implementation Network (RAIN): Amsterdam; MetaMeta: Wageningen; The International Fund for Agricultural Development (IFAD): Rome.

Micro-catchment systems

Micro-catchment water harvesting requires development of small structures across mild land slopes, and is designed to trap and collect runoff from a relatively small catchment area (10–500 m²) to support the growth of plants (Fig. 5; Ali et al., 2010; Mekdaschi and Liniger, 2013; Reddy, 2016). This system targets increased water delivery and storage to the root zone, with application to the growth of arable and fodder crops, shrubs, or trees, with added benefits of groundwater recharge. Water availability to plants depends on the micro-catchment runoff yield and water storage capacity of both the plant basin and the soil profile in the plant root zone (Ali et al., 2010). The ratio between the catchment (collection) area to the cultivated (application) area can vary between 2:1 and 10:1 (Mekdaschi and Liniger, 2013). The size of the catchment can be easily controlled by the farmer, which makes the system easy to adapt and replicate.

Micro-catchment water harvesting has the greatest impact in semi-arid to arid areas with high rainfall variability within seasons but works in many environments. Almost any slope can accommodate micro-catchments, including nearly level plains. Soils need

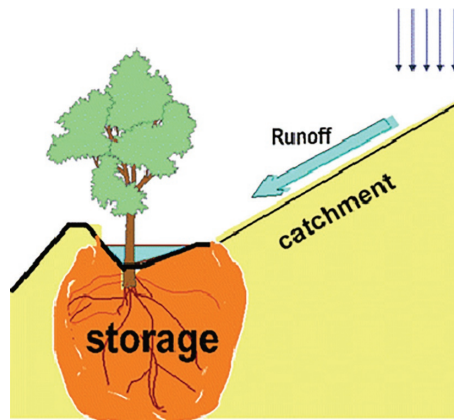


Fig. 5 Micro-catchment water harvesting system. Image credit Reddy PP (2016) Micro-catchment rainwater harvesting. In: *Sustainable Intensification of Crop Production*. Springer: Singapore.

to be deep enough to construct holes or pits that store the collected water. Soils susceptible to sealing and crusting are particularly suitable for inducing runoff in the catchment area. Furthermore, micro-catchment water harvesting can be applied on highly degraded soils where it can be used in the productive rehabilitation process, reducing erosion and flooding. A few examples of technologies that can be categorized under micro-catchment water harvesting systems are summarized in Table 2.

Table 2 Examples of micro-catchment water harvesting technologies (African Development Bank, 2009).

Technologies	Description
Planting pits	Planting pits are used as a water harvesting method to prevent water runoff and erosion and increase infiltration. Pits are dug 20–30 cm wide and 20–30 cm deep and are designed to capture water to plant few seeds of annual or perennial crops. Pits are often found in combination with stone bunds to rehabilitate degraded and crusted lands, and to bring them back into productive land. Planting pits are applied on flat to gently sloped terrain (0–5%) that receives annual rainfall of 350–600 mm. This form of micro-catchment is best suited for land with low permeability, such as silt and clay soils. To further increase crop production, organic matter (such as compost or manure) can be placed in the pits as fertilizer. Many different names are used for planting pits, such as Zai (Zay), Chololo, Mantengo, Ngoro and Tassa. They are widely applied in Sahelian countries, such as Mali, Burkina Faso, and in the Yatanga and Niger regions.
Micro-basins	A micro-basin is a small structure with the shape of a half or a full circle, excavated to obtain a small basin for planting a tree. Micro-basins vary in size; they are small in moist agroecological zones (e.g. 1-m diameter) and large in dry ones (e.g. 2 m). Micro-basins must be carefully lined out on the slope with intermittent placements for runoff control and proper spacing. The procedures in implementing a micro-basin includes cutting of soil as per site requirement, excavating the earth work in a half moon shape throughout its required length and filling the excavated earth in the upper slope for cultivation. After tree planting, micro-basins require little maintenance but weed control may be needed.
Permanent vegetative barriers and strips	Permanent vegetative barriers and strips are made of grasses, shrubs, or trees (often combined) to reduce soil loss and increase infiltration. They are mainly used to replace physical structures on soil with good infiltration (sandy, silty) found on gentle slopes. They are suitable for slopes with gradients of less than 15%. The width of grass strips ranges from 0.5 m to 1.5 m. On steep slopes the width ranges from 2 m to 4 m. Trees and shrub strips are generally wider up to 5–10 m. Since vegetative strips are usually laid along the contours, the distance between them is dictated by the slope of the land. On gentle sloping land, the strips are spaced at 20–30 m, while on steep land the spacing may be 10–15 m. This practice works well in small-scale as well as in large-scale systems. Vegetative strips can also provide firewood and fodder for livestock if palatable varieties of grass or densely spaced bushes are used (cut and carry). They are the least costly or labor demanding type of cross-slope barriers. They are a popular and easy way to ultimately “terrace” land, especially in sub-humid areas, as over time soil eroded upslope is trapped by the vegetation.
Tied earth ridges	Small earth ridges, with furrows between them, blocked with earth ties every 0.5–1.0 m are termed tied ridges. On gentle slopes typical dimensions of the ridges are 20–25 cm height and 0.5–1.5 m spacing between ridges depending on rainfall and the crop to be grown. On flat land compacted bunds and ridges are constructed to generate runoff with a spacing of 1.2–10 m and a ridge height of 30–100 cm, depending on rainfall, crop to be grown and soil characteristics. Runoff is collected between the ridges. The catchment to application area ratio ranges from 1:1 to 5:1. This type of ridge needs maintenance every cropping season and has to be rebuilt about every 5–6 seasons. Tied ridges are often considered an in situ form of water conservation, though where the ridges are large it can be argued that they act themselves as a form of micro-catchment.
Contour earth bunds	WOCAT database on sustainable land management (SLM) measures indicates that Bunds and Fanya Juu are techniques used in agriculture to collect surface run-off, increase water infiltration, and prevent soil erosion. They are suitable for slopes with gradients ranges from 3% to 50%. To avoid lateral water flow and breaching of the bunds cross ties are constructed at regular intervals. Such bunding systems are used in areas with 300–600 mm annual rainfall. They can be applied on all types of relatively permeable soils (e.g., alluvial, red, laterite, brown and, shallow and medium black soils). While this system is often used for the cultivation of annual crops such as maize (<i>Zea mays</i>), tef (<i>Eragrostis tef</i>) or beans (<i>Vicia faba</i> L.), more water-demanding crops, such as bananas, fruits and vegetables may be planted where runoff concentrates immediately above (and sometimes just below) the bund. Contour bunds may be used specifically as a water harvesting technology for tree establishment.

External/Macro-catchment system

Macro-catchment/external catchment systems collect runoff water collection from large areas far from where it is being used (Fig. 6; Mekdaschi and Liniger, 2013). The water may generate from roads, hillsides, and pastures. Several technologies are categorized under this water harvesting system (Table 3), usually designed to control large runoff flows (Kimani et al., 2015). The catchment size ranges from 200 ha to 50 km² and the catchment area cropping area ratio of macro-catchment water harvesting system ranges from 100:1 to 10,000:1 (Kimani et al., 2015).

Macro-catchment water harvesting systems are more applicable in arid and semi-arid areas where the amount of rainfall during the dry season not enough to meet the crop demand. Agricultural production is not possible without irrigation, and in some instances, there is not enough water for household consumption and livestock. The most common sources of runoff water for this system include runoff generated on steep slopes and uncultivated land.

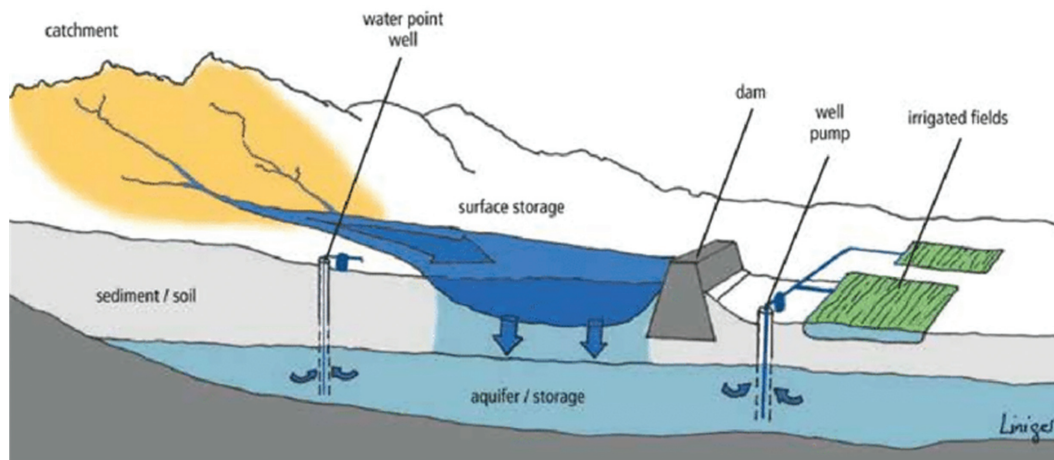


Fig. 6 An example of a macro-catchment water harvesting system linked to a pumping and small-scale irrigation system. Source Mekdaschi SR and Liniger H (2013) *Water Harvesting: Guidelines to Good Practice*. Centre for Development and Environment (CDE): Bern; Rainwater Harvesting Implementation Network (RAIN): Amsterdam; MetaMeta: Wageningen; The International Fund for Agricultural Development (IFAD): Rome.



Fig. 7 Small earthen reservoir constructed in Melissa Watershed, Ethiopia.



Fig. 8 Geomembrane water harvesting ponds in Amhara region, Ethiopia.

Table 3 Most common macro-catchment water harvesting techniques (Kimani et al., 2015; Gebreyess and Amare, 2019).

Technologies	Description
Small earthen Reservoirs	These are earthen structures which are useful to store the runoff water that comes from upstream catchment areas. They are important water harvesting structures that store water for human consumption, crops, and animals. In most cases, the stored water is not free of debris and sediments.
Sand Dams	Simple, easy to maintain and reproducible water harvesting techniques that collect and store clean water for homestead consumption and irrigation. This structure is usually constructed along riverbeds and is used to store water in the sand, prevent run-off, conserve soil, and increase the soil moisture content for crop production. Water from the sand dams can be extracted through holes, or through a pipe that conveys to a tank or a shallow well and the water can be used for multiple purposes. Protecting the sand dams from evaporation prolongs use.
Cistern and Water Tanks	Underground storage tanks that collect water from bare lands, cultivated hillslopes or road catchments during the rainy season. To avoid regular cleaning, small sediment settling basins are connected at the inlet of the tank. The cisterns have plastered walls and lined surfaces. Cisterns are more appropriate in areas where the evaporation is high, skilled labor is available, water scarcity is high, and quality of water is required.
Ponds	Small structures that can store up to 20,000 m ³ of water and have less than 5 m depth below the top-soil surface (Fig. 7). Storage capacity of commonly constructed ponds ranges from 30 m ³ to 50 m ³ . Open ponds are suitable in areas where unskilled labor is available, the evaporation loss is low, efficient water conveyance and applications are used, and when the capital investment is limited. However, they are not suitable in areas where there is a land shortage or where ponds risk malaria disease. The water collected with ponds can be used for livestock and irrigation purposes. Evaporation losses, leaching and sedimentation are the most important factors affecting the sustainability of ponds. In addition, the lack of access to simple water-lifting technologies and watering cans constrains the effective use of water stored in ponds for multiple purposes. Further, lack of fencing and protection of ponds can lead to water pollution. Cultivated hillslopes, natural watercourses, and footpaths are the major water sources for constructed ponds. Water loss due to evaporation and seepage can be decreased in open ponds with shade and lining with clay, plastic and cement as needed (Fig. 8).
Recharge well	Serve to increase water flow to deep aquifers. This method is often applied to recharge deep aquifers where application of water to the land surface are not effective at recharging these aquifers. Also, this technique is more practical in arid and semi-arid areas where drip irrigation is common. The technique is common across the globe and practiced widely in Africa, Southeast Asia, and United States of America.

Floodwater harvesting

Floodwater harvesting is defined as a collection of stream flows for use in irrigation and flood prevention. Streambed (refers to as floodwater harvesting) within streams (wadi) beds and floodwater diversion (refers to as off wadi harvesting) are the two types of floodwater harvesting systems (Oweis et al., 2012). Streambed floodwater harvesting involves building structures across wadi beds to partially or completely dam floodwater and to store it either in surface reservoirs or in channel sediments (Fig. 9). A widespread type of this techniques called terraced wadi system, which are commonly built for agricultural purposes. Terraced wadi systems consist of a series of small dams (check dams) that intersect parts of a wadi course. The check dams lower the runoff velocity of the floods and thereby their transport capacity. This in turn leads to the accumulation of sediments behind the dams and gradually build a terrace or sediment reservoir upstream that might be cultivated (Beckers et al., 2013). Other types of this technique are groundwater dams such as sand storage dams and subsurface dams. Sand storage dams’ function with the same principle as the

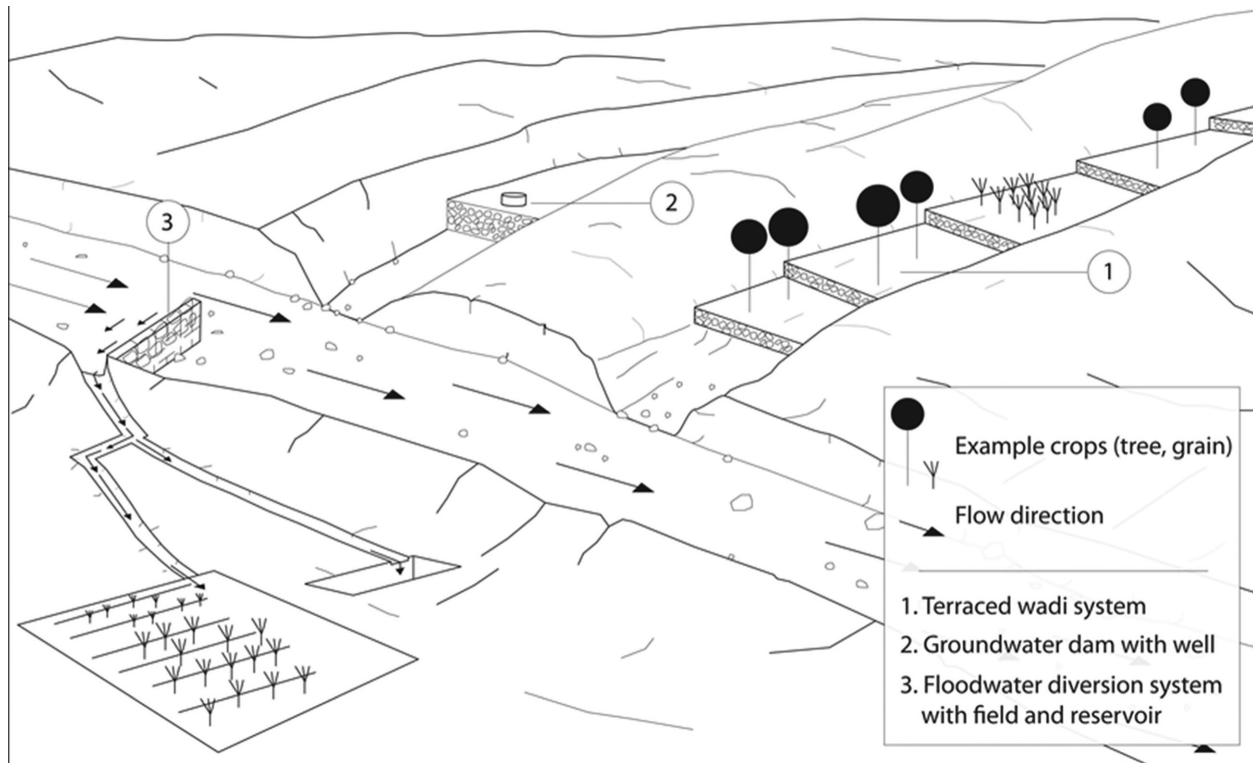


Fig. 9 Examples of floodwater harvesting techniques. Slopes are exaggerated (Beckers et al., 2013). From Beckers B, Berking J and Schütt B (2013) Ancient water harvesting methods in the drylands of the Mediterranean and Western Asia. *Journal for Ancient Studies* 2: 145–164.

terraced wadi system. Subsurface dams are built into the alluvial fill of streams by excavating a trench down to an impervious layer and building a wall in the trench which is subsequently backfilled with the excavated material (Beckers et al., 2013).

Floodwater diversion systems are built to deflect floods from a stream (wadi) channel to convey the water to adjacent storage devices or fields (Fig. 9). This is either accomplished by damming parts of the wadi or blocking the entire channel. The retaining structures are called diversion dams. Floodwater diversion system usually categorized as: (i) spate diversion where flash floods are diverted from ephemeral rivers, (ii) inundation canals that starts to flow when the flow reaches at a certain level, and (iii) recession irrigation where perennial river overtops to its banks.

Floodwater harvesting is most suitable in arid and semi-arid environments where there is high variability of rainfall and water shortages (Ghahari et al., 2014). However, floodwater harvesting is also used in temperate climate for flood prevention. This technique is widely used in North and East Africa, Italy, Israel, Jordan, Tunisia, Libya and Spain for agricultural purpose, and in the Netherlands and Scotland mainly for flood prevention (Alcoforado, 2018). Spate irrigation, for example, is one of the forms of floodwater harvesting used widely in Eastern Africa (Gebreyess and Amare, 2019). Spate irrigation helps to harvest and reuse runoff water and sediments, produce virtual water and recharge groundwater (Ghahari et al., 2014). Direct runoff farming technique is used for immediate use of water in the agriculture areas such as bed preparation, and to make for ridge preparations especially to for the upcoming runoff and floodwater. The water can be collected from roads, hillsides, soil ridges, and short-lived streams. This water is used for growing cereal crops (wheat, barley, maize), vegetables (onion, cabbage, garlic, tomato, etc. and fruit trees).

Rooftop harvesting

Rooftop water harvesting is a system of water harvesting from roofs and paved courtyards during the rainy season (Fig. 10). The harvested water is mainly used for domestic purposes, very rarely for garden crops, too. Rooftop water harvesting yields good quality of water and hence, it is the most common water harvesting system around homestead and schools.

This water harvesting technique is applicable for farmers who would like to harvest water using ponds and tanks for later use during the peak dry season where water shortage affects the community. Rooftop catchments are other options preferred and feasible for farmers (Tadesse et al., 2008). Rooftop area and average rainfall in an area determines the total quantity of water that can be collected or harvested. This is small compared to other forms of water harvesting, but large volumes of water can still be collected. For example, using rooftop water harvesting just on 19 buildings at South Indian University could collect about 114000 m³ of water to help resolve scarcity in the non-monsoon season (Anchan and Prasad, 2021). A study of rooftop rainwater harvesting on the slopes of Mt. Elgon, Uganda estimated 163,063 m³year⁻¹ of captured water, giving the potential to save 4–7% of the annual

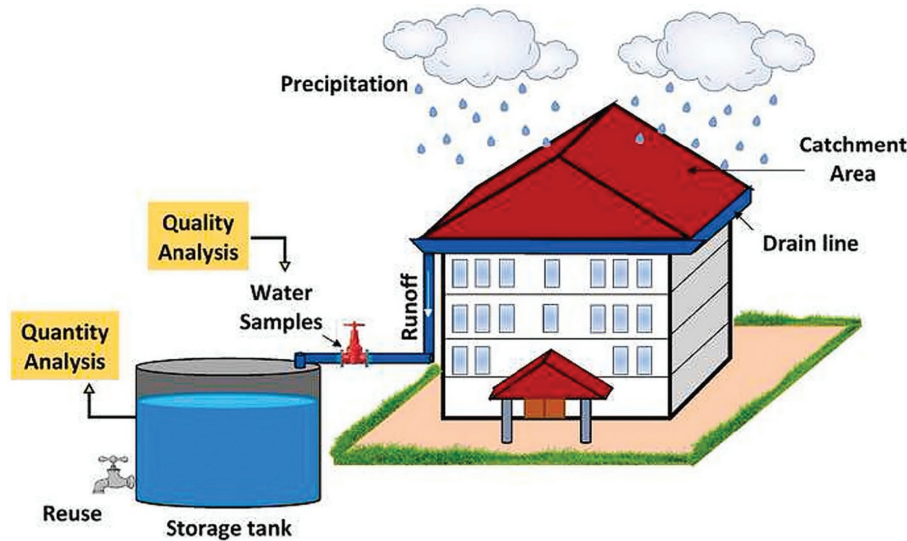


Fig. 10 Rooftop water harvesting. Image credit Anchan SS and Prasad HS (2021) Feasibility of roof top rainwater harvesting potential-A case study of South Indian University. *Cleaner Engineering and Technology* 4: 100206.

household water demand (Bernard and Joyfred, 2020). Most of the water collected from the rooftop is used for drinking, domestic and livestock consumption.

Rooftop water harvesting can be considered as a technique that helps to harvest rainwater. Various systems or mechanisms can be used to collect rainwater harvested from rooftops. For example, underground water holding tanks, ferrocement jar, polyethylene laminated earth holding tank and concrete tanks can be used to store water harvested from rooftop (Fig. 11). The stored water can be used for drinking, washing, irrigation and livestock (Gebreyess and Amare, 2019).

Groundwater harvesting

Groundwater harvesting involves extracting water from qanats, underground dams and wells (Fig. 12). Qanat systems consists of a horizontal tunnel that taps underground water in deposited fans, which brings it to the surface without using pumps or any other pressurized systems. These systems have an inclination of 1–2%, classified by length as short, medium, and long size qanats and



Fig. 11 Rooftop water harvesting using water tanks in Oromia region, Ethiopia.



Fig. 12 Shallow wells and recharging methods in Amhara region, Ethiopia.

reaching 30 km long (Prinz and Singh, 2000). Underground dams store water below the ground, often using sediment (sand dams) to reduce water loss due to evaporation. These systems filter water so help to minimize water pollution (Prinz and Singh, 2000).

The water harvested through this system can be delivered to the fields steadily and used for agricultural production and homestead consumption (Prinz and Singh, 2000). Ponds, large sized water tanks and ground water dams are needed to extract water from the ground and store for later use while there is a water shortage during the dry season or to supplement when there is not sufficient rainfall.

Groundwater harvesting systems are suitable in areas where the annual rainfall ranges from 500 to 1400 mm (Tadesse et al., 2008). They are also suitable in areas where the evaporation is high, skilled labor is available, water scarcity is high, and good water quality is required (Lulseged et al., 2011). If a farmer uses wells as a water source and ponds for collection of water for irrigation, they could grow crops two times per year and meet the household consumption needs.

Environmental and socio-economic benefits of water harvesting and limitations

Water harvesting is used globally for both socio-economic and environmental reasons (Díaz-Pereira et al., 2016). Socio-economic benefits of water harvesting include increased crop yield and farm income leading to improved food security, decreased soil erosion and conservation, and better social networks. As part of broader strategies to improve agricultural sustainability, water harvesting can have added impacts. For example, a study in Zimbabwe demonstrated that the combined use of inorganic fertilizers and water harvesting technologies helps to close yield gap (Kahinda et al., 2007). Supplemental irrigation due to the use of water harvesting technologies reduced the risks of complete crop failure from 20% down to 7% on average and also enhances transpiration water productivity from 1.75 kg m^{-3} up to 2.3 kg m^{-3} on average, by mitigating intra-seasonal dry spells (Kahinda et al., 2007). A study in the semiarid loess region of China reported water harvesting technologies led to a 10–20% increase in crop yield by enabling supplemental irrigation during critical growth stages (Li, 2003). A regional study in Africa by Lebel et al. (2015) indicated that water harvesting can bridge the yield gaps of maize due to water deficit by 40% of the yield gap under the current condition and 31% under future (2050s) climatic conditions, hence providing an alternative to irrigation from scarce or inaccessible groundwater resources. A study in Ethiopia demonstrated that adoption of rainwater harvesting technology had increased adopters' annual farm income by 35% and daily calorie intake per adult equivalent by 16% compared to the annual farm income of non-adopters

(Mekuria et al., 2020). The cited case studies indicate that water harvesting technologies can have considerable contributions to achieve food security and improve the livelihoods of local communities.

In addition to food and feed production (direct benefits), water harvesting technologies have substantial environmental and social returns, such as combating land degradation and migration from rural to urban areas and employment (Oweis and Hachum, 2009; Díaz-Pereira et al., 2016). In the Southern part of the West Bank, water harvesting technologies such as stone terraces and semi-circle bunds were found to reduce the amount of total runoff by 73–80%, depending on climatic conditions (Al-Seekh and Mohammad, 2009). This cause soil moisture content to significantly increase under these water harvesting techniques compared to the control. A study in Southern Ethiopia indicated that water harvesting technologies such as trenches increased tree biomass (on average by 7.7–25.5 kg/tree), which have positive implications to improving regulating ecosystem services (Alem et al., 2020).

Water harvesting systems can provide many advantages over other irrigation schemes. For example, micro-catchment water harvesting systems are simple and inexpensive to construct and can be built rapidly using local materials and manpower. The runoff water has a low salt content and because it does not have to be transported or pumped, it is relatively inexpensive. Micro-catchments enhance leaching and often reduce soil salinity. An example is the use of micro-catchment techniques in Arizona that have returned land retired from agriculture due to high salt content from groundwater irrigation to productive use (Fidelibus and Bainbridge, 1995). Water harvesting systems, particularly micro-catchment water harvesting techniques have higher runoff collection efficiency and negligible conveyance losses. Water harvesting techniques contribute positively to achieving resilient environment and community through adapting climate change and improving food security.

Despite their many perceived potential benefits, the main constraints regarding the adoption of water harvesting techniques at local level include availability of active farm labor, implementation and maintenance costs, the loss of productive land, equitable access to runoff and related resources, changes of the runoff conveyance channels due to erosion and deposition, extension service and training, level of monthly income, main source of livelihood, and land tenure (Mangisoni et al., 2019). Related studies have shown that land size, access to financial services, access to information and contact with extension officers are some of the variables that have a significantly positive effect on the adoption of in field rainwater harvesting technology, while hired labor has significantly negative correlation with adoption (Badisa, 2011).

Table 4 summarizes the required earthworks or stoneworks for various water harvesting systems. The labor costs for constructing the water harvesting systems are expressed in person days. In general, larger structures (e.g., trapezoidal bunds) may take 50% more labor per unit volume of earthworks than smaller structures (e.g., Negarim micro-catchments) because of increased earthmoving required. Typical rates for earthwork range from 1.0 to 3.0 m³ person/day whereas typical rates for stonework is 0.5 m³ person/day.

The limitations or constraints highlights the need for multiple solutions including incentives for implementation for most effective water harvesting techniques adapted to local environmental and socioeconomic conditions. Also, the planning and implementation of water harvesting techniques should not only deal with hydro-climatic challenges but also looks on the socio-economic challenges for efficient and equitable distributions of resources following the implementation of water harvesting infrastructures. The other important socio-economic reality that should be considered is that the necessary to cooperate in construction as most of the technologies or structures cannot be made by individual farmers.

Conclusion

Water harvesting improves the use of available water from precipitation and run-off by concentrating it for immediate use and storage. The control systems that are implemented can divert water to decrease erosion and flood risk. Under climate change, water harvesting will improve resilience to stresses from droughts and extreme rainfall. Water harvesting can also recharge depleted groundwater sources. Water harvesting involves very diverse methods of collecting and managing floodwaters and surface runoff. Techniques can be grouped as: micro-catchment water harvesting, macro-catchment water harvesting, floodwater harvesting,

Table 4 Earthwork/Stonework for various water harvesting systems.

System name and number	Earthwork (m ³ /ha treated)						Stonework (m ³ /ha treated)	
	Negarim micro-catchments (trees)	Contour bunds (trees)	Semi-circular bunds (grass)	Contour ridges (crops)	Trapezoidal bunds (crops)	Water spreading bunds (crops)	Contour stone bunds (crops)	Permeable rock dams (crops)
Slope %	(1)	(2)	(3)	(4)	(5)	(8)	(6)	(7)
0.5	500	240	105	480	370	305	40	70
1.0	500	360	105	480	670	455	40	140
1.5	500	360	105	480	970	NR ^a	40	208
2.0	500	360	210	480	NR ^a	NR ^a	55	280
5.0	835	360	210	480	NR ^a	NR ^a	55	NR ^a

^aNot recommended.

rooftop and courtyard water harvesting and groundwater water harvesting. These techniques of water harvesting have diverse socio-economic and environmental benefits including increased crop yield and farm income as well as improved food security, knowledge of soil erosion and conservation and social networks, increased soil moisture content and water availability, reduced soil loss and reduced downstream flooding and siltation, improved infiltration and groundwater recharge, and improved soil fertility. Despite the many potential benefits of water harvesting, the main constraints to implementing at local level include labor, construction costs, loss of productive land, maintenance, equitable access to runoff and related resources, and change of the runoff conveyance channels due to erosion and deposition. The limitations or constraints highlight the need for multiple solutions including incentives for implementation for most effective water harvesting techniques adapted to local environmental and socioeconomic conditions.

References

- African Development Bank (2009) *Rainwater Harvesting Handbook: Assessment of Best Practices and Experience in Water Harvesting*. Tunis: African Development Bank. <http://www.rwsn.ch/documentation/skatdocumentation>.
- Alcoforado FAG (2018) Flood control and its management. *HSPA Journal of Atmospheric & Earth Sciences* 1: 005.
- Alem S, Némec P, and Habrová H (2020) Effects of a trench as a moisture harvesting structure on the biomass production and growth of trees planted to restore degraded land, Southern Ethiopia. *Applied Sciences* 10(23): 8560.
- Ali A, Yazar A, Aal AA, Oweis T, and Hayek P (2010) Micro-catchment water harvesting potential of an arid environment. *Agricultural Water Management* 98: 96–104.
- Al-Seekh SH and Mohammad AG (2009) The effect of water harvesting techniques on runoff, sedimentation, and soil properties. *Environmental Management* 44: 37–45.
- Anchan SS and Prasad HS (2021) Feasibility of roof top rainwater harvesting potential-A case study of South Indian University. *Cleaner Engineering and Technology* 4: 100206.
- Araya A and Stroosnijder L (2010) Effects of tied ridges and mulch on barley (*Hordeum vulgare*) rainwater use efficiency and production in Northern Ethiopia. *Agricultural Water Management* 97: 841–847.
- Badisa KT (2011) *Socio-economic Factors Determining In-Field Rainwater Harvesting Technology Adoption for Cropland Productivity in Lambani Village: A Case Study of Thulamela Local Municipality of the Vhembe District in Limpopo province*. MSc thesis South Africa: University of Limpopo.
- Beckers B, Berking J, and Schütt B (2013) Ancient water harvesting methods in the drylands of the Mediterranean and Western Asia. *Journal for Ancient Studies* 2: 145–164.
- Bernard B and Joyfred A (2020) Contribution of rainfall on rooftop rainwater harvesting and saving on the slopes of Mt. Elgon, East Africa. *The Scientific World Journal* 2020: 7196342.
- Bossio D, Critchley W, Geheb K, van Lynden G, and Mati B (2007) Conserving land protecting water. In: *Water for Food, Water for Life: A Comprehensive Assessment of Water Management in Agriculture*, pp. 551–583. London: Earth scan and Colombo, International Water Management Institute.
- Dagnaw DC, Guzman CD, Zegeye AD, Akal AT, Moges MA, Tebebu TY, Mekuria W, Ayana EK, Tilahun SA, and Steenhuis TS (2017) Sediment loss patterns in the sub-humid Ethiopian highlands. *Land Degradation & Development* 28(6): 1795–1805.
- Diaz-Pereira E, Asunción Romero-Díaz M, and de Vente J (2016) Environmental and socioeconomic benefits and limitations of water harvesting techniques in semiarid regions. In: *EGU General Assembly Conference Abstracts*. pp. EPSC2016-13505.
- Duveskog D, Nyagaka D, Mweri B, Shiribwa M, and Kaumbutho P (2003) Soil and water conservation: With a focus on water harvesting and soil moisture retention. In: *A Study Guide for Farmer Field Schools and Community-Based Study Groups*. Harare, Zimbabwe: The Farm Level Applied Research Methods for East and Southern Africa (FARMESA).
- FAO (2020) *The State of Food and Agriculture 2020*. Rome: Overcoming Water Challenges in Agriculture. <https://doi.org/10.4060/cb1447en>.
- Fidelibus MW and Bainbridge DA (1995) Micro-catchment water harvesting for desert revegetation. In: *Soil Ecology and Restoration Bulletin*, p. 11. San Diego: Soil Ecology and Restoration Group.
- Garg KK, Anantha KH, Venkataradha A, Dixit S, Singh R, and Ragab R (2021) Impact of rainwater harvesting on hydrological processes in a fragile watershed of South Asia. *Groundwater*. <https://doi.org/10.1111/gwat.13099>.
- Gebreyess BF and Amare A (2019) Water harvesting technologies in semi-arid and arid areas. *Journal of Degraded and Mining Lands Management* 7(1): 1921–1928.
- Ghahari G, Hashemi H, and Berndtsson R (2014) Spate irrigation of barley through floodwater harvesting in Thegareh-Bygone Plain, Iran. *Irrigation and Drainage* 63: 599–611.
- Hofman-Caris R, Bertelkamp C, de Waal L, van den Brand T, Hofman J, van der Aa R, and van der Hoek JP (2019) Rainwater harvesting for drinking water production: A sustainable and cost-effective solution in the Netherlands? *Watermark* 11(3): 511.
- Kahinda JMM, Rockström J, Taigbenue AE, and Dimes J (2007) Rainwater harvesting to enhance water productivity of rainfed agriculture in the semi-arid Zimbabwe. *Physics and Chemistry of the Earth, Parts A/B/C* 32(15–18): 1068–1073.
- Kim DG, Grieco E, Bombelli A, Hickman JE, and Sanz-Cobena A (2021) Challenges and opportunities for enhancing food security and greenhouse gas mitigation in smallholder farming in sub-Saharan Africa, A review. *Food Security* 13: 457–476.
- Kimani MW, Gitau AN, and Ndunge D (2015) Rainwater harvesting technologies in Makueni County, Kenya. www.researchinventy.com *Research Inventy: International Journal of Engineering and Science* 5: 39–49.
- LaFevor MC and Ramos-Scharron CE (2021) Effects of hillslope trenching on surface water infiltration in subalpine forested catchments. *Hydrology* 8: 147. <https://doi.org/10.3390/hydrology8040147>.
- Lebel S, Fieskens L, Forster PM, Jackson LS, and Lorenz S (2015) Evaluation of in situ rainwater harvesting as an adaptation strategy to climate change for maize production in rainfed Africa. *Water Resources Management* 29(13): 4803–4816.
- Lulseged M, Seleshi BA, Gayathree J, Fitsum H, and Teklu E (2011). Inventory, sustainability assessment, and upscaling of best agricultural water management practices. In Awulachew SB; Erkossa T; Balcha Y. (Comps.). *Irrigation and Water for Sustainable Development: Proceedings of the Second Forum*, Addis Ababa, Ethiopia, 15-16 December 2008, Colombo, Sri Lanka, International Water Management Institute (IWMI). pp.153–183.
- Li X-Y (2003) Rainwater harvesting for agricultural production in the semiarid loess region of China. *Journal of Food Agriculture and Environment* 1: 282–285.
- Mady B, Lehmann P, Gorelick SM, and Or D (2020) Distribution of small seasonal reservoirs in semi-arid regions and associated evaporative losses. *Environmental Research Communications* 2(6), 061002.
- Mangisoni JH, Chigowo M, and Katengeza S (2019) Determinants of adoption of rainwater-harvesting technologies in a rain shadow area of southern Malawi. *AFBM Journal* 14: 106–119.
- Mekdaschi SR and Liniger H (2013) *Water Harvesting: Guidelines to Good Practice*. Bern: Centre for Development and Environment (CDE).
- Mekuria KZ, Kassegn Amede A, and Endris Mekonnen E (2020) Adoption of rainwater harvesting and its impact on smallholder farmer livelihoods in Kutaber district, South Wollo Zone, Ethiopia. *Cogent Food & Agriculture* 6(1): 1834910.
- Oweis T and Hachum A (2009) Water harvesting for improved rainfed agriculture in the dry environments. Rainfed agriculture: Unlocking the potential. In: Kijne WJ, Barker R, and Molden D (eds.) *Water Productivity in Agriculture: Limits and Opportunities for Improvement*, pp. 179–197. Wallingford, UK: CAB.
- Oweis Y, Prinz D, and Hachum Y (2012) *Water Harvesting for Agriculture in the Dry Area*. Leiden, the Netherlands: ICARDA, CRC Press/Balkema266.
- Payus C, Ann Huey L, Adnan F, Besse Rimba A, Mohan G, Kumar Chapagain S, Roder G, Gasparatos A, and Fukushi K (2020) Impact of extreme drought climate on water security in North Borneo: Case study of Sabah. *Watermark* 12(4): 1135.

- Piemontese L, Castelli G, Fetzer I, Barron J, Liniger H, Harari N, Bresci E, and Jaramillo F (2020) Estimating the global potential of water harvesting from successful case studies. *Global Environmental Change* 63: 102121.
- Prinz D (2011) The Concept, Components and Methods of Rainwater Harvesting (presentation). In: *2nd Arab Water Forum "Living With Water Scarcity", 20–23 November 2011, Cairo, Egypt*.
- Prinz D and Singh AK (2000) Water resources in arid regions and their sustainable management. *Annals of Arid Lands* 39: 251–252.
- Reddy PP (2016) Micro-catchment rainwater harvesting. In: *Sustainable Intensification of Crop Production*. Singapore: Springer.
- Tadesse N, Berhane A and Bheemalingeswara A (2008) *Initiatives, Opportunities and Challenges in Shallow Groundwater Utilization: A Case Study From Debrekidane Watershed, Hawzien Woreda, Tigray Region, Northern Ethiopia*, Agricultural Engineering International: The CIGR Ejournal, Manuscript LW 08 008. Vol. X.
- Tolossa TT, Abebe FB, and Girma AA (2020) Rainwater harvesting technology practices and implication of climate change characteristics in Eastern Ethiopia. *Cogent Food & Agriculture* 6(1): 1724354.
- Ward M, Poleacovschi C, and Perez M (2021) Using AHP and spatial analysis to determine water surface storage suitability in Cambodia. *Watermark* 13: 367.
- Young SL, Frongillo EA, Jamaluddine Z, Melgar-Quinonez H, Pérez-Escamilla R, Ringler C, and Rosinger AY (2021) Perspective: The importance of water security for ensuring food security, good nutrition, and well-being. *Advances in Nutrition* 12: 1058–1073.

Irrigation methods

Shmulik P Friedman, Department of Environmental Physics and Irrigation, Institute of Soil, Water and Environmental Sciences, Agricultural Research Organization, The Volcani Center, Rishon LeZion, Israel

© 2023 Elsevier Ltd. All rights reserved.

This is an update of D.L. Bjorneberg, R.E. Sojka, IRRIGATION I Methods, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 273-280, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00276-9>.

Introduction	608
Design and management	609
Irrigation uniformity and efficiency	610
Surface irrigation	610
Furrow irrigation	611
Basin and border irrigation	612
Sprinkler irrigation	613
Solid set sprinkler systems	614
Set-move sprinkler systems	615
Self-moving sprinkler systems	615
Lawn sprinkler systems	618
Micro-irrigation	618
Drip irrigation	619
Bubblers and microsprayers	619
Subirrigation	620
Fertigation (chemigation)	620
Salinity hazards and management	621
Choosing an irrigation method	621
Conclusion	621
References	622
Relevant websites	623

Abstract

Irrigation increases yields and ensures the reliability and continuity of crop production. Approximately 40% of the world's food is grown on irrigated land, which accounts for only about 20% of the world's land used for crop production. Various irrigation methods have been developed over time to meet the irrigation needs of certain crops in specific areas. The three main methods of irrigation are surface, sprinkler, and micro-irrigation. In basin, border and furrow—surface irrigation, water flows over the soil by gravity. Sprinkler irrigation applies water to soil by sprinkling or spraying water droplets from fixed or moving systems. Micro-irrigation applies frequent, small water volumes by localized dripping, bubbling, or spraying.

Key points

- Different surface, sprinkler and micro-irrigation methods are described.
- The hydraulic and agronomic features of the different irrigation methods are discussed.
- Their merits and limitations are elucidated in terms of the wetting patterns and water use efficiency.

Introduction

Irrigation is the process of applying water to soil, to meet the water needs of growing plants, and to assist in crops production and in the wellbeing of landscape plants and lawns. The goal is to apply the water to the land as uniformly as possible, so that each plant can take up the amount of water it needs, neither too much nor too little. Water from rivers, reservoirs, lakes, or aquifers is pumped or flows by gravity through pipes, canals, ditches, or even natural streams, and then applied to the irrigated field (Bjorneberg and Sojka, 2005; Brouwer et al., 1988; Hoffman et al., 2007; Israelsen, 1950; Pereira and Trout, 1999; Rohwer et al., 2007). As natural water sources become depleted, the use of recycled, wastewater is growing in prominence. This can create increased challenges, due to soil salinization and the presence of organic compounds or other contaminants that affect the soil and environment.

Applying water to fields and orchards enhances the magnitude, quality, and reliability of crop production—approximately 40% of the world's food is grown on irrigated land, which accounts for only about 20% of the world's land used for crop production. Namely, the areal productivity of irrigated agriculture is almost threefold that of rain-fed agriculture (i.e., that does not use irrigation but instead relies only on direct rainfall).

Various irrigation methods have been developed over time to meet the irrigation needs of certain crops in specific areas. The three main methods of irrigation are surface, sprinkler, and micro-irrigation. In surface (basin, border and furrow) irrigation, water flows over the soil by gravity. Sprinkler irrigation applies water to soil by sprinkling or spraying water droplets from fixed or moving systems. Micro-irrigation applies frequent, small water volumes by dripping, bubbling, or spraying. A fourth, minor, irrigation method is subirrigation, in which the water table is raised to or held near the plant root zone using ditches or subsurface drains to supply the water.

Surface irrigation is thousands of years old. The impact sprinkler, was invented in the US in the 1930s, and quickly found widespread use. Modern drip irrigation with plastic emitters was developed in Israel in the early 1960s. Both sprinkler and drip irrigation offer a practical, usually superior alternative to surface irrigation. Worldwide, about 300 Mha of cropland is equipped for irrigation, out of which about 85% is actually irrigated. This area doubled since 1970, and about 68% of this area is in Asia, 17% in the Americas, 9% in Europe, 5% in Africa and 1% in Oceania. China, India, United States, Pakistan, and Iran account for ca 60% of global irrigated land. About 75% of this land is surface-irrigated, about 18% is sprinkler-irrigated and in about 7% micro-irrigation is practiced. In developed countries sprinkler and micro-irrigation are used on a greater percentage of irrigated cropland, primarily to reduce labor and improve control of water application (ICID: <https://icid-ciid.org/home>).

Several criteria can be used to categorize the irrigation methods. Sprinkler and drip are pressurized irrigation methods in the sense that the water is supplied to the irrigated land at pressures larger than atmospheric via pipes and emitters (sprinklers, sprayers, bubblers drippers). In gravitationally-driven surface irrigation, water is at atmospheric pressure, and the subirrigation water supplied to the root zone is under tension, i.e., at a pressure smaller than atmospheric. Soil water flow dimensionality is another criterion discriminating one-dimensional, vertical, sprinkler, basin (or border) and sub-irrigation from two (or three)-dimensional, vertical and horizontal, soil water flow in furrow and drip (or micro) irrigation. As to the dimensionality of surface streams (waves) flow: for example, in wide basins or in a border strip with cross slope, it is not always that a longitudinal direction dominates, thus the non-negligible transverse flow and variations in water-surface elevation makes these surface flows two-dimensional. A third criterion, related to all irrigation methods, is whether irrigation is supplementary to rainfall as an additional water source, or whether it is full irrigation, whereby crops rarely depend on any contribution from rainfall. Full irrigation is practiced in arid regions experiencing very little rainfall or when crops are grown in semi-arid areas outside of rainy seasons. Supplementary irrigation is used in humid and semi-arid regions when rainfall in the growing season does not meet the full plant water demand. Moreover, in temperate regions, supplementary irrigation is used for high value crops and the production of sensitive root crops, such as potatoes. Erratic weather, associated with climate change, has increased the demand for irrigation across many climatic regions on Earth.

Although manual operation of irrigation systems is still used, most agricultural, residential and landscape irrigation systems, especially the pressurized sprinkler and drip systems, are operated automatically using irrigation controllers and electric solenoid valves. Each irrigation zone has one or more valves that are wired to the controller. When the controller sends power to the valve, the valve opens, allowing water to flow to the lateral lines and water emitters in that zone. This is sometimes triggered via soil moisture sensors installed in the field.

In variable rate irrigation (VRI) practices (within “site-specific” farming), the irrigated fields are delineated to separate management zones (MZ), based on previous-years yield maps and on proxy and remote, soil and plant, sensing, and each MZ is irrigated separately with different irrigation depths. Irrigation depth (or ‘application depth’ in millimeters or inches) is the volume of water applied to a given irrigated area.

Design and management

The design and management of irrigation systems involves agronomic, hydraulic and economic considerations. The irrigation system capacity should meet the crop water demand at critical growth stages and climatological conditions, and also allow for downtime, when the system can be repaired and maintained. The designed system capacity can be decreased by accounting for the use of stored soil water to meet the crop water demand during the peak-use period, and if relevant, also the expected rainfall events during that period. In cases where the water supply cannot match the peak-use rate, water storage in open or closed reservoirs is required (Benami and Ofen, 1984; Cuenca, 1989; Hoffman et al., 2007; Waller and Yitayew, 2016).

Irrigation scheduling refers to the irrigation timing (or frequency) and dosing (daily irrigation depths). The irrigation frequency and the geometry of two or three-dimensional irrigation modes, should account for plant and soil properties such as the depth of root systems and the soil hydraulic conductivity (or capillary length) and water retention characteristic. Crops with shallow root systems planted in coarse-textured, sandy soils should be irrigated more frequently than deep-rooted crops planted in fine-textured, clayey soils. Irrigation dosing is determined mostly by the crop and climatic conditions, varying with the crop growth stages and along the annual seasons. A good understanding of water behavior in the soil-plant-atmosphere continuum (SPAC), of the irrigation system hydraulics, and of the ability to influence them, allows rational decisions, balancing irrigation system, water and other costs with the costs (and benefits) associated with under or over-irrigation (Hoffman et al., 2007).

Irrigation uniformity and efficiency

Irrigation uniformity (IU) quantifies the evenness of water distribution to individual plants, determined mainly by the variation of pressures in a pressurized irrigation system or water surface elevations in a gravity-fed system, responding to and affecting the flow rates at the various locations in the irrigation system.

Non-uniform irrigation means excess water should be applied in certain areas, such that enough water is applied in other areas to minimize plant water stress. The excess water may cause surface runoff and also oxygen deficiency (hypoxia), if the soil profile gets too wet. Several mathematical definitions, which account for the spatial variation of the applied irrigation depths in various irrigation systems, are used to quantify its distribution uniformity (Hoffman et al., 2007).

Irrigation efficiency (IE) can be defined in different scales and contexts. On larger scales it accounts for water losses in reservoirs and during its conveyance to the irrigated fields (conveyance efficiency). At the field scale it accounts primarily for the uniformity of water distribution to its different zones (distribution efficiency) and for operational and accidental spills. Conveyance efficiency and distribution efficiency both refer to the water delivery from the source to a specific outlet emitter. At the soil profile scale, it is usually defined as the ratio between the productive evapotranspiration (ET) water and the total irrigation water that also ends up as runoff and excessive deep percolation losses. Another option is to differentiate between consumptive transpiration (T) water and non-consumptive, irrigation efficiency—reducing evaporation (E) water, although evaporation for cooling and frost protection, as well as deep percolation for salt leaching, are beneficial to the plants (Hoffman et al., 2007).

Surface irrigation

Surface irrigation entails water flowing by gravity over soil. Water is usually supplied by gravity through canals, above or below ground pipes and ditches, with gates, breaches or siphon tubes from the water source to the cropped field. In some locations, however, water may need to be pumped from the source to a field at a higher elevation. Runoff is stopped by a berm, and water ponds behind it; otherwise, flow runs off at the end, and the tailwater can be collected for reuse (Hoffman et al., 2007; Walker and Skogerboe, 1987). If water is reused, surface irrigation efficiencies can be greater than those listed in Table 1. Surface irrigation is also termed ‘controlled flooding’, and its types include furrow (also surge), level basin, and border strip irrigation. Surface irrigation systems are typically used for field crops, pastures, and orchards. A major characteristic of surface irrigation is that the same flow channels conveying water to their distal ends are also the sources of water infiltrating into the soil and being taken up by plant roots. Another distinct characteristic is that the depth of flow is not prescribed, but responds to the difficult-to-characterize hydraulic resistances of the flow channels. Efficiency of surface irrigation systems varies tremendously because of variations in soil type, field uniformity, and management. Surface irrigation is often considered less efficient than sprinkler irrigation or micro-irrigation, because soil, not a pipe, is the water-conveyance system for surface irrigation (Table 1). Nevertheless, a well-managed surface irrigation system on a uniform soil with a runoff-reuse system can approach 90% application efficiency.

The surface irrigation process is comprised of four phases, starting with an advance phase in which the surface water stream propagates over the field from the upper point of entry to the bottom of the irrigated section. Next, during the soaking phase, inflow continues until a sufficient, pre-planned irrigation water depth has been introduced. After cutting off the water inflow, in the depletion phase, the depth of the surface stream is reduced while infiltration, and possibly runoff, also continue. At some stage, the soil surface below the stream is exposed and the recession phase begins, usually from the upstream end (recession can also occur at the leading edge of the water stream, or from its both ends, either prior to completion of advance or after). The time interval between the surface stream arrival and its departure at a given location is termed the infiltration-opportunity time.

Table 1 Potential on-farm application efficiencies of irrigation systems.

<i>System type</i>	<i>Application efficiency (%)</i>
Surface irrigation	
Furrow	60–75
Level basin	60–85
Border	60–80
Sprinkler irrigation	
Solid set	70–85
Set-move	60–75
Moving ^a	75–95
Traveling gun	55–70
Micro-irrigation^b	80–95
Subirrigation	50–80

^aIncludes center-pivot and linear move systems with also low-energy precision application (LEPA).

^bIncludes drip, microsprayers and microsprinklers.

Furrow irrigation

In furrow irrigation, used mostly for row crops (such as cotton, maize and sugar cane), water flows in evenly spaced furrows or corrugates that are typically 0.1–0.3 m wide, spaced at 0.75–2 m (Fig. 1), on fields with slopes of 0.1–3%. The lateral furrow spacing should be smaller for sandy soils and larger for clayey soils. Furrow lengths may range anywhere from tens to hundreds of meters, depending on the soil type, slope and crop type. Shorter furrows are commonly associated with higher uniformity of water application, but result in increasing potential for runoff losses. Usually, the furrows are run in the direction of the predominant slope, normal to the contours, although this should be avoided on steep slopes where soil erosion may be severe. Inflow to irrigation furrows may be supplied from gated pipes or ditches (earthen or concrete). Siphon tubes are frequently used to convey and regulate water flow from ditches to individual furrows (Fig. 2). By creating a siphon, water flows through the tube, over the ditch bank, and into the furrow, as long as the tube outlet is lower than the water elevation in the ditch. Furrow inflow rate is controlled by tube diameter and the elevation difference between the ditchwater level and tube outlet. Gated pipes distribute water to furrows through evenly spaced outlets on the pipes. With earthen ditches, water can flow through a breach or other opening in the ditch bank to individual furrows or a smaller feed ditch that distributes water to several furrows. It is much more difficult to regulate flow through a breach in an earthen ditch than through siphon tubes or pipe gates. Sometimes only one of two neighboring furrows is continuously or alternately irrigated during consecutive irrigations. It can take a considerable period of time for the water to reach the furrow bottom end (the advance phase), meaning water has been infiltrating for a longer period of (opportunity) time at the top end of the field. This results in poor water application uniformity with high application at the top end and lower application at the bottom end. In most cases, the water application uniformity can be improved by increasing the speed at which water advances along the furrow through increasing flow rates. Increasing the advance rate not only improves the uniformity, but also allows application at smaller irrigation depths. Water commonly flows in furrows for short durations of several hours to full days during an irrigation, but shorter or longer durations may be used depending on furrow length and soil, water, and management considerations. Inflow rates for individual furrows vary in the range of approximately 10–100 L min⁻¹, again depending on soil (texture, surface roughness), slope, field length, and management. Ideally, water should advance across the field in approximately 25% of the total irrigation time to irrigate the field uniformly. The goal is to minimize tailwater loss while still allowing sufficient soaking time at the lower end of the furrow. Since soil erosion increases as field slope and inflow rate increase, flow rate must be carefully managed on fields with steeper slopes (greater than 1%). Low inflow rates and long irrigation durations may be needed to apply the desired amount of water during an irrigation on soils with low infiltration rates. Conversely, higher inflow rates are often needed on fields with low slopes or high infiltration rates, for the water to flow across the field and uniformly irrigate its upper and lower ends.

The advance phase of surge furrow irrigation consists of a series of about five surges (or cycles, e.g., on for 1 h and off for 1.5 h), each advancing farther than its predecessor, to get the water to the end of the furrow. Surge flow irrigation is often more efficient and can achieve reduced infiltration rates and faster water advance down the furrow than conventional furrow irrigation, thus increasing the water application uniformity along the furrows and reducing the minimal water application depths.

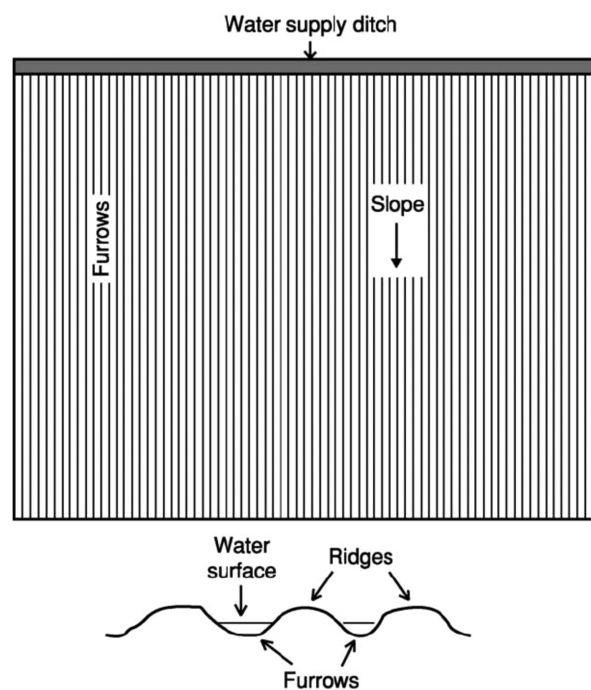


Fig. 1 Schematic of a furrow-irrigated field, where water flows in evenly spaced furrows or corrugates. From the 2005 edition.



Fig. 2 Furrow-irrigated sugar beet field, with water supplied by siphon tubes from a concrete ditch. From the 2005 edition.

Furrow irrigation requires lower capital investment, but greater labor than most other irrigation systems. Furrow irrigation is not well suited to automation as water flow rate must be adjusted for each furrow for each irrigation. However, automated valves and programmable controllers are used, including special surge (butterfly type) valves alternating the water supply to two sets of furrows.

Basin and border irrigation

Basin and border (sometimes termed 'level basin' 'border strip') irrigation systems are similar in that both involve a uniform sheet of water flowing over the soil. The main difference is that basin irrigation applies water to a nearly level field and may include ponding for extended time periods. With border irrigation, water flows between dikes that divide a sloping field into rectangular strips with free drainage at the end. The purpose of the dikes is to contain water as it flows across the field, unlike basin irrigation where the dikes pond the water (Fig. 3). In this respect, border strip irrigation can be considered as a hybrid of level basin and furrow irrigation.

Basins can be as small as a few square meters for a single tree or as large as several hectares with greater than 100 L s^{-1} inflow rates (Fig. 4). Basin size is a balance of soil infiltration rate, slope and water supply. Water depth in flooded basins varies about 5–20 cm, with typical depths of 10–15 cm. Efficient basin irrigation requires a level soil surface with uniform, not too coarse, soil texture and

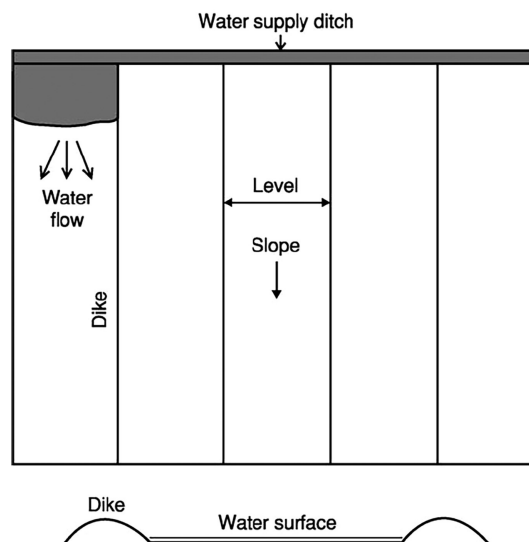


Fig. 3 Schematic of a border-irrigated field, where a uniform sheet of water flows between dikes. From the 2005 edition.



Fig. 4 Level Basin flood irrigation of wheat.

adequate water supply, so the basin is quickly and uniformly covered with water. If the basin is not level, the higher elevation areas will receive less water than the low areas. If the basin infiltration rate is too large, water will slowly advance, causing large differences in infiltration opportunity time within the basin. Rice is cultivated mostly, in either rain-fed or gravity-irrigated, permanently-ponded, flat paddies (or terraces in mountainous terrains).

A special type of basin irrigation is a drain-back level basin. Drain-back level basins have a series of parallel basins that receive inflow from a shallow, 5–10 m wide ditch. After the first basin is filled, a gate opens to start filling the adjacent basin, which is at a lower elevation. Water near the inflow end of the first basin drains back to the inflow ditch and flows to the next basin. This procedure is repeated until every basin has been irrigated. The drain-back phase improves uniformity by reducing the amount of water that infiltrates near the inflow end and initially increasing the inflow rate to the next basin, which increases the surface stream advance rate.

Border irrigation systems are better suited for sloping fields than basin systems because water flows between dikes rather than being ponded within basins. The irrigated areas between dikes may be 3–30 m wide and up to 400 m long. The field slope between dikes (perpendicular to water flow direction) should be nearly level so water flows uniformly down the field. The slope along the dikes can be similar to furrow irrigation, but border systems often have slopes less than 0.5%. Typical inflow rates vary from 10 to 100 L s⁻¹, but vary widely depending on the size of the basin or border, soil texture and slope. If water inflow is cut off before the stream has reached the bottom end of the field, the stream continues to advance for some time, if it is deep enough, and if the slope is large enough, and the infiltration rate, soil roughness, and vegetative drag are small enough. This operating strategy increases the irrigation efficiency in level basins and in sloping border strips, where it reduces the amount of runoff.

Border and basin irrigation require less labor than furrow irrigation because water is supplied to a larger area with a single outlet.

Sprinkler irrigation

Sprinkler (sometimes termed ‘overhead’) irrigation applies water to soil by sprinkling or spraying water through the air on to the soil surface or on to the plants canopy. Water is pressurized and delivered to the irrigation system by a mainline pipe, which is often buried, so it does not interfere with farming operations. Sprinkler irrigation is used for a wide variety of plants, including field crops, orchard trees, lawns (golf courses), and pasture. Sprinkler systems are also installed for frost and freeze protection, for cooling plants, for the control of airborne dust and minimization of wind erosion, and for land application of wastewater (Hoffman et al., 2007; Keller and Bliesner, 1990).

Sprinkler Irrigation systems can be categorized based on their spraying mechanisms and on their portability. Many kinds of sprinkler devices are in use, such as impact sprinkler, spray pads, rotating pads and LEPA (low energy precise application) systems. They use a nozzle to control their discharge rate, its diameter and shape of its orifice, and the water pressure at the nozzle control the flow rate, size of water drops and the wetting pattern. These and the height of the sprinkler device determine the diameter of coverage and the distribution of water within the wetted region. The impact sprinkler in which the sprinkler head, driven in a

circular motion by the force of the outgoing water, pivots on a bearing on top of its threaded attachment nut, was invented in the US in the 1930s, and quickly found widespread use.

Three main portability categories of sprinkler irrigation systems are solid-set, set-move, and self-moving.

Solid-set systems may be installed for a single season for certain field crops or permanently for lawns, orchards, or permanent crops. Set-move systems are manually or mechanically moved to another part of the field after the irrigation set is complete in the present location. Self-moving systems such as center pivots or traveling guns apply water as the system slowly travels through the field.

Small drops are prone to drift in windy conditions and to evaporative losses. Large drops, on the other hand, resist wind drift, however, their impact energy may breakdown soil aggregates and form a seal (crust) on the soil surface which reduces infiltration and leads to runoff.

The application rate, determined by the sprinklers flow rate and their areal coverage, should not exceed the soil infiltration rate to avoid runoff, accounting for also the ground slope, i.e., lower on steeper slopes. Its recommended range is in between about 2 mm h^{-1} in 12% slope clayey soils to 50 mm h^{-1} in flat sandy soils. The time of operation per set should be selected to satisfy crop water requirements throughout the irrigation interval while minimizing deep percolation losses.

Sprinkler irrigation is often more efficient than surface irrigation, because water application is more controlled (Table 1). In hot or windy areas, however, sprinkler irrigation can have significant losses from evaporation and wind drift. Maintenance is also important for efficient sprinkler irrigation. Worn nozzles and leaking pipe connections reduce application uniformity and system efficiency.

Solid set sprinkler systems

Most solid-set sprinkler irrigation systems are designed to apply frequent (e.g., daily), small amounts of water to meet plant water needs. Water application rates can vary from approximately $2\text{--}6 \text{ mm h}^{-1}$ for field crops to $5\text{--}30 \text{ mm h}^{-1}$ for lawns applications. Capital costs are greater for solid-set systems than other sprinkler systems, because the entire irrigated area must be equipped with sprinklers and pipe. However, permanently installed systems can be automated to reduce labor and allow irrigation at any hour of the day, which reduces the risks for plants to be stressed. When properly designed, solid-set systems have high application uniformity. While solid-set systems are most commonly used with lawns, and landscape and permanent crops, these systems are also used for some high-value annual crops with low tolerance for water stress.

Solid-set system designs are as varied as the applications: small sprinklers may irrigate 20 m^2 , or large, gun-type sprinklers may be spaced 50 m apart. Plastic pipe is frequently used for buried applications, but it is also used in some aboveground applications. Aluminum, rigid and flexible plastic pipes (50–100 mm in diameter) are often used for field crops when the system is installed after planting and removed before harvest (Fig. 5). Most systems are divided into zones so a portion of the area is irrigated at one time. Solid-set systems for use in frost and freeze prevention, however, must be designed to water the entire area simultaneously.

The water application depth of a single sprinkler decreases with distance in a pattern in between elliptical and triangular which necessitates proper overlap between the wetted areas to achieve a good irrigation uniformity. Usually, the spacing between the laterals is larger than between the sprinklers along the lateral, creating a rectangular sprinkler arrangement. In such arrangement, and if the application pattern of a single sprinkler is distorted by winds, orienting laterals perpendicular to the prevailing wind direction is generally recommended for best overlap of the wetted areas. In no wind conditions recommended sprinkler and lateral spacing are about 45% and 65% of effective wetted diameters of the sprinklers (at average pressure along the lateral). At wind speeds exceeding 16 km h^{-1} , these both percentages reduce to about 30%.



Fig. 5 Solid-set sprinklers near Rio Vista, California, US.

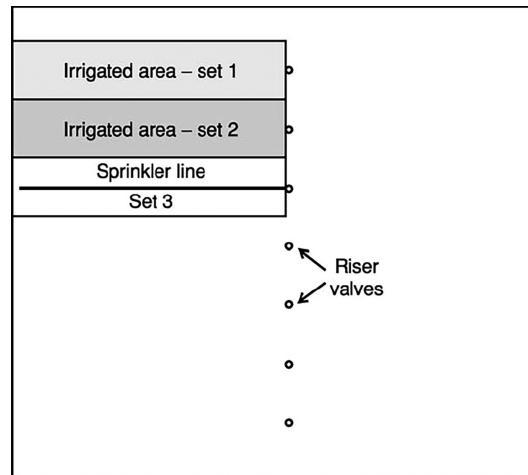


Fig. 6 Schematic of a set-move-irrigated field, where the irrigation system is moved to the adjacent area after completing an irrigation set. From the 2005 edition.

Set-move sprinkler systems

Set-move sprinkler irrigation systems are designed to apply water slowly during the irrigation set (e.g., $4\text{--}6\text{ mm h}^{-1}$), which often lasts 8–24 h. After completing the irrigation set, the sprinkler system is moved to an adjacent area for the next set (Fig. 6). Adequate water has to be applied during an irrigation set to meet crop water needs until the system is moved back to the area, often in 7–10 days.

The common types of set-move irrigation systems are hand-move and towline systems. Hand-move systems can be a single sprinkler or a line of sprinklers. A line of hand-move sprinklers, sometimes called handlines is typically composed of 9- or 12-m long pieces of 75- or 100-mm-diameter aluminum pipe with a sprinkler mounted on one end or in the center. Individual pipes are connected to form an irrigation line, usually not more than 400 m long. After an irrigation set is completed, the line is disconnected, and each piece is moved by hand 10–20 m to the next set. A slight variation to the handline is the ‘dragline’ (‘towline’ or ‘skid-tow’ or ‘end-pull’). These systems have special connections between sprinkler pipes that allow the irrigation line to be pulled by a tractor to the next set. They were developed to reduce labor required to reposition laterals. The rigid couplers of the laterals are fitted with skids or a pair of wheels so the line can be pulled with a tractor. Stabilizers, outriggers or wheel-mounts, are used to keep the lateral from tipping and keep the sprinklers upright. The flexibility of the pipe and the couplers permit the lateral to curve slightly while being pulled to its new position.

Self-moving sprinkler systems

Self-moving irrigation systems include center-pivot, linear-move, and traveling gun systems. Center-pivot and linear-move systems are similar in design and appearance. These systems consist of one or more spans of sprinkler pipe (usually galvanized steel or aluminum) elevated by ‘A-frame’ wheeled towers. Span length varies from 30 to 65 m. The towers, powered by hydraulic or electric motors, elevate the sprinkler pipe 2–4 m above the ground (Fig. 7).

The center pivot has a stationary pivot point so the wheeled towers move in a circle, creating a typical circular pattern with, complete or sectioned, cultivated circles of same or different diameters and various colors viewed from above (Fig. 8). Crops are planted either in straight rows or in circles, conforming to the travel of the wheeled towers. Water and power are supplied to the system through the pivot point. Originally, most center pivots were water-powered. Later on, these systems were replaced by ones driven by electric motors, usually having a motor mounted at each tower. The outside set of wheeled towers sets the pace for the rotation. The inner towers between two segments use angle sensors to detect when the bend at the joint exceeds a certain threshold. When the angle is too large, the wheels rotate to keep the segments aligned. A typical center pivot in the USA (where they were invented in 1940) has eight spans, a total length of approximately 400 m ($\frac{1}{4}$ mi), and irrigates 50–60 ha. Typical periods for a full rotation last about 3 days, the outer edge of the system moving at a speed of about 1 m min^{-1} . Some center pivots are equipped with an additional corner span (arm) or an end-gun that irrigate additional areas in between the main wetted circular area and its inscribing square.

Center pivots are extremely popular, because water is uniformly applied to a large area with little labor. Furthermore, once a circular field has been irrigated, the center pivot is in position to start the next irrigation. System cost per irrigated area is reduced by increasing the total length of a center pivot, because irrigated area per unit length increases with distance from the pivot point. A 50 m span at the pivot point irrigates 0.8 ha, while a 50 m span that is 350 m from the pivot point irrigates 11 ha. Consequently, water application rate increases with distance from the pivot point, usually by increasing the nozzles diameter. Application rates often exceed the infiltration rate of the soil, generating runoff after consuming some surface storage in flat fields, which may limit



Fig. 7 Center pivot with drop tube sprayers.



Fig. 8 A satellite image of circular fields, of either $\frac{1}{4}$ or $\frac{1}{2}$ miles diameter, characteristic of center pivot irrigation, Kansas, US. Credit NASA.

center pivots use in fine-textured, clayey soils. Center pivots can also irrigate fields with undulating terrain that are difficult or impossible to irrigate by surface-irrigation methods; however, these conditions can create additional management challenges.

Early center pivots had impact sprinklers mounted on top of the irrigation pipe that required 5–6 atm (500–600 kPa). Most new systems have low-pressure nozzles (0.7–2 atm, 70–200 kPa) mounted on tubes ('goosenecks') that extend below the irrigation pipe (drop sprinkler heads), so nozzle height varies from 1 to 3 m above the soil (Fig. 7), thus limiting evaporative and wind drift losses. Drop tubes can also be used with drag hoses or bubblers that release the water directly on the ground between crops, sometimes associated with the construction of small dams along the furrow length ('furrow diking').

Manufacturers make numerous types of low-pressure nozzles with fixed or spinning spray plates (LEPA) that provide a wide range of application rates and water-droplet sizes to meet field conditions and operator preferences. A common feature to all nozzle types is a pressure regulator, which maintains constant nozzle pressure as the system travels across a field with varied elevation. Many modern center pivots feature GPS devices for applying VRI.

Linear-move systems (also termed 'lateral move') have a control unit on one end, or in the center on longer systems, that moves the towers in a straight line to irrigate rectangular shaped fields. Power is supplied by an electrical drag cord or by an engine-powered generator mounted on the control unit. Water is typically supplied to the drive unit by a drag hose connected to a buried or aboveground pipe. Drive units can be equipped with a pump so water can be supplied from an open ditch flowing parallel to the travel direction, but this is not commonly used, because it requires a nearly level ditch. System cost per irrigated area is reduced by increasing the distance the system travels. Since hose length is limited to approximately 150 m, the system can move 300 m before the hose is connected to the next riser. Similar to set-move and center-pivot systems, adequate water must be applied to meet crop needs until the linear-move can irrigate the area again.

A manual, old version of a linear-move system and an upgraded version of a handline, still in use, is a wheel line (or 'side roll') comprised of a series of aluminum pipes (100–125 mm in diameter), usually each with a large-diameter wheel of 1.5–2 m mounted in its center or on end, and self-levelling sprinklers along its length, between the wheels (Fig. 9). Water is supplied at one end using a hose. The pipes serve both for conveying the water and as an axle for rotating all the wheels. A motorized drive system (often located near the center of the wheel line) rotates the clamped-together pipe sections as a single axle, rolling the whole wheel line. Manual adjustment of individual wheel positions may be necessary if the system becomes misaligned. After sufficient irrigation depth has been applied to one strip of the field, the hose is removed, the water drained from the system, and the assembly rolled, so that the sprinklers are moved to a different position across the field. This system is less expensive than modern center pivot and linear-move systems, but much more labor-intensive to operate as it does not travel automatically across the field. Wheel line systems are limited (by the wheels' radius) in the height of crops that can be irrigated.

A traveling gun has a large-capacity sprinkler on a cart that is pulled across the field by a cable, by the water-supply hose or by a tractor. Guns are similar to rotor sprinklers, except that their larger nozzle diameters are in the range of 10–50 mm, they generally operate at very high pressures of 2.75–9 atm (275–900 kPa) and their flow rates are 3–76 L s⁻¹.

These systems irrigate an area 50–100 m wide and up to 400 m long. A traveling gun can be considered a set-move system because water is applied as the cart moves across the field and then the system is moved to another area in the field for the next irrigation set. For cable-tow systems, a winch on the cart winds the cable, pulling the cart and a soft hose across the field. A hose-reel system pulls the cart as a plastic hose is wound around a reel on a trailer anchored at the end of the run. The reel or winch is powered by an engine or a water turbine. Smaller versions of traveling guns are available for irrigating sports fields and small pastures.



Fig. 9 Side-roll (wheel line) irrigation system in Idaho, US.

Lawn sprinkler systems

Lawn sprinkler systems are permanently installed in residential lawns, gardens of public buildings, public parks, cemeteries, and golf courses. Usually, most of the components of these irrigation systems are buried, to not interfere with the lawn users and for aesthetic purposes. The two main types of sprinklers used in lawn irrigation are pop-up spray heads and rotors. Spray heads have a fixed spray pattern, while rotors have one or more streams that rotate. Spray heads are used to cover smaller areas, while rotors are used for larger areas. On lawns, pop-up sprayers or sprinklers are usually installed with their tops flush with the soil surface. On activation of the irrigation, the head will pop up out of the ground and water the lawn until the valve closes, the water pressure in the lateral line declines, and the sprayer (sprinkler) head retracts back into the ground. In flower beds or shrub areas, sprinklers are usually mounted on above ground risers. Garden irrigation systems usually consist of few zones divided by their type of plants and water emitters, possibly including also zones with drip irrigation, bubblers, or other types of equipment besides sprinklers. Nowadays, most lawn and garden irrigation systems operate automatically using an irrigation controller and electric solenoid valves.

As opposed to the permanently-installed systems described above, many types of moveable (traveling) hose-end sprinklers are also used for watering lawns and gardens. Some have a spiked base temporarily anchoring them in the ground, while others have a sled base designed to be dragged while attached to the hose, usually connected to an outdoor water faucet.

Micro-irrigation

Another type of pressurized irrigation is micro-irrigation (MI or 'localized irrigation'). Water is applied frequently (daily, or even few times a day—'pulsed irrigation') at low flow rates directly to the plants root zone with minimal evaporation, deep percolation and runoff losses. Water drips from emitters in plastic pipe or tape, or bubbles or sprays from small emitters that only wet a portion of the soil surface (Dasberg and Or, 1999; Hoffman et al., 2007; Lamm et al., 2006; Nakayama and Bucks, 1986). These systems are typically automated, which facilitates the frequent water application that maintains optimum soil-water contents near the plants, reduces deep percolation losses and increases water use efficiency. Filtration is important for MI, because mineral deposits, algae and bacteria can plug the narrow openings on drip emitters, bubblers, and micro-sprayers (several optional sand separators, media and screen (or disc) filters are depicted in Fig. 10). The required level of filtration depends on the size of the water paths (labyrinths and orifices) of the water emitters and on the irrigation water quality. Chemical treatment may also be necessary to reduce mineral or biofilm deposits.

MI is popular for permanently installed systems that irrigate trees, vineyards, orchards, and shrubs, but in the last decades its application for field crops and vegetables irrigation is increasing. MI systems are appropriate for undulating terrains and fields with irregular shapes, providing the highest field-scale, water distribution uniformity. Most MI emitters are also available as pressure-compensating devices. MI can be adapted to soils of both very low and very high infiltration rates, and to almost any

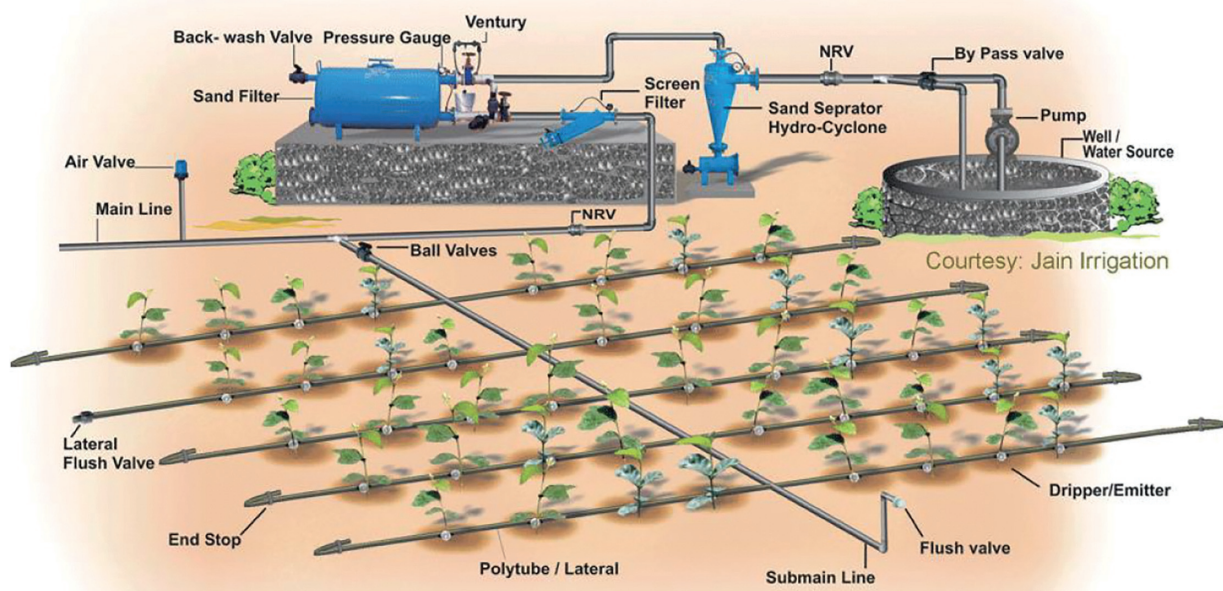


Fig. 10 Layout of a drip irrigation system and its main components.

cropping technology and climatic conditions. However, since the installation costs of MI systems are higher than those of surface and sprinkler irrigation systems, at present MI is used for irrigating mostly high-value crops, i.e., vegetables and perennial crops, mainly orchards and vineyards. With increasing demands and competition on dwindling water resources and the need to minimize environmental consequences of irrigation, MI is expected to play a more important role in the future.

Drip irrigation

Drip irrigation (DI, or trickle irrigation) emitters (drippers) are either integrated (in-line) within polyethylene pipes (laterals, driplines) at regular intervals or attached to the outside (on-line, 'button drippers') of the pipe at desired locations. In addition to the filters mentioned above, other system components include the main, submain and lateral lines, pressure regulators, automatically or manually operated water valves, air (and vacuum) relief valves, end stoppers, and flush valves (Fig. 10). Modern automatic systems include also an irrigation controller used for controlling the solenoid valves of the various irrigated sections and filter back flushing, and a fertilizer application system with a backflow prevention device. Drippers regulate their water flow rates by friction and turbulence head losses created by their narrow labyrinth-type flow paths and orifices. Typical flow rates vary from 0.5 to 8 L h⁻¹. Pressure-compensating emitters maintain a constant flow rate as pressure varies from approximately 0.7 to 2 atm (70–200 kPa), and a non-drain feature keeps the driplines water-full at the end of irrigation, preventing also vacuum and suction of soil particles into the drippers.

Drip tape is usually thin-walled (0.1–0.5 mm, recyclable) plastic tubing (10–30 mm in diameter), with outlets at 10–60 cm intervals. Flow rates usually vary from 1 to 4 L h⁻¹ m⁻¹. Typical operating pressures for drip tape is 0.35–1 atm (35–100 kPa).

In orchards and gardens, trickle rings, i.e., circular driplines with evenly distributed drippers are sometimes placed around the base of the trees or shrubs.

Microtubing irrigation is another type of drip irrigation common in gardening, nurseries and greenhouses. Flow rates are regulated by the plastic tubing's length and diameter, and stakes are usually attached to the end of the tubing to anchor it to the ground. Sometimes the (split) tubes are also used with drippers ('spider drip').

In Subsurface Drip Irrigation (SDI) driplines or drip tapes are installed below the soil surface at depths of 10–50 cm, where there is less opportunity for damage. SDI also reduces evaporation losses and improves weed suppression, facilitates shallow soil tillage and other planting, spraying and harvesting operations. Thus, the buried driplines and drip tapes can be used for several seasons or retrieved after a single season, depending on crop types and farming practices. SDI is more suitable for use with treated wastewater (TWW) than on-surface DI. On the other hand, plant germination and establishment can be challenging in sandy soils with too larger dripline burial depths. Sometimes, also with on-surface DI, a special germination irrigation with usually a sprinkler irrigation system is required. An alternative practice with on-surface DI is to shift the driplines to the planting rows for the germination irrigations. Additional disadvantages of SDI are potential dripper plugging by intruding plant roots and rodent damage. In many circumstances yield increases have been observed under SDI when compared to surface, sprinkler, and even on-surface DI systems.

Driplines and drip tapes can be regarded as water point or line sources depending on the ratio between the drippers' spacing and the soil capillary length; if smaller than one, it resembles a line source. The soil capillary lengths of most agricultural soils vary between approximately 10 cm for sandy soils to 100 cm for structureless, clayey soils with common values of 20–40 cm for loams and fine sands.

The agronomic design of drip irrigation systems involves decisions regarding the spacing between the drippers along driplines and between driplines and the burial depths of SDI driplines, all should be smaller in coarse texture, sandy soils, and for plants with small root systems (DIDAS: <https://app.agri.gov.il/didas/>). The spacings between driplines and between drippers, and the number of driplines per planting bed or tree row should correspond to also the planting configuration and density (Fig. 11). Irrigation frequency should be usually higher in sandy soils and for plants with small root systems, and irrigation dosing should account for the plants water requirements and for proper salt leaching of the localized wetted bulb hosting the active root system. Drip irrigation is often combined with fertigation and plastic mulching, further reducing evaporation and increasing irrigation and fertilizer application efficiency. In addition, DI minimizes the incidence of many diseases that are spread through water contact with the foliage and highly-saline water drops injury in sprinkler irrigation.

Bubblers and microsprayers

Bubblers, microsprayers and other non-drippers low-pressure, MI devices (microsprinklers, foggers, and jets) are often used for irrigating trees, shrubs, or ornamental plants, applying water to larger areas than drippers. Bubblers apply a small stream of water to the soil surface at flow rates of up to 250 L h⁻¹ irrigating a few meters diameter areas, depending on orifice size, type, and pressure. Usually, a small basin (pit) is used to control the distribution of the ponded water. Microsprayers (and microsprinklers) apply a fine spray or mist with similar flow rates and wetted areas as bubblers. Nozzle sizes typically range from 0.5 mm to 2 mm and plugging problems are usually less common as compared to DI. The MI emitters are typically connected to polyethylene pipes with a barbed connector and are attached to hard plastic risers or mounted on stakes. As for drip irrigation, the other MI devices are used to irrigate mostly sandy soils of limited lateral, capillary water flow.



Fig. 11 Drip irrigation of a young mango plantation with two driplines for a tree row.

Subirrigation

Subirrigation applies water below the soil surface by raising the water table into or just below the plant root zone, so that water flows upward into the root zone by capillarity (Hoffman et al., 2007). As capillary rise is usually limited to few tens of centimeters, subirrigation can only be used on relatively flat lands, where also permanent shallow water tables already exist. Thus, it is typically used in conjunction with subsurface drainage, mostly in humid regions without salinity problems. Subsurface drainage lowers the water table and removes excess water through open ditches or perforated pipes (Skaggs and van Schilfgaarde, 1999; Waller and Yitayew, 2016). Water-table depth can be controlled by regulating water levels in open field drainage ditches or by controlling the head on subsurface drainage pipes with adjustable weirs. During wet periods, the water table is lowered so the root zone remains unsaturated. During dry periods, water is pumped into the drainage system to raise the water table and provide additional water for plant growth. In some situations, drained water can be stored for use when irrigating.

Fertigation (chemigation)

Fertigation is the processes of applying water-soluble, NPK and micronutrient, fertilizers via the irrigation system, and chemigation refers to other agro-chemicals such as insecticides, herbicides, and fungicides (Bar-Yosef, 1999). The chemicals may be added, with the aid of a dedicated tank and liquids injectors such as diaphragm pumps, piston pumps, or Venturi aspirators, constantly whenever the system is irrigating or at specified times, when needed. A proportional (to the water discharge rate) injection mode maintains constant concentrations in the irrigation water. It is common to stop the fertigation before irrigation ends in order to clear the fertilizers out of the irrigation system. The flexible and continuous application of plant nutrients, combined with water application, both corresponding to plant needs, increase their nutrient use efficiency, thus, decreasing the amounts of fertilizers (and other agro-chemicals) needed and consequently their leaching into groundwater and surface water resources.

Fertigation (chemigation) is possible with mostly drip and sprinkler irrigation as these irrigation systems can be managed to quickly and uniformly irrigate small amounts of water, which communicates with continuous application of fertilizers (and other agro-chemicals). Drip irrigation is more suitable for fertigation (chemigation), since the direct contact of the sprinkled (sprayed) water with the plants canopy can be problematic. Sometimes, in rainy regions, DI systems are used just for fertigation, even if water application is not required.

Subsurface drip irrigation systems can be used for also improving soil aeration by air injection with a compressor (or blower) or by adding air or oxygen (nano) bubbles or hydrogen peroxide to the irrigation water (a practice termed oxygation or oxygenation) (Ben-Noah and Friedman, 2018).

Salinity hazards and management

Even if using non-saline, good quality irrigation water, evaporation and selective water uptake causes salinization of the soil solution, which may cause osmotic stress to plants and impedes their water uptake, growth and yield production (Hoffman et al., 2007). Salinity problems are more closely associated with climatic conditions, soil and water chemistry than with the type of irrigation system; however, the irrigation system can accentuate salinity problems and affect how salinity is managed. Basin and border irrigation can uniformly leach salts from the soil when infiltration is uniform. Sprinkler irrigation can also uniformly leach salts, but sprinkling can injure sensitive plants when highly saline water is applied to foliage. Furrow irrigation leaches salt from soil below the furrows, while increasing salt concentrations in the beds (ridges). Tilling before planting crops tends to minimize salinity problems by mixing salts in the soil. Salt concentration tends to increase radially from a point-source dripper or laterally from a line source. Well-managed drip irrigation can minimize salt-induced stress by maintaining a high soil-water content with frequent irrigations.

Using all irrigation methods an additional irrigation depth should be applied, the leaching requirement (LR, or leaching fraction, LF), either in each irrigation or periodically, for leaching the saline soil solution. Yet, excessive leaching is undesirable, as it wastes water, and as it introduces more salts into the soil, which may eventually reach groundwater and surface waters, degrading their agricultural and domestic use quality and harming ecosystems. In semi-arid regions, with also a rainy season, a possible, beneficial salinity management strategy is to accumulate salts in the root zones during the dry irrigation seasons and to (partly) rely on rainfall events to leach the salts and prevent inter-annual salinity buildup.

Capillary rise of saline groundwater from shallow water tables is another source of root zone salinity. These circumstances are usually associated with water contents being too high and aeration problems (hypoxia). The solution to both salinity and poor aeration problem is to control the water table depth with an artificial drainage system of open ditches or buried perforated pipes.

Another problem often associated with salinity in arid and semi-arid regions is soil sodicity, caused by using irrigation water rich in sodium and resulting in destruction of soil structure and reduction of the soil infiltration capacity, which limits the applied irrigation rates.

Choosing an irrigation method

Choosing an irrigation method, among the above-described various methods and their variants, is a difficult task. The right selection depends on soil, water, and climatic conditions, as well as crop types, user knowledge and preference, capital and operating costs, and infrastructure availability. No system is best for all situations. In a broader sense, an irrigation method includes not only the system's hydraulic hardware, but also 'software' aspects of operation mode, technology and support infrastructure. These may include guidance for system design, installation, operation, and maintenance, and for irrigation scheduling and other irrigation-related agricultural practices. In addition to the grower benefits and welfare, regional, hydrological and environmental, considerations should also pay a role, within the imposed constraints (Hoffman et al., 2007).

Some typical advantages and disadvantages of irrigation systems are shown in Table 2. Sprinkler or micro-irrigation are often better choices than surface irrigation on sandy soils, where excessive deep percolation is a problem. Surface irrigation may be better in arid, windy areas, where wind and evaporation losses can be significant. Surface irrigation offers less control of application depth, so small, frequent irrigations are not practical for water-sensitive crops, which are better suited to micro-irrigation, solid-set, or center-pivot systems.

Conclusion

Irrigation increases yields and ensures the reliability of crop production—approximately 40% of the world's food is grown on irrigated land, which accounts for only about 20% of the world's land used for crop production. Various irrigation methods have been developed over time to meet the irrigation needs of certain crops in specific areas. The three main methods of irrigation are surface, sprinkler, and micro-irrigation (including drip). In surface irrigation, water flows over the soil by gravity. Sprinkler irrigation applies water to soil by sprinkling or spraying water droplets from fixed or self-moving systems. In micro-irrigation, used at present in only few percent of irrigated land, small water volumes are applied by dripping, bubbling, or spraying. With increasing demands and competition on dwindling water resources and the need to minimize environmental consequences of irrigation, sprinkler, micro and especially drip irrigation are expected to play a more important role in the future. In addition to transition from surface to sprinkler or drip irrigation, irrigation system improvements, better irrigation design and scheduling, and more intense, proxy and remote, soil and plant sensing should be practiced.

Table 2 Advantages and disadvantages of common irrigation methods^a.

System type	Advantages	Disadvantages
<i>Surface irrigation</i>	Avoidance of canopy wetting; suitable for short-term water supplies; low capital, energy and maintenance costs	Low distribution uniformity; high evaporation, deep percolation and runoff losses; conveyance losses; large irrigation depths; unsuitable for sloppy terrain; requires initial landforming operations; not suitable for sandy soils
Furrow	Water flows in small channels: less evaporation losses	High labor and initial landforming operation costs; requires experience to operate; possible runoff and deep percolation losses; possible soil erosion; salt accumulates on ridges
Level basin	Less labor costs than furrow	Ponded water; high evaporation losses; possible plant diseases
Border strip	Less labor and less runoff than furrow; easier to manage infiltration depth	Water flows over entire soil surface; requires repair of ridges and supervision during irrigation
<i>Sprinkler irrigation</i>	Good distribution uniformity; smaller irrigation depths; less deep percolation and runoff losses; allows frost and freeze protection and cooling; simulates natural rainfall; better suited for mechanization and automation	Evaporation losses; wind drift; not suitable for steep slopes; not suitable on soils with low infiltration rates; risk of soil surface sealing, runoff and soil erosion; canopy and fruit wetting; problematic with saline water; requires higher water pressure; high capital, energy and maintenance costs
Solid set	Fits odd-shaped fields; low labor costs	High capital cost; system may interfere with field operations
Set-move	Lower capital costs compared to solid set	More labor than other sprinkler systems; poor uniformity in windy conditions; greater irrigation depth
Self-moving ^b	High distribution uniformity; less wind drift effect; low capital and labor costs	Not suitable for odd-shaped fields; circular center-pivot—irrigated fields leave unused land
Traveling gun	Less interference with soil tillage and agricultural operations; low energy and labor costs	Low distribution uniformity; not suitable for clayey soils; possible soil surface sealing and runoff; not suitable for odd-shaped fields; requires high water pressure
<i>Micro-irrigation</i>	High distribution uniformity; supports frequent applications; low irrigation depths; maintains favorable root zone conditions; allows fertigation; reduced evaporation and runoff losses and soil erosion; suitable for sloping terrain and all soil types; suitable for greenhouses and nurseries; suitable for automation	Not suitable for densely planted crops; requires clean water or treatment and filtration; interferes with soil tillage and agricultural operations; vulnerability to mechanical and animal damages; high capital and maintenance costs
Micro-sprayers	Allows frost and freeze protection and cooling	Evaporation losses; wind drift; canopy wetting
Bubblers	Avoidance of canopy wetting; good weed suppression; improved trafficability	Poor flow rate control and distribution uniformity
On-surface drip	Allows mulching; avoidance of canopy wetting; good weed suppression; suitable for fertigation; improved trafficability	Restricted root zone; low soil water storage; salt accumulation near plants
Sub-surface drip	Lowest evaporation losses; allows mulching; less interference with shallow soil tillage and agricultural operations; avoidance of canopy wetting; best weed suppression; best for TWW application; best for fertigation, chemigation and oxygation; saves driplines removal labor; best trafficability	Restricted root zone; low soil water storage; salt accumulation above drippers; problematic seed germination and plant establishment; increased deep percolation losses; limited dripper application rates; limitations for crop rotation; no visual inspection; potential rodent damage

^aThe general information related to the three main irrigation systems is given in italics.^bIncludes center-pivot and linear move systems with also low-energy precision application (LEPA).

References

- Bar-Yosef B (1999) Advances in fertigation. *Advances in Agronomy* 65: 2–77.
- Benami A and Ofen A (1984) *Irrigation Engineering*. Haifa: Irrigation Engineering Scientific Publications, Israel Institute of Technology.
- Ben-Noah I and Friedman SP (2018) Review and evaluation of root respiration and of natural and agricultural processes of soil aeration. *Vadose Zone Journal* 17. <https://doi.org/10.2136/vzj2017.06.0119>.
- Bjorneberg DL and Sojka RE (2005) Irrigation/methods. In: Hillel (ed.) *Encyclopedia of Soils in the Environment*, 1st edn, pp. 273–280. St. Amsterdam: Elsevier.
- Brouwer C, Prins K, Kay M, and Heibloem M (1988) *Irrigation Water Management: Irrigation Methods*. Training Manual No. 5 Rome: Food and Agriculture Organization of the UN.
- Cuenca RH (1989) *Irrigation System Design: An Engineering Approach*. Englewood Cliffs, NJ: Prentice-Hall.
- Dasberg S and Or D (1999) *Applied Agriculture: Drip Irrigation*. New York, NY: Springer-Verlag.
- Hoffman GJ, Evans RG, Jensen ME, Martin DL, and Elliott RL (eds.) (2007) *Design and Operation of Farm Irrigation Systems*, 2nd edn St. Joseph, MI: American Society of Agricultural and Biological Engineers.
- Israelsen OW (1950) *Irrigation Principles and Practices*, 2nd edn New York, NY: John Wiley and Sons, Wolters Kluwer Health, Inc.
- Keller J and Bliesner RD (1990) *Sprinkle and Trickle Irrigation*. New York, NY: Van Nostrand Reinhold.
- Lamm FR, Ayars JE, and Nakayama FS (eds.) (2006) *Microirrigation for Crop Production: Design, Operation, and Management*. Amsterdam: Elsevier.
- Nakayama FS and Bucks DA (eds.) (1986) *Trickle Irrigation for Crop Production: Design, Operation, and Management*. Amsterdam: Elsevier Science.
- Pereira LS and Trout TJ (1999) Irrigation methods. In: van Lier HN, Pereira LS, and Steiner FR (eds.) *CIGR Handbook of Agricultural Engineering. Land and Water Engineering*, vol. I, pp. 297–379. St. Joseph, MI: ASAE.
- Rohwer J, Gerten D, and Lucht WW (2007) *Development of Functional Irrigation Types for Improved Global Crop Modelling*. PIK Report No. 104 Potsdam: Potsdam Institute for Climate Impact Research.

Skaggs RW and van Schilfgaarde J (eds.) (1999) *Agricultural Drainage. Monograph 38*. Madison, WI: American Society of Agronomy.
 Walker WR and Skogerboe GV (1987) *Surface Irrigation. Theory and Practice*. Englewood Cliffs, NJ: Prentice-Hall.
 Waller P and Yitayew M (2016) *Irrigation and Drainage Engineering*. Switzerland: Springer Science.

Relevant websites

<https://www.fao.org/aquastat/en/>—AQUASTAT, Food and Agriculture Organization of the United Nations (FAO).
<https://www.asce.org/>—American Society of Civil Engineers.
<https://www.asabe.org/>—American Society of Agricultural and Biological Engineers.
<https://www.nass.usda.gov/AgCensus/>—Census of Agriculture, USDA-NASS.
<https://app.agri.gov.il/didas/>—DIDAS-Drip Irrigation Design and Scheduling. Israeli Agricultural Research Organization (ARO).
https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Farm_and_Ranch_Irrigation/—Farm and ranch irrigation survey, USDA-NASS.
<https://icid-ciid.org/home>—International Commission on Irrigation and Drainage (ICID).
<https://www.gao.gov/products/gao-20-128sp>—Irrigated Agriculture: Technologies, Practices, and Implications for Water Scarcity, U.S. Government Printing Office.
<https://www.irrigation.org/>—Irrigation Association
<https://en.wikipedia.org/wiki/Irrigation>—Irrigation, Wikipedia

Soil–plant–atmosphere continuum

John M Norman^{a,*}, Martha C Anderson^b, and William P Kustas^b, ^aUniversity of Wisconsin-Madison, Department of Soil Science, Madison, WI, United States; ^bUSDA-ARS, Hydrology and Remote Sensing Lab, Beltsville, MD, United States

© 2023 Elsevier Ltd. All rights reserved.

This is an update of J.M. Norman, M.C. Anderson, SOIL–PLANT–ATMOSPHERE CONTINUUM, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 513–521, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00416-1>.

Introduction	625
Fundamental principles	625
Water in the soil–plant–atmosphere continuum	626
Small-scale observations and model building	628
Upscaling SPAC observations from plot to regional and global scales	630
Advancing SPAC research: Integration of earth observations, modeling, and ground measurements	631
References	632

Abstract

In this chapter, we describe research for understanding how living systems function in the soil–plant–atmosphere continuum (SPAC). Through process-based modeling and earth observations it is evident that whole ecosystems function as living entities. Process-based models and observations at the cellular to plant canopy level of water, energy and carbon exchange have given us a significant understanding of how living systems function through the SPAC—creating and maintaining their biomes. However, a major and continuing challenge is how SPAC processes scale up to the regional and global scales. We provide an insight to this scaling challenge by describing how water movement, key to sustaining life on Earth, is being quantified at multiple scales, and by discussing the need to incorporate remote sensing as a means of bridging from plant level water use to whole ecosystems and up to global scales. To accomplish this scaling, SPAC research requires integrating multi-sensor remote sensing data over the full electromagnetic spectrum, incorporating commercial satellite observations, and employing not only process-based models but data assimilation, and machine learning techniques with cloud computing.

Key points

- Water is an essential ingredient for life on Earth.
- The interaction of this water with energy is the cornerstone of our understanding of how soil-plant-atmosphere continuum (SPAC) interacts with its environment.
- The greatest value of complex SPAC models lies in improving our understanding of key processes and using this knowledge to derive intermediate models with fewer parameters and wider applicability.
- Scaling from the small spatial scale of direct measurements (meters to tens of meters) up to large regions (thousands of kilometers) requires use of numerous remote-sensing systems.
- Studies of the SPAC are interdisciplinary (involving research in areas that are not claimed by existing disciplines) and multidisciplinary (requiring collaboration among scientists from numerous disciplines).
- The future in advancing SPAC research will be strongly tied to the integration of spatial and temporal information from Earth observations, as this is the only means of bridging both spatial and temporal scales from plant level processes up to the global scale.

Abbreviations

ALEX	Atmosphere-Land EXchange model
ET	EvapoTranspiration
LTAR	Long-Term Agro-ecosystem Research network
LTER	Long-Term Ecological Research network
MODIS	MODerate-resolution Imaging Spectroradiometer
SIF	Solar-Induced Fluorescence
SPAC	Soil-Plant-Atmosphere Continuum
TSEB	Two-Source Energy Balance model
UAV	Unmanned Aerial Vehicle
VIIRS	Visible Infrared Imaging Radiometer Suite

Introduction

The dominant distinguishing feature of the Earth, in comparison with all other celestial bodies we observe, is the presence of life. Humans, as a part of life on Earth, attempt to understand all of the life system, albeit from an anthropocentric perspective. Lovelock (1987) proposed a perspective that was unsettling to some and appealing to others when he suggested that the entire Earth may be functioning as a living organism. As our integrative understanding of living systems within the soil–plant–atmosphere continuum (SPAC) grows, the idea that a whole ecosystem can function as a living entity becomes more acceptable.

Consider a terrestrial ecosystem on land: The most essential ingredient for life on Earth is water and most of the water is not on the land. However, a rapid, long-distance system for transporting water from oceans to land is available through the atmosphere, much like the xylem system in a tree, which transports water long distances (relative to cell sizes) from the soil to the leaves high above the ground. This long-distance atmospheric transport system is driven unceasingly by the uneven, solar heating of the Earth. On the land, where many organisms reside, water moves much more slowly through the soil than through the atmosphere and this provides an essential storage reservoir that maintains the critical continuous flow of water to plants in the presence of fluctuations in the atmospheric supply. The rapid atmospheric transport system that is so critical to moving water from the ocean to the land also fluctuates wildly over the life span of most organisms. The combination of a small-storage, highly fluctuating system for rapid global transport in the atmosphere and a large-storage, stable, slow-flow buffer in the soil redistributes water over the Earth to sustain life on land. This is a highly structured system, like an organism. Furthermore, the living systems exert a strong influence on this larger system to enhance their functioning, just as we humans modify our environment to improve our lot.

Water is the most obvious component of the SPAC, and the interaction of this water with energy is the cornerstone of our understanding of how life interacts with its environment. Every schoolchild learns the water cycle: evaporation of liquid water from surfaces supplies water vapor to the atmosphere, where air currents lift this water to heights that cool it and cause condensation of the water vapor back to the liquid form of droplets, which fall back to the Earth to be recycled. The extraction of energy from the surface by evaporation and the release of energy to the atmosphere by condensation represent the primary means by which the energy from the sun ultimately warms the atmosphere. This 'latent heat,' which is associated with the change of phase of water between liquid and vapor states, and not a temperature change, provides the key feedback between water and energy cycles that stabilizes the soil–plant–atmosphere system, permitting it to sustain life.

Materials such as water flow through the ecosystem seamlessly, moving from place to place, serving innumerable functions, and consuming or releasing energy to achieve a phase suitable to the medium that contains it. This is the SPAC; 'continuum' here does not refer to the material storage forms so apparent to our eyes such as plant leaves or roots, lakes, raindrops, the soil itself, or even the air we breathe; rather, it refers to the unimpeded flow of many forms of material and energy throughout an elegant system for sustaining life on land. Animals are as much a part of this system as plants, but in this article, plants are our emphasis. This ceaseless flow of material and energy sustains life at every level; from the molecular interactions that sustain and mutate the genetic 'memory,' to cellular processes that make up whole organisms, to the global level of a living, breathing planet.

Living systems cycle matter through the SPAC and themselves to create and maintain their structure and extract energy for powering their life functions. All material necessary for sustaining life is cycled on all spatial levels from subcellular to global. The connectedness and functioning of the SPAC can be illustrated by reference to water, the most critical mass constituent to life on Earth. However, understanding the SPAC through water will necessarily involve some considerations of at least carbon and energy.

Fundamental principles

Two general principles are indispensable for studying the SPAC: (1) the conservation principles that take the form of mass and energy 'budgets,' and (2) the transport principles that relate the flow of some quantity to the difference or gradient of other quantities that influence or 'force' the flow and describe the 'state' of the exchange process.

The principles of conservation of mass and energy are the backbone of integrative studies of the SPAC. Because mass and energy can take many forms as they move throughout the SPAC, and even interact with each other, budgets are constructed to quantify the stores and flows of important life-enabling constituents such as water, carbon, or energy. A budget is simply the application of the conservation principles (mass or energy) to a specific system that must be carefully defined. This system may be defined as a leaf, a community of plants, or even the entire Earth. Like balancing a bank account, a budget is a formal statement of the following: Incoming quantity minus outgoing quantity equals the increase in storage of the quantity in the system. In the case of water (W), carbon (C), and energy (E), the budgets become the following:

$$\begin{aligned}W_{IN} - W_{OUT} &= \Delta W \\C_{IN} - C_{OUT} &= \Delta C \\E_{IN} - E_{OUT} &= \Delta E\end{aligned}$$

The intertwined budgets of water, carbon and energy inflow, outflow and storage within the soil–plant–atmosphere system are shown in Fig. 1. Understanding the SPAC through the role of water will involve simultaneous consideration of carbon and energy budgets on a spatial scale appropriate for living systems.

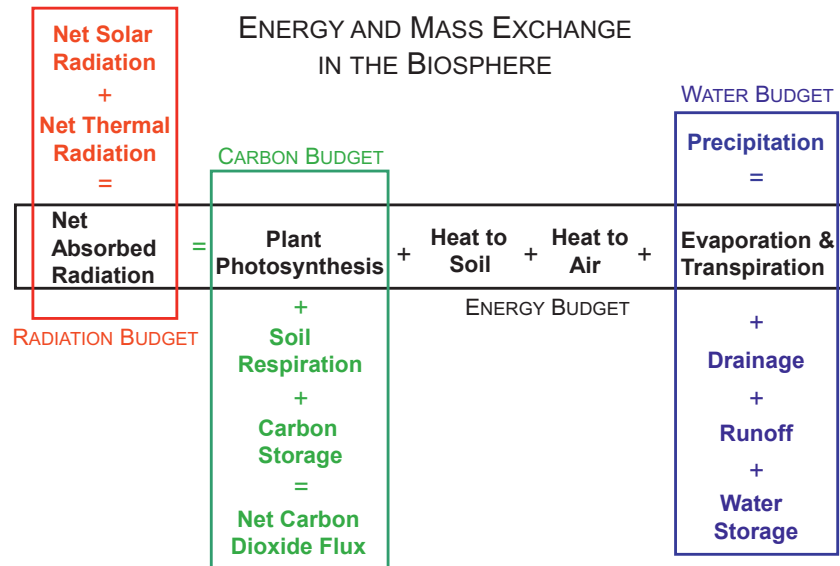


Fig. 1 The interlocking energy and mass budgets for water, carbon, and energy in the soil–plant–atmosphere continuum. Altering individual components can cascade through the system and affect many seemingly unrelated factors.

Consider the energy budget: energy can exist in many forms as it flows through the soil–plant–atmosphere system, and the energy budget represents a way to keep track of the energy, no matter what its form. The rate of energy flow (joules per second = watts) per unit of area (meters squared) is a convenient way to describe the energy budget for a particular land area, because the size of the area does not matter and the emphasis can be on the energy exchanges. Some fraction of the radiant energy from the sun and surroundings can be absorbed by a leaf (this is referred to as the net radiation) and be transformed to the following: (1) sensible heat energy that is convected away from the leaf by moving air that is cooler than the leaf; (2) Sensible heat that is conducted into the leaf itself altering the temperature of the leaf; (3) latent energy in molecules of water that are converted from liquid to vapor inside the leaf as they are transpired (evaporated) from the leaf; and (4) biochemical energy in the form of organic compounds such as sugars and starches created through photosynthesis and other physiological processes to sustain life. These components of the leaf energy budget are shown in Fig. 1.

The transport principle simply states that the flow of some quantity such as water is equal to an appropriate ‘driving force’ divided by a resistance to transport exerted by the various media through which the quantity moves. The transport principle is more difficult to use than the conservation principle because the ‘driving forces’ and transport resistances depend on the characteristics of the system and mechanism of transport. For example, the diffusion of liquid water depends on the water potential difference across a hydraulic transport resistance, whereas the transport of water vapor depends on the water vapor pressure difference across a diffusive transport resistance. In addition, movement of liquids or vapor by mass flow is ‘driven’ by pressure differences across transport resistances. In general, diffusion is an effective mode of transport over small distances, but mass flow can move materials and energy over longer distances. Heat transport is driven by temperature differences. The relevant thermal resistance depends on whether the heat is moving through a solid such as soil (conduction, which is analogous to diffusion), or through a moving fluid such as air (convection, which is analogous to mass flow and is a much more rapid transport mechanism than diffusion). Heat transport by radiation depends on still-different formulations. Thus, much research on the SPAC relates to determining appropriate transport resistances and state variables for characterizing the flow of important quantities throughout the system. A complete understanding of this soil–plant–atmosphere system is a formidable task, but such an understanding will improve the ability of humans to make choices that will sustain a diverse and resilient ecosystem, upon which our survival depends. The fundamental concepts of energy and mass conservation and transport relations constitute the theoretical backbone of environmental biophysics (the study of the SPAC) (Campbell and Norman, 1998).

Water in the soil–plant–atmosphere continuum

The smallest spatial scale that is usually associated with the SPAC is a field plot (square meters to thousands of square meters) from the bottom of a root zone (a few meters deep) to a few tens to hundreds of meters above the top of the vegetation. Although individual studies may use smaller systems such as individual leaves or potted plants to reveal mechanisms, the scale of at least a small field is required to encompass most of the essential components of the natural SPAC and yet be accessible to direct measurement. Researchers generally use formulations nearly identical to those measured in field plots to represent idealized patches of the Earth’s surface thousands of square kilometers in size; then, they combine these patches to understand better the

human influence on global climate processes. Typical time scales for small-scale studies vary from minutes to months and years. On these spatial and temporal scales, researchers attempt to measure all the important flow and state variables, which are used to validate models that can be applied in systems that cannot be measured directly. After all, society cannot afford to measure everything everywhere, but will support measuring some things in many places and many things in a few places. In such a system with unbounded complexity, quantification must proceed with an attitude of humility and openness to unexpected possibilities.

The transport and storage of water in the SPAC are shown in Fig. 2. The flow of liquid water from the soil into the plant depends on the plant having a lower water potential than the soil; thus, the soil water potential determines the availability of soil water for diffusion into the plant. The water potential is the amount of energy (joules) that can be obtained from moving a mass of water (kilograms) to a pool of pure water at atmospheric pressure and specified elevation; and in the soil and plant it is usually negative, indicating that energy must be expended to move the water. As soils dry, the water potential decreases, making it more difficult for plants to obtain the water they need to grow. The soil water potential is related to the water content of the soil through the ‘moisture-release curve,’ which varies from soil to soil. Moisture-release curves have been measured for many different soil types. Typically, soils can provide an amount of water to plants from storage that is equivalent to about 5% (sands) to 20% (silt loams) of the plants’ rooting depth. Thus, for a plant rooted to a depth of 1 m, soils can provide approximately 50–200 mm of water to the plant. With evapotranspiration rates of 2–10 mm per day, soils can store enough water to sustain plants for weeks to months without rainfall or irrigation.

Once water enters the plant through its roots, it moves through the plant vascular system to the leaves if the leaves have lower water potentials than the roots. Since the transport resistance to mass flow in the stem xylem is usually less than the transport resistance to diffusion through the root, and the flows through both root and xylem must be equal according to conservation principles, the water potential difference from the soil to the root xylem will usually be greater than the water potential difference from the root xylem to the leaf.

Water is lost from plants through pores in the leaf surfaces, called stomata. Evaporation from cells just below the stomatal pores (referred to as transpiration) draws water from the leaf cells, which in turn are hydraulically connected to a continuous water column down to the roots. As water is transpired through the stomata, additional water is drawn through the xylem to replace it, very much like suction applied to a straw in a soda. Of course, this evaporation that occurs just below the leaf surface consumes most of the Sun’s energy that is absorbed by the leaf, because water vapor molecules carry away latent energy that was not contained in the liquid water that came up from the roots. Although plants lose most of their water through these stomatal pores, it is unavoidable because these pores must open for carbon dioxide to enter into the leaf and supply photosynthetic tissue with the carbon needed to ultimately build the plant itself and supply animals with a source of life-sustaining food. Furthermore, leaf transpiration can also serve to moderate leaf temperature changes as the environmental air temperature varies over a wide range; this allows leaves to

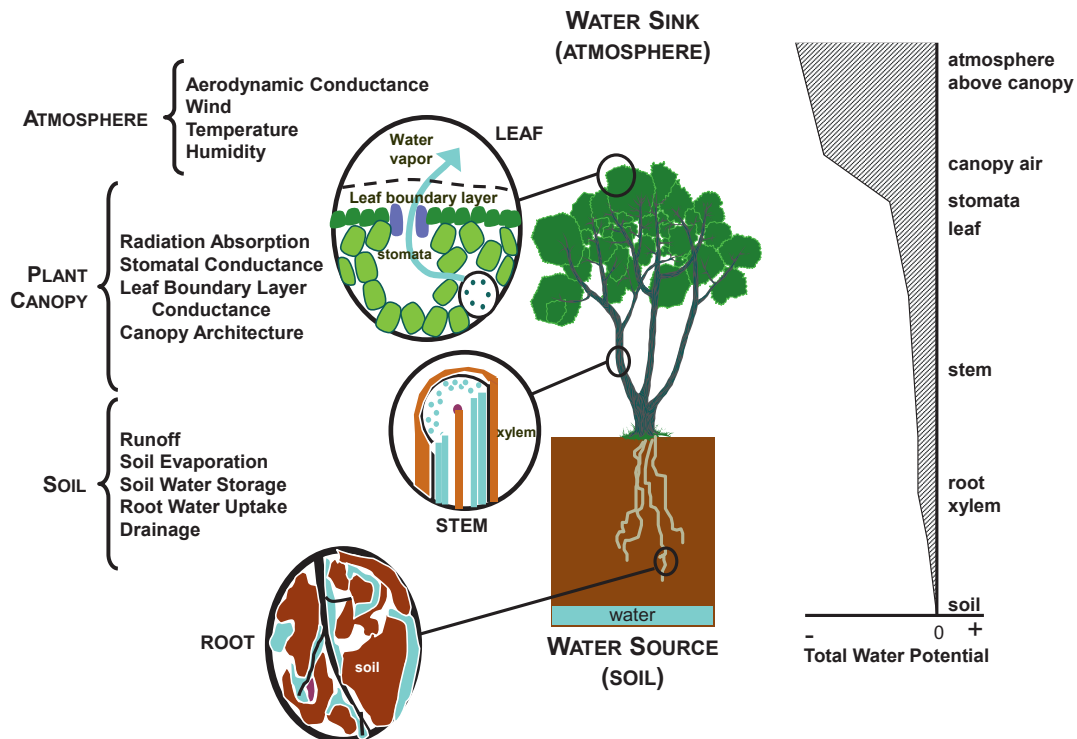


Fig. 2 Water movement and storage in the soil–plant–atmosphere continuum. The graph on the right indicates the decrease in water potential that occurs as water moves from the soil to the atmosphere. On the left the three components of the soil–plant–atmosphere system are listed with the major factors for each that should be considered in a model of intermediate complexity.

function above air temperature under cool conditions, and below air temperature when it is hot so that leaf biochemistry, including photosynthesis, operates closer to optimal conditions. While the flow of liquid water from the soil to the leaf is proportional to water potential differences, the flow of water vapor from inside the leaf to the leaf surface is proportional to water vapor pressure differences. At the water-air interface inside the leaf, a near-equilibrium exists between the liquid water and the water vapor in contact with it. Although different processes are used to describe water movement in different parts of the system, the flow of water is continuous across these discontinuities of processes.

A very small amount of the water that passes from the soil to the atmosphere through the plant remains in the plant as storage as the plant grows (less than 5%). An even smaller amount of the hydrogen and oxygen in the water taken up from the soil remains in the plant in the form of organic molecules synthesized by the plant.

Water that exits the stomata must pass through a thin, still-air layer adjacent to the leaf, called the leaf boundary layer, and be transported through the canopy space by turbulent mixing along the path of decreasing water vapor pressure. Ultimately the water vapor exits the plant canopy, passes through the planetary boundary layer in the lower few thousand meters of the atmosphere, and is lifted high into the atmosphere, where it condenses into clouds and eventually falls back to the Earth as precipitation to repeat the cycle.

Small-scale observations and model building

In plant-environment studies, research derives and validates the transport principles on small, manageable systems in the laboratory or in field plots where direct measurements are possible. These principles are then assumed to apply at the ecosystem and global levels, where direct measurements are not possible, to obtain estimates of the important flow or state variables to quantify human impacts.

Consider the measurements that are necessary to understand the cycling of water in the SPAC. Precipitation falls on a field and either runs off the soil surface to streams or infiltrates into the soil itself (studied in the discipline of hydrology). The infiltrated water may be stored in the soil matrix (soil physics), evaporated from the soil surface (micrometeorology), transpired from plants (plant physiology), or drain below the plant root zone. Water that drains below the plant roots becomes part of the groundwater system (hydrogeology) and may reenter rivers (hydrology). Plants are sustained by the water stored in the soil, and the quantity of this storage depends on the type and condition of the soil. Nearly all the water taken up by plants is transpired to the atmosphere at a rate that depends on solar radiation absorbed, weather variables, and the amount and health of the vegetation. The amount of vegetation is estimated from the fraction of ground area covered by plants (vegetative cover). By keeping track of the daily water budget (Fig. 1 and Table 1), the amount of water stored in the soil can be estimated, and in turn the influence of this stored water on direct evaporation from the soil and plant transpiration (together referred to as evapotranspiration) can be quantified. The maximum amount of water available to plants depends not only on soil properties, but also on plant rooting depth. Plant physiologists have measured rooting depths for the major classes of vegetation so that, given vegetation type and soil maps, the amount of water available to the plants for growth and transpiration can be estimated. Depth to groundwater must also be known in regions of shallow water table to ascertain whether phreatophytic extraction may be occurring by plant roots.

Table 1 contains an example of a 40-day-long water budget for a corn crop that begins about a month after emergence, during the period when the crop is growing rapidly. The incoming solar radiation column represents the maximum amount of evapotranspiration that could occur if all the solar radiation were used to evaporate water; of course, some radiation is reflected so the actual evapotranspiration is the sum of the transpiration and direct evaporation from the soil, which is approximately one-third of the maximum possible evapotranspiration. The maximum amount of water that can be stored in the rooting zone is 200 mm for this soil, and clearly the corn crop is using water faster than the rainfall is replenishing it, so that at the end of 70 days of growth approximately half of the maximum soil water storage has been used up. At this point the water depletion in the soil will begin to cause the growth rate of the corn to decrease unless there is more rainfall. This buffering capacity of soil is crucial to growing crops in continental climates all over the world.

Measurements of water transport and storage are difficult and expensive to make, so they can only be done at a few sites under a few conditions. Models that capture the fundamental conservation and transport principles are therefore essential to applying results of limited SPAC studies to a wider range of conditions than could possibly be measured directly. A wide array (hundreds) of computer models is available for quantifying the role of water in the SPAC and the range of applicability of such models depends on having a minimal, faithful representation of critical factors in each component of the SPAC; namely, of the soil, the plant, and the atmosphere. SPAC models can be divided into three classes: simple, intermediate, and complex.

Simple SPAC models are the easiest to create, generally emphasize the knowledge base of one discipline with a narrowly focused purpose and are the most limited in applicability. For example, the Simple Inverse Yield Model (Morgan et al., 2003), for precision agriculture, and the CERES Maize model (Jones and Kiniry, 1986), for crop yield and water-use prediction, neglect the effect of vegetation roughness and wind on evapotranspiration, reflecting their agricultural origins and limitations in applicability. There is also a simple micrometeorological model of evapotranspiration that omits soil processes, reflecting its atmospheric origins and limited applicability.

Intermediate SPAC models incorporate at least a minimal consideration of key soil, plant, and atmospheric components to provide wide applicability without undue complexity. These intermediate models incorporate atmospheric factors (radiation, wind,

Table 1 Simulation of corn growth and water use from 31 to 70 days after emergence, during a rapid growth phase.

<i>Days after emergence</i>	<i>Rainfall (mm)</i>	<i>Radiation (mm day⁻¹)</i>	<i>Vegetative-cover fraction</i>	<i>Biomass (kg m⁻²)</i>	<i>Transpiration (mm day⁻¹)</i>	<i>Soil evaporation (mm day⁻¹)</i>	<i>Water stored in soil (mm)</i>
31	0	11.02	0.17	0.047	1.16	2.91	177
32	0	10.10	0.21	0.061	1.26	2.53	173
33	0	6.76	0.24	0.073	0.97	1.13	171
34	0	11.70	0.26	0.083	1.83	1.53	168
35	0	6.93	0.31	0.101	1.23	0.72	166
36	0	11.17	0.34	0.113	2.13	0.99	163
37	0	7.62	0.38	0.135	1.61	0.57	161
38	0	6.04	0.41	0.151	1.35	0.39	159
39	0	11.63	0.43	0.164	2.72	0.67	155
40	0	10.94	0.47	0.191	2.75	0.5	152
41	20	5.60	0.51	0.219	1.49	0.74	170
42	10	6.5	0.53	0.234	1.78	0.82	177
43	0	11.41	0.55	0.252	3.20	1.45	173
44	11	5.82	0.59	0.284	1.70	0.62	181
45	0	9.4	0.60	0.301	2.81	0.95	178
46	0	11.61	0.63	0.329	3.55	1.07	173
47	0	10.70	0.66	0.364	3.36	0.89	169
48	0	11.03	0.69	0.398	3.54	0.83	164
49	0	11.49	0.71	0.433	3.76	0.78	160
50	0	11.31	0.73	0.471	3.76	0.69	155
51	0	11.1	0.75	0.509	3.75	0.62	151
52	0	10.16	0.77	0.546	3.45	0.52	147
53	0	9.88	0.78	0.581	3.39	0.46	143
54	0	12.13	0.79	0.614	4.19	0.38	139
55	0	10.51	0.81	0.656	3.65	0.24	135
56	0	10.83	0.82	0.693	3.79	0.20	131
57	0	12.19	0.8	0.731	4.28	0.19	126
58	0	12.76	0.84	0.77	4.50	0.17	122
59	0	9.08	0.85	0.818	3.21	0.10	118
60	20	6.70	0.85	0.851	2.38	0.19	136
61	0	11.61	0.86	0.874	4.13	0.32	131
62	0	11.98	0.86	0.916	4.27	0.31	127
63	0	11.04	0.87	0.958	3.94	0.27	123
64	0	11.12	0.87	0.998	3.97	0.26	118
65	0	11.05	0.87	1.037	3.95	0.25	114
66	0	12.14	0.88	1.077	4.35	0.27	109
67	0	10.8	0.88	1.120	3.89	0.23	105
68	0	10.71	0.88	1.159	3.84	0.23	101
69	0	9.87	0.88	1.198	3.54	0.20	98
70	0	8.32	0.88	1.233	2.99	0.17	94

The solar radiation is in units of millimeters per day of evaporation that could occur if all solar radiation were to evaporate water. Vegetative cover increases as the crop grows and crop biomass increases. Soil evaporation tends to decrease as the soil surface dries out after rainfall and as vegetative cover increases; evapotranspiration (transpiration plus soil evaporation) exceeds rainfall so soil water storage is gradually depleted.

temperature, humidity, and precipitation), vegetation factors (plant type, vegetative-cover fraction and height, rooting depth), and soil properties (soil evaporation, root-zone water transport and storage). Examples of Intermediate SPAC models include the Two-Source Energy Balance (TSEB) model (Norman et al., 1995), for remote sensing applications, and the ALEX (Atmosphere–Land EXchange) model (Anderson et al., 2000), for local site-specific studies.

Complex SPAC models incorporate most of the processes known to be important at small (leaf to field) scales and can be applied to a wide range of cover types and climatic conditions. Because they are very detailed, they generally involve a large array of coefficients related to specific plant and soil types. For this reason, complex models are often not well-suited to regional scale applications, where these many parameters need to be defined at all locations. The greatest value of complex SPAC models lies in improving our understanding of key processes and using this knowledge to derive intermediate models with fewer parameters and wider applicability. Thus different, but related, models tend to be used to assess phenomena on different spatial scales.

An example of this use of a complex SPAC model is the derivation of the TSEB and ALEX models from the complex Cupid model (Norman and Campbell, 1983). The Cupid model predicts plant–atmosphere exchanges of carbon, water, and energy in a very detailed mechanistic way, and it requires dozens of species-specific coefficients related to plant physiology. By studying the behavior

of the Cupid model, the ALEX model could be created using only a few carefully chosen physiological coefficients and yet achieve a surprising level of generality compared with its parent, Cupid. Not only does ALEX link growth and water use in a way nearly as fundamental as Cupid, but it does so with a widely used ‘light-use efficiency’ coefficient that is already tabulated for most vegetation types around the world (Norman and Arkebauer, 1991). The result of this process is a robust model suitable for continental-scale applications.

Complex SPAC models have also been used to study the relation between the architecture of vegetation canopies and their remotely sensed signatures from aircraft and satellites. Since remote sensing is the only way to characterize vegetative cover at the regional and continental scales, we must understand how to interpret remotely sensed land observations to accommodate the profound influence of vegetation density on energy and material exchanges in the SPAC. For example, in Table 1, the increased importance of transpiration (and extraction of soil water to great depths by plant roots) over evaporation from the soil surface (water extraction to only a few centimeters of soil depth) is apparent as the vegetation cover increases from 17% to almost 90% during the 40 days of growth.

Determining the level of complexity appropriate for an application depends on the objective of the endeavor; typically, the simplest model that provides adequate results is the model of choice. Our understanding of energy and water flows in the SPAC is sufficiently well developed to extend this knowledge to larger scales, where direct measurements are only rarely, if ever, possible and evaluate implications of human activities on ecosystem health.

Upscaling SPAC observations from plot to regional and global scales

The impact of humans on the global ecosystem has become so profound that extending our understanding of biophysical processes from small fields to large regions is becoming critical for our survival. The extension of process-level understanding of energy and material flows in the SPAC to larger scales imposes a severe constraint that is not present in plot-scale studies; namely, that the allowable input quantities are limited to state variables that can be monitored over large areas. Only a few state variables are routinely measured widely from ground-based networks across the world and thus can be considered generally available for monitoring the SPAC. For example, near-surface weather variables (air temperature, humidity, wind speed and direction, precipitation, and cloud cover), solar radiation, atmospheric carbon dioxide, soil maps, vegetation-type maps, river and stream monitoring, and satellite images are routinely obtained at many locations in many countries of the world. This means that these are the main quantities that can be assumed to be generally available everywhere (not always true, but a reasonable assumption). All other inputs required for SPAC models must be computed from fundamental principles or from quantitative relations derived from validation experiments at small scales. This pursuit has been exceedingly challenging, but systematic research over decades in numerous disciplines has identified many of the critical processes and surface parameters, and integrative computer modeling has synthesized much of this information so that monitoring water transport and storage on continental scales may now be possible.

Fig. 3 gives an example of how information about SPAC functioning can be transferred across spatial and temporal scales, here using the TSEB surface energy balance modeling system to estimate evapotranspiration (ET). From Fig. 3 it is apparent that scaling from the small spatial scale of direct measurements (meters to tens of meters) up to large regions (thousands of kilometers) requires use of numerous remote-sensing systems from aircraft and Unmanned Aerial Vehicles (UAVs) to various satellites with very different characteristics. The goal of monitoring daily ET at field scales over large regions may require fusion of data from multiple sensor systems. For example, field measurements at the plot and flux-tower scale have been used to develop simplified descriptions of turbulent heat, mass and momentum exchange based on land-surface temperature and vegetation cover, surface roughness parameters related to vegetation height, and atmospheric stability conditions derived iteratively with the surface fluxes. Surface variability at these scales can be well-mapped at the meter scale using UAVs and low-flying aircraft, but these monitoring systems are too costly to deploy with the time and space sampling required for regional applications. Instead, the insights collected at these finer scales can be tested with coarser resolution, but more routine, imagery collected by operational satellite systems like Landsat and Sentinel-2 at the 10–30 m spatial scale. However, the 5–18 day revisit intervals of these current medium-resolution satellites may miss critical changes in the land-surface status—for example, due to rainfall, irrigation, rapid onset drought, agricultural management, or phenological changes in the vegetation. This is even more likely in humid regions where cloud cover reduces the temporal frequency of usable imagery in the visible to thermal wave bands. Moderate resolution (500–1000 m) polar orbiting sensors like the Moderate Resolution Imaging Spectroradiometer (MODIS) or Visible Infrared Imaging Radiometer Suite (VIIRS) provide near daily information, but at scales that may average over multiple management units or landcover types in a single pixel. In many cases, fusion of medium and moderate resolution ET image timeseries can better approximate field-scale fluxes at daily timesteps (Xue et al., 2022). Geostationary satellites provide excellent temporal sampling (15 min) and broad spatial coverage (thousands of kilometers) with kilometer sized pixels, and measure diurnal variations in solar radiation and surface temperature, as well as atmospheric profiles of temperature and humidity—all critical drivers of SPAC functioning that can help to improve model estimates at the finer spatial scales as well (Anderson et al., 2019). At the continental and global scales, the diagnostic information from remote sensing models can be used to assess the performance of prognostic SPAC models used in climate and biophysical studies, and to identify processes that may not be well-represented within these more complex modeling systems (Hain et al., 2015).

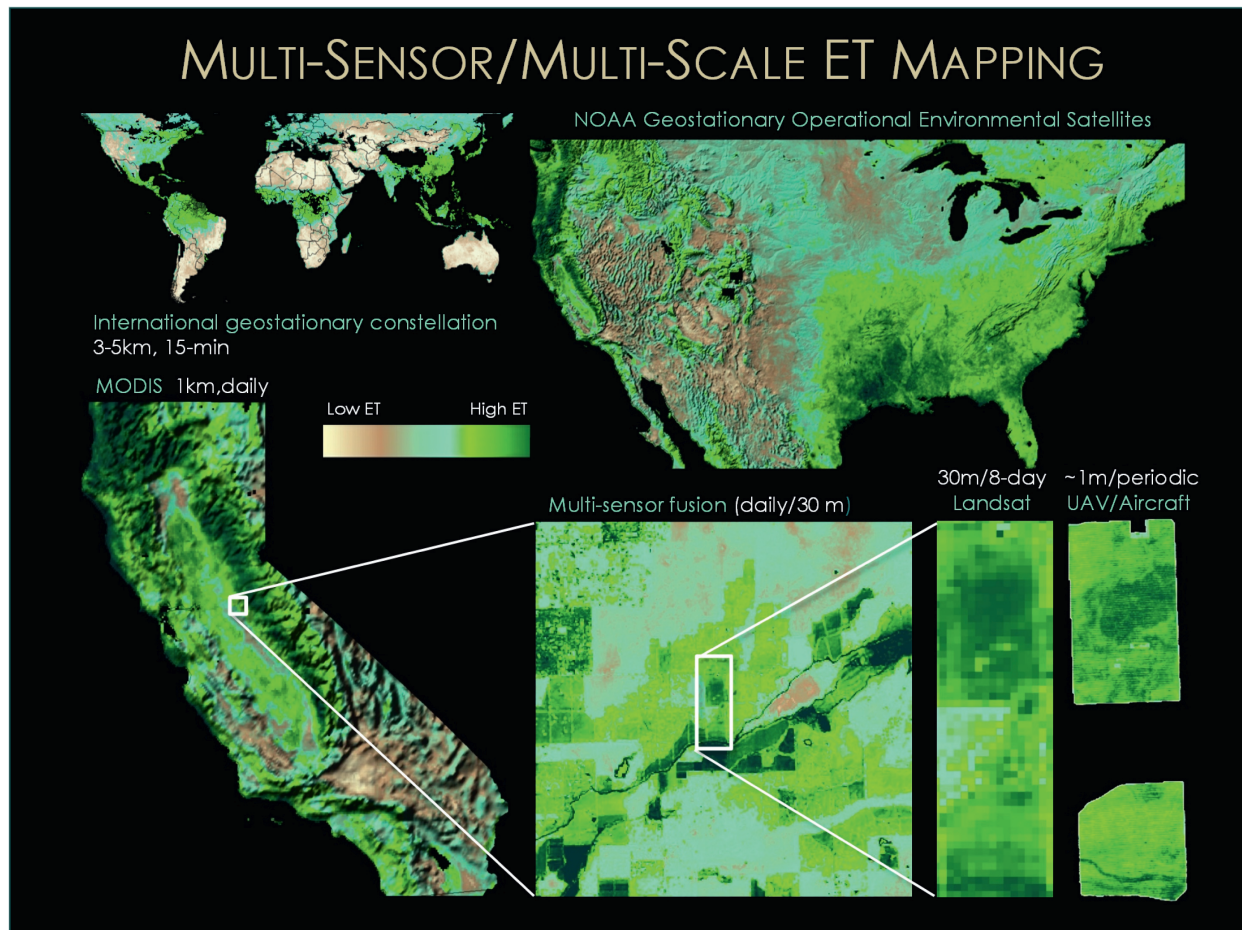


Fig. 3 Mapping ET from plot to global scales using multi-scale thermal remote sensing data within a TSEB-based energy balance modeling system.

Advancing SPAC research: Integration of earth observations, modeling, and ground measurements

Studies of the SPAC are interdisciplinary (involving research in areas that are not claimed by existing disciplines) and multi-disciplinary (requiring collaboration among scientists from numerous disciplines). For example, to understand the influence of vegetation on weather forecasts involves the following activities: meteorologists study the effect of the surface energy and water exchanges on the lower few thousand meters of the atmosphere called the planetary boundary layer; physiologists characterize the dependence of photosynthesis and stomatal processes on atmospheric and soil environmental factors; soil scientists measure the heat- and water-holding characteristics of the soil as well as carbon-exchange processes; micrometeorologists relate vegetation-canopy characteristics and atmospheric variables to the exchange of heat, mass, and momentum with the atmosphere; and hydrologists consider the runoff characteristics of the soil surface and the recharge of groundwater by percolation. The study of vegetative-canopy architecture, within-canopy processes, and the interaction between remote sensing and vegetative cover are key interdisciplinary areas that are not associated with any particular discipline. These multi- and interdisciplinary boundaries can be formidable obstacles to an integrative or holistic quantification of the SPAC, and because so little is known about interactions across these boundaries it is often easier to justify assumptions than engage the unknown knowledge gap.

The use of data from remote sensing retrievals of multiple surface properties across the electromagnetic spectrum is leading to fundamental advances in modeling flux exchanges in the SPAC. For example, while significant efforts with passive and active microwave remote sensing focused on soil moisture retrieval and accounting for the effects of vegetation (e.g., plant water content) on the soil moisture signal (Dorigo et al., 2010), vegetation water content is now seen as useful information in understanding SPAC processes such as plant stomatal behavior to both atmospheric demand and soil moisture conditions (Konings et al., 2017) and with potential applications in quantifying the terrestrial carbon cycle (Konings et al., 2019). Remote sensing of solar-induced fluorescence (SIF), which is a proxy for photosynthesis, is now available from satellites and is being used for constraining global transpiration estimates from SPAC models and other land surface states and fluxes (Jonard et al., 2020; Pagán et al., 2019). Combining the full spectrum of Earth observations is already being utilized in agronomic applications for crop yield estimation (Guan et al., 2017).

Advances in SPAC research in quantifying ET with remote sensing have led to the development of application tools in water resource management; namely, the OpenET {<https://openetdata.org/>, (Melton et al., 2021) effort in the United States, and in Europe the SEN-ET tool {<https://www.esa-sen4et.org/>, (Guzinski et al., 2020) developed by European Space Agency's Sentinel Application Platform. Future satellite missions from both government and commercial enterprises are likely to greatly advance our capability to model the SPAC (McCabe et al., 2017). Indeed, for ET and plant monitoring, the promise of high resolution daily thermal and multispectral infrared observations in the near future from commercial satellites (<https://www.hydrosat.com/>) could provide information necessary for near-real time monitoring of plant water use and stress and lead to a greater understanding of SPAC dynamics.

Clearly, to make such breakthroughs will also require having significant resources devoted to ground truth SPAC-related data. This involves supporting long-term observational networks used by multiple scientific disciplines, such as the Long-Term Ecological Research network (LTER, Kratz et al., 2003) or the U.S. Department of Agriculture experimental watersheds (Goodrich et al., 2021) and Long-Term Agro-ecosystem Research network (LTAR, Bean et al., 2021), providing continuous and reliable data to assess trends in SPAC processes and to validate remotely sensed retrievals from current and future satellite missions. There also need to be dedicated field campaigns, which use remote sensing to address SPAC-related science questions through intensive sampling that long-term networks are not able to provide. Indeed, much of the initial advances in the application of remote sensing for the SPAC came from such multidisciplinary field experiments. Examples in natural ecosystems include FIFE (Sellers et al., 1988), and BOREAS (Sellers et al., 1997), while in agro-ecosystems examples include MONSOON '90 (Kustas and Goodrich, 1994), JORNEX (Havstad et al., 2000), SGP97 (Kustas et al., 2001), SMACEX (Kustas et al., 2005), BEAREX08 (Evelt et al., 2012), and GRAPEX (Kustas et al., 2018).

A recent paper by Durand et al. (2021) provides recommendations on how to better make use of Earth observations to advance hydrologic science. These same recommendations are relevant to advancing SPAC research. They include integrating multi-sensor data over the full electromagnetic spectrum, leveraging commercial satellite observations, and employing not only processed-based models but incorporating data assimilation, making use of cloud computing resources, and incorporating machine learning techniques. The future in advancing SPAC research will be strongly tied to the integration of spatial and temporal information from Earth observations as this is the only means of bridging both spatial and temporal scales from plant level processes up to the global scale.

References

- Anderson MC, Norman JM, Meyers TP, and Diak GR (2000) An analytical model for estimating canopy transpiration and carbon assimilation fluxes based on canopy light-use efficiency. *Agricultural and Forest Meteorology* 101: 265–289.
- Anderson MC, Diak GR, Gao F, Knipper K, Hain CR, Eichelmann E, Hemes KS, Baldocchi DD, Kustas WP, and Yang Y (2019) Impact of insolation data source on remote sensing retrievals of evapotranspiration over the California Delta. *Remote Sensing* 11: 216.
- Bean AR, Coffin AW, Arthur DK, Baffaut C, Holfield Collins C, Goslee SC, Ponce-Campos GE, Sclater VL, Strickland TC, and Yasarer LM (2021) Regional frameworks for the USDA long-term agroecosystem research network. *Frontiers in Sustainable Food Systems* 4.
- Campbell GS and Norman JM (1998) *An Introduction to Environmental Biophysics*. New York: Springer-Verlag.
- Dorigo WA, Scipal K, Parinussa RM, Liu YY, Wagner W, De Jeu RAM, and Naeimi V (2010) Error characterisation of global active and passive microwave soil moisture datasets. *Hydrology and Earth System Sciences* 14: 2605–2616.
- Durand M, Barros A, Dozier J, Adler RF, Cooley S, Entekhabi D, Forman BA, Konings AG, Kustas W, Lundquist JD, Pavelsky TM, Rodell M, and Steele-Dunne S (2021) Achieving breakthroughs in global hydrologic science by unlocking the power of multisensor, multidisciplinary Earth observations. *AGU Advances* 2: e2021AV000455.
- Evelt SR, Kustas WP, Gowda PH, Anderson MC, Prueger JH, and Howell TA (2012) Overview of the Bushland Evapotranspiration and Agricultural Remote sensing EXperiment 2008 (BEAREX08): A field experiment evaluating methods quantifying ET at multiple scales. *Advances in Water Resources* 50: 4–19.
- Goodrich DC, Heilman P, Anderson M, Baffaut C, Bonta J, Bosch D, Bryant R, Cosh M, Endale D, Veith TL, Havens SC, Hedrick A, Kleinman PJ, Langendoen EJ, McCarty G, Moorman T, Marks D, Pierson F, Rigby JR, Schomberg H, Starks P, Steiner J, Strickland T, and Tsegaye T (2021) The USDA-ARS experimental watershed network: Evolution, lessons learned, societal benefits, and moving forward. *Water Resources Research* 57.
- Guan K, Wu J, Kimball JS, Anderson MC, Frolking S, Li B, Hain CR, and Lobell DB (2017) The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields. *Remote Sensing of Environment* 199: 333–349.
- Guzinski R, Nieto H, Sandholt I, and Karamitilios G (2020) Modelling high-resolution actual evapotranspiration through Sentinel-2 and Sentinel-3 data fusion. *Remote Sensing* 12: 1433. <https://doi.org/10.3390/rs12091433>.
- Hain CR, Crow WT, Anderson MC, and Yilmaz MT (2015) Diagnosing neglected moisture source/sink processes with a thermal infrared-based Two-Source Energy Balance model. *Journal of Hydrometeorology* 16: 1070–1086.
- Havstad KM, Kustas WP, Rango A, Ritchie JC, and Schugge TJ (2000) Jornada Experimental Range: A unique arid land location for experiments to validate satellite systems. *Remote Sensing of Environment* 74: 13–25.
- Jonard F, De Cannière S, Brüggemann N, Gentine P, Short Gianotti DJ, Lobet G, Miralles DG, Montzka C, Pagán BR, Rascher U, and Vereecken H (2020) Value of sun-induced chlorophyll fluorescence for quantifying hydrological states and fluxes: Current status and challenges. *Agricultural and Forest Meteorology* 291.
- Jones CA and Kiniry JR (1986) *CERES-Maize: A Simulation Model of Maize Growth and Development*. College Station: Texas A&M University Press.
- Konings AG, Williams AP, and Gentine P (2017) Sensitivity of grassland productivity to aridity controlled by stomatal and xylem regulation. *Nature Geoscience* 10: 284–288.
- Konings AG, Rao K, and Steele-Dunne SC (2019) Macro to micro: microwave remote sensing of plant water content for physiology and ecology. *New Phytologist* 223: 1166–1172.
- Kratz TK, Deegan LA, Harmon ME, and Lauenroth WK (2003) Ecological variability in space and time: Insights gained from the US LTER program. *BioScience* 53: 57–67.
- Kustas WP and Goodrich DC (1994) Preface, MONSOON '90 multidisciplinary experiment. *Water Resources Research* 30: 1211–1225.
- Kustas WP, Jackson TJ, French AN, and MacPherson JI (2001) Verification of patch- and regional-scale energy balance estimates derived from microwave and optical remote sensing during SGP97. *Journal of Hydrometeorology* 2: 254–273.
- Kustas WP, Hatfield J, and Prueger JH (2005) The soil moisture atmosphere coupling experiment (SMACEX): Background, hydrometeorological conditions and preliminary findings. *Journal of Hydrometeorology* 6: 791–804.

- Kustas WP, Anderson MC, Alfieri JG, Knipper K, Torres-Rua A, Parry CK, Nieto H, Agam N, White A, Gao F, McKee L, Prueger JH, Hipps LE, Los SO, Alsina M, Sanchez L, Sams B, Dokoozlian N, McKee M, Jones S, McElrone A, Heitman JL, Howard AM, Post K, Melton FS, and Hain C (2018) The grape remote sensing atmospheric profile and evapotranspiration experiment (GRAPEX). *Bulletin of the American Meteorological Society* 99: 1791–1812.
- Lovelock J (1987) *Gaia: A New Look at Life on Earth*. New York: Oxford University Press.
- McCabe MF, Rodell M, Alsdorf DE, Miralles DG, Uijlenhoet R, Wagner W, Lucieer A, Houborg R, Verhoest NEC, Franz TE, Shi J, Gao H, and Wood EF (2017) The future of Earth observation in hydrology. *Hydrology and Earth System Sciences* 21: 3879–3914.
- Melton FS, Huntington J, Grimm R, Herring, J, Hall M, Rollison, D, Erickson T, et al. (2021) OpenET: Filling a Critical Data Gap in Water Management for the Western United States. *Journal of the American Water Resources Association* 1–24. <https://doi.org/10.1111/1752-1688.12956>.
- Morgan CLS, Norman JM, and Lowery B (2003) Estimating plant-available water across a field with an inverse yield model. *Soil Science Society of America Journal* 67: 620–629.
- Norman JM and Arkebauer TJ (1991) Predicting canopy light-use efficiency from leaf characteristics. In: *Modeling Plant and Soil Systems, Agronomy Monograph No. 31*, pp. 125–143. Madison: ASA-CSSA-SSSA.
- Norman JM and Campbell GS (1983) Application of a plant-environment model to problems in irrigation. In: Hillel DL (ed.) *Advances in Irrigation*. vol. II, pp. 155–188. New York: Academic Press.
- Norman JM, Kustas WP, and Humes KS (1995) A two-source approach for estimating soil and vegetation energy fluxes from observations of directional radiometric surface temperature. *Agricultural and Forest Meteorology* 77: 263–293.
- Pagán BR, Maes WH, Gentile P, Martens B, and Miralles DG (2019) Exploring the potential of satellite solar-induced fluorescence to constrain global transpiration estimates. *Remote Sensing* 11.
- Sellers PJ, Hall FG, Asrar G, Strebel DE, and Murphy RE (1988) The first ISLSCP field experiment (FIFE). *Bulletin of the American Meteorological Society* 69: 22–27.
- Sellers PJ, Hall FG, Kelly RD, Black A, Baldocchi D, Berry J, Ryan M, Ranson KJ, Crill PM, Lettenmaier DP, Margolis H, Cihlar J, Newcomer J, Fitzjarrald D, Jarvis PG, Gower ST, Halliwell D, Williams D, Goodison B, Wickland DE, and Guertin FE (1997) BOREAS in 1997: Experiment overview, scientific results, and future directions. *Journal of Geophysical Research* 102(D24): 28731–28769.
- Xue J, Anderson MC, Gao F, Hain C, Knipper KR, Yang Y, Kustas WP, Bambach NE, McElrone AJ, Castro SJ, Alfieri JG, Prueger JH, McKee LG, Hipps LE, and Alsina M (2022) Improving the spatiotemporal resolution of remotely sensed ET information for water management through Landsat, Sentinel-2, ECOSTRESS and VIIRS data fusion. *Irrigation Science*. <https://doi.org/10.1007/s00271-022-00799-7>.

The Penman-Monteith equation

Richard G Allen^a, Ayse Kilic^b, Guido D'Urso^c, and Simona Consoli^d, ^aDepartment of Soil and Water Systems, University of Idaho, Kimberly, ID, United States; ^bUniversity of Nebraska-Lincoln, Lincoln, NE, United States; ^cUniversita' degli Studi di Napoli Federico II, Portici, Italy; ^dUniversità degli Studi di Catania, Catania, Italy

© 2023 Elsevier Ltd. All rights reserved.

This is an update of R. Allen, Penman-Monteith equation, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, pp. 180–188, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00399-4>.

Introduction	634
Derivation of the Penman-Monteith equation	635
Other parameters in the PM equation	638
Aerodynamic and surface parameters	638
Parameterization of surface resistance	639
Multi-layer ET models	640
Net radiation	640
Soil heat flux	641
Application of the PM equation	642
Limitations to applying the PM equation	642
Applying PM with gridded weather forecast data sets	643
Conclusion	643
References	643
Further reading	644

Abstract

The Penman-Monteith is an elegant, physical expression of the process that water vapor follows when transported from inside the leaf of a plant into the atmospheric boundary layer above. The original Penman combination equation (1948) combined the energy input of radiation from the sun and atmosphere with convective transport of temperature and humidity from the vegetation surface. The Penman-Monteith formulation (1965) included resistance terms for bulk stomates and aerodynamic transport. The equation has gained popularity through its use in estimating reference ET from a grassed surface that is then transformed into ET for specific crops through the crop coefficient.

Key points

- The Penman-Monteith (PM) evapotranspiration (ET) equation estimates the rate of total evaporation and transpiration from the earth's surface
- The PM equation uses commonly measured weather data (solar radiation, air temperature, vapor content, and wind speed)
- The PM equation is derived from a surface energy balance combined with convective transport terms for heat and vapor
- The PM equation is most commonly applied as a reference crop ET equation, for example, for grass and alfalfa surfaces, where assumptions of the equation that surface temperature is near air temperature apply

Introduction

The Penman-Monteith (PM) evapotranspiration (ET) equation (Monteith, 1965) estimates the rate of total evaporation (E) and transpiration (T) from the earth's surface using commonly measured weather data (solar radiation, air temperature, vapor content, and wind speed). The PM equation follows a single-layer or "big leaf" approach, where single surface resistance and single aerodynamic resistance terms represent the transport properties of the vegetated surface. The PM procedure provides scientists, engineers, and managers with the ability to evaluate effects of modifications to cultural practices and crop variety improvements on ET, such as increased plant height, increased or reduced leaf size or density, reduced row spacing, reduced stomatal conductance, etc., by varying these parameters in the calculation algorithms. The PM method is useful in providing direction in determining changes in crop coefficients under new cropping conditions. Determination of crop coefficients is an expensive and time-consuming process, as it requires additional lysimeter or micrometeorological studies.

The PM equation has the form:

$$E \text{ or } ET = \frac{\Delta(R_n - G) + \rho_a c_p (e_s - e_a)/r_a}{\left(\Delta + \gamma \left(1 + \frac{r_s}{r_a}\right)\right) \lambda \rho_w} \quad (1)$$

where Δ is the slope of the saturation vapor pressure vs. temperature curve, R_n is the net radiation flux at the surface, G is the sensible heat exchange from the surface to the soil (positive if the soil is warming), ρ_a is air density, c_p is specific heat of dry air, e_s is the saturation vapor pressure of the air at some height above the surface, e_a is the actual vapor pressure of the air, r_a is aerodynamic resistance to turbulent heat and/or vapor transfer from the surface to some z height above the surface, γ is the psychrometric constant (defined later), r_s is a bulk surface resistance that describes the resistance to flow of water vapor from inside the leaf, vegetation canopy or soil to outside the surface, λ is the latent heat of vaporization, defined as the energy required to convert a mass of liquid water into vapor (having typical units of J kg^{-1}), and ρ_w is density of liquid water. All parameter units in (1) must cancel so that the remaining units for E or ET (depending on application to E from soil or ET from soil and vegetation) are presented as L T^{-1} , for example, mm h^{-1} or mm d^{-1} .

Surface resistance r_s may represent the resistance to vapor flow through plant leaf stomates, or resistance to vapor flow within soil to the soil surface or both in combination. For a wet (saturated) surface, r_s is by definition zero. When r_s is zero, Eq. (1) reverts to the original Penman equation.

Derivation of the Penman-Monteith equation

It is useful to review the derivation of the PM equation to understand how and why it functions. There are two fundamental approaches to estimate E (or ET). These are the surface energy balance equation and the aerodynamic equation. Each equation uses a different approach and each has difficulties for solution when using typical weather data comprised of solar radiation, air temperature, vapor pressure, and wind speed. Penman (1948) combined these two approaches into the so-called combination or Penman equation that could be solved using standard weather data only. Monteith (1965) improved on the Penman (1948) equation by including a surface resistance term and a more rigorous term for aerodynamic transfer.

The surface energy balance equation relates to the various ways in which net radiation energy from the sun and atmosphere is transformed at the surface into outputs of energy in the form of non-radiative parameters. The vertical energy balance at the surface is the sum of sensible heat flux to or from the air and to or from the soil, along with latent heat flux, net radiation and other miscellaneous fluxes. The major components of the vertical energy balance are expressed as:

$$R_n = G + \lambda E + H \quad (2)$$

where λE is the latent heat flux (positive during evaporation) and H is the sensible heat exchange from the surface to air (positive if the air is warming). In Eq. (2), E (or ET) is expressed in terms of the energy required to convert the amount of liquid evaporated into vapor by multiplying by λ . The miscellaneous terms involved in the surface energy balance, such as heat storage within foliage, photosynthesis and respiration are generally insignificant relative to magnitudes of R_n , λE , H and G , and are commonly neglected. Under some conditions, change in heat storage in a canopy may need attention, especially in forest canopies. If R_n and sensible heat flux densities H and G can be measured or estimated reliably, then the latent heat flux density, λE , can be computed from Eq. (2).

$$\lambda E = R_n - G - H \quad (3)$$

The challenge to the solution of Eq. (3) using standard weather data is in solving for H . The traditional aerodynamic expression H is:

$$H = \frac{\rho c_p (T_o - T_a)}{r_{ah}} \quad (4)$$

where T_o is surface temperature, T_a is air temperature and r_{ah} is the aerodynamic resistance to heat transfer between some mean height within the surface to some height a few m above the surface. Generally, r_{ah} is assumed equal to r_a of Eq. (1).

Eq. (4) is analogous to Ohm's law for electrical current flow, where T_o and T_a are equivalent to voltages or potentials, r_{ah} is equivalent to an electrical resistor, and H is analogous to current. Thus, the equation behaves as a linear electrical circuit, so that if the difference between T_o and T_a is doubled, H will double. If resistance r_{ah} is halved (through a doubling of wind speed), then H will double.

The challenge of Eq. (4), and therefore of Eq. (3), is in obtaining surface temperature T_o . T_o can be measured using infrared thermometers, if surface emissivity is known, or it can be measured using a fine-thermocouple or other temperature sensor placed just beneath the surface skin. However, these measurements are not common due to the high operational and maintenance requirements of the instruments and difficulty in transfer of measured values to other locations and surface conditions, even those nearby.

The traditional equation for r_{ah} (and r_a) is:

$$r_{ah} = \frac{\left(\ln \left(\frac{z_u - d}{z_{om}} \right) - \psi_m \right) \left(\ln \left(\frac{z_T - d}{z_{oh}} \right) - \psi_h \right)}{K^2 u_z} \quad (5)$$

where z_u is the height above the ground surface for the wind speed measurement, d is the zero plane displacement of the logarithmic wind profile, z_{om} is a roughness length governing the transfer of momentum from the surface, ψ_m is a correction factor for momentum transfer to account for buoyant instability or stability of the boundary layer, z_T is the height of the air temperature measurement above the ground surface, z_{oh} is an assumed roughness length governing the transfer of sensible heat from the surface, ψ_h is a correction factor for sensible heat transfer to account for instability or stability of the boundary layer, k is the von-karman constant (0.41), and u_z is the wind speed measured at the z_u height. Eq. (5) is sometimes criticized for its use of z_{oh} , commonly expressed as some constant fraction of z_{om} , whereas in reality the ratio of z_{oh}/z_{om} can vary substantially, especially for dry, sparse vegetation (Qualls and Brutsaert, 1996). Other forms of Eq. (5) can be written that rely on differences in T_a and wind speed at two different heights above the surface. Although these forms eliminate the need for z_{oh} , measurement of T_a and wind speed at two heights is not common and it is challenging to produce an unbiased pair of measurements. A variety of studies have investigated effects of foliage density and geometry on apparent bulk roughness of vegetation (see Brutsaert, 1975).

The aerodynamic-based expression for λE is:

$$\lambda E = \frac{\rho c_p (e_o - e_a)}{\gamma (r_{av} + r_s)} \quad (6)$$

where e_o is vapor pressure “inside” the surface, for example inside a leaf stomatal cavity or beneath the soil surface, e_a is vapor pressure of the air at some z height above the ground surface, γ is the psychrometric constant (defined later), r_{av} is aerodynamic resistance to turbulent vapor transfer from the surface to some z height above the surface, and r_s represents any surface resistance to flow of vapor (defined following Eq. 1). Aerodynamic resistance r_{av} is computed exactly as for r_{ah} , i.e., using Eq. (5), except that z_T becomes z_v (height of vapor pressure measurement above the ground), z_{oh} becomes z_{ov} , which is an assumed roughness length governing the transfer of vapor from the surface, and ψ_h becomes ψ_v , a correction factor to account for effect of instability or stability of the boundary layer on vapor flow. Generally, $z_T = z_v$ (air temperature and humidity sensors are usually collocated at the same height), and z_{ov} is assumed equal to z_{oh} and ψ_v is assumed equal to ψ_h . The stability parameters ψ_v , ψ_h and ψ_m can generally be set to zero when calculating ET from well-watered (moist) surfaces because of nearly neutral stability (buoyancy) for those conditions caused by H being close to zero due to conversion of most R_n into λE .

Again, the aerodynamic vapor transfer equation is analogous to Ohm’s law, where e_o and e_a are equivalent to voltages or potentials, r_{av} and r_s are equivalent to electrical resistors *in series*, and λE (vapor flux) is analogous to electrical current.

Vapor pressure of air can be determined from paired measurements of relative humidity and air temperature, which are commonly available or from measurement of dewpoint temperature also commonly available. Vapor pressure inside the soil or plant surface is rarely measured, however, similar to T_o , so that as with Eq. (3), Eq. (6) is challenging to apply in routine practice. Hence the need for the Penman or PM combination equation.

Fig. 1 is a schematic showing the relative locations of resistances r_a (or r_{ah} and r_{av}) and r_s where r_s represents the internal resistance of the soil (if no vegetation) or of the soil, leaves and canopy *in parallel*.

To create the PM equation, one needs to define what is known as the Bowen ratio which is simply the ratio between H and λE . Expanded into the equations for H and λE and written for transfer from *inside* the surface to the air above:

$$\beta = \frac{H}{\lambda E} = \frac{\rho c_p (T_o - T_a) \gamma (r_{av} + r_s)}{\rho c_p (e_o - e_a) (r_{ah})} \quad (7)$$

Cancelling similar terms:

$$\beta = \frac{H}{\lambda E} = \gamma \frac{(T_o - T_a) (r_{av} + r_s)}{(e_o - e_a) (r_{ah})} \quad (8)$$

The energy balance equation can be rearranged and written as:

$$\lambda E = \frac{R_n - G}{1 + \beta} = \frac{R_n - G}{1 + \gamma \frac{(T_o - T_a) (r_{av} + r_s)}{(e_o - e_a) (r_{ah})}} \quad (9)$$

The next step in the derivation of the PM equation is to define the term Δ , which is defined as the slope of the saturation vapor pressure vs. temperature relation. Saturation vapor pressure $e_o(T)$ is a singular function of T and describes the ability of air to hold more vapor as temperature increases. The slope Δ of the $e_o(T)$ vs. T curve can be expressed as a linear approximation for a segment of the saturation curve between two temperatures T_o and T_a :

$$\Delta = \frac{e_o - e_s}{T_o - T_a} \quad (10)$$

where e_o , e_s , T_o and T_a were defined previously. Vapor pressure inside the surface, e_o , is assumed to be saturated and is therefore equivalent to $e_o(T_o)$. e_s is the saturation vapor pressure associated with T_a , so that $e_s = e_o(T_a)$. In practice, Δ is often calculated as the

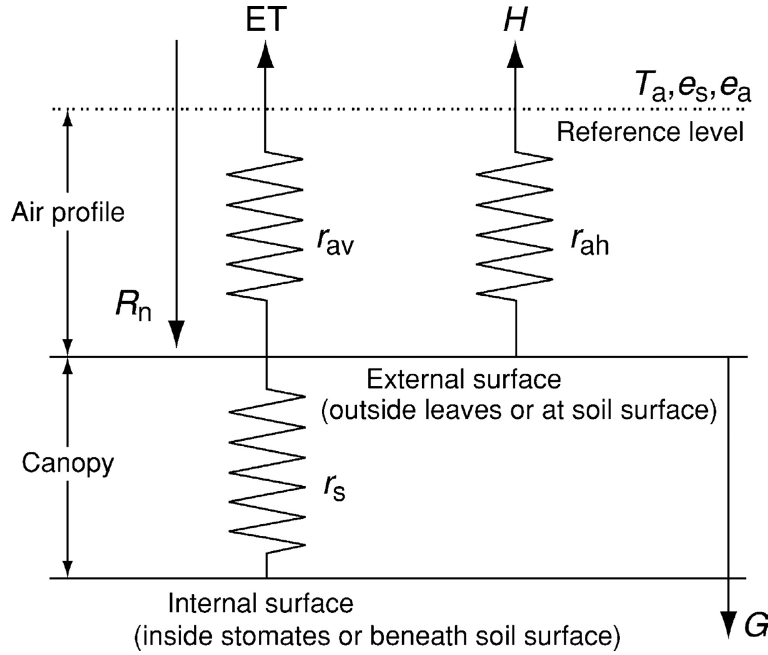


Fig. 1 Schematic showing linkage between resistance terms in the PM equation relative to the surface and elevation of temperature and humidity measurements. ET, evapotranspiration; e_s , saturation vapor pressure at mean air temperature; e_a , actual vapor pressure of the air; G , heat exchange from surface to soil; H , heat exchange from surface to air; r_a , aerodynamic resistance; r_s , bulk surface resistance; R_n , net radiation flux at the surface; T_a , air temperature; r_{av} , aerodynamic resistance to vapor transfer; r_{ah} , aerodynamic resistance to heat transfer.

slope of $e_o(T)$ vs. T at T_a only, since T_o is usually not known and can be substantially warmer or cooler than T_a . More is discussed on this later.

Using the relationship from Eq. (10) to define $(e_o - e_s)/\Delta = T_o - T_a$ and parsing $(e_o - e_s)$ into an equivalent expression of $(e_o - e_s) = (e_o - e_a) - (e_s - e_a)$ so that $T_o - T_a = [(e_o - e_a) - (e_s - e_a)]/\Delta$, this expression is substituted into Eq. (9) so that:

$$\lambda E = \frac{R_n - G}{1 + \beta} = \frac{R_n - G}{1 + \frac{\gamma}{\Delta} \frac{[(e_o - e_a) - (e_s - e_a)](r_{av} + r_s)}{(e_o - e_a)(r_{ah})}} = \frac{R_n - G}{1 + \frac{\gamma}{\Delta} \left[1 - \frac{(e_s - e_a)}{(e_o - e_a)} \right] \frac{(r_{av} + r_s)}{r_{ah}}} \quad (11)$$

We define one more temporary term, λE_a , following Penman (1948):

$$\lambda E_a = \frac{\rho c_p (e_s - e_a)}{\gamma r_{ah}} \quad (12)$$

where λE_a is similar to Eq. (6), the aerodynamic expression for λE , except that r_s is omitted and $(e_o - e_a)$ is replaced with $(e_s - e_a)$. E_a is a fictitious term that allows the elimination of surface parameters T_o and e_s in the final PM equation. The term $(e_s - e_a)$ is known as the vapor pressure deficit (VPD) of the air, because it represents the difference between the saturation vapor pressure at air temperature ($e_s = e_o(T_a)$) and actual vapor pressure of the air (e_a). Both e_s and e_a can be determined from commonly available humidity and air temperature data measured at some height above the surface, generally 1.5–2 m. Because

$$\frac{e_s - e_a}{e_o - e_a} = \frac{\lambda E_a}{\lambda E} \frac{r_{ah}}{r_{av} + r_s} \quad (13)$$

Then Eq. (11) becomes

$$\lambda E = \frac{R_n - G}{1 + \frac{\gamma}{\Delta} \left[1 - \frac{\lambda E_a}{\lambda E} \frac{r_{ah}}{r_{av} + r_s} \right] \frac{r_{av} + r_s}{r_{ah}}} \quad (14)$$

Rearranging Eq. (14) and setting $r_{av} = r_{ah}$ results in the PM equation:

$$\lambda E = \frac{\Delta (R_n - G) + \gamma \lambda E_a}{\Delta + \gamma \left(1 + \frac{r_s}{r_{ah}} \right)} = \frac{\Delta (R_n - G) + \rho c_p (e_s - e_a)/r_{ah}}{\Delta + \gamma \left(1 + \frac{r_s}{r_{ah}} \right)} \quad (15)$$

where, in Eq. (15), E is expressed in terms of energy, rather than in terms of depth of liquid water as in Eq. (1). The PM equation has the extremely valuable advantage of requiring only commonly available weather data, namely solar radiation (used to calculate R_n), air temperature (used to calculate ρ , e_s and Δ), air humidity (used to calculate e_a and assist with R_n), and wind speed (for r_{ah}).

Parameter G is generally estimated as a function of R_n or by direct measurement and r_s is generally estimated as a function of amount of vegetation or amount of surface wetness, of soil. The PM (and Penman equation where r_s is assumed zero and $1/r_{ah}$ is replaced by a simpler empirical “wind function”) are widely used around the world as standard means to predict E and ET . As discussed later, estimation of Δ from T_a only is mostly valid under well-watered (reference crop) conditions where $T_o \sim T_a$. This was the approach of Penman who developed his combination equation for grass and water surfaces.

Other parameters in the PM equation

The psychrometric constant γ is computed as:

$$\gamma = \frac{c_p P}{\lambda \varepsilon} \quad (16)$$

where, $c_p = 1.013 \text{ kJ kg}^{-1} \text{ K}^{-1}$ for moist air, P is atmospheric air pressure, λ is the latent heat of vaporization in kJ kg^{-1} , and $\varepsilon = 0.622$ is the ratio of molecular weights of water vapor to air. For P in kPa, γ will have units of $\text{kPa } ^\circ\text{C}^{-1}$. Mean atmospheric air pressure, P , can be routinely computed as a function of surface elevation using the ideal gas law (Allen et al., 1998) as:

$$P = P_{os} \left(\frac{T_{os} - \alpha_a (\text{elev} - \text{elev}_{os})}{T_{os}} \right)^{g/(\alpha_a R)} \quad P = P_{os} \left(\frac{T_{os} - \alpha (\text{elev} - \text{elev}_{os})}{T_{os}} \right)^{\frac{g}{(\alpha_a R)}} \quad (17)$$

where P_{os} and T_{os} are standard atmospheric pressure in kPa and absolute temperature in K at elevation elev_{os} , and elev is elevation of the ground surface above mean sea level, α_a is the assumed constant adiabatic lapse rate, normally taken as 0.0065 K m^{-1} for saturated air or sometimes as 0.01 K m^{-1} for non-saturated air, g is gravitational acceleration, 9.807 m s^{-2} and R is the specific gas constant for dry air, $287.0 \text{ J kg}^{-1} \text{ K}^{-1}$. Values for P_{os} , T_{os} and elev_{os} are commonly taken for the standard atmosphere at sea level, which are 101.3 kPa, 288 K, and 0 m, respectively. Eq. (17) is relatively insensitive to the value of α_a for elevations up to 3000 m.

Air density, ρ , can be computed as

$$\rho = \frac{1000 P}{T_v R} \quad (18)$$

where P is in kPa, R is $287.0 \text{ J kg}^{-1} \text{ K}^{-1}$ and T_v is virtual temperature, in K. T_v can be computed as:

$$T_v = \frac{T}{1 - 0.378 \frac{e_a}{P}} \quad (19)$$

where T is air temperature in K and e_a is mean vapor pressure of the air, kPa.

Latent heat of vaporization, λ , in MJ kg^{-1} can be computed as:

$$\lambda = 2.501 - 0.002361 T \quad \lambda = 2.501 - 0.002361 T \quad (20)$$

where T is mean air temperature in $^\circ\text{C}$. If available, mean surface temperature or wet bulb temperature can be used to compute the value of λ which better represents conditions at the evaporating surface.

The PM equation can be applied for short periods of 1 h or less, and in fact is intended to represent nearly instantaneous fluxes of ET . However, the equation has traditionally been applied for daily and longer (up to 1 month) calculation timesteps due to paucity of weather data. For daily and longer timesteps, e_s in the PM should be computed as:

$$e_s = \frac{e^o(T_{max}) + e^o(T_{min})}{2} \quad (21)$$

where T_{max} and T_{min} are maximum and minimum daily air temperature for the period and e^o is the saturation vapor pressure function. Actual vapor pressure of the air, e_a , can be computed from relative humidity data, from wet-bulb and dry-bulb psychrometer data, or from dewpoint temperature measurements:

$$e_a = e^o(T_d) \quad (22)$$

where T_d is mean daily or early morning dewpoint temperature.

Aerodynamic and surface parameters

When used to estimate ET from dense vegetation, aerodynamic resistance in PM can be calculated using parameters for d , z_{om} , and z_{oh} determined as a function of vegetation height:

$$d = 0.67 H \quad (23)$$

$$z_{om} = 0.12 H \quad (24)$$

$$z_{oh} \approx 0.1 z_{om} \quad (25)$$

where z_{om} , z_{oh} , d , and mean plant height, H , are in m. For bare soils, common values for z_{om} range from 0.00001 m for mud flats to 0.005 m for smooth soil to 0.01 m or higher for very roughly ploughed soil.

Parameterization of surface resistance

The bulk surface resistance r_s for a crop is an integration of effects of transpiration from various locations within the vegetation and contributions from soil evaporation, along with transfers of radiation and sensible heat within the canopy. Simple approaches determine soil evaporation indirectly by estimating the distribution of net radiation below and between a plant canopy having partial ground cover. For crops that completely shade the ground, total surface resistance can be computed following Grant (1975) where surface resistances from soil and vegetation are added in parallel according to leaf area and radiation extinction coefficients. For an immature row crop or natural vegetation having exposed bare soil between plant rows and incomplete wetting of the soil surface during irrigation, r_s can be expressed as an extension of Grant's approach through combination of various surface resistances acting in parallel. The combination is made according to the fractions of surface C represented by each parallel resistance component (Fig. 2):

$$\frac{1}{r_s} = \frac{C (1 - K_r^{LAI/C})}{r_{sc}} + K_r^{LAI/C} \left(\frac{C_{sW}}{r_{ssw}} + \frac{C_{sD}}{r_{ssd}} \right) + \frac{C_{eW}}{r_{sseW}} + \frac{C_{eD}}{r_{sseD}} \quad (26)$$

where C is the effective fraction of ground shaded by vegetation and is computed as a function of sun angle, row height and width, leaf density, and row orientation, and C_{sW} , C_{sD} , C_{eW} , and C_{eD} are the fractions of the soil surface that are shaded and wet, shaded and dry, sunlit and wet, and sunlit and dry ($C_{sW} + C_{sD} + C_{eW} + C_{eD} = 1$ and $C_{sW} + C_{sD} = C$). Resistance r_{sc} is the surface resistance of the canopy, r_{ssw} is the surface resistance of wet, shaded soil, r_{ssd} is the surface resistance of dry, shaded soil, r_{sseW} is the surface resistance of wet, sunlit soil, and r_{sseD} is the surface resistance of dry, sunlit soil. Parameter K_r is a radiation extinction coefficient for the vegetation canopy having a value of approximately 0.7 (Grant, 1975) and where LAI is leaf area index defined as the aggregated surface area (one-side) of leaves per unit of ground surface area.

Thompson et al. (1981) computed C as $LAI/4$. Brisson (1996) applied separate PM equations for evaporation and transpiration and then recoupled them, similar to the procedure for multi-layer ET models. For sun angles other than directly overhead, C can be estimated for a row crop following Allen et al. (1998) as:

$$C = \frac{W + H |\sin(\eta + \beta)| \tan \theta_z}{S} \quad (27)$$

where H is mean plant height, θ_z is the solar zenith angle ($0 = \text{nadir (overhead)}$), η is the sun azimuth angle (in the northern hemisphere, $\eta = 0$ when sun is due south in the northern hemisphere, η is negative when sun is east of south and η is positive when

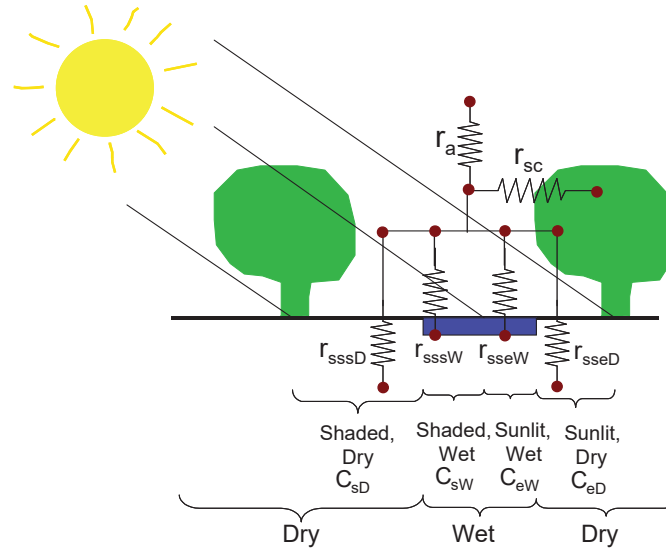


Fig. 2 Schematic showing breakdown and definition of surface resistances for soil when vegetation is at partial cover. r_a , aerodynamic resistance; r_{sc} , surface resistance for the vegetation cover; r_{ssw} , surface resistance of wet, shaded soil; r_{ssd} , surface resistance of dry, shaded soil; r_{sseW} , surface resistance of wet, sunlit soil; r_{sseD} , surface resistance of dry, sunlit soil; C_{sD} , fraction of the soil surface that is shaded and dry; C_{sW} , fraction of the soil surface that is shaded and wet; C_{eW} , fraction of the soil surface that is sunlit and wet; C_{eD} , fraction of the soil surface that is sunlit and dry.

sun is west of south), and β is the row orientation relative to north ($\beta = 0$ for north-south orientation, $\beta = 90^\circ$ for east-west row orientation), W is row width and S is row spacing. Eq. (27) presumes a nearly rectangular row shape of height H and width W and has an upper limit of $C = 1$. In some situations, W may not be measured or available, but an estimate of LAI may be available. In this case, C may be approximated during incomplete row closure as:

$$C = \frac{LAI}{LAI_C} + \frac{H}{S} |\sin(\eta + \beta)| \tan \theta_z \quad (28)$$

where LAI_C is the LAI at complete row closure and shading of the ground. The value for C and, subsequently, for r_s changes with time of day and interacts with R_n , G , wind speed and temperature. For 24-h computation time steps, an average daily value for C is determined for computing soil moisture balances for shaded and unshaded soil. An average daily value for C for east-west rows can be determined by weighting Eq. (27) for sunlight intensity and integrating between the limits of sunrise and sunset. For north-south row orientation, all soil is sunlit and shaded at various times of the day.

Surface resistance, r_{sc} , for densely growing vegetation is often computed in the “single-layer” PM as a function of leaf area by assuming all leaves function as resistors in parallel:

$$r_{sc} = \frac{r_l}{LAI_{eff}} \quad (29)$$

where r_l is the bulk stomatal (or surface) resistance of the vegetation per unit LAI ($s\ m^{-1}$) and LAI_{eff} is the effective LAI contributing to ET. Parameter r_l is the inverse of g_l , the stomatal conductance per unit leaf area. The value of r_l has been taken as $100\ s\ m^{-1}$ for many well-watered agricultural crops (Monteith, 1965; Allen et al., 1989) when calculations are made on a 24-h basis. When ET calculations are made on an hourly or shorter basis, r_l for agricultural crops has a value of about $70\text{--}80\ s\ m^{-1}$ (Allen et al., 2006). The value for r_l increases with environmental stresses including soil moisture deficit.

Surface resistance, r_{sc} , plays an important role in controlling the flux of transpiration from plants into the atmosphere and must be accurately predicted for routine application of the PM and multilayer ET models. r_{sc} represents the ET-averaged daily resistance for 24-h calculation time steps or the mean canopy resistance for hourly time steps for a fully developed canopy. Algorithms are used to modify r_{sc} as leaf area (LAI) and environmental parameters change. General, simple environmental stomatal regulation algorithms have been proposed by several authors, including Lecina et al. (2003) where bulk surface resistance is expressed in terms of multiplicative products of linear functions of solar radiation, air temperature, vapor pressure deficit and soil water deficit.

Soil resistance is generally assumed to increase with drying of soil. Exponential functions describe the process relatively well. A moisture-based soil resistance function that describes experimental data from a Portneuf silt loam soil near Kimberly, Idaho is:

$$r_{ss} = 2000e^{-5x_s} \quad (30)$$

where x_s is the remaining proportion of water in the evaporating layer of the soil surface and e is the natural number. x_s is computed as $x_s = W_{ei}/W_{emax}$, where W_{ei} is the total evaporable moisture near the soil surface on day i and W_{emax} represents the upper limit of evaporable water. A daily or more frequent water balance of the C_{sw} , C_{SD} , C_{EW} , and C_{ED} fractions of the soil surface layer need to be conducted to predict x_s , and thus r_{ss} , for each surface fraction. The E from each fraction is predicted in some proportion to ratio of the inverse of r_{ss} for that fraction to the total bulk r_s for the soil-canopy complex as in Eq. (26). The soil surface layer is typically presumed to be $0.1\text{--}0.15\ m$ thick (Allen et al., 1998). Eq. (30) predicts values for r_{ss} ranging from about $15\ s\ m^{-1}$ directly after wetting to about $2000\ s\ m^{-1}$ for a relatively dry surface soil layer. Or and Lehmann (2019) developed a soil-dependent evaporative characteristic length defines an active surface evaporative capacitor depth below which soil water is sheltered from capillary pull to the evaporating surface. They tested and applied their model globally.

Multi-layer ET models

The PM equation is basically a single layer model where the resistances for vegetation and soil are assumed to reside in parallel, i.e., there are no intermediate aerodynamic or convective resistances separating them or within them. More complicated multi-layer ET models have been employed that contain sophisticated parameter algorithms, intermediate resistances, and relatively complex calculation mechanisms than PM. Multi-layer models decompose the computation of r_s , r_{sc} , and r_a for discrete layers within the vegetation canopy (e.g., Wallace (1996), and Colaizzi et al. (2012)). Because of their increased complexity, multi-layer models can provide more capability and flexibility in predicting ET under variable cropping and weather conditions without the need for calibration, depending on the physical basis of internal mechanisms. Estimated ET may be more accurate than the PM, especially for short time steps (hourly). On the other hand, these models sometimes require substantial calibration of empirical coefficients internal to the model and may use questionable assumptions regarding values for canopy and soil temperature used in estimating spatial distribution of R_n as discussed later.

Net radiation

Net radiation, R_n , can be measured using a net radiometer that measures downward and upward fluxes of short-wave and long-wave radiation. The surface beneath the net radiometer must be very similar in wetness and vegetation type and amount to that being predicted in the PM equation. In many situations, R_n data is not available for a specific surface or condition, and therefore must be

estimated from solar radiation data, air temperature, and possibly vapor pressure. R_n for row crops is often calculated by separately considering the reflectances (albedo) for soil and vegetation. When calculated, the general formula for R_n is:

$$R_n = (1 - \alpha)R_s + R_{nl}R_n = (1 - \alpha)R_s + R_{nl} \quad (31)$$

where α is the albedo of the surface, representing the crop-soil mixture, R_s is incoming short-wave radiation (measured by a pyranometer), and R_{nl} is net (incoming-outgoing) long-wave radiation. R_{nl} is commonly estimated from air temperature and humidity data from a local weather station (Allen et al., 1998), however R_{nl} should be based partly on T_o of the surface that emits the long-wave thermal radiation:

$$R_{nl} = R_{L\downarrow} - R_{L\uparrow} - (1 - \epsilon_o)R_{L\downarrow} = \epsilon_o\epsilon_a\sigma T_a^4 - \epsilon_o\sigma T_o^4 \quad (32)$$

where $R_{L\downarrow}$ is incoming long-wave radiation from the atmosphere, $R_{L\uparrow}$ is outgoing (emitted) long-wave radiation from the surface having temperature T_o (in Kelvin), ϵ_o is broad-spectrum thermal emissivity of the surface, generally between 0.95 and 0.99 and ϵ_a is broad-spectrum thermal emissivity of the near surface atmosphere. The $(1 - \epsilon_o) R_{L\downarrow}$ term represents the fraction of incoming long-wave radiation reflected from the surface. T_a used in the Stephen-Boltzman expression in Eq. (32) represents the temperature of the lower near surface atmosphere that is responsible for emitting the majority of thermal radiation to the surface. One can note that differences in surface temperature, T_o , between wet, evaporating and dry surfaces can be as much as 30 K. Those differences in T_o can cause differences in emitted long-wave radiation from the surface and consequently R_n to be as large as 100 W m^{-2} between the two conditions.

A general formula for α of a crop-soil mixture used by Thompson et al. (1981) is:

$$\alpha = C\alpha_c + (1 - C)\alpha_s \quad (33)$$

where α_c is plant albedo, α_s is soil albedo and C = relative fraction of vegetation cover (defined previously) that considers effects of sun angle. Plant albedo can be computed by choosing typical values at the key growth stages. For example, Vanderkimpfen (1991) found values of α_c to be 0.25, 0.23, 0.23 and 0.30 for beans at Kimberly at emergence, effective cover, ripening and harvest. Values were 0.25, 0.15, 0.15 and 0.35 for wheat for the same stages due to the more vertical leaf structure for wheat and penetration of radiation into the canopy. For overcast conditions where most shortwave radiation is diffuse, the value for α_c increases, often to about 0.35.

Soil albedo in (32) can be computed as a composite of wet soil and dry soil albedo:

$$\alpha_s = \frac{\alpha_{sM} C_{eW} + \alpha_{sD} C_{eD}}{C_{eW} + C_{eD}} \quad (34)$$

where α_{sM} is albedo of moist soil and α_{sD} is albedo of dry soil. The albedo of a moist, drying soil can be assumed to vary linearly with water content between the value for a fully wet soil and a fully dry soil over a given drying period:

$$\alpha_{sM} = \alpha_{sW} + (\alpha_{sD} - \alpha_{sW})(1 - x_{skin}) \quad (35)$$

where α_{sW} is albedo of wet soil and x_{skin} is the fraction of remaining water in the soil skin (top $\sim 1 \text{ cm}$). Example values for α_{sW} and α_{sD} for the Portneuf silt loam near Kimberly, Idaho (Vanderkimpfen, 1991) are $\alpha_{sW} = 0.10$ – 0.20 and $\alpha_{sD} = 0.35$.

Soil heat flux

The magnitude of soil heat storage or release can be significant over a few hours, but is usually small from day to day because heat stored early in the day as soil warms is lost late in the day and at night as the soil cools. For application of PM to short-periods, e.g., using hourly data, measurements or estimates for G are required.

One method commonly used in research is the use of heat flux plates installed at shallow depths below the soil surface. (e.g., at 0.01 m). Installation and measurement with heat flux plates is described by Campbell Scientific Inc (1999). Measurements by soil heat flux plates, G_{zs} at some depth z_s are adjusted to represent G at the soil surface by incorporating the change in heat storage between the soil surface and heat flux plate:

$$G = G_{zs} + C_s \int_0^{z_s} \frac{\partial T_s}{\partial t} dz \quad (36)$$

where $\partial T_s / \partial t$ is the change in temperature with time within the z_s thick layer above the heat flux plate. Temperature is usually measured with thermocouples or buriable thermistors.

Soil heat flux density for dense, short vegetation such as for grass, wheat or alfalfa can be approximated during daylight periods as (Choudhury et al., 1987):

$$G = 0.4 e^{-0.5 LAI} R_n \quad (37)$$

where e is the natural number and G has the same units as R_n . Eq. (37) predicts $G = 0.1 R_n$ for an $LAI = 2.8$, which is typical for clipped grass. A very small fraction of G/R_n (0.036) has been adopted during daylight, even for full-cover alfalfa (0.04) and Verma et al. (1986) used 0.036 for a deciduous forest. For $LAI = 0$ (i.e., bare soil), Eq. (36) estimates $G/R_n = 0.4$.

For nighttime under grass forage, Allen et al. (1998) recommend $G = 0.5 R_n$ for grass and 0.2 for full-cover alfalfa. During late evening and early morning, values for G will be negative as heat flows upward toward the surface. In these reported ratios, the effects of seasonal variations are not addressed. The seasonal lag of soil temperature behind solar radiation can be expected to produce higher ratios.

Application of the PM equation

Besides direct application of the PM equation to a complex surface of soil and vegetation (Ortega-Farias et al., 2006; Leuning et al., 2008) or to soil only, the equation has been widely used to represent ET from a reference vegetation crop for use in irrigation management (e.g., Pruitt and Doorenbos, 1977; Allen et al., 1998, 2006; ASCE, 2016). The reference crop ET (ET_o) represents ET from either a cool season clipped grass or a tall full-cover alfalfa crop, depending on the parameters specified for aerodynamic and surface resistances (Allen et al., 1989; ASCE, 2016).

When the PM equation is used as an ET reference, actual ET for a soil or soil-vegetation surface is computed by multiplying ET_o by a crop coefficient, K_c , where K_c represents the ratio of actual ET to ET_o . The value for K_c of soil is varied according to the moisture level of the soil. Detailed information and values for K_c for many crops are included in Allen et al. (1998).

Fig. 3 shows a comparison of hourly ET estimated by the ASCE version of the PM equation and ET measured by a precision weighing lysimeter (essentially a large flower pot suspended below ground via mechanical balances). The PM estimates, based on measured solar radiation, air temperature, humidity, and wind speed, agreed very well with field measurements. Results are not always this good, however, and depend on the representativeness of values for roughness and for surface resistance. Sensitivity of the PM to each of the four weather components depends on climate and the relative strengths of each component relative to the others (Irmak et al., 2006; Estévez et al., 2009).

Limitations to applying the PM equation

The PM equation functions best when applied as a reference ET equation that represents a near maximum ET from vegetated surfaces. This is because when ET is near maximum, then H , from Eq. (3) is nearly zero so that T_o is close in value to T_a . When this occurs, Δ in Eq. (1) can be estimated accurately from T_a only. In addition, when $T_o \sim T_a$, R_n can be estimated correctly using only T_a . However, when ET flux from a surface is less than a reference or near maximum rate, then H becomes increasingly large and, from Eq. (4), $T_o > T_a$. This condition then requires the following: (i) Calculation of Δ requires the use of both T_o and T_a (Eq. 10); (ii) the R_{nl} calculation (Eq. 32) requires an estimate for surface temperature T_o rather than assuming $T_o \sim T_a$; (iii) the value for G is often larger than when H is nearly 0 and should consider the magnitude of T_o ; and (iv) numerical correction to r_{ah} should be performed when T_o is substantially different from T_a using near-surface boundary layer stability corrections (Eq. 5).

Because T_o is not generally measured, when T_o is substantially different from T_a , it must be estimated by iteratively calculating its value using the four components of the energy balance (Eq. 3). This requires iterative and consecutive recalculation for

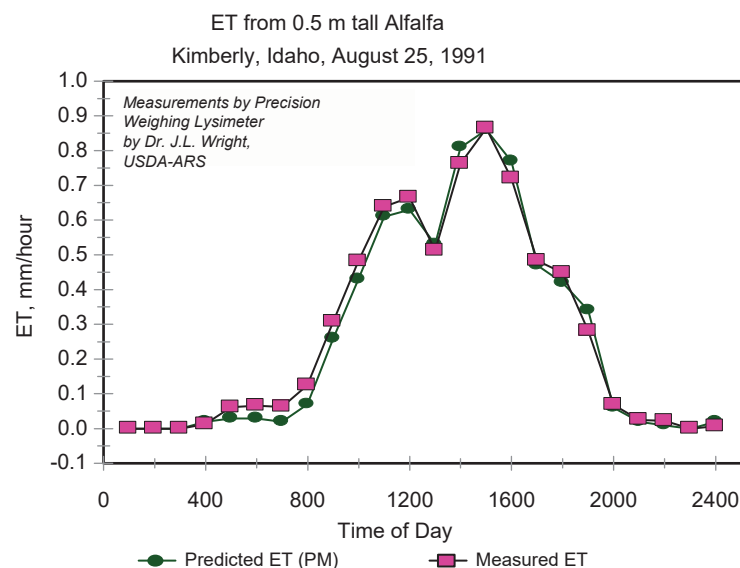


Fig. 3 Comparison between evapotranspiration (ET) during hourly periods predicted by the Penman–Monteith (PM) equation (using parameters from the American Society of Civil Engineers (ASCE)) and ET measured by a precision-weighing lysimeter.

(i) R_n (Eqs. 31 and 32); (ii) G (Eq. 36) or an equation based on T_o (ASCE, 2016); (iii) λE (Eq. 6); (iv) T_o using an inversion of Eq. (4) with H from an inversion of Eq. (3); (v) r_{ah} from Eq. (5), using boundary layer stability corrections that are based on H ; and (vi) Δ from Eq. (10). Iteration proceeds until a value for T_o has stabilized. When this occurs, then appropriate values for Δ , R_n , r_{ah} and G have been calculated and the iterated values can be used in PM (Eq. 1) so that accuracy of the equation is preserved. Ironically, when the above iterative solution, known as the AFIB approach (aerodynamic fluxes via iterative solution of the energy balance) from ASCE (2016), is applied, the need for the PM application “evaporates,” because λE has already been calculated from Eq. (6). Deployment of the AFIB solutions for T_o , Δ , R_n , r_{ah} and G into PM (Eq. 1) produces equivalent estimates to Eq. (6). As an example, applying PM (Eq. 1) in southern Idaho over dry rangeland conditions and using only T_a for the PM equation components overstates ET by about 30 W m^{-2} or 20% (ASCE, 2016, Table 11-6).

The PM method has been applied successfully using parameterization of R_n , r_{ah} and r_s parameters from satellite imagery (Consoli et al., 2006; Ciraolo et al., 2012; D’Urso et al., 2010, 2021). The utility of using remote sensing imagery is that spatial variation in LAI, albedo, plant density and maturity can be captured at a within-field parcel scale. Local calibration of parameters may be required.

The PM equation has even been applied to estimate ET in greenhouses (Seginer, 2002; Yan et al., 2019). However, the aerodynamic terms and potentially the net radiation terms in the PM may deviate substantially from equilibrium boundary layer conditions assumed during development of the PM.

Applying PM with gridded weather forecast data sets

Gridded weather data from weather forecast models are commonly used for ET, because they provide the necessary information on a spatial scale to estimate ET using PM, including the calculation of reference ET, ET_o . Data common to gridded weather data sets include solar radiation or sky cover, near-surface air temperature, humidity, and wind speed. ET_o is a vital parameter in agricultural water management and for most remote sensing (RS) approaches to ET estimation. However, a concern with using gridded data is that, in dry regions, the gridded data represent ambient weather conditions where T and humidity profiles extending above the surface have elevated T and lower e than what occurs over a well-watered “reference” surface as previously described. The standardized PM for computing ET_o requires that the near-surface profiles for air temperature, humidity, and wind speed are in equilibrium with the actively transpiring well-watered surface represented by ET_o . This is a basic requirement of applying the PM equation (ASCE, 2016; Blankenau et al., 2020; Paredes et al., 2018). As a consequence, if gridded weather data fail to represent the near-surface air properties of well-watered agricultural areas, then ET_o will tend to be overstated.

Current applications of the PM equation as a grass-based ET_o estimate include the global WaPOR Daily Reference Evapotranspiration 2.0 produced by the UN-FAO (2020), by the FAO AQUASTAT modeling (UN-FAO, 2021) and for the USA as part of the Climate Engine (Huntington et al., 2017).

Empirical approaches have been suggested for adjusting T and e data to better reflect well-watered reference conditions prior to ingestion into the PM (Jensen et al., 1997; ASCE, 2016). In addition, theoretical approaches have been offered that are based on boundary layer profile theory (Allen et al., 2021).

Conclusion

The Penman-Monteith evapotranspiration model was developed in 1965 to enable estimation of ET from a wide range of vegetation using only weather parameters collected in the air above the vegetation surface. This eliminated the need to measure surface temperature. The PM method has found most application to a well-watered grass reference surface where the ET estimate represents near maximum ET resulting from observed weather. The reference ET estimate is used as a basis to estimate ET from a range of vegetation and crop types.

References

- Allen RG, Jensen ME, Wright JL, and Burman RD (1989) Operational estimates of evapotranspiration. *Agronomy Journal* 81: 650–662.
- Allen RG, Pereira LS, Raes D, and Smith M (1998) Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements. United Nations FAO, Irrigation and Drainage Paper 56. Rome, Italy. 300 p <http://www.fao.org/docrep/X0490E/X0490E00.htm>.
- Allen RG, Pruitt WO, Wright JL, et al. (2006) A recommendation on standardized surface resistance for hourly calculation of reference ET by the FAO56 Penman-Monteith method. *Agricultural Water Management* 81(1–2): 1–22.
- Allen RG, Dhungel R, Dhungana B, Huntington J, Kilic A, and Morton C (2021) Conditioning point and gridded weather data under aridity conditions for calculation of reference evapotranspiration. *Agricultural Water Management* 245: 106531.
- ASCE, (2016). Evaporation, Evapotranspiration, and Irrigation Water Requirements, ASCE Manuals and Reports on Engineering Practice No. 70, (American Society of Civil Engineers, Reston 2016), M. E. Jensen, R. G. Allen (ed), 744 p.
- Blankenau PA, Kilic A, and Allen R (2020) An evaluation of gridded weather data sets for the purpose of estimating reference evapotranspiration in the United States. *Agricultural Water Management* 242: 106376.
- Brisson N (1996) On the importance of soil evaporation and rainfall interception in irrigation scheduling. In: *Evapotranspiration and Irrigation Scheduling, Proc. Int. Conf. of ASAE, San Antonio, TX*, 281–288.
- Brutsaert W (1975) Comments on surface roughness parameters and the height of dense vegetation. *Journal of the Meteorological Society of Japan* 53: 96–98.

- Campbell Scientific Inc (1999) *HFT3 Soil Heat Flux Plate*. 4 p.
- Choudhury BJ, Idso SB, and Reginato RJ (1987) Analysis of an empirical model for soil heat flux under a growing wheat crop for estimating evaporation by an infrared temperature based energy balance equation. *Agricultural and Forest Meteorology* 39: 283–297.
- Ciraolo G, Capodici F, D'Urso G, La Loggia G, and Maltese A (2012, April) Mapping evapotranspiration on vineyards: The Sentinel-2 potentiality. *First Sentinel-2 Preparatory Symposium*, Vol. 707, p. 32.
- Colaizzi PD, Kustas WP, Anderson MC, Agam N, Tolk JA, Evett SR, Howell TA, Gowda PH, and O'Shaughnessy SA (2012) Two-source energy balance model estimates of evapotranspiration using component and composite surface temperatures. *Advances in Water Resources* 50: 134–151.
- Consoletti S, D'Urso G, and Toscano A (2006) Remote sensing to estimate ET-fluxes and the performance of an irrigation district in southern Italy. *Agricultural Water Management* 81(3): 295–314.
- D'Urso G, Richter K, Calera A, et al. (2010) Earth observation products for operational irrigation management in the context of the PLEIADES project. *Agricultural Water Management* 98(2): 271–282.
- D'Urso G, Bolognesi SF, Kustas WP, et al. (2021) Determining evapotranspiration by using combination equation models with sentinel-2 data and comparison with thermal-based energy balance in a California irrigated vineyard. *Remote Sensing* 13(18): 3720.
- Estévez J, Gavilán P, and Berengena J (2009) Sensitivity analysis of a Penman–Monteith type equation to estimate reference evapotranspiration in southern Spain. *Hydrological Processes* 23(23): 3342–3353.
- Grant DR (1975) Comparison of evaporation from barley with Penman estimates. *Agricultural Meteorology* 15: 49–60.
- Huntington JL, Hegewisch KC, Daudert B, et al. (2017) Climate engine: Cloud computing and visualization of climate and remote sensing data for advanced natural resource monitoring and process understanding. *Bulletin of the American Meteorological Society* 98(11): 2397–2410.
- Irmak S, Payero JO, Martin DL, Irmak A, and Howell TA (2006) Sensitivity analyses and sensitivity coefficients of standardized daily ASCE-Penman-Monteith equation. *Journal of Irrigation and Drainage Engineering* 132(6): 564–578.
- Jensen DT, Hargreaves GH, Temesgen B, and Allen RG (1997) Computation of ETo under nonideal conditions. *Journal of Irrigation and Drainage Engineering* 123(5): 394–400.
- Lecina S, Martínez-Cob A, Pérez PJ, Villalobos FJ, and Baselga JJ (2003) Fixed versus variable bulk canopy resistance for reference evapotranspiration estimation using the Penman–Monteith equation under semiarid conditions. *Agricultural Water Management* 60(3): 181–198.
- Leuning R, Zhang YQ, Rajaud A, Cleugh H, and Tu K (2008) A simple surface conductance model to estimate regional evaporation using MODIS leaf area index and the Penman-Monteith equation. *Water Resources Research* 44(10).
- Monteith JL (1965) Evaporation and the environment. In: *The State and Movement of Water in Living Organisms, XIXth Symposium. Soc. For Exp. Biol., Swansea, Cambridge University Press*, 205–234.
- Or D and Lehmann P (2019) Surface evaporative capacitance: How soil type and rainfall characteristics affect global-scale surface evaporation. *Water Resources Research* 55: 519–539. <https://doi.org/10.1029/2018WR024050>.
- Ortega-Farías SO, Olioso A, Fuentes S, and Valdés H (2006) Latent heat flux over a furrow-irrigated tomato crop using Penman–Monteith equation with a variable surface canopy resistance. *Agricultural Water Management* 82(3): 421–432.
- Paredes P, Martins DS, Pereira LS, Cadima J, and Pires C (2018) Accuracy of daily estimation of grass reference evapotranspiration using ERA-Interim reanalysis products with assessment of alternative bias correction schemes. *Agricultural Water Management* 210(2018): 340–353. <https://doi.org/10.1016/j.agwat.2018.08.003>.
- Penman HL (1948) Natural evaporation from open water, bare soil and grass. *Proceedings. Royal Society of London A* 193: 120–146.
- Pruitt WO and Doorenbos J (1977) Background and development of methods to predict reference crop evapotranspiration (ET_o). In: *Appendix II, Crop Water Requirements, Irrigation and Drainage Paper No. 24. (rev.)*, pp. 108–119. Rome, Italy: FAO.
- Qualls R and Brutsaert W (1996) Effect of vegetation density on the parameterization of scalar roughness to estimate spatially distributed sensible heat fluxes. *Water Resources Research* 32(3): 645–652.
- Segner I (2002) Structures and environment: The Penman–Monteith evapotranspiration equation as an element in greenhouse ventilation design. *Biosystems Engineering* 82(4): 423–439.
- Thompson N, Barrie IA, and Ayles M (1981) *The Meteorological Office Rainfall and Evaporation Calculation System: MORECS. Great Britain Hydrological Memorandum No. 45*. 67 p.
- UN-FAO (2020) *WaPOR Database Methodology, V2 Release (Not Yet Published)*. Rome, Italy: Food and Agricultural Organization of the United Nations.
- UN-FAO (2021) Reference Evapotranspiration—AgERA5 Derived (Global—Monthly—~10km). https://data.apps.fao.org/static/data/index.html?prefix=static%2Fdata%2Fc3s%2FAGERA5_ET0_M.
- Vanderkemp PJ (1991) *Estimation of Crop Evapotranspiration by Means of the Penman-Monteith Equation*. Unpublished Ph.D. dissertation Logan, UT 84322-4105: Dept. of Agric. and Irrig. Engineering, Utah State University. 242 p.
- Verma SB, Baldocchi DD, Anderson DE, Matt DR, and Clement RJ (1986) Eddy fluxes of CO₂, water vapour, and sensible heat over a deciduous forest. *Boundary-Layer Meteorology* 36: 71–91.
- Wallace JS (1996) Modelling component transpiration in mixed plant communities. In: *Evapotranspiration and Irrigation Scheduling, Proc. Int. Conf. of ASAE, San Antonio, TX*, 585–591.
- Yan H, Acquah SJ, Zhang C, Wang G, Huang S, Zhang H, Zhao B, and Wu H (2019) Energy partitioning of greenhouse cucumber based on the application of Penman-Monteith and bulk transfer models. *Agricultural Water Management* 217: 201–211.

Further reading

- Allen RG, Smith M, Pereira LS, and Perrier A (1994) An update for the calculation of reference evapotranspiration. *ICID Bulletin* 43(2): 35–92.
- Alves I, Perrier A, and Pereira LS (1998) Aerodynamic and surface resistances of complete cover crops: How good is the “big leaf”? *Transactions of ASAE* 41(2): 345.
- ASCE (2005) *The ASCE Standardized Reference Evapotranspiration Equation. Report of the Task Committee on Standardization of Reference Evapotranspiration, Environmental and Water Resources Institute of the American Society of Civil Engineers*. Reston, VA: ASCE. 171 p.
- Choudhury BJ (1989) Estimating evaporation and carbon assimilation using infrared temperature data: Vistas in modeling. In: Ghassem A (ed.) *Theory and Applications of Optical Remote Sensing*, pp. 628–690. New York: John Wiley & Sons.
- Clothier BE, Clawson KL, Pinter PJ, Moran MS, Reginato RJ, and Jackson RD (1986) Estimates of soil heat flux from net radiation during the growth of alfalfa. *Agricultural and Forest Meteorology* 37: 319–329.
- Jagtap SS and Jones JW (1989) Evapotranspiration model for developing crops. *Transactions of ASAE* 32: 1342–1350.
- Jarvis PG (1976) The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field. *Philosophical Transactions of the Royal Society of London. Series B* 273: 593–610.
- Jensen ME, Burman RD, and Allen RG (1990) *Evapotranspiration and Irrigation Water Requirements. ASCE Manuals and Reports on Engineering Practice No. 70*. 350 p.
- Lafleur PM and Rouse WR (1990) Application of an energy combination model for evaporation from sparse canopies. *Agricultural and Forest Meteorology* 49: 135–153.
- Leuning R, Condon AG, Dunin FX, Ziegler S, and Denmead OT (1994) Rainfall interception and evaporation from soil below a wheat canopy. *Agricultural and Forest Meteorology* 67(3–4): 221–238.
- Lhomme JP (1988) A generalized combination equation derived from a multi-layer micrometeorological model. *Boundary-Layer Meteorology* 45: 103–115.

- Martin DL, Gilley JR, Carmack WJ, and Hardy LA (1993) Chapter 2: Irrigation water requirements. In: *Part 623 National Engineering Handbook*. National Resources Conservation Service. 284 p.
- Perrier A (1982) Land surface processes: Vegetation. In: Eagleson P (ed.) *Land Surface Processes in Atmospheric General Circulation Models*, pp. 395–448. Cambridge University Press.
- Price DT and Black TA (1989) Estimation of forest transpiration and CO₂ uptake using the Penman-Monteith equation and a physiological photosynthesis model. In: Black TA, Spittlehouse DA, Novak MD, and Price DT (eds.) *Estimation of Areal Evapotranspiration*, pp. 213–227. IAHS Pub. No. 177.
- Seginer I (1973) Aerodynamic roughness of vegetated surfaces. *Boundary-Layer Meteorology* 5: 383–393.
- Shuttleworth WJ and Wallace JS (1985) Evaporation from sparse crops—An energy combination theory. *Quarterly Journal of the Royal Meteorological Society* 111: 839–853.
- Shuttleworth WJ (2007) Putting the “vap” into evaporation. *Hydrology and Earth System Sciences* 11(1): 210–244.
- Stewart JB (1989) On the use of the Penman-Monteith equation for determining areal evapotranspiration. In: Black TA, Spittlehouse DL, Novak MD, and Price DT (eds.) *Estimation of Areal Evapotranspiration*, pp. 3–12. IAHS Pub. No. 177.
- Szeicz G and Long IF (1969) Surface resistance of crop canopies. *Water Resources Research* 5: 622–633.
- Thom AS (1971) Momentum absorption by vegetation. *Quarterly Journal of the Royal Meteorological Society* 97: 414–428.
- Thom AS (1972) Momentum, mass and heat exchange of vegetation. *Quarterly Journal of the Royal Meteorological Society* 98: 124–134.
- Wallace JS, Roberts JM, and Sivakumar MVK (1990) The estimation of transpiration from sparse dryland millet using stomatal conductance and vegetation area indices. *Agricultural and Forest Meteorology* 51: 35–49.
- Wright JL (1982) New evapotranspiration crop coefficients. *Journal of the Irrigation and Drainage Division* 108: 57–74.. ASCE.
- Zhang L and Lemeur R (1992) Effect of aerodynamic resistance on energy balance and Penman-Monteith estimates of evapotranspiration in greenhouse conditions. *Agricultural and Forest Meteorology* 58(3–4): 209–228.

Infiltration

Ty Ferre^a and Art Warrick^b, ^aHydrology and Atmospheric Sciences, University of Arizona, Tucson, AZ, United States; ^bEnvironmental Science, University of Arizona, Tucson, AZ, United States

© 2023 Elsevier Ltd. All rights reserved.

Two perspectives on infiltration	646
Steady-state infiltration with minimal surface ponding	646
An agricultural perspective – delivering water to the subsurface	647
Transient infiltration	647
References	651

Abstract

Infiltration is the flow of water across the ground surface into the soil. This process provides a critical link between surface hydrologic processes and subsurface soil and groundwater systems. While this flux occurs at the ground surface, full understanding of infiltration requires the energy status of water in the shallow subsurface, as well as insight into how these conditions change as infiltration continues. This article explains the basic concepts, and provides the foundational equations, needed to understand infiltration. These insights apply both to those who have a primary interest in infiltration as a mechanism that removes water from the surface as well as for those who are interested in the delivery of water to the soil.

Key points

- Two perspectives of infiltration: rate that water leaves above-ground systems and rate that water adds to the subsurface
- Infiltration described with Darcy's law.
- Equations presented for water infiltration across moisture gradients and from ponded surfaces.

Two perspectives on infiltration

Infiltration refers to the flux of water across the ground surface into the soil below. This simple process depends on many of the most complex aspects of water flow including the dependences of water content on water pressure and of hydraulic conductivity on water content and the significant storage capacity associated with changes in water content with time. Because it relates to the connection between two physical realms – above and below ground – infiltration also provides useful insight into the ways that a single phenomenon can be viewed differently from different practical and scientific perspectives. As an example, two views are presented here: infiltration as a control on the rate that water will leave above-ground systems and as a control on the rate at which water can be added to the subsurface. The latter is the traditional focus of soil scientists while the former informs hydrology, global circulation models, and many engineering disciplines.

For many practical applications and scientific fields, infiltration is a hydrologic process that removes water from the domain of interest. For example, if the objective is to predict flooding during a rainstorm, then infiltration determines how much of the rainwater will contribute to runoff and flooding. From this perspective, the most important system characteristic is the infiltration capacity. Namely, given the soil type, initial conditions, and water application rate at the soil surface, the infiltration capacity determines the fraction of the supplied water that infiltrates across the surface interface as a function of time.

Steady-state infiltration with minimal surface ponding

If there are no other impediments to flow in the subsurface (an impermeable layer or a shallow water table), the soil surface can admit infiltration at a rate close to its saturated hydraulic conductivity, (K_s [LT^{-1}]), continuously with no ponding. The qualification that infiltration is often less than K_s is due to the impacts of entrapped air in the soil, which can reduce the steady infiltration rate to between $K_s/3$ and $K_s/2$. This steady infiltration rate, q follows directly from Darcy's Law at the surface:

$$q = -K_s \frac{d(z + \varphi)}{dz} \quad (1)$$

where z is depth and φ is the pressure head. Specifically, the infiltration flux (q [LT^{-1}]) is equal to the saturated hydraulic conductivity such that flow is driven only by gravity. As a result, the pressure head is zero everywhere and there is no ponding of water above the ground surface. It follows from this that any applied flux that is less than K_s will not result in ponding, either. Rather, the soil water content will increase to a level at which the hydraulic conductivity at that water content is equal to the applied flux. The water content will again be constant with depth (assuming a homogeneous subsurface) and the pressure head of the water will be less than zero.

An agricultural perspective – delivering water to the subsurface

In agriculture-related fields, there is often interest in determining how much water can be supplied to the subsurface, with the clearest example being planning for irrigation. As explained above, water that is supplied at a rate that is equal to or less than K_s will not result in ponding. Therefore, regardless of the initial water content, the infiltration rate will be equal to the rate at which water is applied. The cumulative infiltration (I [L]) will equal the infiltration rate multiplied by the time for which water is applied. This would apply to spray or sprinkler irrigation that is maintained at a rate less than K_s , for example.

Transient infiltration

Because the hydraulic conductivity increases with increasing water content, it may seem counter-intuitive that a drier soil can admit infiltration at a faster rate than a wetter soil. This occurs because of the capillary forces that draw water into the drier soil. The first mathematical model of this influence was provided by [Green and Ampt \(1911\)](#). They consider a wetting front (zone of constant water content with depth extending from the ground surface) that advances due to water supplied at a constant pressure head φ_s [L] at the ground surface ([Fig. 1](#)). They define a pressure head at the wetting front, φ_{wf} [L], which has a value between the initial pressure head, φ_i [L], and φ_s . For any time since the beginning of infiltration, Δt , the cumulative infiltration to that time, I [L], must equal the depth of the wetting front, z_{wf} [L], multiplied by the change in volumetric water content across the wetting front, $\Delta\theta$ [$-$].

$$I = z_{wf} \Delta\theta \quad (2)$$

Darcy's Law can be applied over the entire wetting front at that time. Taking the datum elevation to be the ground surface, the hydraulic head at the wetting front, H_{wf} [L], and the ground surface, H_s [L], respectively, are:

$$H_{wf} = -z_{wf} + \varphi_{wf} \quad (3)$$

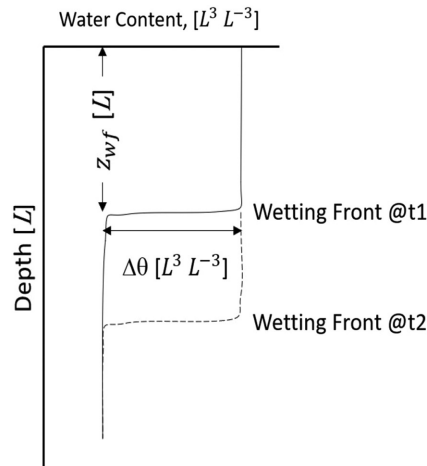


Fig. 1 Schematic representation of a wetting front with key variables.

$$H_s = 0 + \varphi_s \quad (4)$$

The infiltration flux, I [LT^{-1}] is then:

$$i = -K(\theta_{wf}) \frac{(0 + \varphi_s) - (-z_{wf} + \varphi_{wf})}{z_{wf}} = -K(\theta_{wf}) \frac{z_{wf} + (\varphi_s - \varphi_{wf})}{z_{wf}} \quad (5)$$

At long times, this states that the infiltration rate approaches the hydraulic conductivity of the medium at the water content above the wetting front. This shows steady state equivalence between conditions for which the rate of water supply or the pressure head of water at the ground surface are defined. However, at early times during the infiltration process, when the depth to the wetting front is relatively small, the difference in the pressure of the water over the wetting front leads to an infiltration rate that is greater than this asymptotic rate. Plainly stated, the infiltration rate in response to a time invariant pressure head at the ground surface will be highest initially and it will decrease exponentially to a constant value equal to the hydraulic conductivity of the medium at the water content that is associated with the maintained surface water pressure. Due to this change in infiltration rate with time, the wetting front will initially advance relatively rapidly, slowing to a constant rate at late time. Finally, note that the specific results depend on the soil hydraulic properties, which control $K(\theta_{wf})$, and the initial water pressure head, φ_i , which contributes to the value of $\Delta\theta$ over the wetting front. This analysis is commonly used to predict the infiltration rate with time beneath ponded water (positive values of ψ_s). But, some soil science instruments, such as a disk permeameter, have been developed that impose a constant negative water pressure; these also rely on this analysis.

Historically, [Green and Ampt \(1911\)](#) were the first to propose a model to describe infiltration rate as a function of the soil physical parameters. They assumed that a sharp wetting front exists, which separates two regions of constant water content above and below the front, that there is a constant, intermediate pressure head at the wetting front, and that the medium is homogeneous. This leads to following to description of the distance to the wetting front, (L_f [L]), with time as a function of K_s and the change in pressure head, $\Delta\varphi$, and the change in water content, $\Delta\theta$, across the wetting front:

$$L_f = \sqrt{\frac{2K_s t \Delta\varphi}{\Delta\theta}} \quad (6)$$

Example advances of a wetting front during horizontal infiltration following this model are shown in [Fig. 2](#) for a base case of: $K_s = 0.0001 \text{ cm min}^{-1}$; $\Delta\varphi = 30 \text{ cm}$, and $\Delta\theta = 0.15 \text{ cm}^3 \text{ cm}^{-3}$. While changes in any of the three parameters alter the calculated horizontal distance to the sharp wetting front with time, all conditions show similar patterns. The wetting front advances most quickly at early time, slowing as the wetting front moves farther from the source.

A similar expression can be developed for vertical infiltration, considering the influence of gravity. But, it requires root finding or implementation in a numerical model to apply it:

$$L_f - \Delta\varphi \ln\left(1 + \frac{L_f}{\Delta\varphi}\right) = \frac{K_s t}{\Delta\theta} \quad (7)$$

For some agricultural and engineering problems, including surface irrigation, it can be useful to predict the cumulative infiltration that will occur due to the supply of water at a constant pressure for a given time. Widely used relations have been proposed by [Philip \(1957, 1969\)](#), [Knight \(1973\)](#), [Warrick \(1975\)](#), [Brutsaert \(1977\)](#), and [Parlange et al. \(1985\)](#), and others. The simplest, foundational example, developed by [Philip \(1957\)](#) states:

$$i = St^{-0.5} + \gamma K_s \quad (8)$$

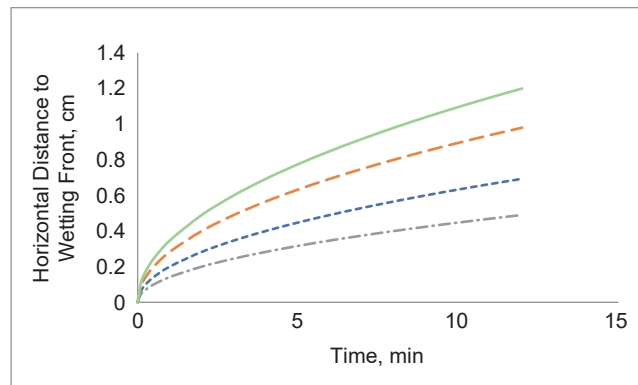


Fig. 2 Horizontal advance of a wetting front as a function of changes in the three variables included in the [Green and Ampt \(1911\)](#) equation. Blue dash is the base value. Orange dash is for large K_s . Gray dash is for large $\Delta\varphi$. Green line is for a large change in $\Delta\theta$.

Showing that there is a contribution that is driven by the sorptivity, $S [LT^{-0.5}]$ of the soil, which dominates at early time, and a steady contribution from gravity-driven infiltration that becomes dominant at late time. Note here the factor, γ , which reduces the steady-state infiltration rate due to the impacts of entrapped air, as presented previously.

Example infiltration rates through time for vertical infiltration are shown in Fig. 3 for a base case of: $K_s = 0.0001 \text{ cm min}^{-1}$; $S = 0.1 \text{ cm s}^{-0.5}$, and $\gamma = 0.8 \text{ cm}^3 \text{ cm}^{-3}$. Changes in any of the three parameters alter the time series of infiltration. But, the nature of the curves is highly similar, with large initial infiltration reducing relatively quickly to a near constant rate.

Other parametric expressions have been proposed that capitalize on the predictable transition of infiltration rate from high to steady values. Lewis (1937) proposed a power law decrease in infiltration rate with time:

$$i(t) = akt^{a-1} + i_0 \quad (9)$$

where a is a unitless empirical parameter and $k [LT^{-1}]$ is also treated as empirical with units that strictly vary with the value of a . The addition of the second term, often referred to as the modified Lewis-Kostiakov formula, accounts for a non-zero infiltration rate at late time.

Example infiltration rates through time for vertical infiltration are shown in Fig. 4 for a base case of: $i_0 = 0.0001 \text{ cm min}^{-1}$; $k = 0.1$, and $a = 0.5$. As for the Phillip equation, changing any single parameter changes the specific shape of the curve, but all of the time series show high initial infiltration reducing relatively quickly to a near constant rate.

A similar equation, based on exponentially decaying infiltration rate, was proposed by Horton (1933):

$$i(t) = (i_i - i_0)e^{-kt} + i_0 \quad (10)$$

where i_i is the initial infiltration rate and k is a soil-specific decay rate $[T^{-1}]$ (Fig. 5).

Like the preceding equations, it shows rapid initial infiltration with a steady decay through time. Historically, Horton was among the first to link the behavior of infiltration at the soil profile scale, with watershed scale overland flows (see review by Beven, 2021).

In many cases, infiltration delays the onset of overland flow (consider surface irrigation where applied water must travel over the surface to the end of the field). For these applications, it can be of particular interest to determine how long a given soil, at a given

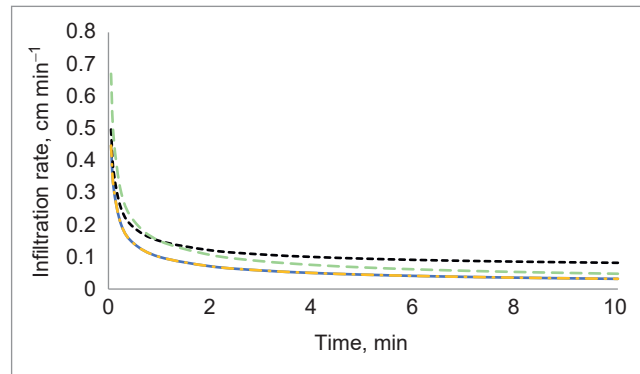


Fig. 3 Vertical advance of a wetting front as a function of changes in the three variables included in the Philip (1957) equation. Blue line is the base value. Black dash line is for large K_s . Green dash line is for large S . Orange dash line is large γ .

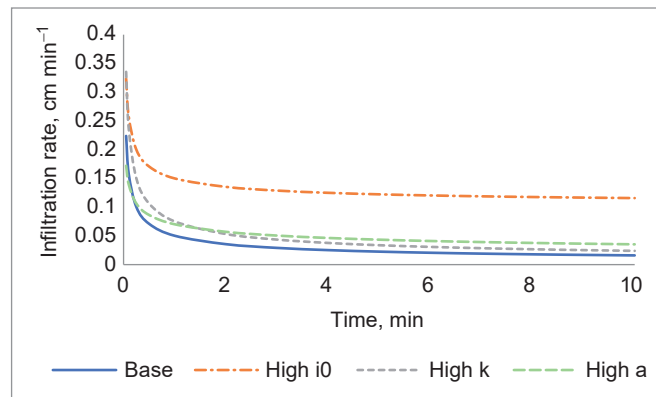


Fig. 4 Vertical advance of a wetting front as a function of changes in the three variables included in the Lewis-Kostiakov (Lewis, 1937) equation. Blue line is the base value. Orange dash line is large i_0 . Gray dash line is large k . Green dash line is large a .

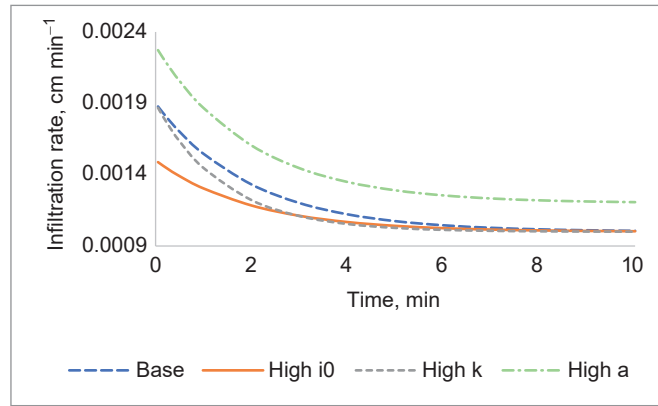


Fig. 5 Vertical advance of a wetting front as a function of changes in the three variables included in the Horton (1941) equation. Blue dash line is the base value. Orange line is large i_0 . Gray dash line is large k . Green dash line is large a .

initial water content, can accept water supplied at a given rate without producing overland flow. As explained above, ponding will only occur if water is applied at a constant rate that exceeds K_s . However, the preceding discussion of infiltration in response to water held at a constant water pressure at the ground surface indicates that a soil should be able to maintain an infiltration rate higher than K_s for some time; the time to ponding would depend on the soil hydraulic properties, the initial conditions, and the application rate. Based on the Philip (1969) expression for infiltration, the time to ponding, t_p [T], can be written as:

$$t_p = \frac{K_s \left(-\varphi_{wf} \right) \Delta\theta}{r(r - K_s)} \quad (11)$$

where r is the applied water rate [LT^{-1}]. Note that the saturated hydraulic conductivity is used because ponding will only occur if $r > K_s$, leading to fully saturated conditions. Some formulations, which depend on soil suction as -1 multiplied by the soil pressure head, do not include the negative sign on the wetting front pressure head. Furthermore, note that – all else being held constant – the time to ponding increases with increased K_s and change in water content across the wetting front and with decreased initial water pressure and water application rate. This analysis has been extended by Assouline et al. (2007) to predict the time of ponding for variable rate application (such as during a natural rainstorm). Considering the discussion of the time to ponding, irrigation could be supplied at a rate that exceeds K_s for some time without causing overland flow; the maximum time would depend on the soil hydraulic properties, initial water pressure in the soil, and the applied rate.

While the early equations describing infiltration focused on one-dimensional horizontal or vertical representations, many real conditions lead to infiltration that is two-dimensional (e.g., beneath an irrigation trench) or three dimensional (e.g., beneath a drip emitter). If water is supplied at a constant rate, such as by an emitter, then horizontal and vertical flow can be thought of as ‘competing’. Water is drawn horizontally by differences in pressure across the wetting front. Vertical water flow is driven by both pressure gradients and gravity. As a result, there should always be more vertical flow than horizontal. But, the degree of difference depends on the relative contributions of pressure and gravity to driving flow. Closed form equations describing multi-dimensional flow are complicated and it is currently more common to solve these problems using numerical models. As a general guideline, conditions that are driven predominantly by gravity (e.g., small pressure differences across the wetting front, relatively wet conditions, and relatively coarse-grained soils) tend to form vertically elongated wetted areas beneath the source. In contrast, conditions that lead to larger contributions from pressure differences (e.g., large pressure differences across the wetting front, relatively dry conditions, and fine-grained soils) result in more spherical wetted areas during infiltration. Regardless of the shape of the wetted area, a wetting front that experiences two- or three-dimensional flow will not progress as far in a given time as would be calculated using a vertical infiltration equation.

If water is supplied at a constant pressure head, then horizontal and vertical infiltration can be thought to combine. Conceptually, as vertical infiltration proceeds, horizontal infiltration draws some water away. This retards the vertical advance of the wetting front, maintaining higher infiltration rates for a longer time. The net result is that two- and three-dimensional infiltration beneath a constant pressure boundary, such as an irrigation furrow, will be greater than that calculated based on a one-dimensional vertical infiltration equation.

Infiltration is a process that occurs at relatively small scales. It can be described by water content and pressure differences that occur on the scale of meters. However, infiltration has impacts across scales. It is often useful to consider infiltration in the context of other hydrologic processes. For example, evapotranspiration describes the removal of water from the soil by plant roots and evaporation. Considering what drives infiltration, you could surmise that the rate of evapotranspiration (the combined result of the plant cover and the evapotranspiration rate of the plants present) would impact the rate of infiltration (Fig. 6). Furthermore, given that precipitation is partitioned into infiltration and runoff, plant cover should affect runoff. Comparing bare ground and a vegetated surface, it is reasonable to expect that the plant cover would reduce the water content in the subsurface, thereby

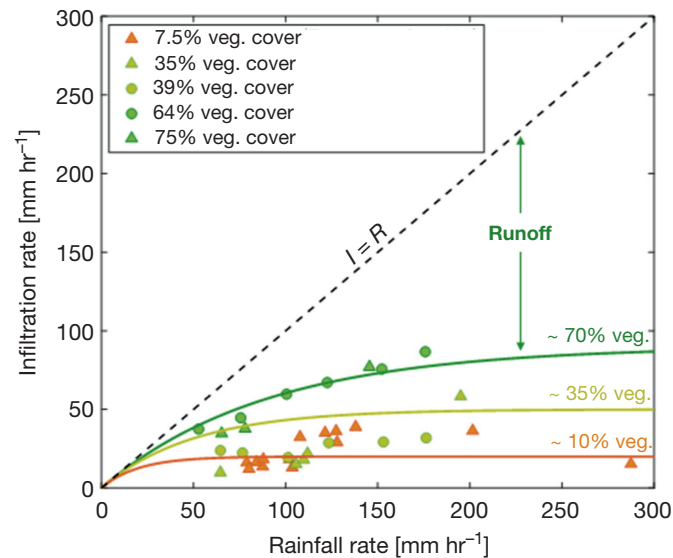


Fig. 6 Impact of vegetation on the ratio of infiltration to runoff as shown by Bonetti et al. (2021).

maintaining higher infiltration rates for longer. But, Bonetti et al. (2021) showed a more pronounced effect; increased vegetative cover led to changes in the soil structure, due to root growth and decay and associated biological activity, which led to a larger fraction of precipitation infiltrating, leading to a reduced rate of runoff. Similar lines of reasoning will help to predict how other conditions, from soil heterogeneity, to the presence of surface mulch, to surface compaction, to a sloping ground surface may impact infiltration rates. But, in many of these cases, quantitative predictions require the use of unsaturated flow models.

References

- Assouline S, Selker JS, and Parlange J-Y (2007) A simple accurate method to predict time of ponding under variable intensity rainfall. *Water Resources Research* 43: W03426.
- Beven K (2021) The era of infiltration. *Hydrology and Earth System Sciences* 25: 851–866. <https://doi.org/10.5194/hess-25-851-2021>.
- Bonetti S, Wei Z, and Or D (2021) A framework for quantifying hydrologic effects of soil structure across scales. *Communications Earth & Environment* 2: 107. <https://doi.org/10.1038/s43247-021-00180-0>.
- Brutsaert W (1977) Vertical infiltration in dry soil. *Water Resources Research* 13: 363–368. <https://doi.org/10.1029/WR013i002p00363>.
- Green WH and Ampt GA (1911) Studies on soil physics: 1, Flow of air and water through soils. *Journal of Agricultural Research* 4: 1–24.
- Horton RE (1933) The role of infiltration in the hydrologic cycle. *Transactions of the American Geophysical Union* 14: 446–460.
- Horton RE (1941) An approach toward a physical interpretation of infiltration-capacity. *Soil Science Society of America Journal* 5: 399. <https://doi.org/10.2136/sssaj1941.036159950005000C0075x>.
- Knight JH (1973) *Solutions of the Nonlinear Diffusion Equation: Existence, Uniqueness, and Estimation*. PhD Thesis, Canberra: Australian National University.
- Lewis MR (1937) The rate of infiltration of water in irrigation practice. *Transactions of the American Geophysical Union* 18: 361–368.
- Parlange J-Y, Haverkamp R, and Touma J (1985) Infiltration under ponded conditions: 1. Optimal analytical solution and comparison with experimental observation. *Soil Science* 139(4): 305–311.
- Philip JR (1957) The theory of infiltration, 1. The infiltration equation and its solution. *Soil Science* 83: 345–357. <https://doi.org/10.1097/00010694-195705000-00002>.
- Philip JR (1969) Theory of infiltration. *Advances in Hydroscience* 5: 215–296.
- Warrick AW (1975) Analytical solutions to the one-dimensional linearized moisture flow equation for arbitrary input. *Soil Science* 120(1975): 79–84.

Solute transport

MC Sukop^a and M Dentz^b, ^aFlorida International University, Miami, FL, United States; ^bSpanish National Research Council (IDAEA-CSIC), Barcelona, Spain

© 2023 Elsevier Ltd. All rights reserved.

Introduction	652
Transport processes	653
Advection	653
Diffusion	653
Dispersion	653
Dimensionality	654
Longitudinal dispersivity	655
Breakthrough curves	655
Adsorption and exclusion	657
Advection-dispersion equation and other solute transport models	659
Two-region advection-dispersion equation	660
Multi-rate mass transfer models	661
Continuous time random walk models	662
Fractional advection-dispersion models	663
Time-fractional advection-dispersion equations	663
Space-fractional advection-dispersion equation	664
Conclusion/summary/outlook	665
References	665
Further reading	665

Abstract

Solute transport, defined here as the movement of chemicals in soils and groundwater aquifers, plays a central role in biogeochemical cycling and many other processes. Solute transport can be dominated by advection, diffusion, or dispersion, or by any combination of these. The interaction of a chemical with soil or aquifer solids can slow transport by retardation. The Advection-Dispersion Equation and its two-region or two-site variants have traditionally been applied to solute transport. More modern approaches generalize the Advection-Dispersion Equation. Despite decades of intensive research, our a priori capability to predict solute transport in complex porous media such as soils remains limited.

Key points

- Solute transport plays a major role in soil processes.
- Advection, diffusion, and dispersion are the three modes of solute transport.
- Depending on flow conditions, dispersion can be dominated by diffusion or variations in solute velocity over a broad range of scales.
- Interaction of solutes with soil surfaces by physical or chemical adsorption can slow transport in a process referred to as retardation.
- In many soil applications, a one-dimensional treatment of solute transport may be adequate.
- The Advection-Dispersion Equation (sometimes including enhancements for two-region or two-site conceptualizations) has traditionally been applied to solve solute transport problems.
- Observed dispersivity based on fitting to the Advection-Dispersion Equation is known to increase with the scale of the transport process.
- More recent solute transport models anticipate different types of scaling depending on the structure of a medium's heterogeneity.

Introduction

The movement of chemicals in soils and groundwater aquifers (solute transport) plays a central role in global biogeochemical cycling including carbon emission and sequestration. Human activities have placed many new chemicals into the environment and population pressure continues to increase the loading of wastes and other chemicals to the subsurface. Salinization and other types

of soil degradation can impact food security. Applications of an understanding of solute transport in porous media include the efficacy and fate of agricultural chemicals (fertilizers and pesticides), impacts of irrigation, the treatment and disposal of wastewaters and wastewater treatment products from human, animal, and industrial sources, and contamination resulting from landfills, mining operations, industrial activities, and high-level radioactive waste disposal. Given its relevance and importance for the maintenance of safe water supplies, biogeochemical cycling, nourishment of plants, regulation of the fate of soil-applied chemicals, soil development, and many other critical hydrologic processes, solute transport has long been the focus of intense research in soil science, groundwater hydrology, and chemical engineering. Nevertheless, our a priori capability to predict solute transport remains limited.

Pores may contain water, soil gases, and introduced materials, such as non-aqueous-phase liquids. Chemicals can be transported in any or all of these three fluid phases. The general aspects of transport are similar in each possible phase, so the emphasis here and in much of the solute transport literature is on the transport of chemicals in aqueous solution. A solute is a chemical that is dissolved in a solvent to form a solution.

Three mechanisms operate to transport solutes in soils and aquifers. 'Advection' (or synonymously 'convection' in the solute transport literature) refers to the movement of chemicals with soil fluids and groundwater. Advection often dominates the solute transport process. Diffusion is the omnidirectional spreading of a chemical that results from thermal (Brownian) motion of the fluid and solute molecules. Dispersion results primarily from variations in fluid velocity that occur at all scales. Dispersion is anisotropic and enhanced in the overall direction of flow.

Breakthrough curves display the variation in solute concentration as a function of time or distance. These measurements and the advection-dispersion equation are the primary tools used in the evaluation of solute transport.

Solutes can interact strongly with soil surfaces by physical or chemical adsorption and their transport can be appreciably slowed in a process known as retardation. Repulsion or exclusion of solutes from soil surfaces can result in transport more rapid than that of the carrier fluid.

Transport processes

Advection

In the most simplistic sense, dissolved solutes are carried along with moving soil fluids. Here, 'fluid' refers to either a gas or a liquid phase; simultaneous treatment of interacting gas and liquid phases requires coupling of the treatment described below. A fluid flowing through a porous medium at a flux q (volume of fluid per unit area of soil per unit time) with a solute concentration C (mass of solute per unit volume of fluid) carries with it a solute flux $J_c = qC$ (mass of solute per unit area of soil per unit time). The mean velocity of the flow v (and the solute, assuming no losses, production, or interaction with the solid phase) can be computed from knowledge of the flux q and the volumetric fluid content θ via $v = q/\theta$. Under steady flow, $x = vt$ (where v is the magnitude of v), and the distance x traveled by an average solute molecule in time t , or the time t needed for the solute to traverse a given distance x , can be calculated. The advective solute flux J_c can also be written in terms of the mean velocity, the concentration, and the fluid content: $J_c = v\theta C$. In some instances, for example when the solute source is continuous and the solute concentration is uniform in the flowing fluid, advection dominates the solute transport process and may be the only process that need be considered.

Diffusion

Diffusion is the spreading of a solute due to thermally induced Brownian motion. It is generally considered to be omnidirectional and driven by concentration gradients according to concentration-independent coefficients, though close study reveals that it is more complicated when solute concentrations are large. Under the typical assumptions, diffusion can be treated with Fick's first law, which gives the diffusive solute flux J_s as a function of the concentration gradient ∇C and the bulk diffusion coefficient D_b , i.e., $J_s = -D_b \nabla C$.

In soils, the diffusion paths are obstructed by solid particles and solutes are constrained to follow the tortuous, connected pore network. The connected network changes as the fluid content changes. The effective diffusion coefficient D_e must therefore be computed from the bulk diffusion coefficient and the soil water content and porosity. A commonly used empirical expression for the estimation of the effective diffusion coefficient is $D_e = D_b \theta^{10/3} / \phi^2$, where ϕ is the porosity.

Dispersion

Consideration of saturated water flow in perhaps the simplest possible pore geometry—a straight capillary tube with circular cross section—provides insight into the processes that take place in the more tortuous and convoluted pore networks that characterize natural porous media. The velocity of fluid in such a tube depends on the pressure drop between the ends of the tube (a distance L apart), the tube's radius r , and the dynamic viscosity η . In general, the velocity is not uniform across the tube's cross section. In the laminar flow regime, generally thought to apply to water movement in most soils, the velocity profile in a capillary tube is parabolic:

$$v(y) = 2v \left(1 - \frac{y^2}{r^2} \right) \quad (1)$$

where $v(y)$ is the velocity as a function of radial distance y from the tube center.

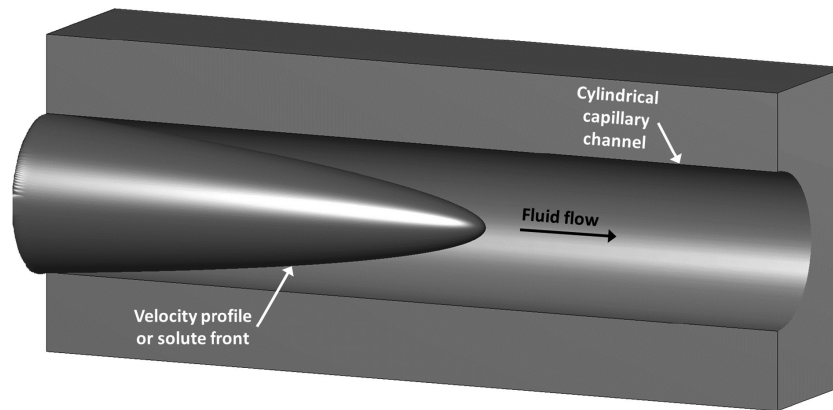


Fig. 1 Schematic velocity profile or solute front position in an idealized cylindrical capillary tube. Flow from left to right.

The mean velocity v can readily be determined from Stokes law in the form:

$$v = \frac{r^2 \Delta P}{8\eta L} \quad (2)$$

According to this model, the maximum velocity at the center of the tube ($y = 0$) is $2v$, and there is zero velocity at the tube's inner surface ($y = r$).

A solute introduced at concentration C_0 at the upstream end of the tube at time zero ($t = 0$) will be convected rapidly in the center of the tube and more slowly near the tube's surface (Fig. 1). Taking integrated samples from various cross sections along the tube at a later time ($t > 0$) and assuming no diffusion results in the collection of fluid containing no solute and fluid containing solute at the introduced concentration C_0 in various proportions. Dilution of the solute with solute-free fluid in the sample occurs according to the relative cross-sectional areas of the solute paraboloid and the solute-free fluid. The relative concentration as a function of time t or distance along the pipe x can be computed from $\frac{C}{C_0} = 1 - \frac{4\eta x}{Gr^2 t}$ (for $t > \frac{4\eta x}{Gr^2}$ and $x < \frac{Gr^2 t}{4\eta}$), where G is the pressure gradient over a length of the pipe ($\Delta P/L$) or a body force like the product of gravity and density (ρg) for a vertical tube. Thus, the solute front is spread out or dispersed due to velocity variations in the pipe.

In addition to the dispersion introduced by velocity variations within individual pores, the variable sizes and path lengths of pores in a natural porous medium and mixing at pore intersections also contribute to dispersion. Buoyancy factors may cause fingering and attendant mixing of a dense solution into a less-dense one. If the dispersion arises from such velocity and path-length variations, and hydrodynamic mixing alone, it is referred to as 'hydrodynamic dispersion.' The impact of variable velocities, path lengths, and mixing accumulates through the entire hierarchy of scales pertinent to the particular porous medium under consideration. Unsaturated conditions found in most soils generally add additional complexity to flow fields within individual pores and increase the tortuosity. Hence, the degree of dispersion is usually greater under partially saturated conditions than in the fully saturated case.

Diffusion results in a spreading of solute fronts similar to that induced by dispersion. It is generally not possible to separate out the effects of diffusion and dispersion. They are typically considered to be additive with a lumped diffusion–dispersion coefficient D_d given by $D_d = D_h + D_e$, where D_h is the coefficient of hydrodynamic dispersion. The diffusive–dispersive solute flux J_d is then treated assuming a Fick's law-like formulation: $J_d = -D_d \nabla C$.

The diffusion-dispersion coefficient has been notoriously difficult to estimate a priori and is generally determined via inverse modeling. Excellent computer programs for its determination from transport measurements are available (Parker and van Genuchten, 1984; Toride et al., 1995). However, this reliance on measurements of the actual transport process for parameter estimation is unfortunate, because the measurements are difficult and time-consuming, and can rarely be carried out at the scale of interest.

Inside a pore, the dimensionless Peclet number ($Pe \equiv vl/D_b$, with l a characteristic length) indicates the relative importance of diffusion and advection; large values of Pe indicate an advection-dominated process, while small values of Pe indicate the dominance of diffusion. Similarly, the dimensionless diffusion-dispersion coefficient $D^* \equiv D_d/D_b$ reflects the relative importance of hydrodynamic dispersion and diffusion. For porous media with well-defined characteristic lengths (i.e., bead diameter in packed beds of uniformly sized glass beads), D^* can be estimated from Pe . Different classes of behavior have been proposed based on the observed relationship between Pe and D^* (Fig. 2). Class I is characterized by very slow flow and the dominance of diffusion. Class II is transitional, with approximately equal and additive hydrodynamic dispersion and diffusion. Hydrodynamic dispersion dominates the dispersion process in classes III and above. In class III, the role of diffusion is still nonnegligible, but it ceases to be significant in class IV. Finally, in class V, the velocity is so high that the flow of many fluids is turbulent.

Dimensionality

The dimensionality of the solute transport process is an important factor. In many soil science applications, for example the uniform application of a pesticide to the soil surface, the process can frequently be treated as one-dimensional. However, in other instances

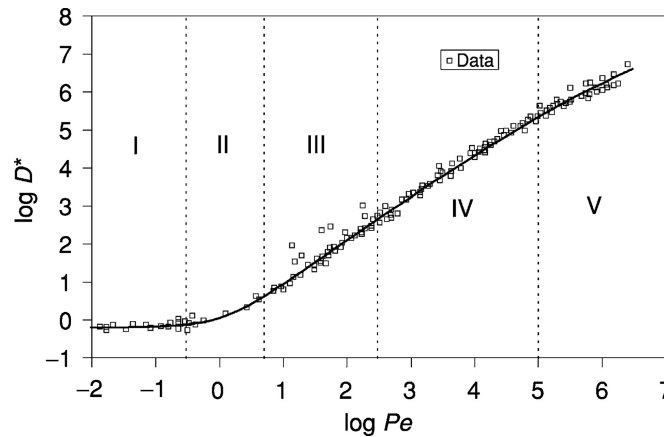


Fig. 2 Empirical relationship between dimensionless dispersion coefficient (D^*) and Peclet number (Pe) with data for uniformly sized particle beds. Roman numerals indicate behavioral class (see text). Adapted from Fried JJ and Combarous MA (1971) Dispersion in porous media. *Advances in Hydroscience* 7: 169–282, with permission.

two- or three-dimensional aspects of the transport process cannot be neglected. The problem can be viewed in terms of the geometrical configurations of the source and sink of interest. In the case of uniform soil application, the source is considered planar and the sink may be a planar water table. In contrast, a chemical spill limited to a small area overlying a thick vadose zone may need to be considered a point source in a fully three-dimensional system where solute may be dispersed both horizontally and vertically.

Even in a flow field that is macroscopically one-dimensional, dispersion is an anisotropic process and is enhanced in the direction of flow relative to transverse directions. A fully three-dimensional treatment of dispersion therefore requires a tensorial representation. For simplicity, we will restrict our discussion of solute transport to one-dimensional flow, where the effects of dispersion are embodied in a single scalar, the longitudinal dispersion coefficient, and the related longitudinal dispersivity.

Longitudinal dispersivity

The relationship between the one-dimensional diffusion-dispersion coefficient and mean pore-water velocity is frequently modeled with the empirical equation $D_d = m + \alpha v^n$, where m and n are fitting parameters, and α is the longitudinal dispersivity. In the hydrodynamic dispersion-dominated regime, the magnitude of m is negligible and n approaches unity. Under these conditions, the longitudinal dispersivity has dimensions of length and can be defined as $\alpha \equiv D_d/v$.

The dispersivity ideally would be a constant for a particular porous medium, and efforts have been made to relate it to other medium properties. In general, it has been observed that under saturated conditions the magnitude of the dispersivity depends upon the scale of the domain under consideration (Fig. 3). The reasons for this scale dependency are unclear, but, at a macroscopic scale, the phenomenon may be related in part to spatial variations in the underlying saturated hydraulic conductivity field. New approaches for treating the relationship between α and scale have recently appeared and are described further below.

Another important influence on dispersivity under saturated conditions is soil structure, as influenced by differences in texture. In general, sandy soils have a narrower range of pore sizes and a smaller dispersivity than finer-grained soils. The combined effects of measurement scale and soil texture on α for soil columns (based on empirical relationships between dispersivity and soil hydraulic parameters and power law scaling of dispersivity with length) are shown in Fig. 4.

Breakthrough curves

Plotting the downstream changes in solute concentration against time at a fixed position or against distance from the source at a fixed time results in a breakthrough curve. Breakthrough curves generated from a continuous source will be approximately sigmoidal (resembling a cumulative Gaussian distribution function) provided the column Peclet number ($Pe_c \equiv v L/D_d$, where L is the column length) is large enough. Complete breakthrough occurs when the effluent concentration (C) is equal to the influent concentration (C_0), giving a relative concentration $C/C_0 = 1$. If the solute is introduced as a short pulse, complete breakthrough may not be attained and the solute concentrations will eventually fall and return to background levels. In this case, a breakthrough curve will resemble a Gaussian density function for adequately large column Peclet numbers. Small column Peclet numbers (usually due to short columns or high dispersion coefficients) result in asymmetrical breakthrough curves. The full behavior of the breakthrough curve is described by an inverse Gaussian distribution.

It is common to plot breakthrough curves from soil columns as relative concentration against the number of pore volumes eluted from the column. Because the computation of eluted pore volume incorporates time and column length, it is a means of making the x -axis non-dimensional. The pore volume in a column of length L and cross-sectional area A is $LA\theta$. The volume of fluid eluted over time t is $Atv\theta$. Thus, the number of pore volumes T eluted at any given time t is $(Atv\theta)/(LA\theta)$ or $T = tv/L$.

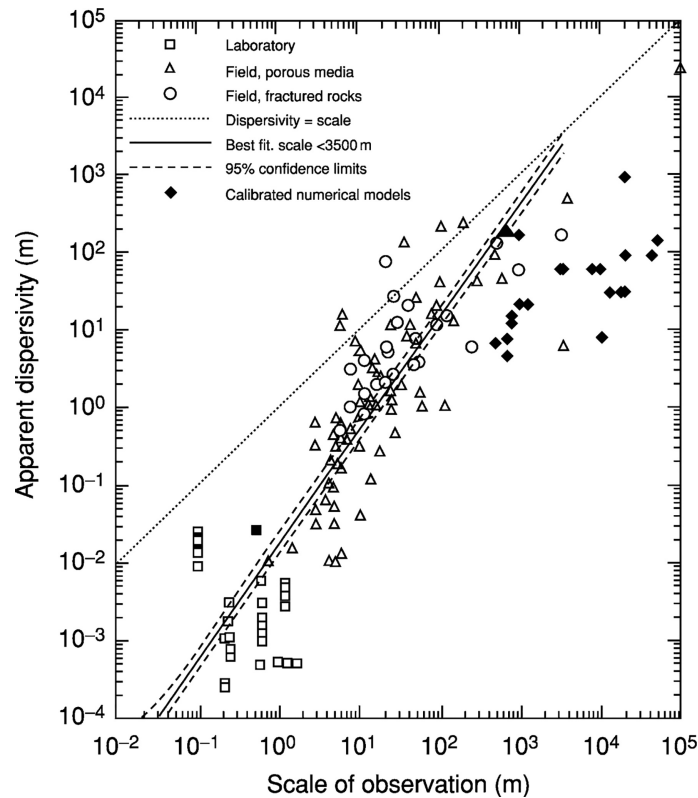


Fig. 3 Apparent dispersivity versus scale for laboratory and field measurements. Adapted from Neuman SP (1995) On advective transport in fractal permeability and velocity fields. *Water Resources Research* 31: 1455–1460, American Geophysical Union, with permission.

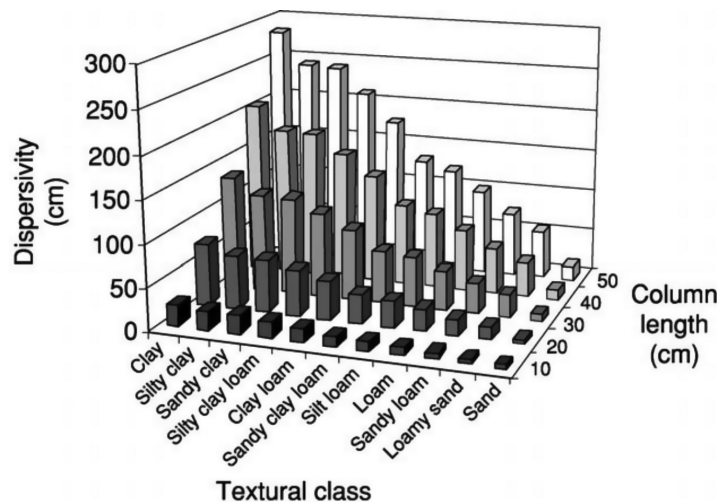


Fig. 4 Approximate effects of measurement scale and soil texture on dispersivity for soil columns based on empirical relationships between dispersivity and soil hydraulic parameters and power-law scaling of dispersivity with length.

This representation has instructive and conceptual benefits because, during piston-type flow where the solute front is not dispersed, it is fairly obvious that, after the input of one pore volume of solution containing solute at concentration C_0 into an initially solute-free column, the concentration of the effluent will immediately jump from $C = 0$ to $C = C_0$. This is shown in Fig. 5 where it corresponds to the curve with infinite column Peclet number. Decreasing the column Peclet number (e.g., by increasing the dispersion coefficient or shortening the column) causes spreading of the solute elution. Some solute travels faster than the mean velocity and elutes before one pore volume of fluid has entered the soil, while some travels slower than the mean and elutes after one pore volume of fluid is injected. When the column Peclet number is relatively high (Fig. 5; $Pe_c = 1000$ and $Pe_c = 100$), the early- and late-arriving solute masses are approximately equal, and the breakthrough curve is approximately sigmoidal and well described by a cumulative Gaussian distribution. The value of C/C_0 at one pore volume on this curve is approximately 0.5; this is considered a

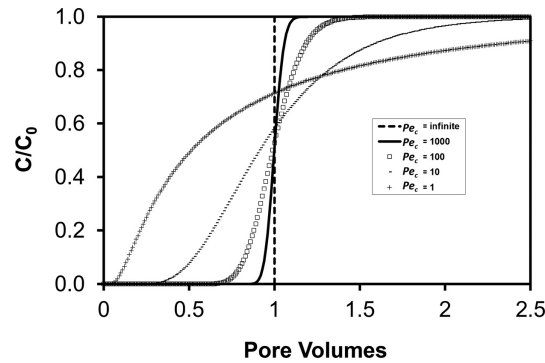


Fig. 5 Effect of column Peclet number (Pe_c) on relative concentration (C/C_0) in column effluent for a step-change application of solute.

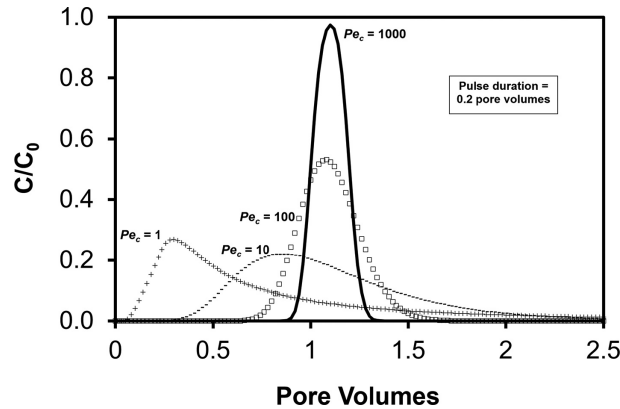


Fig. 6 Effect of column Peclet number on concentration in column effluent for a pulse application of solute. Pulse duration is 0.2 pore volumes.

hallmark of an ‘ideal’ solute transport process by many practitioners, with deviations calling for elaborate modifications to standard theory. Inspection of the gross asymmetry introduced by decreases in the column Peclet number (Fig. 5; $Pe_c = 10$ and $Pe_c = 1$) indicates that such deviations should be expected under many conditions. This effect is pronounced for solute pulses, and is perhaps somewhat counterintuitive in that smaller column Peclet numbers (more dispersion) can actually lead to higher peak concentrations than larger ones (Fig. 6; $Pe_c = 1$ vs. $Pe_c = 10$).

Adsorption and exclusion

Adsorption of solutes on to soil solids generally causes them to move at a fraction of the velocity of the fluid. The most simplistic (and common) treatment of adsorption involves the assumption of a linear adsorption isotherm that relates the adsorbed concentration on soil solids (C_s , mass of solute per unit mass of soil) to the concentration in solution at equilibrium via $C_s = K_d C$, where K_d is the soil/water partitioning or distribution coefficient with dimensions $L^3 M^{-1}$. The linear isotherm is primarily applicable to low solute concentrations and adsorption of a nonspecific physical nature.

K_d can be measured independently in the laboratory by mixing soil and solution and determining the equilibrium partitioning of the solute between the solid and fluid. In practice, however, the K_d of organic solutes is often estimated from the organic carbon partitioning coefficient $K_{oc}(L^3 M^{-1})$ using $K_d = K_{oc} f_{oc}$, where $f_{oc}(MM^{-1})$ is the mass fraction of organic carbon in a soil. There are many tables of K_{oc} values available (e.g., Table 1) and f_{oc} is a commonly measured soil property. Solutes with low K_d values have a low affinity for soil surfaces and, for a conservative solute, $K_d = 0$. Other solutes can be strongly bound to soil and have very high K_{oc} and K_d values.

Exclusion, in which a solute is repelled from solid surfaces and hence concentrated in more rapidly flowing fluid away from pore walls, has also been observed. In particular, negatively charged anions are often excluded from negatively charged soil mineral surfaces. Thus, they may elute earlier from a column than the carrier fluid with which they were introduced.

Adsorption and exclusion phenomena are often incorporated into solute transport models through the use of a retardation factor, R , given by:

$$R = 1 + \frac{\rho_b K_d}{\theta} \quad (3)$$

where ρ_b is the soil bulk density. For conservative solutes, $R = 1$, for sorbing solutes, $R > 1$, and, for excluded solutes, $R < 1$.

Table 1 Organic carbon partitioning coefficients for nonionizable organic compounds.

<i>Compound</i>	<i>Mean K_{oc} (L kg⁻¹)</i>
Acenaphthene	5028
Aldrin	48,686
Anthracene	24,362
Benz(a)anthracene	459,882
Benzene	66
Benzo(a)pyrene	1,166,733
Bis(2-chloroethyl)ether	76
Bis(2-ethylhexyl)phthalate	114,337
Bromoform	126
Butylbenzylphthalate	14,055
Carbon tetrachloride	158
Chlordane	51,798
Chlorobenzene	260
Chloroform	57
DDD	45,800
DDE	86,405
DDT	792,158
Dibenz(a,h)anthracene	2,029,435
1,2-Dichlorobenzene (<i>o</i>)	390
1,4-Dichlorobenzene (<i>p</i>)	687
1,1-Dichloroethane	54
1,2-Dichloroethane	44
1,1-Dichloroethylene	65
<i>trans</i> -1,2-Dichloroethylene	38
1,2-Dichloropropane	47
1,3-Dichloropropene	27
Dieldrin	25,604
Diethylphthalate	84
Di- <i>n</i> -butylphthalate	1580
Endosulfan	2040
Endrin	11,422
Ethylbenzene	207
Fluoranthene	49,433
Fluorene	8906
Heptachlor	10,070
Hexachlorobenzene	80,000
α -HCH (α -BHC)	1835
β -HCH (β -BHC)	2241
γ -HCH (Lindane)	1477
Methoxychlor	80,000
Methylbromide	9
Methylchloride	6
Methylene chloride	10
Naphthalene	1231
Nitrobenzene	141
Pentachlorobenzene	36,114
Pyrene	70,808
Styrene	912
1,1,2,2-Tetrachloroethane	79
Tetrachloroethylene	272
Toluene	145
Toxaphene	95,816
1,2,4-Trichlorobenzene	1783
1,1,1-Trichloroethane	139
1,1,2-Trichloroethane	77
Trichloroethylene	97
<i>o</i> -Xylene	241
<i>m</i> -Xylene	204
<i>p</i> -Xylene	313

Adapted from US Environmental Protection Agency, Soil Screening Guidance: Technical Background Document.

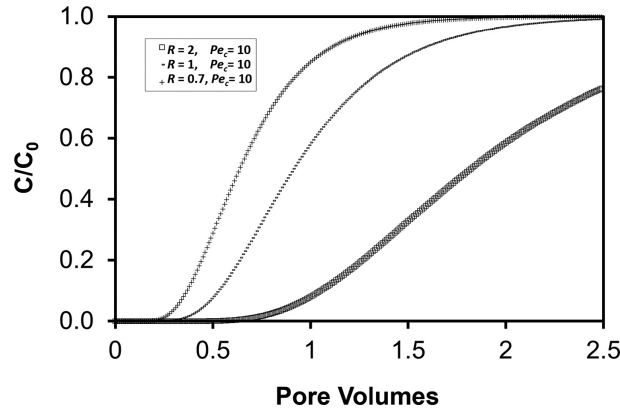


Fig. 7 Effects of retardation (R) and exclusion on breakthrough curve at $Pe_c = 10$.

Fig. 7 shows the effects of solute adsorption and exclusion relative to the breakthrough curve for a conservative solute ($R = 1$, $Pe_c = 10$, in Fig. 7). The retarded curve ($R = 2$, $Pe_c = 10$, in Fig. 7) shows a considerable delay in solute elution. Note also that the degree of solute dispersion is enhanced, even though the dispersion coefficient itself is unchanged. Conversely, the curve representing anion exclusion ($R = 0.7$, $Pe_c = 10$, in Fig. 7) shows early elution of the solute and less dispersion of the solute front.

For a continuous source under piston flow conditions, the solute front advances without dispersion and the velocity of the front (v_s) relative to the mean velocity of the fluid is $v_s = v/R$. With moderate dispersion and adequate transport length, the front assumes a sigmoidal shape, and v_s represents the position of the front defined by $C/C_0 = 0.5$. For a pulse source, v_s characterizes the velocity of the solute center of mass. In some instances, these simple relations may be adequate for estimating the solute front position or peak arrival. However, dispersion can play a critical role in determining the peak concentration and its duration. Evidence suggests that brief exposures to high concentrations may be exceptionally detrimental to human health in some instances.

Advection-dispersion equation and other solute transport models

The advection-dispersion equation (ADE, or convection-dispersion equation, CDE) has long been used as a model for solute transport. It can be derived from the flux terms described above (here taken as one-dimensional in the x direction) and a solute mass balance. The flux of solute due to convection is $J_c = v\theta C$, while the diffusive-dispersive flux is $J_d = -D_d \partial C / \partial x$. Combining these two fluxes, we have $J = J_c + J_d = v\theta C - D_d \partial C / \partial x$. Consider a volume of soil (Fig. 8) and steady one-dimensional flow. If the inflow of solute mass at $x = 0$ and the outflow at $x = L$ are not equal, then mass must accumulate in or be lost from the soil volume. Assuming for the moment that the solute can only be stored in the fluid, then either the volume of fluid at a constant concentration—given by $AL\theta$ —must change, or the solute concentration (C) in a fixed volume of fluid must change. The dimensions of the volume AL are constant and, when the flow is steady, the fluid content θ is constant. Therefore, only the concentration can change. The time rate of change in solute mass inside the volume is $AL\theta dC/dt$.

The difference in the inflow and outflow of mass per unit time can be written as $(\text{mass}_{\text{out}} - \text{mass}_{\text{in}})/\Delta t$. The flux J is the mass per unit area and time, so an equivalent expression is $A(J_{x=L} - J_{x=0})$. The spatial rate of change in solute flux is $\partial J / \partial x$, and the difference between the outflow and inflow is $A[J_{x=0} - (J_{x=0} + L \partial J / \partial x)]$. Equating this difference to the change in the stored mass gives:

$$AL\theta \frac{dC}{dt} = A \left[J_{x=0} - \left(J_{x=0} + L \frac{\partial J}{\partial x} \right) \right] \quad (4)$$

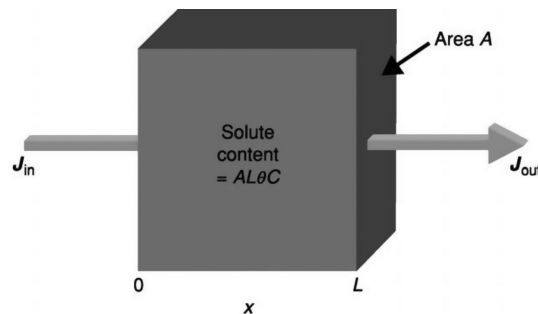


Fig. 8 Elemental volume, with one-dimensional solute fluxes. J , flux; C , concentration; L , length; θ , volumetric fluid content.

which readily simplifies to:

$$\theta \frac{\partial C}{\partial t} = - \frac{\partial J}{\partial x} \quad (5)$$

Incorporating our earlier expression for the solute flux J , we have:

$$\theta \frac{\partial C}{\partial t} = - \frac{\partial (v\theta C - D_d \frac{\partial C}{\partial x})}{\partial x} \quad (6)$$

Finally, with D_d , v , and θ constant, the ADE is:

$$\frac{\partial C}{\partial t} = -v \frac{\partial C}{\partial x} + \frac{D_d}{\theta} \frac{\partial^2 C}{\partial x^2} \quad (7)$$

Eq. (7) is often written with D_d/θ replaced by an apparent diffusion-dispersion coefficient D .

The ADE can readily be modified to incorporate adsorption by accommodating solute storage on the soil solids in addition to storage in the fluid phase. Thus, the storage term becomes $AL\theta C + AL\rho_b C_s$ or, in terms of the solute concentration in the fluid and assuming a linear isotherm, $AL(\theta C + \rho_b K_d C)$. Then Eq. (6) becomes:

$$(\theta + \rho_b K_d) \frac{\partial C}{\partial t} = - \frac{\partial (v\theta C - D_d \frac{\partial C}{\partial x})}{\partial x} \quad (8)$$

Note that this formulation also assumes that the adsorption equilibrium is attained instantaneously. Alternative derivations that remove this restriction are available. Removing the assumed constant variables (v , q and D_d) from the space derivative and dividing by θ yields:

$$R \frac{\partial C}{\partial t} = -v \frac{\partial C}{\partial x} + \frac{D_d}{\theta} \frac{\partial^2 C}{\partial x^2} \quad (9)$$

where R is the retardation factor given by Eq. (3).

Other isotherms, including the Langmuir (which introduces an adsorption maximum) and the Freundlich (which introduces a nonlinear concentration dependence), have been used to describe adsorption during solute transport. Realistic transport simulations of reactive species must often rely on considerably more complex, coupled reaction-transport models.

Many chemical species are susceptible to production or decay in the soil. Zero-order decay proceeds at a rate that is independent of the concentration of the substance where $\partial C/\partial t = kC$, where k is the rate constant (T^{-1}). Such potential sinks and sources, as well as physical sinks (e.g., root uptake) can readily be incorporated into the ADE.

Estimates of the parameters in the ADE can be obtained by moment analysis of solute breakthrough curves. A number of analytical solutions for the ADE are also available. Depending on the concentration detection mode (column effluent or 'resident' pore water) the following solutions have been recommended for a step change in concentration from C_0 to C at time $t = 0$ (Kreft and Zuber, 1978):

$$\text{Flux mode : } \frac{C}{C_0} = \frac{1}{2} \operatorname{erfc} \left[\frac{Rx - vt}{2(DRt)^{\frac{1}{2}}} \right] + \frac{1}{2} \exp \left(\frac{vx}{D} \right) \operatorname{erfc} \left[\frac{Rx - vt}{2(DRt)^{\frac{1}{2}}} \right] \quad (10)$$

$$\text{Resident mode : } \frac{C}{C_0} = \frac{1}{2} \operatorname{erfc} \left[\frac{Rx - vt}{2(DRt)^{\frac{1}{2}}} \right] + \frac{1}{2} \left(\frac{v^2 t}{\pi D r} \right)^{\frac{1}{2}} \operatorname{erfc} \left(- \frac{(Rx - vt)^2}{4DRt} \right) - \frac{1}{2} \left(1 + \frac{vx}{D} + \frac{v^2 t}{DR} \right) \exp \left(\frac{vx}{D} \right) \operatorname{erfc} \left[\frac{Rx + vt}{2(DRt)^{\frac{1}{2}}} \right] \quad (11)$$

Here 'exp' and 'erfc' are the exponential function and complementary error function, respectively. The differences between the two solutions become appreciable in highly dispersive systems. Fig. 9 shows the solutions at $Pe_c = 1$. More complex and multidimensional formulations of the ADE generally require numerical solutions.

Two-region advection-dispersion equation

This equation, also referred to as the 'mobile-immobile water model' (MIM), is a special case of the ADE, in which the liquid phase is conceptually divided into two distinct (mobile and immobile) regions. Advective-dispersive solute flux occurs within the mobile water region, with solutes entering and leaving the immobile water region by diffusion. The ADE for solute transport within the mobile water region can then be written as:

$$\frac{\partial C_m}{\partial t} = D \frac{\partial^2 C_m}{\partial x^2} - v_m \frac{\partial C_m}{\partial x} - \left(\frac{\theta_{im}}{\theta_m} \right) \frac{\partial C_{im}}{\partial t} \quad (12)$$

where the subscripts 'm' and 'im' refer to the mobile and immobile regions, respectively. The kinetic term $\theta_{im} dC_{im}/dt$ is usually expressed in terms of a first-order rate constant k and the concentration difference $C_m - C_{im}$. Eq. (12) can be solved analytically with or without making any assumptions about the spatial arrangement of solids and voids. In the first case, the mobile and immobile

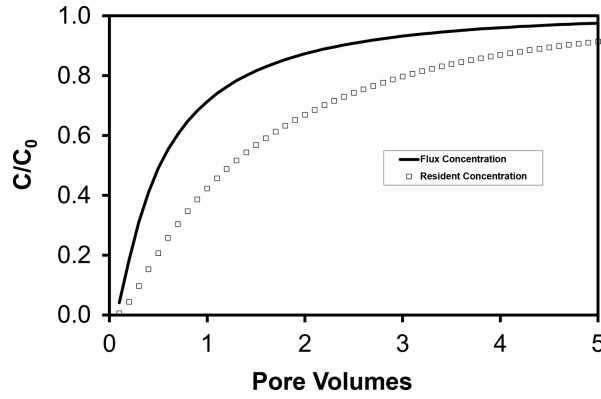


Fig. 9 Effect of concentration detection mode on breakthrough curve at $Pe_c = 1$.

liquid regions are explicitly defined in terms of pore-space geometry. The majority of these models are for packed beds of uniform-sized, spherical soil aggregates; here θ_m is the macropore or inter-'aggregate' water content, and θ_{im} is the intra-'aggregate' water content. While it is possible to extend this approach to nonuniform and non-spherical aggregates, its predictive capability is limited. This is because the relationship between packed beds of sieved aggregates and the pore characteristics of undisturbed soil is far from clear. In the second case, the mobile-immobile water fraction is simply a fitting parameter.

Eq. (12) is commonly employed to parameterize asymmetric solute breakthrough curves with 'heavy tails.' Deviations from the Gaussian distribution associated with the ADE are often observed in breakthrough curves where more of the observed distribution is in the tails. Improvements in fit associated with the MIM as compared to the ADE can probably often be attributed to the presence of additional model parameters. Since it is rarely possible to measure the mobile and immobile water fractions independently, this parameter is usually estimated inversely. A two-site conceptualization of the soil, where a fraction of the adsorption sites are considered to have non-instantaneous adsorption, leads to an equivalent mathematical formulation.

Multi-rate mass transfer models

The multi-rate mass transfer (MRMT) approach (Haggerty and Gorelick, 1995; Villermanx, 1987) generalizes the MIM presented in the previous Section in that it accounts for different types of linear mass transfer processes that can occur at different rates. This includes diffusion into immobile regions, and kinetic first-order surface reactions. The MRMT model describes advective-dispersive transport in a single mobile continuum, as the MIM, and linear mass transfer between the mobile and an ensemble of immobile continua. The weight of the j th immobile continuum is denoted by p_j . The total immobile concentration c_{im} is obtained by the weighted sum over the local immobile concentrations $c_{im}^{(j)}$,

$$C_{im} = \sum_j p_j \beta_j C_{im}^{(j)} \quad (13)$$

where $\beta_j = R_{im}\theta_{im}/R_m\theta_m$ is the capacity coefficient of the j th immobile zone, with R_m the retardation coefficient of the mobile region, and $R_{im}^{(j)}$ and $\phi_{im}^{(j)}$ the retardation coefficients and saturations of the immobile regions. The immobile concentrations are related to the mobile concentration through the convolution product.

$$C_{im}^{(j)} = \int_0^t dt' \varphi^{(j)}(t-t') C_m(t'). \quad (14)$$

The memory function $\varphi^{(j)}(t)$ encodes the information on the physics of the linear mass transfer process. Combining Eqs. (13) and (14), one can write the total immobile concentration as.

$$C_{im} = \beta \int_0^t dt' \varphi(t-t') C_m(t') \quad (15)$$

where the global memory function is defined by $\varphi(t) = \sum_j p_j \beta_j \varphi^{(j)}(t)$ and the capacity coefficient by $\beta = \sum_j p_j \beta_j$. The governing equation for the mobile solute concentration is then given by

$$\frac{\partial C_m}{\partial t} + \beta \frac{\partial}{\partial t} \int_0^t dt' \varphi(t-t') C_m(t') + v \frac{\partial C_m}{\partial x} - D \frac{\partial^2 C_m}{\partial x^2} = 0. \quad (16)$$

For linear first-order kinetic mass transfer, the memory function is given by $\varphi^{(j)}(t) = k_i \exp(-k_i t)$, where k_i denotes the mass transfer rate. For diffusive mass transfer into layered immobile regions, the memory function can be expressed in Laplace space by (Harvey and Gorelick, 1995)

$$\bar{\varphi}^{(j)}(\lambda) = \frac{1}{\sqrt{\lambda \tau_D^{(j)}}} \tanh\left(\sqrt{\lambda \tau_D^{(j)}}\right), \quad (17)$$

where $\tau_D^{(j)} = l_j^2/D$ is the characteristic diffusion time within the immobile layer of thickness l_j . The Laplace transform is defined in (Abramowitz and Stegun, 1972), Laplace transformed quantities are marked by an overbar, the Laplace variable is denoted by λ . Memory functions for spherical and cylindrical immobile geometries can be found for example in Harvey and Gorelick (1995). Memory functions for diffusion into immobile zones characterized by heterogeneous diffusion properties and irregular geometries are discussed in Gouze et al. (2008). The MRMT model can in principle be derived by volume averaging of a composite heterogeneity model that distinguishes a connected mobile region and unconnected immobile regions. The averaging volume is large enough such that it contains a representative sample of immobile zones that does not vary in space and can be characterized by the probability distribution p_j . In other words, the heterogeneity distribution is assumed to be stationary. This is why at each point in space, the MRMT model quantifies in principle a distribution of concentration values and not only a single value as advection-dispersion models do.

Based on this phenomenology, the MRMT model can in principle be parameterized based on the characterization of the geometry and mass transfer processes of the immobile regions. Examples are layered media such as single fracture- matrix systems (Maloszewski and Zuber, 1985). In this case, the memory function decays as $\varphi(t) \sim t^{-1/2}$, which implies a decay of the breakthrough curve in response to a pulse injection as $t^{-3/2}$, which is characteristic of matrix diffusion. Fig. 10 shows breakthrough data from a tracer experiment in fractured rock that shows these signatures. The MRMT model captures the long time tailing and the delay in the peak behavior (Hadermann and Heer, 1996). Depending on the distributions of characteristic mass transfer time scales, the MRMT approach can model power-law tailing of solute breakthrough curves (Haggerty et al. 2000), and general breakthrough curve tailing.

Continuous time random walk models

The continuous time random walk (CTRW) approach (Berkowitz et al. 2006) accounts for the fact that mass transfer times in heterogeneous media are distributed. It has found broad application for the modeling of non-Fickian transport in porous and fractured media from pore to regional scale. The CTRW approach is based on a similar phenomenology as the MRMT approach. In fact, it has been shown, that the two approaches are, under certain conditions, mathematically equivalent (Dentz and Berkowitz 2003).

The CTRW approach models solute transport through the propagation of solute particles over distributed space and time increments. Note that the spatial particle distribution is a measure for the solute concentration. The distribution of space and time increments reflects the medium and microscale transport properties. To illustrate this approach, we consider media that are

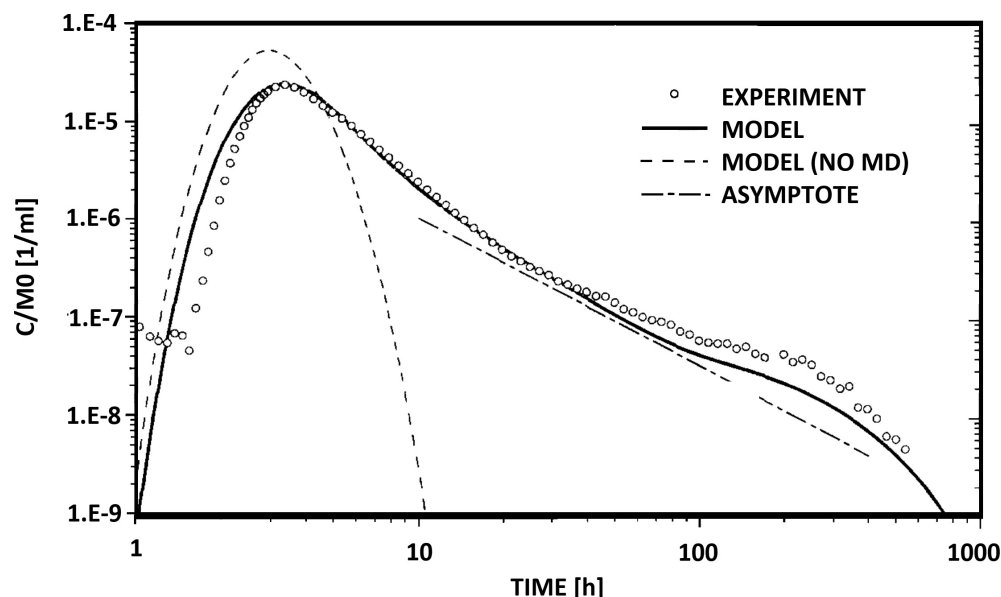


Fig. 10 Experimental breakthrough curve for uranine (circles) fitted by a MRMT model for matrix diffusion (MD, solid line), and ADE estimate (dashed line). Adapted from Fig. 3 in Hadermann J and Heer W (1996) The Grimsel (Switzerland) migration experiment: Integrating field experiments, laboratory investigations and modelling. *Journal of Contaminant Hydrology* 21(1–4): 87–100. doi: 10.1016/0169-7722(95)00035-6.

characterized by a single characteristic heterogeneity length scale l_c , which may be the pore length for pore-scale applications, or the correlation scale of log-hydraulic conductivity for Darcy-scale applications. Mass transfer properties such as velocity or retardation, are assumed to be constant over this length scale. Thus, as a solute particle moves through the medium, its motion can be modeled through spatial transitions over the length l_c associated with a transition time τ . Subsequent transition times are modeled as independent identically distributed random variables. Their distribution $\psi(t)$ can in principle be related to microscopic mass transfer processes such as spatially variable advection or retardation properties. In many applications, the transition time distribution has been modeled by parametric functions such as truncated power-law distributions whose parameters are then fitted, for example to observed breakthrough data (Berkowitz et al. 2006). The phenomenological basis of CTRW reflects that for heterogeneous media, the transport properties are imprinted in the medium structure. Thus, it is able to describe and predict heterogeneity-induced large scale transport phenomena such as breakthrough curve scaling, forward and backward tails in spatial solute distribution as well as non-linear evolution of solute dispersion.

As mentioned above, the CTRW approach propagates solute particles in space and time according to

$$x_{n+1} = x_n + l_c, \quad t_{n+1} = t_n + \tau_n, \quad (18)$$

where x_n and t_n are the particle position and time after n random walk steps. The particle position at a given time t is given by $x(t) = x_{n_t}$, where $n_t = \max(n | t_n \leq t)$ is the number of random walk steps before time t . It can be shown (Berkowitz et al., 2006) that the distribution c of particles that move according to Eq. (18) satisfy the non-local advection dispersion equation

$$\frac{\partial C}{\partial t} + \int_0^t dt' M(t-t') \left[v \frac{\partial C(t')}{\partial x} - D \frac{\partial^2 C(t')}{\partial x^2} \right] = 0. \quad (19)$$

where the velocity is $v = l_c/\tau_c$ and the dispersion coefficient is $D = l_c^2/2\tau_c$. If ψ has a finite mean, the characteristic time scale is equal to the mean transition time $\tau_c = \tau$. The memory function M can be written in Laplace space as

$$\bar{M} = \frac{\tau_c \lambda \bar{\psi}}{1 - \bar{\psi}} \quad (20)$$

For the exponential transition time distribution $\psi = \exp(-t/\tau_c)$, the memory function reduces to $M = \delta(t)$ and the non-local Eq. (19) reduces to the advection-dispersion equation. Solute transport based on the CTRW approach can be modeled numerically using random walk particle tracking methods based on Eq. (18) or using grid-based methods based on the non-local advection-dispersion equation (Eq. 19). Analytical results can be obtained for the long-time scaling of breakthrough curves and spatial moments.

For a transition time distribution that behaves asymptotically as $\psi \sim t^{-1-\gamma}$, the solute breakthrough curves in response to a pulse show the same scaling. The time evolution of the first and second centered moments of the concentration distribution for a pulse-like initial condition depends on γ -range (Shlesinger, 1974). For $0 < \gamma < 1$, the center of mass position scales as t^γ and the second centered moment as $t^{2\gamma}$. For $1 < \gamma < 2$, the scalings are linear for the center of mass and $t^{3-\gamma}$ for the second centered moments. For $\gamma > 2$, both center of mass position and second-centered moment evolve linearly at asymptotic times. Note that the transition time distribution ψ in general does not behave as a power-law. Its shape is determined by the heterogeneity distribution, as for example the distribution of flow speeds or retardation coefficients. If ψ scales as a power-law in a certain time regime, the corresponding behavior of the breakthrough curve and spatial moments can be estimated from the above scaling relations.

Fig. 11 shows complementary cumulative breakthrough data from tracer experiments in Berea sandstone. The CTRW fit reproduces the long tail, which is a manifestation of solution retention mechanisms due to small scale medium heterogeneities (Cortis and Berkowitz, 2004).

Fractional advection-dispersion models

Fractional advection-dispersion models are generalizations of the advection-dispersion equation based on fractional derivatives in space or time. They have been used for the modeling of non-Fickian transport features in porous and fractured media (Benson et al. 2013). Time-fractional approaches are linked to the MRMT and CTRW approaches for memory functions and transition time distributions that decay as pure power-laws. Space-fractional advection-dispersion models describe random walks that are characterized by broad distributions of transition lengths. A comprehensive treatise on time and space fractional transport models and their relation to stochastic processes can be found in the textbook by Meerschaert and Sikorskii (2011). In the following, we briefly summarize time- and space-fractional advection-dispersion models.

Time-fractional advection-dispersion equations

The time-fractional advection-dispersion equation (t-FADE) is linked to a CTRW that describes particle motion through the equations

$$x_{n+1} = x_n + v\Delta s + \sqrt{2D\Delta s}\xi_n, \quad t_{n+1} = t_n + \Delta s^{1/\gamma}\eta_n \quad (21)$$

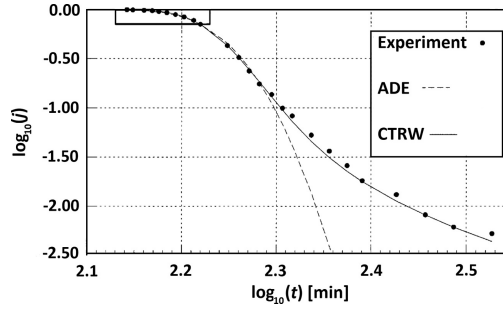


Fig. 11 CTRW versus ADE fits of experimental BTC data of a Berea sandstone core. Symbols denote experimental data from Scheidegger (1959), the solid line the CTRW fit, and the dashed line the ADE fit. Adapted from Fig. 2 in Cortis A and Berkowitz B (2004) Anomalous transport in “classical” soil and sand columns. *Soil Science Society of America Journal* 68(5): 1539–1548. doi: 10.2136/sssaj2004.1539.

where v and D are a velocity and a dispersion coefficient, Δs is termed an operational time increment and η_n is an identical independently distributed random variable whose distribution $\psi_\eta(t)$ is Lévy alpha-stable. Its Laplace transform is given by $\psi_\eta(\lambda) = \exp(-\alpha\lambda^\gamma)$ with $0 < \gamma < 1$. α is a scale parameter of units T^γ . It decays as the power law $t^{-1-\gamma}$. It can be shown that the concentration c satisfies the fractional advection-dispersion equation (Meerschaert and Sikorskii 2011).

$$\alpha \frac{\partial^\gamma C}{\partial t^\gamma} + v \frac{\partial C}{\partial x} - D \frac{\partial^2 C}{\partial x^2} = \alpha \frac{t^{-\gamma} C(0)}{\Gamma(1-\gamma)} \quad (22)$$

in the limit $\Delta s \rightarrow 0$. The Riemann-Liouville fractional derivative on the left side is a non-local operator that is defined by

$$\frac{\partial^\gamma C}{\partial t^\gamma} = \frac{\partial}{\partial t} \int_0^t dt' \frac{(t-t')^{-\gamma}}{\Gamma(1-\gamma)} C(t'). \quad (23)$$

The t-FADE describes power-law decay of breakthrough curves at $t^{-1-\gamma}$. Generalizations of this approach for exponents $1 < \gamma < 2$ can be found in Meerschaert et al. (2010).

Time-fractional mobile-immobile models (Benson et al. 2013) are related to the MRMT approach discussed above. For the memory function $\varphi(t) = \alpha t^{-\gamma}/\Gamma(1-\gamma)$, the governing equation (Eq. 4) is then by definition of the Riemann-Liouville fractional derivative given by

$$\frac{\partial C_m}{\partial t} + \beta \alpha \frac{\partial^\gamma C_m}{\partial t^\gamma} + v \frac{\partial C_m}{\partial x} - D \frac{\partial^2 C_m}{\partial x^2} = 0. \quad (24)$$

For $\gamma = 1/2$, this equation describes transport under fracture-matrix interaction with a semi-infinite matrix.

Space-fractional advection-dispersion equation

The phenomenology of space-fractional advection dispersion models is different from the time-fractional models discussed above, which account for the existence of broad distributions of mass transfer times in heterogeneous media as expressed by power-law distributions of transition times $\psi(t)$, and power-law tails of the memory function $\phi(t)$.

Space-fractional approaches, in contrast, model solute transport through long range spatial interaction. The space-fractional advection-dispersion model can be written as (Benson et al. 2000)

$$\frac{\partial C}{\partial t} + v \frac{\partial C}{\partial x} - D \frac{1+p}{2} \frac{\partial^\gamma C}{\partial x^\gamma} - D \frac{1-p}{2} \frac{\partial^\gamma C}{\partial (-x)^\gamma} = 0, \quad (25)$$

where $0 \leq p \leq 1$ describes the skewness of c , D is a generalized diffusion coefficient of unit L^γ , and the order of the fractional derivative is $0 < \gamma < 2$. This equation is equivalent to the following Langevin equation for the motion of solute particles (Meerschaert and Sikorskii 2011)

$$x_{n+1} = x_n + v\Delta t + \Delta t^{1/\gamma} \xi_n(\gamma, p), \quad (26)$$

where $\xi_n(\gamma, p)$ is distributed according to a Lévy alpha-stable distribution with skewness parameter p and γ the stability parameter. This means that the distribution of the space increment has a long tail and decays as $|x|^{-1-\gamma}$. The space-fractional approach models distributions with non-Gaussian heavy tails in space due to the broad distribution of spatial transition length. This, however, leads to divergent second centered moment. This means the actual width of the modeled distributions is not finite. The phenomenology of this approach is based on a medium structure that is characterized by a power-law distribution of length scales, and transport mechanisms that allow for traversing an arbitrary space increment within a constant time increment Δt .

Conclusion/summary/outlook

Solute transport, defined here as the movement of chemicals in soils and groundwater aquifers, plays a central role in global biogeochemical cycling. Population pressures increase loading of wastes and other chemicals to the subsurface and soil degradation can impact food security. Solute transport affects the efficacy and fate of agricultural chemicals, the treatment and disposal of wastewaters and wastewater treatment products, and the fate of many other potential contaminants. Solute transport is also important for maintenance of safe water supplies, nourishment of plants, and many other processes.

Advection, diffusion, and dispersion are the three modes of solute transport and, depending on flow conditions, transport can be dominated by one or a combination of these. Retardation is a process that slows the movement of a solute. It results from the interaction of solutes with soil surfaces by physical or chemical adsorption. In many soil applications, where the transport of a uniformly applied chemical on the land surface to a water table may be of primary interest, a one-dimensional treatment of solute transport may be adequate.

The Advection-Dispersion Equation and its two-region or two-site enhancements have traditionally been applied to solve solute transport problems. Fitting to the Advection-Dispersion Equation has revealed that the dispersivity increases with the scale of the transport process. More recent models anticipate different types of scaling depending on the structure of the medium heterogeneity. Despite decades of intensive research in soil science, groundwater hydrology, and chemical engineering, our *a priori* capability to predict solute transport in complex porous media such as soils remains limited.

References

- Abramowitz M and Stegun IA (1972) *Handbook of Mathematical Functions*. New York: Dover Publications.
- Benson DA, Wheatcraft SW, and Meerschaert MM (2000) Application of a fractional advection-dispersion equation. *Water Resources Research* 36: 1403–1421.
- Benson DA, Meerschaert MM, and Revielle J (2013) Fractional calculus in hydrologic modeling: A numerical perspective. *Advances in Water Resources* 51: 479–497. <https://doi.org/10.1016/j.advwatres.2012.04.005>.
- Berkowitz B, Cortis A, Dentz M, and Scher H (2006) Modeling non-fickian transport in geological formations as a continuous time random walk. *Reviews of Geophysics* 44: RG2003.
- Cortis A and Berkowitz B (2004) Anomalous transport in “classical” soil and sand columns. *Soil Science Society of America Journal* 68(5): 1539–1548. <https://doi.org/10.2136/sssaj2004.1539>.
- Dentz M and Berkowitz B (2003) Transport behavior of a passive solute in continuous time random walks and multirate mass transfer. *Water Resources Research* 39(5): 1111.
- Gouze P, Melean Z, Le Borgne T, Dentz M, and Carrera J (2008) Non-fickian dispersion in porous media explained by heterogeneous microscale matrix diffusion. *Water Resources Research* 44: W11416.
- Hadermann J and Heer W (1996) The Grimsel (switzerland) migration experiment: Integrating field experiments, laboratory investigations and modelling. *Journal of Contaminant Hydrology* 21(1–4): 87–100. [https://doi.org/10.1016/0169-7722\(95\)00035-6](https://doi.org/10.1016/0169-7722(95)00035-6).
- Haggerty R and Gorelick SM (1995) Multiple-rate mass transfer for modeling diffusion and surface reactions in media with pore-scale heterogeneity. *Water Resources Research* 31(10): 2383–2400.
- Haggerty R, McKenna SA, and Meigs LC (2000) On the late time behavior of tracer test breakthrough curves. *Water Resources Research* 36(12): 3467–3479.
- Harvey CF and Gorelick SM (1995) Temporal moment-generating equations: Modeling transport and mass transfer in heterogeneous aquifers. *Water Resources Research* 31(8): 1895–1911.
- Maloszewski P and Zuber A (1985) On the theory of tracer experiments in fissured rocks with a porous matrix. *Journal of Hydrology* 79: 333–358.
- Meerschaert MM and Sikorskii A (2011) *Stochastic Models for Fractional Calculus*. De Gruyter.
- Meerschaert MM, Zhang Y, and Baeumer B (2010) Particle tracking for fractional diffusion with two time scales. *Computers & Mathematics with Applications* 59(3): 1078–1086.
- Scheidegger AE (1959) An evaluation of the accuracy of the diffusivity equation for describing miscible displacement in porous media. In: *Proceeding of Theory of Fluid Flow in Porous Media Conference*, pp. 101–116. University of Oklahoma.
- Shlesinger M (1974) Asymptotic solutions of continuous-time random walks. *Journal of Statistical Physics* 10(5): 421–434.
- Villerman J (1987) Chemical engineering approach to dynamic modeling of linear chromatography, a simple method for representing complex phenomena from simple concepts. *Journal of Chromatography* 406: 11–26.

Further reading

- Adler PM (1992) *Porous Media: Geometry and Transports*. Boston, MA: Butterworth-Heinemann.
- Bear J (1972) *Dynamics of Flow in Porous Media*. New York: American Elsevier.
- Bear J and Bachmat Y (1990) *Introduction to Modelling of Transport Phenomena in Porous Media*. Dordrecht, the Netherlands: Kluwer Academic.
- Dullien FAL (1992) *Porous Media: Fluid Transport and Pore Structure*, 2nd edn New York: Academic Press.
- Elrick DE and Clothier BE (1990) *Solute Transport and Leaching*. Madison, WI: American Society of Agronomy 93–126. Agronomy Monograph 30.
- Fried JJ and Combarnous MA (1971) Dispersion in porous media. *Advances in Hydroscience* 7: 169–282.
- Hillel D (1998) *Environmental Soil Physics*. New York: Academic Press.
- Jury WA and Flüher H (1992) Transport of chemicals through soil: Mechanisms, models, and field applications. *Advances in Agronomy* 47: 141–201.
- Kreft A and Zuber A (1978) On the physical meaning of the dispersion equation and its solutions for different initial and boundary conditions. *Chemical Engineering Science* 33(11): 1471–1480. [https://doi.org/10.1016/0009-2509\(78\)85196-3](https://doi.org/10.1016/0009-2509(78)85196-3).
- Kutilek M and Nielsen DR (1994) *Soil Hydrology*. Reiskirchen, Germany: Catena-Verlag.
- Leij FJ and Dane JH (1989) *A Review of Physical and Chemical Processes Pertaining to Solute Transport*. Auburn, AL: Auburn University Press. Agronomy and Soils Department Series No. 132, Alabama Agricultural Experiment Station.
- Parker JC and van Genuchten MT (1984) *Determining Transport Parameters from Laboratory and Field Tracer Experiments*. Blacksburg, VA: Virginia Agricultural Experiment Station. Bulletin 84–3.
- Sahimi M (1995) *Flow and Transport in Porous Media and Fractured Rock*. Weinheim: VCH-Verlag.

- Selim HM and Sparks DL (eds.) (2001) *Physical and Chemical Processes of Water and Solute Transport/Retention in Soils*. Madison, WI: Soil Science Society of America. Special Publication 56.
- Toride N, Leij FJ, and van Genuchten MT (1995) *The CXTFIT Code for Estimating Transport Parameters From Laboratory or Field Tracer Experiments*. Riverside, CA: USDA-ARS. Research Report No. 137, US Salinity Laboratory.
- van Genuchten MT and Wierenga PJ (1986) Solute dispersion coefficients and retardation factors. In: Klute A (ed.) *Methods of Soil Analysis*, 2nd edn, pp. 1025–1054. Madison, WI: American Society of Agronomy. Part 1.

Lysimetry

Thomas Pütz^a and Jannis Groh^{a,b,c}, ^aForschungszentrum Jülich GmbH, Institute of Bio- and Geoscience IBG-3: Agrosphere, Jülich, Germany; ^bInstitute of Crop Science and Resource Conservation (INRES) – Soil Science and Soil Ecology, University of Bonn, Bonn, Germany; ^cLeibniz Centre for Agricultural Landscape Research (ZALF), Landscape Functioning, Müncheberg, Germany

© 2023 Elsevier Ltd. All rights reserved.

This is an update of T.A. Howell, LYSIMETRY, Editor(s): Daniel Hillel, Encyclopedia of Soils in the Environment, Elsevier, 2005, Pages 379–386, ISBN 9780123485304, <https://doi.org/10.1016/B0-12-348530-4/00391-X>.

Introduction	667
History	668
Lysimeter types, possibilities, and limitations	669
Parameters and variables measured in lysimeters	674
Water transport	674
Solute transport	675
Heat transport	675
Plant (above and below ground)	676
Supporting lysimeter measurements	676
Site conditions	677
Operation of lysimeter installations	677
Lysimeter as a “gold standard”	678
Acknowledgment	678
References	678

Abstract

Lysimeters are experimental systems for the investigation of water and solute fluxes in the complex soil-plant-atmosphere continuum. Innovations in weighing technology, in the control of the lower boundary condition as well as in data processing make lysimeters a representative ecosystem section that can map all translocation processes. However, the quality of the data is associated with a high technical effort and intensive, and regular maintenance. High precision weighable lysimeters represent the “gold standard” for determining water and solute transport, to calibrate and validate models and to improve process understanding in the critical zone.

Key points

- Lysimeters are an experimental system to control and monitor water and solute fluxes in soil systems.
- History of lysimeters, covering the period from the 1800s to the technological innovations made over the last 20 years.
- Lysimeter types, possibilities, and limitations
- Parameters and variables measured in lysimeters
 - Water transport
 - Solute transport
 - Plant (above and below ground)
 - Supporting lysimeter measurements
 - Site conditions
 - Operation of lysimeters
- Lysimeter can serve as a “Gold Standard”

Introduction

For many studies in the fields of soil and environmental sciences, such as hydrology, agriculture, ecology, meteorology, and the corresponding intersections of these disciplines, conventional field trials are unsuitable, as variables cannot be adequately controlled or measured experimentally. Particularly in the very complex soil-plant-atmosphere continuum, lysimeters represent

an integrative experimental approach that enables precise measurements of water and matter fluxes in combination with field-like conditions such as plant stands. The experimental conditions can be integrated into an ecological framework with realistic conditions to obtain data on in-situ-processes with limited bias.

Lysimeters are vessels of various dimensions filled with repacked or naturally layered soil preserving ecosystem properties. There is a holistic approach with lysimeters, because each compartment interacts dynamically within the biosphere (i.e., soil, water, flora, fauna, and atmosphere). For measuring the different components of the terrestrial water cycle, non-weighable or weighable lysimeters are used to measure soil water storage change, evaporation, actual evapotranspiration, drainage, upward directed water fluxes, precipitation, non-rainfall (e.g., dew) and irrigation water. Only with weighable high precision lysimeters can all the terms of the water balance equation be measured directly (e.g., Forstner et al., 2021; Groh et al., 2019). A special feature of these high precision lysimeters is the direct determination of the hydrological interactions between soil, plant, and atmosphere.

The matter fluxes can be precisely balanced using lysimeters because the volume of the soil or the ecosystem section is exactly defined (Fig. 1). The advantage of balancing is used for investigations on the fate of fertilizers, pesticides, or other anthropogenic substance inputs, which can be measured in the percolation water. Observations on water and matter fluxes in lysimeters provide very reliable data sets for calibration and validation of ecosystem models on different scales (Schelle et al., 2012; Groh et al., 2018b).

Lysimeters can be used in a laboratory, greenhouse, climate chamber or integrated in natural or managed ecosystems. They may also be integrated into Ecotron ecosystem analyzers to produce a fully balanceable and controllable experimental facility (Roy et al., 2021).

History

The term lysimeter is derived from the Greek word “lysis,” that means “solution, dissolved” and the Greek word “metron,” that means “to measure.” Lysimeter experiments have been conducted and published in literature for 300 years. In the 18th century, the first lysimeter studies were carried out in the Netherlands and France. These had more the character of large-scale pot experiments than of modern lysimeter approaches used today. The studies in those early days were primarily aimed at determining the soil water balance and the consumption by plants. A comprehensive overview of historical lysimeter experiments is provided by Kohnke et al. (1940) and Goss et al. (2010). In the last 20 years, several innovations have been introduced into lysimeter technology that have significantly advanced measurement accuracy, directly measurable target variables, realistic integration of lysimeters in ecosystems, lysimeter filling, and data processing. The website of the Lysimeter Research Group (LRG) provides a broad overview of European lysimeter stations representing the different stages of technical development, with various historical, and technical information (www.lysimeter.at/lysimeterplatform.html).

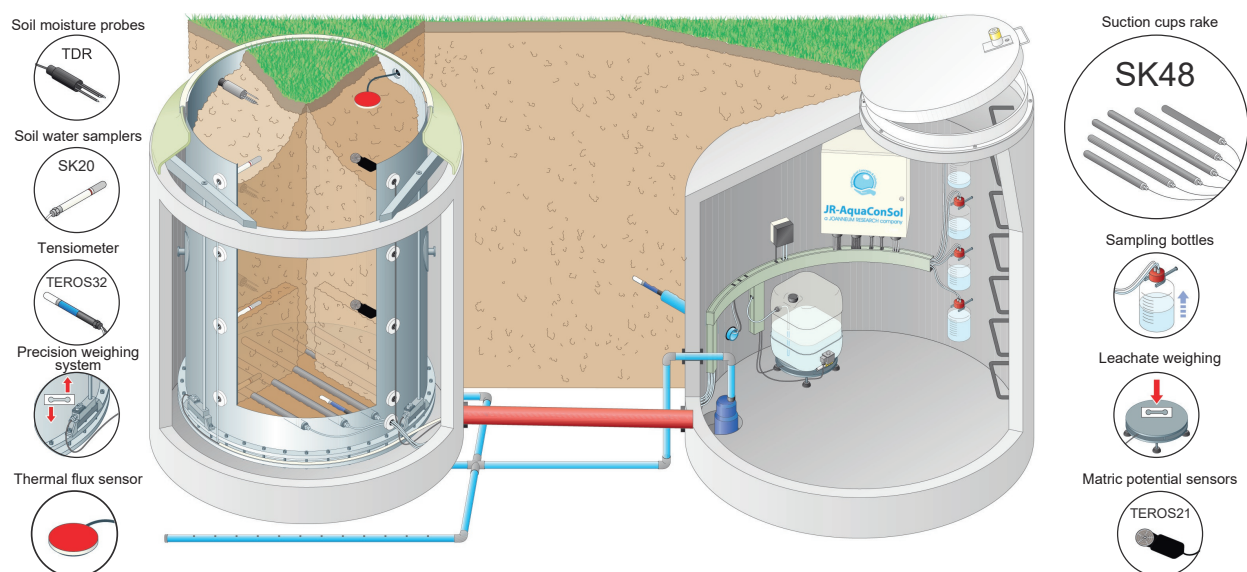


Fig. 1 Cross-sectional drawings of a weighable, circular lysimeter with controlled lower boundary condition and a separate service pit, © G. Klammler, JR-AquaConSol, Austria.

Lysimeter types, possibilities, and limitations

As a general definition, a lysimeter consists of a vessel filled with soil. The lysimeter vessel can be placed above ground or inserted into the ground (Fig. 1). If larger weighable lysimeters are used, the lysimeters at field sites must be placed in a container to allow the installation of load cells. The different lysimeter types (Table 1) can be classified according to the following criteria:

- Size
- Filling procedure
- Design
- Bottom boundary condition of the lysimeter

The size of a lysimeter has considerable significance for its functionality and, depending on its size, may capture soil and plant heterogeneity of the experimental site. The dimensions of lysimeters should therefore be based on the processes to be investigated and the spatial situation, so that processes as well as their spatial variability or heterogeneity can be represented by the selected size. Therefore, under conditions with large soil and plant heterogeneity, it is recommended that the size of the soil core should be greater to encompass the heterogeneity. However, the size of monolithic filled lysimeters is limited by the technical feasibility. For example, large-scale lysimeters with soil volumes of >12 cubic meters require enormous technical effort to establish (Fig. 2). The size and shape of a lysimeter (soil surface, depth, round or square) should be adapted to the type of vegetation. To avoid dominance of flowerpot effects of preferential root growth along the lysimeter walls, there must be a minimum depth or volume of soil that allows natural root development by the vegetation.

The surface areas and depths of lysimeters used worldwide vary enormously in their dimensions and there are no uniform standards. Commercially available lysimeters vary in size from very small, weighable lysimeters (so-called smart field lysimeters) with a surface area of 0.07 m², to larger (0.25 m²) and deeper 0.30–0.90 m (down to 3.0 m depth) lysimeters (Riedl et al., 2022). Purpose-built lysimeters with a surface area of 1–3 m² and depths between 1.0 and 2.5 m, which are also weighable and filled with soil monoliths, are widely used (Marek et al., 2006; Pütz et al., 2016; Allen et al., 1991). With disturbed filled soil, lysimeters can be produced in any size, with dimensions that allow the investigation of entire forest ecosystems or artificial hillslopes. The Landscape Evolution Observatory at biosphere 2 hosts three artificial weighable lysimeters with dimensions of 330 m² and 1 m depth to study hillslope processes under controlled boundary conditions (Arevalo et al., 2020). Other non-weighable lysimeters with surfaces up to 660 m² and depths up to 5.0 m have been built and operated for decades (e.g. Britz, Eberswalde, Germany, St. Arnold; Colbitz-Letzlinger Heide, Germany; Glugla et al., 1982; Wyrwich and Brandenburg, 2018).

The filling procedure is the transfer of a representative soil section of an ecosystem into a lysimeter. A distinction must be made here between filling with disturbed or monolithic soil cores. In disturbed soil cores, soil is repacked into the lysimeter to form layers mimicking the natural horizons found in the field. The disadvantages of this method are the high effort, the disturbed soil structure, the artificial bulk density of the soil, and the stimulation of mineralization processes through the aeration or loosening of the soil. For monolithic lysimeter filling, a distinction is made between the preparative procedure, the cutting method, and the milling method (Fig. 2). In the preparative method, a soil monolith is carved out of the undisturbed soil according to the dimensions of the lysimeter vessel and then transferred into the lysimeter vessel. This method can be used for almost all soil type, especially very stony ones.

Table 1 Characteristics of different lysimeter types.

Type	Description and characteristics
Volumetric/drainage lysimeter	simplest, most common, lysimeter vessel is set up to collect and/or to measure the seepage water volumetrically, non-weighable, water fluxes measured indirectly by seepage water volume at the bottom, precipitation and/or irrigation by separate measuring instruments, various designs of the vessel bottom are used;
Gravimetric lysimeter	weighable lysimeter vessel in combination with a weighable seepage water tank, different weighing techniques available (mechanical, electronic, . . .) and different bottom boundary conditions, very expensive due to complex installation and construction; short measure intervals provide a direct measurement of the water balance components (i.e., precipitation, non-rainfall water, actual evapotranspiration and drainage)
Groundwater lysimeter	weighable/non-weighable lysimeters; characteristic is a device that adjusts the groundwater level in the lysimeter tank constantly or variably; external water addition is necessary during operation; measurement of groundwater recharge possible; use of monolithically filled lysimeters possible;
Large-scale lysimeter	dimensions approx. 15 m × 15 m × 3 m (length x width x depth) and larger, large lysimeter basins, usually made of concrete, filled with disturbed soil, non-weighable; undisturbed variants also possible, if impermeable soil layers are present, only lateral boundaries have to be drawn in; because of the size, special devices for measuring the seepage water are necessary; valid measurements after an adaptation period of 1–2 hydrological years; very expensive; investigation of large vegetation;
Smart field lysimeter	miniaturized form of a fully functional lysimeter, weighable, controlled bottom boundary condition, seepage water and upward direct water measurement, sensor installations;
Special lysimeter	to measure run-off and/or lateral flow



Fig. 2 Filling of 2 m², 2.5 m depth lysimeter with soil monoliths – cutting method: lysimeter with the lysimeter press on top (1st row left), pressing of the lysimeter completed (1st row center), two filled lysimeters and pushing the bottom plate under a filled lysimeter (1st row right), lift out the filled lysimeter (2nd row left), Insertion of a 2 m², 2.5 m long lysimeter into a lysimeter basement (2nd row center), and view into the lysimeter basement with lysimeters protruding through the basement floor, platform scales, and containers for drainage sampling (2nd row right).

In the cutting method, the lysimeter container is equipped with a cutting edge and then pressed into the soil with a centrally acting hydraulic cylinder (Figs. 3 and 4). Once the monoliths have been transferred to the lysimeters, a base plate is pushed underneath the lysimeter. In the milling method, a milling tool is placed on the bottom side according to the inner diameter of the lysimeter container and this tool, which works in a rotating manner, is then pressed into the soil together with the lysimeter container (Reth et al., 2021) (Fig. 5). The advantages of the cutting and milling methods are undisturbed soil structure and natural stratification. The disadvantages of monolith lysimeter filling are the lack of information on the integrity of the soil monolith and the fact that soils with high stone content cannot be filled in via cutting or milling. However, only monolith-filled lysimeters include the inherent heterogeneity of the soil that was present at the start of monitoring. But even with monolithically filled lysimeters, there are challenges to their longevity and accuracy at representing the surrounding environment due to the new boundary conditions they impose. Generally clayey soils cannot be investigated in long-term studies with lysimeters, as their shrinkage and swelling, depending on the soil water content, may prevent sufficient connectivity between the soil and the lysimeter wall. This can create artificial preferential flow paths along the lysimeter wall (wall effect).

The lysimeter design can be easily differentiated between cube-shaped, cuboid or cylindrical lysimeters. Due to the stability of the cylindrical shape, this is very well suited for monolithic filling. With the angular variants, deformation of the lysimeter vessel may occur or, depending on the soil type, soil compaction in the corners is possible.



Fig. 3 Filling of lysimeter - cutting method: careful pressing of a lysimeter vessel into the ground with four hydraulic cylinders and an excavator as abutment (left), preparation of the next driving section with inspection of the next section (center), sheared lysimeter was lifted out of the excavation pit for further processing (right) © G. Klammler, JR-AquaConsol, Austria.



Fig. 4 Reasonably priced lysimeter filling – cutting method: pressing and driving the lysimeter vessel into the ground with an excavator (upper row), shearing off the filled lysimeter with a bottom plate integrated in a transport frame (bottom row).



Fig. 5 Equipment for the vertical filling of soil monoliths with the milling method – starting the filling procedure, rotary cutting tool at the bottom of the lysimeter vessel (left), vertical filling procedure completed (mid), and bottom plate attached below the lysimeter and raising up the lysimeter (right) © S. Reth, UGT GmbH, Germany.



Fig. 6 TERENO-SOILCan lysimeter station Rollesbroich, Germany with six weighable, monolithic lysimeters in a hexagonal arrangement with a pluviometer (in the foreground) and a radiation sensor above a lysimeter (right side). The lysimeters can be recognized by the light-colored sealing lips that close the space between them and the lysimeter container against water penetration. In the center of the hexagon is the cover of the service pit.

Another constructive feature is the arrangement of lysimeters. Here, a basic distinction is made between the placement of a lysimeter in a slightly larger container (Fig. 1) or the installation in a basement-like building (Fig. 2). The lysimeter container completely accommodates the lysimeter vessel and the lysimeter surface is on an equal level with the surrounding field surface. The container requires a slightly larger volume to allow a temperature transfer from the surrounding soil into the lysimeter to be as interference-free as possible. This needs to be balanced against the space required to install the necessary equipment (load cells, brackets, cables, lifting device, etc.) to monitor environmental processes. The resulting gap is then closed by a suitable cover to exclude the intrusion of rain, soil particles or animals. However, this cover must not interfere with the measurements, such as the weighing. When installed in basement-like buildings, the structures have openings in the ceiling through which the lysimeters pass to the surface (Fig. 2). Often, the lysimeters are placed in the room on pedestals or similar, on which the weighing equipment is installed, and seepage water is recorded. The advantage of a lysimeter basement is the facilitated access to the lysimeters, so that maintenance as well as measurement operation are very easy. Disadvantages are the changed temperature profile in the lysimeters caused by the ambient air in the basement deviating from soil temperature and the high construction costs.

In modern lysimeters with a suction controlled bottom boundary, seepage water and rising soil water can be measured in a defined soil volume. To compensate for the hydraulic disconnection of the lysimeters, soil cores from the underlying soil layers, porous plates, or suction candles and tensiometers are installed at the bottom of the lysimeter and tensiometers at a similar soil depth in the surrounding field (Fig. 1 top, Figs. 6 and 7, Fig. 8, left). An adjustable control algorithm minimizes the matric potential difference between the lysimeter and field using a bidirectional pumping system to transport water upwards (capillary rise) or downwards (drainage) via porous plates or suction candles at the lysimeter bottom and seepage water tank (Fig. 9, Fig. 8 left). This technique ensures that the water dynamics of the lysimeter correspond to the observed field dynamics (Groh et al., 2016). As a variant, there are also groundwater lysimeters in which a hanging water column is used to set the groundwater depth at the selected level.



Fig. 7 Lysimeter integrated in a field with the same vegetation (left); Lysimeter exposed shortly before harvest with visible sealing lip (right).



Fig. 8 A TERENO-SOILCan lysimeter turned upside down during installation of the suction rake to control the lower boundary conditions (left) and raised above ground for maintenance with various sensors at different installation depths (right).

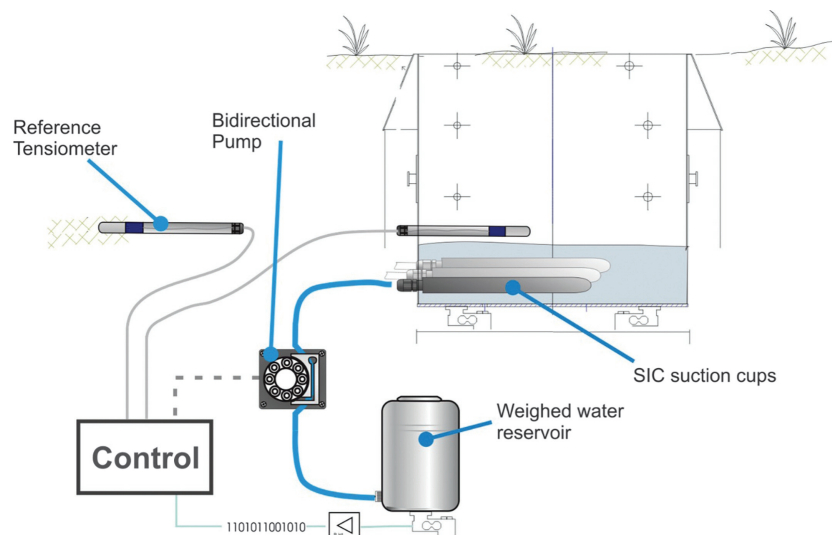


Fig. 9 Controlled lower boundary by suction rake, reference tensiometer, and bi-directional pump. © G. Klammler, JR-AquaConsol, Austria.

Table 2 Advantages and disadvantages of the different lysimeter properties.

<i>Properties</i>	<i>Advantage</i>	<i>Disadvantage</i>
General	water balance, matter fluxes measurable, actual evapotranspiration measurable; use of tracers long-term experiments experimental repetitions (by number of lysimeters); representative ecosystem sections; weighable;	no horizontal flow; expensive; not applicable to all soil types; limited size;
Free-drainage	low construction cost; no additional technical components necessary;	precipitation and irrigation are measured by independent rain gauges; no active controlled bottom boundary, no upward directed water flux laborious maintenance
Suction-controlled	nearly undisturbed field situation; seepage water measurable; capillary rise measurable; complete solute transport (up- and downward direct)	
Monolith	natural soil structure, natural porosity,	complex, expensive filling technology
Weighable	complete water balance measurable; high temporal resolution of the measurement;	expensive

Other lysimeter bottom boundary conditions have an artificial gravel/sand layer, a perforated plate, or an outlet at the lowest point of the bottom, where soils are exposed to the atmosphere (Fig. 1 – center and bottom). Under these bottom boundary conditions, seepage water only occurs when the soil above the bottom plate of a lysimeter is saturated (i.e., matric potential ≥ 0 hPa) and thus deviates in the most cases from the water dynamics in the field. In addition, temporary anaerobic conditions at the lysimeter bottom affect plants and the fate of solutes, which might not reflect conditions in the surrounding field and thus lead to biased observations of water and solute fluxes with lysimeters. Water fluxes change under these artificial boundary conditions because either water is temporarily more accessible to plants due to the nearly saturated soil zone or upward directed water flow (capillary rise) is neglected, both of which affect evapotranspiration processes (Groh et al., 2016).

The different characteristics of lysimeter types and the advantages and disadvantages of the several lysimeter properties are compiled in Tables 1 and 2.

Parameters and variables measured in lysimeters

Water transport

The essential advantage of a lysimeter compared to other experimental settings to investigate water and matter fluxes is the exact knowledge of its volume. At constant soil mass, the change in weight of a lysimeter can only occur by the water phase and to a very small extent also through the production of biomass. The respective components of the water balance are measured across the surface and bottom section of the lysimeter.

In the past, weighable lysimeters were used to determine the change in soil water storage (ΔW). Here measurement frequency was relatively coarse (>30 min) and thus actual evapotranspiration (ET_a) was estimated indirectly by solving the water balance equation:

$$\Delta W = P + I + NRW - SR - D - ET_a + CR$$

where all units are given in millimeters ($\text{kg m}^{-2} = \text{L m}^{-2} = \text{mm}$) at a given time and ΔW is the change in the lysimeter weight over a consecutive time step, P is precipitation, I is irrigation, NRW is non-rainfall water (i.e., dew, hoar frost, fog, water vapor adsorption), SR is surface runoff, D is seepage water, and CR is upward direct water flow from capillary rise (Fig. 9). SR and D will be sampled with a separate container which is either measured by a balance or mounted on the lysimeter itself. Whereby for the latter case, water balance can only be solved on the corresponding drainage sampling time step.

The use of external rain gauge measurements is a clear disadvantage as it neglects NRW and introduces, depending on the rain gauge design and functionally, a systematic measuring error for P , caused by wind and evaporation (Fuchs et al., 2001), when estimating ET_a . Due to high precision ($\leq 10 \text{ g m}^{-2}$) and high temporal measurements (≤ 1 min) with modern lysimeters, water balance components such as ET_a , P , and NRW can be measured directly, when assuming no simultaneous ET_a and P or NRW during a specific short time interval (e.g., 1 min.) with the lysimeter balance (Fig. 10). Collection of high-frequency measurements (< 1 min) is negatively affected by noise induced by wind when determining land surface water fluxes and require a noise reduction filter to be applied (Peters et al., 2017).

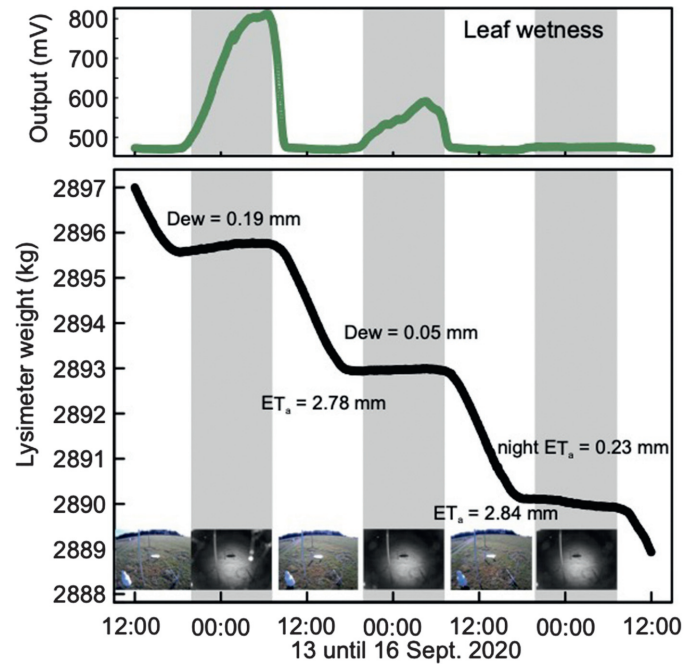


Fig. 10 Evaluation of the 72-h weight course of a lysimeter with a controlled lower boundary condition and separate recording of drainage: leaf wetness measured by a sensor (above); periods with dew, actual evapotranspiration (ET_a), and actual night time evapotranspiration (night ET_a) (bottom), measured at the TERENO-SOILCan lysimeter station Rollesbroich, Germany (Fig. 6).

In many regions, *NRW* inputs contribute additional water to precipitation to the ecosystems (Dawson and Goldsmith, 2018). *NRW* consists of dew (Kaseke et al., 2012), hoar frost (Groh et al., 2018a), distillation (Agam and Berliner, 2006), fog (Klemm et al., 2018), rime (Meissner et al., 2007), and water vapor adsorption (Saaltink et al., 2020). The different *NRW* fractions are typically small at the daily scale compared to precipitation amounts, but they provide a constant source of water and thus help in maintaining ecosystem functions. Despite the importance of *NRW* inputs in soils for terrestrial ecosystem, they are the least studied and characterized components and processes of the terrestrial water cycle. Weighable high precision lysimeters can be used to quantify all *NRW* inputs and their various fractions using other measuring devices (e.g., visibility and leaf wetness sensors; for more details see (Zhang et al., 2019)).

Thus modern weighable lysimeter systems with high precision and high temporal resolution of the weighing data allow direct measurements of ET_a , P , and *NRW* and equipped with a dynamic bottom boundary control provides an holistic determination of the hydrological interactions between soil and atmosphere (Fig. 1 - top; Fig. 6) (Pütz et al., 2016).

Solute transport

The fate and the translocation of a solute in a certain period can be balanced in a lysimeter. The solute can either be applied to the soil or based on a known quantity of the solute in the soil, factoring in biomass uptake and the natural input via the air pathway or volatilization. Solute flux can be measured with lysimeters only directly at the surface (i.e., amount of water) and at the bottom (i.e., water and solute). The measurement of the solute fluxes in the different soil layers of a lysimeter is only possible with additional, considerable effort since D and CR can only be determined indirectly compared to the solute concentration there. Lysimeters are often equipped with sensors at different soil depth to measure water content, matric potential, temperature, and chemical parameters (Fig. 9 right). The leached and upward directed solute flux (S_{flux} , $mg\ m^{-2}$) of modern lysimeter with dynamic controlled bottom boundary can be measured (Pütz et al., 2018):

$$S_{flux} = \begin{cases} C * D, & D > 0 \\ C * CR, & CR > 0 \end{cases}$$

with C the measured solute concentration in D or CR ($mg\ L^{-1}$).

Heat transport

To measure the water and material balances of the reference ecosystem in lysimeters as realistically as possible, no artificial temperature conditions may prevail. Lysimeters that are embedded in the soil should have a close horizontal and vertical connection to the surrounding, naturally layered soil. Especially in the case of weighable lysimeters, the air space between the

weighable lysimeter and the container surrounding the lysimeter should be kept to a minimum (Fig. 1). This facilitates the horizontal temperature transition between the surrounding soil and the soil in the lysimeter. The lysimeter materials that are exposed to the atmosphere (radiation) and are in direct contact with the lysimeter soil lead to a distorting effect on the soil temperature profile in the lysimeter. All lysimeter parts protruding above the soil surface must be as small as possible with a high albedo to minimize this effect. The main temperature changes occur via the surface of the soil in a lysimeter and cause natural temperature gradients within it, which then continue to depth. Horizontal disturbances of the heat flow caused by constructional deficiencies of the lysimeter system can lead to cone-shaped temperature profiles in the lysimeter soil, which means that the temperature profile in the center of the lysimeter corresponds to the natural profile. To a certain depth, the edge effects overlap the natural temperature profile and cause artificial temperature profiles. These effects can be particularly pronounced in lysimeters that are operated in basement lysimeter stations (Fig. 2, Fig. 11). Here, the lysimeter must be insulated to decouple it from the cellar climate. Ideally, the cellar temperature should be adapted to the ground temperature and, if necessary, additional temperature control must be provided via the lysimeter base (Podlasly and Schwärzel, 2013).

Plant (above and below ground)

The vegetation plays a pivotal role in the terrestrial water balance. However, transpiration measurements with weighable lysimeters are not directly possible, so further ecophysiological measurements are needed for a better understanding and modeling of the processes in the soil-water-plant continuum. Ecophysiological aboveground measurements include the canopy height, plant development stages, leaf area index, sap-flow, stomatal conductivity, photosynthesis, and soil respiration. In addition to recording the fresh mass or dry mass of the plants from a vegetated lysimeter, grain yield, single plant parts or their various components can also be balanced. Observational systems to measure root growth indirectly, such as by rhizotrons, can be included in the lysimeter design to include monitoring of the belowground activities of plants.

Supporting lysimeter measurements

The time scale of a lysimeter experiment (short-term or long-term) has an essential influence on the experimental setup. Here, especially in the initial phase of an experiment, so-called “priming effects” can occur. These are processes that have been unnaturally intensified by the lysimeter setup and thus lead to artificial results. In contrast to experimental approaches whose boundary conditions are artificially controlled (laboratory experiments, Ecotrons, etc.), the lysimeter in a field experiment is influenced exclusively by the natural boundary conditions. Due to the complex installation, lysimeter experiments tend to be designed as long-term experiments. Exceptions here are smart field lysimeters, which can be used more flexibly in the field due to their small size, fast soil excavation and operated by power from a solar panel (Eberbach et al., 2013). In studies on the fate of anthropogenic or natural substances under consideration of the water balance, at least two hydrological years or two seepage periods are used in humid climatic zones based on empirical values. In this period, a seepage water volume corresponding to two to three times the soil pore volume should be obtained to be able to describe the transport of these materials.



Fig. 11 Lysimeter facility with basement and control plot (in the foreground) and lysimeter in containers embedded in the ground surrounded by a fence and paved paths (in the background). Negative impacts of the surrounding environment cannot be excluded here.

The objectives of lysimeter experiments can be either specific questions or monitoring approaches that record processes or changes in ecosystem sections within the framework of long-term measurement series.

Site conditions

Lysimeters should be integrated into an ecosystem in such a way that no lateral influences, such as from trees, hedges, bushes, buildings, or fences, etc. in the immediate vicinity of the lysimeters can affect the measurements (Fig. 6, Fig. 7 – optimal, Fig. 11 – non-optimal). A lysimeter that is perfectly integrated into the ecosystem is indistinguishable from the surroundings and should also have a plant population or vegetation that corresponds to that of the surrounding environment. This means that vegetation on the lysimeter and the surrounding area is managed in a similar way (e.g., fertilizer, pesticides, growth regulator, plant density and rows etc.) to prevent an “oasis effect” (Allen et al., 1991). In the early work with lysimeters, it was already recognized that lysimeter construction requires the soil moisture of the reference soil outside the lysimeter to correspond to that inside the lysimeter (Fig. 8). The lateral separation of the lysimeter from its surrounding disturbs lateral in- and outflows. However, this limitation can also be used to determine such processes when comparing measurements within the lysimeter and the field, e.g., for lateral subsurface flow (Ehrhardt et al., 2021) or soil physical properties (Bravo et al., 2020).

A lysimeter interacts directly with its surroundings; changes in solar radiation, shading, the local wind field, and the surrounding crop have a direct effect on the measurement results of a lysimeter (Hagenau et al., 2015). The type of vegetation on a lysimeter, whether representing a natural or an agricultural ecosystem, or bare soil, will determine further management. In the case of agricultural land, soil cultivation, sowing, plant variety, fertilizer application, plant protection activities and harvesting of the surrounding area must be same as for the lysimeters.

Additionally, high-quality, sensitive precipitation gauges and weather stations, which record air temperature, wind speed, solar radiation, and humidity, complete a lysimeter station and enable a full data set to analyze or simulate ecosystems (Fig. 6).

Operation of lysimeter installations

Associated with a lysimeter experiment, a comprehensive number of soil physical, hydrological parameters must be recorded, preferably at high temporal resolution. A horizontal installation of probes should be established in the soil to avoid artificial preferential flow paths. To achieve a better understanding of water fluxes within the soil profile of a lysimeter, measurement of the matric potential and water content is recommended to determine the soil water retention characteristic. However, consideration must be given to the possibility that installed probes can influence processes of the soil water flux (e.g., disturbance of the pore space, interruption of flow paths). The measurements can be compared to water retention curves obtained from soil cores sampled close to the lysimeter.

Suction cups have been used in experiments on the fate of organic or inorganic substances in lysimeters, such as plant nutrients, pesticides, or other anthropogenic chemical inputs (Fig. 12). The management of the soil water sampling strategy (applied pressure head, sampling frequency) might have an effect on the transport water and solutes (Weihermüller et al., 2007).



Fig. 12 Additional measurements and work in lysimeter experiments: Measurement of plant height (left), sampling of soil solution (center), and seepage water (right).

Of course, any maintenance activities or ancillary measurements carried out while the experiment is in progress should have minimal influence on the results of the study. This is a very special challenge, if not impossible, with high precision weighable lysimeters, because the very sensitive balances with an accuracy of approx. 10 g or less can be affected by wind, animals, and maintenance. To minimize disturbances of the measurements, recording the weighing results with the highest possible frequency is useful, with data then averaged and stored in manageable time units, such as every minute. All other measurement data should be recorded at least in 10-min intervals to document the dynamics of the various processes, from which, it may be deemed advisable to shorten the interval to detect the presence of fast processes and interactions.

Technical maintenance and control of a lysimeter is of central importance to the functionality and the quality of the results from a lysimeter experiment. In addition to the daily check of the actual measurement data, now more permissible via online access, a weekly on-site check can identify any external influences or faults. Furthermore, continuously recorded images of a camera on site can be very helpful for the data quality management and to document the seasonal vegetation development.

Only through this intensive control and maintenance work can complete and valid data sets be obtained, thereby allowing reliable and versatile analyses. In addition to the personnel costs for at least one technician, those for repairs or replacement of probes and measuring instruments must also be identified. As lysimeter experiments are often designed as long-term experiments, probe and sensor failures are to be expected with increasing experiment duration.

Lysimeter as a “gold standard”

Since lysimeters can be used to measure both the water balance and the mass balance of a selected variable, lysimeters have been considered for many years as the “gold standard” in various scientific fields. The prerequisite for all comparisons between the real field situation and the lysimeter is to ensure optimal experimental conditions of the lysimeter.

Weighable precision lysimeters are used to determine the actual evapotranspiration, although a variety of structural and experimental parameters can falsify the measurement results (Allen et al., 1991, Groh et al., 2019). When the experiences and technical innovations of Lysimetry described in this article are consistently implemented, lysimeters record the actual evapotranspiration highly precisely and can be used to evaluate the Eddy Covariance method (Gebler et al., 2015). Precision lysimeters are also used as a reference for the measurement of precipitation and non-rainfall water (Groh et al., 2018a; Haselow et al., 2019; Herbrich and Gerke, 2016).

Lysimeters were also used as a “gold standard” to assess the environmental fate of pesticides or nutrients. In addition to the fate of the test compounds (degradation, sorption, etc.), their soil transport has also been explored (Führ et al., 2003). Based on 34 lysimeter experiments with 28 pesticide active ingredients and 35 metabolites, simulation models could be evaluated (Hardy et al., 2008). The reduction of nitrate contamination in water catchment areas in combination with the use of models and appropriately coordinated agricultural management advice could also be achieved with the inclusion of lysimeter measurements (Klammler and Fank, 2014).

Groh et al. (2022) used the “Data set Classification software” developed by Kersebaum et al. (2015) to quantitatively analyze and classify a typical data set of high-precision weighable lysimeters for the calibration and validation of agro-ecological models. Their results classify the dataset with a score of >200 points as highly suitable for parameterizing and improving agro-ecosystems with models.

Observations from weighable, high precision lysimeters thus provide a holistic data set on the ecosystem and can truly be considered a gold standard that can be used either to calibrate and validate models or to improve process understanding in the critical zone.

Acknowledgment

The authors would like to thank Gernot Klammler (JR-AquaConSol GmbH, 8010 GRAZ Austria) for the figures and photos (Figs. 1, 8, and 3) and Sascha Reth (Umwelt-Geräte-Technik GmbH, 15374 Müncheberg, Germany) for the photos in Fig. 5. Jannis Groh was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation—project no. 460817082).

References

- Agam N and Berliner PR (2006) Dew formation and water vapor adsorption in semi-arid environments - A review. *Journal of Arid Environments* 65: 572–590.
- Allen RG, Howell TA, Pruitt WO, Walter IA, and Jensen ME (1991) Lysimeters for evapotranspiration and environmental measurements. In: *Proceedings of the International Symposium on Lysimetry*, Honolulu, Hawaii, July 23–25, 1991, p. 456. New York, NY: American Society of Civil Engineers.
- Arevalo J, Zeng X, Durcik M, Sibayan M, Pangle L, Abramson N, Bugaj A, Ng W-R, Kim M, Barron-Gafford G, Van Haren J, Niu G-Y, Adams J, Ruiz J, and Troch PA (2020) Highly sampled measurements in a controlled atmosphere at the Biosphere 2 Landscape Evolution Observatory. *Scientific Data* 7: 306.
- Bravo S, González-Chang M, Dec D, Valle S, Wendroth O, Zúñiga F, and Dörner J (2020) Using wavelet analyses to identify temporal coherence in soil physical properties in a volcanic ash-derived soil. *Agricultural and Forest Meteorology* 285–286: 107909.
- Dawson TE and Goldsmith GR (2018) The value of wet leaves. *New Phytologist* 219: 1156–1169.

- Eberbach PL, Hoffmann J, Moroni SJ, Wade LJ, and Weston LA (2013) Rhizo-lysimetry: facilities for the simultaneous study of root behaviour and resource use by agricultural crop and pasture systems. *Plant Methods* 9: 3.
- Ehrhardt A, Groh J, and Gerke HH (2021) Wavelet analysis of soil water state variables for identification of lateral subsurface flow: Lysimeter vs. field data. *Vadose Zone Journal* 20: e20129.
- Forstner V, Groh J, Vremec M, Herndl M, Vereecken H, Gerke HH, Birk S, and Pütz T (2021) Response of water fluxes and biomass production to climate change in permanent grassland soil ecosystems. *Hydrology and Earth System Sciences* 25: 6087–6106.
- Fuchs T, Rapp J, Rubel F, and Rudolf B (2001) Correction of synoptic precipitation observations due to systematic measuring errors with special regard to precipitation phases. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere* 26: 689–693.
- Führ F, Burauel P, Mittelstaedt W, Pütz T, and Wanner U (2003) The lysimeter concept: A comprehensive approach to study the environmental behaviour of pesticides in agroecosystems. In: *Environmental Fate and Effects of Pesticides*. American Chemical Society.
- Gebler S, Hendricks Franssen HJ, Pütz T, Post H, Schmidt M, and Vereecken H (2015) Actual evapotranspiration and precipitation measured by lysimeters: A comparison with eddy covariance and tipping bucket. *Hydrology and Earth System Sciences* 19: 2145–2161.
- Glugla G, Fischer D, Höhne U, Kortüm F, and Helbig A (1982) Lysimeteruntersuchungen in der Letzlinger Heide - Wichtiger Beitrag zur Bestimmung der Wasserressourcen bewaldeter Gebiete. *Wasserwirtschaft Wassertechnik* 9: 319–322.
- Goss MJ, Ehlers W, and Unc A (2010) The role of lysimeters in the development of our understanding of processes in the vadose zone relevant to contamination of groundwater aquifers. *Physics and Chemistry of the Earth, Parts A/B/C* 35: 913–926.
- Groh J, Vanderborght J, Pütz T, and Vereecken H (2016) How to control the lysimeter bottom boundary to investigate the effect of climate change on soil processes? *Vadose Zone Journal* 15.
- Groh J, Slawitsch V, Herndl M, Graf A, Vereecken H, and Pütz T (2018a) Determining dew and hoar frost formation for a low mountain range and alpine grassland site by weighable lysimeter. *Journal of Hydrology* 563: 372–381.
- Groh J, Stump C, Lücke A, Pütz T, Vanderborght J, and Vereecken H (2018b) Inverse estimation of soil hydraulic and transport parameters of layered soils from water stable isotope and lysimeter data. *Vadose Zone Journal* 17.
- Groh J, Pütz T, Gerke HH, Vanderborght J, and Vereecken H (2019) Quantification and prediction of nighttime evapotranspiration for two distinct grassland ecosystems. *Water Resources Research* 55: 2961–2975.
- Groh J, Diamantopoulos E, Duan X, Ewert F, Heinlein F, Herbst M, Holbak M, Kamali B, Kersebaum KC, Kuhnert M, Nendel C, Priesack E, Steidl J, Sommer M, Pütz T, Vanderborght J, Vereecken H, Wallor E, Weber TKD, Wegehenkel M, Weihermüller L, and Gerke HH (2022) Same soil, different climate: Crop model intercomparison on translocated lysimeters. *Vadose Zone Journal*: e20202.
- Hagenau J, Meissner R, and Borg H (2015) Effect of exposure on the water balance of two identical lysimeters. *Journal of Hydrology* 520: 69–74.
- Hardy I, Gottesbüren B, Huber A, Jene B, Reinken G, and Ressler H (2008) Comparison of lysimeter results and leaching model calculations for regulatory risk assessment. *Journal für Verbraucherschutz und Lebensmittelsicherheit* 3: 364–375.
- Haselow L, Meissner R, Rupp H, and Miegel K (2019) Evaluation of precipitation measurements methods under field conditions during a summer season: A comparison of the standard rain gauge with a weighable lysimeter and a piezoelectric precipitation sensor. *Journal of Hydrology* 575: 537–543.
- Herbrich M and Gerke HH (2016) Autocorrelation analysis of high resolution weighing lysimeter time series as a basis for determination of precipitation. *Journal of Plant Nutrition and Soil Science* 179: 784–798.
- Kaseke KF, Mills AJ, Brown R, Esler KJ, Henschel JR, and Seely MK (2012) A method for direct assessment of the “non rainfall” atmospheric water cycle: Input and evaporation from the soil. *Pure and Applied Geophysics* 169: 847–857.
- Kersebaum KC, Boote KJ, Jorgenson JS, Nendel C, Bindi M, Frühauf C, Gaiser T, Hoogenboom G, Kollas C, Olesen JE, Rötter RP, Ruget F, Thorburn PJ, Trnka M, and Wegehenkel M (2015) Analysis and classification of data sets for calibration and validation of agro-ecosystem models. *Environmental Modelling & Software* 72: 402–417. <https://doi.org/10.1016/j.envsoft.2015.05.009>.
- Klammler G and Fank J (2014) Determining water and nitrogen balances for beneficial management practices using lysimeters at Wagna test site (Austria). *Science of the Total Environment* 499: 448–462.
- Klemm O, Gonçalves FLT, Katata G, Lin N-H, Eugster W, and Scholl M (2018) Preface to the AAQR special issue “fog, fog collection and dew” *Aerosol and Air Quality Research* 18: I–II.
- Kohnke H, Dreilbeis F, and Davidson J (1940) *A Survey and Discussion of Lysimeters and a Bibliography on their Construction and Performance*, p. 37. U.S. Department of Agriculture Miscellaneous Publication.
- Marek T, Piccini G, Schneider A, Howell T, Jett M, and Dusek D (2006) Weighing lysimeters for the determination of crop water requirements and crop coefficients. *Applied Engineering in Agriculture* 22: 851–856.
- Meissner R, Seeger J, Rupp H, Seyfarth M, and Borg H (2007) Measurement of dew, fog, and rime with a high-precision gravitation lysimeter. *Journal of Plant Nutrition and Soil Science* 170: 335–344.
- Peters A, Groh J, Schrader F, Durner W, Vereecken H, and Pütz T (2017) Towards an unbiased filter routine to determine precipitation and evapotranspiration from high precision lysimeter measurements. *Journal of Hydrology* 549: 731–740.
- Podlasky C and Schwärzel K (2013) Development of a continuous closed pipe system for controlling soil temperature at the lower boundary of weighing field lysimeters. *Soil Science Society of America Journal* 77(6): 2157–2163.
- Pütz T, Kiese R, Wollschläger U, Groh J, Rupp H, Zacharias S, Priesack E, Gerke HH, Gasche R, Bens O, Borg E, Baessler C, Kaiser K, Herbrich M, Munch J-C, Sommer M, Vogel H-J, Vanderborght J, and Vereecken H (2016) TERENO-SOILCan: A lysimeter-network in Germany observing soil processes and plant diversity influenced by climate change. *Environmental Earth Sciences* 75: 1–14.
- Pütz T, Fank J, and Flury M (2018) Lysimeters in vadose zone research. *Vadose Zone Journal* 17(1): 180035.
- Reth S, Perez-Priego O, Coners H, and Nolz R (2021) Lysimeter. In: Foken T (ed.) *Springer Handbook of Atmospheric Measurements*. Cham: Springer.
- Riedl A, Li Y, Eugster J, Buchmann N, and Eugster W (2022) Technical note: High-accuracy weighing micro-lysimeter system for long-term measurements of non-rainfall water inputs to grasslands. *Hydrology and Earth System Sciences* 26: 91–116.
- Roy J, Rineau F, De Boeck HJ, Nijs I, Pütz T, Abiven S, Arnone JA III, Barton CVM, Beenaerts N, Brüggemann N, Dainese M, Domisch T, Eisenhauer N, Garré S, Gebler A, Ghirardo A, Jasoni RL, Kowalchuk G, Landais D, Larsen SH, Leemans V, Le Galliard J-F, Longdoz B, Massol F, Mikkelsen TN, Niedrist G, Piel C, Ravel O, Sauze J, Schmidt A, Schnitzler J-P, Teixeira LH, Tjoelker MG, Weisser WW, Winkler B, and Milcu A (2021) Ecotrons: Powerful and versatile ecosystem analysers for ecology, agronomy and environmental science. *Global Change Biology* 27: 1387–1407.
- Saaltink MW, Kohfahl C, and Molano-Leno L (2020) Analysis of water vapor adsorption in soils by means of a lysimeter and numerical modeling. *Vadose Zone Journal* 19: , e20012.
- Schelle H, Iden SC, Fank J, and Durner W (2012) Inverse estimation of soil hydraulic and root distribution parameters from lysimeter data. *Vadose Zone Journal* 11. <https://doi.org/10.2136/vzj2011.0169>.
- Weihermüller L, Siemens J, Deurer M, Knoblauch S, Rupp H, Göttelein A, and Pütz T (2007) In situ soil water extraction: A review. *Journal of Environmental Quality* 36: 1735–1748.
- Wyrwich D and Brandenburg M (2018) 50 Jahre Großlysimeteranlage St. Arnold - Entwicklung der Klima- und Wasserhaushaltsgrößen bei unterschiedlicher Vegetation (1965–2014). In: *LANUV-Fachbericht*, LANUV-Fachbericht ed Recklingshausen, Germany: Landesamt für Natur, Umwelt und Verbraucherschutz NRW.
- Zhang Q, Wang S, Yue P, and Wang R (2019) A measurement, quantitative identification and estimation method(QINRW) of non-rainfall water component by lysimeter. *MethodsX* 6: 2873–2881.

Recognition of notable past soil scientists

Michael J Goss^a, Evgenia Blagodatskaya^b, Ariena HC van Bruggen^{c,*}, Monika Joschko^d, Stefan Schrader^e, and Christoph Tebbe^e, ^aSchool of Environmental Sciences, University of Guelph, Guelph, ON, Canada; ^bDepartment of Soil Ecology, Helmholtz-Zentrum für Umweltforschung GmbH – UFZ, Halle (Saale), Germany; ^cDepartment of Plant Pathology and the Emerging Pathogens Institute, University of Florida, Gainesville, FL, United States; ^dLeibniz-Zentrum für Agrarlandschaftsforschung (ZALF) e. V., Müncheberg, Germany; ^eInstitute of Biodiversity, Braunschweig, Germany

© 2023 Elsevier Ltd. All rights reserved.

Introduction	680
References	689
Further reading	690

Abstract

The chapter highlights the achievements of notable scientists who have made considerable contributions to our understanding of soils.

Introduction

The technical chapters in this edition of the Encyclopedia of Soils in the Environment very clearly demonstrate the impact over the last 20 years that the application of new technologies has had on our knowledge of soil, the material from which, with sunlight, we derive our daily food. Although the varied characteristics of soil have long been recognized, it would be fair to claim that modern soil science started with Nicolaus of Cusa (1401–1464), who designed a theoretical experiment (in 1450) that Johannes Baptista van Helmont (1579–1644) carried out in what is claimed to be first quantitative experiment in history. He grew a willow tree (*Salix* sp., L.) for a period of 5 years in a pot of soil, enclosed with a perforated lid, and established that the plant removed water from the soil and there was little or no loss of soil solids (van Helmont, 1648). Building on this Stephen Hales (1677–1761) distinguished between the weight of water lost from the soil during the growth of sunflower (*Helianthus annuus*, L.) and the mass transpired (Hales, 1727), essentially separating transpiration through the plant from evaporation from the soil. Hales also estimated the quantity of water added to soil from dew. Philippe de la Hire (1640–1718), who is generally recognized as pioneering the use of lysimeters, and Edmond Halley (1656–1742), now more associated with astronomy and planetary motion, investigated the water balance following precipitation reaching the soil surface. Firstly, de la Hire (1703) demonstrated links between precipitation, drainage and plant water use and then Halley (1687, 1691), recognizing the large volume of water in the atmosphere, emphasized that a quantitative estimate of evaporation was essential to formulate the hydrological cycle between sea and land. However, Halley (1691) did not believe that rain (plus dew) was the source for all springs and rivers. John Dalton (1766–1844), now better known for his atomic theory (Dalton, 1805), also made use of a lysimeter in correctly establishing the components of the hydrological cycle and applying his understanding to derive a water balance for England and Wales (Dalton, 1802a). This required an estimate of the loss by evaporation (E) which he determined as a function of the temperature of the water and the vapor pressure of the atmosphere, together with wind speed such that

$$E = c (e^* - e)$$

where c is a factor depending on wind speed, e^* is the saturated vapor pressure at the temperature of the water and e is vapor pressure of the air over the period of observation (Dalton, 1802b). This provides an example of the development of a topic, in the period of the natural philosophers, that is so critical to the understanding of soil and ecosystem functioning in this era of climate change. In this chapter, the contribution of some past scientists to soils in the environment is recognized.

In the 1st edition of this encyclopedia, the biographical accounts of the activities of J.B. Lawes (1814–1900) and J.H. Gilbert (1817–1901), at what is now Rothamsted Research (Harpenden, UK) focused on their field-based investigations, which earned them recognition as the founding fathers of the scientific method in agricultural research (Johnson, 2005). What is not mentioned is their establishment of the first lysimeters built to retain the natural structure and profile of the soil and its use in determining the quantity and chemical quality of the deep drainage component of the soil water balance (Lawes et al., 1881a,b). The contribution of H. L. Penman (1909–1984), also at Rothamsted, to the science of evaporation of water from land, either cropped or uncropped and rainfed or irrigated, was described in the 1st edition of this publication by Monteith, who contributed to the equation for evaporation from cropped land that bears both their names. Penman's 1948 equation overcame contemporary equipment limitations to measurement of the temperature of an evaporating surface by combining thermodynamic and aerodynamic equations for heat and mass transfer from a surface, where water was freely available. The equation included a parameter

*Emeritus Professor of Plant Pathology.

representing the change in saturated vapor pressure with temperature, which could be treated as a constant over the range of temperatures involved in most systems involving natural evaporation.

The contribution of Monteith (1929–2012) was to extend the Penman formula for application to vegetation (Monteith, 1965). His approach was to rewrite the Penman formulation in a form applicable to single leaves, developing the idea that diffusion from saturated surfaces within leaves was a component of the transfer pathway of water to the atmosphere. This required addition of a parameter to describe stomatal restriction of the diffusion of water. Another parameter was required to represent the external diffusion resistance controlling the flux of water vapor from the leaf driven by the vapor pressure difference between the leaf surface and the free atmosphere. This modified Penman equation has become known as the Penman–Monteith (PM) equation. In 1998 it was taken up as the basis of a globally valid standard for crop water requirement calculations for the Food and Agriculture Organization (FAO) of the United Nations (Allen et al., 1998). The FAO PM equation has the form that only requires meteorological data as input and the predicted evapotranspiration applies to a hypothetical crop, 0.12 m tall, having a fixed canopy resistance and albedo. It is envisaged as resembling an extensive area of green grass of uniform height, that completely shades the soil and has an adequate water supply. It has also become the standard approach to establishing irrigation needs (Allen et al., 2023).

In 1967, Monteith was appointed Professor of Environmental Physics in the School of Agriculture at Nottingham University, creating a strong team with both teaching and research responsibilities. The research had a dual focus: first, microclimate and carbon balance of barley crops, including physiological, agronomic and hydrological components; second, the physics of the animal environment, particularly animal heat balance. In 1976, the work on microclimatology was extended to include tropical crops. Based on the notes for his greatly appreciated class lectures, Monteith published the highly influential text “Principles of environmental physics” in 1973, which extended to four editions, the last three being jointly authored with Michael Unsworth (now Professor Emeritus, College of Earth, Ocean and Atmospheric Sciences, Oregon State University). Some 20 y after moving to Nottingham, during which he served as Head of the Department of Physiology and Environmental Science (1970–73 and 1979–82) and Dean of the School of Agriculture (1985–86), Monteith took up the position of Director of Resource Management Division at the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) in Hyderabad, India. His responsibilities also extended to outstations, including one in Niger in the sub-Sahel. Monteith effectively changed the way in which agronomists and crop physiologists considered the influence of weather on dry matter production by focusing on solar radiation, temperature and humidity. First, the rate of crop dry matter production was proportional to the rate at which a crop canopy intercepted radiant energy (the radiation use efficiency, e), when light was a major limiting factor, with crop yield being a linear function of the cumulative radiation intercepted over the growing season. Second, the dry matter produced per unit of water transpired (transpiration efficiency, q) is almost proportional to the mean value of the saturation vapor pressure deficit of the atmosphere (D) to which the canopy is exposed during the day.

Monteith then introduced the concept of normalized rainfall or irrigation (P_n), where

$$P_n = \frac{P}{D} \cdot D_0$$

P is the rainfall or irrigation in mm and D_0 is the normalizing factor, which has the same units as D (for D in kPa, $D_0 = 1$ kPa).

The dry matter produced per unit of water transpired can be similarly normalized and the slope of the ratio $j = e/qD$ provided a new climatic index and could be used to estimate the potential rate of transpiration (E), rather than by summing a radiation term and an aerodynamic term, as in the Penman equation.

The rate of dry matter accumulation per unit area is given by qE and by efS , where f is the fraction of solar radiation, S , incident per unit area

$$efS = qE$$

$$E = fjSD$$

Monteith plotted normalized monthly rainfall, j^* , against incident radiation over the same time period and indicated on the graph the relationship for a constant value of j per MJ m^{-2} . He argued that during those months where values of normalized rainfall, j^* , exceeded the value of j , crop production was energy limited but when j^* was less than j , it was water limited.

When water was limiting, Monteith's contribution was to demonstrate how crop transpiration depends on the rate at which the root system grows down and proliferates within the soil profile. The period available to the crop for growth is dependent on both soil and climate.

John Monteith was born on the west coast of Scotland but moved to Edinburgh at the age of eleven. Quietly spoken and friendly, he had a great sense of humor. At one workshop on thermocouple psychrometry led by his friend Gaylon Campbell, Monteith circulated a spoof menu for the dinner, offering “Campbell's soup, Welded joints à la Wescor (the main manufacturer at the time), with a sweet of chromel caramel (chromel is an alloy much used in thermocouple manufacture) with cooling currants all washed down with Sauterne Chateau Spanner” (Spanner had been a major proponent of thermocouple psychrometry) (Fig. 1). He is also reported to have jokingly attributed the origin of *albedo* to a fictitious American astronomer, Al Bedo, whom Monteith described as studying the reflectivity of stars!

Organic matter content is a critical component of soil, being crucial for structure, the chemistry and biology that underpin its functioning and associated ecosystem services. However, the chemical investigation of the material as a whole was based on its extraction from soil using a strong alkaline solution (now typically sodium hydroxide at pH 13) that was subsequently protonated

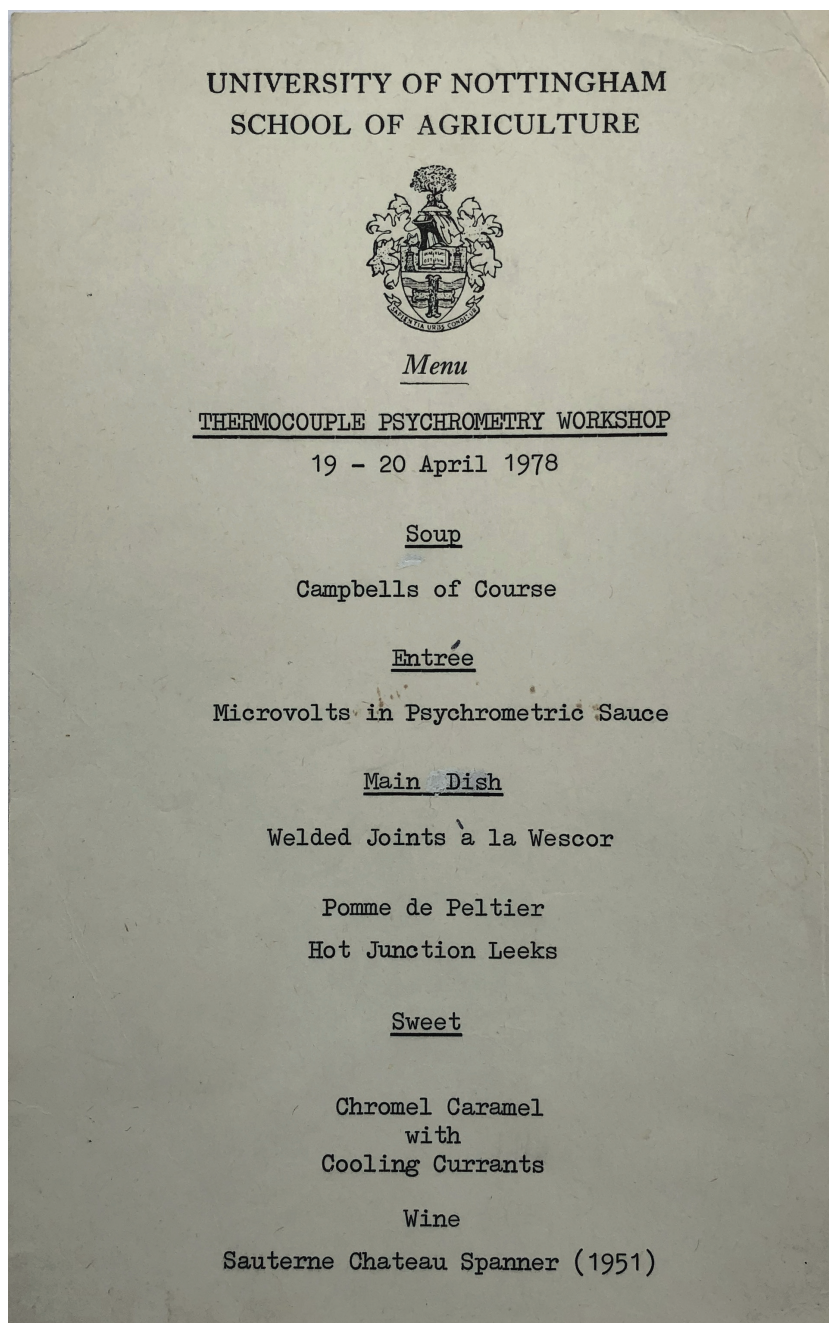


Fig. 1 Menu composed for dinner at a workshop on soil psychrometers by John Monteith.

to enhance its solubility. There have been a number of additional assumptions about the nature of soil organic matter (SOM) that have been associated with the analysis of the extracted material, including its recalcitrance and hence long residence time in soil. It has also been assumed that SOM has been formed following microbial breakdown of plant and animal remains, and the decomposed material being subject to secondary synthesis of compounds of large molecular weight under the influence of metals or extra-cellular enzymes. Recent assessments indicate that most of this picture of SOM chemistry is wrong, and only the role of soil microorganisms remains as key focus (Lehmann and Kleber, 2015).

The development of our understanding of the overall dynamics associated with SOM and the organisms involved owes much to Sir Edward John Russell (1872–1965), working first at Wye College, the School of Agriculture of the University of London, and then at Rothamsted. His research in the early years of the 20th Century provided new and broader understanding of microscopic soil organisms. His starting hypothesis was that the oxygen absorbed by uncropped soil was a measure of the total activity of microbes involved in decomposition processes and the release of plant nutrients. He devised a method to determine the uptake of oxygen into soil and used it to show that the rate of utilization in various soils of similar type (clayey, sandy, calcareous) correlated with

their fertility. Furthermore, this rate was greater in surface soils where conditions of temperature and moisture were more likely to encourage microbial activity. However, oxygen uptake was greatly reduced after autoclaving the soil at 130 °C for 30 min, supporting the claim that it was the microorganisms that were creating the oxygen demand.

Serendipitously, when a temperature of only 95 °C was accidentally used instead of 130 °C, the rate of oxygen uptake into soil greatly increased instead of reducing. In collaboration with F.V. Darbishire, Russell followed up this observation of the effect of incomplete sterilization, and was able to obtain similar results after heating soil to 100 °C or treating it with volatile biocides (Darbishire and Russell, 1907). Importantly, they investigated the growth of plants in pot experiments comparing the effects of the incomplete sterilization of soil with untreated soil from the field. Growth was better and the uptake of N, P and K was greater in pots containing the treated soil; a result attributed to changes in the microbe population. Similar heat (95 °C) and biocide treatments also resulted in a considerable increase in the formation of ammonia as well as in bacterial numbers (based on plate counts) relative to the untreated control soil (Russell and Hutchinson, 1909). They attributed their results to the removal by the incomplete sterilization of a 'harmful factor' that limited the number and activity of soil bacteria. The factor was likely an organism or group of organisms that were bigger than and predatory on the bacteria, possibly protozoa that were larger and more susceptible to the treatments, and their absence would allow the large increase in number and activity of the soil bacteria. Russell and Hutchinson suggested that the ciliates and amoebae present in untreated soil fitted the experimental evidence. Their premise was that soil normally contains an active population of protozoa that prey on the bacteria, but which are more readily killed by partial sterilization, and this further stimulates the beneficial activities of the bacteria into quicker release of plant nutrients.

The discovery that soils typically contained a wide range of different active microorganisms opened up a new field of research within soil science. But Russell was keen to apply research findings to practical agriculture. Recognizing that it was not a practical option for arable agriculture and aware that soils used in the glasshouse industry had a limited life before they developed 'soil sickness' and became unusable for cucumber or tomato production. The nature of the phenomenon was investigated on commercial operations in the Lea Valley on the northeast fringe of London. The soils were found to have very limited bacterial activity, releasing very limited amounts of plant nutrients. However, partial sterilization of the soil again resulted in the stimulation of the bacteria with protozoan predation being the suspected cause of the 'soil sickness'. Heat or biocide treatment became standard practice for maintain good growing conditions and an industry-sponsored research station was established to support growers.

As already noted, the partial sterilization treatment affected the nitrogen available in the soil, which stimulated a survey of seasonal fluctuations in the NO_3^- -N content of field soils together with linked data on weather, soil treatment and cropping. Russell and A. Appleyard then attempted to relate bacterial numbers to the decomposition of organic matter and the formation of CO_2 and NO_3^- -N in the soil. Plate counts of bacteria were made, and the content of NO_3^- -N and CO_2 determined in soil and soil atmosphere, respectively, from five plots on experiments at Rothamsted at a two-week interval over three cropping seasons. Russell and Appleyard argued that the fluctuations in numbers of bacteria matched those for CO_2 and NO_3^- -N, sufficiently to indicate a causal relationship, with temperature and rain being the main factors controlling microbial activity.

Scarcity of fertilizer materials and the imperative for greater food production in the British Isles during World War 1 increased the requirement for organic amendments, such as animal manure and composts. Russell then developed practical schemes for the formation and use of organic amendments, which he identified as contributing to improved soil structure, providing food for the beneficial soil microbes as well as their activity providing the nutrients required by plants. From this effort, two publications resulted: "A student's book on soils and manures" (Russell, 1915) and "Manuring for higher crop production" (Russell, 1916). Both went into second editions by the end of hostilities in 1919! His most well-known publication, entitled *Soil conditions and plant growth*, was first published in 1912 and he wrote 7 editions before his son E. W. Russell produced three more editions, the last in 1973. Since then, there have been two more versions, one in 1988 edited by Alan Wild and the other in 2013 jointly edited by Peter J. Gregory and Stephen Nortcliff.

Russell was forced to give up his education at a technical school in Birmingham at the age of 14, when his family moved to London. Understanding that chemistry (a favorite subject) was practiced in chemists' shops (pharmacies), he opted to work in one and save money that he could use to go to college. He was quickly disillusioned by the reality of his duties but embarked on educating himself by reading and attending lectures and enrolling in courses at the newly-opened philanthropic venues for education and culture. After 3 years his father moved the family to Yorkshire and Russell, unable to find work with a chemistry link, first assisted in his father's Unitarian Chapel before embarking on a course in Wales to become a non-conformist minister. There his education was guided by the college Principal, who encouraged Russell to study for examinations that would eventually end in an open scholarship in science at Aberystwyth University. This he achieved in under 2 years. During his first year in Aberystwyth, he at last received direction in his scientific studies and, when he sat for the London Intermediate B.Sc., he came out top in chemistry and obtained an additional two-year scholarship. After his second year, Russell transferred to Owen's College, Manchester, where his two mentors gave him expert tutelage in chemistry. At the end of his second year at Manchester, he was awarded his B.Sc. in Chemistry and was then appointed Lecturer and Demonstrator. His post-graduate research was on the rates of explosion in gases and the reduced reactivity resulting from desiccation of some combustible gas mixtures, for which he was awarded the degree of D.Sc. by London University.

Russell was aware of the poverty that many were experiencing, causing him to consider the possibilities for resettling the urban poor in rural areas and providing training necessary for them to become contributing members of society. It also made Russell reassess his own career direction, leading him to considered the chemical changes induced by microorganisms. As a result, he obtained a lectureship in chemistry at Wye Agricultural College, where he recognized that building a team of scientists and allowing them to apply their different disciplines to the investigation of soil, plants and animals, presenting their results in a way that could

allow experts, teachers and producers to use them, could help develop and improve agricultural practice. The possibility became a step nearer, when the Principal of Wye (A D Hall) became the Director of Rothamsted Experimental Station and subsequently in 1907 he invited Russell to become the 'Goldsmith's Company Soil Chemist'. Five years later, Russell was appointed Director to succeed Hall who had become a Commissioner of the Development Fund set up by the British Government to support the scientific development of afforestation, agriculture and fisheries. In the 31 years of his Directorship, Russell was able to increase the scientific staff from 6 to 70, with a similar increase in the number of support staff. He was able to oversee the purchase of the estate and the integration of the small experimental station at Woburn belonging to the Royal Agricultural Society. This allowed comparative studies on light sandy to the more clayey soils at Rothamsted. He put in place various arrangements to ensure that each department knew what the others were doing and encouraged cooperation between them. Russell also encouraged visitors to the station, including farmers, members of scientific societies and parties of schoolchildren. Individual visitors were accommodated for shorter or longer periods as 'voluntary workers'. Some came to discuss problems arising from their work or to learn some new experimental technique. Many people, especially from overseas, carried out research on a specific problem, often as part of their postgraduate studies.

Russell received the award of the Order of the British Empire in 1918, and was elected a Fellow of the Royal Society in 1917. In 1922 he was created Knight Bachelor.

In 1927 he was invited to a conference of the recently formed International Society of Soil Science, held in the USA, and decided to become more involved in such international activities, which he saw as helping some of the poorer countries to increase their food production. He was President of the International Soil Congress held at Oxford in 1935.

Russell's influence was not limited to Britain but included its colonies and other overseas territories for which it had responsibility. He became aware that there was little exchange of information on current research between local researchers and those in Britain, resulting in an over-dependence on US work and leadership. Following a conference he initiated, specialist information bureaux were established and added to the Imperial Bureaux of Entomology and Mycology, with the task of publishing abstracts of literature, produce monographs and to provide those seeking information with the relevant list of source publications appropriate for their research. This became the Commonwealth Agricultural Bureaux in 1947 and CAB International in 1986.

In 1957, some 14 years after Russell retired, David Stewart Jenkinson (1928–2011) began working at Rothamsted Experimental Station. After gaining first-class honors in Experimental Science for both BA and BSc degrees, he remained at Trinity College, Dublin and took his PhD in organic chemistry. After a short period in industry Jenkinson decided on a more academic approach to science and obtained a 3-year appointment in soil science at the University of Reading, investigating the chemical structure of 'humic acids' in soil organic matter. He quickly realized that even using the cutting-edge methods of the time, including electrophoresis and infra-red or ultra-violet spectrometry, the nature of the formation of humic materials and their linkage with mineral surfaces meant their composition was likely to be extremely variable. However, his efforts were not without reward as Jenkinson and his colleagues, Goulden and Tinsley, were able to demonstrate for the first time that humic material contained peptide bonds but there were no chemical signatures for lignin. These results refuted the hypothesis then being advanced that the compounds were lignoproteins. Jenkinson also learned from Tinsley how important was painstaking attention to detail in quantitative analytical chemistry! Throughout his career he passed on this lesson to all who worked with him.

Jenkinson began his career at Rothamsted with a study the decomposition and turnover of organic matter in soil incubated under ambient conditions in bottles buried upside down in an unplanted field. He first had to develop the means of uniformly labelling plant material with ^{14}C and of determining the amount of label remaining in the soil after periods ranging from weeks to years. He solved both challenges, designing a growth chamber into which $^{14}\text{CO}_2$ could be introduced and developing new techniques to determine the ^{14}C remaining in the soil. He was able to show the rapid breakdown of both root and shoot material over the first year, with little difference between them after the first 3 months, and slow decomposition over subsequent years. Jenkinson noted that the loss of carbon from the soil did not occur at a uniform rate throughout the year but reflected the temperature in the soil at 10 cm depth. Ten years later, working in Nigeria he and A. Ayanaba showed that the rate of decomposition in soil of similar clay content was four times faster under tropical conditions but followed the same trajectory as at Rothamsted. In addition to temperature, the soil water regime influenced decomposition, with the short-term rate being slower in the dry than in the wet season (Jenkinson and Ayanaba, 1977).

During the incubation experiments at Rothamsted, Jenkinson also started investigating the effects of "partial sterilization" of the soil on the decomposition of organic matter. In his classic 1966 paper he assessed the various hypotheses to explain the greater microbial activity that E.J. Russell had noted and others had continued to investigate. Jenkinson argued that the process used to effect partial sterilization of the soil might release previously unavailable organic matter but this could also result from the killing of some of the microorganisms present. Jenkinson's experiment consisted of mixing soil samples taken from his field incubation study containing ^{14}C -labelled organic material and subjecting the mixture to a wide range of sterilizing treatments of varying severity and then determining the carbon dioxide (CO_2) evolved during incubation, adding a 1 mL suspension of fresh soil as inoculum.

The treatments ranged from air-drying at 20 °C to autoclaving at 120 °C, γ -irradiation, fumigating with alcohol-free chloroform (CHCl_3) or bromomethane (CH_3Br), or oven drying at 80 or 100 °C for 24 h. In all treatments the total amount of CO_2 evolved during incubation increased relative to untreated (Control) soil as did the labelled $^{14}\text{CO}_2$, with the two fumigant treatments resulting in a greater proportion of $^{14}\text{CO}_2$ than the others and being more than threefold that of the control. Diluting the soil oven-dried at 80 °C with fresh soil had no effect on the $^{14}\text{CO}_2$ evolved. Jenkinson concluded that this latter result was inconsistent

with proposed hypotheses about partial sterilization being associated with the presence of toxins in fresh soil, to prey-predator impacts, or to the changed presence of antagonistic microbial relationships. The results from all treatments were also inconsistent with hypotheses of changed availability of previously protected organic material, changed susceptibility to breakdown or to increased vigor of a surviving population following partial sterilization. The only “reasonably satisfactory explanation” was that the increased flush of CO_2 after partial sterilization was due to the decomposition of microbes killed by the treatment. Jenkinson then carried out some additional studies, using repeated partial sterilizations and showed that this initial conclusion was justified and that only a very small part of the $^{14}\text{CO}_2$ evolved came from a constant background production; heat and irradiation make some parts of the soil organic matter (both microbial and non-microbial material) more susceptible to decomposition but partial sterilization treatments generally do not prime non-microbial organic matter for decomposition. Killed or damaged microbes decompose more quickly than those unaffected.

With this background Jenkinson then developed the concept of a method to determine the quantity of C held in living cells in soil, termed the microbial biomass, requiring a treatment to “give complete or near complete” sterilization but does not change the susceptibility of non-microbial organic matter to decomposition and leaves no residues. He settled on chloroform fumigation as this was easier to use than bromomethane. Then by determining the increased CO_2 released by fumigation relative to untreated soil (F), the microbial biomass (B) is given by:

$$B = \frac{F}{k}$$

Where k is the proportion of microbial biomass carbon that decomposes in the post treatment incubation, which he determined experimentally for the ammonia oxidizing bacterium *Nitrosomonas europaea* as approximately 0.3 (based on a 10-day incubation at 25 °C).

The value for k suggested in Jenkinson (1976) was 0.5, taking account of a wide range of organisms. The method was investigated in detail by Jenkinson and Powlson (1976), in part stimulated by criticisms of the underlying assumptions made by E.A. Paul and colleagues in Canada – an example of scientific disagreement leading to additional in-depth investigation.

The limitation of the method for use as a monitoring or survey tool was clearly the long incubation time. Later work that Jenkinson undertook with Brookes and Vance, led to the introduction of extraction of chloroform-released organic-C using 0.5 M potassium sulfate (K_2SO_4) as an alternative to 10 days incubation.. Although the comparison with an unfumigated control was important, other than in very acid soils, and less carbon was extracted than released as CO_2 , the procedure took hours rather than 10 days. Subsequently the fumigation-extraction method has been very widely used. The estimate of microbial biomass carbon was robust and the basic technique could be adapted to determine the biomass nitrogen (Brookes et al., 1985) with 0.5 M NaHCO_3 at pH 8.5 being used as the extractant for the phosphorus content of the biomass (Brookes et al., 1982).

In 1978 Jenkinson, in collaboration with J.M. Oades at the Waite Agricultural Research Institute in Australia, recognized the potential of determining the energy present in biomass by its adenosine-triphosphate (ATP) content. The main problem was the ability to extract ATP from cells without its breakdown and the nature of the relationship between ATP and the microbial biomass present. Jenkinson and Oades (1979) solved the problem of the extract, adopting 0.5 M trichloroacetic acid (TCA), coupled with 0.25 M disodium hydrogen orthophosphate (Na_2HPO_4) and 0.1 M 1,1'-dimethyl-4,4'-bipyridylium dichloride (paraquat) (now replaced by 0.6 M Imidazole), where Na_2HPO_4 and paraquat serve to prevent ATP from becoming fixed to negative and positive charged sites in the soil during extraction.

Other extractants appeared to remove less ATP than did TCA, which also correlated well with the biomass estimated by the post-fumigation CO_2 flush determination. Results have also indicated that despite being a population that is largely inactive because of limited substrate availability, it maintains ATP levels that are more typical of actively growing populations in laboratory culture. There is no evidence that ATP from roots in soil being sieved prior to microbial biomass determination contribute to the total extracted.

Jenkinson made important contributions to the understanding of the nitrogen cycle in agricultural fields, particularly in constructing a nitrogen balance for the long-term arable experiment on Broadbalk at Rothamsted, recognizing the very significant contribution from the atmosphere as wet and dry deposition and encouraging its direct measurement. He also used ^{15}N -labelled fertilizer to identify that which was quickly taken up by the crop and that which entered the soil organic matter, observing its long-term fate, identifying losses to the environment via the drainage system as well as the factors that reduced fertilizer use efficiency. Similar investigations were aimed at understanding the fate of N under permanent grass.

The information gained was put to use in developing computer models that could help advise the agricultural community, both on managing N (the SUNDIAL model) and on managing soil organic matter (the RothC model). The latter has been widely adopted for use in conjunction with global and regional circulation models to identify effects on soils and their productivity as well as what changes in inputs might be required to maintain carbon stocks.

Born in Beverley Hills, California, Jenkinson moved to a small farm in County Armagh, Northern Ireland when he was 4 years old after his parents' fortune was adversely affected by the Wall Street crash. He was kindly, supportive to many scientific colleagues and quietly spoken until roused. He rarely gave direct instructions to his junior collaborators but if he “strongly suggested” a course of action he would be distinctly annoyed if it was not followed! He read widely and had interests in Irish folk music. It is possible to hear him talk of his experiences in a recorded interview at <http://cadensa.bl.uk/uhtbin/cgisirsi/?ps=aCwKfzqhqr/WORKS-FILE/172200051/18/X246/XTITLE/Professor+David+Jenkinson+interviewed+by+Paul+Merchant>.

In recognition of his research achievement Jenkinson was elected a Fellow of the Royal Society in 1991 and received Massey Ferguson National Agricultural Award in 1993 for his research on nitrogen fertilizers. He was elected an Honorary Member of the Soil Science Society of America in 1995 and of the British Society of Soil Science in 2007. He retired from Rothamsted in 1988 after 31 years.



Prof. Dr. Klaus Domsch (1926–2022) was an outstanding scientist with very broad human and scientific interests, who enhanced environmental sciences by bridging microbiology, plant physiology, soil science, biogeochemistry, and agroecology. His contribution to the development of soil biology as individual branch of science is impossible to overestimate.

Born in Chemnitz, he attended the princely school and “St. Afra” boarding school in Meissen. In 1944, at the age of 18, he was drafted as an auxiliary in the German Luftwaffe. After receiving an ‘emergency Abitur’, he trained as a mountain infantryman in Innsbruck. In 1947 he enrolled at the Humboldt University in Berlin and initially studied teaching. In 1949, he transferred to the University of Göttingen to study phytopathology. In 1953, he obtained his doctorate at the Institute of Plant Pathology. From 1954 to 1967, he held his first position at the Institute for Cereal-Oil Fruit and Forage Plant Diseases of the then Federal Biological Research Station in Kiel-Kitzeberg.

Since 1967, the name Domsch has been closely associated with soil biology. That was when he founded and became Director of the Institute of Soil Biology at the Federal Research Centre of Agriculture (FAL) in Braunschweig. Up to his retirement in January 1991, he had authored several books and more than 165 publications. As head of the institute, he was particularly concerned with mentoring new colleagues through scientific stimulation, personal empowerment, regular “eye-to-eye” discussions and pointing out the possible perspectives. During the period of his 24-years as Director, the Institute of Soil Biology obtained a world-wide international reputation for its excellent research and broad networking activities.

From the first publication in 1953, Domsch focused his research on soil fungi, the influence of pesticides on soil micro-organisms and on microbial degradation of pesticides in soil. The research was summarized in several books including a “Compendium of Soil Fungi” (Academic Press, 1980), a book that was the bible for soil mycologists for many years and was cited more than 7000 times.

Domsch made a great contribution to soil biology with his instigation of physiological approaches for microbiology in soil research. In 1962 he introduced soil respiration as a basic parameter for environmental research and related the CO_2 evolved from soil to the activity of its resident microbes. This was a revolutionary step enabling quantitative estimation of soil microbial biomass by substrate-induced respiration (SIR method in cooperation with J.P.E. Anderson) (Anderson and Domsch, 1978); differentiation between fungal and bacterial activity (the fungal:bacterial ratio) and finally a development of an eco-physiological approach (Anderson and Domsch, 2010) based on microbial metabolic indexes – (the $q\text{CO}_2$ and $C_{\text{mic}}:C_{\text{org}}$ ratio). Visiting scientists came from all continents to study these new methods. Publication of eco-physiological approach was awarded as a Citation Classic in Soil Biology & Biochemistry in 2010. For his great contribution to international science, Domsch was awarded the Gregor Johann Mendel Honorary Medal for Merit in Biological Sciences by the Czech Academy of Sciences in 1995.

His far-sightedness influenced a whole generation of scientists who are looking to link excellent basic research and practical agriculture in the field of soil biology; something so very important today if we are to make agriculture more sustainable, but also efficient.

Klaus was in great demand internationally and developed close and productive research involvements in the Netherlands (Centraalbureau voor Schimmelcultures, Baarn) and the USA (University of California at Riverside and at the Department of Plant Pathology Oregon State University at Corvallis).

As a scientist, he took enormous pride in and responsibility for the results of his research. Socially he impressed by the depth of his intellect, inexhaustible energy, wisdom, kindness, and charisma.

Domsch was an unconventional man and scientist. His motto in life was “Carpe Diem” (seize the day!). He was a living demonstration to junior colleagues that scientific excellence and enjoying life are not contradictions. His personal celebrations were unforgettable: full of (scientific) discourse, culture (music) and joie de vivre - an incentive for youth. His deep-rooted beliefs in classical philosophy and his humor were Inspiring!

After his retirement in 1991, he investigated his family history and the complexities of genealogy replaced that of the soil microbial world. On February 25, 2022, one day after Russia's invasion of Ukraine, Prof. Dr. Klaus Domsch, died in Wittmar, Wolfenbüttel County, at the age of 96.

Barbara Gertrud Hedwig Mosse (1919–2010) completed a Diploma and BSc in Horticulture before obtaining a First class Honours BSc in Botany, the latter while working as an Assistant Chemist for the Lawes' Chemical Company. In 1944 she moved to the Pomology Department at the East Malling Research Station in Kent, where she worked on rootstock incompatibility. Her results were reported in a technical communication for the Commonwealth Agricultural Bureau (Mosse, 1962). She published her first paper on mycorrhizas based on work that also formed part of her PhD, awarded in 1956. The paper in the journal *Nature* (Mosse, 1953) consists of four short paragraphs and a composite of three micrographs showing the hyphal connection between a “fructification” and a strawberry root, an enlargement of the point of entry of a hypha into the root and the colonization inside the root. The last paragraph describes her current efforts to “to synthesize mycorrhiza under controlled (aseptic) conditions.” She states that she had produced typical mycorrhizal infections in seedlings grown in autoclaved soil in the test tube by taking spores from the sporocarp of a fungus associated with strawberry roots in the field. That was very much a break-through moment in the field of mycorrhizal research as prior to that time researchers could not even be sure whether the fungus they were observing was a mycorrhizal fungus. Furthermore, it allowed pot cultures of a mycorrhizal fungus to be established on different host plants. Now there was reason to separate spores from the soil to have access to different fungi. Mosse next established that mycorrhizal apple seedlings and clonal leaf bud cuttings were significantly larger than their non-mycorrhizal counterpart and recognized that this had implications for the nutrient levels in the soil and in the plants. The shoot content of not all nutrients were enhanced, with some such as copper, iron and potassium showing positive increases but manganese was reduced.

In 1960 Mosse transferred to Rothamsted Experimental Station. She continued to investigate the potential for inoculating plants with mycorrhizal fungi under aseptic and non-sterile conditions. The importance of early inoculation to obtain the greatest impact on growth was recognized for applicability in annual plants. The role that some bacteria (MHB) can play in helping mycorrhizal fungi enter a host-plant root was also identified.

Mosse surveyed plots on the Broadbalk experiment for spores of mycorrhizal fungi in 1961–2, 1966 and again in 1968. Together with G.D. Bowen from Adelaide She investigated soils from Australia and New Zealand, producing a key for identification of mycorrhizal spore types as well as assessing the conditions most conducive to the presence of spores. Moist conditions and well-worked soils, such as in gardens, and elevated nutrient levels were associated with fewer spores, but soils under natural vegetation were less numerous and diverse than in cropped land. They also concluded that where root growth and development was more sporadic there were more spores.

Much of the next decade was spent investigating the potential benefits from mycorrhizal inoculation. Several studies indicated that the benefits to the host plant could be attributed to the improved supply of phosphorus (P). In work with D.S. Hayman, it was confirmed using ^{32}P -labelling that the fungal partner did not access different sources of P compared with the host-plant roots but was more effective in absorbing it from soils with smaller concentrations and could capture it from outside the soil volume available to a root. In one experiment, there was some evidence that inositol could be a potential source, which later researchers have showed can occur if bacteria are available to break it down and release mineral P for the fungus to take up. Her work also demonstrated that increased concentrations of P and N in the soil could have negative effects on the level of colonization within the root, even to the extent of apparently excluding the fungus, which has been a major cause of concern and even for dismissing the relevance of mycorrhiza in intensively managed crops. The effectiveness of the fungus in supplying P varied between different hosts and fungi, results that subsequently have been extensively investigated by the research community in recent years. There was also evidence that the laboratory isolates were not always more effective than the indigenous mycorrhizal fungi. Mosse's work also foreshadowed the ability of roots to attract specific beneficial microbes from the bulk soil to the rhizosphere. She saw the potential for inoculation in parts of the world with soils with low levels of available P.

Another important aspect that attracted her attention was the potential for interaction between mycorrhizal fungi and rhizobia in N-fixing legumes. It was observed in one experiment on an acid soil that rhizobia only formed sufficient nodules when the plants had also been colonized by mycorrhizal fungi. This kind of information provided useful background when the significance of signaling between microbe and host plant as part of the symbiosis formation was being explored some 20-years later.

Mosse had many people visit her at Rothamsted to gain experience in her laboratory and one with a particularly significant outcome was with M. Giovannetti. Their paper (Giovannetti and Mosse, 1980) compared 4 methods for assessing the infection by mycorrhizal fungi in host-plant roots. The methods chosen were:

- 1 the Gridline intersect method. This was used to estimate the proportion of roots infected and the total length of root and counting the number of interceptions that stained roots made with grid-lines.
- 2 the Visual assessment. Quick estimate of what percentage of the root in the dish showed as stained. Both these methods could be done under a microscope.
- 3 Slide method 1. The roots were cut into 1 cm long sections and 10, randomly chosen, put on a single microscope slide and the length of stained root in millimeters determined for each root and the percentage of the total root calculated.

- 4 Slide method 2. On similar samples of roots the presence or absence of stain was recorded and then expressed as a percentage of the whole.

Both these latter methods were used under the microscope. The results were subjected to statistical assessment that allowed the minimum number of roots or root pieces needed to be used for each method to be determined as well as the sensitivity. In addition, Giovannetti and Mosse reviewed the potential sources of error in each method. It was the first time that the standardization of methodology for making the key assessment in such studies had been attempted.

Mosse managed to travel widely, obtaining several Fellowships that she took at the University of California, Berkeley (1956); the Department of Arboriculture at the University of Pisa, Italy (1960) and the University of Otago, New-Zealand (1966) and was Visiting Scientist in the CSIRO Division of Soils in Adelaide, Australia (1966). She gave courses on mycorrhiza in Brazil, Nigeria, Malaysia and Hawaii.

Nicolson and Gerdemann (1968) published an account of the mycorrhizal fungi identified as members of the genus *Endogone*, including a new species *E. mosseae* (now *Funneliformis mosseae*), named in her honor.

It is clear that she was very passionate about her work. When José-Miguel Barea was visiting Rothamsted, on being told he was working on phosphate solubilizing bacteria she was very clear that he should work on mycorrhiza. Her enthusiasm for the subject and her influence on colleagues around the world have, as one person stated "I believe that she was a fundamental pillar for the development of mycorrhizal research in Europe and throughout the world, at a difficult time, especially as a woman. I think she can be fairly considered as 'the mother of mycorrhizae'."



José-Miguel Barea (1942–2018) was a native of Granada and studied Pharmacy at the University of Granada but even before he completed his degree in 1965, he was working with microbes in the Department of Microbiology. He completed his PhD in Microbiology in 1968, and his thesis work laid the foundation for the further study in Spain of the application of microorganisms that promote the growth of plants by supplying essential nutrients. For both the quality and novelty of his work he was given an Extraordinary Degree Award by the University of Granada.

In 1972 Barea spent a few months working on phosphate solubilizing bacteria (PSB) at Rothamsted Experimental Station, where at one teatime he had a chance meeting with Barbara Mosse, who asked what he was doing. He replied that he was working with PSB. Her response was to tell him that he needed to work with mycorrhiza, which were much more likely to help plants improve their P nutrition. His subsequent focus on mycorrhiza was therefor somewhat accidental but he considered it a stroke of fortune.

Barea obtained a position at the Consejo Superior de Investigaciones Científicas (CSIC), Zaidin Experimental Station (called 'El Zaidin' by Barea). He developed the work begun at Rothamsted on the effects of diazotrophs on the growth of higher plants. The evidence suggested that in *Azotobacter paspali* there was no transfer of nitrogen to host plant roots but its enhanced growth was associated with the release of plant growth regulating compounds, such as auxins gibberellins and cytokinins. He then developed a rapid test using yeasts to identify such substances and went on to show that these microbial products were important in the early process of root colonization by the fungi and now known to be a key part of the process of establishing the symbiosis. Barea included mycorrhizal fungi in his studies and, linked to work with D.S. Hayman at Rothamsted, reported on possible synergistic interactions between the mycorrhizal fungus *Endogone* and phosphate-solubilizing bacteria in low-phosphate soils.

Retrospectively it is possible to identify three strands to his work on mycorrhiza: the interaction between beneficial fungi and rhizobacteria, the role of mycorrhizas in restoration of marginal and degraded soils and the biodiversity of mycorrhiza in natural habitats. The first of these developed from his early interest in PSB and the interaction between rhizobia and mycorrhiza (de Aguilar et al., 1979) which provided a basis for the use of microbes in the supply of nutrients and protection against root pathogens in sustainable crop production. One aspect was to enhance the supply of P from rock phosphate for plants by the joint use of PSB and mycorrhizal fungi. In one investigation it was found that although the inoculation of alfalfa (*Medicago sativa* L.) with the mycorrhizal fungus *Funneliformis mosseae* was more effective than the indigenous fungi, in combination they were even better. Another part of the focus was on enhancing the colonization of host plants using rhizobia as mycorrhizal helper bacteria and part was using mycorrhizal fungi to enhance the supply of P to support biological N-fixation by rhizobia, such as in non-crop woody legumes. The later aspect also contributed to restoring degraded land in semi-arid regions, including the Mediterranean basin.

Barea's interest in the restoration of degraded soil not only covered revegetation with fast-growing trees and shrubs but included improving soil structure through the use of mycorrhizal fungi and reclaiming soils contaminated with heavy metals, using mycorrhiza to protect plants from the toxicity. His investigations also led to an evaluation of the activity of indigenous mycorrhiza in desertified semi-arid land, which suggested that there was sufficient to support the recovery of vegetation cover. He also considered the potential importance of the protection provided by mycorrhiza against drought and studied their effects on preventing premature aging in plants under such conditions.

He developed a general interest in the conservation of endangered native flora found in Granada's Sierra Nevada, and in other natural habitats. It was typical that he would seek to evaluate the potential role for mycorrhiza, including their biodiversity. One of the products of this line of research was the identification of a new mycorrhizal fungus *Entrophospora nevadensis* (Palenzuela et al., 2010). Barea recognized that native mycorrhizal fungi from saline locations could be important in the recovery of areas affected by salinity in natural or agricultural environments. In the Cabo de Gata Natural Park, Barea investigated the community of AMF in the rhizosphere of *Asteriscus maritimus* L. a plant adapted to Mediterranean saline areas. He and his colleagues found from a sample of 40 plants, 30 AMF species belonging to three classes, five orders, nine families and 13 genera of the phylum Glomeromycota (Estrada et al., 2013).

Barea saw the significance the molecular era for mycorrhizal research, using it to combine his interests in the biodiversity of organisms with better understanding their potential functioning within their local habitat.

Barea served as the Head of the Microbiology Department and Vice-Director of the Institute between 1983 and 1987, and as Director between 1989 and 1998. He served as an advisor for the Joint FAO/IAEA (International Atomic Energy Agency) in Vienna, where he promoted the use of $^{15}\text{NO}_3^-$ as a tracer for nitrogen metabolism in microbial and plant studies. He was a charismatic and inspiring teacher and scientific leader.

Barea was an advisor for the International Foundation for Science (IFS), which is based in Sweden, and supports research and research training projects in developing countries – a project close to his heart.

He was elected to the Academy of Mathematics, Physics, Chemistry and Natural Sciences of Granada in 2004. In the same year, he was appointed as a distinguished Professor at the

University of Buenos Aires in Argentina. In 2013, he received the President's Medal from the University of La Frontera (Chile) in recognition of his outstanding contributions to the formation of advanced human capital at the institute.

Barea had the reputation of working hard and playing hard. He enjoyed singing and was a member of a folk group which performed at many of the meetings that he organized.

Vladimir Victorovich Zelenev (1961–2021) was a multi-talented scientist combining unique expertise in soil science, soil microbiology, ecological modeling and environmental ecology. His academic career spanned a wide range of topics in the course of time. He received his MSc in Agrochemistry and Soil Science with honors from the Faculty of Soil Science of Moscow State University in 1983. Zelenev successfully defended his PhD thesis entitled "Spatial and temporal fluctuations in bacteria, microfauna and mineral nitrogen in response to a nutrient impulse in soil" at Wageningen University in the Spring of 2004. He received several fellowships from the International Agricultural Center and from the Graduate School PE&RC for PhD research at Wageningen University. His research on the production and emission of various greenhouse gases from soils in natural ecosystems (Van Bruggen et al., 2017), including carbon dioxide (CO_2) and methane (CH_4), led to Zelenev being given a prestigious fellowship from the Austrian Government to work at the International Institute for Applied Systems Analysis at Laxenburg.

Zelenev was the first scientist to write a simulation model for oscillating populations of bacteria interacting with their substrate (Zelenev et al., 2000, 2005). The unique aspect of this model was the realization that predators or parasites were not needed to initiate wave-like fluctuations in bacterial populations. There were many models describing oscillations in predators and their prey, but regulation of bacterial densities by interaction with their substrate was new. Zelenev introduced the concept of readily utilizable substrate. He realized that oscillations would ensue if growth and death rate curves crossed over at a particular substrate level, referred to as "scissors."

Zelenev was one of the recognized experts in the field of modeling decomposition of organic matter (Shahbaz et al., 2017). He was one of few modelers who combined a bottom-up and top-down approach to simulating organic matter decomposition. This resulted in the model "BACWAVE-WEB" with the necessary detail at the base and yet sufficient functional groups in the food web to be able to test some hypotheses about the role of these groups in organic matter decomposition and mineral nitrogen release (Zelenev et al., 2006). The "BACWAVE-WEB" model has excellent potential to predict responses of microbial communities to a disturbance, which could be used to characterize soil health: oscillations with greater amplitudes indicating poorer soil health.

Zelenev was an amiable and helpful friend and colleague to everyone who crossed his path. He had vast knowledge and interests, yet he always remained a modest person. He was very fond of literature, art and classical music, was fond of mountain skiing. Zelenev had many friends all over the world. He sang well, played the guitar, he was pleasant to talk to, witty, with an amazing sense of humor. He was a dedicated researcher and teacher, always and anywhere.

References

- Allen RG, Pereira LS, Raes D, and Smith M (1998) Crop evaporation: Guidelines for computing water requirements. In: *Food and Agriculture Organization of the United Nations FAO Irrigation Paper 56*. FAO Rome.
- Allen RG, Kilic A, D'Urso G, and Consoli S (2023) The penman-Monteith equation. In: Goss MJ and Oliver MA (eds.) *Encyclopedia of Soils in the Environment*, 2nd edn Elsevier.

- Anderson JPE and Domsch KH (1978) A physiological method for the quantitative measurement of microbial biomass in soils. *Soil Biology & Biochemistry* 10: 215–221.
- Anderson T-H and Domsch KH (2010) Soil microbial biomass: The eco-physiological approach. *Soil Biology & Biochemistry* 42(12): 2039–2043. <https://doi.org/10.1016/j.soilbio.2010.06.026>.
- Brookes PC, Landman A, Pruden G, and Jenkinson DS (1985) Chloroform fumigation and the release of soil nitrogen: a rapid direct extraction method for measuring microbial biomass nitrogen in soil. *Soil Biology & Biochemistry* 17: 837–842. [https://doi.org/10.1016/0038-0717\(85\)90144-0](https://doi.org/10.1016/0038-0717(85)90144-0).
- Brookes PC, Powlson DS, and Jenkinson DS (1982) Measurement of microbial biomass phosphorus in soil. *Soil Biology & Biochemistry* 14: 319–329. [https://doi.org/10.1016/0038-0717\(82\)90001-3](https://doi.org/10.1016/0038-0717(82)90001-3).
- Dalton J (1802a) Experiments and observations to determine whether the quantity of rain and dew is equal to the quantity of water carried off by the rivers and raised by evaporation; with an enquiry into the origin of springs. *Memoirs of the Proceedings of the Literary and Philosophical Society of Manchester* 5 Part 2: 346–372.
- Dalton J (1802b) Experimental essays on the constitution of mixed gases; on the force of steam or vapour from water and other liquids in different temperatures, both in a Torricellian vacuum and in air; on evaporation expansion of gases by heat. *Memoirs of the Proceedings of the Literary and Philosophical Society of Manchester* 5 Part 2: 535–602.
- Dalton J (1805) On the absorption of gases by water and other liquids. In: *Memoirs of the Literary and Philosophical Society of Manchester. 2nd Series*. 6 271–287.
- Darbishire FV and Russell EJ (1907) Oxidation in soils, and its relation to productiveness. *Journal of Agricultural Science* 2: 305–326. <https://doi.org/10.1017/S0021859600000605>.
- de Aguiar CAG, Azcón R, and Barea JM (1979) Endomycorrhizal fungi and Rhizobium as biological fertilisers for Medicago sativa in normal cultivation. *Nature* 279: 325–327. <https://doi.org/10.1038/279325a0>.
- de la Hire P (1703) Remarques sur l'eau de la pluie, et sur l'origine des fontaines; avec quelques particularités sur la construction des cisternes. *Memoires de l'Academie Royale* 1703: 56–69.
- Estrada B, Beltran-Hermoso M, Palenzuela J, Iwase K, Ruiz-Lozano JM, Barea J-M, and Oehl F (2013) Diversity of arbuscular mycorrhizal fungi in the rhizosphere of *Asteriscus maritimus* (L.) Less., a representative plant species in arid and saline Mediterranean ecosystems. *Journal of Arid Environments* 97: 170–175. <https://doi.org/10.1016/j.jaridenv.2013.05.019>.
- Giovannetti M and Mosse B (1980) An evaluation of techniques for measuring vesicular arbuscular mycorrhizal infection in roots. *New Phytologist* 84(3): 489–500.
- Hales S (1727) *Vegetable statics*. London: W. and J. Innes and T. Woodward. <http://www.biodiversitylibrary.org/page/288153>. Accessed 10 July 2023.
- Halley E (1687) An estimate of the quantity of vapour raised out of the sea by the warmth of the sun; derived from an experiment shown before the Royal Society, at one of their late meetings. *Philosophical Transactions of the Royal Society* 16: 366–370. <https://doi.org/10.1098/rstl.1686.0067>.
- Halley E (1691) An account of the circulation of the watry vapours of the sea, and of the cause of springs, presented to the Royal Society. *Philosophical Transactions of the Royal Society* 17: 468–473. <https://doi.org/10.1098/rstl.1686.0084>.
- Jenkinson DS (1976) The effects of biocidal treatments on metabolism in soil-IV. The decomposition of fumigated organisms in soil. *Soil Biology & Biochemistry* 8: 203–208.
- Jenkinson DS and Ayanaba A (1977) Decomposition of Carbon-14 labeled plant material under tropical conditions. *Soil Science Society of America Journal* 41: 912–915. <https://doi.org/10.2136/sssaj1977.03615995004100050020x>.
- Jenkinson DS and Oades JM (1979) A method for measuring adenosine triphosphate in soil. *Soil Biology & Biochemistry* 11: 193–199.
- Jenkinson DS and Powlson DS (1976) The effects of biocidal treatments on metabolism in soil-v. A method for measuring soil biomass. *Soil Biology & Biochemistry* 8: 209–213.
- Johnson AE (2005) Lawes, John Bennet and Gilbert, Joseph Henry. In: *Encyclopedia of Soils in the Environment*, 328–336.
- Lawes JB, Gilbert JH, and Warington R (1881a) On the amount and composition of the rain and drainage-waters collected at Rothamsted. Part I and II. The amount and composition of rainfall. *Journal of the Royal Agricultural Society of England*, 17: 241–279.
- Lawes JB, Gilbert JH, and Warington R (1881b) On the amount and composition of the rain and drainage-waters collected at Rothamsted. Part II cont. The amount and composition of the drainage-waters from unmanured fallow land. *Journal of the Royal Agricultural Society of England*, 17: 311–350.
- Lehmann J and Kleber M (2015) The contentious nature of soil organic matter. *Nature* 528: 60–68. <https://doi.org/10.1038/nature16069>.
- Monteith JL (1965) Evaporation and environment. *Symposia of the Society for Experimental Biology* 19: 205–234.
- Mosse B (1953) Fructifications associated with mycorrhizal strawberry roots. *Nature* 171: 974.
- Mosse B (1962) *Graft-Incompatibility in Fruit Trees: With Particular Reference to Its Underlying Causes*. Commonwealth Agricultural Bureaux, Farnham Royal, Bucks, England.
- Nicolson TH and Gerdemann JW (1968) Mycorrhizal endogone species. *Mycologia* 60: 313–325. <https://doi.org/10.1080/00275514.1968.12018572>.
- Palenzuela J, Barea JM, Ferrol N, Azcón-Aguilar C, and Oehl F (2010) *Entrophospora nevadensis*, a new arbuscular mycorrhizal fungus from Sierra Nevada National Park (southeastern Spain). *Mycologia* 102(3): 624–632. <https://doi.org/10.3852/09-145>.
- Russell EJ (1915) *A student's Book on Soils and Manures*. Cambridge University Press.
- Russell EJ (1916) *Manuring for Higher Crop Production*. Cambridge University Press.
- Russell EJ and Hutchinson HB (1909) The effect of partial sterilization of soil on the production of plant food. *Journal of Agricultural Science (Cambridge)* 3: 111–144.
- Shahbaz M, Kuzyakov Y, Sanaullah M, Heitkamp F, Zelenev V, Kumar A, and Blagodatskaya E (2017) Microbial decomposition of soil organic matter is mediated by quality and quantity of crop residues: mechanisms and thresholds. *Biology and Fertility of Soils* 53(3): 287–301.
- Van Bruggen AHC, He MM, Zelenev VV, Semenov AM, Semenova EV, Khodzhaeva AK, Kuznetsov AM, and Semenov MV (2017) Relationships between greenhouse gas emissions and cultivable bacterial populations in conventional, organic and long-term grass plots as affected by environmental variables and disturbances. *Soil Biology and Biochemistry* 114: 145–159.
- van Helmont JB (1648) *Ortus Medicinæ*. Id est initia physicae inaudita. In: van Helmont FM (ed.) *Progressus Medicinæ Novus, in Morborum Ulitionem, ad Vitam Longam*, p. 939. Amsterdam: Elsevirium.
- Zelenev VV, van Bruggen AHC, and Semenov AM (2000) "BACWAVE", a spatial-temporal model for traveling waves of bacterial populations in response to a moving carbon source in soil. *Microbial Ecology* 40: 260–272.
- Zelenev VV, van Bruggen AHC, and Semenov AM (2005) Modeling wavelike dynamics of oligotrophic and copiotrophic bacteria along wheat roots in response to nutrient input from a growing root tip. *Ecological Modelling* 188: 404–417.
- Zelenev VV, van Bruggen AHC, Leffelaar PA, Bloem J, and Semenov AM (2006) Oscillating dynamics of bacterial populations and their predators in response to fresh organic matter added to soil: The simulation model 'BACWAVE-WEB'. *Soil Biology & Biochemistry* 38: 1690–1711.

Further reading

- He MM, Ma W, Zelenev VV, Khodzhaeva AK, Kuznetsov AM, Semenov AM, Semenov VM, Blok W, and van Bruggen AHC (2017) Short-term dynamics of greenhouse gas emissions and cultivable bacterial populations in response to induced and natural disturbances in organically and conventionally managed soils. *Applied Soil Ecology* 119: 294–306.
- Narouei Khandan HA, Harmon Lapaire C, Harmon P, Olmstead J, Zelenev VV, van der Werf W, Worner SP, Senay SD, and van Bruggen AHC (2017) Potential global and regional geographic distribution of *Phomopsis vaccinii* on *Vaccinium* species projected by two species distribution models. *European Journal of Plant Pathology*. <https://doi.org/10.1007/s10658-017-1146-4>.
- van Bruggen AHC, West JS, van der Werf W, Potting RPJ, Gardi C, Koufakis I, Zelenev VV, Narouei-Khandan H, Schilder A, and Harmon P (2018) Input data needed for a risk model for the entry, establishment and spread of a pathogen (*Phomopsis vaccinii*) of blueberries and cranberries in the EU. *The Annals of Applied Biology* 172: 126–147.

INDEX

This index is in letter-by-letter order, whereby hyphens and spaces within index headings are ignored in the alphabetization, and it is arranged in set-out style, with a maximum of four levels of heading. Location references refer to the volume number, in bold, followed by the page number. Page numbers suffixed by *f*, *b*, *ge* and *t* refers to figures, boxes, glossary and tables terms respectively. *vs.* indicates a comparison. **Bold locators** refers to main articles.

A

Aberration-corrected electron microscope (AC-EM), 2:587
 Abiotic controls
 root exudates, 1:350–351
 soil organo-mineral interactions, 1:393–395
 Abiotic dissolution of minerals, 2:287–289
 Abiotic habitats, 1:408–412
 Abiotic potentials, 3:4
 Abiotic stress, 1:336–337, 1:443, 1:465
 Abnormal soils, 4:188
 Abrasion, 3:213–214
 Absorption, 1:553, 2:316
 Absorption-diffraction method, 2:646–648
 Abstract models, colloid transport, 5:350
 Abundance distribution models (ADM), 4:731–732, 4:731*f*
 Acari (mites), 1:77–79
 ACC deaminase, 1:423
 Accumulation of organic matter, SUITMAs, 4:241
 Accuracy assessment
 soil and elevation data for, 4:748
 soil-landscape modeling, 4:739, 4:748
 Acetoclastic methanogenesis, 2:693–694
 Acetylene-inhibition technique (AIT), 1:159
 Acetylene reduction method, 1:159
 Acid drainage, in soils, 2:2
 Acidic forest soils, cation exchange capacity (CEC), 2:39–41
 Acidification, 2:245, 3:168
 Acidity in soils
 acid drainage, 2:2
 acid precipitation, 2:3
 amelioration of, 2:6
 elemental cycling, 2:3–4
 fertilizer application, 2:4
 measurement and effects of, 2:4–6, 2:4*t*
 processes generating, 2:2–6
 proton generation and consumption processes, soil-plant system, 2:2*t*
 Acid precipitation, in soils, 2:3
 Acid soils, 1:327

Acid sulfate soils, 4:178–179
 Acoustic complexity index (ACI), 3:514
 Acrisols, 4:464, 4:465*t*
 Acrisols and alisols formation
 argilluviation/lessivage, 4:348
 clay fraction, 4:350, 4:351*f*
 climate factor, 4:345
 desilification, 4:347, 4:347*f*
 erosion, 4:349
 ferrolysis, 4:348
 hydroxy-interlayered vermiculite (HIV), 4:346
 leaching, 4:345–346
 leucinization, 4:347, 4:347*f*
 losses, 4:346–347
 melanization, 4:347, 4:347*f*
 mineralogical transformations, 4:346
 pedogenetic processes, 4:346
 properties, 4:349–351
 soil drainage control, 4:346
 surface horizons, 4:346
 sustainable use and management, 4:351–352
 transfer reactions, 4:347
 transformations, 4:347
 vegetation and fauna, 4:346
 weathering, 4:346
 xanthization, 4:349
 Acroxic andosol, 4:467*t*
 Actinides
 anthropogenic sources, 2:229–230
 chemical fractionation, 2:231
 distribution coefficient, 2:230
 electronic configuration, 2:228–229
 properties, 2:229*t*
 transfer factor (TF), 2:230
 uranium and thorium, 2:229
 Actinorhizal symbiosis, 1:273–274
 Action plans
 capacity-building and technology adoption, 3:465–466
 degraded land restoration, 3:464
 land and secure land tenure access, 3:465
 land loss reduction, 3:464
 for policymakers, 3:464*f*

Active acidity, 2:17
 Active concentration profilers, 3:230
 Active layer, ground
 and permafrost, 5:422–423
 seasonal development, 5:424
 Added value of uncertainty information, 4:680–682
 Additionality, 3:553
 Additivity of variances, 4:491
 Alpha diversity richness, 4:729
 Adjusted or actual crop evapotranspiration, 5:571
 Adsorbate, 2:368*ge*
 Adsorbed ion charge density, 2:368*ge*
 Adsorbent, 2:368*ge*
 Adsorption, 2:316, 2:336, 2:343–344, 2:368*ge*, 5:155–156
 and chemical precipitation to P removal, 1:572
 and desorption of N₂, 2:26*f*
 dissolved organic carbon (DOC), 2:422, 2:428, 2:429*f*
 DOM, 2:428
 gas, 2:25–27
 isotherm, 2:330–331, 2:330*f*
 Langmuir isotherms, 2:25*f*
 and microbial processing, 2:424
 organic contaminants biodegradation, 1:553
 organic substances, 1:389*f*
 pesticides, 2:472, 2:474
 phosphorus, 2:156–157, 2:157*f*
 and precipitation, wastewater treatments systems, 1:573
 proteins, 2:397–399, 2:398*f*
 solute transport, 5:657–659
 from solution, 2:24–25
 wastewater treatments systems, 1:573
 Adsorption and desorption of sulfate by soil, 2:195
 Adsorption-desorption reaction, 2:197
 Adsorption edge, 2:331–332, 2:331*f*
 Advanced oxidation processes (AOPs), 3:420
 Advanced thermal treatment (ATT) technologies, 3:328

- Advection, 5:653
- Advection-dispersion equation (ADE), 5:353ge, 5:354, 5:357
- adsorption equilibrium, 5:660
- concentration detection mode, 5:661f
- fractional advection-dispersion models, 5:663–664
- moment analysis, 5:660
- outflow and inflow difference, 5:659–660
- production and decay, 5:660
- solute flux, 5:659f, 5:660
- solute storage, 5:660
- two-region, 5:660–661
- Aeolian EROsion (AERO) model, 3:216
- Aeolian influx, 4:72–74, 4:73–74f
- Aeolian sediments
- role of, 4:335–336
 - sediments and soil formation in semi-arid regions, 4:337–338
 - transport, 4:337
- Aeration, 5:1, 5:14, 5:216, 5:221–222
- Aerial photography, 4:702
- Aerobic chemolithotrophic nitrification, 1:254f, 1:255
- Aerobic transformation process, 1:548–549
- Aerobic vs. anaerobic respiration, 1:371
- Aerodynamic approach, evapotranspiration, 5:503
- Aerodynamic resistance, 5:572, 5:635, 5:637f, 5:639f
- Aeroponic systems, 1:341
- Afanasyev's soil classification, 4:190
- Afforestation, 3:366
- Aggregates, 5:5
- fractionation methods, 2:72
 - soil water repellency (SWR), 5:212
 - stability, 4:372–373
- Aggregation
- biogeochemical functions, 5:14
 - biotic and abiotic feedback, 5:12
 - characterization, 5:12–13
 - dry aggregate size distribution, 5:12
 - flow and transport, 5:14
 - fractionation of organic matter in, 5:13
 - functions and interpretations, 5:13–14
 - hierarchical aggregates, 5:10
 - imaging, 5:13
 - life cycle of hierarchical, 5:10–12
 - mechanical aggregates (fragments), 5:10
 - mechanisms of soil, 5:9f
 - organic matter forms and transformation relevant to, 5:11, 5:11f
 - resistance to degradation, 5:14
 - sources of organic matter for, 5:10–11
 - strength, 5:13
 - water stable, 5:12–13
- Aging effect, 2:349–350
- Aging of nanoparticles, 2:687
- Agricultural landscapes, 3:450
- Agricultural practices and land use on C stock
- direct/indirect actions, 1:236, 1:237f
 - negative impact, 1:236
 - positive impact, 1:237–238
 - uncertain effects, 1:238
- Agricultural practices and soil organic matter (SOM) formation, 2:391
- Agricultural soils
- arsenic in, 2:203–204
 - N₂O emissions from, 3:159–160, 3:160f
- Agricultural systems, leguminous (N-fixing) trees in, 1:477
- Agricultural terracing, 3:451
- geographic distribution of, 3:452
 - traditional slope management, 3:452
- Agricultural waste management
- programs and acts, 3:462–463
 - waste generation, 3:462, 3:462f
- Agriculture, 4:58–60, 4:261
- Agri-environmental payments, 3:553
- Agro-ecology (AE), 1:1ge, 3:16
- Agroforestry, 1:473–478, 1:474t, 1:476f, 3:121, 3:140
- biodiversity conservation, 3:79
 - carbon sequestration, 3:429
 - carbon sequestration and climate change mitigation, 3:79–81, 3:80t, 3:80f
 - carbon sequestration in riparian buffer systems (RBS) in southern Ontario, Canada, 3:83
 - carbon stocks, 3:84f
 - Coffea* sp. with *Erythrina* in Costa Rica, 3:76f
 - concepts of, 3:69–70
 - distribution and area of, 3:71–72
 - and ecosystem services, 3:78–81
 - emergence of, 3:69
 - global distribution and area of, 3:73t
 - homegardens of Jamaica, 3:75f
 - intercropping, 3:72–74, 3:73–74f
 - internal and external influence on, 3:86f
 - as land-use approach, 3:69
 - multistrata systems of, 3:74–76
 - Parkland system of sub-Saharan Africa, 3:75f
 - practices and systems, 3:70–71
 - principles of, 3:69–70
 - protective agroforestry systems, 3:76
 - Riparian buffer, 3:77f
 - silvopasture, 3:76–78, 3:77f
 - soil improvement, 3:78–79
 - system-level carbon sequestration, 3:81–83
 - tree-based intercropping, 3:81–83, 3:82f
 - trees and shrubs in, 3:69–70
 - tree woodlots and specialty crops, 3:78
 - in tropics and the temperate regions, 3:71t
 - water quality enhancement and nutrient interception, 3:79
- Agrogeological approach, soil classification, 4:187–188
- Agronomy, 4:182–183
- practices in, 1:448, 1:454, 1:467
 - water use efficiency in, 5:529
- Air capacity, green roofs, 3:480f
- Air, environmental components, 3:10
- Air filters and regulators of soil, 3:7
- Air-non-aqueous phase liquid-water-soil system
- hysteresis, 5:279
 - interfacial tension, 5:277
 - oil-water, air-oil, and air-water capillary pressure, 5:279
 - scaling parameters, 5:279
 - spreading phenomena, 5:277–278
 - vertical equilibrium, 5:278, 5:278f
 - water relative permeability, 5:281–282
- Air pollution, exposure to, 3:224–226
- Air quality, 3:213, 3:216, 3:486
- Air relative permeability, of air-NAPL-water system, 5:281
- Air-water interface (AWI), colloid retention, 5:338
- colloid mobilization mechanisms, 5:346
 - steady-state regime, 5:345, 5:346f
 - transient flow regime, 5:345, 5:346f
- Akaganéite, 4:141
- Albedo, 5:395–397, 5:396t
- Albolls, 4:229
- ALEX model, 5:629–630
- Alfisols, 4:55ge, 4:229, 4:296, 4:327
- Alkali feldspars, 4:97
- Alkali metals, 2:116–117
- Alkaline earth metals, 2:126, 2:214–215
- atmospheric deposition, 2:129
 - Ca and Mg soil amendments, 2:132
 - climate, parent material, and time control on, 2:130–131
 - as environmental tracers, 2:133
 - partitioning and measurement of, 2:129–130
 - plant requirements, uptake, and deficiencies, 2:131–132
 - rocks and minerals, 2:128–129, 2:128–129t
 - speciation, dissolution, and precipitation, 2:127–128, 2:127f
- Alkaline soils, 1:506, 2:693
- Allelopathic, 3:439ge
- Alley cropping, 3:126–127
- Allophane, 2:267f, 4:119, 4:121, 4:122f, 4:458ge. *See also* Short-range ordered (SRO) aluminosilicates
- Alluvial deposits, 4:35
- Alluvial soils
- amount of, 4:420–421
 - classification, 4:422
 - human uses of, 4:422–423
 - older, 4:421–422
- Alluvial terrace, 4:420, 4:420f
- Alluvium
- floodplains and, 4:420
 - Holocene Green Canyon alluvial fan cut, 4:418–419f
 - landforms of, 4:419–420
 - mineralogical composition of, 4:418
 - particle-size distribution of, 4:418
- Alpha (α) radiation, 2:257
- 1:1 Al silicates, 2:152ge
- Alu-andic, 4:458ge
- Aluminosilicate, 2:265–266, 2:266f, 2:268, 2:270, 2:273, 2:642–643
- Aluminum (Al), 2:132, 2:215
- characteristics, 2:168t
 - concentration in soil solution, 2:172–174
 - dissolution constants, 2:173t
 - elemental properties, 2:168–169, 2:169f
 - exposure vs. Alzheimer's disease, 3:227
 - extractants, 2:168t

- with inorganic ligands in solution phase of soils, 2:169–171
 octahedral, 2:168–170, 2:170f
 with organic ligands in solution and solid phases of soils, 2:171–172, 2:171t, 2:172f
 representative constituents, 2:168t
 solubility of, 2:168
 speciation in soil environments, 2:174–175
 species in solid phase of soils, 2:169
 tetrahedral, 2:168, 2:170f, 2:174–175
 Aluminum hydroxides, 4:119, 4:145
 Aluminum oxides, 4:144t
 adsorption, 4:144
 boehmite, 4:144–145
 gibbsite, 4:144
 nordstrandite and doyleite, 4:144–145
 oxyhydroxide diaspore, 4:144–145
 Aluminum tridecamer, 2:170, 2:170f
 Amelioration, of acidity in soils, 2:6
 Aminization, 2:142ge
 Amitrole, 5:332
 Ammonia (NH₃), 2:143, 2:145, 3:371, 3:417
 aerobic oxidation of, 1:262–263
 anaerobic oxidation of, 1:263
 dissimilatory nitrate reduction to, 1:267
 removal from wastewater, 1:572
 Ammonia oxidizers, 1:28
 Ammonia-oxidizing bacteria (AOB), 1:262
 Ammonification, 1:158, 2:142ge, 3:369ge
 Ammonium, 1:7–8, 2:145
 Ammonium acetate, metal extraction, 2:52
 Amorphous aluminum hydroxide (AAH), 2:347–348
 Amorphous silica, 2:267f
 Amphiboles, 4:92–93, 4:103–107
 Amphids, 1:105, 1:106f
 Amplicon-based sequencing, 1:152–153, 1:152f, 1:164
 Anabolic reactions, 1:234
 Anaerobic digestion, 1:561, 3:328
 Anaerobic soils
 agronomic significance, 1:385–386
 artificial field drainage, 4:411–412
 biogeochemical processes, 1:382–385
 biological characteristics, 1:380–381
 chemical characteristics, 1:381–382
 classification, 4:413–415, 4:414t
 climate, 4:399
 ecological and environmental significance, 1:385–386
 FAO/IIUSS WRB system, 4:414–415
 geography, 4:415
 local and regional groundwater bodies, 4:404–406
 microbial populations, 1:380–381
 morphology, 4:401–411, 4:403f
 oxygen supply, 1:379
 parent material, 4:400
 physical characteristics, 1:380
 relief factor, 4:399–400, 4:401f
 seepage, 4:406
 soil anaerobiosis, 1:379
 soil forming factors, 4:399–400
 soil water logging regime, 4:401–402f
 superficial saturation, 4:406–411
 tidal saltmarshes and mangrove swamps, 4:406
 upland environments, 1:380
 USDA soil taxonomy, 4:413–414
 water saturation, 1:379
 wetland environments, 1:379–380
 Anaerobic transformation process, 1:549–550
 Analysis of variance, 4:494–496, 4:494t
 Anammox, 1:262, 1:255–256, 2:148
 Ancient agriculture
 climate change, 3:33–35
 sustainability and degradation, 3:32–33
 Andic properties, 4:458ge
 Andisols, 4:226–227, 4:328, 4:458ge, 4:464
 Andosol, 4:458ge, 4:461f
 extreme erosion of, 4:468f
 of non-volcanic origin, 4:464–469
 of volcanic origin, 4:464
 Anecic worms, 1:115
 Angular pores, capillarity, 5:231–232
 Anion exchange capacity (AEC)
 biochar, 2:411
 organic soils, 4:390
 Anionic clays, 2:307
 Anisotropy, 4:661–662, 4:662f
 air and water flow, 5:295
 air-filled porosity, 5:296
 gas diffusivity and air-permeability, 5:295–296
 heterogenous structured materials, 5:301
 saturated and unsaturated hydraulic conductivity, 5:297–301, 5:300f
 structured soils, 5:301–304
 swelling and shrinkage, 5:295
 water flow vectors, 5:296f
 Anisotropy ratios (AR), 5:296, 5:301
 Annuals, 3:442
 Anorthite, 2:267f
 Antagonistic microorganisms, 1:443
 Antarctic Cold Desert, 4:274–275
 Antarctic Polar Desert, 4:273
 Anthric saturation, 4:413
 Anthropized area soil
 anthropization, 3:44, 3:45f
 characteristics, 3:41–44
 cities, 3:37–38, 3:38f
 classification, 3:44, 3:46–47f
 composition, 3:41, 3:42f
 contaminants, 3:43–44, 3:43f
 diversity of soils, 3:44
 ecosystem services, 3:37
 features, 3:39
 human activities, 3:37
 industrial soils, 3:40
 military soils, 3:40
 mining soils, 3:40
 (pseudo)natural soils, 3:40
 organic matter, 3:42
 pH, 3:42, 3:43f
 soil construction, 3:39
 soil formation, 3:40–41, 3:41f
 soil sealing, 3:39
 soils under traffic, 3:40
 technogenic material, 3:39
 urban soils, 3:38–40, 3:38f, 3:44, 3:46f
 Anthropocene, 4:55
 Anthropogenic activities, 2:237, 3:290
 Anthropogenic organic compounds (AOCs), 2:355–356
 functional groups, 2:364
 molecular sizes, 2:364
 Anthropogenic soil change, 4:57t, 4:58, 4:67–68
 Anthrosols, 3:24–26, 3:44, 3:46f
 Antibiosis, 1:444
 Antibiotic resistance genes (ARGs), 1:124, 3:569
 Antibiotics, 1:423
 human health and soil, 3:567–569
 Antimony (Sb), 2:215
 biological function, 2:218
 concentration in soil, 2:219t
 contamination of soil, 2:219
 Antitumor and antimicrobial compounds, 1:508
 Ants and termites in tropical and arid regions, bioturbation, 5:108–109
 Apatite, 2:227
Aporcelaimellus obscurioides, 1:108f
 Aqualfs, 4:229
 Aquands, 4:226
 Aquatic plastic toxicology, 3:314
 Aquents, 4:230
 Aqueous film-forming foams (AFFFs), 3:298
 Aqueous media, iron (Fe), 2:183–184
 Aquepts, 4:229
 Aquerts, 4:227
 Aquic soil moisture condition, 4:356
 Aquifers, 5:366
 Aquods, 4:227
 Aquolls, 4:229
 Aquoxs, 4:227
 Aquults, 4:229
 Arable soils, 4:187
 Arachnida, 1:77–87
 Araneae (spiders), 1:79–84
 Araneomorphae, 1:81, 1:83f
 Arbuscular fungi, 2:429–430
 Arbuscular mycorrhizal fungi (AMF), 1:429, 5:548–549
 prerequisites for managing, 1:464–467
 soil organic matter (SOM) decomposition, 2:390
 tillage regimes, 1:457t
 Archaea, 1:264, 2:422
 adaptation, 1:26
 ammonia oxidizers, 1:28
 archaeal cell, 1:26, 1:26t
 biotechnological potential, 1:28–29
 ecology and distribution, 1:25–26
 edaphic and environmental factors, 1:26
 energy metabolism, 1:26
 inoculation of plant growth promoting, 1:452–453
 methanogens, 1:27–28, 1:27t
 taxonomy and ecology, 1:27, 1:27f
 tillage regimes, 1:457t
 Archaeal ammonia oxidizers, 1:254–256

- Archaeognatha
diversity, 1:92
food and habitat, 1:86f, 1:92
interactions, 1:93
Archaeology, 3:20, 4:183
absolute method, 3:23
relative method, 3:23
Archives of human–nature relations, 4:234.
See also Soils of Urban, Industrial,
Traffic, Mining and Military Areas
(SUITMAs)
Arctic permafrost
borehole measurements, 5:410
characteristics, 5:411–414
gas transport, 5:415–416
hydrology, 5:414–415
Arctic urban soil microbiome (Apatity,
Russia), 1:487–490, 1:490f
Area–species relationship (SPARs), 4:730
Arenosols, 2:423
Argids, 4:229
Argillic horizon, 4:54ge
Argilloturbation, 4:181
Argilluviation/lessivage, 4:348
Aridisols, 4:227–229, 4:327
Aridity/humidity index (AHI), 4:16f
Arid landscapes, 3:582
Arid regions
characteristics, 4:333
climate categories, 4:333f
processes of soil formation, 4:334–338
and semi-arid climates, 4:333t
Aromatic sulfonates, 1:306
Arsenic (As), 2:215, 2:506, 3:289t,
3:293–294
in agricultural soil, 2:203–204
clay mineralogy, 2:205
concentration, 2:218t
contamination of soil, 2:215–216
distribution of, 2:202–203
geochemistry of, 2:202–203
interaction with soil nutrients, 2:205
microorganisms, 2:204–205
nitrate, 2:205
occurrence of, 2:202–203
phosphorus, 2:205
silicon, 2:205
soil organic matter, 2:204
soil pH and redox potential, 2:204
sulfur, 2:205
transformation and transport in soil-plant
system, 2:204–205
transport of, 2:204
Arthropods
Acari (mites), 1:77–79
Arachnida, 1:77–87
Araneae (spiders), 1:79–84
Archaeognatha, 1:86f, 1:92–93
basal hexapods, 1:86f, 1:87–94
camouflage, mimicry and defense,
1:99–101
class and order, 1:72, 1:72t
collection method, 1:75–76
Collembola (springtails), 1:87–89
Crustacea, 1:94, 1:95f
detrital food web (DWB), 1:71, 1:73
Diplura, 1:86f, 1:89–90
Myriapoda, 1:95–101
non-insect arthropods, 1:72, 1:73f, 1:75f
Opiliones, 1:84–86, 1:85f
Protura, 1:92
Pseudoscorpiones, 1:86–87
soil horizons, 1:72, 1:72t
Zygentoma, 1:94
Artificial drainage, 3:262
Artificial neural networks (ANNs), 4:479,
4:514
Artistic works in soil
local ecologies and communities, 3:510
materiality and visual imagery, 3:510
multisensory absorption, 3:517–518
sensory experience, 3:510
soil awareness and soil literacy, 3:509–510
soil chromology, 3:510–512
soil scenting, 3:514–515, 3:515f
sounding soil, 3:512–514
terroir, 3:515–517
Art, soil, 3:506
Arylsulfatases, 1:306–307
Asbolane, 4:143
asf gene cluster, 1:306
Ash, 2:433
Assimilation, sulfur, 1:305–307
Assimilatory enzyme, 1:306–307
Associative nitrogen-fixing interaction, 1:273
Astatine (At), 2:215, 2:235
Atmospheric deposition
biogeochemical processes, 2:239
chlorine in, 2:236
fluorine in, 2:236
iodine in, 2:241
Atmospheric nitrogen, 2:144–145
Atomic force microscopy (AFM), 2:590
application to mineral dissolution, 2:591
contact mode, 2:591
mineral-associated organo-mineral
(MAOM), 1:401
modes of operation, 2:591–597
ats gene cluster, 1:306
Attachment detachment approaches, 1:518,
1:520
Attenuated total reflectance (ATR)
spectroscopy
accessories, 2:515
advantage, 2:515
example, 2:515, 2:516f
penetration depth, 2:515
soil organic matter (SOM) analysis, 2:517,
2:517f
technique, 2:515
typical experimental configuration, 2:513f,
2:514
Australian Land Erodibility Model
(AUSLEM), 3:216
Autotrophic and heterotrophic respiration,
5:219
Autotrophic protists, 1:40
Auxiliary metabolic genes (AMGs), 1:16ge,
1:17, 1:20
Availability of sulfur in soil, 2:195–196
Available water capacity (AWC), 3:186
Awareness, 3:506–507
Axonchium micans, 1:108f
Azonal soils, 4:14–15, 4:188–189, 4:246–247
Azospirillum, 1:429
- ## B
- Bacillariophyceae, 1:65
Bacillus thuringiensis proteins, 2:403
Bacteria, 1:264, 2:422
commercialization of plant growth
promoters, 1:426
cooperative interactions, 1:36–37
diversity, 1:32
functional diversity, 1:32
functional redundancy, 1:32–33
horizontal gene transfer, 1:32–33
life strategies and survival, 1:33–36
microscale distribution, 1:33, 1:34f
phylogenetic and physiological diversity,
1:32–33
phylo-phenetic species, 1:32
physiological and ecologic classification,
1:33–34t
removal of, 1:572
soil organic matter (SOM) decomposition,
2:390
spatial patterns, 1:33
survival of, 1:514
tertiary wastewater treatment, 1:569
tillage regimes, 1:457t
ultramicrobacteria, 1:32
yield-acquisition-stress (Y-A-S) strategies,
1:35t
Bacteria-bacteria, 1:425
Bacterial and fungal oxidation, 2:251–252
Bacterial cells, 2:248
Bacterial movement, 1:520
Bacterial sorption and retention in soil
systems
cation exchange capacity (CEC), 1:518
organic waste properties and components
affecting bacterial transport, 1:519
surface charge properties of soil particles
and bacterial cells, 1:518–519
Bacterial transport in soil systems
biological processes, 1:516
impact of change in water content, 1:516f
physical processes, 1:516
physicochemical processes, 1:516
soil type (texture and structure) and
preferential flow, 1:517–518
soil water regime, 1:518
temperature impact, 1:518
transport pathways, 1:516–518
Bacteria-mycorrhizal fungi, 1:425
Bacteria survival and persistence in soil
systems
competing microorganisms, 1:515
factors affecting, 1:514f
nutrient availability, 1:515
organic wastes properties and storage
conditions, 1:515
pH, 1:515
soil carbonate content, 1:515
soil type, 1:514

- soil water status, 1:514
 temperature, 1:514–515
 type of organic waste and application
 method, 1:515
- Bacteria-trichoderma, 1:425
- Bacteriophages (viruses), 1:16ge
 CH₄ dynamics, 1:22–23
 climate change, 1:22f
 CO₂ release, 1:22–23
 diversity, 1:17–18
 DNA stable isotope probing (SIP)
 technique, 1:18
 life strategies, 1:17–18
 lysogenic infections, 1:18
 metagenomic method, 1:18
 single-virus genomics (SVG) method,
 1:19–20
 soil pH, 1:22
 temperature change, 1:22
 terrestrial ecosystem, 1:20–21
 virome method, 1:18–19
- Balanced sampling, 4:542–543, 4:543t,
 4:543f
- Baltimore Harbor
 habitat for minnows and stabilization,
 3:494f
 multi-use natural grass and sport field
 system, 3:494f
- Barea, José-Miguel, 5:688–689
- Bare soil, 3:443
- Barium (Ba), 2:132–133
 biological function, 2:220
 concentration of soil, 2:221t
 contamination of soil, 2:220
- Basal crop coefficient, 5:574, 5:581
- Basal hexapods, 1:86f, 1:87
 Archaeognatha, 1:86f, 1:92–93
 Collembola (springtails), 1:87–89
 Diplura, 1:86f, 1:89–90
 Protura, 1:92
 Zygentoma, 1:94
- Basal respiration, 1:376–377
- Base-cation saturation ratio (BCSR), 1:1ge
- Baseflow, 4:751ge
- Base saturation (BS), 2:39
- Basic cation saturation ratio (BCSR), 2:39
- Basin irrigation system, 5:612–613, 5:612f
- Bayesian approach
 model-based optimization, 4:545
 sample size determination, 4:544–545
- Becquerel (Bq), 2:97
- Bench terrace, 3:202–203, 3:203f
- Beneficial elements, 1:312
- Beneficial water use fraction (BWUF), 5:572
- Benzene, 2:426–428
- Benzenepolycarboxylic acids (BPCAs), 2:441
- Beryllium (Be), 2:132–133, 2:215
- β diversity index, 4:732
- β diversity richness, 4:729
- Beta (β) radiation, 2:257
- Bicarbonate, in soil, 2:263–265
- Bidentate, 2:328
- Biennials, 3:442
- Bioaugmentation, 1:538
- Bioavailability
 antimony (Sb), 2:219
 arsenic (As), 2:216
 barium (Ba), 2:220–221
 lead (Pb), 2:220
 of phosphorus, 2:269–270
 of radionuclides, 2:258–259, 2:258f
 selenium (Se), 2:198, 2:218
 of silicon, 2:271
- Biochar (BC), 2:433, 3:414, 3:484
 applications, 3:416–420
 ash, 2:408
 biological wastes composting, 2:418
 bulk density, 2:411–412
 carbon sequestration, 3:428
 contaminated soil remediation, 2:417–418
 crop yield, 2:415–416
 elemental composition, 2:409t
 fertilizer benefits, 2:414
 hydrogen, 2:408–410
 long-term carbon storage, 2:413–414
 major and micro-nutrient elements, 2:410t
 nitrogen and oxygen, 2:408–410
 organic matter preservation, 2:406–407
 pore space, 2:411–412
 production, 2:407
 properties, 2:407–412
 soil microbial community, 2:416
 soil physical properties, 2:414–415, 2:414t
 soluble salts, 2:407–408
 standardization and quality assurance,
 2:416–417
 surface charge characteristics, 2:411
 total carbon, 2:408–410
 typical properties of, 3:415–416
 unintended consequences, 3:420–422,
 3:421f
- Biochemical oxygen demand (BOD), of
 sewage, 1:571
- Biocrusts, 1:61
 assisted recovery, 1:136
 carbon fixation, 1:133
 climate change, 1:135
 composition and global biodiversity,
 1:131–132
 definition of, 1:128
 functional properties and roles in
 ecosystems, 1:132–134
 global distribution, 1:129–130,
 1:129–130f
 hydrological functions, 1:134
 interactions with biota, 1:134
 management considerations, 1:136–137
 multifunctionality, 1:134
 natural recovery from disturbance, 1:136
 nitrogen fixation and nutrient cycling,
 1:133
 physical disturbance, 1:135
 recognition of, 1:128
 soil surface albedo and heat storage, 1:134
 soil surface stability, 1:133
 temperature, 1:133
 threats and recovery, 1:134–136
 tolerance to stresses, 1:132–133
 ultraviolet radiation, 1:132–133
- Biodegradable (BIO) mulches, 3:313
- Biodiversity, 1:472, 1:530, 3:149–150,
 3:190–191, 3:254
 above-ground, 1:472
 and abundance, 1:10–12, 1:11t
 bioinformatics, 1:12
 conservation, agroforestry, 3:79
 dispersal mechanisms, 1:13
 distribution, 1:12–13
 ecological corridors for, 1:473
 ecosystem functioning, 1:13–14
 ecosystem processes, 1:10
 global change and impacts, 1:14–15
 green roofs, 3:486
 invertebrates, 1:10–12
 mites and nematodes, 1:12
 natural habitat role in, 1:473
 preservation, 4:728
 relations between pedodiversity and, 4:728
 restoration of, 1:478
 stable isotope technology, 1:12
 sustaining, 1:15
 vertebrates, 1:10–12
- Biofilm filtration, 1:569
- Bio-fortification, 3:390
- Biogenic methane, 2:693
- Biogeochemical cycle, 4:84–85, 4:757
 of arsenic (As), 2:216
 of C and Si, 2:265, 2:272–273
 of chlorine, 2:238
- Biogeochemical element cycles, 1:20
- Biogeochemical process
 anaerobic soils, 1:382–385
 soil water content, 5:472
- Biogeography, 1:42–43
- Biological activity
 green roofs, 3:482
 soil degradation, 3:169–170
- Biological availability, of sulfur, 1:304, 1:305f
- Biological components of soil, 1:448–449
- Biological control agents, 1:443–444
- Biological crusts, 5:57
- Biological mechanisms, wastewater
 treatments systems, 1:573
- Biological nitrogen fixation (BNF),
 1:271–272, 2:145
- Biological soil degradation, rangelands,
 3:99–100
- Biological transformations, of sulfur, 1:303,
 1:304f
- Biomarker-based technique, 1:148
- Biomass production, 3:146–147
- Biomethylation, 2:245–246
- Biophysical modeling, granular transport by
 ants and termites, 5:112–113
- Biopolymers, 1:508
- Bioreactors, 3:263
- Biorecovery, 2:254
- Bioremediation, 2:253–254
 extremophiles, 1:508
 metals and organic contaminants,
 3:336–337, 3:337f
 microbial inoculants, 1:431–432
- Biosolids, 3:291
 municipal solid waste, 3:330–331
 municipal wastewater treatment plants,
 1:569
- Biosorption, 2:248, 2:248f
- Biostimulation, 1:537, 1:537f

- Biota, 4:10, 4:10f
- Biotic controls
- root exudates, 1:351
 - soil organo-mineral interactions, 1:393–395
- Biotic degradation, 2:475–476
- Biotic dissolution of minerals, 2:289–291
- Biotic habitats, 1:412–416
- Biotic potentials, 3:4
- Biotic stress, 1:466
- and root mucilage, 1:336t, 1:337
- Biotransformation processes, 1:234, 1:235f
- Biotreatability studies, 1:542
- Bioturbating activities, earthworms, 1:117
- Bioturbation, 2:434, 4:180
- by ants and termites in tropical and arid regions, 5:108–109
 - definition, 5:100
 - by earthworm burrowing, 5:103–104
 - habitat formation and nesting activity, 5:107–108
 - mechanical processes, 5:101–102
 - models for, 5:104–105
 - by penetrating plant roots, 5:103
 - primary processes, 5:102f
 - rodent excavation and burrowing activity, 5:106–107
 - root-soil entanglement, 5:104f
 - soil, 5:100–101
 - soil physical constraints to, 5:102
 - termite mounds physical functioning, 5:109–112, 5:111f
- Birnessite, 4:142
- Bismuth (Bi), 2:215
- Black carbon, 2:65–66. *See also* Pyrogenic carbon (PyC)
- functional groups of, 2:363, 2:363f
 - pyrolysis temperatures of, 2:362–363
 - structures of, 2:361–362, 2:362f
- Black nitrogen (BN), 2:439
- Black plastic mulch, 3:396
- Black soldier fly composting, 1:528–529
- Blocking, colloid transport, 5:349
- Blue and Green water, 5:485, 5:485f
- Bocage landscape, northern France, 1:475f
- Boehmite, 4:144–145
- Bogs, 4:384f, 4:392
- Bond number, 5:230
- Bootstrapping, 4:513
- Border irrigation system, 5:612–613, 5:612f
- Boreal and subarctic regions
- classification and carbon stocks, 4:267
 - climate, 4:264
 - vegetation, 4:264
- Boreal forest soils
- classification, 4:280–285
 - and climate change, 4:286–287
 - cryosols, 4:285
 - distribution, 4:280–285, 4:281f
 - genesis, 4:280–285
 - gleysols, 4:284, 4:284f
 - histosols, 4:284, 4:285f
 - land use and disturbances in, 4:287
 - peatland landscape, 4:285f
 - plant cover in, 4:281–282f
 - podzols, 4:283, 4:283–284f
 - properties, 4:280–285
 - retisols, 4:282
 - turbic cryosol profile, 4:286f
- Boreal forest systems, 4:30–31
- Born–Oppenheimer (BO) approximation, 2:486, 2:488
- Boron (B), 2:215
- Bottom boundary conditions, 5:674
- Boundary condition, 4:739
- Bragg equation, 2:643–644
- Brazilian soil classification system, 4:427
- Breakthrough curve (BTC), 1:518
- Broad spectrum herbicide, 3:439ge
- Bromine (Br), 2:214–215, 2:235
- in soils, 2:236, 2:239
- Brownian motion, 5:248
- Brunauer–Emmett–Teller (BET) equation, 2:25f, 2:26, 2:441
- Bryophytes, 1:128
- Bubblers, 5:619
- Buckingham–Darcy equation, 5:386
- Buffering, 3:255
- buffer capacity, 2:15
 - buffers and filters, 3:262
 - capacity factor, 2:15
 - cation exchange reactions and pH buffering, 2:18
 - concept, 2:15
 - dissolution and precipitation, 2:18
 - ecological theory, 2:15–16
 - of ion activities, 2:20–21
 - lime requirement, 2:17–18
 - pH-dependent surface charge and pH buffering, 2:19, 2:19t
 - samovar model, 2:16–17, 2:16–17f
 - in soil acidity, 2:17–18
 - of soil oxidation-reduction status, 2:19–20, 2:20t
- Buffering capacity
- green roofs, 3:481
 - of soils, 2:273
- Bulk density, 1:5–6, 3:480f
- Bundle of cylindrical capillaries (BCC) model, 5:188–189, 5:189f
- Buried soils, 4:214
- Burrowing activity, bioturbation
- earthworm, 5:103–104
 - rodent excavation and, 5:106–107
- Burst respiration, 1:377
- Buserite, 4:142
- Bush encroachment, 3:97f
- ## C
- Cadmium (Cd), 2:5, 3:293–294
- Calcids, 4:229
- Calcisols, 2:264–265
- Calcite, 2:281
- Calcium (Ca), 2:132, 2:264
- Calcium carbonate (CaCO₃), 4:48ge, 4:178
- accumulation and time effect, 4:49
 - in soil, 2:263–265
- Calcium chloride, metal extraction, 2:52
- Calcium deficiency, 2:132
- Calcium oxalate, 2:251
- Calibration, 4:587–588, 4:619, 4:748–749
- Calorimetric analysis
- differential scanning calorimetry (DSC), 2:667–668
 - heat changes, 2:662
 - isothermal microcalorimetry (IMC), 2:663–665
 - isothermal titration calorimetry (ITC), 2:665–667
 - microcalorimetry, 2:663
 - soil and environmental science, 2:663f
- Cambids, 4:229
- Cambisol, 5:141f
- Camden Yards Sports Complex in the City of Baltimore, Maryland, 3:491, 3:492f
- Camouflage, mimicry and defense
- chemical defense, 1:99–101, 1:100f
 - debris, 1:99, 1:100f
 - mimicry, 1:101, 1:101f
- Canonical denitrification, 1:264–266
- Capacitance probes, water content, 5:173
- Capacitance sensor
- soil moisture, 4:593
 - soil temperature on, 5:173
- Capillarity
- angular pores, 5:231–232
 - capillary rise model, 5:230
 - contact angle, 5:228
 - and curved surfaces, 5:229
 - dynamic aspects, 5:233–235
 - heterogeneous surfaces, 5:236–237
 - historical perspective, 5:226, 5:226f
 - microscale hysteresis, 5:238
 - porous media, 5:226
 - and soil cohesion, 5:237–238
 - in soils, 5:230–231
 - surface tension, 5:226–228
- Capillary flow paths, continuity of, 5:493–494
- Capillary forces, 5:233
- vs.* gravity, 5:494
- Capillary osmosis, 5:267
- Capillary pressures, in fingertip, 5:318f
- Capillary rise method (CRM), 5:207–208, 5:230, 5:233–235
- Capillary *vs.* viscous flow, 5:494–495
- Capping, 3:334, 3:335f
- Carbon (C), 2:215
- availability, organic, 1:268
 - boreal forest soils and climate change, 4:286–287
 - dynamics and greenhouse gases, 3:61–62
 - and energy source, 1:267
 - fixation, biocrusts, 1:133
 - in forest soils, 3:62
 - inorganic forms of, 2:266f, 2:273
 - movement in Cryosols under influence of land use change, 4:286f
 - phosphorylation of, 1:267
 - properties, 2:263
 - sequestration, 2:124, 4:381
- Carbon and nitrogen metabolism, extreme soil microorganisms, 1:500, 1:502t
- Carbonates, 2:251
- alkalinity, 2:265
 - pedogenic, 2:268
 - in soil, 2:263–265

- Carbon calibration dependency, 2:634, 2:634f
- Carbon cycle
- agricultural practices, 1:226t
 - basic, 1:219
 - C pools, 1:219–220
 - DOC in global, 2:430
 - fluxes, 1:220
 - functions, 1:221–222
 - soil algae, 1:67
 - soil-forming factors, 1:220–221
 - soil organic carbon dynamics, 1:223
 - terrestrial, 1:220f
 - vs.* water use efficiency, 5:530
- Carbon dioxide (CO₂), 2:263
- emissions of, 2:692–693, 3:62
- Carbon farming, 3:431
- Carbon fluxes, 1:220
- Carbon isotopes
- labelling approaches to study SOM in cropped soils, 2:87–88
 - natural ¹³C abundance, 2:87
- Carbon metabolism, symbiotic nitrogen fixation, 1:276
- Carbon sequestration, 1:229, 3:148–149, 3:189–190, 3:428
- carbon inputs, 3:426
 - carbon outputs, 3:426–427
 - carbon saturation, 3:427
 - climate change mitigation, 3:429–430
 - and climate change mitigation, 3:79–81, 3:80t, 3:80f
 - crop rotations, 3:128–129
 - potential and constraints, 3:425
 - in riparian buffer systems (RBS) in southern Ontario, Canada, 3:83
 - soil management practices, 3:428–429
 - soil organic matter turnover, 3:425–426
 - soil water repellency (SWR), 5:212
 - strategies, 3:427–428
 - system-level, 3:81–83
- Carbon–silicon interactions, 2:265
- Carbon source for rhizobiome, 1:339
- Carbon stabilization, 1:222, 1:222f
- Carbon stocks
- agroforestry, 3:84f
 - alpine and plateau soils, 4:275
 - Antarctic Cold Desert, 4:274
 - Antarctic Polar Desert, 4:273
 - boreal and subarctic regions, 4:267
 - high Arctic soils, 4:270
 - low Arctic soils, 4:268
- Carbon storage, 3:49, 3:49–50f
- agricultural practices and land use, 1:236–238
 - biochar, 2:413–414
 - carbon age distribution, 1:233, 1:233f
 - carbon inputs to soil, 1:233–234
 - C content and soil density, 1:229–230
 - CO₂ certificates, 1:240
 - C sequestration, 1:240
 - C stock calculation, 1:230–231
 - definition, 1:229, 1:232
 - dry bulk density of soil/apparent density, 1:230
 - ecosystem services and United Nations Sustainable Development Goals (SDG), 1:239, 1:239f
 - greenhouse gas emission, 1:240
 - inorganic C, 1:229
 - mineral-organic associations, 1:235–236
 - organic inputs and soil food chain, 1:234
 - organic matter biotransformation, 1:234, 1:235f
 - organic matter in soil, 1:229, 1:230f
 - photosynthesis, 1:229
 - precautions during estimation, 1:232–233
 - from societal viewpoint, 1:238–240
 - soil and agronomic research, 1:239–240
 - soil fertility and biodiversity, 1:238–239
 - soil organic C, 1:229
 - theoretical C stock trends, 1:232, 1:232f
 - total soil C content (TOC), 1:229–230, 1:231f
- Carbon turnover models, 2:65–66
- Carbon use efficiency (CUE), 2:429
- Catabolic reactions, 1:234
- Catch crops, 3:446–447
- Catchments to global analyses, 5:497–498, 5:498f
- Cation exchange capacity (CEC), 1:314–315, 2:120–121, 2:120f, 4:498
- acidic forest soils, 2:39–41
 - acidic soils, 2:33
 - alkaline soils, 2:33
 - base saturation (BS), 2:39
 - basic cation saturation ratio (BCSR), 2:39
 - biochar, 2:411, 2:411f
 - cation exchange equations, 2:35–36
 - charge sources, 2:33–34
 - concept and models, 2:34–35
 - effective, 2:38
 - exchangeable and non-exchangeable aluminum, 2:37–38
 - green roofs, 3:481
 - in highly weathered tropical soils, 2:41
 - hydrogen ions, 2:38
 - organic soils, 4:390
 - permanent negatively charged mineral surfaces, 2:33–34
 - soil organic matter functional groups, 2:34
 - total, 2:38
 - variably charged mineral surfaces, 2:34
 - variably charged soils, 2:37
- Cation exchange effect, 4:378
- Cell quota model, 1:173, 1:173f, 1:176
- Cellulose, 1:524, 2:389
- Cellulose absorption index (CAI), 4:639
- Cellulose and hemicellulose, 2:438
- Cell wall degrading enzymes, 1:424
- Center pivots, 5:615
- CENTURY ecosystem model, 1:208–209, 1:209f
- Cereal-legume rotation, 3:126–127
- Cereals, legume symbiosis engineering, 1:280
- CERES Maize model, 5:628
- Cerium (Ce), 2:214–215
- Cesium, 2:117, 2:260–261
- Chalcogens, 2:214–215
- Chalcophanite, 4:142
- Change of support
- concept, 4:522–523
 - point variable, 4:525–526
 - punctual kriging, 4:525
 - regularization and de-regularization, 4:526, 4:526f
 - rescaling/upscaling procedure, 4:525
 - uncertainty assessment, 4:679
 - variance of block, 4:526
- Chao1 richness index, 1:541
- Charge-assisted H-bond (CAHB), 2:317
- Charring intensity, 2:434
- Check dam, 3:117, 3:117f
- Chemical amelioration, 4:378–379
- Chemical amendments, 4:378
- Chemical and enzyme kinetics, 1:174–175, 1:174t
- Chemical control, soil-borne plant disease control, 1:440
- Chemical crop protection, 1:463–464
- Chemical enhancement (CE) theory, 2:499–500
- Chemical equilibrium
- calcite, 2:281
 - complex cement materials, 2:282
 - energy barrier, 2:278
 - equilibrium, 2:278
 - free-energy definition, 2:275–276
 - geochemical models, 2:279, 2:281–282
 - kinetic definition, 2:275
 - lead and phosphate interactions, contaminated soils, 2:280
 - and Le Chatelier's principle, 2:276
 - manganese(III) phosphate, 2:281
 - model inputs, 2:279
 - organic compounds, 2:278
 - reaction orders and half-lives, 2:277t
 - solid-phase dissolution and precipitation, 2:277–278
 - system variables, 2:279
 - thermodynamic data/equilibrium constants, 2:278
- Chemical force microscopy (CFM), 2:600–603, 2:603f
- Chemical-free physical fractionations, 2:77–78
- Chemical index of alteration (CIA), 2:109–110, 2:109f
- Chemical kinetics, 2:353
- Chemical leaching, 4:78f
- Chemical reactions, wastewater treatments systems, 1:573
- Chemical reduction, 3:339, 3:340f
- Chemical requirements, of green roofs, 3:481–482
- Chemical shift, 2:618, 2:619f
- Chemical soil degradation, 3:99, 3:168–169
- Chemical stability, green roofs, 3:481–482
- Chemical stabilization, 3:336
- Chemical weathering, 2:127
- Chemisorption, 2:316
- electron paramagnetic resonance (EPR) studies, 2:530
 - persistent organic pollutants (POPs), 3:302
- Chemoattractant, 1:359–360, 1:359f, 1:360t
- Chemodenitrification, 2:147–148

- Chernozem soils, 4:190
 Chilopoda
 diversity, 1:96–97
 food, 1:97
 habitat, 1:97
 interactions, 1:97
 Chinese Loess Plateau
 annual suspended sediment load, 3:345*t*, 3:346
 boundary, 3:350*f*
 check dam farmland, 3:346–348, 3:347*f*
 geological erosion, 3:345–348
 low-disturbance tillage methods, 3:346
 perennial grasses and legumes, 3:348
 restoration, 3:348
 terracing, 3:346
 tree planting programs, 3:348
 wind-borne dust, 3:345
 Chitin, 1:524
Chlorella, 1:65
 Chlorinated solvents, 1:550
 Chlorine (Cl), 2:214–215, 2:235–239
 Chlorite, 4:117
Chlorobium, 1:307
 Chlorophyta, 1:65
 Chlorpyrifos (CPF), 2:504, 2:505*f*
Chromatium, 1:307
 Chromology of soil
 bio-chemical characteristics, soil, 3:512, 3:512*f*
 biodynamic farming, 3:511
 diversity of soil color, 3:510–511
 histories, vulnerabilities and potentials
 narration, 3:511
 iGEM (International Genetic Engineered Machine), 3:512, 3:512*f*
 Munsell Color System, 3:510–511
 portraits, 3:511–512, 3:511*f*
 reliability, 3:511–512
 site assessment, 3:511
 soil properties, 3:512, 3:512*f*
 Chronosequence, soil, 4:51–52, 4:52*f*
 Circular lysimeters, 5:668*f*
 Cities and industry, 4:60–63, 4:61–62*f*
 Citizen science projects, 3:576
 Citric and oxalic acids, 2:245
 Civilizations, soil management
 Africa, 3:31
 America, 3:31
 ancient agriculture, 3:31–33
 Asia and Oceania, 3:30
 Europe, 3:30–31
 locations of, 3:30*f*
 Classification and regression tree (CART) model, 4:505, 4:513
 Classification error, 4:682
 Clausius-Clapeyron equation, 5:420
 Clay dispersion, 4:371–372, 4:371–372*f*
 Clay illuviation, 4:180
 Clay mineralogy, 2:22
 Clay minerals, 2:264, 2:291, 2:337–338, 2:642–643, 3:507
 absorption of, 2:569, 2:569*t*
 allophane and imogolite, 4:119
 aluminum hydroxide minerals, 4:119
 ball-and-stick model, 4:111*f*
 carbonates, sulfates, and soluble salts, 4:120
 crystal structure, 4:110–111
 human health and soil, 3:569–570
 iron oxide minerals, 4:119
 major element composition, 4:110
 manganese oxide minerals, 4:120
 octahedral sheet, 4:112–113, 4:112–113*f*
 1:1-type minerals, 4:113–115
 polyhedral model, 4:111*f*
 soil development sequence, 4:51
 sphere-packing model, 4:111*f*
 structural concepts, 4:110–111
 surfaces, 2:493
 tetrahedra and octahedra, 4:110, 4:111*t*, 4:111*f*
 tetrahedral sheet, 4:111, 4:112*f*
 and time effect, 4:51
 titanium oxide minerals, 4:120
 2:1-type minerals, 4:115–118
 weathering and formation, 2:109–110
 zeolites, 4:119
 Clays, 1:408
 attraction forces, 2:379*f*, 2:380, 2:381*f*
 cation exchange capacity (CEC), 2:377, 2:377*t*
 charge characteristics, 2:377, 2:377*t*
 charge density, 2:380
 dispersion and hydraulic properties, 2:380, 2:384, 2:385*f*
 electrical double layer, 2:377–378, 2:378*f*
 electrostatic attraction forces, 2:380
 flocculation, 2:378, 2:379*f*
 hydration repulsion forces, 2:378
 hydraulic conductivity, 2:384, 2:385*f*
 infiltration rates, 2:384, 2:385*f*
 mechanical dispersion, 2:380
 mineralogy, 2:377, 2:377*t*
 Na/Ca clays (*see* Sodium/calcium montmorillonite)
 platelets, 2:377–378, 2:380
 repulsion forces, 2:377–378, 2:378*f*
 shape and charge, 2:377, 2:377*t*
 spontaneous dispersion, 2:380
 swelling, 4:371
 translocation, 4:324, 4:324–325*f*
 van der Waals attractive force, 2:380
 Clean Water Act, 4:425*t*
 Climate, 4:319–321, 4:320*t*, 4:333–334
 and dissolved organic carbon, 2:422–423
 soil-forming factors of, 4:334*f*
 SUITMAs, 4:239–240
 zone classifications, 4:333–334
 Climate change, 3:379, 4:8, 4:9*f*, 4:83
 alpine and plateau soils, 4:275
 anaerobic soils, 4:399
 ancient agriculture, 3:33–35
 Antarctic Cold Desert, 4:274
 Antarctic Polar Desert, 4:273
 bacteria, fungi and archaea, 1:581
 biocrusts, 1:135
 boreal and subarctic regions, 4:264
 boreal forest soils and, 4:286–287
 carbon cycle, 1:221
 carbon sequestration, 3:429–430
 changing climates, 4:22–23
 cold region soils, 4:260
 cold temperate climates, 4:20–21
 direct *vs.* indirect effects, 1:579–580
 drivers, 1:584
 drying, 2:696–697
 ecological stability, 1:580
 elevated CO₂ effects, 1:583
 evapotranspiration, 4:14
 experiments, 1:580
 extreme soil microorganisms, 1:507
 vs. fossil C emissions, 3:430
 freezing and thawing cycles, 5:407
 high Arctic soils, 4:269
 humid and sub-humid tropical climates, 4:16–17
 low Arctic soils, 4:267
 mitigation, 3:416–417
 organic contaminants biodegradation, 1:555–556
 organism interaction, 1:582–583
 organo-mineral interactions, 1:404–405
 permafrost hydrology, 5:425–426, 5:425*f*
 precipitation, 1:584, 2:697
 predictions, 4:83
 pulse and press changes, 1:580
 seasonally frozen soil hydrology, 5:425
 semi-arid and arid climates, 4:17–18
 soil fauna, 1:581–582
 and soil formation, 4:67
 soil organic matter (SOM) formation, 2:391
 and soil organism interaction, 1:579, 1:579*f*
 on soil processes, 1:583–584
 soil properties, 4:14–15
 soil structure and soil organic matter (SOM) degradation, 3:252–255
 solar radiation, 4:14
 subantarctic forest and tundra, 4:272
 sub-arctic and arctic zones, 4:21–22, 4:22*f*
 temperature, 4:14
 temperature and moisture changes, 1:578
 viruses, 1:582
 warming effects, 1:583, 2:695–696
 warm temperate climates, 4:18–20
 wildfire effects, 2:697–698, 2:698*t*
 wind, 4:14
 zones, 4:15, 4:16*f*
 Climate change mitigation, carbon sequestration and, 3:79–81, 3:80*t*, 3:80*f*
 Climate related issues, 1:531
 Climatic water content, 5:487, 5:487*f*
 Clods in cohesive soils, 5:10
 Clone library construction method, 1:150
 Clorpt approach, 4:477
 Clustering approaches, 4:516
 Cluster random sampling, 4:541–542, 4:541–542*f*
 Co-adsorption effect, 2:332–333
 Coal-based energy production, 3:293
 Coarse fragments, 4:572, 4:572*f*
 Codenitrification, 2:147
 Coefficient of linear extensibility (COLE), 5:141
 Coefficient of variation (CV), 4:491

- Coffea* sp. with *Erythrina* in Costa Rica, 3:76f
 Cognition soil maps, 4:737
 Cohesion, 5:115–116
 Co-kriging, 4:479, 4:669
 Cold region soils
 agriculture, 4:261
 Alpine and plateau soils, 4:275–278
 biological and chemical properties, 4:260
 chemical transformations, 4:260
 climate change, 4:260
 cryoturbation, 4:254–257
 distribution, 4:251–252
 engineering and development, 4:260–261
 environmental change and pressures, 4:250–251
 freezing point depression, 4:253
 genesis and classification, 4:263–278
 land surface model, 4:759
 land use, 4:260–261
 microbial communities, 1:501–504
 morphological properties, 4:253–259
 non-extractable fractions (NEF), 4:260
 northern circumpolar region, 4:251
 Northern Circumpolar Region (NCR), 4:264–270
 patterned ground, 4:259f
 permafrost, 4:250–252
 redoximorphic features, 4:257
 soil organic matter (SOM), 4:260
 soil temperature regime (STR), 4:251f
 southern circumpolar region (SCR), 4:251–252, 4:270–275
 structure, 4:253, 4:254–255f
 temperature and moisture regimes, 4:251, 4:260
 water movement, 4:253
 Cold tolerance mechanisms, of microorganisms, 1:503–504
 Collaboration gradient, 1:324
 Collagen, 1:524
 Collecting Arthropods
 aspirator, 1:76f
 entomology text, 1:75, 1:75f
 organism size and investigation type, 1:74f, 1:75
 pitfall traps, 1:75–76, 1:76f
 shaker/sieve, 1:76, 1:76f
 Tullgren funnel extraction, 1:75, 1:75–76f
 Collective goods, 3:547ge, 3:549–550
 Collembola (springtails)
 diversity, 1:87, 1:88f
 food, 1:87, 1:89–90f
 habitat, 1:89, 1:89f
 interactions, 1:89, 1:91f
 life cycle, 1:87
 soil organic matter (SOM) decomposition, 2:390
 Colligative properties of solutions, 5:153
 Colloidal probe force microscopy (CPFM), 2:598–600
 Colloid filtration theory (CFT), 5:337
 Colloid mobilization mechanisms, 5:346
 Colloid retention mechanisms
 air-water interface (AWI), 5:345–346
 deposition, 5:344
 release, 5:344–345
 saturated and unsaturated porous media, 5:343–344, 5:344f
 straining, 5:345
 Colloid transport
 batch experiments, 5:347
 biocolloids, 5:337
 colloidal particles, 5:337
 colloid retention mechanisms, 5:343–346, 5:344f
 colloid shape, 5:349
 column experiments, 5:347
 contaminant transport, 5:337
 deposition process, 5:344
 Derjaguin approach (DA), 5:340
 Derjaguin-Landau-Verwey-Overbeek (DLVO) forces, 5:337–339
 forces and interactions, 5:337–343
 modeling, 5:349–350
 nanoparticles (NP), 5:337
 non-Derjaguin-Landau-Verwey-Overbeek (DLVO) forces, 5:339–340
 pore-scale experiments, 5:347
 porous media, 5:337
 solid-water interface (SWI) and straining, 5:343–344
 surface element integration (SEI) technique, 5:340
 surface roughness and charge heterogeneity, 5:348–349
 Colluvial debris, 4:34–35
 Colored film mulch, 3:397f
 Colored plastic mulch, 3:397
 Comammox, 1:262, 1:263, 1:266f
 Combination equation
 for grass and water surfaces, 5:637–638
 Penman-Monteith equation, 5:636
 Combined Radar and Radiometer (ComRAD) simulator, 4:607
 Combustion (incineration) with energy recovery, 3:328
 Cometabolism, 1:548–549
 Common International Classification of Ecosystem Services (CICES), 3:548–549
 Communication and dissolved organic carbon, 2:428–430
 Community Bureau of Reference (BCR) procedure, 2:54–55
 Community-level physiological profiling (CLPP), 1:142, 1:483, 1:484f
 Compaction, 3:251–252, 3:252f
 aeration characteristics, 5:91
 amelioration, compacted soils, 5:98
 bulk density, porosity, and packing state, 5:91
 causes, 5:86
 in crop production, 5:92–95
 crop responses to subsoil compaction, 5:93–94, 5:94f
 on environmental components, 5:95–96, 5:96t, 5:96f
 factors and processes affecting, 5:86, 5:87f
 fertilizer application, 5:93, 5:94f
 in forestry, 5:98
 hydraulic properties, 5:91
 interactive crop responses, 5:93, 5:94f
 loading characteristics, 5:88–89
 machinery-soil interactions, 5:89
 mean ground contact pressure, 5:88
 modeling, 5:94–95, 5:95f
 occurrence, 5:86
 peak stresses, 5:88
 plant growth, crop yield and soil biota, 5:93, 5:93f
 prevention, 5:97
 problems, 5:86
 process, 5:86
 root growth and root distribution, 5:92–93, 5:92f
 under running gear, 5:87, 5:87–88f
 seed germination and establishment, 5:92
 soil compactability, 5:89–91, 5:90f
 on soil physical and mechanical properties, 5:91–92
 soil profile, 5:88–89, 5:89f
 soil strength characteristics, 5:91
 stress measurement, 5:97
 Compaction and turfgrass cultural intensity, 3:498
 Companion cropping. *See* Living mulches
 Competitive behavior, 2:331f, 2:332
 Competitor-stress tolerator-ruderal (C-S-R) concept, 1:33–34
 Complexation by ligands, 2:245
 Complexation reaction, 2:198
 Complex SPAC model, 5:629–630
 Complex substrate biodegradation assays, 1:158
 Component-wise gradient boosting algorithm, 4:512–513
 Composition, municipal solid waste, 3:323–331
 Compost
 definition of, 1:523
 supplementation of, 1:525
 Composting
 benefits to soil, 1:529
 black soldier fly, 1:528–529
 cooling phase, 1:526
 curing phase, 1:526
 definition of, 1:523
 effects on soils, 1:529–530, 1:529f
 high temperature phase, 1:526
 initiation phase, 1:525–526
 manure management, 1:561
 maturation phase, 1:526
 maturity tests, 1:526–527
 quality assessment, 1:529
 Comprehensive multiphase (CMP) NMR, 2:610, 2:620, 2:622f
 Computed tomography anisotropy (A_{CT}), 5:301
 Computed tomography, soil process
 gamma-ray tomography, 5:464
 X-ray computed tomography (XCT), 5:460–463
 Computer simulation of reality, 2:484–485, 2:484f
 Conditioned Latin hypercube sampling (cLHS), 4:547
 Cone penetrometer, 5:81–82
 Confidence, 4:490

- Coniferous forest systems, 4:30, 4:31f
- Conservation agriculture (CA), 3:118–119, 3:119f, 3:393. *See also* Mulch
- Conservation Reserve Program, 3:355–356
- Conservation tillage, 3:203, 3:205f
- Conservative behavior, 2:236
- Consistency, soil morphology, 4:161
- Contact angle (CA), 5:205, 5:205f, 5:228, 5:279
- Contact mode, AFM, 2:591
- Contaminants, 1:574–577, 3:255.
See also Metals and organic contaminants remediation
- Contamination
groundwater, 5:368–369
soil degradation, 3:170
- Continuous grazed system, rangeland management, 3:94
- Continuous time random walk (CTRW)
approach, 5:353ge, 5:360
mass transfer properties, 5:662–663
phenomenological basis, 5:662–663
space and time increments, 5:662–663
transition time distribution, 5:663
- Continuum-scale models, colloid transport, 5:350
- Contour or sloping terrace, 3:452
- Controlled drainage, 3:261–262, 3:261f
- Controlled lower boundary by suction rake lysimeters, 5:673f
- Convection, 5:221
- Convection-dispersion equation (CDE), 5:324, 5:357
- Convective lognormal transfer (CLF), 5:353ge
- Conventional free drainage, 3:260, 3:260f
- Convention on Biological Diversity (CBD), 3:537
- Convolutional neural network (CNN), 2:571
- Co-operative phenomena, 2:299, 2:299f
- Copernicus Data and Information Access Services (DIAS), 4:640
- Coping with soil salinization, 3:243
- Copper sulfate, 3:291–292
- Co-precipitation, organo-mineral interactions, 1:390
- Coregionalization
modelling, 4:666–668, 4:667f
simultaneous variation in two or more variables, 4:664–666
- Coronadite, 4:143
- Coronavirus disease 2019. *See* COVID-19
- Correlation, 4:496–497, 4:497f
- Correlation range, 4:660
- Cosmic-ray neutron probe (CRNP), soil moisture, 4:591–592, 4:595
- Cosmic-ray neutron sensing (CRNS), 4:584–585, 4:585f, 5:174, 5:427–428, 5:432–434, 5:433f, 5:475, 5:479
- Cosmic ray soil moisture observing system (COSMOS), 5:432
- Co-solutes, 2:365–366, 2:366f
- Coulombic force, 2:317
- Coupled heat and mass transfer models, 5:391
- Covariance, 4:496
- Covariate-supported geotransfer functions (CoGTFs), 5:191
- Cover crops, 1:441, 1:463, 3:119, 3:262, 3:392, 3:394–395, 3:446–447
Crotalaria, 3:395f
limitations of, 3:398–399
policy decisions, 3:411–412
runoff and soil erosion control, 3:411
soil quality, 3:411
types of, 3:395
- COVID-19 pandemic, 3:285, 3:574
- Cowardin classification, wetlands, 4:438
- Crassulacean acid metabolism (CAM)
photosynthesis, 3:582
- Cretinism, 3:561
- Criminal justice system. *See* Forensic soil science
- Cristobalite, 4:93, 4:145–146
- Critical flocculation concentration (CFC), 4:371
- Critical zone, 5:365
- Crop and pasture mixtures, 3:448
- Crop-based biofuels and nitrous oxide, 3:162
- Crop coefficients, 5:574–575, 5:574–575f
- Crop evapotranspiration, 5:571–573, 5:572–573t
- Crop health regulation, soil phages, 1:21
- Cropland soil management, 3:428–429
- Crop management-dry framing
agroforestry, 3:121
foliar sprays, 3:120
integrated farming system, 3:121
intercrops, 3:120, 3:120f
tolerant crops, 3:120–121, 3:121t
- Cropping, 1:463, 4:577
diversity, 3:126
in sodic soils, 4:379
- Crop production, tillage systems, 3:433–434
- Crop rotations, 1:225
alley cropping, 3:126–127
benefits and drawbacks, 3:128
biodiversity, 3:132
cereal-legume rotation, 3:126
direct drilling, 3:127–128
diversified organic crop rotation system, 3:138
fertility and yield, 3:129–131
forage legumes, 3:139
grassland, 3:127
historic overview, 3:125–126
integrated crop-livestock systems, 3:126–127
intercropping, 3:139
pest control, 3:131–132
social and economic aspects, 3:132–133
soil-borne plant disease control, 1:441
soil carbon sequestration, 3:128–129
systems perspective, 3:138
types, 3:127f
worldwide, 3:127–128
- Crop tolerance to sodicity, 4:377, 4:377t
- Crop water productivity, water use efficiency, 5:531–533, 5:532f, 5:532t
- Crop water requirements (CWR), 5:571
- Crop water stress index (CWSI), 4:607–608, 5:520–522
- Crop yield, 3:240–241, 3:241f
- Cross-channel terrace, 3:452, 3:452f
- Cross-correlation, uncertainty assessment, 4:677
- Crushed volcanic rock mulch, 3:398
- Crustacea, 1:94, 1:95f
- Crusts, 5:57, 5:65–66
- Cryals, 4:229
- Cryands, 4:226
- Cryepts, 4:229
- Cryerts, 4:227
- Cryids, 4:229
- Cryods, 4:227
- Cryogenic electron microscopy (cryo-EM), 2:584–586, 2:585f
- Cryolls, 4:229
- Cryo-pedogenetic process, 5:414
- Cryosols, 4:285
- Cryotic taliks, 5:424
- Cryoturbation, 5:414
boreal and subarctic soils, 4:266
definition, 4:254
ice-cemented permafrost-affected soils, 4:255f
mechanisms, 4:254–257
morphological patterns, 4:254–257
in Subarctic and Low Arctic, 4:256f
- Cry proteins, 2:403
- Cryptobiotic, 1:128
- Cryptogamic, 1:128
- Cryptomelane, 4:143
- Cryptopodzols, 4:458ge
- Crystalline phases, of soil particles, 2:581–583
- Cultural services, earthworms, 1:117
- Culturomics, 1:143, 1:144f, 1:144–145t
- Cumulative dissolution constant, aluminum species, 2:170
- Cupid model, 5:629–630
- Cuticle, foliar nutrients application
cucumber leaf, 3:385f
hydrophilic solutes penetration, 3:385–386
injuries, 3:384
leaf veins and trichomes, 3:386
lipophilic solutes penetration, 3:385
pathways, 3:384
sunflower leaves, 3:386
- Cutting method, lysimeters, 5:670, 5:670f
- Cyanobacteria
biotic interactions, 1:68
cell size, 1:63
extracellular polymeric substances (EPS)
production, 1:64
nitrogen cycle, 1:62, 1:67
oxygenic photosynthesis, 1:62–63
phosphorus cycle, 1:67–68
pigments, 1:62–63
soil moisture, 1:66
solar radiation, 1:66
- Cytokinin synthesis, 1:423

D

- Darcy-Buckingham equation, 5:244
- Darcy-Buckingham law, 4:751ge
- Darcy equation, 5:257, 5:260
- Darcy model, 5:240–241

- experiment, 5:242, 5:243f
 isotropy and scale (in)variance, 5:244
 linear flow models, 5:241–242, 5:241t
 non-Darcian flow and non-uniform flow, 5:245
 representative elementary volume (REV), 5:244–245
 saturated filter flow, 5:241f
 soil water and unsaturated cases, 5:244
 validity and limits, 5:242–243, 5:243f
 Darcy-scale, 5:242, 5:244–245
 Darcy's law, 5:240, 5:279–280, 5:381
 Data acquisition software, 2:629
 Debre Mawi watershed
 hillslope in, 5:290f
 runoff coefficients, 5:289, 5:290f
 saturation excess overland flow, 5:289
 yearly rainfall in, 5:289
 Decentralized/on-site wastewater treatments systems, 1:568
 Deciduous forest systems, 4:30, 4:31f
 Deciphering memory of past human activities, in SUITMAS, 4:235–237
 Decision making, 4:671–672, 4:682
 Decision trees, machine learning, 4:513
 Decomposition
 CENTURY model, 1:209f
 follic histosols, 4:390
 half-life, 1:210, 1:210f
 mean residence time (MRT), 1:210, 1:210f
 peat soils, 4:391–392
 of plant litter, 1:215
 of pure cellulose, 1:211–215, 1:214f
 as self-decay, 1:208–210, 1:209f
 of simple individual compounds, 1:211, 1:212–213t
 soil organic matter (SOM), 2:388–390
 turnover time, 1:210
 Dedifferentiate, 3:439ge
 Deep soil, land surface model, 4:757–758
 Degradation, 3:420
 and dissolved organic carbon, 2:428–430
 rangeland, 3:96–100
 Degradation, soil
 biological, 3:169–170
 chemical, 3:168–169
 definition, 3:165
 drivers, threats and consequences of, 3:167f
 factors controlling, 3:166–167, 3:166f
 physical, 3:167–168
 soil remediation, 3:170
 types of, 3:167, 3:168f
 Degradative technique, pyrogenic carbon (PyC), 2:435–436
 Deliquescence relative humidity (DRH), 3:387
 Delta notation, 2:83
 Delta plains, 3:266
Dendrodrilus rubidus, 1:113f
 Denitrification, 2:142ge, 2:146–147, 2:149, 2:694
 in anaerobic soils, 1:383–384
 canonical, 1:264–266
 Denitrification, microbial
 aerobic oxidation of ammonia, 1:262–263
 anaerobic oxidation of ammonia, 1:263
 anaerobic oxidation of methane, 1:264
 canonical denitrification, 1:264–266
 definition of, 1:262
 dissimilatory nitrate reduction to ammonia/nitrate ammonification, 1:267
 environmental factors, 1:267–269
 under field conditions, 1:266–267
 genetic insights into, 1:265–266
 heterotrophic denitrification to N₂O, 1:267
 in manure management, 1:563–564
 mechanisms, 1:262–267
 nitrate supply, 1:268
 nitrifier denitrification, 1:267
 O₂ availability, controls on, 1:268
 organic C availability, 1:268
 organisms, 1:262–267
 pathways, 1:262
 pH, 1:269
 reductases of oxides of nitrogen used in, 1:265t
 substrates, 1:262–267
 temperature, 1:269
 Denitrifiers, 1:383–384
 Dense non-aqueous phase liquids (DNAPLs), 5:273
 Density
 and compressibility
 of free water, 5:156–157
 of soil water, 5:157, 5:157f
 fluid, 5:257
 flux, 5:257, 5:259f
 fractionation, mineral associated organic matter isolation, 1:397
 separations, soil organic matter, 2:75–77
 Density coefficient, 5:581–583
 Density functional theory (DFT), 2:487–488
 Derjaguin approach (DA), 5:340
 Derjaguin–Landau–Veerwey–Overbeek (DLVO) interactions, 2:598–600
 Derjaguin–Landau–Verwey–Overbeek (DLVO) forces
 electrostatic/double layer force, 5:338–339
 profiles and maps, 5:340–343
 classic DLVO theory, 5:341–343
 particle and collector surface interactions, 5:343
 separation distance, 5:342f
 Van der Waals force/energy, 5:337–338
 Desert fringe loess, 4:336, 4:337f
 Desertification, 1:14, 4:87, 4:340–341
 Desert pavements, 4:335–336
 Deserts, 4:44
 Desert soil systems, 4:29
 Desiccation tolerance, 1:132
 Desilification acrisols and alisols formation, 4:347, 4:347f
 Desorption, 2:343–344
 behavior in soils, 2:348–350
 phosphorus, 2:156–157
Desulfomicrobium norvegicum, chromium toxicity, 2:665
Desulfovibrio vulgaris, chromium toxicity, 2:665
 Detrital food web (DWB), 1:71, 1:73
 Detritusphere, 1:415–416, 1:416f
 Dew, 5:675
 Dewpoint hygrometer method, 5:200
 Dew point method, 5:178, 5:180
 Diagnostic horizons, for mineral and organic soils, 4:199
 Diatoms, 1:61–62
 Diazotrophic rhizobacteria, 1:273
 Diazotrophs, 1:271
 associative, 1:273
 carbon metabolism, 1:276
 culturable taxa, 1:287
 endophytic, 1:273
 endosymbiotic, 1:273
 host-genetic factors, 1:290
 molecular approaches, 1:289
 nifH gene-based analysis, 1:288, 1:288f
 nitrogenase activity regulation, 1:274–276
 nitrogenase enzyme complex, 1:287–288
 nitrogen regulation, 1:274–275
 oxygen regulation, 1:275–276
 oxygen status, 1:278–279
 phyllosphere diazotroph community, 1:288
 post-transcriptional regulation, 1:276
 sequencing analysis, 1:288
 soil moisture and temperatures, 1:278
 2,4-Dichlorophenoxyacetic acid
 decomposition, 2:665
 Dielectric constant, 5:152, 5:153t, 5:437, 5:439
 Dielectric matrix potential sensors, 5:178
 Diethylene triamine pentaacetic acid (DTPA) extraction, 2:51
 Differential scanning calorimetry (DSC), 2:667–668
 clay and organic matter content, 2:668
 heat-flux, 2:667–668
 humic substances, 2:668
 power compensation, 2:667–668
 Diffuse double layer (DDL) theory, 2:34–35, 4:371
 Diffuse layer surface-charge density, 2:368ge
 Diffuse reflectance infrared Fourier transform (DRIFT) spectroscopy, 2:513f, 2:514
 Diffuse reflectance spectroscopy, 4:702
 Diffusion, 5:153, 5:220–221, 5:653
 across dry soil layer, 5:495
 coefficients in soil-liquid and soil-gaseous phases, 5:248f
 current, 5:267
 definition of, 5:247
 Fick's laws of diffusion, 5:248–249
 Maxwell-Stefan diffusion equation, 5:249
 modeling soil-gas, 5:252–253
 modeling solute diffusion, 5:250–251
 origin of, 5:248
 regimes in soil, 5:250
 soil-gas diffusion coefficient, 5:251–252, 5:252f
 in soil, importance of, 5:249
 solute diffusion coefficient in soil, 5:250
 unified modeling approach, 5:253–254, 5:254f
 Diffusion-based NMR, 2:618–619
 Diffusion-based techniques, 2:618–619
 Diffusion chambers, 1:143
 Diffusion ordered spectroscopy (DOSY), 2:618–619, 2:620f
 Diffusive gradients in thin films (DGT), 2:55

- Diffusive mass transfer mechanisms, 3:301–302
- Digestate, 1:561
- Digestive track of invertebrates, 1:190
- Digital agriculture, 4:722–724
- Digital elevation models (DEMs), 4:500, 4:517, 4:686, 4:752*ge*
- Digital PCR (dPCR), 1:149
- Digital soil mapping (DSM), 4:481, 4:701*f*, 4:705–706, 4:720–721, 4:720*f*
- decision tree, 4:691–692, 4:691*f*
- generalized additive model (GAM), 4:691
- knowledge-based inference, 4:688–689
- map quality assessment, 4:687
- mechanistic modelling, 4:688–689
- operationalization, 4:692–694
- scorpan framework, 4:685
- and soil depth functions, 4:689–691
- soil information, 4:685
- soil spatial prediction functions, 4:686
- space-time modelling, 4:689
- spatial and temporal soil variation, 4:684
- uncertainties, 4:687
- uses and users, 4:693–694
- Digital soil morphometrics, 4:166, 4:568
- applications, 4:575–577
- soil profile attributes, 4:569–572
- soil profile variation, 4:573–575
- Digital soil spectra, 2:569
- Digital surface models (DSMs), 4:752*ge*
- Dilutions of high salt water, 4:378–379
- Dimensionality, solute transport, 5:654–655
- Dimension reduction, 4:517
- application, 4:501–502, 4:502*t*, 4:503*f*
- non-metric multidimensional scaling (nMDS), 4:502
- principal component analysis (PCA), 4:501, 4:501*f*
- Dinitrogen, 1:261–262, 1:265–266, 1:266*f*
- Diplopoda
- diversity, 1:95–96*f*, 1:97–98
- food, 1:98
- habitats, 1:98
- interactions, 1:98
- Diplura, 1:86*f*, 1:89–90
- Directional behavior, soil physical properties. *See* Anisotropy
- Direct labeling approach, 2:104, 2:104–105*f*
- Direct runoff, 5:288
- Direct shear tests, 5:116*f*, 5:117
- Discontinuous pressure, 5:317
- Disease suppressive soils, 1:464
- Disinfection by-products (DBP)
- during rapid infiltration, 1:575*t*
- wastewater treatments systems, 1:574–575
- Dispersal mechanisms, 1:13
- Dispersal methods, soil organic matter, 2:72–74
- Dispersion, 2:317, 5:653–654
- Dispersion coefficients, 5:354, 5:356
- Disproportionation, of sulfur, 1:307–308
- Dissimilatory metal-reduction, 2:246, 2:252
- Dissimilatory nitrate reduction to ammonia (DNRA), 1:267, 1:383–384, 2:148
- Dissociation constant, pesticides, 2:471
- Dissolution, 2:284–285
- of alkaline earth metals, 2:127–128
- chemical weathering, 4:38
- of nanoparticles, 2:686
- and precipitation of soil minerals, 2:285–287
- Dissolution constant
- of aluminum-containing minerals in soils, 2:173*t*
- overall, 2:172
- Dissolution diagram, 2:175
- Dissolution-precipitation reaction, 2:197–198
- Dissolved ‘atmosphere’, 5:218–219
- Dissolved organic carbon (DOC)
- adsorption, 2:428, 2:429*f*
- climate, 2:422–423
- communication, 2:428–430
- composition of, 2:426–427
- concentrations, 2:422–423
- definition of, 2:422
- degradation, 2:428–430
- distribution of, 2:424*f*
- in domestic, secondary-treated wastewater, 1:574
- by forest type, 2:425*f*
- in global carbon cycle, 2:430
- impact of, 2:421
- inputs and production, 2:422–426
- microbial processing, 2:428–430
- in mineral soils, 2:423*f*
- origins of, 2:422
- pH, 2:424, 2:425*f*
- quantity and quality of organic material, 2:424
- soil, 2:421
- soil type, 2:423
- soil use and management, 2:424–426
- transport of, 2:426–427
- Dissolved organic matter (DOM), 2:69, 2:71, 2:686
- composition on organo-mineral interactions, 2:427–428
- cycling of, 2:427*f*
- fluorescence anisotropy, 2:556
- fluorescence lifetime, 2:556, 2:556*t*
- fluorescent properties, 2:556
- input and output pathways, 2:555*f*
- movement of, 2:426
- Dissolved organic nitrogen (DON), 1:244, 1:249–250, 2:424–426
- Dissolved oxygen (DO) concentration, 5:367–368
- Distortionless enhancement by polarization transfer (DEPT), 2:618
- Distributed temperature sensing (DTS), 5:454
- Divalent copper, electron paramagnetic resonance (EPR) spectroscopy, 2:528
- Diversity
- analysis and Willis curve, 4:735–737
- concept of, 4:727
- fractals and multifractals, 4:734–735
- nested subsets theory, 4:733–734, 4:733*f*
- soil sciences, fractals and multifractals in, 4:735
- DNA fingerprinting method, 1:150–151
- DNA stable isotope probing (SIP) technique, 1:18
- Dokuchaev’s paradigm, soil classification
- early period, 4:189–190
- genetic approach, 4:188–189
- immediate predecessors, 4:187–188
- late period, 4:191–195
- Domestic livestock, rangeland management, 3:90–91
- Drainage
- climate change, 3:258, 3:262
- environmental effects, 3:262
- functions, 3:258
- practices, 3:257–258
- precipitation, 3:258
- soil hydrology, 3:259
- subsurface drainage, 3:259–261
- surface drainage, 3:259
- system management, 3:258
- temperature, 3:258
- types of, 3:497
- water quality and peak flows, 3:262–263
- water recycling, 3:263
- Drainage fluxmeter (DFM), 5:327, 5:329*f*
- Drilodefensins, 1:114–115
- Drip irrigation, 5:618*f*, 5:619
- Driplines and drip tapes, 5:619
- Droughts
- atmospheric, 5:536–537
- extreme, 5:536–537
- resistance, 5:533
- root form and function, 1:325–326, 1:325*f*
- stress amelioration, 3:417–418
- survival, turfgrass, 3:498–502
- Dryland, biocrusts, 1:129. *See also* Biocrusts
- Dryland farming
- agro climatic characteristics, 3:115, 3:115*t*
- climate change, 3:114–115
- climate variation, 3:115
- crop management, 3:120–123
- crops and cropping system, 3:116, 3:116*t*
- dryland distribution, 3:114
- global land use systems, 3:114, 3:114*f*
- in India, 3:121–123, 3:122*t*
- livelihood challenges, 3:116
- problems, 3:114–115
- recommended management practices (RMPs), 3:118
- soil degradation, 3:115
- soil management, 3:118–120
- soils, 3:115
- water technology, 3:117–118
- Drylands, land surface model, 4:758–759
- DSMART algorithm, 4:692
- Dual crop coefficient approach, 5:581–584
- Dual-permeability models, macropore flow, 5:312
- Dufour effect, 5:267–268
- Durids, 4:229
- Dust clouds, 3:209–210
- Dust fixation, 4:335–336
- Dust production model (DPM), 3:216
- Dust storm, 3:209*f*
- Dynamic contact angle, 5:235
- Dynamic force spectroscopy (DFS), 2:600–603, 2:603*f*

- mineral-associated organo-mineral (MAOM), 1:401
 - Dynamic global vegetation models (DGVM), 3:367
 - Dynamic vapor sorption (DVS), 5:200
 - Dynamic viscosity, 5:257
- E**
- Earth Room, 3:517f, 3:518–519
 - Earth's critical zone, 1:1ge
 - Earth's crust, 4:92, 4:92f
 - Earth system models (ESMs), 4:752ge, 4:760 simulations, 5:410
 - Earthworms
 - bioturbating activities, 1:117
 - body size, 1:113
 - burrowing, bioturbation, 5:103–104
 - climatic factors, 1:113
 - clitellum, 1:112–113
 - copulation of, 3:5f
 - cryoprotectant substances, 1:113
 - cultural services, 1:117
 - ecological categories, 1:114
 - ecosystem services, 1:115–117, 1:116f
 - external morphology, 1:114f
 - habitat and dietary requirements, 1:114–115
 - intensive agricultural practices, 1:114
 - provisioning, 1:115–116
 - regulating service, 1:117
 - soil organic matter (SOM) decomposition, 2:390
 - soil organisms interactions, 1:115
 - soil properties, 1:113–114
 - supporting services, 1:116–117
 - Ecohydrology, 5:536–537
 - E. coli*, 1:513–514
 - Ecological functions of soil
 - biomass production, 3:146–147
 - carbon sequestration, 3:148–149
 - functional and intrinsic biodiversity provision, 3:149–150
 - nutrient cycling, 3:147–148
 - potentially harmful elements and compounds, 3:150
 - soil quality, 3:146, 3:147f
 - water purification and regulation, 3:149
 - Ecology, 1:1ge
 - history of soil, 3:11
 - protists, 1:40–46
 - Economic valuation, of soil-based ecosystem services, 3:550–552
 - Economic value of soil functions, 3:507
 - Economy and soil evaluation, 3:507
 - Ecosystem, 1:3
 - in agricultural landscapes, 1:473
 - biocrusts functional properties and roles in, 1:132–134, 1:132f
 - delivery, 1:473
 - functional diversity and, 1:470
 - and landscapes, soil in, 3:9–10
 - microbial activity, 1:470
 - services, agroforestry, 3:78–81
 - services, soil, 3:9–10
 - soil microbial communities to, 1:474t
 - terrestrial, 1:480
 - urban, 1:480 (*see also* Urban soil microbiome)
 - urgency of, 1:471
 - Ecosystem services, 3:548
 - in anthropized areas, 3:37 (*see also* Anthropized area soil; Soils of Urban, Industrial, Traffic, Mining and Military Areas (SUITMAS))
 - earthworms, 1:115–117, 1:116f
 - soil, 4:85–86
 - soil water content, 5:473–474
 - wetlands, 4:438, 4:439t
 - Ecozones, soils and, 4:246
 - Ectomycorrhizal fungi, 2:429–430
 - Edaphic factors
 - affecting plant nutrients, 1:7–8
 - affecting root and microbial growth, 1:8–9
 - affecting water and air, 1:5–6
 - Edaphic properties, 1:4, 1:4–5t
 - Edaphology, 1:1ge, 1:2–3
 - Eddy covariance (EC) method, 4:612f, 5:502
 - advantages, 4:615
 - challenges, 4:612–613
 - in cotton field, 4:613f
 - measurement principle of, 4:612f
 - scintillometers, 4:615
 - variables measured with, 4:614t
 - Education, 3:506
 - world reference base for soil resources (WRB) in, 4:216
 - Effective porosity, 5:366
 - Effective saturation, 5:380
 - Effective stress, 5:73, 5:80
 - Effervescence, soil morphology, 4:162
 - Efflorescence humidity (ERH), 3:387
 - Eidgenössische Technische Hochschule (ETH) Zürich L-band Radiometer (ELBARA), 4:598, 4:605–607
 - Ekasilicon, 2:271–272
 - Elastoplastic soil behavior, under compressive stresses, 5:77–78
 - Electrical conductivity (EC), 1:1ge, 1:8, 2:408, 2:409f, 5:167
 - Electrical double layer (EDL) model, 2:34–35
 - Electrical force microscopy (EFM), 2:595
 - Electrical resistivity tomography (ERT), 5:446–447, 5:448f
 - Electric conductivity, sensor technology, 5:173
 - Electric dipole on epitaxial growth, 2:592–594
 - Electric permittivity, 5:171–173
 - Electro fluorescence polarization spectroscopy, 2:563
 - Electrokinetic remediation technique, 3:341, 3:341f
 - Electrolyte concentration, sodicity and, 4:373–374, 4:374f
 - Electromagnetic field (EM) theory, 2:499–500
 - Electromagnetic induction (EMI), 4:591–592, 4:616, 5:448–449
 - proximal sensing, 4:596–597
 - PSS in field, 4:582–584, 4:583f, 4:583t
 - Electromagnetic (EM) proximal soil moisture sensors, 4:615–616
 - Electromagnetic sensors
 - commercial water content sensors, 4:594f
 - for soil moisture, 4:593–594
 - Electromagnetic spectrum, 2:567, 2:568f
 - Electromagnetic (EM) technique
 - distribution of, 5:440
 - soil water content, 5:475
 - using radar principles, 5:437
 - wave propagation theory, 5:439
 - Electron acceptor, 1:303–304, 1:307
 - Electron activity concept, 2:19
 - Electron microscopes, 2:576
 - applications, 2:580–584
 - fundamental principles, 2:576–577, 2:577f
 - Electron paramagnetic resonance (EPR) spectroscopy
 - Boltzmann distribution law, 2:534
 - determination and interpretation, 2:523–524
 - electron angular momentum, 2:534
 - environmentally persistent free radicals (EPFRs), 2:541
 - experiment, 2:523
 - external magnetic field, 2:535
 - humic free radicals, 2:522–523
 - instrument, 2:522–523
 - instrumentation, 2:535–536
 - local magnetic fields, 2:535
 - magnitude, 2:534
 - metal binding, 2:538–539
 - paramagnetic transition metal ions, 2:526–529
 - principles, 2:521–522, 2:533–536
 - resonance phenomenon, 2:521
 - sample preparation, 2:523
 - sensitivity and resolution, 2:523
 - soils and soil organic matter, 2:536–538
 - spectral parameters, 2:521–522
 - spin, 2:534
 - spin derivatization studies, 2:529–531
 - spin labels, 2:539–540
 - spin-orbital contribution, 2:534–535
 - whole soil samples, 2:524
 - X-band EPR spectrometer, 2:522f
 - Zeeman effect, 2:534
 - Electro-osmosis, 5:265, 5:267
 - Electro-phoresis, 5:267
 - Electrostatic/double layer force, 5:338–339
 - Electrostatic forces, 2:317
 - Elemental analysis, of pyrogenic carbon (PyC), 2:434, 2:435f
 - Elemental composition, of soil particles, 2:581–583
 - Elemental cycling, acidity in soils, 2:3–4
 - Elemental imbalances, 3:168
 - Elemental mapping, 2:675
 - Element cycling, 2:465–466, 2:466t
 - Element imbalance, soil change
 - eutrophication, 3:357–358
 - fertilizer application, 3:358
 - osmotic stress, 3:356
 - salinity problems, 3:356
 - soil fertility changes, 3:357
 - toxic elements, 3:357
 - trace metals, 3:356–357
 - water quality deterioration, 3:357–358

- Elevated CO₂ effects, 1:583
- Elimination, of municipal solid waste, 3:323–331
- Eluviation, 4:292
- Emerging pollutants, 3:284–285
- Emerging soil disciplines, 3:505–507
- Emission factor (EF), 3:162
- EMMA, 3:228
- Empirical erosion models, 4:706
- Endocrine disruptors, wastewater treatments systems, 1:575
- Endophytes, 1:361, 1:361*f*, 1:420
soil-borne plant disease control, 1:443
- Endophytic nitrogen-fixing associations, 1:273
- Endosaturation, 4:413
- Endosymbiotic nitrogen-fixing associations
actinorhizal symbiosis, 1:273–274
cyanobacterial endophytic associations, 1:274
Parasponia-rhizobium symbiosis, 1:274
rhizobium-legume symbiosis, 1:273
- Energy balance, 5:635
based proximal sensing, 4:612–615
components of, 5:642–643
- Energy balance Bowen ratio system, 5:502
- Energy consumption and energy balance of
buildings, green roofs, 3:485–486
- Energy dispersal spectroscopy (EDS), 2:578
- Energy recovery, of municipal solid waste, 3:326–331
- Energy transfer methods, soil water content monitoring, 5:474–475
- Engineered containment cells (ECCs), 3:335, 3:335*f*
- Engineered nanoparticles (ENPs), 2:682, 2:684, 2:685*f*
- Enhancing effect, 2:332–333, 2:333*f*
- Ensemble Kalman filtering, 4:689
- Ensembles, 4:514
- Entisols, 4:55*ge*, 4:230, 4:296, 4:326, 4:422
- Entrapped air, 5:317
- Enumeration techniques, 1:147–148
- Environmental clean and green, 3:489, 3:497, 3:502
- Environmental components, soil, 3:10
- Environmental dependencies, water use efficiency, 5:534*f*
- Environmental education, 3:576
- Environmental effects of irrigation, 3:268
metals, 3:273
pesticides, 3:272–273
plant nutrients (fertilizers), 3:271–272, 3:272*f*
pollutants attributable to irrigation, 3:271–273
public health concerns, 3:273–274
sustainability, 3:274
waterlogging and salinity, 3:270, 3:271*f*
water shortage, 3:268–270, 3:268–269*f*
- Environmental factors, microbial
denitrification, 1:267–269
- Environmental filters and buffers, 3:9
- Environmentally persistent free radicals (EPFRs), 2:541, 2:541*f*
- Environmental quality, sulfur, 1:308–309
- Environmental scanning electron microscope (ESEM), 2:586, 2:586*f*
- Environmental sciences, 3:505
- Enzymatic mineralization, 1:305
- Episaturation, 4:413
- Equilibrium, 2:274*ge*
- Equilibrium isotope fractionation, 2:108
- Equilibrium tension lysimeters, 5:327
- Equivalent porous medium (EPM), 5:177, 5:180
- Erosion, 3:166–167, 4:59, 4:75–76
acrisols and alisols formation, 4:349
present-day and past rates, 4:76
rates in natural soils, 4:75–76, 4:76*f*
and soil flux, 4:743
soil water repellency (SWR), 5:212
and suspended sediments, 4:380
tillage, 4:346
wind (*see* Wind erosion)
- Erosion control measures
agronomic methods, 3:201, 3:202*f*
approaches, 3:201
mechanical methods, 3:202–203, 3:202–204*f*
soil management methods, 3:203, 3:204–205*f*, 3:205, 3:206*t*
- Erosivity, 3:197–198
- Errors and uncertainty in soil class maps, uncertainty assessment, 4:677–679
- Errors in spatial soil variables, 4:675
- Error standard deviation from replicates, 4:674, 4:674*f*
- Essential climate variable (ECV), 4:591–592
- Essential elements, 1:311–312
- Estimation variance, 4:490
- Estuarine wetlands, 4:438
- Ethylenediamine tetraacetic acid (EDTA), 2:51
- Ethylene glycol monoethyl ether (EGME), 2:22, 2:23*t*, 2:29–30
retention, 2:23*t*
- Eukaryote Tree of Life, 1:40, 1:41*f*
- Eukaryotic algae, 1:128
- Eukaryotic micro-algae, 1:64–65
- Eulerian methods, colloid transport, 5:350
- European Chemicals Agency (ECHA), 2:426
- Eutric andosol, 4:466*t*
- Eutrophication, phosphorus loss, 2:162
- E-value, 2:99–100
- Evaporation, 4:48*ge*, 4:610–612
capillary flow paths, continuity of, 5:493–494
capillary forces *vs.* gravity, 5:494
capillary *vs.* viscous flow, 5:494–495
from catchments to global analyses, 5:497–498, 5:498*f*
characteristic length and stage-2, 5:495
characteristic length, stage-1, 5:493–495
and diffusion across dry soil layer, 5:495
first and second stage of, 5:492–496
in hydrologic cycle and surface energy balance, 5:491, 5:491*f*
imaging at column scale, 5:497
mechanisms sustaining stage-1, 5:493
planes and drying fronts, 5:492*f*, 5:493
pore scale considerations, 5:497
- processes at field scale, 5:497–498
quantification of, 5:497–498
rate dynamics, 5:492
from shallow water tables, 5:495
soil water characteristic (SWC) curve, 5:199–200
soil water repellency (SWR), 5:211
structure formation and land use on, 5:496, 5:496*f*
of water from bare soil, 5:490
- Evapotranspiration (ET), 4:14, 4:611–612, 4:614, 5:483, 5:628, 5:630
aerodynamic approach, 5:503
for agricultural fields, 4:615
atmospheric CO₂ concentration, 5:506
climate model simulations, 5:506
components, 5:504–505*t*
dimensions, 5:501
diurnal variations, surface energy balance components, 4:614*f*
at ecosystem scale, 5:529
eddy covariance (EC) method, 4:612–613, 4:612–613*f*, 4:614*t*
energy balance approach, 5:502
energy balance-based proximal sensing of, 4:612–615
equation (*see* Penman–Monteith (PM) equation)
factors that control, 4:611–612
field measurements, 5:502
historical development, 5:502
land surface, 5:501
land surface energy balance and, 4:611, 4:611*f*
land surface temperature (LST), 4:608
land surface temperature space models, 5:503–504
land use and land cover changes, 5:506
lysimeters, 5:668
magnitude, 5:504–505
mass balance methods, 5:502
measurement, 5:502
micrometeorological methods, 5:502
models for, 5:502–504
multi-layer ET models, 5:640–641
Penman model, 5:503
potential, 5:531, 5:532*f*, 5:536
scintillometer method, 4:614–615, 4:615*f*
soil water availability, 5:470–471
surface energy balance (SEB) models, 5:503–504
terrestrial water balance and, 4:611
vegetation index (VI), 5:503–504
water balance-based proximal sensing of, 4:615–616
water management, 5:506
- Excavated bench terrace, 3:452
- Excess pore water, 3:259
- Exchangeable sodium percentage (ESP), 4:369, 4:374–376*f*, 4:376
- Excitation-emission matrix (EEM), 2:547
- Excitation-emission matrix-fluorescence regional integration (EEM-FRI), 2:554
- Existence value, 3:551–552
- Exponential function, 4:660
- Exponential growth, 1:175

- Externality, 3:547*ge*, 3:548
- External/Macro-catchment system, 5:600, 5:600*f*, 5:601*t*
- Extracellular DNA inactivates infectious root pathogens, 1:339
- Extracellular polymeric substances (EPS), 1:120–123
- microbe–OM–mineral interfaces, 2:460–461, 2:463
- Extracellular polysaccharide production
- bacteria, 1:36
- bacteria life strategies and survival, 1:36
- Extraction, proteins, 2:399–400
- Extraradical mycelium (ERM), 1:453, 1:465, 1:466*f*
- Extreme soil microorganisms
- alkaline soils, 1:506
- astrobiology, 1:507–508
- biodiversity conservation, 1:508
- biological networks and dispersal, 1:501–506
- characteristics, 1:497–498*t*
- climate change, 1:507
- cold soils, 1:501–504
- colonization patterns, 1:495
- hot soils, 1:505–506
- industry, agriculture, and medicinal applications, 1:508
- microbial adaptive potential, 1:496–500
- microbial colonizers, 1:507
- microbial community structure, 1:496
- radioactive soils, 1:506
- saline and hot desert soils, 1:504–505
- stochastic resuscitation model, 1:500*f*
- subsurface environments, 1:506
- viruses, 1:506–507
- Extremophiles, 1:508. *See also* Extreme soil microorganisms
- ecological, and genetic strategies, 1:502*t*
- physiological strategies, 1:502*t*
- Extremozymes, 1:508
- Extrinsic thresholds, 4:76
- Exudates
- composition, 1:472
- C-rich, 1:470
- from fine roots, 1:472
- F**
- Factor of Safety (FoS), 5:121
- Fallow, 3:439*ge*
- Fanya juu terraces, 3:202, 3:202*f*
- Farmed soils, 3:24–26
- Farmer-centric policy implementation, 3:466, 3:467*f*
- Farmer-to-farmer training initiative, 3:576
- Farming system
- agro-ecology (AE), 3:16
- current and future aspirational land use, 3:17, 3:18*t*
- food production *vs.* population growth, 3:12
- food system transformation, 3:17
- innovative farming system, 3:13–14, 3:14*f*
- natural resources, 3:13
- nature-positive farming system, 3:15–16
- productivity by 2050, 3:13
- regenerative agriculture (RA), 3:15–16
- returning land to nature, 3:15, 3:17
- rights-of-soil, 3:16–17
- soil restorative farming system, 3:13–14, 3:14*f*, 3:15*t*
- strategies for 5 R's, 3:15
- sustainability indices, 3:14, 3:15*t*
- Farm ponds, 3:117, 3:117*f*
- Farm yard manure (FYM) as mulch, 3:394
- Fast neutrons, 2:627
- Fast-slow gradient, 1:324
- Fate and transport, 2:422
- Fatty acids-based approaches, 1:149–150
- Fauna, soil organic matter (SOM) decomposition, 2:390
- Fecal contamination, 1:513
- Feedstock composition, 3:416
- Feldspars, 4:92
- chemical formulae, 4:98*t*
- composition and nomenclature, 4:97*f*
- specific gravity, 4:98*t*
- structure, 4:96–97, 4:96*f*
- unit-cell dimensions, 4:98*t*
- weathering, 4:99–100
- Fens, 4:384*f*, 4:392
- Ferralsol (Oxisol), 4:247*f*, 4:461–464, 4:463*t*
- Ferric iron, 2:527–528
- Ferrihydrite, 2:292, 4:140
- Ferro-andic, 4:458*ge*
- Ferrolysis, 4:348
- Fertigation, 5:620–621
- Fertilizers, 1:7, 3:370–371
- application, acidity in soils, 2:4
- application on soil organic matter, 3:374
- biochar, 2:414
- and climate change, 3:379
- effects on soil erodibility, 3:378
- environmental effects of irrigation, 3:271–272, 3:272*f*
- and ground waters, 3:379
- and heavy metal contamination of soils, 3:378
- nitrogen, 3:371*f*
- nutrients and soil properties, 3:379
- on nutrient supplying capacity of soils, 3:372–374
- positive or negative effects, 3:376–377
- recommendations and effects on crops and soil, 2:120–121
- on soil microbial biomass, 3:376–377
- soil physical properties, 3:378
- trace elements, 3:291
- use and soil acidification, 3:375–376, 3:375*f*
- use on enzyme activity in soil, 3:377
- use on pollution of surface, 3:379
- Fertilizers/pesticides hybrid, 2:689
- Fiber bundle model (FBM), 5:120, 5:121*f*
- Fibrist, 4:226
- Fick's law of diffusion, 5:248–249, 5:387
- Field capacity (FC), 5:510–511
- Field sampling
- artificial contamination, 4:553
- climatic conditions and soil humidity, 4:552
- clod samples, 4:558, 4:559*f*, 4:560
- coarse content, 4:560
- conditioning of sample, 4:560–561
- cultural constraints, 4:552
- field notes, 4:560
- humus horizons sample, 4:554, 4:555*f*
- optimal sampling conditions, 4:552
- organo-mineral and mineral sample by augering, 4:553–554
- sampling report, 4:560
- sampling zones, 4:553
- soil pit, 4:554
- volumetric samples, 4:554–558
- Field-scale simulation models, immiscible fluid, 5:280
- Filling, of lysimeters, 5:671*f*
- Filter-paper method, 5:178
- Filtration
- bacterial transport in soil systems, 1:516
- soil water regime, 1:518
- wastewater treatments systems, 1:572–573
- Fingered flow
- in dry, wettable sandy soils, 5:317
- modeling gravity-driven fingers, 5:318–319
- unstable, 5:317
- Fire, 3:95, 4:88
- Fire-affected soils, microbiology, 1:505–506
- Fire effects
- combustion and soil heating, 2:449–450
- direct effects, 2:448–454
- indirect, 2:448–449
- nutrients, 2:452
- soil biological characteristics, 2:454
- soil carbon stocks, 2:450–451
- soil mineralogical composition, 2:452
- soil organic matter, 2:450–451
- soil pH, 2:452
- soil physical characteristics, 2:452–453
- Firewood, 3:6
- Fixed effects, 4:495–496
- Flame atomic absorption spectrophotometer (FAAS), 2:55–56
- Flickering cluster model, 5:150, 5:150*f*
- Floating ore hypothesis, 4:361
- Flocculation and dispersion. *See* Clays
- Flood management, 1:478
- Floodplain, 4:419–420
- Flood risk reduction, green roofs, 3:486
- Floodwater diversion systems, 5:602
- Floodwater harvesting, 5:601–602, 5:602*f*
- Flow and transport
- case study, 5:135
- coupling-interaction parameters, 5:134, 5:134*f*
- equations governing, 5:134–135
- general implications, 5:131–133
- interacting solutes, 5:134–135
- Flow cytometry (FCM), 1:142
- Flow modelling
- homogenization and continuum-scale, 5:26*f*
- pore-scale modelling (*see* Pore-scale modelling)
- Fluid entrapment, hysteresis, 5:279

- Fluorescence correlation spectroscopy (FCS), 2:549, 2:550*f*, 2:556–557
- Fluorescence cross-correlation spectroscopy (FCCS), 2:549
- Fluorescence imaging, root exudates, 1:354
- Fluorescence in situ hybridization (FISH), 1:149
- Fluorescence microscopy, 2:677
- Fluorescence quenching
- polycyclic aromatic hydrocarbons (PAHs) sorption, 2:559–561
 - soil organic components, 2:557–559
- Fluorescence spectroscopy
- classical monodimensional, 2:544
 - clay minerals, 2:552
 - fluorescence correlation spectroscopy (FCS), 2:549
 - fluorescence intensity, 2:547
 - fluorescence lifetime, 2:546–547
 - fluorescence spectra, 2:546–547
 - fluorescence spectrophotometer, 2:545–546
 - frequency-domain method, 2:546
 - instruments, 2:545–546
 - laser induced fluorescence spectroscopy (LIFS), 2:549–550
 - limitations, 2:547–548
 - molecular and environmental factors, 2:548
 - polarization, 2:549
 - quantum yield, 2:547
 - quenching process, 2:548–549
 - relaxation process, 2:544–545
 - soil dissolved organic components, 2:556
 - soil mineral components, 2:561–566
 - soil organic components, 2:552–554
 - temperature control devices, 2:546
 - time correlated single photon counting (TCSPC), 2:550–551
 - time-resolved fluorescence (TRF) spectra, 2:546
 - tridimensional, 2:544
 - X-ray fluorescence spectroscopy, 2:550
- Fluorescence zone integration (FZI) method, 2:547
- Fluoride ions, 2:171
- Fluorine (F), 2:235–237
- Fluorophores, soil organic matter (SOM), 2:553–554
- Fluvents, 4:230
- Fluvisol, 4:248*f*, 4:422
- Flux balance analysis (FBA), 1:177
- Fock operator (FO), 2:487
- Foliar nutrition application
- adjuvants, 3:389
 - agronomic and physiology-related conditions promoting, 3:383*t*
 - barrier properties, leaf surfaces, 3:383
 - bio-fortification, 3:390
 - dry topsoil, 3:389–390
 - efficiency, 3:390–391
 - environmental effects, 3:389
 - horticultural commodities quality, 3:390
 - humidity effects, 3:387–388
 - large nutrient demand, 3:390
 - leaves, 3:384–387
 - poor soil nutrient availability, 3:389
- Foliar sprays, 3:120
- Folic histosols, 4:384–385, 4:384*f*
- base saturation (BS), 4:390
 - characteristics and properties, 4:390
 - C:N ratio, 4:390
 - C:P ratios, 4:390
 - formation, 4:386
 - pH, 4:390
- Folists, 4:226
- Food and Agriculture Organization (FAO)/International Union of Soil Science (IUSS) World Reference Base, 4:414–415
- Food safety, 3:293–294
- Food security, 3:538
- availability and quality, 3:557
 - definition, 3:557
 - essential mineral nutrients, 3:558*t*
 - factors and systems, 3:557*f*
 - inherent soil fertility, 3:557
 - insufficient food, 3:557
 - nutritional value, 3:559
 - salinization and desertification, 3:559
 - under-nutrition, 3:557
- Food system transformation, 3:17
- Force field (FF) method, 2:485–486, 2:488–490
- Force field parameterization, 2:490
- Forensic science, 3:505, 4:183
- Forensic soil science, 3:507
- chain-of-custody, 3:528
 - communication, 3:530
 - history, 3:523
 - human-made soils, 3:528
 - initial description and characterization of sample, 3:528–529
 - random sampling, 3:524–525
 - soil characteristics, 3:523
 - soil classification, 3:528
 - soil databases, 3:526–528
 - soil evidence, 3:526
 - soil examination and methods of analysis, 3:525–526, 3:525*f*, 3:527*f*
 - soil information evaluation, 3:529–530
 - soil inorganic analysis, 3:529
 - soil organic characterization, 3:529
 - soil particle characterization, 3:529
 - soil sampling at crime scene, 3:523–525, 3:524*f*
 - targeted sampling, 3:524–525
 - theory of soil transfer, 3:523
- Forest fires, 2:434
- Forest resilience, soil and, 3:63–64
- Forest soils
- characteristics, 3:57
 - global carbon dynamics and greenhouse gases, 3:61–62
 - global distribution of forests by climatic zones, 3:57, 3:57*f*
 - human use of forests and, 3:58–59
 - intensive management of forest and, 3:59–60
 - litterfall and forest floor, 3:60–61, 3:61*f*
 - management practices, 3:61*f*
 - microbiome, 3:60
 - monitoring network, 3:175–176
 - natural and planted forests, 3:58*t*
 - organic matter recycling in forests, 3:60–61
 - watershed nutrient dynamics and hydrology, 3:62–63
- Forest type, dissolved organic carbon by, 2:425*f*
- Formal diversity indices, 4:729
- Formation constant
- of Al, 2:170, 2:170*t*
 - of Al-LMWOA, 2:171, 2:171*t*
- Formation of oxide or ferrallitic materials, 4:181
- Forsterite, 4:102*f*
- Forward modelling
- for active remote sensing, 4:621–623
 - and inversion, 5:455, 5:456*f*
 - for passive remote sensing, 4:623–624
- Fossil soils, 4:2, 4:5*f*
- Fougerite, 2:177, 2:181–182
- Fourier-transform infrared-photoacoustic spectroscopy (FTIR-PAS), 2:501, 2:502*f*
- Fourier transform infrared (FTIR) spectroscopic method
- attenuated total reflectance (ATR) experiments, 2:514–515, 2:516–517*f*, 2:517
 - diffuse reflectance infrared Fourier transform (DRIFT) experiments, 2:513*f*, 2:514
 - infrared peaks of interest, 2:511, 2:511*f*
 - infrared spectral analysis, 2:511–512, 2:512*f*
 - instrumentation, 2:513
 - mineral-associated organo-mineral (MAOM), 1:400
 - molecular vibration, 2:510
 - as quantitative tool, 2:513
 - transmission-mode measurements, 2:513*f*, 2:514
- Four-tiered waste management hierarchy, municipal solid waste, 3:324*f*
- Fractals and multifractals
- diversity, 4:734–735
 - in soil sciences, 4:735
- Fractional advection-dispersion equation (FADE), 5:353*ge*, 5:359–360
- Fractional advection-dispersion models, 5:663–664
- space-fractional advection-dispersion equation, 5:664
 - time-fractional advection-dispersion equation (t-FADE), 5:663–664
- Fracture, soil mechanics, 5:80–81
- Fragipan, 4:54*ge*
- Fragmentation, in tillage systems, 3:437
- Framework law and sectoral provisions, 3:540
- Frankia*, 1:273–274
- Free-energy definition, of chemical equilibrium, 2:275–276
- Free-living nematodes, 1:105–107
- Free scaling laws, 4:734
- Free silt minerals, 4:44
- Freezing and thawing cycles

- climate change, 5:407
 climatic dependence, 5:406–407
 consequences, 5:400–401
 constitutive behavior, 5:403–405
 dominant controls, 5:401
 external loads, 5:405–406
 geometrical controls, 5:405
 ice-liquid equilibrium, 5:403–404
 liquid films, 5:405
 liquid supply, 5:403
 mechanical disruption, 5:400–401
 mitigation strategies, 5:407–408
 permeability reductions, 5:400–401
 soil freezing curve, 5:404–405
 solutes, 5:406
 thermal conditions, 5:401–403
 transient heat flow, 5:400–401
 Freezing point depression, 4:253
 French classification, of soil, 4:191
 Frequency domain, 5:441
 Frequency domain electromagnetics (FDEM), 5:448, 5:448f
 Frequency domain reflectometry (FDR), 5:173
 Freshwater ecosystems, 3:277
 Freundlich equation, 2:319–320
 Friability, 3:438
 Frost action, 4:180–181
 Frost boils, 4:257–258
 Frost-cracking, low Arctic soils, 4:267
 Frozen soil
 hydraulic conductivity, 5:420
 infiltration, macropores, 5:420–422
 land-use effects, 5:422
 snowcover, 5:422
 Fulvic acids (FAs), fluorescence spectra of, 2:553
 Fumigation-based techniques, 1:148
 Functional metagenomics, 1:163
 Functions of soil. *See* Ecological functions of soil
 Fungal aggregate-mineral interactions, soil organic matter (SOM) formation, 2:392
 Fungal cell walls, 2:248
 Fungal-derived calcium oxalate, 2:251
 Fungi, 2:243, 2:251, 2:422
 activity in soil, 1:58
 classification, 1:49–50
 commercialization of plant growth promoters, 1:426
 direct observation, 1:57–58, 1:57f
 exploitation, 1:58–59
 with fauna, 1:54
 isolation methods, 1:56–57
 microbiome and holobiome, 1:54–55
 morphology and form, 1:50, 1:51f
 nutrient cycling, 1:55
 with other microbes, 1:52
 physiological and molecular methods, 1:58
 physiology and reproduction, 1:50
 with plants, 1:52–54, 1:53f, 1:54t
 soil architecture, 1:55–56, 1:56f
 soil formation and pedogenesis, 1:55
 soil organic matter (SOM) decomposition, 2:390
 Fungicides, 2:471
 Funnel flow, 5:307
 Furrow irrigation, 5:611–612, 5:611–612f
 Fuzzy clustering approach, 4:506
- ## G
- Gallium (Ga), 2:215
 Gamma detectors, 2:628
 γ diversity richness, 4:729
 Gamma (γ) radiation, 2:257
 Gamma-radiometers (GR), 4:717
 Gamma-ray spectrometry (GRS), 4:584
 Gamma-rays, PSS in field, 4:584
 Gamma-ray tomography, 5:464
 Gapon equation, 2:36
 Gas chromatography, root exudates, 1:353
 Gas diffusivity, anisotropy, 5:295–296
 Gas exchange mechanisms, 5:220–221
 Gasification, 3:328
 Gaussian function, 4:661
 Gaussian solute transport, 5:355
 Gedroits soil classification, 4:190
 Gelands, 4:226
 Gelepts, 4:229
 Gelisols, 4:55ge, 4:226
 Gelods, 4:227
 Gelolls, 4:229
 Generalized additive model (GAM), 4:691
 Genesis, boreal forest soils, 4:280–285
 Genetic horizons, 4:198
 Genetic insights into microbial denitrification, 1:265–266
 Genome-scale models (GEM), 1:176–178, 1:177f
 Geoarchaeology, 4:183
 archaeology and soil science, 3:20
 methodological approaches, 3:22–26
 paleosols, 3:24
 pedo-stratigraphic sequences, 3:23
 research and techniques, 3:26–27
 soil formation process, 3:22–23
 soil genesis, 3:23–24
 soil profiles, 3:21–22
 Geochemical models, 2:279
 Geochemistry, 4:237, 5:367–368
 Geographical information systems (GIS), 4:582
 cloud-based portals and processing engines, 4:697
 databases and processing applications, 4:697
 decision support systems, 4:706–708
 digital soil mapping and modeling, 4:705–706
 ethics and privacy issues, 4:697
 pedometrics, 4:706–708
 simulation models, 4:706–708
 soil-environmental geospatial data acquisition, 4:697
 Geological erosion, Chinese Loess Plateau, 3:345–348
 Geomorphic surface, 4:417
 Geomorphology, 4:182
 Geophagy, 3:516, 3:516f
 Geophysical methods
 applications, 5:445–446
 data standardization and open science, 5:456
 distributed temperature sensing (DTS), 5:454
 electrical resistivity tomography (ERT), 5:446–447
 electromagnetic induction (EMI), 5:448–449
 forward modelling and inversion, 5:455, 5:456f
 ground penetrating radar (GPR), 5:450–451
 induced polarization (IP), 5:451–452
 magnetic soil mapping, 5:449–450
 magnetometry, 5:449–450
 nuclear magnetic resonance (NMR) measurements, 5:452
 seismic methods, 5:452–454
 variables, 5:445
 Geoportals, 4:697
 Geospatial Soil Sensing System (GEOS3), 4:638
 Geospatial technologies, 4:697
 Geostatistical modelling, 4:691
 Geostatistics, 4:478, 4:478f, 4:656
 applications, 4:527
 and new developments, 4:669–670
 block kriging, 4:527
 change of support, 4:527
 classification as external drift, 4:663–664
 cokriging, 4:527
 combining trend and random fluctuation, 4:663
 coregionalization, 4:664–666
 estimating variogram, 4:658–659, 4:658f
 fitting the models, 4:662
 framework, 4:523–524
 heterogeneous spatial data, 4:527
 models for variograms, 4:659–661, 4:659f
 multi-scaling, 4:524, 4:525f, 4:527
 random variables and random functions, 4:656–658, 4:656f, 4:657t
 reality, 4:524, 4:524f
 residual maximum likelihood estimation, 4:663
 scale of measurement and spatial variation, 4:523–524
 spatial dependence, 4:523
 spatial support, 4:523
 unbounded models, 4:661, 4:661f
 Geothermal soils, microbiology, 1:505
 German Act on Emission Control, 3:542
 Germanium (Ge), 2:215
 atomic Ge/Si ratio, 2:272–273
 properties, 2:263
 in soil–plant system, 2:272
 German Soil Protection Act, 3:543
 German system, soil classification, 4:191
 Gibberellin synthesis, 1:423
 Gibbs-Duhem equation, 5:263
 Gibbs energy, 5:263

- Gibbs free energy, 5:160
- Gibbsite, 2:267f, 2:268–269, 4:100, 4:119, 4:144
- Glass dust, 2:117
- Gleization, boreal and subarctic soils, 4:266
- Gleyed/gleying, 1:1ge
- Gley morphology, anaerobic soil, 4:401, 4:405f, 4:409f
- development, 4:403
- impeded vertical drainage, soil profiles, 4:409f
- mineral soil profiles, 4:408–410
- types, 4:403f
- waterlogging duration, 4:410, 4:415
- Gley soil colors, 4:430
- Gleysols, 4:284, 4:284f
- Global carbon cycle, dissolved organic carbon in, 2:430
- Global change, 4:740
- afforestation, 3:366
- agriculture, 3:361
- climate change, 3:361–362, 3:361f
- cold and humid environments, 3:365
- deterministic modelling, 3:367
- empirical/statistical predictions, 3:367
- irrigation, 3:365–366
- land use and land cover changes, 3:362–363
- land use and management, 3:366
- N-fertilization/nutrient export, 3:366
- pedogenesis, human effects, 3:366
- pedogenesis modelling, 3:367
- primary minerals weathering, 3:365–366
- soil properties, 3:366–367
- soil stocks and fluxes, 3:363–365
- zonal soil distribution, 3:365
- Global Climate Observing System (GCOS), 4:591–592
- Global cycle of sulfur, 1:304
- Global Energy and Water Cycle Exchanges Project (GEWEX), 4:752ge
- Global GHG balance, 1:371
- Global Navigation Satellite System Reflectometry (GNSS-R), 5:475
- Global navigation satellite systems (GNSS), 4:713–715
- Global positioning system (GPS), 4:582, 4:713
- Global Soil Doctors Programme, 3:576
- Global Soil Hydraulic Properties (GSHP) database, 5:192
- Global soil laboratory (GLOSOLAN) initiative, 4:639
- Global Soil Map (GSM), 4:703–705
- GlobalSoilMap initiative, 4:693
- Global soil moisture regimes map, 4:223f
- Global soil regions map, 4:221f, 4:222
- Global soil resources, 4:207
- Global Terrestrial Network for Permafrost (GTN-P), 4:759
- Global warming potentials (GWP), 3:154
- Global water cycle, 5:470
- anthropogenic effects, 5:485
- biomass distribution, 5:483f
- hydrologic process measurement and model, 5:485–486
- machine learning approaches, 5:488
- precipitation and, 5:483
- soil mediated fluxes, 5:484–485
- soil properties, 5:486
- water level monitoring systems, 5:488
- Global wind erosion, 3:210
- Glomalin-related soil protein (GRSP), 2:401–403
- Gluconic acid, 2:427–428
- Glucosamine, soil organic matter (SOM), 2:392–393
- Glucose, 2:427–428
- Glyphosate, 2:494, 2:494f
- Goethite, 2:177, 2:179, 4:137–139
- Goiter, 3:561
- Golf courses
- green maintenance in, 3:497
- natural ecosystem with water features in, 3:493
- planning, 3:497
- slicing and spiking, 3:498
- surface quality and safety design of, 3:489
- sustainable maintenance planning, 3:489
- traditional maintenance methods, 3:503
- vegetation cover on, 3:497
- Grain-for-Green program, 3:348
- Granular transport by ants and termites, biophysical modeling, 5:112–113
- Graphite furnace atomic absorption spectrophotometer (GFAAS), 2:55–56
- Grass breeding, 3:110
- Grassland
- crop rotations, 3:127
- definition, 3:102–103
- global distribution in 2000, 3:103f
- grazing on soil properties, 3:106–107
- hay production, 3:107–110
- impact of vegetation/species, 3:110
- inorganic fertilizers, application of, 3:108
- irrigation of grassland soils, 3:109
- lime on grassland soils, 3:109
- mowing, 3:107–108
- organic amendments, 3:108–109
- production systems, 3:104–105
- re-seeding, 3:109–110
- silvopasture, 3:110
- soil carbon sequestration in, 3:104
- soil processes, 3:103–104
- on soil properties, 3:105–107
- Grassland soil management, carbon sequestration, 3:429
- Grassland soils of cool-arid-temperate ecozone. *See also* Mollisols-Chernozems
- bioturbation and structure development, 4:303
- characteristics, 4:299–300
- classification, 4:299–300, 4:305, 4:306f
- clay translocation, 4:301
- grassland soil formation, 4:300
- hydrology and its influence, 4:304–305
- Kastanozems, 4:307, 4:310–311f
- location and their soil profiles, 4:302f
- mineral weathering, 4:301
- organic matter, 4:301–304
- Phaeozems (Udolls and Aquolls), 4:306, 4:308–309f
- secondary carbonate mineral formation, 4:301
- Grassland soil systems, 4:29
- Grass waterways, 3:203, 3:204f
- Gravel mulch, 3:397
- Gravimetric measurements, water content, 5:170–171
- Gravitationally-driven surface irrigation, 5:609
- Gravitational potential, 5:163, 5:518
- Grazing systems, 1:191, 4:341
- management, 3:448
- rangeland management, 3:94–95
- on soil properties, 3:106–107
- Great Plains prairie grasslands
- biomass production, 3:352
- government policies, 3:353
- restoration, 3:355–356
- soil organic matter loss, 3:353–355
- Great plate count anomaly, 1:140
- Greenhouse gas emissions, 3:416–417
- carbon dioxide emissions, 2:692–693
- methane emissions, 2:693–694
- nitrous oxide emissions, 2:694–695
- Greenhouse gases (GHGs), 3:154
- carbon dynamics and, 3:61–62
- emissions of, 3:62
- Green infrastructure (GI), 3:474
- Green manures, 3:446–447
- Green roofs, 3:51
- aging and early pedogenesis, 3:484–485
- air capacity, 3:480f
- air quality, 3:486
- benefits, 3:485–486, 3:485f
- biodiversity, 3:486
- biological activity, 3:482
- buffer capacity, 3:481
- bulk density (rb), 3:480f
- cation exchange capacity, 3:481
- characteristics, 3:475t
- chemical requirements of, 3:481–482
- chemical stability, 3:481–482
- constructed technosols for, 3:478–482
- energy consumption and energy balance of buildings, 3:485–486
- flood risk reduction, 3:486
- and modular systems of living walls, 3:480
- natural and recycled organic materials, 3:484
- natural soils and natural mineral materials, 3:482–483
- parent materials and constructed technosol composition, 3:482–485
- pH, 3:481
- physical characteristics, 3:480f
- processed natural materials, 3:483
- profile of, 3:476f
- recycled mineral materials, 3:483–484
- semi-intensive and intensive, 3:480
- substrate composition, 3:484
- technical aspects, 3:475–476
- and vertical green systems, 3:474–478
- Green rust, 2:181–182, 4:141–142

- Green-to-red ratio index (GRI), 5:521
 Green water, 4:473
 Groenevelt, Pieter Hendrik, 5:269–270
 Gross irrigation requirements (GIR), 5:572
 Gross mineralization, 2:102, 2:102f
 Gross primary production, 5:529, 5:535f
 Groundmass, 4:174–175
 Ground penetrating radar (GPR), 4:585, 4:616, 4:717, 5:450–451
 cross-borehole measurements, 4:596
 off-ground, 4:596
 on-ground (surface), 4:596, 4:597f
 soil moisture, 4:591–592, 4:595–596
 Groundwater, 3:379, 4:751ge
 contamination, 5:368–369
 definition, 5:365
 flow, 5:366–367
 geochemistry of, 5:367–368
 harvesting, 5:603–604, 5:603f
 hydrology, 5:365–370
 hypothesis, 4:361
 impact of soils on, 5:370–374
 as natural and anthropogenic resource, 5:369–370
 occurrence, 5:366
 quantity and quality of, 5:364–365
 recharge, 3:189, 4:292–293
 regional flow system, 5:367f
 rooting zone for terrestrial vegetation, 5:365
 soil influence on, 5:364
 storages and conduits, 5:424
 vadose zones to, 5:365f
 Group 14 elements, 2:263
 Group method of data handling (GMDH), 4:648
 Guggenheim-Andersen-de Boer (GAB) model, 2:28f
 Gully erosion control, 3:205
 Guoy-Chapman theory, 2:34–35
 Gypsids, 4:229
 Gypsum, 2:132, 4:178, 4:378
 Gypsum blocks, 5:177–178
 Gypsum rock-colonizing microbes, 1:505
- ## H
- Habitat formation and nesting activity, bioturbation, 5:107–108
 Habitats. *See also* Microhabitats
 abiotic conditions, 1:408–412
 biotic conditions, 1:412–416
 detritusphere, 1:415–416, 1:416f
 food procurement, 1:408
 light, 1:409
 pH, 1:411–412, 1:412f
 porosphere, 1:413–415, 1:414f
 properties, 1:407
 rhizosphere, 1:415
 soil inhabitants, 1:407–408
 solid phase, 1:408–409, 1:409f
 temperature, 1:409–410
 water, 1:410–411, 1:410f
 Hagen-Poiseuille equation, 5:311
 Haines jump concept, 5:20f, 5:21
 Half-cell method, 5:250–251
 Halites, 1:505
 Halloysite, 2:174f, 4:114–115, 4:115f
 Halogens, in soils, 2:262
 abundance, 2:235–236
 geochemistry, 2:235
 properties, 2:235t
 total halogen stocks in foliar biomass, 2:241, 2:241f
 Hamaker approach, 5:337–338
 Hani rice system, 3:453
 Hansen-Hurwitz estimator, 4:538
 Harmful metalloids in soil, 2:506
 Harrowing, 3:435f
 Harvard Forest fulvic acid sample, 2:612f
 Harvestmen, 1:84–86, 1:85f
 Hausmannite, 4:143
 Haymaking, 3:105f
 Hay production, 3:107–110
 HCN, 1:424
 Heat and vapor transfer, 5:385
 Heat and water movement
 applications, 5:390
 continuum model limitations, 5:390
 Heat dissipation matric potential sensor, 5:164–166
 Heat-dissipation sensors (HDS), 5:178
 Heat flux, 5:388
 Heat-flux differential scanning calorimetry, 2:667–668
 Heat storage, 5:388
 Heat transfer
 governing equation, 5:388
 mechanisms, 5:388
 Heat transport, lysimeters, 5:675–676
 Heat treatment temperature (HTT), 3:415, 3:416f
 Heavy coarse organic matter fraction (hcOM), 2:71
 Heavy metals
 contamination of soils, 3:378
 organo-mineral interactions, 1:403–404
 sorption, kinetics of, 2:347
 Hedge networks, 1:473
 Hedgerows, 1:473, 1:474t
 Hele-Shaw cell, 5:316
 Helminths, removal of, 1:572
 Hematite, 2:179, 4:140
 Hemicelluloses, 1:524
 Hemists, 4:226
 Herbicides, 2:471, 3:441, 3:445
 and mechanical weeding, 3:446
 resistance, 3:439ge
 tolerance, 3:439ge
 for weed control, 1:463–464
 Heterogeneity, 2:298, 2:298f, 4:747
 Heteronuclear single quantum coherence (HSQC) spectrum, 2:610f
 Heterotrophic denitrification to N₂O, 1:267
 Heterotrophic nitrification, 1:255–256, 1:262
 Heterotrophic nitrifying bacteria (HNB), 1:262
 Heterotrophic protists, 1:39
 Hexaaquo aluminum ion, 2:170t
 Hexane, 2:427–428
 Hexanol, 2:427–428
 High Arctic soils
 classification and carbon stock, 4:270
 climate, 4:269
 soil formation, 4:270
 vegetation, 4:269
 High frequency electron paramagnetic resonance (HFEP) spectroscopy, 2:535
 metal binding, 2:538–539
 resolution advantage, 2:538f
 High Line Park, New York City, 3:490–491, 3:491f
 High-resolution magic angle spinning NMR, 2:610
 High-resolution soil moisture estimation, 4:625, 4:625f
 High-resolution TEM (HR-TEM), 2:577, 2:577f, 2:582–583, 2:583f
 High-throughput sequencing (HTS)
 amplicon-based sequencing, 1:152–153, 1:152f
 metagenomics (shotgun sequencing), 1:153
 scope and methodology, 1:151, 1:152f
 single-cell sequencing (SCS), 1:153
 Histels, 4:226
 Histogram, 4:489, 4:489f
 History of capillarity, 5:226, 5:226f
 History of water potential, 5:159–160
 Histosols, 4:55ge, 4:226, 4:284, 4:285f, 4:296, 4:385, 4:460–461
 formation, 4:386f
 hyper-humid and humid climates, 4:16
 Hjulström curve, 3:195, 3:195f
 Hollandite, 4:143
 Hollow curve, 4:731–732
 Homegardens of Jamaica, agroforestry, 3:75f
 Homogenization, pore-scale modelling, 5:31–33
 Honghe Hani Rice Terraces, 3:453
 Hooke's law, 2:590, 2:590f
 Horizons, 4:48ge, 4:150–152, 4:164–166, 4:165f, 4:404t
 Horizontal gene transfer, 1:31ge, 1:124
 Hormogonia, 1:274
 Hornblende, 4:104f, 4:105
 Hortonian runoff concept, 5:287–288
 Horvitz-Thompson estimator, 4:538
 Hot-spot habitats, 1:415
 Houston black smectite, 2:174f
 Ht-index, 4:736
 'Human-altered' and 'human-transported' soils (HAHT), 3:528
 Human and animal health, 3:507
 Human-caused soil change, 4:56
 Human health and soil, 3:560t
 antibiotics, 3:567–569
 clay, 3:569–570
 food security, 3:557–559
 ingestion, 3:562–563
 inhalation, 3:563
 medicines, 3:567–570
 nutrient deficiencies and health problems, 3:561t
 nutrition and trace elements, 3:559–562

- Human health and soil (*Continued*)
 skin lesions, soil absorption, 3:563–564
 soil contamination, 3:570
 soil history, 3:556–557
 soil organisms, 3:564–566, 3:564–565*t*
 soil pollutants, 3:566–567
- Human impacts on soil formation
 evaluating, 4:56–58
 scales and scope of, 4:55–56
 significance and future of, 4:67–68
- Human-induced climate change, 4:67
- Human-induced modifications on factors of evolution, SUITMAs, 4:237–240
- Human interactions with volcanic soils, 4:469
- Human-modified soils, 4:195*f*, 4:234
- Human–soil relationships, in human-affected areas, 4:233*f*
- Human use of forests and forest soils, 3:58–59
- Humectants, 3:389
- Humic acids, 2:553
 fluorescence polarization, 2:556, 2:557*f*
 fluorescence spectra, 2:553
 three-dimensional fluorescence contour map, 2:561*f*
- Humic free radicals, electron paramagnetic resonance spectroscopy
 factors affecting, 2:525–526
 interpretation, 2:525
 magnetic field, 2:524
 organic chemicals and metal ion interactions, 2:526
 spectral parameters, 2:524–525
 spin concentration, 2:525
 structural implications, 2:526
- Humic substances (HSs), 2:320, 2:359–361, 2:490–491, 2:492*f*
 functional groups of, 2:361
 and kerogen sorption, 2:321–323
 paradigm, 2:63
 plasticization of, 2:360–361, 2:361*f*
 rubbery and glassy domains of, 2:359–360, 2:360*f*
- Humid and sub-humid tropical climates, 4:16–17
- Humid and sub-humid tropical soils
 biotic diversity and interrelation, 4:357–358, 4:358–359*f*
 climate, 4:356, 4:356–357*f*
 delineation, 4:353–355
 factors governing, 4:355–358
 future research, 4:366
 human influences, 4:365–366
 humid tropical uplands, 4:363, 4:363*f*
 humid zone river soilscapes, 4:361–362, 4:362*f*
 lateritic weathering, 4:358–361
 lowlands, uplands and mountain ranges, 4:357
 old continents and young landscapes, 4:357
 parent material, 4:357
 relief, 4:357
 seasonal river soilscapes, 4:363–364
 sub-humid tropical uplands, 4:364, 4:364*f*
 time, 4:357
 young volcanic territories, 4:364, 4:365*f*
 zones, 4:353–355, 4:354*t*, 4:355*f*
- Humid cool temperate (HCT) soils, 4:289
 classification of, 4:296, 4:296*t*
 environmental characteristics, 4:290–291
 geographic scope, 4:290*f*, 4:290–291
 groundwater recharge, 4:292–293
 impacts of land use and management on, 4:297
 land use, 4:290–291
 temperature influences, 4:291
 wetlands in, 4:292–293
- Humidity effects, foliar nutrition application, 3:387–388
- Humid subtropical ecozone
 characterization, 4:343–344
 vegetation types, 4:344
 worldwide distribution, 4:344*f*
- Humification, 1:394
- Humins, 2:553
- Humods, 4:227
- Humults, 4:229
- Humus, 1:1*ge*, 2:58–60
- Hybridization-based method, 1:150
- Hybrid of scorpan and geostatistics, 4:478–479, 4:480*f*
- Hydration, 5:152, 5:152*f*
 chemical weathering, 4:39
- Hydraulic conductivity (HC), 4:373–375, 4:643, 5:6, 5:257, 5:260, 5:366.
See also Pedotransfer functions (PTFs)
- Hydraulic head and hydraulic gradient, 5:366
- Hydraulics, pore-size distribution, 5:20–21
- Hydric soils, 4:443
 Cowardin system, 4:424–425
 definition, 4:424
 field indicators, 4:432–433, 4:432*t*, 4:433*f*
 formation, 4:427–432
 iron and manganese-based morphological features, 4:429–430
 organic-carbon based morphological features, 4:428–429
 oxidation-reduction reactions, 4:427–428
 redoximorphic features, 4:430–432
 reduced soil, 4:428
 soil classification systems, 4:425–427
 taxonomic groups, 4:426*t*
 technical standard, 4:433–434, 4:434*t*
 upper part, 4:424
- Hydrochar, 2:433
- Hydrodynamics
 steady-state flow through saturated medium, 5:257–258, 5:258*f*
 steady-state flow through unsaturated medium, 5:260
 transient flow through saturated medium, 5:258–259, 5:259*f*
 transient flow through unsaturated medium, 5:261
- Hydrogels, 3:118
- Hydrogen (H), 2:215
- Hydrogen bonding, organo-mineral interactions, 1:390
- Hydrological fluxes, 5:478
- Hydrological functions, biocrusts, 1:134
- Hydrological processes, water harvesting and, 5:595–597
- Hydrologic cycle and surface energy balance, 5:491, 5:491*f*
- Hydrologic process, 5:471–472
- Hydrology
 of groundwater, 5:365–370
 wetlands, 4:444
- Hydrolysis reaction, 2:169–170, 2:170*t*
 chemical weathering, 4:38
- Hydrometer method, 5:37, 5:38*f*
- Hydromorphic features, 4:4
- Hydrophilic soil conditioners, 3:203
- Hydrophilic solutes, cuticular penetration, 3:385–386
- Hydrophobic interactions
 colloid transport, 5:340
 organo-mineral interactions, 1:390
- Hydrophytic plants, 1:381
- Hydroxide minerals, 2:250
- Hydroxy-interlayered vermiculite (HIV), 4:350
 and smectite, 4:117
- HYDRUS-2D and HYDRUS-3D model, 2:481
- Hyperfine splitting interaction, EPR signal, 2:522
- Hyphenated NMR, 2:611–612
- HYPROP benchtop evaporation system, 5:199–200
- Hysteresis
 in capillary behavior, 5:233–235, 5:238*f*
 soil water characteristic (SWC) curve, 5:197, 5:198*f*
- Hysteresis indices (HIs), 2:357–358

I

- IAA production, 1:422–423
- Ice segregation, 5:414
- iGEM (International Genetic Engineered Machine) Exhibition, 3:512, 3:512*f*
- Illite, 4:100–101
- Illuviation, 4:292
- Imitated natural marshland in Southeastern Louisiana, 3:492–493
- Immiscible fluid
 Darcy's law, 5:279–280
 field-scale simulations, 5:280
 non-aqueous phase liquids (*see* Non-aqueous phase liquids (NAPLs))
 pressure, matric head, 5:280
 relative permeability, 5:280–281
 three-phase relative permeability, 5:281–283
- Immobilization, 2:142*ge*, 2:145, 2:248–252
 of inorganic sulfur, 2:195
 of toxic metals, 3:418–419, 3:418*f*
- Imogolite, 2:174*f*, 4:119, 4:121, 4:122*f*, 4:458*ge*. *See also* Short-range ordered (SRO) aluminosilicates
- Impedance sensors, soil moisture, 4:593
- Inactivation viruses, 1:576
- Incentive-based instruments, for soil protection, 3:552–553
- Incentive-based policy instruments, 3:547*ge*

- Inceptisols, 4:55*ge*, 4:229, 4:296, 4:327
- Incineration
with energy recovery, 3:328
municipal solid waste, 3:330
- Incoming longwave radiation, 5:397
- In-crop weed management, 3:445–447
- India
soil quality protection laws and policies, 3:470–471, 3:470*f*
waste management laws and programs, 3:469*f*
- Indicator and disjunctive kriging, 4:669
- Indicator of Reduction In Soil (IRIS) tubes, 4:445
- Indigenous soil knowledge, 4:186–188
- Indigenous technical knowledge (ITK), 3:394
- Indirect labeling approach, 2:104–105
- Induced polarization (IP), 5:451–452, 5:451*f*
- Induced systemic resistance, 1:424
- Induction, 2:317
- Infiltration, 1:571–572
agriculture-related fields, 5:647
excess, 5:289–291
macropores, 5:420–422
minimal surface ponding, 5:646–647
perspectives, 5:646
process, 5:646
snowcover, 5:422
soil water repellency (SWR), 5:209–210
steady-state infiltration, 5:646–647
transient infiltration (*see* Transient infiltration)
- Infiltration rate (IR), 4:376
- Infochemicals, 1:361–362
- Information asymmetry, 3:547*ge*
- Infrared radiation, soil algae, 1:66
- Infrared spectral analysis, 2:511–512, 2:512*f*.
See also Fourier transform infrared (FTIR) spectroscopic method
- Infra-red transmitting mulches (IRT), 3:397
- Inherent water use efficiency (IWUE), 5:529
- Inner-sphere complex (ISC), 2:152*ge*, 2:339, 2:368*ge*, 2:491, 2:492*f*
- Inoculants. *See* Microbial inoculants
- Inoculation
with mycorrhizal fungi, 1:453–454
of plant growth promoting rhizosphere bacteria and archaea, 1:452–453
reliability of, 1:454
of rhizobia, 1:449–452
- Inorganic carbon, 2:58, 2:59*f*
- Inorganic C forms, in soil, 2:263–265
- Inorganic contaminants in soil, 2:506–507
- Inorganic fertilizers, application of, 3:108
- Inorganic mulch, 3:396–398
- Inorganic sulfur, immobilization of, 2:195
- Inositol hexakisphosphate (IHP), 2:347–348, 2:349*f*
- Insecticides, 2:471
- Insensitive nuclei enhancement by polarization transfer (INEPT), 2:618
- In situ chemical oxidation (ISCO) treatment, 3:339, 3:340*f*
- In situ chemical reduction (ISCR) technique, 3:339, 3:340*f*
- Instability
in Hele-Shaw cell, 5:316
pattern development, 5:317
for wetting front, 5:316
- Installations, lysimeters, 5:677–678
- Institutions, 3:547*ge*
- Instrumentation and wave-guides, time-domain reflectometry, 5:440–441, 5:441*f*
- Integral suspension pressure (ISP) method, 5:37–38
- Integrated crop-livestock systems (ICLS), 1:225, 3:126–127
- Integrated disease management, 1:440, 1:443
- Integrated farming system, 3:121
- Integrated nutrient management, 3:119–120
- Integrated Water and Environment Management (IWEM), 5:506
- Integrated weed management (IWM), 3:441, 3:445–446, 3:446*f*
- Intensity-capacity concept, 5:301
- Intensive management of forest and forest soils, 3:59–60
- Inter-aggregate pores, 5:18
- Intercalation, organo-mineral interactions, 1:390
- Intercropping, 3:72–74, 3:73–74*f*, 3:120, 3:120*f*, 3:448
- Interdisciplinary soil studies, 3:29
- Interference competition, 1:444
- Intergovernmental Panel on Climate Change (IPCC), 2:430
- Inter-granular pores, 5:18
- Intermediate complexity model, 1:178–180, 1:179*f*
- Intermediate SPAC models, 5:628–629
- Intermediate transitions, 2:299
- International Committee on Taxonomy of Viruses (ICTV), 1:17
- International customary law, 3:536
- International level soil governance
food security, 3:538
global soil partnership, 3:538
international customary law, 3:536
organizational entities, 3:536
recommendation, 3:538–539
soft law, 3:536
sustainable development goals (SDGs), 3:536–537
sustainable soil management, 3:538
treaty provisions, 3:537
- International soil classification system, 4:206
- International Soil Modeling Consortium (ISMC), 4:753*ge*
- International Union of Soil Sciences (IUSS), 3:574, 4:1–2
- Interpolation error, 4:682
- Interstratified phyllosilicates, 4:117–118
- Intra-class correlation, 4:495–496
- Intraorganic matter diffusion (IOMD), 3:303–304
- Intrazonal soils, 4:14–15, 4:190, 4:248
- Intrinsic surface-charge density, 2:368*ge*
- Intrinsic thresholds, 4:77
- Inundation, 4:751*ge*
- Invasive species, 1:14, 3:97
- Invertebrates, fire effects, 2:454
- Iodine (I), 2:235
cycle, 3:562*f*
deficiency, 3:561
human health and soil, 3:561
in soils, 2:236, 2:239–240
- ¹²⁷Iodine, 2:236
- Ion chromatography, root exudates, 1:353
- Ion exchange
chemical weathering, 4:39
organo-mineral interactions, 1:388
- Ionic strength, anthropogenic organic compounds, 2:365
- Ionizable organic compounds (IOCs), 3:420
- Ion product of water, 5:151
- Iron (Fe), 4:293
aqueous media, 2:183–184
azonal soils, 2:180–182
biogeochemical marker of hydromorphic soils, 2:181–182, 2:182*f*
human health and soil, 3:559–561
as indicator of aridity/humidity, 2:179–180, 2:180*f*
in situ dynamics of iron minerals, 2:186
on landscape evolution, 2:186, 2:187*f*
lateritic formations, typology of, 2:177–179
living organisms, 2:187–190
mineral formation, pathways of, 2:182–186
natural forms of iron oxides, 2:182
oxide formations, 2:177–182
plant system, 2:188
processes *vs.* soil properties, 2:181*t*
properties of chemical element, 2:177, 2:178*f*
rhizosphere, 2:188
sequestration, 1:422
siderophores, 2:188
size of, 2:184–185
surface properties of iron “oxides”, 2:185–186
typology of hydromorphic soils, 2:180–181, 2:181*f*
up-take by organisms, 2:188–190, 2:189*f*
zonal soils, 2:177–180
- Iron and aluminium oxi(hydr)oxides, 2:152*ge*
- Iron and manganese-based morphological features, hydric soils
gley soil colors, 4:430
redox concentrations, 4:430
redox depletions, 4:430
reduced matrix, 4:430
- Iron-deficiency anemia (IDA), 3:559–561
- Iron masses, 4:430
- Iron-only nitrogenase, 1:271
- Iron oxides, 4:119, 4:329
akaganéite, 4:141
crystal structures, 4:137, 4:138–140*f*
dissolution, 4:137
ferrihydrite, 4:140
formation and transformation pathways, 4:141*f*
fougerite, 4:141–142
goethite, 4:137–139
hematite, 4:140

Iron oxides (*Continued*)

- lepidocrocite, 4:140
- magnetite and maghemite, 4:141
- reducing soil conditions, 4:136–137
- schwertmannite, 4:141
- in soil environments, 4:137*t*
- Irragic horizon or soil, 4:54*ge*
- "Irragic" soils, 4:59–60
- Irreversible thermodynamics, 5:264
- Irrigation, 5:608
 - advantages and disadvantages, 5:622*t*
 - choosing, 5:621
 - design and management, 5:609
 - development projects, 3:266–267
 - environmental effects of, 3:268
 - metals, 3:273
 - pesticides, 3:272–273
 - plant nutrients (fertilizers), 3:271–272, 3:272*f*
 - pollutants attributable to irrigation, 3:271–273
 - public health concerns, 3:273–274
 - sustainability, 3:274
 - waterlogging and salinity, 3:270, 3:271*f*
 - water shortage, 3:268–270, 3:268–269*f*
 - fertilization, 5:620–621
 - of grassland soils, 3:109
 - micro-irrigation, 5:609, 5:618–619
 - on-farm application efficiencies, 5:610*t*
 - salinity hazards and management, 5:621
 - sprinkler, 5:609, 5:613–618
 - subirrigation, 5:620
 - surface, 5:609–613
 - uniformity and efficiency, 5:610
- Irrigation water requirement (IWR), 5:572, 5:590–591, 5:591*t*
- Isothermal microcalorimetry (IMC)
 - heavy metals and metalloid toxicity, 2:665
 - isothermal microcalorimeter, 2:663
 - pesticide 2,4-dichlorophenoxyacetic acid decomposition, 2:665
 - soil microbial activity and biomass, 2:663–665
 - toxicity monitoring, 2:665
- Isothermal titration calorimetry (ITC)
 - cation exchange, 2:666
 - enthalpy changes, 2:666–667
 - heavy metal adsorption, 2:666
 - mineral–microorganism interfacial reaction, 2:666
 - phosphate sorption, 2:666
 - reaction kinetics and thermodynamic parameters, 2:665–666
- Isotope exchange kinetics (IEK) approach, 2:100–102, 2:101*f*
- Isotope incorporation techniques, 1:191
- Isotopes
 - analytical methods, 2:84–86
 - carbon, 2:86–88
 - definition of, 2:82
 - effect and isotopic fractionation, 2:83–84
 - fractionation in numbers, 2:83*t*
 - Isotope phi*, 2:84*b*
 - of light elements in soil science, 2:86–91
 - natural abundance, 2:83

- natural abundance and labelling
 - approaches, 2:86
 - in soil science and relative international standards, 2:82*t*
- Isotope tracing methods, root exudates, 1:354
- Isotopic dilution principle
 - in soil–plant system, 2:103
 - in soil–solution systems, 2:98–102
- Isotopic fractionation, 1:161, 2:83–84

J

- Jablonski energy diagram, 2:545*f*
- Jenkinson, David Stewart, 5:684–685
- Jorgensen's oxidizability, 2:301

K

- Kaolinite, 2:174*f*, 2:267*f*, 2:582, 2:582*f*, 4:114, 4:115*f*
 - acrisols and alisols formation, 4:350
- Kaolinite–water interface of periodic model, 2:493*f*
- Karst hypothesis, 4:361
- Kastanozems, 2:264–265, 4:307, 4:310–311*f*
- K_c adjustment for climate, 5:580
- K_{cb} for unknown conditions, 5:582–584, 5:583–584*f*, 5:585*t*, 5:586–587*t*, 5:588–590*t*
- Kelvin probe force microscopy (KPFM), 2:595–597
- Keratin, 1:524
- K-feldspar, 4:97–98
- K fertilizers, 3:371
- Kinetic definition, chemical equilibrium, 2:275
- Kinetic energy, 5:159
- Kinetic isotope fractionation, 2:108
- Kinetic limitation, 4:79
- Kinetics relevant to soil systems, 2:344–346, 2:345*f*
- K-means clustering
 - application, 4:505, 4:506–507*f*
 - precision agriculture and sampling site selection, 4:505–506, 4:506–507*f*
 - spectroscopy, 4:505
 - theory, 4:505
- Knowledge-based inference, 4:688–689
- Kossovich's soil classification, 4:190
- Kriging, 4:510, 4:525, 4:668–669
 - with external drift, 4:669
 - indicator and disjunctive, 4:669
- K-strategists, SOC mineralization, 2:394
- Kubelka-Munk equation, 2:514

L

- Labeling technique, 2:86, 2:104–105, 2:104–105*f*
- Laboratory sample handling
 - archiving, 4:566, 4:567*f*

- contamination and traceability check, 4:562
- crumbling, 4:563
- diagonal method, 4:564, 4:565*f*
- drying, 4:563, 4:563*f*
- from organic horizons, 4:564
- quartering method, 4:564, 4:565*f*
- reception and registration of samples, 4:561
- sample storage, 4:566, 4:567*f*
- sample sub-setting, 4:564
- sending sample for analyses, 4:564–566, 4:566*f*
- sieving, 4:563, 4:564*f*
- soil preparation, 4:562–564
- weighing, 4:562, 4:562*f*
- Lagrangian methods, colloid transport, 5:350
- Land application of effluent, from wastewater treatments systems, 1:570–571
- Land capability, 3:182
- Land degradation, 3:344, 4:86, 4:87*f*
- Land degradation neutrality (LDN), 3:536–537
- Land drainage, 3:257–258. *See also* Drainage
- Land evaluation, 4:481, 4:482*f*
- Landfarming
 - activities, 1:533–534
 - under aerobic conditions, 1:534–535
 - bacteria, 1:538–539, 1:539*f*
 - bioaugmentation, 1:538
 - biostimulation, 1:537, 1:537*f*
 - climatic condition, 1:536
 - concentration of contaminants, 1:536
 - contaminant chemical properties, 1:536, 1:536*f*
 - cost-effective biological treatment, 1:535
 - dehydrogenase indices, 1:537*f*
 - efficacy, 1:542, 1:543–544*f*
 - equipment, 1:534
 - ex-situ process, 1:533–534
 - extremozymes, 1:538–539
 - factors to be considered, 1:534–535
 - field-scale landfarming, 1:540*t*
 - fungi, 1:539, 1:540*t*
 - intermediate products, 1:535
 - life cycle assessment (LCA), 1:542
 - microbial activities, 1:533–534
 - microbiological features, 1:538
 - molecular microbial diversity and abundance, 1:541–542, 1:541*f*
 - nanozymes, 1:538–539
 - oil spills, 1:537
 - organic contaminants, 1:533–534
 - physicochemical processes, 1:535
 - scheme/typical set-up, 1:534, 1:534*f*
 - soil properties, 1:535–536
 - treatability studies, 1:542
- Landfill gas (LFG) utilization, municipal solid waste, 3:327, 3:327*f*
- Landfill, municipal solid waste, 3:329–330
- Landforms, of alluvium, 4:419–420
- Land leveling, 3:118
- Land management practices, 3:200
- Land-Potential Knowledge System (LandPKS), 4:697
- Landscape, 3:506–507, 4:742–743

- evolution, 4:418
 evolution models, 4:748
 planning
 arid landscapes, 3:582
 carbon dioxide emission reduction, 3:584–586
 land-use layer, 3:584
 plant-soil interaction, China, 3:586
 political and economic life, 3:581
 socio-ecological dimension, 3:583
 soil erosion reduction, Blaikie's model, 3:584, 3:585f
 soil taxonomy, 3:581
 traditional map overlay method, 3:581, 3:582f
 tropical soils, 3:582
 urban soil, 3:583–584
 values, 3:580
 stability, 4:55
 water requirements, 5:584–589
 Land surface energy balance
 components, diurnal variations of, 4:614f
 and evapotranspiration, 4:611, 4:611f
 Land surface models (LSMs), 4:753ge
 biogeochemical cycling models, 4:757
 coupled model environments, 4:760
 deep soil, 4:757–758
 extreme environments, 4:758–760
 global-scale, 4:753
 hyperresolution, 4:760
 lower soil profile, 4:754
 semi-permeable/impermeable layers, 4:755–756
 soil heat balance, 4:758
 soil moisture model components, 4:756–757
 spatially-explicit soil simulation model, 4:753–754
 standard conceptual levels, 4:754f
 thermal dynamics models, 4:758
 upper soil profile, 4:754
 Land surface temperature (LST)
 proximal, 4:601–605
 soil microbial activity and controls, 4:600
 space models, 5:503–504
 thermal radiation, 4:600–601
 Land-treatment processes, slow-rate process, 1:570f, 1:571
 Land use, 3:247–249, 3:249f, 4:338–340, 4:339f
 approach, 3:69
 change, 1:470–471, 1:477
 history, 4:240
 and land cover
 water induced soil erosion, 3:199–200, 3:200f, 3:201t
 Land use change, 1:14
 carbon sequestration, 3:428
 and land cover changes, 3:362–363
 and management, 3:366
 Langmuir adsorption, 2:24, 2:25f
 Lanthanides
 abundance, 2:226t
 applications, 2:226–227
 concentration, 2:227
 contraction, 2:225–226
 geochemical fractionation studies, 2:228
 global demand for, 2:227f
 hyperaccumulators, 2:228
 mineral resources, 2:227
 mobility, 2:228
 soil profile, 2:228
 technological applications, 2:227–228
 Lanthanoid, 2:214–215
 Laplace transform, 5:662
 Laser diffractometry (LDM), 5:38, 5:39f
 Laser-induced breakdown spectroscopy (LIBS), 2:655, 2:655f
 advantages, 2:655
 vs. conventional methods, 2:660, 2:660f
 disadvantages, 2:655–656
 limitations, 2:660, 2:660f
 organic carbon content, 2:658–659
 plant nutrients, 2:656t, 2:657
 precision and digital agriculture, 2:656
 soil characteristics, 2:654–655
 soil physicochemical properties, 2:659–660, 2:659t
 steps for soil analysis, 2:655, 2:655f
 toxic elements, 2:657–658, 2:658t
 vs. traditional methods, 2:655
 Laser induced fluorescence spectroscopy (LIFS), 2:546, 2:549–550
 Latent heat, 5:625
 Latent heat flux, evapotranspiration, 4:611–612, 4:614, 4:614t
 Latent heat of vaporization, 5:388
 Lateritic weathering
 characteristics, 4:359, 4:360f
 description, 4:358–359
 elements and minerals distribution, 4:359–361, 4:360f
 hypotheses, 4:361
 idealized laterite profile, 4:359, 4:359f
 sequences, 4:359, 4:360f
 Latin-American Soil Science Society (SLCS), 3:574–575
 Latin hypercube sampling, 4:546
 Lattice Boltzmann models (LBM), 5:196, 5:196f
 Lawn sprinkler systems, 5:618
 Layered double hydroxides (LDHs)
 abiotic nitrate reduction, 2:309–310, 2:309–310f
 auto-organization of reactions, 2:310–311, 2:311f, 2:312–313f
 structure, 2:307–309, 2:307t, 2:308f
 Leaching, 2:236
 biophysical mechanisms, 5:322
 of bromine, 2:239
 chemical exchange, 5:322
 of chlorine, 2:238
 of dissolved solutes, 4:292
 drainage quantity, 5:323–324
 environment, 4:3, 4:3f
 field measurement devices, 5:327
 of fluorine, 2:236
 leachate quality, 5:324
 mechanistic models, 5:333
 modelling and risk assessment, 5:330–332, 5:331f
 nitrogen, 5:327–328
 nitrogen measurements, 5:333
 pesticide leaching, 5:332
 preferential leaching, 5:324–326
 and redistribution of calcite and silica, 4:323–324, 4:323f
 salts, 5:322, 5:328–330
 spatial upscaling, 5:333
 Leaching fraction (LF), 5:328–330
 Lead (Pb)
 biological function, 2:219
 concentration of soil, 2:221t
 contamination of soil, 2:220
 pollution, 2:113
 in roadside soil, 2:506–507
 Leaf area index (LAI), 5:529
 Leaf transpiration, 5:627–628
 Le Chatelier's principle, 2:276
 Legacy soil data, 4:691–692
 Leonardite humic acid (LHA), 2:491
 Lepidocrocite, 4:16, 4:140
 Leucization, 4:347, 4:347f
 Lichens, 1:128
 Life cycle assessment (LCA), 1:542
 Life strategy
 biokinetic diversity, 1:207–208
 concept, 1:204–205
 growth recalcitrance, 1:208
 illustration, 1:202f, 1:204–205
 K-selected organisms, 1:206–207
 L-selected organisms, 1:207
 r-selected organisms, 1:206
 survival triangle, 1:205, 1:205t, 1:206f
 types, 1:204
 Ligand, 2:168
 complexation by, 2:245
 formation constants of Al with, 2:175
 inorganic, 2:169–171
 organic, 2:171–172
 reactive, 2:174
 Ligand exchange, 1:388
 Ligand-promoted mineral dissolution, 2:566
 Lightning, 2:143
 Light non-aqueous phase liquids (LNAPLs), 5:273
 Lignin, 1:524, 2:389, 2:439
 Lime-induced iron chlorosis, 2:14
 Lime on grassland soils, 3:109
 Limiting hydraulic conditions, 5:511–512
 Linear discriminant analysis (LDA)
 application, 4:504, 4:504t
 random forests, 4:505
 theory, 4:503–504
 Linear flow models, 5:241–242, 5:241t
 Linear model of co-regionalization (LMC), 4:528–529, 4:530f
 Linear-move systems, 5:617
 Linear regression model, 4:510
 Linear superposition approximation (LSA), 5:339
 Linuron, 5:332
 Lipids, pyrogenic carbon (PyC), 2:439
 Liquid and water vapor
 governing equation, 5:387
 transport mechanisms, 5:386–387
 Liquid chromatography, of root exudates, 1:353

- Liquid retention and pore shape, 5:189, 5:189f
- Liquid-solid-gas contact angles, 5:228f
- Liquid-vapor interfacial tensions, liquids, 5:227t
- Liquid water flux, 5:388
- Liquid water molecule model, 5:147–148b, 5:148f
- Literacy, soil, 3:506
- Lithiophorite, 4:142
- Lithium (Li), 2:109, 2:117
- Lithobiont microbial communities, 1:505
- Lithology and relief, 4:322
- Litterfall and forest floor, 3:60–61, 3:61f
- Liverworts, 1:128
- Livestock management, rangeland
domestic livestock, 3:90–91
fire, 3:95
grazing system, 3:94–95
monitoring strategies, 3:96
multiple species grazing, 3:92
stocking rate, 3:92–93
water supply, 3:95
wildlife, 3:92
- Livestock manure, 1:559, 1:561, 1:564
- Living mulches, 3:447
- Local endemism, 1:42–43
- Locard Exchange Principle, 3:523
- Loess, 4:42, 4:43f
distribution, 4:44
dust, soil, (paleo)environments, 4:46–47
genesis, 4:43–44, 4:43f
properties, 4:44–45
temperate-zone loess soils, 4:45–46, 4:45f
- Loess-Paleosol-sequence, 4:248f
- Logistic growth, 1:175
- London-van der Waals forces, 2:317
- Longitudinal dispersivity, solute transport, 5:655
- Long-Term Ecological Research network (LTER), 5:632
- Long-term tillage, 3:435f
- Low Arctic soils
classification and carbon stock, 4:268
climate, 4:267
soil formation, 4:267
vegetation, 4:267
- Lower cost-miniaturized spectrometer, 2:573
- Low field NMR, 2:611–612
- Low-molecular-weight organic acids (LMWOA), 2:52
Al with, 2:171t
deprotonated forms of, 2:172f
- Low-molecular-weight organic substances (LMWOS), 2:389
- Low-quality water (LQW) irrigation, 5:124–125
- Lucas-Washburn-Rideal (LWR) equation, 5:233
- L-value, 2:103, 2:103f
- Lysimeters, 5:327
advantages and disadvantages of, 5:674t
with basement and control plot, 5:676f
characteristics, 5:669t
circular, 5:668f
controlled lower boundary by suction rake, 5:673f
cutting method, 5:670, 5:670f
in field with same vegetation, 5:673f
filling of, 5:671f
as “gold standard”, 5:678
heat transport, 5:675–676
history of, 5:668
installations, 5:677–678
limitations, 5:669–674
measurements, 5:676–678
measurements and work in, 5:677f
parameters and variables measured in, 5:674–676
plant (above and below ground), 5:676
site conditions, 5:677
solute transport, 5:675
TERENO-SOILCan, 5:673f
types, 5:669–674
uses, 5:668
vertical filling of soil monoliths, 5:672f
water transport, 5:674–675
- Lysogeny, 1:507
- ## M
- Machine learning (ML), 4:479, 4:686–687, 4:694
artificial neural networks (ANNs), 4:514
challenges, 4:519
and classical statistics, 4:510
definition, 4:510
feature engineering and data fusion, 4:517
geostatistics and, 4:675–676
large data processing, 4:517–518
methods, 4:677
parametric, 4:510
pedotransfer functions (PTFs), 4:649
reinforcement learning, 4:510–511
sensors, 4:519
spatial and spatio-temporal, 4:518
strategies, 4:512–514
supervised learning, 4:510–516
transfer functions, 4:518–519
unsupervised learning, 4:510–511, 4:516–517
- Macroaggregates, 1:390
- Macro- and mesofauna, manure management, 1:562–563
- Macrofauna, 1:234
- Macronutrients, 1:1ge, 1:310, 1:458t
- Macropore flow, 5:306ge
- Macropores, 1:5, 5:1, 5:306ge
drainage flow, 5:306–307
hydropedological approaches, 5:312
and macropore flow
modeling approaches, 5:311–312
network characteristics, 5:308–311
non-equilibrium flow pathways, 5:308
non-equilibrium water flow, 5:308
and preferential flows, 5:307–308
soil structure formation mechanisms, 5:307
surface-vented macropores, 5:307
surface water quality, 5:308
- Mafic rocks, 4:346
- Maghemite, 4:141
- Magnesium (Mg), 2:132, 2:214–215, 2:264
- Magnetic force microscopy (MFM), 2:595–597
- Magnetic resonance imaging (MRI), 2:611–612, 2:613f, 5:464–466
- Magnetic soil mapping, 5:449–450
- Magnetite, 4:141
- Magnetometry, 5:449–450
- Main soil-forming processes, in SUITMAS, 4:241–242
- Malnutrition, 3:13, 3:17
- Manganese (Mn)
concentrations from wind erosion of dust roads, 3:231f
respiratory intake of, 3:233f
toxicity in legumes, 1:466–467
- Manganese oxides, 4:120, 4:142–143, 4:143t
chain structures, 4:143
phyllomanganates, 4:142–143
tectomanganates, 4:143
- Manganese(III) phosphate, 2:281
- Mangrove swamps, 4:406
- Manure deposition, in wrong place, 5:292f
- Manure management
abundance of soil bacterial phyla with, 1:563t
anaerobic digestion, 1:561
composition as affected by storage and treatment, 1:559–561
composting, 1:561
denitrification, microbes involved in, 1:563–564
macro- and mesofauna, 1:562–563
nitrification, microbes involved in, 1:563–564
nitrogen fixation, microbes involved in, 1:563–564
nutrients and management details in animal, 1:560t
production and application, 1:559
soil borne pathogens and antibiotic resistant genes, 1:564–565
on soil conditions and biodiversity, 1:561–565
storage, 1:559–560
treatment, 1:560–561
- Marginal value, 3:547ge
- Markers, human activities in SUITMAS, 4:234–235
- Market failure, 3:547ge
- Mass balance methods, soil water content monitoring, 5:474–475
- Matheronian geostatistics, 4:478
- Matric potential, 5:161, 5:164–166, 5:169–170, 5:172f, 5:175, 5:518
sensors, 5:185t
soil water retention, 5:188
- Maxwell relations, 5:264–265
- Maxwell-Stefan diffusion equation, 5:249
- Meal Culture, 3:516
- Mean, 4:489
- Mean annual soil temperature (MAST), 4:251–252
- Measurement error, 4:672–674, 4:682

- Measurements
 lysimeters, 5:676–678
 and work, in lysimeters, 5:677f
 Mechanical dispersion, 5:355, 5:357f
 Mechanical shaking methods of soil dispersal, 2:72–74
 Mechanical weeding, 3:446
 Mechanistic modelling, digital soil mapping (DSM), 4:688–689
 Mediterranean soils
 chemical properties, 4:329–330
 mineralogical properties, 4:329
 physical properties, 4:328
 Melanization, acrisols and alisols formation, 4:347, 4:347f
 Mendeleyev's classification, 2:214–215
 Mercury, 3:289t, 3:292
 Mercury porosimetry, 5:21
 Mesopores, 5:1
 Mesopotamia, 3:32
 Metabarcoding, 1:31ge
 Metabolism, of sulfur, 1:304–305
 Metabolomics, 1:165
 Metagenomics, 1:18, 1:31ge, 1:144–145t, 1:153, 1:541
 Metal and metalloids, 3:283–284, 3:283t
 Metal extraction and analysis
 diffusive gradients in thin films (DGT), 2:55
 sequential extraction, 2:53–55
 single extraction, 2:51–53
 Metalloid elements, 2:214–215
 Metal–organic frameworks (MOFs), 2:503–504
 Metal oxides, 2:250
 aluminum oxides, 4:143–145
 basic structural unit, 4:136
 iron oxides, 4:136–142
 manganese oxides, 4:142–143
 phyllosulfates, 4:142–143
 silicon oxides, 4:145–146
 titanium oxides, 4:146–147
 Metals
 environmental effects of irrigation, 3:273
 properties of, 2:337, 2:338t
 and trace elements, wastewater treatments systems, 1:574, 1:574t
 Metals and organic contaminants
 remediation
 approaches, 3:334, 3:334t
 bioremediation, 3:336–337, 3:337f
 capping, 3:334, 3:335f
 categories, 3:334
 chemical reduction, 3:339, 3:340f
 chemical stabilization, 3:336
 conventional technologies, 3:334
 electrokinetic remediation, 3:341, 3:341f
 engineered containment cells (ECCs), 3:335, 3:335f
 human activities, 3:333
 in situ chemical oxidation (ISCO), 3:339, 3:340f
 phytoremediation, 3:337–338
 soil flushing, 3:339, 3:339f
 soil vapor extraction (SVE), 3:341, 3:342f
 soil washing, 3:338–339, 3:338f
 stabilization/solidification (S/S), 3:335–336
 sustainable risk-based approach, 3:334
 thermal treatment, 3:341–342
 Metal surface precipitation, 2:340–341
 Metaproteomics, 1:165, 2:404
 Metatranscriptomics, 1:164–165
 Methane (CH₄)
 anaerobic oxidation of, 1:264
 efforts to reduce, emissions, 3:155–158
 emissions, 2:693–694, 4:456
 in soil, 2:263
 soil sink for, 3:155–158
 soil sources, 3:154–155, 3:154t, 3:155–157f
 sources and sinks, 3:154
 Methane monooxygenase (MMO), 1:548–549
 Methanogens, 1:27–28, 1:27t, 1:385
 Methanotrophs, 1:385, 2:694
 Method of known addition, 2:646
 Micas, 4:92, 4:115–116
 dioctahedral and trioctahedral, 4:100, 4:101t
 Fe contents, 4:101
 structure, 4:100f
 weathering, 4:101–102
 Michigan Microwave Canopy Scattering (MIMICS) model, 4:622–623
 Microaggregates, 1:390, 2:493
 Micro- and meso-fauna, 1:234
 Microarrays, 1:150
 Microbe–OM–mineral interfacial reactions
 adhesion, 2:459
 cell-to-mineral attachment mechanism, 2:461–462
 coulombic interaction, 2:461–462
 element cycling, 2:465–466, 2:466t
 examples, 2:458, 2:459f
 extracellular compounds and extracellular polymeric substances (EPS), 2:460–461, 2:463
 forces involved, 2:462, 2:463f
 functional group chemistry, 2:461
 future research, 2:467–468
 impacts, 2:463
 microbial cell surface structure, 2:459–460, 2:460f
 microbially-mediated mineral weathering, 2:464–465, 2:465f
 multiple binding mechanism, 2:462
 non-coulombic interaction, 2:462
 soil aggregate formation, 2:466–467, 2:467f
 Microbes, 1:33
 exploitation of indigenous, 1:454
 in farming systems, 1:464
 farming systems and agronomic management practices, 1:454–464
 groups of, 1:449t
 indigenous arbuscular mycorrhizal fungi, 1:452f
 inoculation, 1:449–454
 land degradation, 1:452t
 management in context of soil, 1:449
 vs. minerals, 2:427–430
 nutrient management, 1:456–463
 potential approaches to soil microbe management, 1:449–454
 potential impacts on, 1:455t
 pre-requisites for ensuring appropriate levels of biological and functional diversity in, 1:454–467
 rhizosphere microorganism groups, 1:450–451t
 tillage systems, 1:454–456
 Microbial activity, soil water repellency (SWR), 5:212
 Microbial biokinetics-general principles
 absolute growth rate, 1:170–171
 amount of microorganisms, 1:170–171
 anabolic substrates, 1:171
 apparent growth rate, 1:171
 catabolic substrates, 1:171
 cell quota model, 1:173, 1:173f, 1:176
 chemical and enzyme kinetics, 1:174–175, 1:174t
 conserved substrates, 1:171
 description, 1:169
 deterministic model, 1:170
 end products, 1:173
 energy overflow, 1:172
 environmental factors, 1:180–184
 exponential growth, 1:175
 genome-scale models (GEM), 1:176–178, 1:177f
 growth parameters, 1:170, 1:170t
 growth yield, 1:171
 intermediate complexity model, 1:178–180, 1:179f
 intermediates, 1:173
 logistic growth, 1:175
 maintenance requirements, 1:172, 1:172f
 microbial biomass, 1:170–171
 modifying factors, 1:180
 Monod model, 1:175–176
 near-zero growth (NZG) kinetics, 1:183–184, 1:183f
 nutrient substrates, 1:171
 nutritional factor, 1:180
 pH effect, 1:180–182, 1:181f
 products, 1:173
 specific growth rate (SGR), 1:171
 specific rate of substrate consumption, 1:171
 state variables, 1:170, 1:170t
 stochastic model, 1:170
 structured models, 1:176
 temperature effect, 1:182–183, 1:182f
 wasteful respiration, 1:172
 Microbial biokinetics-soil biology
 cell-division cycle analysis, 1:189–190
 characteristic microbial habitats, 1:200, 1:201f
 colonial growth, 1:201–202, 1:202f
 cryptic microbial growth, 1:196–197, 1:196t
 decomposable organic compounds (DOC) input to soil, 1:197
 decomposition kinetics, 1:208–215
 digestive track of invertebrates, 1:190
 diversity of soil microhabitats, 1:200–204

- Microbial biokinetics-soil biology
(*Continued*)
DNA synthesis rate, 1:191
factors to be considered, 1:216
future research, 1:215–216
growth dynamics, 1:202–204, 1:202t, 1:203–204t
growth stoichiometry, 1:195
habitats classification, 1:200, 1:200t
heterotrophic CO₂ fixation, 1:191
IBP renaissance, 1:192
isotope incorporation techniques, 1:191
lesson from kinetic studies, 1:216
life strategy concept, 1:204–208
maintenance requirements, 1:196
microbial growth rate, 1:216
microbial periodicity, 1:192
microbial productivity, 1:192–193, 1:194f, 1:195
nonreplicating genetic markers, 1:190
on-site direct microscopy, 1:189
periodic fluctuations, 1:193, 1:194f
population dynamics, 1:200
protein synthesis rate, 1:191
pulse-chase method, 1:197, 1:198f, 1:199–200
respiration rate, 1:191–192
seasonal organic matter decomposition, 1:195–200
SGR calculation, 1:193, 1:194f, 1:195
soil complexity, 1:215–216
soil decomposition rate, 1:195
steady-state populations, 1:190–191
stopping microbial grazing, 1:191
true vs. apparent growth rates, 1:189
water streams, 1:190
- Microbial biomass, 2:422
- Microbial communities-culture-dependent approach
challenges, 1:140
community-level physiological profiling (CLPP), 1:142
culturomics, 1:143, 1:144f, 1:144–145t
diffusion chambers, 1:143
enumeration method, 1:141–142, 1:141f
flow cytometry (FCM), 1:142
microbial culturing and culture media, 1:140
most probable number (MPN) techniques, 1:141–142
optical tweezer (OT), 1:142–143
single-cell manipulation and isolation, 1:142–143
standard plate count, 1:141, 1:141f
- Microbial communities-culture-independent approach
amplicon-based sequencing, 1:152–153, 1:152f
biomarker-based technique, 1:148
biomass, 1:148
clone library construction method, 1:150
culture-dependent method, 1:147
enumeration techniques, 1:147–148
fatty acids-based approaches, 1:149–150
fluorescence in situ hybridization (FISH), 1:149
fumigation-based techniques, 1:148
high-throughput sequencing (HTS), 1:151–153
hybridization-based method, 1:150
metagenomics (shotgun sequencing), 1:153
microarrays, 1:150
microscopy, 1:147–148
molecular techniques, 1:150–151
PCR-based fingerprinting method, 1:150–151
polymerase chain reaction (PCR) methods, 1:149
single-cell sequencing (SCS), 1:153
substrate-induced respiration (SIR), 1:148
traditional techniques, 1:147–149
- Microbial communities, in anaerobic soils, 1:382
- Microbial consortia, 1:425
- Microbial denitrification
aerobic oxidation of ammonia, 1:262–263
anaerobic oxidation of ammonia, 1:263
anaerobic oxidation of methane, 1:264
canonical denitrification, 1:264–266
definition of, 1:262
dissimilatory nitrate reduction to ammonia/nitrate ammonification, 1:267
environmental factors, 1:267–269
under field conditions, 1:266–267
genetic insights into, 1:265–266
heterotrophic denitrification to N₂O, 1:267
mechanisms, 1:262–267
nitrate supply, 1:268
nitrifier denitrification, 1:267
O₂ availability, controls on, 1:268
organic C availability, 1:268
organisms, 1:262–267
pathways, 1:262
pH, 1:269
reductases of oxides of nitrogen used in, 1:265t
substrates, 1:262–267
temperature, 1:269
- Microbial diversity
ecosystem services, 1:478
maintenance of, 1:470
- Microbial efficiency-matrix stabilization (MEMS), 2:429
- Microbial function in soil
acetylene-inhibition technique (AIT), 1:159
acetylene reduction method, 1:159
biochemical techniques, 1:156–157
biogeochemical cycling, 1:156
catabolic capacity and biodegradation, 1:157–158
complex substrate biodegradation assays, 1:158
dehydrogenase activity, 1:159
denitrification, 1:159
enzymatic activity, 1:159–160, 1:160t
fluorescein diacetate (FDA) hydrolysis, 1:159
isotopic techniques, 1:161–163
microbial respiration, 1:157
nitrogen fixation, 1:159
nitrogen mineralization and nitrification, 1:158
omics techniques, 1:163–166
radioisotope techniques, 1:162–163
stable isotopes techniques, 1:161–162, 1:162f
substrate-induced respiration (SIR), 1:157–158
- Microbial immobilization, phosphorus, 2:158–159
- Microbial inoculants
beneficial properties, 1:429, 1:430f
biofertilizers, 1:429
bioremediation, 1:431–432
climate change mitigation, 1:432
competition with resident soil microbial communities, 1:433–434
definition, 1:429
emerging approaches, 1:436–437
field efficacy, 1:435–436
food demand vs. environmental impacts, 1:428–429
function, 1:429–432
genetically modified microorganisms (GMMs), 1:437
microbial consortia, 1:432–433
novel formulation and delivery system, 1:437
plant growth promotion, 1:431
plant stress tolerance enhancement, 1:431
protection against soil-borne pests and diseases, 1:431
quality and delivery, 1:434–435
safety, 1:436
soil and environmental factors, 1:433
soil characteristics enhancement, 1:432
targeted design of microbial consortia, 1:436
- Microbially-induced metal transformations, 2:244, 2:244f
biotechnological applications of, 2:253–254
- Microbially-mediated mineral dissolution, 2:289
- Microbially-mediated mineral weathering, 2:464–465, 2:465f
- Microbial metabolites, 1:303
- Microbial mobility and transport in soils
indicator organisms, pathogen contamination, 1:513–514
microbial pathogens in organic wastes, 1:513f
- Microbial necromass
phosphorus (P), 1:295–297, 1:296f
soil organic matter (SOM), 2:390
- Microbial processing, dissolved organic carbon, 2:428–430
- Microbial respiration pathways, in anaerobic soils, 1:384t
- Microbial selection, root exudates, 1:352
- Microbial transport studies
approaches to, 1:519–520
field studies, 1:519
laboratory studies, 1:520
model studies, 1:520
wastewater treatments systems, 1:577t

- Microbiome
 Arctic urban soil microbiome (Apatity, Russia), 1:487–490, 1:490f
 factors influencing urban soil, 1:482f
 forest soils, 3:60
 from landscape variability to soil microbial diversity, 1:485–487, 1:485f
 map of soil sampling sites, RUDN University, 1:483f
 microbial carbon indicators of Technosols, 1:489f
 microbial community characteristics, 1:481t
 topsoil (0–10 cm) sampling campaign, 1:483, 1:484f, 1:484t
 total microbial biomass, 1:488f
 typology and composition of, 1:482–485, 1:482f
 urban soil, impact of, 1:480
- Microbiotic, 1:128
- Microcalorimetry, 2:663
- Micro-catchment systems, 5:598–599, 5:599f, 5:599t
- Microclimate adjustments, 5:588
- Microdiversity, 1:16ge
- Microfloral, 1:128
- Microfluidics, 1:124–125
- Microhabitats
 examples, 1:200–202, 1:201–202f
 features, 1:200
 growth dynamics, 1:202–204, 1:202t, 1:203–204t
 habitats classification, 1:200
 ordination axes, 1:200, 1:200t
- Micro-irrigation, 3:118
 bubblers and microsprayers, 5:619
 drip irrigation, 5:618f, 5:619
- Micromorphology, 4:166, 4:169f
- Micronutrient cycling, 1:68
- Micronutrients, 1:1ge, 1:310, 3:374
- Microorganisms, 2:243, 3:376
 arsenic, 2:204–205
 fire effects, 2:454
 SUITMAs, 4:240
 wastewater treatments systems, 1:575–577
- Microphytic, 1:128
- Microplastics, 2:668, 3:169. *See also* Plastics
 abiotic transport, 3:317
 aquatic plastic debris, 3:314
 beneficial microorganisms, 3:313–314
 biodegradable (BIO) mulches, 3:313
 biotic transport, 3:317, 3:318f
 direct and indirect sources, 3:312–313, 3:313t, 3:313f
 earthworm mortality, 3:314
 effects, 3:315–317
 and humans, 3:318–319, 3:318f
 low density polyethylene (LDPE) plastic mulch, 3:313
 non-renewable resource, 3:309
 occurrence, 3:313
 and plants, 3:315–316, 3:316f
 plastic content assessment, 3:314
 pro-oxidant additive containing (PAC) plastics, 3:313
 soil animals, 3:314
 and soil ecosystem services, 3:317
 and soil invertebrates and vertebrate, 3:316
 and soil microbiome, 3:315
 and soil physicochemical properties, 3:315, 3:315f
 spectral method, 3:314–315
 stereomicroscope visual selection (SMVS), 3:314
 transport, 3:317
- Micropores, 5:1
- MicroRNA (miRNA), root exudates, 1:355
- Microscale water configuration, soil water repellency (SWR), 5:210–211
- Microsprayers, 5:619
- Microstructure, 4:174–175
- Microtubing irrigation, 5:619
- Microwave antennae measure radiation, 4:621
- Microwave instruments, 4:619
- Microwave radiometry, 4:598, 4:605–607
- Mid-infrared spectroscopy (MIR)
 spectroscopy, field and laboratory soil sensing, 4:702
- Mikulino paleosol, 3:24
- MILESD soilscape evolution model, 4:743–744
- Milligrams, 1:1ge
- Millimeters, 1:1ge
- Millville silt loam soil, angular pore-space model, 5:193–194, 5:194f
- MINEQL model, 2:279
- Mineral amendments, 3:429
- Mineral-associated organic matter (MAOM), 1:235–236, 1:390, 1:393f, 2:69, 2:71, 2:677
 characterization, 1:399–402
 chemical isolation procedures, 1:397–398
 combined physical and chemical isolation, 1:398
 formation, 2:391–392
 Fourier transform infrared (FTIR) spectroscopy, 1:400
 micro-imaging, 1:401
 natural carbon isotope, 1:400
 nuclear magnetic resonance (NMR) spectroscopy, 1:399–400
 physical isolation procedures, 1:397
 radiocarbon dating, 1:400–401
 synchrotron-based techniques, 1:401–402
 X-ray photoelectron spectroscopy (XPS), 1:400
- Mineral deficiencies, 3:559
- Mineral dissolution, 2:285
- Mineral fertilizers, 3:370–372, 3:379–380
- Mineral intensity factor method, 2:648
- Mineralization, 2:142ge
 organic sulfur, 2:195
 phosphorus, 2:158–159
 sulfur, 1:304–305
- Mineralogy, 1:6, 4:237
- Mineral-organic matter-microbiota interactions, 1:391–393, 1:392f
- Mineral-organic-microbial interactions, 1:387
- Mineral particles and aggregates, 2:674–676
- Minerals, 2:580, 2:580f
- Mineral soils
 dissolved organic carbon in, 2:423f
 structures, 2:670–671
- Mineral transformations, SUITMAs, 4:241–242
- Mineral weathering, 4:294–295
- Minimum tillage, 3:444–445
- Mining, 4:63–66, 4:64–65f
 lanthanides, 2:227–228
 trace elements, 3:292
- Mites
 food, 1:77
 habitat, 1:77
 interaction, 1:77
 morphologies and life strategies, 1:77, 1:78–82f
 non-insect soil and litter-dwelling arthropods, 1:77
 phoresy, 1:77–79, 1:80–82f
- Mobile-immobile water model (MIM), 5:353ge, 5:358–359, 5:660–661
- Mobility, arsenic (As), 2:216
- Mobilization, 2:245–248
- Model-assisted estimators, 4:538–539
- Model error, 4:682
- Model inversion, 4:624
- Model of Water Flow and Solute Transport in Macroporous Soil (MACRO), 2:479–481, 2:480t
- Modern drip irrigation, 5:609
- Modular systems of living walls, green roofs and, 3:480
- Mohr-Coulomb failure criterion, 5:77, 5:79–80, 5:79–80f
- Moisture (hydrogen) calibration dependency, 2:635
- Moisture distribution map, 2:640–641, 2:640f
- Moisture equivalent, 5:511
- Molecular modeling (MM) methods, 2:484–490, 2:485f
- Molecular techniques, 1:150–151
- Mollisols, 4:55ge, 4:229, 4:296, 4:328
- Mollisols-Chernozems
 in Europe, 4:310, 4:312–313, 4:314–315f
 features, 4:305
 grain crops, 4:305
 Haplic Chernozem, 4:307f
 in North America, 4:308–309, 4:312f
 in Northeast China, 4:313–316, 4:316f
 prevalence, 4:308
 regions, 4:308, 4:311t
 in South America, 4:309–310, 4:313–314f
 tall grass steppe, 4:305
- Mo-nitrogenase, 1:271–272, 1:272f
- Monocalcium phosphate, 3:376
- Monodentate, 2:328
- Monod model, 1:175–176
- Monomeric xenobiotics, 2:64–65
- Mononuclear aluminum species, 2:173–174f
- Monosilicic acid (H₄SiO₄), adsorption of, 2:268–270
- Monte Carlo methods, 4:677, 4:680
- Monteith, John, 5:681, 5:682f
- Montmorillonite, 2:174f

- Moraine deposits, of Fürkeleferner glacier, 4:4f
- Morphology of soil particle, 2:580–581
- Mosse, Barbara Gertrud Hedwig, 5:688
- Mosses, 1:128
- Most probable number (MPN) techniques, 1:141–142
- Mottles, 1:2ge, 4:157, 4:158f
- Mowing, 3:107–108
- Mucilage, 5:547–548, 5:548f
- Mucilage-microorganism interaction. *See* Rhizobiome
- Mulch, 3:118–119, 5:53
 application, 3:401–402
 black plastic, 3:396
 characteristics, 3:392
 climate change, 3:401
 colored film, 3:397f
 colored plastic, 3:397
 crop growth, 3:408–411
 crushed volcanic rock, 3:398
 on environment, 3:407–408
 evaporation control, 3:402, 3:403f, 3:403t
 farm yard manure (FYM) as, 3:394
 in fields, 3:311, 3:311f
 gravel, 3:397
 infiltration improvement, 3:402–403, 3:403t
 inorganic, 3:396–398
 input-use efficiency, 3:409–410, 3:410t
 limitations of, 3:398–399
 methods of application, 3:398
 moisture retention, 3:403
 nutrient-use efficiency, 3:409–410, 3:410t
 organic, 3:393–395
 organic residue mulch, 3:393–394
 plastic, 3:396–397, 3:396f
 policy decisions, 3:411–412
 recycled paper, 3:395
 root growth, 3:408, 3:409t
 runoff and soil erosion control, 3:404–407
 seedling emergence, 3:408
 soil health, 3:401–402
 soil moisture, 3:402–404
 soil temperature, 3:311
 sprayable preparations, 3:398
 straw mulching, 3:393f, 3:405, 3:408
 transparent/lear plastic, 3:397
 types of, 3:393–398
 wastes from cattle sheds as, 3:394
 water condensation at night, 3:404
 water-use efficiency (WUE), 3:311, 3:409–410, 3:410t
 weed competition, 3:408
 weed growth, 3:311
 white plastic, 3:397
- Mulch farming, 3:118–119
- Multidimensional NMR, 2:615–616, 2:617f
- Multidisciplinary approach, 4:244
- Multiomics, 1:165–166
- Multiple species grazing, rangeland management, 3:92
- Multi-rate mass transfer (MRMT) approach, 5:353ge
 averaging volume, 5:662
 Laplace transform, 5:662
- memory function, 5:662
- mobile solute concentration, 5:661–662
- single fracture-matrix systems, 5:662, 5:662f
 total immobile concentration, 5:661
- Multi-sensor data fusion, 4:588
- Multisensory absorption
 concept, 3:517–518
 Earth Room, 3:517f, 3:518
 neologism, 3:518
 reconstruction, 3:517–518
- Multivariate analysis
 digital elevation model (DEM), 4:500
 digital soil mapping, 4:500
 dimension reduction, 4:501–502
 remote sensing data, 4:500
 soil spectral inference, 4:500
 soil spectral visible near-infrared (vis-NIR) measurement, 4:499, 4:500f
 supervised classification, 4:503–505
 unsupervised classification, 4:505–506
- Multivariate curve alternative least-square fitting (MCR-ALS) method, 2:346
- Municipal solid waste (MSW)
 biosolids, 3:330–331
 challenges, 3:331
 combustion (incineration) with energy recovery, 3:328
 composition, 3:323–331
 definition of, 3:321–322
 elimination, 3:323–331
 energy recovery, 3:326–331
 four-tiered waste management hierarchy, 3:324f
 history of, 3:322–323
 incineration, 3:330
 landfill, 3:329–330
 landfill gas (LFG) utilization, 3:327, 3:327f
 management in United States, 3:324f
 non-thermal waste-to-energy (WtE) (anaerobic digestion), 3:328
 ongoing strategies, 3:323–331
 organic waste and compost strategies, 3:325–326
 production, 3:323–331
 pyrolysis and gasification, 3:328
 recyclable and reclamation products, 3:325
 recycling in 2018, 3:325f
 thermal waste-to-energy (WtE), 3:327–328
 treatment and disposal, 3:329–330
- Munsell color system, 3:510–511
- Munsell soil color chart, 4:155f
- Muramic acid (MurN), 2:392–393
- Murein, 1:524
- Muscovite, 4:100–101
- Museums, soil profile collections, 3:575t
- Mutualism, 1:52, 1:53f, 1:54t
- Mutualistic microorganism, 1:362
- Mutualistic symbiotic, 1:420
- Mycology, 1:49. *See also* Fungi
- Mycorrhizae, 2:152ge, 2:429–430
- Mycorrhiza helper bacteria (MHB), 1:453–454
- Mycorrhizal fungi, 1:424, 5:548–549
 indigenous arbuscular, 1:452f
 inoculation with, 1:453–454
- soil organic matter (SOM) decomposition, 2:390
- Mycorrhizal interaction, 1:279
- Mycorrhizal symbioses, 1:327
- Myxomorphae, 1:81, 1:83f
- Mylonchulus montanus*, 1:108f
- Myo- and scyllo-inositol hexaphosphate, 2:152ge
- Myriapoda
 Chilopoda, 1:93f, 1:96–97
 classes, 1:96
 Diplopoda, 1:95–96f, 1:97–98
 Pauropoda, 1:93f, 1:98–99
 Symphyla, 1:96f, 1:99
- ## N
- Naming soils, 4:208–214
- Nano-fertilizers/pesticides, 2:689
- Nanoparticles (NPs), 2:681
 aging, 2:687
 assembling, 2:687
 beneficial application, 2:688–689
 dissolution, 2:686
 embedding, 2:688
 fate of, 2:685–687
 identification in soil, 2:683–685
 impacts of, 2:688–690
 interactions between soil components and, 2:687–688
 natural occurrence, 2:682
 separation, 2:683
 surface coating of organic matter on, 2:687–688, 2:688f
 techniques in, 2:683–684
 transport, 2:686–687
 two-blade sword to soil system, 2:690, 2:690f
 types, 2:684f
 uptake of, 2:689–690
- Nano particle tracking analysis (NTA), 2:683
- Nanopore cavity (periodic) models, 2:493
- Nano-scale secondary ion mass spectrometer (NanoSIMS), 2:670–672, 2:676f
 coaxial configuration, 2:673–674
 interfaces of soils and biota, 2:676–677
 measurement, 2:673
 microbiota in soil environment, 2:678–679
 technical principle, 2:671, 2:672f
- Naphthalene, 2:426–428
- National Action Programs (NAP), 3:537
- National and regional level soil governance, 3:539–541
 administrative arrangements, 3:541–542
 complementary provision, 3:540
 culture and ethics, 3:542
 framework law and sectoral provisions, 3:540
 Germany, 3:542–543
 legally binding standards, 3:541
 methodological observations, 3:543
 procedural rights and courts, 3:541
 social and ecological services, 3:539–540
 soil data and information, 3:540
 tenure rights, 3:541

- National Biogas and Manure Management Program (NBMMP), 3:469–470
- National Food Security Act, 4:425*t*
- National Green Tribunal Act, 3:469
- National Mission for Sustainable Agriculture (NMSA), 3:469
- National Policy for the Management of Crop Residue (NPMCR), 3:469–470
- National Programme on Organic Production (NPOP), 3:469
- National soil classification systems, 4:191–195
- National Soil Inventory (NSI), 3:175
- National Technical Committee for Hydric Soils (NTCHS), 4:433–434
- National Wetlands Inventory (NWI), 4:424–425
- Natural abundance, nitrogen isotopes, 2:89
- Natural and planted forests, 3:58*t*
- Natural and recycled organic materials, green roofs, 3:484
- Naturally occurring nanoparticles (NNPs), 2:682
- Natural organic matter (NOM), 2:592, 2:687–688
- Natural soils and natural mineral materials, green roofs, 3:482–483
- Nature-positive farming system, 3:15–16
- Navier–Stokes equation, 5:28
- Near- and mid-infrared spectroscopy, 2:571
- Near-edge X-ray absorption fine structure (NEXAFS) spectroscopy, 1:401–402
- Near-zero growth (NZG) kinetics, 1:183–184, 1:183*f*
- Necromass, 2:60, 2:62*t*
- Nematicides, 2:471
- Nematoda, 1:109
- Nematodes, 1:105
- distribution and abundance, 1:109, 1:109*t*
- ecology, 1:110
- monitoring, 1:110
- physiology, 1:109
- in soil, 1:449*t*
- soil organic matter (SOM) decomposition, 2:390
- Neolithic site, desert pavement, 4:335*f*
- Neosoil, 3:517*f*, 3:518
- Nested subsets theory, 4:733–734, 4:733*f*
- Net primary productivity, 3:253, 3:253*f*
- Net radiation, 5:635, 5:640–641
- Net radiation balance, 5:393–395, 5:394*f*
- Neural networks (NNs), 4:649
- Neustruev's soil classification, 4:190
- Neutron-gamma analysis, 2:627
- general principles, 2:627
- modifications, 2:627–628
- Neutron moisture meter (NMM), 5:427–429, 5:429*f*
- access tubes and depth control, 5:429–430
- calibration, accuracy, and representativity, 5:431
- interferences, 5:428
- safety and use considerations, 5:431–432
- statistics, 5:430
- Neutrons, 2:627
- probes, 4:594, 4:595*f*, 5:174
- scattering, 4:594, 5:427–428
- sources, 2:628
- tomography, 5:464
- Neutron thermalization method, 5:428
- New Zealand pastures
- deforestation, 3:350–351
- reforestation, 3:351–352
- Waipaoa Basin, 3:351
- Next-generation sequencing (NGS).
- See also* High-throughput sequencing (HTS); Omics techniques
- for advancing microbial transport studies, 1:520
- denitrification genes, 1:265–266
- Nitrapyrin, 1:259
- Nitrate (NO₃), 1:262, 2:44, 2:46
- ammonification, 1:267
- arsenic, 2:205
- oxygen in, 2:90–91
- pollution, 3:272
- reduction in anaerobic soils, 1:383–384
- supply, microbial denitrification, 1:268
- Nitric phosphate fertilizers, 3:371
- Nitrification, 1:262–263, 1:263*t*, 2:146, 2:694
- in agricultural systems, 1:258–259, 1:258*f*
- ammonia oxidizing bacteria, 1:255–256, 1:256*t*
- controlling factors, 1:256, 1:257*t*
- definition, 1:254–255
- inhibitors, 1:259
- management, 1:258, 1:258*f*
- microbes involved in manure management, 1:563–564
- nitrification rate, 1:257
- organisms, 1:254*f*, 1:255–256, 1:256*t*
- pathway and enzymes in, 1:264*f*
- role, 1:254, 1:254*f*
- trace gas production, 1:255
- Nitrification–denitrification, 1:572
- Nitrifier denitrification, 1:267
- Nitrite (NO₂), 1:262
- Nitrobacter*, 1:255–257
- Nitrogen (N), 2:143, 3:372, 3:372*f*, 4:378
- application, 1:458
- compounds in soils, 2:144
- fixation, 1:421–422
- leaching, 5:327–328
- metabolism, 1:277
- mineralization, 2:145
- need for, 2:143
- removal from wastewater, 1:572
- soil hydrology and biogeochemistry, 3:63
- soil testing, 2:46–47
- transformations, 2:144, 2:144*f*
- on turfgrass species, 3:500–501
- Nitrogenase, 1:271–272, 1:287–288
- Nitrogen cycles, 4:26–27
- anthropogenic impacts on, 2:148–149
- and formation of nitrous oxide, nitric oxide and dinitrogen, 1:266*f*
- human impacts on, 2:148–149
- soil algae, 1:67
- Nitrogen cycling, 1:246, 1:246*f*.
- See also* Organic nitrogen
- ¹⁵N enriched fertilizers, 2:88–89
- Nitrogen fixation, 2:142*ge*, 2:143–145
- microbes involved in manure management, 1:563–564
- and nutrient cycling, biocrusts, 1:133
- Nitrogen-fixing plant-microbe associations, 1:272*f*
- associative nitrogen-fixing interaction, 1:273
- endophytic nitrogen-fixing associations, 1:273
- endosymbiotic nitrogen-fixing associations, 1:273–274
- Nitrogen isotopes, 2:88–89
- and clumped isotopes to study N₂O emissions, 2:92
- enriched fertilizers, 2:88–89
- natural abundance, 2:89
- Nitrogen use efficiency (NUE), 3:162
- Nitrosomonas/Nitrospirilla*, 1:254–256
- Nitrous oxide (N₂O), 2:143
- crop-based biofuels and, 3:162
- emissions, 2:694–695, 3:159–160, 3:160*f*, 3:162, 3:163*f*
- potential for reducing, emissions, 3:162
- soil emissions determinants, 3:160–162, 3:161*f*
- soil sources, 3:158–159, 3:159*f*, 3:159*t*
- Non-aqueous phase liquids (NAPLs), 1:553
- air-NAPL-water systems, 5:277–279
- air relative permeability, 5:281
- aqueous solubility, 5:274*t*
- capillary pressure, 5:275
- contact angle, 5:275
- dense, 5:273
- density and viscosity, 5:274*t*
- fluid distribution, 5:274–279
- interfacial tension, 5:274*t*
- light, 5:273
- macroscopic capillary pressure, 5:277
- mass transfer, 5:283
- NAPL-air partitioning, 5:284
- NAPL-water partitioning, 5:285
- pressure distribution, 5:273
- rate-limited interphase mass transfer, 5:285
- relative permeability, 5:282–283, 5:283*f*
- representative elementary volume (REV), 5:275–276
- simulation models, 5:273
- soil-water partitioning sorption, 5:284
- volatile components, 5:273
- water-air partitioning (volatilization), 5:284
- water saturation, 5:277
- wettability, 5:274
- Young–Laplace equation, 5:275
- Non-cash crops, 3:446
- Non-contact (NC)-AFM imaging, 2:591–597
- Non-cryotic taliks, 5:424
- Non-Darcian flow, 5:245
- Non-Derjaguin–Landau–Verwey–Overbeek (DLVO) forces
- capillary force, 5:340
- empirical model, 5:339
- hydration energy, 5:339
- hydrophobic interaction, 5:340
- steric repulsion, 5:339–340

- Non-equilibrium flow, 5:306*ge*
 Non-extractable (bound) residues (NER),
 2:476–478, 2:478*f*
 Non-invasive imaging, soil process
 applications, 5:467
 computed tomography, 5:460–464
 magnetic resonance imaging (MRI),
 5:464–466
 positron emission tomography (PET),
 5:466–467
 Nonionic compounds sorption
 mineral solids, 2:318, 2:319*f*
 soil organic matter, 2:318–324
 Nonmarket valuation, 3:551–552
 Non-metric multidimensional scaling
 (nMDS), 4:502
 Non-probability sampling, 4:481
 adapted experimental designs, 4:546–547
 geometric sampling designs, 4:546
 grid spacing optimization, 4:547–548,
 4:548*f*
 sampling pattern optimization, 4:548–549
 semivariogram estimation and mapping,
 4:549
 Non-rainfall water, 5:674, 5:678
 Non-symbiotic N₂ fixation (NS-N₂ fixation)
 accurate crop season and annual estimates,
 1:290
Azotobacter chroococcum, field experiments,
 1:287*f*
 diazotrophs, 1:284–285, 1:286*t*,
 1:287–289
 ecology, 1:284–285
 endophytic diazotrophs, 1:290
 energy availability, 1:289
 microsites, 1:284*f*
 nitrogenase enzyme, 1:284
 nutrient availability, 1:289
 and plant relationship, 1:285
 rhizosphere, 1:289
 soil aggregates, 1:290
 soils and terrestrial environments,
 1:285–287
 Non-thermal waste-to-energy (WtE)
 (anaerobic digestion), 3:328
 Non-traditional isotope geochemistry
 equilibrium isotope fractionation, 2:108
 high-precision isotope values, 2:108
 isotope fractionation, 2:108
 kinetic isotope fractionation, 2:108
 mass spectrometry, 2:108
 post-analysis correction, 2:108
 Non-transition chemical elements,
 2:214–215
 Non-transition metals and metalloids,
 2:215*t*, 2:216*f*
 Normalized difference vegetation index
 (NDVI), 4:500, 5:521
 Normalized relative canopy temperature
 (NRCT), 4:607–608
 Normal soils, 4:188
 Northern Circumpolar Region (NCR)
 boreal and subarctic regions, 4:264–267
 high Arctic soils, 4:268–270
 low Arctic soils, 4:267–268
 parent materials, 4:264
 pedological zonation, 4:265*f*
 permafrost extent and soil temperature
 regimes, 4:265*f*
 Northern Hemisphere Permafrost, 5:411
Nostoc, 1:64. *See also* Cyanobacteria
 No-tillage systems, 3:444–445
 Nuclear magnetic resonance (NMR)
 spectroscopy, 2:605–606, 4:591–592,
 5:452
²⁷Al, 2:174–175
 common nuclei in, 2:606–607
 experiment preparation, 2:607–609
 nucleus selection, 2:607
 sensitivity improvements in, 2:622–623
 soil interactions via, 2:618–621
 for soil moisture, 4:597–598, 4:598*f*
 solution-state, 2:609–610, 2:621
 for structural elucidation, 2:612–614,
 2:613–615*t*
 techniques, 2:607, 2:608–609*t*, 2:611
 Nuclear methods, sensor technology for,
 5:174
 Nuclear Overhauser effect spectroscopy
 (NOESY), 2:619–620
 Nuclear quadrupole interaction, EPR signal,
 2:522
 Nucleocyto-plasmic large DNA viruses
 (NCLDVs), 1:19–20
 Nugget variance, 4:660
 Numerical classification, 4:474, 4:476
 Numerical weather prediction (NWP), 4:627
 Nutrient absorption, root mucilage,
 1:337–338
 Nutrient availability
 ecosystem net primary productivity (NPP),
 2:693
 soil algae, 1:66
 soil respiration, 1:373
 Nutrient carriers, foliar spray formulations,
 3:388*t*
 Nutrient cycles
 disruption, 1:556
 protists, 1:45–46
 soil algae, 1:67–68
 Nutrient cycling, 3:147–148
 Nutrient management, 3:500
 in animal manure, 1:560*t*
 on bacterial community structure, 1:460*f*
 liming, 3:142
 mineral fertilizers, 3:141–142
 organic fertilizers, 3:141
 systems approach, 3:141
 Nutrients, 3:280–281, 3:281*t*
 deficiencies, 3:389
 malnutrition, 3:559
 mobilization, root exudates, 1:352
 supply, 3:188
 Nutrients uptake
 arbuscular mycorrhizal (AM) association,
 1:319
 beneficial elements, 1:312
 cation exchange capacity (CEC),
 1:314–315
 clays, 1:314–315
 C:N ratio, 1:314, 1:314*t*
 essential elements, 1:311–312
 ion uptake, 1:312
 microorganisms in, 1:313–314
 mineralization, 1:313
 mineral nutrient cycling, 1:313
 mycorrhizal association, 1:319
 new roots, 1:318, 1:318*f*
 nutrient ion-root interaction, 1:315–316,
 1:316*f*
 pathway, 1:313–314
 pH, 1:314–315
 for plant lifecycle, 1:310
 proliferation of roots, 1:317–318,
 1:317–318*f*
 root length density (RLD), 1:316–317
 root traits, 1:318–319
 root weight ratio (RWR), 1:316–317
 soil animals in, 1:314
 soil development, 1:313
 soil formation factors, 1:312
 soil organic matter, 1:314–315
 Nutrition, potassium in soils, 2:119
 Nutrition sensitive agriculture (NSA), 3:17
- ## O
- Octahedral aluminum, 2:168–170, 2:170*f*
 Octanol-water partition coefficient, 2:471
 Olivines, 4:102–103
 Ombric histosols, 4:392
 Omics techniques, 1:163, 1:164*f*
 amplicon sequencing, 1:164
 functional gene prediction, 1:164
 functional metagenomics, 1:163
 metabolomics, 1:165
 metaproteomics, 1:165
 metatranscriptomics, 1:164–165
 multiomics, 1:165–166
 sequence-based metagenomics, 1:163
 One-chamber diffusion apparatus, 5:251,
 5:252*f*
 “One Health” concept, 3:16–17
 One Health Initiative, 3:507
 1:1-type clay minerals
 halloysite, 4:114–115, 4:115*f*
 kaolinite, 4:114, 4:115*f*
 1:1 layer structure, 4:113–114, 4:114*f*
 Online resuspension chamber, 3:229
 Onsager Reciprocal Relations (ORRs), 5:266
 On-site direct microscopy, 1:189
 On-surface drip irrigation, 5:619
 On-the-go sensors, 4:582–585, 4:583*f*
 Opal, 4:93, 4:146
 Operational electronics, 2:629
 Opiliones, 1:84–86, 1:85*f*
 Optical reflectance, 4:166
 Optical sensors, 2:567–568
 Optical tweezer (OT), 1:142–143
 Ordinary kriging, 4:668–669
 Ore leaching, 2:253
 Organic acids, 2:245
 Organic amendments, 3:108–109
 Organic carbon, 2:58, 2:59*f*, 3:281–282
 availability, 1:268
 dynamics, 3:425–427
 morphological features, 4:428–429

- partitioning coefficients, nonionizable
 - organic compounds, 5:658*t*
- in soil precipitates, 2:293–295
- Organic chemicals sorption
 - charge-assisted H-bond (CAHB), 2:317
 - Gibb's free energy, 2:317
 - H-bond, 2:317
 - hydrophobic effect, 2:318
 - nonionic compounds, 2:318–324
 - organic ions, 2:324–325
 - physisorption process, 2:316
 - sorption distribution ratio, 2:317
 - thermodynamics, 2:316–317
- Organic compounds, wastewater treatments
 - systems, 1:574
- Organic contaminant interactions,
 - 2:494–495
- Organic contaminants biodegradation, 1:547
 - aerobic transformation, 1:548–549
 - anaerobic transformation, 1:549–550
 - bioavailability, 1:552–553
 - chemical properties, 1:550–552
 - climate change, 1:555–556
 - co-contaminants, 1:555
 - contaminants, 1:551*f*
 - environmental factors, 1:552–555
 - groundwater contamination, 1:548
 - microbial degradation, 1:548
 - microorganisms, 1:554
 - mineralization, 1:548
 - nutrient availability, 1:553–554
 - pH, 1:554
 - plants, 1:555
 - predation, 1:554–555
 - temperature, 1:554
 - water content, 1:554
- Organic contaminants in soil, 2:503–506
- Organic farming (OF)
 - vs.* conventional system, 3:143*t*
 - crop rotation system, 3:138–139
 - definition, 3:136
 - elements, 3:137, 3:137*f*
 - farming system impact, 3:137–138
 - findings, 3:142–144, 3:143*t*
 - future research, 3:144
 - nutrient and organic matter system,
 - 3:141–142
 - obligatory and facultative practices, 3:136,
 - 3:136*t*
 - practices, 3:136, 3:136*t*
 - soil health and soil fertility, 3:137, 3:137*f*
 - soil tillage system, 3:140–141
 - tree based systems, 3:140
- Organic fertilizers, 1:529, 3:428
- Organic manures, 1:513, 3:291–292,
 - 3:379–380
- Organic matter (OM), 1:6, 2:22, 2:236,
 - 2:494–495, 3:281–282, 4:74–75,
 - 4:75*f*, 5:10
 - anaerobic soils, decomposition in, 1:383*f*
 - biotransformation continuum model,
 - 1:234, 1:235*f*
 - bromination, 2:239
 - chlorination, 2:238–239
 - content, soil quality and, 3:497
 - decomposition, 2:143
 - fluorination, 2:237
 - pesticide sorption, 2:472
 - polymers, 3:303
 - recycling in forests, forest soils, 3:60–61
 - role, 4:373
- Organic matter loss, soil change
 - litter and organic debris decomposing,
 - 3:351*f*
 - New Zealand pastures, 3:350–352
- Organic mulch, 3:393–395
- Organic nitrogen
 - depolymerization, 1:247, 1:247*f*
 - distribution, 1:244, 1:244*f*
 - forms and properties, 1:244–245, 1:244–245*f*
 - fractionation, 1:250
 - high molecular weight (HMW)
 - compounds, 1:245, 1:245*f*
 - identification, 1:250
 - leaching, 1:249
 - low molecular weight (LMW) compounds,
 - 1:245, 1:245*f*
 - in managed system, 1:251
 - microbial cell uptake, 1:247
 - mineralization, 1:247
 - nitrogen cycling, 1:246, 1:246*f*
 - plant/root uptake, 1:248–249
 - quantification, 1:249–250
 - reactive N effects, 1:244
 - sampling, 1:250
 - soil organic matter persistence, 1:247–248,
 - 1:248*t*
 - total organic nitrogen, 1:249–250
- Organic pollutants, removal of, 3:419–420,
 - 3:421*f*
- Organic production, 1:3–4
- Organic residue mulch, 3:393–394
- Organic root exudates
 - origin and transport, 1:347–348
 - secretion, 1:348–349
- Organic soils
 - anthropogenic use, 4:393–394
 - biomass production, 4:383–384
 - bogs, 4:392
 - characteristics and properties, 4:390–392
 - classification, 4:384–385, 4:385*t*
 - crop production, 4:392
 - distribution, 4:387–390
 - fens, 4:392
 - formation, 4:386–387
 - horizons, 4:384*f*
 - peatlands, 4:384–385
 - properties, 4:385
 - restoration and sustainable use, 4:395–396
 - storage, 4:383
 - utilization, 4:392–396
- Organic sulfur, mineralization of, 2:195
- Organic wastes, 4:379
 - and application method, type of, 1:515
 - and compost strategies, 3:325–326
 - properties and storage conditions, 1:515
- Organismic carbon
 - categories, 2:61*f*
 - decomposition, 2:63
 - humic substances paradigm, 2:63
 - microbial necromass, 2:62*t*
 - residence time, 2:62–63
- Organisms, SUITMAs, 4:240
- Organofluorine compounds in soil, sorption,
 - 2:237
- Organ of perennation, 3:439*ge*
- Organo-mineral aggregates, 2:493
- Organo-mineral associations, 1:236,
 - 2:674–675
- Organo-mineral interactions
 - adsorption, 1:388, 1:389*f*
 - co-precipitation, 1:390
 - hydrogen bonding, 1:390
 - hydrophobic interactions, 1:390
 - intercalation, 1:390
 - ion exchange, 1:388
 - ligand exchange, 1:388
 - physical protection, SOM, 1:390
 - polyvalent cation, 1:388–390
 - and terrestrial environment, 1:403–405
 - Van der Waals forces, 1:390
- Organophosphorus pesticides, 2:503–504,
 - 2:505*f*
- Organosols, 4:427
- Organo-sulfur assimilation genes, 1:306–307
- Organo-sulfur compounds, 1:306–307
- Orthels, 4:226
- Orthents, 4:230
- Orthods, 4:227
- Orthophosphate, 1:294
- Orthopyroxene, 4:104*f*
- Osmosis and osmotic pressure, 5:153–154,
 - 5:153*f*
- Osmotic potential, 5:162, 5:166–167,
 - 5:518
- Outer-sphere complex (OSC), 2:152*ge*, 2:491
- Outer-sphere complex surface-charge density,
 - 2:368*ge*
- Outreach, soil, 3:506
- Overland-flow process, 5:287–291
 - definition, 5:287
 - models, 5:291–292
 - saturation excess runoff models,
 - 5:291–292
 - with SCS-CN, 5:291
 - wastewater treatments systems, 1:572
 - water quality impacts of saturation excess,
 - 5:292
- Oxalates, 2:251
- Oxford Nanopore Technologies (ONT),
 - 1:19–20
- Oxidation
 - in anaerobic soils, 1:385
 - chemical weathering, 4:39
 - ponds, 1:569
 - of sulfur, 1:307, 2:195
- Oxidation-reduction reaction, 2:198
- Oxidative phosphorylation, 2:300, 2:300*f*
- Oxides, 2:263, 2:266*f*, 2:268–269
- Oxisols, 2:41, 4:55*ge*, 4:227, 4:228*f*,
 - 4:461–464
- Oxo-degradable/oxo-biodegradable plastics,
 - 3:313
- Oxyanion, 2:92
 - attenuated total
 - reflectance–Fourier-transformed
 - infrared (ATR–FTIR) spectroscopy,
 - 2:329

- Oxyanion (*Continued*)
 co-adsorption effect, 2:332–333
 competitive behavior, 2:331*f*, 2:332
 detection and quantification in solution, 2:328
 diffuse reflectance FTIR (DRIFT), 2:329
 dissolved electrolytes, 2:332
 enhancing effect, 2:332–333, 2:333*f*
 ion effect on adsorption, 2:332–333
 ionic strength, 2:332
 level, 2:327
 at mineral–water interfaces, 2:328, 2:329*f*
 Raman spectroscopy, 2:329
 reactivity of background anion, 2:332
 site preference and site reactivity, 2:334
 in soils, 2:327
 sorption mechanisms, 2:327–328, 2:333–334
 sorption strength, 2:330–332
 translocation, 2:328
 vibrational spectroscopy, 2:329
 X-ray absorption spectroscopy (XAS), 2:330
- Oxygen (O₂), 2:168, 2:214–215
 availability, controls on microbial denitrification, 1:268
 isotopes in soil water–plant–atmosphere continuum, 2:89–90
 in nitrate and sulfate, 2:90–91
 organic contaminants biodegradation, 1:556
 in phosphate to study P cycling, 2:90
 supply, 3:188
 symbiotic nitrogen fixation, 1:275–276, 1:278–279
- Oxygen diffusion rate (ODR), 5:221
- P**
- Pacific Biosciences (PacBio), 1:19–20
- Paddy soils, 2:202
 in Asia's lowlands, 4:447–448
 blanket fertilizer recommendations, 4:453
 cropping environments, 4:447–448, 4:448*t*, 4:449*f*
 distribution and characteristics, 4:448–450, 4:451*f*
 environmental issues, 4:456
 indigenous N supply and fertilizer-N, 4:452–453, 4:452*f*
 input-output balance for P and K, 4:453, 4:454*t*
 irrigated and rainfed lowland rice, 4:454, 4:455*f*
 lowland rice system and water availability, 4:454, 4:455*f*
 nutrient-supplying capacity, 4:450–452
 submergence, 4:447
 sustainability, 4:454–456
 water productivity at lowlands, 4:454
 water regime, 4:447
 yields *vs.* nutrients supply, 4:453, 4:453*t*
 zinc and sulphur deficiency, 4:453
- Paired sensors, in SWC measurement, 5:200
- Paleosols, 3:24, 4:54*ge*
- Paludification, 4:386–387, 4:386*f*
- Palustrine wetlands, 4:438
- Palygorskite and sepiolite, 4:118, 4:118*f*
- P and K solubilization, 1:422
- Pan lysimeters, 5:327
- Parallel Data Assimilation Framework (PDAF), 5:478–480
- Paramagnetic relaxation probes, 2:540
- Paramagnetic transition metal ions
 divalent copper, 2:528
 divalent manganese, 2:529
 electron paramagnetic resonance (EPR) spectra and parameters, 2:527–529, 2:528*f*
 ferric iron, 2:527–528
 molybdenum (V) and (III), 2:529
 vanadyl ion, 2:528
- Parametric soil water characteristic (SWC) model, 5:190
- Parasites, 1:45
- Parasitism, 1:52–54
- Parasponia*-rhizobium symbiosis, 1:274
- Parent material, 4:8, 4:9*f*
 alluvial deposits, 4:35
 anaerobic soils, 4:400
 classification, 4:35*t*
 and constructed technosol composition, 3:482–485
 residual, 4:34
 soil classification, 4:40
 soil formation, 4:39–40
 soil properties, 4:39
 SUITMAs, 4:239
 total metal concentrations, 3:290*t*
 windblown deposits, 4:35–36
- Parkland system of sub-Saharan Africa, 3:75*f*
- Parks
 heavy metal contamination in public, 3:503
 inventory and analysis of, 3:489
 locations for, 3:489
 master plans for, 3:489
 soil contamination and remediation, 3:503
 water features in, 3:493
- Partial least squares regression (PLSR), 2:569
- Particle size. *See also* Texture analysis
 added value of continuous data, 5:40
 advantages and disadvantages, 5:40
 effective particle diameter, 5:36
 fraction smaller than 2 mm, 5:37
 grain size distribution, 5:36
 hydrometer method, 5:37, 5:38*f*
 integral suspension pressure (ISP) method, 5:37–38
 light scattering, laser diffractometry (LDM), 5:38, 5:39*f*
 measurement effort and cost, 5:40
 measurement theories, 5:38–39
 particle shape, 5:36
 pipette method, 5:37, 5:38*f*
 sample pre-treatment, 5:37
 schemes, 5:40, 5:41*f*
- Particle-size distribution, of alluvium, 4:418
- Particulate organic matter (POM), 2:69–71, 2:391–392, 5:11
- Partitioning in cellular environment, 2:248–250, 2:249–250*f*
- Passive tension fluxmeter, 5:328*f*
- Past human activities, footprints of, 4:233–237
- Pasture, 3:104, 3:108–110
- Pasture weeds, 3:442
- Pathogens, 3:285
 prevention, root contact, 1:339
 survival in soils, wastewater treatments systems, 1:576*t*
- Pauling's Rules, 'electrostatic valency principle' of, 2:175
- Pauropoda, 1:93*f*, 1:98–99
- Payment for ecosystem services (PES), 3:467–468
- Peat-based formulation, 1:434
- Peatlands, 4:287, 4:384–385, 4:751*ge*
 acrotelm, 4:391
 decomposition, 4:391–392
 formation, 4:386–387
 global distribution, 4:388*f*
 landscape, 4:285*f*
 land surface model, 4:759–760
 rewetting, 4:395
 water content, 4:391
- Peclet number, 5:654
- Pedestal erosion, 3:197
- Pedodiversity, 4:476–477
 measurements, 4:727–728
 relations between biodiversity and, 4:728
 scientific concept of, 4:727
- Pedofeatures, 4:176–178, 4:176–177*f*
- Pedogenesis, 4:2–5, 4:46, 4:70, 4:235–236, 4:243–244
 boreal and subarctic regions, 4:266
 human effects, 3:366
 modelling, 3:367
- Pedogenic formation processes in arid soils, 4:178
- Pedology, 1:2*ge*, 4:218, 4:472, 4:472*f*
 definition, 4:1–2
 importance of research, 4:5
- Pedometrics, 4:471–472, 4:472*f*, 4:481–483, 4:706–708
 principles and hypotheses, 4:472–473
 quantitative pedogenesis, 4:483–485, 4:484*f*
 quantitative soil geography, 4:474–477
 soil measurements and properties, 4:473–474
 soil utility and quality, 4:481–483
 space and time prediction, 4:477–481
- Pedon models, 4:163, 4:739, 4:741, 4:748
- Pedon-scale soil development model, 4:741*f*
- Pedo- or petrophysical models, 5:445
- Pedosphere, 3:3*f*
- Pedotransfer functions (PTFs), 2:30, 4:474, 4:475*f*, 4:756, 5:44–45, 5:191
 background and assumptions, 4:651, 4:652*f*
 calibration metrics, 4:650
 class and continuous PTFs, 4:648
 definition, 4:643
 development strategies, 4:647–649
 estimation and simulation, 4:651, 4:653

- evaluation criteria, 4:650
 generalized approach, 4:647, 4:648f
 hierarchical predictors, 4:646–647, 4:646–647t
 hydraulic conductivity, 4:643
 hydraulic PTFs, 4:644
 land qualities, 4:643–644
 in land surface/vadose zone study, 4:650–653
 machine learning approach, 4:649
 output, 4:644–645
 predictors, 4:645–647
 regression approach, 4:648, 4:649t
 replicas and ensembles, 4:649
 representation, 4:644, 4:644f
 R3H1 Class PTF, 4:646–647, 4:647t
 Rosetta 3, 4:646–647, 4:646t
 semi-physical approach, 4:647–648
 types, 4:644, 4:644f
 Peltier effect, 5:265, 5:267
 Penetrating plant roots, bioturbation, 5:103
 Penetrometer of American Society of Agricultural Engineers, 5:82f
 Penicillin, 3:567–568
 Penman model, 5:503
 Penman–Monteith (PM) equation, 5:503, 5:572, 5:591
 aerodynamic and surface parameters, 5:638–639
 application, 5:642
 definition, 5:634–635
 derivation, 5:635–638, 5:637f
 with gridded weather forecast data sets, 5:643
 limitations, 5:642–643, 5:642f
 net radiation, 5:640–641
 parameterization of surface resistance, 5:639–640, 5:639f
 parameters in, 5:638–640
 soil heat flux, 5:641–642
 Per- and poly-fluoroalkyl substances (PFAS), 1:551–552, 2:65, 3:297–298
 applications, 3:298
 chemical structure, 3:300–301
 concentrations, 3:298t
 degradation process, 3:300–301
 environmental compartments, 3:298
 physicochemical properties, 3:299–300, 3:299t
 sorption mechanisms, 3:304
 Perchloroethylene (PCE), 1:550f
 Percolating pore space, 5:306ge
 Percolation theory, 4:77, 4:79, 4:79f, 5:309, 5:353ge, 5:361
 Percolation threshold, 5:250–251, 5:306ge
 Perfluoroalkyl acids (PFCAs), 3:306
 Perfluorooctane amidoalkyl betaine (PFOAB), 3:300
 Permafrost, 1:501, 4:252
 Permafrost-affected soils
 characteristics, 5:411–414
 coarse-scale feedbacks, 5:415
 cryo-pedogenetic process, 5:414
 degradation, 5:416
 energy and water balances, 5:414–415
 gas transport, 5:415–416
 hydrology, 5:414–415
 irreversible process, 5:416
 physical characteristics, 5:411–414
 snow-vegetation feedbacks, 5:415
 subsidence, 5:416
 thermal and hydraulic dynamics, 5:411
 Permafrost hydrology
 catchment-scale effects, 5:424
 climate-change effects, 5:425–426, 5:425f
 seasonal development, 5:424
 springs and frost mounds, 5:424
 taliks, 5:424
 water distribution, 5:422–424, 5:423f
 Permanent charge minerals, 2:337–338
 Permanent negatively charged mineral surfaces, 2:33–34
 Permanent wilting point (PWP), 1:372, 5:511
 Permanganate oxidizable carbon (POXC), 1:373
 Permeability (intrinsic permeability), 5:366
 pore-scale modelling, 5:29–30, 5:30f
 steady-state flow through saturated medium, 5:257
 steady-state flow through unsaturated medium, 5:260
 Permittivity, 5:437–438
 Peroxs, 4:227
 Persistence, 2:399
 Persistent free radicals (PFRs), 3:415, 3:420
 Persistent organic pollutants (POPs), 2:506, 3:169
 ageing, 3:301, 3:301f
 atmospheric transport, 3:298
 availability and bioaccumulation to earthworms, 3:307
 to grazing animals, 3:306
 to plants, 3:306
 to soil microflora, 3:306
 bioavailability, 3:297
 conventional, 3:298
 per- and poly-fluoroalkyl substances (PFAS), 3:297–298
 pesticides, 3:297
 physicochemical properties, 3:298–300
 sorption mechanisms, 3:301–304
 Stockholm Convention (SC), 3:567, 3:568t
 Peruvian Andes, 3:454
 Pesticide Emission model At the Regional and Local scales (PEARL), 2:479–481
 Pesticide Leaching Model (PELMO), 2:479–481
 Pesticide Root Zone Model (PRZM), 2:479–481
 Pesticides, 2:470
 abiotic degradation, 2:475
 agricultural uses, 2:471
 biotic degradation, 2:475–476
 contamination, 2:471
 environmental effects of irrigation, 3:272–273
 fate models, 2:479–481
 leaching, 5:332
 mobility and transport, 2:472
 non-extractable (bound) residues (NER), 2:476–478, 2:478f
 persistence, 2:475–476
 physicochemical properties, 2:471
 plant uptake, 2:478
 soil organisms, 2:479
 sorption, 2:472
 trace elements, 3:290
 transfer to atmosphere, 2:474
 transfer to groundwater, 2:474
 transfer to surface water, 2:472–474
 water solubility, 2:471
 Petroplinthite, 4:54ge
 PFTNA field mobile system configuration, 2:629–630, 2:629f
 pH, 1:411–412, 1:412f
 acidity-alkalinity status, 2:8
 anthropogenic organic compounds, 2:365
 carbon dioxide concentration, 2:12–13
 dissolved organic carbon, 2:424, 2:425f
 factors affecting, 2:12–13
 for grass types, 3:501t
 green roofs, 3:481
 importance, 2:14
 indicators, 2:11
 and ion selectivity effects, on metal sorption, 2:339, 2:340f
 measurement, 2:10
 microbial denitrification, 1:269
 optodes, 2:11
 pH-sensitive dyes, 2:11
 vs. pH_{statv}, 2:305, 2:306f
 plant selection on soil, 3:501–502
 potentiometric method, 2:10–11, 2:10f
 power of, 3:501t
 salt concentration, 2:13
 scale, 2:9, 2:9f
 soil pH measurement, 2:11–12
 soil respiration, 1:373
 suspension effect, 2:13, 2:13f
 temperature, 2:305–306, 2:306f
 theoretical basis, 2:8–9, 2:9f
 Phaeozems (Udolls and Aquolls), 4:306, 4:308–309f
 Pharmaceuticals
 after rapid infiltration, 1:575t
 wastewater treatments systems, 1:575
 Phase transitions, 2:299, 2:299f, 2:300t
 Phlogopite, 4:101
 Phosphate availability
 on AMF colonization, 1:461f
 on colonization and uptake of P, 1:462f
 Phosphate ions, 2:171
 Phosphatic fertilizers, 3:371
 Phospholipid fatty acid analysis, 2:392–393
 Phosphorus (P), 2:152, 1:293, 3:372–373, 4:378
 abiotic process, 2:154–158
 adsorption, acrisols and alisols, 4:352
 arsenic, 2:205
 availability, 2:153, 2:165f
 biotic processes, 2:158–159
 cycle, 1:67–68, 1:297, 2:90, 2:153, 2:153f
 dynamics, 2:153
 and environment, 1:300
 forms and sources of soil, 1:295f
 global distribution, 2:154f
 inorganic, 2:154
 losses, 2:162–163

- Phosphorus (P) (*Continued*)
- microbial necromass, 1:295–297, 1:296f
 - microbial solubilization of inorganic, 1:299
 - mineralization and immobilization, 1:298
 - organic forms, 2:154, 2:155f
 - plant nutrition and strategies, 2:159–161
 - P-solubilizing microorganisms and plant growth, 1:299–300
 - resources and global fluxes, 2:163–164
 - in soils, 2:153–154, 2:503
 - soil testing, 2:45–46, 2:46f
- Phosphorus-containing ammonium phosphate fertilizers, 3:371
- Photosynthesis
- ground evaporation and, 5:533
 - land, 5:537–539
 - soil moisture, 5:536–537
 - stomata and, 5:530
- Phototrophic bacteria, 1:307
- Phyllo-manganates, 4:142–143
- Phyllosilicates/sheet silicates, 2:109, 4:113–118
- Phyllosphere, 1:16ge, 1:20, 1:420
- Physical crusts, 5:57
- Physical disturbance, biocrusts, 1:135
- Physical fractionation methods, soil organic matter
- aggregate fractionation methods, 2:72
 - chemical-free physical fractionations, 2:77–78
 - combined separations, 2:77, 2:78b
 - density separations, 2:75–77
 - dispersal methods, 2:72–74
 - dissolved organic matter (DOM), 2:71
 - heavy coarse organic matter fraction (hcOM), 2:71
 - isolatable soil organic matter fractions, 2:70–72
 - mineral-associated organic matter (MAOM), 2:71
 - particulate organic matter (POM), 2:70–71
 - primary particle fractionations, 2:72–78
 - size fractionations, 2:74
 - soil aggregates, 2:71
- Physically-based soil moisture retrieval, 4:621–624
- Physically-based soil water characteristic models, 5:193–194
- Physical soil degradation, 3:98–99, 3:167–168
- Physical sorption, 2:316–317
- Physiography, 1:2ge
- Phytolith, 2:265, 2:266f, 2:271
- Phytoremediation, 3:337–338
- Picacism, 3:516–517
- Piezometer, 5:164
- Pigments, 1:508
- Pipette method, 5:37, 5:38f
- Pitches, 3:493–496
- Plaggen horizon, 4:54ge, 4:59f
- Plagioclase, 4:97
- Planes and drying fronts, 5:492f, 5:493
- Planetary boundaries, 3:16
- Planning, 3:506–507
- Plant available water, 5:509
- limitations to, 5:512
 - lower limits of, 5:511–512
 - role of soil management, 5:513
 - role of soil texture and structure, 5:513
 - upper limit of, 5:510–511, 5:510f
 - use of, 5:513–514
- Plant defense, root exudates, 1:353
- Plant diversity, 3:442–443
- Plant-feeding nematodes, 1:105
- Plant growth promoters
- bacteria, 1:420
 - commercialization, 1:426
 - mechanisms, 1:421–426, 1:421f
- Plant-growth-promoting bacteria (PGPB). *See also* Microbial inoculants
- ACC deaminase, 1:423
 - antibiotics, 1:423
 - cell wall degrading enzymes, 1:424
 - cytokinin synthesis, 1:423
 - Fe sequestration, 1:422
 - gibberellin synthesis, 1:423
 - HCN, 1:424
 - IAA production, 1:422–423
 - indirect mechanisms, 1:423–424
 - induced systemic resistance, 1:424
 - N fixation, 1:421–422
 - P and K solubilization, 1:422
 - siderophores, 1:424
- Plant growth promoting fungi
- microbial consortia, 1:425
 - mycorrhizal fungi, 1:424
 - trichoderma, 1:424–425
- Plant-growth-promoting rhizobacteria (PGPR), 1:273
- Plant growth, soil water repellency (SWR), 5:212
- Plant-microbe interactions, 1:556
- Plant root-derived C dynamics, 2:390
- Plants
- nutrients (fertilizers), 3:271–272, 3:272f
 - nutrition in sodic soils, 4:378
 - physiology, root exudates, 1:349–350
 - residue, 3:215, 3:221
 - root architecture and sensing, 1:351–352
 - root defenses, 1:340f
 - selection on soil pH to minimize need for fertilizer and pH adjustment, 3:501–502
 - and soil health, 3:442–443
 - water stress, 5:529
- Plant-soil feedbacks, 1:470, 1:472
- Plant-soil-water relations
- averaged sink term distribution, 5:558–559, 5:560f
 - conductance, 5:555
 - downscaling, 5:561–562
 - evapotranspiration (ET), 5:552–553
 - flow process, 5:556, 5:557f
 - hydraulic conductivity, 5:555–556, 5:555f
 - hydraulic lift, 5:556, 5:557f
 - membrane reflection coefficient, 5:567–569
 - non-saline soils, uptake in, 5:558–561
 - parallel root system, 5:564–565
 - partial potentials, 5:554–555, 5:554t
 - relative transpiration *vs.* stress index, 5:565–566, 5:566f
 - resistance networks, 5:564, 5:564f
 - root length density, 5:562, 5:563f
 - root redistribution, 5:556, 5:557f
 - root system conductance and water treatment, 5:561, 5:561f
 - root water uptake compensation, 5:556
 - saline soils, uptake in, 5:567–569
 - simulated transpiration rate reduction, 5:566–567, 5:567f
 - standard uptake density (SUD), 5:558f
 - standard uptake fraction (SUF), 5:559–560
 - stress factor, 5:565, 5:565t
 - total water potential, 5:554–555, 5:554t
 - transpiration and root system conductance, 5:560, 5:561f
 - transpiration reduction function, 5:567, 5:568f
 - upscaling and reducing model complexity, 5:564–567
 - uptake including soil hydraulics, 5:561–563
 - water flow path, 5:553–554, 5:553f
 - water fluxes in 3D root system, 5:558–559, 5:559f
 - zero-dimensional (0D) approach, 5:567
- Plant-water relations
- direct indicators, 5:518–519
 - evapotranspiration (ET), 5:517
 - expansive growth, 5:523
 - Höfner-Thoday diagram, 5:519f
 - indirect indicators, 5:519–521
 - in irrigation scheduling, 5:521–522
 - leaf/canopy temperature, 5:520–521, 5:521f
 - leaf water potential, 5:517
 - plant water status, 5:517–518
 - relative water content (RWC), 5:518
 - root signals and application, 5:524–525
 - sap flow, 5:519–520, 5:520f
 - soil water status, 5:519
 - spectral reflectance techniques, 5:520–521, 5:521f
 - stem diameter variations, 5:520, 5:520f
 - stomatal conductance, 5:519
 - stomatal responses, 5:524
 - transpiration, 5:516–517
 - water deficit, 5:516, 5:522, 5:523f, 5:523t
 - water potential and components, 5:518, 5:519f
 - water stress and sensitivity, 5:523t
- Plasticization, of humic substances, 2:360–361, 2:361f
- Plastic mulch, 3:396–397, 3:396f
- Plastics. *See also* Microplastics
- chemical structure, 3:309, 3:310f
 - conventional, 3:309
 - debris, 3:312, 3:312f
 - definition, 3:309
 - degradation, 3:311–312, 3:312f
 - direct and indirect sources, 3:312–313, 3:313t, 3:313f
 - mulches, 3:311, 3:311f
 - use in agriculture, 3:310, 3:310f
- Plate count method, 1:141, 1:141f

- Playgrounds, surface quality and safety design of, 3:489
- Pleistocene eolian deposits, 4:247f
- Plinthite, 4:359
- Plowing, 3:434–435f
- Podzolization, 4:181, 4:266
- Podzols, 2:110, 2:423, 4:283, 4:283–284f
- Point-based sensors, 4:585–587
- Point of zero charge, 2:152ge, 2:368ge, 2:372–373, 2:372t
- Point of zero net charge, 2:368ge, 2:373–375
- Point of zero net proton charge, 2:368ge, 2:373
- Poise, 2:19–20, 2:20t
- Poisson–Boltzmann (PB) equation, 5:338–339
- Poisson probability distribution, 5:430
- Polarization, 2:549
- Pollutants, attributable to irrigation, 3:271–273
- Pollution indices, 3:567t
- Polonium (Po), 2:215
- Polyaromatic hydrocarbons (PAHs), 2:426, 2:548, 3:415
- Polychlorinated biphenyls (PCBs), 2:506
- Polychlorinated dibenzo-p-dioxins and -furans (PCDD and Fs), 2:417, 3:420–422
- Polygenetic soils, 4:214f
- Polymerase chain reaction (PCR) methods, 1:149
- Polymeric xenobiotics, 2:65
- Polynuclear aluminum species, 2:168, 2:170
- Polyphasic approach, soil algae, 1:65
- Polyvalent cation, 1:388–390
- Ponding, 4:751ge
- Pool
chlorine, 2:237
of forest floor, 2:240–241
iodine, 2:240
- Population, soil, 4:487
- Pore-geometry effects, hysteresis, 5:279
- Pore network models, 5:27–30, 5:27f, 5:33, 5:197
- Pores
classification, 5:18
dynamics, 5:23–24
nature, 5:18–19
size characterizations, 5:18–19
size distribution, 5:19
space, 2:411–412, 5:16
structure, 5:26
- Pore-scale modelling, 5:389
application, 5:30
considerations, 5:497
existing codes, 5:30–31
homogenization, upscaling and beyond, 5:31–33
input data and characterization, 5:26–27
numerical methods, 5:28–30
pore-network extraction procedure, 5:27f
resources, 5:32t
- Pore-size distribution, 5:298, 5:298f
features, 5:22–23, 5:22f
hydraulics, 5:20–21
measured geometry, 5:19
soil process, 5:23–24, 5:24t
- Porosity, 5:366
measurement, 5:18
in natural soils, 5:17–18
structural component, 5:18
textural component, 5:18
- Porous media, unstable preferential flow
checking stability criteria using experimental evidence, 5:317
continuum models, 5:319
dynamic contact angle, 5:317–318
finger flow pathways, 5:316f
fingers formation and development, 5:315f
modeling gravity-driven fingers fingered flow, 5:318–319
observations, 5:315–316
pore network models, 5:319
semi-continuum modeling, 5:319
wetting front instability, conditions for, 5:316
- Portable in situ wind erosion laboratory, 3:229
- Portable X-ray fluorescence (PXRF) spectroscopy, 4:166, 4:586
- Positional uncertainty, 4:677
- Positron emission tomography (PET), 5:466–467
- Post-transcriptional regulation, 1:276
- Potassium (K), 1:7–8, 2:118–119, 2:214–215, 2:261, 3:373, 4:378
availability, 2:120, 2:120–121f
cycle, 2:118f
fertilizers, 3:373
fixation, 2:118f
in plant nutrition, 2:119
structure and physical properties, 2:119
- Potassium chloride, 3:371
- Potentially hazardous proteins in soil, 2:403
- Potentially toxic elements (PTE), 2:417
- Potentiometric method, 2:10–11, 2:10f
- Power compensation differential scanning calorimetry, 2:667–668
- Power laws, 4:734
- Prairies. *See* Grassland soils of cool-arid–temperate ecozone
- Praseodymium (Pr), 2:214–215
- Pratylenchus agilis*, 1:106f
- Precipitation, 1:584, 2:284–285, 5:668, 5:675
adsorption and, wastewater treatments systems, 1:573
of alkaline earth metals, 2:127–128
climate change, 2:697
of metals and minerals, 2:248–251
phosphorus, 2:158
of soil minerals, 2:291
wastewater treatments systems, 1:573
- Precision agriculture (PA), 4:519, 4:587
active *vs.* passive sensors, 4:716
animal data, 4:719
attribute mapping, 4:713–714
crop data, 4:718, 4:718–719f
decision-making, 4:721, 4:723–724f
decision support systems (DSSs), 4:714
definition, 4:711–712
differential action, 4:714
vs. digital agriculture, 4:722–724
- digital soil mapping (DSM), 4:720–721, 4:720f
emergence, 4:710–711
geo-referencing, 4:713
global navigation satellite systems (GNSS), 4:714–715
interpolation, 4:720
management units (MUs), 4:722, 4:723–724f
meteorological data, 4:719–720
pedometrics, 4:720–721, 4:720f
physical *vs.* chemical *vs.* biological sensors, 4:715
practical implementation, 4:713–714, 4:714f
production system, 4:712, 4:712–713f
proximal *vs.* remote-sensing, 4:716
sensing systems, 4:715–716
soil data, 4:716–717, 4:717f
variability, 4:712, 4:712–713f
- Precompression stress, 5:78, 5:78f
- Predation, 1:44–45
- Predictive models for solute and gas diffusivity, 5:250–251
- Preferential flow, unstable
checking stability criteria using experimental evidence, 5:317
continuum models, 5:319
dynamic contact angle, 5:317–318
finger flow pathways, 5:316f
fingers formation and development, 5:315f
modeling gravity-driven fingers fingered flow, 5:318–319
observations, 5:315–316
pore network models, 5:319
semi-continuum modeling, 5:319
wetting front instability, conditions for, 5:316
- Preferential leaching process
aggregated soil, 5:325–326
macroscopic theory, 5:324
mobile fraction, 5:326
non-equilibrium process, 5:324
non-mechanistic transfer-function schemes, 5:325
pedogenic and biogenic process, 5:325
process-based transport schemes, 5:325
- Prescribed (controlled) fires and post-harvest fires, 2:434
- Pressure potential, 5:162, 5:164
- Primary minerals, 4:91
feldspars, 4:96–100
micas, 4:100–102
olivines, 4:102–103
parent materials, 4:92
pyroxenes and amphiboles, 4:103–107
silica minerals, 4:93–96
weathering, 4:91
- Primary wastewater treatments systems, 1:568
- Priming effect, 1:235, 1:339
- Principal component analysis (PCA), 2:346, 4:517
application, 4:501–502, 4:502t, 4:503f
theory, 4:501, 4:501f

- Prions, 2:403
- Probability distribution, 4:672–673, 4:673*f*, 4:676–677, 4:680
- Probability sampling method
- balanced and well-spread sampling, 4:542–544
 - cluster random sampling, 4:541–542
 - design-based approach, 4:537
 - estimators, 4:538–539
 - model-based approach, 4:536–537
 - model-based optimization, 4:545
 - vs.* non-probability sampling, 4:536–537
 - sample size determination, 4:544–545
 - sampling designs, 4:537
 - simple random sampling, 4:539
 - statistical inference, 4:536
 - stratified simple random sampling, 4:539–540
 - systematic random sampling, 4:540–541
- Processed natural materials, green roofs, 3:483
- Processes at field scale, 5:497–498
- Production systems, 3:104–105, 3:323–331
- Prokaryotes, 1:563, 2:243
- Prokaryotic algae, 1:62–64
- Propagules of arbuscular mycorrhizal fungi, 1:453, 1:466–467
- Property rights, 3:549–550, 3:550*f*
- Prophage, 1:16*ge*
- Prophylactic fertilizer, 3:369*ge*
- Protactinium (Pa), 2:214–215
- Protective agroforestry systems, 3:76
- Protective barrier, root mucilage, 1:337
- Protein misfolding cyclic amplification (PMCA), 2:403
- Proteins
- adsorption, 2:397–399, 2:398*f*
 - characterization, 2:399–400
 - extraction, 2:399–400
 - glomalin-related soil protein (GRSP), 2:401–403
 - potentially hazardous, 2:403
 - quantification, 2:399–400
 - in soil, 2:397
 - soil enzymes, 2:400–401
 - soil metaproteomics, 2:404
 - study of adsorbed, 2:399
- Protists
- abundance and diversity, 1:42
 - agricultural systems, 1:46
 - agroecosystems, 1:40
 - autotrophic, 1:40
 - as bioindicators, 1:47
 - community composition, 1:43–44
 - deterministic processes, 1:43–44
 - diversity and ecology, 1:40–46
 - ecological functions, 1:39
 - free-living soil protists, 1:40*f*
 - geographical distribution and diversity patterns, 1:42–44
 - heterotrophic, 1:39
 - morphotypes, 1:39
 - nutrient cycling, 1:45–46
 - parasites, 1:45
 - plant performance, 1:46
 - predation, 1:44–45
 - soil organic matter (SOM) decomposition, 2:390
 - supergroups, 1:40
- Proton generation and consumption processes, soil-plant system, 2:2*t*
- Proton surface-charge density, 2:370*f*
- Protura
- anatomy, 1:86*f*, 1:92
 - developmental stage, 1:92
 - food, 1:92
 - habitat, 1:72*t*, 1:92
 - interactions, 1:92
 - interesting fact, 1:92
- Proximal land surface temperature (LST)
- brightness temperature, 4:601–602
 - factors affecting, 4:602–603
 - measurements and application, 4:607–608
 - microwave radiometry, 4:605–607
 - periodic variations, 4:603
 - thermal infrared sensing, 4:603–605
- Proximal sensing, 4:58
- electromagnetic induction, 4:596–597
 - energy balance-based, 4:612–615
 - evapotranspiration, 4:610
 - methods (spectroscopy), 2:30
 - of soil moisture, 4:593
 - water balance-based, 4:615–616
- Proximal soil sensing (PSS) in field
- advantages, 4:581*t*, 4:582
 - calibration and spectral libraries, 4:587–588
 - COSMOS, 4:584–585
 - definition, 4:580
 - disadvantages, 4:582
 - electromagnetic induction (EMI), 4:582–584, 4:583*f*, 4:583*t*
 - gamma-rays, 4:584
 - ground-penetrating radar (GPR), 4:585
 - multi-sensor data fusion, 4:588
 - on-the-go sensors, 4:582–585, 4:583*f*
 - point-based sensors, 4:585–587
 - portable X-ray fluorescence spectrometry, 4:586
 - sensor classification, 4:580–582
 - soil compaction and strength, 4:585
 - soil gamma emission, 4:584*t*
 - soil moisture sensors, 4:587
 - stop-and-go sensors, 4:587
 - time-domain reflectometry (TDR), 4:585
 - types of sensors, 4:582–587
 - visible, near and mid-infrared diffuse reflectance spectroscopy, 4:585–586
- Proximal thermal infrared sensing, 4:603–605
- Psammets, 4:230
- Pseudomonas protegens* Pf-5, 1:281
- Pseudoscorpiones, 1:86–87
- P-solubilizing bacteria (PSB), 1:429
- Psychrometer, 5:164, 5:164*f*, 5:178
- equation, 5:180
 - field thermocouple, 5:180*f*
- Public areas, 3:490, 3:502
- Public health concerns, environmental effects of irrigation, 3:273–274
- Public preference, 3:507
- Pulse and press changes, 1:580
- Pulse-chase method, 1:197, 1:198*f*, 1:199–200
- Pulsed fast/thermal neutron analysis (PFTNA), 2:627–628
- acquiring data for elemental mapping, 2:637–638
 - calibration dependencies, 2:631–635
 - comparison of chemical analysis and, 2:636, 2:636*t*, 2:636*f*
 - determination of total field soil carbon content, 2:638
 - measurement time, 2:637
 - peak area determinations, 2:631, 2:633*f*
 - raw data, 2:630, 2:630–631*f*
 - soil volume associated with, 2:637
 - system background spectra, 2:631, 2:632–633*f*
- Pulse length-based concentration determination (PULCON), 2:618
- Purifying capacity and regulation, of contaminants, 3:189
- PXRF. *See* Portable X-ray fluorescence (PXRF) spectroscopy
- Pyrite oxidation, 2:1–2
- Pyrochar, 2:433
- Pyrogenic carbon (PyC), 2:65–66, 2:451
- black carbon continuum, 2:433
 - black nitrogen (BN) formation, 2:439
 - cellulose and hemicellulose transformation, 2:438
 - chemical structure characterization, 2:434–437
 - conceptual chemical structure model, 2:439–440
 - as C sink, 2:444
 - decomposition temperature, 2:437–438
 - definition, 2:433
 - degradation pattern, 2:442, 2:442*t*
 - degradative technique, 2:435–436
 - distribution, 2:433
 - dynamics, 2:433
 - elemental analysis, 2:434, 2:435*f*
 - Fourier transform infrared (FTIR) spectroscopy, 2:436
 - gas chromatography and mass spectrometry (GC, MS), 2:435
 - during heating, 2:437–440
 - identification and quantification, 2:441
 - lignin alteration, 2:439
 - lipids fate, 2:439
 - nitrogen and phosphorus availability for plants, 2:440
 - origin, 2:433–434
 - porosity, 2:441
 - preservation, 2:443–444, 2:443*f*
 - soil color, 2:440
 - soil organic matter (SOM) quality, 2:437*f*, 2:440
 - soil pH, 2:440
 - soil properties, 2:433, 2:440–441
 - solid-state ¹³C nuclear magnetic resonance (NMR) spectroscopy, 2:436–437*f*, 2:437
 - spectroscopic techniques, 2:436–437, 2:436–437*f*
- Terra Preta de Índios soils, 2:433

thermogravimetric analysis (TGA) with differential scanning calorimetry (DSC), 2:435–436
 Van Krevelen diagram, 2:434, 2:435f
 vegetation fires, 2:433
 Pyrogenic carbonaceous matter (PCM), 2:323–324
 Pyrogenic organic matter (PyOM), 2:71–72, 2:389
 Pyrolysis, 2:66, 2:407, 3:328
 Pyrolysis temperatures, of black carbon, 2:362–363
 Pyrophyllite, 4:115
 Pyroxenes, 4:92–93, 4:103–107, 4:105t

Q

Qinghai-Tibetan Plateau (QTC), 4:275–278
 Qualifiers, 4:207
 Quantification, proteins, 2:399–400
 Quantitative models, for soil studies, 4:472t
 Quantitative NMR, 2:616–618
 Quantitative PCR (qPCR), 1:149
 Quantitative pedogenesis, 4:483–485, 4:484f
 soil-landscape evolution, 4:483–485, 4:484f
 soil processes, 4:483
 Quantitative soil data, 2:567
 Quantitative soil geography, 4:474–477
 classification paradigm, 4:476
 pedodiversity, 4:476–477
 soil classification, 4:474–476
 soil map units, 4:474–476
 Quantitative trait locus (QTL) analysis, 1:328
 Quantity (Q), 2:116, 2:120, 2:121f
 Quantity and quality of dissolved organic carbon, 2:424
 Quantum chemistry (QC) method, 2:485–488, 2:486f
 Quartz, 2:267f, 4:92, 4:145–146, 4:350
 Quaternary geology, 4:182
 Quenching process, 2:548–549
 Quick X-ray absorption fine structure (Quick-XAFS) spectroscopy, 2:346
 Quinoline degradation, 2:565
 Quorum-sensing, 1:37

R

Radar systems, 4:620
 Radiation balance, 5:393–395, 5:398–399
 Radiative transfer models (RTMs), 4:622
 Radioactive decay, 2:257
 Radioactive isotopes
 available nutrients/phosphorus in soil, 2:99–100, 2:103
 basic principles, 2:97–98
 becquerel, 2:97
 classes, 2:97
 detection and measures, 2:98
 direct labeling approach, 2:104, 2:104–105f
 disintegrations per minute (DPM), 2:98
 E-value, 2:99–100

fertilizer fate, 2:103–105
 future research, 2:105–106
 gross phosphorus mineralization, 2:102, 2:102f
 half-life of radioactive isotope, 2:97
 indirect labeling approach, 2:104–105
 isotope exchange kinetics (IEK), 2:100–102, 2:101f
 isotopic dilution principle, 2:98–99
 limitation, radiotracers, 2:105
 liquid scintillation counter (LSC), 2:98
 L-value, 2:103, 2:103f
 mineral fertilizers, 2:97
 nutrient pools and processes, 2:96, 2:98
 radioactive decay, 2:97
 soil phosphorus dynamics, 2:100–102, 2:101f
 in soil-plant systems, 2:103
 in soil science applications, 2:97, 2:98t
 in soil-solution systems, 2:98–102
 stable isotopes, 2:97
 units of radioactivity, 2:97
 unstable isotopes, 2:97
 Radioactive soils, 1:506
 Radiometers, 4:605–607, 4:606f
 Radionuclides
 bioavailability, 2:258–259, 2:258f
 chemistry, 2:257
 concentration effect, 2:257
 decay, 2:257
 definitions, 2:256–257
 dynamics, 2:257–258
 external and internal exposure, 2:258
 remediation of contaminated soils, 2:259
 soil ingestion, 2:259
 soil microbiological activity, 2:260
 soil-plant transfer, 2:259
 soil to human food chain, 2:259
 Radiotracer techniques, 1:162–163
 Radium, 2:132–133, 2:231
 Radon, 2:231, 2:232f, 2:261–262, 3:563
 Rainsplash erosion, 3:197
Ralstonia solanacearum, 1:21
 Raman scattering, 2:497–498
 Raman spectroscopy, 2:497–499
 in soil components analysis, 2:500–503
 in soil contaminants analysis, 2:503–507
 Ramsar convention, 4:437
 Random effects, 4:495–496
 Random forests, 4:505
 Random functions, 4:656–658, 4:656f, 4:657t
 Randomization, 4:487–488, 4:489f
 Random sampling in mines, 4:657
 Random variables, 4:656–658, 4:656f, 4:657t, 4:672–673
 Random walk method, 5:360
 Range, 4:660
 Rangelands, 3:88
 characteristics, 3:88
 definition, 3:88
 degradation, 3:96–100
 management, 3:89–96
 pastoral systems, 3:89
 spatial and temporal heterogeneity, 3:88
 types, 3:89

Raoult's law, 5:284
 Rapeseed, 3:6, 3:7f
 Rapid-infiltration process, 1:571–572
 Rashtriya Krishi Vikas Yojana (RKVY), 3:469–470
 Rate of heating, 3:416
 Ratio estimator, 4:538
 Rayleigh scattering, 2:497–498, 2:498f
 Reactions
 heterogeneity, 2:298, 2:298f
 layered double hydroxides (*see* Layered double hydroxides (LDHs))
 oxido-reduction (*see* Redox reactions)
 phase transitions, 2:299, 2:299f, 2:300t
 Reactive oxygen species, 2:541
 Readily available water (RAW), 5:513
 as ecological indicators, 5:514
 in hydrological models, 5:514
 in irrigation management, 5:513
 Recalcitrance, 2:389
 Recalcitrant compounds, 2:389
 Reclamation, of salt affected soils, 3:243–244
 Recommended fertilizer rate, 3:369ge
 Recyclable and reclamation products, 3:325
 Recycled mineral materials, green roofs, 3:483–484
 Recycled paper mulch, 3:395
 Redox depletions, 4:430–431, 4:431f
 Redoximorphic features
 cold region soils, 4:257, 4:258f
 hydric soils, 4:429–432
 Redox potential
 anaerobic soils, 1:381
 wetlands, 4:445
 Redox reactions, 2:246–248, 2:246–247f, 2:251–252, 4:180. *See also* Layered double hydroxides (LDHs)
 critical electron potential, 2:302, 2:303f, 2:303t
 at equilibrium, 2:301
 oxidative phosphorylation, 2:300, 2:300f
 pH values, 2:305–306, 2:305–306f
 redox couples classification, 2:301, 2:302f
 rH measurement, 2:304, 2:304f
 theory, 2:301
 Redox-responsive RegB-RegA two-component system, 1:276
 Reduced tillage/no-till systems, 3:429
 Reductases, of oxides of nitrogen, 1:265t
 Reference evapotranspiration, 5:573–574, 5:643
 Reference soil groups, 4:207, 4:208f, 4:209t
 Reflectance spectroscopy, 4:632
 Reforestation, 1:238
 Regenerative agriculture (RA), 3:15–16
 Regions of interest (ROI), 2:673
 Regosols, 2:423, 3:44
 Regression approach, 4:497–498, 4:648, 4:649t
 Regression-kriging (RK) approach, 4:479, 4:690
 Reinforcement learning, 4:510–511
 Reinforcement, soil
 branching intensity, 5:120–121
 Factor of Safety (FoS), 5:121
 Fiber Bundle Model (FBM), 5:120, 5:121f

- Reinforcement, soil (*Continued*)
 geotechnical centrifuges, 5:120–121
 Mohr–Coulomb theory, 5:118–119
 plant roots, 5:118–121
 root area ratio (RAR), 5:115
 shallow landslide, 5:115, 5:116f, 5:121f
 soil cohesion, 5:119–120
 soil shear strength, 5:115–118
 vegetation and hydraulic properties, 5:118, 5:119f
- Relative humidity (RH), 3:387
- Relative percentage difference (RPD) values, 2:501
- Relative permeability, 5:280–281
- Relative scarcity, 3:547ge
- Relative water content (RWC), 5:518
- Relaxation time, of compound, 2:618, 2:619t
- Relay cropping, 3:447
- Reliability, of inoculation, 1:454
- Relief, 4:10–11, 4:11f
- Remediation, 2:584
 soil, 3:170, 5:222
 with vegetation, 4:379, 4:380f
- Remote sensing of soil moisture
 applications, 4:627, 4:628f, 4:628–629t
 instrument characteristics, 4:620–621
 model inversion, 4:624
 physically-based retrieval algorithms, 4:624
 physical retrieval models, 4:621–624
 platforms and instruments, 4:619–621
 retrieval, 4:621–627
 space-borne systems, 4:620, 4:620f
 tower and airborne systems, 4:619
- Rendolls, 4:229
- Repellency index (RI), 5:208
- Replication, 4:487–488, 4:489f
- Re-seeding, 3:109–110
- Reserve acidity, 2:17
- Residence time, 2:348, 3:416
- Residual accumulation hypothesis, 4:361
- Residual herbicide, 3:439ge
- Residual maximum likelihood (REML), 4:663
- Resilience, 1:580
- Respiration, 5:219. *See also* Soil respiration
- Retardation effect, 5:131–132
- Retardation factor, 5:657, 5:659f
- Retention
 of bromine, 2:239
 of chlorine, 2:238
 of fluorine, 2:237
 of iodine, 2:240
 mechanism in soil, 2:236
- Retisols, 4:282
- Reverse heteronuclear saturation transfer difference (RH-STD), 2:620
- Reverse osmosis, 5:266–267
- Rheic histosols, 4:392
- Rhizobia, inoculation of, 1:449–452
- Rhizobiome, 1:448–449, 1:453–454
 moist microenvironment for, 1:338
 mucilage and root-associated bacteria, 1:338–339
 mucilage transfers signals from plant roots to, 1:338
- Rhizobium-legume symbiosis, 1:273
- Rhizodeposition, 1:448, 1:454, 5:547.
See also Root water uptake
- Rhizodeposition flux, 1:233
- Rhizosheath, 5:11
- Rhizosphere, 1:2ge, 1:31ge, 1:343–344, 1:415, 1:420, 2:429–430, 2:676–677, 5:11
 of crops and suppress root pathogens, 1:464
 definition, 1:472
 effects, 1:469
 hydraulic properties, 5:549–550
 interaction, 1:472, 1:472f
 iron (Fe), 2:188
 microorganism groups, 1:450–451t
 organic carbon exudates within, 1:477
 plant growth inoculation promoting, 1:452–453
 positron emission tomography (PET), 5:466–467
- Rice cultivation, 3:155
- Rice Terraces of the Philippine Cordilleras, 3:453
- Richardson–Richards equation, 4:751ge, 4:756, 5:261, 5:324, 5:382–383
- Richness–area relationships, 4:730
- Ridges, 3:203, 3:204f
- Rights-of-soil, 3:16–17
- Rill erosion, 3:197, 3:198f
- Riparian buffer systems (RBS) in southern Ontario, Canada, 3:77f, 3:83
- Ripening, 5:349
- Risk analysis, 4:682
- Road dust, 3:223
 chemical compounds in, 3:225t
 from dust to dose, 3:231–232, 3:233f
 emission factors for road wind erosion, 3:231, 3:231f
 emissions, 3:227–228
 health effects, 3:226–227
 particle deposition in respiratory system, 3:226f
 roadside sampling methodologies, 3:228–230
 road traffic resuspension emissions, 3:227
 road wind erosion, 3:227–228
- Robust estimators, 4:659
- Rock fragments, 4:572
- Rock weathering, 2:464–465, 2:465f, 4:460
- Rocky deserts, lithic microbial colonization, 1:504–505
- Rodent excavation and burrowing activity, bioturbation, 5:106–107
- Rods, 5:437, 5:440–442
- Rooftop water harvesting, 5:602–603, 5:603f
- Root–brain hypothesis, 3:60
- Root exudates, 1:344
 characterization, 1:353
 composition, 1:344
 environmental factors, 1:344
 factors affecting, 1:349–351, 1:350f
 fluorescence imaging, 1:354
 functions, 1:351–353
 gene knockout procedures, 1:354
 inorganic atoms and compounds, 1:344t, 1:345–346
 isotope tracing methods, 1:354
- and microbial responses, 1:344
 microRNA (miRNA), 1:355
 organic compounds, 1:345t, 1:347–348
 production, 1:345–349
 release, 1:348f
 as signal molecules, 1:344
 zymography, 1:354
- Root form and function.
See also Reinforcement, soil
 carbon sequestration, 1:328
 drivers, 1:328, 1:329f
 ecosystem process, 1:328–330
 genetic analysis, 1:328
 nutrient cycling, 1:328–329
 plant growth, 1:321
 productivity, 1:328
 quantitative trait locus (QTL) analysis, 1:328
 root anatomy, 1:322, 1:322f
 root classification, 1:323–324, 1:323f
 root development, 1:322–323
 soil aggregation and stability, 1:330
 stress tolerance, 1:325–327
 trade-offs in root traits, 1:324–325, 1:324f
 water budgets and hydraulic redistribution, 1:329
- Root length density (RLD), 1:316–317
- Root mucilage
 abiotic stress and, 1:336–337
 binding of trace metals, 1:338
 biotic stress and, 1:336t, 1:337
 carbon source for rhizobiome, 1:339
 components, 1:334t
 extracellular DNA inactivates infectious root pathogens, 1:339
 factors affecting production, 1:336
 functions, 1:337–338
 nutrient absorption, 1:337–338
 pathogens prevention from root contact, 1:339
 plant root defenses, 1:340f
 production and chemistry, 1:333–335
 protective barrier, 1:337
 research methods for collection, 1:339–341
 secretion, 1:335f
 water retention, 1:337
- Root-rhizobiome interaction
 chemical gradients, 1:363–365, 1:364f, 1:365–366t, 1:366f
 chemical signaling, 1:358–362
 ecological interactions, 1:361–362
 endophytes, 1:361, 1:361f
 nutrient availability, 1:365, 1:366f
 photosynthates, 1:358
 physical gradients, 1:362–363
 quorum-sensing molecules, 1:360
 redox condition, 1:365
 rhizosphere soil pH, 1:363–364, 1:364f, 1:365t
 salt tolerance, 1:365–367, 1:366t
 signals from plants to microbes, 1:359–360, 1:359f, 1:360t
 symbiosis pathway, 1:358
 volatile chemical molecules release, 1:359–360, 1:360t
 water acquisition, 1:363

- Roots and mycorrhizae, 3:60
 Root signals, 5:524–525
 Root-soil entanglement, bioturbation, 5:104f
 Root water uptake
 conductance per root surface, 5:546
 controlled laboratory studies, 5:546–547
 in ecological and environmental context, 5:550
 effective conductance, 5:547
 field evidence, 5:547
 hydraulic conductivity, 5:544–545, 5:545f
 hydrophobic root diffusion barriers, 5:546
 mucilage, 5:547–548, 5:548f
 mycorrhizal fungi, 5:548–549
 osmolality, 5:546
 pore size in cell walls, 5:546
 radial water flow, 5:546
 rhizosphere hydraulic properties, 5:549–550
 rhizosphere matric conductance, 5:545
 root hairs, 5:546–549
 root strategies for water acquisition, 5:550
 root surface area, 5:546
 soil water flow, 5:544–546
 soil water retention curve, 5:544
 transpiration rate, 5:543, 5:550
 volumetric root water uptake, 5:544–545
 water flow in root, 5:546
 water flow rate per membrane surface, 5:546
 water flux, 5:544–545
 water holding capacity, 5:544
 Root weight ratio (RWR), 1:316–317
 Root-zone soil moisture estimation, 4:624–625
 Rotating biological contactors, 1:569
 Rotational grazing systems, 3:94, 3:94t
 Rotational nuclear Overhauser effect spectroscopy (ROESY), 2:619–620
 Rotations, cropping, 3:448
 Rough surface scattering, 4:622f
 Roxarsone, 3:291–292
 Rubbery and glassy domains, of humic substances, 2:359–360, 2:360f
 Rubefaction, 4:23, 4:324–326, 4:325–326f
 Rubidium (Rb), 2:116–117
 Runoff
 concept, 5:287–288
 farming, 5:594, 5:595f
 generation, spatial scale, 5:422
 snowcover, 5:422
 Runoff and soil erosion control
 crop residue, 3:404, 3:404f
 Manning equation for flow velocity, 3:405
 mulch factor (MF), 3:405
 mulch rates *vs.* runoff, 3:404, 3:404t
 soil aeration, 3:405–406
 soil biological activity, 3:407
 soil chemical environment, 3:407
 soil structure, 3:406, 3:407t
 soil thermal regime, 3:405, 3:406f
 stover, 3:405
 wind erosion, 3:405
 Russian soil classification system (RSCS), 4:191–195, 4:193–194t, 4:194f, 4:427
 Rutile, 4:146–147
- S**
- Saharan dust, 4:44
 Salids, 4:227
 Salination, 3:168, 4:87–88
 of cultivated soils, 3:237–238
 management, 3:242–244
 on plants, 3:240–241
 and pollutants, 4:381
 on soil, 3:239–240
 Saline soils, 4:369
 Salinity, 1:8, 3:417–418, 5:437. *See also* Soil salinity
 environmental effects of irrigation, 3:271f
 on pore-water measurement, 5:442
 in soils, 3:237, 3:237f
 on water content measurement, 5:441
 Salt, 1:8, 3:282–283, 4:178
 dissolution, 3:239, 3:239f
 leaching, 5:328–330
 monitoring of affected soils, 3:242
 precipitation, 3:239, 3:239f
 sieving, 5:267
 transport, 3:239, 3:239f
 Saltation, 3:212–213, 3:213f, 3:215–216, 3:218f
 Samovar model, 2:16–17, 2:16–17f
 Sample handling procedure
 application, 4:552
 in field (*see* Field sampling)
 in laboratory (*see* Laboratory sample handling)
 transport from field to laboratory, 4:561
 Sampling design, 4:481
 Sampling for soil testing, 2:44–45
 Sand layer in flow slab experiment, 5:315f
 Sandy deserts, 1:504
 Sanitation, 1:441
 Sap flow, 5:519–520, 5:520f
 Sapristis, 4:226
 Satellite-based sensors, 5:475
 Satellite missions, 4:620
 Saturated areas, overland flow, 5:288
 Saturated buffers, 3:263
 Saturation excess
 hydrology, 5:288f
 runoff models, 5:291–292
 Saturation transfer difference (STD)
 techniques, 2:620, 2:621f
 Saturation transfer double difference (STDD), 2:620, 2:622f
 Savanna, 4:29
 Scaling of transport properties
 advection-dispersion equation (ADE), 5:354, 5:357
 continuous-time random walk (CTRW) framework, 5:360
 dispersion, 5:354
 fractional advection-dispersion equation (FADE), 5:359–360
 Gaussian solute transport, 5:355
 heterogeneity, 5:361–362
 mobile-immobile (MIM) model, 5:358–359
 percolation theory, 5:361
 scale, 5:354–355
 solute spreading, 5:356
 transfer function model, 5:359
 Scanning electron microscopy (SEM), 2:578
 aggregation, 5:13
 instrumentation, 2:578, 2:579f
 mineral-associated organo-mineral (MAOM), 1:401
 strengths and limitations, 2:579–580
 Scanning transmission electron microscope (STEM), 2:584–586
 Scanning transmission X-ray microscope (STXM), 1:401–402
 Scenting, soil, 3:514–515, 3:515f
 School yards, surface quality and safety design of, 3:489
 Schrödinger equation (SE), 2:484
 Schwertmannite, 4:141
 Scintillometer method
 advantages, 4:615
 evapotranspiration, 4:614–615
 limitations, 4:615
 operation principle, 4:615f
 SCORPAN model, 4:477–479, 4:700–702
 Sealed soils, 5:62
 Seal formation, 5:59
 factors affecting, 5:59
 modeling, 5:60–61
 Seal hydraulic properties, 5:61
 Seal layers, 5:58–59
 polymer application and, 5:62
 for rainfall-runoff-infiltration partitioning, 5:59
 Seals, 5:57
 Seasonally frozen soil hydrology
 climate-change effects, 5:425
 hydraulic conductivity, 5:420
 hydrological fluxes, 5:421f
 land-use effects, 5:422
 macropores, 5:420–422
 runoff generation, spatial scale, 5:422
 snowcover effects, 5:422
 Secondary ion mass spectrometry (SIMS), 2:671
 Secondary soil minerals, 2:642–643
 Secondary wastewater treatment
 activated sludge, 1:569
 systems, 1:568–569
 trickling filter, 1:569
 Secretion, of root mucilage, 1:335f
 Sediment production, 3:189
 Seebeck effect, 5:267
 Seedbed preparation, 3:434, 3:434f
 Seeding, 3:434f
 Seed treatment, 1:435, 1:435f
 Seepage, 4:406, 5:672, 5:674
 Seismic methods, 5:452–454
 Selected area electron diffraction (SAED), 2:577, 2:578f
 Selective dissolution, 2:172
 Selective herbicide, 3:440ge
 Selenium (Se), 2:192–193, 2:215
 chemical properties, 2:193
 deficiency, 3:561
 human health and soil, 3:561–562
 physicochemical properties, 2:193t
 in soil, 2:196–199

- Selenium (Se) (*Continued*)
 soil contamination, 2:218
 species in soil, 2:196
 toxicity, 3:562
 transformation processes in soil, 2:197–198
- Self-moving sprinkler systems, 5:615–617, 5:616f
- Semi-arid and arid climates, 4:17–18
- Semi-arid regions, sediments and soil formation in, 4:337–338
- (Semi-)empirical retrieval models, 4:624
- Semi-intensive and intensive green roofs, 3:480
- Semi-physical approach, pedotransfer functions (PTFs), 4:647–648
- Semivariance, 4:658
- SEM–Raman system, 2:586, 2:587f
- Sensible heat flux, evapotranspiration, 4:611, 4:613–614
- Sensors, 5:442, 4:519
 based on electric conductivity, 5:173
 classification, 4:580–582
 dielectric properties, 5:171–174
 for nuclear methods, 5:174
 selection of water content, 5:172f
 for thermal properties, 5:173
 types, 4:582–587
- Septic systems, 1:568. *See also* Wastewater treatment systems
- Sequence-based metagenomics, 1:163
- Sequential extraction procedures (SEPs), 2:56
 Community Bureau of Reference (BCR) procedure, 2:54–55
 Tessier procedure, 2:53–54, 2:54f
- Sequestration of heavy metals, 2:293–295
- Sessile drop method (SDM), 5:208
- Set-move sprinkler systems, 5:615, 5:615f
- Sewage biosolids, land application of, 1:569–577
- Sewage effluent wastewater treatments systems, 1:571t
- Shallow soils, 1:9
- Shallow water tables, evaporation, 5:495
- Shannon diversity index, 1:541, 3:369ge
- Shape match algorithms, 2:569
- Shear strength
 definition, 5:115–116
 direct shear tests, 5:116f, 5:117
 methods, 5:115–116, 5:116f
 triaxial testing, 5:116f, 5:117–118, 5:118t
- Shielding and construction materials, 2:629
- Short-range ordered (SRO) aluminosilicates
 abundance, 4:121
 for agricultural productivity, 4:132, 4:133t
 Al³⁺ and H₄SiO₄ activity in soil solution, 4:124
 allophane, 4:121, 4:122f
 aluminum solubility, 4:131, 4:131f
 andosols, 4:123
 characteristics, 4:121–122
 chemisorption, 4:129
 differential thermal analysis (DTA), 4:127, 4:127f
 electron microscopy, 4:125, 4:126f
 formation and distribution, 4:122–124
 identification, 4:124–127
 imogolite, 4:121, 4:122f
 infrared spectra, 4:127, 4:128f
 ligand exchange, 4:130
 NMR spectra, 4:127, 4:128f
 physical properties, 4:132
 podzols, 4:124
vs. precipitation, 4:123–124, 4:125f
 quantity, 4:125–126
 reaction with soil organic matter, 4:130
 stability diagram, 4:122–123, 4:123f
 structure and charge, 4:127–129
 surface charging reactions, 4:127
vs. temperature, 4:123–124, 4:125f
vs. time, 4:123–124, 4:124f
 variable-charge behavior, 4:127–129, 4:129f
 X-ray diffraction, 4:124–125, 4:126f
- Shortwave radiation, 5:393–395
- Shotgun sequencing, 1:153
- Shrinkage
 basics, 5:141–142, 5:141f
 deterministic models, 5:142
 displacement transducers, 5:142, 5:142f
 extent, 5:139
 factors affecting, 5:143
 fluid transfer models, 5:143
 measurement, 5:142, 5:142f
 modeling, 5:142
 origin, 5:140–141, 5:140f
 shrinkage curve (SC), 5:140–141f
 shrink-swell processes, 5:139
 significance, 5:143
 soil structural stability, 5:139–140
 Van Genuchten equation, 5:142
 water content *vs.* pore volume, 5:139
- Shrubs, 1:472
- Siderophores, 1:424, 2:245
 bacterial interaction, 1:36–37
 iron (Fe), 2:188
- Silage, 3:104
 collecting and wrapping, 3:106f
 management, 3:108–110
- Sil-andic, 4:458ge
- Silica minerals
 quartz, 4:93–94
 as soil parent material, 4:96
 solubility, 4:94–95
 structures and properties, 4:93, 4:93t
- Silica sorption, 2:269–270
- Silicate, 2:264f, 2:265, 2:287
- Silicic magma, 4:459
- Silicon (Si), 2:168, 2:215
 arsenic, 2:205
 cycle, 1:45
 distribution, 2:265–266
 inorganic forms, 2:266f
 isotope system, 2:109–110
 mineral weathering and clay formation, 2:267–268
 pedogenic, 2:268
 properties, 2:263
 in soil impacts pH buffering, 2:270–271, 2:270f
 in soil–plant system, 2:265–271
 terrestrial biological cycle, 2:271
- Silicon oxides, 4:145t
 opal, 4:146
 quartz and cristobalite, 4:145–146
- Silvopasture, 1:476, 3:76–78, 3:77f, 3:110
- Simazine, 5:332
- Simple Inverse Yield Model, 5:628
- Simple random sampling, 4:539
- Simulation approach, probability sampling, 4:545
- Simulation lower boundary (SLB), 4:754
- Single-cell sequencing (SCS), 1:153
- Single crop coefficient approach, 5:575–580, 5:575f, 5:577t, 5:578–579t, 5:579f, 5:579–580t
- Single metal extraction
 complex extraction, 2:53
 diluted mineral acid, 2:52
 ligands, 2:51
 low molecular weight organic acids (LMWOA), 2:52
 salts, 2:52
- Single soil measurements, 4:672
- Single-virus genomics (SVG) method, 1:19–20
- Site conditions, lysimeters, 5:677
- Site taphonomy, 3:22
- Size fractionations, soil organic matter, 2:74
- Slope gley hypothesis, 4:361
- Slope management, 3:452
- Slope steepness and slope length, 3:199
- Smectite, 2:267f, 4:116–117, 4:345–346
- Smellscape, 3:514–515, 3:515f
- Smoldering fires, 2:434
- Smoothed particle hydrodynamics (SPH), 5:197
- Smother crops, 3:447
- SNIFFER, 3:228, 3:229f
- Snowcover, 5:422, 5:425
- Social dilemma, 3:547ge
- Sodicity, 4:370, 4:372–373
 and conditions prevailing in soil, 4:375, 4:375f
 crop tolerance to, 4:377, 4:377t
 and electrolyte concentration, 4:373–374, 4:374f
 and soil productivity, 4:377–378
- Sodic soils, 2:124
 definitions, 4:369–370
 environmental aspects, 4:380–381
 mechanisms responsible for behavior, 4:371–372
 origin and distribution, 4:370
 plant nutrition in, 4:378
 reclamation and management, 4:378–379
 structural stability, 4:372–373
 water flow in, 4:373–377
- Sodium (Na), 2:214–215, 4:369
 sequestration, 2:124
 and soil productivity, 2:122–124, 2:123f
 in soils, 2:121–124, 2:122f
 toxicity symptoms, 2:123f
- Sodium adsorption ratio (SAR), 1:2, 4:369, 4:370f
- Sodium-affected soils. *See* Sodic soils
- Sodium/calcium montmorillonite

- critical flocculation concentrations (CFCs), 2:380–381, 2:381*t*
- dispersion and flocculation, 2:382, 2:382–383*f*
- electrophoretic mobility, 2:382, 2:382*f*
- exchangeable sodium percentage (ESP), 2:382, 2:382*f*
- flocculation value (FV), 2:380–381, 2:381*f*, 2:383*f*
- peptizing agent effect, 2:382, 2:384*t*
- pH effect, 2:381
- soil clay systems, 2:382, 2:384*f*
- Soil, 2:483–484, 3:236, 4:82
- accretion, 4:71–72
- aeration, 3:498, 5:216
- age, 4:51, 4:293
- as agricultural resource, 3:24–26
- air content and composition, 5:217–218
- air filters and regulators, 3:7
- amendment, 3:414–415, 3:422
- anaerobiosis, 1:379
- analysis of variance, 4:494–496, 4:494*t*
- applied herbicides, 3:445
- as archives of natural and cultural history, 3:3
- art, 3:506
- bacterial phyla, abundance with manure management, 1:563*t*
- biodiversity, 1:472, 3:190
- biogeochemical cycling, 4:84–85
- biological properties, 2:572
- bulk density, 3:188
- carbon, 2:429, 2:634, 2:638–639, 2:639*f*
- carbon sequestration in, 3:104
- characteristics of variation, 4:489–491
- chemical properties, 2:571–572
- chemometrics, 2:569–571
- chlorpyrifos in, 2:504, 2:505*f*
- as chronological control, 3:23–24
- chronosequence, 4:51–52, 4:52*f*, 4:71–72, 4:325–326, 4:748
- classification, 4:181–182, 4:198–199, 4:200–203*t*, 4:202*f*, 4:474–476
- color, 4:153–155, 4:569, 4:570*f*
- compaction and strength, 4:585
- compaction-soil management and cultivation, 3:498
- conditions and biodiversity, manure management on, 1:561–565
- construction, 3:39
- contaminants, 2:572–573
- contamination, 2:583*f*, 2:584
- covariance, correlation and regression, 4:496–498
- definition, 4:197
- depth functions, 4:573–574, 4:573–574*f*
- descriptive statistics, 4:489
- ecology, 3:11
- in ecosystems and landscapes, 3:9–10
- ecosystem services, 3:9–10, 4:85–86
- elemental field distribution maps, 2:638–639
- environmental components, 3:10
- environmental filters and buffers, 3:9
- enzymes, 2:400–401
- evolution, 4:70*f*, 4:72, 4:73*f*
- field properties, 4:473
- filter and buffer, 3:7–8
- food and fodder production, 3:5
- as four-dimensional body, 3:2–10
- functions, 3:548–549, 4:85–86, 4:481–483, 4:483*f*
- genesis and paleopedology, 4:178–181
- as genetic resource, 3:7
- global change pressures on, 4:83*f*
- as habitats for organisms, 3:5
- hydrology, 4:756–757
- improvement, 3:78–79, 3:417
- landscape evolution, 3:24
- load capacity, 3:8
- major types, 4:326–328, 4:329*t*
- map units, 4:474–476
- measurements and properties, 4:473–474
- multifunctionality, 3:547–548
- as natural aesthetic objects, 3:4
- as natural bodies, 3:3
- organic contaminants in, 2:503–506
- organophosphorus pesticides in, 2:503–504, 2:505*f*
- paleoenvironmental reconstructions, 3:24
- pedotransfer function, 4:474, 4:475*f*
- phosphorus, 2:503
- physical properties, 1:5, 2:571, 3:378
- pollution, 3:224
- population, units and samples, 4:487
- potentials by, 3:4–5
- probe, 5:437, 5:440
- processes, 3:103–104
- productivity, 3:2
- properties, 3:105–107, 3:379–380, 4:2, 4:84–85
- quality and organic matter content, 3:497
- raw materials, 3:6, 3:8
- as recreation areas, 3:8
- replication and randomization, 4:487–488, 4:489*f*
- residues, 3:8
- rules for classifying, 4:199–203
- salination on, 3:239–240
- salinity in, 3:237, 3:237*f*
- samples, 4:487
- soil spectroscopy, 4:473–474, 4:474*f*
- statistical significance, 4:491–492
- structure development, SUITMAs, 4:242
- support energy production, 3:6
- taxonomy, WRB and, 4:212*f*, 4:215
- thiabendazole in, 2:504–506
- thiram in, 2:503
- transformations, 3:7, 4:492–493, 4:493–494*f*
- for transport, 3:8
- variation, 4:487
- water cycle, 3:7–8
- water regime control, 3:9
- Soil acidification, 2:148, 3:369*ge*
- Soil aggregation, 2:466–467, 2:467*f*, 3:214, 3:220–221, 3:220–221*f*
- Soil algae, 1:63*f*
- biological soil crusts, 1:61
- biotic interactions, 1:68
- carbon (C) cycle, 1:62
- cyanobacteria, 1:61–62
- diatoms, 1:61–62
- distribution in soil profile, 1:65
- eukaryotic, 1:64–65
- genetic approaches, 1:65
- habitat effects, 1:65–66
- lifestyles, 1:65–66
- microscopic technique, 1:65
- nutrient cycles, 1:67–68
- photo-litho-autotrophic metabolism, 1:62
- prokaryotic algae, 1:62–64
- as soil conditioners, 1:67
- soil moisture, 1:66
- soil pH and nutrient availability, 1:66
- solar radiation, 1:66
- world-wide distribution, 1:64*t*
- Soil and Landscape Grid of Australia (SLGA), 4:703–705
- Soil art. *See* Artistic works in soil
- Soil artificialization, 3:364
- Soil backscattering, 4:621–622
- Soil-based climate change mitigation, 4:88–89
- Soil-based ecosystem services, 3:548–552, 3:549*f*, 3:551*f*
- Soil biofilms, 1:120
- active microbial biomass, 1:122*f*
- biogeographical assessments, 1:124
- ecological functions, 1:123–124
- extracellular polymeric substances (EPS), 1:120–123
- human disease diagnosis, 1:122
- microaerobic microenvironment, 1:124
- microbial biomass carbon, 1:122
- microfluidic chips, 1:124–125
- microscopic observations, 1:120, 1:121*f*
- non-invasive planar optode imaging, 1:124
- Soil-borne disease suppression, 1:456, 1:466
- Soil-borne pathogens and antibiotic resistant genes, 1:564–565
- Soil-borne plant disease control, 1:439
- biological control, 1:442–445
- chemical control, 1:440
- crop rotation, 1:441
- resistant varieties, 1:441–442
- sanitation, 1:441
- solarization, 1:440–441
- Soil carbon sequestration, 1:224*t*
- Soil changes, 3:344, 4:56–58, 4:57*t*
- element imbalance, 3:356–358
- erosion, 3:345–348
- human activities, 3:344
- organic matter loss, 3:349
- Soil chemistry
- carbon dioxide emissions, 2:692–693
- climate change, 2:695–698
- electron microscope-based new technologies, 2:584–587
- greenhouse gas emissions, 2:692–695
- methane emissions, 2:693–694
- nitrous oxide emissions, 2:694–695
- Soil classification, 2:501, 2:501*f*, 4:219
- agronomic classification, 4:187
- American approach, 4:187–188
- chemical classification, 4:187
- Dokuchaev's paradigm, 4:187–189
- history and principles, 4:185

- Soil classification (*Continued*)
 and indigenous soil knowledge, 4:186–188
 and parent material, 4:40
 petrographic classification, 4:187
 post-Dokuchaev early period, 4:189–190
 post-Dokuchaev late period, 4:191–195
- Soil composite mapping processor (ScMAP), 4:638
- Soil Condition Analysis System (SCANS), 4:587, 4:588f
- Soil Conservation Service Curve Number method (SCS-CN), 5:291
- Soil continuum model (SCM), 1:394
- Soil degradation, 1:561, 3:186, 4:328
 biological, 3:99–100, 3:169–170
 chemical, 3:99, 3:168–169
 definition, 3:165
 drivers, threats and consequences of, 3:167f
 factors controlling, 3:166–167, 3:166f
 physical, 3:98–99, 3:167–168
 soil remediation, 3:170
 types, 3:167, 3:168f
- Soil development, 4:740
 climate, 4:49
 models, 4:740f, 4:741
 time effect, 4:50f, 4:51
- Soil ecosystems, 3:239
 biology, 4:26–27
 carbon to nitrogen ratios, 4:26t, 4:27f
 climatic zones, 4:26–27
 microbial biomass, 4:27–28
 organo-mineral interactions, 1:403–404
 savanna and grassland soil systems, 4:29
 tropical soil systems, 4:28–29
- Soil emission, 4:623
- Soil-environmental geospatial data
 acquisition, 4:697–705
- Soil erosion, 3:195f, 3:198–199, 3:210, 3:345–348, 3:369ge, 3:378, 4:86–87, 4:377, 5:596
- Soil failure
 Coulomb failure line, 5:79–80
 criteria, 5:76
 Mohr–Coulomb failure criterion, 5:77, 5:79–80f
 shear, 5:78–80, 5:79f
 soil shear, 5:82
- Soil fauna, 1:562–563
- Soil fertility, 3:182, 4:338
 carbon supply, 3:430
 organic matter supply, 3:425
- Soil flux
 erosion and, 4:743
 treatment, 3:339, 3:339f
- Soil food webs, 1:14–15
- Soil formation, 4:2, 4:2f, 4:70, 4:71f, 4:77, 4:194, 4:235–236, 4:319–326, 4:460–464
 alpine and plateau soils, 4:275
 Antarctic Cold Desert, 4:274
 Antarctic Polar Desert, 4:273
 biota, 4:10, 4:10f
 boreal and subarctic regions, 4:266
 boreal forest systems, 4:30–31
 clay translocation, 4:324, 4:324–325f
 climate, 4:8, 4:9f, 4:319–321, 4:320t
 climate change and, 4:67
 desert soil systems, 4:29
 development of vertic features, 4:326
 ecosystems, 4:26–29
 environmental formula, 4:37
 expression, 4:7–8
 and functions, 1:480
 high Arctic soils, 4:270
 leaching and redistribution of calcite and silica, 4:323–324, 4:323f
 lithology and relief, 4:322
 low Arctic soils, 4:267
 mass additions and losses, 4:323
 parent material, 4:8, 4:9f (*see also* Parent material)
 relief/topography, 4:10–11, 4:11f
 rubefaction, 4:324–326, 4:325–326f
 soil diversity, 4:7
 soil nature, 4:37
 soil organic matter, 4:26
 at spatial and temporal scales, 4:8
 subantarctic forest and tundra, 4:272
 subtropical and temperate systems, 4:30
 surface plant residue, 4:26
 time, 4:11, 4:12f, 4:322–323
 tundra systems, 4:32
 vegetation, 4:321–322
 weathering, 4:37–39
- Soil fragmentation, in tillage systems, 3:437
- Soil freezing characteristic curve (SFCC), 5:413
- Soil gamma emission, 4:584t
- Soil gas chemical composition, 5:222
- Soil-gas diffusion coefficient, 5:249, 5:251–252, 5:252f
- Soil gas-phase dynamics, 5:219–220
 gas production processes, 5:220
 heterogeneity and variability, 5:219–220
 respiration and oxygen uptake, 5:219
- Soil genesis, thresholds in, 4:76–77
- Soil governance, 3:507
 challenges, 3:534–535
 definition, 3:534
 international level, 3:535–539
 international trade, 3:544
 land and soil degradation status, 3:534
 national and regional level, 3:539–541
 soil threats and drivers, 3:535
- SoilGrids, 4:703–705
- Soil health, 1:2, 3:182–183, 3:427–428, 5:14
- Soil health protection acts and programs, in United States, 3:460t
- Soil heat flux, 5:641–642
- Soil horizons and horizon boundaries, 4:570–572, 4:571f
- Soil hydraulic properties, temperature effects
 diurnal fluxes in soil water, 5:380
 long-term implications, 5:383
 soil water flow modeling, 5:382–383
 soil water retention, 5:380–381
 unsaturated hydraulic conductivity, 5:381–382
- Soil indicators, 3:187t
 across spatial and temporal scales, 3:185–186, 3:185f
 different approaches on, 3:183–184, 3:183f
 selection of, 3:184–185
- Soil influence, on groundwater, 5:364
- Soil infrared spectra, 2:569
- Soil inversion/fragmentation, tillage systems, 3:434–436
- Soil irradiation, 2:637
- Soil Land Inference Model (SoLIM), 4:688
- Soil-landscape evolution, 4:483–485, 4:484f
- Soil-landscape modeling
 identifying and minimizing uncertainty, 4:748–749
 landscape, 4:742–743
 pedon, 4:741, 4:741–742f
 soilscape, 4:743–744, 4:744–745f
 space–time boundary conditions for, 4:745–747
- Soil loss
 rangeland, 3:98–99
 ratio, 3:205, 3:206t
- Soil management
 carbon sequestration, 3:428–429
 in early civilizations, 3:30–31
 of public green spaces, 3:489
 soil aeration and, 3:498
- Soil map errors, multivariate distribution of, 4:676–677
- Soil mapping, 4:215–216, 4:575, 4:576f, 4:580
- Soil matric potential, 5:471–472, 5:474
- Soil mechanics
 behavior in field, 5:81–82
 constitutive relationships for soil materials, 5:75–77
 definition, 5:71
 elastoplastic soil behavior, 5:77–80
 fracture, 5:80–81
 history, 5:71
 and land management, 5:83–84
 linear-elastic soil behavior, 5:77
 principle strains, 5:75f
 rotational shear vane, 5:83f
 shear, 5:78–80, 5:79f
 soil shear, 5:82
 soil sinkage, 5:82
 strain, 5:73–75
 stress, 5:71–73
- Soil mediated fluxes, 5:484–485
- Soil memory, 4:234–236
- Soil mesofauna composition, 3:191
- Soil metaproteomics, 2:404
- Soil microbes, 1:472f, 4:84
 to ecosystem services, 1:474t
 in farming systems, 1:464
vs. plants, 1:420
 and system complexity, 1:470–471, 1:471f
- Soil micromorphology, 4:172
 applications, 4:178–183
 history, 4:172–173
 observation techniques, 4:173–174
 sampling, 4:173–174
 techniques, 4:173–174
 thin section preparation, 4:173–174
- Soil mineral components
 aggregation and mobility, 2:565
 clay surfaces, biochemical transformation, 2:565
 fluorescent probes, 2:561
 ligand-promoted mineral dissolution, 2:566

- molecular distribution and conformation, 2:565
- organic chemicals adsorption, 2:563
- polarity and charge adsorption, 2:564–565
- structural quenching, 2:562–563
- surface acidity, 2:563–564
- Soil mineralogy, 1:7
- Soil mineral qualification, 2:646
- Soil moisture (SM), 4:289, 4:752*ge*
- cosmic-ray neutron probe, 4:595
- definition, 4:592–593
- electromagnetic induction, 4:596–597
- electromagnetic sensors, 4:593–594
- ground penetrating radar, 4:595–596, 4:597*f*
- importance of, 4:591–592
- microwave radiometry, 4:598
- model components, 4:756–757
- neutron probes (NP), 4:594, 4:595*f*
- neutron scattering, 4:594
- nuclear magnetic resonance imaging, 4:597–598, 4:598*f*
- proximal sensing, 4:593
- sensors, 4:587, 5:174
- soil algae, 1:66
- spatiotemporal variations in, 4:598
- surface temperature measurements, 4:608
- Soil Moisture Active Passive (SMAP) system, 4:607, 5:486
- Soil Moisture and Ocean Salinity (SMOS) system, 5:486
- Soil-moisture regimes (SMRs), 4:219–222, 4:356
- Soil moisture remote sensing. *See* Remote sensing of soil moisture
- Soil moisture retrieval, 4:621–627, 4:626*f*
- Soil monitoring (SM)
- developments in, 3:178–179
- fundamental questions, 3:174
- importance of, 3:173–174
- networks, 3:175–176
- organizational and practical questions, 3:174–175
- principles, 3:173
- protocols, 3:176
- quality control, 3:178
- remote sensing methods, 3:172
- sites, 3:176–177
- soil sampling, 3:177–178
- Soil morphology, 4:568–569
- chemical response, 4:161–162
- color, 4:153–155
- compound structure, 4:161
- consistence, 4:161
- definition, 4:150
- digital soil morphometrics, 4:166
- grade, 4:161
- horizons, 4:150–151, 4:164–166, 4:165*f*
- micromorphological descriptions, 4:166, 4:169*f*
- mottles, 4:157
- pedon features, 4:163, 4:163*f*, 4:164*t*
- ped surface features, 4:163, 4:164*t*
- redoximorphic features, 4:155–157
- roots and pores, 4:161, 4:162*f*
- shape, 4:159, 4:160*f*
- site characterization, 4:150
- size, 4:159–160
- soil structure, 4:157–159, 4:160*t*
- texture, 4:153
- Soil nitrogen transformations, 2:144–148, 2:144*f*
- emissions, 2:148–149*f*, 2:149
- nitrogen compounds in soils, 2:144
- quantification, 2:148–149*f*, 2:149
- research priorities, 2:150
- Soil nutrients, 1:278
- Soil organic carbon (SOC), 1:470, 1:471*f*, 2:124, 4:291, 4:753*ge*
- airborne platforms, 4:634
- biological soil degradation, rangelands, 3:99–100
- carbon cycle, 1:223
- composition, 1:223
- imaging spectroscopy, 4:632–634
- large spatial variation, 4:632
- on populations of bacteria and fungi, 1:458*f*
- predictions of, 4:586*f*
- remote sensing (RS), 4:637–638
- hyperspectral satellites, 4:637–638
- multispectral satellites, 4:637–638
- Sentinel 2 MSI, 4:638
- soil composite mapping processor (ScMAP), 4:638
- unmanned aerial systems (UAS) (*see* Unmanned aerial systems (UAS))
- spectral models, 4:632
- stocks, 2:388*f*
- and time effect, 4:49
- Soil organic matter (SOM), 1:448, 2:68–69, 2:356, 2:422, 2:483–484, 2:491–492, 2:501–503, 2:581, 3:189–190, 3:374, 4:373, 4:753*ge*. *See also* Organic nitrogen
- abiotic drivers, 2:391–392
- aggregate formation, 2:69
- arsenic, 2:204
- biological activity and cycling of, 3:190
- biomarkers, 2:392–393
- carbon to nitrogen (C:N) ratios, 4:26
- cold region soils, 4:260
- components, 4:26
- composition, 2:58
- decomposition, 2:388–390, 4:26
- definitions, 2:60*t*
- dissolved organic matter (DOM), 2:69
- electron paramagnetic resonance (EPR) spectroscopy, 2:536–538
- fluorescence polarization, 2:559
- fluorescence quenching, 2:557–561
- fluorescence spectra, 2:553
- fluorophores, 2:553–554
- formation, 2:390–392
- genesis and pathways, 2:390–391
- global water cycle, 5:486
- and humus, 2:58–60
- isotopic approaches, 2:393
- microbial accessibility, 2:393
- mineral-associated organic matter (MAOM), 2:69
- mineral interaction, 2:583
- molecular and environmental factors, 2:554–555, 2:555*f*
- organic and inorganic carbon, 2:58
- organismic carbon, 2:60–64
- particulate organic matter (POM), 2:69
- physical fractionation methods, 2:68
- physical protection, 1:390
- recalcitrant nature, 2:393
- soil fertility, 3:352–356
- and soil organic carbon, 1:562
- soil water repellency (SWR), 5:203–204, 5:206
- thermally altered C, 2:65–66
- traits-based strategy, 2:394, 2:394*f*
- turnover, 3:425–426
- xenobiotics, 2:64–65
- Soil organisms, 2:479
- Soil organo-mineral interactions
- abiotic and biotic controls, 1:393–395
- mineral phase properties, 1:394–395
- organic matter dynamics, 1:393–394
- soil solution chemistry, 1:395
- Soil particles, 2:580–581, 3:212–213, 3:213*f*
- Soil penetration resistance, 3:186
- Soil pH
- biochars, 2:407*f*
- methane production, 2:694
- nitrous oxide emissions, 2:695
- and redox potential, arsenic, 2:204
- soil algae, 1:66
- soil morphology, 4:161
- soil organic matter (SOM) formation, 2:391
- symbiotic nitrogen fixation, 1:277–278
- wetlands, 4:445
- Soil physics theory, 5:308
- Soil-plant-atmosphere continuum (SPAC), 5:553*f*, 5:609
- CERES Maize model, 5:628
- complex models, 5:629–630
- computer models, 5:628
- energy budget, 5:626
- evapotranspiration (ET), 5:630, 5:631*f*
- intermediate models, 5:628–629
- principles, 5:625–626
- remote sensing, 5:631–632
- Simple Inverse Yield Model, 5:628
- solar radiation, 5:628
- terrestrial ecosystem, 5:625
- transport principle, 5:626
- vegetation-canopy characteristics, 5:631
- water, 5:625–628
- Soil Plant Atmosphere System Model (SPASMO), 5:328, 5:332
- Soil-plant system
- germanium in, 2:272
- silicon in, 2:265–271
- Soil process, 4:483, 5:23
- Soil production, 4:70
- Soil profile attributes, 4:569–572
- coarse fragments, 4:572, 4:572*f*
- other soil properties, 4:572
- soil color, 4:569, 4:570*f*
- soil horizons and horizon boundaries, 4:570–572, 4:571*f*
- soil structure, 4:569–570
- Soil profile mapping, 4:574–575, 4:576*f*

- Soil profile variation, 4:573–575
 soil depth functions, 4:573–574, 4:573–574f
 soil mapping, 4:575, 4:576f
 soil profile mapping, 4:574–575, 4:576f
- Soil properties, 3:240, 4:374–375, 4:572
- Soil Protection Act (SPA), 3:542–543
- Soil quality, 1:2, 3:181, 4:230
 and health indicators, 3:186–191
 to soil health, 3:182–183
- Soil quality protection policies and laws, 3:457
 action plans, 3:463–465
 agricultural waste management, 3:462–463
 European Unions, 3:459t
 financial incentives needed, 3:467–468
 in Greece, 3:460t
 in India, 3:470–471
 in-situ crop residue management (CRM), 3:463f
 international treaties, 3:459
 planned activities, International Decade of Soils, 3:458f
 policy implementation, 3:466
 private organizations for investment, 3:468–469
 sustainable agricultural practices, 3:457–458
 sustainable soil management (SSM), 3:458–459
 in United States, 3:461t
- Soil reaction, 1:7, 2:8. *See also* pH
- Soil respiration
 aerobic *vs.* anaerobic respiration, 1:371
 in agronomy and ecology, 1:375
 and climate change, 1:377
 CO₂ quotient (qCO₂), 1:372–373
 cover crops, 1:375
 drainage, 1:374–375
 drivers and controls, 1:371–374
 ex-situ measurements, 1:376–377
 and global GHG balance, 1:371
 hydrology and temperature interplay, 1:374
 as indicator of biological activity, 1:370–371, 1:376
 isotopes, 1:376
 management practices, 1:374–375
 measurement, 1:376–377
vs. moisture, 1:370, 1:370f, 1:372
 nutrient availability, 1:373
 nutrient management, 1:375
 pH, 1:373
 on site measurement, 1:376
 soil carbon pools, 1:373
 soil organic matter supply and quality, 1:374
 soil quality and health indices, 1:375
vs. temperature, 1:370–371, 1:370f, 1:372f
 tillage, 1:375
- Soil restorative farming system, 3:13–14, 3:14f, 3:15t
- Soil salinity, 1:2, 3:188, 3:240–241
 b(h,R,C₀) and N(h,R,C₀) function, 5:128–129
 capillary pressure head *vs.* water content, 5:125, 5:126f
 case study, 5:135
 chloride concentration mean values and SD, 5:135, 5:136f
 flow and transport (*see* Flow and transport)
 hydraulic conductivity mean values and SD, 5:136, 5:136f
 low-quality water (LQW) irrigation, 5:124–125
 modelling irrigation with, 5:128
 root form and function, 1:327
 sodium adsorption ratio (SAR) *vs.* water content, 5:127, 5:127f
 soil solution-soil matrix (SS-SM) interaction, 5:125, 5:131–132, 5:137
 solute mean breakthrough curve (BTC), 5:137
 $\theta(h,R,C_0)$ and $K(h,R,C_0)$ function, 5:129–131
 tracer solute velocity, 5:136–137, 5:137f
 water content *vs.* hydraulic conductivity, 5:126–127, 5:127f
 water retention and swelling, 5:125–126, 5:126f
- Soil Sampling Programme (SSP), 3:176
- Soilscape, 4:739, 4:743–744, 4:744–745f, 4:748
- Soil science, 2:626–627, 3:20
 archaeological research, 3:22f
 children and teenagers, 3:578
 delivery modes, 3:574
 diversity, fractals and multifractals in, 4:735
 educator interaction, 3:576
 environmental awareness, 3:577
 examples for applications in, 2:490–495, 2:491f
 formal classroom settings, 3:572–573
 inter- and trans-disciplinarity practice, 3:578
 knowledge teaching, 3:576
 professional groups, 3:574–575
 soil preservation, 3:576
 statistics, 4:487
 teachers, support and training, 3:578
- Soil Science Society of America (SSSA), 3:575
- Soil seal, 3:39, 3:44, 3:47f, 3:51, 5:58–59
 infiltration and runoff, 5:62–63
 modeling, 5:60–62
 structure, 5:60
 water flow in, 5:62
- Soil seed bank, fire effects, 2:454
- Soil sink for methane, 3:155–158
- Soils of Urban, Industrial, Traffic, Mining and Military Areas (SUITMAs), 3:583, 4:233–237. *See also* Anthropized area soil
 accumulation of organic matter, 4:241
 carbon storage, 3:49, 3:49–50f
 climate, 4:239–240
 deciphering memory of past human activities in, 4:235–237
 development of soil structure, 4:242
 direction of pedogenesis, 4:243–244
 ecosystem services, 3:49, 3:50–51f, 3:51
 evolution of, 4:241–244
 footprints of past human activities in, 4:233–237
 functions, 3:44–49
 human-induced modifications on factors of evolution, 4:237–240
 main soil-forming processes observed in, 4:241–242
 mineral transformations, 4:241–242
 organisms, 4:240
 parent materials, 4:239
 pedogenic processes, 3:40–41, 3:41f
 prediction of evolution, 4:243–244
 reservoir of biodiversity, 3:48, 3:48f
 restoration of soil functions, 3:51–53, 3:52–53f
 soil classification, 3:44, 3:46–47f
 soils as archives of human–nature relations, 4:234
 soil sealing, 3:44, 3:47f
 soils of industrial brownfields as archives of past industrial activities, 4:236–237
 spatial variation and soil diversity, 4:244
 specific pedogenetic trends in, 4:242–243
 support for roots, 3:44, 3:47–48f
 temporal discontinuities, 4:244
 topography and hydrology, 4:239
 transfers of soluble compounds and particles, 4:242
 transformation of organic matter, 4:241
 urban soil stratification, 4:236
- Soil solution-soil matrix (SS-SM) interaction, 5:125, 5:131–132, 5:137. *See also* Soil salinity
- Soil sources
 methane, 3:154–155, 3:154t, 3:155–157f
 nitrous oxide, 3:158–159, 3:159t, 3:159f
- Soil spectroscopy, 4:473–474, 4:474f
- Soil structure, 5:1, 4:569–570
 bulk density, 5:2–3
 classification, 5:5f
 comparative scales, 5:2f
 dynamics, 5:3–4
 interactions, 5:3f
 modeling, 5:6
 non-invasive methods, 5:5
 plant available water and, 5:513
 pores, 5:1
 soil fragments, 5:5
 and soil functions, 5:2–3
 stability, 4:372
 three-dimensional visualization and quantification, 5:463f
 2-D thin sectioning approaches, 5:6
 visual assessments, 5:5
- Soil structure and soil organic matter (SOM) degradation
 amenity uses, 3:255
 biodiversity support, 3:254
 buffering and contaminant filtering, 3:255
 causes, 3:247–253, 3:248f
 climate change effects, 3:252–253
 climate change mitigation, 3:254–255
 compaction, 3:251–252, 3:252f
 healthy soil, 3:247
 land use, 3:247–249, 3:249f

- net primary productivity, 3:253, 3:253f
- soil forming factors, 3:247
- tillage/cultivation, 3:249–251, 3:250f
- water and air regulation, 3:253f, 3:254
- Soil surface
 - erodibility, 3:213–215, 3:214t
 - roughness, 3:213–214
 - sealing, 5:57
- Soil survey
 - non-probability sampling, 4:546–549
 - probability sampling (*see* Probability sampling method)
- Soil taxonomy (ST), 4:198–199, 4:737
- Soil taxonomy classification
 - andisols, 4:226–227
 - aridisols, 4:227–229
 - categories, 4:219, 4:220t
 - entisols, 4:230
 - gelisols, 4:226
 - generalized soil map, 4:221f
 - global distribution, 4:219–230
 - histosols, 4:226
 - inceptisols, 4:229
 - mollisols, 4:229
 - oxisols, 4:227
 - soil grouping, 4:222t
 - soil orders, 4:222f
 - spodosols, 4:227
 - ultisols, 4:229
 - vertisols, 4:227
- Soil temperature, 1:9, 4:289
- Soil temperature regimes (STR), 4:219–222, 4:356
- Soil testing
 - agricultural soils, 2:43–44
 - calibration, 2:47–48
 - for crop production, 2:47–48
 - extractants and methodologies, 2:45–47
 - nitrogen, 2:46–47
 - phosphorus, 2:45–46, 2:46f
 - quality, 2:48
 - sampling, 2:44–45
- Soil texture, 1:5, 3:221, 4:48ge, 4:153
 - global water cycle, 5:486
 - plant available water and, 5:513
- Soil thermal conductivity, 5:389f
- Soil tilth
 - aggregates, 5:50–51
 - history, 5:49–50
 - indices, 5:54, 5:55t
 - in situ evaluation, 5:53–54
 - mechanical resistance, 5:54
 - seasonal dynamics, 5:51, 5:52f
 - soil management effects, 5:53
 - soil organic matter, 5:52
 - soil properties, 5:48–49, 5:49f
 - tillage effects, 5:53
 - tilled and zero-tillage arable soils, 5:50–51, 5:51f
- Soil utility and quality, 4:481–483
 - land evaluation, 4:481, 4:482f
 - soil functions, 4:481–483, 4:483f
- Soil vapor extraction (SVE), 3:341, 3:342f
- Soil washing treatment, 3:338–339, 3:338f
- Soil water balance, 4:321t, 5:590
- Soil water characteristic (SWC) curve
 - bundle of cylindrical capillaries (BCC) model, 5:188–189
 - dewpoint hygrometer method, 5:200
 - dry-end SWC relations, 5:195
 - evaporation method, 5:199–200
 - fractal representations, soil porous system, 5:193
 - global datasets, 5:191–192
 - hysteresis, 5:197
 - matric potential, 5:187–188
 - measurement, 5:198
 - paired sensors, 5:200
 - parameters, 5:190–191
 - parametric model, 5:190
 - physically based models, 5:193–194
 - pore network models, 5:197
 - pore scale models, 5:195–197
 - pressure plate apparatus, 5:198–199, 5:199f
 - representing, 5:190
 - water adsorption, 5:195
- Soil water content, 5:437. *See also* Soil moisture (SM)
 - biogeochemical process, 5:472
 - crop productivity and carbon fluxes, 5:478
 - different scales, 5:478–480
 - ecosystem services, 5:473–474
 - energy related process, 5:470–471
 - global water cycle, 5:470
 - hydrologic process, 5:471–472
 - measurements and modelling, 5:476–480
 - measurement techniques, 5:475–476
 - monitoring, 5:474–476
 - using TDR, 5:437–441
 - watershed management, 5:472–473
- Soil water diffusivity, 5:300
- Soil Water Index (SWI), 4:624–625
- Soil waterlogging regime, 4:400, 4:402f
- Soil-water partitioning sorption, 5:284
- Soil water repellency (SWR), 5:203
 - capillary rise method (CRM), 5:207–208
 - drop infiltration, 5:207
 - fire effects, 2:452–453, 2:452f
 - high SOM content, 5:206
 - microbial biofilms and fungal activity, 5:206
 - mitigation, 5:214
 - modelling, 5:213–214
 - quantification, 5:207–209
 - repellency index (RI), 5:208
 - sandy soils, 5:206
 - sessile drop method (SDM), 5:208
 - soil process, 5:209–212
 - surface free energy components, 5:209
 - surfactants, 5:214
 - wetting agents, 5:214
 - wildfires, 5:206
 - Wilhelmy plate method (WPM), 5:208–209
- Soil-water retention, 5:380–381
 - capillary water pressure, 5:380
 - matric potential, 5:380
 - soil water capacity, 5:381
 - suction forces, 5:380
 - temperature effects, 5:380–381
- Soil water retention curve (SWRC), 5:413
- Soil water sensing, 5:427–428
- Solarization, 1:440–441
- Solar radiation, 1:66, 4:14, 5:395, 5:396f
- Solid phase, 1:408–409, 1:409f
- Solid phase sensors (SPS), 5:174–175, 5:177–178
- Solid set sprinkler systems, 5:614, 5:614f
- Solid-state NMR, 2:610, 2:621
- Solifluction, 4:258
- Solonchaks, 1:506
- Solubility of gases, 5:154, 5:154t
- Solubilization, 2:157–158
- Solute diffusion coefficient in soil, 5:250–251
- Solute transport, 5:306–307
 - adsorption and exclusion, 5:657–659
 - advection, 5:653
 - advection-dispersion equation (ADE), 5:659–664
 - applications, 5:652–653
 - breakthrough curves, 5:655–657
 - chemical transport, 5:653
 - continuous time random walk (CTRW) approach, 5:662–663
 - diffusion, 5:653
 - dimensionality, 5:654–655
 - dispersion, 5:653–654
 - fractional advection-dispersion models, 5:663–664
 - longitudinal dispersivity, 5:655
 - lysimeters, 5:675
 - mechanisms, 5:653
 - multi-rate mass transfer (MRMT) approach, 5:661–662
 - pore-size distribution, 5:23
- Solution-diffusion model, 3:385–386
- Solution-state NMR, 2:609–610
- Soret effect, 5:267
- Sorption, 2:336, 2:337f, 2:343–346
 - behavior in soils, 2:347–348
 - of bromine, 2:239
 - chemisorption, 2:316
 - of fluorine, 2:237
 - of iodine, 2:240
 - by metals and oxyanions, 2:347
 - molecular exchange, 2:316
 - organic chemicals (*see* Organic chemicals sorption)
 - organic contaminants biodegradation, 1:553
 - organofluorine compounds in soil, 2:237
 - pesticides, 2:472, 2:476f
 - pH and ion selectivity effects on, 2:339, 2:340f
 - physical, 2:316
 - properties of metals, 2:337, 2:338t
 - surface characteristics, 2:337
 - terminology and role of, 2:336–337
- Sorption-based methods, 2:24–30
- Sorption-desorption hysteresis, 2:356–357, 2:357f
 - anthropogenic organic compounds on, 2:364
 - effects of environmental factors on, 2:364–366
 - effects of properties of SOM on, 2:359–363
 - humic substances, 2:359–361

- Sorption–desorption hysteresis (*Continued*)
 indices, 2:357–358
 physiochemical properties, 2:364
 possible mechanisms, 2:358–359, 2:358–359f
- Sorption–desorption reactions, 2:346f, 2:350–353
- Sorption retarded intraorganic matter diffusion (SROMD), 3:302
- Sorption retarded intraparticle diffusion (SRPD), 3:302
- Sounding soil
 acoustic complexity index (ACI), 3:514
 acoustic ecology, 3:513
 biological activity, 3:513
 soundscapes, 3:513–514, 3:513f
- Southern Circumpolar Region (SCR) soil
 Antarctic Polar Desert, 4:273
 soil parent materials, 4:270
 subantarctic forest and tundra, 4:270–273
 zones, 4:270
- Space and time prediction, 4:477–481
 clorpt approach, 4:477
 geostatistics, 4:478, 4:478f
 hybrid of scorpan and geostatistics, 4:478–479, 4:480f
 machine learning, 4:479
 sampling design, 4:481
 scorpan approach, 4:477–478
 space-time prediction, 4:479–480, 4:481f
- Space-borne systems, 4:620, 4:620f
- Space–time boundary conditions, for
 soilscape models, 4:745–747
 heterogeneity, 4:747
 initial condition, 4:745–746
 time-variant lower boundary conditions, 4:747
 time-variant upper boundary condition, 4:746–747, 4:746–747t
- Space–time prediction, 4:479–480, 4:481f
- Spatial dependence, 4:523
- Spatially Referenced Regression on Watershed Attributes (SPARROW), 5:333
- Spatial prediction, 4:668–669
- Spatial scale
 aim and research significance, 4:522
 change of support, 4:522–523, 4:525–526
 definition, 4:521–522
 geostatistics, 4:523–524, 4:527
 heterogeneous spatial data fusion, 4:527
 models, 4:523
 multispectral data analysis (case study), 4:529–530
 scale, 4:522
 scale triplet concept, 4:521–522
 scaling, 4:522
 spatial information technology (SIT), 4:521–522
 uncertainty in models, 4:533
 variogram, 4:524, 4:525f
- Spatial soil information, uncertainty assessment, 4:675
- Spatial support and resolution, uncertainty assessment, 4:679
- Spatial variation and soil diversity, SUITMAs, 4:244
- Spatio-temporal modelling, 4:689
- Spearman rank correlation, 4:496–497
- Speciation
 of alkaline earth metals, 2:127
 antimony (Sb), 2:219
 arsenic (As), 2:216
 barium (Ba), 2:220–221
 lead (Pb), 2:220
 selenium (Se), 2:218
- Specific storage, 5:366
- Specific surface area (SSA)
 biochar, 2:411–412
 of clay minerals, 2:24
 definition, 2:22
 determination methods, 2:24–30
 EGME retention, 2:29–30
 empirical models/pedotransfer functions, 2:30
 indirect methods, 2:30
 mass-volume estimation methods, 2:24
 of porous media, 2:23
 proximal sensing methods (spectroscopy), 2:30
 sorption-based methods for, 2:24–30
 water-sorption derived, 2:23f
 water vapor sorption, 2:27–29, 2:28–29f
- Specific yield, 5:366
- Spectral libraries, 4:587–588
- Spectroscopy in soil science, 2:567–568
- Sphagnum pulchrum*, 4:387f
- Spherical function, 4:660
- Spiders
 food, 1:82
 habitats, 1:82–83
 infraorders, 1:81
 interactions, 1:83–84, 1:84f
 silk webs, 1:79–81
 venomous bites, 1:79–81
- Spiral-shaped chemosensory organs, 1:106f
- Spodic horizon, 4:54ge
- Spodosol, 4:55ge, 4:227, 4:296
- Sports fields
 pitch's surface, 3:493–496
 soil considerations for, 3:493–497
 soil management considerations on, 3:495–496t
 surface quality and safety design of, 3:489
 sustainable maintenance planning for, 3:489
- Sprayable preparations, mulch, 3:398
- Sprinkler irrigation systems, 5:609
 lawn sprinkler systems, 5:618
 self-moving sprinkler systems, 5:615–617, 5:616f
 set-move sprinkler systems, 5:615, 5:615f
 solid-set systems, 5:614
 sprinkler devices, 5:613–614
- Sputtering, 2:671
- ssu* gene cluster, 1:306
- Stabilization ponds, 1:569
- Stabilized nano-fertilizers/pesticides, 2:689
- Stable isotope probing (SIP), 1:162
- Stable isotopes, 2:82
 light elements analysis, 2:107
 mass spectrometry, 2:107
 metal pollution source, 2:113
- non-traditional isotope geochemistry, 2:108
 pedogenetic process, 2:109–111
- Stagnosol–Gleysol hypothesis, 4:361
- Standard error, 4:490
- States of water
 crystalline structure of ice, 5:149–150, 5:149f
 flickering cluster model, 5:150, 5:150f
 water phase diagram, 5:150–151, 5:150f
- State-wise legislation, United States, 3:461t
- Statistical inference, 4:536
- Statistical significance, 4:491–492
- Steady state, 2:275ge
- Steady-state flow
 through saturated medium, 5:257–258, 5:258f
 through unsaturated medium, 5:260
- Stefan–Boltzmann law, 4:601
- STEP-AWBH model, 4:700–702
- Steppes. *See* Grassland soils of cool-arid–temperate ecozone
- Steric repulsion, 5:339–340
- Stocking rate
 rangeland management, 3:92–93
 in South African Highveld, 3:93f
- Stoichiometry, 2:275ge
- Stomata, foliar nutrition application, 3:386–387
- Stomatal conductance modeling and water use efficiency, 5:537
- Stop-and-go sensors, 4:587
- Storage, manure management, 1:559–560
- Storativity, 5:366
- Strain, concept of, 5:73–75
- Stratified simple random sampling, 4:539–540
- Stratigraphy, 4:418
- Straw mulch, 3:393f
- Streambed floodwater harvesting, 5:601–602
- Streaming current, 5:267
- Streamline, 5:355
- Streamtube model, 5:353ge, 5:359
- Strength, soil, 5:81–82
- Stress
 concept of, 5:71–73
 normal and sheer, 5:72f
 octahedral shear stress, 5:73
 principal stresses, 5:72f
 soil mechanics, 5:71–73
 temperature, 1:133
 tolerance to, 1:132–133
 total *vs.* effective, 5:73
 ultraviolet radiation, 1:132–133
- Stress strain, soil structure changes, 5:302–303
- Stress tolerance
 abiotic stress, 1:325
 acid soils, 1:327
 bacteria, 1:35
 compaction, 1:326–327
 drought, 1:325–326, 1:325f
 mycorrhizal symbioses, 1:327
 nutrient bioavailability, 1:326, 1:326f
 salinity, 1:327
 waterlogged/flooded soils, 1:327

- Strip cropping, 3:220
- Strontium (Sr), 2:132–133, 2:214–215, 2:261
- Structural equation modelling (SEM), 4:689
- Structured models, 1:176
- Structured soils, anisotropy
land sliding process, 5:303–304
shear stress components, 5:302–303
stress strain process, 5:301–302
tensorial stress strain, 5:302–303
- Structure formation and land use,
evaporation, 5:496, 5:496f
- Subantarctic forest and tundra
classification and carbon stocks, 4:273
climate, 4:272
landscapes and soils, 4:272f
soil formation, 4:272
vegetation, 4:272
- Sub-arctic and arctic zones, 4:22f
- Subcritical water repellent solid surface,
5:205f
- Subirrigation, 5:620
- Substrate-induced respiration (SIR), 1:148,
1:157–158, 1:377
- Substrates, 1:523–525
composition, green roofs, 3:484
microbial denitrification, 1:262–267
- Substrate symmetry on epitaxial growth,
2:592–594
- Subsurface drainage
controlled drainage, 3:261, 3:261f
conventional free drainage, 3:260, 3:260f
excess pore water, 3:259
- Subsurface drip irrigation (SDI), 5:619
- Subsurface-flow, 4:752ge
- Subsurface-water, 4:752ge
- Subtropical and temperate systems
coniferous forest systems, 4:30
deciduous forest systems, 4:30
- Suction cups, leaching loss, 5:327
- Sufficiency nutrient levels, 1:2
- SUITMAs. *See* Soils of Urban, Industrial,
Traffic, Mining and Military Areas
(SUITMAs)
- Sulfate
esters, 1:306–307, 1:306f
ions, 2:171
oxygen in, 2:90–91
reduction in anaerobic soils, 1:385
- Sulfur (S), 2:192, 2:214–215
arsenic, 2:205
assimilation, 1:305–307
biological availability, 1:304, 1:305f
biological transformations, 1:303, 1:304f
contents in soil, 2:194
disproportionation, 1:307–308
global cycle, 1:304
isotopes, 2:91
metabolism, 1:304–305
mineralization, 1:304–305
oxidation, 1:307
physicochemical properties, 2:193t
reduction, 1:308
in soil, 2:194–196, 2:198–199
sulfate-reducing bacteria, 1:303,
1:308–309
sulfur-oxidizing bacteria, 1:307
- transformation processes in soil,
2:194–195
- transformations and environmental
quality, 1:308–309
- Superficial saturation
excessive rainfall, 4:410–411
impeded vertical drainage, 4:406–410,
4:408f
paddy cultivation, 4:411
- Superhyperfine splitting interaction, EPR
signal, 2:522
- Superinfection exclusion, 1:16ge, 1:21
- Supervised classification
application, 4:504, 4:504t
classification and regression tree (CART)
model, 4:505
linear discriminant analysis (LDA),
4:503–505
random forests, 4:505
- Supervised learning, 4:510–516
artificial neural networks (ANNs), 4:514
bagging, 4:513
component-wise gradient boosting,
4:512–513
decision trees, 4:513
ensembles, 4:514
model assessment and uncertainty, 4:515
model interpretation, 4:515–516
model selection, 4:514–515
quantitative response variable, 4:511
regularization, 4:512
support vector machines (SVMs), 4:514
topsoil carbon measurements, 4:511
- Supplementary irrigation, 5:609
- Support vector machines (SVMs), 4:514,
4:649
- Suppressive soil, 1:445
- Surface charge, 2:368
balance, 2:369–372
biochar, 2:411
components, 2:369
density, 2:368ge, 2:369
Surface complexation, 2:339
- Surface drainage, 3:259
- Surface element integration (SEI) technique,
5:340
- Surface energy balance (SEB), 5:503–504
components, 4:611, 4:611f
diurnal variations, 4:614f
and evapotranspiration, 4:611, 4:611f
- Surface-enhanced Raman spectroscopy
(SERS), 2:499–500, 2:500f
in soil components analysis, 2:500–503
in soil contaminants analysis, 2:503–507
- Surface-flow, 4:752ge
- Surface fractal models, 5:193
- Surface free energy, 5:228
- Surface functional groups, 2:339
- Surface horizons, 4:59
- Surface irrigation, 5:609
basin and border, 5:612–613, 5:612f
characteristics, 5:610
efficiency, 5:610
furrow irrigation, 5:611–612, 5:611–612f
phases, 5:610
- Surface mulch, 5:53
- Surface plasmon polariton (SPP), 2:503–504
- Surface resistance, 5:572–573, 5:639–640,
5:639f
- Surface roughness and charge heterogeneity,
5:348–349
- Surface runoff, 5:596
- Surface soil texture, 3:213, 3:214t
- Surface tension, 5:156, 5:156f, 5:226–228
- Surface tension-viscous flow (STVF) theory,
5:380–382
- Surface-water, 4:752ge
- Surfactants, 3:389
- Suspension effect, 2:13, 2:13f
- Sustainability
environmental effects of irrigation, 3:274
and environmental safety solutions,
3:489
outcomes, 3:497
- Sustainable agriculture, 1:2–4
- Sustainable Development Goals (SDGs),
3:457, 3:459f
- Sustainable forest management, 3:64–65
- Sustainable soil management (SSM),
3:458–460
- SWAT (Soil and Water Assessment Tool)
model, 5:291–292
- Swelling behaviour. *See* Shrinkage
- Symbiotic nitrogen fixation
biological nitrogen fixation, 1:271–272
engineering strategies, 1:279–281, 1:280f
environmental factors, 1:277–279
nitrogenase activity regulation,
diazotrophs, 1:274–276
nitrogen-fixing plant-microbe associations,
1:272–274
nitrogen regulation, 1:274–275, 1:275f
nutrient exchange, 1:276–277
oxygen regulation, 1:275–276
post-transcriptional regulation, 1:276
soil fertility, 1:279
- Symphyla, 1:96f, 1:99
- Synchrotron-based techniques, 1:401–402
- Synchrotron radiation XRD (SR-XRD),
2:650–651
- Synthetic chemostat model (SCM),
1:178–180, 1:179f
- Systematic random sampling, 4:540–541
- System complexity, soil microbes and,
1:470–471, 1:471f
- System for continuous aerosol measurement
of particulate emissions from roads
(SCAMPER), 3:228
- Systemic herbicide, 3:440ge
- System-level carbon sequestration, 3:81–83
- ## T
- Tagged neutron method, 2:628
- Taiga, 4:388
- Taliks, 5:424
- Tapping mode (TM)-AFM imaging,
2:591–597, 2:593f
- Taxa-range size distributions, 4:734
- Taxonomic pedodiversity, 4:727–728
- Technogenic material, 3:39, 4:36–37

- Technosol, 3:44, 3:47f
for green roofs and vertical green systems, 3:478–482
microbial carbon indicators, 1:489f
physical requirements, 3:478–481
- Tectomanganates, 4:143
- Tellurium (Te), 2:215
- Temperate grasslands, 3:103–104
- Temperate plant-soil interface, 4:295–296
- Temperate zone
landscapes, 4:389f
organic soils, 4:388
- Temperate-zone loess soils, 4:45–46, 4:45f
- Temperature, 1:409–410
bacterial transport in soil systems, 1:518
bacteria survival and persistence in soil systems, 1:514–515
biocrusts, 1:133
microbial denitrification, 1:269
phase, composting, 1:526
soil organic matter, 2:364–365
stresses, 1:133
- Tennessee (Ts), 2:235
- Tensiometer, 5:161, 5:164–165
advanced pressure transducer, 5:176f
disadvantage, 5:176
water potential measurement, 5:175–181
- Tensorial stress strain, soil structure changes, 5:302–303
- Tenure rights, 3:541
- TERENO-SOILCan, lysimeters, 5:673f
- Termite mounds physical functioning, bioturbation, 5:109–112, 5:111f
- Termites, 4:357–358, 4:358–359f
- Terraced wadi system, 5:601–602
- Terraces, 3:202, 3:202–203f
- Terracing, 3:451, 3:451f
the Andes, 3:454
in Asia, 3:453
in Latin America, 3:454
in Mexican Mixteca Alta UNESCO Geopark, 3:454
- Terra firme, 4:752ge
- Terra preta concept, 4:362
- Terrestrial biosphere model, 5:539
- Terrestrial cosmogenic nuclide (TCN), 4:743
- Terrestrial ecosystem, 5:625
- Terrestrial flux, 2:237–238
- Terrestrial longwave radiation, 5:397–398
- Terrestrial microalgae, 1:61
- Terrestrial nitrogen (N) cycle, 1:243
- Terrestrial Observation Panel for Climate (TOPC), 4:591–592
- Terrestrial System Modelling Platform (TerrSysMP), 5:478–479
- Terrestrial water balance and evapotranspiration, 4:611
- Terroir
food production, 3:516
geophagy, 3:516, 3:516f
picacism, 3:516–517
- Tertiary wastewater treatments systems, 1:569
- Tessier sequential extraction procedures, metal, 2:53–54, 2:54f
- Testing re-entrained aerosol kinetic emissions from roads (TRAKER) and EMMA, 3:228
- Tetrahedral aluminum, 2:168, 2:170f, 2:174–175
- Tetramethylpiperidin-1-oxyl (TEMPO), 2:539–540
- Textural classification system, 5:41, 5:42f
- Textural porosity, 5:18
- Texture analysis
air and fluids storage and exchange, 5:44
bulk densities and porosities, 5:36t
chemical properties and response to environment, 5:44
climatic effects, 5:45
data conversion, 5:41–42, 5:43f
definition, 5:36
drainage capacity, 5:45
hydraulic conductivity, 5:45
mechanical properties, 5:45
particle size and textural classification schemes, 5:40–44
pore-size distribution (PSD), 5:44
significance, 5:44–45
soil hydraulic properties, 5:44–45
soil particle size measurement techniques, 5:36–40
textural classification system, 5:41, 5:42f
tumble effect, 5:45
unification of soil particle-size measurement, 5:43–44
- Texture, soil, 4:153
- Thaumarchaeota, 1:28
- Thermal analysis, 2:662
- Thermal infrared sensors
physical basis, 4:603
stationary, 4:604–605
thermal infrared (TIR) camera, 4:604, 4:604–605f
- Thermally altered C, 2:65–66
- Thermal properties, sensor technology for, 5:173
- Thermal radiation, 4:600–601
- Thermal soil remediation, 3:341–342
- Thermal waste-to-energy (WtE), 3:327–328
- Thermo-diffusive process, 5:267
- Thermodynamic constants, EPR parameters, 2:531
- Thermodynamic cycle of sorption free energy, 2:495f
- Thermodynamics
constant, 2:174–175
hydrostatics of swelling soils, 5:264–265
 L_{DE} and L_{ED} , 5:267
“linearity” argument, 5:267–268
linear non-equilibrium, 5:268
 L_{TD} and L_{DT} , 5:267
 L_{TE} and L_{ET} , 5:267
 L_{TV} and L_{VT} , 5:266–267
 L_{VE} and L_{EV} , 5:267
Maxwell relations, 5:264–265
vs. mechanistic theory, 5:268
modern development, 5:268
non-equilibrium, 5:268
non-equilibrium thermodynamics of soil water, 5:265–267
soil water, classical (equilibrium) of, 5:263–265
tensiometer, 5:264
- Thermo-electric process, 5:267
- Thermo-filtration, 5:266–267
- Thermo-osmosis, 5:266–267
- Therophyte, 3:440ge
- Thiabendazole (TBZ), 2:504–506
- Thiram in soil, 2:503, 2:504f
- Thorium (Th), 2:214–215, 2:229
- Thornthwaite approach, 4:334
- Three-dimensional air flow, 5:296
- 3D digital structure model
extraction from, 5:30
multi-flow simulation in, 5:27
pore-network extraction procedure, 5:27f
porous space, 5:27
segmented pore space, 5:28
stochastic reconstruction techniques, 5:27
structure of soil sample, 5:31
- Thresholds in soil genesis, 4:76–77
- Thymidine and leucine incorporation, 1:163
- Tidal saltmarshes, 4:406
- Tillage systems, 1:238, 3:249–251, 3:250f, 3:444–445, 3:445t, 4:379, 4:577, 5:10–11
aims of, 4:346
at Aspeberget, Sweden, Bronze Age, 3:437f
characteristics, 3:435t
crop production, 3:433–434
definition, 3:433–436
diversified soil, 3:140–141
intensity, 3:434f
in long-term tillage, 3:435f
microbes, 1:454–456
parameters of microbial functioning, 1:457t
regimes, archaea, 1:457t
soil fragmentation, 3:434–437
soil inversion, 3:434–436
on soil physical properties, 3:434f
soil respiration, 1:375
systems approach, 3:140
tilth, 5:50–51, 5:51f, 5:53
with various tillage intensity, 3:434f
water range for, 3:438, 3:438t
- Tilth, 5:83. *See also* Soil tilth
- Time, 4:11, 4:12f, 4:322–323
and soil formation, 4:51–52
soil properties variation, 4:49–51
- Time correlated single photon counting (TCSPC), 2:550–551
- Time-domain reflectometry (TDR), 4:585
bulk electrical conductivity measurement using, 5:442
instrumentation and wave-guides, 5:440–441, 5:441f
measurement of soil water content using, 5:437–439, 5:438f
salinity impact on water content measurement, 5:441
SM sensor, 4:593, 4:595–596
water content, 5:171–173
impact on pore-water salinity measurement, 5:442
- Time domain transmission (TDT), 5:475

- Time-fractional advection-dispersion equation (t-FADE), 5:663–664
- Time-of-flight (TOF), 2:671
- Time resolved fluorescence (TRF), 2:546, 2:553
- Time-variant lower boundary conditions, in soil(-scape) models, 4:747
- Time-variant upper boundary conditions, in soil(-scape) models, 4:746–747, 4:746–747t
- Titanium oxides, 4:146–147, 4:146t
- Titanomagnetites, 4:147
- Todorokite, 4:143
- Tolerant crops, 3:120–121, 3:121t
- Topography and hydrology, SUITMAs, 4:239
- Torrands, 4:226
- Torrerts, 4:227
- Torroxs, 4:227
- Total available water (TAW), 5:509, 5:513
as ecological indicators, 5:514
in hydrological models, 5:514
in irrigation management, 5:513
- Total cation concentration (TCC), 4:378
- Total dissolved solids (TDS), 5:367–368
- Total soil metagenomes, 1:16ge
- Total stress, 5:73
- Total water potential, 5:163–164
- Tower and airborne systems, 4:619
- Toxic metals, immobilization of, 3:418–419, 3:418f
- Trace elements, 3:168–169
characteristics, 3:289t
contaminated soil environments, 3:292t
definition, 3:290
energy production industries, 3:293
fertilizers, 3:291
manufacturing and recycling industries, 3:293
mining processes, 3:292
organic manures, 3:291–292
pesticides, 3:290
soils and potential risk pathways, 3:293–294, 3:293t
wastewater, 3:292
- Trace metals binding, root mucilage, 1:338
- Tracing experiments, 2:83
- Tradeoffs of wastewater irrigation, 3:280, 3:280f
- Traditional slope management, 3:452
- Traits-based strategy, soil organic carbon determination, 2:394, 2:394f
- Transdisciplinary soil studies, 3:29
- Transfer function model, 5:359
- Transformations, 4:4, 4:492–493, 4:493–494f
of organic matter, SUITMAs, 4:241
of precipitated soil minerals, 2:292
quality, sulfur, 1:308–309
in soil-plant system, 2:204–205
- Transient flow
colloid retention, 5:345, 5:346f
through saturated medium, 5:258–259, 5:259f
through unsaturated medium, 5:261
- Transient infiltration
at constant pressure head, 5:650
cumulative infiltration, 5:648–649
vs. evapotranspiration, 5:650–651, 5:651f
horizontal infiltration, 5:648, 5:648f
hydraulic head, 5:647–648
infiltration delays, 5:649–650, 5:650f
infiltration flux, 5:648
multi-dimensional water flow, 5:650
vertical infiltration, 5:648–649, 5:649f
wetting front, 5:647, 5:647f
- Transition metals in soil
cadmium, 2:139
chromium, 2:137
cobalt, 2:138
copper, 2:138–139
definition, 2:135–136
environmental fate, 2:136–137
gold, 2:140
iridium, 2:140
iron, 2:138
manganese, 2:137–138
mercury, 2:140
nickel, 2:138
osmium, 2:140
palladium, 2:139
periodic table, 2:141f
platinum, 2:140
scandium, 2:137
silver, 2:139
sources, 2:136
tantalum, 2:140
technetium, 2:139
terminology, 2:136
titanium, 2:137
trace element deficiencies, 2:137
tungsten, 2:140
vanadium, 2:137
- Translocation, 4:4, 4:48ge
- Transmission electron microscopy (TEM), 2:577, 2:577f
instrumentation, 2:578, 2:579f
mineral-associated organo-mineral (MAOM), 1:401
strengths and limitations, 2:579–580
- Transmission-mode Fourier transform infrared spectroscopy, 2:513f, 2:514
- Transmissivity, 5:366
- Transparent/lear plastic mulch, 3:397
- Transpiration, 4:48ge, 4:610–612, 5:639–640
- Transport
arsenic, 2:204
colloid, 5:350
dissolved organic carbon, 2:426–427
kinetics, 2:353
limitation, 4:79
in soil-plant system, 2:204–205
- Transport, pore-scale modelling
application, 5:30
existing codes, 5:30–31
homogenization, upscaling and beyond, 5:31–33
input data and characterization, 5:26–27
numerical methods, 5:28–30
pore-network extraction procedure, 5:27f
resources, 5:32t
- Treatment plants, wastewater treatments systems, 1:568–569
- Tree-based intercropping, 3:81–83, 3:82f
- Tree based system, 3:140
- Trees, 1:472, 1:474t
- Tree woodlots and specialty crops, 3:78
- Triangular silver nanoparticles (TSNPs), 2:503
- Triaxial testing, 5:116f, 5:117–118, 5:118t
- Trichoderma, 1:424–425
- Tridymite, 4:93
- Triple collocation analysis (TCA), 4:627
- Triple superphosphate fertilizer, 3:371
- Tropical moist and wet forests, 4:759
- Tropical soil. *See* Humid and sub-humid tropical soils
- Tropical soil systems, 4:28–29
- Tuller-Or (TO) model, 2:28
- Tundra, 4:32, 4:387–388
- Turbels, 4:226
- Turbic cryosol profile, 4:286f
- Turf
brown/straw colored (dormant) lawn-turf strategies, 3:499–500
dormant, 3:499–500
management, 3:497, 3:502
pitch's surface, 3:493
- Turfgrass drought survival, 3:498–502
control root-related stresses, 3:499
management, 3:499
mowing and nutrition, 3:499
water/irrigation efficiency, 3:499
- Two-chamber diffusion apparatus, 5:251
- Two-dimensional correlation spectroscopy (2D-COS), 2:549
- Two-dimensional (2-D) ^1H - ^{13}C heteronuclear correlation (HETCOR) NMR experiment, 2:437
- Two-phase flow soil experiments, 5:25–26
- Two-region advection-dispersion equation, 5:660–661
- 2:1-type clay minerals
chlorite, 4:117
hydroxy-interlayered vermiculite and smectite, 4:117
interstratified phyllosilicates, 4:117–118
mica, 4:115–116
palygorskite and sepiolite, 4:118, 4:118f
pyrophyllite, 4:115
smectite, 4:116–117
2:1 layer structure, 4:115
vermiculite, 4:116
- ## U
- Udalfs, 4:229
- Udands, 4:227
- Udepts, 4:229
- Uderts, 4:227
- Udolls, 4:229
- Udoxs, 4:227
- Udults, 4:229
- Ultisol, 4:55ge, 4:229, 4:296, 4:328, 4:464
- Ultrasonic vibration, soil dispersion, 2:74
- Ultraviolet radiation, biocrusts, 1:132–133
- U-manometer tensiometers, 5:175–176

- Uncertainty assessment
 added value of uncertainty information, 4:680–682
 change of support, 4:679
 cross-correlation, 4:677
 errors and uncertainty in soil class maps, 4:677–679
 geostatistics and machine learning, 4:675–676
 measurement error standard deviation
 from replicates, 4:674, 4:674f
 multivariate distribution of soil map errors, 4:676–677
 positional uncertainty, 4:677
 single soil measurements, 4:672
 spatial soil information, 4:675
 spatial support and resolution, 4:679
 statistical model of errors in spatial soil variables, 4:675
 statistical model of measurement error, 4:672–674
 uncertainty of categorical soil variables, 4:677
 uncertainty propagation, 4:680
- Uncertainty of categorical soil variables, 4:677
- Uncertainty propagation, 4:680
- Uncontrolled wildfires, 2:434
- UN Convention on Biological Diversity (UNCBD), 3:459
- UN Convention to Combat Desertification (UNCCD), 3:537
- Uncorrected matrix correlation (UMC)
 method, 2:547
- Under-nutrition, 3:13
- Undersowing, 3:447
- UNESCO Global Geoparks, 3:454
- UN Framework Convention on Climate Change (UNFCCC), 3:537
- Unified diffusivity modeling, 5:253–254
- United Nations Convention to Combat Desertification (UNCCD), 3:533
- United Nations Sustainable Development Goals (SDGs), 3:64f
- Units, soil, 4:487
- Universal kriging, 4:479, 4:669
- Universal solvent, 5:152. *See also* Water properties
- Unmanned aerial systems (UAS)
 area covered and data limitation, 4:639
 bare soil detection, threshold limitation, 4:639
 fixed wing UAS, 4:635f
 flight campaigns, 4:636–637
 infrastructure capabilities, 4:640
 light weight spectrometers and multispectral cameras, 4:635
 mosaic images, 4:636
 parsimonious approach, 4:636
 spatial resolution, 4:635
 UAS-borne Tetracam Mini-MCA6 camera, 4:635–636
- Unpaved roads, dust emissions, 3:224
- Unsaturated hydraulic conductivity
 Darcy's law, 5:381
 functional models, 5:382
 relative hydraulic conductivity, 5:382
 scaling factor, 5:381
 soil water diffusivity, 5:382
- Unsaturated soil, 5:316
- UNsaturated SOil hydraulic DATABASE (UNSODA), 5:190–191
- Unstable preferential flow, porous media
 checking stability criteria using
 experimental evidence, 5:317
 continuum models, 5:319
 dynamic contact angle, 5:317–318
 finger flow pathways, 5:316f
 fingers formation and development, 5:315f
 modeling gravity-driven fingers fingered flow, 5:318–319
 observations, 5:315–316
 pore network models, 5:319
 semi-continuum modeling, 5:319
 wetting front instability, conditions for, 5:316
- Unsupervised classification
 application, 4:505, 4:506–507f
 fuzzy clustering approach, 4:506
k-means clustering, 4:505
- Unsupervised learning, 4:510–511
 clustering, 4:516
 dimension reduction, 4:517
- Upscaling, pore-scale modelling, 5:31–33
- Uranium, 2:229, 2:230f
²³⁸U, decay chain, 2:261f
- Urban forests, soils in, 3:65, 3:66f
- Urbanization, 3:474, 3:486
- Urban land and classifications, 3:583f
- Urban soil microbiome
 Arctic (Apatity, Russia), 1:487–490, 1:490f
 factors influencing, 1:482f
 impact of, 1:480
 from landscape variability to soil microbial diversity, 1:485–487, 1:485f
 map of soil sampling sites, RUDN University, 1:483f
 microbial carbon indicators of Technosols, 1:489f
 microbial community characteristics, 1:481t
 topsoil (0–10 cm) sampling campaign, 1:483, 1:484t, 1:484f
 total microbial biomass, 1:488f
 typology and composition, 1:482–485, 1:482f
- Urban soil mosaic, 3:44, 3:46f
- Urban soil stratification, SUITMAs, 4:236
- Urea, 3:371
- Urea hydrolysis, 3:369ge
- USA Environmental Protection Agency 4 step risk assessment process, 3:226f
- US Environmental Protection Agency (USEPA) priority list, 2:218
- Ustalfs, 4:229
- Ustands, 4:227
- Ustepts, 4:229
- Usterts, 4:227
- Ustolls, 4:229
- Ustoxs, 4:227
- Ustults, 4:229
- V**
- Vadose zone, 5:249, 5:315
- Validation, 4:619
- Valley of Colca River, 3:454
- Vanadyl ion, 2:528
- Van der Waals forces, 1:390
- Van Genuchten model parameters, 5:191t
- Van Krevelen diagram, 2:434, 2:435f
- Vanselow equation, 2:36
- Vapor pressure, 2:471, 5:154–155, 5:155t
- Variability and environmental dependencies, water use efficiency, 5:533–537, 5:534–536f
- Variable rate irrigation (VRI) practices, 5:609
- Variable source area (VSA), 5:288, 5:288–289f
- Variance, 4:490–491, 4:673–675
- Variograms, 4:478, 4:524, 4:525f.
See also Spatial scale
 estimating, 4:658–659, 4:658f
 models for, 4:659–661, 4:659f
- Vegetation, 2:238, 3:215, 4:321–322, 5:64–65
 alpine and plateau soils, 4:275
 Antarctic Cold Desert, 4:274
 Antarctic Polar Desert, 4:273
 boreal and subarctic regions, 4:264
 clearance, 3:444–445
 emission and scattering, 4:623–624, 4:623f
 fires, 2:433–434, 2:443, 2:448–449
 high Arctic soils, 4:269
 impact of, 3:110
 low Arctic soils, 4:267
 rangelands, 3:96–98
 residues, 3:218–219, 3:219f
 scattering, 4:622–623
 subantarctic forest and tundra, 4:272
 succession, 4:77
 and vegetative residues, 3:218–219, 3:219f
- Vegetation index (VI), 5:503–504
- Vermicomposting, 1:527–528
- Vermiculite, 4:116
- Vernacular soil, 4:186
- Vernadite, 4:142
- Vertebrate nematode parasites, 1:105
- Vertical filling of soil monoliths, lysimeters, 5:672f
- Vertical green systems, 3:477f. *See also* Green roofs
 constructed technosols, 3:478–482
 green roofs and, 3:474–478
 intraporous structure, 3:478
 physical requirements, 3:478–481
 technical aspects, 3:476–478
 water holding capacity and drainage, 3:478–480
- Vertical profile, 3:229, 3:230f
- Vertisols, 2:264–265, 4:227, 4:328, 4:461, 4:462t, 5:139–140
- Vienna soil organic matter modeler (VSOMM), 2:491–492
- Viral ecology, 1:16ge, 1:17
- Viral lysis process, 1:17
- Viral operational taxonomic units (vOTU), 1:17–18

- Viral shunt, 1:16*ge*, 1:20
 Virgin Compression Curve, 5:86, 5:87*f*
 Viromes, 1:16*ge*
 Virulent phages, 1:18
 Viruses
 bacteriophages (*see* Bacteriophages (viruses))
 extreme soil environments, 1:506–507
 Viscosity of water, 5:157–158, 5:158*t*
 Visible, near and mid-infrared diffuse reflectance spectroscopy, 4:585–586
 Visible-near-infrared (VNIR) spectroscopy, 2:568*f*, 2:569, 4:702, 4:704*f*
 Visual evaluation of soil structure (VESS), 3:184, 5:53–54
 Vitrandes, 4:226
 Vitric andosols, 4:460–461
 Vitric soil properties, 4:458*ge*
 Vivianite arsenates in soil, 2:506
 V-nitrogenase, 1:271
 Volatile organochlorines (VOCs), 2:238
 Volatilization, 2:198, 2:236
 of chlorine, 2:238, 2:241
 of iodine, 2:240
 pesticide, 2:474
 wastewater treatments systems, 1:573
 Volcanic rocks, 4:460*t*
 Volcanic soils, 2:423
 human interactions with, 4:469
 microbiology, 1:505
 Volume of fluid method (VOF), 5:197
 Volumetric samples
 clod method, 4:558
 cylinder method/core method, 4:555, 4:556*f*
 moist soil conditions, 4:554
 sand method, 4:558, 4:558*f*
 water method/excavation method, 4:555–558, 4:557*f*
 Volumetric water content, 5:169
 Volume wetness, 5:169
- W**
- War, 4:66, 4:66*f*
 Warming effects, 1:583, 2:695–696
 Wassents, 4:230
 Wassists, 4:226
 Waste deposit, 3:10*f*
 Wastes from cattle sheds, as mulch, 3:394
 Waste-to-energy (WtE), 3:327–328
 Wastewater
 quality, 3:282
 salts in, 3:282–283
 use for irrigation, 3:278–280, 3:279*f*
 Wastewater treatment plants (WWTP), 3:322
 Wastewater treatments systems
 adsorption and precipitation, 1:573
 biological mechanisms, 1:573
 chemical reactions, 1:573
 contaminants, 1:574–577
 decentralized/on-site, 1:568
 disinfection by-products, 1:574–575
 endocrine disruptors, 1:575
 fate of disinfection by-products (DBP) during rapid infiltration, 1:575*t*
 fate of pharmaceuticals after rapid infiltration, 1:575*t*
 filtration, 1:572–573
 land application of effluent from, 1:570–571
 metals and trace elements, 1:574, 1:574*t*
 microbial transport, 1:577*t*
 microorganisms, 1:575–577
 organic compounds, 1:574
 overland-flow process, 1:572
 pathogen survival in soils, 1:576*t*
 pharmaceuticals, 1:575
 primary, 1:568
 rapid-infiltration process, 1:571–572
 secondary, 1:568–569
 sewage effluent, 1:571*t*
 tertiary, 1:569
 treatment mechanisms, 1:572–573
 treatment plants, 1:568–569
 volatilization, 1:573
 WATEQ model, 2:279
 Water, 1:410–411, 1:410*f*
 and air quality, 3:181
 and air regulation, 3:253*f*, 3:254
 and biogeochemical cycling, 3:63*f*
 environmental components, 3:10
 erosion, 1:473
 holding capacity, 1:470
 quality, 1:473
 solubility, 2:471
 spreading, 5:594, 5:596*f*
 storage, 3:263, 3:264*t*
 streams, 1:190
 supply, 3:186–188
 tables, evaporation, 5:495
 transport, lysimeters, 5:674–675
 Water balance, 5:491
 accuracy of, 4:615
 based on proximal sensing, 4:615–616
 components, 4:611–612
 and evapotranspiration, terrestrial, 4:611
 lysimeters, 5:668
 Water balance index (WABI), 5:521
 Water band index (WBI), 5:521
 Water Cloud Model (WCM), 4:622–623
 Water content
 capacitance probes, 5:173
 definition, 5:169
 frequency domain reflectometry (FDR), 5:173
 gravimetric measurements, 5:170–171
 measurement of, 5:170–174
 noninvasive techniques, 5:174
 pressure head, 5:170
 reference method, 5:170–171
 sensors, 5:172*f*, 5:182–184*t*
 dielectric properties, 5:171–174
 electric conductivity, 5:173
 time domain reflectometry (TDR), 5:171–173
 volumetric measurements, 5:171
 Water cycles, 3:7–8, 5:377, 5:530
 Water deficiency, 4:59–60
 Water drop penetration time (WDPT) test, 5:207
 Water entry pressure, 5:207–208
 Water-ethanol droplet (MED) test, 5:207
 Water extracted organic carbon (WEOC), 2:422
 Water flow in sodic soils, 4:373–377
 Water fluxes in soils, 4:77
 Water from bare soil, evaporation, 5:490
 Water harvesting, 3:117, 5:593, 5:597*f*
 categories, 5:597–604
 components, 5:595, 5:598*f*
 concepts, 5:595
 earthwork/stonework, 5:605*t*
 environmental and socio-economic benefits, 5:604–605
 history, 5:594–595
 and hydrological processes, 5:595–597
 principles, 5:595, 5:598*t*
 technology, 5:594
 Water induced soil erosion
 control measures, 3:201–205
 factors affecting, 3:197–200
 financial value, 3:197
 hillslopes, 3:194
 land management practices, 3:200
 land use and land cover, 3:199–200, 3:200*f*, 3:201*t*
 local, regional and global rates, 3:194
 measurement and monitoring, 3:194
 off-site consequences, 3:196, 3:197*f*
 on-site consequences, 3:196, 3:196*f*
 process, 3:194–195, 3:195*f*
 rainfall erosivity, 3:197–198
 slope steepness and slope length, 3:199
 soil erodibility, 3:195*f*, 3:198–199
 total costs, 3:197
 types, 3:197, 3:198*f*, 3:198*t*
 Water-ion interaction
 adsorption, 5:155–156
 ionization and pH, 5:151–152
 osmotic pressure, 5:153–154, 5:153*f*
 solubility of gases, 5:154, 5:154*t*
 solvent properties of water, 5:152–153, 5:152*f*, 5:153*t*
 surface tension, 5:156, 5:156*f*
 vapor pressure, 5:154–155, 5:155*t*
 Waterlogged/flooded soils, 1:327
 Waterlogging, environmental effects of irrigation, 3:270, 3:271*f*
 Water phase diagram, 5:150–151, 5:150*f*
 Water-plant relations. *See* Plant-water relations
 Water potential, 5:518, 5:519*f*
 advanced pressure transducer tensiometers, 5:176*f*
 components, 5:161–163
 definition, 5:160–161, 5:170
 dry range, 5:178–180
 energy forms, 5:159
 field thermocouple psychrometer, 5:180*f*
 historical perspective, 5:159–160
 matric potential, 5:164–166
 measurement, 5:174–181
 osmotic potential, 5:166–167
 pressure potential, 5:164

- Water potential (*Continued*)
 psychrometers, 5:178
 sensors, 5:179f
 solid-phase sensors (SPS), 5:177–178
 tensiometers, 5:175–181
 total water potential, 5:163–164
 units, 5:171t
 WP4C Dewpoint PotentiaMeter, 5:181f
- Water properties
 density and compressibility, 5:156–157
 dynamic and kinematic viscosity, 5:157–158, 5:158t
 of free water, 5:156–157
 hydrogen bonding, 5:148–149, 5:149f
 importance of water, 5:146–147
 liquid water molecule model, 5:147–148b, 5:148f
 molecular structure, 5:147–148, 5:148f
 of soil water, 5:157, 5:157f
 states of water, 5:149–151
 water-ion interaction, 5:151–154
- Water purification and regulation, 3:149
- Water quality
 bioreactors, 3:263
 buffers and filters, 3:262
 cover crops, 3:262
 drainage water recycling, 3:263
 enhancement and nutrient interception, 3:79
 impacts, saturation excess overland flow, 5:292
 by salinity, 3:237–238, 3:238f
 saturated buffers, 3:263
 water storage, 3:263, 3:264t
 wetlands, 3:262
- Water range, tillage systems, 3:438, 3:438t
- Water regime control, 3:9
- Water relative permeability, 5:281–282
- Water repellency, 5:317
- Water retention, root mucilage, 1:337
- Water scarcity, 3:279, 5:594
- Watershed management, 5:472–473
- Watershed nutrient dynamics and hydrology, 3:62–63
- Watershed storage considerations, 5:289–291
- Water shortage, environmental effects of irrigation, 3:268–270, 3:268–269f
- Water-soluble organic compounds (WSOCs), 3:422
- Water table/phreatic surface, 4:752ge
- Water use efficiency (WUE)
 in agronomy, 5:529
 crop water productivity, 5:531–533, 5:532t, 5:532f
 definitions, 5:528–529
 for linking water and carbon cycles, 5:530
 observations and simulations, 5:530–531
 stomatal conductance modeling and, 5:537
 trends in, 5:537–539, 5:538f
 values of, 5:528f
 variability and environmental dependencies, 5:533–537, 5:534–536f
- Water vapor
 density, 5:387
 flux, 5:387
 sorption, 2:27–29, 2:28–29f
- Wave function methods, 2:487
- Weather, forecast data sets, 5:643
- Weathering, 2:177, 2:284, 3:59, 4:2–5, 4:77, 4:78f, 4:743
 chemical, 4:38, 4:337–338
 feldspars, 4:99–100
 fluorine in, 2:236
 olivines, 4:102–103
 physical, 4:38
 pyroxenes and amphiboles, 4:106–107
 rate of dissolution, 4:99f
 of rock fragments, 3:64
- Weed
 control, herbicides for, 1:463–464
 definition, 3:441
 management, 3:440, 3:442
 and soil management interactions, 3:443–445
- Weed seedbank, 3:440–441, 3:440f, 3:443–444, 3:444f
- Weighing lysimeter, 5:502, 5:668, 5:678
- Wet dust sampler (WDS), 3:228
- Wetland rice. *See* Paddy soils
- Wetlands, in humid cool temperate zone, 4:292–293
- Wetland soils, 1:379–380, 4:424–425, 4:752ge. *See also* Anaerobic soils
 chemistry, 4:439–440
 Cowardin classification, 4:438
 definition, 4:437
 ecological character, 4:437
 ecosystem services, 4:438–439, 4:439t
 ferrolysis process, 4:442f
 gleyic color pattern, 4:442f
 gradient, 4:441
 human perception, 4:437
 hydric soils, 4:443
 hydrology, 4:444
 hydrophytic vegetation, 1:381
 indicators, 4:445
 inland and coastal, 4:437
 iron and manganese minerals, 4:440
 juxtaposition, 4:442
 land surface model, 4:759–760
 measurement, 4:444–445
 mottles, 4:442
 nodules and concretions, 4:442–443
 organic matter, 4:439
 organic matter decomposition, 1:383f
 oxidation-reduction potential, 1:381f, 1:384f
 Ramsar Convention, 4:437–438, 4:438t
 redox concentrations and depletions, 4:440
 redox reactions, microbial groups, 1:383t
 reducing and oxidizing half reactions, 4:439, 4:439t
 reducing conditions, 4:440
 reductants accumulation, 1:384f
 simplified oxidation half reaction, 4:439
 stagnic color pattern, 4:441f
 traditional attitudes, 4:437
 vegetation, 4:437
- Wettability, 5:204–205, 5:236–237
- Wettable silt soil surface, 5:228f
- White plastic mulch, 3:397
- Wildfires, 5:206
- Wildlife, rangeland management, 3:92
- Wilhelmy plate method (WPM), 5:208–209
- Willis's curve, 4:731–732
 broad universe of, 4:735–736
 and long tailed distributions, 4:736
- Wilting coefficient, 5:511
- Wind, 3:212
- Windblown deposits, 4:36
- Wind erosion, 3:209–210, 3:209f, 3:227–228
 barriers, 3:219–220, 3:219–220f
 chemical treatments, 3:221
 control, 3:218–221
 effects of, 3:210, 3:212f
 emergency control, 3:221
 measurement, 3:216–218, 3:217–218f
 models, 3:215–216
 processes, 3:210–215
 vulnerability, 3:211f
- Wind erosion assessment model (WEAM), 3:216
- Wind erosion equation (WEQ), 3:215
- Wind erosion prediction system (WEPS), 3:215–216
- Woody plant encroachment (WPE), 3:97–98
- World Climate Research Programme (WCRP), 4:752ge
- World reference base (WRB), 4:198
- World reference base for soil resources (WRB), 3:528, 4:186
 application, 4:215–216
 buried soils, 4:214
 development, 4:207
 diagnostics, 4:208, 4:210–211t, 4:212f
 in education, 4:216
 field workshop, 4:213–214, 4:213f
 major changes in WRB 2022 edition, 4:214–215
 map legends, 4:216
 naming soils, 4:208–214
 polygenetic soils, 4:214f
 qualifiers, 4:207
 reference soil groups, 4:207, 4:208f, 4:209t
 and soil taxonomy, 4:215
 structure, 4:207–208
- World soil map, 4:221f
- Worms, 1:2
- WP4C Dewpoint PotentiaMeter, 5:181f
- WP4C psychrometer, 5:164f

X

- Xanthization, 4:349
- Xanthophyceae, 1:65
- Xenobiotic organic compounds, 2:64–65
- Xeralfs, 4:229
- Xerands, 4:226
- Xerepts, 4:229
- Xererts, 4:227
- Xerolls, 4:229
- Xerults, 4:229
- X-ray, 2:643
- X-ray absorption fine structure spectroscopy (XAFS) technique, 2:346
- X-ray computed tomography (XCT)
 detector, 5:462
 image analysis, 5:462–463
 industrial, 5:460–461

manipulator stage, 5:462
 medical, 5:460–461
 principles, 5:460
 synchrotron based, 5:460–461
 types, 5:460–461, 5:461*f*
 X-ray source, 5:461–462
 X-ray diffraction (XRD), 2:642–643
 analysis for soil mineral qualification,
 2:646
 application in soil evolution, 2:648–651
 based quantification methods for soil
 minerals, 2:646–648
 factors affecting, signal strength, 2:646
 instrumentation, 2:644–645
 intensity of diffraction line, 2:645

X-ray fluorescence spectroscopy, 2:550
 X-ray μ -CT images, of stable macroaggregates,
 5:11
 X-ray photoelectron spectroscopy (XPS),
 1:400, 2:483–484
 X-ray tomography (XCT) imaging, 5:26, 5:28

Y

Yield-acquisition-stress (Y-A-S) strategies,
 1:34, 1:35*t*
 Young–Laplace equation, 5:188, 5:229,
 5:275
 Young soils, 2:131

Z

Zeeman effect, 2:534
 Zeolites, 4:119
 Zero point of charge (ZPC), 2:41
 Zero tension lysimeters, 5:327
 Zero-tillage arable soils, 5:50–51, 5:51*f*
 Zinc (Zn), 2:215
 deficiency, 3:562
 pollution, 2:113
 in soil, 2:507
 Zonal-azonal-intrazonal soils, 4:249
 Zonal soils, 4:14–15, 4:188–189, 4:399
 Zymography, 1:354

Encyclopedia of SOILS IN THE ENVIRONMENT

Volume Five: Soil Physics

EDITORS-IN-CHIEF

Michael J Goss

Professor Emeritus, School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G 2W1, CANADA

Margaret Oliver

Visiting professor, Soil Research Centre, Department of Geography and Environmental Science, Russell Building, University of Reading, Whiteknights, Reading, RG6 6DR, United Kingdom.

SECTION EDITORS

Dr Evgenia Blagodatskaya

Dept. of Soil Ecology, Helmholtz-Zentrum für Umweltforschung GmbH – UFZ, Theodor-Lieser Str. 4, D-06120 Halle (Saale), GERMANY

Professor Adrian Unc

Boreal Ecosystems and Agricultural Sciences, Grenfell Campus, Memorial University of Newfoundland, 20 University Drive, Corner Brook, Newfoundland A2H 5G4, CANADA

Dr. Siobhán Staunton

INRA, UMR Eco&Sols, Montpellier SupAgro, bâtiment 12, 2 Place Viala, 34060 Montpellier Cedex 2, Montpellier, FRANCE

Professor Baoshan Xing

University of Massachusetts – Amherst, College of Natural Sciences, Stockbridge School of Agriculture, Environmental & Soil Chemistry, Program, 410 Paige Laboratory, Amherst, MA 01003, USA

Dr Karin Mueller

New Zealand Institute for Plant & Food Research Ltd., Ruakura Research Centre, Private Bag 3123, Hamilton, New Zealand

Professor Christina Siebe

Instituto de Geología, Circuito de la Investigación Científica s/n, Ciudad Universitaria, Ciudad de México, CP 04510, México

Professor Rupert Baeumler

Institute of Geography, Friedrich-Alexander-University of Erlangen-Nuremberg, Wetterkreuz 15, 91058 Erlangen, Germany

Dr Uta Stockmann

Soil Processes and Function Organisational Unit, CSIRO Agriculture and Food, Black Mountain, GPO BOX 1700, Canberra ACT 2601, AUSTRALIA

Professor Paul Hallett

Room 2.23, Cruickshank Building, School of Biological Sciences, University of Aberdeen, Aberdeen, AB24 3UU, UNITED KINGDOM

Professor Dani Or

Department of Environmental Systems Science (D-USYS), Institute of Biogeochemistry and Pollutant Dynamics (IBP), Soil and Terrestrial Environmental Physics (STEP), CHN F 29.1 Universitaetstrasse 16, 8092 Zurich, SWITZERLAND



ELSEVIER

elsevier.com/books-and-journals

Set ISBN



Volume ISBN

